Journal of Medical Internet Research

Journal Impact Factor (JIF) (2023): 5.8
Volume 28 (2026) ISSN 1438-8871 Editor in Chief: Gunther Eysenbach, MD, MPH

Contents

Original Papers

Explainable and Interpretable AI for Voice and Speech Analysis in Clinical Care: Systematic Review (e83790)
Mohamed Ebraheem, Jamie Toghranegar, Bridge2AI-Voice Consortium, Yael Bensoussan, John Templeton. 57

Securing Federated Learning With Blockchain in the Medical Field: Systematic Literature Review (e79052)
Xudong Wang, Yi Xie, Xiaoliang Chen, Jiaming Yang, Ruiyuan Li, Weihang Gao, Zineng Yan, Hong Zhou, Zhewei Ye. 1861

The Phases of Living Evidence Synthesis Using AI: Living Evidence Synthesis (Version 1) (e76130)
Xuping Song, Zhenjie Lian, Rui Wang, Ruixin Li, Zhenzhen Yang, Xufei Luo, Lei Feng, Zhiming Ma, Zhen Pu, Qi Wang, Long Ge, Caihong Li, Yaolong Chen, Kehu Yang, John Lavis. 2050

Artificial Intelligence, Connected Care, and Enabling Digital Health Technologies in Rare Diseases With a Focus on Lysosomal Storage Disorders: Scoping Review (e73612)
Alberta Spreafico, Luca Neri, Kim Angel, Cinzia Bellettato, Maurizio Scarpa. 2946

Approaches to Collect Comprehensive Electronic Patient Data Across Multiple Providers and Payers for Research: Landscape Analysis (e86330)
Robert Rudin, Nihar Chhatiawala, Joshua Mandel, William Gordon, David McCallie, Daniel Gottlieb. 3887

Association Between Telemedicine Adoption and Physician Job Satisfaction: Cross-Sectional Study (e82285)
Xiaohui Zhai, Zhongliang Zhou, Zhichao Wang, Guanping Liu, Peter Coyte. 3903

Facilitators of and Barriers to Global Digital Oral Health: Mixed Methods Study (e76236)
Elham Emami, Pascaline Kengne Talla, Camille Inquimbert, Nicolas Giraudeau. 3918

User Profiles and Engagement in a Hypertension Self-Management App: Cross-Sectional Survey (e83075)
Felix Muehlensiepen, Susann May, Frances Seifert, Eileen Wengemuth, Olen Johannsen, Martin Middeke, Martin Heinze, Pascal Petit, Nicolas Vuillerme, Sebastian Spethmann, Dunja Bruch. 3929

Associations Between Short-Video Platform Use and Health Across Health Distribution and Usage Behaviors in China: Cross-Sectional Questionnaire Study (e86526)
Yangyang Pan, Kangkang Zhang, Yilin Wei, Yangzhen Huang, Chengxu Long, Chenxin Yang, Shangfeng Tang. 3943

Health Equity Analysis of Awareness and Use of GetCheckedOnline, British Columbia's Digital Intervention for Sexually Transmitted and Blood-Borne Infection Testing in 5 Urban, Suburban, and Rural Communities: Cross-Sectional Survey Study (e78561)
Rodrigo Sierra Rosales, Aidan Ablona, Hsiu-Ju Chang, Devon Haag, Heather Pedersen, Catherine Worthington, Daniel Grace, Rod Knight, Devon Greyson, Mark Gilbert. 3959

XSL•FO
RenderX

Prevalence of Social Media Addiction and Associations With Usage Patterns, Burnout, and Health Conditions Among Medical Trainees in China: Cross-Sectional Study (e75675)
Zexu Guan, Ni Tang, Guoshuai Luo, Xiao Zhang. 3972

Digital Health Literacy, Technology Acceptance, and Competence Among Older Adults Aged ≥65 Years: Cross-Sectional Study Investigating Differences Between Women and Men (e85846)
Franziska Jung, Melanie Luppa, Matthias Reusche, Kerstin Wirkner, Melanie Eberl, Yvonne Dietz, Christoph Engel, Steffi Riedel-Heller. 3988

Structural Inequalities in Online Health Information Seeking: Cross-National Multilevel Study (e88110)
Petra Raudenská, Elena Link. 4002

Women's Engagement With Different Internet-Enabled Technologies to Access Digital Menopause Information: Mixed Methods, Multiphase Sequential Study (e78215)
Alison Osborne, Elizabeth Sillence, Caroline Claisse, Abigail Durrant. 4018

Operationalizing Digital Health Equity in Artificial Intelligence–Enabled Patient Decision Aids for Older Adults: Mixed Methods Study (e89011)
Cindy Tian, Xiaochen Yang, Kailu Wang, Annie Cheung, Jonathan Ma, Canjie Lu, Jasmine Yu, Crystal Chan, Jiamin Chen, Kun Ouyang, Ivan Lin, Tim Pang, Shi Zhao, Yingwei Wang, Eliza Wong. 4033

An Internet-Based Mindfulness- and Compassion-Based Intercare Program for Reducing Parental Burnout: Randomized Controlled Trial (e87416)
Francisco Villalón López, Maximiliano Escaffi-Schwarz. 4049

Understanding University Students' Experiences of Engaging With AI and Apps for Their Mental Health and Well-Being: Qualitative Study (e75381)
Julia Groot, Ben Ainsworth, Nathalia Gjersoe, Christopher Clarke, Richard Hamshaw, Mark Brosnan. 4065

Examining the Association Between Internet Addiction and Nonsuicidal Self-Injury Among Chinese Middle School Students: Prospective Cohort Study (e86427)
Tianqing Fan, Lintong Song, Sihong Li, Xingyue Jing, Mireille Twayigira, Chunxiang Huang, Xuerong Luo, Yanmei Shen. 4084

Promoting Psychological Resilience and Well-Being in Youth With a Smartphone-Based Ecological Momentary mHealth Intervention: Secondary Analysis of a Microrandomized Trial (e85552)
Eva Wierzba, Anita Schick, Christian Rauschenberg, Janik Fechtelpeter, Selina Hiller, Christian Götzl, Daniel Durstewitz, Silvia Krumm, Georgia Koppe, Ulrich Reininghaus. 4095

The Mediating Role of Interpersonal Needs in Perceived Parenting Styles and Social Media Addiction Among University Students: Cross-Sectional Study (e91861)
Yujie Liu, Suping Wang, Ting Wei, Birong Wu, Xin Ge, Shangbin Liu, Chen Xu, Shunyu Tao, Xiaohong Fan, Fan Hu, Ying Wang, Xue Yang, Yong Cai. 4161

A Preventive Social Media Intervention for Perinatal Depression and Anxiety in Regional, Rural, and Remote Communities: Participatory Co-Design Study (e91778)
Kacey Lynch, Adrian Shatte, Jessica Muller, Kendall George, Gisele Rossini, Angela Anson, Courtney Hala, James Dimmock, Delyse Hutchinson, Samantha Teague. 4176

Optimization of University Counseling Consent Forms With Large Language Models: Multidimensional Comparative Evaluation (e86502)
Jianchen Luo, Jing Ma, Danni Zhan, Yuhong Zhou, Jiayu Li, Lan Zhang, Wentao Wang. 4197

Well-Being and Cognitive Factors Influencing Health Care Workers' Adherence to Internet-Based Stress Management: Mixed Methods Analysis of a Nonrandomized Controlled Study (e73306)
Miqi Li, Quan Wang, Danli Zheng, Danni Feng, Xiaorong Lang. 4214

XSL•FO
RenderX

Changes in Workplace Productivity and Estimated Cost Savings During Internet-Based Cognitive Behavioral Therapy in the Irish National Health Service: Naturalistic, Repeated-Measures, Retrospective Survey Study (e80689)
Chi Lee, Siobhán Harty, Angel Enrique, Alba Jiménez-Díaz, Garrett Hisler, Daniel Duffy, Derek Richards. 4228

Goal Setting and Anchoring Effects on Meditation Using a Digital Platform: Large-Scale Digital Field Study (e85801)
Michael Bowen, Michael Beam. 4244

Identifying Evidence-Based Strategies in a Digital Mental Health Intervention for Depression: Qualitative Content Analysis (e84030)
Geneva Jonathan, Jenna Sung, Heyli Arcese, Phoebe Holz, Kathryn Bentley, Emily Bernstein, Hilary Weingarden, Jennifer Greenberg, Oliver Harrison, Sabine Wilhelm. 4255

Long-Term Outcomes, Moderators, and Predictors in Online Mindfulness–Based Cognitive Therapy for People With Cancer: Secondary Analysis of a Randomized Controlled Trial (e79928)
Nasim Badaghi, Jelle Lubbers, Judith Prins, Saskia Kelders, Anne Speckens, Linda Kwakkenbos. 4274

Stimulating Future-Oriented Thinking and Goal-Achievement Through the Future Self Using Virtual Reality and a Smartphone App: Randomized Controlled Trial (e84420)
Jean-Louis van Gelder, Tiffany Tettero, Aniek Siezenga, Esther Mertens. 4289

A Self-Guided App-Based Mindfulness Intervention for Racially and Ethnically Minoritized Individuals Who Experience Discrimination-Related Mental Health Symptoms: Randomized Controlled Trial (e84328)
Giovanni Ramos, Amanda Montoya, Adrian Aguilera, Anna Lau, Craig Enders, Yinyin Wen, Denise Chavira. 4307

Bidirectionality Between Perceived Immediate and Long-Term Benefits and Losses and Internet Gaming Disorder Among Chinese Adolescent Gamers: Prospective Longitudinal Study (e74030)
Siman Li, Jianxin Zhang, Ji-bin Li, Joseph Lau, Yanqiu Yu. 4325

Interactions of Technology and Obsessive-Compulsive Disorder Symptomatology in Adults: Qualitative Interview Study (e85033)
Lucas Occhino-Moede, Kaitlyn Sulivan-Pascual, Kendall Phelan, Harrison Wang, Daniel Mokhtar, Elisa Liu, Erica Schug, Megan Mirkis, Thomas Baek, Tamerlane Visher, Ujjwal Pasupulety, Adam Frank. 4339

Assessing Usage and Usability of a Narrative-Based Psychoeducational Digital Intervention to Improve Medication Adherence Among Individuals With Schizophrenia in a Stable Phase: Mixed Methods Study (e59175)
Dian Zhu, Fangyuan Chang, Hongyi Yang, Yiwen Wei, Zhao Liu. 4352

Effectiveness of and Mechanisms of Change in a Self-Help Web- and App-Based Resilience Intervention on Perceived Stress in the General Working Population: Randomized Controlled Trial (e78335)
Sandy Hannibal, Dörte Behrendt, Michèle Wessa, Sarah Schäfer, Nina Dalkner, Dirk Lehr. 4367

Predictors of Professional Responses in Nonprofit Mental Health Forums: Interpretable Machine Learning Analysis (e74359)
Shuang Geng, Yanghui Li, Jie Wang, Peixuan Chen, Xusheng Wu, Zhiqun Zhang. 4391

Effect of AI-Based Natural Language Feedback on Engagement and Clinical Outcomes in Fully Self-Guided Internet-Based Cognitive Behavioral Therapy for Depression: 3-Arm Randomized Controlled Trial (e76902)
Mirai So, Yoichi Sekizawa, Sora Hashimoto, Masami Kashimura, Hajime Yamakage, Norio Watanabe. 4411

Leisure Screen Time, Internet Gaming Disorder, and Mental Health Among Chinese Adolescents: Large-Scale Cross-Sectional Study (e80737)
Qin Deng, Linna Sha, Jiaojiao Hou, Xunying Zhao, Rong Xiang, Jiangbo Zhu, Yang Qu, Jinyu Zhou, Ting Yu, Xin Song, Sirui Zheng, Tao Han, Bin Yang, Mengyu Fan, Xia Jiang. 4431

XSL•FO
RenderX

Longitudinal Between- and Within-Person Associations Among Screen Time, Bedtime, and Daytime Sleepiness Among Adolescents: Three-Wave Prospective Panel Study (e78972)
Michał Tkaczyk, Albert Ksinan, David Smahel. 4445

Tailored Internet-Delivered Mindfulness-Based Interventions for Patients With Hepatocellular Carcinoma After Transarterial Chemoembolization: Qualitative Study (e78337)
Zengxia Liu, Min Li, Yong Jia, Li Chen. 4461

Associations Among Cyberbullying Victimization, Inhibitory Control, Neural Activation of Error Processing, and Mental Health Problems in Adolescents: Neuroimaging, Retrospective Longitudinal Cohort Study Using the Adolescent Brain Cognitive Development Data (e75126)
Xuanyu Zhang, Chengyan Xie, Yu Chen, Boyu Qiu. 4475

Adherence, Acceptability, and Sexual Health Outcomes of the Odeya App–Based Intervention for Sexual Distress in Women With Endometriosis: Randomized Controlled Mixed Methods Trial (e86042)
Selina Kronthaler, Eden Kosman, Tatjana Tissen-Diabaté, Elena Mühle, Luzia Weber- Schallauer, Therese Pross, Maria Karsten, Klaus Beier, Laura Hatzler. 4492

Mental Health Professionals' Perceptions of Benefits and Disadvantages of Telehealth: International Mixed Methods Study (e75905)
Madeline Montoya, Cary Kogan, Brigitte Khoury, José García-Pacheco, Rebeca Robles, Tahilia Rebello, Maya Kulygina, Geoffrey Reed. . . .
4 5 1 7

Artificial Intelligence Centrality in Psychotic Delusions and Violence Risk in Forensic Psychiatry: Retrospective Observational Study of Judicial Decisions (e93349)
Alexandre Hudon, Chanel Pagé. 4537

Patterns of Digital Media Engagement During Wartime and Their Associations With Stress, Resilience, and Well-Being: Cross-Sectional Survey Study (e91484)
Moti Zwilling, Galit Klein, Sharon Barak, Noa Shtainmetz, Riki Tesler, Avi Zigdon. 4551

Dietary Inflammatory Index and Depressive Symptoms in Chinese University Students Leveraging an Intelligent Ordering System: 3-Year Longitudinal Prospective Cohort Study (e83784)
Peng Hong, Chen Hao, Zhou Weiqiang, Qian Jie, Zhang Yimeng, Ding Jingyun, Qian Haihong, Jia Yingnan. 4571

Development of a Validated Lay Checklist (Info Without Side Effects) for Assessing Health Information on Websites: Mixed Methods Study (e69529)
Ursula Griebler, Irma Klerings, Christina Koscher-Kien, Benedikt Lutz, Eva Krczal, Dominic Ledinger, Iris Mair, Robert Emprechtinger, Filiz Keser Aschenberger, Bernd Kerschner. 4586

The Effect of Informational Podcasts on Shared Decision-Making, Anxiety, and Patient Satisfaction in Hospital Visits: Intervention Study (e81485)
Jannie Frølund, Anders Løkke, Hanne Jensen, Flemming Skjøth, Ingeborg Farver-Vestergaard. 4609

The Effect of Text Messaging on the Postoperative Pain Experience in Pediatric Patients Undergoing Thoracic Surgery: Randomized Controlled Trial (e81806)
Yun Shi, Qianqiu Wang, Aihua Liu, Li Jiang. 4626

Designing Electronic Problem-Solving Training for Individuals With Traumatic Brain Injury: Mixed Methods, Community-Based, Participatory Research Case Study (e83995)
Matthew Schmidt, Yueqi Weng, Shannon Juengst, Alexandra Holland. 4638

XSL•FO
RenderX

Toward Patient-Centric Digital Monitoring of Obstructive Sleep Apnea: Mixed Methods Study (e82460)
James Timmis, Kerstin Schorr, Rana Yüksel, Tim van den Broek, Sebastiaan Overeem, Dagmar Smid, Willem van den Brink, Nina Haring. . . 4679

Effectiveness of Educational Videos in Encouraging Preferences for Guideline-Based Cancer Screening in Japan: Three-Arm Pseudorandomized Controlled Trial (e82322)
Kosuke Sakai, Yuko Furuya, Shoko Nakazawa, Kota Fukai, Kei Sano, Masayuki Tatemichi. 4679

Seeking But Not Discussing Online Health Information With Physicians: Cross-Sectional Survey Study of eHealth Literacy–Empowerment Profiles and Patient-Centered Communication (e78836)
Qianfeng Lu, Wen Jiao, Angela Chang, Peter Schulz. 4693

Impact of GPT-4–Generated Discharge Letters on Patients' Medical Comprehension: Prospective Crossover Study (e81243)
Friederike Holderried, Alessandra Sonanini, Christian Stegemann–Philipps, Anne Herrmann–Werner, Philipp Spitzer, Martina Guthoff, Nils Heyne, Konstantin Sering, Martin Holderried, Felix Eisinger. 4711

The Real-Time Support Role of Augmented Reality Technology in Shared Decision-Making in Neurosurgery Under the SEGUE Framework: Randomized Controlled Trial (e87198)
Zhengbo Yuan, Zhongjie Shi, Zhanxiang Wang. 4728

Digital Patient Decision Aid for Antiobesity Medications: Mixed Methods Study of Human-Centered Design and Usability Evaluation (e89428)
Li-Jen Wang, Yi-Jen Wang, Yu-Lun Cheng, Wen-Liang Fang, Weu Wang, Meng-Cong Zheng. 4750

Engagement Trends in Online Vaccine Content: Longitudinal YouTube Study (e88519)
Michele Tizzani, Yelena Mejova. 4770

Understanding Public Perceptions of Lung Cancer in China: Infodemiology Study of Baidu Index and Weibo Posts (e85058)
Zongyuan Li, Jian Zhang, Yiming Fan, Nan Chen, Lunxu Liu. 4791

Why People Conceal Mental Health Problems: Qualitative Analysis of TikTok Posts (e88244)
Chloe Roske, Kael Ragnini, Qinchun Zhu, Ashari Palmer, Meredith Kells, Heather Davis, Matthew Nock. 4808

Misinformation in Social Media Narratives on Highly Pathogenic Avian Influenza: Systematic Content Analysis of Facebook and Instagram Posts (e73275)
Ahmed Al-Rawi, Betty Ackah, Abdelrahman Fakida, Kelley Lee. 4822

Patient Perceptions of Ozempic (Semaglutide) for Weight Loss: Mixed Methods Analysis of Online Medication Reviews (e78391)
Abanoub Armanious, Rachel-Mae Hunter, Kristi Griffiths, Hannah Bowrey, Robyn Brown, Morgan James. 4837

Public Emotional and Thematic Responses to Major Emergencies on Social Media, 2024-2025: Cross-Sectional Convergent Mixed Methods Study (e84648)
Xingrong Guo, Yiqian Fan, Yiming Guo. 4852

Global Sentiment Toward Health AI at the Dawn of the ChatGPT Era: Empirical Analysis of Twitter (X) Discourse (e80346)
Lily Wass, Zhengdong Wu, José Vizoso, Joseph Wu, Leesa Lin. 4880

Self-Reported Health Outcomes in Metabolic Health YouTube Comments: Cross-Sectional Study and Rule-Based Natural Language Processing Framework Development and Validation (e94855)
Ricardo Ribeiro, Aneesh Zutshi. 4897

XSL•FO
RenderX

User-Centered Diabetes Self-Management App (DiabAid Nexus) in Sub-Saharan Africa: Development and Usability Study (e87110)
Genet Aboye, Ermias Gebremichael, Gizeaddis Simegn, Jean-Marie Aerts. 4922

Development of an mHealth Intervention for Reducing Sedentary Behavior in Older Adults: Delphi Study (e83302)
Siqing Chen, Xiaofei Nie, Jiekai Yang, Yaqin Li, Lili Yang. 4937

Exploring Vaping Cessation App Use Among Youth: Qualitative Descriptive Study (e85778)
Karlee Fonteyne, Mischa Taylor, Roula Nawara, Danielle Rodberg, Tina Afshar, Laura Struik. 4953

Chinese Mobile Health Apps for Preventing and Managing Pelvic Floor Dysfunction: Quality Assessment and Content Analysis (e80126)
Yuqing Song, Xue Deng, Ting Hu, Sijie Feng, Lu Xing. 4966

Long-Term Mobile-Based Glycemic Intervention for Secondary Prevention in Patients With Diabetes Undergoing Surgical Revascularization: Multicenter Randomized Controlled Trial (e72226)
Yangwu Song, Yifeng Nan, Xieraili Tiemuerniyazi, Ziang Yang, Shicheng Zhang, Xi Li, Wei Feng. 4978

Acceptability of Technologies to Support Early Dementia Detection: Qualitative Study With the Boston University Alzheimer's Disease Center Cohort (e84004)
Sarah Wilson, Emily Beswick, Zachary Popp, Salman Rahman, Sharandeep Bhogal, Tim Whitfield, Spencer Low, Raiyan Khan, Clare Tolley, Zuzana Walker, Rhoda Au, Sarah Slight. 4995

Association Between Wearable Device Adoption and Health-Related Lifestyle Behaviors: Retrospective Cohort Study (e88276)
Eunjeong Choi, Seoyeong Choi, Suk-Yong Jang. 5008

Experiences and Perceived Influence of the Artificial Intelligence–Based Health Education Accurately Linking System (AI-HEALS) on Health Behaviors Among Patients With Type 2 Diabetes: Qualitative Study (e84605)
Jing Wang, Yibo Wu, Yang Jiang, Rantong Bao, Xinbao Gu, Bingyang Kong, Xinying Sun. 5017

Preferences and Willingness to Pay for Smart Bracelets Among Chinese Pregnant and Postpartum Women: Discrete Choice Experiment (e55941)
Jing Wang, Mingyue Cheng, Jialiang Feng, Wenzhuo Li, Xiaotao Zhou, Yufang Cao, Haiyan Feng, Fan Yang, Xiaoyun Wang, Yibo Wu. 5035

Assessing the Use of Wearable Mobile-Monitoring Devices Among Individuals With Serious Mental Illness: Qualitative Acceptability and Feasibility Study (e85087)
Aubrey Freitas, Jesus Chavez, Melissa Chinchilla, Ronald Calderon, Stephanie Chassman, Lauren Hoffmann, Alexander Young. 5052

Efficacy of Digital Speech Therapy for Poststroke Dysarthria: Randomized Noninferiority Trial (e81938)
Yuyoung Kim, Minjung Kim, Saebyeol Kim, Jinwoo Kim, Joon-Ho Shin, Yoonkyung Chang, Ji Na, JungWan Kim, Tae-Jin Song. 5064

Messages and Notifications for the "OA Coach" Knee Osteoarthritis Self-Management Mobile App: Codevelopment and Evaluation Using a Participatory Research Design With Focus Groups and Surveys (e83507)
Kate Bryce, Jocelyn Bowden, David Hunter, Robin Huang, Naomi Bloul, Vicky Duong, Jillian Eyles. 5081

Characterizing Physical Activity Trajectories Preceding Incident Major Depressive Disorder Diagnosis With Consumer Wearable Devices in the All of Us Research Program: Retrospective Nested Case-Control Study (e93164)
Yuezhou Zhang, Amos Folarin, Hugh Ellis, Rongrong Zhong, Callum Stewart, Heet Sankesara, Hyunju Kim, Shaoxiong Sun, Abhishek Pratap, Richard Dobson. 5104

XSL•FO
RenderX

A Proposed Participatory Framework for Explainable AI in mHealth: Mixed Methods Study Integrating User and Stakeholder Requirements (e87158)
Farzana Islam, Ashraful Islam, M Amin, Moinul Zaber. 5118

Smartphone App–Based Music-Facilitated Pulmonary Rehabilitation Program Integrating Rhythm-Guided Walking and Singing for Patients With Chronic Obstructive Pulmonary Disease: Multicenter Randomized Controlled Trial (e81707)
Minghui Shi, Shiwei Qumu, Xiaoning Bu, Jie Zhang, Kai Liu, Jinxiang Wang, Yinan Zhang, Jieping Lei, Siyuan Wang, Wei Li, Xingyao Tang, Jisong Yan, Yaodie Peng, Chen Wang, Ting Yang, Ke Huang. 5151

Perspectives Regarding the Privacy, Security, and Confidentiality of Data Collected via mHealth Apps in Saudi Arabia: Qualitative Analysis (e83381)
Nasser Alhammad, Mohannad Alajlani, Alaa Abd-alrazaq, Theodoros Arvanitis, Gregory Epiphaniou. 5171

A Gamified Mobile Health Intervention to Promote Physical Activity, Executive Function, and Mental Health in College Students: Randomized Controlled Trial (e82769)
Ming Yang, Yingjun Guo, Zhen Li, Lei Jiang, Yanan Gao, Sen Yang, Zhixiong Zhou. 5185

Preferences for Personalized Text Message Appointment Reminders Among Outpatients in a Universal Health System: Cross-Sectional Study (e81324)
Shih-Huai Hsiao, Yu-Chen Chuang, Shu-Ling Ma, Hui-Tzu Lin, I-Chen Lee, Pei-Shih Chen. 5206

Personalized Smartphone Messaging for Secondary Prevention After Percutaneous Coronary Intervention: Randomized Controlled Trial (e81524)
Sunki Lee, Ji Kim, Seung Shin, Jeonghoon Ahn, Hyun-Ghang Jeong, Kyung Jung, Younghoon Song, Eung Kim. 5220

Digital Peer Support to Increase Walking Among Older Adults: Cluster Randomized Trial (e75708)
Atsushi Nakagomi, Noriyuki Abe, Takayuki Ueno, Gemmei Izuka, Yohei Kawasaki, Katsunori Kondo. 5238

Enhancing Adherence to Home-Based Expiratory Muscle Strength Training in Parkinson Disease: Randomized Controlled Trial of an mHealth Intervention (e78022)
Martin Srp, Martina Hoskovcova, Rebeka Lagnerova, Katerina Dvorakova, Radim Kliment, Jan Muzik, Radim Krupicka, Ota Gal, Evzen Ruzicka.
5 2 5 1

Comparing the Associations of Internet Addiction and Internet Gaming Disorder With Psychopathological Symptoms: Cross-Sectional Study of Three Independent Adolescent Samples (e82414)
Ying-ying Li, A-qian Hu, Ling-li Yi, Zi-xin Mao, Qiu-yue Lü, Juan Wang, Wei Wei, Yue-qi Huang, Shu Huang, Wen-jing Dai, Meng-xuan Qiao, Jia-jun Xu, Qiang Wang, Xiao-jing Li, Fu-gang Luo, Wei Deng, Yu-zheng Hu, Tao Li, Wan-jun Guo. 5267

Digital Phenotyping for Adolescent Mental Health: Feasibility Study Using Machine Learning to Predict Mental Health Risk From Active and Passive Smartphone Data (e72501)
Balasundaram Kadirvelu, Teresa Bellido Bel, Aglaia Freccero, Martina Di Simplico, Dasha Nicholls, A Faisal. 5282

The Effectiveness of the Headspace App for Improving Sleep: Randomized Controlled Trial (e56287)
Zoltan Torok, Larisa Gavrilova, Amish Patel, Matthew Zawadzki. 5309

Associations Between Smartphone-Based Finger Tapping and Cognitive Performance in Older Adults: Observational Study (e82463)
Huitong Ding, Taylor Orwig, Jian Rong, Heaven Tatere, Chunyu Liu, Eric Schramm, Chathurangi Pathiravasan, Joanne Murabito, Honghuang Lin. 5323

Impact of Push Notifications on Physical Activity and Sodium Intake Among Patients with Hypertension: Microrandomized Trial of a Just-in-Time Adaptive Intervention (e78218)
Jessica Golbus, Michael Dorsch, Yuxuan Chen, Tanima Basu, Evan Luff, Predrag Klasnja, Mark Newman, Lesli Skolarus, Walter Dempsey, Brahmajee Nallamothu. 5335

XSL•FO
RenderX

Effects of an eHealth Cardiac Exercise Rehabilitation Platform for Patients After Percutaneous Coronary Intervention Based on the Persuasive Systems Design Model: Randomized Controlled Trial (e71450)
Yang Liu, Xiting Huang, Ziying Dai, Zhili Jiang, Wenxiao Wu, Jing Wang, Zhiqian Wang, Luyao Yu, Hanyu Li, Lihua Huang. 5346

Feasibility, Acceptability, and Perspectives Regarding the Use of Activity Tracking Wearable Devices Among Home Health Aides: Mixed Methods Study (e77510)
Ian Solano-Kamaiko, Michael Dicpinigaitis, Melissa Tan, Irene Yang, Kexin Cheng, Ronica Peramsetty, Michelle Shum, Yanira Escamilla, Jennifer Bayly, Meghan Turchioe, Ariel Avgar, Aditya Vashistha, Nicola Dell, Madeline Sterling. 5368

Characteristics Influencing Support for the National Health Service COVID-19 App in England and Wales: Findings From a Longitudinal Survey (e76863)
Josephine Exley, Paul Boadu, Kasim Allel, Bob Erens, Nicholas Mays, Mustafa Al-Haboubi. 5379

Patterns and Characteristics of Mobile App Use to Promote Wellness and Manage Illness: Cross-Sectional Study (e71363)
Hayriye Gulec, David Smahel, Yi Huang. 5395

Effects of an 8-Week App-Based Mindfulness Intervention on Mental Health in Working Women: Randomized Controlled Trial (e62814)
Riko Uwagawa, Koichiro Adachi, Mariko Shimoda, Ryu Takizawa. 5417

Changing Habits With the Happy Hands App: Qualitative Focus Group Study of a Hand Osteoarthritis Self-Management Intervention (e82773)
Kristine Fjeldstad, Anne Tveter, Eivor Rasmussen, Lena Olden, Sissel Nyheim, Thalita Blanck, Rikke Killingmo, Ingvild Kjeken. 5432

Facilitators and Barriers to Implementing Mobile Mental Health Interventions: Qualitative Study of the Consolidated Framework for Implementation Research in Pediatric Oncology Providers (e87533)
Shannon Hong, Michaela Patton, Krysta Barton, Tonya Palermo, Karen Mulholland, Eric Chow, Nancy Lau. 5444

Mobile Health App Attitudes and Adoption Among Oncology Providers: Cross-Sectional National Survey (e85583)
Nancy Lau, Kavin Srinakarin, Shannon Hong, Homer Aalfs, Michael Roth, Karly Ingram, Amy Berkman, Sherilynn Chan, Joanna Patten, Wade Iwata, Mallory Taylor, Jesse Fann, Eric Chow, Tonya Palermo. 5458

Patient Sharing of Digital Health Data in the Veterans Health Administration: Cross-Sectional Analysis (e80517)
Mark Zocchi, Jessica Lipschitz, Felicia Bixler, Stephanie Robinson, Bella Etingen, Timothy Hogan, Ndindam Ndiwane, Saige Calkins, Terry Newton, Nilesh Shah, Ben Kragen, Alexander Young, Stephanie Shimada. 5476

Use and Influencing Factors of mHealth Services Among Adult Survivors of Cancer: Cross-Sectional Survey Study (e82902)
Taiguo Liu, Lina Duo, Musi Zhang, Jianjun Zhang, Jingna Zhang, Xiao Sun, Haoyue Deng, Ji Cheng, Ya Sun. 5488

Effectiveness and Cost-Effectiveness of an App-Based Mindfulness Breast Care Program for Reducing Body Image Distress and Stigma Among Survivors of Breast Cancer: Randomized Controlled Trial (e85913)
Huicong Lin, Yujing Zhong, Shuang Zheng, Xiaoying Jiang, Yun Xu, Sally Chan, Shengjie Liu, Ka Chow, Qingmo Yang, Wenhe Huang, Jiemin Zhu. 5503

Agile Development and Testing of a Gamified Human Milk Feeding Education Mobile App for Participants of the Special Supplemental Nutrition Program for Women, Infants, and Children: Co-Design Approach (e80330)
Yarisbel Melo Herrera, Shannon Mitchell, Amy Malinowski, Catherine Wright, Marguerite Dibble, Patricia Cassi, Naina Qayyum, Erin Hennessy.
5 5 2 7

XSL•FO
RenderX

Posttraumatic Symptoms as Predictors of Engagement With a Mobile App for Coping After Military Sexual Trauma: Public Usage Data Analysis Study (e85098)
Shilpa Hampole, Shannon McCaslin-Rodrigo, Valerie Carr, David Schuster, Sofia Reyes, Sarah Senti, Kelly Ramsey, Jason Owen, W Skidmore, Amy Street, Katherine Taylor, Margaret-Anne Mackintosh. 5544

Use of Health and Well-Being Technology, Basic Psychological Needs, and the Mediating Role of Technological Identity in 6 European Countries: Prospective Longitudinal Survey Study (e83054)
Moona Heiskari, Aki Koivula, Magdalena Celuch, Iina Savolainen, Teijo Osma, Atte Oksanen. 5562

Daily Stress and Heart Rate Variability Among Mindfulness Meditation Practitioners: mHealth Observational Study (e78244)
Jo Takezawa, Shixian Geng, Masahiro Fujino, Mika Miyake, Kazutoshi Sasahara, Koji Yatani, Atsushi Niida. 5578

Understanding mHealth Engagement Among Patients With 30-Day Hospital Revisits: Secondary Analysis of a Randomized Clinical Trial (e89067)
Susan Landon, Angira Mondal, Aiden Ahn, Klea Profka, Anna Morgan, Eric Bressman. 5598

Smartphone Keystroke Biomarkers as Predictors of Adverse Neuropsychiatric Sequelae After Trauma in Trauma Survivors: Prospective Observational Cohort Study (e73771)
Nicole Short, Xinming An, Yinyao Ji, Qinghua Li, Thomas Neylan, Gari Clifford, Stacey House, Francesca Beaudoin, Jennifer Stevens, Sarah Linnstaedt, Laura Germine, John Haran, Alan Storrow, Christopher Lewandowski, Paul Musey Jr, Phyllis Hendry, Sophia Sheikh, Christopher Jones, Brittany Punches, Jose Pascual, Mark Seamon, Erica Harris, Claire Pearson, Roland Merchant, Robert Domeier, Niels Rathlev, Brian O'Neil, Paulina Sergot, Leon Sanchez, Steven Bruce, Ronald Kessler, Karestan Koenen, Kerry Ressler, Samuel McLean. 5606

Associations Between Mukbang Watching and Appetite, Nutrition, and Quality of Life in Pediatric Patients with Cancer: Intensive Longitudinal Study (e80932)
Yuxue Xiao, Jiani Wang, Ting Zhong, Xiao Wang, Dongyan Tang, Wei Xia. 5629

Large Language Model–Based Analysis of Statin Therapy Discussions and Sentiment on Social Media: Cross-Sectional Observational Study (e85057)
Siru Liu, Jialin Liu. 5674

How Online Scheduling Platforms Affect Insurance-Based Disparities in Access to Specialist Outpatient Care in Berlin, Germany: Cross-Sectional Audit Study (e82452)
Cordula Kreuzenbeck, Silvia Angerer. 5686

Social Support Mechanisms in an Online Type 1 Diabetes Community: Social Network Analysis of Stakeholder Diversity and Disease Duration (e82996)
Yujia Zhu, Yanxiang Zhang, Chenxiao Zhao, Xifeng Zeng, Xueying Zheng. 5699

Reddit Discussions During the 2022 Mpox Outbreak: Observational Analysis of Sentiment, Topics, and Audience Engagement (e90152)
Xi Luo, Zahra Movahedi Nia, Jude Kong. 5719

Characteristics of Human-Like Virtual Profiles in Relation to Audience Reach and Engagement on Instagram: Secondary Data Analysis (e86233)
Alexandra Lebrecht, Winze Tam, Omar Merlo, Andreas Eisingerich. 5740

Human Papillomavirus Vaccine Discourse and Sentiment on Reddit Before and After COVID-19: Mixed Methods Retrospective Cross-Sectional Study (e83558)
Marcin Marciniak, Sean Setzen, Deeya Bhattacharya, Jamie Masliah, Robin Powszok, Erich Sturgis, Vanessa Stubbs, Mihir Bhayani. 5758

Interpretations of Menstrual Blood Appearance and Diagnostic Potential Among Social Media Users: Cross-Sectional Mixed Methods Social Media Listening Study (e85550)
Ruth Niedermeier, Noemi Castelletti, Andreas Bender, Michael Hoelscher, Olena Ivanova. 5773

XSL•FO
RenderX

Leveraging Social Media to Achieve Population-Level Reach of Lung Cancer Screening-Eligible Individuals: A RE-AIM Framework Perspective (e80281)
Lisa Carter-Bawa, Jamie Ostroff, Susan Rawl, Erin Hirsch, Smita Banerjee, Andrew Ciupek, Robert Comer, Minal Kale, Katherine Leopold, Patrick Monahan, James Slaven Jr, Francis Valenzona, Renda Wiener, Ana Vielma. 5790

Diabetes Technologies in Ultra-Endurance Type 1 Diabetes: Qualitative Study (e86815)
Jean-Charles Vauthier, Lucie Choley, Delphine Arduini, Patrick Mas, Bernard Kabuth. 5803

Cocreating Principles for Digital Health Equity: Cross-Sectional, Qualitative Study for Participatory Human-Centered Design in Catalonia (e84129)
Jordi Piera-Jiménez, Núria Vilarasau Creus, Ada Maymó Costa, Xabier Michelena, Andrea Climent Fageda, Alèxia Farré, László Herczeg, Lekshmy Parameswaran, Gerard Carot-Sans, Luis Valle. 5815

Moving Research to Practice: A Qualitative Study Exploring Patient Perspectives of the Implementation and Sustainability of a Digital Health Intervention for Chronic Kidney Disease (e91966)
Courtney Lightfoot, Roseanne Billany, Gurneet Sohansoha, Thomas Wilkinson, Melanie Davies, Thomas Yates, Alice Smith, Matthew Graham-Brown. 5838

Digital Health Intervention to Promote Lifelong Specialized Care in Adults With Congenital Heart Disease: Theory-Driven Community Co-Designed Study (e75867)
Anushree Agarwal, Joseph Valente, Karina Buenrostro, Katelyn Macholl, Juhi Mehta, Keerthana Reddy, Karina Manayan, Parang Kim, Aleah Sparks, Kunyi Li, Pranav Ahuja, Kevin Sun, Kimberly Payton, Mark Norris, Katia Bravo-Jaimes, Leigh Reardon, Philip Moons, Megumi Okumura, Gregory Marcus, Michelle Gurvitz. 5856

A Multilingual Digital Microlearning Intervention for Oral Health in Refugee Shelters: Randomized Controlled Trial (e95562)
Maxi Müller, Fabian Schulz, George Jogho, Simone Steffens, Kirstin Vach, Ahmad Baradee, Daniel Reissmann, Benedikt Spies, Anna-Lena Hillebrecht. 5873

Text Messaging Support for Patients Diagnosed With Impaired Glucose Tolerance During Pregnancy: Nonrandomized Pre-Post Implementation Study Assessing Impact on Postpartum Transitions of Care (e76493)
Natasha Kumar, Whitney Bender, Celeste Durnwald. 5887

Predictors of Outcomes of Parent Training Targeting Disruptive Behavior in Children Aged 4 Years at 6-Month Follow-Up: Results From a Large Prospective Cohort Implementation Study (e79592)
Minja Westerlund, Susanna Hinkka-Yli-Salomäki, Terja Ristkari, Malin Kinnunen, Amit Baumel, Andre Sourander. 5902

Leveraging Text Messaging and Behavior Theory to Improve Colorectal Cancer Screening in Federally Qualified Health Centers: Cohort Study (e86408)
Tracy Angelocci, Tonghui Xu, Tushar Talaparthy, Cecilia Corral, Omolola Adepoju. 5919

A Digital Diabetes Self-Management Education and Support Program Integrated With Continuous Glucose Monitoring for Type 2 Diabetes: Randomized Controlled Trial (e78321)
Ashley Berthoumieux, Jeanean Naqvi, Sean Zion, Jenna Napoleone, Amanda McGuill, Christian Cerrada, Hyun Lee, Timothy Dunn, David Kerr, Carolyn Jasik, Sarah Linke. 5934

Nicotine Dependence and Quit Self-Confidence in a Smoking Cessation Program Using a Group-Based Digital Peer-Supported App and Cigarette Consumption–Adjusted Nicotine Aids Among Japanese Workers: Retrospective Cohort Study (e84792)
Shota Yoshihara, Kayoko Takahashi, Chiaki Uemura, Shin Murakami, Satoru Amano, Daichi Harada, Ying Jiang, Hiroshi Yamato. 5949

XSL•FO
RenderX

A Digital Toolkit for Weight Loss Maintenance in European Adults (NoHoW): 2×2 Factorial Randomized Controlled Trial (e69634)

R Stubbs, Cristiana Duarte, Clarissa Dakin, António Palmeira, Falko Sniehotta, Graham Horgan, Sofus Larsen, Marta Marques, Jorge Encantado, Elizabeth Evans, Jake Turicchi, Ruari O'Driscoll, Sarah Scott, Beth Pearson, Lauren Ramsey, Marie-Louise Mikkelsen, Inês Santos, Marcela Matos, Pedro Teixeira, Berit Heitmann. 5967

Effect of Integrated Internet-Based Acceptance and Commitment Therapy and Behavioral Activation Among Ethnic Minority Young Adults With Alcohol Use Disorder in Hong Kong: Pilot Randomized Controlled Trial (e83896)

Getaneh Belay, Katherine Lam, Qi Liu, Ting Mao, Yim Mak, Ka Ho. 5990

Cardiac Arrest Survivors' Perspectives to Inform the Co-Design of a Web-Based Support and Learning Platform: Qualitative Content Analysis (e84432)

Annette Waldemar, Anders Bremer, Johan Israelsson, Katarina Heimburg, Per Nordberg, Erik Blennow Nordström, Kristofer Årestedt, Ingela Thylén. 6014

Effectiveness of a Blended Intervention to Promote Physical Activity Among Office Employees: Randomized Controlled Trial (e80249)

Yan Sun, Yang Gao, Alison Ou, Xueying Wang, Yaojie Xie, Xiang-yan Chen, Shirley Fong, Wendy Huang, Chun-Qing Zhang, Julien Baker, Martin Hagger. 6031

Self-Guided Internet-Based Mindfulness-Informed Stress Management for Generalized Anxiety Disorder: Randomized Controlled Trial With Longitudinal Network Analysis (e91751)

Ziwei Wang, Hui Tong, Jianrong Yue, Siyan Wu, Ye Xia, Han Zhang, Yuan Yang. 6046

Social Media Intervention Based on the Information-Motivation-Behavioral Skills Model Promotes HIV Testing and Reduces High-Risk Behaviors Among Men Who Have Sex With Men in Resource-Limited Settings in China: Randomized Controlled Trial (e84279)

Huanzhuo Mai, Jianyuan Liu, Liping Cheng, Hao Liang, Ping Cui, Qin Yao, Qin Cao, Heng Zhang, Wenping Liao, Shiran He, Xing Yang, Ping Cen, Jiegang Huang. 6064

Usability of iSupport Swiss, a World Health Organization Digital Intervention for Caregivers of People With Dementia: Mixed Methods Study (e81247)

Anna Messina, Rebecca Amati, Anna Annoni, Emiliano Albanese, Maddalena Fiordelli. 6084

Determinants of the Uptake and Frequency of Use of a Web Portal Digital Health Intervention in Patients With Type 2 Diabetes and/or Coronary Heart Disease: Secondary Analysis of a Randomized Controlled Trial (e80895)

Maximilian Scholl, Claas Lendt, Sebastian Appelbaum, Bianca Biallas, Katja Brenk-Franz, Chloé Chermette, Friederike Frank, Angeli Gawlik, Lisa Giesen, Martina Heßbrügge, Lucas Küppers, Larisa Pilic, Marcus Redaèlli, Lara Schneider, Frank Vitinius, Stefan Wilm, Uwe Konerding. . . .
6 1 1 8

Internet-Based Problem Management Plus Intervention for Antenatal Depression: Randomized Controlled Trial (e81998)

Shuanghui Zheng, Shiyu Wang, Xiaoyu Shui, Jing Zhao, Ying Lin, Qian Huang, Siqi Li, Yang Chen, Zhijie Zou, Jin Zheng, Xiaoli Chen. 6138

Digital Intervention (MiVacunaLA 2.0) to Promote COVID-19 Vaccine Acceptance Among Hispanic Children: Community-Based Randomized Controlled Trial (e78103)

Luisa Blanco, Alexandra Klomhaus, Rebecca Dudovitz, Denise Marquez, Ray Lopez-Chang, Evelyn Aleman, Bas Weerman, Michael Moldoff, Keith Norris, Yelba Castellon-Lopez. 6156

Designing a Carbohydrate Counting App for Young Adults With Type 1 Diabetes: Usability Testing Interview Study (e86024)

Asmaa Housni, Aidan Shulkin, Alexandra Katz, Giuliana Giannini, Amélie Roy-Fleming, Meranda Nakhla, Courtney South, Anne-Sophie Brazeau. 6172

XSL•FO
RenderX

Effectiveness of Postdischarge Telephone Calls in Reducing Hospital Utilization: Quasi-Randomized Controlled Trial (e80529)
Sandie Du, Abigail Anada, Maria Montenegro, Maurie Maitland, Sean Chuan, Sonia Jose, Harveer Sihota, Delia Cooper, Joanna Rivera, Dianne Burditt, Minetaro Naruki-van Velzen, Megan MacPherson. 6186

Effects of Self-Compassion and Mindfulness Interventions on Mental Health and Work-Related Outcomes Among Japanese Workers: Randomized Controlled Trial (e79991)
Takumu Kurosawa, Koichiro Adachi, Ryu Takizawa. 6200

Effects of Internet-Based Dementia Risk Reduction Education on Risk and Protective Factor Knowledge, Intentions, and Health Behaviors: Randomized Controlled Trial (e79405)
Anthony Levinson, Stephanie Ayers, Sandra Clark, Rebekah Woodburn, Amy Schneeberg, Dima Hadid, Nick Kates, Gillian Strudwick, Roland Grad, Alexandra Papaioannou, Maureen Dobbins, Henry Siu, Dante Duarte, Karen Saperson, Sharon Marr, Doug Oliver, Sarah Neil-Sztramko.
6 2 2 0

Enhancing LGBTQ+ Inclusivity in an AI-Powered Sexual Health Chatbot: User-Centered Design Approach Through a Nonprofit and Academic Partnership (e78621)
William Liem, Elizabeth Casline, Julianna Lorenzo, Jacob Gordon, Andrés Avila, Attia Taylor, Nicole Levitz, Michael O'Keefe, Karen Shum, Kathryn Macapagal. 6240

The Impact of a Health Coaching App on the Subjective Well-Being of Individuals With Multimorbidity: Mixed Methods Study (e78738)
Isabelle Symes, Alexandra Burton, Daniela Mercado, Feifei Bu. 6261

Problems and Barriers Regarding the Admission, Financing, and Service Provision of Digital Health Apps: Qualitative Stakeholder Survey (e73332)
Felix Plescher, Godwin Giebel, Carina Abels, Silke Neusser, Bonnie Kampka, Kirstin Börchers, Jürgen Wasem, Nikola Blase. 6279

Effects of Using a Smartphone App Combined With Behavior Change Techniques on the Level of Physical Activity Among Adults and Older Adults: Sequential Multiple Assignment Randomized Trial (e73388)
Maria Simoes, Neli Proença, Vinícius Lauria, Matheus do Nascimento, Ricardo Padovani, Victor Dourado. 6293

Effect of Lung Cancer Screening, Smoking Cessation, and Cessation Smartphone App to Health-Related Quality of Life Among Heavy Smokers: Randomized Controlled Trial (e81687)
Antti Kurtti, Sanna Iivanainen, Riitta Kaarteenaho, Heidi Andersen, Antti Jekunen, Tuula Vasankari, Jussi Koivunen. 6313

Patient and Care Team Perspectives of Barriers to and Facilitators for the Implementation of a Digital Health Program for Depression in Primary Care: Qualitative Study (e72003)
Andrea Nederveld, Elise Robertson, Angela Lanigan, Elisabeth Callen, Tarin Clay, Ben Fehnert, Lambros Chrones, Michael Martin, Margaret McCue, Christina Hester, Melissa Filippi. 6331

Internet Health Care Service Use Behavioral Pattern Among Older Adults and the Role of the Technology Acceptance and Social Ecological Theory Model: Cross-Sectional Survey (e78037)
Rui Li, Xinyu Xu, Qingsong Li, Haobiao Liu, Ting Zhou, Abebe Amhare, Peiyu Liu, Jing Tang, Wei Wang, Fuju Zheng, Jing Han. 6343

Therapeutic Effects of a WeChat Mini-Program on Metabolic Dysfunction–Associated Fatty Liver Disease: Randomized Controlled Trial (e76204)
Chao Sun, Guangyu Chen, Cuicui Shi, Haixia Cao, Ruixu Yang, Jing Zeng, Xiaoyan Duan, Xin Sun, Jian-Gao Fan. 6359

Cognitive-Behavioral Therapy−Based Massed Brief Psychoeducational Group via Videoconference for Social Anxiety: Randomized Controlled Trial (e79825)
Lele Feng, Wei Liu, Liechuan Cui, Deborah Dobson, Xinfeng Tang. 6380

XSL•FO
RenderX

Smartphone-Based Digital Eczema Education Program for Atopic Dermatitis in Children Aged 0 to 6 Years: Multicenter, Randomized, Parallel Controlled Clinical Study (e79559)
Huan Yang, Hong Shu, Liu-hui Wang, Ping Li, Yun-ling Li, Qin-feng Li, Xiu-ping Han, Jing Tian, Jing Chang, Hua Qian, Jing-ping Chen, Xin-qiang Ding, Pan-qian Wu, Li-min Dou, Zhen Luo, Wei Li, Yang-yang Lin, Lin Li, Shu-zhen Yue, Yang Gu, Li Yang, Xiao-hong Sun, Xiao-yan Luo, Lin Ma, Hua Wang. 6396

Bayesian-Based Pharmacokinetic Framework Integrated with Therapeutic Drug Monitoring for Assessing Adherence to Antiseizure Medications: A Clinical Trial Simulation Study (e77917)
Xiao-Qin Liu, Zi-Ran Li, Wei-Wei Lin, Juan Wang, Fu-Qing Gu, Jun-Jie Ding, Zheng Jiao. 6409

A Real-Life Digital Intervention for Personalized Nutrition in Adults With Overweight or Obesity: Remote Randomized Controlled Trial (e73367)
Jelle de Jong, Femke Hoevenaars, Lotte Peters, Charlotte Berendsen, Wilrike Pasman, Martien Caspers, Remon Dulos, Suzan Wopereis. . .
6 4 2 3

A Web-Based Cancer Prevention Intervention for Rural Emerging Adults: Mixed Methods Development and Pilot-Testing Study (e80803)
Echo Warner, Alishia Kinsey, Barbara Walkosz, Julia Berteletti, Kayla Nuss, Annelise Small, W Woodall, Deanna Kepka, Douglas Taren, Meghan Skiba, Dolores Guest, Cindy Blair, Judith Gordon, David Wetter, Evelinn Borrayo, Kimberly Henry, Andrew Sussman, David Buller. 6440

Effects of Artificial Intelligence Recognition–Based Telerehabilitation on Exercise Capacity in Patients With Hypertension: Randomized Controlled Trial (e81400)
Qiuru Yao, Baizhi Qiu, Longlong He, Qin Wang, Jihua Zou, Donghui Liang, Shuyang Wen, Yingchao Liu, Gege Li, Jinjing Hu, Huan Ma, Guozhi Huang, Qing Zeng. 6459

Digital Engagement Significantly Enhances Weight Loss Outcomes in Adults With Obesity Treated With Tirzepatide: Retrospective Cohort Study of a Digital Weight Loss Service (e83718)
Hans Johnson, Ashley Clift, Daniel Reisel, David Huang. 6477

The Relationship Between Physician Self-Disclosure and Patient Acquisition in Digital Health Markets: Cross-Sectional Study (e84963)
Quanchen Liu, Pengqing Yin, Jing Fan. 6493

Website Use and Associations With Behavior Change and Weight Loss in Cancer Survivors and Their Partners: Secondary Analysis of a Randomized Controlled Trial (e86908)
Harleen Kaur, Dori Pekmezi, Tracy E Crane, David Farrell, Laura Q Rogers, Wendy Demark-Wahnefried. 6509

Traditional Rehabilitation Experiences, Unmet Needs, and Perspectives on Virtual Reality–Based Rehabilitation Among Patients With Stroke in China: Qualitative Thematic Analysis and Semistructured Interview Study (e84532)
Xite Zheng, Lu Xing, Haitao Lu, Shimeng Hao, Fen Liu. 6526

Effectiveness of the Mobile-Based Diabetes Little Helper Video Intervention on Medication Adherence Among Older Adults Living With Type 2 Diabetes Mellitus, Henan, China: Randomized Controlled Trial (e78731)
Meng Wang, Khuan Lee, Hui Thew, Siti Sowtali, Qiuhuan Jiang, Zhengyan Zhao, Yu Liu, Poh Lim. 6542

Added Technology for Weight Management in Cardiac Rehabilitation: Biopsychosocial Factors and Rehabilitees' Talk in a Mixed Method Study (e78347)
Heli Lahtio, Teemu Paajanen, Hilkka Korpi, Ari Heinonen, Tuulikki Sjögren. 6560

Effectiveness of the Components of a Digital Multiple Health Behavior Intervention Among University Students (Buddy): Factorial Randomized Trial (e88884)
Katarina Åsberg, Oskar Lundgren, Hanna Henriksson, Pontus Henriksson, Ann Eldh, Preben Bendtsen, Marie Löf, Marcus Bendtsen. 6577

XSL•FO
RenderX

Telepathology and Mobile Health System for Province-Wide Pathology Consultation in Henan, China: Retrospective Evaluation Study (e75172)
Jinming Shi, Ming Ye, Dongxu Sun, Xianying He, Yaoen Lu, Linlin Wang, Haotian Chen, Wenchao Wang, Jie Zhao, Fangfang Cui. 6600

Creation of EmpowerMe Website to Promote Self-Efficacy in Survivors of Stroke: Co-Design Study (e76756)
Elizabeth Lynch, Adrian O'Malley, Zoe Adey-Wakeling, Billie Bonevski, Dominique Cadilhac, Saran Chamberlain, Leonid Churilov, Robyn Clark, Richard Cullen, Erin Godecke, Gillian Harvey, Fiona Jones, Natasha Lannin, Stacy Larcombe, Jarrad Law, Annette McGrath, Lisa Murphy, Katie Nesbitt, Karly Zacharia, Niranjan Bidargaddi, Coralie English. 6617

Insights and Recommendations From Moderators and Community Members for Keeping Online Peer Support Safe: Thematic Analysis (e81943)
Hannah Jones, Grace Lavelle, Elly Aylwin-Foster, Ciara Regan, Alan Simpson, Ewan Carr, Matthew Hotopf, Vanessa Lawrence. 6632

Effect of a Digital-Driven Physician-Pharmacist Collaborative Model for Diabetes in Primary Health Care: Cluster Randomized Trial (e77470)
Jie Xiao, Qing Wang, Shenglan Tan, Lei Chen, Daxiong Xiang, Haiyan Yuan, Xia Li, Shuting Huang, Bingjie Tang, Yan Guo, Haiying Huang, Danhui Zhao, Yue Li, Li Wang, Qun Li, Juan Liu, Ping Xu. 6652

Coproducing an Online Platform for People With Long-Term Physical Health Conditions: Development and Usability Study (e79666)
Hannah Jones, Grace Lavelle, Elly Aylwin-Foster, Ewan Carr, Alan Simpson, Matthew Hotopf, Vanessa Lawrence. 6668

Effectiveness of the Components of a Digital Multiple Health Behavior Change Intervention Among Individuals Seeking Help Online (Coach): Factorial Randomized Trial (e88881)
Joel Crawford, Katarina Åsberg, Jenny Blomqvist, Oskar Lundgren, Hanna Henriksson, Pontus Henriksson, Preben Bendtsen, Marie Löf, Marcus Bendtsen. 6686

Exploring Patient Empowerment and Health System Outcomes Associated With MyHealthNB, a Provincial Personal Health Record System: Exploratory Mixed Methods Study (e87521)
Paula Voorheis, Stephan Dombrowski, Frances Bruno, Carolyn Steele Gray. 6705

Clinical Effects of a Digital Health Intervention for Adults With Type 2 Diabetes in the United States: Retrospective Cohort Study (e66911)
David Kerr, Nita Thingalaya, Praveen Potukuchi, Keni Lee, Edward Han-Burgess, Darren Frey, Alison Edwards, Xinyan Yu, Laura Wilson, Felix Lee. 6726

Evaluating Co-Designed vs Researcher-Driven Personalized Feedback Formats in a Brief Digital Alcohol Use Intervention: Mixed Methods Study (e87393)
Antoinette Poulton, Joanna Bowen, Gezelle Dali, Cindy Chew, Katrina Prior, Emma Devine, Louise Birrell, Lexine Stapinski, Robert Hester. . 6 7 4 3

A Web-Based Virtual Environment Behavioral Intervention as Cardiovascular Disease and Metabolic Disease Prevention Education in Persons With HIV: Evaluation of the LEARN Randomized Controlled Trial (e91145)
S Ramos, Lindsie Boerger, Arjee Restar, Manali Phadke, Sangchoon Jeon, Constance Johnson, Gail Melkus, Trace Kershaw, Harmony Reynolds, Allison Vorderstrasse. 6765

Acceptability and Preliminary Evaluation of a Campus-Integrated Digital Platform (Fruto) for University Students' Mental Health Help-Seeking: Sequential Mixed Methods Study (e78930)
Myungsung Kim, Hyorim Kim, Jeong-in Heo, Seonmi Lee, Orane Lahcine, Sangil Lee, Hwang Kim, Dooyoung Jung. 6787

Digital Media for Health Outcomes: Evaluation Study of a Massive Online Open Course (e85016)
Hannah Melchinger, Sarah Christie, Abigail Titus, Alyssa Cruz, Chelsey Lepage, Tony Foleno, Colleen Thompson-Kuhn, Adam Abebe, G. Gweh, Donewell Bangure, Stephen Maina, Judy Mwangi, Peninah Thande, Maike Winters, Amyn Malik, Aleksandra Kuzmanovic, Surangani Abeyesekera, Sara Epperson, Whittney Tom, Kylie Holmes, Angus Thomson, Saad Omer. 6807

XSL•FO
RenderX

Combined Internet-Based Cognitive Behavioral Therapy and Face-to-Face Physiotherapy in Primary Health Care for Chronic Widespread Pain: Randomized Controlled Trial (e86792)
Anna Bergenheim, Maria Larsson, Chan-Mei Ho-Henriksson, Anna Grimby-Ekman, Anna Larsson, Marie Persson, Sandra Weineland, Lena Nordeman. 6827

Effectiveness of a Home-Based and Group-Based Tele-Exercise Program for Breast Cancer Survivors: Pilot Randomized Controlled Trial (e79564)
Yang Gao, Yan Sun, Xueying Wang, Oi Chun, Brigitte Fung, Wai Fung. 6844

Digital Literacy and Patient Satisfaction in Telemedicine Follow-Up With In-Person App Instruction Versus Outpatient Department Follow-Up After Upper-Extremity Surgery: A Randomized Controlled Trial (e86918)
Werapat Ngowroongrueng, Sitthiphong Suwannaphisit, Chanon Malaikritsanachalee, Sutee Thaveepunsan, Sirisak Chaitantipongse, Theephop Teeragananan, Kanokporn Imsakul, Supaira Puntumabumrung, Krit Prasittichok. 6860

Telehealth Scale and Artificial Intelligence Adoption Tiers Across Clinical and Operational Domains in US Hospitals: Cross-Sectional Study (e96762)
Lingbo Liu. 6872

Clinician-Patient Asynchronous Text Messaging Communication in Hospital-at-Home Care: Qualitative Study (e83530)
Jeremy Seow, Stephanie Ko, Shi Low, Shefaly Shorey. 6883

Telehealth Usability, Engagement Patterns, and Technical Infrastructure in Managing Noncommunicable Diseases Among Health Care Professionals in Brazil, Ghana, Honduras, and the United Kingdom: Multinational Cross-Sectional Study (e64070)
Ken Brackstone, Roberta Gonçalves, Adriana Pagano, Zilma Reis, Seth Afagbedzi, Lysien Zambrano, Tainá Lopes, Sarah Cordeiro, Julia Nunes, Wagner Meira Jr, James Batchelor, Antonio Ribeiro. 6897

Telerehabilitation in Community Stroke Services: Mixed Methods Evaluation of Current Practice and Lessons for Sustained Use (e87741)
Elizabeth Chandler, Charlotte Dorer, Valerie Pomeroy, Nicola Hancock. 6909

Development and Preliminary Validation of ORCA-PD, an Online Rapid Cognitive Assessment for Parkinson Disease: Mixed Methods Study (e90057)
Avigail Lithwick Algon, Sarah Brisman, Chi-Ying Lin, William Saban. 6924

Safety of Telemedicine Versus In-Person Care for Patients With Tracheal Devices: Propensity Score–Matched Cohort Study (e95479)
Pedro Ciaralo, Paulo Cardoso, Helio Minamoto, Benoit Bibas, Carlos Carvalho, Paulo Pego-Fernandes. 6939

Digital Monitoring of Pre-Exposure Prophylaxis Users Through Electronic Patient-Reported Outcome Measures and Electronic Patient-Reported Experience Measures: Multicenter Prospective Study on Feasibility, Safety, and Predictive Modeling of Digital Engagement (e87592)
Gabriel Mercadal-Orfila, Joaquín Serrano, Tilman Mijares, Lluís Vidal, Adrian Curran, Salvador Herrera-Pérez. 6952

Accuracy of Visual Inspection Alone to Assess Joint Effusions of the Hand: Cross-Sectional Study (e86261)
Carrie Ye, J Mitchell, Shahad Al-Matar, Mena Bishay, Niall Jones, Dalton Sholter, Dylan Johnson, Steven Katz, Anthony Russell, Mohamed Abdalla. 6973

Rapid Deployment of Telemedicine in HIV Care: Mixed Methods Study of Providers' Attitudes and Perceptions (e75933)
Shannon Galvin, Artur Queiroz, Juan Rivera, Mia Calamari, Katherine Wright, Lisa Hirschhorn, Laura Rusie, Patrick Janulis, Rebecca Kumar, Frank Palella, Mary Masters, Jacqueline Bannon, Susheel Reddy, Claudia Hawkins. 6982

XSL•FO
RenderX

Characteristics of Patients Accessing Outpatient Oncology Services Virtually and Predictors of Subsequent Unplanned Emergency Department Presentations in 78,323 Adults in Australia: Retrospective Cohort Study (e87694)
Ramya Walsan, Reema Harrison, Johanna Westbrook, Ashfaq Chauhan, Michelle Moscova, Anita Vandyke, Elizabeth Manias, Tracey Webster, Natalie Taylor, Prince Peprah, Rebecca Mitchell. 7036

Understanding How a Digital Platform for Chronic Disease Management Can Enable and Limit Patient Self-Care: Qualitative Study (e71875)
Lysanne Lessard, Mark de Reuver, Kerri-Anne Mullen. 7046

Telemedicine Adoption for Managing Chronic and Rare Diseases in Indonesia During and Beyond the COVID-19 Era: Qualitative Study (e83462)
Christine Pratama, Rima Nurlianti, Cornelius Herstatt, Moritz Goeldner. 7061

Telehealth Intervention to Reduce Sedentary Behavior in Older Adults With Type 2 Diabetes: Development and Feasibility Study (e80827)
Xiaoyan Zhang, Dan Yang, Sihan Chen, Lei Wang, Xin Li, Ziyan Wang, Meiqi Meng, Xuejing Li, Hongzhan Jiang, Yufang Hao, Jiayin Luo. . . 7 0 7 7

Association Between Physician Communication Features and Patient Outcomes in Telemedicine: Retrospective Cross-Sectional Observational Study (e86977)
Yipei Wang, Shu Wang, Ke Zhang, Zhijie Liu, Qingbian Ma, Hong Ji, Zheng Hou, Tracy Liu, Xuedong Xu, Xinxia Wu, Changxiao Jin. 7092

Digital Primary Health in Rwanda: Qualitative Study of User Experiences and Implementation Lessons From Babyl's Telemedicine Platform (e84832)
Sabine Musange Furere, Gashaija Absolomon, Nathalie Umutoni, Stella Matutina Umuhoza, Felix K Rubuga, Yvonne Nsaba-Uwera, Patrick Singa Muhoza, Regis Hitimana, Jeanine Condo, James Humuza. 7110

Telehealth Use and Modality Choice Among US Adults: Shorrocks-Shapley Decomposition of a 2022 Cross-Sectional National Survey (e81879)
Corneliu Bolbocean, Corey Hayes, Cari Bogulski. 7127

Attitudes Toward Video Consultations From the Perspective of Physicians and Psychotherapists in German Outpatient Care After the COVID-19 Pandemic: Survey Study (e73757)
Lara Kleinschmidt, Juergen Wasem, Nikola Blase, Beatrice Nauendorf, Juliane Malsch, Matthias Brittner, Paul Brandenburg, André Aeustergerling, Theresa Hüer. 7142

Communication Challenges and Mitigation Strategies in Primary Care Virtual Consultations: Qualitative Study (e79399)
Ahmed Alboksmaty, Tetiana Lunova, Ara Darzi, Ana-Luisa Neves. 7158

Gender Concordance and Patient Outcomes in Indian Telemedicine: Retrospective Cross-Sectional Quantitative Study of 286,000 Consultations (e78311)
Nafisa Vaz, Vishalkumar Jani. 7172

Virtual Delivery of Early Psychosis Care: Retrospective Cohort Study of Factors Associated With Initial Engagement (e81313)
Trinity Vey, Nicole Kozloff, George Foussias, Wei Wang, Albert Wong, Aristotle Voineskos, Nicole Davis-Faroque, Alexia Polillo. 7185

Rising to the Challenge of Early Screening in Primary Health Care Through the Web Italian Network for Autism Spectrum Disorder (Win4ASD) in the Pediatric Population: Retrospective Observational Study (e74302)
Noemi Buo, Eleonora Rosi, Silvia Busti Ceccarelli, Mariarosa Ferrario, Valerio Maiorca, Erika Morandi, Nicole Viganó, Ivan Limosani, Laura Falcone, Massimo Molteni, Paola Colombo. 7196

XSL•FO
RenderX

The Effect of Veteran Race and Socioeconomic Status on Enrollment in Remote Patient Monitoring for Hypertension: Retrospective Observational Cross-Sectional Study (e78423)
Hannah Friedman, Hillary Mull, Jenice Guzman-Clark, Daniel Sturgeon, Marva Foster. 7208

Impact of Tele-Intensive Care Units on the Clinical Outcomes of Critically Ill Patients With COVID-19: Retrospective Cohort Study (e64996)
Thais Midega, Fabio Hohmann, Renato Chaves, Guacyra Almeida, Vivian Leoneza, Jennifer Cabral, Bianca Vitalino, Emanuelle Camboim, Nelma Machado, Ricardo Melo, Jorge Feliciano, Breno Cardoso, Leonardo Ferraz, Thiago Corrêa, Maura Santos, Renata Albaladejo Morbeck, Adriano Pereira. 7223

Supporting Access to Care Through Peripheral Devices and Patient-Generated Health Data: Qualitative Study (e88405)
Ashley Griffin, Caroline Gray, Victoria Ngo, Camila Chaudhary, Stephanie Shimada, Amanda Midboe, Donna Zulman. 7234

Telehealth Delivery of the Homeostasis–Enrichment–Plasticity Approach for Premature Infants With Developmental Risks: Exploratory Feasibility Study (e86883)
Aymen Balikci, Gamze Sirma, Izgi-Miray Demirbag, Teresa May-Benson, Hatice Sozen, Ayse Aracikul Balikci, Gul Ilbay. 7248

Effectiveness of Telemedicine vs Face-to-Face Consultation in Fighting COVID-19: Retrospective Cohort Study of Adult Patients With COVID-19 in a Primary Care Setting (e74046)
Fangfang Jiao, Ka Ho, Lapkin Chiang, Siu Ko, Catherine Chen. 7266

Virtual Consultations for People With Intellectual Disabilities in General Practice and Community Care: Mixed Methods Qualitative Study (e81173)
Freda Mold, Anna Cox, Vicki Tsianakas, Harm van Marwijk, Paul Shanahan, Treena Parsons, Jo Armes. 7282

Mode Effects Between Mobile Web and Telephone Surveys on Patient Experience Scores in South Korea: Secondary Analysis of a Randomized Controlled Trial Under Various Missingness Scenarios (e79398)
Young-Geun Choi, Bon Koo, Yeongchae Song, Young Do. 7302

Fusing Specialized Surveys of Rare Populations to Larger Surveys for Generalized Inference: Cross-Sectional Survey Study (e86059)
Karilynn Rockhill, Elizabeth Bemis, Nicole Schow, Heather Olsen, Kyle Beekman, Evelyn Fox, Andrew Monte, Joshua Black. 7314

After-Hours Use of the Electronic Health Record Among Medical and Surgical Specialists After Implementation of a System-Wide Integrated Clinical Information System in Alberta, Canada: Longitudinal Descriptive Study (e76872)
Robert Pauly, Tania Stafinski, Helia Koosha, Melita Avdagovska, Maliheh Hadizadeh, David Bigam, Timothy Graham, Craig Kuziemsky, Narmin Kassam, Devidas Menon. 7329

Trust Building in Internet-Based Home Care Among Loyal Patients: Qualitative Study (e88860)
Yan Yang, Fei Lu, Xinxin Wang, Leiwen Tang, Luchen Pan, Yu Zhang, Hongling Sun, Jia Feng, Chenling Zhu, Meijuan Lan. 7339

Child Vaccination Status and Behavioral and Social Drivers of Vaccination Among Their Caregivers in the Philippines: Cross-Sectional Survey Study Comparison of Household, Mobile, and Online Modes (e81059)
Kimberly Bonner, Mikka Hipol, Dominique Sy, Rivandra Royono, Douglas Johnson, Isabel del Rosario, Alice Redfern, Darahlyn Biel-Romualdo, Man Wong, Eugene Lam, Shibani Kulkarni, Kirsten Ward, Romel Lacson, Meng-Yu Chen, James Miraflor, Devon Pacial, Rowena Bunoan, Hugo Catan, Talya Shragai. 7354

Illness Experiences in Young and Middle-Aged Patients With Dilated Cardiomyopathy on Chinese Social Media: Qualitative Study (e76918)
Weijing Chen, Hanyu Xiao, Chenchen Feng, Jialin Li, Jinhua Jin, Ruimin Mu. 7368

XSL•FO
RenderX

Sociodemographic Paradoxes and Enrollment Differences in In-Person Versus Online Recruitment to a Mobile Health Smoking Cessation Intervention for Food-Insecure Adults: Secondary Analysis of a Randomized Controlled Trial (e80530)
Charles Hoogland, Steven Sutton, Sarah Jones, Bence Nagy, Samuel Brockway, David Himmelgreen, Thomas Mantz, Michael Businelle, Ya-Chen Shih, Jennifer Vidrine, Damon Vidrine. 7381

Willingness to Share Internet Use Data for Research on Early Disease Detection: Cross-Sectional Survey (e85637)
Christina Derksen, Mel Ramasawmy, Sanjula Arora, Wynsee Lau, Suzanne Scott. 7399

From Searching to Coping, How Chinese Patients With Breast Cancer Navigate Web-Based Health Information: Semistructured Interview Study (e80363)
Jiahui Liu, Xingfeng Li, Xuying Li, Jiejun Chen, Dan Zhang, Cuiling Qiu. 7420

Sociodemographic Drivers of Recruitment and Attrition in Digital Neurological Research: Longitudinal Cohort Study (e83432)
Peyman Nejat, Ashley Bachman, Vicki Stubbs, Joseph Duffy, John Stricker, Vitaly Herasevich, David Jones, Rene Utianski, Hugo Botha. . . . 7 4 3 6

Using Epidemiological Test Diagnostics to Select Fraud Detection Methods: Secondary Analysis of Quantitative Cross-Sectional Survey Data (e85161)
Rachel Willard-Grace, Tali Klima, Mansi Dedhia, Emily Lo, Annie Nisnevich, Allison Gray, Holly Henry. 7450

The Feasibility of Smartwatch Micro–Ecological Momentary Assessment for Tracking Eating Patterns of Malaysian Children and Adolescents in the South-East Asian Community Observatory Child Health Update 2020: Cross-Sectional Study (e73435)
Richard Lane, Louise Millard, Ruth Salway, Chris Stone, Andy Skinner, Sophia Brady, Jeevitha Mariapun, Sutha Rajakumar, Amutha Ramadas, Hussein Rizal, Laura Johnson, Tin Su, Miranda Armstrong. 7461

Integrating a Large Language Model to Streamline Nursing Handover Documentation Across Multiple Hospitals in Taiwan: Development and Implementation Study (e81604)
Ray-Jade Chen, Mai-Szu Wu, Lung-Wen Tsai, Shy-Shin Chang, Shu-Tai Shen Hsiao, Yu-Sheng Lo. 7475

Digital Health Technology Use Among Rehabilitation Professionals in China: Multi-Province Cross-Sectional Survey (e90623)
Shuning Duan, Zi-Ru Wang, Xier Chen, Guoxin Ni. 7497

Understanding User Intent in Code-Mixed Sexual and Reproductive Health Queries in Urban India: Hierarchical Classification Approach Using Large Language Models (e86545)
Sumon Dey, Manvi S, Aradhana Thapa, Meet Shah, Zeel Mehta, Shraddha Kapile, Tanvi Divate, Suhani Jalota, Azra Ismail. 7515

Combining Machine Learning Models and Screening to Enhance Suicide Risk Identification for American Indian Patients: Retrospective Cohort Study (e82669)
Novalene Goklish, Emily Haroz, Rohan Dayal, Valentín Sierra, Roy Adams, Francene Sinquah, Paul Rebman, Jacob Taylor. 7537

Longitudinal Effects of a Smartphone Game (Tumaini) for HIV Prevention Among Kenyan Adolescents: 45-Month Trajectories of Condom Use–Related Proximal Outcomes From a Randomized Controlled Trial (e83982)
Kate Winskell, Gaëlle Sabben, Haowen Qin, Calvin Mbeda, Sophie Goldenberg, Ken Ondeng'e, Richard Ndivo, Judith Arego, Xinwei He, Robert Bednarczyk, Kelli Komro, Robert Lyles, Victor Mudhune. 7548

Detection of Microbehavior Intervals for Predicting Mental Health: Clinically Relevant and Advanced Multimodal Temporal Analysis (e87049)
Sapir Gershov, Charlotte Hilberdink, Yiwen Zhao, Sarah Birnbaum, Victoria Mueller, Stephen Wall, Katharina Schultebraucks. 7565

XSL•FO
RenderX

Digital Illness Narratives of Young Chinese Adults With Diabetes on RedNote: Qualitative Narrative Analysis (e96543)
Zikun Liu, Donghan Fu, Yingjie Liu, Ying Meng. 7579

Patient-Reported Experiences With Viewing and Understanding Test Results in Patient Portals: Cross-Sectional Survey Analysis (e94098)
Chelsea Richwine, Bryan Steitz, Jordan Everson. 7591

Impact of Financial Incentives on Electronic Health Record–Driven Recruitment of Underrepresented Communities in Research: Randomized Controlled Trial (e86110)
Christina Ping, Andrew Davis, Waishun Gema, Foluso Ademuyiwa, Mark Fiala. 7606

Understanding Patient-Reported Offenses in Electronic Health Records: Cross-Sectional Mixed Methods Survey (e86178)
Saija Simola, Sari Kujala, Åsa Cajander, Anna Kharko, Bridget Kane, Bo Wang, Rose-Mharie Åhlfeldt, Maria Hägglund. 7616

Use, Utility, and User Experience of Cloud-Based Medical Imaging in Pulmonary Nodule Care in China: Mixed Methods Study (e86745)
Ziguo Chen, Junhan Wu, Yunying Chen, Weitao Zhuang, Jiaxuan Huang, Weifeng Zhong, Zijie Li, Shuhuan Xie, Chaofan Liu, Guojie Lu, Guibin Qiao. 7657

Perspectives on Health Data Sharing Among Patients With Somatic and Mental Health Diseases: Focus Group Study (e79990)
Sabrina Fesl, Caroline Lang, Falk Verhees, Jochen Schmitt, Madlen Scheibe. 7670

Physicians' Perceptions and Demand Regarding Clinical and Secondary Use of Patient-Generated Health Data: Cross-Sectional Survey (e86368)
Ye-Eun Park, Sang Beck, Yura Lee. 7693

Adaptation and Validation of the Social Media Cyberbullying Victimization Scale Among Chinese College Students: Cross-Sectional Study (e88857)
Chen Xu, Suping Wang, Yujie Liu, Xin Ge, Xue Yang, Xiaohong Fan, Li Li, Ying Wang, Yong Cai. 7709

Development of a Digital Health Literacy Assessment Framework for Older Adults: Delphi-Based Study (e82334)
Xingxia Zhang, Yongqing Yuan, Jie Jiang. 7726

The WONE Index as a Multidimensional Assessment of Stress Resilience: A Development and Validation Study (e81714)
Lydia Roos, Destiny Gilliland, Kelsey Julian, Reeva Misra. 7742

Program Theory and Core Outcome Set Development for a Technology-Assisted Counseling Intervention in Dementia: Multimethods Study (e81669)
Dorothee Bauernschmidt, Anja Bieber, Ronja Hubrich, Janina Wittmann, Gabriele Meyer. 7772

Validation of the Simplified Chinese Clear Communication Index Using Diabetes Education Materials: Instrument Adaptation and Validation Study (e83935)
Lu Liu, Yi Chen, Xiaoqi Li, Hanyu Du, Rong Gui, Huiqing Jin, Fengfei Chen, Yujun Lin, Qiuhuan Huang, Yingge Tong. 7798

A Digital Tool for Assessing the Distinct Effects of Depression, Anxiety, and Attention-Deficit/Hyperactivity Disorder (ADHD) on Children's Emotional Cognitive Bias: Cross-Sectional Study (e86286)
Sang-Eon Park, Jisu Chung, Sang Lee. 7814

XSL•FO
RenderX

Authoritative Textbook-Augmented Large Language Models for High-Altitude Public Health Medical Education in the Xizang Autonomous Region: Cross-Sectional Comparative Evaluation Study (e92852)
Kun He, Qiming Xiao, Wangyang Chen, Lisha Jing, Yabing Wang, Shuai Li, Daiyu Yang, Hemiao Xu, Ke Pang, Ruoyu Xiao, Zhashilamu Liu, Deji Zhuoga, Ruxuan Chen, Jingyi Li, Long Chang, Yangzhong Zhou, Zhe Zhang, Ran Li, Lujing Ying, Rutong Li, Hongwei Wang, Xin Yin, Ge Zhen, Siyi Cai, Qijun Shan, Qiang Wang, Danzeng Zhuoga, Ciren Yangjin, Gesang Luobu, Tu Ji, Dong Wu. 7830

Training Clinicians in Private Practice in Family-Based Treatment for Anorexia Nervosa: Randomized Controlled Trial Comparing Two Online Approaches (e89999)
James Lock, Daniel Le Grange, Brittany Matheson, Bohye Kim, Cara Bohon, Booil Jo. 7848

The Influencing Factors of Medical Postgraduates' Usage Intention Toward Artificial Intelligence–Generated Content Tools in Academic Research: Qualitative Analysis (e77928)
Chen Wang, Liu Wang, Xuejiao Zhang, Huiying Qi. 7861

Opportunities and Concerns of Gamified, Extended Reality for Home-Based Motor Rehabilitation for Children With Brain Injury: Qualitative Case Study on Design Elements Related to the Engagement and Fatigue Perspectives (e84013)
Eivind Kolstad, Nenad Pavel, Alexis Cartajenas, Trust Saidi, Ingvild Hurum Rosseland, Åse Bergheim, Nora Synnøve Mørk, Kathinka Granum Selmer-Olsen, Parisa Gazerani, Shefaly Shorey, Minna Pikkarainen. 7875

A Peer-Led, Nurse-Involved Blended Online and Offline Peer Support Program (PNO2PSP) for Psychosocial Adjustment in Young- to Middle-Aged Patients With Breast Cancer: Cluster Randomized Clinical Trial (e86097)
Yiheng Zhang, Haiyan He, Ye Zhang, Youxian Zhong, Chunhong Luo, Juanjuan Chen, Lili Chen, Ni Zhang, Shihao Sun, Baoyi Zhang, Jia Fang, Jingwen Yan, Meifen Zhang. 7927

Digital Health–Based Peer Support Ecosystem for Gestational Diabetes Mellitus in Vietnam (VALID II Study): Multistakeholder Cocreation and Pilot Study (e82434)
Ngoc Dang, Dan Meyrowitsch, Ditte Linde, Christina Vinter, Thanh Nguyen, Dang Nguyen, Thi Nguyen, Hieu Le, Thi Vu, Thi Tran, Tine Gammeltoft, Jens Søndergaard. 7957

Factors Influencing Continuance Intention for Online Consultations Among Survivors of Cancer: Grounded Theory Study (e84644)
Yutang Yao, Musi Zhang, Shanshan Peng, Zhuzhong Cheng, Yun Duan. 7975

The Structure of Psychopathology on Reddit: Network Analysis of Mental Health Communities in Relation to the ICD Diagnostic System (e80958)
Bojan Evkoski, Srebrenka Letina, Petra Kralj Novak. 7988

Challenges of Standard Pediatric Epilepsy Monitoring and the Potential Benefits of Contactless Sensor Technologies: Exploratory Qualitative Study (e83009)
Natalie Öhl, Dilbar Strack, Stefan Kusnik, Isabell Link, Marlene Hofmann, Tobias Steigleder, Alina Weisser, Christoph Ostgathe, Regina Trollmann, Maria Heckel. 8005

Development and User-Centered Evaluation of Smart Systems for Loneliness Monitoring in Older Adults: Mixed Methods Study (e81027)
Yi Zhou, Jessica Rees, Faith Matcham, Ashay Patel, Michela Antonelli, Anthea Tinker, Sebastien Ourselin, Wei Liu. 8019

Attitudes and Willingness to Participate in Drug Clinical Trials Among Patients With Cancer: Multistage Qualitative Study (e88758)
Sijia Yu, Mengmeng Lyu, Haiping Ma, Keying Xu, Haiyan Li, Lili Ma, Mengting Ji. 8041

The Implementation of a Business Process Model and Notation for Modeling Patient Health Care Trajectories: Systematic Review (e78506)
Jean-Baptiste Gartner, Paolo Landa, Matthew Haren, Célia Lemaire, Elena Tanfani, Catherine Paquet, Frédéric Bergeron, André Côté. 8060

XSL•FO
RenderX

Experiences of Alert Fatigue and Its Contributing Factors in Hospitals: Qualitative Study (e78676)
Nicki Newton, Adeola Bamgboje-Ayodele, Rowena Forsyth, Amina Tariq, Jing Huang, Ramyasri Yannam, Daniel Lalor, Andrew Sobey, Melissa Baysari. 8083

Improving the Understandability of Clinical Guidelines: Development and Evaluation of a GPT-4–Based Pipeline (e81915)
Matthew Jones, Melissa Torgbi, Harish Tayyar Madabushi. 8097

Interorganizational Mechanisms for Developing and Implementing Clinical Decision Support Systems in Primary Care: Exploratory, Qualitative Case Study (e83084)
Jesse Santema, Jeroen van Wijngaarden, Eric Hiddink, Fleur Deken, Maaike Kleinsmann, Hendrikus van Os. 8108

Assessing Health Care Professionals' Perceptions of a New System in Clinical Workflows: Systems Engineering Initiative for Patient Safety–Based Consensual Qualitative Research (e86166)
Ye-Eun Park, Minsu Ock, Jae-Ho Lee, Dae-Hyun Ko, Hak-Jae Lee, Taezoon Park, Junsang Yoo, Yura Lee. 8126

Advancing Nursing Data Integration Through a Nursing Minimum Dataset for the Conceptual and Technical Development of a "Fall Prevention" Data Module: Development Study (e82417)
Sarah Milkov, Antonia Schmidt, Anja Burmann, Niklas Tschorn, Marcel Klötgen, Wolfgang Deiters, Christian Potthoff, Kirsten Neveling, Yvonne Weber, Maren Keuchel, Daniela Holle. 8161

Classifying American Society of Anesthesiologists Physical Status With a Low-Rank–Adapted Large Language Model: Development and Validation Study (e89540)
Min-Chia Chen, Shanq-Jang Ruan, Jo-Hsin Wu, Pei-fu Chen. 8174

Uptake of Clinical Decision Support Systems Among Health Care Professionals in Six European Countries and the United States: Cross-Sectional Survey Within the I-CARE4OLD Project (e85071)
Collin Exmann, Anna-Maria Hiltunen, Ira Haavisto, Anna Salminen, Maikki Messo, Mikko Nuutinen, Mark Hoogendoorn, Wiebe Boorsma, Elizabeth Howard, Vanja Peši , Mor Alon, Federica Mammarella, Rosa Liperoti, Olena Švihnosová, Daniela Fialová, Natalia Drapała, Katarzyna Szczerbi ska, Anja Declercq, Hein van Hout, Johanna De Almeida Mello. 8195

Designing Emotion-Integrated Visualizations for Kidney Function Self-Management: User-Centered Design Study With Youth Transplant Recipients and Caregivers (e80481)
Chelsea Ng, Julia Dunbar, Lily Jeffs, Sanaa Syed, Nathan Pan, Hyeyoung Ryu, Jodi Smith, Jaime Snyder, Wanda Pratt, Ari Pollack. 8208

High Patient Willingness to Grant Broad Consent for Real-World Data Use in Rheumatology—Implications for Real-World Data Platform Governance: Cross-Sectional Study (e80347)
Jutta Richter, Antonia Becker, Tim Filla, Hasan Acar, Waldemar Ockert, Dominykas Kriauciunas, Edith Aggarwal, Markus Schröder, Ralf Hansen, Jörg Distler, Matthias Schneider. 8227

Guide on Selection of Optimal Motivational Themes for Use in a Clinical Trial Recruiting Black US Adults: Survey Study (e75857)
Tamunotonye Harry, Jingyi Cao, Zaib Hussain, Ruth-Alma Turkson-Ocran, Stephen Juraschek, Timothy Lahey, Timothy Plante, Yuanyuan Feng. 8241

Methods for Detecting Suspicious Information From Individual Transactions of Pharmaceutical Products via Twitter (now X): Retrospective Observational Study (e91103)
Ruoyu Zhang, Kazuko Kimura, Naoko Yoshida. 8262

Exploring Attitudes Toward AI-Based Contactless Sensors in Health Among Five Stakeholder Groups: Qualitative Study (e75783)
Eva Hille, Hannah Bleher, Birte Coppers, Sophie Engelhardt, Bjoern Eskofier, Lorenz Garbe, Maria Heckel, Anna-Maria Liphardt, Alina Manlig, Nicolas Rohleder, David Samhammer, Carla Spielmann, Matthias Braun. 8276

Ethical Handling of Occupational Health and Safety Data in the Fire Service: Empirical Interview and Focus Group Study of Firefighter and Fire Service Leadership Privacy Preferences (e84465)
Rachel Topazian, Aleksandra Wec, Joseph Ali, Shannon Frattaroli, Cassandra Crifasi. 8297

XSL•FO
RenderX

The Ethics of Leveraging Routinely Collected Patient Data for AI Development: Mixed Methods Study (e79863)
Menno Maris, Joanna Klopotowska, Ronald Cornet, Mariëtte Van den Hoven, Joris Lieverse, Daniel Fernández-Llaneza, Marieke Bak, LEAPfROG Consortium. 8309

Ethical Knowledge, Challenges, and Institutional Strategies Among Medical AI Developers and Researchers: Focus Group Study (e79613)
Sophia Fantus, Jinxu Li, Tianci Wang, Lu Tang. 8330

Educating Students About Digital Health Research Ethics: Curricula Review and Expert Interview Study (e82861)
Yier Zhu, Brian McInnis, Tanya Punater, Brittany York, Camille Nebeker. 8342

Strategy for Hepatitis B and C Virus Testing Campaigns Through Web Services and Digital Advertising in Japan: Nationwide Cross-Sectional Study With Correspondence Analysis (e89585)
Kosuke Sakai, Shoko Nakazawa, Yuko Furuya, Kota Fukai, Kei Sano, Masaaki Korenaga, Masayuki Tatemichi. 8357

Factors Influencing Universal Coverage of AI-Assisted Cervical Cancer Screening: Qualitative Study Based on the Macro Model of Health System (e75372)
Lu Ji, Xinke Zhou, Lan Yao. 8371

Opposition to Youth e-Cigarette Prevention Campaigns on Twitter and TikTok: Cross-Platform Observational Mixed Methods Analysis (e83791)
Chandler Carter, Simon Page, Mateusz Borowiecki, Ganna Kostygina, Sherry Emery, Miao Feng. 8385

Suicidal Thoughts and Behaviors Among Chinese Adolescents in Relation to Negative Life Events, Internet Addiction, and Sexual Abuse: Cross-Sectional Study (e85371)
Juanfang Zhu, Qinqin Jiang, Feng Zhang, Hao Liu, Shaojie Yu, Jinhai Sun, Zhe Zhao, Lijuan Liu, Lei Yuan. 8398

Embracing Co-Design and Interprofessional Teamwork to Build an Innovative Dashboard for a National Social Needs Screening and Referral Clinical Intervention in the Veterans Health Administration: Design and Development Study (e81846)
Lauren Russell, Sarah Leder, Jaime Boris, Alicia Cohen, Christopher Halladay, Kathleen Mitchell, Sydney Ruggles, Andrea Berkheimer, Chava Sonnier, Michelle Wilcox, Ernest Moy, Meaghan Kennedy. 8414

Developing and Validating a Machine Learning Algorithm to Predict the Risk of Incident Opioid Use Disorder Among OneFlorida+ Patients: Prognostic Modeling Study (e79482)
Jabed Faysal, Weihsuan Lo-Ciganic, Walid Gellad, Yonghui Wu, Christopher Harle, Khoa Nguyen, James Huang, Gerald Cochran, Debbie Wilson, Stephanie Staras, Siegfried Schmidt, Eric Rosenberg, Danielle Nelson, Shunhua Yan, Gary Reisfield, William Greene, Courtney Kuza, Md Hasan. 8429

Breast Cancer Screening Knowledge and Sentiments in Singaporean Women: Mixed Methods Study Using Topic Modeling, Sentiment Analysis, and Structured Questionnaire Data (e78439)
Peh Ho, Zi Lim, Jenny Liu, Nur Mohamed Riza, Ying Chew, Yi Lim, Hui Tan, Su-Ann Goh, Han Oh, Chi Chin, Sing Kwek, Zhi Zhang, Desmond Ong, Swee Quek, Sujith Wijerathne, Philip Iau, Mikael Hartman, Jingmei Li. 8444

Integration of Digital Therapeutics Into Occupational Rehabilitation in Germany: Multilevel Simulation Study (e93793)
Christoph Wallner, Sonja Schmidt, Felix Reinkemeier, Dagmar-Maria Dick, Marcus Lehnhardt, Marius Drysch. 8464

Rapid Epidemiological Data Collection on Social Media for COVID-19: Comparative Study Between Online Surveys and Conventional Cohorts (e80311)
Maged Mortaga, Hendrik Nunner, Sydney Paltra, Leonard Stellbrink, Jens Friedel, Manuela Harries, Jessica Krepel, Berit Lange, MuSPAD Study Group, Viola Priesemann, André Calero Valdez. 8478

XSL•FO
RenderX

Experiences of Home-Dwelling Older Adults and Their Family Caregivers With Digital Health Services: Qualitative Study (e89496)
JiaYi Ruan, CuiXia Guan, QingQing Chen, Ting Yao, Yun Liu, YuXia Zhang, LunFang Xie. 8498

Harnessing Participatory Surveillance Cohorts and Proxy Indicators to Dynamically Track Epidemic Trends and Undiagnosed COVID-19 Infections in Singapore: Longitudinal Observational Study (e85608)
Sheng Soh, Aung Aung, Wei Ong, Kangwei Zeng, Jean-Marc Chavatte, Lin Cui, Raymond Lin, Vanessa Lim, May Lwin, I-Cheng Chen. 8513

National Acceptance and Determinants of Immersive Extended Reality in Health Care in China: Cross-Sectional Study (e88000)
Jiaying Li, Patricia Davidson, Helen Chan, Cho Wong, Xiang Qi, Zengjie Ye, Ankie Cheung, Daniel Fong, Yibo Wu, Junxin Li. 8531

The Open Syndrome Definition as a Machine-Readable Standard for Public Health: Design and Implementation Study (e86249)
Ana Ferreira, Aleksandar Anžel, Izabel Marcilio, Helen Hughes, Alex Elliot, Jude Kong, Madlen Schranz, Alexander Ullrich, Georges Hattab. . 8 5 5 2

Exploring Reddit Discourse and Information Needs Surrounding Extreme Heat: Topic, Sentiment, and Engagement Analysis (e82426)
Melissa MacKay, Soroush Zamani Alavijeh, Sydney Gosselin, Fattane Zarrinkalam, Jennifer McWhirter. 8574

Detection of Antithrombotic-Related Bleeding in Older Inpatients: Multicenter Retrospective Study Using Structured and Unstructured Electronic Health Record Data (e77809)
Claire Coumau, Frederic Gaspar, Mehdi Zayene, Elliott Bertrand, Lorenzo Alberio, Christian Lovis, Patrick Beeler, Fabio Rinaldi, Monika Lutters, Marie-Annick Le Pogam, Chantal Csajka, SwissMADE Collaborators. 8589

Behavioral Dynamics of AI Trust and Health Care Delays Among Adults: Integrated Cross-Sectional Survey and Agent-Based Modeling Study (e82170)
Xueyao Cai, Weidong Li, Wenjun Shi, Yuchen Cai, Jianda Zhou. 8608

Screen Time and Chronic Pain Health: Mendelian Randomization Study (e78233)
Jiahui Jiang, Chunyan Pu, Jiarui Cai, Chuan Yu, Zhenmi Liu, Chenghan Xiao. 8628

Genre-Specific Gaming Addiction and Flourishing in Adolescents: Cross-Sectional Survey Study (e89319)
Yan Luo, Xiaofei Chen, Yaoyu Liu, Wanjia Hong, Miner Fan, David Thomas, Sean Mullen. 8645

Implementation of Reference Framework for Preventive Care for Older Adults: Cross-sectional Qualitative Study (e78814)
Claire Zhong, Mingtao Chen, Chung Lo, Man YIM, Xiaoshu Zhang, William Wong, Junjie Huang, Martin Wong. 8665

Virtual Care and Health Care Access: Pragmatic Evaluation of Implementation, Acceptance, and Use in General Practice and Aged Care Homes (e89019)
Adeola Bamgboje-Ayodele, Tamasha Jayawardena, Anita Higgins, Margaret Watkiss, Fiona Robinson, Susan Kurrle, Pan Teng, Melissa Baysari, Meredith Makeham. 8686

Institutionalizing Digital Parenting Programs in Low Resource Settings in China: Comparative Case Study of Health Care and Education Sectors Using the RE-AIM Framework (e79848)
Xinyu Shi, Ruochen Ruan, Yi Qie, Jamie Lachman, Na Zhong, Zuyi Fang. 8703

A Complex Digital Health Intervention to Support People With HIV: Organizational Readiness Survey Study and Preimplementation Planning for a Hybrid Effectiveness-Implementation Study (e76327)
Jacqueline Hodges, Wendy Cohn, Amanda Castel, Tabor Flickinger, Ava Waldman, Michelle Hilgart, Olivia Kirby, Sylvia Caldwell, Karen Ingersoll. 8724

XSL•FO
RenderX

Using the NASSS–Complexity Assessment Tool to Evaluate the Implementation of “Cadê O Kauê?”: Chat-Story Intervention for Youth Participation in Mental Health Promotion in Brazil (e79106)
Sakshi Setia, Gabriela Pavarini, Mona Koshkouei, Sheila Giardini Murta. 8743

Explaining the Use Behavior of Digital Technologies in Pediatric Rehabilitation: Structural Equation Modeling Analysis of a Cross-Sectional European Survey (e80355)
Johanne Mensah Gourmel, Sylvain Brochard, Saranda Bekteshi, Elegast Monbaliu, Gwenaël Cornec, Anca Grigoriu, Christopher Newman, Marco Konings, Javier De La Cruz, Christelle Pons. 8759

Health Communication Campaign Performance During the HEALing Communities Study: Cross-Sectional Examination of Digital Advertising Methods (e75458)
Nicky Lewis, Jennifer Reynolds, Diane Krause, Philip Reeves, Jamie Luster, Michael Stein, Sharon Walsh, Amy Farmer, Michelle Lofwall, Monica Roberts, Hilary Surratt, Brooke Crockett, Kara Stephens, Kelli Bursey, Kristin Mattson, Michael Slater. 8774

Engagement With Meditation Apps: Cross-Sectional Survey of Use and Associations (e71960)
Julia Adams, Jonathan Davies, Prai Wattanatakulchat, Julieta Galante, Felicity Miller, Simon D'Alfonso, Nicholas Van Dam. 8790

A Stepped Care Approach to Promoting Engagement in a Digital Health Intervention: Secondary Analysis of a Randomized Controlled Trial (e80921)
Esha Dwibedi, Wen You, Lee Ritterband, Donna-Jean Brock, Annie Reid, Christina Frederick, Jamie Zoellner. 8813

Development of Virtual Reality Health Literacy: Delphi Expert Consensus Study (e85842)
Junghee Yoon, Mangyeong Lee, Dokyoon Kim, Joungwon Park, Su Kim, Jiyoon Han, Juhee Cho. 8830

Virtual Exposure With Response Prevention for Obsessive-Compulsive Disorder: Randomized Controlled Trial (e79326)
Lara Rolvien, Lena Jelinek, Luzie Lohse, Steffen Moritz, Lisa Borgmann, Jürgen Gallinat, Simone Kühn, Franziska Miegel. 8846

Automated Inference of Social Anxiety From Behavior in Social Virtual Reality: Cross-Sectional Observational Study (e79147)
Gayoung Son, Marius Rubo. 8870

Automated Identification of Nursing Diagnoses and Interventions From Nursing Records Using a Retrieval-Augmented Large Language Model Approach: Quantitative Study (e89850)
Minye Li, Dianjie Chen, Qun Xiao, Zhengji He, Yanyan Zhang, Jinlan Zhong, Yiwei Luo, Hui Ma. 8888

Temporal Annotation of German Clinical Language in Real and Synthetic Clinical Documents: Corpus Development and Baseline Tagger Validation Study (e71458)
Luise Modersohn, Udo Hahn. 8903

Extracting and Classifying Drug Discontinuations From Estonian Electronic Health Records: Development and Validation Study (e86183)
Hendrik Šuvalov, Nikita Umov, Maria Malk, Markus Haug, Sven Laur, Marek Oja, Sirli Tamm, Sulev Reisberg, Jaak Vilo, Raivo Kolde. 8929

Investigating the Effect of Hospital Infection Control Informatization on Optimizing Microbiological Specimen Submission Before Antibiotic Therapy: Failure Mode and Effects Analysis (e78118)
Jianxiong Wu, Weijiang Zhan, Jun Hua, Qinling Ge, Yuexian Zhu, Huixian Yu, Min Zhao, Xiaoyan Zhan, Bingwei Zhu, Tongqi Xiang, Longxi Lu, Tieying Dai. 8942

Evaluating Encoder and Decoder Models for Extended Clinical Concept Recognition in Japanese Clinical Texts: Comparative Study With Weighted Soft Matching (e78681)
Yuya Tsukiji, Satoshi Kataoka, Masafumi Itokazu, Ryozo Nagai, Takeshi Imai. 8994

XSL•FO
RenderX

Human-AI Interaction in Kidney Transplant Decision Support Systems: Qualitative Study of Patient and Support Person Expectations (e83195)
Zeineb Sassi, Sascha Eickmann, Roland Roller, Bilgin Osmanodja, Jakob Spencker, Ömer Ömero lu, Aljoscha Burchardt, Michael Hahn, Peter Dabrock, Sebastian Möller, Klemens Budde, Anne Herrmann. 9016

Detecting Uncoded Self-Harm in Veterans' Electronic Health Records Using Positive and Unlabeled Learning: Retrospective Cohort Study (e89071)
Praveen Kumar, Alexandria Viszolay, Rajesh Upadhayaya, Fariha Moomtaheen, Donald Greer, Cristian Bologa, Kristan Schneider, Sharon Davis, Michael Matheny, David van der Goes, Gerardo Villarreal, Yiliang Zhu, Mauricio Tohen, Scott Malec, Jeremy Yang, Elliot Fielstein, Christophe Lambert. 9029

Addressing Data Quality Challenges in Lung Cancer Data Within the Observational Medical Outcomes Partnership Common Data Model: Observational Study (e90246)
Jens Declerck, Mieke Deschepper, Kirsten Colpaert, Dipak Kalra, Pascal Coorevits. 9046

Quality of Cancer-Related Clinical Coding in Primary Care in North Central London: Mixed Methods Quality Improvement Project (e73205)
Afsana Bhuiya, Graham Roberts, Katie Tucker, Stefanie Bonfield, Georgia Black. 9056

Multimodal Large Language Models for Cystoscopic Image Interpretation and Bladder Lesion Classification: Comparative Study (e87193)
Yung-Chi Shih, Cheng-Yang Wu, Shi-Wei Huang, Chung-You Tsai. 9077

Evaluation of Large Language Models for Structured Data Extraction From Interstitial Lung Disease Clinical Notes: Comparative Study (e90547)
Stephanie Chen, Manoj Maddali, Curtis Langlotz, Christian Bluethgen, Jonathan Chen, Rishi Raj. 9098

Digital and Navigational Health Literacy in Swiss Cancer Survivors Compared With the General Population: Cross-Sectional Questionnaire Study (e84228)
Julia Seinsche, Elena Alder, Tami Wehrmann, Jörg Haslbeck, Karin Ribi, Saskia De Gani. 9113

Validation of the Updated Digital Health Literacy Instrument and Development of a Short Form: Online Survey Study of the General Population (e86879)
Rosalie van der Vaart, Inge Tuitert, Job van 't Veer, Constance Drossaert. 9133

Relationship Between Online Health Information Acquisition and Shared Decision-Making Among Patients With Diabetes: Cross-Sectional Survey Study (e86137)
Yue Ming, Miriam Matteson, Yao Zhang. 9147

Association of Electronic Health Literacy With Self-Care and Health Outcomes Among Patients With Type 2 Diabetes Mellitus: Cross-Sectional Study (e77856)
Phoenix Mo, Alice Kong, Luyao Xie, Virginia Chan, Joseph Lau. 9161

The Relationship Between Electronic Health Literacy and Health-Related Quality of Life Among Chinese Older Adults: Cross-Sectional Study (e84700)
Yongqiang Wang, Baozhen Dai, Jiazhen Yao. 9174

Digital Engagement and Cognitive Function Among Older Adults in China: Cross-Sectional Questionnaire Study and Moderated Mediation Model Analysis (e83955)
Yongqi Du, Qing Niu, Gangrui Tan, Jianqian Chao, Shengxuan Jin, Leixia Wang. 9195

AI-Based Triage Decision Support: Multisite Economic Evaluation in the United States (e95213)
Scott Levin, Ben Steinhart, Rohit Sangal, Lesley Meng, K Frick, Cabrina McGinn, Jeremiah Hinson, Andrew Taylor. 9216

XSL•FO
RenderX

Cost-Effectiveness Analysis of Chatbot-Supported Remote Patient Monitoring for Anticoagulation Management: Health Economic Evaluation Within a Pilot Crossover Trial (e85430)
Ana Santos, Filipa Sampaio, Federico Guede-Fernández, Julian Perelman, Ana Londral. 9227

Remote Patient Monitoring in Louisiana Medicare Beneficiaries With Diabetes or Hypertension: Retrospective Cohort Study of 2016-2020 Claims Data (e80782)
Debra Winberg, Charles Stoecker, Jian Li, Lizheng Shi. 9245

EPOCA Tele-Monitoring System for Older Adults at High Risk of Hospitalization: Budget Impact and Cost-Effectiveness Analysis (e80302)
Henri Leleu, Damien Testa, Mireille Dutech, Etienne Minvielle, Elise Cabanes. 9259

Extended Grammar of Systematized Nomenclature of Medicine – Clinical Terms for Semantic Representation of Clinical Data: Methodological Study (e80314)
Christophe Gaudet-Blavignac, Julien Ehrsam, Monika Baumann, Adel Bensahla, Mirjam Mattei, Yuanyuan Zheng, Christian Lovis. 9272

Establishment and Optimization of a Patient-Reported Outcome–Based Electronic-Diary for Symptoms Evaluation in Patients With Gastroesophageal Reflux Disorder: Prospective Cohort Study (e83680)
Yun-Chun Chen, Yen-Po Wang, Jui-Hsuan Hung, Da-Wei Wang, Shang-Liang Wu, Li-Fen Chen, Yueh-Hsin Ping, Mei-Lien Pan, Ching-Liang Lu. 9287

Quality of Life Trajectories With Integration Into Electronic Health Records for High-Resolution Patient Outcomes: Algorithm Development and Validation Study (e79834)
Martin Dugas, Robin Fleige, Max Blumenstock, Stephan Feder, Tobias Dittrich, Niels Siegel, Celine Bergmann, Luis Wettach, Sandra Sauer, Pavlina Lenga, Sandro Krieg, Susanne Dugas-Breit, Lucy Kessler, Kosima Kosmalla, Fleur Fritz-Kebede, Martin Loos, Hans-Christoph Friederich, Thomas Pausch, Matthias Ganzinger. 9300

Artificial Intelligence for Identifying Patient-Reported Outcome and Experience Measures in Oncology: Retrospective Cross-Sectional Study Using ClinicalTrials.gov (e84533)
Jessica Soyer, Akram Hecini, Sylvain Juchet, Céline Desvignes-Gleizes, Maxime Thiebaut, Jean-Philippe Bertocchio. 9311

Cognitive Dissonance–Based Priming Intervention: Randomized Encouragement With in-the-Wild Phishing Simulation Attack in Health Care (e68051)
Prosper Yeng, Muhammad Fauzi, Arnstein Vestad, Bian Yang, Katrien De Moor, Christian Jacobsen, John-Bosco Diekuu, Meriem Bettayeb. . 9 3 2 6

Communicative Behaviors in an Internet-Based Intervention for Individuals With Autism: Mixed Methods Analysis (e76527)
Britta Westerberg, Karin Jacobson, Maria Unenge Hallerbäck, Susanne Bejerot, Fredrik Holländare. 9347

Quality Evaluation of Large Language Model–Assisted Generation of Initial Senior Physician Ward Round Records for Patients With Acute Poisoning: Cross-Sectional Study (e91222)
Junping Zhu, Wei Pan, Yonghong Wang, Kui Yan, Zhicheng Fang, Xianyi Yang. 9361

Hepatic Steatosis Severity Prediction in Nonobese Individuals: Machine Learning Model Development and Validation (e82529)
Yitong Zhu, Yongshuai Wang, Shenyu Zhang, Jian Yang, Feng Zhang, Yan Liu, Jun Shang, Yongliang Zhang, Jizhou Wang, Lianxin Liu. 9 3 7 2

Patient Cognitive Bias in Large Language Model–Supported Health Consultations: Simulation-Based Comparative Study (e85770)
Yi Zuo, Qifeng Wan, Shalong Wang. 9394

XSL•FO
RenderX

Benchmark Integrity and Reasoning-Trace Errors in Medical Question Answering With Large Language Models: Mixed Methods Study With Sparse Autoencoders (e90061)
Jialin Liu, Siru Liu, Adam Wright. 9462

An Evaluation of Pretrained Generative Models for Augmenting Small Health Data: Comparative Modeling Study (e88678)
Margerie Huet-Dastarac, Fida Dankar, Dan Liu, Samer El Kababji, Lisa Pilgram, Khaled El Emam. 9480

Alignment Between Cardiologists and AI-Driven Diagnostic Systems: Mixed Methods Study (e83541)
Mahdi Mahdavi, Sarah White, Sandeep Hothi, Chris Flood, Rosica Panayotova, Daniel Frings. 9497

Patients' Perspectives on the Implementation of AI in Radiological Diagnostics: Focus Group Study (e89178)
Cornelia Karger. 9510

Artificial Intelligence in Patient-Centered Care and Macro-, Meso-, and Micro-Level Determinants of Rehumanization and Dehumanization: Qualitative Interview Study (e82774)
Dora Horvath, Noemi Lorincz. 9522

Performance of Vision-Enabled Large Language Models in Image-Based Electrocardiogram Interpretation: Exploratory Evaluation (e86692)
Nibras Soubh, Eva Rasenack, Helge Haarmann, Felix Wiedmann, Markus Zabel, Constanze Schmidt, Rayan Suliman, Leonard Bergau. . . . 9 5 3 9

A Large Language Model–Powered Multiagent Framework Emulating Standardized Patients in Clinical Communication Skills Training: Development and Evaluation Study (e84747)
Yufei Qu, Xiaowei Xu, Yunzi Long, Yijie Wang, Jiao Li, Xudong Lu. 9556

Exploring Patient Perspectives on the Use of Artificial Intelligence to Inform Joint Decision-Making for Patients With Multiple Conditions in Primary Care in the United Kingdom: Qualitative Study (e87507)
Sarah Flanagan, Charlotte Spurway, Louise Jackson, Jenny Cooper, Francesca Crowe, Shamil Haroon, Tom Marshall, Leah Fitzsimmons, Eleanor Hathaway, Krishnarajah Nirantharakumar, Thomas Jackson, Sheila Greenfield. 9583

Time-Dynamic AI Models to Predict Quality of Life in Patients With Breast Cancer: Development and Validation Study Using the EORTC BALANCE Cohort (e81424)
Niclas Hubel, Thijs van der Heijden, Benjamin Murauer, Belle de Rooij, Kelly de Ligt, Helena Verkooijen, Sofie Gernaat, Meeke Hoedjes, Volker Arndt, Lonneke van de Poll-Franse, Bernhard Holzner, Jens Lehmann. 9595

Performance of AI Tools in Citing Retracted Literature : Content Analysis (e88766)
Sebastian Labenbacher, Maximilian Niederer, Sascha Hammer, Matthias Bader, Nikolaus Schreiber, Helmar Bornemann-Cimenti. 9613

Artificial Intelligence in Health Professions Education: Qualitative Study of Student Experiences (e82432)
Marina Guirguis, Salomon Fotsing, Joanne Fevry, Christine Landry, Diane Bouchard-Lamothe, Jennifer Lacroix, Alireza Jalali. 9623

Initial Insights Into an Institutional Secure Large Language Model for Magnetic Resonance Imaging Examination Requests: Retrospective Study (e82579)
James Hallinan, Naomi Leow, Yi Low, Aric Lee, Wilson Ong, Matthew Chan, Ganakirthana Devi, Stephanie He, Daniel Loh, Desmond Lim, Xi Low, Mei Lim, Clement Yong, Weizhong Sng, Ee Teo, Jiong Tan, Naresh Kumar, Andrew Makmur, Yonghan Ting. 9637

Views of People With Psychosis About Algorithm-Based Relapse Prediction and Data Sharing: Qualitative Study (e86753)
Emily Eisner, Hannah Ball, John Ainsworth, Matteo Cella, Richard Drake, Daniel Elton, Sophie Faulkner, Kathryn Greenwood, Andrew Gumley, Gillian Haddock, Kimberley Kendall, Alex Kenny, Tor-Ivar Krogsæter, Jane Lees, Shôn Lewis, Alie Phiri, Matthias Schwannauer, Rebecca Turner, Annabel Walsh, James Walters, Til Wykes, Uzma Zahid, Sandra Bucci. 9648

Context-Aware Sentence Classification of Radiology Reports Using Synthetic Data: Development and Validation Study (e86365)
Tomohiro Kikuchi, Yosuke Yamagishi, Kohei Yamamoto, Toshiaki Akashi, Harushi Mori, Hisaki Makimoto, Takahide Kohro. 9671

XSL•FO
RenderX

Error Detection in Emergency Radiology Reports Using a Large Language Model: Multistage Evaluation Study (e86841)
Hui Shen, Tianyang Wu, Fei Wang, Jin Fang, Yuange Li, Xiaoling Wu, Shuyi Liu, Liting Chen, Qiuping Ren, Xiao Meng, Jiatong Xu, Jie Sun, Yujie Zhao, Xin Liu, Liaoyuan Wang, Guipeng Mai, Jingjing You, Zhe Jin, Xuewei Wu, Wenle He, Xue Han, Shuixing Zhang, Dong Zeng, Bin Zhang. 9 6 8 5

Improving Retrieval Augmented Generation for Health Care by Fine-Tuning Clinical Embedding Models: Development and Evaluation Study (e82997)
Kamyar Arzideh, Henning Schäfer, Ahmad Idrissi-Yaghir, Cynthia Schmidt, Bahadir Eryilmaz, Mikel Bahn, Amin Turki, Olivia Pollok, Eva Hartmann, Philipp Winnekens, Katarzyna Borys, Johannes Haubold, Felix Nensa, René Hosch. 9704

Anticipating Moral and Economic Considerations, Opportunities, and Potential Frictions for AI in Medical Imaging: Multistakeholder Cocreation Study (e83407)
Martin Schilder, Alexandra Keyser, Susan van Hees, Alessandro Sbrizzi, Wouter Boon. 9743

Forecasting Waitlist Trajectories for Patients With Metabolic Dysfunction–Associated Steatohepatitis Cirrhosis: A Neural Network Competing Risk Analysis (e68247)
Gopika Punchhi, Yingji Sun, Eunice Tan, Naomi Hlaing, Chang Liu, Sumeet Asrani, Sirisha Rambhatla, Mamatha Bhat. 9762

End-to-End Platform for Electrocardiogram Analysis and Model Fine-Tuning: Development and Validation Study (e81116)
Lucas Bickmann, Lucas Plagwitz, Antonius Büscher, Lars Eckardt, Julian Varghese. 9776

Predicting the Intention to Use Generative Artificial Intelligence for Health Information: Comparative Survey Study (e75648)
Jörg Matthes, Anne Reinhardt, Selma Hodzic, Jaroslava Ka ková, Alice Binder, Ljubisa Bojic, Helle Maindal, Corina Paraschiv, Knud Ryom. . 9 7 9 4

Key Information Influencing Patient Decision-Making About AI in Health Care: Survey Experiment Study (e75615)
Xuan Zhu, Austin Stroud, Sarah Minteer, Dong Yoo, Jennifer Ridgeway, Maryam Mooghali, Jennifer Miller, Barbara Barry. 9808

Intervention in Health Misinformation Using Large Language Models for Automated Detection, Thematic Analysis, and Inoculation: Case Study on COVID-19 (e75500)
Samira Malek, Christopher Griffin, Robert Fraleigh, Robert Lennon, Vishal Monga, Lijiang Shen. 9834

Developing an AI-Assisted Tool That Identifies Patients With Multimorbidity and Complex Polypharmacy to Improve the Process of Medication Reviews: Qualitative Interview and Focus Group Study (e74304)
Aseel Abuzour, Samantha Wilson, Alan Woodall, Frances Mair, Asra Aslam, Andrew Clegg, Eduard Shantsila, Mark Gabbay, Michael Abaho, Danushka Bollegala, Harriet Cant, Alan Griffiths, Layik Hama, Gary Leeming, Emma Lo, Simon Maskell, Maurice O'Connell, Olusegun Popoola, Sam Relton, Roy Ruddle, Pieta Schofield, Matthew Sperrin, Tjeerd Van Staa, Iain Buchan, Lauren Walker. 9856

Evaluating and Validating Large Language Models for Health Education on Developmental Dysplasia of the Hip: 2-Phase Study With Expert Ratings and a Pilot Randomized Controlled Trial (e73326)
Hui Ouyang, Gan Lin, Yiyuan Li, Zhixin Yao, Yating Li, Han Yan, Fang Qin, Jinghui Yao, Yun Chen. 9871

Data Poisoning Vulnerabilities Across Health Care Artificial Intelligence Architectures: Analytical Security Framework and Defense Strategies (e87969)
Farhad Abtahi, Fernando Seoane, Ivan Pau, Mario Vega-Barbas. 9893

Implementing an Artificial Intelligence Decision Support System in Radiology: Prospective Qualitative Evaluation Study Using the Nonadoption Abandonment Scale-Up, Spread, and Sustainability (NASSS) Framework (e80342)
Sundresan Naicker, Paul Schmidt, Bruce Shar, Amina Tariq, Ashleigh Earnshaw, Steven McPhail. 9913

XSL•FO
RenderX

Integrated Prediction System for Individualized Ovarian Stimulation and Ovarian Hyperstimulation Syndrome Prevention: Algorithm Development and Validation (e78245)
Jingjing Chen, Jianjuan Zhao, Huiyu Qiu, Yanhui Liu, Yunqi Zhang, Qicheng Sun, Yan Yi, Hongying Tang, Jing Zhao, Bin Xu, Qiong Zhang, Ge Yang, Hui Li, Junjie Liu, Zhongzhou Yang, Shaolin Liang, Yanping Li, Jing Fu. 9930

Patterns of AI Use in Clinical Work by Hospitalists: Survey Study (e85973)
Prabhava Bagla, Jasmah Hanna, Bhargav Marthambadi, Stacey Watkins. 9945

Psychosis Risk and Generative Artificial Intelligence Use Frequency, Motivations, and Delusion-Like Experiences: Cross-Sectional Survey Study (e85038)
Benjamin Buck, Anne Maheux. 9958

Scalable and Robust Artificial Intelligence for Spine Alignment Assessment: Multicenter Study Enabled by Real-Time Data Transformation (e78396)
Guilin Chen, Nan Meng, Yipeng Zhuang, Zhe Chen, Zhen Bian, Zhaoyang Gong, Jiawei Shi, Tao Huang, Xihe Kuang, Pengyu Lu, Cong Nie, Qifeng Yu, Zefu Chen, Hui Jiang, Zhongmin Zhang, Chaojun Zheng, Yu Liang, Nan Wu, Jason Cheung, Jianguo Zhang, Teng Zhang. 9969

Measuring the Impact of AI on Report-Drafting Efficiency in Chest Computed Tomography Interpretation: Retrospective Analysis (e77967)
Weiqi Liu, You Wu, Wei Yu, Mark Bittle, Zhuozhao Zheng, Hadi Kharrazi. 9983

Transfer Learning and Machine Learning for Training Five-Year Survival Prognostic Models in Early Breast Cancer: Development and Validation Study (e88665)
Lisa Pilgram, Kai Yang, Ana-Alicia Beltran-Bless, Gregory Pond, Lisa Vandermeer, John Hilton, Marie-France Savard, Andreanne LeBlanc, Lois Shepherd, Bingshu Chen, John Bartlett, Karen Taylor, Jane Bayani, Sarah Barker, Melanie Spears, Cornelis van der Velde, Elma Meershoek-Klein Kranenbarg, Luc Dirix, Elizabeth Mallon, Annette Hasenburg, Christos Markopoulos, Lamin Juwara, Fida Dankar, Mark Clemons, Khaled El Emam. 9996

Plain Language Summarization of Environmental Health Research Using Generative AI: Community-Engaged Qualitative Study (e87118)
Katherine Wade, Lauren Anderson, Evelyn Medley, Ted Smith. 10019

Barriers, Facilitators, and Intention to Use AI for Breast Cancer Diagnosis: Mixed Methods Study Among Austrian Physicians With and Without AI Experience (e80274)
Marlene Kritz, Erol Holawatsch, Doris Behrens, Walter Hyll. 10030

Multitask Artificial Intelligence–Based Electrocardiogram Tool for Preoperative Cardiac Testing in Noncardiac Surgery: Retrospective Cohort Study of Health Care Utilization and Costs (e90099)
Hong-Mi Choi, Yerin Kim, Joonghee Kim, Jiesuck Park, In-Chang Hwang, Yun Choi, Ji Lee, Yeonyee Yoon, Il-Young Oh, Goo-Yeong Cho, In-Ae Song, Youngjin Cho. 10045

Comparative Evaluation of ChatGPT, Google Translate, and UD Talk for Chinese-to-Japanese Translation in Cardiology and Pulmonology Outpatient Consultations: Prospective Observational Study (e93082)
Ke-Yun Chao, Chia-Hui Huang, Wei-Lun Liu, Chao-Yu Chen, Ai Lim, Yi-Zhen Chu, Atsuko Sasaoka, Chung-Yu Lin. 10064

A Chatbot to Meet Parents' Information Needs for Sickle Cell Trait Newborn Screening Results: Multiple Methods Formative Study (e86022)
Anne Madeo, Courtney Gauchel, Emerson Borsato, Amy Gaviglio, Kimberly Kaphingst, Guilherme Del Fiol, Karen Eilbeck. 10076

Enhancing Detection of Message Intents in a Mobile Health Smoking-Cessation Intervention Using Large Language Model Fine-Tuning, Data Downsampling, and Error Correction: Algorithm Development and Validation (e83437)
Shagoto Rahman, Cornelia Pechmann, Ian Harris. 10089

XSL•FO
RenderX

Evaluation of an Artificial Intelligence Conversational Chatbot to Enhance HIV Preexposure Prophylaxis Uptake: Development and Usability Internal Testing (e79671)
Jun Tao, Ellie Pavlick, Amaris Grondin, Josue Bustamante, Harrison Martin, Hannah Parent, Natalie Fenn, Alexi Almonte, Amanda Maguire-Wilkerson, Mofan Gu, Jack Rusley, Bryce Perler, Tyler Wray, Amy Nunn, Philip Chan. 10117

Developing a Quality Evaluation Index System for Health Conversational Artificial Intelligence: Mixed Methods Study (e83188)
Weizhen Liao, Meng Li, Chengyu Ma, Youli Han, Dan Wang, Haopeng Liu, Yi Wang, Zijie Feng, Huichao Wang, Yiru Guan. 10133

Developing a Service Quality Index System for AI Health Care Chatbots: Mixed Methods Study (e83051)
Yu Gu, Xinyi Wang. 10154

Comparison of Emotional Content in Text Responses From Physicians and AI Chatbots to Patient Health Queries: Cross-Sectional Study (e85516)
Daniel Burns, Channing Bice, Paul Johnson, Nicholas Chia, Timothy Robinson. 10166

Young Adults' Interactions With Food and Nutrition Content on Social Media and Implications for Intervention Design: Semistructured Interview Study (e89344)
Hao Tang, Amy Ahern, Marie Spreckley, Andrea Smith. 10179

Human and Robot Assistance for Cognitive Load in Younger and Older Adults: Multimodal Within-Subject Experimental Study (e94738)
Simone Varrasi, Roberto Vagnetti, Nicola Camp, John Hough, Alessandro Di Nuovo, Sabrina Castellano, Daniele Magistro. 10194

Evaluating Nursing Work Systems and Identifying Barriers for Robotic Technology Integration: Observational Study (e89409)
Gina Georgadarellis, Ellen Benjamin, Shannon Roberts, Cidalia Vital, Frank Sup IV. 10211

Assistive Robotics for Healthy Aging: A Foundational Phenomenological Co-Design Exercise (e77179)
Stephen Potter, Mark Hawley, Angela Higgins, Farshid Amirabdollahian, Mauro Dragone, Alessandro Di Nuovo, Praminda Caleb-Solly. 10226

Adoption of Internet of Things in Health Care: Weighted and Meta-Analytical Review of Theoretical Frameworks and Predictors (e64091)
Inês Veiga, Tiago Oliveira, Mijail Naranjo-Zolotov, Ricardo Martins, Stylianos Karatzas. 10244

A Review of Medical Data Sharing Initiatives With a Focus on the Use of Blockchain Technologies: Qualitative Comparative Analysis (e63575)
James Cunningham, Nigel Davies, Sarah Devaney, Mike Harding, Søren Holm, Victoria Neumann, Clare Smith, John Ainsworth. 10275

Resistance to Blockchain Adoption in Health Care Organizations: Evidence From a Cross-Sectional Study (e77933)
Wilfried Bazomanza Nzabandora, Sehl Mellouli, Rohit Nishant. 10300

Patient Perceptions of Blockchain-Based Health Information Exchange: User-Centered Design Study (e78849)
Richard Guse, Shanshan Hu, Scott Thiebes, Christina Erler, Catherine Caridia, Wilhelm Stork, Martin Gersch, Ali Sunyaev. 10319

Digital Discourse, Secondary Victimization, and Psychological Harm: Mixed-Methods Analysis of System Justification in the #MeToo Movement (e75533)
Harsh Parekh, Shriya Thakkar, Paras Bhatt, Patricia Akello. 10338

XSL•FO
RenderX

Analysis of Multilevel Factors Mobilizing the Spectrum of Interorganizational Knowledge Sharing for Facilitating Digital Transformation at Scale: Qualitative Study (e83345)
Hajar Mozaffar, Robin Williams, Kathrin Cresswell. 10371

Virtual Multidisciplinary Gastrointestinal Care for Adults With Gastrointestinal Needs: Retrospective Cohort Study (e89061)
Grace Wang, William Chey, Sanskriti Varma, Sameer Berry. 10391

Use of the Online Portal “Embryotox” in Routine Health Care: Mixed Methods Study (e81286)
Marlies Onken, Evelin Beck, Martina Breuning, Christine Holmberg, Franziska Metke, Anne Müller, Christof Schaefer, Katarina Dathe. 10405

Usage and Perceptions of Electronic Patient Records Experienced by Users and Nonusers in the Canton of Vaud, Switzerland: Mixed Methods Study (e83702)
François Bastardot, Catherine LaRocco, Mathilde Robert, Veronika Schoeb. 10421

Effects of a Soft Robotic Exoskeleton for Gait Training on Clinical Outcomes in Patients With Parkinson Disease: Randomized Controlled Pilot Study (e82629)
Shiying Bu, Xiaoqing Luo, Shuangfang Li, Yuqin Shi, Lingjing Jin, Qing Zhao. 10437

Facilitators and Barriers to Implementing a Remote Monitoring Model of Care for Stable Patients With Axial Spondyloarthritis Using the Consolidated Framework for Implementation Research: Qualitative Study (e82480)
Yu Kwan, Xin Chew, Humairah Zainal, Zhonghui Xiong, Melissa Hock, Ka-Hao Zee, Ying-Ying Leung, Truls Østbye, Lian Low, Hendra Goh, Charmaine Sum, Sungwon Yoon, Charmaine Wang, Warren Fong, Julian Thumboo. 10460

A Virtual Integrated General Practitioner–Pediatrician Model of Care Implemented in Metropolitan and Rural Primary Care Settings: Qualitative Analysis of Clinician Perspectives on the SUSTAIN Model of Care (e86707)
Susan Bullock, Raghu Lingam, Karen Wheeler, Tammy Meyers Morris, Corin Miller, Louisa Adams, Ken Peacock, Annemarie Christie, Michael Hodgins, Harriet Hiscock, Lena Sanci, Natalie Taylor, Carmen Crespo-Gonzalez. 10476

Digital Quality Monitoring for Type 2 Diabetes in Swiss Primary Care: Qualitative Interview Study (e82960)
Odile-Florence Giger, Lola Ackermann, Marinja Principe, Simon Meier, Elgar Fleisch, Susanna Gallani, Michael Brändle, Tobias Kowatsch, Mia Jovanova. 10497

Digital Literacy and Interpersonal Trust as Predictors of Willingness to Share Patient-Generated Health Data Among Korean Internet Users: Cross-Sectional Study Using Privacy Calculus and Communication Privacy Management Theories (e75448)
Dongsu Lee, Wonseuk Jang. 10514

Quality Challenges in Municipal Telecare Call Center Services: Qualitative Evaluation Using the Anchored, Realistic, Cocreated, Human, Integrated, and Evaluated (ARCHIE) Framework (e76054)
Linda Grøndal-Eeles, Janne Dugstad, Hilde Eide, Etty Nilsen. 10535

Exploring the Dynamics of Actors, Structural Factors, and Bricolage in the Implementation and Sustainability of eHealth Solutions: Qualitative Multiple-Case Study (e79999)
Susanne Eriksen, Christine Øye, Anne Dahler. 10549

“I Want to Spend My Time Living”—Experiences With a Digital Outpatient Service With a Mobile App for Tailored Care Among Adults With Long-Term Health Service Needs: Qualitative Study Using Thematic Analysis (e79155)
Heidi Holmen, Erik Fosse. 10564

Feasibility and Acceptability of AI-Powered Tools for Early Autism Screening in Egypt: Semistructured Focus Group Study (e82564)
Pratheepan Yogarajah, Ammal Metwally, Priyanka Chaurasia, Ghada Elshaarawy, Shereen El Khateeb, Engy Ashaat, Amal Elsaeid, Nahed Elghareeb, Amira ElRifay. 10576

XSL•FO
RenderX

Facility-Level Associations Between Use of a Digital Health Platform for Voluntary Counseling and Testing and HIV Testing Outcomes in the Urban Primary Health Care Centers of Guangzhou, China: Cross-Sectional Study (e83662)
Ye Chen, Yu-Fei Wang, Yu-Zhou Gu, Xiao-Ru Fan, Yong-Heng Lu, Ju-Shuang Li, Jun Wang, Zhi-Ye Lin, Chun-Li Zhang, Meng-Die Li, Ning-Jun Ren, Zhi-Zhong Luo, Rui-Fei Zuo, Fu-Chuan Xiong, Yu-Tong Kang, Xiu-Rong Lin, Chun Hao. 10593

Usage and Exposure to Content of the NHS Healthy Living Program for People With Type 2 Diabetes: Retrospective Observational Cohort Study (e89690)
Antonia Marsden, Rhiannon Hawkes, Jack Benton, David French, Sarah Cotterill. 10607

Measuring the Quality of Datasets: Development of the IDEFIM Indicator Set for Empirical Health Research (e90482)
Sonja Harkener, Oliver Bott, Christian Draeger, Tobias Hartz, Ekkehart Jenetzky, Matthias Löbe, Stefanie March, Chris Schubert, Jürgen Stausberg. 10622

Online Social Interaction, Neighborhood Perception, and the Mediating Role of Social Capital in Charitable Giving for Seriously Ill Patients: Cross-Sectional Study (e85573)
Jie Chen, Junzhang Hu, Yang Yang, Fangjuan Du, Xuelian Li. 10637

What Patients With Asthma Share When No One Listens: Multimethod Observational Study of Patient Narratives on Reddit (e77027)
Elena Curto-Sánchez, Gabriela Salazar-Palacios, Ana Martín-Varillas, Estela Prieto-Maíllo, Jacinto Ramos-González, Ignacio Dávila-González, Domingo Palacios-Ceña, Juan Cuenca-Zaldivar. 10658

Moderators' Role in Fostering Health Practitioner Engagement Within an Online Active Learning Program (The Community Fracture Capture Learning Hub): Qualitative Relational Content Thematic Analysis (e83764)
Ahmed Fathalla, Shanton Chang, Ralph Audehm, Cherie Chiang, Alexandra Gorelik, Christopher Yates, Steve Snow, Rahul Barmanray, Sarah Price, Lucy Collins, John Wark. 10675

Quantifying Innovation in Stroke: Large Language Model Bibliometric Analysis (e70754)
Adam Marcus, Georgina Lockwood-Taylor, Daniel Rueckert, Paul Bentley. 10692

Automated Prediction of Glasgow Coma Scale Scores From Unstructured Electronic Health Records Using Natural Language Processing: Development and Validation Study (e81245)
Marta Fernandes, Niels Turley, Haoqi Sun, Shibani Mukerji, Lidia Moura, M Westover, Sahar Zafar. 10706

Electrocardiogram-Based Mental Stress Detection Amid Everyday Activities Using Machine Learning: Model Development and Validation Study (e80450)
Buelent Uendes, Alex Antonides, Sjors van de Ven, Denise van der Mee, Eco de Geus, Mark Hoogendoorn. 10721

Examining the Association Between Internet Use and Perceived Stress in Adults: Longitudinal Observational Study Combining Web Tracking Data With Questionnaires (e78775)
Mohammad Belal, Nguyen Luong, Talayeh Aledavood, Juhi Kulshrestha. 10747

The Impact of Evaluation Strategy on Sepsis Prediction Model Performance Metrics in Intensive Care Data: Retrospective Cohort Study (e72083)
Dang-Khoa Do, Patrick Rockenschaub, Sebastian Boie, Oliver Kumpf, Hans-Dieter Volk, Felix Balzer, Falk von Dincklage, Gregor Lichtner. . .
1 0 7 7 3

Sleep and Activity Patterns in Depression From Wearable Data: Unsupervised Clustering Study (e86900)
Carolin Oetzmann, Yuezhou Zhang, Nicholas Cummins, Ewan Carr, Faith Matcham, Sara Siddi, Femke Lamers, Daniel Leightley, Katie White, Amos Folarin, Peter Annas, Josep Haro, Brenda Penninx, Srinivasan Vairavan, Til Wykes, Richard Dobson, Vaibhav Narayan, Matthew Hotopf.
1 0 7 9 1

XSL•FO
RenderX

Digital Rehabilitation Following Ultrasound-Guided Injection for Chronic Rotator Cuff Injury: Randomized Controlled Trial (e79494)
Zifu He, Yuemin Zhang, Liyan Shi, Gaineng Yuan, Xuefeng Liu, Yue Cai, Jue Zhao, Jing Xu. 10812

Physical Activity Recommendations Tailored by a Predictive Model for Adults With High Blood Pressure: Observational Study (e78492)
Yuhui Yang, Manqing Chen, Weiwei Hu, Yifan Fu, Xingyan Li, Zhenli Liao, Hongman Feng, Yaling Zhao, Leilei Pei, Baibing Mi, Fangyao Chen. 10812

Comparative Analysis of Prescriptions and Pharmacy Services in Internet-Based Psychiatric Hospital During and After the COVID-19 Pandemic: Retrospective Cross-Sectional Observational Study (e74059)
Guowei Deng, Hui Xia, De-wei Shang, Yuguan Wen, Jinqing Hu, Yaqian Tan. 10893

Skill-Based Virtual Reality Therapy in High-Impact Chronic Pain: 2-Year Follow-Up Results From a Secondary Analysis of a Randomized Controlled Trial (e90688)
Todd Maddox, Josh Sackman, Robert Bonakdar, Roselani Maddox, Beth Darnall. 10914

Time Savings Through an AI Speech Assistant for Nursing Documentation: Pre-Post Time-Motion Study in German Long-Term Care (e86078)
Katja Schwabe, Drin Ferizaj, Susann Neumann, Sandra Strube-Lahmann, Nils Lahmann. 10925

Medical Consultation and HIV Testing After AI-Based Symptom Check: Retrospective Cohort Study (e90257)
Nao Taguchi, Kunihiro Hirahara, Keisuke Harada, Keisuke Orimo, KuanYeh Lee, Kota Iwahashi, Akifumi Imamura. 10938

Computerized Self-Reported Medical History Taking to Support Early Rule Out of Major Adverse Cardiac Events in Patients With Acute Chest Pain: Post Hoc Analysis of the CLEOS-CPDS Prospective Cohort Study (e76087)
Helge Brandberg, Carl Sundberg, Jonas Spaak, Sabine Koch, Thomas Kahan. 10953

Acceptability, Feasibility, and Perceived Effectiveness of Video-Based Patient Records for Supporting Care Delivery to Older Adults With Frailty: Nonrandomized Mixed Methods Pilot Study (e77318)
Phoebe Averill, Rachael Lear, Ricky Odedra, Susannah Long, Alex Taylor, Pi-Jung Charville, Jessica Fernandes, Uzoamaka Ekeogu, Jessica Leombruno, Sophia Ellis, Erik Mayer. 11019

Associations and Pathways Between Online Health Information–Seeking Behavior and Patient Adherence: Cross-Sectional Study (e91115)
Zhenyu Xu, Jingjing Jia, Chaofan Wu. 11039

Adaptive Fast-Slow Large Language Model Framework for Multidimensional Classification of Prenatal Ultrasound Reports: Comparative Study (e91399)
Wei Zhong, Huihui Yan, Yifan Liu, Yan Liu, Kai Yang, Huimin Gao, Zhengyang Yao, Wenjing Hao, Yousheng Yan, Chenghong Yin. 11050

Confidence Measurement Metrics in Multimodal Large Language Models for Ultrasound-Based Radiology Cases: Comparative Evaluation Study of Self-Reported, Consistency-Based, and Hybrid Methods (e86498)
Taewon Han, Jaeseung Shin, Jeong Lee, Kyowon Gu. 11065

Performance Evaluation of GPT-5, Grok 4, and DeepSeek R1 in Interpreting Complete Blood Count Reports for Hematologic Diseases: Retrospective Comparative Study (e87802)
Xianfei Ye, Xinglun Qi, Lina Fan, Qian Yu, Suming Zhou, Chunyun Ren, Dagan Yang. 11079

Assessing Eligibility for Anticancer Drug Health Insurance Reimbursement Using Large Language Models: Benchmark Development and Comparative Study (e95877)
Junhyuk Seo, Taerim Kim, Ju-Hyun Kim. 11093

Trusting Generative AI for Health Advice: Preregistered Survey Experiment (e97882)
Asheley Landrum, Nitin Verma, Amanda Kehrberg. 11106

XSL•FO
RenderX

Large Language Model Summarization of Physician-to-Physician Calls for Interhospital Transfer of Patients With ST-Elevation Myocardial Infarction: Observational Study (e88834)
Jesse Wrenn, Madelaine Behrens, Mary Hershey, Marc Maldaver, John Mitchell, Trevor Thompson, Austin Triana, Zain Virk, Yasemin Akdas, Michael Cauley, Michael Ward, Ken Monahan. 11121

Disclaimers and Referral Patterns for Medical Advice Across Urgency Levels: Large Language Model Evaluation Study (e84668)
Florian Reis, Louis Agha-Mir-Salim, Richard Hickstein, Moritz Reis, Sophie Piper, Felix Balzer, Sebastian Boie. 11130

Effectiveness of AI-Assisted Patient Health Education Using Voice Cloning and ChatGPT: Prospective Randomized Controlled Trial (e81387)
Yan Sun, Shangqing Xu, Hongying Jin, Xiaoyan Han, Kangqi Jin, Yimei Zhang, Xiaoli Ma, Huaping Wei, Minjie Ma. 11143

Psychotherapists' Trust, Distrust, and Generative AI Practices in Psychotherapy: Qualitative Study (e88932)
Jinyan Kuang, Amber Pope, Yixuan Zhang. 11157

The Alberta Quality Assessment Tool: Risk of Bias (AQAT:RoB) for the Evaluation of Medical Large Language Model Question-Answer Studies: Development and Pilot Validation (e87057)
Carrie Ye, Joseph Mitchell, Daniel Baumgart, Zechen Ma, Angela Fung, Daniela Orellana, Juel Chowdhury, Abdullah Abass, Steven Katz, Jacob Jaremko, Pierre Boulanger, Claire Barber, Gillian Lemermeyer, Hosna Jabbari, Lili Mou, Maryam Mirzaei, Mary Githumbi, Puneeta Tandon, Randy Goebel, Rhys Clark, Whitney Hung, Marjan Abbasi, Farhad Maleki, Scott Klarenbach, Mohamed Abdalla. 11173

Benchmarking Large Language Models and Prompt Engineering Strategies in Microsatellite Instability Cancers: Evaluation Study (e88614)
Yuxin Zhang, Jie Song, Cheng Bi, Xin Zheng, Zhichuan Xu, Dan Cao, Bairong Shen. 11187

Reliability of Large Language Model Generated Clinical Reasoning in Assisted Reproductive Technology: Blinded Comparative Evaluation Study (e85206)
Dou Liu, Ying Long, Sophia Zuoqiu, Di Liu, Kang Li, Yiting Lin, Hanyi Liu, Rong Yin, Tian Tang. 11204

Large Language Model–Generated Patient Instructions for Prescriptions in Primary Health Care: Preclinical Algorithm Validation (e84444)
Zilma Silveira Nogueira Reis, Elisa Tuler Albergaria, Adriana Silvina Pagano, Eura Martins Lage, Flávia Ribeiro de Oliveira, Cristiane dos Santos Dias, Juliana Almeida Oliveira, Gláucia Miranda Varella Pereira, Isaias Jose Ramos de Oliveira, Érico Franco Mineiro, Igor Carvalho Lima Oliveira, Davi dos Reis de Jesus, Antônio Pereira de Souza Júnior, Igor de Carvalho Gomes, Rodrigo André Cuevas Gaete, Ricardo Cruz-Correia, Leonardo Rocha. 11219

Using a Large Language Model to Support Thematic Analysis of Patient Experiences in Chronic Illness Management: Comparative Qualitative Study (e88677)
Sara Kivity, Yechiel Barilan, Reut Noham, Mor Saban. 11237

Health Information Adoption Among Patients With Chronic Disease in China: Qualitative Interview Study of Patient–Platform Coshaping (e85229)
Lin Huang, Guimin Duan, Jiani Tang, Peipei He. 11275

Channel Allocation and Equity in Preventive Campaigns for Older Adults: Agent-Based Modeling Study (e88429)
Jihye Lee, Juyoung Park, Yuna Kim, Duk-Jo Kong. 11480

Reporting Horizon Scanning Studies: Prototype Development Study (e88253)
Sonia Garcia Gonzalez-Moral, Megan Fairweather, Sarah Khan, Ross Fairbairn, Gill Norman. 11498

Forecasting the Impacts of Artificial Intelligence Assistance in Virtual Consultations for Chronic Obstructive Pulmonary Disease: Exploratory Futures Wheel Study (e90208)
Pranavsingh Dhunnoo, Bertalan Meskó, Karen McGuigan, Vicky O'Rourke, Michael McCann. 11510

XSL•FO
RenderX

Sequencing AI Automation and Data Interoperability in Oncology Using a Scenario-Planning Framework Coupled With Discrete-Event Simulation: Proof-of-Concept Study (e92642)
Peter May, Sabine Brookman-May, Edward Garrahy, Johannes von Büren. 11524

Horizon Scanning for Medical Technologies: Methodological Framework Development Study of the Medical Innovation Scanning Techniques (MIST) Framework (e93166)
Sonia Garcia Gonzalez-Moral, Andrew Mkwashi, Fiona Beyer, Dawn Craig. 11534

Mapping Practice-Based Signals of Generative AI in Psychiatric Care: Qualitative Study of Korean Psychiatrists' Experiences, Interpretations, and Implementation Priorities (e96556)
Myungsung Kim, Yoosuk An, Min Jeon, Yunji Lee, Orane Lahcine, Hyorim Kim, Seonmi Lee, Sangil Lee, Jong-Chul Yang, Sang-Won Jeon, Dooyoung Jung, Chul-Hyun Cho. 11566

Framing the Convergence of One Health and Digital Health in the Global South With a Gender-Sensitive Foresight Perspective: Delphi Study Using Latent Semantic Analysis (e78702)
Jude Kong, Nicola Bragazzi. 11582

Reviews

Radiomics-Based AI for the Diagnosis and Prognosis of Vessels Encapsulating Tumor Clusters in Hepatocellular Carcinoma: Systematic Review and Meta-Analysis (e90931)
Xuefeng Hua, Rongdang Fu, Ziwei Yin. 90

Therapeutic Interaction Features of AI Chatbots in Depression Interventions: Systematic Review and Meta-Analysis (e88697)
Ting Huang, Shuangyu Li, Yanzhong Wang, Wei Liu. 115

Care Robots as Emerging Health Technologies: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e95232)
Gaeun Kim, Jinmi Jeon. 141

Digital Interventions Targeting Parents to Improve Early Childhood Movement, Nutrition, and Sleep Behaviors: Systematic Review (e85525)
Johanna Sandborg, Brittany Reese, Sarah Marshall, Kylie Hesketh, Rachel Laws, Katherine Downing. 164

Applications, Challenges, and Future Directions of Large Language Models in Health Care Communication: Scoping Review (e84726)
Jing Chang, Ruotong Peng, Xi Chen, Yishu Zhu, Ruting Miao, Zeng Cao, Hui Feng. 194

Physical Activity Interventions Using Digital Health Interventions for Cancer-Related Fatigue in People With a History of Cancer: Scoping Review (e83727)
Yeeun Kim, Ka Bae, Ji Sung, Yun Ko, Sue Kim. 217

Vision-Based Artificial Intelligence Technologies for Epilepsy Monitoring: Scoping Review and Taxonomy Development Study (e83895)
Mirijana Irnich, Jonas Hammer, Aleksandra Flok, Frank Teuteberg. 233

Accessibility of Digital Financial Applications for People With Visual Impairment: Scoping Review (e82315)
Louise Puli, Lars Kooijman, Tanjila Kanij, Charmine Hartel, Abu Shahriar, Kristian Rotaru. 261

XSL•FO
RenderX

Implementing Digital Respiratory Technologies for People With Respiratory Conditions: Scoping Review (e88325)
Io Hui, Kathleena Condon, Shailesh Kolekar, Nicola Roberts, Katherina Sreter, Sami Simons, Carlos Figueiredo, Zoe McKeough, Hani Salim, Aleksandra Gawlik-Lipinski, Apolline Gonsard, Ay e Aral, Anna Vanoverschelde, Matthew Armstrong, Dario Kohlbrenner, Cátia Paixão, Patrick Stafler, Efthymia Papadopoulou, Adrian Rabe, Milan Mohammad, Izolde Bouloukaki, Shirley Quach, Georgios Kaltsakas, Kate Loveys, Tonje Reier-Nilsen, Anthony Sunjaya, Paul Robinson, Michaela Senek, Amy Chan, Hilary Pinnock. 279

Embodied Intelligence Applications in Health Care Populations: Scoping Review (e83871)
Wan Shu, Ying Zhou, Erhong Sun, Lei Deng, Xuchun Ye. 303

Applications of DeepSeek in Medicine: Bibliometric Analysis and Scoping Review (e93354)
Haoran Zhang, Dawei Wang, Yanliang Xu, Shuming Han, Guangxin Wang. 323

Performance of Deep Learning in Classifying Age-Related Macular Degeneration From Images: Systematic Review and Meta-Analysis (e97174)
Yu Zhu, Yue Niu, Shangye Sun, Wei Liu, Ying Dou, Yu Guo. 345

Effects of Multicomponent Digital Health Interventions on Multidimensional Physical Activity in Older Adults: Systematic Review, Meta-Analysis, and Meta-Regression of Randomized Controlled Trials (e91338)
Jiayi Yao, Haozhe Wang, Shiguan Jia, Hao Chen, Junhao He, Juncheng Long, Shengxian Chen. 373

Effectiveness of eHealth Interventions in Alleviating Burden on Informal Caregivers of People With Dementia: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e78568)
Min Liang, Jinxia Rong, Crina Grosan, Xiyi Wang. 401

Digital and Blended Lifestyle Interventions for Preschool-Aged Children and Families With a Low Socioeconomic Position and the General Population: Scoping Review (e86596)
Lea Hohendorf, Tessa Dekkers, Hanneke Kip, Laurence Alpay, Saskia Kelders. 425

Artificial Intelligence Governance in Health Systems: Systematic Review of Frameworks and Integrative Model Proposal (e87448)
Hassane Alami, Renata Pozelli Sabio, Elsury Pérez, Marie-Pierre Gagnon, Lyse Langlois, Jean-Louis Denis, Kathy Malas, Lysanne Rivard, Mathilde Savoldelli, Mohamed Ag Ahmed, Jean-Paul Fortin. 446

Measuring eHealth Literacy in the European Economic Area, Switzerland, and the United Kingdom: Scoping Review (e87461)
Jonah Weist, Julia Nitsche, Jan Ehlers, Theresa Busse. 469

Exploring Technological Solutions for Interoperability Between Patient Electronic Medical Records and Clinical Registries: Scoping Review (e82380)
Erika Haynes, James Brannigan, Jessica Suna, Reid Malseed, Alana Delaforce, Rachel Mulvenney-Fenner, Katherine Alog-Daroya, Karin Plummer, Craig McBride, Roy Kimble, Bronwyn Griffin. 497

Virtual Reality Interventions for Stress Reduction in the General Population: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e78212)
Hannah Strauch, Isabel Schuil, Stefan Simm, Jens Grubert, Snehanjali Kalamkar, Karin Meissner. 514

Efficacy, User Engagement, and Acceptability of Cognitive Behavioral Therapy–Oriented Psychological Chatbots for Adults With Depressive and/or Anxiety Symptoms: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e82677)
Bingyan Gong, Nisha Yao, Hangxin Xie, Chuncheng Huang, Tomoko Kishimoto, Howard Berenbaum, Wenting Mu. 548

XSL•FO
RenderX

Barriers and Facilitators to Patient Acceptance of Artificial Intelligence in Health Care: Systematic Review (e80581)
Huiqin Shi, Jingying Huang, Jin Yang, Mengbo Han. 570

Efficacy of Technology-Based Interventions on the Reduction of Loneliness: Systematic Review and Meta-Analysis (e80059)
Zdenek Meier, Marie Buchtova, Jan Sandora, Lukas Novak, Jakub Helvich, Ondrej Buchta, Jana Furstova, Klara Malinakova, Peter Tavel. . .
5 9 7

Effectiveness of Digital Health Interventions in Older Adults With Frailty and Sarcopenia: Systematic Review and Meta Analysis of Randomized Controlled Trials (e88374)
Ting Dai, Changsheng Guo, Lingli Gao, Jinyu Huang, Yan Chen, Yujie Zhang, Jing Gao, Xiaodong Feng. 615

AI-Based Automation for Medication Reconciliation: Scoping Review (e86760)
Juan Tabja Bortesi, Maria Becerra, Jonathan Ranisau, Bonnie Wen, Praveen Nadesan, P Devereaux, Michael McGillion, Jeremy Petch. 640

Machine Learning and Deep Learning Models for Predicting Future Falls in Community-Dwelling Older Adults: Systematic Review and Meta-Analysis of Longitudinal Evidence (e84844)
Ying Gao, Doudou Xu, Xinru Li, Jue Wang, Linbin Wang, Beiwen Wu, Haifeng Zhao, Xian Qiu, Weiyi Zhu. 659

Concerns of Using Large Language Models in Health Care Research and Practice: Umbrella Review (e87804)
Feyza Yarar, Pauline Addis, Megan Fairweather, Dawn Craig, Hannah O'Keefe. 692

Understanding Remission of Long-Term Conditions Through Electronic Health Records: Scoping Review (e80796)
Hilda Hounkpatin, Benjamin Barton, Margaret Ogden, Rohini Mathur, Beth Stuart, Hajira Dambha-Miller. 718

Barriers and Facilitators to the 3 Sides of Extended Reality-Rehabilitation Adoption: Scoping Review (e80055)
Luuk Baltissen, Miranda Stienstra, Lien Denoo, Joris Knoben. 737

Explainable AI in Cancer Imaging: Scoping Review of Methods, Modalities, and Clinical Integration (e80645)
Dimitris Fotopoulos, Ioannis Ladakis, Dimitrios Filos, Pedro Moreno-Sánchez, Mark van Gils, Ioanna Chouvarda. 757

Receipt of Medicines Information From the Internet and Other Information Sources Among Adult Medicine Users in Developed Economies, 2010-2025: Systematic Review (e71984)
Niina Mononen, Milla-Maria Salo, Daria Timonen, Marika Pohjanoksa-Mäntylä, Marja Airaksinen. 782

Efficacy of Internet-Based Self-Help Interventions for Irritable Bowel Syndrome: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e87216)
Xiaohong Xu, Ting Liu, Fang Wang, Hongyue He, Jiao Chen, Jingxian Liu. 808

Toward Participatory Precision Health With Co-Designed Recommendations: Systematic Review of Just-in-Time Adaptive Interventions in Adolescents and Young Adults (e84422)
Kathleen Guan, Mohammed Amara, Viviana Cortiana, Imran Khan, Sıla Gürbüz, Carmine Iorio, Rayyan Ali, Christopher Adlung, Crystal Smit, Annabel Vreeker, Eva Thalassinou, Eeske van Roekel, Loes Keijsers, Mark de Reuver, Caroline Figueroa. 828

Large Language Models in Colorectal Cancer Care and Clinical Decision Support: Systematic Review (e89862)
Jinglei Tian, Qifeng Lou, Xue Wang, Hangying Xu, Huiting Mei, Yanli Yu. 850

Effectiveness of mHealth Interventions Aimed at Promoting Physical Activity and Reducing Sedentary Behavior on Work-Related Outcomes Among Workers: Systematic Review (e80540)
Takako Miki, Michiko Nohara, Nami Homma, Daiki Nagamine, Shinji Yamaguchi, Kyoko Nomura. 871

XSL•FO
RenderX

GenAI-Supported Virtual Patients in Health Care Education: Systematic Review (e82756)
Juming Jiang, Megan Ye, Tyrone Kwok, Janet Wong. 893

Applications of Natural Language Processing and Large Language Models for Social Determinants of Health: Systematic Review (e83793)
Swati Rajwal, Avinash Pandey, Ziyuan Zhang, Yankai Chen, Michael Liu, Sudeshna Das, Hannah Rogers, Abeed Sarker, Yunyu Xiao. 936

Machine Learning in HIV Care and Antiretroviral Therapy: Systematic Review (e79219)
Thamina Boudra, Arafate Idrissou, Oussama Barakat, Siamak Davani, Marie-Blanche Valnet Rabier, Jennifer Lagoutte-Renosi. 960

Patient Concerns Regarding Artificial Intelligence Applications in Health Care: Systematic Review and Meta-Synthesis Based on Social Ecological Theory (e85663)
Jiayu Hou, Zhiqiao Zhang, Xuan Cheng, Weihong Wang. 981

Online Community Support for Stroke Survivors and Caregivers: Scoping Review (e71190)
Putu Handayani, Kamila Alifia, Juliana Sutanto, Erlin Erlina, Sharyn Davies, Arthur Mawuntu, Muhammad Jusuf, Narelle Warren. 1015

Image-Based Deep Learning for Cataract Diagnosis: Systematic Review and Meta-Analysis (e78869)
Ruixi Li, Hongyi Li, Chong Li, Shuo Li, Linhong Lei, Dan Tao. 1043

Hybrid Care Modifications in the Delivery of Nonpandemic Care During the COVID-19 Pandemic: Scoping Review (e84756)
Natalia Sanchez Villalobos, Tessa van Loenen, Lem Ngongalah, Máire Connolly, Mart Stein, Chantal Rovers, Aura Timen. 1086

Use of Mobile Technologies for Smoking Cessation Among Smokers and Former Smokers: Systematic Review (e83072)
Meret Lakeberg, Falk Hoffmann, Jannik Ohmes, Birte Berger-Höger, Jasmin Helbach. 1109

Effects of Stand-Alone Digital Lifestyle Interventions on Weight-Related Outcomes in Adults With Overweight or Obesity: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e81070)
Si-An Lee, Jin-Hyuck Park. 1130

Deep Learning Algorithms Versus Radiologists in Digital Breast Tomosynthesis for Breast Cancer Detection: Systematic Review and Meta-Analysis (e91659)
Shewen Lyu, Zepeng Wang, Yujing Mu, Luyao Wang, Xiaohua Pei. 1155

Multidimensional Effects of Telemedicine on Patients With Spinal Cord Injury: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e87088)
Wen Zhong, Liyi Huang, Xin Sun, Rui Liu, Lu Wang, Quan Wei. 1170

Comparing Pulmonary Telerehabilitation and Center-Based Pulmonary Rehabilitation for Effectiveness and Adherence in Chronic Obstructive Pulmonary Disease: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e80500)
Ya Li, Hailong Zhang, Guixiang Zhao, Jiansheng Li, Haoxuan Huang, Longyu Wang, Peilin Jia. 1202

Clinical Effectiveness of Immersive Virtual Reality Exercise Interventions: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e87542)
Riley Brown, Megan Ross, Trevor Russell. 1231

Recognizing the Importance of Design, Content, and Delivery Features of Health Animations for Preventive Health Behaviors: Realist Review (e79769)
Kathleen McCorry, Gregory Maniatopoulos, Linda Errington, Ellie Land, Michael Craig, Santosh Vijaykumar, Till Bärnighausen, Nicola O'Brien.
1 2 6 5

XSL•FO
RenderX

Effectiveness of Digital Serious Games on Knowledge and Attitudes in Public Health Education: Systematic Review and Bayesian Network Meta-Analysis of Randomized Controlled Trials (e89281)
Di Huang, Dongjun Wu, Rene Hexel, Christine Brown-Wilson, Jing Zhou, Wendy Moyle. 1281

Accuracy of Deep Learning for Detecting Axillary Lymph Node Metastasis in Breast Cancer: Systematic Review and Meta-Analysis (e77593)
Xueying Wang, Tiantian Li, Xiaohang Wang, Deyuan Fu. 1320

Machine Learning in Predicting the Risk of Esophagogastric Variceal Bleeding Among Patients With Liver Cirrhosis: Systematic Review and Meta-Analysis (e78203)
Yuan Lian, Xinping Qiu, Geli Wang, Nannan Liu, Jing Zhao, Shanshan Song, Shiqi Wang, Mingjun Sun. 1341

Innovations in Deaf Health Care Communication: Systematic Review of Sign Language Recognition Systems (e70417)
Milena Marcolino, Lucca Oliveira, Lucas Valle, Luiza Rosa, Gabriela Sanches, Natalia Santos, Michelle Costa, Elidea Bernardino, Raniere Cordeiro, Raquel Prates, Zilma Reis, Mario Campos. 1364

Using Group Chats to Drive Behavior Change in Digital Health Interventions: Scoping Review and Realist Synthesis (e88911)
Hein Thu, Felicia Chan, Jumana Hashim, Shahmir Ali. 1383

Digital Technologies and Biomarkers for Locomotor Capacity Assessment in Older Adults: Systematic Review (e83814)
Shuhan Zhou, Jundan Huang, Jia Yu, Xiaoyang Li, Chi Zhang, Yinan Zhao, Mingyue Hu, Hui Feng. 1404

A Health-Related Digital Ecological Momentary Assessment in Children (Aged 5– 11 Years): Systematic Review (e79291)
Sydney Charitos, Lauren Thompson, Amberly Brigden, Jon Bird. 1425

The Role of Digital Biomarkers in Physiological Signal-Based Depression Assessment: Systematic Review and Meta-Analysis (e76432)
Hyeongsuk Lee, Seung-Gul Kang, SeonHeui Lee. 1452

Accuracy of Radiomics-Based Machine Learning for Predicting Risk of Recurrence in Non–Small Cell Lung Cancer: Systematic Review and Meta-Analysis (e84223)
Junpei Wu, Ye Zhang, Jiaye Wang, Chengshui Chen, Shiyu Hu, Wenyu Chen. 1477

Multimodal AI for Alzheimer Disease Diagnosis: Systematic Review of Datasets, Models, and Modalities (e85414)
Ziwen Yu, Anthony Mulholland, Tianyan Huang, Qiang Liu. 1491

Effectiveness of Mobile Health for Improving Medication Adherence in Patients With Cancer: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e85949)
Wei Liu, Li Wang, Qian Zhang, Jing Yang. 1527

Artificial Intelligence Tools for Automating Evidence Synthesis: Scoping Review (e81597)
Sashika Harasgama, Helen Pearce, Cameron Appel, Liam Loftus, Helena Painter, Isla Kuhn, Justine Karpusheff, Aji Ceesay, John Ford. . . .
1 5 5 3

Smartphone-Based Digital Phenotyping Across Health Conditions: Scoping Review (e84146)
Arlen Dumas, Joanne Hokayem, Georgia Goodman, Krishna Venkatasubramanian, Peter Chai. 1566

Effects of Virtual Reality–Based Interventions for Promoting Physical Activity in Patients With Heart Failure: Systematic Review (e86567)
Jeong-Ah Ahn, Jung Lee, Kyoung-A Kim. 1596

XSL•FO
RenderX

The Current Landscape of Remote Digital Symptom Monitoring for Patients With Lung Cancer: Scoping Review (e83666)
Hao Huang, Naomi Takemura, Hammoda Abu-Odah, Shuhan Li, Ting Mao, Zengjie Ye, Janelle Yorke. 1611

Association Between Short-Form Video Use and Mental Health: Systematic Review and Meta-Analysis (e82503)
Di Tang, Xin Zhang, Pengpeng Gou, Jie Feng, Rui Hu, Kim-wai Sum. 1627

Telehealth Approaches for Pediatric Otitis Media and Clinical Outcomes: Scoping Review (e85416)
Masao Noda, Ryohei Akiyoshi, Makoto Hosoya, Chikako Shinkawa, Ryota Koshu, Hidekane Yoshimura, Yukifumi Monden, Hiroaki Fushiki, Yasuhiro Tanaka, Makoto Ito. 1647

The Performance of Artificial Intelligence in Classifying Molecular Markers in Adult-Type Gliomas Using Histopathological Images: Systematic Review (e78377)
Obada Almaabreh, Rukaya Al-Dafi, Aliya Tabassum, Ahmad Othman, Alaa Abd-alrazaq. 1671

eHealth Literacy and Type 2 Diabetes Prevention Among At-Risk Populations: Mechanistic Systematic Review Using Theory-Driven Thematic Analysis (e77788)
Jingyi Li, Arina Azlan, Nurzihan Hassim, Yuan Wang, Ruina Guo. 1691

Artificial Intelligence Applications in Medical Devices for Personalized Health Care Solutions: Systematic Review (e72410)
Hemanth Ponnambalath Mohanadas, A Manikandan, Ahmad Ismail, Nick Tucker, Saravana Jaganathan. 1726

Effectiveness of Interventions for Internet, Smartphone, and Gaming Addictions: Umbrella Review and Meta–Meta-Analysis (e81705)
Minggang Zhang, Tong Lan, Tao Song, Xiaochun Wang. 1750

Effectiveness of Virtual Reality–Based Early Rehabilitation Strategies on Pain, Sleep, Anxiety, Balance, Cognition, and Limb Motor Function in Adult Intensive Care Unit Patients: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e81865)
Fei Wu, Yunting Wu, Yana Xing, Weixin Cai, Ran Zhang. 1764

Artificial Intelligence for Medicines Information: Scoping Review of Clinical Applications and Digital Health Inequalities (e77747)
Shahd Al-Arkee, Josephine Falade, Vibhu Paudyal. 1785

Machine Learning in Left Ventricular Hypertrophy Detection: Systematic Review and Meta-Analysis (e76637)
Yilin Li, Ke Zhao, Jing Wu. 1799

Accuracy of Medical Image–Based Deep Learning for Detecting Microvascular Invasion in Hepatocellular Carcinoma: Systematic Review and Meta-Analysis (e82000)
Wei Feng, Bo Qu, Shuo Han. 1817

The Predictive Value of Machine Learning for Postoperative Delirium in Cardiac Surgery: Systematic Review and Meta-Analysis (e72304)
Yi Guo, Hong Xu, Ankui Wang, Mingming Zhang, Shuai Zhang, Peng Xie. 1846

Digital Inequalities in the Use of eHealth Services in European Public Health Care Systems: Systematic Review of Observational Studies (e81841)
Garazi Monasterio, Marcos Fernández-López, Erika Valero, Unai Martin, Amaia Ayala-García. 1896

XSL•FO
RenderX

Data-Driven Guideline Adherence in Data Representation and Compliance Measurement: Scoping Review (e79937)
Minh Hoang, Candice Donnelly, Christina Igasto, Amith Shetty, Malcolm Pradhan, Tim Shaw. 1919

Effects of Digital Health Interventions to Promote Safer Sex Behaviors Among Youth: Systematic Review and Bayesian Network Meta-Analysis (e87071)
Yiran Zhu, Wenwen Peng, Die Hu, Edmond Choi, Maritta Välimäki, Ci Zhang, Xianhong Li. 1934

Characterization of Models for Identifying Physical and Cognitive Frailty in Older Adults With Diabetes: Systematic Review and Meta-Analysis (e84617)
Xia Wang, Shujie Meng, Xiang Xiao, Liu Lu, Hongyan Chen, Yong Li, Rong Zhang, Qiwu Jiang, Shan Liu, Ru Gao. 1967

Predictive Performance of Artificial Intelligence Algorithms for Gestational Diabetes Mellitus in Pregnant Women: Systematic Review and Meta-Analysis (e79729)
Yingni Liang, Anran Dai, Meiyan Luo, Zhuolian Zheng, Jiayu Shen, Yinhua Su, Zhongyu Li. 1988

Barriers to Digital Health Adoption in Older Adults: Scoping Review Informed by Innovation Resistance Theory (e75591)
Yosefa Birati, Roy Tzemah-Shahar. 2011

Machine Learning Techniques Used for the Identification of Sociodemographic Factors Associated With Cancer: Systematic Literature Review (e79187)
Liz González-Infante, Gaston Marquez, Solange Parra-Soto, Mónica Cardona-Valencia, Carla Taramasco. 2035

Patient Benefits in the Context of Sepsis-Related AI-Based Clinical Decision Support Systems: Scoping Review (e76772)
Pascal Raszke, Godwin Giebel, Jürgen Wasem, Michael Adamzik, Hartmuth Nowak, Lars Palmowski, Philipp Heinz, Nina Timmesfeld, Marianne Tokic, Frank Brunkhorst, Nikola Blase. 2066

Diagnostic Performance of Deep Learning and Radiomics in Extracranial Carotid Plaque Detection: Systematic Review and Meta-Analysis (e77092)
Lingjie Ju, Yongsheng Guo, Haiyong Guo, Ruijuan Liu, Yiyang Wang, Siyu Wang, Na Ma, Junhong Ren. 2081

Evidence for Digital Health Tools Designed to Support the Triage of Musculoskeletal Conditions in Primary, Urgent, and Emergency Care Settings: Scoping Review (e81578)
Linda Truong, James Wrightson, Raphaël Vincent, Eunice Lui, Jamon Couch, Ellen Wang, Cobie Starcevich, Dean Giustini, Alex Haagaard, Elena Lopatina, Niels van Berkel, Michael Rathleff, Clare Ardern. 2107

The Diagnostic Value of Image-Based Machine Learning for Osteoporosis: Systematic Review and Meta-Analysis (e75965)
Rui Zhao, Haolin Yang, Yangbo Li, Xiaoyun Li, Zhijie Yang, Yanping Lin, Jiachun Huang, Lei Wan, Hongxing Huang. 2128

The Application of Mobile Health in Self-Management Among Patients Undergoing Dialysis: Scoping Review (e76880)
Qin Xu, Yulin Xu, Xiaoqin Liu, Xiaolin Ma. 2151

Accuracy of Deep Learning in Diagnosing Chronic Obstructive Pulmonary Disease: Systematic Review and Meta-Analysis (e83459)
Hui Yang, Yijiu Wu, Tong Wu, Jingyan Ji, Sitao Lei, Weibin Xu. 2174

Assessment of the Diagnostic Performance and Clinical Impact of AI in Hepatic Steatosis: Systematic Review and Meta-Analysis (e78310)
Jiamei Song, Dan Liu, Jitong Li, Haoru Cong, Ruixue Deng, Yihan Lu, Jiayi Sun, Jingzhou Zhang. 2192

AI for Detecting and Predicting Postpartum Depression: Scoping Review (e77376)
Mais Alkhateeb, Ajisha Nayeem, Arfan Ahmed, Mohammed Alsahli, Javaid Sheikh, Alaa Abd-Alrazaq. 2219

XSL•FO
RenderX

Face-to-Face Versus Digital, Telephone-Delivered, and Self-Help Cognitive Behavioral Therapy for Irritable Bowel Syndrome: Systematic Review and Bayesian Indirect Treatment Comparison Meta-Analysis (e75833)
Qing-Feng Tao, Can Hua, Xiao Zhuo, Jian-Jiao Mou, Chao-Rong Xie, Yu-Xin Zhang, Bei Lv, Xin-Ying Niu, Min Chen, Hui Zheng. 2246

Digital Interventions Targeting Healthy and Sustainable Eating Behavior: Systematic Review and Meta-Analysis (e80821)
Käbi Vanwinkelen, Bram Spruyt, Tim Smits. 2262

Diet-Related Health Recommender Systems for Patients With Chronic Health Conditions: Scoping Review (e77726)
Xiaolan Dong, Bei Yun, Anni Pakarinen, Zhuting Zheng, Hao Niu, Tian Jin, Changrong Yuan, Jingting Wang. 2287

AI-Supported Digital Microscopy Diagnostics in Primary Health Care Laboratories: Scoping Review (e78500)
Joar von Bahr, Antti Suutala, Vinod Diwan, Andreas Mårtensson, Johan Lundin, Nina Linder. 2307

Effectiveness of Machine Learning in Detecting Vessels Encapsulating Tumor Clusters in Hepatocellular Carcinoma: Systematic Review and Meta-Analysis (e82839)
Huili Shui, Wenyu Wu, Zhenming Xie, Bing Yang, Jia Deng, Dongxin Tang. 2327

Machine Learning Prediction Models for Preeclampsia: Systematic Review and Meta-Analysis (e78714)
Lu Liu, Qixuan Zhu, Yichi Zong, Xueyuan Chen, Wei Zhang, Jun Wang. 2339

Communication Strategies to Promote Patient Engagement in Telemedicine: Systematic Review (e85456)
Yangna Hu, Cindy Ngai, Rui Jiang. 2365

Products, Performance, and Technological Development of Ambulatory Oxygen Therapy Devices: Scoping Review (e81077)
Shohei Kawachi, Mariana Hoffman, Lorena Romero, Magnus Ekström, Jerry Krishnan, Anne Holland. 2387

The Development and Use of AI Chatbots for Health Behavior Change: Scoping Review (e79677)
Lingyi Fu, Ryan Burns, Yuhuan Xie, Jincheng Shen, Shandian Zhe, Paul Estabrooks, Yang Bai. 2405

Efficacy of Brain-Computer Interface Therapy for Upper Limb Rehabilitation in Chronic Stroke: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e79132)
HongJie Chen, GuoJun Yun. 2425

Behavioral Determinants and Effectiveness of Digital Behavior Change Interventions for the Prevention of Sexually Transmitted Infections and HIV: Overview of Systematic Reviews (e74201)
Giuliano Duarte-Anselmi, Susana Sanduvete-Chaves, Salvador Chacón-Moscoso, Daniel López-Arenas. 2448

Effect of Digital Health Interventions on College Students' Lifestyle Behaviors: Systematic Review (e82192)
Qingyuan Zhou, Jiajun Jiang, Zhihua Yin, Ruishi Fan. 2471

Key Components and Barriers in Web-Based Suicide Prevention Gatekeeper Training: Systematic Narrative Review (e81572)
Olivier Ferlatte, Emmanuelle Gareau, Keven Lee, Kinda Wassef, John Oliffe, Hannah Kia, Brock Dumville. 2496

Interventions That Use Highly Visual Social Media Platforms to Tackle Unhealthy Body Image in Adolescents and Young Adults: Systematic Review of Randomized Controlled Trials and Quasi-Experimental Studies (e80141)
Anthony Copez-Lonzoy, Janina Bazalar-Palacios, Nikol Mayo-Puchoc, Gustavo Calderon de la Cruz, Liliana Cruz-Ausejo, Evelyn Goicochea-Ríos, Juan Bazo-Alvarez. 2516

XSL•FO
RenderX

Text-Based Depression Estimation Using Machine Learning With Standard Labels: Systematic Review and Meta-Analysis (e82686)
Shengming Zhang, Chaohai Zhang, Jiaxin Zhang. 2531

AI and Wearables for Early Detection of Cognitive Impairment and Dementia: Systematic Review (e86262)
Ander Cejudo, Markel Arrojo, Cristina Martín, Aitor Almeida. 2551

Design Practices for Data Dashboards in Health Care: Scoping Review (e77361)
Heike Vornhagen, Stephen Barrett, Ciara Carroll, Lydia Iladiva, Gregory Martin, Declan McKeown, Jennifer Martin. 2577

The Effects of Digital Health Interventions for Family Members in Intensive Care Units: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e83294)
Chuchu Zhang, Weijing Sui, Weilin Jiang, Yibing Shen, Zhenzhen Huang, Ran Yan, Jia Yi, Jiayu Zhang, Chaoyi Zhang, Yiyu Zhuang, Xiaoyan Gong. 2588

AI-Enabled Wearables for Motor Function Assessment and Rehabilitation in Parkinson Disease: Scoping Review (e85596)
Shengting Li, Siqi Chen, Xiaosong Yu, Huixiang Shang, Tingting Tu, Mingtao Quan. 2615

Effectiveness of Telerehabilitation Interventions for Self-Management of Tinnitus: Update of a Systematic Review (e83529)
Sara Demoen, Elise Van Kerchove, Annick Timmermans, Vincent Van Rompaey, Sarah Michiels, Annick Gilles. 2640

Digital Health Interventions to Promote Physical Activity Among Adolescents: Systematic Review (e82395)
Rui Fan, Jia Jiang, Qing Zhou, Xin Zhang, Zhou Wu, Liu Ji. 2677

Digital Technologies That Support Meaningful Connections in Care Homes: Scoping Review (e88181)
Deborah Muldrew, Rosemary Bradley, Kelly Conway. 2704

Association Between Digital Biomarkers of Health and Anxiety: Systematic Review and Meta-Analysis (e73812)
Yolanda Lau, Natalia Chemas, Heema Ajeet Gokani, Rachel Morrell, Harisd Phannarus, Claudia Cooper, Zuzana Walker, Harriet Demnitz-King, Natalie Marchant. 2728

Effects of Digital Health Interventions on Functional and Psychological Outcomes in Older Patients With Hip Fractures: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e79563)
Wei Fan, Qi Zhang, Qunfeng Lu. 2756

Use of Health and Welfare Technology in Palliative Care: State-of-the-Art Review (e79637)
Viktoria Zander, Maja Holm, Monir Mazaheri, Christine Gustafsson, Sara Landerdahl Stridsberg, Ragnhild Hedman. 2780

The Effects of Digital Health Interventions on Motor Symptoms, Nonmotor Symptoms, and Quality of Life in Patients With Parkinson Disease: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e79935)
Ruwen Liu, Sirui Zhang, Yi Xiao, Yangfan Cheng, Huifang Shang. 2798

Effects of Wearable Devices on Parkinson Disease: Systematic Review and Meta-Analysis of Randomized Controlled Trials Within the International Classification of Functioning, Disability, and Health Framework (e85914)
Jiarong Wu, Wanli Zang, Mingqing Fang, Ningkun Xiao, Xianzuo Zhang, Su Wang, Qiuxia Zhang. 2826

Determinants and Health Outcomes of Digital Health Literacy in Patients With Cardiovascular Disease: Systematic Review and Meta-Analysis (e89102)
Eun-Jung Kim, Da-Young Kim, Ha-Jin Kim, Youn-Jung Son. 2850

XSL•FO
RenderX

Robot-Assisted Therapy for Upper Limb Rehabilitation After Stroke: Umbrella Review (e79363)
Sijia Liu, Xianghong Zhang, Lijun Zhou, Jiashan Zhang, Jialin Liu, Chengqi He. 2872

Technology-Based Interventions for Prevention of Type 2 Diabetes Following Gestational Diabetes: Systematic Review and Meta-Analysis (e78841)
Claire Eades, Anh Nguyen-Hoang, Louise Hoyle, Dawn Cameron, Josie Evans. 2894

Predictive Value of Machine Learning for Poststroke Mortality Risk: Systematic Review and Meta-Analysis (e83821)
Yujie Chen, Zhujing Ou, Yiting Deng, Anling Luo, Xuzi Li, Yujia Yang, Qin Yang, Xintong Wu, Muke Zhou. 2918

Measuring Technology-Facilitated Sexual Violence and Abuse: Scoping Review of Existing Measures (e90068)
Sharon Pak, Linruo Zhang, Chanchan Wu, Pui Chau, Caroline Bradbury-Jones, Edmond Choi. 2972

Multimodal Intelligent Monitoring of Parkinson Disease: Scoping Review of Progress and Translational Challenges (e89057)
Jianhui Tan, Xiuzhen Deng, Cuilan Wu, Molin Yao, Jingyi Liao, Haoli Zheng, Changlin Lian. 2992

Toward a Common Set of Interface Requirements for Genomic Data Management: Scoping Review (e78405)
Valeria Resendez, Funda Yıldırım, Eugenio Gaeta, Giuseppe Fico, Simone Borsci. 3016

Media Exposure and Its Association With Vaccine Attitudes, Intentions, and Hesitancy: Systematic Review (e74280)
Francesco Leonforte, Vito Nicosia, Paola Comite, Giustino Morlino, Antonio Mistretta. 3037

Current Practices and a Novel Operational Framework for Planning Research on Digital Health Promotion Interventions From Development to Implementation: Scoping Review (e82611)
Claire Collin, Corinne Alberti, Philippe Martin, Graham Moore, Clara Eyraud, Enora Le Roux. 3062

Automated Approaches of Text Simplification of Patient Education Materials: Scoping Review (e88365)
Cornelia Krenn, Christine Loder, Natalie Berger, Klaus Jeitler, Thomas Semlitsch, Andrea Siebenhofer, Denise Wilfling. 3080

Reporting of Telehealth Implementation in Cystic Fibrosis: Scoping Review Using a Novel Theory-Based Evaluation Lens (e86194)
Tamara Vagg, Robyn Doherty, Sarath Ranganathan, Barry Plant. 3103

Blockchain-Enabled Self-Sovereign Identity Applications in Health Care: Scoping Review (e89574)
Abha Pokharel, Surya Kathayat. 3125

Effectiveness of Digital Health Interventions to Improve Self-Care in Patients With Chronic Diseases: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e88708)
Jessica Longhini, Daniel Pedrotti, Federica Foladori, Melania Stedile, Francesca Stefani, Michela Dal Ben, Alessandro Froner, Marta Proietti Pesci, Stefano Toccoli, Anna Brugnolli. 3143

Navigating Diabetes Management in the Digital Era: Scoping Review of Online Health Information-Seeking Behavior (e82081)
Pin-Heng Tiao, Hsun-Yu Chan, Yen-Ming Huang. 3166

The Emerging Roles of AI in Self-Directed Stress Management: Systematic Review (e90709)
Mary Reyes, Shauna Teo, Andree Hartanto. 4116

XSL•FO
RenderX

Thematic Mapping and Evolution of Social Media Mining in Health Research: Hybrid Bibliometric Synthesis (e86200)
Mia Yang, Sabine Bohnet-Joschko. 5644

Effectiveness of mHealth Interventions for Improving eHealth Literacy Among Patients With Chronic Diseases: Meta-Analysis and Systematic Review (e82004)
Xueqin Yang, Yingping Ma, Yueting Wang, Mengting Liu, Xingyu Liu, Xueping Jiao, Fanghong Yan, Yuxia Ma, Lin Han, Yanan Zhang. 6098

Application of Digital Health Technologies in Cardiac Rehabilitation for Patients With Coronary Heart Disease: Scoping Review (e85917)
Xinyu Zhu, Lei Liu, Yingjie Wang, Hongyuan Li, Min Zang, Jiayu Wang, Yu Tian, Zihan Li. 6998

Patient Experiences With Online Laboratory Test Presentations From Access to Activation: Systematic Review (e88259)
Mia Lustria, Lovinta Atrinawati, Obianuju Aliche, Kyunghye Kim, Zhe He. 7632

How Cancer Online Support Groups Work, for Whom, and in What Circumstances: Realist Review (e77445)
Jacqueline Bender, Sarah Scruton, Geoff Wong, Lauren Squires, Stephanie Babinski, Marina Englesakis, Alaina Cyr, Henry Potts, Andrea Tricco, Richard Wassersug, Mary Esplen, Victoria Forster, Olga Perski, Holly Witteman, Charlene Soobiah, Janet Papadakos, Colleen Young, Colleen Fox. 7893

Machine Learning for Predicting Stroke Risk Stratification Using Multiomics Data: Systematic Review (e85654)
Hae Yoo, Hyerim Shin, Eun-Jung Kim, Youn-Jung Son. 8142

Immersive Technologies for Cognitive Rehabilitation in Dementia and Mild Cognitive Impairment: Systematic Review (e84349)
Jaromir Konecny, Giuseppe Lanza, Serafino Buono, Raffaele Ferri, Antonina Luca, Radek Martinek, Sabrina Musso, Aurora Palmigiano, Barbora Prauzkova, Angelica Quercia, Francesco Rundo, Ramachandran Subramanian, Alessandro Serretti, Michal Prauzek. 8955

Application Value of Radiomics-Based Machine Learning for Preoperative Risk Stratification of Bladder Cancer: Systematic Review and Meta-Analysis (e81084)
Zirong He, Yinghua Liu, Qin Jiang, Fang Hou, Bing Xu. 9409

Artificial Intelligence for Predicting Treatment Response in Patients With Anxiety Disorders After Cognitive Behavioral Therapy: Systematic Review and Meta-Analysis (e86079)
Jiawen Liu, Junhui Wang, Zhaobin Wu, Mohamad Adam Assim. 9722

Viewpoints

The Warmth of Medicine: The Irreplaceability of Doctors in the Era of Artificial Intelligence (e97412)
Yu-ying Yang, Jian-min Liu. 3191

Solidarity or Segregation? ChatGPT Health and US Health Care Disparities (e94972)
Andrew Barnhart, Giuseppe Comerci, Barbara Prainsack, Matthias Braun. 3195

Enhancing Physician Resilience to Generative AI: Multilevel Framework for Shared Authority, Verification, and Skill Preservation (e88058)
Hongxia Pan, Jialin Liu, Siru Liu. 3203

Personal Health Large Language Models and the Negotiation of Medical Authority in Clinical Care: Opportunities, Risks, and Governance (e91727)
Wenyi Xie, Jialin Liu, Siru Liu. 3214

XSL•FO
RenderX

Lessons Learned From Implementing a Culturally Tailored Virtual Support Program for Asian American Breast Cancer Survivors With Pain and Depression (e81187)
Seulgi Ryu, Dongmi Kim, Wonshik Chee, Eun-Ok Im. 3228

Unique Digital Images as Incentives in Clinical Trials: A Digital Shift Toward Meaningful Participation (e88022)
Xavier Tadeo, Gyula Seres, Peter Wang, Yoann Sapanel, Alexandria Remus, Jasmine Eyal, Jason Lee, Simon Chesterman, Julian Savulescu, John Wong, Christopher Asplund, R Stone, Benjamin Tee, Reza Shokri, Marlena Raczkowska, Li Chong, Siong Kwek, Agata Blasiak, Dean Ho. 3239

Using Ultra-Abridged Individual Difference Scales for Personalization in Digital Mental Health to Improve Uptake, Engagement, and Experiences: Three-Tiered Decision Framework for Scale Shortening (e80662)
Siu Yeung, Alan Tong, Han Zhao, Winnie Mak. 3251

Ethical Governance of Large Language Models in Health Care: Trust, Responsibility, and Equity in Routine Use (e93470)
Xiongwen Yang, Chuan Xu. 3267

Nontraditional Data in Pandemic Preparedness and Response: Identifying and Addressing First- and Last-Mile Challenges (e85540)
Mattia Mazzoli, Irma Varela-Lasheras, Sónia Namorado, Constantino Pereira Caetano, Andreia Leite, Lisa Hermans, Niel Hens, Polen Türkmen, Kyriaki Kalimeri, Leo Ferres, Ciro Cattuto, Daniela Paolotti, Stefaan Verhulst. 3275

Rheumatic Digital Twin: Proposed Machine Learning–Based Multimodal Framework to Inform Clinical Decision-Making (e86763)
Daniyal Selani, Rachel Knevel, Marcel Reinders, Erik van den Akker. 3292

AI in Clinical Decision Support Systems: Promising Applications and Strategies for Managing Data Challenges (e71532)
Jennifer Daly, Dursun Delen, Zheng Han, River Smith, Jacqueline Honerlaw, Kelly Cho, Bridget Bennett, Jennifer Sippel. 3308

From Pilot Trap to Institutional Capacity: A Governance Framework for Sustainable Clinical AI Implementation in Health Systems (e92680)
Jin Tian, Zengren Zhao, Longmei Tang, Yongzhao Song, Yuchang Li, Nan Jiang. 3322

Governing Patient-Facing AI-Generated Video in Digital Health: A Risk-and-Ethics Matrix for Deployment, Monitoring, and Change Control (e91940)
Yongzheng Hu, Wei Jiang. 3339

The Actionable Innovation Day Approach: Participatory Model for Advancing Critical Care Innovation (e73614)
Brett Hryciw, Cecilia Tran, Rashi Ramchandani, Cameron Love, Christine Caron, Aimee Sarti, Annelise Miller, Suzanne Madore, Michael Chasse, Andy Pan, Simon Didcote, Scott Millington, Heather Galley, Kwadwo Kyeremanteng, Andrew Seely. 3354

The Digital Exposome: A Life Course Framework for Health in the Digital Age (e90153)
Pascal Petit, Nicolas Vuillerme. 3366

Co-Lifecycle Governance for Learning Medical AI: A Hybrid Convergence Framework for Adaptive Regulatory Oversight (e90654)
Jae Lee, Boram Choi, Kwunho Jeong, Sang Suh, Ju Kim, Dae-Soon Son. 3375

XSL•FO
RenderX

Frailty-Based Remote Monitoring in Older Adults With Heart Failure: Conceptual Framework for Adaptive Digital Health Strategies (e88681)
Rémi Esser, Olivier Maurou, Marine Larbaneix, Alejandro Mondragon, Marlène Esteban, Christine Farges, Nicolas Pages, Sophie Nisse Durgeat, Marc Harboun. 3387

The Use of Speculative Fiction in Future-Focused Health Care Research (e89699)
Sjors Groeneveld, Jobke Wentzel, Melissa Laurens, Lisette Gemert-Pijnen, Verdaasdonk, Harmieke Os-Medendorp, Marjolein Ouden. 3395

Rerouting Eye Care: How AI and Telemedicine Are Reshaping Ophthalmology Patient Journeys in the United Kingdom and Germany (e93170)
Broder Poschkamp, Ariel Ong, Justin Engelmann, Michael Grün, Andreas Stahl, Pearse Keane, David Merle. 3407

Extrinsic Trust as a Contractual Framework for Accountable AI in Health Care: Viewpoint (e83903)
Anthony Kelly. 3417

Digital Twins for Just-in-Time Adaptive Interventions (JITAIs): Framework for Optimizing and Continually Improving JITAIs (e72830)
Asim Gazi, Daiqi Gao, Susobhan Ghosh, Ziping Xu, Anna Trella, Predrag Klasnja, Susan Murphy. 3429

AI Agents and Epidemic Intelligence on Respiratory Infectious Diseases: Toward a Conceptual Framework Integrating Decision Support (e86936)
Liuyang Yang, Liyu Shan, Xiaolin Cao, Jinzhao Cui, Michael Tong, Yan Niu, Ting Zhang. 3450

Deep Research Agents: Major Breakthrough or Incremental Progress for Medical AI? (e88195)
Matthew Wong, Ariel Ong, David Merle, Pearse Keane. 3458

From Knowledge Graphs to Digital Twins: Perspectives on Modeling Patient Outcomes for Health Care Quality Assessment (e81946)
Anna-Katharina Nitschke, Juan Diaz Ochoa, Simone Neumaier, Markus Knott. 3470

Integrating Generative AI Into Patient-Centered Clinical Decision Support: Viewpoint on Research and Practice Considerations (e81628)
Prashila Dullabh, Courtney Zott, Nicole Gauthreaux, Caroline Peterson, Abigail Aronoff, Kistein Monkhouse, Dean Sittig. 3492

Beyond Missingness: Systematizing Methods for Comprehensive Data Fitness Assessment in Clinical Research (e76398)
Hanieh Razzaghi, Kaleigh Wieand, Kimberley Dickinson, Michael Kahn, Jason Roy, Clair Blacketer, Dimitri Christakis, Christopher Forrest, Jane Greenberg, Harold Lehmann, Keith Marsolo, Jennifer Sciolla, Mark Weiner, Nicole Weiskopf, L Bailey. 3503

Scalable Agile Framework for Execution in AI for Medical AI Ethics Policy Design in Small- and Medium-Sized Enterprises (e80028)
Ion Nemteanu, Adir Mancebo Jr, Leslie Joe, Ryan Lopez, Patricia Lopez, Warren Pettine. 3516

From Agents to Governance: Essential AI Skills for Clinicians in the Large Language Model Era (e86550)
Weiping Cao, Qing Zhang, Jialin Liu, Siru Liu. 3539

Collaborative and Cooperative Hospital “In-House” Medical Device Development and Implementation in the AI Age: The European Responsible AI Development (EURAID) Framework Compatible With European Values (e80754)
Anett Schönfelder, Maria Eberlein-Gonska, Manfred Hülsken-Giesler, Florian Jovy-Klein, Jakob Kather, Elisabeth Kohoutek, Thomas Lennefer, Elisabeth Liebert, Myriam Lipprandt, Rebecca Mathias, Hannah Muti, Julius Obergassel, Thomas Reibel, Ulrike Rösler, Moritz Schneider, Larissa Schlicht, Hannes Schlieter, Malte Schmieding, Nils Schweingruber, Martin Sedlmayr, Reinhard Strametz, Barbara Susec, Magdalena Wekenborg, Eva Weicken, Katharina Weitz, Anke Diehl, Stephen Gilbert. 3548

AI Triage in Primary Care: Building Safer and More Equitable Real-World Evidence (e88396)
Aymn Alamoudi, Evangelos Kontopantelis, Salwa Zghebi, Benjamin Brown. 3571

XSL•FO
RenderX

Digital Health as an Enabler of Personalized Medicine in Ghana: Gaps, Opportunities, and Future Directions (e70696)
Godsway Sackey, Babajide Owoyele, Frank Baiden, Stefan Konigorski. 3583

Rethinking Trust in Synthetic Health Data: Lessons From 7 European Research Initiatives (e83369)
Jens Declerck, Dipak Kalra, Antti Airola, Ahmed Amer, Christos Chatzichristos, Maria del Mar Mañu Pereira, Bruno M. de Brito Robalo, Francesco Ghini, Alberto Gutierrez-Torre, Sem Hoogteijling, Susanne Hultsch, Jan Ramon, Sara Reidel, Francesco Regazzoni, Luís Silva, Inês Silveira, Tsekeridou Sofia, Christophe Maes. 3589

Clinical AI is Not (Yet) Trustworthy-But It Could Be (e85433)
Ali Saad, Sofia Dias, Ghada Alhussein, David Lyreskog, Ioannis Gerasimou, Beatriz Alves, Maarten de Vos, Ioannis Drivas, John Zaras, Andreas Stergioulas, Iskanter Bensenousi, Leontios Hadjileontiadis, Christos Chatzichristos, Stelios Hadjidimitriou. 3596

Open-Source Large Language Models and AI Health Equity: A Health Service Triangle Model Perspective (e86769)
Shaolong Wu, Wenxin Zou, Jiong Tu, Chunxiao Wang, Cheng Jin, Jing Liao, Kwok Tang, Ning Liu, Chun Hao. 3609

Pediatric Clinical Images Without Consent: A Governance Gap in the Long-Term Reuse of Health Data in Digital Health Ecosystems (e94952)
Iwona Bujek. 3620

Training AI Models for Aesthetic Facial Evaluation: Focused Review and Framework to Mitigate Homogenizing Bias (e95452)
Anisha Kumar, Lav Varshney. 3626

Ethical Considerations in Personal Health Large Language Models (e92240)
Jialin Liu, Siru Liu. 3644

Policy Considerations for National Virtual Hospitals: Global Evidence and the Seha Virtual Hospital Model (e89276)
Sohaib Shujaat, Hawazin Almutairi, Saleh Alhuraibi, Hongyang Ma. 10356

Assessing the Evolution and Influence of Medical Open Databases on Biomedical Research and Health Care Innovation: A 25-Year Perspective With a Focus on Privacy and Privacy-Enhancing Technologies (e58954)
Albert Yang, Mei-Lien Pan, Henry Lu, Chung-Yueh Lien, Da-Wei Wang, Chih-Hsiung Chen, Der-Cherng Tarng, Dau-Ming Niu, Shih-Hwa Chiou, Chun-Ying Wu, Ying Sun, Shih-Ann Chen, Shuu-Jiun Wang, Wayne Sheu, Chi-Hung Lin. 11252

Africa's Digital Health Revolution: The Digital Fit-Viability Model to Move From Innovation to Scaled Implementation (e63495)
Afra Jiwa, Antony Ngatia, Karim Benali, Niclas Boehmer, Sangu Delle, Patrick Emedom-Nnamdi, Chris Fofie, Christine O'Brien, Tobi Olatunji, Kate Obayabgona, Milind Tambe, Richard Fletcher, Adeline Boatin, Bethany Hedt-Gauthier. 11267

Backcasting the Trust Gap: A Strategic Road Map for Clinician Adoption of AI Diagnostics by 2040 (e94234)
Yunguo Yu. 11556

A Futures Framework for Clinical AI Governance: Anticipating Emerging Risks, Shifting Roles, and Regulatory Challenges (e96152)
Yi Yang, Jialin Liu, Siru Liu. 11596

XSL•FO
RenderX

Tutorials

Integration of Federated Learning and Blockchain in Health Care: Tutorial on Medical Data, Architectures, Privacy, Security, and Regulatory Compliance (e80178)
Yahya Shahsavari, Yaser Baseri, Abdelhakim Hafid, Oussama Dambri, Dimitrios Makrakis. 3661

Building EVA (Educación, Vinculación, y Autoayuda): Tutorial for the Development of a Digital Mental Health Chatbot for Adolescents Living With HIV (e75351)
Diego Vasquez, Neil Rupani, Carmen Contreras, Lenka Kolevic, Molly Franke, Jerome Galea. 3724

Co-Develop-IT! Unifying Methodological Guideline for the Co-Design, Development, and Evaluation of Individually Tailored Technology-Enhanced Training and Rehabilitation Concepts: Consensus Development Study and Tutorial (e84163)
Patrick Manser, Lotte Hardeman, Andreas Wallin, David Moulaee Conradsson, Breiffni Leavy, Franziska Albrecht, Melvyn Roerdink, Jean-Jacques Temprado, Eling de Bruin, Erika Franzén. 3744

Using an Adult-Designed Wearable for Pediatric Monitoring: Practical Tutorial and Application in School-Aged Children With Obesity (e76166)
Alessandra Angelucci, Andrea Aliverti, Matteo Vandoni, Daniela Lucini, Cristiana Larizza, Vittoria Carnevale Pellino, Gianvincenzo Zuccotti, Valeria Calcaterra. 3766

SMARTCLOTH Prototype for Dietary Management in Patients With Diabetes Mellitus: Tutorial on Human-Centered Design Methodology for Health Care Hardware Development (e75744)
Jose Palomares, Rafael Molina-Luque, Fernando León-García, Irene Casares-Rodríguez, María García-Rodríguez, María Villena Esponera, Guillermo Molina-Recio. 3786

Developing a Trauma-Informed Social Media Campaign to Disseminate Endometriosis-Specific Qualitative Art-Based Research Findings: Tutorial (e83491)
Kerry Marshall, Hargun Dhillon, A Howard, Heather Noga, Grace Yang, William Zhu, Jessica Sutherland, Sarah Lett, Anna Leonova, Paul Yong, Natasha Orr. 3812

Team-Based Analysis of Large-Scale Qualitative Data: Tutorial Using a Nationwide SMS Text Messaging Poll of Youth (e72526)
Melissa DeJonckheere, Samantha Chuisano, Marika Waselewski, Kendrin Sonneville, Tammy Chang. 3834

Commentaries

Translating Telehealth Communication Research Into Patient-Centered, Implementable Practice (e93690)
Rachel Pittmann, Paula Koppel, David Barrett. 3845

Beyond GPT-4: The Rapidly Evolving Potential of Large Language Models for Clinical Guideline Improvement (e95004)
Scott Nelson, Adam Wright. 3850

Why We Need Patients and Community at the Center of AI Health Communication Research (e97577)
Julie Ayre, Lexuan Shao, Adam Dunn. 3854

Rethinking Postdischarge Intervention Evaluation (e98435)
Brigitte Woo. 3857

Moving From Keywords to Contextual Meaning: A Commentary on Hybrid Bibliometric Synthesis in Health Research (e102159)
Dimitrios Zikos. 3860

XSL•FO
RenderX

From Data Stewardship to Model Stewardship: Extending Governance Frameworks for AI Era Health Data Use (e97859)
Leon Rozenblit, Steven Labkoff, Charles Safran. 3864

Maturity, Safety, and Equity of AI-Enabled Systems and Triage in Integrated Primary Care (e97341)
Siaw-Teng Liaw. 3868

Hospitalists Are Already Using AI—Why Implementation Will Determine Its Impact (e97419)
Anna Maw, Aakriti Pandita, Marisha Burden. 3872

Beyond Time Saved: Implementation, Equity, and the Utility Threshold for Nursing AI Scribes (e101190)
Charlene Ronquillo. 3876

From Design to Accountable Impact for Data Dashboards in Health Care (e98272)
Barbara-Jo Achuff. 3880

Digital Health Interventions for Family Members of ICU Patients: Current Evidence and Future Directions (e95876)
Tanya Abenaim, Dan Poenaru, Michael Goldfarb. 3883

Letters to the Editor

Beyond Visual Consensus: Tiered Reference Framework for AI Cystoscopy Studies (e101910)
Ahmet Bayraktar, Bilgi ler. 10846

Authors' Reply: Beyond Visual Consensus: Tiered Reference Framework for AI Cystoscopy Studies (e103335)
Chung-You Tsai, Shi-Wei Huang. 10848

Human-in-the-Loop as a Safety Guardrail: Clinical Accountability in the Large Language Model Era (e85726)
Isaac Zablah, Yolly Molina, Antonio Garcia-Loureiro. 10850

Corrigenda and Addendas

Correction: Characterization of Models for Identifying Physical and Cognitive Frailty in Older Adults With Diabetes: Systematic Review and Meta-Analysis (e92717)
Xia Wang, Shujie Meng, Xiang Xiao, Liu Lu, Hongyan Chen, Yong Li, Rong Zhang, Qiwu Jiang, Shan Liu, Ru Gao. 10853

Correction: User Profiles and Engagement in a Hypertension Self-Management App: Cross-Sectional Survey (e93515)
Felix Muehlensiepen, Susann May, Frances Seifert, Eileen Wengemuth, Olen Johannsen, Martin Middeke, Martin Heinze, Pascal Petit, Nicolas Vuillerme, Sebastian Spethmann, Dunja Bruch. 10855

Correction: Leveraging Social Media to Achieve Population-Level Reach of Lung Cancer Screening-Eligible Individuals: A RE-AIM Framework Perspective (e94664)
Lisa Carter-Bawa, Jamie Ostroff, Susan Rawl, Erin Hirsch, Smita Banerjee, Andrew Ciupek, Robert Comer, Minal Kale, Katherine Leopold, Patrick Monahan, James Slaven Jr, Francis Valenzona, Renda Wiener, Ana Vielma. 10857

XSL•FO
RenderX

Correction: Hybrid Care Modifications in the Delivery of Nonpandemic Care During the COVID-19 Pandemic: Scoping Review (e101728)
Natalia Sanchez Villalobos, Tessa van Loenen, Lem Ngongalah, Máire Connolly, Mart Stein, Chantal Rovers, Aura Timen. 10859

Correction: Integrating Text and Image Analysis: Exploring GPT-4V's Capabilities in Advanced Radiological Applications Across Subspecialties (e91415)
Felix Busch, Tianyu Han, Marcus Makowski, Daniel Truhn, Keno Bressem, Lisa Adams. 10861

Correction: Call for Decision Support for Electrocardiographic Alarm Administration Among Neonatal Intensive Care Unit Staff: Multicenter, Cross-Sectional Survey (e101652)
Xiaoli Tang, Xiaochen Yang, Jiajun Yuan, Jie Yang, Qian Jin, Hanting Zhang, Liebin Zhao, Weiwei Guo. 10863

Correction: Multicriteria Optimization of Language Models for Heart Failure With Preserved Ejection Fraction Symptom Detection in Spanish Electronic Health Records: Comparative Modeling Study (e97621)
Jacinto Mata, Victoria Pachón, Ana Manovel, Manuel Maña, Manuel de la Villa. 10865

Correction: Enhancing LGBTQ+ Inclusivity in an AI-Powered Sexual Health Chatbot: User-Centered Design Approach Through a Nonprofit and Academic Partnership (e96639)
William Liem, Elizabeth Casline, Julianna Lorenzo, Jacob Gordon, Andres Avila, Attia Taylor, Nicole Levitz, Michael O'Keefe, Karen Shum, Kathryn Macapagal. 10867

Correction: Securing Federated Learning With Blockchain in the Medical Field: Systematic Literature Review (e95788)
Xudong Wang, Yi Xie, Xiaoliang Chen, Jiaming Yang, Ruiyuan Li, Weihang Gao, Zineng Yan, Hong Zhou, Zhewei Ye. 10869

Correction: Evidence for Digital Health Tools Designed to Support the Triage of Musculoskeletal Conditions in Primary, Urgent, and Emergency Care Settings: Scoping Review (e92722)
Linda Truong, James Wrightson, Raphaël Vincent, Eunice Lui, Jamon Couch, Ellen Wang, Cobie Starcevich, Dean Giustini, Alex Haagaard, Elena Lopatina, Niels van Berkel, Michael Rathleff, Clare Ardern. 10871

Correction: Acceptability of Health Information Technology by Health Care Professionals: Where We Are Now and How We Can Fill the Gap (e89383)
Corinne Isnard-Bagnis, Stéphane Mouchabac, Riadh Lebib, Hervé Bismut, Pierre Geoffroy. 10873

Correction: Combining Artificial Intelligence and Human Support in Mental Health: Digital Intervention With Comparable Effectiveness to Human-Delivered Care (e88640)
Clare Palmer, Emily Marshall, Edward Millgate, Graham Warren, Michael Ewbank, Elisa Cooper, Samantha Lawes, Alastair Smith, Chris Hutchins-Joss, Jessica Young, Malika Bouazzaoui, Morad Margoum, Sandra Healey, Louise Marshall, Shaun Mehew, Ronan Cummins, Valentin Tablan, Ana Catarino, Andrew Welchman, Andrew Blackwell. 10875

Correction: Effectiveness of a Web-Based Medication Education Course on Pregnant Women's Medication Information Literacy and Decision Self-Efficacy: Randomized Controlled Trial (e91835)
Suya Li, Hui-Jun Chen, Jie Zhou, Yi-Bei Zhouchen, Rong Wang, Jinyi Guo, Sharon Redding, Yan-Qiong Ouyang. 10877

Correction: Culturally Adapted Guided Internet-Based Cognitive Behavioral Therapy for Hong Kong People With Depressive Symptoms: Randomized Controlled Trial (e88495)
Jia-Yan Pan, Jonas Rafi. 10879

Correction: Ingroup Favoritism Surrounding COVID-19 Vaccinations in the Hispanic Communities: Experimental Study (e90652)
Juwon Hwang, Asya Cooley, Skye Cooley, Robert Hinck. 10883

Correction: A Supervised Explainable Machine Learning Model for Perioperative Neurocognitive Disorder in Liver-Transplantation Patients and External Validation on the Medical Information Mart for Intensive Care IV Database: Retrospective Study (e88952)
Zhendong Ding, Linan Zhang, Yihan Zhang, Jing Yang, Yuheng Luo, Mian Ge, Weifeng Yao, Ziqing Hei, Chaojin Chen. 10885

XSL•FO
RenderX

Correction: Infoveillance of COVID-19 Infections in Dentistry Using Platform X: Descriptive Study (e86469)
Alghalia Al-Mansoori, Ola Al Hayk, Sharifa Qassmi, Sarah Aziz, Fatima Haouari, Tawanda Chivese, Faleh Tamimi, Alaa Daud. 10887

Correction: AI-Enabled Wearables for Motor Function Assessment and Rehabilitation in Parkinson Disease: Scoping Review (e97418)
Shengting Li, Siqi Chen, Xiaosong Yu, Huixiang Shang, Tingting Tu, Mingtao Quan. 10889

Correction: Effects of an eHealth Cardiac Exercise Rehabilitation Platform for Patients After Percutaneous Coronary Intervention Based on the Persuasive Systems Design Model: Randomized Controlled Trial (e99326)
Yang Liu, Xiting Huang, Ziying Dai, Zhili Jiang, Wenxiao Wu, Jing Wang, Zhiqian Wang, Luyao Yu, Hanyu Li, Lihua Huang. 10891

Research Letters

Crowdsourcing Medical School Admissions Data: Development and Analysis of the CycleTrack Platform (e83087)
Daniel Amusin, Isabelle Hua, Arthi Kozhumam. 10974

Evaluating the Potential of Reasoning Large Language Models to Perpetuate Racial and Gender Disease Stereotypes in Health Care (e82256)
Joshua Docking, Lee Li, Bradley Menz, Stephen Bacchi, Ashley Hopkins, Michael Sorich. 10980

Large Language Model Versus Multidisciplinary Team: Feasibility Study of Pancreatic Cancer Management Recommendations (e95411)
Zhuoran Liu, Xin Zhao, Tianyang Mao, Heng Zhou, Kangyi Jiang. 10985

Quality of Conventional versus Artificial Intelligence Oral Surgery Consent Forms: Comparative Analysis (e59851)
Jan Gaessler, Bernhard Remschmidt, Ann-Kathrin Jopp, Behrouz Arefnia, Adrian Franke, Marcus Rieder. 10990

National Institutes of Health–Funded Artificial Intelligence and Machine Learning Research, 2019 2023: Cross-Sectional Study (e84861)
Joshua Le, Joseph Morrison, Atul Malhotra, Shamim Nemati, Gabriel Wardi, James Ford. 10994

Transformer-Based Topic Modeling: Characterizing Cannabis Product Adverse Experiences Self-Reported as Requiring Medical Attention on Reddit (e82661)
Tim Mackey, Matthew Nali, Meng Larsen, Zhuoran Li, Cassandra Taylor, Beverly Wolpert, Catharine Trice. 10999

Sex Differences in Medical Specialist Physicians' Electronic Health Record In-Basket Workloads and the Implications for Compensation and Equity: Retrospective Cross-Sectional Study (e79172)
Ashwini Nadkarni, Laura Glick, Molly Becker, Michael Healey, David Stein, Mahyar Heydarpour, Lisa Rotenstein. 11004

Trends in Venture Capital Investment in AI-Driven Biopharmaceutical Startups (e84968)
Abhishek Bazaz, Yunan Ji, Mariana Socal, So-Yeon Kang. 11009

The Impact of Ontario's Virtual Care Payment Model on Cancer Care: A Natural Policy Experiment (e89151)
Adam Suleman, Zhihui Liu, Anna Sinaiko, Anca Prica, Danielle Rodin. 11014

XSL•FO
RenderX

News and Perspectivess

When Old Diseases Return: Cholera, Crisis, and Digital Surveillance in Fragile Settings (e94818)
Isabelle Basbouss-Serhal. 11293

Data Governance Lessons From an Unvalidated Dataset (e94518)
Cliff Dominy. 11296

The Right to Understand in Health Care AI (e95090)
Anshu Ankolekar. 11299

Wearable Air Samplers Reveal How Wildfire Shapes the Exposome (e93193)
Virginia Gewin. 11303

Influencing the Influencers: How Health Experts Are Partnering With Content Creators to Fight Misinformation Online (e93450)
Wendy Glauser. 11306

Is Going Analog Good for Children and Teens' Mental Health and Well-Being? (e94018)
Sara Novak. 11309

Physician, Restore Thyself? The Digital Gap in Physician Well-Being Support (e93338)
Jenny Cato. 11312

Using AI to Assess Potential Zoonotic Threats (e93261)
Cliff Dominy. 11315

Bridging the Gender Gap in Health Care Innovation: The Evolution of FemTech (e103487)
Jenny Cato. 11318

Sorting Science From Marketing in the Era of Data-Driven Biological Aging Clocks (e102951)
Jenna Congdon. 11321

As Social Media Scales Back Fact-Checking, Can Technologies Fill the Gap? (e95730)
Wendy Glauser. 11324

Further Promise and Potential for Precision Medicine in Oncology (e95657)
Shalini Narang. 11327

Emerging Risks of AI-to-AI Interactions in Health Care: Lessons From Moltbook (e96199)
Tejas Athni. 11330

Our AI-Powered Discoveries Are Trapped in a Predigital System (e96018)
Boon-How Chew. 11333

How Digital Diaries Are Changing Communication and Recovery in the Intensive Care Unit and Beyond (e92661)
Jenna Congdon. 11336

Digital Narrative Medicine: Translating Stories Into Medical Education and Better Patient Care (e93037)
Candice Sage. 11339

How Digital Health Technology Is Extending Independent Living (e93767)
Jenna Congdon. 11343

XSL•FO
RenderX

What Health Care Organizations Have Learned From Telecommunication Outages (e91456)
Catharine Solomon. 11346

Hospital-at-Home: New Technology Brings Acute Care to Patients' Homes (e98143)
Jenna Congdon. 11349

The Need for Continued Investment in Digital Pain Assessment (e97777)
Vanessa Nirode. 11352

From Innovation to Infrastructure: Why Digital Behavioral Health Still Struggles to Scale (e97118)
Trevor van Mierlo. 11355

Biocomputing: Beyond the Hype (e100949)
Simon Spichak. 11359

Printing Personalized Medicine: 3D Models Bring Better Surgical Outcomes (e100950)
Jenna Congdon. 11362

Policymakers and Researchers Zero In On the Impact of AI Toys (e102064)
Simon Spichak. 11366

How Machine Learning Can Help Close Evidence Gaps for Drug Safety in Pregnant Women (e101042)
Michelle Falci. 11369

After the Prescription: The Clinical Support Gap in Telehealth-Based GLP-1 Care (e101874)
Anna Zucker. 11373

Can Digital Tools Fix Bias in Mental Health Triage? (e100947)
Beth Rush. 11376

Meet the New Health Care Gatekeeper: Your Wearable (e101881)
Blythe Karow. 11380

Affordable GLP-1? When Digital Platforms Meet Policy Reform (e102069)
Xiangming Zhan. 11383

Remote Patient Monitoring Is Scaling Without a Clear Organizational Model: Insights From the Netherlands (e101887)
Vincent Peters. 11386

From Virtual Molecules to Clinical Trials: How AI Is Reshaping Preclinical Drug Discovery (e101366)
Benedette Cuffari. 11390

US Food and Drug Administration Shifts to AI-Enhanced Regulatory Review With Elsa 4.0 and HALO (e101884)
Tejas Athni. 11393

Can Humanlike Reasoning Be Replicated in Large Language Models for Clinical Decision-Making? (e103526)
Shalini Narang. 11396

How Health Care Workers Can Manage Digital Fatigue (e104196)
Sara Novak. 11399

XSL•FO
RenderX

You Can't Launch This: Trust as Infrastructure in Digital Behavioral Health (e104731)
Trevor van Mierlo. 11402

Can AI in Health Care Be Truly Inclusive? (e104527)
Beth Rush. 11406

The Future of Online Misinformation Research: Tackling the Landscape With Integrity and Urgency (e104526)
Wen-Ying Chou. 11409

Potential of Digital Tools for Chronic Pain Management (e104524)
Vanessa Nirode. 11413

The Ethics of Extending Life: Longevity Medicine and Health Inequity (e104414)
Jenna Congdon. 11416

Written Exposure Therapy as Transformative, Scalable Care for Posttraumatic Stress Disorder (e103501)
Candice Sage. 11420

When Lived Experience Designs the Intervention (e91371)
Trevor van Mierlo. 11423

A Frontline Worker's Take on Hybrid Care Implementation in the Hospital Setting (e90879)
Jenna Congdon. 11426

WHOOP, There It Is: Lessons From WHOOP's FDA Warning Letter (e90882)
Blythe Karow. 11429

UnitedXR Europe 2025: Aligning Health Care Extended Reality (e90727)
Jose Costa. 11433

How Does That Large Language Model Make You Feel? (e105105)
Simon Spichak. 11437

Building a Malaria Intelligence System for Real-Time Prediction and Data-Driven Intervention Planning (e105472)
Sharon Muzaki. 11440

Transformation Versus Innovation in Digital Health Care and the Future of Clinical AI (e105359)
Boon-How Chew. 11444

American Medical Association Shares Framework to Address the Escalating Risk of Physician Deepfakes (e104953)
Shalini Narang. 11447

Identity Theft in Academia: Predatory Journals, AI, and the Rise of False Authorship (e104934)
Cliff Dominy. 11450

Centers for Medicare & Medicaid Services to Launch Landmark ACCESS Program (e105562)
Delaney Rebernik. 11454

Regulating AI Chatbot Impersonations of Medical Professionals (e104835)
Tejas Athni. 11458

How Digital Twins Can Improve Health System Operations (e98113)
Mark Crawford. 11461

XSL•FO
RenderX

CRISPR Diagnostics, in Your Pocket (e98572)
Cliff Dominy. 11464

Big Tech and the Rise of Consumer-Facing Health AI Assistants (e99230)
Tejas Athni. 11467

Keeping Clinical Trials on an Inclusive Track (e99011)
Virginia Gewin. 11471

Immortal AI, Mortal Life: Long-Term Perspectives on AI and Human Knowledge (e98707)
Hyunjin Shim. 11474

AI-Quantified Breast Arterial Calcification Can Predict Heart Disease Risk From Mammograms (e99154)
Shalini Narang. 11477

Retractions

Retraction: Assessing the Causal Association Between COVID-19 and Graves Disease: Mendelian Randomization Study (e101185)
JMIR Publications Editorial Office. 11496

Retraction: Adherence of Female Health Care Workers to the Use a Web-Based Tool for Improving and Modifying Lifestyle: Prospective Target Group Pilot Study (e101897)
JMIR Publications Editorial Office. 11497

XSL•FO
RenderX

Explainable and Interpretable AI for Voice and Speech Analysis in Clinical Care: Systematic Review

Mohamed Ebraheem[1], MSCS; Jamie Toghranegar[2], SLPD; Bridge2AI-Voice Consortium[3,6]; Yael Bensoussan[2*], MD; John Michael Templeton[1,3*], PhD

[1]Bellini College of Artificial Intelligence, Cybersecurity and Computing, University of South Florida, 4202 East Fowler Avenue, Tampa, FL, United States

[2]Department of Otolaryngology Head and Neck Surgery, USF Health Voice Center, University of South Florida, Tampa, FL, United States

[3]Medical Engineering, College of Engineering, University of South Florida, Tampa, FL, United States

6

[*]these authors contributed equally

Corresponding Author:
Mohamed Ebraheem, MSCS
Bellini College of Artificial Intelligence, Cybersecurity and Computing, University of South Florida, 4202 East Fowler Avenue, Tampa, FL, United States

Abstract

Background: Driven by recent advances in artificial intelligence (AI), particularly in medicine, audio-based voice and speech biomarkers are increasingly investigated for various medical applications as a complementary or even alternative modality to traditional medical devices. The adoption of deep learning techniques in recent literature is motivated by their superior performance compared to classical machine learning methods. However, ethical and regulatory concerns regarding the black-box nature of these models have limited their integration into clinical workflows. Consequently, explainable artificial intelligence (XAI) has recently been used to address this issue by generating explanations for opaque model outputs. Ideally, medical XAI systems aim to provide human-understandable, clinically grounded explanations essential for enhanced AI trustworthiness and, thereby, facilitate adoption into real-world clinical settings.

Objective: We conduct a systematic literature review of XAI methods applied for explaining deep learning techniques in audio-based voice and speech clinical applications. We aim to identify what XAI methods have been used to explain the decisions of deep learning voice and speech AI systems in health care, as well as XAI-informed insights. Additionally, we aim to contextualize these findings with respect to clinical applicability and stakeholder relevance. Lastly, we identify opportunities and recommendations for future clinical audio XAI design.

Methods: We used PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). Six electronic databases (IEEE Xplore, ACM Digital Library, Scopus, PubMed, Web of Science, and Nature) were searched for papers published between January 2015 and February 2025. Eligible studies applied explainability or interpretability methods to deep learning models for voice or speech audio in health care contexts. Risk of bias was assessed using PROBAST+AI (Prediction Model Risk of Bias Assessment Tool). The results were thematically synthesized across explainability categories, input representations, clinical domains, validation strategies, and stakeholder considerations.

Results: A total of 30 studies met the inclusion criteria. These studies used a range of explainability approaches, including gradient-based methods, perturbation-based techniques, surrogate model–based methods, model-internal representation analyses, concept-based detectors, and attention-based explanations. Applications spanned diverse clinical domains, including voice disorders, neurodegenerative diseases, psychiatric conditions, and traumatic brain injury. Overall, results indicate that most studies relied primarily on qualitative interpretation of explainability outputs, with limited quantitative validation of explanation consistency across external datasets. Furthermore, none of the included studies explicitly conducted human-in-the-loop evaluations with relevant stakeholders, highlighting a substantial gap in stakeholder alignment.

Conclusions: Current XAI practices in clinical voice and speech analysis are limited by insufficient validation, lack of domain-specific design, and misalignment with clinical stakeholder needs. This review highlights opportunities for developing validated, audio-aware, and stakeholder-centered XAI approaches to support trustworthy clinical deployment. Interpretation of these findings should consider limitations related to single-reviewer study selection, potential high-risk of bias, and the repeated use of benchmark datasets.

(J Med Internet Res 2026;28:e83790) doi:10.2196/83790

XSL•FO
RenderX

KEYWORDS

explainable artificial intelligence; clinical voice analysis; speech biomarkers; deep learning; interpretability; medical decision support; trustworthy AI; artificial intelligence

Introduction

Overview

Voice is a rich modality that has garnered the interest of the research community for its potential in numerous medical applications [1,2]. Voice and speech biomarkers have recently been used for voice pathology detection [3-5], voice quality assessment [6,7], neurodegenerative disease diagnosis [8-13], mental health monitoring [14,15], cardio-respiratory condition classification [16,17], as well as automatic speech recognition (ASR) for disordered speech [18-20]. The main reason behind this interest in audio-based voice and speech biomarkers is the costliness and invasiveness of traditional voice evaluation techniques, such as laryngoscopy, stroboscopy, laryngeal electromyography, and imaging technologies such as MRI (magnetic resonance imaging) and CT (computed tomography) scans, which limit accessibility for many. Alternatively, artificial intelligence (AI)–driven audio-based medical systems pave the way for broader access to medical services for marginalized and underprivileged populations [21].

Yet, AI integration into real clinical settings is limited, in part due to the scarcity of high-quality data needed to train reliable and fair models [22]. Consequently, myriad large-scale projects are underway for the purpose of collecting extensive and representative voice audio datasets. National Institutes of Health–funded Bridge2AI is a multi-institution, large-scale project that aims to collect standardized, AI-ready, and ethically sourced voice data across various health conditions, where voice and speech samples have been collected from 442 participants [23]. Similarly, AphasiaBank (part of TalkBank Project, Carnegie Mellon University) is another National Institutes of Health–funded endeavor that has amassed multimodal data from 306 persons with aphasia [24]. Launched in 2023 and funded through 2028, SpeechDx is a global initiative dedicated to creating an extensive dataset of Alzheimer's speech biomarkers and has recruited about 2000 participants [25]. These efforts demonstrate the general recognition of the potential of voice and speech AI in routine clinical practice.

Recently, the trustworthiness of AI systems has been a central issue for clinical integration of AI, particularly the obscurity of the decision-making processes of deep learning models to end users [22,26-29]. Known as the "black-box" problem, this issue is especially critical in medicine, where decisions directly impact patient safety. Yet, regulatory frameworks have struggled to keep pace with AI's rapid development, leaving unresolved questions of liability and accountability [30]. Clinicians and regulators must understand how models operate, how reliable they are, and under what conditions they fail, before AI systems are integrated into clinical workflows in a meaningful way [31]. Only then can patients be guaranteed safe, high-quality medical care. On one end, there are opinions against the use of black-box models at all for high-stakes environments such as medicine [32]. While white-box, classical machine learning (ML) models such as decision trees or support vector machines trained on interpretable, hand-crafted features offer greater transparency, deep learning models consistently achieve superior performance; this is known as the interpretability-accuracy trade-off [33].

Explainable artificial intelligence (XAI) has emerged to ameliorate this challenge, aiming to make black-box models more transparent [22,29,31,34]. However, many XAI techniques have been developed for image-based domains, such as overlaying saliency maps on brain MRI or chest CT scans. In voice analysis, raw audio is often transformed into time-frequency representations such as spectrograms before being input into deep learning models. Mapping a region on a spectrogram back to an intuitively understandable auditory event (eg, a tremor on a specific syllable) is inherently more complex than identifying a visible tumor on an x-ray. The relationship between spectral features and perceived voice quality or pathology is often highly nonlinear, making clinical interpretation challenging.

Furthermore, the intuitiveness of model explanations varies substantially with respect to end users [35-37]. Computer scientists and ML researchers may be able to interpret technical visualizations such as activation maps or feature attribution plots, whereas clinicians, patients, and regulators have different knowledge bases, priorities, and constraints. A "one-size-fits-all" approach to XAI design is therefore inappropriate, especially when explainability is positioned as a pathway to increasing trust. Thereby, multidisciplinary effort in the design of XAI methods is vital for providing appropriate presentations of explanations suited for the diverse backgrounds and needs of the respective stakeholders [37,38].

In this context, this paper presents a systematic literature review of XAI methods applied to deep learning models for clinical voice and speech analysis. In this review, we use "clinical audio" and "clinical voice and speech" to refer to voice or speech data derived from individuals with a clinical condition used for a health care–relevant application or outcome, regardless of whether the recording occurred in a controlled clinical environment. The review addresses the following research questions:

- What XAI methods have been used to explain the decisions of deep learning voice and speech AI systems in health care?
- What insights are derived from the application of these XAI methods?
- What are the limitations of these XAI methods in the context of clinical audio in terms of clinical applicability and stakeholder relevance?

The remainder of this paper is organized as follows. The rest of the Introduction section presents background for XAI concepts and different perspectives regarding the definitions of explainability and interpretability. We also outline the objectives of XAI and provide a discussion on the broad application of

XSL•FO
RenderX

XAI in diverse medical domains. The Methods section describes the systematic review methodology. The Results section presents explainability approaches, explainability validation strategies, and human-centered evaluation. Then, we discuss the limitations of current approaches, stakeholder alignment, and the impact of audio representation on interpretability, upon which future directions are presented.

Background

Explainability vs Interpretability

There is no clear consensus in the literature on the definitions of explainability and interpretability in the context of AI, and the two terms are often used interchangeably. Nonetheless, various works have attempted to provide distinctions between them.

Linardatos et al [39] highlight the persistent ambiguity surrounding these concepts and review the different ways they have been differentiated in literature. They describe interpretability as relating to the intuition behind a model's outputs, such that a more interpretable model makes it easier to identify causal relationships between inputs and outputs. On the other hand, explainability is concerned with the internal logic and mechanics of the model. They conclude that interpretability is the broader term, and that a model can be interpretable without being explainable.

Gilpin et al [40] approach the distinction differently, defining explainability as a means of answering the questions "why" or "why not" a system behaves in a particular way. In contrast, they describe interpretability as the ability to represent the model's internal processes in a human-understandable form, emphasizing that this is dependent on the knowledge and needs of the target user.

Das and Rad [41] define interpretability as a quality of a system in which its expressions convey human-understandable insights into how it works. They differentiate explanations as additional metadata—produced either by the model or an external algorithm—that clarify the relationship between inputs and outputs.

The National Institute of Standards and Technology also provides guidance on explainability, outlining four principles for explainable systems, the most important of which is that they should deliver accompanying justifications or reasons for model outputs [42]. The work further distinguishes between self-interpretable models, which are inherently understandable to humans, and post hoc explanations, which are generated by an explainability algorithm to provide insight into otherwise opaque models.

In summary, interpretability relates to the inherent transparency of the model itself; white-box models, such as decision trees and linear regression, are generally considered interpretable, while explainability refers to the generation of additional information (ie, explanations) that clarify the reasoning behind a model's decision, regardless of the model's inherent transparency. In this work, we adopt the latter definition.

Objectives of Explainable Artificial Intelligence

Overview

XAI serves multiple partially overlapping objectives, which vary according to the target application and the perspective of the stakeholder [43,44]. The following categories are directly derived from other studies [43,44].

Debugging and Monitoring

Explanations help AI developers identify model errors, biases, and spurious correlations, thereby revealing opportunities for performance improvement. XAI can also be used to monitor performance drift during deployment, ensuring the model continues to operate as intended over time.

Evaluation and Validation

XAI enables stakeholders to assess whether a model is appropriate, reliable, and clinically valid for a given application. It supports both predeployment evaluation and ongoing validation, ensuring that model decisions remain aligned with intended clinical outcomes.

Justification and Transparency

XAI fosters trust among stakeholders by providing context-relevant rationales for model decisions. This includes justifying individual predictions and improving transparency in the decision-making process, allowing clinicians, patients, and regulators to audit and verify the system's outputs.

Improvement and Learning

Explanations support iterative model refinement through collaboration with domain experts, enhancing alignment with clinical reasoning. XAI can also contribute to the discovery of new domain knowledge, such as identifying previously unknown biomarkers.

Governance and Compliance

In high-stakes domains such as medicine, XAI facilitates compliance with ethical, legal, and regulatory requirements, including provisions such as the European General Data Protection Regulation "right to explanation" [45,46]. It enables model auditability, supports liability attribution, and strengthens governance processes.

Characteristics of XAI Methods

Overview

Traditionally, XAI methods are categorized according to their scope of application (ie, model-specific and model-agnostic). Classical taxonomy groups XAI methods into intrinsic methods, where the model contains native interpretable or explainable components, and post hoc techniques, where a secondary system generates explanations. The resultant explanations provide either local insight for individual sample instances or globally justify model behavior. Accordingly, methods in the literature are categorized in the Results section (Table 1).

XSL•FO
RenderX

Table . Summary of XAI[a] methods reported in the included studies, classified according to model dependency (model-agnostic vs model-specific), scope of explanation (local vs global), relationship to the predictive model (post hoc vs intrinsic), and explanation modality. The table shows ablation analysis and latent space interpretation dominating the literature, along with Grad-CAM[b].

XAI methods	Model agnostic/specific	Global/local	Post hoc/intrinsic	Explanation modality	Implemented in
Grad-CAM	Specific	Local	Post hoc	Visual	[3,6,10,47-49]
Guided backpropagation	Specific	Local	Post hoc	Visual	[50]
Saliency maps	Specific	Local	Post hoc	Visual	[4]
Eigen-CAM[c]	Specific	Local	Post hoc	Visual	[10]
SHAP[d]	Agnostic	Local	Post hoc	Tabular	[9,51,52]
GradientSHAP[e]	Specific	Local	Post hoc	Tabular/visual	[53]
LIME[f]	Agnostic	Local	Post hoc	Tabular/visual	[9]
xDMFCC[g]	Agnostic	Local	Post hoc	Visual/tabular	[54]
Ablation studies	Specific	Global	Post hoc	Tabular	[11,12,14,15,18,20,55,56]
Simple attention	Specific	Local	Intrinsic	Visual	[52]
Attention rollout	Specific	Local	Post hoc	Visual	[57]
Concept detectors network	Specific	Global	Intrinsic	Concept-level	[7,58]
Sinc filters	Specific	Global	Intrinsic	Conceptual (filter shape)	[5,53]
Feature map analysis	Specific	Global	Post hoc	Visual (CNN[h] filters)	[13,20]
t-SNE[i]	Agnostic	Global	Post hoc	Visual	[3,5,19,20,49,56,57,59,60]

[a]XAI: explainable artificial intelligence.
[b]Grad-CAM: gradient-weighted class activation mapping.
[c]Eigen-CAM: Eigen class activation mapping.
[d]SHAP: Shapley Additive Explanations.
[e]GradientSHAP: gradient Shapley Additive Explanations.
[f]LIME: local interpretable model-agnostic explanation.
[g]xDMFCC: explainable deep learning mel-frequency cepstral coefficients.
[h]CNN: convolutional neural network.
[i]t-SNE: t-distributed stochastic neighbor embedding.

Model-Specific vs Model-Agnostic

Model-specific methods exploit structural aspects particular to model type or architecture to, in theory, produce higher-quality explanations. For example, Grad-CAM (gradient-weighted class activation mapping) is applicable to convolutional neural networks (CNNs) and takes advantage of their spatial feature maps to highlight input features that are most influential for the output.

In contrast, model-agnostic methods, such as SHAP (Shapley Additive Explanations), are generalizable to any ML model regardless of architecture. This universality enables versatile implementation across many applications. Unlike model-specific techniques that benefit from internal model architecture, model-agnostic methods often exhibit lower fidelity and efficiency, especially when dealing with high-dimensional data.

Local vs Global

Local explainability methods aim to clarify the reasoning behind a model's prediction for a single input instance. These explanations are instance-specific and do not describe the model's overall decision-making process. For example, visualization techniques such as Class Activation Maps can generate heatmaps that indicate the regions of an input spectrogram most relevant to the model's prediction. Local explanations are particularly valuable in clinical decision support scenarios where individual case justification is critical.

Global explainability methods, on the other hand, aim to describe the model's general behavior across the entire dataset. This can involve identifying the most influential features for distinguishing between classes, mapping decision boundaries, or summarizing feature interactions. Such methods are crucial for understanding systematic model biases, validating clinical relevance, and ensuring that the model's logic aligns with domain knowledge.

XSL•FO
RenderX

Intrinsic vs Post Hoc

As mentioned earlier, intrinsic (or self-) interpretability refers to the degree to which a model's decision-making process is transparent and human-understandable by design. White-box models are examples of intrinsically interpretable models. However, they often underperform compared to more complex black-box architectures such as deep neural networks.

Post hoc methods, in contrast, are applied after the training process is completed to explain the model's behavior without altering its internal mechanics. These techniques, ranging from saliency maps to perturbation-based analyses, are particularly prevalent for explaining black-box models.

XAI in Medicine

With recent advances in medical AI, the implementation of XAI has become increasingly critical to ensure that AI models are reliable, ethical, legally compliant, and clinically aligned, particularly in high-stakes environments such as health care. The literature contains numerous surveys discussing the role of XAI across diverse medical applications.

Several recent reviews have broadly examined the adoption of XAI in medicine, emphasizing its potential to improve transparency, foster trust, and facilitate regulatory compliance [61-65]. These works cover a variety of data modalities, including medical imaging, electronic health records (EHR), genomics, and time-series analysis. Commonly reported methods include model-agnostic techniques, most notably LIME (local interpretable model-agnostic explanations) and SHAP, alongside post hoc visualization approaches. Nonetheless, these reviews consistently highlight persistent challenges relating to faithfulness, stability, and standardized evaluation of explanations.

Domain-specific surveys further illustrate these trends. For example, van der Velden et al [66] reviewed over 200 studies applying XAI to medical imaging, noting the predominance of visualization techniques, followed by perturbation-based, textual, and example-based approaches. While visual explanations have been extensively validated, the authors stress the need for equivalent validation of textual and example-based methods. Muhammad and Bendechache [67] similarly highlight the interpretive ambiguity and perturbation sensitivity of visual methods. In the context of medical time-series data, Caterson et al [68] present a scoping review of XAI for EHR, finding feature attribution methods to be the most widely used. Salih et al [69] review XAI in cardiology for ECG (electrocardiogram) and EHR, reporting the frequent use of SHAP and Grad-CAM, followed by LIME, but also note that nearly half (47%) of the studies did not use any formal evaluation of the explanations produced. In the mental health and psychiatry domain, Joyce et al [70] reviewed XAI applied to neuroimaging, interview transcripts, and physiological data, underscoring the importance of human-centered design for producing clinically useful explanations beyond raw saliency maps. With respect to clinical audio, Chen et al [16] reviewed XAI approaches for vocal biomarker-based lung disease detection, discussing several issues, including the utility of explanations defined in terms of informativeness and user understanding as being an important criterion for evaluating explanations.

Despite the breadth of these surveys, to the best of our knowledge, no systematic reviews have comprehensively examined the use of XAI for clinical voice and speech biomarkers more broadly. This gap highlights the need for a dedicated synthesis in this domain, particularly given the unique challenges and interpretive demands of audio-based clinical decision support systems.

Unique Challenges of Clinical Audio for XAI

Audio Abstraction and Representation

In the time domain, audio is represented as waveforms that fully capture the acoustic signal but are difficult for humans to visually interpret directly. Consequently, audio is often transformed into spectrograms or mel-frequency cepstral coefficients (MFCCs). While clinicians such as speech-language pathologists and audiologists are trained to interpret spectrograms, visual explanations based on time-frequency representations remain suboptimal, particularly for other stakeholder groups. This issue is further highlighted when explanations are inconsistent across samples or vary substantially between different explainability methods, undermining their reliability and interpretive stability [71,72]. This can be attributed to human auditory perception of sound, rather than of the visual [73]. MFCCs, a compact nonlinear "spectrum-of-the-spectrum" representation of audio, are even more obscure to clinicians, limiting their clinical interpretability [47,74]. Additionally, speaker characteristics are usually spread across multiple frequency bands, making the problem of localizing relevant frequency information too broad to solve effectively with traditional vision-based XAI methods [75]. Nonetheless, assuming relevant spatiotemporal regions are identifiable through traditional vision-based methods, understanding why these regions are important remains obscure, at least partially, and requires further analysis to answer such questions. Ultimately, the nonvisual perceptual nature of voice and speech makes current visual explanations a nonideal solution to the overarching problem of explainability and trustworthy clinical AI.

Temporal Dynamics

Unlike medical images, where both axes carry spatial meaning, audio has an inherently temporal structure. Identifying when a clinically relevant event occurs (eg, a stutter or pause) is as important as identifying which acoustic features are involved. This mismatch makes it difficult for visualization-based methods such as saliency maps or Grad-CAM, designed for spatial data, to yield clinically actionable explanations in audio [73,75].

Annotation Scarcity

Phonetic, prosodic, disfluency, and voice quality annotations are essential for aligning model explanations with known biomarkers; however, such granular annotations, typically performed by multiple trained experts, are resource-intensive and, consequently, scarce. Without such annotations, validation of model explanations' alignment to medically grounded features becomes increasingly daunting, limiting their utility in practice.

XSL•FO
RenderX

In summary, these challenges highlight that explainability methods developed for domains such as imaging or EHR are not directly transferable to clinical audio. Audio-native evaluation frameworks and human-centered approaches are needed to ensure that XAI methods drive actionable clinical insights.

Methods

Systematic Review Search Strategy

We conducted our systematic review according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guideline. A review protocol was prepared, but not registered before this study. Our goal was to survey papers that perform explainability or interpretability analyses of deep learning models applied to clinical voice and speech audio. A comprehensive search was carried out across IEEE Xplore, ACM, Scopus, PubMed, Web of Science, and Nature.

The following search terms were selected across three domains: explainability and interpretability, voice, speech, and acoustics, and clinical and health care context. These terms were used to create a query, which was adapted for each database. The exact search strings for each database can be found in Multimedia Appendix 1. The search was conducted across all databases on February 7, 2025, except ACM, which was searched on February 11, 2025. The search included studies published between January 1, 2015, and the date of the search. In total, 1426 records were retrieved, which were reduced to 1348 after removing 78 duplicates.

Eligibility Criteria

Papers were selected if they met the following criteria of being (1) focused on deep learning models applied to voice and/or speech data for health care applications, (2) applied explainability or interpretability techniques to the model, (3) reported empirical results or experimental validation, and (4) published in peer-reviewed journals or conferences.

Papers were excluded for any of the following reasons: (1) explainability and/or interpretability analysis performed for non–deep learning models; (2) used audio transcripts only (ie, purely natural language processing methods); (3) used nonclinical audio datasets; (4) not published in English; and (5) reviews, theses, or non–peer-reviewed papers.

Study Selection

The results of the search strategy were exported to the RAYYAN tool [76] for screening. One reviewer conducted the initial screening of titles and abstracts against the eligibility criteria, and a second reviewer independently verified the filtering decisions. Papers were excluded only if there was clear evidence in the title or abstract that the paper did not meet the criteria; otherwise, they proceeded to full-text review. The screening process resulted in 187 papers deferred for full-text review, of which 2 papers were retracted, and 2 papers were inaccessible. The final set for full-text review was 183 papers.

Full-text screening was performed by two reviewers, with deliberation and discussion used to resolve disagreements. Reasons for exclusion at this stage were documented. The final included set consisted of 30 papers that met all eligibility criteria.

Data Extraction

Data extraction was carried out by a single reviewer using a standardized spreadsheet. A second reviewer reviewed the extracted data and deliberated on any uncertainties or discrepancies to ensure accuracy and completeness. The following information was extracted from the final set of included papers: (1) bibliographic information (authors, and year), (2) dataset information (clinical condition, number of subjects, and acoustic tasks), (3) clinical application, (4) deep learning methodology (model, hyperparameters, and training or validation strategy), (5) model performance, (6) explainability or interpretability strategy, (7) insights gained from explainability or interpretability analysis, and (8) validation or support for the explainability or interpretability results.

Data Synthesis

Overview

Data were synthesized across explainability method type, explainability input-output characteristics, explainability validation strategies, and stakeholder involvement.

Explainability Methods

As existing surveys of XAI demonstrate a lack of consensus on a unified XAI taxonomy, we adopted a set of commonly used XAI method categories reported across prior surveys and taxonomy literature, including work in medical XAI [77-79]. Our work does not aim to propose a new categorization or taxonomy, as this is outside the scope of this study. Accordingly, explainability methods in the included studies were grouped using the following widely adopted categories:

- Gradient saliency-based methods: these methods derive explanations by analyzing gradients of the model output with respect to the input or intermediate feature maps, characterizing the influence of different parts of the input signal on model predictions.
- Perturbation-based techniques: these methods generate explanations by systematically modifying, masking, or removing parts of the input or feature space and observing the resulting change in model output.
- Surrogate model-based methods: this category includes methods that use simpler models to approximate the local behavior of complex models.
- Model-internal representation analysis: this category encompasses methods that provide explainability through inspection and/or visualization of a model's internal structures.
- Concept-based methods: these methods include techniques that provide explanations by relating model behavior to predefined, higher-level, semantically human-understandable concepts or clinically meaningful attributes.
- Attention-based explanations: this family of methods encompasses techniques that rely on attention weights to generate explanations.

XSL•FO
RenderX

Explainability Input-Output Representation

For each included study, we recorded the input representation over which explainability was applied and the corresponding form of the explainability output, as these factors directly determine the interpretability and modality of explanations in clinical audio systems. Input representation refers to the signal or feature representation interrogated by the explainability method (which may differ from that used by the predictive model), while output characteristics refer to how explanations were presented (eg, explanation modality). These attributes were extracted to support the synthesis of explainability practices across studies and to contextualize differences in explanation form, granularity, and clinical interpretability.

Explainability Validation Strategies

In this work, an explainability validation strategy is defined as any technical, model-centric procedure used to assess the faithfulness, consistency, or robustness of explainability outputs with respect to the underlying model behavior. This definition explicitly refers to model-centered validation and excludes human judgment or interpretive assessment.

The following explainability validation strategies were recorded for the included studies:

- Perturbation-based validation: input features, time segments, or frequency regions identified as salient by the explainability method are systematically modified, masked, or removed to assess the resulting impact on model predictions or performance.
- Ground-truth or annotation-based verification: salient regions or features identified by the explainability method are compared against externally defined references, such as expert annotations, labeled phonetic events, disorder-specific acoustic markers, or task-related temporal boundaries, when used to evaluate correspondence with model behavior.
- Stability or cross-dataset consistency analysis: explanation patterns are examined across different datasets, cohorts, recording conditions, or evaluation splits to assess the robustness and consistency of explanations under data variation or dataset shift.

Domain-Specific Explanation Patterns

To support structured synthesis of explainability findings, study-derived insights were grouped and analyzed according to the clinical application domain targeted by each study (as defined by each study's objective). Explainability outputs were examined within each domain to identify recurring explanation patterns. We also noted potential risk of circular validation, particularly in cases where multiple studies relied on the same dataset, which may limit the generalizability of observed explanation trends.

Human-Centered Analysis and Stakeholder Alignment

Beyond technical explainability validation, human-centered evaluation is critical for assessing explanation quality, the clinical relevance, and practical utility of explainability methods in real-world health care settings. Accordingly, stakeholder alignment was assessed separately from technical validation. Stakeholder alignment was defined as explicit involvement of domain experts, such as clinicians, speech-language pathologists, or regulatory stakeholders, in interpreting, evaluating, or providing feedback on explainability outputs. For each included study, the presence or absence of human-in-the-loop evaluation was recorded, along with the reported form of involvement.

Risk of Bias and Applicability Assessment

A formal quality and risk-of-bias assessment was conducted to evaluate the methodological robustness of prediction models estimating health-related outcomes. PROBAST+AI (Prediction Model Risk of Bias Assessment Tool for Artificial Intelligence) [80], an extension of the original PROBAST [81] framework developed to address advances in AI- and ML-based prediction models, was used for this purpose.

A review-level PICOTS (Population; Index Model; Comparator; Outcome; Timing; Setting) framework was predefined to guide the scope and intended applicability of the assessment. The tool was applied to studies that developed or evaluated prediction models estimating health-related outcomes. Studies that did not fit these criteria were not suited for quality assessment by the PROBAST+AI tool. Each eligible study was classified as model development, model evaluation, or both. In accordance with PROBAST+AI guidance, development and evaluation components were assessed separately.

The tool evaluates four domains: (1) participants and data sources: addressing data origin, collection procedures, and representativeness; (2) predictors: examining input definition, preprocessing, and availability at the time of intended use; (3) outcome: assessing outcome definition, measurement, and timing; and (4) analysis: evaluating sample size adequacy, validation strategy, handling of missing data, risk of data leakage, and mitigation of overfitting.

Domains were rated as low, high, or unclear for development (quality concern) and evaluation (risk of bias), respectively. Applicability was assessed for the first three domains only and not for the analysis domain. The fourth domain does not include applicability considerations under the PROBAST+AI framework, as applicability refers to the assessor's review question or intended use of a model, including the target population and setting. In accordance with PROBAST+AI guidance, if at least one domain was rated high, the overall judgment was classified as high. Detailed signaling criteria are described in Multimedia Appendix 2 [3-6,9-15,47-57,59,60,82].

The assessment was independently conducted by two reviewers, with disagreements resolved by discussion. The PICOTS framework definition and domain-specific ratings are provided in Multimedia Appendix 2.

Results

Overview of Included Studies

A total of 30 studies met the inclusion and exclusion criteria. Although the search spanned publications from 2015 to 2025, all included studies were published from 2020 onward, except for a single paper. This can be explained by the relatively recent adoption of deep learning-based approaches and XAI for voice

XSL•FO
RenderX

and speech in health care research. A PRISMA flow diagram summarizing the selection process is shown in Figure 1. The results are summarized in Table 2.

Figure 1. PRISMA flowchart. DL: deep learning; ML: machine learning; NLP: natural language processing; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; XAI: explainable artificial intelligence.

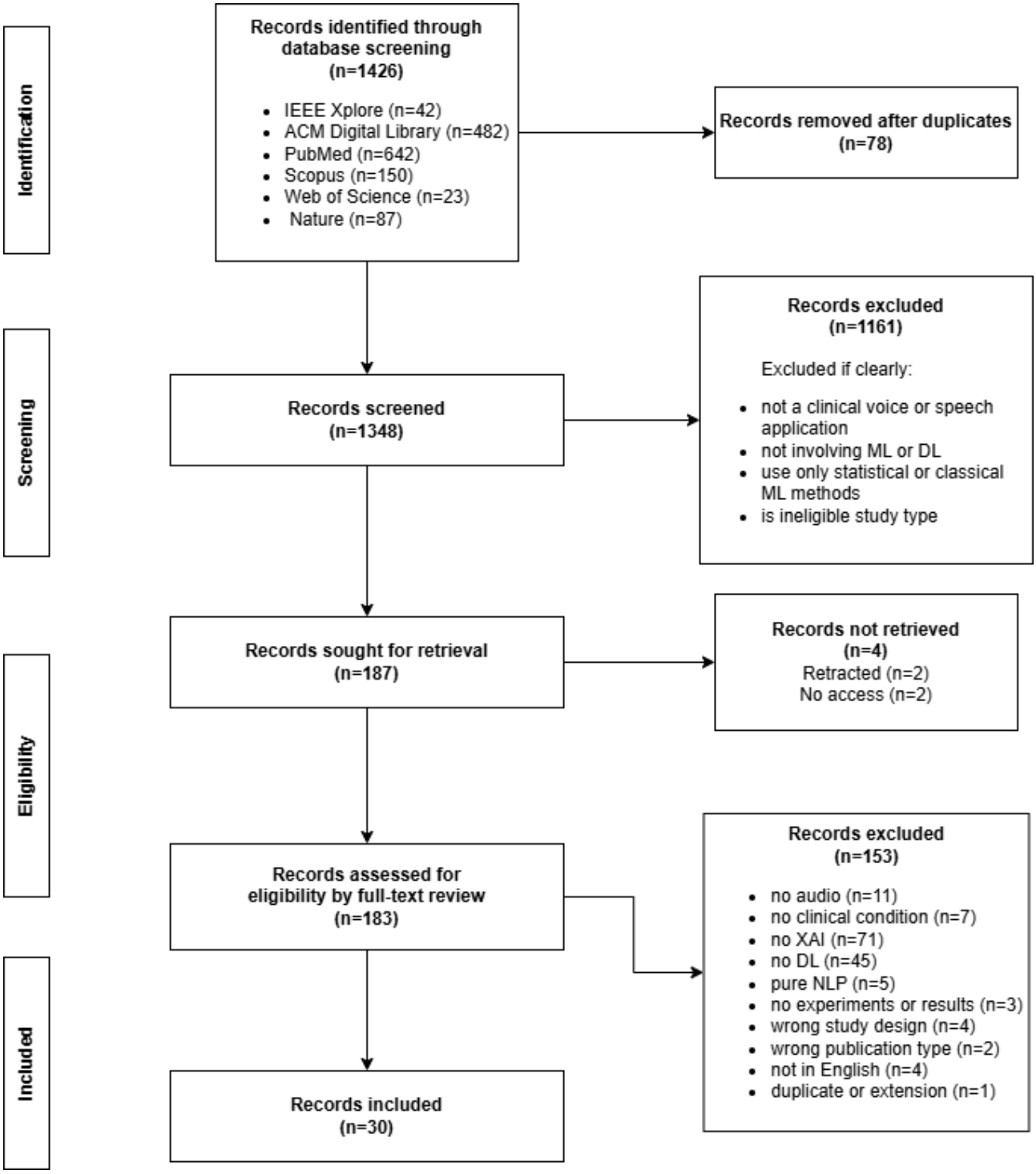

XSL•FO
RenderX

Table . Overview of the 30 included studies applying explainable and interpretable deep learning methods to clinical voice and speech analysis. The table summarizes clinical application domains, datasets, modeling approaches, and explainability techniques. Across studies, post hoc, local explainability methods predominated, with latent representation analysis, most commonly t-SNE[a], being the most frequently used approach. Overall, explanations were primarily interpreted qualitatively, with limited quantitative validation, external consistency assessment, or human-centered evaluation reported.

Work	Application	Model	Model performance	XAI[b] method	XAI insight	Quantitative XAI output validation
Shaikh et al [4]	Voice disorder classification	MLP[c], 1D-CNN[d]	97.1% accuracy, 99.8% recall, 97% F1	Saliency maps	Top five LLD[e] features were extracted per condition; saliency maps emphasized low amplitude, high-frequency spectrogram regions, though no consistent pattern emerged.	Cross-dataset evaluation
Gupta et al [50]	Multiclass dysarthria severity classification	ResNet-14[f]	98.8% accuracy, 98.9% F1	Guided backpropagation	For low-severity cases, the model focused on high-energy vowel regions (clearer phoneme boundaries); for high-severity cases, activations were diffuse, consistent with phonatory instability and temporal smearing.	—[g]
Fu et al [48]	Schizophrenia vs healthy (binary)	Sch-Net (CNN[h] with skip connections, CBAM[i])	97.68% accuracy, 99.1% recall, 97.7% F1	Grad-CAM[j]	Reduced high-frequency energy focus (<5 kHz) and emphasis on low-frequency formant stripes (<2 kHz), suggesting articulation errors (voiced instead of unvoiced consonants) aligned with blunted affect.	—
Lee et al [6]	Postoperative vocal recovery (GRBAS[k])	EfficientNet-B4 (CNN)+ LSTM[l]	0.379 RMSE[m] (regression), 91.8% AUC[n] (breathiness, binary)	Grad-CAM	Different activation patterns by GRBAS level: attention to low bands (0 - 2 kHz, formants), mid bands (2 - 4 kHz, harmonics/noise), and temporal regions (pauses/breathy segments).	—
Peng et al [3]	Multiclass voice disorder classification	OpenL3+ SVM[o]	99.5% accuracy, 99.6% recall, 99.6% F1	Grad-CAM, t-SNE	Disorder-specific band focus: healthy in low-frequency regions; hyperkinetic dysphonia in both high and low bands; reflux laryngitis in high bands; hypokinetic dysphonia weak across bands.	—

XSL•FO
RenderX

Work	Application	Model	Model performance	XAI[b] method	XAI insight	Quantitative XAI output validation
Rojas et al [47]	Mild traumatic brain injury detection	ResNet	67.4% accuracy, 76.1% recall, 69.9% F1	Grad-CAM	Model down-weighted low frequencies and emphasized high-frequency regions for mTBI[p] predictions.	—
Shen and Zhang [49]	Speech disfluency detection	Multiadversarial neural network	58.7% UAR[q]	Time-related Grad-CAM, t-SNE	Highlighted frames aligned with annotated disfluencies; distinct disfluency types exhibited different temporal Grad-CAM patterns; results supported wav2vec capturing meaningful temporal cues.	—
Jeong et al [10]	Parkinson disease classification	AST[r], EfficientNet-based	92.15% accuracy, 91.5% recall, 92.15% F1	Eigen-CAM[s]	Emphasis on higher-frequency bands associated with muffled/degraded speech in PD[t].	Annotation-based verification
Schultebraucks et al [51]	PTSD[u]/major depressive disorder (binary and severity) at 1 month	Multimodal DNN[v]	90% AUC, 84% recall, 83 F1 (PTSD), 86% AUC, 82% recall, 82% F1 (depression)	SHAP[w]	Key predictors: reduced pitch/intensity (voice), negative affect/self-focus (language), flat affect (face).	—
Ditthapron et al [53]	TBI[x] vs healthy (binary)	pSinc+ cGRU[y]	83.8% balanced accuracy, 92.9% recall, 85.1% F1	GradientSHAP[z], Sinc filters	High attribution to filler words and high-frequency spectral patterns; formants were salient for TBI detection.	—
Zhang et al [9]	Dementia detection	BiLSTM[aa]+ multi-head attention (audio); Distil-BERT[ab]+1D-CNN+ cross-modal attention	80.8% accuracy, 77.57% recall, 83.23% F1 (external validation)	SHAP (LLDs), LIME[ac] (text/audio)	Linguistic features (eg, noun phrase rate, word rate) were most predictive among LLDs; AD[ad] speech showed more fillers/pronouns/function words, lower energy, and slower rate; attention emphasized disfluent/low-energy segments.	—
Gutiérrez-Serafín et al [54]	Brain lesion detection	CNN	73% accuracy, 75% recall, 75% F1	xDMFCCs[ae] (LIME)	MFCC[af]-1/2 (energy/clarity) were most important; controls showed earlier/clearer articulation with more discriminative energy in higher-order MFCCs; patients showed delayed onset, slower articulation, and longer phoneme duration.	—

XSL•FO
RenderX

Work	Application	Model	Model performance	XAI[b] method	XAI insight	Quantitative XAI output validation
Liu et al [18]	Automatic dysarthric speech recognition	TDNN-HMM[ag]; CTC[ah]; LAS[ai], encoder–decoder	25.5% WER[aj]	Ablation studies	TDNN[ak] outperformed CTC on moderate-severe dysarthria (highlighting the importance of temporal dependencies); speaker adaptation substantially reduced WER (per-speaker customization is beneficial).	—
Huang et al [14]	Schizophrenia severity (classification/regression)	Transformer embeddings (BERT[al], ELECTRA[am], TERA[an])+ BiLSTM + FC[ao]	88% accuracy, 80% F1 (severity classification)	Ablation studies	BERT is most crucial for the TLC[ap] scale and for PANSSs[aq] except PANSS-General; ELECTRA contributed moderately; TERA was low for TLC but important for PANSS, especially positive symptoms and general psychopathology.	—
Herath et al [82]	Aphasia severity classification	DNN	98.5% accuracy, 97.3% recall, 97.4% F1	Ablation studies	MFCC-DNN performed best; ZCR-DNN[ar] performed worst.	—
He et al [55]	Schizophrenia detection	WNSA-Net[as]	98.16% accuracy, 98.72% F1 (TORGO[at])	Ablation studies	Wideband and narrowband spectrograms provided complementary information; dilated convolutions captured micro-level (pitch, formants) and macro-level (rate, prosody) cues.	Cross-dataset evaluation
Lahoti et al [11]	Parkinson disease detection	Multihead attention BiLSTM	85.02% accuracy, 84.9% F1	Ablation studies	Augmenting cepstral features with shifted delta cepstra improved performance over single-frequency filtering cepstral coefficients alone, highlighting long-term temporal dependencies for PD detection.	—

XSL•FO
RenderX

Work	Application	Model	Model performance	XAI[b] method	XAI insight	Quantitative XAI output validation
Zhang et al [15]	Depression detection	Wav2vec+1D-CNN+ LSTM	90.9% accuracy, 90.7% F1, 95.6% AUC (binary)	Ablation studies	Models with self-attention performed better; wav2vec embeddings were superior; 7-second segments worked best, suggesting emotion is concentrated in short spans.	—
Laguarta and Subirana [12]	Alzheimer disease detection	Open voice brain model (GNN[au])	93.3% accuracy, 95% AUC	Ablation studies	Memory/fluency features dominated early-stage AD detection, with sentiment/prosody also contributing; AD often showed high saliency in respiratory control, disfluency, or memory-related patterns.	—
Joshy and Rajan [56]	Dysarthria severity classification	DNN, CNN, GRU[av]	93.97% accuracy (speaker dependent), 70.52% accuracy (speaker independent)	Ablation studies; t-SNE	MFCCs performed best in a speaker-dependent setup; CQCCs[aw] generalized better to unseen speakers; articulatory features were strongest among disorder-specific sets but weaker in a speaker-independent setup; MFCC-based i-vectors showed clearer class clustering in t-SNE.	Multidataset evaluation
Yue et al [20]	Automatic dysarthric speech recognition	Multistream CNN + LiGRU[ax]	30.3% WER (dysarthric), 11% WER (typical)	Ablation; CNN filter analysis; t-SNE	Best WER resulted from combining spectrogram magnitude with vocal tract and excitation streams; speed perturbation without F0 fixing improved WER; filters fed with vocal-tract signals emphasized low quefrencies, whereas excitation filters suppressed them; t-SNE showed progressive dysarthric/typical separation and reduced gender clustering over training.	—

XSL•FO
RenderX

Work	Application	Model	Model performance	XAI[b] method	XAI insight	Quantitative XAI output validation
Wang et al [52]	Auditory verbal hallucination detection	Uni-modal Bi-GRU[ay]; multi-modal self-attention DNN	84% F1 (overall), 78% F1 (audio, text)	Simple attention visualization, SHAP	Attention prioritized clauses describing distress, influence, or interference, aligning higher weights with higher auditory verbal hallucination severity.	—
Lau et al [57]	Voice disorder detection	AST	81.9% UAR, 91.1% AUC	t-SNE, attention rollout	Model focused on specific phonemes (eg, / / and the segment "/e/ /s/ /i/ /n/") rather than merely high-energy regions.	—
Abderrazek et al [58]	Head and neck cancer intelligibility	CNN	0.91 PCC[az]	Concept detector network	No neurons detected phonetic features in the first dense layer; from subsequent layers to output, the number of phonetic feature detectors increased by a factor of 1.75.	Cross-dataset evaluation
Mathad et al [7]	Hypernasality assessment (children with cleft palate)	DNN	0.797 PCC	Concept detector network	A DNN nasality model estimated posterior probabilities for nasal consonants, oral consonants, nasalized vowels, and oral vowels; these were combined into an objective hypernasality measure that quantified detected nasality against expected nasality per phrase.	Cross-dataset evaluation
Hung et al [5]	Voice disorder classification	SincNet (CNN-based)	83.3% accuracy, 77.31% UAR	Sinc filter analysis, t-SNE	SincNet filters emphasized F1/F2 more clearly than standard CNN filters, preserving formant structure and energy in 500 - 3000 Hz bands.	—
Vasquez-Correa et al [13]	Parkinson disease assessment	CNN	97.6% accuracy, 98.7% AUC (multimodal)	Feature map analysis	Feature maps showed filters highlighting speech transitions (syllable onsets/offsets); many filters in layers 2 and 4 differentiated PD from healthy controls.	—

XSL•FO
RenderX

Work	Application	Model	Model performance	XAI[b] method	XAI insight	Quantitative XAI output validation
Lee et al [60]	ASD[ba] detection in infants	BiLSTM with autoencoder	68.18% accuracy, 65.1% UAR, 54.57% F1	t-SNE	Autoencoder embeddings yielded clearer ASD vs TD[bb] separation than eGeMAPS[bc] with BiLSTM.	—
Geng et al [19]	Automatic dysarthric speech recognition	TDNN; conformer	25.5% WER	t-SNE	SVD[bd]-based spectrotemporal deep embeddings showed better separation of dysarthric vs typical speech than i-vectors/x-vectors.	Cross-language, multidataset evaluation

XSL•FO
RenderX

Work	Application	Model	Model performance	XAI[b] method	XAI insight	Quantitative XAI output validation
Kim et al [59]	Laryngeal disease classification	ResNet-50	92.15% accuracy, 91.53% recall, 92.15% F1	t-SNE	Pooled CNN features for benign disease overlapped with cancer and vocal cord paralysis, explaining reduced multiclass performance.	—

[a]t-SNE: t-distributed stochastic neighbor embedding.
[b]XAI: explainable artificial intelligence.
[c]MLP: multilayer perceptron.
[d]1D-CNN: one-dimensional convolutional neural network.
[e]LLD: low-level descriptor.
[f]ResNet-14: residual network.
[g]Not available.
[h]CNN: convolutional neural network.
[i]CBAM: convolutional block attention module.
[j]Grad-CAM: gradient-weighted class activation mapping.
[k]GRBAS: Grade, Roughness, Breathiness, Asthenia, Strain.
[l]LSTM: long short-term memory.
[m]RMSE: root mean squared error.
[n]AUC: area under the curve.
[o]SVM: support vector machine.
[p]mTBI: mild traumatic brain injury.
[q]UAR: unweighted average recall.
[r]AST: audio spectrogram transformer.
[s]Eigen-CAM: Eigen class activation mapping.
[t]PD: Parkinson disease.
[u]PTSD: posttraumatic stress disorder.
[v]DNN: deep neural network.
[w]SHAP: Shapley Additive Explanations.
[x]TBI: traumatic brain injury.
[y]cGRU: cascading gated recurrent unit.
[z]GradientSHAP: gradient Shapley Additive Explanations.
[aa]BiLSTM: bidirectional long short-term memory.
[ab]DistilBERT: distilled version of Bidirectional Encoder Representations from Transformers.
[ac]LIME: local interpretable model-agnostic explanations.
[ad]AD: Alzheimer disease.
[ae]xDMFCC: explainable deep learning mel-frequency cepstral coefficients.
[af]MFCC: mel-frequency cepstral coefficient.
[ag]TDNN-HMM: time-delay neural network—hidden Markov model
[ah]CTC: connectionist temporal classification.
[ai]LAS: Listen, Attend, and Spell model architecture.
[aj]WER: word error rate.
[ak]TDNN: time delay neural network.
[al]BERT: Bidirectional Encoder Representations From Transformer.
[am]ELECTRA: Efficiently Learning an Encoder That Classifies Token Replacements Accurately.
[an]TERA: Transformer Encoder Representations From Alteration.
[ao]FC: fully connected layer.
[ap]TLC: Thought, Language, and Communication.
[aq]PNASS: Positive and Negative Syndrome Scale.
[ar]ZCR-DNN: deep neural network with zero-crossing rate features as input.
[as]WNSA-Net: axial-attention-based network using wideband and narrowband spectrograms.

XSL•FO
RenderX

[at]TORGO: database of acoustic and articulatory speech from speakers with dysarthria (University of Toronto).

[au]GNN: graph neural network.

[av]GRU: gated recurrent unit.

[aw]CQCC: constant Q cepstral coefficients.

[ax]LiGRU: light gated recurrent unit.

[ay]BiGRU: bidirectional gated recurrent unit.

[az]PCC: Pearson correlation coefficient.

[ba]ASD: autism spectrum disorder.

[bb]TD: typically developing.

[bc]eGeMAPS: extended Geneva Minimalistic Acoustic Parameter Set.

[bd]SVD: Saarbrücken Voice Database.

The included studies applied explainability methods across a broad range of voice- and speech-related health domains, including voice and structural laryngeal pathology [3-6,58,59], Parkinson disease (PD) [10,11,13], dysarthria and automatic dysarthric speech recognition [18-20,50,51], dementia [9] and Alzheimer disease (AD) detection [12], psychiatric and mental health conditions [15,48,52,55], traumatic brain injury (TBI) and focal brain lesions [47,53,54], aphasia [82], cleft palate-related hypernasality assessment [7], head and neck cancer–related intelligibility assessment [58], and autism spectrum disorder [60].

Datasets used across the included studies varied in quality, scale, language, and provenance. Most datasets were collected in controlled clinical environments, with only a single study [52] relying on remotely collected speech data. Dataset sizes ranged from approximately 15 participants to several hundred. Table 3 reports dataset sizes and tasks reported in their respective data collection protocols, not the studies included in this literature review. The datasets used represent a range of linguistic backgrounds, with English, Chinese, and Korean datasets being the most prevalent. Publicly available benchmark datasets were predominantly used in studies focusing on neurodegenerative diseases and motor speech disorders, whereas studies targeting psychiatric conditions relied exclusively on institution-specific or private datasets (eg, TORGO, UASpeech (Universal Access Speech), and PC-GITA [Parkinson Corpus – Grupo de Investigación en Telecomunicaciones Aplicadas]). Speech elicitation tasks varied across studies and included sustained vowel phonation, diadochokinesis, read speech, picture description, and free or spontaneous speech.

XSL•FO
RenderX

Table . Overview of speech and voice datasets used in the reviewed studies, including clinical domain, language, participant population, elicited speech tasks, and study usage. The datasets encompass a wide range of domains, sizes, and languages. UASpeech[a] was the most frequently used publicly available benchmark, while other studies relied on their own data collection protocols.

Dataset name	Application/domain	Language	Population	Tasks	Used by
ADReSS[b] (DementiaBank) [83]	Alzheimer disease	English	156 subjects (78 AD[c], 78 controls)	Cookie theft picture	[9,12]
Pitt Corpus (DementiaBank) [84]	Alzheimer disease	English	500 (253 AD, 247 controls)	Cookie theft picture, word fluency task, spontaneous interviews	[9]
AVH[d] Voice Diaries Dataset [52]	Auditory verbal hallucinations	English	384 participants	30-day audio diary recordings	[52]
SNUBH[e] Infant Dataset [60]	Autism	Korean	39 infants (10 ASD[f], 29 TD[g])	Clinical vocalizations during ASD assessment	[60]
LANNA[h] Speech Corpus [85]	Specific language impairment	Czech	188 children (118 SLI[i], 70 controls)	Vowels, consonants, syllables, words, sentences, picture description	[48]
Max-Planck Brain Lesion Dataset [54]	Brain lesions	Dutch	16 patients with lesions, 16 controls	Emotion-elicited word production	[54]
Americleft Database [86]	Cleft palate	English	60 children with CP[j], 10 controls	Sentences	[7]
NMCPC[k] Database [87]	Cleft palate	English	32 children with CP, 9 controls	Sentences	[7]
JCCOCC[l] MoCA[m] Cantonese Speech corpus [88]	Cognitive impairment	Cantonese	469 speakers	Cognitive assessment interviews	[19]
Bellevue Trauma Dataset [89]	Depression or PTSD[n]	English, Spanish, and Mandarin	377 (first round) 221 (second round)	Clinical interviews	[51]
CMDC[o] [90]	Depression	Cantonese	78 speakers	Structured interviews (audio, video, text)	[15]
DAIC-WOZ[p] [91]	Depression, PTSD, anxiety	English	189 speakers	Semistructured clinical interviews (audio, video, text)	[15]
CUDYS[q] Corpus [92]	Dysarthria	Cantonese	27 dysarthric speakers	Short sentence recordings	[18]
TORGO[r] [93]	Dysarthria	English	15 speakers (7 dysarthric, 7 controls)	Nonwords, words, sentences	[20,56]
UASpeech[a] [94]	Dysarthria	English	29 speakers (16 dysarthric, 13 controls)	Isolated words	[18-20,50,56]
KSoF[s] [95]	Fluency/stuttering	German	37 speakers	Therapy-based speech recordings	[49]
SEP-28k-E[t] [96]	Fluency/stuttering	English	21,857 three-second clips (23 h)	Spontaneous speech (podcasts)	[49]
BREF[u] [97]	Healthy speech	French	≈120 speakers (100 h)	Newspaper reading	[58]
C2SI-LEC[v] [98]	Head and neck cancer	French	94 patients, 41 controls	Pseudo-words, image description, read speech	[58]
DIRAMS[w] dataset [6]	Thyroidectomy speech impairment	Korean	114 patients (preoperation, 2-wk postoperation, 3 mo postoperation)	1 - 20 s utterance	[6]

XSL•FO
RenderX

Dataset name	Application/domain	Language	Population	Tasks	Used by
PC-GITA[x] [99]	PD[y]	Spanish	100 subjects (50 patients with PD, 50 controls)	Sustained vowel, DDK[z], words, sentences, read passage, free speech	[11,13]
Sangmyung University PD Dataset [10]	PD	Korean	200 speakers (100 PD, 100 controls)	Vowels, consonants, DDK	[10]
Ruhr University PD Dataset [100]	PD	German	168 idiopathic PDs	DDK, read passage	[13]
CzechPD [101]	PD	Czech	46 speakers (23 PD, 23 controls)	Sustained vowel, DDK, read passage, free speech	[13]
Sichuan University Schizophrenia Dataset [48]	Schizophrenia	Mandarin	28 patients, 28 controls	Emotion-elicited reading	[48]
NTUH[aa] Schizophrenia Dataset [14]	Schizophrenia	Taiwanese	26 patients	Clinical interviews	[14]
Coelho Corpus [102]	TBI[ab]	English	55 TBI, 52 controls	Memory task, picture description	[53]
Adolescent mTBI[ac] Dataset [47]	mTBI	English	72 concussion, 93 controls	Multisyllabic word reading	[47]
FEMH[ad] Speech Disorder Database [103]	Voice disorders	Mandarin	1061 samples (101 neoplasm, 100 functional dysphonia, 124 vocal palsy, 718 phonotrauma, 100 normal)	Sustained vowel	[5]
Saarbrücken Voice Database [104]	Voice disorders	German	687 healthy, 1356 patients	Sustained vowels, pitch glides, read phrase	[4,57]
VOICED[ae] [105]	Voice disorders	Italian	208 samples (150 pathological, 58 healthy)	Sustained vowel	[3,4]

XSL•FO
RenderX

Dataset name	Application/domain	Language	Population	Tasks	Used by
Yeouido St. Mary Hospital of the Catholic University Voice Dataset [59]	Voice disorder	Korean	30 laryngeal cancer, 97 vocal fold paralysis, 81 benign mucosal disease, 155 controls	Sustained vowel	[59]

[a]UASpeech: Universal Access Speech.

[b]ADReSS: Alzheimer's Dementia Recognition Through Spontaneous Speech.

[c]AD: Alzheimer disease.

[d]AVH: auditory verbal hallucination.

[e]SNUBH: Seoul National University Bundang Hospital.

[f]ASD: autism spectrum disorder.

[g]TD: typically developing.

[h]LANNA: Laboratory of Artificial Neural Network Applications.

[i]SLI: specific language impairment.

[j]CP: cleft palate.

[k]NMCPC: New Mexico Cleft Palate Center.

[l]JCCOCC: Jockey Club Centre for Osteoporosis Care and Control.

[m]MoCA: Montreal Cognitive Assessment.

[n]PTSD: posttraumatic stress disorder.

[o]CMDC: Chinese Multimodal Depression Corpus.

[p]DAIC-WOZ: Distress Analysis Interview Corpus – Wizard of Oz.

[q]CUDYS: Chinese University of Hong Kong Dysarthric Speech.

[r]TORGO: database of acoustic and articulatory speech from speakers with dysarthria (University of Toronto).

[s]KSoF: Kassel State of Fluency.

[t]SEP-28k-E: Stuttering Events in Podcasts (extended).

[u]BREF: a large read-speech corpus for French (Computer Science Laboratory for Mechanics and Engineering Sciences – French National Centre for Scientific Research).

[v]C2SI-LEC: Carcinologic Speech Severity Index corpus – short text reading task.

[w]DIRAMS: Dongnam Institute of Radiological and Medical Sciences.

[x]PC-GITA: Parkinson Corpus – Grupo de Investigación en Telecomunicaciones Aplicadas.

[y]PD: Parkinson disease.

[z]DDK: diadochokinesis.

[aa]NTUH: National Taiwan University Hospital.

[ab]TBI: traumatic brain injury.

[ac]mTBI: mild traumatic brain injury.

[ad]FEMH: Far Eastern Memorial Hospital.

[ae]VOICED: Voice Icarfederico II.

Explainability Methods

Explainability methods used in the included literature are summarized in Table 1. Across the 30 included studies, gradient-based saliency methods, input perturbation techniques, and model-internal representation analysis were the most frequently used explainability approaches.

Gradient-based saliency methods were reported in 7 studies [3,4,6,47-50], with Grad-CAM being the most commonly used technique, appearing in 5 studies. Grad-CAM [106] is a model-specific technique that uses gradients of convolutional feature maps to generate class-discriminative explanations as coarse (ie, non–pixel-level) heatmaps. Other gradient-based approaches, including guided backpropagation and vanilla saliency maps, were rarely used [4,50]. Vanilla saliency [107] computes pixel-level importance by calculating gradients of the model output (ie, logits) with respect to the input features, where the magnitude of the gradient at each feature indicates its importance to the model's decision. Guided backpropagation [108] differs from vanilla saliency maps in that gradients are only propagated when both the forward activation and the backward gradient are positive, resulting in cleaner, less noisy, and more visually interpretable maps. Unlike Grad-CAM, vanilla saliency and guided backpropagation operate directly at the input level and do not provide explicit, spatially localized class-discriminative explanations.

Input perturbation-based techniques were also widely adopted across applications. SHAP was implemented in 4 studies spanning dementia detection [9], auditory verbal hallucination assessment [52], psychiatric disorder classification [51], and TBI [53]. SHAP [109] is a model-agnostic explainability method grounded in cooperative game theory that attributes to each feature a contribution score, quantifying how the presence or absence of that feature influences a model's prediction. Ablation-based analysis was the most prevalent perturbation method, reported in 9 studies, and was applied across multiple

XSL•FO
RenderX

clinical tasks [11,12,14,15,18,20,55,56,82]. Ablation studies usually involve systematic removal of specific parts of the model or input features and observing the resulting impact on model performance.

Model-internal representation analysis was reported in 12 studies, where interpretability was derived through inspection or visualization of learned internal model representations. Although primarily used as a dimensionality reduction technique, t-distributed stochastic neighbor embedding (t-SNE) [110] dominated this category, appearing in 9 studies [3,5,19,20,49,56,57,59,60], for its utility in mapping high-dimensional latent representations into low-dimensional space while preserving local neighborhood relationships. Additional representation-level analysis included inspection of learned convolutional filters [13,20], parameterized filter structures (ie, Sinc filters) [5], and dimensionality reduction of feature maps (ie, Eigen class activation maps [Eigen-CAM]) [10], providing insight into how internal model components responded to speech signals. Sinc filters are parameterized to directly model frequency bands, allowing them to more effectively emphasize formant frequencies. As Sinc filters effectively function as bandpass filters, they offer conceptually more interpretable representations compared to the abstract patterns learned by arbitrarily shaped CNN filters [5]. Unlike Grad-CAM, Eigen-CAM [111] is a gradient-free method that calculates saliency maps by taking the first principal component of flattened feature maps.

Surrogate model-based methods were less frequently used among the included literature [9,54]. Most notably, LIME [112] is a model-agnostic explainability method that learns a simple, interpretable surrogate model, such as linear models or decision trees, that locally approximates the predictions of a black-box model. An adaptation of LIME, explainable deep learning mel-frequency cepstral coefficients (xDMFCC), is another surrogate-based model designed for interpreting MFCC-based audio representations [54].

Concept-based methods, which link internal model activations to predefined, human-interpretable concepts to explain predictions in domain-relevant terms, were implemented in 2 studies [7,58].

Lastly, attention-based explanations were reported in 2 studies, where interpretability was derived either through direct visualization of attention weights [113] or through attention rollout-based aggregation of attention across layers [57].

Explainability Input-Output Representation

Reported explainability methods operated over a range of input representations, which ultimately defined the form and interpretability of the resulting explanations.

Time-frequency and cepstral representations, including wideband and narrowband spectrograms, mel-spectrograms, and MFCCs, were commonly interrogated by gradient-based saliency methods [3,4,6,10,47,49,50]. In these cases, explanations were presented as heatmaps highlighting salient temporal segments and spectral bins of the input representation. Surrogate-based methods and input perturbation techniques typically operate over cepstral representations and acoustic low-level descriptors (LLDs) [9,51]. These approaches produced explanations in the form of numerical feature importance values and ranking scores. Various audio representations were used for ablation analysis, where interpretable insight was based on tabular performance changes, following feature removal or modification [11,12,14,15,18,20,55,56,82].

Model-internal representation analyses, such as t-SNE, derived interpretability from latent representations learned by deep models, with explanations expressed as 2D projections revealing structure or separation in the latent space [3,5,19,20,49,56,57,59,60]. In a small number of studies, parameterized filter models operating directly on raw waveforms were used, where interpretability was provided through filter frequency responses identifying salient frequency bands [5,53].

Finally, concept-based approaches operated on frame-level cepstral representations, where neural activations were inspected in relation to predefined, human-interpretable semantic or clinically meaningful concepts [7,58].

Domain-Specific Explanation Patterns

Overview

Outputs of explainability and interpretability methods were synthesized according to the target clinical application, thus identifying recurring clinical interpretability themes across the literature for each outcome domain.

Voice and Structural Laryngeal Pathology

Explanations within the voice disorder classification and laryngeal pathology literature indicate saliency across low- to midfrequency regions associated with formant structure and harmonic organization, as well as high-frequency components indicative of noise and phonatory instability. For example, for voice disorders classification, gradient-based methods emphasized low-amplitude, high-frequency spectrotemporal regions [4], while disorder-specific band emphasis patterns were reported in the study by Peng et al [3]. On a similar note, t-SNE analysis for pooled CNN features revealed overlap for benign and malignant laryngeal diseases [59]. On the other hand, parameterized Sinc filters were found to more explicitly capture F1 and F2 frequencies and energy bands compared to traditional CNN filters [5]. Postoperative recovery assessment using GRBAS (Grade, Roughness, Breathiness, Asthenia, Strain) scores showed distinct activation patterns within the 0‑2 kHz region typical of formant structure, the 2‑4 kHz band associated with harmonics and noise components, and temporally localized regions corresponding to pauses and breathiness [6]. In head and neck cancer, articulatory attributes were found to be important for intelligibility assessment through the discovery of progressively increasing phonetic feature detectors (neurons) across deeper layers of the concept detector network [58]. Ultimately, explanations consistently conveyed the importance of phonatory control, harmonic structure, and articulatory clarity as commonly highlighted cues.

While the studies by Peng et al [3] and Shaikh et al [4] relied on the VOICED (Voice Icarfederico II) dataset, other studies

XSL•FO
RenderX

based their investigations on distinct public or locally collected datasets. Thus, the aforementioned explanation patterns do not seem to be confined to a single dataset.

Parkinson Disease

Insights drawn from explainability analysis revealed degraded spectral patterns and impaired articulatory transitions consistent with hypokinetic dysarthria. Eigen-CAM emphasized higher frequencies typical of muffled and degraded speech [10]. CNN filter analysis indicated slower and less distinct phoneme transitions for patients with PD [13]. Similarly, ablation studies underscored the importance of long-term temporal dependencies and cepstral dynamics demonstrating the altered articulatory dynamics of patients with PD [11].

It is worth noting that the studies by Lahoti et al [11] and Vasquez-Correa et al [13] used the same dataset (PC-GITA); therefore, convergent findings between these 2 studies should be interpreted with appropriate context.

Dysarthria and Dysarthric Speech Recognition

Ablation analyses in dysarthria severity classification and automatic dysarthric speech recognition applications identified various spectral and temporal features as significant contributors to model decisions. Guided backpropagation revealed diffuse and less localized activation patterns in high-severity dysarthric cases, in contrast to more focused vowel-centered saliency in milder cases [50]. In ASR models, ablation and latent representation analyses emphasized the importance of temporal dependency modeling through time delay neural network architectures [18], speaker adaptation mechanisms [20], and low-quefrency components reflecting vocal tract characteristics [19]. Additionally, MFCC-, constant Q cepstral coefficients–, and articulatory-based feature sets demonstrated varying performance across speaker-dependent and speaker-independent ASR systems [56]. Notably, the findings across these studies do not constitute independent validation, as all dysarthria-focused experiments were conducted using the same dataset (UASpeech), as shown in Table 3. Therefore, these findings are data-specific and are not generalizable conclusions.

Dementia and Alzheimer Disease

In dementia detection, SHAP and LIME analysis identified noun phrase rate, empty word rate, and hesitation ratio as influential linguistic and fluency features indicative of decreased lexical complexity and increased disfluency in AD speech, alongside voice-related features reflecting lower vocal energy [9]. In addition to linguistic attributes, graph-based models in the study by Laguarta and Subirana [12] highlighted memory-related biomarkers in early-stage AD. It is important to note that both studies used the ADReSS (Alzheimer's Dementia Recognition Through Spontaneous Speech) dataset; therefore, their convergent findings should be interpreted within the context of the same dataset.

Psychiatric Disorders

Explainability and interpretability analyses across psychiatric disorders were found to emphasize the importance of vocal-spectral attributes, linguistic content, and temporally localized affective states. Grad-CAM heatmaps revealed reduced high-frequency energy in schizophrenic speech and highlighted altered formant contours compared to controls, typical of articulatory disruption [48]. Ablation analysis in the study by He et al [55] showed that wideband spectrograms, capturing transient events and articulatory changes, and narrowband spectrograms, emphasizing pitch variation and voice quality, provided complementary representations for schizophrenia detection. SHAP analyses of multimodal models identified reduced voice intensity alongside linguistic markers of negative affect and self-focus as important predictors for posttraumatic stress disorder and major depressive disorder [51]. Similarly, voice quality and prosodic features ranked among the top SHAP predictors for auditory verbal hallucination detection, while attention heatmaps prioritized textual clauses reflecting distress and interference [52]. In the study by Zhang et al [15], self-attention mechanisms applied to shorter audio segments enhanced depression detection, suggesting the temporal locality of depressive state-related cues. Across psychiatric applications, explanations frequently reflected prosodic flattening, altered spectral energy distribution, and content-level emotional markers.

Traumatic Brain Injury and Brain Lesion Detection

In TBI, GradientSHAP identified high-frequency spectral patterns and filler words as salient markers [53], and Grad-CAM highlighted high-frequency components in mild traumatic brain injury classification [47]. xDMFCC analysis demonstrated the importance of lower-order cepstral coefficients as indicators of spectral sharpness and speech clarity in brain lesion detection [54]. In contrast, higher-order cepstral coefficients did not clearly capture phoneme transitions in patients. The study concluded that brain lesion speech was characterized by delayed onset, slower articulation, and prolonged phoneme duration. Explanations across these studies allude to reduced articulatory precision and spectral clarity as discriminatory markers.

Other Clinical Domains

Several additional clinical applications were represented by single studies and therefore did not permit cross-study pattern synthesis.

The work by Mathad et al [7] performed a hypernasality assessment using a concept detector network to estimate posterior probabilities of nasal and oral phoneme classes, which were subsequently combined into an objective hypernasality measure. Similarly, Abderrazek et al [58] used a concept detector framework for head and neck cancer intelligibility assessment, using a French phone classifier to identify phonetic feature detectors (ie, internal neurons) that informed a quantitative intelligibility metric. Ablation analysis in the study by Herath et al [82] demonstrated the superiority of MFCCs over alternative spectral and temporal representations for aphasia severity classification. For autism spectrum disorder detection, t-SNE plots revealed clearer separation of autoencoder embeddings compared to handcrafted acoustic features [60]. Finally, in the study by Shen and Zhang [49], they applied time-related Grad-CAM to highlight temporally localized activations aligned with manually-annotated stutter disfluency segments.

XSL•FO
RenderX

Explainability Validation Strategies

Of the 30 included studies, 6 performed model-centric explainability validation using external data [4,7,19,55,56,58], including 1 study [19] that explicitly assessed the cross-language consistency of explanation patterns. Ground-truth or annotation-based verification was reported in a single study [49], where salient temporal regions identified by the explainability method were compared against reference labels provided by nonclinical annotators.

The remaining studies did not conduct quantitative validation of explainability outputs. Instead, 10 studies relied solely on qualitative interpretation of explanations and comparison with findings reported in prior literature [9,10,12,13,48,50-52,54,57]. In these cases, explanations were assessed narratively for plausibility or consistency with known clinical or acoustic characteristics without formal evaluation of faithfulness, robustness, or stability. Zhang et al [9] and Fu et al [48] performed cross-dataset evaluation not as consistency analysis for explainability but as external validation for the underlying model. Lastly, 12 studies reported explainability outputs without conducting quantitative analysis or qualitative literature comparisons [3,5,6,11,14,15,18,20,53,59,60,82].

Human-Centered Analysis and Stakeholder Alignment

None of the included studies explicitly reported formal human-centered evaluation of explainability outputs, such as structured assessment of explanations by clinicians, speech-language pathologists, or regulatory stakeholders.

To contextualize potential stakeholder involvement, we additionally recorded the domain expertise of study authors. Approximately half of the included studies [3,4,9,11,12,15,18-20,49,50,53,55,56,82] were authored exclusively by technical researchers (eg, computer scientists, engineers, or biomedical engineers), while the remaining studies included at least one author with nontechnical domain expertise (eg, clinical, medical, or speech-language pathology background). However, the presence of nontechnical coauthors did not correspond to explicit reporting of human-in-the-loop evaluation of explainability outputs.

Quality and Risk of Bias Assessment

A total of 25 studies were eligible for PROBAST+AI assessment, while 5 studies were not assessed because their models did not predict health-related outcomes. Further, 3 of these studies developed automatic dysarthric speech recognition models, and 2 studies developed models predicting phonetic features. Although the outputs of these models were later used to derive clinical measures, the models themselves did not constitute health-related prediction tasks within the scope of PROBAST+AI.

For model development, 80% (20/25) of studies were judged to have an overall high-quality concern, 16% (4/25) of studies had an overall low-quality concern, and 4% (1/25) of studies received an overall unclear rating. Domain 4 (analysis) was the primary driver of great concern, followed by domain 1 (participants and data sources). In domain 4, concerns were mainly related to small development datasets relative to model complexity and insufficient safeguards against overfitting. In domain 1, concerns stemmed from unclear or restrictive inclusion and exclusion criteria, limiting representativeness. Domains 2 (predictors) and 3 (outcome) were rated as low concern in nearly all studies. Applicability concerns across domains 1 - 3 were rated as low for all assessed studies.

For model evaluation, 92% (23/25) of studies were judged to have a high risk of bias, 4% (1/25) of studies a low risk, and 4% (1/25) of studies unclear risk. As in development, domains 4 and 1 were the main contributors to elevated risk. Evaluation splits were frequently small and unlikely to be representative, particularly in relation to model complexity. Data leakage was identified in 6 studies. Furthermore, performance assessment rarely extended beyond standard discrimination metrics; calibration assessment and decision-analytic measures were largely absent. Domains 2 and 3 were rated as low risk of bias across all studies. Detailed domain-level ratings for each study are provided in Multimedia Appendix 2.

Discussion

Methodological Quality of Underlying Prediction Models

The concentration of quality concerns within the analysis domain, as identified by PROBAST+AI [80], has important implications for XAI applications and the interpretability of explanations. The use of complex, high-capacity deep learning models trained and evaluated on limited clinical voice datasets, often without external validation or calibration assessment, increases the risk of overfitting to dataset-specific characteristics rather than generalizable, condition-relevant attributes.

This is especially important because post hoc explainability methods are contingent on the model's internal representations and input-output relationships. Explanations derived from overfitted or insufficiently validated models may appear coherent while reflecting confounding artifacts or spurious correlations [32,72]. Thus, the credibility of XAI methods does not rest solely on the faithfulness of the explainability technique, but also on the methodological rigor of the development and evaluation of the underlying predictive models.

These considerations are particularly relevant for the analysis of the domain-specific explanation patterns. Although recurring themes were observed across the clinical domains, the risk of bias of the underlying models and circular validation across studies due to the use of the same dataset should motivate the reader to view the findings or explanations with caution, as they may reflect data-specific characteristics rather than generalizable disease-related insights. Accordingly, the purpose of presenting domain-specific explanation patterns is not to establish definitive clinical explanatory signatures, but rather to characterize the current landscape of explainability practice in voice and speech AI and to motivate more rigorous, validated, and clinically grounded approaches in future work.

Interpretation of Explanations

Although current XAI methods provide preliminary insight into the inner workings of clinical audio models, the interpretation

XSL•FO
RenderX

of explainability outputs is rarely subjected to rigorous validation. Many studies rely on a limited number of illustrative examples to interpret local explanations (eg, saliency maps or attention weights) without quantitative assessment or statistical analysis [3,4,6,48,50,57]. This increases the risk of overinterpretation such that visually compelling or anecdotal explanations are inferred to be clinically meaningful despite limited evidence of generalizability.

Some studies attempt to contextualize explanation outputs by comparison with established clinical or acoustic knowledge [9,10,12,48,50-52,54]. While such comparisons may enhance face validity, they do not guarantee faithfulness of the explanation to the model's true decision process. Prior work has demonstrated, for example, that attention weights are not inherently faithful indicators of feature importance and may exhibit weak or inconsistent correspondence with gradient-based relevance measures, leading to potentially misleading interpretations [114-116]. Interpreting explanations primarily through the lens of existing medical literature may therefore introduce confirmation bias, whereby explanations that align with prior expectations are accepted uncritically while alternative patterns are overlooked [35].

These concerns are reinforced by the fact that most reviewed studies [3,6,9,10,12,15,18,47,48,50-54] did not perform any form of quantitative explainability validation. In the absence of systematic evaluation, explanatory interpretations are more vulnerable to confirmation bias and may overstate their clinical relevance. This pattern might be indicative of underlying issues relevant to clinical voice and speech research. The scarcity of qualitative explanation validation is possibly driven by the high cost and, subsequently, the scarcity of annotated, high-quality clinical voice and speech datasets. The same reason might also explain the high risk of bias for most included studies, even for otherwise methodologically sound ones.

Complexity-Transparency Trade-Off

The reviewed studies suggest an inherent trade-off between model complexity, input representation, and explainability in clinical audio-based AI systems [32]. In particular, the interpretability of model explanations is closely linked to the degree of semantic transparency in the input features. Audio representation in the literature included raw audio [49], low-level acoustic descriptors (eg, jitter, shimmer, and harmonic-to-noise ratio) [51,60], raw time-frequency representations (eg, spectrograms and mel-spectrograms) [6,48,50,82], coefficient-based features (eg, MFCCs) [7,11,56,82], and deep neural network–transformer-based embeddings (eg, i-vectors, wav2vec, and HuBERT [Hidden-Unit Bidirectional Encoder Representations From Transformers]) [14,18,19,60].

Models trained on manually extracted LLDs offer the highest degree of interpretability, as these features are directly related to established clinical biomarkers or constructs. Explanations from models trained on LLDs can be relatively easily understood using feature attribution methods such as SHAP or LIME. This alignment between model explanations and clinical knowledge contributes to the popularity of such methods in the medical AI domain.

While MFCCs are also hand-engineered features, they present interpretability challenges due to their abstract and decorrelated nature [117]. The coefficients do not correspond directly to intuitive phonetic or physiological phenomena. Although recent techniques such as xDMFCC attempt to provide coefficient-level explanations across time, and Tracey et al [117] sought to demystify MFCCs as vocal biomarkers by correlating them with known LLDs, the clinical relevance of these explanations remains limited. Despite efforts to position MFCCs as vocal biomarkers, they largely obscure the internal workings of deep models and contribute to the opacity of clinical audio systems [118].

While raw audio exhibits energy and temporal information, spectrograms offer a more interpretable alternative. As visual time-frequency representations, they capture dynamic changes in energy that are often physically and clinically meaningful, such as vocal formants or pauses. Trained speech pathologists and clinicians can interpret spectrograms directly, making visualization-based explanations (eg, saliency maps) more accessible and potentially clinically actionable compared to MFCCs. Although visual explanations of spectrograms might inform about the time segment most important for the detection of a condition and can highlight relevant frequency bands, more granular visual explanations are typically not feasible.

Finally, state-of-the-art systems increasingly rely on transformer-based embeddings such as wav2vec and HuBERT, which yield task-optimized, high-dimensional representations learned from raw waveforms. While these embeddings deliver substantial performance gains, they are the most difficult to interpret due to high-dimensional, nonlinear abstraction from both low-level acoustic features and clinically grounded descriptors. As a result, explanations tend to be less granular, less transparent, and harder to align with clinical reasoning.

Misalignment of Explanations With Stakeholder Needs

A major theme in the literature is the misalignment of the explanation form or modality and the needs of the end users or stakeholders. All included studies produce explanations targeted for highly technical audiences (ie, AI researchers and developers), without taking into consideration the interpretive frameworks of clinicians, patients, regulators, and policymakers. While such explanations offer insight to AI developers and researchers, critical for understanding model behavior, enhancing performance, and ensuring reliability, these explanations are often too technical, abstract, or detached from domain-specific language to be directly actionable in clinical decision-making. Figure 2 highlights the central challenge of stakeholder–explanation misalignment in clinical voice and speech audio AI, illustrating how diverse stakeholders hold distinct expectations and informational needs, and emphasizing the necessity of tailoring explanation strategies accordingly.

XSL•FO
RenderX

Figure 2. Although XAI aims to address the black-box issue of deep learning models, current XAI methods do not cater to the diverse expectations and needs of the different stakeholders. AI: artificial intelligence; XAI: explainable artificial intelligence.

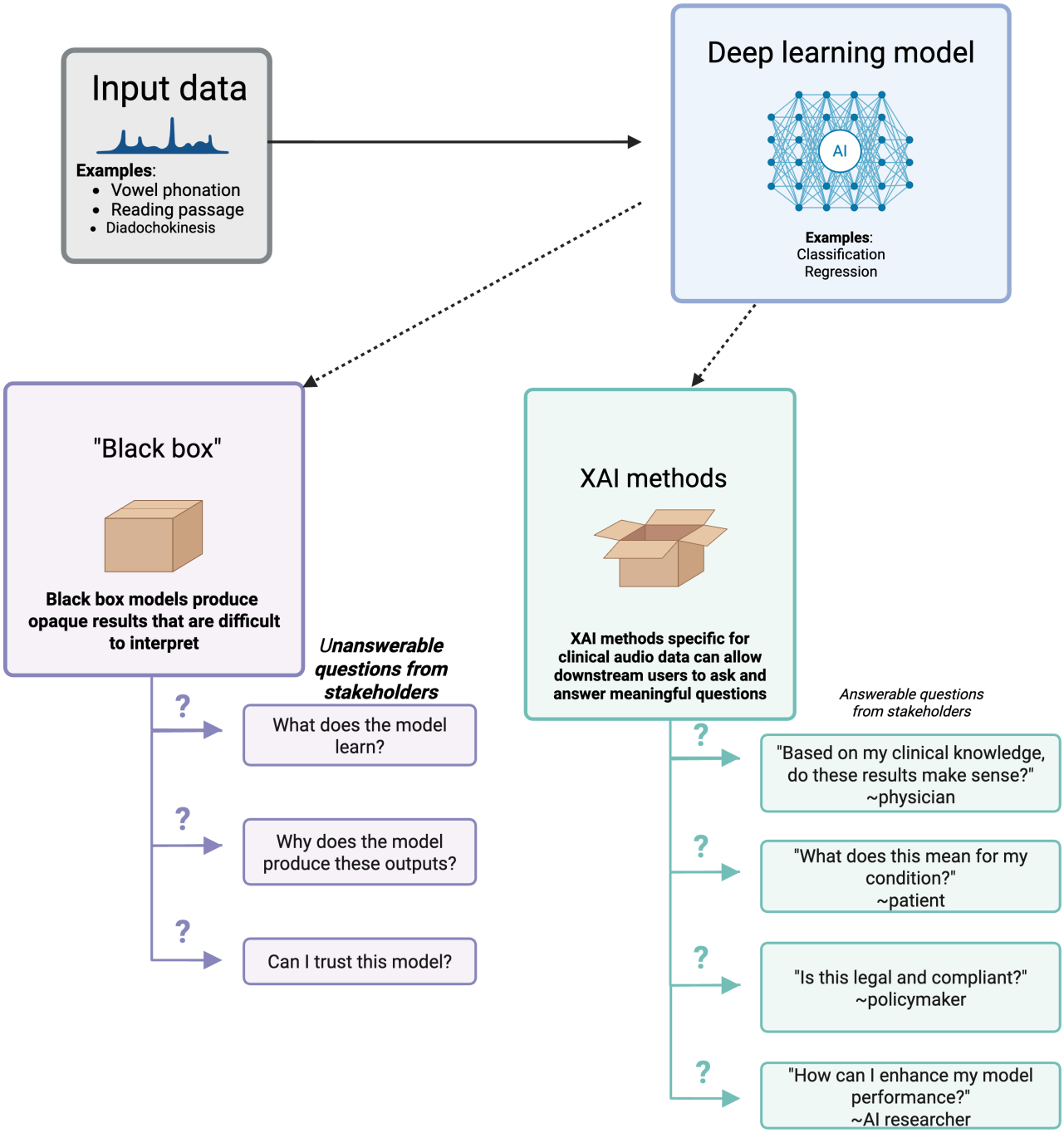

Explanations that focus on algorithmic mechanisms poorly convey information in terms of established diagnostic criteria or clinical reasoning processes [35]. Patients, with their highly limited technical and clinical knowledge, are less likely to draw relevant insights from such technical explanation modalities such as activation maps and feature attribution plots [119]. Similarly, regulators and policymakers prefer explanations that offer transparent, auditable decision pathways to assess compliance, fairness, and accountability [120].

This gap is partly a result of explanation design being driven primarily by XAI method availability rather than user requirements [121]. Most of the reviewed literature adapts generic explainability techniques to clinical audio tasks without taking into consideration the information needs, domain expertise, or cognitive constraints of their target users. This highlights the imperative need for human-centered, context-aware design of clinical audio explainability methods in which concerned stakeholders participate and provide valuable input for suitable explanation modalities and content.

Future Directions

In this work, we discussed several limitations of current methods and identified opportunities for advancing interpretability and explainability in clinical audio-based deep learning systems. Accordingly, this section outlines future directions and

XSL•FO
RenderX

recommendations for advancing XAI specifically for clinical audio applications, as summarized in Figure 3.

Figure 3. The future of voice and speech XAI in health care lies in the integration of perceptually aligned explanation methods, robust evaluation frameworks, and stakeholder-centered design, enabling explanations that are both faithful to model behavior and meaningful in real-world clinical practice. XAI: explainable artificial intelligence.

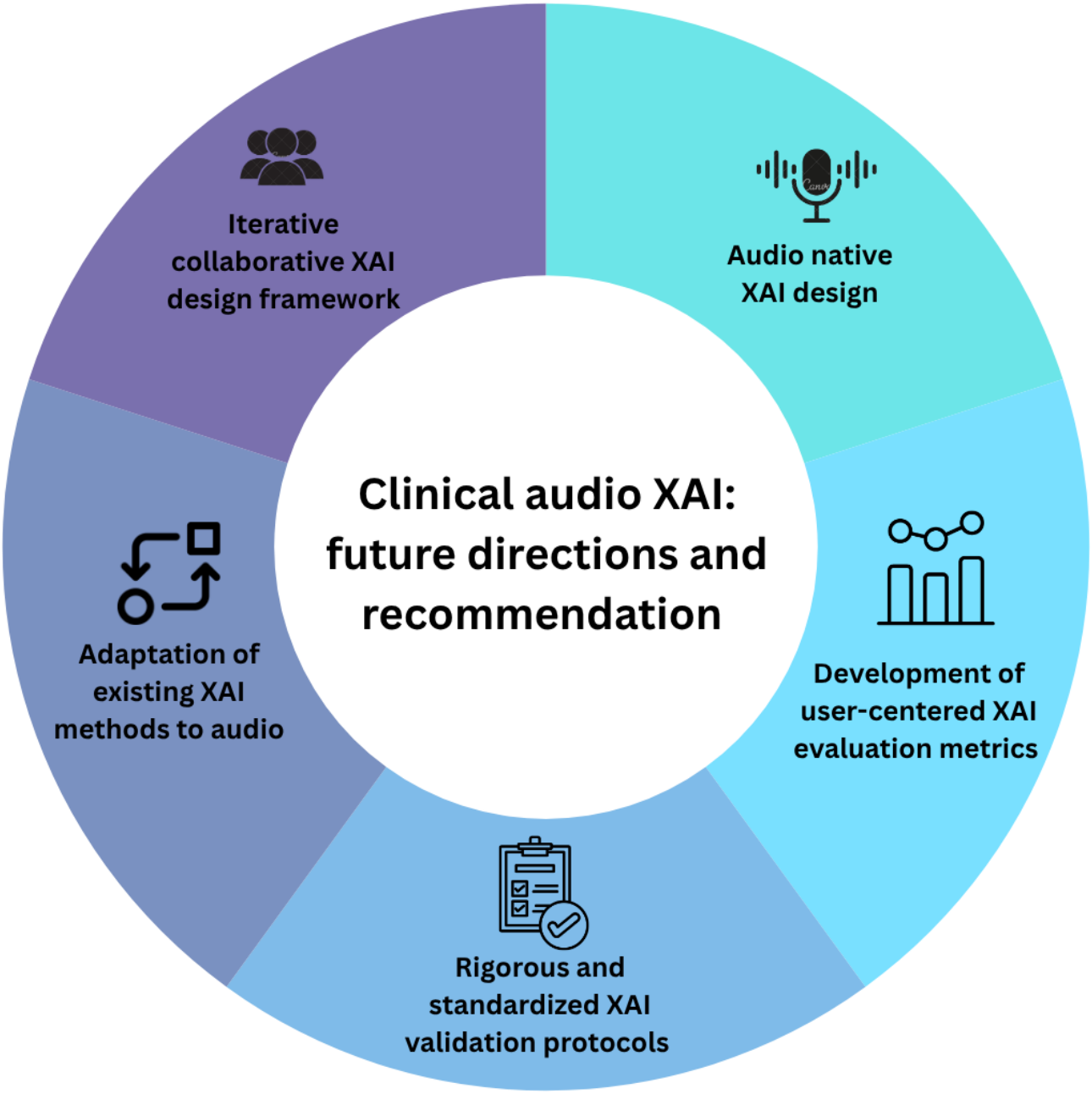

Although the literature used various explainability strategies, several techniques that are well established in related domains have rarely been adapted for clinical audio-based systems. For instance, example-based and counterexample-based explainability methods are widely used in audio-based emotion recognition [73] to generate sonified explanations. Similarly, in the text-to-audio domain, AudioGenX quantifies the importance of textual tokens corresponding to generated audio using factual and counterfactual techniques [16]. Integrating these approaches into clinical voice and speech analysis can enhance explanation fidelity and stakeholder interpretability.

Moreover, many of the XAI techniques surveyed in this review are originally developed for image or tabular data, underscoring the need for domain-specific approaches tailored to the temporal-spectral nature of audio signals and the heterogeneous manifestations of speech and voice disorders. This is especially important given the perceptual nature of audio. These methods should aim to map abstract representations (eg, deep transformer-based embeddings) to established clinical constructs, thereby bridging the gap between model outputs and actionable clinical insight. For example, CoughLIME (Cough Local Interpretable Model-Agnostic Explanations) [122] extended LIME to generate sonified explanations for COVID-19 cough analysis. Future work should aim to develop sonified

XSL•FO
RenderX

XAI that aligns with the perceptive ability of clinicians (eg, speech-language pathologists). Recent progress in large audio language models is garnering interest due to chain-of-thought reasoning and their capability to identify environmental sounds, speech characteristics, and respiratory and heart sounds [123,124]. These models can be used to extract salient acoustic features and present them in structured, report-style textual summaries through a stakeholder-friendly interface.

Our review also identified a lack of rigorous validation of XAI explanations. Future work should use both quantitative measures (eg, fidelity, sensitivity, perturbation-based testing, and cross-dataset explanation consistency) and qualitative, human-centered evaluation strategies (eg, expert annotation comparison and interrater agreement). For example, a recent work introduces a frequency band perturbation framework for quantitatively evaluating the faithfulness of various XAI techniques [125].

Additionally, the misalignment of explanation formats with stakeholder needs highlights the importance of iterative, collaborative design processes in which AI developers engage clinicians, patients, and regulators throughout system development. Such collaboration can help ensure that explanations are understandable, clinically relevant, and operationally feasible, while supporting auditability, transparency, and accountability. For instance, the work by Pizzimenti et al [126] describes a Delphi study aimed at unifying and standardizing vocal biomarker research where clinicians, statisticians, audio signal processing experts, AI researchers, and ethicists are involved in this endeavor. Such efforts highlight the importance of structured, interdisciplinary validation frameworks for clinical audio research.

Finally, throughout this systematic review, we encountered substantial difficulty in quantifying the degree of explainability and practical utility of reported methods, particularly across different stakeholder groups. Consistent with prior findings [44], only a small number of studies formally evaluated the effectiveness of XAI in clinical settings. Future work should therefore develop composite evaluation frameworks that integrate objective indicators (eg, performance improvement and error detection) with subjective measures (eg, perceived clarity, trust, and usability), enabling comparison of explainability methods in terms of both model alignment and real-world clinical impact. This is achieved by addressing limitations identified under PROBAST+AI, such that future work prioritizes study designs with representative cohorts, sufficient sample sizes, and appropriate model evaluation of diagnostic and prognostic modeling. Such practices will ensure the development of methodologically sound models, such that clinical outcomes derived from explainability methods are of high quality and clinical value, thus increasing the trustworthiness of AI in clinical practice.

Limitations

This review has several limitations that should be considered when interpreting the findings. Study selection and data extraction were conducted by a single reviewer. Although established PRISMA procedures and predefined inclusion criteria were followed, the absence of a second independent screening may increase the risk of selection bias or missed studies.

Second, conclusions regarding domain-specific explanation patterns are constrained by the methodological quality of the underlying prediction models. A substantial proportion of studies exhibited a high risk of bias, particularly in model development and evaluation. In addition, several studies relied on repeated use of the same benchmark datasets across dysarthria, PD, and AD domains, raising the risk of circular validation and inflating apparent consistency of explanation patterns. The domain-specific synthesis presented in this review is therefore intended to characterize current practice rather than establish definitive clinical explanatory signatures.

Finally, given the rapid growth of this field, studies published after the search period (February 2025) may provide additional insights and are not reflected in this review.

Conclusions

In this systematic review, we presented current practices of explainability and interpretability for deep learning–based voice and speech analysis in clinical care. Across 30 eligible studies, we identified a diverse set of explainability methods, which we organized into commonly adopted categories. Our findings indicate that, although explainability techniques are increasingly applied across a wide range of clinical speech and voice applications, their use is largely exploratory and rarely supported by rigorous validation. Explanations were predominantly assessed through qualitative interpretation, with limited evaluation of faithfulness, robustness, or consistency across datasets, and no explicit human-in-the-loop assessment involving clinical or regulatory stakeholders. Additionally, the quality of underlying models limits the validity of the reported explanation patterns and, subsequently, the applicability of these models for real-world clinical applications. These findings highlight the need for domain-specific, clinically grounded explainability methods, standardized validation protocols, and stakeholder-aware explanation design to support the safe and effective integration of voice and speech AI into clinical practice.

Acknowledgments

ME was responsible for conception, design, and main manuscript preparation. JT, YB, and JMT were responsible for supplemental materials and overall formal review and editing. All authors reviewed and approved the final version of this paper. OpenAI ChatGPT 5.2 (Feb 2026) was used for manuscript editing, correcting grammatical errors, and enhancing readability in the Abstract, Methods, Results, and figure and table captions following the first round of peer reviews.

XSL•FO
RenderX

Bridge2AI-Voice Consortium: Baycrest Centre: Amanda Chao; Linda Ma; Gayathiri Rajkumar Boston Children's Hospital: Kathy Jenkins; Stacy Jo; Elizabeth Silberholz Boston Chilren's Hospital: John Costello CENIDET: Enrique Diaz-Ocampo Dalhousie University & Vector Institute: Xijie Zeng Dalhousie University; Vector Institute: Frank Rudzicz Florida Atlantic University: Elijah Moothedan Harvard University; Massachusetts Institute of Technology: Rahul Brito Hennick Bridgepoint Hospital: Omar Ghaffar Hospital for Sick Children: Jennifer Siu; Justin Levinsky; Laurie Russell; Joyce Samuel; Lala Su Massachusetts Institute of Technology: Isaac Bevers; Kaley Jenney; Jordan Wilke; Satrajit Ghosh Mount Sinai Hospital: Julie Tu; Madeleine Zanin Mount Sinai Hospital, Sinai Health: Selina Casalino Mount Sinai Hospital, Sinai Health, Toronto; Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto: Radhika Mahajan NIH Clinical Center, U.S. National Institutes of Health; Institute of Biomedical Engineering, University of Oxford: James Anibal Oregon Health & Science University: David Dorr; Steven Bedrick; Abhijeet Dalal; William Hersh; LeAnn Michaels; Venkata Swarna Mukhi Talluri Sick Kids: Anna Goldenberg; Siyu Miao Simon Fraser University: Jean-Christophe Bélisle-Pipon; Dona Amraei; Alexander Bernier; Alden Blatter; Léo Cadillac; Amanda Doherty-Kirby; Renee English; Hortense Gallois; C. Gaelyn Garrett; Zoha Khawaja; Chloe Loewith; Marie-Françoise Malo; Pablo Montoya Varela; Michaela Pnacekova; Jaiden Potter; Claire Premi-Bortolotto; Luka Taylor; Gavin Victor; Claire Wilson Sinai Health: Lochana Jayachandran; Elisa Lapadula The Hastings Center: Vardit Ravitsky The Hospital for Sick Children: Evan Ng Trillium Health Partners and University of Toronto: Amer Ghavanini UT Health, Houston: Toufeeq Ahmed Syed University of Central Florida: Shaheen Awan University of Florida: Donald Bolser University of South Florida: Yael Bensoussan; Ruth Bahr; Stephanie Watts; Micah Boyer; Yassmeen Abdel-Aty; Kirollos Armosh; Ana Sophia Avila Martinez; Helena Beltran; Moroni Berrios; John Brown; Iris De Santiago; Mohamed Ebraheem; Ellie Eiseman; Mahmoud Elmahdy; Emily Evangelista; Karim Hanna; Jennifer Jain; Brenda Juan Guardela; Ayush Kalia; Megha Kalia; Cynthia Kostelnik; Alisa Krause; Genelle Leo; Vrishni Maharaj; Marian Mikael; Yosef Nafii; Tempestt Neal; Karlee Newberry; Christopher Nickel; Trevor Pharr; Parnaz Rafatjou; JM Rahman; Jillian Rossi; John Stark; Shrramana Ganesh Sudhakar; Jamie Toghranegar; Megan Urbano; Theresa Zesiewicz University of Toronto: Jordan Lerner-Ellis Vanderbilt University: Alexander Gelbard Vanderbilt University Medical Center: Maria Powell; Amy Brown; Kenneth Fletcher; Kenji Kobayashi; Amanda Peltier; Matthew Pontell; Sarah Rohde; Michael de Riesthal; Samantha Salvi Cruz; Kimberly Vinson Washington University in St. Louis: Andrea Krussel Washington University in St. Louis School of Medicine: Phillip Payne Weill Cornell Medicine: Alexandros Sigaras; Anaïs Rameau; Olivier Elemento; John Ramos; Jeffrey Tang; Robin Zhao; Pantelis Zisimopoulos

Funding

This project is part of the Bridge2AI-Voice program funded by the NIH (National Institutes of Health) Common Fund #3OT2OD032720-01S2. YB is the principal investigator for this grant.

Data Availability

Data extraction spreadsheet and ratings of risk of bias signaling questions are available on request through the authors' email.

Conflicts of Interest

JMT and YB are PhD advisors for ME. YB, ME, JMT, and JT are members of the Bridge2AI-Voice Consortium.

Multimedia Appendix 1
Database search queries.
[DOCX File, 20 KB - jmir_v28i1e83790_app1.docx]

Multimedia Appendix 2
PROBAST+AI risk assessment.
[DOCX File, 42 KB - jmir_v28i1e83790_app2.docx]

Checklist 1
PRISMA checklist.
[PDF File, 206 KB - jmir_v28i1e83790_app3.pdf]

References

1. Fagherazzi G, Fischer A, Ismael M, Despotovic V. Voice for health: the use of vocal biomarkers from research to clinical practice. Digit Biomark 2021;5(1):78-88. [doi: 10.1159/000515346] [Medline: 34056518]
2. Ramanarayanan V, Lammert AC, Rowe HP, Quatieri TF, Green JR. Speech as a biomarker: opportunities, interpretability, and challenges. Perspect ASHA SIGs 2022 Feb 11;7(1):276-283. [doi: 10.1044/2021_PERSP-21-00174] [Medline: 41341425]
3. Peng X, Xu H, Liu J, Wang J, He C. Voice disorder classification using convolutional neural network based on deep transfer learning. Sci Rep 2023 May 4;13(1):7264. [doi: 10.1038/s41598-023-34461-9] [Medline: 37142759]

XSL•FO
RenderX

4. Shaikh AAS, Bhargavi MS, Naik GR. Unraveling the complexities of pathological voice through saliency analysis. Comput Biol Med 2023 Nov;166:107566. [doi: 10.1016/j.compbiomed.2023.107566] [Medline: 37857135]
5. Hung CH, Wang SS, Wang CT, Fang SH. Using SincNet for learning pathological voice disorders. Sensors (Basel) 2022 Sep 2;22(17):6634. [doi: 10.3390/s22176634] [Medline: 36081092]
6. Lee JH, Lee CY, Eom JS, Pak M, Jeong HS, Son HY. Predictions for three-month postoperative vocal recovery after thyroid surgery from spectrograms with deep neural network. Sensors (Basel) 2022 Aug 24;22(17):6387. [doi: 10.3390/s22176387] [Medline: 36080847]
7. Mathad VC, Scherer N, Chapman K, Liss JM, Berisha V. A deep learning algorithm for objective assessment of hypernasality in children with cleft palate. IEEE Trans Biomed Eng 2021 Oct;68(10):2986-2996. [doi: 10.1109/TBME.2021.3058424] [Medline: 33566756]
8. Liu N, Yuan Z, Tang Q. Improving Alzheimer's disease detection for speech based on feature purification network. Front Public Health 2021;9:835960. [doi: 10.3389/fpubh.2021.835960] [Medline: 35310782]
9. Zhang Z, Wang T, Hu Z, Yang LZ, Li H. DEMENTIA: a hybrid attention-based multimodal and multi-task learning framework with expert knowledge for Alzheimer's disease assessment from speech. IEEE J Biomed Health Inform 2025 Apr;29(4):2957-2968. [doi: 10.1109/JBHI.2024.3509620] [Medline: 40030727]
10. Jeong SM, Kim S, Lee EC, Kim HJ. Exploring spectrogram-based audio classification for Parkinson's disease: a study on speech classification and qualitative reliability verification. Sensors (Basel) 2024 Jul 17;24(14):4625. [doi: 10.3390/s24144625] [Medline: 39066023]
11. Lahoti A, Gurugubelli K, Arroyave JRO, Vuppala AK. Shifted delta cepstral coefficients with RNN to improve the detection of Parkinson's disease from the speech. IC3-2022 2022 Aug 4:284-288. [doi: 10.1145/3549206.3549258]
12. Laguarta J, Subirana B. Longitudinal speech biomarkers for automated Alzheimer's detection. Front Comput Sci 2021;3:624694. [doi: 10.3389/fcomp.2021.624694]
13. Vasquez-Correa JC, Arias-Vergara T, Orozco-Arroyave JR, Eskofier B, Klucken J, Noth E. Multimodal assessment of Parkinson's disease: a deep learning approach. IEEE J Biomed Health Inform 2019 Jul;23(4):1618-1630. [doi: 10.1109/JBHI.2018.2866873] [Medline: 30137018]
14. Huang YJ, Lin YT, Liu CC, et al. Assessing schizophrenia patients through linguistic and acoustic features using deep learning techniques. IEEE Trans Neural Syst Rehabil Eng 2022;30:947-956. [doi: 10.1109/TNSRE.2022.3163777] [Medline: 35358049]
15. Zhang X, Zhang X, Chen W, Li C, Yu C. Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments. Sci Rep 2024 Apr 25;14(1):9543. [doi: 10.1038/s41598-024-60278-1] [Medline: 38664511]
16. Chen Z, Liang N, Li H, et al. Exploring explainable AI features in the vocal biomarkers of lung disease. Comput Biol Med 2024 Sep;179:108844. [doi: 10.1016/j.compbiomed.2024.108844] [Medline: 38981214]
17. Bauser M, Kraus F, Koehler F, et al. Voice assessment and vocal biomarkers in heart failure: a systematic review. Circ Heart Fail 2025 Aug;18(8):e012303. [doi: 10.1161/CIRCHEARTFAILURE.124.012303] [Medline: 40270235]
18. Liu S, Geng M, Hu S, et al. Recent progress in the CUHK dysarthric speech recognition system. IEEE/ACM Trans Audio Speech Lang Process 2021;29:2267-2281. [doi: 10.1109/TASLP.2021.3091805]
19. Geng M, Xie X, Ye Z, et al. Speaker adaptation using spectro-temporal deep features for dysarthric and elderly speech recognition. IEEE/ACM Trans Audio Speech Lang Process 2022;30:2597-2611. [doi: 10.1109/TASLP.2022.3195113]
20. Yue Z, Loweimi E, Christensen H, Barker J, Cvetkovic Z. Acoustic modelling from raw source and filter components for dysarthric speech recognition. IEEE/ACM Trans Audio Speech Lang Process 2022;30:2968-2980. [doi: 10.1109/TASLP.2022.3205766]
21. Di Cesare MG, Perpetuini D, Cardone D, Merla A. Assessment of voice disorders using machine learning and vocal analysis of voice samples recorded through smartphones. BioMedInformatics 2024;4(1):549-565. [doi: 10.3390/biomedinformatics4010031]
22. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
23. Bensoussan YS, Rameau A, Elemento O, et al. Bridge2AI-Voice: an ethically-sourced, diverse voice dataset linked to health information. PhysioNet 2020;101(23):e215-e220. [doi: 10.13026/gzjs-0535]
24. Macwhinney B, Fromm D, Forbes M, Holland A. AphasiaBank: methods for studying discourse. Aphasiology 2011;25(11):1286-1307. [doi: 10.1080/02687038.2011.589893] [Medline: 22923879]
25. What if speech could unlock early detection of alzheimer's disease? The Alzheimer's Drug Discovery Foundation. URL: https://www.alzdiscovery.org/research-and-grants/speechdx [accessed 2026-05-20]
26. Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J Med Ethics 2021 Mar 18;47(5):329-335. [doi: 10.1136/medethics-2020-106820] [Medline: 33737318]
27. Poon AIF, Sung JJY. Opening the black box of AI-medicine. J Gastroenterol Hepatol 2021 Mar;36(3):581-584. [doi: 10.1111/jgh.15384] [Medline: 33709609]
28. London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Cent Rep 2019 Jan;49(1):15-21. [doi: 10.1002/hast.973] [Medline: 30790315]

XSL•FO
RenderX

29. Raposo VL. The fifty shades of black: about black box AI and explainability in healthcare. Med Law Rev 2025 Jan 4;33(1):fwaf005. [doi: 10.1093/medlaw/fwaf005] [Medline: 39916325]
30. Aravazhi PS, Gunasekaran P, Benjamin NZY, et al. The integration of artificial intelligence into clinical medicine: trends, challenges, and future directions. Dis Mon 2025 Jun;71(6):101882. [doi: 10.1016/j.disamonth.2025.101882] [Medline: 40140300]
31. Lauritsen SM, Kristensen M, Olsen MV, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat Commun 2020 Jul 31;11(1):3852. [doi: 10.1038/s41467-020-17431-x] [Medline: 32737308]
32. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019 May;1(5):206-215. [doi: 10.1038/s42256-019-0048-x] [Medline: 35603010]
33. Teng Q, Liu Z, Song Y, Han K, Lu Y. A survey on the interpretability of deep learning in medical diagnosis. Multimed Syst 2022;28(6):2335-2355. [doi: 10.1007/s00530-022-00960-4] [Medline: 35789785]
34. Marey A, Arjmand P, Alerab ADS, et al. Explainability, transparency and black box challenges of AI in radiology: impact on patient care in cardiovascular radiology. Egypt J Radiol Nucl Med 2024;55(1):183. [doi: 10.1186/s43055-024-01356-2]
35. Wysocki O, Davies JK, Vigo M, et al. Assessing the communication gap between AI models and healthcare professionals: explainability, utility and trust in AI-driven clinical decision-making. Artif Intell 2023 Mar;316:103839. [doi: 10.1016/j.artint.2022.103839] [Medline: 41550460]
36. Bienefeld N, Boss JM, Lüthy R, et al. Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals. NPJ Digit Med 2023 May 22;6(1):94. [doi: 10.1038/s41746-023-00837-4] [Medline: 37217779]
37. Kim M, Kim S, Kim J, Song TJ, Kim Y. Do stakeholder needs differ? - Designing stakeholder-tailored explainable artificial intelligence (XAI) interfaces. Int J Hum Comput Stud 2024 Jan;181:103160. [doi: 10.1016/j.ijhcs.2023.103160]
38. Singh A, Sengupta S, Lakshminarayanan V. Explainable deep learning models in medical image analysis. J Imaging 2020 Jun 20;6(6):52. [doi: 10.3390/jimaging6060052] [Medline: 34460598]
39. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. Entropy (Basel) 2020 Dec 25;23(1):18. [doi: 10.3390/e23010018] [Medline: 33375658]
40. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: an overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA: IEEE; 2018:80-89. [doi: 10.1109/DSAA.2018.00018]
41. Das A, Rad P. Opportunities and challenges in explainable artificial intelligence (XAI): a survey. arXiv. Preprint posted online on Jun 23, 2020. [doi: 10.48550/arXiv.2006.11371]
42. Phillips PJ, Hahn CA, Fontana PC, Yates AN, Greene K, Broniatowski DA, et al. Four principles of explainable artificial intelligence. : National Institute of Standards and Technology; 2021 Sep. [doi: 10.6028/NIST.IR.8312]
43. Bhatt U, Xiang A, Sharma S, et al. Explainable machine learning in deployment. FAT* '20 2020 Jan 27:648-657. [doi: 10.1145/3351095.3375624]
44. Jung J, Lee H, Jung H, Kim H. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: a systematic review. Heliyon 2023 May;9(5):e16110. [doi: 10.1016/j.heliyon.2023.e16110] [Medline: 37234618]
45. Nannini L, Balayn A, Smith AL. Explainability in AI policies: a critical review of communications, reports, regulations, and standards in the EU, US, and UK. FAccT '23 2023 Jun 12:1198-1212. [doi: 10.1145/3593013.3594074]
46. UK GDPR guidance and resources. Information Commissioner's Office. URL: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/ [accessed 2026-05-20]
47. Rojas F, Madanian S, Templeton JM, Poellabauer C, Schneider SL. Exploring deep learning and grad-CAM for speech-based detection of mild traumatic brain injury. In: 2024 IEEE International Conference on Big Data (BigData): IEEE; 2024:6108-6116. [doi: 10.1109/BigData62323.2024.10825360]
48. Fu J, Yang S, He F, et al. Sch-net: a deep learning architecture for automatic detection of schizophrenia. Biomed Eng Online 2021 Aug 3;20(1):75. [doi: 10.1186/s12938-021-00915-2] [Medline: 34344372]
49. Shen J, Zhang X. Individual-independent and cross-language detection of speech disfluencies in stuttering based on multi-adversarial tasks and self-training. Biomed Signal Process Control 2025 Feb;100:107051. [doi: 10.1016/j.bspc.2024.107051]
50. Gupta S, Patil AT, Purohit M, et al. Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments. Neural Netw 2021 Jul;139:105-117. [doi: 10.1016/j.neunet.2021.02.008] [Medline: 33684609]
51. Schultebraucks K, Yadav V, Shalev AY, Bonanno GA, Galatzer-Levy IR. Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood. Psychol Med 2022 Apr;52(5):957-967. [doi: 10.1017/S0033291720002718] [Medline: 32744201]
52. Wang W, Xu W, Chander A, Nepal S, Buck B, Pakhomov S, et al. The power of speech in the wild: discriminative power of daily voice diaries in understanding auditory verbal hallucinations using deep learning. Proc ACM Interact Mob Wearable Ubiquitous Technol 2023;7(3):133. [doi: 10.1145/3610890] [Medline: 38737573]
53. Ditthapron A, Lammert AC, Agu EO. Continuous TBI monitoring from spontaneous speech using parametrized sinc filters and a cascading GRU. IEEE J Biomed Health Inform 2022 Jul;26(7):3517-3528. [doi: 10.1109/JBHI.2022.3158840] [Medline: 35290191]

XSL•FO
RenderX

54. Gutiérrez-Serafín B, Andreu-Perez J, Pérez-Espinosa H, Paulmann S, Ding W. Toward assessment of human voice biomarkers of brain lesions through explainable deep learning. Biomed Signal Process Control 2024 Jan;87:105457. [doi: 10.1016/j.bspc.2023.105457]
55. He L, Fu J, Li Y, Xiong X, Zhang J. WNSA-net: an axial-attention-based network for schizophrenia detection using wideband and narrowband spectrograms. IEEE/ACM Trans Audio Speech Lang Process 2023;31:721-733. [doi: 10.1109/TASLP.2022.3209941]
56. Joshy AA, Rajan R. Automated dysarthria severity classification: a study on acoustic features and deep learning techniques. IEEE Trans Neural Syst Rehabil Eng 2022;30:1147-1157. [doi: 10.1109/TNSRE.2022.3169814] [Medline: 35452390]
57. Lau HS, Huntly M, Morgan N, Iyenoma A, Zeng B. Interpreting pretrained speech models for automatic speech assessment of voice disorders. In: Bashford T, editor. Artif Intell Healthcare 2024;14975:59-72. [doi: 10.1007/978-3-031-67278-1_5]
58. Abderrazek S, Fredouille C, Ghio A, Lalain M, Meunier C, Woisard V. Interpreting deep representations of phonetic features via neuro-based concept detector: application to speech disorders due to head and neck cancer. IEEE/ACM Trans Audio, Speech, Lang Process 2023;31:200-214. [doi: 10.1109/TASLP.2022.3221039]
59. Kim HB, Song J, Park S, Lee YO. Classification of laryngeal diseases including laryngeal cancer, benign mucosal disease, and vocal cord paralysis by artificial intelligence using voice analysis. Sci Rep 2024 Apr 23;14(1):9297. [doi: 10.1038/s41598-024-58817-x] [Medline: 38654036]
60. Lee JH, Lee GW, Bong G, Yoo HJ, Kim HK. Deep-learning-based detection of infants with autism spectrum disorder using auto-encoder feature representation. Sensors (Basel) 2020 Nov 26;20(23):6762. [doi: 10.3390/s20236762] [Medline: 33256061]
61. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst 2021 Nov;32(11):4793-4813. [doi: 10.1109/TNNLS.2020.3027314] [Medline: 33079674]
62. Chaddad A, Peng J, Xu J, Bouridane A. Survey of explainable AI techniques in healthcare. Sensors (Basel) 2023 Jan 5;23(2):634. [doi: 10.3390/s23020634] [Medline: 36679430]
63. Sadeghi Z, Alizadehsani R, Cifci MA, et al. A review of explainable artificial intelligence in healthcare. Comput Electr Eng 2024 Aug;118:109370. [doi: 10.1016/j.compeleceng.2024.109370]
64. Mienye ID, Obaido G, Jere N, et al. A survey of explainable artificial intelligence in healthcare: concepts, applications, and challenges. Inf Med Unlocked 2024;51:101587. [doi: 10.1016/j.imu.2024.101587]
65. Sheu RK, Pardeshi MS. A survey on medical explainable AI (XAI): recent progress, explainability approach, human interaction and scoring system. Sensors (Basel) 2022 Oct 21;22(20):8068. [doi: 10.3390/s22208068] [Medline: 36298417]
66. van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal 2022 Jul;79:102470. [doi: 10.1016/j.media.2022.102470] [Medline: 35576821]
67. Muhammad D, Bendechache M. Unveiling the black box: a systematic review of explainable artificial intelligence in medical image analysis. Comput Struct Biotechnol J 2024 Dec;24:542-560. [doi: 10.1016/j.csbj.2024.08.005] [Medline: 39252818]
68. Caterson J, Lewin A, Williamson E. The application of explainable artificial intelligence (XAI) in electronic health record research: a scoping review. Digit HEALTH 2024;10:20552076241272657. [doi: 10.1177/20552076241272657] [Medline: 39493635]
69. Salih AM, Galazzo IB, Gkontra P, et al. A review of evaluation approaches for explainable AI with applications in cardiology. Artif Intell Rev 2024;57(9):240. [doi: 10.1007/s10462-024-10852-w] [Medline: 39132011]
70. Joyce DW, Kormilitzin A, Smith KA, Cipriani A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. NPJ Digit Med 2023 Jan 18;6(1):6. [doi: 10.1038/s41746-023-00751-9] [Medline: 36653524]
71. Kindermans PJ, Hooker S, Adebayo J, Alber M, Schütt KT, Dähne S, et al. The (un)reliability of saliency methods. In: Explainable AI: Interpreting, Explaining and Visualizing Deep: Springer-Verlag; 2022:267-280. [doi: 10.1007/978-3-030-28954-6_14]
72. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. Proc 32nd Int Conf Neural Inf Proc Syst 2018:9525-9536 [FREE Full text]
73. Zhang W, Lim BY. Towards relatable explainable AI with the perceptual process. CHI '22 2022 Apr 29:181. [doi: 10.1145/3491102.3501826]
74. Petti U, Nyrup R, Skopek JM, Korhonen A. Ethical considerations in the early detection of Alzheimer's disease using speech and AI. FAccT '23 2023 Jun 12:1062-1075. [doi: 10.1145/3593013.3594063]
75. Li P, Li L, Hamdulla A, Wang D. Reliable visualization for deep speaker recognition. arXiv. Preprint posted online on Apr 12, 2022. [doi: 10.48550/arXiv.2204.03852]
76. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 5;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
77. Speith T. A review of taxonomies of explainable artificial intelligence (XAI) methods. FAccT '22 2022 Jun 21:2239-2250. [doi: 10.1145/3531146.3534639]
78. Samek W, Montavon G, Lapuschkin S, Anders CJ, Muller KR. Explaining deep neural networks and beyond: a review of methods and applications. Proc IEEE 2021;109(3):247-278. [doi: 10.1109/JPROC.2021.3060483]

XSL•FO
RenderX

79. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv 2019 Sep 30;51(5):1-42. [doi: 10.1145/3236009]
80. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ 2025 Mar 24;388:e082505. [doi: 10.1136/bmj-2024-082505] [Medline: 40127903]
81. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019 Jan 1;170(1):51-58. [doi: 10.7326/M18-1376] [Medline: 30596875]
82. Herath H, Weraniyagoda W, Rajapaksha RTM, Wijesekara P, Sudheera KLK, Chong PHJ. Automatic assessment of aphasic speech sensed by audio sensors for classification into aphasia severity levels to recommend speech therapies. Sensors (Basel) 2022 Sep 14;22(18):6966. [doi: 10.3390/s22186966] [Medline: 36146316]
83. Martinc M, Pollak S. Tackling the ADReSS challenge: a multimodal approach to the automated recognition of Alzheimer's dementia. In: Martinc M, Pollak S, editors. Presented at: Interspeech 2020; Oct 25-29, 2020; Shanghai, China p. 2157-2161. [doi: 10.21437/Interspeech.2020-2202]
84. Becker JT, Boller F, Lopez OL, Saxton J, McGonigle KL. The natural history of Alzheimer's disease. Description of study cohort and accuracy of diagnosis. Arch Neurol 1994 Jun;51(6):585-594. [doi: 10.1001/archneur.1994.00540180063015] [Medline: 8198470]
85. Grill P, Tučková J. Speech databases of typical children and children with SLI. PLoS One 2016;11(3):e0150365. [doi: 10.1371/journal.pone.0150365] [Medline: 26963508]
86. Chapman KL, Baylis A, Trost-Cardamone J, et al. The Americleft Speech Project: a training and reliability study. Cleft Palate Craniofac J 2016 Jan;53(1):93-108. [doi: 10.1597/14-027] [Medline: 25531738]
87. Javid MH, Gurugubelli K, Vuppala AK. Single frequency filter bank based long-term average spectra for hypernasality detection and assessment in cleft lip and palate speech. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): IEEE; 2020:6754-6758. [doi: 10.1109/ICASSP40776.2020.9054684]
88. Xu SS, Ke X, Mak MW, et al. Speaker-turn aware diarization for speech-based cognitive assessments. Front Neurosci 2023;17:1351848. [doi: 10.3389/fnins.2023.1351848] [Medline: 38292896]
89. Schultebraucks K, Shalev AY, Michopoulos V, et al. A validated predictive algorithm of post-traumatic stress course following emergency department admission after a traumatic stressor. Nat Med 2020 Jul;26(7):1084-1088. [doi: 10.1038/s41591-020-0951-z] [Medline: 32632194]
90. Zou B, Han J, Wang Y, et al. Semi-structural interview-based chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders. IEEE Trans Affective Comput 2023;14(4):2823-2838. [doi: 10.1109/TAFFC.2022.3181210]
91. Gratch J, Artstein R, Lucas G, et al. The distress analysis interview corpus of human and computer interviews. ELRA 2014:3123-3128. [doi: 10.63317/3o7bccg9xequ]
92. Wong KH, Yeung YT, Chan EHY, Wong PCM, Levow GA, Meng H. Development of a cantonese dysarthric speech corpus. Proc Interspeech 2015 2015:329-333. [doi: 10.21437/Interspeech.2015-149]
93. Rudzicz F, Namasivayam AK, Wolff T. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Lang Resour Eval 2012 Dec;46(4):523-541. [doi: 10.1007/s10579-011-9145-0]
94. Kim H, Hasegawa-Johnson M, Perlman A, et al. Dysarthric speech database for universal access research. Proc Interspeech 2008 2008:1741-1744. [doi: 10.21437/Interspeech.2008-480]
95. Bayerl S, Wolff von Gudenberg A, Hönig F, Noeth E, Riedhammer K. KSoF: the kassel state of fluency dataset – a therapy centered dataset of stuttering. arXiv. Preprint posted online on Jun 16, 2022. [doi: 10.48550/arXiv.2203.05383]
96. Lea C, Mitra V, Joshi A, Kajarekar S, Bigham JP. SEP-28k: a dataset for stuttering event detection from podcasts with people who stutter. In: Bigham JP, editor. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021:6798-6802. [doi: 10.1109/ICASSP39728.2021.9413520]
97. Larnel LF, Gauvain JL, Eskenazi M. BREF, a large vocabulary spoken corpus for french. Proc 2nd Eur Conf Speech Commun Technol (Eurospeech 1991) 1991:505-508. [doi: 10.21437/Eurospeech.1991-126]
98. Woisard V, Astésano C, Balaguer M, et al. C2SI corpus: a database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers. Lang Resour Eval 2021 Mar;55(1):173-190. [doi: 10.1007/s10579-020-09496-3]
99. Orozco-Arroyave JR, Arias-Londoño JD, Vargas-Bonilla JF, González-Rátiva MC, Nöth E. New spanish speech corpus database for the analysis of people suffering from Parkinson's disease. ELRA 2014:342-347. [doi: 10.63317/3c643i6hcfzo]
100. Skodda S, Visser W, Schlegel U. Gender-related patterns of dysprosody in Parkinson disease and correlation between speech variables and motor symptoms. J Voice 2011 Jan;25(1):76-82. [doi: 10.1016/j.jvoice.2009.07.005]
101. Rusz J, Cmejla R, Ruzickova H, Ruzicka E. Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. J Acoust Soc Am 2011 Jan 1;129(1):350-367. [doi: 10.1121/1.3514381]
102. Coelho CA, Youse KM, Le KN. Conversational discourse in closed-head-injured and non-brain-injured adults. Aphasiology 2002 Apr;16(4-6):659-672. [doi: 10.1080/02687030244000275]
103. Bhat C, Kopparapu SK. FEMH voice data challenge: voice disorder detection and classification using acoustic descriptors. In: 2018 IEEE International Conference on Big Data (Big Data): IEEE; 2018:5233-5237. [doi: 10.1109/BigData.2018.8622543]

XSL•FO
RenderX

104. Pützer M, Barry WJ. Saarbruecken voice database. : Zenodo; 2008. [doi: 10.5281/zenodo.16874898]
105. Cesari U, De Pietro G, Marciano E, Niri C, Sannino G, Verde L. A new database of healthy and pathological voices. Comput Electr Eng 2018 May;68:310-321. [doi: 10.1016/j.compeleceng.2018.04.008]
106. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis 2020 Feb;128(2):336-359. [doi: 10.1007/s11263-019-01228-7]
107. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv. Preprint posted online on Apr 19, 2014. [doi: 10.48550/arXiv.1312.6034]
108. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: the all convolutional net. arXiv. Preprint posted online on Apr 13, 2015. [doi: 10.48550/arXiv.1412.6806]
109. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. arXiv. Preprint posted online on Nov 25, 2017. [doi: 10.48550/arXiv.1705.07874]
110. Maaten L, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res 2008;9(86):2579-2605 [FREE Full text]
111. Muhammad MB, Yeasin M. Eigen-CAM: class activation map using principal components. In: Muhammad MB, Yeasin M, editors. 2020 International Joint Conference on Neural Networks (IJCNN): IEEE; 2020:1-7. [doi: 10.1109/IJCNN48605.2020.9206626]
112. Ribeiro MT, Singh S, Guestrin C. “Why should I trust you?”: Explaining the predictions of any classifier. Proc 22nd ACM SIGKDD Int Conf Knowl Discovery Data Min 2016:1135-1144. [doi: 10.1145/2939672.2939778]
113. Wang P, Lu W, Lu C, Zhou R, Li M, Qin L. Large language model for medical images: a survey of taxonomy, systematic review, and future trends. Big Data Min Anal 2025;8(2):496-517. [doi: 10.26599/BDMA.2024.9020090]
114. Liu Y, Li H, Guo Y, Kong C, Li J, Wang S. Rethinking attention-model explainability through faithfulness violation test. In: Kamalika C, Stefanie J, Le S, Csaba S, Gang N, Sivan S, editors. PMLR 2022;162:13807-13824 [FREE Full text]
115. Lopardo G, Precioso F, Garreau D. Attention meets post-hoc interpretability: a mathematical perspective. In: Ruslan S, Zico K, Katherine H, Adrian W, Nuria O, Jonathan S, editors. PMLR 2024;235:32781-32800 [FREE Full text]
116. Bibal A, Cardon R, Alfter D, et al. Is attention explanation? An introduction to the debate. In: François T, editor. Proc 60th Ann Meeting Assoc Comput Linguist (Vol 1) 2022:3889-3900. [doi: 10.18653/v1/2022.acl-long.269]
117. Tracey B, Volfson D, Glass J, et al. Towards interpretable speech biomarkers: exploring MFCCs. Sci Rep 2023 Dec 21;13(1):22787. [doi: 10.1038/s41598-023-49352-2] [Medline: 38123603]
118. Jiao Y, Berisha V, Liss J. Interpretable phonological features for clinical applications. In: Jiao Y, Berisha V, Liss J, editors. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): IEEE; 2017:5045-5049. [doi: 10.1109/ICASSP.2017.7953117]
119. Aranovich TDC, Matulionyte R. Ensuring AI explainability in healthcare: problems and possible policy solutions. Inf Commun Technol Law 2023 May 4;32(2):259-275. [doi: 10.1080/13600834.2022.2146395]
120. Moorthy UMK, Muthukumaran AMJ, Kaliyaperumal V, Jayakumar S, Vijayaraghavan KA. Explainability and regulatory compliance in healthcare. In: Explainable Artificial Intelligence in the Healthcare Industry: John Wiley & Sons, Ltd; 2025:521-561. [doi: 10.1002/9781394249312]
121. Chen H, Gomez C, Huang CM, Unberath M. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. NPJ Digit Med 2022 Oct 19;5(1):156. [doi: 10.1038/s41746-022-00699-2] [Medline: 36261476]
122. Wullenweber A, Akman A, Schuller BW. CoughLIME: sonified explanations for the predictions of COVID-19 cough classifiers. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC): IEEE; 2022:1342-1345. [doi: 10.1109/EMBC48229.2022.9871291]
123. Wang TN, Chen LL, Zeghidour N, Saeed A. CaReAQA: a cardiac and respiratory audio question answering model for open-ended diagnostic reasoning. In: X XO, C E, S P, G W, T S, A M, editors. PMLR 2025;287:231-246 [FREE Full text]
124. Ma Z, Chen Z, Wang Y, Chng ES, Chen X. Audio-cot: exploring chain-of-thought reasoning in large audio language model. arXiv. Preprint posted online on Jan 13, 2025. [doi: 10.48550/arXiv.2501.07246]
125. Buck A, Cosma G, Phillips I, Conway P, Baker P. A framework for evaluating faithfulness in explainable AI for machine anomalous sound detection using frequency-band perturbation. arXiv. Preprint posted online on Jan 26, 2026. [doi: 10.48550/arXiv.2601.19017]
126. Pizzimenti M, Kalia A, Toghranegar JA, et al. Consensus-based definitions for VOCAL biomarkers: the international VOCAL initiative. medRxiv. Preprint posted online on Dec 3, 2025. [doi: 10.1101/2025.10.23.25338518] [Medline: 41404273]

Abbreviations

AD: Alzheimer disease
ADReSS: Alzheimer's Dementia Recognition Through Spontaneous Speech
AI: artificial intelligence
ASR: automatic speech recognition
CNN: convolutional neural network

XSL•FO
RenderX

CoughLIME: Cough Local Interpretable Model-Agnostic Explanations
CT: computed tomography
ECG: electrocardiogram
EHR: electronic health record
Eigen-CAM: Eigen class activation maps
Grad-CAM: gradient-weighted class activation mapping
GRBAS: Grade, Roughness, Breathiness, Asthenia, Strain (voice scale)
HuBERT: Hidden-Unit Bidirectional Encoder Representations from Transformers
LIME: local interpretable model-agnostic explanations
LLD: low-level descriptor
MFCC: mel-frequency cepstral coefficients
ML: machine learning
MRI: magnetic resonance imaging
PC-GITA: Parkinson Corpus – Grupo de Investigación en Telecomunicaciones Aplicadas
PD: Parkinson disease
PICOTS: Population; Index Model; Comparator; Outcome; Timing; Setting
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST+AI: Prediction Model Risk of Bias Assessment Tool for Artificial Intelligence
SHAP: Shapley Additive Explanations
t-SNE: t-distributed stochastic neighbor embedding
TBI: traumatic brain injury
UASpeech: Universal Access Speech
VOICED: Voice Icarfederico II
XAI: explainable artificial intelligence
xDMFCC: explainable deep learning mel-frequency cepstral coefficients

Edited by A Coristine; submitted 09.Sep.2025; peer-reviewed by A Amanna, S Pradhan, X Wang; revised version received 17.Apr.2026; accepted 20.Apr.2026; published 24.Jun.2026.

Please cite as:
Ebraheem M, Toghranegar J, Bridge2AI-Voice Consortium, Bensoussan Y, Templeton JM
Explainable and Interpretable AI for Voice and Speech Analysis in Clinical Care: Systematic Review
J Med Internet Res 2026;28:e83790
URL: https://www.jmir.org/2026/1/e83790
doi:10.2196/83790

© Mohamed Ebraheem, Jamie Toghranegar, Bridge2AI-Voice Consortium, Yael Bensoussan, John Michael Templeton. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Radiomics-Based AI for the Diagnosis and Prognosis of Vessels Encapsulating Tumor Clusters in Hepatocellular Carcinoma: Systematic Review and Meta-Analysis

Xuefeng Hua[1*], PhD; Rongdang Fu[2*], PhD; Ziwei Yin[3], PhD

[1]Department of Hepatobiliary and Pancreatic Surgery, Guangzhou First People's Hospital, South China University of Technology, Guangzhou, Guangdong, China

[2]First People's Hospital of Foshan (Foshan Hospital affiliated to Southern University of Science and Technology), School of Medicine, Southern University of Science and Technology, Foshan, Guangdong, China

[3]Department of Oncology, Third Xiangya Hospital of Central South University, 138 Tongzipo Rd, Yuelu District, Changsha, Hunan, China

[*]these authors contributed equally

Corresponding Author:
Ziwei Yin, PhD
Department of Oncology, Third Xiangya Hospital of Central South University, 138 Tongzipo Rd, Yuelu District, Changsha, Hunan, China

Abstract

Background: Vessels encapsulating tumor clusters (VETCs), a CD34-positive vascular pattern in hepatocellular carcinoma (HCC), are linked to aggressive biology, early recurrence, and poor survival. Because pathologic VETC assessment requires postoperative immunohistochemistry and may be affected by sampling, preoperative noninvasive prediction remains clinically important. Radiomics-based artificial intelligence (AI) applied to routine contrast-enhanced imaging may provide a surrogate marker, but evidence across has not been comprehensively appraised.

Objective: This study aimed to evaluate the diagnostic accuracy and prognostic value of radiomics-based AI models for noninvasive VETC prediction in HCC using PICOTS (patient population, intervention, comparator, outcomes, timing, and setting) and PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Diagnostic Test Accuracy) frameworks.

Methods: We searched PubMed, Embase, Web of Science, and the Cochrane Library, gray literature, and citations (original search July 11, 2025; updated April 17, 2026). The original search was completed on July 11, 2025, and the reconstructed strategy was rerun on April 17, 2026. Eligible retrospective cohort studies developed or validated radiomics or deep-learning models using contrast-enhanced magnetic resonance imaging (CEMRI), contrast-enhanced computed tomography (CECT), contrast-enhanced ultrasound (CEUS), or 2-[^{1} F]fluoro-2-deoxy-D-glucose positron emission tomography or computed tomography ([^{18}F]FDG PET/CT) to predict CD34-confirmed VETC and reported 2×2 diagnostic data and/or hazard ratios (HRs) for early recurrence. Mutually exclusive cohorts were treated as separate datasets only when patient overlap was absent. Risk of bias was assessed with the Prediction model Risk Of Bias Assessment Tool+AI, and certainty with GRADE (Grading of Recommendations, Assessment, Development, and Evaluation). Diagnostic accuracy was synthesized with bivariate random-effects models; prognostic HRs were pooled with restricted maximum likelihood+Hartung-Knapp-Sidik-Jonkman random-effects models.

Results: In total, 15 studies (729 internal-validation and 613 external-validation patients) were included; 14 were from China and 1 from Japan. Moreover, 10 studies used CEMRI, 3 CECT, 1 CEUS, and 1 [^{18}F]FDG PET/CT. CEMRI-based AI showed the best performance: sensitivity=0.84 (95% CI 0.73-0.93; 95% prediction interval [PI] 0.45-1.00), specificity=0.79 (95% CI 0.70-0.86; 95% PI 0.50-0.97), and area under the curve=0.87. Meta-regression suggested that center type, validation design, algorithm class, and magnetic resonance imaging field strength contributed to specificity heterogeneity. CECT, CEUS, and [^{18}F]FDG PET/CT evidence was limited. AI-predicted VETC positivity was associated with early recurrence (HR 2.34, 95% CI 1.93-2.84). GRADE certainty ranged from low to moderate, mainly due to imprecision, risk of bias, and heterogeneity.

Conclusions: This review is innovative because it integrates diagnostic accuracy, modality comparison, algorithm performance, and recurrence prognosis for AI-based VETC prediction. Unlike previous modality-specific reviews, it clarifies what AI brings to the field: a potential preoperative bridge between imaging phenotypes and biologically aggressive HCC. Real-world use should remain cautious and decision-supportive, given retrospective designs, geographically concentrated cohorts, limited external validation, heterogeneity, risk of bias, and low-to-moderate GRADE certainty, rather than replacing histopathology or multidisciplinary clinical judgment in practice.

XSL•FO
RenderX

Trial Registration: PROSPERO CRD420251167155; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251167155

(J Med Internet Res 2026;28:e90931) doi:10.2196/90931

KEYWORDS

radiomics; artificial intelligence; hepatocellular carcinoma; vessels encapsulating tumor clusters; meta-analysis; PRISMA; Preferred Reporting Items for Systematic reviews and Meta-Analyses

Introduction

Hepatocellular carcinoma (HCC) is the most common primary malignant liver tumor, accounting for approximately 90% of liver cancer cases. According to the latest global statistics, liver cancer remains a major health burden with approximately 865,269 new cases and 757,948 deaths recorded in 2022. It is characterized by an age-standardized mortality rate of 7.4 per 100,000, making it the third leading cause of cancer-related mortality worldwide [1,2]. The poor prognosis associated with HCC is largely driven by its high recurrence rate and limited therapeutic window, which underscores the urgent need for early risk stratification to optimize treatment outcomes [1]. Vessels encapsulating tumor clusters (VETCs) are a specific vascular pattern characterized by tumor clusters being completely enveloped by CD34-positive endothelial cells [3]. Research has demonstrated that VETC positivity is an independent poor prognostic factor in HCC, associated with higher early recurrence rates and shorter disease-free and overall survival compared with that in patients with negative VETC [4]. Consequently, the noninvasive and accurate prediction of VETC status before treatment is crucial for identifying high-risk patients, tailoring personalized treatment strategies, and guiding postoperative surveillance [5]. Moreover, early identification of VETC status has the potential to influence clinical decisions regarding surgical versus nonsurgical management, thereby further improving treatment planning [6-8].

The traditional gold standard for VETC diagnosis relies on immunohistochemical analysis (eg, CD34 staining) of postoperative pathological specimens [9]. However, this method is invasive, cannot be performed preoperatively, and due to intratumoral heterogeneity, biopsy sampling may be subject to bias, potentially missing VETC-positive areas [10]. To overcome these limitations, noninvasive imaging techniques have been explored. Although conventional imaging techniques, such as computed tomography (CT) [11], magnetic resonance imaging (MRI), ultrasound [12], and 2-[^{1}F]fluoro-2-deoxy-D-glucose positron emission tomography or computed tomography ([^{1}F]FDG PET/CT) [13], are used for preoperative HCC evaluation and have been explored for predicting VETC status, their accuracy is limited by tumor heterogeneity, interobserver variability in interpretation, and the lack of standardized assessment criteria [14].

In recent years, artificial intelligence (AI) technologies based on medical images, particularly deep learning and radiomics, have shown significant potential for the noninvasive prediction of VETC and prognosis assessment in HCC [15]. Radiomics enables the high-throughput extraction of quantitative features from medical images, while deep learning models can automatically learn complex, subvisual imaging patterns associated with the VETC phenotype [16]. Unlike biopsy, which is limited by invasiveness and sampling bias, and unlike conventional imaging, which relies on subjective visual interpretation, AI-based approaches can comprehensively capture and integrate high-dimensional tumor characteristics to overcome these limitations [17]. However, existing studies exhibit considerable variability in model performance, suffer from small sample sizes, and lack methodological standardization, which collectively motivate the need for this systematic review and meta-analysis [18,19]. Moreover, the impact of different imaging modalities and algorithms on AI model performance remains controversial, lacking systematic synthesis [20].

Several systematic reviews have addressed related questions, but leave critical gaps that this study fills. Existing meta-analyses synthesizing radiomics models for HCC microvascular invasion have examined a different histopathological end point and therefore do not inform VETC-specific preoperative decision-making. A recent modality-specific meta-analysis by Wang et al [21] evaluated only MRI-based radiomics for VETC and did not pool evidence across CT, ultrasound, or PET/CT, nor did it quantify the prognostic value of AI-predicted VETC for early recurrence. Narrative reviews cataloging qualitative imaging features of VETC have not provided pooled diagnostic estimates, risk-of-bias appraisal using the recent Prediction model Risk Of Bias Assessment Tool (PROBAST)+AI tool, or GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) certainty ratings [3]. To our knowledge, no previous quantitative synthesis has simultaneously (1) pooled diagnostic accuracy of AI-based VETC prediction across all 4 major contrast-enhanced modalities (contrast-enhanced magnetic resonance imaging [CEMRI], contrast-enhanced computed tomography [CECT], contrast-enhanced ultrasound [CEUS], and [^{18}F]FDG PET/CT), (2) integrated PROBAST+AI risk-of-bias appraisal with GRADE certainty grading, and (3) meta-analyzed the hazard ratio (HR) of AI-predicted VETC for early recurrence. This review is designed to address these 3 gaps and therefore provides incremental, decision-relevant evidence beyond the existing literature.

Therefore, this systematic review and meta-analysis aims to synthesize the existing evidence to rigorously evaluate the overall diagnostic accuracy and sources of heterogeneity in image-based AI models for predicting VETC status and its prognostic value in HCC.

XSL•FO
RenderX

Methods

Overview

This systematic review and meta-analysis adhered to the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Diagnostic Test Accuracy) guidelines [22], CHARMS (Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies) checklists, PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist, and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) checklist, and was prospectively registered on PROSPERO (CRD420251167155). These 4 checklists were provided in Checklists 1-4.

Search Strategy

A comprehensive literature search was conducted across 4 major databases—PubMed (via NCBI), Embase (via Ovid), Web of Science (via Clarivate), and the Cochrane Library (via Wiley). The original search strategy was completed on July 11, 2025. During revision, the strategy was reconstructed to improve reproducibility and database-specific syntax; it was rerun on March 17, 2026, and the finalized strategy was last updated on April 17, 2026. Thus, July 11, 2025, refers to the initial search cutoff, whereas April 17, 2026, refers to the most recent rerun of the reconstructed search strategy. The reconstruction did not change the review question or eligibility criteria; it made the previously used AI-related terms, VETC-specific terms, and HCC-specific terms explicit for each database. To ensure comprehensiveness, no limits were placed on language or publication year. Search results from all databases were exported to EndNote X20 (Clarivate Analytics) for systematic deduplication, followed by manual verification. Moreover, 2 independent reviewers (XH and RF) performed initial screening of titles and abstracts, followed by full-text assessment.

Inclusion and Exclusion Criteria

Studies were included based on the PITROS framework:

1. Population (P): patients with HCC scheduled for hepatectomy were included. Studies involving nonsurgical patients were excluded due to the lack of pathological confirmation, as definitive identification of the VETC vascular pattern currently requires microscopic evaluation of the resected tumor tissue.
2. Index test (I): radiomics-based AI, including CEMRI, CECT, CEUS, and [^{1}F]FDG PET/CT, extracting high-dimensional quantitative features to predict VETC.
3. Target condition (T): positive group defined as VETC-positive, negative group defined as VETC-negative.
4. Reference standard (R): pathological biopsy of VETC for result validation.
5. Outcomes (O): sensitivity, specificity, area under the curve (AUC), and early recurrence HR.
6. Setting (S): retrospective or prospective case data from public databases or local hospitals.

Studies were systematically excluded based on the following comprehensive criteria: (1) titles and abstracts irrelevant to the review question; (2) nonoriginal literature types, including reviews, case reports, conference abstracts, meta-analyses, and editorial correspondence; (3) studies without AI algorithms; (4) research not evaluating VETC; (5) studies lacking essential diagnostic data; or (6) studies with overlapping patient data. For publications reporting more than 1 cohort, such as training, internal-validation, and external-validation cohorts, each cohort was treated as a separate dataset only when the patients were mutually exclusive. Cohorts with overlapping participants were not entered separately, in accordance with the principle that double counting occurs when the same participants or evidence are included more than once in a synthesis. The screening process was independently conducted by 2 reviewers (XH and RF), who first assessed title and abstract relevance, then evaluated full-text articles against inclusion and exclusion criteria. Any disagreements between reviewers were resolved through discussion, with a third reviewer (ZY) consulted if consensus could not be reached.

Quality Assessment and GRADE Certainty

We used the latest PROBAST+AI quality assessment tool [23], which supersedes the 2019 PROBAST version. This comprehensive evaluation framework comprises 2 distinct phases—model development and model assessment, with each phase encompassing 7 critical domains addressing participants, data sources, predictive factors, outcome assessment, and analytical methodologies. The assessment categorizes each domain's evaluation results into 3 levels—low (L), high (H), and unclear (U), determined through specific signaling questions. These questions were graded using a nuanced scale: "Yes" (Y), "Probably Yes" (PY), "Probably No" (PN), "No" (N), "No Information" (NI), and occasionally "Not Applicable" (NA).

To ensure methodological rigor and minimize potential bias, 2 independent reviewers (XH and RF) conducted a comprehensive risk of bias assessment using the PROBAST+AI tool. Any disagreements between reviewers were systematically resolved through in-depth discussion and meticulous analysis, ensuring high reliability and consistency in the final assessment. Detailed signaling questions and comprehensive evaluation tables are provided in Tables S2 and S3 in Multimedia Appendix 1.

We used the GRADE framework to assess the certainty of the evidence for sensitivity and specificity. This method evaluates 5 key domains—risk of bias, indirectness, inconsistency, imprecision, and small-study effects—to systematically rate the evidence quality. The detailed assessment criteria and protocols are fully documented in Table S4 in Multimedia Appendix 1 to ensure transparency.

Outcome Measures

The primary outcome measures encompassed sensitivity, specificity, and AUC for AI-based approaches across different imaging modalities (CEMRI, CECT, CEUS, and [^{18}F] FDG PET/CT), alongside their early recurrence HR values. Sensitivity, calculated as true positive (TP)/(TP+false negative [FN]), quantifies the AI model's ability to accurately identify true positive cases, while specificity, calculated as true negative (TN)/(TN+false positive [FP]), evaluates the model's proficiency

XSL•FO
RenderX

in correctly recognizing negative cases. AUC represents the comprehensive discriminative performance of the receiver operating characteristic (ROC) curve, providing a holistic assessment of the model's diagnostic capabilities. For early recurrence HR values, we directly extracted reported HRs with their 95% CIs from the original studies. In cases where direct reporting was absent, we planned to indirectly calculate HR values using published survival curves through precise data extraction and reconstruction techniques. An aggregated HR>1 with statistical significance would indicate the AI model's significant prognostic stratification capability, demonstrating that patients categorized as "high-risk" by the AI model have a multiplicative early recurrence risk compared with the "low-risk" group. Early recurrence was defined according to the original studies; definitions varied across studies (eg,≤1 y or ≤2 y after surgery). For studies presenting multiple nonoverlapping patient datasets, we considered the contingency tables independent and extracted them comprehensively. When multiple AI algorithms were reported, we selected the highest-performing model based on AUC for analysis.

Data Extraction

Two independent reviewers (XH and RF) systematically extracted data from full-text articles to assess their eligibility, with any discrepancies resolved through consultation with a third reviewer (ZY). The comprehensive data extraction process encompassed multiple critical domains, including (1) patient and study baseline information: authors, publication year, country, study design, imaging modalities (CEMRI, CECT, CEUS, and [^{18}F] FDG PET/CT), reference standard, analysis methodology, total patient numbers (training and internal and external validation sets), and patient counts with positive VETC; (2) technical parameters: research center, data source, AI methodology, AI model, optimal AI algorithms, data splitting approach, and TP, FP, FN, and TN across different validation sets; and (3) imaging modality information: manufacturer, scanner modality, evaluation time, regions of interest, sequence parameters, enhancement phases, contrast agents, and radiotracer details. Given that most studies did not provide complete diagnostic contingency tables, we reconstructed TP, FP, FN, and TN values by systematically analyzing sensitivity, specificity, reference standard positive VETC cases, and total patient populations.

Statistical Analysis

For the diagnostic meta-analysis, we addressed the observed between-study heterogeneity by using a bivariate random-effects model using the restricted maximum likelihood (REML) method. To enhance the robustness of our pooled estimates and account for uncertainty in variance, the Hartung-Knapp-Sidik-Jonkman (HKSJ) adjustment was applied to calculate the pooled sensitivity and specificity along with their corresponding 95% CIs. We calculated 95% prediction intervals (PIs) for the main diagnostic syntheses and for modality-specific subgroup syntheses only when random-effects pooling was performed with sufficient independent datasets. These PIs were reported as descriptive estimates of how performance might vary in a future clinical setting and were not used alone to justify additional subgroup analyses. Because subgroup analyses were prespecified by imaging modality, subgroup PIs were intended to contextualize heterogeneity within clinically defined modality groups rather than to generate new post hoc subgroup analyses [24]. The overall diagnostic accuracy was summarized by the AUC derived from the hierarchical summary receiver operating characteristic model, and the interpretation of these values was adjusted to acknowledge the impact of the substantial heterogeneity identified. Pooled HRs for early recurrence were estimated using a single prespecified REML+HKSJ random-effects model.

For CEMRI-based studies with more than 10 studies, meta-regression and bivariate boxplots were performed to investigate potential sources of heterogeneity. The prespecified covariates for meta-regression included center (single vs multi), validation type (internal vs external), data splitting method (hold-out vs independent validation), AI model type (radiomics vs radiomics and clinical), AI algorithms (deep learning vs machine learning), and MRI field strength (1.5T&3.0T vs 3.0T). The impact of different algorithms on diagnostic performance was also assessed. A bubble plot visualized the diagnostic odds ratio (DOR) over time. The clinical utility of CEMRI-based models was appraised using Fagan's nomogram, and small-study effects were assessed using Deeks' funnel plot asymmetry test (P value<.05 for the slope coefficient, indicating potential bias). A chord diagram illustrated the frequency of different imaging modalities and algorithms. All analyses were conducted using Stata (version 15.1; StataCorp LLC; with the *Midas* and *Metadta* packages) and R (version 4.3.1; R Core Team; using ggplot2 and *tidyverse*). Several deviations from the original registered protocol should be noted; the HKSJ adjustment and the reporting of 95% PIs were not prespecified but were incorporated post hoc to strengthen methodological rigor. Additionally, certain meta-regression covariates, such as MRI field strength and AI algorithm type, were added based on clinical relevance identified during the review process. No changes were made to the original inclusion or exclusion criteria.

Results

Study Selection

The initial database search identified 549 potentially relevant publications. Following the removal of 281 duplicate records, 268 articles proceeded to the preliminary screening phase. During this stage, 243 records were excluded based on title and abstract review, primarily due to clear irrelevance or nonconforming publication types. This resulted in 25 articles sought for full-text retrieval. Out of 25 articles, 1 was excluded as the full text was inaccessible. The remaining 24 articles underwent a detailed, full-text eligibility assessment. Of these, 9 studies were excluded due to insufficient or incomplete diagnostic data, with reasons including (1) without using AI algorithms, (2) nonevaluated VETC, (3) data (TP, FP, TN, and FN) not available, or (4) patients overlap. Consequently, 15 studies [6-8,12,24-34] met the predefined inclusion criteria and were incorporated into the final meta-analysis. The entire study selection process adhered to the PRISMA guidelines and is detailed in Figure 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram illustrating the systematic identification and selection process for studies on radiomics-based AI models predicting vessels encapsulating tumor clusters in hepatocellular carcinoma. AI: artificial intelligence; FN: false negative; FP: false positive; TN: true negative; TP: true positive; VETC: vessels encapsulating tumor cluster.

XSL•FO
RenderX

Study Description and Quality Assessment

A total of 15 eligible studies were included in this meta-analysis (Tables 1 and 2). The validation cohorts consisted of 12 studies [6-8,12,26-28,31-35] providing internal validation data (encompassing 729 patients) and 7 studies [6,7,27,29,30,32,36] providing external validation data (encompassing 613 patients). The distribution by primary imaging modality was as follows: 3 used CECT [6,27,36], 1 used [^{18}F]FDG PET/CT [8], 1 used CEUS [12], and the remaining 10 studies used CEMRI [7,26,28-35]. All 15 studies, published between 2021 and 2025, were retrospective in design. The reference standard for confirming VETC status was pathological biopsy. With the exception of 1 study from Japan [35], all others originated from China. All analyses were conducted on a per-patient basis. When a publication reported training, internal-validation, and/or external-validation cohorts, these cohorts were extracted as separate datasets only if the original article indicated that the patients did not overlap.

XSL•FO
RenderX

Table . Study and patient characteristics of the included studies.

Author	Year	Country	Study design	Type of imaging	Temporal range	Reference standard	Analysis	Total patients or lesions, n			No. of VETC[a] patients or lesions
								Training	Internal validation	External validation	
Zhao et al [6]	2025	China	Retro[b]	CECT[c]	2014 - 2022	Pathology	PB[d]	169	72	152	• Training: 39 • Internal validation: 14 • External validation: 53
Xu et al [12]	2024	China	Retro	CEUS[e]	2018 - 2021	Pathology	PB	195	47	—[f]	• Training: 80 • Internal validation: 16
Matsuda et al [33]	2025	Japan	Retro	CEMRI	2008 - 2020	Pathology	PB	153	51	—	• Training: 33 • Internal validation: 12
Hu et al [8]	2025	China	Retro	PET/CT[g]	2015 - 2022	Pathology	PB	103	46	—	• Training: 55 • Internal validation: 23

XSL•FO
RenderX

Author	Year	Country	Study design	Type of imaging	Temporal range	Reference standard	Analysis	Total patients or lesions, n			No. of VETC[a] patients or lesions
								Training	Internal validation	External validation	
Gu et al [30]	2025	China	Retro	CEMRI	2015 - 2022	Pathology	PB	144	54	32	• Training: 73 • Internal validation: 28 • External validation: 16
Che et al [7]	2025	China	Retro	CEMRI	2014 - 2021	Pathology	PB	253	108	144	• Training: 82 • Internal validation: 34 • External validation: 54
Zhang et al [24]	2024	China	Retro	CECT	2017 - 2023	Pathology	PB	106	47	37	• Training: 53 • Internal validation: 23 • External validation: 18
Dong et al [31]	2023	China	Retro	DCE-MRI[i]	2015 - 2021	Pathology	PB	154	67	—	• Training: 46 • Internal validation: 22
Zhang et al [25]	2024	China	Retro	CEMRI[h]	2013 - 2020	Pathology	PB	163	71	—	

XSL•FO
RenderX

Author	Year	Country	Study design	Type of imaging	Temporal range	Reference standard	Analysis	Total patients or lesions, n			No. of VETC[a] patients or lesions
								Training	Internal validation	External validation	
											• Training: 71 • Internal validation: 30
Yang et al [27]	2024	China	Retro	DCE-MRI	2015 - 2021	Pathology	PB	219	—	101	• Training: 68 • External validation: 25
Qu et al [29]	2023	China	Retro	CEMRI	2014 - 2021	Pathology	PB	168	72	—	• Training: 74 • Internal validation: 31
Chu et al [32]	2022	China	Retro	CEMRI	2017 - 2020	Pathology	PB	93	40	—	• Training: 35 • Internal validation: 16
Yu et al [26]	2021	China	Retro	CEMRI	2015 - 2020	Pathology	PB	128	54	—	• Training: 72 • Internal validation: 25
Wang et al [28]	2025	China	Retro	CEMRI	2016 - 2023	Pathology	PB	305	—	115	• Training: 161 • External validation: 55
	2025	China	Retro	CECT		Pathology	PB	130	—	32	

XSL•FO
RenderX

Author	Year	Country	Study design	Type of imaging	Temporal range	Reference standard	Analysis	Total patients or lesions, n			No. of VETC[a] patients or lesions
								Training	Internal validation	External validation	
Wang et al [34]					2017 - 2022						• Training: 68 • External validation: 14

[a]VETC: vessels encapsulating tumor cluster.

[b]Retro retrospective.

[c]CECT: contrast-enhanced computed tomography.

[d]PB: patient-based.

[e]CEUS: contrast-enhanced ultrasonography.

[f]Not available.

[g]PET/CT: positron emission tomography/computed tomography.

[h]CEMRI: contrast-enhanced magnetic resonance imaging.

[i]DCE-MRI: dynamic contrast-enhanced magnetic resonance imaging.

XSL•FO
RenderX

Table . Technical aspects of included studies.

Author	Year	Center	Data source	AI[a] method	AI model	Optimal AI algorithms[b]	Data splitting method	Internal validation sets				External validation sets			
								TP[c]	FP[d]	FN[e]	TN[f]	TP	FP	FN	TN
Zhao et al [6]	2025	Multiple centers	Local hospital	ML[g]	R&CM[h]	LR[i]	Hold-out	12	17	2	41	40	21	13	78
Xu et al [12]	2024	Single center	Local hospital	DL[j]	RM[k]	CNN[l]	Hold-out	13	6	3	25	—[m]	—	—	—
Matsuda et al [33]	2025	Single center	Local hospital	ML	R&CM	LASSO[n]	Hold-out	10	9	2	30	—	—	—	—
Hu et al [8]	2025	Multiple centers	Local hospital	ML	R&CM	LR	Independent validation	—	—	—	—	17	6	6	17
Gu et al [30]	2025	Multiple centers	Local hospital	DL	R&CM	MLP[o]	Independent validation	—	—	—	—	24	4	4	22
Gu et al [30]	2025	Multiple centers	Local hospital	DL	R&CM	MLP	Independent validation	—	—	—	—	15	2	1	14
Che et al [7]	2025	Multiple centers	Local hospital	ML	R&CM	LR	Hold-out	26	21	8	53	43	27	11	63
Zhang et al [24]	2024	Multiple centers	Local hospital	ML	R&CM	LR	Hold-out	19	6	4	18	11	3	7	16
Dong et al [31]	2023	Single center	Local hospital	DL	RM	DNN[p]	Time-series split	17	8	5	37	—	—	—	—
Zhang et al [25]	2024	Single center	Local hospital	ML	R&CM	LR	Hold-out	30	12	0	29	—	—	—	—
Yang et al [27]	2024	Multiple centers	Local hospital	DL	RM	CNN	Independent validation	—	—	—	—	20	30	5	46
Qu et al [29]	2023	Single center	Local hospital	ML	R&CM	LR	Hold-out	27	15	4	26	—	—	—	—
Chu et al [32]	2022	Single center	Local hospital	DL	R&CM	CNN	Hold-out	7	0	9	24	—	—	—	—
Yu et al [26]	2021	Single center	Local hospital	ML	R&CM	RF[q]	Hold-out	25	4	0	25	—	—	—	—
Wang et al [28]	2025	Multiple centers	Local hospital	DL	R&CM	CNN	Independent validation	—	—	—	—	37	8	18	52

XSL•FO
RenderX

Author	Year	Center	Data source	AI[a] method	AI model	Optimal AI algorithms[b]	Data splitting method	Internal validation sets				External validation sets			
								TP[c]	FP[d]	FN[e]	TN[f]	TP	FP	FN	TN
Wang et al [34]	2025	Single center	Local hospital	ML	RM	RF	Hold-out	—	—	—	—	8	3	6	15

[a]AI: artificial intelligence.
[b]The optimal artificial intelligence algorithm means the algorithm with the highest area under the curve value.
[c]TP: true positive.
[d]FP: false positive.
[e]FN: false negative.
[f]TN: true negative.
[g]ML: machine learning.
[h]R&CM: radiomic and clinical model.
[i]LR: logistic regression.
[j]DL: deep learning.
[k]RM: radiomic model.
[l]CNN: convolutional neural network.
[m]Not available.
[n]LASSO: least absolute shrinkage and selection operator.
[o]MLP: multilayer perceptron.
[p]DNN: deep neural network.
[q]RF: random forest.

The methodological quality of the included studies was appraised using the PROBAST+AI tool, with the detailed signaling-question tables provided in Tables S2 and S3 in Multimedia Appendix 1. For model development, the overall quality judgment was high-risk in 6.7% (1/15) of studies and low-risk in 93.3% (14/15). Applicability concerns for model development were high in 13.3% (2/15) of studies and low in 86.7% (13/15). For model validation, the overall risk of bias was high in 20% (3/15) of studies, unclear in 20% (3/15), and low in 60% (9/15). Applicability concerns for validation cohorts were low in all included studies. The main high-risk judgments were driven by analysis-related issues, whereas unclear judgments mainly reflected insufficient reporting of outcome assessment. These risk-of-bias findings are presented in the main manuscript together with the PRISMA flow diagram (Figure 1) and the prognostic forest plot (Figure 2), with full details retained in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 2. Rescaled forest plot of hazard ratios for early recurrence risk (early recurrence as defined by the original studies, eg, ≤1 or ≤2 years post hepatectomy or transplantation) in patients with hepatocellular carcinoma stratified by artificial intelligence–predicted vessels encapsulating tumor cluster status across multiple imaging modalities. The x-axis is displayed on a logarithmic scale with an adjusted range to improve visualization of CIs and the prediction interval. Rows from the same article indicate mutually exclusive training, internal-validation, or external-validation cohorts when reported. HKSJ: Hartung-Knapp-Sidik-Jonkman; HR: hazard ratio; REML: restricted maximum likelihood [6,7,12,25-27,31].

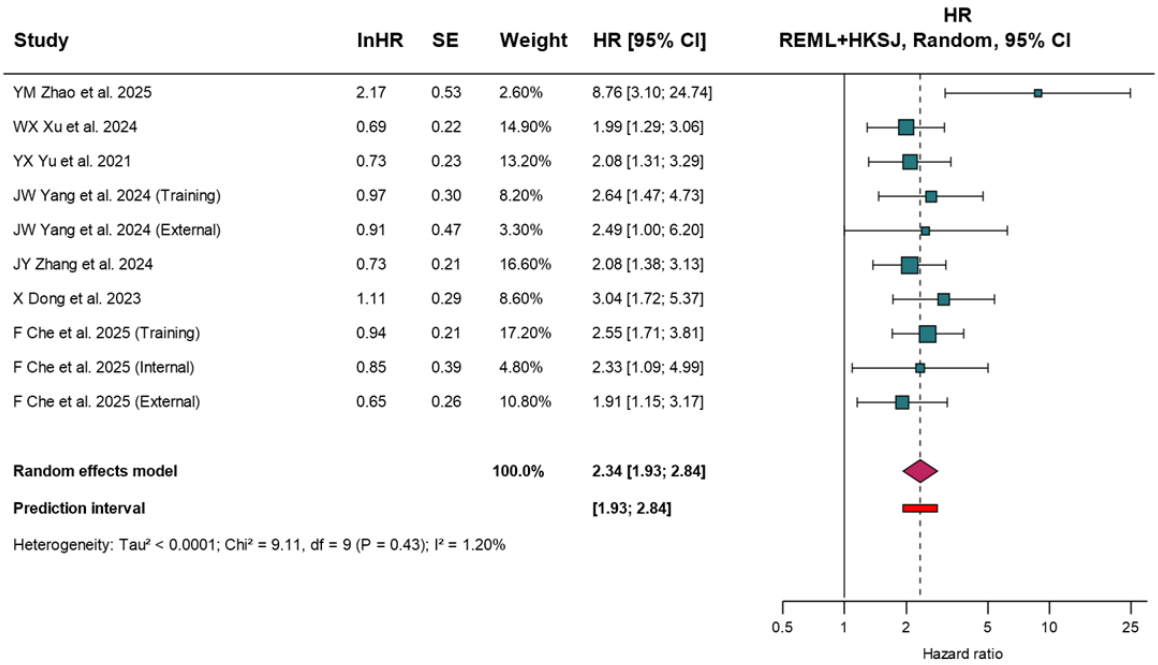

The certainty of evidence was evaluated according to the GRADE framework and is now summarized in a main-text diagnostic Summary of Findings table. The table presents expected TP, FN, TN, and FP results per 1000 patients tested at the median VETC prevalence of 38%, together with the number of participants and studies, certainty ratings, and concise comments. Detailed domain judgments remain in Table S4 in Multimedia Appendix 1. Across outcomes, certainty ranged from low to moderate, with downgrading mainly driven by validation-related risk of bias, imprecision, and heterogeneity.

The Predictive Performance and Temporal Development Trends of Different AI Algorithms

As illustrated in Figure 3B, the multilayer perceptron (MLP) algorithm demonstrated the highest sensitivity and specificity values at 0.89 (95% CI 0.26‑1.00) and 0.86 (95% CI 0.71‑0.97). Furthermore, Figure 3C indicates a slight decline in the DOR from 2021 to 2025, with random forest (RF) achieving relatively high DOR values while other algorithms maintained lower levels.

XSL•FO
RenderX

Figure 3. Comprehensive visualization of the interactions between imaging modalities and artificial intelligence (AI) algorithms, and the diagnostic performance of AI models for predicting vessels encapsulating tumor clusters in hepatocellular carcinoma. It includes: (A) a chord diagram illustrating the interconnections between 4 imaging modalities (contrast-enhanced magnetic resonance imaging, contrast-enhanced computed tomography, contrast-enhanced ultrasound, and 2-[¹ F]fluoro-2-deoxy-D-glucose positron emission tomography or computed tomography) and 9 AI algorithms; (B) violin plots comparing the predictive performance (area under the curve) of different AI algorithms; (C) a bubble plot showing the diagnostic odds ratio of the optimal-performing AI algorithm stratified by publication year; and (D) violin plots summarizing the diagnostic performance (sensitivity and specificity) of various imaging modalities. AI: artificial intelligence; AUC: area under the curve; CECT: contrast-enhanced computed tomography; CEMRI: contrast-enhanced magnetic resonance imaging; CEUS: contrast-enhanced ultrasound; CNN: convolutional neural network; DNN: deep neural network; LASSO: least absolute shrinkage and selection operator; LR: logistic regression; MLP: multilayer perceptron; RF: random forest.

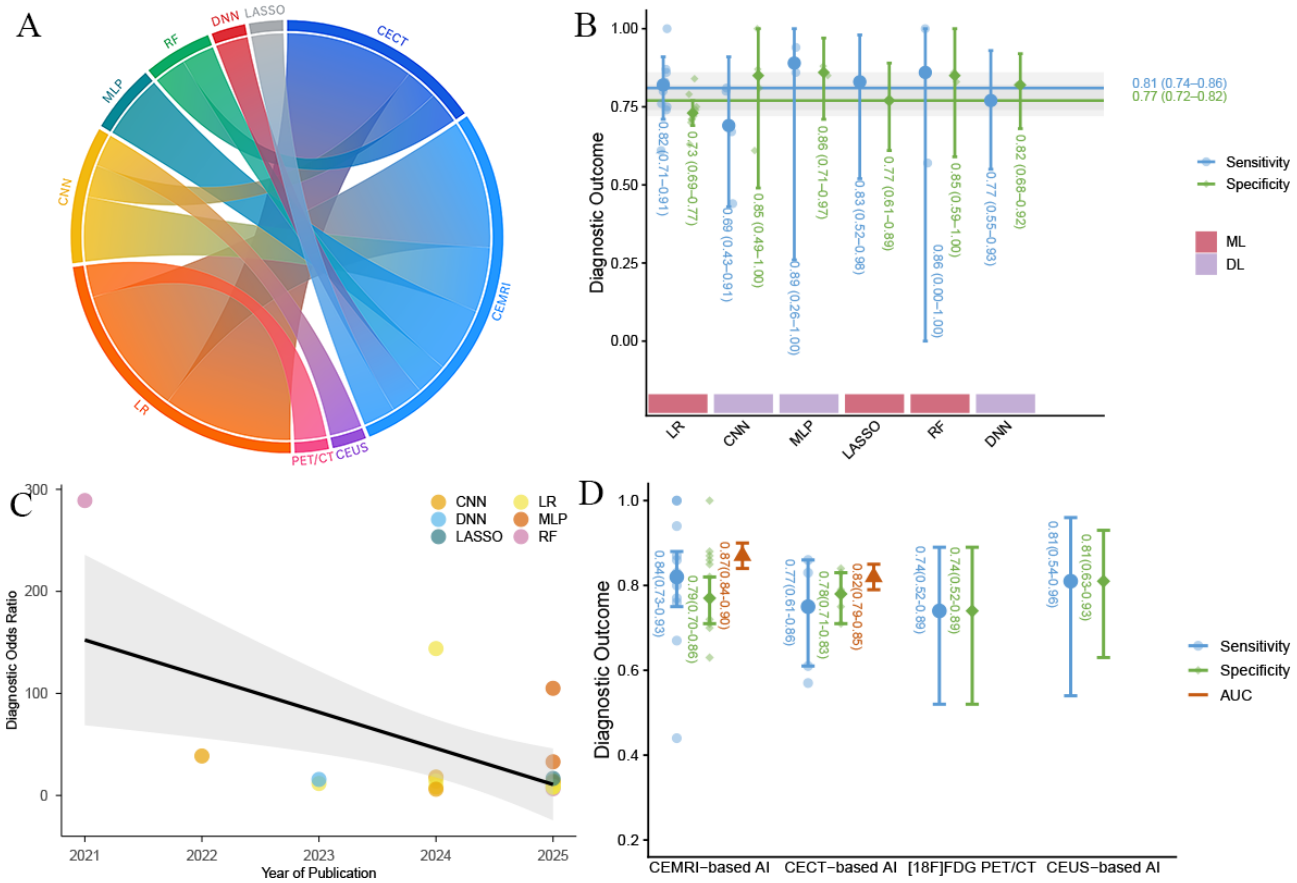

The Predictive Performance of Different Imaging-Based AI Model

The results indicated that no significant threshold effect was observed for AI models based on CEMRI (Spearman ρ=−0.03, *P*=.93). The primary pooled analysis combined both internal and external validation datasets. For CEMRI-based AI models, the average pooled sensitivity was 0.84 (95% CI 0.73 - 0.93; 95% PI 0.45 - 1.00; τ^2=0.03) with low certainty, and the average pooled specificity was 0.79 (95% CI 0.70 - 0.86; 95% PI 0.50 - 0.97; τ^2=0.02) with low certainty. The 95% CIs describe uncertainty around the average pooled estimates, whereas the 95% PIs describe how accuracy may be distributed in a future setting; therefore, the wide CEMRI PIs indicate clinically important between-setting variability despite favorable average accuracy. For CECT-based AI models, sensitivity was 0.75 (95% CI 0.61 - 0.86; 95% PI 0.62 - 0.85; τ^2<0.001) with moderate certainty, and specificity was 0.78 (95% CI 0.71 - 0.83; 95% PI 0.69 - 0.85; τ^2=0) with moderate certainty. For [^{18}F]FDG PET/CT and CEUS, only 1 study was available for each modality; therefore, their average estimates were not used to infer a distribution of effects across settings. The [^{18}F]FDG PET/CT-based AI model had sensitivity and specificity of 0.74 (95% CI 0.52 - 0.90), both with low certainty. The CEUS-based AI model had sensitivity of 0.81 (95% CI 0.54 - 0.96) and specificity of 0.81 (95% CI 0.63 - 0.93), both with low certainty.

The Predictive Value of Early Recurrence

The GRADE summary of findings, including the diagnostic summary of findings estimates and expected results per 1000 patients tested, is presented in Table 3.

As illustrated in Figure 2, regarding the prognostic value for predicting early recurrence, the pooled analysis of 10 nonoverlapping datasets from 7 studies [6,7,12,25-27,31] revealed an average HR of 2.34 (95% CI 1.93 - 2.84) with the REML+HKSJ method. The CI reflects uncertainty around the average prognostic effect. Because between-study heterogeneity was very low (τ^2<0.0001), the PI was similar to the CI, suggesting limited observed dispersion for the pooled HR; however, modality-specific HRs from single studies should not be interpreted as distributions across settings. In this survival-HR forest plot, Che et al [7] contributed 3 mutually exclusive patient cohorts (training, internal validation, and external validation), and Yang et al [27] contributed 2 mutually exclusive cohorts (training and external validation). Because these patient groups

XSL•FO
RenderX

did not overlap, they were treated as separate datasets and did not constitute double-counting. When stratified by imaging modality, the HR was 2.32 (95% CI 2.04 - 2.63; $\tau^2=0$) for CEMRI-based AI (8 datasets from 5 studies [7,25-27,31]), 8.76 (95% CI 3.10 - 24.74) for CECT-based AI (1 study [6]), and 1.99 (95% CI 1.29 - 3.06) for CEUS-based AI (1 study [12]). The extremely high HR observed in the CECT subgroup should be interpreted cautiously, as it was derived from only a single study with a wide CI. Similarly, the CEUS-based estimate was based on a single study.

XSL•FO
RenderX

Table . GRADE[a] summary of findings table for AI[b]-based imaging prediction of VETC[c] in HCC[d]. Expected results were calculated per 1000 patients tested at the median VETC prevalence of 38% and rounded to the nearest whole patient. Certainty was judged using the GRADE framework for diagnostic test accuracy evidence; downgrading reflected validation-related risk of bias, imprecision, heterogeneity or inconsistency, limited external validation, and small-study or single-study evidence where applicable. Prediction intervals were not estimable for modalities represented by a single study.

Insex test and test result	Number of results per 1000 patients tested (95% CI), at 38% VETC prevalence	Participants and studies or datasets	Certainty of the evidence (GRADE)	Comments
CEMRI[e]-based AI[b]		909 participants (10 studies or datasets)	Low	Favorable average accuracy, but certainty was downgraded for validation-related risk of bias, imprecision, and heterogeneity/inconsistency; prediction intervals were wide.
True positives	319 (277-353)			
False negatives	61 (27-103)			
True negatives	490 (434-533)			
False positives	130 (87-186)			
CECT[f]-based AI		340 participants (3 studies or datasets)	Moderate	Average accuracy was moderate-to-good; certainty was mainly limited by imprecision and limited external validation.
True positives	285 (232-327)			
False negatives	95 (53-148)			
True negatives	484 (440-515)			
False positives	136 (105-180)			
[^{18}F]FDG PET/CT[g]-based AI		46 participants (1 study or dataset)	Low	Single-study evidence was preliminary; certainty was downgraded for imprecision and limited validation, and prediction intervals were not estimable.
True positives	281 (198-342)			
False negatives	99 (38-182)			
True negatives	459 (322-558)			
False positives	161 (62-298)			
CEUS[h]-based AI		47 participants (1 study or dataset)	Low	Single-study evidence suggested potential accuracy but requires independent validation; prediction intervals were not estimable.
True positives	308 (205-365)			
False negatives	72 (15-175)			
True negatives	502 (391-577)			
False positives	118 (43-229)			

[a]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.
[b]AI: artificial intelligence.
[c]VETC: vessels encapsulating tumor cluster.
[d]HCC: hepatocellular carcinoma.
[e]CEMRI: contrast-enhanced magnetic resonance imaging.
[f]CECT: contrast-enhanced computed tomography.
[g][^{18}F]FDG PET/CT: fluorine-18 fluorodeoxyglucose positron emission tomography/computed tomography.
[h]CEUS: contrast-enhanced ultrasonography.

XSL•FO
RenderX

Meta-Regression and Bivariate Box Plots

The results of the meta-regression are illustrated in Table 4 and indicate that the heterogeneity in CEMRI-based AI models was primarily attributable to variations in center (specificity: *P*=.01); validation (specificity: *P*=.01); AI algorithms (specificity: *P*<.001); MRI field strength (specificity: *P*=.05). Specificity was significantly higher in single-center studies compared with multicenter studies (0.81 vs 0.75, *P*=.01). Similarly, internal validation cohorts demonstrated significantly higher specificity than external validation cohorts (0.81 vs 0.75, *P*=.01). Deep learning models achieved significantly higher specificity than machine learning models (0.83 vs 0.73, *P*<.001). Furthermore, studies using a 3.0T MRI field strength showed significantly higher specificity than those using a combination of 1.5T and 3.0T scanners (0.79 vs 0.77, *P*=.05).

it The bivariate box plot analysis identified the studies by Yu et al [26], Zhang et al [25], Yang et al [27], and Chu et al [32] as outliers, suggesting they may be potential sources of heterogeneity (Figure 4A).

Table . Subgroup analysis and meta-regression analysis of CEMRI[a]-based AI[b] performance.

Subgroup	Studies, n	Sensitivity (95%CI)	Meta-regression *P* value	Specificity (95%CI)	Meta-regression *P* value
Center			.06		.01
Single center	6	0.87 (0.77 - 0.96)		0.81 (0.72 - 0.90)	
Multicenter	6	0.81 (0.69 - 0.92)		0.75 (0.65 - 0.85)	
Validation			.06		.01
Internal validation	6	0.87 (0.77 - 0.96)		0.81 (0.72 - 0.90)	
External validation	6	0.81 (0.69 - 0.92)		0.75 (0.65 - 0.85)	
Data splitting			.31		.18
Hold out splitting	7	0.86(0.75,0.96)		0.77 (0.68 - 0.86)	
Independent validation	4	0.83 (0.68 - 0.98)		0.79 (0.67 - 0.91)	
AI[b] model			.95		.56
Radiomic	2	0.80 (0.57 - 1.00)		0.71 (0.55 - 0.88)	
Radiomic and clinical	10	0.85 (0.76 - 0.93)		0.79 (0.72 - 0.86)	
AI algorithms			.98		<.001
Deep learning	6	0.77 (0.64 - 0.89)		0.83 (0.75 - 0.91)	
Machine learning	6	0.89 (0.82 - 0.97)		0.73 (0.64 - 0.82)	
MRI[c] field strength			.37		.05
1.5 T and 3.0 T	5	0.85 (0.73 - 0.97)		0.77 (0.67 - 0.88)	
3.0 T	7	0.83 (0.72 - 0.94)		0.79 (0.70 - 0.87)	

[a]CEMRI: contrast-enhanced magnetic resonance imaging.

[b]AI: artificial intelligence.

[c]MRI: magnetic resonance imaging.

XSL•FO
RenderX

Figure 4. Heterogeneity evaluation, small effect test, and clinical utility analysis of artificial intelligence models for predicting vessels encapsulating tumor clusters in hepatocellular carcinoma. (A) Bivariate boxplot of sensitivity and specificity, visualizing the distribution and heterogeneity of study-level estimates. (B) Deek's funnel plot for the assessment of potential small-study effects. (C) Fagan plot for Bayesian analysis of clinical utility, using the median vessels encapsulating tumor cluster prevalence (38%) across included studies as the prior probability. AI: artificial intelligence; CEMRI: contrast-enhanced magnetic resonance imaging; LR: logistic regression [25-27,32].

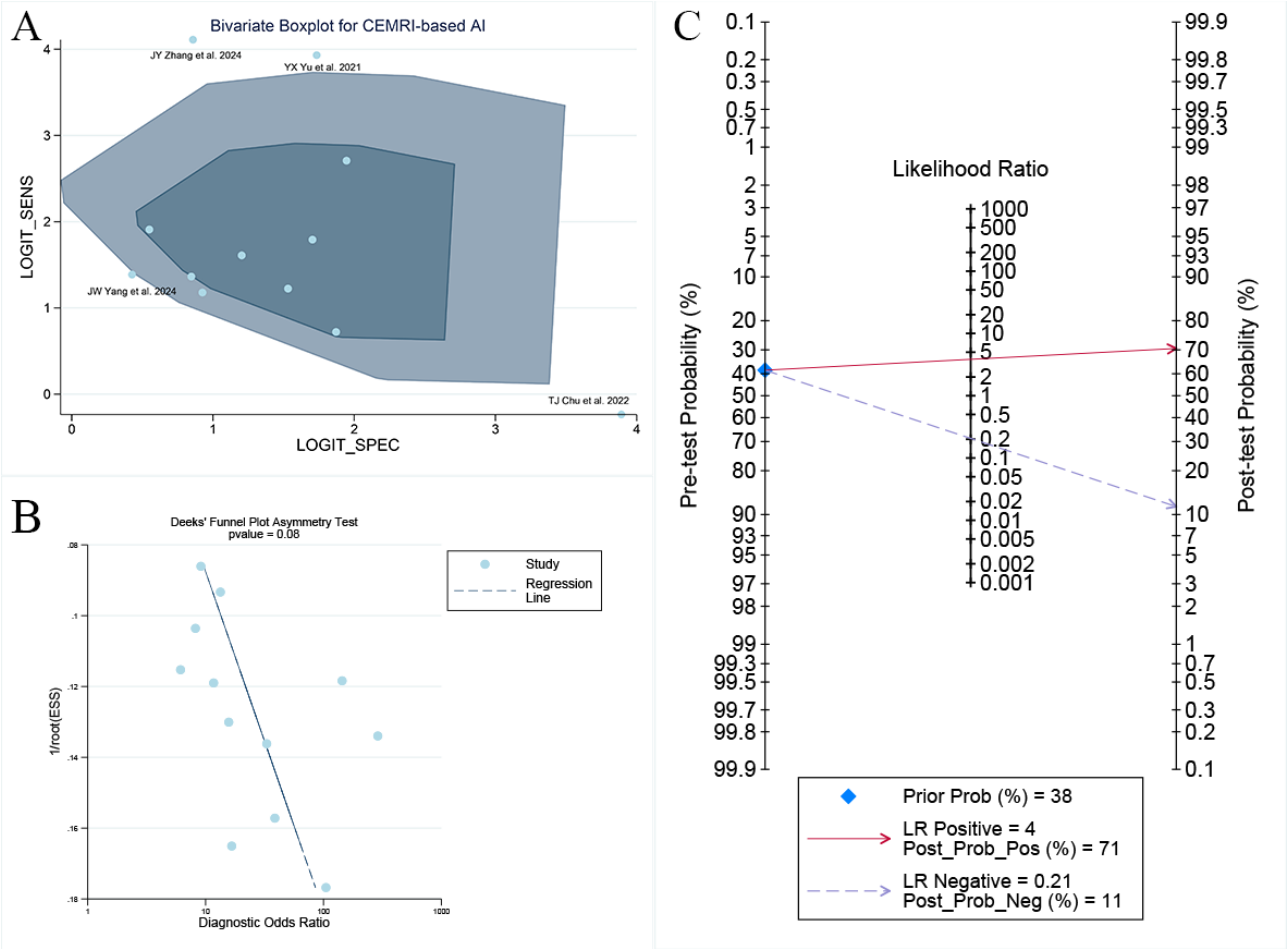

Sensitivity Analysis

Sensitivity analysis was performed by excluding the 4 outlier studies [25-27,32] to assess their impact on the pooled estimates for CEMRI-based AI models. After exclusion, the pooled sensitivity was 0.80 (95% CI 0.73 - 0.87), specificity was 0.77 (95% CI 0.70 - 0.84), and AUC was 0.85 (95% CI 0.82 - 0.88), showing no substantial changes compared with the primary results. Additionally, a subgroup analysis was conducted by including only pure radiomics studies. In this analysis, the pooled sensitivity, specificity, and AUC for CEMRI-based models were 0.81 (95% CI 0.68 - 0.92), 0.68 (95% CI 0.50 - 0.84), and 0.75 (95% CI 0.71 - 0.79), indicating a slight decrease in diagnostic performance. We also performed an analysis restricted to studies with external validation. In this analysis, the pooled sensitivity, specificity, and AUC of CEMRI-based models were 0.80 (95% CI 0.67 - 0.90), 0.77 (95% CI 0.61 - 0.90), and 0.84 (95% CI 0.80 - 0.87), respectively.

Small-Study Effects and Clinical Application Value

Deeks' funnel plot asymmetry test indicated no statistically significant small-study effects for CEMRI-based AI models (P=.08; Figure 4B). Furthermore, using the prespecified median prevalence of 38% as the pretest probability, the Fagan nomogram demonstrated a positive posttest probability of 71% and a negative posttest probability of 11% (Figure 4C).

Discussion

Main Findings

Our meta-analysis demonstrated that AI models based on CEMRI yielded the highest sensitivity (0.84) and AUC (0.87) for predicting VETC in HCC, while models based on CEUS achieved the highest specificity (0.81). The superior sensitivity and AUC of CEMRI-based AI can be attributed to the ability of CEMRI to provide high soft-tissue contrast and detailed visualization of intratumoral heterogeneity and peritumoral vasculature [26]. This allows AI models to extract a large number of high-dimensional radiomic features, potentially maximizing the capture of subtle image phenotypes associated with the VETC pattern [35]. Conversely, the high specificity of CEUS-based AI likely stems from its capacity for real-time assessment of tumor hemodynamics and microvascular perfusion [36]. AI models can leverage this dynamic information to accurately identify the characteristic perfusion patterns of VETC-positive HCCs, thereby reducing false-positive diagnoses [37]. Furthermore, the typical enhancement patterns of non-VETC HCCs on CEUS may allow AI models to more

XSL•FO
RenderX

reliably exclude VETC-negative cases [12]. However, as this finding was based on a single study, it should be interpreted with caution and requires validation in larger, independent cohorts.

Detailed Discussion and Comparison

Regarding the impact of different algorithms, our subgroup analysis revealed that the RF algorithm demonstrated the highest sensitivity (1.00), while both MLP and RF achieved the highest specificity values (0.86). The robust performance of RF, characterized by high sensitivity and specificity, can be explained by its ensemble nature. By constructing a multitude of decision trees and aggregating their predictions, RF effectively mitigates overfitting, often seen in single models, thereby enhancing its ability to correctly identify both positive and negative cases [38]. Furthermore, RF's inherent capability to automatically assess feature importance allows it to select the most discriminative radiomics features for classification, improving its adaptability and discriminatory power when dealing with complex, high-dimensional medical imaging data [39]. However, the perfect sensitivity of 1.00 reported for the RF model should be interpreted with caution, as this finding may reflect study-specific characteristics, such as small sample sizes, overfitting, or the use of internal validation only. The high specificity attained by MLP, an artificial neural network, likely stems from its capacity to model complex, nonlinear relationships within the data. When the feature distribution of negative (non-VETC) samples is distinct, MLP can leverage its multiple hidden layers to learn highly specific decision boundaries, effectively separating them from positive cases and resulting in a low false-positive rate [40]. However, conclusions regarding MLP performance remain preliminary due to the limited number of studies.

Our pooled prognostic analysis, based on 10 nonoverlapping datasets from 7 studies [6,7,12,25-27,31], indicated that patients with HCC classified as VETC-positive by AI models had a significantly higher risk of early recurrence, with an HR of 2.34 (95% CI 1.93‑2.84). This suggests that AI-based VETC prediction offers prognostic capability for early recurrence risk. Notably, the AI model based on CECT yielded the highest HR of 8.76, although its wide 95% CI (3.10‑24.74) warrants caution in interpretation, likely attributable to the limited number of studies available for this modality. Given the small number of contributing studies, this estimate may be unstable. Previous research findings indicated that in the pathological classification, the VETC-positive group had an HR (95% CI 1.44‑2.07) for early recurrence [21], and the results of this study suggest that the AI model's predictive capability for early recurrence risk is not inferior to that of traditional pathological risk classification.

A recent meta-analysis by Wang et al [41] specifically evaluated the performance of MRI-based radiomics for predicting VETC in HCC, reporting a sensitivity of 0.88, specificity of 0.86, and an AUC of 0.93. This meta-analysis, which included a broader scope of imaging modalities, found a comparatively lower pooled predictive performance for CEMRI-based AI. This discrepancy may be explained by our inclusion of a larger number of studies investigating MRI-based AI models, which provides a more generalizable, albeit slightly less optimistic, estimate of real-world performance. Our study extends the findings of Wang et al [41] by providing a comprehensive comparative analysis of AI performance across multiple imaging modalities, including CEUS, CECT, and [^{18}F] FDG PET/CT, in addition to CEMRI. We further investigated the influence of different AI algorithms on predictive performance. Most significantly, our work provides a novel contribution by quantitatively synthesizing the prognostic value of AI-predicted VETC status, demonstrating that the AI model exhibits good prognostic capability for predicting early recurrence.

Significant heterogeneity was observed among the included studies, which is an acknowledged challenge in radiomics meta-analyses. We accounted for this inherent variability a priori by using a bivariate random-effects hierarchical summary ROC model with HKSJ adjustment for diagnostic metrics, and a prespecified REML+HKSJ random-effects model for the prognostic HR synthesis, with DerSimonian-Laird and uncorrected REML reported only as sensitivity analyses. Meta-regression and bivariate boxplot analysis identified several potential sources of this heterogeneity. The significantly higher specificity in single-center versus multicenter studies (0.81 vs 0.75, P=.01) may reflect the greater standardization in imaging protocols and patient populations within a single institution, whereas multicenter studies encompass variations in scanners, acquisition parameters, and patient demographics that can challenge model generalizability [42]. The higher specificity in internal validation cohorts compared with external validation sets (0.81 vs 0.75, P=.01) underscores the performance drop commonly encountered when models are applied to independent, external data with different distributions [43]. The superior specificity of deep learning models over machine learning models (0.83 vs 0.73, P<.001) may be attributed to their ability to automatically learn complex, hierarchical feature representations directly from images, potentially capturing more discriminative patterns to rule out negative cases [44]. Furthermore, studies using 3.0T MRI scanners showed higher specificity than those using a mix of 1.5T/3.0T devices (0.79 vs 0.77, P=.05), likely due to the superior signal-to-noise ratio and spatial resolution of 3.0T, leading to more consistent and detailed image data [45]. The bivariate boxplot also pinpointed specific studies as potential outliers contributing to heterogeneity [24,27,28,33].

Future Direction

The findings of this meta-analysis hold several implications for clinical practice. The identified strengths of specific imaging-modality and algorithm combinations—such as the high sensitivity of CEMRI-based AI and the high specificity of CEUS-based AI—could inform the development of noninvasive tools for preoperative VETC prediction and risk stratification [46]. The demonstrated prognostic value of AI-classified VETC status for early recurrence, which was comparable with pathological classification, suggests its potential role in guiding postoperative surveillance and adjuvant therapy decisions, thereby facilitating timely management of HCC [47]. It is crucial to emphasize that any implemented AI should function as a decision-support tool, augmenting rather than replacing clinician judgment [48]. A key consideration for clinical implementation

XSL•FO
RenderX

is model generalizability; the fact that only 7 included studies [6,7,24,27,28,30,34] reported external validation performance underscores the need for future research to prioritize robust, multicenter external validation [42]. Current unimodal AI models also have inherent limitations, as they may not capture the comprehensive information available from a multiparametric imaging workup [43]. Future AI systems may need to integrate findings across multiple imaging modalities, correlate them with clinical context, and communicate synthesized insights [44]. Other barriers to widespread adoption include the scarcity of expert-annotated data, regulatory hurdles, and challenges regarding model interpretability and transparency [45]. The "black-box" nature of many complex models remains a significant obstacle to clinical trust and integration [49]. Advances in techniques, like few-shot learning, self-supervised models, and explainable AI, alongside collaborative platforms, are needed to build a robust and trustworthy AI ecosystem in radiology [50]. AI-assisted VETC prediction may be particularly beneficial for high-risk surgical candidates, such as patients with large tumors, borderline resectable disease, or elevated AFP levels [51].

Limitations

Several limitations of this meta-analysis should be acknowledged when interpreting the results. First, the predominantly retrospective design and relatively small sample sizes of the included studies may introduce potential biases in patient selection and data collection. Therefore, large-scale, prospectively designed studies are warranted to validate these findings [52]. Second, to avoid potential patient overlap across multiple algorithms reported in a single study, we extracted data only for the best-performing algorithm from each publication. While this approach mitigates bias, it prevents a comprehensive evaluation of the performance spectrum across all developed algorithms and may present an overly optimistic view; future studies with independent test sets for all models are needed [53]. Third, the generalizability of our conclusions may be limited by the geographical distribution of the included literature. Specifically, 14 out of the 15 included studies (93.3%) were conducted in China, representing a significant geographic concentration within Asian populations. While these findings are robust for the studied cohorts, validation in more ethnically and geographically diverse populations outside of China is necessary to ensure the widespread applicability and diagnostic stability of these AI models across different health care systems and genetic backgrounds [54]. Fourth, the average pooled diagnostic effects should be distinguished from the distribution of effects across settings. The 95% CIs describe uncertainty around the average sensitivity or specificity, whereas the 95% PIs describe how performance may vary in future settings. For example, the CEMRI sensitivity CI was 0.73 - 0.93, but the PI was 0.45 - 1.00, indicating that substantial heterogeneity may remain despite favorable average performance, as recommended by Borenstein [55] for the interpretation of heterogeneity. Fifth, the GRADE certainty of evidence was low for most outcomes and moderate only for CECT sensitivity and specificity, mainly because of imprecision, validation-related risk of bias, and limited external validation. Given these limitations, the pooled results should be interpreted cautiously to avoid overstatement. Future studies with standardized imaging protocols, unified reporting guidelines, prospective designs, and geographically diverse external validation are needed to reduce heterogeneity and improve certainty. Finally, differences in the definition of early recurrence across studies may have influenced the pooled HR estimates [56], and other potential sources of heterogeneity—such as variations in patient demographics, tumor stage, and imaging acquisition protocols—may also contribute to the observed variability.

Conclusion

In conclusion, the available evidence indicates that AI models derived from contrast-enhanced imaging can identify VETC status with clinically meaningful average accuracy, particularly for CEMRI, and that model-predicted VETC positivity is associated with a higher risk of early recurrence. The main contribution of this synthesis is to define the current evidentiary boundary for preoperative VETC risk stratification; AI outputs may help clinicians recognize tumors with aggressive vascular architecture, select patients for intensified follow-up, and generate hypotheses for individualized perioperative planning. These applications remain conditional. Most included cohorts were retrospective and geographically concentrated, external validation was sparse, and uncertainty was amplified by heterogeneity, imprecision, and low-to-moderate GRADE certainty. Before routine clinical adoption, future multicenter prospective studies should use prespecified thresholds, standardized image acquisition and reporting, independent external validation, calibration assessment, and evaluations of whether AI-informed decisions improve patient outcomes. Until such evidence is available, AI-based VETC prediction should be used as investigational support within multidisciplinary assessment rather than as a standalone basis for management.

Acknowledgments

During the preparation of this work, the authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing and translation. The GenAI tool used was: DeepSeek-V3.2. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes (declaration submitted by: ZY).

Funding

This study is supported by Guangzhou Science and Technology Program (Project 2023A03J0952 and Project 2023A03J0953).

XSL•FO
RenderX

Data Availability

The original findings of this study are encompassed within the article. For additional inquiries, please contact the corresponding authors.

Authors' Contributions

XH: conceptualization, formal analysis, software, writing – original draft, final approval.
RF and ZY: data curation, formal analysis, investigation, methodology, writing – review & editing.
All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables.
[DOCX File, 52 KB - jmir_v28i1e90931_app1.docx]

Checklist 1
PRISMA-DTA checklist.
[PDF File, 180 KB - jmir_v28i1e90931_app2.pdf]

Checklist 2
CHARMS checklist.
[DOCX File, 21 KB - jmir_v28i1e90931_app3.docx]

Checklist 3
PRISMA checklist for abstracts.
[DOCX File, 17 KB - jmir_v28i1e90931_app4.docx]

Checklist 4
PRISMA-S checklist.
[DOCX File, 18 KB - jmir_v28i1e90931_app5.docx]

References

1. Singal AG, Kanwal F, Llovet JM. Global trends in hepatocellular carcinoma epidemiology: implications for screening, prevention and therapy. Nat Rev Clin Oncol 2023 Dec;20(12):864-884. [doi: 10.1038/s41571-023-00825-3] [Medline: 37884736]
2. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74(3):229-263. [doi: 10.3322/caac.21834] [Medline: 38572751]
3. Zhu Y, Wang M, Cao J, Song H, Zhang P. Vessels encapsulating tumor clusters in hepatocellular carcinoma: a distinct metastatic pathway with diagnostic and therapeutic significance. J Transl Med 2026 Jan 19;24(1):185. [doi: 10.1186/s12967-025-07354-w] [Medline: 41555423]
4. Wang YY, Dong K, Wang K, Sun Y, Xing BC. Effect of vessels that encapsulate tumor clusters (VETC) on the prognosis of different stages of hepatocellular carcinoma after hepatectomy. Dig Liver Dis 2023 Sep;55(9):1288-1294. [doi: 10.1016/j.dld.2023.03.008] [Medline: 37037766]
5. Xiong SP, Wang CH, Zhang MF, Yang X, Yun JP, Liu LL. A multi-parametric prognostic model based on clinicopathologic features: vessels encapsulating tumor clusters and hepatic plates predict overall survival in hepatocellular carcinoma patients. J Transl Med 2024 May 18;22(1):472. [doi: 10.1186/s12967-024-05296-3] [Medline: 38762511]
6. Zhao YM, Xie SS, Wang J, et al. VETC predicting model based on CE-CT can predict prognosis and assisting treatment plan for solitary HCC: better together with radiomics. BMC Cancer 2025 Jul 1;25(1):1033. [doi: 10.1186/s12885-025-14408-1] [Medline: 40596967]
7. Che F, Gao F, Li Q, et al. Fractal analysis based on Gd-EOB-DTPA-enhanced MRI for prediction of vessels that encapsulate tumor clusters in patients with hepatocellular carcinoma. Int J Surg 2025 Jul 1;111(7):4389-4399. [doi: 10.1097/JS9.0000000000002547] [Medline: 40441719]

XSL•FO
RenderX

8. Hu S, Zou Q, Shen Z, et al. Preliminary study on the ability of 18F-fluorodeoxyglucose positron emission tomography/computed tomography radiomics to predict vessels that encapsulate tumor clusters and prognosis in hepatocellular carcinoma. Quant Imaging Med Surg 2025 Jul 1;15(7):6217-6233. [doi: 10.21037/qims-2024-2734] [Medline: 40727341]
9. Liu K, Dennis C, Prince DS, et al. Vessels that encapsulate tumour clusters vascular pattern in hepatocellular carcinoma. JHEP Rep 2023 Aug;5(8):100792. [doi: 10.1016/j.jhepr.2023.100792] [Medline: 37456680]
10. Gama H, Albuquerque R, Campos Wanderley D, et al. Cd34 immunostaining adds specificity to microvascular invasion analysis in hepatocellular carcinoma. J Hepatocell Carcinoma 2023;10(91-8):91-98. [doi: 10.2147/JHC.S389836] [Medline: 36711001]
11. Pan J, Huang H, Zhang S, et al. Intraindividual comparison of CT and MRI for predicting vessels encapsulating tumor clusters in hepatocellular carcinoma. Eur Radiol 2025 Jan;35(1):61-72. [doi: 10.1007/s00330-024-10944-9] [Medline: 38992109]
12. Xu W, Zhang H, Zhang R, et al. Deep learning model based on contrast-enhanced ultrasound for predicting vessels encapsulating tumor clusters in hepatocellular carcinoma. Eur Radiol 2025 Feb;35(2):989-1000. [doi: 10.1007/s00330-024-10985-0] [Medline: 39066894]
13. Xu H, Wang H, Yu SZ, et al. Prognostic and diagnostic value of [18F]FDG, 11C-acetate, and [68Ga]Ga-FAPI-04 PET/CT for hepatocellular carcinoma. Eur Radiol 2025 Jul;35(7):4121-4131. [doi: 10.1007/s00330-025-11352-3] [Medline: 39838091]
14. Chen FM, Du M, Qi X, et al. Nomogram estimating vessels encapsulating tumor clusters in hepatocellular carcinoma from preoperative gadoxetate disodium-enhanced MRI. J Magn Reson Imaging 2023 Jun;57(6):1893-1905. [doi: 10.1002/jmri.28488] [Medline: 36259347]
15. Xia T, Zhao B, Li B, et al. MRI-based radiomics and deep learning in biological characteristics and prognosis of hepatocellular carcinoma: opportunities and challenges. J Magn Reson Imaging 2024 Mar;59(3):767-783. [doi: 10.1002/jmri.28982] [Medline: 37647155]
16. Jiang X, Hu Z, Wang S, Zhang Y. Deep learning for medical image-based cancer diagnosis. Cancers (Basel) 2023 Jul 13;15(14):3608. [doi: 10.3390/cancers15143608] [Medline: 37509272]
17. Zhang YP, Zhang XY, Cheng YT, et al. Artificial intelligence-driven radiomics study in cancer: the role of feature engineering and modeling. Mil Med Res 2023 May 16;10(1):22. [doi: 10.1186/s40779-023-00458-8] [Medline: 37189155]
18. Yang Y, Zhang H, Gichoya JW, Katabi D, Ghassemi M. The limits of fair medical imaging AI in real-world generalization. Nat Med 2024 Oct;30(10):2838-2848. [doi: 10.1038/s41591-024-03113-4] [Medline: 38942996]
19. Fountzilas E, Pearce T, Baysal MA, Chakraborty A, Tsimberidou AM. Convergence of evolving artificial intelligence and machine learning techniques in precision oncology. NPJ Digit Med 2025 Jan 31;8(1):75. [doi: 10.1038/s41746-025-01471-y] [Medline: 39890986]
20. Pellat A, Barat M, Coriat R, Soyer P, Dohan A. Artificial intelligence: a review of current applications in hepatocellular carcinoma imaging. Diagn Interv Imaging 2023 Jan;104(1):24-36. [doi: 10.1016/j.diii.2022.10.001] [Medline: 36272931]
21. Wang M, Cao L, Wang Y, Huang H, Tian X, Lei J. The prognostic value of vessels encapsulating tumor clusters (VETC) in patients with hepatocellular carcinoma: a systematic review and meta-analysis. Clin Transl Oncol 2024 Aug;26(8):2037-2046. [doi: 10.1007/s12094-024-03427-2] [Medline: 38523240]
22. Salameh JP, Bossuyt PM, McGrath TA, et al. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. BMJ 2020 Aug 14;370:m2632. [doi: 10.1136/bmj.m2632] [Medline: 32816740]
23. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ 2025 Mar 24;388:e082505. [doi: 10.1136/bmj-2024-082505] [Medline: 40127903]
24. Zhang C, Zhong H, Zhao F, Ma ZY, Dai ZJ, Pang GD. Preoperatively predicting vessels encapsulating tumor clusters in hepatocellular carcinoma: machine learning model based on contrast-enhanced computed tomography. World J Gastrointest Oncol 2024 Mar 15;16(3):857-874. [doi: 10.4251/wjgo.v16.i3.857] [Medline: 38577448]
25. Zhang J, Liu M, Qu Q, et al. Radiomics analysis of gadoxetic acid-enhanced MRI for evaluating vessels encapsulating tumour clusters in hepatocellular carcinoma. Front Oncol 2024;14:1422119. [doi: 10.3389/fonc.2024.1422119] [Medline: 39193385]
26. Yu Y, Fan Y, Wang X, et al. Gd-EOB-DTPA-enhanced MRI radiomics to predict vessels encapsulating tumor clusters (VETC) and patient prognosis in hepatocellular carcinoma. Eur Radiol 2022 Feb;32(2):959-970. [doi: 10.1007/s00330-021-08250-9] [Medline: 34480625]
27. Yang J, Dong X, Wang F, et al. A deep learning model based on MRI for prediction of vessels encapsulating tumour clusters and prognosis in hepatocellular carcinoma. Abdom Radiol 2024 Apr;49(4):1074-1083. [doi: 10.1007/s00261-023-04141-3]
28. Wang J, Shi C, Fan Y, Wu Q, Zhang T, Zhang J, et al. The value of Gd-EOB-DTPA enhanced MRI deep learning in preoperative prediction of vessels completely encapsulating tumor clusters of hepatocellular carcinoma [article in Chinese]. Chinese Journal of Radiology 2025;59(6):657-664. [doi: 10.3760/cma.j.cn112149-20240624-00350]
29. Qu Q, Liu Z, Lu M, et al. Preoperative gadoxetic acid-enhanced MRI features for evaluation of vessels encapsulating tumor clusters and microvascular invasion in hepatocellular carcinoma: creating nomograms for risk assessment. J Magn Reson Imaging 2024 Sep;60(3):1094-1110. [doi: 10.1002/jmri.29187] [Medline: 38116997]

XSL•FO
RenderX

30. Gu M, Zou W, Chen H, et al. Multilayer perceptron deep learning radiomics model based on Gd-BOPTA MRI to identify vessels encapsulating tumor clusters in hepatocellular carcinoma: a multi-center study. Cancer Imaging 2025 Jul 7;25(1):87. [doi: 10.1186/s40644-025-00895-9] [Medline: 40624579]
31. Dong X, Yang J, Zhang B, et al. Deep learning radiomics model of dynamic contrast-enhanced MRI for evaluating vessels encapsulating tumor clusters and prognosis in hepatocellular carcinoma. J Magn Reson Imaging 2024 Jan;59(1):108-119. [doi: 10.1002/jmri.28745] [Medline: 37078470]
32. Chu T, Zhao C, Zhang J, et al. Application of a convolutional neural network for multitask learning to simultaneously predict microvascular invasion and vessels that encapsulate tumor clusters in hepatocellular carcinoma. Ann Surg Oncol 2022 Oct;29(11):6774-6783. [doi: 10.1245/s10434-022-12000-6] [Medline: 35754067]
33. Matsuda K, Ueno A, Tsuzaki J, et al. Vessels encapsulating tumor clusters contribute to the intratumor heterogeneity of HCC on Gd-EOB-DTPA-enhanced MRI. Hepatol Commun 2025 Jan 1;9(1):e0593. [doi: 10.1097/HC9.0000000000000593] [Medline: 39670871]
34. Wang W, Lu J, Yan Y, et al. A radiomic model for noninvasive prediction of PD-L1 and VETC expression in hepatocellular carcinoma using enhanced abdominal CT. Front Oncol 2025;15:1696376. [doi: 10.3389/fonc.2025.1696376] [Medline: 41458583]
35. Fringuello Mingo A, Colombo Serra S, Macula A, et al. Amplifying the effects of contrast agents on magnetic resonance images using a deep learning method trained on synthetic data. Invest Radiol 2023 Dec 1;58(12):853-864. [doi: 10.1097/RLI.0000000000000998] [Medline: 37378418]
36. Faccia M, Garcovich M, Ainora ME, et al. Contrast-enhanced ultrasound for monitoring treatment response in different stages of hepatocellular carcinoma. Cancers (Basel) 2022 Jan 18;14(3):481. [doi: 10.3390/cancers14030481] [Medline: 35158749]
37. Tiwari A, Mishra S, Kuo TR. Current AI technologies in cancer diagnostics and treatment. Mol Cancer 2025 Jun 2;24(1):159. [doi: 10.1186/s12943-025-02369-9] [Medline: 40457408]
38. Chen X, Ishwaran H. Random forests for genomic data analysis. Genomics 2012 Jun;99(6):323-329. [doi: 10.1016/j.ygeno.2012.04.003] [Medline: 22546560]
39. Ganaie MA, Tanveer M, Suganthan PN, Snasel V. Oblique and rotation double random forest. Neural Netw 2022 Sep;153:496-517. [doi: 10.1016/j.neunet.2022.06.012] [Medline: 35816861]
40. Castro W, Oblitas J, Santa-Cruz R, Avila-George H. Multilayer perceptron architecture optimization using parallel computing techniques. PLoS ONE 2017;12(12):e0189369. [doi: 10.1371/journal.pone.0189369] [Medline: 29236744]
41. Wang M, Wang Y, Cao L, et al. Diagnostic performance based on MRI for preoperative of vessels encapsulating tumor clusters in hepatocellular carcinoma: a systematic review and meta-analysis. Radiol Med 2025 Dec;130(12):1921-1935. [doi: 10.1007/s11547-025-02100-0] [Medline: 40991162]
42. Kouvari M, Valenzuela-Vallejo L, Guatibonza-Garcia V, et al. Liver biopsy-based validation, confirmation and comparison of the diagnostic performance of established and novel non-invasive steatotic liver disease indexes: results from a large multi-center study. Metabolism 2023 Oct;147:155666. [doi: 10.1016/j.metabol.2023.155666] [Medline: 37527759]
43. Lipkova J, Chen RJ, Chen B, et al. Artificial intelligence for multimodal data integration in oncology. Cancer Cell 2022 Oct 10;40(10):1095-1110. [doi: 10.1016/j.ccell.2022.09.012] [Medline: 36220072]
44. Rao VM, Hla M, Moor M, et al. Multimodal generative AI for medical image interpretation. Nature 2025 Mar;639(8056):888-896. [doi: 10.1038/s41586-025-08675-y] [Medline: 40140592]
45. Sosna J, Joskowicz L, Saban M. Navigating the AI landscape in medical imaging: a critical analysis of technologies, implementation, and implications. Radiology 2025 Jun;315(3):e240982. [doi: 10.1148/radiol.240982] [Medline: 40552997]
46. Mózes FE, Lee JA, Vali Y, et al. Performance of non-invasive tests and histology for the prediction of clinical outcomes in patients with non-alcoholic fatty liver disease: an individual participant data meta-analysis. Lancet Gastroenterol Hepatol 2023 Aug;8(8):704-713. [doi: 10.1016/S2468-1253(23)00141-3] [Medline: 37290471]
47. Chen ZY, Guo ZX, Lu LH, et al. The predictive value of vessels encapsulating tumor clusters in treatment optimization for recurrent early-stage hepatocellular carcinoma. Cancer Med 2021 Aug;10(16):5466-5474. [doi: 10.1002/cam4.4102] [Medline: 34212527]
48. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
49. Castelvecchi D. Can we open the black box of AI? Nature 2016 Oct 6;538(7623):20-23. [doi: 10.1038/538020a] [Medline: 27708329]
50. Champendal M, Müller H, Prior JO, Dos Reis CS. A scoping review of interpretability and explainability concerning artificial intelligence methods in medical imaging. Eur J Radiol 2023 Dec;169:111159. [doi: 10.1016/j.ejrad.2023.111159] [Medline: 37976760]
51. Renne SL, Woo HY, Allegra S, et al. Vessels encapsulating tumor clusters (VETC) is a powerful predictor of aggressive hepatocellular carcinoma. Hepatology 2020 Jan;71(1):183-195. [doi: 10.1002/hep.30814] [Medline: 31206715]
52. Van den Bergh O, Walentynowicz M. Accuracy and bias in retrospective symptom reporting. Curr Opin Psychiatry 2016 Sep;29(5):302-308. [doi: 10.1097/YCO.0000000000000267] [Medline: 27427854]

XSL•FO
RenderX

53. Hanna MG, Pantanowitz L, Jackson B, et al. Ethical and Bias Considerations in Artificial Intelligence/Machine Learning. Mod Pathol 2025 Mar;38(3):100686. [doi: 10.1016/j.modpat.2024.100686] [Medline: 39694331]
54. de Hond AAH, Shah VB, Kant IMJ, Van Calster B, Steyerberg EW, Hernandez-Boussard T. Perspectives on validation of clinical predictive algorithms. NPJ Digit Med 2023 May 6;6(1):86. [doi: 10.1038/s41746-023-00832-9] [Medline: 37149704]
55. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023 Dec;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
56. Leonhardt CS, Gustorff C, Klaiber U, et al. Prognostic factors for early recurrence after resection of pancreatic cancer: a systematic review and meta-analysis. Gastroenterology 2024 Oct;167(5):977-992. [doi: 10.1053/j.gastro.2024.05.028] [Medline: 38825047]

Abbreviations

[^{18}F]FDG PET/CT: 2-[^{1} F]fluoro-2-deoxy-D-glucose positron emission tomography or computed tomography
AI: artificial intelligence
AUC: area under the curve
CECT: contrast-enhanced computed tomography
CEMRI: contrast-enhanced magnetic resonance imaging
CEUS: contrast-enhanced ultrasound
CHARMS: Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies
CT: computed tomography
DOR: diagnostic odds ratio
FN: false negative
FP: false positive
GRADE: Grading of Recommendations, Assessment, Development, and Evaluation
HCC: hepatocellular carcinoma
HKSJ: Hartung-Knapp-Sidik-Jonkman
HR: hazard ratio
MLP: multilayer perceptron
MRI: magnetic resonance imaging
PI: prediction interval
PITROS: population, index test, target condition, reference standard, outcomes, and setting
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Diagnostic Test Accuracy
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
PROBAST: Prediction model Risk Of Bias Assessment Tool
REML: restricted maximum likelihood
RF: random forest
TN: true negative
TP: true positive
VETC: vessels encapsulating tumor cluster

Edited by S Brini; submitted 06.Jan.2026; peer-reviewed by A Hassankhani, S Sukumar; revised version received 07.May.2026; accepted 08.May.2026; published 02.Jul.2026.

Please cite as:
Hua X, Fu R, Yin Z
Radiomics-Based AI for the Diagnosis and Prognosis of Vessels Encapsulating Tumor Clusters in Hepatocellular Carcinoma: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e90931
URL: https://www.jmir.org/2026/1/e90931
doi:10.2196/90931

© Xuefeng Hua, Rongdang Fu, Ziwei Yin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Therapeutic Interaction Features of AI Chatbots in Depression Interventions: Systematic Review and Meta-Analysis

Ting Huang[1], BEng, MSE; Shuangyu Li[2], PhD; Yanzhong Wang[3], PhD; Wei Liu[1], PhD

[1]Department of Engineering, King's College London, S2.20, Strand Building, Strand Campus, Strand, London, United Kingdom

[2]Department of Interdisciplinary Humanities, Faculty of Arts and Humanities, King's College London, London, United Kingdom

[3]Department of Population Health Sciences, School of Life Course and Population Sciences, Faculty of Life Sciences & Medicine, King's College London, London, United Kingdom

Corresponding Author:
Wei Liu, PhD
Department of Engineering, King's College London, S2.20, Strand Building, Strand Campus, Strand, London, United Kingdom

Abstract

Background: Depression is a prevalent mental health disorder and a leading cause of disability worldwide, creating substantial personal and societal burdens. Digital mental health interventions have emerged as accessible and scalable solutions, with artificial intelligence (AI)–driven chatbots increasingly applied to deliver therapeutic content, monitor symptoms, and provide personalized support. However, limited evidence exists on how chatbot interaction features influence treatment adherence and clinical outcomes in depression.

Objective: This systematic review aimed to evaluate the clinical effectiveness of AI-driven chatbots for depression and to examine the associations between chatbot characteristics, treatment outcomes, and user adherence.

Methods: A systematic review and meta-analysis were conducted following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines, searching 6 databases (Web of Science, Scopus, PubMed, IEEE Xplore, Embase, and APA PsycINFO) for randomized controlled trials (RCTs) published before May 30, 2025. Eligible studies involved individuals with depression or depressive symptoms receiving AI-driven chatbot, conversational agent, or virtual therapist interventions, with outcomes assessed using the Patient Health Questionnaire-9 (PHQ-9). Data extraction included chatbot type, interaction features, adherence, and standardized mean differences (SMDs) for symptom change. Risk of bias was assessed using the Cochrane Risk of Bias tool version 2 (RoB 2). Random-effects meta-analyses were performed with the Hartung-Knapp-Sidik-Jonkman adjustment. This review was preregistered on the Open Science Framework.

Results: A total of 11 RCTs involving 2220 participants (1091 in the intervention and 1129 in the control groups) were included. Using a random-effects model with Hartung-Knapp-Sidik-Jonkman adjustment, AI-driven chatbots showed a small-to-moderate reduction in depressive symptoms compared with control conditions, but the effect was not statistically significant (SMD=−0.46, 95% CI −1.02 to 0.10; P=.01; 95% prediction interval −1.50 to 0.58). Subgroup analyses of adherence did not show significant differences across the reported chatbot-type subgroups. In contrast, exploratory analyses of interaction features revealed more consistent patterns for adherence. Emotional responsiveness, structured feedback strategies, and interaction frequency were associated with higher adherence in high-scoring subgroups, whereas dialogue depth, self-disclosure encouragement, and user agency level showed weaker or inconsistent associations. For clinical outcomes, associations with interaction features were less consistent and more heterogeneous.

Conclusions: This systematic review provides an interaction-focused synthesis of AI-driven chatbot interventions for depression, examining how interaction features relate to clinical outcomes and user adherence. Although overall effects were not statistically significant, emotional responsiveness, structured feedback, and interaction frequency were consistently associated with higher adherence. Engagement and outcomes may be influenced by distinct mechanisms. Limitations include the small number of RCTs, heterogeneity, reliance on study-reported descriptions, and potential publication bias. These findings highlight the importance of interaction design in developing scalable digital mental health interventions.

(J Med Internet Res 2026;28:e88697) doi:10.2196/88697

KEYWORDS

depression; AI-driven chatbot; user adherence; digital mental health intervention; interaction design; meta-analysis

XSL•FO
RenderX

Introduction

Depression is one of the leading causes of disability worldwide. According to the World Health Organization (WHO), more than 280 million people worldwide are affected, with a prevalence of 3.8% in the general population and 5.7% among adults aged 60 years or older [1]. Depression contributes substantially to the global burden of disease and has serious effects on quality of life, productivity, and physical health [2]. Although face-to-face psychotherapy and pharmacological treatments are effective, access remains limited due to workforce shortages, stigma, and geographical barriers [3,4].

To address these barriers, digital mental health interventions (DMHIs) have rapidly emerged as scalable, accessible alternatives [5]. For example, internet-based cognitive behavioral therapy (iCBT) allows remote treatment, while mobile apps support mood tracking and self-management [6]. Artificial intelligence (AI)–driven chatbots, which have attracted growing attention in this field, offer continuous support and characterize sustained, language-based interaction.

Early DMHIs relied on web-based psychoeducation and structured cognitive behavioral therapy programs with fixed content and limited interactivity [7,8]. With the expansion of mobile technologies, these interventions shifted toward app-based formats, providing more flexible, on-demand support through features such as reminders, mood tracking, and self-help tools [9]. However, these features largely remained task-based and relied on predefined responses rather than sustained, context-sensitive interaction.

A meta-analysis by Lattie et al [10], examined a wide range of DMHIs, including iCBT, app-based interventions, messaging systems, and virtual reality platforms. The findings showed that DMHIs can achieve clinical outcomes comparable to traditional therapies for mild to moderate depression [10]. However, the review provided limited differentiation between interaction types and did not examine their specific roles in shaping clinical outcomes. This suggests that interaction has often been treated as a secondary feature rather than a core mechanism in DMHIs.

Despite advances, DMHIs still face high dropout rates and low retention. This persistent issue is known as the “Law of Attrition,” which highlights the challenge of maintaining long-term engagement in digital interventions [11]. Recent work suggests evaluating both engagement and clinical effectiveness rather than focusing on a single metric [12]. These concerns are especially relevant for chatbot-based interventions, which rely on ongoing conversational interaction for continued use.

Existing studies show that traditional DMHIs, such as web-based psychoeducation and iCBT, struggle with user retention and lack personalized, interactive support. These shortcomings underscore the importance of interaction for improving therapeutic effectiveness and adherence. In response, computerized cognitive behavioral therapy emerged in the late 20th century. Since the mid-2010s, AI has been integrated into health care, leading to the rise of AI-driven chatbots [13].

Unlike earlier DMHIs, AI chatbots are fundamentally interaction-centered, with therapeutic support delivered primarily through ongoing conversational exchange. Consistent with prior definitions of conversational agents in digital health, they are distinguished from nonconversational digital interventions by their capacity for sustained, multiturn dialogue, which forms the core therapeutic mechanism [14]. A meta-analysis by Li et al [15] found that AI-based conversational agents significantly reduced depressive symptoms and further suggested that user experience depends on factors such as therapeutic alliance with AI, content engagement, and communication quality. This points to the need to examine the mechanisms through which chatbot interactions produce clinical benefit. However, the role of specific interaction features in shaping these mechanisms remains insufficiently examined.

By leveraging natural language processing (NLP) and machine learning (ML), chatbots can simulate human conversation and provide round-the-clock access to psychoeducation, cognitive behavioral therapy exercises, and emotional support [16,17]. Recent studies have shown that chatbots such as Woebot (Woebot Health, Inc) and Wysa (Wysa Health) can reduce depressive symptoms across diverse populations [18,19]. At the same time, more recent work suggests that integrating large language models (LLMs) into mental health care remains at an early stage. A UK-based evaluation involving 132 participants found that although many users were familiar with systems such as ChatGPT (OpenAI) and Doubao (ByteDance), their clinical use in mental health care was still limited [20]. Similarly, semistructured interviews with German adolescents experiencing depressive symptoms showed that participants generally held cautiously positive attitudes toward chatbots, while also expressing diverse and sometimes conflicting expectations regarding personalization [21]. Together, these findings suggest that LLMs may extend chatbot capabilities in contextual awareness, empathy, and personalized interaction, but their clinical role is still evolving [22].

Recent research has highlighted the importance of aligning chatbot design with user preferences and needs in mental health care. For instance, Kim et al [23] used a mixed logit model to analyze user choice data, showing that preferences for mental health chatbots are broadly consistent with those observed in traditional counseling contexts. They therefore emphasized the importance of human-centered design in the development of health care chatbots. Personalization is also critical across different user groups. For example, chatbots designed for older adults should incorporate age-friendly interfaces that account for age-related physiological characteristics [24]. In addition, user personality traits and usage contexts have been identified as important factors in shaping chatbot interaction. Furini et al [25], based on data from multiple user profiles and scenarios, highlighted the need to integrate personality and health conditions into chatbot interactions to improve engagement and outcomes. Together, these studies underscore the importance of interaction design in mental health chatbots. However, their findings remain fragmented and difficult to synthesize across different interaction features and user groups.

Trust is a key factor in mental health care, as therapeutic effectiveness depends on the development of rapport, empathy, and credibility between clinician and patient. In traditional therapy, clinicians build trust through empathic listening,

XSL•FO
RenderX

contingent feedback, and adaptive communication. Translating these mechanisms into chatbot interactions remains challenging. Dong and Wu [26] explored how the perceived status of a health care chatbot influences patient trust. Their findings suggest that when chatbots assume a high-status role and provide contextually contingent responses, users report lower anxiety when interacting with AI systems. However, how specific interaction features such as empathy expression or feedback strategies can be adaptively adjusted to promote trust remains insufficiently studied [27-29]. This is particularly relevant given the skepticism that both patients and clinicians often express toward AI in health care, which may compromise trust-building and acceptance. Existing studies suggest that embedding personalization mechanisms into interactional features could strengthen user trust, thereby improving adherence and clinical outcomes [30-32]. Overall, trust represents a key interactional mechanism in mental health chatbots. However, the effects of specific interaction features on trust, adherence, and clinical outcomes have not been systematically examined.

This understanding of how AI-driven chatbots support depression care remains fragmented. Existing research has largely focused on overall clinical effectiveness, while systematic examination of how specific interaction features (eg, dialogue depth, feedback strategies, and emotional responsiveness) and content types (eg, self-disclosure prompts and goal-setting) relate to therapeutic outcomes remains limited. Moreover, evidence on how these interactional characteristics influence user adherence is sparse and often indirect. Addressing these gaps is essential to inform the design of chatbot-based interventions that not only reduce depressive symptoms but also sustain engagement and trust over time. Against this background, the present study synthesizes existing evidence to examine how interaction and content features of AI-driven chatbots relate to clinical effectiveness and user adherence in depression care.

This systematic review therefore aims to address the following research questions (RQs):

1. RQ1: What is the overall clinical effectiveness of AI-driven chatbots in depression interventions?
2. RQ2: How does user adherence vary across AI-driven chatbots based on different AI technologies?
3. RQ3: How do the interaction features of AI-driven chatbots influence both treatment outcomes and user adherence?

Methods

Search Strategy

This systematic literature search was conducted to identify research on AI-driven chatbot interventions for the treatment of depression. The objective was to evaluate the influence of chatbot-based digital tools on clinical outcomes, including symptom improvement, treatment effectiveness, and user adherence. Studies were included only if chatbots were used as therapeutic tools, while those focusing exclusively on diagnosis, screening, or prediction were excluded.

The search was conducted across 6 major academic databases, including Scopus, Web of Science, PubMed, IEEE Xplore, APA PsycINFO (Ovid), and Embase (Ovid). The strategy incorporated 2 primary concept domains, namely depression and chatbot or conversational systems. Only peer-reviewed journal articles and conference proceedings published in English, with coverage up to May 30, 2025, were included, as detailed in Textbox 1.

The search strategy was developed in accordance with guidance from the Cochrane Handbook. Consistent conceptual blocks were applied across all databases, with syntax adapted for each platform. Full database-specific search strategies are provided in Multimedia Appendix 1. Reporting of the search strategy and process adhered to PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) guidelines to ensure transparency and reproducibility [33]. The PRISMA-S checklist is presented in Checklist 1.

Textbox 1. Search strategy.

Search topic

1. Depression and depressive disorders (eg, “depression,” “major depressive disorder,” “major depressive disorder,” “dysthymia”)
2. Chatbot and conversational agent systems (eg, “chatbot,” “conversational agent,” “virtual agent,” “dialogue system,” “Woebot,” “Wysa”)
3. Intervention and treatment-related terms (eg, “intervention,” “therapy,” “treatment,” “counselling,” “psychotherapy,” “cognitive behavioral therapy,” “CBT,” “digital mental health,” “digital therapy,” “psychological intervention”)

Search example

TS=((depress* OR “major depressive disorder” OR “MDD” OR dysthymi* OR “persistent depressive disorder” OR “depressive disorder*” OR “depressive symptom*” OR “depressive episode*” OR “recurrent depressive disorder” OR “unipolar depression” OR “mood disorder*” OR “affective disorder*” OR “subclinical depression” OR “subthreshold depression”)

AND

(chatbot* OR “chat bot*” OR “conversational agent*” OR “conversational AI” OR “dialogue system*” OR “dialog system*” OR “virtual therapist*” OR “virtual agent*” OR “relational agent*” OR “embodied agent*” OR “mental health bot*” OR Woebot OR Wysa OR Tess OR Youper OR Replika OR Ellie)

AND

(intervention* OR therap* OR treatment* OR counsel* OR psychotherap* OR “cognitive behavioral therapy” OR CBT OR “digital mental health” OR “digital therap*” OR “psychological intervention*”))

XSL•FO
RenderX

Selection Criteria

Inclusion Criteria

Studies were included if they met all of the following criteria:

1. Population: participants were individuals experiencing depression or related affective conditions, including major depressive disorder, dysthymia, clinical depression, or comorbid anxiety symptoms.
2. Intervention: the study examined a digital intervention in which an AI-driven chatbot, conversational agent, or virtual therapist played a central role in delivering therapeutic content (eg, psychoeducation, cognitive behavioral therapy, counseling support, or mood regulation exercises).
3. Purpose of intervention: the chatbot was used with the explicit aim of reducing depressive symptoms, improving psychological well-being, or supporting behavioral change. Both standalone and blended interventions (chatbot plus human support) were eligible.
4. Outcome: the study reported at least one outcome related to treatment effectiveness, symptom improvement, or adherence.
5. Study type: randomized controlled trials (RCTs).
6. Publication status and language: full-text available in English; published as a journal article or conference proceeding.

Exclusion Criteria

Studies were excluded if they met any of the following:

1. Nontherapeutic use: the chatbot was used solely for screening, diagnosis, symptom monitoring, or predictive modeling without a therapeutic component.
2. Non-AI or rule-based systems: the system used was not AI-driven (static decision-tree chatbots without learning capacity).
3. Nondepressive focus: the intervention targeted conditions unrelated to depression, such as bipolar disorder, schizophrenia, psychosis, dementia, or autism spectrum disorder.
4. Theoretical or technical papers: studies describing only the design, technical architecture, or conceptual framework of a chatbot without user evaluation or outcome reporting.
5. Lack of baseline data: studies that did not report baseline outcome measures for depression.
6. Inconsistent outcome measurement: pre-post comparisons not based on validated depression scales, specifically the Patient Health Questionnaire-9 (PHQ-9).
7. Gray literature: Editorials, protocols, opinion pieces, preprints, or dissertations.

Data Extraction

All records identified from the 6 databases, including Scopus, Web of Science, PubMed, IEEE Xplore, APA PsycINFO (Ovid), and Embase (Ovid), were imported into EndNote (version 21; Clarivate) for management and initial filtering based on 11 predefined bibliographic fields, including author, year, abstract, and keywords. A total of 3372 records were retrieved. After removing 4 erroneous records, 3368 records were retained for deduplication, resulting in 2097 unique articles. Titles and abstracts were independently screened by 2 reviewers (TH and WL) against the eligibility criteria. Any discrepancies were resolved through discussion with a third reviewer. In total, 87 articles were selected for full-text assessment, of which 8 met the inclusion criteria. An additional 3 eligible studies were identified through snowballing. Overall, 11 studies were included in the final analysis and subjected to meta-analysis. The screening and selection process is summarized in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses; Checklist 2) flow diagram.

In addition to extracting general study characteristics, 6 interaction features of the chatbots were assessed, including interaction frequency, emotional responsiveness, self-disclosure encouragement, dialogue depth, feedback strategy, and user agency level. Each feature was rated on a 5-point scale (1=very low-5=very high) by 2 independent reviewers (TH and WL). A third reviewer adjudicated when ratings differed by ≥1 point, and the final score for each feature was calculated as the mean of the available ratings. Interrater reliability across all ratings was good (overall intraclass correlation coefficient [ICC]=0.71); detailed results are presented in the Multimedia Appendix 2.

For clarity, the 6 features were defined as follows: interaction frequency (the intensity and regularity of user-chatbot exchanges), emotional responsiveness (the chatbot's ability to adaptively provide empathetic responses), self-disclosure encouragement (prompts guiding users to share personal experiences or emotions), dialogue depth (the richness and reflectiveness of conversations), feedback strategy (the presence of timely and tailored prompts or evaluative responses), and user agency level (the degree of control and choice available to the user).

User adherence was operationalized as intervention completion, defined as the proportion of participants who completed the intervention protocol relative to the number initially enrolled. This completion-based measure was consistently reported across the included studies and was therefore adopted to enable quantitative synthesis of adherence outcomes.

Data Quality

The methodological quality of the included studies was assessed independently by 2 reviewers (TH and WL) using the Cochrane Risk of Bias 2 (RoB 2) tool for RCTs. The tool evaluates potential biases across five domains: (1) bias arising from the randomization process, (2) bias due to deviations from intended interventions, (3) bias due to missing outcome data, (4) bias in the measurement of the outcome, and (5) bias in the selection of the reported result. Each domain was rated as "low risk," "some concerns," or "high risk" according to the signaling questions provided by RoB 2 guidelines.

Any discrepancies in assessments between the 2 reviewers (TH and WL) were resolved through discussion; if disagreement persisted, a third reviewer was consulted. The final risk-of-bias assessments were summarized in tabular and graphical form.

Statistical Analysis

Meta-analyses were conducted using random-effects models. Standardized mean differences (SMDs) were calculated for

XSL•FO
RenderX

continuous outcomes, and odds ratios (ORs) were calculated for dichotomous outcomes. Heterogeneity was quantified using the I^2 and statistics [34]. Prediction intervals (PIs) were calculated to estimate the range of true effects in future comparable settings [35]. To provide more robust CI estimates, particularly given the relatively small number of included studies, the Hartung-Knapp-Sidik-Jonkman (HKSJ) adjustment was applied [36]. The DerSimonian-Laird estimator was also used for comparison [37]. To assess potential small-study effects, funnel plots were visually inspected. Egger regression test [38], the Begg-Mazumdar rank correlation test [39], and the Duval and Tweedie trim-and-fill procedure were applied where appropriate (≥10). All analyses were conducted in R (version 4.5.1; R Foundation for Statistical Computing).

Results

Overview

The study selection process is shown in Figure 1. After full-text screening, a total of 11 articles were included in the analysis. In total, 2220 participants (1091 in the intervention and 1129 in the control groups) were included in our analysis. A summary of the characteristics of the studies is presented in Table 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart. Study selection for systematic review.

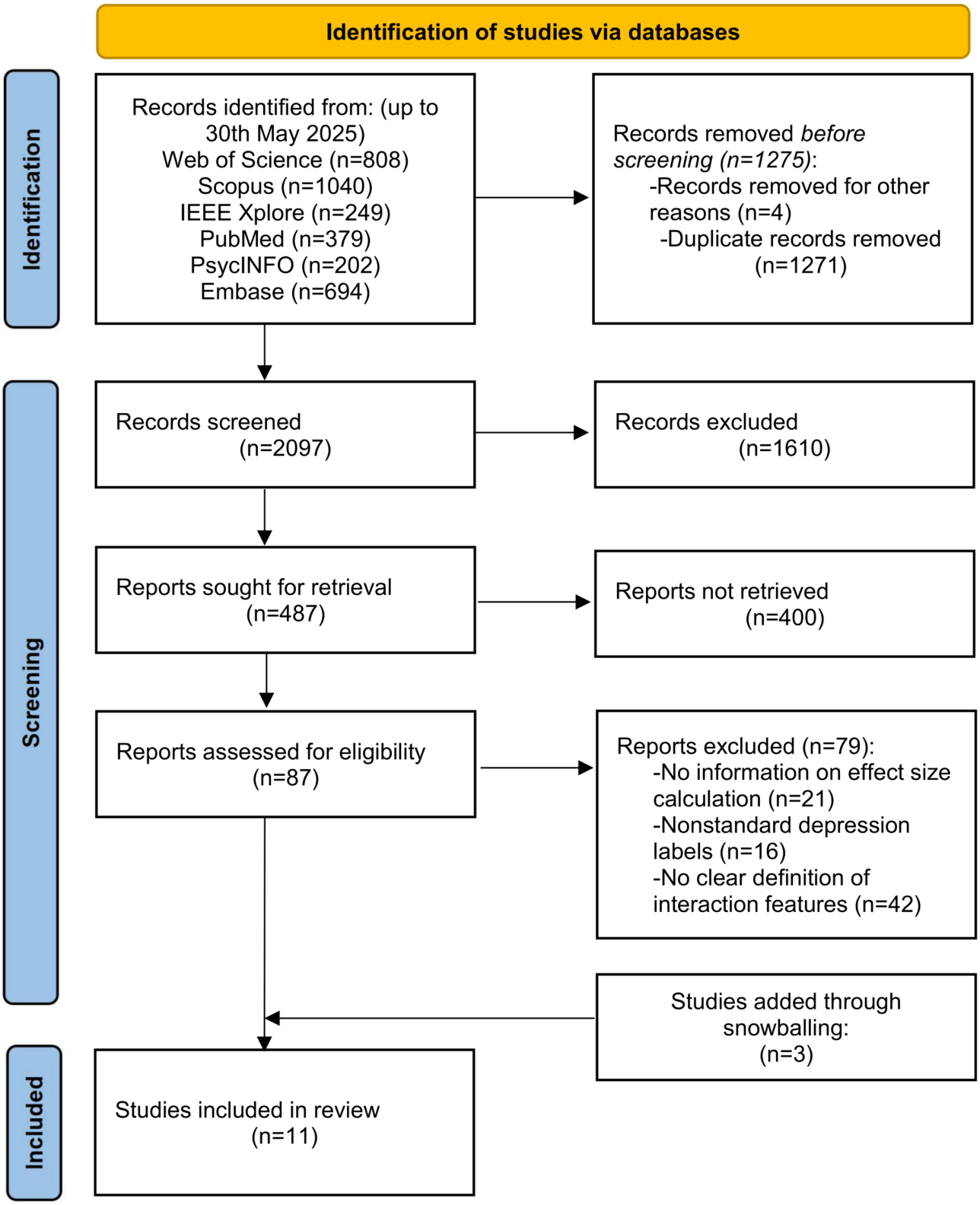

XSL•FO
RenderX

Table . Summary of all studies included in the review.

Item	Study and year	Country	Initial participants (n)		Used for effect size (n)		Age (years), mean (SD) or median (IQR)	Study methodology	Scales	Chatbot name or platform
			Intervention	Control	Intervention	Control				
1	Chen et al (2025) [40]	China	62	41	62	41	__[a]	2-armed RCT[b] with 2 parallel groups recruited from Hong Kong	PHQ-9[c] GAD-7[d]	COVID-19 information chatbot (University of Hong Kong)
2	Fitzpatrick et al (2017) [41]	United States	34	36	31	25	• 22.2 (2.33)	2-armed RCT with 2 groups recruited from a university community social media site	PHQ-9 GAD-7	Woebot (Woebot Health, Inc)
3	He et al (2022) [42]	China	49	49	44	32	• 18.78 (3.18)	3-arm RCT performed at a university in Tianjin, China	PHQ-9	XiaoE (Tianjin University; technical support from Xiaomi Corporation)
4	Kang and Hong (2024) [43]	South Korea	22	10	15	3	• Experimental: 23.5 (1.78) • Control: 22.9 (1.85)	2-armed RCT with participants recruited from Sungkyunkwan University's Colleges of Natural Sciences and Humanities and Social Sciences in Seoul, South Korea	UCLA[e] PHQ-9	Woebot (Woebot Health, Inc)
5	Karkosz et al (2024) [44]	Poland	40	41	33	35	• Experimental: 26.60 (5.06) • Control: 24.76 (4.01)	2-armed RCT with participants recruited via Facebook (Meta) and Instagram (Meta) advertisements	CESD-R[f] PANAS[g] PHQ-9 PSWQ[h] R-UCLA[i] STAI[j] SWLS[k]	Fido (Szkoła Wyższa Psychologii Społecznej University research team)

XSL•FO
RenderX

Item	Study and year	Country	Initial participants (n)		Used for effect size (n)		Age (years), mean (SD) or median (IQR)	Study methodology	Scales	Chatbot name or platform
			Intervention	Control	Intervention	Control				
6	Liu et al (2022) [45]	China	41	42	33	30	• 23.08 (1.76)	2-armed RCT with participants recruited from 3 different universities in China	PHQ-9 GAD-7	XiaoNan (South China University of Technology)
7	Sabour et al (2023) [46]	China	90	121	70	105	—	3-arm RCT with participants recruited from social media platforms	PHQ-9 GAD-7 PANAS ISI[l]	ES-Bot[m] (part of Emohaa, Beijing Lingxin Intelligent Technology Co, Ltd)

XSL•FO
RenderX

Item	Study and year	Country	Initial participants (n)		Used for effect size (n)		Age (years), mean (SD) or median (IQR)	Study methodology	Scales	Chatbot name or platform
			Intervention	Control	Intervention	Control				
8	Tong et al (2024) [47]	China	140	145	118	132	• 26.45 (8.37)	2-armed RCT with participants recruited from social media platforms	SUPPH[n] eTAP[o] SCBI[p] MHLS[q] PHQ-9 GAD-7 MAAS[r] PERMA[s]	Boon (Chinese University of Hong Kong)
9	Ulrich et al (2024) [48]	Switzerland	70	70	42	56	• 26.7 (6.3)	2-armed RCT with participants recruited from a population of university students in Switzerland	PHQ-9 GAD-7 PHQ-15 HAPA[t]	MISHA (Szkoła Wyższa Psychologii Społecznej University research team)
10	Verescha-gin et al (2024) [49]	Canada	743	746	591	619	• 20 (19-23)	2-armed RCT with participants recruited from the University of British Columbia (UBC) Vancouver campus	GAD-7 PHQ-15 USAUDIT-C[u]	Minder (University of British Columbia)
11	Yasukawa et al (2024) [50]	Japan	74	75	52	51	• 41.4 (11.1)	2-armed RCT with participants recruited from Japan	PHQ-9 GAD-7 CBT[v] skills SWLS[w] WHO-5[x] WSAS[y] UWES[z]	EPO/LINE (Sony Group Corporation)

[a]Not applicable.
[b]RCT: randomized controlled trial.
[c]PHQ: Patient Health Questionnaire.
[d]GAD-7: Generalized Anxiety Disorder 7-item scale.
[e]UCLA: UCLA Loneliness Scale.
[f]CESD-R: Center for Epidemiologic Studies Depression Scale Revised.
[g]PANAS: Positive and Negative Affect Scale.
[h]PSWQ: Penn State Worry Questionnaire.
[i]R-UCLA: Revised UCLA Loneliness Scale.
[j]STAI: State-Trait Anxiety Inventory.
[k]SWLS: Satisfaction With Life Scale.
[l]ISI: Insomnia Severity Index.

XSL•FO
RenderX

[m]ES: emotional support.

[n]SUPPH: strategies used by people to promote health.

[o]eTAP: e-Therapy Attitude and Process Questionnaire.

[p]SCBI: Self-Care Behaviors Inventory.

[q]MHLS: Mental Health Literacy Scale.

[r]MAAS: Mindful Attention Awareness Scale.

[s]PERMA: positive emotion, engagement, relationships, meaning, and accomplishment.

[t]HAPA: health action process approach.

[u]USAUDIT-C: US Alcohol Use Disorders Identification Test–Consumption Scale.

[v]CBT: cognitive behavioral therapy.

[w]SWLS: Satisfaction with Life Scale.

[x]WHO-5: World Health Organization-Five Well-Being Index.

[y]WSAS: Work and Social Adjustment Scale.

[z]UWES: Utrecht Work Engagement Scale.

Overall Clinical Effectiveness of AI-Driven Chatbots in Depression Interventions

A total of 11 [40-50] RCTs involving 2220 participants (1091 in the experimental group and 1129 in the control group) were included in the meta-analysis. The pooled results indicated a small-to-moderate effect of AI-driven chatbots on depressive symptoms, compared with control conditions (SMD −0.46, 95% CI −1.02 to 0.10; *P*=.01; as shown in Figure 2). Negative SMD values indicate greater symptom reduction in the chatbot intervention groups.

To further quantify the real-world implications of heterogeneity, 95% PIs were calculated. The PI ranged from −1.50 to 0.58, indicating that the true effect in a future comparable setting could vary substantially and may include no effect. Heterogeneity among studies was substantial (I^2=87%), suggesting considerable variability in intervention effects across trials.

Sensitivity analyses excluding Fitzpatrick et al [41] reduced heterogeneity from 87% to 60%, while the direction of the pooled effect remained unchanged. Detailed results are provided in Multimedia Appendix 3. To further explore potential sources of heterogeneity, subgroup analyses were conducted to examine differences across AI chatbot types.

Figure 2. Forest plot of the overall clinical effect of artificial intelligence (AI)–driven chatbots on depressive symptoms. A random-effects meta-analysis with Hartung-Knapp-Sidik-Jonkman (HKSJ) adjusted 95% CIs is presented. The prediction interval is also shown [40-50]. AI: artificial intelligence; HKSJ: Hartung-Knapp-Sidik-Jonkman.

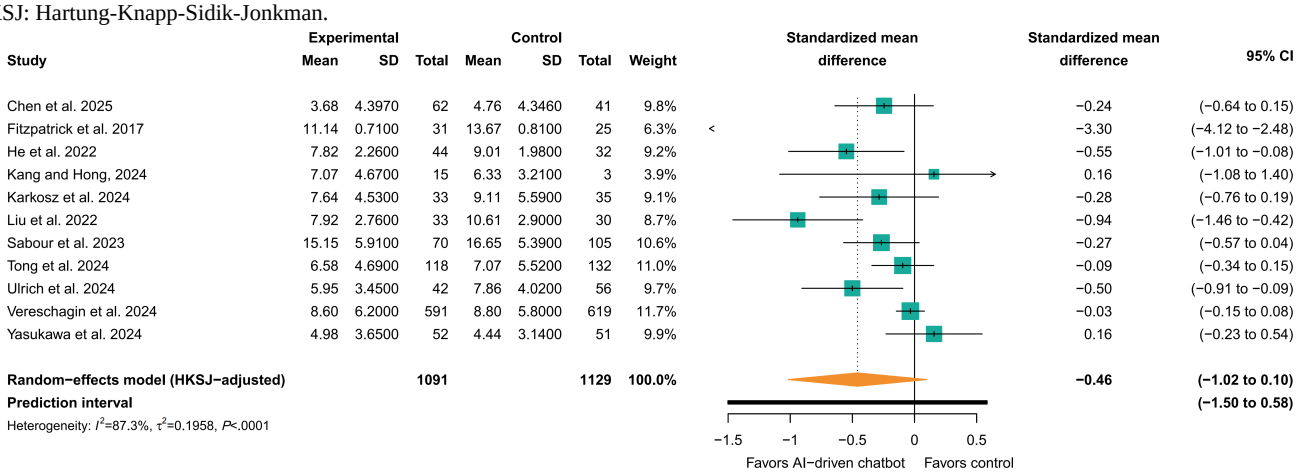

Overall Adherence of AI-Driven Chatbots in Depression Interventions

Across all 11 [40-50] included studies, the pooled analysis showed no significant difference in adherence between intervention and control groups, as shown in Figure 3 (OR 1.22, 95% CI 0.57 - 2.62; *P*=.57). To further interpret heterogeneity in real-world settings, the review calculated the 95% PI for the overall adherence outcome. The 95% PI ranged from 0.24 to 6.18, indicating substantial between-study variability. This suggests that the true adherence effect in a comparable future setting could range from lower to substantially higher engagement than in control conditions. Heterogeneity was considerable (I^2=74.8%).

XSL•FO
RenderX

Figure 3. Forest plot of the overall effect of artificial intelligence (AI)–driven chatbots on user adherence in depression interventions. A random-effects meta-analysis with Hartung-Knapp-Sidik-Jonkman (HKSJ) adjusted 95% CIs is presented. The prediction interval is also shown [40-50]. AI: artificial intelligence; HKSJ: Hartung-Knapp-Sidik-Jonkman; OR: odds ratio.

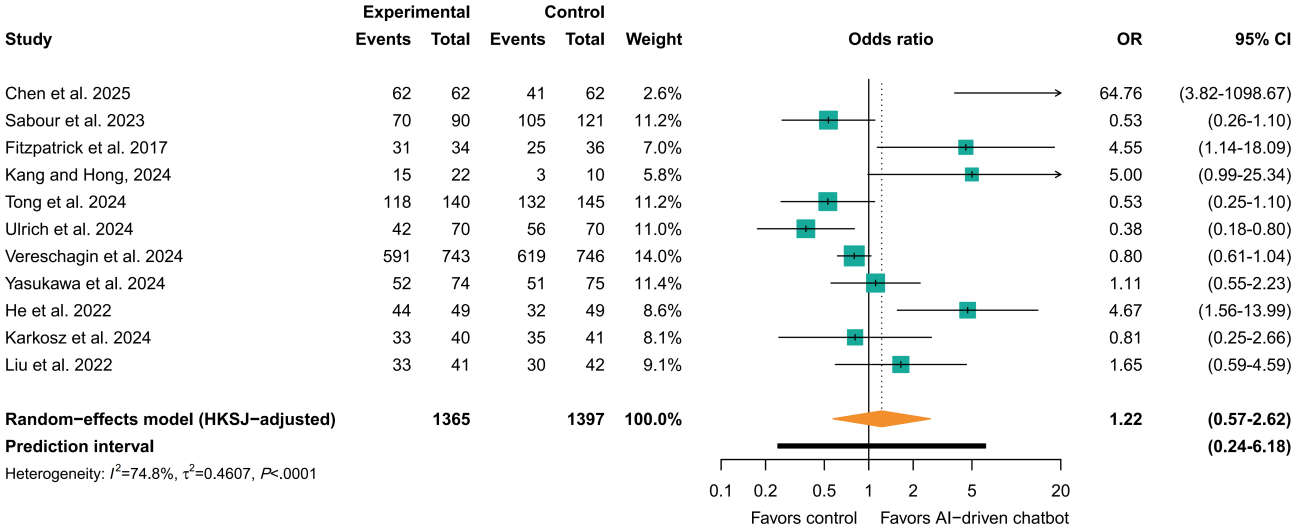

Risk of Bias and Certainty of Evidence Assessment

All studies meeting the inclusion criteria reported depressive symptom outcomes at the end of the intervention, assessed using the PHQ-9. Overall, 7 studies [40,43-46,48,49] were judged to have a low risk of bias, while 4 studies [41,42,47,50] were assessed as having some concerns. The most common problem was missing outcome data (Domain 3), with insufficient reporting of outcomes in some studies. In contrast, all studies were judged to be of low risk across domains related to randomization, deviations from intended interventions, outcome measurement, and selective reporting. No study was considered to be at high risk of bias in any domain. Detailed domain-level assessments are presented in Figure 4 and Multimedia Appendix 4.

Figure 4. Risk of bias assessment of included studies using the Cochrane Risk of Bias (RoB) 2 tool [40-50].

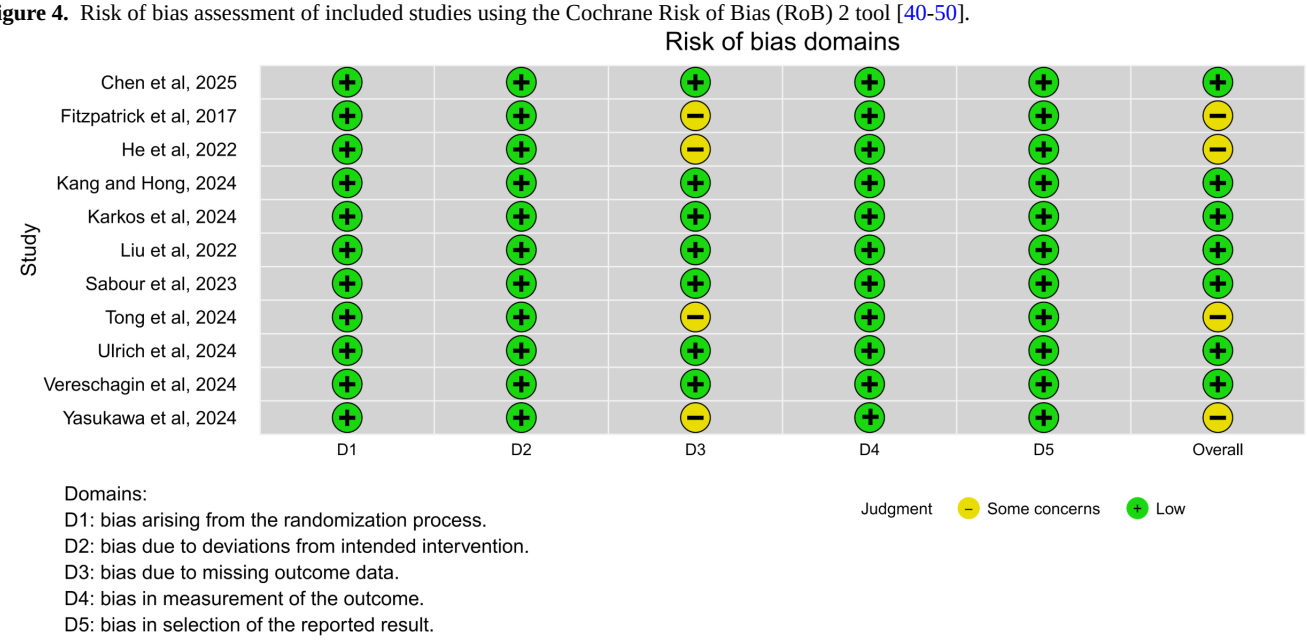

Certainty of evidence was assessed using the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach. As all included studies were RCTs, the evidence started at high certainty. For clinical effectiveness, certainty was downgraded because of very serious inconsistency, serious imprecision, and suspected publication bias, resulting in very low-certainty evidence. For user adherence, certainty was downgraded because of serious inconsistency and serious imprecision, resulting in low-certainty evidence. A summary of the GRADE assessments for the main outcomes is presented in Table 2, and a more detailed GRADE evidence profile is provided in Multimedia Appendix 5.

XSL•FO
RenderX

Table . Grading of Recommendations Assessment, Development, and Evaluation (GRADE) summary of findings for the main outcomes of AI-driven chatbot interventions for depression.

Outcome	Number of studies (participants)	Effect estimate	Certainty of evidence	Reasons for downgrading
Clinical effectiveness (depressive symptom reduction)	11 RCTs[a] (n=2220)	SMD[b] –0.46 (95% CI –1.02 to 0.10)	Very low	Very serious inconsistency, serious imprecision, and suspected publication bias
User adherence	11 RCTs (n=2762)	OR[c] 1.22 (95% CI 0.57-2.62)	Low	Serious inconsistency and serious imprecision

[a]RCT: randomized controlled trial

[b]SMD: standardized mean difference

[c]OR: odds ratio

User Adherence Across Different AI-Driven Chatbot Types

Before presenting the subgroup analyses, this review clarifies the classification of AI-driven chatbots. Although LLMs are technically a subset of NLP, in this review we distinguish LLM-based chatbots as systems built on large pretrained generative models (eg, GPT, Gemini, and Claude) that directly generate responses in an open-ended way. In contrast, NLP-/ML-based chatbots are systems that use more traditional NLP or ML methods, such as intent classifiers, decision trees, or response selection within constrained conversational flows. Rule-based chatbots are systems that rely on fixed scripts or expert-crafted rules. This operational categorization aligns with how the primary studies describe their systems and is consistent with recent surveys that distinguish LLM paradigms from traditional NLP approaches [51,52].

To further explore heterogeneity in adherence, subgroup analyses were conducted by AI type (Figure 5; Figure 6). The pooled results indicated no significant overall difference in adherence between experimental and control groups (OR 1.22, 95% CI 0.57 - 2.62; *P*=.57). To further interpret heterogeneity in real-world settings, we calculated the 95% PI for the overall adherence outcome. The 95% PI ranged from 0.24 to 6.18, indicating substantial between-study variability. This suggests that the true adherence effect in a comparable future setting could range from lower to substantially higher engagement than in control conditions.

Figure 5. Subgroup analysis of user adherence to artificial intelligence (AI)–driven chatbot interventions for depression by AI type. Random-effects meta-analysis with Hartung-Knapp-Sidik-Jonkman (HKSJ) adjusted 95% CIs is presented for each subgroup. The large language model (LLM)–based subgroup was excluded because one included study reported complete adherence in the intervention group, resulting in an extreme and clinically uninterpretable pooled odds ratio estimate [41-45,47-49]. AI: artificial intelligence; HKSJ: Hartung-Knapp-Sidik-Jonkman; ML: machine learning; NLP: natural language processing; OR: odds ratio.

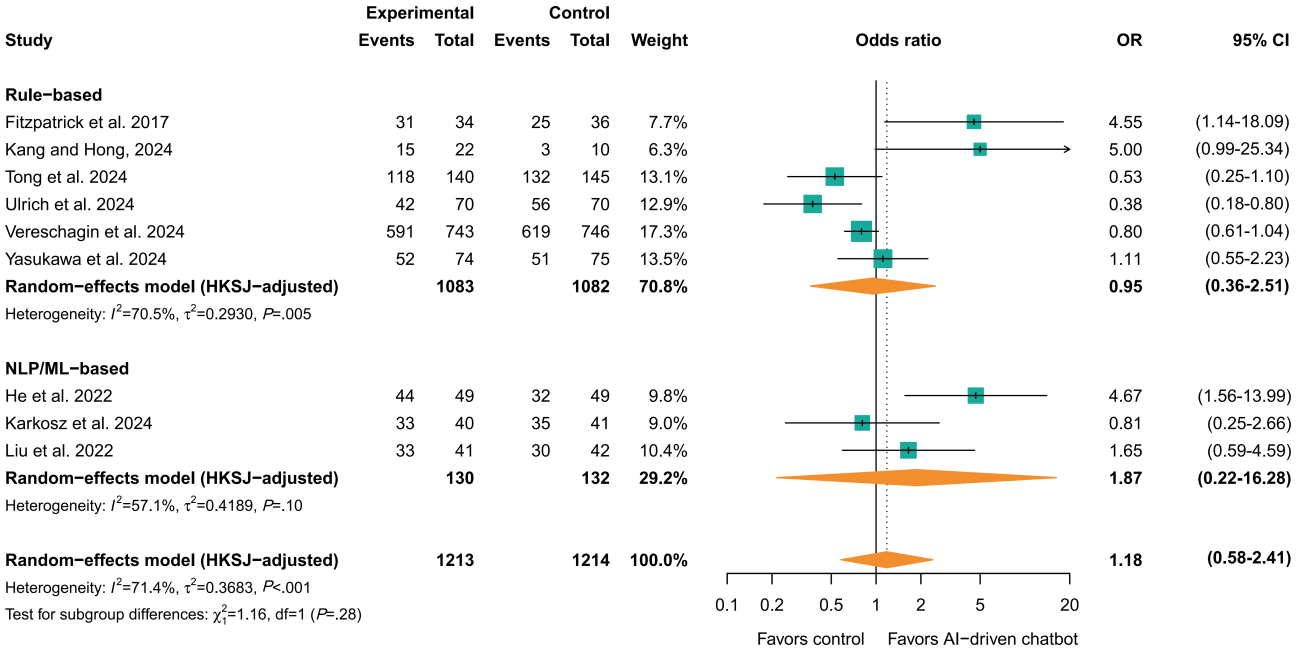

XSL•FO
RenderX

Figure 6. Subgroup analysis of the clinical effectiveness of artificial intelligence (AI)–driven chatbots in depression interventions by AI type. Random-effects meta-analysis with Hartung-Knapp-Sidik-Jonkman (HKSJ) adjusted 95% CIs is presented for each subgroup [40-50]. AI: artificial intelligence; HKSJ: Hartung-Knapp-Sidik-Jonkman; LLM: large language model; ML: machine learning; NLP: natural language processing.

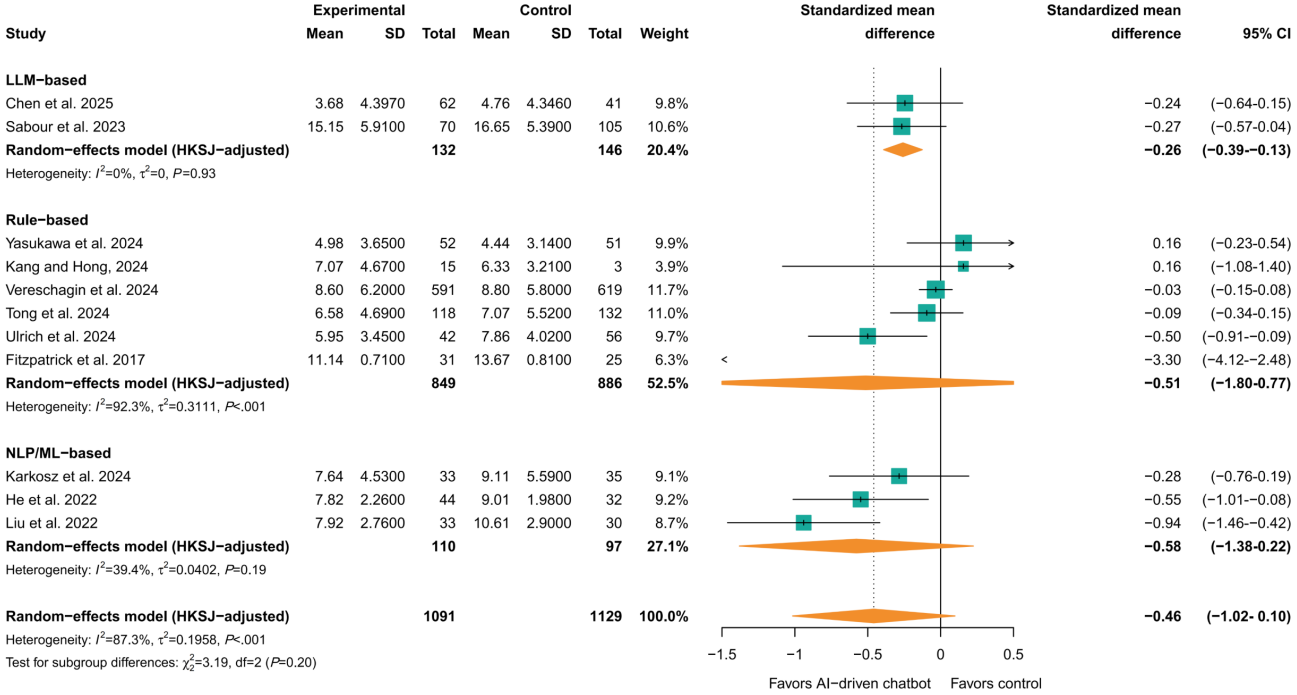

At the subgroup level, no statistically significant effects were observed. Rule-based chatbots showed a nonsignificant pooled effect (OR 0.95, 95% CI 0.36 - 2.51), while NLP/ML-based chatbots also did not reach statistical significance (OR 1.87, 95% CI 0.22 - 16.28). The test for subgroup differences was not statistically significant (P=.47). The LLM-based subgroup was excluded from the adherence subgroup figure because one included study reported complete adherence in the intervention group, resulting in an extreme and clinically uninterpretable pooled OR estimate.

These findings suggest that chatbot type alone does not consistently predict user adherence across studies. Given the substantial heterogeneity, the results should be interpreted with caution. The findings further indicate that factors beyond chatbot type, such as specific interaction features, may play a more important role in sustaining user engagement. In the following section, we therefore turn to an analysis of 6 key interaction features, including interaction frequency, emotional responsiveness, self-disclosure encouragement, dialogue depth, feedback strategy, and user agency to clarify their contribution to adherence.

Interaction Features in Relation to Clinical Effectiveness and User Adherence

This section examines 6 interaction features of AI-driven chatbots: interaction frequency, emotional responsiveness, self-disclosure encouragement, dialogue depth, feedback strategy, and user agency level. Studies were classified into high (≥3.75) and low (≤3.5) groups based on expert ratings. The detailed scoring table is provided in Multimedia Appendix 2. These ratings were derived from structured coding of published study descriptions and were used as operational proxies of interaction characteristics, rather than as direct measurements of chatbot behavior. The analyses in this section were therefore conducted in an exploratory, hypothesis-generating manner. Unlike the chatbot-type subgroup analysis presented earlier, the analyses in this section represent stratified meta-analyses based on study-level coding of interaction features. For each feature, studies were categorized as high or low according to predefined criteria, and pooled estimates were calculated separately within these strata. These analyses examine whether variation in interaction design characteristics is associated with differences in clinical effectiveness and user adherence, rather than comparing participant-level subgroups within individual trials. Section “Interaction Features and Clinical Effectiveness” reports the relationship between interaction features and clinical effectiveness, while section “Interaction Features and User Adherence” analyzes their association with user adherence.

Interaction Features and Clinical Effectiveness

As shown in Figure 7 and summarized in Table 3, differences were observed between high and low-scoring subgroups across several interaction features. In general, high-scoring subgroups tended to show more consistent patterns of treatment effects, whereas low-scoring subgroups showed more heterogeneous and less stable estimates. These findings should be interpreted as exploratory contrasts rather than confirmatory evidence. Full model outputs and sensitivity analyses are provided in Multimedia Appendix 6 [40-50].

For dialogue depth, the high group was significantly associated with better outcomes (SMD=−0.35, 95% CI −0.61 to −0.10; P=.007; I^2=0%), while the low group failed to reach significance and showed substantial heterogeneity (SMD=−0.54, 95% CI −1.20 to 0.11; P=.10; I^2=97.0%). This pattern was largely influenced by the trial of Fitzpatrick et al [41], which used the Woebot platform and reported an exceptionally large effect size (SMD=−3.30, 95% CI −4.12 to −2.48). The authors noted frequent misunderstandings and repetitive dialogues as

XSL•FO
RenderX

limitations, which may have shaped participants' engagement and contributed substantially to the heterogeneity in this subgroup.

By contrast, He et al [42] was consistently classified into high groups across all 6 interaction features, with particularly high scores in emotional responsiveness (4.75). He et al [42] found that enhanced emotional awareness significantly predicted superior therapeutic outcomes ($F_{2, 145}$=3.636; *P*=.03), a finding corroborated by the meta-analysis (SMD=−0.55, 95% CI −1.01 to −0.08; *P*=.02; I^2=80%). This provides indicative evidence that empathetic and adaptive chatbot responses may enhance clinical effectiveness. In addition, Liu et al [45] was classified into the low group for interaction frequency (score=3). The intervention demonstrated a significant negative effect (SMD −0.94, 95% CI −1.46 to −0.42), suggesting that insufficient interaction intensity may limit sustained therapeutic gains.

Taken together, these exploratory findings suggest that dialogue depth, emotional responsiveness, and interaction frequency may be associated with variation in clinical effectiveness across studies. However, given the limited number of included trials and the substantial heterogeneity observed in several subgroups, these patterns should be interpreted with caution and regarded as hypothesis-generating rather than definitive evidence.

Figure 7. Stratified meta-analytic estimates of clinical effectiveness according to interaction feature level (high vs low study-level coding) [40-50]. SMD: standardized mean difference.

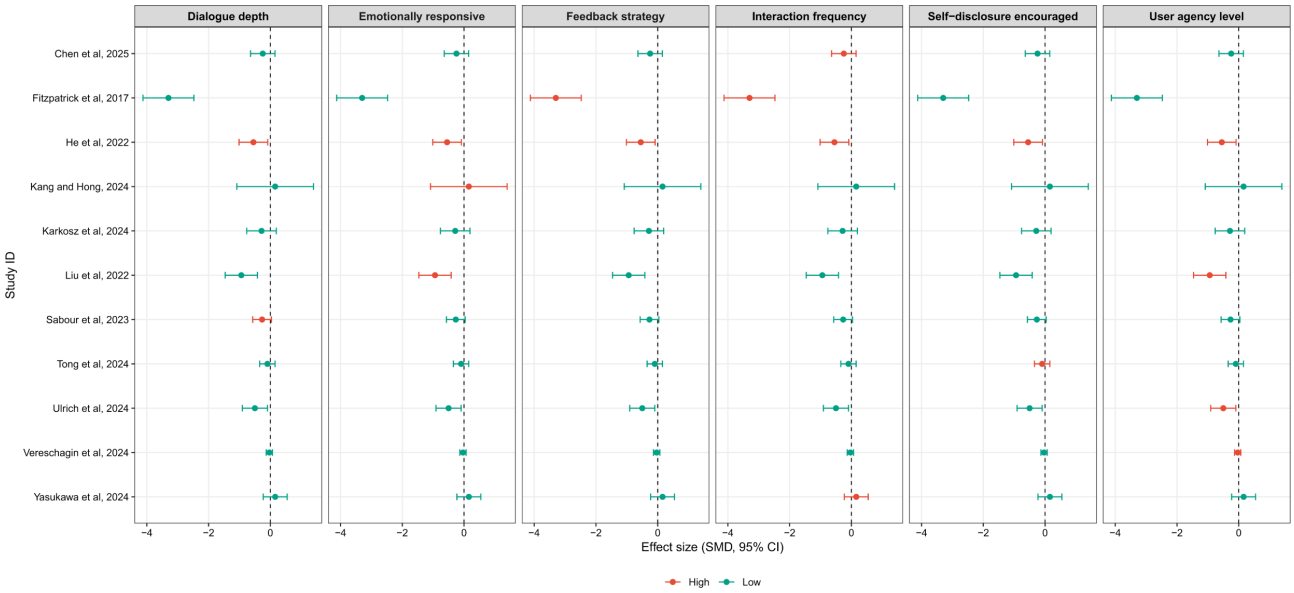

XSL•FO
RenderX

Table . Subgroup analyses of 6 interaction features and clinical effectiveness.

Interaction feature and subgroup	Pooled effect, SMD[a] (95% CI)	Z[b]	*P* value	I^2 (%)
Dialogue depth				
Low	–0.49 (–0.85 to –0.12)	2.63	.009	87
High	–0.35 (–0.61 to –0.10)	2.71	.007	0
Emotionally responsive				
Low	–0.42 (–0.76 to –0.09)	2.48	.01	90
High	–0.63 (–1.07 to –0.19)	2.81	.01	33
Feedback strategy				
Low	–0.21 (–0.38 to –0.04)	2.36	.02	59
High	–1.90 (–4.60 to –0.79)	1.98	.05	97
Interaction frequency				
Low	–0.26 (–0.48 to –0.05)	2.46	.01	65
High	–0.92 (–1.95 to 0.11)	1.74	.08	93
Self-disclosure encouraged				
Low	–0.52 (–0.90 to –0.14)	2.67	.008	90
High	–0.28 (–0.71 to 0.16)	1.25	.21	65
User agency level				
Low	–0.49 (–1.00 to 0.01)	1.92	.05	90
High	–0.46 (–0.90 to - 0.03)	2.08	.04	84

[a]SMD: standardized mean difference.

[b]Z denotes the Wald test statistic used for pooled odds ratios.

Interaction Features and User Adherence

Subgroup analyses were conducted to examine the association between 6 interaction features and user adherence, using the completion-based adherence definition described in the Methods (as shown in Figure 8 and Table 4). Overall, differences were observed between high- and low-scoring subgroups across several interaction features. In contrast to the clinical effectiveness outcomes reported in Figure 7, effect estimates for user adherence showed greater variability in magnitude and precision across studies.

For emotional responsiveness, the high-scoring subgroup showed a statistically significant association with adherence (OR 3.03, 95% CI 1.45 - 6.36; P=.003; I^2=14%), whereas the low-scoring subgroup did not reach statistical significance (OR 0.87, 95% CI 0.53 - 1.44; P=.59; I^2=70%). He et al [42], who were classified in the high group (score of 4.75), reported a statistically significant association between emotional awareness and adherence. He et al [42] illustrates how emotionally responsive chatbot interactions may be associated with adherence outcomes in specific contexts, rather than providing confirmatory evidence of a causal relationship.

In the domain of feedback strategy, high-scoring studies yielded robust and consistent associations (OR 4.62, 95% CI 1.96 - 10.91; P<.001; I^2=0%). Fitzpatrick et al [41] and He et al [42] both applied structured feedback mechanisms, and their findings largely drove the statistical significance of this subgroup. While these findings highlight a consistent pattern within the available data, they should be interpreted cautiously given the small number of contributing studies.

With regard to interaction frequency, the high group demonstrated a significant advantage (OR 4.18, 95% CI 1.10 - 15.87; P=.04; I^2=78%), whereas the low group did not (OR 0.75, 95% CI 0.49 - 1.13; P=.17; I^2=53%). Chen et al [40], and Yasukawa et al [50], which were included in the high-frequency subgroup, both reported patterns consistent with sustained adherence under more frequent interactions. Rather than implying a causal relationship, this contrast illustrates variability in adherence outcomes across different interaction intensity profiles.

XSL•FO
RenderX

Figure 8. Stratified meta-analytic estimates of user adherence according to interaction feature level (high vs low study-level coding) [40-50].

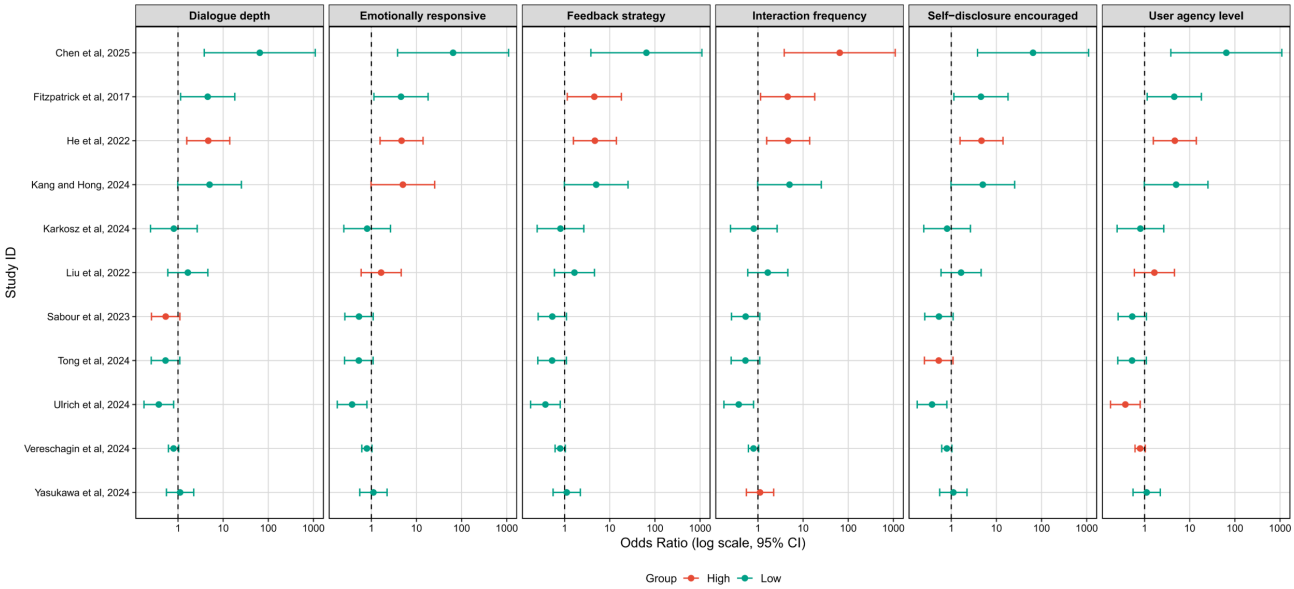

Table . Subgroup analyses of 6 interaction features and user adherence.

Interaction feature and subgroup	Pooled effect, OR[a] (95% CI)	Z[b]	*P* value	I^2 (%)
Dialogue depth				
Low	1.17 (0.68-2.01)	0.56	.57	72
High	1.52 (0.18-12.80)	0.38	.70	91
Emotionally responsive				
Low	0.87 (0.53-1.44)	0.55	.59	70
High	3.03 (1.45-6.36)	2.93	.003	14
Feedback strategy				
Low	0.90 (0.57-1.44)	0.43	.67	67
High	4.62 (1.96-10.91)	3.50	<.001	0
Interaction frequency				
Low	0.75 (0.49-1.13)	1.38	.17	53
High	4.18 (1.10-15.87)	2.10	.04	78
Self-disclosure encouraged				
Low	1.17 (0.68-2.01)	0.56	.57	72
High	1.51 (0.18-12.83)	0.38	.70	91
User agency level				
Low	1.48 (0.65-3.35)	0.94	.35	76
High	1.11 (0.49-2.53)	0.25	.80	81

[a]OR: odds ratio.

[b]Z denotes the Wald test statistic used for pooled odds ratios.

Taken together, these exploratory results suggest that emotional responsiveness, feedback strategy, and interaction frequency may be associated with variation in adherence outcomes across studies. In contrast, self-disclosure encouragement, dialogue depth, and user agency level showed weaker or nonsignificant associations. Given the heterogeneity of adherence estimates and the limited number of included trials, these patterns should be interpreted as hypothesis-generating and contingent on contextual and design-specific factors.

Small-Study Effects

Funnel plots were generated to assess potential small-study effects for the primary outcomes (Multimedia Appendix 7). For clinical effectiveness, visual inspection suggested possible funnel plot asymmetry. Egger regression test (*P*=.05) and the

XSL•FO
RenderX

Begg-Mazumdar rank correlation test (P=.04) both indicated statistically significant asymmetry. The trim-and-fill procedure imputed 3 potentially missing studies, and the adjusted pooled estimate was attenuated (SMD=−0.13, 95% CI −0.47 to 0.21) and no longer statistically significant. However, substantial heterogeneity was present (I^2=87%), and funnel plot asymmetry in this context may reflect genuine between-study variability rather than publication bias alone. For user adherence, statistical testing did not indicate significant funnel plot asymmetry (Begg test P=.07). Although the trim-and-fill method imputed 3 studies, the adjusted pooled estimate remained nonsignificant (OR 0.84, 95% CI 0.49 - 1.45), suggesting that small-study effects did not materially alter the overall conclusion for adherence outcomes.

Discussion

Summary of Key Findings

This systematic review evaluated the clinical effectiveness of AI-driven chatbots for depression and examined how interaction features relate to treatment outcomes and user adherence. AI-driven chatbots tended to reduce depressive symptoms, but after a more conservative analysis, this effect lost statistical significance, and studies remained heterogeneous.

No stable or statistically significant differences in user adherence were observed across chatbot types. This suggests that chatbot type alone may not explain variation in engagement patterns across studies. In addition, user adherence and clinical effectiveness did not show a stable one-to-one relationship.

By contrast, the exploratory analyses of interaction features revealed more informative patterns. Emotional responsiveness, feedback strategy, and interaction frequency showed more consistent associations with user adherence, whereas their relationships with clinical effectiveness were more mixed and heterogeneous. Dialogue depth, self-disclosure encouragement, and user agency showed weaker or more context-dependent associations. Overall, these findings suggest that interaction design may offer more explanatory value than chatbot type alone in understanding variation across studies.

Overall Interpretation

The findings of this review are broadly consistent with previous research on DMHIs. Previous studies suggest that technology-assisted interventions may reduce depressive symptoms to some extent, but their effect sizes are often small to moderate and vary across studies [53-55]. A similar pattern was observed in the present review. Although the pooled effect on depressive symptoms remained in a favorable direction, it was no longer statistically significant after applying the more conservative HKSJ method, and substantial heterogeneity remained. The wide PI further suggests that the true effect in a comparable future setting could vary considerably, potentially including no meaningful benefit. Taken together, these findings suggest that the overall clinical effectiveness of AI-driven chatbot interventions remains uncertain and may be influenced by differences in intervention design, implementation context, and study populations [56,57]. This cautious interpretation is also consistent with the GRADE assessment, which rated the certainty of evidence for clinical effectiveness as very low.

Another important finding of this review is that user adherence and clinical effectiveness did not show a stable or directly corresponding relationship. In this review, some studies reported relatively high levels of sustained engagement at the descriptive level [46,48,50], but this pattern did not consistently correspond to greater overall symptom improvement. This finding is consistent with previous research showing that user engagement is an important condition for the success of digital interventions, but it does not reliably predict clinical benefit [58,59]. In the present review, the overall adherence analysis also showed no significant difference between intervention and control groups. Heterogeneity remained high, and the PI was wide. This suggests that adherence outcomes may also vary substantially across studies and implementation settings. In this sense, adherence may be understood as an important condition for intervention success but not a sufficient one [55,59]. At the same time, the findings of this review suggest that chatbot type alone does not provide a stable explanation for variation in effectiveness or adherence across studies.

Building on this distinction, the present review further examined how different interaction features may relate to these divergent patterns. The exploratory analyses suggest that different interaction design elements may relate differently to user engagement and symptom change. In particular, the associations between interaction features and user adherence appeared to be clearer, whereas their relationships with clinical effectiveness were more mixed and heterogeneous. Taken together, these findings suggest that variation across studies may be better understood by focusing on interaction design features rather than chatbot type alone. In other words, differences across studies may be more closely related to how interactions are designed and implemented than to the underlying technical category itself [55,56,59]. At the same time, given the low certainty of evidence for the main outcomes, these interpretations should remain cautious. A more fine-grained understanding of interaction design may help explain variability across studies and inform future system development.

Potential Design Implications of Interaction Features

The exploratory analyses suggested that interaction features may provide a more informative lens than chatbot type alone for understanding variation across studies. To clarify these findings, the potential design implications are discussed separately for clinical effectiveness and user adherence. Given the heterogeneity of the evidence and the low certainty of the main outcomes, these implications should be understood as cautious and exploratory rather than prescriptive design recommendations.

Interaction Features and Clinical Effectiveness

The exploratory analyses suggested that the relationships between interaction features and clinical effectiveness were mixed and heterogeneous. Across the 6 interaction features, no single feature showed a uniformly stable association with symptom improvement. Instead, different features appeared to relate to treatment outcomes in different ways, and their

XSL•FO
RenderX

potential value seemed to depend on intervention context, therapeutic structure, and user characteristics. Taken together, these findings suggest that interaction design may contribute to clinical effectiveness, but the current evidence does not support simple or universal design conclusions.

Among the 6 features, dialogue depth showed one of the clearest patterns in relation to treatment outcomes. Deeper dialogue was associated with more consistent symptom improvement, whereas lower dialogue depth was linked to more variable treatment effects. These findings suggest that dialogue depth may support therapeutic benefit when it is appropriately structured. Existing research provides mixed evidence regarding the value of open-ended dialogue in DMHIs [51]. Many AI-driven chatbots rely on structured and guided conversational flows to maintain clarity and therapeutic focus [12]. While more open dialogue may increase perceived empathy and human-likeness [52], it may also increase cognitive load or lead to topic drift if not carefully designed [60]. Evidence from cross-cultural studies suggests that reflective and emotionally expressive dialogue can be beneficial, but its impact depends on contextual relevance and timing [61]. From a design perspective, dialogue depth may be better understood as an adaptive feature rather than a fixed attribute [62-64]. Integrating deeper dialogue within structured therapeutic components, such as journaling or behavioral activation tasks, may help maintain alignment with therapeutic objectives [65]. Overall, these observations highlight an important tension between expressive interaction and cognitive manageability. Dialogue depth may contribute to therapeutic alliance and perceived empathy under certain conditions [66], but its effectiveness likely depends on user characteristics, emotional state, and intervention structure.

Other interaction features showed less consistent relationships with symptom change. Emotional responsiveness was more consistently associated with user adherence and, to a lesser extent, with clinical outcomes. Interventions that incorporated more consistent and contextually appropriate emotional feedback tended to show more stable effects, but increasing emotional expressiveness without moderation is unlikely to improve outcomes directly. Research suggests that user engagement is influenced not only by emotional tone but also by the relevance and structure of therapeutic content [67]. Experimental work further indicates that improvements in emotional response mechanisms may enhance user trust and cognitive restructuring processes [68]. Emotional responsiveness may therefore be better understood as a process of calibration rather than intensity [69,70].

Feedback strategies also appeared to have a less stable relationship with clinical effectiveness than with adherence. Interventions that incorporated structured and personalized feedback tended to show more stable engagement patterns, but their influence on symptom improvement likely depends on how feedback is implemented and integrated within the intervention. Within internet-delivered cognitive behavioral therapy, individualized feedback has been associated with lower dropout even when symptom change is comparable [71]. This suggests that feedback may support treatment delivery, but its direct clinical impact is likely to vary across contexts.

Interaction frequency similarly showed different patterns for symptom change and sustained engagement. Lower-frequency interventions were associated with more consistent symptom improvement, whereas higher-frequency contact did not necessarily correspond to better short-term outcomes. Some studies have found that increased conversational exchange is associated with symptom improvement [72], while others report that gains may stabilize or diminish over longer periods of exposure [73]. This suggests that the effects of interaction frequency may not be linear. From a design perspective, interaction frequency should therefore be considered alongside timing, tone, and user context, and adaptive scheduling may be preferable to fixed high-frequency contact.

By comparison, self-disclosure encouragement and user agency showed weaker and more context-dependent relationships with clinical effectiveness. Encouraging self-disclosure was not consistently associated with improvements in symptoms. This contrasts with prior evidence indicating that self-disclosure is a key mechanism for building therapeutic alliance and enhancing engagement [74-76]. In face-to-face care, disclosure helps reduce stigma and promotes help-seeking, and digital interventions have attempted to replicate these processes through structured prompts for emotional expression and narrative sharing [77,78]. Taken together, these findings suggest that self-disclosure may not function in the same way across digital and in-person settings. Similarly, user agency was not consistently associated with clinical outcomes. Previous research supports the importance of perceived control in digital mental health systems [41,79], but agency may shape how users experience and engage with the intervention rather than directly improve outcomes.

Taken together, the exploratory findings suggest that interaction features may contribute to clinical effectiveness, but their relationships with symptom improvement are mixed and strongly shaped by context. Dialogue depth appeared to show the clearest potential relevance to therapeutic benefit, whereas emotional responsiveness, feedback strategy, and interaction frequency showed less stable associations with clinical outcomes. Self-disclosure encouragement and user agency showed weaker and more context-dependent patterns. These observations suggest that potential design implications for clinical effectiveness should be interpreted cautiously. At present, the evidence is better suited to generating conceptual implications than to supporting fixed design recommendations.

Interaction Features and User Adherence

The exploratory analyses suggested that interaction features showed clearer and more consistent patterns for user adherence than for clinical effectiveness. Across the 6 interaction features, emotional responsiveness, feedback strategy, and interaction frequency appeared to be more consistently associated with sustained engagement, whereas dialogue depth, self-disclosure encouragement, and user agency showed weaker or more context-dependent relationships. Taken together, these findings suggest that interaction design may be particularly important for understanding continued participation in chatbot-based interventions.

XSL•FO
RenderX

Emotional responsiveness showed one of the clearest relationships with user adherence. Interventions that incorporated more consistent and contextually appropriate emotional feedback tended to show more stable effects. However, this does not suggest that increasing emotional expressiveness without moderation will necessarily improve outcomes. Rather, emotional responsiveness may support engagement when it is calibrated appropriately, whereas excessive or poorly timed amplification may increase emotional burden or cognitive load and thereby undermine sustained engagement [80-82]. Qualitative studies have shown that users value personalized emotional support delivered at an appropriate pace [83]. Concerns about fully automated systems often extend beyond privacy and safety to include whether the system responds in a socially and emotionally appropriate manner [84]. Personalization therefore remains important, and adjusting tone, timing, and response frequency in relation to recent mood patterns may enhance usability and satisfaction [85].

Feedback strategy also appeared to play an important role in supporting user adherence. Interventions that incorporated structured and personalized feedback tended to show more stable engagement patterns. Timely and context-aware prompts can increase short-term engagement [86], whereas generic reminders may be less effective for sustaining engagement in real-world settings [87]. Broader research on guided digital interventions indicates that formats incorporating responsive elements or human support tend to achieve better retention than unguided approaches [88-90]. Methodological reviews further identify adherence and attrition as central determinants of overall effectiveness in DMHIs [91,92]. Feedback may therefore serve as a reinforcement cue that helps stabilize engagement over time, although excessive or poorly timed prompts may contribute to notification fatigue [93].

Interaction frequency similarly showed a clearer relationship with sustained engagement than with symptom change. Higher-frequency contact appeared more closely linked to continued participation. Interaction frequency may operate as both a structural and behavioral cue. A predictable rhythm of contact can reduce decision burden and support habit formation by transforming prompts into routine action cues [94-96]. From a design perspective, interaction frequency should be considered alongside timing, tone, and user context. Adaptive scheduling based on user behavior or mood patterns may be preferable to fixed high-frequency contact, and allowing users to adjust contact frequency may further support autonomy and reduce fatigue.

By comparison, dialogue depth showed a less stable relationship with sustained engagement. Although deeper dialogue was associated with more consistent symptom improvement, its relationship with continued use was less clear, and lower dialogue depth was linked to inconsistent adherence patterns. Existing research provides mixed evidence regarding the value of open-ended dialogue in DMHIs [51]. Many AI-driven chatbots rely on structured and guided conversational flows to maintain clarity and therapeutic focus [12]. While more open dialogue may increase perceived empathy and human-likeness [52], it may also increase cognitive load or lead to topic drift if not carefully designed [60]. Evidence from cross-cultural studies suggests that reflective and emotionally expressive dialogue can be beneficial, but its impact depends on contextual relevance and timing [61]. Research on digital behavior change interventions also suggests that early interactions should minimize cognitive demands to support initial engagement [97], and providing users with options to regulate conversational depth may reduce interaction fatigue [98].

Encouraging self-disclosure and increasing user agency also showed weaker and more context-dependent relationships with adherence. Although prior evidence indicates that self-disclosure can support therapeutic alliance and engagement [74-76], the present findings suggest that its effects in digital interventions may depend more strongly on timing, pacing, and context. Digital interventions have attempted to introduce structured prompts for emotional expression and narrative sharing [77,78], but willingness to disclose sensitive information may vary across regions and populations [99,100]. Privacy and ethical concerns remain important barriers, as fear of data misuse or personal information leakage can directly undermine trust and weaken adherence [101-103]. Similarly, user agency was not consistently associated with sustained engagement, although providing an appropriate degree of choice and control may still contribute to perceived engagement and satisfaction. Previous research supports the importance of perceived control in digital mental health systems [79]. At the same time, excessive freedom may increase interactional burden, whereas overly constrained interaction may reduce perceived control and engagement [104]. A balanced approach may therefore be more acceptable across different users and contexts [105].

Taken together, the exploratory findings suggest that interaction features may offer greater explanatory value for user adherence than for clinical effectiveness. In particular, emotional responsiveness, feedback strategy, and interaction frequency appeared to be more consistently related to sustained engagement, whereas dialogue depth, self-disclosure encouragement, and user agency seemed more dependent on timing, structure, and user readiness. However, given the heterogeneity of the evidence and the low certainty of the main outcomes, these patterns should be interpreted cautiously as conceptual implications rather than prescriptive design rules.

Limitations

This systematic review has several limitations that should be considered when interpreting the findings. First, the number of included studies was relatively small (n=11), particularly for subgroup analyses by AI chatbot type. This may have limited statistical power and reduced the generalizability of the findings. Second, substantial heterogeneity was observed across studies. Variations in intervention duration, chatbot design, delivery format, and target populations may have contributed to the variability and uncertainty in effect estimates, despite the use of random-effects models. The wide PIs in the main analyses further suggest that effects may vary across comparable future settings.

Third, interaction features were extracted and scored based on descriptions reported in the included studies rather than direct inspection of chatbot behavior. Although a structured coding protocol was applied and ratings were conducted independently

XSL•FO
RenderX

by 2 human-computer interaction experts, with adjudication by a third expert, some degree of subjectivity in feature interpretation was unavoidable. The feature-level findings should therefore be interpreted as exploratory. Finally, this meta-analysis relied exclusively on published studies, which may have introduced publication bias, as studies reporting nonsignificant or negative results are less likely to be published. In addition, statistical assessment indicated evidence of small-study effects for clinical effectiveness, and trim-and-fill adjustment attenuated the pooled estimate. However, given the substantial heterogeneity across studies, funnel plot asymmetry may partly reflect genuine between-study variability rather than publication bias alone. Taken together, these limitations indicate that the main findings should be interpreted cautiously.

Future Directions

Future research should adopt more standardized reporting of interaction features, clinical outcomes, and adherence measures. Studies that directly test the causal impact of specific interaction strategies, ideally within comparable therapeutic frameworks, are needed to clarify how different design elements relate to symptom change and sustained engagement. In addition, adaptive and personalized interaction models warrant further investigation to better accommodate diverse user needs, intervention contexts, and patterns of use. More transparent reporting of chatbot interaction design and more standardized documentation of intervention characteristics would also improve comparability across studies and support stronger evidence synthesis in the future.

Conclusion

The main contribution of this systematic review and meta-analysis is that it not only evaluated the clinical effectiveness of AI-driven chatbots for depression but also examined user adherence and interaction features within the same analytic framework. This allowed the review to move beyond the question of whether AI-driven chatbots may work and to explore possible reasons why findings vary across studies. The results showed a favorable trend for depressive symptom reduction, but the overall evidence remained uncertain. In addition, chatbot type alone did not provide a stable explanation for differences in user adherence. By contrast, interaction features, especially those related to sustained participation, appeared to offer a more informative perspective for understanding user engagement. Compared with previous reviews that mainly focused on overall effectiveness or differences between chatbot types, this study places greater emphasis on the role of interaction design in explaining variation in both outcomes and adherence. In this way, it offers a more fine-grained interpretive framework for the field. In practical terms, the findings suggest that the value of AI-driven chatbots for depression depends not only on the underlying technical architecture but also on how interactions are designed, structured, and supported over time. Future system development, clinical evaluation, and real-world implementation should therefore consider clinical outcomes and sustained engagement together, rather than relying only on short-term symptom change or chatbot type as the main basis for evaluation.

Acknowledgments

The authors declare that generative artificial intelligence (GAI) tools were used in a limited capacity to assist with language editing during manuscript preparation. According to the GAIDeT taxonomy (2025), the following task was delegated to GAI tools under full human supervision: proofreading and editing. The GAI tool used was ChatGPT (OpenAI). All aspects of the study design, data analysis, and interpretation were conducted by the authors. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets used and analyzed during this systematic review are available from the corresponding author upon reasonable request.

Authors' Contributions

TH contributed to data curation, investigation, formal analysis, and writing the original draft. SL contributed to validation and writing – review & editing. YW contributed to methodology and validation. WL contributed to methodology and validation. All authors contributed to writing – review & editing and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File, 23 KB - jmir_v28i1e88697_app1.docx]

Multimedia Appendix 2

XSL•FO
RenderX

Ratings of chatbot interaction features across included studies.
[DOCX File, 18 KB - jmir_v28i1e88697_app2.docx]

Multimedia Appendix 3
Leave-one-out sensitivity analysis of between-study heterogeneity (I^2).
[DOCX File, 1952 KB - jmir_v28i1e88697_app3.docx]

Multimedia Appendix 4
Risk of Bias (RoB) 2.
[XLSX File, 10 KB - jmir_v28i1e88697_app4.xlsx]

Multimedia Appendix 5
Grading of Recommendations Assessment, Development, and Evaluation (GRADE) summary.
[DOCX File, 14 KB - jmir_v28i1e88697_app5.docx]

Multimedia Appendix 6
Sensitivity analyses using the Hartung-Knapp-Sidik-Jonkman adjustment.
[DOCX File, 14 KB - jmir_v28i1e88697_app6.docx]

Multimedia Appendix 7
Funnel plot.
[DOCX File, 104 KB - jmir_v28i1e88697_app7.docx]

Checklist 1
PRISMA-S checklist.
[DOCX File, 17 KB - jmir_v28i1e88697_app8.docx]

Checklist 2
PRISMA checklist
[DOCX File, 272 KB - jmir_v28i1e88697_app9.docx]

References

1. Depressive disorder (depression). World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/depression [accessed 2026-06-05]
2. Vigo D, Thornicroft G, Atun R. Estimating the true global burden of mental illness. Lancet Psychiatry 2016 Feb;3(2):171-178. [doi: 10.1016/S2215-0366(15)00505-2] [Medline: 26851330]
3. Kazdin AE, Rabbitt SM. Novel models for delivering mental health services and reducing the burdens of mental illness. Clin Psychol Sci 2013 Apr;1(2):170-191. [doi: 10.1177/2167702612463566]
4. Clement S, Schauman O, Graham T, et al. What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies. Psychol Med 2015 Jan;45(1):11-27. [doi: 10.1017/S0033291714000129] [Medline: 24569086]
5. Mohr DC, Riper H, Schueller SM. A solution-focused research approach to achieve an implementable revolution in digital mental health. JAMA Psychiatry 2018 Feb 1;75(2):113-114. [doi: 10.1001/jamapsychiatry.2017.3838] [Medline: 29238805]
6. Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry 2019 Oct;18(3):325-336. [doi: 10.1002/wps.20673] [Medline: 31496095]
7. Karyotaki E, Efthimiou O, Miguel C, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021 Apr 1;78(4):361-371. [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
8. Garrido S, Millington C, Cheers D, et al. What works and what doesn't work? A systematic review of digital mental health interventions for depression and anxiety in young people. Front Psychiatry 2019;10:759. [doi: 10.3389/fpsyt.2019.00759] [Medline: 31798468]
9. Graham AK, Lattie EG, Powell BJ, et al. Implementation strategies for digital mental health interventions in health care settings. Am Psychol 2020 Nov;75(8):1080-1092. [doi: 10.1037/amp0000686] [Medline: 33252946]
10. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019 Jul 22;21(7):e12869. [doi: 10.2196/12869] [Medline: 31333198]

XSL•FO
RenderX

11. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11. [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
12. Boucher EM, Raiker JS. Engagement and retention in digital mental health interventions: a narrative review. BMC Digit Health 2024;2(1):52. [doi: 10.1186/s44247-024-00105-9]
13. Olawade DB, Wada OZ, Odetayo A, David-Olawade AC, Asaolu F, Eberhardt J. Enhancing mental health with artificial intelligence: current trends and future prospects. J Med Surg Public Health 2024 Aug;3:100099. [doi: 10.1016/j.glmedi.2024.100099]
14. Laranjo L, Dunn AG, Tong HL, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018 Sep 1;25(9):1248-1258. [doi: 10.1093/jamia/ocy072] [Medline: 30010941]
15. Li H, Zhang R, Lee YC, Kraut RE, Mohr DC. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med 2023 Dec 19;6(1):236. [doi: 10.1038/s41746-023-00979-5] [Medline: 38114588]
16. Wind TR, Rijkeboer M, Andersson G, Riper H. The COVID-19 pandemic: the “black swan” for mental health care and a turning point for e-health. Internet Interv 2020 Apr;20:100317. [doi: 10.1016/j.invent.2020.100317] [Medline: 32289019]
17. Abd-Alrazaq AA, Rababeh A, Alajlani M, Bewick BM, Househ M. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. J Med Internet Res 2020 Jul 13;22(7):e16021. [doi: 10.2196/16021] [Medline: 32673216]
18. Prochaska JJ, Vogel EA, Chieng A, et al. A therapeutic relational agent for reducing problematic substance use (Woebot): development and usability study. J Med Internet Res 2021 Mar 23;23(3):e24850. [doi: 10.2196/24850] [Medline: 33755028]
19. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. Front Digit Health 2022;4:847991. [doi: 10.3389/fdgth.2022.847991] [Medline: 35480848]
20. Guo Z, Lai A, Ive J, et al. Development and evaluation of HopeBot: an LLM-based chatbot for structured and interactive PHQ-9 depression screening. arXiv. Preprint posted online on Jan 14, 2026 URL: https://arxiv.org/abs/2507.05984 [accessed 2026-06-21] [doi: 10.48550/arXiv.2507.05984]
21. Kuhlmeier FO, Bauch L, Gnewuch U, Lüttke S. Designing chatbots to treat depression in youth: qualitative study. JMIR Hum Factors 2025 Jun 19;12:e66632. [doi: 10.2196/66632] [Medline: 40536944]
22. Ferrario A, Sedlakova J, Trachsel M. The role of humanization and robustness of large language models in conversational artificial intelligence for individuals with depression: a critical analysis. JMIR Ment Health 2024 Jul 2;11:e56569. [doi: 10.2196/56569] [Medline: 38958218]
23. Kim M, Oh J, Kim D, Shin J, Lee D. Understanding user preferences in developing a mental healthcare AI chatbot: a conjoint analysis approach. Int J Hum-Comput Interact 2025 Apr 18;41(8):4813-4821. [doi: 10.1080/10447318.2024.2353450]
24. Khamaj A. AI-enhanced chatbot for improving healthcare usability and accessibility for older adults. Alexandria Eng J 2025 Mar;116:202-213. [doi: 10.1016/j.aej.2024.12.090]
25. Furini M, Mariani M, Montagna S, Ferretti S. Conversational skills of LLM-based healthcare chatbot for personalized communications. Presented at: GoodIT ’24; Sep 4-6, 2024; Bremen, Germany p. 429-432 URL: https://dl.acm.org/doi/proceedings/10.1145/3677525 [accessed 2026-06-09] [doi: 10.1145/3677525.3678693]
26. Dong Y, Wu Y. Interacting with healthcare chatbot: effects of status cues and message contingency on AI credibility assessment. Int J Hum-Comput Interact 2025 Jun 3;41(11):6908-6920. [doi: 10.1080/10447318.2024.2387396]
27. Dosovitsky G, Pineda BS, Jacobson NC, Chang C, Escoredo M, Bunge EL. Artificial intelligence chatbot for depression: descriptive study of usage. JMIR Form Res 2020 Nov 13;4(11):e17065. [doi: 10.2196/17065] [Medline: 33185563]
28. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth 2018 Nov 23;6(11):e12106. [doi: 10.2196/12106] [Medline: 30470676]
29. Mantello PA, Ghotbi N, Ho MT, Mizutani F. Gauging public opinion of AI and emotionalized AI in healthcare: findings from a nationwide survey in Japan. AI Soc 2025 Jun;40(5):3735-3749. [doi: 10.1007/s00146-024-02126-4]
30. Jin E, Ryoo Y, Kim W, Song YG. Bridging the health literacy gap through AI chatbot design: the impact of gender and doctor cues on chatbot trust and acceptance. Internet Res 2025 May 27;35(3):1299-1329. [doi: 10.1108/INTR-08-2023-0702]
31. Phan TA, Bui VD. AI with a heart: how perceived authenticity and warmth shape trust in healthcare chatbots. J Mark Commun 2025:1-21. [doi: 10.1080/13527266.2025.2508887]
32. Kuhail MA, Alturki N, Thomas J, Alkhalifa AK, Alshardan A. Human-human vs human-AI therapy: an empirical study. Int J Hum Comput Interact 2025 Jun 3;41(11):6841-6852. [doi: 10.1080/10447318.2024.2385001]
33. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
34. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 6;327(7414):557-560. [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
35. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023 Dec;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]

XSL•FO
RenderX

36. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14(1):25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
37. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986 Sep;7(3):177-188. [doi: 10.1016/0197-2456(86)90046-2] [Medline: 3802833]
38. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634. [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
39. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics 1994 Dec;50(4):1088-1101. [doi: 10.2307/2533446] [Medline: 7786990]
40. Chen C, Lam KT, Yip KM, et al. Comparison of an AI chatbot with a nurse hotline in reducing anxiety and depression levels in the general population: pilot randomized controlled trial. JMIR Hum Factors 2025 Mar 6;12:e65785. [doi: 10.2196/65785] [Medline: 40048637]
41. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 6;4(2):e19. [doi: 10.2196/mental.7785] [Medline: 28588005]
42. He Y, Yang L, Zhu X, et al. Mental health chatbot for young adults with depressive symptoms during the COVID-19 pandemic: single-blind, three-arm randomized controlled trial. J Med Internet Res 2022 Nov 21;24(11):e40719. [doi: 10.2196/40719] [Medline: 36355633]
43. Kang B, Hong M. Digital interventions for reducing loneliness and depression in Korean college students: mixed methods evaluation. JMIR Form Res 2024 Sep 12;8:e58791. [doi: 10.2196/58791] [Medline: 39264705]
44. Karkosz S, Szymański R, Sanna K, Michałowski J. Effectiveness of a web-based and mobile therapy chatbot on anxiety and depressive symptoms in subclinical young adults: randomized controlled trial. JMIR Form Res 2024 Mar 20;8(1):e47960. [doi: 10.2196/47960] [Medline: 38506892]
45. Liu H, Peng H, Song X, Xu C, Zhang M. Using AI chatbots to provide self-help depression interventions for university students: a randomized trial of effectiveness. Internet Interv 2022 Mar;27:100495. [doi: 10.1016/j.invent.2022.100495] [Medline: 35059305]
46. Sabour S, Zhang W, Xiao X, et al. A chatbot for mental health support: exploring the impact of Emohaa on reducing mental distress in China. Front Digit Health 2023;5:1133987. [doi: 10.3389/fdgth.2023.1133987] [Medline: 37214342]
47. Tong ACY, Wong KTY, Chung WWT, Mak WWS. Effectiveness of topic-based chatbots on mental health self-care and mental well-being: randomized controlled trial. J Med Internet Res 2025 Apr 30;27:e70436. [doi: 10.2196/70436] [Medline: 40306635]
48. Ulrich S, Lienhard N, Künzli H, Kowatsch T. A chatbot-delivered stress management coaching for students (MISHA App): pilot randomized controlled trial. JMIR Mhealth Uhealth 2024 Jun 26;12:e54945. [doi: 10.2196/54945] [Medline: 38922677]
49. Vereschagin M, Wang AY, Richardson CG, et al. Effectiveness of the Minder mobile mental health and substance use intervention for university students: randomized controlled trial. J Med Internet Res 2024 Mar 27;26:e54287. [doi: 10.2196/54287] [Medline: 38536225]
50. Yasukawa S, Tanaka T, Yamane K, et al. A chatbot to improve adherence to internet-based cognitive-behavioural therapy among workers with subthreshold depression: a randomised controlled trial. BMJ Ment Health 2024 Jan 10;27(1):e300881. [doi: 10.1136/bmjment-2023-300881] [Medline: 38199786]
51. Joy GV, Joy FE, Nashwan AJ. Between empathy and algorithms: navigating interpersonal dynamics in AI-augmented mental health care- discursive review. Asian J Psychiatr 2026 Feb;116:104816. [doi: 10.1016/j.ajp.2025.104816] [Medline: 41494438]
52. Baik RL, Lee S, Xie SJ, Liao W, Hwang EH, Yuwen W. Adapting communication styles in health chatbot using large language models to support family caregivers from multicultural backgrounds. Presented at: CHI EA '25; Apr 26 to May 1, 2025; Yokohama, Japan p. 1-8 URL: https://dl.acm.org/doi/proceedings/10.1145/3706599 [accessed 2026-06-09] [doi: 10.1145/3706599.3719711]
53. Firth J, Torous J, Nicholas J, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry 2017 Oct;16(3):287-298. [doi: 10.1002/wps.20472] [Medline: 28941113]
54. Fu Z, Burger H, Arjadi R, Bockting CLH. Effectiveness of digital psychological interventions for mental health problems in low-income and middle-income countries: a systematic review and meta-analysis. Lancet Psychiatry 2020 Oct;7(10):851-864. [doi: 10.1016/S2215-0366(20)30256-X] [Medline: 32866459]
55. Plessen CY, Panagiotopoulou OM, Tong L, Cuijpers P, Karyotaki E. Digital mental health interventions for the treatment of depression: a multiverse meta-analysis. J Affect Disord 2025 Jan 15;369:1031-1044. [doi: 10.1016/j.jad.2024.10.018] [Medline: 39419189]
56. Bakhti R, Daler H, Ogunro H, Hope S, Hargreaves D, Nicholls D. Exploring engagement with and effectiveness of digital mental health interventions in young people of different ethnicities: systematic review. J Med Internet Res 2025 Apr 7;27:e68544. [doi: 10.2196/68544] [Medline: 40194267]

XSL•FO
RenderX

57. Zagorscak P, Heinrich M, Bohn J, Stein J, Knaevelsrud C. How individuals change during internet-based interventions for depression: a randomized controlled trial comparing standardized and individualized feedback. Brain Behav 2020 Jan;10(1):e01484. [doi: 10.1002/brb3.1484] [Medline: 31777204]
58. Lipschitz JM, Pike CK, Hogan TP, Murphy SA, Burdick KE. The engagement problem: a review of engagement with digital mental health interventions and recommendations for a path forward. Curr Treat Options Psychiatry 2023 Sep;10(3):119-135. [doi: 10.1007/s40501-023-00297-3] [Medline: 38390026]
59. Forbes A, Keleher MR, Venditto M, DiBiasi F. Assessing patient adherence to and engagement with digital interventions for depression in clinical trials: systematic literature review. J Med Internet Res 2023 Aug 11;25:e43727. [doi: 10.2196/43727] [Medline: 37566447]
60. Hudon A, Stip E. Delusional experiences emerging from AI chatbot interactions or “AI Psychosis”. JMIR Ment Health 2025 Dec 3;12(1):e85799. [doi: 10.2196/85799] [Medline: 41273266]
61. Chin H, Song H, Baek G, et al. The potential of chatbots for emotional support and promoting mental well-being in different cultures: mixed methods study. J Med Internet Res 2023 Oct 20;25:e51712. [doi: 10.2196/51712] [Medline: 37862063]
62. Ahmad R, Siemon D, Gnewuch U, Robra-Bissantz S. Designing personality-adaptive conversational agents for mental health care. Inf Syst Front 2022;24(3):923-943. [doi: 10.1007/s10796-022-10254-9] [Medline: 35250365]
63. Morris RR, Kouddous K, Kshirsagar R, Schueller SM. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. J Med Internet Res 2018 Jun 26;20(6):e10148. [doi: 10.2196/10148] [Medline: 29945856]
64. Kocaballi AB, Sezgin E, Clark L, et al. Design and evaluation challenges of conversational agents in health care and well-being: selective review study. J Med Internet Res 2022 Nov 15;24(11):e38525. [doi: 10.2196/38525] [Medline: 36378515]
65. Latif M, Awan F, Gul M, et al. Preliminary evaluation of a culturally adapted CBT-based online programme for depression and anxiety from a lower middle-income country. Cogn Behav Therapist 2021;14:e36. [doi: 10.1017/S1754470X21000313]
66. Tremain H, McEnery C, Fletcher K, Murray G. The therapeutic alliance in digital mental health interventions for serious mental illnesses: narrative review. JMIR Ment Health 2020 Aug 7;7(8):e17204. [doi: 10.2196/17204] [Medline: 32763881]
67. Yarrington JS, Metts A, Vargas JH, Couto DD, Marafon T, Cohen ZD. Comparative effectiveness and user-rated helpfulness of digital just-in-time adaptive interventions for psychological distress. J Affect Disord 2025 Dec 1;390:119878. [doi: 10.1016/j.jad.2025.119878] [Medline: 40652979]
68. Rządeczka M, Sterna A, Stolińska J, Kaczyńska P, Moskalewicz M. The efficacy of conversational AI in rectifying the theory-of-mind and autonomy biases: comparative analysis. JMIR Ment Health 2025 Feb 7;12(1):e64396. [doi: 10.2196/64396] [Medline: 39919295]
69. Baggett KM, Davis B, Sheeber L, et al. Optimizing social-emotional-communication development in infants of mothers with depression: protocol for a randomized controlled trial of a mobile intervention targeting depression and responsive parenting. JMIR Res Protoc 2021 Aug 18;10(8):e31072. [doi: 10.2196/31072] [Medline: 34406122]
70. Mansoor M, Hamide A, Tran T. Conversational AI in pediatric mental health: a narrative review. Children (Basel) 2025 Mar 14;12(3):359. [doi: 10.3390/children12030359] [Medline: 40150640]
71. Zagorscak P, Heinrich M, Sommer D, Wagner B, Knaevelsrud C. Benefits of individualized feedback in internet-based interventions for depression: a randomized controlled trial. Psychother Psychosom 2018;87(1):32-45. [doi: 10.1159/000481515] [Medline: 29306945]
72. Zhang R, Nicholas J, Knapp AA, et al. Clinically meaningful use of mental health apps and its effects on depression: mixed methods study. J Med Internet Res 2019 Dec 20;21(12):e15644. [doi: 10.2196/15644] [Medline: 31859682]
73. Zhong W, Luo J, Zhang H. The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: a systematic review and meta-analysis. J Affect Disord 2024 Jul 1;356:459-469. [doi: 10.1016/j.jad.2024.04.057] [Medline: 38631422]
74. Doan U, Hong D, Hitchcock C. Please, just talk to me: self-disclosure mediates the effect of autobiographical memory specificity on adolescent self-harm and depressive symptoms in a UK population-based study. J Affect Disord 2025 May 1;376:10-17. [doi: 10.1016/j.jad.2025.01.141] [Medline: 39892759]
75. Gonsalves PP, Nair R, Roy M, Pal S, Michelson D. A systematic review and lived experience synthesis of self-disclosure as an active ingredient in interventions for adolescents and young adults with anxiety and depression. Adm Policy Ment Health 2023 May;50(3):488-505. [doi: 10.1007/s10488-023-01253-2] [Medline: 36738384]
76. Goh M, Jeong H, Yoo JH, Han O. Self-disclosure in digital healthcare: enhancing user engagement. Presented at: 2023 IEEE International Conference on Agents (ICA); Dec 4-6, 2023; Kyoto, Japan p. 63-68. [doi: 10.1109/ICA58824.2023.00020]
77. Cui Y, Lee YJ, Jamieson J, Yamashita N, Lee YC. Exploring effects of chatbot’s interpretation and self-disclosure on mental illness stigma. Proc ACM Hum-Comput Interact 2024 Apr 17;8(CSCW1):1-33. [doi: 10.1145/3637329]
78. Ma X, Hancock J, Naaman M. Anonymity, intimacy and self-disclosure in social media. Presented at: CHI’16; May 7-12, 2016; San Jose, CA p. 3857-3869 URL: https://dl.acm.org/doi/proceedings/10.1145/2858036 [accessed 2026-06-23] [doi: 10.1145/2858036.2858414]

XSL•FO
RenderX

79. Khosravi M, Zare Z, Mojtabaeian SM, Izadi R. Artificial intelligence and decision-making in healthcare: a thematic analysis of a systematic review of reviews. Health Serv Res Manag Epidemiol 2024;11:23333928241234863. [doi: 10.1177/23333928241234863] [Medline: 38449840]
80. Wang L, Bhanushali T, Huang Z, Yang J, Badami S, Hightow-Weidman L. Evaluating generative AI in mental health: systematic review of capabilities and limitations. JMIR Ment Health 2025 May 15;12(1):e70014. [doi: 10.2196/70014] [Medline: 40373033]
81. Kahlon MK, Aksan N, Aubrey R, et al. Effect of layperson-delivered, empathy-focused program of telephone calls on loneliness, depression, and anxiety among adults during the COVID-19 pandemic: a randomized clinical trial. JAMA Psychiatry 2021 Jun 1;78(6):616-622. [doi: 10.1001/jamapsychiatry.2021.0113] [Medline: 33620417]
82. Franke Föyen L, Zapel E, Lekander M, Hedman-Lagerlöf E, Lindsäter E. Artificial intelligence vs. human expert: licensed mental health clinicians' blinded evaluation of AI-generated and expert psychological advice on quality, empathy, and perceived authorship. Internet Interv 2025 Sep;41:100841. [doi: 10.1016/j.invent.2025.100841] [Medline: 40525210]
83. Bisconti N, Odier M, Becker M, Bullock K. Feasibility and acceptability of a mobile app-based TEAM-CBT (testing empathy assessment methods-cognitive behavioral therapy) intervention (feeling good) for depression: secondary data analysis. JMIR Ment Health 2024 May 10;11(1):e52369. [doi: 10.2196/52369] [Medline: 38728080]
84. Lee HS, Wright C, Ferranto J, et al. Artificial intelligence conversational agents in mental health: patients see potential, but prefer humans in the loop. Front Psychiatry 2024;15:1505024. [doi: 10.3389/fpsyt.2024.1505024] [Medline: 39957757]
85. Fouyaxis J, Bidargaddi N, Du W, Looi JCL, Lipschitz J. Critical design decisions and user demographics in enhancing real-time digital mental health interventions: a systematic review. Digit Health 2024;10:20552076241306782. [doi: 10.1177/20552076241306782] [Medline: 39687526]
86. Bidargaddi N, Almirall D, Murphy S, et al. To prompt or not to prompt? A microrandomized trial of time-varying push notifications to increase proximal engagement with a mobile health app. JMIR Mhealth Uhealth 2018 Nov 29;6(11):e10123. [doi: 10.2196/10123] [Medline: 30497999]
87. Teepe GW, Da Fonseca A, Kleim B, et al. Just-in-time adaptive mechanisms of popular mobile apps for individuals with depression: systematic app search and literature review. J Med Internet Res 2021 Sep 28;23(9):e29412. [doi: 10.2196/29412] [Medline: 34309569]
88. Leung C, Pei J, Hudec K, Shams F, Munthali R, Vigo D. The effects of nonclinician guidance on effectiveness and process outcomes in digital mental health interventions: systematic review and meta-analysis. J Med Internet Res 2022 Jun 15;24(6):e36004. [doi: 10.2196/36004] [Medline: 35511463]
89. Mercadal Rotger J, Cabré V. Therapeutic alliance in online and face-to-face psychological treatment: comparative study. JMIR Ment Health 2022 May 2;9(5):e36775. [doi: 10.2196/36775] [Medline: 35499910]
90. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: a systematic and meta-analytic review. J Consult Clin Psychol 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
91. Torous J, Lipschitz J, Ng M, Firth J. Dropout rates in clinical trials of smartphone apps for depressive symptoms: a systematic review and meta-analysis. J Affect Disord 2020 Feb 15;263:413-419. [doi: 10.1016/j.jad.2019.11.167] [Medline: 31969272]
92. Wu A, Scult MA, Barnes ED, Betancourt JA, Falk A, Gunning FM. Smartphone apps for depression and anxiety: a systematic review and meta-analysis of techniques to increase engagement. NPJ Digit Med 2021 Feb 11;4(1):20. [doi: 10.1038/s41746-021-00386-8] [Medline: 33574573]
93. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020 Sep 29;22(9):e20283. [doi: 10.2196/20283] [Medline: 32990635]
94. Malouin-Lachance A, Capolupo J, Laplante C, Hudon A. Does the digital therapeutic alliance exist? Integrative review. JMIR Ment Health 2025 Feb 7;12:e69294. [doi: 10.2196/69294] [Medline: 39924298]
95. Arapakis I, Lalmas M, Cambazoglu BB, Marcos MC, Jose JM. User engagement in online news: under the scope of sentiment, interest, affect, and gaze. J Assoc Inf Sci Technol 2014 Oct;65(10):1988-2005 [FREE Full text] [doi: 10.1002/asi.23096]
96. Lattie EG, Schueller SM, Sargent E, et al. Uptake and usage of IntelliCare: a publicly available suite of mental health and well-being apps. Internet Interv 2016 May;4(2):152-158. [doi: 10.1016/j.invent.2016.06.003] [Medline: 27398319]
97. Pelly M, Fatehi F, Liew D, Verdejo-Garcia A. Novel behaviour change frameworks for digital health interventions: a critical review. J Health Psychol 2023 Sep;28(10):970-983. [doi: 10.1177/13591053231164499] [Medline: 37051615]
98. Dietrich F, Arenz A, Reinecke L. What constitutes experiences of autonomy in digital technology use? A (computational) scoping review through the lens of self-determination theory. Interact Comput 2026 Apr 13;38(3):487-500. [doi: 10.1093/iwc/iwae050]
99. Mozafari N, Weiger WH, Hammerschmidt M. Trust me, I'm a bot – repercussions of chatbot disclosure in different service frontline settings. J Serv Manag 2022 Feb 28;33(2):221-245. [doi: 10.1108/JOSM-10-2020-0380]
100. Wang A, Zhou Y, Ma H, et al. Preparing for aging: understanding middle-aged user acceptance of AI chatbots through the technology acceptance model. Digit Health 2024;10:20552076241284903. [doi: 10.1177/20552076241284903] [Medline: 39381827]

XSL•FO
RenderX

101. Prizeman K, McCabe C, Weinstein N. Stigma and its impact on disclosure and mental health secrecy in young people with clinical depression symptoms: a qualitative analysis. PLoS ONE 2024;19(1):e0296221. [doi: 10.1371/journal.pone.0296221] [Medline: 38180968]
102. Wu Y, Shao J, Zhang D, et al. Pathways from self-disclosure to medical coping strategy among adolescents with moderate and major depression during the COVID-19 pandemic: a mediation of self-efficacy. Front Psychiatry 2022;13:976386. [doi: 10.3389/fpsyt.2022.976386]
103. Kim J, Lee K, Kim W, Jeong N, Kim J, Song H. Empathetic pedagogical agent: mitigating harmful effects of negative feedback through self-disclosure. Int J Hum Comput Interact 2025 Aug 3;41(15):9366-9383. [doi: 10.1080/10447318.2024.2425881]
104. Cross SP, Alvarez-Jimenez M. The digital cumulative complexity model: a framework for improving engagement in digital mental health interventions. Front Psychiatry 2024;15:1382726. [doi: 10.3389/fpsyt.2024.1382726] [Medline: 39290300]
105. Yang Y, Tavares J, Oliveira T. A new research model for artificial intelligence-based well-being chatbot engagement: survey study. JMIR Hum Factors 2024 Nov 11;11:e59908. [doi: 10.2196/59908] [Medline: 39527812]

Abbreviations

AI: artificial intelligence
DMHI: digital mental health intervention
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
HKSJ: Hartung-Knapp-Sidik-Jonkman
iCBT: internet-based cognitive behavioral therapy
LLM: large language model
ML: machine learning
NLP: natural language processing
OR: odds ratio
PHQ-9: Patient Health Questionnaire-9
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
RCT: randomized controlled trial
RoB 2: Risk of Bias tool version 2
RQ: research question
SMD: standardized mean difference
WHO: World Health Organization

Edited by S Brini; submitted 01.Dec.2025; peer-reviewed by D Shin, WH Gu; revised version received 14.May.2026; accepted 19.May.2026; published 30.Jun.2026.

Please cite as:
Huang T, Li S, Wang Y, Liu W
Therapeutic Interaction Features of AI Chatbots in Depression Interventions: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e88697
URL: https://www.jmir.org/2026/1/e88697
doi:10.2196/88697

© Ting Huang, Shuangyu Li, Yanzhong Wang, Wei Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Care Robots as Emerging Health Technologies: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Gaeun Kim[1], PhD; Jinmi Jeon[2,3], MSN

[1]Department of Nursing, College of Nursing, Keimyung University, 1095, Dalgubeol-daero, Dalseo-gu, Daegu, Republic of South Korea

[2]Graduate School, Department of Nursing, Keimyung University, Daegu, Republic of South Korea

[3]Department of Nursing, Keimyung University Dongsan Hospital, Daegu, Republic of Korea

Corresponding Author:
Gaeun Kim, PhD
Department of Nursing, College of Nursing, Keimyung University, 1095, Dalgubeol-daero, Dalseo-gu, Daegu, Republic of South Korea

Abstract

Background: Global aging and health care workforce shortages are increasing demand for therapeutic support among older adult and pediatric populations. Care robots, an umbrella term for socially assistive robots, companion and pet platforms, and therapeutic humanoids sharing a social interaction component, have been proposed as embodied digital health adjuncts, but prior syntheses have remained population- or platform-specific, leaving cross-population and cross-platform effects unquantified.

Objective: This review quantifies the pooled effects of care-robot interventions across 7 prespecified outcome domains, examines robot platform, target population, and intervention-duration moderators, and grades evidence certainty.

Methods: Following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020, we searched 5 databases (PubMed, including MEDLINE, Embase, Cochrane CENTRAL, CINAHL, APA PsycINFO) and 2 trial registries (ClinicalTrials.gov; World Health Organization International Clinical Trials Registry Platform) through April 29, 2026, with no language restriction. Eligible studies were randomized controlled trials of embodied care robots versus standard care, active controls, or waitlist. Random-effects meta-analysis used Hedges *g* with Hartung-Knapp-Sidik-Jonkman correction; for domains with k≥4, we report 95% CIs and prediction intervals (PIs). Risk of bias used the Cochrane Risk of Bias 2 tool; certainty used the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework.

Results: A total of 34 randomized controlled studies (n=2476; 17 countries; 2015‐2024) met inclusion criteria; 20 contributed to at least 1 meta-analysis, and 14 entered narrative synthesis. Under Hartung-Knapp-Sidik-Jonkman pooling, only neuropsychiatric symptoms reached significance: Hedges *g*=0.44 (95% CI 0.03-0.84; 95% PI −0.42 to 1.30; k=6; *P*=.04). The remaining 6 domains were favorable but nonsignificant: quality of life Hedges *g*=0.15 (95% CI −0.41 to 0.71; 95% PI −0.86 to 1.16; k=5); depression Hedges *g*=0.20 (95% CI −0.08 to 0.49; 95% PI −0.36 to 0.76; k=7); agitation Hedges *g*=0.32 (95% CI −0.07 to 0.71; 95% PI −0.26 to 0.89; k=4); stress and pain Hedges *g*=0.53 (95% CI −0.48 to 1.53; 95% PI −1.57 to 2.62; k=6); social-communicative skills Hedges *g*=0.45 (95% CI −0.52 to 1.42; 95% PI −1.66 to 2.56; k=5); and cognitive function Hedges *g*=0.18 (95% CI −0.62 to 0.98; 95% PI −1.11 to 1.47; k=4). All 7 PIs encompassed the null, indicating no assured effect in new settings. The neuropsychiatric symptom result was fragile: significance was lost in 5 of 6 leave-one-out iterations and on excluding Petersen 2016 (data extraction ambiguity), yielding a Hedges *g* of 0.40 (*P*=.10). GRADE certainty was low for neuropsychiatric symptoms and very low for the remaining 6 domains.

Conclusions: Across cross-population, cross-platform randomized controlled trial pooling, the evidence does not yet support routine clinical adoption of care robots. Potential benefits are narrow, of low certainty, and contingent on adequately powered multicenter confirmatory trials with PI-informed sample sizes. Care robots are best framed as facilitator-supported adjuncts that augment, rather than substitute for, human-delivered care.

(J Med Internet Res 2026;28:e95232) doi:10.2196/95232

KEYWORDS

socially assistive robots; care robots; digital health; robotics; systematic review; meta-analysis

Introduction

Health care systems worldwide face 2 parallel and increasingly convergent pressures that are reshaping how direct patient care is organized. The first is demographic: by 2050, the global population aged 60 years and older is projected to reach 2.1 billion, with corresponding rises in the prevalence of chronic disease, cognitive impairment, and dementia [1]. The second is structural: a projected deficit of 12.9 million health care

XSL•FO
RenderX

professionals by 2035 [2] is already constraining the dose and continuity of one-to-one, relational care, most acutely in long-term care and in specialized pediatric therapy. Beyond these aggregate workforce figures, specific populations—children with autism spectrum disorder (ASD) and other neurodevelopmental conditions, pediatric inpatients undergoing anxiety-provoking procedures, and older adults with dementia in residential care—share a common practical problem: they benefit from structured, predictable, attention-capturing human interaction that is difficult to deliver at scale.

Care robots have emerged as one technological response to this shared problem. They are physically embodied platforms designed for direct patient interaction and include socially assistive robots such as PARO, NAO, Pepper, Kabochan, and Kaspar; companion and pet robots such as Joy for All, iRobiQ, and MiRo; and specialized therapeutic humanoids [3]. Unlike disembodied digital health tools—telehealth platforms, mobile health apps, or wearable sensors—care robots offer multimodal physical presence (tactile, visual, auditory, and increasingly conversational) that may elicit social and emotional responses in ways that screen-based interventions do not [4]. In older adults with cognitive impairment, tangible tactile interaction with companion robots is thought to support emotional regulation and person-centered engagement in a manner consistent with Kitwood's model of dementia care [5]. In children with ASD, the simplified, predictable, and repeatable social behavior of humanoid robots has been proposed as a scaffold for practicing social-communicative skills that children may find difficult to rehearse with human partners [6]. Acceptance and usability—rather than efficacy alone—are recognized as critical determinants of whether such platforms translate from trials into practice [7,8]. Beyond stand-alone use, socially assistive robots are increasingly deployed as components of broader technological ecosystems that integrate sensing, data capture, and connected services rather than as isolated devices [9,10].

Despite this shared theoretical rationale, the empirical literature on care robots has developed in siloed streams. Prior systematic reviews and meta-analyses have either restricted themselves to a single population—Pu et al [11] in older adults, Leng et al [12] in dementia care with pet robots, and Yu et al [13] in dementia broadly; or Scassellati et al [6] in children with ASD—or to a single platform, most commonly PARO in dementia care. The collective consequence is that the most clinically actionable question—whether observed robot-intervention effects are driven primarily by platform (eg, PARO vs humanoid) or by population (eg, dementia vs ASD vs pediatric procedural anxiety)—cannot be answered from any one of these reviews alone, because each review holds either platform or population fixed by design. Two further consequences follow for practice. First, procurement, deployment, and health technology assessment decisions are increasingly made across populations (eg, a long-term care operator deciding among PARO, NAO, and companion pet platforms for mixed-acuity settings), yet there is no published synthesis that places the platform effects observed in one population against those observed in another within a single analytic framework. Second, each single-population review has necessarily treated implementation questions—facilitator training, acceptance, setting fit—in isolation, missing the opportunity to identify cross-cutting patterns. Because care-robot randomized controlled trials (RCTs) have grown substantially over the past decade (2015 - 2024), the minimum number of studies (k) required for meaningful moderation analyses of robot platform and target population is now within reach.

Accordingly, this systematic review and meta-analysis had 4 prespecified objectives. First, to quantify the pooled effects of care-robot interventions across 7 prespecified outcome domains in a cross-population sample that explicitly includes both older adult and pediatric populations. Second, to examine the moderating effects of robot platform (PARO vs humanoid vs other companion and pet), target population (dementia or older adult vs pediatric or ASD vs other), and intervention duration (≥10 weeks vs <10 weeks) through prespecified subgroup analyses. Third, to appraise the methodological quality of included studies using the Cochrane Risk of Bias 2 (RoB 2) tool. Fourth, to evaluate the certainty of evidence across domains using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework and to translate the resulting evidence profile into realistic, targeted recommendations for clinical use. To our knowledge, this is among the first quantitative syntheses to pool care-robot RCTs across both older adult and pediatric populations within a common framework; because per-subgroup study counts are small, we frame the resulting comparison across robot platform and populations as exploratory and hypothesis-generating rather than confirmatory.

Methods

Overview

This systematic review and meta-analysis is reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 Statement [14], the PRISMA 2020 for Abstracts extension, the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search) extension for reporting literature searches [15], and the methodological standards of the *Cochrane Handbook for Systematic Reviews of Interventions* [16]. Completed PRISMA 2020 Expanded, PRISMA 2020 for Abstracts, and PRISMA-S checklists are provided as Checklist 1-3. The review was not prospectively registered in PROSPERO (International Prospective Register of Systematic Reviews); an a priori review protocol specifying eligibility criteria, outcome domains, analytic plan, and risk-of-bias assessment was developed and dated before the formal search and is available from the corresponding author on reasonable request.

Eligibility Criteria

Eligibility criteria were defined using the Population, Intervention, Comparison, Outcomes, and Study Design framework. The population comprised patients of any age, in any clinical or community setting, with no restrictions on primary diagnosis or age group, consistent with the cross-population aim of the review. The intervention was any physically embodied care robot designed for direct patient interaction, including socially assistive, companion, pet, and

XSL•FO
RenderX

therapeutic humanoid platforms. Because the eligibility criteria required a social or interactive component, all included platforms are, by definition, social robots or socially assistive robots; we retain "care robots" as the umbrella term while noting that every included device met the social robots or socially assistive robots criterion. Surgical robots, passive rehabilitation exoskeletons without a social interaction component, and clinician-controlled telepresence robots were excluded. Comparators were standard care without robot interaction, attention-matched active controls (eg, reading or tablet-based sessions of equivalent duration), placebo conditions (deactivated robot or plush equivalent), and waitlist controls. Outcomes were restricted to quantitative, validated patient-reported outcome measures within 7 prespecified outcome domains defined in the a priori protocol: neuropsychiatric symptoms (NPS), quality of life, depression, agitation, stress and pain, social-communicative skills, and cognitive function. Other outcomes encountered during extraction (eg, motor activity and biomarkers) were retained in the narrative synthesis but not pooled. Eligible study designs were individual RCTs and cluster RCTs. Single-arm before-after studies, case reports, qualitative-only studies, feasibility studies without an outcome comparator, and conference abstracts lacking full numerical data were excluded.

Information Sources

We searched 5 electronic databases: PubMed (MEDLINE), Embase (Elsevier), the Cochrane Central Register of Controlled Trials (CENTRAL), CINAHL Complete (EBSCO), and APA PsycINFO. Two trial registries (ClinicalTrials.gov and the World Health Organization [WHO] International Clinical Trials Registry Platform [ICTRP]) were searched for completed trials with reported results. Reference lists of all included studies and of 4 prior systematic reviews in the field were screened, and forward citation tracking was performed via Google Scholar. Searches covered database inception through April 29, 2026. No language or publication date restrictions were applied at the search stage. At full-text screening, articles for which no English or Korean full text or professional translation was obtainable despite repeated retrieval attempts were excluded (n=5); this is reported in the PRISMA flow diagram.

Search Strategy

Search strategies were developed iteratively in consultation with a health-sciences librarian and combined controlled vocabulary (MeSH [Medical Subject Headings] terms in PubMed and CENTRAL; EMTREE terms in Embase; CINAHL Subject Headings; APA Thesaurus terms in PsycINFO) with free-text terms mapped with field modifiers (.ti,ab,kw and database-specific equivalents), truncation, and proximity operators. Three concept blocks were combined with AND: (1) care robots, including both generic terms ("social robot," "socially assistive robot," "companion robot," "robotic pet," "humanoid robot," and "therapeutic robot") and platform-specific trade names (PARO, NAO, Pepper, Kaspar, Kabochan, CommU, Joy for All, MiRo, and iRobiQ); (2) health care context terms (eg, "nursing home," "long-term care," "dementia care," "pediatric," "hospital," "rehabilitation," and "outpatient"); and (3) patient-outcome terms covering the 7 prespecified domains (eg, "quality of life," "neuropsychiatric symptoms," "agitation," "anxiety," "pain," "social skills," and "mental status"). Full database-specific search strategies, exact search dates, and the number of records retrieved per database are reported in Checklist 1 (PRISMA-S compliant).

Selection Process

All retrieved records were imported into Covidence systematic review software (Veritas Health Innovation) and deduplicated automatically with manual verification. Two reviewers (GK and JJ) independently screened all unique records by title and abstract and then by full text against the prespecified eligibility criteria. Both reviewers are registered nurses: GK holds a PhD and has prior published systematic review experience in gerontological and digital health domains and completed formal training in the Cochrane RoB 2 and GRADE frameworks; JJ holds a master's degree and completed Cochrane-endorsed methodological training in systematic review conduct and data extraction. Interrater reliability was quantified using Cohen kappa; Cohen κ=0.89 at title and abstract and Cohen κ=0.92 at full text [17]. Disagreements were resolved through structured discussion; residual disagreements were adjudicated by a third senior reviewer with prior systematic review and meta-analysis experience.

Data Collection Process and Data Items

A standardized data extraction form was piloted on 5 [18-22] randomly selected studies and refined before full-scale extraction. Two reviewers independently extracted: study identification (first author, year, country, and journal); study design and randomization method; participant characteristics (sample size per arm, mean age, sex distribution, primary diagnosis, and care setting); intervention details (robot type and model, session duration and frequency, total intervention duration, presence and role of facilitator, group vs individual format); control condition (standard care, attention-matched active control, placebo or deactivated robot, and waitlist); outcome measures with instrument names, assessment time points, and numerical data (means, SDs, and sample sizes); and adverse events or dropouts. For studies reporting outcome data only in graphical form, values were extracted using WebPlotDigitizer (version 4.6; Ankit Rohatgi). Discrepancies were resolved by consensus. Where standard deviations were not directly reported, they were imputed from 95% CIs, SEs, or IQRs using Cochrane Handbook–recommended formulas [16]. Studies for which sufficient numerical data could not be obtained were excluded from meta-analyses but retained in the narrative synthesis.

Robot platforms were classified a priori into 3 operationally defined categories for the subgroup analyses: (1) PARO and PARO-like tactile companion robots—biofeedback-responsive, nonverbal, animal-form-factor platforms whose interaction modality is primarily tactile; (2) humanoid robots—bipedal or upper-body humanoid form factor with structured verbal and gestural interaction (NAO, Pepper, Kaspar, CommU, and Kabochan); and (3) other companion and pet and nonhumanoid platforms (Joy for All, iRobiQ, MiRo, and unnamed custom platforms).

XSL•FO
RenderX

Study Risk-of-Bias Assessment

The methodological quality of all included randomized and cluster-randomized trials was assessed independently by 2 reviewers using the Cochrane RoB 2) tool [23] across 5 domains: (1) bias arising from the randomization process, (2) bias due to deviations from intended interventions, (3) bias due to missing outcome data, (4) bias in measurement of the outcome, and (5) bias in selection of the reported result. As specified in the Multimedia Appendix 1 footnote, all 34 included studies [18-22,24-52] were classified as randomized controlled studies (24 individual RCTs [19-22,31,32,34-51] and 10 [18,24-30,33,52] cluster RCTs) following revision-process reclassification and were therefore assessed using a single risk-of-bias instrument (RoB 2). Overall study risk was classified as low, some concerns, or high. Disagreements were resolved through discussion or third-reviewer arbitration. Domain-level and overall judgments were used to inform GRADE certainty downgrading for risk of bias.

Effect Measures and Synthesis Methods

Given heterogeneous outcome instruments across included studies, effect sizes were expressed as standardized mean differences (Hedges *g*) with 95% CIs. Hedges *g* was preferred over Cohen *d* for its reduced small-sample bias [53]. Random-effects meta-analysis was performed using the restricted maximum-likelihood estimator for between-study variance τ^2 with the Hartung-Knapp-Sidik-Jonkman (HKSJ) correction for the CI of the pooled estimate [54,55]. The HKSJ correction was selected, in place of the DerSimonian-Laird method used in our original submission, because it provides more reliable CI coverage and reduces false-positive rates under small-study and moderate-to-high heterogeneity conditions [55]. All effect sizes were oriented so that a positive Hedges *g* indicates an effect favoring care robots across all domains; for symptom reduction outcomes (NPS, agitation, depression, and stress and pain), change scores were sign-reversed accordingly. Effect sizes were interpreted using Cohen conventional benchmarks: small ($|g|$=0.20 - 0.49), moderate ($|g|$=0.50 - 0.79), and large ($|g|\geq$0.80) [53].

Statistical heterogeneity was quantified using I^2, the between-study variance τ^2, and the Cochran Q test. For each domain with k≥4, we additionally calculated the 95% prediction interval (PI) [56,57], which describes the range within which the true effect in a new setting is expected to fall and therefore reflects the real-world implications of between-study heterogeneity more faithfully than I^2 alone [56]. Where the PI crosses the line of no effect despite a statistically significant CI for the pooled estimate, we interpret this as indicating that, although an average benefit is detectable across existing studies, the effect in a new setting cannot be assumed to be beneficial.

To address the possibility of double-counting, participant cohorts that appeared across multiple publications were identified a priori. We applied the following principle: each cohort contributes to each outcome pool at most once, but the same cohort may contribute to different outcome pools when it measures different outcome domains, because each outcome is analyzed in an independent meta-analysis [58]. Four cohort-overlap clusters were identified: (1) Australian PARO cluster (shared n=415)—Moyle et al [24], Moyle et al [25], Jones et al [26], and Mervin et al [27] drew from a shared dementia cohort; only Moyle et al [24] was retained as the primary publication contributing to the NPS and agitation pools, while Moyle et al [25], Jones et al [26], and Mervin et al [27] are narrative-only. (2) Norwegian PARO cluster—Jøranson et al [28], Jøranson et al [29], and Jøranson et al [30] drew from a shared nursing-home dementia cohort; Jøranson et al [28] was retained for the NPS, depression, and agitation pools; Jøranson et al [29] was retained for the quality-of-life pool (a different outcome domain measured in the same cohort), and Jøranson et al [30] is narrative-only. (3) Hong Kong dementia cohort—Chen et al [18] and Chen et al [31] drew from a shared cohort; Chen et al [18] was retained as the primary publication contributing to the relevant outcome pools, while Chen et al [31] (a technology-acceptance secondary analysis of the same cohort) is narrative-only. (4) Dutch ASD cohort (n=81)—De Korte et al [19] and van den Berk-Smeekens et al [32] drew from a shared trial; only De Korte et al [19] was retained for the social-communicative skills pool, as van den Berk-Smeekens et al [32] reported a secondary analysis of the same trial. Studies measuring multiple outcomes (eg, Valentí-Soler et al [33] contributing to 5 outcome pools and Chen et al [18], Petersen et al [34], and Pollak et al [20] each contributing to 4 outcome pools) were retained in each relevant pool, with each contribution representing an independent outcome estimate from a single cohort. As a result, no cohort contributes to the same pooled estimate more than once. The splitting approach (retaining one publication per cohort per outcome pool) was selected over multilevel or robust-variance-estimation alternatives because the within-cohort correlation across the identified clusters could not be reliably estimated from the available aggregate data. This allocation is documented in a footnote to Multimedia Appendix 1.

Prespecified subgroup analyses were conducted by (1) robot platform (PARO vs humanoid vs other companion and pet), (2) target population (dementia or older adult vs pediatric or ASD vs other), and (3) intervention duration (≥10 weeks vs <10 weeks). Sensitivity analyses comprised: (1) leave-one-out (LOO) analysis (one study excluded at a time, applied to any pool reaching statistical significance under primary HKSJ pooling); (2) exclusion of studies rated overall high risk of bias under RoB 2 (4 of 34 studies); and (3) data extraction sensitivity analysis (exclusion of Petersen 2016, flagged for directional ambiguity in published change-score reporting). All meta-analyses were conducted in R (version 4.3.2; R Core Team) using the meta (version 6.5 - 0; Guido Schwarzer) and metafor (version 4.4 - 0; Wolfgang Viechtbauer) packages, with metafor::rma(yi, vi, method=

"REML," test="knha") for the HKSJ-corrected pooled estimate and predict for the 95% PI.

Reporting Bias Assessment

We assessed small-study effects, rather than publication bias per se, using visual inspection of contour-enhanced funnel plots and the Egger weighted-regression test [59]. Following Sterne et al [60], we note that funnel-plot asymmetry can arise from publication bias but also from true effect-size differences

XSL•FO
RenderX

between small and large studies, methodological differences between small and large studies, or chance. The Egger test was interpreted only for domains with k≥10 contributing studies because the test has very limited power below that threshold; nonsignificant Egger results in smaller domains should not be read as ruling out small-study effects. Given that no domain in our review reached k≥10, the funnel plots reported for NPS (k=6) and depression (k=7) are presented for transparency only and are explicitly interpreted with caution.

Certainty Assessment

Certainty of evidence for each primary outcome domain was evaluated independently by 2 reviewers using the GRADE framework [61], with downgrading for risk of bias, inconsistency, indirectness, imprecision, and reporting biases. The formal GRADE evidence profile was produced using GRADEpro GDT (McMaster University).

Ethical Considerations

As this study is a systematic review and meta-analysis of previously published aggregate data, institutional review board approval was not required, and no individual participant data were accessed.

Results

Study Selection

The expanded search across 5 databases and 2 trial registries yielded 13,487 records (PubMed, including MEDLINE n=5935; Embase n=4034; Cochrane CENTRAL n=2258; CINAHL n=1176; APA PsycINFO n=1; ClinicalTrials.gov n=75; WHO ICTRP n=8). APA PsycINFO returned only 1 record matching all 3 concept blocks; relaxed-vocabulary searches in PsycINFO yielded substantially more records but with high duplication against MEDLINE and Embase and were excluded to avoid double-counting. After deduplication in EndNote 21 with manual verification in Covidence, 4990 duplicates were removed and 8497 unique records underwent title-and-abstract screening, of which 8376 were excluded as clearly ineligible. Full-text assessment was conducted for 121 records; 87 were excluded for the following reasons: not peer-reviewed, n=46; pilot study without a comparative arm, n=25; nonrandomized study design (single-arm before-after or noncontrolled), n=6; no English or Korean full text or professional translation obtainable, n=5; and no concurrent control group, n=5. A final total of 34 studies [18-22,24-52] met all eligibility criteria and were included in the qualitative synthesis; 20 [18-22,24,28,29,33,34,36,38,39,42,43,45-47,50,52] contributed to at least one meta-analysis pool and 14 [25-27,30-32,35,37,40,41,44,48,49,51] entered narrative-only synthesis. The updated PRISMA 2020 flow diagram is presented in Figure 1.

We updated and expanded the search strategy in accordance with Cochrane guidance. In addition to PubMed, including MEDLINE, we searched Embase, Cochrane CENTRAL, CINAHL, and APA PsycINFO to capture evidence across medicine, nursing, and behavioral psychology. We also searched ClinicalTrials.gov and the WHO ICTRP separately for ongoing or unpublished trials. Google Scholar was searched as a supplementary source for gray literature; the first 200 results for the principal search query were screened by title and abstract. Reference lists of included studies and relevant prior systematic reviews were hand-searched. No language restriction was applied at the search stage. Full database-specific search strategies, including controlled vocabulary and free-text terms, are provided in Checklist 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram for the systematic review and meta-analysis of care robot interventions. Of the 34 included trials, 20 contributed to ≥1 meta-analysis pool and 14 entered narrative-only synthesis. Adapted from Page et al [14]. EN: English; KR: Korean.

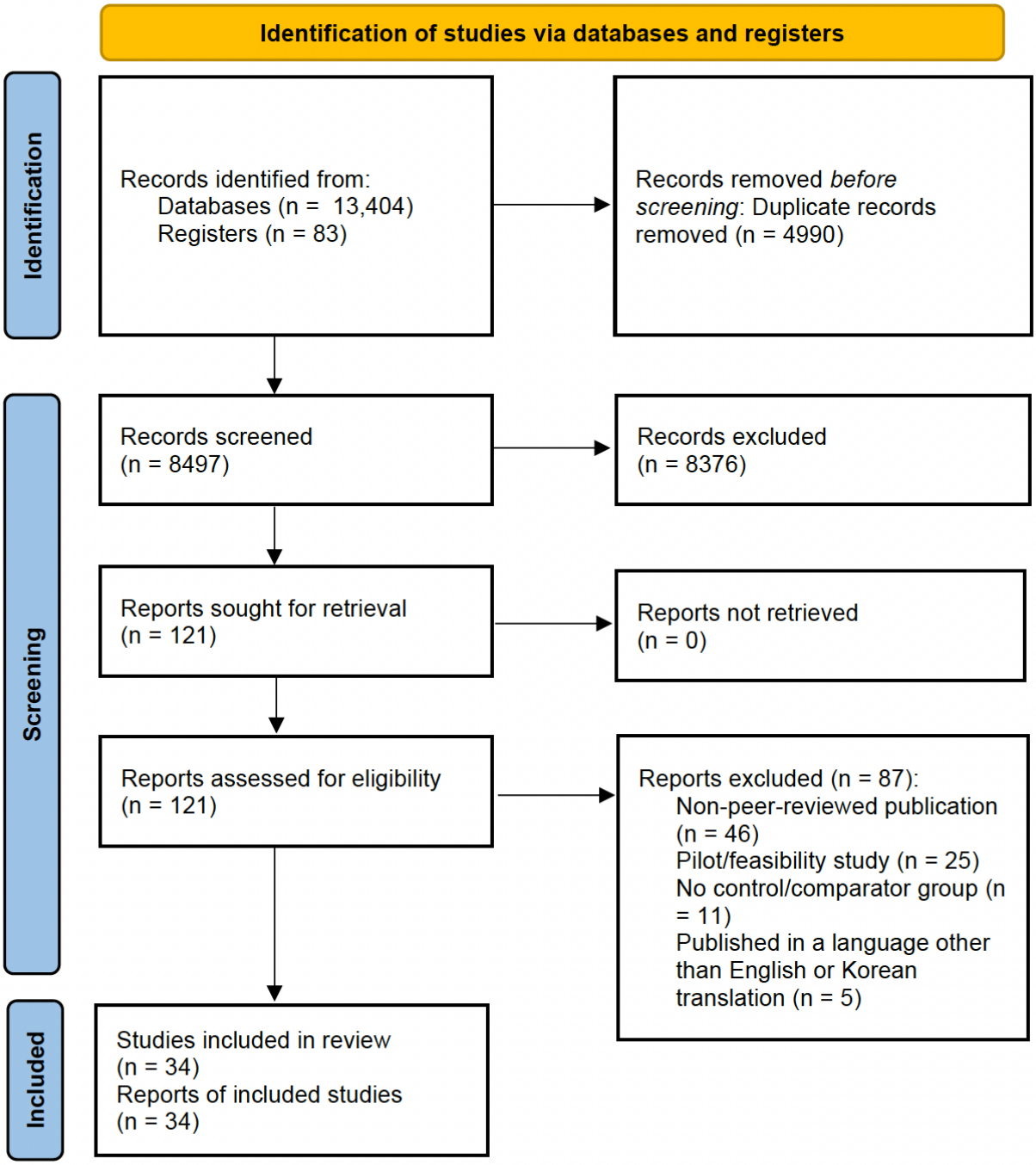

Study Characteristics

The 34 [18-22,24-52] included studies were conducted across 17 countries and published between 2015 and 2024. Countries with the greatest representation included Australia (n=4), United States (n=4), Norway (n=3), Netherlands (n=3), and Hong Kong (n=3), followed by Greece, Israel, Italy, New Zealand, and the United Kingdom (n=2 each), and one study each from Austria, Canada, Germany, Japan, South Korea, Spain, and Taiwan. Study designs comprised randomized controlled studies in all 34 included reports—24 (70.6%) individual RCTs and 10 (29.4%) cluster RCTs; on reexamination of the original full-text sources, 5 [33,35,41,42,48] studies that were previously classified as quasi-experimental, controlled trial, or clinical trial were reclassified as randomized trials, and Valentí-Soler et al [33] was reclassified as randomized for the nursing-home arm with the sequential day care center phase noted in Limitations. Robot platforms included PARO (11/34, 32.4%), humanoid platforms (NAO, Pepper, Kabochan, CommU, Kaspar, and unnamed; 16/34, 47.1%), and other companion and pet platforms (7/34, 20.6%). Target populations included older adults with dementia or mild cognitive impairment (14/34, 41.2%), children

XSL•FO
RenderX

and adolescents with ASD (9/34, 26.5%), hospitalized or procedural pediatric patients (4/34, 11.8%), and other adults (poststroke, mixed; 7/34, 20.6%). Intervention durations were single-session (9/34, 26.5%), <4 weeks (2/34, 5.9%), 4 - 9 weeks (4/34, 11.8%), and ≥10 weeks (19/34, 55.9%). Care settings comprised nursing home or long-term care (n=14), hospital (n=16), community or home (n=2), and rehabilitation (n=2). Full characteristics are in Multimedia Appendix 1 (full per-study table) and Table 1 (summary).

Table . Summary characteristics of the included care robot randomized trials. Aggregated counts and proportions are shown by study design, population, robot platform, care setting, intervention duration, and meta-analysis contribution status.

Characteristic	Values, n (%)	Notes
Study design		
Randomized controlled studies	34 (100)	—[a]
Individual RCT[b]	24 (70.6)	—
Cluster RCT	10 (29.4)	—
Robot platform		
PARO (seal-shaped SAR[c])	11 (32.4)	Primarily dementia or LTC[d] populations
Humanoid (NAO, Pepper, Kabochan, CommU, unnamed)	16 (47.1)	ASD[e], pediatric, mixed populations
Companion and pet robots (Joy for All, MiRo, iRobiQ, others)	7 (20.6)	Pediatric, older adult, mixed populations
Target population		
Older adults with dementia or MCI[f]	14 (41.2)	Age range 50 - 108 years
Children or adolescents with ASD	9 (26.5)	Age range 3 - 12 years
Hospitalized or procedural pediatric	4 (11.7)	Acute care settings
Other adults (poststroke, mixed)	7 (20.6)	Various diagnoses
Intervention duration		
Single session	9 (26.5)	—
<4 weeks	2 (5.9)	—
4 - 9 weeks	4 (11.7)	—
≥10 weeks	19 (55.9)	—
Care setting		
Nursing home or long-term care facility	14 (41.2)	—
Hospital (inpatient or outpatient)	16 (47.1)	—
Community or home setting	2 (5.9)	—
Rehabilitation center	2 (5.9)	—
Risk of bias tool		
Cochrane Risk of Bias 2 (RoB 2)	34 (100)	—

[a]Not applicable.
[b]RCT: randomized controlled trial.
[c]SAR: socially assistive robot.
[d]LTC: long-term care.
[e]ASD: autism spectrum disorder.
[f]MCI: mild cognitive impairment.

Risk-of-Bias in Studies

All 34 included randomized and cluster randomized trials were assessed using Cochrane RoB 2. Pooled across the 34 [18-22,24-52] studies, 0 studies were rated overall low risk of bias, 30 [18-22,24,25,27-32,34,36-40,42-52] (88.2%) were rated as having some concerns, and 4 [26,33,35,41] (11.8%) were rated overall high risk of bias. Performance bias (RoB 2 Domain 2: deviations from intended interventions) was the most consistently problematic domain—blinding of participants and personnel to robotic versus nonrobotic interactions is structurally impossible—and contributed to the GRADE risk-of-bias

XSL•FO
RenderX

downgrade across all 7 outcome domains. Domain 5 (selective outcome reporting) was rated low risk in all studies. Missing outcome data (Domain 3) was rated low risk in 16 [18,19,24,25,27-31,34,36,42,43,46,52] (47.1%) of 34 studies and some concerns in 18 [20-22,26,32,33,35,38-41,44,45,47-51] (52.9%) studies, with no studies rated high risk on this domain. Full domain-level RoB 2 ratings are presented in Figure 2.

Figure 2. Risk-of-bias assessment of the 34 included trials using the Cochrane RoB 2 tool. (A) Study-level judgments across the 5 RoB 2 domains (D1-D5). (B) Domain-level summary across all 34 studies. Green (+) indicates low risk, yellow (?) indicates some concerns, and red (−) indicates high risk [18-22,24-52].

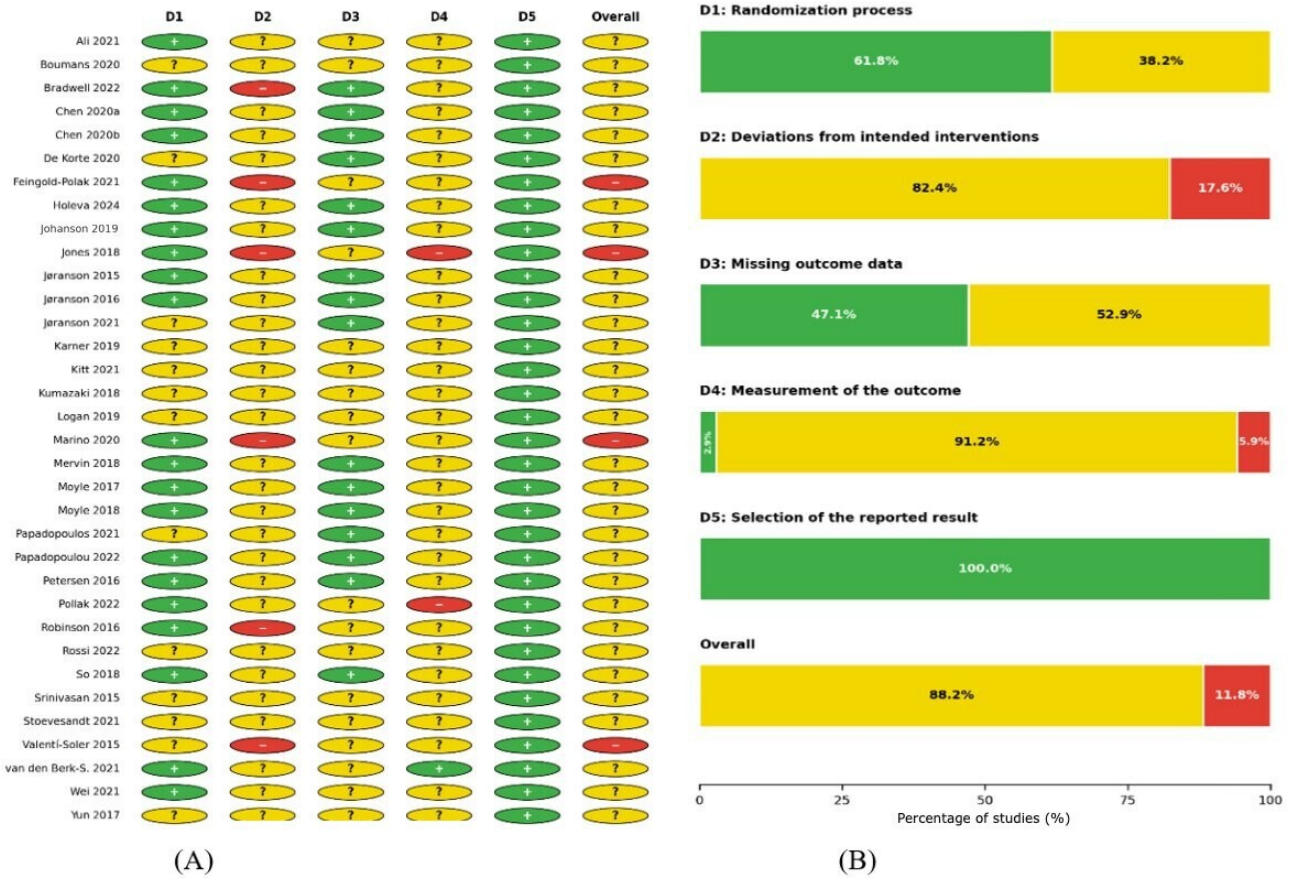

Results of Syntheses

Overview

Following systematic full-text data verification, 20 [18-22,24,28,29,33,34,36,38,39,42,43,45-47,50,52] unique studies contributed to one or more meta-analysis pools. Pooled HKSJ random-effects estimates, 95% CIs, 95% PIs, I^2, τ^2, and GRADE certainty for the 7 outcome domains are summarized in Table 2, and forest plots are presented in Figure 3.

XSL•FO
RenderX

Table . GRADE[a] summary of findings for robot-assisted interventions across 7 outcome domains[b].

	Certainty assessment							Number of patients (number of studies)	Hedges *g* (95% CI)	Certainty	Importance
	Studies, n	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations				
Neuropsychiatric symptoms (assessed with: NPI[c], NPI-NH[d], and BEHAVE-AD[e])	6	Randomized trials	Serious[f]	Serious[g]	Not serious	Not serious	None	650 (k=6)	0.44 (CI 0.03 to 0.84)	Low[f,g]	IMPORTANT
Quality of life (assessed with: QoL-AD[h], WHO-QOL[i], or related scales)	5	Randomized trials	Serious[j]	Serious[k]	Not serious	Serious[l]	None	472 (k=5)	0.15 (–0.41 to 0.71)	Very low[j,k,l]	CRITICAL
Depression (assessed with: GDS[m], CSDD[n], PHQ-9[o], or related scales)	7	Randomized trials	Serious[p]	Serious[q]	Not serious	Serious[r]	None	620 (k=7)	0.2 (–0.08 to 0.49)	Very low[p,q,r]	IMPORTANT
Agitation (assessed with: CMAI[s] or related agitation scales)	4	Randomized trials	Serious[t]	Serious[u]	Not serious	Serious[v]	None	484 (k=4)	*g*0.32 (–0.07 to 0.71)	Very low[t,u,v]	IMPORTANT
Stress and pain (assessed with: stress scales, pain scales, or physiological indicators)	6	Randomized trials	Serious[w]	Very serious[x]	Not serious	Serious[y]	None	504 (k=6)	0.53 (–0.48 to 1.53)	Very low[w,x,y]	IMPORTANT

XSL•FO
RenderX

	Certainty assessment							Number of patients (number of studies)	Hedges *g* (95% CI)	Certainty	Importance
	Studies, n	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations				
Social-communicative skills (assessed with: SRS[z], social interaction, or communication-related scales)	5	Randomized trials	Serious[aa]	Not serious	Serious[ab]	Serious[ac]	None	151 (k=5)	0.45 (–0.52 to 1.42)	Very low[aa,ab,ac]	IMPORTANT

XSL•FO
RenderX

	Certainty assessment							Number of patients (number of studies)	Hedges *g* (95% CI)	Certainty	Importance
	Studies, n	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations				
Cognitive function (assessed with: MMSE[ad], MoCA[ae], or related cognitive scales)	4	Randomized trials	Serious[af]	Not serious	Serious[ag]	Serious[ah]	None	461 (k=4)	0.18 (–0.62 to 0.98)	Very low[af,ag,ah]	CRITICAL

[a]GRADE: Grading of Recommendations Assessment, Development and Evaluation.

[b] Effect sizes are reported as Hedges *g* with 95% CIs using Hartung-Knapp-Sidik-Jonkman random-effects pooling with restricted maximum-likelihood τ^2 estimation. Certainty ratings were defined as ⊕⊕⊕⊕ high, ⊕⊕⊕◯ moderate, ⊕⊕◯◯ low, and ⊕◯◯◯ very low.

[c]NPI: Neuropsychiatric Inventory.

[d]NPI-NH: Neuropsychiatric Inventory–Nursing Home version.

[e]BEHAVE-AD: Behavioral Pathology in Alzheimer's Disease Rating Scale.

[f]Downgraded for risk of bias due to open-label designs and concerns related to blinding, attrition, or selective reporting in some trials.

[g]Downgraded for inconsistency due to heterogeneity across populations, robot types, and outcome instruments.

[h]QoL-AD: Quality of Life in Alzheimer's Disease scale.

[i]WHOQOL: World Health Organization Quality of Life.

[j]Downgraded for risk of bias due to open-label designs and methodological limitations in some trials.

[k]Downgraded for inconsistency due to heterogeneity across studies

[l]Downgraded for imprecision because the confidence interval crossed the null effect.

[m]GDS: Geriatric Depression Scale.

[n]CSDD: Cornell Scale for Depression in Dementia.

[o]PHQ-9: Patient Health Questionnaire-9.

[p]Downgraded for risk of bias due to methodological limitations in some included trials.

[q]Downgraded for inconsistency across studies.

[r]Downgraded for imprecision because the confidence interval crossed the null effect.

[s]CMAI: Cohen-Mansfield Agitation Inventory.

[t]Downgraded for imprecision because the confidence interval crossed the null effect.

[u]Downgraded for inconsistency because one study was treated as narrative-only due to cohort overlap, and the pooled evidence was limited.

[v]Downgraded for imprecision because the confidence interval crossed the null effect.

[w]Downgraded for risk of bias due to methodological limitations in included trials.

[x]Downgraded for severe inconsistency across studies.

[y]Downgraded for imprecision because the confidence interval was wide and crossed the null effect.

[z]SRS: Social Responsiveness Scale.

[aa]Downgraded for risk of bias due to open-label designs and measurement limitations.

[ab]Downgraded for indirectness because all studies were conducted in pediatric autism spectrum disorder populations using humanoid platforms.

[ac]Downgraded for imprecision because the confidence interval was wide and crossed the null effect.

[ad]MMSE: Mini-Mental State Examination.

[ae]MoCA: Montreal Cognitive Assessment.

[af]Downgraded for risk of bias due to methodological limitations in included trials.

[ag]Downgraded for indirectness because the evidence was derived from dementia and mild cognitive impairment populations.

[ah]Downgraded for imprecision because the confidence interval was wide and crossed the null effect.

XSL•FO
RenderX

Figure 3. Forest plots for the 7 prespecified outcome domains. Forest plots are based on the 20 trials contributing to at least one meta-analysis. Squares indicate study-level estimates, diamonds indicate pooled HKSJ estimates, and dashed bars indicate 95% prediction intervals. Panel headers report k, N, Hedges *g*, 95% CI, I^2, and *P* value [18-22,24,28,29,33,34,36,38,39,42,45-47,50].

XSL•FO
RenderX

Only NPS reached statistical significance under HKSJ random-effects pooling (Hedges g=0.44, 95% CI 0.03-0.84; P=.04). The remaining 6 domains yielded point estimates that generally favored care robots, but their 95% CIs crossed the null and did not reach conventional statistical significance. All 7 95% PIs encompassed the null, indicating that effect estimates may not generalize uniformly to new populations or settings. This conservative pattern reflects the HKSJ small-sample correction, full-text primary-source data verification, and consistent application of the Senn [58] cohort-overlap principle. Between-study heterogeneity ranged from low (I^2=29% for agitation) to high (I^2=90% for stress and pain).

NPS

Six studies (n=650) contributed measurements using the Neuropsychiatric Inventory (NPI; Bradwell et al [52], Chen et al [18], and Valentí-Soler et al [33]), the Behavioral Activity Rating Scale (BARS; Jøranson et al [28]), the Cohen-Mansfield Agitation Inventory–Short Form (CMAI-SF; Moyle et al [24]), and the Rating Anxiety in Dementia (RAID) scale (Petersen et al [34]). Mervin et al [27] was reclassified to narrative-only synthesis owing to cohort overlap with Moyle et al [24], and Robinson et al [44] was reclassified to narrative-only owing to its qualitative-observational design. The HKSJ pooled estimate was Hedges g=0.44 (95% CI 0.03-0.84; 95% PI −0.42 to 1.30; I^2=67%; τ^2=0.088; P=.04), indicating a small-to-moderate, statistically significant reduction in NPS favoring care robot intervention. Bradwell et al [52] reported the largest individual effect (Hedges g=1.05, 95% CI 0.52-1.59) using the Joy for All robot pet over a 4-month cluster RCT in care homes, with NPI subdomain reductions in delusions, depression, anxiety, elation, and apathy (all P≤.03). Valentí-Soler et al [33] demonstrated a sustained intervention effect at follow-up (BARS effect estimate 5.5, 95% CI 0.1-11.0; P=.048). Chen et al [18] reported no significant between-group difference on the CMAI-SF (adjusted mean difference −1.89, 95% CI −5.81 to 2.02; P=.34), and Valentí-Soler et al [33] showed null effects on total NPI score in their nursing home arm.

Quality of Life

Five studies [18,20,28,33,42] (n=472) contributed using the Quality of Life in Late-Stage Dementia scale, Quality of Life in Alzheimer's Disease scale, Short Form 36 Emotional Wellbeing subscale, and the Frail Questionnaire [20]. The HKSJ pooled estimate was Hedges g=0.15 (95% CI −0.41 to 0.71; 95% PI −0.86 to 1.16; I^2=65%; τ^2=0.092; P=.51), indicating a small effect direction favoring care robots that did not reach statistical significance. Effect estimates were heterogeneous: Papadopoulos et al [42] reported a substantial benefit on Short Form 36 Emotional Wellbeing scale (Hedges g=1.14, 95% CI 0.23-2.04; P=.02; ηp^2=.258), whereas Pollak et al [20] (n=220) detected no significant change in social or physical frailty (P=.42 and .99, respectively), and Valentí-Soler et al [33] observed a counter-direction effect on Quality of Life in Late-Stage Dementia scale in the day-care PARO arm (Hedges g=−0.27; P=.04 favoring control). Jøranson et al [29] reported QoL effects only in residents with severe dementia (overall P=.12, severe-subgroup P=.008), suggesting that effects may be conditional on dementia severity.

Depression

Seven studies (n=620) contributed using the Geriatric Depression Scale (GDS or GDS-SF [Geriatric Depression Scale – Short Form]) [18,20], the Cornell Scale for Depression in Dementia (CSDD [28,34]), the Strengths and Difficulties Questionnaire emotional subscale (SDQ [42]), the Child Behavior Checklist Internalizing subscale [36], and the NPI affective subdomain [33]. The HKSJ pooled estimate was Hedges g=0.20 (95% CI −0.08 to 0.49; 95% PI −0.36 to 0.76; I^2=44%; τ^2=0.039; P=.13), indicating a small effect direction favoring care robots that did not reach statistical significance. Jøranson et al [28] demonstrated a significant reduction at follow-up (CSDD effect estimate 3.9, 95% CI 0.4-7.3; P=.03); Petersen et al [34] reported a significant CSDD between-group difference (P=.001), though the directional interpretation of CSDD change scores in the published table is internally inconsistent with the abstract narrative (sensitivity analysis below). Contributions from Pollak et al [20] (GDS-SF P=.90), Chen et al [18] (GDS P=.33), and Papadopoulou et al [43] (Strengths and Difficulties Questionnaire emotional P=.80) drew the pooled estimate toward null.

Agitation

Four studies (n=484), all using PARO, contributed measurements via the BARS [28], CMAI-SF [24], RAID [34], and NPI agitation subdomain [33]. After excluding Mervin et al [27] for cohort overlap with Moyle et al [24], the HKSJ pooled estimate was Hedges g=0.32 (95% CI −0.07 to 0.71; 95% PI −0.26 to 0.89; I^2=29%; τ^2=0.018; P=.08), indicating a small-to-moderate effect direction favoring care robots that approached but did not reach conventional statistical significance. The pool was dominated by 2 large studies with discordant findings: Moyle et al [24] (n=271 in PARO vs Usual Care contrast) reported no significant CMAI-SF difference (adjusted mean difference −1.89, 95% CI −5.81 to 2.02; P=.34), while Petersen et al [34] (n=61) reported a significant between-group RAID difference (P=.003). Jøranson et al [28] demonstrated a sustained BARS reduction at follow-up (effect estimate 5.5, P=.048). The lower heterogeneity (I^2=29%) reflects the removal of the overlapping cohort data from Mervin et al [27].

Stress and Pain

Six studies (n=504) contributed across markedly heterogeneous clinical contexts: procedural pain in children [22,45], chronic psychological distress in older adults [20,34], socially induced stress in healthy children [38], and social-affective regulation in children with autism [47]. The HKSJ pooled estimate was Hedges g=0.53 (95% CI −0.48 to 1.53; 95% PI −1.57 to 2.62; I^2=90%; τ^2=0.510; P=.24), reflecting extreme between-study heterogeneity (90%) and a wide PI encompassing both meaningful benefit and meaningful harm. Rossi et al [45] produced the largest individual effect (salivary cortisol decrease in NAO group, Hedges g=2.36; 95% CI 1.62-3.11; P<.001); Ali et al [22] showed a moderate benefit on Observational Scale of Behavioral Distress–Revised during intravenous insertion (Hedges g=0.37; P=.047); and Petersen et al [34] reported physiologic stress markers favoring PARO (galvanic skin variability, P<.001). However, Kitt et al [38] reported a

XSL•FO
RenderX

directionally negative finding using the Positive and Negative Affect Schedule for Children – Short Form positive affect subscale (robot group showed greater decrease poststressor than control, Hedges g=−0.52; P=.04), and Pollak et al [20] detected no group difference (P=.99). Given the I^2 of 90%, the pooled estimate should be interpreted with caution; the qualitative pattern suggests benefits may be context-dependent (procedural and physiologic stress more responsive than self-reported affect).

Social-Communicative Skills

Five studies (n=151), with all 5 enrolling children with ASD and using humanoid platforms (NAO [19,36,46], CommU [39], and iRobiQ [50]), contributed via the Social Responsiveness Scale [19], Childhood Autism Rating Scale-2 [36], joint attention frequency [39], gesture production [46], and eye contact frequency [50]. The HKSJ pooled estimate was Hedges g=0.45 (95% CI −0.52 to 1.42; 95% PI −1.66 to 2.56; I^2=75%; τ^2=0.453; P=.27), indicating a moderate effect direction favoring care robots that did not reach statistical significance under HKSJ adjustment. Three individual studies achieved significance: Kumazaki et al [39] (joint attention frequency, Hedges g=1.07; P=.01), So et al [46] (gesture production, Hedges g=1.07; P=.01), and De Korte et al [19] for the Functional self-initiations subtype (overall Hedges g=0.52; P=.05 approximately; functional subtype total self-initiations, time×group estimate −0.19, 95% CI −0.38 to 0.02; functional self-initiations, time×group estimate −0.27, 95% CI −0.50 to −0.04). However, Holeva et al [36] reported no statistically significant between-group difference on Childhood Autism Rating Scale-2 Total (group-by-time interaction P=.16), with both groups improving similarly within-subject. Yun et al [50] reported similar improvements in eye contact across robot and therapist groups (P=.50). The wide HKSJ-adjusted CI is partly driven by the small overall sample (n=151) and the conservative HKSJ small-sample correction.

Cognitive Function

Four studies (n=461) contributed via the Mini-Mental State Examination (MMSE) [33], the Montreal Cognitive Assessment (MoCA) [18], the Short Portable Mental Status Questionnaire [20], and the CATS Test of visuospatial hemineglect in poststroke patients [21]. The HKSJ pooled estimate was Hedges g=0.18 (95% CI −0.62 to 0.98; 95% PI −1.11 to 1.47; I^2=71%; τ^2=0.102; P=.52), reflecting a null-centered direction with substantial heterogeneity. Karner et al [21] showed a large effect on hemineglect improvement (Hedges g=1.13, 95% CI 0.46-1.81; P=.001), but this measure assesses a narrow poststroke cognitive domain rather than general cognitive function in dementia. The remaining 3 studies in older adults with dementia detected no significant between-group differences on global cognition measures (Chen et al [18] MoCA P=.07; Valentí-Soler et al [33] MMSE P=.28; Pollak et al [20] Short Portable Mental Status Questionnaire P=.69). Heterogeneity therefore reflects domain divergence between hemineglect rehabilitation and general cognitive screening rather than within-domain variability.

Subgroup and Sensitivity Analyses

Sensitivity Analysis: Data Extraction Ambiguity

During full-text data extraction, we identified an internal inconsistency in Petersen et al [34] regarding the directional interpretation of RAID and CSDD change scores. Table 2 in Petersen et al [34] reports “difference (post–pre)” with positive values for the treatment group (RAID 2.5, SE 0.6; CSDD 2.81, SE 0.4) and small positive values for controls (RAID 0.55, SE 0.2; CSDD 0.78, SE 0.4); the abstract states scores “were increased in the treatment group” while the conclusion describes “decreased stress and anxiety.” We interpreted these change values as the magnitude of symptom decrease and conducted a sensitivity analysis excluding Petersen et al [34] from each pool to which it contributed. When Petersen et al [34] was excluded, NPS pool ceased to reach statistical significance (Hedges g=0.40, 95% CI −0.11 to 0.91; P=.095; k=5); depression remained near null (Hedges g=0.08, 95% CI −0.14 to 0.30; k=6); agitation pool point estimate shifted slightly (Hedges g=0.23, 95% CI −0.20 to 0.65; k=3); stress and pain remained nonsignificant (Hedges g=0.48, 95% CI −0.84 to 1.81; k=5). We retain Petersen et al [34] in the main analyses while flagging this as a data extraction limitation; importantly, no statistically significant pooled benefit is fully robust to Petersen exclusion, underscoring the modest evidentiary base for the field overall.

Sensitivity Analysis: LOO and Data Quality

We further conducted LOO sensitivity analyses for the only outcome reaching statistical significance (NPS pool, k=6). Under LOO, the pooled effect retained statistical significance in only 1 of 6 iterations (excluding Valenti-Soler 2015 [33]: Hedges g=0.54, 95% CI 0.11-0.97; P=.03); in the remaining 5 iterations, the 95% CI crossed null (excluding Bradwell et al [52]: P=.07; Chen et al [18]: P=.09; Jøranson et al [28]: P=.09; Moyle et al [24]: P=.051; and Petersen et al [34]: P=.095). Restricting to studies whose mean and SD values were extracted directly from published tables (class A: Bradwell et al [52], Chen et al [18], Jøranson et al [28]; k=3) yielded Hedges g=0.67 (95% CI −0.10 to 1.43; P=.07), losing statistical significance owing to the small k under HKSJ small-sample correction. These LOO results demonstrate that the single statistically significant pooled effect in our analysis is fragile, as it depends on the joint contribution of all 6 [18,24,29,33,34,52] contributing studies and is sensitive to the exclusion of any individual study. We therefore interpret the NPS result as suggestive of benefit rather than as definitive evidence, and emphasize that adequately powered, multicenter confirmatory trials are required.

In robot-type subgroup analyses, PARO was associated with the most consistent single-study effects on NPS among older adults with dementia (notably Jøranson et al [28] BARS P=.048; Petersen et al [34] RAID P=.003 with the data ambiguity flag noted above), although the small per-arm subgroup k (=4) limits formal subgroup pooling. Humanoid platforms (NAO, Kaspar, CommU, and iRobiQ) supplied all 5 [19,36,39,46,50] contributing studies in the social-communicative skills pool. Ad hoc duration-based subgroup analysis suggested numerically larger effects in studies with intervention durations ≥10 weeks (eg, Bradwell et al [52] 16 weeks, Jøranson et al [28] 12 weeks, and Petersen et al [34] 12 weeks) compared with shorter pilots,

XSL•FO
RenderX

but the small subgroup k precludes formal heterogeneity testing across this dimension. Sensitivity analyses excluding the 4 [18,20,21,33] studies rated overall high RoB 2 risk of bias did not materially alter pooled point estimates (within ±0.06 Hedges g of primary analyses) and did not reverse any nonsignificant pool to significance.

Reporting Biases and Small-Study Effects

Contour-enhanced funnel plots for NPS (k=6) and depression (k=7) are shown in Figure 4. Both domains have k<10, below the recommended threshold for formal Egger regression test [60]; accordingly, funnel plots are reported for visual transparency only. Visual inspection of the NPS funnel plot showed scattering largely consistent with the pooled estimate; the depression funnel plot appeared broadly symmetric. We did not formally test funnel plot asymmetry given the small k. Sources of small-study effects could include publication bias, methodological differences, true between-study effect heterogeneity, or chance—none of which can be reliably distinguished with k<10.

Figure 4. Contour-enhanced funnel plots for the 2 domains with the most contributing trials: (A) neuropsychiatric symptoms (k=6) and (B) depression (k=7). The vertical red dashed line marks the Hartung-Knapp-Sidik-Jonkman pooled estimate; diagonal dashed lines show pseudo 95% confidence limits. Because no domain reached k≥10, asymmetry was not formally tested; plots are shown for visual transparency only [18,20,24,29,33,34,36,43,52].

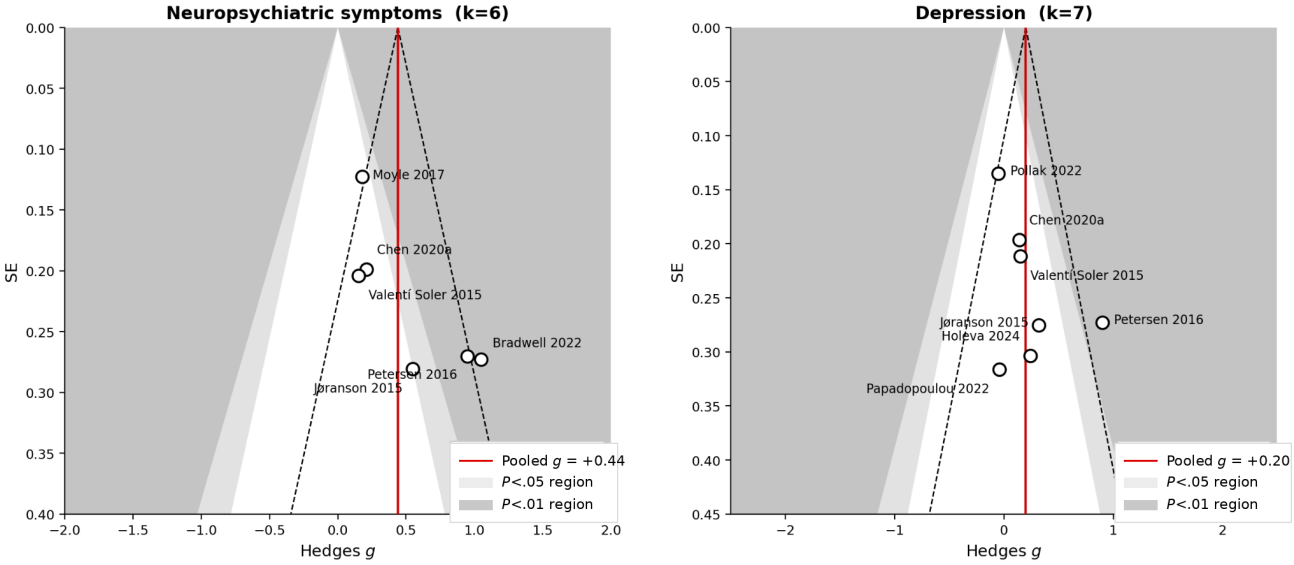

(A) Neuropsychiatric symptoms (B) Depression

Certainty of Evidence

Full GRADE ratings are presented in Table 2. In summary, certainty was rated low for one outcome—NPS—reflecting structural performance bias inherent to unblinded robot trials and substantial inconsistency, partially offset by the directional consistency of effect across most contributing studies. Certainty was rated very low for the remaining 6 outcomes—quality of life, depression, agitation, stress and pain, social-communicative skills, and cognitive function—reflecting a combination of nonsignificant or null pooled estimates, imprecision (95% CIs crossing null and wide PIs relative to effect size), and small contributing sample sizes. The Discussion section is framed in light of these certainty ratings—that is, the single statistically significant pooled effect (NPS, sensitive to one study's inclusion) is interpreted as a suggestive signal supporting careful, selective consideration of care robots in narrow indications, rather than as confirmatory evidence supporting general clinical recommendations or health system integration.

Acceptance and Implementation Outcomes (Secondary)

A minority of included studies reported acceptance- or usability-relevant secondary outcomes (eg, System Usability Scale, Unified Theory of Acceptance and Use of Technology items, observed session completion rates, and facilitator-time data). Because reporting was sparse, inconsistent, and predominantly descriptive, these outcomes were summarized narratively rather than pooled. Across the studies that reported them, acceptance was generally higher in sessions with dedicated facilitator support and in longer-duration protocols.

Discussion

Principal Findings

Relative to the 4 prespecified objectives, this synthesis produced a patterned rather than uniform set of findings. Across the 7 prespecified outcome domains, HKSJ random-effects meta-analysis identified a single domain—NPS—for which the pooled estimate favored care robots and the confidence interval excluded the null. For the remaining 6 domains (quality of life, depression, agitation, stress and pain, social-communicative skills, and cognitive function), the pooled estimates were directionally favorable but did not reach statistical significance. Critically, a distinction must be drawn between the average effect and the distribution of effects across settings; whereas the CI reflects uncertainty about the mean effect, the PI describes the plausible effect in a new setting. In every domain—including NPS—the 95% PI encompassed the null, so although an average benefit is detectable across existing trials for NPS, a beneficial effect cannot be assumed in a new care setting. The neuropsychiatric-symptom signal was, moreover, fragile, losing significance under most LOO exclusions and upon removal of

XSL•FO
RenderX

the single study with documented data extraction ambiguity. Interpreted strictly in light of the substantial between-study heterogeneity observed across pools, the predominance of high or some-concerns RoB 2 ratings, and GRADE certainty ranging from low (NPS) to very low (all other domains), these results indicate that no pooled estimate currently provides a robust basis for clinical decision-making.

Interpretation in the Context of Prior Reviews

Our findings both converge with and extend prior syntheses. For older adults with dementia, the direction and approximate magnitude of the pooled effects we observed for NPS and agitation are consistent with the single-population meta-analyses of Pu et al [11], Leng et al [12], and Yu et al [13], none of which, however, compared these effects with those observed in pediatric populations. For social-communicative skills in children with ASD, our humanoid-platform effect size is directionally consistent with Scassellati et al [6] but, under HKSJ pooling, did not reach statistical significance (Hedges g=0.45, 95% CI −0.52 to 1.42). The principal methodological contribution of this review is the direct juxtaposition of platform–population pairings within a single analytic framework. The pattern of variation across robot platforms and populations observed here is therefore best read as a hypothesis-generating descriptive signal rather than as confirmatory evidence; it nonetheless points to platform-population matching, rather than platform-agnostic deployment, as a useful organizing principle for future trial design and procurement decisions.

Our nonsignificant pooled effects on depression and cognitive function are informative rather than merely negative. Depression in dementia is etiologically complex, with substantial neurobiological determinants that brief psychosocial interventions may be insufficient to modify [62]. The borderline signal in our exploratory subgroup restricted to PARO interventions of ≥10 weeks is hypothesis-generating only and requires prospective adequately powered testing. The null pooled effect for cognitive function is consistent with the broader behavioral-intervention literature and reflects the well-established limitation of noncognitive behavioral inputs in reversing underlying neurodegeneration [40,62-65]; measurement-level sensitivity of global screening instruments (MMSE and MoCA) to detect short-term changes may also contribute.

Care Setting as a Moderator

Setting was not a prespecified subgroup analysis, but the observed effect pattern suggests it warrants interpretation. The domains with the largest pooled effects—NPS and agitation (dementia in long-term care) and social-communicative skills (ASD therapy, predominantly in structured therapy or hospital outpatient settings)—are those in which the intervention was delivered in a structured daily routine with consistent facilitator support. Community and home-setting deployments were underrepresented (n=2) and reported smaller and less consistent effects. The available evidence therefore best supports deployment in settings where the surrounding care structure—scheduled sessions, trained facilitators, documentation routines—can be aligned with the intervention [66], and does not yet speak to unstructured community or home deployment.

Targeted Adjunctive Use Cases

Rather than supporting broad claims about health-system integration, the evidence pattern above—interpreted in the context of low GRADE certainty for the one statistically significant outcome (NPS, sensitive to a single study's inclusion) and very low certainty for the remaining 6 nonsignificant outcomes—is most consistent with 3 narrow adjunctive use cases in which the available evidence and the clinical realities align. These should be regarded as candidate indications meriting further confirmatory trials, not as recommendations for routine clinical adoption.

1. Facilitator-supported calming and engagement sessions for older adults with moderate-to-severe dementia in long-term care. PARO-based sessions of approximately 20‑30 minutes, delivered 2‑3 times per week by trained staff for ≥10 weeks, are the intervention parameters most consistently associated with detectable effects on NPS (agitation pooled effect was directionally similar but did not reach conventional statistical significance). GRADE certainty for this configuration is low. The role is adjunctive—complementing, not replacing, person-centered nursing care—and is most appropriate for units where baseline neuropsychiatric-symptom burden is high and nonpharmacological options are prioritized.
2. Procedural distraction in pediatric inpatient settings. Humanoid robots deployed during time-limited, anxiety-provoking procedures (eg, venipuncture, imaging, and preoperative induction) represent a plausible adjunctive use supported by the stress, pain, and engagement literature [67,68], though with very low GRADE certainty. The time-limited nature of the exposure is itself a strength here: novelty effects, which are a limitation in longer-horizon claims, may be an asset in single-session procedural use.
3. Structured social-communicative practice tasks for children with ASD. Humanoid robot–assisted sessions are best positioned as a structured-practice adjunct to established behavioral therapies such as applied behavior analysis or pivotal response treatment, not as a stand-alone therapeutic modality. GRADE certainty is very low (the social-communicative skills pool did not reach statistical significance under HKSJ pooling), and transfer of gains from robot-mediated to human-mediated social interaction remains an unresolved question that future trials should address [69,70]

Clinical Effectiveness, Technology Acceptance, and Implementation Feasibility

The 3 dimensions of clinical effectiveness (does the intervention work under trial conditions?), technology acceptance (will patients, families, and staff use it?), and implementation feasibility (can the intervention be embedded sustainably in routine care?) are related but distinct, and the evidence base for each differs in maturity. This review quantifies clinical effectiveness and finds a patterned but uneven picture. Technology acceptance is supported by a parallel body of qualitative and survey-based literature [7,8,71] indicating

XSL•FO
RenderX

generally positive acceptance when sessions are facilitator-supported, culturally aligned, and adequately introduced; but acceptance ratings are variable across populations and platforms, and this review did not pool these data. Implementation feasibility remains the least-developed dimension: we did not identify studies that reported cost-effectiveness, institutional readiness, or sustained-deployment outcomes with the rigor required for pooled synthesis. Conflating these 3 dimensions risks overstating the readiness of care-robot deployment; they should be reported separately in future trials and reviews.

Acceptability is decisive for translation yet was not poolable in this effectiveness-focused synthesis, and it is especially salient in older adults, in whom acceptance is shaped by prior attitudes, perceived usefulness, and design rather than being a given. Dedicated work in this population—including the development of instruments to measure acceptance of social robotics and willingness-to-interact studies—underscores that acceptance is a distinct, measurable construct that should be assessed in its own right rather than inferred from efficacy [72,73].

Clinical and Implementation Implications

From a digital health and nursing informatics perspective, these findings, interpreted in light of low or very low GRADE certainty across all 7 outcome domains, suggest several conditional implementation implications rather than firm recommendations. First, where care robots are considered for selective adoption in the narrow indications identified above, they would be most appropriately positioned within the broader digital health ecosystem rather than deployed as isolated devices. PARO and humanoid robots equipped with sensor arrays and data-logging capabilities can generate structured patient engagement metrics—interaction duration, motion activity, physiological indicators—that may be interoperable with electronic health records and clinical decision-support systems where local infrastructure permits. The clinical value of such integration has not been formally evaluated in the trials reviewed here [74].

Strengths and Limitations

Strengths of this review include a cross-population, cross-platform quantitative synthesis; use of the HKSJ random-effects estimator with PIs rather than CIs alone; explicit distinction between average effects and their distribution across settings; PRISMA 2020, PRISMA 2020 for Abstracts, and PRISMA-S adherence with all checklists supplied; dual-reviewer screening, extraction, risk-of-bias assessment, and GRADE rating throughout; and a comprehensive search of 5 databases and 2 trial registries without language or date restrictions.

Limitations should be read in light of the cautious claims above. First, the review was not prospectively registered in PROSPERO. Although an a priori protocol was developed and is available on request, the absence of public prospective registration reduces independent verifiability of prespecification claims. Second, performance bias is structurally inherent to care robot research: blinding of participants and care providers to robotic versus nonrobotic interaction is impossible, and this domain consistently represented the highest RoB 2 ratings, rated high or some concerns in all 34 [18-22,24-52] included studies. Attention-matched active control conditions (Moyle et al [24] plush and Yun et al [50] therapist-guided routine) partially mitigate this concern but cannot eliminate it. Third, considerable clinical and methodological heterogeneity remains across pools, ranging from low to high, reflecting diversity in robot platforms, target populations, intervention dose, and outcome instruments. Fourth, the included sample remains modest, with a median per-study analyzed n=62 and only 4 [20,24,25,27] included studies enrolling >200 participants. Fifth, follow-up durations were generally short, with most studies measuring outcomes at end-of-intervention (≤16 weeks); long-term sustainability and dose-response relationships remain underexplored. Sixth—and new in this revision—primary-source data verification revealed limitations specific to data extraction: (1) one cohort overlap (Mervin et al [27] with Moyle et al [24] cluster RCT, n=415 Australian residents) was identified during full-text reextraction and resolved via reclassification of Mervin et al [27] to narrative-only synthesis under Senn [58]; (2) Robinson et al [44] was confirmed as a qualitative observational follow-up of an external RCT cohort and reclassified to narrative-only; and (3) Petersen et al [34] published a table whose directional interpretation of RAID and CSDD change scores is inconsistent with the abstract narrative, prompting a sensitivity analysis. When Petersen was excluded, the only statistically significant pool (NPS) became nonsignificant under HKSJ pooling, highlighting the modest evidentiary base. Finally, 3 included studies were small pilot RCTs (Marino et al [41] n=14; Yun et al [50] n=15; Srinivasan et al [47] n=24), and several studies analyzed secondary outcomes rather than preregistered primary endpoints. Together, these limitations argue for caution in interpreting any single pooled estimate and for prioritizing future adequately powered, multicenter RCTs with preregistered primary outcomes. A further limitation specific to data extraction quality is that mean and SD values for 9 of 37 study-outcome contributions (24%) were derived from text or figure approximations rather than from published tables; sensitivity analyses excluding these contributions did not change whether any pool reached statistical significance, but produced moderate changes in pooled point estimates (notably a substantial attenuation of the stress and pain estimate when text or figure-approximated contributions were excluded). To make the link to certainty explicit, these limitations map directly onto the GRADE downgrade domains recorded in Table 2. The structural impossibility of blinding and the predominance of high or some concerns RoB 2 ratings drive the risk-of-bias downgrade; the wide range of I^2 across pools drives the inconsistency downgrade; and the modest per-study sample sizes, together with CIs crossing the null and wide PIs, drive the imprecision downgrade. Reliance on text- or figure-approximated data for a subset of contributions bears on both the risk of bias and imprecision judgments. The deliberate cross-population pooling, although a strength for breadth, introduces a degree of indirectness when pooled estimates are applied to any single target population; we flag this as a consideration even though it did not by itself trigger a separate GRADE downgrade in our profile.

XSL•FO
RenderX

Future Research Directions

Several priorities emerge directly from the gaps identified above. Multicenter RCTs of adequate size, with attention-matched active controls and assessor-blinded outcome measurement, are the most immediate methodological need; sample sizes should be informed by the HKSJ-corrected effect estimates reported here, which are typically more conservative than earlier DerSimonian-Laird–based values and therefore produce larger sample size targets. Standardization of outcome measurement through a stakeholder-derived core outcome set, developed along lines analogous to the COMET Initiative [8], would substantially improve comparability across future trials. Follow-up periods of 3, 6, and 12 months postintervention are necessary to distinguish short-term engagement from durable clinical benefit—a distinction that current short-duration evidence cannot address. Formal cost-effectiveness analyses and budget-impact modeling are required to inform health technology assessment. Implementation science studies that explicitly measure staff training requirements, institutional readiness, and patient and family acceptance as implementation outcomes in their own right would begin to build the implementation-feasibility evidence base that is currently absent. Cross-cultural replication, particularly in low- and middle-income settings, and mechanistic studies pairing behavioral outcomes with physiological markers (cortisol and heart rate variability) or neuroimaging where feasible, round out the research agenda.

Conclusions

This systematic review and meta-analysis provides one of the broadest quantitative syntheses to date of care-robot RCTs across patient populations and outcome domains, and is best read as an exploratory, hypothesis-generating synthesis rather than a definitive one. Following systematic full-text data verification of all included randomized controlled studies, a statistically significant pooled effect favoring care robots emerged in only one of the 7 prespecified domains, NPS. That signal proved fragile under LOO and data extraction sensitivity analyses, and its PI, like those of the remaining 6 domains, encompassed the null, so a benefit cannot be assumed in any new care setting. Across the other 6 domains, pooled effects favored care robots in direction but did not reach statistical significance. Interpreted strictly in light of substantial between-study heterogeneity, the predominance of high or some concerns risk-of-bias ratings, and GRADE certainty of low (NPS) to very low (all other domains), these findings indicate that the evidence base is more limited than earlier syntheses suggested, a difference attributable to the HKSJ small-sample correction, primary-source data verification, and PI reporting. The current evidence is therefore promising but not yet sufficient to support routine clinical adoption of care robots across the breadth of populations and outcomes examined. Where benefits may be most plausibly realized, neuropsychiatric symptom reduction in older adults with dementia and procedural-distress mitigation in pediatric settings, they should be regarded as candidate indications requiring confirmation, not as established uses. Accordingly, care robots are best positioned as facilitator-supported adjuncts that augment, rather than substitute for, human-delivered care, and broader implementation claims should await adequately powered, multicenter trials with attention-matched controls, blinded outcome assessment, and extended postintervention follow-up.

Acknowledgments

No generative artificial intelligence tools were used in the conception, design, conduct, data extraction, statistical analysis, interpretation, drafting, or revision of this manuscript. All content is the sole work of the authors, who reviewed and approved the final text and take full responsibility for it.

Funding

This study was funded by Keimyung University Dongsan Medical Center research promotion grant (2023); the funder had no role in this review.

Data Availability

The datasets analyzed during this study are available from the corresponding author on reasonable request. All primary data were extracted from previously published, publicly accessible peer-reviewed reports cited in this article.

Authors' Contributions

GK conceptualized and designed the study, developed the electronic search strategy, conducted database searches, performed screening and full-text eligibility assessment, extracted and verified all data, assessed risk of bias using the RoB 2 tool, performed meta-analyses including subgroup and sensitivity analyses, evaluated evidence certainty using GRADE, and drafted and critically revised the manuscript. JJ contributed to full-text screening, data extraction verification, and manuscript review. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Characteristics of the 34 included randomized and cluster-randomized controlled trials of care robot interventions (per-study; full version), conducted across 17 countries and published 2015-2024. Target populations include older adults with dementia or mild cognitive impairment, children and adolescents with autism spectrum disorder, hospitalized or procedurally distressed pediatric patients, and other adult clinical populations. Robot platforms include PARO, humanoid platforms (NAO, Pepper, Kabochan, CommU, and Kaspar), and other companion and pet platforms. Cohort-overlap allocation per Senn (2009) is documented in the footnote.
[DOCX File, 32 KB - jmir_v28i1e95232_app1.docx]

Checklist 1
PRISMA-S checklist.
[DOCX File, 20 KB - jmir_v28i1e95232_app2.docx]

Checklist 2
PRISMA 2020 for Abstracts checklist.
[DOCX File, 269 KB - jmir_v28i1e95232_app3.docx]

Checklist 3
PRISMA Expanded checklist.
[DOCX File, 25 KB - jmir_v28i1e95232_app4.docx]

References

1. World report on ageing and health. World Health Organization. 2015. URL: https://www.who.int/publications/i/item/9789241565042 [accessed 2026-06-17]
2. Global strategy on human resources for health: Workforce 2030. World Health Organization. URL: https://www.who.int/publications/i/item/9789241511131 [accessed 2026-06-17]
3. Wada K, Shibata T. Living with seal robots—its sociopsychological and physiological influences on the elderly at a care house. IEEE Trans Robot 2007;23(5):972-980. [doi: 10.1109/TRO.2007.906261]
4. Biocca F, Harms C, Burgoon JK. Toward a more robust theory and measure of social presence: review and suggested criteria. Presence (Camb) 2003 Oct;12(5):456-480. [doi: 10.1162/105474603322761270]
5. Kitwood T. Dementia Reconsidered: The Person Comes First: Open University Press; 1997. URL: https://search.worldcat.org/title/1249788841 [accessed 2026-06-25]
6. Scassellati B, Admoni H, Matarić M. Robots for use in autism research. Annu Rev Biomed Eng 2012;14:275-294. [doi: 10.1146/annurev-bioeng-071811-150036] [Medline: 22577778]
7. He Y, He Q, Liu Q. Technology acceptance in socially assistive robots: scoping review of models, measurement, and influencing factors. J Healthc Eng 2022;2022:6334732. [doi: 10.1155/2022/6334732] [Medline: 35911583]
8. Williamson PR, Altman DG, Bagley H, et al. The COMET handbook: version 1.0. Trials 2017 Jun 20;18(Suppl 3):280. [doi: 10.1186/s13063-017-1978-4] [Medline: 28681707]
9. Ciuffreda I, Amabili G, Casaccia S, et al. Design and development of a technological platform based on a sensorized social robot for supporting older adults and caregivers: GUARDIAN ecosystem. Int J Soc Robotics 2025 May;17(5):803-822. [doi: 10.1007/s12369-023-01038-5]
10. Amabili G, Maranesi E, Margaritini A, et al. Managing cognitive decline through a social robot-based intervention: protocol for the engAGE proof of concept and randomized controlled trial. JMIR Res Protoc 2025 Sep 2;14:e67601. [doi: 10.2196/67601] [Medline: 40896828]
11. Pu L, Moyle W, Jones C, Todorovic M. The effectiveness of social robots for older adults: a systematic review and meta-analysis of randomized controlled studies. Gerontologist 2019 Jan 9;59(1):e37-e51. [doi: 10.1093/geront/gny046] [Medline: 29897445]
12. Leng M, Liu P, Zhang P, et al. Pet robot intervention for people with dementia: a systematic review and meta-analysis of randomized controlled trials. Psychiatry Res 2019 Jan;271:516-525. [doi: 10.1016/j.psychres.2018.12.032] [Medline: 30553098]
13. Yu C, Sommerlad A, Sakure L, Livingston G. Socially assistive robots for people with dementia: systematic review and meta-analysis of feasibility, acceptability and the effect on cognition, neuropsychiatric symptoms and quality of life. Ageing Res Rev 2022 Jun;78:101633. [doi: 10.1016/j.arr.2022.101633] [Medline: 35462001]
14. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
15. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]

XSL•FO
RenderX

16. Higgins JPT, Thomas J. In: Chandler J, editor. Cochrane Handbook for Systematic Reviews of Interventions Version 64: Cochrane; 2023. URL: https://training.cochrane.org/handbook [accessed 2026-06-19]
17. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282. [Medline: 23092060]
18. Chen K, Lou VWQ, Tan KCK, Wai MY, Chan LL. Effects of a humanoid companion robot on dementia symptoms and caregiver distress for residents in long-term care. J Am Med Dir Assoc 2020 Nov;21(11):1724-1728. [doi: 10.1016/j.jamda.2020.05.036] [Medline: 32713772]
19. De Korte MW, van den Berk-Smeekens I, van Dongen-Boomsma M, et al. Self-initiations in young children with autism during pivotal response treatment with and without robot assistance. Autism 2020 Nov;24(8):2117-2128. [doi: 10.1177/1362361320935006] [Medline: 32730096]
20. Pollak C, Wexler SS, Drury L. Effect of a robotic pet on social and physical frailty in community-dwelling older adults: a randomized controlled trial. Res Gerontol Nurs 2022;15(5):229-237. [doi: 10.3928/19404921-20220830-01] [Medline: 36113009]
21. Karner S, Stenner H, Spate M, Behrens J, Krakow K. Effects of a robot intervention on visuospatial hemineglect in postacute stroke patients: a randomized controlled trial. Clin Rehabil 2019 Dec;33(12):1940-1948. [doi: 10.1177/0269215519865993] [Medline: 31409126]
22. Ali S, Manaloor R, Ma K, et al. A randomized trial of robot-based distraction to reduce children's distress and pain during intravenous insertion in the emergency department. Can J Emerg Med 2021 Jan;23(1):85-93. [doi: 10.1007/s43678-020-00023-5] [Medline: 33683608]
23. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
24. Moyle W, Jones CJ, Murfield JE, et al. Use of a robotic seal as a therapeutic tool to improve dementia symptoms: a cluster-randomized controlled trial. J Am Med Dir Assoc 2017 Sep 1;18(9):766-773. [doi: 10.1016/j.jamda.2017.03.018] [Medline: 28780395]
25. Moyle W, Jones C, Murfield J, et al. Effect of a robotic seal on the motor activity and sleep patterns of older people with dementia, as measured by wearable technology: a cluster-randomised controlled trial. Maturitas 2018 Apr;110:10-17. [doi: 10.1016/j.maturitas.2018.01.007] [Medline: 29563027]
26. Jones C, Moyle W, Murfield J, et al. Does cognitive impairment and agitation in dementia influence intervention effectiveness? Findings from a cluster-randomized-controlled trial with the therapeutic robot, PARO. J Am Med Dir Assoc 2018 Jul;19(7):623-626. [doi: 10.1016/j.jamda.2018.02.014] [Medline: 29656838]
27. Mervin MC, Moyle W, Jones C, et al. The cost-effectiveness of using PARO, a therapeutic robotic seal, to reduce agitation and medication use in dementia: findings from a cluster-randomized controlled trial. J Am Med Dir Assoc 2018 Jul;19(7):619-622. [doi: 10.1016/j.jamda.2017.10.008] [Medline: 29325922]
28. Jøranson N, Pedersen I, Rokstad AMM, Ihlebaek C. Change in quality of life in older people with dementia participating in Paro-activity: a cluster-randomized controlled trial. J Adv Nurs 2016 Dec;72(12):3020-3033. [doi: 10.1111/jan.13076] [Medline: 27434512]
29. Jøranson N, Pedersen I, Rokstad AMM, Ihlebæk C. Effects on symptoms of agitation and depression in persons with dementia participating in robot-assisted activity: a cluster-randomized controlled trial. J Am Med Dir Assoc 2015 Oct 1;16(10):867-873. [doi: 10.1016/j.jamda.2015.05.002] [Medline: 26096582]
30. Jøranson N, Olsen C, Calogiuri G, Ihlebæk C, Pedersen I. Effects on sleep from group activity with a robotic seal for nursing home residents with dementia: a cluster randomized controlled trial. Int Psychogeriatr 2021 Oct;33(10):1045-1056. [doi: 10.1017/S1041610220001787] [Medline: 32985396]
31. Chen K, Lou VWQ, Tan KCK, Wai MY, Chan LL. Changes in technology acceptance among older people with dementia: the role of social robot engagement. Int J Med Inform 2020 Sep;141:104241. [doi: 10.1016/j.ijmedinf.2020.104241]
32. van den Berk-Smeekens I, de Korte MWP, van Dongen-Boomsma M, et al. Pivotal response treatment with and without robot-assistance for children with autism: a randomized controlled trial. Eur Child Adolesc Psychiatry 2022 Dec;31(12):1871-1883. [doi: 10.1007/s00787-021-01804-8] [Medline: 34106357]
33. Valentí Soler M, Agüera-Ortiz L, Olazarán Rodríguez J, et al. Social robots in advanced dementia. Front Aging Neurosci 2015;7:133. [doi: 10.3389/fnagi.2015.00133] [Medline: 26388764]
34. Petersen S, Houston S, Qin H, Tague C, Studley J. The utilization of robotic pets in dementia care. J Alzheimers Dis 2017;55(2):569-574. [doi: 10.3233/JAD-160703] [Medline: 27716673]
35. Feingold-Polak R, Barzel O, Levy-Tzedek S. A robot goes to rehab: a novel gamified system for long-term stroke rehabilitation using a socially assistive robot-methodology and usability testing. J Neuroeng Rehabil 2021 Jul 28;18(1):122. [doi: 10.1186/s12984-021-00915-2] [Medline: 34321035]
36. Holeva V, Nikopoulou VA, Lytridis C, et al. Effectiveness of a robot-assisted psychological intervention for children with autism spectrum disorder. J Autism Dev Disord 2024 Feb;54(2):577-593. [doi: 10.1007/s10803-022-05796-5] [Medline: 36331688]
37. Johanson DL, Ahn HS, MacDonald BA, et al. The effect of robot attentional behaviors on user perceptions and behaviors in a simulated health care interaction: randomized controlled trial. J Med Internet Res 2019 Oct 4;21(10):e13667. [doi: 10.2196/13667] [Medline: 31588904]

XSL•FO
RenderX

38. Kitt ER, Crossman MK, Matijczak A, Burns GB, Kazdin AE. Evaluating the role of a socially assistive robot in children's mental health care. J Child Fam Stud 2021;30(7):1722-1735. [doi: 10.1007/s10826-021-01977-5] [Medline: 34025101]
39. Kumazaki H, Yoshikawa Y, Yoshimura Y, et al. The impact of robotic intervention on joint attention in children with autism spectrum disorders. Mol Autism 2018;9:46. [doi: 10.1186/s13229-018-0230-8] [Medline: 30202508]
40. Logan DE, Breazeal C, Goodwin MS, et al. Social robots for hospitalized children. Pediatrics 2019 Jul;144(1):e20181511. [doi: 10.1542/peds.2018-1511] [Medline: 31243158]
41. Marino F, Chilà P, Sfrazzetto ST, et al. Outcomes of a robot-assisted social-emotional understanding intervention for young children with autism spectrum disorders. J Autism Dev Disord 2020 Jun;50(6):1973-1987. [doi: 10.1007/s10803-019-03953-x] [Medline: 30852783]
42. Papadopoulos C, Castro N, Nigath A, et al. The CARESSES randomised controlled trial: exploring the health-related impact of culturally competent artificial intelligence embedded Into socially assistive robots and tested in older adult care homes. Int J Soc Robot 2022;14(1):245-256. [doi: 10.1007/s12369-021-00781-x] [Medline: 33907589]
43. Papadopoulou MT, Karageorgiou E, Kechayas P, et al. Efficacy of a robot-assisted intervention in improving learning performance of elementary school children with specific learning disorders. Children (Basel) 2022 Jul 31;9(8):1155. [doi: 10.3390/children9081155] [Medline: 36010046]
44. Robinson H, Broadbent E, MacDonald B. Group sessions with Paro in a nursing home: structure, observations and interviews. Australas J Ageing 2016 Jun;35(2):106-112. [doi: 10.1111/ajag.12199] [Medline: 26059390]
45. Rossi S, Santini SJ, Di Genova D, et al. Using the social robot NAO for emotional support to children at a pediatric emergency department: randomized clinical trial. J Med Internet Res 2022 Jan 13;24(1):e29656. [doi: 10.2196/29656] [Medline: 34854814]
46. So WC, Wong MKY, Lam CKY, et al. Using a social robot to teach gestural recognition and production in children with autism spectrum disorders. Disabil Rehabil Assist Technol 2018 Aug;13(6):527-539. [doi: 10.1080/17483107.2017.1344886] [Medline: 28673117]
47. Srinivasan SM, Park IK, Neelly LB, Bhat AN. A comparison of the effects of rhythm and robotic interventions on repetitive behaviors and affective states of children with autism spectrum disorder (ASD). Res Autism Spectr Disord 2015 Oct 1;18:51-63. [doi: 10.1016/j.rasd.2015.07.004] [Medline: 26251668]
48. Stoevesandt D, Jahn P, Watzke S, et al. Comparison of acceptance and knowledge transfer in patient information before an MRI exam administered by humanoid robot versus a tablet computer: a randomized controlled study. Rofo 2021 Aug;193(8):947-954. [doi: 10.1055/a-1382-8482] [Medline: 34111898]
49. Wei CW, Kao HY, Wu WH, Chen CY, Fu HP. The influence of robot-assisted learning system on health literacy and learning perception. Int J Environ Res Public Health 2021 Oct 21;18(21):11053. [doi: 10.3390/ijerph182111053] [Medline: 34769571]
50. Yun SS, Choi J, Park SK, Bong GY, Yoo H. Social skills training for children with autism spectrum disorder using a robotic behavioral intervention system. Autism Res 2017 Jul;10(7):1306-1323. [doi: 10.1002/aur.1778] [Medline: 28464438]
51. Boumans R, van Meulen F, van Aalst W, et al. Quality of care perceived by older patients and caregivers in integrated care pathways with interviewing assistance from a social robot: noninferiority randomized controlled trial. J Med Internet Res 2020 Sep 9;22(9):e18787. [doi: 10.2196/18787] [Medline: 32902387]
52. Bradwell H, Edwards KJ, Winnington R, Thill S, Allgar V, Jones RB. Implementing affordable socially assistive pet robots in care homes before and during the COVID-19 pandemic: stratified cluster randomized controlled trial and mixed methods study. JMIR Aging 2022 Aug 24;5(3):e38864. [doi: 10.2196/38864] [Medline: 35830959]
53. Cohen J. Statistical Power Analysis for the Behavioral Sciences: Taylor and Francis; 1988. URL: https://search.worldcat.org/title/17877467 [accessed 2026-06-19]
54. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. Stat Med 2001 Dec 30;20(24):3875-3889. [doi: 10.1002/sim.1009] [Medline: 11782040]
55. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
56. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023 Dec;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
57. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019 Jun;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]
58. Senn SJ. Overstating the evidence: double counting in meta-analysis and related problems. BMC Med Res Methodol 2009 Feb 13;9:10. [doi: 10.1186/1471-2288-9-10] [Medline: 19216779]
59. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634. [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
60. Sterne JAC, Sutton AJ, Ioannidis JPA, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011 Jul 22;343:d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
61. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-926. [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]

XSL•FO
RenderX

62. Zhang Y, Cai J, An L, et al. Does music therapy enhance behavioral and cognitive function in elderly dementia patients? A systematic review and meta-analysis. Ageing Res Rev 2017 May;35:1-11. [doi: 10.1016/j.arr.2016.12.003] [Medline: 28025173]
63. Hu M, Zhang P, Leng M, Li C, Chen L. Animal-assisted intervention for individuals with cognitive impairment: a meta-analysis of randomized controlled trials and quasi-randomized controlled trials. Psychiatry Res 2018 Feb;260:418-427. [doi: 10.1016/j.psychres.2017.12.016] [Medline: 29268204]
64. Lyketsos CG, Lopez O, Jones B, Fitzpatrick AL, Breitner J, DeKosky S. Prevalence of neuropsychiatric symptoms in dementia and mild cognitive impairment: results from the cardiovascular health study. JAMA 2002 Sep 25;288(12):1475-1483. [doi: 10.1001/jama.288.12.1475] [Medline: 12243634]
65. Bahar-Fuchs A, Martyr A, Goh AMY, Sabates J, Clare L. Cognitive training for people with mild to moderate dementia. Cochrane Database Syst Rev 2019 Mar 25;3(3):CD013069. [doi: 10.1002/14651858.CD013069.pub2] [Medline: 30909318]
66. Koh WQ, Felding SA, Budak KB, Toomey E, Casey D. Barriers and facilitators to the implementation of social robots for older adults and people with dementia: a scoping review. BMC Geriatr 2021 Jun 9;21(1):351. [doi: 10.1186/s12877-021-02277-9] [Medline: 34107876]
67. Beran TN, Ramirez-Serrano A, Vanderkooi OG, Kuhn S. Reducing children's pain and distress towards flu vaccinations: a novel and effective application of humanoid robotics. Vaccine 2013 Jun 7;31(25):2772-2777. [doi: 10.1016/j.vaccine.2013.03.056] [Medline: 23623861]
68. Jibb LA, Birnie KA, Nathan PC, et al. Using the MEDiPORT humanoid robot to reduce procedural pain and distress in children with cancer: a pilot randomized controlled trial. Pediatr Blood Cancer 2018 Sep;65(9):e27242. [doi: 10.1002/pbc.27242] [Medline: 29893482]
69. Pennisi P, Tonacci A, Tartarisco G, et al. Autism and social robotics: a systematic review. Autism Res 2016 Feb;9(2):165-183. [doi: 10.1002/aur.1527] [Medline: 26483270]
70. Kouroupa A, Laws KR, Irvine K, Mengoni SE, Baird A, Sharma S. The use of social robots with children and young people on the autism spectrum: a systematic review and meta-analysis. PLoS One 2022;17(6):e0269800. [doi: 10.1371/journal.pone.0269800] [Medline: 35731805]
71. Abdi J, Al-Hindawi A, Ng T, Vizcaychipi MP. Scoping review on the use of socially assistive robot technology in elderly care. BMJ Open 2018 Feb 12;8(2):e018815. [doi: 10.1136/bmjopen-2017-018815] [Medline: 29440212]
72. Bevilacqua R, Di Rosa M, Riccardi GR, et al. Design and development of a scale for evaluating the acceptance of social robotics for older people: the robot era inventory. Front Neurorobot 2022;16:883106. [doi: 10.3389/fnbot.2022.883106] [Medline: 35874107]
73. Bevilacqua R, Felici E, Cavallo F, Amabili G, Maranesi E. Designing acceptable robots for assisting older adults: a pilot study on the willingness to interact. Int J Environ Res Public Health 2021 Oct 12;18(20):10686. [doi: 10.3390/ijerph182010686] [Medline: 34682433]
74. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]

Abbreviations

ASD: autism spectrum disorder
BARS: Behavioral Activity Rating Scale
CMAI-SF: Cohen-Mansfield Agitation Inventory–Short Form
CSDD: Cornell Scale for Depression in Dementia
GDS: Geriatric Depression Scale
GDS-SF: Geriatric Depression Scale – Short Form
GRADE: Grading of Recommendations Assessment, Development and Evaluation
HKSJ: Hartung-Knapp-Sidik-Jonkman
ICTRP: International Clinical Trials Registry Platform
LOO: leave-one-out
MeSH: Medical Subject Headings
MMSE: Mini-Mental State Examination
MoCA: Montreal Cognitive Assessment
NPI: Neuropsychiatric Inventory
NPS: neuropsychiatric symptoms
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search
PROSPERO: International Prospective Register of Systematic Reviews
RAID: Rating Anxiety in Dementia
RCT: randomized controlled trial

XSL•FO
RenderX

ROB 2: Risk of Bias 2
WHO: World Health Organization

Edited by S Brini; submitted 13.Mar.2026; peer-reviewed by A Hamdan, G Amabili, V Patil; revised version received 09.Jun.2026; accepted 10.Jun.2026; published 30.Jun.2026.

Please cite as:
Kim G, Jeon J
Care Robots as Emerging Health Technologies: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e95232
URL: https://www.jmir.org/2026/1/e95232
doi:10.2196/95232

© Gaeun Kim, Jinmi Jeon. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Interventions Targeting Parents to Improve Early Childhood Movement, Nutrition, and Sleep Behaviors: Systematic Review

Johanna Sandborg[1], PhD; Brittany L Reese[1], PhD; Sarah Marshall[1,2], PhD; Kylie D Hesketh[1], PhD; Rachel Laws[1], PhD; Katherine L Downing[1], PhD

[1]Institute for Physical Activity and Nutrition (IPAN), Faculty of Health, Deakin University, Locked Bag 20000, Geelong, VIC, Australia

[2]Sydney School of Public Health, The University of Sydney, Sydney, Australia

Corresponding Author:
Johanna Sandborg, PhD
Institute for Physical Activity and Nutrition (IPAN), Faculty of Health, Deakin University, Locked Bag 20000, Geelong, VIC, Australia

Abstract

Background: Early childhood (0 - 5 years) is key for shaping health behaviors, yet optimal behaviors are rarely achieved. Digital health promotion interventions offer scalable support for families; however, most research has focused on childhood more broadly, leaving limited evidence for the early childhood period.

Objective: The primary aim of this systematic review was to examine whether autonomously delivered digital interventions targeting parents are effective at increasing physical activity, reducing sedentary behavior, improving nutrition (breastfeeding, feeding practices), and/or optimizing sleep among children aged 0 - 5 years. The secondary aim was to review the reporting of co-design practices, user engagement, and process evaluation, and to assess how engagement influences intervention effectiveness.

Methods: Seven databases were searched for randomized controlled trials (RCTs) evaluating autonomously delivered digital interventions targeting one or more of the following behaviors: physical activity, sedentary behavior, nutrition, or sleep among children (published to January 2026). Study quality was assessed using the Joanna Briggs Institute Critical Appraisal Checklist for RCTs. Findings were narratively synthesized by target age-group and behavior, and the direction of effect was summarized in structured tables.

Results: Of the 14,352 identified records, 38 interventions (33 RCTs, 4 pilot RCTs, and 1 feasibility RCT) were included. Most studies focused on pregnancy to infancy (n=24; 0 - 1 y), followed by preschoolers (n=8; 3 - 5 y) and toddlers (n=6; 1 - 2 y). Intervention duration ranged from 2 weeks to 1000 days, and various digital formats were used (apps n=11, SMS text messaging n=10, web- or internet-based platforms n=6, WeChat [Tencent] n=3, tablet-based program n=2, a combination of app and SMS text messaging n=1, website and emails n=1, emails and SMS text messaging n=1, automated voice calls n=1, Facebook Messenger Chatbot [Meta] n=1, and online videos n=1). Interventions spanning pregnancy to infancy reported mixed findings for breastfeeding and feeding practices. Studies targeting toddlers showed improvements in sleep, mixed findings for diet and screen time, and no differences in physical activity. Most studies targeting preschoolers reported significant improvements for feeding practices and diet, but no differences in physical activity, sedentary behavior and sleep, and mixed findings for screen time. Most studies reported co-design or engagement (n=24), but few examined the impact of engagement on intervention effectiveness (n=6), and those that did reported mixed findings. Interpretation was limited by heterogeneous designs, inconsistent outcome measures, and mixed risk-of-bias ratings across studies.

Conclusions: This review advances the field by synthesizing evidence on scalable digital interventions that support parents in promoting healthy lifestyle behaviors across the first 2000 days, together with key design and implementation factors that have rarely been reported in previous reviews. Unlike prior work, it focuses exclusively on autonomously delivered digital interventions in early childhood. Findings show heterogeneous designs and mixed effectiveness, and highlight 3 priority evidence gaps: limited studies in toddlers and preschoolers, incomplete reporting of engagement, and limited understanding of how engagement influences outcomes. These findings define priorities for future research to strengthen the evidence for scalable digital interventions in early childhood.

Trial Registration: PROSPERO CRD42022372639; https://www.crd.york.ac.uk/PROSPERO/view/CRD42022372639

(J Med Internet Res 2026;28:e85525) doi:10.2196/85525

KEYWORDS

mHealth; movement behaviours; sedentary behaviour; screen time; physical activity; breastfeeding; diet; digital; eHealth

XSL•FO
RenderX

Introduction

Background

Early childhood (birth through 5 years) is recognized as a critical period during which key health behaviors (diet, physical activity, sedentary behavior, and sleep) are established. However, evidence shows that these behaviors are suboptimal from early life. Despite the well-established benefits of breastfeeding, fewer than half of infants are exclusively breastfed for the first 6 months [1]. Similarly, global evidence indicates that adherence to early childhood dietary [2-4] and movement behavior guidelines is low, with large international reviews reporting that only a small proportion of children meet recommendations across physical activity, sedentary behavior, screen time, and sleep [2,3]. Of particular concern is the fact that these suboptimal behaviors can track into later childhood and adolescence [4,5], underscoring the need for interventions to promote health behaviors from a young age. Recent studies have also highlighted widening socioeconomic inequalities in children's early environments [6,7] and less optimal diet, physical activity, sedentary behavior, screen time, and sleep among children from socioeconomically disadvantaged backgrounds [8-11], reinforcing the need for accessible intervention strategies for families who may have limited access to traditional face-to-face services.

Existing early childhood interventions have shown varied success in improving diet, physical activity, sedentary behavior, and sleep habits [12-15]. Traditionally, these interventions have relied largely on time-consuming and costly face-to-face delivery, with limited consideration of scalability or implementation at scale [16]. In contrast, digital interventions (eHealth and mobile health [mHealth]) have the advantage that they can be delivered anywhere, anytime, maximizing potential reach across diverse socioeconomic, geographical, and cultural backgrounds. Recent years have seen a rapid expansion of digital health solutions, including web-based platforms, mobile apps, and wearable technologies [17], supported by growing evidence of their potential for scalability and cost-effectiveness [18,19].

Reflecting this broader growth in digital health, there has also been a marked increase in digital interventions targeting diet and movement behaviors across all age groups, with most reporting efficacy in changing behavior [20-22]. This also includes a growing interest in the feasibility and effectiveness of these types of interventions for targeting childhood obesity and obesity-related behaviors [23-31]. However, these reviews have largely focused on childhood broadly (0 - 18 years) [23,24,27-30] or have focused solely on preschoolers (3 - 6 years) [25,26]. Many have also examined single behavioral domains such as physical activity or sedentary behavior [27-29], or specific population groups such as Indigenous mothers of young children [31]. In addition, most evaluate digital interventions that involve some degree of human support or multicomponent programs and do not address key design and implementation factors. As such, they offer limited insight into early childhood as a distinct developmental period or the potential of autonomously delivered interventions to support parents across multiple behaviors in the first 2000 days. Given that health behaviors are largely shaped early in life and that this life stage is characterized by unique developmental and parental influence, a review focusing solely on this period is warranted.

Moreover, co-design, engagement, and implementation factors are highly important for successful intervention delivery [32,33]; yet, these elements remain inconsistently assessed and underreported in interventions targeting parents of young children [34]. Thus, this review provides a timely and comprehensive synthesis of autonomously delivered digital interventions targeting parents across the first 2000 days (conception to age 5 years), examining multiple behavioral domains (breastfeeding, feeding practices, diet, physical activity, sedentary behavior, screen time, and sleep) and integrating evidence on co-design, process evaluation, and engagement. This broader scope allows us to identify key gaps, emerging patterns and implications for the design and implementation of scalable digital strategies in early childhood.

Objectives

The primary aim of this systematic review was to examine whether autonomously delivered digital interventions targeting parents are effective at increasing physical activity, reducing sedentary behavior, improving nutrition (breastfeeding, feeding practices, and dietary outcomes), and/or optimizing sleep among children aged 0 - 5 years. The secondary aim was to review the reporting of co-design practices, user engagement, and process evaluation, and to assess how engagement influences intervention effectiveness.

Methods

Eligibility Criteria

We included randomized controlled trials (RCTs) of interventions delivered to parents and caregivers of children in the first 2000 days of life (herein referred to as parents) solely via digital technology (mHealth/eHealth). Interventions needed to aim to improve one or more of the following behaviors: increase physical activity, reduce sedentary behavior, improve nutrition (breastfeeding, food intake, and feeding practices), and/or optimize sleep among children.

We only included interventions that solely used digital technologies to autonomously deliver the intervention (ie, where no personnel were needed to deliver and/or maintain the intervention). We used this definition as we were interested in solutions for intervention delivery that could be more easily and cost-effectively scaled compared to interventions requiring delivery personnel (with or without a digital component). Therefore, we excluded digital interventions that required delivery personnel for one-to-one support, for example, telephone coaching calls or face-to-face counseling sessions with a supplementary online social support group.

We considered direct outcome measures of children's target behaviors (eg, parent-reported or accelerometer-measured physical activity) and indirect outcome measures known to influence children's target behaviors (eg, parental feeding style or changes to food environment). There were 2 key reasons we decided to also include indirect outcome measures: (1) it can

XSL•FO
RenderX

be difficult to accurately measure child behaviors, particularly for younger children (eg, breastmilk intake), and (2) national guidelines for infant feeding [35] and movement behaviors in the early years [36] include parental influences known to impact child behaviors (eg, for establishing healthy sleep habits, parents can set up a calming bedtime routine and consistent sleep and wake-up times).

Studies were also excluded if they were (1) not an RCT design, (2) solely targeting other caregivers (eg, grandparents or childcare providers), (3) among parents with older children (≥6 years old), (4) among parents with children who had clinical health conditions (eg, diabetes and premature birth), and (5) interventions delivered primarily within the antenatal period (noting that those delivered from pregnancy to infancy were included). We limited studies to primary research published in English in the peer-reviewed literature. Literature reviews and meta-analyses, theses, conference proceedings, and gray literature were not included.

Information Sources

We searched 7 electronic databases, including Embase (Elsevier), Academic Search Complete (EBSCO), CINAHL Complete (EBSCO), Global Health, MEDLINE Complete (EBSCO), PsycINFO (EBSCO), and SPORTDiscus (EBSCO). All databases were searched individually rather than simultaneously via a multidatabase platform. We did not search study registries, conference proceedings, websites, or other online sources, nor did we undertake citation searching or contact authors or experts. No supplementary search methods beyond database searching were used, as the review was limited to peer-reviewed RCTs. Searches were conducted in December 2022 and updated in August 2024 and again in January 2026.

Search Strategy

This systematic review was prospectively registered with PROSPERO (International Prospective Register of Systematic Reviews; ID: CRD42022372639) [37]. We made 2 amendments to the registered protocol: (1) to limit study design to RCTs due to the recent increase in publications and (2) the use of a different quality appraisal tool specific to RCTs. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) 2020 statement guidelines [38] and the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension; Checklist 1) extension for reporting search strategies were followed to ensure transparent reporting of the search process [39].

The search strategy was developed by BM, SM, and KD, in consultation with all authors. We ran preliminary searches and sought technical guidance from Deakin University librarians to refine the search strategy. The search strategy included a combination of keywords to capture concepts according to the Population, Intervention, Comparison, Outcomes, and Study design (PICOS) tool: (1) children aged ≤5 years, (2) mHealth or eHealth intervention, (3) intervention behavioral targets (ie, child physical activity, sedentary behavior, nutrition, and/or sleep), and (4) study design (see Multimedia Appendix 1). The search strategy was not adapted from previous reviews and did not undergo a formal peer-review process.

Selection Process

All search results were exported to Covidence and duplicates were removed. Two reviewers (SM or JS and either BM, KD, or KH) independently screened all articles first by the title and abstract; a third independent reviewer resolved discrepancies (BM or KD). Two reviewers screened the full texts (BM and SM or JS and KD) using the inclusion/exclusion criteria described above. Discrepancies were resolved by a third reviewer (KD or KH). During the updated review in 2026, 2 researchers, CS and SR, assisted with screening, data extraction, and evaluation (disagreements were resolved by JS) following the same procedure.

Data Collection Process

Two reviewers (JS and BM or CS and SR) extracted the following information using a prepiloted data collection template developed for this review. Template fields included study characteristics (eg, authors, year of publication, country, study design, study aims, and target behaviors), setting and participants (eg, setting, inclusion/exclusion criteria, recruitment, and sample size), intervention description (eg, technology used, delivery mode, content, onboarding processes, and engagement), and participant outcomes (eg, measures related to target behavioral outcomes, data collection tool and method, and results). Co-design (eg, stakeholder engagement in intervention development), intervention theory (ie, the use of a behavior change theory in intervention development), process evaluation outcomes (eg, acceptability, feasibility, and reach), and intervention engagement data (eg, app analytics, self-reported use, and impact of engagement on intervention effectiveness) were extracted using a combination of the data extraction tool at Elicit.com [40] (Ought; an online platform that automates data extraction) and manual extraction and checked for accuracy and completeness by one author (KD or CS).

Data Items

We extracted data on all relevant behavioral outcomes (breastfeeding, feeding practices, diet, physical activity, sedentary behavior, screen time, and sleep). For each outcome, all reported time points and measurement tools were collected when available. We also extracted additional study variables, including participant characteristics, intervention features, delivery mode, co-design processes, theoretical underpinnings, process evaluation outcomes, and engagement metrics. When information was unclear or missing, assumptions were not made; instead, data were recorded as reported.

Study Risk of Bias Assessment

The risk of bias and quality assessment for each individual study was assessed by 2 authors independently (JS and KD or CS and SR) using the Joanna Briggs Institute (JBI) Critical Appraisal Checklist for RCTs [41]. The checklist includes 13 items, where each item can be scored yes, no, unclear, or not applicable, in the categories of selection and allocation; administration of intervention/exposure; assessment, detection, and measurement of outcomes; participant retention; and statistical conclusion validity. As the interventions were autonomously delivered (eg, via digital platforms), there were no treatment deliverers involved; therefore, the item “Treatment deliverers blinded”

XSL•FO
RenderX

was not applicable in this context. The initial interrater agreement between JS and KD was 81%, and 85% for CS and SR. Discrepancies in assessment between authors were discussed (JS and KD or CS and SR) until consensus was reached.

Synthesis Methods

Consistent with Cochrane guidance [42], due to the heterogeneity in definition, measurement, and reporting of outcomes across studies, meaning they did not estimate the same underlying effect, a meta-analysis was not able to be conducted. Therefore, following PRISMA [38] and synthesis without meta-analysis recommendations [43], a structured narrative synthesis was undertaken, grouping studies by target age group (pregnancy-infancy, toddlers, and preschoolers) and by behavioral domain (breastfeeding, diet, physical activity, sedentary behavior, and sleep). We summarized the direction of effect and patterns in effectiveness using structured tables.

Certainty Assessment

We did not conduct a formal certainty-of-evidence assessment because of the heterogeneity of the included studies and the lack of commensurable effect estimates. Instead, we considered study-level risk of bias and the consistency of findings when interpreting the results.

Results

Study Selection

The literature search yielded 14,352 unique articles. Most records excluded at the title and abstract stage were due to at least one of the following reasons: the study did not examine an autonomously delivered digital intervention (eg, clinical-delivered or face-to-face), targeted populations outside the scope of the review (eg, older children, adolescents, or adults), or did not use an RCT design (eg, observational studies, qualitative studies, pilot feasibility work without randomization, or large noninterventional epidemiological analyses). Following screening, 258 full-text reports were assessed for eligibility, of which 49 studies describing 38 interventions were included in this review. No studies were excluded at full-text review that appeared to meet the inclusion criteria; all exclusions were based on predefined criteria (eg, wrong age group, design, and intervention type). The screening process is presented in Figure 1.

XSL•FO
RenderX

Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram showing the study selection for randomized controlled trials of autonomously delivered digital interventions targeting early childhood health behaviors (0 - 5 years). Studies involved healthy parent-child populations across multiple countries and were identified through searches conducted in December 2022, August 2024, and January 2026. mHealth: mobile health; RCT: randomized controlled trial.

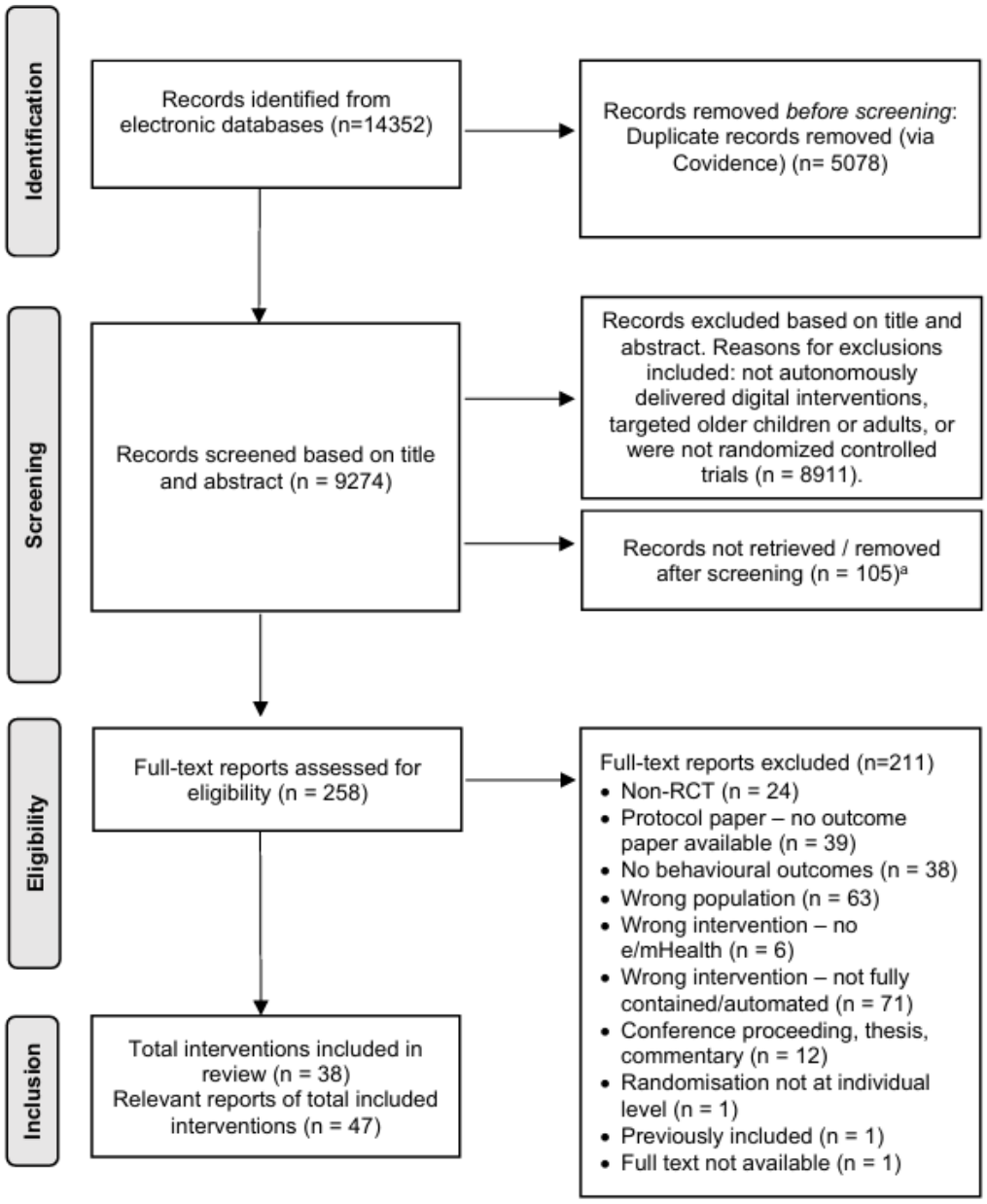

[a] Three database searches were conducted and screened separately in Covidence. Counts were combined for reporting. Minor discrepancies between screening and full-text assessment reflect removal of duplicate records identified across searches after initial screening.

Study Characteristics

An overview of the targeted outcomes and age groups for the 38 included interventions [44-81] is presented in Table 1. Almost half of the included studies (50%, 19/38) [45,48,51,53,55,56,58,59,62,64,65,67,69,70,72,76,77,79,80] focused on improving breastfeeding, 26% (10/38) [44,46,47,49,52,60,61,71,73,81] focused on multiple behaviors, 16% (6/38) [50,54,66,74,75,78] on diet only, 5% (2/38) [63,68] on sleep only, and 3% (1/38) [57] on physical activity only. The multiple behavior interventions targeted different combinations of breastfeeding, feeding practices, diet, physical activity, sedentary behavior, screen time, and sleep. Only one study focused on all behaviors (breastfeeding, feeding practices, physical activity, sedentary behavior, screen time, and sleep) [61], and no study focused solely on sedentary behavior/screen time. One study included children aged 0 - 3 years but is reported under the infant age group because the mean child age fell within infancy [72].

XSL•FO
RenderX

Table . Overview of the 38 randomized controlled trials of autonomously delivered digital interventions for healthy parent-child populations (0 - 5 years), conducted across multiple countries and published up to January 2026.

Age group	Breastfeeding	Combined[a]	Diet	Physical activity	Sedentary behavior	Sleep
Newborn/infants[b,c]	19	2	2	0	0	1
Toddlers[d]	0	3	2	0	0	1
Preschoolers[e]	0	5	2	1	0	0

[a]The combined interventions focused on multiple behaviors eg, diet and physical activity/sedentary behavior.

[b]One of the studies targeted fathers; the others targeted mothers.

[c]0-2 months of age for newborns and 3-11 months of age for infants.

[d]12-35 months of age (≥1 to < 3 years).

[e]36-59 months of age (≥3 to < 5 years).

Tables 2-4 present the study characteristics for the included studies divided by age group. In summary, the included studies comprised RCTs (n=32) [44-63,68,70-80], pilot RCTs (n=4) [47,64-66], one feasibility RCT [67], and one pilot study using a micro-RCT design [81]. Studies were published between 2011 and 2026, with most studies being published since 2020 (27/38, 71%) [51-62,64,65,69-81]. Most studies were conducted in the United States (n=14) [44,45,47,48,57,64-70,76], followed by Australia (n=5) [49,52,53,71,80], China (n=3) [51,61,72], Sweden (n=2) [46,60], Norway (n=2) [50,54], Ethiopia (n=2) [77,79], and Thailand (n=2) [73,74]. Single studies were conducted in India [56], Myanmar [59], Nepal [78], Iran [75], Turkey [58], Vietnam [55], Spain [62] and Canada [81]. The studies included sample sizes ranging from 18 [81] to 5095 [56].

XSL•FO
RenderX

Table . Risk of bias assessment table (using the Joanna Briggs Institute [JBI] Critical Appraisal Checklist for randomized controlled trials [41]).

Author (date)	Selection and allocation			Administration of intervention/exposure			Assessment, detection, and measurement of the outcome			Participant retention	Statistical conclusion validity		
	True randomization	Group allocation concealed	Groups similar at baseline	Participants blinded	Treatment deliverers blinded[a]	Groups treated identically (other than intervention)	Outcome assessors blinded	Outcomes measured the same way for groups	Outcomes measured in a reliable way	Follow-up complete	Intention-to-treat (ITT) analysis	Appropriate statistical analysis	Trial design appropriate
Pregnancy and infancy													
Breastfeeding													
Ahmed et al (2016) [45]	Y[b]	U[c]	Y	N[d]	N	Y	U	Y	Y	Y	Y	Y	Y
Unger et al (2018) [48]	Y	Y	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y
Wu et al (2020) [51]	Y	Y	N	U	N	Y	Y	Y	U	Y	Y	Y	Y
Lewkowitz et al (2020) [69]	Y	Y	Y	Y	—[e]	Y	Y	Y	Y	Y	Y	Y	Y
Scott et al (2021) [53]	Y	N	Y	N	N	Y	U	Y	Y	Y	Y	Y	Y
Saucedo Baza et al (2022) [65]	Y	U	Y	N	N	Y	U	Y	Y	N	Y	Y	Y
Doan et al (2022) [55]	Y	Y	Y	Y	—	Y	Y	Y	Y	Y	Y	Y	Y
LeFevre et al (2022) [56]	Y	Y	Y	N	N	Y	Y	Y	U	Y	Y	Y	Y
Acar and Sahin (2023) [58]	Y	U	Y	N	U	Y	U	Y	U	Y	Y	Y	Y
Hmone et al (2023) [59]	Y	Y	N	N	U	Y	Y	Y	U	Y	Y	Y	Y

XSL•FO
RenderX

Author (date)	Selection and allocation			Administration of intervention/exposure			Assessment, detection, and measurement of the outcome			Participant retention	Statistical conclusion validity		
	True randomization	Group allocation concealed	Groups similar at baseline	Participants blinded	Treatment deliverers blinded[a]	Groups treated identically (other than intervention)	Outcome assessors blinded	Outcomes measured the same way for groups	Outcomes measured in a reliable way	Follow-up complete	Intention-to-treat (ITT) analysis	Appropriate statistical analysis	Trial design appropriate
Vila-Candel et al (2024) [62]	Y	Y	N	U	N	N	Y	Y	U	Y	Y	Y	Y
Henshaw et al (2024) [76]	Y	U	U	N	—	Y	U	Y	Y	Y	N	Y	Y
De Mello et al (2025) [70]	Y	U	Y	N	N	Y	U	Y	U	N	N	Y	Y
Brown et al (2025) [80]	Y	Y	Y	Y	—	Y	U	Y	Y	U	Y	Y	Y
Cherie et al (2025) [79]	Y	U	Y	N	—	Y	Y	Y	Y	Y	Y	Y	Y
Gilano et al (2025) [77]	Y	U	Y	N	—	Y	Y	Y	Y	Y	Y	Y	Y
Breastfeeding and feeding practices													
Palacios et al (2018) [67]	Y	U	Y	U	—	Y	U	Y	Y	Y	Y	Y	Y
Davis et al (2023) [64]	Y	Y	N	N	N	Y	Y	Y	U	Y	Y	Y	Y
Li et al (2024) [72][f]	Y	U	N	N	N	Y	U	Y	Y	N	Y	Y	Y
Combined[g]													
Wen et al (2020) [52]	Y	U	Y	N	—	Y	Y	Y	Y	Y	Y	Y	Y

XSL•FO
RenderX

Author (date)	Selection and allocation			Administration of intervention/exposure			Assessment, detection, and measurement of the outcome			Participant retention	Statistical conclusion validity		
	True randomization	Group allocation concealed	Groups similar at baseline	Participants blinded	Treatment deliverers blinded[a]	Groups treated identically (other than intervention)	Outcome assessors blinded	Outcomes measured the same way for groups	Outcomes measured in a reliable way	Follow-up complete	Intention-to-treat (ITT) analysis	Appropriate statistical analysis	Trial design appropriate
Wu et al (2023) [61]	Y	N	N	N	N	N	N	Y	U	Y	Y	Y	Y
Diet													
Røed et al (2021) [54]	Y	U	Y	N	N	Y	U	Y	Y	Y	Y	Y	Y
Helle et al (2019) [50]	Y	Y	Y	N	—	Y	U	Y	Y	Y	Y	Y	Y
Sleep													
Moon et al (2017) [63]	Y	U	N	U	U	Y	U	Y	U	Y	U	Y	Y
Toddlerhood													
Combined[g]													
Alexandrou et al (2023) [60]	Y	N	Y	N	—	Y	Y	Y	Y	Y	Y	Y	Y
Sandborg et al (2025) [71]	Y	Y	N	N	U	Y	N	Y	Y	N	Y	Y	Y
Jongpaiboonpatana et al (2025) [73]	Y	U	Y	N	U	Y	Y	Y	U	Y	Y	Y	Y
Diet													
Cunningham et al (2023) [78]	Y	Y	Y	N	U	Y	Y	Y	Y	Y	Y	Y	Y
Hunsrisakhun et al (2025) [74]	Y	Y	Y	N	U	Y	Y	Y	Y	Y	N	Y	Y
Sleep													

XSL•FO
RenderX

Author (date)	Selection and allocation			Administration of intervention/exposure			Assessment, detection, and measurement of the outcome			Participant retention	Statistical conclusion validity		
	True randomization	Group allocation concealed	Groups similar at baseline	Participants blinded	Treatment deliverers blinded[a]	Groups treated identically (other than intervention)	Outcome assessors blinded	Outcomes measured the same way for groups	Outcomes measured in a reliable way	Follow-up complete	Intention-to-treat (ITT) analysis	Appropriate statistical analysis	Trial design appropriate
Mindell et al (2011) [68]	U	U	N	Y	—	Y	Y	Y	Y	U	U	N	Y
Preschool													
Combined[g]													
Knowlden et al (2015) [44]	Y	U	Y	Y	U	N	Y	Y	Y	Y	Y	Y	Y
Delisle-Nyström et al (2017) [46]	Y	Y	Y	N	N	Y	Y	Y	Y	Y	Y	Y	Y
Sun et al (2017) [47]	U	U	Y	N	N	Y	Y	Y	Y	Y	Y	Y	Y
Hammersley et al (2019) [49]	Y	Y	Y	Y	N	N	Y	Y	Y	Y	Y	Y	Y
Diet													
Bakirci-Taylor et al (2019) [66]	Y	Y	N	U	N	Y	N	Y	Y	Y	N	Y	Y
Hojati et al (2024) [75]	Y	Y	Y	N	N	Y	Y	Y	Y	Y	U	Y	Y
Physical activity													
Staiano et al (2022) [57]	Y	U	Y	N	—	N	Y	Y	Y	Y	Y	Y	Y

XSL•FO
RenderX

Author (date)	Selection and allocation			Administration of intervention/exposure			Assessment, detection, and measurement of the outcome			Participant retention	Statistical conclusion validity		
	True randomization	Group allocation concealed	Groups similar at baseline	Participants blinded	Treatment deliverers blinded[a]	Groups treated identically (other than intervention)	Outcome assessors blinded	Outcomes measured the same way for groups	Outcomes measured in a reliable way	Follow-up complete	Intention-to-treat (ITT) analysis	Appropriate statistical analysis	Trial design appropriate
Phillips et al (2026) [81]	Y	U	U	N	U	Y	U	Y	Y	N	N	Y	Y

[a]Due to the nature of the interventions under investigation (they were all autonomously delivered), blinding of treatment deliverers was not applicable in this context.

[b]Y: yes.

[c]U: unclear.

[d]N: no.

[e]Not applicable.

[f]This study enrolled children up to 3 years, but the average age in both arms fell within infancy; the study is therefore reported under the pregnancy-infancy category.

[g]The combined interventions focused on multiple behaviors eg, diet and physical activity/sedentary behavior.

XSL•FO
RenderX

Table . Study characteristics and effectiveness of autonomously delivered digital interventions focusing on pregnancy to infancy (0 - 11 months; 0 - 2 months of age for newborns and 3 - 11 months of age for infants): randomized controlled trials published between 2011 - 2026 and across multiple countries (n=24).

Author (date)	Country	N	Target group	Duration[a]	Follow up	Intervention type	Assessment	Breastfeeding	Feeding practices	Diet	PA[b]	SB[c]	ST[d]	Sleep
Breastfeeding														
Ahmed et al (2016) [45]	United States	141	Mother-newborn dyads	30 days	Hospital discharge, 1, 2, and 3 months post partum	Web-based	Survey	S/NS[e]	—[f]	—	—	—	—	—
Unger et al (2018) [48]	United States	300	Pregnant women (<36 wk pregnant)	Pregnancy (26 wk) to 12 weeks post partum	16 and 24 weeks post partum	SMS text messaging	Survey	S[g]	—	—	—	—	—	—
Wu et al (2020) [51]	China	344	Pregnant women (11 - 37 wk pregnant)	Third month of pregnancy to 6 months post partum	BL[h], 0 - 1 month, 2 - 3 months, and 4 - 5 months post partum	WeChat (Tencent)	Interview	S/NS	S/NS[i]	—	—	—	—	—
Lewkowitz et al (2020) [69]	United States	170	Low-income first-time mothers (~36 wk pregnant)	Pregnancy (36 wk) to 6 weeks post partum	BL, postpartum day 2, 6 weeks, 3 and 6 months	App	Survey	NS[j]	—	—	—	—	—	—
Scott et al (2021) [53]	Australia	1426	Fathers	From recruitment in pregnancy until 6 months post partum	BL, 6 - and 26 weeks post partum	App	Survey	NS	NS[i]	—	—	—	—	—
Saucedo Baza et al (2022) [65][l]	United States	36	Pregnant women (beyond 37 wk pregnant)	6 weeks	BL, 4 - 6 wk post partum	App	Survey	NS	—	—	—	—	—	—

XSL•FO
RenderX

Author (date)	Country	N	Target group	Duration[a]	Follow up	Intervention type	Assessment	Breastfeeding	Feeding practices	Diet	PA[b]	SB[c]	ST[d]	Sleep
Doan et al (2022) [55]	Vietnam	568	Mothers who delivered by a cesarian section	Pregnancy to 4 months post partum	BL, 1-, 4- and 6-months	App	Interview	S/NS	—	—	—	—	—	—
LeFevre et al (2022) [56]	India	5095	Hindi speaking women (4 - 7 mo pregnant)	Pregnancy (gestational wk 12 - 34) to 12 months post partum	BL, 12 months	Voice calls	Survey	NS	—	—	—	—	—	—
Acar and Sahin (2023) [58]	Turkey	73	Primiparous mothers	8 weeks (First day to 8 wk post partum)	BL, 4, and 8 weeks	App	Survey	S[g]	—	—	—	—	—	—
Hmone et al (2023) [59]	Myanmar	353	Pregnant women (28 and 34 wk pregnant)	6 months (from gestational wk 28 - 34 to 6 months of age)	BL, 1 - 6 months	SMS text messaging	Interview	S	S	—	—	—	—	—
Vila-Candel et al (2024) [62]	Spain	270	Pregnant women (third trimester)	6 months (third trimester to 6 months of age)	Hospital discharge after delivery, 15 days, 6 weeks, 3, and 6 months	App	Survey	NS	—	—	—	—	—	—
Henshaw et al (2024) [76]	United States	128	Mother and partner	6 weeks	BL, 6 weeks, 6 months	Tablet-based program	Survey	NS	S/NS	—	—	—	—	—
De Mello et al (2025) [70]	United States	36	Pregnant women (32 - 36 wk gestation)	Antenatal to 12 months after birth	BL (32 - 36 wk gestation), 12 months	App	Survey	NS	NS	—	—	—	—	—

XSL•FO
RenderX

Author (date)	Country	N	Target group	Duration[a]	Follow up	Intervention type	Assessment	Breastfeeding	Feeding practices	Diet	PA[b]	SB[c]	ST[d]	Sleep
Brown et al (2025) [80]	Australia	5783	Mothers 1 - 4 wk post partum	24 months	BL, 3 weeks, 3 months, 6 months, 9 months, 1 year, 18 months	SMS text messaging	Survey	NS	NS	—	—	—	—	—
Cherie et al (2025) [79]	Ethiopia	743	Pregnant women (26 - 28 wk gestation)	Antenatal to 42 days post partum	BL, 60 days	SMS text messaging	Survey	U[k]	U	—	—	—	—	—
Gilano et al (2025) [77]	Ethopia	675	Pregnant women (16 - 28 wk gestation)	Antenatal to 6 months after birth	BL, 1 month, 6 months	SMS text messaging	Interview	S	S	—	—	—	—	—
Breastfeeding and feeding practices														
Palacios et al (2018) [67][l]	United States	202	Caregivers of infants 0 - 2 months old participating in the WIC program, Intervention: infant age 0.93 (SD 0.44) months; Control: infant age 0.98 (SD 0.47) months	4 months	BL, 4 months	SMS text messaging	Survey	NS	NS	—	—	—	—	—

XSL•FO
RenderX

Author (date)	Country	N	Target group	Duration[a]	Follow up	Intervention type	Assessment	Breastfeeding	Feeding practices	Diet	PA[b]	SB[c]	ST[d]	Sleep
Davis et al (2023) [64][i]	United States	38	Parent of an infant aged 3 - 30 days	12 months	BL (0 - 2 wk), 2 - 4 months, 6 - 9 mo, and 12 months	SMS text messaging	Survey	NS	NS	—	—	—	—	—
Li et al (2024) [72][m]	China	1332	Caregivers of children 0 - 3 years, Intervention: 8.6 (SD 7.2) months; Control: 9.0 (SD 6.8) months	9 months	BL and 9 months	WeChat	Survey	S/NS	S/NS	NS	—	—	—	—
Combined[n]														
Wen et al (2020) [52]	Australia	1155	Pregnant women (24 - 34 wk pregnant)	Antenatal to 10 months after birth	BL, 6, and 12 months of child age	SMS text messaging	Survey	NS	S	—	NS	—	S	—
Wu et al (2023) [61]	China	1610	Infants and young children aged 6 - 20 months old and their primary caregivers, 36% aged 6 - 11 months, 64% aged 12 - 20 months	2 months	BL, 1 month, and 2 months	WeChat	Survey	NS	S	S	S	NS	S	NS
Diet														

XSL•FO
RenderX

Author (date)	Country	N	Target group	Duration[a]	Follow up	Intervention type	Assessment	Breastfeeding	Feeding practices	Diet	PA[b]	SB[c]	ST[d]	Sleep
Røed et al (2021) [54]	Norway	291	Parents of infants and toddlers, mean age 10.9 (1.2) months old	6 months	BL and 6 months	Website	Survey	—	—	S/NS	—	—	—	—
Helle et al (2019) [50]	Norway	715	Mother of a 3 - 5 month-old infant	12 months	BL (child age: 5 months), 12 months (intervention completion), and one year after the intervention (child age: 24 months)	Website	Survey	—	S/NS	NS	—	—	—	—
Sleep														
Moon et al (2017) [63]	United States	1600	Mothers of infants, mean age 11.2 (SD 4.4) weeks	60 days	BL and infant age 60 days	Email or SMS text messaging	Survey	—	—	—	—	—	—	S

[a]Due to the nature of recruitment in pregnancy, the duration of the intervention was not always clear.

[b]PA: physical activity.

[c]SB: sedentary behavior.

[d]ST: screen time.

[e]S/NS: some significant and some nonsignificant results.

[f]Not applicable,

[g]S: significant.

[h]BL: baseline.

[i]These interventions focused only on breastfeeding, but other infant feeding practice outcomes (eg, introduction to formula and complementary foods, giving dairy or dairy products, water, semisolid, or solid foods) were also assessed.

[j]NS: not significant.

[k]U: percentages reported for each group but insufficient information to determine statistical significance.

[l]These studies were pilot randomized controlled trials (RCT) or feasibility RCT.

[m]This study enrolled children up to 3 years, but the average age in both arms fell within infancy; the study is therefore reported under the pregnancy-infancy category.

[n]The combined interventions focused on multiple behaviors eg, diet and physical activity/sedentary behavior.

XSL•FO
RenderX

Table . Study characteristics and effectiveness of autonomously delivered digital interventions focusing on toddlers (12 - 35 months; ≥1 to -<3 years): randomized controlled trials published between 2011 - 2026 and across multiple countries (n=6).

Author (date)	Country	N	Target group	Duration	Follow-up	Intervention type	Assessment details	Feeding practices	Diet	PA[a]	SB[b]	ST[c]	Sleep
Combined[d]													
Alexandrou et al (2023) [60][e]	Sweden	552	Parents with a 2.5 - 3 y-old child, 2.5 years (n=403), 3 years (n=149)	6 months	BL[f] and 6 months	App	Survey	—	S[g]	NS[h]	-	S	—
Sandborg et al (2025) [71]	Australia	1165	Parents with a child 18 - 35 months (mean age 27.1 [SD 4.1] months)	12 months	BL and 6 months	App +SMS text messaging	Survey	—	—	S	—	NS	S
Jongpaiboonpatana et al (2025) [73]	Thailand	112	Parents of children aged 12 - 36 months; Intervention 23.3 (SD 6.2) months; Control: 22.5 (SD 7.0) months	6 weeks	BL, 6, and 10 weeks	Online videos	Survey + interview	—	—	S	—	NS	—
Diet													
Cunningham et al (2026) [78]	Nepal	2537	Families with children aged 12 - 23 months	1000 days	BL and 1000 days	SMS text messaging	Survey	—	NS	—	—	—	—
Hunsrisakhun et al (2025) [74]	Thailand	303	Parents of a child 6 - 42 months; Group 1: 23.4 (SD 9.9) months; Group 2: 24.0 (SD 10.6) months	6 months	BL, 3, and 6 months	Facebook Messenger Chatbot (Meta)	Survey	S/NS[i]	NS	—	—	—	—
Sleep													

XSL•FO
RenderX

Author (date)	Country	N	Target group	Duration	Follow-up	Intervention type	Assessment details	Feeding practices	Diet	PA[a]	SB[b]	ST[c]	Sleep
Mindell et al (2011) [68]	United States	264	Mothers and their infant or toddler (ages 6 - 36 months, mean age 19.4 [SD 8.9] months)	2 weeks	Days 8 (BL), 15, and 22	Website +emails	Survey	—	—	—	—	—	S

[a]PA: physical activity.

[b]SB: sedentary behavior.

[c]ST: screen time.

[d]The combined interventions focused on multiple behaviors eg, diet and physical activity/sedentary behavior.

[e]This study focused on both the toddler period and preschool age.

[f]BL: baseline.

[g]S: significant.

[h]NS: not significant.

[i]S/NS: some significant and some nonsignificant results.

Risk of Bias in Studies

The susceptibility to bias of the included studies is presented in Table 2. Briefly, all studies measured outcomes the same way for groups and had an appropriate study design (n=38) [44-81]. Most studies used appropriate statistical analyses (n=37) [44-67,69-81], had true randomization (n=36) [44-46,48-67,69-81], had complete follow-up (or, if not, adequately described and analyzed differences between groups in terms of their follow-up; n=31) [44-64,66,67,69,73-79], used intention-to-treat analysis (n=30) [44-62,64,65,67,69,71-73,77-80], and treated groups identically (other than the intervention of interest; n=33) [45-48,50-56,58-60,63-81]. Findings were mixed for other risk of bias items; in particular, concealment of group allocation (yes: n=17 [46,48-51,55,56,59,62,64,66,69,71,74,75,78,80]; unclear: n=18 [44,45,47,52,54,57,58,63,65,67,68,70,72,73,76,77,79,81]; no: n=3 [53,60,61]), similarity of groups at baseline (yes=25 [44-47,49,50,52-58,60,65,67,69,70,73-75,77-80]; no=11 [48,51,59,61-64,66,68,71,72]; unclear=2 [76,81]), blinding of participants (yes=6 [44,49,55,68,69,80]; no=27 [45-48,50,52-54,56-61,64,65,70-79,81]; unclear=5 [51,62,63,66,67]), blinding of outcome assessors (yes=22 [44,46-49,51,52,55-57,59,60,62,64,68,69,73-75,77-79]; no=3 [61,66,71]; unclear=13 [45,50,53,54,58,63,65,67,70,72,76,80,81]), reliability of outcome measures (yes=28 [44-50,52-55,57,60,65-69,71,72,74-81]; unclear=10 [51,56,58,59,61-64,70-73]). The item assessing the blinding of treatment deliverers was not applicable for many studies, as the digital interventions under investigation were all autonomously delivered.

Results of Synthesis

Intervention Characteristics

Key intervention characteristics of the included studies spanning pregnancy to infancy, toddlers, and preschoolers are presented in Tables 2-4, respectively. Additional details on the intervention and control conditions for each age group are available (in Tables S2-S4 in Multimedia Appendix 2). Overall, intervention duration ranged from 2 weeks [68] to 1000 days [78]. The included studies used various delivery channels of digital technologies for the intervention. The most common digital tools used were apps (11/38, 29%) [46,53,55,57,58,60,62,65,69,70,75], followed by SMS text messaging (10/38, 26%) [48,52,59,64,67,77-81], and web- or internet-based (6/38, 16%) [44,45,49,50,54,66]. Other digital tools used included WeChat (Tencent; 3/38, 8%) [51,61,72], online videos (1/38, 3%) [73], a combination of app and SMS text messaging (1/38, 3%) [71], websites and emails (1/38, 3%) [68], or email and SMS text messaging (1/38, 3%) [63], a tablet-based program (2/38, 5%) [47,76], automated voice calls (1/38, 3%) [56], and Facebook Messenger Chatbot (Meta; 1/38, 3%) [74].

Assessment of Target Outcomes

As shown in Tables 2-4, a broad range of methods and definitions of outcomes were used, with varying numbers as well as timing of the follow-up measures. Outcomes were mainly assessed by parent report using a range of different questionnaires, except for 5 studies where objective methods were used. Four studies [46,49,57,81] used accelerometry to assess physical activity and sedentary behavior, and one study [66] used reflective spectroscopy to assess dietary intake (fruit and vegetable consumption).

XSL•FO
RenderX

Intervention Effectiveness

Effectiveness of Studies Focusing on Pregnancy to Infancy

The level of effectiveness of the studies spanning pregnancy to infancy is presented in Table 3 (overview) and Table S5 in Multimedia Appendix 2 (details).

Among the studies that targeted breastfeeding, most reported no difference in breastfeeding outcomes between the groups [53,56,62,65,69,70,80], or mixed results [45,51,55,72,76], while one study [79] reported breastfeeding outcomes as percentages only, with insufficient information to determine statistical significance. Of the interventions that showed an effect on one or more of the breastfeeding outcomes [45,48,51,55,58,59,76,77], most targeted mothers already in pregnancy, ranging from gestational week 11-37 [48,51,55,59,77] with the remainder commencing post-birth [45,58,76] (Table 3 and Table S5 in Multimedia Appendix 2). The duration of the successful interventions ranged between 30 days and up to 6 months post partum, with a variety in digital delivery modes including an app [55,58], SMS text messaging [48,59,77], web-based [45], tablet-based [76], and WeChat [51] (Table S2 in Multimedia Appendix 2).

Three of the studies also reported results for the effectiveness of the intervention on feeding practices, showing a significant reduction in bottle-feeding at 6 months in the intervention group [59] but no difference in the introduction of formula and complementary foods between the groups [53]. One study reported mixed results with a significantly lower rate of giving dairy products to the child 0 - 1 months post partum in the intervention groups but no difference in giving semisolid or solid foods at 0 - 1 month, 2 - 3 months, and 4 - 5 months [51]. Three studies targeted both breastfeeding and feeding practices, with most reporting nonsignificant [64,67] or mixed results [72]. In addition, the studies targeting a combination of behaviors [52,61] reported significant improvements in feeding-related outcomes (eg, appropriate timing of solid foods and diet diversity).

The 2 studies targeting diet during infancy reported mixed findings [50,54]. In more detail, Røed et al [54] reported a larger increase in the frequency of vegetable intake in the intervention group compared to the control group but no difference in child food intake of fruits, vegetables, and discretionary foods between baseline and 6 months. In contrast, Helle et al [50] reported nonsignificant findings for all dietary intake variables, child mealtime habits, frequency of family meals, maternal feeding practices, and maternal feeding style at 12 months. However, impact was reported for other feeding practices, with significantly higher food responsiveness and lower emotional overeating in the intervention group compared to the control [50]. Both interventions were delivered via a website and included intervention content in video as well as recipes. Røed et al [54] also included modules with lessons, activity elements (eg, quizzes), a discussion forum, and weekly emails was half as long as that of Helle et al [50] (6 months vs 12 months), and targeted older infants (mean age 11 months vs 3 - 5 months old).

Only one study targeted sleep in the infancy period and reported significant improvements for infant sleep practices in the intervention group [63]. Sleep practices included a higher prevalence of placing their infant in a supine position, room sharing without bed sharing, no soft bedding use, and any pacifier use. The intervention was 60 days and included health messages and educational videos delivered by email or SMS text messages.

Effectiveness of Studies Focusing on Toddlers

The effectiveness of studies focusing on toddlers is presented in Table 4 (overview) and Table S6 in Multimedia Appendix 2 (details). Among the interventions targeting multiple behaviors, one reported significant improvements in diet (higher vegetable intake and lower intakes of sweet and savory treats and sweet drinks) and less screen time but no differences in physical activity [60]. The other studies reported significant improvements in parental knowledge of child movement behaviors and sleep [71] and increases in parent-child play frequency [73], but no significant improvements related to screen time [71,73].

The 2 studies targeting diet also reported mixed findings [74,78]. One intervention did not demonstrate overall improvements in dietary diversity [78]. The other intervention reported several significant improvements in feeding-related behaviors (eg, reduced bottle use and fewer night feedings) but also nonsignificant outcomes (eg, consuming sweet foods) [74].

The other study in this age group targeted sleep only and reported improved sleep (decreased sleep onset latency, decreased number/duration of night wakings, increased sleep continuity, and increased nighttime sleep) in the intervention group compared to the control group [68]. The intervention duration and digital delivery mode varied; one was 6 months in duration and delivered using an app [60], while the sleep intervention was 2 weeks and delivered via a website and emails [68] (Table 4 and Table S3 in Multimedia Appendix 2).

Effectiveness of Studies Focusing on Preschoolers

The effectiveness of the studies focusing on preschoolers is presented in Table 5 (overview) and Table S7 in Multimedia Appendix 2; details). Among the studies targeting a combination of behaviors, most reported significant improvements for feeding practices (2/2, 100%) [47,49] and diet (3/4, 75%) [44,46,49], for example, lower intakes of sweetened beverages [44,46] and discretionary foods [49], and higher intakes of fruit and vegetables [44], while one study reported null effects for child eating style and eating related to hunger [47]. All studies targeting a combination of behaviors reported null effects for physical activity (3/3, 100%) [46,47,49], sedentary behavior (2/2, 100%) [44,46], and sleep (1/1, 100%) [49]. The 2 studies targeting screen time reported both positive [44] and null results [49]. Similarly, the studies targeting physical activity reported no overall effects on physical activity or sedentary behavior [57,81], except for context-specific effects (eg, presence of a parent, weather-dependent) and increased parental moderate-to-vigorous physical activity [81].

XSL•FO
RenderX

Table . Study characteristics and effectiveness of autonomously delivered digital interventions focusing on the preschool age (36 - 59 months; ≥3 to -<5 y): randomized controlled trials published between 2011 and 2026 and across multiple countries (n=8).

Author (date)	Country	N	Target group	Duration	Follow up	Digital tool	Assessment details	Feeding practices	Diet	PA[a]	SB[b]	ST[c]	Sleep
Combined[d]													
Knowlden et al (2015) [44]	United States	57	Mothers of 4 - 6-year-old children	8 weeks	BL[e], 4, and 8 weeks	Website	Survey	—[f]	S	—	NS[g]	S[h]	—
Delisle-Nyström et al (2017) [46]	Sweden	315	Parents of children aged 4 years, Intervention: 4.5 (SD 0.1) years, 45% female, maternal age 36.0 (SD 4.1) years; Control: 4.5 (SD 0.1) years	6 months	BL, 6 months	App	Direct observation (photos) and accelerometry	—	S	NS	NS	—	—
Sun et al. (2017) [47][i]	United States	32	Chinese mothers and their children 3 - 5-year-old child, mean age 4.3 (SD 0.7) years	8 weeks	BL, 3, and 6 months	Tablet-based program	Survey	S	NS	NS	—	—	—
Hammersley et al (2019) [49]	Australia	86	Parents of 2 - 5 y-old children, mean age 3.5 (SD 0.9) years	11 weeks	BL, 3, and 6 months	Website	Survey and accelerometry	S	S	NS	—	NS	NS
Diet													

XSL•FO
RenderX

Author (date)	Country	N	Target group	Duration	Follow up	Digital tool	Assessment details	Feeding practices	Diet	PA[a]	SB[b]	ST[c]	Sleep
Bakirci-Taylor et al. (2019) [66][i]	United States	30	Parents and children aged 3 - 8 years, Intervention: child age 3.8 (SD 0.8) years; Control: child age 3.6 (SD 1.4) years	10 weeks	BL, 5- and 10 weeks	Website	Survey, direct observation (photos) and reflective spectroscopy	—	S/NS[j]	—	—	—	—
Hojati et al (2024) [75]	Iran	116	Mothers and children aged 2 - 6 y with confirmed undernutrition, Intervention: 50.3 (SD 11.6) months; Control: 46.51 (SD 13.3) months	3 months	BL, 3 months	App	Survey	S	—	—	—	—	—
Physical activity													
Staiano et al (2022) [57]	United States	72	Parents and 3 - 5-year-old children, mean age 4.0 (SD 0.8) years	12 weeks	BL, 12 weeks (end of intervention), and 24-week (follow-up)	App	Direct observation and accelerometry	—	—	NS	NS	—	—

XSL•FO
RenderX

Author (date)	Country	N	Target group	Duration	Follow up	Digital tool	Assessment details	Feeding practices	Diet	PA[a]	SB[b]	ST[c]	Sleep
Phillips et al (2026) [81]	Canada	18	Parents and children aged 3 - 4 years	2 weeks	2 weeks; 7 prompts/day; 60-min post-prompt outcome windows	SMS text messaging	Accelerometry	—	—	S/NS	NS	—	—

[a]PA: physical activity.

[b]SB: sedentary behavior.

[c]ST: screen time.

[d]The combined interventions focused on multiple behaviors eg, diet and physical activity/sedentary behavior.

[e]BL: baseline

[f]Not applicable.

[g]NS: not significant.

[h]S: significant.

[i]These studies were pilot randomized controlled trials.

[j]S/NS: some significant and some nonsignificant results.

The studies targeting diet only reported higher vegetable intake in the intervention group but no difference in the frequency of fruit and vegetable consumption [66] and significant improvements in maternal nutrition knowledge, feeding attitudes, and nutrition practices [75].

The interventions with an effect on diet and/or feeding practices ranged from 8 weeks to 6 months, with the majority being 8 - 11 weeks [44,47,49,66,75], and the digital delivery mode included using a website or web-based intervention [44,49,66], an app [46,75] and a tablet-based program [47] (Table 5 and Table S4 in Multimedia Appendix 2). The study that had a significant intervention effect on screen time targeted 4 - 6-year-olds (n=57) and delivered a web-based intervention over 8 weeks [44]. Hammersley et al [49] evaluated the effectiveness of a website but found no effect on screen time. Their intervention duration was longer (11 wk), and the participating children were younger (mean age ~3.5, SD 0.9 years; n=86).

Intervention Development and Engagement

Co-design, intervention theory, process evaluation, and engagement outcomes for studies focusing on pregnancy to infancy, toddlers, and preschoolers are shown in Tables S8, S9, and S10 in Multimedia Appendix 2, respectively.

Co-Design

Of the 38 interventions, 24 (63%) [45-49,51,53-56,58-61,63,64,67,69,72,74,75,78-80] reported some form of co-design or end-user engagement in the development of the intervention. The extent of co-design or end-user engagement ranged from end users' (ie, parents') views on delivery platforms [45] and pilot testing content with end users (eg, [46]) to intensive qualitative work involving a range of stakeholders to inform all aspects of the intervention [63]. Of the 24 interventions reporting some form of co-design, 16 (67%) showed some evidence of effectiveness. By comparison, 10 of the 14 interventions (71%) that did not report any form of co-design showed some evidence of effectiveness.

Intervention Theory

Twenty-seven of the 38 interventions (71%) [44-47,49,50,52-55,57,59,60,64-67,70-74,76-78,80,81] reported being underpinned by theory. The most commonly used approaches were social-cognitive and learning-based theories, including Social Cognitive Theory [44-46,49,50,53,54,57,60,65,66], the Health Belief Model [52,59,64], social learning theory [73], Protection Motivation Theory [74], and behavior-change communication frameworks [77]. Motivational and self-regulatory models were also used, such as Self-Determination Theory [81], the TransTheoretical model of health behavior change [67], the Information Motivation Behavioral Skills model [47], and breastfeeding self-efficacy frameworks [70,76]. Several interventions used comprehensive design frameworks, including the Behavior Change Wheel [71,80], the Theoretical Domains Framework [80], and the World Health Organization (WHO) comprehensive health literacy model [72], while others used technology-oriented or implementation-focused models such as the Behavioral Intervention Technology Model [55], AI-supported behavior-change models [74], and a formal theory of change [78]. Of the 27 interventions that reported being underpinned by theory, 19 (70%) showed some evidence of effectiveness. By comparison, 6 of the 11 (55%) interventions that did not report any use of theory to underpin the intervention showed some evidence of effect.

Process Evaluation

Process evaluation results were reported for 24 studies (63%). These mostly focused on subjective reports of acceptability, satisfaction, and usefulness (n=17) [44,45,47,49,53,57,58,60,61,65,66,68-70,74,80,81] and/or objective measures of delivery/use (eg, successful delivery of

XSL•FO
RenderX

SMS text messaging, usage data for websites/apps; n=9) [46,48,51-54,64,67,69]. Interventions using mobile apps were most consistently found to be acceptable and useful; apps were predominantly rated highly for ease of use, design, and helpfulness [46,53,57,58,69]. Although SMS text messaging interventions were generally found to be acceptable and useful, some studies reported declining engagement over time [52] or technical issues (eg, issues with SMS text messaging delivery [67] or phone service [64]). Web-based interventions showed mixed results in terms of acceptability and usage, with studies generally finding lower engagement compared to apps; for example, Røed et al [54] reported that 13% of participants never used the website, while Bakirci-Taylor et al [66] reported that participants engaged more with the Facebook page than the mobile website. Other interventions using Facebook reported less positive results; for example, Hammersley et al [49] reported that only 39% of participants found the Facebook component useful. In contrast, the Facebook Messenger Chatbot intervention reported high satisfaction (mean score 4.0 out of 5; SD 0.5-0.6 across groups) [74].

Engagement

Engagement outcomes were reported by 24 studies (63%); n=7 [50,57,58,61,65,70,78] used subjective measures (eg, self-reported use), n=13 [46,48,52,54,56,62,63,66,67,69,71,72,74] used objective measures (eg, app analytics), and n=4 [49,53,60,80] used both. Of these, 6 studies examined the impact of engagement on intervention effectiveness, with mixed findings across behaviors. All breastfeeding-focused interventions reported no impact on the outcomes [56,62,69], whereas studies targeting breastfeeding and feeding practices [72], diet [78], and movement behaviors [71] found that higher engagement (eg, more videos viewed, messages opened, or greater app use) was associated with more favorable outcomes.

Discussion

Principal Findings

This systematic review provides a comprehensive overview of autonomously delivered digital interventions targeting multiple behaviors in the first 2000 days and examines co-design, engagement, and process evaluation, areas rarely assessed in previous reviews. Thirty-eight interventions were included in the review, with most focusing on improving breastfeeding practices by targeting mothers and their youngest children (newborns and infants) and a growing number of studies targeting toddlers and preschoolers. Overall, intervention designs varied considerably, and results were mixed for all targeted age groups with no apparent trend in intervention characteristics (eg, target behavior and digital delivery mode) for interventions shown to be effective. Although most studies reported some form of co-design or end-user engagement, very few examined the impact of engagement on the efficacy of the intervention.

The variability in intervention design is not unique for autonomously delivered digital interventions, and previous reviews on digital interventions for improving breastfeeding have similarly reported heterogeneity; for example, in delivery modes [82-84]. In terms of effectiveness, most of the included studies in our review reported no differences in breastfeeding outcomes between the groups and reported mixed results. These findings are similar to previous systematic reviews and meta-analyses focusing solely on mobile apps [83], remote provision of breastfeeding support education (eg, telephone, SMS text messaging, social media, video call, and email) [82], or mHealth-based interventions to promote breastfeeding [84]. To illustrate, Ziebart et al [83] found insufficient evidence for sustained beneficial effects of breastfeeding promotion and support through mobile apps on breastfeeding rates, while Gavine et al [82] concluded that remote interventions can be effective for improving exclusive breastfeeding at 3 months but with low certainty of evidence due to risk of bias, substantial heterogeneity, and imprecision in some outcomes. In contrast, Qian et al [84] reported, amongst other things, improvements in exclusive breastfeeding rates up to 6 months after delivery in comparison with usual care. These previous reviews included only up to 4 of the studies captured in our review, likely due to differences in search timeframes and inclusion/exclusion criteria. For example, Ziebart et al [83] focused exclusively on mobile apps related to breastfeeding and excluded web-based interventions; Gavine et al [82] limited their search to studies published after 2010 and focused on remote care broadly; and Qian et al [84] included a wider range of mHealth modalities (eg, phone calls, SMS text messaging, and interactive systems) but only considered breastfeeding outcomes. In contrast, our review included RCTs targeting a broader set of health behaviors (breastfeeding, feeding practices, diet, physical activity, sedentary behavior, and sleep) in children aged 0 - 5 years, delivered solely via autonomous digital technologies. Thus, our review provides an important addition to the existing evidence.

Moreover, we report the effectiveness of interventions covering the first 5 years of life, which is recognized as a critical period for establishment of health behaviors. The studies targeting toddlers showed promising results for diet, screen time, and sleep but not physical activity. However, there were only 6 [60,68,71,73,74,78] studies in this age group, highlighting the need for more research targeting toddlers. Similarly, the number of studies targeting preschoolers was limited, and results indicated significant improvements for feeding practices and diet, mixed findings for screen time, and no overall differences for children's physical activity, sedentary behavior, and sleep, although one trial reported context-specific reductions in children's sedentary time and increased parental moderate-to-vigorous physical activity. Previous systematic reviews on digital interventions to promote these health behaviors in preschoolers have also reported mixed findings [25,26]. For example, Zhou et al [26] reported significant improvements for dietary behaviors and sleep but no significant improvements in physical activity for parent-based eHealth interventions. In contrast, a systematic review and meta-analysis on the effectiveness of eHealth interventions for promoting 24-hour movement behaviors in preschoolers [25] reported small but positive effects on physical activity, sedentary time, and sleep. One potential explanation for the discrepancies in results could be differences in intervention delivery, as most of the included studies in the meta-analysis relied on human interaction (eg, courses with staff members, motivational coaching, individual discussion/telephone calls, face-to-face

XSL•FO
RenderX

workshop, and home visits) [25,26]. Although autonomously delivered interventions have greater potential for scalability and potential cost-effectiveness, digital interventions including a delivery personnel component might be more effective; however, considering the paucity of studies in this age group, more research is required to determine this.

Reporting of elements related to the design and evaluation of interventions, including co-design and process evaluation, is important considering that these can help improve intervention effectiveness [85] and identify key components that contribute to the success of interventions [86], respectively. Another important aspect is participant engagement, as it can provide information on levels of intervention exposure and uptake, which are crucial for effectiveness, as well as help identify factors that may optimize engagement [33] and ultimately inform the development of more impactful interventions. Although most of the included studies reported some form of co-design or end-user engagement, only 6 studies [56,62,69,71,72,78] examined the impact of engagement on intervention effectiveness. In contrast with previous findings in other populations (eg, [87,88]) and outcomes such as parenting practices and cognitions [71], the studies included in this review reported mixed findings across behaviors. All breastfeeding-focused interventions reported that higher engagement was not associated with intervention effectiveness on the targeted outcomes (ie, breastfeeding) [56,62,69], while studies targeting breastfeeding and feeding practices [72], diet [78], and movement behaviors [71] found that higher engagement was associated with more favorable outcomes. Nevertheless, the low reporting rate of engagement suggests that the results from this review about the impact may not fully reflect the intended interventions. Although engagement data are supposedly easy to collect in digital interventions (eg, apps), previous research has highlighted a gap in evaluating and reporting how engagement influences intervention effectiveness [89]. This raises critical questions, such as why engagement data are often omitted, and whether low engagement levels could be a contributing factor, with researchers hesitant to report underwhelming results. Despite the assumption that digital interventions increase reach, this may not translate to increased engagement [34]. Ultimately, understanding whether and how engagement mediates outcomes is essential to determine the true value of these interventions, and addressing these gaps is crucial for ensuring that scalable interventions are also impactful and meaningfully engaged with by users. This review highlights key considerations for improving future research, including the evaluation and reporting of the impact of engagement on the effectiveness of digital interventions for promoting health behaviors in early childhood.

Strengths and Limitations

This review has several strengths and limitations. A key strength was the systematic approach used to search, screen, and synthesize the literature, including the PROSPERO registration of the review protocol and the use of the JBI Critical Appraisal Checklist for RCTs. Moreover, we focused on digital interventions that were delivered autonomously and thus, in theory, have the capacity to be scaled up and delivered at large without heavy researcher or staff input. Another strength is that we also considered intervention development (co-design and intervention theory) and participant engagement, which are important elements for intervention effectiveness [85]; however, this broader evaluation was constrained by the extent of reporting within individual studies. The present review also has limitations to acknowledge. First, despite comprehensive searches across multiple databases, it is possible that relevant studies were missed, particularly unpublished or non-English studies. Second, we restricted the review to autonomously delivered digital interventions, which limit our findings to hybrid models that include human support. Third, considerable heterogeneity in study populations, intervention content, delivery formats, and outcome measures prevented and limited our ability to perform a meta-analysis. The wide variety in intervention objectives, settings, methodologies, and delivery modes also made it difficult to compare findings across studies. Finally, digital health interventions evolve rapidly, and more recent innovations may not yet be represented in the current evidence base, as highlighted by recent evidence underscoring the challenges of synthesizing findings in this fast-moving field [22].

In terms of strengths and limitations for the individual studies, although most studies had an appropriate study design and use of statistical methods, findings were mixed for other risk of bias items, including concealment of group allocation, blinding of participants, and outcome assessors, as well as reliability of outcome measures. Considering the nature of the interventions, blinding of participants and researchers might not be feasible, suggesting that checklists assessing study quality specific to digital interventions are warranted. Most studies also used subjective methods to assess outcomes, which are inherently subject to misreporting biases. Finally, most studies were conducted in high-income countries, which limits generalizability to low-income countries.

Conclusion

This review shows that autonomously delivered digital interventions for early childhood are highly heterogeneous and demonstrate mixed effectiveness, making it difficult to identify which components are most impactful. It is innovative in synthesizing evidence across the first 2000 days while simultaneously examining co-design, engagement, and implementation factors, dimensions rarely brought together in previous work. Unlike earlier reviews that focus on older children, single behaviors, or interventions involving human support, this review focuses solely on scalable autonomously delivered digital interventions for children 0‑5 years, a formative period for long-term health. Most importantly, it adds new insight by identifying three priority evidence gaps: (1) the scarcity of studies targeting toddlers and preschoolers, (2) inconsistent and incomplete reporting of engagement, and (3) limited understanding of how engagement influences outcomes. While autonomous digital interventions offer clear advantages in reach and scalability, their usefulness ultimately depends on whether interventions remain engaging, relevant, and effective for families. Together, these findings define priority areas for future research and clarify what is needed to strengthen the evidence base for scalable digital interventions in early childhood.

XSL•FO
RenderX

Acknowledgments

We are grateful to Frances Beard (Health Liaison Librarian, Deakin University) for her support in preparing the search strategy. We sincerely thank Cynthia Smith and Stephanie Renehan for their valuable assistance with supplementary searches, data screening and extraction, and duplicate risk of bias assessments conducted during the updated search and review process. All authors contributed to the writing and critically reviewed and approved the final draft of the submitted manuscript. In addition, KDH is supported by a Heart Foundation Future Leader Fellowship (105929). JS was supported by two post-doctoral fellowships from the Henning and Johan Throne-Holst Foundation and the Erik and Edith Fernström Foundation for Medical Research. The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript. Generative artificial intelligence (AI) tools (Microsoft Copilot) were used only for language refinement (eg, improving clarity and phrasing). AI tools were not used to generate scientific content, interpret results, extract data beyond what is described in the Methods, conduct analyses, or create references. All AI-assisted text was reviewed, edited, or rewritten by the authors, who take full responsibility for the final manuscript.

Funding

No review-specific funding was received. KDH and JS are supported by independent fellowships.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files. The full list of included studies and the data extracted from them are available in Multimedia Appendix 2.

Conflicts of Interest

JS is a founder of Science4Families AB, but this company had no role in this research. No other disclosures were reported.

Multimedia Appendix 1
Full search strategy for all databases included in this systematic review, detailing search terms, Boolean operators, filters, and date limits used across Embase, Academic Search Complete, CINAHL Complete, Global Health, MEDLINE Complete, PsycINFO, and SPORTDiscus during searches conducted in December 2022, August 2024, and January 2026.
[PDF File, 191 KB - jmir_v28i1e85525_app1.pdf]

Multimedia Appendix 2
Detailed characteristics of included studies.
[PDF File, 494 KB - jmir_v28i1e85525_app2.pdf]

Checklist 1
PRISMA-S checklist.
[PDF File, 145 KB - jmir_v28i1e85525_app3.pdf]

References

1. North K, Gao M, Allen G, Lee AC. Breastfeeding in a global context: epidemiology, impact, and future directions. Clin Ther 2022 Feb;44(2):228-244. [doi: 10.1016/j.clinthera.2021.11.017] [Medline: 34973827]
2. Tapia-Serrano MA, Sevil-Serrano J, Sánchez-Miguel PA, López-Gil JF, Tremblay MS, García-Hermoso A. Prevalence of meeting 24-Hour Movement Guidelines from pre-school to adolescence: a systematic review and meta-analysis including 387,437 participants and 23 countries. J Sport Health Sci 2022 Jul;11(4):427-437. [doi: 10.1016/j.jshs.2022.01.005] [Medline: 35066216]
3. McArthur BA, Volkova V, Tomopoulos S, Madigan S. Global prevalence of meeting screen time guidelines among children 5 years and younger: a systematic review and meta-analysis. JAMA Pediatr 2022 Apr 1;176(4):373-383. [doi: 10.1001/jamapediatrics.2021.6386] [Medline: 35157028]
4. Lioret S, Campbell KJ, McNaughton SA, et al. Lifestyle patterns begin in early childhood, persist and are socioeconomically patterned, confirming the importance of early life interventions. Nutrients 2020 Mar 9;12(3):724. [doi: 10.3390/nu12030724] [Medline: 32182889]
5. Craigie AM, Lake AA, Kelly SA, Adamson AJ, Mathers JC. Tracking of obesity-related behaviours from childhood to adulthood: a systematic review. Maturitas 2011 Nov;70(3):266-284. [doi: 10.1016/j.maturitas.2011.08.005] [Medline: 21920682]
6. Cattan S, Fitzsimons E, Goodman A, Phimister A, Ploubidis GB, Wertz J. Early childhood inequalities. Oxf Open Econ 2024 Jul 5;3(Supplement_1):i711-i740. [doi: 10.1093/ooec/odad072]

XSL•FO
RenderX

7. Vicat-Blanc L, Merry L, Harguindéguy-Lincourt MC, Tang Y, Van Hulst A. Co-design of interventions and services with structurally marginalized populations in the context of maternal and early childhood primary care: a rapid scoping review. Prim Health Care Res Dev 2025 Jun 16;26:e48. [doi: 10.1017/S146342362510011X] [Medline: 40518975]
8. Gautam N, Dessie G, Rahman MM, Khanam R. Socioeconomic status and health behavior in children and adolescents: a systematic literature review. Front Public Health 2023;11:1228632. [doi: 10.3389/fpubh.2023.1228632] [Medline: 37915814]
9. Brushe ME, Lynch JW, Melhuish E, Reilly S, Mittinty MN, Brinkman SA. Objectively measured infant and toddler screen time: findings from a prospective study. SSM Popul Health 2023 Jun;22:101395. [doi: 10.1016/j.ssmph.2023.101395] [Medline: 37096246]
10. Nyberg G, Helgadóttir B, Moraeus L, Sipinen JP, Lindroos AK, Fröberg A. A national sample of Swedish young children shows sociodemographic variations in physical activity and screen time. Acta Paediatr 2026 Feb;115(2):298-308. [doi: 10.1111/apa.70321] [Medline: 41039737]
11. Kracht CL, Webster EK, Staiano AE. Sociodemographic differences in young children meeting 24-hour movement guidelines. J Phys Act Health 2019 Oct 1;16(10):908-915. [doi: 10.1123/jpah.2019-0018] [Medline: 31491748]
12. Askie LM, Espinoza D, Martin A, et al. Interventions commenced by early infancy to prevent childhood obesity-the EPOCH collaboration: an individual participant data prospective meta-analysis of four randomized controlled trials. Pediatr Obes 2020 Jun;15(6):e12618. [doi: 10.1111/ijpo.12618] [Medline: 32026653]
13. Blake-Lamb TL, Locks LM, Perkins ME, Woo Baidal JA, Cheng ER, Taveras EM. Interventions for childhood obesity in the first 1,000 days a systematic review. Am J Prev Med 2016 Jun;50(6):780-789. [doi: 10.1016/j.amepre.2015.11.010] [Medline: 26916260]
14. Hodder RK, O'Brien KM, Wyse RJ, et al. Interventions for increasing fruit and vegetable consumption in children aged five years and under. Cochrane Database Syst Rev 2024 Sep 23;9(9):CD008552. [doi: 10.1002/14651858.CD008552.pub8] [Medline: 39312396]
15. Mihrshahi S, Jawad D, Richards L, et al. A review of registered randomized controlled trials for the prevention of obesity in infancy. Int J Environ Res Public Health 2021 Mar 2;18(5):2444. [doi: 10.3390/ijerph18052444] [Medline: 33801485]
16. Grady A, Jackson J, Wolfenden L, Lum M, Yoong SL. Assessing the scalability of healthy eating interventions within the early childhood education and care setting: secondary analysis of a Cochrane systematic review. Public Health Nutr 2023 Dec;26(12):3211-3229. [doi: 10.1017/S1368980023002550] [Medline: 37990443]
17. Liu S, Ma J, Sun M, Zhang C, Gao Y, Xu J. Mapping the landscape of digital health intervention strategies: 25-year synthesis. J Med Internet Res 2025 Jan 13;27(1):e59027. [doi: 10.2196/59027] [Medline: 39804697]
18. Laws RA, Litterbach EKV, Denney-Wilson EA, et al. A comparison of recruitment methods for an mHealth intervention targeting mothers: lessons from the growing healthy program. J Med Internet Res 2016 Sep 15;18(9):e248. [doi: 10.2196/jmir.5691] [Medline: 27634633]
19. Brown V, Tran H, Downing KL, Hesketh KD, Moodie M. A systematic review of economic evaluations of web-based or telephone-delivered interventions for preventing overweight and obesity and/or improving obesity-related behaviors. Obes Rev 2021 Jul;22(7):e13227. [doi: 10.1111/obr.13227] [Medline: 33763956]
20. Fiedler J, Eckert T, Wunsch K, Woll A. Key facets to build up eHealth and mHealth interventions to enhance physical activity, sedentary behavior and nutrition in healthy subjects - an umbrella review. BMC Public Health 2020 Oct 23;20(1):1605. [doi: 10.1186/s12889-020-09700-7] [Medline: 33097013]
21. Alley SJ, Waters KM, Parker F, et al. The effectiveness of digital physical activity interventions in older adults: a systematic umbrella review and meta-meta-analysis. Int J Behav Nutr Phys Act 2024 Dec 18;21(1):144. [doi: 10.1186/s12966-024-01694-4] [Medline: 39696583]
22. McDermott KT, Noake C, Wolff R, et al. Digital interventions to moderate physical inactivity and/or nutrition in young people: a cancer prevention Europe overview of systematic reviews. Front Digit Health 2023;5:1185586. [doi: 10.3389/fdgth.2023.1185586] [Medline: 37534029]
23. Hammersley ML, Jones RA, Okely AD. Parent-focused childhood and adolescent overweight and obesity eHealth interventions: a systematic review and meta-analysis. J Med Internet Res 2016 Jul 21;18(7):e203. [doi: 10.2196/jmir.5893] [Medline: 27443862]
24. Bonvicini L, Pingani I, Venturelli F, et al. Effectiveness of mobile health interventions targeting parents to prevent and treat childhood obesity: systematic review. Prev Med Rep 2022 Oct;29:101940. [doi: 10.1016/j.pmedr.2022.101940] [Medline: 36161123]
25. Jiang S, Ng JYY, Chong KH, Peng B, Ha AS. Effects of eHealth interventions on 24-hour movement behaviors among preschoolers: systematic review and meta-analysis. J Med Internet Res 2024 Feb 21;26(1):e52905. [doi: 10.2196/52905] [Medline: 38381514]
26. Zhou P, Li Y, Lau PW, Yan L, Song H, Shi TL. Effectiveness of parent-based electronic health (eHealth) intervention on physical activity, dietary behaviors, and sleep in preschoolers: a systematic review. J Exerc Sci Fit 2024 Jan;22(1):1-13. [doi: 10.1016/j.jesf.2023.10.004] [Medline: 38021206]
27. Wang JW, Zhu Z, Shuling Z, et al. Effectiveness of mHealth app-based interventions for increasing physical activity and improving physical fitness in children and adolescents: systematic review and meta-analysis. JMIR Mhealth Uhealth 2024 Apr 30;12:e51478. [doi: 10.2196/51478] [Medline: 38687568]

XSL•FO
RenderX

28. Baumann H, Fiedler J, Wunsch K, Woll A, Wollesen B. mHealth interventions to reduce physical inactivity and sedentary behavior in children and adolescents: systematic review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2022 May 11;10(5):e35920. [doi: 10.2196/35920] [Medline: 35544294]
29. He Z, Wu H, Yu F, et al. Effects of smartphone-based interventions on physical activity in children and adolescents: systematic review and meta-analysis. JMIR Mhealth Uhealth 2021 Feb 1;9(2):e22601. [doi: 10.2196/22601] [Medline: 33522980]
30. Singh B, Ahmed M, Staiano AE, et al. Lifestyle eHealth and mHealth interventions for children and adolescents: systematic umbrella review and meta-meta-analysis. J Med Internet Res 2025 Oct 17;27(1):e69065. [doi: 10.2196/69065] [Medline: 41106821]
31. Ishaque S, Ela O, Dowling A, et al. Mobile health interventions for modifying indigenous maternal and child-health related behaviors: systematic review. J Med Internet Res 2025 Apr 30;27:e57019. [doi: 10.2196/57019] [Medline: 40305103]
32. Yardley L, Spring BJ, Riper H, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
33. Schoeppe S, Alley S, Van Lippevelde W, et al. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review. Int J Behav Nutr Phys Act 2016 Dec 7;13(1):127. [doi: 10.1186/s12966-016-0454-y] [Medline: 27927218]
34. Sandborg J, Markides BR, Simmons S, et al. Parental and demographic predictors of engagement in an mHealth intervention: observational study from the let's grow trial. JMIR Mhealth Uhealth 2025 Jul 15;13(1):e60478. [doi: 10.2196/60478] [Medline: 40664374]
35. National Health and Medical Research Council. Infant feeding guidelines. : National Health and Medical Research Council; 2012 URL: https://www.nhmrc.gov.au/about-us/publications/infant-feeding-guidelines-information-health-workers [accessed 2026-06-05]
36. Recommendations for infants, toddlers and preschoolers (birth to 5 years). Australian Government Department of Health, Disability and Ageing. 2021. URL: https://www.health.gov.au/topics/physical-activity-and-exercise/physical-activity-and-exercise-guidelines-for-all-australians/for-infants-toddlers-and-preschoolers-birth-to-5-years [accessed 2025-09-11]
37. Markides B, Marshall S, Laws R, Hesketh K, Downing KL. Systematic review of mHealth interventions targeting parents and caregivers to improve early childhood movement, nutrition and sleep behaviours. PROSPERO. URL: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42022372639 [accessed 2024-12-10]
38. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
39. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
40. Elicit: the AI research assistant. Elicit. URL: https://elicit.com [accessed 2024-12-18]
41. Barker TH, Stone JC, Sears K, et al. The revised JBI critical appraisal tool for the assessment of risk of bias for randomized controlled trials. JBI Evid Synth 2023 Mar 1;21(3):494-506. [doi: 10.11124/JBIES-22-00430] [Medline: 36727247]
42. Deeks JJ, Higgins JP, Altman DG, McKenzie JE, Veroniki AA. Chapter 10: analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions: Cochrane; 2024. URL: https://training.cochrane.org/handbook/current/chapter-10 [accessed 2026-05-18]
43. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ 2020 Jan 16;368:l6890. [doi: 10.1136/bmj.l6890] [Medline: 31948937]
44. Knowlden AP, Sharma M, Cottrell RR, Wilson BRA, Johnson ML. Impact evaluation of enabling mothers to prevent pediatric obesity through web-based education and reciprocal determinism (EMPOWER) randomized control trial. Health Educ Behav 2015 Apr;42(2):171-184. [doi: 10.1177/1090198114547816] [Medline: 25161168]
45. Ahmed AH, Roumani AM, Szucs K, Zhang L, King D. The effect of interactive web-based monitoring on breastfeeding exclusivity, intensity, and duration in healthy, term infants after hospital discharge. J Obstet Gynecol Neonatal Nurs 2016 Mar;45(2):143-154. [doi: 10.1016/j.jogn.2015.12.001]
46. Nyström CD, Sandin S, Henriksson P, et al. Mobile-based intervention intended to stop obesity in preschool-aged children: the MINISTOP randomized controlled trial. Am J Clin Nutr 2017 Jun;105(6):1327-1335. [doi: 10.3945/ajcn.116.150995] [Medline: 28446496]
47. Sun A, Cheng J, Bui Q, Liang Y, Ng T, Chen JL. Home-based and technology-centered childhood obesity prevention for Chinese mothers with preschool-aged children. J Transcult Nurs 2017 Nov;28(6):616-624. [doi: 10.1177/1043659617719139] [Medline: 28826348]
48. Unger JA, Ronen K, Perrier T, et al. Short message service communication improves exclusive breastfeeding and early postpartum contraception in a low- to middle-income country setting: a randomised trial. BJOG 2018 Nov;125(12):1620-1629. [doi: 10.1111/1471-0528.15337] [Medline: 29924912]

XSL•FO
RenderX

49. Hammersley ML, Okely AD, Batterham MJ, Jones RA. An internet-based childhood obesity prevention program (Time2bHealthy) for parents of preschool-aged children: randomized controlled trial. J Med Internet Res 2019 Feb 8;21(2):e11964. [doi: 10.2196/11964] [Medline: 30735139]
50. Helle C, Hillesund ER, Wills AK, Øverby NC. Evaluation of an eHealth intervention aiming to promote healthy food habits from infancy -the Norwegian randomized controlled trial early food for future health. Int J Behav Nutr Phys Act 2019 Jan 3;16(1):1. [doi: 10.1186/s12966-018-0763-4] [Medline: 30606197]
51. Wu Q, Huang Y, Liao Z, van Velthoven MH, Wang W, Zhang Y. Effectiveness of WeChat for improving exclusive breastfeeding in Huzhu county China: randomized controlled trial. J Med Internet Res 2020 Dec 3;22(12):e23273. [doi: 10.2196/23273] [Medline: 33270026]
52. Wen LM, Rissel C, Xu H, et al. Effects of telephone and short message service support on infant feeding practices, “Tummy Time,” and screen time at 6 and 12 months of child age: a 3-group randomized clinical trial. JAMA Pediatr 2020 Jul 1;174(7):657-664. [doi: 10.1001/jamapediatrics.2020.0215] [Medline: 32282034]
53. Scott JA, Burns SK, Hauck YL, et al. Impact of a face-to-face versus smartphone app versus combined breastfeeding intervention targeting fathers: randomized controlled trial. JMIR Pediatr Parent 2021 Apr 12;4(2):e24579. [doi: 10.2196/24579] [Medline: 33843604]
54. Røed M, Medin AC, Vik FN, et al. Effect of a parent-focused eHealth intervention on children’s fruit, vegetable, and discretionary food intake (Food4toddlers): randomized controlled trial. J Med Internet Res 2021 Feb 16;23(2):e18311. [doi: 10.2196/18311] [Medline: 33591279]
55. Doan TTD, Pham NM, Binns C, et al. Effect of a mobile application on breastfeeding rates among mothers who have cesarean deliveries: a randomized controlled trial. Breastfeed Med 2022 Oct;17(10):832-840. [doi: 10.1089/bfm.2022.0088] [Medline: 36083249]
56. LeFevre AE, Shah N, Scott K, et al. The impact of a direct to beneficiary mobile communication program on reproductive and child health outcomes: a randomised controlled trial in India. BMJ Glob Health 2022 Jul;6(Suppl 5):e008838. [doi: 10.1136/bmjgh-2022-008838] [Medline: 35835477]
57. Staiano AE, Newton RL, Beyl RA, et al. mHealth intervention for motor skills: a randomized controlled trial. Pediatrics 2022 May 1;149(5):e2021053362. [doi: 10.1542/peds.2021-053362] [Medline: 35415743]
58. Acar Z, Şahin N. Development of a mobile application -based breastfeeding program and evaluation of its effectiveness. J Pediatr Nurs 2024;74:51-60. [doi: 10.1016/j.pedn.2023.11.011] [Medline: 37995477]
59. Hmone MP, Li M, Agho KE, Alam NA, Chad N, Dibley MJ. Tailored text messages to improve breastfeeding practices in Yangon, Myanmar: the M528 individually randomized controlled trial. Am J Clin Nutr 2023 Mar;117(3):518-528. [doi: 10.1016/j.ajcnut.2023.01.003] [Medline: 36811470]
60. Alexandrou C, Henriksson H, Henström M, et al. Effectiveness of a smartphone app (MINISTOP 2.0) integrated in primary child health care to promote healthy diet and physical activity behaviors and prevent obesity in preschool-aged children: randomized controlled trial. Int J Behav Nutr Phys Act 2023 Feb 21;20(1):22. [doi: 10.1186/s12966-023-01405-5] [Medline: 36810069]
61. Wu Q, Wang X, Zhang J, Zhang Y, van Velthoven MH. The effectiveness of a WeChat-based self-assessment with a tailored feedback report on improving complementary feeding and movement behaviour of children aged 6-20 months in rural China: a cluster randomized controlled trial. Lancet Reg Health West Pac 2023 Aug;37:100796. [doi: 10.1016/j.lanwpc.2023.100796] [Medline: 37273963]
62. Vila-Candel R, Mena-Tudela D, Franco-Antonio C, Quesada JA, Soriano-Vidal FJ. Effects of a mobile application on breastfeeding maintenance in the first 6 months after birth: randomised controlled trial (COMLACT study). Midwifery 2024 Jan;128:103874. [doi: 10.1016/j.midw.2023.103874] [Medline: 37979550]
63. Moon RY, Hauck FR, Colson ER, et al. The effect of nursing quality improvement and mobile health interventions on infant sleep practices: a randomized clinical trial. JAMA 2017 Jul 25;318(4):351-359. [doi: 10.1001/jama.2017.8982] [Medline: 28742913]
64. Davis KE, Klingenberg A, Massey-Stokes M, et al. The baby bites text messaging project with randomized controlled trial: texting to improve infant feeding practices. Mhealth 2023;9:11. [doi: 10.21037/mhealth-22-31] [Medline: 37089270]
65. Saucedo Baza A, Mignacca C, Delgado PE, et al. A technological approach to improved breastfeeding rates and self-efficacy: a randomized controlled pilot study. J Hum Lact 2023 Nov;39(4):679-687. [doi: 10.1177/08903344231190625] [Medline: 37571838]
66. Bakırcı-Taylor AL, Reed DB, McCool B, Dawson JA. mHealth improved fruit and vegetable accessibility and intake in young children. J Nutr Educ Behav 2019 May;51(5):556-566. [doi: 10.1016/j.jneb.2018.11.008] [Medline: 30638880]
67. Palacios C, Campos M, Gibby C, Meléndez M, Lee JE, Banna J. Effect of a multi-site trial using short message service (SMS) on infant feeding practices and weight gain in low-income minorities. J Am Coll Nutr 2018;37(7):605-613. [doi: 10.1080/07315724.2018.1454353] [Medline: 29708471]
68. Mindell JA, Du Mond CE, Sadeh A, Telofski LS, Kulkarni N, Gunn E. Efficacy of an internet-based intervention for infant and toddler sleep disturbances. Sleep 2011 Apr 1;34(4):451-458. [doi: 10.1093/sleep/34.4.451] [Medline: 21461323]

XSL•FO
RenderX

69. Lewkowitz AK, López JD, Carter EB, et al. Impact of a novel smartphone application on low-income, first-time mothers' breastfeeding rates: a randomized controlled trial. Am J Obstet Gynecol MFM 2020 Aug;2(3):100143. [doi: 10.1016/j.ajogmf.2020.100143] [Medline: 33345878]
70. de Mello Sa SR, Wang Z, Sapkalova V, et al. A smartphone-based application to improve breastfeeding duration and self-efficacy: a randomized controlled clinical trial. Women Health 2025(2):1-13. [doi: 10.1080/03630242.2024.2448519]
71. Sandborg J, Downing KL, Orellana L, et al. Six-month intervention effect of a digital movement behavior intervention on parent- and child intermediary outcomes-results from the Let's Grow randomized controlled trial. Int J Behav Nutr Phys Act 2025 Jun 16;22(1):78. [doi: 10.1186/s12966-025-01764-1] [Medline: 40524171]
72. Li Y, Xiao Q, Chen M, et al. Improving parental health literacy in primary caregivers of 0- to 3-year-old children through a WeChat official account: cluster randomized controlled trial. JMIR Public Health Surveill 2024 Jul 4;10(1):e54623. [doi: 10.2196/54623] [Medline: 38989817]
73. Jongpaiboonpatana P, Jirakran K, Trairatvorakul P, Chonchaiya W. Increasing parent-child play frequency via exclusively online, video-based, peer-to-peer modeling program: randomized controlled trial. Pediatr Res 2025 Sep;98(3):997-1005. [doi: 10.1038/s41390-025-03800-z] [Medline: 39815088]
74. Hunsrisakhun J, Naorungroj S, Tangkuptanon W, Wattanasit P, Pupong K, Pithpornchaiyakul S. Impact of oral health chatbot with and without toothbrushing training on childhood caries. Int Dent J 2025 Apr;75(2):1348-1359. [doi: 10.1016/j.identj.2024.09.028] [Medline: 39443262]
75. Hojati A, Abbasalizad Farhangi M. MyKid'sNutrition mobile application: effect on mothers' nutritional knowledge and nutritional status of preschool-aged children with undernutrition – a randomised controlled trial. BMJNPH 2025 Jun;8(1):29-37. [doi: 10.1136/bmjnph-2024-001007]
76. Henshaw E, Cooper M, Wood T, Krishna S, Lockhart M, Doan S. A randomized controlled trial of the happy, healthy, loved personalized text-message program for new parent couples: impact on breastfeeding self-efficacy and mood. BMC Pregnancy Childbirth 2024 Jul 26;24(1):506. [doi: 10.1186/s12884-024-06684-9] [Medline: 39060974]
77. Gilano G, Dekker A, Fijten R. The Effect of mHealth on exclusive breastfeeding and its associated factors among women in South Ethiopia: a cluster randomized controlled trial. Nutrients 2025 Nov 5;17(21):3477. [doi: 10.3390/nu17213477] [Medline: 41228550]
78. Cunningham K, Cech S, Gupta AS, Rana PP, Humphries D, Frongillo EA. Text messages to improve young child diets: results from a cluster-randomized controlled trial in Kanchanpur, Nepal. Matern Child Nutr 2026 Mar;22(1):e13702. [doi: 10.1111/mcn.13702] [Medline: 39016667]
79. Cherie N, Wordofa MA, Debelew GT. The effect of an interactive mobile health intervention to improve community-based essential neonatal care practices among postpartum women in northeast Ethiopia: a cluster randomized controlled trial. Int Health 2025 Sep 3;17(5):820-835. [doi: 10.1093/inthealth/ihae080] [Medline: 39789852]
80. Brown AL, Hudson N, Pinfold J, et al. The impact of dose in an mHealth intervention to support parents and carers via healthy beginnings for hunter new England kids program: pragmatic randomized controlled trial. JMIR Form Res 2025 Oct 1;9(1):e70158. [doi: 10.2196/70158] [Medline: 41032735]
81. Phillips SM, Bourke M, Inniss BV, Ahluwalia M, Tucker P. A pilot effectiveness study of a just-in-time micro-randomized controlled trial on the physical activity and sedentary time of young children and their parents: the active family m-health intervention. PLoS ONE 2026;21(1):e0340687. [doi: 10.1371/journal.pone.0340687] [Medline: 41544066]
82. Gavine A, Marshall J, Buchanan P, et al. Remote provision of breastfeeding support and education: systematic review and meta-analysis. Matern Child Nutr 2022 Apr;18(2):e13296. [doi: 10.1111/mcn.13296] [Medline: 34964542]
83. Ziebart M, Kammermeier M, Koletzko B, Patro-Golab B. Mobile applications for promoting and supporting breastfeeding: systematic review and meta-analysis. Matern Child Nutr 2025 Jan;21(1):e13733. [doi: 10.1111/mcn.13733] [Medline: 39390945]
84. Qian J, Wu T, Lv M, et al. The value of mobile health in improving breastfeeding outcomes among perinatal or postpartum women: systematic review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2021 Jul 16;9(7):e26098. [doi: 10.2196/26098] [Medline: 34269681]
85. Halvorsrud K, Kucharska J, Adlington K, et al. Identifying evidence of effectiveness in the co-creation of research: a systematic review and meta-analysis of the international healthcare literature. J Public Health (Oxf) 2021 Apr 12;43(1):197-208. [doi: 10.1093/pubmed/fdz126] [Medline: 31608396]
86. Skivington K, Matthews L, Simpson SA, et al. A new framework for developing and evaluating complex interventions: update of medical research council guidance. Int J Nurs Stud 2024 Jun;154:104705. [doi: 10.1016/j.ijnurstu.2024.104705] [Medline: 38564982]
87. Henriksson P, Migueles JH, Söderström E, Sandborg J, Maddison R, Löf M. User engagement in relation to effectiveness of a digital lifestyle intervention (the HealthyMoms app) in pregnancy. Sci Rep 2022 Aug 13;12(1):13793. [doi: 10.1038/s41598-022-17554-9] [Medline: 35963935]
88. Lippke S, Corbet JM, Lange D, Parschau L, Schwarzer R. Intervention engagement moderates the dose-response relationships in a dietary intervention. Dose Response 2016;14(1):1559325816637515. [doi: 10.1177/1559325816637515] [Medline: 27069440]

XSL•FO
RenderX

89. Elkes J, Cro S, Batchelor R, et al. User engagement in clinical trials of digital mental health interventions: a systematic review. BMC Med Res Methodol 2024 Aug 24;24(1):184. [doi: 10.1186/s12874-024-02308-0] [Medline: 39182064]

Abbreviations

JBI: Joanna Briggs Institute
mHealth: mobile health
PICOS: Population, Intervention, Comparison, Outcomes, and Study design
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
PROSPERO: International Prospective Register of Systematic Reviews
RCT: randomized controlled trial
WHO: World Health Organization

Edited by S Brini; submitted 08.Oct.2025; peer-reviewed by A Staiano, C Romanzini, SM Stegenga; revised version received 03.Apr.2026; accepted 06.Apr.2026; published 26.Jun.2026.

Please cite as:
Sandborg J, Reese BL, Marshall S, Hesketh KD, Laws R, Downing KL
Digital Interventions Targeting Parents to Improve Early Childhood Movement, Nutrition, and Sleep Behaviors: Systematic Review
J Med Internet Res 2026;28:e85525
URL: https://www.jmir.org/2026/1/e85525
doi:10.2196/85525

© Johanna Sandborg, Brittany L Reese, Sarah Marshall, Kylie D Hesketh, Rachel Laws, Katherine L Downing. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Applications, Challenges, and Future Directions of Large Language Models in Health Care Communication: Scoping Review

Jing Chang[1*], MS; Ruotong Peng[2*], PhD; Xi Chen[1], PhD; Yishu Zhu[1], MS; Ruting Miao[3], MS; Zeng Cao[4*], MS; Hui Feng[5*], PhD

[1]Xiangya School of Nursing, Central South University, Changsha, China
[2]School of Nursing, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China
[3]School of Nursing, Jiangxi Medical College, Nanchang University, Nanchang, China
[4]Xiangya Hospital, Central South University, Changsha, China
[5]Xiangya School of Nursing, Central South University, Changsha, Changsha, China
[*]these authors contributed equally

Corresponding Author:
Hui Feng, PhD
Xiangya School of Nursing, Central South University, Changsha, Changsha, China

Abstract

Background: Effective health care communication is crucial in the medical field. However, effective communication in clinical practice still faces numerous obstacles, and large language models (LLMs) offer various possibilities for improving the quality of medical communication. To date, there are no published reviews on the use of LLMs in health care communication.

Objective: This review sought to summarize the applications and challenges of LLMs in health care communication and to identify directions for future research.

Methods: A comprehensive literature search was conducted in PubMed, Embase, Web of Science, and the Cochrane Library from January 2018 to November 2025. The search and selection process followed the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guideline and the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) checklist. Eligible studies used LLMs to facilitate health care communication among the public, patients, and clinicians. Following rigorous data extraction and cross-checking, we conducted a quantitative analysis of characteristics of the included literature. Furthermore, using communication accommodation theory as a framework, we identified application patterns of LLMs in health care communication and summarized current challenges and future directions.

Results: Ninety-six studies were included in this review, all published between 2023 and 2025, summarizing 4 patterns of LLM application in health care communication: transforming medical information (n=30), facilitating dynamic interaction (n=38), empowering communication capabilities (n=10), and optimizing clinical workflows (n=18). The role of LLMs in health care communication is undergoing a paradigm shift from “static information processing” to “dynamic intelligent interaction.” Although they show great promise for practical applications, current evaluation methods and dimensions exhibit significant heterogeneity. Furthermore, LLMs still face multiple challenges in their practical application in health care communication, including technical reliability issues, social trust and adoption, interaction and access barriers, and clinical integration challenges.

Conclusions: Unlike previous studies that merely touched upon the challenges and future directions, this scoping review uses communication accommodation theory to systematically map the application patterns and developmental landscape of LLM-mediated health care communication. Health care communication powered by LLMs holds significant innovation potential and is currently still in the early stages of rapid development. Future research should focus on optimizing model performance, strengthening ethical governance frameworks, enhancing human-machine collaboration models, and ensuring responsible application of LLMs in health care through rigorous empirical validation.

Trial Registration: OSF Registries 10.17605/OSF.IO/YVXSP; https://osf.io/yvxsp/overview

(J Med Internet Res 2026;28:e84726) doi:10.2196/84726

KEYWORDS

health care communication; communication accommodation theory; scoping review; digital health; large language models

XSL•FO
RenderX

Introduction

A good clinician-patient relationship is the foundation of medical practice [1]. Effective health care communication not only facilitates interprofessional collaboration and high-quality care delivery but also positively impacts patient treatment adherence, health outcomes, and overall health care quality [2-5]. However, achieving effective communication in clinical practice remains hindered by numerous barriers. Research indicates that clinicians spend more than 20% of their working hours on communication activities, with economic inefficiencies costing approximately US $4.9 billion annually [6]. This significant communication burden and efficiency pressure further limit opportunities for patient-centered interactions [5,6]. Additionally, complex medical terminology and language barriers hinder patients' comprehension of health care information, negatively impacting treatment decision-making and medication adherence [7,8].

Large language models (LLMs) offer a transformative paradigm for addressing these challenges. As artificial intelligence (AI) systems trained on massive text corpora, LLMs demonstrate exceptional performance in natural language understanding and generation tasks [9,10]. From text analysis and summary generation to clinical applications, LLMs demonstrate diverse capabilities for clinical support [11-13]. Simultaneously, through domain-specific fine-tuning, LLMs can maintain ongoing engagement with user queries, facilitate interactions, and generate controlled outputs [14]. Furthermore, existing research has highlighted the unique value of LLMs in health care communication by developing chatbots with customized behaviors [14]. Multiple studies indicate that LLMs are key tools for improving information transmission efficiency and alleviating the burden of clinical communication [15-18]. Therefore, leveraging LLMs to optimize information dissemination and communication methods will be a crucial approach to improving the quality of health care communication in the future [19].

In this study, health care communication is defined as the dynamic, interactive process within medical settings that facilitates accurate transfer of clinical information, emotional exchange, and collaborative decision-making through linguistic and nonlinguistic mediation [20-22]. To systematically interpret how LLMs intervene in this complex process, this review introduces the interaction-centered communication accommodation theory (CAT) as an analytical framework [23]. Proposed by Howard Giles in 1973, CAT emphasizes that individuals dynamically adjust speech, intonation, and discourse to manage social distance and interpersonal relationships [24,25]. In clinical settings, CAT reveals how providers adapt their communication styles to foster trust and understanding [26,27]. CAT identifies 5 sociolinguistic strategies, including approximation, explicability, interpersonal control, discourse management, and emotional expression [24]. It provides a systematic lens for analyzing how LLMs empower medical communication [28]. Research underscores that CAT uniquely suits large-scale text-based telemedicine among multiple frameworks that elucidate clinical communication [24].

As an innovative and transformative technology, LLM-based health care communication demonstrates immense potential to advance medical communication toward greater efficiency, precision, and personalization [29,30]. However, to date, only one commentary has briefly explored the role and challenges of LLMs in medical communication, highlighting the field's future potential [31]. At present, the specific use cases, challenges, and future directions of LLMs in current applications remain unclear. Therefore, this study aims to systematically map the current state of research in this field through a scoping review and identify known knowledge gaps. The key issues addressed in this review include (1) systematizing application patterns of LLMs in health care communication based on the CAT, (2) exploring current evaluation methods and dimensions in the health care communication domain, (3) identifying limitations and challenges in applying LLMs to health care communication, and (4) proposing recommendations for future research to inform the better development and application of LLM-based health care communication.

Methods

Study Design, Protocol, and Registration

This scoping review adheres to the methodological framework proposed by Arksey and O'Malley [32], and was conducted in accordance with PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews; Checklist 1) and the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension; Checklist 2) to ensure methodological transparency and reproducibility [33,34]. Our full review protocol is published in the Open Science Framework registries [35].

Information Sources and Search

To identify relevant English-language literature, the research team systematically searched the PubMed, Embase, Web of Science, and Cochrane Library databases, without using additional information sources. The initial search was conducted on July 30, 2024, and the final search on November 20, 2025, using the same search strategy to identify newly published research. Each database was searched via its web interface, with searches conducted separately within each database. No research registries or other online sources (such as websites, conference proceedings, and journal directories) were searched. We did not contact researchers to seek additional sources. All search strategies were developed specifically for this scoping review and were not formally peer-reviewed by independent experts before implementation. A filter for publications released after 2018 was applied to all databases. The search strategy combined MeSH (Medical Subject Headings) terms with free-text keywords, with core concepts covering LLMs (eg, large language model OR ChatGPT OR large language model* OR language neural network* OR generative AI OR AI-Generated OR generative artificial intelligence OR ChatGPT OR Artificial Intelligence Chatbots OR MedPalm OR GPT OR pretrained model* OR conversational AI OR deep learning language model* OR language model* OR language generation model*) and communication (eg, Respon* OR Repl* OR Report* OR

XSL•FO
RenderX

Question* OR Transform* OR Summar* OR Communicat* OR Interpret* OR Explan* OR Inform* OR Answer). The specific search terms were determined by the research team through iterative discussions, with search strategies and keywords adjusted based on results to ensure the retrieval of literature spanning both research domains. We adjusted the search strategy based on each database, and the complete search strategy is detailed in Multimedia Appendix 1 [36-131]. Additionally, we manually screened the reference lists of included studies for additional eligible records. No forward citation searching was performed.

Selection of Sources of Evidence

The retrieved literature was managed and deduplicated using EndNote 21 software (Clarivate). Two systematically trained researchers independently screened the titles and abstracts of each article based on inclusion and exclusion criteria. Similarly, the full text of articles included in the title and abstract screening was independently reviewed by 2 authors (JC and RP) for inclusion in the evaluation. Any discrepancies were resolved through discussion, involving a third author (HF) when necessary. Additionally, this study aims to provide a comprehensive overview of LLMs' applications in health care communication. Therefore, the included studies were not assessed for methodological quality to ensure the breadth of the literature review.

Eligibility Criteria

The review applied the following inclusioncriteria:

1. Peer-reviewed empirical studies applying LLMs to health care communication.
2. Published between 2018 and 2025. This cutoff reflects the introduction of Bidirectional Encoder Representations from Transformers, a novel language representation model widely regarded as the origin of contemporary LLMs [132].

The review applied the following exclusion criteria:

1. Research that focuses solely on LLMs as tools for static knowledge retrieval or question-answering (eg, evaluating only their accuracy in answering questions on a medical licensing exam). This review defines "health care communication" as a collaborative process aimed at achieving information exchange, emotional interaction, and shared decision-making through dynamic interaction. Accuracy assessments based on static question-answering focus solely on the quality of knowledge retrieval; they fail to reflect a model's ability to respond in real time to user feedback and emotional needs, nor can they evaluate the model's effectiveness in applying communicative adaptation strategies for dynamic adjustment (for details on excluded categories, refer to Multimedia Appendix 1).
2. Articles unrelated to health care communication, such as prediction or diagnosis of disease.
3. Study protocols, preprints, trial registrations, editorials, letters, and commentaries.

Data Charting Process and Data Items

The data for each article were extracted through a predesigned data extraction form by the research team. Two authors (JC and RP) independently extracted data from the identified studies using Microsoft Excel, including the author, year, country, study design, medical disciplines, targeted population, research objectives, type of model, evaluation methods, evaluation content, application patterns of LLMs in health care communication, challenges, and other relevant information (eg, barriers). Potential discrepancies in data extraction were discussed by the authors (JC and RP) and resolved. At least one additional author (HF) independently verified the accuracy of each literature record to validate the analysis results.

Synthesis of Results

This study summarizes and analyzes the extracted data. The research team used descriptive statistical methods to summarize the general characteristics, evaluation methods, and evaluation content of the included studies. Additionally, this study used thematic analysis within the CAT framework to identify patterns and strategies for the application of LLMs in health care communication and to outline the existing challenges. This study strictly followed the 3-stage thematic synthesis method proposed by Thomas and Harden [133]. The 2 authors independently performed open coding of the extracted content from the included studies based on the meaning and content of the data. Subsequently, while remaining faithful to the original findings of the included studies, they distilled their core meanings, categorized the open-coded data, and developed descriptive themes. The descriptive themes were subsequently developed into analytical themes focusing on the patterns and challenges of LLM applications in health care communication. To ensure the rigor of the analysis, all initial coding and the thematic framework were independently reviewed and validated by a third researcher (HF). For any discrepancies in understanding or coding, the research team held multiple meetings for in-depth discussion and, when necessary, redefined the coding manual until the discrepancies were resolved and consensus was reached. The final analysis results were presented through a combination of narrative descriptions and charts, aiming to provide a comprehensive reflection of the current state of this field.

Results

Selection of Sources of Evidence

The study screening process is summarized in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram (Figure 1). A total of 13,677 articles were retrieved from the database, and 4908 duplicates were removed. The full text of 754 articles was screened after screening the titles and abstracts. Finally, a total of 96 studies [36-131] met the inclusion criteria.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of article selection.

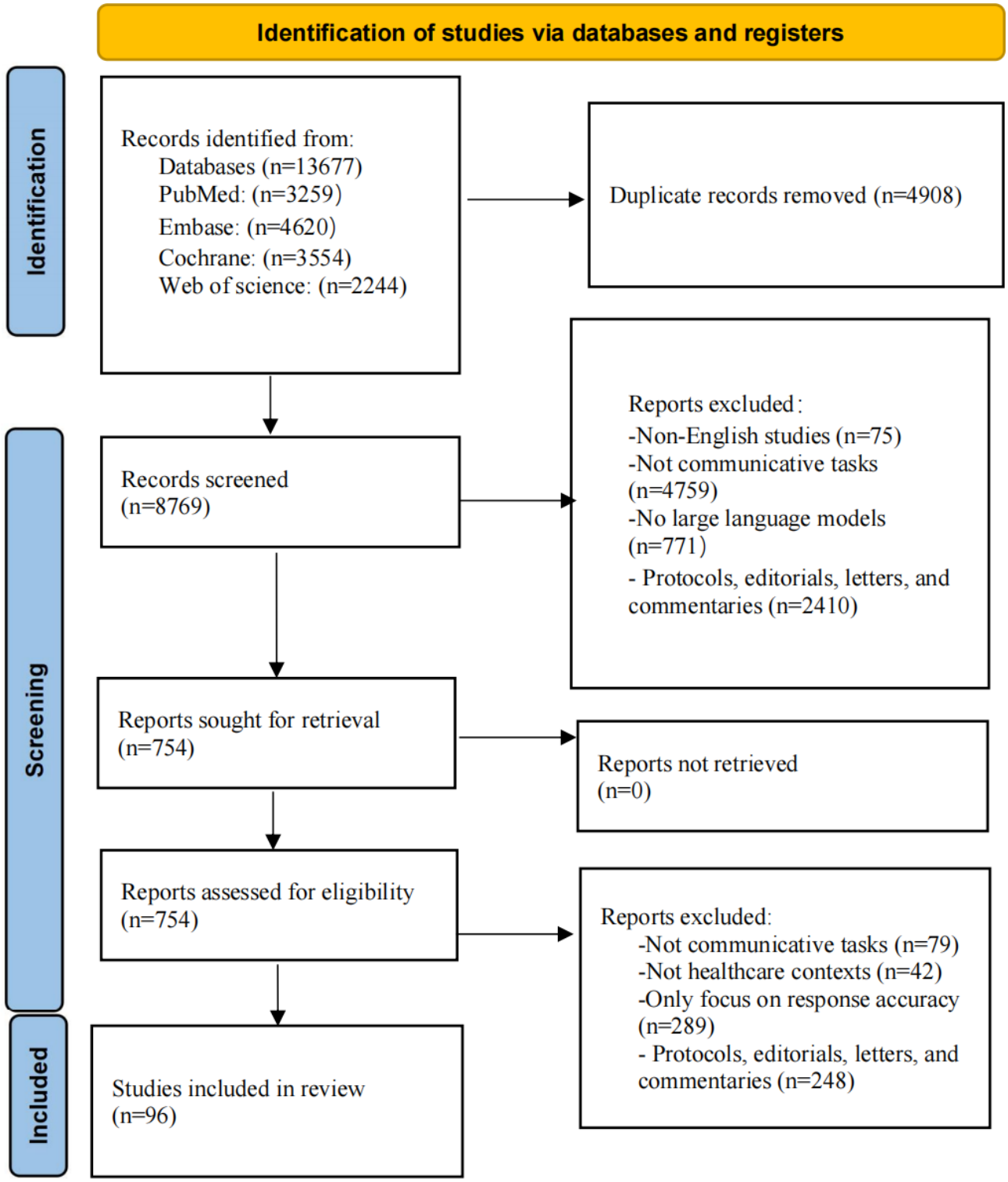

Characteristics of Sources of Evidence

The included studies were published between 2023 and 2025 across 19 countries. The United States had the most studies (n = 46) [37-41,44-47,51,53-55,58-60,62,64-66,68,69,73,78-80,83-86,88,89,94-99,100,102,110,113,116,119,122-131], followed by China (n=13), Turkey (n=10) [70-72,74-76,82,85,89,92], and Germany (n=7) [42,48,50,52,93,103,115]. Other countries included Australia (n=4) [43,101,107,112], South Korea (n=2) [108,117], India (n=2) [49,61], Poland [79], Italy [118], the United Kingdom [36], Ireland [87], Romania [67], the Netherlands [77], Japan [84], Brazil [90], Thailand [120], Sweden [104], Singapore [96], and Colombia [63] (each n=1). The included publications exhibited diverse methodological characteristics, comprising 33 descriptive studies [33,34,41,45,46,49,53,58-60,62,64-66,68,69,71,76,80,82,83,89,94,99,100,102,103,115,116,122,126,127], 24 comparative studies [43,47,48,52,61,70,72-75,77,79,81,84,85,87,88,91-93,95,97,98,117], 8 cross-sectional studies [51,54,55,63,78,86,90,125], 6

XSL•FO
RenderX

technology development studies [67,111-113,120,121], 5 randomized controlled trials [37,40,105,114,118], 6 quality improvement studies [44,123,124,128-130], 3 mixed methods studies [108,119,126], and 1 each of proof-of-concept studies [101], observational studies [131], cohort studies[50], qualitative studies [104], exploratory studies [42], pilot studies [36], feasibility studies [109], usability studies [107], multiphase studies [110], multicenter quantitative studies [57], and survey studies [53]. In terms of the specific models used, 83 studies [36-38,40-42,44-53,55-66,68-97,99,100,102-110,112,115-117,119,122,124-131] examined ChatGPT (OpenAI) performance either on its own or in comparison with other models, 6 studies [39,54,98,101,111,123] did not specify the models used, and 4 studies [67,85,113,114] used models they had developed themselves. Research applications spanned multiple medical disciplines, with radiology being the most common field (n=14) [36,37,39-41,45,46,48-50,52-54,65]. For details on the country, study design, medical disciplines, target population, research objectives, model type, evaluation methods, and evaluation content for each study, please refer to Multimedia Appendix 1.

Synthesis of Results: Different Categories of Application Patterns of LLMs

The systematic search and screening process identified 96 studies [36-131] meeting the inclusion criteria. Based on the nature of communication tasks, these studies were categorized into four primary application patterns: (1) transforming medical information, (2) facilitating dynamic interaction, (3) empowering communication capabilities, and (4) optimizing clinical workflows (Figure 2). For clarity, Table 1 summarizes the strategies and tactics of CAT, listing their classic definitions and applications in this study.

Figure 2. Four application patterns of large language models in health care communication.

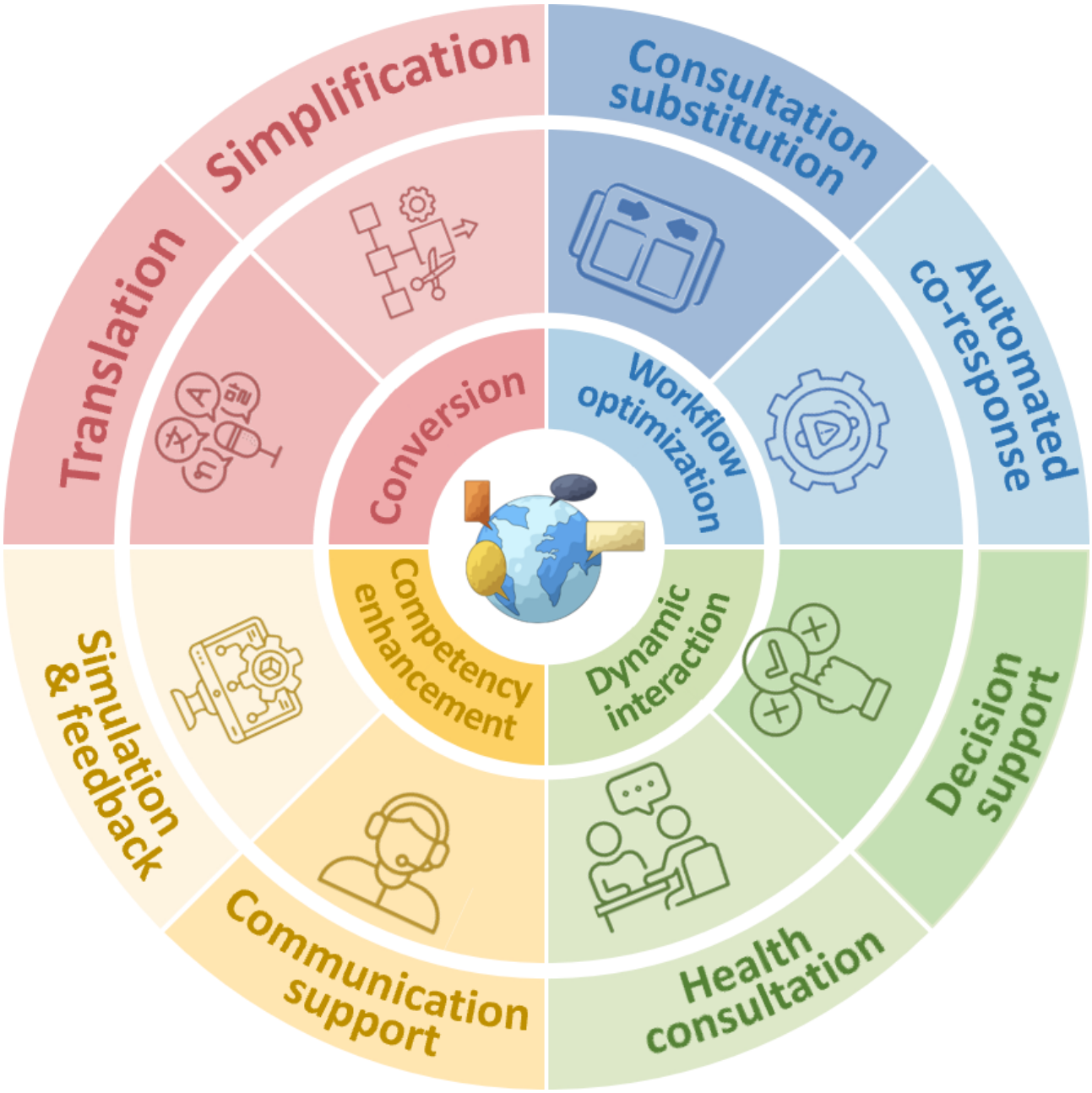

XSL•FO
RenderX

Table . CAT[a] strategies, definitions, and their application [24].

Strategy	Classic CAT definition	Tactics (this study) and application
Approximation	Adjusting one's communicative style to be more similar (convergence), more distinct (divergence), or unchanged (maintenance) relative to the interlocutor.	• Convergence: use language that resonates with users; simplify jargon. • Divergence: using technical jargon without a lay explanation. • Maintenance: maintain a consistent tone and style regardless of user changes.
Interpretability	Ensuring messages are understandable by adjusting complexity, clarity, or explicitness.	• Simplification: breaking down complex medical terms into lay language. • Clarification: providing additional explanations or examples to aid understanding.
Discourse management	Regulating conversational flow, including turn-taking and topic control.	• Topic initiation: guide the conversation toward relevant topics. • Topic shifting: redirect discussions to maintain focus or address new issues.
Interpersonal control	Managing role dynamics and authority in interaction.	• Assertiveness: delivering confident recommendations and guiding the consultation. • Responsiveness: promptly responding to patient questions and concerns.
Emotional expression	Conveying empathy, support, and emotional alignment.	• Empathy: perceiving the patient's emotions, expressing empathy. • Support offer: encouraging the patient, providing emotional support.

[a]CAT: communication accommodation theory.

The application of LLMs in health care communication has shown a significant growth trend and an evolution in functional focus between 2023 and 2025 (Figure 3). In Figure 3, the horizontal axis represents the year of publication, the vertical axis reflects different application patterns of LLMs in health care communication, and the size of the bubbles indicates the number of related studies. Among these, "Transforming medical information" emerged earliest, with the number of related studies steadily increasing from 3 [41,46,49] in 2023 to 17 [36-40,42-45,47-65] in 2025. The "Facilitating dynamic interaction" domain had no recorded applications in 2023, but rapidly increased to 6 in 2024 and further grew to 32 in 2025, becoming the most widely applied domain to date. Meanwhile, "Optimizing clinical workflow" grew from zero applications in 2023 to 10 in 2025. In contrast, "Empowering communication capabilities" started later (with only 2 studies [107,113] in 2024) but had grown to 8 [104-116] by 2025.

XSL•FO
RenderX

Figure 3. Evolution application patterns of large language models in health care communication. Note: The horizontal axis shows the distribution of publication years, while the vertical axis illustrates the application patterns of large language models in health care communication; the size of the bubbles indicates the number of studies.

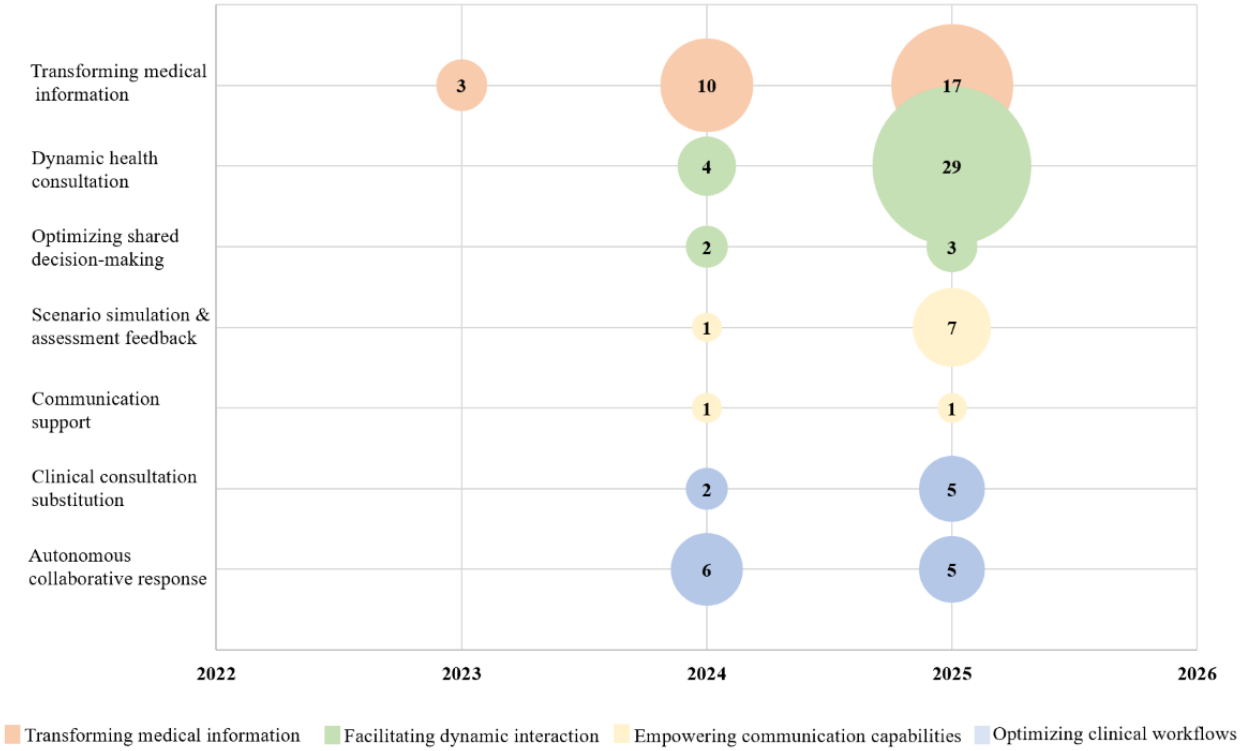

Transforming Medical Information

LLMs bridge the health literacy gap between physicians and patients by converting specialized medical texts into patient-friendly summaries through interpretability and approximation strategies (n=30). Most studies (n=23) focuses on the "deprofessionalization" of medical documentation. Specifically, LLMs were used to simplify radiology reports (n=13), pathology reports (n=5), and discharge summaries (n=5) [36-58]. Through simplification and clarification tactics, clinical terminology is adapted into plain-language information that patients can readily comprehend, reducing comprehension barriers for those with limited medical knowledge. Furthermore, specialized educational materials (eg, kidney stones, hand surgery, shoulder surgery, and palliative care; n=4) [59-62] and cross-language translations (n=3) were generated [63-65]. This approach mimics the expression patterns of nonspecialist audiences, not only eliminating language and comprehension barriers but also ensuring patients can participate equally in health care decision-making during dynamic clinical interactions, thereby promoting equitable dissemination of health information.

Facilitating Dynamic Interaction

Dynamic Health Consultation

LLMs enable dynamic personalized information exchange through multiturn dialogues, transforming health care information access from "static search" to "dynamic interaction" by leveraging the responsiveness substrategy within interpersonal control strategies (n=33). Thirty-three studies used LLMs to capture patient needs and provide expert-level advice by initiating topics within discourse management strategies, addressing disease-related inquiries [66-98]. For instance, He et al [97] investigated whether LLMs could provide patients with inflammatory bowel disease with appropriate advice comparable to that of gastroenterologists. Wu et al [98] evaluated the quality of preventive and therapeutic advice generated by language models for influenza-related queries in online health communities, focusing on their performance in delivering emotional support through emotional expression strategies.

Optimizing Shared Decision-Making

LLMs bridge communication gaps between clinicians and patients by integrating approximation and discourse management strategies (n=5). Two studies used LLMs to provide clinicians with up-to-date clinical guidelines, conversation prompts, and key summaries via approximation strategies, while also helping patients understand disease information, treatment plans, and potential risks to facilitate shared decision-making [99,100]. Three studies examined the use of LLMs to clarify patients' concerns through interactive dialogue, thereby enhancing the quality of the clinical informed consent process [100-102]. For instance, Allen et al [101] pioneered a 4-phase interactive model (precommunication, language model–guided interaction, clarification phase, and physician review) that uses discourse management strategies to optimize informed consent workflows. Furthermore, this model enables patients to clarify preoperative concerns through repeatable, low-pressure dialogues using support-offer tactics, thereby improving the efficiency of

XSL•FO
RenderX

informed consent [101]. The other 2 studies examined LLM support for informed consent in oxytocin-induced labor and total knee arthroplasty [102,103].

Empowering Communication Capabilities

Scenario Simulation and Assessment Feedback

LLMs support patient-centered communication by simulating real-world clinical scenarios, providing constructive feedback, and integrating biosignal technology (n=8). Seven studies used convergence tactics to simulate diverse patient roles, enabling trainees to engage in authentic role-playing across clinical contexts such as pharmacy, emergency medicine, obstetrics and gynecology, nursing education, and pain communication. This approach facilitates mastery of patient-centered communication techniques [104-110]. Some studies integrated virtual reality technology to simulate virtual wards, enabling concurrent training in verbal communication and physical environment management. LLMs further enhance communication capabilities through emotional expression strategies and interpersonal control strategies. Beyond simulation training, LLMs also provide constructive feedback on the quality and emotional resonance of clinical interactions, reinforcing skills such as active listening and empathy. Additionally, advanced systems such as EEG Emotion Copilot integrate biological signals and LLMs to assist clinicians in identifying patients' emotions, thereby delivering personalized, emotionally intelligent treatment recommendations [111].

Communication Support

For individuals with specific language or physical impairments, LLMs provide support by integrating approximation and interpretability strategies, thereby promoting communication equality (n=2). Adikari et al [112] integrated a custom-trained LLM into a conversational system, enabling real-time detection and correction of neologisms and semantic errors, as well as intelligent sentence completion suggestions during interruptions. This effectively leverages strategies to help people with aphasia express themselves in standard contexts, ensuring accurate transmission of meaning. For patients with amyotrophic lateral sclerosis, LLM-driven predictive text functionality significantly reduced the physical burden of eye-tracking operations, while clarification tactics enhanced communication fluency and efficiency in clinical settings [113].

Optimizing Clinical Workflows

Clinical Consultation Substitution

Seven studies explored leveraging discourse management strategies to empower LLMs, either supplementing or replacing health care professionals in specialized applications and triage scenarios, thereby streamlining clinical communication processes (n=7) [113-119]. The "SSPEC" chatbot developed by Wan et al [114] demonstrated its ability to guide patients in recounting medical needs through topic initiation tactics, assist nurses in outpatient reception and triage, and address concerns about model accuracy through response monitoring and early warning mechanisms. In specialized clinical settings, 6 studies integrated interpretability and interpersonal control strategies to enhance the performance of language models in oral hygiene guidance, emergency department management, mental health coping, urology clinical consultations, dental education, and follow-up [115-120]. For example, Chung et al [116] used ChatGPT for preoperative consultations, effectively enhancing patients' disease awareness and optimizing clinical communication efficiency.

Autonomous Collaborative Response

LLMs enhance communication efficiency in health care systems by leveraging interpretability and approximation strategies to automate information generation and enable cross-specialty collaboration (n=11). In prehospital emergency scenarios, LLMs, combined with speech recognition technology, rapidly generate diagnostic summaries to shorten treatment response times [121]. Addressing high-pressure handover environments in emergency departments, Genes et al [122] used LLMs to generate structured transfer summaries, thereby improving emergency department handover processes. Furthermore, this technology enhances interdisciplinary comprehension by adding lay summaries to highly specialized records (eg, ophthalmology notes), ensuring nonspecialist physicians accurately interpret and implement specialist diagnoses during patient referrals [123]. LLMs also demonstrate utility in routine administrative tasks and electronic patient consultations. Specifically, 6 studies explored their capacity for clinical automation through discourse management strategies in automatically responding to patient portal messages and electronic health record inquiries [124-129]. Two additional studies used convergence tactics to ensure generated content aligns with professional norms and expression conventions, examining the "LLM draft-physician review" practice model's role in reducing health care provider burden and optimizing clinical workflows [130,131].

Evaluation Methods and Dimensions of LLM Applications

Of the 96 studies included [36-131], 93 evaluated the practical applications of LLMs in health care communication [36-66,68-100,102-118,120-131]. Evaluation methods primarily included subjective assessments, qualitative interviews with domain experts and patients, and objective metric measurements using standardized assessment tools (Figure 4).

XSL•FO
RenderX

Figure 4. Mapping of evaluation methods and dimensions. LLM: large language model. Note: Validated assessment tools, for example, the Flesch-Kincaid Grade Level is a widely recognized readability formula that estimates the reading level of a text based on average sentence length and vocabulary complexity and generates a corresponding score [77].

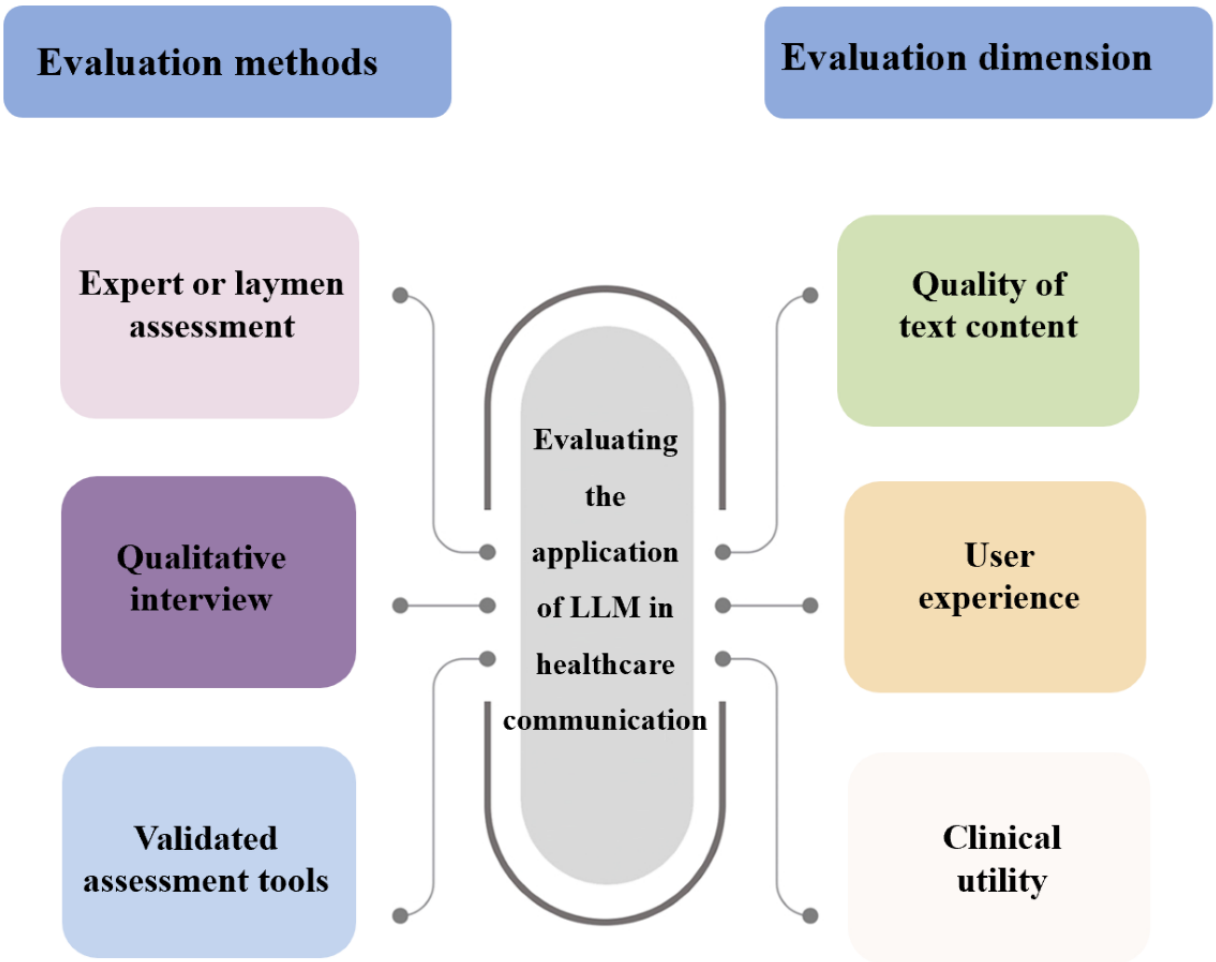

The evaluation tools exhibited a diverse range of characteristics. Existing studies primarily used Likert scales, custom scoring systems, and questionnaires to evaluate content, combined with qualitative interviews to capture user experience [88,89,110]. A small number of metrics use objective measurement tools; readability metrics include the Flesch Reading Ease, Flesch-Kincaid Grade Level, Coleman-Liau Index, Simplified Nonsense Measure, Gunning-Fog Index, and the Flesch-Szigrist formula [77,85,88]. Some researchers used the DISCERN tool (Deborah Charnock), the Global Quality Score, and the Misinformation Rating Scale to assess content quality [75,90]. Furthermore, measurement tools tailored to specific clinical scenarios, such as the Communication Confidence Self-Assessment Scale and the Mother-Infant Care Communication Assessment Scale, have been used to evaluate changes in users' communication skills [105,106].

The evaluation dimensions primarily include content quality, user experience, and clinical utility, with content quality being the core focus of the research (Figure 5). A total of 80 studies used the following criteria to assess content quality: accuracy, consistency, readability, clarity, comprehensiveness, overall quality, safety, and hallucinations [36-66,68-97,99,100,102,103,110,112,114-117,120-123,125-129]. Among these, accuracy (n=48) and readability (n=33) were the most frequently applied evaluation metrics. Twenty-nine studies evaluated users' experiences with LLM-based health care communication [53-55,61,78,79,81,84,86-88,93,98,104,105,108-110,114,117-119,123,125-130]. These evaluations covered not only metrics such as satisfaction, likability, and practicality but also collected user experiences and subjective preferences through qualitative research [54,55,104,109]. Additionally, some studies assessed model performance from an emotional perspective (such as empathy) [78,79]. Regarding clinical utility, 27 studies examined the feasibility of integrating LLMs into clinical workflows and their impact on medical practice [36,42,47,57,61,75,81,85,87,97,104-109,113-115,117,118,122-124,126,128,131]. Evaluation metrics varied by specific clinical communication task, primarily focusing on feasibility, practical impact, and efficiency. For example, studies used metrics such as willingness to use, usability, usefulness, and applicability to assess feasibility [61,109]. The impact of LLMs as assistive tools on clinical practice was evaluated by assessing improvements in communication confidence, medical history collection, empathy,

XSL•FO
RenderX

and communication skills [105,106]. Furthermore, efficiency metrics were primarily measured by quantifying relevant parameters within clinical workflows, including communication efficiency and AI draft usage rates [123,124,129]. For specific details, please refer to Multimedia Appendix 1.

Figure 5. Evaluation dimensions and frequencies. GenAI: generative artificial intelligence.

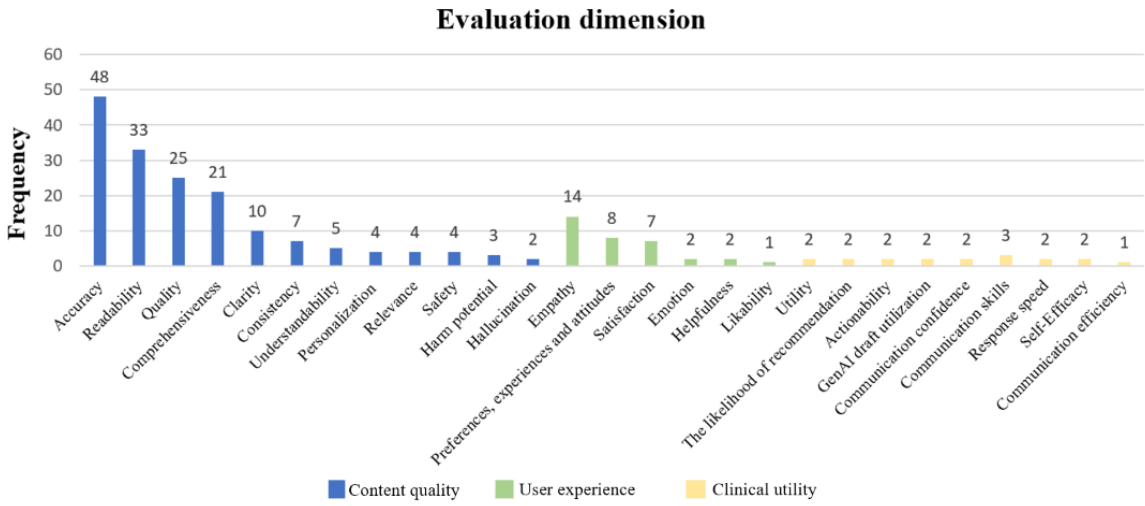

Comparison of Performance Differences

A total of 19 studies evaluated performance differences among mainstream LLMs in health care communication [48,51,61,63,70,71,73-75,77,85,89-93,95,99,119]. In tasks such as dental implant consultations, melanoma management, patient-physician communication regarding rare diseases, and the interpretation of pathology reports, ChatGPT consistently outperformed Bard (now Gemini; Google) in response accuracy; however, in rhinoplasty consultations, its performance was slightly inferior to Claude's performance [51,63,73,77,92,93]. In obstetric care consultations, ChatGPT, Kimi (Moonshot AI), and ERNIE Bot (Baidu) showed similar performance in terms of accuracy and completeness [95]. Although ChatGPT demonstrated high accuracy rates in multiple studies, its readability scores were generally low. In contrast, DeepSeek demonstrated advantages in optimizing the accessibility of medical information, with responses that outperformed GPT and Gemini in clarity, comprehensiveness, and readability [70,74,91,92]. In mental health interactions, Pi outperforms ChatGPT in empathy and user acceptance due to its human-like response style [119]. Regarding multilingual performance, ChatGPT performs best at interpreting pathology reports in Spanish, while Perplexity stands out on English tests [119].

Three studies compared the clinical performance of general-purpose models with domain-specific models [84,93,126]. In generating draft responses to patients, ChatGPT significantly outperformed specialized models in overall scores and responsiveness metrics. However, expert evaluations noted that responses from the specialized model CLAIR better aligned with physicians' professional language styles, whereas ChatGPT's answers were perceived as having a "robotic" tone [126]. In tasks involving doctor-patient communication regarding rare diseases, ChatGPT outperformed BioMistral. However, BioMistral 7B can run locally within a medical setting, offering privacy advantages [93]. Additionally, Özcivelek and Özcan [85] noted that the domain-specific model Dental GPT demonstrated the highest factual accuracy in consultations regarding oral and maxillofacial prosthetics, but performed worst in terms of readability.

Existing Challenges

We have identified four existing challenges for LLMs in health care communication: (1) technical reliability issues, (2) social trust and adoption, (3) interaction and access barriers, and (4) clinical integration challenges (Figure 6).

XSL•FO
RenderX

Figure 6. Existing challenges.

Technical Reliability Issues

Research indicates that, to improve readability for patients, LLMs may oversimplify medical terminology, potentially omitting critical details [58]. Furthermore, constrained by the "hallucination," models may produce associated inferences with ambiguous terminology definitions, erroneous mechanism descriptions, or insufficient evidence [36,82]. Furthermore, the effectiveness of LLM responses depends heavily on the quality and timeliness of training data. Most current models are trained on general internet corpora rather than rigorously vetted specialized medical datasets, and may generate information that is inconsistent with clinical guidelines or outdated [75]. Combined with inherent "black-box characteristics" and data update delays, this further undermines the interpretability and reliability of their outputs [97,98].

Social Trust and Adoption

LLMs lack a holistic understanding of patients' clinical contexts, social backgrounds, and psychological states. Their responses may lack sufficient depth to support the comprehensive situational judgment required in medical practice, making it difficult to address complex clinical issues [87]. At the emotional-interaction level, LLMs cannot perceive nonverbal cues such as facial expressions and tonal shifts, making it difficult to gauge patients' emotional states and support needs [58]. Moreover, the directive tone and verbose expressions in their responses may provoke resistance among individuals experiencing emotional crises [119]. Furthermore, existing models are predominantly built on English-language datasets and Western value systems, exhibiting significant limitations in multicultural health care settings and potentially leading to biased outputs [128]. Furthermore, unclear legal and ethical accountability, coupled with data privacy and security risks, collectively undermine societal trust in LLMs within serious medical contexts [58].

Interaction and Access Barriers

The clinical efficacy of LLMs relies heavily on the quality of user prompts [43]. Existing research predominantly uses standardized question sets for testing; yet, real-world clinical inquiries often exhibit ambiguity, emotionality, or unstructured characteristics. Populations with lower health literacy may struggle to obtain accurate information through effective interaction. Additionally, regional disparities in model deployment may create access barriers for economically constrained areas and populations [98].

Clinical Integration Challenges

Currently, clinical text summaries generated by LLMs often lack consistent formatting and standardization, making them difficult to integrate directly into existing clinical workflows or electronic health record systems [41]. Furthermore, these models tend to provide guideline-based "standardized" responses, struggling to tailor recommendations based on patient history, psychosocial factors, or individual preferences [70,97]. They also fail to dynamically optimize advice based on real-time assessments or rehabilitation progress. More critically, existing models lack mechanisms for medical risk assessment and emergency referral, and are unable to provide timely guidance during patient health crises, posing potential safety risks [119]. The lack of real-world clinical validation further hinders their ability to serve as true clinical substitutes in complex, dynamic doctor-patient interactions [97,121,125].

Discussion

Summary of Evidence

This review systematically explores the application patterns of LLMs in health care communication. Existing research has evaluated content quality, clinical utility, and user experience of LLMs primarily through objective and subjective metrics. Despite promising prospects, the field currently faces multiple challenges, which provide direction and focal points for future research. As technology advances and applications expand, LLMs will play an increasingly vital role in health care communication.

Application Domains and Development Trends of LLMs

Relevant literature was published between 2023 and 2025, coinciding with the November 2022 release of ChatGPT. The geographic distribution of the included studies showed a bias toward developed countries, such as the United States. This may be attributed to the region's thriving IT industry, advanced health care infrastructure, and ample funding support [134]. ChatGPT emerged as the most frequently cited model in this review due to its practicality and ease of use for clinicians [135]. The studies also encompassed general-purpose models such as Gemini and DeepSeek, alongside a few custom-developed models, highlighting the breadth of models explored in the included research. By testing and comparing different models, researchers can systematically evaluate performance variations and identify model biases [136]. Application domains show diversification trends. As research advances and models

XSL•FO
RenderX

optimize, customized professional models focused on clinical specialties are expected to be deployed across multiple scenarios.

The application of LLMs in health care communication is shifting from "static information processing" to "dynamic intelligent interaction." Transforming medical information is the earliest and most steadily growing area. These tasks rely primarily on structured reasoning and linguistic fluency rather than on complex diagnostic reasoning, giving them an early advantage in technological implementation [137]. As the technology continues to advance, the research focus has shifted toward addressing the more challenging demands of real-time feedback and multiturn dialogue. For instance, the sustained growth in the field of facilitating dynamic interaction indicates that LLMs are gradually evolving into "participants" capable of assisting doctor-patient interactions. Concurrently, the emergence of applications aimed at optimizing clinical workflows reflects a trend toward integrating LLM technology, specifically through the automation of communication tasks to alleviate the increasingly severe administrative burden and clinical pressure within health care systems. Although the field of communication enhancement had a relatively late start, its sustained growth demonstrates that the potential of LLMs to strengthen emotional connections between doctors and patients and assist in complex decision-making is gradually being realized. Future research should focus on translating LLMs' efficacy into practical clinical settings, particularly on their real-world performance in multicenter, large-scale scenarios [125].

Main Findings

LLMs empowered by 5 communication accommodation strategies are pioneering new approaches to health care communication. Through strategic language adaptation, they enable dynamic interaction with patients' needs, demonstrating the potential to bridge gaps in health care resources and improve the quality of care [24]. Current research predominantly focuses on low-level tasks such as text translation and basic consultations, reflecting that LLMs remain in the early stages of integrating into health care communication. Research indicates that simplification and clarification tactics are particularly crucial for addressing the inherent limitations of text-based online consultations [24]. Through these strategies, LLMs simplify medical texts and facilitate cross-language translation, thereby enhancing the transparency of communication between doctors and patients. One study indicates that multimodal LLMs are evolving toward context-aware capabilities [137]. This will help LLMs adjust their communication strategies in real time to achieve "communication adaptation," further enhancing the interpretability, clinical accuracy, and empathy of the text generated by the models.

The emergence of specialty-specific response chatbots highlights LLMs' vast potential to improve accessible medical consultations and empower patients' health management. By incorporating interpersonal control strategies, LLMs are driving a shift in medical interactions from static retrieval to interactive consultation. This strategy is essential for adhering to patient-centered care principles in virtual consultations [138]. In stigmatized domains such as sexually transmitted diseases and mental disorders, LLMs leverage anonymity and accessibility to create low-pressure communication environments [139]. This effectively reduces patients' need for impression management and fear of self-disclosure, encouraging authentic and in-depth expression [140]. As communication intermediaries, LLMs optimize the understanding and sharing of information through approximation strategies, thereby supporting shared decision-making between doctors and patients [103]. Research has shown that when communication methods align with patient needs, the effectiveness of consultations is significantly enhanced [138]. Future research should quantify the substantive impact of LLM interventions on decision conflict and clinician-patient trust. Given insufficient clinical validation, LLMs cannot replace clinical judgment. Still, they should serve as "communication copilots," improving medical communication through collaborative methods such as dynamic dialogue monitoring and refining medical history details [103]. Subsequent studies can focus on developing clinically adapted interactive tools to achieve optimal human-machine collaboration [141].

High-fidelity clinical scenario simulations generated by LLMs offer a low-cost, highly scalable solution for communication skills training. By dynamically simulating diverse patient profiles, LLMs adapt their communication to align with users' language and needs, fostering deeper understanding and connection [24]. However, existing systems are still in their early stages of development, exhibiting limitations in processing nonverbal cues and in addressing privacy and security risks [142]. Consequently, virtual training should be conducted under human supervision, with continuous iteration based on professional feedback to achieve rapid improvement. Future efforts must focus on enhancing the reliability, safety, and scientific validity of virtual patients [143]. Furthermore, relying solely on automated assessment for communication skills may overlook students' psychological and emotional needs [143]. Therefore, establishing a "human-machine collaborative" assessment system is essential to balance teaching efficiency with humanistic care [142,144]. Additionally, the development of customized intervention programs for individuals with language or motor impairments that incorporate interpretability strategies holds profound significance for enhancing the communication autonomy of this marginalized patient population, thereby improving their social participation, vocational integration, and quality of life [112,113]. It is crucial to emphasize that patient-facing tools must be designed within reasonable parameters and built with appropriate safeguards to ensure safety [54].

LLMs demonstrate application value in optimizing clinical workflows, with their core lying in deep integration with clinical dialogue. Preconsultation effectively enhances patient disease awareness [145]. LLMs empowered by discourse management strategies can serve as foundational consulting tools to alleviate time pressures in clinical work and enhance the efficiency of in-person consultations [24]. Research has shown that effective discourse management strategies can maintain conversational coherence and address patient concerns, thereby fostering trust between physicians and patients in digital settings [138].

XSL•FO
RenderX

However, the absence of nonverbal cues can hinder effective communication, necessitating greater linguistic flexibility and highlighting the importance of integrating convergence tactics into communication systems [24]. Interdisciplinary summaries and automated responses generated through such strategies have demonstratedtheir potential to enhance collaborative efficiency. Although LLMs show potential in simulated treatment dialogues, there is currently a lack of empirical evaluation of their effectiveness in real clinical settings [146]. Therefore, LLMs should currently be positioned as clinical adjunct tools. Their successful deployment may be constrained by existing clinical workflows and health care systems' capacity to integrate novel communication tools alongside training resources [147,148]. To fully unlock the medical benefits of LLMs, forward-looking policy frameworks and industry standards must be established to ensure their effective integration with clinical practice.

Evaluation Methods and Dimensions

There is significant heterogeneity in the evaluation of health care communication research using LLMs, making it difficult to accurately assess the models' task performance and practical effectiveness and obscuring their potential in clinical practice.

Currently, expert evaluation remains the primary assessment method in this field, with only a few metrics evaluated using validated tools. Most studies focus on subjective metrics and user experience, lacking objective quantitative metrics and evaluation tools, which limits the comparability of research findings. In response, some scholars have proposed advancing the quantification and structuring of evaluation methods. For example, Huo et al [149] suggested developing and applying quantitative metrics to evaluate model outputs. Furthermore, the lack of a unified, standardized evaluation framework has led to significant heterogeneity in existing evaluation tools and dimensions, further hindering effective comparisons across studies [150]. Wei et al [151] provided insights for establishing evaluation guidelines for LLMs in medical response by integrating factors such as model versions and prompt design. Future work should prioritize developing standardized evaluation frameworks tailored to medical contexts and exploring hybrid assessment methods that combine human expert reviews, user feedback, and automated metrics.

Regarding evaluation content, existing research primarily focuses on content quality, clinical utility, and user experience [146]. However, metrics concerning ethical considerations, such as fairness and bias, remain underevaluated. Bedi et al [146] emphasize that incorporating bias into evaluation frameworks can be an effective way to mitigate harmful biases in LLMs. Traditional model evaluation has primarily focused on the accuracy of medical question-answering tasks; however, due to the lack of objective metrics, it is difficult to comprehensively assess the true effectiveness of LLMs in complex medical communication scenarios [68,152]. Overall, current evaluation practices for LLM-based health care communication lack rigor. More controlled methods are needed to enhance scalability and scientific rigor, such as using validated tools, clearly defining evaluation dimensions, standardizing assessment criteria, and systematically examining changes in patient behavior or clinical outcomes [143]. Therefore, future evaluation systems must balance technical performance, clinical effectiveness, and ethical compliance to establish comprehensive and reliable metrics for the responsible application of LLMs in health care.

Comparison of Performance Differences

Research indicates that performance differences among models stem from heterogeneity in their training datasets, algorithmic architectures, and underlying model capabilities [16,153]. Consistent with previous studies, ChatGPT outperforms most general-purpose models in response accuracy, but its limited readability limits its practical application [154]. However, the success of medical communication depends not only on the accurate transmission of information but also on whether that information can be translated into advice that patients can understand and act upon [155].

Across different clinical specialties, model performance exhibits heterogeneity. Compared to domain-specific fine-tuned models, general-purpose models may exhibit reduced reliability in highly specialized contexts due to a lack of specific training [156]. Taking Dental GPT as an example, this model demonstrates high accuracy and relatively low readability in the field of oral and maxillofacial prosthetics [85]. This aligns with previous research indicating that chatbots trained on domain-specific datasets outperform general-purpose language models [157]. This may be because the model was developed specifically for the dental field; the highly specialized training dataset ensures the model's precise grasp of complex medical facts, though it may also introduce comprehension barriers due to specialized terminology. The CLAIR series of models used by Liu et al [126] demonstrated greater empathy in generating draft patient responses, attributed to the model's fine-tuning on real clinical scenario data. This helps the model mimic clinicians' communication styles and care practices in actual practice, achieving a higher degree of clinical realism [126].

Additionally, linguistic differences can also impact model performance. Recent studies indicate that factors influencing LLM performance include not only model capacity but also linguistic diversity, contextual nuances, and the interaction of multimodal content [158,159]. The diversity of clinical specialties and languages underscores the urgent need to develop specialized, multilingual models [160]. Future research should not be limited to increasing model capacity but should also focus on balancing medical authority with public readability through domain-specific fine-tuning, thereby bridging the gap between professional depth and human communication [161].

Challenges and Future Directions

Overall, we identified existing challenges and recommended ways to address them in the future. A summary of these recommendations is presented in Multimedia Appendix 1.

Concerns regarding the clinical reliability of models arise from issues such as "hallucinations," "black-box nature," and information omissions. Research indicates that the quality of model responses is highly dependent on the training data [162]. Therefore, integrating knowledge graphs with multimodal data to construct standardized training corpora holds promise as an effective solution. Knowledge graphs ensure the accuracy and

XSL•FO
RenderX

security of data sources [163]. Thefusion of multimodal data facilitates comprehensive analysis of patient information, enhancing LLMs' capabilities in complex clinical consultations [164]. Furthermore, given the dynamic evolution of medical knowledge, incorporating retrieval-augmented generation frameworks not only dynamically integrates the latest clinical guidelines and medical evidence to improve the timeliness of model outputs but also enhances system transparency [165].

Limitations in emotional interaction and insufficient depth of expression, as well as legal-ethical concerns, represent significant bottlenecks affecting the social adoption and trust in LLMs. Future research may explore multimodal technologies integrating speech recognition, facial expression analysis, and text comprehension with "human-machine collaboration" models to build context-adaptive, empathetic interaction systems [166]. The HAILEY system, developed by Sharma et al [167], provides empathy-based communication suggestions for physicians, precisely supporting medical interactions. Simultaneously, there is an urgent need to build culturally adapted regional corpora to enhance the representation of marginalized groups and eliminate systemic biases [168]. Additionally, multiple studies emphasize the importance of establishing a comprehensive legal framework [169]. Regarding data governance, strict safeguards for patients' rights to informed consent, data access, and data deletion must be implemented alongside rigorous access control and verification mechanisms [170]. Differential privacy technology, which ensures that individual data remains unidentifiable by introducing controlled noise, represents an effective solution [171].

To overcome interaction and access barriers, future research should focus on developing structured question-assistance frameworks and prompt optimization to guide users in providing structured, complete information, thereby reducing cognitive load during model interactions [172]. Digital inclusion remains paramount, as significant racial, gender, and educational disparities persist in internet access and digital literacy [173]. We therefore advocate for developing training programs to enhance clinicians' human-machine collaboration skills and implementing public digital literacy education to advance health equity.

LLMs face multiple limitations in clinical interactions. Model outputs should adhere to standardized templates and structured guidelines to ensure consistency and completeness [174]. With explicit authorization and strict privacy safeguards, granting secure model access to patient electronic health records may be a strategy for personalization [175]. Additionally, real-time, multitiered risk identification and response mechanisms should be established, integrating multimodal interaction capabilities to detect emotional shifts or psychological issues [119]. Ultimately, establishing channels for human intervention in scenarios involving complex decision-making or deep emotional support will systematically enhance the applicability and safety of models in clinical settings [166].

In summary, unlike previous studies that merely touched upon the challenges and future directions, this study marks the first systematic application of CAT to research on the use of LLMs in health care communication. It provides a comprehensive overview of the current state of applications and prospects in this field. The study lays the groundwork to unlock the potential of LLMs to optimize communication and to promote their responsible use and high-quality development in clinical practice.

Limitations

Several limitations of the scoping review must be acknowledged. First, given the inherent conceptual breadth and interdisciplinary nature of LLMs, coupled with the rapid evolution of related concepts, it remains challenging to completely rule out the possibility of omissions despite the comprehensive retrieval strategy used in this study. This study did not use quantitative measures such as Cohen kappa to assess coding consistency; therefore, it has limitations in reflecting statistical reliability among coders [176]. Additionally, excluding non-English literature may introduce selection bias. Second, most studies originate from high-income countries, such as the United States. This not only influences the scope of ethical discussions and the methods used to address specific issues but also raises questions about the applicability of these findings in low- and middle-income countries. Third, this study aims to provide a comprehensive overview of LLM applications in health care communication. To ensure the comprehensiveness of literature inclusion, we did not conduct a quality assessment of the included studies. Fourth, the rapidly evolving nature of LLMs means our findings primarily reflect the landscape as of the search date; subsequent new models and applications may alter current patterns.

Conclusion

This review is the first to systematically summarize the application patterns of LLMs in health care communication by applying CAT. Currently, LLMs are still in the early stages of integration into clinical practice, and their widespread adoption continues to face challenges, including technical reliability, social trust and acceptance, barriers to interaction and access, and clinical integration. Future research should focus on optimizing model performance, strengthening ethical governance frameworks, and refining human-machine collaboration models, while ensuring safe application in the health care sector through rigorous empirical validation. This study highlights the potential of LLMs to optimize health care communication and is expected to promote their responsible application and high-quality development in medical practice.

Acknowledgments

The authors declare that no generative artificial intelligence (AI) was used in the production of this manuscript.

XSL•FO
RenderX

Funding

The research was supported and funded by the Key Research and Development Program of Hunan Province (2025JK2118).

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Database search strategies, types of studies excluded, characteristics of included studies, and existing challenges.
[DOCX File, 53 KB - jmir_v28i1e84726_app1.docx]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 69 KB - jmir_v28i1e84726_app2.docx]

Checklist 2
PRISMA-S checklist.
[DOCX File, 17 KB - jmir_v28i1e84726_app3.docx]

References

1. Singh M. Communication as a bridge to build a sound doctor-patient/parent relationship. Indian J Pediatr 2016 Jan;83(1):33-37. [doi: 10.1007/s12098-015-1853-9] [Medline: 26228061]
2. Medendorp NM, van den Heuvel LM, Han PKJ, Hillen MA, Smets EMA. Communication skills training for healthcare professionals in providing genetic counseling: a scoping literature review. Patient Educ Couns 2021 Jan;104(1):20-32. [doi: 10.1016/j.pec.2020.07.018] [Medline: 32782179]
3. Doyle C, Lennox L, Bell D. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. BMJ Open 2013 Jan 3;3(1):e001570. [doi: 10.1136/bmjopen-2012-001570] [Medline: 23293244]
4. Riedl D, Schüßler G. The influence of doctor-patient communication on health outcomes: a systematic review. Z Psychosom Med Psychother 2017 Jun;63(2):131-150. [doi: 10.13109/zptm.2017.63.2.131] [Medline: 28585507]
5. Höglander J, Holmström IK, Lövenmark A, Van Dulmen S, Eide H, Sundler AJ. Registered nurse-patient communication research: an integrative review for future directions in nursing research. J Adv Nurs 2023 Feb;79(2):539-562. [doi: 10.1111/jan.15548] [Medline: 36534429]
6. Vermeir P, Vandijck D, Degroote S, et al. Communication in healthcare: a narrative review of the literature and practical recommendations. Int J Clin Pract 2015 Nov;69(11):1257-1267. [doi: 10.1111/ijcp.12686] [Medline: 26147310]
7. Meuter RFI, Gallois C, Segalowitz NS, Ryder AG, Hocking J. Overcoming language barriers in healthcare: a protocol for investigating safe and effective communication when patients or clinicians use a second language. BMC Health Serv Res 2015 Sep 10;15(371):371. [doi: 10.1186/s12913-015-1024-8] [Medline: 26357948]
8. Schillinger D, Duran ND, McNamara DS, Crossley SA, Balyan R, Karter AJ. Precision communication: physicians' linguistic adaptation to patients' health literacy. Sci Adv 2021 Dec 17;7(51):eabj2836. [doi: 10.1126/sciadv.abj2836] [Medline: 34919437]
9. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls: a narrative review. Ann Intern Med 2024 Feb;177(2):210-220. [doi: 10.7326/M23-2772] [Medline: 38285984]
10. Min B, Ross H, Sulem E, et al. Recent advances in natural language processing via large pre-trained language models: a survey. ACM Comput Surv 2024 Feb 29;56(2):1-40. [doi: 10.1145/3605943]
11. Bicknell BT, Butler D, Whalen S, et al. ChatGPT-4 omni performance in USMLE disciplines and clinical skills: comparative analysis. JMIR Med Educ 2024 Nov 6;10:e63430. [doi: 10.2196/63430] [Medline: 39504445]
12. Kaczmarczyk R, Wilhelm TI, Martin R, Roos J. Evaluating multimodal AI in medical diagnostics. NPJ Digit Med 2024 Aug 7;7(1):205. [doi: 10.1038/s41746-024-01208-3] [Medline: 39112822]
13. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
14. Wu T, He S, Liu J, et al. A brief overview of ChatGPT: the history, status quo and potential future development. IEEE/CAA J Autom Sinica 2023;10(5):1122-1136. [doi: 10.1109/JAS.2023.123618]
15. Yu E, Chu X, Zhang W, et al. Large language models in medicine: applications, challenges, and future directions. Int J Med Sci 2025;22(11):2792-2801. [doi: 10.7150/ijms.111780] [Medline: 40520893]

XSL•FO
RenderX

16. Xu X, Chen Y, Miao J. Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review. J Educ Eval Health Prof 2024;21(6):6. [doi: 10.3352/jeehp.2024.21.6] [Medline: 38486402]
17. Ning Y, Ong JCL, Cheng H, et al. How can artificial intelligence transform the training of medical students and physicians? Lancet Digit Health 2025 Oct;7(10):100900. [doi: 10.1016/j.landig.2025.100900] [Medline: 41047321]
18. Li S, Li Y, Zhou S, et al. A community-codesigned LLM-powered chatbot for primary care: a randomized controlled trial. Nat Health 2026;1(2):238-250. [doi: 10.1038/s44360-025-00021-w] [Medline: 41659358]
19. O'Brien LM, Lawless M, Ambagtsheer RC. Health literacy focused communication training for primary healthcare providers working with older adults: a scoping review. Geriatr Nurs 2025;66(Pt A):103601. [doi: 10.1016/j.gerinurse.2025.103601] [Medline: 40829532]
20. Momand B, Dubrowski A. Addressing social context in health provider and senior communication training: what can we learn from communication accommodation theory? Cureus 2020 Dec 23;12(12):e12247. [doi: 10.7759/cureus.12247] [Medline: 33505814]
21. Jechna Z. Physician's prosodic behavior and the regulation of patient anxiety: a quantitative study within the communication accommodation theory framework. Patient Educ Couns 2025 Jan;130(108426):108426. [doi: 10.1016/j.pec.2024.108426] [Medline: 39265453]
22. Niederdeppe J, Boyd AD, King AJ, Rimal RN. Strategies for effective public health communication in a complex information environment. Annu Rev Public Health 2025 Apr;46(1):411-431. [doi: 10.1146/annurev-publhealth-071723-120721] [Medline: 39656948]
23. Bylund CL, Peterson EB, Cameron KA. A practitioner's guide to interpersonal communication theory: an overview and exploration of selected theories. Patient Educ Couns 2012 Jun;87(3):261-267. [doi: 10.1016/j.pec.2011.10.006] [Medline: 22112396]
24. Wang F, Wang J, Hu H, Shi W. When language heals: evaluating patient-centered communication in Chinese telemedicine through communication accommodation theory. Digit Health 2025;11:20552076251411220. [doi: 10.1177/20552076251411220] [Medline: 41473837]
25. Hewett DG, Watson BM, Gallois C, Ward M, Leggett BA. Intergroup communication between hospital doctors: implications for quality of patient care. Soc Sci Med 2009 Dec;69(12):1732-1740. [doi: 10.1016/j.socscimed.2009.09.048] [Medline: 19846246]
26. Fullwood D, Fallon E, Bolajoko O, et al. From unheard to empowered: utilizing communication accommodation theory to address Black men's with prostate cancer healthcare preferences. Cancer Control 2024;31:10732748241293987. [doi: 10.1177/10732748241293987] [Medline: 39865591]
27. Hehl J, McDonald DD. Older adults' pain communication during ambulatory medical visits: an exploration of communication accommodation theory. Pain Manag Nurs 2014 Jun;15(2):466-473. [doi: 10.1016/j.pmn.2012.03.007] [Medline: 24882026]
28. Hemsley B, Balandin S, Worrall L. Nursing the patient with complex communication needs: time as a barrier and a facilitator to successful communication in hospital. J Adv Nurs 2012 Jan;68(1):116-126. [doi: 10.1111/j.1365-2648.2011.05722.x] [Medline: 21831131]
29. Tao X, Zhou S, Ding K, et al. An LLM chatbot to facilitate primary-to-specialist care transitions: a randomized controlled trial. Nat Med 2026 Mar;32(3):934-942. [doi: 10.1038/s41591-025-04176-7] [Medline: 41555035]
30. Cui X, Gu Y, Fang H, Zhu T. Development and evaluation of LLM-based suicide intervention chatbot. Front Psychiatry 2025;16(1634714):1634714. [doi: 10.3389/fpsyt.2025.1634714] [Medline: 40838251]
31. Raghu Subramanian C, Yang DA, Khanna R. Enhancing health care communication with large language models-the role, challenges, and future directions. JAMA Netw Open 2024 Mar 4;7(3):e240347. [doi: 10.1001/jamanetworkopen.2024.0347] [Medline: 38466311]
32. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
33. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
34. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
35. Scoping review protocol. OSF. URL: https://osf.io/7vf2u/files/a3p5z [accessed 2026-05-03]
36. Anand E, Ghersin I, Lingam G, et al. AI-generated patient-friendly MRI fistula summaries: a pilot randomised study. J Imaging 2025 Sep 4;11(9):302. [doi: 10.3390/jimaging11090302] [Medline: 41003353]
37. Berzolla E, Gosnell GG, Chen L, Vonck C, Alaia E, Meislin R. Artificial intelligence large language models improve patient comprehension of radiologist magnetic resonance imaging reports. Arthroscopy 2025 Nov;41(11):4607-4614. [doi: 10.1016/j.arthro.2025.04.033] [Medline: 40288466]
38. Bheemireddy S, Leslie SE, Durden JA, et al. The use of ChatGPT-4.0 to simplify breast pathology reports: a study on readability and accuracy. Ann Surg Oncol 2025 Oct;32(11):8400-8408. [doi: 10.1245/s10434-025-17860-2] [Medline: 40690168]

XSL•FO
RenderX

39. Butler JJ, Harrington MC, Tong Y, et al. From jargon to clarity: improving the readability of foot and ankle radiology reports with an artificial intelligence large language model. Foot Ankle Surg 2024 Jun;30(4):331-337. [doi: 10.1016/j.fas.2024.01.008] [Medline: 38336501]
40. Chen AH, Rudin RS, Levine DM, Mehrotra A. Improving patient understanding of radiology reports using generative artificial intelligence: a vignette study of 2000 US adults. J Am Med Inform Assoc 2026 Feb 1;33(2):326-335. [doi: 10.1093/jamia/ocaf187] [Medline: 41241949]
41. Chung EM, Zhang SC, Nguyen AT, Atkins KM, Sandler HM, Kamrava M. Feasibility and acceptability of ChatGPT generated radiology report summaries for cancer patients. Digit Health 2023;9(2):20552076231221620. [doi: 10.1177/20552076231221620] [Medline: 38130802]
42. Eisinger F, Holderried F, Mahling M, et al. What's going on with me and how can I better manage my health? The potential of GPT-4 to transform discharge letters into patient-centered letters to enhance patient safety: prospective, exploratory study. J Med Internet Res 2025 Jan 21;27:e67143. [doi: 10.2196/67143] [Medline: 39836954]
43. Hains L, Kleinig O, Murugappa A, et al. Large language model discharge summary preparation using real-world electronic medical record data shows promise. Intern Med J 2025 May 28;55(7):1188-1192. [doi: 10.1111/imj.70073] [Medline: 40434141]
44. Khanmammadova N, Gevorkyan R, Afyouni AS, et al. Evaluation of prostate cancer pathology reports generated by ChatGPT to enhance patient comprehension. Sci Rep 2025 Oct 2;15(1):34459. [doi: 10.1038/s41598-025-17639-1] [Medline: 41038938]
45. Kuckelman IJ, Wetley K, Yi PH, Ross AB. Translating musculoskeletal radiology reports into patient-friendly summaries using ChatGPT-4. Skeletal Radiol 2024 Aug;53(8):1621-1624. [doi: 10.1007/s00256-024-04599-2] [Medline: 38270616]
46. Li H, Moon JT, Iyer D, et al. Decoding radiology reports: potential application of OpenAI ChatGPT to enhance patient understanding of diagnostic reports. Clin Imaging 2023 Sep;101:137-141. [doi: 10.1016/j.clinimag.2023.06.008] [Medline: 37336169]
47. Li Y, Li F, Hong N, et al. A comparative study of recent large language models on generating hospital discharge summaries for lung cancer patients. J Biomed Inform 2025 Aug;168:104867. [doi: 10.1016/j.jbi.2025.104867] [Medline: 40544901]
48. Prucker P, Busch F, Dorfner F, et al. Performance of open-source and proprietary large language models in generating patient-friendly radiology chest CT reports. Clin Imaging 2025 Sep;125:110557. [doi: 10.1016/j.clinimag.2025.110557] [Medline: 40639135]
49. Sarangi PK, Lumbani A, Swarup MS, et al. Assessing ChatGPT's proficiency in simplifying radiological reports for healthcare professionals and patients. Cureus 2023 Dec;15(12):e50881. [doi: 10.7759/cureus.50881] [Medline: 38249202]
50. Schmidt S, Zimmerer A, Cucos T, Feucht M, Navas L. Simplifying radiologic reports with natural language processing: a novel approach using ChatGPT in enhancing patient understanding of MRI results. Arch Orthop Trauma Surg 2024 Feb;144(2):611-618. [doi: 10.1007/s00402-023-05113-4] [Medline: 37950763]
51. Steimetz E, Minkowitz J, Gabutan EC, et al. Use of artificial intelligence chatbots in interpretation of pathology reports. JAMA Netw Open 2024 May 1;7(5):e2412767. [doi: 10.1001/jamanetworkopen.2024.12767] [Medline: 38776080]
52. Stephan D, Bertsch AS, Schumacher S, et al. Improving patient communication by simplifying AI-generated dental radiology reports with ChatGPT: comparative study. J Med Internet Res 2025 Jun 9;27:e73337. [doi: 10.2196/73337] [Medline: 40489773]
53. Sunshine A, Honce GH, Callen AL, et al. Evaluating the quality and understandability of radiology report summaries generated by ChatGPT: survey study. JMIR Form Res 2025 Aug 27;9:e76097. [doi: 10.2196/76097] [Medline: 40864905]
54. Tang CC, Nagesh S, Fussell DA, et al. Generating colloquial radiology reports with large language models. J Am Med Inform Assoc 2024 Nov 1;31(11):2660-2667. [doi: 10.1093/jamia/ocae223] [Medline: 39178375]
55. Williams CYK, Subramanian CR, Ali SS, et al. Physician- and large language model-generated hospital discharge summaries. JAMA Intern Med 2025 Jul 1;185(7):818-825. [doi: 10.1001/jamainternmed.2025.0821] [Medline: 40323616]
56. Yang X, Xiao Y, Liu D, et al. Enhancing physician-patient communication in oncology using GPT-4 through simplified radiology reports: multicenter quantitative study. J Med Internet Res 2025 Apr 17;27:e63786. [doi: 10.2196/63786] [Medline: 40245397]
57. Yang X, Xiao Y, Liu D, et al. Enhancing doctor-patient communication using large language models for pathology report interpretation. BMC Med Inform Decis Mak 2025 Jan 23;25(1):36. [doi: 10.1186/s12911-024-02838-z] [Medline: 39849504]
58. Zaretsky J, Kim JM, Baskharoun S, et al. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. JAMA Netw Open 2024 Mar 4;7(3):e240357. [doi: 10.1001/jamanetworkopen.2024.0357] [Medline: 38466307]
59. Vallurupalli M, Shah ND, Vyas RM. Optimizing readability of patient-facing hand surgery education materials using chat generative pretrained transformer (ChatGPT) 3.5. J Hand Surg Am 2024 Oct;49(10):986-991. [doi: 10.1016/j.jhsa.2024.05.007] [Medline: 38970600]
60. Chandra K, Ghilzai U, Lawand J, Ghali A, Fiedler B, Ahmed AS. Improving readability of shoulder and elbow surgery online patient education material with Chat GPT (Chat Generative Pretrained Transformer) 4. J Shoulder Elbow Surg 2025 Nov;34(11):e1119-e1124. [doi: 10.1016/j.jse.2025.02.025] [Medline: 40118438]

XSL•FO
RenderX

61. Gondode PG, Khanna P, Sharma P, Duggal S, Garg N. End-of-life care patient information leaflets-a comparative evaluation of artificial intelligence-generated content for readability, sentiment, accuracy, completeness, and suitability: ChatGPT vs Google Gemini. Indian J Crit Care Med 2024 Jun;28(6):561-568. [doi: 10.5005/jp-journals-10071-24725] [Medline: 39130387]
62. Kianian R, Carter M, Finkelshtein I, Eleswarapu SV, Kachroo N. Application of artificial intelligence to patient-targeted health information on kidney stone disease. J Ren Nutr 2024 Mar;34(2):170-176. [doi: 10.1053/j.jrn.2023.10.002] [Medline: 37839591]
63. Juan-Guardela EJJ, Beltrán-España JA, Ravagli-Baquero MP, et al. Reliability and readability assessment of atrial fibrillation patient information delivered by artificial intelligence-based language models (ChatGPT, YouChat, Gemini, and Perplexity AI) in English and Spanish. Clin Med Insights Cardiol 2025;19:11795468251383666. [doi: 10.1177/11795468251383666] [Medline: 41180819]
64. Dias R, Castan A, Gotoff K, et al. ChatGPT 3.5 better improves comprehensibility of English, than Spanish, generated responses to osteosarcoma questions. J Surg Oncol 2025 Jun;131(8):1692-1695. [doi: 10.1002/jso.28109] [Medline: 39898783]
65. Gulati V, Roy SG, Moawad A, et al. Transcending language barriers: can ChatGPT be the key to enhancing multilingual accessibility in health care? J Am Coll Radiol 2024 Dec;21(12):1888-1895. [doi: 10.1016/j.jacr.2024.05.009] [Medline: 38880289]
66. Chandler RD, Warner S, Guillaume D, Wells J. Exploring ChatGPT's role in women's health self-education: a descriptive study comparing responses with public health guidance. Nurs Outlook 2025;73(4):102468. [doi: 10.1016/j.outlook.2025.102468] [Medline: 40614295]
67. Pap IA, Oniga S. eHealth assistant AI chatbot using a large language model to provide personalized answers through secure decentralized communication. Sensors (Basel) 2024 Sep 23;24(18):6140. [doi: 10.3390/s24186140] [Medline: 39338885]
68. Akçay O, Öztürk Ö, Acar T, Gürsoy S. Accuracy and reliability of ChatGPT in answering patient questions about lung cancer and its surgery: an expert panel evaluation by thoracic surgeons. J Cancer Educ 2025 Jul 4. [doi: 10.1007/s13187-025-02682-3] [Medline: 40615772]
69. Belge Bilgin G, Bilgin C, Childs DS, et al. Performance of ChatGPT-4 and Bard chatbots in responding to common patient questions on prostate cancer 177Lu-PSMA-617 therapy. Front Oncol 2024;14:1386718. [doi: 10.3389/fonc.2024.1386718] [Medline: 39070149]
70. Demir Cicek B, Cicek O. Evaluating the response of AI-based large language models to common patient concerns about endodontic root canal treatment: a comparative performance analysis. J Clin Med 2025 Oct 22;14(21):7482. [doi: 10.3390/jcm14217482] [Medline: 41226878]
71. Doğan L, Yılmaz İ. The performance of ChatGPT-4 and Bing Chat in frequently asked questions about glaucoma. Eur J Ophthalmol 2025 Jul;35(4):1323-1328. [doi: 10.1177/11206721251321197] [Medline: 39973162]
72. Gencer B, Arzu U, Orhan SS, Dinçal T, Ekinci M. Evaluation of ChatGPT responses about sexual activity after total hip arthroplasty: a comparative study with observers of different experience levels. J Clin Med 2025 Apr 24;14(9):2942. [doi: 10.3390/jcm14092942] [Medline: 40363974]
73. Goshtasbi K, Best C, Powers B, et al. Comparative performance of the leading large language models in answering complex rhinoplasty consultation questions. Facial Plast Surg Aesthet Med 2025;27(4):378-383. [doi: 10.1089/fpsam.2024.0206] [Medline: 39812574]
74. Gültekin O, Inoue J, Yilmaz B, et al. Evaluating DeepResearch and DeepThink in anterior cruciate ligament surgery patient education: ChatGPT-4o excels in comprehensiveness, DeepSeek R1 leads in clarity and readability of orthopaedic information. Knee Surg Sports Traumatol Arthrosc 2025 Aug;33(8):3025-3031. [doi: 10.1002/ksa.12711] [Medline: 40450565]
75. Guven Y, Ozdemir OT, Kavan MY. Performance of artificial intelligence chatbots in responding to patient queries related to traumatic dental injuries: a comparative study. Dent Traumatol 2025 Jun;41(3):338-347. [doi: 10.1111/edt.13020] [Medline: 39578674]
76. Ibas M, Dursun S, Paksoy M, Ocal R, Karatas E. Accuracy and safety of ChatGPT-4o responses in rhinoplasty postoperative counseling: a panel-based study. Acta Otolaryngol 2025 Sep;145(9):851-856. [doi: 10.1080/00016489.2025.2541612] [Medline: 40772566]
77. Kamminga NCW, Kievits JEC, Plaisier PW, et al. Do large language model chatbots perform better than established patient information resources in answering patient questions? A comparative study on melanoma. Br J Dermatol 2025 Jan 24;192(2):306-315. [doi: 10.1093/bjd/ljae377] [Medline: 39365602]
78. Kuo FH, Fierstein JL, Tudor BH, et al. Comparing ChatGPT and a single anesthesiologist's responses to common patient questions: an exploratory cross-sectional survey of a panel of anesthesiologists. J Med Syst 2024 Aug 22;48(1):77. [doi: 10.1007/s10916-024-02100-z] [Medline: 39172169]
79. Lewandowski M, Łukowicz P, Świetlik D, Jendrysik B, Barańska-Rybak W. Comparing physician and artificial intelligence ChatGPT-4 responses to common patient questions regarding hidradenitis suppurativa: a single-blind study. Postepy Dermatol Alergol 2025;42(5):460-466. [doi: 10.5114/ada.2025.151910] [Medline: 41262860]

XSL•FO
RenderX

80. Li AW, Adelstein JM, Li LT, Sinkler MA, Mistovich RJ. Assessing ChatGPT responses to frequently asked questions regarding pediatric supracondylar humerus fractures. J Pediatr Orthop 2025 Jul 1;45(6):327-331. [doi: 10.1097/BPO.0000000000002923] [Medline: 39917871]
81. Guo S, Li G, Gou J, et al. The performance of ChatGPT-4.0 and ChatGPT-4omni on answering thyroid question: a multicenter study. J Surg Res 2025 Sep;313:500-508. [doi: 10.1016/j.jss.2025.06.066] [Medline: 40706228]
82. Mete U, Özmen Ö. Assessing the accuracy and reproducibility of ChatGPT for responding to patient inquiries about otosclerosis. Eur Arch Otorhinolaryngol 2025 Mar;282(3):1567-1575. [doi: 10.1007/s00405-024-09039-4] [Medline: 39461921]
83. Moothedan E, Jhumkhawala V, Burgoa S, Martinez L, Sacca L. Response to “Qualitatively Assessing ChatGPT Responses to Frequently Asked Questions Regarding Sexually Transmitted Diseases: Considerations”. Sex Transm Dis 2025 Apr 1;52(4):e11-e11. [doi: 10.1097/OLQ.0000000000002121] [Medline: 39681087]
84. Motegi M, Shino M, Kuwabara M, et al. Comparison of physician and large language model chatbot responses to online ear, nose, and throat inquiries. Sci Rep 2025 Jul 1;15(1):21346. [doi: 10.1038/s41598-025-06769-1] [Medline: 40596359]
85. Özcivelek T, Özcan B. Comparative evaluation of responses from DeepSeek-R1, ChatGPT-o1, ChatGPT-4, and dental GPT chatbots to patient inquiries about dental and maxillofacial prostheses. BMC Oral Health 2025 May 31;25(1):871. [doi: 10.1186/s12903-025-06267-w] [Medline: 40450291]
86. Raghunathan R, Jacobs AR, Sant VR, et al. Can large language models address unmet patient information needs and reduce provider burnout in the management of thyroid disease? Surgery 2025 Jan;177:108859. [doi: 10.1016/j.surg.2024.06.075] [Medline: 39424485]
87. Rotem R, Simon C, Rottenstreich M, O’reilly BA, Weintraub AY, O’Sullivan OE. ChatGPT in urogynecology: comparing large language model responses to human experts. Acta Obstet Gynecol Scand 2025 Oct 23. [doi: 10.1111/aogs.70085] [Medline: 41128396]
88. Roy JM, Atallah E, Piper K, et al. Comparison of quality, empathy and readability of physician responses versus chatbot responses to common cerebrovascular neurosurgical questions on a social media platform. Clin Neurol Neurosurg 2025 Aug;255:108986. [doi: 10.1016/j.clineuro.2025.108986] [Medline: 40451125]
89. Şan H, Bayrakcı Ö, Çağdaş B, Serdengeçti M, Alagöz E. Reliability and readability analysis of ChatGPT-4 and Google Bard as a patient information source for the most commonly applied radionuclide treatments in cancer patients. Rev Esp Med Nucl Imagen Mol (Engl Ed) 2024;43(4):500021. [doi: 10.1016/j.remnie.2024.500021] [Medline: 38821410]
90. Scaff SPS, Reis FJJ, Ferreira GE, Jacob MF, Saragiotto BT. Assessing the performance of AI chatbots in answering patients’ common questions about low back pain. Ann Rheum Dis 2025 Jan;84(1):143-149. [doi: 10.1136/ard-2024-226202] [Medline: 39874229]
91. Shao XR, Ruan T, Ju XG, Sun YH, Cui JC. Evaluating artificial intelligence chatbots’ responses to gynecomastia inquiries: comparative study of information quality, readability, and guideline consistency. Digit Health 2025;11:20552076251367645. [doi: 10.1177/20552076251367645] [Medline: 40881062]
92. Tuzlalı M, Baki N, Aral K, Aral CA, Bahçe E. Evaluating the performance of AI chatbots in responding to dental implant FAQs: a comparative study. BMC Oral Health 2025 Oct 8;25(1):1548. [doi: 10.1186/s12903-025-06863-w] [Medline: 41063105]
93. Weber MT, Noll R, Marchl A, et al. MedBot vs RealDoc: efficacy of large language modeling in physician-patient communication for rare diseases. J Am Med Inform Assoc 2025 May 1;32(5):775-783. [doi: 10.1093/jamia/ocaf034] [Medline: 39998911]
94. White CA, Masturov YA, Haunschild E, Michaelson E, Shukla DR, Cagle PJ. Can ChatGPT reliably answer the most common patient questions regarding total shoulder arthroplasty? J Shoulder Elbow Surg 2025 May;34(5):e254-e264. [doi: 10.1016/j.jse.2024.08.025] [Medline: 39419373]
95. Xue D, Liao S. Evaluation of three large language models’ response performances to inquiries regarding post-abortion care in the context of Chinese language: a comparative analysis. Risk Manag Healthc Policy 2025;18:2731-2741. [doi: 10.2147/RMHP.S531777] [Medline: 40862290]
96. Zhang S, Chua IJ, Liang ZC, Pang KC. Can ChatGPT provide personalised responses for frequently asked questions regarding anterior cruciate ligament reconstruction? Knee 2025 Oct;56:495-502. [doi: 10.1016/j.knee.2025.06.018] [Medline: 40680504]
97. He Z, Liu Y, Wang Z, et al. Revolutionizing inflammatory bowel disease healthcare communication: a head-to-head comparison of gastroenterologist and ChatGPT responses. Gastroenterol Rep (Oxf) 2025;13:goaf087. [doi: 10.1093/gastro/goaf087] [Medline: 41035903]
98. Wu H, Li M, Zhang L. Comparing physician and large language model responses to influenza patient questions in the online health community. Int J Med Inform 2025 May;197:105836. [doi: 10.1016/j.ijmedinf.2025.105836] [Medline: 39986122]
99. Muasher-Kerwin C, Hughes MC, Foster ML, Al Azher I, Alhoori H. Exploring large language models for summarizing and interpreting an online brain tumor support forum. Digit Health 2025;11:20552076251337345. [doi: 10.1177/20552076251337345] [Medline: 40297356]

XSL•FO
RenderX

100. Bertges DJ, Beck AW, Schermerhorn M, et al. Testing ChatGPT's ability to provide patient and physician information on aortic aneurysm. J Surg Res 2025 Mar;307:129-138. [doi: 10.1016/j.jss.2025.01.015] [Medline: 40014909]
101. Allen JW, Levy N, Wilkinson D. Empowering patient autonomy: the role of large language models (LLMs) in scaffolding informed consent in medical practice. Bioethics 2026 Feb;40(2):183-193. [doi: 10.1111/bioe.70030] [Medline: 40931965]
102. Grünebaum A, Dudenhausen J, Chervenak FA. Enhancing patient understanding in obstetrics: the role of generative AI in simplifying informed consent for labor induction with oxytocin. J Perinat Med 2025 Jul 28;53(6):688-695. [doi: 10.1515/jpm-2024-0428] [Medline: 39470098]
103. Kienzle A, Niemann M, Meller S, Gwinner C. ChatGPT may offer an adequate substitute for informed consent to patients prior to total knee arthroplasty-yet caution is needed. J Pers Med 2024 Jan 5;14(1):69. [doi: 10.3390/jpm14010069] [Medline: 38248771]
104. Bakhaya S, Lehnbom EC, de Carvalho Filho MA, Ma KY, Svensberg K. Development and evaluation of AI chatbot tool for written communication training in self-care: experiences of pharmacy students and faculty. Curr Pharm Teach Learn 2026 Jan;18(1):102503. [doi: 10.1016/j.cptl.2025.102503] [Medline: 41061515]
105. Chen PJ. Effectiveness of integrating generative artificial intelligence with virtual reality for maternity communication simulation: a randomized controlled trial. Clin Simul Nurs 2025 Aug;105:101786. [doi: 10.1016/j.ecns.2025.101786]
106. Chen PJ, Liou WK. ChatGPT-driven interactive virtual reality communication simulation in obstetric nursing: a mixed-methods study. Nurse Educ Pract 2025 May;85:104383. [doi: 10.1016/j.nepr.2025.104383] [Medline: 40315616]
107. Gutiérrez Maquilón R, Uhl J, Schrom-Feiertag H, Tscheligi M. Integrating GPT-based AI into virtual patients to facilitate communication training among medical first responders: usability study of mixed reality simulation. JMIR Form Res 2024 Dec 11;8:e58623. [doi: 10.2196/58623] [Medline: 39661979]
108. Kim J, Won J, Lee Y. Use of a generative pre-trained transformer-based virtual patient for health assessment and communication training in nursing education: a mixed-methods study. Nurse Educ Pract 2025 Oct;88:104536. [doi: 10.1016/j.nepr.2025.104536] [Medline: 40902283]
109. Wang Z, Fan TT, Li ML, Zhu NJ, Wang XC. Feasibility study of using GPT for history-taking training in medical education: a randomized clinical trial. BMC Med Educ 2025 Jul 10;25(1):1030. [doi: 10.1186/s12909-025-07614-9] [Medline: 40640776]
110. Weisman D, Sugarman A, Huang YM, Gelberg L, Ganz PA, Comulada WS. Development of a GPT-4-powered virtual simulated patient and communication training platform for medical students to practice discussing abnormal mammogram results with patients: multiphase study. JMIR Form Res 2025 Apr 17;9:e65670. [doi: 10.2196/65670] [Medline: 40246299]
111. Chen H, Zeng W, Chen C, et al. EEG emotion copilot: optimizing lightweight LLMs for emotional EEG interpretation with assisted medical record generation. Neural Netw 2025 Dec;192:107848. [doi: 10.1016/j.neunet.2025.107848] [Medline: 40683189]
112. Adikari A, Alahakoon D, Pallewela N, Pierce JE, Hernandez NJ, Rose ML. Reconstructing impaired language using generative AI for people with aphasia. Sci Rep 2025 Nov 19;15(1):40877. [doi: 10.1038/s41598-025-24725-x] [Medline: 41258473]
113. Cai S, Venugopalan S, Seaver K, et al. Using large language models to accelerate communication for eye gaze typing users with ALS. Nat Commun 2024 Nov 1;15(1):9449. [doi: 10.1038/s41467-024-53873-3] [Medline: 39487163]
114. Wan P, Huang Z, Tang W, et al. Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. Nat Med 2024 Oct;30(10):2878-2885. [doi: 10.1038/s41591-024-03148-7] [Medline: 39009780]
115. Carl N, Haggenmüller S, Wies C, et al. Evaluating interactions of patients with large language models for medical information. BJU Int 2025 Jun;135(6):1010-1017. [doi: 10.1111/bju.16676] [Medline: 39967059]
116. Chung D, Dhillon H, Sidhom K, Bal D, Patel P, Jawanda G. (096) Real world utility of chatgpt in pre-vasectomy counselling in an office-based setting: a pilot study. J Sex Med 2024 Aug 12;21(7). [doi: 10.1093/jsxmed/qdae167.094]
117. Park C, An MH, Hwang G, Park RW, An J. Clinical performance and communication skills of ChatGPT versus physicians in emergency medicine: simulated patient study. JMIR Med Inform 2025 Jul 17;13:e68409. [doi: 10.2196/68409] [Medline: 40674718]
118. Santonocito S, Cicciù M, Ronsivalle V. Evaluation of the impact of AI-based chatbot on orthodontic patient education: a preliminary randomised controlled trial. Clin Oral Investig 2025 Apr 30;29(5):278. [doi: 10.1007/s00784-025-06356-8] [Medline: 40304793]
119. Scholich T, Barr M, Wiltsey Stirman S, Raj S. A comparison of responses from human therapists and large language model-based chatbots to assess therapeutic communication: mixed methods study. JMIR Ment Health 2025 May 21;12:e69709. [doi: 10.2196/69709] [Medline: 40397927]
120. Warin K, Taetragoo U, Trachoo V, et al. AI chatbot for post-operative oral surgery information and support. Stud Health Technol Inform 2025 Aug 7;329:1938-1939. [doi: 10.3233/SHTI251288] [Medline: 40776305]
121. Chen C, Hu Y, Cai W, et al. Deep learning-based in-ambulance speech recognition and generation of prehospital emergency diagnostic summaries using LLMs. Int J Med Inform 2025 Nov;203:106029. [doi: 10.1016/j.ijmedinf.2025.106029] [Medline: 40639122]
122. Genes N, Simon G, Koziatek C, et al. Generative artificial intelligence summaries to facilitate emergency department handoff. Appl Clin Inform 2025 Aug;16(4):1185-1191. [doi: 10.1055/a-2681-5008] [Medline: 40795949]

XSL•FO
RenderX

123. Tailor PD, D'Souza HS, Castillejo Becerra CM, et al. Evaluation of AI summaries on interdisciplinary understanding of ophthalmology notes. JAMA Ophthalmol 2025 May 1;143(5):410-419. [doi: 10.1001/jamaophthalmol.2025.0351] [Medline: 40178837]
124. Garcia P, Ma SP, Shah S, et al. Artificial intelligence-generated draft replies to patient inbox messages. JAMA Netw Open 2024 Mar 4;7(3):e243201. [doi: 10.1001/jamanetworkopen.2024.3201] [Medline: 38506805]
125. Kaur A, Budko A, Liu K, Eaton E, Steitz BD, Johnson KB. Automating responses to patient portal messages using generative AI. Appl Clin Inform 2025 May;16(3):718-731. [doi: 10.1055/a-2565-9155] [Medline: 40132987]
126. Liu S, McCoy AB, Wright AP, et al. Leveraging large language models for generating responses to patient messages-a subjective analysis. J Am Med Inform Assoc 2024 May 20;31(6):1367-1379. [doi: 10.1093/jamia/ocae052] [Medline: 38497958]
127. Scott M, Muncey W, Seranio N, et al. Assessing artificial intelligence-generated responses to urology patient in-basket messages. Urol Pract 2024 Sep;11(5):793-798. [doi: 10.1097/UPJ.0000000000000637] [Medline: 39162591]
128. Small WR, Wiesenfeld B, Brandfield-Harvey B, et al. Large language model-based responses to patients' in-basket messages. JAMA Netw Open 2024 Jul 1;7(7):e2422399. [doi: 10.1001/jamanetworkopen.2024.22399] [Medline: 39012633]
129. Soroudi D, Gozali A, Knox JA, et al. Comparing provider and ChatGPT responses to breast reconstruction patient questions in the electronic health record. Ann Plast Surg 2024 Nov 1;93(5):541-545. [doi: 10.1097/SAP.0000000000004090] [Medline: 39445873]
130. English E, Laughlin J, Sippel J, DeCamp M, Lin CT. Utility of artificial intelligence-generative draft replies to patient messages. JAMA Netw Open 2024 Oct 1;7(10):e2438573. [doi: 10.1001/jamanetworkopen.2024.38573] [Medline: 39401041]
131. Mandal S, Wiesenfeld BM, Szerencsy AC, et al. Utilization of generative AI-drafted responses for managing patient-provider communication. NPJ Digit Med 2025 Oct 2;8(1):591. [doi: 10.1038/s41746-025-01972-w] [Medline: 41038966]
132. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on May 24, 2019 URL: https://doi.org/10.48550/arXiv.1810.04805 [accessed 2026-03-20]
133. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008 Jul 10;8(45):45. [doi: 10.1186/1471-2288-8-45] [Medline: 18616818]
134. Sezgin E. Artificial intelligence in healthcare: complementing, not replacing, doctors and healthcare providers. Digit Health 2023;9:20552076231186520. [doi: 10.1177/20552076231186520] [Medline: 37426593]
135. De Angelis L, Baglivo F, Arzilli G, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health 2023;11:1166120. [doi: 10.3389/fpubh.2023.1166120] [Medline: 37181697]
136. Holmes G, Tang B, Gupta S, Venkatesh S, Christensen H, Whitton A. Applications of large language models in the field of suicide prevention: scoping review. J Med Internet Res 2025 Jan 23;27:e63126. [doi: 10.2196/63126] [Medline: 39847414]
137. Al Zaabi A, Alshibli R, AlAmri A, AlRuheili I, Lutfi SL. Trends and trajectories in the rise of large language models in radiology: scoping review. JMIR Med Inform 2025 Dec 9;13:e78041. [doi: 10.2196/78041] [Medline: 41364806]
138. Zolnierek KBH, Dimatteo MR. Physician communication and patient adherence to treatment: a meta-analysis. Med Care 2009 Aug;47(8):826-834. [doi: 10.1097/MLR.0b013e31819a5acc] [Medline: 19584762]
139. Shao X, Ruan T, Ju X, Sun Y, Cui J. Evaluating artificial intelligence chatbots' responses to gynecomastia inquiries: comparative study of information quality, readability, and guideline consistency. Digit Health 2025;11:20552076251367645. [doi: 10.1177/20552076251367645] [Medline: 40881062]
140. Reis M, Reis F, Kunde W. Influence of believed AI involvement on the perception of digital medical advice. Nat Med 2024 Nov;30(11):3098-3100. [doi: 10.1038/s41591-024-03180-7] [Medline: 39054373]
141. Stade EC, Stirman SW, Ungar LH, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. Npj Ment Health Res 2024 Apr 2;3(1):12. [doi: 10.1038/s44184-024-00056-z] [Medline: 38609507]
142. Lin Y, Luo Z, Ye Z, et al. Applications, challenges, and prospects of generative artificial intelligence empowering medical education: scoping review. JMIR Med Educ 2025 Oct 23;11:e71125. [doi: 10.2196/71125] [Medline: 41128430]
143. Zeng J, Qi W, Shen S, et al. Embracing the future of medical education with large language model-based virtual patients: scoping review. J Med Internet Res 2025 Nov 13;27:e79091. [doi: 10.2196/79091] [Medline: 41232097]
144. Yu S, Lee SS, Hwang H. The ethics of using artificial intelligence in medical research. KMJ 2024 Dec;39(4):229-237. [doi: 10.7180/kmj.24.140]
145. Qian H, Dong B, Yuan JJ, et al. Pre-consultation system based on the artificial intelligence has a better diagnostic performance than the physicians in the outpatient department of pediatrics. Front Med (Lausanne) 2021;8:695185. [doi: 10.3389/fmed.2021.695185] [Medline: 34820391]
146. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 2025 Jan 28;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
147. Bayor AA, Li J, Yang IA, Varnfield M. Designing clinical decision support systems (CDSS)-a user-centered lens of the design characteristics, challenges, and implications: systematic review. J Med Internet Res 2025 Jun 20;27:e63733. [doi: 10.2196/63733] [Medline: 40540451]
148. Jung K, Kashyap S, Avati A, et al. A framework for making predictive models useful in practice. J Am Med Inform Assoc 2021 Jun 12;28(6):1149-1158. [doi: 10.1093/jamia/ocaa318] [Medline: 33355350]

XSL•FO
RenderX

149. Huo B, Boyle A, Marfo N, et al. Large language models for chatbot health advice studies: a systematic review. JAMA Netw Open 2025 Feb 3;8(2):e2457879. [doi: 10.1001/jamanetworkopen.2024.57879] [Medline: 39903463]
150. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Med Inform Decis Mak 2025 Mar 7;25(1):117. [doi: 10.1186/s12911-025-02954-4] [Medline: 40055694]
151. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. J Biomed Inform 2024 Mar;151:104620. [doi: 10.1016/j.jbi.2024.104620] [Medline: 38462064]
152. Doshi R, Amin K, Khosla P, Bajaj S, Chheang S, Forman HP. Utilizing large language models to simplify radiology reports: a comparative analysis of ChatGPT3.5, ChatGPT4.0, Google Bard, and Microsoft Bing. Radiology and Imaging. Preprint posted online on Jun 7, 2023. [doi: 10.1101/2023.06.04.23290786]
153. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. J Dent Res 2020 Jul;99(7):769-774. [doi: 10.1177/0022034520915714] [Medline: 32315260]
154. Jiang C, Cheng W, Jiang X, Zhang J, Tang X. A comparative study of ChatGPT-4o and DeepSeek responses to mandibular angle osteotomy questions. J Craniofac Surg 2025 Jul 31;36(7):e1113-e1116. [doi: 10.1097/SCS.0000000000011698] [Medline: 40742907]
155. Forsey J, Ng S, Rowland P, Freeman R, Li C, Woods NN. The basic science of patient-physician communication: a critical scoping review. Acad Med 2021 Nov 1;96(11S):S109-S118. [doi: 10.1097/ACM.0000000000004323] [Medline: 34348382]
156. Zheng S, Zhao N, Wang J, et al. Comparison of a specialized large language model with GPT-4o for CT and MRI radiology report summarization. Radiology 2025 Aug;316(2):e243774. [doi: 10.1148/radiol.243774] [Medline: 40892451]
157. Lehman EP, Hernandez E, Mahajan D, et al. Do we still need clinical language models? Presented at: Conference on Health, Inference, and Learning (CHIL) 2023; Jun 22-24, 2023. [doi: 10.48550/arXiv.2302.08091]
158. Zhang Q, Wu Z, Song J, Luo S, Chai Z. Comprehensiveness of large language models in patient queries on gingival and endodontic health. Int Dent J 2025 Feb;75(1):151-157. [doi: 10.1016/j.identj.2024.06.022] [Medline: 39147663]
159. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023 Mar 4;47(1):33. [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
160. Luo PW, Liu JW, Xie X, et al. DeepSeek vs ChatGPT: a comparison study of their performance in answering prostate cancer radiotherapy questions in multiple languages. Am J Clin Exp Urol 2025;13(2):176-185. [doi: 10.62347/UIAP7979] [Medline: 40400997]
161. Yıldırım A, Cicek O, Genç YS. Can AI-based ChatGPT models accurately analyze hand-wrist radiographs? a comparative study. Diagnostics (Basel) 2025 Jun 14;15(12):1513. [doi: 10.3390/diagnostics15121513] [Medline: 40564836]
162. Chen YQ, Zhang SX. Superior resilience to poisoning and amenability to unlearning in quantum machine learning. Nat Commun 2026 Mar 9;17(1):3716. [doi: 10.1038/s41467-026-70420-4] [Medline: 41803125]
163. Yang L, Chen H, Li Z, Ding X, Wu X. Give us the facts: enhancing large language models with knowledge graphs for fact-aware language modeling. IEEE Trans Knowl Data Eng 2024;36(7):3091-3110. [doi: 10.1109/TKDE.2024.3360454] [Medline: 25079929]
164. Malgaroli M, Hull TD, Zech JM, Althoff T. Natural language processing for mental health interventions: a systematic review and research framework. Transl Psychiatry 2023 Oct 6;13(1):309. [doi: 10.1038/s41398-023-02592-2] [Medline: 37798296]
165. Quer G, Topol EJ. The potential for large language models to transform cardiovascular medicine. Lancet Digit Health 2024 Oct;6(10):e767-e771. [doi: 10.1016/S2589-7500(24)00151-1] [Medline: 39214760]
166. Qian Y, Tang W, Xu T, Zhu D, Hua W. Emotional support powered by artificial intelligence in healthcare settings: a scoping review of technologies, contents, and outcomes. Int J Nurs Stud 2025 Dec;172:105243. [doi: 10.1016/j.ijnurstu.2025.105243] [Medline: 41092641]
167. Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nat Mach Intell 2023;5(1):46-57. [doi: 10.1038/s42256-022-00593-2]
168. Chen M, Li C, Sitaram S, Wang J, Xie X. CultureLLM: incorporating cultural differences into large language models. Presented at: Advances in Neural Information Processing Systems 37; Dec 10-15, 2024. [doi: 10.52202/079017-2693]
169. Shumway DO, Hartman HJ. Medical malpractice liability in large language model artificial intelligence: legal review and policy recommendations. J Osteopath Med 2024 Jul 1;124(7):287-290. [doi: 10.1515/jom-2023-0229] [Medline: 38295300]
170. Ong JCL, Chang SYH, William W, et al. Ethical and regulatory challenges of large language models in medicine. Lancet Digit Health 2024 Jun;6(6):e428-e432. [doi: 10.1016/S2589-7500(24)00061-X] [Medline: 38658283]
171. Alzoubi YI, Mishra A. Differential privacy and artificial intelligence: potentials, challenges, and future avenues. EURASIP J Inf Sec 2025;2025(1). [doi: 10.1186/s13635-025-00203-9]
172. Wang M, Liu Y, Liang X, Li S, Huang Y, Zhang X, et al. LangGPT: rethinking structured reusable prompt design framework for llms from the programming language. arXiv. Preprint posted online on Jun 29, 2024. [doi: 10.48550/arXiv.2402.16929]
173. Perret S, Alon N, Carpenter-Song E, et al. Standardising the role of a digital navigator in behavioural health: a systematic review. Lancet Digit Health 2023 Dec;5(12):e925-e932. [doi: 10.1016/S2589-7500(23)00152-8] [Medline: 38000876]
174. Wang Z, Gao J, Danek B, et al. Compliance and factuality of large language models for clinical research document generation. J Am Med Inform Assoc 2026 Mar 1;33(3):563-572. [doi: 10.1093/jamia/ocaf174] [Medline: 41144289]

XSL•FO
RenderX

175. Hao Y, Holmes J, Waddle MR, et al. Personalizing prostate cancer education for patients using an EHR-integrated LLM agent. NPJ Digit Med 2025 Dec 18;8(1):770. [doi: 10.1038/s41746-025-02166-0] [Medline: 41413170]
176. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282. [Medline: 23092060]

Abbreviations

AI: artificial intelligence
CAT: communication accommodation theory
LLM: large language model
MeSH: Medical Subject Headings
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by S Brini; submitted 25.Sep.2025; peer-reviewed by B Eapen, D Chrimes; revised version received 04.Apr.2026; accepted 06.Apr.2026; published 26.Jun.2026.

Please cite as:
Chang J, Peng R, Chen X, Zhu Y, Miao R, Cao Z, Feng H
Applications, Challenges, and Future Directions of Large Language Models in Health Care Communication: Scoping Review
J Med Internet Res 2026;28:e84726
URL: https://www.jmir.org/2026/1/e84726
doi:10.2196/84726

© Jing Chang, Ruotong Peng, Xi Chen, Yishu Zhu, Ruting Miao, Zeng Cao, Hui Feng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Physical Activity Interventions Using Digital Health Interventions for Cancer-Related Fatigue in People With a History of Cancer: Scoping Review

Yeeun Kim[1,2], MSN; Ka Ryeong Bae[3], PhD; Ji Hyun Sung[4], PhD; Yun Hee Ko[1], PhD; Sue Kim[5], PhD

[1]College of Nursing, Yonsei University, Seoul, Republic of Korea

[2]School of Nursing, University of Texas at Austin, Austin, TX, United States

[3]College of Nursing, Eulji University, Seongnam, Republic of Korea

[4]College of Nursing, Keimyung University, Daegu, Republic of Korea

[5]Mo-Im Kim Nursing Research Institute, College of Nursing, Yonsei University, 50-1, Yonsei-Ro, Seodaemun-gu, Seoul, Republic of Korea

Corresponding Author:
Sue Kim, PhD
Mo-Im Kim Nursing Research Institute, College of Nursing, Yonsei University, 50-1, Yonsei-Ro, Seodaemun-gu, Seoul, Republic of Korea

Abstract

Background: Although exercise has been proven effective in alleviating cancer-related fatigue (CRF), traditional face-to-face programs may not be accessible due to physical, temporal, or geographical barriers. Digital health interventions (DHIs) offer scalable alternatives for promoting physical activity; however, evidence synthesizing DHI-based physical activity interventions specifically targeting CRF and their intervention characteristics remains limited.

Objective: This scoping review aimed to map the types of digital health–based physical activity interventions for managing CRF, to summarize the key characteristics of DHI modalities and CRF outcomes, and to identify knowledge gaps for future research.

Methods: A systematic literature search was conducted across 4 databases (PubMed, EMBASE, Cochrane, and PsycINFO) up to December 2025. Inclusion criteria comprised experimental studies involving adults with a history of cancer, digital exercise interventions, a control group, and fatigue outcomes. Screening and data extraction followed the Joanna Briggs Institute Manual for Evidence Synthesis and PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Review) standards. The protocol was registered in PROSPERO (International Prospective Register of Systematic Reviews; CRD42022304285). Interventions were classified using the Evidence Standards Framework (ESF).

Results: Thirty-three studies comprising 3443 participants were included, representing 32 interventions. Most were randomized controlled trials (n=30, 91%). Interventions were delivered primarily via mobile apps (11/32, 34%) and wearable devices (8/32, 25%), followed by web-based platforms, videoconferencing, exergaming, and augmented reality. Eighteen interventions (reported in 19 studies) demonstrated statistically significant between-group CRF reductions, predominantly at immediate postintervention assessments. Evidence for sustained benefit beyond 12 months was limited, and only one study evaluated ultra-long-term outcomes, which did not demonstrate maintained improvement. Populations with breast cancer accounted for the largest proportion of participants. Fatigue measurement tools varied substantially, potentially contributing to heterogeneity in effect estimates. Most interventions were classified as ESF tier C, indicating a predominant focus on clinical outcome improvement rather than individual-level self-management and system-level implementation. Wearable device-based interventions showed the highest proportion of significant CRF improvement, whereas mobile app–based interventions reported both significant and nonsignificant results. DHIs lasting more than 12 weeks appeared to be associated with more effective CRF outcomes compared to shorter programs. Direct comparisons between in-person and digital delivery were scarce.

Conclusions: Digital health–based physical activity interventions can reduce CRF in people with a history of cancer, with wearable device and longer-duration interventions showing the most favorable outcomes. However, the current evidence is concentrated in populations with breast cancer, and system-level integration remains unexplored. Future research should prioritize diverse populations with cancer, longer follow-up periods, standardized fatigue measurement, and systematic examination of specific intervention components that contribute to CRF reduction. Ultimately, advancing from individual efficacy trials toward scalable, workflow-integrated digital solutions will be key to sustainable CRF management across diverse oncology settings.

(J Med Internet Res 2026;28:e83727) doi:10.2196/83727

XSL•FO
RenderX

KEYWORDS

people with a history of cancer; digital health; exercise; fatigue; neoplasm

Introduction

Cancer-related fatigue (CRF) is defined as "a distressing, persistent, subjective sense of physical, emotional, and/or cognitive tiredness or exhaustion related to cancer or cancer treatments that interfere with usual functioning" [1]. CRF is one of the most common symptoms of cancer and its treatment and persists for several years after cancer treatment. The prevalence of overall CRF was 43% among people with a history of cancer, and the prevalence of mild CRF was 71% [2]. Since one-third of people with a history of cancer have experienced affective, cognitive, or physical fatigue, CRF is associated with comorbidities, behavioral and psychological factors, and should be continuously assessed and addressed throughout the entire course of survivor care [3]. CRF has been associated with pain, depression, education level, and physical activity [3,4]. Accordingly, clinical practice guidelines, including those from the National Comprehensive Cancer Network, American Society of Clinical Oncology–Society for Integrative Oncology, and European Society for Medical Oncology, consistently recommend that CRF be routinely screened and that nonpharmacological interventions, particularly exercise, be offered as management strategies [1,5,6].

Exercise is strongly recommended to alleviate CRF as a premier nonpharmacological intervention [5,7,8]. Multiple meta-analyses consistently suggested that exercise interventions significantly reduce CRF and improve the quality of life across various cancer types and treatment phases [7-9]. However, face-to-face exercise programs often encounter barriers related to accessibility, costs, and physical limitations of survivors [10,11]. These barriers have been further amplified for people with a history of cancer living in rural or underserved areas, and the COVID-19 pandemic has further highlighted the need for remote and flexible exercise delivery models [12]. This underscores the importance of exploring ways to deliver exercise interventions more feasibly and consistently throughout the survivorship period.

Digital health interventions (DHIs), defined as interventions delivered through digital technologies, have emerged as a scalable and accessible alternative, overcoming spatial and temporal constraints in cancer care [13,14]. In cancer care, DHIs have been applied across multiple domains, including symptom-focused interventions [15], promotion of physical activity through mobile apps [16], and delivery of psychological support via mobile health platforms [17]. These DHIs have been delivered through various modalities, such as mobile apps, web pages, and wearable devices. These approaches may be particularly useful for extending supportive care beyond conventional in-person settings and supporting survivorship care. Given the established role of exercise in managing CRF, DHIs designed to promote physical activity are relevant for people with a history of cancer. However, it remains understudied which types of digital physical activity interventions are associated with improvements in CRF.

Several existing reviews have examined related topics, but each has notable limitations. Previous systematic reviews and meta-analyses of e-health interventions have focused on physical activity as the primary outcome without examining its effects on CRF [18,19]. Other reviews have been limited to specific cancer types, such as breast cancer, restricting the generalizability of their findings [19-21]. A recent scoping review indicated that the effectiveness of DHIs on fatigue varies considerably depending on the intervention's type and functionality [22]. However, this review did not provide a granular synthesis of physical activity–specific digital interventions or clarify which components are most consistently associated with the reduction of CRF. To address these gaps, this scoping review specifically focuses on physical activity using DHIs, includes all cancer types, and maps the intervention components in relation to CRF. In addition, given the high degree of heterogeneity in digital platforms, exercise protocols, and fatigue assessment tools, a scoping review is the appropriate methodology for this emerging field [23]. It allows for a comprehensive mapping of the evidence and identifies specific intervention characteristics requiring further rigorous evaluation.

Despite the growing body of evidence supporting DHIs for promoting physical activity in people with a history of cancer, it remains understudied how the specific digital components of physical activity interventions relate to CRF across the diverse cancer population. Therefore, this scoping review aims to map the landscape of DHI-delivered physical activity interventions for managing CRF in people with a history of cancer, to summarize the key characteristics of the DHI modalities, CRF assessment methods, and reported CRF outcomes, and to identify knowledge gaps for future research.

Methods

Protocol and Registration

The protocol for this review was registered on PROSPERO (International Prospective Register of Systematic Reviews; registered number CRD42022304285). Although the protocol was initially planned as a systematic review, the study was conducted as a scoping review due to the high degree of heterogeneity in digital platforms, exercise protocols, and fatigue assessment tools in the included studies, which made a scoping review more appropriate for comprehensively mapping the evidence of physical activity DHI types for CRF.

Study Design

This scoping review selected studies on people with cancer engaging in physical activity DHIs targeting CRF and analyzed the findings. The study adhered to the guidelines of the JBI (Joanna Briggs Institute) Manual for Evidence Synthesis [24] and was reported in accordance with PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [25] (Checklist 1).

XSL•FO
RenderX

Eligibility Criteria

The key questions for the study were selected according to the JBI Manual for Evidence Synthesis [24] for population, concept, and context (Textbox 1).

Textbox 1. Population, concept, context

- Patient: People with a history of cancer (adults aged 18 years and older since diagnosis with cancer, excluding pediatric patients)
- Concept: exercise/physical activity through a digital health intervention (eg, mobile app, web-based, wearable device, etc) and cancer-related fatigue
- Context: open (no limitation)

The specific inclusion criteria were (1) experimental design studies with a control group, (2) using digital delivery methods, (3) employing a physical activity intervention for patients with a history of cancer, (4) reporting fatigue as an outcome, (5) aged 18 years and older at the time of diagnosis of cancer, and (6) studies published in English.

We excluded studies that did not meet one or more of these criteria, such as studies without a physical activity component, lacking digital delivery, not involving people with a history of cancer, without a fatigue outcome, or that were review or protocol papers. Multimodal interventions that included physical activity were included if fatigue outcomes linked to exercise were presented, whereas they were excluded if it was not possible to determine which intervention component influenced the outcome. Studies that used wearable devices solely for self-monitoring purposes and/or used phone follow-ups without interaction focused on the intervention were also excluded.

Information Sources

A systematic review of the major databases, such as PubMed, EMBASE, PsycINFO, and Cochrane, was primarily conducted until June 2024 and updated on January 6, 2026, for the period up to December 2025. References from relevant review papers were screened to identify additional eligible studies. The study registries, online resources, contacting authors, and additional information sources were not searched in line with the research purpose.

Search

The search strategy was reported following the PRISMA-Search (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [26] (Checklist 2). The search formulae comprised MeSH (Medical Subject Headings) terms and text words, with filtering methods employed to increase specificity based on the characteristics of each database. The search period was not limited for the review, which was based on key questions, and inclusion and exclusion criteria were used to select eligible studies. The search and data collection process was initially consulted with a professional librarian and reviewed and discussed in research team meetings. The research team comprised 2 PhDs, 1 graduate student, and 2 doctorally prepared professors of nursing. The existing search strategy was not reused due to inconsistent digital health keywords. After team discussions, a study-specific strategy was developed. The search terms combined the keywords “cancer,” “digital health,” and “cancer-related fatigue,” and the filters were for English and/or adults in each database (see Multimedia Appendix 1 for a full list of search terms). Two research team members (authors YK and YHK) independently conducted at least 1 search in each database individually to check for literature matches, and the retrieved papers were collected and organized using the EndNote 21 program.

Selection of Sources of Evidence

The screening of the collected data was conducted using Covidence (Veritas Health Innovation) [27] through more than 10 research team meetings. Duplicate records were removed using EndNote 21 (Clarivate) and Covidence. The titles and abstracts of the initially retrieved papers were reviewed independently, and then the full text of papers that appeared relevant was assessed. Inclusion and exclusion criteria were used to select eligible studies, but in cases of disagreement, the full text was reviewed with the entire team until consensus was reached through comprehensive discussion.

Data Charting Process

Data charting was conducted using Covidence, with 2 reviewers independently extracting data based on a predefined charting form developed from the study objectives.

The data were organized and categorized into tables according to the analysis framework that was created based on the objectives of the study. Studies were identified, categorized, and analyzed according to general characteristics (author, publication year, country, sample size, cancer type and stage, fatigue scale used, and intervention effectiveness) and intervention characteristics (intervention method and intervention delivery format). The Evidence Standards Framework (ESF) [28] was used to classify the DHI into 3 tiers: tier A targets system services to save health care providers’ time and cost; tier B facilitates health management through communication and information, encompassing functional categories such as health care diaries and promoting good health; tier C involves clinical interventions through informing clinical management, driving clinical management, active monitoring, calculating, diagnosing, and treating conditions. In this study, we categorized as grade A if study outcomes included cost analysis of providers’ time or expenses. If the intervention focused on offering general health information without interaction with experts, it was determined as grade B, and if it focused on personalized information or guidance for self-management or utilized interaction with experts to promote good health and healthy lifestyles, it was categorized as grade C.

XSL•FO
RenderX

Synthesis of Results

The results were synthesized descriptively to map the range and characteristics of digital health–based physical activity interventions targeting CRF. Studies were grouped and summarized according to study characteristics and intervention features, and frequencies were calculated for delivery modalities, intervention duration, ESF tiers, fatigue measurement tools, and intervention effectiveness.

Results

Selection of Sources of Evidence

A total of 5874 studies were identified based on the search strategy, and 3 studies were identified from citation searching. From this total, 1043 studies were removed due to duplications. Titles and abstracts of the remaining 4831 studies were screened based on the inclusion criteria. Consequently, 725 studies were excluded in the following order: no physical intervention (or unable to determine the effect of physical activity components; n=171), no digital health component in the intervention (n=459), no fatigue results (n=32), no control group (n=17), wrong population (n=2), no statistical analysis (n=1), review paper (n=2), and protocol paper (n=10). Ultimately, 33 studies [29-61] were included in the review (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram of study selection.

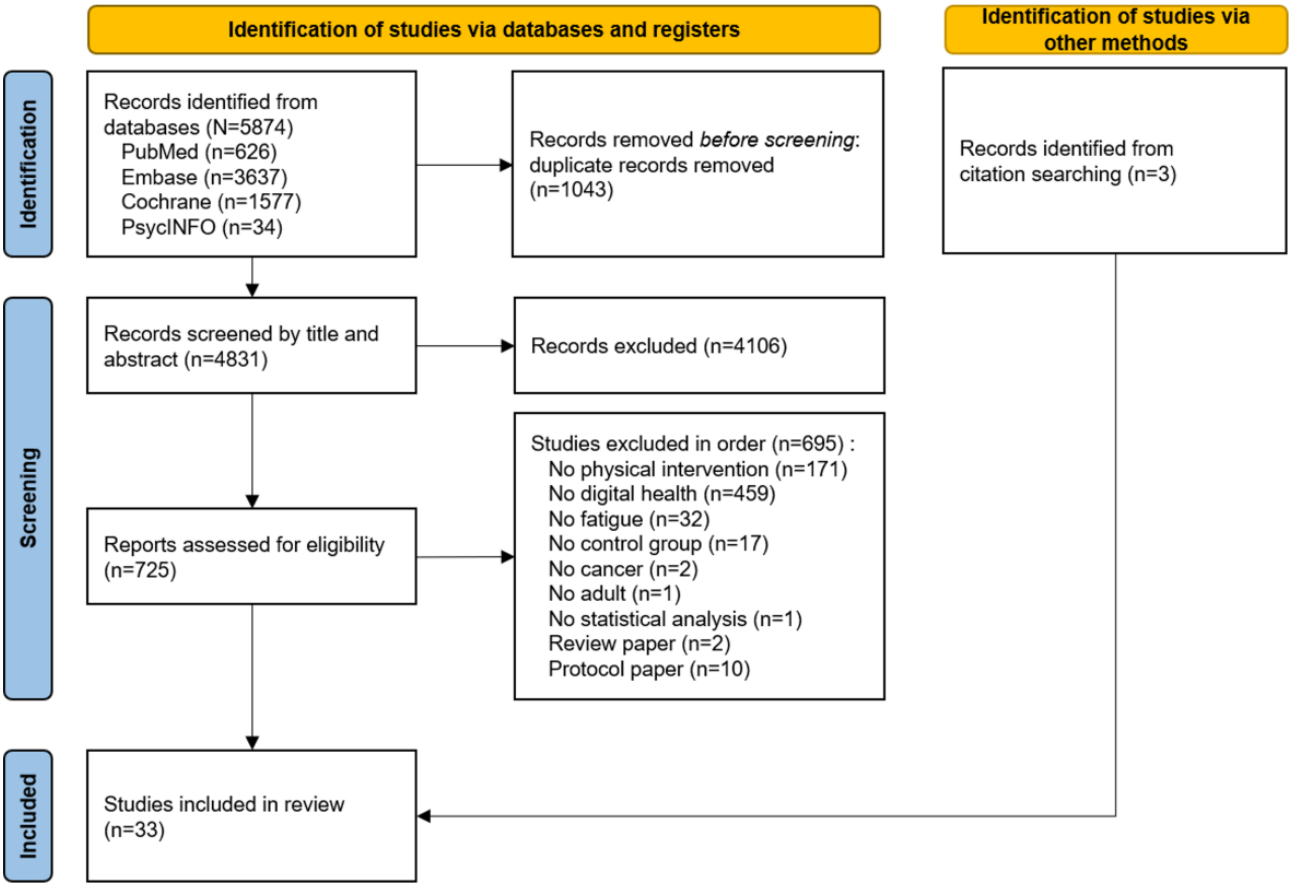

Characteristics of Sources of Evidence

The general characteristics of the 33 selected studies are shown in Table 1 and Multimedia Appendix 2. The publication dates ranged from 2015 to 2025, with 26 (79%) studies published after 2020. The United States accounted for the largest number of papers (n=9, 27%), followed, in order, by China (n=5, 15%), Spain (n=4, 12%), Australia (n=3, 9%), the Netherlands (n=3, 9%), and other countries (n=1, 3% each). In terms of study design, the majority (n=30, 91%) were randomized controlled trials (RCTs), and only 3 (9%) were quasi-experimental designs. Most were 2-arm studies (n=29, 88%).

XSL•FO
RenderX

Table . Publication date, country, study design, and sample characteristics of the 33 selected studies.

Parameter and category	Values, n (%)	References
Publication date		
2019	7 (21)	[29-35]
2020 - 2025	26 (79)	[36-61]
Country		
USA	9 (27)	[37,39,41,45,51,54,56,57,59]
China	5 (15)	[34,48,50,55,61]
Spain	4 (12)	[30,42,46,53]
Australia	3 (9)	[36,49,58]
Netherlands	3 (9)	[32,38,44]
Brazil	2 (6)	[33,43]
Belgium	1 (3)	[43]
Canada	1 (3)	[29]
Denmark	1 (3)	[35]
Japan	1 (3)	[40]
Poland	1 (3)	[47]
South Korea	1 (3)	[31]
Taiwan	1 (3)	[60]
Study design		
RCT[a]	30 (91)	[29,30,32,34-36,38-61]
Non-RCT	3 (9)	[31,33,37]
Cancer type		
Single cancer		
Breast cancer	10 (30)	[30,31,36,40,42,46,50,51,53,60]
Hematologic cancer	3 (9)	[45,54,58]
Lung cancer	2 (6)	[52,55]
Others	6 (18)	[34,35,37,47,48,61]
Mixed cancers	5 (15)	[29,32,38,44,49]
All cancer types	7 (21)	[33,39,41,43,56,57,59]
Completion of cancer treatment		
Yes (survivor)	16 (48.5)	[29-31,36,38-40,42,45,46,49,52,53,57-59]
No (patient)	14 (42.4)	[33-35,37,41,43,47,48,50,51,54-56,61]
Mixed sample	3 (9.1)	[32,44,60]
Control group		
Usual care	18 (55)	[29-31,34,35,38,41,42,47,49,50,53,55-60]
Active care	7 (21)	[37,39,40,46,51,52,61]
Waitlist	5 (15)	[32,36,44,45,48]
Others	3 (9)	[33,43,54]

[a]RCT: randomized controlled trial.

Upon examining key methodological characteristics of the included RCTs, most studies reported randomization and clearly defined control groups, whereas blinding was infrequently reported due to the behavioral nature of the interventions. Sample size and attrition were generally reported, and validated fatigue measures and appropriate statistical analyses were consistently used.

XSL•FO
RenderX

All studies reported aiming to determine the feasibility and suitability of the developed interventions or to evaluate their effectiveness. Two studies [42,44] were conducted to ascertain the long-term effects of the previous studies [30,32].

The total number of participants was 3443, with a range of 20 to 478 participants per study. When classified by cancer types, 21 out of 33 (63%) studies focused on a single type of cancer, of which the most common was breast cancer (10 studies). Five out of 33 (15%) studies included mixed samples of breast, prostate, and/or colorectal cancer, while 7 out of 33 (21%) were nonspecific and recruited all cancers.

Among the 33 studies included, 14 (42%) focused on patients receiving active cancer treatment, whereas 16 (48%) targeted people with a history of cancer who had completed intensive therapy, with less than half noting time since post treatment, which ranged from 1 month [53] to 2 years [45]. Three (9%) studies included participants in both treatment and posttreatment phases.

As presented in Table 1, for the control group, 18 (55%) studies used usual care and 5 (15%) employed a waitlist. Seven (21%) studies employed active care, such as providing control participants with wearable devices without further specification (n=3), allowing access to the mobile app (n=2), and offering a coaching or education program (n=2). The remaining 3 studies were either a noncancer control [33], employed a crossover approach [43], or were factorial designs [54].

Results of Individual Sources of Evidence

Thirty-two interventions were analyzed, excluding studies [42,44] that reported on the same intervention [30,32]. As presented in Table 2 and Multimedia Appendix 3, the most frequent digital therapeutics and delivery method was a mobile app (n=11, 34%), of which most (n=9, 82%) were conducted after 2020, followed by wearable devices (n=8, 25%), such as wristband-type activity monitor devices. Other intervention modalities included a web page (n=5, 16%), videoconferencing (n=3, 9%), Xbox (n=3, 9%), augmented reality (n=1, 3%), and software that could not be classified as strictly mobile app–based or web-based (n=1, 3%). As shown in Figure 2, the number of studies using digital physical activity interventions increased since 2020. Wearable device and videoconferencing interventions emerged after 2020, whereas mobile apps and web page interventions were relatively distributed across the study period.

Table . Intervention characteristics description of the 32 interventions.

Parameter and category	Values, n (%)	References
Intervention modality		
Mobile app	11 (34)	[31,34,40,45,46,48,55-57,61]
Wearable device	8 (25)	[36,37,39,49-52,54]
Web pages	5 (16)	[29,30,32,38,59]
Videoconferencing	3 (9)	[53,58,60]
Xbox	3 (9)	[33,35,43]
Augmented reality device	1 (3)	[47]
Software	1 (3)	[41]
ESF[a] grade		
B	2 (6)	[29,35]
C	30 (94)	[30-34,36-41,43,45-61]
Intervention duration		
<12 weeks	9 (28)	[29,30,33,43,45,46,52,58,61]
12 weeks	15 (47)	[31,35-37,39,40,48-51,56,57,59,60]
>12 weeks	8 (25)	[32,34,38,41,47,53-55]

[a]ESF: Evidence Standards Framework.

The intervention duration ranged from 3 weeks to 6 months, with 12 weeks (n=15, 47%) being the most frequently used period. The detailed intervention characteristics, including physical activity type and digital components, are presented in Multimedia Appendix 3. Most interventions combined digital activity tracking with some form of professional support, such as online coaching, while the specific combination of intervention components varied across studies.

Interventions were graded according to the ESF [28]. Two of the 32 (6%) interventions were grade B, that is, focused on supporting patients' self-management and wellness; the remaining were grade C, that is, treatment or diagnosis-focused interventions. There was no grade A, which aims to save cost and time for staff, which is likely the nature of the selected studies being focused on testing the efficacy of physical activity for people following cancer diagnosis.

XSL•FO
RenderX

Figure 2. Number of studies by year and intervention modality (n=32 interventions).

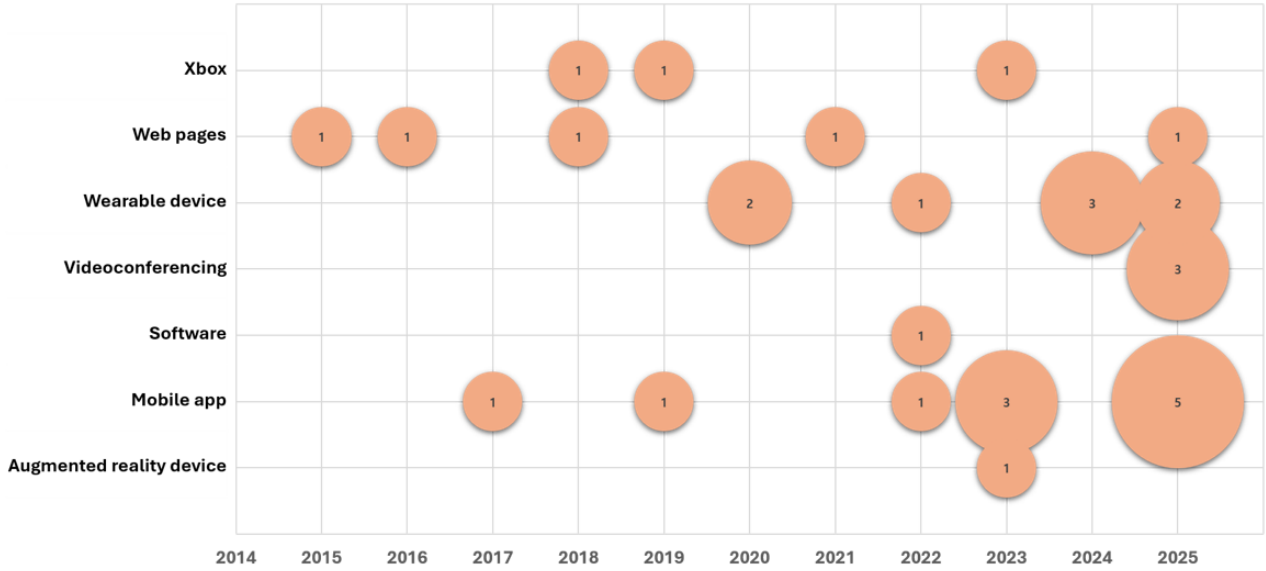

Various fatigue measurements were used in assessments (Table 3). The most frequently utilized scales were the Multidimensional Fatigue Inventory (MFI) and the European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire C-30 (EORTC QLQ C-30) fatigue subscale, each used in 5 (15%) studies. Three (9%) studies each used one of the following: Functional Assessment of Chronic Illness Therapy-Fatigue Scale, Functional Assessment of Chronic Illness Therapy-Fatigue, Piper Fatigue Scale, Patient Reported Outcomes Measurement Information System, or the Functional Assessment of Cancer Therapy-Fatigue or its fatigue subscale. Two (6%) studies each used either the Brief Fatigue Inventory, the Checklist Individual Strength, or the Chalder Fatigue Scale. Three studies used other instruments such as the Fatigue Symptom Inventory, the Multidimensional Fatigue Symptom Inventory—Short Form, and the Schwartz Cancer Fatigue Scale-6. All studies employed fatigue as the primary variable, with the exception of 4 studies [31,49,53,60], where the primary variable was quality of life, and fatigue was reported as a subscale of EORTC QLQ C-30.

XSL•FO
RenderX

Table . Description of fatigue measurements and statistical effects on fatigue between groups.

Parameter and category	Values, n (%)	References
Fatigue scale[a]		
MFI[b]	5 (15)	[37,38,48,52,61]
EORTC QLQ C-30[c]	5 (15)	[31,49,53,55,60]
FACIT-Fatigue[d]	3 (9)	[36,45,47]
FACIT-F[e]	3 (9)	[33,43,58]
FACT-F[f]	3 (9)	[29,35,54]
PFS[g]	3 (9)	[30,42,46]
PROMIS[h]	3 (9)	[51,56,57]
BFI[i]	2 (6)	[34,59]
CIS[j]	2 (6)	[32,44]
CSF[k]	2 (6)	[40,50]
Fatigue Symptom Inventory	1 (3)	[39]
MSFI-SF[l]	1 (3)	[55]
SCFS-6[m]	1 (3)	[41]
Statistically significant		
Yes (effective)	19 (58)	[30,32-34,36,37,39-41,43,44,46,50-55,60]
No (nonsignificant)	14 (42)	[29,31,35,38,42,45,47-49,56-59,61]

[a]As some studies used multiple scales, percentages sum to >100%.
[b]MFI: Multidimensional Fatigue Inventory.
[c]EORTC QLQ C-30: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire C-30.
[d]FACIT-Fatigue: Functional Assessment of Chronic Illness Therapy-Fatigue Scale.
[e]FACIT-F: Functional Assessment of Chronic Illness Therapy-Fatigue.
[f]FACT-F: Functional Assessment of Cancer Therapy-Fatigue.
[g]PFS: Piper Fatigue Scale.
[h]PROMIS: Patient Reported Outcomes Measurement Information System.
[i]BFI: Brief Fatigue Inventory.
[j]CIS: Checklist Individual Strength.
[k]CFS: Chalder Fatigue Scale.
[l]MSFI-SF: Multidimensional Fatigue Symptom Inventory—Short Form.
[m]SCFS-6: Schwartz Cancer Fatigue Scale-6.

In all studies, fatigue was measured a total of 44 times after baseline. Twenty-nine measurements occurred immediately at the conclusion of the intervention [29-31,33-41,43,45-48]. Four studies included midpoint measurements during the intervention [32-34,47], while 4 studies did not assess fatigue during the immediate postintervention period [32,42,44,58]. Twelve measurements focused on longer-term effects within 12 months after the postintervention period [30,32,36,44,46,49,58,59]. Another study evaluated an ultra-long-term effect 5 years later [42] from the original study [30].

Efficacy on Fatigue

As presented in Table 3, a total of 19 studies (18 interventions) reported a statistically significant reduction in CRF in the intervention group compared with the control group. Two studies employing the same intervention presented a positive effect on CRF 3 months before the end of the intervention and at 6 months [32,44] and 12 months after the intervention [44]. Building on the original report that showed significant results at the 6-month mark [30], its follow-up study [42] examined outcomes 5 years after the intervention. However, no statistically significant improvement in fatigue was noted at this ultra-long-term assessment. Of the 18 interventions that reduced CRF, 7 (39%) were wearable devices, 4 (22%) were mobile apps, 2 (11%) each employed videoconferencing, web pages, or Xbox, and 1 (6%) used software as the DHI component.

As illustrated in the evidence gap map (Table 4), wearable device interventions showed the highest proportion of statistically significant CRF improvement (7/8, 88% interventions), whereas mobile app interventions showed the lowest proportion of that (4/10, 40%). No studies examined

XSL•FO
RenderX

web page interventions for hematologic or lung cancer, and videoconferencing interventions were tested only for individuals with breast cancer and hematologic cancer. As shown in Figure 3, interventions lasting more than 12 weeks presented the highest proportion of significant CRF improvements (6/8, 75%), followed by those lasting less than 12 weeks (5/9, 56%) and those lasting 12 weeks (7/15, 47%).

Table . Evidence gap map of intervention modality by cancer type and cancer related fatigue across 33 included studies[a].

Intervention modality	Breast cancer	Hematologic cancer	Lung cancer	Others	Mixed cancers	All cancer types
Mobile app	[31,40,46]	[45]	[55]	[34,48,61]	—[b]	[56,57]
Wearable	[36,50,51]	[54]	[52]	[37]	[49]	[39]
Web pages	[30/42[c]]	—	—	—	[29, 32/44[c], 38]	[59]
Videoconferencing	[53,60]	[58]	—	—	—	—
Xbox	—	—	—	[35]	—	[33,43]
AR[d] device	—	—	—	[47]	—	—
Software	—	—	—	—	—	[41]

[a]The following studies showed statistically significant improvement in CRF between groups: [30,32-34,36,37,39-41,43,44,46,50-55,60].

[b]Not available.

[c]References [30] and [32] report the original intervention, and [42] and [44], respectively report its long-term follow-up.

[d]AR: augmented reality.

Figure 3. Proportion of studies reporting statistically significant cancer-related fatigue (CRF) improvement by intervention duration (n=32 interventions).

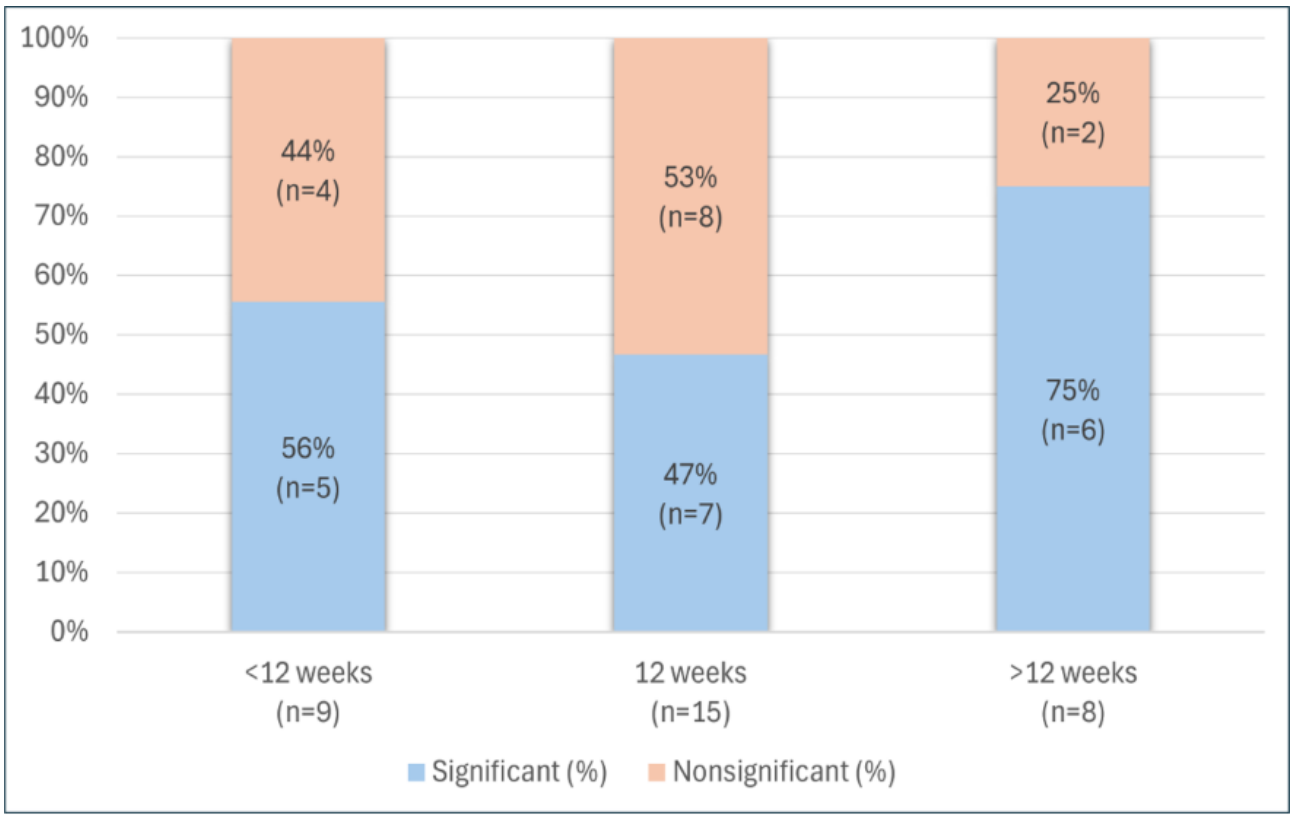

Discussion

Principal Results

This scoping review mapped the current experimental evidence on physical activity programs delivered via DHIs to address CRF in people with a history of cancer and examined which modality of intervention features tend to improve in CRF. Mobile apps and wearable devices were the most commonly used delivery modalities, followed by web pages, videoconferencing, and exergaming. Twelve-week programs were the most frequently adopted intervention duration, and the evidence has expanded rapidly, with the majority of studies published after 2020, reflecting growing interest in digitally delivered exercise programs for people with a history of cancer. A wide range of fatigue assessment tools was used across the

XSL•FO
RenderX

studies, with no single measurement predominating. More than half of the intervention reported statistically significant CRF improvements; however, findings at long-term follow-up were less consistent. Wearable device interventions showed the most consistent decrease in CRF, whereas mobile app interventions demonstrated more mixed findings. Interventions lasting more than 12 weeks were more effective in CRF than those of equal or less than 12 weeks.

Several knowledge gaps were identified. People with a history of breast cancer comprised the largest proportion of the study populations, and evidence for other cancer types, including hematologic or lung cancer, remained limited. Only one study assessed outcomes beyond 5 years, restricting the conclusion regarding the ultra-long-term durability of effects [42]. Furthermore, the high degree of heterogeneity in fatigue assessment tools, control group conditions, and digital delivery modalities made it difficult to directly compare intervention effects across the studies. These gaps highlighted the need for future studies examining diverse populations with cancer, longer follow-up periods, and standardized fatigue assessments to better determine the optimal digital delivery modality and intervention components for CRF reduction.

Comparison With Prior Work

Previous systematic reviews and meta-analyses have primarily synthesized nondigital exercise interventions or relaxation-based approaches [7,8,62,63]. A meta-analysis [64] reported that eHealth interventions improved pain interference and sleep but not pain severity or fatigue among people with a history of cancer; however, exercise-specific digital interventions and sustainability were not systematically examined. Moreover, recent reviews [18,19,22] have typically aggregated digital interventions across heterogeneous behavioral targets (eg, symptom management, general physical activity, or psychosocial support), without distinguishing exercise-based programs or comparing platform-specific intervention characteristics.

The present review extends prior work by characterizing digital delivery modalities, mapping intervention scope using the ESF [28], and identifying heterogeneity across cancer types, treatment phases, and outcome measurements, thereby informing not only efficacy but also the implementation readiness of digital exercise interventions for CRF. ESF mapping shows that most interventions focus on individual-level support and pay relatively little attention to system-level efficiency and workflow integration. Mobile and wearable-based interventions primarily supported continuous self-monitoring and real-time feedback, which have been identified as core behavior change techniques facilitating awareness and engagement in physical activity promotion [13,14]. In contrast, web-based programs tended to rely more on structured educational content and scheduled interactions, which are typical characteristics of traditional web-based health behavior interventions [65], potentially leading to different engagement patterns and intervention intensity. Most interventions were classified as ESF grade C, with no tier A interventions addressing system-level efficiency [28], indicating opportunities for scalable, implementation-oriented development, and integration into routine oncology service workflows. These modality-specific patterns suggest that the way support is delivered may influence participant engagement and the effective dose of the intervention.

These findings align with a recent meta-analysis, which reported that wearable electronic device system–supported physical activity programs significantly improved the quality of life and physical activity outcomes [66]. The continuous feedback mechanism inherent in wearable devices may enhance self-regulatory processes, thereby facilitating more consistent engagement with physical activity compared to other digital modalities. In contrast, although a previous review reported that mobile app–based interventions enhanced physical activity among people with a history of cancer [18], the present review found that mobile app interventions showed mixed results on CRF. This finding suggests that increased physical activity alone may not be sufficient to improve fatigue outcomes.

Fatigue trajectories after breast cancer treatment are influenced by biobehavioral factors, suggesting that intervention responsiveness may differ across diagnoses [67]. This concentration reflects where physical activity programs delivered via DHIs have been most feasible to test to date, underscoring the need to expand recruitment to more diverse populations with cancer. Given that treatment burden can influence baseline fatigue and responsiveness to exercise-based interventions [7,8,62], heterogeneity in cancer treatment exposure, including surgery, chemotherapy, radiotherapy, immunotherapy, and transplantation, may further constrain intervention outcomes—a consideration when analyzing physical activity DHIs. The mixed follow-up findings suggest that structured support initially improves outcomes, but maintenance may require continued reinforcement or transition-to-maintenance components after the program ends [54]. A recent review on digital health and telehealth in cancer care similarly reported that the evidence for fatigue outcomes was mixed [68]. This suggests that the inconsistent long-term results observed in the present review reflect a broader pattern in the field. Further research is needed to determine how long the CRF benefits of DHIs persist after the completion of intervention and what factors contribute to the maintenance or decrease of these effects over time. Direct comparisons between in-person and digital interventions remain limited, however, as only 2 of our selected studies explicitly compared delivery modalities [57,61]. One study comparing personalized exercise programs delivered in-person versus telehealth found that both interventions improved fatigue, without differences between interventions [57]. The other RCT reported improvement in physical activity with DHIs compared to in-person interventions, but no statistically significant fatigue differences between in-person and DHI modalities [61]. This suggests that whether interventions are delivered digitally or in-person may not determine fatigue outcomes, but the intensity, duration, and content of interventions may be the crucial components of reducing fatigue mechanisms. This highlights a critical evidence gap regarding the comparative effectiveness and implementation efficiency of digital versus in-person interventions for CRF. Given the growing evidence supporting the association between physical activity and CRF in several people with a history of cancer, particularly people with a history of breast and prostate cancer [69], future study should examine whether DHIs on

XSL•FO
RenderX

physical activity can produce consistent CRF benefits across more diverse populations with cancer and treatment settings.

Limitations

This review was limited to English-language publications. Methodological quality was examined using simplified descriptive indicators rather than formal risk-of-bias tools. Although multiple major biomedical databases were searched, the omission of engineering-focused databases (eg, IEEE Xplore) may have limited the identification of some technology-driven interventions.

The inclusion of diverse cancer diagnoses and heterogeneous treatment regimens, including differences in treatment intensity and recovery trajectories, may have influenced baseline fatigue severity and intervention responsiveness, limiting interpretability across populations. People with a history of breast cancer comprised the largest proportion of study populations, limiting generalizability to other cancer types and male populations. In addition, long-term follow-up evidence remains scarce, with only one study assessing outcomes beyond 5 years [42], restricting conclusions regarding the durability of effects. Finally, variability in digital access and literacy may further affect real-world applicability of these findings.

Heterogeneity across the included studies may have influenced the observed results in several ways. First, a wide range of fatigue assessment tools was used across studies, with no single instrument predominating, which limits the direct comparison of CRF outcomes across studies. Second, control conditions varied substantially, ranging from usual care and waitlist controls to active comparators using wearable devices or mobile apps, which may have attenuated detectable intervention effects and limited cross-study comparability. Third, the wide age range of participants, spanning young adults to older survivors, may further contribute to variability in engagement and intervention responsiveness due to differences in digital familiarity and physical capacity [70].

Conclusions

This scoping review identified several major types of physical activity DHIs for CRF in people with a history of cancer and found that wearable device interventions and longer-duration interventions represented an effective approach for CRF outcomes. Furthermore, as the demand for DHI modalities that actually prompt people with a history of cancer to move may increase in the future, our findings can be used by practitioners as they consider selecting and applying digital health tools to deliver activity-based interventions for alleviating CRF. As the reviewed DHIs were predominantly on patients with breast cancer, clinicians and researchers can respond to the shortcomings identified in this review by actively exploring strategies to engage and include diverse cancer types and underserved populations, while including CRF as a primary outcome.

Overall, this review supports the potential clinical relevance of digital health–based exercise interventions for managing CRF. Beyond individual-level efficacy, the absence of system-level integration (ESF grade A) across the reviewed interventions points to a critical gap between current research and real-world implementation. To enhance real-world impact, future research should prioritize standardized fatigue measurements, broader populations with cancer, and strategies to bridge the digital health literacy gap. Ultimately, shifting the focus from individual efficacy to scalable, workflow-integrated digital solutions will be essential for the sustainable management of CRF in diverse oncology settings.

Acknowledgments

We gratefully acknowledge Professor Sun Young Park (College of Nursing, Kyungpook National University, Korea) for creating the original search strategy and Dr. Na Won Kim and Ms. Dami Jeong (Yonsei University Medical Library) for reviewing the search process. We used the generative artificial intelligence tool ChatGPT by OpenAI under human supervision for 2 tasks classified according to the CAIDeT taxonomy: validation, which involved cross-checking the tables, and summarizing text, which involved assisting with the first draft of the abstract. Responsibility for the final manuscript lies entirely with the authors. Generative artificial intelligence tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was supported by a 2021 Alumni Research Cluster Grant (number 6-2021-0181) from Yonsei University College of Nursing. No funder had any role in the study design or decision to submit the paper for publication.

Data Availability

All data generated within this scoping review are available in tables and attached as multimedia appendices, such as search strategy and details of the selected studies.

Authors' Contributions

Conceptualization: YK, SK (equal)
Data curation: YK
Formal analysis: YK (lead), YHK (equal), KRB (supporting), JHS (supporting), SK (supporting)
Funding acquisition: SK
Investigation: YK, KRB, JHS, YHK, SK (equal)

XSL•FO
RenderX

Methodology: YK, KRB, JHS, YHK, SK (equal)
Project administration: YK (lead), YHK (supporting)
Supervision: SK
Visualization: YK (lead), YHK (supporting)
Writing – original draft: KRB (lead), YK (supporting)
Writing – review and editing: SK (lead), JHS (supporting), YHK (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full list of search terms.
[XLSX File, 19 KB - jmir_v28i1e83727_app1.xlsx]

Multimedia Appendix 2
General characteristics, study aims, and population characteristics of included studies.
[DOCX File, 37 KB - jmir_v28i1e83727_app2.docx]

Multimedia Appendix 3
Summary of intervention characteristics of physical activity and digital health components.
[DOCX File, 35 KB - jmir_v28i1e83727_app3.docx]

Checklist 1
PRISMA ScR checklist.
[DOCX File, 52 KB - jmir_v28i1e83727_app4.docx]

Checklist 2
PRISMA search checklist.
[DOCX File, 19 KB - jmir_v28i1e83727_app5.docx]

References

1. Berger AM, Mooney K, Alvarez-Perez A, et al. Cancer-related fatigue, version 2.2015. J Natl Compr Canc Netw 2015 Aug;13(8):1012-1039. [doi: 10.6004/jnccn.2015.0122] [Medline: 26285247]
2. Kang YE, Yoon JH, Park NH, Ahn YC, Lee EJ, Son CG. Prevalence of cancer-related fatigue based on severity: a systematic review and meta-analysis. Sci Rep 2023 Aug 7;13(1):12815. [doi: 10.1038/s41598-023-39046-0] [Medline: 37550326]
3. Thong MSY, Doege D, Koch-Gallenkamp L, et al. Fatigue in long-term cancer survivors: prevalence, associated factors, and mortality. A prospective population-based study. Br J Cancer 2025 Oct;133(6):831-843. [doi: 10.1038/s41416-025-03116-z] [Medline: 40665014]
4. Zeilinger EL, Zrnic-Novakovic I, Oppenauer C, et al. Prevalence and biopsychosocial indicators of fatigue in cancer patients. Cancer Med 2024 Jun;13(11):e7293. [doi: 10.1002/cam4.7293] [Medline: 38819432]
5. Bower JE, Lacchetti C, Alici Y, et al. Management of fatigue in adult survivors of cancer: ASCO-Society for Integrative Oncology guideline update. J Clin Oncol 2024 Jul 10;42(20):2456-2487. [doi: 10.1200/JCO.24.00541] [Medline: 38754041]
6. Fabi A, Bhargava R, Fatigoni S, et al. Cancer-related fatigue: ESMO Clinical Practice Guidelines for diagnosis and treatment. Ann Oncol 2020 Jun;31(6):713-723. [doi: 10.1016/j.annonc.2020.02.016] [Medline: 32173483]
7. Chen X, Li J, Chen C, et al. Effects of exercise interventions on cancer-related fatigue and quality of life among cancer patients: a meta-analysis. BMC Nurs 2023 Jun 13;22(1):200. [doi: 10.1186/s12912-023-01363-0] [Medline: 37312185]
8. Herranz-Gómez A, Cuenca-Martínez F, Suso-Martí L, et al. Effectiveness of therapeutic exercise models on cancer-related fatigue in patients with cancer undergoing chemotherapy: a systematic review and network meta-analysis. Arch Phys Med Rehabil 2023 Aug;104(8):1331-1342. [doi: 10.1016/j.apmr.2023.01.008] [Medline: 36736602]
9. Zhou S, Chen G, Xu X, et al. Comparative efficacy of various exercise types on cancer-related fatigue for cancer survivors: a systematic review and network meta-analysis of randomized controlled trials. Cancer Med 2025 Apr;14(7):e70816. [doi: 10.1002/cam4.70816] [Medline: 40145635]
10. Soriano-Maldonado A, Díez-Fernández DM, Esteban-Simón A, et al. Effects of a 12-week supervised resistance training program, combined with home-based physical activity, on physical fitness and quality of life in female breast cancer survivors: the EFICAN randomized controlled trial. J Cancer Surviv 2023 Oct;17(5):1371-1385. [doi: 10.1007/s11764-022-01192-1] [Medline: 35314958]

XSL•FO
RenderX

11. Lin Y, Wu C, He C, et al. Effectiveness of three exercise programs and intensive follow-up in improving quality of life, pain, and lymphedema among breast cancer survivors: a randomized, controlled 6-month trial. Support Care Cancer 2022 Dec 13;31(1):9. [doi: 10.1007/s00520-022-07494-5] [Medline: 36512157]
12. Lopez AM. Telehealth in cancer care: inequities, barriers, and opportunities. Cancer J 2024;30(1):2-7. [doi: 10.1097/PPO.0000000000000694] [Medline: 38265919]
13. Jahnel T, Pan CC, Pedros Barnils N, et al. Developing and evaluating digital public health interventions using the digital public health framework DigiPHrame: a framework development study. J Med Internet Res 2024 Sep 12;26:e54269. [doi: 10.2196/54269] [Medline: 39264696]
14. Kasoju N, Remya NS, Sasi R, et al. Digital health: trends, opportunities and challenges in medical devices, pharma and bio-technology. CSI Trans ICT 2023 Apr;11(1):11-30. [doi: 10.1007/s40012-023-00380-3]
15. Mayer C, Walch O, Dempsey W, et al. A circadian and app-based personalized lighting intervention for the reduction of cancer-related fatigue. Cell Rep Med 2025 Mar 18;6(3):102001. [doi: 10.1016/j.xcrm.2025.102001] [Medline: 40056908]
16. Cheung DST, Kwok TWH, Liu S, et al. A smartphone app (WExercise) to promote physical activity among cancer survivors: randomized controlled trial. J Med Internet Res 2025 Oct 3;27:e75839. [doi: 10.2196/75839] [Medline: 41042984]
17. Barth J, Schläpfer S, Schneider F, et al. Mobile health intervention CanRelax reduces distress in people with cancer in a randomized controlled trial. NPJ Digit Med 2025 May 10;8(1):269. [doi: 10.1038/s41746-025-01688-x] [Medline: 40348854]
18. Xiao K, Tang L, Chen Y, Zhou J, Yang Q, Wang R. The effectiveness of E-health interventions promoting physical activity in cancer survivors: a systematic review and meta-analysis of randomized controlled trials. J Cancer Res Clin Oncol 2024 Feb 2;150(2):72. [doi: 10.1007/s00432-023-05546-9] [Medline: 38305910]
19. Kang H, Moon M. Effects of digital physical activity interventions for breast cancer patients and survivors: a systematic review and meta-analysis. Healthc Inform Res 2023 Oct;29(4):352-366. [doi: 10.4258/hir.2023.29.4.352] [Medline: 37964457]
20. Cooper KB, Lapierre S, Carrera Seoane M, et al. Behavior change techniques in digital physical activity interventions for breast cancer survivors: a systematic review. Transl Behav Med 2023 Apr 15;13(4):268-280. [doi: 10.1093/tbm/ibac111] [Medline: 36694356]
21. Zhang X, Fang J, Hao Y, Yang D, Luo J, Li X. Digital behavior change interventions to promote physical activity and reduce sedentary behavior among survivors of breast cancer: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2025 Jun 19;27:e65278. [doi: 10.2196/65278] [Medline: 40537089]
22. Lee K, Kim S, Kim SH, et al. Digital health interventions for adult patients with cancer evaluated in randomized controlled trials: scoping review. J Med Internet Res 2023 Jan 6;25:e38333. [doi: 10.2196/38333] [Medline: 36607712]
23. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020 Oct;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
24. Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, editors. JBI Manual for Evidence Synthesis: Joanna Briggs Institute (JBI); 2024. [doi: 10.46658/JBIMES-24-01]
25. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
26. Rethlefsen ML, Page MJ. PRISMA 2020 and PRISMA-S: common questions on tracking records and the flow diagram. J Med Libr Assoc 2022 Apr 1;110(2):253-257. [doi: 10.5195/jmla.2022.1449] [Medline: 35440907]
27. Covidence. URL: https://www.covidence.org [accessed 2025-08-24]
28. Evidence standards framework for digital health technologies. : National Institute for Health and Care Excellence (NICE); 2018 URL: https://www.nice.org.uk/corporate/ecd7/resources/evidence-standards-framework-for-digital-health-technologies-pdf-1124017457605 [accessed 2025-04-01]
29. Forbes CC, Blanchard CM, Mummery WK, Courneya KS. Feasibility and preliminary efficacy of an online intervention to increase physical activity in Nova Scotian cancer survivors: a randomized controlled trial. JMIR Cancer 2015 Nov 23;1(2):e12. [doi: 10.2196/cancer.4586] [Medline: 28410166]
30. Galiano-Castillo N, Cantarero-Villanueva I, Fernández-Lao C, et al. Telehealth system: a randomized controlled trial evaluating the impact of an internet-based exercise intervention on quality of life, pain, muscle strength, and fatigue in breast cancer survivors. Cancer 2016 Oct 15;122(20):3166-3174. [doi: 10.1002/cncr.30172] [Medline: 27332968]
31. Uhm KE, Yoo JS, Chung SH, et al. Effects of exercise intervention in breast cancer patients: is mobile health (mHealth) with pedometer more effective than conventional program using brochure? Breast Cancer Res Treat 2017 Feb;161(3):443-452. [doi: 10.1007/s10549-016-4065-8] [Medline: 27933450]
32. Golsteijn RHJ, Bolman C, Volders E, Peels DA, de Vries H, Lechner L. Short-term efficacy of a computer-tailored physical activity intervention for prostate and colorectal cancer patients and survivors: a randomized controlled trial. Int J Behav Nutr Phys Act 2018 Oct 30;15(1):106. [doi: 10.1186/s12966-018-0734-9] [Medline: 30376857]
33. Oliveira PF, Iunes DH, Alves RS, Carvalho JM, Menezes FS, Carvalho LC. Effects of exergaming in cancer related fatigue in the quality of life and electromyography of the middle deltoid of people with cancer in treatment: a controlled trial. Asian Pac J Cancer Prev 2018 Sep 26;19(9):2591-2597. [doi: 10.22034/APJCP.2018.19.9.2591] [Medline: 30256065]

XSL•FO
RenderX

34. Lu Y, Qu HQ, Chen FY, et al. Effect of Baduanjin qigong exercise on cancer-related fatigue in patients with colorectal cancer undergoing chemotherapy: a randomized controlled trial. Oncol Res Treat 2019;42(9):431-439. [doi: 10.1159/000501127] [Medline: 31266043]
35. Villumsen BR, Jorgensen MG, Frystyk J, Hørdam B, Borre M. Home-based "exergaming" was safe and significantly improved 6-min walking distance in patients with prostate cancer: a single-blinded randomised controlled trial. BJU Int 2019 Oct;124(4):600-608. [doi: 10.1111/bju.14782] [Medline: 31012238]
36. Vallance JK, Nguyen NH, Moore MM, et al. Effects of the ACTIVity And TEchnology (ACTIVATE) intervention on health-related quality of life and fatigue outcomes in breast cancer survivors. Psychooncology 2020 Jan;29(1):204-211. [doi: 10.1002/pon.5298] [Medline: 31763746]
37. Xiao C, Beitler JJ, Higgins KA, et al. Pilot study of combined aerobic and resistance exercise on fatigue for patients with head and neck cancer: inflammatory and epigenetic changes. Brain Behav Immun 2020 Aug;88:184-192. [doi: 10.1016/j.bbi.2020.04.044] [Medline: 32330594]
38. van de Wiel HJ, Stuiver MM, May AM, et al. Effects of and lessons learned from an internet-based physical activity support program (with and without physiotherapist telephone counselling) on physical activity levels of breast and prostate cancer survivors: the PABLO randomized controlled trial. Cancers (Basel) 2021 Jul 21;13(15):3665. [doi: 10.3390/cancers13153665] [Medline: 34359567]
39. Johnson AM, Baker KS, Haviland MJ, et al. A pilot randomized controlled trial of a Fitbit- and Facebook-based physical activity intervention for young adult cancer survivors. J Adolesc Young Adult Oncol 2022 Aug;11(4):379-388. [doi: 10.1089/jayao.2021.0056] [Medline: 34677081]
40. Ochi E, Tsuji K, Narisawa T, et al. Cardiorespiratory fitness in breast cancer survivors: a randomised controlled trial of home-based smartphone supported high intensity interval training. BMJ Support Palliat Care 2022 Mar;12(1):33-37. [doi: 10.1136/bmjspcare-2021-003141] [Medline: 34389552]
41. Wilkie DJ, Schwartz AL, Liao WC, et al. Reduced cancer-related fatigue after tablet-based exercise education for patients. Cancer Control 2022;29:10732748221087054. [doi: 10.1177/10732748221087054] [Medline: 35414203]
42. Álvarez-Salvago F, Jiménez-García JD, Martínez-Amat A, et al. Does participation in therapeutic exercise programs after finishing oncology treatment still ensure an adequate health status for long-term breast cancer survivors? A ≥ 5 years follow-up study. Support Care Cancer 2023 May 18;31(6):343. [doi: 10.1007/s00520-023-07801-8] [Medline: 37199790]
43. da Silva Alves R, de Carvalho JM, Borges JBC, Nogueira DA, Iunes DH, Carvalho LC. Effect of exergaming on quality of life, fatigue, and strength and endurance muscle in cancer patients: a randomized crossover trial. Games Health J 2023 Oct;12(5):358-365. [doi: 10.1089/g4h.2022.0161] [Medline: 37155685]
44. Golsteijn RHJ, Bolman C, Peels DA, Volders E, de Vries H, Lechner L. Long-term efficacy of a computer-tailored physical activity intervention for prostate and colorectal cancer patients and survivors: a randomized controlled trial. J Sport Health Sci 2023 Nov;12(6):690-704. [doi: 10.1016/j.jshs.2023.08.002] [Medline: 37591482]
45. Lee K, Shamunee J, Lindenfeld L, et al. Feasibility of implementing a supervised telehealth exercise intervention in frail survivors of hematopoietic cell transplantation: a pilot randomized trial. BMC Cancer 2023 May 1;23(1):390. [doi: 10.1186/s12885-023-10884-5] [Medline: 37127595]
46. Lozano-Lozano M, Galiano-Castillo N, Gonzalez-Santos A, et al. Effect of mHealth plus occupational therapy on cognitive function, mood and physical function in people after cancer: secondary analysis of a randomized controlled trial. Ann Phys Rehabil Med 2023 Mar;66(2):101681. [doi: 10.1016/j.rehab.2022.101681] [Medline: 35671976]
47. Pieczyńska A, Zasadzka E, Pilarska A, Procyk D, Adamska K, Hojan K. Rehabilitation exercises supported by monitor-augmented reality for patients with high-grade glioma undergoing radiotherapy: results of a randomized clinical trial. J Clin Med 2023 Oct 29;12(21):6838. [doi: 10.3390/jcm12216838] [Medline: 37959303]
48. Wen L, Chen X, Cui Y, Zhang M, Bai X. Effects of Baduanjin exercise in nasopharyngeal carcinoma patients after chemoradiotherapy: a randomized controlled trial. Support Care Cancer 2022 Dec 23;31(1):79. [doi: 10.1007/s00520-022-07548-8] [Medline: 36562869]
49. Hardcastle SJ, Leyton-Román M, Maxwell-Smith C, Hince D. Impact of the promoting physical activity in regional and remote cancer survivors intervention on health-related quality of life in breast and colorectal cancer survivors. Front Oncol 2024;14:1368119. [doi: 10.3389/fonc.2024.1368119] [Medline: 39309736]
50. Li H, Sang D, Gong L, et al. Improving physical and mental health in women with breast cancer undergoing anthracycline-based chemotherapy through wearable device-based aerobic exercise: a randomized controlled trial. Front Public Health 2024;12:1451101. [doi: 10.3389/fpubh.2024.1451101] [Medline: 39363984]
51. Phillips SM, Starikovsky J, Solk P, et al. Feasibility and preliminary effects of the Fit2ThriveMB pilot physical activity promotion intervention on physical activity and patient reported outcomes in individuals with metastatic breast cancer. Breast Cancer Res Treat 2024 Nov;208(2):391-403. [doi: 10.1007/s10549-024-07432-5] [Medline: 39014267]
52. Arents E, Haesevoets S, Hermans F, et al. Physical activity telecoaching in post-surgical NSCLC patients: a mixed-methods pilot study exploring feasibility, acceptability and actual usage. Cancers (Basel) 2025 Sep 2;17(17):2886. [doi: 10.3390/cancers17172886] [Medline: 40940983]

XSL•FO
RenderX

53. Lavín-Pérez AM, Collado-Mateo D, Nieto I, et al. Effects of individualized high-intensity online concurrent exercise guided by autonomic modulation on the mental health and quality of life of breast cancer survivors. Psychooncology 2025 Dec;34(12):e70348. [doi: 10.1002/pon.70348] [Medline: 41339114]
54. Lee CY, Gordon MJ, Markofski MM, et al. Optimization of mHealth behavioral interventions for patients with chronic lymphocytic leukemia: the HEALTH4CLL study. J Cancer Surviv 2025 Aug;19(4):1325-1334. [doi: 10.1007/s11764-024-01555-w] [Medline: 38472612]
55. Li G, Zhou X, Deng J, et al. Digital therapeutics-based cardio-oncology rehabilitation for lung cancer survivors: randomized controlled trial. JMIR mHealth uHealth 2025 Feb 25;13:e60115. [doi: 10.2196/60115] [Medline: 39999435]
56. Lukkahatai N, Benjasirisan C, Shen A, et al. Combined technology-enhanced home exercise and acupressure (TEHEplus) program on symptoms among cancer patients receiving immunotherapy: a feasibility study. BMC Cancer 2025 Oct 1;25(1):1481. [doi: 10.1186/s12885-025-14887-2] [Medline: 41034793]
57. Lukkahatai N, Han G, Benjasirisan C, et al. A comparison of in-person and telehealth personalized exercise programs for cancer survivors: a secondary data analysis. Cancers (Basel) 2025 Jul 23;17(15):2432. [doi: 10.3390/cancers17152432] [Medline: 40805134]
58. Ma DD, Liu Z, Au K, et al. Randomized controlled trial of a virtually delivered exercise and stress management program to improve physical performance of hematopoietic cell transplant survivors. J Clin Oncol 2025 Mar 10;43(8):949-959. [doi: 10.1200/JCO.24.00333] [Medline: 39591546]
59. Unick JL, Duffy C, Dizon D, et al. Evaluation of a translatable web-based intervention for increasing physical activity among cancer survivors: pilot randomized trial. JMIR Cancer 2025 Oct 2;11:e79610. [doi: 10.2196/79610] [Medline: 41037739]
60. Yang YH, Chao YL, Lin YF, Liu PC, Chang KJ, Hou IC. Home-based remote dance program with biopsychosocial model improves quality of life in breast cancer patients: a randomized controlled trial. J Bodyw Mov Ther 2025 Sep;43:376-385. [doi: 10.1016/j.jbmt.2025.04.029] [Medline: 40483151]
61. Yu K, Yin B, Zhu Y, et al. Efficacy of a digital postoperative rehabilitation intervention in patients with primary liver cancer: randomized controlled trial. JMIR mHealth uHealth 2025 Apr 7;13:e59228. [doi: 10.2196/59228] [Medline: 40194311]
62. Zhou R, Chen Z, Zhang S, et al. Effects of exercise on cancer-related fatigue in breast cancer patients: a systematic review and meta-analysis of randomized controlled trials. Life (Basel) 2024 Aug 14;14(8):1011. [doi: 10.3390/life14081011] [Medline: 39202753]
63. Wang Y, Yang L, Lin G, et al. The efficacy of progressive muscle relaxation training on cancer-related fatigue and quality of life in patients with cancer: a systematic review and meta-analysis of randomized controlled studies. Int J Nurs Stud 2024 Apr;152:104694. [doi: 10.1016/j.ijnurstu.2024.104694] [Medline: 38281450]
64. Li J, Zhu C, Liu C, Su Y, Peng X, Hu X. Effectiveness of eHealth interventions for cancer-related pain, fatigue, and sleep disorders in cancer survivors: a systematic review and meta-analysis of randomized controlled trials. J Nurs Scholarsh 2022 Mar;54(2):184-190. [doi: 10.1111/jnu.12729] [Medline: 34791779]
65. de Sousa D, Fogel A, Azevedo J, Padrão P. The effectiveness of web-based interventions to promote health behaviour change in adolescents: a systematic review. Nutrients 2022 Mar 16;14(6):1258. [doi: 10.3390/nu14061258] [Medline: 35334915]
66. Wang Z, Li Y, Wang Q, Su Y. The effectiveness of wearable electronic device system-supported physical activity programs for cancer survivors: meta-analysis of randomized controlled trials. J Med Internet Res 2025 Aug 14;27:e74347. [doi: 10.2196/74347] [Medline: 40811695]
67. Bower JE, Wiley J, Petersen L, Irwin MR, Cole SW, Ganz PA. Fatigue after breast cancer treatment: biobehavioral predictors of fatigue trajectories. Health Psychol 2018 Nov;37(11):1025-1034. [doi: 10.1037/hea0000652] [Medline: 30321021]
68. Shaffer KM, Turner KL, Siwik C, et al. Digital health and telehealth in cancer care: a scoping review of reviews. Lancet Digit Health 2023 May;5(5):e316-e327. [doi: 10.1016/S2589-7500(23)00049-3] [Medline: 37100545]
69. Filis P, Markozannes G, Chan DS, et al. Grading the evidence for physical activity and any outcome in cancer survivors: an umbrella review of 740 meta-analytic associations. Crit Rev Oncol Hematol 2025 Mar;207:104602. [doi: 10.1016/j.critrevonc.2024.104602] [Medline: 39730034]
70. Choi E, Ryu S, Chun H, Kwak W, Choi SK. Personal competence factors associated with better access to digital health [Report in Korean]. : Korea Institute for Health and Social Affairs; 2022 URL: https://www.kihasa.re.kr/en/publish/paper/research/view?seq=46808 [accessed 2025-06-08]

Abbreviations

CRF: cancer-related fatigue
DHI: digital health intervention
EORTC QLQ C-30: European for Research and Treatment of Cancer Quality of Life Questionnaire C-30
ESF: Evidence Standards Framework
FACIT: Functional Assessment of Chronic Illness Therapy

XSL•FO
RenderX

JBI: Joanna Briggs Institute
MESH: Medical Subject Headings
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Review
PROSPERO: International Prospective Register of Systematic Reviews
RCTs: randomized controlled trials

Edited by S Brini; submitted 07.Sep.2025; peer-reviewed by G Grigoriadis, Y Hao; revised version received 14.May.2026; accepted 15.May.2026; published 26.Jun.2026.

Please cite as:
Kim Y, Bae KR, Sung JH, Ko YH, Kim S
Physical Activity Interventions Using Digital Health Interventions for Cancer-Related Fatigue in People With a History of Cancer: Scoping Review
J Med Internet Res 2026;28:e83727
URL: https://www.jmir.org/2026/1/e83727
doi:10.2196/83727

© Yeeun Kim, Ka Ryeong Bae, Ji Hyun Sung, Yun Hee Ko, Sue Kim. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Vision-Based Artificial Intelligence Technologies for Epilepsy Monitoring: Scoping Review and Taxonomy Development Study

Mirijana Irnich, MSc; Jonas Hammer, MSc; Aleksandra Flok, MSc; Frank Teuteberg, Prof Dr

Department of Management Accounting and Information Systems, University Osnabrück, Katharinenstraße 3, Osnabrück, Germany

Corresponding Author:
Mirijana Irnich, MSc
Department of Management Accounting and Information Systems, University Osnabrück, Katharinenstraße 3, Osnabrück, Germany

Abstract

Background: Artificial intelligence (AI) technologies for vision-based epilepsy monitoring are advancing rapidly in health care. Despite growing research using various video data sources and analytical approaches, no comprehensive framework exists to classify these technologies.

Objective: This scoping review aimed to develop and validate a taxonomy for AI technologies in vision-based epilepsy monitoring and to characterize visual AI approaches in epilepsy care.

Methods: Using an extended taxonomy development framework, we developed the taxonomy in 5 iterative cycles, drawing on theory and practice. We conducted a scoping review, market analysis, and applicability evaluation with market-ready solutions. We searched Scopus, Web of Science, and PubMed, including MeSH (Medical Subject Headings) terms; the final search was completed on January 16, 2026. We included primary studies from 2013 onward on AI-based or machine learning–based monitoring or prediction of epileptic seizures in humans using visual data. We excluded reviews, non-English publications, nonepilepsy studies, studies focused only on electroencephalography or wearables, animal studies, and pre-2013 publications. Evidence was charted through narrative and tabular synthesis and descriptive frequency analysis. In line with scoping review guidance, we did not conduct a meta-analysis or critical appraisal. To assess validity and practical relevance, 9 domain experts evaluated the taxonomy using a Delphi technique.

Results: We included 40 original studies. Study analysis yielded 16 dimensions, including data acquisition source, tracking target, image processing, classifier type, performance metrics, environment, seizure classification, data privacy, and user interface. Expert feedback added 4 further dimensions, including communication mode and information purpose. The final taxonomy comprises 23 dimensions with 102 characteristics. The review identified structural evidence gaps across settings, evaluation maturity, and reporting practices. Detection and classification in stationary settings predominated, whereas predictive approaches and real-time feedback were limited. Deep learning detection methods were common, but performance reporting was inconsistent, and patient-facing functionalities were limited. Privacy safeguards and standardized metrics were often incompletely reported, reducing comparability and maturity assessment. The taxonomy translates these patterns into guidance for benchmarking, procurement evaluation, user interface, and explainable AI design. We synthesized 5 main findings and 10 implications for research and practice. Key challenges concern standardization, seizure prediction, and real-time applicability.

Conclusions: Vision-based AI technologies for epilepsy monitoring are still dominated by proof-of-concept and pilot evaluations, indicating a gap between technical feasibility and deployment-ready systems. This scoping review presents an implementation-oriented taxonomy integrating application context, system architecture, visual analysis, AI models, performance reporting, and feedback design into a single classification framework. Unlike prior work that mainly maps methods or data sources, the taxonomy provides a shared structure for consistent system-level characterization and comparison across studies and emerging solutions. It may support benchmarking, implementation-focused evaluation, procurement, and translation into clinical and home settings.

(J Med Internet Res 2026;28:e83895) doi:10.2196/83895

KEYWORDS

epilepsy; taxonomy; artificial intelligence; health care; monitoring; computer vision; classification system; digital health; eHealth; remote patient monitoring systems; health information systems

XSL•FO
RenderX

Introduction

Background

Epilepsy, a prevalent neurological disorder affecting approximately 50 million individuals globally, is characterized by recurrent, unprovoked seizures that can significantly impair quality of life [1,2]. Effective monitoring and timely detection of seizures are crucial for optimal management and improved patient outcomes, especially for preventing sudden unexpected death in epilepsy [3]. Traditional monitoring methods, such as electroencephalography (EEG), often require wearable devices that can be intrusive and uncomfortable, potentially limiting patient compliance and continuous monitoring capabilities [4]. Recent advancements in artificial intelligence (AI) have paved the way for innovative, nonwearable monitoring systems that use audiovisual data to predict, detect, and warn of epileptic seizures [5]. These systems offer a contactless approach, enhancing patient comfort and enabling continuous monitoring without the physical constraints associated with wearable devices [6]. For instance, vision-based AI systems have demonstrated efficacy in detecting motor seizures by analyzing patient movements in visual data, most commonly video recordings, providing a nonintrusive alternative to traditional methods [7].

The current state of the art in epilepsy monitoring remains largely centered on EEG due to its high accuracy in detecting seizure events [8,9]. EEG-based systems remain the clinical gold standard, offering precise detection capabilities, but they are often resource-intensive because of long setup times, the need for clinical expertise, and the associated long hospital stays [3]. This traditional approach relies heavily on inpatient settings, wearable devices, or induced seizures through medication withdrawal—methods that can be intrusive, resource-intensive, and distressing for patients [10-12]. Given the high costs of hospital stays and long waiting times for diagnostic assessments, home telemonitoring has gained increasing importance [11]. Consequently, there is an emerging demand for innovative, patient-friendly monitoring systems that address these patient needs. For individuals with epilepsy, this involves reducing dependence on hospital-based diagnostic procedures through the implementation of reliable home-based detection systems. This approach is expected to ensure timely interventions, thereby enhancing the general well-being and quality of life of those affected.

A promising development in the field has been the combination of EEG with video-based seizure detection, which integrates the accuracy of EEG with the nonintrusive nature of vision-based monitoring [12]. Advances in deep learning (DL) have further enhanced vision-based seizure detection, improving system performance and providing additional decision support for epilepsy management [13]. These improvements have facilitated a transition from clinical technologies to home-based solutions relying solely on video monitoring, enabling greater accessibility and continuity of care [8]. The ability to capture seizure events in real-world settings further enhances ecological validity and contributes to more personalized treatment approaches [13]. The integration of AI with video data not only alleviates the discomfort associated with wearable devices but also enhances the accuracy of seizure detection. AI algorithms can analyze subtle visual cues and patterns that indicate seizure activity, facilitating early detection and intervention [4,5]. Moreover, nonwearable systems reduce the burden on patients and caregivers, promoting better adherence to monitoring protocols and improving overall quality of life [3].

AI monitoring systems represent a transformative shift in epilepsy detection and management, with vision-based solutions emerging as a key advancement in the field. These systems follow a similar workflow to process video data and analyze it using AI techniques. An illustrative overview of an exemplary schematic flow of monitoring systems based on visual data with AI models is provided in Multimedia Appendix 1. To facilitate a comprehensive classification of the contents of this work, the 5 steps of the process that serve as a contextual foundation are presented below as examples: monitoring, visual data acquisition, visual data preprocessing, AI model processing, and event analysis. At the start of the schematic flow, patient monitoring is initiated using either mobile or stationary systems to enable continuous observation [3,14]. Visual data are acquired through optical sensors that generate video recordings, which are subsequently preprocessed. The goal of preprocessing is to reduce artifacts, that is, unwanted disturbances or distortions that alter the signal, may compromise analytical quality, and are not part of the original input [15]. Common methods include image denoising, video stabilization, and normalization of video data to ensure high-quality input for subsequent AI-based analysis [14]. AI model processing then uses these optimized data, often through DL or convolutional neural networks (CNNs), to identify patterns and estimate seizure probabilities. Preprocessed data are analyzed with neural networks to detect potential clinical events, such as tonic-clonic seizures. Event analysis involves validating these potential events against predefined thresholds and classification criteria. If thresholds are exceeded (eg, event duration or intensity), the events are systematically classified as clinically relevant [16,17]. Automatically generated reports provide detailed information, including the temporal characteristics of the detected events [18]. In addition, a review process may be conducted in which medical professionals validate automated classifications and diagnoses and, when necessary, correct identified events to enhance the accuracy of future AI evaluations [5,18]. Depending on the objective, these monitoring systems either serve seizure detection or support diagnostic processes related to epileptic seizures [5].

The range of applications extends from contactless monitoring of respiratory and heart rates via ultra-wideband or millimeter-wave radar to continuous sleep analysis, reconstruction of mechanical and electrical cardiac activities, and analysis of nocturnal breathing patterns to detect early signs of neurodegenerative diseases, such as Parkinson disease [19]. Moreover, movement sequences are recognized and classified, which is particularly relevant for automatic fall detection in older or care-dependent individuals [20]. Another application includes monitoring correct medication intake, enabling the early identification of potential errors in administration [19]. As vision-based technologies advance, they are anticipated to

XSL•FO
RenderX

serve as a fundamental component of future epilepsy monitoring, providing a less invasive and more accessible alternative to traditional methods. While challenges remain, such as ensuring system reliability, addressing environmental variability (eg, fluctuating light conditions or disturbances), and improving user training, ongoing advancements in AI and DL are rapidly enhancing system performance [5,11,13,21,22]. With continuous technological progress, vision-based AI systems have the potential to transform epilepsy care by providing accessible, cost-effective, and user-friendly monitoring in real-world environments.

Previous reviews in this domain have provided valuable but largely isolated insights into specific aspects of AI-based epilepsy monitoring. Their categorizations primarily focused on the scope of monitoring applications [13], target groups such as pediatric or adult patients [3], the period of epilepsy (eg, ictal vs interictal) [3,10], data acquisition sources including EEG and video [23], tracking targets [13], video tracking approaches [21], image processing methods [24,25], and the types of classifiers and performance metrics used [23]. While these contributions have mapped important components of the field, they remain fragmented and mainly descriptive. Whereas earlier reviews have summarized the evidence, our work formalizes it by developing a taxonomy that specifies mutually exclusive and collectively exhaustive dimensions and characteristics. Their initiation occurs on the premise of substantiated proof-of-concept and concrete systems, with refinement occurring through iterative design methodologies. Our taxonomy integrates technical, clinical, and contextual dimensions into a single, coherent schema that enables consistent classification across research and market solutions of AI-enabled video monitoring for epilepsy. By synthesizing these disparate elements into a unified conceptual framework, the taxonomy provides a structured foundation for systematically comparing existing systems, identifying research gaps, and guiding the design and evaluation of future AI-driven monitoring solutions.

Objectives

Current research leverages diverse data sources and technological methods to advance the field of vision-based AI monitoring [8-10]. Despite prior studies conducting reviews to provide overviews of these approaches [11-14], a structured taxonomy for categorizing these developments remains absent, and little is known about the dimensions and characteristics of AI technologies in vision-based epilepsy monitoring. To address this research gap, this scoping review developed a taxonomy following the method proposed by Nickerson et al [26], which prescribes iterative cycles to identify dimensions and characteristics, and systematically examined AI technologies in vision-based epilepsy monitoring in both research and market technologies. Specifically, this paper clarified the purpose of developing a taxonomy by articulating why it is needed, how it is constructed, and what it contributes based on the extended taxonomy design process (ETDP) [16]. By reviewing the literature and the health care market and incorporating expert insights, it provides a structured overview of existing solutions and their state-of-the-art developments and offers a foundation for future advancements.

This scoping review aimed to develop and validate a comprehensive taxonomy for vision-based AI technologies in epilepsy monitoring. Specifically, we derived core dimensions and defining characteristics from a scoping review and a market analysis, instantiated and refined the taxonomy through iterative design cycles guided by an ETDP, and evaluated its completeness and practical relevance using expert elicitation. In addressing research question (RQ) 1, the taxonomy's dimensions and characteristics are specified. In addition, RQ2 synthesizes implications and future directions for research and practice. Our final taxonomy serves as a structured framework for classifying AI technologies in vision-based epilepsy monitoring, supporting researchers in organizing the field, developers in creating targeted solutions, and decision-makers in evaluating and selecting appropriate technologies. Furthermore, this taxonomy supports businesses and health care institutions in assessing market-ready systems and regulatory compliance while facilitating a systematic understanding of emerging trends for policymakers.

The following RQs will be addressed:

- RQ1: *What are the dimensions and characteristics of AI technologies in vision-based epilepsy monitoring that can be integrated to develop a comprehensive taxonomy?*
- RQ2: *What implications and future directions can be identified from the existing literature and expert perspectives on AI technologies in vision-based epilepsy monitoring?*

Methods

Overview of Taxonomy Design Process

A taxonomy is built through iterative derivation, instantiation, and evaluation to define the field's structure, yielding coherent dimensions and characteristics for comparison and decision-making. This work followed an integrated single evidence base with dual outputs design. A scoping review was conducted to provide the empirical basis for both (1) evidence charting and gap mapping and (2) taxonomy development. These complementary outputs pursue a shared objective of structuring and interpreting a heterogeneous, rapidly evolving body of literature at different analytical levels. This design avoids duplication while ensuring methodological coherence between evidence synthesis and framework construction. The scoping review synthesizes the breadth of published work and enables transparent reporting of study characteristics and evidence gaps. Building on this empirical synthesis, the taxonomy translates the identified patterns into a structured, reproducible classification framework that supports consistent comparison across research prototypes and market-ready solutions.

The aim of this scoping review was to systematically classify emerging AI technologies in vision-based epilepsy monitoring through the development of a comprehensive taxonomy. This was accomplished by adopting the methodological framework [26] and incorporating the ETDP [27]. The ETDP facilitates the explicit description of the underlying problem, the observed phenomenon, the target user groups, and the intended purpose. It is organized along the 6 Design Science Research

XSL•FO
RenderX

activities—problem identification, objective definition, design and development, demonstration, evaluation, and communication—and introduces iterative entry and exit points to structure cycles of conceptual and empirical work. For evaluation, ETDP adopts the *why-how-what* framing: clarify the purpose (formative or summative), select the paradigm and techniques (artificial or naturalistic), and specify the evaluated properties, for example, usefulness (Multimedia Appendix 2). *Why* clarifies function, whether the evaluation is formative or summative [28]. *How* specifies the environment, if it is an artificial (eg, laboratory studies) or naturalistic evaluation (eg, case studies with real users) and the timing (*ex-ante* or *ex-post*) and the methods [28,29]. *What* focuses on the criteria, including the objective and subjective ending conditions and the defined evaluation criteria [29]. Overall, the integration of ETDP enables rigorous ex-ante and ex-post assessment of taxonomies while maintaining flexibility for evolving phenomena.

The iterative taxonomy development process followed either a conceptual-to-empirical (C2E) or an empirical-to-conceptual (E2C) approach [26]. The deductive C2E method initiated the taxonomy development by systematically analyzing the relevant literature in the field of emerging AI technologies in vision-based epilepsy monitoring. An inductive E2C strategy was used to derive dimensions and characteristics from existing solutions in practice. Multimedia Appendix 3 illustrates the ETDP-aligned process and the 5 iterative cycles conducted. The meta-characteristic of this taxonomy is thus the identification of the defining properties of AI technologies in vision-based epilepsy monitoring. The ending conditions were guided by the objective and subjective criteria [26] (Multimedia Appendix 2). To derive the final taxonomy, a structured 5-iteration development process was conducted, alternating between C2E and E2C approaches. This iterative procedure combined theoretical grounding with practical assessment, with the latter achieved through an analysis of market-ready solutions to ensure relevance to current practice and to confirm practical alignment of this taxonomy (Figure 1).

This methodological approach is particularly suited, given the increasing complexity and heterogeneity of AI- and vision-based monitoring solutions in epilepsy care, which currently lack a unified framework to systematically organize and compare their diverse characteristics. Accordingly, the taxonomy development began with a rigorous articulation of the motivation and necessity for such a structured classification within this domain. The observed phenomenon comprises emerging AI technologies in vision-based epilepsy monitoring, especially those using video data such as video-EEG or stand-alone video-based seizure detection systems, which are playing an increasingly pivotal role in the diagnosis and management of epilepsy. The primary purpose of the taxonomy is to enable a systematic and transparent categorization of these emerging AI technologies, thereby fostering conceptual clarity, facilitating informed clinical and technological decision-making, and uncovering relevant avenues for future research. The taxonomy thus aims to address the needs of health care professionals, researchers, and technology developers by providing a comprehensive overview of existing and emerging solutions. Patients and patient advocacy groups represent an indirect but essential target group, as improved classification ultimately contributes to the development of more tailored and patient-centered technologies.

XSL•FO
RenderX

Figure 1. Methodological approach for taxonomy development modeled on Nickerson et al [26]. AI: artificial intelligence.

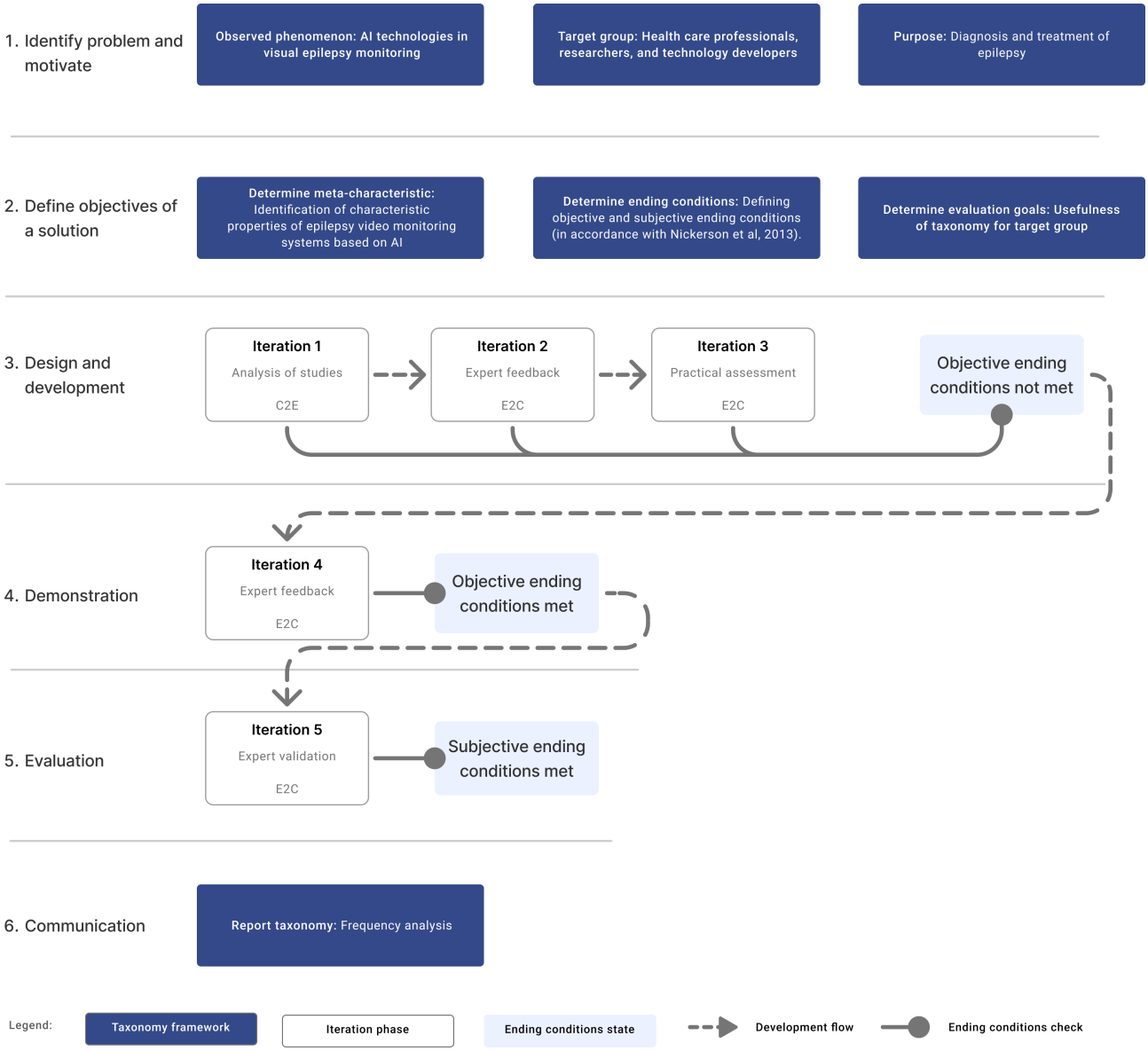

Design and Development

Study Design

The first iteration of the taxonomy development involved a scoping review to identify and synthesize existing taxonomies and empirical studies focusing on AI-enabled epilepsy monitoring according to vom Brocke et al [30]. This scoping review was conducted in accordance with the Joanna Briggs Institute guidelines [31] and the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) [32]. Search reporting follows PRISMA-S [33]. The completed PRISMA-S checklist is provided in Checklist 1, and full reproducible database-specific search strategies, including all update runs, are available in Multimedia Appendix 4.

Eligibility Criteria

We defined eligibility criteria following the Joanna Briggs Institute guidance for scoping reviews using the PCC (population, concept, and context) framework [31].

- Population: Studies involving human participants with epilepsy or epileptic seizures (or human seizure events captured on video) were included; animal studies were excluded.
- Concept: We included studies reporting AI-based or machine learning–based approaches for seizure monitoring, detection, classification, or prediction that used visual or video data (eg, computer vision–based monitoring). Studies focusing solely on EEG or wearable sensing without any visual or video component were excluded. We included only studies reporting the development and/or evaluation of a vision-based AI-ML system or algorithm for epilepsy monitoring. Studies describing rule-based approaches without an AI-ML component were excluded.
- Context: We considered any care context, including clinical environments (eg, epilepsy monitoring units and video-EEG settings), laboratory or sleep settings, and home or ambulatory monitoring. No restrictions were applied by country or geographic region.

XSL•FO
RenderX

Original empirical studies were included as primary evidence, whereas review articles were excluded. Review articles were excluded from the scoping review because the objective was to synthesize and map primary empirical evidence directly, thereby avoiding duplication of data and potential bias introduced by secondary interpretations. We included peer-reviewed, full-text research articles as eligible source types and excluded protocols, editorials, letters, commentaries, and abstract-only records. We restricted inclusion to English-language publications and to studies published from 2013 onward. This restriction aligns with the emergence of modern DL in computer vision and medical imaging. The breakthrough of deep CNNs in the 2012 ImageNet competition fundamentally changed image analysis and triggered the rapid expansion of DL approaches across domains, including medical imaging [34]. Given that our taxonomy aims to characterize current AI- and DL-based video monitoring approaches, restricting inclusion to studies published from 2013 onward is methodologically consistent with the technological evolution of the field.

Information Sources and Search Strategy

To identify relevant studies on AI-based monitoring systems for patients with epilepsy and visual data, the following search string was used to search in the title, abstract, and keywords: (“epilep*” OR “seizure*” OR “epileptic seizure*” OR “seizure disorder*” OR “tonic-clonic” OR “tonic clonic” OR “ictal” OR “convulsions” OR “generalized tonic-clonic” OR “grand mal”) AND (“AI” OR “artificial intelligence” OR “machine learning” OR “deep learning” OR “deep neural network*” OR “convolutional neural network*” OR “CNN”) AND (“predict*” OR “monitor*” OR “forecast*” OR “observ*” OR “recognit*” OR “alert*” OR “track”) AND (“video*” OR “camera*” OR “depth sensor*” OR “3D camera*” OR “vision-based” OR “vision based” OR “visual*” OR “motion*” OR “computer vision” OR “pose estimation”) AND PUBYEAR>2012. Search strategies were developed de novo and were not adapted from prior literature reviews. The information sources included Scopus (Elsevier), Web of Science Core Collection (Clarivate), and PubMed (National Library of Medicine). In PubMed, the search strategy was additionally complemented with relevant MeSH (Medical Subject Headings) terms, which are provided in Multimedia Appendix 4 together with the full database-specific search strategies. Each database was searched separately on its native platform (Scopus, Web of Science Core Collection, and PubMed); we did not conduct simultaneous multidatabase searching on a single platform. We did not search study registries. The searches were initially run on January 6, 2025. During revision, we updated the search by rerunning all database searches with a refined strategy on October 1 and on November 17, 2025. The final search was conducted on January 16, 2026, for all databases, covering publications from January 2013 onward. Each update reran the full search in all databases. No automated alerts were used. Newly retrieved records were deduplicated against the existing library prior to screening. No published search filters were used; we applied database limits for document type (article) and language (English) as reported in the database-specific strategies (Multimedia Appendix 4). The scoping review focused on peer-reviewed literature indexed in the selected databases; additional sources such as trial registries, preprint servers, or gray literature were not included. We did not hand-search websites, journal tables of contents, or conference proceedings. No protocol was registered. No additional data or studies were sought by contacting authors, experts, manufacturers, or others.

Study Selection and Data Charting

The search strategy was peer-reviewed by another information technology expert using the Peer Review of Electronic Search Strategies (PRESS) checklist [35] and revised accordingly. All records were exported to Microsoft Excel. Duplicates were removed in EndNote 2025 (Clarivate) using the multistep deduplication method described by Bramer et al [36] prior to screening. After deduplication, titles and abstracts were screened independently by 3 reviewers against the eligibility criteria. Potentially relevant records underwent full-text assessment. Discrepancies were resolved through discussion until consensus was reached. Comprehensive data extraction tables covering taxonomy dimensions are provided in Multimedia Appendix 5, while study-level characteristics are presented separately in Multimedia Appendix 6. Study-level characteristics were systematically charted using a structured Excel-based coding template (Multimedia Appendix 7).

Synthesis and Taxonomy Development Process

Before synthesis, extracted data were harmonized by standardizing terminology (eg, model types, evaluation metrics, and clinical settings). Consistent with scoping review guidance, critical appraisal of the individual sources of evidence was not performed. Given the heterogeneity of study designs and outcomes, no statistical pooling or meta-analysis was conducted; instead, findings were synthesized narratively and tabulated to highlight recurring dimensions, methodological trends, and research gaps.

The included studies were analyzed in the first iteration, following a C2E approach. Information from the original studies was used to derive meta-characteristics and foundational dimensions that shaped the conceptual basis of the taxonomy by incorporating insights from proof-of-concept and market-ready solutions. Three reviewers extracted data using the same structured Excel-based coding template (Multimedia Appendix 7), and discrepancies were resolved through discussion. The included studies were heterogeneous and frequently lacked components (eg, control groups and patient-level outcomes) necessary for structured bias assessment tools. Instead, we documented methodological limitations narratively and qualitatively considered transparency indicators to support credibility.

To integrate the perspective of potential end users, a *second iteration* was conducted using an E2C approach. The use of a Delphi study, as outlined by Gallego and Bueno [37], was particularly well suited to this work, enabling efficient consensus-building among a specialized group of domain experts within a limited time frame. This approach facilitated iterative refinement of the taxonomy while maintaining methodological rigor, ensuring that the resulting framework was both practically relevant and theoretically sound. By incorporating structured rounds of feedback, the process enhanced the validity of the

XSL•FO
RenderX

taxonomy and aligned it with established quality criteria for taxonomy development. This Delphi study [37] was initiated with domain experts, including clinicians, engineers, and AI developers (Multimedia Appendix 8). Specifically, the panel consisted of AI and clinical experts from practice (P1, P2, and P3) and researchers in AI, eHealth, data security, and taxonomy development (R1, R2, R3, R4, R5, and R6). Their qualitative feedback led to the refinement of the taxonomy and the addition of 4 new dimensions. As a result, the ending conditions for taxonomy development were not yet met, and the process continued.

The *third iteration* returned to an inductive E2C approach. We accumulated the dimensions and characteristics for the subsequent iterations by searching for market-ready solutions of AI technologies for vision-based epilepsy monitoring to align with practice. A market analysis was conducted using the Crunchbase database to examine existing commercial systems. Crunchbase constitutes a leading internet database provider for business information, such as products, venture capital, and industries [38]. The search terms "epilepsy" and "video monitoring" identified only one company, "Neuro Event," which developed the monitoring system "NELLI" [39]. In addition, the video-EEG monitoring system "SEER" was included in the analysis [24]. This step ensured that the taxonomy not only reflected academic research but also captured characteristics of real-world applications and emerging industry trends.

Demonstration and Evaluation

Combining iterations 4 and 5 within this section is warranted by their shared method (E2C, Delphi study), common participants, and uniform data collection instruments. Both iterations operationalize the application of the taxonomy and its *ex-post* assessment. The focus shifts from validation of objective to subjective ending conditions. An integrated report better represents the iterative logic of the taxonomy development, ensures comparability across rounds, and provides a coherent audit trail for how the ending conditions were examined.

In the *fourth iteration*, a second round of the Delphi study [37] was conducted with the same group of experts. As part of this process, the experts received the iteratively developed taxonomy via email, including detailed descriptions of each dimension and characteristic as derived through the previous iterations. They were explicitly instructed to evaluate whether the identified dimensions and characteristics were both necessary and sufficient for classifying AI technologies in vision-based epilepsy monitoring and to determine if any essential aspects were missing, ambiguous, or misclassified. Furthermore, the experts were asked to systematically assess the taxonomy against the taxonomy design principles [26] to examine whether the objective ending conditions were fulfilled.

The *fifth iteration* involved conducting a third round of the Delphi study [37] with the same group of experts. They reviewed the updated taxonomy and provided structured feedback, particularly with regard to the subjective ending conditions [27]. These conditions include *conciseness*, *comprehensiveness*, *robustness*, *explainability*, and *extensibility*. To facilitate structured feedback, experts were provided with a standardized evaluation template that contained explicit questions for each subjective ending condition to elicit qualitative comments and suggestions related to each of these conditions. For *conciseness*, experts judged whether any dimension or characteristic was redundant or could be merged without loss of meaning. The *comprehensiveness* of the taxonomy was investigated through its mapping to typical and rare use cases from vision-based epilepsy monitoring to identify potential gaps. *Robustness* was examined using deliberately atypical vignettes to evaluate whether classification guidance remained reliable under nonstandard conditions. *Explainability* was evaluated by having experts restate the taxonomy's distinctions and indicate whether the aspects, dimensions, characteristics, and naming conventions supported coherent interpretation across stakeholders. *Extensibility* was assessed by requesting proposals for plausible future additions and checking whether these could be accommodated without reworking existing dimensions. Experts then issued an overall judgment on whether the subjective ending conditions were met, with a short rationale. Aligned with the *evaluation goal* of establishing the taxonomy's *usefulness* for health care professionals, researchers, and technology developers, we implemented a summative, ex-post usefulness check within the Delphi study. The same group of experts mapped the taxonomy to typical and edge-case scenarios drawn from vision-based epilepsy monitoring practice and reflected on its applicability for conducting analyses, enabling systematic comparisons, structuring communication across roles, and informing design decisions. Responses were captured within a structured template that elicited brief rationales and judgments of sufficiency [27].

In summary, experts provided a brief overall statement on whether the objective and subjective ending conditions were met, alongside a justification of the usefulness of the taxonomy. The outcomes of these judgments are reported in the *Results* section.

Communication

Reporting the taxonomy through a frequency analysis shows which fields are well served and where the gaps are located. To communicate the state of the field relative to the taxonomy, we conducted a descriptive frequency analysis consistent with step 6 *Report Taxonomy* of our methodological approach [27]. The unit of analysis was the set of studies included in the scoping review and the results of the market analysis for AI technologies for vision-based epilepsy monitoring solutions. A standardized coding template derived from the final taxonomy guided the extraction of aspect, dimension, and characteristic labels. Using a predefined matrix, 3 researchers independently annotated each study for the presence of characteristics within every characteristic, allowing multilabel coding where the taxonomy permits (eg, multiple data acquisition sources) and otherwise enforcing exactly one characteristic per dimension. Ambiguities and missing information were marked as not reported following explicit decision rules, with disagreements resolved by consensus after calculating interrater agreement. We then aggregated binary indicators to produce per-characteristic frequencies at the level of studies and, where available, at the level of distinct systems. The resulting counts were summarized

XSL•FO
RenderX

in tabular form and visualized to communicate coverage across dimensions and characteristics.

Ethical Considerations

This work did not involve medical products or medicinal trials. It combines a Delphi study and a scoping review. According to the guidelines of the research committee of the University of Osnabrück (Germany), scoping reviews do not require ethical approval. Furthermore, the Research Ethics Committee of the University of Osnabrück (Germany) determined that the Delphi study does not constitute human subjects research (decision/date: NHSR/19.09.2025). This Delphi study focused only on professional opinions from adult participants and involved anonymous expert elicitation and no collection of identifiable private information. Therefore, formal ethics approval was not required. Nevertheless, all panelists received written information about the study purpose (including aims, procedures, risks, and data handling) and provided written informed agreement to participate. Participation was voluntary and could be discontinued at any time without penalty. No vulnerable populations were included. All Delphi panelists were adults (aged ≥18 y) and participated voluntarily after receiving information about the study's aims, procedures, risks, and data handling. No clinical information or identifiable patient data were collected. Expert feedback was documented using coded identifiers and analyzed in aggregate. No direct personal identifiers are reported. All files were kept on a secure institutional platform of the University Osnabrück with access restricted to the study team with password-protected login and handled in compliance with the General Data Protection Regulation. Participants received no financial or material compensation for taking part in this work.

Results

Study Characteristics

Selection of Sources of Evidence

After the removal of duplicates, this yielded a total of 1157 hits in Scopus, PubMed, and Web of Science databases. Three authors screened all titles and abstracts independently for eligibility, followed by full-text assessment of eligible records. Discrepancies were resolved through discussion. No automation tools were used during screening or data charting. Following a comprehensive scoping review, 30 publications were excluded due to the absence of video data, and 10 papers were excluded because they did not involve any AI technologies. In addition, 8 review articles (secondary research) were excluded. After screening titles, abstracts, and reading full texts, 40 original studies remained for dimension and characteristic analysis (Figure 2). Three authors collected data from the reports and assessed each study independently. No automation tool was used. On the basis of the scoping review, we initially identified the first categories of the taxonomy analysis: tracking target, period of epilepsy, type of classifier, evaluation metrics, scope, and image processing. The selection process is summarized in the PRISMA flow diagram (Figure 2).

XSL•FO
RenderX

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of the study selection process. AI: artificial intelligence.

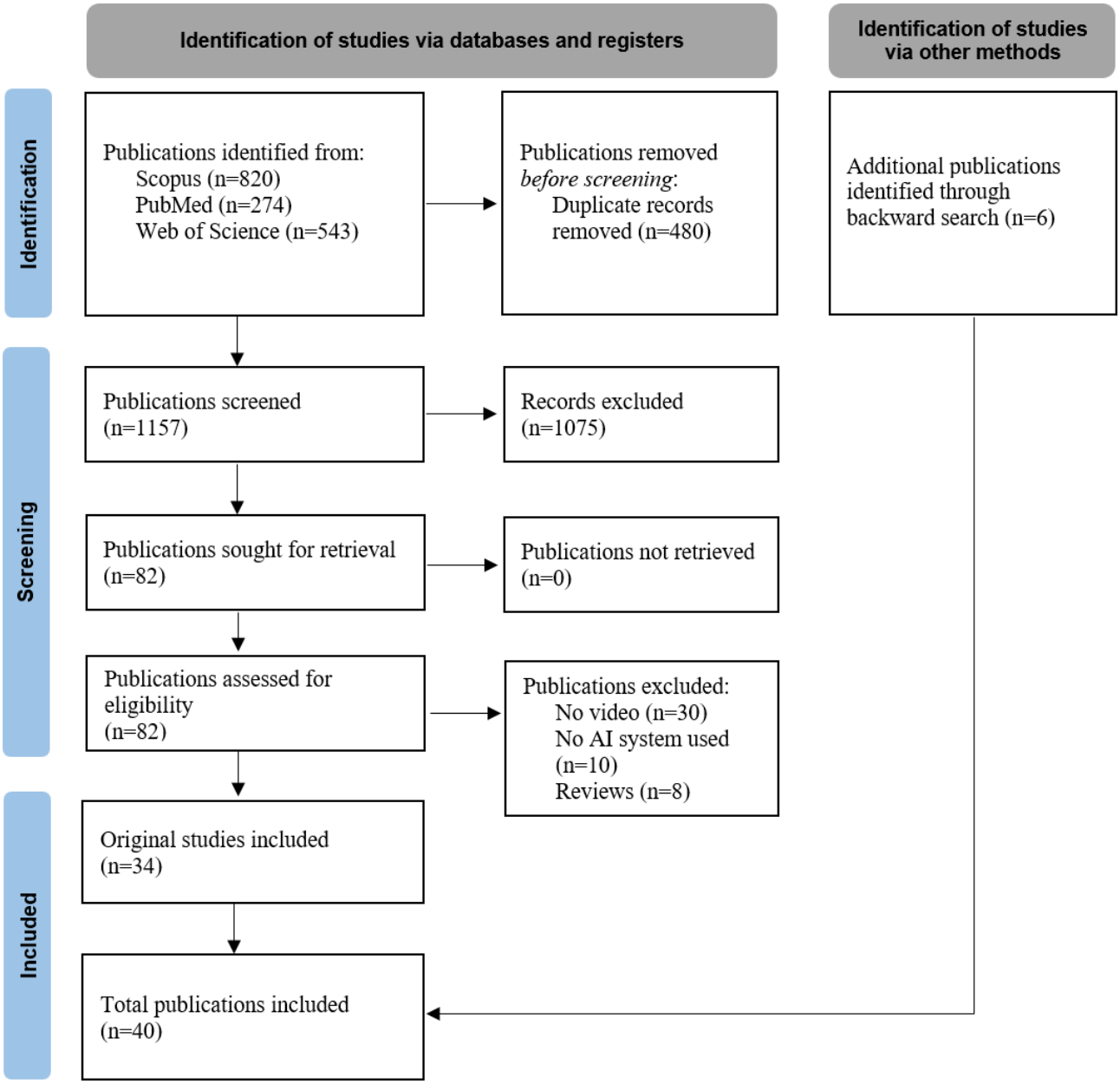

Characteristics of Sources of Evidence

Key study characteristics are transparently charted across the included evidence, enabling comparison of setting, study design, evaluation maturity, data source, sample type, age group, and sample size (Table 1). All extraction study details are provided in Multimedia Appendix 9. Across the original studies (N=40), the evidence base is predominantly situated in hospital wards (n=30, 75%), follows method development designs (n=28, 70%), and reports pilot evaluations in real-world settings (n=32, 80%). Retrospective data use (n=23, 57,5%) is more common than prospective designs, and participant samples are generally small, with most studies including 50 or fewer individuals. Overall, the evidence landscape reflects a field centered on technical feasibility and early-stage validation within controlled clinical environments. These structural patterns informed the subsequent taxonomy development by highlighting recurring dimensions (eg, setting, evaluation maturity, and data source type) and revealing systematic gaps that required explicit conceptual representation within the classification framework.

XSL•FO
RenderX

Table . Study characteristics.

Dimension and characteristic	Sources, n (%)
Setting	
Hospital ward	30 (75)
Home	6 (15)
Simulated	2 (5)
Residential care	2 (5)
Study design	
Method development	28 (70)
Feasibility	6 (15)
Observational study	4 (10)
Dataset development	2 (5)
Evaluation maturity	
Pilot in real setting	32 (80)
Laboratory prototype	6 (15)
Deployed	2 (5)
Data source type	
Retrospective	23 (57.5)
Prospective	11 (27.5)
Staged	3 (7.5)
Mixed	2 (5)
Public dataset	1 (2.5)
Sample type	
Patients	38 (95)
Healthy volunteers	2 (5)
Age group	
Adults	32 (80)
Pediatric	7 (17.5)
Mixed	1 (2.5)
Participant sample size (total N)	
≤10 (very small)	12 (30)
11 - 50 (small)	19 (47.5)
51 - 200 (medium)	4 (10)
>200 (large)	4 (10)
Not reported	1 (2.5)

Synthesis of Results

To support a structured overview of the evidence base and make gaps visible at a glance, we created an evidence map to visualize the distribution of evidence across setting and evaluation maturity (bubble plot; Multimedia Appendix 10). The map highlights 3 recurring gaps: first, prospective real-world evaluations in home settings remain scarce (Figure 3); second, the literature focuses predominantly on detection and classification, while prediction is comparatively underrepresented (Multimedia Appendix 10); and third, heterogeneous and incomplete reporting—particularly regarding classifier type and the use of standardized performance metrics—limits cross-study comparability and makes it difficult to judge methodological maturity and practical readiness (Multimedia Appendix 10).

Taken together, these gaps indicate that the field is not only uneven in where and how systems are evaluated but also in what is targeted and how results are reported. To move beyond gap identification and provide a structured synthesis that can guide future development, evaluation, and reporting, the next section derives a taxonomy from the included sources. This taxonomy

XSL•FO
RenderX

consolidates the observed design and evaluation characteristics into a coherent classification scheme and explicitly incorporates the underrepresented areas identified in the evidence map.

Figure 3. Evidence map of the study results by setting and evaluation maturity.

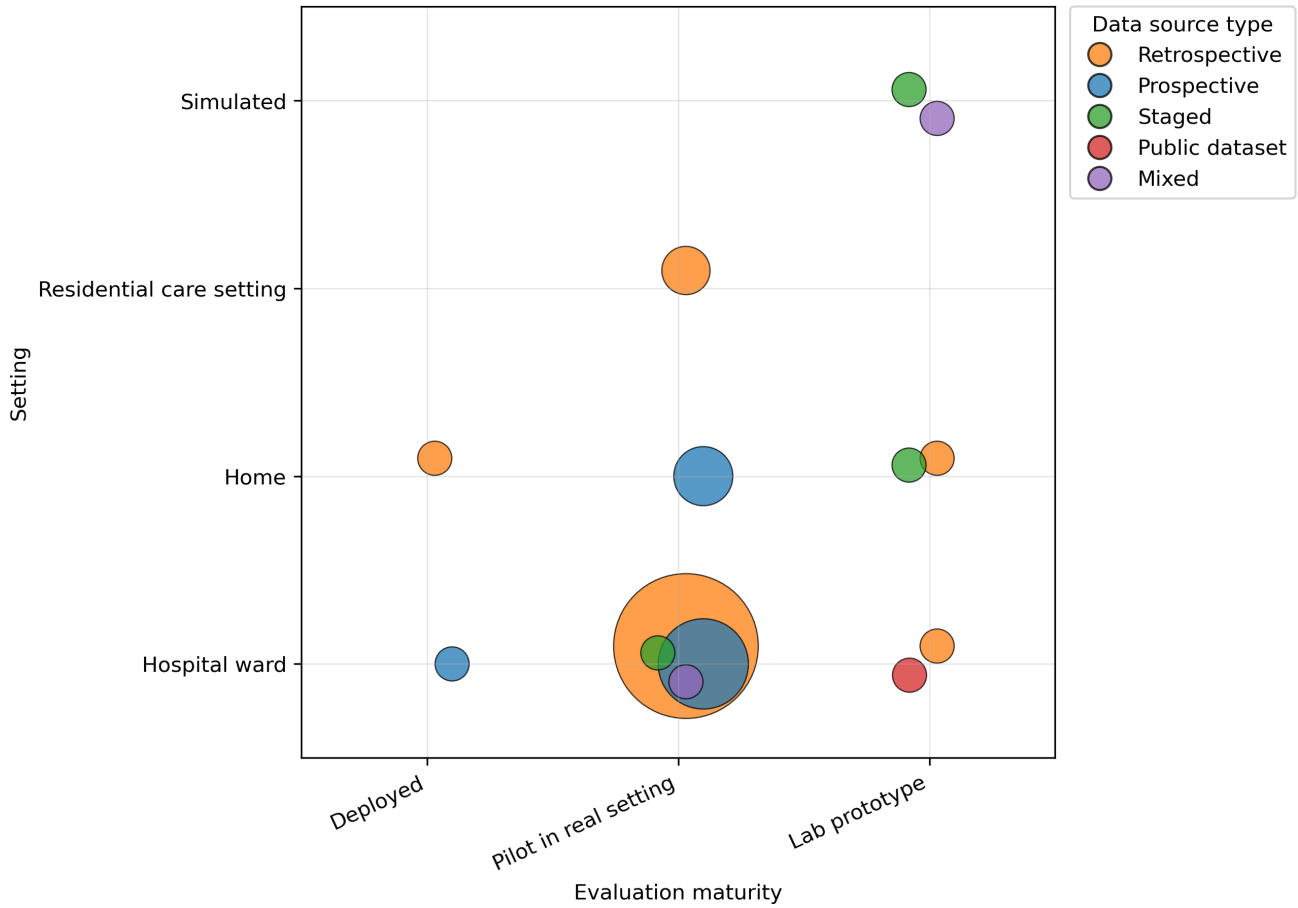

Overview of All Taxonomy Iterations

To ensure a transparent and comprehensible presentation of the taxonomy development process, this section provides an overview of the 5 iterative cycles undertaken to derive the final taxonomy. Each iteration contributed incrementally by identifying, refining, or validating dimensions and characteristics, thereby systematically constructing a robust framework for classifying AI technologies in vision-based epilepsy monitoring.

Developed through an iterative methodology, the final taxonomy consists of 23 dimensions and 102 characteristics, capturing the complexity and heterogeneity of current and emerging solutions in this field. In Multimedia Appendix 3, we provide a schematic overview of the iterations and the dimensions identified at each stage, illustrating the evolution of the taxonomy and the structured expansion of its scope.

During the first iteration, we initially identified 16 dimensions: scope [40], target group [22], period of epilepsy [41,42], data acquisition source [43,44], tracking target [45], video tracking [46,47], image processing [14], type of classifier [42,48], performance metrics [41,49], environment [50], seizure classification [51], medical device [17], salient attributes [5], data privacy [18], user interface [45], and user interaction [3]. These newly uncovered dimensions broadened the taxonomy to better reflect the diverse technical and contextual facets of AI-enabled visual monitoring solutions. As the predefined ending conditions [26] had not yet been fulfilled, iteration 2 used an E2C approach, leveraging a Delphi study to integrate expertise from researchers and practitioners. This strategy led to the addition of 4 further dimensions: cryptographic measure, communication mode, response type, and information purpose. The structured expert feedback not only ensured practical relevance but also helped maintain methodological rigor aligned with established quality criteria for taxonomy development. Despite these advancements, the ending conditions remained unmet, leading to iteration 3. In this stage, practical alignment was prioritized by conducting a market analysis of commercial solutions using Crunchbase and targeted internet research. As a result, 3 new dimensions were identified: computing paradigm, connection type, and support [24,39]. This practical focus ensured that the taxonomy adequately captured real-world implementation aspects and industry trends. The ending conditions were not yet met; another iteration followed. Finally, iterations 4 and 5 consisted of a second round of the Delphi study with the same panel of domain experts to rigorously evaluate the taxonomy against both the objective and subjective ending conditions in accordance with Nickerson et al [26]. The experts confirmed the taxonomy's conciseness, robustness, explainability, comprehensiveness, and extensibility, with no additional dimensions proposed.

XSL•FO
RenderX

Final Taxonomy

Overview

The defined taxonomy met all predefined ending conditions after 5 iterations of developing the aspects, dimensions, and characteristics. Finally, the subjective ending conditions, including conciseness, robustness, explainability, comprehensiveness, and extensibility, were fulfilled. It is imperative to note that the completion of 5 subjective ending conditions indicates a high sufficiency of the taxonomy. However, we present the final outcome of our research: Established through a rigorous iterative methodology, the taxonomy comprises 6 overarching aspects, encompassing 23 dimensions and a total of 102 characteristics. Each aspect integrates specific dimensions and associated characteristics, offering a rigorous foundation for evaluating the functionality and applicability of these systems (Table 2). The target audience of this taxonomy includes clinical researchers, neurologists, health care professionals, decision-makers, and developers who aim to evaluate, design, or implement AI technologies in vision-based epilepsy monitoring.

XSL•FO
RenderX

Table . Final taxonomy of emerging artificial intelligence (AI) technologies in vision-based epilepsy monitoring.

Aspect and dimension	Characteristics
Application and context	
Scope	• Detection • Classification • Prediction
Target group	• Patients with epilepsy • Caregivers • Medical professionals • Data scientists
Environment	• Stationary • Mobile
Seizure classification	• Nonmotor symptoms • Motor symptoms
Period of epilepsy	• Interictal • Preictal • Ictal • Postictal
Data acquisition source	• Depth sensors • Infrared • 2D camera • 3D camera • Video-EEG[a] • ECG[b] • Audio
Visual analysis	
Tracking target	• Body • Face • Sleeping area • Room overview
Video tracking	• MD[c] • BC[d] • MOI[e] • ROI[f] • HMD[g] • SKS[h] • Appearance- and feature-based methods
Image processing	• Optical flow • Frame differencing • Spatiotemporal interest points • Contrast-based analysis
AI model	
Type of classifier	• SVM[i] • RF[j] • MP[k] • CNN[l] • GMM[m] • LSTM[n] • I3D[o] • Other types of classifier

XSL•FO
RenderX

Aspect and dimension	Characteristics
Performance metrics	• Accuracy • Sensitivity • Specificity • F_1-score • Precision • Recall • False-positive rate • Area under the curve
Market identity	
Medical device	• Certified • Not certified • Proof of concept
Salient attribute	• Environmental robustness • Cost-efficiency • Real-time analysis • Ease of use • High system performance
Data privacy	• Anonymization • Pseudonymization • No privacy preserving measures • Synthetic data
Cryptographic measure	• Encryption in transit • Encryption at rest • No encryption
System architecture	
User interface	• Web platform • Voice assistant • Mobile app • Desktop app • Wearable device
User interaction	• Reporting • Interactive • Adaptive • No interaction
Computing paradigm	• Cloud-based platform • Edge-based platform • Local on device
Connection type	• Wi-Fi • Built-in-modem • Ethernet • Bluetooth
Support	• On call • System setup • Help center • Daily technical checks • Expert data review • Chat
Feedback system	

XSL•FO
RenderX

Aspect and dimension	Characteristics
Communication mode	• Real time • Periodic • Event based • On demand
Response type	• Visual • Auditory • Haptic • Text based
Information purpose	• Alerting or warning • Performance evaluation • Recommendation • User learning

[a]EEG: electroencephalogram.
[b]ECG: electrocardiogram.
[c]MD: movement dynamics.
[d]BC: biomechanical characteristics.
[e]MOI: movement of interest.
[f]ROI: region of interest.
[g]HMD: head movement detection.
[h]SKS: simple keypoint system.
[i]SVM: support vector machine.
[j]RF: random forest.
[k]MP: multilayer perceptron.
[l]CNN: convolutional neural network.
[m]GMM: Gaussian mixture model.
[n]LSTM: long short-term memory.
[o]I3D: inflated 3D.

Application and Context

The first aspect delineates the operational domain and intended purpose of AI technologies in vision-based epilepsy monitoring. It considers their primary scope ranging from seizure detection, classification, and prediction [8,13] to the specific target group they serve, including patients with epilepsy, caregivers, medical professionals, and data scientists [3,45]. This aspect further differentiates systems based on their environment, distinguishing between stationary implementations, such as hospital- and home-based systems with mobile configurations that allow for monitoring in diverse environments [3,45]. Moreover, it incorporates distinctions in seizure classification, differentiating between motor and nonmotor symptomatology [23]. The temporal dimension of epilepsy monitoring is also captured, classifying the period of epilepsy as interictal, preictal, ictal, or postictal [8]. Additionally, it accounts for various data acquisition modalities, including depth sensors, infrared imaging, 2D and 3D cameras, video-EEG, electrocardiography, and audio [23,52].

Visual Analysis

The second aspect examines the computational techniques used to track and interpret seizure-related phenomena. This aspect defines key tracking targets, encompassing body, face, sleeping area, and room overview [3,47,52,53]. It further classifies video tracking methodologies, including movement dynamics [17,50,51,54], biomechanical characteristics [54,55], movement of interest [5,45,56], region of interest [16,56-58], head movement detection (HMD) [55,59], simple keypoint system [47,59], and appearance- and feature-based methods leveraging visual attributes [14,54]. Furthermore, it includes image processing techniques such as optical flow [6,16,60-62], frame differencing [63], spatiotemporal interest points [57,62], and contrast-based analysis [6,57,60], all of which contribute to enhancing the accuracy and robustness of seizure identification [6,23].

AI Model

The third aspect addresses the algorithmic paradigms underpinning epilepsy monitoring systems. This aspect details the types of classifier, including support vector machine [48,64], random forest [48], multilayer perceptron [48,59], CNN [6,50,54,57,61,65], Gaussian mixture model [63], long short-term memory (LSTM) [50,53,57,61,65], inflated 3D [18,54,66], and other types of classifiers, for example, automated seizure and spike detection [10]. An essential topic of this aspect is the assessment of system efficacy using well-established performance metrics, including accuracy, sensitivity, specificity, F_1-score, precision, recall, false-positive rate, and area under the curve [8,12,23,67,68].

XSL•FO
RenderX

Market Identity

The fourth aspect captures regulatory, commercial, and privacy-related considerations. It classifies systems based on their certification status, differentiating between certified [45] and not certified systems and proof-of-concept systems [48,60,61,65]. It also identifies salient attributes, such as environmental robustness, cost-efficiency, real-time analysis [55,62], ease of use [16], and high system performance [62]. Additionally, it examines data privacy protocols, distinguishing among anonymization, pseudonymization, no privacy preserving measures (eg, no anonymization or pseudonymization) [14,22,68], and synthetic data (R1 and R4) while also addressing cryptographic security measures, including encryption in transit, encryption at rest (R3), and no encryption [18,46]. Privacy-preserved detection and new frameworks are on the rise [69].

System Architecture

The fifth aspect encompasses the technical structure and interaction modalities of AI technologies in vision-based epilepsy monitoring. This aspect categorizes user interfaces ranging from web platforms [3,45], voice assistants (P1 and R2), mobile and desktop apps, and wearable devices [3,10,16]. It also distinguishes user interaction paradigms, including reporting, interactive [3,10], adaptive (R4 and R5), and no interaction because systems only track patients without involving an interaction opportunity directly for this target group, for example, the system extracts interesting video sequences and saves and uploads these for the medical supervisor [17]. Computing paradigm is classified as cloud-based platform, edge-based platform, and local on device [22,46,67,68], while connection types are categorized into Wi-Fi, built-in modem, Ethernet, and Bluetooth [24,39]. Furthermore, this aspect considers system support, such as on call, system setup, help center, daily technical checks, expert data review, and chat [24,39].

Feedback System

The sixth and final aspect pertains to mechanisms for delivering feedback to users. These dimensions represent topics that are more forward-looking and have not yet been fully established in research or practice (Table 3). This aspect categorizes communication mode based on real-time, periodic, event-based, and on-demand feedback (R1, R2, and R4). It also classifies response type as visual, auditory, haptic, or text-based (R2 and R4). The information purpose marks the last characteristic and describes diverse functional objectives, including alerting and warning users, performance evaluation, recommendations, and user learning to enhance self-management (R2). On the basis of feedback, these elements were included to reflect potential future developments and ensure that the taxonomy remains relevant over time. By incorporating this forward-looking perspective, the taxonomy can serve as a foundation for both current classification and future advancements of AI technologies in vision-based epilepsy monitoring.

XSL•FO
RenderX

Table . Frequency analysis based on the scoping review (n=0-40).

Aspect and dimension	Characteristics
Application and context	
Scope	• Detection (n=34, 85%) • Classification (n=24, 60%) • Prediction (n=2, 5%)
Target group	• Patients with epilepsy (n=8, 20%) • Caregivers (n=2, 5%) • Medical professionals (n=33, 82.5%) • Data scientists (n=17, 42.5%)
Environment	• Stationary (n=35, 87.5%) • Mobile (n=7, 17.5%)
Seizure classification	• Nonmotor symptoms (n=8, 20%) • Motor symptoms (n=37, 92.5%)
Period of epilepsy	• Interictal (n=9, 22.5%) • Preictal (n=3, 7.5%) • Ictal (n=37, 92.5%) • Postictal (n=3, 7.5%)
Data acquisition source	• Depth sensors (n=9, 22.5%) • Infrared (n=20, 50%) • 2D camera (n=36, 90%) • 3D camera (n=5, 12.5%) • Video-EEG[a] (n=21, 52.5%) • ECG[b] (n=3, 7.5%) • Audio (n=6, 15%)
Visual analysis	
Tracking target	• Body (n=30, 75%) • Face (n=13, 32.5%) • Sleeping area (n=23, 57.5%) • Room overview (n=2, 5%)
Video tracking	• MD[c] (n=30, 75%) • BC[d] (n=1, 2.5%) • MOI[e] (n=16, 40%) • ROI[f] (n=17, 42.5%) • HMD[g] (n=3, 7.5%) • SKS[h] (n=8, 20%) • Appearance and feature-based methods (n=31, 77.5%)
Image processing	• Optical flow (n=13, 32.5%) • Frame differencing (n=3, 7.5%) • Spatiotemporal interest points (n=6, 15%) • Contrast-based analysis (n=11, 27.5%)
AI[i] model	
Type of classifier	• SVM[j] (n=8, 20%) • RF[k] (n=2, 5%) • MP[l] (n=4, 10%) • CNN[m] (n=22, 55%) • GMM[n] (n=3, 7.5%) • LSTM[o] (n=11, 27.5%) • I3D[p] (n=4, 10%) • Other types of classifier (n=13, 32.5%)

XSL•FO
RenderX

Aspect and dimension	Characteristics
Performance metrics	• Accuracy (n=20, 50%) • Sensitivity (n=25, 62.5%) • Specificity (n=16, 40%) • F_1-score (n=14, 35%) • Precision (n=15, 37.5%) • Recall (n=8, 20%) • False-positive rate (n=12, 30%) • Area under the curve (n=13, 32.5%)
Market identity	
Medical device	• Certified (n=6, 15%) • Not certified (n=0, 0%) • Proof of concept (n=34, 85%)
Salient attribute	• Environmental robustness (n=4, 10%) • Cost-efficiency (n=4, 10%) • Real-time analysis (n=3, 7.5%) • Ease of use (n=4, 10%) • High system performance (n=3, 7.5%)
Data privacy	• Anonymization (n=7, 17.5%) • Pseudonymization (n=2, 5%) • No privacy-preserving measures (n=0, 0%) • Synthetic data (n=0, 0%)
Cryptographic measure	• Encryption in transit (n=3, 7.5%) • Encryption at rest (n=0, 0%) • No encryption (n=2, 5%)
System architecture	
User interface	• Web platform (n=6, 15%) • Voice assistant (n=0, 0%) • Mobile application (n=4, 10%) • Desktop application (n=3, 7.5%) • Wearable device (n=1, 2.5%)
User interaction	• Reporting (n=6, 15%) • Interactive (n=10, 25%) • Adaptive (n=0, 0%) • No interaction (n=1, 2.5%)
Computing paradigm	• Cloud-based platform (n=7, 17.5%) • Edge-based platform (n=2, 5%) • Local on device (n=14, 35%)
Connection type	• Wi-Fi (n=2, 5%) • Built-in-modem (n=1, 2.5%) • Ethernet (n=3, 7.5%) • Bluetooth (n=1, 2.5%)
Support	• On call (n=1, 2.5%) • System setup (n=3, 7.5%) • Help center (n=0, 0%) • Daily technical checks (n=3, 7.5%) • Expert data review (n=3, 7.5%) • Chat (n=0, 0%)
Feedback system	

XSL•FO
RenderX

Aspect and dimension	Characteristics
Communication mode	• Real time (n=2, 5%) • Periodic (n=2, 5%) • Event-based (n=3, 7.5%) • On demand (n=5, 12.5%)
Response type	• Visual (n=4, 10%) • Auditory (n=1, 2.5%) • Haptic (n=0, 0%) • Text-based (n=4, 10%)
Information purpose	• Alerting and warning (n=2, 5%) • Performance evaluation (n=4, 10%) • Recommendation (n=0, 0%) • User learning (n=0, 0%)

[a]EEG: electroencephalogram.
[b]ECG: electrocardiogram.
[c]MD: movement dynamics.
[d]BC: biomechanical characteristics.
[e]MOI: movement of interest.
[f]ROI: region of interest.
[g]HMD: head movement detection.
[h]SKS: simple keypoint system.
[i]AI: artificial intelligence.
[j]SVM: support vector machine.
[k]RF: random forest.
[l]MP: multilayer perceptron.
[m]CNN: convolutional neural network.
[n]GMM: Gaussian mixture model.
[o]LSTM: long short-term memory.
[p]I3D: inflated 3D.

Demonstration and Evaluation

All 9 experts expressed explicit satisfaction with the taxonomy's scope and structure, confirming that it was well aligned with both theoretical rigor and practical relevance. Every identified object was examined, each characteristic classified at least one object, no further dimensions or characteristics were introduced, and cell combinations were unique and nonredundant, thereby fulfilling the objective ending conditions in iteration 4 and concluding the taxonomy development process. In line with the *Introduction*'s stated purpose (why, how, and what), these results complete the formative, *ex-ante* stage of the *ETDP* and motivate a summative check of utility.

Building on this confirmation, the same expert group proceeded to appraise whether the taxonomy is concise, comprehensive, robust, explainable, and extensible for its intended use in iteration 5 to evaluate the subjective ending conditions. Panelists reported that the taxonomy is *concise*, noting an absence of redundant elements while allowing for focused extensions was helpful. Experts judged the taxonomy as *comprehensive* and the coverage to be largely sufficient for the domain (P1-P3 and R1-R6). One expert described the final taxonomy as "very comprehensive and complete in the categories considered" (R1). Furthermore, experts indicated that guidance remains reliable for nonstandard or edge scenarios, characterizing *robustness* as good and emphasizing maintained flexibility and precision (P1-P3 and R1-R6). The experts reported that the taxonomy's structure supports understanding; distinctions were seen as transparent and traceable, and the systematic differentiation was said to improve *explainability* and provide a comprehensive overview (P1-P3 and R1-R6). It is acknowledged that certain terminology may not be comprehensible to specialists. For instance, medical terminology may not be readily accessible to IT professionals, and conversely, IT terminology may not be readily comprehensible to medical staff. The incorporation of a detailed description or an extended legend accompanied by comprehensive explanations could prove beneficial in this context (R1 and R6). Facing the condition of *extensibility*, the forward compatibility was consistently affirmed. Experts noted that, due to its modular structure, the taxonomy can be extended by adding elements without rewriting existing ones (P1-P3 and R1-R6). In aggregate, experts assessed the subjective ending conditions as fulfilled, indicating that the taxonomy is both theoretically coherent and practically usable for its intended purposes.

Experts judged the taxonomy to be *useful* for its intended target group and purpose (P1-P3 and R1-R6). Panelists provided convergent evidence that the taxonomy is practically useful for analysis, comparison, communication, and decisions. They noted that, in clinical settings, the taxonomy functions as a common language for appraising systems and coordinating with nonclinical stakeholders, ultimately aiding selection and

XSL•FO
RenderX

configuration. Clinicians reportedly view it as a shared frame that supports systematic comparison of alternatives and clearer stakeholder communication, thereby informing practical choices in deployment and workflow integration. "This taxonomy could facilitate a systematic analysis of the landscape of AI-based epilepsy monitoring and drive targeted research and development in areas where it is most required," declares one expert (R5). Researchers indicated that the taxonomy enables structured synthesis and comparability of studies, facilitates the identification of gaps, and provides a stable basis for cumulative evidence building. From a clinical standpoint, experts reported that the taxonomy offers a common structure for evaluating candidate systems and coordinating expectations among care teams and technology partners, which, in turn, supports context-appropriate choices (R4). From a research perspective, it was described as enabling consistent coding, comparison, and aggregation of evidence, while surfacing neglected areas for future work (R2 and R4). In development contexts, experts indicated that it makes market and clinical requirements more transparent and helps align design trade-offs with medical preconditions (R4 and R6). From an industry perspective, experts indicated that the taxonomy clarifies domain constraints and medical prerequisites, translating needs into design requirements and product positioning (R4 and R6). One expert highlighted that the clear delineation of categories creates shared understanding and a robust foundation for decision-making (R1), whereas another underscored that such structure can advance work in areas with the greatest unmet need (R2). Overall, experts characterized usefulness as high and emphasized that the taxonomy's clear demarcations promote shared understanding and provide a dependable decision base (P1-P3 and R1-R6).

Viewed in its entirety, the evaluation examined whether the taxonomy is sufficient, clear, applicable, and extensible for its intended use. Complementing the formative, ex-ante checks applied during construction (objective and subjective ending conditions across iterative C2E and E2C cycles), the Delphi study offered an ex-post assessment by domain experts. Their consensus confirmed that the taxonomy's scope and structure are well aligned with the field and that no further dimensions were required (P1-P3 and R1-R6). This step closes the ETDP loop regarding the demonstration and evaluation phase. The developed taxonomy is theoretically coherent and practically usable for classifying AI technologies in vision-based epilepsy monitoring and for supporting consistent study design and reporting.

Communication: Frequency Analysis

As demonstrated in iteration 3 of the practical assessment, the market analysis revealed that many systems have not yet reached market readiness. Given this constrained evidence base, we included one real-world application as an illustrative use case rather than as the basis for a comparative market analysis. Specifically, we provide a detailed classification example of the market-ready system NELLI (Multimedia Appendix 11) [39]. As the number of real-world applications was insufficient for robust market-level inference, the subsequent synthesis focused on the scoping review evidence. To synthesize insights and highlight gaps, we conducted a frequency analysis across the studies of the scoping review (Multimedia Appendix 7). The analysis included 40 studies with multiple coding permitted per dimension when single studies implemented more than one approach. Therefore, frequencies represent counts of appearances rather than mutually exclusive proportions. The corresponding cases are reported transparently below (Table 3).

The aspect of *application and context* for AI technologies in vision-based epilepsy monitoring reveals the predominant aims to be detection (n=34, 85%) and classification (n=24, 60%), while prediction is almost absent (n=2, 5%). This pattern underscores a focus on immediate recognition of ictal events consistent with clinical priorities, whereas predictive systems remain an open area, likely reflecting their higher methodological complexity. Multiple entries for detection and classification reflect multiscope designs within the same study [17,18,55,56,62]. Target group analysis shows that the overwhelming majority of technologies are tailored to medical professionals (n=33, 82.5%) and secondly to data scientists (n=17, 42.5%), with only marginal attention given to patients with epilepsy themselves (n=8, 20%) or caregivers (n=2, 5%). This suggests a prevailing research trend toward supporting diagnostic and treatment processes within clinical workflows while largely neglecting patient-centered or caregiver-focused designs. The environmental context predominantly involves stationary settings (n=35, 87.5%) compared to mobile scenarios (n=7, 17.5%), reflecting the common use of video-EEG laboratories and controlled observation environments. Similarly, seizure classification heavily favors motor symptoms (n=37, 92.5%) over nonmotor symptoms (n=8, 20%), likely because motor phenomena are more easily captured via visual data, whereas nonmotor events often require multimodal sensing or subjective reporting. Regarding the period of epilepsy, the ictal phase is the principal focus (n=37, 92.5%), followed by preictal (n=3, 7.5%) and postictal (n=3, 7.5%) phases, with limited attention to interictal periods (n=9). This underlines a current emphasis on real-time seizure monitoring over long-term interictal assessment. The data acquisition landscape shows a diverse application of technologies: depth sensors (n=9, 22.5%), infrared (n=20, 50%), 2D cameras (n=36, 90%), 3D cameras (n=5, 12.5%), and video-EEG (n=21, 52.5%) are all well represented. Multimodal systems based on electrocardiography (n=3, 7.5%) and audio (n=6, 15%) are far less frequent, indicating that more multimodal approaches are necessary. Multiple entries for data sources arise when studies use several modalities concurrently [18,42,56,66,70,71].

The *visual analysis* also highlights the clear dominance of body-focused tracking (n=30, 75%) and sleeping area targeting (n=23, 57.5%), which likely reflects the typical contexts of nocturnal seizure detection. Multiple entries occur where similar events span body and sleeping area [44,60,70]. Face tracking (n=13, 32.5%) appears as a secondary priority, while room-wide overviews (n=2, 5%) are notably rare. For video tracking operators, most studies use movement dynamics (n=30, 75%), and many adopt appearance or feature-based pipelines (n=31, 77.5%). Movement of interest is reported in 16 studies (n=16, 40%), region of interest is reported in 17 studies (n=17, 42.5%), head movement detection appears in 3 studies (n=3, 7.5%),

XSL•FO
RenderX

simple keypoint system in 8 studies (n=8, 20%), and biomechanical characteristics in 1 study (n=1, 2.5%). Multiple entries were made where single studies combined several operators [40,47,54,55]. For image processing, optical flow is comparatively common (n=13, 32.5%), followed by contrast-based analysis (n=11, 27.5%), whereas frame differencing (n=3, 7.5%) and spatiotemporal interest points (n=6, 15%) are used infrequently.

Within the aspect of *AI model*, CNNs are the most used classifiers (n=22, 55%), closely followed by LSTMs (n=11, 27.5%), with other models such as support vector machine (n=8), inflated 3D (n=4, 10%), multilayer perceptron (n=4, 10%), Gaussian mixture model (n=3, 7.5%), and random forest (n=2, 5%) used considerably less. This trend underscores the preference for DL approaches suited for complex spatiotemporal patterns in video data. In this dimension, multiple entries were acceptable because of multiple approaches in studies within AI models based on, for example, LSTM and CNN [43,44,51,65]. Performance evaluation predominantly relies on sensitivity (n=25, 62.5%), accuracy (n=20, 50%), specificity (n=16, 40%), precision (n=15, 37.5%), area under the curve (n=13, 32.5%), and F_1-score (n=14, 35%), while more nuanced metrics such as recall (n=8, 20%) and false-positive rates (n=12, 30%) are less consistently reported. This points to a need for more comprehensive benchmarking that captures the practical trade-offs between detection sensitivity and false alarm rates.

Market identity is dominated by proof-of-concept solutions (n=34, 85%), with relatively few certified medical devices (n=6, 15%). Reported salient attributes cluster around environmental robustness (n=4, 10%), cost-efficiency (n=4, 10%), real-time analysis (n=3, 7.5%), ease of use (n=4, 10%), and high system performance (n=3, 7.5%). Privacy and security remain underreported. Anonymization (n=7, 17.5%) and pseudonymization (n=2, 5%) are occasionally specified; synthetic data (n=0, 0%) and explicit no privacy preserving measure (n=0, 0%) do not appear. Encryption in transit is sometimes stated (n=3, 7.5%), whereas encryption at rest (n=0, 0%) is not, and no encryption (n=2, 5%) is rarely reported.

Reporting on *system architecture* is uneven. User interfaces include web platforms (n=6, 15%), mobile apps (n=4, 10%), desktop apps (n=3, 7.5%), and wearables (n=1, 2.5%), with voice assistants (n=0, 0%). User interaction is mainly interactive (n=10, 25%) and reporting (n=6, 15%), less with no interaction (n=1, 2.5%), and adaptive interactions are absent (n=0, 0%). Computing paradigms split between cloud-based (n=7, 17.5%) and local on device (n=14, 35%), with edge-based platform (n=2, 5%). The connection type is rarely specified with Wi-Fi (n=2, 5%), built-in modem (n=1, 2.5%), Ethernet (n=3, 7.5%), and Bluetooth (n=1, 2.5%). Support modalities show system setup (n=3), daily technical checks (n=3, 7.5%), expert data review (n=3, 7.5%), and on call (n=1, 2.5%), but help center and chat are both absent (n=0, 0%).

To complete the results, the *feedback system* indicates that most implementations are geared toward retrospective review rather than immediate intervention. Communication modes are led by on-demand use (n=5, 12.5%), with periodic and event-based configurations occurring less frequently (each n=2, 5%), and true real-time communication remaining uncommon (n=2, 5%). Response types are dominated by visual and text-based outputs (each n=4, 10%), while auditory responses are rarely reported (n=1, 2.5%) and haptic feedback is absent (n=0, 0%). In terms of purpose, the literature primarily documents performance evaluation (n=4, 10%), with comparatively few systems supporting alerting or warning (n=2, 5%) and none providing recommendation or user-learning functions (each n=0).

This frequency analysis provides a structured overview of the current landscape of vision-based AI technologies for epilepsy monitoring and highlights prevailing research emphases as well as underexplored areas. By quantifying recurring dimensions and characteristics, this analysis complements the taxonomy by offering an additional empirical perspective on where the field currently concentrates its efforts and where evidence remains sparse. Beyond summarizing patterns, the frequency analysis increases the practical utility of the taxonomy by enabling systematic comparison across studies and supporting identification of research gaps. In practical terms, the resulting overview can support multiple stakeholder groups: researchers may use it to classify and contrast approaches and to prioritize future investigations, clinicians and decision-makers may use it to appraise the maturity and fit of available technologies for specific care contexts, and developers or companies may use it to benchmark solution profiles and identify opportunities for product development.

Discussion

Principal Findings

In this scoping review of vision-based AI technologies for epilepsy monitoring, we synthesized the evidence into a comprehensive taxonomy structured around application and context, visual analysis, AI model, system architecture, and feedback system, which was subsequently evaluated and refined through domain-expert input. Mapping the included sources to the taxonomy shows that current work is dominated by seizure detection and classification (prediction remains rare) and is largely situated in stationary, hospital ward settings, with comparatively limited evidence from home or residential contexts. Moreover, the field currently shows limited evaluation maturity: most systems remain proof of concept or pilot stage, while deployed solutions are rarely reported.

Interpretation and Implications

Building on these principal findings, we derive a set of implications for research and practice (summarized in Multimedia Appendix 12), aimed at supporting researchers, clinicians, developers, policymakers, and patient advocacy groups.

Predictive Approaches Remain Underrepresented

Across the reviewed literature, vision-based seizure monitoring is predominantly framed as a recognition task focused on detection and classification, whereas prediction is rarely examined. This aligns with prior work describing seizure monitoring as primarily reactive rather than anticipatory (eg, time-to-event) in its current operationalization [3,5,17]. Potential explanations discussed in the literature include the limited

XSL•FO
RenderX

specificity of preictal visual cues, constraints in obtaining sufficiently annotated preictal recordings, and difficulties in achieving robust generalization across patients and real-world settings [13]. Moreover, the most reliable visual markers typically appear at or after ictal onset, inherently biasing research toward detection rather than forecasting [5,51]. Addressing this gap requires methodological advances, multimodal data integration, and careful validation of predictive accuracy in real-world settings, which has been achieved outside vision-based monitoring systems [72].

Clinical and Patient-Centered Translation Remains Uneven

The evidence base emphasizes stationary clinical environments and professional workflows, while home or residential contexts are comparatively less represented. This pattern is consistent with a field in which technical feasibility and controlled evaluation conditions often precede broader implementation, particularly when systems must operate under variable lighting, camera placement, occlusion, and caregiver routines outside specialized clinical units [5,22]. Clinical readiness is a prerequisite for successful out-of-clinic use [22,67]. Establishing efficacy, safety, and operability in clinical workflows should come before extending to home environments, where variability and resource constraints otherwise amplify failure modes [5,22]. Sustained adoption in home environments hinges on usability and human factors, such as simple setup, low cognitive and technical demands, clear feedback, and an alarm burden that is manageable and fits with caregiver routines [3,16,22]. We therefore recommend co-design with patients and caregivers, treating out-of-clinic expansion as a distinct development step, requiring explicit attention to robustness, usability, and operational support.

Nonmotor Seizures and Multimodal Sensing Remain Less Often Addressed

Body-focused tracking and monitoring of the sleeping area are common targets, reflecting the clinical emphasis on nocturnal seizure detection. In contrast, more advanced or multimodal approaches, such as face tracking, room-wide overviews, or recognition of nonmotor phenomena, are less frequently described, indicating that visual monitoring remains narrow in scope and often captures overt motor activity in controlled settings. Current systems rely on visual cues to capture motor symptoms, with less attention to subtler phenomena that require physiological, audio, or multimodal data fusion [5,45,73]. Data fusion (eg, EEG and accelerometers) and adaptive filtering techniques can improve robustness [6,13]. Future research should expand beyond motor seizure detection and explore integrated sensing pipelines capable of handling diverse seizure presentations to increase noise resistance for real-world deployment.

System Maturity Highlights Gaps in Privacy and Regulatory Preparedness

Many solutions remain at the proof-of-concept stage, with limited progress toward certified medical devices. Privacy and security measures are also inconsistently reported: anonymization, pseudonymization, or encryption practices are seldom described as explicit system properties. In the European context, this is notable because data protection obligations under the General Data Protection Regulation and emerging requirements under the EU AI Act place increasing emphasis on governance, transparency, and safeguards for high-risk medical AI systems as they move toward deployment [74]. Work aiming for implementation may therefore benefit from integrating privacy-by-design and documentation-ready development practices earlier in the development.

Real-Time Feedback and Alerting Mechanisms Remain Uncommon

Many systems emphasize retrospective review and post hoc analysis, whereas real-time monitoring and responsive feedback mechanisms are less commonly addressed. This emphasis suggests that current solutions are often optimized for documentation and clinical review rather than time-critical response. For deployment-oriented use cases, future systems should specify the intended feedback objective (eg, documentation vs immediate escalation) and evaluate feedback as a system-level property, including latency, notification logic, and false alarm burden. Some existing systems already report latency and computational inefficiencies [5,45]. Optimizing real-time architectures through edge computing, efficient signal processing, and context-aware AI may reduce latency and support more reliable real-time operation in both clinical and home-based settings [62].

Taken together, these findings indicate a field that is technologically advanced in seizure recognition yet often focused on professional, stationary, and retrospective use cases. Important opportunities for predictive, patient-centered, multimodal, and privacy-preserving solutions remain largely unaddressed. These results suggest the need for standardized reporting of feedback latency, notification channels, escalation logic, and end-user targets to enable comparability and to assess clinical readiness. From a theoretical perspective, these patterns support the need for a maturity-oriented framework that integrates deployment setting, system maturity, data provenance, task type (detection, classification, and prediction), and reporting quality as core explanatory dimensions. The taxonomy developed in this scoping review operationalizes these dimensions and thereby contributes to conceptual standardization in the field. Overall, the implications outline a pathway from detection-focused development toward robust, patient-centered, explainable, and regulation-ready systems with the potential to advance epilepsy care, while preserving the empirical constraints and opportunities identified by the taxonomy.

Limitations

This scoping review has several limitations. Despite a rigorous approach, the scoping review may still reflect selection bias because it was screened and selected by the reviewers, restricted to 3 bibliographic databases, and relied on predefined keywords and inclusion criteria. The exclusion of studies published in languages other than English has the potential to constitute an additional bias. Although the scoping review was conducted across 3 key databases commonly used in health and AI research, scoping reviews may benefit from an even more

XSL•FO
RenderX

extensive search strategy, incorporating additional databases or gray literature, to further enhance methodological completeness. However, expert validation and iterative development ensure its robustness. Furthermore, the iterative E2C methodology inherently relies on interpretative refinements, introducing contextual dependencies. Nonetheless, this approach is well established in taxonomy development, allowing adaptability to technological and clinical advancements. The status quo on market-ready solutions appears limited in the landscape of AI monitoring systems that are based on visual data. NELLI [39] provided a basis for the iteration of practical assessment. To obtain a more comprehensive perspective on additional market solutions in the close field, the focus was directed toward a video-EEG solution known as SEER [24]. The performance metrics within AI models pose another challenge, given the multitude of metrics that can be considered. Furthermore, despite the underreporting of privacy and cryptographic safeguards in the literature, a jurisdiction-specific compliance analysis was not undertaken (eg, AI Act). This work does not include a detailed, jurisdiction-specific compliance mapping, which will be addressed in a dedicated follow-up study. Finally, the validation to date relied on the contributions of domain experts; patient and informal caregiver perspectives, and prospective real-world deployments were not yet included.

Conclusions

This scoping review points to a broader transition problem in vision-based epilepsy monitoring: while algorithmic feasibility for seizure recognition is increasingly demonstrated, translation into deployable, trustworthy systems remains constrained by system-level requirements. Progress toward deployment in this domain will likely depend less on incremental model improvements and more on development and evaluation practices that explicitly address readiness for real-world operation, including robustness under uncontrolled conditions, workflow fit, feedback and escalation design, and governance mechanisms that make privacy protection and safety assurance auditable and maintainable over time.

The developed taxonomy provides a shared reference structure for describing and evaluating vision-based epilepsy monitoring technologies end to end. Its innovation lies in integrating application context with visual analysis, AI modeling, system architecture, and feedback design within a single classification framework. Where prior work has often centered on algorithms, datasets, or modality-specific performance summaries, this taxonomy enables consistent, system-level characterization across research prototypes and emerging solutions. By standardizing how systems are characterized across dimensions and characteristics, it can support benchmarking and implementation-focused evaluation (including procurement considerations) and guide implementation priorities. For implementation stakeholders, the taxonomy can serve as a structured checklist to align a system's intended use and deployment setting with required architecture, feedback design, evaluation evidence, and safeguards, thereby informing selection, integration, and rollout planning. More broadly, the taxonomy supports progress in a diverse and fast-evolving field by making system boundaries, intended use, and evaluation targets explicit and comparable across studies. This can help shift the literature from model-centered reporting toward system-level evidence that can be interpreted across contexts, strengthening the basis for synthesis, replication, and translation. In this sense, the taxonomy is not only a classification tool but also a practical framework to support more consistent study design, reporting, and deployment-oriented evaluation in future research.

Acknowledgments

The authors declare the use of generative artificial intelligence (Gen-AI) in the research and writing process. According to the GAIDeT (Generative Artificial Intelligence Delegation Taxonomy; 2025), the following tasks were delegated to Gen-AI tools under full human supervision: summarizing text, translation, and reformatting. The Gen-AI tool used was ChatGPT (OpenAI; models GPT-5 and GPT-5.2). Responsibility for the final manuscript lies entirely with the authors. Gen-AI tools are not listed as authors and do not bear responsibility for the final outcomes. All Gen-AI outputs were reviewed and verified. This declaration was submitted by the authors. We would like to thank Patricia Kajüter-Rodriguez for reviewing our search strategy in accordance with the Cochrane Handbook and all experts for participating in the Delphi study.

Funding

This research was conducted within the scope of the EPIMON project, which is funded by the Federal Ministry of Research, Technology and Space (BMFTR; grant 13GW0705C). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

Main data generated or analyzed during this scoping review are included in this published article or in the multimedia appendix. The detailed results of the expert opinions are not available due to privacy concerns.

Authors' Contributions

MI, JH, and FT conceptualized the study, with MI providing overall supervision. MI, JH, and AF developed the methodology, performed data curation for the scoping review, and carried out the formal analysis. MI and JH performed validation. MI produced

XSL•FO
RenderX

and interpreted all visualizations and prepared the original draft, and MI, JH, AF, and FT reviewed and edited the manuscript. All authors approved this version and meet the International Committee of Medical Journal Editors authorship criteria.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Exemplary schematic flow of monitoring systems based on visual data with artificial intelligence models.
[PNG File, 322 KB - jmir_v28i1e83895_app1.png]

Multimedia Appendix 2
Synthesis of the extended taxonomy design process: demonstration and evaluation aligned with 3 guiding questions of Design Science Research (table based on Nickerson et al [26]).
[PDF File, 53 KB - jmir_v28i1e83895_app2.pdf]

Multimedia Appendix 3
Overview of all iterations with the identified dimensions during the taxonomy development.
[PNG File, 632 KB - jmir_v28i1e83895_app3.png]

Multimedia Appendix 4
Search method and search strategies for all databases.
[PDF File, 105 KB - jmir_v28i1e83895_app4.pdf]

Multimedia Appendix 5
Comprehensive data extraction tables for taxonomy dimensions.
[PDF File, 98 KB - jmir_v28i1e83895_app5.pdf]

Multimedia Appendix 6
Comprehensive data extraction tables for study levels.
[PDF File, 82 KB - jmir_v28i1e83895_app6.pdf]

Multimedia Appendix 7
Coding template and detailed frequency analysis across all included studies.
[XLSX File, 36 KB - jmir_v28i1e83895_app7.xlsx]

Multimedia Appendix 8
Study sample selection of experts from practice and research.
[PDF File, 34 KB - jmir_v28i1e83895_app8.pdf]

Multimedia Appendix 9
Data charting of the study results.
[XLSX File, 13 KB - jmir_v28i1e83895_app9.xlsx]

Multimedia Appendix 10
Evidence maps of the study results.
[PDF File, 243 KB - jmir_v28i1e83895_app10.pdf]

Multimedia Appendix 11
Taxonomy application example for NELLI.
[PDF File, 124 KB - jmir_v28i1e83895_app11.pdf]

Multimedia Appendix 12
Main findings and implications for research and practice.
[PDF File, 32 KB - jmir_v28i1e83895_app12.pdf]

XSL•FO
RenderX

Checklist 1
PRISMA-S checklist.
[PDF File, 122 KB - jmir_v28i1e83895_app13.pdf]

References

1. Beghi E, Giussani G, Nichols E, et al. Global, regional, and national burden of epilepsy, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol 2019 Apr;18(4):357-375. [doi: 10.1016/S1474-4422(18)30454-X]
2. Epilepsy. World Health Organization. 2025. URL: https://www.who.int/news-room/fact-sheets/detail/epilepsy [accessed 2026-01-22]
3. Meisel C, Holtkamp M, Vock S. Ambulantes langzeit-video-EEG als neuer diagnostischer Ansatz in Deutschland: ergebnisse einer Machbarkeitsstudie. Nervenarzt 2023 Jun;94(6):519-524. [doi: 10.1007/s00115-022-01412-0]
4. Sheehan TA, Winter-Potter E, Dorste A, Meisel C, Loddenkemper T. Veni, vidi, vici-when is home video seizure monitoring helpful? Epilepsy Curr 2025;25(1):9-16. [doi: 10.1177/15357597241253426] [Medline: 39554274]
5. Rai P, Knight A, Hiillos M, et al. Automated analysis and detection of epileptic seizures in video recordings using artificial intelligence. Front Neuroinform 2024;18. [doi: 10.3389/fninf.2024.1324981]
6. Chen K, Gabriel P, Alasfour A, et al. Patient-specific pose estimation in clinical environments. IEEE J Transl Eng Health Med 2018;6:2101111. [doi: 10.1109/JTEHM.2018.2875464] [Medline: 30483453]
7. Fazli Besheli B, Sha Z, Gavvala JR, et al. Using high-frequency oscillations from brief intraoperative neural recordings to predict the seizure onset zone. Commun Med (Lond) 2024 Nov 25;4(1):243. [doi: 10.1038/s43856-024-00654-0] [Medline: 39587325]
8. Wang Z, Song X, Chen L, et al. Research progress of epileptic seizure prediction methods based on EEG. Cogn Neurodyn 2024 Oct;18(5):2731-2750. [doi: 10.1007/s11571-024-10109-w] [Medline: 39555266]
9. Assim OM, Mahmood AF. A novel universal deep learning approach for accurate detection of epilepsy. Med Eng Phys 2024 Sep;131(1):104219. [doi: 10.1016/j.medengphy.2024.104219] [Medline: 39284648]
10. Nurse ES, Perera T, Hannon T, Wong V, Fernandes KM, Cook MJ. Rates of event capture of home video EEG. Clin Neurophysiol 2023 May;149(12–17):12-17. [doi: 10.1016/j.clinph.2023.02.165]
11. Brunnhuber F, Slater J, Goyal S, et al. Past, present and future of home video-electroencephalographic telemetry: a review of the development of in-home video-electroencephalographic recordings. Epilepsia 2020 Nov;61 Suppl 1(S1):S3-S10. [doi: 10.1111/epi.16578] [Medline: 32683686]
12. Singh AK, Rana A, Kapse V. Non-EEG-based seizure detection: a comprehensive review. Presented at: 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE); May 9-11, 2024; Gautam Buddha Nagar, India p. 31-34. [doi: 10.1109/IC3SE62002.2024.10593241]
13. Ahmedt-Aristizabal D, Armin MA, Hayder Z, et al. Deep learning approaches for seizure video analysis: a review. Epilepsy Behav 2024 May;154:109735. [doi: 10.1016/j.yebeh.2024.109735] [Medline: 38522192]
14. Garção VM, Abreu M, Peralta AR, Bentes C, Fred A, P. da Silva H. A novel approach to automatic seizure detection using computer vision and independent component analysis. Epilepsia 2023 Sep;64(9):2472-2483. [doi: 10.1111/epi.17677]
15. Kaur T, Diwakar A, et al. Artificial Intelligence in epilepsy. Neurol India 2021;69(3):560-566. [doi: 10.4103/0028-3886.317233] [Medline: 34169842]
16. Armand Larsen S, Terney D, Østerkjerhuus T, et al. Automated detection of nocturnal motor seizures using an audio-video system. Brain Behav 2022 Sep;12(9):e2737. [doi: 10.1002/brb3.2737] [Medline: 35939047]
17. Peltola J, Basnyat P, Armand Larsen S, et al. Semiautomated classification of nocturnal seizures using video recordings. Epilepsia 2023 Dec;64 Suppl 4(S4):S65-S71. [doi: 10.1111/epi.17207] [Medline: 35194778]
18. Karacsony T, Loesch-Biffar AM, Vollmar C, Noachtar S, Cunha JPS. DeepEpil: towards an epileptologist-friendly ai enabled seizure classification cloud system based on deep learning analysis of 3D videos. Presented at: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI); Jul 27-30, 2021. [doi: 10.1109/BHI50953.2021.9508555]
19. Chen Y, Wu M. Artificial intelligence-enabled contactless sensing for medical diagnosis. Medical Review 2023 Jun 27;3(3):195-197. [doi: 10.1515/mr-2023-0022]
20. Song Y, Taylor W, Ge Y, Dashtipour K, Imran MA, Abbasi QH. Design and implementation of a contactless AI-enabled human motion detection system for next-generation healthcare. Presented at: 2021 IEEE International Conference on Smart Internet of Things (SmartIoT); Aug 13-15, 2021; Jeju, Republic of Korea p. 112-119. [doi: 10.1109/SmartIoT52359.2021.00027]
21. Karácsony T, Jeni LA, De la Torre F, Cunha JPS. Deep learning methods for single camera based clinical in-bed movement action recognition. Image Vis Comput 2024 Mar;143:104928. [doi: 10.1016/j.imavis.2024.104928]
22. Garção VM, Abreu M, Sá F, et al. Lampsy: an Invisible real-time epilepsy video monitoring and automatic seizure detection device. IEEE Access 2025;13:96018-96030. [doi: 10.1109/ACCESS.2025.3575528]
23. Brown BM, Boyne AMH, Hassan AM, Allam AK, Cotton RJ, Haneef Z. Computer vision for automated seizure detection and classification: a systematic review. Epilepsia 2024 May;65(5):1176-1202. [doi: 10.1111/epi.17926] [Medline: 38426252]
24. EEG monitoring with video. SEER Medical. 2024. URL: https://seermedical.com/ [accessed 2026-01-22]

XSL•FO
RenderX

25. Kerr WT, McFarlane KN, Figueiredo Pucci G. The present and future of seizure detection, prediction, and forecasting with machine learning, including the future impact on clinical trials. Front Neurol 2024;15:1425490. [doi: 10.3389/fneur.2024.1425490] [Medline: 39055320]
26. Nickerson RC, Varshney U, Muntermann J. A method for taxonomy development and its application in information systems. Eur J Inf Syst 2013 May;22(3):336-359. [doi: 10.1057/ejis.2012.26]
27. Kundisch D, Muntermann J, Oberländer AM, et al. An update for taxonomy designers: methodological guidance from information systems research. Bus Inf Syst Eng 2022;64:421-439. [doi: 10.1007/s12599-021-00723-x]
28. Venable J, Pries-Heje J, Baskerville R. FEDS: a framework for evaluation in Design Science Research. Eur J Inf Syst 2016 Jan;25(1):77-89. [doi: 10.1057/ejis.2014.36]
29. Prat N, Comyn-Wattiau I, Akoka J. A taxonomy of evaluation methods for information systems artifacts. J Manag Inf Syst 2015 Jul 3;32(3):229-267. [doi: 10.1080/07421222.2015.1099390]
30. Vom Brocke J, Simons A, Niehaves B, Riemer K, Plattfaut R, Cleven A. Reconstructing the giant: on the importance of rigour in documenting the literature search process. Presented at: 17th European Conference on Information Systems; Jun 8-10, 2009 URL: https://aisel.aisnet.org/ecis2009/161/ [accessed 2026-01-22]
31. Peters M, Godfrey C, McInerney P, Trico A, Khalil H, Munn Z. Scoping reviews. In: JBI Manual for Evidence Synthesis 2020. [doi: 10.46658/JBIRM-20-01]
32. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850]
33. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
34. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017 May 24;60(6):84-90. [doi: 10.1145/3065386]
35. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS peer review of electronic search strategies: 2015 guideline statement. J Clin Epidemiol 2016 Jul;75(40–46):40-46. [doi: 10.1016/j.jclinepi.2016.01.021] [Medline: 27005575]
36. Bramer WM, Giustini D, de Jonge GB, Holland L, Bekhuis T. De-duplication of database search results for systematic reviews in EndNote. J Med Libr Assoc 2016 Jul;104(3):240-243. [doi: 10.3163/1536-5050.104.3.014]
37. Gallego D, Bueno S. Exploring the application of the Delphi method as a forecasting tool in information systems and technologies research. Technol Anal Strateg Manag 2014 Oct 21;26(9):987-999. [doi: 10.1080/09537325.2014.941348]
38. Crunchbase search (keywords: “epilepsy” and “video monitoring”). Crunchbase. 2026. URL: https://www.crunchbase.com/ [accessed 2026-01-22]
39. Smart solution for seizure detection, combining AI and human expertise. Neuroeventlabs. 2024. URL: https://neuroeventlabs.com/ [accessed 2026-01-22]
40. Zwettler GA, Reichhardt A, Stradner M. EpiMon: vision-based early warning system for monitoring uprising epileptic seizures during night. 2021 Presented at: The 10th International Workshop on Innovative Simulation for Healthcare; Sep 15-17, 2021; Genova, Italy p. 17-26. [doi: 10.46354/i3m.2021.iwish.004]
41. Pediaditis M, Ciubotaru AN, Brunschwiler T, et al. Machine learning techniques for personalized detection of epileptic events in clinical video recordings. AMIA Annu Symp Proc 2020;2020:1003-1011. [Medline: 33936476]
42. Cao J, Fang Y, Cui X, Zheng R, Jiang T, Gao F. Synchronized video and EEG based childhood epilepsy seizure detection. IEEE Trans Emerg Top Comput Intell 2024;8(6):3742-3753. [doi: 10.1109/TETCI.2024.3372387]
43. Ding L, Wang H, Fu L. STSE-xlstm: a deep learning framework for automated seizure detection in long video sequences using spatio-temporal and attention mechanisms. Presented at: 2024 10th International Conference on Computer and Communications (ICCC); Dec 13-16, 2024; Chengdu, China p. 781-785. [doi: 10.1109/ICCC62609.2024.10942266]
44. Martini ML, Valliani AA, Sun C, et al. Deep anomaly detection of seizures with paired stereoelectroencephalography and video recordings. Sci Rep 2021 Apr 5;11(1):7482. [doi: 10.1038/s41598-021-86891-y] [Medline: 33820942]
45. Andersson FK, Gauffin H, Lindehammar H, Vigren P. Video-based automatic seizure detection in pharmacoresistant epilepsy: a prospective exploratory study. Epilepsy Behav 2024 Dec;161:110118. [doi: 10.1016/j.yebeh.2024.110118] [Medline: 39536364]
46. Karácsony T, Fearns N, Vollmar C, et al. NeuroKinect4K: a novel 4K RGB-D-IR video system with 3D scene reconstruction for enhanced epileptic seizure semiology monitoring. Presented at: 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); Jul 15-19, 2024. [doi: 10.1109/EMBC53108.2024.10781546]
47. Hou JC, Thonnat M, Bartolomei F, McGonigal A. Automated video analysis of emotion and dystonia in epileptic seizures. Epilepsy Res 2022 Aug;184:106953. [doi: 10.1016/j.eplepsyres.2022.106953]
48. Diop S, Essid N, Jouen F, Bergounioux J, Trabelsi I. Adapting action recognition neural networks for automated infantile spasm detection. IEEE Trans Neural Syst Rehabil Eng 2024;32:3751-3760. [doi: 10.1109/TNSRE.2024.3472088]
49. Hanosh O, Ansari R, Younis K, Cetin AE. Real-time epileptic seizure detection during sleep using passive infrared sensors. IEEE Sensors J 2019 Aug 1;19(15):6467-6476. [doi: 10.1109/JSEN.2019.2907664]

XSL•FO
RenderX

50. Yang Y, Sarkis RA, Atrache RE, Loddenkemper T, Meisel C. Video-based detection of generalized tonic-clonic seizures using deep learning. IEEE J Biomed Health Inform 2021 Aug;25(8):2997-3008. [doi: 10.1109/JBHI.2021.3049649] [Medline: 33406048]
51. Ahmedt-Aristizabal D, Fookes C, Nguyen K, Denman S, Sridharan S, Dionisio S. Deep facial analysis: a new phase I epilepsy evaluation using computer vision. Epilepsy Behav 2018 May;82:17-24. [doi: 10.1016/j.yebeh.2018.02.010]
52. Ahmedt-Aristizabal D, Fookes C, Dionisio S, Nguyen K, Cunha JPS, Sridharan S. Automated analysis of seizure semiology and brain electrical activity in presurgery evaluation of epilepsy: a focused survey. Epilepsia 2017 Nov;58(11):1817-1831. [doi: 10.1111/epi.13907] [Medline: 28990168]
53. Ahmedt-Aristizabal D, Nguyen K, Denman S, Sridharan S, Dionisio S, Fookes C. Deep motion analysis for epileptic seizure classification. Presented at: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); Jul 18-21, 2018. [doi: 10.1109/EMBC.2018.8513031]
54. Karácsony T, Loesch-Biffar AM, Vollmar C, Rémi J, Noachtar S, Cunha JPS. Novel 3D video action recognition deep learning approach for near real time epileptic seizure classification. Sci Rep 2022 Nov 15;12(1):19571. [doi: 10.1038/s41598-022-23133-9] [Medline: 36379994]
55. IEEE staff. In: 2012 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society: IEEE; 2012.
56. Ojanen P, Knight A, Hakala A, et al. An integrative method to quantitatively detect nocturnal motor seizures. Epilepsy Res 2021 Jan;169:106486. [doi: 10.1016/j.eplepsyres.2020.106486] [Medline: 33310414]
57. Ahmedt-Aristizabal D, Fookes C, Denman S, et al. A hierarchical multimodal system for motion analysis in patients with epilepsy. Epilepsy & Behavior 2018 Oct;87:46-58. [doi: 10.1016/j.yebeh.2018.07.028]
58. Ahmedt-Aristizabal D, Nguyen K, Denman S, et al. Vision-based mouth motion analysis in epilepsy: a 3D perspective. Presented at: 2019 41st Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Jul 23-27, 2019. [doi: 10.1109/EMBC.2019.8857656]
59. Lin N, Gao W, Li L, et al. vEpiNet: a multimodal interictal epileptiform discharge detection method based on video and electroencephalogram data. Neural Netw 2024 Jul;175:106319. [doi: 10.1016/j.neunet.2024.106319] [Medline: 38640698]
60. Geertsema EE, Thijs RD, Gutter T, et al. Automated video-based detection of nocturnal convulsive seizures in a residential care setting. Epilepsia 2018 Jun;59 Suppl 1(S1):53-60. [doi: 10.1111/epi.14050] [Medline: 29638008]
61. Ahmedt-Aristizabal D, Denman S, Nguyen K, Sridharan S, Dionisio S, Fookes C. Understanding patients' behavior: vision-based analysis of seizure disorders. IEEE J Biomed Health Inform 2019 Nov;23(6):2583-2591. [doi: 10.1109/JBHI.2019.2895855] [Medline: 30714935]
62. van Westrhenen A, Petkov G, Kalitzin SN, Lazeron RHC, Thijs RD. Automated video-based detection of nocturnal motor seizures in children. Epilepsia 2020 Nov;61 Suppl 1(Suppl 1):S36-S40. [doi: 10.1111/epi.16504] [Medline: 32378204]
63. Ogura Y, Hayashi H, Nakashima S, et al. A neural network based infant monitoring system to facilitate diagnosis of epileptic seizures. Presented at: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); Aug 25-29, 2015; Milan p. 5614-5617. [doi: 10.1109/EMBC.2015.7319665]
64. Li J, Zhen X, Liu X, Ouyang G. Classifying normal and abnormal status based on video recordings of epileptic patients. ScientificWorldJournal 2014;2014:1-6. [doi: 10.1155/2014/459636]
65. Eguchi K, Yaguchi H, Nakakubo S, et al. Video-based detection of epileptic spasms in west syndrome using a deep neural network: a pilot case study. J Neurol Sci 2023 Jun 15;449:120671. [doi: 10.1016/j.jns.2023.120671] [Medline: 37167655]
66. Karacsony T, Loesch-Biffar AM, Vollmar C, Noachtar S, Cunha JPS. A deep learning architecture for epileptic seizure classification based on object and action recognition. Presented at: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 4-8, 2020; Barcelona, Spain p. 4117-4121. [doi: 10.1109/ICASSP40776.2020.9054649]
67. Boyne A, Yeh HJ, Allam AK, et al. Video-based detection of tonic-clonic seizures using a three-dimensional convolutional neural network. Epilepsia 2025 Jul;66(7):2495-2506. [doi: 10.1111/epi.18381] [Medline: 40126380]
68. Hu D, Wu K, Fang Y, Jiang T, Gao F, Cao J. STMemAE: an instance-level based spatio-temporal memory autoencoder for unsupervised vision-based seizure detection. IEEE Transactions on Emerging Topics in Computational Intelligence 2025;9(5):3298-3310. [doi: 10.1109/TETCI.2024.3522208]
69. Mehta D, Sivathamboo S, Simpson H, Kwan P, O'Brien T, Ge Z. Privacy-preserving early detection of epileptic seizures in videos. In: Image Computing and Computer Assisted Intervention: MICCAI; 2023. [doi: 10.1007/978-3-031-43904-9_21]
70. Achilles F, Tombari F, Belagiannis V, Loesch AM, Noachtar S, Navab N. Convolutional neural networks for real-time epileptic seizure detection. Comput Methods Biomech Biomed Eng Imaging Vis 2018 May 4;6(3):264-269. [doi: 10.1080/21681163.2016.1141062]
71. Hu D, Fang Y, Cao J, Jiang T, Gao F. An end-to-end vision-based seizure detection with a guided spatial attention module for patient detection. IEEE Internet Things J 2024 May 15;11(10):18869-18879. [doi: 10.1109/JIOT.2024.3367739]
72. Nasseri M, Pal Attia T, Joseph B, et al. Ambulatory seizure forecasting with a wrist-worn device using long-short term memory deep learning. Sci Rep 2021 Nov 9;11(1):21935. [doi: 10.1038/s41598-021-01449-2] [Medline: 34754043]

XSL•FO
RenderX

73. Lin N, Li L, Gao W, et al. Development and validation of a multimodal automatic interictal epileptiform discharge detection model: a prospective multi-center study. BMC Med 2025 Aug 15;23(1):479. [doi: 10.1186/s12916-025-04316-3] [Medline: 40817241]
74. Busch F, Kather JN, Johner C, et al. Navigating the European Union Artificial Intelligence Act for healthcare. NPJ Digit Med 2024 Aug 12;7(1):210. [doi: 10.1038/s41746-024-01213-6] [Medline: 39134637]

Abbreviations

AI: artificial intelligence
C2E: conceptual-to-empirical
CNN: convolutional neural network
DL: deep learning
E2C: empirical-to-conceptual
EEG: electroencephalogram
ETDP: extended taxonomy design process
GAIDeT: Generative Artificial Intelligence Delegation Taxonomy
Gen-AI: generative artificial intelligence
LSTM: long short-term memory
MeSH: Medical Subject Headings
PCC: population, concept, and context
PRESS: Peer Review of Electronic Search Strategies
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RQ: research question

Edited by S Brini; submitted 10.Sep.2025; peer-reviewed by P Bhatnagar, R Marshall; revised version received 03.Mar.2026; accepted 07.Mar.2026; published 24.Jun.2026.

Please cite as:
Irnich M, Hammer J, Flok A, Teuteberg F
Vision-Based Artificial Intelligence Technologies for Epilepsy Monitoring: Scoping Review and Taxonomy Development Study
J Med Internet Res 2026;28:e83895
URL: https://www.jmir.org/2026/1/e83895
doi:10.2196/83895

© Mirijana Irnich, Jonas Hammer, Aleksandra Flok, Frank Teuteberg. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Accessibility of Digital Financial Applications for People With Visual Impairment: Scoping Review

Louise Puli[1,2], MPH, BPO; Lars Kooijman[3,4], PhD; Tanjila Kanij[5], PhD; Charmine Hartel[3,6], PhD; Abu Zafar M Shahriar[1], PhD; Kristian Rotaru[3,7,8], PhD

[1]Department of Banking and Finance, Monash Business School, Monash University, 900 Dandenong Road, Caufield East, Victoria, Australia

[2]Rehabilitation, Ageing and Independent Living Research Centre, School of Primary and Allied Health Care, Monash University, Frankston, Victoria, Australia

[3]Opportunity Tech Lab, Monash Business School, Monash University, Caufield East, Victoria, Australia

[4]Office of the Chief Forensic Scientist, Victoria Police Forensic Services Department, Macleod, Victoria, Australia

[5]Department of Computing Technologies, School of Science, Computing and Emerging Technologies, Swinburne University of Technology, Hawthorn, Victoria, Australia

[6]Department of Management, Monash Business School, Monash University, Caufield East, Victoria, Australia

[7]Department of Accounting, Monash Business School, Monash University, Caufield East, Victoria, Australia

[8]School of Psychological Sciences, Monash University, Clayton, Victoria, Australia

Corresponding Author:
Louise Puli, MPH, BPO
Department of Banking and Finance, Monash Business School, Monash University, 900 Dandenong Road, Caufield East, Victoria, Australia

Abstract

Background: Routine financial activities are now conducted primarily through digital channels. Many such systems remain inaccessible to more than 2.2 billion people globally living with vision impairment, limiting independent financial management. Constrained access can create financial strain and social disadvantage, reducing access to health-enabling resources, and contributing to avoidable health inequities.

Objective: This scoping review maps evidence on the accessibility of digital financial services for individuals with visual impairment (VI) as a digital determinant of health. We synthesized barriers and facilitators, characterized study designs, settings, and populations, and identified evidence gaps to inform inclusive design, digital health research priorities, and policy.

Methods: A scoping review was conducted using the Joanna Briggs Institute framework and reported in line with PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines. Eight databases (PubMed, MEDLINE, CINAHL, Scopus, Web of Science, Business Source Complete, ProQuest, and IEEE Xplore) were searched for peer-reviewed papers in English published between 1995 and 2026. Searches featured controlled vocabulary and free-text terms structured in 3 conceptual blocks (VI, digital financial services, and accessibility or usability). A random sample of 20% of titles, abstracts, full texts, and included studies was independently screened or charted by 2 reviewers to calibrate decisions; the remainder were screened and charted by a single reviewer. Data were charted using a standardized extraction form, and results were synthesized descriptively and thematically.

Results: Twenty-three studies met the inclusion criteria. Studies were conducted across 12 countries, with the largest number from India (n=7), Indonesia (n=2), Thailand (n=2), and the United States (n=2). Study designs included qualitative studies (n=6), mixed methods studies (n=1), cross-sectional studies (n=4), nonrandomized experimental studies (n=2), and technical or design-focused evaluations (n=6). One study was a large population survey (n=19,136), and the remaining studies with human participants had sample sizes ranging from 4 to 36 participants. Accessibility barriers were reported across all platform types, with authentication-related barriers described in 18 studies and screen reader incompatibility in 17 studies. Reported barriers included reliance on sighted assistance for tasks such as login, verification, and payments, compromising privacy and independence. Facilitators included assistive technology support, logical navigation order, nonvisual feedback mechanisms, and accessible authentication alternatives. Evidence mapping revealed recurrent barrier patterns across Android, iOS, and web platforms. No longitudinal or intervention-based evaluations were identified.

Conclusions: This review provides a focused synthesis of accessibility evidence at the intersection of digital financial services and VI, a domain addressed by neither prior digital accessibility reviews nor financial inclusion for people with disabilities. Authentication methods, interface labeling, and navigation were identified as persistent cross-platform accessibility barriers. The findings carry implications for financial technology developers, accessibility auditors, and policymakers implementing accessibility

XSL•FO
RenderX

legislation and extend the digital determinants of health framework by demonstrating how inaccessible financial technology may compound health inequities.

(J Med Internet Res 2026;28:e82315) doi:10.2196/82315

KEYWORDS

visual impairment; accessibility; financial technology; user-centered design; assistive technology; digital inclusion; PRISMA; consumer empowerment; Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Introduction

Globally, 2.2 billion people live with some form of visual impairment (VI), including blindness or low vision [1]. Individuals with VI consistently report reduced quality of life and social exclusion, impacting access to education, employment, housing, and essential services [1-3]. As digital banking, particularly mobile- and web-based apps, becomes the predominant means of accessing financial services [4], concern is increasing that people with VI face substantial barriers to digital financial inclusion [3].

Access to financial resources is recognized as a social determinant of health [5]. Accordingly, access to financial services mediated through digital technologies can affect health outcomes [6]. These interactions, described as digital determinants of health [6], warrant close examination to identify how digital systems create or reinforce barriers for people with disabilities. Consistent with this perspective, Richardson et al [7] proposed a comprehensive digital health equity framework that identifies access to digital tools, including financial management platforms, as a determinant operating across individual, interpersonal, community, and societal levels. Empirical evidence supports this framing. Naveenan et al [8] showed that digital inclusion moderates the relationship between financial inclusion and health outcomes in developing countries, suggesting that inaccessible digital financial services may attenuate the health benefits of financial participation.

Barriers for people with VI commonly arise from a misalignment between user needs and visually dominant design practices in financial technologies, which are typically developed for visual rather than multisensory interaction [9]. VI spans a wide spectrum, from low vision to complete blindness, necessitating tailored design approaches. For instance, individuals with low vision often rely on adjustable font sizes, high-contrast displays, and customizable layouts to use digital banking apps effectively [3]. In contrast, people who are blind typically navigate digital environments using assistive technologies such as screen readers, voice commands, tactile input devices, or refreshable Braille displays [10]. However, despite the existence of assistive technologies, these tools are frequently unavailable, particularly in low- and middle-income settings, and are often difficult to use in practice [11]. Even when assistive technologies are accessible, many digital financial apps remain poorly designed, with limited compatibility that renders essential financial tasks difficult or impossible to complete independently [1-3,9,11].

Beyond interface-level barriers, engagement with digital financial apps is shaped by user-specific factors such as digital and financial literacy, familiarity with assistive technologies, and privacy concerns [11]. Security mechanisms that rely primarily on visual input, including CAPTCHA, biometric prompts, and multifactor authentication, often lack accessible alternatives. Although intended to enhance security, such measures can inadvertently exclude people with VI and reinforce digital health inequities [3]. In this context, access to financial services represents not only a technical issue but also a matter of human rights and social inclusion [1]. Financial inclusion supports at least 5 United Nations Sustainable Development Goals, including poverty reduction, gender equality, and reduced inequalities [12]. Inaccessible digital platforms risk exacerbating existing disparities and undermining broader development objectives [11].

Inclusive, user-centered design and participatory co-design with people with VI are widely recognized as essential for addressing accessibility barriers in digital platforms [13,14]. Kim [14] demonstrated that designers' mental models often differ fundamentally from those of users with visual disabilities, a disconnect that may help explain persistent accessibility failures in financial apps. Research-informed personas and co-design approaches have been used to guide the development of more inclusive systems, emphasizing simplified navigation, nonvisual authentication options, and consistent assistive technology compatibility [15,16]. To the best of our knowledge, no review has focused specifically on the accessibility of digital financial services for people with VI. Although Puli et al [12] examined financial inclusion for disability more broadly, their review did not address VI-specific barriers.

Several reviews have examined the broader intersection of information and communication technology (ICT) and VI [17-21]. Ashraf et al [17], for example, conducted a systematic review of ICT apps designed to support people with VI across multiple domains, reporting positive effects on self-esteem but without quantifying broader life impacts or addressing financial inclusion specifically. Similarly, Al-Razgan et al [20] reviewed usability barriers in mobile apps for people with VI across themes such as assistive devices, screen layout, and audio guidance, but did not examine financial technologies in detail. Hamideh Kerdar et al [21] identified systemic accessibility barriers in digital apps for people with VI, without focusing on financial services. Jonsson et al [22] found limited consideration of sensory impairments in digital health service design, while Fayyad and Al-Sinnawi [23] reported that banks in Palestine remained reluctant to implement accessible services for people with VI despite policy commitments.

Collectively, this literature indicates that although digital accessibility and financial inclusion have been studied extensively in isolation from each other, their intersection remains underexplored. Existing research has largely examined

XSL•FO
RenderX

accessibility challenges in general digital environments or financial inclusion at a population level without identifying the specific mechanisms through which digital financial systems exclude people with VI. As a result, there is limited evidence on the barriers and facilitators shaping access to digital financial tools, particularly across different platforms, user contexts, and stages of the transaction process. This gap is significant because digital financial tools have the potential to enhance autonomy and financial participation, yet inaccessible systems may instead reinforce dependence, limit privacy, and exacerbate existing social and health inequities [12].

To address this gap, this scoping review aims to (1) map the characteristics, settings, and populations of existing research; (2) synthesize reported barriers and facilitators affecting accessibility of digital financial services; (3) examine methodological approaches in this field; and (4) identify evidence gaps to inform future research and policy. The findings are intended to support more inclusive digital financial and health-enabling technologies.

Methods

Study Design

This scoping review follows the methodological framework outlined by the Joanna Briggs Institute for scoping reviews and is reported in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist (Checklist 1)[24]. The PRISMA-ScR checklist was verified against the paper on revision, and all items are addressed within the Methods and Results sections.

Protocol and Registration

A review protocol was developed a priori by the review team to define the review questions, eligibility criteria, search strategy, and data charting framework. No substantive deviations from the protocol occurred following commencement of the review. The review protocol was not registered prospectively.

Research Questions

This scoping review was designed to systematically map the literature on the accessibility of digital financial services for people with VI. The review addressed the following questions:

1. What types of evidence exist regarding the accessibility of digital financial services for people with VI?
2. What are the primary barriers and facilitators identified in the literature?
3. What are the characteristics of research conducted in this field (eg, study designs, settings, and populations)?
4. What gaps exist in current evidence to guide future research and policy?

Eligibility Criteria

Eligible studies included (1) studies reporting end user data from people with VI and (2) technical or design-focused evaluations directly relevant to the accessibility or usability of digital financial apps (eg, assistive technology compatibility, text scaling, or authentication workflows). Eligible study designs included qualitative studies, cross-sectional studies, experimental or quasi-experimental studies, and technical or prototype development studies relevant to the accessibility of digital financial apps. We included technical evaluations because accessibility barriers in financial apps can arise from implementation choices (eg, text scaling, labeling, and authentication workflows) even when end user testing is absent.

The search encompassed publications from 1995 through February 2026. This extended timeframe was chosen to capture early research on accessibility in digital finance and related sociotechnical systems, such as web-based banking and authentication, that preceded contemporary mobile banking. In practice, included studies clustered from 2011 onward, corresponding with the rise of modern mobile financial apps.

Studies were excluded if they did not focus on VI or digital financial services, lacked an accessibility or usability component, or were not peer-reviewed. Non-English publications were excluded because of the language limitations of the review team. Publications without sufficient methodological detail to allow extraction of study characteristics or accessibility findings (eg, abstracts without full text) were also excluded.

Information Sources

Searches were initially conducted in June 2024 and were rerun on February 27, 2026, to capture newly published studies before final screening. The following databases were searched: PubMed, MEDLINE, CINAHL (EBSCOhost), Scopus, Web of Science, Business Source Complete (EBSCOhost), ProQuest, and IEEE Xplore. No additional information sources (eg, gray literature, hand-searching, or contact with authors) were used.

Search Strategy

The search strategy combined controlled vocabulary (eg, MeSH [Medical Subject Headings] in MEDLINE and CINAHL headings, where appropriate) and free-text terms across 3 conceptual blocks, combined using Boolean operators. The three blocks were (1) VI or blindness; (2) accessibility, usability, assistive technology, and common barriers (including authentication-specific terms such as CAPTCHA, one-time password [OTP], and biometric); and (3) digital financial services and applications (including mobile banking, fintech, digital wallets, and mobile payments).

Within each conceptual block, synonyms and related terms were combined using OR, and the 3 blocks were combined using AND. Truncation and phrase searching were applied where supported by individual databases.

The original strategy used 4 conceptual blocks; this was revised to 3 in line with the Cochrane Handbook guidance (Section 4.4.2) to improve sensitivity by consolidating related concepts and reducing overly restrictive AND combinations. The revised search was rerun across all databases on February 27, 2026.

The complete search strategies for all databases, including search terms and Boolean operators, are provided in Multimedia Appendix 1. No geographic limits were applied. We limited our search to English-language peer-reviewed publications because the review team lacked shared fluency in additional

XSL•FO
RenderX

languages, and the aim was to map published research with reproducible methods and accessible full texts within the review timeframe.

Search Reporting

Search methods were additionally reported in accordance with the PRISMA-S [25] extension for literature search reporting. Full search strategies for all databases, including exact search strings, Boolean operators, limits, and database-specific adaptations, are provided in Multimedia Appendix 1.

Searches were limited to English-language, peer-reviewed publications published between 1995 and 2026. Records were imported into Covidence (Veritas Health Innovation Ltd), and duplicates were removed before screening.

No additional search methods (eg, citation searching, hand-searching of journals, or gray literature searches) were undertaken.

Selection of Sources of Evidence

Search results were imported into Covidence for deduplication and screening. To ensure consistent application of eligibility criteria, a random 20% sample of titles and abstracts was independently screened by 2 reviewers. Discrepancies were discussed, and the screening approach was calibrated before screening the remaining records, which were assessed by a single reviewer.

Full-text papers deemed potentially eligible were retrieved and evaluated against the inclusion criteria. As in the title and abstract screening stage, 20% of full-text papers were independently assessed by 2 reviewers to verify consistency, with the remainder reviewed by a single reviewer. Disagreements during dual screening were resolved through discussion, with consultation from a third reviewer when consensus could not be reached.

Full search strategies for each database, including exact search strings, limits, and dates searched, are provided in Multimedia Appendix 1.

Data Charting Process

A standardized data charting form was developed through team consultation and piloted on 2 included studies before full data extraction. To ensure consistency, 20% of studies were independently charted by 2 reviewers, with the remaining studies charted by a single reviewer using the finalized framework. Any discrepancies identified during dual charting were resolved through discussion and consensus.

Data Items

Data items were defined a priori and extracted using a standardized charting form. Extracted variables included bibliographic details (author, year, and country); study characteristics (study design); population details (type and severity of VI and demographics); type of financial application studied; accessibility features examined; reported user needs, barriers, and outcomes; and evaluation focus (end user study, technical evaluation, or mixed).

These variables were selected to support descriptive mapping of accessibility barriers and facilitators across platforms and study contexts.

Critical Appraisal of Individual Sources of Evidence

Consistent with scoping review methodology, formal critical appraisal of individual sources of evidence was not undertaken [24]. The aim of this review was to map the extent, characteristics, and nature of the available evidence rather than assess the methodological quality or risk of bias of the included studies.

Data Synthesis

Charted data were synthesized descriptively by one author (LP) and discussed with the full research team to inform interpretation. Study characteristics were summarized using counts and frequencies. Accessibility barriers and facilitators were grouped using inductive thematic categorization based on reported findings, without formal qualitative coding. Evidence mapping was used to visualize the distribution of barrier types across platform contexts.

Ethical Considerations

Ethics approval was not required for this study because it involved the analysis and synthesis of data from publicly available published literature and did not involve the collection of primary data from human participants.

Results

Selection of Sources of Evidence

The final search, conducted on February 27, 2026, identified 1719 records, including 1134 from Scopus, 484 from PubMed, 75 from ProQuest, 10 from CINAHL, 9 from Business Source Complete, 6 from Web of Science, and 1 from IEEE Xplore. There were 218 duplicates, leaving 1501 records for title and abstract screening.

During title and abstract screening, 1436 records were excluded, and 65 full-text papers were sought and assessed for eligibility; all were retrieved. Of these, 42 full-text papers were excluded. A total of 23 studies were included in the review. There were no ongoing studies and no studies awaiting classification. The study selection process is summarized in the PRISMA-ScR flow diagram (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) flow diagram summarizing the scoping review search results.

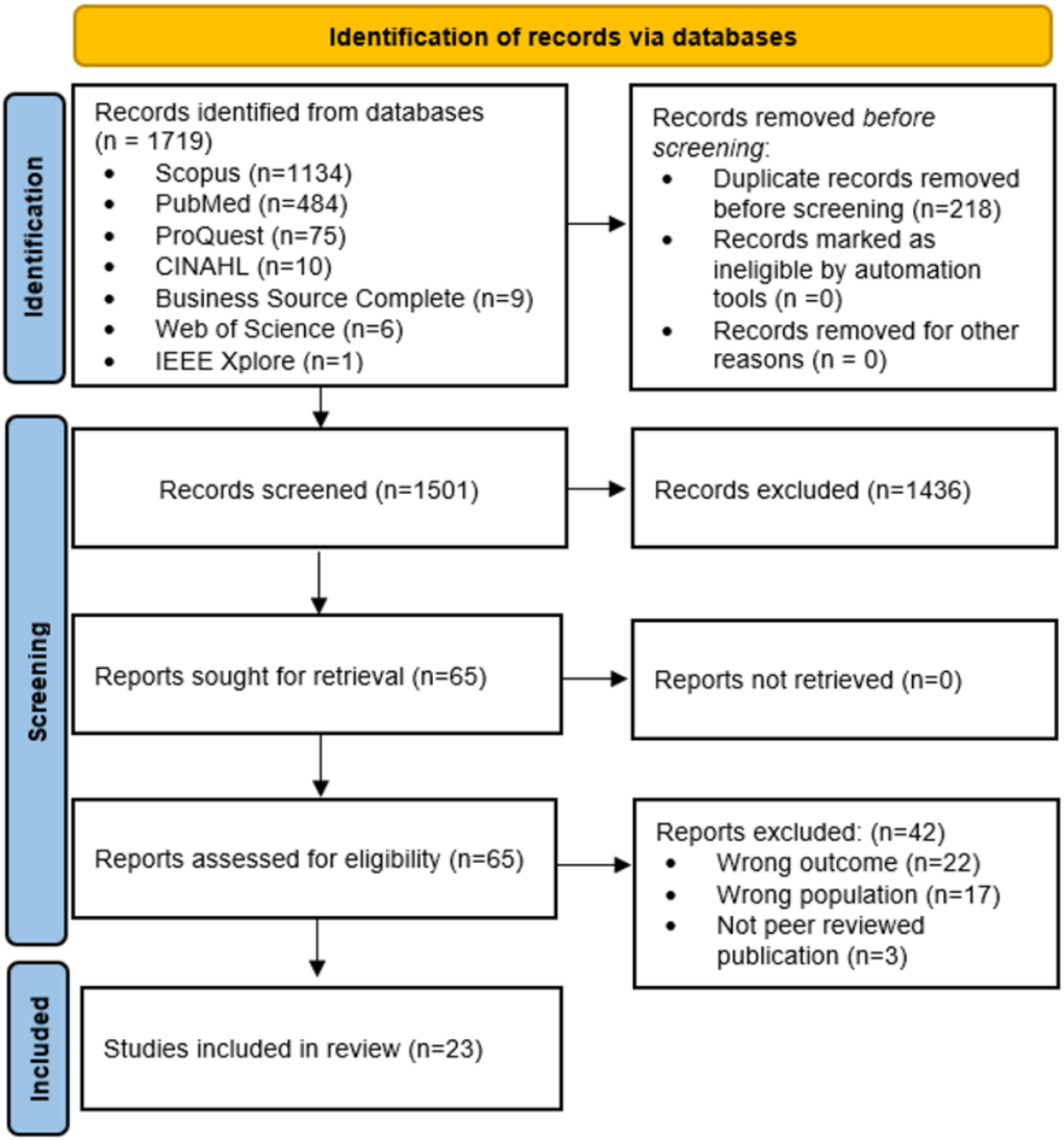

Characteristics of Sources of Evidence

Table 1 provides the characteristics of the included studies [9,10,13,15,26-44]. The 23 included studies were published between 2011 and 2026. Figure 2A provides the geographic distribution of the included studies. As shown in Figure 2B, the largest number of studies was conducted in India (n=7). Studies were also conducted in the United States (n=2), Thailand (n=2), Indonesia (n=2), Japan (n=1), Belgium (n=1), Nigeria (n=1), the Netherlands (n=1), Saudi Arabia (n=1), the United Kingdom (n=1), Singapore (n=1), Canada (n=1), and Hungary (n=1), alongside 1 study described as global in scope.

Study designs included qualitative studies (n=6), mixed methods study (n=1), cross-sectional studies (n=4), nonrandomized experimental studies (n=2), design and prototype evaluations (n=4), and technical evaluations without human participants (n=6).

XSL•FO
RenderX

Table . Characteristics of the 23 included studies on the accessibility of digital financial apps for people with visual impairment, published from 2011 to 2026. Studies were identified through a systematic search of 8 databases. For each study, the table provides the lead author and year, country, and setting; study design; participant number; characteristics of participants (or data sources for technical studies), type of visual impairments addressed, description of the financial application or system evaluated, and assistive technology used or assessed.

Study and year	Country	Study design	Participant number	Visual impairment (as described by the study)	Financial application description	Assistive technology use
Afandi et al (2026) [26]	Indonesia	Qualitative	4	Blind and partially sighted	Inability to use mobile banking apps due to a lack of screen reader compatibility	Limited screen reader use
Alayed (2025) [27]	Saudi Arabia	Mixed methods	12	Blind	SNB[a], AlRajhi, and Riyad Bank mobile banking apps were evaluated for accessibility	Participants used VoiceOver (iOS) or TalkBack (Android) screen readers
Alnfiai and Sampalli (2019) [10]	Canada	Nonrandomized experimental	10	Blind or visually impaired	BraillePassword: an accessible web authentication technique on touchscreen devices (mobile, tablet, or computer)	Participants used screen readers and headphones to prevent aural eavesdropping
Alshayban and Malek (2022) [15]	Global	Empirical study	No end user participants	Low vision	Automated software tool AccessiText, developed by Alshayban and Malek [15], for detecting text accessibility issues in Android apps arising from incompatibility with the Text Scaling Assistive Service	Android text scaling accessibility (data analysis), iOS Dynamic Type referenced where applicable, AccessiText tool developed for Android only
Angsupanich and Matayong (2024) [28]	Thailand	Prototype development	5	Visual impairments	A mobile app for currency recognition	Google text-to-speech
Braeken (2017) [29]	Belgium	Cryptographic system design and formal security evaluation	No end user participants	Visual impairments	Smartphone used as a payment proxy device	Not reported
Branham and Kane (2015) [30]	United States	Qualitative	40 (20 sighted and 20 blind participants)	Blind and sighted pairs	Collaborative accessibility using a sighted peer support	All participants with visual impairment used screen readers
Christy and Pillai (2021) [31]	India	Cross-sectional	15	Blind and low vision	The study included 57 apps across various categories, such as online shopping and transport	Assistive gestures used in iOS and Android; screen readers (TalkBack and KNFB Reader), voice assistants (Siri and Google Assistant), and magnifiers

XSL•FO
RenderX

Study and year	Country	Study design	Participant number	Visual impairment (as described by the study)	Financial application description	Assistive technology use
Fenwick et al (2022) [32]	Singapore	Cross-sectional	27	Age-related macular degeneration, including polypoidal choroidal vasculopathy.	Identifying impact of age-related macular degeneration on age-related quality of life, including financial well-being, using computerized adaptive testing	Not reported
Gopalakrishnan et al (2022) [33]	India	Pilot study with focus groups	15	Peripheral and central field loss	A virtual reality bank to assess the functional performance of people with low vision compared with those with normal vision	Virtual reality headset to access virtual bank
Kameswaran et al (2018) [34]	India	Qualitative	32	Blind and low vision	Digital ride-hailing services	All participants used smartphones; 25 used Android phones with TalkBack screen readers, and 7 iPhones with VoiceOver and Google Maps
Kameswaran and Hulikal Muralidhar (2019) [35]	India	Qualitative	30	Visual impairments	Digital payment interface while accessing payment on rideshare apps in India	Twenty-five participants used screen readers built into their Android phone (TalkBack or VoiceOver)
Kameswaran et al (2023) [36]	India	Qualitative	30	Blind and low vision	Internet banking, mobile banking, and third-party banking. Among participants, 25 used internet banking and mobile banking, 27 used digital payments, and 3 did not use digital banking because they were uncomfortable with technology	All participants used screen readers for computers and mobile phones. Among participants, 20 used Android mobile phones (with TalkBack), 7 used iPhones (with VoiceOver), and 3 used both an iPhone and an Android phone
Maesaka (2015) [37]	Japan	Design case study	5	Color vision weakness (color vision deficiency), age-related vision deterioration, glaucoma, cataract, and general accessibility considerations.	Smartphone-oriented internet banking system designed by NEC[b] compliant with JIS X 8341-3 classes A and AA	Voice reading software

XSL•FO
RenderX

Study and year	Country	Study design	Participant number	Visual impairment (as described by the study)	Financial application description	Assistive technology use
Okonji et al (2020) [38]	Nigeria	Cross-sectional	98	Visually impaired with best corrected visual acuity lower than 6/36 (20/120), indicating moderate-to-severe vision loss.	Online banking, shopping, and bill payment	Screen readers such as Dolphin-Supernova (Dolphin Computer Access Ltd), ZoomText (Freedom Scientific), and JAWS (Freedom Scientific)
Radványi et al (2011) [39]	Hungary	Nonrandomized experimental	5	Visually impaired	A mobile app for banknote recognition	Bionic Eyeglass prototype
Ramani et al (2023) [40]	India	Descriptive design and prototype development	No end user participants	Blind	Mobile app prototype developed to recognize Indian currency in real time and provide audio output of the detected note	Smartphone or tablet with camera
Reyes-Cruz et al (2022) [13]	United Kingdom	Ethnomethodological study	10	Blind and partially sighted	Analyzed video-recorded demonstrations involving participants with visual impairments, HCI[c] to identify accessibility issues	VoiceOver (Apple), TalkBack (Google), JAWS (Freedom Scientific), Siri (Apple), Seeing AI (Microsoft), KNFB Reader (Sensotec NV), text detection, facial recognition, and light detection
Singh et al (2024) [9]	India	Qualitative	15	Visually disabled	Application to support digital payments on Google Pay, PhonePe, and Paytm (One97 Communications)	All participants used smartphones with screen reader assistive tools
Sofiyati et al (2024) [41]	Indonesia	Prototype development	No end user participants	Low vision	CurrencyNet is a Rupiah banknote recognition application	LoVi[d] application provides voice feedback and uses a smartphone camera; designed for use on common smartphones

XSL•FO
RenderX

Study and year	Country	Study design	Participant number	Visual impairment (as described by the study)	Financial application description	Assistive technology use
van der Cruijsen and Reijerink (2024) [42]	Netherlands	Cross-sectional	19,136 respondents	Blind or visually impaired (2.1% of respondents), deaf or hearing impaired, limited or no hand function, mild intellectual disability, and mobility impairments (walking difficulty or wheelchair).	Description of payment environment: high digitalization (contactless cards and smartphones); 20% of POS[e] transactions paid with cash (2022); near-universal card and internet access; high acceptance of cash (96%) and debit cards (92%). Dependent variables (cash importance measures): share of POS transactions paid with cash, uses cash only, cash-dependent ("cannot do without cash"), and cash preference	Not reported
Yang et al (2014) [43]	United States	Prototype development	No end user participants	General visually impaired	A Money Reader Android app for recognizing US dollar banknotes, custom-built by Yang et al [43]	Android accessibility features (TalkBack and Explore by Touch) and a text-to-speech engine for audio output
Yawai et al (2024) [44]	Thailand	Prototype development	No end user participants	General visually impaired	A Thai banknote and coin recognition application	Real-time voice output and audio feedback

[a]SNB: Saudi National Bank.

[b]NEC: Nippon Electric Company.

[c]HCI: human-computer interaction.

[d]LoVi: low vision.

[e]POS: points of sale.

XSL•FO
RenderX

Figure 2. Geographic distribution of the 23 included studies. (A) World map showing the countries in which the included studies were conducted. Colors indicate study count per country: red (n=7), blue (n=2), and green (n=1). Custom map image created with Mapchart [45]. (B) Corresponding bar chart of study counts by country. India accounted for nearly one-third of all included studies (n=7), followed by Indonesia, Thailand, and the United States (n=2 each). The remaining 9 countries contributed 1 study each. One additional study (Alshayban and Malek [15]) was global in scope and is not mapped to a specific country.

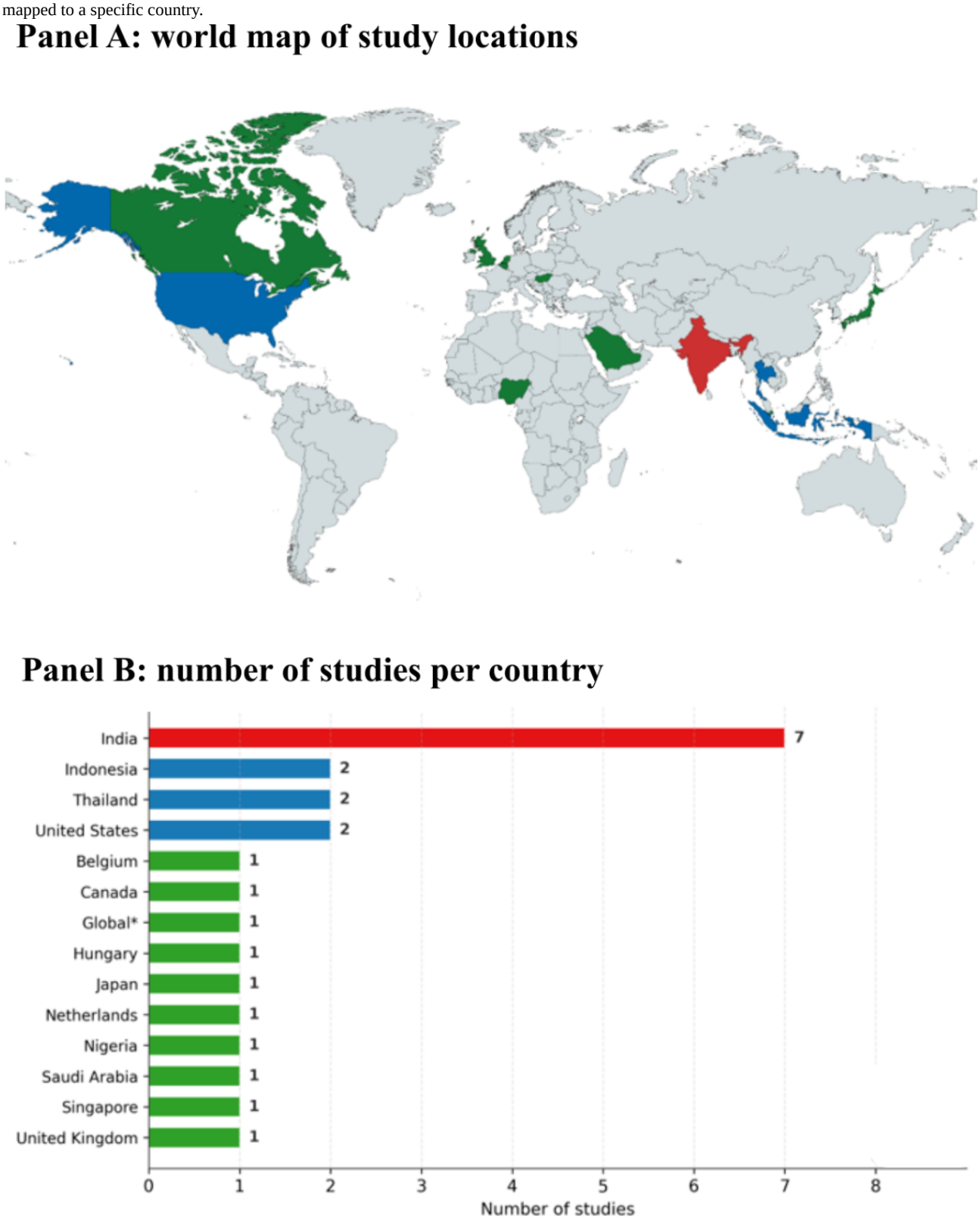

Participant Characteristics and Data Sources

Of the 23 studies, 17 included human participants, and 6 were technical evaluations without human participants. Among the 17 studies with human participants, 14 focused directly on people with VI and 3 enrolled broader samples, such as blind and sighted pairs, respondents with VI as a subgroup, or related clinical populations. As shown in Figure 3, publication activity in this field was sparse until 2022, after which a marked increase

XSL•FO
RenderX

occurred, with more than 70% of the included studies published from 2022 onward.

Figure 3. Distribution of the included studies (n=23) by publication year and study type. Studies are grouped by type: end user studies involving human participants (n=17; solid blue bars) and technical or design-focused evaluations without human participants (n=6; hatched orange bars). The cumulative total is shown as a line (pink). Publication activity was sparse from 2011 to 2020, with no more than 2 studies per year. A marked increase occurred from 2022 onward, with more than 73.9% (17/23) of the included studies published between 2022 and 2026.

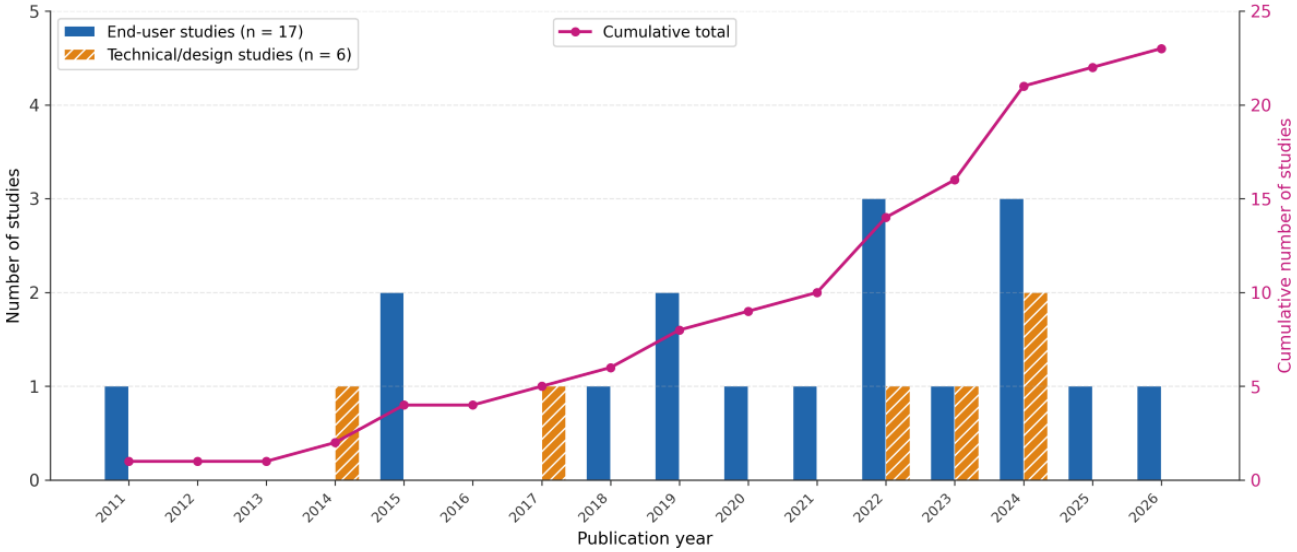

Excluding 1 large population survey (van der Cruijsen and Reijerink [42], n=19,136 and 2.1% with VI), the remaining studies with human participants had sample sizes ranging from 4 to 98. The population survey is reported separately from the smaller end user studies to avoid distorting the participant summary. Low-vision participants were included in 6 of 17 studies, 3 enrolled participants who were blind and participants with low vision, and 4 used nonspecific descriptors (eg, “visually impaired”) without severity detail.

Specific diagnoses were inconsistently reported. Where described, conditions included age-related macular degeneration, glaucoma, cataracts, retinitis pigmentosa, peripheral field loss, and central field loss.

Synthesis of Results

Overview

Of the 23 included studies [9,10,13,15,26-44], 6 studies used qualitative research designs, including semistructured interviews, focus groups, and related observational approaches. One study (Afandi et al [26]) was classified as qualitative based on its interview-based methodology, although the original extraction described it as a case-control study. One study used a mixed methods design combining qualitative approaches with observational analyses. Four studies used cross-sectional designs, including a population survey on digital payment behavior and surveys of technology use among people with VI. Two studies used nonrandomized experimental designs to evaluate accessible authentication methods and tactile currency recognition tools. Four studies used design or prototype evaluation approaches involving human participants, including prototype development with end user testing, virtual reality pilot testing, a design case study, and an ethnomethodological analysis of assistive technology use. The remaining 6 studies were technical evaluations without human participants, including prototype development for currency recognition and accessible payments (n=4), automated accessibility evaluation (n=1), and cryptographic system design (n=1).

Assistive Technology Use

Assistive technology use was reported in 14 of the 17 end user studies. Screen readers were the most frequently described assistive technologies, with TalkBack (Android; Google) explicitly reported in 6 studies and VoiceOver (iOS; Apple) reported in 5 studies; 4 studies reported the use of both platforms. In addition to screen readers, several studies described the use of other assistive technologies, including optical character recognition readers, seeing artificial intelligence, KNFB Reader, refreshable Braille displays, screen magnification tools, and voice assistants.

The 6 technical evaluations without human participants assessed a variety of tools, such as authentication techniques, automated accessibility detection tools, currency recognition algorithms, prototype apps, and payment system architectures. These studies often evaluated how these technologies interacted with assistive technology. Furthermore, these studies predominantly examined system design or functionality relevant to accessibility but did not always involve direct participation of users with VI.

Accessibility and Usability Barriers

Barriers related to screen reader compatibility and interface labeling were reported in 17 of 23 studies [9,10,13,15,26-44]. These included missing or incorrect screen reader labels, unlabeled buttons or form fields, and unclear prompts. Issues related to navigation flow and focus order were described in 15 studies, including difficulties moving sequentially through screens or completing multistep workflows. Complex interface layouts that limited efficient nonvisual navigation were reported in 12 studies.

Authentication-related barriers were reported in 18 studies. Visual CAPTCHA challenges were described in 14 studies, and

XSL•FO
RenderX

OTP workflows requiring app switching were reported in 13 studies. Biometric authentication mechanisms that were not consistently accessible were described in 9 studies. Additional authentication-related issues included session timeouts in 6 studies, personal identification number entry difficulties in 7 studies, and inconsistent or inaccessible error messaging in 8 studies.

Other usability barriers described across studies included missing alternative text for images in 10 studies, layout distortions when text scaling was applied in 7 studies, auto-scrolling content or advertisements in 5 studies, limited audio or haptic feedback in 6 studies, and technical failures such as application crashes or freezing in 4 studies.

Accessibility Features and Facilitators

Accessibility features were reported in 19 [9,10,13,15,26-28,30,31,34-38,41-45] of 23 included studies [9,10,13,15,26-44]. Table 2 organizes the reported barriers, enablers, and evidence gaps across 5 stages of the user journey (ie, authentication, navigation, transaction, confirmation, and ongoing use), showing where the evidence is concentrated and where gaps remain. Among the most frequently reported features, screen reader compatibility was described in 15 studies, while provision of alternative text for images or interface elements was reported in 9 studies. Keyboard or gesture-based navigation options were described in 7 studies, and voice-based interaction features were reported in 6 studies.

Haptic feedback for confirmation of transactions or authentication steps was described in 5 studies. Alternative authentication mechanisms designed to reduce reliance on visual input (eg, tactile or nonvisual authentication approaches) were evaluated in 4 studies.

Table . Accessibility barriers, enablers, and evidence gap matrix for digital financial services across the user journey. The matrix maps accessibility barriers, enablers, and evidence gaps identified across the 23 included studies along 5 stages of the digital financial services user journey: authentication, navigation, transaction, confirmation, and ongoing use. For each stage, the specific barriers reported in the literature, the enablers or design strategies proposed or evaluated, and the gaps where empirical evidence remains limited or absent are presented. The framework illustrates how barriers compound across the user journey rather than occurring in isolation, and highlights the stages where future research is most needed.

Factor	Authentication	Navigation	Transaction	Confirmation	Ongoing use
Barriers	CAPTCHA[a], OTP[b] requiring app switching, and inaccessible biometric authentication.	Poor focus order, complex layouts, multistep workflows, and screen reader incompatibility.	Unlabeled fields, inaccessible forms, and missing alternative text.	Lack of audio or haptic feedback and unclear error messages.	Updates are breaking accessibility, and app crashes or freezing.
Enablers	Gesture-based authentication and alternative CAPTCHA.	Logical navigation structure and screen reader optimization.	Proper labeling and accessible form design.	Audio feedback, haptic confirmation, and clear error messaging.	Ongoing accessibility testing and compatibility monitoring.
Evidence gaps	Lack of evaluation of accessible authentication methods in real-world financial apps; limited testing of alternatives to CAPTCHA, OTP, and biometrics; and absence of security–accessibility trade-off analysis.	Limited empirical evaluation of navigation improvements (eg, focus order and layout simplification) on task completion and efficiency, and lack of standardized accessibility benchmarks across platforms.	Minimal evidence on end-to-end transaction completion success, limited testing of accessible form design in real financial workflows, and absence of error recovery and usability studies.	Lack of research on feedback mechanisms (audio and haptic) and their impact on user confidence, error detection, and transaction safety.	No longitudinal studies examining how app updates affect accessibility over time, and the absence of evidence on regression testing practices or long-term compatibility monitoring with assistive technologies.

[a]CAPTCHA: Completely Automated Public Turing test to tell Computers and Humans Apart.

[b]OTP: one-time password.

Reported Recommendations and Proposed Strategies

Recommendations to improve accessibility were reported in 18 [9,10,13,15,26-28,30,31,34-38,41,43-45] of the 23 included studies [9,10,13,15,26-44]. For example, design-level recommendations were described in 14 studies and included the adoption of universal design principles, collaborative or participatory development processes, and routine accessibility audits during application development.

Several studies recommended improvements to screen reader compatibility, clearer labeling of interface elements, improved navigation flow, and enhanced nonvisual feedback mechanisms. Alternative authentication strategies designed to reduce reliance on visual input were proposed in 5 studies, including tactile- or haptic-based authentication approaches.

User-centered testing and iterative design processes were recommended in 9 studies, particularly to support accessibility maintenance following application updates. Policy- or organizational-level recommendations were reported in 7 studies, including calls for stronger enforcement of accessibility standards, training for financial service providers, and improved regulatory oversight and enforcement for digital financial platforms. Table 3 provides an evidence map summarizing the distribution of reported accessibility barriers across platform types.

XSL•FO
RenderX

Table . Evidence map of accessibility barriers by platform type and study frequency. The matrix shows the number of included studies reporting each accessibility barrier across different platform types (iOS, Android, web, multiple platforms, and not specified). Cell values indicate study counts. Common barriers include authentication issues (Completely Automated Public Turing test to tell Computers and Humans Apart and one-time password [n=18]), assistive technology compatibility (n=17), and navigation issues (n=15). Multiple studies may contribute to multiple cells because some examine multiple platforms or barrier types. Counts therefore represent the number of studies contributing to each cell rather than mutually exclusive totals.

Barrier description	iOS	Android	Web	Multiple platforms	Not specified	Total
Authentication barriers (CAPTCHA[a] and OTP[b])	4	6	5	3	0	18
Assistive technology compatibility issues	5	6	4	2	0	17
Navigation issues	4	7	3	1	0	15
Complex interface layouts	2	5	3	2	0	12
Compatibility issues following updates	2	3	1	1	0	7
Text scaling and readability issues	3	3	1	0	0	7
Insufficient audio feedback	2	3	1	0	0	6
Social or contextual barriers	0	1	0	2	1	4
Physical interface barriers	0	0	0	3	1	4
Technical failures	2	2	0	0	0	4

[a]CAPTCHA: Completely Automated Public Turing test to tell Computers and Humans Apart.

[b]OTP: one-time password.

Discussion

Principal Findings

This scoping review mapped the published evidence on the accessibility of digital financial apps for people with VI. The evidence reveals persistent, cross-platform accessibility barriers in authentication workflows, screen reader compatibility, and navigation design, alongside facilitators such as consistent assistive technology support, structured navigation, and nonvisual feedback mechanisms that are directly relevant to digital health equity. However, the evidence base remains small, geographically concentrated, and dominated by cross-sectional and qualitative designs. The accessibility barrier-enabler-gap matrix (Table 2) illustrates how these challenges manifest across 5 stages of the user journey, from authentication through ongoing use, and reveals that evidence gaps are most pronounced at the confirmation and ongoing-use stages, where almost no empirical evaluation exists. No study evaluated the effectiveness of an accessibility intervention over time, and the technical evaluation literature has developed largely in isolation from end user experience research. These findings directly address the review objectives by mapping study characteristics, identifying barriers and facilitators, and highlighting evidence gaps.

Authentication represents the most pervasive barrier domain. Visual CAPTCHA, OTP workflows requiring app switching, session timeouts, and inaccessible biometric verification were described across the majority of included studies and across all platform types. This finding is consistent with Schmeelk and Petrie [46], who surveyed visually disabled users in the United Kingdom and the United States and found that password management, multifactor authentication, and biometric systems each present distinct accessibility barriers. Authentication barriers did not operate in isolation. In combination with screen reader incompatibility and complex navigation, these barriers produced cascading failures that prevented independent transaction completion, a prerequisite for secure and autonomous participation in digitally mediated health-relevant services. Multiple studies documented how this forced reliance on sighted intermediaries for tasks as fundamental as logging in, verifying transactions, and making payments [35,36]. Drawing on the concept of “moneywork,” Kameswaran et al [36] characterized this additional, invisible labor required for people with VI to participate in financial systems that sighted users navigate with minimal friction. Griffith et al [47] provided empirical quantification of these costs, demonstrating the time disparity experienced by blind users encountering web accessibility barriers, findings subsequently referenced by the US Department of Justice in its 2024 Americans with Disabilities Act (ADA) Title II rule on web accessibility. The privacy implications are substantial: sharing account details or transaction amounts with intermediaries undermines the financial autonomy digital banking is intended to support.

XSL•FO
RenderX

As outlined in the Introduction, access to financial services via digital technology constitutes a digital determinant of health [6]. Richardson et al [7] proposed a comprehensive framework for digital health equity, identifying digital tool access as a key determinant across multiple ecological levels. The barriers identified in this review, particularly those affecting authentication and independent transaction management, have implications that extend beyond financial autonomy to downstream health-relevant outcomes. When people with VI cannot independently manage digital payments, they face reduced access to health-related financial functions, including insurance management, medication payments, and health savings. Paik et al [48] demonstrated that underrepresentation in digital health data creates feedback loops worsening health disparities, a mechanism that may also operate through financial data exclusion. Several included studies documented participants forgoing digital transactions or reverting to cash specifically because of accessibility barriers [35,42], suggesting that digital financial inaccessibility may function as an upstream determinant of both financial and health inequity. Ha et al [49], in JMIR, demonstrated the feasibility of user-centered design for disability-inclusive digital health platforms, establishing a model that could be extended to financial technology. Integrating financial application accessibility into digital health equity frameworks would acknowledge the role of financial autonomy in health outcomes and align with the World Health Organization (WHO) global strategy on digital health 2020-2025.

Facilitators ranged from platform-level assistive technology support to design-level innovations. Among the most concrete contributions, Alnfiai and Sampalli [10] proposed BraillePassword, a gesture-based authentication method leveraging Braille literacy that directly addresses password-entry barriers. A related innovation, TapCAPTCHA [50], demonstrated that gesture-based CAPTCHA alternatives using audio clips achieved substantially better success rates for users with VIs compared to conventional audio CAPTCHA, offering a concrete solution to the visual CAPTCHA barrier identified repeatedly across included studies. Several studies demonstrated viable offline currency recognition apps integrating text-to-speech output, addressing persistent cash-handling barriers in contexts where tactile denomination features are unreliable. At the system level, Alshayban and Malek [15] demonstrated automated accessibility evaluation at scale, suggesting a pathway toward continuous monitoring rather than reliance on postrelease user complaints. These innovations sit within a growing ecosystem of mobile apps functioning as assistive technologies for people with VI [51], though translation into mainstream financial services remains limited. User-generated workarounds, such as memorizing button locations, preorganizing currency notes by denomination, and switching between multiple apps, also emerged as an important theme, offering practical insight into unmet design needs [35,36].

The barriers identified in this review are consistent with broader digital accessibility research, but their interaction with financial transaction workflows creates a distinctive pattern of risk. In general app accessibility, an unlabeled button may cause frustration; in banking apps, an inaccessible authentication screen can prevent account access entirely, creating financial and privacy vulnerability. Puli et al [12] reviewed financial inclusion for people with disabilities broadly and identified digital inaccessibility as a barrier; this review extends this by mapping the specific mechanisms through which inaccessibility operates for people with VI in financial apps. Hamideh Kerdar et al [21] reviewed digital technology accessibility for people with VI across 49 studies but did not focus on financial apps, confirming the gap this review addresses. In the European context, Borowska-Beszta et al [52] framed digital payment services as a form of assistive technology that can empower consumers with disabilities, a perspective that reframes financial app accessibility not merely as a compliance issue but as an enabler of autonomy. Wilson et al [53], in a systematic review of strategies to advance digital health equity, found that people with disabilities face compounding barriers across multiple equity dimensions, consistent with our finding that authentication, navigation, and labeling barriers compound rather than occur in isolation. Notably, the 2 research streams identified in this review, end user experience studies and technical evaluations, appear largely disconnected, echoing Ortiz-Escobar et al [11] who found limited implementation of user-centered design standards in assistive technology development.

The geographic concentration of evidence in India warrants examination. India's prominence likely reflects the convergence of a large population with VI, rapid digitalization of financial services through platforms such as Unified Payments Interface, and the postdemonetization context in which the withdrawal of high-denomination notes in 2016 made digital payment adoption nearly mandatory [9,35]. The relative absence of studies from Europe and North America is notable given established regulatory frameworks such as the European Accessibility Act (compliance deadline 2025) and the ADA. Recent evidence from other contexts, however, suggests that the barriers identified in this review are global rather than India-specific: Fayyad and Al-Sinnawi [23] documented persistent financial exclusion of people with VI in Palestine despite national policy commitments; Huang et al [54] identified insufficient legal safeguards for digital financial accessibility in China; Goundar and Sathye [55] found marketplace barriers to financial service access for people with VI in Fiji; and Muuo et al [56] reported that while persons with disabilities in Kenya embrace mobile money, financial institutions lack the knowledge to adapt their products. These findings suggest that regulatory coverage alone does not guarantee accessible implementation and that comparative studies across regulatory environments could help identify which policy mechanisms translate into measurable improvements.

The evidence gaps identified suggest several priorities. First, no included study evaluated the effectiveness of an accessibility intervention over time; longitudinal or pre-post studies assessing the impact of specific design changes on transaction success rates would provide intervention-level evidence. Second, integrating the disconnected technical and end user research streams is essential, as Ara et al [57] found that automated accessibility evaluation alone is insufficient and must be

XSL•FO
RenderX

complemented by testing with assistive technology users. Third, participatory design involving users with VIs, guided by emerging best practices [14], should be embedded in financial application development rather than limited to post hoc usability testing. Fourth, the intersectionality of VI with age, digital literacy, and gender remains underexplored; most qualitative studies recruited predominantly male participants. Fifth, low vision was included in fewer studies than blindness, despite representing the majority of the global VI population. Finally, with the emergence of artificial intelligence–powered accessibility solutions [58], future research should examine whether conversational agents, intelligent interfaces, and automated remediation tools can meaningfully reduce the barriers identified in this review.

Limitations

This scoping review has several limitations. The included studies were predominantly qualitative and conducted in diverse settings, limiting direct comparability. The geographic concentration of many studies in India suggests potential context-specific findings, potentially limiting global generalizability. Additionally, sample sizes in the included studies were often small, and limited quantitative data were available, restricting robust evaluation of the effectiveness of accessibility features. Finally, publication bias toward positive outcomes and the exclusion of gray literature may have influenced the completeness of the evidence. The initial search strategy used a more restrictive 4-block structure. Following peer review, this was revised to 3 blocks (consistent with the Cochrane Handbook Section 4.4.2) and rerun across all databases on February 27, 2026, yielding 23 studies compared to 12 in the original analysis. While the revision strengthened the search, relevant studies indexed under nonstandard terminology or in databases not included in the search may not have been captured. The predominance of qualitative and cross-sectional designs among the included studies also limits the ability to draw causal inferences about the effectiveness of specific accessibility features or interventions.

Conclusions

This scoping review provides the first focused synthesis of accessibility evidence at the intersection of digital financial services, VI, and digital health equity. Unlike prior reviews examining either ICT accessibility for users with VIs broadly [21] or financial inclusion for disability generally [12], this review identifies the specific mechanisms through which inaccessible financial apps undermine independence, privacy, and financial autonomy. By mapping 23 studies across 12 countries, the review identifies authentication, interface labeling, and navigation as persistent cross-platform barriers; reveals a disconnect between technical design literature and end user experience research; and highlights the absence of longitudinal or intervention-based evidence. The findings carry direct implications for multiple stakeholders. Financial technology developers should prioritize accessible authentication workflows, and evidence that alternatives such as gesture-based CAPTCHAs outperform standard audio CAPTCHAs [50] demonstrates that viable solutions exist. Accessibility auditors should evaluate the full transaction lifecycle, not individual screens. Policymakers implementing the European Accessibility Act, ADA, and comparable frameworks should consider mandatory accessibility certification for financial apps, noting that emerging evidence from China [54], Palestine [23], and Kenya [56] demonstrates that legislative intent does not guarantee accessible implementation. More broadly, integrating financial application accessibility into digital health equity frameworks [6,7] would acknowledge the role of financial autonomy in health outcomes and position this work within the broader digital determinants of health agenda.

Acknowledgments

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following task was delegated to GenAI tools under full human supervision: proofreading and editing. The GenAI tool used was ChatGPT-5 (OpenAI). Responsibility for the final paper lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the outcomes. This declaration was submitted by Louise Puli.

Funding

The funder had no involvement in the study design, data collection, analysis, interpretation, or manuscript writing.

Authors' Contributions

Conceptualization – LP, TK, AZS; Data curation – LP, LK, KR; Formal analysis - LP, LK, KR, TK; Funding acquisition – KR, AZMS; Investigation – LP, LK, TK, AZMS, KR; Methodology – LP, LK, TK, CH, KR; Project administration –AZMS, KR; Supervision – AZMS, KR; Visualization – LP, LK, KR; Writing – original draft – LP; Writing – review & editing – LK, TK, CH, AZMS, KR

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy including full search strings for each database.

XSL•FO
RenderX

[DOCX File, 22 KB - jmir_v28i1e82315_app1.docx]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 55 KB - jmir_v28i1e82315_app2.docx]

References

1. World report on vision. : World Health Organization; 2019 URL: https://www.who.int/publications/i/item/world-report-on-vision [accessed 2026-04-13]
2. Brown RL, Barrett AE. Visual impairment and quality of life among older adults: an examination of explanations for the relationship. J Gerontol B Psychol Sci Soc Sci 2011 May;66(3):364-373. [doi: 10.1093/geronb/gbr015] [Medline: 21402645]
3. Global report on assistive technology. : World Health Organization; 2022 URL: https://www.who.int/publications/b/64373 [accessed 2026-06-05]
4. Wewege L, Lee J, Thomsett MC. Disruptions and digital banking trends. J Appl Finance Bank 2020;10(6):15-56 [FREE Full text]
5. World report on social determinants of health equity. : World Health Organization; 2025 URL: https://www.who.int/publications/i/item/9789240107588 [accessed 2026-04-13]
6. Chidambaram S, Jain B, Jain U, et al. An introduction to digital determinants of health. PLOS Digit Health 2024 Jan;3(1):e0000346. [doi: 10.1371/journal.pdig.0000346] [Medline: 38175828]
7. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119. [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
8. Naveenan RV, Liew CY, Kijkasiwat P. Nexus between financial inclusion, digital inclusion and health outcomes: evidence from developing economies. Soc Indic Res 2024 Aug;174(1):367-408. [doi: 10.1007/s11205-024-03391-y]
9. Singh S, Jatana N, Sehgal S, Anand R, Arunkumar B, Ramesh JVN. Making digital payments accessible beyond sight: a usability study of UPI-based smartphone applications. IEEE Access 2024;12:6830-6841. [doi: 10.1109/ACCESS.2023.3348840]
10. Alnfiai M, Sampalli S. BraillePassword: accessible web authentication technique on touchscreen devices. J Ambient Intell Human Comput 2019 Jun;10(6):2375-2391. [doi: 10.1007/s12652-018-0860-x]
11. Ortiz-Escobar LM, Chavarria MA, Schönenberger K, et al. Assessing the implementation of user-centred design standards on assistive technology for persons with visual impairments: a systematic review. Front Rehabil Sci 2023;4:1238158. [doi: 10.3389/fresc.2023.1238158] [Medline: 37744430]
12. Puli L, Layton N, Bell D, Shahriar AZ. Financial inclusion for people with disability: a scoping review. Glob Health Action 2024 Dec 31;17(1):2342634. [doi: 10.1080/16549716.2024.2342634] [Medline: 38726584]
13. Reyes-Cruz G, Fischer JE, Reeves S. Demonstrating interaction: the case of assistive technology. ACM Trans Comput-Hum Interact 2022 Oct 31;29(5):1-37. [doi: 10.1145/3514236]
14. Kim HN. Guidance for best practices in participatory design involving people with visual impairment. Ergon Des 2024 Apr;32(2):48-54. [doi: 10.1177/10648046211058248]
15. Alshayban A, Malek S. AccessiText: automated detection of text accessibility issues in Android apps. Presented at: ESEC/FSE 2022: Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering; Nov 14-18, 2022; Singapore p. 984-995. [doi: 10.1145/3540250.3549118]
16. Kanij T, Du X, Grundy J, Madugalla A, Karolita D. An approach to generating diverse personas for children and the elderly for software development. Presented at: 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC); Jun 26-30, 2023. [doi: 10.1109/COMPSAC57700.2023.00120]
17. Ashraf MM, Hasan N, Lewis L, Hasan MR, Ray P. A systematic literature review of the application of information communication technology for visually impaired people. Int J Disabil Manag 2016;11:e6. [doi: 10.1017/idm.2016.6]
18. Griffin-Shirley N, Banda DR, Ajuwon PM, et al. A survey on the use of mobile applications for people who are visually impaired. J Vis Impair Blind 2017 Jul;111(4):307-323. [doi: 10.1177/0145482X1711100402]
19. Senjam SS, Manna S, Bascaran C. Smartphones-based assistive technology: accessibility features and apps for people with visual impairment, and its usage, challenges, and usability testing. Clin Optom (Auckl) 2021;13:311-322. [doi: 10.2147/OPTO.S336361] [Medline: 34866955]
20. Al-Razgan M, Almoaiqel S, Alrajhi N, et al. A systematic literature review on the usability of mobile applications for visually impaired users. PeerJ Comput Sci 2021;7:e771. [doi: 10.7717/peerj-cs.771] [Medline: 34901428]
21. Hamideh Kerdar S, Bächler L, Kirchhoff BM. The accessibility of digital technologies for people with visual impairment and blindness: a scoping review. Discov Computing 2024 Aug 1;27(1):24. [doi: 10.1007/s10791-024-09460-7]
22. Jonsson M, Johansson S, Hussain D, Gulliksen J, Gustavsson C. Development and evaluation of eHealth services regarding accessibility: scoping literature review. J Med Internet Res 2023 Aug 17;25:e45118. [doi: 10.2196/45118] [Medline: 37590050]

XSL•FO
RenderX

23. Fayyad M, Al-Sinnawi AR. Challenges of achieving financial inclusion for individuals with visual impairments. Heliyon 2024 Aug 30;10(16):e35573. [doi: 10.1016/j.heliyon.2024.e35573] [Medline: 39220932]
24. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
25. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
26. Afandi MA, Hartarto RB, Wicaksono FS, et al. Advancing financial inclusion for persons with disabilities in Indonesia: evidence from Yogyakarta Special Region. J Popul Soc Stud 2025;34:755-775. [doi: 10.25133/JPSSv342026.038]
27. Alayed A. Are Saudi Arabian banks' mobile applications accessible for blind or partially sighted users?: A customers' perspective and evaluation. Univ Access Inf Soc 2025 Mar;24(1):585-606. [doi: 10.1007/s10209-023-01082-y]
28. Angsupanich S, Matayong S. A prototype real-time mobile application for Thai banknote recognition to support visually impaired persons. Univ Access Inf Soc 2024 Nov;23(4):1987-2002. [doi: 10.1007/s10209-023-01016-8]
29. Braeken A. An improved E-payment system and its extension to a payment system for visually impaired and blind people with user anonymity. Wireless Pers Commun 2017 Sep;96(1):563-581. [doi: 10.1007/s11277-017-4184-5]
30. Branham SM, Kane SK. Collaborative accessibility: how blind and sighted companions co-create accessible home spaces. Presented at: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems; Apr 18-23, 2015; Seoul, South Korea p. 2373-2382. [doi: 10.1145/2702123.2702511]
31. Christy B, Pillai A. User feedback on usefulness and accessibility features of mobile applications by people with visual impairment. Indian J Ophthalmol (Poona City) 2021;69(3):555-558. [doi: 10.4103/ijo.IJO_1042_20]
32. Fenwick EK, Lee EPX, Man REK, et al. Identifying the content for an item bank and computerized adaptive testing system to measure the impact of age-related macular degeneration on health-related quality of life. Qual Life Res 2022 Apr;31(4):1237-1246. [doi: 10.1007/s11136-021-02989-w] [Medline: 34562188]
33. Gopalakrishnan S, Jacob CES, Kumar M, Karunakaran V, Raman R. Development of immersive virtual reality environment for assessment of functional vision in people with low vision: a pilot study. Nep J Oph 2022 May 23;14(1):19-30. [doi: 10.3126/nepjoph.v14i1.31199]
34. Kameswaran V, Gupta J, Pal J, et al. "We can go anywhere": understanding independence through a case study of ride-hailing use by people with visual impairments in metropolitan India. Proc ACM Hum Comput Interact 2018 Nov 1;2(CSCW):1-24. [doi: 10.1145/3274354]
35. Kameswaran V, Hulikal Muralidhar S. Cash, digital payments and accessibility: a case study from metropolitan India. Proc ACM Hum Comput Interact 2019 Nov 7;3(CSCW):1-23. [doi: 10.1145/3359199]
36. Kameswaran V, Y V, Marathe M. Advocacy as access work: how people with visual impairments gain access to digital banking in India. Proc ACM Hum-Comput Interact 2023 Apr 14;7(CSCW1):1-23. [doi: 10.1145/3579596]
37. Maesaka H. Easy-to-use, smartphone-oriented internet banking, featuring Color Universal Design. NEC Technical Journal 2015 Jan 1;9:44-47 [FREE Full text]
38. Okonji PE, Okiki OC, Ogwezzy D. Perceived importance and difficulty of online activities among visually impaired persons in Nigeria. Assist Technol 2020;32(1):16-22. [doi: 10.1080/10400435.2018.1457574] [Medline: 29580182]
39. Radványi M, Solymár Z, Stubendek A, Karacs K. Mobile banknote recognition: topological models in scene understanding. Presented at: ISABEL '11: Proceedings of the 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies; Oct 26-29, 2011; Barcelona, Spain p. 1-5. [doi: 10.1145/2093698.2093883]
40. Ramani K, Suneetha I, Pushpalatha N, Reddy KB, Harish P. A mobile application for currency denomination identification for visually impaired. Presented at: International Conference on Information Technology; Dec 19-21, 2019; Bhubaneswar, India p. 105-112. [doi: 10.1007/978-981-99-5994-5_11]
41. Sofiyati M, Wiranata FA, Shalanannda W, et al. LoVi App: Android application-based image classification for low vision. J ICT Res Appl 2024 Sep 30;18(2):108-129. [doi: 10.5614/itbj.ict.res.appl.2023.18.2.3]
42. van der Cruijsen C, Reijerink J. Uncovering the digital payment divide: understanding the importance of cash for groups at risk. J Consum Aff 2024 Jun;58(2):486-505. [doi: 10.1111/joca.12591]
43. Yang Z, Schafer JM, Ganz A. Reliable and user friendly U.S banknote recognition application for visually impaired users on Android smartphones. Int J E-Health Med Commun 2014 Jul;5(3):1-16 [FREE Full text] [doi: 10.4018/ijehmc.2014070101]
44. Yawai W, Saejia M, Limsiroratana N, Thammasang R. Thai currency classification application for visually impaired persons. J Curr Sci Technol 2024 Sep 1;14(3):64. [doi: 10.59796/jcst.V14N3.2024.64]
45. Homepage. MapChart. URL: https://www.mapchart.net/ [accessed 2026-05-25]
46. Schmeelk S, Petrie H. Digital authentication for visually disabled people: initial results of an online survey. Presented at: International Conference on Computers Helping People with Special Needs; Jul 11-15, 2022; Lecco, Italy p. 41-50. [doi: 10.1007/978-3-031-08645-8_6]
47. Griffith M, Wentz B, Lazar J. Quantifying the cost of web accessibility barriers for blind users. Interact Comput 2022 Nov 1;34(6):137-149. [doi: 10.1093/iwc/iwad004]
48. Paik KE, Hicklen R, Kaggwa F, et al. Digital determinants of health: health data poverty amplifies existing health disparities-a scoping review. PLOS Digit Health 2023 Oct;2(10):e0000313. [doi: 10.1371/journal.pdig.0000313] [Medline: 37824445]

XSL•FO
RenderX

49. Ha S, Ho SH, Bae YH, et al. Digital health equity and tailored health care service for people with disability: user-centered design and usability study. J Med Internet Res 2023 Nov 28;25:e50029. [doi: 10.2196/50029] [Medline: 38015589]
50. Alnfiai M, Alassery F. TapCAPTCHA: non-visual CAPTCHA on touchscreens for visually impaired people. J Multimodal User Interfaces 2022 Dec;16(4):385-398. [doi: 10.1007/s12193-022-00394-2]
51. Pundlik S, Shivshanker P, Luo G. Impact of apps as assistive devices for visually impaired persons. Annu Rev Vis Sci 2023 Sep 15;9(1):111-130. [doi: 10.1146/annurev-vision-111022-123837] [Medline: 37127283]
52. Borowski-Beszta M, Borowska-Beszta B, Polasik M. Digital payment services in Poland as assistive technology and empowering tool for consumers with disabilities. Disabil Soc 2024 Dec 2;39(12):3051-3073. [doi: 10.1080/09687599.2023.2254468]
53. Wilson S, Tolley C, Mc Ardle R, et al. Recommendations to advance digital health equity: a systematic review of qualitative studies. NPJ Digit Med 2024 Jun 29;7(1):173. [doi: 10.1038/s41746-024-01177-7] [Medline: 38951666]
54. Huang Z, Lin C, Liu X, Feng Z. Navigating digital financial exclusion: legal imperatives for disabled people in China. Disabil Soc 2025 Oct 3;40(10):2928-2933. [doi: 10.1080/09687599.2024.2411530]
55. Goundar S, Sathye M. Exploring access to financial services by visually impaired people. J Risk Financial Manag 2023 Feb;16(2):96. [doi: 10.3390/jrfm16020096]
56. Muuo SW, Virendrakumar B, Okello G, et al. Financial inclusion for persons with disabilities: experiences of providers and users of financial products and services in Kenya. PLoS One 2025;20(4):e0321493. [doi: 10.1371/journal.pone.0321493] [Medline: 40233056]
57. Ara J, Sik-Lanyi C, Kelemen A. Accessibility engineering in web evaluation process: a systematic literature review. Univ Access Inf Soc 2024 Jun;23(2):653-686. [doi: 10.1007/s10209-023-00967-2]
58. Chemnad K, Othman A. Digital accessibility in the era of artificial intelligence-bibliometric analysis and systematic review. Front Artif Intell 2024;7:1349668. [doi: 10.3389/frai.2024.1349668] [Medline: 38435800]

Abbreviations

ADA: Americans with Disabilities Act
ICT: information and communication technology
MeSH: Medical Subject Headings
OTP: one-time password
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
VI: visual impairment
WHO: World Health Organization

Edited by S Brini; submitted 13.Aug.2025; peer-reviewed by J Owuor, W Hu; revised version received 16.Apr.2026; accepted 17.Apr.2026; published 23.Jun.2026.

Please cite as:
Puli L, Kooijman L, Kanij T, Hartel C, Shahriar AZM, Rotaru K
Accessibility of Digital Financial Applications for People With Visual Impairment: Scoping Review
J Med Internet Res 2026;28:e82315
URL: https://www.jmir.org/2026/1/e82315
doi:10.2196/82315

© Louise Puli, Lars Kooijman, Tanjila Kanij, Charmine Hartel, Abu Zafar M Shahriar, Kristian Rotaru. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Implementing Digital Respiratory Technologies for People With Respiratory Conditions: Scoping Review

Io Chi-Yan Hui[1], PhD, CEng; Kathleena Condon[2], MPH; Shailesh Kolekar[3], MBBS, MD; Nicola J Roberts[4], PhD; Katherina Bernadette Sreter[5], MD, PhD; Sami O Simons[6], MD, PhD; Carlos Figueiredo[7], MD; Zoe McKeough[8], BAppSc(Physio), PhD; Hani Salim[9], PhD; Aleksandra Gawlik-Lipinski[10], MSc; Apolline Gonsard[11], MD; Ayşe Önal Aral[12], MD; Anna Vanoverschelde[13], PhD, PharmD; Matthew Armstrong[14], PhD; Dario Kohlbrenner[15], PhD; Cátia Paixão[16], PhD; Patrick Stafler[17], MD; Efthymia Papadopoulou[18], MD; Adrian Paul Jaravata Rabe[19], MD; Milan Mohammad[20], MD; Izolde Bouloukaki[21], PhD; Shirley Quach[22], PhD, RRT; Georgios Kaltsakas[23], PhD; Kate Loveys[24], PhD; Tonje Reier-Nilsen[25], MD, PhD; Anthony Paulo Sunjaya[26], MD, PhD; Paul Robinson[2], PhD; Michaela Senek[27], PhD, MPH; Amy Hai Yan Chan[28], BPharm(Hons), PhD; Hilary Pinnock[1], MBChB, MD

[1]School of Population Health Sciences, Usher Institute, The University of Edinburgh, 5 Little France Rd, Edinburgh, United Kingdom
[10]Department of Respiratory Sciences, University of Leicester, Leicester, United Kingdom
[11]University Hospital Necker-Enfants Malades, AP-HP, Paris, France
[12]Pulmonary Diseases Clinic, Ankara Gölbaşı State Hospital, Gölbaşı/Ankara, Turkey
[13]General Hospital Zeno, Knokke-Heist, Belgium
[14]Department of Sport and Exercise Sciences, University of Durham, Durham, United Kingdom
[15]Epidemiology, Biostatistics, and Prevention Institute, University of Zurich, Zurich, Switzerland
[16]Respiratory Research and Rehabilitation Laboratory (Lab3R), School of Health Sciences (ESSUA) and Institute of Biomedicine (iBiMED), University of Aveiro, Aveiro, Portugal
[17]Schneider Children's Medical Center of Israel and Gray Faculty of Life Sciences, Tel Aviv University, Petah Tikva, Israel
[18]Department of Respiratory Failure, G. Papanikolaou General Hospital of Thessaloniki, Thessaloniki, Greece
[19]School of Public Health, Imperial College London, London, United Kingdom
[2]Faculty of Health, Medicine and Behavioural Sciences, University of Queensland, Brisbane, Queensland, Australia
[20]Centre for Physical Activity Research, Copenhagen University Hospital, Copenhagen, Hovedstaden, Denmark
[21]Department of Social Medicine, School of Medicine, University of Crete, Heraklion, Crete, Greece
[22]School of Rehabilitation Sciences, Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada
[23]Centre for Human & Applied Physiological Sciences, King's College London, London, United Kingdom
[24]Department of Paediatrics: Child and Youth Health, School of Medicine, the University of Auckland, Auckland, New Zealand
[25]Oslo Sports Trauma Research Center, Insitute of Sports Medicine, Norwegian School of Sport Sciences, Oslo, Norway
[26]School of Population Health, University of New South Wales, Sydney, NSW, Australia
[27]School of Medicine& Population Health, University of Sheffield, Sheffield, United Kingdom
[28]School of Pharmacy, The University of Auckland, Auckland, New Zealand
[3]Department of Medicine Helsingborg Hospital, Lund University, Region Skane, Sweden
[4]School of Health and Social Care, Edinburgh Napier University, Edinburgh, United Kingdom
[5]Department of Pulmonology, University Hospital Centre "Sestre Milosrdnice", Zagreb, Croatia
[6]NUTRIM Institute for Nutrition and Translational Research in Metabolism, Faculty of Health, Medicine and Life Sciences, Maastricht University Medical Center+, Maastricht, The Netherlands
[7]Hospital Santa Marta, ULS São José, Lisboa, Portugal
[8]Sydney School of Health Sciences, University of Sydney, Sydney, NSW, Australia
[9]Department of Family Medicine, Universiti Putra Malaysia, Serdang, Selangor, Malaysia

Corresponding Author:
Hilary Pinnock, MBChB, MD
School of Population Health Sciences, Usher Institute, The University of Edinburgh, 5 Little France Rd, Edinburgh, United Kingdom

Abstract

XSL•FO
RenderX

Background: Digital health offers opportunities for safe, equitable, and accessible care, and its integration into respiratory care is a strategic priority for the European Respiratory Society. However, sustainable implementation remains complex, and guidance for health care systems is limited.

Objective: This study aimed to undertake a scoping review of the published initiatives that have implemented digital respiratory technologies into real-world routine clinical practice over the past decade, identify the technologies used, implementation strategies used, the challenges and supports they encountered, and the lessons they reported for making care more equitable, strengthening patient-professional relationships, improving the patient journey, and reducing environmental impact.

Methods: Following Arksey and O'Malley's methodology, we searched ten databases (December 2013 - 2023 [updated April 2025 and February 2026]): MEDLINE, Embase, CINAHL, PsycINFO, Cochrane Library, Web of Science, Scopus, IEEE Xplore, CABI Global Health, and WHO Medicus; and used key domains in the commonly used implementation frameworks such as the Consolidated Framework for Implementation Research (CFIR), Nonadaptation, Abandonment, and Challenges to the Scale-up, Spread, and Sustainability of Health and Care Technologies (NASSS), and the Reach, Effectiveness, Adoption, Implementation, and Maintenance (RE-AIM) framework to categorize results and understand methodologies used. As a scoping review, we mapped the available evidence, rather than synthesizing outcomes, appraising study quality, or estimating effectiveness. To broaden coverage and strengthen interpretation, we crowdsourced additional studies and sought feedback on our preliminary findings from a network of respiratory experts across 17 countries.

Results: Overall, 24,672 studies were identified; after deduplication, 14,811 were screened; 84 studies from 31 countries were included in the final review. The digital respiratory technologies comprised apps, platforms, chatbots, and smart devices. Reported technological functionalities encompassed remote consultation, clinician monitoring, video directly observed therapy, remote rehabilitation training, self-management support, education, monitoring medication adherence, and a school-based remote clinic. CFIR, RE-AIM, and the plan-do-study-act (PDSA) cycle were the most widely used frameworks. Successful implementation used simple technologies that fitted existing workflows and avoided additional workload. Co-development and trust-building with end-users influenced motivation and adoption, while leadership, team cohesion, and communication facilitated success. Barriers included insufficient resources, poor interoperability, lack of funding and reimbursement, and limited technical support.

Conclusions: This scoping review provides a cross-condition review of digital respiratory technologies implemented in routine clinical practice. Unlike previous disease-specific or experimental-focused reviews, our innovative approach used established implementation Theories, Models, and Frameworks (TMFs) to identify shared barriers and enablers across diverse populations and health care systems. We summarize key implementation domains in state-of-the-art digital respiratory care and identify major gaps related to health equity, patient–clinician trust, continuity of support, and environmental sustainability. These findings emphasize the value of using implementation TMFs for scaling effective, patient-centered digital respiratory care in real-world settings.

International Registered Report Identifier (IRRID): RR2-10.1371/journal.pone.0314914

(J Med Internet Res 2026;28:e88325) doi:10.2196/88325

KEYWORDS

digital health; respiratory; artificial intelligence; implementation; review

Introduction

Chronic respiratory diseases affect more than 545 million people worldwide, have a significant impact on individuals' quality of life, and are a major burden to health care systems [1]. Digital health tools that can contribute to regular monitoring of respiratory conditions are prone to exacerbation and support better health outcomes by enabling timely interventions. Conventionally, telemedicine, telecare, e-health, and mobile health have been defined as the use of electronic means, such as information and communication technologies, or mobile devices and remote sensors, to deliver health services [2,3]. Digital therapeutics are "interventions that are driven by high-quality software programs to treat, manage, or prevent a disease or disorder" [4]. Digital respiratory health is an umbrella term for all these modalities, potentially incorporating artificial intelligence (AI), to support routine respiratory care [5] (Figure 1). Apps, platforms, and smart wearable or portable devices can be interconnected to support diagnosis, underpin personalized medicine, and enable self-management. Real-time video facilitates clinician-supervised remote rehabilitation by enabling health care professionals to guide patients through prescribed exercises within their homes [6]. It also allows clinicians to directly observe patients' medication adherence and inhaler technique, thereby supporting more accurate assessment and timely intervention. By integrating environmental data and using AI as the "brain," the system can advise on actions to prevent exacerbations and hospitalizations. Routine data collected from these tools aids interpretation of medical images and public health decisions [7].

XSL•FO
RenderX

Figure 1. Digital respiratory health landscape. Image created using ICOGRAMS EDUCATION, under CC-BY license [8,9]. AI: artificial intelligence; GP: general practitioner.

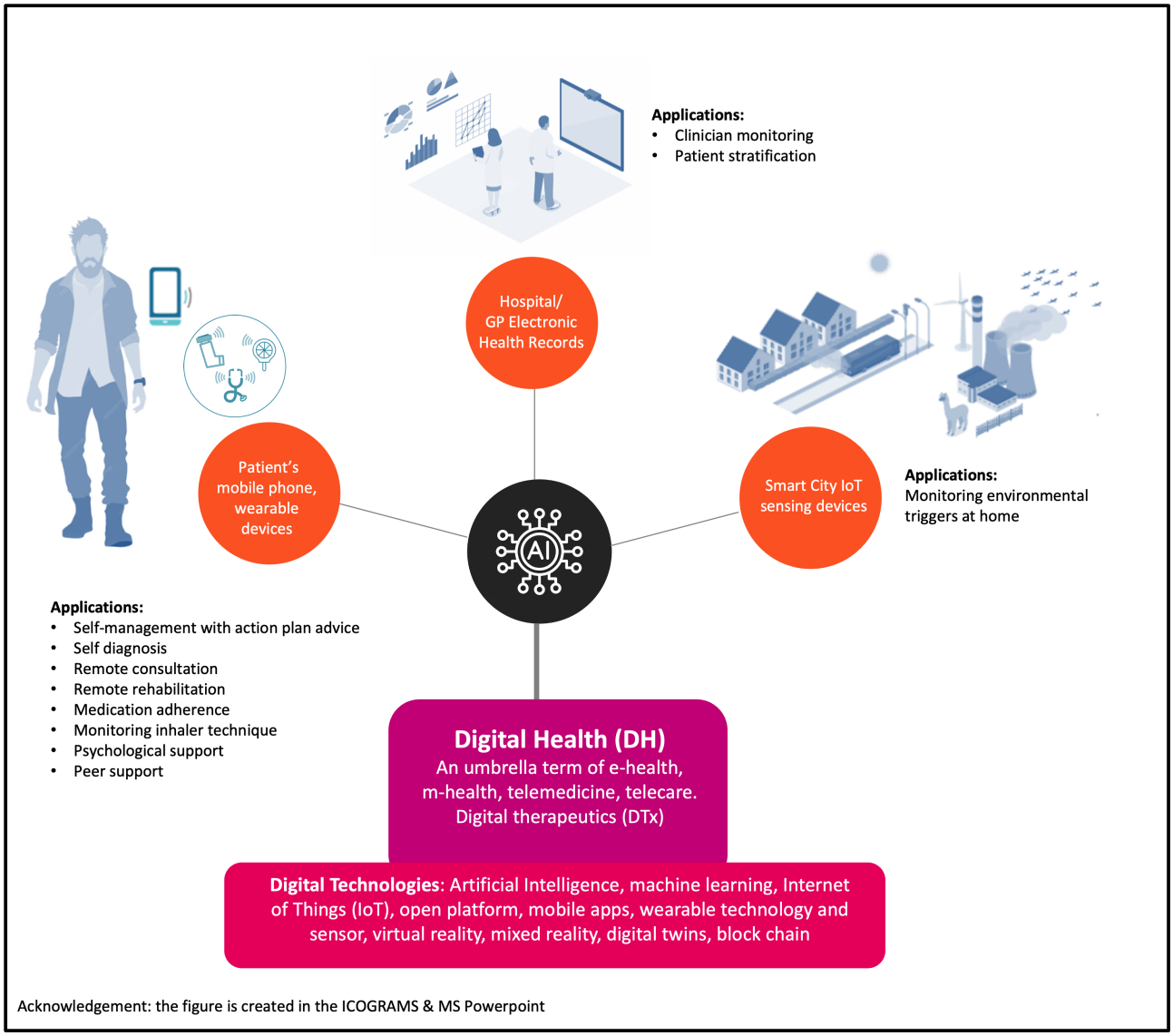

The global COVID-19 pandemic accelerated adoption of digital health technologies, enabling rapid deployment potential [5-7]. However, integrating digital health technologies into routine care is complex. Short-term success does not guarantee long-term sustainability. To address this challenge, implementation Theories, Models, and Frameworks (TMFs) were developed. They provide considerations ("domains") and a systematic approach to guide planning and evaluation of the digitization process, helping organizations to identify barriers and facilitators for successful implementation. They also share similar descriptions of the domains, but terminologies vary. Widely used TMFs are Consolidated Framework for Implementation Research (CFIR), Nonadaptation, Abandonment, and Challenges to the Scale-up, Spread, and Sustainability of Health and Care Technologies (NASSS), Reach, Effectiveness, Adoption, Implementation, and Maintenance (RE-AIM), Exploration, Preparation, Implementation, and Sustainment (EPIS), Unified Theory of Acceptance and Use of Technology (UTAUT), Technology Acceptance Model (TAM), and Theoretical Domains Framework [10]. CFIR and NASSS framework are determinant frameworks that delineate categories and specific domains that function as either barriers or facilitators, thereby influencing the outcomes of implementation processes [11,12]. RE-AIM is an evaluation framework that provides specific aspects of domains that can be evaluated to determine the success of an implementation [13]. EPIS is a process model and determinant framework that outlines the domains and sequential steps involved in translating interventions into routine practice [14]. UTAUT and TAM are technology-focused models. They conceptualize technology adoption as domains of perceived usefulness and perceived ease of use [15,16]. The UTAUT extends this foundation by integrating 8 prior acceptance frameworks and identifying 4 primary determinants—performance expectancy, effort expectancy, social influence, and facilitating conditions—moderated by user characteristics such as age and experience. Theoretical domains framework is a classic theory that can be applied to enhance understanding and/or explanation of various aspects of implementation [17]. Of these well-known implementation TMFs, CFIR is the most frequently used TMF. CFIR, NASSS, and RE-AIM are more widely used TMFs as they are relatively comprehensive, covering at least 2 phases of

XSL•FO
RenderX

implementation (planning, early implementation, and late implementation) [18]. Despite the similarity in the coverage of the domains, the language and glossary are not standardized. Lack of harmonization of domains across borders and sectors [19] obstructs the digital transformation of health care, preventing it from providing safe, fair, and accessible care for everyone [20].

The World Health Organization (WHO) considers digital transformation to be essential for achieving universal health coverage, the EU (European Union) endorses implementation of technological advances to improve patient outcomes and promote patient-centered care, and adoption of digital health is a strategic priority for the European Respiratory Society (ERS) [21,22]. However, only a minority of digital health studies progress beyond the pilot or local intervention stage [23], and harmonization across borders and sectors is rare [19], with the welcome exception of the European Health Data Space, which has been greeted by the ERS as an opportunity to push forward health care, research, and policymaking [24,25].

Digital respiratory technologies are increasingly being adopted across a wide range of respiratory conditions and in diverse health systems. Exploring and adapting to clinical and socioeconomic context is crucial for successful local implementation, though shared barriers and enablers can provide useful learnings that help shape the feasibility and uptake of digital interventions [1,4]. Adopting a global perspective enables the learnings from successful implementation efforts to inform less developed technological initiatives. Using TMFs can support this process by ensuring the challenges are comprehensively considered; this is the underlying rationale for this global scoping review.

The CONNECT Clinical Research Collaboration (CRC) was launched by the ERS in 2023 with the goal of addressing the overarching challenges of whole systems implementation of respiratory digital health care. The core, overarching objective of CONNECT is to bridge the gap between existing disease- or location-focused digital respiratory health initiatives and sustainable, equitable, connected implementation into clinical respiratory practice in diverse health care systems [26]. CONNECT has 5 objectives, one of which focuses on systematically reviewing existing digital respiratory health care embedded in routine clinical practice. In this scoping review (WPIII), we explored published initiatives that integrated and implemented digital respiratory technologies into routine clinical workflows. Our objectives were to scope and describe:

1. Characteristics of respiratory digital health initiatives implemented in routine clinical practice in the last 10 years.
2. TMFs are used to develop and evaluate the implementation strategies.
3. Barriers and enablers to implementation.

Insights relevant to reducing inequity, enhancing patient and professional relationships, supporting the patient journey, and reducing adverse environmental impact.

Methods

Overview

We followed Arksey and O'Malley's methodology for scoping reviews [27] and used Covidence software to manage the review process, including to de-duplicate the studies [28]. We used the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) to guide reporting [29] and PRISMA-S to report the review methodologies and findings [30].

Published Protocol

Detailed methodology is described below with additional details available in our published protocol [31]. There were no deviations from the protocol.

Information Sources

We included ten databases (MEDLINE, Embase, CINAHL, PsycINFO, Cochrane Library, Web of Science, Scopus, IEEE Xplore, CABI Global Health, and WHO Medicus). We browsed reference lists manually with forward searches on included studies using the International Statistical Institute Proceedings. To ensure robust representation of research from low- and middle-income countries, we incorporated the low- and middle-income country (LMIC)–focused databases CABI Global Health and WHO Medicus. In addition, we consulted 28 colleagues in the CONNECT group to identify the missing studies in their networks (see Peer Review and consultation exercise section below for details).

Search Strategy

The search strategy was newly developed for this review to reflect its defined scope and objectives. We developed this search strategy with the Academic Support Librarian at the University of Edinburgh, used the key search terms of "respiratory condition" AND "digital technology" AND "implementation" to search ten databases over a 10-year period from December 12, 2013, to December 12, 2023 (update search performed on April 4, 2025, and February 7, 2026). Given the substantial number of studies, we applied a time limit filter to capture only those falling within the defined search window. In the update searches, we re-ran the same search terms and methods during the update search to identify the newly published studies. See Multimedia Appendix 1 for detailed search strategy.

Eligibility Criteria

We included any digital health intervention implemented to support routine patient care. Patient care included but was not limited to digital support for diagnosis, self-management, monitoring, medication adherence and compliance, education, psychological support, social support, remote consultation, or health professional-facing interventions. For example, a decision support system that was designed to support delivery of care was included, but an administrative system for scheduling appointments would be excluded.

Implementation studies are not always accurately indexed, so we defined the key features of "implementation" to guide the selection process.

XSL•FO
RenderX

The inclusion and exclusion criteria are outlined in Table 1. See Multimedia Appendix 2 and our published protocol [24] for further details of the criteria and definitions.

Table . Inclusion and exclusion criteria.

Criterion	Definition	Operationalization and examples
Inclusion criteria		
Population	Patients with any respiratory conditions (short or long-term) of all ages	Where there was doubt about whether a condition was "respiratory," we included those conditions covered by the Assemblies and Groups of the ERS[a]
Intervention	The intervention should be delivered within the existing service and available to all clinically-eligible individuals (specifically not only to people who consent to the research)	Existing staff may be upskilled for the purpose, but interventions that require additional staff or resources for day-to-day delivery are unlikely to be sustained beyond the end of the project and were therefore excluded
Setting	Any health care setting was eligible	—[b]
Comparator	Typically usual care	Many studies used before-and-after observational designs
Outcomes	Outcomes should reflect the uptake and impact of the intervention within the population	for example, using routine data
Duration:	Any duration	We did not specify a duration for the evaluation, but only included studies in which the intervention was (or was intended to be) embedded for the long term in routine care
Study design	All study types	No geographical or language restrictions
Location	Any location	We did not exclude by language. CONNECT is a global network with members fluent in most major languages
Exclusion criteria		
Administrative only	Digital health interventions that did not directly involve patient care	for example, interventions such as workflow management, appointments, or triage systems with exclusively administrative purposes
Education	Health professional education	for example, online conferences and online courses
Pandemic management	Population-level COVID-19 initiatives, such as contact tracing or vaccination programs	COVID-19 was considered as a respiratory condition when studies focused on individual-level care, such as hospital-at-home programs or management of COVID-19 in people with chronic respiratory disease
Effectiveness research	Randomized controlled trials conducted to test the effectiveness of the intervention	for example, those requiring participant consent process to individual randomization) were excluded as our interest is in implementation in routine clinical care
Research types		
Conference abstracts	Potentially relevant abstracts were tagged and monitored for subsequent publications.	—
Review articles	Reviews and other relevant papers that did not meet inclusion criteria were retained for background reading and reference checking	—
Gray literature	We did not search for gray literature	—

[a]ERS: European Respiratory Society.

[b]Not applicable.

XSL•FO
RenderX

Selection Process

This review was conducted by twenty-seven volunteers from the ERS CONNECT CRC, with a few hours funded by CONNECT for a post-doctoral researcher (ICYH) to provide training, oversight, and ensure quality control. Independent duplicate screening was carried out by the trained volunteers, with disagreements resolved by ICYH, AHYC, and HP in discussion with the 6 volunteers (ZM, AOA, NJR, KBS, SOS, and CF) who reviewed most studies (see protocol for further description of this process [31]). Divergent interpretations (eg, of implementation-related concepts, and which exclusion criterion took precedence) were clarified in discussion, and we subsequently achieved 100% agreement.

Data Charting, Extraction, and Data Items

Using a similar model of selected volunteers (AOA, SOS, SK, and CF) overseen by the lead author (ICYH, with AC and HP), data extraction was conducted in duplicate, with a final check on accuracy by ICYH. Data extraction was completed using a piloted, refined data extraction sheet with headings reflecting the four objectives (see Multimedia Appendix 3 for the full data extraction table).

Collating, Summarising and Reporting the Results

We used a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram to report the selection process [29,30] and collated a narrative summary of qualitative data with charted results from the quantitative data to address the review objectives:

1. Objective 1: we used a bubble plot of the technology components and their associated respiratory conditions to visualize the features integrated into digital health initiatives across various respiratory conditions.
2. Objective 2: we identified the commonly used frameworks (see description in the introduction) and their usage in different countries and income groups. Initially, we used NASSS domains (Condition Complexities, Technology, Value Proposition, Adopters, Organization, Wider System, Embedding, and Adaptation Over Time) to categorize reporting, subsequently integrating features from CFIR and RE-AIM to develop an enriched model that comprehensively captures the complex dynamics influencing the implementation of digital technology within routine clinical practice.
3. Objectives 3 and 4: we used narrative synthesis to categorize the implementation barriers and enablers identified by the authors, and their insights related to health inequities, patient–professional relationships, patient care pathways, and environmental impacts.

Peer Review and Consultation Exercise

We engaged with colleagues from the CONNECT CRC collaboration from the outset and maintained ongoing interaction throughout the review process. This group includes academic researchers, clinicians with experience in implementing digital respiratory health technologies, and collaborators with technological expertise. Representing 17 countries across both high-income and low- and middle-income settings, 28 colleagues provided diverse perspectives on the review process. We conducted online workshops at key stages of the review.

1. Beginning of the review: to explore topical themes and practical challenges and refine research questions.
2. Title and abstract, full text screening stages: to fine-tune the inclusion criteria, exclusion criteria, and discuss selection conflicts, crowdsourced potential studies to identify gaps in the included studies and suggest additional papers for consideration.
3. Data charting and collation stages: we recorded a short video summarizing preliminary findings and used an online memo board to gather asynchronous feedback, facilitating engagement across different time zones. The feedback was collected in parallel with the discussions after our abstract presentation at the ERS Congress to finalize the findings and the practical interpretation.

Results

Overview

Figure 2 is the PRISMA-ScR flow diagram. A total of 24,672 records were identified. After removal of duplicates, the titles and abstracts of 14,800 records were screened. Of these, 14,443 records were excluded, primarily because they focused exclusively on COVID-19 without examining an underlying respiratory condition (eg, chronic obstructive pulmonary disease [COPD] or asthma), or because they were conference abstracts, systematic reviews, or scoping reviews. Reviews were retained at the search stage rather than filtered automatically to ensure that conclusions of relevant reviews could be considered when interpreting and comparing results.

XSL•FO
RenderX

Figure 2. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for scoping reviews) flow diagram.

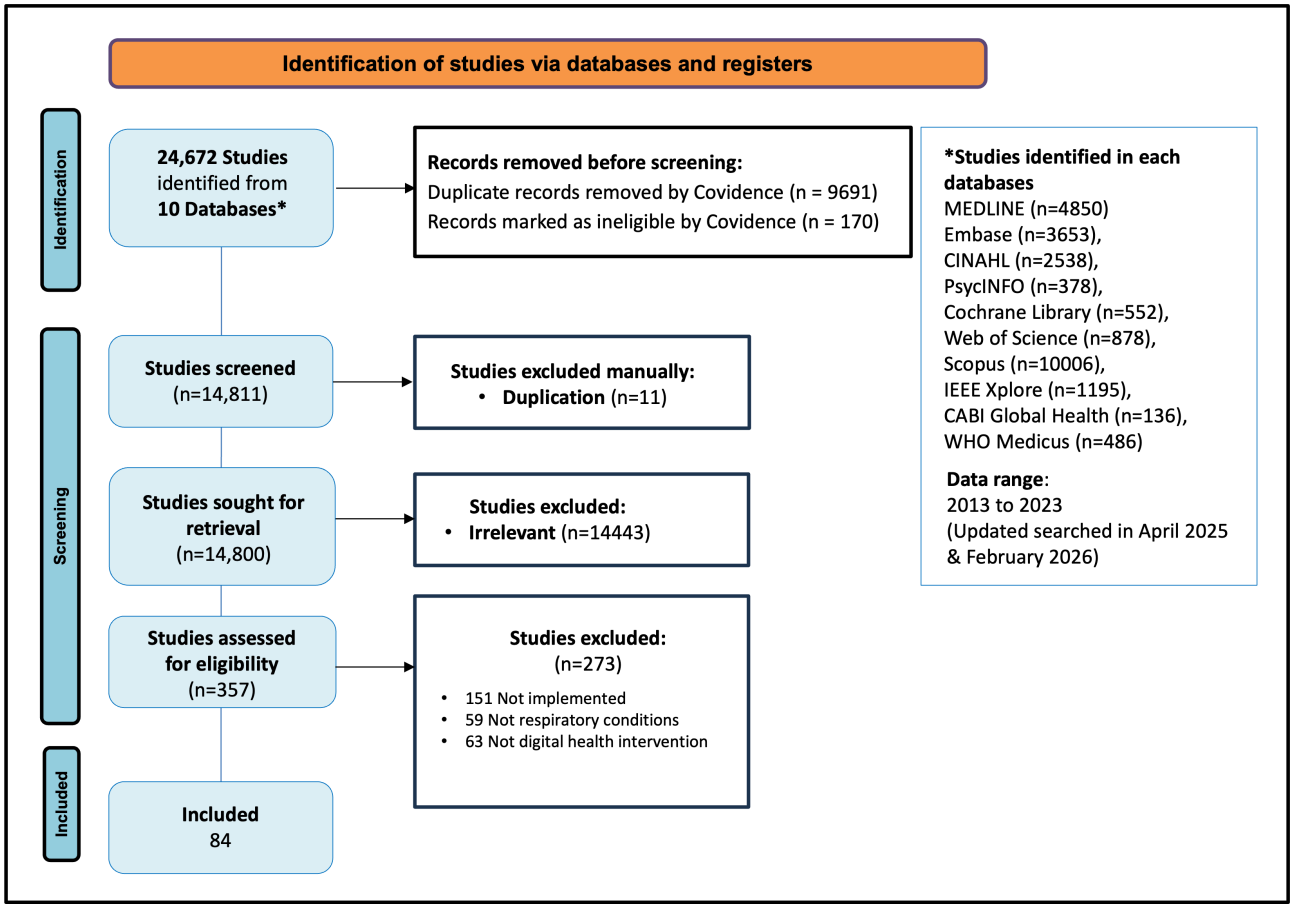

A total of 357 full-text articles were assessed for eligibility. Following full-text screening, 273 articles were excluded. The most common reasons for exclusion of full-text papers were lack of implementation in routine care (n=151; 55%), no inclusion of respiratory patients (n=59, 22%), or initiatives focused on administrative systems for management, or clinicians rather than providing patient care (n=63, 23%). A total of 84 studies [32-115] met the eligibility criteria and were included in the review.

Study Characteristics (Objective 1)

Characteristics of the 84 included digital respiratory initiatives are described below, summarized in Table 2, and detailed in Multimedia Appendix 3 [32-115].

XSL•FO
RenderX

Table . Characteristics of the included studies.

Characteristics of the Included Program (n=84)	
Conditions	No of program (n=84), n (%)
Asthma, asthma, and AR[a]	18 (21)
COPD[b], COPD and ILD[c], COPD and HF[d]	23 (27)
Cystic fibrosis	6 (7)
Lung cancer	8 (10)
Lung transplant	3 (4)
Sleep-related breathing disorders	3 (4)
Tuberculosis (TB)/HIV	10 (12)
Multiple respiratory conditions	13 (15)
Year	
Pre–COVID-19	47 (56)
During COVID-19	32 (38)
Post–COVID-19	5 (6)
Participant age	
Adult (16 years or older)	61 (73)
Pediatric (younger than 16 years)	14 (17)
All ages	9 (11)
Countries	
High-income countries (IHCs)	71 (85)
Low and middle-income countries (LMICs)	7 (8)
Rural, low-resourced regions	6 (7)
Implementation scale	
Country	20 (24)
Regional	33 (39)
Service	31 (37)

[a]AR: allergic rhinitis.
[b]COPD: chronic obstructive pulmonary disease.
[c]ILD: interstitial lung disease.
[d]HF: heart failure.

Participants

The initiatives provided care services for pediatric (n=14; 17%) and adult (n=61; 73%) patients, or both (n=9; 11%). They mostly targeted specific respiratory conditions, but 15% (13/84) included multiple conditions. Age ranges reflected different disease profiles, with interventions for asthma [32-49] and cystic fibrosis including children [73-78]; while COPD [50-72], lung cancer [79-86], lung transplant [87-89], and sleep-related breathing disorders targeted older adults [90-92]. Tuberculosis or multiple respiratory conditions involved all age groups [93-115].

Interventions

Features of the mobile apps, digital platforms, and monitoring devices are shown in Figure 3.

Most commonly, clinicians used an app, either alone or connected with multiple monitoring devices (n=36; 42%), videoconference, SMS, phone communication features for health coaching, follow-up consultations, discussing action plans, triage, and escalating care to other providers (n=48; 57%). Four specific categories of applications were identified:

1. Video directly observed therapy for remote monitoring of adherence to tuberculosis medication [93,95-98,101,102].
2. Diagnosis of sleep apnea and home monitoring of continuous positive airway pressure therapy [90-92].
3. Virtual rehabilitation is used to provide remote exercise training for respiratory conditions at home or community settings [46,50,68,89,111].
4. School-based telemedicine services to provide care for children with asthma in rural areas [37,41].

XSL•FO
RenderX

Figure 3. Bubble plot of the technology features and their associated respiratory conditions [32-115]. APAP: automatic positive airway pressure; AR: allergic rhinitis; CDSS: clinical decision support system; COPD: chronic obstructive pulmonary disease; CPAP: continuous positive airway pressure; EHR: electronic health record; HF: heart failure; ILD: interstitial lung disease; EMM: electronic medication monitor; vDOT: video directly observed therapy.

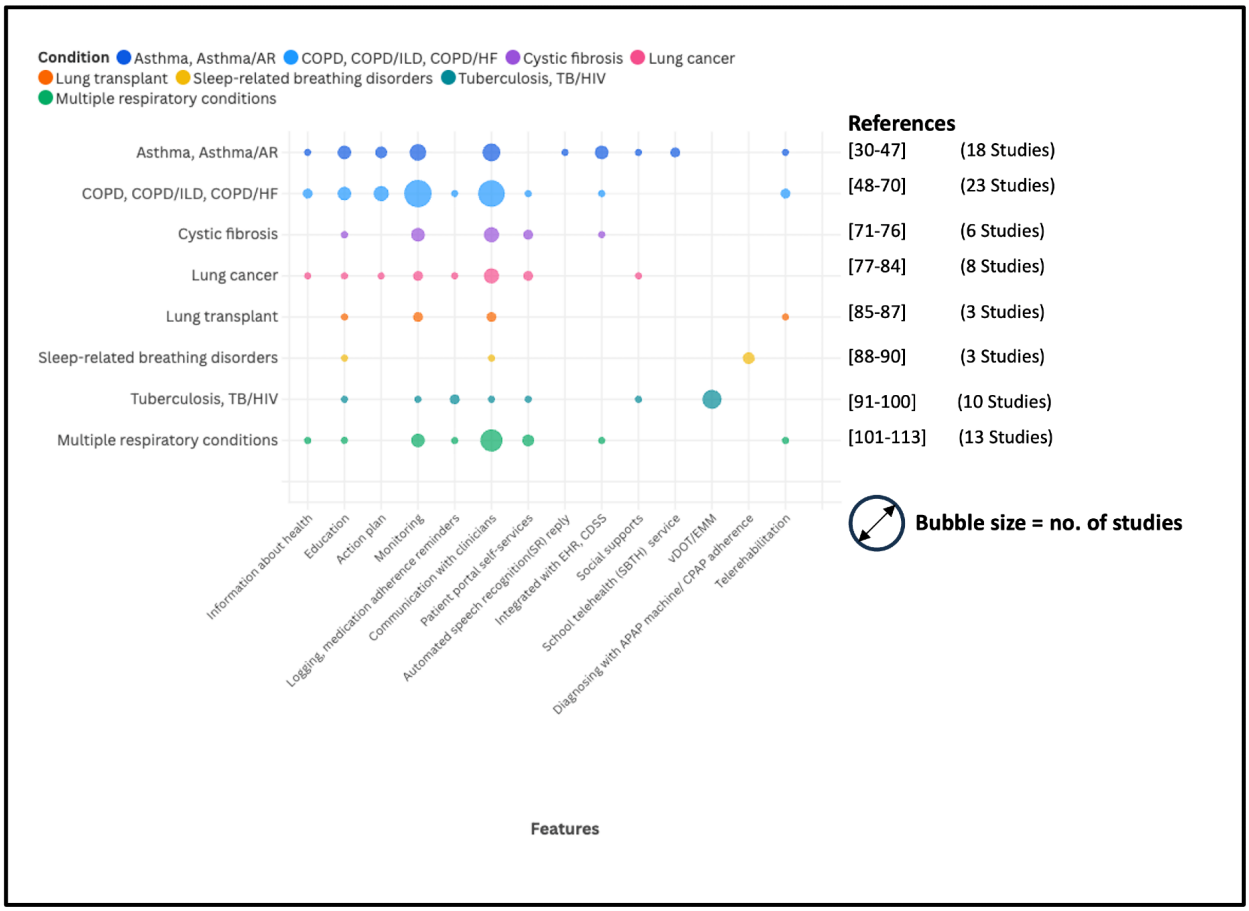

Setting for the Technology Service

Most initiatives (n=64; 76%) were implemented within individual services (hospitals, GP clinics, and community centers) or regional clusters. Country-wide initiatives (n=20; 24%) included one application that was deployed in 22 countries. Of the 84 initiatives, most were deployed outside the EU region, including 36% (n=30) from the United States and 32% (n=27) from other countries such as Australia, Japan, Canada, China, India, Russia, South Africa, South Korea, and Singapore. Only 32% (n=27) originated from the EU.

Methodology

Of the 84 studies, 81% (n=68) [33,34,36-39,41-46,49,50,52,54,56-72,80,81,84-91,93-107,111-115] were service evaluations using pre- and poststudy designs, longitudinal observational, interrupted time-series, participatory action, or case study methodology, 19% (n=16) [32,35,40,47,48,51,53,55,73,79,82,83,92,108-110] studies were a narrative summary of their implementation experiences and learning. Overall, 38% (n=32) [37,42,43,46,47,51,56-60,64,67-71,73,75-77,84-86,91,94-97,99,100,105] used interviews with patient and professional stakeholders to explore the implementation process.

Outcomes

We did not synthesize outcomes or summarize effectiveness, as scoping review methodology is designed to map the available evidence rather than evaluate study quality or determine effectiveness. Overall, 50 studies reported clinical outcomes [32-34,36,39,45-54,56,57,59,61-65,71,72,74,77,78,81,87,89-96,98,99,101-106,109-111,113-115], with each study assessing between one and four outcomes. Of these, health care service quality-related outcomes were reported in more than half of the initiatives (n=35; 42%), followed by prescribing or treatment adherence (n=18; 21%), and health-related outcomes, including respiratory-specific outcomes such as disease control and exacerbations, quality of life, and the mental health benefits (n=6; 7%).

Two initiatives used real-world observational data from an app to explore correlations between treatment usage and asthma control [49] or risk of needing care escalation [52].

Use of Implementation Theories, Models, and Frameworks (TMFs; Objective 2)

Of the 84 initiatives, 30% (n=25; and all the 18 country-level evaluations) used one or more TMF to guide development or evaluation (Figure 4).

XSL•FO
RenderX

Figure 4. Implementation theories, models, and frameworks (TMFs) used by the high-income and low middle-income countries [34,37,38,41,43,44,52,53,59,60,66,69,71,75,76,79,83-86,97,101,105,114,115]. AMA: American Medical Association; CFIR: Consolidated Framework for Implementation Research; EPIS: Exploration, Preparation, Implementation, and Sustainment framework; NASSS: Nonadoption, abandonment, scale-up, spread, and sustainability framework; ReAIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance framework; TAM: Technology Acceptance Model; TDF: Theoretical Domains Framework; TMF: Theories, Models, and Framework; UTAUT: Unified Theory of Acceptance and Use of Technology.

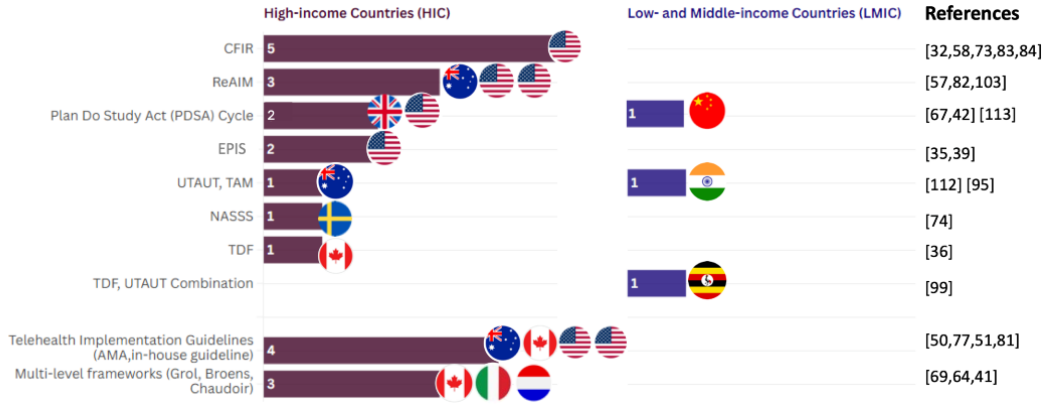

The remaining 70% (n=59) of studies did not report using an implementation framework or describe their approach to initiative development. The most commonly used frameworks were CFIR (n=5; 20%) [34,60,75,85,86], RE-AIM (n=3; 12%) [59,84,105], and the plan-do-study-act (PDSA) cycle (n=3; 12%) [44,69,115]. Of the initiatives that used TMF, three were from the LMICs (China, India, Uganda). Seven service- and regional-level initiatives used a range of guidelines and multilevel models [43,52,53,66,71,79,83] to observe the digitization process. Typically, they focused on the benefits of an agile approach that incorporated feedback from patients and clinical staff.

Common Implementation Domains

Of the 7 interconnected implementation domains from the NASSS, 6 were addressed frequently; they were “condition complexities,” “technology,” “value proposition,” “adopters,” “organization,” and “wider system” (Table 3). Only “embedding and adoption over time” was not explicitly reported in the included studies, though the need to incorporate long-term monitoring was mentioned.

XSL•FO
RenderX

Table . Common implementation domains.

Commonly reported implementation domains (n=84)	n (%)
Technology	
Integration (vendor partnerships, use of existing infrastructure, and interoperability)	36 (42.9)
System accuracy and reliability, security and privacy, confidentiality	11 (13.4)
User-friendly interface and features	6 (7.1)
Development supported by evidence	2 (2.4)
Value proposition	
Values to end users (cost, COVID-19 emergency, and service advancement)	29 (35.4)
Adopters	
Acceptability and satisfaction	38 (45.2)
Adoption, usability, and adherence rate	39 (47.6)
Co-develop with end users	13 (15.9)
Motivation, perceptions from end users	12 (14.6)
Technical support and devices to patients	11 (13.4)
Self-efficacy, knowledge	5 (6.1)
Identify early adopters	4 (4.9)
Staff willingness and organizational readiness to change	4 (4.9)
Trust from end users	2 (2.4)
Organization	
Staff training, supports, and extra workload due to implementation	26 (31)
Develop interprofessional team and workflow	26 (31)
Use opinion leaders, subject matter expertise, advisory boards, workgroups	10 (12.2)
Procurement processes, supply chain, device logistics	3 (3.7)
Wider system	
Reimbursement, insurance, and funding	16 (19)
Policy, social benefits, guideline compliance, “tech push” as driver for implementation	14 (17.1)

Within the widely reported domain of “condition complexities,” examples were both organizational (the rapid transition and unprecedented strain on health care systems of the COVID-19 pandemic [87] as well as incomplete registries in lung cancer [82,84]). Features of specific clinical contexts included multimorbidity in COPD [68], poor adherence in asthma [33,34,42], tuberculosis [93-102], and obstructive sleep apnea [90-92], as well as the high degree of disciplined self-care required in cystic fibrosis [76]. Technology domains such as choosing in-house development or a proprietary solution, or integration with the electronic health record were mostly considered in the development phases. Value proposition domains were described as the incentive for launching the initiative. Adoption and organizational factors were influenced by inter-relationships, communication, and culture. Opinion leaders and subject expertise could drive implementation. Wider system domains included both internal and external service goals, political or policy aims, intervention costs (money; time), and funding (grants; reimbursement).

Implementation Barriers and Facilitators (Objective 3)

More than half (n=54; 64%) of the studies reported barriers and/or enablers to implementing digital health initiatives in routine care [33-34,36,47,50-52,55-57,60,64-67,71,73,75,76,79,80,82,84-89,92,93,95-97,100-102,106-108,110,113,114]. These are summarized in Figures 5 and 6; with details in Multimedia Appendix 3.

XSL•FO
RenderX

Figure 5. Facilitators for effective implementation.

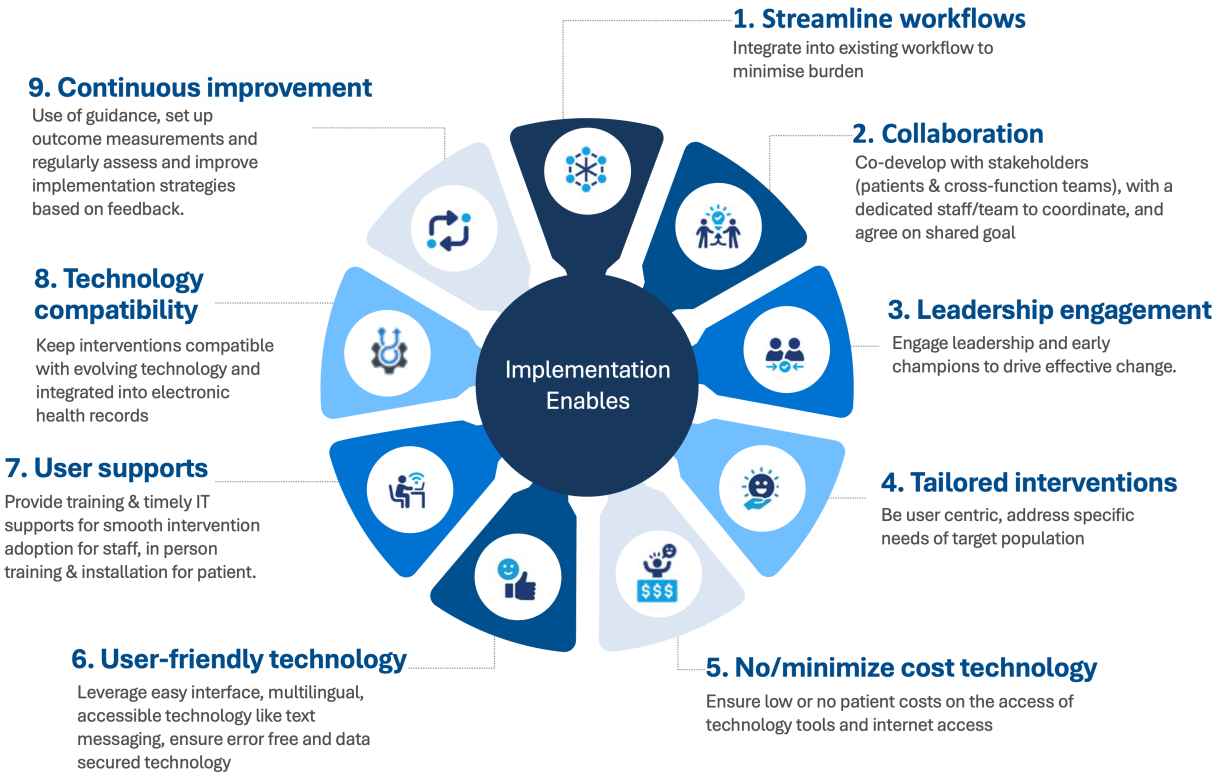

Figure 6. Barriers for effective implementation.

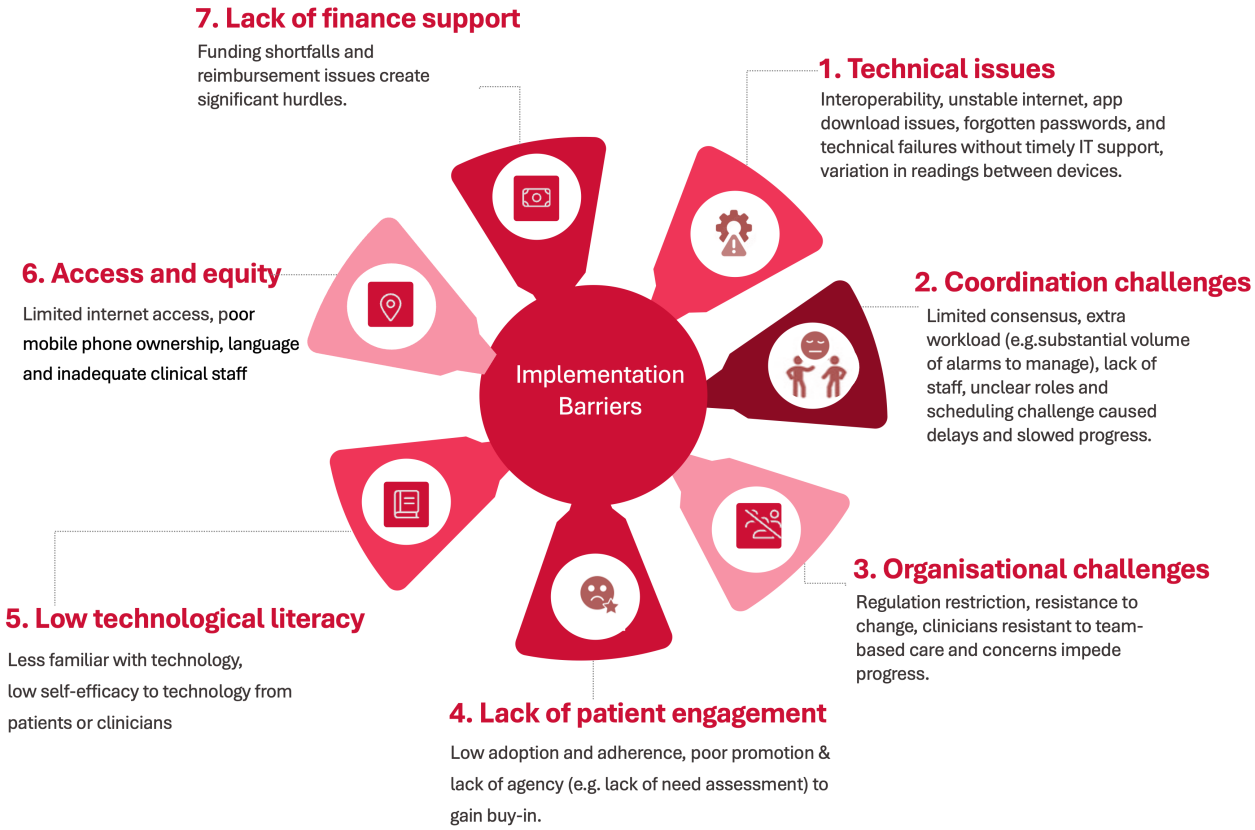

Key issues included:

1. Technical barriers (poor integration with existing technology; software installation problems; password barriers) could be reduced by error-free, user-friendly

XSL•FO
RenderX

interfaces, tailored to the needs of the target population and offering easy access to timely IT support.

2. Coordination challenges (implementation teams with unclear roles; difficulty scheduling meetings) could be mitigated by facilitating co-development from the outset (common goals; clear expectations) and allowing sufficient committed staff time.
3. Streamlining processes and integrating into existing workflow to avoid additional workload (eg, duplicate data entry) reduced resistance to adopting new ways of working.
4. Engaged leadership and local champions helped convince staff of the value of the innovation and overcome reservations (eg, about third-party providers of systems; or risk to valued interpersonal contact).
5. Optimizing patient engagement by raising awareness, promoting buy-in, and improving technological literacy by using co-design, developing innovations that were patient-friendly, multilingual, addressed actual needs, and used data-secured technology.
6. Removing structural barriers by minimizing cost to the patient (free devices and internet access), thereby reducing unaffordability and the risk of increasing inequities.
7. Funding models that did not reimburse the full costs of a technology-based service (eg, required upgrades of existing legacy systems to maintain compatibility, necessitated multidisciplinary support staff costs) created significant hurdles for clinicians to adopt and sustain the initiative.

Studies reported that the COVID-19 pandemic was a key facilitator in eliminating some previous barriers (eg, additional funding streams and coordination support) as the traditional care model was changed rapidly to digital. Attitudes shifted as both patients and providers recognized the need for social distancing and remote care delivery overriding preferences for valued direct patient interactions.

Insights (Objective 4)

Health Inequity

Digital health initiatives that incorporated remote monitoring and brought services closer to the point of need such as providing remote consultations within schools for students with respiratory conditions would be an effective strategy for improving access to care, but initiatives from LMICs [82,93-97,100,101,113,115], rural regions, and low-income populations [37,41,52,58,85,112] faced significant barriers such as poor transportation, no insurance, poor mobile phone coverage, and health care provider shortage. Tailored interventions and language support facilitated the needs of diverse subpopulations.

Patient and Professional Relationships

Remote consulting could reduce clinicians' workloads but raises concerns about dehumanization and reduced patient engagement compared to face-to-face interactions [34]. Remote pediatric consultations struggled to involve the child (though easier if the child already knew the clinician), and even parents were less active than in face-to-face appointments [44,46]. Building rapport was hindered by clinicians' difficulty sensing body language and responding to emotional cues [43,58,64,86]. Initiatives with 2-way messaging improved access, but ensuring timely responses was difficult [43]. Practical challenges included difficulty checking objective signs [70,73,75,88]. Retaining the option of in-person consultations was deemed safe and respectful for individual preferences [46,68].

Patient Journey

Digital initiatives allowed patients to receive support in their own homes and could improve continuity of care [47,53,107], empower self-care, and ensure that care plans reflected individual needs and preferences [55,81], while reducing barriers of travel time and costs [34,67,91,108,111]. Connected care models, such as a platform connecting multiple monitoring devices or enabling data sharing with specialists across departments, helped to coordinate support, simplify the patient journey, and improve access to multiple specialists [73,77].

Environmental Impact

Remote monitoring and consultation minimized travel [42,58,95,112], especially in rural areas, potentially improving attendance at pulmonary rehabilitation [111]. By decreasing reliance on carbon-intensive travel for health care appointments, digital approaches supported both environmental sustainability and economic efficiency.

We have summarized the do's and don'ts for implementation in Figure 7.

XSL•FO
RenderX

Figure 7. Do's and don'ts for implementation. UI: user interface; UX: user experience.

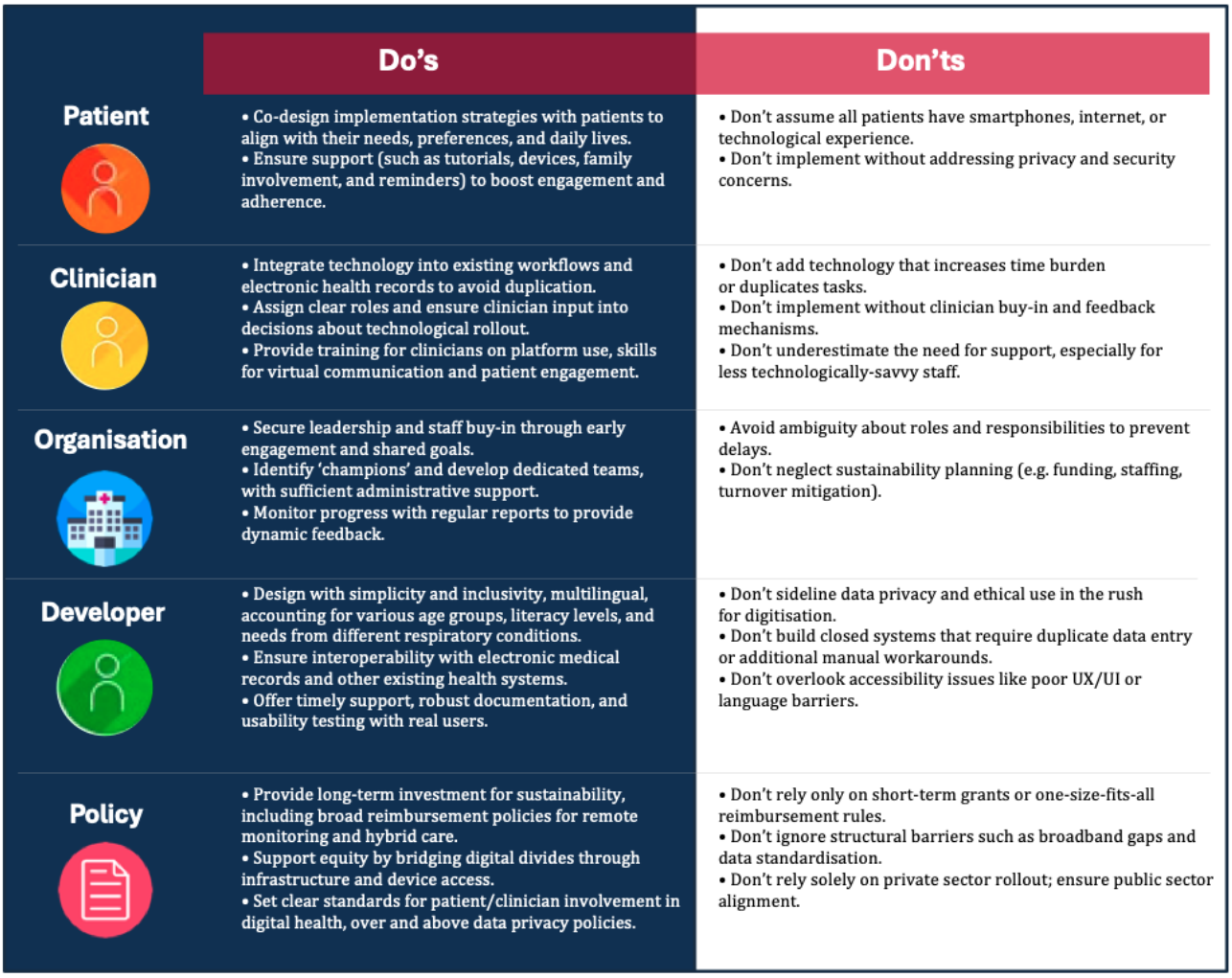

Discussion

Summary of Findings

This review provides an overview of evidence on the implementation of digital technologies from 84 digital respiratory care initiatives across diverse populations and settings. The most targeted population was adult patients with a range of respiratory conditions. The initiatives used apps, digital platforms, and smart devices to enable remote monitoring and facilitate communication with clinicians for care planning and management. Health care service quality, prescribing or treatment adherence, and respiratory-specific clinical outcomes were the most reported outcome measurements. Fewer than a third of initiatives, typically national initiatives in high-income countries, used TMFs such as CFIR, RE-AIM, and PDSA cycles to address condition complexities, technology, value proposition, adopters, organization, and wider system factors. Barriers to implementation included limited technical integration, low patient engagement, funding shortfalls, and organizational challenges, mitigated by early involvement of stakeholders, leadership and champions, user-friendly design, and providing training and support.

Interpretation and Implications in Relation to Published Literatures

Use of Implementation Frameworks

Fewer than a third of the studies used a TMF to guide the implementation and evaluation of their initiative—and those that did were typically country-level initiatives in high-income settings (Figure 4). The comprehensive approaches to development and evaluation embodied by the CFIR and the RE-AIM frameworks may be perceived as resource-intensive, or too complex for local initiatives, or not adapted to the specific policies, infrastructure, and social-cultural adjustments required in LMICs [116]. However, it is notable that the least reported domains (Table 3) are "organizational" and the "wider context;" they emerge as key barriers and enablers, suggesting that a structured approach to digital implementation could be helpful in adapting to diverse clinical, health care systems and socioeconomic contexts.

Complexities in Implementation

Clinical complexity challenged the implementation of digital health care for respiratory conditions. For example, multimorbidity is common in COPD, and initiatives required technology developers to integrate systems for multiple conditions in often deprived social contexts whilst providing error-free technology that is user-friendly and compliant with

XSL•FO
RenderX

global data privacy regulations [117]. The organizational enablers and barriers resonated with those found in a recent survey of stakeholders who facilitated innovation adoption across 18 UK National Health Service sites [118]. Ensuring timely technical support and involving cross-functional teams and patients early in the design phase were critical (but complex) tasks requiring strong leadership to align stakeholders and manage resources effectively [119]. Additionally, the challenges and enablers were dynamic; initial barriers like staff resistance due to perceived workload could evolve into facilitators as confidence in the intervention grew [120]. The COVID-19 pandemic accelerated adoption by easing regulations and reimbursement barriers, shifting preferences toward remote care [5].

Complexity in implementation reflects the multiple stakeholders involved in developing, using, and maintaining digital care and the need to align organizational change with existing workflows and routines [121,122].

Future Challenges for Implementation

Two key challenges that remain under-explored are avoiding increasing inequity and considering environmental impacts. Initiatives prioritized clinical outcomes and service outcomes over assessing differential access, usability, or benefits across diverse socioeconomic, geographic, and demographic groups.

Digital Divide

A substantial number of studies showed that, despite their promise, digital health technologies can inadvertently widen existing health inequities—“the digital divide”—particularly in LMICs, rural areas with poor infrastructure, or other resource-constrained settings [123]. About one in five studies in this review highlighted inequity as a concern [37,41,52,58,82,85,93-97,100,101,112,113,115]. Disparities in broadband availability, digital literacy, and culturally responsive technology design disproportionately burden underserved groups, increasing uneven uptake and consequently exacerbating differential health outcomes [124]. Digital-only interventions heighten vulnerability in these populations and may entrench, rather than alleviate, socioeconomic barriers to care [125]. In response, care-coordination models that provide structured, personalized support, encompassing routine screening for digital-health readiness, proactive patient navigation (eg, including community health workers), and coordinated cross-sector referrals, offer plausible countermeasures by embedding digital tools within human-facilitated, inclusive workflows [126]. However, realizing the benefits of these equity-first hybrid models requires explicit operational definitions and clear delineation of responsibilities between technological components and coordinating staff, alongside sustained solutions for root-cause constraints such as reliable access to devices (eg, mobile phones, wearables, and AI), affordable and stable connectivity [127].

Environmental Sustainability

Within the studies in this review, 5 considered environmental sustainability to be an important benefit of the technology initiative [42,58,95,111,112]. Remote care is often framed as a means of reducing carbon emissions and generating economic benefits [128]. However, the energy-intensive infrastructure required to develop and operate AI systems raises concerns about increased energy consumption, making the long-term net carbon impact on health care uncertain [129]. Training large-scale AI models, such as deep neural networks and generative models, demands substantial computational power [130]. In the meanwhile, AI combined with internet of things and the things network smart-city sensors offers new ways to address environmental challenges, including disaster forecasting, climate modeling, and operational energy optimization. Examples include sentiment analysis to shape environmental messaging, reinforcement learning to optimize water and energy use, and AI simulations to evaluate conservation strategies under varied ecological conditions [131]. Furthermore, the emerging concept of “Green AI” calls for quantifying embodied carbon and operational energy use across the full lifecycle of AI infrastructure, from construction to decommissioning data centers, together with associated Scope 1 - 3 (direct and indirect) supply-chain emissions, to support meaningful carbon reduction efforts [132,133]. However, deploying AI in routine respiratory care by adding health care-specific supply chains (eg, inhaler recycling pathways) can contribute to an overall beneficial footprint [134]. Further research is needed to develop pragmatic key performance indicators and measurement frameworks that capture both carbon impacts and health benefits.

Trust

Reduced personal interaction in remote consultations raises trust concerns. Strong patient-clinician relationships require confidence in technology, clear motives, and empathetic responses [135]. Despite adoption being widely discussed, “trust from end users” was only mentioned in 2% of the included studies, suggesting the need to increase the focus on this core attribute (Table 3). As generative AI becomes increasingly integrated into daily life, evaluating its perceived empathy and trustworthiness in health care is increasingly crucial. Moreover, perceptions of algorithmic bias, data security, and transparency can significantly influence whether patients feel comfortable engaging with AI-supported care [136,137]. Establishing robust guidelines for how AI systems communicate with patients, who is responsible when things go wrong, and how humans supervise AI decisions could help reduce worries about trust, safety, and reliability [138,139]. Ultimately, systematically assessing how different populations experience trust in AI-enabled health care will be essential for equitable and sustainable implementation.

Strengths and Limitations

We used standard methodology PRISMA-ScR and PRISMA-S methodology, while considering the reporting items suggested by Arksey and O’Malley [27] to ensure a transparent and systematic approach to this mixed-methods scoping review [23]. The collaborative process enabled input throughout the review from a multidisciplinary global team with academic and clinical interest in digital respiratory health.

There were some limitations. Despite the use of broad search terms, poor indexing of implementation research may have excluded relevant studies. Gray literature was omitted due to concerns regarding the methodological quality and limited resources. We set a 10-year search time limit to ensure our

XSL•FO
RenderX

review reflected current technology and its implementation, but may have missed valuable lessons from older, well-established initiatives. We included studies from both LMICs and high-income countries to ensure global representation; however, the diverse health care contexts meant the included studies were heterogeneous. To facilitate contextual interpretation, we grouped the findings into low- and high-income countries based on the World Bank's income classification. Most of the initiatives were deployed in non-EU countries, where barriers, such as regulatory requirements and funding structures, were substantially different from those in the EU or low-resource settings. Nevertheless, this broad geographical representation provided valuable insight into common barriers and facilitators at an international level, and our focus on TMFs is a strength, as a comprehensive framework can support the adaptation required to transfer ideas between diverse health care systems.

Our scoping review provides an overview of digital respiratory initiatives rather than evaluating their effectiveness. Our analysis is limited by reported data, and could not detect unreported factors, but key implementation challenges and facilitators were highlighted in most papers.

Conclusions

This scoping review provides insights into factors affecting the implementation of digital respiratory health care, summarizing 84 digital respiratory initiatives, the frameworks they used, and the barriers and facilitators to successful implementation in routine care. The findings provide a contemporary overview of the use of TMFs across LMICs and high-income countries, highlighting the value of structured implementation of TMFs in supporting adoption across diverse clinical settings, health care systems, and socioeconomic contexts. It offers an empirical basis to inform future policy statements and guidance on standardized approaches to the development, evaluation, and reporting of digital health care interventions. Meaningful engagement of multiple stakeholders throughout the project lifecycle is essential to ensure alignment with patient needs and organizational workflows, thereby minimizing additional workload and safeguarding time for direct patient care. Addressing persistent and emerging challenges, including the digital divide, environmental sustainability, and the establishment of trust between technologies such as AI, patients, and clinicians, will be essential to support effective and scalable delivery of digital respiratory health care services.

Acknowledgments

We gratefully acknowledge the advice from Dr Marshall Dozier, Academic Support Librarian, University of Edinburgh, for her help with developing the search strategy. No generative artificial intelligence tools were used in any portion of the manuscript.

Funding

The CONNECT Clinical Research Collaboration is funded by the European Respiratory Society Clinical Research Collaboration (ERS CRC) Fund. This scoping review had limited funding from CONNECT which was used to provide a small honorarium for one of the lead researchers (ICYH). All other work was undertaken by volunteers from the CONNECT collaboration. The CONNECT CRC funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

HP and ICYH conceptualized this study. HP, ICYH, and AHYC led the development of the scoping review with the advice of all authors. ZM, AOA, NJR, KBS, SOS, and CF contributed to the conflict resolutions. AOA, SOS, SK, and CF contributed to data curation, with the final check on accuracy by ICYH. ICYH performed the formal analysis, managed the project, software (created and tested the search codes on the databases, managed Covidence), resources, validated the data, prepared the visualizations, drafted and edited the original paper, which was reviewed, validated, and revised by HP critically. All authors approved the final version, contributed to discussions or meetings.

Conflicts of Interest

Io Chi Yan Hui is an Honorary Fellow in the University of Edinburgh and a Senior Consultant at the PM Group. She receives the European Respiratory Society (ERS) Clinical Research Collaboration fund to develop this study. This study and her research with the University of Edinburgh are independent from, and not financially supported by the PM Group. Her views in this publication are her own, and not those of the PM Group. Neither she, nor PM Group, stand to gain financially from this work. She is the WPIII co-lead of the CONNECT working group and the Chair of the ERS Group 1.04 ehealth/mhealth. Aleksandra Gawlik-Lipinski has received honoraria for lectures from AstraZeneca, support to attend meetings from AstraZeneca and Chiesi. Kate Loveys is a former consultant to and employee of Soul Machines Ltd. Patrick Stafler has received consulting fees from TytoCare for freelance advisory work, outside the submitted work. Shirley Quach has received fellowship and research grants from the Canadian Lung Association, Lung Health Foundation, Asthma Canada, and Canadian Society of Respiratory Therapists, unrelated to the submitted work. Anthony Sunjaya reports research grants from Australia's Medical Research Future Fund and

XSL•FO
RenderX

the National Health and Medical Research Council (NHMRC) outside of the submitted work and all paid to his institution, and reports consultancy fees from the Australian Commission on Safety and Quality in Health Care. AS is also co-founder of a digital health startup, BantingMed, unrelated to this submitted work. Amy Hai Yan Chan reports research grants from Health Research Council of New Zealand, Auckland Medical Research Foundation, Asthma UK, University of Auckland, Oakley Mental Health Foundation, World Health Organization, and Hong Kong University, outside the submitted work and all paid to her institution (the University of Auckland); previously was supported by the Robert Irwin Postdoctoral Fellowship; reports consultancy fees from AcademyeX and Spoonful of Sugar Ltd. and travel support from AstraZeneca; advisory board fees from AstraZeneca, GSK, CSL. Amy Hai Yan Chan is the clinical director of Asthma NZ, and is a member of the Respiratory Effectiveness Group, the scientific advisory board for Asthma Respiratory Foundation NZ, and working group lead for the European Respiratory Society Clinical Research Collaboration "CONNECT." Tonje Reier-Nilsen has received honoraria from ALK, AstraZeneca, MediTuner, and Sanofi for lectures, travel fees, and advisory boards. SOS reports grants from Roche, Dutch Research Council (NWO), and Dutch Lung Foundation (Longfonds); consulting fees from AstraZeneca and Chiesi; honoraria from AstraZeneca, Chiesi, and Janssen-Cilag; support for attending meetings from AstraZeneca and Chiesi; all outside the scope of the submitted work and paid to his institution. SOS reports a patent pending (published 12 March 2026); SOS reports roles as board member of the P4O2 consortium, and national delegate and member of the management committee of the COST ACTION eVoiceNet (CA24128). Hilary Pinnock has research grants from multiple, noncommercial sources for conducting applied respiratory research, all paid to her university and outside of the submitted work. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Detailed search strategy.
[DOCX File, 52 KB - jmir_v28i1e88325_app1.docx]

Multimedia Appendix 2
Detailed criteria and definitions.
[DOCX File, 35 KB - jmir_v28i1e88325_app2.docx]

Multimedia Appendix 3
Full data extraction table.
[XLSX File, 47 KB - jmir_v28i1e88325_app3.xlsx]

Checklist 1
PRISMA-ScR reporting checklist.
[DOCX File, 114 KB - jmir_v28i1e88325_app4.docx]

References

1. Chronic respiratory diseases and health equity by 2050. : World Health Organization (WHO) European Region, European Respiratory Society (ERS); 2025 URL: https://www.knowledge-action-portal.com/sites/default/files/2025-06/WHO-EURO-2025-12340-52112-79990-eng.pdf [accessed 2026-05-29]
2. Pinnock H, Poberezhets V, Drummond D. Digital Respiratory Healthcare: European Respiratory Society; 2023. [doi: 10.1183/2312508X.erm10223]
3. Labaki WW, Han MK. Chronic respiratory diseases: a global view. Lancet Respir Med 2020 Jun;8(6):531-533. [doi: 10.1016/S2213-2600(20)30157-0] [Medline: 32526184]
4. Pinnock H, Hui CY, van Boven JFM. Implementation of digital home monitoring and management of respiratory disease. Curr Opin Pulm Med 2023 Jul 1;29(4):302-312. [doi: 10.1097/MCP.0000000000000965] [Medline: 37132298]
5. Pinnock H, Murphie P, Vogiatzis I, Poberezhets V. Telemedicine and virtual respiratory care in the era of COVID-19. ERJ Open Res 2022 Jul;8(3):00111-02022. [doi: 10.1183/23120541.00111-2022] [Medline: 35891622]
6. Cox NS, Dal Corso S, Hansen H, et al. Telerehabilitation for chronic respiratory disease. Cochrane Database Syst Rev 2021 Jan 29;1(1):CD013040. [doi: 10.1002/14651858.CD013040.pub2] [Medline: 33511633]
7. Poberezhets V, Kasteleyn MJ, Aardoom JJ. The use of digital health solutions for COVID-19: pandemic surveillance, remote monitoring and big data. In: Hilary P, Vitalii P, David D, editors. Digital Respiratory Health (ERS Monograph): European Respiratory Society; 2023:160-169. [doi: 10.1183/2312508X.10001623]
8. Icograms commercial license agreement. ICOGRAMS EDUCATION. URL: https://education.icograms.com/license-agreement-commercial [accessed 2026-06-16]
9. CC BY 4.0 attribution 4.0 international deed. Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/deed.en [accessed 2026-06-16]
10. Atzmon O, Crowther ME, Bei B, O'Connor DA. The use of implementation science theoretical approaches in hybrid effectiveness-implementation type 1 randomised trials of healthcare interventions: a scoping review. Implement Sci 2025 May 16;20(1):23. [doi: 10.1186/s13012-025-01435-6] [Medline: 40380198]

XSL•FO
RenderX

11. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated consolidated framework for implementation research based on user feedback. Implement Sci 2022 Oct 29;17(1):75. [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]
12. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
13. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
14. Aarons GA, Hurlburt M, Horwitz SM. Advancing a conceptual model of evidence-based practice implementation in public service sectors. Adm Policy Ment Health 2011 Jan;38(1):4-23. [doi: 10.1007/s10488-010-0327-7] [Medline: 21197565]
15. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view1. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]
16. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
17. Richardson M, Khouja CL, Sutcliffe K, Thomas J. Using the theoretical domains framework and the behavioural change wheel in an overarching synthesis of systematic reviews. BMJ Open 2019 Jun 22;9(6):e024950. [doi: 10.1136/bmjopen-2018-024950] [Medline: 31229999]
18. Lukersmith S, Woods C, Sarma H, de Miquel C, Salvador-Carulla L. Determining the process components of impact assessment in health and social program implementation: a scoping review of theories, models and frameworks. Public Health (Fairfax) 2025 Mar;240:41-47. [doi: 10.1016/j.puhe.2024.12.056] [Medline: 39862623]
19. Istepanian RSH. Mobile health (m-Health) in retrospect: the known unknowns. Int J Environ Res Public Health 2022 Mar 22;19(7):3747. [doi: 10.3390/ijerph19073747] [Medline: 35409431]
20. Sunyer-Vidal JG, Rodes M, Lupiáñez-Villanueva F. Policy Context for Digital Transformation: Benefits and Challenges of Implementing EU Digital Health Regulation: Digital respiratory healthcare; 2023:132-141. [doi: 10.1183/2312508X.10001423]
21. World Health Organization. Exploring the Digital Health Landscape in the WHO European Region: Digital Health Country Profiles: World Health Organization; 2024.
22. Gappa M, Ryan S, Garcia-Aymerich J, et al. The future of the European Respiratory Society: strategy update 2025. Eur Respir J 2025 Jan;65(1):2402327. [doi: 10.1183/13993003.02327-2024] [Medline: 39819825]
23. Perrin Franck C, Babington-Ashaye A, Dietrich D, et al. iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res 2023 May 10;25:e46694. [doi: 10.2196/46694] [Medline: 37163336]
24. European health data space. European Commission. URL: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en [accessed 2026-04-12]
25. European Respiratory Society. ERS Submission to the European Commission's Proposal on the European Health Data Space [Internet]: European Respiratory Society. URL: https://www.ersnet.org/wp-content/uploads/2022/08/ERS-submission-EHDS.pdf [accessed 2026-04-12]
26. van Boven JFM, Drummond D, Chan AHY, et al. ERS "CONNECT" clinical research collaboration - moving multiple digital innovations towards connected respiratory care: addressing the over-arching challenges of whole systems implementation. Eur Respir J 2023 Nov;62(5):2301680. [doi: 10.1183/13993003.01680-2023] [Medline: 37945041]
27. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
28. Covidence - better systematic review management. Veritas Health Innovation. 2023. URL: https://www.covidence.org/ [accessed 2026-04-12]
29. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
30. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
31. Hui CY, Condon K, Kolekar S, et al. Implementing digital respiratory technologies for people with respiratory conditions: a protocol for a scoping review. PLoS ONE 2024;19(12):e0314914. [doi: 10.1371/journal.pone.0314914] [Medline: 39729438]
32. Benfante A, Principe S, Cicero MN, et al. Management of severe asthma during the first lockdown phase of SARS-CoV-2 pandemic: tips for facing the second wave. Pulm Pharmacol Ther 2022 Jun;73-74:102083. [doi: 10.1016/j.pupt.2021.102083] [Medline: 34628048]
33. Bereznicki B, Peterson G, Jackson S, Walters EH, Gee P. The sustainability of a community pharmacy intervention to improve the quality use of asthma medication. J Clin Pharm Ther 2011 Apr;36(2):144-151. [doi: 10.1111/j.1365-2710.2010.01165.x] [Medline: 21366642]
34. Cvietusa PJ, Goodrich GK, Shoup JA, et al. Implementing health care technology research into practice to improve adult asthma management. J Allergy Clin Immunol Pract 2019 Mar;7(3):908-914. [doi: 10.1016/j.jaip.2018.08.029] [Medline: 30201160]

XSL•FO
RenderX

35. Ainsworth B, Bruton A, Thomas M, Yardley L. One year later: highlighting the challenges and opportunities in disseminating a breathing-retraining digital behaviour change intervention. Digit Health 2020;6:2055207620936441. [doi: 10.1177/2055207620936441] [Medline: 32728474]
36. Harada N, Harada S, Ito J, Atsuta R, Hori S, Takahashi K. Mobile health app for Japanese adult patients with asthma: clinical observational study. J Med Internet Res 2020 Aug 14;22(8):e19006. [doi: 10.2196/19006] [Medline: 32795993]
37. Johnson EE, MacGeorge C, King KL, et al. Facilitators and barriers to implementation of school-based telehealth asthma care: program champion perspectives. Acad Pediatr 2021;21(7):1262-1272. [doi: 10.1016/j.acap.2021.04.025] [Medline: 33940203]
38. Kouri A, Yamada J, Sale JEM, Straus SE, Gupta S. Primary care pre-visit electronic patient questionnaire for asthma: uptake analysis and predictor modeling. J Med Internet Res 2020 Sep 18;22(9):e19358. [doi: 10.2196/19358] [Medline: 32945779]
39. Kuhn L, Reeves K, Taylor Y, et al. Planning for action: the impact of an asthma action plan decision support tool integrated into an electronic health record (EHR) at a large health care system. J Am Board Fam Med 2015;28(3):382-393. [doi: 10.3122/jabfm.2015.03.140248] [Medline: 25957371]
40. Lowery CL, Bronstein JM, Benton TL, Fletcher DA. Distributing medical expertise: the evolution and impact of telemedicine in Arkansas. Health Aff 2014 Feb;33(2):235-243. [doi: 10.1377/hlthaff.2013.1001]
41. MacGeorge CA, King K, Andrews AL, et al. School nurse perception of asthma care in school-based telehealth. J Asthma 2022 Jun;59(6):1248-1255. [doi: 10.1080/02770903.2021.1904978] [Medline: 33730979]
42. Radhakrishnan D, Higginson A, Thipse M, Tessier M, Radhakrishnan A. Optimizing pediatric asthma education using virtual platforms during the COVID-19 pandemic. Allergy Asthma Clin Immunol 2022 Aug 7;18(1):72. [doi: 10.1186/s13223-022-00713-y] [Medline: 35934694]
43. van den Wijngaart LS, Geense WW, Boehmer AL, et al. Barriers and facilitators when implementing web-based disease monitoring and management as a substitution for regular outpatient care in pediatric asthma: qualitative survey study. J Med Internet Res 2018 Oct 30;20(10):e284. [doi: 10.2196/jmir.9245] [Medline: 30377147]
44. Van Houten L, Deegan K, Siemer M, Walsh S. A telehealth initiative to decrease no-show rates in a pediatric asthma mobile clinic. J Pediatr Nurs 2021;59:143-150. [doi: 10.1016/j.pedn.2021.04.005] [Medline: 33887684]
45. Teresi RK, Hendricks AC, Moraveji N, Murray RK, Polsky M, Maselli DJ. Clinical interventions following escalations from a continuous respiratory monitoring service in patients with chronic obstructive pulmonary disease. Chronic Obstr Pulm Dis 2024 Nov 22;11(6):558-568. [doi: 10.15326/jcopdf.2023.0475] [Medline: 39357503]
46. Prior SJ, Bassett G, Van Beek E, et al. Evaluation of a telehealth paediatric asthma and allergy clinic patient follow-up during the COVID-19 pandemic. Children (Basel) 2024 Dec 11;11(12):1507. [doi: 10.3390/children11121507] [Medline: 39767936]
47. Allenbrand R, DiDonna A, Marshall J, Kennedy K. Establishing a virtual home assessment program: from concept to implementation as a result of the COVID-19 pandemic. Curr Allergy Asthma Rep 2023 Sep;23(9):531-540. [doi: 10.1007/s11882-023-01099-6] [Medline: 37405680]
48. Joo KR, Brewer TL, Matcham WA. Implementation of text message reminder system in the pediatric urgent care to improve compliance with asthma care follow-up. J Inform Nurs 2019;4(1):6-12 [FREE Full text]
49. Bédard A, Basagaña X, Anto JM, et al. Mobile technology offers novel insights into the control and treatment of allergic rhinitis: the MASK study. J Allergy Clin Immunol 2019 Jul;144(1):135-143. [doi: 10.1016/j.jaci.2019.01.053] [Medline: 30951790]
50. Alwakeel AJ, Sicondolfo A, Robitaille C, Bourbeau J, Saad N. The accessibility, feasibility, and safety of a standardized community-based tele-pulmonary rehab program for chronic obstructive pulmonary disease a 3-year real-world prospective study. Ann Am Thorac Soc 2022 Jan;19(1):39-47. [doi: 10.1513/AnnalsATS.202006-638OC] [Medline: 34170802]
51. Cheung NW, Crampton M, Nesire V, Hng TM, Chow CK, Western Sydney Integrated Care Program Investigators. Model for integrated care for chronic disease in the Australian context: Western Sydney Integrated Care Program. Aust Health Rev 2019 Oct;43(5):565-571. [doi: 10.1071/AH18152] [Medline: 30862349]
52. Chow JSF, Gonzalez-Arce V, Knight A, Kohler F. Retrospective analysis of telemonitoring in Wollondilly, Australia. JICA 2018 Apr 16;26(2):150-157. [doi: 10.1108/JICA-10-2017-0032]
53. Coffey JD, Christopherson LA, Williams RD, et al. Development and implementation of a nurse-based remote patient monitoring program for ambulatory disease management. Front Digit Health 2022;4:1052408. [doi: 10.3389/fdgth.2022.1052408] [Medline: 36588748]
54. Dyrvig AK, Gerke O, Kidholm K, Vondeling H. A cohort study following up on a randomised controlled trial of a telemedicine application in COPD patients. J Telemed Telecare 2015 Oct;21(7):377-384. [doi: 10.1177/1357633X15572202] [Medline: 25761469]
55. Eeles J, Ellison S, Jones C, Huntington C. Early experiences of telehealth monitoring for patients with COPD and implementation of person-centred care plans. Br J Community Nurs 2023 Apr 2;28(4):172-178. [doi: 10.12968/bjcn.2023.28.4.172] [Medline: 36989197]

XSL•FO
RenderX

56. Hofer F, Schreyögg J, Stargardt T. Effectiveness of a home telemonitoring program for patients with chronic obstructive pulmonary disease in Germany: evidence from the first three years. PLoS ONE 2022;17(5):e0267952. [doi: 10.1371/journal.pone.0267952] [Medline: 35551546]
57. Howard O, Thomas A, Henry H, Wallace J. Impact of a pharmacist-led outpatient telemedicine clinic on chronic obstructive pulmonary disease in a veteran population. J Am Pharm Assoc 2022 Nov;62(6):1919-1924. [doi: 10.1016/j.japh.2022.06.011]
58. Price-Haywood EG, Arnold C, Harden-Barrios J, Davis T. Stop the divide: facilitators and barriers to uptake of digital health interventions among socially disadvantaged populations. Ochsner J 2023;23(1):34-42. [doi: 10.31486/toj.22.0101] [Medline: 36936477]
59. Raeside R, Singleton AC, Todd A, et al. Lung support service: implementation of a nationwide text message support program for people with chronic respiratory disease during the COVID-19 pandemic. Int J Environ Res Public Health 2022 Dec 19;19(24):24. [doi: 10.3390/ijerph192417073] [Medline: 36554954]
60. Sibbald SL, Van Asseldonk R, Cao PL, Law B. Lessons learned from inadequate implementation planning of team-based chronic disease management: implementation evaluation. BMC Health Serv Res 2021 Feb 12;21(1):134. [doi: 10.1186/s12913-021-06100-4] [Medline: 33579271]
61. van Berkel C, Almond P, Hughes C, Smith M, Horsfield D, Duckworth H. Retrospective observational study of the impact on emergency admission of telehealth at scale delivered in community care in Liverpool, UK. BMJ Open 2019 Jul 31;9(7):e028981. [doi: 10.1136/bmjopen-2019-028981] [Medline: 31371293]
62. van Buul AR, Derksen C, Hoedemaker O, van Dijk O, Chavannes NH, Kasteleyn MJ. eHealth program to reduce hospitalizations due to acute exacerbation of chronic obstructive pulmonary disease: retrospective study. JMIR Form Res 2021 Mar 18;5(3):e24726. [doi: 10.2196/24726] [Medline: 33734091]
63. Cheung CY, Yam WS, Palmer MD, et al. Is videoconference pulmonary rehabilitation associated with improvements in knowledge in people living with COPD? A propensity-matched service-evaluation. Chron Respir Dis 2025;22:14799731241310895. [doi: 10.1177/14799731241310895] [Medline: 39840462]
64. Schmidt CW, Borgnakke K, Frølich A, Kayser L. Preferences, needs, and values of patients with chronic obstructive pulmonary disease attending a telehealth service: qualitative interview study. JMIR Hum Factors 2024 Jun 21;11:e53131. [doi: 10.2196/53131] [Medline: 38905629]
65. Taylor A, Cushing A, Dow M, et al. Long-term usage and improved clinical outcomes with adoption of a COPD digital support service: key findings from the RECEIVER trial. Int J Chron Obstruct Pulmon Dis 2023;18:1301-1318. [doi: 10.2147/COPD.S409116] [Medline: 37378275]
66. Segato F, Masella C. Telemedicine services: how to make them last over time. Health Policy Technol 2017 Sep;6(3):268-278. [doi: 10.1016/j.hlpt.2017.07.003]
67. Odeh B, Kayyali R, Nabhani-Gebara S, Philip N. Implementing a telehealth service: nurses' perceptions and experiences. Br J Nurs 2014;23(21):1133-1137. [doi: 10.12968/bjon.2014.23.21.1133] [Medline: 25426527]
68. Verweel L, Packham T, Goldstein R, Brooks D, MacKay C. Implementation of virtual pulmonary rehabilitation during the COVID-19 pandemic: experiences and perceptions of patients and healthcare providers. Respir Med 2024;225:107588. [doi: 10.1016/j.rmed.2024.107588] [Medline: 38460709]
69. Taylor J, Coates E, Wessels B, Mountain G, Hawley MS. Implementing solutions to improve and expand telehealth adoption: participatory action research in four community healthcare settings. BMC Health Serv Res 2015 Dec 1;15:529. [doi: 10.1186/s12913-015-1195-3] [Medline: 26626564]
70. Bond CS, Worswick L. Self management and telehealth: lessons learnt from the evaluation of a Dorset Telehealth Program. Patient 2015 Aug;8(4):311-316. [doi: 10.1007/s40271-014-0091-y] [Medline: 25315192]
71. Hunting G, Shahid N, Sahakyan Y, et al. A multi-level qualitative analysis of telehomecare in Ontario: challenges and opportunities. BMC Health Serv Res 2015 Dec 9;15:544. [doi: 10.1186/s12913-015-1196-2] [Medline: 26645639]
72. Sánchez FM, Bommatty MS, Haro MQ, et al. Telemonitoring in patients with COPD: a prospective study with results from the AIRE project. Respir Med 2025 Nov;248:108307. [doi: 10.1016/j.rmed.2025.108307] [Medline: 40849078]
73. Perkins R, Davis J, Bailey J, et al. Telehealth implementation in cystic fibrosis care during COVID-19: the clinician experience. Pediatr Pulmonol 2020;55(S2):S359-S359.
74. Somerville L, List R, Compton M, et al. 110 Real-world outcomes in cystic fibrosis telehealth clinical care during the COVID-19 pandemic. J Cyst Fibros 2021 Nov;20(Supplement 2):S55-S56. [doi: 10.1016/S1569-1993(21)01535-6]
75. Van Citters AD, Dieni O, Scalia P, et al. Barriers and facilitators to implementing telehealth services during the COVID-19 pandemic: a qualitative analysis of interviews with cystic fibrosis care team members. J Cyst Fibros 2021 Dec;20 Suppl 3(101128966):23-28. [doi: 10.1016/j.jcf.2021.09.004] [Medline: 34930537]
76. Mazzocato P, Luckhaus JL, Malmqvist Castillo M, et al. A patient-driven mobile health innovation in cystic fibrosis care: comparative cross-case study. J Med Internet Res 2024 Jul 31;26:e50527. [doi: 10.2196/50527] [Medline: 39083342]
77. Nazareth D, Sumner C, Daniels J, Mathews J, Hyde G, Walshaw M. Managing an adult CF population in the COVID-19 pandemic - 1 year on. QJM 2022 Aug 13;115(8):521-524. [doi: 10.1093/qjmed/hcab228] [Medline: 34432060]
78. Opipari-Arrigan L, Kouril M, Connor J, et al. Evaluating feasibility, acceptability, and effectiveness of a digital behavioral nutrition intervention in cystic fibrosis care. Clin Pract Pediatr Psychol 2025 Sep;13(3):269-279. [doi: 10.1037/cpp0000543]

XSL•FO
RenderX

79. Ahmed S, LePage K, Benc R, et al. Lessons learned from the implementation of a person-centred digital health platform in cancer care. Curr Oncol 2022 Sep 29;29(10):7171-7180. [doi: 10.3390/curroncol29100564] [Medline: 36290841]
80. Darcourt JG, Aparicio K, Dorsey PM, et al. Analysis of the implementation of telehealth visits for care of patients with cancer in Houston during the COVID-19 pandemic. JCO Oncol Pract 2021 Jan;17(1):e36-e43. [doi: 10.1200/OP.20.00572] [Medline: 33026951]
81. Demedts I, Himpe U, Bossuyt J, et al. Clinical implementation of value based healthcare: impact on outcomes for lung cancer patients. Lung Cancer 2021 Dec;162:90-95. [doi: 10.1016/j.lungcan.2021.10.010] [Medline: 34763159]
82. Ye X, Wei J, Li Z, et al. Design and implementation of a mobile system for lung cancer patient follow-up in China and initial report of the ongoing patient registry. Oncotarget 2017 Jan 17;8(3):5487-5497. [doi: 10.18632/oncotarget.13720]
83. Jaidar N, Thomas H, Nahm ES. Advancing telemedicine outreach within a comprehensive cancer center: lessons learned from rapid transformation of care practice and operational changes associated with the COVID-19 pandemic. J Oncol Navig Surviv 2021;12(4):111-117 [FREE Full text]
84. Lowery J, Fagerlin A, Larkin AR, Wiener RS, Skurla SE, Caverly TJ. Implementation of a web-based tool for shared decision-making in lung cancer screening: mixed methods quality improvement evaluation. JMIR Hum Factors 2022 Apr 1;9(2):e32399. [doi: 10.2196/32399] [Medline: 35363144]
85. Mackwood M, Butcher R, Vaclavik D, et al. Adoption of telemedicine in a rural US cancer center amid the COVID-19 pandemic: qualitative study. JMIR Cancer 2022 Aug 16;8(3):e33768. [doi: 10.2196/33768] [Medline: 35895904]
86. Turner K, Bobonis Babilonia M, Naso C, et al. Health care providers' and professionals' experiences with telehealth oncology implementation during the COVID-19 pandemic: a qualitative study. J Med Internet Res 2022 Jan 19;24(1):e29635. [doi: 10.2196/29635] [Medline: 34907900]
87. Kayser MZ, Valtin C, Greer M, Karow B, Fuge J, Gottlieb J. Video consultation during the COVID-19 pandemic: a single center's experience with lung transplant recipients. Telemed J E Health 2021 Jul;27(7):807-815. [doi: 10.1089/tmj.2020.0170] [Medline: 33054671]
88. Odisho AY, Liu AW, Maiorano AR, et al. Design and implementation of a digital health home spirometry intervention for remote monitoring of lung transplant function. J Heart Lung Transplant 2023 Jun;42(6):828-837. [doi: 10.1016/j.healun.2023.01.010] [Medline: 37031033]
89. Wickerson L, Helm D, Gottesman C, et al. Telerehabilitation for lung transplant candidates and recipients during the COVID-19 pandemic: program evaluation. JMIR Mhealth Uhealth 2021 Jun 17;9(6):e28708. [doi: 10.2196/28708] [Medline: 34048354]
90. Patel SR, Bakker JP, Stitt CJ, Hardy W, Aloia MS, Nouraie SM. Impact of an extended telemonitoring and coaching program on continuous positive airway pressure adherence. Ann Am Thorac Soc 2022 Dec;19(12):2070-2076. [doi: 10.1513/AnnalsATS.202204-310OC] [Medline: 35858100]
91. Kosky C, Madeira N, Boulton K, et al. Telemedicine compared to standard face-to-face care for continuous positive airway pressure treatment: real-world Australian experience. Sleep 2022 Oct 10;45(10):1-11. [doi: 10.1093/sleep/zsac157] [Medline: 35830509]
92. Hirshkowitz M, Sharafkhaneh A. A telemedicine program for diagnosis and management of sleep-disordered breathing: the fast-track for sleep apnea tele-sleep program. Semin Respir Crit Care Med 2014 Oct;35(5):560-570. [doi: 10.1055/s-0034-1390069] [Medline: 25353101]
93. Kerschberger B, Daka M, Shongwe B, et al. The introduction of video-enabled directly observed therapy (video-DOT) for patients with drug-resistant TB disease in Eswatini amid the COVID-19 pandemic – a retrospective cohort study. BMC Health Serv Res 2024 Jun 3;24(1):699. [doi: 10.1186/s12913-024-11151-4] [Medline: 38831356]
94. Hodges J, Zhdanova S, Koshkina O, et al. Implementation of a mobile health strategy to improve linkage to and engagement with HIV care for people living with HIV, tuberculosis, and substance use in Irkutsk, Siberia. AIDS Patient Care STDS 2021 Mar;35(3):84-91. [doi: 10.1089/apc.2020.0233] [Medline: 33538649]
95. Prabhu A, Agarwal U, Tripathy JP, et al. "99DOTS"techno-supervision for tuberculosis treatment - a boon or a bane? Exploring challenges in its implementation at a tertiary centre in Delhi, India. Indian J Tuberc 2020 Jan;67(1):46-53. [doi: 10.1016/j.ijtb.2019.08.010] [Medline: 32192617]
96. Thekkur P, Kumar AN, Chinnakali P, et al. Outcomes and implementation challenges of using daily treatment regimens with an innovative adherence support tool among HIV-infected tuberculosis patients in Karnataka, India: a mixed-methods study. Glob Health Action 2019;12(1):1568826. [doi: 10.1080/16549716.2019.1568826] [Medline: 30712507]
97. Thomas BE, Kumar JV, Onongaya C, et al. Explaining differences in the acceptability of 99DOTS, a cell phone-based strategy for monitoring adherence to tuberculosis medications: qualitative study of patients and health care providers. JMIR Mhealth Uhealth 2020 Jul 31;8(7):e16634. [doi: 10.2196/16634] [Medline: 32735220]
98. Bachina P, Lippincott CK, Perry A, et al. Programmatic adoption and implementation of video-observed therapy in Minnesota: prospective observational cohort study. JMIR Form Res 2022 Aug 5;6(8):e38247. [doi: 10.2196/38247] [Medline: 35834671]
99. Ortiz Laza N, Lopez Aranaga I, Toral Andres J, et al. Latent tuberculosis infection treatment completion in Biscay: differences between regimens and monitoring approaches. Front Med 2023;10. [doi: 10.3389/fmed.2023.1265057]

XSL•FO
RenderX

100. Wang N, Shewade HD, Thekkur P, et al. Electronic medication monitor for people with tuberculosis: implementation experience from thirty counties in China. PLoS ONE ;15(4):e0232337. [doi: 10.1371/journal.pone.0232337]
101. Kiwanuka N, Kityamuwesi A, Crowder R, et al. Implementation, feasibility, and acceptability of 99DOTS-based supervision of treatment for drug-susceptible TB in Uganda. PLOS Digit Health 2023 Jun;2(6):e0000138. [doi: 10.1371/journal.pdig.0000138] [Medline: 37390077]
102. Perry A, Chitnis A, Chin A, et al. Real-world implementation of video-observed therapy in an urban TB program in the United States. Int J Tuberc Lung Dis 2021 Aug 1;25(8):655-661. [doi: 10.5588/ijtld.21.0170] [Medline: 34330351]
103. Heo S, Jung W, Hwang SY, et al. The Impact of Introducing a Temporary Telemedicine Policy on Primary Care Visits: An Analysis of National Claims Data in South Korea. Telemed J E Health 2025 Jun;31(6):726-736. [doi: 10.1089/tmj.2024.0511] [Medline: 39982782]
104. Schnoor K, Versluis A, Bakema R, et al. A pharmacy-based eHealth intervention promoting correct use of medication in patients with asthma and COPD: nonrandomized pre-post study. J Med Internet Res 2022 Jun 8;24(6):e32396. [doi: 10.2196/32396] [Medline: 35675120]
105. Steinman L, Chadwick K, Chavez Santos E, et al. Remote evidence-based programs for health promotion to support older adults during the COVID-19 pandemic and beyond: mixed methods outcome evaluation. JMIR Aging 2024 Jun 13;7:e52069. [doi: 10.2196/52069] [Medline: 38869932]
106. Braun M, Schmidt O, Schultz T, Woehrle H, Sundrup MG, Schöbel C. Experiences with digital care of patients with chronic and acute lung diseases during the SARS-CoV-2 pandemic. Internist (Berl) 2022 Mar;63(3):255-265. [doi: 10.1007/s00108-022-01266-3] [Medline: 35181796]
107. Capusan KY, Fenster T. Patient satisfaction with telehealth during the COVID-19 pandemic in a pediatric pulmonary clinic. J Pediatr Health Care 2021;35(6):587-591. [doi: 10.1016/j.pedhc.2021.07.014] [Medline: 34417077]
108. Denker K, Dean M, Chapman D, Sump C. Using a centralized nursing team to implement multi-specialty pediatric remote patient monitoring programs. J Pediatr Nurs 2023;69:10-17. [doi: 10.1016/j.pedn.2022.12.028] [Medline: 36592607]
109. Vilendrer S, Patel B, Chadwick W, et al. Rapid deployment of inpatient telemedicine in response to COVID-19 across three health systems. J Am Med Inform Assoc 2020 Jul 1;27(7):1102-1109. [doi: 10.1093/jamia/ocaa077] [Medline: 32495830]
110. Zhang JR, Chua QHA, Shen XM, Viseikumaran JA, Teoh T, Tan NG. A systematic implementation of telemedicine in Singapore's COVID-19 community recovery facilities. Telemedicine and E-Health 2022 Nov;28(11):1587-1594. [doi: 10.1089/tmj.2021.0466]
111. Knox L, Dunning M, Davies CA, et al. Safety, feasibility, and effectiveness of virtual pulmonary rehabilitation in the real world. Int J Chron Obstruct Pulmon Dis 2019;14(101273481):775-780. [doi: 10.2147/COPD.S193827] [Medline: 31040656]
112. Mashru J, Kirlew M, Saginur R, Schreiber YS. Management of infectious diseases in remote northwestern Ontario with telemedicine videoconference consultations. J Telemed Telecare 2017 Jan;23(1):83-87. [doi: 10.1177/1357633X15625136] [Medline: 26748393]
113. Severini RDS, Oliveira PD, Couto TB, et al. Fast, cheap and feasible: Implementation of pediatric telemedicine in a public hospital during the Covid-19 pandemic. J Pediatr (Rio J) 2022;98(2):183-189. [doi: 10.1016/j.jped.2021.05.007] [Medline: 34181889]
114. Tan SC, Cross A. Virtual visiting in the intensive care unit: implementation, utilisation and provider acceptance in an Australian metropolitan health service. Aust Health Rev 2022 Jun;46(3):284-288. [doi: 10.1071/AH22032] [Medline: 35546423]
115. Xiao YZ, Chen XJ, Sun XL, et al. Effect of implementing an informatization case management model on the management of chronic respiratory diseases in a general hospital: retrospective controlled study. JMIR Med Inform 2024 Jun 19;12:e49978. [doi: 10.2196/49978] [Medline: 38904478]
116. Hui CY, Abdulla A, Ahmed Z, et al. Mapping national information and communication technology (ICT) infrastructure to the requirements of potential digital health interventions in low- and middle-income countries. J Glob Health 2022 Dec 29;12:04094. [doi: 10.7189/jogh.12.04094] [Medline: 36579436]
117. Jawad LA. Security and privacy in digital healthcare systems: challenges and mitigation strategies. Abhigyan 2024 Mar;42(1):23-31. [doi: 10.1177/09702385241233073]
118. McKenzie F, Hewitt C, Holmes HL, Johnson MR, Clarke T. Facilitating innovation in healthcare: insights from the frontline of adoption. BMJ Innov 2025 Jul;11(3):126-132. [doi: 10.1136/bmjinnov-2025-001407]
119. Pfitzer E, Bitomsky L, Nißen M, Kausch C, Kowatsch T. Success factors of growth-stage digital health companies: systematic literature review. J Med Internet Res 2024 Dec 11;26(1):e60473. [doi: 10.2196/60473] [Medline: 39661978]
120. Rouleau G, Wu K, Ramamoorthi K, et al. Mapping theories, models, and frameworks to evaluate digital health interventions: scoping review. J Med Internet Res 2024 Feb 5;26:e51098. [doi: 10.2196/51098] [Medline: 38315515]
121. Greenhalgh T. Role of routines in collaborative work in healthcare organisations. BMJ 2008 Nov 17;337(nov17 1):a2448. [doi: 10.1136/bmj.a2448] [Medline: 19015186]
122. Katonai G, Arvai N, Mesko B. AI and primary care: scoping review. J Med Internet Res 2025 Aug 15;27:e65950. [doi: 10.2196/65950] [Medline: 40836794]
123. Sylla B, Ismaila O, Diallo G. 25 Years of digital health toward universal health coverage in low- and middle-income countries: rapid systematic review. J Med Internet Res 2025 May 29;27:e59042. [doi: 10.2196/59042] [Medline: 40440696]

XSL•FO
RenderX

124. Wilson S, Tolley C, Mc Ardle R, et al. Recommendations to advance digital health equity: a systematic review of qualitative studies. NPJ Digit Med 2024 Jun 29;7(1):173. [doi: 10.1038/s41746-024-01177-7] [Medline: 38951666]
125. Beltrán V, Abarza L, Acuña-Mardones P, et al. Effectiveness of digital and remote health interventions in pediatric populations from underserved or rural areas: a systematic review of randomized controlled trials. Int J Telemed Appl 2026;2026(1):5586850. [doi: 10.1155/ijta/5586850] [Medline: 41705276]
126. Endalamaw A, Zewdie A, Wolka E, Assefa Y. A scoping review of digital health technologies in multimorbidity management: mechanisms, outcomes, challenges, and strategies. BMC Health Serv Res 2025 Mar 15;25(1):382. [doi: 10.1186/s12913-025-12548-5] [Medline: 40089752]
127. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119. [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
128. Donald N, Irukulla S. Greenhouse gas emission savings in relation to telemedicine and associated patient benefits: a systematic review. Telemed J E Health 2022 Apr 20;28(11):1555-1563. [doi: 10.1089/tmj.2022.0047] [Medline: 35446668]
129. Ueda D, Walston SL, Fujita S, et al. Climate change and artificial intelligence in healthcare: review and recommendations towards a sustainable future. Diagn Interv Imaging 2024 Nov;105(11):453-459. [doi: 10.1016/j.diii.2024.06.002] [Medline: 38918123]
130. Bouza L, Bugeau A, Lannelongue L. How to estimate carbon footprint when training deep learning models? A guide and review. Environ Res Commun 2023 Nov 1;5(11):115014. [doi: 10.1088/2515-7620/acf81b] [Medline: 38022395]
131. Kan Y. AI for environmental sustainability: advances, challenges, and future directions. IJAI4S 2025 Mar 17;1(1). [doi: 10.63619/ijais.v1i1.006]
132. Oliveira AP, Carraquico T, Martinez-Perez C. Beyond efficiency: a systematic review of energy consumption and carbon footprint across the AI lifecycle. Sustainability 2026 Jan 29;18(3):1359. [doi: 10.3390/su18031359]
133. Hasan SM, Islam T, Saifuzzaman M, Ahmed KR, Huang CH, Shahid AR. Carbon emission quantification of machine learning: a review. IEEE Trans Sustain Comput 2025 Jun 11;10(6):1085-1102. [doi: 10.1109/TSUSC.2025.3578834]
134. Onasanya AA, Haider Y, Peaston G, Ignatowicz A, Turner AM. Inhaler sustainability in asthma and COPD care: a systematic review. BMJ Open 2025 Jul 25;15(7):e098052. [doi: 10.1136/bmjopen-2024-098052] [Medline: 40713041]
135. Senek M, Drummond D, Pinnock H, et al. Impact of digital health on patient-provider relationships in respiratory secondary care based on qualitative and quantitative evidence: systematic review. J Med Internet Res 2025 May 30;27:e70970. [doi: 10.2196/70970] [Medline: 40446293]
136. Moy S, Irannejad M, Manning SJ, et al. Patient perspectives on the use of artificial intelligence in health care: a scoping review. J Patient Cent Res Rev 2024;11(1):51-62. [doi: 10.17294/2330-0698.2029] [Medline: 38596349]
137. Kim C, Gadgil SU, Lee SI. Transparency of medical artificial intelligence systems. Nat Rev Bioeng 2026 Jan;4(1):11-29. [doi: 10.1038/s44222-025-00363-w] [Medline: 42027791]
138. M Astobiza A, Alonso M, Ortega Lozano R. Trust and AI in healthcare: a systematic review. Monash Bioeth Rev 2025 Nov 11;11:1-21. [doi: 10.1007/s40592-025-00272-z] [Medline: 41217688]
139. Palmieri S, Robertson CT, Cohen IG. New guidance on responsible use of AI. JAMA 2026 Jan 20;335(3):207-208. [doi: 10.1001/jama.2025.23059] [Medline: 41335417]

Abbreviations

AI: artificial intelligence
CFIR: Consolidated Framework for Implementation Research
COPD: chronic obstructive pulmonary disease
CRC: Clinical Research Collaboration
EPIS: exploration, preparation, implementation, and sustainment
ERS: European Respiratory Society
EU: European Union
LMIC: low and middle income countries
NASSS: Nonadaptation, Abandonment, and Challenges to the Scale-up, Spread, and Sustainability of Health and Care Technologies
PDSA: plan-do-study-act
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for scoping reviews
RE-AIM: reach, effectiveness, adoption, implementation, and maintenance
TAM: Technology Acceptance Model
TMF: Theories, Models, and Framework
UTAUT: Unified Theory of Acceptance and Use of Technology
WHO: World Health Organization

XSL•FO
RenderX

Edited by S Brini; submitted 23.Nov.2025; peer-reviewed by S Mclean, T Vagg; revised version received 09.May.2026; accepted 11.May.2026; published 16.Jun.2026.

Please cite as:
Hui ICY, Condon K, Kolekar S, Roberts NJ, Sreter KB, Simons SO, Figueiredo C, McKeough Z, Salim H, Gawlik-Lipinski A, Gonsard A, Aral AÖ, Vanoverschelde A, Armstrong M, Kohlbrenner D, Paixão C, Stafler P, Papadopoulou E, Rabe APJ, Mohammad M, Bouloukaki I, Quach S, Kaltsakas G, Loveys K, Reier-Nilsen T, Sunjaya AP, Robinson P, Senek M, Chan AHY, Pinnock H
Implementing Digital Respiratory Technologies for People With Respiratory Conditions: Scoping Review
J Med Internet Res 2026;28:e88325
URL: https://www.jmir.org/2026/1/e88325
doi:10.2196/88325

©Io Chi-Yan Hui, Kathleena Condon, Shailesh Kolekar, Nicola J Roberts, Katherina Bernadette Sreter, Sami O Simons, Carlos Figueiredo, Zoe McKeough, Hani Salim, Aleksandra Gawlik-Lipinski, Apolline Gonsard, Ayşe Önal Aral, Anna Vanoverschelde, Matthew Armstrong, Dario Kohlbrenner, Cátia Paixão, Patrick Stafler, Efthymia Papadopoulou, Adrian Paul Jaravata Rabe, Milan Mohammad, Izolde Bouloukaki, Shirley Quach, Georgios Kaltsakas, Kate Loveys, Tonje Reier-Nilsen, Anthony Paulo Sunjaya, Paul Robinson, Michaela Senek, Amy Hai Yan Chan, Hilary Pinnock. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Embodied Intelligence Applications in Health Care Populations: Scoping Review

Wan Shu*, PhD; Ying Zhou*, MS; Erhong Sun*, PhD; Lei Deng, MS; Xuchun Ye, PhD

School of Nursing, Naval Medical University, No. 800, Xiangyin Road, Shanghai, China

*these authors contributed equally

Corresponding Author:
Xuchun Ye, PhD
School of Nursing, Naval Medical University, No. 800, Xiangyin Road, Shanghai, China

Abstract

Background: Embodied intelligence—artificial intelligence instantiated in physical or virtual bodies that can perceive, communicate, and interact with users and their environments—has been increasingly applied in health care. However, the evidence base remains fragmented because of inconsistent terminology, diverse embodiment forms, and limited synthesis of application domains, target populations, care settings, acceptability, and effectiveness. This fragmentation constrains conceptual clarity and translation into routine health care practice.

Objective: This scoping review aimed to systematically map the applications of embodied intelligence in health care by classifying embodiment forms, identifying major functional domains, describing target populations and implementation settings, and synthesizing the available evidence on acceptability and effectiveness.

Methods: This scoping review followed the Arksey and O'Malley framework, with enhancements by Levac et al, and was reported in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) guidelines. Seven electronic databases were searched from database inception to December 2025, supplemented by gray literature searches and backward citation screening. Eligible studies were primary empirical studies published in English or Chinese that examined embodied intelligence in health care contexts. Two reviewers independently screened records and charted data using a pilot-tested standardized form. Descriptive statistics and thematic synthesis were applied. No formal critical appraisal was conducted because the aim was to map the breadth and characteristics of the evidence base.

Results: A total of 83 studies were included. Five embodiment forms were identified: virtual humanoid agents (32/83, 38.6%), physical humanoid robots (32/83, 38.6%), virtual animal-shaped agents (1/83, 1.2%), physical animal robots (13/83, 15.7%), and mechanical robots (5/83, 6%). Applications clustered into 3 functional domains: health management and health education (40/83, 48.2%), mental health promotion (37/83, 44.6%), and physiological health promotion (6/83, 7.2%). Older adults were the most frequently targeted population (45/83, 54.3%). Interventions were mainly implemented in home settings, care homes, laboratories, and hospitals. Twenty-two randomized controlled trials reported generally beneficial effects on health behaviors, mental health outcomes, or cognitive function, although outcome measures were heterogeneous. Twelve studies examined acceptability and generally reported favorable user acceptance.

Conclusions: This scoping review provides the first comprehensive synthesis of embodied intelligence in health care using a unified classification of forms, functional domains, populations, and application settings. The findings indicate that embodied intelligence is most mature in "health management and health education" and "mental health promotion," with increasing real-world deployment in home and care home settings. By consolidating fragmented evidence and standardizing terminology, this review offers a practical foundation for clinicians, nurses, and policymakers to support the implementation of embodied intelligence in routine health care. Evidence is limited by heterogeneous outcome measures, many lab-based evaluations, and the absence of formal quality appraisal, underscoring the need for standardized outcome measures, rigorous randomized controlled trials, and longitudinal evaluations to enable scalable and ethically grounded real-world adoption.

(J Med Internet Res 2026;28:e83871) doi:10.2196/83871

KEYWORDS

embodied intelligence; artificial intelligence; health care; scoping review; robotics

XSL•FO
RenderX

Introduction

What Is Already Known

Embodied intelligence has been piloted in health care, but prior research remains fragmented and lacks a systematic classification of embodiment forms, functional roles, and application contexts. Small-scale studies suggest that embodied intelligence may support rehabilitation and provide companionship; however, there is still no comprehensive map of where, how, and for whom these systems have been implemented.

What This Paper Adds

We systematically classified embodied intelligence in health care into 5 embodiment forms and 3 core functional domains. By synthesizing evidence from 22 randomized controlled trials (RCTs) and 12 mixed methods acceptability studies, we summarized the reported effectiveness of embodied intelligence in health management and health education and mental health promotion, as well as patterns of user acceptability across settings. These findings provide an evidence base to inform future standardized evaluation and implementation of embodied intelligence in health care, with particular relevance to nursing practice.

Rationale

As global populations age and chronic diseases become increasingly prevalent, health care systems face unprecedented challenges [1,2]. Conventional care models often struggle to address diverse and individualized patient needs, prompting growing interest in artificial intelligence (AI)–enabled solutions in health care [3]. In recent years, AI has been widely applied to tasks such as diagnosis, risk prediction, and clinical decision support, contributing to improvements in efficiency and accuracy across health care systems [4,5].

Embodied intelligence represents a distinct branch of AI that integrates computational intelligence with physical or virtual embodied entities, enabling direct interaction with users and their environments [6]. In this review, embodied intelligence refers to AI systems instantiated in physical or virtual forms that interact with users through perception, communication, and action, including virtual humanoid agents, physical humanoid robots, animal-shaped agents, and mechanical robots applied in health care contexts.

In this context, embodiment refers to the integration of intelligence with a physical or virtual form that enables situated perception, action, and interaction within an environment [7]. In this review, embodied intelligence is used as an umbrella term for embodied AI systems that exhibit interactive and adaptive behaviors through their physical or virtual embodiment. Embodied robots represent a subset of embodied intelligence, referring specifically to physically instantiated robotic systems with intelligent capabilities, whereas virtual embodied agents (eg, virtual humanoid or animal-shaped agents) represent nonphysical yet embodied forms of intelligent systems. To avoid ambiguity, this review adopts embodied intelligence as the overarching term and classifies systems based on embodiment form rather than using related terms interchangeably.

Although AI has been increasingly integrated into health care, many widely deployed AI applications remain disembodied, functioning primarily as background analytical tools such as diagnostic algorithms, risk prediction models, or data-driven decision support systems [8]. These systems typically operate at the level of information processing and decision-making, with limited direct interaction with patients or clinicians.

In contrast, embodied intelligence emphasizes embodiment as a core design principle, integrating intelligence not only algorithmically but also through physical or virtual presence, environmental context, and interactive engagement. By enabling multimodal interaction—such as dialogue, movement, facial expressions, or touch—embodied intelligence can extend AI from passive analytical support toward more active participation in care processes. This interactive and socially situated nature makes embodied intelligence particularly relevant for health care domains that rely on sustained engagement, trust, and relational interaction, including health education, chronic disease management, mental health support, and nursing care [6]. Despite growing interest in embodied intelligence for health care applications, existing studies remain fragmented with respect to embodiment types, functional capacities, and the evaluation of effectiveness and acceptability.

Several recent reviews have examined AI and robotics in health care from specific perspectives. For example, Fiske et al [9] explored the ethical and societal implications of embodied AI in mental health–related fields such as psychiatry, psychology, and psychotherapy. Huang et al [10] conducted a systematic review of intelligent physical robots in health care, focusing on antecedents, consequences, and organizational factors associated with robot implementation. Kiuchi et al [11] provided a meta-review emphasizing psychological insights into embodied conversational agents, chatbots, and social assistive robots, with particular attention to interaction design and user experience.

While these reviews offer valuable insights, they primarily concentrate on specific embodiment forms, disciplinary viewpoints, or outcome domains. In contrast, this scoping review adopts a broader and integrative perspective by examining multiple types of embodied intelligence—both physical and virtual—across diverse health care contexts. By systematically mapping embodiment forms, functional roles, application settings, and evidence of acceptability and effectiveness, this review aims to clarify the current landscape of embodied intelligence in health care and identify priority research gaps to inform future research and practice.

Despite the growing body of research on embodied intelligence in health care, there remains a lack of integrative evidence that systematically examines how different forms of embodied intelligence are applied across health care contexts, what functional roles they serve, and how acceptability and effectiveness are evaluated [12]. Existing studies are often fragmented by technology type, application domain, or outcome focus, making it difficult to obtain a comprehensive understanding of the field.

XSL•FO
RenderX

Objectives

To address these gaps, this scoping review synthesizes applications of embodied intelligence in health care by mapping embodiment forms, functional domains, application settings, and evidence of acceptability and effectiveness. By providing an integrated overview of current practices and identifying priority research gaps, this review aims to inform future research and support the responsible integration of embodied intelligence into health care and nursing practice.

Accordingly, this scoping review addresses the following research questions:

1. What types of embodied intelligence have been applied in health care, and how are they classified based on their embodiment forms?
2. What functional roles do different types of embodied intelligence serve in health care applications?
3. In which health care contexts and settings have embodied intelligence systems been implemented?
4. What evidence has been reported regarding the acceptability and effectiveness of embodied intelligence in health care?

Methods

Protocol and Registration

A scoping review was conducted according to the framework described by Arksey and O'Malley [13] with an extended version by Levac et al [14]. This review is reported based on the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines (Checklist 1) [15]. The protocol was registered on the Open Science Framework (OSF) [16]. The review methods were implemented before OSF registration and were subsequently documented in the registered protocol. The review title and wording were refined during manuscript development to better reflect the scope of the evidence map; however, no substantive changes were made to the eligibility criteria, information sources, screening procedures, or synthesis approach.

Eligibility Criteria

The eligibility criteria were developed using the PCC (population, concept, and context) framework [17] to ensure alignment with the review objectives. The population of interest included any human participants, such as patients or healthy individuals, who interacted with or were affected by embodied intelligence systems in a health care context. Specific population characteristics, including age and health status, were not used as eligibility criteria but were charted during data extraction.

The core concept was embodied intelligence, defined in this review as AI systems integrated with a physical or virtual form capable of interacting with users and/or the environment through situated perception, communication, or action. Eligible systems included, for example, virtual embodied conversational agents, humanoid robots, animal-shaped robots, and mechanical robots. Embodied intelligent robotics was considered a subset of embodied intelligence referring specifically to physically instantiated robotic systems with intelligent and interactive capabilities; the two terms were not used interchangeably.

The context included any health care–related setting, such as hospitals, clinics, long-term care facilities, residential settings, research laboratories, and community-based programs.

Primary empirical studies of any design were eligible for inclusion, including RCTs, pilot studies, quasi-experiments, observational studies, qualitative studies, mixed methods studies, case reports, and crossover trials. Studies published in English or Chinese were included. Studies were excluded if they were nonprimary research, such as reviews, editorials, letters, or study protocols, or if they were not available as full peer-reviewed journal articles with sufficient methodological detail for screening and data charting, such as conference abstracts without full papers, dissertations, theses, or other gray literature items.

For study selection, embodiment was operationally defined as the presence of a physical or virtual form that enabled situated interaction with users and/or the environment, such as visual representation, movement or gesture, facial expressions, voice interaction, or tactile interaction.

Information Sources

A comprehensive search was conducted in 7 electronic databases: PubMed (National Library of Medicine interface), Web of Science Core Collection (Web of Science platform), Cochrane Library (Wiley platform), MEDLINE (via Ovid), APA PsycNet (APA PsycNet platform), CQVIP (CQVIP platform), and China National Knowledge Infrastructure (CNKI) (CNKI platform). No separate search update was undertaken beyond the final searches completed in December 2025. All databases were searched from database inception to December 2025. No simultaneous multidatabase searching on a single platform was undertaken.

All retrieved records were imported into EndNote (Clarivate) for deduplication. Duplicate records were removed using automated matching followed by manual verification. The most recent search was conducted in December 2025.

Search Strategy

A 3-step search strategy was applied to identify relevant studies from database inception to December 2025, with search reporting guided by the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) guideline (Checklist 2) [18]. First, 7 electronic databases were searched. The search strategy was adapted for each database and is presented in full (line by line) in Multimedia Appendix 1.

Second, purposeful searching of online resources for gray literature was conducted using OpenGrey and Google Scholar. Third, backward citation searching was performed by screening the reference lists of included studies and relevant reviews to identify additional potentially eligible records.

No date limits were applied at the database search stage, and all databases were searched from inception to December 2025. Language restrictions were limited to English and Chinese to match the review scope and the language capacity of the review

XSL•FO
RenderX

team. Eligibility restrictions to primary research studies reporting original empirical data were implemented during study selection. No published search filters were used. The search strategies were developed for this review and were not directly adapted from previous literature reviews. No study registries were searched as part of this review. No additional studies or data were sought by contacting authors, experts, manufacturers, or other stakeholders. The PRESS (Peer Review of Electronic Search Strategies) guideline was not conducted and is acknowledged as a limitation.

Selection of Sources of Evidence

Study selection was conducted in two stages by two independent reviewers. In the first stage, titles and abstracts were screened against the eligibility criteria. In the second stage, the full texts of potentially eligible studies were retrieved and assessed for final inclusion.

Any discrepancies between reviewers at either stage were resolved through discussion and consensus. All screening decisions were documented using a standardized screening protocol to ensure consistency and transparency throughout the selection process.

Data Charting Process

Data charting was conducted using a standardized data extraction form that was developed and pilot-tested prior to formal extraction. Two reviewers independently extracted data from all included studies, and any discrepancies were resolved through discussion until consensus was reached.

To support consistency and traceability, all coding decisions and charting procedures were documented in a shared Microsoft Excel file. Chinese-language information extracted from included studies was translated into English by one reviewer and cross-checked by another reviewer before synthesis.

Data Items

The data charted from the included studies included the authors, year of publication, country or region where the study was conducted, study design, study population or target group, type of embodied intelligence, functional role, application setting, and key outcomes.

These variables were selected to capture both the technological characteristics of embodied intelligence, such as embodiment form and role, and the application-oriented characteristics relevant to interpretation, such as population, setting, and outcomes. For the purposes of charting and synthesis, embodied intelligence was classified along 2 dimensions: material instantiation, defined as virtual versus physical, and morphology, defined as humanoid, animal-shaped, or mechanical. Functional domains were classified inductively according to the primary aim reported in each study.

When studies reported overlapping objectives, classification was based on the dominant aim described by the authors and the primary outcomes emphasized in the study. These operational decisions were made to facilitate consistent comparison across heterogeneous sources of evidence.

Critical Appraisal of Individual Sources of Evidence

A formal critical appraisal of individual sources of evidence was not conducted. This review was designed as a scoping review to map the breadth, characteristics, and application patterns of the available evidence rather than to estimate pooled effects or exclude studies based on methodological quality.

Given the heterogeneity of study designs and the exploratory nature of the field, all eligible primary studies were retained for charting and narrative synthesis. The absence of formal critical appraisal is acknowledged as a limitation and is considered when interpreting the findings, particularly those related to acceptability and effectiveness.

Synthesis of Results

To facilitate the synthesis and analysis of the included studies, a standardized data extraction form was developed, and the application of embodied intelligence and the roles played in health care were then grouped thematically through an inductive approach. This process involved two independent reviewers (WS and YZ) who independently read through the full text of each study to assign initial codes to relevant segments describing the technology's use, function, or context. These initial codes were subsequently organized into broader categories or themes; for example, codes related to health education, disease self-management, and medication adherence were grouped under the overarching theme of "health management and education," while codes concerning emotional support, cognitive stimulation, and reduction of loneliness were clustered under "mental health promotion." The reviewers met regularly to discuss their coding and theme development, resolving any discrepancies through consensus. All coding decisions and theme definitions were documented in a shared coding sheet (Microsoft Excel) to support consistency and traceability; no qualitative analysis software was used. Additionally, countries, participants, settings, and study designs were mapped, and trends in publication numbers were analyzed to identify patterns and gaps in the existing literature. Mapping outputs (including evidence distribution across embodiment forms, functional domains, and settings) were generated through descriptive frequency summaries and cross-tabulation.

Results

Selection of Sources of Evidence

The detailed study selection process is illustrated in Figure 1. A total of 3797 records were initially identified through electronic database searches, and an additional 41 records were identified via other methods, including gray literature and citation searching, bringing the total to 3838 records. After removing 745 duplicates, 3093 records remained for screening. Of these, 2743 records (2740 from databases and 3 from other methods) were excluded during title and abstract screening, leaving 350 records for full-text assessment. At the title/abstract stage, records were excluded if they clearly did not meet the PCC-based eligibility criteria (eg, not involving embodied intelligence as defined in this review, not in a health care context, not primary research, not involving human participants, or not published in English/Chinese). Following a

XSL•FO
RenderX

comprehensive evaluation of the full texts, 267 records were excluded (235 from databases and 32 from other methods) for reasons such as not focusing on embodied intelligence or health care, inappropriate study designs, or lack of available full text. This resulted in a final total of 83 studies included in this scoping review.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram illustrating the study selection process for a scoping review on embodied intelligence in health care. CNKI: China National Knowledge Infrastructure; CQVIP: VIP Chinese Journal Database.

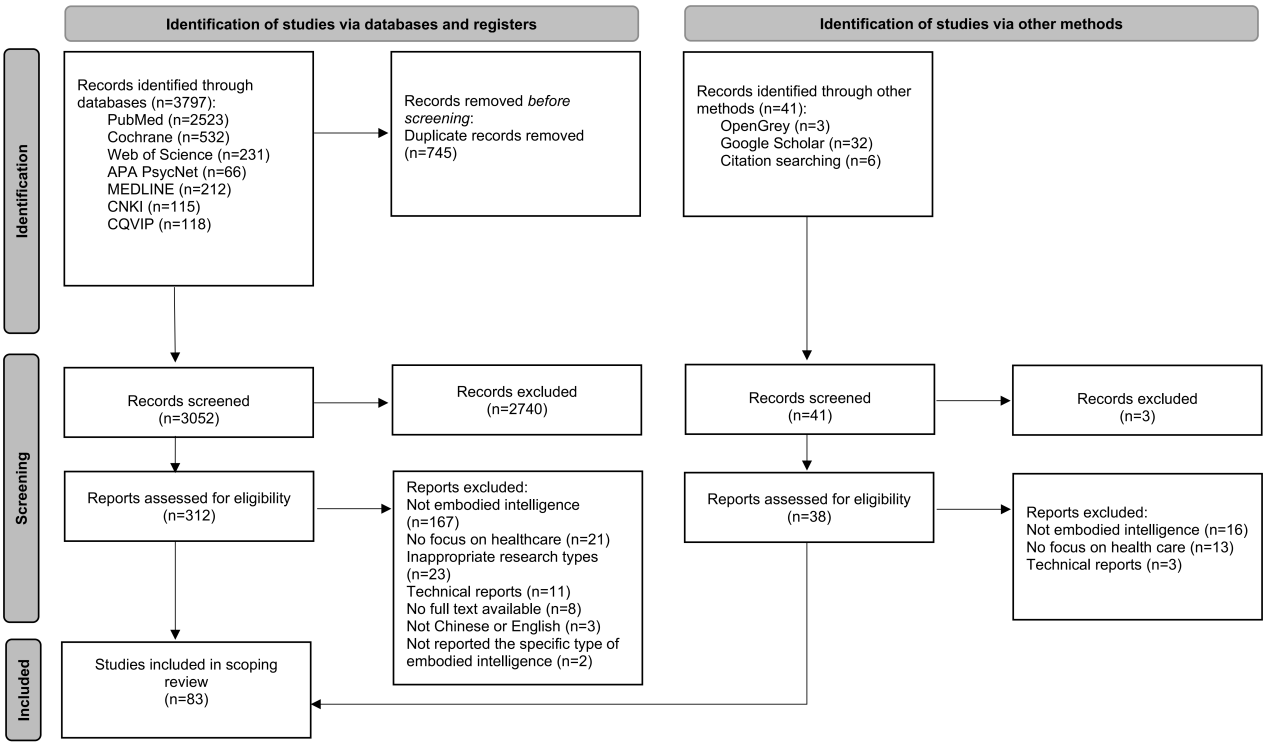

*Consider, if feasible to do so, reporting the number of records identified from each database or register searched (rather than the total number across all databases/registers).
**If automation tools were used, indicate how many records were excluded by a human and how many were excluded by automation tools.

From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71.

Characteristics of Sources of Evidence

There were 83 eligible studies published in English or Chinese according to the inclusion criteria. They were conducted in 20 countries, with more than half of the studies originating from Europe and the United States (Figure 2A and Table 1). Overall, the number of publications increased over time, peaking in 2021, followed by a marked decline in 2022 (Figure 2B). Study designs were diverse, with RCTs accounting for the largest proportion (nearly one-third) of all studies (Table 2). The characteristics of the included studies are presented in Multimedia Appendix 2 [18-99].

XSL•FO
RenderX

Figure 2. Characteristics of included studies in the scoping review. Panel A shows the geographic distribution of the 83 included studies by study setting (country/region where the study was conducted). Panel B presents the trends in publication volume by year and publication period.

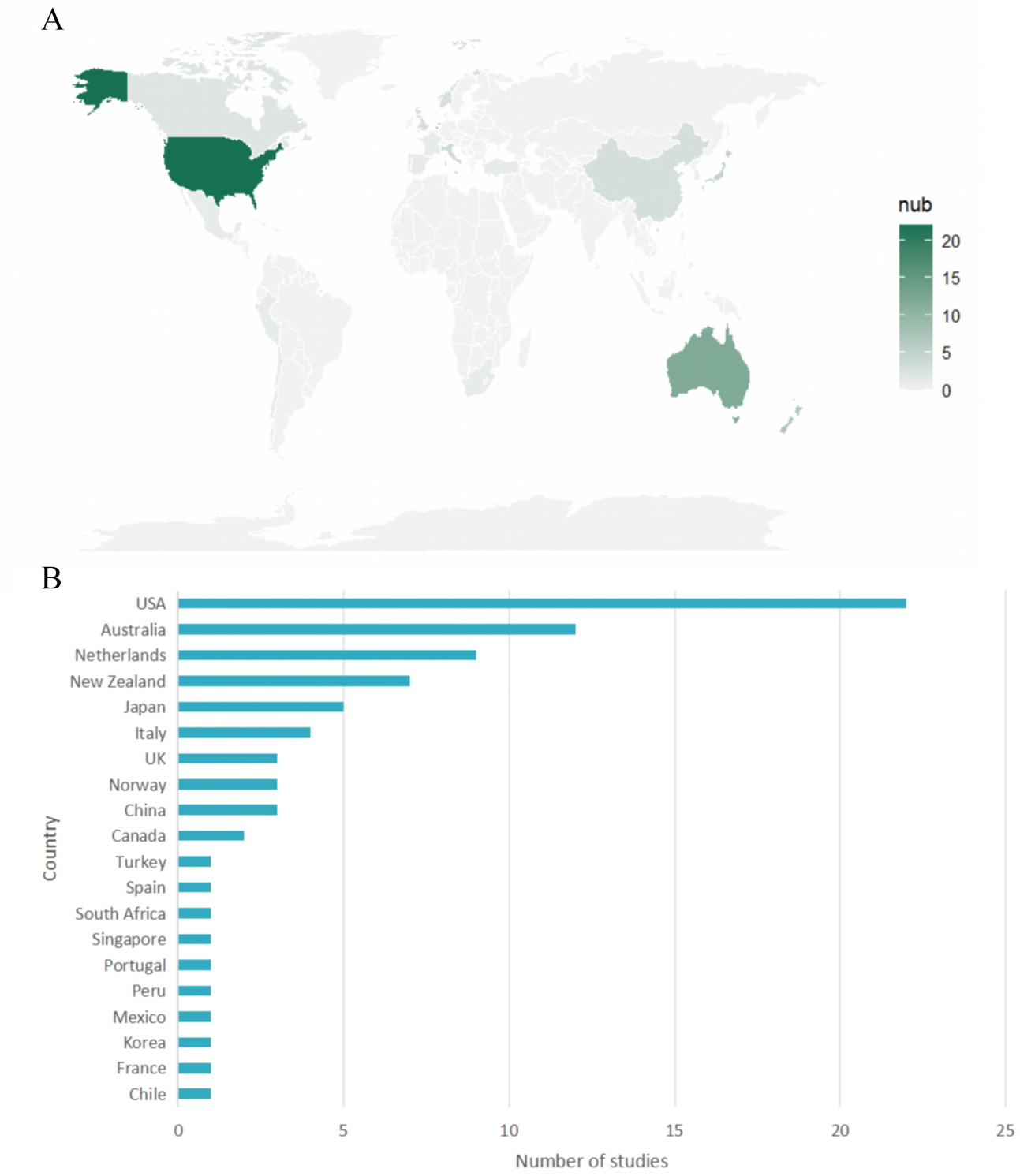

XSL•FO
RenderX

Table . Geographic distribution of the 83 included studies by study setting.

Country	Included studies, n
United States	22
Australia	12
Netherlands	9
New Zealand	7
Japan	5
China	4
Italy	4
Norway	3
United Kingdom	3
Canada	2
Spain	2
Chile	1
France	1
Korea	1
Mexico	1
Peru	1
Portugal	1
Singapore	1
South Africa	1
South Korea	1
Turkey	1

Table . The distribution of study designs.

Study design	Count, n
Randomized controlled trials	22
Pilot study	18
Quasi-experiments	15
Mixed methods studies	11
Qualitative study	7
Case studies	3
Crossover trials	2
3×3 factorial design	1
Descriptive observational pilot study	1
Observational cohort study	1
Pre-post study	1
Quasi-experimental research design	1

A review of the literature indicates that embodied intelligence can be categorized into virtual embodied intelligence and physical embodied intelligence based on whether the system has a physical form. Virtual embodied intelligence is predominantly composed of human-shaped embodied conversational agents (32/33, 96.9%), whereas within physical embodied intelligence, human-shaped robots account for 64% (32/50), followed by animal-shaped robots (13/50, 26%) and mechanical robots (5/50, 10%) (Figure 3A). Physical embodied intelligence represents the majority of health care applications (50/83, 66.2%). As shown in the publication trend graph (Figure 3B), over the past 5 years, publications on physical embodied intelligence have generally exceeded those on virtual embodied intelligence.

XSL•FO
RenderX

Figure 3. Categories and publication trends of embodied intelligence applied in health care. Panel A displays the distribution of embodied intelligence categories (virtual vs physical, including humanoid, animal-shaped, and mechanical forms) across the 83 included studies. Panel B illustrates temporal trends in publications comparing virtual embodied intelligence and physical embodied intelligence. The data derive from studies conducted in multiple countries and health care settings between database inception and December 2025.

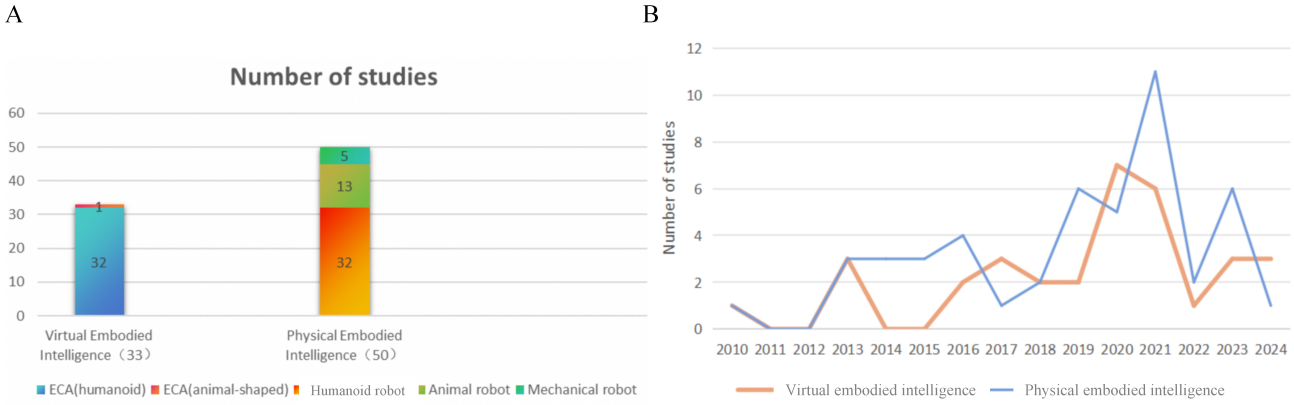

This review indicates that, although embodied intelligence has been applied across a range of health care scenarios, it is predominantly studied in older adults, who account for 54.4% (46/83) of the target populations (Figure 4A). Among these studies involving older adults, 69.6% (32/46) focused on healthy older adults and 30.4% (14/46) on individuals with cognitive impairment. This pattern likely reflects the increasing demand for elderly care services associated with rapid population aging. Implementation settings spanned 4 categories (Figure 4B): domestic residences (22/83, 26.5%), long-term care facilities (27/83, 32.5%), laboratory-based evaluation contexts (22/83, 26.5%), and hospital units (12/83, 14.5%). In this review, "laboratory" refers to controlled research environments (eg, university or clinical research labs) used primarily for prototyping and evaluation (eg, usability, interaction quality, feasibility testing, or preliminary outcomes), rather than real-world care delivery settings. While 73.5% (61/83) of studies were conducted in real-world care settings, over a quarter (22/83, 26.5%) were confined to controlled laboratory environments, which may limit ecological validity and the generalizability of findings to routine practice.

Figure 4. Study populations and application settings of embodied intelligence in health care. Panel A shows the distribution of target populations across included studies (eg, older adults, patients with chronic conditions, and children). Panel B summarizes the health care application settings, including home, care home, hospital, and laboratory environments. The figure reflects findings from 83 studies published between database inception and December 2025.

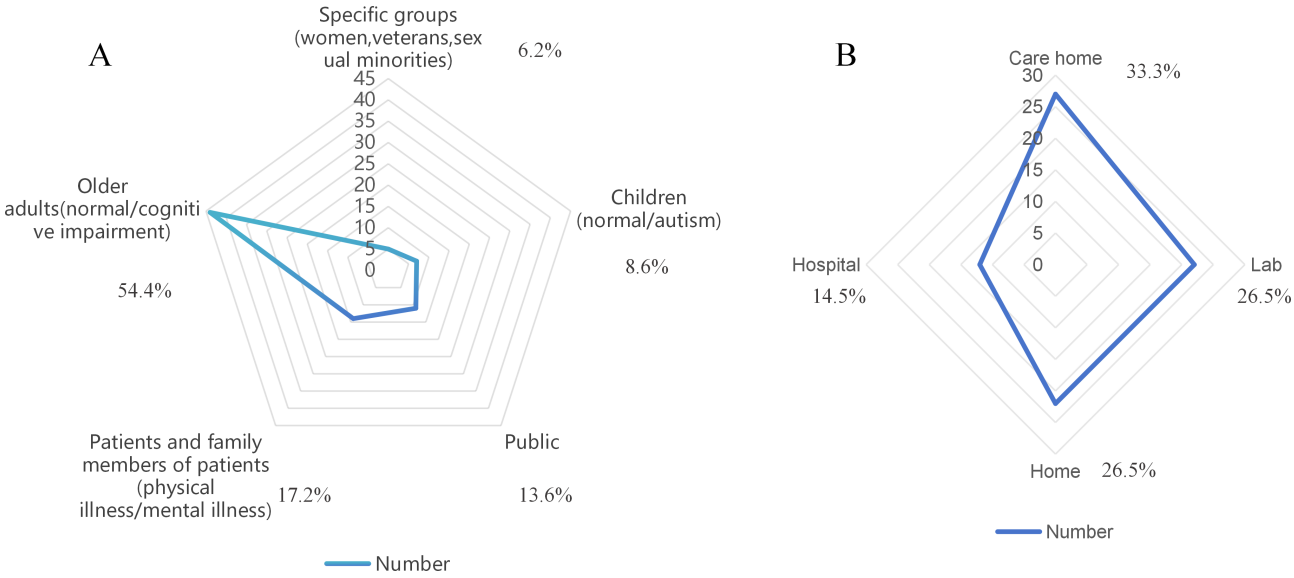

Critical Appraisal Within Sources of Evidence

No critical appraisal within sources of evidence is reported because a formal critical appraisal of included studies was not undertaken in this scoping review. Consistent with the review aim of mapping the breadth and characteristics of the evidence base, all eligible studies were retained for data charting and synthesis regardless of methodological quality.

Results of Individual Sources of Evidence

The Three Major Roles of Embodied Intelligence Applied in Health Care

Analysis of the 83 included studies indicates that embodied intelligence in health care serves 3 principal functional domains: health management and health education, mental health promotion, and physiological health promotion. In this review, the distinction between "health management" and "health

XSL•FO
RenderX

promotion" is based on the primary intent of the intervention. Health management denotes structured, ongoing support aimed at controlling a specific disease, condition, or modifiable risk factor, typically through monitoring, symptom tracking, medication or lifestyle adherence support, and self-management guidance. Health education denotes interventions primarily intended to improve health-related knowledge, skills, and decision-making (eg, patient education or training), and was grouped with health management when education served a disease- or risk-focused management goal. In contrast, health promotion denotes interventions primarily intended to enhance well-being, function, or health capacity (eg, emotional well-being, cognitive engagement, physical function, or rehabilitation outcomes) without being anchored to the management of a single disease alone.

These categories were derived inductively during thematic analysis and were not predefined. Specifically, interventions centered on disease self-management, adherence, and condition-focused education were grouped under "health management and health education"; those centered on emotional support, anxiety reduction, or cognitive engagement were grouped under "mental health promotion"; and those centered on improving physical activity, functional performance, or rehabilitation outcomes were grouped under "physiological health promotion." When studies reported overlapping objectives (eg, monitoring functions embedded within a rehabilitation program), classification was determined by the dominant aim described by the authors and the primary outcomes emphasized.

As summarized in Table 3, these functions were operationalized through 46 distinct embodied intelligence devices. Health management and health education (40/83, 48.2%) and mental health promotion (37/83, 44.6%) accounted for most applications. Most studies reported beneficial outcomes within their respective domains; 2 experimental studies did not report improvements and are marked (※) in Table 3.

Table . Roles of embodied intelligence in health care.

Types of embodied intelligence	Role		
	Health management and health education	Mental health promotion	Rehabilitation training
Virtual embodied intelligence (animal-shaped; n=1)		✓	
Virtual embodied intelligence (humanoid-shaped; n=25)	✓※[a]	✓	
Humanoid robot (n=12)	✓	✓	✓
Animal robot (n=2)	✓	✓※[a]	
Mechanical robot (n=6)	✓	✓	

[a]※ indicates studies that did not report beneficial effects consistent with the stated role.

Health Management and Health Education

Among the 83 included studies, 40 examined the application of embodied intelligence in health management and health education.

First, in 30 studies, embodied intelligence served health management functions, covering both professional health management and self-management. Professional health management was mainly applied in medical services, experimental research, and other specialized scenarios. Specifically, these systems supported health data collection, disease screening, and assessments such as attention level evaluation in populations including people with diabetes or chronic obstructive pulmonary disease, pregnant women, and older adults. Based on these assessments, the systems delivered targeted and personalized management advice [18-41]. Reported roles included assisting clinicians with documentation, reducing clinical risks, and improving patients' adherence to recommendations. Notably, ter Stal et al [40] quasi-experimental study reported low adherence, which was attributed to insufficient personalization and limited information accuracy.

In contrast, self-management functions were mainly implemented in home or nursing home settings to support recipients (eg, older adults or people with diabetes) with exercise guidance, smoking cessation support, and disease self-management outside clinical encounters [42-47]. These applications primarily aimed to improve individuals' quality of life.

Furthermore, 10 studies focused on the application of embodied intelligence in health education. Researchers integrated educational content into embodied intelligence systems to deliver health education to diverse groups, including individuals with chronic conditions, people with alcohol use problems, older adults, and the general public. The findings suggest that embodied intelligence may improve participants' health literacy. Compared with traditional education delivered by health professionals, embodied intelligence offered greater flexibility and convenience, mainly through reduced time and location constraints and potentially lower learning burden. Consequently, participants reported higher satisfaction and better acceptance of education delivered via embodied intelligence systems [48-57].

Mental Health Promotion

There were 37 studies focusing on the application of embodied intelligence in mental health promotion, addressing 3 main tasks: psychological intervention, cognitive enhancement, and emotional companionship.

XSL•FO
RenderX

Psychological Intervention

In recent years, embodied intelligence has attracted increasing attention as a delivery modality for psychological interventions. Of the 37 studies, 6 integrated cognitive behavioral therapy and/or mindfulness-based practices into embodied intelligence systems. Findings suggest that these interventions may enhance subjective well-being and reduce distress among university students and members of the general public with psychological difficulties [12,58-62].

Cognitive Enhancement

Embodied intelligence–supported cognitive interventions have also been explored as a nonpharmacological approach. Among the included studies, 3 focused on using embodied intelligence to deliver cognitive training for older adults [63-65]. These studies reported that cognitive stimulation games and other embodied intelligence–based training may improve memory and executive function, while also promoting engagement in activities and social interaction. For example, Park et al [66] implemented a 6-week embodied intelligence–assisted multimodal cognitive training program and observed significant improvements in indicators such as memory and executive function, supporting the potential of embodied intelligence to enhance cognitive functioning in older adults.

Emotional Companionship

A total of 28 studies examined embodied intelligence for emotional companionship. Of these, 22 (78.6%) targeted older adults. The findings suggest that embodied intelligence can provide emotional support through multimodal interactions (eg, speech, facial expressions, and movement) and that older adults are generally receptive to embodied intelligence in a companion role. Reported benefits can be summarized at 3 levels.

First, at the psychological level, emotional bonds formed between embodied intelligence and older adults were associated with reduced depressive symptoms, loneliness, and psychological distress, and may even modulate perceived pain [24,58,66-76].

Second, at the social level, embodied intelligence was reported to facilitate social participation and improve social integration among older adults [77-83].

Third, at the family level, embodied intelligence may act as a shared social medium that supports intergenerational communication and strengthens family interactions through shared use [84,85]. Overall, embodied intelligence may elicit positive emotional experiences in older adults, which can motivate them to share these experiences with others and potentially increase opportunities for social engagement. As a novel technology, embodied intelligence may also function as a social catalyst and conversational topic in peer interactions.

Beyond older adult populations, 6 studies involved children or individuals with depression. Four studies focused on pediatric populations; 3 reported that embodied intelligence could reduce anxiety and enhance social interaction by engaging children (including those with autism) through voice-based interaction [86-88]. However, 1 study reported different findings. Kitt et al [89] quasi-experimental evaluation of an animal-like robot (Paro) in pediatric health care settings found no stress-reducing effect. This discrepancy may reflect differences in role framing and study design: in Crossman et al [87], the robot provided comforting interaction after children completed a stressful task, whereas in Kitt et al [89], the robot was involved during the stressful task itself, which may have been perceived as disruptive rather than supportive.

Two studies involving individuals with depression suggested that embodied intelligence, through verbal interaction tailored to users' emotional states and needs, may provide emotional companionship and reduce perceived social stigma and internalized self-stigma [89,90].

Physiological Health Promotion

Embodied intelligence was applied for physiological health promotion in 6 included studies, covering sensory training for children with autism and motor rehabilitation for individuals with mobility impairments; however, most of these studies (5/6) were conducted in laboratory settings.

Three laboratory-based studies involving children with autism embedded traditional sensory training into embodied intelligence systems and used multimodal interaction (eg, voice and tactile interaction). These studies reported improvements in auditory comprehension, emotional comprehension, and attention [91-93].

In addition, for patients requiring exercise rehabilitation after cerebrovascular events or heart disease, embodied intelligence was used to monitor exercise performance in real time and provide timely feedback, helping patients track training effects and progress [94-96]. One study reported that patients' acceptance of embodied intelligence–assisted rehabilitation was higher than that of traditional exercise-based rehabilitation programs [97].

Evidence Distribution and Gaps Across Embodied Intelligence Types and Functional Domains

As shown in Figure 5, the evidence base was unevenly distributed across embodied intelligence types and functional domains. Bubble size and the number displayed within each bubble represent the number of included studies in each type-domain combination. The largest concentration of evidence was observed for virtual humanoid agents in health management and health education (n=25), followed by physical humanoid robots in mental health promotion (n=17). Physical humanoid robots were the only type represented across all 3 functional domains, including physiological health promotion (n=6). In contrast, virtual animal-shaped agents were represented by only 1 study, and mechanical robots remained limited overall. Across functional domains, physiological health promotion was the least represented area (6/83, 7.2%), highlighting a clear crosscutting evidence gap in the current literature.

The bubble map shows the distribution of the 83 included studies across embodied intelligence types and functional domains. Bubble size and the number shown within each bubble represent the number of studies in each category combination. Empty cells indicate that no studies were identified for that specific type-domain combination.

XSL•FO
RenderX

Figure 5. Bubble evidence gap map of embodied intelligence applications in health care.

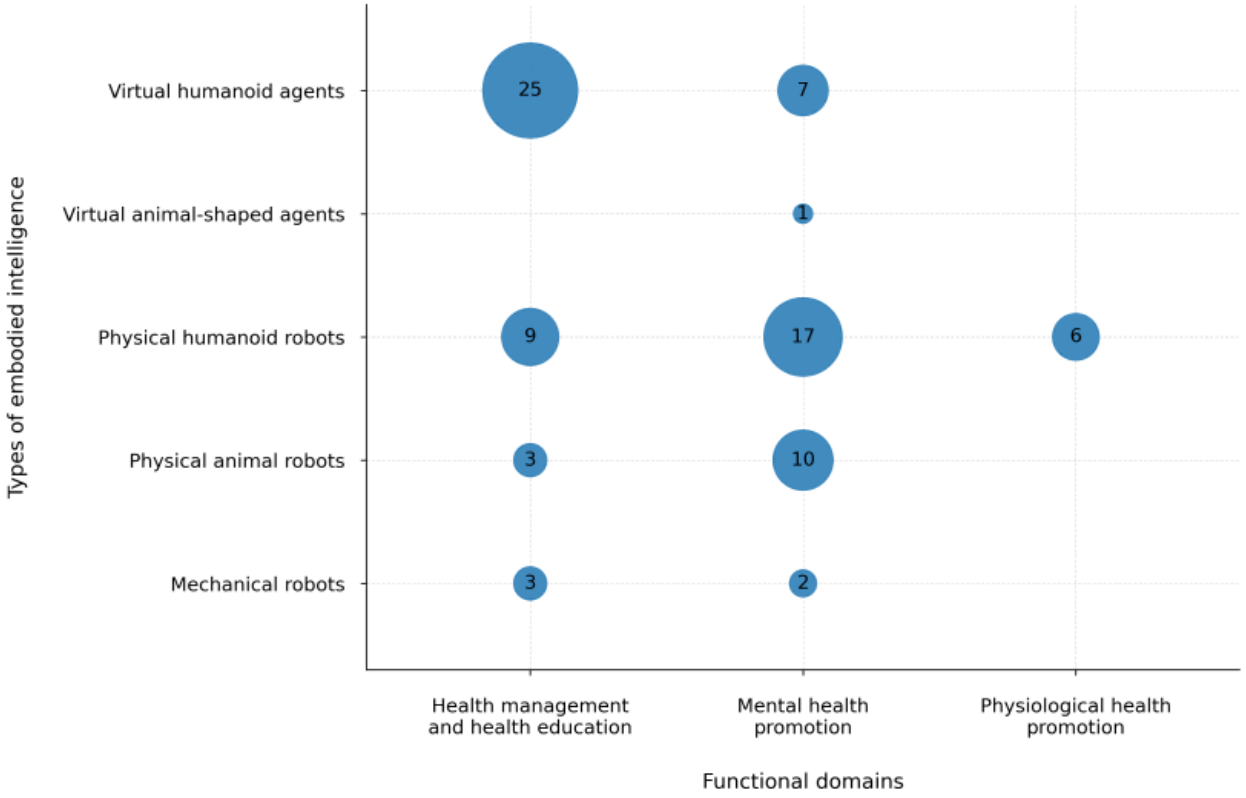

Acceptability and Effectiveness of Embodied Intelligence in Health Care

To strengthen coherence, Table 4 presents a role-based framework mapping the available evidence on acceptability and effectiveness of embodied intelligence across the 3 domains (health management and health education, mental health promotion, and physiological health promotion).

Table . A role-based framework mapping evidence on the acceptability and effectiveness of embodied intelligence in health care (N=83).

Role category	Studies, n (%)	Key findings on effectiveness	Key findings on acceptability
Health management and health education	40 (48.2)	Demonstrated efficacy in improving patient knowledge, medication adherence, and self-management skills	Generally high acceptance, particularly when systems were perceived as helpful and nonintrusive
Mental health promotion	37 (44.6)	Shown to reduce anxiety and depression symptoms, improve cognitive function, and enhance social interaction in older populations	High acceptance reported, especially for anthropomorphic designs; some studies noted concerns about privacy and overreliance
Physiological health promotion	6 (7.2)	Limited evidence, but promising results in promoting physical activity and assisting with rehabilitation exercises	Mixed acceptance; users valued functionality but sometimes found the technology complex or impersonal

Evidence of Acceptability

Evidence of acceptability refers to users' perceptions, satisfaction, usability, engagement, and willingness to adopt embodied intelligence systems.

As summarized in Table 4, about 12 studies examined the acceptability of embodied intelligence in health care applications, predominantly using mixed methods designs. Most focused on health management and education, with fewer addressing mental health promotion; however, heterogeneity in acceptability measures limited direct cross-study comparisons. These findings primarily reflect user experience and perceived feasibility, rather than measured health or behavioral outcomes.

In health management and education, acceptability appears to be shaped by an interplay of human, device-related, and environmental factors. Device-related considerations include anthropomorphic design features and interactive capabilities. Baptista et al [43] reported that acceptability increased when embodied intelligence was framed as a supportive coach rather than an authoritative medical expert in diabetes care. Similarly, Heffner et al [47] found that anthropomorphic appearance and movement fluidity were associated with higher acceptance,

XSL•FO
RenderX

underscoring the importance of anthropomorphic design. Consistent with this, Boustani et al [52] and Orejana et al [34] reported that voice interaction and light-based feedback can strengthen social presence and reduce perceived isolation, which was associated with enhanced acceptability. Regarding interactive functions, Reilly et al [36] observed in focus groups that an animal-like robot (Paro) elicited high acceptability by eliciting positive nonverbal responses (eg, petting and smiling), although one participant with animal aversion showed lower engagement. Human and environmental factors were also influential. Zhang [42], Boumans et al [21], and Broadbent et al [23] reported that health literacy, familiarity with technology, education level, and health status were associated with acceptability. Clinical versus home settings also emerged as an important contextual factor shaping acceptability.

Mental health promotion studies showed similar influences of individual and environmental factors on acceptability. Cobo Hurtado et al [68] found that older adults' initial acceptability of cognitive intervention functions was affected by technophobia; as proficiency increased, the negative influence of anxiety on acceptability diminished. In cross-setting comparisons, Jegundo et al [85] noted that embodied intelligence was more readily accepted in institutional health care settings, whereas home environments required personalized adaptations (eg, functional customization and optimized interaction modes) to enhance acceptability.

Evidence of Effectiveness

Overview

In this review, evidence of effectiveness refers to reported impacts of embodied intelligence interventions on health-related, behavioral, cognitive, or psychological outcomes.

As shown in Table 4, effectiveness evidence was primarily derived from RCTs (22 studies), concentrated in health management and education (12 studies) and mental health promotion (10 studies).

Health Management and Education

In this domain, embodied intelligence–based interventions generally reported favorable effects on health behaviors, clinical indicators, and health literacy. Studies by Bickmore et al [51], Robinson and Kavanagh [57], Reilly et al [36], National Clinical Trial [32], and Bickmore et al [50] compared embodied intelligence interventions with traditional management approaches and reported improvements in management goals such as increasing monthly average steps, reducing high-calorie food intake, and weight control. Gardiner et al [53] also reported improvements in stress management and healthy eating behaviors.

Studies by Moyle et al [31], King et al [55], and Jack et al [28] compared embodied intelligence–based health education with traditional health manual–based education and reported better outcomes. Specifically, patients reported higher self-rated quality of life and better adherence to healthy behaviors [54,98], and pregnant women had fewer clinic visits and lower prepregnancy risk ratios [55]. These findings are consistent with the results reported by Gong et al [45].

Bickmore et al [19] compared voice-based embodied intelligence with traditional web-based health information searching and reported higher search success rates and user satisfaction in the embodied intelligence group. In addition, Boumans et al [21] conducted an RCT with 42 older adults and found that embodied intelligence achieved accuracy comparable to health care providers in collecting disease-related information while reporting higher user acceptability.

Mental Health Promotion

In this domain, the evidence suggests that embodied intelligence may improve mental health outcomes and provide emotional support. The reported effects can be summarized in 3 dimensions: improvement of psychological symptoms, emotional support, and multimodal intervention.

For the improvement of psychological symptoms, several RCTs reported the benefits of embodied intelligence. For example, Karhiy et al [61] 3-arm trial reported that the embodied intelligence group achieved reductions in perceived stress comparable to human therapists and showed higher compliance than both the human-therapist and chatbot groups. Chen et al [100] and Jøranson et al [81] reported reductions in psychological distress in people with dementia and improvements in agitation and depressive symptoms in older adults. In addition, an RCT by Park et al [66] involving 135 older adults with mild cognitive impairment reported that embodied intelligence–assisted training improved cognitive function and reduced negative emotions.

Regarding emotional support, Robinson et al [82,83] and Lavin et al [73] reported that embodied intelligence can establish emotional connections with older adults and function as a social catalyst, alleviating loneliness and depressive feelings through interaction elements such as touch and conversation.

For multimodal intervention, several studies suggested added value relative to traditional approaches. Sebastian and Richards [91] reported that multimodal interaction (eg, language and expressions) reduced psychological stigmatization more than conventional staff-delivered interventions. Robinson et al [12] reported that mindfulness guidance delivered via embodied intelligence was more effective than traditional audio-guided methods, and Crossman et al [87] reported reduced anxiety in children through multimodal features such as facial expressions, sound, and touch.

Unlike acceptance-related evidence, these studies primarily evaluated intervention effectiveness using objective or validated outcome measures.

Synthesis of Results

Overall, the charted evidence indicates that embodied intelligence in health care is concentrated in 5 embodiment forms and 3 principal functional domains. The evidence base is most developed for virtual humanoid agents and physical humanoid robots, particularly in health management and health education and in mental health promotion. By contrast, virtual animal-shaped agents, mechanical robots, and applications in physiological health promotion remain comparatively underrepresented.

XSL•FO
RenderX

The evidence gap map further showed that the distribution of evidence across embodiment types and functional domains was uneven. Darker concentrations of evidence were observed for virtual humanoid agents in health management and health education and for physical humanoid robots in mental health promotion, whereas several type-domain combinations contained very limited or no evidence. Across embodiment types, physiological health promotion was consistently the least represented domain, indicating a cross-cutting research gap.

To support an integrative presentation of the findings, Table 4 summarizes the evidence on acceptability and effectiveness across the 3 major functional domains. Taken together, the results suggest that embodied intelligence has been most extensively investigated in aging-related and psychosocial care contexts, that acceptability is generally reported as favorable but context-dependent, and that evidence of effectiveness is promising but concentrated in a limited number of domains and study designs.

Discussion

Summary of Evidence

This scoping review systematically synthesizes the research progress and application landscape of embodied intelligence in health care. The literature shows sustained growth in publications and increasingly diversified implementation scenarios. Two recurring development directions emerge across the included studies: anthropomorphism and intelligence. Anthropomorphic design involves attributing human-like physical characteristics, motivations, intentions, and even emotional states to nonhuman entities [12,101]. Empirical work suggests that anthropomorphic features can enhance user experience via multimodal interaction (eg, verbal, facial, and textual modalities) and may improve perceived trust and safety [98].

In parallel, advances in AI—particularly large language models—have expanded the potential for personalized interaction. However, current human-agent interactions remain constrained by limitations in multimodal fusion, natural language understanding, and affective computing, which can affect fluency, accuracy, and recognition of users' emotions and needs. As shown by ter Stal et al [40], suboptimal personalization mechanisms and inaccurate information can reduce user compliance when embodied agents are used for health management. Taken together, these findings suggest that "more human-like" and "more intelligent" designs are promising but must be evaluated against safety, accuracy, and person-centered interaction quality before large-scale deployment.

Consistent with the thematic mapping, embodied intelligence in health care clusters into 3 functional domains: health management and education, mental health promotion, and physical health promotion. The predominance of studies in health management/education and mental health promotion indicates a current research focus on domains where conversational interaction, engagement, and relational support are central and where embodied agents may plausibly address workforce shortages and access barriers.

For health management and education, embodied intelligence appears advantageous because it can deliver multimodal guidance, provide tailored interaction, and support access independent of time and location. Reported levels of acceptance and compliance indicate feasibility for supporting—rather than uniformly replacing—clinical tasks in management and educational workflows, especially under resource constraints and uneven geographic distribution. In health education, multimodal interaction (voice, text, and facial expressions) has been reported to deliver personalized health knowledge and behavior guidance with efficacy comparable to medical professionals [55], while potentially reducing labor demands. Future research should prioritize clinically safe personalization (eg, accurate information provision, appropriate role framing, and escalation pathways) to ensure that convenience does not compromise care quality.

For mental health promotion, the literature highlights 2 particularly salient mechanisms: psychological intervention and emotional companionship. Regarding psychological intervention, evidence suggests that users—especially those with mental health challenges—may disclose more when interacting with a nonhuman agent than with traditional human-delivered approaches, potentially due to greater perceived privacy protection and reduced psychological defensiveness [102]. Regarding emotional companionship, the potential value is pronounced in older adults, where increasing numbers of seniors living alone and limited companionship resources create an urgent gap. Embodied intelligence may contribute through functional substitution (eg, daily assistance) and social substitution (eg, companionship), with anthropomorphic features increasing social presence and facilitating trust and emotional bonding [103]. Studies also indicate that emotion recognition and responsive interaction may alleviate loneliness and depressive symptoms among older adults [103,104], including those with cognitive decline. Importantly, acceptance and experience vary by setting (eg, nursing homes vs home care) [77]. Given the global shift toward home-based aging [105], future studies should prioritize ecologically valid home deployments, with design choices that fit household routines and caregiver workflows, to strengthen real-world transferability.

By contrast, embodied intelligence remains less represented in physical health promotion, which may reflect higher technical thresholds, stricter safety/access standards, and greater demands for robust performance in physical interaction [106]. Users may accept robot-assisted rehabilitation more readily than conventional training, potentially because gamified interfaces and real-time feedback enhance engagement [107]. Compared with traditional assistive devices, anthropomorphic interaction and emotional intelligence may improve communication and sustained participation [108]. To strengthen evidence, future work should benchmark interventions against relevant standards and usual-care comparators and report outcomes in a way that supports cross-study synthesis.

XSL•FO
RenderX

This review also synthesizes evidence on acceptability and effectiveness. Acceptability varies across studies and appears to be shaped by individual characteristics, robot-related factors, and environmental context. Unlike typical consumer products, acceptance in health care is tied not only to cost but also to perceived health value, trust in the technology, alignment with the agent's role, and interaction experience, as well as perceived fit and personalization within a specific care scenario [109]. Evidence of effectiveness is currently concentrated in health management/education and mental health promotion. Although only 22 RCTs were identified, all 22 reported positive effects, and some suggested benefits beyond traditional approaches. However, effectiveness assessment remains hindered by no standardized outcomes and heterogeneous methods, limiting comparability and generalizability. Establishing more consistent evaluation criteria would facilitate evidence synthesis. In addition, 2 quasi-experimental studies [97,110] highlight how technical limitations—particularly nonpersonalized or inappropriate feedback—can undermine perceived usability, trust, and acceptance, reinforcing the importance of optimizing interaction quality and safety.

This scoping review is innovative in providing a data-driven map of embodied intelligence in health care across embodiment forms, functional roles, settings, and evidence types. Unlike prior reviews that focus on a single device category or narrow clinical domain, this review integrates both virtual and physical embodiments and synthesizes acceptability and effectiveness evidence within a unified framework. This synthesis clarifies where evidence is most mature (health management and education; mental health promotion) and where major gaps remain (physiological health promotion, standardized outcomes, and longitudinal real-world evaluations). Our findings also highlight recurring determinants of implementation success, including role framing, interaction quality, and contextual fit across settings. In practice, these results can guide stakeholders to match embodiment and function to care needs and settings, while prioritizing safe personalization and ethically grounded deployment. Future research should focus on rigorous, ecologically valid evaluations and standardized reporting to support translation into scalable health care practice.

Limitations

This study limited its literature inclusion to publications in Chinese and English, which may have introduced selection bias by excluding relevant evidence published in other languages. Publication bias is also a concern, as studies reporting positive outcomes are more likely to be published than those reporting null or negative findings; this may lead to an overrepresentation of RCTs demonstrating the efficacy of embodied intelligence and an underrepresentation of potential limitations. In addition, a substantial proportion of included studies were conducted in laboratory-based evaluation contexts, which may limit ecological validity and the generalizability of findings to real-world care settings. Moreover, consistent with the purpose of a scoping review, we did not conduct a formal critical appraisal of study quality. Therefore, findings—particularly those related to acceptability and effectiveness—should be interpreted cautiously, as the methodological rigor of included studies may vary and could influence the strength of conclusions.

Conclusions

This scoping review provides a comprehensive synthesis of current evidence on the applications of embodied intelligence in health care, highlighting both emerging trends and critical gaps in the literature. The findings indicate that existing research has predominantly focused on health management and mental health promotion, while evidence for physiological health promotion remains limited. In addition, certain forms of embodied intelligence—particularly virtual animal-shaped agents and mechanical robots—are markedly underrepresented, and studies conducted in real-world clinical settings, especially hospitals, are scarce.

Based on these findings, future research should prioritize addressing these gaps through rigorous and longitudinal study designs, including RCTs and implementation-focused studies conducted in routine health care environments. Such efforts are essential to better understand not only the effectiveness of embodied intelligence interventions but also their sustainability, scalability, and integration into existing care systems. From the perspective of health care and nursing practice, embodied intelligence has the potential to serve as a complementary tool for health education, chronic disease management, and psychosocial support. However, successful adoption will depend on careful attention to usability, workflow integration, and ethical considerations, as well as on clearly defining the roles of nurses and other health care professionals as facilitators, supervisors, and mediators of human-technology interaction. Addressing these practical considerations alongside methodological gaps will be crucial for translating embodied intelligence from experimental settings into meaningful, patient-centered health care practice.

Acknowledgments

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. Translation tasks were delegated to GenAI tools under full human supervision. The GenAI tool used was Qwen3.5-Plus. Responsibility for the final manuscript lies entirely with the authors.

Funding

This work was supported by the University-Industry Collaborative Education Program (project no. 230905329045253) and the Key Laboratory of Geriatric Long-Term Care, Ministry of Education (Naval Medical University). The funders had no involvement in the study design, data collection, analysis, interpretation of results, or the writing of the manuscript.

XSL•FO
RenderX

Data Availability

All data supporting the findings of this scoping review are included in the main text and its supplementary information. For further details or additional datasets, contact the corresponding author.

Authors' Contributions

Conceptualization: WS, YZ, ES, XY
Data curation: WS, YZ, ES
Formal analysis: WS, YZ
Investigation: WS, YZ
Methodology: WS, YZ, ES, XY
Project administration: ES, XY
Supervision: ES, XY
Funding acquisition: XY
Writing – original draft: WS, YZ
Writing – review & editing: WS, YZ, ES, LD, XY
WS, YZ, and ES are the co-first authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File, 24 KB - jmir_v28i1e83871_app1.docx]

Multimedia Appendix 2
Characteristics of all included studies.
[DOCX File, 53 KB - jmir_v28i1e83871_app2.docx]

Checklist 1
PRISMA-ScR checklist.
[PDF File, 608 KB - jmir_v28i1e83871_app3.pdf]

Checklist 2
PRISMA-S checklist.
[PDF File, 144 KB - jmir_v28i1e83871_app4.pdf]

References

1. Ni Y, Zhou Y, Kivimäki M, et al. Socioeconomic inequalities in physical, psychological, and cognitive multimorbidity in middle-aged and older adults in 33 countries: a cross-sectional study. Lancet Healthy Longev 2023 Nov;4(11):e618-e628. [doi: 10.1016/S2666-7568(23)00195-2] [Medline: 37924843]
2. Tong X, Xu J, Gong E, et al. Frailty as a breakthrough point for multimorbidity management among older adults: challenges and opportunities in China. BMJ 2024 Oct;387:e076767. [doi: 10.1136/bmj-2023-076767]
3. Nichol B, McCready J, Erfani G, et al. Exploring the impact of socially assistive robots on health and wellbeing across the lifespan: an umbrella review and meta-analysis. Int J Nurs Stud 2024 May;153:104730. [doi: 10.1016/j.ijnurstu.2024.104730] [Medline: 38430662]
4. Huang X, Gu L, Sun J, Eils R. Bridging the gaps: overcoming challenges of implementing AI in healthcare. Med 2025 Apr 11;6(4):100666. [doi: 10.1016/j.medj.2025.100666] [Medline: 40220747]
5. Mohajer-Bastami A, Moin S, Ahmad S, et al. Artificial intelligence in healthcare: applications, challenges, and future directions: a narrative review informed by international, multidisciplinary expertise. Front Digit Health 2025;7:1644041. [doi: 10.3389/fdgth.2025.1644041] [Medline: 41306935]
6. Liu Y, Cao X, Chen T, et al. From screens to scenes: a survey of embodied AI in healthcare. Information Fusion 2025 Jul;119:103033. [doi: 10.1016/j.inffus.2025.103033]
7. Kwon W, Baek S, Baek J, et al. Reinforced intelligence through active interaction in real world: a survey on embodied AI. Int J Control Autom Syst 2025 Jun;23(6):1597-1612. [doi: 10.1007/s12555-025-0127-1]
8. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]

XSL•FO
RenderX

9. Fiske A, Henningsen P, Buyx A. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J Med Internet Res 2019 May 9;21(5):e13216. [doi: 10.2196/13216] [Medline: 31094356]
10. Huang R, Li H, Suomi R, Li C, Peltoniemi T. Intelligent physical robots in health care: systematic literature review. J Med Internet Res 2023 Jan 18;25:e39786. [doi: 10.2196/39786] [Medline: 36652280]
11. Kiuchi K, Otsu K, Hayashi Y. Psychological insights into the research and practice of embodied conversational agents, chatbots and social assistive robots: a systematic meta-review. Behav Inf Technol 2024 Nov 17;43(15):3696-3736. [doi: 10.1080/0144929X.2023.2286528]
12. Robinson NL, Connolly J, Suddrey G, Kavanagh DJ. A brief wellbeing training session delivered by a humanoid social robot: a pilot randomized controlled trial. Int J of Soc Robotics 2024 May;16(5):937-951. [doi: 10.1007/s12369-023-01054-5]
13. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
14. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5:69. [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
15. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
16. Shu W. Embodied intelligence in healthcare: a scoping review. . 2025 Aug 5. [doi: 10.17605/OSF.IO/UYNS4]
17. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc 2015 Sep;13(3):141-146. [doi: 10.1097/XEB.0000000000000050] [Medline: 26134548]
18. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
19. Bickmore TW, Utami D, Matsuyama R, Paasche-Orlow MK. Improving access to online health information with conversational agents: a randomized controlled experiment. J Med Internet Res 2016 Jan 4;18(1):e1. [doi: 10.2196/jmir.5239] [Medline: 26728964]
20. Blindheim K, Solberg M, Hameed IA, Alnes RE. Promoting activity in long-term care facilities with the social robot Pepper: a pilot study. Inform Health Soc Care 2023 Apr 3;48(2):181-195. [doi: 10.1080/17538157.2022.2086465] [Medline: 35702818]
21. Boumans R, van Meulen F, Hindriks K, Neerincx M, Olde Rikkert MGM. Robot for health data acquisition among older adults: a pilot randomised controlled cross-over trial. BMJ Qual Saf 2019 Oct;28(10):793-799. [doi: 10.1136/bmjqs-2018-008977] [Medline: 30894423]
22. Boumans R, van Meulen F, Hindriks K, Neerincx M, Olde Rikkert M. A feasibility study of a social robot collecting patient reported outcome measurements from older adults. Int J of Soc Robotics 2020 Jan;12(1):259-266. [doi: 10.1007/s12369-019-00561-8]
23. Broadbent E, Peri K, Kerse N, et al. Robots in older people's homes to improve medication adherence and quality of life: a randomised cross-over trial. In: Beetz M, Johnston B, Williams MA, editors. Social Robotics: Springer; 2014:64-73. [doi: 10.1007/978-3-319-11973-1_7]
24. Broadbent E, Kerse N, Peri K, et al. Benefits and problems of health-care robots in aged care settings: a comparison trial. Australas J Ageing 2016 Mar;35(1):23-29. [doi: 10.1111/ajag.12190] [Medline: 26364706]
25. Chen YC, Yeh SL, Huang TR, Chang YL, Goh JOS, Fu LC. Social robots for evaluating attention state in older adults. Sensors (Basel) 2021 Oct 28;21(21):7142. [doi: 10.3390/s21217142] [Medline: 34770448]
26. Curumsing MK, Rivera Villicana J, Vouliotis A, et al. Talk with Ted: an embodied conversational agent for caregivers. Gerontol Geriatr Educ 2025;46(2):186-203. [doi: 10.1080/02701960.2024.2302584] [Medline: 38252487]
27. Demange M, Pino M, Kerhervé H, Rigaud AS, Cantegreil-Kallen I. Management of acute pain in dementia: a feasibility study of a robot-assisted intervention. J Pain Res 2019;12:1833-1846. [doi: 10.2147/JPR.S179640] [Medline: 31289446]
28. Jack BW, Bickmore T, Yinusa-Nyahkoon L, et al. Improving the health of young African American women in the preconception period using health information technology: a randomised controlled trial. Lancet Digit Health 2020 Sep;2(9):e475-e485. [doi: 10.1016/S2589-7500(20)30189-8] [Medline: 33328115]
29. Law M, Sutherland C, Ahn HS, et al. Developing assistive robots for people with mild cognitive impairment and mild dementia: a qualitative study with older adults and experts in aged care. BMJ Open 2019 Sep 24;9(9):e031937. [doi: 10.1136/bmjopen-2019-031937] [Medline: 31551392]
30. Martínez-Miranda J, Martínez A, Ramos R, et al. Assessment of users' acceptability of a mobile-based embodied conversational agent for the prevention and detection of suicidal behaviour. J Med Syst 2019 Jun 25;43(8):246. [doi: 10.1007/s10916-019-1387-1] [Medline: 31240494]
31. Moyle W, Cooke M, Beattie E, et al. Exploring the effect of companion robots on emotional expression in older adults with dementia: a pilot randomized controlled trial. J Gerontol Nurs 2013 May;39(5):46-53. [doi: 10.3928/00989134-20130313-03] [Medline: 23506125]
32. Usage and health effects of embodied conversational agents among older adults. National Clinical Trial. URL: https://clinicaltrials.gov/show/NCT04510883 [accessed 2026-05-19]

XSL•FO
RenderX

33. Olde Keizer RACM, van Velsen L, Moncharmont M, et al. Using socially assistive robots for monitoring and preventing frailty among older adults: a study on usability and user experience challenges. Health Technol 2019 Aug;9(4):595-605. [doi: 10.1007/s12553-019-00320-9]
34. Orejana JR, MacDonald BA, Ahn HS, Peri K, Broadbent E. Healthcare robots in homes of rural older adults. In: Tapus A, André E, Martin JC, Ferland F, Ammi M, editors. Social Robotics: Springer International Publishing; 2015, Vol. 9388:512-521. [doi: 10.1007/978-3-319-25554-5_51]
35. Pu L, Moyle W, Jones C, Todorovic M. The effect of a social robot intervention on sleep and motor activity of people living with dementia and chronic pain: a pilot randomized controlled trial. Maturitas 2021 Feb;144:16-22. [doi: 10.1016/j.maturitas.2020.09.003] [Medline: 33358203]
36. Reilly ED, Kathawalla UK, Robins HE, et al. An online acceptance and mindfulness intervention for chronic pain in veterans: development and protocol for a pilot feasibility randomized controlled trial. JMIR Res Protoc 2023 Mar 7;12:e45887. [doi: 10.2196/45887] [Medline: 36881446]
37. Richards D, Caldwell P. Improving health outcomes sooner rather than later via an interactive website and virtual specialist. IEEE J Biomed Health Inform 2018 Sep;22(5):1699-1706. [doi: 10.1109/JBHI.2017.2782210] [Medline: 29990229]
38. Stara V, Vera B, Bolliger D, et al. Toward the integration of technology-based interventions in the care pathway for people with dementia: a cross-national study. Int J Environ Res Public Health 2021 Oct 2;18(19):10405. [doi: 10.3390/ijerph181910405] [Medline: 34639704]
39. Stara V, Vera B, Bolliger D, et al. Usability and acceptance of the embodied conversational agent Anne by people with dementia and their caregivers: exploratory study in home environment settings. JMIR Mhealth Uhealth 2021 Jun 25;9(6):e25891. [doi: 10.2196/25891] [Medline: 34170256]
40. ter Stal S, Sloots J, Ramlal A, op den Akker H, Lenferink A, Tabak M. An embodied conversational agent in an eHealth self-management intervention for chronic obstructive pulmonary disease and chronic heart failure: exploratory study in a real-life setting. JMIR Hum Factors 2021 Nov 4;8(4):e24110. [doi: 10.2196/24110] [Medline: 34734824]
41. Yoshii K, Kimura D, Kosugi A, et al. Screening of mild cognitive impairment through conversations with humanoid robots: exploratory pilot study. JMIR Form Res 2023 Jan 13;7:e42792. [doi: 10.2196/42792] [Medline: 36637896]
42. Zhang ZZ. Medical shared decision making with embodied conversational agents [Dissertation]. : Northeastern University; 2020 Mar. [doi: 10.17760/D20350241]
43. Baptista S, Wadley G, Bird D, Oldenburg B, Speight J, My Diabetes Coach Research Group. Acceptability of an embodied conversational agent for type 2 diabetes self-management education and support via a smartphone app: mixed methods study. JMIR Mhealth Uhealth 2020 Jul 22;8(7):e17038. [doi: 10.2196/17038] [Medline: 32706734]
44. Bickmore TW, Silliman RA, Nelson K, et al. A randomized controlled trial of an automated exercise coach for older adults. J Am Geriatr Soc 2013 Oct;61(10):1676-1683. [doi: 10.1111/jgs.12449] [Medline: 24001030]
45. Gong E, Baptista S, Russell A, et al. My diabetes coach, a mobile app-based interactive conversational agent to support type 2 diabetes self-management: randomized effectiveness-implementation trial. J Med Internet Res 2020 Nov 5;22(11):e20322. [doi: 10.2196/20322] [Medline: 33151154]
46. Görer B, Salah AA, Akın HL. An autonomous robotic exercise tutor for elderly people. Auton Robot 2017 Mar;41(3):657-678. [doi: 10.1007/s10514-016-9598-5]
47. Heffner JL, Watson NL, Serfozo E, et al. An avatar-led digital smoking cessation program for sexual and gender minority young adults: intervention development and results of a single-arm pilot trial. JMIR Form Res 2021 Jul 30;5(7):e30241. [doi: 10.2196/30241] [Medline: 34328430]
48. Nkabane-Nkholongo E, Mokgatle M, Bickmore T, Julce C, Jack BW. Adaptation of the Gabby conversational agent system to improve the sexual and reproductive health of young women in Lesotho. Front Digit Health 2023;5:1224429. [doi: 10.3389/fdgth.2023.1224429] [Medline: 37860039]
49. Beinema T, Op den Akker H, Hurmuz M, Jansen-Kosterink S, Hermens H. Automatic topic selection for long-term interaction with embodied conversational agents in health coaching: a micro-randomized trial. Internet Interv 2022 Mar;27:100502. [doi: 10.1016/j.invent.2022.100502] [Medline: 35198412]
50. Bickmore TW, Pfeifer LM, Byron D, et al. Usability of conversational agents by patients with inadequate health literacy: evidence from two clinical trials. J Health Commun 2010;15(Suppl 2):197-210. [doi: 10.1080/10810730.2010.499991] [Medline: 20845204]
51. Bickmore TW, Schulman D, Sidner C. Automated interventions for multiple health behaviors using conversational agents. Patient Educ Couns 2013 Aug;92(2):142-148. [doi: 10.1016/j.pec.2013.05.011] [Medline: 23763983]
52. Boustani M, Lunn S, Visser U, Lisetti C. Development, feasibility, acceptability, and utility of an expressive speech-enabled digital health agent to deliver online, brief motivational interviewing for alcohol misuse: descriptive study. J Med Internet Res 2021 Sep 29;23(9):e25837. [doi: 10.2196/25837] [Medline: 34586074]
53. Gardiner PM, McCue KD, Negash LM, et al. Engaging women with an embodied conversational agent to deliver mindfulness and lifestyle recommendations: a feasibility randomized control trial. Patient Educ Couns 2017 Sep;100(9):1720-1729. [doi: 10.1016/j.pec.2017.04.015] [Medline: 28495391]

XSL•FO
RenderX

54. Hurmuz MZM, Jansen-Kosterink SM, Op den Akker H, Hermens HJ. User experience and potential health effects of a conversational agent-based electronic health intervention: protocol for an observational cohort study. JMIR Res Protoc 2020 Apr 3;9(4):e16641. [doi: 10.2196/16641] [Medline: 32242517]
55. King AC, Campero I, Sheats JL, et al. Testing the comparative effects of physical activity advice by humans vs. computers in underserved populations: the COMPASS trial design, methods, and baseline characteristics. Contemp Clin Trials 2017 Oct;61:115-125. [doi: 10.1016/j.cct.2017.07.020] [Medline: 28739541]
56. Kramer LL, Blok M, van Velsen L, Mulder BC, de Vet E. Supporting eating behaviour of community-dwelling older adults: co-design of an embodied conversational agent. Design Health (Abingdon) 2021;5(1):120-139. [doi: 10.1080/24735132.2021.1885592] [Medline: 34381936]
57. Robinson NL, Kavanagh DJ. A social robot to deliver a psychotherapeutic treatment: qualitative responses by participants in a randomized controlled trial and future design recommendations. Int J Hum Comput Stud 2021 Nov;155:102700. [doi: 10.1016/j.ijhcs.2021.102700]
58. Ma G, Dong L, Jiao M, et al. Application of interactive mobile health education using intelligent robots in the perioperative care of esophageal cancer patients. Nurs Pract Res 2023;20(22):3322-3327. [doi: 10.3969/j.issn.1672-9676.2023.22.016]
59. Cho M, Kim D, Jang M, et al. Evaluating human-care robot services for the elderly: an experimental study. Int J of Soc Robotics 2024 Jul;16(7):1561-1587. [doi: 10.1007/s12369-024-01157-7]
60. Hudlicka E. Virtual training and coaching of health behavior: example from mindfulness meditation training. Patient Educ Couns 2013 Aug;92(2):160-166. [doi: 10.1016/j.pec.2013.05.007] [Medline: 23809167]
61. Karhiy M, Sagar M, Antoni M, Loveys K, Broadbent E. Can a virtual human increase mindfulness and reduce stress? A randomised trial. Comput Human Behav Artif Hum 2024 Jan;2(1):100069. [doi: 10.1016/j.chbah.2024.100069]
62. Shidara K, Tanaka H, Adachi H, Kanayama D, Kudo T, Nakamura S. Adapting the number of questions based on detected psychological distress for cognitive behavioral therapy with an embodied conversational agent: comparative study. JMIR Form Res 2024 Mar 14;8:e50056. [doi: 10.2196/50056] [Medline: 38483464]
63. Suganuma S, Sakamoto D, Shimoyama H. An embodied conversational agent for unguided internet-based cognitive behavior therapy in preventative mental health: feasibility and acceptability pilot trial. JMIR Ment Health 2018 Jul 31;5(3):e10454. [doi: 10.2196/10454] [Medline: 30064969]
64. Chan J, Nejat G. Promoting engagement in cognitively stimulating activities using an intelligent socially assistive robot. In: 2010 IEEE/ASME International Conference on Advanced Intelligent Mechatronics: IEEE; 2010:533-538. [doi: 10.1109/AIM.2010.5695806]
65. Fan J, Mion LC, Beuscher L, Ullal A, Newhouse PA, Sarkar N. SAR-Connect: a socially assistive robotic system to support activity and social engagement of older adults. IEEE Trans Robot 2022 Apr;38(2):1250-1269. [doi: 10.1109/tro.2021.3092162] [Medline: 36204285]
66. Park EA, Jung AR, Lee KA. The humanoid robot Sil-Bot in a cognitive training program for community-dwelling elderly people with mild cognitive impairment during the COVID-19 pandemic: a randomized controlled trial. Int J Environ Res Public Health 2021 Aug 3;18(15):8198. [doi: 10.3390/ijerph18158198] [Medline: 34360490]
67. Bott N, Wexler S, Drury L, et al. A protocol-driven, bedside digital conversational agent to support nurse teams and mitigate risks of hospitalization in older adults: case control pre-post study. J Med Internet Res 2019 Oct 17;21(10):e13440. [doi: 10.2196/13440] [Medline: 31625949]
68. Cobo Hurtado L, Viñas PF, Zalama E, Gómez-García-Bermejo J, Delgado JM, Vielba García B. Development and usability validation of a social robot platform for physical and cognitive stimulation in elder care facilities. Healthcare (Basel) 2021 Aug 19;9(8):1067. [doi: 10.3390/healthcare9081067] [Medline: 34442204]
69. Fields N, Xu L, Greer J, Murphy E. Shall I compare thee…to a robot? An exploratory pilot study using participatory arts and social robotics to improve psychological well-being in later life. Aging Ment Health 2021 Mar;25(3):575-584. [doi: 10.1080/13607863.2019.1699016] [Medline: 31851830]
70. Huisman C, Kort H. Two-year use of care robot Zora in Dutch nursing homes: an evaluation study. Healthcare (Basel) 2019 Feb 19;7(1):31. [doi: 10.3390/healthcare7010031] [Medline: 30791489]
71. Khosla R, Chu MT. Embodying care in Matilda: an affective communication robot for emotional wellbeing of older people in Australian residential care facilities. ACM Trans Manage Inf Syst 2013 Dec;4(4):18. [doi: 10.1145/2544104]
72. Lane GW, Noronha D, Rivera A, et al. Effectiveness of a social robot, “Paro,” in a VA long-term care setting. Psychol Serv 2016 Aug;13(3):292-299. [doi: 10.1037/ser0000080] [Medline: 27195530]
73. Lavin P, Lesage M, Monroe E, et al. Humanoid robot intervention vs. treatment as usual for loneliness in long-term care homes: study protocol for a pilot randomized controlled trial. Front Psychiatry 2022;13:1003881. [doi: 10.3389/fpsyt.2022.1003881] [Medline: 36339874]
74. Pérez-Zuñiga G, Arce D, Gibaja S, et al. Qhali: a humanoid robot for assisting in mental health treatment. Sensors (Basel) 2024 Feb 18;24(4):1321. [doi: 10.3390/s24041321] [Medline: 38400478]
75. Pu L, Moyle W, Jones C. How people with dementia perceive a therapeutic robot called PARO in relation to their pain and mood: a qualitative study. J Clin Nurs 2020 Feb;29(3-4):437-446. [doi: 10.1111/jocn.15104] [Medline: 31738463]
76. Ruiz-del-Solar J, Salazar M, Vargas-Araya V, et al. Mental and emotional health care for COVID-19 patients: employing Pudu, a telepresence robot. IEEE Robot Automat Mag 2021 Mar;28(1):82-89. [doi: 10.1109/MRA.2020.3044906]

XSL•FO
RenderX

77. Tanioka T, Betriana F, Yokotani T, et al. The experience of older persons with mental health conditions who interact with healthcare robots and nurse intermediaries: the qualitative case studies. Belitung Nurs J 2021;7(4):346-353. [doi: 10.33546/bnj.1541] [Medline: 37484892]
78. Birks M, Bodak M, Barlas J, Harwood J, Pether M. Robotic seals as therapeutic tools in an aged care facility: a qualitative study. J Aging Res 2016;2016:8569602. [doi: 10.1155/2016/8569602] [Medline: 27990301]
79. D'Onofrio G, Sancarlo D, Raciti M, et al. MARIO project: validation and evidence of service robots for older people with dementia. J Alzheimers Dis 2019;68(4):1587-1601. [doi: 10.3233/JAD-181165] [Medline: 30958360]
80. Hudson J, Ungar R, Albright L, Tkatch R, Schaeffer J, Wicker ER. Robotic pet use among community-dwelling older adults. J Gerontol B Psychol Sci Soc Sci 2020 Oct 16;75(9):2018-2028. [doi: 10.1093/geronb/gbaa119] [Medline: 32789476]
81. Jøranson N, Pedersen I, Rokstad AMM, Ihlebæk C. Effects on symptoms of agitation and depression in persons with dementia participating in robot-assisted activity: a cluster-randomized controlled trial. J Am Med Dir Assoc 2015 Oct 1;16(10):867-873. [doi: 10.1016/j.jamda.2015.05.002] [Medline: 26096582]
82. Jøranson N, Pedersen I, Rokstad AMM, Aamodt G, Olsen C, Ihlebæk C. Group activity with Paro in nursing homes: systematic investigation of behaviors in participants. Int Psychogeriatr 2016 Aug;28(8):1345-1354. [doi: 10.1017/S1041610216000120] [Medline: 27019225]
83. Robinson H, Macdonald B, Kerse N, Broadbent E. The psychosocial effects of a companion robot: a randomized controlled trial. J Am Med Dir Assoc 2013 Sep;14(9):661-667. [doi: 10.1016/j.jamda.2013.02.007] [Medline: 23545466]
84. Robinson H, Broadbent E, MacDonald B. Group sessions with Paro in a nursing home: structure, observations and interviews. Australas J Ageing 2016 Jun;35(2):106-112. [doi: 10.1111/ajag.12199] [Medline: 26059390]
85. Jegundo AL, Dantas C, Quintas J, et al. Perceived usefulness, satisfaction, ease of use and potential of a virtual companion to support the care provision for older adults. Technologies (Basel) 2020 Jul;8(3):42. [doi: 10.3390/technologies8030042]
86. Leung AYM, Zhao IY, Lin S, Lau TK. Exploring the presence of humanoid social robots at home and capturing human-robot interactions with older adults: experiences from four case studies. Healthcare (Basel) 2022 Dec 22;11(1):39. [doi: 10.3390/healthcare11010039] [Medline: 36611499]
87. Crossman MK, Kazdin AE, Kitt ER. The influence of a socially assistive robot on mood, anxiety, and arousal in children. Prof Psychol Res Pr 2018;49(1):48-56. [doi: 10.1037/pro0000177]
88. Rossi S, Larafa M, Ruocco M. Emotional and behavioural distraction by a social robot for children anxiety reduction during vaccination. Int J of Soc Robotics 2020 Jul;12(3):765-777. [doi: 10.1007/s12369-019-00616-w]
89. Kitt ER, Crossman MK, Matijczak A, Burns GB, Kazdin AE. Evaluating the role of a socially assistive robot in children's mental health care. J Child Fam Stud 2021;30(7):1722-1735. [doi: 10.1007/s10826-021-01977-5] [Medline: 34025101]
90. O'Leary TKZ. Co-design and evaluation of a smartphone-based mental health promotion and anti-stigma embodied conversational agent for church-affiliated black adults [Dissertation]. : Northeastern University; 2023. [doi: 10.17760/D20620618]
91. Sebastian J, Richards D. Changing stigmatizing attitudes to mental health via education and contact with embodied conversational agents. Comput Human Behav 2017 Aug;73:479-488. [doi: 10.1016/j.chb.2017.03.071]
92. Costa S, Lehmann H, Dautenhahn K, Robins B, Soares F. Using a humanoid robot to elicit body awareness and appropriate physical interaction in children with autism. Int J of Soc Robotics 2015 Apr;7(2):265-278. [doi: 10.1007/s12369-014-0250-2]
93. Louie WYG, Korneder J, Abbas I, Pawluk C. A study on an applied behavior analysis-based robot-mediated listening comprehension intervention for ASD. Paladyn J Behav Robot 2021 Feb 10;12(1):31-46. [doi: 10.1515/pjbr-2021-0005]
94. Robins B, Dautenhahn K. Tactile interactions with a humanoid robot: novel play scenario implementations with children with autism. Int J of Soc Robotics 2014 Aug;6(3):397-415. [doi: 10.1007/s12369-014-0228-0]
95. Casas J, Irfan B, Senft E, et al. Social assistive robot for cardiac rehabilitation: a pilot study with patients with angioplasty. In: HRI '18: Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction: ACM; 2018:79-80. [doi: 10.1145/3173386.3177052]
96. Hun Lee M, Siewiorek DP, Smailagic A, Bernardino A, Bermúdez I Badia S. Design, development, and evaluation of an interactive personalized social robot to monitor and coach post-stroke rehabilitation exercises. User Model User-adapt Interact 2023;33(2):545-569. [doi: 10.1007/s11257-022-09348-5] [Medline: 37123108]
97. Nomura T, Kanda T, Yamada S, Suzuki T. The effects of assistive walking robots for health care support on older persons: a preliminary field experiment in an elder care facility. Intel Serv Robotics 2021 Mar;14(1):25-32. [doi: 10.1007/s11370-020-00345-4]
98. Zhang T, Au Yeung C, Aurelia E, et al. Prompting an embodied AI agent: how embodiment and multimodal signaling affects prompting behaviour. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems: ACM; 2025:1-25. [doi: 10.1145/3706598.3713110]
99. Rodríguez-Domínguez MT, Bazago-Dómine MI, Jiménez-Palomares M, et al. Interaction assessment of a social-care robot in day center patients with mild to moderate cognitive impairment: a pilot study. Int J of Soc Robotics 2024 Mar;16(3):513-528. [doi: 10.1007/s12369-024-01106-4]
100. Chen K, Lou VWQ, Tan KCK, Wai MY, Chan LL. Effects of a humanoid companion robot on dementia symptoms and caregiver distress for residents in long-term care. J Am Med Dir Assoc 2020 Nov;21(11):1724-1728. [doi: 10.1016/j.jamda.2020.05.036] [Medline: 32713772]

XSL•FO
RenderX

101. Ackermann H, Henke A, Chevalère J, et al. Physical embodiment and anthropomorphism of AI tutors and their role in student enjoyment and performance. NPJ Sci Learn 2025 Jan 8;10(1):1. [doi: 10.1038/s41539-024-00293-z] [Medline: 39779711]
102. Papneja H, Yadav N. Self-disclosure to conversational AI: a literature review, emergent framework, and directions for future research. Pers Ubiquit Comput 2025 Apr;29(2):119-151. [doi: 10.1007/s00779-024-01823-7]
103. Chen SC, Moyle W, Jones C, Petsky H. A social robot intervention on depression, loneliness, and quality of life for Taiwanese older adults in long-term care. Int Psychogeriatr 2020 Aug;32(8):981-991. [doi: 10.1017/S1041610220000459] [Medline: 32284080]
104. Li Q, Luximon Y, Zhang J. The influence of anthropomorphic cues on patients' perceived anthropomorphism, social presence, trust building, and acceptance of health care conversational agents: within-subject web-based experiment. J Med Internet Res 2023 Aug 10;25:e44479. [doi: 10.2196/44479] [Medline: 37561567]
105. Haferkamp T, Ing S, Weyer G, Gleason LJ. Home-based primary care (HBPC): aging in place in 2025. Current Geriatrics Reports 2025 Apr;14(1):3. [doi: 10.1007/s13670-025-00429-y]
106. Iakovidis DK, Vartholomeos P, Gall AL, et al. Medical & healthcare robotics: a roadmap for enhanced precision, safety, and efficacy. Meas Sci Technol 2025 Oct 31;36(10):103001. [doi: 10.1088/1361-6501/ae09bf]
107. Mohebbi A. Human-robot Interaction in rehabilitation and assistance: a review. Current Robotics Reports 2020 Sep;1(3):131-144. [doi: 10.1007/s43154-020-00015-4]
108. Sharma A, Rathi Y, Patni V, Sinha DK. A systematic review of assistance robots for elderly care. In: 2021 International Conference on Communication Information and Computing Technology (ICCICT): IEEE; 2021:1-6. [doi: 10.1109/ICCICT50803.2021.9510142]
109. Wang Y, Chen H, Cheng H, et al. The attitudes and acceptance of functional assistive robots among older adults with disabilities: a mixed-methods study. Innov Aging 2025;9(9):igaf097. [doi: 10.1093/geroni/igaf097] [Medline: 41140795]
110. Wainer J, Dautenhahn K, Robins B, Amirabdollahian F. A pilot study with a novel setup for collaborative play of the humanoid robot KASPAR with children with autism. Int J of Soc Robotics 2014 Jan;6(1):45-65. [doi: 10.1007/s12369-013-0195-x]

Abbreviations

AI: artificial intelligence
CNKI: China National Knowledge Infrastructure
CQVIP: VIP Chinese Journal Database
OSF: Open Science Framework
PCC: population, concept, and context framework
PRESS: Peer Review of Electronic Search Strategies
PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
RCT: randomized controlled trial

Edited by S Brini; submitted 10.Sep.2025; peer-reviewed by CK Tan, R Suomi, R Huang, S Sharma; revised version received 09.Apr.2026; accepted 10.Apr.2026; published 12.Jun.2026.

Please cite as:
Shu W, Zhou Y, Sun E, Deng L, Ye X
Embodied Intelligence Applications in Health Care Populations: Scoping Review
J Med Internet Res 2026;28:e83871
URL: https://www.jmir.org/2026/1/e83871
doi:10.2196/83871

© Wan Shu, Ying Zhou, Erhong Sun, Lei Deng, Xuchun Ye. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Applications of DeepSeek in Medicine: Bibliometric Analysis and Scoping Review

Haoran Zhang[1,2*], MS; Dawei Wang[3*], PhD; Yanliang Xu[4], BSc; Shuming Han[2], MS; Guangxin Wang[2], MD, PhD

[1]School of Clinical Medicine, Shandong Second Medical University, Weifang, Shandong, China

[2]Shandong Innovation Center of Intelligent Diagnostic Technology, Central Hospital Affiliated to Shandong First Medical University, 105 Jiefang Road, Jinan, Shandong, China

[3]Key Laboratory of Endocrine Glucose & Lipids Metabolism and Brain Aging, Ministry of Education; Department of Endocrinology, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, Shandong, China

[4]Library, Shandong Second Medical University, Weifang, Shandong, China

[*]these authors contributed equally

Corresponding Author:
Guangxin Wang, MD, PhD
Shandong Innovation Center of Intelligent Diagnostic Technology, Central Hospital Affiliated to Shandong First Medical University, 105 Jiefang Road, Jinan, Shandong, China

Abstract

Background: The integration of large language models (LLMs) into medicine has reshaped health care delivery, education, and research. Although proprietary models face challenges such as data privacy, regulation, and adaptability, DeepSeek, an open-source LLM, has emerged as a customizable and cost-effective alternative with significant potential for clinical and operational applications. However, the rapid expansion of research in this area necessitates a systematic mapping of its landscape, applications, and challenges.

Objective: This study combines bibliometric analysis with a scoping review to systematically map and characterize the literature on DeepSeek's medical applications. The aims were to (1) analyze publication trends, leading contributors, and research themes and (2) identify primary application domains, strengths, limitations, and future directions.

Methods: Following the framework by Arksey and O'Malley and the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines, a systematic search was conducted using PubMed, Web of Science, and Scopus from January 20, 2025, to November 30, 2025. Bibliometric analysis was then used to quantify publication trends, productivity, and research themes across 371 papers. The scoping review thematically synthesized the applications, strengths, and limitations of 353 original articles.

Results: The publication output showed a progressive increase, with China (n=163), Turkey (n=52), and the United States (n=48) as leading contributors. Keyword co-occurrence analysis formed 7 clusters; the 3 most frequent keywords were "large language model," "artificial intelligence," and "patient education." DeepSeek has shown promising yet preliminary performance across multiple domains, including patient education, clinical decision support, medical education, workflow optimization, and medical research. The evidence base remains predominantly low in quality, with 66.6% (235/353) of original articles classified as low-quality evidence, consisting largely of unvalidated benchmarking, simulated cases, and single-center retrospective analyses. Only 6.8% (24/353) of studies met the criteria to be considered high quality, and prospective randomized trials assessing patient-relevant outcomes were notably absent.

Conclusions: Publications on DeepSeek's medical applications increased progressively from January 2025 through November 2025, with China, Turkey, and the United States as the leading contributors. The scoping review found that DeepSeek has been evaluated across 5 domains (patient education, clinical decision support, medical education, workflow optimization, and research), with variable but often competitive performance relative to proprietary models. Strengths included readability, diagnostic accuracy in select specialties, cost-efficiency, and local deployability. Limitations included inconsistent cross-specialty performance, hallucinations, ethical concerns, data privacy issues, and regulatory gaps. The evidence base is predominantly low-quality and simulation-based, with few prospective trials or randomized controlled trials. These findings indicate that DeepSeek's clinical readiness varies, and future research should address prospective validation, multimodal capabilities, bias mitigation, human oversight, and equitable access.

(J Med Internet Res 2026;28:e93354) doi:10.2196/93354

XSL•FO
RenderX

KEYWORDS

DeepSeek; large language model; artificial intelligence in medicine; clinical decision support; medical education; scoping review; biomedical ethics; PRISMA

Introduction

The integration of artificial intelligence (AI), particularly large language models (LLMs), into medicine has prompted a paradigm shift in health care delivery, education, and research. LLMs, such as OpenAI's GPT series, have demonstrated considerable capabilities for processing complex medical data, supporting clinical decision-making, and improving patient communication. However, the widespread adoption of proprietary LLMs in clinical settings faces substantial challenges, including data privacy concerns, regulatory constraints, and limited adaptability to institutional requirements [1]. In this context, DeepSeek, an open-source LLM developed by Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co Ltd, has emerged as a promising alternative, distinguished by its customizability, cost-effectiveness, and alignment with data governance standards [1-3]. This model represents a significant advancement in AI, particularly for its sophisticated reasoning capabilities and its impact on AI research and applications.

DeepSeek's architecture, especially in reasoning-enhanced iterations such as DeepSeek-R1, incorporates innovative training approaches, including Group Relative Policy Optimization (GRPO). This rule-based reinforcement learning paradigm, which functions without task-specific supervised fine-tuning during the reasoning alignment phase and builds upon a pretrained base model, fosters emergent reasoning behaviors that are particularly valuable for complex medical reasoning tasks [4,5]. This open-weight nature enables local deployment, making it particularly attractive in health care settings where data security and privacy are paramount [1,6]. Since its release, DeepSeek and its associated intelligent agents have been implemented in multiple tertiary hospitals across China, resulting in measurable improvements in clinical and operational workflows, including patient follow-up, imaging analysis, and administrative automation [7-9]. Such real-world implementations underscore the potential for redefining AI-driven health care delivery.

The growing corpus of studies evaluating DeepSeek medical applications has revealed several strengths. In clinical diagnostics, DeepSeek-R1 achieved a diagnostic accuracy comparable to that of GPT-4 in complex clinicopathological cases [10]. In specialized areas, such as ophthalmology, it has exhibited diagnostic and management performance on par with OpenAI o1 while reducing token-related costs by approximately 15-fold [11]. Moreover, DeepSeek excels in Chinese-language medical contexts, outperforming ChatGPT at delivering prostate cancer radiotherapy information in Chinese and demonstrating superior results on Chinese medical licensing examinations [12,13]. Beyond clinical decision support, DeepSeek shows promise in medical education, patient communication, and administrative tasks, with documented deployments across multiple Chinese tertiary hospitals supporting applications ranging from imaging interpretation to automated administrative workflows [9]. However, these promising benchmarking results warrant further examination in real-world clinical settings, which are now emerging primarily in China.

The rapid integration of DeepSeek into clinical practice, particularly within Chinese hospital systems [2,9], underscores the necessity for a thorough evaluation of its applications, limitations, and future directions. The existing literature lacks a comprehensive assessment of publication trends and emerging research fronts in this rapidly evolving domain. Evidence remains fragmented across medical specialties, and the heterogeneous methodologies and outcomes limit a holistic understanding of the model's clinical utility, safety profile, and readiness for broader implementation. Therefore, a comprehensive synthesis of available evidence is essential to guide health care institutions, policymakers, and developers in evaluating DeepSeek's realistic capabilities, optimal deployment strategies, and associated risks.

To address this gap and systematically map the research landscape, this study adopted an integrated methodological approach that combined bibliometric analysis with a scoping review. Bibliometric analysis quantitatively characterizes the field at the macro level, examining publication trends over time, core authors and institutions, high-frequency keywords, and journal distributions. This enables the objective identification of research hot spots and evolutionary trajectories [14,15]. Simultaneously, a scoping review is a systematic methodology designed to map key concepts, evidence types, and knowledge gaps within a broad or emerging field. Rather than synthesizing evidence for definitive conclusions, it uses qualitative or descriptive methods to identify existing research themes, methodological characteristics, and underexplored areas, thereby clarifying the overall research landscape [16]. Given that literature on DeepSeek in medicine is growing rapidly and includes highly heterogeneous publications, such as proof-of-concept studies, preclinical research, preliminary clinical trials, and technical descriptions, a scoping review is more suitable than a systematic review for this context, as it focuses on comprehensively mapping the domain without mandating formal quality appraisal. The combination of these two methods leveraged their complementary strengths: Bibliometric analysis provides an objective, structured quantitative overview, while the scoping review delivers a nuanced, contextualized conceptual map. This integrated analysis provided a more powerful and multidimensional understanding of the field's scope, developmental dynamics, and future directions from both quantitative and qualitative perspectives.

Guided by this integrated approach, the study was structured as follows. First, a bibliometric analysis was conducted to examine relevant original articles and reviews, addressing the following questions: (1) What are the volume, growth trajectory, and geographic distribution of publications? (2) Which countries/regions, institutions, and authors are leading the

XSL•FO
RenderX

research? and (3) What are the key research themes and their evolution? Second, a scoping review was performed to critically evaluate the literature content, focusing on the following questions: What are the primary medical application domains of DeepSeek, and how do trends vary across different health care fields? Finally, the discussion synthesizes findings from both methods to highlight implementation challenges, identify major research gaps, and suggest future directions for the effective integration of DeepSeek into global health care systems.

Methods

Overview

This study used an integrated approach that combined bibliometric analysis and a scoping review to provide complementary insights. The bibliometric method examined the current application of DeepSeek in medicine from multiple dimensions, analyzed researcher characteristics and journal distributions, and identified research hot spots and trends. The bibliometric analysis was conducted based on the framework proposed by Cobo et al [17], following the guidelines for reporting bibliometric reviews of biomedical literature (BIBLIO) [18]. This scoping review systematically extracted and synthesized the applications, challenges, and future research directions of DeepSeek in medicine. The study was conducted according to the framework by Arksey and O'Malley [19] and reported following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines (Checklist 1) [20].

Databases, Search Strategy, and Screening Process

To ensure a comprehensive retrieval of the literature on the applications of DeepSeek in medicine, a systematic search was conducted on December 16, 2025, in PubMed, Web of Science Core Collection (WoSCC), and Scopus. The search strategy (Multimedia Appendix 1) used both controlled vocabularies (MeSH, Web of Science Categories, and SUBJAREA) and free-text terms tailored to each database to optimize retrieval.

Given that the public release of DeepSeek's reasoning model, DeepSeek-R1, on January 20, 2025 [21,22], marked the beginning of subsequent research into its applications, including in medicine, the search encompassed the period from January 20, 2025, to November 30, 2025.

To ensure comprehensive retrieval, the inclusion criteria were as follows: (1) studies investigating the application of DeepSeek in medicine, (2) document types limited to original articles and reviews for bibliometric analysis and original articles only for scoping review, (3) studies published in peer-reviewed academic journals, and (4) no language restrictions.

The exclusion criteria were as follows: (1) duplicate publications; (2) literature that proposed only speculative or hypothetical uses without substantive analysis or findings; (3) non-peer-reviewed journal items, including books, editorials, preprints, commentaries, conference abstracts, case reports, and retracted articles; and (4) studies with insufficient information for bibliometric analysis or whose full text was unavailable for in-depth content extraction during the scoping review.

After receiving professional training, two authors (HZ and DW) independently screened the titles and abstracts and excluded irrelevant studies based on the aforementioned criteria. The interrater agreement was almost perfect (Cohen κ=0.93). Any disagreements during screening were resolved through discussion or, when necessary, arbitration by a third reviewer (GW).

Bibliometric Analysis

The final bibliometric analysis included 371 papers. Full records of the selected publications were exported and stored in Excel 2021 (Microsoft Corp) and EndNote Desktop (Clarivate). Bibliographic metadata such as authors' names, affiliations, countries/regions, and keywords were standardized in a uniform format.

Excel 2021 was used to generate tables highlighting the top 10 authors, institutions, and countries/regions based on their publication output, whereas VOSviewer (version 1.6.19) was used for data visualization of bibliometric mapping, including keyword co-occurrence analysis. Keyword co-occurrence analysis examined the fundamental characteristics of keywords, such as their frequency and temporal evolution. This method helped identify research hot spots and track developmental trends within specialized fields. The three common types of visualizations used in the keyword co-occurrence analysis were the network, density, and overlay maps. In the network map, nodes represented keywords, and the connecting lines represented keyword co-occurrence relationships. The size of a node indicates its frequency, the thickness of a line represents the strength of co-occurrence, and the nodes are clustered together by color to reveal distinct research themes or subfields. The overlay map chronologically visualized the keyword trajectories by assigning chromatic codes corresponding to the computationally derived average publication years (APYs). The density map emphasizes the "research density" or concentration of keywords in the knowledge landscape. Areas with numerous closely located keywords appear as warm-colored regions, such as purple, indicating core well-developed research fronts. Cooler-colored areas such as blue or white represent sparser, potentially peripheral, or emerging topics. The centrality of keywords, which reflects their capacity to bridge different parts of the research network, was derived using CiteSpace (version 7.0.0).

Scoping Review

This scoping review included a total of 353 publications. A data extraction form was created using Excel to extract in-depth content from the papers. This form included items such as paper title, research objectives, key findings, research design types, DeepSeek's strengths, limitations and challenges, future recommendations, DeepSeek model version, quality tier, and application areas. It should be noted that, although quality assessment is not obligatory for scoping reviews, the methodological quality of all included studies was categorized into 3 tiers (high, moderate, and low) based on the criteria (Multimedia Appendix 2) in order to characterize the strength of the available evidence. Data extraction was conducted independently by 2 authors (HZ and DW). Both authors independently extracted data from all 353 included articles in

XSL•FO
RenderX

duplicate using the data extraction form created in Excel. After independent extraction, the 2 authors compared their results. Disagreements were resolved through discussion or by consulting a third author (GW) when consensus could not be reached. The extracted data (Multimedia Appendix 3) were then critically analyzed and organized thematically to address the research question, thereby mapping the key application areas of DeepSeek in medicine. The discussion section elaborates on the challenges, research gaps, and future work for the application of DeepSeek in the medical field.

Ethical Considerations

Since this study was a bibliometric and scoping review of previously published literature, ethical approval from an ethics committee is not required.

Results

Bibliometric Analysis of DeepSeek Applications in Medicine

A systematic search of PubMed, Scopus, and WoSCC yielded 371 publications on the application of DeepSeek in medicine for bibliometric analysis (Figure 1). Among these, the majority (363/371, 97.8%) were categorized as original articles, while the remaining (8/371, 2.2%) were reviews. In terms of publication languages, 358 papers were written in English, and 13 were written in Chinese.

Figure 1. The diagram depicting the paper selection process. WoSCC: Web of Science Core Collection.

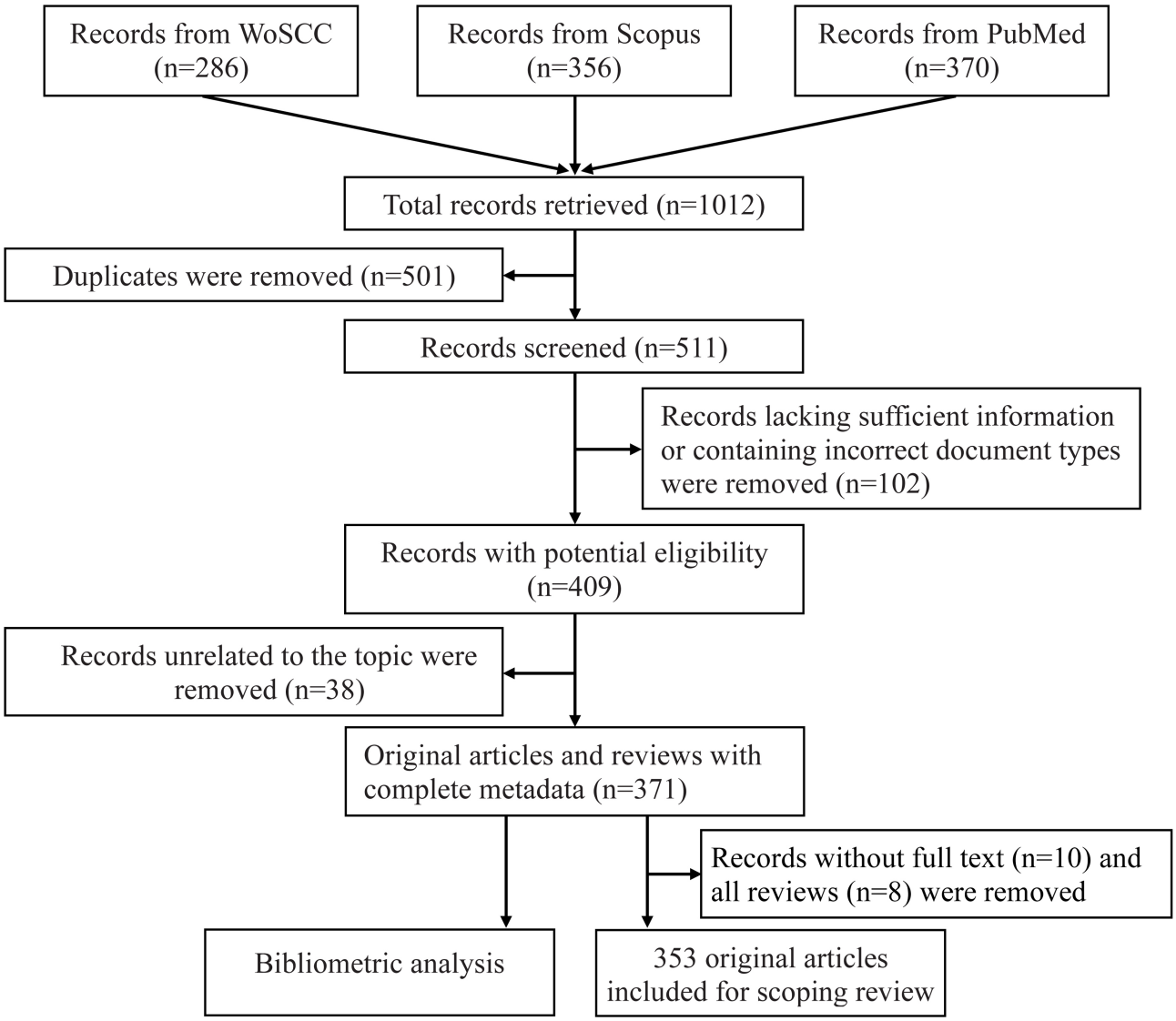

Monthly Publication Output

The monthly publication output increased progressively over time. From January to November 2025, the number of papers rose from 0 to 70, with the highest output (70 papers) observed in November (Figure 2).

XSL•FO
RenderX

Figure 2. Monthly count of publications on DeepSeek's medical applications identified in this review.

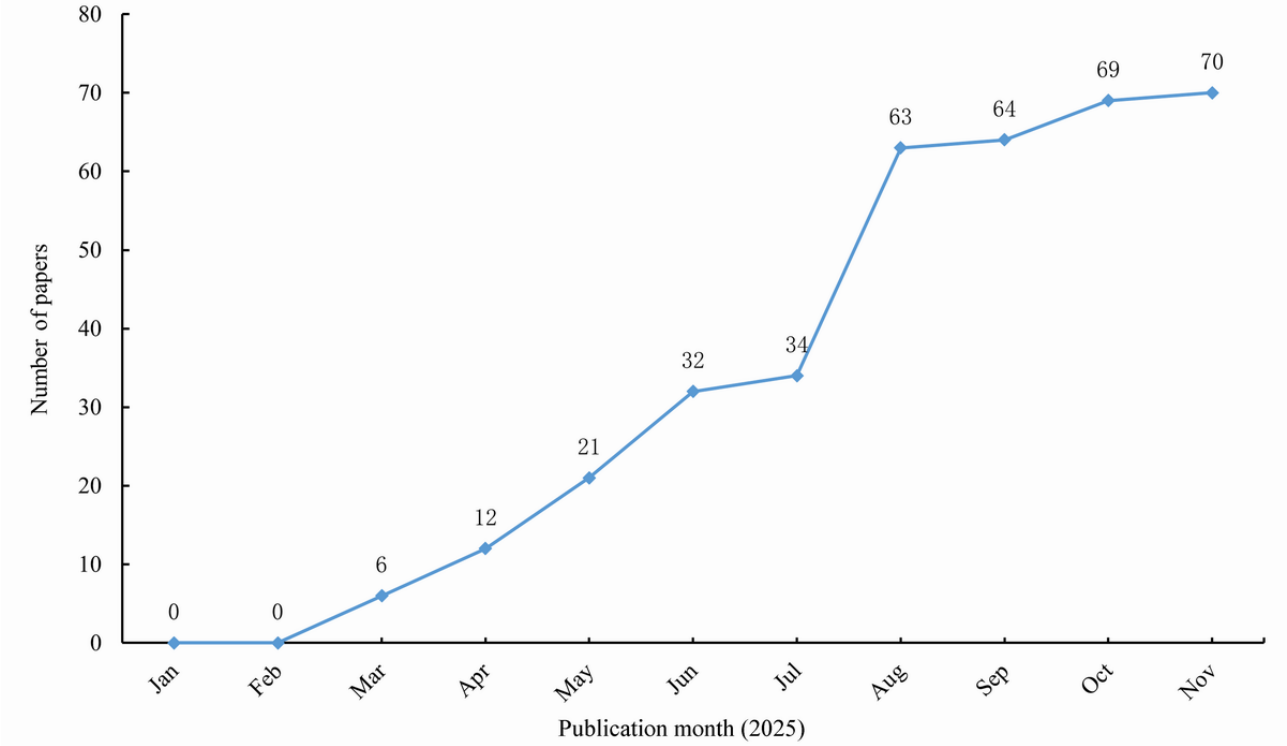

Analysis of Source Journals

Of the 216 journals that published papers on the applications of DeepSeek in medicine, 12 published more than 5 papers each. The 10 most active journals collectively contributed 90 publications, accounting for 24.3% (90/371) of the total output. *Cureus* was the most productive journal with 19 publications, followed by *Scientific Reports* (n=10), *BMC Oral Health* (n=9), *International Journal of Medical Informatics* (n=9), *BMC Medical Education* (n=8), *Frontiers in Artificial Intelligence* (n=7), *Frontiers in Public Health* (n=7), *JMIR Medical Informatics* (n=7), *Journal of Medical Internet Research* (n=7), and *Journal of Medical Systems* (n=7).

The Top 10 Authors, Institutions, and Nations/Regions Ranked by Publication Count

Table 1 presents the top 10 authors, institutions, and countries/regions ranked by their respective number of publications on the applications of DeepSeek in medicine.

XSL•FO
RenderX

Table . The top 10 authors, organizations, and countries ranked by the number of papers.

Rank	Authors[a]		Organizations[a]		Countries/Regions[a]	
	Name	Papers, n	Name	Papers, n	Name	Papers, n
1	Liu Y	6	Shanghai Jiao Tong University	16	China	163
2	Zhang J	5	Chinese Academy of Medical Sciences	10	Turkey	52
3	Li J	5	Sichuan University	10	United States	48
4	Wang J	5	Zhejiang University	9	Germany	24
5	Wang Y	5	Capital Medical University	9	India	23
6	Xu L	4	University of Health Sciences, Turkey	9	United Kingdom	20
7	Rozen WM	3	Southern Medical University	8	Italy	14
8	Cuomo R	3	Soochow University	7	Saudi Arabia	14
9	Marcaccini G	3	Sun Yat-sen University	7	Australia	9
10	Chen S	3	Tsinghua University	6	Canada	8

[a]These 3 categories are independent of each other.

Most Cited Papers on the Medical Applications of DeepSeek

Table 2 lists the 10 most-cited publications on the medical applications of DeepSeek: 9 were original articles, while 1 was a review [12,23-31].

XSL•FO
RenderX

Table . Top 10 most-cited publications on the medical applications of DeepSeek.

Rank	Authors	Publication date	Total citations, n	Research focus
1	Zhou et al [23]	June 2025	50	Comparative evaluation of DeepSeek and ChatGPT models
2	Deng et al [24]	May 2025	38	DeepSeek's advances, applications, and challenges across various domains, including health care
3	Kaygisiz and Teke [25]	April 2025	29	DeepSeek's diagnostic performance in oral pathologies
4	Rasool et al [26]	March 2025	28	DeepSeek's emotion-aware embedding fusion for responses
5	Yilmaz et al [27]	April 2025	16	Comparative performance of LLMs[a] on oral pathology multiple-choice questions
6	Marcaccini et al [28]	March 2025	16	DeepSeek and AI[b] in hand fracture management
7	Luo et al [12]	April 2025	16	DeepSeek versus ChatGPT in multilingual prostate cancer radiotherapy
8	Özcivelek and Özcan [29]	May 2025	15	Comparative evaluation of AI chatbots on dental and maxillofacial prostheses
9	Gültekin et al [30]	August 2025	14	Comparative evaluation of AI models for patient education
10	Seth et al [31]	March 2025	12	Evaluating DeepSeek and AI in hand surgery decisions

[a]LLMs: large language models.

[b]AI: artificial intelligence.

Keyword Co-Occurrence Analysis

A keyword co-occurrence analysis was performed to map predominant research hot spots. Synonyms were consolidated prior to analysis; specifically, "large language model(s)" was standardized as "large language model," and "generative artificial intelligence/AI" was standardized as "generative artificial intelligence." The top 10 keywords by frequency are listed in Table 3. Notably, "generative artificial intelligence" ranked seventh in frequency but third in centrality. From an initial set of 968 keywords, 41 occurring more than 4 times were included in the keyword co-occurrence analysis. These formed 7 well-defined clusters, visualized in the network map (Figure 3A).

The temporal overlay map (Figure 3B) illustrates the evolution of research focus, with keywords colored by their APYs. Purple nodes represent earlier themes, while crimson indicates more recent activity. Early research concentrated primarily on medical education. The keywords "retrieval-augmented generation" and "oncology" showed the highest APY, reflecting a rising interest in these areas.

The density map (Figure 3C) displays keywords according to their average frequency of occurrence. Crimson regions correspond to the most frequently occurring keywords, followed by blue and then white areas, in descending order.

XSL•FO
RenderX

Table . The top 10 keywords regarding DeepSeek's applications in medicine.

Rank	Keywords	Frequency of occurrence, n	Centrality
1	Large language model	227	1.00
2	Artificial intelligence	197	0.55
3	Patient education	30	0.02
4	Medical education	28	0.01
5	Clinical decision support	19	0.01
6	Machine learning	19	0.05
7	Generative artificial intelligence	19	0.07
8	Natural language processing	9	0.01
9	Prompt engineering	8	0.00
10	Diagnostic accuracy	8	0.03

XSL•FO
RenderX

Figure 3. Keyword co-occurrence analysis visualization using VOSviewer [32,33]: (A) network visualization, with keywords grouped into 6 distinct thematic clusters; (B) overlay map colored by the average publication year of each keyword, ranging from purple (earlier) to crimson (recent); and (C) density map based on keyword co occurrence frequency, where color intensity reflects occurrence rate: crimson (highest), blue (moderate), and white (lowest).

XSL•FO
RenderX

Summary of Extracted Data in the Scoping Review: Study Quality, Model Versions, Comparative Performance, and Documented Limitations

Of the 353 original articles, 24 (6.8%) met the criteria for high quality. These were primarily prospective evaluations and studies with external validation. A further 94 studies (94/353, 26.6%) were classified as moderate quality. The majority (235/353, 66.6%) were classified as low quality, reflecting the exploratory nature of the current evidence base, which is dominated by invalidated benchmarking using examination questions and single-center retrospective analyses.

Analysis of DeepSeek-specific versions revealed that DeepSeek-R1 was the most frequently studied (mentioned in 197 papers, 55.8% of the 353 articles), followed by DeepSeek-V3 (114/353, 32.3%) and unspecified versions of DeepSeek (61/353, 17.3%).

A total of 283 studies compared DeepSeek with other LLMs, primarily ChatGPT, in medical applications. Among these, 126 studies (126/283, 44.5%) reported positive results in which DeepSeek outperformed or showed significant advantages; 84 studies (84/283, 29.7%) reported neutral results with comparable performance, no statistically significant difference, or mixed strengths and limitations; and 73 studies (73/283, 25.8%) reported negative results in which DeepSeek underperformed relative to other models.

DeepSeek's primary weaknesses included inconsistent domain performance in 61 papers, incomplete answers in 47 papers, poor readability in 42 papers, and hallucinations in 38 papers. Ethical risks, though fewer in absolute count at 57 papers, were severe; specifically, non-maleficence was documented in 22 papers with potential patient harm, autonomy was documented in 15 papers with privacy and informed consent concerns, beneficence was documented in 8 papers with lack of empathy and impaired therapeutic relationship, and justice was documented in 12 papers highlighting bias and inequity. Other barriers reported in 55 papers further hindered clinical adoption.

Application Domains of DeepSeek in Medicine

Based on the scoping review of 353 full-text papers, the medical applications of DeepSeek can be summarized into the primary domains discussed in the following sections. Because a single study often evaluated DeepSeek in multiple domains, the sum of article counts across these domains exceeds 353.

DeepSeek in Patient Education and Communication

The applications of DeepSeek in patient education and communication were addressed in 105 articles. Among these, 91 were cross-sectional studies, 5 were descriptive studies, 4 were prospective studies including 1 randomized controlled trial (RCT), and the remaining 5 used other design types.

DeepSeek can generate patient-facing materials that are both readily comprehensible and clinically accurate. This capability has been empirically validated; for example, in generating patient education materials for spinal surgeries, DeepSeek-R1 achieved the lowest Flesch-Kincaid Grade Level scores, indicating content accessible to a broader audience including those with limited health literacy [23]. Similarly, in orthopedics, DeepSeek-R1 provided clearer and more easily understandable explanations of anterior cruciate ligament surgery than ChatGPT, which offered greater comprehensiveness but at a higher reading level [30]. This emphasis on linguistic accessibility is critical in patient-facing materials because improved readability enhances patient engagement, reduces anxiety, and supports informed decision-making [23,34]. Furthermore, DeepSeek has performed strongly in multilingual contexts, effectively generating patient education content in both Chinese and English, which is vital for serving diverse linguistic populations [12,35].

Although DeepSeek excels in readability, its responses sometimes lack comprehensive detail or sufficient citations of sources, and occasional inaccuracies or AI hallucinations have been noted [29,36,37]. Furthermore, some studies found that DeepSeek performed similarly to, or even less accurately than, ChatGPT when generating patient education materials [38,39].

DeepSeek in Clinical Decision Support and Treatment Planning

Of the 176 articles addressing DeepSeek in clinical decision support and treatment planning, 120 were cross-sectional studies, 22 were retrospective studies, 9 were prospective studies (including 2 RCTs), 2 were mixed-design studies, 14 were proof-of-concept studies, and the remaining 9 articles comprised expert consensus and other designs.

Regarding diagnostic accuracy, DeepSeek models have achieved notable results. In a dual-phase retrospective-prospective study classified as high methodological quality (n=300 liver lesions in the retrospective cohort and 126 liver lesions in the prospective cohort), DeepSeek-V3 demonstrated higher Liver Imaging Reporting and Data System (LI-RADS) classification accuracy than junior radiologists and achieved performance comparable with that of senior radiologists for hepatocellular carcinoma diagnosis [40]; however, this finding awaits replication in larger, multicenter settings. In a moderate-quality historical control study, DeepSeek-R1 demonstrated diagnostic accuracy comparable to that of GPT-4 in complex clinicopathologic cases [10]. In a low-quality cross-sectional study, Jiao et al [11] found that diagnostic accuracy in diagnosing corneal diseases varied significantly among LLMs (P=.001). GPT-4o achieved the highest accuracy (80%), while DeepSeek R1 achieved only 65%; both had accuracies that were significantly lower than that of human experts (92.5%; (P<.001).

For treatment planning, DeepSeek-V3 demonstrated statistically superior accuracy compared with ChatGPT-o1 in head and neck cancer management [41], and DeepSeek-R1 outperformed OpenAI o1 in diagnostic accuracy and next-step decision-making in ophthalmology [42]. These models have demonstrated strengths in specialized domains, including hand fracture management [28], urinary incontinence management [43], and postprostatectomy urinary incontinence guidelines [44], although they have limitations in complex scenarios. Notably, DeepSeek's clinical reasoning capabilities are enhanced through its reinforcement learning framework, which enables emergent reasoning patterns, such as self-reflection and verification [5], contributing to its strong performance in clinical decision support tasks. However, although DeepSeek shows

XSL•FO
RenderX

promising capabilities for clinical decision support, it cannot replace multidisciplinary tumor boards or human expertise, as it lacks contextual clinical judgment, physical examination capabilities, and the ability to negotiate complex trade-offs among specialists; instead, it streamlines clinical workflows by rapidly organizing patient data [41]. The integration of few-shot prompting has been shown to substantially enhance DeepSeek's accuracy in specialized tasks, such as Coronary Artery Disease Reporting and Data System (CAD-RADS) category assignment [42], suggesting that optimal prompt engineering is crucial for clinical implementation.

Overall, DeepSeek has emerged as a scalable tool to support treatment decisions, streamline workflows, and reduce diagnostic errors; however, integration requires careful validation and human oversight to mitigate risks.

DeepSeek in Medical Education and Benchmarking

Of 109 articles addressing the applications of DeepSeek in medical education and benchmarking, 93 were cross-sectional studies, 6 were retrospective studies, 5 were perspective studies, and 5 were descriptive studies.

On the Chinese National Medical Licensing Examination, DeepSeek-R1 achieved 92% accuracy, significantly outperforming ChatGPT-4o (87.2%) and demonstrating strength on low-difficulty questions [13]. Similarly, in the gastroenterology board examinations, both the base R1 model (77.1%) and search-augmented version (81.5%) surpassed the passing threshold and significantly outperformed the offline ChatGPT-3 (65.1%) and ChatGPT-4 (62.4%) models [45]. Cross-specialty comparisons revealed consistent patterns: In basic medical sciences, DeepSeek-R1 scored 78.33% alongside ChatGPT-4, whereas in clinical sciences, it scored 87.5%, demonstrating robust knowledge integration [46]. When evaluated against other reasoning-enhanced models on ophthalmology board-style questions, DeepSeek-R1 (72.5%) and its lighter variant R1-Lite (76.5%) performed competitively with OpenAI o1 Pro (83.4%), suggesting a balanced trade-off between performance and computational efficiency [47]. The model also demonstrated strong anatomical knowledge, achieving 89.2% accuracy on Turkish Dental Specialty Admission Exam anatomy questions, comparable with other major models, though below ChatGPT-4o's 98.6% [48]. These benchmark studies collectively indicate that DeepSeek provides a cost-effective, open-weight alternative for medical education, with utility in knowledge assessment and examination preparation. However, performance gaps persist in specialized domains and image-based questions, highlighting areas for future development and the continued need for human oversight in comprehensive medical education frameworks.

DeepSeek for Clinical Workflow Optimization

A total of 63 articles described DeepSeek for clinical workflow optimization, including 26 cross-sectional studies, 2 descriptive studies, 17 retrospective studies, 4 prospective studies, 10 proof-of-concept studies, and 4 articles with other study designs.

The integration of DeepSeek models into health care systems offers significant potential to enhance operational efficiency and streamline clinical workflows, primarily by automating routine and time-consuming tasks. A prominent example is the locally deployed closed-loop system powered by DeepSeek for quality control of electronic nursing documentation. This system implements a comprehensive framework spanning the real-time, final, and vertical dimensions of quality assurance. The results include a dramatic reduction in documentation omission rates from 7.19% to just 1.79%; a decline in logical inconsistencies from 9.35% to 0.72%; and the complete elimination of timeliness errors, which previously stood at 8.63%. Concurrently, the quality control time per record decreased by 3.2-fold, reallocating nursing efforts toward direct patient care [6].

In dyslipidemia management, DeepSeek, alongside Claude-3 and GPT-4, optimized guideline-based workflows across 30 standardized cases, boosting accuracy from 72% for physicians to 91% with AI. Integration with human experts further raised simulated low-density lipoprotein cholesterol target attainment to 92%, demonstrating its utility in minimizing guideline deviations while enhancing workflow efficiency [49]. However, one moderate-quality study found that DeepSeek R1 achieved an accuracy of only 48.4% in a noncritical emergency department triage task, which is significantly lower than that of another LLM, Gemini 2.0 flash (73.8%) [50].

The large-scale deployment of DeepSeek across nearly 90 Chinese tertiary hospitals has reportedly increased patient follow-up efficiency 40-fold, marking a transformative impact on hospital administration and clinical workflow automation [9]. By managing labor-intensive tasks with high consistency and speed, DeepSeek enables a paradigm shift from reactive to proactive operational governance. This transition enabled health care professionals to focus their expertise on more complex clinical decision-making responsibilities.

Medical Research and Data Analysis

Medical research and data analysis were mentioned in 73 articles. Among these, 41 had a cross-sectional design, 6 were descriptive studies, 2 were perspective studies, 9 were proof-of-concept studies, 9 were retrospective studies, 1 had a mixed design, and the remaining 5 used other design types.

DeepSeek models have demonstrated significant utility in accelerating and refining medical research and data analysis workflows. DeepSeek facilitates the reading of medical literature, information extraction, and screening. Several studies have developed AI-powered screening tools using DeepSeek to identify relevant studies for systematic reviews, reporting high accuracy and a significant reduction in manual workload [51-53]. For example, the LitAutoScreener tool, which integrates DeepSeek, achieved high accuracy and significantly improved screening efficiency, reducing the processing time to seconds per article [51]. Similarly, other evaluations have confirmed that DeepSeek-based tools can reduce manual workload while maintaining high recall rates in literature screening for meta-analyses [53]. In fields such as aging research, DeepSeek-R1 is part of a multi-LLM ensemble that successfully extracts protocol details from clinical trial records, doubling the yield of conventional search methods and achieving expert-level accuracy for core data points [54]. Second, DeepSeek assists with generating and refining research topics

XSL•FO
RenderX

and study designs. It helps researchers analyze cutting-edge trends, funding guidelines, and successful grant applications, thereby validating the novelty of the proposed research questions [3]. For instance, DeepSeek-R1 has been used to explore novel research ideas and generate systematic review topics in fields such as oral and maxillofacial surgery [55]. Similarly, in biomedical research, DeepSeek models show promise in extracting structured pre-analytical variability data from the scientific literature, facilitating standardized reporting and systematic evaluation [56]. Furthermore, DeepSeek serves as a valuable tool for peer review and for critiquing research proposals. Its capacity to generate high-quality evidence-based responses enables a preliminary assessment of a proposal's feasibility and soundness. This function is particularly beneficial in multidisciplinary contexts where the model's ability to synthesize information from diverse sources significantly enhances the evaluation process [57,58]. Third, DeepSeek demonstrated substantial potential as an assistant for drafting, editing, and refining the content of medical research papers. Its capabilities span various domains of medical research and practice, making it a versatile tool for enhancing the quality and efficiency of academic writing. The model's proficiency at generating structured, clear, and comprehensible content is particularly valuable in medical research, where precision and clarity are paramount [59].

Other Application Domains

In other application domains, 25 articles were identified, comprising 18 cross-sectional studies, 2 perspective articles, 2 descriptive studies, and 3 proof-of-concept studies.

Beyond the primary domains discussed, DeepSeek has been explored in several niche but critical areas, including treatment outcome prediction, drug development assistance, and suicide risk prediction. Instead of reactive question-answering, DeepSeek is integrated into predictive analytics platforms. It can proactively flag at-risk patients, suggest personalized screening intervals, and predict individual responses to therapies based on electronic health records and real-time data [60,61]. In nasopharyngeal carcinoma, DeepSeek-V3-0324 demonstrated superior performance in treatment response evaluation compared with ChatGPT-4o-latest (96.5% vs 82.9%) and showed stronger agreement with expert annotations [62].

In drug discovery, DeepSeek aids with predicting drug-drug interactions and molecular property modeling, achieving superior performance in regression and classification tasks critical to drug discovery [63,64].

The model's chain-of-thought enabled analysis of factors associated with correct predictions, such as substance abuse and age-related comorbidities. This application underscores DeepSeek's potential for mental health risk assessment, though further validation is needed [65].

Discussion

Main Findings

This integrated bibliometric and scoping review provided a comprehensive early-stage mapping of the rapidly evolving research landscape concerning DeepSeek's applications in medicine. This field is characterized by explosive growth, global engagement, and exploration across a remarkably diverse spectrum of clinical and operational domains. The findings collectively underscore DeepSeek's emergence not merely as another LLM but as a potent, open-weight contender with specific capabilities that address critical needs in modern health care, including cost-effectiveness, linguistic accessibility, and scalability.

Bibliometric data showed a research frontier that has been intensively explored. The increased publication output regarding applications of DeepSeek in medicine is clear. Our results align with those of an analysis of the global research profile of another LLM, ChatGPT, conducted by Alessandri-Bonetti et al [66], who revealed explosive growth in publications during the first 7 months after its release. This pattern is also consistent with a broader LLM systematic review by Chen et al [67], which reported that, between January 2022 and September 2025, approximately 3.2 clinical LLM studies were published per day, with a linear increase of 7.04 studies per month following the release of ChatGPT. Notably, DeepSeek was not included in the analysis by Chen et al [67], underscoring the gap and the need for our focused review.

The geographical and institutional productivity led by China, followed by Turkey and the United States, reflects widespread international interest of DeepSeek's potential, with major academic medical centers driving early investigations. Papers on DeepSeek's applications in medicine have been published in various journals, ranging from well-known open-access journals such as *Cureus* and *Scientific Reports* to professional medical informatics and medical education journals such as the *Journal of Medical Internet Research*. This publication pattern indicates that the research reaches both broad scientific and specialized clinical audiences. Keyword co-occurrence analysis effectively identified the core themes of this research trend. The temporal overlay, which revealed a shift from foundational medical education topics toward more specialized areas such as "retrieval-augmented generation" and "oncology," illustrates the field's rapid maturation and deepening focus. Synthesizing the scoping review findings, DeepSeek as a medical tool initially gained attention for its strength in democratizing medical information. For instance, in patient education, it can generate outputs with higher readability than its counterparts, such as ChatGPT.

Perhaps the most striking finding is that DeepSeek has demonstrated competitive and sometimes superior performance compared with existing proprietary models in clinical decision support tasks. The bibliometric analysis revealed that "clinical decision support" formed the largest cluster, while the scoping review further indicated that these studies primarily focused on three specific tasks: "aiding diagnosis," "differential diagnosis," and "treatment plan formulation." The evidence that DeepSeek-V3 can match senior radiologists at specialized diagnostic classifications or that DeepSeek-R1 rivals GPT-4 and OpenAI o1 in diagnostic accuracy across ophthalmology and complex clinicopathological cases challenges the assumption that superior capability is the exclusive domain of closed, commercial models. This "performance parity" achieved through an open-weight architecture has profound implications.

XSL•FO
RenderX

Specifically, it suggests a pathway toward breaking the monopoly of advanced AI in clinical support, potentially fostering innovation, reducing costs, and allowing for better adaptation to local health care contexts and linguistic needs.

The utility of this model in medical education and benchmarking further supports its position as a disruptive and cost-effective tool [68]. For institutions and learners worldwide, particularly in resource-constrained settings, DeepSeek offers a viable, high-quality alternative for exam preparation, simulation, and curriculum development, potentially lowering the barriers to accessing advanced medical training aids.

Beyond its direct clinical and educational applications, this review highlighted DeepSeek's transformative potential across broader health care operations. Documented case studies have demonstrated reductions in documentation error rates in nursing and lower specimen return rates in gynecological examinations and enabled large-scale patient follow-up. By automating a vast array of low-complexity tasks, DeepSeek can free human resources to provide higher quality care and reduce systemic inefficiencies across the health care continuum.

Of the 353 papers included in this scoping review, only 6.8% (24/353) met the criteria for high quality, whereas the majority (235/353, 66.6%) were classified as low quality, consisting predominantly of invalidated benchmarking using examination questions, single-center convenience samples, and proof-of-concept studies. This distribution reflects a critical gap in the current literature: The rapid proliferation of DeepSeek in medicine has been accompanied by an abundance of exploratory studies with limited external validity. Although such benchmarking studies offer valuable insights into the model's technical capabilities and serve as initial performance indicators, they do not directly inform real-world diagnostic accuracy, patient safety, or clinical utility [69].

In head-to-head comparisons with other LLMs, DeepSeek demonstrated predominantly favorable or comparable performance: Positive outcomes (126/283, 44.5%) were more frequent than negative ones (73/283, 25.8%), and a substantial proportion of studies (84/283, 29.7%) showed no clear superiority of either model. However, because these results derived predominantly from low-quality (235/353, 66.6%) or moderate-quality evidence, with only 6.8% (24/353) meeting high methodological standards, performance claims should be considered preliminary and hypothesis-generating rather than definitive. Clinically, it excels in open-source accessibility, low cost, readability, Chinese language proficiency, and structured reasoning; nonetheless, limitations, including occasional inaccuracies, lower reliability in certain tasks, and the absence of prospective clinical trials, necessitate continued validation and human oversight.

Comparison With Prior Reviews on Other LLMs in Medicine

To contextualize the novel and distinct contributions of our work, we compared this review with existing reviews of other LLMs in medicine, such as ChatGPT, GPT-4, LLaMA, and Gemini. Several prior reviews have documented the rapid adoption of proprietary LLMs in health care, highlighting their utility in clinical reasoning, medical education, and patient communication [66,70,71]. However, most existing reviews have primarily focused on closed-source models, which are characterized by limited transparency, restricted capacity for local deployment, and substantial cost barriers. These limitations hinder their scalability and reduce their adaptability across diverse institutional settings. In contrast, this review specifically focused on DeepSeek, an open-weight LLM, and identified several distinctive features that differentiate it from the patterns reported in previous LLM reviews.

First, methodologically, we combined bibliometric analysis with a scoping review to provide both quantitative mapping of research trends and qualitative synthesis of applications and challenges of DeepSeek in medicine, a dual approach rarely applied in prior LLM reviews, which have tended to rely on either bibliometric or narrative synthesis alone [66,71].

Second, geographically, the research landscapes differ substantially. For ChatGPT, early publications were predominantly led by institutions in the United States and Europe, with a wide distribution across high-income countries [66,70]. In contrast, our analysis identified China as the dominant contributor to DeepSeek medical research (163 papers), followed by Turkey and the United States. This pattern aligns with DeepSeek's country of origin and its rapid deployment across Chinese tertiary hospitals [2,9]. Notably, the early and substantial involvement of Turkish researchers (52 papers) in DeepSeek research is a distinctive feature not observed in early ChatGPT literature.

Third, previous reviews focused predominantly on proprietary models such as ChatGPT, GPT-4, LLaMA, and Gemini. In contrast, our study addressed a significant gap by examining an open-source alternative with distinct architectural advantages and greater deployment flexibility. In terms of real-world deployment, the deployment of DeepSeek across nearly 90 tertiary hospitals in China has resulted in measurable improvements in workflow efficiency and documentation quality. This scale of implementation has not been reported in similar reviews of other LLMs, which have largely focused on simulated or benchmarking studies [6,9]. In terms of application areas, prior work on ChatGPT and other proprietary LLMs identified medical education, clinical decision support, and patient communication as core areas [70,71]. Our keyword co-occurrence analysis confirmed that these are also central themes for DeepSeek. However, DeepSeek's open-weight architecture introduces distinctive features not emphasized in proprietary LLM reviews: on-premises deployability, data privacy, cost-effectiveness, and superior performance in Chinese-language medical tasks. These features represent unique contributions of DeepSeek to the medical LLM landscape and are not simply typical of any newly introduced LLM. Regarding performance and utility, our findings demonstrated that DeepSeek achieved competitive or superior performance compared with proprietary models in clinical diagnostics, medical licensing examinations, and patient education while substantially reducing costs, advantages that prior reviews have identified as critical unmet needs in AI integration [71,72].

XSL•FO
RenderX

Challenges in the Applications of DeepSeek in Medicine

Guided by an ethical framework, the efficacy and safety of any medical intervention must be carefully calibrated in modern medical practice [73]. As aforementioned, DeepSeek demonstrates significant potential for enhancing medical workflows, medical education, and research. However, its application faces numerous challenges in terms of effectiveness and safety, including accuracy issues, data privacy concerns, ethical uncertainties, and diverse global regulations governing AI.

Accuracy and Variable Performance Across Medical Domains and Specialties

Although DeepSeek has demonstrated diagnostic accuracy comparable to that of specialist clinicians and proprietary models in certain areas [40,74-76], its overall efficacy remains inconsistent [42,77,78]. The model exhibits strong zero-shot and few-shot learning capabilities in general tasks; however, the rapid evolution of medical knowledge necessitates continuous pretraining on extensive volumes of high-quality, domain-specific data. In data-scarce specialties, particularly those lacking sufficient fine-tuning datasets, DeepSeek often fails to effectively acquire new features and patterns, leading to model hallucinations, defined as the generation of seemingly plausible but factually incorrect or unsupported information [36,79]. Such limitations are particularly severe in domains involving rare diseases and complex, nonclassical clinical scenarios, where available pretraining data are often insufficient and clinically unvalidated [36,37,80,81]. Furthermore, as a fundamentally text-based model, DeepSeek exhibits inherent limitations in processing specialized nontextual medical data, such as medical images, complex laboratory metrics, and genomic data [74,82-84]. These constraints collectively contribute to inconsistent model performance across specific medical domains and hinder its generalization.

Ethical and Safety Risks

The integration of DeepSeek into medical practice raises ethical challenges that implicate all 4 foundational principles of biomedical ethics, namely autonomy, nonmaleficence, beneficence, and justice, which were originally proposed by Beauchamp and Childress in 1979 [85].

Autonomy: Challenges to Patient Self-Determination and Informed Consent

The application of DeepSeek in medicine may undermine the principle of autonomy in medical ethics. As an open-source model, DeepSeek can be deployed on-premises in a hospital environment, which facilitates compliance with data privacy requirements [1,9,81]. However, its broader adoption is complicated by varying regulatory frameworks across regions, such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) [3]. The Italian data protection authority, for instance, has restricted DeepSeek over concerns that its data handling methods fail to meet the strict privacy rules of the European Union [81]. Although techniques such as chain-of-thought have enhanced the interpretability of decision-making, the model's fundamental "black-box" nature persists, posing practical challenges to informed consent in clinical applications [60,86-90].

Nonmaleficence: Risks of Novel and Amplified Harms

The rapid, cost-effective integration of DeepSeek in Chinese hospitals underscores a central paradox in medicine: how to seize the opportunity for transformative innovation while mitigating the risks of undue haste and still upholding the principle of "first, do no harm" [2]. However, this model may provide overly definitive recommendations, potentially suggesting unnecessary tests or harmful treatments without adequate contextual warnings [91,92]. If clinicians over-rely on AI outputs, effectively delegating core cognitive tasks such as comprehensive analysis, differential diagnosis, and clinical judgment to the machine, it may lead to the erosion of clinical skills and their independent clinical reasoning. Furthermore, however data-driven its suggestions may be, DeepSeek may lack the nuanced and holistic understanding of a patient's psychosocial context that an experienced physician integrates. Collectively, these issues challenge the ethical principle of nonmaleficence.

Beneficence: The Challenge of Defining and Delivering "Good"

The principle of beneficence obligates health care providers to act in ways that promote patients' well-being and enhance clinical outcomes [93]. However, an emphasis on AI-driven efficiency may unintentionally marginalize the irreplaceable human dimensions of medicine, such as empathy, compassion, and the therapeutic physician-patient relationship. Although systems like DeepSeek are adept at optimizing measurable, data-informed endpoints, the concept of "good" in medical practice encompasses psychosocial, spiritual, and qualitative aspects of care that resist easy quantification [89,94]. Overreliance on algorithmic pathways designed to maximize metrics neglects the holistic components of beneficence [95]. Consequently, the physician's role as a compassionate interpreter of illness, which lies at the heart of medical beneficence, may be subordinate to the pursuit of algorithmic efficiency.

Justice: Amplifying Inequities in Algorithmic Health Care

The principle of justice concerns fair and equitable distribution of health care benefits and burdens. Despite the use of data preprocessing techniques and fairness-aware algorithms, DeepSeek can still perpetuate and potentially amplify societal or health care biases present in its historical medical training data, including the underdiagnosis of certain conditions within specific demographic groups, thereby harming marginalized populations [80,88,96]. Furthermore, because DeepSeek's training framework is primarily optimized for English and Chinese, it carries inherent lexical and cultural biases that may limit its applicability to global health care contexts [12,35,97]. Additionally, the benefits of advanced AI, such as DeepSeek, are likely to accrue disproportionately to well-resourced tertiary-care urban hospitals equipped with the necessary infrastructure and specialized personnel for local deployment.

XSL•FO
RenderX

Such unequal access exacerbates existing health disparities across regions and socioeconomic groups.

Other Challenges

In addition to challenges such as accuracy, variable performance across medical domains and specialties, and medical ethics and safety issues, the application of DeepSeek in medicine faces other obstacles, including the redesign of clinical workflows, delineation of liability, regulatory lag, and trust and adoption. The deployment of DeepSeek challenges some clinicians' work habits and creates a demand for professionals who understand both clinical practice and AI. A shortage of talent limits its wider adoption. When errors in DeepSeek-assisted decision-making lead to medical incidents, how should legal responsibility be defined? Should it fall on the operating physician, the hospital that adopted the AI, or the model developers? Currently, global regulations in this field generally lag, and this uncertainty greatly dampens hospitals' willingness to implement such technologies. Trust remains another challenge; although DeepSeek is easy to use, concerns about risks affect its acceptance [87].

Future Work in the Applications of DeepSeek in Medicine

Based on the aforementioned challenges, future research and development should prioritize the directions highlighted in the following sections to advance the reliable, ethical, and equitable integration of DeepSeek into medical practice.

From Benchmarking to Clinical Validation: Prospective and Pragmatic Studies

The current evidence base is dominated by low-quality, simulation-based studies. Future work should move beyond examination-style benchmarks and retrospective analyses toward prospective, multicenter, and pragmatic clinical trials. Specifically, RCTs are urgently needed to compare DeepSeek-assisted care against standard practice using both proximal performance metrics, such as diagnostic accuracy, and patient-relevant outcomes, including treatment adherence, adverse events, and quality of life [98,99]. Such trials should also evaluate human-AI interaction models, for example, human-in-the-loop versus fully automated approaches, to determine the optimal balance between efficiency and safety [100,101]. Furthermore, real-world implementation science frameworks should be applied to assess scalability, usability, and unintended consequences across diverse health care settings.

Strengthening Governance, Explainability, and Safety

To address ethical and regulatory gaps, future work should co-develop clinically interpretable explainability methods tailored to DeepSeek's reasoning architecture. Techniques such as structured audit trails, uncertainty quantification, and natural language rationales can support informed consent and clinician oversight [89,102]. On the governance front, clear liability and accountability frameworks are required to delineate responsibilities among developers, health care institutions, and clinicians when AI-assisted errors occur [88,96]. Additionally, the "human-in-command" principle, which mandates that DeepSeek's recommendations serve as decision support rather than replacement for clinician judgment, should be embedded into clinical workflows and professional guidelines [98,103]. As articulated in the concept of AI-assisted medicine introduced by Wang et al [104], a discipline that uses AI technologies to assist with disease research, prevention, diagnosis, and treatment as well as to promote health maintenance, clinicians must retain ultimate decision-making authority and accountability [100,101]. This conceptual foundation reinforces that AI remains a tool to augment, not supplant, human expertise.

Mitigating Bias and Promoting Equitable Access

Despite DeepSeek's open-weight advantage, bias and inequity remain critical challenges. Future research should conduct systematic bias audits across demographic subgroups such as sex, socioeconomic status, and ethnicity using multi-institutional and multilingual datasets [105,106]. To avoid perpetuating health care disparities, developers should expand medically validated support beyond English and Chinese to other major world languages while adapting outputs to local clinical guidelines and cultural contexts [12,107].

Redefining Medical Education and Workforce Development

The rapid adoption of DeepSeek demands a parallel evolution in medical curricula. Future educational interventions should cultivate "AI literacy": the ability to critically appraise AI-generated recommendations; recognize hallucinations and bias; and integrate AI outputs with compassionate, patient-centered communication [98,108]. Institutions should develop interdisciplinary training programs that bridge clinical practice and data science to build a workforce capable of deploying, auditing, and improving medical AI systems. Finally, professional societies should establish certification and continuing education standards for AI-augmented clinical practice.

Unexplored Domains and Long-Term Monitoring

Most current research focuses on diagnosis, medical education, and workflow efficiency, leaving prevention and long-term care underexplored. Future investigations should prioritize disease prevention, population health management, and long-term care [103,109]. Additionally, postdeployment surveillance systems should be established to monitor real-world performance, detect emergent harms, and enable continuous model improvement, closing the loop from evidence generation to sustained safe implementation [9,90].

Limitations of the Study

Several limitations of this study should be considered when interpreting the findings. First, the review covered literature published over a relatively short and recent timeframe. Consequently, the observed surge in publications may reflect early enthusiasm rather than sustained scientific progress. Second, although a language-agnostic search strategy was used, most included studies were published in English, with only a small number (n=13) in Chinese. This linguistic imbalance, coupled with the predominance of contributions from researchers based in China, indicates a notable geographical concentration of the available evidence. As a result, the findings may not be directly generalizable to health care systems operating within

XSL•FO
RenderX

different regulatory, cultural, or infrastructural contexts. Third, the included studies exhibited substantial heterogeneity in methodologies, medical specialties, evaluation metrics, comparator models, and DeepSeek model versions—for example, R1 versus V3, which differ in parameter counts, training data, and reasoning depth. This variability precluded quantitative synthesis of outcomes and hindered direct cross-study comparisons. Although we reported version-specific findings where available, direct comparisons of performance should be interpreted with caution. Future research should adopt standardized version reporting and benchmark against fixed model checkpoints to enhance comparability and reproducibility. Finally, much of the evidence is derived from benchmarking studies, simulated cases, or retrospective analyses, with a formal quality appraisal showing that 66.6% (235/353) of included original articles were of low quality and only 6.8% (24/353) met the criteria to be considered high quality. Prospective clinical trials or RCTs assessing DeepSeek's impact on tangible patient health outcomes in real-world clinical settings remain notably scarce. Consequently, the overall quality of the evidence base is inherently preliminary, and the reviewed corpus carries a high risk of bias. The reported strengths of DeepSeek should be interpreted with caution, as these findings predominantly derive from low-quality, controlled, nongeneralizable settings.

Conclusion

This integrated bibliometric and scoping review synthesized the available evidence on DeepSeek's applications in medicine. The bibliometric analysis revealed a progressive increase in publication output from January 2025 through November 2025, with China, Turkey, and the United States as the leading contributors. Keyword co-occurrence analysis formed 7 clusters; the 3 most frequent keywords were "large language model," "artificial intelligence," and "patient education."

The scoping review found that DeepSeek has been evaluated across 5 primary application domains: patient education and communication, clinical decision support and treatment planning, medical education and benchmarking, clinical workflow optimization, and medical research and data analysis. In these domains, DeepSeek demonstrated variable but often competitive performance compared with proprietary models, with documented strengths in readability of patient education materials, diagnostic accuracy in select specialties, cost-efficiency, and local deployability. Nevertheless, it should be noted that most included studies were of moderate or low quality, and the evidence base is predominantly composed of benchmarking and simulation studies, with a notable scarcity of prospective clinical trials or RCTs assessing patient-relevant outcomes. Additionally, the review identified consistent limitations, including variable performance across medical specialties, model hallucinations, ethical concerns, data privacy challenges, and regulatory gaps. Future integration will require robust prospective clinical validation, expansion of multimodal capabilities, bias mitigation strategies, human-in-the-loop governance frameworks, and equitable access strategies.

Acknowledgments

The authors would like to acknowledge Editage for English language editing [110].

Funding

This work was supported by the Science and Technology Project of Jinan Health Commission (grant 2020-3-02).

Data Availability

All data generated or analyzed during this study are included in this published article and its multimedia appendices.

Authors' Contributions

Conceptualization: GW
Data curation: HZ, DW
Formal analysis: YX, SH
Funding acquisition: GW
Methodology: HZ, DW
Resources: HZ, DW, YX, SH
Software: HZ, DW, GW
Supervision: GW
Visualization: HZ, SH
Writing – original draft: HZ, DW, YX, SH
Writing – review & editing: GW
All the authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Search strategy.
[DOC File, 48 KB - jmir_v28i1e93354_app1.doc]

Multimedia Appendix 2
Quality assessment criteria for studies included in the scoping review.
[DOC File, 33 KB - jmir_v28i1e93354_app2.doc]

Multimedia Appendix 3
The extracted data for scoping review.
[XLSX File, 246 KB - jmir_v28i1e93354_app3.xlsx]

Checklist 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist.
[DOCX File, 73 KB - jmir_v28i1e93354_app4.docx]

References

1. Sandmann S, Hegselmann S, Fujarski M, et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. Nat Med 2025 Aug;31(8):2546-2549. [doi: 10.1038/s41591-025-03727-2] [Medline: 40267970]
2. Zeng D, Qin Y, Sheng B, Wong TY. DeepSeek's "low-cost" adoption across China's hospital systems: too fast, too soon? JAMA 2025 Jun 3;333(21):1866-1869. [doi: 10.1001/jama.2025.6571] [Medline: 40293869]
3. MohanaSundaram A, Sathanantham ST, Ivanov A, Mofatteh M. DeepSeek's readiness for medical research and practice: prospects, bottlenecks, and global regulatory constraints. Ann Biomed Eng 2025 Jul;53(7):1754-1756. [doi: 10.1007/s10439-025-03738-7] [Medline: 40272697]
4. Jin I, Tangsrivimol JA, Darzi E, et al. DeepSeek vs. ChatGPT: prospects and challenges. Front Artif Intell 2025;8:1576992. [doi: 10.3389/frai.2025.1576992] [Medline: 40612384]
5. Guo D, Yang D, Zhang H, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature New Biol 2025 Sep;645(8081):633-638. [doi: 10.1038/s41586-025-09422-z] [Medline: 40962978]
6. Lv J, Xu Y, Jiang M, et al. A DeepSeek-powered locally deployed closed-loop system for enhancing quality control in electronic nursing documentation: development and clinical validation. J Am Med Inform Assoc 2025 Oct 1;32(10):1526-1532. [doi: 10.1093/jamia/ocaf109] [Medline: 40668938]
7. Wang Y, Tan W, Cheng S, et al. Large language model agent for managing patients with suspected hypertension. Hypertension 2026 Jan;83(1):212-224. [doi: 10.1161/HYPERTENSIONAHA.125.25305] [Medline: 41064862]
8. Miao Y, Wen J, Luo Y, Li J. MedARC: Adaptive multi-agent refinement and collaboration for enhanced medical reasoning in large language models. Int J Med Inform 2026 Feb;206:106136. [doi: 10.1016/j.ijmedinf.2025.106136] [Medline: 41109093]
9. Chen J, Miao C. DeepSeek deployed in 90 Chinese tertiary hospitals: how artificial intelligence is transforming clinical practice. J Med Syst 2025 Apr 24;49(1):53. [doi: 10.1007/s10916-025-02181-4] [Medline: 40272650]
10. Chan L, Xu X, Lv K. DeepSeek-R1 and GPT-4 are comparable in a complex diagnostic challenge: a historical control study. Int J Surg 2025;111(6):4056-4059. [doi: 10.1097/JS9.0000000000002386]
11. Jiao C, Rosas E, Asadigandomani H, et al. Diagnostic performance of publicly available large language models in corneal diseases: a comparison with human specialists. Diagnostics (Basel) 2025 May 13;15(10):1221. [doi: 10.3390/diagnostics15101221] [Medline: 40428214]
12. Luo PW, Liu JW, Xie X, et al. DeepSeek vs ChatGPT: a comparison study of their performance in answering prostate cancer radiotherapy questions in multiple languages. Am J Clin Exp Urol 2025;13(2):176-185. [doi: 10.62347/UIAP7979] [Medline: 40400997]
13. Wu J, Wang Z, Qin Y. Performance of DeepSeek-R1 and ChatGPT-4o on the Chinese National Medical Licensing Examination: a comparative study. J Med Syst 2025 Jun 3;49(1):74. [doi: 10.1007/s10916-025-02213-z] [Medline: 40459679]
14. Huang Y, Wan Y, Chen J, Qin M, Wang J, Liang H. Knowledge mapping of biomarkers in amyotrophic lateral sclerosis: a comprehensive bibliometric and visual analysis. Neurodegener Dis Manag 2026 Apr;16(2):191-207. [doi: 10.1080/17582024.2025.2554525] [Medline: 40905501]
15. Chen X, Yang Y, Yun D, et al. Current status and solutions for AI ethics in ophthalmology: a bibliometric analysis. NPJ Digit Med 2025 Oct 2;8(1):594. [doi: 10.1038/s41746-025-01976-6]
16. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5:69. [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
17. Cobo MJ, López-Herrera AG, Herrera-Viedma E, Herrera F. Science mapping software tools: review, analysis, and cooperative study among tools. J Am Soc Inf Sci 2011 Jul;62(7):1382-1402 [FREE Full text] [doi: 10.1002/asi.21525]

XSL•FO
RenderX

18. Montazeri A, Mohammadi S, M Hesari P, Ghaemi M, Riazi H, Sheikhi-Mobarakeh Z. Preliminary guideline for reporting bibliometric reviews of the biomedical literature (BIBLIO): a minimum requirements. Syst Rev 2023 Dec 15;12(1):239. [doi: 10.1186/s13643-023-02410-2] [Medline: 38102710]
19. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
20. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
21. Gibney E. China's cheap, open AI model DeepSeek thrills scientists. Nature New Biol 2025 Feb 6;638(8049):13-14. [doi: 10.1038/d41586-025-00229-6]
22. Conroy G, Mallapaty S. How China created AI model DeepSeek and shocked the world. Nature New Biol 2025 Feb 13;638(8050):300-301. [doi: 10.1038/d41586-025-00259-0]
23. Zhou M, Pan Y, Zhang Y, Song X, Zhou Y. Evaluating AI-generated patient education materials for spinal surgeries: comparative analysis of readability and DISCERN quality across ChatGPT and DeepSeek models. Int J Med Inform 2025 Jun;198:105871. [doi: 10.1016/j.ijmedinf.2025.105871] [Medline: 40107040]
24. Deng Z, Ma W, Han QL, et al. Exploring DeepSeek: a survey on advances, applications, challenges and future directions. IEEE/CAA J Autom Sinica 2025 May;12(5):872-893. [doi: 10.1109/JAS.2025.125498]
25. Kaygisiz Ö, Teke MT. Can DeepSeek and ChatGPT be used in the diagnosis of oral pathologies? BMC Oral Health 2025 Apr 25;25(1):638. [doi: 10.1186/s12903-025-06034-x] [Medline: 40281436]
26. Rasool A, Shahzad MI, Aslam H, Chan V, Arshad MA. Emotion-aware embedding fusion in large language models (Flan-T5, Llama 2, DeepSeek-R1, and ChatGPT 4) for intelligent response generation. AI 2025 Mar 13;6(3):56. [doi: 10.3390/ai6030056]
27. Yilmaz BE, Gokkurt Yilmaz BN, Ozbey F. Artificial intelligence performance in answering multiple-choice oral pathology questions: a comparative analysis. BMC Oral Health 2025 Apr 15;25(1):573. [doi: 10.1186/s12903-025-05926-2] [Medline: 40234873]
28. Marcaccini G, Seth I, Xie Y, et al. Breaking bones, breaking barriers: ChatGPT, DeepSeek, and Gemini in hand fracture management. J Clin Med 2025 Mar 14;14(6):1983. [doi: 10.3390/jcm14061983] [Medline: 40142791]
29. Özcivelek T, Özcan B. Comparative evaluation of responses from DeepSeek-R1, ChatGPT-o1, ChatGPT-4, and dental GPT chatbots to patient inquiries about dental and maxillofacial prostheses. BMC Oral Health 2025 May 31;25(1):871. [doi: 10.1186/s12903-025-06267-w] [Medline: 40450291]
30. Gültekin O, Inoue J, Yilmaz B, et al. Evaluating DeepResearch and DeepThink in anterior cruciate ligament surgery patient education: ChatGPT-4o excels in comprehensiveness, DeepSeek R1 leads in clarity and readability of orthopaedic information. Knee Surg Sports Traumatol Arthrosc 2025 Aug;33(8):3025-3031. [doi: 10.1002/ksa.12711] [Medline: 40450565]
31. Seth I, Marcaccini G, Lim K, et al. Management of Dupuytren's disease: a multi-centric comparative analysis between experienced hand surgeons versus artificial intelligence. Diagnostics (Basel) 2025 Feb 28;15(5):587. [doi: 10.3390/diagnostics15050587] [Medline: 40075834]
32. VOSviewer. URL: https://www.vosviewer.com/ [accessed 2026-06-10]
33. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 2010;84:523-538. [doi: 10.1007/s11192-009-0146-3] [Medline: 20585380]
34. Lau JYS, Gerald Sng GR, Cao R, Chen J. A comparative study of ChatGPT and DeepSeek in spinal cord injury patient education: can artificial intelligence "speak" spinal cord injury? J Spinal Cord Med 2026 May;49(3):618-623. [doi: 10.1080/10790268.2025.2554013] [Medline: 40938207]
35. Liu Y, Yu F, Zhang X, et al. Assessing the role of large language models between ChatGPT and DeepSeek in asthma education for bilingual individuals: comparative study. JMIR Med Inform 2025 Aug 13;13:e65365. [doi: 10.2196/65365] [Medline: 40802989]
36. Uldin H, Saran S, Gandikota G, et al. A comparison of performance of DeepSeek-R1 model-generated responses to musculoskeletal radiology queries against ChatGPT-4 and ChatGPT-4o - a feasibility study. Clin Imaging 2025 Jul;123:110506. [doi: 10.1016/j.clinimag.2025.110506] [Medline: 40381536]
37. Wu H, Yao S, Bao H, Guo Y, Xu C, Ma J. ChatGPT-4.0 and DeepSeek-R1 does not yet provide clinically supported answers for knee osteoarthritis. Knee 2025 Oct;56:386-396. [doi: 10.1016/j.knee.2025.06.007] [Medline: 40618549]
38. Alluri AA, Khan Z, Krithika V, et al. Assessing the suitability of ChatGPT and DeepSeek AI for patient education on common rheumatological disorders. Cureus 2025 Aug;17(8):e90600. [doi: 10.7759/cureus.90600] [Medline: 40984935]
39. Gurbuz S, Bahar H, Yavuz U, Keskin A, Karslioglu B, Solak Y. Comparative efficacy of ChatGPT and DeepSeek in addressing patient queries on gonarthrosis and total knee arthroplasty. Arthroplast Today 2025 Jun;33:101730. [doi: 10.1016/j.artd.2025.101730] [Medline: 40521295]
40. Zhang J, Liu J, Guo M, Zhang X, Xiao W, Chen F. DeepSeek-assisted LI-RADS classification: AI-driven precision in hepatocellular carcinoma diagnosis. Int J Surg 2025;111(9):5970-5979. [doi: 10.1097/JS9.0000000000002763]
41. Vural Camalan B, Doluoglu S, Taraf NH, Gunay MM, Ozlugedik S. ChatGPT versus DeepSeek in head and neck cancer staging and treatment planning: guideline-based study. Eur Arch Otorhinolaryngol 2025 Sep;282(9):4815-4824. [doi: 10.1007/s00405-025-09524-4] [Medline: 40523995]

XSL•FO
RenderX

42. Mikhail D, Farah A, Milad J, et al. DeepSeek-R1 vs OpenAI o1 for ophthalmic diagnoses and management plans. JAMA Ophthalmol 2025 Oct 1;143(10):834-842. [doi: 10.1001/jamaophthalmol.2025.2918] [Medline: 40906471]
43. Cao H, Hao C, Zhang T, et al. Battle of the artificial intelligence: a comprehensive comparative analysis of DeepSeek and ChatGPT for urinary incontinence-related questions. Front Public Health 2025;13:1605908. [doi: 10.3389/fpubh.2025.1605908] [Medline: 40771241]
44. Pinto VBP, Ataídes RJC, do Nascimento LAP, et al. Performance of ChatGPT and DeepSeek in the management of postprostatectomy urinary incontinence. Int Braz J Urol 2025;51(6):e20250325. [doi: 10.1590/S1677-5538.IBJU.2025.0325] [Medline: 40857549]
45. Ibrahim AF, Danpanichkul P, Hayek A, et al. Artificial intelligence in gastroenterology education: DeepSeek passes the gastroenterology board examination and outperforms legacy ChatGPT models. Am J Gastroenterol 2026 Apr 1;121(4):1041-1043. [doi: 10.14309/ajg.0000000000003552] [Medline: 40392256]
46. Meo SA, Abukhalaf FA, ElToukhy RA, Sattar K. Exploring the role of DeepSeek-R1, ChatGPT-4, and Google Gemini in medical education: how valid and reliable are they? Pak J Med Sci 2025 Jul;41(7):1887-1892. [doi: 10.12669/pjms.41.7.12183] [Medline: 40735572]
47. Shean R, Shah T, Pandiarajan A, et al. A comparative analysis of DeepSeek R1, DeepSeek-R1-Lite, OpenAi o1 Pro, and Grok 3 performance on ophthalmology board-style questions. Sci Rep 2025 Jul 2;15(1):23101. [doi: 10.1038/s41598-025-08601-2] [Medline: 40595291]
48. Tassoker M. Who knows anatomy best? A comparative study of ChatGPT-4o, DeepSeek, Gemini, and Claude. Clin Anat 2026 Jan;39(1):25-29. [doi: 10.1002/ca.70012] [Medline: 40708277]
49. Ucdal M, Yurtsever K, Yildiz P, Akalin A, Mert KU, Guven GS. Comparison of artificial intelligence models and human experts in managing dyslipidemia: assessment of adherence to clinical guidelines. Cureus 2025 Aug;17(8):e91363. [doi: 10.7759/cureus.91363] [Medline: 40904968]
50. Lee S, Jung S, Park JH, Cho H, Moon S, Ahn S. Performance of ChatGPT, Gemini and DeepSeek for non-critical triage support using real-world conversations in emergency department. BMC Emerg Med 2025 Sep 1;25(1):176. [doi: 10.1186/s12873-025-01337-2] [Medline: 40890624]
51. Tao Y, Li X, Yisha Z, Yang S, Zhan S, Sun F. LitAutoScreener: development and validation of an automated literature screening tool in evidence-based medicine driven by large language models. Health Data Sci 2025;5:0322. [doi: 10.34133/hds.0322] [Medline: 40904687]
52. Ruan M, Fan J, Liu M, Meng Z, Zhang X, Zhang C. Artificial intelligence for the science of evidence synthesis: how good are AI-powered tools for automatic literature screening? BMC Med Res Methodol 2025 Aug 25;25(1):199. [doi: 10.1186/s12874-025-02644-9] [Medline: 40855531]
53. Cai X, Geng Y, Du Y, et al. Utilizing large language models to select literature for meta-analysis shows workload reduction while maintaining a similar recall level as manual curation. BMC Med Res Methodol 2025 Apr 28;25(1):116. [doi: 10.1186/s12874-025-02569-3] [Medline: 40295957]
54. Young RJ, Matthews AM, Poston B. Benchmarking multiple large language models for automated clinical trial data extraction in aging research. Algorithms 2025;18(5):296. [doi: 10.3390/a18050296]
55. Grillo R, Llanos AH, Costa C, Melhem-Elias F. Comparison of large language models in oral and maxillofacial surgery. Br J Oral Maxillofac Surg 2026 Jan;64(1):43-49. [doi: 10.1016/j.bjoms.2025.08.015] [Medline: 41076417]
56. Scholz V, Bichtemann S, Bott OJ, Illig T, Haag S. AI for extracting pre-analytical variability data from biomedical literature: feasibility and validation. Stud Health Technol Inform 2025 Sep 3;331:52-62. [doi: 10.3233/SHTI251379] [Medline: 40899527]
57. Wu X, Cai G, Guo B, et al. A multi-dimensional performance evaluation of large language models in dental implantology: comparison of ChatGPT, DeepSeek, Grok, Gemini and Qwen across diverse clinical scenarios. BMC Oral Health 2025 Jul 28;25(1):1272. [doi: 10.1186/s12903-025-06619-6] [Medline: 40721763]
58. Li Y, Dong J, Liu D, et al. Systematic benchmarking of large language models in programmed cell death-oriented gastric cancer research: a comparative analysis of DeepSeek-V3, DeepSeek-R1, and Claude 3.5. Discov Onc 2025 Jul 1;16(1):1227. [doi: 10.1007/s12672-025-02911-7]
59. Kayaalp ME, Gültekin O, Akçaalan S, Kahraman H, Topçu HN, Kavrul Kayaalp G. Artificial intelligence in medical and biological research: promise and perils of ChatGPT and DeepSeek in advancing healthcare. Turk J Biol 2025;49(5):585-599. [doi: 10.55730/1300-0152.2765] [Medline: 41246235]
60. Abuabara A, do Nascimento T, Trentini SM, et al. Evaluating the accuracy of generative artificial intelligence models in dental age estimation based on the Demirjian's method. Front Dent Med 2025;6:1634006. [doi: 10.3389/fdmed.2025.1634006] [Medline: 40800006]
61. AlShahwan N, Fetyani IM, Beyari MB, et al. Comparative performance analysis of AI engines in answering American Board of Surgery in-training examination questions: a multi-subspecialty evaluation. Surg Innov 2025 Dec;32(6):502-506. [doi: 10.1177/15533506251361664] [Medline: 40664612]
62. Yang Y, Yang F, Xiao S, et al. Application of large language models in TN staging and treatment response evaluation for patients with nasopharyngeal carcinoma: a comparative performance analysis of ChatGPT-4o-Latest and DeepSeek-V3-0324. J Magn Reson Imaging 2025 Dec;62(6):1793-1801. [doi: 10.1002/jmri.70140] [Medline: 41045017]

XSL•FO
RenderX

63. Yan MY, Qin JA, Yan D. Performance evaluation and application value of large language models in the prediction of drug-drug interactions. Yaoxue Xuebao 2025;60(7):2122-2131. [doi: 10.16438/j.0513-4870.2025-0590]
64. Xie L, Jin Y, Xu L, Chang S, Xu X. Fusing domain knowledge with a fine-tuned large language model for enhanced molecular property prediction. J Chem Theory Comput 2025 Jul 22;21(14):6743-6758. [doi: 10.1021/acs.jctc.5c00605] [Medline: 40631446]
65. McCoy TH, Perlis RH. Reasoning language models for more transparent prediction of suicide risk. BMJ Ment Health 2025 May 11;28(1):e301654. [doi: 10.1136/bmjment-2025-301654] [Medline: 40350181]
66. Alessandri-Bonetti M, Liu HY, Giorgino R, Nguyen VT, Egro FM. The first months of life of ChatGPT and its impact in healthcare: a bibliometric analysis of the current literature. Ann Biomed Eng 2024 May;52(5):1107-1110. [doi: 10.1007/s10439-023-03325-8] [Medline: 37482572]
67. Chen SF, Alyakin A, Seas A, et al. LLM-assisted systematic review of large language models in clinical medicine. Nat Med 2026 Mar;32(3):1152-1159. [doi: 10.1038/s41591-026-04229-5] [Medline: 41776077]
68. Anusitviwat C, Suwannaphisit S, Bvonpanttarananon J, Tangtrakulwanich B. Comparing ChatGPT and DeepSeek for assessment of multiple-choice questions in orthopedic medical education: cross-sectional study. JMIR Form Res 2025 Dec 19;9:e75607. [doi: 10.2196/75607] [Medline: 41418321]
69. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023 Mar 4;47(1):33. [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
70. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
71. He K, Mao R, Lin Q, et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. Information Fusion 2025 Jun;118:102963. [doi: 10.1016/j.inffus.2025.102963]
72. Liu F, Zhou H, Gu B, et al. Application of large language models in medicine. Nat Rev Bioeng 2025 Jul;3(6):445-464. [doi: 10.1038/s44222-025-00279-5]
73. Unger JP, Morales I, De Paepe P, Roland M. Integrating clinical and public health knowledge in support of joint medical practice. BMC Health Serv Res 2020 Dec 9;20(Suppl 2):1073. [doi: 10.1186/s12913-020-05886-z] [Medline: 33292211]
74. Hassanein FEA, El Barbary A, Hussein RR, et al. Diagnostic performance of ChatGPT-4o and DeepSeek-3 differential diagnosis of complex oral lesions: a multimodal imaging and case difficulty analysis. Oral Dis 2025 Dec;31(12):3361-3371. [doi: 10.1111/odi.70007] [Medline: 40589366]
75. Goyal A, Sulaiman SA, Alaarag A, et al. Comparison of ChatGPT and DeepSeek large language models in the diagnosis of pericarditis. World J Cardiol 2025 Aug 26;17(8):110489. [doi: 10.4330/wjc.v17.i8.110489] [Medline: 40949931]
76. He Q, Tan Z, Niu W, et al. From algorithms to operating room: can large language models master China's attending anesthesiology exam? A cross-sectional evaluation. Int J Surg 2026 Jan 1;112(1):190-201. [doi: 10.1097/JS9.0000000000003406] [Medline: 40905848]
77. Karataş G, Karataş ME. Artificial intelligence in pediatric ophthalmology: a comparative study of ChatGPT-4.0 and DeepSeek-R1 performance. Strabismus 2026 Mar;34(1):61-67. [doi: 10.1080/09273972.2025.2536782] [Medline: 40726359]
78. Smith A, Liebrenz M, Bhugra D, Grana J, Schleifer R, Buadze A. Are clinical improvements in large language models a reality? Longitudinal comparisons of ChatGPT models and DeepSeek-R1 for psychiatric assessments and interventions. Int J Soc Psychiatry 2026 Feb;72(1):91-102. [doi: 10.1177/00207640251358071] [Medline: 40741928]
79. Harada Y, Kawamura R, Yokose M, Singh H, Shimizu T. Atypical presentations at risk for diagnostic errors in internal medicine: a scoping review. J Gen Intern Med 2026 May;41(7):1937-1956. [doi: 10.1007/s11606-025-09901-z] [Medline: 41085962]
80. Brohi S, Mastoi QUA, Jhanjhi NZ, Pillai TR. A research landscape of agentic AI and large language models: applications, challenges and future directions. Algorithms 2025;18(8):499. [doi: 10.3390/a18080499]
81. Temsah A, Alhasan K, Altamimi I, et al. DeepSeek in healthcare: revealing opportunities and steering challenges of a new open-source artificial intelligence frontier. Cureus 2025 Feb;17(2):e79221. [doi: 10.7759/cureus.79221] [Medline: 39974299]
82. Liu Q, Xin Y, Wu C, et al. Diagnostic value of combining ultrafast cine MRI and morphological measurements on gastroesophageal reflux disease. Abdom Radiol 2025;50(10):4495-4506. [doi: 10.1007/s00261-025-04890-3]
83. Diniz-Freitas M, Diz-Dios P. DeepSeek: another step forward in the diagnosis of oral lesions. J Dent Sci 2025 Jul;20(3):1904-1907. [doi: 10.1016/j.jds.2025.02.023] [Medline: 40654453]
84. ElSayed A, Updegrove GF. Limitations of broadly trained LLMs in interpreting orthopedic Walch glenoid classifications. Front Artif Intell 2025;8:1644093. [doi: 10.3389/frai.2025.1644093] [Medline: 40951327]
85. Beauchamp T, Childress J. Principles of Biomedical Ethics: marking its fortieth anniversary. Am J Bioeth 2019 Nov;19(11):9-12. [doi: 10.1080/15265161.2019.1665402] [Medline: 31647760]
86. Wang M, Shen Y, Zhao B, Zhou X, Sun L, Liu X. Enhancing LLM-based clinical reasoning in anesthesiology via graph-augmented retrieval and explainable generation. Health Inf Sci Syst 2025 Dec;13(1):62. [doi: 10.1007/s13755-025-00379-x] [Medline: 41041605]
87. Choudhury A, Shahsavar Y, Shamszare H. User intent to use DeepSeek for health care purposes and their trust in the large language model: multinational survey study. JMIR Hum Factors 2025 May 26;12:e72867. [doi: 10.2196/72867] [Medline: 40418796]

XSL•FO
RenderX

88. Wang Z, Zhou H, Song T. A bibliometric analysis of large language model-based AI chatbots in surgery. Annals of Medicine & Surgery 2025;87(7):4127-4138. [doi: 10.1097/MS9.0000000000003234]
89. Moëll B, Sand Aronsson F, Akbar S. Medical reasoning in LLMs: an in-depth analysis of DeepSeek R1. Front Artif Intell 2025;8:1616145. [doi: 10.3389/frai.2025.1616145] [Medline: 40607450]
90. Cao Y, Wang J, Li Y, Zhang Y, Zhong G, Song P. Expert consensus on the deployment of DeepSeek in medical institutions. Chinese Medical Ethics 2025;38(5):674-678. [doi: 10.12026/j.issn.1001-8565.2025.05.19]
91. Si Y, Meng Y, Chen X, et al. Quality safety and disparity of an AI chatbot in managing chronic diseases: simulated patient experiments. NPJ Digit Med 2025 Sep 25;8(1):574. [doi: 10.1038/s41746-025-01956-w] [Medline: 40999038]
92. Dong C, Qiu X, Deng J, et al. Comparative evaluation of large language models in delivering guideline-compliant recommendations for topical NSAID use in musculoskeletal pain: a multidimensional analysis. Clin Rheumatol 2025 Nov;44(11):4703-4710. [doi: 10.1007/s10067-025-07640-4] [Medline: 40952435]
93. Rowland SP, Fitzgerald JE, Holme T, Powell J, McGregor A. What is the clinical value of mHealth for patients? NPJ Digit Med 2020;3:4. [doi: 10.1038/s41746-019-0206-x] [Medline: 31970289]
94. Watts E, Patel H, Kostov A, Kim J, Elkbuli A. The role of compassionate care in medicine: toward improving patients' quality of care and satisfaction. J Surg Res 2023 Sep;289:1-7. [doi: 10.1016/j.jss.2023.03.024] [Medline: 37068438]
95. Thomas RL, Uminsky D. Reliance on metrics is a fundamental challenge for AI. Patterns (N Y) 2022 May 13;3(5):100476. [doi: 10.1016/j.patter.2022.100476] [Medline: 35607624]
96. Su H, Sun Y, Li R, et al. Large language models in medical diagnostics: scoping review with bibliometric analysis. J Med Internet Res 2025 Jun 9;27:e72062. [doi: 10.2196/72062] [Medline: 40489764]
97. Zhou H, Wang Z, Wang R, et al. DeepSeek versus GPT: evaluation of large language model chatbots' responses on orofacial clefts. J Craniofac Surg 2025 Sep 1;36(6):2197-2201. [doi: 10.1097/SCS.0000000000011399] [Medline: 40245329]
98. Zhou J, Cheng Y, He S, Chen Y, Chen H. Large language models for transforming healthcare: a perspective on DeepSeek‐R1. MedComm – Future Medicine 2025 Jun;4(2):e70021 [FREE Full text] [doi: 10.1002/mef2.70021]
99. Wang Q, Chen Z, Zhang H, et al. Large language models could be applied in personalized out-of-hospital management for breast cancer: a prospective randomized single blind study. Sci Rep 2025 Sep 29;15(1):33589. [doi: 10.1038/s41598-025-18759-4]
100. Sahni NR, Carrus B. Artificial intelligence in U.S. health care delivery. N Engl J Med 2023 Oct 12;389(15):1442-1443. [doi: 10.1056/NEJMc2310288]
101. Finkenberg J. NASS 2023 presidential address: artificial intelligence and its effect on the art of medicine and the physician-patient relationship. Spine J 2024 Feb;24(2):191-194. [doi: 10.1016/j.spinee.2023.11.001] [Medline: 37944759]
102. Hui L, Khosa F. Artificial intelligence in action: racial and gender disparities in academic radiology. Cureus 2025 Sep;17(9):e92382. [doi: 10.7759/cureus.92382] [Medline: 41103889]
103. Dai X, Xi Y, Hu Y, et al. LLM evaluation for thyroid nodule assessment: comparing ACR-TIRADS, C-TIRADS, and clinician-AI trust gap. Front Endocrinol 2025;16:1667809. [doi: 10.3389/fendo.2025.1667809]
104. Wang G, Meng X, Zhang F. Past, present, and future of global research on artificial intelligence applications in dermatology: a bibliometric analysis. Medicine (Baltimore) 2023;102(45):e35993. [doi: 10.1097/MD.0000000000035993]
105. Huang Y, Yang G, Shen Y, et al. Application of large language models in complex clinical cases: cross-sectional evaluation study. JMIR Med Inform 2025 Aug 14;13:e73941. [doi: 10.2196/73941] [Medline: 41055081]
106. Sallam M, Alasfoor IM, Khalid SW, et al. Chinese generative AI models (DeepSeek and Qwen) rival ChatGPT-4 in ophthalmology queries with excellent performance in Arabic and English. Narra J 2025 Apr;5(1):e2371. [doi: 10.52225/narra.v5i1.2371] [Medline: 40352182]
107. Lim ECN, Cheng NCL, Lim CED. The art of medical synthesis: where Chinese medical wisdom intersects with artificial intelligence. Journal of Traditional Chinese Medical Sciences 2026 Jan;13(1):51-59. [doi: 10.1016/j.jtcms.2025.08.001]
108. Patil NG, Kou NL, Baptista‐Hon DT, Monteiro O. Artificial intelligence in medical education: a practical guide for educators. MedComm – Future Medicine 2025 Jun;4(2):e70018 [FREE Full text] [doi: 10.1002/mef2.70018]
109. Yan W, Liu J, Liang W. DeepSeek empowers general medicine: potential application and prospect. Chinese General Practice 2025 Jun;28(17):2065-2069. [doi: 10.12114/j.issn.1007-9572.2025.0023]
110. Editage [FREE Full text]

Abbreviations

AI: artificial intelligence
APY: average publication year
BIBLIO: bibliometric reviews of biomedical literature
CAD-RADS: Coronary Artery Disease Reporting and Data System
GDPR: General Data Protection Regulation
GRPO: Group Relative Policy Optimization
HIPAA: Health Insurance Portability and Accountability Act
LI-RADS: Liver Imaging Reporting and Data System

XSL•FO
RenderX

LLM: large language model
MeSH: medical subject headings
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
RCT: randomized controlled trial
WoSCC: Web of Science Core Collection

Edited by A Coristine; submitted 11.Feb.2026; peer-reviewed by F Liu, S Chen, ZA Nazi; revised version received 19.May.2026; accepted 20.May.2026; published 15.Jun.2026.

Please cite as:
Zhang H, Wang D, Xu Y, Han S, Wang G
Applications of DeepSeek in Medicine: Bibliometric Analysis and Scoping Review
J Med Internet Res 2026;28:e93354
URL: https://www.jmir.org/2026/1/e93354
doi:10.2196/93354

© Haoran Zhang, Dawei Wang, Yanliang Xu, Shuming Han, Guangxin Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Performance of Deep Learning in Classifying Age-Related Macular Degeneration From Images: Systematic Review and Meta-Analysis

Yu Zhu[1], MD; Yue Niu[2], BS; Shangye Sun[3], MD; Wei Liu[1], BS; Ying Dou[1], MD; Yu Guo[4], MD

[1]Department of Ophthalmology, Jilin Province FAW General Hospital, Changchun, China

[2]Department of Human Resources, Jilin Province FAW General Hospital, Changchun, China

[3]Department of CT, Jilin Province FAW General Hospital, Changchun, China

[4]Department of Otolaryngology, Jilin Province FAW General Hospital, 2643 Dongfeng Street, Changchun, Jilin Province, China

Corresponding Author:
Yu Guo, MD
Department of Otolaryngology, Jilin Province FAW General Hospital, 2643 Dongfeng Street, Changchun, Jilin Province, China

Abstract

Background: Age-related macular degeneration (AMD) is a leading cause of irreversible blindness worldwide. Retinal imaging and deep learning (DL) may support scalable screening, but deployment requires evidence on pooled performance. This is important because missed neovascular disease may delay treatment, whereas excessive false positives may overload referral pathways.

Objective: This study aimed to compare the diagnostic performance of DL algorithms with ophthalmologists for detecting AMD and differentiating wet AMD (wAMD) from dry AMD (dAMD) and to identify factors that influence DL performance.

Methods: PubMed, Embase, Web of Science, and the Cochrane Library were searched through October 5, 2025, and updated on April 19, 2026. Eligible studies applied DL to classify AMD from normal retinas or wAMD from dAMD using retinal images. Two reviewers (MHT and XL) independently extracted data and assessed risk of bias using the Prediction model Risk Of Bias Assessment Tool for Artificial Intelligence (PROBAST+AI) tool. Pooled sensitivity, specificity, accuracy, and area under the curve were estimated using bivariate random-effects models. Clinician comparisons were stratified by experience (junior vs senior). Small-study effects were assessed via Deeks' funnel plot asymmetry test. Evidence certainty was appraised using the Grading of Recommendations, Assessment, Development, and Evaluation framework. The protocol was registered in the International Prospective Register of Systematic Reviews (PROSPERO; CRD420251243276).

Results: Overall, 28 studies were included, comprising 77,485 samples for AMD detection and 28,705 samples for wAMD versus dAMD classification. For AMD detection, DL achieved a pooled sensitivity of 0.98 (95% CI 0.96 - 0.99; prediction interval [PI] 0.95 - 0.99), specificity of 0.98 (95% CI 0.95 - 0.99; PI 0.95 - 0.99), accuracy of 0.97 (95% CI 0.96 - 0.99), and area under the curve of 1.00 (95% CI 0.99 - 1.00). For wAMD versus dAMD, DL showed sensitivity of 0.95 (95% CI 0.91 - 0.97; PI 0.89 - 0.97), specificity of 0.95 (95% CI 0.93 - 0.97; PI 0.92 - 0.97), accuracy of 0.95 (95% CI 0.92 - 0.97), and area under the curve of 0.99 (95% CI 0.97 - 0.99). DL showed higher sensitivity than senior ophthalmologists for AMD (0.98 vs 0.75; $P<.001$) and higher specificity and accuracy than junior ophthalmologists for wAMD classification. Optical coherence tomography–based models performed more consistently than color fundus photography or multimodal models. Evidence certainty was moderate.

Conclusions: Compared with ophthalmologists, DL algorithms demonstrated superior and more balanced diagnostic performance in the available head-to-head evidence, potentially providing a consistent decision-support baseline that mitigates human threshold-dependent trade-offs. However, high heterogeneity, wide PIs, predominantly retrospective designs, and possible performance inflation from internal validation mean that these relative performance findings remain preliminary rather than deployment ready. DL should be viewed as a triage adjunct requiring local calibration, not an autonomous diagnostic replacement. Prospective, multicenter, patient-level external validation with prespecified human comparison arms is required.

Trial Registration: PROSPERO CRD420251243276; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251243276

(J Med Internet Res 2026;28:e97174) doi:10.2196/97174

KEYWORDS

age-related macular degeneration; deep learning; artificial intelligence; optical coherence tomography; meta-analysis

XSL•FO
RenderX

Introduction

Age-related macular degeneration (AMD) remains a leading cause of irreversible blindness in older individuals globally [1]. Clinically, the disease is classified into dry AMD (dAMD), characterized by the progressive accumulation of drusen and geographic atrophy, and wet AMD (wAMD), which involves rapid vision loss due to macular neovascularization [2]. As the global population ages, the prevalence of AMD is projected to rise significantly; recent estimates indicate that the number of individuals with AMD-related vision impairment will increase from 8.06 million in 2021 to approximately 21.34 million by 2050 [1]. Consequently, early and accurate diagnosis is paramount. Timely detection allows for appropriate intervention, which is critical for slowing disease progression, preserving visual function, and improving overall patient prognosis.

Conventionally, color fundus photography (CFP) and optical coherence tomography (OCT) serve as the cornerstones for AMD screening and diagnosis. However, reliance on these modalities presents distinct challenges. CFP is frequently limited by image quality; issues such as media opacities or small pupils can render images ungradable, with rates as high as 47.6% in some screening contexts, and CFP often lacks sensitivity for detecting subtle early-stage structural changes or neovascular activity [3,4]. Conversely, while OCT offers high gradability (up to 97.7%) and detailed cross-sectional visualization, it is constrained by a limited field of view and reduced efficacy in identifying pigmentary abnormalities compared to CFP [3,4]. Beyond these technical constraints, the manual interpretation of vast imaging datasets is inherently labor-intensive, subjective, and prone to interobserver variability, creating a scalability bottleneck for population-wide screening.

In response to these challenges, deep learning (DL) algorithms using OCT, CFP, or multimodal imaging have emerged as a transformative approach, offering the potential for automated, high-throughput classification [5,6]. While these algorithms demonstrate theoretical superiority in efficiency and feature extraction, the current literature reveals substantial heterogeneity in performance outcomes [7,8]. Discrepancies regarding model generalization to real-world settings and the comparative performance of DL algorithms against ophthalmologists remain unresolved [7,8]. Two pivotal questions persist: First, how does the diagnostic performance of DL models quantitatively compare against ophthalmologists of varying expertise? Second, what factors, such as imaging modality, type of validation, database source, study centers, and unit of analysis, influence DL performance? Existing literature offers fragmented and sometimes contradictory insights, lacking a comprehensive quantitative synthesis.

Several previous meta-analyses have evaluated DL performance in AMD diagnosis. Leng et al [9] reported a pooled sensitivity of 94% and specificity of 97% for convolutional neural network–based algorithms, while Chen et al [10] highlighted the overall superiority of artificial intelligence (AI) over retinal specialists. However, these prior reviews have notable limitations: they did not stratify human-AI comparisons by clinician experience level, used conventional bias assessment tools rather than the recently developed Prediction model Risk Of Bias Assessment Tool for Artificial Intelligence (PROBAST+AI) instrument [11], and did not separately evaluate the clinically critical task of differentiating wAMD from dAMD. Moreover, the rapid advancement of DL architectures, particularly vision transformers, necessitates an updated quantitative synthesis incorporating the latest evidence.

Importantly, our review was designed to address several evidence gaps that were not fully covered in previous meta-analyses. First, instead of evaluating DL algorithms in isolation, we directly compared DL performance with ophthalmologists and further stratified these comparisons by clinician experience level. This is clinically relevant because screening and referral decisions are often made by clinicians with different levels of expertise. Second, we separately evaluated the classification of wAMD versus dAMD, a task with immediate therapeutic implications because delayed recognition of wAMD may postpone anti–vascular endothelial growth factor treatment. Third, we incorporated PROBAST+AI, a recently developed tool tailored to prediction models using AI, thereby providing a more AI-specific assessment of bias than conventional quality appraisal tools [12]. Fourth, we examined prediction intervals (PIs), validation strategy, imaging modality, and other sources of heterogeneity to move beyond average pooled performance and assess the likely robustness of DL algorithms across clinical settings. These features make the current review not only an update of the evidence base, but also a more deployment-oriented synthesis of the clinical value and limitations of DL for AMD image classification.

Therefore, this systematic review and meta-analysis addressed these clinically relevant and deployment-oriented evidence gaps. Its objective was to evaluate the diagnostic performance of DL algorithms compared with ophthalmologists of varying experience levels for detecting AMD and differentiating its subtypes (wAMD vs dAMD), and to assess potential factors influencing DL diagnostic performance through subgroup analyses and meta-regressions.

Methods

Overview

This systematic review and meta-analysis was conducted in strict accordance with the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy) guidelines [13], with the specific reporting items detailed in Table S1 in Multimedia Appendix 1. The abstract was reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 for abstracts checklist, as shown in Table S2 in Multimedia Appendix 1. This systematic review and meta-analysis adhered to the preregistered protocol (PROSPERO [International Prospective Register of Systematic Reviews] CRD420251243276). At the request of reviewers, PIs for sensitivity and specificity were calculated as an additional analysis to provide estimates of expected performance in new clinical settings [13].

XSL•FO
RenderX

Search Strategy

A comprehensive literature search was conducted across PubMed, Embase, Web of Science, and the Cochrane Library databases, with coverage extending through April 19, 2026. The initial search was conducted through October 5, 2025, and subsequently updated on April 19, 2026, to capture any recently published studies. Two independent reviewers (YZ and YN) performed the preliminary screening of titles and abstracts, followed by a full-text assessment. The search strategy used a combination of free-text terms and Medical Subject Headings focusing on four distinct domains: AMD-related terminologies (eg, "Macular Degeneration"), DL concepts (eg, "Artificial Intelligence" and "Deep Learning"), imaging modalities (eg, "Optical Coherence Tomography" and "fundus photograph"), and diagnostic performance metrics (eg, "sensitivity" and "specificity"). No restrictions regarding language or publication year were applied during the initial retrieval. To ensure exhaustiveness, reference lists of included studies and relevant meta-analyses were manually scrutinized for additional literature. Detailed search queries were provided in Table S3 in Multimedia Appendix 1.

Eligibility Criteria

The study inclusion adhered to the Patient, Index test, Target condition, Reference standard, Outcome, and Setting (PITROS) framework, which is detailed in Table 1. To ensure the analysis focused on diagnostic test accuracy, several exclusion criteria were systematically applied. We excluded publications with clearly irrelevant titles and abstracts, and specific noneligible study types, including reviews, cross-sectional surveys, case reports, conference abstracts, meta-analyses, letters, and studies with unavailable full texts. Furthermore, studies were excluded if their primary aim was not the classification of AMD versus Normal or wAMD versus dAMD, or if they lacked sufficient data to extract or calculate a 2×2 contingency table (true positives [TPs], false positives [FPs], false negatives [FNs], and true negatives [TNs]). The screening was performed independently by 2 reviewers (YZ and SYS). Any discrepancies were resolved through discussion or, if necessary, by consultation with a third senior reviewer (YG) to reach a final consensus.

Table . Summary of inclusion criteria using the PITROS[a] framework.

Criteria	Details
Participants (P)	Adults undergoing retinal imaging (CFP[b] or OCT[c]) with confirmed ocular status (normal, dAMD[d], or wAMD[e]) based on clinical diagnosis or standard imaging protocols.
Index test (I)	DL[f] algorithms using retinal images (CFP, OCT, or both) for automated diagnosis or classification.
Target conditions (T)	The study addressed two classification tasks: first, AMD[g] versus normal, comparing confirmed AMD cases (positive) against healthy controls (negative); and second, wAMD versus dAMD, distinguishing exudative or neovascular AMD (positive) from nonexudative or atrophic AMD (negative).
Reference standard (R)	Clinical diagnosis by ophthalmologists based on multimodal imaging (eg, CFP, OCT, and fluorescein angiography) and/or longitudinal follow-up.
Outcomes (O)	Diagnostic performance metrics, including sensitivity, specificity, accuracy, and AUC[h]. Data extraction focused on contingency tables (TP[i], FP[j], FN[k], and TN[l]).
Settings (S)	Retrospective or prospective studies using single-center, multicenter clinical datasets, or public databases (eg, AREDS[m] and Kaggle).

[a]PITROS: Patient, Index test, Target condition, Reference standard, Outcome, and Setting.
[b]CFP: color fundus photography.
[c]OCT: optical coherence tomography.
[d]dAMD: dry age-related macular degeneration.
[e]wAMD: wet age-related macular degeneration.
[f]DL: deep learning.
[g]AMD: age-related macular degeneration.
[h]AUC: area under the curve.
[i]TP: true positive.
[j]FP: false positive.
[k]FN: false negative.
[l]TN: true negative.
[m]AREDS: Age-Related Eye Disease Study.

XSL•FO
RenderX

Quality Assessment and Certainty of Evidence

The methodological quality and risk of bias of the included studies were assessed using the PROBAST+AI tool [11], an updated version replacing PROBAST 2019. This tool evaluates two distinct phases, model development and model evaluation, across seven domains each, encompassing participants, predictors, outcomes, and analysis. Each domain is judged as having a low, high, or unclear risk of bias based on a series of tailored signaling questions. These questions are rated as "yes," "probably yes," "probably no," "no," "no information," or "not applicable". The complete set of signaling questions and detailed ratings were provided in Tables S4 and S5 in Multimedia Appendix 1. To ensure objectivity and accuracy, 2 reviewers (YZ and WL) independently performed this assessment for all included studies.

To appraise the certainty of the evidence for the pooled sensitivity, specificity, and diagnostic accuracy, we used the Grading of Recommendations, Assessment, Development, and Evaluations (GRADE) framework for diagnostic studies. This approach focuses on five key domains: risk of bias, indirectness, inconsistency, imprecision, and small-study effects [13]. The GRADE summary of findings table was formatted according to the diagnostic test accuracy template recommended by the GRADE working group and used a pretest probability of 20% for expected results per 1000 tested. The full GRADE assessment criteria and the final judgments for each outcome were detailed in Table S6 in Multimedia Appendix 1.

Data Extraction

Two reviewers (YZ and YD) independently performed data extraction from the full-text articles, and disagreements were resolved by discussion with a third reviewer (YG). Extracted information included study design, patient or sample size, imaging modality, data source, validation design, reference standard, target condition, AI architecture, diagnostic contingency data or reconstructed diagnostic data, ophthalmologist experience level when available, and risk-of-bias and certainty judgments. The data extraction tables are included in the paper.

As most studies did not report the full binary diagnostic contingency table (2×2 table), we used an indirect derivation approach. Specifically, TP, FP, FN, and TN values were extracted by merging categories from multiclass confusion matrices (three- or four-class tables) reported in the included studies. In a few cases where such matrices were unavailable, the values were indirectly calculated using reported sensitivity, specificity, AMD sample size, and total sample size. When studies provided multiple, nonoverlapping validation sets, the corresponding contingency tables were assumed independent, and all were extracted. However, when multiple DL algorithms were presented within a single study, only the major model defined by the authors was extracted to avoid patient overlap in data pooling. To ensure the independence of data in our meta-analyses and avoid double counting, this approach guaranteed that the same patient sample was not counted more than once.

Outcome Measures

The primary outcome measures were pooled sensitivity, specificity, accuracy, and area under the curve (AUC). Sensitivity and specificity describe threshold-dependent diagnostic performance, accuracy summarizes the proportion of correctly classified samples in the analyzed dataset and may be affected by class balance, and AUC reflects threshold-independent discrimination across possible decision thresholds. These metrics were therefore interpreted as complementary rather than interchangeable indicators of model behavior.

Statistical Analysis

Considering the inherent heterogeneity anticipated among studies, a bivariate random-effects model was used to pool the sensitivity, specificity, and AUC values for both DL algorithms and ophthalmologists [14]. For the diagnostic accuracy metric, a generalized linear mixed model with a random-effects framework was used following a rlogit transformation. A 2-sample Z-test was used to compare the differences in pooled sensitivity, specificity, diagnostic accuracy, and AUC, with statistical significance defined as a P value <.05. It is important to note that this bivariate model uses restricted maximum likelihood estimation, which differs fundamentally from the DerSimonian-Laird approach used in standard pairwise meta-analyses. While the Hartung-Knapp-Sidik-Jonkman adjustment is recommended for DL-based analyses to reduce false positives [15], it is not directly applicable to the bivariate diagnostic framework, as the restricted maximum likelihood–based bivariate model already provides more accurate variance estimation that inherently guards against inflated type I error rates. To characterize the distribution of true effects across different populations and settings, 95% PIs were calculated where a sufficient number of studies (≥3) were available, complementing the confidence intervals for the pooled average effects [16]. While confidence intervals quantify the precision of the average effect, PIs estimate the range within which the true diagnostic performance of a future study is expected to fall.

For DL algorithm outcomes demonstrating substantial heterogeneity, a bivariate boxplot and multivariable meta-regression were performed to explore potential sources. As prespecified, subgroup analyses were conducted based on imaging modality (OCT vs CFP vs multimodal), with between-subgroup differences compared and visualized using violin plots. The potential clinical impact of the DL algorithms was assessed using a Fagan nomogram. Small-study effects were evaluated using Deeks' funnel plot asymmetry test, with a P value <.10 indicating potential asymmetry [17]. It should be noted that funnel plot asymmetry can arise from multiple sources beyond small-study effects, including differences in study quality, true heterogeneity, and chance [17,18]. All statistical analyses were performed using Stata 15.1 (StataCorp LLC, with the *midas* and *metadta* packages) and R (version 4.5.1; R Core Team, using the *ggplot2* and *tidyverse* packages). All statistical tests were 2-tailed.

XSL•FO
RenderX

Use of Large Language Models

During the preparation of this work, we used OpenAI Codex (GPT-5) to assist with text generation, proofreading and editing, summarizing text, formulation of conclusions, translation, and reformatting under full human supervision. The tool was not used to make eligibility decisions, extract data, perform statistical analyses, or draw independent scientific conclusions. After using this tool, we reviewed and edited the content as needed and took full responsibility for the content of the publication.

Results

Study Selection

The initial database search identified 3586 potentially relevant records. After removing duplicates, 1980 unique records underwent title and abstract screening. During this phase, 1861 records were excluded due to obvious irrelevance or ineligible publication types (eg, reviews and conference abstracts). Subsequently, 119 full-text articles were assessed for eligibility. Following a detailed review, 61 studies were excluded as they did not primarily focus on distinguishing AMD from normal retinas or discriminating wAMD from dAMD. One cross-sectional study and 32 studies lacking sufficient or complete diagnostic data (TP, FP, FN, and TN) were further excluded. An additional 3 articles identified from other nondatabase sources (eg, reference lists) were also included [19-21]. Consequently, 28 studies [19-46] met all predefined inclusion criteria and were included in the meta-analysis. The study selection process followed the PRISMA guidelines, as shown in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram illustrating the study selection process for the systematic review and meta-analysis. AMD: age-related macular degeneration; dAMD: dry age-related macular degeneration; FN: false negative; FP: false positive; TN: true negative; TP: true positive; wAMD: wet age-related macular degeneration.

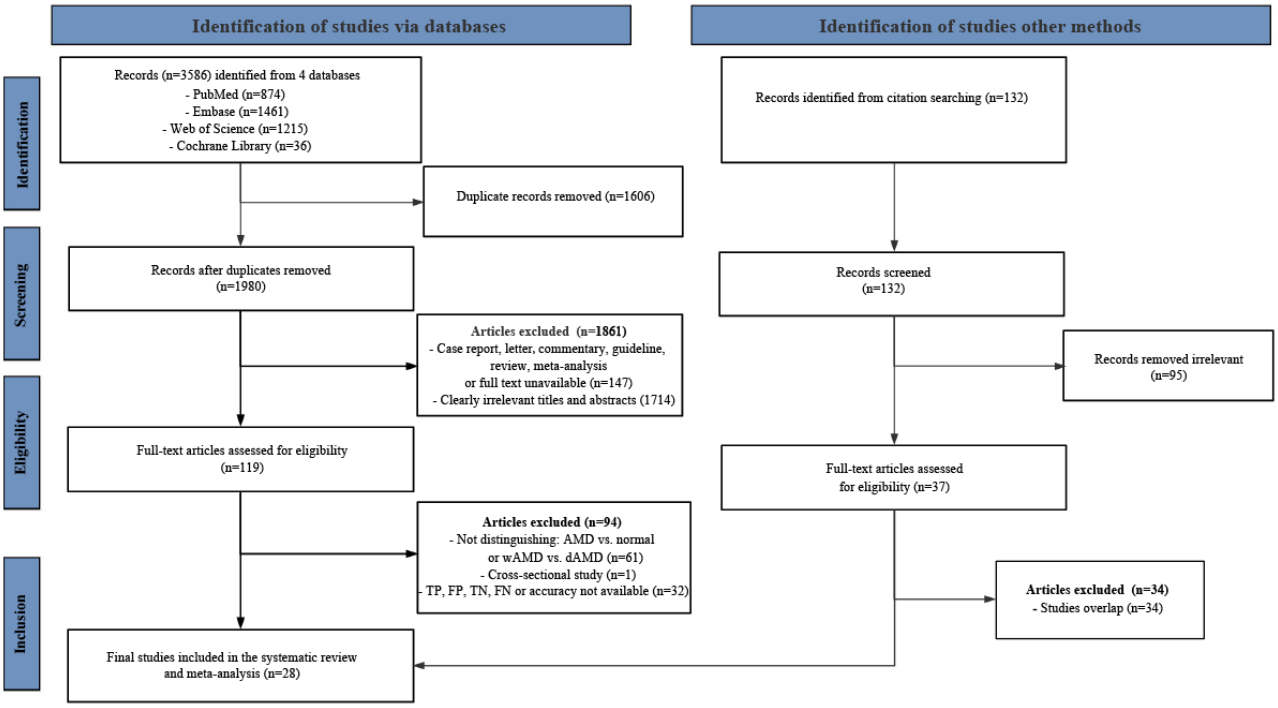

Study Characteristics

A total of 27 studies [19-36,38-46], comprising a validation set of 77,485 samples, targeted the classification of AMD versus normal. Among these, diagnostic performance data were available for junior ophthalmologists in 1 study [24], and for senior ophthalmologists in 3 studies [20,24,36]. Ten studies were based on OCT imaging [23,25-28,32,34,35,40,44], 12 on CFP [19-21,24,29-31,33,36,38,39,41], and 4 on multimodal (OCT+CFP) inputs [42,43,45,46]. Detailed study, patient, and technical characteristics were presented in Table 2 and Tables S7-S8 in Multimedia Appendix 1.

For the classification of wAMD versus dAMD, 16 studies (validation set: 28,705 samples) were included [19,23-29,31,33,37,41-44,46]. Data for head-to-head comparisons with ophthalmologists were available from two studies each for junior and senior practitioners [19,24], enabling stratified analysis by experience level. Six studies used OCT [23,25-28,44], 7 used CFP [19,24,29,31,33,37,41], and 3 used multimodal imaging [42,43,46]. Corresponding detailed characteristics were provided in Table 2 and Tables S9-S10 in Multimedia Appendix 1. The complete diagnostic data for ophthalmologists were presented in Table S11 in Multimedia Appendix 1.

XSL•FO
RenderX

Table . Study and patient characteristics of the included studies.

Author	Year	Country	Study design	Analysis	Reference standard	Target condition	Number of total sample size			Number of positive sample size
							Training	IV[a]	EV[b]	
Abdelhalim et al [22]	2025	Multiple countries	Retrospective	IB[c]	—[d]	AMD[e] versus normal	• 1011	• 127	—	• Training: 589 • IV: 74
Bao et al [24]	2025	China	Retrospective	IB	—		• 894			
						AMD versus normal		• IV1: 223	• EV1: 1395 • EV2: 59	• Training: 406 • IV1: 101 • EV1: 194 • EV2: 40
						wAMD[f] versus dAMD[g]		• IV1: 92	• EV1: 136 • EV2: 40	• Training: 267 • IV1: 60 • EV1: 36 • EV2: 21
Durmaz Engin et al [28]	2025	NA	Retrospective	IB	Expert consensus	AMD versus normal	• 1200	• 300	—	• Training: 800 • IV: 200
Zhen et al [46]	2025	China	Retrospective	PB[h]	—		• Training 1: 664		—	
						AMD versus normal		• 134		• Training 1: 509 • IV: 101
						wAMD versus dAMD		• 101		• Training: 386 • IV: 75
Alenezi et al [23]	2024	NA	Retrospective	IB	Expert consensus		• 554		—	
						AMD versus normal		• 554		• Training: 367 • IV: 367

XSL•FO
RenderX

Author	Year	Country	Study design	Analysis	Reference standard	Target condition	Number of total sample size			Number of positive sample size
							Training	IV[a]	EV[b]	
						wAMD versus dAMD		• 361		• Training: 173 • IV: 173
García-Floriano et al [30]	2024	Multiple countries	Retrospective	IB	Expert consensus	AMD versus normal	• 250	• 250	• 22	• Training: 128 • IV: 128 • EV: 11
Le et al [33]	2024	Multiple countries	Retrospective	IB	Expert consensus		• 2359			
						AMD versus normal		• 2359	• 750	• Training: 1144 • IV: 1144 • EV: 500
						wAMD versus dAMD		• 1095	• 440	• Training: 599 • IV: 588 • EV: 239
Oliveira et al [36]	2024	Multiple countries	Retrospective	IB	Expert consensus	AMD versus normal	• 6896	• 210	• 80	• Training: 275 • IV: 105 • EV: 40
Wan et al [41]	2024	China	Retrospective	IB	Clinical classification manifestations and expert consensus		• 516			
						AMD versus normal		• 129	• 100	• Training: 312 • IV: 78 • EV: 60
						wAMD versus dAMD		• 78	• 60	• Training: 214 • IV: 53 • EV: 41
	2024	Turkey		IB	—				—	

XSL•FO
RenderX

Author	Year	Country	Study design	Analysis	Reference standard	Target condition	Number of total sample size			Number of positive sample size
							Training	IV[a]	EV[b]	
Yusufoglu et al [44]			Retrospective							

XSL•FO
RenderX

Author	Year	Country	Study design	Analysis	Reference standard	Target condition	Number of total sample size			Number of positive sample size
							Training	IV[a]	EV[b]	
						AMD versus normal	• Training 1: 1622 • Training 2: 2240	• IV1: 347 • IV2: 491		• Training 1: NA • IV1: 214 • Training 2: NA • IV2: 248
						wAMD versus dAMD	• 1622	• 213		• Training 1: NA • IV1: 110
Celebi et al [26]	2023	Turkey	Retrospective	IB	Expert consensus		• Training 1: 4067 • Training 2: 59,139		—	
						AMD versus normal		• IV1: 1741 • IV2: 25,345		• Training 1: 2577 • IV1: 1103 • Training 2: 40,544 • IV2: 17,375
						wAMD versus dAMD		• IV1: 1085 • IV2: 17,369		• Training 1: 812 • IV1: 345 • Training 2: 26,219 • IV2: 11,234
El-Den et al [29]	2023	Multiple countries	Retrospective	IB	—		• 605		—	
						AMD versus normal		• 128		• Training: 454 • IV: 98
						wAMD versus dAMD		• 97		• Training: 151 • IV: 32
Leingang et al [35]	2023	Multicountries	Retrospective	PB	Expert consensus	AMD versus normal	• 1733	• 96	—	• Training: 1620 • IV: 90

XSL•FO
RenderX

Author	Year	Country	Study design	Analysis	Reference standard	Target condition	Number of total sample size			Number of positive sample size
							Training	IV[a]	EV[b]	
Chen et al [27]	2022	China	Retrospective	EB[i]	Expert consensus		• 612			
						AMD versus normal		• 153	• 214	• Training: 356 • IV: 89 • EV: 162
						wAMD versus dAMD		• 87	143	• Training: 189 • IV: 47 • EV: 96
He et al [32]	2022	China	Retrospective	IB[c]	—	AMD versus normal	• 77,568	• 750	• 2130	• Training: 36,656 • IV: 500 • EV: 723
Skevas et al [21]	2022	Germany	Prospective	PB[h]	Expert consensus	AMD versus normal	—	—	• 598	• EV: 69
Wang et al [42]	2022	China	Retrospective	EB	Expert consensus		—		—	

XSL•FO
RenderX

Author	Year	Country	Study design	Analysis	Reference standard	Target condition	Number of total sample size			Number of positive sample size
							Training	IV[a]	EV[b]	
						AMD versus normal		• 91		• Training: NA • IV: 71
						wAMD versus dAMD		• 71		• Training: NA • IV: 33
Tak et al [37]	2021	United States	Retrospective	IB	Expert consensus	wAMD versus dAMD	• 350	• 72	—	• Training: NA • IV: 28
Takhchidi et al [38]	2021	Russia	Retrospective	IB	Expert consensus	AMD versus normal	• 994	• 206	—	• Training: 475 • IV: 100
Thomas et al [40]	2021	India	Retrospective	IB	Expert consensus	AMD versus normal	• 87,264	• 750	• 30	• Training: 41,238 • IV: 500 • EV: 15
Heo et al [19]	2020	Korea	Retrospective	IB	Expert consensus		• 279	• 279	—	
						AMD versus normal				• Training: 191 • IV: 191
						wAMD versus dAMD				• Training: 99 • IV: 99
Zapata et al [45]	2020	Multicountries	Retrospective	IB	Expert consensus	AMD versus normal	• 7949	• 2208	—	• Training: NA • IV: 1082
Bhatia et al [25]	2019	Multiple countries	Retrospective	EB	—		—	—		

XSL•FO
RenderX

Author	Year	Country	Study design	Analysis	Reference standard	Target condition	Number of total sample size			Number of positive sample size
							Training	IV[a]	EV[b]	
						AMD versus normal			• EV1:98 • EV1:75	• EV1: 48 • EV2: 50
						wAMD versus dAMD			• EV2:50	• EV2: 25
Matsuba et al [20]	2019	Japan	Retrospective	IB	Expert consensus	AMD versus normal	• 5000	• 111	—	• Training: 870 • IV: 42
Yoo et al [43]	2019	Korea	Retrospective	IB	Pathological examination					

XSL•FO
RenderX

Author	Year	Country	Study design	Analysis	Reference standard	Target condition	Number of total sample size						Number of positive sample size
							Training		IV[a]		EV[b]		
						AMD versus normal	• Training 1: 2100		• IV1: 900		• EV2: 83		• Training 1: 1400 • IV1: 600 • EV2: 48
						wAMD versus dAMD	—		—		• EV2:48		• EV2: 36
Grassmann et al [31]	2018	Multiple countries	Prospective	IB	Expert consensus		• 83,653						
						AMD versus normal			11,618		• 1677		• Training: 53,375 • IV: 7571 • EV: 220
						wAMD versus dAMD			• 6631		• 123		• Training: 9357 • IV: 1432 • EV: 4
Tan et al [39]	2018	Multicountries	Retrospective	IB	—	AMD versus normal	• 1110		• 1100		—		• Training: 708 • IV: 700
Lee et al [34]	2017	United States	Retrospective	IB	Expert consensus	AMD versus normal	• 80,839		• 20,163		—		• Training: 41,074 • IV: 11,616

[a]IV: internal validation.

[b]EV: external validation.

[c]IB: image-based.

[d]NA: not available.

[e]AMD: age-related macular degeneration.

[f]wAMD: wet age-related macular degeneration.

[g]dAMD: dry age-related macular degeneration.

[h]PB: patient-based.

[i]EB: eye-based.

Quality Assessment and GRADE Certainty

The risk of bias and applicability concerns, as assessed by the PROBAST+AI tool, were summarized in Figure 2 and Tables S4-S5 in Multimedia Appendix 1. For the model development phase, 14% (4/28) of studies were judged to have a high overall risk of bias regarding quality [23,25,28,30], while none (0/28) raised high applicability concerns. For the model validation or testing phase, 25% (7/28) of studies were rated as having a high overall risk of bias [23,25,28,30,35,37,40], again with no studies (0/28) presenting high applicability concerns. Overall, the proportion of high-risk ratings was low, with most domains assessed as low risk, indicating an acceptable overall quality of the included literature.

XSL•FO
RenderX

Figure 2. Risk of bias and applicability concerns of the included studies regarding model development and model evaluation domains using the Prediction model Risk Of Bias ASsessment Tool for artificial intelligence tools. The colors represent the proportion of studies with low, high, or unclear risk. PROBAST+AI: Prediction model Risk Of Bias ASsessment Tool for artificial intelligence.

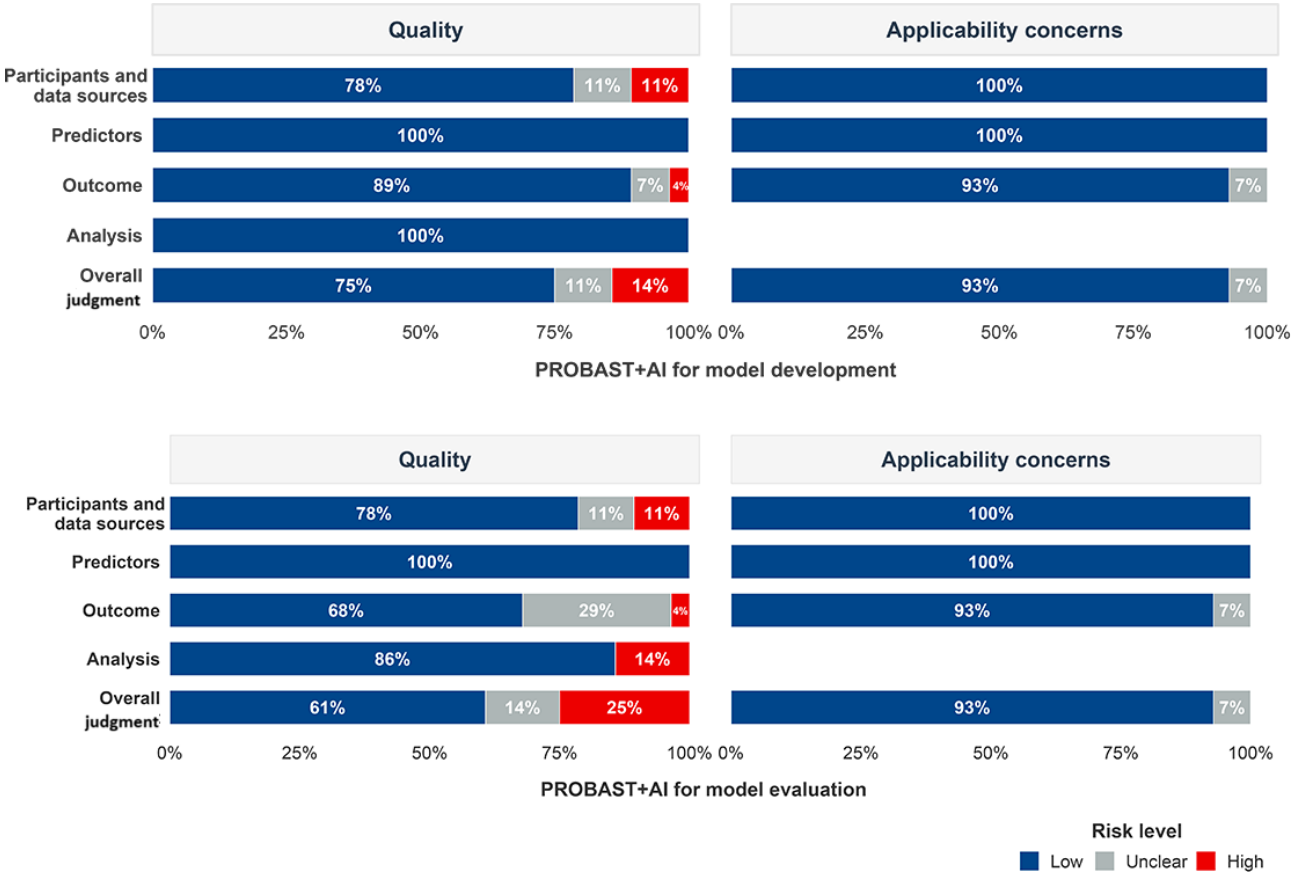

Using the GRADE framework, the certainty of evidence for the two primary diagnostic tasks was rated as moderate (Table 3 and Table S6 in Multimedia Appendix 1). Following the GRADE Summary of Findings format for diagnostic test accuracy, Table 3 reports pooled sensitivity and specificity, expected results per 1000 tested at a 20% pretest probability, PIs, certainty ratings, and plain-language interpretation. The detailed data extraction tables and PROBAST+AI risk-of-bias tables are retained in the Supplementary Materials because of their length (Tables S4-S5 and S7-S10 in Multimedia Appendix 1).

In Table 3, expected results were calculated per 1000 tested at an illustrative pretest probability of 20% (200 with the target condition and 800 without); this assumed prevalence may vary across clinical settings. Diagnostic accuracy was treated as a surrogate for patient-important consequences because direct evidence on whether artificial intelligence–assisted testing improves visual outcomes, referral burden, or treatment timing was unavailable. For age-related macular degeneration versus normal, the target condition was age-related macular degeneration; for wet age-related macular degeneration versus dry age-related macular degeneration, the target condition was wet age-related macular degeneration.

XSL•FO
RenderX

Table . GRADE[a] summary of findings table for deep learning–based AMD[b] image classification.

x test	No. of studies and validation datasets	Sensitivity (95% CI)	Specificity (95% CI)	Expected results per 1000 tested at 20% prevalence	95% prediction interval	Certainty	Plain-language interpretation
DL[c] for AMD versus normal	27 studies [19-36,38-46] and 37 validation datasets	0.98 (0.96 - 0.99)	0.98 (0.95 - 0.99)	• TP[d] 196 • FN[e] 4 • TN[f] 784 • FP[g] 16	• Sensitivity 0.95 - 0.99 • Specificity 0.95 - 0.99	Moderate[h]	At a 20% pretest probability, DL would correctly identify most AMD cases and correctly rule out most non-AMD eyes or images. The prediction intervals indicate that performance may vary across settings.
DL for wAMD[i] versus dAMD[j]	16 studies [19,23-29,31,33,37,41-44,46] and 22 validation datasets	0.95 (0.91 - 0.97)	0.95 (0.93 - 0.97)	• TP 190 FN 10 • TN 760 • FP 40	• Sensitivity 0.89 - 0.97 • Specificity 0.92 - 0.97	Moderate[h]	At a 20% pretest probability, DL would correctly identify most wAMD cases and correctly classify most dAMD cases. The wider prediction interval for sensitivity supports cautious local validation before deployment.

[a]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.

[b]AMD: age-related macular degeneration.

[c]DL: deep learning.

[d]TP: true positive.

[e]FN: false negative.

[f]TN: true negative.

[g]FP: false positive.

[h]Downgraded one GRADE level for risk of bias because most included studies were retrospective, many relied on internal validation or incompletely reported patient-level separation or reference standards, and PROBAST+AI identified high risk of bias in a subset of validation or testing studies; detailed domain-level judgments are provided in Table S6 in Multimedia Appendix 1.

[i]wAMD: wet age-related macular degeneration.

[j]dAMD: dry age-related macular degeneration.

Goodness-of-Fit and Model Diagnostics

The goodness-of-fit and bivariate normality of the bivariate mixed-effects regression models were evaluated through graphical inspection (Figures S1 and S2 in Multimedia Appendix 1). The goodness-of-fit plots (Figures S1A and S2A in Multimedia Appendix 1), illustrating the normal probability of deviance residuals, demonstrated that the observed data points adhered closely to the reference diagonal line, indicating a robust model fit. Similarly, the bivariate normality plots (Figures S1B and S2B in Multimedia Appendix 1) revealed a linear alignment of data points within the chi-square probability plots, confirming that the random effects of sensitivity and specificity followed a bivariate normal distribution. Collectively, these diagnostic assessments substantiate the validity and statistical robustness of the models used in this systematic review and meta-analysis.

DL Algorithms Versus Ophthalmologists for AMD Versus Normal Classification

The pooled diagnostic performance of image-based DL algorithms for distinguishing AMD from normal retinas was high across threshold-dependent metrics and threshold-independent discrimination. Sensitivity and specificity quantify performance at the diagnostic thresholds reported by individual studies, accuracy reflects overall correct classification in the analyzed datasets, and AUC describes discrimination across possible thresholds. The primary forest plots with PIs are shown in Figure 3, and the intervals should be interpreted as indicating expected between-setting variability rather than only statistical uncertainty around the pooled estimate.

Comparative analysis revealed that the pooled sensitivity of DL algorithms was significantly higher than that of senior

XSL•FO
RenderX

ophthalmologists (0.98 vs 0.75; Z=4.94; *P*<.001), as was the pooled accuracy (0.97 vs 0.83; Z=4.43; *P*<.001). These results were detailed in Figure 4 and Figures S5 and S6 in Multimedia Appendix 1. Notably, the comparison with junior ophthalmologists for AMD detection was limited to a single study [24], precluding robust statistical inference for this subgroup.

Figure 3. Primary forest plots for pooled sensitivity, specificity, and accuracy of deep learning algorithms for classifying age-related macular degeneration from normal retinas. Prediction intervals are printed in the plots to show expected between-setting variability [19-36,38-46].

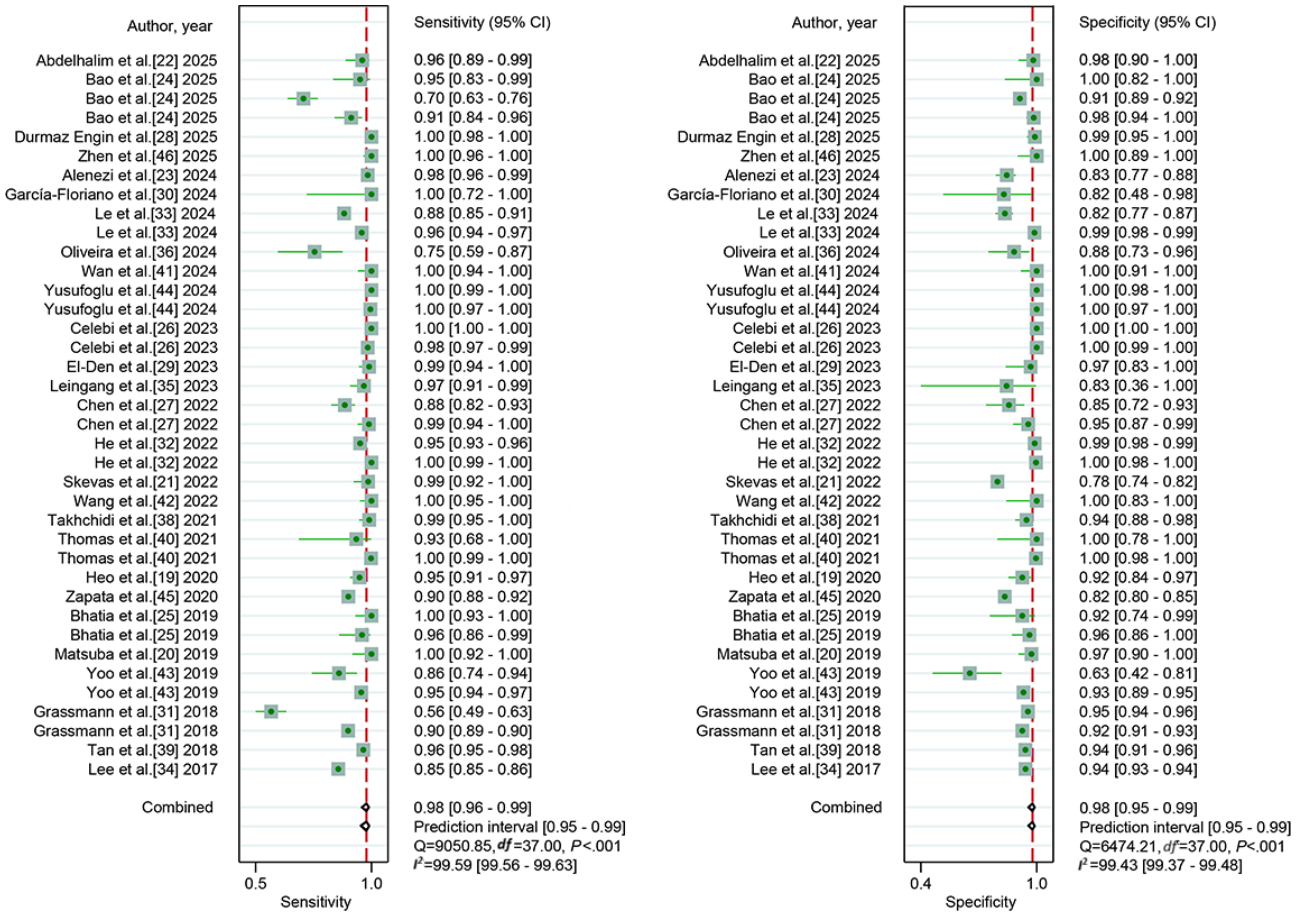

XSL•FO
RenderX

Figure 4. Violin plots comparing the diagnostic performance of deep learning algorithms versus junior and senior ophthalmologists. The top row displays the sensitivity, specificity, and accuracy for classifying age-related macular degeneration versus normal, while the bottom row displays the performance for classifying wet age-related macular degeneration versus dry age-related macular degeneration. The scattered dots represent individual study estimates, and the internal box plots indicate the median and interquartile range. *P* values indicate the statistical significance of the comparisons. AMD: age-related macular degeneration; dAMD: dry age-related macular degeneration; DL: deep learning; wAMD: wet age-related macular degeneration.

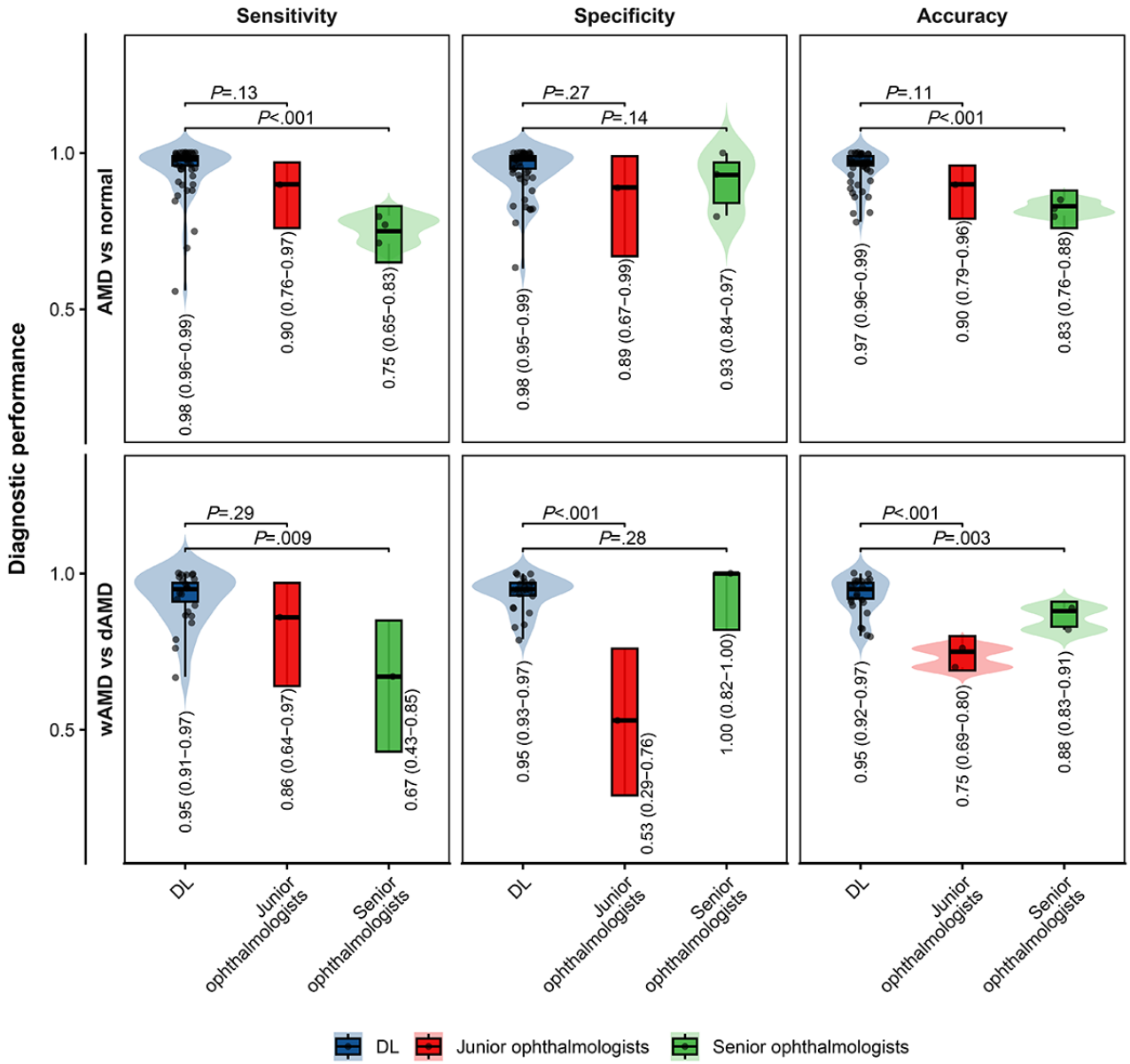

DL Algorithms Versus Ophthalmologists for wAMD Versus dAMD Classification

For the classification of wAMD versus dAMD, DL algorithms again showed high pooled sensitivity, specificity, accuracy, and AUC, but these metrics describe different behaviors. Accuracy summarizes correct classification within the included datasets, whereas AUC reflects discrimination across thresholds and may remain high even when real-world threshold selection, disease spectrum, or image quality differs. The primary forest plots with PIs are shown in Figure 5.

Comparative analyses showed that, within the limited head-to-head datasets, DL algorithms had higher pooled specificity (0.95 vs 0.53; Z=3.49; *P*<.001) and diagnostic accuracy (0.95 vs 0.75; Z=6.48; *P*<.001) than junior ophthalmologists. Compared with senior ophthalmologists, DL algorithms had higher pooled sensitivity (0.95 vs 0.67; Z=2.58; *P*=.009) and diagnostic accuracy (0.95 vs 0.88; Z=2.90; *P*=.003). All comparisons are shown in Figure 4 and Figures S5 and S6 in Multimedia Appendix 1. These findings suggest a possible relative performance advantage for DL in selected metrics, but the small number of clinician-comparison studies means that the results should be interpreted as preliminary rather than definitive.

XSL•FO
RenderX

Figure 5. Primary forest plots for pooled sensitivity, specificity, and accuracy of deep learning algorithms for classifying wet age-related macular degeneration from dry age-related macular degeneration. Prediction intervals are printed in the plots to show expected between-setting variability [19,23-29,31,33,37,41-44,46].

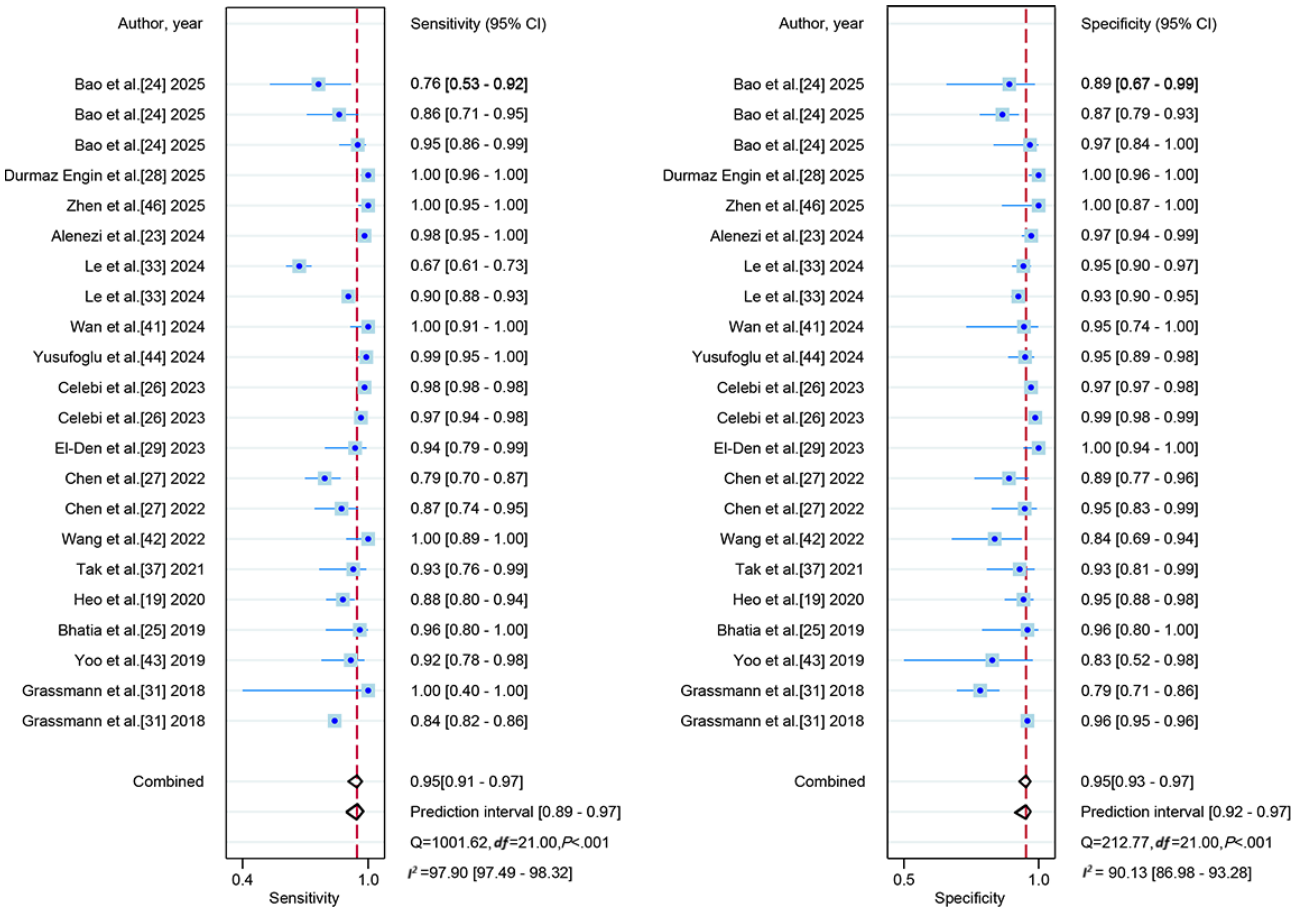

Subgroup Analysis for AMD Versus Normal Classification

Subgroup analysis based on imaging modality, as illustrated in Figure 6 and Figures S9-S17 in Multimedia Appendix 1, revealed statistically significant performance differences among DL algorithms. Specifically, OCT-based DL algorithms demonstrated significantly higher pooled specificity compared to CFP-based algorithms (0.99 vs 0.94; Z=2.92; *P*=.003). Furthermore, OCT-based algorithms achieved significantly higher pooled accuracy (0.99 vs 0.94; Z=3.36; *P*<.001) and a significantly higher pooled AUC value (1.00 vs 0.98; Z=3.50, *P*<.001) than CFP-based DL algorithms.

XSL•FO
RenderX

Figure 6. Subgroup analysis of deep learning algorithms based on imaging modalities. The violin plots illustrate the distribution of sensitivity, specificity, accuracy, and area under the curve for optical coherence tomography-based, color fundus photography-based, and multimodal models in classifying age-related macular degeneration versus normal (top row) and wet age-related macular degeneration versus dry age-related macular degeneration (bottom row). The scattered dots represent individual study estimates, and the internal box plots indicate the median and interquartile range. *P* values indicate the statistical significance of the comparisons. AMD: age-related macular degeneration; AUC: area under the curve; CFP: color fundus photography; dAMD: dry age-related macular degeneration; DL: deep learning; OCT: optical coherence tomography; wAMD: wet age-related macular degeneration.

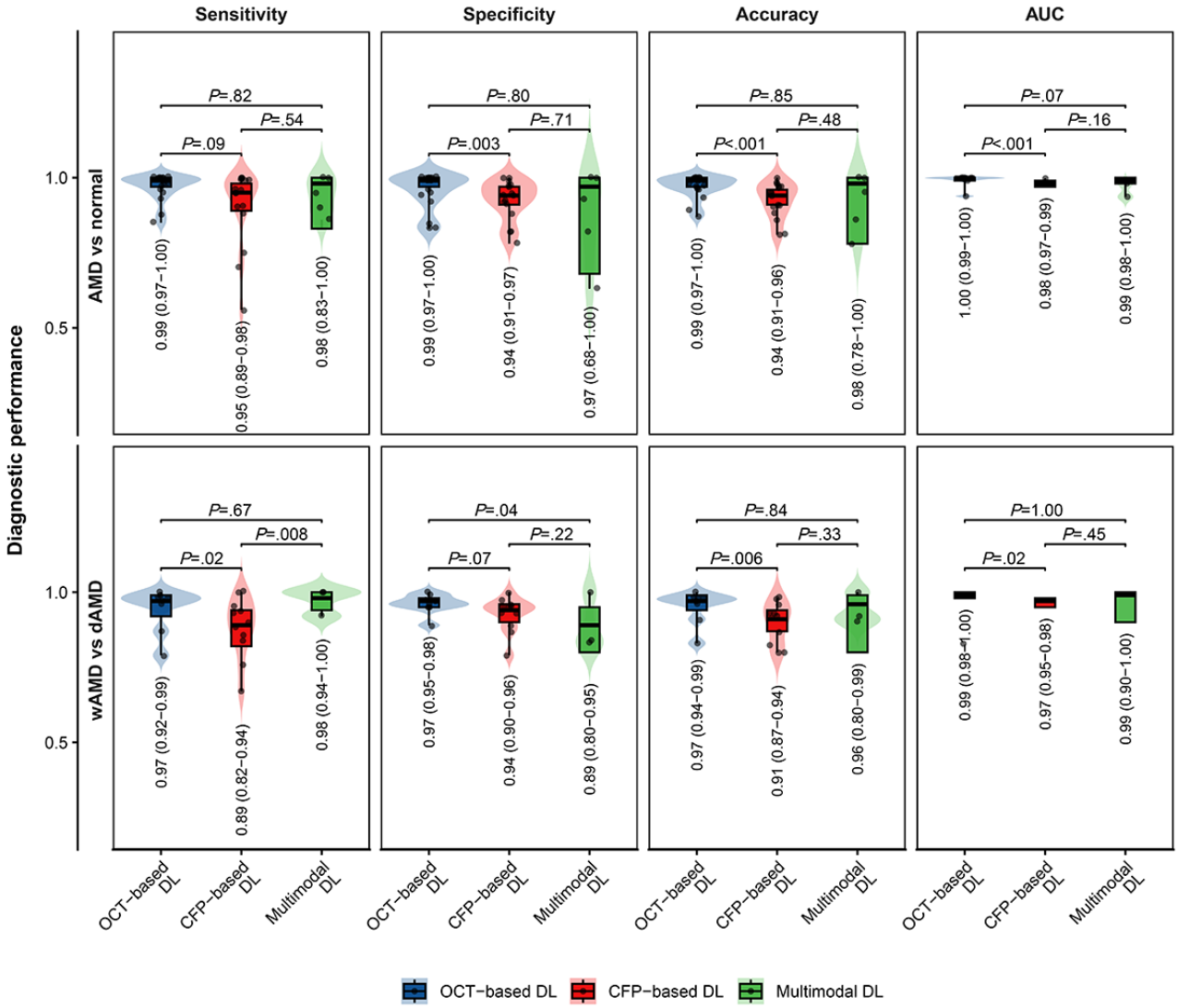

Subgroup Analysis for wAMD Versus dAMD Classification

The subgroup analysis by imaging modality for wAMD versus dAMD classification was presented in Figure 6 and Figures S18-S27 in Multimedia Appendix 1. OCT-based DL algorithms demonstrated significantly higher pooled sensitivity than CFP-based algorithms (0.97 vs 0.89; Z=2.25; *P*=.02). Their pooled accuracy was also significantly higher (0.97 vs 0.91; Z=2.73; *P*=.006), as was their pooled AUC value (0.99 vs 0.97; Z=2.17; *P*=.02).

Heterogeneity Investigation: Bivariate Boxplot and Meta-Regression

To explore the substantial statistical heterogeneity observed, meta-regression and bivariate boxplot analyses were conducted. For the AMD versus normal classification, meta-regression indicated that the type of validation (internal validation vs external validation), database source (open database vs private database), and study centers (single center vs multicenter) were potential sources of heterogeneity (Table S12 in Multimedia Appendix 1). The bivariate boxplot suggested that the studies by Skevas et al [21], Celebi et al [26], Grassmann et al [31], and Yoo et al [43], might be influential outliers contributing to the heterogeneity (Figure 7A).

For the wAMD versus dAMD classification, meta-regression identified the type of validation (internal validation vs external validation), type of imaging (unimodal vs multimodal), database source (open database vs private database), and study centers (single center vs multicenter) design as potential moderators explaining heterogeneity (Table S13 in Multimedia Appendix 1). The corresponding bivariate boxplot highlighted the studies by El-Den et al [29], Le et al [33], and Wang et al [42] as potential outliers influencing the pooled estimates (Figure 7B).

XSL•FO
RenderX

Figure 7. Assessment of heterogeneity and small-study effects. (A and B) Bivariate boxplots identifying potential outliers and influential studies for age-related macular degeneration versus normal and wet age-related macular degeneration versus dry age-related macular degeneration classifications. Studies falling outside the colored ellipses are considered outliers. (C and D) Deeks' funnel plots evaluating small-study effects for the two classification tasks; a *P* value >.10 indicates no significant small-study effects. AMD: age-related macular degeneration; dAMD: dry age-related macular degeneration; wAMD: wet age-related macular degeneration [21,26,29,31,33,42,43].

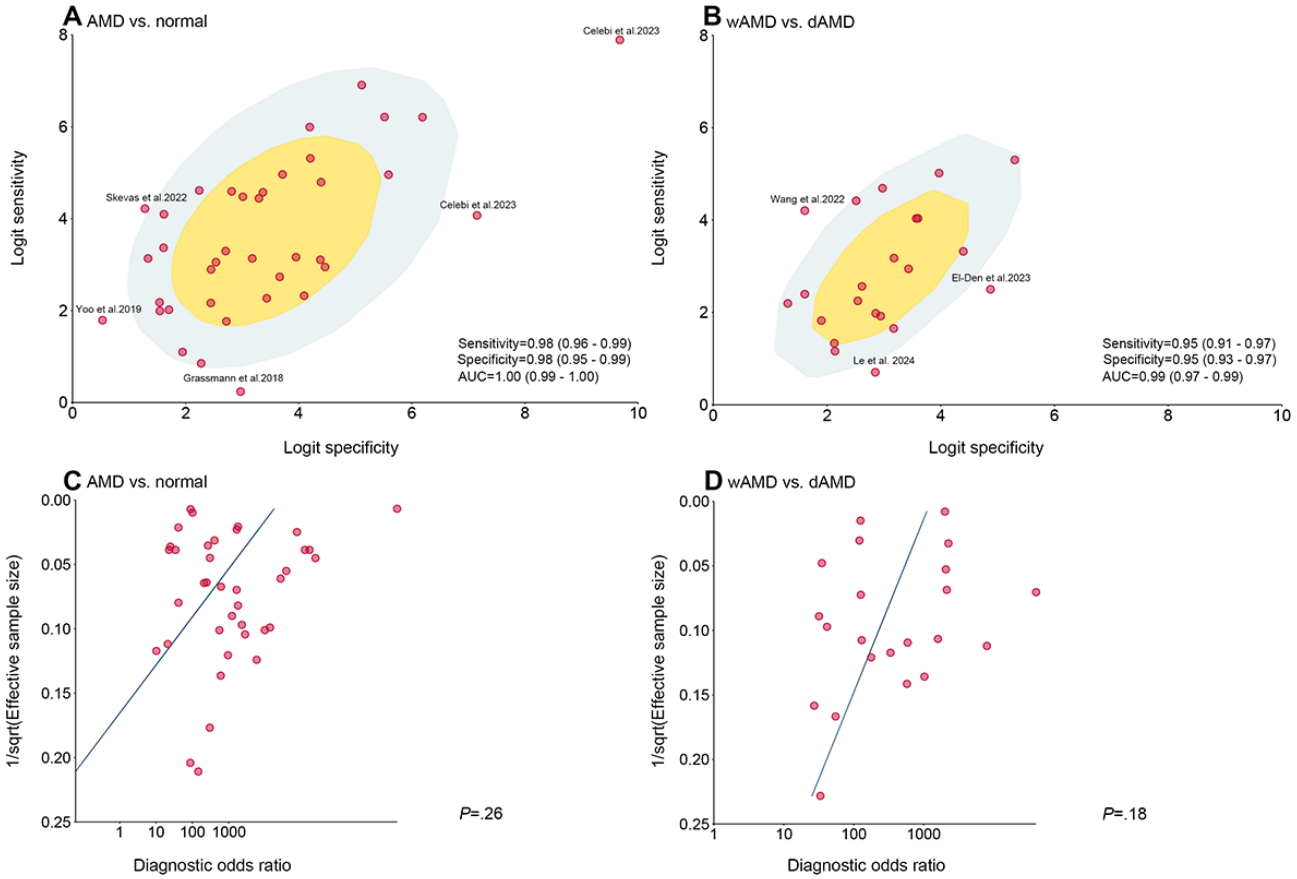

Sensitivity Analysis

Sensitivity analyses were conducted to assess the robustness of our findings. Whether excluding studies flagged as high-risk in the validation set by the PROBAST+AI tool, or removing the identified outliers and influential data points, the variations in diagnostic performance were stable (Tables S14 and S15 in Multimedia Appendix 1). The results remained consistent with the primary analysis, demonstrating that the conclusions are robust and not disproportionately driven by these extreme studies.

Small-Study Effects and Clinical Applicability

Deeks' funnel plot asymmetry test indicated no evidence of significant small-study effects for either AMD classification task (*P*=.26 and 0.18; Figures 7C and 7D). Assuming a pretest probability of 20%, Fagan nomogram analysis demonstrated that a positive DL test result increased the posttest probability to 91% for AMD versus normal and 84% for wAMD versus dAMD, while a negative result reduced it to 1% for both tasks (Figures 8A and 8B).

XSL•FO
RenderX

Figure 8. Fagan's nomograms evaluating the clinical utility of deep learning algorithms. (A) Clinical utility for classifying age-related macular degeneration versus normal. (B) Clinical utility for classifying wet age-related macular degeneration versus dry age-related macular degeneration. The left axis represents the pretest probability (set at 20%), the middle axis represents the likelihood ratio, and the right axis represents the posttest probability. AMD: age-related macular degeneration; dAMD: dry age-related macular degeneration; wAMD: wet age-related macular degeneration.

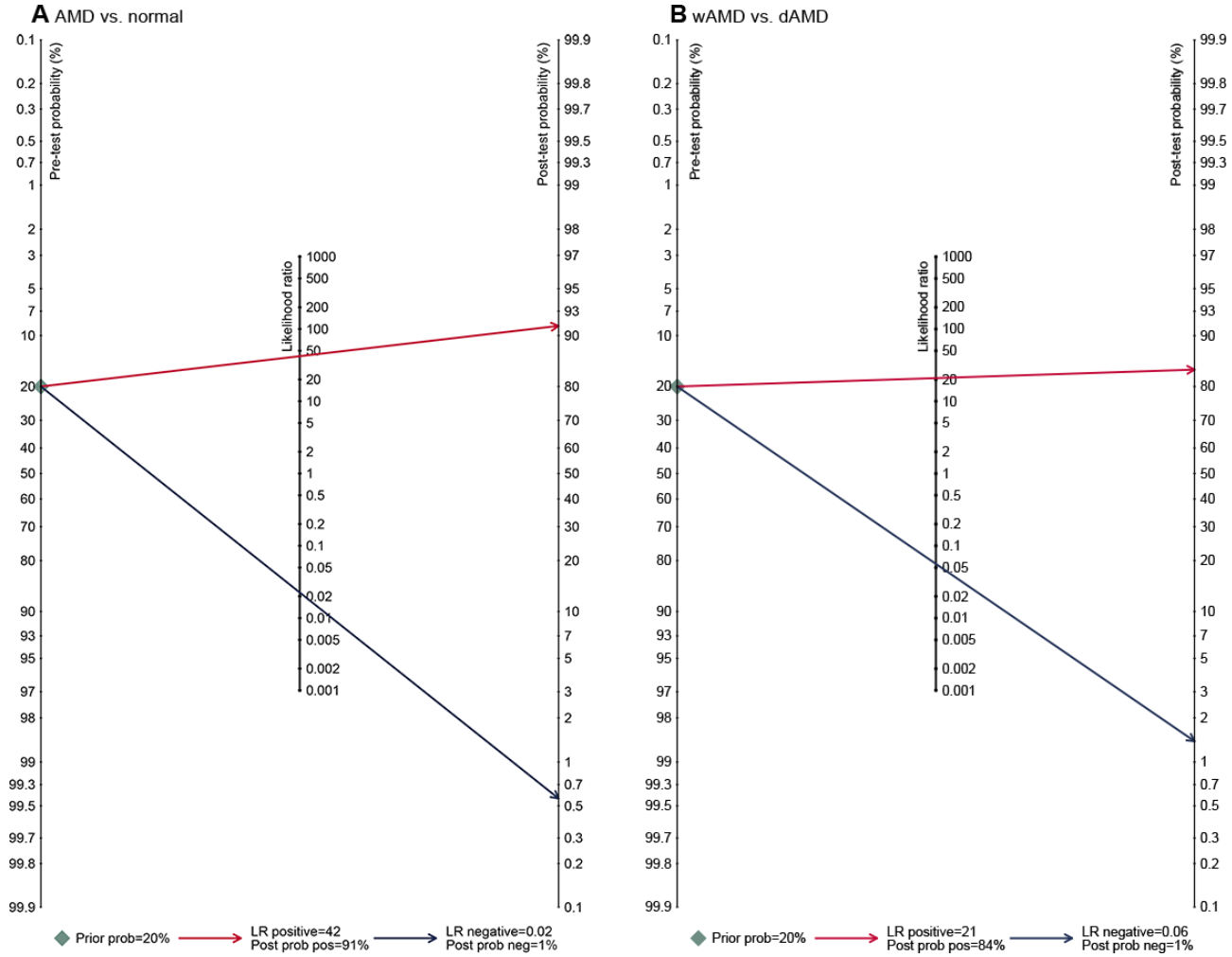

Discussion

In relation to our objective of comparing DL algorithms with ophthalmologists and identifying factors that influence diagnostic performance, the main finding is that DL models showed strong pooled performance for both AMD detection and wAMD versus dAMD classification, but the strength of this evidence differs across comparisons. The clinician comparisons suggest a possible role for DL as a consistent decision-support baseline, yet the sparse junior-ophthalmologist data and wide between-study variability mean that these findings should be interpreted as hypothesis-generating rather than definitive for deployment.

Our meta-analysis reveals a deployment-relevant pattern. Within the analyzed datasets, DL algorithms demonstrated significantly higher pooled sensitivity and accuracy compared to the available metrics for senior ophthalmologists in distinguishing AMD from normal controls; however, no significant differences were observed between DL and junior ophthalmologists across any diagnostic metrics. The higher pooled performance of DL suggests that these models possess an enhanced capability to detect subtle, pixel-level morphological changes and nonlinear feature interactions, such as early exudative signs, that may elude the visual inspection of even the most experienced clinicians [20,24,36]. Conversely, the lower sensitivity observed in senior ophthalmologists appears to be behaviorally driven by specific decision thresholds rather than skill deficiencies. As evidenced by the data from studies such as Matsuba et al [20], Bao et al [24], and Oliveira et al [36], senior experts exhibited a distinct preference for conservative diagnostic thresholds, achieving high specificity (summary 0.93) but notably reduced sensitivity (summary 0.75). This indicates a clinical preference for "rule-in" strategies to strictly avoid false positives, a constraint that DL algorithms do not possess. However, the unexpected parity between DL algorithms and junior practitioners should be interpreted with caution; this finding is attributable to data sparsity rather than clinical equivalence, as the analysis is restricted to only one study for AMD detection and two studies for wAMD classification, introducing a high risk of small-sample bias [20,24,36]. Similarly, the comparisons with senior ophthalmologists, while more robust, remain limited in number (three studies for AMD detection, two for wAMD classification), and the resulting estimates should be considered preliminary rather than definitive. This data-driven limitation, however, stands in contrast to the wAMD versus dAMD task, where sufficient

XSL•FO
RenderX

head-to-head comparisons (two studies) enabled a stratified analysis, revealing distinct behavioral patterns across experience levels [20,24,36].

The very high AUC values observed in our analysis warrant careful interpretation and should not be equated with flawless real-world diagnostic performance. Restricted test distributions, curated image quality, internal validation, repeated use of public datasets, model selection based on the best-performing algorithm, and insufficiently documented patient-level splitting may inflate discrimination [25,38,43,46]. Potential data leakage cannot be excluded in studies that did not clearly separate patients, eyes, or images across training, validation, and test sets. Therefore, AUCs close to 1.0 indicate excellent discrimination within the analyzed datasets, not proof of flawless performance in prospective clinical workflows [47,48].

In the clinically critical task of differentiating wAMD from dAMD, automated DL systems demonstrated robustness across varying datasets. Our results indicated that the pooled metrics of DL algorithms not only showed higher specificity and accuracy compared to junior ophthalmologists but also indicated higher sensitivity and accuracy relative to senior ophthalmologists. Rather than framing this as DL mitigating an "experience gap" or correcting specific human errors, these findings suggest that DL algorithms offer a more consistent and objective diagnostic baseline that balances sensitivity and specificity. These findings advocate for a collaborative clinical paradigm: DL algorithms could serve as a triage filter to enhance specificity for primary care providers while functioning as a high-sensitivity "second reader" for specialists resolving equivocal wAMD cases [49].

Interestingly, our subgroup analysis highlights the relatively consistent performance of OCT-based models for automated AMD classification. OCT-based models significantly outperformed CFP-based approaches, driven by the capture of pathognomonic cross-sectional features—such as intraretinal fluid and pigment epithelial detachment—that are often obscured in 2D fundus photography [50]. Intriguingly, multimodal DL (OCT + CFP) did not significantly surpass standalone OCT models. This suggests a "saturation effect," where the rich structural data of OCT capture the vast majority of diagnostic signals, rendering the incremental value of CFP marginal [35]. In practice, the underperformance of multimodal models relative to standalone OCT may also stem from feature redundancy and fusion noise [51]; when OCT and CFP capture substantially overlapping diagnostic information [43], their combination can paradoxically introduce variance through misaligned spatial features, registration errors, and conflicting feature representations, ultimately degrading rather than enhancing the decision boundary. From a translational perspective, this finding is pivotal; it implies that the computational cost and technical challenges of multimodal alignment (eg, fusion noise and registration errors) may currently outweigh the clinical benefits [51,52]. Therefore, unless fusion strategies are substantially optimized, OCT-based workflows currently appear to be a practical foundation for clinical deployment [35].

Building upon the baselines established by previous meta-analyses, this systematic review advances the understanding of DL in AMD diagnosis. In 2023, Leng et al [9] reported a pooled sensitivity of 94% and specificity of 97% for convolutional neural network algorithms. More recently, Chen et al [27] highlighted the superiority of AI over retinal specialists. Our analysis incorporates the latest studies using advanced architectures, such as Vision Transformers. This inclusion yields modestly higher pooled metrics, reflecting the field's technological maturation [33]. Most significantly, this systematic review distinguishes itself through four methodological innovations that enhance clinical relevance: (1) a stratified comparison of AI versus ophthalmologists, explicitly differentiating by experience level; (2) the application of the PROBAST+AI tool for bias assessment, complemented by the GRADE framework; (3) a rigorous subgroup analysis by imaging modality (OCT, CFP, multimodal) to isolate technical performance drivers; and (4) a granular evaluation extending beyond binary detection to the specific classification of wAMD versus dAMD. Collectively, these advancements establish a more robust evidence base than prior reviews.

Heterogeneity is central to the interpretation of these findings [53]. Extreme between-study variability should not be treated only as a statistical descriptor; it indicates that pooled estimates may not transfer reliably to clinics with different devices, acquisition protocols, labeling rules, disease spectra, or patient populations [53-55]. By using a bivariate random-effects model and multivariable meta-regression, we identified that differences in validation strategies (internal vs external), database sources (open vs. private), and study settings (single-center vs. multicenter) significantly influence diagnostic performance. Studies relying solely on internal validation frequently reported inflated metrics, illustrating a generalization gap when models face domain shifts in image acquisition or demographics [50]. Similarly, single-center studies risk overfitting to specific center features arising from uniform protocols, whereas multicenter designs typically demonstrate greater robustness through exposure to diverse image qualities [50]. Consequently, our analysis suggests that database diversity and annotation quality are likely more critical determinants of generalizability than mere data accessibility. Furthermore, specific outliers in the bivariate box plot (Skevas et al [21], Celebi et al [26], Grassmann et al [31], El-Den et al [29], Le et al [33], Wang et al [42], and Yoo et al [43]) highlight how methodological divergences, such as algorithm architecture and data curation strategies, can materially affect performance. This indicates that future improvements in AI reliability will depend less on novel model architectures and more on the curation of diverse, multicenter external validation datasets. The substantial heterogeneity observed in this systematic review and meta-analysis warrants careful interpretation. Rather than reflecting routine statistical noise, this level of heterogeneity signals the pooling of fundamentally diverse data sources. Specifically, the included studies used different imaging hardware (eg, Heidelberg Spectralis, Topcon, Zeiss Cirrus OCT devices; various fundus camera systems), acquisition protocols (varying image resolutions, fields of view, and scan patterns), and ground-truth labeling methodologies (ranging from consensus grading by multiple retinal specialists to single-expert annotation or semi-automated classification systems) [32,35,45]. These technical and methodological differences fundamentally

XSL•FO
RenderX

influence the feature space available to DL algorithms and likely account for much of the observed heterogeneity. The PIs (eg, sensitivity: 0.95 - 0.99; specificity: 0.95 - 0.99 for distinguishing AMD from normal retinas; sensitivity: 0.89 - 0.97; specificity: 0.92 - 0.97 for classifying wAMD vs dAMD) further underscore that the average pooled performance, while encouraging, may not be representative of performance in any individual deployment setting. This finding has important implications for clinical deployment: site-specific validation using local imaging equipment and patient populations remains essential before implementing any DL-based AMD screening system.

These heterogeneity findings provide concrete guidance for future study design [56]. Investigators should prioritize external validation on datasets from institutions and populations distinct from the training data [53], use multicenter designs incorporating diverse imaging devices and acquisition protocols, report patient-level separation between training and testing data, and stratify performance by imaging device, acquisition protocol, labeling method, and patient demographics [56]. Because PIs indicate that local performance may differ from pooled estimates, summary metrics alone should not be used as a deployment decision rule [54].

Translating these findings into practice, DL algorithms exhibit the potential to augment the diagnostic workflow rather than replace it. The superior accuracy of DL in classifying wAMD versus dAMD suggests potential use in resource-limited settings or tele-ophthalmology screening. However, considering current algorithms are primarily trained on isolated OCT or CFP images, they often lack integration with other imaging modalities or clinical parameters; future models should therefore evaluate multimodal imaging and patient clinical contexts to emulate comprehensive diagnoses [57]. Beyond these technical and clinical considerations, significant implementation barriers persist, including the scarcity of expert-annotated data, regulatory hurdles, and technical challenges regarding data availability, model interpretability, transparency, and generalization capability [36,52,58]. Advances in few-shot learning, self-supervised models, and centralized platforms may support a more integrated AI ecosystem, requiring sustained multidisciplinary efforts to optimize AI safety and support safe clinical practice [52,58].

Beyond diagnostic performance metrics, the successful clinical translation of DL algorithms requires addressing practical implementation challenges. These include seamless integration into existing electronic health record systems and ophthalmic imaging workflows, real-time processing capabilities compatible with clinical time constraints, and intuitive user interfaces that present AI-generated results in a manner that supports rather than disrupts clinical decision-making [46]. Furthermore, clinician trust and acceptance—shaped by model interpretability, transparency of AI reasoning, and consistent performance across diverse clinical scenarios—are prerequisites for successful adoption [23,24]. Future validation of DL tools must therefore extend beyond accuracy benchmarks to encompass usability studies, clinician acceptance evaluations, and workflow efficiency assessments in real-world clinical settings [41].

Our findings should be interpreted considering several limitations. First, the predominance of retrospective study designs (26 of 28 included studies) represents a fundamental limitation that must be carefully considered when interpreting the strong pooled performance metrics. Retrospective datasets are typically curated from clinical archives, which may systematically exclude poor-quality images, atypical presentations, and diagnostically challenging cases that are routinely encountered in prospective clinical workflows. This selection inherently inflates the apparent diagnostic performance and limits the generalizability of our findings to real-world screening and clinical deployment settings [48]. Second, to address potential patient overlap and maintain statistical independence, we extracted performance metrics exclusively from the primary AI algorithm within each study, omitting data from suboptimal models. While methodologically sound for meta-analysis, this approach inherently reflects a “best-case scenario” that likely inflates the pooled performance estimates compared to average algorithmic performance. This reporting bias is an inherent limitation of the current DL literature in ophthalmology and should be carefully considered by clinicians and policymakers when interpreting these results for clinical implementation decisions [12]. Future research should therefore granularly evaluate performance variances across different algorithmic architectures, including less optimal models, to ensure a more balanced and realistic assessment of the DL landscape. Third, direct head-to-head comparisons between DL algorithms and ophthalmologists were small, particularly for the AMD versus normal task where only one study provided junior ophthalmologist data, limiting the statistical power of these specific subgroups [50]. Future research must prioritize prospective, multicenter trials with prespecified human comparison arms to definitively validate these retrospective results [12].

In conclusion, this systematic review suggests that, compared with ophthalmologists, DL algorithms demonstrate superior and more balanced diagnostic performance for AMD image classification, providing a consistent decision-support baseline that mitigates the threshold-dependent trade-offs observed in human graders. However, these relative-performance findings remain preliminary because head-to-head evidence is sparse, especially for junior ophthalmologists, and because wide PIs, high heterogeneity, retrospective designs, and possible inflation from restricted datasets, internal validation, or leakage limit clinical transportability. DL systems should therefore be locally calibrated and prospectively validated as triage adjuncts rather than autonomous replacements. Before implementation, prospective multicenter studies should test representative patients, use strict patient-level external validation [56], include prespecified human comparison arms [59], and evaluate workflow integration, interpretability, and safety [60].

XSL•FO
RenderX

Acknowledgments

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision:

- Text generation

- Proofreading and editing

- Summarizing text

- Formulation of conclusions

- Translation

- Reformatting

The GenAI tool used was: OpenAI Codex (GPT-5).

Responsibility for the final manuscript lies entirely with the authors.

GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Declaration submitted by: The primary author (YZ)

Funding

The authors declared no financial support was received for this work.

Data Availability

All data generated or analyzed during this systematic review and meta-analysis are included in this published article and its supplementary information files.

Authors' Contributions

Conceptualization, methodology, software, formal analysis, investigation (screening, quality assessment, and data extraction), data curation, writing – original draft, visualization: YZ
Investigation (literature search and screening): YN
Investigation (study selection): SYS
Investigation (quality assessment): WL
Data curation, Investigation (data extraction): YD
Supervision, Validation, Writing – review and editing: YG
All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed methodology, search strategies, quality assessments, subgroup analyses, and sensitivity analyses.
[PDF File, 10176 KB - jmir_v28i1e97174_app1.pdf]

Checklist 1
PRISMA checklist.
[PDF File, 220 KB - jmir_v28i1e97174_app2.pdf]

References

1. Jeong YD, Park S, Kim MS, et al. Global burden of vision impairment due to age-related macular degeneration, 1990–2021, with forecasts to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet Glob Health 2025 Jul;13(7):e1175-e1190. [doi: 10.1016/S2214-109X(25)00143-3]
2. Marchesi N, Capierri M, Pascale A, Barbieri A. Different therapeutic approaches for dry and wet AMD. Int J Mol Sci 2024 Dec 4;25(23):13053. [doi: 10.3390/ijms252313053] [Medline: 39684764]
3. Midena E, Frizziero L, Torresin T, Boscolo Todaro P, Miglionico G, Pilotto E. Optical coherence tomography and color fundus photography in the screening of age-related macular degeneration: a comparative, population-based study. PLoS ONE 2020;15(8):e0237352. [doi: 10.1371/journal.pone.0237352] [Medline: 32797085]

XSL•FO
RenderX

4. Trinh M, Cheung R, Nam J, Ng D, Nivison-Smith L, Ly A. High risk does not guarantee high accuracy-evaluating the prognostic accuracy of OCT biomarkers for predicting late AMD. Ophthalmic Physiol Opt 2025 Sep;45(6):1293-1301. [doi: 10.1111/opo.13547] [Medline: 40557781]
5. Moradi M, Chen Y, Du X, Seddon JM. Deep ensemble learning for automated non-advanced AMD classification using optimized retinal layer segmentation and SD-OCT scans. Comput Biol Med 2023 Mar;154(106512):106512. [doi: 10.1016/j.compbiomed.2022.106512] [Medline: 36701964]
6. Neri G, Rebecchi C, Oakley JD, et al. Deep learning model for automated classification of macular neovascularization subtypes in AMD. Invest Ophthalmol Vis Sci 2025 Jul 1;66(9):55. [doi: 10.1167/iovs.66.9.55] [Medline: 40689724]
7. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019 Oct;1(6):e271-e297. [doi: 10.1016/S2589-7500(19)30123-2] [Medline: 33323251]
8. Pandey PU, Ballios BG, Christakis PG, et al. Ensemble of deep convolutional neural networks is more accurate and reliable than board-certified ophthalmologists at detecting multiple diseases in retinal fundus photographs. Br J Ophthalmol 2024 Feb 21;108(3):417-423. [doi: 10.1136/bjo-2022-322183] [Medline: 36720585]
9. Leng X, Shi R, Wu Y, et al. Deep learning for detection of age-related macular degeneration: a systematic review and meta-analysis of diagnostic test accuracy studies. PLoS ONE 2023;18(4):e0284060. [doi: 10.1371/journal.pone.0284060] [Medline: 37023082]
10. Chen KY, Chan HC, Chan CM. Can artificial intelligence with multimodal imaging outperform traditional methods in predicting age-related macular degeneration progression? A systematic review and exploratory meta-analysis. BMC Med Inform Decis Mak 2025 Sep 1;25(1):321. [doi: 10.1186/s12911-025-03119-z] [Medline: 40890721]
11. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ 2025 Mar 24;388:e082505. [doi: 10.1136/bmj-2024-082505] [Medline: 40127903]
12. Vrudhula A, Kwan AC, Ouyang D, Cheng S. Machine learning and bias in medical imaging: opportunities and challenges. Circ Cardiovasc Imaging 2024 Feb;17(2):e015495. [doi: 10.1161/CIRCIMAGING.123.015495] [Medline: 38377237]
13. Salameh JP, Bossuyt PM, McGrath TA, et al. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. BMJ 2020 Aug 14;370:m2632. [doi: 10.1136/bmj.m2632] [Medline: 32816740]
14. Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MGM, Stijnen T. Bivariate random effects meta-analysis of ROC curves. Med Decis Making 2008;28(5):621-638. [doi: 10.1177/0272989X08319957] [Medline: 18591542]
15. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
16. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023 Dec;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
17. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol 2005 Sep;58(9):882-893. [doi: 10.1016/j.jclinepi.2005.01.016] [Medline: 16085191]
18. Lau J, Ioannidis JPA, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. BMJ 2006 Sep 16;333(7568):597-600. [doi: 10.1136/bmj.333.7568.597] [Medline: 16974018]
19. Heo TY, Kim KM, Min HK, et al. Development of a deep-learning-based artificial intelligence tool for differential diagnosis between dry and neovascular age-related macular degeneration. Diagnostics (Basel) 2020 Apr 28;10(5):261. [doi: 10.3390/diagnostics10050261] [Medline: 32354098]
20. Matsuba S, Tabuchi H, Ohsugi H, et al. Accuracy of ultra-wide-field fundus ophthalmoscopy-assisted deep learning, a machine-learning technology, for detecting age-related macular degeneration. Int Ophthalmol 2019 Jun;39(6):1269-1275. [doi: 10.1007/s10792-018-0940-0] [Medline: 29744763]
21. Skevas C, Weindler H, Levering M, Engelberts J, van Grinsven M, Katz T. Simultaneous screening and classification of diabetic retinopathy and age-related macular degeneration based on fundus photos-a prospective analysis of the RetCAD system. Int J Ophthalmol 2022;15(12):1985-1993. [doi: 10.18240/ijo.2022.12.14] [Medline: 36536981]
22. Abdelhalim I, Nadmid N, Elsharkawy M, Ghazal M, Mahmoud AH, El-Baz A. Mask-unmask regions (MUMR) framework for classifying AMD grades using inter-regional interaction analysis. IEEE ACCESS 2025;13:8286-8296. [doi: 10.1109/ACCESS.2025.3526948]
23. Alenezi A, Alhamad H, Brindhaban A, Amizadeh Y, Jodeiri A, Danishvar S. Enhancing readability and detection of age-related macular degeneration using optical coherence tomography imaging: an AI approach. Bioengineering (Basel) 2024 Mar 22;11(4):300. [doi: 10.3390/bioengineering11040300] [Medline: 38671722]
24. Bao S, Yang Z, Zhang Z, Qu J, Sun J. AttResAMD: an attention-driven deep learning framework for expert-level automated classification of age-related macular degeneration from fundus photography. Interdiscip Sci 2025 Aug 30. [doi: 10.1007/s12539-025-00763-x] [Medline: 40885885]

XSL•FO
RenderX

25. Bhatia KK, Graham MS, Terry L, et al. Disease classification of macular optical coherence tomography scans using deep learning software: validation on independent, multicenter data. Retina (Philadelphia, Pa) 2020 Aug;40(8):1549-1557. [doi: 10.1097/IAE.0000000000002640] [Medline: 31584557]
26. Celebi ARC, Bulut E, Sezer A. Artificial intelligence based detection of age-related macular degeneration using optical coherence tomography with unique image preprocessing. Eur J Ophthalmol 2023 Jan;33(1):65-73. [doi: 10.1177/11206721221096294] [Medline: 35469472]
27. Chen M, Jin K, Yan Y, et al. Automated diagnosis of age-related macular degeneration using multi-modal vertical plane feature fusion via deep learning. Med Phys 2022 Apr;49(4):2324-2333. [doi: 10.1002/mp.15541] [Medline: 35172022]
28. Durmaz Engin C, Beşenk U, Özizmirliler D, Selver MA. Comparative analysis of automated vs. expert-designed machine learning models in age-related macular degeneration detection and classification. Turk J Ophthalmol 2025 Jun 25;55(3):120-126. [doi: 10.4274/tjo.galenos.2025.74780] [Medline: 40560103]
29. El-Den NN, Naglah A, Elsharkawy M, et al. Scale-adaptive model for detection and grading of age-related macular degeneration from color retinal fundus images. Sci Rep 2023 Jun 13;13(1):9590. [doi: 10.1038/s41598-023-35197-2] [Medline: 37311794]
30. García-Floriano A, Ventura-Molina E. Age-related macular degeneration detection in retinal fundus images by a deep convolutional neural network. MATHEMATICS 2024;12(10):1445. [doi: 10.3390/math12101445]
31. Grassmann F, Mengelkamp J, Brandl C, et al. A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. Ophthalmology 2018 Sep;125(9):1410-1420. [doi: 10.1016/j.ophtha.2018.02.037] [Medline: 29653860]
32. He T, Zhou Q, Zou Y. Automatic detection of age-related macular degeneration based on deep learning and local outlier factor algorithm. Diagnostics (Basel) 2022 Feb 18;12(2):532. [doi: 10.3390/diagnostics12020532] [Medline: 35204621]
33. Le NT, Le Truong T, Deelertpaiboon S, et al. ViT - AMD: a new deep learning model for age - related macular degeneration diagnosis from fundus images. INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS 2024 Jan;2024(1). [doi: 10.1155/2024/3026500]
34. Lee CS, Baughman DM, Lee AY. Deep learning is effective for the classification of OCT images of normal versus Age-related Macular Degeneration. Ophthalmol RETINA 2017;1(4):322-327. [doi: 10.1016/j.oret.2016.12.009] [Medline: 30693348]
35. Leingang O, Riedl S, Mai J, et al. Automated deep learning-based AMD detection and staging in real-world OCT datasets (PINNACLE study report 5). Sci Rep 2023 Nov 9;13(1):19545. [doi: 10.1038/s41598-023-46626-7] [Medline: 37945665]
36. Oliveira GC, Rosa GH, Pedronette DCG, et al. Robust deep learning for eye fundus images: bridging real and synthetic data for enhancing generalization. Biomed Signal Process Control 2024 Aug;94:106263. [doi: 10.1016/j.bspc.2024.106263]
37. Tak N, Reddy AJ, Martel J, Martel JB. Clinical wide-field retinal image deep learning classification of exudative and non-exudative age-related macular degeneration. Cureus 2021 Aug;13(8):e17579. [doi: 10.7759/cureus.17579] [Medline: 34646633]
38. Takhchidi K, Gliznitsa P, Svetozarskiy S, Bursov A, Shusterzon K. Labelling of data on fundus color pictures used to train a deep learning model enhances its macular pathology recognition capabilities. BRSMU 2021 Aug(2021(4). [doi: 10.24075/brsmu.2021.040]
39. Tan JH, Bhandary SV, Sivaprasad S, et al. Age-related macular degeneration detection using deep convolutional neural network. FUTURE GENERATION COMPUTER SYSTEMS 2018 Oct;87:127-135. [doi: 10.1016/j.future.2018.05.001]
40. Thomas A, Harikrishnan PM, Gopi VP, Palanisamy P. An automated method to detect age-related macular degeneration from optical coherence tomographic images. Biomed Eng Appl Basis Commun 2021 Oct;33(5). [doi: 10.4015/S1016237221500368]
41. Wan C, Zhao JN, Hong XQ, Yang WH, Zhang SC. HCSP-Net: a novel model of age-related macular degeneration classification based on color fundus photography. CMC 2024;79(1):391-407. [doi: 10.32604/cmc.2024.048307]
42. Wang W, Li X, Xu Z, et al. Learning two-stream CNN for multi-modal age-related macular degeneration categorization. IEEE J Biomed Health Inform 2022 Aug;26(8):4111-4122. [doi: 10.1109/JBHI.2022.3171523] [Medline: 35503853]
43. Yoo TK, Choi JY, Seo JG, Ramasubramanian B, Selvaperumal S, Kim DW. The possibility of the combination of OCT and fundus images for improving the diagnostic accuracy of deep learning for age-related macular degeneration: a preliminary experiment. Med Biol Eng Comput 2019 Mar;57(3):677-687. [doi: 10.1007/s11517-018-1915-z] [Medline: 30349958]
44. Yusufoğlu E, Fırat H, Üzen H, et al. A Comprehensive CNN model for age-related macular degeneration classification using OCT: integrating inception modules, SE blocks, and ConvMixer. Diagnostics (Basel) 2024 Dec 17;14(24):2836. [doi: 10.3390/diagnostics14242836] [Medline: 39767197]
45. Zapata MA, Royo-Fibla D, Font O, et al. Artificial intelligence to identify retinal fundus images, quality validation, laterality evaluation, macular degeneration, and suspected glaucoma. Clin Ophthalmol 2020;14(419-29):419-429. [doi: 10.2147/OPTH.S235751] [Medline: 32103888]
46. Zhen B, Qi Y, Tang Z, et al. Low-rank fine-tuning meets cross-modal analysis: a robust framework for age-related macular degeneration categorization. J Digit Imaging Inform med 2025;39(1):714-731. [doi: 10.1007/s10278-025-01513-7]
47. Rosenblatt M, Tejavibulya L, Jiang R, Noble S, Scheinost D. Data leakage inflates prediction performance in connectome-based machine learning models. Nat Commun 2024 Feb 28;15(1):1829. [doi: 10.1038/s41467-024-46150-w] [Medline: 38418819]

XSL•FO
RenderX

48. Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ Digit Med 2022 Apr 12;5(1):48. [doi: 10.1038/s41746-022-00592-y] [Medline: 35413988]
49. Yim J, Chopra R, Spitz T, et al. Predicting conversion to wet age-related macular degeneration using deep learning. Nat Med 2020 Jun;26(6):892-899. [doi: 10.1038/s41591-020-0867-7]
50. Miladinović A, Biscontin A, Ajčević M, et al. Evaluating deep learning models for classifying OCT images with limited data and noisy labels. Sci Rep 2024 Dec 5;14(1):30321. [doi: 10.1038/s41598-024-81127-1] [Medline: 39638854]
51. Li Y, El Habib Daho M, Conze PH, et al. A review of deep learning-based information fusion techniques for multimodal medical image classification. Comput Biol Med 2024 Jul;177(108635):108635. [doi: 10.1016/j.compbiomed.2024.108635]
52. Sosna J, Joskowicz L, Saban M. Navigating the AI landscape in medical imaging: a critical analysis of technologies, implementation, and implications. Radiology 2025 Jun;315(3):e240982. [doi: 10.1148/radiol.240982] [Medline: 40552997]
53. Rashidisabet H, Sethi A, Jindarak P, et al. Validating the generalizability of ophthalmic artificial intelligence models on real-world clinical data. Transl Vis Sci Technol 2023 Nov 1;12(11):8. [doi: 10.1167/tvst.12.11.8] [Medline: 37922149]
54. Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021 Jul 15;385(3):283-286. [doi: 10.1056/NEJMc2104626] [Medline: 34260843]
55. Choi Y, Yu W, Nagarajan MB, et al. Translating AI to clinical practice: overcoming data shift with explainability. Radiographics 2023 May;43(5):e220105. [doi: 10.1148/rg.220105] [Medline: 37104124]
56. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024 Apr 16;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
57. Rao VM, Hla M, Moor M, et al. Multimodal generative AI for medical image interpretation. Nature New Biol 2025 Mar;639(8056):888-896. [doi: 10.1038/s41586-025-08675-y] [Medline: 40140592]
58. Champendal M, Müller H, Prior JO, Dos Reis CS. A scoping review of interpretability and explainability concerning artificial intelligence methods in medical imaging. Eur J Radiol 2023 Dec;169:111159. [doi: 10.1016/j.ejrad.2023.111159] [Medline: 37976760]
59. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ 2022 May 18;377:e070904. [doi: 10.1136/bmj-2022-070904] [Medline: 35584845]
60. Sandhu S, Lin AL, Brajer N, et al. Integrating a machine learning system into clinical workflows: qualitative study. J Med Internet Res 2020 Nov 19;22(11):e22421. [doi: 10.2196/22421] [Medline: 33211015]

Abbreviations

AI: artificial intelligence
AMD: age-related macular degeneration
AUC: area under the curve
CFP: color fundus photography
dAMD: dry age-related macular degeneration
DL: deep learning
FN: false negative
FP: false positive
GRADE: Grading of Recommendations, Assessment, Development, and Evaluations
OCT: optical coherence tomography
PI: prediction interval
PITROS: Patient, Index test, Target condition, Reference standard, Outcome, and Setting
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy
PROBAST+AI: Prediction model Risk Of Bias Assessment Tool for Artificial Intelligence
PROSPERO: International Prospective Register of Systematic Reviews
TN: true negative
TP: true positive
wAMD: wet age-related macular degeneration

XSL•FO
RenderX

Edited by S Brini; submitted 04.Apr.2026; peer-reviewed by MH Tsai, X Liang; revised version received 11.May.2026; accepted 12.May.2026; published 15.Jun.2026.

Please cite as:
Zhu Y, Niu Y, Sun S, Liu W, Dou Y, Guo Y
Performance of Deep Learning in Classifying Age-Related Macular Degeneration From Images: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e97174
URL: https://www.jmir.org/2026/1/e97174
doi:10.2196/97174

© Yu Zhu, Yue Niu, Shangye Sun, Wei Liu, Ying Dou, Yu Guo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effects of Multicomponent Digital Health Interventions on Multidimensional Physical Activity in Older Adults: Systematic Review, Meta-Analysis, and Meta-Regression of Randomized Controlled Trials

Jiayi Yao[1], MS; Haozhe Wang[1], MS; Shiguan Jia[1], MS; Hao Chen[2], MS; Junhao He[3], MS; Juncheng Long[1], MS; Shengxian Chen[4], PhD

[1]School of Physical Education, China University of Mining and Technology, No. 1 University Road, Xuzhou, Jiangsu, China

[2]School of Marxism, Jiangsu Normal University, Xuzhou, Jiangsu, China

[3]School of Physical Education and Training, Capital University of Physical Education and Sports, Beijing, China

[4]Department of Physical Education and Research, Beijing City University, No. 26 Yangzhen Section, Muyan Road, Yangzhen, Shunyi District, Beijing, China

Corresponding Author:
Shengxian Chen, PhD
Department of Physical Education and Research, Beijing City University, No. 26 Yangzhen Section, Muyan Road, Yangzhen, Shunyi District, Beijing, China

Abstract

Background: The comprehensive effects of multicomponent digital health interventions (DHIs) on multidimensional physical activity indicators and sedentary behavior (SB) remain controversial.

Objective: This systematic review aimed to evaluate the impact of multicomponent DHIs on daily steps, moderate-to-vigorous physical activity (MVPA), light physical activity, total physical activity, and SB in older adults.

Methods: PubMed, Web of Science, Embase, The Cochrane Library, and CINAHL were searched up to February 20, 2026. Randomized controlled trials concerning multicomponent DHIs for promoting exercise behavior in older adults were included. RoB 2.0 was used to evaluate study quality. Meta-analyses were performed using the Hartung-Knapp-Sidik-Jonkman random-effects model, and 95% prediction intervals (PIs) were calculated via Nagashima adjustment to evaluate effect dispersion. The GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) system was used to evaluate evidence certainty.

Results: A total of 26 randomized controlled trials (n=4129) were included. The results showed that multicomponent DHIs significantly improved daily steps (mean difference [MD] 822.8, 95% CI 198.3 to 1447.3 steps/d; 95% PI –1452.4 to 3098.0) and MVPA (MD 45.9, 95% CI 23.9 to 67.9 min/wk; 95% PI –9.4 to 101.2). However, the improvements in SB (MD –283.7, 95% CI –610.8 to 43.5 min/wk; 95% PI –984.5 to 417.1), total physical activity (MD 104.4; 95% CI –109.2 to 318.0 min/wk; 95% PI –444.4 to 653.2), and light physical activity (MD 39.3, 95% CI –96.2 to 174.7 min/wk; 95% PI –227.6 to 306.2) did not reach statistical significance. As some included studies combined digital tools with human support, the independent contribution of digital technology remains uncertain. PIs indicated a certain degree of dispersion across different clinical contexts. Subgroup analysis showed higher effect sizes for standalone wearables, human-assisted interventions, and populations with chronic disease risks. Meta-regression showed that effect sizes remained stable across different ages and durations. The trim-and-fill method confirmed the robustness of MVPA results. GRADE assessment indicated “moderate” certainty for MVPA and “low” for daily steps and other indicators.

Conclusions: This systematic review suggests that multicomponent DHIs may serve as an effective means for enhancing daily steps and MVPA in older adults. The innovation lies in evaluating the true effect distribution of multicomponent DHIs through Hartung-Knapp-Sidik-Jonkman random-effects models and Nagashima PIs. Compared with previous studies, this review identified the impact of population characteristics and control group differences on effect estimates using PI and subgroup models, confirming that advanced age did not significantly diminish the good adaptability of older adults to DHIs. Evidence limitations include high heterogeneity, lack of long-term follow-up, and differences between objective and subjective measurement tools. In practice, priority should be given to hardware carriers with simplified interaction and integrated human support, with tailored strategies developed for different risk subgroups.

Trial Registration: PROSPERO CRD420261323151; https://www.crd.york.ac.uk/PROSPERO/view/CRD420261323151

(J Med Internet Res 2026;28:e91338) doi:10.2196/91338

XSL•FO
RenderX

KEYWORDS

older adults; digital health interventions; physical activity; sedentary behavior; randomized controlled trials; meta-analysis; meta-regression

Introduction

With the accelerating process of global population aging, maintaining the physical function and independent living capacity of the older population has become a core element of public health strategies [1,2]. Physical activity (PA) is widely recognized as a key protective factor for preventing chronic noncommunicable diseases, maintaining muscle mass, and improving mental health [3,4]. However, there is a significant imbalance in participation across different intensities of PA among older adults. Most older individuals struggle to reach recommended levels of moderate-to-vigorous physical activity (MVPA), which directly limits the maintenance of cardiorespiratory fitness and metabolic health [5]. Meanwhile, sedentary behavior (SB) accounts for 65% to 80% of the waking hours of older adults; this pattern of prolonged physical inactivity is closely associated with an increased risk of all-cause mortality [6]. Although light physical activity (LPA) is considered an effective alternative for reducing SB and improving health outcomes in older adults, systematic attention toward LPA remains relatively scarce in existing intervention studies and guideline recommendations [7].

Multicomponent digital health interventions (DHI) provide important means for promoting improvements in exercise behavior among older adults by integrating behavior change techniques such as goal setting, self-monitoring, automated feedback, and social support [8,9]. Currently, intervention formats encompass various digital technology carriers, including mobile apps, standalone wearable electronic devices, and web-based interactive platforms, with applications expanding from general community-dwelling healthy older adults to specific groups with health risks [10,11]. Furthermore, as intervention programs often comprise multidimensional components, they exhibit significant heterogeneity in delivery agents [12]. One category is “technology-driven,” where automated algorithms (such as fully automated app programs) serve as the primary interactive entity to achieve real-time goal adjustment and feedback. The other category is “technology-assisted,” where digital technology primarily serves as an auxiliary monitoring tool, while the core behavior change support still relies on human involvement (such as health care professionals or coaches). Compared with traditional face-to-face intervention programs, multicomponent digital interventions demonstrate significant advantages in reducing resource consumption, increasing accessibility, and enhancing the convenience of long-term management [13]. Additionally, an increasing number of digital programs are being designed based on established theoretical frameworks, such as Social Cognitive Theory and the Health Action Process Approach, aiming to enhance exercise adherence in older adults through precise monitoring data and personalized interactive experiences [14].

Although several meta-analyses have confirmed the positive effects of multicomponent DHIs on increasing daily steps and MVPA, existing reviews still possess various limitations [15]. Currently, the composition of the participant populations included in most studies is extremely complex, often mixing generally healthy individuals, those at risk for chronic diseases, and various patients already in clinical treatment stages. This leads to ambiguity regarding the guiding significance of research conclusions for healthy community-dwelling older adults or specific at-risk populations [16]. Such population heterogeneity limits a deep understanding of intervention suitability, making it difficult to clarify the precise efficacy of multicomponent digital technology in populations transitioning from a healthy state to a diseased state. Simultaneously, existing reviews provide an imbalanced assessment of different PA intensity indicators. In particular, regarding the effects on reducing SB and increasing LPA, conclusions across different studies vary significantly and lack a systematic summary [17]. Due to the diversity of technical carriers and interaction methods, existing research rarely directly compares the advantages and disadvantages of standalone wearable devices vs mobile apps and web platforms in terms of intervention effects [18]. Furthermore, most studies primarily report the average pooled effect after intervention, failing to fully explain the sources of significant interstudy heterogeneity. Although participants’ age, sex ratio, and intervention duration are considered key moderating variables affecting intervention outcomes, in-depth meta-regression analyses targeting these specific covariates remain scarce in this field. Consequently, the internal laws governing how intervention efficacy evolves or varies by participant characteristics have not been fully revealed [19]. The application of tools to assess the certainty of pooled evidence remains insufficiently widespread in existing reviews, which also limits the applied value of research conclusions.

Therefore, this systematic review aims to comprehensively examine the multidimensional effects of multicomponent DHIs on daily steps, MVPA, SB, LPA, and total physical activity (TPA) in older adults through systematic evaluation and meta-analysis, thereby providing a complete picture of changes in exercise behavior among older adults. This systematic review not only focuses on the average pooled effect generated by interventions but also evaluates the distribution and generalizability of intervention effects across different research contexts through prediction intervals (PIs). Meanwhile, this systematic review will compare differences in intervention efficacy between standalone wearable devices and models involving mobile apps or web platforms. It will also use meta-regression analysis to quantitatively explore the moderating effects of age, intervention duration, and prompt frequency on effect sizes, thereby assessing the stability of multicomponent DHI effects across different older populations and implementation schemes. Finally, this systematic review will rigorously evaluate the certainty of evidence for each outcome indicator in conjunction with the GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) system, aiming to provide rigorous evidence-based

XSL•FO
RenderX

support for developing precise digital health management strategies for older adults.

Methods

Study Design

This systematic review protocol was registered with PROSPERO (CRD420261323151). After the research team conducted the initial presearch in December 2025, it was found that the originally designed inclusion or exclusion criteria and intervention definitions required refined adjustments based on preliminary evidence. To ensure high consistency between the registered protocol and the final implementation plan, the team updated and resubmitted the registration application in early 2026 after perfecting the protocol. Although it constitutes retrospective registration in terms of procedure due to the presearch, the final implementation plan, outcome measure definitions, and statistical analysis plan of this systematic review are entirely consistent with the design initially submitted to the platform, with no protocol deviations occurring. The implementation and reporting of this systematic review strictly follow the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement [20], and the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) guideline [21] was adopted for the reporting of the search strategy to ensure the transparency and reproducibility of the search process.

Search Strategy

The search process for this systematic review strictly follows the PRISMA-S extension statement. A systematic and comprehensive literature search was conducted on February 20, 2026, across the PubMed (via NCBI), Web of Science Core Collection (via Clarivate), The Cochrane Library (via Wiley), Embase (via Elsevier), and CINAHL (via EBSCOhost) databases, with the search timeframe spanning from the inception of each database to the date of the search. The search scheme was constructed based on 4 core dimensions: population, intervention, outcome, and study design, using Boolean logic operations combining subject headings (eg, MeSH [Medical Subject Headings], Emtree, and CINAHL Subject Headings) and free-text terms. To ensure search depth and precision, differentiated adaptations were made according to the search characteristics of each database. For example, expanded subject heading searches were combined with TIAB (title and abstract) field limitations in PubMed and Embase; the TS (topic) field was used for exact term matching in Web of Science; the NEXT proximity operator was applied in The Cochrane Library; and the XB (exact bound) qualifier was introduced in CINAHL to enhance the accuracy of subject heading matching. Specific terms encompassed “older adults” (“aged,” “older adults,” “elderly,” and “aging”), “digital health interventions” (“telemedicine,” “mobile app,” “wearable devices,” “mHealth,” and “eHealth”), and “physical activity indicators” (“PA,” “exercise,” “daily steps,” “MVPA,” and “SB”), with randomized controlled trial (RCT) filters applied for precise screening. Additionally, this systematic review identified extra records by manually tracking the reference lists of included studies (backward citation searching) and attempted to contact original authors via email to obtain unpublished raw data or missing information. No additional searches were performed for manually browsed online resources, conference proceedings, or specific manufacturers; meanwhile, the search strategy was developed and proofread internally by the research team without undergoing an external peer review process. Given that this systematic review design was limited to RCTs, this systematic review did not conduct specific searches of gray literature or preprint databases. Complete search steps and Boolean logic formulas for each database are detailed in Multimedia Appendix 1.

Inclusion and Exclusion Criteria

The literature screening process strictly followed the PICOS principles: participants (P) included older adults with a mean age of ≥60 years, encompassing community-dwelling healthy individuals as well as specific populations with frailty, fall risks, or chronic disease risks; interventions (I) were defined as programs aimed at promoting PA that used multicomponent digital health technologies (eg, mobile apps, interactive web platforms, and smart wearable devices) as the core interactive carrier, with interventions categorized based on whether the digital technology served as the primary automated interactive entity or as an auxiliary tool for human-supported interventions; comparisons (C) were set as groups receiving usual care, basic health monitoring, pure health education, or those on a waitlist; outcomes (O) were required to include at least one of the following: daily steps, weekly MVPA, SB, LPA, or TPA volume; and study type (S) was limited to peer-reviewed, published RCTs, including randomized pilot trials and feasibility studies with clear randomization designs. Exclusion criteria were defined as (1) studies with nonrandomized controlled designs, case reports, or reviews; (2) research consisting only of study protocols or conference abstracts that lacked complete empirical results; (3) literature where original valid data could not be obtained or converted, and attempts to contact the authors were unsuccessful; and (4) intervention studies where digital technology served only as an auxiliary recording and measurement tool and did not involve any behavior change feedback or interactive functions.

Data Collection Process and Data Items

Data extraction was performed independently by two researchers (JY and HW), referencing the Cochrane Handbook guidelines to record information. The extracted content encompassed basic study information, participant characteristics (age, sex, and health status), outcome indicators (daily steps, MVPA, LPA, TPA, and SB), and key details of the multicomponent digital interventions, such as technical carriers, theoretical frameworks, core behavior change techniques, and the level of human interaction. Any discrepancies during the extraction process were resolved through discussion or by consulting a third independent researcher (SC).

Choice of Effect Measures and Handling of Missing Data

During the quantitative data extraction process, this systematic review followed a specific hierarchy of priorities: preference was given to effect estimates adjusted by models such as

XSL•FO
RenderX

analysis of covariance; if unavailable, change scores from baseline to post intervention were calculated; if neither of the former was obtainable, postintervention means and SDs were used. For missing SDs during the calculation of change scores, this systematic review referred to methods recommended in the Cochrane Handbook, setting the correlation coefficient at 0.5 for estimation [22]; in cases of missing data or unclear reporting, the original authors were contacted via email to obtain supplementary information.

Assessment of Risk of Bias

The quality of the included studies was evaluated using the Cochrane RoB 2.0 tool [23]. The assessment covered bias arising from the randomization process, bias due to deviations from intended interventions, bias due to missing outcome data, bias in the measurement of the outcome, and bias in the selection of the reported results. Each domain, as well as the overall risk, was rated as low risk, some concerns, or high risk. SJ and HC performed the evaluations independently, resolving disagreements through discussion.

Evidence Certainty Assessment

This systematic review used the GRADE system to assess the certainty of evidence for pooled outcome indicators, comprehensively considering the risk of bias, inconsistency, indirectness, imprecision, and publication bias. In evaluating inconsistency, the initial heterogeneity encompassing all relevant studies served as the basis for downgrading decisions. The quality of evidence was ultimately classified into four levels: high, moderate, low, or very low [24].

Data Analysis

All statistical analyses for this systematic review were performed using R (R Foundation) software via the *metafor* and *forestplot* packages. Although individual studies varied in measurement tools and intervention protocols, given that daily steps, MVPA, LPA, TPA, and SB all possess unified physical units and clear clinical interpretative value, this systematic review used the mean difference (MD) and its 95% CI values for effect size pooling. To ensure the rigor of MD pooling, all data were converted to standard dimensions before analysis. Furthermore, this systematic review conducted sensitivity analyses using the standardized mean difference (SMD) to evaluate the impact of measurement tool diversity on the robustness of the pooled results. Considering the potential clinical and methodological heterogeneity among studies, this systematic review consistently adopted a random-effects model for effect size pooling. For multiarm studies involving multiple intervention groups compared against a single control group, to avoid overstating the evidence due to double-counting of samples, this systematic review followed Cochrane Handbook recommendations by equally dividing the control group sample size among the comparison groups while keeping the means and SDs constant. Additionally, to reduce the false positive rate in the random-effects model and provide more robust CIs, the Hartung-Knapp-Sidik-Jonkman (HKSJ) method was used to adjust all pooled effect sizes and *P* value calculations [25].

Statistical heterogeneity was quantitatively evaluated using the Q-test and the I^2 index. Notably, given that I^2 only reflects the proportion of observed variation caused by true differences in effect, to further quantify the clinical significance of heterogeneity in real-world settings, 95% PIs were calculated when the number of included studies was ≥3. Unlike the 95% CI, which reflects the precision of the average effect, the 95% PI quantifies the range of variation of the true effect across different populations or settings [26]. For significant outcome indicators, sensitivity analyses were conducted using the "leave-one-out" method to assess the robustness of pooled results and identify sources of heterogeneity. This systematic review preset subgroup classification criteria across five dimensions to address sources of heterogeneity. First, intervention duration was divided into short-term and medium-to-long-term using a 12-week threshold; second, population types were categorized into those with specific chronic disease risks and generally healthy older adults; third, control group settings were divided into active controls, providing routine exercise guidance or time-matched education, and passive controls, maintaining the status quo or joining a waitlist; fourth, technical carriers were categorized into standalone wearable devices and mobile apps; finally, delivery agent models were classified based on the nature of human-computer interaction into technology-driven and technology-mediated types, with the former providing automated feedback via algorithms or sensors without real-time human intervention, and the latter using digital platforms with remote guidance or personalized feedback provided by professionals [27]. Simultaneously, random-effects meta-regression analysis was performed on primary outcome indicators using age, intervention duration, female proportion, and prompt frequency as covariates [28]. This systematic review assessed the presence of small-study effects by observing funnel plot symmetry combined with Egger linear regression test (when the number of included studies k was >10) [29]. It should be noted that funnel plot asymmetry may stem from various factors such as publication bias, differences in study quality, or heterogeneity. If significant asymmetry was detected, this systematic review used the trim-and-fill method to adjust the pooled effect size, further evaluating the impact of potential bias on the robustness of the conclusions.

Results

Literature Search and Study Selection

The search process for this systematic review strictly followed the PRISMA-S extension criteria (Figure 1). Through systematic searches of the PubMed (via NCBI), Web of Science Core Collection (via Clarivate), The Cochrane Library (via Wiley), Embase (via Elsevier), and CINAHL (via EBSCOhost) databases, a total of 14,180 original records were obtained. After using the automated tool EndNote (Clarivate) to remove 6548 duplicate records, a preliminary screening of titles and abstracts was conducted for the remaining 7632 records, resulting in the exclusion of 7485 irrelevant documents. The primary reasons for excluding these records included a clear deviation in the research topic, nonoriginal research types, or a distinct mismatch with the target population. During the full-text retrieval of the remaining 147 reports, 5 reports could not be obtained in full. Subsequently, researchers performed a rigorous eligibility

XSL•FO
RenderX

assessment on 142 full-text papers, excluding a total of 116 documents for specific reasons: participants in acute rehabilitation phases or noncommunity or rural-dwelling groups (n=31); lack of digital carriers or non-PA interventions (n=35); missing outcome measures or unusable data (n=18); non-RCTs or study protocols only (n=24); and duplicate publications or studies with overlapping data (n=8). Following this screening process, 26 studies were ultimately included in the systematic review, of which 24 studies were included in the meta-analysis due to having sufficient quantitative data.

Figure 1. PRISMA 2020 flowchart of literature search and selection for multicomponent digital health interventions to improve PA levels in older adults. PA: physical activity; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; RCT: randomized controlled trial.

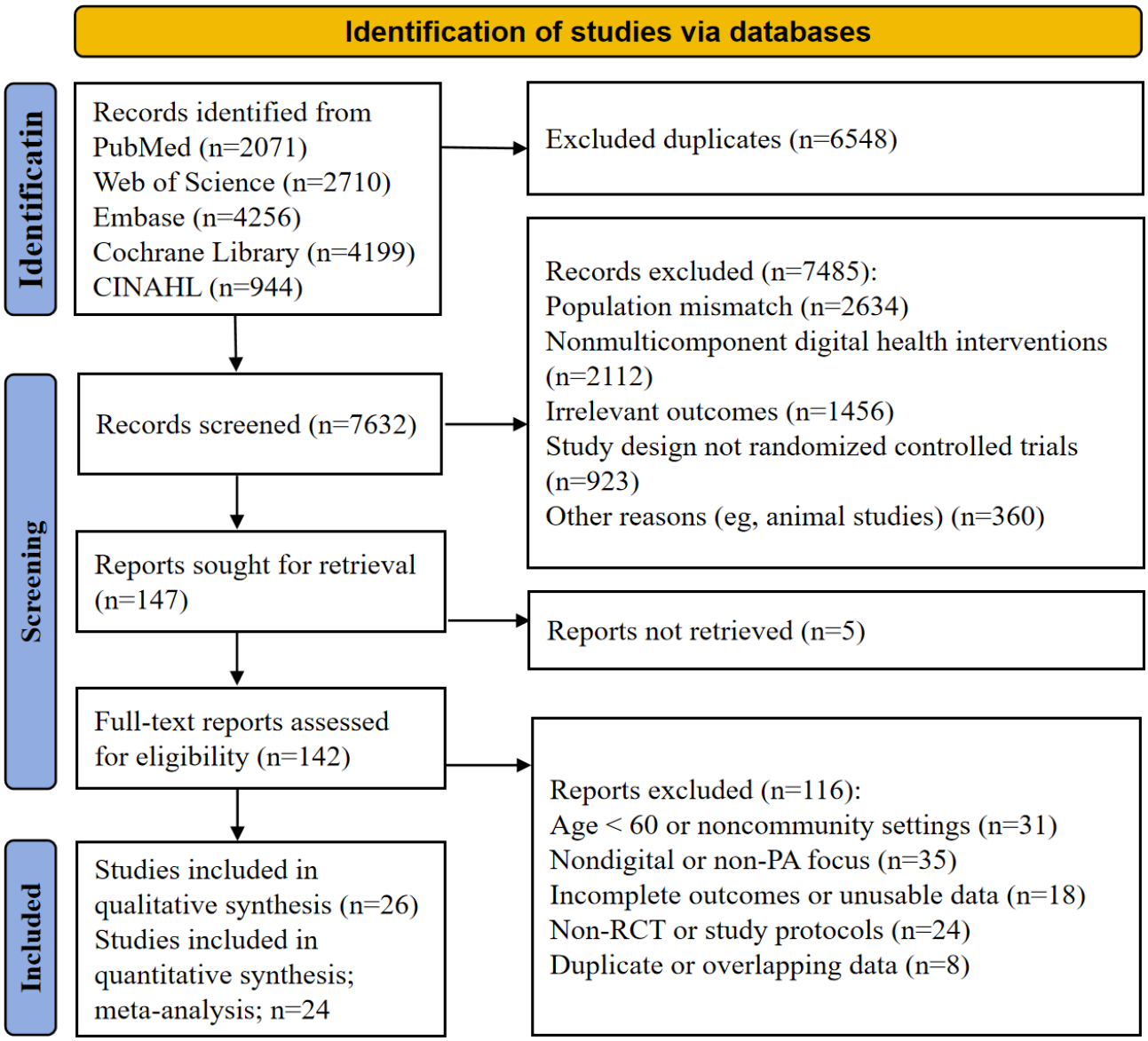

Description of Included Study Characteristics

The 26 RCTs ultimately included in this systematic review were published between 2009 and 2025 [30-55], encompassing various design types such as cluster RCTs, crossover RCTs, and randomized pilot studies. These RCTs involved a total of 4129 participants, with research areas spanning Australia [35,36,43,45,49], the United States [37,40,42,44], China [33,38,41,46,48,51], the United Kingdom [30,54], Japan [47,52], and other countries [31,32,34,39,50,53,55]. Intervention durations varied significantly, ranging from 4 weeks [38] to 12 months of continuous intervention or follow-up [39,43,50,54], with short-to-medium-term interventions of 12 and 24 weeks being the most common. The mean age of participants was over 60 years, ranging from 60.0 [37] to 77.4 [35] years. Sample sizes per study ranged from 13 [32] to 605 [53], with a notable female predominance. Beyond healthy community-dwelling older adults, this systematic review populations widely involved groups at high risk for cardiovascular disease [44], impaired glucose tolerance [54], frailty or prefrailty [33,38,46], cognitive decline risk [51], fall risk [30], and overweight or obese sedentary high-risk groups [37,41,42].

In terms of intervention measures and technology implementation, digital health carriers showed a diverse trend, primarily including mobile apps [31,33,38,41,48,51,52], wearable electronic devices [32,36,37,42-46,49,53], and web-based interactive platforms [34,35,39,40,47,48,53]. Most intervention protocols were based on established behavior change theories, such as Social Cognitive Theory

XSL•FO
RenderX

[32,42-45,47,48], the Health Action Process Approach [34,48], the Theory of Planned Behavior [45], and the Behavior Change Wheel [33]. Core intervention components typically included goal setting, self-monitoring, automated feedback, and social support. Regarding control group settings and outcome evaluation, control groups were mostly set to receive standard health education [31-33,35,48,49,55], usual care [30,36,37,41,46,54], email-based information [47], or waitlist status [38,40,45,53]. Outcomes were primarily measured using objective tools, such as ActiGraph GT9X accelerometers [45], ActivPAL sensors (PAL Technologies Ltd) [42,46], or smart pedometers [30,48-50,54]. Regarding data sources for effect size calculation among the 24 studies [32-55] included in the quantitative synthesis, 16 (66.7%) studies [32-38,40,42-47,49,53] used change scores from baseline; 6 (25%) studies [41,50-52,54,55] converted medians and IQRs into means and SDs; and the remaining 2 (8.3%) studies [39,48] used unadjusted postintervention data. Two studies, Audsley et al (2020) [30] and Saçıkara and Cingil (2024) [31], were excluded from the quantitative synthesis because they only reported the magnitude of change or qualitative trends in PA without providing the raw mean and SD values required for effect size calculation (details are shown in Table 1).

XSL•FO
RenderX

Table . Characteristics of the 26 included RCTs[a] on multicomponent digital health interventions for improving physical activity levels in older adults.

Study (author/year)	Int/FU[b] duration	Sample size (IG/CG[c] initial/completed)	Age (years), mean (SD)	Sex (% female)	Health status/population	Digital platform and theoretical framework	Core BCTs[d]	Control setting	Outcomes and tools
Yates et al, 2009 [54]	12 months / 12 months	33+31/34 (initial) 29+28/26 (complete)	65 (8)	34	IGT[e]	Pedometer; diabetes prevention or self-efficacy theory	Education, personalized goals or steps, diary, action planning	Usual care (brief)	Steps; obj[f]: NL-800 pedometer
Bickmore et al, 2013 [50]	12 months / 12 months	263 (initial)	71.3	61	Sedentary urban older adults	Tablet+virtual agent; ECA[g] theory	Daily automated dialogue, daily goal setting, automated coaching	Pedometer group	Steps; pedometer (obj)
Ashe et al, 2015 [32]	6 months / 6 months	13/13 (initial) 12/8 (complete)	64.1 (4.5)	100	Sedentary older women	Fitbit One (Fitbit Inc); SCT[h]/social ecological model	Goal setting, self-monitoring, feedback, social support, planning	Health education	MVPA[i]; ActiGraph (obj; ActiGraph LLC)
Cadmus-Bertram et al, 2015 [37]	16 weeks / N/A	25/26 (initial) 24/25 (complete)	60.0 (5.2)	100	Overweight/obese sedentary older women	Fitbit One+web; CALO-RE[j] framework	Self-monitoring, goal setting, behavioral feedback, phone counseling	Usual care	MVPA; ActiGraph (obj)
Mouton and Cloes, 2015 [39]	3 months / 12 months	52+3/50 (initial) 149 completed	65.0 (6.0)	64	Insufficiently active older adults	Web platform+email; TTM[k]/ecological model	SMART[l] goals, PA[m] diary, feedback, forum social support	No intervention	Total PA; IPAQ-S[n] (Sub[o])
Muellmann et al, 2019 [53]	10 weeks / 10 weeks	104+107/107 (initial) 85+93/90 (complete)	68.6 (5.5)	64	Insufficiently active older adults	Web platform+Fitbit (Fitbit Inc); self-regulation theory	Goal setting, action or coping planning, feedback, self-monitoring	Waitlist control	Steps/MVPA/SB[p]; ActiGraph (obj)
Oliveira et al, 2019 [49]	6 months / 6 months	64/67 (initial) 54/57 (complete)	71.1 (7.1)	65	General older adults	Fitbit/pedometer; BCTs	GAS[q], telephone coaching, feedback	Standard health education	MVPA/steps/LPA[r]/total PA; ActiGraph (obj)
Roberts et al, 2019 [44]	20 weeks / 20 weeks	20/20 (initial) 18/19 (complete)	72.0 (7.4)	60	CVD[s] risk /high sedentary behavior	Fitbit Zip (Fitbit Inc); cognitive behavioral theory	Monitoring feedback, weekly SMS, goal setting	Structured exercise counseling	Steps/SB; ActiGraph (obj)
Van Dyck et al, 2019 [34]	5 weeks / 3 months	38/34 (initial) 34/31 (complete)	70.9 (4.8)	49	General older adults	MyPlan 2.0 Web (Ghent University); self-regulation or HAPA theory	Goal setting, action planning, coping planning, feedback	No intervention	Total PA/MVPA; IPAQ[t] (sub)
Audsley et al, 2020 [30]	6 months / 6 months	20/30 (initial) 20/25 (complete)	IG: 76.9 (7.0)/CG: 73.8 (6.4)	80	Older adults at risk of falling	Pedometer+telephone; MI[ac]	MI counseling, PA diary, goal setting, action planning	Usual care	MVPA; Phone-FITT (sub; McGill University)

XSL•FO
RenderX

Study (author/year)	Int/FU[b] duration	Sample size (IG/CG[c] initial/completed)	Age (years), mean (SD)	Sex (% female)	Health status/population	Digital platform and theoretical framework	Core BCTs[d]	Control setting	Outcomes and tools
Kwan et al, 2020 [51]	12 weeks / 13 weeks	16/17 (initial) 15/15 (complete)	71.0 (9.0)[v]	85	Risk of cognitive decline	App+WhatsApp (WhatsApp LLC); persuasive technology	Personalized goals, monitoring, electronic reminders, video feedback	Routine behavior intervention	Steps/MVPA/total PA; ActiGraph (obj)
Rosenberg et al, 2020 [42]	12 weeks / 12 weeks	29/31 (initial) 29/25 (complete)	68.4 (4.9)	68	Overweight/obese sedentary older adults	Jawbone handband (Jawbone Inc); SCT/habit formation	Telephone coaching, goal setting, environment reminders, self-monitoring	Healthy life handbook	Steps/total PA; activPAL (obj; PAL Technologies Ltd)
Brickwood et al, 2021 [36]	12 weeks / 12 months	37+38/42 (initial) 24+25/26 (complete)	72.4 (6.1)	64	Risk of chronic diseases	Jawbone UP24 app (Jawbone Inc); SCT theory	Goal setting, app or SMS feedback, self-monitoring	Usual care	Steps; activPAL (obj)
Delbaere et al, 2021 [35]	24 months / 24 months	254/249 (initial) 201/212 (complete)	77.4 (5.4)	67	General older adults	Standing Tall app (NeuRA[w]); user-centered design/BCTs	Customized exercises, goal setting, feedback, educational manual	Electronic health education	LPA; MoveMonitor (obj; McRoberts)
Lin et al, 2025 [33]	14 weeks / 3 months	22/18 (initial) 22/18 (complete)	IG: 72.1 (3.7), CG: 80.4 (6.8)	85	Frail/prefrail older adults	Fitbit+app; BCTs	Goal setting, incentive badges, monitoring, action cues	Training+ lectures	Steps/MVPA; ActiGraph (obj)
Zhou et al, 2021 [41]	12 weeks / N/A	40/41 (initial) 34/34 (complete)	67.4 (5.1)	66	Overweight/obese older adults	App+smartband; BCTs	Goal setting, diet or PA monitoring, feedback, WeChat social support	Pedometer hand manual	Steps/MVPA; smartband (obj)
Alley et al, 2022 [45]	12 weeks / 24 weeks	78+96/69 (initial) 56+59/51 (complete)	69.34 (4.32)	78.6	Insufficiently active older adults	Web-based+Fitbit; TPB[x]/SCT	6 tailored suggestion modules, action planning, exercise video library	Waitlist control	MVPA/SB/Total PA/Steps; ActiGraph
Cai et al, 2022 [48]	3 months / N/A	36/36 (initial) 34/30 (complete)	66.9 (4.2)	64	Rural older adults	WeChat (Tencent Holdings Ltd)+pedometer; SCT/HAPA[y] model	Peer support, planning, microexperience sharing	Standard health education	Steps; pedometer (obj)
Pischke et al, 2022 [40]	9 months / 9 months	242 (initial) 160 (complete)	68.7 (5.4)	62	Insufficiently active older adults	Web/app+Fitbit; self-regulation theory	Monitoring, goals, feedback, small group coaching, social support	Waitlist control	MVPA/SB; ActiGraph (obj)

XSL•FO
RenderX

Study (author/year)	Int/FU[b] duration	Sample size (IG/CG[c] initial/completed)	Age (years), mean (SD)	Sex (% female)	Health status/population	Digital platform and theoretical framework	Core BCTs[d]	Control setting	Outcomes and tools
Recio-Rodríguez et al, 2022 [55]	3 months / 3 months	81/79 (initial)	70.8 (4.0)	61.3	General older adults	App+smartband; behavior change theory	Lifestyle counseling, self-monitoring, dietary logs, automated feedback	Short behavior intervention	Steps; accelerometer/smartband
Kawaguchi et al, 2024 [52]	12 weeks / 12 weeks	87/94 (initial) 80/85 (complete)	IG: 69.9 (6.0), CG: 70.1 (6.7)	53	General older adults	ESP app (LINE Corporation)+LINE (LINE Corporation); BCTs	GPS trajectory monitoring, rank-based rewards, educational columns	Basic monitoring (Google Fit [Google LLC])	Steps; Google Fit (obj)
Oliveira et al, 2024 [43]	12 months / 12 months	290/315 (initial) 258/252 (complete)	74.0 (8.0)	70	General older adults	Fitbit+telephone; SDT theory	Risk assessment, goal setting, coaching, problem solving	Nutritional coaching	Steps; ActiGraph (obj)
Saçıkara and Cingil, 2024 [31]	8 weeks / 6 months	33/33 (initial) 33/27 (complete)	66.3 (4.2)	50	General older adults	Google Fit app; health promotion model	Health education, self-monitoring, small-group social interaction, rewards	No intervention	Total PA; scale (sub)
Uemura et al, 2024 [47]	12 weeks / 36 weeks	15/14 (initial) 14/12 (complete)	IG: 73.9 (3.9)/CG: 69.4 (3.2)	50	General older adults	Zoom (Zoom Video Communications, Inc) meeting; self-directed/SCT	Exploratory tasks, small group discussion, goals, monitoring	Mail information push	Total PA/MVPA/LPA; ActiGraph
Lee et al, 2025 [38]	4 weeks / 3 months	19/19 (initial) 16/18 (complete)	71.8 (9.34)	71	Frail/prefrail older adults	Rehab-Fit app (Fujitsu Limited); Self-efficacy theory	Educational workshops, app guidance, video/audio feedback	Waitlist control	MVPA; ActiGraph (obj)

XSL•FO
RenderX

Study (author/year)	Int/FU[b] duration	Sample size (IG/CG[c] initial/completed)	Age (years), mean (SD)	Sex (% female)	Health status/population	Digital platform and theoretical framework	Core BCTs[d]	Control setting	Outcomes and tools
Li et al, 2025 [33]	6 months / 6 months	67/67 (initial) 52/54 (complete)	69.3 (5.1)	75	Prefrail older adults	mHealth app; LIFE[z] model/BCW[aa] theory	Personalized plans, video guidance, remote monitoring, feedback	Health education	MV-PA/SB/LPA; ActiGraph (obj)

[a]RCT: randomized controlled trial.
[b]Int/FU: intervention/follow-up.
[c]IG/CG: intervention group/control group.
[d]BCT: behavior change technique.
[e]IGT: impaired glucose tolerance.
[f]Obj: objective measurement.
[g]ECA: embodied conversational agent.
[h]SCT: Social Cognitive Theory.
[i]MVPA: moderate-to-vigorous physical activity.
[j]CALO-RE: Coventry, Aberdeen, and London—Refined.
[k]TTM: transtheoretical model.
[l]SMART: Sequential Multiple Assignment Randomized Trial.
[m]PA: physical activity.
[n]IPAQ-S: International Physical Activity Questionnaire-Short Form.
[o]Sub: subjective or self-reported measurement frameworks.
[p]SB: sedentary behavior.
[q]GAS: goal attainment scaling.
[r]LPA: light physical activity.
[s]CVD: cardiovascular disease.
[t]IPAQ: International Physical Activity Questionnaire.
[u]MI: motivational interviewing.
[v]Median (IQR).
[w]NeuRA: Standing Tall App (Neuroscience Research Australia.
[x]TPB: Theory of Planned Behavior.
[y]HAPA: health action process approach.
[z]LIFE: Lifestyle Interventions and Independence for Elders.
[aa]BCW: behavior change wheel.

Risk-of-Bias Assessment

This systematic review performed a systematic methodological quality assessment of the 26 included studies [30-55], revealing that most research performed well across key evaluation domains (Figure 2 [30-55] and Figure 3). Evaluation across five dimensions, including the randomization process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported results, found that most studies were overall rated as low risk, with single uncertainties appearing only in the domain of deviations from intended interventions. Specifically, studies such as Alley et al (2022) [45], Brickwood et al (2021) [36], Cai et al (2022) [48], Li et al (2025) [33], Oliveira et al (2024) [43], and Yates et al (2009) [54] provided detailed descriptions of random sequence generation and allocation concealment, and used objective tools such as accelerometers or pedometers to avoid measurement bias potentially arising from nonblinded designs. Meanwhile, Delbaere et al (2021) [35], Kawaguchi et al (2024) [52], Lee et al (2025) [38], Recio-Rodríguez et al (2022) [55], Uemura et al (2024) [47], and Mouton et al (2015) [39] further ensured the robustness of the overall evidence quality through comprehensive protocol registration and transparent results reporting, reflecting the rigor of current DHI research designs.

XSL•FO
RenderX

Figure 2. Risk of bias summary for the 26 included RCTs evaluating multicomponent digital health interventions for older adults. RCT: randomized controlled trial [30-55].

XSL•FO
RenderX

Figure 3. Risk-of-bias graph presented as cumulative percentages across the 26 included RCTs on multicomponent digital health interventions for older adults. RCT: randomized controlled trial.

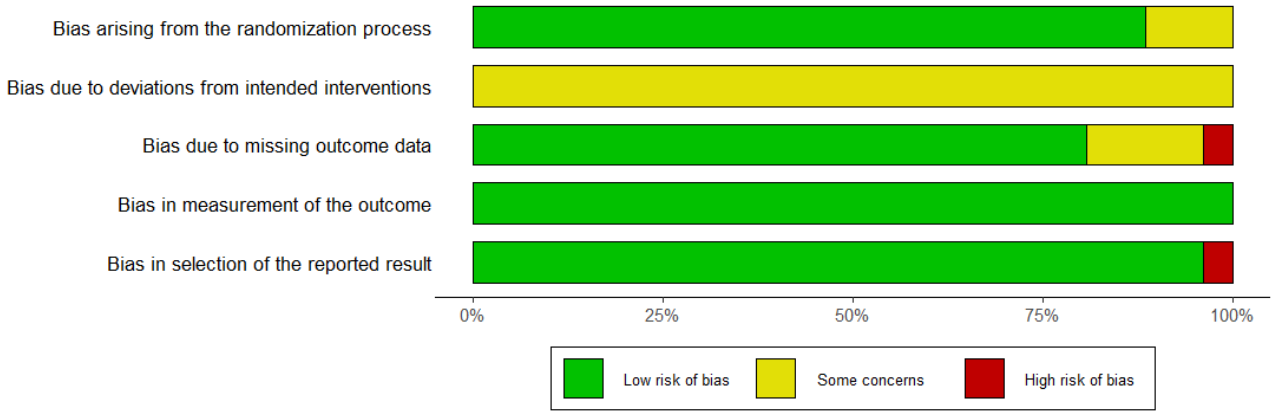

Some studies were rated as having some concerns or high risk due to simultaneous methodological limitations across multiple dimensions. Roberts et al (2019) [44], Muellmann et al (2019) [53], and Bickmore et al (2013) [50] were marked as having some concerns in both randomization and intervention deviation due to unclear allocation concealment details and the inability to blind the intervention process, resulting in a downgrade of the overall rating. Additionally, Ashe et al (2015) [32], Kwan et al (2020) [51], Pischke et al (2022) [40], and Van Dyck et al (2019) [34] accumulated multiple bias risks in data integrity due to high attrition rates or lack of clear intention-to-treat analysis for missing data, and were overall judged as having some concerns. Regarding studies judged as high risk, Zhou et al (2021) [41] showed significant defects in the missing outcome data domain, with a high proportion of participant dropout in the later stages without reasonable explanation, reducing evidence reliability. Audsley et al (2020) [30] demonstrated a high risk of bias in the selection of the reported results, as the final reported indicators deviated significantly from the prespecified trial registration protocol; this tendency for selective reporting directly pushed this systematic review's overall risk rating to high, suggesting caution when interpreting its pooled effects.

Meta-Analysis Pooled Results

Meta-Analysis of the Effects of Multicomponent DHIs on Daily Steps in Older Adults

The meta-analysis results showed that among the 14 included studies (n=1651), multicomponent DHIs significantly improved daily steps in older adults [32,36,41-44,46,48-52,54,55]. The weighted pooled mean estimate for daily steps merged using the HKSJ random-effects model was 822.8 (95% CI 198.3-1447.3) steps/d, with a statistical test result of t_{13}=2.85 (P=.01). Moderate-to-high statistical heterogeneity existed between studies (I^2=67.5%; P<.001). The 95% PI calculated based on the Nagashima method was −1452.4 to 3098.0 steps/d. This PI crossed the line of no effect, quantifying the potential distribution range of the true effect size in future similar clinical settings and suggesting a certain degree of volatility in intervention effects across different research contexts. To evaluate the robustness of the results, this systematic review conducted multidimensional sensitivity analyses (Table S1 in Multimedia Appendix 2). The "leave-one-out" analysis showed that the pooled MD values after excluding studies one by one fluctuated between 677.84 and 938.34 steps/d, and the 95% CI of all iteration results did not cross 0. The analysis identified that Zhou et al (2021) [41] and Yates et al (2009) [54] contributed significantly to the model heterogeneity; excluding Zhou et al (2021) [41] reduced I^2 to 60.1%. Furthermore, sensitivity analysis based on the SMD (Hedges g) further confirmed the robustness of the results. The pooled effect size merged using the HKSJ random-effects model was 0.4 (95% CI 0.1 to 0.6; t_{13}=3.44; P=.004), and its heterogeneity level (I^2=68.7%) was highly consistent with the primary analysis results. The calculated 95% PI was −0.4 to 1.1. The MD and SMD results exhibited consistency in effect direction and significance (Figure S1 in Multimedia Appendix 2).

Meta-Analysis of the Effects of Multicomponent DHIs on MVPA in Older Adults

The meta-analysis results indicated that among the 13 included studies (n=1523), multicomponent DHIs increased MVPA duration in older adults [32-34,37,38,40,41,45-47,49,51,53]. The weighted pooled mean estimate merged using the HKSJ random-effects model was 45.9 (95% CI 23.9-67.9) min/wk, with a statistical test result of t_{12}=4.1 (P<.001). The statistical heterogeneity test between studies showed I^2=18.6% (P=.26). The 95% PI calculated based on the Nagashima correction was −9.4 to 101.2 min/wk, quantifying the distribution range of true effect sizes in similar clinical settings. Sensitivity analysis addressing the diversity of measurement tools in the original studies showed that the pooled effect size calculated based on the SMD was 0.3 (95% CI 0.2-0.5; t_{12}=4.2; P<.001), with an interstudy heterogeneity level of I^2=18.2%. The calculated 95% PI was −0.1 to 0.7 (Figure S2 in Multimedia Appendix 2). The results showed that the summary statistics expressed as MD and SMD demonstrated consistency in effect direction and significance levels.

Meta-Analysis of the Effects of Multicomponent DHIs on Weekly Sedentary Time in Older Adults

The meta-analysis results showed that among the 5 included studies (n=741), the effect of multicomponent DHIs on weekly SB in older adults did not reach statistical significance

XSL•FO
RenderX

[33,40,44,45,53]. The weighted pooled mean estimate merged using the HKSJ random-effects model was −283.7 (95% CI −610.8 to 43.5) min/wk, with a statistical test result of t_{13}=−2.41 (*P*=.07). Significant statistical heterogeneity existed between studies (I^2=76.2%; *P*=.002). The calculated 95% PI was −984.5 to 417.1 min/wk. This indicator quantified the distribution range of true effect sizes in future similar intervention scenarios, suggesting that intervention effects fluctuate considerably across different individuals and research contexts.

Meta-Analysis of the Effects of Multicomponent DHIs on Weekly TPA Duration in Older Adults

The meta-analysis results showed that among the 7 included studies (n=886) [32,34,39,42,45,47,49], the effect of multicomponent DHIs on weekly TPA duration in older adults did not reach statistical significance [34,39,42,45,47,49,51]. The weighted pooled mean estimate merged using the HKSJ random-effects model was 104.4 (95% CI −109.2 to 318.0) min/wk, with a statistical test result of t_6=1.20 (*P*=.28). Moderate statistical heterogeneity existed between studies (I^2=55.3%; *P*=.04). The calculated 95% PI was −444.4 to 653.2 min/wk. This indicator quantified the distribution range of true effect sizes in similar clinical settings.

Meta-Analysis of the Effects of Multicomponent DHIs on Weekly LPA Duration in Older Adults

The meta-analysis results indicated that among the 4 included studies (n=1153) [33,35,47,49], the effect of multicomponent DHIs on weekly LPA duration in older adults did not reach statistical significance [33,47,49,56]. The weighted pooled mean estimate merged using the HKSJ random-effects model was 39.3 (95% CI −96.2 to 174.7) min/wk, with a statistical test result of t_3=0.92 (*P*=.42). Moderate-to-high statistical heterogeneity existed between studies (I^2=60.3%; *P*=.06). The calculated 95% PI was −227.6 to 306.2 min/wk. This indicator quantified the distribution range of true effect sizes in similar clinical settings.

Subgroup Analysis

Subgroup Analysis of the Effects of Multicomponent DHIs on Daily Steps in Older Adults

The subgroup analysis results for daily steps are summarized in Table 2. Regarding intervention duration, the pooled MD for short-term interventions (≤12 wk, k=6) and medium-to-long-term interventions (>12 wk, k=8) were 789.6 (95% CI −941.2 to 2520.5; 95% PI −3414.0 to 4993.3) and 838.6 (95% CI 208.6-1468.6; 95% PI −1020.9 to 2698.1), respectively. The test for subgroup differences showed *P*=.95, suggesting that intervention duration is not a key variable moderating step-count outcomes (Figure S3 in Multimedia Appendix 2). Regarding population characteristics, the difference between groups was statistically significant (*P*=.049), with the intervention effect in the high-risk or specific population group (MD=1256.1, 95% CI 258.6-2253.6) being significantly higher than that in the general healthy or community population group (MD=260.5, 95% CI −459.6 to 980.5). This reflects that DHIs may be more effective in motivating populations with specific health needs (Figure S4 in Multimedia Appendix 2). Regarding control group type, the pooled MD for the active control group (k=10) and the routine care group (k=4) were 599.0 (95% CI −101.0 to 1299.1) and 1478.1 (95% CI −674.6 to 3630.9), respectively. Although the routine care group numerically showed a higher increase in steps, the test for subgroup differences showed *P*=.24, and heterogeneity levels were highly similar between the 2 groups (I^2 of 65.6% and 66.1%, respectively), indicating that the control setting has limited moderating effect on step-count intervention outcomes (Figure S5 in Multimedia Appendix 2).

XSL•FO
RenderX

Table . Summary table evaluating the effects of multicomponent digital health interventions on daily steps (steps/d) in older adults across different subgroup characteristics based on the HKSJ[a] random-effects model.

Subgroup classification	k	Pooled MD[b] (95% CI)	95% PI[c]	I^2 (%)	$P_{subgroup}$
Duration					.95
Short-term (≤12 wk)	6	789.6 (−941.2 to 2520.5)	−3414.0 to 4993.3	77.3	
Medium-to-long-term (>12 wk)	8	838.6 (208.6 to 1468.6)	−1020.9 to 2698.1	60.8	
Health status					.049[d]
General healthy/community-dwelling	6	260.5 (−459.6 to 980.5)	−1376.5 to 1897.4	44.3	
High-risk/specific populations	8	1256.1 (258.6 to 2253.6)	−1587.4 to 4099.6	76	
Control group setting					.24
Active control/health education	10	599 (−101.0 to 1299.1)	−1731.0 to 2929.1	65.6	
Routine care/basic monitoring	4	1478.1 (−674.6 to 3630.9)	−3819.5 to 6775.8	66.1	
Technical carrier					.2
Mobile app/interactive platform	9	618.4 (−305.4 to 1542.2)	−2340.5 to 3577.3	67.9	
Standalone wearable device	5	1307.3 (299.1 to 2315.4)	−750.8 to 3365.3	31.4	
Delivery agent					.21
Technology-driven	7	459.9 (−381.1 to 1300.9)	−2040.6 to 2960.3	58.4	
Technology-mediated	7	1185.5 (43.4 to 2327.6)	−2150.3 to 4521.3	72.1	

[a]HKSJ: Hartung-Knapp-Sidik-Jonkman.

[b]MD: mean difference.

[c]PI: prediction interval.

[d]*P* values <.05 were considered statistically significant.

Regarding technical carriers, the pooled MD for the standalone wearable device group (1307.3, 95% CI 299.1-2315.4) was numerically superior to the mobile app or interactive platform group (618.4, 95% CI −305.4 to 1542.2), and its intragroup heterogeneity decreased to 31.4% (subgroup *P*=.20), suggesting this technical carrier is an important factor in explaining daily step-count heterogeneity (Figure S6 in Multimedia Appendix 2). Finally, regarding technical implementation mode, the pooled MD was 459.9 (95% CI −381.1 to 1300.9) for technology-driven interventions (k=7) and 1185.5 (95% CI 43.4 to 2327.6) for technology-mediated interventions (k=7; subgroup *P*=.21). Although the difference did not reach significance, the technology-mediated type demonstrated stronger statistical significance (*P*=.046; Figure S7 in Multimedia Appendix 2).

Subgroup Analysis of the Effects of Multicomponent DHIs on MVPA in Older Adults

The results of the subgroup analysis for MVPA are summarized in Table 3. Regarding the dimension of intervention duration, the pooled MD values for short-term interventions (≤12 wk, k=7) and medium-to-long-term interventions (>12 wk, k=6) were 53.2 (95% CI 24.2-82.2; 95% PI −1.9 to 108.3) and 41.0 (95% CI 4.6-77.4; 95% PI −40.3 to 122.3), respectively. The test for subgroup differences showed *P*=.55, with intragroup heterogeneity of I^2=0% and I^2=24.8%, respectively (Figure S8 in Multimedia Appendix 2). Regarding population characteristics, the pooled MD for the general healthy or community-dwelling older adult group (k=7) and the high-risk or specific population group (k=6) were 40.1 (95% CI 6.1-74.0; 95% PI −40.6 to 120.7) and 57.0 (95% CI 28.8-85.1; 95% PI 14.9-99.0), respectively. The test for subgroup differences

XSL•FO
RenderX

showed P=.45, with intragroup heterogeneity of I^2=27.2% and I^2=0%, respectively (Figure S9 in Multimedia Appendix 2). Regarding control group type, the pooled MD for the active control group (k=7) and the routine care group (k=6) were 39.6 (95% CI 3.3-76.0; 95% PI −41.7 to 120.9) and 54.3 (95% CI 28.2-80.4; 95% PI 5.1-103.5), respectively; no statistically significant difference was observed between the two groups (P=.52), with intragroup heterogeneity of I^2=14.3% and I^2=0%, respectively (Figure S10 in Multimedia Appendix 2).

Table . Summary table evaluating the effects of multicomponent digital health interventions on weekly moderate-to-vigorous physical activity time (min/wk) in older adults across different subgroup characteristics based on the HKSJ[a] random-effects model.

Subgroup classification	k	Pooled MD[b] 95% CI	95% PI[c]	I^2 (%)	$P_{subgroup}$
Duration					.55
Short-term (≤12 wk)	7	53.2 (24.2 to 82.2)	−1.9 to 108.3	0	
Medium-to-long-term (>12 wk)	6	41 (4.6 to 77.4)	−40.3 to 122.3	24.8	
Health status					.45
General healthy/community-dwelling	7	40.1 (6.1 to 74.0)	−40.6 to 120.7	27.2	
High-risk/specific populations	6	57 (28.8 to 85.1)	14.9 to 99.0	0	
Control group setting					.52
Active control/health education	7	39.6 (3.3 to 76.0)	−41.7 to 120.9	14.3	
Routine care/basic monitoring	6	54.3 (28.2 to 80.4)	5.1 to 103.5	0	
Technical carrier					.66
Mobile app/interactive platform	10	52 (28.1 to 76.0)	8.1 to 96.0	0	
Standalone wearable device	3	38.8 (−78.0 to 155.6)	−234.3 to 311.9	38.6	
Delivery agent					.25
Technology-driven	6	58.3 (16.6 to 100.1)	−23.4 to 140.0	3.2	
Technology-mediated	7	36.8 (13.1 to 60.4)	−23.4 to 97.0	20.9	

[a]HKSJ: Hartung-Knapp-Sidik-Jonkman.

[b]MD: mean difference.

[c]PI: prediction interval.

Regarding the dimension of technical carriers, the pooled MD for the mobile app or interactive platform group (k=10) was 52.0 (95% CI 28.1-76.0; 95% PI 8.1-96.0), while the pooled MD for the standalone wearable device group (k=3) was 38.8 (95% CI −78.0-155.6; 95% PI −234.3 to 311.9). The test for subgroup differences showed P=.66, with I^2 values of 0% and 38.6%, respectively (Figure S11 in Multimedia Appendix 2). Finally, regarding the dimension of technical implementation mode, the pooled MD for the technology-driven type (k=6) was 58.3 (95% CI 16.6-100.1; 95% PI −23.4 to 140.0), and the pooled MD for the technology-mediated type (k=7) was 36.8 (95% CI 13.1-60.4; 95% PI −23.4 to 97.0). The test for subgroup differences showed P=.25, with intragroup heterogeneity of I^2=3.2% and I^2=20.9%, respectively (Figure S12 in Multimedia Appendix 2).

Meta-Regression Analysis

For the daily steps indicator, the meta-regression analysis included 14 studies (k=14) [32,36,40-46,48-51,54,55]. The results indicated that none of the 3 examined covariates could significantly explain the interstudy heterogeneity (Figure 4). Specifically, participants' mean age (Figure 4A: P=.25; R^2=2.84%; τ^2=8.25 times 105), intervention duration (Figure

XSL•FO
RenderX

4B: P=.615; R^2=0%; τ^2=1.07 times 106), and prompt frequency (Figure 4C: P=.55; R^2=0%; τ^2=1.08 times 106) showed no significant linear correlation with the postintervention increment in steps. The regression lines appeared flat, with observed values from individual studies distributed on both sides of the lines. Influence diagnostics for these models (detailed in Figures S13-S15 in Multimedia Appendix 2) showed that Cook distance for all included studies remained at low levels, and no strong influential points significantly driving the regression trajectories were found. This suggests that the effect of multicomponent DHIs on increasing daily steps maintained good stability across different age groups, implementation periods, and reminder intensities.

Figure 4. Univariate meta-regression scatter plots evaluating the effects of participants' mean age, total intervention duration, and intervention prompt frequency on daily steps and weekly MVPA in multicomponent digital health interventions. Each bubble in the figure represents an original study, with the bubble size proportional to the weight assigned to that study in the regression analysis (typically the inverse of the sample size). The solid center line represents the fitted meta-regression line, reflecting the correlation trend between the independent variable and the effect size. The shaded area represents the 95% CI of the regression line. P values are used to determine whether the regression coefficients are statistically significant. MD: mean difference; MVPA: moderate-to-vigorous physical activity.

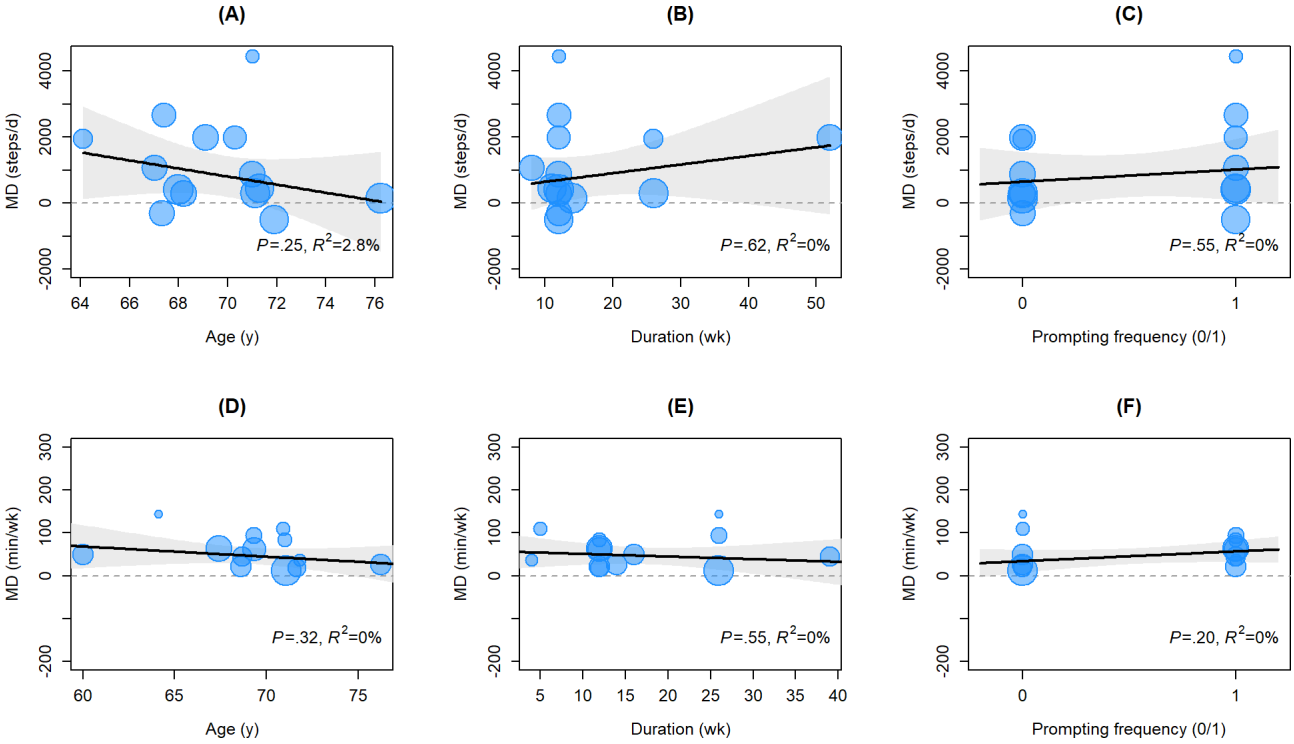

For the MVPA indicator, the meta-regression analysis covered 13 studies (k=13) [32,34,37,38,40,41,43,45-47,49,51,53]. The results demonstrated that the intervention effect for this indicator was likewise not significantly influenced by the preset moderators. Participants' age (Figure 4D: P=.32; R^2=0%; τ^2=601.43), intervention duration (Figure 4E: P=.55; R^2=0%; τ^2=614.94), and prompt frequency (Figure 4F: P=.20; R^2=0%; τ^2=574.23) all failed to reach statistical significance. Influence diagnostics identified a few studies with relatively high Cook distances (detailed in Figures S16-S18 in Multimedia Appendix 2), but after applying the Sidik-Jonkman estimator and Hartung-Knapp adjustment, the driving effect of these strong influential points on the regression slopes was effectively suppressed. Although some residual heterogeneity (τ^2) remained in the models, the R^2 values were all 0%, indicating that the variation in MVPA effect sizes could not be explained by the aforementioned variables.

Analysis of Small-Study Effects and Robustness Testing

The analysis of small-study effects (Figure 5) showed that for the daily steps indicator, the Egger test result was P=.06 (>.05), and the distribution of study points within the funnel plot was basically symmetrical. This suggests that there is no obvious asymmetry caused by small-study effects for this indicator. Conversely, for the MVPA indicator, the Egger test result was P=.01 (<.05), and the funnel plot exhibited significant asymmetry. Furthermore, there was a lack of small-sample, negative-result studies distributed to the left of the centerline at the bottom of the funnel. This indicates a significant risk of small-study effects for this indicator, which may stem from publication bias, differences in study quality, or high heterogeneity.

XSL•FO
RenderX

Figure 5. Egger linear regression test funnel plots of the effects of multicomponent digital health interventions on daily steps and weekly MVPA in older adults. Each scatter point in the figure represents an original study. Given that this systematic review included a large total sample size, the distribution of the funnel plot can more reliably reflect the authenticity of the evidence. The horizontal axis represents the effect size (mean difference), and the vertical axis represents the SE. The center vertical solid line is the pooled effect size, and the diagonal lines on both sides represent the 95% CIs. The highly symmetrical distribution of points within the funnel, combined with the Egger test *P* value, demonstrates whether significant small-sample effects or publication bias exist. MVPA: moderate-to-vigorous physical activity.

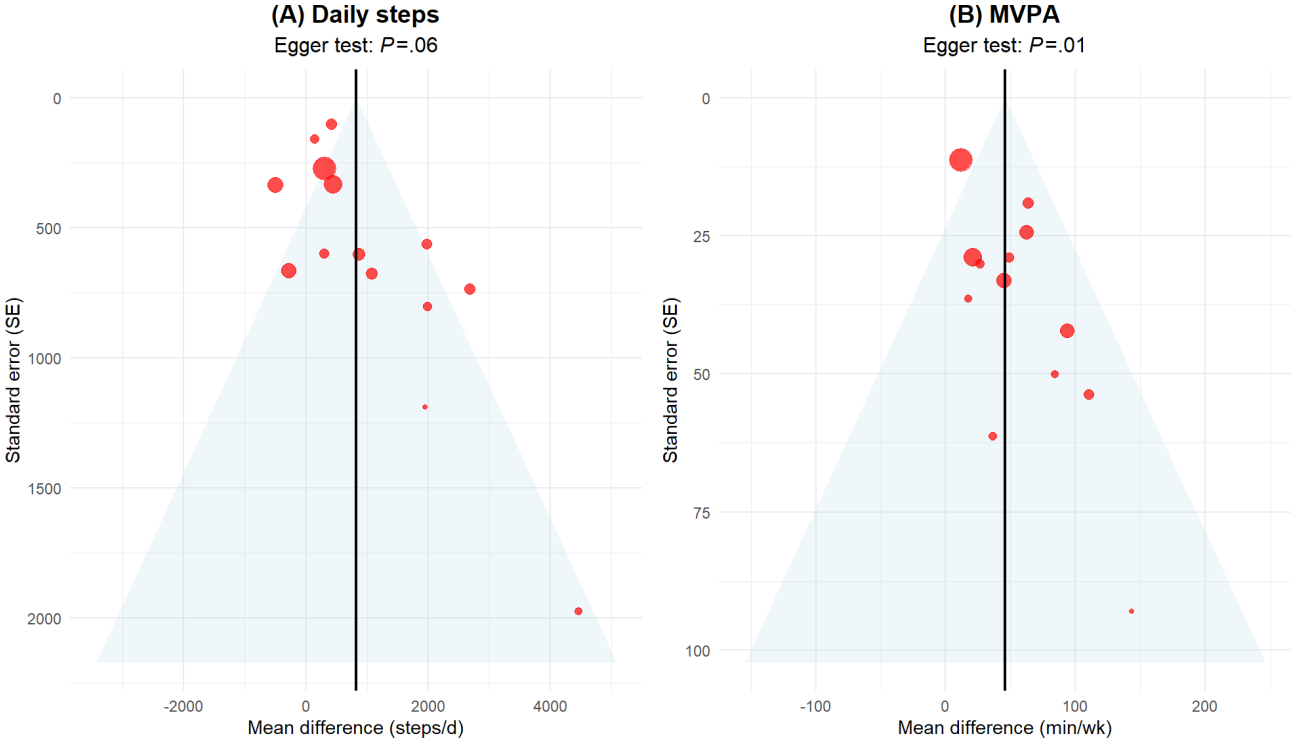

Regarding the detected asymmetry, this systematic review used the trim-and-fill method for robustness analysis (Figure 6). For the daily steps indicator, the model identified and filled in 3 potentially missing studies. The adjusted pooled effect size was MD=521.4 (95% CI −239.3 to 1282.0; *P*=.18) steps/d. The results indicate that after considering potential small-study effects, the significant increase in daily steps observed in the original analysis shifted to nonsignificant. For the MVPA indicator, the model identified and filled in 4 potentially missing studies. The adjusted pooled effect size was MD=36.5 (95% CI 11.2-61.8; *P*=.005) min/wk. Although the adjusted effect size decreased compared to the original value (MD=45.9), its statistical significance remained robust. In summary, the distribution after filling was more symmetrical than the original funnel plot. Despite the identified potential asymmetry, MVPA maintained significant clinical meaning after adjustment.

XSL•FO
RenderX

Figure 6. Funnel plots based on the trim-and-fill method correcting for publication bias regarding the effects of digital health interventions on daily steps and weekly MVPA in older adults. MVPA: moderate-to-vigorous physical activity.

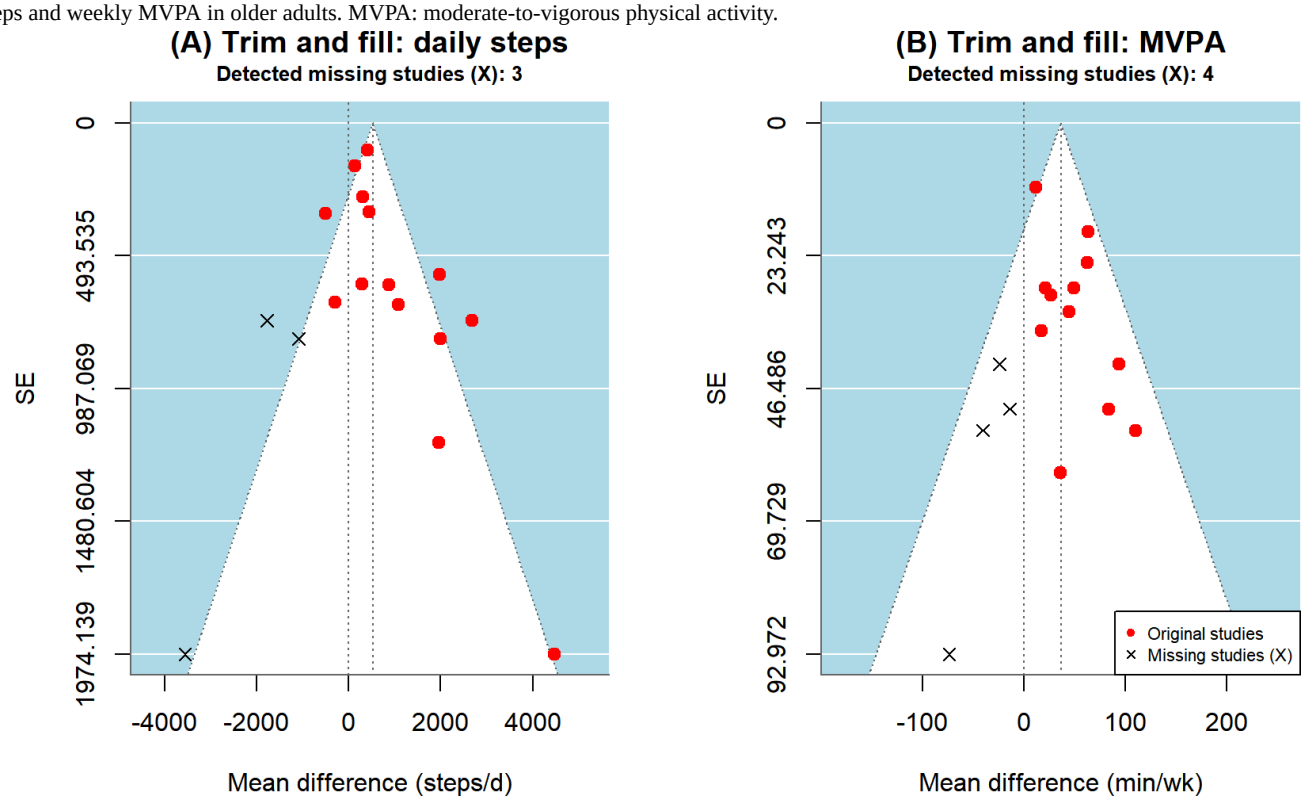

Evidence Certainty Assessment

This systematic review used the GRADE system to assess the quality of evidence for various indicators of DHIs affecting PA in older adults (Table 4). The assessment results showed that only the quality of evidence for MVPA duration was rated as "moderate." Although this indicator was downgraded by one level due to significant publication bias (Egger test P=.01), its pooled effect remained robust after correction via the trim-and-fill method. In contrast, the quality of evidence for daily steps was rated as "low." This indicator was downgraded by one level due to significant inconsistency between studies (I^2=67.5%) and further downgraded in the publication bias dimension, because it lost statistical significance during the trim-and-fill stress test (adjusted P=.18), suggesting that the results were severely interfered with by small-study effects. Furthermore, the quality of evidence for sedentary time, TPA, and LPA were all rated as "low." Specifically, sedentary time was downgraded by one level each in the dimensions of inconsistency and imprecision due to extremely high heterogeneity (I^2=76.2%) and a limited number of included studies (k=5). TPA and LPA not only exhibited moderate-to-high heterogeneity (I^2 of 55.3% and 60.3%, respectively), but their pooled effect 95% CIs both crossed the zero line, indicating nonsignificant statistical meaning and a serious risk of imprecision.

XSL•FO
RenderX

Table . Summary of the quality of evidence grading for outcome indicators (daily steps, MVPA[a], LPA[b], TPA[c], and SB[d]) related to physical activity affected by multicomponent digital health interventions based on the GRADE[e] framework.

Outcome	Studies (n, k)	Sample size (N)	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Publication bias	Certainty
MVPA time (min/wk)	13	1523	MD[f] 45.9 (23.9 to 67.9)	Not serious	Not serious	Not serious	Not serious	Serious[g]	Moderate ⊕⊕⊕
Daily steps (steps/d)	14	1651	MD 822.8 (198.3 to 1447.3)	Not serious	Serious[h]	Not serious	Not serious	Serious[i]	Low ⊕⊕
Sedentary time (min/wk)	5	741	MD −283.7 (−610.8 to 43.5)	Not serious	Serious[j]	Not serious	Serious[k]	Not detected	Low ⊕⊕
TPA (min/wk)	7	886	MD 104.4 (−109.2 to 318.0)	Not serious	Serious[l]	Not serious	Serious[m]	Not detected	Low ⊕⊕
LPA (min/wk)	4	1153	MD 39.3 (−96.2 to 174.7)	Not serious	Serious[n]	Not serious	Serious[o]	Not detected	Low ⊕⊕

[a]MVPA: moderate-to-vigorous physical activity.

[b]LPA: light physical activity.

[c]TPA: total physical activity.

[d]SB: sedentary behavior.

[e]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.

[f]MD: mean difference.

[g]Publication bias: Egger test showed significant publication bias (P=.014); although results remained robust after trim-and-fill correction, the quality was downgraded by one level according to standard protocols.

[h]Inconsistency: significant heterogeneity existed among studies (I^2=67.5%), and the prediction interval (95% PI –1452.4 to 3098.0) crossed the null line, indicating substantial variation in intervention effects across studies, thus the quality was downgraded by one level.

[i]Publication bias: the trim-and-fill stress test showed that the corrected results lost statistical significance (P=.18), suggesting the findings were severely affected by small-study effects, therefore the quality was downgraded by one level in the publication bias domain .

[j]Inconsistency: the original pooled results showed high heterogeneity (I^2=76.2%) and the prediction interval crossed the zero line, thus the quality was downgraded by one level.

[k]Imprecision: the number of included studies was small (k=5), the sample size was insufficient, and the 95% CI crossed the null line, indicating a risk of imprecision, therefore the quality was downgraded by one level.

[l]Inconsistency: moderate-to-high heterogeneity existed (I^2=55.3%), thus the quality was downgraded by one level.

[m]Imprecision: the effect size was nonsignificant, its 95% CI crossed the zero line, and the sample size was small, thus the quality was downgraded by one level.

[n]Inconsistency: moderate heterogeneity existed (I^2=60.3%), thus the quality was downgraded by one level.

[o]Imprecision: the effect size was not statistically significant, the CI crossed the zero line, and the sample size was limited, thus the quality was downgraded by one level.

Discussion

Main Findings and Evidence Overview

This systematic review evaluated the effects of multicomponent DHIs on PA in older adults through a systematic review and meta-analysis of 26 RCTs. The results found that multicomponent DHI can significantly increase the weekly duration of MVPA in older adults; this result remained robust after correction using the trim-and-fill method and is supported by moderate-quality evidence. In contrast, although the original analysis showed a significant increase in daily steps, the quality of evidence for this indicator was rated as “low” because it lost statistical significance under the trim-and-fill stress test and was affected by significant interstudy heterogeneity. However, due to high heterogeneity and low-quality evidence ratings, it is currently not possible to confirm the statistical significance of DHI in reducing SB, or increasing LPA and TPA. In the subgroup analysis of daily steps, using standalone wearable devices as a technical means showed a superior effect compared to mobile apps or web platforms and significantly reduced intragroup heterogeneity; simultaneously, the effect of interventions on increasing steps in high-risk or specific diseased populations was superior to that in healthy older populations. Other subgroup results for daily steps showed that technology-mediated interventions (involving human participation) and studies using routine care as a control observed larger effect sizes, while different intervention durations (≤12 wk vs >12 wk) showed no significant difference in their impact on step increases. In the subgroup analysis of MVPA, no significant differences in effects were observed

XSL•FO
RenderX

across different durations, population characteristics, or technical means, demonstrating good consistency. Meta-regression analysis further confirmed that age, intervention duration, and prompt frequency did not exert significant moderating effects on the effect sizes of steps and MVPA, indicating that after excluding strong influential points, the intervention effect maintained stability across different participant characteristics and implementation periods. Finally, this systematic review clearly distinguished between the average effect and the distribution of effects; although the CIs of core indicators were significant, the PIs for all outcome indicators crossed the line of no effect, reflecting obvious dispersion of intervention effects across different research settings and contexts. This suggests that in future-specific implementation environments, there remains a possibility that interventions may not produce the expected benefits. Therefore, the potential impact of participant characteristics and intervention environments should be fully considered when promoting multicomponent DHI.

Analysis of Influence Patterns and Moderating Effects of Multicomponent DHIs on Exercise Behavior

The effects of multicomponent DHIs in increasing daily steps and MVPA in older adults stem from the enhancement of participants' self-efficacy by behavior change techniques, such as real-time monitoring, goal setting, and automated feedback [56]. From a public health perspective, the increase in steps observed in this systematic review (+822.8 steps/d) holds clear clinical benefit value [31]. Previous research has indicated that for the older adult population, an increase of even approximately 500 to 1000 daily steps is significantly associated with a reduced risk of all-cause mortality and can effectively delay the onset of mobility impairment [57]. Meanwhile, the weekly increment of approximately 46 minutes of MVPA reaches nearly one-third of the level recommended by the World Health Organization, which carries important translational medicine significance for improving cardiorespiratory fitness and insulin sensitivity, and reducing cardiovascular disease risk in older adults [58,59]. The step enhancement effect in this systematic review is consistent with the research direction of Wu et al (2023) [18] regarding the use of wearable devices by older adults (SMD 0.59) [18]. In contrast, Wang et al (2025) [15] pointed out that the improvement in steps by application-based interventions (SMD 0.18) was relatively modest, while the promotion effect on MVPA (SMD 0.48) was more prominent [15]. However, when interpreting the findings, a clear distinction must be made between CIs and PIs. Although the CIs for steps and MVPA were significant, their PIs both crossed the null line. This indicates that, influenced by moderate statistical heterogeneity between studies, the intervention effects exhibit significant dispersion across different clinical settings.

This systematic review found that although multicomponent DHIs increased steps and MVPA, their effectiveness in improving SB was limited, which may reflect limitations in the design logic of current intervention protocols [32,39]. Most multicomponent DHI protocols feature monitoring and feedback mechanisms that focus on capturing and motivating "planned exercise" that reaches intensity thresholds, while lacking immediate disruption strategies for prolonged SB [60]. This results in participants being more inclined to increase conscious exercise increments, yet a significant evidence gap remains regarding the reconstruction of all-day sedentary patterns [18]. Furthermore, the lack of a simultaneous significant increase in TPA further reveals the complexity of exercise behavior change in older adults [43]. Despite the significant increase in MVPA, the improvement in overall activity levels remains limited because multicomponent DHIs often lack immediate incentives for trivial, low-intensity daily activities [61,62]. Future protocol designs should optimize reminder strategies [63] and attempt to expand the intervention focus from pure exercise increments to all-day sedentary interruptions, aiming to achieve a more comprehensive reshaping of behavioral patterns in older adults [64]. Additionally, it is noteworthy that the regression coefficients of all meta-regression models were not statistically significant. This result suggests that the preset clinical and methodological covariates do not explain the significant heterogeneity observed between studies; fluctuations in effect size may stem more from deep-level variables that were not captured, such as participants' baseline digital literacy, the specific interaction logic of the intervention protocol, or the degree of social support [42,65]. Meanwhile, considering the limited number of studies included in the models, insufficient statistical power may have restricted the ability of the models to detect weak moderating effects [66]; therefore, the findings of the meta-regression should be cautiously interpreted as a failure to identify significant moderators within the existing evidence scale. After excluding influential points using the HKSJ model, this finding proves that the positive effects of multicomponent DHIs maintained good stability in the older population aged 60 to 77 years. This challenges the traditional assumption that older age groups inevitably face a "digital divide bottleneck" leading to intervention failure [67], suggesting that as long as the protocol design is reasonable, the response of older adults to core behavior change techniques has cross-population reference value. Subgroup analysis results provide preliminary clues for understanding the heterogeneity of intervention effects. Regarding the daily steps dimension, the technical carrier was identified as a potential source of heterogeneity, with intragroup heterogeneity reduced in the standalone wearable device group. Although the statistical power for testing between-group differences was limited due to the small number of studies within subgroups, standalone wearable devices showed a better effect size trend than the mobile app group, which may be attributed to a more direct interaction method that reduces participants' dependence on digital literacy [68]. Furthermore, the significant effect on step indicators for populations at high risk of chronic diseases also suggests that when intervention goals are highly correlated with an individual's sense of health security, multicomponent digital health technology can generate stronger behavioral drivers [69]. Notably, this systematic review also observed opposing trends in the effects of delivery agents across different indicators: step increases leaned toward technology-mediated human support, while MVPA enhancement was more deeply influenced by technology-driven real-time feedback [48,56], further demonstrating the necessity for precision design in multicomponent digital protocols.

Significantly, publication bias tests further revealed the essential difference in evidence certainty between daily steps and MVPA.

XSL•FO
RenderX

After applying the trim-and-fill method to adjust for potential small-study effects, the pooled effect for daily steps exhibited clear vulnerability as it lost statistical significance [70]. This finding raises doubts about the robustness of the original effect size for this indicator and directly led to its GRADE evidence quality being downgraded from "moderate" to "low" [24]. In contrast, the intervention effect on MVPA remained robust under stress testing, supporting its evidence-based value as "moderate" quality evidence. This contrast clarifies that the driving effect of multicomponent DHIs on high-intensity exercise behavior in older adults is more certain than simply increasing steps. Combined with the GRADE evidence quality assessment, although the presence of heterogeneity lowered the certainty rating of some outcome indicators, the observed effect sizes reached the aforementioned clinical benefit benchmarks, making their reference value for clinical decision-making remain clear [31,57]. When discussing the sustainability of intervention effects, it must be clarified that the current meta-regression reflects the stability of effect sizes during the intervention implementation period. However, in the current digital health field, distinguishing between "effectiveness during long-term intervention" and "effect maintenance after intervention withdrawal" is crucial, and the latter still faces a significant research evidence gap in existing literature [47,48]. This inference suggests that future research should focus on how to transform digital external incentives into participants' lasting internal behavioral inertia [71] and conduct age-friendly modifications of digital interfaces for older adults to maintain participation momentum, which still requires more long-term tracking studies for confirmation [72].

Practical Implications

This systematic review provides evidence-based support for the management of community-dwelling older adults' health and the development of multicomponent digital health exercise intervention protocols. At the public health policy level, multicomponent DHIs can serve as an effective supplement to traditional face-to-face interventions; particularly in resource-limited community settings, their features of wide coverage and remote support offer high cost-effectiveness [36]. Given the significant benefits observed among populations at high risk of chronic diseases, it is recommended that multicomponent digital health monitoring measures be positioned forward into the risk screening stage, aiming to delay the transition of older adults toward pathological states through early behavioral interventions [33]. The practical path should not only pursue an increase in exercise volume but should also attempt to guide older adults to reduce SB by increasing light activities such as walking and housework. This approach is more consistent with the physical condition of the advanced older adults and is easier to maintain over the long term than simply pursuing high-intensity exercise goals [73]. However, given that the PIs for the core indicators in this systematic review all crossed the null line, relevant institutions must remain cautious when translating the evidence-based findings of this systematic review. The effectiveness of multicomponent DHIs highly depends on the specific implementation environment, the intensity of technical support, and the digital literacy of the participants; in contexts lacking resources or necessary human assistance, there is a risk that intervention effects may fail to reach targets [74]. Therefore, practice should emphasize personalized adaptation based on local conditions.

In terms of technical carrier selection, the efficacy advantages demonstrated by standalone wearable devices suggest that the research and development of age-friendly digital products should prioritize simplifying human-computer interaction logic. Compared to mobile apps that require higher digital skills, hardware carriers with passive monitoring processes and intuitive feedback are more aligned with the operating habits of the older population [36]. Addressing the phenomenon of diminishing effects in advanced older individuals, product design needs to provide differentiated interface displays and guidance modes [10]. Simultaneously, intervention protocols should focus on the trend of effect attenuation over extended cycles [75]. Clinicians and managers need to dynamically adjust goal settings based on participant feedback, and be timely in introducing social support or professional involvement to mitigate the decline in adherence during long-term application [42]. Furthermore, intervention protocols should break through single step-count assessments and emphasize diversified goals such as "sedentary interruptions" and the "embedding of trivial physical activities," achieving a comprehensive reshaping of the daily behavior patterns of older adults [46].

Research Limitations and Future Outlook

Although this systematic review evaluated the efficacy of multicomponent DHIs through multidimensional indicators, some limitations remain in terms of methodology and evidence application. Since most included studies combined digital tools with human support, the independent contribution of digital technology remains uncertain [31,36]. At the study design level, the specificity of digital interventions meant that most original studies could not achieve double-blinding of participants and implementers, increasing the risk of performance bias [39,46]. Furthermore, the attribution analysis of intervention effects is similarly limited; because most protocols adopted a multicomponent model, with differences in core interaction agents compared to auxiliary monitoring tools and the intensity of accompanying human support. The complexity of these components makes it difficult to completely isolate the respective contributions of the automated features of the digital technology itself from the accompanying social support or professional guidance [39,76]. Meanwhile, the significant variation in control group settings, ranging from passive waiting to active health education, may have diluted the estimation of the "incremental value" of digital technology [32,47]. From a systemic perspective of behavior monitoring, this systematic review was not yet able to fully incorporate the daily behaviors of older adults into a 24-hour PA comprehensive framework; since the total daily time for older adults is constant, changes in specific exercise indicators caused by digital interventions are inevitably accompanied by the reallocation of sleep or trivial daily activities [77,78], and this systematic review struggled to reveal the deep impact of digital interventions on all-day time-use patterns. Additionally, considering that the follow-up periods of most studies coincided with the intervention cycles, the effects revealed by this systematic review primarily reflect immediate efficacy during the implementation period; as the

XSL•FO
RenderX

original literature generally lacks long-term tracking data after the cessation of interventions, it remains difficult to confirm whether digital interventions can be transformed into the maintenance of lasting behavioral habits once technical support is removed [33,41]. Although this systematic review covered multiple indicators, the number of studies targeting LPA and TPA was relatively small, which perhaps resulted in insufficient statistical power for the pooled effect sizes. Furthermore, differences existed among original studies in the choice of PA measurement tools, including both objective monitoring devices and self-report questionnaires; this inconsistency in measurement methods may have introduced a degree of heterogeneity during the data pooling process [35,44]. The participant populations were mainly concentrated on healthy community-dwelling older adults or groups at high risk of chronic diseases; therefore, caution should be exercised when extrapolating the conclusions to groups with severe functional impairments or specific clinical conditions [36,46]. Although the results showed that the moderating effects of age, intervention duration, and prompting frequency were all nonsignificant, this may be limited by insufficient statistical power due to the small number of included studies. Moreover, the residual heterogeneity (τ^2) still present in the models suggests that unobserved variables, such as baseline digital skills and socioeconomic status, may still contribute potential influences [79].

Based on the aforementioned limitations, future research needs to deepen in the following directions. First, emphasis should be placed on distinguishing "effectiveness during the intervention period" from "effect maintenance after intervention withdrawal," evaluating the potential of digital technology in promoting the long-term sustainability of behavior change in older adults through long-term follow-up after a washout period [41]. Given the trend of effect attenuation that may occur as cycles extend, exploring dynamic feedback mechanisms with lasting driving force and age-friendly interface optimization for the advanced older adults is crucial [75]. Second, subsequent studies should actively adopt 24-hour PA comprehensive guidelines to explore how multicomponent digital health technology can optimize health outcomes for older adults through more scientific time reallocation strategies by simultaneously monitoring sleep, SB, and PA of various intensities [78]. Regarding the complexity of intervention components, it is recommended to conduct head-to-head studies or component-dismantling trials to clarify the efficacy differences between automated technology components and different intensities of human support [80]. Additionally, conducting precise exercise prescription research targeting specific health risk subgroups will help provide more systemic evidence-based support for the digital health management of older adults [81].

Conclusions

This review systematically evaluated the impact of multicomponent DHIs on the PA behavior of older adults. The research suggests that digital interventions may effectively improve daily steps and MVPA levels in older adults, but there is currently no significant statistical evidence to support their role in reducing sedentary time or improving LPA. Since most original studies combined digital tools with human support, the independent driving role of digital technology within these interventions requires further verification. Meta-regression and subgroup analyses further revealed that intervention effects maintained good stability across different age gradients, intervention cycles, and frequencies, and that standalone wearable devices possessed an efficacy advantage over mobile apps in driving behavior change in older adults. Unlike previous reviews focusing on general adult populations or single exercise indicators, this systematic review focuses on the uniqueness of the older population; through in-depth comparison with existing evidence, this systematic review not only verifies the immediate effectiveness of multicomponent digital interventions, but also evaluates the distribution of effects in real-world settings through more cautious statistical models, identifying that intervention benefits possess clear uncertainty across different contexts. This systematic review brings key evidence regarding the good adaptability of the older population to multicomponent digital interventions, suggesting that advanced age does not significantly weaken the efficacy of technical interventions, and clearly identifies "effect maintenance after intervention withdrawal" as an urgent gap to be filled in the current evidence chain. In terms of real-world significance, this systematic review confirms that promoting hardware carriers with simple interaction and high automation is an effective path for reshaping health behavior patterns in older adults, providing core evidence-based guidance for formulating low-cost, broad-coverage, and high-adherence precise public health intervention strategies in an aging society.

Acknowledgments

We thank all original study authors included in this meta-analysis for their contributions to the literature. The authors state that no generative artificial intelligence or artificial intelligence–assisted technologies were used in the writing of this paper or in the performance of the underlying research.

Funding

This review received no external funding. The authors received no financial support for the research, authorship, and/or publication of this paper.

Data Availability

The datasets supporting the conclusions of this meta-analysis are included in this paper and its supplementary files. Search strategies and data extraction forms are available from the corresponding author upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

JY and SC conceived and designed the research. JY, HW, and SJ were responsible for the literature search and data extraction. HC and JH performed the quality assessment and statistical analysis. JL contributed to data visualization and interpretation. JY drafted the manuscript. SC provided supervision, project administration, and critical revision of the manuscript for important intellectual content.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Comprehensive search strategies and Boolean logic for all 5 databases updated as of February 20, 2026, including the detailed deduplication process.
[DOCX File, 29 KB - jmir_v28i1e91338_app1.docx]

Multimedia Appendix 2
Evidence and robustness analyses, including leave-one-out sensitivity analysis, HKSJ random-effects forest plots with 95% PIs for primary and subgroup outcomes, and meta-regression influence diagnostics. HKSJ: Hartung-Knapp-Sidik-Jonkman; PI: prediction interval.
[DOCX File, 1392 KB - jmir_v28i1e91338_app2.docx]

Checklist 1
PRISMA 2020 checklist which specifies the precise manuscript page and line numbers where each of the 27 reporting items—from title and abstract to methods, results, and discussion—is addressed.
[PDF File, 193 KB - jmir_v28i1e91338_app3.pdf]

Checklist 2
PRISMA 2020 for Abstracts checklist, which confirms that the study's abstract follows the 12-item reporting standard covering essential elements such as eligibility criteria, synthesis of results, and funding information.
[PDF File, 148 KB - jmir_v28i1e91338_app4.pdf]

Checklist 3
PRISMA-S checklist which details the reporting of literature search strategies across 16 items, specifically mapping information sources, search strings, and deduplication processes to their respective locations in the manuscript and Multimedia Appendix 1.
[PDF File, 102 KB - jmir_v28i1e91338_app5.pdf]

References

1. Beard JR, Officer AM, Cassels AK. The world report on ageing and health. Gerontologist 2016 Apr;56 Suppl 2(Suppl 2):S163-S166. [doi: 10.1093/geront/gnw037] [Medline: 26994257]
2. Amuthavalli Thiyagarajan J, Mikton C, Harwood RH, et al. The UN decade of healthy ageing: strengthening measurement for monitoring health and wellbeing of older people. Age Ageing 2022 Jul 1;51(7):afac147. [doi: 10.1093/ageing/afac147] [Medline: 35776669]
3. Bull FC, Al-Ansari SS, Biddle S, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. Br J Sports Med 2020 Dec;54(24):1451-1462. [doi: 10.1136/bjsports-2020-102955] [Medline: 33239350]
4. Cruz-Jentoft AJ, Bahat G, Bauer J, et al. Sarcopenia: revised European consensus on definition and diagnosis. Age Ageing 2019 Jul 1;48(4):601-601. [doi: 10.1093/ageing/afz046] [Medline: 30312372]
5. Guthold R, Stevens GA, Riley LM, Bull FC. Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1·9 million participants. Lancet Glob Health 2018 Oct;6(10):e1077-e1086. [doi: 10.1016/S2214-109X(18)30357-7] [Medline: 30193830]
6. Harvey JA, Chastin SFM, Skelton DA. How sedentary are older people? a systematic review of the amount of sedentary behavior. J Aging Phys Act 2015 Jul;23(3):471-487. [doi: 10.1123/japa.2014-0164] [Medline: 25387160]
7. Yerrakalva D, Yerrakalva D, Hajna S, Griffin S. Effects of mobile health app interventions on sedentary time, physical activity, and fitness in older adults: systematic review and meta-analysis. J Med Internet Res 2019 Nov 28;21(11):e14343. [doi: 10.2196/14343] [Medline: 31778121]
8. Mair JL, Salamanca-Sanabria A, Augsburger M, et al. Effective behavior change techniques in digital health interventions for the prevention or management of noncommunicable diseases: an umbrella review. Ann Behav Med 2023 Sep 13;57(10):817-835. [doi: 10.1093/abm/kaad041] [Medline: 37625030]

XSL•FO
RenderX

9. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
10. Di Pumpo M, Miatton A, Riccardi MT, et al. Digital health interventions to promote physical activity in community-dwelling older adults: a systematic review and semiquantitative analysis. Int J Public Health 2024;69:1607720. [doi: 10.3389/ijph.2024.1607720] [Medline: 39830161]
11. Chen W, Yao J, Wang H, Jia S, Zhu X, Mao L. Effects of digital health interventions on muscle mass, muscle strength, and physical function in older adults with sarcopenia: a systematic review and meta-analysis. Front Public Health 2025;13:1711514. [doi: 10.3389/fpubh.2025.1711514] [Medline: 41383324]
12. Marcolino MS, Oliveira JAQ, D'Agostino M, Ribeiro AL, Alkmim MBM, Novillo-Ortiz D. The impact of mHealth interventions: systematic review of systematic reviews. JMIR mHealth uHealth 2018 Jan 17;6(1):e23. [doi: 10.2196/mhealth.8873] [Medline: 29343463]
13. Zangger G, Bricca A, Liaghat B, et al. Benefits and harms of digital health interventions promoting physical activity in people with chronic conditions: systematic review and meta-analysis. J Med Internet Res 2023 Jul 6;25:e46439. [doi: 10.2196/46439] [Medline: 37410534]
14. Boxall C, Dennison L, Miller S, et al. Implementing a digital physical activity intervention for older adults: qualitative study. JMIR Aging 2025 Aug 21;8:e64953. [doi: 10.2196/64953] [Medline: 40840434]
15. Wang G, Xiang R, Yang X, Tan L. Digital technology empowers exercise health management in older adults: a systematic review and meta-analysis of the effects of mHealth-based interventions on physical activity and body composition in older adults. Front Public Health 2025;13:1661028. [doi: 10.3389/fpubh.2025.1661028] [Medline: 41103467]
16. Park Y, Kim EJ, Park S, Lee M. Digital health intervention effect on older adults with chronic diseases living alone: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2025 Mar 31;27:e63168. [doi: 10.2196/63168] [Medline: 40163849]
17. Franssen WMA, Franssen GHLM, Spaas J, Solmi F, Eijnde BO. Can consumer wearable activity tracker-based interventions improve physical activity and cardiometabolic health in patients with chronic diseases? A systematic review and meta-analysis of randomised controlled trials. Int J Behav Nutr Phys Act 2020 May 11;17(1):57. [doi: 10.1186/s12966-020-00955-2] [Medline: 32393357]
18. Wu S, Li G, Du L, Chen S, Zhang X, He Q. The effectiveness of wearable activity trackers for increasing physical activity and reducing sedentary time in older adults: a systematic review and meta-analysis. Digit Health 2023;9:20552076231176705. [doi: 10.1177/20552076231176705] [Medline: 37252261]
19. Kwan RYC, Salihu D, Lee PH, et al. The effect of e-health interventions promoting physical activity in older people: a systematic review and meta-analysis. Eur Rev Aging Phys Act 2020;17:7. [doi: 10.1186/s11556-020-00239-5] [Medline: 32336996]
20. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
21. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
22. Luo D, Wan X, Liu J, Tong T. Optimally estimating the sample mean from the sample size, median, mid-range, and/or mid-quartile range. Stat Methods Med Res 2018 Jun;27(6):1785-1805. [doi: 10.1177/0962280216669183] [Medline: 27683581]
23. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
24. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-926. [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
25. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14(1):25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
26. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019 Jun;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]
27. Stockwell S, Schofield P, Fisher A, et al. Digital behavior change interventions to promote physical activity and/or reduce sedentary behavior in older adults: a systematic review and meta-analysis. Exp Gerontol 2019 Jun;120:68-87. [doi: 10.1016/j.exger.2019.02.020] [Medline: 30836130]
28. Muellmann S, Forberger S, Möllers T, Bröring E, Zeeb H, Pischke CR. Effectiveness of eHealth interventions for the promotion of physical activity in older adults: a systematic review. Prev Med 2018 Mar;108:93-110. [doi: 10.1016/j.ypmed.2017.12.026] [Medline: 29289643]
29. Sterne JAC, Sutton AJ, Ioannidis JPA, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011 Jul 22;343:d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]

XSL•FO
RenderX

30. Audsley S, Kendrick D, Logan P, Jones M, Orton E. A randomised feasibility study assessing an intervention to keep adults physically active after falls management exercise programmes end. Pilot Feasibility Stud 2020;6:37. [doi: 10.1186/s40814-020-00570-9] [Medline: 32161660]
31. Saçıkara Z, Cingil D. The effect of multiple nursing interventions on physical activity and health promotion in the elderly: a randomized controlled trial. Geriatr Nurs 2024;59:150-158. [doi: 10.1016/j.gerinurse.2024.06.036] [Medline: 39002505]
32. Ashe MC, Winters M, Hoppmann CA, et al. “Not just another walking program”: Everyday Activity Supports You (EASY) model-a randomized pilot study for a parallel randomized controlled trial. Pilot Feasibility Stud 2015;1:4. [doi: 10.1186/2055-5784-1-4] [Medline: 27175291]
33. Li N, Wang N, Xu Y, et al. The impacts of a mHealth platform-enabled lifestyle-integrated multicomponent exercise program on reversing pre-frailty in community-dwelling older adults: a randomized controlled trial. Int J Nurs Stud 2025 Jul;167:105072. [doi: 10.1016/j.ijnurstu.2025.105072] [Medline: 40222237]
34. Van Dyck D, Herman K, Poppe L, Crombez G, De Bourdeaudhuij I, Gheysen F. Results of MyPlan 2.0 on physical activity in older Belgian adults: randomized controlled trial. J Med Internet Res 2019 Oct 7;21(10):e13219. [doi: 10.2196/13219] [Medline: 31593541]
35. Delbaere K, Valenzuela T, Lord SR, et al. E-health StandingTall balance exercise for fall prevention in older people: results of a two year randomised controlled trial. BMJ 2021 Apr 6;373:n740. [doi: 10.1136/bmj.n740] [Medline: 33824131]
36. Brickwood KJ, Ahuja KDK, Watson G, O’Brien JA, Williams AD. Effects of activity tracker use with health professional support or telephone counseling on maintenance of physical activity and health outcomes in older adults: randomized controlled trial. JMIR mHealth uHealth 2021 Jan 5;9(1):e18686. [doi: 10.2196/18686] [Medline: 33399541]
37. Cadmus-Bertram LA, Marcus BH, Patterson RE, Parker BA, Morey BL. Randomized trial of a fitbit-based physical activity intervention for women. Am J Prev Med 2015 Sep;49(3):414-418. [doi: 10.1016/j.amepre.2015.01.020] [Medline: 26071863]
38. Lee JLC, Wong AYL, Ng PHF, et al. Outdoor exercise facility-based integrative mobile health intervention to support physical activity, mental well-being, and exercise self-efficacy among older adults with prefrailty and frailty in Hong Kong: pilot feasibility randomized controlled trial study. JMIR mHealth uHealth 2025 Jun 5;13:e69259. [doi: 10.2196/69259] [Medline: 40471669]
39. Mouton A, Cloes M. Efficacy of a web-based, center-based or combined physical activity intervention among older adults. Health Educ Res 2015 Jun;30(3):422-435. [doi: 10.1093/her/cyv012] [Medline: 25772974]
40. Pischke CR, Voelcker-Rehage C, Ratz T, et al. Web-based versus print-based physical activity intervention for community-dwelling older adults: crossover randomized trial. JMIR mHealth uHealth 2022 Mar 23;10(3):e32212. [doi: 10.2196/32212] [Medline: 35319484]
41. Zhou M, Zhang N, Zhang Y, et al. Effect of mobile-based lifestyle intervention on weight loss among the overweight and obese elderly population in China: a randomized controlled trial. Int J Environ Res Public Health 2021 Aug 21;18(16):8825. [doi: 10.3390/ijerph18168825] [Medline: 34444573]
42. Rosenberg DE, Anderson ML, Renz A, et al. Reducing sitting time in obese older adults: the I-STAND randomized controlled trial. J Aging Phys Act 2020 Dec 1;28(6):864-874. [doi: 10.1123/japa.2019-0470] [Medline: 32498040]
43. Oliveira JS, Sherrington C, Rissel C, et al. Effect of a coaching intervention to enhance physical activity and prevent falls in community-dwelling people aged 60+ years: a cluster randomised controlled trial. Br J Sports Med 2024 Mar 21;58(7):382-391. [doi: 10.1136/bjsports-2023-107027] [Medline: 38253435]
44. Roberts LM, Jaeger BC, Baptista LC, et al. Wearable technology to reduce sedentary behavior and CVD risk in older adults: a pilot randomized clinical trial. Clin Interv Aging 2019;14:1817-1828. [doi: 10.2147/CIA.S222655] [Medline: 31695350]
45. Alley SJ, van Uffelen J, Schoeppe S, et al. The effectiveness of a computer-tailored web-based physical activity intervention using Fitbit activity trackers in older adults (active for life): randomized controlled trial. J Med Internet Res 2022 May 12;24(5):e31352. [doi: 10.2196/31352] [Medline: 35552166]
46. Liu JYW, Kwan RYC, Yin YH, Lee PH, Siu JYM, Bai X. Enhancing the physical activity levels of frail older adults with a wearable activity tracker-based exercise intervention: a pilot cluster randomized controlled trial. Int J Environ Res Public Health 2021 Sep 30;18(19):10344. [doi: 10.3390/ijerph181910344] [Medline: 34639644]
47. Uemura K, Kamitani T, Watanabe A, Okamoto H, Saho K, Yamada M. Active learning through video conferencing to maintain physical activity among older adults: a pilot randomized controlled trial. J Aging Phys Act 2024 Dec 1;32(6):730-738. [doi: 10.1123/japa.2023-0180] [Medline: 38830609]
48. Cai X, Qiu S, Luo D, et al. Effects of peer support and mobile application-based walking programme on physical activity and physical function in rural older adults: a cluster randomized controlled trial. Eur Geriatr Med 2022 Oct;13(5):1187-1195. [doi: 10.1007/s41999-022-00682-w] [Medline: 36001254]
49. Oliveira JS, Sherrington C, Paul SS, et al. A combined physical activity and fall prevention intervention improved mobility-related goal attainment but not physical activity in older adults: a randomised trial. J Physiother 2019 Jan;65(1):16-22. [doi: 10.1016/j.jphys.2018.11.005] [Medline: 30581138]
50. Bickmore TW, Silliman RA, Nelson K, et al. A randomized controlled trial of an automated exercise coach for older adults. J Am Geriatr Soc 2013 Oct;61(10):1676-1683. [doi: 10.1111/jgs.12449] [Medline: 24001030]

XSL•FO
RenderX

51. Kwan RY, Lee D, Lee PH, et al. Effects of an mHealth brisk walking intervention on increasing physical activity in older people with cognitive frailty: pilot randomized controlled trial. JMIR mHealth uHealth 2020 Jul 31;8(7):e16596. [doi: 10.2196/16596] [Medline: 32735218]
52. Kawaguchi K, Nakagomi A, Ide K, Kondo K. Effects of a mobile app to promote social participation on older adults: randomized controlled trial. J Med Internet Res 2024 Sep 30;26:e64196. [doi: 10.2196/64196] [Medline: 39348180]
53. Muellmann S, Buck C, Voelcker-Rehage C, et al. Effects of two web-based interventions promoting physical activity among older adults compared to a delayed intervention control group in Northwestern Germany: results of the PROMOTE community-based intervention trial. Prev Med Rep 2019 Sep;15:100958. [doi: 10.1016/j.pmedr.2019.100958] [Medline: 31410347]
54. Yates T, Davies M, Gorely T, Bull F, Khunti K. Effectiveness of a pragmatic education program designed to promote walking activity in individuals with impaired glucose tolerance: a randomized controlled trial. Diabetes Care 2009 Aug;32(8):1404-1410. [doi: 10.2337/dc09-0130] [Medline: 19602539]
55. Recio-Rodríguez JI, Gonzalez-Sanchez S, Tamayo-Morales O, et al. Changes in lifestyles, cognitive impairment, quality of life and activity day living after combined use of smartphone and smartband technology: a randomized clinical trial (EVIDENT-age study). BMC Geriatr 2022 Oct 6;22(1):782. [doi: 10.1186/s12877-022-03487-5] [Medline: 36203135]
56. Lyons EJ, Lewis ZH, Mayrsohn BG, Rowland JL. Behavior change techniques implemented in electronic lifestyle activity monitors: a systematic content analysis. J Med Internet Res 2014 Aug 15;16(8):e192. [doi: 10.2196/jmir.3469] [Medline: 25131661]
57. Lee IM, Shiroma EJ, Kamada M, Bassett DR, Matthews CE, Buring JE. Association of step volume and intensity with all-cause mortality in older women. JAMA Intern Med 2019 Aug 1;179(8):1105-1112. [doi: 10.1001/jamainternmed.2019.0899] [Medline: 31141585]
58. Nelson ME, Rejeski WJ, Blair SN, et al. Physical activity and public health in older adults: recommendation from the American College of Sports Medicine and the American Heart Association. Med Sci Sports Exerc 2007 Aug;39(8):1435-1445. [doi: 10.1249/mss.0b013e3180616aa2] [Medline: 17762378]
59. Ekelund U, Tarp J, Steene-Johannessen J, et al. Dose-response associations between accelerometry measured physical activity and sedentary time and all cause mortality: systematic review and harmonised meta-analysis. BMJ 2019 Aug 21;366:l4570. [doi: 10.1136/bmj.l4570] [Medline: 31434697]
60. Dumuid D, Pedišić Ž, Palarea-Albaladejo J, Martín-Fernández JA, Hron K, Olds T. Compositional data analysis in time-use epidemiology: what, why, how. Int J Environ Res Public Health 2020 Mar 26;17(7):2220. [doi: 10.3390/ijerph17072220] [Medline: 32224966]
61. Sullivan AN, Lachman ME. Behavior change with fitness technology in sedentary adults: a review of the evidence for increasing physical activity. Front Public Health 2016;4:289. [doi: 10.3389/fpubh.2016.00289] [Medline: 28123997]
62. Gomersall SR, Rowlands AV, English C, Maher C, Olds TS. The ActivityStat hypothesis: the concept, the evidence and the methodologies. Sports Med 2013 Feb;43(2):135-149. [doi: 10.1007/s40279-012-0008-7] [Medline: 23329607]
63. Cleland V, Squibb K, Stephens L, et al. Effectiveness of interventions to promote physical activity and/or decrease sedentary behaviour among rural adults: a systematic review and meta-analysis. Obes Rev 2017 Jul;18(7):727-741. [doi: 10.1111/obr.12533] [Medline: 28401687]
64. Bellettiere J, LaMonte MJ, Evenson KR, et al. Sedentary behavior and cardiovascular disease in older women: the objective physical activity and cardiovascular health (OPACH) study. Circulation 2019 Feb 19;139(8):1036-1046. [doi: 10.1161/CIRCULATIONAHA.118.035312] [Medline: 31031411]
65. Zhao YC, Zhao M, Song S. Online health information seeking behaviors among older adults: systematic scoping review. J Med Internet Res 2022 Feb 16;24(2):e34790. [doi: 10.2196/34790] [Medline: 35171099]
66. Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions: John Wiley & Sons; 2008. [doi: 10.1002/9780470712184]
67. Li W, Chen Z. Bridging the urban-rural divide: digital literacy as a catalyst for enhancing physical exercise participation in China. Front Public Health 2025;13:1630850. [doi: 10.3389/fpubh.2025.1630850] [Medline: 41063961]
68. Moore K, O'Shea E, Kenny L, et al. Older adults' experiences with using wearable devices: qualitative systematic review and meta-synthesis. JMIR mHealth uHealth 2021 Jun 3;9(6):e23832. [doi: 10.2196/23832] [Medline: 34081020]
69. Hamine S, Gerth-Guyette E, Faulx D, Green BB, Ginsburg AS. Impact of mHealth chronic disease management on treatment adherence and patient outcomes: a systematic review. J Med Internet Res 2015 Feb 24;17(2):e52. [doi: 10.2196/jmir.3951] [Medline: 25803266]
70. Shi L, Lin L. The trim-and-fill method for publication bias: practical guidelines and recommendations based on a large database of meta-analyses. Medicine (Baltimore) 2019 Jun;98(23):e15987. [doi: 10.1097/MD.0000000000015987] [Medline: 31169736]
71. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000 Jan;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
72. Bertolazzi A, Quaglia V, Bongelli R. Barriers and facilitators to health technology adoption by older adults with chronic diseases: an integrative systematic review. BMC Public Health 2024 Feb 16;24(1):506. [doi: 10.1186/s12889-024-18036-5] [Medline: 38365698]

XSL•FO
RenderX

73. Brawley LR, Rejeski WJ, King AC. Promoting physical activity for older adults: the challenges for changing behavior. Am J Prev Med 2003 Oct;25(3 Suppl 2):172-183. [doi: 10.1016/s0749-3797(03)00182-x] [Medline: 14552942]
74. Schreiweis B, Pobiruchin M, Strotbaum V, Suleder J, Wiesner M, Bergh B. Barriers and facilitators to the implementation of eHealth services: systematic literature analysis. J Med Internet Res 2019 Nov 22;21(11):e14197. [doi: 10.2196/14197] [Medline: 31755869]
75. Amagai S, Pila S, Kaat AJ, Nowinski CJ, Gershon RC. Challenges in participant engagement and retention using mobile health apps: literature review. J Med Internet Res 2022 Apr 26;24(4):e35120. [doi: 10.2196/35120] [Medline: 35471414]
76. Skivington K, Matthews L, Simpson SA, et al. Framework for the development and evaluation of complex interventions: gap analysis, workshop and consultation-informed update. Health Technol Assess 2021 Sep;25(57):1-132. [doi: 10.3310/hta25570] [Medline: 34590577]
77. Kracht CL, Burkart S, Groves CI, et al. 24-hour movement behavior adherence and associations with health outcomes: an umbrella review. J Act Sedentary Sleep Behav 2024;3(1):25. [doi: 10.1186/s44167-024-00064-6] [Medline: 39399355]
78. Ross R, Chaput JP, Giangregorio LM, et al. Canadian 24-hour movement guidelines for adults aged 18-64 years and adults aged 65 years or older: an integration of physical activity, sedentary behaviour, and sleep. Appl Physiol Nutr Metab 2020 Oct;45(10 (Suppl. 2):S57-S102. [doi: 10.1139/apnm-2020-0467] [Medline: 33054332]
79. Badr J, Motulsky A, Denis JL. Digital health technologies and inequalities: a scoping review of potential impacts and policy recommendations. Health Policy 2024 Aug;146:105122. [doi: 10.1016/j.healthpol.2024.105122] [Medline: 38986333]
80. Collins LM, Murphy SA, Strecher V. The multiphase optimization strategy (MOST) and the Sequential Multiple Assignment Randomized Trial (SMART): new methods for more potent eHealth interventions. Am J Prev Med 2007 May;32(5 Suppl):S112-S118. [doi: 10.1016/j.amepre.2007.01.022] [Medline: 17466815]
81. Buffart LM, Kalter J, Sweegers MG, et al. Effects and moderators of exercise on quality of life and physical function in patients with cancer: an individual patient data meta-analysis of 34 RCTs. Cancer Treat Rev 2017 Jan;52:91-104. [doi: 10.1016/j.ctrv.2016.11.010] [Medline: 28006694]

Abbreviations

DHI: digital health intervention
GRADE: Grading of Recommendations, Assessment, Development, and Evaluation
HKSJ: Hartung-Knapp-Sidik-Jonkman
LPA: light physical activity
MD: mean difference
MeSH: Medical Subject Headings
MVPA: moderate-to-vigorous physical activity
PA: physical activity
PI: prediction interval
PICOS: Participants, Interventions, Comparisons, Outcomes, Study Type
PRISMA : Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
RCT: randomized controlled trial
SB: sedentary behavior
SMD: standardized mean difference
TIAB : title and abstract
TPA: total physical activity
TS: topic
XB: exact bound

Edited by S Brini; submitted 13.Jan.2026; peer-reviewed by FJ Yang, S Lu; revised version received 04.Apr.2026; accepted 06.Apr.2026; published 29.May.2026.

Please cite as:
Yao J, Wang H, Jia S, Chen H, He J, Long J, Chen S
Effects of Multicomponent Digital Health Interventions on Multidimensional Physical Activity in Older Adults: Systematic Review, Meta-Analysis, and Meta-Regression of Randomized Controlled Trials
J Med Internet Res 2026;28:e91338
URL: https://www.jmir.org/2026/1/e91338
doi:10.2196/91338

XSL•FO
RenderX

© Jiayi Yao, Haozhe Wang, Shiguan Jia, Hao Chen, Junhao He, Juncheng Long, Shengxian Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effectiveness of eHealth Interventions in Alleviating Burden on Informal Caregivers of People With Dementia: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Min Liang[1,2], BSc; Jinxia Rong[2], BSc; Crina Grosan[3], PhD; Xiyi Wang[1,4], PhD

[1]School of Nursing, Shanghai Jiao Tong University, 1 Banxia Road, Pudong New Area District, Shanghai, China

[2]School of Nursing, Fudan University, Shanghai, China

[3]Division of Digital Health and Applied Technology Assessment, Florence Nightingale Faculty of Nursing, Midwifery & Palliative Care, King's College London, London, United Kingdom

[4]Department of Nursing, Renji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Corresponding Author:
Xiyi Wang, PhD
School of Nursing, Shanghai Jiao Tong University, 1 Banxia Road, Pudong New Area District, Shanghai, China

Abstract

Background: Informal caregivers of people with dementia frequently experience substantial psychological burden, including elevated stress and depressive symptoms. eHealth interventions have emerged as a scalable solution to support caregivers. However, their effectiveness and the influence of intervention characteristics remain unclear.

Objective: This study aimed to evaluate the effectiveness of eHealth interventions in reducing caregiver burden and depressive symptoms among informal caregivers of people with dementia, and to explore potential moderators of intervention effectiveness, including delivery modality, human support, and intervention duration.

Methods: This systematic review and meta-analysis was conducted in accordance with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Search extension) guidelines. Eight electronic databases and clinical trial registries were searched from inception to March 10, 2026. Eligible studies were randomized controlled trials involving informal caregivers of people with dementia receiving interactive eHealth interventions compared with usual care or inactive controls. Study selection, data extraction, and risk-of-bias assessment were conducted independently by 2 reviewers (ML and JXR). Meta-analyses were conducted using a random-effects model with the Hartung-Knapp-Sidik-Jonkman adjustment. Heterogeneity was assessed using the I^2 statistic, and 95% prediction intervals (PIs) were calculated for the primary analyses. The certainty of evidence was evaluated using GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) criteria.

Results: Thirty-five trials (N=3388) were included. eHealth interventions were associated with a statistically significant reduction in caregiver burden (k=35, standardized mean difference [SMD] −0.26, 95% CI −0.42 to −0.10; P=.002; τ^2=0.16, τ=0.40; 95% PI −1.10 to 0.58) and a modest reduction in depressive symptoms (k=23, SMD −0.27, 95% CI −0.53 to −0.01; P=.04; τ^2=0.31, τ=0.55; 95% PI −1.45 to 0.91). While average effects were statistically significant, wide PIs indicate substantial between-study heterogeneity and variability in real-world effectiveness. Subgroup analyses suggested that short-term (≤8 weeks) intervention was associated with borderline significantly stronger effects for caregiver burden, while human-supported and mobile-based interventions demonstrated larger point estimates with substantial heterogeneity. Meta-regression indicated that caregiver age was a potential moderator (P=.05) in univariate analysis, with decreasing effectiveness observed in older populations. Sensitivity analysis of 27 high-quality studies confirmed robust effects (SMD −0.31, P=.002). No significant small-study effects were detected (Egger test: P>.35). The overall certainty of evidence was moderate.

Conclusions: eHealth interventions provide modest but variable benefits in reducing burden and depressive symptoms among informal caregivers of people with dementia. This review is timely, given the rapidly expanding digital dementia care and extends prior evidence using a more conservative Hartung-Knapp-Sidik-Jonkman framework with PIs, showing that effects vary from minimal to substantial. While average effects support effectiveness, wide PIs and moderate certainty evidence indicate that outcomes are context-dependent and influenced by heterogeneity and methodological limitations. Hybrid digital-human models should be prioritized to enhance consistency and real-world impact.

Trial Registration: PROSPERO CRD42024625273; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024625273

(J Med Internet Res 2026;28:e78568) doi:10.2196/78568

XSL•FO
RenderX

KEYWORDS

caregiver burden; dementia; meta-analysis; systematic review; eHealth

Introduction

Rationale

Dementia is a progressive neurodegenerative disorder and a major contributor to disability and dependency among older adults worldwide, representing a major and growing global public health priority [1]. Recent meta-analyses have demonstrated the substantial burden experienced by dementia caregivers, particularly in relation to psychological distress and care demands [2]. In parallel, global epidemiological evidence indicates that the prevalence and societal burden of dementia are rapidly increasing, placing escalating pressure on health and long-term care systems worldwide [3]. Estimates from the China Alzheimer Report 2025 indicate that the prevalence has reached 13 million, representing a substantial and rapidly escalating public health and socioeconomic challenge, particularly in aging societies where long-term care systems remain underdeveloped [4]. Beyond the clinical burden of dementia itself, the associated caregiving demands represent a critical yet often underrecognized component of the overall disease burden. Importantly, the burden associated with informal caregiving has been increasingly recognized as a parallel public health challenge, with substantial implications for both caregiver health and health system sustainability [2]. Despite growing recognition, effective and scalable solutions to support caregivers remain limited.

Most people living with dementia receive care in home and community settings. In China, cultural norms emphasizing family responsibility, combined with the high cost and limited availability of institutional care, mean that informal caregivers, typically spouses or adult children, provide the majority of long-term care [5,6]. Recent large-scale randomized controlled trials suggest that existing dementia care models may have limited impact on key caregiver outcomes such as burden, despite improvements in care satisfaction [7]. These caregivers frequently manage progressive cognitive decline, neuropsychiatric symptoms, and functional deterioration, often without formal training or adequate support [8]. Informal caregivers are often described as the “second invisible patients” of the dementia care system [9]. Caregiver burden is a multidimensional construct encompassing psychological distress, physical strain, social restriction, and financial pressure. Elevated caregiver burden is associated with adverse outcomes including depressive symptoms, reduced quality of life, and increased health care utilization. In addition to objective caregiving demands, psychosocial factors, such as coping resources and access to support services, also influence caregiver outcomes [10,11].

eHealth is defined by the World Health Organization as the use of information and communication technologies for health and health-related fields, including health care services, health education, and research [12]. Building on this broad conceptualization, eHealth interventions, including web-based programs, mobile apps, and telehealth services, offer advantages in terms of accessibility, flexibility, and scalability compared with traditional face-to-face interventions [13-15]. Recent systematic reviews suggest that such interventions can improve caregiver outcomes, although the magnitude and consistency of effects vary across intervention types and delivery formats [16]. However, despite recognition of caregiver burden, effective and scalable intervention strategies remain insufficient.

Although recent reviews have examined the effectiveness of eHealth interventions for dementia caregivers, findings remain inconsistent. Reported effects on caregiver burden range from moderate reductions to nonsignificant outcomes [16-19]. These discrepancies may reflect heterogeneity in study populations, variation in intervention components, and the inclusion of nonrandomized designs. Furthermore, many meta-analyses have relied on conventional random-effects models, which may underestimate uncertainty when the number of included studies is limited. More conservative methods, such as the Hartung-Knapp-Sidik-Jonkman (HKSJ) approach, are recommended in such contexts to reduce the risk of inflated type I errors [20,21]. In addition, the relative contribution of specific intervention characteristics, such as delivery modality, intervention duration, and the inclusion of human-supported guidance, remains unclear. Elucidating these factors is essential to optimize the design and implementation of scalable digital caregiver support programs, particularly in the context of the rapid expansion of digital health services.

To address these gaps, this study focuses exclusively on randomized controlled trials (RCTs) and applies more conservative statistical approaches, including the HKSJ method, to provide more robust and reliable estimates. Furthermore, this study systematically examines key intervention characteristics, including delivery modality, intervention duration, and guidance level, to identify features associated with improved effectiveness. Compared with previous reviews that included heterogeneous study designs or did not explore intervention-level moderators, this study aims to provide a more rigorous and mechanism-oriented synthesis. By identifying the specific components that drive effectiveness, this work offers practical guidance for the design, implementation, and scaling of eHealth interventions for dementia caregivers.

Objectives

This review aimed to evaluate the effectiveness of eHealth interventions in reducing caregiver burden among informal caregivers of individuals with dementia. The secondary objective was to assess the effects of these interventions on caregivers’ depressive symptoms. An additional aim was to examine potential moderators of intervention effectiveness through subgroup analyses and meta-regression, including intervention delivery modality, intervention duration, and the presence of human-supported guidance. Understanding these factors may inform the development of more targeted and scalable digital interventions for dementia caregiver support.

XSL•FO
RenderX

Methods

Study Design and Registration

This systematic review and meta-analysis was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement [22], PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Search extension) recommendations [23], and the Cochrane Handbook for Systematic Reviews of Interventions. The study protocol was prospectively registered with PROSPERO (ID: CRD42024625273).

Information Sources

A comprehensive search was performed across 8 electronic databases: PubMed (via NCBI), Embase and Scopus (via Elsevier), Web of Science Core Collection (via Clarivate), CENTRAL (Cochrane Central Register of Controlled Trials), ProQuest Dissertations & Theses Global, and CINAHL and PsycINFO (via EBSCOhost). In addition, ClinicalTrials.gov was searched to identify ongoing or unpublished studies. Reference lists of included studies were manually screened to identify additional relevant studies. All databases were searched from inception to March 10, 2026. Reference lists of included studies and relevant systematic reviews were manually screened to identify additional eligible studies. No additional sources such as conference proceedings, organizational websites, or gray literature databases were searched, as the database strategy was considered sufficient to capture relevant peer-reviewed evidence. No contact was made with study authors, experts, or manufacturers for additional or unpublished data.

Search Strategy

The literature search was conducted and reported in accordance with the PRISMA-S extension to ensure transparency and reproducibility [23]. The search strategy combined controlled vocabulary terms (eg, MeSH [Medical Subject Headings] and Emtree) and free-text keywords related to dementia, caregivers, and eHealth or mHealth interventions. The strategy was developed de novo by the research team and iteratively refined. The search was limited to English-language studies due to resource constraints for translation and to RCTs to ensure methodological rigor. No date restrictions were applied. No validated external search filters were used; instead, study design restrictions were implemented through a combination of indexing terms and keywords. The full search strategies for all databases are provided in Table S1 in Multimedia Appendix 1. Any deviations from PRISMA-S checklist items due to methodological constraints are explicitly reported in the manuscript.

Eligibility Criteria

Eligibility criteria were determined according to the PICOS (Population, Intervention, Comparison, Outcomes, and Study Design) framework. (1) Population (P): studies were included if they involved informal caregivers (eg, family members, spouses, or friends) providing unpaid care for individuals with a clinical diagnosis of dementia. (2) Intervention (I): were defined as eHealth-related interventions (eg, web-based platforms, mobile apps, or telehealth programs) targeting informal caregivers of individuals with dementia. (3) Comparison (C): eligible comparators included usual care, waitlist controls, or active nondigital interventions (eg, printed educational materials). (4) Outcomes (O): the primary outcome was caregiver burden, assessed using validated standardized instruments such as the Zarit Burden Interview. To avoid selective reporting bias, all studies were included providing extractable data for burden, regardless of its status as a primary or secondary outcome in the original trial. If a study reported multiple burden scales, the Zarit Burden Interview was prioritized for pooling to enhance consistency. Secondary outcomes included depressive symptoms measured by validated scales such as the Center for Epidemiological Studies Depression Scale or equivalent, following the same inclusion as the primary outcome. (5) Study design (S): peer-reviewed RCTs published in English.

Studies were excluded if they (1) focused solely on formal or professional health care providers, (2) provided only passive web-based informational content without interactive components, or (3) were conference abstracts, editorials, or protocols without extractable outcome data.

Study Selection and Data Extraction

All retrieved records were imported into EndNote (version 21; Clarivate Analytics) for deduplication using both automated and manual processes. Title and abstract were screened independently by 2 reviewers (ML and JXR). Screening was conducted in a 2-stage process to manage the high volume of retrieved records (n=7449). In the first stage, 2 reviewers (ML and JXR) independently performed a preliminary triage based on titles to remove clearly irrelevant records (eg, animal studies and unrelated medical conditions). To ensure a highly sensitive and overinclusive process, records were excluded at this stage only if both reviewers independently deemed them unequivocally unrelated to the research objective. In the second stage, all remaining records underwent a formal title-and-abstract screening process, in which both title and abstract information was considered, conducted independently by 2 reviewers (ML and JXR) in accordance with Cochrane recommendations [24]. This stage aimed to ensure an overinclusive and sensitive selection process. Any record considered potentially relevant by either reviewer was retained for full-text assessment. Full-text papers of potentially eligible studies were then assessed for inclusion. Discrepancies were resolved through discussion or consultation with a third reviewer (XYW).

Data extraction was conducted independently by 2 researchers (ML and JXR) using a standardized extraction form. Disagreements were resolved through discussion with a third reviewer (XYW). Extracted data included (1) study characteristics (author, year, and country), (2) participant demographics (age, gender, and relationship to care recipient), (3) intervention characteristics (modality, duration, guidance, and theoretical framework), and (4) outcome data (sample size, means, and SDs). For studies with multiple intervention arms, the control group was proportionally divided to prevent double-counting.

XSL•FO
RenderX

Study Risk-of-Bias Assessment

The risk of bias for each included RCT was assessed using the Cochrane Risk of Bias 2 (RoB 2) tool. Five domains were evaluated: bias arising from the randomization process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported result. Two reviewers (ML and JXR) independently assigned a rating of "Low risk," "Some concerns," or "High risk" to each domain. Two reviewers (ML and JXR) conducted assessments independently, with disagreements resolved through consensus.

Statistical Analysis and Meta-Analysis

Effect Size Calculation

Meta-analyses were conducted using a random-effects framework with inverse-variance weighting. Effect sizes were expressed as Hedges *g* (bias-corrected standardized mean difference [SMD]) to account for potential bias in small sample sizes and to enable pooling across studies that used different assessment scales for caregiver outcomes. All analyses were performed using R (meta package).

Heterogeneity and Random-Effects Model

Pooled estimates were calculated using the HKSJ method, which produces more conservative confidence intervals and reduces the risk of inflated type I error when the number of included studies is small [20]. The Hartung-Knapp adjustment was applied and explicitly labeled in the forest plots. The Sidik-Jonkman method was used to estimate between-study variance (τ^2). Statistical heterogeneity was assessed using the I^2 statistic and the Cochran Q test. Furthermore, 95% prediction intervals (PIs) were calculated for the primary meta-analysis to reflect the expected range of true effects in future settings. PIs were not calculated for subgroup analyses due to methodological limitations [25].

Subgroup and Meta-Regression Analyses

Prespecified subgroup analyses were conducted to explore sources of heterogeneity, focusing on caregiver dominant type, intervention modality (Web-based, App/Instant Messaging (IM) (Mobile), or Video/Tele (Videoconferencing/[VC]), interaction type (Asynchronous, Synchronous), intervention guidance category (Self-guided vs Human-supported), comparator type, outcome hierarchy, monitoring, and intervention duration. Meta-regression was performed to examine the influence of continuous moderators, including mean age, proportion of female participants, and study attrition rate.

Small-Study Effects and Sensitivity Analysis

Potential small-study effects were evaluated using a contour-enhanced funnel plot and Egger linear regression test [26]. Sensitivity analyses were conducted by sequentially excluding studies assessed as having a high risk of bias to evaluate the robustness of the pooled estimates.

Certainty of Evidence Assessment

The overall certainty of evidence for each outcome was evaluated using the GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) approach, considering risk of bias, inconsistency, indirectness, imprecision, and publication bias. The certainty of evidence was categorized as high, moderate, low, or very low.

Synthesis Methods and Visualization

Study characteristics were summarized in tabular form (Table 1). Forest plots were generated to present individual study effect sizes alongside pooled estimates, including SMDs, 95% CIs, and 95% PIs.

XSL•FO
RenderX

Table . Characteristics of included randomized controlled trials evaluating eHealth interventions for informal caregivers of people with dementia (abridged edition; *k*=35).

Study	Sample size (I/C)	Delivery/Duration	Intervention component	Outcome
Beauchamp et al [27], 2005 (United States)	150/149	Web/Asynch[a]/4w	Self-Guided: Multicomponent (Education + CBT[b] skills)	Strain[c] (CSI[d]) Dep[e] (CES-D[f])
Gitlin et al [28], 2010 (United States)	117/122	VC[g]/Synch/[h]24w	Human-Supported: Caregiver Training (ACT[i] model)	Burden (ZBI[j]) Dep (CES-D)
Kwok et al [29], 2013 (China)	18/20	Mobile/Synch/12w	Self-Guided: Psychoeducation	Burden (ZBI)
Torkamani et al [30], 2014 (United Kingdom, Spain, and Greece)	27/30	Web/Asynch/24w	Self-Guided: Monitoring & Alerts	Burden (ZBI)
Kales et al [31], 2018 (United States)	26/30	Mobile/Asynch/4w	Human-Supported: Symptom Management (Tool-based)	Dep (CES-D)[c]; Burden (ZBI)
Meichsner et al [32], 2019 (Germany)	15/15	Web/Asynch/8w	Human-Supported: Online CBT Modules	Burden (VAS[k])[c]; Dep (CES-D)[c]
Metcalfe et al [33], 2019 (United Kingdom, France, and Germany)	29/29	Web/Asynch/6w	Self-Guided: RHAPSODY Program	Burden (BSFC-10)[l] Dep (PSS-10[m])
Williams et al [34], 2019 (United States)	43/41	VC/Hybrid/12w	Human-Supported: Video-based Feedback	Burden (ZBI)[c]; Dep (CES-D)[c]
James et al [35], 2021 (United States)	5/5	Mobile/Asynch/2w	Self-Guided: Mind–Body + HRV[n] Monitoring	Burden (ZBI)
Baruah et al [36], 2021 (India)	29/26	Web/Asynch/12w	Self-Guided: iSupport (Multi-module)	Burden (ZBI)[c]; Dep (CES-D)[c]
Fossey et al [37], 2021 (United Kingdom)	53/101/54	Web/Asynch/26w	Human-Supported: CBT-based Program	Dep (HADS[o]); Burden (RSS[p])
Teles et al [38], 2022 (Portugal)	11/20	Web/Asynch/12w	Self-Guided: Online Training (iSupport)	Burden (ZBI)[c]; Dep (CES-D)
Hepburn et al [39], 2022 (United States)	96/165	VC/Synch/7w	Human-Supported: Psychoeducation + Coaching	Burden (ZBI)[c]; Dep (HADS-D)
Bodenstein [40], 2022 (Canada)	16/14	VC/Synch/8w	Human-Supported: Mindful Yoga	Burden (ZBI) Dep (PHQ-9[q])
Han et al [41], 2023 (United States)	9/10	VC/Synch/8w	Human-Supported: ACT + Behavioral Activation	Burden (ZBI) Dep (DASS-21[r])
Rhodus et al [42], 2023 (United States)	9/9/7	VC/Synch/6w	Human-Supported: Sensory-based Intervention	Burden (ZBI)
Rodriguez et al [43], 2023 (United States)	23/22	Mobile/Asynch/24w	Human-Supported: Digital Coaching + Monitoring	Burden (NPI-Q[s])
Hu et al [44], 2024 (China)	30/30	Mobile/Synch/8w	Human-Supported: MBSR[t]	Burden (ZBI)[c]; Dep (DASS-21)
Salehinejad et al [45], 2024 (Iran)	24/23	Web/Asynch/8w	Self-Guided: Education + Behavioral Management	Burden (ZBI)
Xie et al [46], 2024 (China)	33/33	Web/Asynch/24w	Self-Guided: Caregiver Training (Nurse-led)	Burden (ZBI)
Jain [47], 2024 (India)	8/8	VC/Synch/3w	Human-Supported: CBT+ Mindfulness	Burden (ZBI)[c]
Song et al [48], 2024 (United States)	13/14	VC/Synch/5w	Human-Supported: CBT-I + Behavioral Activation	Burden (ZBI)

XSL•FO
RenderX

Study	Sample size (I/C)	Delivery/Duration	Intervention component	Outcome
Yuan et al [49], 2025 (China)	51/55	Mobile/Synch/20w	Human-Supported: iSupport + Follow-up	Burden (ZBI)
Nguyen pilot et al [50], 2025 (Vietnam)	27/30	Mobile/Asynch/7w	Human-Supported: Psychoeducation + Discussion	Burden (ZBI); Dep (DASS-21)
Nguyen et al [51], 2025 (Vietnam)	80/81	Mobile/Asynch/7w	Human-Supported: Video Psychoeducation + Messaging	Dep (DASS-21)[c]; Burden (ZBI)
Han et al [52], 2025 (United States)	16/17	VC/Synch/10w	Self-Guided: ACT-based	Dep (PHQ-9)[c]; Burden (ZBI)
Nichols et al [53], 2025 (United States)	56/54	Mobile/Synch/12w	Human-Supported: Problem-Solving + Coping	Burden (ZBI)[c]; Dep (PHQ-9)
Stevens et al [54], 2025 (United States)	79/88	Web/Asynch/24w	Human-Supported: Psychoeducation + Skills + Network	Burden (ZBI)[c]; Dep (CES-D)
Pfaff et al [55], 2025 (Germany)	83/87	VC/Hybrid/24w	Human-Supported: Needs Assessment + Care Planning	Burden (ZBI)
Windle et al [56], 2025 (United Kingdom)	175/177	Web/Asynch/24w	Self-Guided: iSupport (5 modules)	Burden (ZBI)[c]; Dep (CES-D)[c]
Sun et al [57], 2026 (China)	98/103	Mobile/Asynch/12w	Self-Guided: Personalized Care + Consultation	Burden (ZBI)[c]; Dep (SDS[u])
Yuan et al [58], 2026 (Singapore)	27/26	Mobile/Asynch/4w	Self-Guided: Knowledge Base + Peer Support	Dep (CES-D)[c]; Burden (ZBI)
Brijnath et al [59], 2026 (Australia)	47/46	Web/Asynch/12w	Self-Guided: WHO[v] iSupport Lite (Multimedia)	Burden (ZBI)[c]; Dep (CES-D)
Cheng and Ng [60] 2026 (43 countries)	132/142	Web/Synch/12w	Self-Guided: Multi-app + AI[w] Support	Dep (PHQ-9)[c]; Burden (ZBI)
Durepos et al [61], 2026 (Canada)	14/15	VC/Synch/16 - 24w	Human-Supported: ACT	Burden (ZBI); Dep (DASS-21)

[a] Asynch: asynchronous.
[b]CBT: cognitive behavioral therapy.
[c]Primary outcome as defined in the original study.
[d]CSI: caregiver Strain Index.
[e]Dep: depression.
[f]CES-D: Center for Epidemiologic Studies Depression Scale.
[g]VC: videoconferencing.
[h]Synch: synchronous.
[i]ACT: advancing caregiver training.
[j]ZBI: Zarit Burden Interview.
[k]VAS: Visual Analogue Scale.
[l]BSFC-10: Burden Scale for Family Caregivers.
[m]PSS-10: Perceived Stress Scale.
[n]HRV: heart rate variability.
[o]HADS: Hospital Anxiety and Depression Scale.
[p]RSS: Relative Stress Scale.
[q]PHQ-9: Patient Health Questionnaire-9.
[r]DASS-21: Depression, Anxiety and Stress Scale.
[s]NPI-Q: Neuropsychiatric Inventory Questionnaire (Caregiver Distress subscale).
[t]MBSR: mindfulness-based stress reduction.
[u]SDS: Zung Self-Rating Depression Scale.
[v]WHO: World Health Organization.
[w]AI: artificial intelligence.

XSL•FO
RenderX

Result

Study Selection and Flow Diagram

The systematic search across 8 electronic databases and registers identified 7444 records. After removing duplicates and records marked as ineligible by automation tools, 2846 records remained for screening. Following title and abstract screening, 113 reports were sought for retrieval, of which 7 could not be retrieved. A total of 106 reports from databases and registers were assessed for eligibility. In addition, 5 reports were identified through citation searching and assessed separately for eligibility. After excluding 75 reports from database and register searches and 1 report identified through citation searching, 35 RCTs were included in the review. The study selection process is detailed in the PRISMA 2020 flow diagram (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram illustrating the systematic selection process of RCTs evaluating eHealth interventions for informal caregivers of people with dementia across 8 electronic databases. RCT: randomized controlled trial.

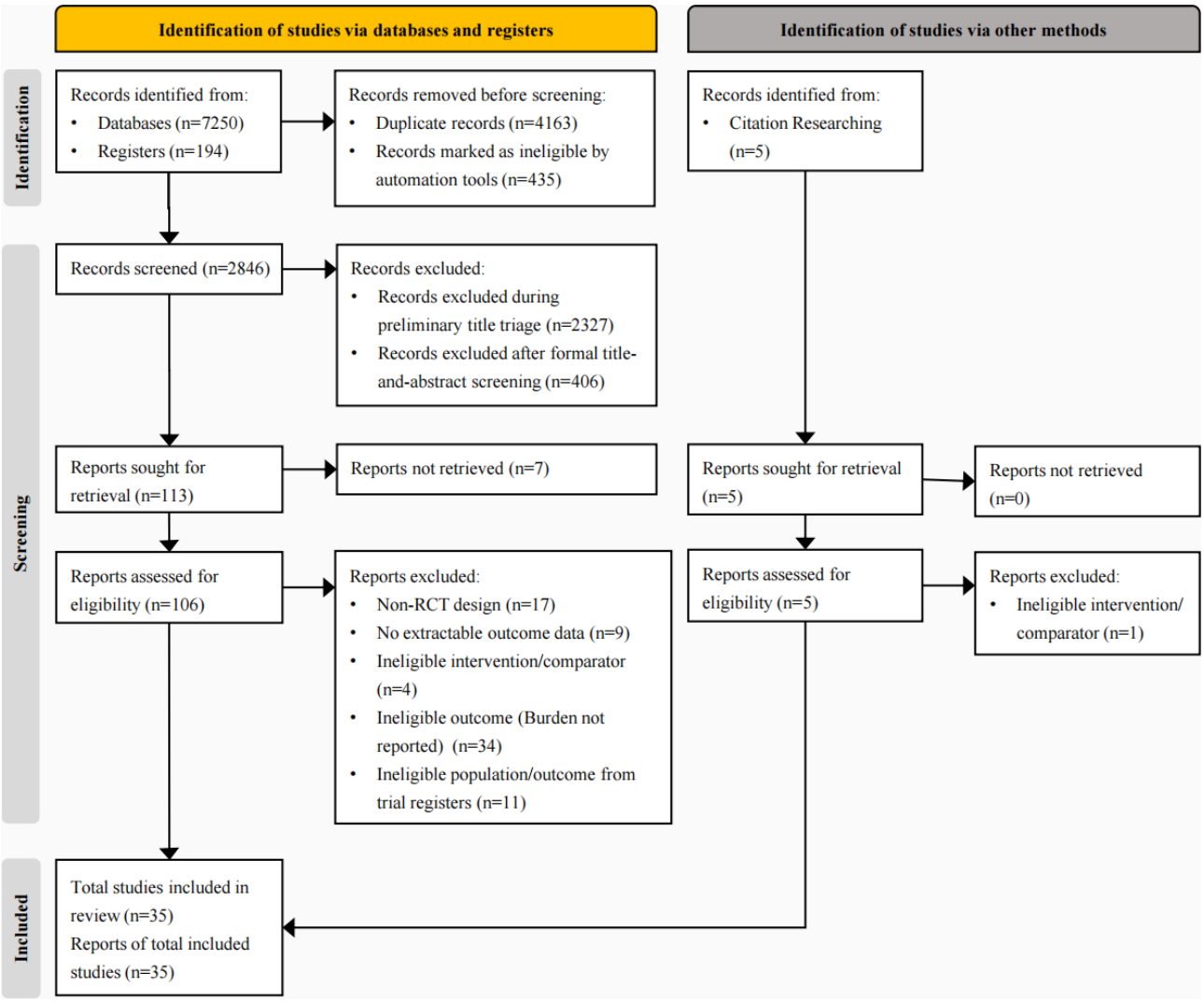

Study Characteristics

The 35 included RCTs, published between 2005 and 2026, were conducted across North America, Asia, Europe, and Oceania [27-61] (Table 1; details in Table S2 in Multimedia Appendix 1 [27-61]). The trials collectively enrolled 3388 informal caregivers (intervention: n=1700; control: n=1688). Caregivers were predominantly female (mean proportion 73.06%, SD 14.73%) and were generally middle-aged to older adults, with reported mean or median ages ranging from approximately 30 to 73 years. Most participants were spouses or adult children of individuals with dementia. Interventions varied in format and duration. Early studies primarily used web-based platforms, whereas more recent trials increasingly adopted mobile or hybrid digital delivery models. Intervention duration ranged from 2 to 26 weeks. Approximately 85.71% (30/35) of interventions were explicitly grounded in established theoretical frameworks, such as the Stress and Coping Model or the World Health Organization's iSupport program. Detailed study characteristics are presented in Table 1 and Table S2 in Multimedia Appendix 1.

Risk-of-Bias Assessment

Risk of bias was assessed using the RoB 2 tool (Figure 2 [27-61]). Three (8.6%) studies were rated as low risk, 24 (68.6%) studies had some concerns, and 8 (22.9%) studies were judged to have high risk of bias. The most common sources of bias were related to deviations from intended interventions and

XSL•FO
RenderX

outcome measurement. These issues primarily reflected the inherent difficulty of blinding participants in behavioral and digital interventions and the reliance on self-reported outcome measures. Most studies showed low risk of bias for randomization and selective outcome reporting.

Figure 2. Risk-of-bias assessment for 35 randomized controlled trials using the RoB 2 tool, evaluating internal validity across 5 domains for eHealth interventions targeting dementia caregiver burden and depression [27-61].

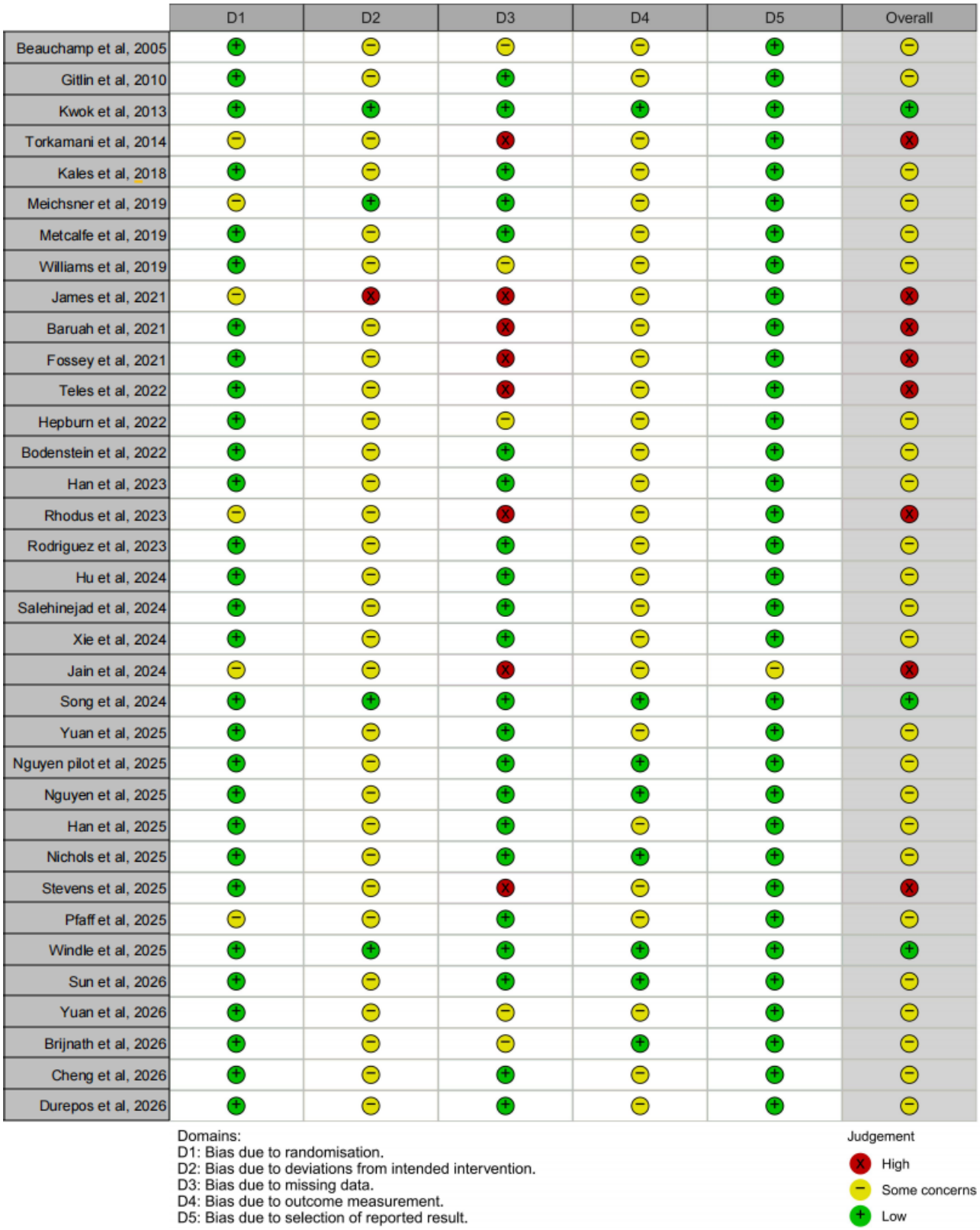

XSL•FO
RenderX

Synthesis of Primary Outcome: Caregiver Burden

Overall Effect and PI

Thirty-five studies involving 3388 participants (intervention: n=1700; control: n=1688) were included in the meta-analysis. Using a random-effects model with the HKSJ adjustment, eHealth interventions were associated with a statistically significant reduction in caregiver burden compared with control conditions (SMD −0.26, 95% CI −0.42 to −0.10; t_{34}=−3.32; P=.002; τ^2=0.16, τ=0.40; 95% PI −1.10 to 0.58; Figure 3 [27-61]). Substantial heterogeneity was observed across studies (I^2 73.6%, 95% CI 63.2%-81.0%; Q=128.61; df=34; P<.001). The 95% PI indicates considerable variability in the potential effects of eHealth interventions across different contexts and implementations.

XSL•FO
RenderX

Figure 3. Forest plot of the effect of eHealth interventions on caregiver burden in informal caregivers of people with dementia (*k*=35), using the Hartung-Knapp-Sidik-Jonkman random-effects model. Red lines represent the 95% prediction intervals [27-61], SMD: standardized mean difference.

XSL•FO
RenderX

Subgroup Analyses for Caregiver Burden

Prespecified subgroup analyses were conducted to explore the potential sources of heterogeneity (Table 2 and Figure S1 in Multimedia Appendix 1 [27-61]).

Table . Subgroup analyses of eHealth interventions on caregiver burden (k=35)[a].

Variables and levels	K	SMD[b]	95% CI	I^2	P_{within}
Caregiver dominant type					
Adult-child-dominant	15	−0.284	−0.583 to 0.015	84.20%	.06
Spousal-dominant	10	−0.183	−0.295 to −0.072	0.00%	*.005*
Mixed	10	−0.305	−0.720 to 0.110	74.00%	.13
Modality					
Web-based	13	−0.183	−0.410 to 0.043	62.60%	.10
Video/Tele (VC)	11	−0.086	−0.240 to 0.069	0.00%	*.24*
App/IM (Mobile)	11	−0.505	−0.914 to −0.097	85.20%	*.02*
Interaction					
Asynchronous	19	−0.292	−0.544 to −0.040	82.20%	*.03*
Synchronous	16	−0.224	−0.429 to −0.018	45.50%	*.04*
Guidance category					
Self-guided	15	−0.143	-0.348 to 0.062	57.00%	.16
Human-supported	20	−0.354	−0.596 to −0.112	78.30%	*.006*
Comparator					
Inert control	15	−0.178	−0.307 to −0.050	5.80%	*.01*
Enhanced control	20	−0.322	−0.592 to −0.053	83.00%	*.02*
Duration					
Short-term	16	−0.456	−0.782 to −0.131	82.10%	*.009*
Long-term	19	−0.13	−0.254 to −0.006	35.10%	*.04*
Outcome hierarchy					
Primary	19	−0.215	−0.428 to −0.002	63.50%	*.049*
Secondary	16	−0.316	−0.582 to −0.049	78.10%	*.02*
Mentoring					
Passive tracking	12	−0.224	−0.507 to 0.060	84.00%	.11
Active monitoring	23	−0.286	−0.498 to −0.075	62.80%	*.01*

[a]Italicized values indicate statistically significant results (P<.05).

[b]SMD: standardized mean difference.

Significant and Borderline Moderators

Intervention duration was identified as a significant moderator of efficacy (Q=3.98, P=.05). Short-term interventions (≤8 weeks) demonstrated a significant and more pronounced reduction in caregiver burden (SMD −0.456, 95% CI −0.782 to −0.131; P=.009; τ^2=0.276, τ=0.526), whereas long-term interventions showed a smaller, though still significant, effect (SMD −0.130, 95% CI −0.254 to −0.006; P=.04; τ^2=0.044, τ=0.209). Furthermore, Delivery Modality showed a strong moderating trend (Q=4.71, P=.10). Mobile-based interventions (Apps/IM) yielded the largest pooled effect size (SMD −0.505, 95% CI −0.914 to −0.097; P=.02; τ^2=0.303, τ=0.551), followed by Web-based platforms (SMD −0.183, 95% CI −0.410 to 0.043; P=.10; τ^2=0.106, τ=0.326), while Video/Teleconsultation (VC) showed the smallest and nonsignificant effect (SMD −0.086, 95% CI −0.240 to 0.069; P=.24; τ^2=0.037, τ=0.193). Notably, VC-based subgroups exhibited zero within-group heterogeneity (I^2=0%), suggesting a highly consistent, albeit modest, effect across diverse settings.

XSL•FO
RenderX

Impact of Guidance and Monitoring Strategies

Regarding guidance and monitoring, neither Guidance Category (P=.16) nor Clinical Monitoring (P=.70) reached statistical significance for between-group differences; however, clear patterns of within-group efficacy emerged. Interventions incorporating Human-Supported components (SMD −0.354, 95% CI −0.596 to −0.112; P=.006; τ^2=0.203, τ=0.450) and Active Monitoring (SMD −0.286, 95% CI −0.498 to −0.075; P=.01; τ^2=0.172, τ=0.415) both demonstrated significant burden reduction, whereas Self-guided programs (SMD −0.143, 95% CI −0.348 to 0.062; P=.16; τ^2=0.097, τ=0.311) and Passive Tracking (SMD −0.224, 95% CI −0.507 to 0.060; P=.11; τ^2=0.162, τ=0.402) failed to achieve statistical significance. Similarly, both forms of Interaction yielded significant results, with Asynchronous delivery (SMD −0.292, 95% CI −0.544 to −0.040; P=.03; τ^2=0.217, τ=0.466) remaining more robustly supported than Synchronous delivery (SMD −0.224, 95% CI −0.429 to −0.018; P=.04; τ^2=0.105, τ=0.324). Analysis of Comparator type revealed that eHealth interventions were effective regardless of whether they were compared against Inert Controls (SMD −0.178, 95% CI −0.307 to −0.050; P=.01; I^2=5.8%; τ^2=0.042, τ=0.204) or Enhanced Controls (SMD −0.322, 95% CI −0.592 to −0.053; P=.02; τ^2=0.258, τ=0.508). The remarkably low heterogeneity in the Inert Control subgroup (I^2=5.8%).

Caregiver and Outcome Characteristics

No significant differences were observed across Caregiver Dominant Types (P=.67) or Outcome Hierarchy (P=.53). Analysis of caregiver profiles revealed that interventions targeting Spousal-dominant groups yielded a remarkably consistent and significant effect (k=10, SMD −0.183, 95% CI −0.295 to −0.072; P=.005; τ^2=0.013, τ=0.113) with zero heterogeneity (I^2=0%), whereas Adult-child (SMD −0.284, 95% CI −0.583 to 0.015; P=.06; τ^2=0.237, τ=0.487) and Mixed (SMD −0.305, 95% CI −0.720 to 0.110; P=.13; τ^2=0.248, τ=0.498) groups exhibited high variability (I^2>74%). Finally, eHealth interventions were effective regardless of whether burden was assessed as a Primary (SMD −0.215, 95% CI −0.428 to −0.002; P=.049; τ^2=0.153, τ=0.391) or Secondary (SMD −0.316, 95% CI −0.582 to −0.049; P=.02; τ^2=0.182, τ=0.427) outcome.

Synthesis of Secondary Outcome: Depressive Symptoms

Overall Effect and PI

Twenty-three studies involving 2467 participants (k=23; intervention: n=1261; control: n=1206) reported depressive symptoms among caregivers. Meta-analysis using a random-effects model with the HKSJ adjustment indicated a statistically significant reduction in depressive symptoms associated with eHealth interventions (SMD −0.27, 95% CI −0.53 to −0.01; P=.04; τ^2=0.31, τ=0.55; 95% PI −1.45 to 0.91; Figure 4 [27,31-34,36-41,44,50-54,56-61]). Substantial heterogeneity was observed across studies (I^2=85.7%, 95% CI 79.8%-89.9%; Q=154.07, df=22; P<.001).

XSL•FO
RenderX

Figure 4. Forest plot of depressive symptoms, showing Hartung-Knapp-Sidik-Jonkman–adjusted random-effects estimates and 95% prediction intervals [27,31-34,36-41,44,50-54,56-61]. SMD: standardized mean difference.

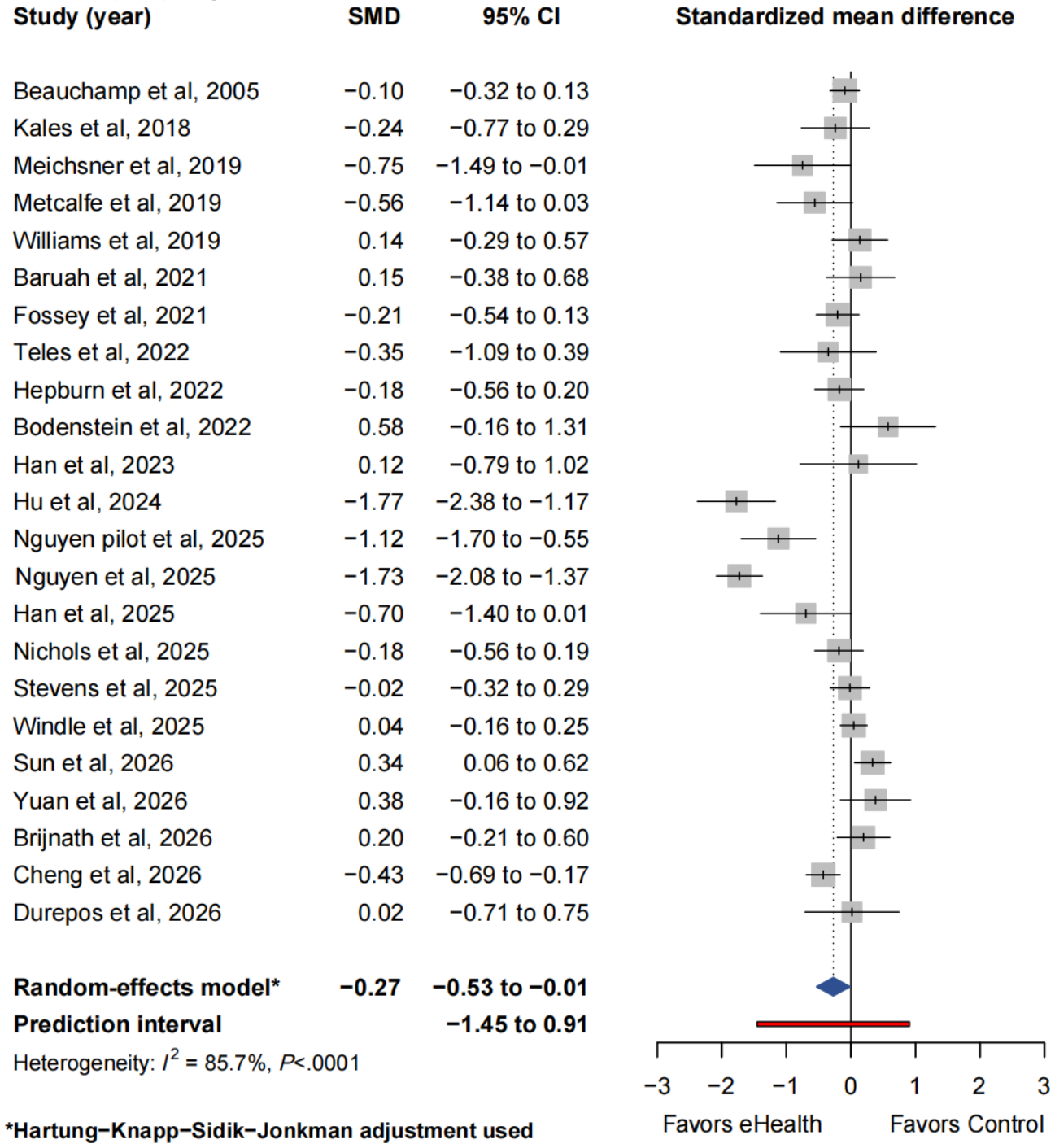

Subgroup Analyses for Depressive Symptoms

Prespecified subgroup analyses were conducted to investigate the sources of substantial heterogeneity observed in the secondary outcome (Table 3 and Figure S2 in Multimedia Appendix 1 [27,31-34,36-41,44,50-54,56-61]).

XSL•FO
RenderX

Table . Subgroup analyses of eHealth interventions on depressive symptoms (k=23).

Variables and levels	K	SMD[a]	95% CI	I^2	P_{within}
Caregiver dominant type					
Adult-child-dominant	12	−0.214	−0.623 to 0.194	90.40%	.27
Spousal-dominant	6	−0.234	−0.532 to 0.063	16.10%	.10
Mixed	5	−0.42	−1.499 to 0.658	88.00%	.34
Modality					
Web-based	10	−0.15	−0.347 to 0.047	47.30%	.12
App/IM (Mobile)	7	−0.611	−1.461 to 0.239	94.80%	.13
Video/Tele (VC)	6	−0.019	−0.433 to 0.394	32.00%	.91
Interaction					
Asynchronous	14	−0.267	−0.612 to 0.078	88.70%	.12
Synchronous	9	−0.281	−0.786 to 0.224	78.30%	.24
Guidance category					
Self-guided	10	−0.064	−0.322 to 0.193	67.20%	.59
Human-supported	13	−0.423	−0.859 to 0.013	88.70%	.06
Comparator					
Inert control	10	−0.071	−0.359 to 0.218	69.30%	.59
Enhanced control	13	−0.432	−0.852 to −0.012	89.70%	*.045* [b]
Duration					
Short-term	11	−0.504	−1.034 to 0.027	90.40%	.06
Long-term	12	−0.052	−0.229 to 0.125	54.40%	.53
Outcome hierarchy					
Secondary	15	−0.225	−0.536 to 0.087	76.90%	.14
Primary	8	−0.357	−0.945 to 0.230	92.10%	.19
Monitoring					
Passive tracking	11	−0.187	−0.591 to 0.217	90.50%	.33
Active monitoring	12	−0.361	−0.757 to 0.035	75.80%	.07

[a]SMD: standardized mean difference.

[b]Italicized value indicates statistically significant result (P<.05).

Key Drivers and Borderline Moderators

Although no single variable reached the traditional threshold for subgroup differences, Intervention Duration emerged as a key moderator with a strong trend toward significance (P=.07). Short-term interventions (≤8 weeks) demonstrated a larger potential effect size (SMD −0.504, 95% CI −1.034 to 0.027; P=.06; τ^2=0.541, τ=0.735) than Long-term programs (SMD −0.052, 95% CI −0.229 to 0.125; P=.53; τ^2=0.051, τ=0.225). Notably, the Comparator Type and Guidance Category also exhibited meaningful patterns. Interventions showed significant within-group efficacy when compared against Enhanced Controls (SMD −0.432, 95% CI −0.852 to −0.012; P=.045; τ^2=0.412, τ=0.642), whereas comparisons with Inert Controls were nonsignificant (SMD −0.071, 95% CI −0.359 to 0.218; P=.59; τ^2=0.126, τ=0.355). Similarly, Human-Supported guidance (SMD −0.423, 95% CI −0.859 to 0.013; P=.06; τ^2=0.448, τ=0.669) showed a stronger trend toward within-group efficacy than Self-guided programs (SMD −0.064, 95% CI −0.322 to 0.193; P=.59; τ^2=0.096, τ=0.310).

Platform and Technical Implementation

Analysis of Modality (P=.30) and Interaction Type (P=.96) confirmed that the digital delivery format did not significantly alter intervention outcomes. While Mobile-based apps/IM (SMD −0.611, 95% CI −1.461 to 0.239; P=.13; τ^2=0.781, τ=0.884) showed the largest point estimate among platforms, high within-group heterogeneity (I^2=94.8%) resulted in wide CIs (Web-based: SMD −0.150, 95% CI −0.347 to 0.047; P=.12; τ^2=0.054, τ=0.232; Video/Tele (VC): SMD −0.019, 95% CI −0.433 to 0.394; P=.91; τ^2=0.095, τ=0.308). Furthermore, Active Monitoring (eg, real-time tracking with feedback) showed a

XSL•FO
RenderX

stronger trend toward reduction (SMD −0.361, 95% CI −0.757 to 0.035; P=.07; τ^2=0.314, τ=0.560) than Passive Tracking (SMD −0.187, 95% CI −0.591 to 0.217; P=.33; τ^2=0.316, τ=0.563).

Caregiver Characteristics and Generalizability

The efficacy of eHealth interventions appeared remarkably consistent across different caregiver demographics. No significant differences were found for Caregiver Dominant Type (P=.89), with Adult-child (SMD −0.214, 95% CI −0.623 to 0.194; P=.27; τ^2=0.350, τ=0.592), Spousal (SMD −0.234, 95% CI −0.532 to 0.063; P=.10; τ^2=0.048, τ=0.218), and Mixed (SMD −0.420, 95% CI −1.499 to 0.658; P=.34; τ^2=0.667, τ=0.817) caregiver groups all showing similar directions of effect. Of note, the Spousal-dominant subgroup exhibited the highest level of evidence consistency (I^2=16.1%). Similarly, no meaningful differences were observed across interaction formats, with comparable effects for Asynchronous (SMD −0.267, 95% CI −0.612 to 0.078; P=.12; τ^2=0.304, τ=0.551) and Synchronous interventions (SMD −0.281, 95% CI −0.786 to 0.224; P=.24; τ^2=0.356, τ=0.596). Finally, the impact remained stable regardless of whether depressive symptoms were evaluated as a Primary (SMD −0.357, 95% CI −0.945 to 0.230; P=.19; τ^2=0.430, τ=0.656) or Secondary (SMD −0.225, 95% CI −0.536 to 0.087; P=.14; τ^2=0.260, τ=0.510) outcome (P=.65).

Exploration of Heterogeneity: Meta-Regression

Univariate and multivariate meta-regression analyses were conducted examining caregiver age, proportion of female participants, and attrition rates (Table 4).

For the primary outcome (Caregiver Burden; k=35), caregiver age emerged as a significant moderator (β=.02, 95% CI 0.0008 to 0.0392; Z=2.044, P=.049; Figure 5). In contrast, neither the proportion of female participants (P=.44) nor the study attrition rate (P=.36) significantly influenced the treatment effect. In the multivariate model, age showed borderline significance (β=.0209, 95% CI −0.0009 to 0.0427; P=.07), and the intercept remained significant (β=−1.542, 95% CI −2.7495 to −0.3347; P=.02).

Table . Meta-regression analysis of potential moderators for eHealth intervention efficacy.

Moderator	Coefficient, β	SE	95% CI	Z value	P value
Caregiver burden (k[a]=35)					
Age (years), mean	0.0200	0.0098	0.0008 to 0.0392	2.044	*.049* [b]
Female, %	0.0046	0.0059	−0.0070 to 0.0162	0.788	.44
Attrition, %	0.0043	0.0047	−0.0049 to 0.0135	0.922	.36
Multivariate model					
Intercept	−1.5421	0.616	−2.7495 to −0.3347	−2.504	.02
Age (years)	0.0209	0.0111	−0.0009 to 0.0427	1.877	.07
Female, %	−0.0001	0.0064	−0.0126 to 0.0124	−0.023	.98
Attrition rate	0.0050	0.0045	−0.0038 to 0.0138	1.100	.28
Depressive symptoms (k=23)					
Age (years), mean	0.0162	0.0184	−0.0199 to 0.0523	0.877	.39
Female, %	0.0023	0.0104	−0.0181 to 0.0227	0.222	.83
Attrition, %	0.0087	0.0068	−0.0046 to 0.0220	1.276	.22
Multivariate model					
Intercept	−1.4264	1.1837	−3.7464 to 0.8936	−1.205	.24
Age (years)	0.0149	0.0199	−0.0241 to 0.0539	0.748	.46
Female, %	0.0018	0.0112	−0.0201 to 0.0237	0.161	.87
Attrition rate	0.0089	0.0072	−0.0052 to 0.0230	1.237	.23

[a]k: number of studies.

[b]Italicized value indicates statistically significant result (P<.05).

XSL•FO
RenderX

Figure 5. Meta-regression bubble plot of mean age of caregivers on the intervention effect (standardized mean difference). The red line indicates the fitted meta-regression trend line.

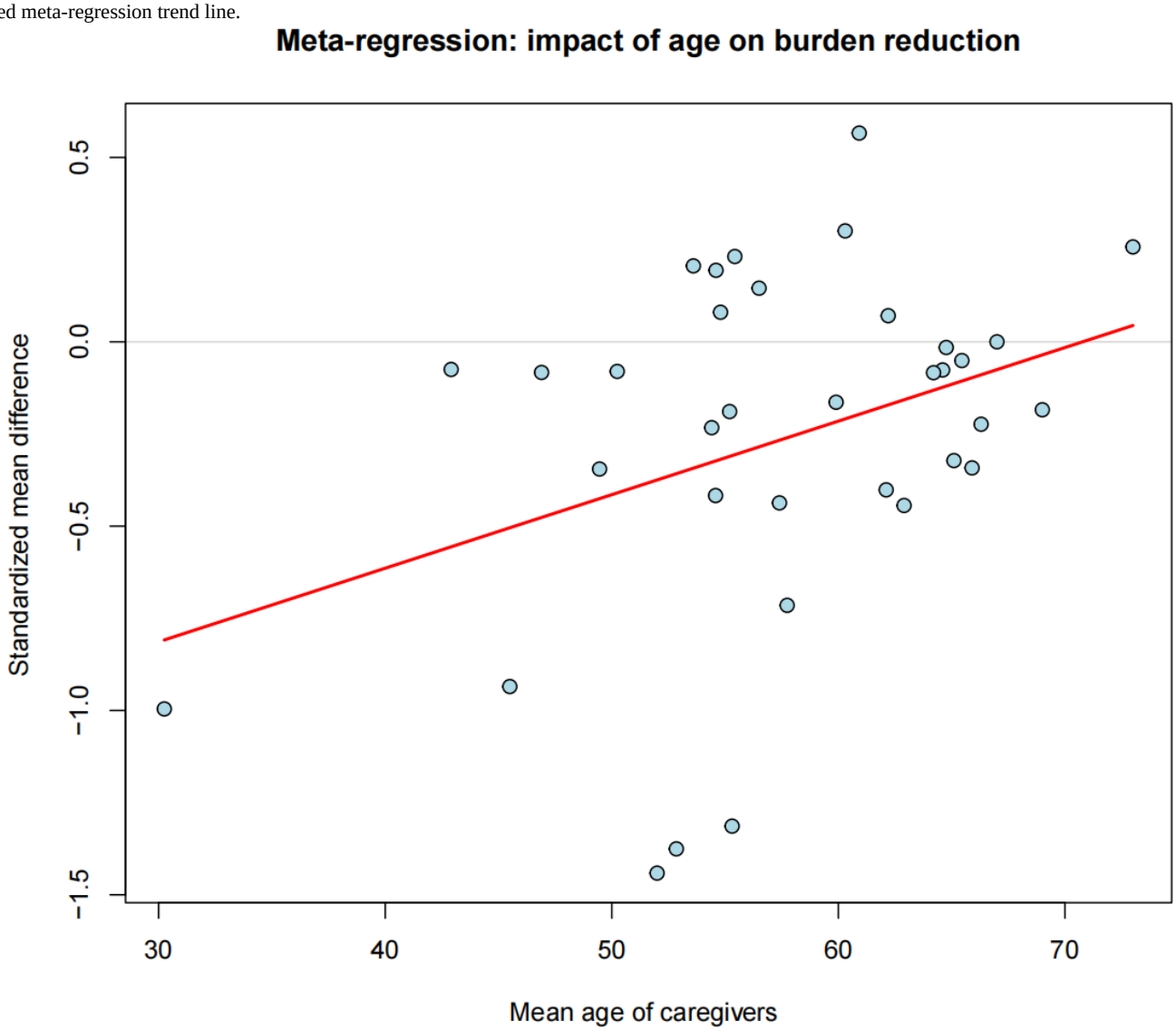

For the secondary outcome (Depressive Symptoms) (k=23), the meta-regression results were consistent with the burden analysis in terms of direction but did not reach statistical significance. The intervention effect on depression was not significantly moderated by caregiver age (β=.0162, 95% CI −0.0199 to 0.0523; P=.39), female percentage (P=.83), or attrition rate (P=.22). The multivariate model confirmed that these factors collectively did not explain the substantial residual heterogeneity.

Small-Study Effects and Publication Bias

Egger tests indicated no significant funnel plot asymmetry for caregiver burden (t_{33}=−0.808; P=.42) or depressive symptoms (t_{21}=−0.954; P=.35). Contour-enhanced funnel plots (Figure 6A for caregiver burden and Figure 6B for depressive symptoms) confirmed symmetric distributions, suggesting that observed heterogeneity is attributable to study characteristics rather than publication bias.

XSL•FO
RenderX

Figure 6. Contour-enhanced funnel plot for assessment of publication bias. (A) primary outcome (caregiver burden) and (B) secondary outcome (depressive symptoms). The plots indicate symmetric distributions across both outcomes.

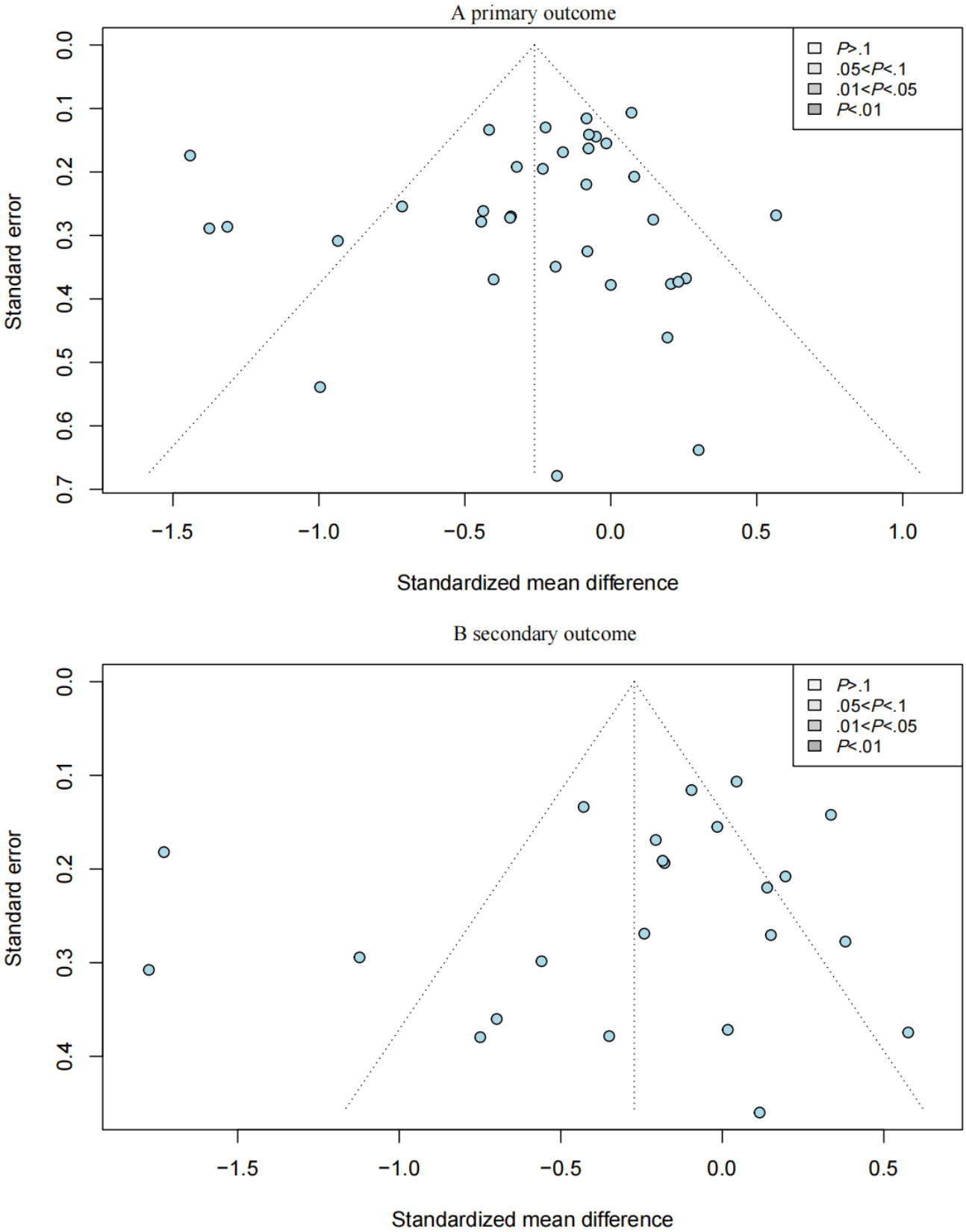

Sensitivity Analysis and Robustness

Sensitivity analyses were conducted by focusing on 27 high-quality studies (N=2885) for the primary outcome and 19 studies (N=2059) for the secondary outcome, after excluding those identified with a high risk of bias. For the primary outcome (Caregiver Burden), the pooled effect for caregiver burden remained statistically significant and increased in magnitude (SMD −0.31, 95% CI −0.50 to −0.13; t_{26}=−3.47; P=.002; 95% PI −1.28 to 0.59; τ^2=0.169, τ=0.412) (Table S3 in Multimedia

XSL•FO
RenderX

Appendix 1). Heterogeneity remained substantial (I^2=77.0%; Q=113.19; P<.001). This PI indicates that while the average effect is beneficial, the outcome in a specific clinical setting could vary.

Regarding the secondary outcome (Depressive Symptoms), the pooled effect yielded a borderline significant reduction (SMD −0.31, 95% CI −0.63 to 0.01; t_{18}=−2.06; P=.05; 95% PI −1.63 to 1.01; τ^2=0.371, τ=0.609). While heterogeneity remained high (I^2=88.1%, P<.001), the 95% PI and the borderline significant P value support that eHealth interventions provide a nonsignificant trend for depressive symptoms. Collectively, these results derived from the conservative HKSJ model confirmed that the observed clinical benefits of eHealth interventions are not driven by low-quality studies and remain robust even under the most rigorous inclusion criteria.

Quality of Evidence (GRADE)

Using the GRADE approach, the certainty for both caregiver burden and depressive symptoms was rated as moderate (Table 5). Evidence was downgraded 1 level due to risk of bias associated with lack of participant blinding and reliance on self-reported outcomes. Despite high heterogeneity, consistent effect direction across studies supported a moderate rating.

Table . Summary of findings and Grading of Recommendations, Assessment, Development, and Evaluation certainty of evidence.

Certainty assessment							Patients, n		Effect (95% CI)		Certainty	Importance
Studies, N	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	eHealth interventions	Normal care	Relative	Absolute		
Caregiver burden												
35	Randomized trials	Serious[a]	Not serious	Not serious	Not serious	None	1700	1688	N/A[b]	SMD[c] −0.26 SD lower (−0.42 lower to −0.10 lower)	◯[d] Moderate[a]	Critical
Depressive symptoms												
23	Randomized trials	Serious[a]	Not serious	Not serious	Not serious	None	1261	1206	N/A[b]	SMD −0.27 SD lower (−0.53 lower to 0.01 lower)	◯[d] Moderate[a]	Critical

[a]Downgraded 1 level due to lack of blinding (D2) and subjective self-reporting (D4) in most randomized controlled trials.

[b]Not applicable.

[c]SMD: standardized mean difference.

[d] indicates level of evidence; ◯ indicates no downgrade in certainty within the GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) framework.

Discussion

Principal Findings

This systematic review and meta-analysis of 35 RCTs suggests that eHealth interventions are associated with a small but statistically significant reduction in caregiver burden and depressive symptoms among informal caregivers of people with dementia. Although pooled estimates favor eHealth interventions, the 95% PI indicates that the effect in a future specific clinical context could range from a substantial benefit to no effect. This distinction suggests that while eHealth interventions are validated at the population level, their effectiveness in real-world settings is highly context-dependent. The moderate certainty of this evidence, assessed by GRADE, reflects consistent effects direction alongside recognized limitations of psychosocial trials, particularly the inability to blind participants and outcome assessors.

A key observation is that eHealth interventions incorporating human support or active monitoring tend to yield larger effect sizes than fully self-guided approaches. Although subgroup analyses were not statistically significant, consistent within-group patterns suggest that human involvement (eg, professional feedback, coaching, or monitoring) may enhance engagement and adherence [62,63]. The magnitude of effect observed in this review exceeds the negligible effects reported in some prior meta-analyses (eg, SMD≈−0.06) [19], yet remains smaller than those reported for structured, education-based interventions (eg, g=−0.45) [2], and is consistent with recent syntheses of home-based digital interventions reporting small-to-moderate but heterogeneous effects [64]. The wider literature, however, remains inconsistent, with some studies reporting moderate short-term benefits and others demonstrating small or nonsignificant effects, particularly for caregiver burden and depressive outcomes [17-19,21]. This heterogeneity likely reflects variation in intervention intensity, degree of human support, and methodological differences across trials, including

XSL•FO
RenderX

attrition, outcome measurement, and duration of follow-up. Importantly, while confidence intervals represent the average effect, the wide PIs underscore substantial variability in outcomes across contexts. Collectively, these findings indicate that eHealth interventions yield modest and heterogeneous effects, driven more by intervention characteristics than by uniformly strong efficacy.

Subgroup analyses further suggest that short-term (≤8 weeks) and mobile-based interventions tend to demonstrate relatively stronger effects on caregiver burden reduction. While consistent with prior evidence that time-limited digital interventions can improve proximal psychosocial outcomes, effects on caregiver burden remain inconsistent [19]. Compared with earlier reports of larger effects (eg, d=−0.65) [17], and more recent estimates showing smaller effects (eg, SMD=−0.21) [18], the pooled estimates in this review are comparatively conservative, potentially reflecting methodological differences, including the use of the HKSJ adjustment. Although shorter-duration (≤2 months) and mobile- or web-based interventions appeared more likely to yield statistically significant reductions in caregiver burden, these findings were not robust under substantial heterogeneity (I^2≈75% - 92%) and variations in study design [17,18,65]. Overall, effects on caregiver burden remain small or nonsignificant, particularly for long-term outcomes and quality of life. One plausible explanation is that short-term or mobile-based interventions primarily influence proximal outcomes, while failing to produce sustained changes in more complex outcomes such as caregiver burden and quality of life, potentially due to limited intensity or insufficient personalization [66]. Taken together, these findings suggest that intervention effectiveness is determined more by design features, outcome selection, and intensity than by delivery modality alone [16,67].

Meta-regression analyses demonstrated a potential moderating effect of caregiver age, with reduced effectiveness observed in older populations in univariate models. However, these findings suggest that intervention effectiveness is determined more by design features, outcome selection, and intensity than by delivery modality alone. This pattern aligns with existing evidence suggesting that digital interventions for caregivers generally yield small or nonsignificant effects and limited improvements in quality of life [66], and that benefits may not be uniformly distributed across subgroups. Age-related differences may be attributable to broader digital divide factors, including variability in digital literacy, usability challenges, and technology-related self-efficacy [68]. Evidence from gerontechnology and digital health research indicates that adoption and sustained engagement are strongly influenced by usability and contextual fit, with digital literacy demands and cognitive load representing persistent barriers [69,70]. In addition, broader syntheses of digital therapeutics adoption identify mistrust, usability constraints, and insufficiently tailored onboarding processes as key barriers among older populations [71]. Accordingly, differential effectiveness by age is more plausibly explained by variation in engagement and usability rather than by inherent differences in treatment response.

Overall, the findings of this review are consistent with recent umbrella and meta-analytic evidence suggesting that intervention heterogeneity, rather than delivery modality per se, is the principal driver of variability in effect estimates across studies [17,18,21]. This reinforces the interpretation that effectiveness is not determined solely by the presence of digital components but by how interventions are designed, implemented, and aligned with user needs. Accordingly, standardized, one-size-fits-all digital approaches are unlikely to achieve optimal outcomes. Future interventions should instead prioritize adaptability and personalization, alongside the deliberate integration of human support elements (eg, guidance, feedback, or monitoring), to enhance user engagement, adherence, and overall effectiveness.

Limitations

Several limitations should be considered when interpreting these findings. First, the wide PI indicates substantial between-study variability, suggesting that intervention effects may differ considerably across real-world settings. Second, as common in psychosocial and digital health research, most studies were rated as having “some concerns” or “high risk” of bias, primarily due to the inability to blind participants (RoB 2 Domain 2) and reliance on self-reported outcomes (Domain 4). Third, high attrition rates in several pilot trials may have affected internal validity [35,38], although meta-regression did not identify attrition as a significant moderator. Fourth, while subgroup analyses identified potentially important moderators (eg, intervention duration, delivery modality, and level of human support), these findings should be interpreted cautiously due to limited statistical power and persistent heterogeneity. Furthermore, although the screening process was rigorous, the preliminary title-based triage may have introduced a minimal risk of inadvertently excluding relevant studies. Finally, this review was restricted to English-language peer-reviewed publications, potentially excluding relevant gray literature or region-specific implementations.

Innovation, Contributions, and Real-World Implications

This review extends the existing literature in several important ways. Methodologically, the application of the HKSJ approach, alongside the reporting of PIs, provides more conservative and clinically informative estimates of intervention effects. Analytically, it moves beyond evaluating overall effectiveness by exploring potential moderators, offering insights into the conditions under which eHealth interventions may be more effective. Although subgroup differences were not statistically significant, the observed patterns highlight the potential importance of intervention characteristics such as human support and duration [10].

From a practical perspective, these findings have important implications for clinical practice, digital health development, and health policy. Clinically, guided eHealth interventions could be integrated into dementia care pathways as scalable support tools, particularly during high-stress periods. For digital health developers, platforms should be grounded in evidence-based psychological frameworks [17], and our results underscore the potential value of embedding human-supported components, such as coaching, feedback, or monitoring, within digital platforms, rather than relying on fully self-guided designs. At the health system level, the evidence points toward the potential of a hybrid care model in which digital infrastructure is

XSL•FO
RenderX

complemented by training human facilitators or "digital coaches," whose involvement appears linked to intervention effectiveness [62,63]. Usability and embedded support appear more critical than baseline digital literacy for successful adoption [71].

Conclusions

This review provides timely evidence that eHealth interventions are associated with modest reductions in caregiver burden and improvements in depressive symptoms among informal caregivers of people with dementia. By applying the HKSJ framework with PIs, this study offers more robust and practice-relevant estimates than prior reviews. Interpretation of these findings requires integrating multiple dimensions of evidence, as statistically significant average effects based on 95% CIs coexist with wide 95% PIs, indicating substantial heterogeneity and suggesting that real-world effectiveness may range from meaningful benefit to minimal impact. This variability, together with the identified risk of bias, particularly the lack of participant blinding, and an overall moderate level of certainty based on GRADE, indicates that intervention effectiveness is context-dependent and influenced by both methodological and implementation factors. Subgroup analyses indicate potential benefits of human-supported and shorter-duration interventions, although substantial heterogeneity persists. Future research should focus on optimizing hybrid digital-human intervention models, improving accessibility across diverse populations, and generating high-quality evidence with longer follow-up to enhance consistency and real-world applicability.

What Is Already Known

These include the following:

- Informal caregivers of people with dementia experience high levels of psychological burden; eHealth interventions offer a promising scalable approach for caregiver support.
- Previous meta-analyses often relied on conventional statistical models, which may underestimate uncertainty and yield overly narrow CIs in the presence of substantial heterogeneity.

What This Paper Adds

These include the following:

- Methodological rigor: This review applies PRISMA-S reporting standards and uses the HKSJ method and PIs, providing a more conservative and practice-relevant estimate of intervention effectiveness.
- Intervention characteristics insights: Interventions incorporating human support and mobile-based delivery were associated with more consistent effects, whereas fully self-guided approaches showed less consistent evidence of benefit.
- Demographic moderators: Meta-regression reveals that caregiver age appears to influence intervention effectiveness, with reduced effects observed in older populations, highlighting the importance of age-sensitive intervention design.

Acknowledgments

The authors confirm that no generative artificial intelligence (AI) or AI-assisted technologies were used in the preparation of this manuscript.

Funding

This study was funded by Shanghai Science and Technology Foundation Soft Science Research Project (grant 25692114800) and National Natural Science Foundation of China (grant 72304183).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

ML, XW, and CG designed the study and revised the manuscript. ML and XW drafted and revised the manuscript. ML and JR conducted the database search and data extraction. ML and CG conducted the data analysis. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables and figures including search strategy (Table S1), characteristics of included randomized controlled trials of eHealth interventions for informal caregivers of people with dementia (Table S2), comparison of primary and sensitivity meta-analyses (Table S3), and forest plots of subgroup analyses for caregiver burden and depressive symptoms (Figures S1 and S2).
[DOCX File, 33951 KB - jmir_v28i1e78568_app1.docx]

XSL•FO
RenderX

Checklist 1
PRISMA 2020 expanded checklist.
[PDF File, 1381 KB - jmir_v28i1e78568_app2.pdf]

Checklist 2
PRISMA-S checklist.
[PDF File, 63 KB - jmir_v28i1e78568_app3.pdf]

Checklist 3
PRISMA 2020 for abstracts checklist.
[PDF File, 464 KB - jmir_v28i1e78568_app4.pdf]

References

1. Hao M, Chen J. Trend analysis and future predictions of global burden of Alzheimer's disease and other dementias: a study based on the global burden of disease database from 1990 to 2021. BMC Med 2025 Jul 1;23(1):378. [doi: 10.1186/s12916-025-04169-w] [Medline: 40597083]
2. Pradana AA, Susanto H, Chiu HL. Effectiveness of dementia literacy interventions for caregivers of people with dementia: a meta-analysis of randomized controlled trials. Int J Nurs Stud 2025 Nov;171:105191. [doi: 10.1016/j.ijnurstu.2025.105191] [Medline: 40885079]
3. Livingston G, Huntley J, Liu KY, et al. Dementia prevention, intervention, and care: 2024 report of the Lancet standing Commission. Lancet 2024 Aug 10;404(10452):572-628. [doi: 10.1016/S0140-6736(24)01296-0] [Medline: 39096926]
4. Zhi N, Ren R, Qi J, et al. The China Alzheimer Report 2025. Gen Psychiatr 2025;38(4):e102020. [doi: 10.1136/gpsych-2024-102020] [Medline: 40792123]
5. Eggleston K, Kim D. Dementia care in a rapidly aging society. Alzheimers Dement 2025 May;21(5):e70268. [doi: 10.1002/alz.70268] [Medline: 40407100]
6. Duangjina T, Hershberger PE, Gruss V, Fritschi C. Resilience in family caregivers of Asian older people with dementia: an integrative review. J Adv Nurs 2025 Jan;81(1):156-170. [doi: 10.1111/jan.16272] [Medline: 38863175]
7. Reuben DB, Stevens AB, Gill TM, et al. Patient and caregiver outcomes of health system, community-based, and usual dementia CARE: a prespecified analysis of the Dementia CARE study (D-CARE) randomized clinical trial. JAMA Intern Med 2025 Oct 1;185(10):1227-1236. [doi: 10.1001/jamainternmed.2025.4247] [Medline: 40892399]
8. Garrido SC, Teixeira S, Mora-Lopez G, Sampaio F. Factors associated with anxiety, stress, depression and burden among informal caregivers of patients with dementia: a cross-sectional study. BMC Nurs 2025 Nov 11;24(1):1384. [doi: 10.1186/s12912-025-04014-8] [Medline: 41219763]
9. Bui LK, Park M, Giap TT. eHealth interventions for informal caregivers of people with dementia. Alzheimers Dement 2023 Jun;19(S5). [doi: 10.1002/alz.059865]
10. Cattaneo A, Vitali A, Regazzoni D, Rizzi C. The burden of informal family caregiving in Europe, 2000-2050: a microsimulation modelling study. Lancet Reg Health Eur 2025 Jun;53:101295. [doi: 10.1016/j.lanepe.2025.101295] [Medline: 40255934]
11. Su YY, Wang WF, Jhang KM, Mau LW, Chen CM. Enhancing care service utilization and reducing burden: the role of needs assessments for dementia caregivers in long-term care. BMC Nurs 2025 Aug 1;24(1):1004. [doi: 10.1186/s12912-025-03686-6]
12. eHealth. World Health Organization Eastern Mediterranean Regional Office. URL: https://www.emro.who.int/health-topics/ehealth [accessed 2026-03-30]
13. Abdulazeem H, Borges do Nascimento IJ, Weerasekara I, et al. Use of digital health technologies for dementia care: bibliometric analysis and report. JMIR Ment Health 2025 Feb;12:e64445. [doi: 10.2196/64445]
14. Bui LK, Park M, Giap TTT. eHealth interventions for the informal caregivers of people with dementia: a systematic review of systematic reviews. Geriatr Nurs (Lond) 2022 Nov;48:203-213. [doi: 10.1016/j.gerinurse.2022.09.015]
15. Zhi S, Ma D, Song D, et al. The influence of web‐based decision aids on informal caregivers of people with dementia: a systematic mixed‐methods review. Int J Mental Health Nurs 2023 Aug;32(4):947-965. [doi: 10.1111/inm.13177]
16. Zhu L, Xing Y, Jia H, Xu W, Wang X, Ding Y. Effects of telehealth interventions on the caregiver burden and mental health for caregivers of people with dementia: a systematic review and meta-analysis. Aging Ment Health 2024 Nov;28(11):1427-1439. [doi: 10.1080/13607863.2024.2371480]
17. Lumini MJ, França D, Sousa MR, et al. Efficacy of digital technology-based interventions for reducing caregiver burden and stress: a systematic review and meta-analysis. Front Digit Health 2025;7:1636084. [doi: 10.3389/fdgth.2025.1636084] [Medline: 41281293]
18. Hua M, Su J, Zhang F, Chen S, Li J, Yang L. Effects of digital psychological interventions for family caregivers of people with dementia: a systematic review and meta-analysis. Age Ageing 2026 Jan 3;55(1):afaf373. [doi: 10.1093/ageing/afaf373] [Medline: 41505195]

XSL•FO
RenderX

19. Li Q, Zhang D, Quinn É, Walsh P, Hynes D. Online interventions for dementia caregiver burden and self-efficacy: a systematic review and meta-analysis. Int J Nurs Stud Adv 2026 Jun;10:100490. [doi: 10.1016/j.ijnsa.2026.100490]
20. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
21. Naunton Morgan B, Windle G, Sharp R, Lamers C. eHealth and web-based interventions for informal carers of people with dementia in the community: umbrella review. J Med Internet Res 2022 Jul 22;24(7):e36727. [doi: 10.2196/36727]
22. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ :n71. [doi: 10.1136/bmj.n71]
23. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z]
24. Lefebvre C, Glanville J, Briscoe S, et al. Chapter 4: searching for and selecting studies (chapter last updated September 2024). In: Higgins JPT, Thomas J, Chandler J, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 65. URL: https://training.cochrane.org/handbook [accessed 2026-04-18]
25. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023 Dec;12(4):101014. [doi: 10.1016/j.imr.2023.101014]
26. Sterne JAC, Sutton AJ, Ioannidis JPA, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011 Jul 22;343:d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
27. Beauchamp N, Irvine AB, Seeley J, Johnson B. Worksite-based internet multimedia program for family caregivers of persons with dementia. Gerontologist 2005 Dec;45(6):793-801. [doi: 10.1093/geront/45.6.793] [Medline: 16326661]
28. Gitlin LN, Winter L, Dennis MP, Hodgson N, Hauck WW. Targeting and managing behavioral symptoms in individuals with dementia: a randomized trial of a nonpharmacological intervention. J Am Geriatr Soc 2010 Aug;58(8):1465-1474. [doi: 10.1111/j.1532-5415.2010.02971.x] [Medline: 20662955]
29. Kwok T, Wong B, Ip I, Chui K, Young D, Ho F. Telephone-delivered psychoeducational intervention for Hong Kong Chinese dementia caregivers: a single-blinded randomized controlled trial. Clin Interv Aging 2013;8:1191-1197. [doi: 10.2147/CIA.S48264] [Medline: 24072965]
30. Torkamani M, McDonald L, Saez Aguayo I, et al. A randomized controlled pilot study to evaluate a technology platform for the assisted living of people with dementia and their carers. J Alzheimers Dis 2014;41(2):515-523. [doi: 10.3233/JAD-132156] [Medline: 24643137]
31. Kales HC, Gitlin LN, Stanislawski B, et al. Effect of the WeCareAdvisorTM on family caregiver outcomes in dementia: a pilot randomized controlled trial. BMC Geriatr 2018 May 10;18(1):113. [doi: 10.1186/s12877-018-0801-8] [Medline: 29747583]
32. Meichsner F, Theurer C, Wilz G. Acceptance and treatment effects of an internet-delivered cognitive-behavioral intervention for family caregivers of people with dementia: a randomized-controlled trial. J Clin Psychol 2019 Apr;75(4):594-613. [doi: 10.1002/jclp.22739] [Medline: 30597537]
33. Metcalfe A, Jones B, Mayer J, et al. Online information and support for carers of people with young-onset dementia: a multi-site randomised controlled pilot study. Int J Geriatr Psychiatry 2019 Oct;34(10):1455-1464. [doi: 10.1002/gps.5154] [Medline: 31111516]
34. Williams KN, Perkhounkova Y, Shaw CA, Hein M, Vidoni ED, Coleman CK. Supporting family caregivers with technology for dementia home care: a randomized controlled trial. Innov Aging 2019 Jul;3(3):igz037. [doi: 10.1093/geroni/igz037] [Medline: 31660443]
35. James TA, James D, Larkey LK. Heart-focused breathing and perceptions of burden in Alzheimer's caregivers: an online randomized controlled pilot study. Geriatr Nurs 2021;42(2):397-404. [doi: 10.1016/j.gerinurse.2021.02.006] [Medline: 33639543]
36. Baruah U, Varghese M, Loganathan S, et al. Feasibility and preliminary effectiveness of an online training and support program for caregivers of people with dementia in India: a randomized controlled trial. Int J Geriatr Psychiatry 2021 Apr;36(4):606-617. [doi: 10.1002/gps.5502] [Medline: 33491811]
37. Fossey J, Charlesworth G, Fowler JA, et al. Online education and cognitive behavior therapy improve dementia caregivers' mental health: a randomized trial. J Am Med Dir Assoc 2021 Jul;22(7):1403-1409. [doi: 10.1016/j.jamda.2020.10.009] [Medline: 33288467]
38. Teles S, Ferreira A, Paúl C. Feasibility of an online training and support program for dementia carers: results from a mixed-methods pilot randomized controlled trial. BMC Geriatr 2022 Mar 1;22(1):173. [doi: 10.1186/s12877-022-02831-z] [Medline: 35232389]
39. Hepburn K, Nocera J, Higgins M, et al. Results of a randomized trial testing the efficacy of Tele-Savvy, an online synchronous/asynchronous psychoeducation program for family caregivers of persons living with dementia. Gerontologist 2022 Apr 20;62(4):616-628. [doi: 10.1093/geront/gnab029]
40. Bodenstein KC. Virtual mindful chair yoga (MCY) intervention for older adults with dementia and caregivers: pilot randomized controlled trial [Master's thesis]. : McGill University; 2022.

XSL•FO
RenderX

41. Han AR, Yuen HK, Jenkins J. The feasibility and preliminary effects of a pilot randomized controlled trial: videoconferencing acceptance and commitment therapy in distressed family caregivers of people with dementia. J Health Psychol 2023 May;28(6):554-567. [doi: 10.1177/13591053221141131]
42. Rhodus EK, Baum C, Kryscio R, et al. Feasibility of telehealth occupational therapy for behavioral symptoms of adults with dementia: randomized controlled trial. Am J Occup Ther 2023 Jul 1;77(4):7704205010. [doi: 10.5014/ajot.2023.050124]
43. Rodriguez MJ, Kercher VM, Jordan EJ. Technology caregiver intervention for Alzheimer's disease (I‐CARE): feasibility and preliminary efficacy of Brain CareNotes. J Am Geriatr Soc 2023 Dec;71(12):3836-3847. [doi: 10.1111/jgs.18591]
44. Hu M, Liu ZY, Xiong J, et al. Impact of 'web-based mindfulness based stress reduction combined with solution-focused intervention on dementia caregivers' negative affect. NeuroAsia 2024;29(2):435-443. [doi: 10.54029/2024fpi]
45. Salehinejad S, Jannati N, Azami M, Mirzaee M, Bahaadinbeigy K. A web-based information intervention for family caregivers of patients with Dementia: a randomized controlled trial. J Inform Sci 2024 Feb;50(1):104-115. [doi: 10.1177/01655515221081353]
46. Xie Y, Shen S, Liu C, Hong H, Guan H, Zhang J, et al. Internet-based supportive interventions for family caregivers of people with dementia: randomized controlled trial. JMIR Aging 2024 Oct 4:e50847. [doi: 10.2196/69493]
47. Jain S. Efficacy of brief cognitive behavior therapy with mindfulness techniques on perceived stress, rumination, life satisfaction and burden among caregivers of dementia: a tele-health trial [MPhil thesis]. : Central Institute of Psychiatry; 2024.
48. Song Y, Papazyan A, Lee D, et al. The feasibility of a sleep education program for informal dementia care dyads: a pilot randomized controlled trial. J Am Geriatr Soc 2024 Apr;72(4):1207-1215. [doi: 10.1111/jgs.18720]
49. Yuan S, Zhang J, Wang Z, Wang H, Xia M. Supporting informal dementia caregivers through an iSupport web-based primary health care intervention: hybrid effectiveness-implementation mixed methods study. J Med Internet Res 2025 Sep 9;27:e77688. [doi: 10.2196/77688]
50. Nguyen TTT, Vu LDT, Truong KT, et al. Smartphone app-based psychoeducation for caregivers of people with dementia in Vietnam: a pilot randomized controlled trial. Clin Interv Aging 2025;20:1647-1660. [doi: 10.2147/CIA.S551994] [Medline: 41000210]
51. Nguyen TTT, Vu LDT, Nguyen TC, et al. Mobile app psychoeducation for dementia caregivers in Vietnam: a randomized controlled trial. J Multidiscip Healthc 2025;18:6871-6885. [doi: 10.2147/JMDH.S557573] [Medline: 41158517]
52. Han AR, Oster R, Yuen H, Jenkins J, Hawkins J, Edwards L. Videoconference-delivered acceptance and commitment therapy for family caregivers of people with dementia: pilot randomized controlled trial. JMIR Form Res 2025 Mar 31;9:e67545. [doi: 10.2196/67545]
53. Nichols LO, Martindale-Adams J, Seel RT, Zuber JK, Perera R, Perrin PB. Reducing burden and anxiety for caregivers of veterans with traumatic brain injury and dementia: randomized controlled trial of the resources for enhancing all caregivers' health-Hope intervention. Rehabil Psychol 2025 Jul 31. [doi: 10.1037/rep0000632] [Medline: 40742769]
54. Stevens AB, Cho J, Birchfield T, et al. A randomized trial of two online platforms for dementia family caregivers: GamePlan4Care and Resources4Care. Alzheimers Dement 2025 Oct;21(10):e70690. [doi: 10.1002/alz.70690] [Medline: 41085219]
55. Pfaff M, Hoffmann W, Boekholt M, et al. Cost-Effectiveness of a Digitally Supported Care Management Program for Caregivers of People With Dementia. Value Health 2025 Apr;28(4):527-535. [doi: 10.1016/j.jval.2025.01.011] [Medline: 39922302]
56. Windle G, Flynn G, Hoare Z, Goulden N, Edwards RT, Anthony B, et al. Evaluating the effects of the World Health Organization's online intervention 'iSupport'to reduce depression and distress in dementia carers: a multi-centre six-month randomised controlled trial in the UK. Lancet Reg Health Eur 2025:101125. [doi: 10.1016/j.lanepe.2025.101125]
57. Sun Y, Ji M, Leng M, Chen B, Liu S, Wang Z. The effects of a mobile application-based dementia care intelligent recommender system on caregivers of people with dementia: a randomized controlled trial. Age Ageing 2026 Feb 1;55(2):afag021. [doi: 10.1093/ageing/afag021] [Medline: 41642176]
58. Yuan Q, Lee YT, Samari E, et al. Evaluating the feasibility and potential effectiveness of a mobile-based intervention to promote the mental health of informal caregivers of persons with dementia in Singapore: results from a mixed-methods two-arm pilot randomized controlled trial. J Affect Disord 2026 Apr;399:121059. [doi: 10.1016/j.jad.2025.121059]
59. Brijnath B, Dang TH, Enticott J, et al. Evaluating the effects of the World Health Organisation's online intervention "iSupport lite" to reduce burden in multilingual dementia carers: results from the Australian Draw-Care randomised clinical trial. Age Ageing 2026 Feb 1;55(2):afag030. [doi: 10.1093/ageing/afag030] [Medline: 41719424]
60. Cheng ST, Ng PHF. Intervention without borders—an automated self-guided AI-enhanced psychoeducation intervention for dementia caregivers: parallel-group randomized waitlist-controlled trial. Am J Geriatr Psychiatry 2026 Jan;PMID. [doi: 10.1016/j.jagp.2026.01.001]
61. Durepos P, Ricketts N, Olthuis JV, et al. Acceptance and commitment therapy tele-counselling program with care partners of persons living with dementia: a pilot feasibility RCT. J Appl Gerontol 2026 Feb;45(2):225-236. [doi: 10.1177/07334648251338303]
62. Greenway FT, Weal M, Palmer-Cooper EC. Hybrid mHealth care: patient perspectives of blended treatments for psychosis. A systematic review. Schizophr Res 2024 Dec;274:1-10. [doi: 10.1016/j.schres.2024.08.017] [Medline: 39244945]

XSL•FO
RenderX

63. Alvarez-Jimenez M, Nicholas J, Valentine L, et al. A national evaluation of a multi-modal, blended, digital intervention integrated within Australian youth mental health services. Acta Psychiatr Scand 2025 Mar;151(3):317-331. [doi: 10.1111/acps.13751] [Medline: 39260824]
64. Albarqi MN. Home-based digital healthcare interventions for dementia: a systematic review of patient and family caregiver outcomes. Healthcare (Basel) 2026;14(7):854. [doi: 10.3390/healthcare14070854]
65. de-Moraes-Ribeiro FE, Moreno-Cámara S, da-Silva-Domingues H, Palomino-Moral P, Del-Pino-Casado R. Effectiveness of internet-based or mobile app interventions for family caregivers of older adults with dementia: a systematic review. Healthcare (Basel) 2024 Jul 27;12(15):1494. [doi: 10.3390/healthcare12151494] [Medline: 39120195]
66. Kim M, Tan KR, Coombs LA. Effects of web-based interventions on cancer caregivers' burden and quality of life: a systematic review and meta-analysis. Palliat Support Care 2025 Jul 18;23:e134. [doi: 10.1017/S1478951525100370] [Medline: 40676719]
67. Zhai S, Chu F, Tan M, Chi NC, Ward T, Yuwen W. Digital health interventions to support family caregivers: an updated systematic review. Digit Health 2023;9:20552076231171967. [doi: 10.1177/20552076231171967] [Medline: 37223775]
68. Yang R, Gao S, Jiang Y. Digital divide as a determinant of health in the U.S. older adults: prevalence, trends, and risk factors. BMC Geriatr 2024;24(1):1027. [doi: 10.1186/s12877-024-05612-y]
69. Jutai JW, Hatoum F, Bhardwaj D, Hosseini M. Implementation of digital health technologies for older adults: a scoping review. Front Aging 2024;5:1349520. [doi: 10.3389/fragi.2024.1349520] [Medline: 38784681]
70. Moreno A, Scola MC, Sun H, et al. A systematic review of gerontechnologies to support aging in place among community-dwelling older adults and their family caregivers. Front Psychol 2024;14:1237694. [doi: 10.3389/fpsyg.2023.1237694]
71. Azizan A. Barriers and motivators of geriatric engagement in digital therapeutics: a narrative review. Ageing Res Rev 2026 Jan;114:102963. [doi: 10.1016/j.arr.2025.102963] [Medline: 41318100]

Abbreviations

CENTRAL: Cochrane Central Register of Controlled Trials
GRADE: Grading of Recommendations, Assessment, Development, and Evaluation
HKSJ: Hartung-Knapp-Sidik-Jonkman
IM: Instant Messaging
MeSH: Medical Subject Headings
PI: prediction interval
PICOS: Population, Intervention, Comparison, Outcomes, and Study Design
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Search extension
RCT: randomized controlled trial
RoB 2: Cochrane Risk of Bias 2
SMD: standardized mean difference
VC: videoconferencing

Edited by S Brini; submitted 04.Jun.2025; peer-reviewed by B Wolff, H Susanto; revised version received 17.Apr.2026; accepted 17.Apr.2026; published 03.Jun.2026.

Please cite as:
Liang M, Rong J, Grosan C, Wang X
Effectiveness of eHealth Interventions in Alleviating Burden on Informal Caregivers of People With Dementia: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e78568
URL: https://www.jmir.org/2026/1/e78568
doi:10.2196/78568

© Min Liang, Jinxia Rong, Crina Grosan, Xiyi Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 3.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital and Blended Lifestyle Interventions for Preschool-Aged Children and Families With a Low Socioeconomic Position and the General Population: Scoping Review

Lea Hohendorf[1,2], MSc; Tessa Dekkers[1], PhD; Hanneke Kip[1,3], PhD; Laurence Alpay[2], PhD; Saskia M Kelders[1,4], PhD

[1]Health Psychology & Technology, Faculty of Behavioural, Management and Social Sciences, University of Twente, Drienerlolaan 5, Enschede, Overijssel, The Netherlands

[2]Medical Technology Research Group, Faculty of Healthcare, Sport & Well-being, Inholland University of Applied Sciences, Haarlem, The Netherlands

[3]Department of Research, Stichting Transfore, Deventer, The Netherlands

[4]Optentia Research Focus Area, Vaal Triangle Campus, North-West University, Vanderbijlpark, South Africa

Corresponding Author:
Lea Hohendorf, MSc
Health Psychology & Technology, Faculty of Behavioural, Management and Social Sciences, University of Twente, Drienerlolaan 5, Enschede, Overijssel, The Netherlands

Abstract

Background: Many unhealthy habits develop in early childhood and can lead to long-term health risks, which disproportionately affect children with low socioeconomic position (SEP). Digital and blended lifestyle interventions can promote healthier lifestyles, yet families with lower SEP remain underrepresented and face unique barriers to healthy behaviors and intervention access. As a result, it remains unclear which intervention characteristics are most effective for these populations.

Objective: This study aimed to identify and map digital and blended lifestyle interventions targeting preschool-aged children and explore how intervention characteristics and reported effectiveness patterns differ between interventions for the general population and low SEP families.

Methods: A search across Scopus, Web of Science, Cochrane Library, ERIC, and ACM Digital Library was conducted. Studies were eligible if they (1) targeted preschool-aged children (6 months to 5 years) or caregivers, (2) evaluated a digital or blended lifestyle intervention, and (3) addressed at least one behavioral domain (nutrition, physical activity, sedentary behavior, sleep, and oral health). Studies focusing on pregnant women, children aged >5 years, or interventions delivered solely nondigitally were excluded. Screening and prioritization were supported by artificial intelligence–assisted software (ASReview; Utrecht University). Data covered intervention targets, populations, theoretical and guideline foundations, delivery modes and settings, and persuasive systems design features. Intervention characteristics and effectiveness categories were synthesized descriptively. The methodological quality of quasi-experimental studies was assessed using the Joanna Briggs Institute critical appraisal tools.

Results: A total of 77 studies describing 54 interventions were included. Of these, 35 targeted the general population, and 19 focused solely on low SEP families. Interventions were similar across groups: typically parent-focused, targeting multiple lifestyle domains, and informed by theories, frameworks, or evidence-based guidelines. Low SEP interventions more often used text messaging, included fewer persuasive design features, and tended toward single delivery channels. Effectiveness findings were mixed; no consistent patterns emerged when interventions were grouped by their characteristics. Among low SEP interventions only, less effective interventions more often targeted multiple behaviors and included more persuasive systems design features.

Conclusions: To our knowledge, this is one of the first reviews to map digital and blended lifestyle interventions for preschool-aged children while comparing intervention characteristics and descriptive effectiveness patterns across general and low SEP populations. The review highlights heterogeneity in intervention characteristics and outcomes and identifies gaps where evidence is limited. These findings underscore the need to better represent low SEP families in intervention research, strengthen understanding of feasibility and contextual fit, and support multilevel approaches that reflect families' everyday contexts. However, the findings should be interpreted in light of several limitations, including the exclusion of gray literature, partial double-screening, the possibility that artificial intelligence–assisted prioritization may have omitted relevant studies, and variation in the reporting of intervention components across studies.

(J Med Internet Res 2026;28:e86596) doi:10.2196/86596

XSL•FO
RenderX

KEYWORDS

eHealth; lifestyle interventions; digital health interventions; health behaviors; preschool-aged children; low socioeconomic position; intervention characteristics; effectiveness; intervention evaluation; persuasive systems design

Introduction

Unhealthy lifestyle behaviors are already common in preschool-aged children (6 months to 5 years), raising concern as they may shape long-term health trajectories [1,2]. Across regions, only a minority of preschool-aged children, typically fewer than 1 in 10, meet all 24-hour movement guidelines for physical activity, screen time, and sleep [3,4]. Likewise, frequent consumption of sugary snacks and drinks, which contributes to early childhood caries, is already common in this age group [5-7]. These behaviors are linked to an increased risk of overweight, obesity, and other noncommunicable diseases later in life [8-12]. Because routines are established early, intervening in this age group is critical to prevent negative health outcomes and promote healthier lifelong habits.

This need for early prevention is especially urgent among children from families with a low socioeconomic position (SEP), who are disproportionally affected by these health disparities. For example, European studies show that children with less-educated parents are more than twice as likely to be overweight as peers from higher-educated families [13]. SEP, defined by income, education, occupation, and perceived social class [14], strongly shapes health behaviors, as structural barriers such as financial constraints, neighborhood safety, and access to resources can hinder adoption of healthier lifestyles [15,16]. Language barriers and low literacy may also limit access to interventions and their content [17]. Without accounting for families' contexts, interventions may result in poor engagement and may widen disparities [18]. Despite their greater need, low SEP families remain underrepresented in research and prevention efforts. A systematic overview of existing interventions can help clarify the extent to which this group has been targeted and reveal important gaps in prevention approaches.

More specifically, limited knowledge exists on the characteristics of interventions for preschool-aged children, as well as how such characteristics have been incorporated into their design and delivery. For example, parents are often assumed to be the main target, which makes sense given their central role in shaping children's behaviors [19]. Yet, questions remain about whether and how other potential stakeholders, such as children themselves or grandparents, have been addressed and what this might mean for intervention relevance and effectiveness [20]. Delivery and design strategies raise similar uncertainties. Traditional formats can create financial and logistical barriers [21,22], while digital approaches offer broader reach and tailoring but face challenges such as low literacy, limited trust, and poor connectivity [18,23-25]. Such barriers can undermine adherence. Blended formats may help address these barriers. Design features such as reminders, feedback, and social support, as described in the persuasive systems design (PSD) framework, may further enhance support [24,26-28]. Yet it remains unclear how these strategies have been implemented in interventions for preschool-aged children or what lessons might be drawn from their application. Taken together, these gaps highlight the need for a systematic overview that consolidates knowledge on intervention characteristics, clarifies which strategies have been attempted, and identifies where important opportunities remain.

To address these gaps, this scoping review provides a comprehensive overview of existing digital and blended lifestyle interventions for preschool-aged children, considering both the general population and families with a lower SEP. low SEP families are of particular interest, as they face greater structural barriers to healthy behaviors but also have the most to gain from effective, well-designed interventions. More insight is needed into what "well-designed" might mean in this context, for example, which intervention characteristics have been most frequently applied, how different target groups have been addressed, and how effectiveness has been reported. An overview may help identify patterns or gaps in these areas, offering useful insights to guide the development of future interventions. Therefore, this review seeks to answer the following research questions: (1) What are the characteristics of existing digital and blended lifestyle interventions aimed at improving lifestyle behaviors in preschool-aged children? (2) What descriptive effectiveness patterns are reported in digital and blended lifestyle interventions targeting preschool-aged children? For both questions, additional investigation was conducted regarding how characteristics and effectiveness levels differ between interventions targeting the general population and those designed for low SEP families.

Methods

Scoping Review Methodology

A scoping review was chosen as it is well suited for mapping key concepts, identifying knowledge gaps, and systematically assessing the breadth of existing research in an emerging field [29]. Given that this review aimed to map intervention characteristics and provide a descriptive overview of the effectiveness patterns across a heterogeneous body of studies, a scoping review design was considered most appropriate to provide an overview of the current state of affairs. This review follows the Joanna Briggs Institute (JBI) framework, which expands upon the methodology proposed by Arksey and O'Malley [30] and refined by Levac et al [31]. Additionally, this review adheres to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) reporting guidelines, and the full checklist is provided in Checklist 1. No protocol was registered for this scoping review.

Eligibility Criteria

The inclusion and exclusion criteria for this review are provided in Table 1. These criteria were formulated to identify studies examining digital and blended interventions that promote healthy lifestyle behaviors in preschool-aged children (6 months to 5

XSL•FO
RenderX

years). The criteria were developed following the JBI population, concept, and context framework to ensure methodological rigor. Language criteria restricted inclusion to studies published in English. Population criteria focused on interventions targeting preschool-aged children (6 months to 5 years) and/or their caregivers, with additional stakeholders (eg, parents-to-be, educators, and health care professionals) included only when the intervention ultimately targeted improvements in children's lifestyle behaviors. The concept domain covered two elements: (1) the intervention aim, representing one or more targeted lifestyle behaviors (nutrition, physical activity, sedentary behavior, sleep, and oral health), and (2) the delivery strategy requiring interventions to be delivered digitally or in a blended format. Digital interventions were defined as those including at least one digital component (eg, text messages, app, and website), while blended interventions combined digital and nondigital (eg, in-person sessions and paper-based materials) components. Study type criteria permitted empirical peer-reviewed research reporting effectiveness, development, or implementation of interventions (eg, randomized controlled trials [RCTs], protocols, and feasibility studies). Context criteria were left unspecified to allow inclusion of studies conducted in diverse settings.

Table . Eligibility criteria used during title, abstract, and full-text screening. The criteria outline the included population (preschool-aged children and/or caregivers), the targeted lifestyle behaviors (nutrition, physical activity, sedentary behavior, sleep, and oral health), the required delivery strategy, and eligible study types.

Aspect	Inclusion	Exclusion
Language	• English	• Any other language
Population	• Target preschool-aged children aged >6 months • Target parents of preschool-aged children • Target pregnant women and parents-to-be (preventatively) • Target any other caretaker of preschool-aged children • Family-based • Target preschool or kindergarten teachers • Target health care professionals • Interventions targeting other stakeholders besides children were included only if they indirectly targeted preschool-aged children's health outcomes	• Targets children aged >5 years • Targeted pregnant women for the prenatal health of the fetus
Aim	• Nutrition • Physical activity • Sedentary behavior • Sleep • Oral health	• Mental health • Management of (chronic) illnesses or disorders (eg, autism and arthritis) • Supporting healthier lifestyles in pregnant women • Promoting (exclusive) breastfeeding
Delivery strategy	• Delivered digitally • Delivered in a blended way (digital components combined with in-person or paper-based materials)	• Delivered solely in person • Delivered solely paper-based
Study type	• Peer-reviewed journal papers reporting effectiveness, development, or implementation studies (including RCTs[a], protocols, pilot studies, and feasibility studies)	• Systematic reviews • Reports of process evaluations

[a]RCT: randomized controlled trial.

Information Sources

A systematic electronic literature search of five databases, namely Scopus, Web of Science, ERIC, ACM Digital Library, and Cochrane Library, was conducted. Each database was searched individually, and no multidatabase vendor platform was used. No study registries (eg, ClinicalTrials.gov), gray literature sources, handsearching, website browsing, or other supplementary search methods were used. Backward citation searching of included studies was performed to identify additional potentially relevant records. No study authors or experts were contacted to identify additional studies or obtain additional data. The original searches were executed in November 2023, with updated searches using the same strategy and databases conducted in May 2025 and January 2026 to ensure inclusion of recently published studies. No date limits were applied to the initial search, and the update ranges were restricted to periods not previously covered.

XSL•FO
RenderX

Search Strategy

The search strategy and its reporting followed the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) guidelines [32]; the complete PRISMA-S checklist is provided in Checklist 2. An information specialist from the University of Twente was consulted for the construction of the search strategy and the selection of databases. No search strategies from prior reviews were reused or adapted. The search strategy included keywords related to the concepts of digital and blended interventions, lifestyle behaviors, lifestyle interventions, and the target group (eg, preschool children and parents). No methodological search filters were applied. Search results were imported into Mendeley (Elsevier; initial search and the May 2025 update) and Zotero (Corporation for Digital Scholarship; January 2026 update), where duplicates were identified, manually verified, and removed. The complete search strategy used for all databases is available in Multimedia Appendix 1.

Selection of Sources of Evidence

The search results from all databases were combined and exported into the reference manager Mendeley, where duplicate records were removed. The final set of search results was then imported into ASReview (version 1.2.1; Utrecht University) for the selection of eligible studies. ASReview is an artificial intelligence (AI)-assisted tool designed to aid researchers in systematically reviewing literature [33]. To train the AI model, it was provided with prior knowledge in the form of 14 preselected papers (refer to Multimedia Appendix 2), labeled as relevant or irrelevant by the investigating researcher. ASReview uses an active learning algorithm that continuously reprioritizes records to present those most likely to be relevant first. As the model stabilizes, the probability of identifying additional relevant papers becomes very low. To balance comprehensiveness and efficiency, we used a mixed stopping rule: screening was terminated after a minimum of 10% of the dataset had been reviewed, and 150 consecutive records were irrelevant. This pragmatic rule follows a similar rationale to those used in other AI-assisted screening procedures (eg, SAFE heuristic) [34,35].

The updated search was conducted using the same search strategy and databases, and newly identified records were screened using the same procedure. The selection process consisted of a two-step screening process, including (1) title and abstract screening and (2) full-text screening. During the title and abstract screening, 2 authors (LH and TD) reviewed 200 papers and made inclusion and exclusion decisions on 187 of these. Agreement was observed for 97% of overlapping records (κ=0.90), indicating almost perfect agreement [36]. Because the first 100 abstracts were reviewed jointly before independent screening began, Cohen kappa should be interpreted as an upper-bound estimate of interrater reliability. Disagreements or uncertainties were discussed until consensus was reached; if no agreement could be found, a third author (HK) was consulted. The remaining abstracts were then screened by the first author (LH) until a total of 928 abstracts had been reviewed, and 150 consecutive papers were deemed irrelevant. In the full-text screening phase, a random selection of 12 full-text papers (10% of all papers) was independently reviewed against the inclusion and exclusion criteria by 2 authors (LH and TD), with any doubts or disagreements discussed and resolved. Cohen κ was 0.71, indicating substantial agreement between reviewers [35]. If consensus could not be reached, the third author (HK) was consulted. Finally, data extraction from the remaining eligible full-text papers was conducted by the first author (LH).

Data Charting Process

A data extraction form was developed to systematically collect key information relevant to the review questions (refer to Table 2). The first author (LH) conducted the data charting, with uncertainties or inconsistencies discussed and resolved with the second author (TD). Initially, data were charted deductively using predefined data items.

XSL•FO
RenderX

Table . Data items coded for data extraction. The table summarizes the population, study, and intervention characteristics extracted from each included paper, with accompanying explanations detailing how each item was defined and operationalized for this review.

Data item	Explanation
Population	Coded as "low SEP[a]" intervention only if the studies explicitly stated an aim to reach vulnerable populations; otherwise, they were coded as "general population."
Study design	For example, a randomized controlled trial or protocol.
Study research objectives	Research aims or objectives.
Study outcomes	Primary outcomes.
Intervention name	Name of intervention.
Intervention objectives	The intervention objectives of each intervention were recorded.
Target behaviors	Moreover, the specific target behaviors (refer to Table 1) were documented under their behavior categories.
Target group	For example, "parents," "caregivers," or "family."
Theories, frameworks, and evidence-based guidelines	For example, behavioral frameworks and theories, behavior change techniques, or evidence-based guidelines.
Strategy and channel	Delivery strategy refers to interventions being delivered either fully digitally or in a blended way. Blended refers to a combination of digital components and in-person or paper-based aspects. Concerning the delivery channel, we documented the usage of any tool to deliver the intervention content, such as a "smartphone app" or "website."
Delivery environments	For example, "daily life" or "in a primary care clinic."
Persuasive features	The persuasive features implemented in the interventions were documented and coded according to the persuasive systems design framework by Oinas-Kukkonen and Harjumaa [28]. Identified features were coded into four categories, namely, primary task support, dialogue support, credibility support, and social support.

[a]SEP: socioeconomic position.

Data Items

The predefined data items included target group, persuasive features, and theoretical and conceptual basis. A complete overview of all extracted variables is provided in Table 2.

Critical Appraisal of Individual Sources of Evidence

The methodological quality of studies was assessed using the JBI critical appraisal tools for RCTs and quasi-experimental studies. These checklists evaluate the extent to which a study minimizes bias in its design, conduct, and analysis. The RCT checklist contained 13 questions, while the quasi-experimental checklist had 9 questions. To ensure reliability, 10% of studies were independently screened by both the first and second authors, with discrepancies resolved through discussion. Papers were classified based on the approach used by Helmyati et al [37]: (1) low risk of bias (RCTs with 10 - 13 "yes" answers or quasi-experimental studies with 7 - 9), (2) medium risk of bias (RCTs with 7 - 9 "yes" answers or quasi-experimental studies with 5 - 6), and (3) high risk of bias (RCTs with 0 - 6 "yes" answers or quasi-experimental studies with 0 - 4). Critical appraisal was conducted to describe study quality and was not used as a basis for excluding studies.

Synthesis of Results

The extracted information was coded inductively to identify and group similar characteristics across interventions. To avoid overrepresenting interventions with multiple components, each characteristic was recorded only once per intervention, even if it was implemented in several ways (eg, personalization delivered via both email and app was coded as a single instance). To ensure consistent and reproducible categorization of target populations, interventions were classified as targeting low SEP families only when this was explicitly stated by the study authors. This criterion was chosen to minimize subjective inference and enable mutually exclusive grouping across studies. The synthesis was descriptive in nature. To address the second study objective, interventions evaluated using a quasi-experimental design were classified as more effective, less effective, or ineffective based on the criteria established by Morrison et al [38]. Using this framework, interventions were categorized as more effective if they led to improvements in most outcome measures and/or were at least as effective as comparison groups and more effective than no-intervention groups. Interventions were considered less effective if they led to improvements in only a minority of outcomes or were less effective than comparison groups but still outperformed waiting-list groups. Finally, interventions were classified as ineffective if they resulted in no measurable improvements or change. This categorization was conducted based on the outcome patterns reported in the included papers. No statistical reanalysis was performed, and the categorization is intended to provide an overview of reported patterns rather than to support causal conclusions or claims of statistical predominance.

XSL•FO
RenderX

PSD Features

Persuasive elements within the intervention design were analyzed using the PSD framework [28] to assess how digital and blended interventions encourage behavior change. The framework classifies persuasive features into four principles: primary task support, dialogue support, system credibility, and social support. PSD features were coded deductively using the framework. Even when authors did not explicitly identify intervention features as persuasive, they were documented if present based on the description of the interventions and available screenshots. A full list of PSD features and their respective categories is available in Multimedia Appendix 3.

Ethical Considerations

Ethical approval was not required for this scoping review because the study analyzed publicly available published literature and did not involve human participants or identifiable personal data.

Results

Selection of Sources of Evidence

The full search strategy, number of included papers, and reasons for exclusion are provided in Figure 1. The initial and updated search together yielded more than 13,000 records, from which 77 papers describing 54 interventions were included. The most common reasons for exclusion were targeting children outside the 6-month to 5-year age range, trial registrations, duplicate records, and a focus on breastfeeding or parental health (physical or mental).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analyses) flow diagram presenting identification and selection of papers for the scoping review of digital and blended lifestyle interventions for preschool-aged children. The search process consisted of an initial search (November 2023), an updated search (May 2025), and a second updated search (January 2026); numbers are combined for clarity.

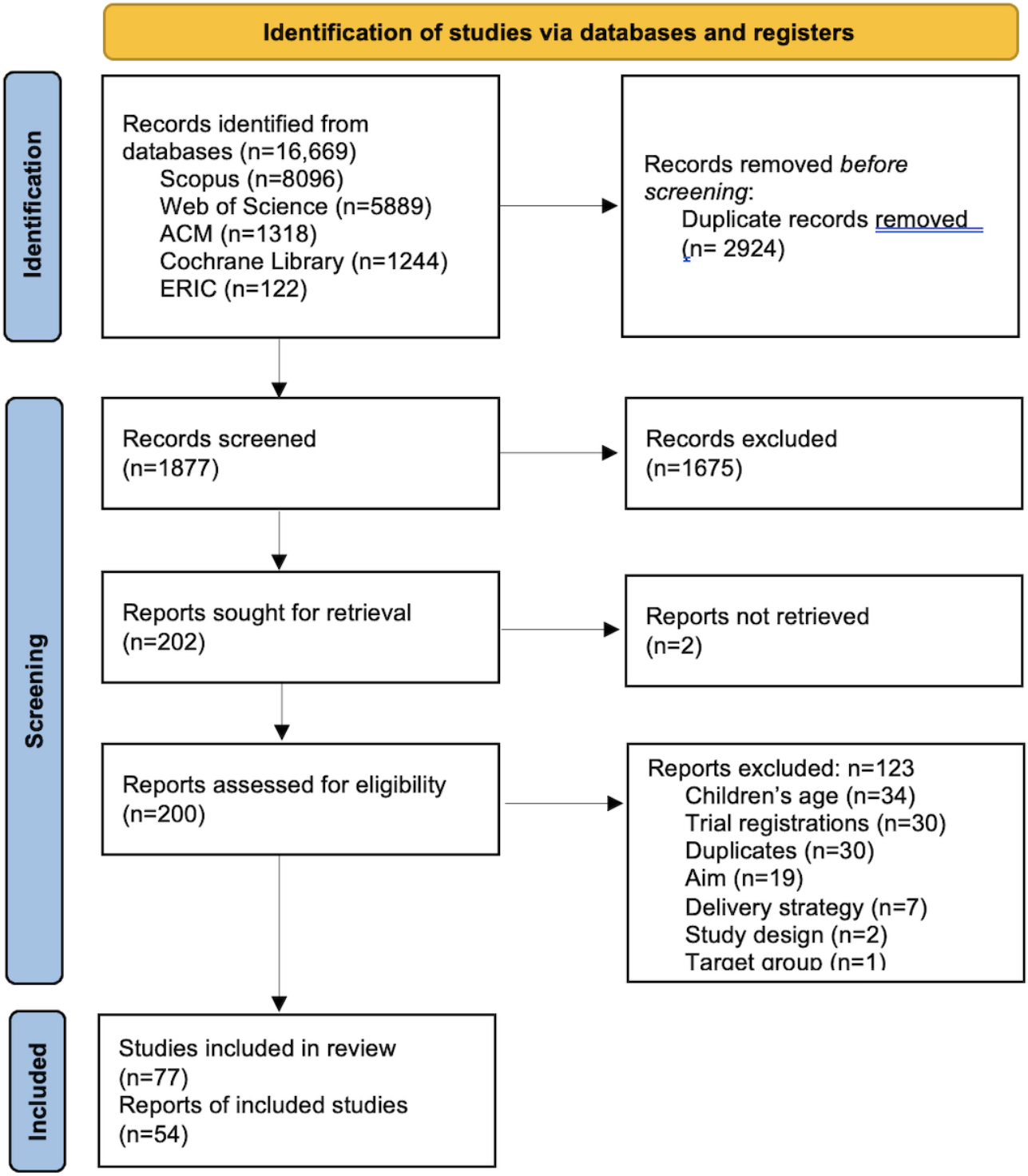

Characteristics of Sources of Evidence

An overview of the 77 included studies is provided in Multimedia Appendix 4, while Table 3 provides a summary of study and intervention characteristics across population groups. Of these, 34 (44%) were protocols for experimental studies, 35 (46%) were pre- and quasi-experimental studies, and 8 (10%) were feasibility, acceptability, and usability studies. The protocols primarily described intervention development and outlined the design and methodology of planned effectiveness trials. Among the 35 pre- and quasi-experimental studies, most used an RCT design (n=22, 63%), followed by cluster RCTs (n=7, 20%). These studies mainly evaluated intervention effects on targeted lifestyle behaviors and anthropometric outcomes, such as BMI.

XSL•FO
RenderX

Table . Overview of study and intervention characteristics across population groups (general population and low socioeconomic position [SEP]), including study design; target behaviors; target group; theories, frameworks, and evidence-based guidelines; strategy; channel; delivery environment; and persuasive features. Values are reported as numbers (n) and percentages (%). Counts represent how often a characteristic was reported or present across the included studies and interventions.

Intervention characteristic	General population, n (%)	Low SEP[a], n (%)	Total, n
Study design			
Protocols	25 (74)	9 (26)	34
Pre- and quasi-experimental studies	24 (67)	12 (33)	36
Feasibility, acceptability, and usability studies	4 (57)	3 (43)	7
Target behaviors			
Nutrition	11 (61)	7 (39)	18
Physical activity	1 (100)	0 (0)	1
Sedentary behavior	1 (100)	0 (0)	1
Sleep	3 (100)	0 (0)	3
Oral health	1 (100)	0 (0)	1
Combination	18 (60)	12 (40)	30
Target group			
Parents or caregivers	28 (67)	14 (33)	42
Others	7 (58)	5 (42)	12
Theories, frameworks, and evidence-based guidelines			
Evidence-based recommendations, guidelines, and previous research	16 (70)	7 (30)	23
Social cognitive theory	14 (78)	4 (22)	18
BCTs[b]	15 (83)	3 (17)	18
Nothing reported	3 (50)	3 (50)	6
Others	37 (70)	16 (30)	53
Strategy			
Digital	18 (64)	10 (36)	28
Blended	17 (65)	9 (35)	26
Channel			
Application-based components	12 (71)	5 (29)	17
Browser-based components	13 (72)	5 (28)	18
Human-delivered components	14 (61)	9 (39)	23
Voice or text messages	14 (58)	10 (42)	24
Emails or e-newsletters	2 (67)	1 (33)	3
Social media	5 (42)	7 (58)	12
Materials	6 (60)	4 (40)	10
Delivery environment			
Daily life	24 (67)	12 (33)	36
Others	11 (61)	7 (39)	18
PSD[c] features			
Primary task support	36 (78)	10 (22)	46
Dialogue support	31 (74)	11 (26)	42

XSL•FO
RenderX

Intervention characteristic	General population, n (%)	Low SEP[a], n (%)	Total, n
System credibility support	8 (80)	2 (20)	10
Social support	9 (53)	8 (47)	17

[a]SEP: socioeconomic position.

[b]BCT: behavior change technique.

[c]PSD: persuasive systems design.

Synthesis of Results

Overview of Intervention Characteristics

Overall, 54 interventions were included in the scoping review. Their key characteristics are summarized in Multimedia Appendix 5. The majority targeted families in the general population (n=35, 65%), while 19 (35%) interventions focused solely on low SEP families. A typical intervention, regardless of the target group, addressed multiple behavior categories, was parent-focused, informed by theories or evidence-based guidelines, delivered digitally, and integrated into daily-life contexts. For instance, Wu et al [39] (general population) and Lee et al [40] (low SEP) targeted nutrition and physical activity through digital channels such as social media apps, text messages, and websites to engage parents and caregivers in their daily routines. Both interventions were based on evidence-based recommendations, guidelines, and previous research or social cognitive theory. However, a few interventions differed from this general pattern. For example, the Barns matmot (version 2.0) intervention targeted only nutrition-related behaviors and involved kindergarten staff alongside caregivers, combining web components with in-person delivery, such as providing warm meals [41].

A gap map was created to support the descriptive synthesis of intervention characteristics across the general population and low SEP interventions (refer to Figure 2). Nutrition was the most frequently targeted domain in both groups. Physical activity and sedentary behavior were also common targets, whereas sleep and oral health were addressed less frequently, particularly in low SEP interventions. The complementary bar chart in Figure 3 illustrates the distribution of digital vs blended delivery strategies across the 2 population groups. Overall, digital interventions were slightly more common than blended ones, and the distribution was nearly equivalent within each population group.

Figure 2. Evidence gap map showing the distribution of targeted lifestyle behavior domains in digital and blended lifestyle interventions for preschool-aged children across population groups (general population and low socioeconomic position [SEP] families).

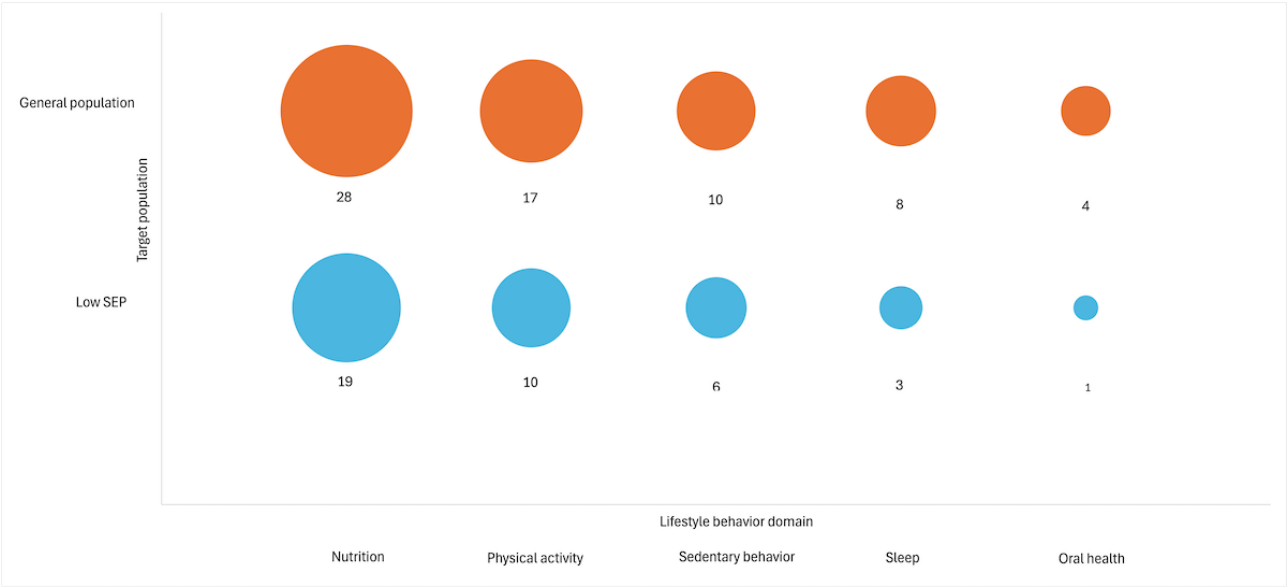

XSL•FO
RenderX

Figure 3. Distribution of included interventions by delivery mode (digital vs blended) stratified by population group (general population vs low socioeconomic position [SEP]).

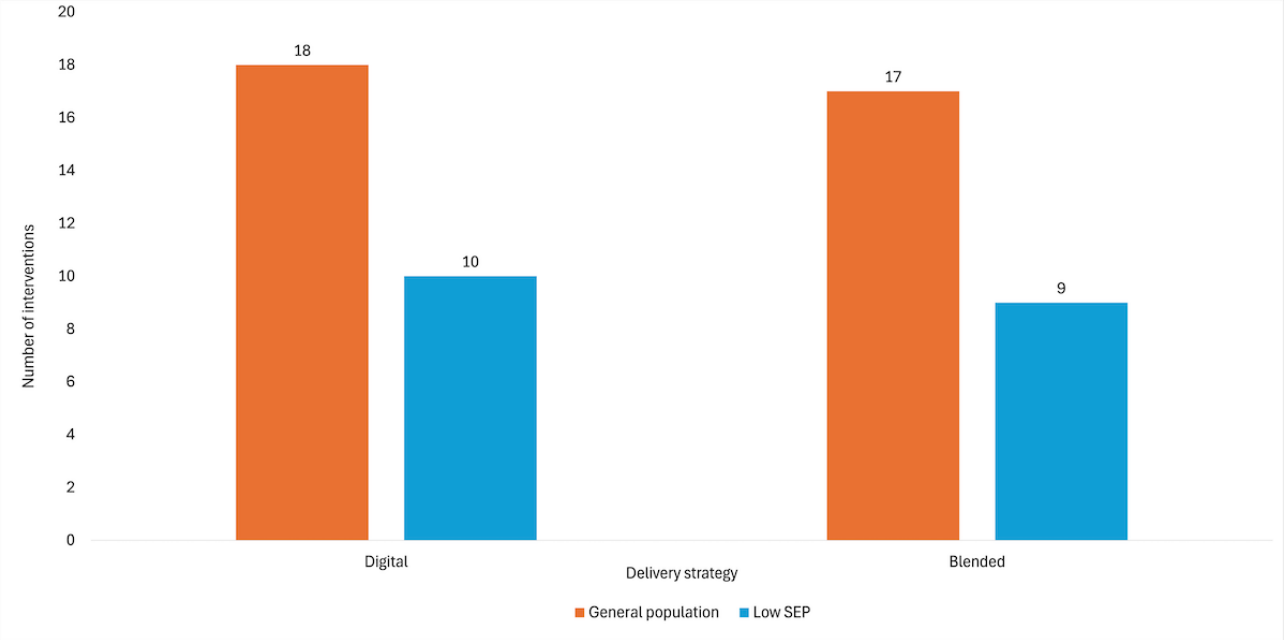

Target Behaviors

Many interventions in both groups targeted multiple behavior categories, ranging from 2 to 4 (general population: 18/35, 51% and low SEP: 12/19, 63%). Illustrative examples of how these target behaviors were presented within interventions are provided in Multimedia Appendix 6. For example, MINISTOP (version 2.0) [42] addressed nutrition, physical activity, sedentary behavior, sleep, and oral health in the general population, while Futuros Fuertes [43] focused on nutrition, sedentary behavior, and sleep in low SEP families. A total of 17 (49%) interventions in the general population group and 7 (37%) interventions in the low SEP group focused on a single behavior. Within single-behavior interventions, nutrition-related behaviors were the most common focus (general population: 11/17, 65% and low SEP: 7/7, 100%), with objectives such as reducing sugar-sweetened beverage intake, limiting unhealthy snacking, promoting fruit and vegetable consumption, and supporting responsive feeding practices. Only a few general population interventions exclusively addressed other behaviors, including sleep (n=3, 17%), oral health (n=1, 6%), physical activity (n=1, 6%), and sedentary behavior (n=1, 6%).

Target Group

In both groups, interventions primarily targeted parents as agents of behavior change for their children (general population: 28/35, 79% and low SEP: 15/19, 79%). A smaller number of interventions also involved other caregivers or intermediaries. These included health care professionals (general population: 1/35, 3%), early childhood educators (general population: 2/35, 6%), preschool staff together with parents or children (general population: 3/35, 9% and low SEP: 1/19, 5%), and multiple caregivers such as parents and grandparents (general population: 1/35, 3% and low SEP: 3/19, 16%). For example, the general population interventions Barns matmot 2.0 [41] and Go NAPSACC [44] engaged preschool staff alongside parents and children, while FUNS [45] targeted mothers, fathers, and grandparents in low SEP families.

Theories, Frameworks, and Evidence-Based Guidelines

The majority of interventions reported using theories, frameworks, and guidelines to inform their content and delivery (general population: 32/35, 91%, and low SEP: 16/19, 84%), with many incorporating multiple (general population: 22/32, 69%, and low SEP: 9/16, 56%). Among the reported approaches, evidence-based guidelines and previous research were the most frequently documented, forming the basis for intervention content and rationale in nearly half of all studies (general population: 16/32, 50%, and low SEP: 7/16, 44%). For instance, both MINISTOP (version 2.0; general population) and the Healthy Future Program (low SEP) relied on evidence-based guidelines from organizations such as the World Health Organization (WHO) and the American Academy of Pediatrics [42,46]. Social cognitive theory was the second most frequently cited framework (general population: 14/32, 44%, and low SEP: 4/16, 25%). The Greenlight Plus intervention [47] (general population), for example, combined social cognitive theory, active learning, and empirical evidence and guidelines, alongside behavior change techniques (BCTs) such as goal-setting and self-monitoring. Meanwhile, the GROW intervention [48] (low SEP) drew on the Centers for Disease Control and Prevention's theory and social cognitive theory and incorporated BCTs such as self-monitoring, problem solving, and goal-setting.

In addition to these broader frameworks, a subset of interventions reported using BCTs, drawing on frameworks such as the Coventry, Aberdeen, and London-Refined taxonomy [49], the behavior change wheel [50], or the BCTs taxonomy v1 [51] (general population n=13/32, 41%, and low SEP n=3/16, 19%). A total of 9 interventions (general population n=6/13, 46%, and low SEP n=3/3, 100%) reported on the BCTs they used. The most frequently used BCTs were goal-setting (general population n=4/6, 67%, and low SEP n=2/3, 67%),

XSL•FO
RenderX

self-monitoring (general population n=3/6, 50%, and low SEP n=2/3, 67%), and feedback (general population n=3/6, 50%, and low SEP n=1/3, 33%). For example, Go NAPSACC (general population) included both self-monitoring and feedback techniques [44] (refer to Multimedia Appendix 6).

Delivery Strategy and Channel

Slightly more than half of the interventions in both groups used a digital-only strategy (general population: 18/35, 51%, and low SEP: 10/19, 53%), while the other half adopted blended approaches, combining digital and nondigital components. General population interventions predominantly used multiple delivery channels (21/35, 60%), including websites, apps, online modules, and text messages, whereas low SEP interventions were almost evenly split between single (9/19, 47%) and multiple channels (10/19, 53%). In the low SEP group, commonly used digital tools included voice or text messages (10/19, 53%) and social media platforms (7/19, 37%) such as Facebook (Meta Platforms, Inc) groups and forums. For example, Food4Toddlers (general population) delivered content through a web-based platform and mobile app, providing parents with interactive modules, videos, recipes, and a discussion forum to support healthy food and eating environments in their daily routines [52]. Meanwhile, IIMAANJE (low SEP) or TEXT2COPE (general population; refer to Multimedia Appendix 6) used weekly voice and text messages to engage both mothers and fathers, offering practical guidance on infant and young child feeding practices that could easily be integrated into daily caregiving [53,54].

Among interventions using blended approaches, in-person components, such as counseling sessions, educational workshops with professionals or researchers, or delivery of content to children in classes, were common (general population: 14/35, 49%, and low SEP: 8/19, 47%). Other blended approaches provided additional physical materials, including pamphlets, measuring cups, water bottles, or magnets, to reinforce digital content and support behavior change in daily routines. The Ready, Set, Gulp! intervention by Lewis et al [55], for example, combined a smartphone app with a water promotion toolkit, including water bottles, stickers, and a booklet (refer to Multimedia Appendix 6).

Delivery Environment

Most interventions were delivered in a single environment, primarily daily life, for both the general population (24/35, 69%) and low SEP groups (12/19, 63%). For example, Early Food for Future Health [56] sent parents monthly emails linking to a secure website with videos and recipes designed for flexible, at-home viewing, while Samen Happie! [57] provided parents with an app containing age-based modules and interactive challenges to seamlessly integrate into daily routines. However, some interventions also incorporated additional delivery settings, including primary care clinics, daycare centers or preschools, community or organizational centers, or other unspecified locations for in-person sessions (refer to Multimedia Appendix 5).

Persuasive Features

None of the included studies explicitly reported using PSD features. However, 45 interventions (general population: 30/35, 86%, and low SEP: 15/19, 79%) provided sufficiently detailed descriptions to enable deductive coding based on the PSD framework [28] (refer to Multimedia Appendix 5). The number of PSD features implemented ranged from 1 to 7 in general population interventions (mode=3) and 1 to 6 in low SEP interventions, where a single feature was most common (mode=1). Multiple PSD features were more frequent in general population interventions (22/30, 73%), whereas low SEP interventions were almost evenly split between using one and multiple features (one: 7/15, 47%, and multiple: 8/15, 53%). For example, MINISTOP (version 2.0; general population) combined reminders, reduction, and rewards, while Cooking Matters (low SEP) relied solely on social learning to promote healthy behaviors [42,58]. Illustrative screenshots demonstrating how these PSD features were implemented are provided in Multimedia Appendix 6. Across both groups, the most commonly incorporated features were suggestions and reminders, which are both part of the dialogue support category (general population: 14/30, 47%, and low SEP: 5/15, 33%). For example, the Nenne Navi (general population) included suggestions by providing advice to the parent [59], while the TEXT2COPE (general population) intervention included reminders as part of its text messages [53]. Notably, low SEP interventions also frequently included social facilitation (8/15, 53%), which was exclusively implemented through social media components.

Effectiveness Categorization and Intervention Characteristics

Overall, 36 quasi-experimental studies were assessed for methodological quality using the JBI appraisal tools. The full appraisal details are provided in Multimedia Appendices 7 and 8. Out of these, 16 studies had a low risk of bias, 18 had a medium risk, and 2 studies had a high risk. The most frequent sources of bias were self-reported outcomes, lack of blinding (often not applicable to digital interventions), and absence of intention-to-treat analyses. No clear pattern was observed between study quality and effectiveness classifications: both low- and medium-risk studies included a mix of more effective, less effective, and ineffective interventions (low risk: more effective 7/16, 44%, less effective 7/16, 44%, and ineffective 2/16, 12%; medium risk: more effective 8/18, 44%, less effective 7/18, 39%, and ineffective 3/18, 17%). The 2 high-risk studies were classified as more effective and less effective. Importantly, methodological quality appraisal and effectiveness categorization represent distinct assessments: trial rigor was evaluated using JBI tools, whereas effectiveness reflects only the outcome patterns reported by the original studies.

A total of 36 interventions were assessed for their effectiveness in quasi-experimental trials, alongside 2 follow-up studies evaluating sustainability. As noted in the Methods section, this effectiveness categorization reflects descriptive outcome patterns reported in the original studies. No statistical reanalysis was performed, and the categorizations should not be interpreted as inferential evidence or claims of causal predominance. An

XSL•FO
RenderX

overview of the effectiveness categories per target population is provided in Table 4. Overall, interventions targeting the general population were more often categorized as more effective (13/24, 54%) compared to interventions targeting low SEP families (3/12, 25%). Meanwhile, a larger proportion of low SEP interventions were categorized as less effective (9/12, 75%) compared to general population interventions (6/24, 25%). Of the 2 follow-up studies assessing longer-term outcomes, neither MINISTOP [60] nor Early Food for Future Health [61] maintained effects at 12-month follow-up and were therefore coded as ineffective. No single study reported both short- and long-term effects within the same publication; long-term outcomes were only available through separate follow-up studies and were therefore classified independently.

Table . Effectiveness categorizations of included interventions stratified by target population (general population vs low socioeconomic position [SEP]), reported as number and percentage of interventions in each category.

Effectiveness categorization	General population, n (%)	Low SEP,[a] n (%)	Total, n (%)
More effective	12 (60)	3 (25)	15 (47)
Less effective	5 (25)	9 (75)	14 (44)
Ineffective	3 (15)	0 (0)	3 (9)
Total	20 (63)	12 (37)	32 (100)

[a]SEP: socioeconomic position.

When comparing intervention characteristics across the 3 effectiveness categories (more effective, less effective, and ineffective), few consistent descriptive patterns were observed. However, some differences were seen, particularly in interventions targeting low SEP families. For instance, all 3 more effective low SEP interventions (100%) each incorporated only a single PSD feature, with 2 relying exclusively on social learning. In contrast, less effective low SEP interventions were more frequently characterized by multiple PSD features (5/8, 63%). Among these, suggestions and reminders were the most common (3/8, 38%). For example, FirstStep2Health [62] used social learning through a Facebook-delivered program, virtual meetings, and text messages to engage parents and caregivers, while the intervention by Lee et al [40], which combined suggestions, reminders, and self-monitoring, was categorized as less effective. These observations represent an exploratory descriptive pattern and should not be interpreted as causal or prescriptive design conclusions.

Target behaviors appeared to vary descriptively across effectiveness categories. Less effective low SEP interventions more frequently targeted multiple behavior categories (6/9, 67%), whereas interventions in other effectiveness categories, including more effective low SEP interventions, more effective general population interventions, less effective general population interventions, and ineffective general population interventions, were more evenly split between single and multiple target behaviors. For example, Samen Happie! [57], Healthy Children, Strong Families 2 [63], and Futuros Fuertes [43] targeted multiple behaviors such as nutrition, sleep, sedentary behavior, and physical activity. Next, more effective low SEP interventions used blended approaches slightly more often (2/3, 67%). Less effective interventions were predominantly digital-only (6/9, 67%), often relying on voice or text messages (4/6, 67%). For example, communities for healthy living [64] used a blended approach combining education sessions, various materials, online components, and a Facebook group to support parents and caregivers in obesity prevention. In contrast, Cooking Matters [58], Downs et al [65], and NUTRES [66] relied solely on digital channels such as text, voice messages, and social media.

Feasibility, Acceptability, and Usability

Overall, 13 studies assessed feasibility, acceptability, or usability outcomes, including 8 targeting the general population (MINISTOP [version 2.0] [67], TEXT2COPE [53], Nenne Navi [68], Mini Movers [69], Lewis et al [55], He et al [70], Ferdous et al [71], and Ihab et al [72]) and 5 focusing on families from low socioeconomic backgrounds (FUNS [45], Samen Happie! [73], Downs et al [65], Happy Family, Healthy Kids [74], and Ling et al [75]). Regarding delivery, 5 interventions were fully digital and 8 used blended strategies that combined digital components with in-person or paper-based elements (refer to Table 3). Generally, both digital and blended programs were well received, though blended formats were more frequently evaluated and tended to yield more detailed insights.

Across studies, participants emphasized several positive aspects. Digital interventions were praised for their accessibility, convenience, and ease of integration into daily routines. MINISTOP (version 2.0) [67] and Nenne Navi [68] demonstrated high usability, cultural reach, and user satisfaction, while FUNS [45] showed strong preimplementation acceptability and cultural relevance among low SEP parents. Blended programs such as TEXT2COPE [53], Mini Movers [69], and Lewis et al [55] were valued for structured goal-setting, reminders, and opportunities for interaction and feedback. Interventions for low SEP families, including Happy Family, Healthy Kids [74] and Downs et al [65], achieved high acceptability when materials were culturally tailored and presented through accessible or low-literacy-friendly formats.

Challenges differed by population and context. For general population interventions, issues most often related to time or scheduling demands associated with interactive or in-person components [69-71] and the need for sustained motivation after initial enthusiasm [67,68]. Technical constraints, such as device and connectivity, occasionally affected blended delivery [71]. Among low SEP families, barriers were more structural, including limited phone ownership, unstable network access, and competing work or caregiving demands [65,74]. Engagement sometimes declined after initial uptake, as seen in Samen Happie! [73], despite moderate acceptability scores.

XSL•FO
RenderX

Most studies assessed feasibility and acceptability after participants had used the intervention within a pilot or trial context, focusing on user experience rather than real-world uptake. Only FUNS [45] and Ihab et al [72] examined these aspects formatively to inform later optimization. Additionally, Ling et al [75] reported that children continued using the mindfulness and movement activities after the intervention period, indicating sustained engagement with the child-directed components. Overall, the evidence suggests that interventions for families with young children are feasible and acceptable across populations, although positive evaluations during trials do not necessarily predict sustained use once research involvement or structured support ends.

Discussion

Summary of Evidence

This scoping review identified and compared the characteristics of digital and blended lifestyle interventions for preschool-aged children from 6 months to 5 years. Additionally, it examined differences between interventions for the general population and those targeting low SEP families. Across 54 interventions, characteristics were largely similar across populations and effectiveness categories, though more programs targeted the general population. A typical intervention addressed multiple behaviors (eg, nutrition, physical activity, and sleep) and centered on parents as primary agents of change. Most drew on evidence-based guidelines, such as the WHO recommendations on physical activity, sedentary behavior, and sleep for children aged <5 years [8], and theories such as social cognitive theory. Delivery mainly relied on digital channels, including websites, apps, and text messages. Overall, only minor descriptive tendencies were observed. General population interventions tended to include more persuasive features, whereas more effective low SEP interventions used fewer PSD features, and less effective low SEP interventions more often targeted multiple behavior categories. Since no consistent patterns emerged, observed effectiveness categories may relate to additional factors, such as how well interventions align with families' everyday contexts, routines, and resources. Although most interventions were considered feasible and acceptable in trial settings, these findings may not fully reflect their usability or sustainability in real-world conditions. Understanding these contextual influences, particularly in low SEP settings, might be essential for creating usable and sustainable interventions in practice. While contextual fit was not directly assessed in this review, health-behavior frameworks emphasize that families' engagement with interventions is shaped by everyday contextual conditions such as routines, competing demands, and environmental or structural constraints [50,76]. From this perspective, contextual fit offers a useful lens for interpreting the heterogeneous descriptive patterns observed in this review. However, this should be viewed as a theory-informed interpretation rather than an empirically tested determinant. Therefore, more research in which contextual fit is used as a determinant is necessary.

This review found that most digital and blended interventions for preschool-aged children target the general population, with limited focus on socioeconomically disadvantaged families. This imbalance echoes earlier observations: despite recognition of health inequalities, disadvantaged groups appear to remain underrepresented in digital health research [77]. It may seem counterintuitive, given that they face higher rates of preventable health issues and could particularly benefit from intervention efforts [13,78]. Limited focus may reflect persistent challenges such as recruitment and retention difficulties, funding priorities, and assumptions about digital access [25]. Yet, greater representation on its own may not be sufficient. Interventions developed without attention to families' everyday contexts risk offering limited benefit or even unintentionally reinforcing inequalities [79,80]. Our review also suggests that feasibility, acceptability, and usability studies have so far focused mainly on general population interventions. Only a small number have paid attention to these aspects in low SEP interventions, leaving little insight into how low SEP families might perceive and engage with such tools. Future research could therefore place more emphasis not only on improving recruitment, but also on systematically examining these families' experiences. Involving them early in development may help ensure that interventions are more likely to reflect lived realities and contribute to equitable outcomes. Co-design methods that actively engage these families early in the development process offer a promising avenue for incorporating their experiences and preferences and ensuring that interventions are contextually relevant.

Building on this, the review also considered intervention effectiveness. Interventions developed specifically for low SEP families were more often classified as less effective than those for the general population, though there were exceptions. Notably, some interventions originally designed for the general population appeared more effective in disadvantaged groups once adapted to families' preferences and contexts. For example, the MINISTOP intervention was initially rated as less effective [81]. A revised version (MINISTOP version 2.0), adapted and translated to be more culturally sensitive and accessible, was more effective in a multiethnic, socioeconomically diverse sample [42]. This may indicate that adapting existing programs can sometimes be a more efficient route than creating entirely new ones. At the same time, interventions focusing solely on individual behavior change, which was the case for most interventions in this review, may be overly optimistic, since families' choices are shaped by broader social and environmental circumstances. The social ecological model highlights how factors at multiple levels—individual, family, community, and policy—interact to influence health [76,82]. From this perspective, individual-level strategies alone may be insufficient in disadvantaged contexts, as suggested by prior research [83,84]. While some low SEP interventions in this review were categorized as less effective, these observations are descriptive and should not be interpreted as causal evidence. Many low SEP individuals describe their lifestyles as "logical" responses to structural constraints [15], underscoring the need to address surrounding conditions alongside individual choices. Future efforts could therefore focus on adapting the content and design to support individual behavior change more effectively by fitting families' contexts. Additionally, more attention could be paid to developing strategies at the community and policy levels to create conditions for sustainable impact. More

XSL•FO
RenderX

specifically, more evidence is needed on how community- and policy-level strategies can complement and be combined with family-based interventions to create more supportive environments for sustainable behavior change.

Most interventions in this review focused on parents or caregivers as agents of change, consistent with social cognitive theory, which emphasizes observational learning, modeling, and the reciprocal influence between parents' and children's behaviors. This approach is supported by evidence of parental influence on child health behaviors [85-87]. While parents play a crucial role, involving children more directly may help them internalize positive behaviors and build lasting habits. Developmental frameworks emphasize children's active participation in routines as critical for independence and self-regulation [88-90]. Digital tools offer ways to support such engagement. Applying principles from design thinking can make these tools more effective by grounding them in empathy, co-design, and iterative testing with children. Such approaches can lead to more engaging and developmentally appropriate experiences that may also foster a sense of ownership and sustained motivation [91-93]. Playful, interactive formats such as gamified apps have shown promise in early childhood settings. For example, Mole's Veggie Adventures engaged preschoolers and promoted healthier eating [94]. While these approaches may show promise, sustaining behavior change remains a well-recognized challenge [95,96], as also reflected in this review. Looking beyond the home to settings such as preschools or daycare centers can offer additional opportunities to embed interventions [97-99]. Some of the included studies in this review have already adopted blended formats, including preschools or clinic settings, with mixed results, suggesting that effectiveness may be influenced by implementation quality and contextual fit. Aligning components across home, care, and community settings may help reinforce habits over time [100]. Future research should therefore examine how collaboration among several stakeholders, such as parents, children, educators, health care professionals, and community leaders, can best support intervention sustainability and lasting impact.

Strengths and Limitations

A key strength of this review is its tailored search strategy. Because evidence on digital and blended lifestyle interventions for preschool-aged children is spread across multiple disciplines, we used a multidisciplinary search strategy and searched databases spanning public health, psychology, medicine, education, and technology (Scopus, Web of Science, Cochrane, ERIC, and ACM Digital Library). The search strategy was carefully developed with all authors and an information specialist and refined collaboratively, increasing the likelihood that relevant studies were captured from fragmented literature.

Several limitations should be noted. First, restricting the search to peer-reviewed literature may have introduced bias by excluding gray literature, such as program evaluations and organizational reports, which often provide valuable but inconsistently reported insights. Second, only a subset of records (200 papers during title and abstract screening and approximately 10% of full-text papers) were double-screened, though inclusion and exclusion decisions were discussed in detail to improve consistency. The use of AI-assisted screening (ASReview) may also have deprioritized relevant studies, and it cannot be ruled out that some were missed. Nevertheless, subsequent snowball searches of the reference lists of included papers did not yield many eligible studies, suggesting that the screening process likely captured the most relevant evidence. Third, the reporting of intervention components in the included studies, particularly PSD features, was often unclear, making it difficult to extract and code these elements. Inadequate reporting is a well-known issue in digital health research, limiting comparison of interventions, identification of success factors, and adaptation across populations. Although reporting guidelines such as CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) [101] improve transparency, they do not explicitly cover feature-level or design-specific elements such as persuasive features. Consistent reporting of these aspects remains essential to ensure that knowledge about digital health interventions is generalizable and replicable. Despite these challenges, applying the PSD framework here provided a structured lens for analyzing intervention design and helped provide a more systematic understanding of how persuasive features are used in child health interventions. Fourth, although a clear-cut criterion was necessary to categorize interventions consistently and reproducibly into general population and low SEP groups, this approach also has limitations. For example, classifying interventions as targeting low SEP only when this was explicitly stated by study authors may have underrepresented or misclassified interventions delivered in implicitly disadvantaged contexts.

Conclusion

In conclusion, this review contributes to the evidence base by summarizing digital and blended lifestyle interventions for preschool-aged children across multiple behavioral lifestyle domains, addressing the limited attention given to early childhood in digital health research. To our knowledge, it is also among the few reviews to compare interventions developed for the general population with those targeting families with low SEP, offering a descriptive perspective on similarities and differences across populations. Although only limited descriptive patterns emerged across intervention characteristics and effectiveness categories, the review nevertheless provides an overview of existing approaches and documents heterogeneity in characteristics and reporting. Moreover, it highlights areas where evidence remains sparse or unevenly distributed. These descriptive findings, combined with known barriers experienced by families with fewer resources, invite further reflection on how feasibility, acceptability, and contextual relevance may influence whether and how families are able to engage with interventions. Taken together, these insights point toward several directions for future work. low SEP families remain underrepresented in intervention development and evaluation, and greater attention is needed to understand feasibility, contextual fit, and acceptability from their perspectives. Approaches that consider multistakeholder and multilevel influences may better support families' opportunities to adopt healthy lifestyle behaviors. Additionally, children themselves

XSL•FO
RenderX

may need to be more directly involved in future intervention development. Sustainable and equitable progress may require complementing individual-focused strategies with broader structural and environmental supports. Future research could therefore examine how individual-level interventions can be embedded within wider community, service, and policy systems. Ultimately, advancing this field may depend on designing interventions that are contextually grounded, system-aware, and responsive to the diverse realities of families' everyday lives.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing, summarizing text, adapting and adjusting emotional tone, and reformatting. The GAI tool used was ChatGPT (OpenAI), including GPT-4 (used in 2024) and ChatGPT (version 5.2, used during 2025-2026). Exact minor versions were not recorded because of continuous platform updates. Responsibility for the final paper lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. The declaration was submitted by Lea Hohendorf.

Funding

This publication is part of the project "METAHEALTH: Health in a Microbial, Sociocultural and Care Context in the First 1000 Days of Life" (project NWA 1389.20.080), funded by the Dutch Research Agenda–Research along Routes by Consortia (NWA-ORC) 2020/21 and financed by the Dutch Research Council (NOW). The funding body had no role in the study design; data collection, analysis, or interpretation; or paper writing.

Data Availability

The data extracted and analyzed in this scoping review are openly available in the 4TU.ResearchData repository. The deposited materials include the full data extraction sheet, the list of included studies, and the list of excluded studies with reasons for exclusion. The dataset is accessible via the following DOI: 10.4121/e4152aeb-acd8-4ffb-8c8f-13e0ba001c48. No access restrictions apply.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full electronic search strategies for all databases used in the scoping review.
[DOCX File, 16 KB - jmir_v28i1e86596_app1.docx]

Multimedia Appendix 2
List of papers used as prior knowledge to train ASReview (Utrecht University).
[DOCX File, 19 KB - jmir_v28i1e86596_app2.docx]

Multimedia Appendix 3
Overview of persuasive systems design features.
[DOCX File, 19 KB - jmir_v28i1e86596_app3.docx]

Multimedia Appendix 4
Overview of the characteristics of included studies, including authors, publication year, study design and aim, intervention name, and quality assessment.
[DOCX File, 119 KB - jmir_v28i1e86596_app4.docx]

Multimedia Appendix 5
Overview of the characteristics of included interventions, including population; intervention name and goal; target behaviors; target group; theories, frameworks, and evidence-based guidelines; strategy and channel; delivery environment; persuasive features; and effectiveness.
[DOCX File, 45 KB - jmir_v28i1e86596_app5.docx]

Multimedia Appendix 6
Screenshots of included interventions.
[DOCX File, 2589 KB - jmir_v28i1e86596_app6.docx]

XSL•FO
RenderX

Multimedia Appendix 7
Overview of the critical appraisal of included randomized controlled trials.
[DOCX File, 59 KB - jmir_v28i1e86596_app7.docx]

Multimedia Appendix 8
Critical appraisal table for quasi-experimental studies.
[DOCX File, 23 KB - jmir_v28i1e86596_app8.docx]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 87 KB - jmir_v28i1e86596_app9.docx]

Checklist 2
PRISMA-S checklist.
[DOCX File, 17 KB - jmir_v28i1e86596_app10.docx]

References

1. Wang L, Jansen W, van Grieken A, et al. Identifying patterns of lifestyle behaviours among children of 3 years old. Eur J Public Health 2020 Dec 11;30(6):1115-1121. [doi: 10.1093/eurpub/ckaa109] [Medline: 32653911]
2. Dalal M, Cazorla-Lancaster Y, Chu CG, Agarwal N. Healthy from the start-lifestyle interventions in early childhood. Am J Lifestyle Med 2022;16(5):562-569. [doi: 10.1177/15598276221087672] [Medline: 36072692]
3. Decraene M, Verbestel V, Cardon G, et al. Compliance with the 24-hour movement behavior guidelines and associations with adiposity in European preschoolers: results from the ToyBox-study. Int J Environ Res Public Health 2021 Jul 14;18(14):7499. [doi: 10.3390/ijerph18147499] [Medline: 34299954]
4. Cliff DP, McNeill J, Vella SA, et al. Adherence to 24-hour movement guidelines for the early years and associations with social-cognitive development among Australian preschool children. BMC Public Health 2017 Nov 20;17(Suppl 5):857. [doi: 10.1186/s12889-017-4858-7] [Medline: 29219104]
5. De Craemer M, Lateva M, Iotova V, et al. Differences in energy balance-related behaviours in European preschool children: the ToyBox-study. In: Gillison F, editor. PLoS One 2015;10(3):e0118303. [doi: 10.1371/journal.pone.0118303] [Medline: 25785738]
6. Molinari AHW, Grap ME, Pierce SL, et al. Caregiver-reported sugar-sweetened beverage consumption and cavities in children aged 1 to 5 years, National Survey of Children's Health 2021-2022. Prev Chronic Dis 2025 Sep 11;22:E58. [doi: 10.5888/pcd22.250183] [Medline: 40934341]
7. Uribe SE, Innes N, Maldupa I. The global prevalence of early childhood caries: a systematic review with meta-analysis using the WHO diagnostic criteria. Int J Paediatr Dent 2021 Nov;31(6):817-830. [doi: 10.1111/ipd.12783] [Medline: 33735529]
8. Guidelines on physical activity, sedentary behaviour and sleep for children under 5 years of age. : World Health Organization; 2019 URL: https://www.who.int/publications/i/item/9789241550536 [accessed 2025-06-20]
9. Noncommunicable diseases: childhood overweight and obesity. World Health Organization. 2025. URL: https://www.who.int/news-room/questions-and-answers/item/noncommunicable-diseases-childhood-overweight-and-obesity [accessed 2025-06-20]
10. Oral health. World Health Organization. 2025. URL: https://www.who.int/news-room/fact-sheets/detail/oral-health [accessed 2025-06-20]
11. Physical activity guidelines for children (under 5 years). National Health Service. 2022. URL: https://www.nhs.uk/live-well/exercise/physical-activity-guidelines-children-under-five-years/ [accessed 2025-06-20]
12. Wu X, Ohinmaa A, Veugelers PJ, Maximova K. Health-related quality of life in relation to health behaviour patterns among Canadian children. Children (Basel) 2024 Mar 14;11(3):346. [doi: 10.3390/children11030346] [Medline: 38539381]
13. Hermeling L, Steinacker JM, Kobel S. Beyond correlates: the social gradient in childhood overweight. Arch Public Health 2024 Jan 9;82(1):3. [doi: 10.1186/s13690-023-01232-x] [Medline: 38195594]
14. Socioeconomic status. American Psychological Association. 2025. URL: https://www.apa.org/topics/socioeconomic-status/ [accessed 2025-06-20]
15. Bukman AJ, Teuscher D, Feskens EJM, van Baak MA, Meershoek A, Renes RJ. Perceptions on healthy eating, physical activity and lifestyle advice: opportunities for adapting lifestyle interventions to individuals with low socioeconomic status. BMC Public Health 2014 Oct 4;14(1):1036. [doi: 10.1186/1471-2458-14-1036] [Medline: 25280579]
16. Mantziki K, Renders CM, Vassilopoulos A, et al. Inequalities in energy-balance related behaviours and family environmental determinants in European children: changes and sustainability within the EPHE evaluation study. Int J Equity Health 2016 Sep 29;15(1):160. [doi: 10.1186/s12939-016-0438-1] [Medline: 27680964]

XSL•FO
RenderX

17. Whitehead L, Talevski J, Fatehi F, Beauchamp A. Barriers to and facilitators of digital health among culturally and linguistically diverse populations: qualitative systematic review. J Med Internet Res 2023 Feb 28;25:e42719. [doi: 10.2196/42719] [Medline: 36853742]
18. Al-Dhahir I, Reijnders T, Faber JS, et al. The barriers and facilitators of eHealth-based lifestyle intervention programs for people with a low socioeconomic status: scoping review. J Med Internet Res 2022 Aug 24;24(8):e34229. [doi: 10.2196/34229] [Medline: 36001380]
19. Tomayko EJ, Tovar A, Fitzgerald N, et al. Parent involvement in diet or physical activity interventions to treat or prevent childhood obesity: an umbrella review. Nutrients 2021 Sep 16;13(9):3227. [doi: 10.3390/nu13093227] [Medline: 34579099]
20. Kinsner K, Parlakian R, MacLaughlin S. Grandparents who care: literature review. : ZERO TO THREE; 2017 URL: https://www.zerotothree.org/wp-content/uploads/2022/09/Grandparents-Who-Care_-A-Literature-Review.pdf [accessed 2025-06-20]
21. McKeon G, Mastrogiovanni C, Teychenne M, Rosenbaum S. Barriers and facilitators to participating in an exercise referral scheme among women living in a low socioeconomic area in Australia: a qualitative investigation using the COM-B and theoretical domains framework. Int J Environ Res Public Health 2022 Sep 28;19(19):12312. [doi: 10.3390/ijerph191912312] [Medline: 36231623]
22. Oudbier SJ, Souget-Ruff SP, Chen BSJ, Ziesemer KA, Meij HJ, Smets EMA. Implementation barriers and facilitators of remote monitoring, remote consultation and digital care platforms through the eyes of healthcare professionals: a review of reviews. BMJ Open 2024 Jun 10;14(6):e075833. [doi: 10.1136/bmjopen-2023-075833] [Medline: 38858155]
23. Lustria MLA, Noar SM, Cortese J, Van Stee SK, Glueckauf RL, Lee J. A meta-analysis of web-delivered tailored health behavior change interventions. J Health Commun 2013;18(9):1039-1069. [doi: 10.1080/10810730.2013.768727] [Medline: 23750972]
24. Hengst TM, Lechner L, Dohmen D, Bolman CA. The facilitators and barriers of mHealth adoption and use among people with a low socio-economic position: a scoping review. Digit Health 2023;9:20552076231198702. [doi: 10.1177/20552076231198702] [Medline: 37691766]
25. Borghouts J, Eikey E, Mark G, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387. [doi: 10.2196/24387] [Medline: 33759801]
26. Santarossa S, Kane D, Senn CY, Woodruff SJ. Exploring the role of in-person components for online health behavior change interventions: can a digital person-to-person component suffice? J Med Internet Res 2018 Apr 11;20(4):e144. [doi: 10.2196/jmir.8480] [Medline: 29643048]
27. Kelders SM, Kok RN, Ossebaard HC, Van Gemert-Pijnen JEWC. Persuasive system design does matter: a systematic review of adherence to web-based interventions. J Med Internet Res 2012 Nov 14;14(6):e152. [doi: 10.2196/jmir.2104] [Medline: 23151820]
28. Oinas-Kukkonen H, Harjumaa M. Persuasive systems design: key issues, process model, and system features. Commun Assoc Inf Syst 2009;24. [doi: 10.17705/1CAIS.02428]
29. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020 Oct;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
30. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
31. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5(1):69. [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
32. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
33. Active learning explained. ASReview. 2022. URL: https://asreview.nl/blog/active-learning-explained/ [accessed 2025-06-20]
34. van de Schoot R, de Bruin J, Schram R, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell 2021 Feb;3(2):125-133. [doi: 10.1038/s42256-020-00287-7]
35. Boetje J, van de Schoot R. The SAFE procedure: a practical stopping heuristic for active learning-based screening in systematic reviews and meta-analyses. Syst Rev 2024 Mar 1;13(1):81. [doi: 10.1186/s13643-024-02502-7] [Medline: 38429798]
36. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]
37. Helmyati S, Lusmilasari L, Sandhi A, et al. Systematic review on supplementation, fortification, and food-based interventions for preventing iron deficiency anemia in low- and middle-income countries. Asia Pac J Clin Nutr 2025 Feb;34(1):10-35. [doi: 10.6133/apjcn.202502_34(1).0002] [Medline: 39828255]
38. Morrison LG, Yardley L, Powell J, Michie S. What design features are used in effective e-health interventions? A review using techniques from critical interpretive synthesis. Telemed J E Health 2012 Mar;18(2):137-144. [doi: 10.1089/tmj.2011.0062] [Medline: 22381060]
39. Wu Q, Wang X, Zhang J, Zhang Y, van Velthoven MH. The effectiveness of a WeChat-based self-assessment with a tailored feedback report on improving complementary feeding and movement behaviour of children aged 6-20 months in rural China: a cluster randomized controlled trial. Lancet Reg Health West Pac 2023 Aug;37:100796. [doi: 10.1016/j.lanwpc.2023.100796] [Medline: 37273963]

XSL•FO
RenderX

40. Lee H, Oldewage-Theron W, Dawson J. Effects of a theory-based, multicomponent eHealth intervention for obesity prevention in young children from low-income families: a pilot randomized controlled study. Nutrients 2023 May 13;15(10):2296. [doi: 10.3390/nu15102296]
41. Blomkvist EAM, Helland SH, Hillesund ER, Øverby NC. A cluster randomized web-based intervention trial to reduce food neophobia and promote healthy diets among one-year-old children in kindergarten: study protocol. BMC Pediatr 2018 Jul 14;18(1):232. [doi: 10.1186/s12887-018-1206-8] [Medline: 30007401]
42. Alexandrou C, Henriksson H, Henström M, et al. Effectiveness of a smartphone app (MINISTOP 2.0) integrated in primary child health care to promote healthy diet and physical activity behaviors and prevent obesity in preschool-aged children: randomized controlled trial. Int J Behav Nutr Phys Act 2023 Feb 21;20(1):22. [doi: 10.1186/s12966-023-01405-5] [Medline: 36810069]
43. Beck AL, Mora R, Joseph G, et al. A multimethod evaluation of the Futuros Fuertes intervention to promote healthy feeding, screen time, and sleep practices. Acad Pediatr 2023;23(7):1351-1360. [doi: 10.1016/j.acap.2023.05.006] [Medline: 37211275]
44. Willis EA, Burney R, Hales D, et al. “My wellbeing-their wellbeing”- an eHealth intervention for managing obesity in early care and education: protocol for the Go NAPSACC cares cluster randomized control trial. In: Shen L, editor. PLoS One 2023;18(7):e0286912. [doi: 10.1371/journal.pone.0286912] [Medline: 37418363]
45. Ramírez AS, Ayala GX, Murillo M, Glik DC, Guerrero AD. Integrating theory with a user-centered design approach to maximize mHealth acceptability and usability. Health Educ Behav 2025 Jun;52(3):329-339. [doi: 10.1177/10901981241311232] [Medline: 39772995]
46. Chen Y, Wu Y, Dill SE, et al. Effect of the mHealth-supported healthy future programme delivered by community health workers on maternal and child health in rural China: study protocol for a cluster randomised controlled trial. BMJ Open 2023 Jan 20;13(1):e065403. [doi: 10.1136/bmjopen-2022-065403] [Medline: 36669837]
47. Heerman WJ, Rothman RL, Sanders LM, et al. A digital health behavior intervention to prevent childhood obesity: the greenlight plus randomized clinical trial. JAMA 2024 Dec 24;332(24):2068-2080. [doi: 10.1001/jama.2024.22362] [Medline: 39489149]
48. Po’e EK, Heerman WJ, Mistry RS, Barkin SL. Growing right onto wellness (GROW): a family-centered, community-based obesity prevention randomized controlled trial for preschool child-parent pairs. Contemp Clin Trials 2013 Nov;36(2):436-449. [doi: 10.1016/j.cct.2013.08.013] [Medline: 24012890]
49. Michie S, Ashford S, Sniehotta FF, Dombrowski SU, Bishop A, French DP. A refined taxonomy of behaviour change techniques to help people change their physical activity and healthy eating behaviours: the CALO-RE taxonomy. Psychol Health 2011 Nov;26(11):1479-1498. [doi: 10.1080/08870446.2010.540664] [Medline: 21678185]
50. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6(1):42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
51. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
52. Røed M, Hillesund ER, Vik FN, Van Lippevelde W, Øverby NC. The Food4toddlers study - study protocol for a web-based intervention to promote healthy diets for toddlers: a randomized controlled trial. BMC Public Health 2019 May 14;19(1):563. [doi: 10.1186/s12889-019-6915-x] [Medline: 31088438]
53. Militello L, Melnyk BM, Hekler EB, Small L, Jacobson D. Automated behavioral text messaging and face-to-face intervention for parents of overweight or obese preschool children: results from a pilot study. JMIR Mhealth Uhealth 2016 Mar 14;4(1):e21. [doi: 10.2196/mhealth.4398] [Medline: 26976387]
54. Downs SM, Gueye D, Sall M, et al. The impact and implementation of an mHealth intervention to improve infant and young child feeding in Senegal: IIMAANJE protocol for a cluster randomized control trial. Front Public Health 2023;11:1258963. [doi: 10.3389/fpubh.2023.1258963] [Medline: 37818304]
55. Lewis KH, Hsu FC, Block JP, et al. A technology-driven, healthcare-based intervention to improve family beverage choices: results from a pilot randomized trial in the United States. Nutrients 2023 Apr 29;15(9):2141. [doi: 10.3390/nu15092141] [Medline: 37432293]
56. Helle C, Hillesund ER, Wills AK, Øverby NC. Evaluation of an eHealth intervention aiming to promote healthy food habits from infancy-the Norwegian randomized controlled trial Early Food for Future Health. Int J Behav Nutr Phys Act 2019 Jan 3;16(1):1. [doi: 10.1186/s12966-018-0763-4] [Medline: 30606197]
57. Karssen LT, Vink JM, de Weerth C, et al. An app-based parenting program to promote healthy energy balance-related parenting practices to prevent childhood obesity: protocol using the intervention mapping framework. JMIR Form Res 2021 May 14;5(5):e24802. [doi: 10.2196/24802] [Medline: 33988510]
58. Zhang Q, Panichelli J, Hall LA. Assessment of cooking matters Facebook platform to promote healthy eating behaviors among low-income caregivers of young children in the United States: a pilot study. Nutrients 2021 Aug 4;13(8):2694. [doi: 10.3390/nu13082694] [Medline: 34444854]
59. Yoshizaki A, Murata E, Yamamoto T, et al. Improving children’s sleep habits using an interactive smartphone app: community-based intervention study. JMIR Mhealth Uhealth 2023 Feb 10;11:e40836. [doi: 10.2196/40836] [Medline: 36641237]

XSL•FO
RenderX

60. Delisle Nyström C, Sandin S, Henriksson P, Henriksson H, Maddison R, Löf M. A 12-month follow-up of a mobile-based (mHealth) obesity prevention intervention in pre-school children: the MINISTOP randomized controlled trial. BMC Public Health 2018 May 24;18(1):658. [doi: 10.1186/s12889-018-5569-4] [Medline: 29793467]
61. Helle C, Hillesund ER, Wills AK, Øverby NC. Examining the effects of an eHealth intervention from infant age 6 to 12 months on child eating behaviors and maternal feeding practices one year after cessation: the Norwegian randomized controlled trial Early Food for Future Health. In: Simeoni U, editor. PLoS One 2019;14(8):e0220437. [doi: 10.1371/journal.pone.0220437] [Medline: 31442241]
62. Ling J, Suriyawong W, Robbins LB, Zhang N, Kerver JM. FirstStep2Health: a cluster randomised trial to promote healthy behaviours and prevent obesity amongst low-income preschoolers. Pediatr Obes 2024 Jul;19(7):e13122. [doi: 10.1111/ijpo.13122] [Medline: 38622494]
63. Tomayko EJ, Prince RJ, Cronin KA, et al. Healthy children, strong families 2: a randomized controlled trial of a healthy lifestyle intervention for American Indian families designed using community-based approaches. Clin Trials 2017 Apr;14(2):152-161. [doi: 10.1177/1740774516685699] [Medline: 28064525]
64. Beckerman JP, Aftosmes-Tobio A, Kitos N, et al. Communities for healthy living (CHL) - a family-centered childhood obesity prevention program integrated into Head Start services: study protocol for a pragmatic cluster randomized trial. Contemp Clin Trials 2019 Mar;78:34-45. [doi: 10.1016/j.cct.2019.01.002] [Medline: 30630109]
65. Downs SM, Sackey J, Kalaj J, Smith S, Fanzo J. An mHealth voice messaging intervention to improve infant and young child feeding practices in Senegal. Matern Child Nutr 2019 Oct;15(4):e12825. [doi: 10.1111/mcn.12825]
66. Lozada-Tequeanes AL, Théodore FL, Kim-Herrera E, et al. Effectiveness and implementation of a text messaging mHealth intervention to prevent childhood obesity in Mexico in the COVID-19 context: mixed methods study. JMIR Mhealth Uhealth 2024 Apr 9;12:e55509. [doi: 10.2196/55509] [Medline: 38592753]
67. Alexandrou C, Rutberg S, Johansson L, Lindqvist AK, Müssener U, Löf M. User experiences of an app-based mHealth intervention (MINISTOP 2.0) integrated in Swedish primary child healthcare among Swedish-, Somali- and Arabic-speaking parents and child healthcare nurses: a qualitative study. Digit Health 2023;9:20552076231203630. [doi: 10.1177/20552076231203630] [Medline: 37766904]
68. Yoshizaki A, Mohri I, Yamamoto T, et al. An interactive smartphone app, Nenne Navi, for improving children's sleep: pilot usability study. JMIR Pediatr Parent 2020 Dec 1;3(2):e22102. [doi: 10.2196/22102] [Medline: 33122163]
69. Downing KL, Salmon J, Hinkley T, Hnatiuk JA, Hesketh KD. Feasibility and efficacy of a parent-focused, text message-delivered intervention to reduce sedentary behavior in 2- to 4-year-old children (mini movers): pilot randomized controlled trial. JMIR Mhealth Uhealth 2018 Feb 9;6(2):e39. [doi: 10.2196/mhealth.8573] [Medline: 29426816]
70. He Q, Ha ASC, Zheng B, Okely AD. Feasibility and potential efficacy of a family-based intervention on promoting physical activity levels and fundamental movement skills in preschoolers: a cluster randomised controlled trial. Appl Psychol Health Well Being 2024 Aug;16(3):1266-1288. [doi: 10.1111/aphw.12527] [Medline: 38240635]
71. Ferdous TE, Jaman MJ, Siddique AB, et al. Feasibility of employing mHealth in delivering preventive nutrition interventions targeting the first 1000 days of life: experiences from a community-based cluster randomised trial in rural Bangladesh. Nutrients 2024 Oct 10;16(20):3429. [doi: 10.3390/nu16203429] [Medline: 39458424]
72. Ihab M, Abdelaziz W, Hassan W, El Tantawi M. Development and acceptability of behavioral interventions promoting mothers' brushing of pre-school children's teeth: the preparation phase of the multi-phase optimization strategy framework. BMC Oral Health 2023 Aug 31;23(1):616. [doi: 10.1186/s12903-023-03351-x] [Medline: 37653476]
73. Karssen LT, Larsen JK, Burk WJ, et al. Process and effect evaluation of the app-based parenting program Samen Happie! on infant zBMI: a randomized controlled trial. Front Public Health 2022;10:1012431. [doi: 10.3389/fpubh.2022.1012431] [Medline: 36620259]
74. Chen S, Ling J, Buhlman R, Tadavich S, Kao TSA. Acceptability and satisfaction of a mindfulness-based healthy eating and stress management program targeting economically marginalized families in a pilot trial. J Pediatr Psychol 2025 May 1;50(5):402-411. [doi: 10.1093/jpepsy/jsaf010] [Medline: 40096702]
75. Ling J, Ashley A, Zahry N, et al. A mindfulness-based lifestyle intervention among economically marginalized caregiver-preschooler dyads: feasibility, acceptability, and satisfaction. School Ment Health 2025;17(2):715-732. [doi: 10.1007/s12310-025-09767-w] [Medline: 40656841]
76. McLeroy KR, Bibeau D, Steckler A, Glanz K. An ecological perspective on health promotion programs. Health Educ Q 1988;15(4):351-377. [doi: 10.1177/109019818801500401] [Medline: 3068205]
77. König LM, Western MJ, Denton AH, Krukowski RA. Umbrella review of social inequality in digital interventions targeting dietary and physical activity behaviors. NPJ Digit Med 2025 Jan 6;8(1):11. [doi: 10.1038/s41746-024-01405-0] [Medline: 39762352]
78. Libuy N, Bann D, Fitzsimons E. Inequalities in body mass index, diet and physical activity in the UK: longitudinal evidence across childhood and adolescence. SSM Popul Health 2021 Dec;16:100978. [doi: 10.1016/j.ssmph.2021.100978] [Medline: 34950761]
79. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]

XSL•FO
RenderX

80. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. J Am Med Inform Assoc 2018 Aug 1;25(8):1080-1088. [doi: 10.1093/jamia/ocy052] [Medline: 29788380]
81. Nyström CD, Sandin S, Henriksson P, et al. Mobile-based intervention intended to stop obesity in preschool-aged children: the MINISTOP randomized controlled trial. Am J Clin Nutr 2017 Jun;105(6):1327-1335. [doi: 10.3945/ajcn.116.150995] [Medline: 28446496]
82. Golden SD, Earp JAL. Social ecological approaches to individuals and their contexts: twenty years of health education & behavior health promotion interventions. Health Educ Behav 2012 Jun;39(3):364-372. [doi: 10.1177/1090198111418634] [Medline: 22267868]
83. Hillier-Brown FC, Bambra CL, Cairns JM, Kasim A, Moore HJ, Summerbell CD. A systematic review of the effectiveness of individual, community and societal-level interventions at reducing socio-economic inequalities in obesity among adults. Int J Obes (Lond) 2014 Dec;38(12):1483-1490. [doi: 10.1038/ijo.2014.75] [Medline: 24813369]
84. Adams J, Mytton O, White M, Monsivais P. Why are some population interventions for diet and obesity more equitable and effective than others? The role of individual agency. PLoS Med 2016 Apr;13(4):e1001990. [doi: 10.1371/journal.pmed.1001990] [Medline: 27046234]
85. Bandura A. Social Foundations of Thought and Action: A Social Cognitive Theory: Pearson; 1986.
86. Yee AZH, Lwin MO, Ho SS. The influence of parental practices on child promotive and preventive food consumption behaviors: a systematic review and meta-analysis. Int J Behav Nutr Phys Act 2017 Apr 11;14(1):47. [doi: 10.1186/s12966-017-0501-3] [Medline: 28399881]
87. Vaughn AE, Ward DS, Fisher JO, et al. Fundamental constructs in food parenting practices: a content map to guide future research. Nutr Rev 2016 Feb;74(2):98-117. [doi: 10.1093/nutrit/nuv061] [Medline: 26724487]
88. Vygotsky LS. Mind in Society: The Development of Higher Psychological Processes: Harvard University Press; 1978.
89. Montessori M. The Montessori Method: Scientific Pedagogy as Applied to Child Education in “The Children’s Houses” with Additions and Revisions by the Author: Stokes Company; 1912. [doi: 10.1037/13054-000]
90. Ling J, Robbins LB, Wen F, Zhang N. Lifestyle interventions in preschool children: a meta-analysis of effectiveness. Am J Prev Med 2017 Jul;53(1):102-112. [doi: 10.1016/j.amepre.2017.01.018] [Medline: 28237633]
91. Scholten H, Granic I. Use of the principles of design thinking to address limitations of digital mental health interventions for youth: viewpoint. J Med Internet Res 2019 Jan 14;21(1):e11528. [doi: 10.2196/11528] [Medline: 31344671]
92. Voorheis P, Zhao A, Kuluski K, et al. Integrating behavioral science and design thinking to develop mobile health interventions: systematic scoping review. JMIR Mhealth Uhealth 2022 Mar 16;10(3):e35799. [doi: 10.2196/35799] [Medline: 35293871]
93. Van Mechelen M, Laenen A, Zaman B, Willems B, Abeele VV. Collaborative design thinking (CoDeT): a co-design approach for high child-to-adult ratios. Int J Hum Comput Stud 2019 Oct;130:179-195. [doi: 10.1016/j.ijhcs.2019.06.013]
94. Vepsäläinen H, Skaffari E, Wojtkowska K, et al. A mobile app to increase fruit and vegetable acceptance among Finnish and Polish preschoolers: randomized trial. JMIR Mhealth Uhealth 2022 Jan 4;10(1):e30352. [doi: 10.2196/30352] [Medline: 34982718]
95. Bailey D, Duncan GJ, Odgers CL, Yu W. Persistence and fadeout in the impacts of child and adolescent interventions. J Res Educ Eff 2017;10(1):7-39. [doi: 10.1080/19345747.2016.1232459] [Medline: 29371909]
96. Jeong J, Pitchik HO, Fink G. Short-term, medium-term and long-term effects of early parenting interventions in low- and middle-income countries: a systematic review. BMJ Glob Health 2021 Mar;6(3):e004067. [doi: 10.1136/bmjgh-2020-004067] [Medline: 33674266]
97. Lum M, Wolfenden L, Jones J, et al. Interventions to improve child physical activity in the early childhood education and care setting: an umbrella review. Int J Environ Res Public Health 2022 Feb 10;19(4):1963. [doi: 10.3390/ijerph19041963] [Medline: 35206152]
98. Neshteruk C, Burkart S, Flanagan EW, Melnick E, Luecking C, Kracht CL. Policy, systems, and environmental interventions addressing physical activity in early childhood education settings: a systematic review. Prev Med 2023 Aug;173:107606. [doi: 10.1016/j.ypmed.2023.107606] [Medline: 37414226]
99. Barnes C, Yoong SL, Nathan N, et al. Feasibility of a web-based implementation intervention to improve child dietary intake in early childhood education and care: pilot randomized controlled trial. J Med Internet Res 2021 Dec 15;23(12):e25902. [doi: 10.2196/25902] [Medline: 34914617]
100. Lally P, Gardner B. Promoting habit formation. Health Psychol Rev 2013 May;7(sup1):S137-S158. [doi: 10.1080/17437199.2011.603640]
101. Eysenbach G, Group CE. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]

Abbreviations

AI: artificial intelligence
BCT: behavior change technique

XSL•FO
RenderX

CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
JBI: Joanna Briggs Institute
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
PSD: persuasive systems design
RCT: randomized controlled trial
SEP: socioeconomic position
WHO: World Health Organization

Edited by S Brini; submitted 27.Oct.2025; peer-reviewed by MBR Mahmoud, T Baranowski; revised version received 10.Apr.2026; accepted 11.Apr.2026; published 05.Jun.2026.

Please cite as:
Hohendorf L, Dekkers T, Kip H, Alpay L, Kelders SM
Digital and Blended Lifestyle Interventions for Preschool-Aged Children and Families With a Low Socioeconomic Position and the General Population: Scoping Review
J Med Internet Res 2026;28:e86596
URL: https://www.jmir.org/2026/1/e86596
doi:10.2196/86596

© Lea Hohendorf, Tessa Dekkers, Hanneke Kip, Laurence Alpay, Saskia M Kelders. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Artificial Intelligence Governance in Health Systems: Systematic Review of Frameworks and Integrative Model Proposal

Hassane Alami[1,2], MSc PH, MSc HP, PhD; Renata Pozelli Sabio[1], MSc, PhD; Elsury Johanna Pérez[1], MSc, PhD; Marie-Pierre Gagnon[3], MSc, PhD; Lyse Langlois[4], PhD; Jean-Louis Denis[5], MSc, PhD; Kathy Malas[1,6], MPO; Lysanne Rivard[1], MA, PhD; Mathilde Savoldelli[1,7], MSc, DrPHc; Mohamed Ali Ag Ahmed[1,8], MD, MPH, PhD; Jean-Paul Fortin[9], MD, MPH, MBA

[1]Department of Health Management, Evaluation and Policy, School of Public Health & Center for Public Health Research of the University of Montreal, P.O. Box 6128, Branch Centre-Ville, Montreal, QC, Canada

[2]Institute for Data Valorization (IVADO), Montreal, QC, Canada

[3]Faculty of Nursing, Laval University, Quebec, QC, Canada

[4]Department of Industrial Relations, Laval University, Quebec, QC, Canada

[5]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada

[6]Integrated Health and Social Services University Network for West-Central Montreal (CIUSSS West-Central Montreal), Montreal, QC, Canada

[7]Territorial Professional Health Community, Lauragais, Occitanie, France

[8]Health Services Management, School of Advanced Public Studies, University of Moncton, Moncton, NB, Canada

[9]Faculty of Medicine, Laval University, Quebec, QC, Canada

Corresponding Author:
Hassane Alami, MSc PH, MSc HP, PhD
Department of Health Management, Evaluation and Policy, School of Public Health & Center for Public Health Research of the University of Montreal, P.O. Box 6128, Branch Centre-Ville, Montreal, QC, Canada

Abstract

Background: Several artificial intelligence (AI) governance frameworks have emerged to help health systems (HS) address AI-related risks. However, most fail to capture the multidimensional and evolving nature of real-world governance.

Objective: This systematic review aimed to synthesize existing AI governance frameworks for HS and to propose an integrative AI governance model identifying key components to guide AI-related policy, practice, and research in HS.

Methods: A comprehensive search was conducted in 8 academic databases (PubMed, MEDLINE, Embase, ACM Digital Library, Web of Science, Scopus, Social Sciences Abstracts, and PsycINFO), gray literature databases, and international organization web portals in October 2024 (updates: July 2025 and March 2026) and limited to studies published from November 2014 to March 2026 in English, French, Spanish, or Portuguese. Eligible documents included peer-reviewed articles and reports proposing AI governance frameworks for HS. Two reviewers independently selected the frameworks, assessed their quality using the Appraisal of Guidelines for Research and Evaluation for Health Systems, and extracted data. Results were synthesized using thematic analysis.

Results: The research retrieved 10,175 records, among which 19 AI governance frameworks were identified. Most were published between 2022 and 2024 (n=13, 68%), half (n=10, 53%) were developed by authors based in North America, and only one-third (n=6, 32%) were derived from primary studies. The frameworks focused on 4 levels of AI governance: international (n=3, 16%), national (n=5, 26%), local (n=3, 16%), and organizational (n=8, 42%). All of them underline the crucial role of multidisciplinary bodies in the structure of AI governance in HS. Six key AI governance processes in HS emerged as critical: (1) need and/or problem identification (n=14, 74%), (2) data governance (n=17, 89%), (3) risk assessment and management (n=17, 89%), (4) validation and/or evaluation (n=18, 95%), (5) maintenance and monitoring (n=16, 84%), and (6) integration (n=9, 47%). Additionally, 4 pivotal relational mechanisms were identified: (1) ethical principles and/or values (n=17, 89%), (2) education and training (n=14, 74%), (3) communication (n=12, 63%), and (4) standards and regulations (n=13, 68%).

Conclusions: Our study provides a comprehensive synthesis of existing AI governance frameworks for HS across 4 levels (local, regional, national, and international), underpinned by a quality assessment of the 19 identified frameworks. It differs from existing studies that concentrate on specific dimensions or settings by contributing an integrative AI governance model for HS comprising 2 dimensions and 4 relational mechanisms across the 4 levels, explicitly modeling their interactions. Future research should test and operationalize the proposed model to enhance its practical applicability. Strengthening the methodological rigor of AI governance frameworks will be essential for the responsible integration of AI in HS. As the framework is primarily grounded in Global North and English-language literature, validation in other contexts is warranted.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e87448) doi:10.2196/87448

KEYWORDS

artificial intelligence; digital health; governance; health system; machine learning; deep learning; natural language processing

Introduction

Background

The rapid and widespread integration of artificial intelligence (AI) into health systems (HS) has raised high expectations while also creating significant uncertainties and challenges [1-4]. AI refers to "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments" [5]. As AI is increasingly viewed as a lever to improve efficiency amidst workforce shortages [6], health organizations are implementing AI tools to support precision diagnostics, personalized medicine, clinical decision-making, and virtual assistance, among other applications [7,8]. The deployment of these complex technologies introduces new pressures and risks, including ethical, sociopolitical, economic, organizational, professional, and clinical challenges, with particular sensitivities around patient safety and privacy [1,3,4,9]. Additional concerns include bias and inequities across subgroups, regulatory uncertainty, liability, vendor lock-in, technical debt, data quality and interoperability gaps, inadequate management and training, and misalignment between procurement incentives and demonstrated real-world value [10,11]. In response, stakeholders have called for robust and adaptive AI governance frameworks for HS to safeguard the system and enable responsible, sustainable, and effective integration of AI-based technologies [1,12-15].

Although several ethical AI frameworks for HS define fundamental values (eg, trust, transparency, accountability, safety, justice, fairness, nonmaleficence, and privacy) and normative objectives [16,17], they generally do not provide benchmarks or operational guidance for achieving them [17-19]. While the role of ethical frameworks is to articulate principles, the role of governance frameworks is to provide the operational structures, processes, and mechanisms (the "how") required to translate principles (the "what") into practice [20-24]. To this end, governance frameworks identify relevant actors, allocate decision-making authority, and clarify responsibilities [20,21,23-25]. They also guide core functions such as data governance, model validation and auditing, and risk assessment and management [20,21,23-25]. To support a practical and consistent application of AI ethical principles [23], governance frameworks should aim to operationalize the following: (1) roles, rules, norms, competencies, and policies applicable throughout the AI lifecycle; (2) value-sharing arrangements, procurement and contracting requirements, remuneration and reimbursement mechanisms, and intellectual property considerations; (3) auditable requirements regarding data governance, human oversight, explainability, model validation, and cybersecurity; (4) postdeployment monitoring and incident reporting; assessments of clinical effectiveness, safety, cost implications, and equity impacts; and liability and risk allocation; and (5) stakeholder engagement, transparent public communication, accountability mechanisms, and criteria for model replacement or retirement [23,26-32].

Although AI governance in HS presents distinct characteristics and requirements, it nonetheless forms an integral component of the broader governance architecture for digital health [1,12,25,33-35]. It is grounded in the institutional, regulatory, and infrastructural arrangements and structures that shape how digital technologies are designed, procured, deployed, used, evaluated, and monitored in real-world care settings and that ultimately determine whether these tools can be integrated safely, effectively, efficiently, sustainably, and equitably at scale [1,12,25,34]. Accordingly, we adopt a governance perspective consistent with digital health governance and define AI governance in HS as "the exercise of political, administrative, and technical authority to manage everything associated with the health information system, in all areas of a national system. The structure of this governance consists of the mechanisms, processes, and institutions through which all stakeholders articulate their interest, exercise their rights, meet their obligations, and resolve their differences" [33]. These governance elements are interconnected, forming a coherent whole that ensures AI-based technologies meet clinical, legal, ethical, and regulatory requirements and are used in alignment with HS strategies, objectives, principles, and values [35].

The pressing need for comprehensive AI governance frameworks in HS has led to a proliferation of proposals over the past decade. Yet, these initiatives often address sociopolitical, ethical, clinical, or data management issues separately, rather than treating them as interdependent dimensions of governance [36]. This literature remains limited by 2 persistent conceptual shortcomings: the absence of a shared definition of AI governance in HS and the frequent conflation of governance with related notions such as ethics, responsibility, or regulation. For example, a scoping review by Stogiannos et al [37] mapped the available literature on AI governance in medical imaging and radiology and proposed a governance framework. However, the authors did not clearly define what they meant by governance; their analysis primarily examined the factors relevant to governance in the United Kingdom's imaging and radiology sectors, and most of the documents reviewed emphasized ethical and regulatory considerations at the organizational level rather than governance in a broader institutional sense [37]. This fragmentation, combined with a tendency to underestimate the structural and operational demands of governance, has contributed to persistent misalignment between national ambitions for AI integration and the strategies, institutional arrangements, and resources actually deployed to support such integration [36].

Objective

To the best of our knowledge, no systematic review has yet identified and synthesized AI governance frameworks in HS, despite the fact that such analysis is essential for identifying

XSL•FO
RenderX

recurrent shortcomings and informing strategies for the responsible integration of AI-based technologies into HS [38]. This systematic review aimed to synthesize existing AI governance frameworks for HS and to propose an integrative AI governance model identifying key components to guide AI-related policy, practice, and research in HS.

Methods

Overview

This systematic review was conducted and reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analyses) Literature Search Extension statement (Checklist 1) [39,40]. The full description of methods is provided in the protocol registered on the Open Science Framework prior to data collection [41]. No modifications or additional information were made to the protocol following its registration (PRISMA item 24c). In addition, the following PRISMA checklist items were considered not applicable to this review: 10b, 12, 13e, 13f, 15, 19, 20b, 20d, and 22 (Checklist 1). Given the qualitative nature of this systematic review, these items were not relevant to the methodological approach adopted in this study. Ethical approval was not required for this systematic review.

Eligibility Criteria

The evidence map included published peer-reviewed articles or reports proposing a new framework, guideline, standard, or position statement of AI governance in HS. Eligible documents had to be explicitly labeled as AI governance in HS by the authors. Only records published in English, French, Spanish, or Portuguese between November 2014 and March 2026 were included. We selected 2014 as the starting point because the mid-2010s marks the period when AI began transitioning from primarily experimental applications to more concrete and integrated uses within HS [42-44]. This period also coincides with the expansion of large-scale digital health data infrastructures and the growing involvement of major technology companies in health AI ecosystems. From a governance perspective, it corresponds to the emergence of more structured global discussions on AI ethics, responsible AI, and digital health governance, as well as the proliferation of national digital health and AI strategies [43,44]. We therefore chose 2014 to focus the analysis on frameworks most relevant to current policy and implementation contexts. Abstracts, letters to the editor, commentaries, essays, viewpoints, conference proceedings, and unavailable articles were excluded. Two reviewers (RPS and EJP) independently screened all references for inclusion using *Covidence* (a web-based software platform) [45]. Disagreements were resolved through discussion. If the 2 primary reviewers could not reach a consensus, a third reviewer (HA) adjudicated the final decision.

Information Sources and Search Strategy

A comprehensive literature search was conducted in 8 databases (PubMed, MEDLINE [Ovid], Embase [Ovid], ACM Digital Library, Web of Science [Clarivate], Scopus, Social Sciences Abstracts [EBSCO], and PsycINFO [EBSCO]) in October 2024 (updated in July 2025 and March 2026). Database searches were conducted and limited to human studies published since November 2014. The strategy combined free-text keywords and controlled vocabulary terms structured around the concepts of AI (“Artificial Intelligence” or “AI” or “machine learning” or “deep learning” or “natural language processing”) AND governance (“governance” or “regulation” or “framework” or “model” or “policy” or “guideline” or “ethic” or “standard” or “principles” or “regulatory” or “responsible”) AND health system (“health system” or “healthcare system” or “health services” or “health sector” or “hospital” or “healthcare delivery” or “public health systems” or “health sector”). The search strategy was developed by the authors through a rapid literature review and group discussion. We conducted searches across Google Scholar, ProQuest Dissertations and Theses Database (ProQuest), Social Science Research Network (SSRN eLibrary), and Evidence for Informed Health Policymaking (PDQ-Evidence). The electronic search was complemented by manual searches of the reference lists of relevant articles and web portals of institutions including the World Health Organization (WHO), United Nations, the Organisation for Economic Co-operation and Development, World Bank, European Union, National Institute for Health and Care Excellence, and National Institutes of Health. Details on the search strategies are available in Multimedia Appendix 1. All references retrieved from the searches were imported into *EndNote20* (Clarivate) to manage citations and remove duplicates. The refined list of citations was imported into *Covidence*.

Data Collection Process

Data were extracted using a standardized form in *Microsoft Excel*, including (1) document characteristics (author names, year of publication, first author country, objective, type of document, targeted audience, and definition of AI governance); (2) methodology (research question, study design, sample characteristics, and data collection procedures); (3) framework characteristics (name, level of application, design, actors, roles and accountabilities, governance structure functions, processes, mechanisms, barriers, and challenges); and (4) funding sources. Data were extracted by 2 reviewers (RPS and LR) and checked by 2 others (EJP and HA).

Selection Process: Study Quality Appraisal Using AGREE-HS

Framework quality was assessed using a modified version of the Appraisal of Guidelines for Research and Evaluation for HS (AGREE-HS). The original AGREE-HS and the modifications are presented in Multimedia Appendix 2 [46]. As in the original AGREE-HS, 5 quality items were rated: (1) topic (clarity and relevance of the AI governance challenge); (2) participants (stakeholder inclusion, multidisciplinary representation, and conflicts of interest); (3) methods (transparency and rigor of framework development processes); (4) recommendations (clarity, feasibility, ethical principles, equity, and sociocultural and political factors); and (5) implementability (feasibility, barriers and facilitators, resource requirements, sustainability, and transferability). Regarding modifications, 21 criteria of the original AGREE-HS applicable

XSL•FO
RenderX

to frameworks were also included, as no tool currently exists for these.

Additionally, we assigned scores to each criterion in advance to improve assessment reliability. Two authors (RPS and EJP) independently applied the explicit scoring criteria to each item using a Microsoft Excel spreadsheet. Disagreements in each item score were resolved by consensus. Each domain score was calculated by aggregating its items' scores and was rated on a 7-point scale, where 1 indicated the lowest quality and 7 indicated the highest quality. The framework overall score was calculated following the AGREE-HS manual as follows [47]: (obtained score–minimum score)/(maximum score–minimum score)×100%, with a minimum total score of 5 and a maximum total score of 35. We performed a tertile split of the frameworks' overall scores using the 33rd- and 67th-percentile cut points to classify them as high quality (overall scores >51%), moderate quality (overall scores between >41% and 51%), or low quality (overall scores ≤41%).

Data Synthesis

Data were analyzed and integrated through a narrative synthesis approach, which allowed for a coherent combination of the varied elements within the AI governance frameworks for HS. Framework levels were categorized as international, national, local, or organizational based on their scope. The analysis was informed by the framework of Van Grembergen et al [48] for IT governance adapted to AI-based technologies. This framework encompasses 3 interconnected dimensions of governance: structures, processes, and relational mechanisms. The structures are the actors and their roles that ensure a clear chain of responsibility for the decisions and operations around AI-based technologies. The processes are the formal methods and workflows for making and monitoring these technologies, including establishing a framework for planning, prioritizing, and monitoring AI initiatives. Relational mechanisms are elements that foster trust, participation, partnership, and mutual understanding in the processes [48]. Whenever it is possible and pertinent, the components of the processes and relational mechanisms identified in the frameworks are presented according to the different stages of the AI lifecycle.

Results

Framework Selection

A total of 10,175 records were retrieved from the bibliographic databases, of which 91% (n=9210) remained after removing duplicates. Following title and abstract screening, 163 (2%) documents were full-text screened. Of these, 12 (0.1%) articles proposed an AI governance framework for HS. One (0.009%) article was excluded, as it was a scoping review [37], resulting in the inclusion of 11 (0.1%) peer-reviewed articles. Another 87 additional full-text records were identified through websites and reference lists, adding 8 publications to the systematic review. A total of 19 AI governance frameworks for HS were identified in 15 peer-reviewed articles and 4 reports (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) for literature screening [49]. AI: artificial intelligence; HS: health systems.

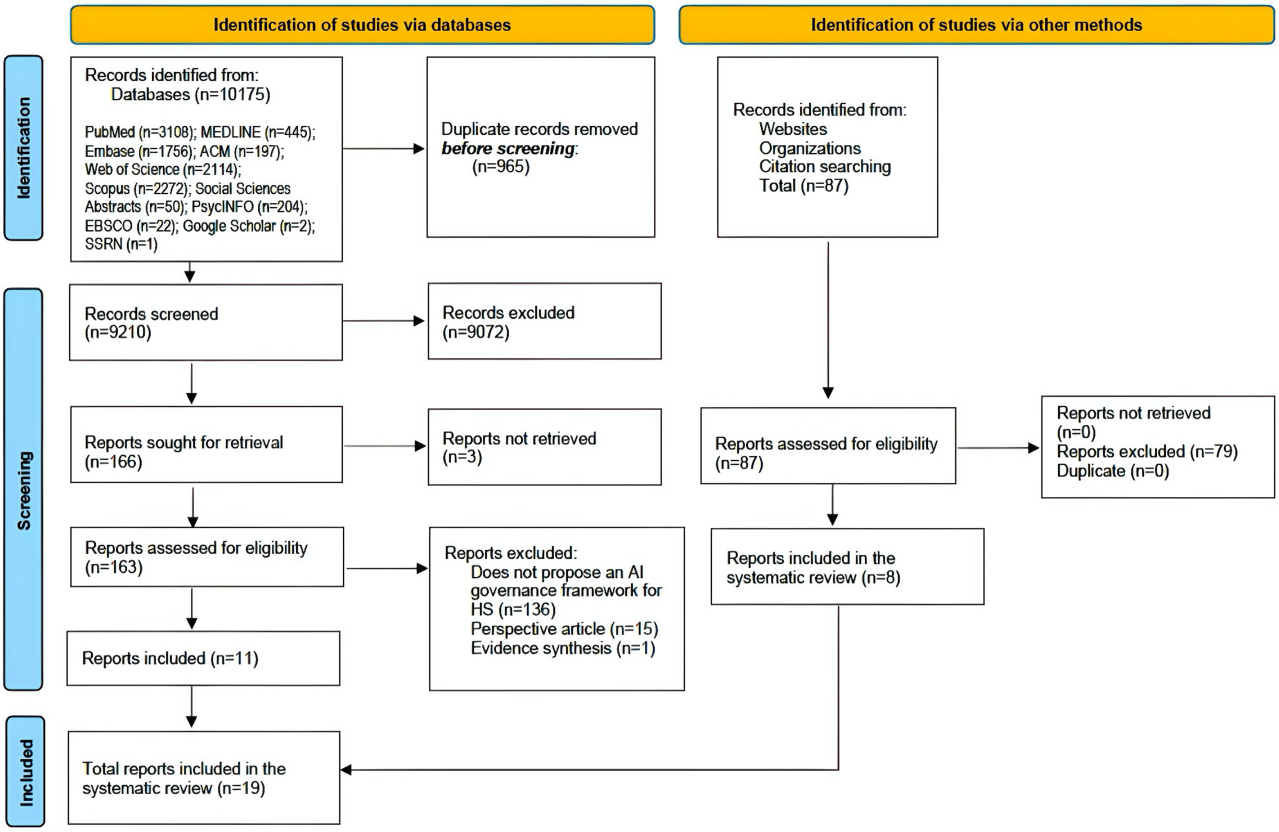

XSL•FO
RenderX

Framework Characteristics

An overview of included documents is presented in Multimedia Appendix 3. Six documents resulted from primary studies [50-55], and only 6 provided a definition of AI governance [1,56-60]. The geographical distribution of first authors and the scale of intended audiences are presented in Figure 2. Most frameworks (13/19, 68%) were published between 2022 and 2024 [50-57,59,61-64]. Ten (53%) frameworks were developed by authors based in North America—7 (37%) from the USA [50,55,57,59-61,63] and 3 (16%) from Canada [58,62,65]—3 (16%) in Australia [30,53,64], 2 (10%) in Switzerland [1,56], 1 (5%) in the United Kingdom [51], 1 (5%) in Germany [52], 1 (5%) in New Zealand [54], and 1 (5%) in Qatar [66]. Details of the included frameworks are presented in Multimedia Appendix 4. They targeted audiences at 4 levels: 3 international [1,51,56], 5 national [30,59,64-66], 3 local [53,54,62], and 8 organizational [50,52,55,57,58,60,61,63]. Ten frameworks were developed through collaboration between academics and practitioners [1,50,55,56,59,60,63-66].

Figure 2. Geographical distribution of first authors and level of intended audience (organizational, local, national, and international).

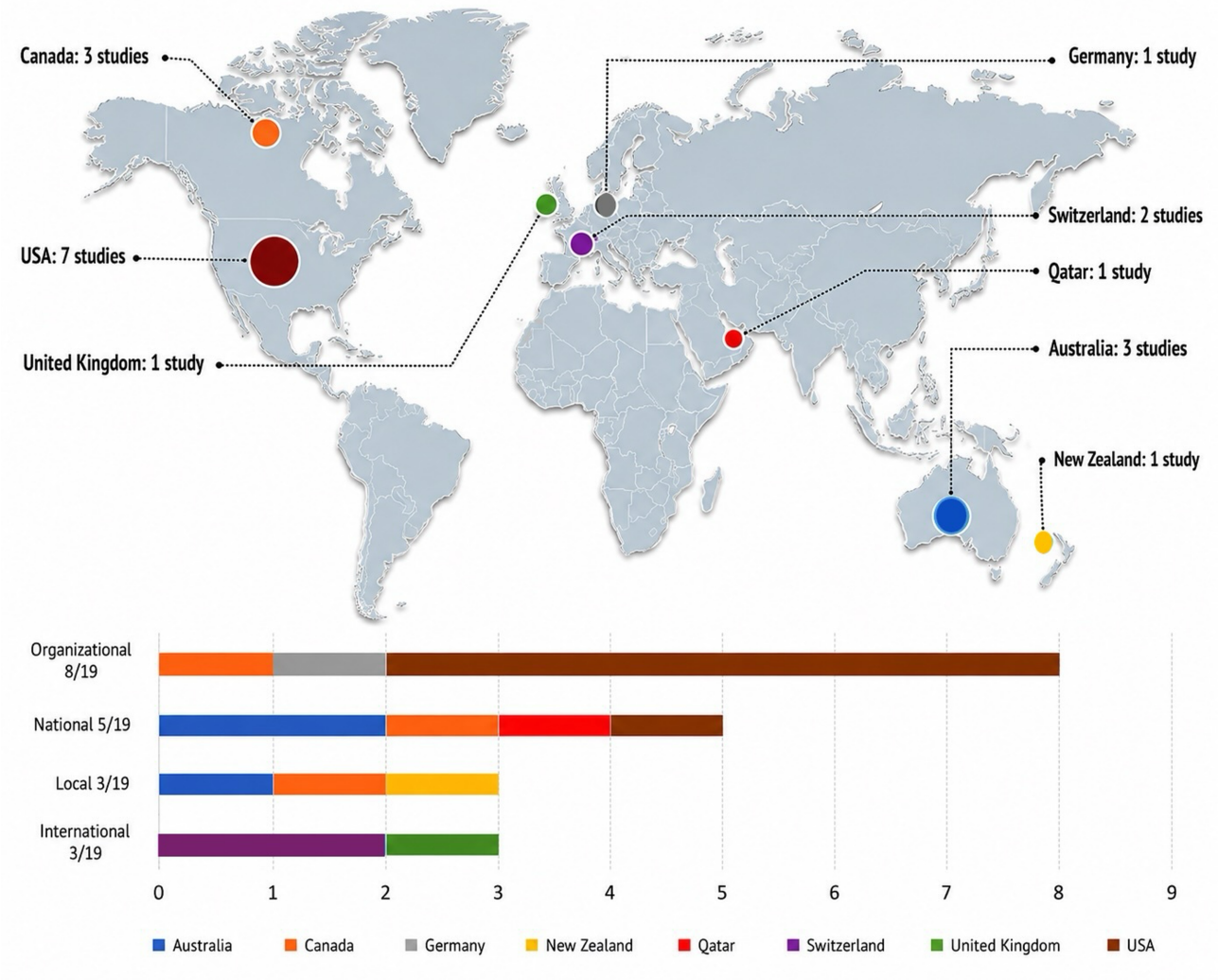

Framework Critical Appraisal

Table 1 presents the results of the adapted AGREE-HS. Seven frameworks were rated as high quality [1,30,50,51,56,60,62], 6 were rated as moderate quality [53,54,58,61,63,66], and 6 were rated as low quality [52,55,57,59,64,65].

While 9 frameworks received a score of 6 or higher for the information about the framework development team (Table 1—Participants) [50,51,53,54,56,61-63,65], only 7 reached this threshold for describing the AI governance challenge, its causes, and the priority attributed to it (Table 1—Topic) [1,51,56,58,60,62,66]. Most of the authors do not provide a complete description of systematic and transparent methods to review the evidence and to agree upon the framework components. Therefore, 14 frameworks scored 3 or lower on the dimension assessing the methods used to develop them (Table 1—Methods) [1,30,50,52,53,55,57,59,61-66]. Finally, most authors did not include information on the framework implementability (17 scored 3 or less; Table 1—Implementability) [1,30,51-56,58-66].

XSL•FO
RenderX

Table . Results of the Appraisal of Guidelines for Research and Evaluation for Health Systems (AGREE-HS) and classification of the frameworks according to their quality[a].

References	Participants	Topic	Methods	Recommendations	Implementability	Overall score (%)	Quality
International level							
WHO[b], 2024 [56]	7	7	6	6	2	77	High
Morley et al, 2022 [51]	6	7	7	5	2	73	High
WHO, 2021 [1]	4	7	3	5	3	57	High
National level							
Reddy et al, 2020 [30]	5	5	3	5	3	53	High
Jaremko et al, 2019 [65]	6	5	2	2	1	37	Low
Parker et al, 2024 [59]	4	3	3	4	2	37	Low
AAAiH[c], 2023 [64]	4	4	2	3	1	30	Low
Solaiman 2025 [66]	4	7	1	5	3	50	Moderate
Local level							
Arnaout et al, 2024 [62]	7	7	2	4	1	53	High
Whittaker et al, 2023 [54]	6	4	4	4	1	47	Moderate
Carter et al, 2024 [53]	7	2	3	5	1	43	Moderate
Organizational level							
Liao et al, 2022 [50]	6	4	2	3	6	53	High
Bedoya et al, 2022 [61]	7	5	2	3	2	47	Moderate
Hassan et al, 2025 [58]	3	6	5	4	1	47	Moderate
Economou-Zavlanos et al, 2024 [63]	6	3	3	3	3	43	Moderate
Daye et al, 2022 [57]	4	4	3	2	4	40	Low
Kim et al, 2023 [55]	4	4	3	4	1	37	Low
Apfelbacher et al, 2024 [52]	2	3	2	1	1	31	Low
Kim et al, 2026 [60]	5	7	5	4	3	67	High

[a]Score ranging from 1 to 7: score of 1 (lowest quality): a score of 1 should be given if there is no information that is relevant to the AGREE-HS item, if the criteria are very poorly reported in the HSG document or if the authors explicitly state that it was not done. Score of 7 (highest quality): a score of 7 should be given if the information related to the AGREE-HS item was exceptionally well reported, all criteria related to the item have been considered during the development of the guidance, and the information related to the item is applicable in its context. The frameworks were classified into high quality (overall scores >51%), low quality (overall scores ≤41%), and moderate quality (overall scores between >41% and 51%). The details about the evaluation components can be found in Multimedia Appendix 2.

XSL•FO
RenderX

[b]WHO: World Health Organization.

[c]AAAiH: Australian Alliance for Artificial Intelligence in Healthcare.

Governance Structures

AI governance structures encompass functions and the roles and responsibilities of involved actors. Multimedia Appendix 5 presents authors' statements on the functions of AI governance structures. Seventy-four percent (14/19) of frameworks explicitly define these functions [1,30,50,51,53,54,56,59-63,65,66]. Two international-level frameworks emphasize that these structures are intended to support the achievement of national health objectives [1,51].

According to Morley et al [51] and the WHO [1], AI governance structures also play a crucial role in establishing shared standards of safety, privacy, and efficacy to ensure compliance with minimum requirements. Furthermore, the WHO indicates that these structures can help prevent regulations that could create unfair competitive advantages or disadvantages for both companies and governments. Such structures also ensure that governments develop regulatory frameworks that uphold ethical principles, human rights, and international law [56].

At a national level, 4 frameworks indicate that AI governance structures have the function of establishing guidelines to evaluate AI-based technologies intended for use in the HS [30,59,65,66]. According to Reddy et al [30], Parker et al [59], and Solaiman [66], these guidelines help to ensure that AI-based technologies are implemented safely, ethically, and in compliance with regulatory standards. Likewise, for 3 local-level frameworks, the function of AI governance structures is to ensure adherence to ethical principles in the deployment of AI, such as equity, privacy, and transparency [53,54,62]. Arnaout et al [62] and Whittaker et al [54] highlighted that these structures also play a crucial role in identifying gaps for AI solutions and ensuring the appropriateness, safety, and effectiveness of AI development and deployment. Additionally, 6 organizational-level frameworks underscore the pivotal role of AI governance structures in providing guidance and support throughout the AI lifecycle [50,57,58,60,61,63]. For some, these structures are responsible for supporting the development of AI-based technologies [57,58,60] and for establishing evaluation, monitoring, and adoption processes [50,57,58,60].

The literature on actors and their roles and responsibilities is notably more heterogeneous. Across the 19 frameworks, actor composition within governance structures at all levels varied substantially, reflecting differences in contextual scope and sectoral representation. These actors are categorized by subgroup and scale in Figure 3; however, the categories overlap. All frameworks underlined the crucial role of multidisciplinary bodies or teams in governing AI within HS across these different levels. Notably, only the international-level frameworks explicitly highlight the importance of synergy between governance structures operating at different scales [1,51,56].

Owing to the large number of actors involved, the frameworks present significant variations in the roles and responsibilities assigned to each actor, with many responsibilities shared among multiple actors. For instance, while Morley et al [51] suggested that "collaboration and multidisciplinary working by decision-makers, technologists, healthcare professionals, and academics are needed to ensure appropriate expertise throughout the AI life cycle," the WHO [56] stated that "governments could require developers to ensure that the design and development of a general-purpose foundation model achieve certain outcomes throughout its life cycle." For this reason, it was not feasible to synthesize the roles and responsibilities comprehensively. Other examples of actors' roles and responsibilities are provided in Multimedia Appendix 6.

XSL•FO
RenderX

Figure 3. Artificial intelligence (AI) governance actors identified across the included frameworks. OECD: Organisation for Economic Co-operation and Development; UNESCO: United Nations Educational, Scientific and Cultural Organization; WHO: World Health Organization.

Governance Processes

Overview

Governance processes involve the formal methods and workflows for developing and monitoring AI-based technologies. Within the included frameworks, 6 processes were identified as critical: need and/or problem identification, data governance, risk assessment and management, validation and/or evaluation, maintenance and monitoring, and integration. Table 2 presents the frameworks that included these processes as crucial in AI governance within HS.

XSL•FO
RenderX

Table . Artificial intelligence governance processes identified across the included frameworks.

Author and year	Need and/or problem identification[a]	Data governance[b]	Risk assessment and management[c]	Validation and/or evaluation[d]	Maintenance and monitoring[e]	Integration[f]
International level						
WHO[g], 2024 [56]	X	X	X	X		
Morley et al, 2022 [51]	X			X	X	
WHO, 2021 [1]	X	X	X	X	X	
National level						
Reddy et al, 2020 [30]		X	X	X	X	X
Jaremko et al, 2019 [65]		X				
Parker et al, 2024 [59]	X	X	X	X	X	
AAAiH[h], 2023 [64]		X	X	X	X	
Solaiman, 2025 [66]		X	X	X	X	X
Local level						
Arnaout et al, 2024 [62]	X	X	X	X	X	
Whittaker et al, 2023 [54]	X	X	X	X	X	X
Carter et al, 2024 [53]	X	X	X	X	X	
Organizational level						
Liao et al, 2022 [50]			X	X	X	
Bedoya et al, 2022 [61]	X	X	X	X	X	X
Hassan et al, 2025 [58]	X	X	X	X		X
Economou-Zavlanos et al, 2024 [63]	X	X	X	X	X	X
Daye et al, 2022 [57]	X	X	X	X	X	
Kim et al, 2023 [55]	X	X	X	X	X	X
Apfelbacher et al, 2024 [52]	X	X	X	X	X	X
Kim et al, 2026 [60]	X	X	X	X	X	X

[a]Total frameworks=74%.

[b]Total frameworks=89%.

[c]Total frameworks=89%.

[d]Total frameworks=95%.

[e]Total frameworks=84%.

[f]Total frameworks=47%.

[g]WHO: World Health Organization.

XSL•FO
RenderX

[h]AAAiH: Australian Alliance for Artificial Intelligence in Healthcare.

Need and/or Problem Identification

Fourteen (74%) of 19 frameworks emphasized the importance of identifying the needs and/or problems that the AI solution aims to address during the initial phases of technology development [1,51-63]. However, none of the frameworks provided any tools or guidelines to conduct this needs assessment. All international-level frameworks stated that AI in HS should enhance system capacity or improve health outcomes [1,51,56]. At the local level, 2 frameworks explicitly asserted that the use of AI should be based on a clearly defined need and an appropriate contextual understanding [54,62]. In a similar manner, 7 organizational-level frameworks stated that the need or problem an AI technology is intended to address should be clearly identified during the initial stage, prior to the development of the technology [52,55,57,58,60,61,63]. One of these frameworks also underscored the importance of understanding the context in which the problems arise, which should rely on the inclusion of frontline workers to describe the problems in an effective way [55]. In contrast, at the national level, only Parker et al [59] highlighted the importance of developers understanding the needs of HS to develop tailored AI-based technologies.

Data Governance

Although data security was addressed in 14 (74%) of 19 frameworks [1,30,52,54,56-60,62-66], only 11 (58%) frameworks included information on forms of consent [1,30,53,56-58,62-66], and merely 6 (32%) considered data quality [53-55,60,61,63]. For instance, all international-level frameworks emphasized the need for a well-established data protection structure, including clear laws and regulations frameworks to ensure transparency and individuals' privacy [1,51,56]. The WHO's frameworks underscore the importance of establishing clear procedures for obtaining informed consent from data subjects [1,56]. Similarly, all national-level frameworks refer to the development of procedures or frameworks to ensure data security [30,59,64-66]. Four of them mention concerns related to obtaining informed consent and patient privacy [30,64-66]. Nevertheless, none of the international or national frameworks reviewed address the issue of data quality. Local-level frameworks addressed at least two critical elements of data governance. Arnaout et al [62] and Whittaker et al [54] mentioned the need for ensuring appropriate and secure environments for data storage and sharing. Carter et al [53] and Arnaout et al [62] underscored the importance of explaining the use of data to patients and of ensuring that they understand and approve it. Additionally, Carter et al [53] and Whittaker et al [54] emphasized that data quality requires the use of data that are both complete and representative of the targeted population.

In the same vein, 5 organizational-level frameworks encompassed key data governance components related to data security [52,57,58,60,63]. Some of these components are privacy protection [57,58,63], secure storage mechanisms, management of datasets [60], and data traceability [58]. However, explicit recognition of the critical role of consent in protecting patients' rights is found in only 2 organizational-level frameworks [57,63]. Four organizational-level frameworks highlight the importance of ensuring data quality [55,60,61,63], including the reduction of bias, as well as ensuring data diversity and appropriate data cleaning [55].

Risk Assessment and Management

Eighty-nine percent (17/19) of frameworks included risk assessment and management as part of AI governance and underscore the need for both proactive (predeployment) and reactive (postimplementation) strategies [1,30,50,52-64,66]. At the international level, the 2 WHO frameworks addressed privacy risks associated with the use of AI technologies in HS [1,56]. One of them highlighted that providers or developers should inform deployers about the potential risks of the AI technology, thereby allowing for an informed decision regarding its implementation [56]. Three national-level frameworks emphasized the importance of national standards and postmarket safety monitoring to ensure the timely identification and reporting of risks to patients [30,64,66], while for another 3 frameworks, risk assessment should be an ongoing process, supported by continuous monitoring of safety and risks [30,59,66]. Likewise, 2 local-level frameworks emphasized the need for assessing and managing risks prior to the deployment of AI-based technologies [54,62], and 2 frameworks pointed out the need for ongoing monitoring after deployment [53,62].

The same trend is observed in all 8 organizational-level frameworks [50,52,55,57,58,60,61,63]. Three of them bring forward the importance of implementing continuous monitoring of the algorithm [50,55,60]. Bedoya et al [61] outlined a risk-based triage system algorithm review, ranging from "full committee evaluation" to "no further review." Economou-Zavlanos et al [63] emphasized an iterative approach that supports continuous feedback between reviewers and developers to strengthen risk-benefit assessments. Furthermore, Kim et al [55] advocated for organizations' early identification and mitigation of potential AI-related risks.

Validation and/or Evaluation

Most frameworks highlighted validation and/or evaluation processes as core components of effective AI governance in HS [1,30,50-64,66]. The 3 international-level frameworks included these components [1,51,56]. While 2 of them highlighted the evaluation of AI performance as a crucial step prior to deployment [1,51], 1 emphasized the importance of validation in assessing the suitability of the data used in the models [56]. Equally, 4 national-level frameworks incorporated the validation and evaluation of AI-based technologies as essential steps prior to deployment and/or integration into HS [30,59,64,66]. For instance, Parker et al [59] noted that the evaluation process, which should be conducted by governance teams at the HS level but also by third parties providing assurance reviews, may encompass considerations such as performance, privacy, compliance, legal issues, patient safety, clinical integration, IT integration, and validation using internal datasets.

Three local-level [53,54,62] and 8 organizational-level [50,52,55,57,58,60,61,63] frameworks emphasized the

XSL•FO
RenderX

importance of continuously evaluating AI-based technologies following implementation. Two local-level frameworks highlighted that validation and evaluation processes are essential for assessing the performance of AI-based technologies prior to their deployment in HS [53,62]. Three organizational-level frameworks emphasized that this evaluation should include the analysis of bias, potential integration into the clinical workflow, and data handling procedures [57,58,60]. In addition, 3 organizational-level frameworks underscored that an initial evaluation should be conducted to assess regulatory approval needs, strategic alignment with the organization, and resource requirements, including funding, time, data environment, infrastructure, and human capital [55,58,60].

In the same manner, all organizational-level frameworks included validation or evaluation of AI-based technologies in the development and/or deployment stages. These frameworks also shared a consensus on the importance of validating datasets to determine the appropriateness of the AI technology for the target population [50,52,55,57,58,60,61,63]. However, not all organizational-level frameworks specified the strategies for doing so [50,57,58,61,63]. For instance, 3 of them supported the use of "real-world clinical data" for evaluation, recommending an approach in which the AI technology is used, but its results remained hidden from clinical users during the assessment phase [55,61,63]. Finally, 3 organizational-level frameworks outlined a set of metrics for postdeployment evaluation, such as algorithm sensitivity and specificity, patient outcomes, reduction in burden on HS providers, and overall cost efficiency [50,55,60].

Maintenance and Monitoring

Eighty-four percent (16/19) of the frameworks emphasized maintenance and monitoring in AI governance for HS [1,30,50-55,57,59-64,66]. At the international level, the WHO's framework underlines that robust postmarketing surveillance could enhance regulatory oversight by helping to identify biases [1]. Similarly, Morley et al [51] highlighted that collective intelligence at this level can be supported through national oversight procedures, as establishing such knowledge sharing mechanisms would enable countries to access information about AI-based technologies they are considering or have already implemented. However, the manner to integrate monitoring varies among national-level frameworks [30,59,64,66]. While 2 frameworks suggested integrating regular audits and reporting to enable ongoing safety monitoring [30,59], 2 others emphasized the implementation of a national postmarket safety monitoring system [64,66]. Furthermore, local-level frameworks emphasized the need to establish reporting mechanisms following the implementation of AI-based technologies [53,54], as well as continuous monitoring for biases, errors, and performance assessment [53,62].

Correspondingly, 7 organizational-level frameworks highlighted maintenance and monitoring in AI governance [50,52,55,57,60,61,63]. Four frameworks emphasized that performance and impact should be monitored regularly to detect performance drift and assess clinical outcomes after implementation [50,55,61,63]. Likewise, 2 organizational-level frameworks underlined the crucial role of monitoring outcomes related to safety, as well as clinical and technical performance [61,63]. Nevertheless, only 2 frameworks pointed out that monitoring should also include equity in care, economic impact, and workforce satisfaction [55,60].

Integration

Although 47% (9/19) of the frameworks emphasized integrating AI-based technologies into HS as part of governance processes at national [30,66], local [54], and organizational levels [52,55,58,60,61,63], none of the international frameworks mentioned this integration. In all these frameworks, "integration" refered to the analysis of how the new AI technology will align with or affect existing clinical workflows [30,52,54,55,58,60,61,63,66]. The national-level frameworks recommend establishing a clinical governance committee to support the integration of AI-based technologies into clinical workflows [30,66]. At the local level, Whittaker et al [54] pointed out that involving clinicians in the conception and development phases helps to better understand how seamlessly new technologies can integrate into clinical workflows. According to them, without insight into how local health services currently operate, new developments may entirely fail to meet the practical needs of those systems. Likewise, organizational-level frameworks emphasized the importance of understanding current workflows [63], designing intuitive user interfaces, supporting frontline workers [55], and adapting AI-based technologies during deployment to ensure its seamless integration and optimal use [52,60].

Governance Relational Mechanisms

Overview

Relational mechanisms are process elements that foster trust, participation, partnership, and mutual understanding. Within the included frameworks, 4 relational mechanisms were identified as critical: ethical principles and/or values, education and training, communication, and standards and regulations. Table 3 summarizes the percentage of frameworks that included these relational mechanisms.

XSL•FO
RenderX

Table . Artificial intelligence governance relational mechanisms identified across the included frameworks.

Author and year	Ethical principles and/or values[a]	Education and training[b]	Communication[c]	Standards and regulations[d]
International level				
WHO[e], 2024 [56]	X	X	X	X
Morley et al, 2022 [51]		X		X
WHO, 2021 [1]	X	X	X	X
National level				
Reddy et al, 2020 [30]	X	X		X
Jaremko et al, 2019 [65]		X		
Parker et al, 2024 [59]	X	X	X	
AAAiH[f], 2023 [64]	X	X	X	X
Solaiman, 2025 [66]	X		X	X
Local level				
Arnaout et al, 2024 [62]	X	X	X	
Whittaker et al, 2023 [54]	X	X		X
Carter et al, 2024 [53]	X	X	X	
Organizational level				
Liao et al, 2022 [50]	X			X
Bedoya et al, 2022 [61]	X	X	X	X
Hassan et al, 2025 [58]	X	X		X
Economou-Zavlanos et al, 2024 [63]	X		X	X
Daye et al, 2022 [57]	X		X	X
Kim et al, 2023 [55]	X		X	
Apfelbacher et al, 2024 [52]	X	X		
Kim et al, 2026 [60]	X	X	X	X

[a]Total frameworks: 89%.
[b]Total frameworks: 74%.
[c]Total frameworks: 63%.
[d]Total frameworks: 68%.
[e]WHO: World Health Organization.
[f]AAAiH: Australian Alliance for Artificial Intelligence in Healthcare.

Ethical Principles and/or Values

Although 89% (17/19) of the frameworks integrated principles and/or values into AI governance for HS [1,30,50,52-64,66], most commonly emphasizing the importance of ethical principles, such as privacy, inclusiveness, fairness, and transparency [1,30,50,52,54,56-60,62-64,66], they reflected a heterogeneous understanding of how these principles are embedded within AI governance. At the international level, the 2 WHO frameworks placed ethics and human rights at the core of AI governance and highlighted that achieving this goal requires both government interventions and actions from companies [1,56]. Therefore, they recommend companies to adopt a paradigm that grounds technology design in “the values of human dignity, freedom, equality, and solidarity” [1,56]. Additionally, these frameworks also emphasized the importance of inclusiveness as a core principle for achieving ethics-centered AI governance. This includes integrating community oversight [1,56] and ensuring the participation of low- and middle-income countries [56], as well as indigenous communities [1], in shaping international AI governance standards. Finally, 1 framework also highlighted human autonomy and responsiveness principles [1].

At the national level, 3 frameworks suggested different approaches to integrate ethics into AI governance [30,59,64]. While the Australian Alliance for Artificial Intelligence in Healthcare (AAAiH) advocated for the establishment of a national ethical framework [64], Parker et al [59] emphasized that a comprehensive AI governance, guided by an

XSL•FO
RenderX

interdisciplinary committee including expertise in ethics, can foster adherence to ethical standards. Likewise, Reddy et al [30] argued that a governance model based on the principles of fairness, transparency, trustworthiness, and accountability can address ethical challenges such as biases, privacy concerns, and trust of both the clinicians and the public in the use of AI in HS.

At the local level, 2 frameworks emphasized the importance of using training datasets that are representative and inclusive of population diversity. They also highlighted the need for AI governance that considers societal impacts, extending beyond individual risks and benefits [53,54]. While one of them emphasized that ethics should be embedded throughout the design and development of AI-based technologies [54], the other underscored the need for ethical review processes to assess how clinicians and patients are likely to engage with model outputs in clinical decision-making [53]. Additionally, Whittaker et al [54] asserted that AI governance should demonstrate respect for traditional knowledge and cultural protocols.

At the organizational level, all frameworks incorporated ethical principles in AI governance [50,52,55,57,58,60,61,63]. Three of these frameworks explicitly emphasized the importance of integrating ethical principles before and/or during the deployment of AI-based technologies [50,57,58]. Therefore, AI governance bodies within the organization are accountable for ensuring that both internal or external developers integrate these principles. For instance, Daye et al [57] argued that during the preimplementation assessment process, the committee must consider fairness and other ethical issues and address systemic biases, including those affecting disadvantaged communities. Similarly, Economou-Zavlanos et al [63] incorporated ethical and quality principles across the development lifecycle by introducing "evaluation checkpoints" where principles of clinical value, safety, fairness, and equity are systematically reevaluated throughout the development process. Regarding equity, 4 frameworks explicitly identified it as an outcome to be measured in AI governance [50,55,57,60]. Liao et al [50] and Kim et al [55] emphasized the importance of ensuring an equitable distribution of the benefits across population subgroups [50,55]. Furthermore, 2 frameworks highlighted the need for equitable resource allocation for AI implementation to avoid exacerbating existing disparities between units or departments [55,57].

Education and Training

Seventy-four percent (14/19) of the frameworks incorporated education and training as key mechanisms of AI governance. All these frameworks emphasized the need to educate and train professionals in the use of AI-based technologies within HS [1,30,51-54,56,58-62,64,65]. Two international-level [1,56] and 3 national-level [30,64,65] frameworks also highlighted the importance of improving AI literacy among the general public, including patients. Moreover, 1 international-level framework highlighted that providing education and training in ethics to designers and developers is critical to AI governance [1]. At the local level, Whittaker et al [54] advocated for technical guidance that reflects the specific needs and conditions of the local context. In addition, 2 organizational-level frameworks underscored the importance of implementing training plans, including the availability of materials to educate end users on the appropriate use of AI-based technologies [61,63].

Communication

Communication is identified as a key mechanism in 63% (12/19) of the frameworks [1,53,55-57,59-64,66]. Although communication concerns vary across all levels of governance, 5 frameworks recognize the role of effective communication in improving transparency about the use of AI in HS [53,56,59,62,64]. For instance, 1 national-level framework suggested that HS must ensure clear communication strategies to inform when AI-based technologies are being considered, so they can bring them into the governance process [59]. Likewise, 1 local-level framework stressed that AI users should receive clear, transparent, and context-specific information to reduce the potential risk of misuse and foster trust in the AI technology [62]. At the organizational level, 5 frameworks highlighted the need to establish clear and consistent communication between developers and the governance committee [55,57,60,61,63]. Two of them included the end users in this dynamic of communication to improve the AI technology [55,57].

Standards and Regulations

Sixty-eight percent (13/19) of the frameworks incorporate standards and regulations in AI governance for HS [1,30,50,51,54,56-58,60,61,63,64,66]. Two international-level frameworks emphasized the importance of establishing international standards to reduce risks to patient safety [1,51]. Likewise, 3 international-level and 2 national-level frameworks brought forward the need to create national standards to support the development and deployment of AI [1,51,56,64,66], ensuring safety, quality, and ethical standards. For instance, Reddy et al [30] argued that normative standards include "how AI models will be designed and deployed" in the context of HS. At the organizational level, of the 6 frameworks that address standards and regulations concerns [50,57,58,60,61,63], 3 proposed a "risk-based" approach to evaluation of AI models, with regulatory requirements that vary according to technology type and potential risks [57,61,63].

Governance Barriers and Challenges

Fifty-three percent (10/19) of the frameworks identified barriers and challenges to AI governance for HS [1,50,51,54,55,57,59,60,63,66]. International-level frameworks highlighted challenges arising from the disconnection between different yet highly interrelated policy domains across governance levels, the lack of international coordination in AI governance for HS [51], and concerns about the potential emergence of a "market" in which individuals can "buy and sell health data" [1]. Across other levels, 6 frameworks emphasized challenges related to the resources (eg, costs, infrastructure, and time) required for the effective implementation of AI governance [54,57,59,60,63,66]. Additionally, 3 organizational-level frameworks raised concerns about the ongoing resource demands necessary for the continuous monitoring of AI-based technologies [55,57,60].

An Integrative AI Governance Model for HS

None of the reviewed frameworks provided a comprehensive approach that fully captures the multidimensional and dynamic

XSL•FO
RenderX

nature of AI governance in HS. We therefore synthesized the dimensions and components reported across the 19 included frameworks to propose an integrative AI governance model that identifies key elements to guide AI-related policy, practice, and research in HS. The resulting model is presented in Figure 4. The model advances the field by explicitly articulating AI governance as a multilevel, interrelated system, in contrast to existing health information system or responsible AI governance frameworks that typically address only selected components or specific settings. Importantly, this integrative model reflects the current state of the literature and is not intended as a ready-to-implement governance framework for decision-makers. Rather, it provides a foundational scaffold that requires further refinement through expert validation, stakeholder and end user consultation, pilot testing, and evaluation.

The integrative AI governance model for HS shows how the key governance components identified across the included frameworks are deeply interconnected, as well as the permeability of different levels of AI governance. The framework comprised 2 governance dimensions (structures and processes) and 4 relational mechanisms (ethical principles and/or values, education and training, communication, and standards and regulations) in 4 interconnected levels (organizational, local, national, and international). The darker blue at the organizational level reflects a higher concentration of well-established and more specific AI governance frameworks for HS. For instance, the AI governance functions vary across levels but follow a complementary pattern.

The double-headed arrows highlight the dynamic nature of interactions between the dimensions and their components. For instance, the bidirectional arrow linking the structural elements (functions, actors, and roles and responsibilities) illustrates a reciprocal relationship. AI governance functions are shaped by actors who hold distinct yet complementary (and sometimes conflicting) roles and responsibilities within multidisciplinary bodies. These actors play a decisive role in defining, operationalizing, and adapting governance functions. Conversely, the governance structure and its functions can also shape, redefine, or shift the roles, relative importance, and influence of these actors over time. Similarly, the 2-way arrows between data governance, risk assessment and management, and validation and/or evaluation illustrate the interdependence among these components across the entire AI lifecycle. These 3 components are intended to ensure that AI-based technologies fulfill their goals while safeguarding public health, and individual privacy and safety.

Figure 4. The Integrative AI Governance Model for HS. Lighter colors show more heterogeneity in the findings, while darker colors emphasize more consistency in the literature.

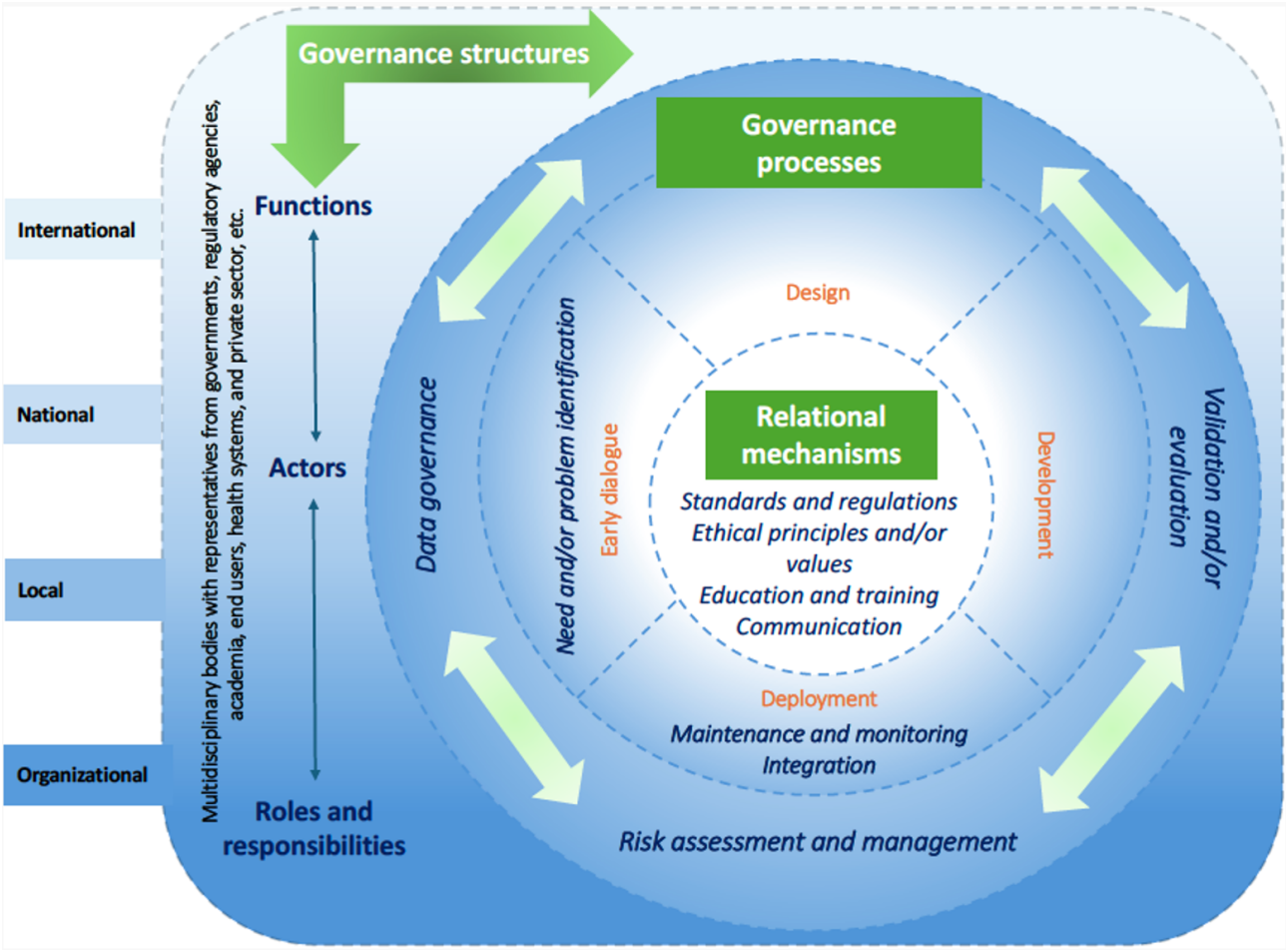

The gradient from blue to white at the center of the processes represents the adaptation and the interaction between the relational mechanisms arising from the interplay among all components of the AI governance framework for HS. Thus,

XSL•FO
RenderX

establishing ethical principles, standards, and regulations, together with clear, context-sensitive, and transparent communication strategies, as well as tailored education and training programs and initiatives adapted to different audiences (eg, health professionals, policy-makers, managers, patients, and the general population), is essential at international, national, local, and organizational levels to guide the responsible development and deployment of AI-based technologies within HS [1,30,56,64], particularly given their potential to reflect and reproduce structural inequities [67,68]. Nevertheless, their adoption and adaptation over time can differ depending on the context characteristics and resources, potentially giving rise to new relational mechanisms within AI governance processes. The 4 relational mechanisms underpin all processes, especially need and/or problem identification, maintenance and monitoring, and integration, which are critical to ensure AI and HS sustainability.

Discussion

Principal Findings

This study synthesized existing AI governance frameworks for HS published over the last decade and reviewed 19 frameworks distributed across international, national, local, and organizational levels. The frameworks articulated operational arrangements meant to translate ethical principles and legal and regulatory requirements into actionable structures, processes, and mechanisms. Consistent with prior literature reviews [10,69,70], our findings highlighted that despite the growing number of AI governance frameworks for HS, quality and implementability are largely lacking.

Contribution to the Literature

Integrative AI Governance Model for HS

This study contributed an integrative AI governance model for HS, which gathers 6 processes and 4 relational mechanisms across the 4 governance levels, explicitly modeling their interactions. Although no previous integrative framework has been proposed against which comparisons can be made, prior published general AI governance frameworks underscored similar structures, processes, and mechanisms [10,35,71]. For instance, the responsible AI governance framework of Papagiannidis et al [10] includes dimensions of privacy, data governance, safety, and bias, which are aligned with our integrative framework. Furthermore, the governance framework of Stogiannos et al [37] emphasized that HS staff training should include core AI concepts, including basic principles, model validation and evaluation, clinical applications, ethical considerations, regulatory requirements, and AI model limitations. Although this framework shared certain features with ours, it is primarily designed to inform procurement decisions [37].

Theoretical Grounding Between AI Governance and HS Governance

By proposing a multidimensional and multilevel framework, our integrative model provides a foundation to better integrate an interdisciplinary analytical lens to AI governance in HS, namely, governance theories and models anchored in the social sciences that can inform current debates on the most appropriate governance arrangements for HS [72-74].

First, robust governance is widely recognized as a central determinant of HS performance and resilience [72,75-77]. A large body of literature suggests that, among settings with comparable resources, differences in HS performance are often explained less by financing, infrastructure, or workforce levels than by the quality of governance mechanisms [72,75-77]. Conversely, dysfunctions, such as corruption, poor coordination, weak accountability mechanisms, and competing agendas, can severely undermine HS effectiveness, even when substantial financial and technical investments are made [72,75-77]. The analysis of AI governance in HS must therefore pay closer attention to the multilevel dynamics that structure these systems, particularly the institutional logics shaping governance arrangements and practices within the AI ecosystem [78-80].

Second, a multilevel analytical perspective is also warranted as HS are characterized by institutional pluralism, bringing together diverse stakeholders, such as governments, regulatory bodies, technology providers, investors, scientists, health professionals, patients, and civil society [74]. While interactions among these actors are shaped by power relations and trade-offs, several studies indicate that current AI governance in HS remains largely technocratic, overlooking the sociopolitical, economic, and even ideological dynamics inherent in AI development and use [1,81]. In this context, governance transcends the mere regulation of a technological tool; it requires arbitrating among competing (sometimes antagonistic) objectives and priorities to protect the public interest [73,82-85]. Finally, HS research offers conceptual frameworks that support a more structured analysis of interactions among stakeholders across multiple levels of governance [72,75-77]. These approaches also help examine how institutions shape relationships among actors and how resources are allocated, accessed, and mobilized within the governance system [72,75-77].

Further Research and Key Improvement Opportunities

While the proposed integrative model constitutes an important synthesis of knowledge about the governance of AI in HS, there remains significant scope for improvements.

Improving Quality

To improve quality, further research should use systematic and transparent methods, such as case studies, network analysis, and stakeholder mapping, to guide framework development and implementation and to analyze specific actors' roles and responsibilities [86-90]. In addition, data governance should be expanded beyond data security to encompass patient consent and data quality, as even the most advanced AI-based technologies depend on high-quality datasets to function effectively [91]. Data security overshadowing data quality in current governance frameworks is a key insight of this study, as dataset quality can have a significant impact on mitigating bias and ensuring equity in the delivery of care using AI technologies [92-94]. Data security should also be broadened to proactively address cyberattacks, particularly ransomware targeting health data, as well as issues of data sovereignty

XSL•FO
RenderX

[1,7,25,26,56]. Given that such threats and issues are expected to increase, integrating them into AI governance within HS is essential [95-97]. Finally, to improve comparability and rigor, further research should make AI validation and evaluation and maintenance and monitoring metrics available across all AI model levels, and include equity, economic, and environmental indicators [98-100].

Improving Processes

Further research is needed to better understand the governance process of "integration" at a broader level as AI requires a systemic transformation, which goes beyond individualized concerns for workflow redesign or workforce adaptation [7,101]. While these elements are important, they should be situated within a broader (architectural) reorganization of HS [13,25]. For example, a review examining the application of AI in radiology showed that, despite AI's substantial potential to optimize patient care, its effective use requires a reorganization of professional roles and responsibilities, as well as a redesign of regulatory frameworks [102]. Specifically, clinicians need to learn to develop competencies to supervise and appropriately use AI systems, while regulatory frameworks must evolve to balance requirements for human oversight with the increasing autonomy and agency of AI models [102]. The current emphasis on adaptation rather than systemic transformation appears misaligned with the broader ambitions often attributed to AI-based technologies [101]. This underscores the need for future studies to analyze and inform the systemic transformation within HS to fully benefit from AI.

Improving Mechanisms

Regarding the relational mechanisms, 4 main areas for improvement were identified. First, there is a need to better conceptualize and operationalize the ethical principles and/or values specific to AI in HS [103]. As many AI-based technologies reproduce and reinforce health inequities [104], scholars have underscored the need for tools that help stakeholders integrate ethical principles and/or values in emerging technologies [103,105]. Future governance frameworks could build a more constructive dialogue with previous reviews about AI ethical principles and/or values to strengthen their conceptualization and operationalization [16,17,103,106]. For instance, according to Fields et al [106], addressing racial bias in AI-based technologies requires concrete measures, including the meaningful inclusion of Black researchers, other minority scholars, and community members throughout the AI development process. It also entails conducting audits explicitly designed to detect and mitigate racial (and other) bias, strengthening transparency, and establishing robust accountability mechanisms [106].

Second, there is a need to clarify whether standards and regulations should be legally binding or merely aspirational [7]. On this issue, the WHO calls for the urgent development of governance frameworks for AI that safeguard the rights of individuals and communities, promote equity and justice, while guiding innovation and strengthening the resilience of HS [1]. Such governance must be grounded in local sociopolitical, institutional, and cultural realities, supported by robust legal, regulatory, and accountability mechanisms, and responsive to the operational challenges associated with integrating AI into health care practice [20-22].

Third, AI education and training programs within HS governance will require adaptable strategies that can keep pace with rapid developments in AI, ensuring that professionals are equipped with up-to-date knowledge and skills. These initiatives should also equip policymakers, managers, patients, the general public, and other stakeholders directly or indirectly affected by the use of AI in health [1,7,14,56]. Fourth, further research should clarify the actors engaged in communication, the direction and flow of that communication, and the specific target audiences (eg, clinicians, patients, and public), as this dimension remains insufficiently analyzed and developed [107,108]. In sum, to strengthen the model's institutional and practical relevance, future research should prioritize the validation and operationalization of the integrative AI governance model. This agenda could be advanced through Delphi studies involving diverse stakeholders across multiple HS levels, complemented by case studies conducted in varied geographic, sociopolitical, cultural, and economic contexts to refine the model and support its broader applicability [90,109-111].

Strengths and Limitations of the Study

To the best of our knowledge, this systematic review is the first to synthesize existing AI governance frameworks across 4 governance levels while simultaneously assessing their quality. Yet, this study has certain limitations. First, although our search strategy was comprehensive and extended to gray literature, it remains possible that frameworks in progress or those disseminated in other languages were not identified. Regarding language restrictions, we acknowledged that relevant frameworks may exist in other languages, particularly in countries that are advanced in AI development (eg, China and Germany). However, we did not conduct literature searches in these languages as our team does not possess the linguistic expertise required to analyze documents in these languages. Relying on automated translation tools was not considered an appropriate option, as these tools pose a substantial risk of misinterpreting concepts and terms whose meaning depends heavily on context. Consequently, our model carries a Western bias, as the underlying evidence is primarily sourced from the Global North. This disparity underscores a critical need for future research to investigate AI governance within contexts of the Global South.

Second, a limitation arises from the use of an adapted version of AGREE-HS to evaluate the methodological quality of each included framework. Although this tool is the best option to evaluate whether the framework's authors explicitly define the topic (what), participants (who), and methods (how), several AGREE-HS items are interpretative in nature requiring contextual judgment. However, we ensured the robustness and consistency of data extraction and appraisal through independent coding, calibration, and consensus-based adjudication. This approach is commonly adopted in reviews incorporating qualitative appraisal, where calibration, discussion, and consensus are recognized as appropriate strategies to ensure methodological rigor [112]. Finally, the heterogeneity of this body of literature may have constrained the extent to which the

XSL•FO
RenderX

distinctive nuances of each framework were fully captured in the results. To mitigate this limitation, our review and our model were developed systematically, following rigorous methodological guidance [39], to better examine diverse frameworks spanning multiple AI governance components and contexts.

Conclusions

This systematic review of AI governance frameworks for HS highlights key elements of AI governance structures, processes, and mechanisms distributed across international, national, local, and organizational levels. Rather than focusing on specific governance dimensions or settings, the review innovates by integrating all governance components and their interactions across contexts. This study thus differs from other reviews by contributing a multilevel analysis that elucidates the strengths and gaps within current frameworks, adding to this expanding body of scholarship. The study brings to the field an integrative AI governance model for HS. The proposed model may inform future scholarly work as well as real-world policy and practice-oriented discussions on ensuring the responsible and sustainable integration of AI-based technologies into HS.

Acknowledgments

We would like to thank research collaborators who provided us with insightful comments throughout the study: Gloria Liliana Palomino, Racha Soubra, Soutongnoma Safiata Kabore. The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT (Generative Artificial Intelligence Delegation Taxonomy) [113], the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing. The GenAI tool used was ChatGPT-4 and ChatGPT-5.x. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

The study was supported by the International Observatory on the Societal Impacts of Artificial Intelligence and Digital Technologies (grant FRQSC- FS12207); and the Institute for Data Valorization (IVADO), Canada. The funders were not involved in the study design, data collection, analysis, interpretation, decision to publish, or the writing of the manuscript.

Data Availability

The datasets used and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

HA contributed to conceptualization, methodology, supervision, investigation, data curation, formal analysis, writing of the original draft, manuscript review and editing, project administration, and funding acquisition. RPS, EJP, and LR contributed to conceptualization, methodology, investigation, data curation, formal analysis, writing of the original draft, and manuscript review and editing. MPG, JLD, LL, and KM contributed to conceptualization, methodology, investigation, manuscript review and editing, and funding acquisition. MS, MAAA, and JPF contributed to conceptualization, methodology, validation, and manuscript review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Literature search strategy.
[PDF File, 154 KB - jmir_v28i1e87448_app1.pdf]

Multimedia Appendix 2
Modifications to the Appraisal of Guidelines for Research and Evaluation for HS (AGREE-HS).
[PDF File, 141 KB - jmir_v28i1e87448_app2.pdf]

Multimedia Appendix 3
Study characteristics (n=19).
[PDF File, 238 KB - jmir_v28i1e87448_app3.pdf]

Multimedia Appendix 4
Characteristics of the AI governance frameworks for HS.
[PDF File, 185 KB - jmir_v28i1e87448_app4.pdf]

XSL•FO
RenderX

Multimedia Appendix 5
AI governance structure functions identified across the included frameworks.
[PDF File, 238 KB - jmir_v28i1e87448_app5.pdf]

Multimedia Appendix 6
Actors' roles and responsibilities.
[PDF File, 275 KB - jmir_v28i1e87448_app6.pdf]

Checklist 1
PRISMA checklist.
[PDF File, 186 KB - jmir_v28i1e87448_app7.pdf]

References

1. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240029200 [accessed 2025-01-11]
2. Ennis-O'Connor M, O'Connor WT. Charting the future of patient care: a strategic leadership guide to harnessing the potential of artificial intelligence. Healthc Manage Forum 2024 Jul;37(4):290-295. [doi: 10.1177/08404704241235893] [Medline: 38441043]
3. Baig MA, Almuhaizea MA, Alshehri J, Bazarbashi MS, Al-Shagathrh F. Urgent need for developing a framework for the governance of AI in healthcare. Stud Health Technol Inform 2020 Jun 26;272:253-256. [doi: 10.3233/SHTI200542] [Medline: 32604649]
4. Upadhyay U, Gradisek A, Iqbal U, Dhar E, Li YC, Syed-Abdul S. Call for the responsible artificial intelligence in the healthcare. BMJ Health Care Inform 2023 Dec 21;30(1):e100920. [doi: 10.1136/bmjhci-2023-100920] [Medline: 38135293]
5. Recommendation of the council on artificial intelligence. Organisation for Economic Co-operation and Development. 2019. URL: https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449 [accessed 2025-01-22]
6. Bodnari A, Travis J. Scaling enterprise AI in healthcare: the role of governance in risk mitigation frameworks. NPJ Digit Med 2025 May 13;8(1):272. [doi: 10.1038/s41746-025-01700-4] [Medline: 40360659]
7. Esmaeilzadeh P. Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: a perspective for healthcare organizations. Artif Intell Med 2024 May;151:102861. [doi: 10.1016/j.artmed.2024.102861] [Medline: 38555850]
8. Ramezani M, Takian A, Bakhtiari A, Rabiee HR, Ghazanfari S, Sazgarnejad S. Research agenda for using artificial intelligence in health governance: interpretive scoping review and framework. BioData Min 2023 Oct 31;16(1):31. [doi: 10.1186/s13040-023-00346-w] [Medline: 37904172]
9. Shipton L, Vitale L. Artificial intelligence and the politics of avoidance in global health. Soc Sci Med 2024 Oct;359:117274. [doi: 10.1016/j.socscimed.2024.117274] [Medline: 39217716]
10. Papagiannidis E, Mikalef P, Conboy K. Responsible artificial intelligence governance: a review and research framework. The Journal of Strategic Information Systems 2025 Jun;34(2):101885. [doi: 10.1016/j.jsis.2024.101885]
11. You Z, Wang Y, Xiao Y. Analysing the suitability of artificial intelligence in healthcare and the role of AI governance. Health Care Anal 2025 Dec;33(4):413-445. [doi: 10.1007/s10728-025-00514-5] [Medline: 40048068]
12. Alami H, Lehoux P, Auclair Y, et al. Artificial intelligence and health technology assessment: anticipating a new level of complexity. J Med Internet Res 2020 Jul 7;22(7):e17707. [doi: 10.2196/17707] [Medline: 32406850]
13. Alami H, Lehoux P, Denis JL, et al. Organizational readiness for artificial intelligence in health care: insights for decision-making and practice. J Health Organ Manag 2020 Dec 3;ahead-of-print(ahead-of-print):106-114. [doi: 10.1108/JHOM-03-2020-0074] [Medline: 33258359]
14. Stogiannos N, Litosseliti L, O'Regan T, et al. Black box no more: a cross-sectional multi-disciplinary survey for exploring governance and guiding adoption of AI in medical imaging and radiotherapy in the UK. Int J Med Inform 2024 Jun;186:105423. [doi: 10.1016/j.ijmedinf.2024.105423] [Medline: 38531254]
15. Wang Y, Li N, Chen L, et al. Guidelines, consensus statements, and standards for the use of artificial intelligence in medicine: systematic review. J Med Internet Res 2023 Nov 22;25:e46089. [doi: 10.2196/46089] [Medline: 37991819]
16. Mohsin Khan M, Shah N, Shaikh N, Thabet A, Alrabayah T, Belkhair S. Towards secure and trusted AI in healthcare: A systematic review of emerging innovations and ethical challenges. Int J Med Inform 2025 Mar;195:105780. [doi: 10.1016/j.ijmedinf.2024.105780] [Medline: 39753062]
17. Korukoglu FB, Rishan RM, Elhamer Z, Lakamana S, Zhang J, Bozkurt S. Mapping ethical guidelines for AI in healthcare: a global perspective. Stud Health Technol Inform 2025 May 15;327:1358-1362. [doi: 10.3233/SHTI250624] [Medline: 40380727]
18. Shiferaw KB, Roloff M, Balaur I, Welter D, Waltemath D, Zeleke AA. Guidelines and standard frameworks for artificial intelligence in medicine: a systematic review. JAMIA Open 2025 Feb;8(1):ooae155. [doi: 10.1093/jamiaopen/ooae155] [Medline: 39759773]

XSL•FO
RenderX

19. Jackson BR, Rashidi HH, Lennerz JK, de Baca ME. Ethical and regulatory perspectives on generative artificial intelligence in pathology. Arch Pathol Lab Med 2025 Feb 1;149(2):123-129. [doi: 10.5858/arpa.2024-0205-RA] [Medline: 39299704]
20. National eHealth Strategy Toolkit. World Health Organization. 2012. URL: https://www.who.int/publications/i/item/national-ehealth-strategy-toolkit [accessed 2026-01-10]
21. Marcelo A, Medeiros D, Ramesh K, Roth S, Wyatt P. Transforming health systems through good digital health governance. ADB. URL: https://www.adb.org/publications/transforming-health-systems-good-digital-health-governance [accessed 2026-01-29]
22. Khoja S, Durrani H, Nayani P, Fahim A. Scope of policy issues in eHealth: results from a structured literature review. J Med Internet Res 2012 Feb 17;14(1):e34. [doi: 10.2196/jmir.1633] [Medline: 22343270]
23. Perry B, Uuk R. AI governance and the policymaking process: key considerations for reducing AI risk. BDCC 2019;3(2):26. [doi: 10.3390/bdcc3020026]
24. Dafoe A. AI governance: a research agenda. : University of Oxford; 2018 URL: https://cdn.governance.ai/GovAI-Research-Agenda.pdf [accessed 2026-01-30]
25. Alami H, Lehoux P, Papoutsi C, Shaw SE, Fleet R, Fortin JP. Understanding the integration of artificial intelligence in healthcare organisations and systems through the NASSS framework: a qualitative study in a leading Canadian academic centre. BMC Health Serv Res 2024 Jun 3;24(1):701. [doi: 10.1186/s12913-024-11112-x] [Medline: 38831298]
26. Alami H, Rivard L, Lehoux P, et al. Conflicts and complexities around intellectual property and value sharing of artificial intelligence healthcare solutions in public-private partnerships: a qualitative study. SSM - Health Systems 2025 Dec;5:100093. [doi: 10.1016/j.ssmhs.2025.100093]
27. Freeman S, Wang A, Saraf S, et al. Developing an AI governance framework for safe and responsible AI in health care organizations: protocol for a multimethod study. JMIR Res Protoc 2025 Jul 28;14:e75702. [doi: 10.2196/75702] [Medline: 40720809]
28. Kim JY, Hasan A, Kueper J, et al. Establishing organizational AI governance in healthcare: a case study in Canada. NPJ Digit Med 2025 Aug 15;8(1):522. [doi: 10.1038/s41746-025-01909-3] [Medline: 40817280]
29. Reddy S. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. Implement Sci 2024 Mar 15;19(1):27. [doi: 10.1186/s13012-024-01357-9] [Medline: 38491544]
30. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. J Am Med Inform Assoc 2020 Mar 1;27(3):491-497. [doi: 10.1093/jamia/ocz192] [Medline: 31682262]
31. Scott IA, van der Vegt A, Canaris S, Nolan P, Pointon K. Preparing healthcare organisations for using artificial intelligence effectively. Aust Health Rev 2025 Jul;49(4):AH25102. [doi: 10.1071/AH25102] [Medline: 40717026]
32. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
33. Carnicero J, Serra P. Governance for digital health: the art of health systems transformation. : Inter-American Development Bank; 2020 URL: https://publications.iadb.org/en/governance-digital-health-art-health-systems-transformation [accessed 2025-03-20]
34. Lupiáñez-Villanueva F, Gunderson L, Vitiello S, Febrer N, Folkvord F, Chabanier L, et al. Study on health data, digital health and artificial intelligence in healthcare. : Publications Office of the European Union; 2022 URL: https://data.europa.eu/doi/10.2875/702007 [accessed 2026-01-29]
35. Mäntymäki M, Minkkinen M, Birkstedt T, Viljanen M. Defining organizational AI governance. AI Ethics 2022 Nov;2(4):603-609. [doi: 10.1007/s43681-022-00143-x]
36. Ekeland AG, Linstad LH. Elaborating models of eHealth governance: qualitative systematic review. J Med Internet Res 2020 Oct 28;22(10):e17214. [doi: 10.2196/17214] [Medline: 33112247]
37. Stogiannos N, Malik R, Kumar A, et al. Black box no more: a scoping review of AI governance frameworks to guide procurement and adoption of AI in medical imaging and radiotherapy in the UK. Br J Radiol 2023 Dec;96(1152):20221157. [doi: 10.1259/bjr.20221157] [Medline: 37747285]
38. Crossnohere NL, Elsaid M, Paskett J, Bose-Brill S, Bridges JFP. Guidelines for artificial intelligence in medicine: literature review and content analysis of frameworks. J Med Internet Res 2022 Aug 25;24(8):e36823. [doi: 10.2196/36823] [Medline: 36006692]
39. Page MJ, Moher D, Bossuyt PM, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021 Mar 29;372:n160. [doi: 10.1136/bmj.n160] [Medline: 33781993]
40. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
41. Alami H, Pérez EJ, Sabio RP, Denis JL, Gagnon MP, Langlois L, et al. Artificial intelligence governance in the health system: protocol for a systematic review of frameworks. OSF 2025 Nov 9. [doi: 10.2196/preprints.87448]
42. Chesterman S, Gao Y, Hahn J, Sticher V. The evolution of AI governance. Computer (Long Beach Calif) 2024;57(9):80-92. [doi: 10.1109/MC.2024.3381215]

XSL•FO
RenderX

43. Hussain W, Mabrok M, Gao H, Rabhi FA, Rashed EA. Revolutionising healthcare with artificial intelligence: a bibliometric analysis of 40 years of progress in health systems. Digit Health 2024;10:20552076241258757. [doi: 10.1177/20552076241258757] [Medline: 38817839]
44. Mouloudj K, Bouarar A, Bouarar AC, Asanza DM, Srivastava M. Adopting artificial intelligence in healthcare: a narrative review. In: Teixeira S, J R, editors. The Use of Artificial Intelligence in Digital Marketing: Competitive Strategies and Tactics IGI Global 2023:1-20. [doi: 10.4018/978-1-6684-9324-3.ch001]
45. The world's #1 systematic review tool. Covidence. 2025. URL: www.covidence.org [accessed 2026-03-24]
46. Brouwers MC, Lavis JN, Spithoff K, et al. Assessment of health systems guidance using the Appraisal of Guidelines for Research and Evaluation - Health Systems (AGREE-HS) instrument. Health Policy 2019 Jul;123(7):646-651. [doi: 10.1016/j.healthpol.2019.05.004] [Medline: 31160062]
47. The appraisalof guidelines research and evaluation -the AGREE II instrument. AGREE Next Steps Consortium. 2017. URL: http://www.agreetrust.org [accessed 2026-03-25]
48. Van Grembergen W, De Haes S, Guldentops E. Structures, processes and relational mechanisms for IT governance. In: Strategies for Information Technology Governance: IGI Global; 2004:1-36. [doi: 10.4018/978-1-59140-140-7.ch001]
49. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
50. Liao F, Adelaine S, Afshar M, Patterson BW. Governance of Clinical AI applications to facilitate safe and equitable deployment in a large health system: key elements and early successes. Front Digit Health 2022;4:931439. [doi: 10.3389/fdgth.2022.931439] [Medline: 36093386]
51. Morley J, Murphy L, Mishra A, Joshi I, Karpathakis K. Governing data and artificial intelligence for health care: developing an international understanding. JMIR Form Res 2022 Jan 31;6(1):e31623. [doi: 10.2196/31623] [Medline: 35099403]
52. Apfelbacher T, Koçman SE, Prokosch HU, Christoph J. A governance framework for the implementation and operation of AI applications in a university hospital. Stud Health Technol Inform 2024 Aug 22;316:776-780. [doi: 10.3233/SHTI240527] [Medline: 39176908]
53. Carter SM, Aquino YSJ, Carolan L, et al. How should artificial intelligence be used in Australian health care? Recommendations from a citizens' jury. Med J Aust 2024 May 6;220(8):409-416. [doi: 10.5694/mja2.52283] [Medline: 38629188]
54. Whittaker R, Dobson R, Jin CK, et al. An example of governance for AI in health services from Aotearoa New Zealand. NPJ Digit Med 2023 Sep 1;6(1):164. [doi: 10.1038/s41746-023-00882-z] [Medline: 37658119]
55. Kim JY, Boag W, Gulamali F, et al. Organizational governance of emerging technologies: AI adoption in healthcare. 2023 Jun 12 Presented at: FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency; Jun 12-15, 2023; Chicago IL USA p. 1396-1417 URL: https://dl.acm.org/doi/proceedings/10.1145/3593013 [doi: 10.1145/3593013.3594089]
56. World Health Organization. Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models: World Health Organization; 2024. URL: https://www.who.int/publications/i/item/9789240084759 [accessed 2025-01-11]
57. Daye D, Wiggins WF, Lungren MP, et al. Implementation of clinical artificial intelligence in radiology: who decides and how? Radiology 2022 Dec;305(3):555-563. [doi: 10.1148/radiol.212151] [Medline: 35916673]
58. Hassan M, Borycki EM, Kushniruk AW. Artificial intelligence governance framework for healthcare. Healthc Manage Forum 2025 Mar;38(2):125-130. [doi: 10.1177/08404704241291226] [Medline: 39470044]
59. Parker VJ, Economou NJ, Silcox C. AI governance in health systems: aligning innovation, accountability and trust. Duke Health. 2024. URL: https://healthaigovernance.duke.edu/news/white-paper-ai-governance-health-systems-aligning-innovation-accountability-and-trust [accessed 2025-01-11]
60. Kim JY, Hasan A, Balu S, Sendak M. People process technology and operations framework for establishing AI governance in healthcare organizations. npj Digit Med 2026;9(1). [doi: 10.1038/s41746-026-02419-6]
61. Bedoya AD, Economou-Zavlanos NJ, Goldstein BA, et al. A framework for the oversight and local deployment of safe and high-quality prediction models. J Am Med Inform Assoc 2022 Aug 16;29(9):1631-1636. [doi: 10.1093/jamia/ocac078] [Medline: 35641123]
62. Arnaout A, Gill P, Virani A, et al. Shaping the future of healthcare in British Columbia: establishing provincial clinical governance for responsible deployment of artificial intelligence tools. Healthc Manage Forum 2024 Sep;37(5):320-328. [doi: 10.1177/08404704241264819] [Medline: 39030752]
63. Economou-Zavlanos NJ, Bessias S, Cary MP Jr, et al. Translating ethical and quality principles for the effective, safe and fair development, deployment and use of artificial intelligence technologies in healthcare. J Am Med Inform Assoc 2024 Feb 16;31(3):705-713. [doi: 10.1093/jamia/ocad221] [Medline: 38031481]
64. A national policy roadmap for artificial intelligence in healthcare. Australian Alliance for Artificial Intelligence in Healthcare (AAAiH). 2023. URL: https://aihealthalliance.org/wp-content/uploads/2023/11/AAAiH_NationalPolicyRoadmap_FINAL.pdf [accessed 2025-05-05]

XSL•FO
RenderX

65. Jaremko JL, Azar M, Bromwich R, et al. Canadian Association of Radiologists White Paper on ethical and legal issues related to artificial intelligence in radiology. Can Assoc Radiol J 2019 May;70(2):107-118. [doi: 10.1016/j.carj.2019.03.001] [Medline: 30962048]
66. Solaiman B. From bench to bedside: governing health care artificial intelligence (AI) through a “true lifecycle approach”. Am J Law Med 2025 Dec;51(3-4):452-478. [doi: 10.1017/amj.2025.10091] [Medline: 41700404]
67. Benjamin R. Assessing risk, automating racism. Science 2019 Oct 25;366(6464):421-422. [doi: 10.1126/science.aaz3873] [Medline: 31649182]
68. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
69. Birkstedt T, Minkkinen M, Tandon A, Mäntymäki M. AI governance: themes, knowledge gaps and future agendas. INTR 2023 Dec 18;33(7):133-167. [doi: 10.1108/INTR-01-2022-0042]
70. Batool A, Zowghi D, Bano M. AI governance: a systematic literature review. In Review: Research Square Preprint posted online on 2025. [doi: 10.21203/rs.3.rs-4784792/v1]
71. Sharma S. Trustworthy artificial intelligence: design of AI governance framework. Strategic Analysis 2023 Sep 3;47(5):443-464. [doi: 10.1080/09700161.2023.2288994]
72. Olmen JV, Criel B, Bhojani U, et al. The Health System Dynamics Framework: the introduction of an analytical model for health system analysis and its application to two case-studies. hcs 2012;2(1):1-21. [doi: 10.5195/hcs.2012.71]
73. Ulnicane I. Governance fix? Power and politics in controversies about governing generative AI. Policy and Society 2025 Jan 4;44(1):70-84. [doi: 10.1093/polsoc/puae022]
74. Denis JL, Langley A, Rouleau L. Strategizing in pluralistic contexts: rethinking theoretical frames. Human Relations 2007 Jan;60(1):179-215. [doi: 10.1177/0018726707075288]
75. Mikkelsen-Lopez I, Wyss K, de Savigny D. An approach to addressing governance from a health system framework perspective. BMC Int Health Hum Rights 2011 Dec 2;11(1):13. [doi: 10.1186/1472-698X-11-13] [Medline: 22136318]
76. Blanchet K, Nam SL, Ramalingam B, Pozo-Martin F. Governance and capacity to manage resilience of health systems: towards a new conceptual framework. Int J Health Policy Manag 2017 Aug 1;6(8):431-435. [doi: 10.15171/ijhpm.2017.36] [Medline: 28812842]
77. Kruk ME, Ling EJ, Bitton A, et al. Building resilient health systems: a proposal for a resilience index. BMJ 2017 May 23;357:j2323. [doi: 10.1136/bmj.j2323] [Medline: 28536191]
78. Lounsbury M, Steele CWJ, Wang MS, Toubiana M. New directions in the study of institutional logics: from tools to phenomena. Annu Rev Sociol 2021 Jul 31;47(1):261-280. [doi: 10.1146/annurev-soc-090320-111734]
79. Denis JL. Transformation de la gouvernance des systèmes de santé: analyse de la littérature. CSBE. 2023. URL: https://www.csbe.gouv.qc.ca/publication/transformation-gouvernance-systemes-de-sante-analyse-litterature.html [accessed 2026-01-22]
80. Gjaltema J, Biesbroek R, Termeer K. From government to governance…to meta-governance: a systematic literature review. Public Management Review 2020 Dec 1;22(12):1760-1780. [doi: 10.1080/14719037.2019.1648697]
81. Shaffer J, Alenichev A, Faure MC. The Gates Foundation’s new AI initiative: attempting to leapfrog global health inequalities? BMJ Glob Health 2023 Nov;8(11):e013874. [doi: 10.1136/bmjgh-2023-013874] [Medline: 37923320]
82. Taeihagh A. Governance of generative AI. Policy and Society 2025 Jan 4;44(1):1-22. [doi: 10.1093/polsoc/puaf001]
83. Wen Y. Public interest v.s. special interest: the strategic framing tactics of technologies in the political arena. Res Policy 2024 Oct;53(8):105071. [doi: 10.1016/j.respol.2024.105071]
84. Farrell H. AI as governance. Annu Rev Polit Sci (Palo Alto) 2025 Jun 17;28(1):375-392. [doi: 10.1146/annurev-polisci-040723-013245]
85. Cohen T, Suzor NP. Contesting the public interest in AI governance. IPR 2024;13(3). [doi: 10.14763/2024.3.1794]
86. Walker DHT, Bourne LM, Shelley A. Influence, stakeholder mapping and visualization. Construction Management and Economics 2008 Jun;26(6):645-658. [doi: 10.1080/01446190701882390]
87. Cottafava D, Corazza L. Co-design of a stakeholders’ ecosystem: an assessment methodology by linking social network analysis, stakeholder theory and participatory mapping. K 2021 Mar 27;50(3):836-858. [doi: 10.1108/K-12-2019-0861]
88. Tabassum S, Pereira FSF, Fernandes S, Gama J. Social network analysis: an overview. WIREs Data Min & Knowl 2018 Sep;8(5):e1256. [doi: 10.1002/widm.1256]
89. Scott J, Carrington P. The SAGE Handbook of Social Network Analysis: SAGE Publications; 2011:11-25. [doi: 10.4135/9781446294413.n2]
90. Yin RK. Case Study Research and Applications: Design and Methods: SAGE Publications; 2017.
91. Whang SE, Roh Y, Song H, Lee JG. Data collection and quality challenges in deep learning: a data-centric AI perspective. VLDB J 2023 Jul;32(4):791-813. [doi: 10.1007/s00778-022-00775-9]
92. Barbosa da Silva J Jr, Birminghamm M, Rivière Cinnamond A, et al. Bias-mitigated AI as a foundation for resilient and effective health systems. JMIR Public Health Surveill 2026 Feb 23;12(1):e88457. [doi: 10.2196/88457] [Medline: 41730169]
93. Chinta SV, Wang Z, Palikhe A, et al. AI-driven healthcare: fairness in AI healthcare: a survey. PLOS Digit Health 2025 May;4(5):e0000864. [doi: 10.1371/journal.pdig.0000864] [Medline: 40392801]

XSL•FO
RenderX

94. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: implications for clinical decision-making. PLOS Digit Health 2024 Nov;3(11):e0000651. [doi: 10.1371/journal.pdig.0000651] [Medline: 39509461]
95. Di Palma G, Scendoni R, Ferorelli D, De Benedictis A, Tambone V, De Micco F. AI-induced cybersecurity risks in healthcare: a narrative review of blockchain-based solutions within a clinical risk management framework. Risk Manag Healthc Policy 2025;18:3479-3497. [doi: 10.2147/RMHP.S544523] [Medline: 41185742]
96. Mariettou S, Koutsojannis C, Triantafillou V. Artificial intelligence and algorithmic approaches of health security systems: a review. Algorithms 2025;18(2):59. [doi: 10.3390/a18020059]
97. Li S, Surineni K, Prabhakaran N. Cyber-attacks on hospital systems: a narrative review. The American Journal of Geriatric Psychiatry: Open Science, Education, and Practice 2025 Sep;7:30-39. [doi: 10.1016/j.osep.2025.03.002] [Medline: 41404441]
98. Zhao L, Liu S, Xin T, et al. AI agent in healthcare: applications, evaluations, and future directions. npj Artif Intell 2026;2(1):31. [doi: 10.1038/s44387-026-00076-4]
99. Ahadian P, Xu W, Liu D, Guan Q. Ethics of trustworthy AI in healthcare: challenges, principles, and practical pathways. Neurocomputing 2026 Jan;661:131942. [doi: 10.1016/j.neucom.2025.131942]
100. Chin MH, Afsar-Manesh N, Bierman AS, et al. Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. JAMA Netw Open 2023 Dec 1;6(12):e2345050. [doi: 10.1001/jamanetworkopen.2023.45050] [Medline: 38100101]
101. Bekbolatova M, Mayer J, Ong CW, Toma M. Transformative potential of AI in healthcare: definitions, applications, and navigating the ethical landscape and public perspectives. Healthcare (Basel) 2024 Jan 5;12(2):125. [doi: 10.3390/healthcare12020125] [Medline: 38255014]
102. Mahedi RA, Iqbal H, Azmee R, et al. Current trends and future prospects of artificial intelligence in transforming radiology. J Curr Health Sci 2024 Nov;4(2):95-104. [doi: 10.47679/jchs.202487]
103. Lopez-Ramos LM, Pluktaite G, Bui CKT, et al. Operationalizing AI ethics in medicine—a co-creation workshop study. BMC Med Ethics 2025;26(1):150. [doi: 10.1186/s12910-025-01317-y]
104. Owens K, Walker A. Those designing healthcare algorithms must become actively anti-racist. Nat Med 2020 Sep;26(9):1327-1328. [doi: 10.1038/s41591-020-1020-3] [Medline: 32908272]
105. Mathews DJH, Balatbat CA, Dzau VJ. Governance of emerging technologies in health and medicine — creating a new framework. N Engl J Med 2022 Jun 9;386(23):2239-2242. [doi: 10.1056/NEJMms2200907]
106. Fields CT, Black C, Thind JK, et al. Governance for anti-racist AI in healthcare: integrating racism-related stress in psychiatric algorithms for Black Americans. Front Digit Health 2025;7:1492736. [doi: 10.3389/fdgth.2025.1492736] [Medline: 40444183]
107. McWhirter JE, MacKay M, McAlpine D, Grant LE, Papadopoulos A. The Public Health Communication Competency Framework: Results from a multi-method and consensus-building process. Can J Public Health 2025 Nov 12;2025:1-18. [doi: 10.17269/s41997-025-01107-4] [Medline: 41225276]
108. Niederdeppe J, Boyd AD, King AJ, Rimal RN. Strategies for effective public health communication in a complex information environment. Annu Rev Public Health 2025 Apr;46(1):411-431. [doi: 10.1146/annurev-publhealth-071723-120721] [Medline: 39656948]
109. Linstone HA, Turoff M. The Delphi Method: Techniques and Applications: Addison-Wesley Pub. Co; 1975.
110. Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. J Adv Nurs 2000 Oct;32(4):1008-1015. [doi: 10.1046/j.1365-2648.2000.t01-1-01567.xDigital] [Medline: 11095242]
111. Okoli C, Pawlowski SD. The Delphi method as a research tool: an example, design considerations and applications. Information & Management 2004 Dec;42(1):15-29. [doi: 10.1016/j.im.2003.11.002]
112. O'Connor C, Joffe H. Intercoder reliability in qualitative research: debates and practical guidelines. Int J Qual Methods 2020 Jan 1;19:1609406919899220. [doi: 10.1177/1609406919899220]
113. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (Generative AI Delegation Taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res 2026 Apr;33(3):2544331. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]

Abbreviations

AAAiH: Australian Alliance for Artificial Intelligence in Healthcare
AGREE-HS: Appraisal of Guidelines for Research and Evaluation for Health Systems
AI: artificial intelligence
HS: health systems
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SSRN: Social Science Research Network
WHO: World Health Organization

XSL•FO
RenderX

Edited by S Brini; submitted 09.Nov.2025; peer-reviewed by H Mukhlis, T Jimoh; revised version received 21.Apr.2026; accepted 01.May.2026; published 08.Jun.2026.

Please cite as:
Alami H, Pozelli Sabio R, Pérez EJ, Gagnon MP, Langlois L, Denis JL, Malas K, Rivard L, Savoldelli M, Ag Ahmed MA, Fortin JP
Artificial Intelligence Governance in Health Systems: Systematic Review of Frameworks and Integrative Model Proposal
J Med Internet Res 2026;28:e87448
URL: https://www.jmir.org/2026/1/e87448
doi:10.2196/87448

© Hassane Alami, Renata Pozelli Sabio, Elsury Johanna Pérez, Marie-Pierre Gagnon, Lyse Langlois, Jean-Louis Denis, Kathy Malas, Lysanne Rivard, Mathilde Savoldelli, Mohamed Ali Ag Ahmed, Jean-Paul Fortin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Measuring eHealth Literacy in the European Economic Area, Switzerland, and the United Kingdom: Scoping Review

Jonah Valentin Weist[1,2], MSc; Julia Nitsche[2], Dr rer medic; Jan Peter Ehlers[2], Prof Dr med vet; Theresa Sophie Busse[1], Prof Dr rer medic

[1]Junior Professorship for Digital Health, Faculty of Health, Witten/Herdecke University, Alfred-Herrhausen-Straße 50, Witten, Germany

[2]Chair of Didactics and Educational Research in Healthcare, Faculty of Health, Witten/Herdecke University, Witten, Germany

Corresponding Author:
Jonah Valentin Weist, MSc
Junior Professorship for Digital Health, Faculty of Health, Witten/Herdecke University, Alfred-Herrhausen-Straße 50, Witten, Germany

Abstract

Background: Digital tools continue to evolve and have the potential to improve health care delivery. However, they are associated with challenges, including accessibility issues and health misinformation. Individuals need eHealth literacy (eHL) to reliably use these tools, and providers require appropriate eHL measurement approaches to offer targeted solutions. For around 2 decades, researchers have been operationalizing and measuring eHL.

Objective: This paper aims to provide an up-to-date overview of how eHL has been assessed in recent years in the European Economic Area, Switzerland, and the United Kingdom and which methodological limitations need to be considered.

Methods: A scoping review was conducted. Records were searched via CINAHL, PubMed, and Google Scholar on January 31, 2025, and January 28, 2026. Peer-reviewed empirical papers published in German or English since 2020 that measured eHL in the European Economic Area, Switzerland, or the United Kingdom were included. The synthesis covered the publication trend, eHL measurement approaches and associated limitations reported in the included papers, eHL measurement frequency, countries and languages, and samples.

Results: In the final analysis, 132 papers published between 2020 and 2025 were included. The number of publications per year showed an overall upward trend (2020: 11/132, 8.33%; 2025: 35/132, 26.52%). Nine self-report eHL measurement instruments were used, the eHealth Literacy Scale being the most frequent (94/132, 71.21%). All included papers (132/132, 100%) reported data collection via surveys, and digital surveys were common (71/132, 53.79%). Reported limitations included potential self-report biases (37/132, 28.03%), selection biases due to the data collection modes (26/132, 19.70%), and limitations specific to 5 eHL measurement instruments (18/132, 13.64%). Most included papers (121/132, 91.67%) reported eHL results from a single measurement per participant. Data were collected in 22 countries within the target regions. The distribution of eHL measurement instruments varied considerably between countries. Data were collected in 22 languages within the target regions. Patients or individuals with health problems were a frequent target group (64/132, 48.48%). Most papers (104/132, 78.79%) described studies covering broad adult age ranges. Sample sizes ranged from ≤50 (7/132, 5.30%) to >2000 participants (8/132, 6.06%).

Conclusions: To our knowledge, this is the first scoping review synthesizing eHL measurement limitations reported in empirical papers from the European Economic Area, Switzerland, and the United Kingdom. The identified limitations reported in the included papers potentially lead to biased results. Therefore, health care providers and researchers should take various factors into account when selecting eHL measurement approaches, such as eHL measurement purpose, target population, and data collection setting. Future research should address these constraints by adapting and developing new or revised eHL measurement instruments, including translated and culturally adapted versions. Policymakers should encourage health care providers to conduct methodologically well-founded eHL measurements as a basis for targeted solutions.

Trial Registration: PROSPERO CRD42025642890; https://www.crd.york.ac.uk/PROSPERO/view/CRD42025642890

(J Med Internet Res 2026;28:e87461) doi:10.2196/87461

KEYWORDS

assessment; digital health; eHealth; electronic health literacy; health literacy; measurement instruments; operationalization; scoping review; telehealth; telemedicine

XSL•FO
RenderX

Introduction

Rationale

The internet is used by many individuals in the context of health. In a 2024 survey conducted in the European Union, 58% of the participants aged between 16 and 74 years reported that they had used the internet to search for health-related information within the last 3 months [1]. The internet offers more than access to health-related information. Today, a variety of digital tools offer further use cases in the context of health. For example, users can generate health information via generative artificial intelligence (GenAI)–based chatbots [2] and share health information or interact with others via social media platforms [3]. In addition, dedicated health applications such as Apple Health [4] and Google Fit [5] help individuals track health goals via smartphones and smart watches. The World Health Organization provides an overview of digital interventions for individuals in the context of health that enable communication, tracking, reporting, financial transactions, and consent management [6].

However, digital tools are also associated with obstacles for individuals who lack access to hardware and internet or digital literacy [7]. Subgroups, such as older adults, are particularly at risk of digital exclusion [8,9]. Furthermore, users are confronted with misinformation on health topics. A review of studies that analyzed health-related posts on social media platforms such as Facebook, Instagram, and Twitter, now called X, indicates that health misinformation is widespread [10]. Similarly, studies show that GenAI models generate unreliable health information in some cases [11,12]. Users who follow incorrect advice generated by ChatGPT may be exposed to serious health risks [12].

As often argued in the literature, sufficient eHealth literacy (eHL) is required for the reliable use of digital tools in the context of health [13-16]. Norman and Skinner [13] proposed an early and influential definition of eHL in 2006, long before the GenAI era. They defined eHL as “the ability to seek, find, understand, and appraise health information from electronic sources and apply the knowledge gained to addressing or solving a health problem” ([13], p. 1). Information on eHL levels is relevant for providers of digital interventions to develop targeted solutions or physicians to suggest appropriate digital health interventions and to evaluate the level of support required by individual patients [17].

Different instruments have been developed to measure eHL and used to varying degrees in empirical research. The most common eHL measurement instrument is the eHealth Literacy Scale (eHEALS) [18-20]. The eHEALS is an 8-item self-report measure developed by Norman and Skinner [17] and published in 2006. It was initially tested among students aged 13 to 21 years [17] and has been used in a variety of studies, involving internet users and the general population. However, researchers pointed out limitations of the eHEALS, including its one-dimensionality [21,22], insufficient capture of actual eHL levels due to the risk of overestimation and underestimation [14,23], and lack of items related to shifting use cases, such as social media [14,24]. Researchers also argued that the underlying Lily model, named after its visual arrangement of 6 literacies around eHL—traditional literacy, information literacy, media literacy, computer literacy, scientific literacy, and health literacy [13,17]—does not cover important aspects of eHL, such as the ability to formulate health questions; know-how in using digital technologies; and contextual, cultural, and communicative competencies [15] or individual, situational, and environmental factors [25]. Another aspect missing in the Lily model is mentioned in the eHealth Literacy Framework (eHLF) by Norgaard et al ([26], p. 533): “safety and control,” particularly related to personal health data. In light of the conceptual discourse and shifting use cases, further eHL measurement instruments have been developed. Prominent examples are the multidimensional eHealth Literacy Questionnaire (eHLQ) [27], which is based on the eHLF [26], and the Digital Health Literacy Instrument (DHLI), which covers skills related to social media [14]. Similar to the eHEALS [17], these instruments provide eHL scores based on self-reports [14,27]. However, the DHLI is at least supplemented by a set of performance-based items [14]. A review indicates that performance-based approaches are less common in empirical research [18]. In March 2025, Norman, Skinner, and colleagues [28] published an updated definition of eHL, which takes “effectiveness, safety and helpfulness” into account and is as follows: “the ability to engage with digital technologies in effective, safe, and helpful ways to achieve health goals” ([28], p. 1). They also updated the Lily model, which now considers contextual factors surrounding eHL, and developed a revised 10-item version of the eHEALS based on the new Lily model, which has not yet been published [28].

Overall, the outlined historical development shows that the definition and operationalization of eHL has changed since the initial definition by Norman and Skinner [13] was published in 2006. Measuring eHL in a fast-paced digital environment is a challenging task. Researchers and health care providers not only need suitable eHL measurement instruments but also need to select appropriate data collection methods and modes, especially if they aim to assess the eHL of vulnerable subgroups. For example, the use of online surveys can lead to an underrepresentation of non-internet users [29]. To provide guidance for researchers, practitioners, and policymakers, there is a need for an up-to-date overview of current eHL measurement instruments, data collection methods and modes, and associated limitations. The latter is particularly important, as existing reviews in the field of eHL measurement [eg, 18-20] did not systematically analyze eHL measurement–related limitations reported in the included papers.

Objectives

This paper has two central objectives: (1) to provide an overview of eHL measurement instruments as well as data collection methods and modes recently used in empirical research in the European Economic Area, Switzerland, and the United Kingdom (ie, in countries that are well comparable in terms of health care system standards, culture, socioeconomic aspects, and data protection) and (2) to examine the limitations of the eHL measurement instruments and data collection approaches reported in the included papers. In addition, it aims to analyze the overall publication trend, eHL assessment frequency per

XSL•FO
RenderX

participant, countries of data collection, geographical distribution of eHL measurement instruments, data collection languages, target groups, and sample sizes.

The following research question served as the basis for this paper: how has eHL been assessed in the European Economic Area, Switzerland, and the United Kingdom in recent years, and which methodological limitations need to be considered?

Methods

Given the objective of this paper, to provide a broad overview of empirical papers in the field of eHL measurement and associated limitations, a scoping review was conducted. The scoping review is reported according to the PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses Extension for Scoping Reviews) (Checklist 1)[30].

Protocol and Registration

The study protocol was published on PROSPERO on February 07, 2025 (CRD42025642890). It was updated on November 06, 2025, as the scope and methods—including the eligibility criteria to reduce the number of relevant records—were refined, and again on March 05, 2026, following the updated searches conducted on January 28, 2026. As can be seen in the protocol, the study was initially planned as a rapid review but evolved into a scoping review, which maps existing eHL measurement approaches and associated limitations.

Eligibility Criteria

Multimedia Appendix 1 provides an overview of the inclusion and exclusion criteria developed by TSB and JVW in consultation with JPE and JN. Papers that described empirical studies and were published in peer-reviewed academic journals were included. Papers had to be published in 2020 or later for two reasons: (1) rapid technological advances and associated new use cases in the context of health, for example, in the field of GenAI [2] and (2) regulatory changes. For example, Germany introduced the Digital Care Act (Digitale-Versorgung-Gesetz [DVG]) in 2019, which has regulated and enabled the reimbursement of prescribable digital health applications (Digitale Gesundheitsanwendungen [DiGA]) [31]. Regulatory measures regarding the regulation and reimbursement of digital health interventions have been taken in other European countries as well [32,33]. In addition, papers were included only if abstracts and full texts were available in either English or German. Papers had to describe studies with a clear geographic focus on the European Economic Area, Switzerland, and the United Kingdom. These countries are well comparable in terms of culture and socioeconomic aspects and have developed health care systems. Furthermore, data protection is strictly regulated: in the European Economic Area by the General Data Protection Regulation (GDPR) [34], in Switzerland by the Federal Act on Data Protection (FADP) [35], and in the United Kingdom by the UK GDPR and the Data Protection Act 2018 [36]. Papers had to report the use of all items or subscales of an original eHL measurement instrument intended for calculating eHL scores.

Papers were excluded if they were reviews. Study protocols, conference papers, gray literature, and book chapters were also excluded. Moreover, cross-country studies without a clear focus on the target regions were excluded. In addition, papers were excluded if they reported the use of a measurement instrument not capturing the skills described in Norman and Skinner's [13] 2006 definition of eHL, the use of a measurement instrument solely for eHL subdimensions(eg, traditional literacy, health literacy, information literacy, scientific literacy, media literacy, and computer literacy [13]), or the use of a context-specific eHL measurement instrument (eg, COVID-19 or health conditions). Furthermore, studies that had a clear focus on adolescents or children but not on adults were excluded. Studies involving individuals with a professional perspective (eg, health professionals, IT specialists, teachers, and trainers, as well as students in health, IT, education, or communication study programs) were also excluded given the focus of this paper on eHL measurement instruments that are suitable for the general population, patients, parents, or informal caregivers.

It should be noted that it was not initially planned to exclude papers without a specific geographic focus on the European Economic Area, Switzerland, and the United Kingdom. However, after the title-abstract screening of the records from the initial searches, many records from a variety of countries with differences in culture, socioeconomic conditions, data protection regulation, and health care quality remained. To improve the comparability and facilitate the synthesis, the geographic focus eligibility criterion was added. The search strategy itself would not have changed if this criterion had been defined from the beginning.

Search Process

The reporting of the search process adheres to the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) checklist [37].

Information Sources

To identify a wide range of potentially relevant records from different subject areas, the databases PubMed and CINAHL were separately searched (ie, not simultaneously on a single platform). Supplementary searches were conducted via Google Scholar. Study registries, online or print sources, and cited references or citing references were not searched, browsed, or examined. No additional studies were retrieved through contacts or others, and no additional information sources or search methods were used.

Search Strategy

All authors collaboratively developed the search strategy. Broad search strings in English and German were used to identify a wide range of possibly relevant records via CINAHL and PubMed. Shorter search strings were used for the supplementary search via Google Scholar, as the search engine does not process long search queries [38]. The search strings were developed on the basis of the population, concept, and context (PCC) framework [39]. As the aim was to identify empirical papers that described studies involving individuals without a professional perspective (population), no subgroup-specific terms were included. Instead, the search strings contained terms and available MeSH terms in CINAHL and PubMed related to measurement instruments (concept) and eHL (context). The

XSL•FO
RenderX

initial searches were conducted on January 31, 2025. In these initial searches, records published since 2020 were filtered. The searches were updated on January 28, 2026, to ensure that the review was up to date. The searches were identical, with the exception of the filters for the publication period, which were set to the period starting January 31, 2025 (PubMed), and January 2025 (CINAHL and Google Scholar), to avoid overlap with records that had already been identified in the initial searches. During both the initial and updated searches, filters were applied in CINAHL and PubMed to identify only English and German records with an available abstract. The "[a]pply equivalent subjects" expander was used in CINAHL to identify additional relevant records. Due to the large number of results retrieved in Google Scholar, the "[s]orted by relevance" function was applied, and the first 100 results retrieved via the English search string and the first 100 retrieved via the German search string were included at both search dates (ie, 400 records in total). Multimedia Appendix 2 presents the full search strategies for the CINAHL, PubMed, and Google Scholar (ie, search strings, filters, and expanders). The search strategies were not peer-reviewed.

Selection of Sources of Evidence

The screening of the records identified via the initial searches and updated searches was carried out separately at 2 different points in time. The method was identical except for one difference. The geographical focus eligibility criterion was applied only during the full-text screening of the records identified via the initial searches and not during the title-abstract screening. During the screening of the records identified via the updated searches, this criterion was applied from the beginning.

In both cases, EndNote was used for automatic deduplication. Further duplicates were later removed manually. ASReview, an open-source machine learning tool, which ranks records based on the estimated relevance, was seen as a solution to handle the high number of records identified [40]. Van de Schoot et al ([40], p. 125) showed that ASReview "can yield far more efficient reviewing than manual reviewing while providing high quality." The records were divided by language because ASReview may have difficulty processing different languages: (1) English titles and abstracts and (2) German titles or abstracts. As most records were in English, only records with English titles and abstracts were screened in ASReview, while the smaller number of records with German titles or abstracts were screened manually.

To train the ASReview model, an initial training set containing relevant and irrelevant records was required [40]. JVW imported all English records into Rayyan, a digital tool for collaborative screening [41], created a 5% sample using the sampling function, screened the titles and abstracts and identified relevant records: 33 from the sample of the initial searches, and 2 from the sample of the updated searches. To reduce potential bias in the training set, the same number of irrelevant records was randomly selected. TSB and JN reviewed and confirmed both sets of relevant and irrelevant records. The English records were imported into ASReview, and the training set was used to train the ASReview model. ASReview default settings were used (ie, feature extraction technique: term frequency-inverse document frequency (TF-IDF), classifier: naïve Bayes, query strategy: maximum, and balance strategy: dynamic resampling/double). JVW conducted the title-abstract screening in ASReview. As recommended, the records were screened until a predefined stopping criterion was reached [40]. In line with previous review processes [42-44], the ASReview stopping criterion was set at 100 consecutive irrelevant records. The remaining English records were excluded.

Records with German titles or abstracts identified via both searches were manually screened in Rayyan. In both cases, a 25% sample was selected using the sampling function. TSB and JVW independently screened the records in both samples. Conflicts were resolved through discussion between the two authors. JVW screened the remaining records.

English and German full-text reports were retrieved and imported into Rayyan. In both cases, a 25% sample was created using the sampling function; TSB and JVW independently screened these records. Most conflicts were resolved through discussion between both authors. Some conflicts were discussed in the whole research team. JVW screened the remaining full-text reports and consulted TSB in cases of uncertainty to ensure a consistent and accurate selection.

Data Charting Process

All authors developed a standardized extraction form, tested it in advance, and modified it on the basis of a sample of 5 included papers [39] and reviewer feedback. JVW extracted data from the included papers using MAXQDA (version 24; VERBI Software Consult Sozialforschung GmbH) to manually code relevant text passages according to predefined categories. TSB then verified the correctness of all extracted data.

Data Items

The final extraction form contained the following deductive categories: (1) author and publication year, (2) title, (3) full-text language, (4) main objective, (5) objective to validate an eHL measurement instrument, (6) data collection period, (7) eHL measurement frequency per participant, (8) (target) age groups, (9) specific participants' perspectives, (10) country focus, (11) data collection language, (12) sample size in the central eHL measurement part(s), (13) eHL measurement instrument, (14) data collection method and mode, and (15) reported limitations regarding the eHL measurement instrument and data collection method and modes. Additional subcategories were developed inductively based on the material. Information that was not explicitly reported in the included papers was inferred from the context where reasonable, as indicated in Multimedia Appendix 3.

Synthesis of Results

The synthesis focused on the following eight themes and questions:

1. Publication trend: how did the number of publications develop?
2. eHL measurement instruments: which eHL measurement instruments were used, and which versions were validated in the included papers?

XSL•FO
RenderX

3. Data collection methods and modes: which data collection methods and modes were used?
4. Reported limitations: which limitations regarding eHL measurement instruments and data collection methods and modes were reported?
5. eHL measurement frequency per participant: was eHL measured once or at multiple time points?
6. Countries and languages: in which countries and languages were data collected, how often were the eHL measurement instruments used per country, and in which languages were data collected?
7. Target groups: what were the target groups in terms of health-related perspective and age?
8. Sample sizes: how large were the samples in the central eHL measurement part(s)?

As the primary objective of this paper was to analyze how eHL was measured, the role of eHL as a variable in each included paper was not analyzed. Because the reporting was sometimes insufficiently transparent and traceable, it was not always possible to determine whether the included papers analyzed the same datasets as other included papers. Therefore, to ensure a consistent and objective approach, each included paper was considered in the synthesis, regardless of whether analyses were based on the same datasets.

Results

Selection of Sources of Evidence

Through database searches, 13,657 records were identified (CINAHL: n=3525; PubMed: n=10,132). Including the 400 records from Google Scholar, the total number of records was 14,057. Of these, 10,256 records were identified during the initial searches (January 31, 2025) and 3801 records during the updated searches (January 28, 2026). Of the records from the initial searches, 1944 duplicates were automatically removed using EndNote, resulting in 8312 unique records. Of the records from the updated searches, 509 duplicates were removed, resulting in 3292 unique records. Therefore, the total number of records to be potentially screened was 11,604. While all unique German records from the initial searches (113/113, 100%) and the updated searches (100/100, 100%) were manually screened in Rayyan, 2523 of the 8199 unique English records from the initial searches (30.77%) and 1115 of the 3192 unique English records from the updated searches (34.93%) were screened in ASReview until the predefined stopping criterion was reached. In total, 3076 irrelevant records and 7753 that remained after the ASReview stopping criterion was reached were excluded. The full texts of the remaining 775 records were retrieved and assessed for eligibility. For various reasons, 643 full-text reports were excluded. A total of 132 papers were included in the final analysis. Figure 1 shows the combined results of both selection points in time (initial and updated searches).

Some papers might appear to be eligible for inclusion but were excluded. For instance, some papers [eg, 45-48] reported only the use of selected subscales or items. As inclusion in the final analysis of this study required the reported use of all items or subscales of a version of an eHL measurement instrument intended for calculating eHL scores, these papers were excluded. Other examples of excluded papers include those that reported the use of context-specific versions of eHL measurement instruments (eg, related to COVID-19 [eg, 49-54] or chronic pain [55]).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) flow diagram of the paper selection showing the combined results of 2 selection time points (initial and updated searches) (adapted from Page et al [56], which is published under Creative Commons Attribution 4.0 International License [57]).

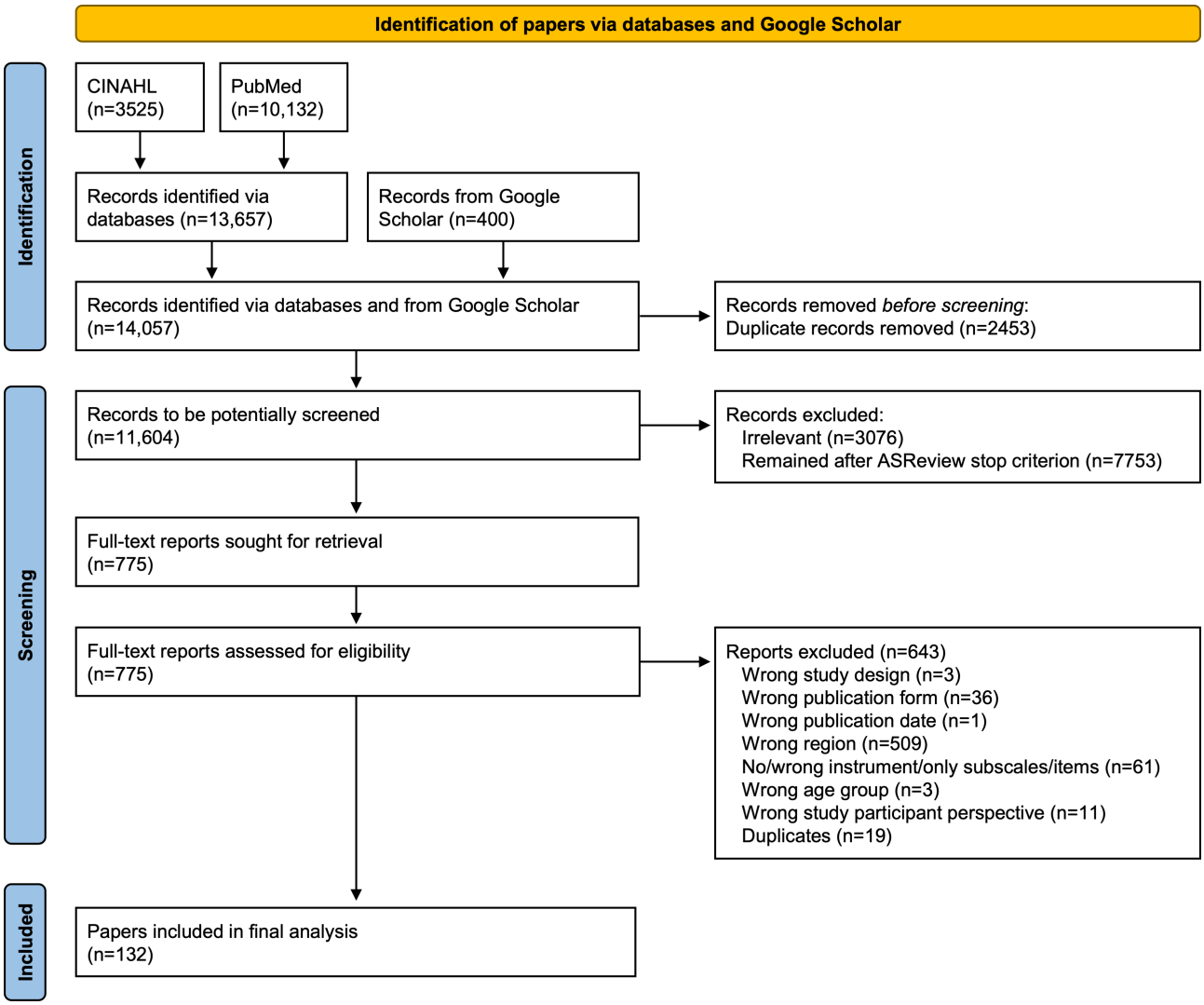

Characteristics and Results of Sources of Evidence

Multimedia Appendix 3 presents the metadata extracted from the included papers [22,58-188].

Synthesis of Results

Publication Trend

The included papers were published between 2020 and 2025, with the average number of publications per year being 22 (SD 8.72). As shown in Figure 2, the number of publications per year showed an overall upward trend. Data collection took place between 2016 and 2025.

XSL•FO
RenderX

Figure 2. Distribution of the included papers by publication year (N=132).

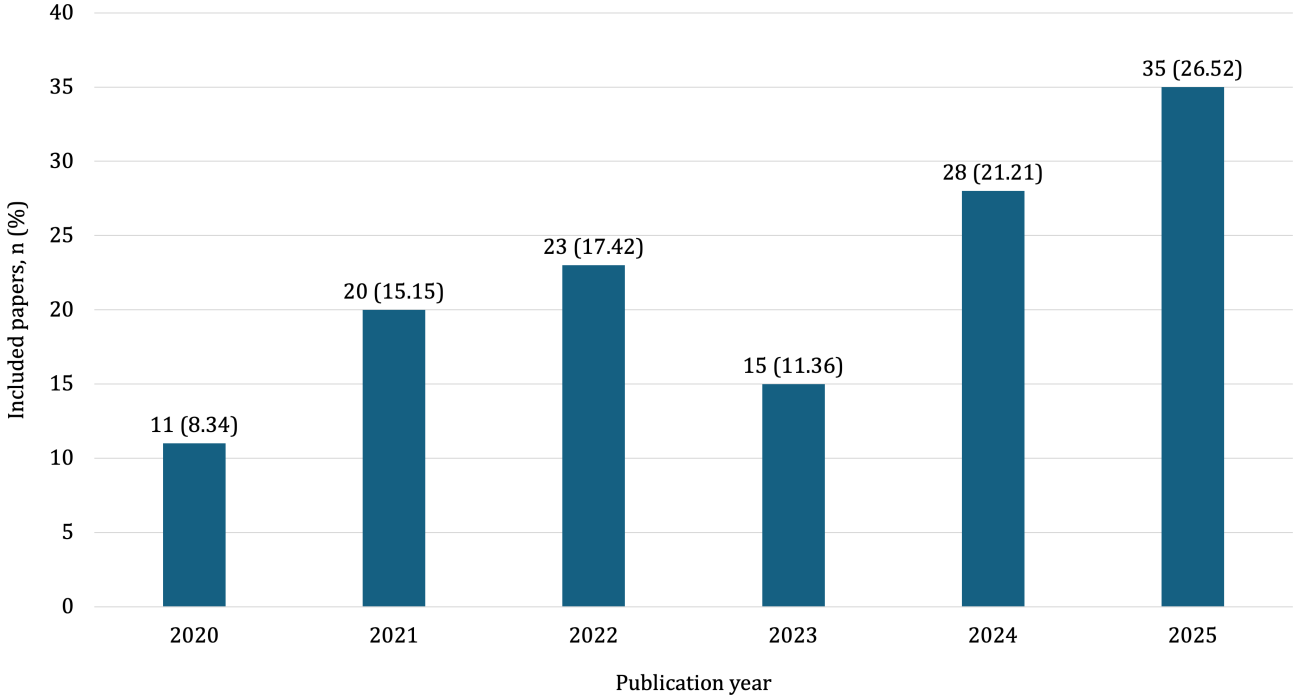

eHealth Literacy Measurement Approaches and Associated Limitations

Measurement Instruments

The included papers reported the use of all items or subscales intended for calculating eHL scores of versions of a total of 9 measurement instruments in various languages:

- the eHealth Literacy Scale (eHEALS),
- the eHealth Literacy Questionnaire (eHLQ),
- the digital health literacy module of the Health Literacy Survey 2019 - 2021 (HLS_{19}-DIGI),
- the Readiness and Enablement Index for Health Technology (READHY),
- the Digital Health Literacy Instrument (DHLI),
- the eHealth Literacy and Use Scale (eHLUS),
- the revised eHealth Literacy Scale-Extended (eHEALS-E),
- the eHealth Literacy Scale for Carers of People with Chronic Diseases (eHEALS-Carer), and
- the Transactional eHealth Literacy Instrument (TeHLI)

The eHEALS was by far the most common eHL measurement instrument (94/132, 71.21%). The eHLQ was a distant second (20/132, 15.15%), followed by the HLS_{19}-DIGI (9/132, 6.82%) and the READHY (7/132, 5.30%). Of the included papers, 2 each reported the use of the DHLI (2/132, 1.52%), the eHLUS (2/132, 1.52%), and the revised eHEALS-E (2/132, 1.52%). One paper each reported the use of the eHEALS-Carer (1/132, 0.76%) and the TeHLI (1/132, 0.76%). Six of the 132 included papers(4.54%) reported the use of all items or subscales intended for calculating eHL scores of 2 of the identified eHL measurement instruments: the eHEALS and the eHLUS (2/132, 1.52%) [91,166], the eHEALS and the eHLQ (2/132, 1.52%) [97,98], the eHEALS and the revised eHEALS-E (1/132, 0.76%) [86], and the eHEALS and the TeHLI (1/132, 0.76%) [160]. Table 1 lists the original eHL measurement instruments, including the initial developer and the year of development and the underlying concept or instrument.

XSL•FO
RenderX

Table . eHealth literacy measurement instruments identified in the included papers (N=132).[a]

Instrument	Initial developer and year	Underlying concept or instrument	Papers, n (%)	Citations
eHEALS[b]	Norman and Skinner, 2006 [17]	Lily model [13]	94 (71.21)	[illegible]
eHLQ[c]	Kayser et al, 2018 [27]	eHLF[d] [26]	20 (15.15)	[illegible]
HLS_{19}-DIGI[e]	HLS_{19} Consortium of the WHO Action Network M-POHL, 2022 [189]	DHLI[f] [14], eHEALS [17]	9 (6.82)	[69,108,117,136,150-152,171,177]
READHY[g]	Kayser et al, 2019 [190]	eHLQ [27], heiQ[h] [191], HLQ[i] [192]	7 (5.30)	[64,144,148,161,169,170,175]
DHLI	Van der Vaart et al, 2017 [14]	Skills identified in a qualitative bottom-up study [193]	2 (1.52)	[154,180]
eHLUS[j]	Stephan et al, 2025 [166]	Theoretical framework developed based on literature search, German eHEALS [194]	2 (1.52)	[91,166]
Revised eHEALS-E[k]	Petrič and Atanasova, 2024 [22]	Initial eHEALS-E [21], eHEALS [17], Norman and Skinner's [13] definition of eHL	2 (1.52)	[22,86]
eHEALS-Carer[l]	Efthymiou et al, 2019 [195]	eHEALS [17]	1 (0.76)	[85]
TeHLI[m]	Paige et al, 2019 [196]	TMeHL[n] [197]	1 (0.76)	[160]

[a]Six included papers (6/132, 4.54%) appear twice in the table, as they reported the use of all items or subscales intended for calculating scores of versions of 2 eHL measurement instruments [86,91,97,98,160,166].

[b]eHEALS: eHealth Literacy Scale.

[c]eHLQ: eHealth Literacy Questionnaire.

[d]eHLF: eHealth Literacy Framework.

[e]HLS_{19}-DIGI: digital health literacy module of the Health Literacy Survey 2019 - 2021.

[f]DHLI: Digital Health Literacy Instrument.

[g]READHY: Readiness and Enablement Index for Health Technology.

[h]heiQ: Health Education Impact Questionnaire.

[i]HLQ: Health Literacy Questionnaire.

[j]eHLUS: eHealth Literacy and Use Scale.

[k]eHEALS-E: eHealth Literacy Scale-Extended.

[l]eHEALS-Carer: eHealth Literacy Scale for Carers of People with Chronic Diseases.

[m]TeHLI: Transactional eHealth Literacy Instrument.

[n]TMeHL: Transactional Model of eHealth Literacy.

Multimedia Appendix 4 summarizes the characteristics of the identified original eHL measurement instruments. Notably, all identified measurement instruments provide eHL scores based on self-reports. The DHLI is the only identified eHL measurement instrument, which contains supplementary performance-based items [14]. Both included papers reporting the use of the DHLI (2/132, 1.52%) did not report the use of these items [154,180].

Of the 132 included papers, 23 (17.42%) reported the objective to validate an eHL measurement instrument. These papers were grouped into two clusters:

1. Two of the 132 included papers (1.52%) reported the objective to validate a new eHL measurement instrument: the eHLUS (1/132, 0.76%) [166] and the revised eHEALS-E (1/132, 0.76%) [22]. Although an earlier version of the eHEALS-E was published in 2017 [21], the revised version validated in the included paper [22] is considered a new eHL measurement instrument in this scoping review because it contains different items.
2. Twenty-one of the 132 included papers (15.91%) reported the objective to validate a version of at least 1 already existing eHL measurement. These were versions of the eHEALS (15/132, 11.36%) [62,71-73,75,86,92,93,104,122-124,131,181,182], the eHLQ (4/132, 3.03%) [96,97,142,158], the revised eHEALS-E (1/132, 0.76%) [86], and the TeHLI (1/132, 0.76%) [160]. One included paper (1/132, 0.76%) reported the objective to validate such a version of 2 of the identified instruments: the eHEALS and the revised eHEALS-E [86]. In addition,

XSL•FO
RenderX

1 included paper (1/132, 0.76%) aimed to validate the HLS_{19}-DIGI in various languages across 13 countries [117].

Multimedia Appendix 5 provides an overview of characteristics of the validation papers and the eHL measurement instruments.

Some translated or adapted versions used or validated in the included papers differ from the original instruments in key aspects, such as dimensionality (eg, versions of the eHEALS showed a 2-factor [62,73,75,86,92,123,194] or 3-factor structure [71] unlike the original eHEALS [17]) and number of items (eg, the Greek version of the revised eHEALS-E [86] comprises 30 items from the initial item pool of the revised eHEALS-E [22], and the Greek version of the eHEALS [86] showed a better fit with 6 items). Further examples of deviations include the reported use of the eHEALS with a different response scale—for example, the reported use of a "7-point Likert scale" ([139], p. 9) instead of the original 5-point Likert scale [17]—and 2 supplementary eHEALS items for calculating eHL scores [105,134].

Data Collection Methods and Modes

All included papers (132/132, 100%) reported that data on eHL were collected via surveys. As shown in Table 2, most included papers (71/132, 53.79%) reported or implied that surveys were administered digitally (eg, online or on a tablet). Twenty-three of the 132 included papers (17.42%) reported or implied that 2 or 3 data collection modes were used. One of these (1/132, 0.76%) described a cross-country study in which different modes (1 or 2) were used depending on the country [117].

Table . Data collection modes identified in the included papers (N=132).[a]

Mode	Papers, n (%)	Citations
Digital	71 (53.79)	[illegible]
Paper-based	44 (33.33)	[illegible]
Face-to-face	16 (12.12)	[22,64,88-90,98,117,122,128,144,147,150-152,170,177]
Telephone-based	9 (6.82)	[69,76-78,117,118,137,144,162]
Not reported	18 (13.64)	[58,68,74,94,110,145,148,157,163,165,168,169,174,176,179-182]

[a]Of the 132 included papers, 23 (17.42%) are listed more than once as they reported or implied that 2 or 3 data collection modes were used [22,63,64,88-90,98,104,117,118,126,127,132,133,144,161,162,167,170,172,178,183,184].

Reported Limitations

Clusters of Reported Limitations

By reviewing the *Limitations* sections or, if these were not available, the *Discussion* sections, it was identified that 65 of the 132 included papers (49.24%) reported at least 1 limitation related to the applied eHL measurement instruments and eHL-related data collection methods and modes [illegible]. Four clusters were identified: (1) self-report biases, (2) selection biases due to the data collection modes, (3) instrument-specific limitations, and (4) other limitations associated with the data collection methods and modes.

Self-Report Biases

Notably, 37 of the 132 included papers (28.03%) reported the risk of self-report biases associated with the use of self-report instruments [illegible]. Of these, some explicitly mentioned particular forms or causes, including overestimation or underestimation of actual skills (8/132, 6.06%) [73,77,111,157,159,171,172,181], inaccurate or socially desirable responses (6/132, 4.54%) [22,77,111,131,134,171], recall difficulties (3/132, 2.27%) [77,135,137], comprehension issues (1/132, 0.76%) [75], and perspective shifts (1/132, 0.76%) [104].

Selection Biases Due to the Data Collection Modes

Twenty-six of the 132 included papers (19.70%) reported potential selection biases due to the data collection modes [62,66,67,79,81,82,92,97,98,109,115-117,120,123,124,129,130,143,146,156,167,173,183,186,187]. Most of these (22/132, 16.67%) noted that data collection via online surveys potentially led to an overrepresentation or underrepresentation of subgroups with specific characteristics and provided specific examples of potentially underrepresented subgroups [62,79,81,82,92,97,98,109,115-117,120,123,124,129,130,143,156,167,173,186,187], including low digital literacy, less digital experience, or lack of access to hardware or the internet (19/132, 14.39%) [62,79,81,82,92,97,98,109,115,116,120,123,124,129,130,143,167,173,187]; older age (6/132, 4.54%) [79,82,109,130,173,187]; low education or functional literacy (4/132, 3.03%) [130,156,173,186]; low eHL (2/132, 1.52%) [117,167]; low income or wealth (2/132, 1.52%) [186,187]; and residence in rural areas (1/132, 0.76%) [186] (if characteristics were reported only for potentially overrepresented subgroups, these were inverted). Two of the 132 included papers (1.52%) addressed the possibility of selection biases associated with the use of self-administered, paper-based questionnaires, mentioning the exclusion of individuals with a lower level of functional literacy [66,67]. One of the 132 included papers (0.76%) noted the possible overrepresentation or underrepresentation of individuals with low eHL "even though we used paper questionnaires" ([146], p. 682). One of the 132 included papers (0.76%) reported unspecified selection biases because of fewer responses to paper-based questionnaires than to online questionnaires [183].

Instrument-Specific Limitations

Eighteen of the 132 included papers (13.64%) reported at least 1 limitation specific to an eHL measurement instrument; however, limitations regarding the validation process of translated and adapted instruments were not analyzed in this

XSL•FO
RenderX

scoping review. Eleven of the 132 papers (8.33%) reported eHEALS-specific limitations, including limited comprehensiveness in times of evolving digital use cases (8/132, 6.06%) [73,86,87,120,122,125,143,167] in areas such as "Health 2.0" ([73], p. 8), "Web 2.0" ([122], p. 15), "Web 2.0 and 3.0" ([86], p. 14; [87], p. 9), and "online communities, social media, and AI-based health applications" ([87], p. 9); insufficient coverage of skills related to evaluation of digital health information (1/132, 0.76%) [107]; insufficient explanation of the connection between the eHEALS and the underlying Lily model (1/132, 0.76%) [75]; lack of clarity whether "insufficient health literacy, digital literacy, or a combination hereof" leads to low eHL (1/132, 0.76%) ([75], p. 246); insufficient distinction between individuals with different levels of digital literacy (1/132, 0.76%) [75]; lack of "uniformly used cutoff scores reported in the literature" (1/132, 0.76%) ([128], p. 11); the potential "inability to capture true differences between participants achieving the highest possible score" due to a ceiling effect found in the study (1/132, 0.76%) ([167], p. 9); and the difficulty for participants to perceive "differences between the individual items" ([128], p. 11). Three of the 132 included papers (2.27%) reported eHLQ-specific limitations: the complexity in analyses due to the multidimensional structure (2/132, 1.52%) [163,173], "increasing the risk of serendipitous results" (1/132, 0.76%) ([163], p. 10), and the less extensive psychometric testing compared to the eHEALS (1/132, 0.76%) [146]. Two of the included papers (1.52%) mentioned *READHY*-specific limitations: one of these (1/132, 0.76%), which was published in 2023 and reported the use of an English version of the READHY, addressed the overall complexity and length, potentially leading to selection bias and limiting its practicality [144], whereas the other (1/132, 0.76%), which was published in 2020, mentioned a lack of validated translations into languages other than Danish at the time of the survey [170]. Two of the included papers (2/132, 1.52%) that reported the objective to validate a version of the revised eHEALS-E mentioned the insufficient distinction between 2 dimensions (ie, awareness of sources and recognition of quality and meaning) [22,86]. Further criticism included the following: potential presence of obsolete items and overall length, which could impair the practicality (1/132, 0.76%) [22]; reliance of the items in one dimension ("Being smart on the Net") on the technological progress, potentially requiring ongoing updates (1/132, 0.76%) [22]; the need to integrate further items concerning "Web 3.0" (1/132, 0.76%) ([86], p. 13); lack of cognitive interviews for some Slovenian items and all English items aimed at improving the wording (1/132, 0.76%) [22]; the availability of validated versions in Slovenian and Greek only at the time of the publication of the included paper (1/132, 0.76%) [86]; and the requirement of further testing of items to avoid social desirability bias (1/132, 0.76%) [22]. One paper (1/132, 0.76%) that aimed to initially validate the eHLUS reported that it is based on eHEALS and pointed out limitations of the development and validation process: potentially limited diversity of experts who participated in the development and validation process; operational challenges during the expert interviews; and the context-specific validation, which may not be generally applicable [166].

Other Limitations Associated With the Data Collection Methods and Modes

Other limitations identified to be associated with the data collection methods and modes included smaller sample sizes due to the exclusive use of paper-based surveys (2/132, 1.52%) [158,159], less thoughtful responses due to the use of a telephone-based survey (1/132, 0.76%) [78], and unspecified "obvious inherent limitations" of surveys (1/132, 0.76%) ([180], p. 7). In addition, one included paper (1/132, 0.76%) reported the use of different modes to collect data in different countries, limiting comparability of the results between countries [117].

Other Findings

eHealth Literacy Measurement Frequency Per Participant

Most included papers (121/132, 91.67%) reported eHL results from a single measurement per participant [22,58-67,69,70,72,74-102,104-130,132-159,161,162,164,165,167-175,177-180,183-188]. Seven of the 132 included papers (5.30%) aiming to validate a version of an eHL measurement instrument used a test-retest design [71,73,131,160,166,181,182]. Three of the 132 included papers (2.27%) reported eHL results from baseline eHL measurement and 1 follow-up (1/132, 0.76%) [68] or 2 follow-ups (2/132, 1.52%) [103,176], and 1 included paper (1/132, 0.76%) reported only the results from 2 follow-ups due to a technical error [163].

Countries and Languages

Countries

The described studies were conducted across 22 countries within the target regions of this study, most frequently in Germany (39/132, 29.54%). Five included papers (5/132, 3.79%) reported multicountry data collection, with 1 paper each reporting data collection in 2 countries (1/132, 0.76%) [85], 4 countries (1/132, 0.76%) [143], 6 countries and additional unspecified countries within the European Union or the European Economic Area (1/132, 0.76%) [59], 7 countries (1/132, 0.76%) [58], and 12 countries within the target regions of this study and 1 country (Israel) located outside the target region of this study (1/132, 0.76%) [117]. Table 3 lists the countries of data collection identified in the included papers.

XSL•FO
RenderX

Table . Countries of data collection identified in the included papers (N=132).[a]

Country	Papers, n (%)	Citations
Germany	39 (29.54)	[illegible]
Poland	15 (11.36)	[72,78-84,107,134,135,160,165,187,188]
Sweden	15 (11.36)	[65-67,93,94,112,143,157-159,167,168,176,181,182]
Norway	14 (10.61)	[61,64,71,75,96,103,117,118,126,132,133,145,146,172]
Hungary	9 (6.82)	[58,59,69,101,102,117,138,185,186]
United Kingdom	9 (6.82)	[115,116,121,127,129,143,144,162,173]
Denmark	8 (6.06)	[63,117,148,161,163,169,170,179]
Italy	8 (6.06)	[68,119,139,143,149,155,171,180]
Portugal	8 (6.06)	[58-60,90,117,122,131,177]
Spain	6 (4.54)	[58,59,88,89,97,98]
Greece	4 (3.03)	[85,86,105,110]
Cyprus	3 (2.27)	[58,59,85]
Czechia	3 (2.27)	[117,140,141]
Finland	3 (2.27)	[58,95,175]
Slovenia	3 (2.27)	[22,58,59]
Austria	2 (1.52)	[113,117]
France	2 (1.52)	[73,117]
Ireland	2 (1.52)	[117,178]
Switzerland	2 (1.52)	[104,117]
Belgium	1 (0.76)	[117]
Netherlands	1 (0.76)	[142]
Slovakia	1 (0.76)	[117]

[a]Five included papers (5/132, 3.79%) are listed 2 or 3 times because they reported multicountry data collection [58,59,85,117,143], with one of these describing that participants were also recruited from Israel, which lies outside the target region of this study and is therefore not listed [117]. One of the included papers (1/132, 0.76%) reported 6 countries of data collection and data collection in additional unspecified countries within the European Union or the European Economic Area [59].

Geographical Distribution of eHealth Literacy Measurement Instruments

The use of the eHL measurement instruments varied considerably by country. Versions of the eHEALS were used across 19 different countries within the target regions of this study. In contrast, the other identified eHL measurement instruments were used much less frequently. Apart from the eHEALS, only the eHLQ, the HLS_{19}-DIGI, and the READHY were each used in more than 2 countries. The distribution of the eHL measurement instruments by country of data collection in the included papers is shown in Figure 3.

XSL•FO
RenderX

Figure 3. eHealth literacy measurement instruments by country of data collection identified in the included papers (N=132). One multicountry paper also included data from Israel, which lies outside the target region of this study and is therefore not listed [117]. Another paper (1/132, 0.76%) reported 6 countries of data collection and other unspecified countries in the European Union or the European Economic Area, which are therefore not listed [59]. DHLI, Digital Health Literacy Instrument; eHEALS, eHealth Literacy Scale; eHEALS-Carer, eHealth Literacy Scale for Carers of People with Chronic Diseases; eHEALS-E, eHealth Literacy Scale-Extended; eHLQ, eHealth Literacy Questionnaire; eHLUS, eHealth Literacy and Use Scale; HLS_{19}-DIGI, digital health literacy module of the Health Literacy Survey 2019 - 2021; READHY, Readiness and Enablement Index for Health Technology; TeHLI, Transactional eHealth Literacy Instrument.

0 31 7–15 3–6 2 1

Country	eHEALS	eHLQ	HLS_{19}-DIGI	READHY	DHLI	eHLUS	Revised eHEALS-E	eHEALS-Carer	TeHLI
Germany	31	1	6	0	1	2	0	0	0
Poland	15	0	0	0	0	0	0	0	1
Sweden	12	3	0	0	0	0	0	0	0
Norway	6	6	1	1	0	0	0	0	0
Hungary	7	0	2	0	0	0	0	0	0
United Kingdom	7	1	0	1	0	0	0	0	0
Denmark	0	3	1	4	0	0	0	0	0
Italy	6	0	1	0	1	0	0	0	0
Portugal	5	1	2	0	0	0	0	0	0
Spain	4	4	0	0	0	0	0	0	0
Greece	3	0	0	0	0	0	1	1	0
Slovenia	2	0	0	0	0	0	1	0	0
Cyprus	2	0	0	0	0	0	0	1	0
Czechia	2	0	1	0	0	0	0	0	0
Finland	2	0	0	1	0	0	0	0	0
Austria	1	0	1	0	0	0	0	0	0
France	1	0	1	0	0	0	0	0	0
Ireland	1	0	1	0	0	0	0	0	0
Switzerland	1	0	1	0	0	0	0	0	0
Belgium	0	0	1	0	0	0	0	0	0
Netherlands	0	1	0	0	0	0	0	0	0
Slovakia	0	0	1	0	0	0	0	0	0

Languages

As shown in Table 4, the described studies collected data within the target regions of this paper in a total of 22 languages, most frequently in German (41/132, 31.06%). Two of the 132 included papers (1.52%) reported data collection exclusively in a language other than the official language of the country of data collection (Sweden), namely Arabic [66,182]. Ten of the 132 included papers (7.58%) described studies that collected data in 2 or more languages, including 1 paper (1/132, 0.76%) that reported data collection in both the country's official language (Spanish) and a regional official language (Catalan) [97], 2 papers (2/132, 1.52%) that described multicountry studies in which data were collected exclusively in the official language(s) of the respective countries within the target regions of this study [117,143] (one of which also reported data collection in nonofficial languages in Israel [117], which lies outside the predefined target regions of this study), and 7 papers (7/132, 5.30%) that reported data collection not only in the official language(s) of the respective countries within the target regions of this study but also in other languages—English (4/132, 3.03%) [58,59,153,179]; Arabic (1/132, 0.76%) [65]; Russian and Turkish (1/132, 0.76%) [152]; and Arabic, English, Russian, Turkish, and Twi (1/132, 0.76%) [128].

XSL•FO
RenderX

Table . Data collection languages identified in the included papers (N=132).[a]

Language	Papers, n (%)	Citations
German	41 (31.06)	[illegible]
English	16 (12.12)	[58,59,115-117,121,127-129,143,144,153,162,173,178,179]
Polish	15 (11.36)	[72,78-84,107,134,135,160,165,187,188]
Norwegian	14 (10.61)	[61,64,71,75,96,103,117,118,126,132,133,145,146,172]
Swedish	13 (9.85)	[65,67,93,94,112,143,157-159,167,168,176,181]
Hungarian	9 (6.82)	[58,59,69,101,102,117,138,185,186]
Italian	9 (6.82)	[68,117,119,139,143,149,155,171,180]
Danish	8 (6.06)	[63,117,148,161,163,169,170,179]
Portuguese	8 (6.06)	[58-60,90,117,122,131,177]
Greek	6 (4.54)	[58,59,85,86,105,110]
Spanish	6 (4.54)	[58,59,88,89,97,98]
Arabic	4 (3.03)	[65,66,128,182]
Czech	3 (2.27)	[117,140,141]
Finnish	3 (2.27)	[58,95,175]
Slovenian	3 (2.27)	[22,58,59]
Dutch	2 (1.52)	[117,142]
French	2 (1.52)	[73,117]
Russian	2 (1.52)	[128,152]
Turkish	2 (1.52)	[128,152]
Catalan	1 (0.76)	[97]
Slovak	1 (0.76)	[117]
Twi	1 (0.76)	[128]

[a]Ten included papers (10/132, 7.58%) are listed 2 or more times because they reported data collection in multiple languages [58,59,65,97,117,128,143,152,153,179]. For 1 paper describing a multicountry study that also included participants from Israel, only the data collection languages used within the predefined target regions of this study are shown [117].

Samples

Health-Related Perspectives

Three major clusters were identified regarding the health-related perspectives of the target groups:

1. Sixty-four of the 132 included papers (48.48%) reported that patients or individuals with current or previous health conditions were the target group [illegible]
2. Eleven of the 132 included papers (8.33%) stated that parents or legal guardians of children or informal caregivers were the target group [85,101,104,112,113,132,133,135,158,172,176].
3. Fifty-nine of the 132 included papers (44.70%) described studies with other target groups (eg, general population or students) [illegible]

Two of the 132 included papers (1.52%) were each assigned to 2 of these clusters. Of these, one study (1/132, 0.76%) reported that patients (assigned to cluster 1) and parents of hospitalized children (assigned to cluster 2) were the target group [158], and the other (1/132, 0.76%) reported that users of symptom checker applications (assigned to cluster 1) and nonusers of symptom checker applications (assigned to cluster 3) were the target group [184]. One of the included papers (1/132, 0.76%) reported that, in addition to patients with hypertension (assigned to cluster 1), physicians were the target group [70]; no target group cluster 4 for physicians was created, as this study did not focus on individuals with a professional perspective.

Age Groups

In terms of age groups, the included papers were grouped into three clusters:

1. Of the 132 included papers, 119 (90.15%) reported that only adults were included. Of these papers, 104 (104/132, 78.79%) stated that adults in general aged ≥18 or >18 years or broad adult age ranges were included [illegible] Furthermore, 10 of the 132 papers (7.58%) focused on older adults, namely adults aged ≥40 years (2/132, 1.52%) [102,186], >45 years (1/132, 0.76%) [63], ≥50 years (1/132, 0.76%) [68], 50 to 69 years (1/132, 0.76%) [161], ≥60 years (1/132, 0.76%) [74], and ≥65 years (4/132, 3.03%) [64,94,122,169]. Two papers (2/132, 1.52%) mentioned

XSL•FO
RenderX

that younger adults were included, namely adults aged 18 to 29 years (1/132, 0.76%) [175] and 18 to 35 years (1/132, 0.76%) [80]; 2 papers (2/132, 1.52%) reported that middle-aged adults were included, namely adults aged 30 to 65 years (1/132, 0.76%) [168] and 34 to 64 years (1/132, 0.76%) [91]; and 1 paper (1/132, 0.76%) reported that both "younger adults" and "older adults" were included, without further specification ([134], p. 1739).

2. Five of the 132 included papers (3.79%) reported that minor and adult participants were included. Of these, 1 paper (1/132, 0.76%) mentioned that individuals with a verified social media account were the target group, regardless of their age, with the actual age range being 14 to 72 years [72]; 3 papers (3/132, 2.27%) specified that individuals aged ≥14 years (1/132, 0.76%) [76] and ≥16 years (2/132, 1.52%) [109,154] were included; and 1 paper (1/132, 0.76%) stated that individuals aged <35 years were included [73].
3. Eight of the 132 included papers (6.06%) did not provide a target age, a qualitative description regarding the age, an actual age range, or age groups [60,104,132,135,140,141,176,179].

Sample Sizes

As shown in Table 5, the included papers were divided into 7 clusters on the basis of the sample sizes reported for the central eHL measurement(s), or total sample if not available, ranging from <50 (7/132, 5.30%) to >2000 participants (8/132, 6.06%).

Table . Sample sizes reported for the central eHealth literacy measurement(s) or, if unavailable, the overall study sample sizes in the included papers (N=132).

Sample size, n	Papers, n (%)	Citations
<50	7 (5.30)	[74,91,95,112,113,121,169]
51 - 100	12 (9.09)	[58,59,63,68,128,136,148,157,161,163,165,177]
101 - 250	36 (27.27)	[47,78,79,90,100,106,108,110,111,118,134,143,144,147,156,158,159,164,166,167,170,174,178,184]
251 - 500	30 (22.73)	[61,64,67,73,86,94,98,115,116,120,122,124,126,130,134,138,146,153,155,167,172,178,182]
501 - 1000	20 (15.15)	[60,65,69,76-78,93,97,104,105,129,135,137,139-141,154,183,185,186]
1001 - 2000	19 (14.39)	[22,71,72,79,80,83,84,102,107,109,119,125,127,142,145,149,151,152,160]
>2000	8 (6.06)	[81,82,117,143,150,173,187,188]

Discussion

This paper aimed to provide an overview of how eHL has been assessed in the European Economic Area, Switzerland, and the United Kingdom in recent years and which methodological limitations need to be considered.

Principal Findings

Summary of Evidence

A total of 132 empirical papers were included in the final analysis. The publication trend indicated growing research activity in the field of eHL measurement. Across the included papers, the use of versions of 9 eHL measurement instruments that provide eHL scores based on self-reports was reported, most frequently the use of the eHEALS (94/132, 71.21%). All included papers (132/132, 100%) reported the use of surveys to collect data on eHL, most of which were conducted digitally (71/132, 53.79%). Reported limitations include potential self-report biases associated with the use of self-report instruments (37/132, 28.03%), selection biases due to the data collection mode (26/132, 19.70%), and limitations specific to 5 of the 9 identified eHL measurement instruments (18/132, 13.64%). Most included papers (121/132, 91.67%) reported eHL results from a single measurement per participant. Data were collected across 22 countries within the target regions, most frequently in Germany (39/132, 29.54%). The use of the eHL measurement instruments varied considerably between the countries. In total, 22 different data collection languages were identified, with German being the most common (41/132, 31.06%). Patients or individuals with current or previous health problems were a commonly reported target group (64/132, 48.48%). Broad adult age ranges were very common (104/132, 78.79%). Sample sizes ranged from ≤50 (7/132, 5.30%) to >2000 (8/132, 6.06%) participants.

Self-Report eHealth Literacy Measurement Instruments

Self-report biases are a key limitation of self-report measurement instruments [198,199], as also noted in various papers included in this scoping review. A study, which was also cited in some included papers, found a weak relationship between the eHEALS—the most frequently used eHL measurement instrument—and performance-based task [14,23]. However, no performance-based eHL measurement instruments were identified via this scoping review. One reason may be that we did not consider instruments solely assessing only some of the skills described in the 2006 definition of eHL by Norman and Skinner [13] or focusing on specific subdimensions of eHL. Crocker et al ([18], p. 1) identified "29 unique performance-based eHealth literacy measurement tools," using less strict criteria. One of these was used in a study also included in this scoping review: Schulz et al [155] reported the use of the eHEALS alongside additional tasks. Specifically, participants rated the quality of 2 simulated websites (1 with high quality and 1 with low quality) using a 7-point semantic differential scale (eg, "accurate" vs "inaccurate") ([155], p. 4); furthermore, they were shown the right treatment options (eg, "seeking help from a doctor") ([155], p. 4]) and wrong treatment options (eg, "treating depression with St John's wort or vitamins or yoga without any mention of antidepressant medications or psychotherapy") ([155], p. 4]) and had to make a selection. While Crocker et al [18] considered this approach an eHL

XSL•FO
RenderX

measurement instrument, it does not capture the relevant aspects of the eHL definition by Norman and Skinner [13] (eg, searching health information) and was therefore not considered an eHL measurement instrument in this paper. Another instrument, identified by Crocker et al [18] and via this scoping review, is the DHLI [14]. It comprises 21 self-report items intended for calculating eHL scores and 7 supplementary performance-based items, which did not demonstrate acceptable validity in the validation study [14]. Thus, the DHLI [14] is a self-report eHL measurement instrument with supplementary performance-based items rather than a performance-based eHL measurement instrument. Nevertheless, Crocker et al [18] found that performance-based eHL measurement instruments are rarely used and highlighted that several were not completely published. They also mention other challenges of performance-based eHL measurement instruments, such as the requirement of regular updates considering new scientific evidence, varying topic relevance for different target groups, and the time demand for researchers and participants [18].

Exclusion of Subgroups in Online Surveys

Selection biases were mentioned in the included papers, particularly in the context of frequent use of online surveys. Online surveys potentially introduce self-selection biases and lead to an undercoverage of subgroups, such as individuals with limited or no access to the internet [200]. Despite growing internet adoption, 6% of households in the European Union still had no internet access in 2024 [201]. Hardware and internet access, in particular, are prerequisites for using digital tools [202]. However, even individuals who do not yet use the internet may benefit from digital health interventions if they are equipped with the necessary hardware and an internet connection. In addition, digital interventions need to reflect eHL-related needs [203]. Therefore, it would be important to ensure inclusion of subgroups who are not yet using the internet, and other potentially underrepresented subgroups, which were mentioned in the included papers. Selection biases can be addressed through methodological strategies, such as probability sampling [200]. Another approach is mixed-mode designs (ie, different modes are offered to different target groups)—which, however, introduces further (selection) biases [204]—as well as sequential mixed-mode designs (ie, different survey modes are offered to the same target group one after another) [205].

Alignment With Technological Innovations

This study found that some included papers criticized eHL measurement instruments for not fully aligning with technological developments and shifting use cases. As the internet has been getting more interactive since the eHEALS was published in 2006, skills related to social media, where users potentially encounter health misinformation [10], are considered relevant in the context of eHL [23,24]. Although social media is not a new phenomenon, only some identified eHL measurement instruments contain items explicitly related to interactive digital tools: the revised eHEALS-E [22], the TeHLI [196], and the DHLI [14]. Also, the HLS_{19}-DIGI [189], which is based on the DHLI [14], contains such items, although not in the main part intended for calculating eHL scores (HL-DIGI). eHL can also be considered essential in the context of health information provided by GenAI, as studies [11,12] show that health misinformation is a serious issue. Yet, GenAI has been established in recent years, and this may explain why it is not explicitly mentioned in older eHL measurement instruments. However, not even the revised eHEALS-E [22] and the eHLUS [166] published in the period after the launch of ChatGPT in 2022 mention GenAI tools explicitly. However, the developers in the revised eHEALS-E took GenAI into account, pointing out "that the Being smart on the Net dimension is particularly important in the context of the rapidly evolving internet technologies based on artificial intelligence algorithms that require critical awareness, distance, and understanding" ([22], p. 8). Dedicated measurement instruments were developed in recent years to measure artificial intelligence literacy in general [206,207]. Furthermore, as outlined, newer theoretical models such as the eHLF [26] and the new Lily model [28] also include the aspect "safety." However, safety in terms of data protection is only considered in the DHLI [14], the eHLF-based eHLQ [27], and the eHLUS [166]. Overall, the findings indicate the need to continuously review and update eHL measurement instruments to ensure alignment with technological innovations. Developing an eHL measurement instrument that does not become obsolete in a rapidly changing digital environment can be described as a key challenge for future research.

Complexity of eHealth Literacy Measurement Instruments

In terms of complexity or length, the eHEALS-E [22], the eHLQ [27], and the READHY [190], which contains the eHLQ [27], were criticized in the included papers. Similar to the DHLI [14], the TeHLI [196], and the eHLUS [166], these eHL measurement instruments capture a broader range of dimensions of eHL. Efthymiou et al [86], who validated the Greek versions of the eHEALS and the revised eHEALS-E, provided guidance in selecting the appropriate eHL measurement instrument. Specifically, they noted that "[t]he eHeals could be used as an additional tool when eHealth Literacy is not the core concept measured and the revised eHeals-Extended can be used when researchers wish to measure eHealth Literacy concept more thoroughly" ([86], p. 2). Following this argumentation, the use of eHL measurement instruments with a broader range of dimensions should be considered when the aim is to generate deeper insights, including specific needs of individuals. However, it is debatable whether health professionals will follow this advice. For example, Norman ([24], p. 2) noted that the eHEALS was developed as a short scale, as "health professionals […] said they would not use a long instrument in practice."

eHealth Literacy Measurement Instruments for Older Adults

In addition, we want to highlight that none of the eHL measurement instruments identified in this paper was specifically developed for older adults. Also, Wang et al [208] noted a lack of such eHL measurement instruments. The researchers recently introduced the Digital Health Literacy Questionnaire for Older Adults "designed to align with the realities of older adults' daily lives and health care experiences, focusing on potential challenges they may face when using digital technologies to access, process, communicate, and

XSL•FO
RenderX

understand health information and services“ ([208], p. 12). The instrument is not included in the final analysis of this scoping review, as it was used to assess eHL of older adults in China [208]. Older age groups differ from younger individuals in certain characteristics (eg, physical and cognitive conditions, literacy levels, technology usage, and privacy preferences), requiring special methodological considerations, for example, in terms of item complexity [209]. As older adults are particularly at risk of digital exclusion [9,210], there is a need to evaluate the appropriateness of eHL measurement instruments for them.

Limitations

The review process used is subject to some limitations. First, it needs to be noted that, due to the use of ASReview to facilitate the title and abstract screening of the English records, 30.77% (2523/8199) of the English records from the initial searches and 34.93% (1115/3192) of the English records from the updated searches were screened. The remaining English records remained after the ASReview stopping criterion (ie, 100 consecutive irrelevant records) was reached in both cases. Thus, despite the demonstrated performance of ASReview [40], there is a risk that relevant papers were not screened and thus not included in the final analysis.

Second, the exclusion of papers without a clear focus on the target regions resulted in the omission of papers and eHL measurement instruments. As German and English papers were included, there may be an overrepresentation of papers from German- and English-speaking regions and an underrepresentation of papers from non–German-speaking and non–English-speaking regions within the target regions.

Third, the analysis covered limitations reported in the *Limitations* section or, if a *Limitations* section was unavailable, the *Discussion* section of the included papers. Therefore, limitations reported elsewhere in the included papers were not analyzed. Moreover, the analysis was limited to reported limitations regarding eHL measurement instruments as well as eHL-related data collection methods and modes. Other limitations reported in the included papers (eg, those related to cross-sectional study designs or sampling or recruitment strategies and limitations specific to translated and culturally adapted versions of eHL measurement instruments) were not extracted and analyzed in this study.

Implications for Practice, Future Research, and Policy

This paper provides health care practitioners and researchers with a structured overview of limitations related to eHL measurement instruments and data collection methods and modes. The identified eHL measurement instruments differ in key aspects, including length, complexity, and alignment with rapidly evolving digital use cases. Particularly, online surveys might lead to an exclusion of subgroups. Thus, it clearly shows that selecting both a suitable eHL measurement instrument and suitable data collection methods and modes is a key decision, which should be based on factors such as the purpose of the planned eHL measurement, the target population, and the data collection setting.

The overview could encourage researchers to adopt eHL measurements that are less common or were not used in specific countries yet. This potentially requires translation, cultural adaptation, and validation efforts. Future research should also focus on revising existing eHL measurement instruments and developing new eHL measurement instruments to capture the skills required for an ever-growing range of digital tools in the context of health, including social media and GenAI. This may require new underlying operationalization approaches that manage to avoid becoming obsolete. Use case–specific eHL measurement instruments may also be relevant for health care providers and researchers who only want to assess specific aspects (eg, skills that are necessary when using a specific digital health application) considering setting and time constraints (eg, in health care facilities). The fact that only self-report eHL measurement instruments were identified and the associated risk of self-report biases reported in the included papers indicate a need for performance-based eHL measurement approaches. Furthermore, revised or new eHL measurement instruments tailored to older adults are highly relevant for gaining a better understanding of the specific eHL-related needs.

Policymakers should encourage health care providers to conduct eHL measurements using carefully selected eHL measurement instruments and data collection methods and modes to provide targeted solutions based on individual eHL-related needs.

Conclusions

To our knowledge, this is the first scoping review synthesizing eHL measurement limitations reported in empirical papers from the European Economic Area, Switzerland, and the United Kingdom. The analysis of 132 empirical papers showed that eHL was assessed using 9 eHL measurement instruments in various languages, especially the eHEALS, and that digital surveys were a common data collection approach. The identified eHL measurement–related limitations, such as self-reported eHL measurement, exclusion of subgroups, and lack of alignment of technological developments and changing use cases in the context of health, potentially lead to biased results. Therefore, selecting suitable eHL measurement approaches is a critical decision. Health care professionals and researchers should consider a variety of factors, such as measurement purpose, target population, and data collection setting. There is a need for translated and culturally adapted as well as revised or newly developed eHL measurement instruments that keep pace with technological change, cover diverse use cases or focus on specific use cases, and contain performance-based components. Policymakers should encourage health care providers to thoughtfully select appropriate eHL measurement approaches and use eHL-related insights as a basis for targeted solutions.

XSL•FO
RenderX

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: JVW, JN, JPE, TSB
Formal Analysis: JVW, TSB
Methodology: JVW, JN, JPE, TSB
Supervision: TSB, JN, JPE
Visualization: JVW
Writing – original draft: JVW
Writing – review & editing: JVW, JN, JPE, TSB

Conflicts of Interest

None declared.

Multimedia Appendix 1
Inclusion and exclusion criteria.
[DOCX File, 26 KB - jmir_v28i1e87461_app1.docx]

Multimedia Appendix 2
Search strings, filters, and expanders.
[DOCX File, 18 KB - jmir_v28i1e87461_app2.docx]

Multimedia Appendix 3
Metadata of the included papers.
[DOCX File, 180 KB - jmir_v28i1e87461_app3.docx]

Multimedia Appendix 4
Characteristics of the identified original eHealth literacy measurement instruments.
[DOCX File, 29 KB - jmir_v28i1e87461_app4.docx]

Multimedia Appendix 5
Characteristics of the eHealth literacy measurement instruments validated in the included papers.
[DOCX File, 38 KB - jmir_v28i1e87461_app5.docx]

Checklist 1
Preferred Reporting Items for Systematic reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) Checklist.
[DOCX File, 88 KB - jmir_v28i1e87461_app6.docx]

References

1. Individuals using the internet for seeking health-related information. : Eurostat; 2025 URL: https://ec.europa.eu/eurostat/databrowser/view/tin00101/default/table [accessed 2026-05-01] [doi: 10.2908/TIN00101]
2. Bautista JR, Herbert D, Farmer M, De Torres RQ, Soriano GP, Ronquillo CE. Health consumers' use and perceptions of health information from generative artificial intelligence chatbots: a scoping review. Appl Clin Inform 2025 Aug;16(4):892-902. [doi: 10.1055/a-2647-1210] [Medline: 40602776]
3. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res 2021 May 12;23(5):e17917. [doi: 10.2196/17917] [Medline: 33978589]
4. Apple Health: meaningful insights: backed by science. Apple. 2025. URL: https://www.apple.com/health/ [accessed 2025-05-03]
5. Google Fit: coaching you to a healthier and more active life. Google. 2025. URL: https://www.google.com/intl/en_us/fit/ [accessed 2025-05-03]

XSL•FO
RenderX

6. Classification of Digital Interventions, Services and Applications in Health: A Shared Language to Describe the Uses of Digital Technology for Health: World Health Organization; 2023. URL: https://www.who.int/publications/i/item/9789240081949 [accessed 2026-05-06]
7. Livieri G, Mangina E, Protopapadakis ED, Panayiotou AG. The gaps and challenges in digital health technology use as perceived by patients: a scoping review and narrative meta-synthesis. Front Digit Health 2025;7:1474956. [doi: 10.3389/fdgth.2025.1474956] [Medline: 40212901]
8. Friemel TN. The digital divide has grown old: determinants of a digital divide among seniors. New Media Soc 2016 Feb;18(2):313-331. [doi: 10.1177/1461444814538648]
9. Ge H, Li J, Hu H, Feng T, Wu X. Digital exclusion in older adults: a scoping review. Int J Nurs Stud 2025 Aug;168:105082. [doi: 10.1016/j.ijnurstu.2025.105082] [Medline: 40279791]
10. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021 Jan 20;23(1):e17187. [doi: 10.2196/17187] [Medline: 33470931]
11. Ayers JW, Zhu Z, Poliak A, et al. Evaluating artificial intelligence responses to public health questions. JAMA Netw Open 2023 Jun 1;6(6):e2317517. [doi: 10.1001/jamanetworkopen.2023.17517] [Medline: 37285160]
12. Andrikyan W, Sametinger SM, Kosfeld F, et al. Artificial intelligence-powered chatbots in search engines: a cross-sectional study on the quality and risks of drug information for patients. BMJ Qual Saf 2025 Jan 28;34(2):100-109. [doi: 10.1136/bmjqs-2024-017476] [Medline: 39353736]
13. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9. [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
14. van der Vaart R, Drossaert C. Development of the digital health literacy instrument: measuring a broad spectrum of Health 1.0 and Health 2.0 skills. J Med Internet Res 2017 Jan 24;19(1):e27. [doi: 10.2196/jmir.6709] [Medline: 28119275]
15. Gilstad H. Toward a comprehensive model of eHealth literacy. In: Proceedings of the 2nd European Workshop on Practical Aspects of Health Informatics (PAHI 2014): CEUR-WS; 2014:63-72 URL: http://ceur-ws.org [accessed 2025-04-26]
16. Griebel L, Enwald H, Gilstad H, Pohl AL, Moreland J, Sedlmayr M. eHealth literacy research: quo vadis? Inform Health Soc Care 2018 Dec;43(4):427-442. [doi: 10.1080/17538157.2017.1364247] [Medline: 29045164]
17. Norman CD, Skinner HA. eHEALS: the eHealth Literacy Scale. J Med Internet Res 2006 Nov 14;8(4):e27. [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
18. Crocker B, Feng O, Duncan LR. Performance-based measurement of eHealth literacy: systematic scoping review. J Med Internet Res 2023 Jun 2;25:e44602. [doi: 10.2196/44602] [Medline: 37266975]
19. Karnoe A, Kayser L. How is eHealth literacy measured and what do the measurements tell us? A systematic review. Knowl Manag E-Learn 2015 Dec 15;7(4):576-600. [doi: 10.34105/j.kmel.2015.07.038]
20. Wang C, Chang L, Chen X, Kong J, Qi H. eHealth literacy assessment instruments: scoping review. J Med Internet Res 2025 Aug 20;27:e66965. [doi: 10.2196/66965] [Medline: 40835422]
21. Petrič G, Atanasova S, Kamin T. Ill literates or illiterates? Investigating the eHealth literacy of users of online health communities. J Med Internet Res 2017 Oct 4;19(10):e331. [doi: 10.2196/jmir.7372] [Medline: 28978496]
22. Petrič G, Atanasova S. Validation of the extended e-health literacy scale: structural validity, construct validity and measurement invariance. BMC Public Health 2024 Jul 25;24(1):1991. [doi: 10.1186/s12889-024-19431-8] [Medline: 39054541]
23. van der Vaart R, van Deursen AJ, Drossaert CH, Taal E, van Dijk JA, van de Laar MA. Does the eHealth Literacy Scale (eHEALS) measure what it intends to measure? Validation of a Dutch version of the eHEALS in two adult populations. J Med Internet Res 2011 Nov 9;13(4):e86. [doi: 10.2196/jmir.1840] [Medline: 22071338]
24. Norman C. eHealth literacy 2.0: problems and opportunities with an evolving concept. J Med Internet Res 2011 Dec 23;13(4):e125. [doi: 10.2196/jmir.2035] [Medline: 22193243]
25. Levin-Zamir D, Bertschi I. Media health literacy, eHealth literacy, and the role of the social environment in context. Int J Environ Res Public Health 2018 Aug 3;15(8):1643. [doi: 10.3390/ijerph15081643] [Medline: 30081465]
26. Norgaard O, Furstrand D, Klokker L. The e-health literacy framework: a conceptual framework for characterizing e-health users and their interaction with e-health systems. Knowl Manag E-Learn 2015 Dec 15;7(4):522-540. [doi: 10.34105/j.kmel.2015.07.035]
27. Kayser L, Karnoe A, Furstrand D, et al. A multidimensional tool based on the eHealth literacy framework: development and initial validity testing of the eHealth Literacy Questionnaire (eHLQ). J Med Internet Res 2018 Feb 12;20(2):e36. [doi: 10.2196/jmir.8371] [Medline: 29434011]
28. Milanti A, Norman C, Chan DNS, So WKW, Skinner H. eHealth literacy 3.0: updating the Norman and Skinner 2006 model. J Med Internet Res 2025 Mar 11;27:e70112. [doi: 10.2196/70112] [Medline: 40068166]
29. Hsia J, Zhao G, Town M. Estimating undercoverage bias of internet users. Prev Chronic Dis 2020 Sep 10;17:E104. [doi: 10.5888/pcd17.200026] [Medline: 32915129]
30. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]

XSL•FO
RenderX

31. Gesetz für eine bessere Versorgung durch Digitalisierung und Innovation [Article in German]. Bundesgesetzblatt. 2019 Dec 19. URL: https://www.bgbl.de/xaver/bgbl/start.xav?startbk=Bundesanzeiger_BGBl&jumpTo=bgbl119s2562.pdf#/text/bgbl119s2562.pdf?_ts=1758500669727 [accessed 2025-09-22]
32. van Kessel R, Srivastava D, Kyriopoulos I, et al. Digital health reimbursement strategies of 8 European countries and Israel: scoping review and policy mapping. JMIR Mhealth Uhealth 2023 Sep 29;11:e49003. [doi: 10.2196/49003] [Medline: 37773610]
33. Tarricone R, Petracca F, Weller HM. Towards harmonizing assessment and reimbursement of digital medical devices in the EU through mutual learning. NPJ Digit Med 2024 Oct 1;7(1):268. [doi: 10.1038/s41746-024-01263-w] [Medline: 39354125]
34. Legal framework of EU data protection. European Commission. URL: https://commission.europa.eu/law/law-topic/data-protection/legal-framework-eu-data-protection_en [accessed 2026-02-18]
35. Federal Act on Data Protection (Data Protection Act, FADP). Fedlex: the publication platform for federal law. 2020. URL: https://www.fedlex.admin.ch/eli/cc/2022/491/en [accessed 2026-02-19]
36. Data protection. GOV.UK. 2026. URL: https://www.gov.uk/data-protection [accessed 2026-05-01]
37. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
38. Gusenbauer M, Haddaway NR. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. Res Synth Methods 2020 Mar;11(2):181-217. [doi: 10.1002/jrsm.1378] [Medline: 31614060]
39. Pollock D, Peters MDJ, Khalil H, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. JBI Evid Synth 2023 Mar 1;21(3):520-532. [doi: 10.11124/JBIES-22-00123] [Medline: 36081365]
40. van de Schoot R, de Bruin J, Schram R, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell 2021;3(2):125-133. [doi: 10.1038/s42256-020-00287-7]
41. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan: a web and mobile app for systematic reviews. Syst Rev 2016 Dec 5;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
42. Adamse I, Eichelsheim V, Blokland A, Schoonmade L. The risk and protective factors for entering organized crime groups and their association with different entering mechanisms: a systematic review using ASReview. Eur J Criminol 2024 Nov;21(6):859-886. [doi: 10.1177/14773708241250278]
43. van den Berg RL, van der Landen SM, Keijzer MJ, et al. Smartphone- and tablet-based tools to assess cognition in individuals with preclinical Alzheimer disease and mild cognitive impairment: scoping review. J Med Internet Res 2025 May 27;27:e65297. [doi: 10.2196/65297] [Medline: 40424609]
44. van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, van der Palen J, Doggen CJM, Lenferink A. Artificial intelligence in systematic reviews: promising when appropriately used. BMJ Open 2023 Jul 7;13(7):e072254. [doi: 10.1136/bmjopen-2023-072254] [Medline: 37419641]
45. Álvarez-Pérez Y, Perestelo-Pérez L, Rivero-Santanta A, et al. Co-creation of massive open online courses to improve digital health literacy in pregnant and lactating women. Int J Environ Res Public Health 2022 Jan 14;19(2):913. [doi: 10.3390/ijerph19020913] [Medline: 35055733]
46. Martens L, Brand T, Zeeb H, Graff H, Wiersing R, Muellmann S. Digitale Gesundheitsinformationen und -angebote: Eine Fokusgruppenstudie zu Nutzungsbarrieren mit digital exkludierten Bevölkerungsgruppen [Article in German]. Präv Gesundheitsf 2025 Aug 28. [doi: 10.1007/s11553-025-01266-9]
47. Schulz PJ, Rothenfluh F. Influence of health literacy on effects of patient rating websites: survey study using a hypothetical situation and fictitious doctors. J Med Internet Res 2020 Apr 6;22(4):e14134. [doi: 10.2196/14134] [Medline: 32250275]
48. Zangger G, Mortensen SR, Tang LH, Thygesen LC, Skou ST. Association between digital health literacy and physical activity levels among individuals with and without long-term health conditions: data from a cross-sectional survey of 19,231 individuals. Digit Health 2024;10:20552076241233158. [doi: 10.1177/20552076241233158] [Medline: 38410789]
49. Bak CK, Krammer J, Dadaczynski K, et al. Digital health literacy and information-seeking behavior among university college students during the COVID-19 pandemic: a cross-sectional study from Denmark. Int J Environ Res Public Health 2022 Mar 19;19(6):3676. [doi: 10.3390/ijerph19063676] [Medline: 35329363]
50. Dadaczynski K, Okan O, Messer M, et al. Digital health literacy and web-based information-seeking behaviors of university students in Germany during the COVID-19 pandemic: cross-sectional survey study. J Med Internet Res 2021 Jan 15;23(1):e24097. [doi: 10.2196/24097] [Medline: 33395396]
51. Lorini C, Velasco V, Bonaccorsi G, et al. Validation of the COVID-19 digital health literacy instrument in the Italian language: a cross-sectional study of Italian university students. Int J Environ Res Public Health 2022 May 20;19(10):6247. [doi: 10.3390/ijerph19106247] [Medline: 35627789]
52. Martins S, Augusto C, Martins MRO, et al. Adaptation and validation of the Digital Health Literacy Instrument for Portuguese university students. Health Promot J Austr 2022 Oct;33 Suppl 1:390-398. [doi: 10.1002/hpja.580] [Medline: 35124876]

XSL•FO
RenderX

53. Rosário R, Fronteira I, Martins MRO, et al. Infodemic preparedness and COVID-19: searching about public health and social measures is associated with digital health literacy in university students. Int J Environ Res Public Health 2022 Sep 28;19(19):12320. [doi: 10.3390/ijerph191912320] [Medline: 36231618]
54. Bonaccorsi G, Gallinoro V, Guida A, et al. Digital health literacy and information-seeking in the era of COVID-19: gender differences emerged from a Florentine University experience. Int J Environ Res Public Health 2023 Jan 31;20(3):2611. [doi: 10.3390/ijerph20032611] [Medline: 36767976]
55. Castarlenas E, Sánchez-Rodríguez E, Roy R, et al. Electronic health literacy in individuals with chronic pain and its association with psychological function. Int J Environ Res Public Health 2021 Nov 28;18(23):12528. [doi: 10.3390/ijerph182312528] [Medline: 34886254]
56. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
57. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2026-04-29]
58. Ahmed F, Romero Saletti S, D'Souza E, et al. Assessing the user experience of the EU mobile app for cancer prevention: mixed methods study. JMIR Form Res 2025 Sep 19;9:e73844. [doi: 10.2196/73844] [Medline: 40971579]
59. Ahmed F, Romero Saletti SM, D'Souza E, et al. Piloting a cancer awareness app across six European countries: a pre-post study. Front Public Health 2025;13:1648212. [doi: 10.3389/fpubh.2025.1648212] [Medline: 41089856]
60. Almeida S, Pinto E, Correia M, Veiga N, Almeida A. Evaluating e-Health literacy, knowledge, attitude, and health online information in Portuguese university students: a cross-sectional study. Int J Environ Res Public Health 2024 Feb 27;21(3):271. [doi: 10.3390/ijerph21030271] [Medline: 38541272]
61. Andersen MH, Hermansen Å, Dahl KG, et al. Profiles of health literacy and digital health literacy in clusters of hospitalised patients: a single-centre, cross-sectional study. BMJ Open 2024 May 20;14(5):e077440. [doi: 10.1136/bmjopen-2023-077440] [Medline: 38772592]
62. Bäuerle A, Marsall M, Jahre LM, et al. Psychometric properties of the German revised version of the eHealth literacy scale in individuals with cardiac diseases: validation and test of measurement invariance. Digit Health 2023;9:20552076231194915. [doi: 10.1177/20552076231194915] [Medline: 37588160]
63. Bendtsen MG, Schönwandt BMT, Rubæk M, Hitz MF. Evaluation of an mHealth app on self-management of osteoporosis: prospective survey study. Interact J Med Res 2024 Apr 1;13:e53995. [doi: 10.2196/53995] [Medline: 38557362]
64. Bergh S, Benth J, Høgset LD, Rydjord B, Kayser L. Assessment of technology readiness in Norwegian older adults with long-term health conditions receiving home care services: cross-sectional questionnaire study. JMIR Aging 2025 Feb 7;8:e62936. [doi: 10.2196/62936] [Medline: 39918862]
65. Bergman L, Nilsson U, Dahlberg K, Jaensson M, Wångdahl J. Health literacy and e-Health literacy among Arabic-speaking migrants in Sweden: a cross-sectional study. BMC Public Health 2021 Nov 25;21(1):2165. [doi: 10.1186/s12889-021-12187-5] [Medline: 34823499]
66. Bergman L, Nilsson U, Dahlberg K, Jaensson M, Wångdahl J. Validity and reliability of the Arabic version of the HLS-EU-Q16 and HLS-EU-Q6 questionnaires. BMC Public Health 2023 Feb 10;23(1):304. [doi: 10.1186/s12889-023-15226-5] [Medline: 36765302]
67. Bergman L, Nilsson U, Dahlberg K, Jaensson M, Wångdahl J. Validity and reliability of the Swedish versions of the HLS-EU-Q16 and HLS-EU-Q6 questionnaires. BMC Public Health 2023 Apr 20;23(1):724. [doi: 10.1186/s12889-023-15519-9] [Medline: 37081538]
68. Bevilacqua R, Strano S, Di Rosa M, et al. eHealth literacy: from theory to clinical application for digital health improvement: results from the ACCESS training experience. Int J Environ Res Public Health 2021 Nov 10;18(22):11800. [doi: 10.3390/ijerph182211800] [Medline: 34831555]
69. Bíró É, Vincze F, Nagy-Pénzes G, Ádány R. Investigation of the relationship of general and digital health literacy with various health-related outcomes. Front Public Health 2023;11:1229734. [doi: 10.3389/fpubh.2023.1229734] [Medline: 37588120]
70. Breil B, Salewski C, Apolinário-Hagen J. Comparing the acceptance of mobile hypertension apps for disease management among patients versus clinical use among physicians: cross-sectional survey. JMIR Cardio 2022 Jan 6;6(1):e31617. [doi: 10.2196/31617] [Medline: 34989683]
71. Brørs G, Wentzel-Larsen T, Dalen H, et al. Psychometric properties of the Norwegian version of the electronic Health Literacy Scale (eHEALS) among patients after percutaneous coronary intervention: cross-sectional validation study. J Med Internet Res 2020 Jul 28;22(7):e17312. [doi: 10.2196/17312] [Medline: 32720900]
72. Burzyńska J, Rękas M, Januszewicz P. Evaluating the psychometric properties of the eHealth Literacy Scale (eHEALS) among Polish social media users. Int J Environ Res Public Health 2022 Mar 29;19(7):4067. [doi: 10.3390/ijerph19074067] [Medline: 35409753]
73. Chaniaud N, Sagnier C, Loup-Escande E. Translation and validation study of the French version of the eHealth Literacy Scale: web-based survey on a student population. JMIR Form Res 2022 Aug 31;6(8):e36777. [doi: 10.2196/36777] [Medline: 36044264]

XSL•FO
RenderX

74. Chatsatrian M, Kunde K, Bosompem J, et al. Usability evaluation of digital health applications for older people with depressive disorders: prospective observational study in a mixed methods design. JMIR Hum Factors 2025 Nov 28;12:e66271. [doi: 10.2196/66271] [Medline: 41328453]
75. Dale JG, Lüthi A, Fundingsland Skaraas B, Rundereim T, Dale B. Testing measurement properties of the Norwegian version of Electronic Health Literacy Scale (eHEALS) in a group of day surgery patients. J Multidiscip Healthc 2020;13:241-247. [doi: 10.2147/JMDH.S242985] [Medline: 32210568]
76. De Santis KK, Jahnel T, Sina E, Wienert J, Zeeb H. Digitization and health in Germany: cross-sectional nationwide survey. JMIR Public Health Surveill 2021 Nov 22;7(11):e32951. [doi: 10.2196/32951] [Medline: 34813493]
77. De Santis KK, Muellmann S, Pan CC, et al. Digitisation and health: second nationwide survey of internet users in Germany. Digit Health 2024;10:20552076241301457. [doi: 10.1177/20552076241301457] [Medline: 39600383]
78. Duplaga M. The acceptance of key public health interventions by the Polish population is related to health literacy, but not eHealth literacy. Int J Environ Res Public Health 2020 Jul 29;17(15):5459. [doi: 10.3390/ijerph17155459] [Medline: 32751180]
79. Duplaga M. The determinants of conspiracy beliefs related to the COVID-19 pandemic in a nationally representative sample of internet users. Int J Environ Res Public Health 2020 Oct 26;17(21):7818. [doi: 10.3390/ijerph17217818] [Medline: 33114542]
80. Duplaga M. The use of fitness influencers' websites by young adult women: a cross-sectional study. Int J Environ Res Public Health 2020 Sep 1;17(17):6360. [doi: 10.3390/ijerph17176360] [Medline: 32882887]
81. Duplaga M. A nationwide natural experiment of e-health implementation during the COVID-19 pandemic in Poland: user satisfaction and the ease-of-use of remote physician's visits. Int J Environ Res Public Health 2022 Jul 8;19(14):8338. [doi: 10.3390/ijerph19148338] [Medline: 35886190]
82. Duplaga M. The roles of health and e-health literacy, conspiracy beliefs and political sympathy in the adherence to preventive measures recommended during the pandemic. Int J Environ Res Public Health 2022 Jul 8;19(14):8346. [doi: 10.3390/ijerph19148346] [Medline: 35886198]
83. Duplaga M, Grysztar M. The association between future anxiety, health literacy and the perception of the COVID-19 pandemic: a cross-sectional study. Healthcare (Basel) 2021 Jan 5;9(1):43. [doi: 10.3390/healthcare9010043] [Medline: 33466487]
84. Duplaga M, Turosz N. User satisfaction and the readiness-to-use e-health applications in the future in Polish society in the early phase of the COVID-19 pandemic: A cross-sectional study. Int J Med Inform 2022 Dec;168:104904. [doi: 10.1016/j.ijmedinf.2022.104904] [Medline: 36332522]
85. Efthymiou A, Middleton N, Charalambous A, Papastavrou E. Health literacy and eHealth literacy and their association with other caring concepts among carers of people with dementia: a descriptive correlational study. Health Soc Care Community 2022 May;30(3):1109-1119. [doi: 10.1111/hsc.13341] [Medline: 33956368]
86. Efthymiou A, Kalaitzaki A, Rovithis M, Petrič G. Validation of the eHealth literacy scales: comparison between the shorter and longer versions. Inform Health Soc Care 2025 Jan 2;50(1):33-47. [doi: 10.1080/17538157.2025.2451427] [Medline: 39823355]
87. Fugmann D, Holsteg S, Schäfer R, Niegisch G, Dinger U, Karger A. Electronic health literacy, psychological distress, and quality of life in urological cancer patients: a longitudinal study during transition from inpatient to outpatient care. Curr Oncol 2025 Nov 13;32(11):637. [doi: 10.3390/curroncol32110637] [Medline: 41294699]
88. García-García D, Ajejas Bazán MJ, Pérez-Rivas FJ. Factors influencing eHealth literacy among Spanish primary healthcare users: cross-sectional study. Int J Environ Res Public Health 2022 Nov 23;19(23):15497. [doi: 10.3390/ijerph192315497] [Medline: 36497572]
89. García-García D, Bazán MJA, Pérez-Rivas FJ. Correlation between health and eHealth literacy and a healthy lifestyle: a cross-sectional study of Spanish primary healthcare patients. Healthcare (Basel) 2023 Nov 18;11(22):2980. [doi: 10.3390/healthcare11222980] [Medline: 37998472]
90. Garcia M, Machado R, Serra I, João AL. eHealth literacy in a migrant community and its association with chronic disease. Front Public Health 2025;13:1668262. [doi: 10.3389/fpubh.2025.1668262] [Medline: 41089873]
91. Gehrmann J, Stephan J, Dehner J, Stullich A, Richter M. Pilot study of an app-supported psychosocial prevention intervention: a mixed-methods approach. Pilot Feasibility Stud 2025 Dec 1;11(1):155. [doi: 10.1186/s40814-025-01737-y] [Medline: 41327486]
92. Geiger S, Esser AJ, Marsall M, et al. Association between eHealth literacy and health outcomes in German athletes using the GR-eHEALS questionnaire: a validation and outcome study. BMC Sports Sci Med Rehabil 2024 May 24;16(1):117. [doi: 10.1186/s13102-024-00902-9] [Medline: 38790069]
93. Georgsson M, Odzakovic E, Björk M, et al. Validation of the eHealth Literacy Scale instrument in a restless legs syndrome population: classical test theory and Rasch analysis study. J Med Internet Res 2025 Sep 10;27:e68474. [doi: 10.2196/68474] [Medline: 40930108]
94. Ghazi SN, Berner J, Anderberg P, Sanmartin Berglund J. The prevalence of eHealth literacy and its relationship with perceived health status and psychological distress during Covid-19: a cross-sectional study of older adults in Blekinge, Sweden. BMC Geriatr 2023 Jan 4;23(1):5. [doi: 10.1186/s12877-022-03723-y] [Medline: 36597040]

XSL•FO
RenderX

95. Giunti G, Yrttiaho T, Guardado-Medina S, et al. Feasibility and usability evaluation of a gamified fatigue management mobile application for persons with multiple sclerosis in everyday life. Mult Scler Relat Disord 2025 May;97:106379. [doi: 10.1016/j.msard.2025.106379] [Medline: 40073696]
96. Hermansen Å, Andersen MH, Borge CR, et al. Preliminary validity testing of the eHealth Literacy Questionnaire (eHLQ): a Confirmatory Factor Analysis (CFA) in Norwegian hospitalized patients. BMC Psychol 2023 Nov 23;11(1):409. [doi: 10.1186/s40359-023-01449-z] [Medline: 37996858]
97. Hernández Encuentra E, Robles N, Angulo-Brunet A, Cullen D, Del Arco I. Spanish and Catalan versions of the eHealth Literacy Questionnaire: translation, cross-cultural adaptation, and validation study. J Med Internet Res 2024 May 10;26:e49227. [doi: 10.2196/49227] [Medline: 38728072]
98. Hernández Encuentra E, González Caballero JL, Montagni I, Fernández Gutiérrez M, Bas Sarmiento P. Digital health literacy among the Spanish population: a descriptive and latent class analysis study. Eur J Public Health 2025 Aug 1;35(4):617-623. [doi: 10.1093/eurpub/ckaf016] [Medline: 40101757]
99. Holderried M, Hoeper A, Holderried F, et al. Attitude and potential benefits of modern information and communication technology use and telemedicine in cross-sectoral solid organ transplant care. Sci Rep 2021 Apr 27;11(1):9037. [doi: 10.1038/s41598-021-88447-6] [Medline: 33907269]
100. Holderried TAW, Stasik I, Schmitz MT, et al. Unleashing the potential of eHealth in outpatient cancer care for patients undergoing immunotherapy: a quantitative study considering patients' needs and current healthcare challenges. Front Digit Health 2024;6:1414442. [doi: 10.3389/fdgth.2024.1414442] [Medline: 39498102]
101. Hölgyesi Á, Luczay A, Tóth-Heyn P, et al. The impact of parental electronic health literacy on disease management and outcomes in pediatric type 1 diabetes mellitus: cross-sectional clinical study. JMIR Pediatr Parent 2024 Mar 20;7:e54807. [doi: 10.2196/54807] [Medline: 38506893]
102. Hölgyesi Á, Zrubka Z, Gulácsi L, et al. Robot-assisted surgery and artificial intelligence-based tumour diagnostics: social preferences with a representative cross-sectional survey. BMC Med Inform Decis Mak 2024 Mar 27;24(1):87. [doi: 10.1186/s12911-024-02470-x] [Medline: 38553703]
103. Holmen H, Holm AM, Falk RS, et al. A digital outpatient service with a mobile app for tailored care and health literacy in adults with long-term health service needs: multicenter nonrandomized controlled trial. J Med Internet Res 2025 Apr 28;27:e60343. [doi: 10.2196/60343] [Medline: 40294411]
104. Juvalta S, Kerry MJ, Jaks R, Baumann I, Dratva J. Electronic health literacy in Swiss-German parents: cross-sectional study of eHealth literacy scale unidimensionality. J Med Internet Res 2020 Mar 13;22(3):e14492. [doi: 10.2196/14492] [Medline: 32167476]
105. Katsaliaki K. Factors influencing use of eHealth services during and after the COVID-19 pandemic. Health Serv Manage Res 2025 May;38(2):97-106. [doi: 10.1177/09514848241275777] [Medline: 39177510]
106. Knitza J, Simon D, Lambrecht A, et al. Mobile health usage, preferences, barriers, and eHealth literacy in rheumatology: patient survey study. JMIR Mhealth Uhealth 2020 Aug 12;8(8):e19661. [doi: 10.2196/19661] [Medline: 32678796]
107. Kobryn M, Duplaga M. Does health literacy protect against cyberchondria: a cross-sectional study? Telemed J E Health 2024 Apr;30(4):e1089-e1100. [doi: 10.1089/tmj.2023.0425] [Medline: 38016126]
108. Kokwaro L, Krüger H, Stratmann D, et al. Digital health literacy among people with bipolar disorder in Germany: a cross-sectional survey. Front Psychiatry 2025;16:1651938. [doi: 10.3389/fpsyt.2025.1651938] [Medline: 41164090]
109. König L, Kuhlmey A, Suhr R. Digital health literacy of the population in Germany and its association with physical health, mental health, life satisfaction, and health behaviors: nationally representative survey study. JMIR Public Health Surveill 2024 Feb 21;10:e48685. [doi: 10.2196/48685] [Medline: 38381497]
110. Kostagiolas P, Koumbouli M, Theodorou P, Stamou S. Investigation of the information-seeking behavior of hospitalized patients at the general hospital of Corfu. J Hosp Librariansh 2021 Jan 2;21(1):36-60. [doi: 10.1080/15323269.2021.1860456]
111. Kretzschmar C, Knitza J, Pietschner R, Atreya R, Neurath MF, Orlemann T. mHealth use, preferences, barriers, and eHealth literacy among patients with inflammatory bowel disease: survey study. JMIR Hum Factors 2025 Nov 13;12:e64471. [doi: 10.2196/64471] [Medline: 41232095]
112. Kristjánsdóttir Ó, Welander Tärneberg A, Stenström P, Castor C, Kristensson Hallström I. eHealth literacy and socioeconomic and demographic characteristics of parents of children needing paediatric surgery in Sweden. Nurs Open 2023 Feb;10(2):509-524. [doi: 10.1002/nop2.1316] [Medline: 36053815]
113. Kubb C, Foran HM. Online health information seeking for self and child: an experimental study of parental symptom search. JMIR Pediatr Parent 2022 May 9;5(2):e29618. [doi: 10.2196/29618] [Medline: 35532970]
114. Lambrecht A, Vuillerme N, Raab C, et al. Quality of a supporting mobile app for rheumatic patients: patient-based assessment using the user version of the Mobile Application Scale (uMARS). Front Med (Lausanne) 2021;8:715345. [doi: 10.3389/fmed.2021.715345] [Medline: 34368202]
115. Lear R, Freise L, Kybert M, Darzi A, Neves AL, Mayer EK. Patients' willingness and ability to identify and respond to errors in their personal health records: mixed methods analysis of cross-sectional survey data. J Med Internet Res 2022 Jul 8;24(7):e37226. [doi: 10.2196/37226] [Medline: 35802397]

XSL•FO
RenderX

116. Lear R, Freise L, Kybert M, Darzi A, Neves AL, Mayer EK. Perceptions of quality of care among users of a web-based patient portal: cross-sectional survey analysis. J Med Internet Res 2022 Nov 17;24(11):e39973. [doi: 10.2196/39973] [Medline: 36394922]
117. Levin-Zamir D, Van den Broucke S, Bíró É, et al. Measuring digital health literacy and its associations with determinants and health outcomes in 13 countries. Front Public Health 2025;13:1472706. [doi: 10.3389/fpubh.2025.1472706] [Medline: 40182520]
118. Linnestad AM, Skogestad IJ, Gay CL, et al. Exploring digital health literacy clusters in a Norwegian stroke survivor population: a cross-sectional study (NORFAST). Digit Health 2025;11:20552076251380049. [doi: 10.1177/20552076251380049] [Medline: 41041464]
119. Lo Moro G, Scaioli G, Bert F, Zacchero AL, Minutiello E, Siliquini R. Exploring the the relationship between COVID-19 vaccine refusal and belief in fake news and conspiracy theories: a nationwide cross-sectional study in Italy. Int J Environ Res Public Health 2022 Jul 30;19(15):9350. [doi: 10.3390/ijerph19159350] [Medline: 35954711]
120. Lortz J, Rassaf T, Johannsen L, et al. Patient acceptance of video consultations in cardiology. Eur Heart J Digit Health 2025 Nov;6(6):1273-1281. [doi: 10.1093/ehjdh/ztaf089] [Medline: 41267832]
121. Lowe C, Browne M, Marsh W, Morrissey D. Usability testing of a digital assessment routing tool for musculoskeletal disorders: iterative, convergent mixed methods study. J Med Internet Res 2022 Aug 30;24(8):e38352. [doi: 10.2196/38352] [Medline: 36040787]
122. Luz S, Nogueira P, Costa A, Henriques A. Psychometric analysis of the eHealth Literacy Scale in Portuguese older adults (eHEALS-PT24): instrument development and validation. J Med Internet Res 2025 Feb 26;27:e57730. [doi: 10.2196/57730] [Medline: 40009846]
123. Marsall M, Engelmann G, Skoda EM, Teufel M, Bäuerle A. Measuring electronic health literacy: development, validation, and test of measurement invariance of a revised German version of the eHealth Literacy Scale. J Med Internet Res 2022 Feb 2;24(2):e28252. [doi: 10.2196/28252] [Medline: 35107437]
124. Marsall M, Dinse H, Schröder J, Skoda EM, Teufel M, Bäuerle A. Assessing electronic health literacy in individuals with the post-COVID-19 condition using the German revised eHealth Literacy Scale: validation study. JMIR Form Res 2024 Apr 25;8:e52189. [doi: 10.2196/52189] [Medline: 38662429]
125. Marsall M, Weigl M, Lüttel D, Müller H. Digital health literacy: a cross-sectional survey study among patients after hospitalization in Germany. Z Evid Fortbild Qual Gesundhwes 2025 Apr;193:18-25. [doi: 10.1016/j.zefq.2025.01.004] [Medline: 40023719]
126. Maurud S, Lunde L, Moen A, Opheim R. Mapping conditional health literacy and digital health literacy in patients with inflammatory bowel disease to optimise availability of digital health information: a cross-sectional study. Scand J Gastroenterol 2025 Jun;60(6):536-547. [doi: 10.1080/00365521.2025.2497952] [Medline: 40314186]
127. Moon Z, Zuchowski M, Moss-Morris R, Hunter MS, Norton S, Hughes LD. Disparities in access to mobile devices and e-health literacy among breast cancer survivors. Support Care Cancer 2022 Jan;30(1):117-126. [doi: 10.1007/s00520-021-06407-2] [Medline: 34236506]
128. Muellmann S, Wiersing R, Zeeb H, Brand T. Digital health literacy in adults with low reading and writing skills living in Germany: mixed methods study. JMIR Hum Factors 2025 May 22;12:e65345. [doi: 10.2196/65345] [Medline: 40402567]
129. Neves AL, Smalley KR, Freise L, Harrison P, Darzi A, Mayer EK. Determinants of use of the care information exchange portal: cross-sectional study. J Med Internet Res 2021 Nov 11;23(11):e23481. [doi: 10.2196/23481] [Medline: 34762063]
130. Nurtsch A, Teufel M, Jahre LM, et al. Drivers and barriers of patients' acceptance of video consultation in cancer care. Digit Health 2024;10:20552076231222108. [doi: 10.1177/20552076231222108] [Medline: 38188860]
131. Oliveira L, Zandonadi RP, Nakano EY, et al. From validation to assessment of e-health literacy: a study among higher education students in Portugal. Healthcare (Basel) 2024 Aug 15;12(16):1626. [doi: 10.3390/healthcare12161626] [Medline: 39201184]
132. Olsbø S, Kiserud SG, Hermansen Å, Hamilton Larsen M, Bjørnland K. Health literacy in parents of children with Hirschsprung disease: a novel study. Pediatr Surg Int 2024 Dec 5;41(1):21. [doi: 10.1007/s00383-024-05917-4] [Medline: 39633025]
133. Olsbø S, Larsen MH, Kiserud SG, Hagen TS, Hermansen Å, Bjørnland K. Parental health literacy in anorectal malformation: needs and challenges. Pediatr Surg Int 2025 Jul 16;41(1):214. [doi: 10.1007/s00383-025-06096-6] [Medline: 40670693]
134. Olszewski R, Watros KM, Brzeziński J, et al. COVID-19 health communication strategies for older adults: chatbots and traditional media. Adv Clin Exp Med 2025 Oct;34(10):1739-1747. [doi: 10.17219/acem/195242] [Medline: 39679656]
135. Pacut A, Duplaga M, Więcek S, et al. The relationship between stress, anxiety, and health literacy in parents of children with chronic gastroenterological diseases: a multi-center cross-sectional study. BMC Public Health 2025 Oct 6;25(1):3357. [doi: 10.1186/s12889-025-24329-0] [Medline: 41053680]
136. Palisi A. Digitale Gesundheitskompetenz bei chronischen, nicht-spezifischen Rückenschmerzen [Article in German]. MSK-Muskuloskel Phys 2024 Jul;28(3):176-184. [doi: 10.1055/a-2302-0159]
137. Pan CC, De Santis KK, Muellmann S, et al. Sociodemographics and digital health literacy in using wearables for health promotion and disease prevention: cross-sectional nationwide survey in Germany. J Prev (2022) 2025 Jun;46(3):371-391. [doi: 10.1007/s10935-024-00821-y] [Medline: 39692799]

XSL•FO
RenderX

138. Papp-Zipernovszky O, Horváth MD, Schulz PJ, Csabai M. Generation gaps in digital health literacy and their impact on health information seeking behavior and health empowerment in Hungary. Front Public Health 2021;9:635943. [doi: 10.3389/fpubh.2021.635943] [Medline: 34055714]
139. Piper L, De Cosmo LM, Benvenuto M, Viola C. How do the determinants of collaborative consumption influence its use in healthcare? A managerial perspective. Int J Health Policy Manag 2024;13:8453. [doi: 10.34172/ijhpm.8453] [Medline: 39620511]
140. Pisl V, Volavka J, Chvojkova E, Cechova K, Kavalirova G, Vevera J. Dissociation, cognitive reflection and health literacy have a modest effect on belief in conspiracy theories about COVID-19. Int J Environ Res Public Health 2021 May 11;18(10):5065. [doi: 10.3390/ijerph18105065] [Medline: 34065023]
141. Pisl V, Volavka J, Chvojkova E, Cechova K, Kavalirova G, Vevera J. Willingness to vaccinate against COVID-19: the role of health locus of control and conspiracy theories. Front Psychol 2021;12:717960. [doi: 10.3389/fpsyg.2021.717960] [Medline: 34744880]
142. Poot CC, Meijer E, Fokkema M, Chavannes NH, Osborne RH, Kayser L. Translation, cultural adaptation and validity assessment of the Dutch version of the eHealth Literacy Questionnaire: a mixed-method approach. BMC Public Health 2023 May 30;23(1):1006. [doi: 10.1186/s12889-023-15869-4] [Medline: 37254148]
143. Qiu CS, Lunova T, Greenfield G, et al. Determinants of digital health literacy: International cross-sectional study. J Med Internet Res 2025 Jun 30;27:e66631. [doi: 10.2196/66631] [Medline: 40587845]
144. Ramjee S, Mohamedthani H, Patel AU, et al. The effect of remote digital services on health care inequalities among people under long-term dermatology follow-up: cross-sectional questionnaire study. JMIR Dermatol 2023 Dec 8;6:e48981. [doi: 10.2196/48981] [Medline: 38064259]
145. Ramstad KJ, Brørs G, Pettersen TR, et al. eHealth technology use and eHealth literacy after percutaneous coronary intervention. Eur J Cardiovasc Nurs 2023 Jul 19;22(5):472-481. [doi: 10.1093/eurjcn/zvac087] [Medline: 36190843]
146. Rognsvåg T, Nordmo IK, Bergvad IB, et al. Digital health literacy in Norwegian patients with hip and knee arthroplasty: normative data from a cross-sectional study. Acta Orthop 2024 Nov 25;95:677-684. [doi: 10.2340/17453674.2024.42304] [Medline: 39584821]
147. Rokohl AC, Pine NS, Adler W, et al. Health literacy in patients wearing prosthetic eyes: a prospective cross-sectional study. Curr Eye Res 2023 Jan;48(1):11-17. [doi: 10.1080/02713683.2022.2141783] [Medline: 36305786]
148. Rosenmeier N, Busk D, Dichman C, Nielsen KM, Kayser L, Wagner MK. Technology readiness level and self-reported health in recipients of an implantable cardioverter defibrillator: cross-sectional study. JMIR Cardio 2025 Feb 6;9:e58219. [doi: 10.2196/58219] [Medline: 39913910]
149. Scacchi A, Lo Moro G, Giacomini G, et al. Trust levels toward health care and government: insights from TrustMe, an Italian cross-sectional study. J Prev Med Hyg 2024 Jun;65(2):E125-E133. [doi: 10.15167/2421-4248/jpmh2024.65.2.3231] [Medline: 39430985]
150. Schaeffer D, Klinger J, Berens EM, et al. Health literacy in Germany before and during the COVID-19 pandemic. Gesundheitswesen 2021 Oct;83(10):781-788. [doi: 10.1055/a-1560-2479] [Medline: 34547817]
151. Schaeffer D, Gille S, Berens EM, et al. Digitale Gesundheitskompetenz der Bevölkerung in Deutschland: Ergebnisse des HLS-GER 2 [Article in German]. Gesundheitswesen 2023 Apr;85(4):323-331. [doi: 10.1055/a-1670-7636] [Medline: 34905785]
152. Schaeffer D, Klinger J, Berens EM. Digitale Gesundheitskompetenz von Personen mit und ohne Migrationserfahrung – Ein Vergleich von zwei Querschnittbefragungen [Article in German]. Präv Gesundheitsf 2024 Nov;19(4):552-559. [doi: 10.1007/s11553-024-01112-4]
153. Schmieding ML, Kopka M, Bolanaki M, et al. Impact of a symptom checker app on patient-physician interaction among self-referred walk-in patients in the emergency department: multicenter, parallel-group, randomized, controlled trial. J Med Internet Res 2025 Apr 2;27:e64028. [doi: 10.2196/64028] [Medline: 40173434]
154. Schomakers EM, Lidynia C, Vervier LS, Calero Valdez A, Ziefle M. Applying an extended UTAUT2 model to explain user acceptance of lifestyle and therapy mobile health apps: survey study. JMIR Mhealth Uhealth 2022 Jan 18;10(1):e27095. [doi: 10.2196/27095] [Medline: 35040801]
155. Schulz PJ, Pessina A, Hartung U, Petrocchi S. Effects of objective and subjective health literacy on patients' accurate judgment of health information and decision-making ability: survey study. J Med Internet Res 2021 Jan 21;23(1):e20457. [doi: 10.2196/20457] [Medline: 33475519]
156. Sippel A, Riemann-Lorenz K, Pöttgen J, et al. Validation of the German eHealth impact questionnaire for online health information users affected by multiple sclerosis. BMC Med Inform Decis Mak 2022 Aug 16;22(1):219. [doi: 10.1186/s12911-022-01968-6] [Medline: 35974395]
157. Sjöström A, Hajdarevic S, Hörnsten Å, Öberg U, Isaksson U. Experiences of online COVID-19 information acquisition among persons with type 2 diabetes and varying eHealth literacy. Int J Environ Res Public Health 2021 Dec 15;18(24):13240. [doi: 10.3390/ijerph182413240] [Medline: 34948852]
158. Sjöström AE, Hajdarevic S, Hörnsten Å, Kristjánsdóttir Ó, Castor C, Isaksson U. The Swedish version of the eHealth Literacy Questionnaire: translation, cultural adaptation, and validation study. J Med Internet Res 2023 Apr 12;25:e43267. [doi: 10.2196/43267] [Medline: 37043268]

XSL•FO
RenderX

159. Sjöström A, Hajdarevic S, Hörnsten Å, Isaksson U. eHealth literacy and health-related internet use among Swedish primary health care visitors: cross-sectional questionnaire study. JMIR Form Res 2024 Dec 5;8:e63288. [doi: 10.2196/63288] [Medline: 39637377]
160. Smoła P, Zwierczyk U, Duplaga M. Transactional e-health literacy and its association with e-health services use in Polish adults: a cross-sectional study. Front Digit Health 2024;6:1458650. [doi: 10.3389/fdgth.2024.1458650] [Medline: 39650753]
161. Sollie M, Hansen M, Thomsen JB. Health technology readiness amongst patients with suspected breast cancer using the READHY-tool: a cross-sectional study. J Med Syst 2023 Nov 16;47(1):118. [doi: 10.1007/s10916-023-02016-0] [Medline: 37971517]
162. Spanakis P, Lorimer B, Newbronner E, et al. Digital health literacy and digital engagement for people with severe mental ill health across the course of the COVID-19 pandemic in England. BMC Med Inform Decis Mak 2023 Sep 26;23(1):193. [doi: 10.1186/s12911-023-02299-w] [Medline: 37752460]
163. Spindler H, Dyrvig AK, Schacksen CS, et al. Increased motivation for and use of digital services in heart failure patients participating in a telerehabilitation program: a randomized controlled trial. Mhealth 2022;8:25. [doi: 10.21037/mhealth-21-56] [Medline: 35928510]
164. Springer F, Hambsch PK, Mehnert-Theuerkauf A, Nicolay NH. Digital support and artificial intelligence in cancer patients undergoing radiation therapy: patient utilization, acceptance and attitudes. Front Oncol 2025;15:1546221. [doi: 10.3389/fonc.2025.1546221] [Medline: 41079088]
165. Stajszczyk M, Świerkowska G, Smolik K, Domysławska I, Charkiewicz K, Samborski W. The perspective of Polish patients with rheumatoid arthritis: treatment expectations, patient-reported outcomes, and digital literacy (the SENSE study). Reumatologia 2023;61(5):331-338. [doi: 10.5114/reum/171625] [Medline: 37970121]
166. Stephan J, Gehrmann J, Dehner JC, Stullich A, Richter M. Development and validation of the eHealth Literacy and Use Scale (eHLUS) to measure medical app literacy. Public Health 2025 Mar;240:27-32. [doi: 10.1016/j.puhe.2024.12.057] [Medline: 39854851]
167. Stephen DA, Nordin A, Johansson UB, Nilsson J. eHealth literacy and its association with demographic factors, disease-specific factors, and well-being among adults with type 1 diabetes: cross-sectional survey study. JMIR Diabetes 2025 Mar 31;10:e66117. [doi: 10.2196/66117] [Medline: 40163628]
168. Sylwander C, Wahl AK, Andersson MLE, Haglund E, Larsson I. Health literacy in individuals with knee pain: a mixed methods study. BMC Public Health 2023 Aug 29;23(1):1656. [doi: 10.1186/s12889-023-16585-9] [Medline: 37644536]
169. Terp R, Kayser L, Lindhardt T. Older patients' competence, preferences, and attitudes toward digital technology use: explorative study. JMIR Hum Factors 2021 May 14;8(2):e27005. [doi: 10.2196/27005] [Medline: 33988512]
170. Thorsen IK, Rossen S, Glümer C, Midtgaard J, Ried-Larsen M, Kayser L. Health technology readiness profiles among Danish individuals with type 2 diabetes: cross-sectional study. J Med Internet Res 2020 Sep 15;22(9):e21195. [doi: 10.2196/21195] [Medline: 32930669]
171. Totaro M, Cicolini G, Bianconi A, et al. Digital health literacy in patients with hypertension: a cross-sectional study. J Clin Nurs 2026 Jan;35(1):158-171. [doi: 10.1111/jocn.17811] [Medline: 40369694]
172. Tschamper MK, Wahl AK, Hermansen Å, Jakobsen R, Larsen MH. Parents of children with epilepsy: characteristics associated with high and low levels of health literacy. Epilepsy Behav 2022 May;130:108658. [doi: 10.1016/j.yebeh.2022.108658] [Medline: 35358855]
173. Turnbull J, Prichard J, MacLellan J, Pope C. eHealth literacy and the use of NHS 111 online urgent care service in England: cross-sectional survey. J Med Internet Res 2024 Jun 4;26:e50376. [doi: 10.2196/50376] [Medline: 38833297]
174. Ullrich G, Bäuerle A, Vogt H, et al. Digital health literacy and attitudes toward eHealth technologies among patients with cardiovascular disease and their implications for secondary prevention: survey study. JMIR Form Res 2025 Mar 19;9:e63057. [doi: 10.2196/63057] [Medline: 40106277]
175. Vahteristo A, Jylhä V, Kuusisto H. The use and readiness for eHealth and eWelfare among young adults. Health Informatics J 2025;31(1):14604582241307208. [doi: 10.1177/14604582241307208] [Medline: 39901610]
176. Valan L, Isaksson U, Hörnsten A, Carlsund A. Evaluating the impact of digital support on parental stress in Swedish child health care: results from an intervention study. Int J Pediatr 2025;2025(1):8780069. [doi: 10.1155/ijpe/8780069] [Medline: 40599560]
177. Valentim P, Arriaga M, Nogueira P, Costa A. Digital and navigational health literacy in surgical patients: vulnerabilities in the transition to post-discharge care. Healthcare (Basel) 2025 Dec 10;13(24):3227. [doi: 10.3390/healthcare13243227] [Medline: 41464296]
178. Van Rhoon L, McSharry J, Byrne M. Development and testing of a digital health acceptability model to explain the intention to use a digital diabetes prevention programme. Br J Health Psychol 2022 Sep;27(3):716-740. [doi: 10.1111/bjhp.12569] [Medline: 34719099]
179. Villadsen SF, Hadi H, Ismail I, Osborne RH, Ekstrøm CT, Kayser L. eHealth literacy and health literacy among immigrants and their descendants compared with women of Danish origin: a cross-sectional study using a multidimensional approach among pregnant women. BMJ Open 2020 May 7;10(5):e037076. [doi: 10.1136/bmjopen-2020-037076] [Medline: 32385065]
180. Vitolo M, Ziveri V, Gozzi G, et al. DIGItal health literacy after COVID-19 outbreak among frail and non-frail cardiology patients: the DIGI-COVID study. J Pers Med 2022 Dec 31;13(1):99. [doi: 10.3390/jpm13010099] [Medline: 36675760]

XSL•FO
RenderX

181. Wångdahl J, Jaensson M, Dahlberg K, Nilsson U. The Swedish version of the Electronic Health Literacy Scale: prospective psychometric evaluation study including thresholds levels. JMIR Mhealth Uhealth 2020 Feb 24;8(2):e16316. [doi: 10.2196/16316] [Medline: 32130168]
182. Wångdahl J, Dahlberg K, Jaensson M, Nilsson U. Arabic version of the Electronic Health Literacy Scale in Arabic-speaking individuals in Sweden: prospective psychometric evaluation study. J Med Internet Res 2021 Mar 22;23(3):e24466. [doi: 10.2196/24466] [Medline: 33749614]
183. Wecker H, Höllerl L, Schick TS, Biedermann T, Zink A, Ziehfreund S. Patient journey and disease-related digital media usage: a cross-sectional study among dermatology patients across Germany. J Dtsch Dermatol Ges 2024 Mar;22(3):377-387. [doi: 10.1111/ddg.15311] [Medline: 38308182]
184. Wetzel AJ, Klemmt M, Müller R, Rieger MA, Joos S, Koch R. Only the anxious ones? Identifying characteristics of symptom checker app users: a cross-sectional survey. BMC Med Inform Decis Mak 2024 Jan 23;24(1):21. [doi: 10.1186/s12911-024-02430-5] [Medline: 38262993]
185. Zrubka Z, Brito Fernandes Ó, Baji P, et al. Exploring eHealth literacy and patient-reported experiences with outpatient care in the Hungarian general adult population: cross-sectional study. J Med Internet Res 2020 Aug 11;22(8):e19013. [doi: 10.2196/19013] [Medline: 32667891]
186. Zrubka Z, Vékás P, Németh P, et al. Validation of the PAM-13 instrument in the Hungarian general population 40 years old and above. Eur J Health Econ 2022 Nov;23(8):1341-1355. [doi: 10.1007/s10198-022-01434-0] [Medline: 35102464]
187. Zwierczyk U, Sowada C, Duplaga M. Eating choices-the roles of motivation and health literacy: a cross-sectional study. Nutrients 2022 Sep 28;14(19):4026. [doi: 10.3390/nu14194026] [Medline: 36235678]
188. Zwierczyk U, Kobryn M, Duplaga M. The awareness of the role of commercial determinants of health and the readiness to accept restrictions on unhealthy food advertising in Polish society. Nutrients 2023 Nov 10;15(22):4743. [doi: 10.3390/nu15224743] [Medline: 38004137]
189. The HLS19 Consortium of the WHO Action Network M-POHL. The HLS19-DIGI Instrument to Measure Digital Health Literacy: Austrian National Public Health Institute; 2022. URL: https://m-pohl.net/sites/m-pohl.net/files/inline-files/Factsheet%20HLS19-DIGI.pdf [accessed 2026-05-01]
190. Kayser L, Rossen S, Karnoe A, et al. Development of the multidimensional Readiness and Enablement Index for Health Technology (READHY) tool to measure individuals' health technology readiness: initial testing in a cancer rehabilitation setting. J Med Internet Res 2019 Feb 12;21(2):e10377. [doi: 10.2196/10377] [Medline: 30747717]
191. Osborne RH, Elsworth GR, Whitfield K. The Health Education Impact Questionnaire (heiQ): an outcomes and evaluation measure for patient education and self-management interventions for people with chronic conditions. Patient Educ Couns 2007 May;66(2):192-201. [doi: 10.1016/j.pec.2006.12.002] [Medline: 17320338]
192. Osborne RH, Batterham RW, Elsworth GR, Hawkins M, Buchbinder R. The grounded psychometric development and initial validation of the Health Literacy Questionnaire (HLQ). BMC Public Health 2013 Jul 16;13(1):658. [doi: 10.1186/1471-2458-13-658] [Medline: 23855504]
193. van der Vaart R, Drossaert CHC, de Heus M, Taal E, van de Laar M. Measuring actual eHealth literacy among patients with rheumatic diseases: a qualitative analysis of problems encountered using Health 1.0 and Health 2.0 applications. J Med Internet Res 2013 Feb 11;15(2):e27. [doi: 10.2196/jmir.2428] [Medline: 23399720]
194. Soellner R, Huber S, Reder M. The concept of eHealth literacy and its measurement: German translation of the eHEALS. J Media Psychol 2014 Jan 1;26:29-38. [doi: 10.1027/1864-1105/a000104]
195. Efthymiou A, Middleton N, Charalambous A, Papastavrou E. Adapting the eHealth literacy scale for carers of people with chronic diseases (eHeals-Carer) in a sample of Greek and Cypriot carers of people with dementia: reliability and validation study. J Med Internet Res 2019 Nov 28;21(11):e12504. [doi: 10.2196/12504] [Medline: 31778120]
196. Paige SR, Stellefson M, Krieger JL, Miller MD, Cheong J, Anderson-Lewis C. Transactional eHealth literacy: developing and testing a multi-dimensional instrument. J Health Commun 2019;24(10):737-748. [doi: 10.1080/10810730.2019.1666940] [Medline: 31583963]
197. Paige SR, Stellefson M, Krieger JL, Anderson-Lewis C, Cheong J, Stopka C. Proposing a transactional model of eHealth literacy: concept analysis. J Med Internet Res 2018 Oct 2;20(10):e10175. [doi: 10.2196/10175] [Medline: 30279155]
198. Paulhus DL, Vazire S. The self-report method. In: Handbook of Research Methods in Personality Psychology: Guilford Press; 2007, Vol. 1:224-239.
199. Podsakoff PM, MacKenzie SB, Lee JY, Podsakoff NP. Common method biases in behavioral research: a critical review of the literature and recommended remedies. J Appl Psychol 2003 Oct;88(5):879-903. [doi: 10.1037/0021-9010.88.5.879] [Medline: 14516251]
200. Bethlehem J. Selection bias in web surveys. Int Statistical Rev 2010 Aug;78(2):161-188. [doi: 10.1111/j.1751-5823.2010.00112.x]
201. Digital economy and society statistics - households and individuals. Eurostat. 2025. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Digital_economy_and_society_statistics_-_households_and_individuals [accessed 2026-05-01]
202. Medero K, Merrill K Jr, Ross MQ. access across the digital divide for intersectional groups seeking web-based health information: national survey. J Med Internet Res 2022 Mar 15;24(3):e32678. [doi: 10.2196/32678] [Medline: 35289761]

XSL•FO
RenderX

203. Cheng C, Beauchamp A, Elsworth GR, Osborne RH. Applying the electronic health literacy lens: systematic review of electronic health interventions targeted at socially disadvantaged groups. J Med Internet Res 2020 Aug 13;22(8):e18476. [doi: 10.2196/18476] [Medline: 32788144]
204. Vannieuwenhuyze J, Loosveldt G, Molenberghs G. A method for evaluating mode effects in mixed-mode surveys. Public Opin Q 2010 Jan 1;74(5):1027-1045. [doi: 10.1093/poq/nfq059]
205. Tran AD, White AE, Torok MR, Jervis RH, Albanese BA, Walter EJS. Lessons learned from a sequential mixed-mode survey design to recruit and collect data from case-control study participants: formative evaluation. JMIR Form Res 2024 May 27;8:e56218. [doi: 10.2196/56218] [Medline: 38801768]
206. Koch MJ, Wienrich C, Straka S, Latoschik ME, Carolus A. Overview and confirmatory and exploratory factor analysis of AI literacy scale. Comput Educ Artif Intell 2024 Dec;7:100310. [doi: 10.1016/j.caeai.2024.100310]
207. Lintner T. A systematic review of AI literacy scales. NPJ Sci Learn 2024 Aug 6;9(1):50. [doi: 10.1038/s41539-024-00264-4] [Medline: 39107327]
208. Wang X, Zhang C, Qi Y, et al. Digital health literacy questionnaire for older adults: instrument development and validation study. J Med Internet Res 2025 Mar 19;27:e64193. [doi: 10.2196/64193] [Medline: 40106815]
209. Quinn K. Methodological considerations in surveys of older adults: technology matters. Int J Emerg Technol Soc 2010;8(2):114-133. [doi: 10.13140/2.1.3897.9209]
210. Kebede AS, Ozolins LL, Holst H, Galvin K. Digital engagement of older adults: scoping review. J Med Internet Res 2022 Dec 7;24(12):e40192. [doi: 10.2196/40192] [Medline: 36477006]

Abbreviations

DHLI: Digital Health Literacy Instrument
DiGA: Digital Health Applications (Digitale Gesundheitsanwendungen)
DVG: Digital Care Act (Digitale-Versorgung-Gesetz)
eHEALS: eHealth Literacy Scale
eHEALS-Carer: eHealth Literacy Scale for Carers of People with Chronic Diseases
eHEALS-E: eHealth Literacy Scale-Extended
eHL: eHealth literacy
eHLF: eHealth Literacy Framework
eHLQ: eHealth Literacy Questionnaire
eHLUS: eHealth Literacy and Use Scale
GDPR: General Data Protection Regulation
GenAI: generative artificial intelligence
HLS_{19}-DIGI: digital health literacy module of the Health Literacy Survey 2019‐2021
PCC: population, concept, and context framework
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
READHY: Readiness and Enablement Index for Health Technology
TeHLI: Transactional eHealth Literacy Instrument
TF-IDF: term frequency-inverse document frequency

Edited by S Brini; submitted 09.Nov.2025; peer-reviewed by E Brainin, F Welzel; revised version received 10.Apr.2026; accepted 13.Apr.2026; published 22.May.2026.

Please cite as:
Weist JV, Nitsche J, Ehlers JP, Busse TS
Measuring eHealth Literacy in the European Economic Area, Switzerland, and the United Kingdom: Scoping Review
J Med Internet Res 2026;28:e87461
URL: https://www.jmir.org/2026/1/e87461
doi:10.2196/87461

© Jonah Valentin Weist, Julia Nitsche, Jan Peter Ehlers, Theresa Sophie Busse. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN

XSL•FO
RenderX

1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring Technological Solutions for Interoperability Between Patient Electronic Medical Records and Clinical Registries: Scoping Review

Erika Haynes[1,2], BN; James Brannigan[3], MD; Jessica Suna[4], MHS; Reid Malseed[2], BCA; Alana Delaforce[5], PhD; Rachel Mulvenney-Fenner[2], MN; Katherine Alog-Daroya[2], MN; Karin Plummer[1,2], PhD; Craig McBride[2,3], PhD; Roy Kimble[2,3], DMed; Bronwyn Griffin[1,2], PhD

[1]School of Nursing and Midwifery, Griffith Health, Griffith University, 170 Kessels Road, Nathan, Brisbane, Queensland, Australia
[2]Children's Health Queensland, Hospital and Health Service, Brisbane, Australia
[3]The University of Queensland, Brisbane, Australia
[4]Central Queensland University, Rockhampton, Australia
[5]CSIRO, Brisbane, Australia

Corresponding Author:
Erika Haynes, BN
School of Nursing and Midwifery, Griffith Health, Griffith University, 170 Kessels Road, Nathan, Brisbane, Queensland, Australia

Abstract

Background: The use of electronic medical records (EMRs) and clinical registries has transformed health care delivery by improving data management, care coordination, and research capacity. However, the full potential of these technologies can only be realized through effective interoperability, thereby reducing the burden of manual data entry and enhancing the use of real-world clinical data.

Objective: This review examines technologies that enable automated data extraction and transfer, which promote interoperability between EMRs and clinical registries.

Methods: A search of PubMed, CINAHL, Embase, and Web of Science, including studies published between January 2013 and April 2025, was registered with Open Science Framework a priori and involved three key concepts: (1) “registry,” (2) “electronic medical records,” and (3) “interoperability.” A 2-phase screen identified studies evaluating technologies that facilitate automated data extraction or interoperability. Automation was defined as fully automated, where data are extracted and transferred without human intervention, or semiautomated, where extraction or transfer is predominantly automated but may include manual validation. Only technologies supporting ongoing database integration were eligible for inclusion. Screening, data extraction, and synthesis were conducted by multiple independent reviewers. Technology experts provided extensive input and guidance throughout to ensure the accuracy and relevance of the extracted information.

Results: Overall, 36 studies met the inclusion criteria, representing 12 countries across 5 continents and addressing a wide range of acute and chronic health conditions. Epic was the most frequently reported EMR system, while the most common registry platforms were REDCap (Research Electronic Data Capture; Vanderbilt University), structured query language (SQL) server database, and EMR-embedded solutions. Most approaches centered around extracting data from structured formats (n=18), or a combination of both structured and unstructured formats (n=10), emphasizing the central role of structured EMR data in current automated extraction approaches.

Conclusions: This review advances understanding of interoperability between EMRs and clinical registries by uniquely examining automated and sustainable solutions for data exchange, extending beyond prior work that has largely focused on technologies designed for isolated systems or study-specific data extraction. A novel contribution of this review is the synthesis of context-specific considerations derived from reported implementations, providing a comprehensive overview of how technology selection and implementation are shaped by the context in which they are deployed. While these advancements have reduced reliance on inefficient, error-prone, and resource-intensive manual processes, ongoing challenges in data standardization, seamless integration, and long-term sustainability are compounded by poor and inconsistent reporting across studies. Future efforts should follow comprehensive reporting guidelines, adhere to robust governance principles, and incorporate implementation science frameworks, to not only enable meaningful comparison and synthesis in future research, but also to ensure that technologies can be effectively, feasibly, and sustainably integrated within health care contexts, while upholding the ethical and equitable use of health care data.

(J Med Internet Res 2026;28:e82380) doi:10.2196/82380

XSL•FO
RenderX

KEYWORDS

registry; electronic medical record; interoperability; data extraction; data transfer

Introduction

In health care, the integration of innovative technology has driven unprecedented progress in communication and information systems, resulting in significant impacts on the efficiency, accuracy, and quality of health care delivery [1]. Electronic medical records (EMRs) and clinical registries exemplify the sophisticated application of technology to streamline data management, enhance patient care coordination, and support robust clinical research [2]. Clinical registries in particular are increasingly leveraged for research, quality assurance, and benchmarking, allowing health care providers and researchers the ability to maximize resources and improve patient outcomes in an era of growing demands and limited funding [3,4]. However, the full potential of these systems can only be realized through effective interoperability, which ensures seamless data exchange and integration across diverse health care platforms [5].

Interoperability is broadly defined as "the ability of two or more systems to work together, regardless of different interfaces, platforms, and technologies adopted" in a way that facilitates data-sharing and data-use in improving health care delivery [6]. In this context, the promise of interoperability between patient EMRs and clinical registries represents enormous potential to minimize the burden of data input, management, and maintenance, while maximizing research productivity in line with real-world experiences [7-9]. The issue therein not only stems from technical limitations but also encompasses concerns regarding maintaining information fidelity alongside data standardization, ensuring data privacy and security, and navigating regulatory constraints [9,10]. Overcoming these challenges is pivotal to harnessing the full potential of registries and EMRs, moving toward a more integrated and efficient health care technology ecosystem.

In contemporary health care, EMRs play a crucial role in capturing comprehensive patient data across various clinical encounters [11]. The data, documented in either a structured format (eg, predefined fields such as diagnosis codes, lab results, and medication lists) or an unstructured format (eg, free-text clinical notes, summaries, and narrative reports) [12], facilitates patient care by centralizing all patient information in one location. This centralization keeps multidisciplinary team members informed, supports clinical decision-making, and ensures continuity of care across diverse health care settings [13,14]. Typically, these same data elements are also leveraged in clinical registries to systematically collect, study, and interpret patient data relevant to specific medical conditions [15,16], thus filling broad knowledge and evidence gaps that are challenging and costly to capture using traditional research methods [15,17]. At present, the population of registry data predominantly relies on manual methods, where data are extracted from EMRs and transcribed into the correct format for registry use [18-20]. Despite the translational benefits of clinical registry data use, this current process is time-consuming, error-prone, resource-intensive, and often constrained by the availability of ongoing funding [20-24].

Automated data extraction and transfer technologies have emerged as critical tools to improve interoperability, reduce data entry burden, and enhance the timeliness and completeness of health information for research, quality improvement, and policy development [25,26]. However, while these technological advancements offer significant benefits, there is currently no universally accepted "gold standard" methodology, and challenges surrounding adoption are not yet fully understood. Successful integration of these tools requires not only technical accuracy in extracting and transmitting standardized data, but also robust measures to protect patient privacy, ensure data security, and comply with regulatory requirements [25,27,28]. Adoption is further shaped by organizational, technical, and cultural barriers, underscoring the value of implementation science frameworks to guide effective and sustainable integration [29,30]. Within this context, this review primarily aims to explore existing automated data extraction and transfer processes from EMRs to clinical registries, with a focus on technologies used, data fidelity, and governance.

The specific review questions are:

1. What technologies are being used to facilitate interoperability through automated data extraction and transfer from EMRs to clinical registries or databases?
2. To what degree has the implementation of automated data extraction and transfer tools successfully enabled accurate and complete transmission of data in a standardized format?
3. What measures have been implemented to safeguard data privacy, ensure security, and comply with regulatory requirements during automated data extraction and transfer?
4. What implementation challenges have been reported in the adoption of these technologies, and what implementation frameworks have been used to support their effective integration into clinical data systems?

Methods

Study Design

The PRISMA-ScR (Preferred Reporting Items of Systematic Reviews and Meta-Analyses extension for Scoping Reviews) was used to conduct this review [31,32]. The review protocol was registered with Open Science Framework prior to study commencement [33].

Search Strategy

The search strategy was developed in consultation with a university health librarian to find published studies relevant to the review questions. The search strategy comprised three key concepts: (1) registry, (2) electronic health records (EHR), and (3) interoperability (Table S1 in Multimedia Appendix 1). While both terms, EMR and EHR, were included in the search strategy to ensure comprehensiveness, this study uses EMR to

XSL•FO
RenderX

collectively refer to both terms, except where the original technology name explicitly uses EHR.

The Systematic Review Accelerator Polyglot Search Translator automation tool [34] was used to convert this search strategy into the appropriate formats for 4 databases, including PubMed, CINAHL, Embase, and Web of Science. The reference lists of all included sources were hand-searched. Non-indexed publications were considered if identified through reference list hand-searching, but were not systematically sought beyond the selected databases. Gray literature was neither searched nor eligible, as the review focused exclusively on peer-reviewed studies. The search included all studies published between January 1, 2013, and April 30, 2025, inclusive.

Eligibility Criteria

The core concept of interest was any study exploring technology enabling interoperability between EMRs and research databases, such as clinical registries. Automation was defined as either fully automated, in which data are extracted and transferred without human intervention, or semiautomated, in which extraction or transfer is predominantly automated but may include manual validation steps. This ability to automate the process of extracting and transferring data for research purposes was key for inclusion and, as such, any study using a manual extraction or transfer process was excluded. Publications focusing on technology to extract data for a single study, rather than for integration into ongoing databases or registries, were also excluded. Eligibility within this scoping review was not limited by a specific type of participant, population, or health care context. Similarly, the study design and the specific EMR or registry used did not affect inclusion. Only full-text, peer-reviewed, primary research studies available in English were considered for inclusion, ensuring sufficient detail on the relevant technology was available.

Study Selection

Studies were uploaded into Covidence (Covidence systematic review software) for screening and de-duplication [35]. Four independent screeners (ED, JB, RM, and KA) completed a title and abstract screen. Conflicts were resolved by discussion between 2 reviewers (ED and JB). Once all conflicts were resolved, the same screeners independently completed the full-text screen, recording the rationale for study exclusion to ensure transparency.

Data Extraction, Synthesis, and Quality Appraisal

Data were entered into Microsoft Excel using a data extraction tool developed a priori [36]. The data extracted are presented in Table S1 in Multimedia Appendix 2.

Each paper was extracted by at least 2 reviewers (ED, JB, RM, and KA), with 2 additional technology experts (JS and RM) contributing to the extraction process and providing technological guidance. Quality appraisal of individual sources was not conducted, as inclusion was based solely on relevance to the review aims, regardless of methodological rigor or risk of bias. This approach aligns with PRISMA-ScR guidance and reflects the broader intent of scoping reviews to provide a comprehensive overview of existing literature while mapping the breadth and nature of available evidence [32].

Results

Overview

The search yielded 12,815 studies for screening. Overall, 36 studies [14,18-24,37-64] met the inclusion criteria and were included in the final review. A summary of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow chart is displayed in Figure 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram illustrating the identification, screening, eligibility assessment, and inclusion of studies published between January 2013 and April 2025 inclusive.

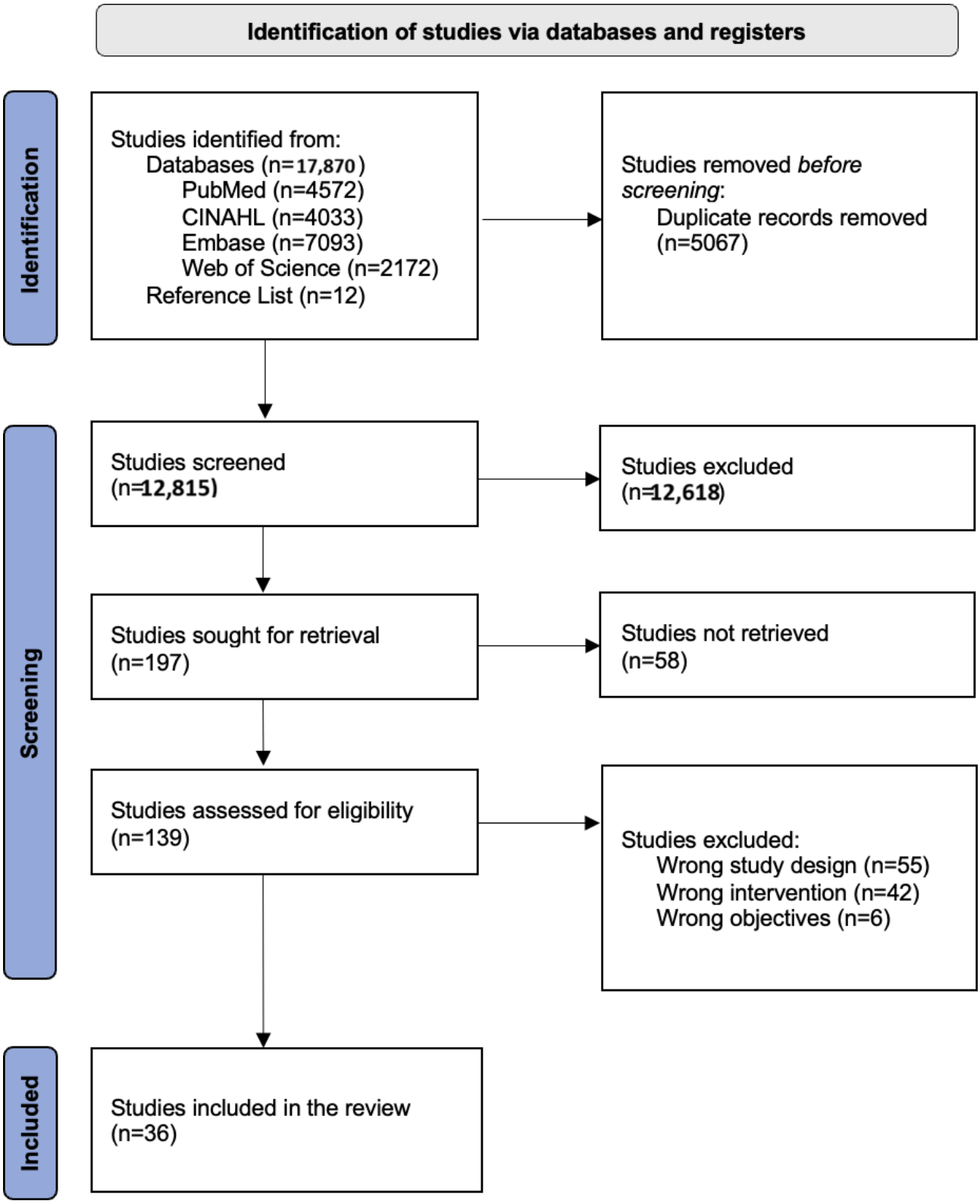

Study Characteristics

A total of 36 heterogenous studies [14,18-24,37-64] were included in the review (Table S1 in Multimedia Appendix 3). Twelve countries from 5 continents were represented, with the majority of studies originating from the United States. Twenty studies [14,18,19,24,37-52] were single-site. The studies addressed a wide range of acute and chronic health conditions, though 3 studies [41,53,54] did not specify the registry's clinical focus. Patient data were extracted from between 50 and 1.4 million patients per study, with 16 studies [22,24,39,41,45,49-54,57-60,64] not clearly reporting the number of unique patients included. The most predominant health care setting type was outpatient care (n=7)

XSL•FO
RenderX

[14,18,19,21,22,55,61], followed by a combination of settings (n=6) [52-54,57,62,63], inpatient care (n=4) [37,38,45,56], surgical (n=3) [47,49,59], and telehealth (n=1) [39]; 15 studies [20,23,24,40-44,46,48,50,51,58,60,64] did not clearly describe the setting type. The predominant study design was descriptive technical reports (n=20) [18-23,37,40,41,45,48,53-56,58,60,62-64], which detailed the development or integration of automated extraction technologies with EMR and registry systems, and pilot validation studies (n=13) [38,39,42-44,46,47,49-52,56,59], which evaluated data extraction systems on a smaller scale to assess accuracy or feasibility.

A variety of EMRs were used, with Epic (Epic Systems Corporation) being the most frequently used system, appearing in 12 studies [18,20,21,23,41,42,45,46,52,53,55,56]. Eight studies [39,40,43,44,50,57-59] did not specify the system used, with an additional 4 studies [60-63] reporting that multiple EMR solutions were used without specifying which systems. Similarly, 12 studies [18,24,38-41,43,57-59,62,63] did not specify the registry or database systems used. The most commonly used database systems included REDCap, structured query language (SQL) server database, and EMR-embedded solutions.

Technological Approaches to Data Extraction and Transfer

Various technologies were used to facilitate the automated extraction and transfer of data from EMRs onto clinical registries (Table S1 in Multimedia Appendix 4), with approaches often tailored to align with data formats and institutional requirements. Figure 2 provides a schematic overview of the main types of technological solutions identified in this review, and how they interface with EMRs and clinical registries. A total of 18 studies [14,19-21,23,24,45-48,53,55-58,60,61,63] extracted data from structured formats, 10 studies [18,22,37,40-42,50,52,62,64] extracted data from mixed formats (structured and unstructured), 6 studies [38,39,43,44,51,59] extracted data from unstructured formats, and the remaining 2 studies [49,54] did not report the format of the data source.

Figure 2. Common technological approaches supporting electronic medical record-to-registry data extraction and transfer. EMR: electronic medical record; REDCap: Research Electronic Data Capture; SSMIX-2: Standardized Structured Medical Information exchange version 2.

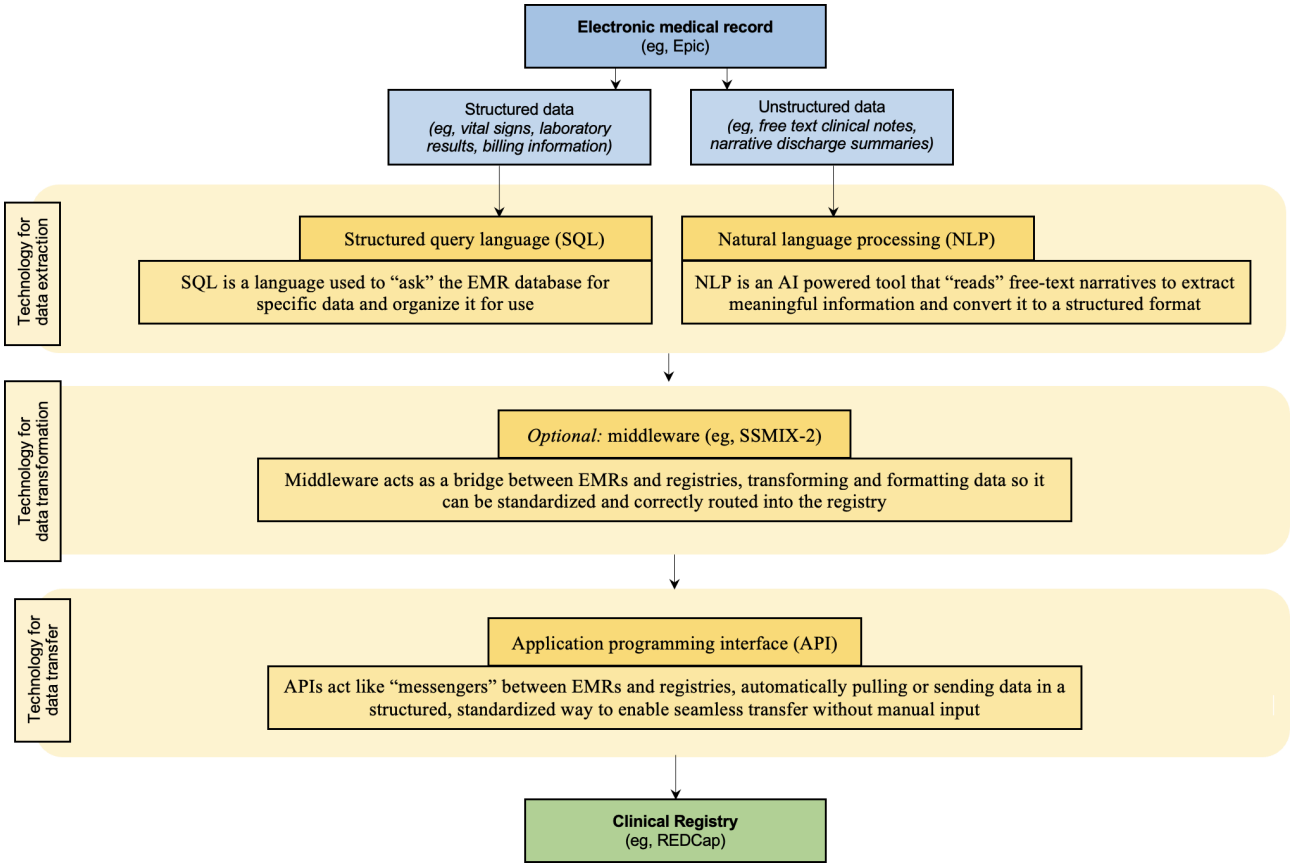

Data extraction from structured fields was most commonly performed using SQL querying methods, as reported explicitly in Abu-Rish Blakeney et al [37] (SQL database extraction, near real-time transfer), Bodagh et al [14] (SQL queries via Cerner Health Information Exchange and National Health Service Data Transfer Services, near real-time), and Kannan et al [21] (custom SQL-based extract, transform, load [ETL] pipeline, weekly extraction). Similar SQL-based approaches were inferred to have been used by Kariuki et al [19] and Nasir et al [55], based on their description of SQL-driven ETL pipelines, although detailed technical implementation was not fully described.

SQL enables automated extraction of clinical data from EMR databases by querying structured fields such as diagnosis codes, laboratory results, and medication lists. These data elements are typically stored in relational tables within the EMR. SQL scripts are written to select specific fields based on predefined criteria (eg, diagnosis and date range) and export them in a standardized format for registry ingestion.

XSL•FO
RenderX

Queries can be scheduled to run at regular intervals (eg, daily or weekly), supporting repeatable, automated data extraction without manual handling. In many cases, the extracted datasets are further processed through ETL pipelines to harmonize field names, apply validation rules, and align data structures with the registry schema, ensuring interoperability. ETL pipelines automate the process of moving data from source systems like EMRs into target systems such as registries by first extracting relevant data, then transforming it to meet predefined standards (eg, renaming fields, standardizing units, and applying validation rules), and finally loading it into the registry database in the correct format.

Built-in EMR functionalities, such as Epic Clarity and embedded data mart solutions, were used for automated routine extraction and transfer by Milinovich and Kattan [41] (Epic Clarity, weekly extraction), Nathan et al [45] (Epic Clarity, daily extraction), Mou et al [43] (Epic Clarity, daily extraction), and Salati et al [49] (built-in EMR functionality, near real-time updates).

For unstructured or mixed-format data sources, natural language processing (NLP) methodologies were applied to automate the extraction of clinical information from free-text documentation. NLP automates the extraction of information from unstructured free-text fields in EMRs by converting narrative data into structured formats suitable for registry ingestion.

Two common NLP techniques are named entity recognition (NER) and pattern matching. NER algorithms identify and classify clinical concepts, such as diagnoses, medications, or procedures, often mapping them to standardized terminologies such as *ICD-10* (*International Statistical Classification of Diseases and Related Health Problems, Tenth Revision*) or SNOMED-CT (Systematized Nomenclature of Medicine - Clinical Terms), enabling consistent structuring of extracted data. Pattern matching, by contrast, uses predefined rules or regular expressions to detect specific text patterns (eg, medication names, dosages, and dates) without requiring machine learning models.

In EMR-to-registry integration, these methods allow clinically relevant information to be automatically detected, extracted, and formatted into predefined registry fields, reducing the need for manual chart review and enabling large-scale, systematic use of free-text clinical documentation. Bacchi et al [38] explicitly described using Python-based NLP libraries to extract stroke metrics, while Heider et al [39] reported use of Apache Unstructured Information Management Architecture for COVID-19 data extraction, with daily updates. Munzone et al [44] used a rule-based NLP algorithm, and Mou et al [42] used open- and closed-source large language models (LLMs) for oncology data extraction. Tavabi et al [59] developed a lightweight NLP pipeline to build orthopedic registries from free-text clinical notes. Wang et al [50] deployed a LLM (ChatGLM) to automate real-world data extraction in Chinese hospital settings.

Several studies extracted both structured and unstructured data, using a combination of technologies [18,22,37,40-42,50,52,62,64]. Two studies [49,54] did not clearly specify the technologies used for data extraction. One study [62] described the use of a centralized repository for aggregation of primary care data, but did not detail the specific technological mechanisms for extraction and transfer.

Middleware and standardized interoperability frameworks were also used in many studies. Nakagawa et al [61] and Sugiyama et al [60] used the Standardized Structured Medical Information eXchange version 2 (SS-MIX2) middleware framework to enable standardized extraction across hospital systems. SS-MIX2 is a standardized middleware framework developed in Japan that enables the extraction, storage, and sharing of structured clinical data from EMRs by converting it into a unified format, supporting interoperability across different health care systems. In terms of interoperability frameworks, Cheng et al [53] used REDCap Clinical Data Interoperability Services (CDIS) with daily data transfer via a FHIR (Fast Healthcare Interoperability Resources)-based application programming interface (API). CDIS enables automated extraction of structured clinical data from EMRs by packaging information into standardized FHIR bundles. FHIR organizes health care data into modular, machine-readable resources, while APIs provide the secure connection that transmits these resources to the target registry.

Technologies used to facilitate the transfer of data onto clinical registries varied across studies, with APIs and built-in EMR functionalities being the most commonly reported. Direct API integrations (eg, FHIR APIs) were reported in Stevens et al [52], Cheng et al [53], and Goel et al [58] to support structured and secure registry population, while Pittman et al [20] leveraged the Informatics for Integrating Biology & the Bedside (i2b2) middleware platform for scheduled weekly updates via an API. In other studies, such as Pan et al [46] and Salati et al [49], the specific transfer protocol was not explicitly stated; however, the use of embedded EMR functions suggests that internal secure batch transfers were used.

The frequency of data transfer varied depending on the system architecture, institutional requirements, and operational considerations, ranging from near real-time updates to intervals as long as 6 months. Near real-time updates were achieved in studies such as Abu-Rish Blakeney et al [37], Bodagh et al [14], and Salati et al [49], particularly where SQL or built-in EMR pipelines were leveraged. Daily transfers were reported by Nathan et al [45], Mou et al [43], and Heider et al [39]. Weekly or monthly batch processes were described in Kannan et al [21] and Nasir et al [55], while longer biannual update cycles were used in Garies et al [62], reflecting resource constraints. One study customized the frequency of data transfer based on an assigned schedule (eg, daily, weekly, or monthly) for each patient [52]. Twelve studies [24,38,43,44,48,50,51,56-60] did not specify the frequency of transfers used, and 17 studies [18,38,39,41,43,44,46-48,50,54,56,57,59,62-64] did not explicitly detail the technology used to transfer data onto the clinical registry.

Evaluating Data Quality in Automated EMR-to-Registry Technologies

Data quality was assessed over three key domains: the completeness of data input, the accuracy of data extracted, and the semantic consistency of data transferred (Table S1 in Multimedia Appendix 5) [65]. Twenty-one studies [19-22,24,37,38,40,43,45-50,55,57,58,60,62,63] evaluated the

XSL•FO
RenderX

completeness of data inputted into EMR systems. Among these, 13 studies [19-21,24,45-48,55,57,58,60,63] extracted exclusively from structured formats, 5 studies [22,37,40,50,62] derived data from mixed formats, 2 studies [38,43] derived data from unstructured formats, and 1 study [49] did not specify the data format. The most common methods of assessing data completeness post input were manual validation, completion rate assessments, and use of mandatory fields within structured data capture tools. Of these studies, only 8 quantified completeness of data input as a percentage, with completeness ranging from 50% to 100% [19,20,24,40,43,48-50]. One study [50] presented completeness between structured and unstructured formats, finding 100% completeness for structured data versus less than 20% for unstructured data. The remaining studies did not report completeness of data input.

Thirty studies [18-21,23,24,37-50,52,54-57,59-63] evaluated the accuracy of technology-driven data extraction. Of these, only 10 studies [19,20,23,40,43,44,49,50,52,56] expressed accuracy as a percentage, all demonstrating performance exceeding 90%, irrespective of data format. One study [43] further assessed accuracy using sensitivity and specificity metrics, while another compared accuracy of data extraction between open- and closed-source large language models [56]. Manual validation emerged as the predominant method for assessing extraction accuracy.

Several studies described alternative methods of assessing data quality that did not align with the predefined metrics summarized in Table S1 in Multimedia Appendix 5. One study applied a qualitative research-grade rating, incorporating both accuracy and completeness, without explicitly reporting percentages [37]. One study used rejection rates of XML files due to validation errors (eg, missing or inaccurate data) as an indicator of data quality [63]. Several studies used mandatory fields within structured data capture forms, implying full completeness of data input [21,22,45,58]. Three studies [20,46,48] reported data quality at the variable or domain level, such as the availability of key laboratory values or the presence of diagnosis codes, rather than providing comprehensive dataset-level metrics. In one study, data quality was assessed through subjective user perception rather than an objective measurement [55].

In addition, a variety of methods were used to maintain semantic consistency post transfer into the clinical registry. Of the 21 studies [21-23,37,38,40-43,45-50,52-58,60-64] that reported a method, the most commonly used approach was the application of standardized terminologies and coding systems. This method was reported in 11 studies [23,41,43,48,52,54,55,58,60-62], and most frequently involved *ICD-10*, Logic Observation Identifiers Names and Codes (LOINC), and SNOMED CT to ensure the consistent interpretation of clinical terms and concepts across systems. Six studies [23,37,40,55,56,63] implemented a shared data dictionary to define data elements and align their definitions across systems. Five studies [50,53,60,61,64] adhered to data exchange standards, notably Health Level 7, FHIR, CDISC, and JOIA, which define the structure and protocols for encoding and transmitting data between systems to preserve syntactic and semantic integrity. Three studies [21,42,43] applied data models to maintain structural and relational consistency across datasets. One study used standardized nomenclature [22], and 5 studies [23,43,55,60,61] used multiple approaches. The remaining 15 studies [14,18-20,24,38,39,44-47,49,51,57,59] did not specify a method for maintaining semantic consistency post transfer.

Approaches to Data Privacy, Security, and Regulatory Compliance

To uphold privacy standards, patient de-identification prior to registry transfer was conducted in 19 studies [18,20,22,23,37,40,43,44,50,54-58,60-64] (Table S1 in Multimedia Appendix 6). Among these, 15 studies [18,20,37,40,43,44,50,54-58,61,63,64] did not report the method used. The remaining studies used various approaches, including pattern-matching algorithms [62], automated server-based anonymization [60], use of an EHR-R-REDCap pipeline [23], and manual redaction [22].

Reported security measures included the use of secure transmission protocols between EMRs and registries, infrastructure and storage protections, and access controls to safeguard registry data. The majority of studies (n=27) [14,18-24,37,38,40-45,47-55,57,59] did not report a secure transfer method. Seven studies [56,58,60-64] reported transferring the data over HTTPS, secure file transfer protocol, or virtual private network, and 2 studies [39,46] reported applying security protocols such as secure sockets layer to encrypt and protect data during transfer. Two further studies [37,53] reported using a secure encrypted channel without specifying the method.

Nine studies [22,37,44,53,57,60,62-64] adopted secure infrastructure and storage measures to protect registry data. These included Health Insurance Portability Accountability Act (HIPAA)–compliant storage (n=3) [22,37,53], centralized secure housing (n=2) [60,62], secure cloud infrastructure (n=2) [44,64], an intranet-based system (n=1) [57], and a secure web-based interface (n=1) [63]. One study stated registry storage was secure, but did not specify the measures taken [61].

Methods used to control and authenticate access to registry data included audit logs (n=4) [46,53,62,63], role-based access control (n=3) [20,53,63], multifactor authentication (MFA) (n=2) [22,44], password and USB key security (n=2) [57,62], token-based access control (n=1) [53], single sign-on (n=1) [22], and permission set access control (n=1) [40]. While some studies implemented multiple measures (n=6) [22,39,53,57,62,63], others restricted access without specifying the methods used (n=8) [37,42,47,52,55,60,61,64].

Various measures were implemented to ensure registry regulatory compliance, including participant consent procedures, ethical oversight, restrictions on data use, and adherence to regulatory standards (Table S1 in Multimedia Appendix 7). Consent procedures were reported in 9 studies [40,42,44,55-57,60-62], including a waiver of consent (n=5) [40,42,55-57], opt-out procedures (n=3) [60-62], and explicit opt-in consent (n=1) [44]. Ethics approval for the registry was obtained in 23 studies [14,19,20,23,24,37,38,40,44,46,47,50,52,53,55-57,59-64], with one study receiving an exception [42]. To restrict registry data

XSL•FO
RenderX

use, institutional review board (IRB) oversight (n=19) [20,21,23,24,37,40,42,46,47,50,53,55-57,59-62,64], research approval processes (n=8) [40,44,53,60-64], and data sharing agreements (n=5) [21,37,44,62,63] were used.

Only a few studies referenced compliance with jurisdiction-specific regulatory standards. Of the studies conducted in the United States, only 4 explicitly reported adherence to the HIPAA [22,37,53,55]. One study based in Japan cited compliance with the Act on the Protection of Personal Information (APPI) [64], while one European study referenced alignment with the General Data Protection Regulation [44]. The remaining 30 studies [14,18-21,23,24,38-43,45-52,54,56-63] did not reference adherence to a jurisdiction-specific regulatory standard, though some demonstrated efforts to ensure data privacy and security, suggesting alignment with regulatory principles.

Implementation Challenges and Frameworks for Clinical Data Technologies

Thirty-four studies identified a range of challenges associated with implementation (Table S1 in Multimedia Appendix 8) [14,18-24,37-40,42-56,58-64]. Each challenge aligned with at least one of the identified thematic categories: (1) data quality, (2) data mapping, standardization and semantic harmonization, (3) technical, infrastructure and resource availability, (4) workflow integration and adoption, (5) interoperability and data integration across systems, (6) privacy, security and governance, (7) registry scope, generalizability and maintenance, and (8) semantic and temporal issues. The most commonly reported challenges included missing or incomplete data [18,19,38,40,47,55], inaccurate data input or extraction [42-44,49-51], significant time and resource demands associated with system development and implementation [23,24,50,54,56,59], poor staff buy-in, uptake, or adoption [14,21,45,47,53], and the inability to capture events occurring outside the institution [18,37,40,44,55]. One study broadly highlighted technical, policy, and governance challenges without providing specific details [53].

Despite the breadth and volume of challenges identified, no studies explicitly reported the use of an implementation science framework to guide technological development and implementation. Instead, 3 studies referenced the use of formal software development methodologies, specifically the Agile Methodology [14,21] and a modified Software Development Life Cycle [37]. An additional 6 studies incorporated elements commonly aligned with implementation science frameworks, including iterative development and co-design, stakeholder involvement and collaborative problem-solving, embedded feedback loops, and rapid-cycle improvement processes [45,49,53,56,63,64].

Mapping Evidence and Contextual Factors for EMR-to-Registry Interoperability

To understand the current landscape of EMR-to-registry integration, evidence on reported technologies, data formats, and implementation challenges was synthesized. Table S1 in Multimedia Appendix 9 presents an evidence gap map summarizing these practices, highlighting the heterogeneity of solutions implemented across studies and the prevalence of implementation challenges, revealing areas that require targeted improvement to support successful integration. Importantly, the map also underscores gaps in reporting technologies used to transfer extracted data from EMRs to clinical registries, a critical component for achieving interoperability. These findings are contextualized in Figure 3, which provides a consolidated overview of factors influencing technological selection, development, and implementation.

Figure 3. Key contextual considerations influencing technology selection for electronic medical record-to-registry interoperability. *ICD-10*: *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision*; OMOP CDM: observational medical outcomes partnership common data model; PREM: patient reported experience measures; PROM: patient reported outcome measures; SNOMED CT: Systematized Nomenclature of Medicine - Clinical Terms.

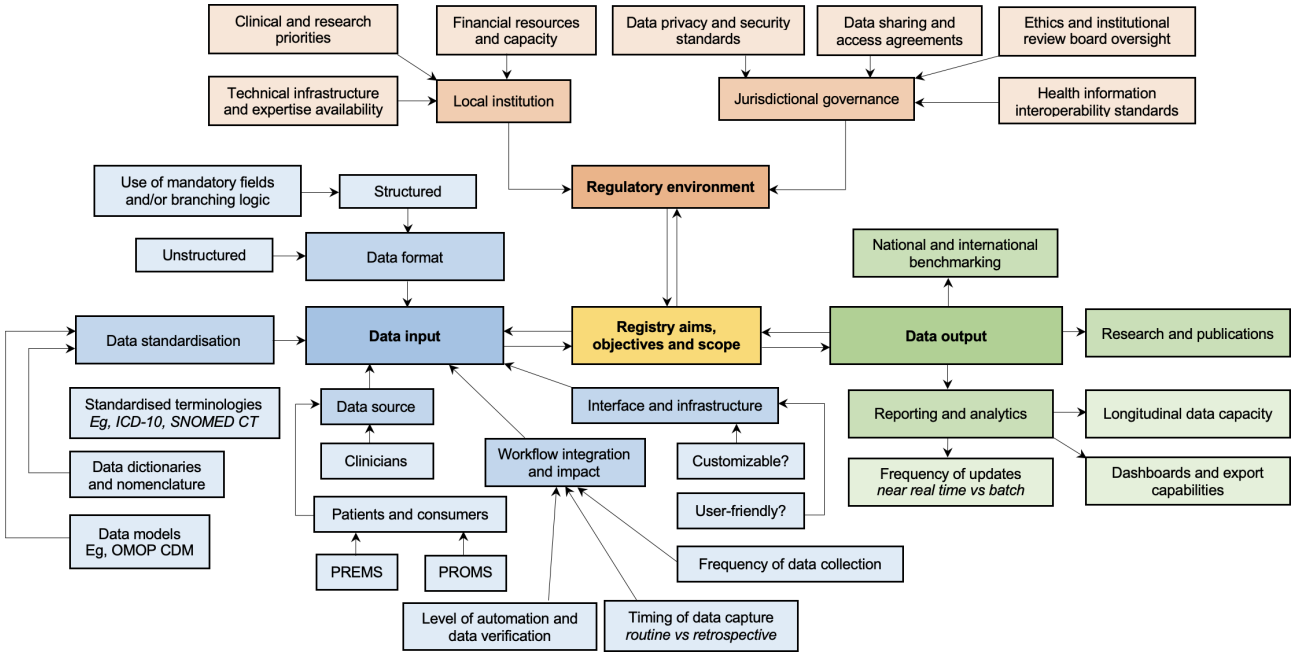

XSL•FO
RenderX

Discussion

Principal Findings

This review advances understanding of interoperability between EMRs and clinical registries by uniquely examining automated and sustainable solutions for data exchange, extending beyond prior work that has largely focused on technologies designed for isolated or study-specific data extraction. Through a synthesis of studies examining existing technological approaches, this review reveals that current approaches are highly variable, context-dependent, and inconsistently evaluated across studies. While existing studies provide important insight into how these processes can be tailored to meet the unique needs and constraints of each health care context, the absence of consistent quality reporting, robust evaluation, and methodological rigor meant no single approach could be recommended. Notably, none of the included studies used a rigorous or comparative study design to evaluate system effectiveness or impact, with most relying on descriptive approaches. This is further compounded by inconsistent and non-standardized reporting on data quality metrics, privacy, security, and regulatory compliance processes, as well as limited application of implementation science frameworks despite reported challenges. Collectively, these limitations restrict meaningful comparisons across studies, and by extension, the technological insights that could inform future implementation.

A novel contribution of this review was the synthesis of the context-specific considerations that influenced technology selection, as informed by reported implementations in the literature and summarized in Figure 3. This synthesis revealed that a registry's purpose and scope not only shaped decisions regarding what data was collected (input) and how it was structured, managed, and used (output), but was itself reciprocally influenced by the broader institutional and jurisdictional regulatory environment. Collectively, these interrelated factors not only influenced the feasibility of implementation, but also had profound implications for data quality, degree of automation and interoperability, scalability, and long-term sustainability. These insights have direct implications for future implementations of interoperable technologies, guiding strategic design and governance decisions that promote sustainability, security, and regulatory compliance, while enabling the ethical and efficient reuse of health data for clinical care, quality improvement, and research.

One key factor influencing both the selection and performance of technologies was the data format, ranging from highly structured fields to unstructured free text. Most studies extracted data from structured formats, benefiting from their standardization, compatibility with coding systems, and designated entry fields that enhance data quality and completeness, especially when supported by mandatory fields and branching logic. However, more than 80% of clinical data exists in unstructured text [12], which remains underused, highlighting the missed potential for richer data capture and integration. Unlike structured fields, free text enables nuanced narrative expressivity in clinical documentation, making their effective use crucial for comprehensive patient records and research insights [66]. While some studies used NLP for unstructured data extraction, wider adoption and continued advancement of artificial intelligence (AI)–driven techniques could further enhance the ability to extract meaningful insights from such data, ultimately improving both clinical decision-making and research potential [67].

A key consideration in EMR research is the widespread concern regarding the reliability and usability of EMR-derived data for clinical research, as noted in previous literature [68-70]. The challenge remains in using an imperfect system, originally designed to support clinical care, financial billing, and insurance claims, for clinical research [8]. Missing data, varying levels of standardization, and a lack of harmonized terminologies can introduce the risk of biases, limit reproducibility, and hinder the integration of EMR-derived data into high-quality research and evidence-based practice. A balancing axiom suggests that data fields actively used, frequently accessed, and regularly reviewed in clinical workflows are more likely to demonstrate improved quality and reliability over time [21]. However, routine clinical use alone does not guarantee data quality and therefore does not negate the need for standardized and transparent evaluation practices. Addressing these foundational issues is essential to improving the credibility, use, and long-term value of EMR-linked registry systems in high-quality clinical research.

At a minimum, studies should report key data quality metrics such as completeness of data input into EMR systems, accuracy of data extracted using technological methods, and semantic consistency of data transferred onto registry systems. The use of existing frameworks which aim to assess data quality [71-73] and semantic consistency [6,74] can help standardize these processes and enhance methodological rigor. Where appropriate, these metrics should also be reported using quantifiable measures (eg, the percentage of missing data) and be accompanied by a clear description of the methods used to calculate them. This ensures that technologies in future research can be evaluated consistently, enabling reliable cross-study comparisons and supporting the development of safe, transparent, and evidence-based recommendations for improving EMR-to-registry interoperability.

Alongside data quality concerns, inconsistent reporting on measures ensuring data privacy, security, and regulatory compliance highlight additional issues surrounding patient confidentiality and the protection of their sensitive information. While these registries enable collaborative data use that enhances both research and clinical outcomes which reciprocally benefit patients, they introduce legal and ethical considerations regarding patient rights and consent, particularly when data is repurposed for research without explicit authorization [75-77]. Common use of opt-out procedures and ethical waivers of consents, particularly when de-identified data is used, has facilitated large-scale data access for research while minimizing direct patient burden [78,79]. However, these approaches raise ethical concerns pertaining to transparency, patient autonomy, and privacy risks, especially when the distinction between de-identified and re-identifiable data has not been clearly articulated [76]. True de-identification is not possible in registries where data pertaining to the same patient is entered

XSL•FO
RenderX

at multiple timepoints. In such cases, pseudonymization is commonly used, replacing patient identifiers with coded keys that enable longitudinal linkage of records while preserving patient confidentiality. The secure management of these re-identification keys is therefore a crucial governance consideration, balancing the need for data use in longitudinal research with the ethical imperative to protect patient privacy and autonomy. Notably, no patient perspectives were captured in any of the studies reviewed, revealing a significant gap regarding patient experiences and their views on the reuse of their data in clinical registries [80]. Similarly, inconsistencies in the use of secure transfer methodologies and poor reporting of jurisdiction-specific regulatory compliance highlight potential risks in integrating and harmonizing data from EMRs to clinical registries. These issues underscore the critical necessity for improved transparency and rigorous adherence to governance frameworks to ensure that data exchange between EMRs and clinical registries remains secure, ethical, and aligned with regulatory standards across diverse health data ecosystems.

Given the sensitivity of health information, technical safeguards must be in place to maintain patient privacy, comply with regulatory requirements, and build trust among stakeholders. Several established and emerging technical solutions are being used to address these concerns. Encryption, both at rest and in transit, is a foundational measure that ensures data cannot be accessed or intercepted without authorization [81]. Role-based access controls and audit logging further enhance security by restricting access to authorized personnel and maintaining traceability of data interactions [82]. In some cases, federated data models and privacy-preserving record linkage allow for data analysis across institutions without centralizing identifiable information, thereby minimizing exposure [83]. The use of secure cloud infrastructure, compliant with standards such as ISO 27001, HIPAA, and General Data Protection Regulation, has also become more prevalent, offering scalable and resilient environments for data storage and processing [81]. Additionally, data governance frameworks play a critical role in defining policies for data access, sharing, and reuse, ensuring that ethical and legal considerations are upheld [84]. Emerging technologies such as homomorphic encryption and differential privacy offer promising avenues for enabling secure computation on sensitive data without compromising individual privacy [85]. While these approaches are still maturing, they represent important innovations in the field of health data protection.

Efforts to improve data quality and regulatory compliance must also minimize clinical burden [86,87]. The variety of implementation challenges reported across studies highlights the complexity of integrating these tools into clinical and registry workflows, with common barriers including data quality concerns, limited staff engagement, high resource demands for system development, and poor alignment with existing institutional infrastructure and workflows. Despite these challenges, there was a paucity of studies explicitly applying an implementation science framework. This theoretical gap likely contributes to the fragmented nature of reporting and hinders the successful adoption, integration, and long-term sustainability of EMR to registry integration. The use of process-driven models and determinant frameworks is a well-established antecedent to successful technology adoption [88-91]. A multidisciplinary, co-design, implementation-informed approach is needed to support alignment with clinical workflows and maximize research use while minimizing disruptions to patient care [92,93].

While automated technologies offer significant advantages over manual data entry, their feasibility is heavily dependent on local financial and technical capacity. Most studies originated from high-income countries, with mature EMR systems and infrastructure. In contrast, low-middle income countries reportedly faced persistent barriers stemming from limited funding and infrastructure capacity [19,63]. The implementation of EMRs and the subsequent development and maintenance of clinical registries is well recognized as a costly and complex process, requiring substantial investment in system design, infrastructure setup, maintenance, staff training, data security, and interoperability [94-96]. Beyond the initial setup, registries also incur ongoing costs that are not typically funded by traditional research project grants, adding an additional element of complexity to their implementation and long-term sustainability across organizations. This presents a major barrier for low-middle income countries, where differences in health system models, funding structures, and technological infrastructure further limit the feasibility of widespread EMR integration. Even in high-income countries such as Australia, the roll-out of interoperable registries is further complicated by the fact that most states have adopted different EMR solutions with varying levels of implementation across hospitals, despite the availability of funding. These disparities threaten to widen global gaps in clinical research capacity, data-driven decision-making, and ultimately patient outcomes.

Limitations of This Study

This review has several limitations. Heterogeneity in study designs and inconsistent reporting of technology characteristics constrained the potential for direct comparison and meta-analysis. Only English studies published between January 2013 and April 2025 were included. Excluding gray literature may result in the underrepresentation of relevant real-world technological implementations, particularly outside high-income countries, and could potentially bias findings toward well-resourced settings. Nonindexed sources may also have been missed as they were not systematically searched outside the selected databases. Technologies designed for single study use were also excluded, which may have provided additional insights. Study authors were not contacted for clarification, and all interpretations relied solely on the information presented in the published manuscripts. Furthermore, many included studies are situated within specific national contexts (eg, SS-MIX2 in Japan), which may limit the broader international generalizability of the findings.

Conclusions

This review provides insight into the diverse strategies that facilitate interoperability between EMRs and clinical registries. While technological advancements have negated the need for manual data extraction and transfer onto clinical registries, challenges remain in ensuring access, consistency, security, and sustainability across diverse health care settings. Given the

XSL•FO
RenderX

ethical, legal, and regulatory complexities of reusing clinical data for research, future initiatives must operate within robust governance frameworks to safeguard data security, protect patient privacy, and ensure compliance with both institutional and jurisdictional policies while promoting transparency, accountability, and equitable access to research opportunities. Implementation efforts should be tailored to local contexts, guided by implementation science frameworks and informed by meaningful engagement with end users to support adoption. Overcoming these barriers, alongside continued innovation in interoperability solutions, will be essential to maximizing the potential and impact of EMR-linked registries to drive high-quality research, strengthen clinical decision-making, and ultimately improve patient outcomes.

Acknowledgments

We would like to acknowledge Justin Clarke, Senior Research Information Specialist from the Faculty of Health Sciences and Medicine at Bond University, Queensland, Australia, for his valuable contribution in refining the search strategy and applying the Systematic Review Accelerator Polyglot Search Translator automation tool, which was used to convert the final search strategy into appropriate formats for four databases.

Funding

This review was supported by the Walker Pediatric Research Fellowship awarded by the Pediatric Surgical Department within Children's Health Queensland, Hospital and Health Service. This support was presented in the form of a higher degree by research stipend, with no involvement in the study design, data collection and analysis, decision to publish or preparation of the manuscript. The authors have no conflicts of interest to disclose.

Authors' Contributions

EH conceptualized and designed the study and search strategy; led screening, data collection, synthesis, and interpretation of findings; drafted the initial manuscript, reviewed and revised the manuscript; and approved the final manuscript as submitted. JB contributed to the design of the search strategy, conducted study screening, data collection, analysis, synthesis, and interpretation; contributed to the initial manuscript, critically reviewed and revised the manuscript; and approved the final manuscript as submitted. JS and RM contributed to the technological understanding, synthesis, and interpretation of findings; critically reviewed and revised the manuscript; and approved the final manuscript as submitted. RMF and KAD conducted screening and data extraction, critically reviewed the manuscript; and approved the final manuscript as submitted. AD, KP, CMB, and RK contributed to the conceptualization, synthesis, and interpretation of the study; critically reviewed and revised the manuscript and approved the final manuscript as submitted. BG contributed to the conceptualization and design of the study, supervised screening, data collection, analysis, and interpretation, critically reviewed and revised the manuscript, and approved the final manuscript as submitted.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search terms and strategy.
[DOCX File, 17 KB - jmir_v28i1e82380_app1.docx]

Multimedia Appendix 2
Data extraction categories and description.
[DOCX File, 15 KB - jmir_v28i1e82380_app2.docx]

Multimedia Appendix 3
Participant and study characteristics of included studies.
[DOCX File, 70 KB - jmir_v28i1e82380_app3.docx]

Multimedia Appendix 4
Summary of electronic medical record and registry infrastructure, technologies, and interoperability approaches.
[DOCX File, 71 KB - jmir_v28i1e82380_app4.docx]

Multimedia Appendix 5
Summary of data quality measures following data input, extraction, and transfer.
[DOCX File, 71 KB - jmir_v28i1e82380_app5.docx]

XSL•FO
RenderX

Multimedia Appendix 6
Summary of data privacy and security measures.
[DOCX File, 71 KB - jmir_v28i1e82380_app6.docx]

Multimedia Appendix 7
Summary of measures for registry regulatory compliance.
[DOCX File, 70 KB - jmir_v28i1e82380_app7.docx]

Multimedia Appendix 8
Summary of implementation challenges identified in establishing electronic medical record-to-registry interoperability.
[DOCX File, 109 KB - jmir_v28i1e82380_app8.docx]

Multimedia Appendix 9
Evidence gap map of technologies, data formats, and implementation challenges in electronic medical record-to-registry integration.
[DOCX File, 88 KB - jmir_v28i1e82380_app9.docx]

Checklist 1
PRISMA-ScR checklist.
[PDF File, 85 KB - jmir_v28i1e82380_app10.pdf]

References

1. Yeung AWK, Torkamani A, Butte AJ, et al. The promise of digital healthcare technologies. Front Public Health 2023;11:1196596. [doi: 10.3389/fpubh.2023.1196596] [Medline: 37822534]
2. Alzu'bi AA, Watzlaf VJM, Sheridan P. Electronic health record (EHR) abstraction. Perspect Health Inf Manag 2021;18(Spring):1g. [Medline: 34035788]
3. Klaiman T, Pracilio V, Kimberly L, Cecil K, Legnini M. Leveraging effective clinical registries to advance medical care quality and transparency. Popul Health Manag 2014 Apr;17(2):127-133. [doi: 10.1089/pop.2013.0021] [Medline: 24152057]
4. Hoque DME, Kumari V, Hoque M, Ruseckaite R, Romero L, Evans SM. Impact of clinical registries on quality of patient care and clinical outcomes: a systematic review. PLoS ONE 2017;12(9):e0183667. [doi: 10.1371/journal.pone.0183667] [Medline: 28886607]
5. Olaronke I, Soriyan H, Gambo I, Olaleke J. Interoperability in healthcare: benefits, challenges and resolutions. Int J Innov Appl Stud 2013 Apr;3:2028-9324 [FREE Full text]
6. de Mello BH, Rigo SJ, da Costa CA, et al. Semantic interoperability in health records standards: a systematic literature review. Health Technol (Berl) 2022;12(2):255-272. [doi: 10.1007/s12553-022-00639-w] [Medline: 35103230]
7. Manion FJ, Harris MR, Buyuktur AG, Clark PM, An LC, Hanauer DA. Leveraging EHR data for outcomes and comparative effectiveness research in oncology. Curr Oncol Rep 2012 Dec;14(6):494-501. [doi: 10.1007/s11912-012-0272-6] [Medline: 22948276]
8. Nordo AH, Levaux HP, Becnel LB, et al. Use of EHRs data for clinical research: historical progress and current applications. Learn Health Syst 2019 Jan;3(1):e10076. [doi: 10.1002/lrh2.10076] [Medline: 31245598]
9. Pop B, Fetica B, Blaga ML, et al. The role of medical registries, potential applications and limitations. Med Pharm Rep 2019 Jan;92(1):7-14. [doi: 10.15386/cjmed-1015] [Medline: 30957080]
10. Filkins BL, Kim JY, Roberts B, et al. Privacy and security in the era of digital health: what should translational researchers know and do about it? Am J Transl Res 2016;8(3):1560-1580. [Medline: 27186282]
11. Cowie MR, Blomster JI, Curtis LH, et al. Electronic health records to facilitate clinical research. Clin Res Cardiol 2017 Jan;106(1):1-9. [doi: 10.1007/s00392-016-1025-6] [Medline: 27557678]
12. Negro-Calduch E, Azzopardi-Muscat N, Krishnamurthy RS, Novillo-Ortiz D. Technological progress in electronic health record system optimization: systematic review of systematic literature reviews. Int J Med Inform 2021 Aug;152:104507. [doi: 10.1016/j.ijmedinf.2021.104507] [Medline: 34049051]
13. Honavar SG. Electronic medical records - the good, the bad and the ugly. Indian J Ophthalmol 2020 Mar;68(3):417-418. [doi: 10.4103/ijo.IJO_278_20] [Medline: 32056991]
14. Bodagh N, Archbold RA, Weerackody R, et al. Feasibility of real-time capture of routine clinical data in the electronic health record: a hospital-based, observational service-evaluation study. BMJ Open 2018 Mar 8;8(3):e019790. [doi: 10.1136/bmjopen-2017-019790] [Medline: 29523565]
15. D'Agnolo HM, Kievit W, Andrade RJ, Karlsen TH, Wedemeyer H, Drenth JP. Creating an effective clinical registry for rare diseases. United European Gastroenterol J 2016 Jun;4(3):333-338. [doi: 10.1177/2050640615618042] [Medline: 27403298]
16. Solomon DJ, Henry RC, Hogan JG, Van Amburg GH, Taylor J. Evaluation and implementation of public health registries. Public Health Rep 1991;106(2):142-150. [Medline: 1902306]

XSL•FO
RenderX

17. Dreyer NA, Garner S. Registries for robust evidence. JAMA 2009 Aug 19;302(7):790-791. [doi: 10.1001/jama.2009.1092] [Medline: 19690313]
18. Chen AM, Kupelian PA, Wang PC, Steinberg ML. Development of a radiation oncology-specific prospective data registry for research and quality improvement: a clinical workflow-based solution. JCO Clin Cancer Inform 2018 Dec;2:1-9. [doi: 10.1200/CCI.17.00036] [Medline: 30652556]
19. Kariuki JM, Manders EJ, Richards J, et al. Automating indicator data reporting from health facility EMR to a national aggregate data system in Kenya: an interoperability field-test using OpenMRS and DHIS2. Online J Public Health Inform 2016;8(2):e188. [doi: 10.5210/ojphi.v8i2.6722] [Medline: 28149444]
20. Pittman CA, Miranpuri AS. Neurosurgery clinical registry data collection utilizing informatics for integrating biology and the bedside and electronic health records at the University of Rochester. Neurosurg Focus 2015 Dec;39(6):26621414. [doi: 10.3171/2015.9.FOCUS15382] [Medline: 26621414]
21. Kannan V, Fish JS, Mutz JM, et al. Rapid development of specialty population registries and quality measures from electronic health record data*. An agile framework. Methods Inf Med 2017 Jun 14;56(99):e74-e83. [doi: 10.3414/ME16-02-0031] [Medline: 28930362]
22. Kapoor R, Sleeman WC 4th, Nalluri JJ, et al. Automated data abstraction for quality surveillance and outcome assessment in radiation oncology. J Appl Clin Med Phys 2021 Jul;22(7):177-187. [doi: 10.1002/acm2.13308] [Medline: 34101349]
23. Shalhout SZ, Saqlain F, Wright K, Akinyemi O, Miller DM. Generalizable EHR-R-REDCap pipeline for a national multi-institutional rare tumor patient registry. JAMIA Open 2022 Apr;5(1):ooab118. [doi: 10.1093/jamiaopen/ooab118] [Medline: 35156001]
24. Williams A, Goedicke W, Tissera KA, Mankarious LA. Leveraging existing tools in electronic health record systems to automate clinical registry compilation. Otolaryngol Head Neck Surg 2020 Mar;162(3):408-409. [doi: 10.1177/0194599820901713] [Medline: 31961772]
25. Lee M, Kim K, Shin Y, Lee Y, Kim TJ. Advancements in electronic medical records for clinical trials: enhancing data management and research efficiency. Cancers (Basel) 2025 May 2;17(9):1552. [doi: 10.3390/cancers17091552] [Medline: 40361478]
26. Mueller C, Herrmann P, Cichos S, et al. Automated electronic health record to electronic data capture transfer in clinical studies in the German health care system: feasibility study and gap analysis. J Med Internet Res 2023 Aug 4;25:e47958. [doi: 10.2196/47958] [Medline: 37540555]
27. Holmes JH, Beinlich J, Boland MR, et al. Why Is the electronic health record so challenging for research and clinical care? Methods Inf Med 2021 May;60(1-02):32-48. [doi: 10.1055/s-0041-1731784] [Medline: 34282602]
28. van Velthoven MH, Mastellos N, Majeed A, O'Donoghue J, Car J. Feasibility of extracting data from electronic medical records for research: an international comparative study. BMC Med Inform Decis Mak 2016 Jul 13;16(1):90. [doi: 10.1186/s12911-016-0332-1] [Medline: 27411943]
29. Trinkley KE, Maw AM, Torres CH, Huebschmann AG, Glasgow RE. Applying implementation science to advance electronic health record-driven learning health systems: case studies, challenges, and recommendations. J Med Internet Res 2024 Oct 7;26:e55472. [doi: 10.2196/55472] [Medline: 39374069]
30. Bauer MS, Kirchner J. Implementation science: what is it and why should I care? Psychiatry Res 2020 Jan;283:112376. [doi: 10.1016/j.psychres.2019.04.025] [Medline: 31036287]
31. Sucharew H, Macaluso M. Progress notes: methods for research evidence synthesis: the scoping review approach. J Hosp Med 2019 Jul 1;14(7):416-418. [doi: 10.12788/jhm.3248] [Medline: 31251164]
32. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
33. Dulay E. Current capacity of interoperability between clinical registries and patient electronic medical records: a scoping review. Open Science Framework (OSF). Preprint posted online on Oct 23, 2023 URL: https://doi.org/10.17605/OSF.IO/83W76 [accessed 2026-05-20]
34. Clark JM, Sanders S, Carter M, et al. Improving the translation of search strategies using the Polyglot Search Translator: a randomized controlled trial. J Med Libr Assoc 2020 Apr;108(2):195-207. [doi: 10.5195/jmla.2020.834] [Medline: 32256231]
35. Veritas Health Innovation. Covidence - better systematic review management. Covidence. URL: https://www.covidence.org/ [accessed 2026-05-20]
36. Microsoft excel. Microsoft 365. URL: https://office.microsoft.com/excel [accessed 2026-05-20]
37. Abu-Rish Blakeney E, Wolpin S, Lavallee DC, Dardas T, Cheng R, Zierler B. Developing and implementing a heart failure data mart for research and quality improvement. Inform Health Soc Care 2019;44(2):164-175. [doi: 10.1080/17538157.2018.1455202] [Medline: 29672242]
38. Bacchi S, Gluck S, Koblar S, Jannes J, Kleinig T. Automated information extraction from free-text medical documents for stroke key performance indicators: a pilot study. Intern Med J 2022 Feb;52(2):315-317. [doi: 10.1111/imj.15678] [Medline: 35187820]

XSL•FO
RenderX

39. Heider PM, Pipaliya RM, Meystre SM. A natural language processing tool offering data extraction for COVID-19 related information (DECOVRI). Stud Health Technol Inform 2022 Jun 6;290:1062-1063. [doi: 10.3233/SHTI220268] [Medline: 35673206]
40. Li N, Zhu Q, Dang Y, et al. Development and implementation of a dynamically updated big data intelligence platform using electronic medical records for secondary hypertension. Rev Cardiovasc Med 2024 Mar;25(3):104. [doi: 10.31083/j.rcm2503104] [Medline: 39076957]
41. Milinovich A, Kattan MW. Extracting and utilizing electronic health data from Epic for research. Ann Transl Med 2018 Feb;6(3):42. [doi: 10.21037/atm.2018.01.13] [Medline: 29610734]
42. Mou Z, Sitapati AM, Ramachandran M, Doucet JJ, Liepert AE. Development and implementation of an automated electronic health record-linked registry for emergency general surgery. J Trauma Acute Care Surg 2022 Aug 1;93(2):273-279. [doi: 10.1097/TA.0000000000003582] [Medline: 35195091]
43. Mou Y, Lehmkuhl J, Sauerbrunn N, et al. Improving the quality of unstructured cancer data using large language models: a German oncological case study. Stud Health Technol Inform 2024 Aug 22;316:685-689. [doi: 10.3233/SHTI240507] [Medline: 39176835]
44. Munzone E, Marra A, Comotto F, et al. Development and validation of a natural language processing algorithm for extracting clinical and pathological features of breast cancer from pathology reports. JCO Clin Cancer Inform 2024 Aug;8:e2400034. [doi: 10.1200/CCI.24.00034] [Medline: 39137368]
45. Nathan JK, Foley J, Hoang T, et al. The stroke navigator: meaningful use of the electronic health record to efficiently report inpatient stroke care quality. J Am Med Inform Assoc 2018 Nov 1;25(11):1534-1539. [doi: 10.1093/jamia/ocy102] [Medline: 30124956]
46. Pan HY, Shaitelman SF, Perkins GH, Schlembach PJ, Woodward WA, Smith BD. Implementing a real-time electronic data capture system to improve clinical documentation in radiation oncology. J Am Coll Radiol 2016 Apr;13(4):401-407. [doi: 10.1016/j.jacr.2015.09.036] [Medline: 26681164]
47. Rayman S, Benvenisti H, Westrich G, Schtrechman G, Nissan A, Segev L. Colorectal surgery surveillance: a novel method for composing an automated real-time prospective registry. Isr Med Assoc J 2021 Apr;23(4):239-244. [Medline: 33899357]
48. Rubio-Mayo P, Ojeda-Thies C, Jiménez-Cerezo MJ, Garcia-Barrio N, Cruz-Bermúdez JL, Pedrera-Jiménez M. HCE2RNFC: an efficient methodology for reusing the EHR in the Spanish National Hip Fracture Registry. Stud Health Technol Inform 2024 Aug 22;316:1422-1426. [doi: 10.3233/SHTI240679] [Medline: 39176648]
49. Salati M, Pompili C, Refai M, Xiumè F, Sabbatini A, Brunelli A. Real-time database drawn from an electronic health record for a thoracic surgery unit: high-quality clinical data saving time and human resources. Eur J Cardiothorac Surg 2014 Jun;45(6):1017-1019. [doi: 10.1093/ejcts/ezt577] [Medline: 24394554]
50. Wang B, Lai J, Cao H, et al. Enhancing the interoperability and transparency of real-world data extraction in clinical research: evaluating the feasibility and impact of a ChatGLM implementation in Chinese hospital settings. Eur Heart J Digit Health 2024 Nov;5(6):712-724. [doi: 10.1093/ehjdh/ztae066] [Medline: 39563908]
51. Wulff A, Mast M, Hassler M, Montag S, Marschollek M, Jack T. Designing an openEHR-Based pipeline for extracting and standardizing unstructured clinical data using natural language processing. Methods Inf Med 2020 Dec;59(S 02):e64-e78. [doi: 10.1055/s-0040-1716403] [Medline: 33058101]
52. Stevens A, Karki S, Shivers E, et al. SmartChart Suite: a Fast Healthcare Interoperability Resources-based framework for longitudinal syphilis surveillance using structured and unstructured data. JAMIA Open 2025 Feb;8(1):ooae145. [doi: 10.1093/jamiaopen/ooae145] [Medline: 39735787]
53. Cheng AC, Duda SN, Taylor R, et al. REDCap on FHIR: clinical data interoperability services. J Biomed Inform 2021 Sep;121:103871. [doi: 10.1016/j.jbi.2021.103871] [Medline: 34298155]
54. González L, Pérez-Rey D, Alonso E, et al. Building an I2B2-based population repository for clinical research. Stud Health Technol Inform 2020 Jun 16;270:78-82. [doi: 10.3233/SHTI200126] [Medline: 32570350]
55. Nasir K, Gullapelli R, Nicolas JC, et al. Houston Methodist cardiovascular learning health system (CVD-LHS) registry: methods for development and implementation of an automated electronic medical record-based registry using an informatics framework approach. Am J Prev Cardiol 2024 Jun;18:100678. [doi: 10.1016/j.ajpc.2024.100678] [Medline: 38756692]
56. Valencia Morales DJ, Bansal V, Heavner SF, et al. Validation of automated data abstraction for SCCM discovery VIRUS COVID-19 registry: practical EHR export pathways (VIRUS-PEEP). Front Med (Lausanne) 2023;10:1089087. [doi: 10.3389/fmed.2023.1089087] [Medline: 37859860]
57. Dong Y, Fang K, Wang X, et al. The network of Shanghai Stroke Service System (4S): a public health-care web-based database using automatic extraction of electronic medical records. Int J Stroke 2018 Jul;13(5):539-544. [doi: 10.1177/1747493018765492] [Medline: 29561219]
58. Goel AK, Campbell WS, Moldwin R. Structured data capture for oncology. JCO Clin Cancer Inform 2021 Feb;5:194-201. [doi: 10.1200/CCI.20.00103] [Medline: 33591796]
59. Tavabi N, Pruneski J, Golchin S, et al. Building large-scale registries from unstructured clinical notes using a low-resource natural language processing pipeline. Artif Intell Med 2024 May;151:102847. [doi: 10.1016/j.artmed.2024.102847] [Medline: 38658131]

XSL•FO
RenderX

60. Sugiyama T, Miyo K, Tsujimoto T, et al. Design of and rationale for the Japan Diabetes compREhensive database project based on an Advanced electronic Medical record System (J-DREAMS). Diabetol Int 2017 Nov;8(4):375-382. [doi: 10.1007/s13340-017-0326-y] [Medline: 30603343]
61. Nakagawa N, Sofue T, Kanda E, et al. J-CKD-DB: a nationwide multicentre electronic health record-based chronic kidney disease database in Japan. Sci Rep 2020 Apr 30;10(1):7351. [doi: 10.1038/s41598-020-64123-z] [Medline: 32355258]
62. Garies S, Cummings M, Forst B, et al. Achieving quality primary care data: a description of the Canadian Primary Care Sentinel Surveillance Network data capture, extraction, and processing in Alberta. Int J Popul Data Sci 2019 Jul 29;4(2):1132. [doi: 10.23889/ijpds.v4i2.1132] [Medline: 34095540]
63. Dalhatu I, Aniekwe C, Bashorun A, et al. From paper files to web-based application for data-driven monitoring of HIV programs: Nigeria's journey to a national data repository for decision-making and patient care. Methods Inf Med 2023 Sep;62(3-04):130-139. [doi: 10.1055/s-0043-1768711] [Medline: 37247622]
64. Miyake M, Akiyama M, Kashiwagi K, Sakamoto T, Oshika T. Japan Ocular Imaging Registry: a national ophthalmology real-world database. Jpn J Ophthalmol 2022 Nov;66(6):499-503. [doi: 10.1007/s10384-022-00941-0] [Medline: 36138192]
65. AbuHalimeh A. Improving data quality in clinical research informatics tools. Front Big Data 2022;5:871897. [doi: 10.3389/fdata.2022.871897] [Medline: 35574572]
66. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc 2011;18(2):181-186. [doi: 10.1136/jamia.2010.007237] [Medline: 21233086]
67. Rubinger L, Gazendam A, Ekhtiari S, Bhandari M. Machine learning and artificial intelligence in research and healthcare. Injury 2023 May;54 Suppl 3:S69-S73. [doi: 10.1016/j.injury.2022.01.046] [Medline: 35135685]
68. Syed R, Eden R, Makasi T, et al. Digital health data quality issues: systematic review. J Med Internet Res 2023 Mar 31;25:e42615. [doi: 10.2196/42615] [Medline: 37000497]
69. Garza M, Myneni S, Nordo A, et al. eSource for standardized health information exchange in clinical research: a systematic review. Stud Health Technol Inform 2019;257(115-24):115-124. [Medline: 30741183]
70. Gianfrancesco MA, Goldstein ND. A narrative review on the validity of electronic health record-based research in epidemiology. BMC Med Res Methodol 2021 Oct 27;21(1):234. [doi: 10.1186/s12874-021-01416-5] [Medline: 34706667]
71. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 1;20(1):144-151. [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
72. Reimer AP, Milinovich A, Madigan EA. Data quality assessment framework to assess electronic medical record data for use in research. Int J Med Inform 2016 Jun;90:40-47. [doi: 10.1016/j.ijmedinf.2016.03.006] [Medline: 27103196]
73. Lewis AE, Weiskopf N, Abrams ZB, et al. Electronic health record data quality assessment and tools: a systematic review. J Am Med Inform Assoc 2023 Sep 25;30(10):1730-1740. [doi: 10.1093/jamia/ocad120] [Medline: 37390812]
74. Facile R, Chronaki C, van Reusel P, Kush R. Standards in sync: five principles to achieve semantic interoperability for TRUE research for healthcare. Front Digit Health 2025;7:1567624. [doi: 10.3389/fdgth.2025.1567624] [Medline: 40538570]
75. Kotwal S, Webster AC, Cass A, Gallagher M. A review of linked health data in Australian nephrology. Nephrology (Carlton) 2016 Jun;21(6):457-466. [doi: 10.1111/nep.12721] [Medline: 26748448]
76. Jabour AM. Putting patients at the center of health information exchange design: An exploration of patient preferences for data sharing. Health Informatics J 2024;30(3):14604582241277029. [doi: 10.1177/14604582241277029] [Medline: 39142341]
77. Tertulino R, Antunes N, Morais H. Privacy in electronic health records: a systematic mapping study. J Public Health (Berl) 2024 Mar;32(3):435-454. [doi: 10.1007/s10389-022-01795-z]
78. de Man Y, Wieland-Jorna Y, Torensma B, et al. Opt-in and opt-out consent procedures for the reuse of routinely recorded health data in scientific research and their consequences for consent rate and consent bias: systematic review. J Med Internet Res 2023 Feb 28;25:e42131. [doi: 10.2196/42131] [Medline: 36853745]
79. Md Emdadul Hoque D, Ruseckaite R, Lorgelly P, McNeil JJ, Evans SM. Cross-sectional study of characteristics of clinical registries in Australia: a resource for clinicians and policy makers. Int J Qual Health Care 2018 Apr 1;30(3):192-199. [doi: 10.1093/intqhc/mzx196] [Medline: 29385457]
80. Fesl S, Lang C, Schmitt J, et al. Factors influencing patients' willingness to share their digital health data for primary and secondary use: a theory- and evidence-based overview of reviews. Digit Health 2025;11:20552076251340254. [doi: 10.1177/20552076251340254] [Medline: 40599675]
81. Mehrtak M, SeyedAlinaghi S, MohsseniPour M, et al. Security challenges and solutions using healthcare cloud computing. J Med Life 2021;14(4):448-461. [doi: 10.25122/jml-2021-0100] [Medline: 34621367]
82. de Carvalho Junior MA, Bandiera-Paiva P. Health information system role-based access control current security trends and challenges. J Healthc Eng 2018;2018:6510249. [doi: 10.1155/2018/6510249] [Medline: 29670743]
83. Baumgartner M, Kreiner K, Lauschensky A, et al. Health data space nodes for privacy-preserving linkage of medical data to support collaborative secondary analyses. Front Med (Lausanne) 2024;11:1301660. [doi: 10.3389/fmed.2024.1301660] [Medline: 38660421]
84. Abraham R, Schneider J, vom Brocke J. Data governance: a conceptual framework, structured review, and research agenda. Int J Inf Manage 2019 Dec;49:424-438. [doi: 10.1016/j.ijinfomgt.2019.07.008]

XSL•FO
RenderX

85. Shin H, Ryu K, Kim JY, Lee S. Application of privacy protection technology to healthcare big data. Digit Health 2024;10:20552076241282242. [doi: 10.1177/20552076241282242] [Medline: 39502481]
86. Alumran A, Aljuraifani SA, Almousa ZA, et al. The influence of electronic health record use on healthcare providers burnout. Informatics in Medicine Unlocked 2024;50:101588. [doi: 10.1016/j.imu.2024.101588]
87. Murad MH, Vaa Stelling BE, West CP, et al. Measuring documentation burden in healthcare. J Gen Intern Med 2024 Nov;39(14):2837-2848. [doi: 10.1007/s11606-024-08956-8] [Medline: 39073484]
88. Delaforce A, Li J, Niven P, et al. Using the CFIR-ERIC approach to enhance the uptake of a digital fall prevention platform: a quasi-experimental pre-post implementation study. 2025. [doi: 10.21203/rs.3.rs-6372140/v1]
89. Rouleau G, Wu K, Ramamoorthi K, et al. Mapping theories, models, and frameworks to evaluate digital health interventions: scoping review. J Med Internet Res 2024 Feb 5;26:e51098. [doi: 10.2196/51098] [Medline: 38315515]
90. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
91. Li J, Maddock E, Hosking M, et al. Identifying and optimizing factors influencing the implementation of a fast healthcare interoperability resources accelerator: qualitative study using the consolidated framework for implementation research-expert recommendations for implementing change approach. JMIR Med Inform 2025 May 27;13:e66421. [doi: 10.2196/66421] [Medline: 40424614]
92. Griffin BR, Trenoweth K, Dimanopoulos TA, et al. Co-design of a paediatric post-trauma electronic psychosocial screen. J Pediatr Nurs 2024;76:52-60. [doi: 10.1016/j.pedn.2024.02.004] [Medline: 38359545]
93. Dimanopoulos MTA, Trenoweth MK, De Young AC, Kimble R, Griffin BR. The acceptability, feasibility and adoption of a co-designed electronic post-injury psychosocial screening tool for carers of children admitted to hospital following injury. J Pediatr Nurs 2025;81:155-164. [doi: 10.1016/j.pedn.2025.01.027] [Medline: 39908860]
94. Modi S, Feldman SS. The value of electronic health records since the health information technology for economic and clinical health act: systematic review. JMIR Med Inform 2022 Sep 27;10(9):e37283. [doi: 10.2196/37283] [Medline: 36166286]
95. Aguirre RR, Suarez O, Fuentes M, Sanchez-Gonzalez MA. Electronic health record implementation: a review of resources and tools. Cureus 2019 Sep 13;11(9):e5649. [doi: 10.7759/cureus.5649] [Medline: 31700751]
96. Al Ani M, Garas G, Hollingshead J, Cheetham D, Athanasiou T, Patel V. Which electronic health record system should we use? A systematic review. Med Princ Pract 2022;31(4):342-351. [doi: 10.1159/000525135] [Medline: 35584616]

Abbreviations

AI : artificial intelligence
API : application programming interface
APPI: Act on the Protection of Personal Information
CDIS: Clinical Data Interoperability Services
CDISC: Clinical Data Interchange Standards Consortium
CDISC: Clinical Data Interchange Standards Consortium
EHR: electronic health record
EMR: electronic medical record
ETL: extract, transform, load
FHIR: Fast Healthcare Interoperability Resources
HIPAA: Health Insurance Portability and Accountability Act
i2b2: Informatics for Integrating Biology & the Bedside
ICD-10: *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision*
IRB: institutional review board
LLM: large language model
LOINC: Logic Observation Identifiers Names and Codes
MFA: multifactor authentication
NER: named entity recognition
NLP: natural language processing
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items of Systematic Reviews and Meta-Analyses extension for Scoping Reviews
REDCap: Research Electronic Data Capture
SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms
SQL: structured query language
SS-MIX2: Standardized Structured Medical Information exchange version 2

XSL•FO
RenderX

Edited by S Brini; submitted 14.Aug.2025; peer-reviewed by A Shivanna, G Costa; revised version received 14.Jan.2026; accepted 15.Jan.2026; published 25.May.2026.

Please cite as:
Haynes E, Brannigan J, Suna J, Malseed R, Delaforce A, Mulvenney-Fenner R, Alog-Daroya K, Plummer K, McBride C, Kimble R, Griffin B
Exploring Technological Solutions for Interoperability Between Patient Electronic Medical Records and Clinical Registries: Scoping Review
J Med Internet Res 2026;28:e82380
URL: https://www.jmir.org/2026/1/e82380
doi:10.2196/82380

© Erika Haynes, James Brannigan, Jessica Suna, Reid Malseed, Alana Delaforce, Rachel Mulvenney-Fenner, Katherine Alog-Daroya, Karin Plummer, Craig McBride, Roy Kimble, Bronwyn Griffin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Virtual Reality Interventions for Stress Reduction in the General Population: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Hannah Strauch[1], MSc; Isabel Schuil[1], MSc; Stefan Simm[1], Prof Dr; Jens Grubert[2], Prof Dr; Snehanjali Kalamkar[2], BE, MSc; Karin Meissner[1], Prof Dr Med

[1]Faculty of Applied Natural Sciences and Health, Coburg University of Applied Sciences and Arts, Friedrich-Streib-Straße 2, Coburg, Germany
[2]Faculty of Electrical Engineering and Computer Sciences, Coburg University of Applied Sciences and Arts, Coburg, Germany

Corresponding Author:
Karin Meissner, Prof Dr Med
Faculty of Applied Natural Sciences and Health, Coburg University of Applied Sciences and Arts, Friedrich-Streib-Straße 2, Coburg, Germany

Abstract

Background: Increasing mental demands across multiple life domains underscore the importance of effective individual stress management to mitigate the adverse health consequences of chronic stress. Growing evidence suggests that virtual reality (VR) interventions constitute an effective approach to stress reduction.

Objective: This systematic review and meta-analysis aimed to examine and compare application areas of VR interventions for stress reduction in the general population and to identify potential predictors of effectiveness based on sample characteristics and intervention design.

Methods: Five databases (MEDLINE, CINAHL, CENTRAL, PsycInfo, and Web of Science) were systematically searched for randomized controlled trials investigating the effectiveness of VR interventions for stress reduction in the general population. Studies were included if they primarily focused on stress reduction, included a neutral control condition, and reported a validated measurement of perceived stress. Trials targeting mental disorders or those conducted in the context of medical procedures were excluded. Two reviewers independently screened the literature, extracted data, and assessed the risk of bias using the Cochrane Collaboration's tool. Effects were synthesized using pooled standardized mean differences, and relevant predictors were evaluated through subgroup analyses and meta-regressions.

Results: A total of 55 relevant studies met the inclusion criteria, with 37 investigating single-session and 18 multisession interventions (ranging from 1 to 42 sessions over 2 days to 6 months). The meta-analysis included 39 studies with 4024 participants (sample sizes 24 - 409; mean ages 19.2 - 70.6 years). Intervention types included VR-based nature exposure (21), biophilic architectural elements (6), guided meditation (9), interactive tasks (4), and other approaches (1). On average, VR interventions significantly reduced perceived stress level (−0.55, 95% CI −0.70 to −0.40; P<.001; I^2=76%; 95% prediction interval [PI] −1.32 to 0.23), anxiety (−0.88, 95% CI −1.23 to −0.54; P<.001; I^2=87%; 95% PI −2.07 to 0.31), and depression (−0.34, 95% CI −0.47 to −0.21; P<.001; I^2=0%; 95% PI −0.47 to −0.21), and enhanced positive emotion (−0.64, 95% CI −0.84 to −0.46; P<.001; I^2=62%; 95% PI −1.28 to −0.02). In addition, the decrease in systolic blood pressure was significant (−0.20, 95% CI −0.37 to −0.04; P=.02; 95% PI −0.37 to −0.04). Univariate linear regression analyses identified multiple sessions (R^2=15%, P=.02) as a significant predictor of VR-based stress reduction. Further analyses of technical and content-related characteristics revealed that a higher image refresh rate (t_{15}=−2.36, 2-tailed; P=.04) was associated with stronger intervention effects.

Conclusions: This review highlights the considerable potential of VR interventions for reducing psychological stress in the general population and extends recent evidence by quantifying the effects of a broad range of VR-based stress reduction approaches and, for the first time, comparing diverse intervention characteristics. VR represents a promising alternative, particularly in contexts where access to real-life nature or conventional relaxation methods is limited or resource-intensive.

Trial Registration: PROSPERO CRD42024592600; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024592600

(J Med Internet Res 2026;28:e78212) doi:10.2196/78212

KEYWORDS

virtual reality; VR; immersion; stress reduction; mental health; positive emotion; general population; stress management; predictor; effectiveness; meta-analysis; systematic review; PRISMA

XSL•FO
RenderX

Introduction

Stress can influence individual health in multiple ways. In general, stress is defined as an organism's reaction to diverse stimuli categorized as harmful or straining by the brain [1]. These stressors can be temporary or recurring and result in an activation of the hormonal and nervous system leading to physical, cognitive, and sociopsychological adjustments [1,2]. While acute stress, for example, time pressure, can have beneficial effects on performance and cognition, chronic stress, characterized by ongoing demands without adequate recovery, is associated with negative consequences for both mental (eg, sleeping or memory issues) and physical health (eg, hypertension). These effects may occur indirectly, by affecting health behavior [3], as well as direct risk factors for several diseases such as cancer, stroke, anxiety disorder, or dementia [2].

The prevalence of both general mental health issues, such as negative emotions, worries, or stress, and mental disorders, such as depression, has increased in European countries over time [4,5]. Global crises, for example, the COVID-19 pandemic, the ongoing climate change, and various armed conflicts, act as collective stressors, negatively affecting the mental well-being of the general population [4,6,7]. In addition, individuals encounter unique challenges that can affect their perceived stress levels (PSLs). According to the Flash Eurobarometer 530, a representative survey on mental health commissioned by the European Commission and conducted across 27 EU countries among citizens over the age of 15 years in 2023, 46% of the participants reported experiencing emotional or psychological problems in the previous 12 months [5]. Among these, 69% reported feeling sad or down, 50% experienced excessive fears and worries, and 34% struggled with daily stress.

The current rise in mental load and its detrimental effects on psychological and physiological health highlight the need for effective strategies to help individuals in coping with daily demands and reducing PSL. According to the Flash Eurobarometer 530, mental health has become a major public concern [5]. The majority of respondents either "totally agreed" (56%) or "tended to agree" (33%) that mental health promotion is just as important as the promotion of physical health. There is strong evidence that stress management techniques, such as mindfulness-based stress reduction [8], progressive muscle relaxation [9], and autogenic training [10], can help individuals in coping with stressful situations and prevent negative health consequences [11]. In addition, exposure to nature has been shown to be an effective method for reducing stress [12,13]. Enhancing scientific evidence, the Flash Eurobarometer 530 found that 35% of the respondents experience mental health benefits from contact with nature and green spaces, while 27% indicated that relaxation techniques improve their mental well-being [5].

Over the past years, virtual reality (VR) has gained significant attention as a promising technology in various health care domains [14,15]. As the quality and accessibility of VR applications are steadily improving, the number of publications on its potential health benefits has noticeably increased. The possibility to expose users to specific virtual environments or situations has given rise to two major fields of VR research and application: the utilization as an alternative and location-independent treatment for mental disorders and, since 2019, as a tool for individual health promotion, including stress reduction and meditation [15,16]. Common application areas with proven effectiveness include VR exposure therapy for severe anxiety disorders [17], such as social anxiety disorder [18,19], posttraumatic stress disorder [20,21], and other psychiatric diagnoses [22,23]. In addition to the therapy of specific disorders, systematic reviews provide evidence for the stress-reducing effects of VR-based relaxation and stress management interventions for both individuals with mental health conditions [24,25] and the general population [26-28].

Recognizing stress management as a key factor in overall health, considerable efforts have been made to translate proven stress reduction techniques into VR, with the aim of creating cost-effective and easily accessible alternatives [24]. For instance, research has demonstrated the efficacy of VR-based stress management based on the exposure to natural environments [29], breathing exercises [30], and mindfulness interventions [26]. Evidence also supports the stress-reducing effects of these interventions in specific target groups and settings, including patients with cancer [31] and cardiovascular disease [32], students and young adults [28], and workplace environments [33]. Overall, current research highlights the potential of VR for stress reduction for specific application areas. Nevertheless, given the heterogeneity of stressors and challenges among different target groups, there is a corresponding variety in requirements and treatment goals for VR-based stress interventions. As most recent reviews are limited to specific target groups (eg, patients with psychiatric disorders [17-20] or young adults [28]) or interventions (eg, nature exposure [29,34] or breathing interventions [30]), a broader meta-analysis comparing different VR intervention approaches regarding their effectiveness for separate target populations is lacking. The effectiveness of these interventions may be influenced by factors such as intervention design, VR content, and technological features, as well as individual characteristics including age, sex, prior experiences, and impairments [14,27,35].

As demonstrated, VR-based interventions have been shown to be effective in reducing PSL in the general population. When daily stress becomes more common among the general population, easily applicable tools for stress reduction for everyone become important for global health promotion. Consequently, the effects of VR interventions for stress reduction, especially in the general population, should be investigated in more detail. To exploit the potential of VR as an easily accessible and purposive method for individual self-practiced stress reduction and to create expedient VR interventions addressing individual requirements and conditions, it is crucial to investigate predictors that enhance stress-reducing effects for different target groups. Therefore, the objective of this systematic review was to synthesize the most common application areas and intervention characteristics in order to quantify and compare the effectiveness of VR for stress reduction across various settings and target groups by

XSL•FO
RenderX

meta-analyses. As other studies focus on the general effectiveness of VR interventions in distinct target populations and the predictors of an enhanced effectiveness in stress reduction remain largely underexplored [28-30], this meta-analysis focused on differences in the effectiveness of VR-based interventions for reducing PSL, comparing subgroups based on sociodemographic characteristics (eg, age, sex, or prior VR experience) and across various application areas (eg, clinical settings, workplaces, universities, and the general population). Furthermore, this meta-analysis exploratively investigated which technical features and methodological design aspects of VR interventions could influence and predict their effectiveness in reducing PSL, so that further VR approaches can be designed more expediently, resulting in more successful VR applications for stress reduction.

Methods

Guidelines and Register

This systematic review was conducted and reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [36], the PRISMA-S (Extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews) guidelines [37], and the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) approach [38]. The study protocol was preregistered in PROSPERO (CRD42024592600).

Literature Search

A systematic search of multiple databases was conducted (HS) on July 24, 2024, to identify randomized controlled trials (RCTs) of immersive VR-based interventions for stress reduction in the general population. The information sources searched included the databases MEDLINE PubMed, CINAHL EBSCOhost, the Cochrane Controlled Trials Register (CENTRAL), PsycInfo EBSCOhost, and Web of Science. Study registries, citation searching, or further resources were not considered. The search strings combined different keywords and text terms related to VR and stress, completed by validated filters [39] for RCTs. There were no further limits applied, for example, for date or language. Search strategies were peer reviewed by a second researcher (KM) regarding PRESS (Peer Review of Electronic Search Strategies) guidelines [40]. The search string used for MEDLINE is provided as an example in Table 1, and the complete search strategies for each database are provided in Multimedia Appendix 1. The search was updated on February 13, 2026, using the same strategy with publication date filters applied.

XSL•FO
RenderX

Table . Search strategy of the systematic search for the systematic review and meta-analysis on virtual reality interventions for stress reduction in the database MEDLINE.

No.	Search terms
#1	“virtual reality”[MeSH Terms]
#2	“virtual reality”[All Fields]
#3	“virtual reality”[Title/Abstract:~2] or “simulated reality”[Title/Abstract:~2] or “digital reality”[Title/Abstract:~2] or “spatial reality”[Title/Abstract:~2]
#4	“virtual environment”[Title/Abstract:~2] or “simulated environment”[Title/Abstract:~2] or “immersive environment”[Title/Abstract:~2] or “digital environment”[Title/Abstract:~2]
#5	“virtual nature”[Title/Abstract:~2] or “simulated nature”[Title/Abstract:~2] or “digital nature”[Title/Abstract:~2] or “computer-generated nature”[Title/Abstract:~2]
#6	“virtual world”[Title/Abstract:~2] or “simulated world”[Title/Abstract:~2] or “immersive world”[Title/Abstract:~2] or “computer-generated world”[Title/Abstract:~2]
#7	“virtual spaces”[Title/Abstract:~2] or “3D environment”[Title/Abstract:~2] or “virtual immersion”[Title/Abstract:~2] or “immersive technology”[Title/Abstract:~2] or “3D simulation”[Title/Abstract:~2] or “3D video”[Title/Abstract:~2] or “360 degree simulation”[Title/Abstract:~2] or “360 degree video”[Title/Abstract:~2] or “virtual exposure”[Title/Abstract:~2] or “virtual experience”[Title/Abstract:~2] or “immersive experience”[Title/Abstract:~2] or “simulated experience”[Title/Abstract:~2]
#8	1 or 2 or 3 or 4 or 5 or 6 or 7
#9	“stress, psychological”[MeSH Terms] or “subjective stress”[MeSH Terms] or “relaxation”[MeSH Terms]
#10	“stress*“[Title/Abstract] or “distress”[Title/Abstract] or “coping”[Title/Abstract] or “relax*“[Title/Abstract] or “resilience”[Title/Abstract] or “restorat*“[Title/Abstract] or “meditat*“[Title/Abstract] or “positive affect”[Title/Abstract] or “negative affect”[Title/Abstract]
#11	9 or 10
#12	“randomized controlled trial”[Publication Type] or “controlled clinical trial”[Publication Type]
#13	“placebo”[Title/Abstract] or “random*“[Title/Abstract] or “trial”[Title/Abstract] or “within subject*“[Title/Abstract] or “between subject*“[Title/Abstract] or “drug therapy”[MeSH Subheading]
#14	“control*“[Title/Abstract] or “comparison”[Title/Abstract] or “compare*“[Title/Abstract] or “study”[Title/Abstract] or “different”[Title/Abstract] or “experiment*"[Title/Abstract]
#15	“participant*“[Title/Abstract] or “patient*“[Title/Abstract] or “group*“[Title/Abstract] or “intervention*“[Title/Abstract] or “individual*“[Title/Abstract] or “subject*“[Title/Abstract] or “therapy”[Title/Abstract] or “adult*"[Title/Abstract]
#16	14 and 15
#17	12 or 13 or 16
#18	8 and 11 and 17
#19	“animals”[MeSH Terms]
#20	“humans”[MeSH Terms]
#21	19 not 20
#22	“review”[Publication Type] or “meta-analysis”[Publication Type]
#23	21 or 22
#24	18 not 23

XSL•FO
RenderX

Eligibility Criteria

Inclusion was limited to RCTs that used either a within-subjects or between-subjects design and were published in English or German. Study protocols were excluded. Included studies were required to focus on human participants aged 18 years or older without any psychiatric diagnoses (eg, anxiety disorder, posttraumatic stress disorder, or depression) due to special demands regarding mental health. No further restrictions were applied regarding demographics such as gender, age, income, or intervention setting (eg, workplace or university, clinical or private setting). This review exclusively encompassed interventions based on VR with a primary focus on stress reduction, delivered through immersive technologies such as cave automatic virtual environments (CAVE) or head-mounted displays (HMD). In this context, a CAVE was defined as an enclosed room with virtual images projected onto the walls, floor, and ceiling. An HMD, in contrast, is a headset worn on the head, providing virtual images and sounds while shielding external stimuli. There was no further restriction regarding intervention content, for example, nature exposure, meditation task, or game. VR-based interventions with other aims than reducing PSL, such as stress inoculation training, pain management, physiological recovery, skill training, or the treatment of psychiatric disorders, were excluded. Furthermore, studies that used VR interventions for anxiety reduction during medical procedures, such as chemotherapy or dental surgery, were excluded to prevent potential biases caused by pain or medication effects. However, studies conducted in a clinical setting or among adults with a physical health condition (eg, cancer) were included if the intervention aimed at a reduction of PSL in general, independent of medical treatment, and no psychiatric diagnosis was involved. Further, inclusion required that the study design incorporated at least one neutral control condition or group, such as no intervention, treatment as usual, waitlist control, or a neutral environment (eg, urban or empty VR) as a placebo. To focus on the primary outcome of this review, only studies that provided PSLs, including arousal or negative affect, were considered. PSL had to be assessed using a validated questionnaire, a visual analog scale (VAS), or a numeric rating scale, with measurements taken at least postintervention. Secondary outcomes encompassed PSL in comparison to an active control group, as well as physiological stress parameters such as heart rate (HR), heart rate variability (HRV), skin conductance level (SCL), and systolic blood pressure (SYS). Secondary psychological outcomes included anxiety, depression, feelings of restoration, and positive emotions. Studies that were limited to physical stress outcomes were excluded.

Study Selection

All records identified through the literature search were first screened for duplicates using the online tool Rayyan [41], followed by a manual verification by one reviewer (HS) to ensure accuracy. After the removal of duplicates, two researchers (HS and IS) independently screened the titles and abstracts to exclude clearly irrelevant studies. Any disagreements were resolved through discussion. Next, full texts of the remaining articles were retrieved and assessed for eligibility based on the above-mentioned predefined eligibility criteria. This process was conducted independently by two researchers (HS and IS). Disagreements were resolved through discussion, and if consensus could not be reached, a third reviewer (KM) was consulted. If relevant information was missing to clearly apply the eligibility criteria, the authors of the respective studies were contacted to provide further details.

Data Collection

Relevant study information and data were extracted from eligible studies by 2 of 3 independent researchers (HS, IS, and KM) using a standardized spreadsheet form. To ensure accuracy, extracted data were compared and any discrepancies during the data extraction process were resolved through discussion. Extracted data items included (1) bibliographic information (first author, year, and country) and study design aspects (within-subjects or between-subjects, sample size, setting, and target population); (2) participant demographics, including age, gender, educational level, family status, race, health conditions, and prior VR experience; (3) numbers of participants in the intervention and control groups, as well as dropout numbers; (4) details of the VR intervention, such as conducted procedure, treatment duration, number and length of VR sessions, VR content and environment (provided relaxation technique, environment style and realism, content motion, technical aspects of the output device [device type, refresh rate, and field of view]), and level of user interactivity; (5) details of passive and, if applicable, active control conditions (procedure, treatment duration, intervention content); (6) details of measurements and analyses; and (7) reported outcome effects. The primary outcome variable of PSL, including negative affect or arousal, was addressed using a predefined hierarchical structure based on use frequency in the included studies to ensure standardized data extraction: Negative Affect Scale of Positive and Negative Affect Schedule (PANAS-NA), Profile of Mood States (POMS), State Scale of State-Trait Anxiety Inventory (STAI-S), Stress Scale of Depression, Anxiety, and Stress Scale (DASS-S), VAS for stress (VAS-S), Perceived Stress Scale (PSS), Warr's Mood Scale (Warr's) and miscellaneous measures. Secondary outcomes included physiological parameters of SCL, HR, HRV (high frequency component [HF-HRV], root mean square of successive differences [rMSSD], low frequency to high frequency ratio [LF/HF], and HRV response fraction), and SYS, as well as measures of restoration (Perceived Restorativeness Scale [PRS], Restoration Outcome Scale [ROS], and Recovery Experience Questionnaire [REQ]), positive emotion (positive affect scale of Positive and Negative Affect Schedule [PANAS-PA] and Zuckerman Inventory of Personal Reactions [ZIPERS]), anxiety (STAI-S, anxiety scale of Depression, Anxiety, and Stress Scale [DASS-A], anxiety subscale of Hospital Anxiety and Depression Scale [HADS-A], Generalized Anxiety Disorder [GAD] scale, VAS for anxiety [VAS-A], trait subscale of State-Trait Anxiety Inventory [STAI-T]), and depression (Self-Rating Depression Scale [SDS], depression subscale of Hospital Anxiety and Depression Scale [HADS-D], depression scale of Depression, Anxiety, and Stress Scale [DASS-D], depression subscale of Profile of Mood States [POMS-D], depression subscale of Warr's Mood Scale [Warr's-D], depression scale of Patient-Reported Outcomes Measurement Information System [PROMIS-D]). Additionally,

XSL•FO
RenderX

measurements of perceived presence (IGroup Presence Questionnaire [IPQ], Presence Questionnaire [PQ], ITC-Sense of Presence Inventory [ITC-SOPI]), cybersickness (Simulator Cybersickness Questionnaire [SSQ]), and user experience with the VR intervention were considered. All effect measures were primarily extracted for postintervention time point in the form of mean and SD. If studies reported only standard errors, CIs, or IQRs, we calculated the SD according to methods described in the *Cochrane Handbook* [42]. In case of multisession interventions, the measure after the first session was extracted when reported. For between-subjects design studies that encompassed multiple relevant VR interventions (eg, an interactive VR game and a noninteractive VR exposure to nature), analyses were based on aggregated means and SDs across VR intervention groups to prevent double-counting of the neutral control. For within-subjects design studies, the VR intervention with the highest "dose," such as the greatest level of biodiversity or stimulus intensity, was selected for inclusion. If relevant information on outcomes or intervention design was missing, the authors of the respective studies were contacted to obtain missing data. In case of nonresponse, studies with incomplete data on the PSL measure were excluded from the meta-analysis. Studies providing data on the primary outcome but limited information on intervention design or intervention characteristics were included in the main analyses. Further analyses only included studies providing the necessary information.

Risk of Bias Assessment

The risk of bias (RoB) was assessed by two independent researchers (HS and KM) for each included study using the Cochrane Collaboration's Risk of Bias tool for randomized trials [42]. Each study was evaluated as having a low, high, or unclear RoB across the following domains: random sequence generation (selection bias), allocation concealment (selection bias), blinding of participants and personnel (performance bias), blinding of outcome assessment (detection bias), incomplete outcome data (attrition bias), and selective reporting (reporting bias). Any discrepancies between reviewers were resolved through discussion. An overall RoB judgment was not conducted.

Certainty of Evidence Assessment

The quality of evidence was assessed by two independent researchers (HS and KM) following the GRADE approach [38]. Therefore, the certainty of evidence was rated for each analyzed outcome (PSL, positive emotion, anxiety, depression, restoration, HR, HRV, SCL, and SYS) and categorized as high, moderate, low, or very low for the domains RoB, inconsistency, indirectness, imprecision, and publication bias. The overall certainty of evidence was determined by the GRADE classification system as high, moderate, low, or very low.

Data Synthesis

To investigate potential predictors and differences in stress-reducing effects based on variations in VR interventions, studies were categorized for the domain setting, intervention type, environment realism, environment style, content motion, and user interactivity. Descriptions of the domains and the applied labels are provided in Multimedia Appendix 2. Ladakis et al [14] synthesize in a scoping review that most studies on VR interventions for stress reduction focus on the following application fields: the general population, students or working adults, groups with high daily stress, and people with special demands in a clinical setting. Therefore, labels for the domain setting were determined as general population, workplace/university, and clinical setting. Environmental realism was poorly described in most studies. Therefore, a binary labeling in real-world recording or computer simulation was conducted. As a dichotomous division (yes or no) did not adequately capture the variety of the included studies in content motion as well as in user interactivity, an intermediate stage was implemented. Consequently, there were studies with no (simple 3D photos or simulations without animated elements), low (a static scene from a consistent perspective with some moving events like leaves in the wind), or high (a moving scene with changing perspective like on a pathway) content motion. User interactivity was also divided into 3 levels, based on Zhang et al [43] and Gras [44]: low control (only watching and exploring scene by head movement), medium control (option to change perspective while moving in the environment by controller or body movement), or high control (option to interact with elements due to an interactive task or game). The assignment of the included studies to the different labels of domains is provided in the Results section. If a study evaluated the characteristics of multiple relevant VR interventions, the effect sizes were pooled and included as one intervention in subgroup analyses. If intervention arms differed in methodological or technical aspects and categorization was not clearly applicable for a domain, the study was excluded from analyses to prevent double counting.

Statistical Analysis

The primary meta-analysis was conducted and visualized by forest plots using Cochrane's Review Manager (RevMan), version 9.11.0. As PSL was measured differently by the included studies, for each study, standardized mean differences (SMDs) between intervention and control groups were calculated for the primary outcome of PSL after the intervention, along with corresponding 95% CIs. Due to substantial variation in intervention design, a random-effects model following the restricted maximum likelihood method was applied to pool the results. As the approach described by Hartung, Knapp, Sidik, and Jonkman (HKSJ) is proven to perform better when trials of similar size are combined and to result in more adequate error rates, the HKSJ correction was applied to calculate CIs. Secondary outcomes were also analyzed using random-effects meta-analyses of SMDs. Following Cohen's recommendations [45], SMDs can be categorized as none (SMD<0.2), small (0.2≤SMD<0.5), medium (0.50≤SMD<0.80), or large (SMD≥0.80) effect.

Statistical precision of the average intervention effect was assessed by I^2 statistic. Further, heterogeneity and the practical robustness of observed intervention effects were explored by 95% prediction intervals (PIs). As the PI for the primary outcome of PSL was not significant, sensitivity analyses were conducted to identify possible subgroups with significant PIs and thus more robust results. For this purpose, studies were

XSL•FO
RenderX

subdivided based on study design, including gender ratio (men or women dominance), age group (<40 years or ≥40 years), setting (clinical setting or general population), target group, and intervention duration (single- or multisession intervention). Further subgroups were formed based on features of the VR intervention, including the used technical device (HMD or CAVE), intervention type, environment style and realism, content motion, and level of user interactivity. The presence of reporting bias was explored by examining funnel plot asymmetry and conducting the Egger test in IBM SPSS Statistics, version 30.0. To assess the robustness of the findings and to evaluate the potential influence of the study design, a series of sensitivity analyses were conducted.

Potential predictors of the effectiveness of VR-based interventions for stress reduction were examined through meta-regressions, conducted in IBM SPSS Statistics, version 30.0, using the inverse-variance method. As the number of studies providing all data for an overall multivariate regression model was inadequate for the number of observed predictors, predictors were purpose-guided and allocated to two subsets defined by study design and technical or content-related VR characteristics for analyses. The assignment of the separate predictors for different meta-regressions is provided in Table 2 as well as the total number of included studies and participants. Only studies providing information on all observed predictors in each subset were included in analyses. Subsequently, all ordinal and continuous explanatory variables regarding study design (mean age of participants, percentage of men, intervention duration in days, number of VR sessions, single session length) were separately tested in univariate linear regression analyses (32 studies). As the chi-square statistic (Multimedia Appendix 3) revealed that a higher age of participants and multisession interventions were more likely in the clinical setting, the setting was dummy coded and integrated in regression analyses (clinical setting=yes/no). Further, multivariate regression analyses examining technical and content-related VR characteristics were conducted in a subset of 17 studies that provided sufficient data. As statistical power was low, the number of investigated predictors had to be limited. The chi-square test revealed a high correlation of the intervention type with the level of content motion and user interactivity, wherefor this categorial variable was not considered. Meta-regressions were visualized using bubble plots, generated in R Studio (version 2026.01.00). For all statistical tests, a *P* value of <.05 was considered the threshold for statistical significance.

Table . Overview of the assignment of predictors to different meta-regression analyses to investigate possible predictors of the effectiveness of virtual reality interventions on perceived stress level in the general population, including the total number of included studies and participants for each analysis.

	Involved predictors	Included studies (references)	Studies, n	Participants, n
Study design	Mean age (years), males (%), multisession intervention (1,0), intervention duration (days), number of VR[a] sessions (1-42), session length (min), clinical setting (1,0)	[46-77]	32	3023
Technical and content factors of VR	Content motion (1-3), user interactivity (1-3), environment realism (1,0), refresh rate (Hz), FoV[b] (degree)	[46-48,50,51,53-55,57,60,64,66,68,69,76,78,79]	17	1831

[a]VR: virtual reality.

[b]FoV: field of view.

Results

Literature Search

The PRISMA flow diagram (Figure 1) provides an overview of the study selection process. The database search yielded 13,187 results, of which 4971 were identified as duplicates and removed. Following the screening of the titles and abstracts of the remaining 8217 papers, 296 articles were retrieved and screened in full text. In total, 55 studies met the predefined and in methods section described inclusion criteria and were included in the systematic review. Of these, 16 publications did not provide complete data (ie, mean and SD postintervention for both intervention and control group) for the primary outcome of PSL and were excluded from the meta-analysis (Table 3, MA="no"), resulting in 39 studies included in the quantitative synthesis. Studies that seemed to fit at first glance but were excluded during full-text screening are provided in Multimedia Appendix 4.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the study selection process for the systematic review and meta-analysis on virtual reality interventions for stress reduction in the general population (searches conducted in June 2024 and February 2026). VR: virtual reality.

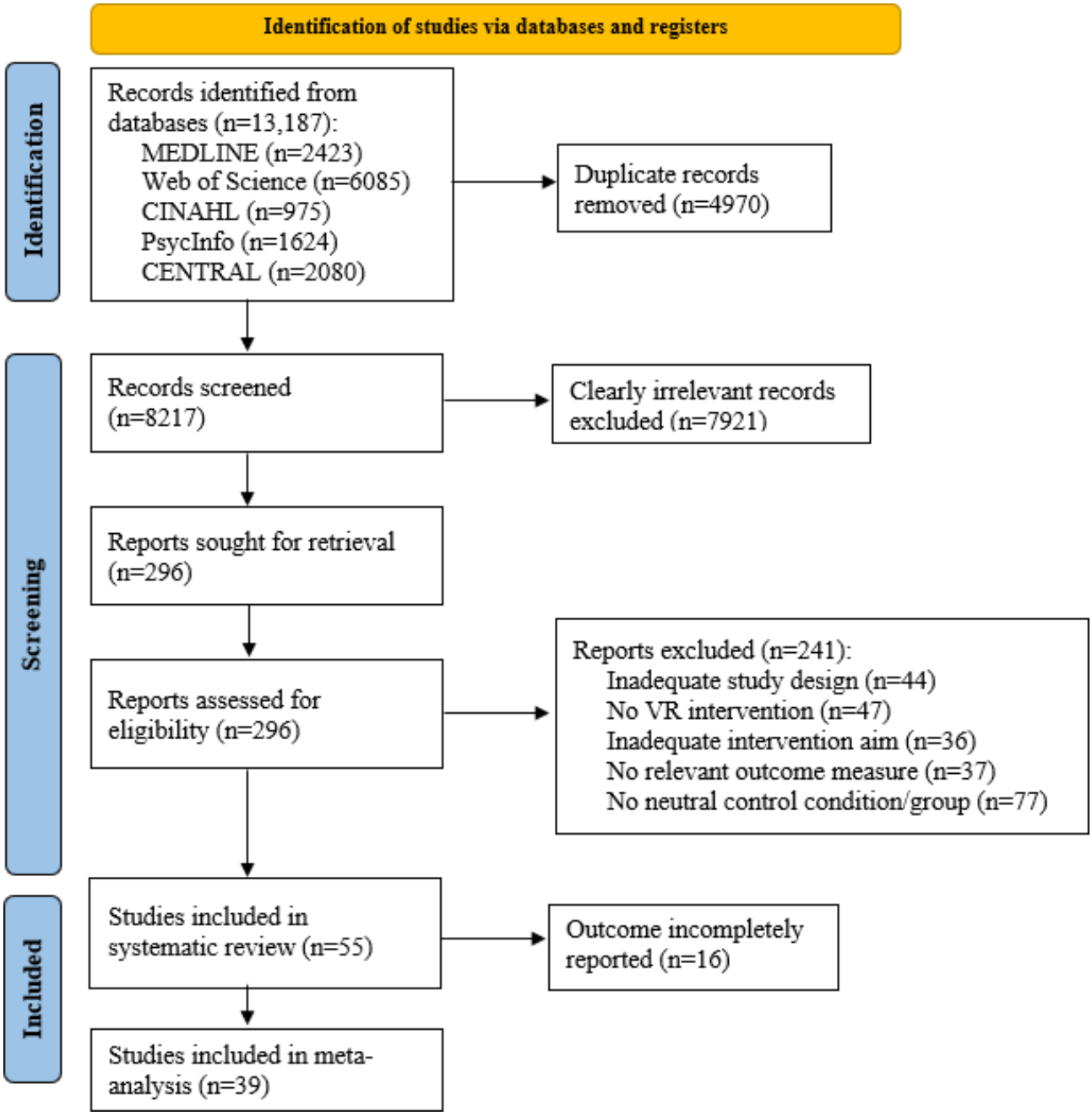

XSL•FO
RenderX

Table . Overview of the main characteristics of the included studies in the systematic review and meta-analysis on virtual reality interventions for stress reduction for the general population, providing details on study design, intervention setting, participants, intervention type and duration, and virtual reality output device.

Study ID	MA[a]	Country	Study design	Setting	Randomized, n	Men, %	Mean age, years	Intervention type	Control (active control)[b]	Output device	Duration of intervention	Session length (number)
Ahn 2025 [46]	Yes	Korea	Between	Workplace/university	41	26.8	21.3	Nature exposure	WLC (TSM)	HMD[c]	2 weeks	20 min (6)
Anderson 2017 [80]	No	USA	Within	General population	18	50.0	32.0	Nature exposure	eVR	HMD	1 time	15 min (1)
Bodet-Contentin 2023 [78]	Yes	France	Within	Workplace/university	88	19.3	—[d]	Meditation	TAU/NI	HMD	1 time	8 min (1)
Browning 2019 [47]	Yes	USA	Between	Workplace/university	98	43.9	20.2	Nature exposure	TAU/NI (RWE)	HMD	1 time	6 min (1)
Chand 2024 [48]	Yes	India	Between	Workplace/university	44	70.45	24.4	Miscellaneous	TAU/NI	HMD	6 days	15 min (6)
Chen 2024 [49]	Yes	USA	Within	General population	173	49.1	37.0	Nature exposure	bVR	CAVE[e]	1 time	5 min (1)
Da 2024 [81]	No	China	Between	General population	273	24.0	22.3	Biophilic design	bVR	HMD	1 time	3 min (1)
Emanjomeh 2020 [50]	Yes	USA	Within	Workplace/university	35	77.1	23.5	Biophilic design	eVR (RWE)	HMD	1 time	5 min (1)
Gao 2019 [82]	No	China	Between	Workplace/university	116	50.0	20.7	Nature exposure	bVR (aVR)	HMD	1 time	5 min (1)
Gao 2024 [83]	Yes	China	Between	Workplace/university	195	—	—	Nature exposure	eVR	HMD	1 time	3 min (1)
Gentile 2024 [51]	Yes	USA	Between	Workplace/university	409	39.1	19.4	Meditation	ot (TSM)	HMD	1 time	20 min (1)
Hessabi 2020 [52]	Yes	Iran	Between	Clinical setting	60	50.0	51.0	Nature exposure	TAU/NI	HMD	4 days	15 min (2)
Hooyberg 2023 [53]	Yes	Belgium	Within	General population	164	32.3	34.9	Nature exposure	bVR (aVR)	HMD	1 time	16 min (2)
Huang 2020 [54]	Yes	China	Between	Workplace/university	89	49.4	23.0	Biophilic design	bVR	HMD	1 time	10 min (1)
Hung 2025 [84]	No	Taiwan	Between	General population	111	50.0	24.0	Nature exposure	bVR	HMD	1 time	5 min (1)
Jiang 2025 [85]	No	China	Between	Workplace/university	66	—	26.9	Meditation	TAU/NI (TSM)	HMD	8 days	20 min (8)

XSL•FO
RenderX

Study ID	MA[a]	Country	Study design	Setting	Randomized, n	Men, %	Mean age, years	Intervention type	Control (active control)[b]	Output device	Duration of intervention	Session length (number)
Jimenez 2025 [86]	Yes	USA	Between	General population	200	19.0	—	Meditation	TAU/NI	HMD	1 time	6 min (1)
Jimenez-Barragan 2025 [55]	Yes	Spain	Between	Clinical setting	70	0.0	31.9	Meditation	TAU/NI	HMD	6 weeks	14 min (42)
Kawai 2024[f] [56]	Yes	Switzerland	Between	General population	96	48.9	39.4	Nature exposure	bVR	HMD	1 time	20 min (1)
Kerr 2023 [57]	Yes	Swiss	Between	Workplace/university	87	51.7	22.9	Meditation	WLC	HMD	4 weeks	20 - 26 min (4)
Kılıç 2023 [58]	Yes	Turkey	Between	Clinical setting	131	0.0	23.3	Nature exposure	TAU/NI	HMD	2 days	5 min (6)
Kim 2024 [59]	Yes	Korea	Between	Workplace/university	60	31.7	21.9	Meditation	TAU/NI (TSM)	HMD	5 days	30 min (5)
Knaust 2022 [87]	No	Germany	Within	General population	102	59.8	36.5	Nature exposure	TAU/NI (2D)	HMD	1 time	5 min (1)
Kosa 2024 [60]	Yes	USA	Between	Workplace/university	202	54.5	19.2	Game/activity	ot (2D, aVR)	HMD	1 time	15 min (1)
Kumpulainen 2024 [61]	Yes	Finland	Within	General population	57	33.3	40.0	Nature exposure	TAU/NI	CAVE	1 time	10 min (1)
Lepilkina 2023 [62]	Yes	Russia	Between	Clinical setting	68	44.1	42.9	Meditation	TAU/NI	HMD	5 days	20 - 30 min (5)
Li 2025 [63]	Yes	China	Between	Clinical setting	98	39.8	60.5	Meditation	TAU/NI	HMD	1 time	25 min (1)
Liszio 2018 [64]	Yes	Germany	Between	General population	62	41.9	22.6	Nature exposure	TAU/NI (2D)	HMD	1 time	7 min (1)
Liszio 2019[f] [65]	Yes	Germany	Between	General population	57	28.1	23.7	Game/activity, nature exposure	TAU/NI	HMD	1 time	9 min (1)
Liu 2023[f] [66]	Yes	China	Within	Workplace/university	24	41.7	21.8	Biophilic design	bVR	HMD	1 time	6 min (1)
Ma 2025 [67]	Yes	China	Between	Workplace/university	170	33.5	19.9	Game/activity	TAU/NI (oVR)	HMD	1 time	25 min (1)
Ma 2026[f] [68]	Yes	China	Between	General population	61	54.1	24.7	Biophilic design	bVR	HMD	1 time	6 min (1)
Mahmud 2022 [88]	No	Malaysia	Between	General population	67	37.3	33.2	Nature exposure	WLC (TSM)	HMD	2 weeks	5 min (6)

XSL•FO
RenderX

Study ID	MA[a]	Country	Study design	Setting	Randomized, n	Men, %	Mean age, years	Intervention type	Control (active control)[b]	Output device	Duration of intervention	Session length (number)
Martínez Manchón 2024[f] [89]	Yes	Croatia	Between	Workplace/university	64	7.8	—	Game/activity, nature exposure	eVR	HMD	1 time	6 min (1)
Newman 2022[f] [69]	Yes	England	Between	General population	120	13.3	20.0	Nature exposure	bVR	HMD	1 time	10 min (1)
Plante 2003 [90]	No	USA	Between	Workplace/university	154	33.8	—	Nature exposure	TAU/NI (RWE)	HMD	1 time	20 min (1)
Richesin 2021 [91]	No	USA	Between	Workplace/university	44	18.2	21.2	Game/activity	eVR (2D)	HMD	1 time	15 min (1)
Şansal 2024 [70]	Yes	Turkey	Within	General population	60	40.0	70.6	Nature exposure	bVR	HMD	1 time	6 min (1)
Schebella 2019 [92]	No	Australia	Between	General population	52	46.2	37.6	Nature exposure	bVR	HMD	1 time	5 min (1)
Schutte 2017 [71]	Yes	Australia	Between	Workplace/university	26	38.5	34.5	Nature exposure	bVR	HMD	1 time	6 min (1)
Song 2022 [72]	Yes	China	Between	Clinical setting	70	68.3	60.4	Nature exposure	bVR	HMD	1 week	5:20 min (3-5)
Sun 2023[f] [73]	Yes	China	Between	Clinical setting	63	0.0	31.8	Nature exposure, biophilic design	bVR	HMD	1 time	5 min (1)
Suppakittpaisarn 2023 [79]	Yes	Taiwan, Thailand, USA	Between	General population	270	51.9	—	Nature exposure	bVR	HMD	1 time	1 - 15 min (1)
Valtchanov 2010 [93]	Yes	Canada	Between	Workplace/university	69	46.4	—	Nature exposure	eVR (aVR)	HMD	1 time	10 min (1)
Villani 2008 [94]	No	Italy	Between	Workplace/university	64	46.9	24.5	Meditation	TAU/NI (2D, TSM)	HMD	1 week	— (2)
Wu 2025 [95]	No	China	Between	Clinical setting	85	0.0	46.4	Meditation	bVR	HMD	2 weeks	15 min (3-5)
Xiaoxue 2024 [96]	No	China	Within	General population	23	30.4	62.7	Biophilic design	bVR	HMD	1 week	5 min (1)
Yang 2024 [74]	Yes	Ukraine	Between	Workplace/university	120	60.8	22.0	—	TAU/NI	HMD	6 months	60 min (26)
Yildirim 2024 [75]	No	Australia	Within	Workplace/university	32	50.0	29.2	Biophilic design	bVR	HMD	1 time	6 min (1)
Yin 2024 [97]	No	China	Between	Workplace/university	320	43.8	21.1	Nature exposure	TAU/NI	HMD	1 time	3 min (1)

XSL•FO
RenderX

Study ID	MA[a]	Country	Study design	Setting	Randomized, n	Men, %	Mean age, years	Intervention type	Control (active control)[b]	Output device	Duration of intervention	Session length (number)
Yin 2020[f] [76]	Yes	USA	Between	Workplace/university	100	37.0	29.0	Biophilic design	eVR	HMD	1 time	6 min (1)
Yu 2018 [98]	No	Taiwan	Within	Workplace/university	30	43.3	—	Nature exposure	bVR	HMD	1 time	9:30 min (1)
Yuan 2022 [99]	No	China	Between	Clinical setting	63	33.3	82.0	Nature exposure	TAU/NI	HMD	3 days	5 min (3)
Zhang 2022 [77]	Yes	China	Between	Clinical setting	90	0.0	51.7	Nature exposure	TAU/NI	HMD	3 months	30 min (6)
Zheng 2024[f] [100]	Yes	China	Between	Workplace/university	60	26.7	20.2	Meditation, nature exposure	TAU/NI	HMD	2 weeks	10 - 15 min (—)

[a]Study included (yes) or excluded (no) from meta-analyses (MA).

[b]Neutral control condition was categorized as “TAU/NI” (treatment as usual or no intervention), “eVR” (empty or blank VR), “bVR” (built or urban VR), “WLC” (waitlist control), or “ot” (other task). Active control condition was categorized as “2D” (2D simulation), “RWE” (real-world exposure), “TSM” (traditional stress management), or “aVR” (alternative VR intervention).

[c]HMD: head-mounted display.

[d]Not available.

[e]CAVE: cave automatic virtual environment.

[f]Study observed more than one VR intervention of interest (eg, noninteractive and interactive group, different levels of biodiversity/realism). Primary VR intervention and other VR intervention effects were pooled for analyses.

Characteristics of Studies Included in the Systematic Review

Table 3 provides an overview of the main characteristics of the 55 studies included in this review [46-52,54-100]. Most studies were published within the past 5 years, while 8 were conducted between 2017 and 2019, and 3 prior to 2011. Among them, 43 studies used a between-subjects and 12 a within-subjects design. The 55 studies included one of the following neutral control conditions: 20 trials built/urban VR environment, 6 trials empty/blank VR control, 3 trials waitlist control, 23 trials treatment as usual/no intervention, and 2 trials other nonrelaxing tasks without VR. A total of 17 studies included an additional active control condition: 5 trials digital 2D simulations (2D), 3 trials real-world nature exposure (RWE), 6 trials traditional stress management (TSM) techniques, and 5 trials an alternative VR intervention that was not part of the main intervention focus. About 18 trials focused on the general population [49,53,56,61,64,65,68-70,79-81,84,86-88,92,96], 27 trials were conducted in work-related settings (2 in workplace, 23 in university, and 2 combined) [46-48,50,51,54,57,59,60,66,67,71,74-76,78,82,83,85,89-91,93,94,97,98,100], and 10 trials targeted subgroups with specific demands in clinical settings (eg, pregnant women, older adults, or clinical residents) [52,55,58,62,63,72,73,77,95,99].

Most studies (37 trials) implemented single-session interventions (Table 3). The remaining 18 studies involved multisession interventions lasting from 2 days to 6 months (1 trial 2 days, 1 trial 3 days, 1 trial 4 days, 2 trial 5 days, 1 trial 6 days, 3 trials 1 week, 1 trial 8 days, 3 trials 2 weeks, 1 trials 4 weeks, 1 trial 5 weeks, 1 trial 6 weeks, 1 trial 3 months, and 1 trial 6 months). The number of VR sessions ranged from 1 to 42 (37 trials 1 session, 9 trials 2 - 5 sessions, 7 trials 6 - 8 sessions, 2 trials more than 26 sessions). The duration of individual VR sessions ranged from 1 to 60 minutes (29 trials <10 min, 25 trials 10 - 30 min, 1 trial>30 min).

A total of 31 studies investigated the effects of pure exposure to natural environments in VR (eg, forest, beach, or park) on stress reduction [46,47,49,52,53,56,58,61,64,65,70-73,77,79,80,82-84,87-90,92,93,97-100]. About 12 studies combined natural VR environments with a meditation task [51,55,57,59,62,63,78,85,86,94,95,100]. Another 9 studies studied the effect of biophilic design elements integrated into environments such as offices, streets, or courtyards in VR [50,54,66,68,73,75,76,81,96]. Five studies focused on games or interactive tasks as the main VR content [60,65,67,89,91], and 1 study [48] investigated miscellaneous content in VR. About 17 studies included more than one VR intervention to examine differences in factors such as environment type and biodiversity [49,53,54,66,68,69,72,73,76,80,82,92,96,97], meditation guidance [100], or level of user interactivity [65,89]. In nearly all studies (n=53), the VR intervention was delivered via HMD, while only 2 studies used a CAVE as an output device [49,61]. Around 28 studies used real-world recordings (photos or videos) as VR content [46-49,52,53,56,58,61,66,70-73,79,80,82-85,87,89,90,92,95,97-99], whereas 25 used a computer-generated simulation

XSL•FO
RenderX

[50,51,54,55,57,59,60,62-65,67-69,75-78,81,86,88,93,94,96]. Two studies did not provide any technical details about the VR intervention.

Meta-Analysis of Included Studies

Characteristics of Studies Included

A total of 39 studies provided sufficient data on the primary outcome of PSL and were included in the meta-analysis (Table 4) [46-53,55-79,83,86,89,93,100]. In total, the meta-analysis included 4024 participants, of whom 2179 received a VR-based intervention for stress reduction and 1845 were assigned to a neutral control condition. Participants' mean ages ranged from 19.2 to 70.6 years (27 trials <40 years, 7 trials ≥40 years, 5 trials not reporting age). Further characteristics of the included studies are available in Table 3.

Possible predictors for stress reduction were analyzed based on both general study characteristics and further technical and content-related aspects of VR interventions. The detailed information used for categorization as well as reported outcome parameters for studies included in the meta-analysis are provided in Table 4. Most studies used a simple nature exposure intervention and offered limited user interaction, typically restricted to head movement for scene exploration.

XSL•FO
RenderX

Table . Technical details and reported outcome measures of the included randomized controlled trials in the quantitative synthesis of the systematic review and meta-analysis on virtual reality interventions for stress reduction in the general population.

Study ID	Session length, min (number)	Environment realism (style)[a]	Content motion[b]	FoV[c], degree (refresh rate, Hz)	User interactivity[d]	Stress (PSL) outcome[e]	Anxiety / depression[f]	Positive emotion outcome[g]	Restorativeness outcome[h]	Physiological outcomes[i]
Ahn 2025 [46]	20 (6)	Recording (dl)	Low	134.38 (120)	Medium	PSS	PROMIS-A/PROMIS-D	—[j]	—	—
Bodet-Contentin 2023 [78]	8 (1)	Simulation (gs)	Low	126.57 (60)	Low	VAS-S	VAS-A/—	—	—	—
Browning 2019 [47]	6 (1)	Recording (gs)	Low	96.0 (60)	Low	PANAS-NA	—	PANAS-PA	PRS	—
Chand 2024 [48]	15 (6)	Recording (bs)	Low	134.38 (120)	Low	DASS-S	DASS-A/DASS-D	—	—	LF/HF
Chen 2024 [49]	5 (2)	Recording (gs)	Low	— (—)	Low	POMS	STAI-S/POMS-D	—	—	HR, HRV response fraction, SCL
Emamjomeh 2020 [50]	5 (1)	Simulation (ba)	No	138.59 (90)	Low	PANAS-NA	—	PANAS-PA	—	—
Gao 2024 [83]	3 (1)	Recording (gs)	—	150.57 (120)	Low	PANAS-NA	—/BPOMS-D	PANAS-PA	—	HR, SYS
Gentile 2024 [51]	20 (1)	Simulation (dl)	Low	145.17 (90)	Low	Other	—	Other	—	—
Hessabi 2020 [52]	15 (2)	Recording (—)	—	— (—)	—	STAI-S	—	—	—	—
Hooyberg 2023 [53]	16 (2)	Recording (bs)	Low	124.45 (80)	Low	PANAS-NA	—	PANAS-PA	—	HR, SCL
Huang 2020 [54]	10 (1)	Simulation (ba)	No	135.83 (75)	—	PANAS-NA	—	PANAS-PA	—	SCL
Jimenez 2025 [86]	6 (1)	Simulation (dl)	—	126.57 (60)	Low	STAI-S	VAS-A/—	—	—	—
Jimenez-Barragan 2025 [55]	14 (42)	Simulation (bs)	Low	126.57 (60)	Low	STAI-S	—	—	—	HR
Kawai 2024[k] [56]	20 (1)	Recording (gs)	Low	134.38 (120)	Low	Other	—	—	ROS	SCL
Kerr 2023 [57]	21.5[l] (4)	Simulation (gs)	Low	131.52 (75)	Low	DASS-S	DASS-A/DASS-D	—	—	HR, HF-HRV, SYS
Kılıç 2023 [58]	5 (6)	Recording (gs)	High	134.38 (120)	—	DASS-S	STAI-S/—	—	—	—
Kim 2024 [59]	30 (5)	Simulation (—)	—	134.38 (120)	Low	PSS	—	—	—	HRV
Kosa 2024 [60]	15 (1)	Simulation (gs)	High	134.38 (120)	High	PANAS-NA	STAI-S/—	PANAS-PA	REQ	—
Kumpulainen 2024 [61]	10 (1)	Recording (gs)	High	— (—)	Low	Warr's	— /Warr's-D	Other	—	HR, HF-HRV
Lepilkina 2023 [62]	25[l] (5)	Simulation (dl)	—	124.45 (80)	—	STAI-S	—	—	—	—
Li 2025 [63]	25 (1)	Simulation (dl)	—	— (90)	Low	STAI-S	—	—	—	HR, SYS

XSL•FO
RenderX

Study ID	Session length, min (number)	Environment realism (style)[a]	Content motion[b]	FoV[c], degree (refresh rate, Hz)	User interactivity[d]	Stress (PSL) outcome[e]	Anxiety / depression[f]	Positive emotion outcome[g]	Restorativeness outcome[h]	Physiological outcomes[i]
Liszio 2018 [64]	7 (1)	Simulation (uw)	Low	123.75 (90)	Low	PANAS-NA	STAI-S/—	PANAS-PA	—	SDSD
Liszio 2019[k] [65]	9 (1)	Simulation (bs)	Low	123.75 (90)	Low/high	PANAS-NA	STAI-S/—	PANAS-PA	—	SDSD
Liu 2023 [66]	6 (1)	Recording (ba)	High	138.59 (90)	Low	POMS	—	—	ROS	LF/HF
Ma 2025 [67]	25 (1)	Simulation (dl)	High	— (70)	High	PANAS-NA	STAI-S	PANAS-PA	—	HR
Ma 2026[k] [68]	6 (1)	Simulation (bd)	High	138.59 (90)	Medium	STAI-S	—	—	—	HR, rMSSD, SCL
Martínez Manchón 2024[k] [89]	6 (1)	Recording (dl)	Low/high	134.38 (120)	Low/high	VAS-S	—	—	—	—
Newman 2022 [69]	10 (1)	Simulation (gs)	High	145.17 (90)	Medium	PANAS-NA	—	PANAS-PA	—	—
Şansal 2024 [70]	6 (1)	Recording (gs)	Low	— (—)	Low	PANAS-NA	—	PANAS-PA	PRS	—
Schutte 2017 [71]	6 (1)	Recording (gs)	Low	— (—)	Low	PANAS-NA—		PANAS-PA	PRS	—
Song 2022 [72]	5.33[l] (4)	Recording (dl)	No	130.11 (70)	—	PANAS-NA	—/SDS	PANAS-PA	—	HR, SYS
Sun 2023[k] [73]	5 (1)	Recording (gs/ba)	Low	— (—)	Low	PANAS-NA	—	PANAS-PA	—	HR, SCL, SYS
Suppakittpaisarn 2023 [79]	7[l] (1)	Recording (gs)	Low	131.52 (72)	Low	VAS-S	—	—	—	—
Valtchanov 2010 [93]	10 (1)	Simulation (bs)	High	— (—)	Medium	VAS-S	—	ZIPERS	—	HR
Yang 2024 [74]	60 (26)	—	—	145.17 (90)	Medium	STAI-S	STAI-T/—	—	—	—
Yildirim 2024 [75]	6 (1)	Simulation (ba)	—	134.38 (120)	—	Other	—	—	—	HR, SCL, SYS
Yin 2020 [76]	6 (1)	Simulation (ba)	No	145.17 (90)	Medium	STAI-S	—	—	—	HR, rMSSD, SCL, SYS
Zhang 2022 [77]	30 (6)	Simulation (bs)	High	— (—)	Medium	DT	SAS/SDS	—	—	—

XSL•FO
RenderX

Study ID	Session length, min (number)	Environment realism (style)[a]	Content motion[b]	FoV[c], degree (refresh rate, Hz)	User interactivity[d]	Stress (PSL) outcome[e]	Anxiety / depression[f]	Positive emotion outcome[g]	Restorativeness outcome[h]	Physiological outcomes[i]
Zheng 2024[k] [100]	12.5[l] (—)	— (dl)	—	— (—)	—	PSS	GAD/—	—	—	—

[a]Environment style was categorized as “gs” (green space, eg, garden, forest, park), “bs” (blue space, eg, island or beach), “dl” (different landscapes), “ba” (biophilic architecture, eg, office room, courtyard, street scenes), or “uw” (under water).

[b]Content motion was ranked as no motion (simple photos or simulations without animated elements), low (a static scene from a consistent perspective with some moving events like leaves in the wind), or high (a moving scene with changing perspective like on a pathway).

[c]Diagonal field of view (FoV) in degree: (horizontal fov)2+(vertical fov)2

[d]User interactivity was ranked as low control (only watching and exploring scenes by head movement), medium control (option to change perspective while moving in the environment by controller or body movement), or high control (option to interact with elements due to an interactive task or game).

[e]DASS-S: stress scale of Depression, Anxiety, and Stress Scale; DT: distress thermometer; PANAS-NA: negative affect scale of Positive and Negative Affect Scale; POMS: Profile of Mood States; PSL: Perceived Stress Level; PSS: Perceived Stress Scale; STAI-S: state scale of State-Trait Anxiety Inventory; VAS-S: Visual Analog Scale for Stress; Warr’s: Warr’s Mood Scale.

[f]BPOMS-D: depression subscale of Brief Profile of Mood States; DASS-A: anxiety scale of Depression, Anxiety, and Stress Scale; DASS-D: depression scale of Depression, Anxiety, and Stress Scale; GAD: Generalized Anxiety Disorder Scale; POMS-D: depression subscale of Profile of Mood States; PROMIS-A: anxiety scale of Patient-Reported Outcomes Measurement Information System; PROMIS-D: depression scale of Patient-Reported Outcomes Measurement Information System; SAS: Self-Rating Anxiety Scale; SDS: Self-Rating Depression Scale; STAI-T: trait subscale of State-Trait Anxiety Inventory; VAS-A: Visual Analog Scale for Anxiety; Warr’s-D: depression subscale of Warr’s Mood Scale.

[g]PANAS-PA: positive affect scale of Positive and Negative Affect Scale; ZIPERS: Zuckerman Inventory of Personal Reactions.

[h]PRS: Perceived Restorativeness Scale; REQ: Recovery Experience Questionnaire; ROS: Restoration Outcome Scale.

[i]HR: heart rate; HRV: heart rate variability; LF/HF: low frequency to high frequency ratio; rMSSD: root mean square of successive differences; SCL: skin conductance level; SDSD: standard deviation of successive differences; SYS: systolic blood pressure.

[j]Not available.

[k]Study observed more than one VR interventions of interest (eg, noninteractive and interactive group, different levels of biodiversity/realism). Primary VR intervention and other VR intervention effects were pooled for analyses. Subgroup analyses and meta-regression studies were excluded if categorization was not applicable.

[l]Study used different intervention durations or session lengths of VR without a separate data report. For analyses, the mean value was used.

Risk of Bias Assessment

Summary graphs illustrating the RoB in the included studies are shown in Figure 2. The risk of selection bias was rated as low or unclear in all studies, with the latter primarily due to insufficient reporting on randomization and allocation procedures. Owing to the unblinded nature of the interventions, most studies were rated as having an unclear or high risk of performance and detection bias. Reporting bias was generally assessed as low for most of the studies.

XSL•FO
RenderX

Figure 2. Risk of bias summary graph for the included studies in the meta-analysis on virtual reality interventions for stress reduction in the general population, generated using Cochrane's Review Manager (RevMan), version 9.11.0. Risk of bias was rated as low (green), high (red), or unclear (yellow) for each domain [46-74,76-79,81,83,86,89,93,100].

XSL•FO
RenderX

Perceived Psychological Stress

The forest plot for the postintervention effect on perceived stress is shown in Figure 3. On average, the pooled effect of VR-based interventions on PSL was significant (SMD −0.55, 95% CI −0.70 to −0.40; $P<.001$), considering the moderate-to-high level of variation in study outcomes ($I^2=76\%$, $P<.001$). However, the nonsignificant 95% PI (−1.32 to 0.23) suggests that while the majority of participants experienced a reduction in PSL through the VR intervention, a subset of participants showed no decrease, or even an increase in stress levels.

The Egger test for funnel plot asymmetry was marginally significant (intercept 0.66, $P=.05$). Moreover, inspection of the funnel plot (Figure 4) suggests the potential presence of small-study bias, possibly due to publication bias or the high heterogeneity of included studies.

To evaluate study quality as a potential source of heterogeneity, a series of sensitivity analyses were performed (Table 5). First, a sensitivity analysis only including between-subjects design studies was conducted to exclude possible carryover effects related to within-subjects design studies. To address potential biases from small sample sizes or a high RoB in critical domains, further analyses were restricted to studies with at least 100 participants, low risk of selection bias, and low risk of attrition bias. To examine the impact of potential performance bias, a further analysis only involved studies with low or unclear RoB. An additional sensitivity analysis excluded the two studies using a CAVE system as an output device and thus was limited to HMD. As illustrated in Table 5, all sensitivity analyses yielded significant pooled effect sizes ($P<.001$), while I^2 values remained at moderate to high levels. However, PIs did not reach significance in any of these analyses, indicating that the investigated factors did not contribute substantially to heterogeneity between studies.

To investigate variation in study outcomes and the precision of overall effects of VR-based interventions on PSL in more detail, further exploratory sensitivity analyses were performed. For this purpose, studies were stratified by age group (<40 years or ≥40 years), intervention duration (single-session or multisession), gender dominance (more women or more men), setting (general population or clinical), target population, and intervention type. In all subset analyses, I^2 values remained above 50%, and the range of the 95% PIs could not be reduced (see Multimedia Appendix 5) [46-74,76-79,81,83,86,89,93,100]. Thus, the sensitivity analyses did not explain the observed heterogeneity.

XSL•FO
RenderX

Figure 3. Forest plot for the primary outcome of perceived stress (postintervention) of the meta-analysis on virtual reality interventions for stress reduction in the general population, generated using Cochrane's Review Manager (RevMan), version 9.11.0, and including effect sizes for separate studies [46-74,76-79,81,83,86,89,93,100]. VR: virtual reality.

Study or Subgroup	VR intervention Mean	VR intervention SD	VR intervention Total	Neutral control Mean	Neutral control SD	Neutral control Total	Weight	Std. mean difference IV, Random, 95% CI
Ahn 2025	10.46	4.33	13	19.77	5.4	13	1.4%	-1.84 [-2.78 , -0.90]
Bodet-Contentin 2023	1.23	1.8	88	1.4	1.63	88	3.1%	-0.10 [-0.39 , 0.20]
Browning 2019	1.24	0.46	32	1.2	0.57	32	2.5%	0.08 [-0.41 , 0.57]
Chand 2024	5	2.74	22	10.36	7.45	22	2.1%	-0.94 [-1.56 , -0.31]
Chen 2024	-0.04	12.59	173	4.83	15.09	173	3.3%	-0.35 [-0.56 , -0.14]
Emamjomeh 2020	12	2.67	30	13.53	5.05	30	2.4%	-0.37 [-0.88 , 0.14]
Gao 2024	14.42	6.02	147	18.27	8.79	41	2.9%	-0.57 [-0.92 , -0.22]
Gentile 2024	1.53	0.55	100	1.75	0.6	103	3.2%	-0.38 [-0.66 , -0.10]
Hessabi 2020	34.66	20.1	30	56.56	36.26	30	2.4%	-0.74 [-1.26 , -0.21]
Hooyberg 2023	0.16	0.32	164	0.37	0.47	55	3.1%	-0.58 [-0.89 , -0.27]
Huang 2020	38.37	52.78	59	83.11	120.07	28	2.6%	-0.55 [-1.01 , -0.09]
Jimenez 2025	33.4	11.3	101	41.6	13.9	99	3.1%	-0.65 [-0.93 , -0.36]
Jimenez-Barragan 2025	35.03	8.09	33	33.4	11.3	101	2.8%	0.15 [-0.24 , 0.55]
Kawai 2024[a]	3.3	2.04	44	3.05	2.05	41	2.7%	0.12 [-0.30 , 0.55]
Kerr 2023	8.06	6.08	37	10.42	8.49	41	2.6%	-0.31 [-0.76 , 0.13]
Kim 2024	2.24	0.51	20	2.69	0.51	20	2.0%	-0.86 [-1.52 , -0.21]
Kılıç 2023	4.75	3.4	63	8.25	4.13	63	2.9%	-0.92 [-1.29 , -0.55]
Kosa 2024	1.14	0.24	50	1.51	0.68	51	2.8%	-0.72 [-1.12 , -0.31]
Kumpulainen 2024	2.3	1.8	57	2.6	1.9	57	2.9%	-0.16 [-0.53 , 0.21]
Lepilkina 2023	39.05	9.45	40	38.8	8.76	20	2.3%	0.03 [-0.51 , 0.56]
Li 2025	41.02	6.81	45	44.25	7.57	53	2.8%	-0.44 [-0.85 , -0.04]
Liszio 2018	11.9	2.53	22	16.3	6.87	23	2.1%	-0.83 [-1.44 , -0.22]
Liszio 2019[a]	11.42	2.29	38	11.74	1.59	19	2.3%	-0.15 [-0.70 , 0.40]
Liu 2023	4.13	8.45	24	16.31	10.43	24	2.1%	-1.26 [-1.89 , -0.64]
Ma 2025	12.53	3.5	55	22.48	7.78	61	2.7%	-1.61 [-2.03 , -1.19]
Ma 2026[a]	17.9	3.84	39	19.04	4.31	20	2.3%	-0.28 [-0.82 , 0.26]
Martínez Manchón 2024[a]	2.22	1.54	43	3.36	2.05	19	2.3%	-0.66 [-1.21 , -0.11]
Newman 2022	11.49	2.89	60	12	3.09	60	2.9%	-0.17 [-0.53 , 0.19]
Şansal 2024	10.55	1.25	60	15.33	7.16	60	2.9%	-0.92 [-1.30 , -0.55]
Schutte 2017	1.12	0.25	13	1.17	0.3	13	1.7%	-0.18 [-0.95 , 0.60]
Song 2022	11.28	2.62	32	12.97	2.2	31	2.4%	-0.69 [-1.20 , -0.18]
Sun 2023[a]	11.62	3.32	42	14.38	5.99	21	2.4%	-0.62 [-1.16 , -0.09]
Suppakittpaisarn 2023	0.183	0.65	134	0.208	0.19	135	3.3%	-0.05 [-0.29 , 0.19]
Valtchanov 2010	1.22	0.35	24	1.45	0.62	22	2.2%	-0.45 [-1.04 , 0.13]
Yang 2024	41.72	7.66	60	48.73	9.19	60	2.9%	-0.82 [-1.20 , -0.45]
Yildirim 2024	-7.66	1.15	32	-6.59	1.62	32	2.4%	-0.75 [-1.26 , -0.24]
Yin 2020	1.6	0.46	75	1.7	0.5	25	2.6%	-0.21 [-0.66 , 0.24]
Zhang 2022	2.95	1.85	38	5.49	1.39	39	2.4%	-1.54 [-2.05 , -1.03]
Zheng 2024[a]	23.63	4.14	40	28.4	2.74	20	2.2%	-1.26 [-1.84 , -0.67]
Total (HKSJ[b])			**2179**			**1845**	**100.0%**	**-0.55 [-0.70 , -0.40]**
95% prediction interval								[-1.32 , 0.23]

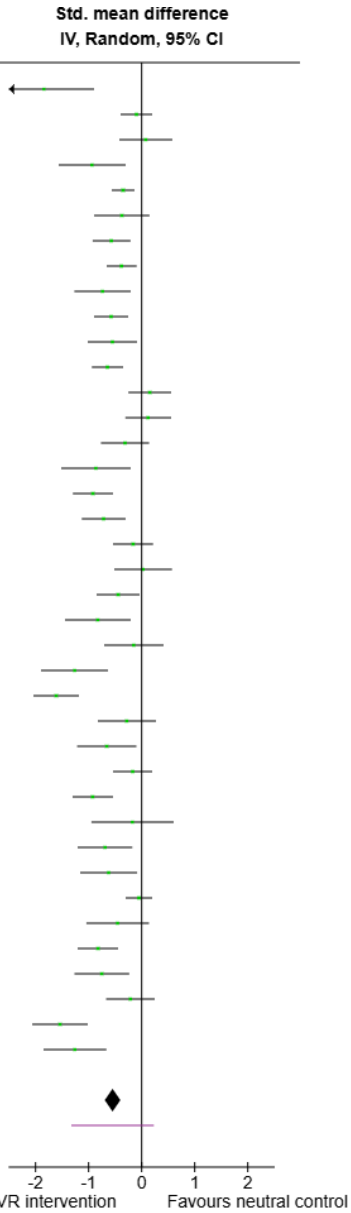

Test for overall effect: T = 7.53, df = 38 (P < 0.00001)

Heterogeneity: Tau² (REML[c]) = 0.14; Chi² = 147.41, df = 38 (P < 0.00001); I² = 76%

Footnotes

[a]Study observed more than one VR interventions of interest. Effects were pooled for main analyses.

[b]CI calculated by Hartung-Knapp-Sidik-Jonkman method.

[c]Tau² calculated by Restricted Maximum-Likelihood method.

XSL•FO
RenderX

Figure 4. Funnel plot for the primary outcome of perceived stress (postintervention) of the meta-analysis on virtual reality interventions for stress reduction in the general population, generated using Cochrane's Review Manager (RevMan), version 9.11.0. SMD: standardized mean difference.

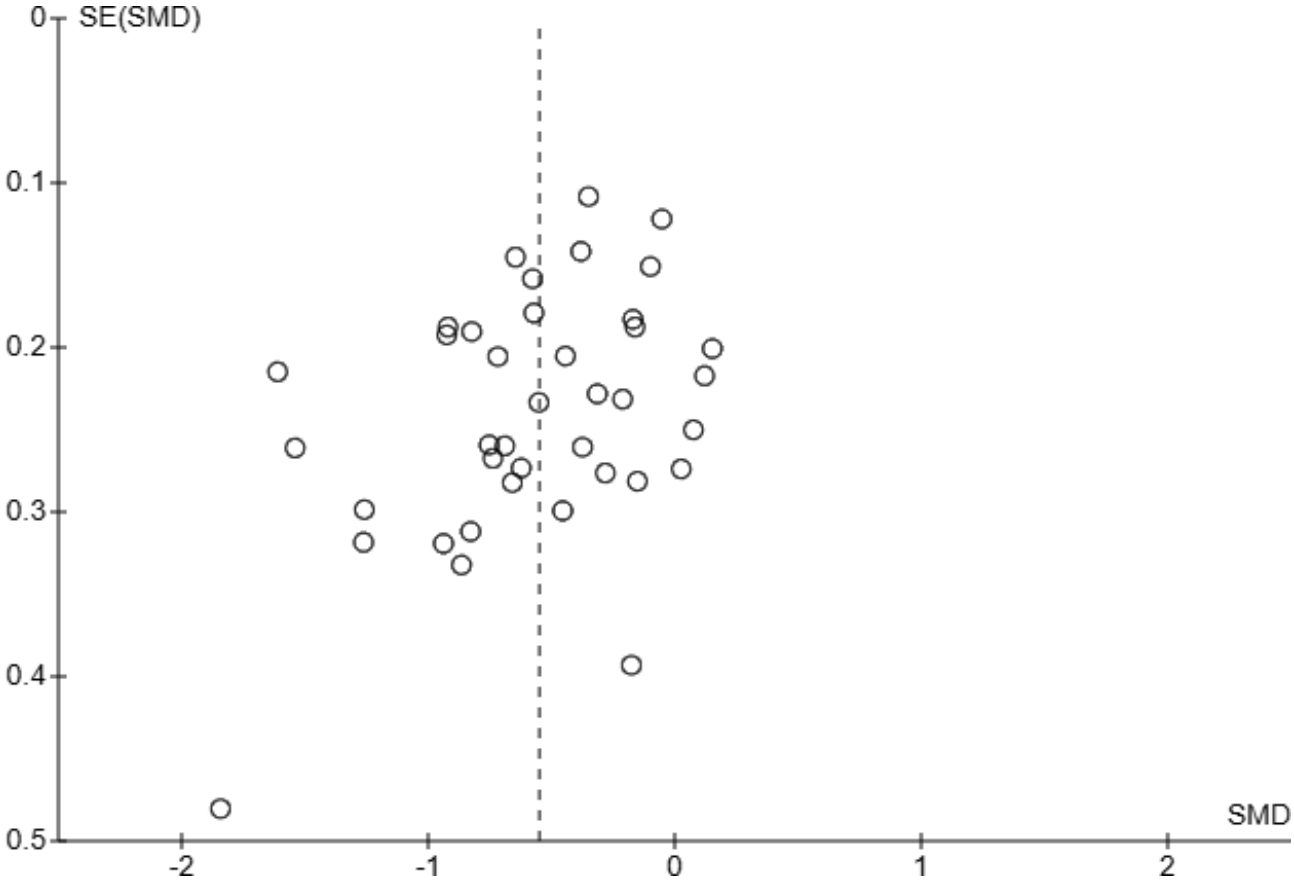

Table . Results of the sensitivity analyses for the primary outcome perceived stress level (postintervention) of the meta-analysis on virtual reality interventions for stress reduction in the general population, investigating effect sizes for randomized controlled trials exclusively with low risk of bias in critical domains, large sample size, or between-subjects study design.

	Random effects, SMD[a] (95% CI)	I^2 value	95% PI[b]
All studies (n=39)	−0.55 (−0.70 to −0.40)	76%	−1.32 to 0.23
Low RoB[c] – selection bias (n=16)	−0.74 (−1.04 to −0.43)	84%	−1.86 to 0.38
Low RoB – attrition bias (n=29)	−0.48 (−0.64 to −0.32)	72%	−1.17 to 0.21
Low or unclear RoB – performance bias (n=25)	−0.47 (−0.63 to −0.31)	64%	−1.07 to 0.13
Large sample size ≥100 (n=12)	−0.57 (−0.83 to −0.31)	83%	−1.41 to 0.27
Output device HMD[d] (n=37)	−0.57 (−0.72 to −0.41)	76%	−1.36 to 0.23
Between-subjects design (n=31)	−0.56 (−0.73 to −0.28)	77%	−1.40 to 0.28

[a]SMD: standardized mean difference.

[b]PI: prediction interval.

[c]RoB: risk of bias.

[d]HMD: head-mounted display.

Objective Stress Measures

A total of 18 studies reported at least one physiological stress parameter (12 trials HR, 10 trials HRV, 8 trials SCL, and 7 trials SYS). Most objective stress outcomes showed no significant intervention effect compared with the neutral control group: HR (SMD −0.08, 95% CI −0.21 to 0.05, *P*=.22; 95% PI −0.26 to 0.11); HRV (SMD 0.23, 95% CI −0.09 to 0.54, *P*=.14; 95% PI −0.59 to 1.04); and SCL (SMD −0.03, 95% CI −0.15 to 0.10, *P*=.60; 95% PI −0.15 to 0.10). A significant intervention effect was observed for SYS (SMD −0.20, 95% CI −0.37 to −0.04, *P*=.02; 95% PI −0.37 to −0.04). Forest plots for the physiological outcomes are provided in Multimedia Appendix 6 [48,49,53-57,59,61,63-68,72,73,75,76,83,93], including RoB assessment.

XSL•FO
RenderX

Secondary Psychological Outcomes

A total of 17 studies reported a measure of positive emotion. To ensure a consistent direction of effects across outcomes, effect sizes were coded such that negative SMDs indicate beneficial effects of VR interventions. The pooled SMD showed a significant effect of VR interventions on positive emotion (−0.65, 95% CI −0.84 to −0.46, *P*<.001), with a moderate level of heterogeneity (I^2=62%, *P*<.001). As the 95% PI (−1.28 to −0.02) reached significance, VR interventions appear to consistently promote positive emotions across participants. Measures of anxiety and depression were reported in 14 and 8 studies, respectively. On average, significant pooled effects were observed for both outcomes: anxiety (SMD −0.88, 95% CI −1.23 to −0.54, *P*<.001; 95% PI −2.07 to 0.31) and depression (SMD −0.34, 95% CI −0.47 to −0.21, *P*<.001; 95% PI −0.47 to −0.21). However, the heterogeneity was high for anxiety (I^2=87%, *P*<.001) and negligible for depression (I^2=0%, *P*=.71). The nonsignificant PI for anxiety suggests that VR interventions did not reduce, or even increase, anxiety in some participants, whereas depression symptoms decreased more consistently. Additionally, 6 studies reported a measure of restoration, for which the pooled average effect was significant (SMD −1.50, 95% CI −2.58 to −0.42, *P*=.02). Again, the nonsignificant 95% PI (95% PI −4.26 to 1.26) indicates even negative effects on restoration for some individuals. Forest plots and effect sizes of each study are provided in Multimedia Appendix 6, including RoB assessment.

Effects of VR-Based Interventions Compared to Active Controls

A total of 7 studies included an active control condition without a VR component (2 trials RWE, 3 trials 2D, 2 trials TSM). The pooled SMD for PSL showed no significant effect of VR interventions in comparison to active controls (SMD −0.70, 95% CI −1.59 to 0.18, *P*=.10; 95% PI −3.12 to 1.71; Figure 5). Five studies also provided data on positive emotion, for which no significant effect of VR interventions was observed when compared to active controls (SMD −0.03, 95% CI −0.80 to 0.73, *P*=.91; 95% PI −1.70 to 1.64).

Figure 5. Forest plot of the meta-analysis on virtual reality interventions for stress reduction in the general population for the primary outcome PSL (postintervention) in comparison to active control condition, generated using Cochrane's Review Manager (RevMan), version 9.11.0 [46,47,50,51,59,60,64]. PSL: perceived stress level; VR: virtual reality.

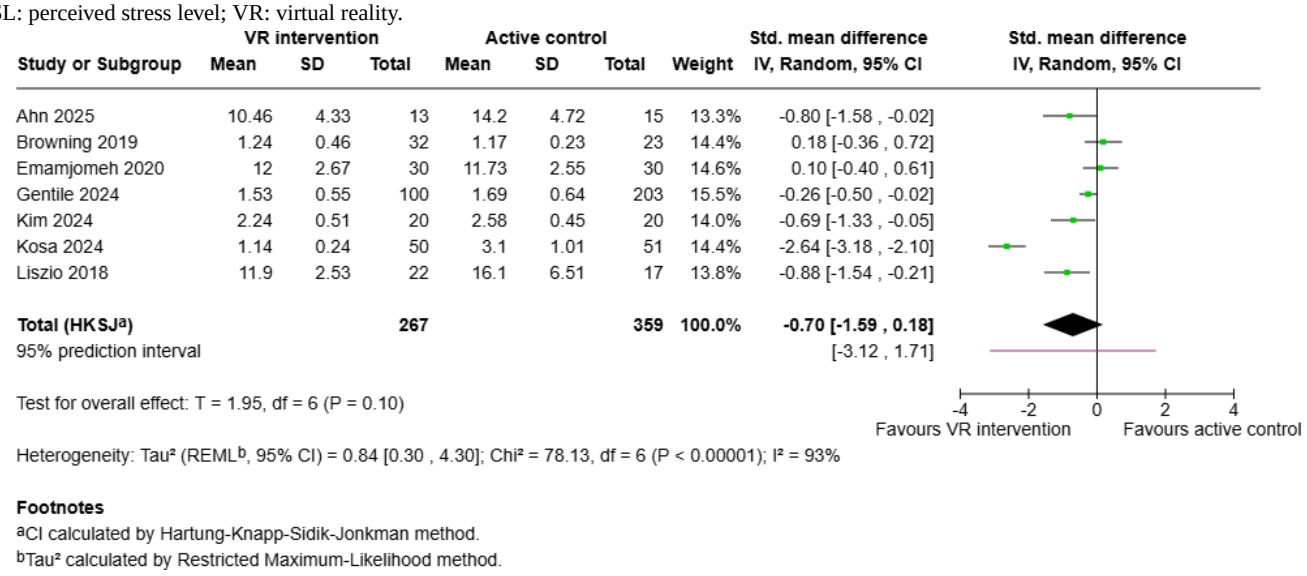

Certainty of Evidence

We identified moderate-quality evidence that VR interventions enhance positive emotion in the general population compared with neutral control conditions (Table 6). The certainty of evidence for PSL and most other psychological outcomes was rated as low to very low. In contrast, all physiological stress parameters showed a moderate or high certainty of evidence.

XSL•FO
RenderX

Table . Summary of findings of the meta-analysis on virtual reality interventions for stress reduction in the general population according to GRADE[a], including certainty of evidence assessment[b].

Outcome	SMD[c] (95% CI)	Participants (studies)	Certainty of the evidence (GRADE)[d]	Comments
Perceived stress level	−0.55 (−0.70 to −0.40)	4024 (39 RCTs)[e]	Very low	In 4 studies, PSL[f] slightly increased on a nonsignificant level, while 35 demonstrated decreased PSL. Nonsignificant PI[g] suggested variety in intervention effects.
Positive emotion	−0.65 (−0.84 to −0.46)	1688 (17 RCTs)	Moderate	Positive emotion was enhanced consistently in all studies, supported by the significant PI.
Anxiety	−0.88 (−1.23 to −0.54)	1472 (14 RCTs)	Very low	All studies showed an average decrease in anxiety.
Depression	−0.34 (−0.47 to −0.21)	936 (8 RCTs)	Low	All studies showed a consistent decrease in depression, supported by the significant PI.
Restoration	−1.50 (−2.58 to −0.42)	440 (6 RCTs)	Very low	—[h]
Heart rate	−0.08 (−0.21 to 0.05)	1398 (12 RCTs)	Moderate	—
Heart rate variability	0.23 (−0.09 to 0.54)	901 (10 RCTs)	Moderate	—
Systolic blood pressure	−0.20 (−0.37 to −0.04)	654 (7 RCTs)	High	6 of 7 studies showed a decrease in SYS[i]. Consistent effects on SYS are underlined by the significant PI.
Skin conductance level	−0.03 (−0.15 to 0.10)	992 (8 RCTs)	Moderate	—

[a]GRADE: Grading of Recommendations Assessment, Development, and Evaluation.

[b]People: adults without psychiatric diagnoses; Settings: general population; Intervention: immersive virtual reality interventions for stress reduction; Comparison: neutral control.

[c]SMD: standardized mean difference.

[d]*High*: This research provides a very good indication of the likely effect. The likelihood that the effect will be *substantially different* is low. *Moderate*: This research provides a good indication of the likely effect. The likelihood that the effect will be *substantially different* is moderate. *Low*: This research provides some indication of the likely effect. However, the likelihood that it will be *substantially different* is high. *Very low*: This research does not provide a reliable indication of the likely effect. The likelihood that the effect will be *substantially different* is very high. *Substantially different* means a large enough difference that it might affect a decision.

[e]RCT: randomized controlled trial.

[f]PSL: perceived stress level.

[g]PI: prediction interval.

[h]Not applicable.

[i]SYS: systolic blood pressure.

Predictors for the Effectiveness of VR-Based Interventions for Psychological Stress Reduction

To identify potential explanatory variables for the stress-reducing effects of VR interventions, several regression analyses were conducted.

Given that only 15 studies provided complete data for all predictors, an overall multivariate regression incorporating all 12 predictors was not feasible due to insufficient statistical power. Consequently, explanatory variables were grouped into study design aspects and technical or content-related characteristics of the VR intervention for the regression analyses. Potential predictors included the percentage of men, mean age of study participants, intervention duration in days, number of VR sessions, length of single VR session, setting, refresh rate, diagonal field of view, level of content motion, environment

XSL•FO
RenderX

realism, level of user interactivity, and intervention type (for details, see Table 2).

Results of the univariate linear regression analyses of study-design related factors as potential predictors for the intervention effects are summarized in Table 7. Including 32 studies with a total of 3023 participants (SMD −0.56, 95% CI −0.73 to −0.39, P<.001; 95% PI −1.38 to 0.26; I^2=76%), no significant effects were found for age (P=.36) or gender (P=.32). Multisession interventions predicted larger stress-reducing effects (R^2=15.0%, $F_{1,30}$=6.66; P=.02) compared with single-session approaches. No significant associations were identified for intervention duration in days (P=.26), number of VR sessions (P=.91), or length of individual sessions (P=.09). Multivariate regression analyses (Table S1 in Multimedia Appendix 7) did not identify any significant predictors, which may be attributable to interdependencies among single factors.

Table . Univariate linear regression analyses for possible predictors of the effectiveness of virtual reality (VR) interventions based on participants and intervention design (n=32) in the framework of the meta-analysis on VR interventions for stress reduction in the general population[a,b].

Predictor	Estimate	SE	2-tailed *t* test (*df*)	*P* value	*F* test (*df*)	τ^2	R^2
Mean age (years)	0.01	0.01	0.94 (30)	.36	0.88 (1, 30)	0.005	0%
Males (%)	0.01	0.01	1.02 (30)	.32	1.04 (1, 30)	0.005	0%
Clinical setting (1,0)[c]	−0.12	0.25	−0.48 (30)	.64	0.23 (1, 30)	0.005	0%
Intervention duration (days)	−0.00	0.00	−1.15 (30)	.26	1.33 (1, 30)	0.005	1.1%
Multisession intervention (1,0)[d]	−0.50	0.19	−2.58 (30)	.02	6.66 (1, 30)	0.004	15%
Number of VR sessions	−0.00	0.02	−0.12 (30)	.91	0.01 (1, 30)	0.005	0%
Single VR session length (min)	−0.02	0.01	−1.75 (30)	.09	3.05 (1, 30)	0.005	6%

[a]Analyses were conducted in IBM SPSS Statistics, version 30.0, using a random-effects model with HKSJ (Hartung, Knapp, Sidik, and Jonkman) correction.

[b]Effect sizes of different VR interventions of one publication were pooled for the main analyses and are included as one intervention group [56,65,68,73].

[c]Clinical setting was dummy coded as “1=clinical setting” and “0=nonclinical setting.”

[d]Intervention duration was dummy coded as “1=multisession intervention” and “0=single-session intervention.”

We therefore performed explorative chi-square tests (Tables S1 and S3 in Multimedia Appendix 3), which revealed codependencies between the predictors age, setting, and intervention duration. Specifically, clinical settings were more likely to involve multisession interventions (χ^2_2=14.49, P<.001) and to include older participants (χ^2_2=9.41, P=.009).

Next, the explanatory role of content-related and technical characteristics of the VR intervention (eg, content motion, user interactivity, environment realism, field of view, and refresh rate) on the stress-reducing effects of VR interventions was investigated in linear regression analyzes based on a subset of 17 studies with a total of 1831 participants (SMD −0.37, 95% CI −0.61 to −0.14, P=.004; I^2=73%; 95% PI −1.13 to 0.38). Univariate regression analyses (Table 8) identified significant results for image refresh rate (R^2=42.1%, $F_{1,15}$=12.91, P=.003) as well as real-world recordings (R^2=23.8%, $F_{1,15}$=6.17, P=.03) in comparison to computer-simulated content. The multivariate model (Table S2 in Multimedia Appendix 7) showed significance and demonstrated a predictive power of 42.9% (R^2=42.9%, $F_{5,11}$=3.43, P=.04). Within this model, a higher image refresh rate (t_{15}=−2.36, 2-tailed; P=.04) was identified as a potential independent predictor of the intervention effects on PSL. Environment realism (P=.43) did not remain a significant predictor, which may be attributable to interdependencies between the environment realism and intervention type, showing real-world recordings more likely in nature exposure interventions in comparison to other types (χ^2_3=17.09, P<.001) with more passive levels of user interactivity (Tables S2 and S4 in Multimedia Appendix 3).

XSL•FO
RenderX

Table . Univariate linear regression analyses for possible predictors of the intervention effectiveness based on content-related and technical aspects (n=17), performed in the framework of the meta-analysis on virtual reality interventions for stress reduction in the general population[a].

Predictor	Estimate	SE	2-tailed *t* test (*df*)	*P* value	*F* test (*df*)	τ^2	R^2
Content motion[b]	−0.13	0.30	−0.42 (15)	.68	0.18 (1, 15)	0.01	0%
User interactivity[c]	−0.15	0.27	−0.55 (15)	.59	0.30 (1, 15)	0.01	0%
Environment realism[d]	−0.73	0.29	−2.49 (15)	.03	6.17 (1, 15)	0.01	23.8%
Refresh rate (Hz)	−0.02	0.01	−3.59 (15)	.003	12.91 (1, 15)	0.01	42.1%
FoV[e] (degree)	−0.02	0.02	−0.94 (15)	.36	0.88 (1, 15)	0.01	0%

[a]Analyses were conducted in IBM SPSS Statistics, version 30.0, using a random-effects model with HKSJ (Hartung, Knapp, Sidik, and Jonkman) correction.

[b]Content motion was categorized as “1=no motion” (simple 3D images without moving elements), “2=low motion” (static scene with some moving elements), or “3=high motion” (dynamic scene with changing perspective).

[c]User interactivity was ranked as “1=low control” (only watching and exploring scene by head movement), “2=medium control” (option to change perspective while moving in environment by controller or body movement), or “3=high control” (option to interact with elements due to an interactive task or game).

[d]Environment realism was dummy coded as dichotomic variable “real-world recording” (1=yes, 0=no) for regression analyses.

[e]Diagonal field of view (FoV) was calculated by the following formula: (horizontal fov)2+(vertical fov)2

Discussion

Summary of Findings

This systematic review synthesizes evidence on the effectiveness of various VR-based interventions for stress reduction across a broad range of application areas and identifies characteristics that may predict this effectiveness. The meta-analyses demonstrated that VR-based interventions led to significantly greater reductions in PSL, anxiety, and depression as well as increases in positive affect, compared to neutral control conditions. However, the nonsignificant PIs for reductions in negative affect pointed to substantial heterogeneity, supporting the notion that VR interventions do not improve psychological stress in all individuals similarly. In contrast, significant PIs for positive affect and depressive symptoms suggest more consistent beneficial effects in these domains. Regarding physiological stress markers, no significant effects were observed for most parameters (eg, HR, HRV, or SCL). Nevertheless, a significant reduction was found for SYS, with significant PIs indicating a robust and consistent intervention effect. Beyond establishing the overall effectiveness of VR interventions in reducing PSL, this meta-analysis also identified several potential predictors of enhanced intervention outcomes. Specifically, multisession interventions, higher image refresh rates, and increased environmental realism, particularly through real-world recordings, were positively associated with stronger stress-reducing effects of VR interventions in univariate regression analyses.

Most of the included studies were published within the past 5 years, reflecting the rapidly growing interest in the use of VR for managing psychological stress. Among the various intervention types identified, virtual exposure to natural environments emerged as the most frequently investigated approach. These interventions were predominantly passive in nature, offering limited interactivity, although some incorporated additional elements such as guided meditation or interactive mini-games to enhance their stress-reducing effects. The majority of studies used HMDs as a delivery method for VR content, were conducted in university or workplace settings, and evaluated single-session interventions. In contrast, RCTs evaluating long-term interventions or targeting specific populations with specific needs, such as pregnant women or older adults, were less common and warrant further investigation. Overall, the findings of this systematic review highlight the considerable diversity of VR-based interventions for stress reduction, characterized by substantial heterogeneity in target populations, methodological and technical design, as well as in reported levels of immersion, presence, and user experience.

Interpretation and Implications of Findings

While previous qualitative literature reviews highlighted the general effectiveness of VR interventions for stress reduction [14,24,26-28,35], the present review extends those findings by quantifying intervention effects for stress reduction and statistically identifying possible predictors of intervention success. Chen et al [34] noted that existing meta-analyses have primarily focused on clinical populations, which motivated their investigation of virtual nature exposure and 2D images on anxiety, stress, and depression in healthy adults. In contrast, our analysis included a broader range of VR-based approaches, especially those delivered through immersive technologies. The meta-analysis of 39 RCTs confirmed that VR interventions can effectively reduce psychological stress in the general population. However, the wide and nonsignificant PIs indicate substantial variability in intervention effects and suggest that not all subgroups may benefit equally in practice. One possible explanation for this heterogeneity is the occurrence of

XSL•FO
RenderX

cybersickness [101], which has been associated with increases in physiological stress markers (eg, cortisol levels and HR) [102], elevated psychological stress [103], and diminished intervention effects (eg, on anxiety) [104]. As cybersickness is likely influenced by characteristics of the VR content [105], it may contribute, together with the heterogeneity of intervention types, to the observed variability in PSL reduction. However, as only 4 of the included studies assessed cybersickness, its impact on intervention outcomes could not be systematically evaluated.

Univariate regression analyses further indicated that multisession interventions were associated with greater effectiveness compared to single-session interventions. This finding is consistent with previous research suggesting that longer-term interventions may be more beneficial for stress reduction [106]. One possible explanation is that studies using multisession designs more frequently targeted populations with elevated baseline stress levels, such as clinical samples [52,55,58,62,72,77] or individuals in workplace and university settings [46,48,57,59,74,100]. In contrast, single-session studies often involved experimentally induced, transient stress in younger participants [56,64,69,79,83,93]. This distinction is relevant, as existing evidence indicates that individuals with chronic conditions [107,108] and university students [109-111] are particularly susceptible to higher and more persistent stress levels. Moreover, while chronic stress typically requires repeated intervention exposure, acute stress may respond to brief relaxation techniques [112,113]. Accordingly, the optimal number, duration, and frequency of VR sessions likely depend on the nature of stress (chronic vs acute) and remain to be clarified by future research.

Notably, age and intervention setting were not identified as significant predictors in our regression analyses. Similarly, Li et al [29] reported no significant influence of age or session duration on the effectiveness of VR-based nature exposure on positive affect, while highlighting an overrepresentation of younger adults in the existing literature. Nevertheless, other studies suggest that older adults may represent a particularly promising target group for VR-based stress reduction [114], potentially benefiting even more than college students [115].

We also examined methodological and technical characteristics of VR content as potential predictors of VR intervention success. Our analyses indicated that the type of environment (eg, blue vs green space) did not significantly influence stress-reducing effects of VR interventions, consistent with the findings of the narrative review by Li et al [116]. In contrast, univariate regression analyses identified environmental realism (eg, computer simulation vs real-world video recordings) as a significant predictor, with stronger stress reduction observed for real-world recordings. However, this effect did not persist in multivariate analyses, possibly due to interdependencies with other factors such as intervention type and user interactivity. Although user interactivity did not emerge as a significant predictor, the observed effect size suggests that a more passive user experience may be associated with greater reductions in PSL. This is in line with findings by Reese et al [117], who reported stronger stress-reducing effects for passive VR nature exposure compared to active engagement within the VR environment. At the same time, the high heterogeneity of effects in studies involving active tasks, along with the limited representation of highly interactive interventions in our dataset, may have reduced statistical power and contributed to the lack of significance [118]. Moreover, previous research indicates that higher interactivity may enhance positive affect [97], suggesting that its impact may differ depending on the outcome domain (eg, negative vs positive emotions) as well as individual user preferences. Accordingly, the role of user interactivity and environmental realism warrants further investigation. The level of content motion was not identified as a significant predictor in our analyses, although individual studies have reported beneficial effects of dynamic content on physiological outcomes such as heart rate [119], as well as higher perceived presence for video-based compared to slideshow-based VR environments [120,121]. Moreover, this meta-analysis identified refresh rate as a promising predictor of stress-reducing effects of VR. However, the refresh rate considered in this meta-analysis reflects only the technically available frequency of image updates of the used HMD, whereas the actual rate at which images appear on the display (framerate) depends on the rendering of raw recordings into the VR content [122], for which insufficient information was available. As a higher refresh rate is generally associated with a better user experience, which is closely linked to interactivity and content motion [95,98,99], the observed interrelations between these variables are to be expected. It is therefore plausible that a smoother and more coherent user experience plays a crucial role for the effectiveness of VR-based stress interventions. Overall, given the substantial heterogeneity across studies, further research is needed to better understand predictors of intervention effects, particularly with regard to potential differences between chronic stress and acute mental load. Although the multivariate model showed significant predictive power, it did not fully account for the variability in outcomes, indicating that additional, as yet unidentified factors are likely to influence intervention success.

Significant effects of VR interventions were also observed for secondary outcomes, particularly increased positive affect and reductions in anxiety and depressive symptoms. Notably, the PIs for positive emotion and depression were significant, suggesting that these beneficial effects are relatively consistent across individuals. In contrast, nonsignificant PIs for other negative emotional outcomes (eg, PSL or anxiety) indicate that improvements in these domains may depend on additional factors. VR interventions were further associated with a significant reduction in SYS, with significant PIs indicating consistent effects across participants. These findings are in line with previous research, such as Yao et al [13], which demonstrated stress-reducing effects of VR interventions on both psychological and physiological parameters. Moreover, the beneficial impact of (virtual) nature exposure on blood pressure has been summarized in prior reviews [123,124]. While Song et al [124] also reported positive effects on HR and HRV, no significant effects on these or other physiological stress markers were observed in our analyses. Several factors may explain this discrepancy. First, the number of studies reporting physiological outcomes was relatively small, potentially limiting statistical power. Second, multisession interventions, which tend to yield stronger effects, were underrepresented in the

XSL•FO
RenderX

available data. Third, substantial variability in the assessment and reporting of physiological measures, as well as potential limitations in measurement sensitivity, may have obscured detectable effects. The considerable heterogeneity in physiological outcomes is also reflected in previous studies [125,126], which attribute inconsistencies to differences in study design and population characteristics [126]. Overall, our findings are consistent with prior meta-analyses reporting reductions in PSL and increases in positive affect following VR interventions, alongside nonsignificant effects on cardiovascular parameters such as HR and HRV when compared to neutral control conditions [29,115].

Consistent with our finding of no substantial differences between VR interventions and active control conditions, such as 2D simulations or real-world nature exposure, previous studies suggest that VR-based nature exposure may be as effective as real-life exposure. For instance, Reese et al [127] reported similar improvements in well-being and PSL following both digital forest bathing and in vivo nature exposure. Likewise, the meta-analysis by Fan and Baharum [128] confirmed VR-based nature exposure as a viable and nearly equivalent alternative to real-world nature experience for stress reduction.

Limitations

Several limitations should be considered when interpreting the results of this meta-analysis. First, no information scientist was involved in developing the search strategy, and relevant studies may therefore have been missed. However, the search followed a systematic and transparent strategy based on the Cochrane guidelines and the PICO framework, and the final search string was pretested and refined according to the PRESS guidelines by 2 independent researchers. Second, due to the nature of VR intervention studies, blinding of group allocation is rarely feasible, and only 3 studies implemented blinded control conditions. As a result, potential biases related to placebo effects or response bias cannot be fully excluded. However, evidence from VR-based pain management suggests that expectation effects may be limited in this context [129]. Third, many included studies exhibited a high risk of selection, performance, and detection bias, highlighting methodological limitations within the existing literature. Fourth, although strict inclusion criteria were applied to reduce heterogeneity, considerable variability between studies remained. Analyses could not identify more homogeneous subsets of studies with more robust effects regarding stress reduction through VR interventions. Prediction intervals were often nonsignificant, and I^2 values ranged from moderate to high, limiting the generalizability of the findings across populations. Fifth, the chi-square test indicated associations between several predictors, suggesting potential confounding. For example, multisession interventions were predominantly conducted in clinical settings and involved participants over 40 years of age. These interdependencies should be taken into account when interpreting the predictive value of related factors such as intervention duration in univariate analyses. Sixth, interventions involving active user engagement, such as interactive tasks or the use of controllers, remain underrepresented in the current literature, with the majority of interventions offering a passive VR experience [14]. Consequently, the stress-reducing potential of more interactive VR interventions cannot yet be adequately assessed. Seventh, limited reporting of participant characteristics (eg, prior VR experience) and insufficient detail on intervention features restricted the feasibility of conducting more robust subgroup and regression analyses. As a result, the statistical power of predictor analyses was limited and should be considered when interpreting these findings. Eighth, data on potential predictors were inconsistently reported, hampering a more comprehensive examination of their influence. It is therefore possible that additional unmeasured factors contribute to intervention effectiveness, which may explain the relatively low proportion of explained variance observed in the analyses.

Conclusions

While previous reviews have demonstrated the general effectiveness of VR interventions for stress reduction without quantifying their effects [14,24,26-28,35], or have focused on specific subgroups [32,130,131] or intervention types [29,30,128], this review highlights the considerable potential of VR-based interventions to reduce psychological stress in the general population. By quantifying effects across a broad range of approaches and, for the first time, directly comparing diverse intervention formats, this meta-analysis provides a more comprehensive understanding of their effectiveness. Given that VR-based interventions appear comparable in effectiveness to established stress management techniques, they represent a promising alternative, particularly in contexts where access to real-life nature or conventional interventions is limited or resource-intensive. While the findings suggest reliable improvements in positive affect and reductions in depressive symptoms, effects on PSLs were more variable, with some individuals experiencing limited benefit. This variability was observed in nearly all subgroups, suggesting that individual characteristics and contextual factors play a critical role in shaping intervention outcomes. Multisession interventions generally yielded stronger effects than single-session designs, whereas intervention duration and total number of VR sessions did not emerge as significant predictors. Thus, the optimal number, duration, and frequency of VR sessions remain unclear and may differ depending on whether stress is acute or chronic. Future studies with larger and more diverse samples are needed to directly compare effects across target groups and application contexts, including short-term stress induction and everyday stress experiences.

This meta-analysis further identified characteristics associated with a more calming user experience, such as higher refresh rates and lower levels of user interactivity, as potential predictors of stronger stress-reducing effects. These findings may inform the design of future VR approaches for stress reduction by highlighting the importance of a high-quality and coherent user experience. However, individual preferences regarding content motion and user interactivity in VR suggest that customizable VR environments may further enhance effectiveness. Advances in VR hardware and content development could facilitate such personalization, raising important questions about how to implement these as accessible, low-threshold tools for everyday stress management.

XSL•FO
RenderX

Despite these insights, a substantial proportion of the variance in intervention outcomes remains unexplained, indicating that the stress-reducing effects of VR are shaped by multiple interacting factors, including participant characteristics, intervention design, and technical features. Future RCTs should therefore systematically assess and report these variables to validate current findings, explore interdependencies, and establish evidence-based guidelines for optimizing VR-based stress reduction across diverse populations.

Funding

This research is part of the Collaborative Research Project "AI4Coping – Artificial Intelligence for Sustainable Support in Stress Management," funded by the Bavarian State Ministry of Science and the Art (grant H.2-F1116.CO/52/2; to KM).

Data Availability

All data generated or analyzed during this study are included in this published article and in Multimedia Appendices 1-8.

Authors' Contributions

HS and KM designed the study. HS reached, selected, extracted, and assessed the literature, performed the statistical analyses, and summarized qualitative and quantitative results. IS supported study selection. JG and SK supported categorizing information on technical devices. KM contributed to data extraction, risk of bias assessment, and statistical analyses. SS supported planning and conducting the statistical analyses. HS and KM drafted the manuscript, which was critically reviewed by SS and IS. All authors read and approved the final manuscript. There was no generative AI used in any portion of the manuscript writing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete search strings.
[DOCX File, 24 KB - jmir_v28i1e78212_app1.docx]

Multimedia Appendix 2
Overview on the different domains and applied labels.
[DOCX File, 22 KB - jmir_v28i1e78212_app2.docx]

Multimedia Appendix 3
Cross tables and chi-square test for variable distribution in separate subgroups.
[DOCX File, 55 KB - jmir_v28i1e78212_app3.docx]

Multimedia Appendix 4
Excluded studies from this systematic review.
[XLSX File, 62 KB - jmir_v28i1e78212_app4.xlsx]

Multimedia Appendix 5
Meta-analyses of different subgroups.
[DOCX File, 2403 KB - jmir_v28i1e78212_app5.docx]

Multimedia Appendix 6
Forest and funnel plots for secondary outcomes.
[DOCX File, 1359 KB - jmir_v28i1e78212_app6.docx]

Multimedia Appendix 7
Multivariate linear regression analyses and bubble plots.
[DOCX File, 194 KB - jmir_v28i1e78212_app7.docx]

Multimedia Appendix 8
Raw data analyzed in meta-analysis.
[XLSX File, 55 KB - jmir_v28i1e78212_app8.xlsx]

XSL•FO
RenderX

Checklist 1
PRISMA checklist.
[DOCX File, 38 KB - jmir_v28i1e78212_app9.docx]

Checklist 2
PRISMA-S checklist.
[DOCX File, 24 KB - jmir_v28i1e78212_app10.docx]

References

1. Werdecker L, Esch T. Stress und Gesundheit. In: Haring R, editor. Gesundheitswissenschaften. Springer Reference Pflege – Therapie – Gesundheit: Springer; 2019. [doi: 10.1007/978-3-662-58314-2_33]
2. Barry V, Stout ME, Lynch ME, et al. The effect of psychological distress on health outcomes: a systematic review and meta-analysis of prospective studies. J Health Psychol 2020 Feb;25(2):227-239. [doi: 10.1177/1359105319842931] [Medline: 30973027]
3. Bamert M, Schaffter Y, Bally F, O'Connor DB, Inauen J. Assessment of stress and its relationship with health behaviour in daily life: a systematic review. Health Psychol Rev 2025 Dec;19(4):906-936. [doi: 10.1080/17437199.2025.2534000] [Medline: 40844372]
4. Walther L, Junker S, Thom J, Hölling H, Mauz E. High-frequency surveillance of mental health indicators in the adult population of Germany: trends from 2022 to 2023. Dtsch Arztebl Int 2023 Oct 27;120(43):736-737. [doi: 10.3238/arztebl.m2023.0180] [Medline: 37997940]
5. Mental health report. : Publications Office of the European Union; 2023. [doi: 10.2875/48999]
6. Richter EP, Brähler E, Zenger M, et al. Compounded effects of multiple global crises on mental health: a longitudinal study of East German adults. J Clin Med 2024 Aug 13;13(16):4754. [doi: 10.3390/jcm13164754] [Medline: 39200894]
7. Kaman A, Devine J, Erhart M, et al. Youth mental health in times of global crises: evidence from the German longitudinal COVID-19 and psychological health study. J Adolesc Health 2025 Dec;77(6):1050-1057. [doi: 10.1016/j.jadohealth.2025.05.010] [Medline: 40693987]
8. Khoury B, Sharma M, Rush SE, Fournier C. Mindfulness-based stress reduction for healthy individuals: a meta-analysis. J Psychosom Res 2015 Jun;78(6):519-528. [doi: 10.1016/j.jpsychores.2015.03.009] [Medline: 25818837]
9. Muhammad Khir S, Wan Mohd Yunus WMA, Mahmud N, et al. Efficacy of progressive muscle relaxation in adults for stress, anxiety, and depression: a systematic review. Psychol Res Behav Manag 2024;17:345-365. [doi: 10.2147/PRBM.S437277] [Medline: 38322293]
10. Seo E, Kim S. Effect of autogenic training for stress response: a systematic review and meta-analysis. J Korean Acad Nurs 2019 Aug;49(4):361-374. [doi: 10.4040/jkan.2019.49.4.361] [Medline: 31477667]
11. Varvogli L, Darviri C. Stress management techniques: evidence-based procedures that reduce stress and promote health. Health Sci J 2011(2):74-89 [FREE Full text]
12. Shuda Q, Bougoulias ME, Kass R. Effect of nature exposure on perceived and physiologic stress: a systematic review. Complement Ther Med 2020 Sep;53:102514. [doi: 10.1016/j.ctim.2020.102514] [Medline: 33066853]
13. Yao W, Zhang X, Gong Q. The effect of exposure to the natural environment on stress reduction: a meta-analysis. Urban For Urban Green 2021 Jan;57:126932. [doi: 10.1016/j.ufug.2020.126932]
14. Ladakis I, Filos D, Chouvarda I. Virtual reality environments for stress reduction and management: a scoping review. Virtual Real 2024 Mar;28(1). [doi: 10.1007/s10055-024-00943-y]
15. Pawassar CM, Tiberius V. Virtual reality in health care: bibliometric analysis. JMIR Serious Games 2021 Dec 1;9(4):e32721. [doi: 10.2196/32721] [Medline: 34855606]
16. Hamad A, Jia B. How virtual reality technology has changed our lives: an overview of the current and potential applications and limitations. IJERPH 2022 Sep 8;19(18):11278. [doi: 10.3390/ijerph191811278]
17. van Loenen I, Scholten W, Muntingh A, Smit J, Batelaan N. The effectiveness of virtual reality exposure-based cognitive behavioral therapy for severe anxiety disorders, obsessive-compulsive disorder, and posttraumatic stress disorder: meta-analysis. J Med Internet Res 2022 Feb 10;24(2):e26736. [doi: 10.2196/26736] [Medline: 35142632]
18. Tan YL, Chang VYX, Ang WHD, Ang WW, Lau Y. Virtual reality exposure therapy for social anxiety disorders: a meta-analysis and meta-regression of randomized controlled trials. Anxiety Stress Coping 2025 Mar;38(2):141-160. [doi: 10.1080/10615806.2024.2392195] [Medline: 39161295]
19. Horigome T, Kurokawa S, Sawada K, et al. Virtual reality exposure therapy for social anxiety disorder: a systematic review and meta-analysis. Psychol Med 2020 Nov;50(15):2487-2497. [doi: 10.1017/S0033291720003785] [Medline: 33070784]
20. Deng W, Hu D, Xu S, et al. The efficacy of virtual reality exposure therapy for PTSD symptoms: a systematic review and meta-analysis. J Affect Disord 2019 Oct 1;257:698-709. [doi: 10.1016/j.jad.2019.07.086] [Medline: 31382122]
21. Kothgassner OD, Goreis A, Kafka JX, Van Eickels RL, Plener PL, Felnhofer A. Virtual reality exposure therapy for posttraumatic stress disorder (PTSD): a meta-analysis. Eur J Psychotraumatol 2019 Dec 31;10(1):1654782. [doi: 10.1080/20008198.2019.1654782]

XSL•FO
RenderX

22. Carl E, Stein AT, Levihn-Coon A, et al. Virtual reality exposure therapy for anxiety and related disorders: a meta-analysis of randomized controlled trials. J Anxiety Disord 2019 Jan;61:27-36. [doi: 10.1016/j.janxdis.2018.08.003]
23. Emmelkamp PMG, Meyerbröker K. Virtual reality therapy in mental health. Annu Rev Clin Psychol 2021 May 7;17(1):495-519. [doi: 10.1146/annurev-clinpsy-081219-115923] [Medline: 33606946]
24. Li Pira G, Aquilini B, Davoli A, Grandi S, Ruini C. The use of virtual reality interventions to promote positive mental health: systematic literature review. JMIR Ment Health 2023 Jul 6;10:e44998. [doi: 10.2196/44998] [Medline: 37410520]
25. Riches S, Jeyarajaguru P, Taylor L, et al. Virtual reality relaxation for people with mental health conditions: a systematic review. Soc Psychiatry Psychiatr Epidemiol 2023 Jul;58(7):989-1007. [doi: 10.1007/s00127-022-02417-5] [Medline: 36658261]
26. Riches S, Azevedo L, Bird L, Pisani S, Valmaggia L. Virtual reality relaxation for the general population: a systematic review. Soc Psychiatry Psychiatr Epidemiol 2021 Oct;56(10):1707-1727. [doi: 10.1007/s00127-021-02110-z] [Medline: 34120220]
27. Velana M, Sobieraj S, Digutsch J, Rinkenauer G. The advances of immersive virtual reality interventions for the enhancement of stress management and relaxation among healthy adults: a systematic review. Appl Sci (Basel) 2022 Jul 20;12(14):7309. [doi: 10.3390/app12147309]
28. Xu J, Khanotia A, Juni S, et al. Effectiveness of virtual reality-based well-being interventions for stress reduction in young adults: systematic review. JMIR Ment Health 2024 Mar 29;11:e52186. [doi: 10.2196/52186] [Medline: 38551625]
29. Li H, Ding Y, Zhao B, Xu Y, Wei W. Effects of immersion in a simulated natural environment on stress reduction and emotional arousal: a systematic review and meta-analysis. Front Psychol 2022;13(1058177). [doi: 10.3389/fpsyg.2022.1058177]
30. Cortez-Vázquez G, Adriaanse M, Burchell GL, Ostelo R, Panayiotou G, Vlemincx E. Virtual reality breathing interventions for mental health: a systematic review and meta-analysis of randomized controlled trials. Appl Psychophysiol Biofeedback 2024 Mar;49(1):1-21. [doi: 10.1007/s10484-023-09611-4] [Medline: 38236355]
31. Zhang T, Ren Z, Wakefield CE, et al. Are digital psychological interventions for psychological distress and quality of life in cancer patients effective? A systematic review and network meta-analysis. Clin Psychol Rev 2025 Feb;115:102520. [doi: 10.1016/j.cpr.2024.102520] [Medline: 39615074]
32. Turan Kavradim S, Yangöz Ş, Özer Z. Effectiveness of virtual reality interventions on physiological and psychological outcomes of adults with cardiovascular disease: A systematic review and meta-analysis. J Nurs Scholarsh 2023 Sep;55(5):949-966. [doi: 10.1111/jnu.12885] [Medline: 36942847]
33. Riches S, Taylor L, Jeyarajaguru P, Veling W, Valmaggia L. Virtual reality and immersive technologies to promote workplace wellbeing: a systematic review. J Ment Health 2024 Apr;33(2):253-273. [doi: 10.1080/09638237.2023.2182428] [Medline: 36919828]
34. Chen L, Yan R, Yu J. Virtual nature, real relief: how exposure to virtual natural environments reduces anxiety, stress, and depression in healthy adults. NPJ Digit Med 2025 Nov 18;8(1):679. [doi: 10.1038/s41746-025-02057-4] [Medline: 41254294]
35. Meshkat S, Edalatkhah M, Di Luciano C, et al. Virtual reality and stress management: a systematic review. Cureus 2024 Jul;16(7):e64573. [doi: 10.7759/cureus.64573] [Medline: 39144853]
36. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev 2021 Dec;10(1). [doi: 10.1186/s13643-021-01626-4]
37. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z]
38. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-926. [doi: 10.1136/bmj.39489.470347.AD]
39. Lefebvre C, Glanville J, Briscoe S, et al. Technical supplement to Chapter 4: Searching for and selecting studies. In: Higgins JP, Thomas J, Cumpston M, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions 2022. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-04-technical-supplement-searching-and-selecting-studies [accessed 2026-05-01]
40. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS Peer Review of Electronic Search Strategies: 2015 guideline statement. J Clin Epidemiol 2016 Jul;75:40-46. [doi: 10.1016/j.jclinepi.2016.01.021]
41. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan – a web and mobile app for systematic reviews. Syst Rev 2016 Dec 5;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
42. Higgins J, Green S. Cochrane Handbook for Systematic Reviews of Interventions: Wiley; 2011.
43. Zhang L, Bowman DA, Jones CN. Exploring effects of interactivity on learning with interactive storytelling in immersive virtual reality. In: 2019 11th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games): IEEE; 2019. [doi: 10.1109/VS-Games.2019.8864531]
44. Gras RA. Categorising virtual reality content. In: Jung T, Tom Dieck MC, Correia Loureiro SM, editors. Extended Reality and Metaverse: Springer; 2023:122-133. [doi: 10.1007/978-3-031-25390-4_11]
45. Cohen J. Statistical Power Analysis for the Behavioral Sciences: Routledge; 2013. [doi: 10.4324/9780203771587]

XSL•FO
RenderX

46. Ahn J, Kim J, Park Y, Kim R, Choi H. Nature-based virtual reality relaxation to improve mental health and sleep in undergraduate students: a randomized controlled trial. Digit Health 2025;11:20552076251365140. [doi: 10.1177/20552076251365140] [Medline: 40799837]
47. Browning M, Mimnaugh KJ, van Riper CJ, Laurent HK, LaValle SM. Can simulated nature support mental health? Comparing short, single-doses of 360-degree nature videos in virtual reality with the outdoors. Front Psychol 2019;10:2667. [doi: 10.3389/fpsyg.2019.02667] [Medline: 32010003]
48. Chand K, Chandra S, Dutt V. Raga Bhairavi in virtual reality reduces stress-related psychophysiological markers. Sci Rep 2024 Oct 22;14(1):24816. [doi: 10.1038/s41598-024-74932-1] [Medline: 39438543]
49. Chen D, Yin J, Yu CP, Sun S, Gabel C, Spengler JD. Physiological and psychological responses to transitions between urban built and natural environments using the cave automated virtual environment. Landsc Urban Plan 2024 Jan;241:104919. [doi: 10.1016/j.landurbplan.2023.104919]
50. Emamjomeh A, Zhu Y, Beck M. The potential of applying immersive virtual environment to biophilic building design: A pilot study. Journal of Building Engineering 2020 Nov;32:101481. [doi: 10.1016/j.jobe.2020.101481]
51. Gentile DA, Kim EL. Comparing virtual reality, video, and audio-guided meditations in fostering positive attitudes toward meditation. Mindfulness (N Y) 2024 Mar;15(3):586-600. [doi: 10.1007/s12671-024-02311-y]
52. Hessabi M, Sajjadi M, Shareinia H, Rouhani M. The effect of virtual reality on anxiety in patients admitted to the cardiac care unit. Int J Pharm Phytopharm Res 2020;10:37-42. [doi: 10.51847/PDul5le]
53. Hooyberg A, Michels N, Roose H, et al. The psychophysiological reactivity to beaches vs. to green and urban environments: insights from a virtual reality experiment. J Environ Psychol 2023 Nov;91:102103. [doi: 10.1016/j.jenvp.2023.102103]
54. Huang Q, Yang M, Jane HA, Li S, Bauer N. Trees, grass, or concrete? The effects of different types of environments on stress reduction. Landsc Urban Plan 2020 Jan;193:103654. [doi: 10.1016/j.landurbplan.2019.103654]
55. Jimenez-Barragan M, Del Pino Gutierrez A, Sauch Valmaña G, et al. Immersive virtual reality eHealth intervention to reduce anxiety and depression in pregnant women: randomized controlled trial. JMIR Hum Factors 2025 Apr 30;12:e71708. [doi: 10.2196/71708] [Medline: 40306641]
56. Kawai C, Georgiou F, Pieren R, Tobias S, Mavros P, Schäffer B. Investigating effect chains from cognitive and noise-induced short-term stress build-up to restoration in an urban or nature setting using 360° VR. J Environ Psychol 2024 Dec;100:102466. [doi: 10.1016/j.jenvp.2024.102466]
57. Kerr JI, Weibel RP, Naegelin M, et al. The effectiveness and user experience of a biofeedback intervention program for stress management supported by virtual reality and mobile technology: a randomized controlled study. BMC Digit Health 2023 Oct 13;1(1):42. [doi: 10.1186/s44247-023-00042-z]
58. Kılıç S, Dereli Yılmaz S. Virtual reality headset simulating a nature environment to improve health outcomes in pregnant women: a randomized-controlled trial. Clin Nurs Res 2023 Nov;32(8):1104-1114. [doi: 10.1177/10547738231184923] [Medline: 37408298]
59. Kim KY, Hur MH, Kim WJ. Effects of virtual reality (VR)-based meditation on sleep quality, stress, and autonomic nervous system balance in nursing students. Healthcare (Basel) 2024 Aug 8;12(16):1581. [doi: 10.3390/healthcare12161581] [Medline: 39201140]
60. Kosa M, Johnson-Glenberg MC. The additive effects of interactivity, immersion, and eudaimonia for stress reduction and mood management: A randomized controlled study on games and virtual reality. Psychology of Popular Media 2024 Oct;13(4):696-708. [doi: 10.1037/ppm0000512]
61. Kumpulainen S, Esmaeilzadeh S, Pesola AJ. Assessing the well-being benefits of VR nature experiences on group: heart rate variability insights from a cross-over study. J Environ Psychol 2024 Aug;97:102366. [doi: 10.1016/j.jenvp.2024.102366]
62. Lepilkina T, Beniashvili AG, Cheremin RA, et al. Efficacy of a relaxation scenario in virtual reality for the comorbid symptoms of anxiety and asthenia in a general hospital setting: a pilot comparative randomized open-label study. Consort Psychiatr 2023 Mar 31;4(1):38-51. [doi: 10.17816/CP221] [Medline: 38239567]
63. Li H, Liu F, Li S, Li C, Tian Z. Effects of virtual reality-guided imagery relaxation intervention in improving anxiety among lung cancer patients: a randomized controlled trial. Asian Nurs Res (Korean Soc Nurs Sci) 2025 Dec;19(5):509-517. [doi: 10.1016/j.anr.2025.08.003]
64. Liszio S, Graf L, Masuch M. The relaxing effect of virtual nature: immersive technology provides relief in acute stress situations. Annu Rev Cyberther Telemed 2018;16:87-93 [FREE Full text]
65. Liszio S, Masuch M. Interactive immersive virtual environments cause relaxation and enhance resistance to acute stress. Annu Rev Cyberther Telemed 2019;17:65 [FREE Full text]
66. Liu C, Zhang N, Li J, Shi J, Gao W, Yu Y. Preliminary data on effects of different street vegetation on stress recovery. Build Simul 2023 Nov;16(11):2109-2121. [doi: 10.1007/s12273-023-1018-0]
67. Ma J, Li H, Zhao M, Williams J. Mind nature: a virtual reality nature therapy for reducing anxiety and negative mood in university students. Int J Clin Health Psychol 2025;25(4):100636. [doi: 10.1016/j.ijchp.2025.100636] [Medline: 41142583]
68. Ma L, Chen J, An Q, et al. Immersive virtual nature environments for stress reduction: a between-subjects experimental study. Ann N Y Acad Sci 2026 Feb;1556(1):e70145. [doi: 10.1111/nyas.70145] [Medline: 41353623]
69. Newman M, Gatersleben B, Wyles KJ, Ratcliffe E. The use of virtual reality in environment experiences and the importance of realism. J Environ Psychol 2022 Feb;79:101733. [doi: 10.1016/j.jenvp.2021.101733]

XSL•FO
RenderX

70. Şansal KE, Şimşek AC, Aktan S, Özbey F, Paksoy A. Restorative effects of virtual nature on the emotional well-being of community-dwelling older adults. ejgg 2024 Apr 1;6(1):12-18. [doi: 10.4274/ejgg.galenos.2023.2023-3-2]
71. Schutte NS, Bhullar N, Stilinović EJ, Richardson K. The impact of virtual environments on restorativeness and affect. Ecopsychology 2017 Mar;9(1):1-7. [doi: 10.1089/eco.2016.0042]
72. Song R, Chen Q, Zhang Y, et al. Psychophysiological restorative potential in cancer patients by virtual reality (VR)-based perception of natural environment. Front Psychol 2022 Oct 10;13:1003497. [doi: 10.3389/fpsyg.2022.1003497]
73. Sun Y, Li F, He T, et al. Physiological and affective responses to green space virtual reality among pregnant women. Environ Res 2023 Jan;216:114499. [doi: 10.1016/j.envres.2022.114499]
74. Yang P. Virtual reality tools to support music students to cope with anxiety and overcome stress. Educ Inf Technol 2024 Sep;29(13):16525-16540. [doi: 10.1007/s10639-024-12464-x]
75. Yildirim M, Globa A, Gocer O, Brambilla A. Multisensory nature exposure in the workplace: exploring the restorative benefits of smell experiences. Build Environ 2024 Aug;262:111841. [doi: 10.1016/j.buildenv.2024.111841]
76. Yin J, Yuan J, Arfaei N, Catalano PJ, Allen JG, Spengler JD. Effects of biophilic indoor environment on stress and anxiety recovery: a between-subjects experiment in virtual reality. Environ Int 2020 Mar;136:105427. [doi: 10.1016/j.envint.2019.105427] [Medline: 31881421]
77. Zhang X, Yao S, Wang M, et al. The impact of VR-CALM intervention based on VR on psychological distress and symptom management in breast cancer survivors. J Oncol 2022 Jun 7;2022:1-10. [doi: 10.1155/2022/1012813]
78. Bodet-Contentin L, Letourneur M, Ehrmann S. Virtual reality during work breaks to reduce fatigue of intensive unit caregivers: a crossover, pilot, randomised trial. Aust Crit Care 2023 May;36(3):345-349. [doi: 10.1016/j.aucc.2022.01.009] [Medline: 35246356]
79. Suppakittpaisarn P, Wu CC, Tung YH, et al. Durations of virtual exposure to built and natural landscapes impact self-reported stress recovery: evidence from three countries. Landscape Ecol Eng 2023 Jan;19(1):95-105. [doi: 10.1007/s11355-022-00523-9]
80. Anderson AP, Mayer MD, Fellows AM, Cowan DR, Hegel MT, Buckey JC. Relaxation with immersive natural scenes presented using virtual reality. Aerosp Med Hum Perform 2017 Jun 1;88(6):520-526. [doi: 10.3357/AMHP.4747.2017] [Medline: 28539139]
81. Da C, Li Y, Gao X, et al. Seasonal variations in psychophysiological stress recovery from street greenery: a virtual reality study on vegetation structures and configurations. Build Environ 2024 Dec;266:112058. [doi: 10.1016/j.buildenv.2024.112058]
82. Gao T, Zhang T, Zhu L, Gao Y, Qiu L. Exploring psychophysiological restoration and individual preference in the different environments based on virtual reality. Int J Environ Res Public Health 2019 Aug 26;16(17):3102. [doi: 10.3390/ijerph16173102] [Medline: 31455015]
83. Gao Y, Zhang Y, Zhang W, Meng H, Zhang Z, Zhang T. Well-being and sustainable utilization of forest environment with diverse vegetation distributions. Sustainability 2024 Sep 29;16(19):8469. [doi: 10.3390/su16198469]
84. Hung SH, Lee TA, Lee AY, Chang CY. Enhancing physiological effects, subjective vitality, and emotional well-being through immersive experiences in biologically diverse environments. Landscape Ecol Eng 2025 Oct;21(4):773-791. [doi: 10.1007/s11355-025-00673-6]
85. Jiang Y, Wang Y, Han L. Does virtual reality forest mindfulness restore individuals' mood after work and the next day? A multilevel modeling experiment. J Environ Psychol 2025 Jun;104:102596. [doi: 10.1016/j.jenvp.2025.102596]
86. Jimenez RT, Zuniga-Hernandez M, Suen MY, et al. Culturally sensitive treatment of caregiver anxiety with virtual reality: a prospective, pragmatic, randomized study. J Patient Exp 2025;12:23743735251326671. [doi: 10.1177/23743735251326671] [Medline: 40151338]
87. Knaust T, Felnhofer A, Kothgassner OD, Höllmer H, Gorzka RJ, Schulz H. Exposure to virtual nature: the impact of different immersion levels on skin conductance level, heart rate, and perceived relaxation. Virtual Real 2022 Sep;26(3):925-938. [doi: 10.1007/s10055-021-00595-2]
88. Mahmud AA, Zakaria H, Mohd Yusoff MZ, et al. Brief virtual reality exposure therapy and its effects on negative and positive emotions among healthy working adults: a feasibility study. Alpha Psychiatry 2022;23(5):223-229. [doi: 10.5152/alphapsychiatry.2022.21781]
89. Martínez Manchón FDA, Šimunić A. Effects of short-term exposure to interactive vs. non-interactive virtual nature on cognitive performance and mental health in college students. Int J Human–Computer Interact 2024 Sep 16;40(18):5443-5453. [doi: 10.1080/10447318.2023.2235122]
90. Plante TG, Aldridge A, Su D, Bogdan R, Belo M, Kahn K. Does virtual reality enhance the management of stress when paired with exercise? An exploratory study. Int J Stress Manag 2003 Aug;10(3):203-216. [doi: 10.1037/1072-5245.10.3.203]
91. Richesin MT, Baldwin DR, Wicks LAM. Art making and virtual reality: a comparison study of physiological and psychological outcomes. Arts Psychother 2021 Sep;75:101823. [doi: 10.1016/j.aip.2021.101823]
92. Schebella MF, Weber D, Schultz L, Weinstein P. The nature of reality: human stress recovery during exposure to biodiverse, multisensory virtual environments. Int J Environ Res Public Health 2019 Dec 19;17(1):56. [doi: 10.3390/ijerph17010056] [Medline: 31861700]

XSL•FO
RenderX

93. Valtchanov D, Ellard C. Physiological and affective responses to immersion in virtual reality: effects of nature and urban settings. J Cyber Ther Rehabil 2010;3(4):359-374 [FREE Full text]
94. Villani D, Riva G. Presence and relaxation: a preliminary controlled study. PsychNol J 2008;6(1):7-25 [FREE Full text]
95. Wu S, Liu G, Yang J, et al. Psychological effects of virtual reality intervention on breast cancer patients with different personalities: a randomized controlled trial. Int J Nurs Sci 2025 Mar;12(2):107-114. [doi: 10.1016/j.ijnss.2025.02.008] [Medline: 40241873]
96. Xiaoxue S, Huang X. Promoting stress and anxiety recovery in older adults: assessing the therapeutic influence of biophilic green walls and outdoor view. Front Public Health 2024;12:1352611. [doi: 10.3389/fpubh.2024.1352611] [Medline: 38686034]
97. Yin M, Li K, Xu Z, Jiao R, Yang W. Exploring the impact of autumn color and bare tree landscapes in virtual environments on human well-being and therapeutic effects across different sensory modalities. In: Petrescu L, editor. PLOS One 2024;19(4):e0301422. [doi: 10.1371/journal.pone.0301422] [Medline: 38635838]
98. Yu CP, Lee HY, Luo XY. The effect of virtual reality forest and urban environments on physiological and psychological responses. Urban Forestry & Urban Greening 2018 Oct;35:106-114. [doi: 10.1016/j.ufug.2018.08.013]
99. Yuan S, Tao F, Li Y. The restorative effects of virtual reality forests on elderly individuals during the COVID-19 lockdown. J Organ End User Comput 2022 Nov;34(6):1-22. [doi: 10.4018/JOEUC.297626]
100. Zheng L, Li W, Song S, et al. Effectiveness of mindfulness-based virtual reality training on stress, anxiety, and depression among Chinese university students. Mindfulness (N Y) 2024 Apr;15(4):899-913. [doi: 10.1007/s12671-024-02335-4]
101. Saredakis D, Szpak A, Birckhead B, Keage HAD, Rizzo A, Loetscher T. Factors associated with virtual reality sickness in head-mounted displays: a systematic review and meta-analysis. Front Hum Neurosci 2020;14:96. [doi: 10.3389/fnhum.2020.00096] [Medline: 32300295]
102. Kim YS, Won J, Jang SW, Ko J. Effects of cybersickness caused by head-mounted display-based virtual reality on physiological responses: cross-sectional study. JMIR Serious Games 2022 Oct 17;10(4):e37938. [doi: 10.2196/37938] [Medline: 36251360]
103. Zielasko D, Rehling B, von Dawans B, Domes G. Do not immerse and drive? Prolonged effects of cybersickness on physiological stress markers and cognitive performance. Virtual Real 2026;30(2). [doi: 10.1007/s10055-026-01335-0]
104. Moon DU, Lütt A, Kim H, et al. Impact of cybersickness and presence on treatment satisfaction and clinical outcomes in virtual reality-based biofeedback for depression and anxiety. J Psychiatr Res 2025 Jul;187:53-61. [doi: 10.1016/j.jpsychires.2025.04.047] [Medline: 40345075]
105. Park SY, Koo DK. The impact of virtual reality content characteristics on cybersickness and head movement patterns. Sensors (Basel) 2025 Jan 2;25(1):215. [doi: 10.3390/s25010215] [Medline: 39797006]
106. Yusufov M, Nicoloro-SantaBarbara J, Grey NE, Moyer A, Lobel M. Meta-analytic evaluation of stress reduction interventions for undergraduate and graduate students. Int J Stress Manag 2019;26(2):132-145. [doi: 10.1037/str0000099]
107. M. S, S. M, Vadakkiniath IJ, A. G. Prevalence and correlates of stress, anxiety, and depression in patients with chronic diseases: a cross-sectional study. Middle East Curr Psychiatry 2023 Aug 16;30(1):66. [doi: 10.1186/s43045-023-00340-2]
108. Samolis S. Stress in medical patients. Ann Gen Psychiatry 2010 Apr;9(S1):S39. [doi: 10.1186/1744-859X-9-S1-S39]
109. Zaharuddin FA, Ibrahim N, Yusof AM. Enhancing user comfort in virtual environments for effective stress therapy: design considerations. Int J Adv Comput Sci Appl 2024 Dec 30;15(12). [doi: 10.14569/IJACSA.2024.0151230]
110. Asif S, Muddassar A, Shahzad TZ, Raouf M, Pervaiz T. Frequency of depression, anxiety and stress among university students. Pak J Med Sci 2020;36(5):971-976. [doi: 10.12669/pjms.36.5.1873]
111. Olson N, Oberhoffer-Fritz R, Reiner B, Schulz T. Stress, student burnout and study engagement – a cross-sectional comparison of university students of different academic subjects. BMC Psychol 2025 Mar 24;13(1):293. [doi: 10.1186/s40359-025-02602-6]
112. Rosemberg MA, Granner J, Li Y, Seng JS. A scoping review of interventions targeting allostatic load. Stress 2020 Sep 2;23(5):519-528. [doi: 10.1080/10253890.2020.1784136]
113. Ghasemi F, Beversdorf DQ, Herman KC. Stress and stress responses: a narrative literature review from physiological mechanisms to intervention approaches. J Pac Rim Psychol 2024 Jan;18. [doi: 10.1177/18344909241289222]
114. Gao D, Su Y, Zhang X, Li H, Luo H. The application of virtual reality meditation and mind–body exercises among older adults. Front Psychol 2024 Mar 19;15:1303880. [doi: 10.3389/fpsyg.2024.1303880]
115. Chan SHM, Qiu L, Esposito G, Mai KP, Tam KP, Cui J. Nature in virtual reality improves mood and reduces stress: evidence from young adults and senior citizens. Virtual Real 2023 Dec;27(4):3285-3300. [doi: 10.1007/s10055-021-00604-4]
116. Li H, Browning M, Rigolon A, et al. Beyond “bluespace” and “greenspace”: a narrative review of possible health benefits from exposure to other natural landscapes. Science of The Total Environment 2023 Jan;856:159292. [doi: 10.1016/j.scitotenv.2022.159292]
117. Reese G, Kohler E, Menzel C. Restore or get restored: the effect of control on stress reduction and restoration in virtual nature settings. Sustainability 2021 Feb;13(4):1995. [doi: 10.3390/su13041995]
118. Higgins JPT, Thompson SG. Controlling the risk of spurious findings from meta-regression. Stat Med 2004 Jun 15;23(11):1663-1682. [doi: 10.1002/sim.1752] [Medline: 15160401]

XSL•FO
RenderX

119. Nicoly JB, Masters R, Gaddy V, Interrante V, Ortega F. The restorative influence of virtual reality environment design. In: SAP '24: ACM Symposium on Applied Perception 2024: Association for Computing Machinery; 2024. [doi: 10.1145/3675231.3675244]
120. Pawełczyk W, Młyńczyk K, Murawski M, Rutkowska A, Rutkowski S. Assessment of spatial presence experiences and symptoms of cybersickness in dynamic and static virtual environments. J Clin Med 2025 Jun 12;14(12):4200. [doi: 10.3390/jcm14124200] [Medline: 40565946]
121. Mostajeran F, Krzikawski J, Steinicke F, Kühn S. Effects of exposure to immersive videos and photo slideshows of forest and urban environments. Sci Rep 2021 Feb 17;11(1):3994. [doi: 10.1038/s41598-021-83277-y] [Medline: 33597577]
122. Wang J, Shi R, Zheng W, Xie W, Kao D, Liang HN. Effect of frame rate on user experience, performance, and simulator sickness in virtual reality. IEEE Trans Visual Comput Graphics 2023 May;29(5):2478-2488. [doi: 10.1109/TVCG.2023.3247057]
123. Ideno Y, Hayashi K, Abe Y, et al. Blood pressure-lowering effect of Shinrin-yoku (forest bathing): a systematic review and meta-analysis. BMC Complement Altern Med 2017 Aug 16;17(1):409. [doi: 10.1186/s12906-017-1912-z] [Medline: 28814305]
124. Song C, Ikei H, Miyazaki Y. Physiological effects of nature therapy: a review of the research in Japan. Int J Environ Res Public Health 2016 Aug 3;13(8):781. [doi: 10.3390/ijerph13080781] [Medline: 27527193]
125. Corazon SS, Sidenius U, Poulsen DV, Gramkow MC, Stigsdotter UK. Psycho-physiological stress recovery in outdoor nature-based interventions: a systematic review of the past eight years of research. Int J Environ Res Public Health 2019 May 16;16(10):1711. [doi: 10.3390/ijerph16101711] [Medline: 31100773]
126. Robbemond LM, Noordzij ML, van Driel CMG, Veling W. Physiological responses to virtual reality-based stress regulation and relaxation interventions: a systematic review. Stress Health 2026 Apr;42(2):e70164. [doi: 10.1002/smi.70164] [Medline: 41830413]
127. Reese G, Stahlberg J, Menzel C. Digital shinrin-yoku: do nature experiences in virtual reality reduce stress and increase well-being as strongly as similar experiences in a physical forest? Virtual Real 2022 Sep;26(3):1245-1255. [doi: 10.1007/s10055-022-00631-9]
128. Fan L, Baharum MR. The effects of digital nature and actual nature on stress reduction: a meta-analysis and systematic review. Internet Interv 2024 Dec;38:100772. [doi: 10.1016/j.invent.2024.100772] [Medline: 39329057]
129. Shafir R, Watson L, Felix RB, et al. Factors influencing the hypoalgesic effects of virtual reality. Pain 2025 Mar 4;166(8):1836-1846. [doi: 10.1097/j.pain.0000000000003549] [Medline: 40043143]
130. Moyer M, Lenz AS, Gonzales-Wong C, Rodriguez G. A meta-analysis of virtual reality interventions for reducing the symptoms of anxiety. Counseling Outcome Research and Evaluation 2025 Jul 3;16(2):124-142. [doi: 10.1080/21501378.2024.2438952]
131. Savoric T, Aziz S, Ling RR, Antlej K, Arnab S, Subramaniam A. Systematic review: the impact of virtual reality interventions on stress and anxiety in intensive care units. J Crit Care 2025 Dec;90:155164. [doi: 10.1016/j.jcrc.2025.155164]

Abbreviations

CAVE: cave automatic virtual environment
DASS: Depression, Anxiety, and Stress Scale
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
HKSJ: Hartung, Knapp, Sidik, and Jonkman approach
HMD: head-mounted display
HR: heart rate
HRV: heart rate variability
PANAS: Positive and Negative Affect Schedule
PI: prediction interval
POMS: Profile of Mood States
PRESS: Peer Review of Electronic Search Strategies
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews
PSL: perceived stress level
RCT: randomized controlled trial
RoB: risk of bias
RWE: real-world exposure
SCL: skin conductance level
SMD: standardized mean difference
STAI-S: State Scale of State-Trait Anxiety Inventory
SYS: systolic blood pressure
TSM: traditional stress management

XSL•FO
RenderX

VR: virtual reality

Edited by S Brini; submitted 30.May.2025; peer-reviewed by B Najand, K Jermsittiparsert, V White; revised version received 24.Mar.2026; accepted 25.Mar.2026; published 25.May.2026.

Please cite as:
Strauch H, Schuil I, Simm S, Grubert J, Kalamkar S, Meissner K
Virtual Reality Interventions for Stress Reduction in the General Population: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e78212
URL: https://www.jmir.org/2026/1/e78212
doi:10.2196/78212

© Hannah Strauch, Isabel Schuil, Stefan Simm, Jens Grubert, Snehanjali Kalamkar, Karin Meissner. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Efficacy, User Engagement, and Acceptability of Cognitive Behavioral Therapy–Oriented Psychological Chatbots for Adults With Depressive and/or Anxiety Symptoms: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Bingyan Gong[1], PhD; Nisha Yao[2], PhD; Hangxin Xie[1]; Chuncheng Huang[1], BS; Tomoko Kishimoto[3], PhD; Howard Berenbaum[4], PhD; Wenting Mu[1], PhD

[1]Department of Psychological and Cognitive Sciences, Tsinghua University, 30 Shuangqing Road, Haidian District, Beijing, China

[2]School of Kinesiology and Health, Capital University of Physical Education and Sports, Beijing, China

[3]Faculty of Psychology, Beijing Normal University, Beijing, China

[4]Department of Psychology, University of Illinois Urbana-Champaign, Champaign, IL, United States

Corresponding Author:
Wenting Mu, PhD
Department of Psychological and Cognitive Sciences, Tsinghua University, 30 Shuangqing Road, Haidian District, Beijing, China

Abstract

Background: Cognitive behavioral therapy (CBT) is the most examined psychotherapy for depression and anxiety, but delivery faces significant barriers such as limited access, cost, and time constraints. CBT-oriented psychological chatbots offer a promising means of addressing these challenges. Yet, their overall efficacy, user engagement, and acceptability have not been systematically synthesized.

Objective: This study aimed to evaluate the efficacy, user engagement, and acceptability of CBT-oriented chatbots for adults with depressive and/or anxiety symptoms.

Methods: A systematic search of 9 databases, including PubMed, Cochrane Central Register of Controlled Trials, Embase, Web of Science, PsycINFO, CINAHL, China National Knowledge Infrastructure, WanFang, and VIP Databases, was conducted from inception to February 2026. Eligibility criteria included randomized controlled trials comparing CBT-oriented chatbots with control groups in adults with depressive and/or anxiety symptoms. Risk of bias (ROB) was assessed using the Cochrane ROB tool. Random-effects meta-analyses (Hartung-Knapp-Sidik-Jonkman adjustment) calculated pooled effect sizes (Hedges *g*), 95% CIs, and 95% prediction intervals (PIs). Heterogeneity was evaluated using the I^2 statistic, and Galbraith plots were used to identify outliers for subsequent sensitivity analyses. Subgroup and meta-regression analyses examined potential moderators. The certainty of evidence was evaluated using the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) approach. Data on user engagement and acceptability were extracted and synthesized using narrative and quantitative methods where available.

Results: Twenty-nine eligible randomized controlled trials were included. CBT-oriented psychological chatbots produced a moderate reduction in depressive symptoms at postintervention (g=−0.55, 95% CI −0.70 to −0.40, 95% PI −1.23 to 0.13) and a small reduction in anxiety symptoms (g=−0.26, 95% CI −0.37 to −0.14, 95% PI −0.67 to 0.15). At follow-up, effects were small for depression (g=−0.32, 95% CI −0.55 to −0.09, 95% PI −0.93 to 0.29) and nonsignificant for anxiety (g=−0.19, 95% CI −0.43 to 0.04, 95% PI −0.84 to 0.46). Subgroup and meta-regression analyses revealed that anxiety outcomes were significantly moderated by clinical profiles—showing distinct advantages for comorbid symptoms—and the proportion of female participants. The CBT-oriented chatbots received an adequate level of engagement that complied with digital intervention standards. Although user satisfaction ratings were generally favorable, technical limitations and repetitive interaction patterns remain to be addressed to enhance overall acceptability. Regarding the limitations of evidence, the overall certainty was rated as very low to low, predominantly driven by high ROB and substantial heterogeneity.

Conclusions: This study innovatively isolates CBT-oriented chatbots from broader digital interventions, providing a precise, methodology-driven evaluation of theoretically grounded therapeutics. This review brings critical evidence to the field that these tools yield significant short-term relief, particularly for comorbid anxiety profiles. In the real world, CBT chatbots offer profound potential as scalable, low-barrier first-line tools. To sustain engagement, future developments must evolve from rigid rule-based scripts toward adaptive, large language model–driven architectures while ensuring clinical safety.

Trial Registration: PROSPERO CRD42024615506; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024615506

XSL•FO
RenderX

(J Med Internet Res 2026;28:e82677) doi:10.2196/82677

KEYWORDS

chatbots; cognitive behavioral therapy; digital mental health; depression; anxiety; systematic review; meta-analysis; engagement; acceptability

Introduction

Mental disorders have become a growing public health concern worldwide and are among the leading causes of the global health–related burden [1]. The Global Burden of Disease Study 2019 found that depressive disorders accounted for the highest proportion (37.3%) of the global burden of disease attributable to mental disorders, followed by anxiety disorders (22.9%) [2]. In addition to causing considerable suffering and impaired functioning, when severe, these illnesses can be life-threatening [3]. Meanwhile, depressive and anxiety disorders impose substantial economic burden through excessive health care expenditures [4,5]. According to *The Lancet Global Health*, every dollar invested in scaled-up treatment for depression and anxiety yields a 4-fold return in improved health and productivity outcomes [6]. Therefore, there is an urgent imperative to develop effective treatments for depressive and anxiety disorders.

Cognitive behavioral therapy (CBT) is the most examined type of psychotherapy for depression and anxiety [7-9]. However, access to CBT remains limited due to resource constraints, long waiting times, stigma, and a preference for self-help strategies [10-12]. These challenges highlight the need for alternative delivery methods that are more accessible, flexible, and cost-effective. Digital mental health interventions (DMHIs) have emerged as promising, scalable solutions capable of delivering standardized, protocol-adherent CBT. This approach enhances accessibility and cost-effectiveness, thereby addressing the escalating demand for mental health services [13,14]. Among these innovations, chatbots have attracted considerable attention due to their capacity to provide immediate, structured therapeutic support through emulation of human-like dialogue [15].

CBT-oriented psychological chatbots are designed to deliver therapeutic content based on CBT principles, guiding users through processes such as cognitive restructuring, mood monitoring, and behavior activation [16]. Although an increasing number of studies have evaluated CBT-oriented chatbots for depressive and anxiety symptoms, the findings remain inconsistent. Some studies have reported significant reductions in depressive and/or anxiety symptoms [17-28], while others have found no significant improvements compared with control groups [29-42]. Mixed results have also been observed, with some interventions showing effectiveness for depression but not anxiety [43-45]. Given these inconsistencies, a systematic synthesis is needed to clarify the overall efficacy of CBT-oriented chatbots and identify factors that may influence therapeutic outcomes.

To date, no systematic review has specifically focused on the efficacy of CBT-oriented chatbots for addressing depression and/or anxiety. Existing reviews typically encompass a wide spectrum of digital psychological technologies or chatbot interventions in general mental health, rather than isolating the specific efficacy of CBT-based chatbots for depression and/or anxiety. Consequently, their findings reflect the aggregated effects of heterogeneous interventions, making it difficult to draw conclusions about the unique contribution of CBT-oriented chatbots. In addition to this broad scope, existing reviews also suffer from several methodological limitations: (1) lack of quantitative synthesis: 2 reviews have only descriptive summaries without conducting meta-analysis, resulting in a lack of specific numerical support for the results and difficulty in determining the magnitude and statistical significance of the effects [46,47]. (2) Limited statistical power: 3 systematic reviews examine chatbots for mental health improvement, but they included a relatively small number of randomized controlled trials (RCTs) targeting anxiety and/or depression (ranging from 2 to 13), resulting in insufficient statistical power [48-50]. This scarcity is primarily due to the fact that the literature search was conducted before the recent surge in digital health publications. (3) Incomplete outcome assessment: 1 systematic review restricts its quantitative synthesis exclusively to depressive symptom measurements across the 11 eligible trials, while failing to analyze anxiety symptoms [51]. (4) Inclusion of nontarget populations: another review [52] includes individuals who did not exhibit symptoms of anxiety or depression prior to intervention, such as Suharwardy et al [53] and Kleinau et al [54], which could potentially lead to underestimation of the intervention's true efficacy and even false-negative results. (5) Limited exploration of moderators: only 4 reviews investigate efficacy moderators using subgroup or meta-regression analyses [48,49,51,52]. Consequently, the evidence regarding the efficacy of CBT-oriented chatbots for depressive and/or anxiety symptoms remains fragmented and inadequately synthesized. The underlying mechanisms affecting their efficacy are still unclear.

Beyond therapeutic efficacy, engagement and acceptability are vital in impacting clinical outcomes. Engagement refers to the extent to which a user is exposed to intervention content, which is critical to realizing the promise of digital interventions in the real world [55]. Acceptability refers to the degree to which individuals' perceived ease of use and perceived usefulness of a technology positively influence their usage intentions, thereby driving its adoption [56]. Nonetheless, prior reviews have not elucidated acceptability and engagement with CBT-oriented chatbots in depressive and anxiety symptoms.

This study is the first to systematically isolate theoretically grounded, CBT-oriented chatbots from the broader landscape of digital conversational agents. The study aimed to (1) evaluate the therapeutic efficacy of CBT-oriented chatbots in adults with depressive and/or anxiety symptoms, (2) systematically explore potential moderators of intervention efficacy, and (3) synthesize evidence on user engagement and acceptability to provide actionable insights for future clinical implementation. Clarifying

XSL•FO
RenderX

the therapeutic potential and user experience of CBT-oriented chatbots is critical, given the urgent global demand for accessible digital mental health solutions. These findings establish an evidence base to advance the development and optimization of scalable, evidence-driven chatbots.

Methods

Study Design

This systematic review and meta-analysis was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [57], and the PRISMA 2020 Expanded Checklist was used. The study protocol was prospectively registered at PROSPERO (CRD42024615506).

Eligibility Criteria

Participants

The included studies involved adults aged 18 years and older with baseline depression and/or anxiety symptom scores meeting or exceeding the defined cutoff for mild severity. Participants were categorized as depression or anxiety-only participants (individuals presenting with symptoms exclusively within the depression and/or anxiety, without documented co-occurring somatic or other psychological symptoms), and participants with comorbid symptoms (individuals with depressive and/or anxiety symptoms accompanied by at least 1 co-occurring somatic symptom or other psychological symptoms).

Intervention

Trials were included if they clearly specified the utilization of CBT-oriented psychological chatbots as an intervention modality. CBT-oriented psychological chatbots were defined as systems integrating CBT techniques—such as cognitive restructuring, behavioral activation, exposure, and problem-solving—into chatbot interactions. The chatbots engaged in 2-way communication with users and were delivered via mobile or computer devices. Studies in which chatbots merely served as reminders to encourage user participation, rather than delivering therapeutic interventions, were excluded.

Control or Comparison

The studies included at least 1 control condition, active (other active interventions, treatment-as-usual, psychological placebo, and attention control) or passive (waitlist, blank control, or assessment only) control group.

Outcomes

The primary outcomes were quantitative measures of depression or anxiety symptom severity. The secondary outcomes included user engagement and acceptability of CBT-oriented psychological chatbots. Engagement in this review was defined as any objective indicator used to quantify the extent of intervention use, including attrition rate, number of logins, module completion, frequency of use, and time spent in the intervention [58]. Acceptability was defined as participants' psychological and emotional attitudes toward the chatbot interventions, including satisfaction, tolerance, and user experience [56,59].

Study Design and Characteristics

The eligible studies were RCTs published in English or Chinese. The review included peer-reviewed papers, dissertations, and conference proceedings, while it excluded preprint papers, conference abstracts, proposals, and editorials. Studies with duplicate published data were excluded.

Search Strategy

The systematic literature search was designed, conducted, and reported in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Search Extension) guideline [60]. The search strategy consisted of 2 steps. First, the following electronic databases were systematically searched from the inception to April 1, 2025, with a final update on February 9, 2026: PubMed (via NCBI), Cochrane Central Register of Controlled Trials (via Wiley), Embase (via Elsevier), Web of Science (via Clarivate), PsycINFO (via APA PsycNet), CINAHL (via EBSCOhost), China National Knowledge Infrastructure, WanFang, and VIP Databases. The search strategy combined MeSH (Medical Subject Headings) and free-text terms related to "depression," "anxiety," and "chatbot." The strategy developed de novo for this review was not adapted from prior searches. No language or date restrictions, nor any predefined search filters, were applied to maximize sensitivity. Second, the reference lists of all eligible studies were manually searched to identify additional relevant studies. No other online resources or study registries were searched. A detailed search strategy is presented as shown in Table S1 in Multimedia Appendix 1.

Selection Process

EndNote 20 (a reference management software; Clarivate) was used to organize and manage the identified records. After removing duplicates, 2 authors (BG and HX) independently screened all titles or abstracts. Subsequently, the full texts of studies selected in the previous step were independently reviewed by the same 2 authors. The disagreements between the 2 authors were discussed until consensus was reached. When necessary, a third reviewer (WM) was consulted to make a final decision. Interrater reliability, assessed using Cohen κ [61], indicated a very good level of agreement for title or abstract screening (0.86) and full-text review (0.85).

Data Collection Process

Data extraction was divided into two categories: (1) basic data describing the study characteristics, and (2) core data used for statistical analysis. Two authors (BG and HX) independently coded key study characteristics and extracted relevant outcome data. The corresponding author (WM) or the first author (BG) was contacted via email if information for calculating effect sizes in a study was not reported. If no response was received after 1 month, or if the authors were unable to provide the required data, the study was excluded as unavailable.

Data Items

Two independent reviewers (BG and HX) extracted the information from eligible studies. The extracted basic information included (1) study characteristics: the name of the first author, the year of publication, country of trial, and study

XSL•FO
RenderX

design; (2) participant characteristics: nature of participant, sample size, mean age, and proportion of females; (3) experimental characteristics: intervention group, control group, depression or anxiety measures; and (4) chatbot characteristics: CBT component, intervention duration, number of sessions, duration of each session, follow-up length, chatbot name, platform, chatbot type, dialogue initiative, interaction frequency, input, and output. The core data include (1) primary outcomes: depressive or anxiety symptom severity, and (2) secondary outcomes: engagement and acceptability.

Risk of Bias Assessment

To assess the risk of bias (ROB) and methodological quality of included studies, 2 rating systems were independently applied by 2 reviewers (BG and HX), with disagreements resolved through discussion and consultation with a third reviewer (WM). The Cochrane Collaboration's tool for assessing ROB [62] was used to evaluate all studies across 7 domains: "random sequence generation," "allocation concealment," "blinding of participants and personnel," "blinding of outcome assessment," "incomplete outcome data," "selective reporting," and "other bias." Each domain was rated as having a "low," "unclear," or "high risk of bias."

Grading the Quality of Evidence

The GRADE (Grading of Recommendations Assessment, Development, and Evaluation) framework was used to evaluate the certainty of evidence provided by this meta-analysis for the different outcomes [63]. The GRADE approach considers 5 domains, including ROB, inconsistency, indirectness, imprecision, and publication bias. Evidence quality was categorized as high, moderate, low, or very low. Two authors (BG and HX) independently evaluated the evidence quality for the following outcomes. Any discrepancies in ratings were resolved through discussion.

Statistical Analyses

Primary outcome measures were converted into standardized mean differences with 95% CIs and 95% prediction intervals (PIs). When SDs were not available, they were estimated through mathematical transformation [64]. If the SD_{change} or the correlation between pre- and postintervention scores was not reported and could not be calculated, we estimated the SD_{change} using imputed correlation coefficients (r) derived from studies that reported complete data. Considering potential variations in temporal stability and clinical constructs, r values were calculated separately for depressive symptoms (Post: r=0.85; Follow-up: r=0.73) and anxiety symptoms (Post: r=0.67; Follow-up: r=0.67). If both intention-to-treat and completer analyses were presented, data from the former were extracted and analyzed. For studies with multiarm designs that included multiple experimental or control groups, we extracted the data from the group most relevant to our research questions. If multiple groups met the inclusion criteria, we combined the means and SDs from these arms, as suggested by the Cochrane guidelines for integrating multiple groups from a single study (Table S2 in Multimedia Appendix 1 [20,23,25,27,35]). When a study had multiple follow-up data, we extracted the longest follow-up for analysis [64]. Given potential baseline differences, the change from baseline scores was computed for each group to represent efficacy. The effect sizes for depressive and anxiety symptoms were calculated separately, using Hedges g to adjust for small sample bias, with 95% CI and z statistics to determine significance (P<.05). Following Cohen's [65] (1988) conventions, effect sizes (Hedges' g) of 0.2, 0.5, and 0.8 were interpreted as small, medium, and large, respectively. All meta-analyses were performed using an inverse variance random-effects model with Hartung-Knapp-Sidik-Jonkman adjustment [66].

Heterogeneity was assessed using Cochrane Q test (P<.10 for significance) and the I^2 statistics, with interpretation thresholds following the Cochrane Handbook [64]. Specifically, an I^2 of 0%-40% was considered to represent unimportant heterogeneity, 30%-60% moderate heterogeneity, 50%-90% substantial heterogeneity, and 75%-100% considerable heterogeneity [64]. To visually explore potential sources of heterogeneity and identify specific outlier studies, a Galbraith plot was generated. Studies positioned outside the 95% confidence bounds of the regression line were classified as outliers. Subsequently, a sensitivity analysis excluding these identified outliers was conducted to evaluate the robustness of the pooled effect sizes and to maintain overall homogeneity [64]. To explore potential moderating factors and identify sources of heterogeneity, subgroup analyses and meta-regression for the primary outcomes (depressive and anxiety symptoms) were conducted. The investigated covariates included participants' symptom profiles, control conditions, intervention duration, delivery platform, chatbot type, country, and ROB. Spearman correlation analyses showed no multicollinearity among moderators (|r| <0.7; Table S3 in Multimedia Appendix 1). To examine the relative importance of different moderator domains, we adopted a multimodel inference approach using information criteria. We compared the 3 prespecified models (model 1: sample characteristics, model 2: technical features, and model 3: study design) using the Akaike information criterion (AIC) [67] and the Bayesian information criterion (BIC) [68]. Lower AIC and BIC values indicate a more parsimonious model that achieves a better trade-off between goodness of fit and complexity [67,68]. Small-study effects were assessed through visual inspection of funnel plot asymmetry and the Egger test [69], which was applied when 10 or more trials were available [70]. For all statistical analyses, a significance level of α<.05 was considered. All analyses were conducted using Stata SE (version 17.0; StataCorp LLC).

Significant heterogeneity was observed across studies in the reporting of chatbot engagement and acceptability. The variability manifested in several key dimensions: (1) diverse operational definitions of engagement (eg, adherence rate, session frequency, and number of messages exchanged) and acceptability (eg, satisfaction and subjective user experience), (2) heterogeneous measurement instruments (eg, standardized questionnaires, study-designed items, and system usage logs), and (3) inconsistent reporting of outcome metrics (eg, means, SDs; medians; proportions; and only descriptive summaries). This variability in measurement and reporting precluded the calculation of comparable effect sizes across studies, making a meta-analysis unfeasible. Therefore, findings related to

XSL•FO
RenderX

engagement and acceptability were qualitatively summarized for each included trial.

Results

Study Selection

The database search yielded 20,782 records and a further 2 records were identified after searching the reference lists of relevant systematic reviews. After removing 8155 duplicates, 12,629 titles or abstracts were assessed for eligibility and 12,232 records were excluded. Following full-text screening, 397 papers were further screened and 368 were excluded. Finally, 29 studies were included in the systematic review and meta-analysis (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. Search and study selection process. CBT: cognitive behavioral therapy; RCT: randomized controlled trial.

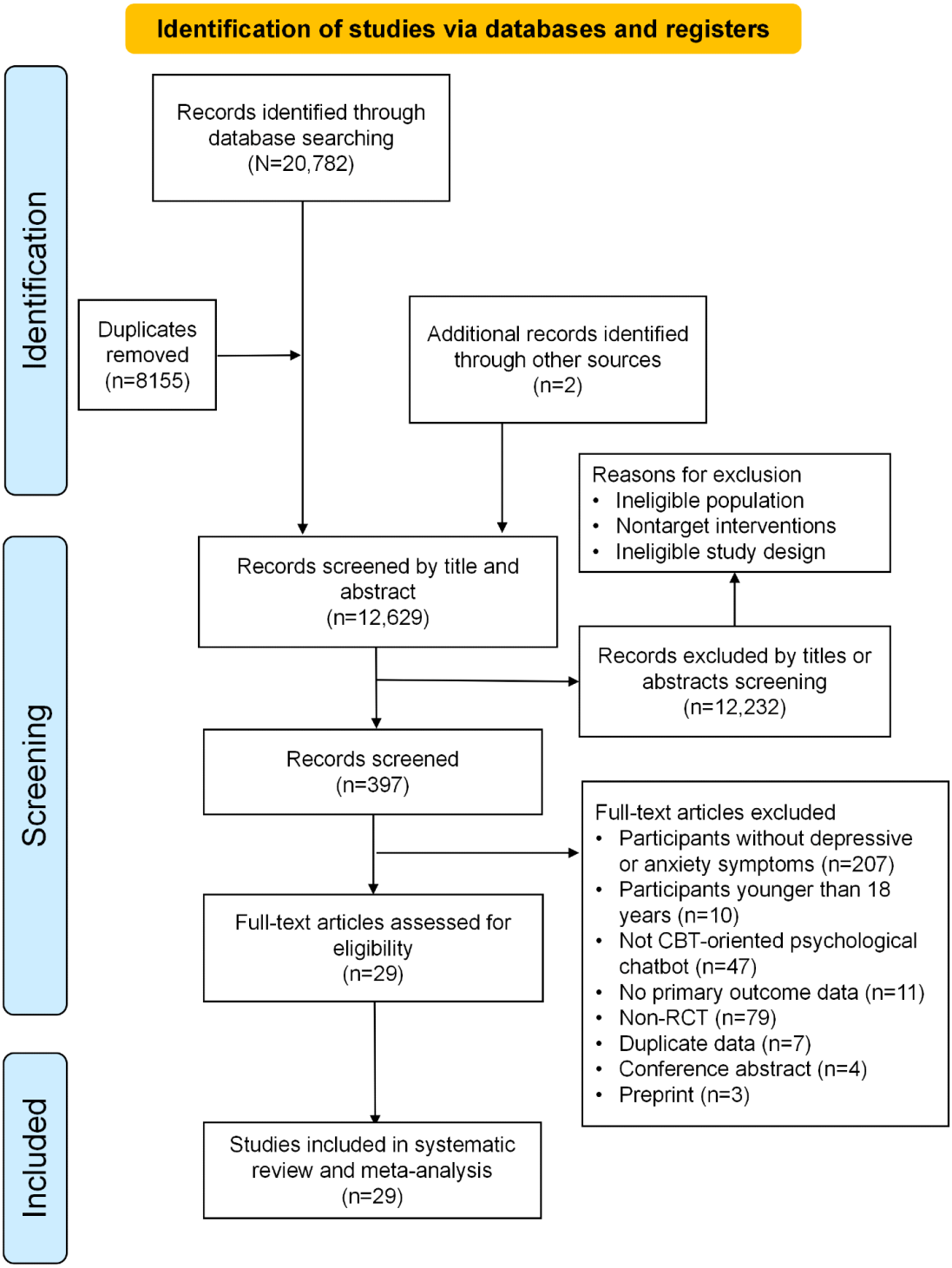

XSL•FO
RenderX

Study Characteristics

The key characteristics of studies and participants are summarized in Table S4 in Multimedia Appendix 1 [17-45]. The studies were published between 2016 and 2026. Nine trials were from the United States [24,27,29,36,38,40,41,44,45], 7 from China [19,20,23,25,26,30,31], 2 from Canada [21,33], 2 from South Korea [37,39], 2 from Jordan [17,32], 1 from the United Kingdom [18], 1 from Switzerland [43], 1 from Poland [34], 1 from Italy [35], 1 from Argentina [42], 1 from Switzerland, Germany, and Austria [22], and 1 from Romania, Spain, and Scotland [40]. Among these trials, 24 trials were 2-arm RCTs [17-19,21,22,24,26,28-34,36-45], 3 trials were 3-arm RCTs [23,25,27], 1 trial was a 4-arm RCT [35], and 1 trial was a 5-arm RCT [20]. Sixteen studies used active controls [17,18,20,25-29,31,32,34,37,39,40,42,45], and 13 studies used passive controls [19,21-24,30,33,35,36,38,41,43,44]. All studies were published in English.

Half of the studies (15/29, 52%) were conducted in only depression or anxiety participants [18,19,21,23-30,34,40,42,45], while the remainder (14/29, 48%) involved participants with comorbid symptoms (eg, arthritis or diabetes, frequent headaches, irritable bowel syndrome, stress symptoms, and attention deficit) [17,20,22,31-33,35-39,41,43,44]. The included studies reported data from a total of 5686 participants. Study sample sizes ranged from 18 to 1489. The average age ranged from 18.9 to 56.5 years, and the proportion of female participants ranged from 36.7% to 100%. The most commonly used measure of depression severity among the included studies was the Patient Health Questionnaire-9 (17/29) [18,20-23,25-27,29-31,33,34,36,42,43,45]. For anxiety severity, the most commonly used measure was the Generalized Anxiety Disorder-7 (17/29) [18-23,26,27,29,30,33,35,38,41-43,45]. Regarding the statistical approach for handling missing data, 13 of the 27 included RCTs used an intention-to-treat analysis [17-20,23,25,26,28-30,35,37,39], while the remaining 16 trials reported results based on completer analysis [21,22,24,27,31-34,36,38-43,45].

Intervention and chatbot characteristics are shown in Table S5 in Multimedia Appendix 1 [17-45]. In the included studies, multiple CBT intervention components for psychological improvement were involved. Psychoeducation provides knowledge about psychological processes and therapeutic principles. Cognitive strategies identify and challenge biases or distortions, often using methods such as the ABC model and cognitive restructuring to modify maladaptive thoughts. Behavioral techniques, including behavioral activation, exposure, and behavioral experiments, were used to adjust maladaptive behavior patterns, increase engagement in valued activities, and reduce avoidance. Emotion-focused approaches targeted the identification, understanding, and regulation of emotions. Problem-solving therapy helps participants address life challenges by enhancing their problem-solving skills. Other components involve sleep interventions, relaxation, mindfulness, gratitude practices, and self-monitoring.

Five studies recommended that participants interact freely with the chatbot [21,22,26,34,39], while 13 studies required daily access [19,20,23,25,27,28,30,31,36-38,44,45]. The duration of chatbot interventions varied considerably, ranging from 1 week to 16 weeks. Eleven studies reported follow-up outcomes: 5 provided follow-up assessments at 1 month [20,23,25,30,34], 2 at 8 weeks [18,44], 3 at 3 months [17,35,36], and 1 at both 3 and 6 months [41]. More than half of the studies (n=16) implemented chatbots as independent, stand-alone systems (eg, MISHA, Wysa, Minder, and TEO) [18-22,28,29,33,35-40,43,45], while 13 studies deployed chatbots within multimodel platform (eg, Website, Facebook message, WeChat, and Amazon's Alexa) [17,23-27,30-32,34,41,42,44]. Thirteen studies used artificial intelligence (AI)–based chatbots that use machine learning, natural language processing, or other AI technologies to interpret users' input and generate responses [19,24-29,33-35,38,42,45]. In contrast, 16 studies implemented rule-based chatbots, where responses were determined by predefined rules or decision trees [17,18,20-23,30-32,36,37,39-41,43,44]. Chatbots led and controlled the conversation in 58.6% (17/29) of the studies [17,20,22-25,27,28,32,33,36,38,41-45]. The input modalities for participants included written language and/or emoji (26/29, 90%) [17,19-23,25,27-45], spoken language only (1/23, 4%) [24], or a combination of written and spoken languages (2/29, 7%) [18,26]. The chatbot output mainly consisted of text (7/29, 24%) [18,23,26,30,33,35,36], voice and text (1/29, 3%) [24], text and emoji (3/29, 10%) [27,34,42], text and/or voice combined with image (2/29, 7%) [25,29], and text combined with other modalities such as audio, video, emojis, images, and infographics (16/29, 55%) [17,19-22,28,31,32,37-41,43-45].

Risk of Bias

Twenty-five studies [17-22,24-36,38,41-45] described adequate random sequence generation, indicating low risk. Allocation concealment was unclear in 16 studies [18,22,26,31-35,37-44]. Five studies [18,30,36,40,42] had an ROB arising from incomplete outcome data, and 4 studies [18,21,24,40] exhibited risk concerning selective reporting. The most prevalent issue involved blinding implementation. Sixteen studies [17-19,21,22,26,28-31,33,34,36,37,43,45] exhibited high ROB in the blinding of participants and personnel, while 9 studies [22,26,29,32,36,37,40,43,45] showed high risk in blinding of outcome assessment. These findings clearly indicate that achieving adequate blinding of participants and personnel poses a particular challenge in psychotherapy research. The ROB graph and ROB summary are shown in Figures 2A and 2B [17-45].

XSL•FO
RenderX

Figure 2. The results of the risk of bias assessment of included studies [17-45]. (**A**) Risk of bias summary and (**B**) risk of bias graph.

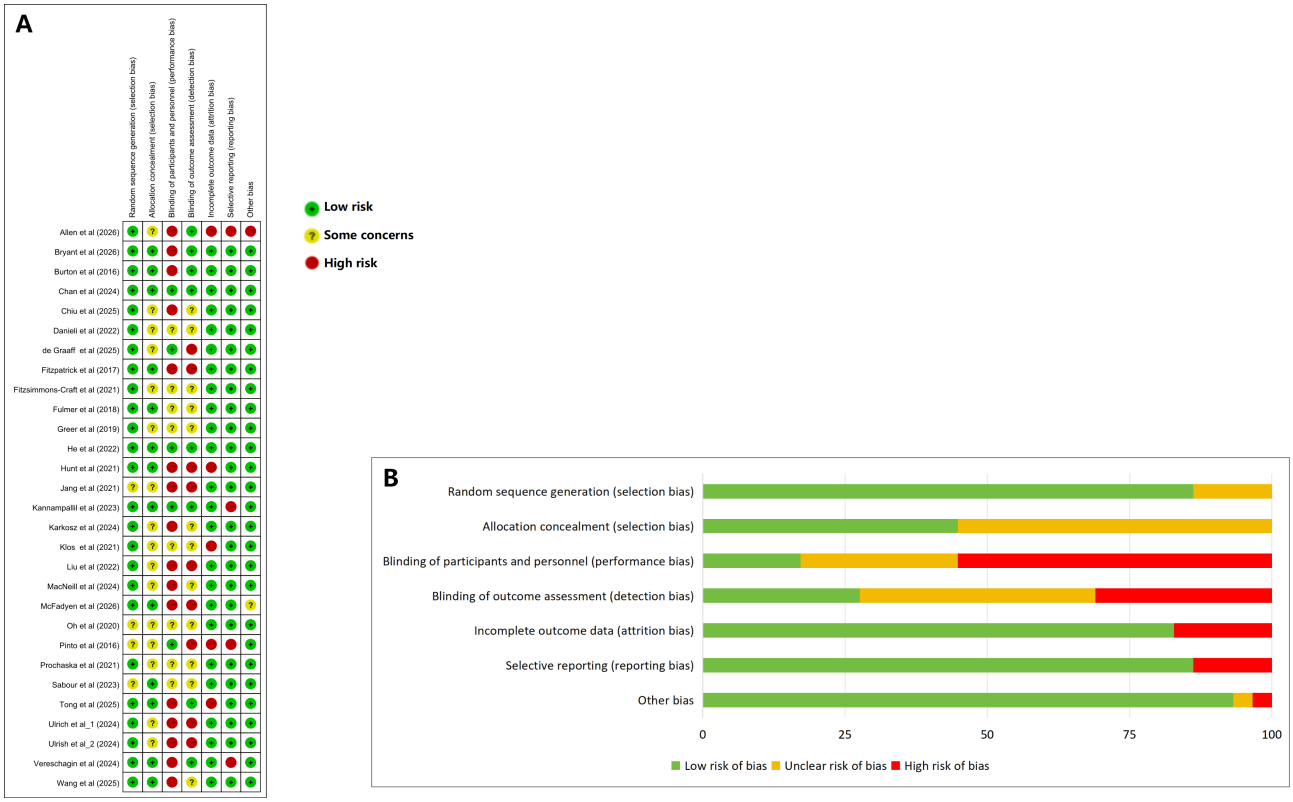

Quality of Evidence

According to the GRADE summary of findings, the main limitations included the high RoB, the substantial heterogeneity, and the imprecision among the RCTs. As a result, the certainty of evidence was rated as low for postintervention depressive and anxiety symptoms, as well as for long-term depressive symptoms, and very low for long-term anxiety symptoms (Table S6 in Multimedia Appendix 1).

Efficacy

Depressive Symptoms

Meta-analyses of 27 trials (4247 participants) [17-40,43-45] demonstrated significant effects of chatbot-based interventions decreasing postintervention depressive symptom scores (I^2=77.8%; P<.001), with a medium effect size (g=−0.55, 95% CI −0.70 to −0.40, 95% PI −1.23 to 0.13; Figure 3A). No evidence of significant small-study effects was found by the Egger test (z=−0.96, P=.335), and visual inspection of the funnel plot confirmed symmetry (Figure S1 in Multimedia Appendix 2). To explore potential sources of heterogeneity, a Galbraith plot was generated, identifying 5 outlier studies [18,21,29,34,38] (Figure S2 in Multimedia Appendix 2). A sensitivity analysis excluding the 5 studies revealed a substantial reduction in interstudy heterogeneity, with I^2 dropping from 77.8% to 45.7%. The repooled analysis demonstrated a slightly stronger and highly consistent medium effect size (g=−0.65, 95% CI −0.79 to −0.51). Notably, the 95% PI (−1.10 to −0.21) no longer crossed zero after the exclusion (Figure 3B). These findings not only verify the strong robustness of the primary results but also suggest a more consistent expected efficacy of chatbot-based interventions across future settings.

XSL•FO
RenderX

Figure 3. Forest plots of the efficacy of cognitive behavioral therapy–oriented chatbots on depressive symptoms at postintervention. (**A**) Primary meta-analysis of all included trials and (**B**) sensitivity analysis excluding 5 outlier studies. The green diamonds represent the pooled overall effect size and its 95% CI. The thick red horizontal lines placed directly underneath the summary diamonds represent the 95% prediction interval. REML; SMD: standardized mean difference [17-40,43-45].

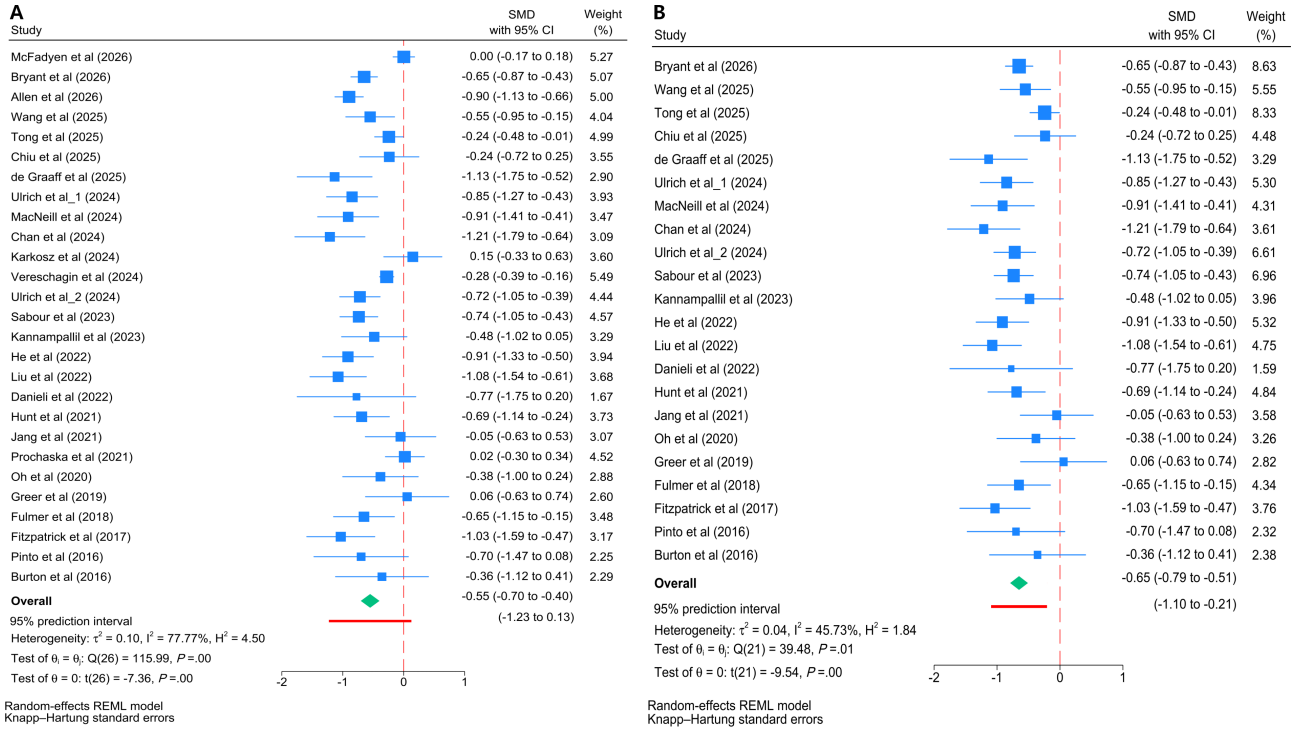

A total of 9 trials (1283 participants) [17,20,25,30,34-36,41,44] were included in the meta-analysis of long-term efficacy on depressive symptoms. The pooled effect size demonstrated a significant improvement favoring chatbot interventions over control conditions on depressive symptoms (I^2=63.6%, P=.004), representing a small effect size (g=−0.32, 95% CI −0.55 to −0.09, 95% PI −0.93 to 0.29; Figure 4). However, these results should be interpreted with caution, as the small number of trials may limit the statistical power of this follow-up analysis.

Figure 4. Forest plot of the long-term efficacy (follow-up) of cognitive behavioral therapy–oriented chatbots on depressive symptoms. The green diamond represents the pooled overall effect size and its 95% CI. The thick red horizontal line placed directly underneath the summary diamond represents the 95% prediction interval. REML: restricted maximum likelihood; SMD: standardized mean difference [17,20,25,30,34-36,41,44].

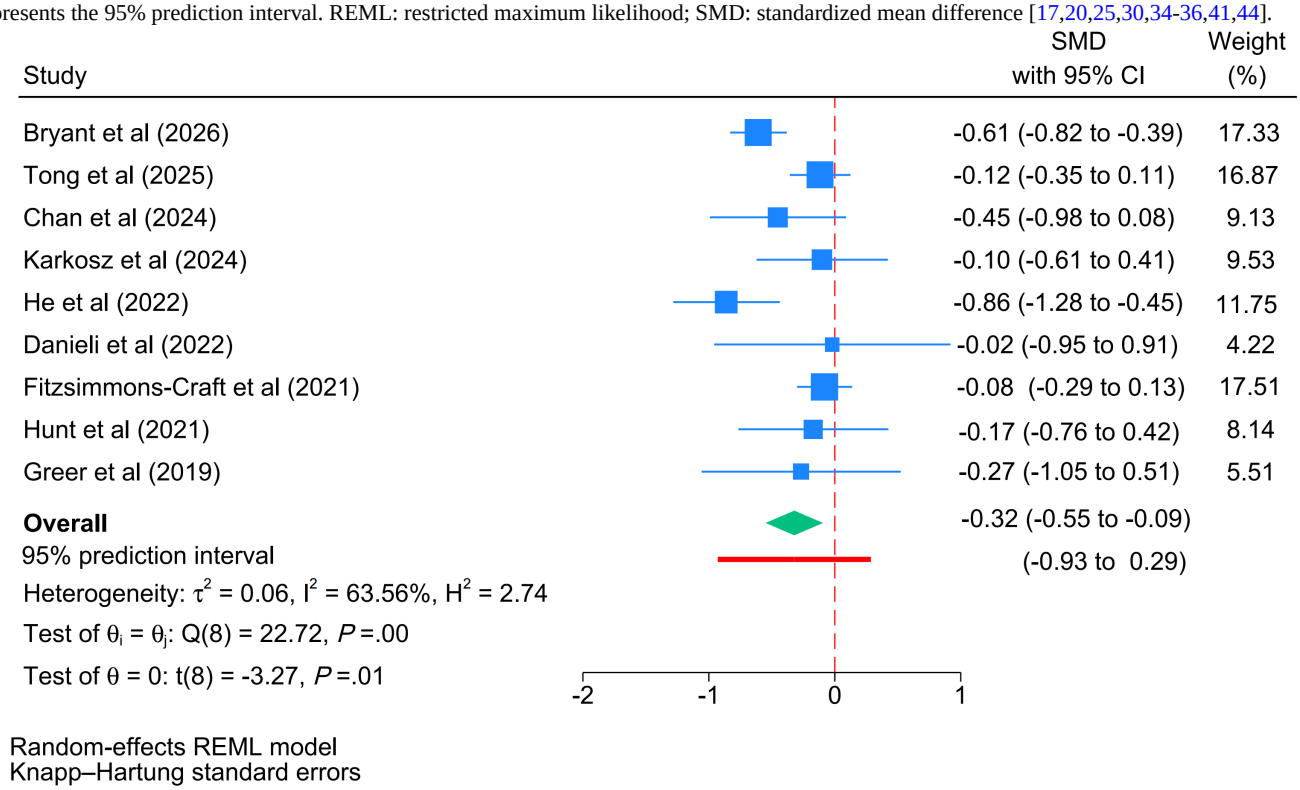

XSL•FO
RenderX

Anxiety Symptoms

Meta-analyses of 25 trials (4158 participants) [17-24,26,27,29-39,42-45] demonstrated significant effects of chatbot-based interventions in decreasing postintervention anxiety symptom scores (I^2=56.8%, P<.001), with a small effect size (g=−0.26, 95% CI −0.37 to −0.14, 95% PI −0.67 to 0.15; Figure 5 [17-24,26,27,29-39,42-45]). No significant small-study effects were indicated by the Egger test (z=−0.92, P=.356), and visual inspection of the funnel plot confirmed symmetry (Figure S1 in Multimedia Appendix 2). To explore potential sources of heterogeneity, a Galbraith plot was generated, identifying 3 outlier studies [17,29,45] (Figure S2 in Multimedia Appendix 2). A sensitivity analysis excluding the 3 studies revealed a substantial reduction in interstudy heterogeneity, with I^2 dropping from 56.8% to 0.0%. The repooled analysis demonstrated a slightly stronger and highly consistent small effect size (g=−0.28, 95% CI −0.35 to −0.20). Notably, the 95% PI (−0.36 to −0.20) no longer crossed zero after the exclusion (Figure 5). These findings not only verify the strong robustness of the primary results but also suggest a more consistent expected efficacy of chatbot-based interventions across future settings.

Figure 5. Forest plots of the efficacy of cognitive behavioral therapy–oriented chatbots on anxiety symptoms at postintervention. (**A**) Primary meta-analysis of all included trials and (**B**) sensitivity analysis excluding 3 outlier studies. The green diamonds represent the pooled overall effect size and its 95% CI. The thick red horizontal lines placed directly underneath the summary diamonds represent the 95% prediction interval. REML: restricted maximum likelihood; SMD: standardized mean difference [17-24,26,27,29-39,42-45].

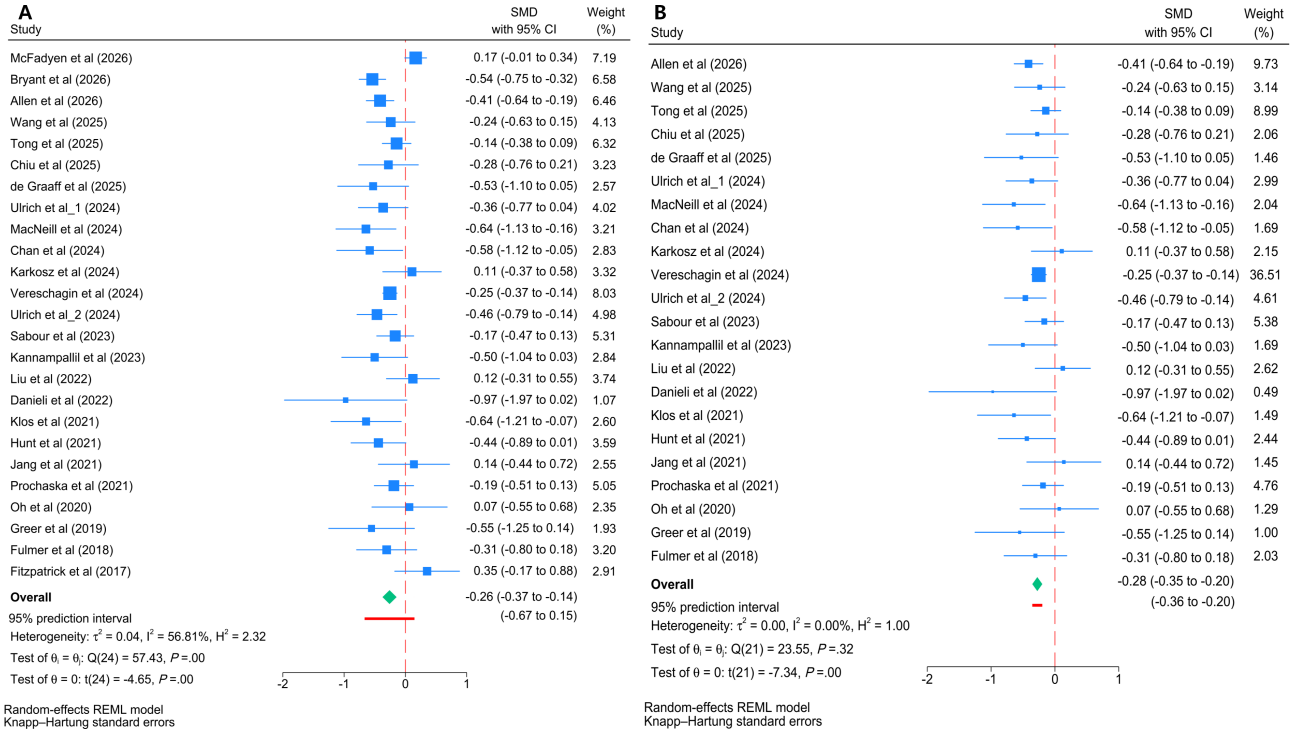

A total of 8 trials (1185 participants) [17,20,30,34-36,41,44] were included in the meta-analysis of long-term efficacy on anxiety symptoms. The pooled effect size demonstrated no significant improvement of chatbot interventions over control conditions on follow-up anxiety symptoms (I^2=65.74%, P=.001), representing a nonsignificant effect (g=−0.19, 95% CI −0.43 to 0.04, 95% PI −0.84 to 0.46; Figure 6 [17,20,30,34-36,41,44]). Given the limited number of studies, these findings should be viewed with caution due to potential constraints on statistical power.

XSL•FO
RenderX

Figure 6. Forest plot of the long-term efficacy (follow-up) of cognitive behavioral therapy–oriented chatbots on anxiety symptoms. The green diamond represents the pooled overall effect size and its 95% CI. The thick red horizontal line placed directly underneath the summary diamond represents the 95% prediction interval. REML: restricted maximum likelihood; SMD: standardized mean difference [17,20,30,34-36,41,44].

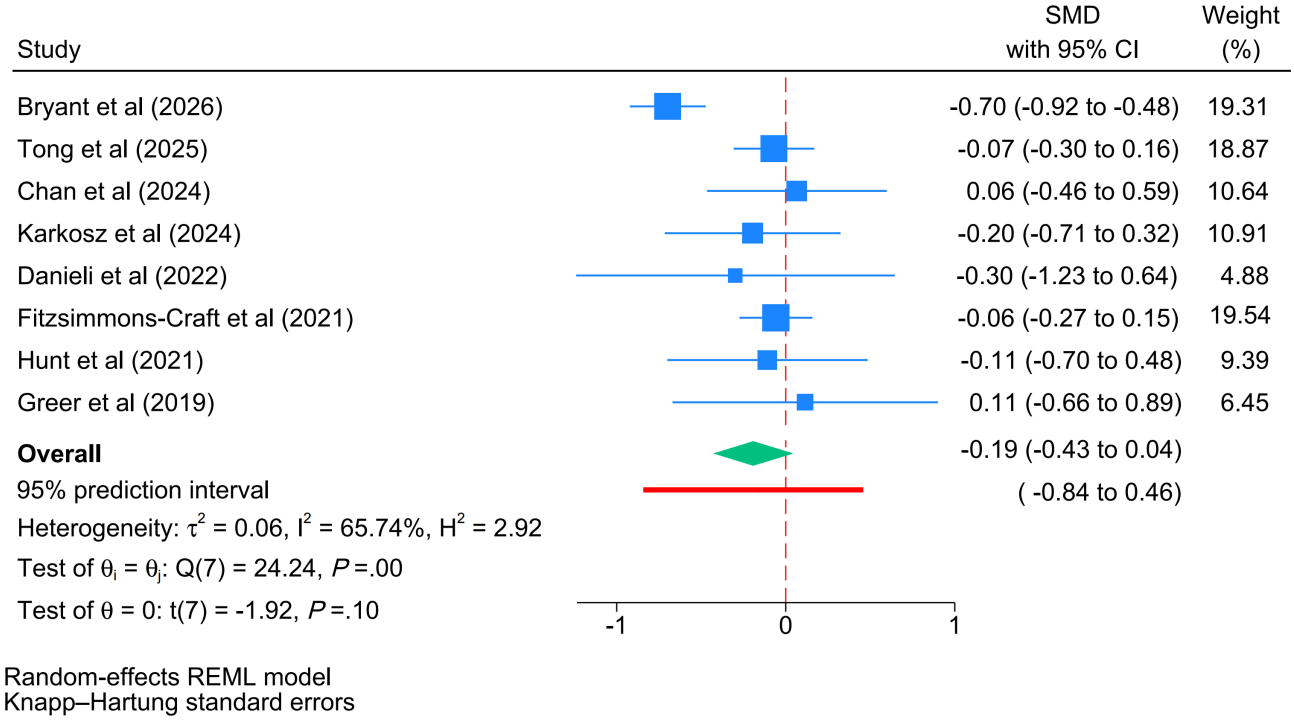

Moderators of Efficacy

Overview

Subgroup analyses were conducted to explore between-study variance and provide a more nuanced understanding of the efficacy of chatbot interventions. The pooled standardized mean difference for each subgroup can be found in Table 1. Meta-regressions were conducted to assess whether age and proportion of females influenced the short-term effect size of CBT-oriented chatbots on depressive and anxiety symptoms. Additionally, to assess the relative importance of these moderators, AIC- or BIC-based multimodel inference was conducted across 3 predefined models (Table S7 in Multimedia Appendix 1).

XSL•FO
RenderX

Table . Subgroup analyses for depressive and anxiety symptoms.

	Number of studies	Heterogeneity		Meta-analysis		
		I^2 (%)	*P* value	*g*	95% CI	*P* value
Total depressive symptoms: EOT[a]						
Participants symptom profiles						.80
Depression or anxiety-only	14	82.99	<.001	−0.54	−0.76 to −0.31	
Comorbid symptoms	13	67.65	<.001	−0.57	−0.82 to −0.32	
Control condition						.46
Active control	15	79.13	<.001	−0.60	−0.85 to −0.35	
Passive control	12	70.05	<.001	−0.49	−0.69 to −0.29	
Intervention duration						.40
1-4 weeks	15	74.03	<.001	−0.49	−0.69 to −0.29	
5-8 weeks	9	84.17	<.001	−0.61	−0.95 to −0.28	
9-16 weeks	3	34.43	.249	−0.78	−1.59 to 0.03	
Platform						.94
Multimodel	11	73.50	<.001	−0.55	−0.82 to −0.27	
Stand-alone	16	80.00	<.001	−0.56	−0.76 to −0.35	
Chatbot type						.64
Rule-based	13	77.37	<.001	−0.59	−0.81 to −0.37	
AI[b]-based	14	72.80	<.001	−0.52	−0.76 to −0.27	
Country						.32
Western	16	80.33	<.001	−0.49	−0.70 to −0.28	
Eastern	11	68.05	.001	−0.64	−0.89 to −0.39	
ROB[c]						.11
Low risk	4	33.34	.178	−0.78	−1.23 to −0.32	
Unclear	10	61.72	.004	−0.39	−0.66 to −0.12	
High risk	13	83.81	<.001	−0.59	−0.83 to −0.36	
Total depressive symptoms: FU[d]						
Follow-up duration						.75
<3 months	5	59.56	.036	−0.36	−0.77 to 0.06	
≥3 months	4	70.57	.006	−0.28	−0.74 to 0.17	
Total anxiety symptoms: EOT						
Participants symptom profiles						*.008*[e]
Depression or anxiety-only	12	64.70	.001	−0.15	−0.32 to 0.02	
Comorbid symptoms	13	5.88	.387	−0.41	−0.54 to −0.28	
Control condition						.37

XSL•FO
RenderX

	Number of studies	Heterogeneity		Meta-analysis		
		I^2 (%)	*P* value	*g*	95% CI	*P* value
Active control	13	68.86	<.001	−0.18	−0.39 to 0.02	
Passive control	12	0.00	.535	−0.28	−0.37 to −0.19	
Intervention duration						.40
1-4 weeks	13	5.16	.162	−0.23	−0.34 to −0.12	
5-8 weeks	10	66.81	<.001	−0.38	−0.59 to −0.17	
9-16 weeks	2	69.13	.072	−0.17	−4.13 to 3.79	
Platform						.71
Multimodel	11	41.17	.074	−0.29	−0.46 to −0.11	
Stand-alone	14	66.76	<.001	−0.24	−0.42 to −0.07	
Chatbot type						.13
Rule-based	13	26.30	.257	−0.33	−0.44 to −0.23	
AI-based	12	59.68	.003	−0.16	−0.39 to 0.07	
Country						.67
Western	15	66.34	<.001	−0.29	−0.46 to −0.12	
Eastern	10	42.98	.063	−0.24	−0.42 to −0.06	
ROB						*.02*
Low risk	2	0.00	.882	−0.55	−0.74 to −0.35	
Unclear	10	0.00	.445	−0.24	−0.40 to −0.07	
High risk	13	69.43	<.001	−0.22	−0.40 to −0.04	
Total anxiety symptoms: FU						
Follow-up duration						.24
<3 months	4	0.00	.876	−0.06	−0.21 to 0.09	
≥3 months	4	77.40	<.001	−0.32	−0.84 to 0.21	

[a]EOT: end of treatment.

[b]AI: artificial intelligence.

[c]ROB: risk of bias.

[d]FU: follow up.

[e]Italicized values are significant at *P*<.05.

Depressive Symptoms

No significant between-subgroup differences on depressive symptoms were observed across all subgroup factors (all *P*>.05). For participants' symptom profiles, both adults with depression or anxiety only and those with comorbid symptoms showed significant improvements in depression, with moderate effects (*g*=−0.54 and −0.57, respectively). Under control conditions, interventions compared with an active control group yielded a larger effect (*g*=−0.60) than those compared with a passive control (*g*=−0.49). Regarding intervention duration, effects were strongest for interventions lasting 9 - 16 weeks (*g*=−0.78), followed by 5 - 8 weeks (*g*=−0.61), and 1 - 4 weeks (*g*=−0.49). In terms of platform types, both multimodel and stand-alone platform groups significantly affected depression, with effect sizes of −0.55 and −0.56, respectively. For chatbot types, rule-based and AI-based groups showed similar effect sizes (*g*=−0.59 and −0.52, respectively). Regarding the study location, interventions conducted in Eastern countries showed a larger effect size (*g*=−0.64) than those conducted in Western countries (*g*=−0.49). In terms of ROB, studies with a low risk demonstrated the largest effect (*g*=−0.78), followed by those with a high risk (*g*=−0.59) and an unclear risk (*g*=−0.39). Meta-regression analyses suggested that neither the mean sample age (β=.003, *P*=.715) nor the proportion of females (β=.002, *P*=.804) had a significant impact on depressive symptoms (Figure S3 in Multimedia Appendix 2).

For studies with a follow-up period of less than 3 months, a small but nonsignificant effect was observed (*g*=−0.36, 95% CI −0.77 to 0.06). Similarly, for studies with a follow-up period of 3 months or longer, the effect size was also not statistically

XSL•FO
RenderX

significant (g=−0.28, 95% CI −0.74 to 0.17). Furthermore, no significant difference was observed between these 2 follow-up durations (P=.746). This suggests that intervention effects were not statistically significant at follow-up, regardless of whether the duration was shorter or longer than 3 months.

Anxiety Symptoms

Significant between-subgroup differences on anxiety symptoms were observed for participants' symptom profiles (P=.008) and ROB (P=.022), while no significant differences were found across the other subgroup factors (all P>.05). As for participants' characteristics, a larger effect size was observed in adults with comorbid symptoms (g=−0.41) than in adults with depression or anxiety-only symptoms (g=−0.15). Under control conditions, the therapeutic effect of chatbots was more pronounced for passive control groups (g=−0.28) in contrast to active control groups (g=−0.18). Regarding intervention characteristics, the effect was stronger for interventions lasting 5 - 8 weeks (g=−0.38) than for interventions lasting 1 - 4 weeks (g=−0.23) or 9 - 16 weeks (g=−0.17). Additionally, interventions delivered through multimodal platforms yielded a slightly larger effect (g=−0.29) than those using stand-alone platforms (g=−0.24). Similarly, the rule-based group showed a marginally larger effect size (g=−0.33) compared with the AI-based group (g=−0.16). Regarding the study location, interventions conducted in Western countries showed a slightly larger effect (g=−0.29) than those in Eastern countries (g=−0.24). For ROB, studies with a low risk demonstrated the largest effect (g=−0.55), followed by those with an unclear risk (g=−0.24) and a high risk (g=−0.22). Meta-regression analyses suggested that mean sample age (β=−.006, P=.394) had no impact on anxiety symptoms. However, a negative association was identified between the proportion of female participants and the pooled effect size (β=−.014, P=.022), with a higher proportion of females associated with a larger absolute effect size (Figure S3 in Multimedia Appendix 2). Subgroup analysis by follow-up length showed that chatbot interventions had no significant long-term effect on anxiety symptoms, whether the follow-up period was less than 3 months (g=−0.06, 95% CI −0.21 to 0.09) or 3 months and longer (g=−0.32, 95% CI −0.84 to 0.21). Additionally, there was no significant difference observed between these 2 follow-up durations (P=.237).

To explore the importance of moderators, we conducted a multimodal inference analysis using information criteria. For depressive symptoms, the model including study design characteristics (model 3: control type, country, and ROB) emerged as the best-fitting model. It explained the highest explained variance (R^2=16.96%) and yielded the lowest AIC (33.30) and residual heterogeneity (τ^2=0.08, I^2=70.19%). For anxiety symptoms, model 3 (study design) again demonstrated the best fit, with the highest R^2 (63.53%), lowest AIC (8.16), and lowest residual heterogeneity (τ^2=0.01, I^2=28.32%). Notably, all 3 models explained substantially more variance in anxiety symptoms (R^2 range: 39.52%-63.53%) than in depressive symptoms (R^2 range: 4.26%-16.96%).

Engagement

Engagement outcomes, synthesized from commonly reported metrics (reported by at least 2 studies), revealed distinct interaction patterns (Figure 7). Although the mean attrition was moderate (22%), participants engaged on average for 12.84 days over the 4-week intervention period (range 9 - 16 days). Program completion rates varied widely (2.8% - 94%), and reports of interaction duration were inconsistent (50 - 134 minutes). Notably, 1 study demonstrated intensive daily interaction (25.54 times per day), while only 10% of studies have reported interaction session metrics (mean sessions: 8.87). In line with these findings, the mean number of messages exchanged was substantial (494 messages per user), and participants completed an average of 7.59 modules.

XSL•FO
RenderX

Figure 7. Engagement outcomes for commonly reported metrics. This graph displays the average values for the main reported (at least 2 studies) engagement metrics, calculated from the means reported by the studies.

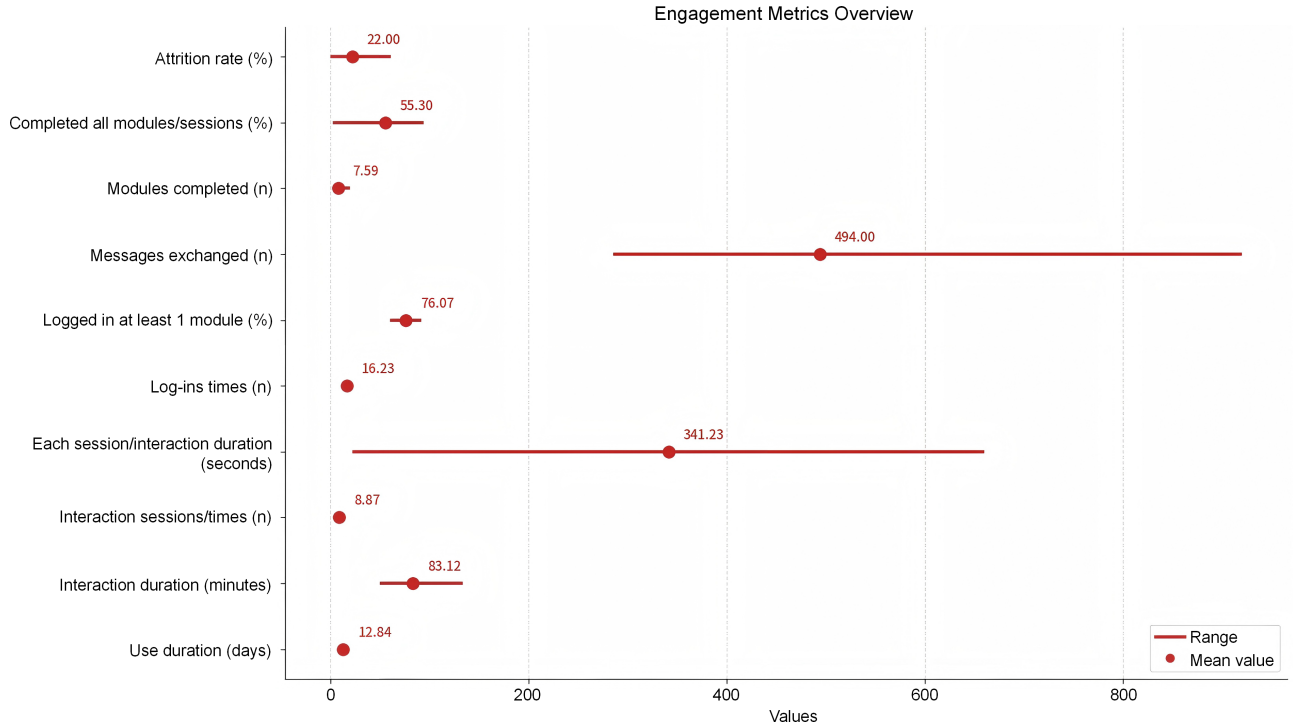

Acceptability

Two studies [22,43] assessed overall acceptability using the user version of the Mobile App Rating Scale (total score range 1 - 5), reporting an average score of 3.99. One study [38] measured overall acceptability via the Usage Rating Profile—Intervention (total score range 6 - 36), reporting a mean score of 29, which indicated high acceptability. User satisfaction with chatbot interventions was assessed in 9 studies (9/21, 42%) [23,25,27,29,32,35,37,38,45], which included overall, content, functional, and design satisfaction. Among these, 5 studies [25,29,35,37,45] rated overall and content satisfaction using a 5-point Likert scale, and the average scores were 3.72 and 3.69, respectively. Two studies [27,38] reported rates of overall satisfaction with chatbots, with 86% and 83% of participants expressing satisfaction, respectively. One study [27] reported that 80% of participants expressed satisfaction with the therapeutic content delivered by chatbots. Two studies [32,38] evaluated satisfaction using the 8-item Client Satisfaction Questionnaire (total score range 6 - 36), the average score was 25.9. Four studies [23,28,38,44] assessed participants' intent to recommend the chatbots to others in the future, and a majority of participants (71% - 88%) demonstrated a positive likelihood of recommendation. Five studies [25,27,29,37,45] reported that participants felt that they had learned something new from the chatbot, 4 studies [25,27,29,37] noted that participants gained information relevant to daily life, 2 studies [27,37] described participants feeling comfortable with the therapeutic process, and 3 studies [25,27,45] reported increased emotional awareness among participants.

Qualitative feedback on user experience was examined in 14 studies. Participants highlighted several positive features of chatbots, including ease of access, empathy and friendliness, encouragement, usefulness and helpfulness of the information provided, convenience, interesting, and reliability. Negative feedback included technical issues (ie, bugs, glitches, lags, system crashes, and inflexible), content repetitiveness, unnatural or impersonal conversation, misunderstood replies, nonintuitive interfaces, not enough interactivity, and the perceptions that the chatbot was annoying or tedious.

Discussion

Main Findings

This systematic review and meta-analysis of RCTs provides a comprehensive synthesis of the efficacy, user engagement, and acceptability of CBT-oriented chatbots for adults with depressive and anxiety symptoms. The findings demonstrate that these interventions significantly alleviate both depressive and anxiety symptoms in the short term, yielding a medium effect size for depression and a small effect size for anxiety. Efficacy was largely consistent across various demographic and clinical factors for depression, whereas for anxiety, chatbots showed greater benefits for individuals with comorbid symptoms, in participant samples with a higher proportion of females, and in trials with higher methodological quality. Furthermore, our synthesis revealed high user acceptability and adequate engagement levels, although technical and personalization barriers remain. Overall, these findings support the usefulness of CBT-oriented chatbots as accessible DMHIs. However, to provide a clinically responsible interpretation, these promising results must be explicitly weighed against the substantial heterogeneity, prevalent ROBs, and the resulting low certainty of evidence.

Efficacy

The meta-analysis provides robust evidence supporting the short-term efficacy of CBT-oriented psychological chatbots in

XSL•FO
RenderX

alleviating depressive symptoms among adults. The observed moderate effect is notably larger than the magnitude reported in a recent broad-spectrum review that included a variety of chatbot interventions [52]. This discrepancy likely stems from our refined inclusion criteria, which were restricted to chatbot interventions based on CBT technology. Given that CBT has consistently demonstrated superiority over other psychotherapeutic interventions and is considered the first-line psychosocial treatment of choice for anxiety and depressive disorders [71,72], this specialized focus may explain the enhanced therapeutic impact. The efficacy for anxiety symptoms was relatively small, which is consistent with previous meta-analyses [49,52]. Our finding indicated that structured cognitive-behavioral components, such as cognitive restructuring, psychoeducation, and exposure techniques, can be effectively delivered through digital interfaces to target anxiety-related symptoms [73,74]. When compared with traditional face-to-face or internet-delivered CBT, which typically yields moderate to large effect sizes for depression and anxiety [75-78], the effects of CBT chatbots are slightly more modest. However, considering the low-barrier and highly scalable nature of chatbots, this moderate efficacy remains clinically highly valuable, especially among individuals reluctant to pursue traditional face-to-face therapy [79].

While the 95% CIs confirmed a significant average therapeutic effect for both depression and anxiety across all included trials, the initial 95% PIs—which estimate the distribution of true effects in future or individual clinical contexts—were wide and crossed zero. This critical distinction indicates that although chatbot interventions are effective on average, their actual therapeutic impact in a new, specific setting might range from highly beneficial to negligible. However, once distinct outlier studies were excluded in the sensitivity analyses, the PIs stabilized and no longer crossed zero. Notably, a closer inspection of these outliers revealed striking methodological consistencies: the majority exhibited a high ROB and predominantly used active control conditions rather than passive controls. In meta-analytic research, comparing an intervention with an active control typically yields smaller relative effect sizes, a finding that, when compounded by high risks of bias, likely drove these studies' extreme statistical deviations [80,81]. This underscores that the observed heterogeneity is heavily influenced by specific study design choices, particularly the selection of comparator groups and methodological limitations, rather than purely clinical variance among target populations. Consistent with GRADE assessment, this implies that the real-world expected effectiveness of CBT chatbots depends profoundly on rigorous clinical implementation and appropriate baseline comparisons.

Specifically, interpreting these overall efficacy findings requires explicit consideration of 3 interconnected methodological constraints. First, the GRADE results indicate that the overall certainty of evidence ranged from low (for short-term outcomes and long-term depression) to very low (for long-term anxiety). This highlights that our confidence in the precise effect estimates is limited, and future rigorous trials are highly likely to influence these conclusions. Second, the substantial heterogeneity underscores that chatbot efficacy is not universally uniform; while average effects are significant, the actual therapeutic impact is highly sensitive to specific study designs and unmeasured clinical contexts. Third, the ubiquitous ROB, predominantly driven by the lack of participant and personnel blinding, is an inescapable challenge in digital psychotherapy research that can potentially inflate self-reported outcomes [82]. However, it is noteworthy that in our subgroup analyses, studies rated as having a low overall ROB actually demonstrated the largest effect sizes for both depression and anxiety. This nuanced finding implies that while performance bias is prevalent, the therapeutic effects of chatbots are unlikely to be merely an artifact of poor methodological quality. Together, these 3 factors—low GRADE certainty, high heterogeneity, and inherent blinding limitations—dictate that CBT-oriented chatbots should currently be recommended as supplementary clinical tools rather than stand-alone replacements for standard care [83,84].

Regarding long-term sustainability, the therapeutic benefits for depression attenuated over time, and no significant long-term differences were observed for anxiety. This fragility of long-term effects is a common challenge in digital interventions studies [85,86]. Anxiety and depressive symptoms often require ongoing reinforcement of adaptive coping skills; without booster sessions or the therapeutic alliance inherent in face-to-face therapy, the initial gains achieved via chatbots may diminish [87].

Moderators of Efficacy

For depressive symptoms, the therapeutic efficacy remained remarkably consistent across all examined clinical, demographic, and technological factors. This broad consistency suggests that the core cognitive-behavioral mechanisms delivered via chatbots are robust and highly generalizable for depression. Furthermore, multimodel inference indicated that study design characteristics were the primary drivers of the observed heterogeneity, implying that variance in depression outcomes is largely methodological rather than clinical. While our examined moderators accounted for some of this variance, they did not fully explain it. This residual, unmeasured heterogeneity likely stems from several confounding factors (eg, the specific CBT components) that were inconsistently reported across primary studies [88]. Future studies should systematically report these variables to facilitate more granular meta-analytic investigations.

Conversely, the efficacy for anxiety symptoms was significantly influenced by specific clinical and demographic variables. Notably, chatbots were more effective for adults with comorbid symptoms than for those presenting with anxiety or depression only. This suggests that CBT-oriented chatbots may successfully target transdiagnostic mechanisms—such as cognitive appraisals or emotion exposure—that are prevalent in complex clinical presentations, thereby yielding broader symptom relief [89]. Meta-regression revealed that a higher proportion of female participants was associated with a larger absolute effect size for anxiety reduction. This aligned with previous digital health research indicating that women tend to show greater engagement and adherence to self-guided psychological interventions, possibly due to more favorable attitudes toward emotional expression and help-seeking behavior [90,91]. However, given the limited number of studies included in this meta-regression, these findings should be interpreted with caution. Future

XSL•FO
RenderX

research with larger samples is needed to confirm the moderating role of gender composition and explore underlying mechanisms more thoroughly. Additionally, multimodel inference identified study design characteristics as the primary drivers of heterogeneity in anxiety outcomes, substantially outperforming patient profiles and technical features. Specifically, trials with a low ROB yielded significantly larger effect sizes. This highlights that variance in anxiety efficacy is fundamentally methodological, underscoring the imperative for rigorous trial designs to accurately evaluate these digital interventions.

Engagement

Overall, participant engagement with CBT-oriented chatbots demonstrated promising adherence levels. The average attrition rate in the included studies was 22%. Among these studies, 27 out of 29 (93.1%) studies revealed attrition rates lower than 50%, which is the recommended cutoff threshold for effective DMHIs [92]. This indicates a strong user willingness to interact with CBT-oriented psychological chatbots.

However, synthesizing these engagement outcomes was complicated by widespread inconsistencies in how they were defined and measured, as well as in their levels, which is a common challenge in the field of digital interventions [55,93]. Furthermore, instances where attrition rates were exceptionally high, participant dropout was directly linked to chatbots providing inaccurate or irrelevant responses, which hindered the development of user-chatbot rapport [42].

Given that higher usage rates are generally associated with greater treatment outcomes [94], optimizing engagement remains critical. Future studies should aim to systematically define and standardize engagement metrics, while also exploring user-centered design strategies such as personalized reminders, adaptive pacing, and flexible content delivery.

Acceptability

Overall, CBT-oriented chatbots appeared to be perceived more positively than negatively by users from various backgrounds regarding acceptability. Across the studies that assessed satisfaction, participants generally expressed high levels of satisfaction with the chatbots, including satisfaction of content, functionality, and design. This suggests that participants found the chatbots generally helpful, accessible, and engaging. Study participants appreciated the features such as convenience, empathy, encouragement, and ease of use. Furthermore, many participants indicated that they would recommend chatbot interventions to others, suggesting considerable potential for broader acceptance. These findings are partially aligned with previous systematic reviews reporting that chatbots provided helpful information and were easy to use [95,96].

Nonetheless, several issues identified from participant feedback require attention in future development. Common criticisms included technical problems, content repetitiveness, impersonal interactions, and the lack of natural conversation flow. These issues are critical, as they can significantly undermine user engagement and potentially hinder therapeutic outcomes [96,97]. Therefore, future research should focus on improving the chatbot's conversational quality, making interactions more human-like, and reducing technical errors to improve acceptability.

Implications and Recommendations for Future Studies

The integration of these findings provides actionable guidelines for both clinical implementation and future research: (1) Clinical implementation: CBT chatbots should be positioned within stepped care models as highly scalable, first-line triage interventions, particularly for populations with comorbid mild to moderate symptoms or those reluctant to seek face-to-face care. (2) Standardization of metrics: future RCTs must establish and adhere to standardized engagement metrics (eg, explicit definitions of module completion, log-in frequency, and dropout rates) to allow for more precise meta-analytic syntheses of user adherence. (3) Technological advancement: given user frustrations with rigid, rule-based scripts, future chatbot development should carefully integrate large language models (LLMs). Researchers must design studies to evaluate whether the adaptive, human-like conversational flows of LLMs can improve long-term engagement and efficacy, while ensuring that they strictly adhere to evidence-based CBT frameworks without hallucinating clinical advice. (4) Long-term follow-up: future trials should prioritize extended follow-up periods and test the integration of automated booster sessions to sustain therapeutic gains.

Limitations

Limitations of the Evidence

First, according to the GRADE summary of findings, the certainty of evidence for primary outcomes ranged from very low to low. This was predominantly driven by a high RoB concerning the blinding of participants and personnel, which may inflate self-reported symptom improvements. Second, many studies lacked long-term follow-up data, restricting conclusions about the durability of treatment effects. Third, engagement and acceptability outcomes were inconsistently reported, and key indicators such as log-in frequency, dropout rates, or module completion were available in only a subset of studies, limiting the precision of pooled estimates.

Limitations of the Review

First, despite our rigorous inclusion criteria, there remained substantial methodological and technological heterogeneity across the included interventions (eg, varying intervention durations and chatbot functionalities), which limits the broader generalizability of our findings to all digital mental health contexts. Second, although subgroup and meta-regression analyses were conducted, the number of studies available for some analyses may have been insufficient to detect small moderator effects with adequate statistical power and should be interpreted with caution. Finally, this review predominantly captures rule-based or narrowly trained AI systems. As the field rapidly evolves, the absence of studies evaluating modern LLM-powered chatbots limits our ability to comment on the next generation of digital therapeutics.

Conclusions

This systematic review and meta-analysis is innovative in being the first to systematically isolate theoretically grounded,

XSL•FO
RenderX

CBT-oriented chatbots from the broader landscape of digital conversational agents. Unlike broader review studies that pool disparate interventions, our targeted multidimensional approach provides a precise, methodology-driven evaluation of a specific digital therapeutic. By doing so, this review brings to the field critical evidence that these tailored interventions yield significant short-term relief for both depressive and anxiety symptoms. Notably, we identified a specific therapeutic advantage for anxiety management among individuals with comorbid profiles via transdiagnostic mechanisms. However, our findings also highlight the fragility of long-term effects and demonstrate that outcome consistency is profoundly dependent on rigorous study designs and appropriate active comparators. In the real world, the implications of these findings are profound: CBT chatbots hold immense potential to democratize mental health care as scalable, low-barrier first-line tools within stepped care models. To fulfill this potential and sustain long-term engagement, future developments must evolve from rigid, rule-based scripts toward adaptive, LLM-driven architectures, provided they strictly adhere to clinical safety and evidence-based CBT frameworks.

Acknowledgments

The authors thank all the reviewers for their assistance and support. The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: (1) proofreading and editing, and (2) adapting and adjusting emotional tone. The GAI tool used was: ChatGPT-5.0. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. The declaration has been submitted by the authors.

Funding

The study was funded by the National Natural Science Foundation of China (nos. 32200907 and U2336214), the China Postdoctoral Science Foundation (no. 2025M783475), the Postdoctoral Fellowship Program of China Postdoctoral Science Foundation (no. GZC20252270), the Humanities and Social Science Fund Project from the Ministry of Education of China (no. 25YJCZH055), and the ShuiMu Tsinghua Scholar Program (no. 2024SM190). The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Data Availability

Data sharing is not applicable to this systematic review and meta-analysis as no new data were created in this study.

Authors' Contributions

Conceptualization: BG, WM
Data curation: BG, HX
Formal analysis: BG
Investigation: BG, HX
Supervision: WM
Writing – original draft: BG
Writing – review & editing: BG, NY, HB, TK, CH, WM
All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies, summary of multiarm studies and data extraction or combination strategies, correlation matrix, study and participants characteristics, intervention and chatbot characteristics, GRADE (Grading of Recommendations Assessment, Development, and Evaluation) summary of findings, and multivariable meta-regression.
[DOCX File, 67 KB - jmir_v28i1e82677_app1.docx]

Multimedia Appendix 2
Funnel plots, Galbraith plots, and meta-regression bubble plots.
[DOCX File, 714 KB - jmir_v28i1e82677_app2.docx]

Checklist 1
PRISMA 2020 expanded checklist, PRISMA 2020 for Abstracts Checklist, and PRISMA-S checklist.
[DOCX File, 599 KB - jmir_v28i1e82677_app3.docx]

XSL•FO
RenderX

References

1. Ferrari AJ, Santomauro DF, Aali A. Global incidence, prevalence, years lived with disability (YLDs), disability-adjusted life-years (DALYs), and healthy life expectancy (HALE) for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet 2024 May 18;403(10440):2133-2161. [doi: 10.1016/S0140-6736(24)00757-8] [Medline: 38642570]
2. Collaborators GMD. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022 Feb;9(2):137-150. [doi: 10.1016/S2215-0366(21)00395-3]
3. Kalin NH. Anxiety, depression, and suicide in youth. Am J Psychiatry 2021 Apr 1;178(4):275-279. [doi: 10.1176/appi.ajp.2020.21020186] [Medline: 33789454]
4. Simon GE, Moise N, Mohr DC. Management of depression in adults: a review. JAMA 2024 Jul 9;332(2):141-152. [doi: 10.1001/jama.2024.5756] [Medline: 38856993]
5. Chodavadia P, Teo I, Poremski D, Fung DSS, Finkelstein EA. Prevalence and economic burden of depression and anxiety symptoms among Singaporean adults: results from a 2022 web panel. BMC Psychiatry 2023 Feb 14;23(1):104. [doi: 10.1186/s12888-023-04581-7] [Medline: 36782116]
6. The Lancet Global Health. Mental health matters. Lancet Glob Health 2020 Nov;8(11):e1352. [doi: 10.1016/S2214-109X(20)30432-0] [Medline: 33069297]
7. Cuijpers P, Quero S, Noma H, et al. Psychotherapies for depression: a network meta-analysis covering efficacy, acceptability and long-term outcomes of all main treatment types. World Psychiatry 2021 Jun;20(2):283-293. [doi: 10.1002/wps.20860] [Medline: 34002502]
8. David D, Cristea I, Hofmann SG. Why cognitive behavioral therapy is the current gold standard of psychotherapy. Front Psychiatry 2018;9:4. [doi: 10.3389/fpsyt.2018.00004] [Medline: 29434552]
9. Malhi GS, Bell E, Bassett D, et al. The 2020 Royal Australian and New Zealand College of Psychiatrists clinical practice guidelines for mood disorders. Aust N Z J Psychiatry 2021 Jan;55(1):7-117. [doi: 10.1177/0004867420979353]
10. Goetz CJ, Mushquash CJ, Maranzan KA. An integrative review of barriers and facilitators associated with mental health help seeking among indigenous populations. Psychiatr Serv 2023 Mar 1;74(3):272-281. [doi: 10.1176/appi.ps.202100503] [Medline: 36065579]
11. Haugen PT, McCrillis AM, Smid GE, Nijdam MJ. Mental health stigma and barriers to mental health care for first responders: a systematic review and meta-analysis. J Psychiatr Res 2017 Nov;94:218-229. [doi: 10.1016/j.jpsychires.2017.08.001] [Medline: 28800529]
12. Bhatti MM, Siddiqi N, Jennings H, Afaq S, Arsh A, Khan BA. Effectiveness, barriers, and facilitators of interventions delivered by traditional healers for the treatment of common mental disorders: a systematic review. Transcult Psychiatry 2024 Dec;61(6):885-904. [doi: 10.1177/13634615241273001] [Medline: 39314162]
13. Catania J, Beaver S, Kamath RS, et al. Evaluation of digital mental health technologies in the United States: systematic literature review and framework synthesis. JMIR Ment Health 2024 Aug 30;11:e57401. [doi: 10.2196/57401] [Medline: 39213023]
14. Khan K, Hollis C, Murphy T, Hall CL. Digital and remote behavioral therapies for treating tic disorders: recent advances and next steps. Front Psychiatry 2022;13:928487. [doi: 10.3389/fpsyt.2022.928487] [Medline: 35911244]
15. Xue J, Zhang B, Zhao Y, et al. Evaluation of the current state of chatbots for digital health: scoping review. J Med Internet Res 2023 Dec 19;25:e47217. [doi: 10.2196/47217] [Medline: 38113097]
16. Suganuma S, Sakamoto D, Shimoyama H. An embodied conversational agent for unguided internet-based cognitive behavior therapy in preventative mental health: feasibility and acceptability pilot trial. JMIR Ment Health 2018;5(3):e10454. [doi: 10.2196/10454]
17. Bryant RA, de Graaff AM, Habashneh R, et al. A guided chatbot-based psychological intervention for psychologically distressed older adolescents and young adults: a randomised clinical trial in Jordan. NPJ Digit Med 2026;9(1):57. [doi: 10.1038/s41746-025-02142-8]
18. Allen A, Young AH, Jellesma FC, et al. An exploratory randomized controlled trial of an AI-enabled mental health intervention for generalized anxiety. J Affect Disord 2026 May;401:121275. [doi: 10.1016/j.jad.2026.121275]
19. Wang Y, Li X, Zhang Q, Yeung D, Wu Y. Effect of a cognitive behavioral therapy–based AI chatbot on depression and loneliness in Chinese university students: randomized controlled trial with financial stress moderation. JMIR Mhealth Uhealth 2025;13:e63806-e63806. [doi: 10.2196/63806]
20. Chan WS, Cheng WY, Lok SHC, et al. Assessing the short-term efficacy of digital cognitive behavioral therapy for insomnia with different types of coaching: randomized controlled comparative trial. JMIR Ment Health 2024 Aug 7;11:e51716. [doi: 10.2196/51716] [Medline: 39110971]
21. Vereschagin M, Wang AY, Richardson CG, et al. Effectiveness of the minder mobile mental health and substance use intervention for university students: randomized controlled trial. J Med Internet Res 2024 Mar 27;26:e54287. [doi: 10.2196/54287] [Medline: 38536225]

XSL•FO
RenderX

22. Ulrich S, Gantenbein AR, Zuber V, Von Wyl A, Kowatsch T, Künzli H. Development and evaluation of a smartphone-based chatbot coach to facilitate a balanced lifestyle in individuals with headaches (BalanceUP App): randomized controlled trial. J Med Internet Res 2024 Jan 24;26:e50132. [doi: 10.2196/50132] [Medline: 38265863]
23. Sabour S, Zhang W, Xiao X, et al. A chatbot for mental health support: exploring the impact of Emohaa on reducing mental distress in China. Front Digit Health 2023;5:1133987. [doi: 10.3389/fdgth.2023.1133987] [Medline: 37214342]
24. Kannampallil T, Ajilore OA, Lv N, et al. Effects of a virtual voice-based coach delivering problem-solving treatment on emotional distress and brain function: a pilot RCT in depression and anxiety. Transl Psychiatry 2023 May 12;13(1):166. [doi: 10.1038/s41398-023-02462-x] [Medline: 37173334]
25. He Y, Yang L, Zhu X, et al. Mental health chatbot for young adults with depressive symptoms during the COVID-19 pandemic: single-blind, three-arm randomized controlled trial. J Med Internet Res 2022 Nov 21;24(11):e40719. [doi: 10.2196/40719] [Medline: 36355633]
26. Liu H, Peng H, Song X, Xu C, Zhang M. Using AI chatbots to provide self-help depression interventions for university students: a randomized trial of effectiveness. Internet Interv 2022 Mar;27:100495. [doi: 10.1016/j.invent.2022.100495] [Medline: 35059305]
27. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. JMIR Ment Health 2018 Dec 13;5(4):e64. [doi: 10.2196/mental.9782] [Medline: 30545815]
28. Burton C, Szentagotai Tatar A, McKinstry B, et al. Pilot randomised controlled trial of Help4Mood, an embodied virtual agent-based system to support treatment of depression. J Telemed Telecare 2016 Sep;22(6):348-355. [doi: 10.1177/1357633X15609793] [Medline: 26453910]
29. McFadyen J, Habicht J, Dina LM, Harper R, Hauser TU, Rollwage M. Increasing engagement with cognitive-behavioral therapy (CBT) using generative AI: a randomized controlled trial (RCT). Commun Med (Lond) 2026 Jan 15;6(1):129. [doi: 10.1038/s43856-025-01321-8] [Medline: 41540194]
30. Tong ACY, Wong KTY, Chung WWT, Mak WWS. Effectiveness of topic-based chatbots on mental health self-care and mental well-being: randomized controlled trial. J Med Internet Res 2025 Apr 30;27:e70436. [doi: 10.2196/70436] [Medline: 40306635]
31. Chiu YH, Lee YF, Lin HL, Cheng LC. Using cognitive behavioral therapy-based chatbots to alleviate symptoms of insomnia, depression, and anxiety: a randomized controlled trial. Health Informatics J 2025;31(4):14604582251396428. [doi: 10.1177/14604582251396428] [Medline: 41296540]
32. de Graaff AM, Habashneh R, Fanatseh S, et al. Evaluation of a guided chatbot intervention for young people in Jordan: feasibility randomized controlled trial. JMIR Ment Health 2025 Feb 5;12:e63515. [doi: 10.2196/63515] [Medline: 39908538]
33. MacNeill AL, Doucet S, Luke A. Effectiveness of a mental health chatbot for people with chronic diseases: randomized controlled trial. JMIR Form Res 2024 May 30;8:e50025. [doi: 10.2196/50025] [Medline: 38814681]
34. Karkosz S, Szymański R, Sanna K, Michałowski J. Effectiveness of a web-based and mobile therapy chatbot on anxiety and depressive symptoms in subclinical young adults: randomized controlled trial. JMIR Form Res 2024 Mar 20;8:e47960. [doi: 10.2196/47960] [Medline: 38506892]
35. Danieli M, Ciulli T, Mousavi SM, et al. Assessing the impact of conversational artificial intelligence in the treatment of stress and anxiety in aging adults: randomized controlled trial. JMIR Ment Health 2022 Sep 23;9(9):e38067. [doi: 10.2196/38067] [Medline: 36149730]
36. Hunt M, Miguez S, Dukas B, Onwude O, White S. Efficacy of Zemedy, a mobile digital therapeutic for the self-management of irritable bowel syndrome: crossover randomized controlled trial. JMIR Mhealth Uhealth 2021 May 20;9(5):e26152. [doi: 10.2196/26152] [Medline: 33872182]
37. Jang S, Kim JJ, Kim SJ, Hong J, Kim S, Kim E. Mobile app-based chatbot to deliver cognitive behavioral therapy and psychoeducation for adults with attention deficit: a development and feasibility/usability study. Int J Med Inform 2021 Jun;150:104440. [doi: 10.1016/j.ijmedinf.2021.104440] [Medline: 33799055]
38. Prochaska JJ, Vogel EA, Chieng A, et al. A randomized controlled trial of a therapeutic relational agent for reducing substance misuse during the COVID-19 pandemic. Drug Alcohol Depend 2021 Oct 1;227:108986. [doi: 10.1016/j.drugalcdep.2021.108986] [Medline: 34507061]
39. Oh J, Jang S, Kim H, Kim JJ. Efficacy of mobile app-based interactive cognitive behavioral therapy using a chatbot for panic disorder. Int J Med Inform 2020 Aug;140:104171. [doi: 10.1016/j.ijmedinf.2020.104171] [Medline: 32446158]
40. Pinto MD, Greenblatt AM, Hickman RL, Rice HM, Thomas TL, Clochesy JM. Assessing the critical parameters of eSMART-MH: a promising avatar-based digital therapeutic intervention to reduce depressive symptoms. Perspect Psychiatr Care 2016 Jul;52(3):157-168. [doi: 10.1111/ppc.12112] [Medline: 25800698]
41. Fitzsimmons-Craft EE, Chan WW, Smith AC, et al. Effectiveness of a chatbot for eating disorders prevention: a randomized clinical trial. Int J Eat Disord 2022 Mar;55(3):343-353. [doi: 10.1002/eat.23662] [Medline: 35274362]
42. Klos MC, Escoredo M, Joerin A, Lemos VN, Rauws M, Bunge EL. Artificial intelligence-based chatbot for anxiety and depression in university students: pilot randomized controlled trial. JMIR Form Res 2021 Aug 12;5(8):e20678. [doi: 10.2196/20678] [Medline: 34092548]

XSL•FO
RenderX

43. Ulrich S, Lienhard N, Künzli H, Kowatsch T. A chatbot-delivered stress management coaching for students (MISHA App): pilot randomized controlled trial. JMIR Mhealth Uhealth 2024 Jun 26;12:e54945. [doi: 10.2196/54945] [Medline: 38922677]
44. Greer S, Ramo D, Chang YJ, Fu M, Moskowitz J, Haritatos J. Use of the chatbot "Vivibot" to deliver positive psychology skills and promote well-being among young people after cancer treatment: randomized controlled feasibility trial. JMIR Mhealth Uhealth 2019 Oct 31;7(10):e15018. [doi: 10.2196/15018] [Medline: 31674920]
45. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 6;4(2):e19. [doi: 10.2196/mental.7785] [Medline: 28588005]
46. Limpanopparat S, Gibson E, Harris DA. User engagement, attitudes, and the effectiveness of chatbots as a mental health intervention: a systematic review. CHBAH 2024 Aug;2(2):100081. [doi: 10.1016/j.chbah.2024.100081]
47. Gaffney H, Mansell W, Tai S. Conversational agents in the treatment of mental health problems: mixed-method systematic review. JMIR Ment Health 2019 Oct 18;6(10):e14166. [doi: 10.2196/14166] [Medline: 31628789]
48. Li H, Zhang R, Lee YC, Kraut RE, Mohr DC. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med 2023 Dec 19;6(1):236. [doi: 10.1038/s41746-023-00979-5] [Medline: 38114588]
49. He Y, Yang L, Qian C, et al. Conversational agent interventions for mental health problems: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2023 Apr 28;25:e43862. [doi: 10.2196/43862] [Medline: 37115595]
50. Abd-Alrazaq AA, Rababeh A, Alajlani M, Bewick BM, Househ M. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. J Med Internet Res 2020 Jul 13;22(7):e16021. [doi: 10.2196/16021] [Medline: 32673216]
51. Lim SM, Shiau CWC, Cheng LJ, Lau Y. Chatbot-delivered psychotherapy for adults with depressive and anxiety symptoms: a systematic review and meta-regression. Behav Ther 2022 Mar;53(2):334-347. [doi: 10.1016/j.beth.2021.09.007] [Medline: 35227408]
52. Zhong W, Luo J, Zhang H. The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: a systematic review and meta-analysis. J Affect Disord 2024 Jul 1;356:459-469. [doi: 10.1016/j.jad.2024.04.057] [Medline: 38631422]
53. Suharwardy S, Ramachandran M, Leonard SA, et al. Feasibility and impact of a mental health chatbot on postpartum mental health: a randomized controlled trial. AJOG Glob Rep 2023 Aug;3(3):100165. [doi: 10.1016/j.xagr.2023.100165] [Medline: 37560011]
54. Kleinau E, Lamba T, Jaskiewicz W, et al. Effectiveness of a chatbot in improving the mental wellbeing of health workers in Malawi during the COVID-19 pandemic: a randomized, controlled trial. PLoS One 2024;19(5):e0303370. [doi: 10.1371/journal.pone.0303370] [Medline: 38805444]
55. Nahum-Shani I, Shaw SD, Carpenter SM, Murphy SA, Yoon C. Engagement in digital interventions. Am Psychol 2022 Oct;77(7):836-852. [doi: 10.1037/amp0000983] [Medline: 35298199]
56. Perski O, Short CE. Acceptability of digital health interventions: embracing the complexity. Transl Behav Med 2021 Jul 29;11(7):1473-1480. [doi: 10.1093/tbm/ibab048] [Medline: 33963864]
57. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
58. Linardon J. Rates of attrition and engagement in randomized controlled trials of mindfulness apps: systematic review and meta-analysis. Behav Res Ther 2023 Nov;170:104421. [doi: 10.1016/j.brat.2023.104421] [Medline: 37862854]
59. Nadal C, Sas C, Doherty G. Technology acceptance in mobile health: scoping review of definitions, models, and measurement. J Med Internet Res 2020 Jul 6;22(7):e17256. [doi: 10.2196/17256] [Medline: 32628122]
60. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
61. Li M, Gao Q, Yu T. Kappa statistic considerations in evaluating inter-rater reliability between two raters: which, when and context matters. BMC Cancer 2023;23(1):799. [doi: 10.1186/s12885-023-11325-z]
62. Higgins JPT, Altman DG, Gøtzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011 Oct 18;343:d5928. [doi: 10.1136/bmj.d5928] [Medline: 22008217]
63. Balshem H, Helfand M, Schünemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011 Apr;64(4):401-406. [doi: 10.1016/j.jclinepi.2010.07.015] [Medline: 21208779]
64. Higgins JPT, Thomas J, Chandler J, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 6.5 2024. URL: https://training.cochrane.org/handbook [accessed 2026-02-13]
65. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd ed: Routledge; 1988. [doi: 10.4324/9780203771587]
66. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14(1):25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
67. Akaike H. A new look at the statistical model identification. In: Parzen E, Tanabe K, Kitagawa G, editors. Selected Papers of Hirotugu Akaike: Springer; 1974. [doi: 10.1007/978-1-4612-1694-0_16]

XSL•FO
RenderX

68. Schwarz G. Estimating the dimension of a model. Ann Statist 1978;6(2):461-464 [FREE Full text] [doi: 10.1214/aos/1176344136]
69. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634. [doi: 10.1136/bmj.315.7109.629]
70. Sterne JAC, Sutton AJ, Ioannidis JPA, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011 Jul 22;343:d4002-d4002. [doi: 10.1136/bmj.d4002]
71. Barth J, Munder T, Gerger H, et al. Comparative efficacy of seven psychotherapeutic interventions for patients with depression: a network meta-analysis. PLOS Med 2013;10(5):e1001454. [doi: 10.1371/journal.pmed.1001454] [Medline: 23723742]
72. Tolin DF. Is cognitive-behavioral therapy more effective than other therapies? A meta-analytic review. Clin Psychol Rev 2010 Aug;30(6):710-720. [doi: 10.1016/j.cpr.2010.05.003] [Medline: 20547435]
73. Trent ES, Storch EA. Cognitive behavioral therapy for anxiety disorders. Psychiatr Clin North Am 2024 Dec;47(4):673-688. [doi: 10.1016/j.psc.2024.04.011] [Medline: 39505447]
74. Carl JR, Miller CB, Henry AL, et al. Efficacy of digital cognitive behavioral therapy for moderate-to-severe symptoms of generalized anxiety disorder: a randomized controlled trial. Depress Anxiety 2020 Dec;37(12):1168-1178. [doi: 10.1002/da.23079] [Medline: 32725848]
75. Hofmann SG, Smits JAJ. Cognitive-behavioral therapy for adult anxiety disorders: a meta-analysis of randomized placebo-controlled trials. J Clin Psychiatry 2008 Apr;69(4):621-632. [doi: 10.4088/jcp.v69n0415] [Medline: 18363421]
76. Kambeitz-Ilankovic L, Rzayeva U, Völkel L, et al. A systematic review of digital and face-to-face cognitive behavioral therapy for depression. NPJ Digit Med 2022 Sep 15;5(1):144. [doi: 10.1038/s41746-022-00677-8] [Medline: 36109583]
77. Karyotaki E, Efthimiou O, Miguel C, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021 Apr 1;78(4):361-371. [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
78. Păsărelu CR, Andersson G, Bergman Nordgren L, Dobrean A. Internet-delivered transdiagnostic and tailored cognitive behavioral therapy for anxiety and depression: a systematic review and meta-analysis of randomized controlled trials. Cogn Behav Ther 2017 Jan 2;46(1):1-28. [doi: 10.1080/16506073.2016.1231219]
79. Ahmed A, Hassan A, Aziz S, et al. Chatbot features for anxiety and depression: A scoping review. Health Informatics J 2023 Jan;29(1):14604582221146719. [doi: 10.1177/14604582221146719]
80. Cuijpers P, Miguel C, Harrer M, Ciharova M, Karyotaki E. The overestimation of the effect sizes of psychotherapies for depression in waitlist controlled trials: a meta-analytic comparison with usual care controlled trials. Epidemiol Psychiatr Sci 2024 Nov 6;33:e56. [doi: 10.1017/S2045796024000611] [Medline: 39501660]
81. Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ 2008 Mar 15;336(7644):601-605. [doi: 10.1136/bmj.39465.451748.AD] [Medline: 18316340]
82. Juul S, Gluud C, Simonsen S, Frandsen FW, Kirsch I, Jakobsen JC. Blinding in randomised clinical trials of psychological interventions: a retrospective study of published trial reports. BMJ Evid Based Med 2021 Jun;26(3):109. [doi: 10.1136/bmjebm-2020-111407] [Medline: 32998993]
83. Olawade DB, Wada OZ, Odetayo A, David-Olawade AC, Asaolu F, Eberhardt J. Enhancing mental health with artificial intelligence: current trends and future prospects. J Med Surg Public Health 2024 Aug;3:100099. [doi: 10.1016/j.glmedi.2024.100099]
84. Manole A, Cârciumaru R, Brînza R, Manole F. An exploratory investigation of chatbot applications in anxiety management: a focus on personalized interventions. Information 2025;16(1):11. [doi: 10.3390/info16010011]
85. Mak WWS, Tong ACY, Fu ACM, et al. Efficacy of internet-based rumination-focused cognitive behavioral therapy and mindfulness-based intervention with guided support in reducing risks of depression and anxiety: a randomized controlled trial. Appl Psychol Health Well Being 2024 May;16(2):696-722. [doi: 10.1111/aphw.12512] [Medline: 38073271]
86. Christ C, Schouten MJ, Blankers M, et al. Internet and computer-based cognitive behavioral therapy for anxiety and depression in adolescents and young adults: systematic review and meta-analysis. J Med Internet Res 2020;22(9):e17831. [doi: 10.2196/17831]
87. Reddy Y, Sudhir P, Manjula M, Arumugham S, Narayanaswamy J. Clinical practice guidelines for cognitive-behavioral therapies in anxiety disorders and obsessive-compulsive and related disorders. Indian J Psychiatry 2020;62(8):230. [doi: 10.4103/psychiatry.IndianJPsychiatry_773_19]
88. Nakagami Y, Uwatoko T, Shimamoto T, et al. Long-term effects of internet-based cognitive behavioral therapy on depression prevention among university students: randomized controlled factorial trial. JMIR Ment Health 2024;11:e56691-e56691. [doi: 10.2196/56691]
89. Barlow DH, Farchione TJ, Sauer-Zavala S. Unified Protocol for Transdiagnostic Treatment of Emotional Disorders: Therapist Guide: Oxford University Press; 2017. [doi: 10.1093/med-psych/9780190685973.001.0001]
90. Jagemann I, Stegemann M, von Brachel R, Hirschfeld G. Gender differences in preferences for mental health apps in the general population—a choice-based conjoint analysis from Germany. BMC Psychiatry 2024 Oct 14;24(1):682. [doi: 10.1186/s12888-024-06134-y] [Medline: 39402505]

XSL•FO
RenderX

91. Escoffery C. Gender similarities and differences for e-Health behaviors among U.S. adults. Telemed J E Health 2018 May;24(5):335-343. [doi: 10.1089/tmj.2017.0136] [Medline: 28813630]
92. Nwosu A, Boardman S, Husain MM, Doraiswamy PM. Digital therapeutics for mental health: is attrition the Achilles heel? Front Psychiatry 2022;13:900615. [doi: 10.3389/fpsyt.2022.900615] [Medline: 35982936]
93. Kelders SM, Kip H, Greeff J. Psychometric Evaluation of the TWente Engagement with Ehealth Technologies Scale (TWEETS): evaluation study. J Med Internet Res 2020 Oct 9;22(10):e17757. [doi: 10.2196/17757] [Medline: 33021487]
94. Forbes A, Keleher MR, Venditto M, DiBiasi F. Assessing patient adherence to and engagement with digital interventions for depression in clinical trials: systematic literature review. J Med Internet Res 2023 Aug 11;25:e43727. [doi: 10.2196/43727] [Medline: 37566447]
95. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. J Med Internet Res 2023 Feb 24;25:e40789. [doi: 10.2196/40789] [Medline: 36826990]
96. Milne-Ives M, de Cock C, Lim E, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res 2020 Oct 22;22(10):e20346. [doi: 10.2196/20346] [Medline: 33090118]
97. Borghouts J, Eikey E, Mark G, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387. [doi: 10.2196/24387] [Medline: 33759801]

Abbreviations

AI: artificial intelligence
AIC: Akaike information criterion
BIC: Bayesian information criterion
CBT: cognitive behavioral therapy
DMHI: digital mental health intervention
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
LLM: large language model
MeSH: Medical Subject Headings
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Search Extension
RCT: randomized controlled trial
ROB: risk of bias

Edited by S Brini; submitted 21.Aug.2025; peer-reviewed by D Shin, TB Xu; revised version received 10.Mar.2026; accepted 10.Mar.2026; published 08.May.2026.

Please cite as:
Gong B, Yao N, Xie H, Huang C, Kishimoto T, Berenbaum H, Mu W
Efficacy, User Engagement, and Acceptability of Cognitive Behavioral Therapy–Oriented Psychological Chatbots for Adults With Depressive and/or Anxiety Symptoms: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e82677
URL: https://www.jmir.org/2026/1/e82677
doi:10.2196/82677

© Bingyan Gong, Nisha Yao, Hangxin Xie, Chuncheng Huang, Tomoko Kishimoto, Howard Berenbaum, Wenting Mu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Barriers and Facilitators to Patient Acceptance of Artificial Intelligence in Health Care: Systematic Review

Huiqin Shi[1], BSc; Jingying Huang[1], MSc; Jin Yang[1], MSc; Mengbo Han[2], BSc

[1]Department of Nursing, Intensive Care Unit, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, No. 3 East Qingchun Road, Hangzhou, Zhejiang, China

[2]Department of Nursing, Sir Run Run Shaw Hospital, Hangzhou, China

Corresponding Author:
Jingying Huang, MSc
Department of Nursing, Intensive Care Unit, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, No. 3 East Qingchun Road, Hangzhou, Zhejiang, China

Abstract

Background: Artificial intelligence (AI) in the domain of health care is increasing in prominence. Acceptance is an indispensable prerequisite for the widespread implementation of AI.

Objective: This study aimed to explore barriers and facilitators influencing patients' acceptance of AI.

Methods: We conducted a systematic review following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. Nine databases, including PubMed, Web of Science, and Embase, were comprehensively searched from inception to December 23, 2025. We included qualitative, quantitative, and mixed methods studies investigating adult patients' attitudes toward medical AI. Two researchers independently screened records, extracted data, and appraised methodological quality using the Mixed Methods Appraisal Tool. Following the Joanna Briggs Institute convergent integrated approach, data synthesis was guided by integrating the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) and the Theoretical Domains Framework (TDF). Factors were mapped to behavior change techniques (BCTs) and evaluated for practical feasibility using the Affordability, Practicability, Effectiveness and cost-effectiveness, Acceptability, Side-effects/safety, and Equity criteria.

Results: A total of 61 studies met the inclusion criteria out of 7452 search results. Study designs included qualitative (n=20), quantitative (n=35), and mixed methods (n=6). Performance and effort expectancies were the primary determinants of acceptance. Major barriers comprised perceived operational complexity, lack of algorithmic trust, reduced interpersonal interaction, privacy vulnerabilities, and high costs. Facilitators included transparent data governance, interpretability of AI decisions, improved clinician-patient communication, and human-centered design. Education level and disease severity emerged as key moderating variables. Through UTAUT2-TDF mapping, we identified 25 distinct BCTs (6 high, 14 medium, and 5 low feasibility) and formulated 40 actionable intervention strategies.

Conclusions: This study innovatively integrates the UTAUT2 and TDF frameworks to evaluate patient acceptance of medical AI. Unlike existing reviews that predominantly evaluate isolated psychosocial factors or purely technical attributes, this transtheoretical approach differentiates itself by merging technology adoption mechanisms directly with behavioral drivers. Consequently, it contributes to the field by systematically identifying multilevel factors influencing acceptance, including performance expectancy, effort expectancy, and ethical security, and translating these into 40 actionable BCTs. In real-world clinical practice, these findings provide a feasible, prioritized blueprint for clinicians and administrators to design patient-centered interventions, enhancing the clinical integration and long-term effectiveness of medical AI.

Trial Registration: PROSPERO CRD42024598884; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024598884

(J Med Internet Res 2026;28:e80581) doi:10.2196/80581

KEYWORDS

artificial intelligence; patients; barriers; facilitators; interventions

Introduction

Artificial intelligence (AI) is a transformative technology that integrates machine learning and deep learning, demonstrating vast potential across various medical and scientific fields [1]. AI in medicine relies on technologies such as machine learning and natural language processing to transform medical imaging analysis, clinical decision support, and personalized treatment planning, thereby enhancing diagnostic accuracy, efficiency, and patient outcomes [2,3]. By enhancing diagnostic efficiency and accuracy through intelligent medical image analysis, early disease risk prediction, and personalized treatment planning, AI has significantly improved patient outcomes and health care

XSL•FO
RenderX

experiences [1]. AI has been integrated into the health care sector to offer effective, clinically relevant solutions for patients and medical professionals. An expanding body of literature has demonstrated that AI-based interventions can equal or even surpass physician expertise [4].

Nevertheless, its adoption on a broader scale remains limited. Ethical concerns, standardization gaps, and ambiguous legal responsibilities are among the challenges hindering the widespread implementation of AI in health care today. According to Rogers' Diffusion of Innovations theory, user acceptance is pivotal in determining the adoption rate of new technologies [5]. In technological contexts, acceptance is characterized as the readiness, purpose, and intrinsic motivation to use a technology, stemming from positive attitudes toward the system [6]. The acceptance of AI systems follows a similar dynamic to other new tools, but the unpredictable management of complex scenarios and expectations for human-like interactions often amplify resistance. Accepting change is a complex process, as humans are often averse to altering the familiar status quo in favor of comfort. Yet, for enhancing patient outcomes and experiences throughout the health care journey in the long term, acceptance remains a crucial factor in adopting and integrating newly introduced innovations such as AI into daily practice [7]. Understanding patient acceptance is crucial for improving AI adoption, informing health care policy, and building a more robust and equitable health care system [8].

With the advancement of medical AI, its implementation in clinical practice has evoked complex attitudes and feelings among patients [9,10]. Therefore, many scholars have investigated the willingness to adopt such technology among different patient groups, aiming to identify relevant barriers and facilitating factors. Alsanosi et al [11] used questionnaire surveys to explore the awareness, attitudes, and behaviors toward AI among specific patient populations, finding that education level, age, and self-efficacy positively influenced their willingness to use it. Gundlack et al [12] conducted in-depth interviews through a semistructured approach with a subset of patients, revealing that patients were most concerned with whether AI could reduce medical costs, shorten waiting times, and accelerate the diagnostic and treatment process. Ensuring transparency in the development, data processing, and intended use of AI systems was also considered crucial [13]. Research by Longoni and Morewedge [10] highlights that, despite clear evidence of AI's cost-effectiveness, accuracy, and operational efficiency, many patients remain hesitant about AI systems that operate autonomously or make critical medical decisions without human involvement.

Despite the valuable insights offered by previous studies, there remains a lack of systematic integration and supplementation grounded in a unified theoretical framework. This gap hinders the effective translation of research findings into practical clinical interventions. Given the complexity and multidimensionality of patient perceptions, applying a well-established theoretical model to synthesize existing evidence on the factors influencing AI adoption can not only elucidate the underlying mechanisms but also enhance the generalizability and applicability of research outcomes across diverse clinical contexts. Various theoretical models have been developed to measure the acceptance of technological innovations, such as the technology acceptance model and the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) [14]. These approaches have been widely adopted in prior research, including in the health care field, such as exploring patient adoption of mobile health apps [15], telemedicine technologies [16], and wearable devices [17], demonstrating strong predictive power in explaining patients' behavioral intentions and actual usage behaviors. UTAUT2 integrates 8 technology acceptance-related theories, including the technology acceptance model, by synthesizing key factors influencing behavioral intention and usage behavior: performance expectancy, effort expectancy, social influence, and facilitating conditions [15,18]. It also incorporates moderating variables such as gender, age, voluntariness, and experience.

Therefore, this study uses the UTAUT2 framework to examine the direct determinants of patients' acceptance of medical AI technologies from a technology adoption perspective. However, while this model effectively identifies influencing factors with an emphasis on technological attributes and user psychology, it offers limited explanatory power regarding specific behavior change mechanisms [19]. Furthermore, its theoretical dimensions remain relatively abstract, posing challenges for deriving targeted intervention strategies. To address these limitations, we integrate the Theoretical Domains Framework (TDF), a well-established framework in behavioral science, to supplement and refine the dimensions of UTAUT2 [20]. By leveraging the mapping relationships between theoretical domains and behavior change techniques (BCTs) [21], this integrative approach bridges technology acceptance theory and behavioral science, enabling the identification of tailored BCTs to guide clinical practice and enhance patient acceptance of AI. Building on this dual-theoretical foundation, the primary aim of this systematic review is to synthesize evidence through the combined lenses of technology acceptance and behavioral science to comprehensively identify the barriers and facilitators to patient acceptance of AI in hospital settings.

Methods

Overview

This review was reported following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement [22]. The protocol was registered with PROSPERO in October 2024 (registration no CRD42024598884). A theoretically grounded mixed methods systematic review provides a structured and integrative approach to identifying the barriers and facilitators that influence patient acceptance of AI.

Search Strategy

The literature search was conducted following the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Searching) reporting guidelines [23]. Two researchers (HS and JY) and an evidence-based expert (JH) developed the search strategy using the Population, Interest, and Context framework [24]. Eligibility criteria focused on

XSL•FO
RenderX

patients, AI artifacts, and attitudes and perceptions. Barriers and facilitators were defined as factors preventing or motivating AI use. The search strategy for this study was built around the core terms "artificial intelligence," "patient," and "attitude," combining the use of medical subject headings and keywords. A comprehensive search was conducted across PubMed, Web of Science, Embase, Cochrane Library, CINAHL, CNKI, Wan Fang, VIP, and SinoMed (inception to December 23, 2025). Detailed search strategies are provided in Table S1 in Multimedia Appendix 1. Our comprehensive search strategy was executed across multiple sources to ensure literature saturation. This included searching major study registries, systematic online database searches, manual browsing of key journals and organizational websites, and citation searching (both backward and forward) of all included studies and relevant reviews. No restrictions were applied based on publication status or language during the initial search phase.

Inclusion and Exclusion Criteria

We included qualitative, quantitative, and mixed methods studies that addressed (1) patients' perceptions of using AI in health care (diagnosis and treatment) and (2) barriers and facilitators to patients' willingness to use AI. Studies limited to health care professionals, managers, technology developers, and medical and nursing students, as well as studies that only tested the effectiveness of AI, were excluded. Table 1 shows the inclusion and exclusion criteria for this study.

Table . Inclusion and exclusion criteria for the mixed methods systematic review on patient acceptance of medical artificial intelligence.

	Population	Phenomena of interest	Context	Study type	Limits
Inclusion criteria	• Patients (regardless of disease type) • Patients aged ≥18 years	• Patient perceptions of the use of AI[a] (diagnostic and therapeutic). • Barriers and facilitators to patients' willingness to use AI	• All medical fields	• Qualitative studies • Quantitative studies • Mixed methods studies	• Literature in English or Chinese
Exclusion criteria	• Health care professionals, administrators, or technology developers only • Students of medicine and nursing	• Studies investigating the effectiveness of AI technology in health care only	• Nonmedical scenarios such as sports and fitness	• Gray literature • Conference abstracts • Unpublished studies and dissertations	—[b]

[a]AI: artificial intelligence.

[b]Not available.

Literature Screening and Data Extraction

Two researchers (HS and MH) independently screened papers after removing duplicates in EndNote (version 21; Clarivate), resolving disagreements through discussion or third-party consultation (JH). Data on barriers, facilitators, and interventions were extracted using a data extraction table [24]. Qualitative studies provided themes and participant quotes, while quantitative studies included outcome measures and narrative summaries. Mixed methods studies were analyzed separately. Extracted details included author, year, design, participant characteristics, findings, and interventions, with open-ended survey responses treated as qualitative data, ensuring comprehensive extraction and synthesis. When necessary, corresponding authors were contacted to obtain missing or supplementary data.

Quality Assessment

The Mixed Methods Appraisal Tool (MMAT) [25] was used to assess the methodological quality of the included studies, covering qualitative, quantitative (randomized/nonrandomized), and mixed methods designs. Two researchers (HS and MH) independently conducted assessments, resolving disagreements through consensus or consultation with a third expert (JH).

Synthesis Methods

Overview

This study aims to systematically identify the barriers to and facilitators of patient adoption of AI in health care settings. Given that the research question is well-suited to be addressed through complementary qualitative and quantitative evidence, we adopted the "convergent integrated approach" from the Joanna Briggs Institute methodology for data synthesis [26]. This method involves transforming quantitative data into a qualitative format and integrating it with original qualitative findings, which are subsequently categorized and integrated based on similarity in data meaning. This procedure was initially performed by one researcher and then discussed and cross-checked with a senior researcher experienced in evidence-based practice.

We used narrative synthesis rather than meta-analysis due to substantial heterogeneity in the included studies regarding design, measurement, and AI application contexts [27]. This heterogeneity precluded a statistically meaningful quantitative

XSL•FO
RenderX

pooling of effect sizes. Consequently, certain items on the Synthesis Without Meta-analysis checklist pertaining strictly to meta-analytical data transformations were not applicable to this study. Our primary objective was to explore the complex mechanisms influencing acceptance, an explanatory aim better served by narrative synthesis that integrates qualitative and quantitative evidence, rather than by quantifying isolated effect sizes.

Integrated Analytical Framework

During the thematic synthesis stage, the identified obstacles and facilitators were summarized by theme and mapped into the various domains of the UTAUT2 [28], which is a reliable framework for understanding technology acceptance [29-32]. To deepen the behavioral analysis, these factors based on UTAUT2 were further analyzed through the TDF [33]. The TDF is based on behavioral science and provides a structured perspective with 14 domains, systematically describing the determinants of behavior (refer to Table S2 in Multimedia Appendix 1 for its structure). This step enables a more precise and context-specific understanding of the obstacles and facilitators. Finally, to convert these insights into actionable strategies, the identified TDF domains were linked to specific BCTs through the established mapping method [21,34], leveraging the hierarchical BCT taxonomy (version 1; refer to Table S3 in Multimedia Appendix 1).

Priority Assessment of BCTs

To guide practical implementation, we evaluated the feasibility of the identified BCTs using the Affordability, Practicability, Effectiveness and Cost-Effectiveness, Acceptability, Side-effects or Safety, and Equity criteria [35]. This process involved independent ratings by 2 researchers on 6 dimensions (affordability, practicality, effectiveness, acceptability, side effects, and equity), based on empirical evidence and contextual knowledge. Discrepancies were resolved through discussion with a third expert. Each BCT was subsequently classified as having high, medium, or low feasibility. The specific definitions and classification rules are detailed in Tables S4 and S5 in Multimedia Appendix 1. This feasibility rating was combined with the target level of each BCT (individual, provider, or system).

Results

Literature Search

A comprehensive search of 9 electronic databases using a refined search strategy identified 7452 records. After removing duplicates, 6680 records remained. Following an initial screening of titles and abstracts, 128 records were deemed relevant. Subsequently, full-text screening was performed according to the inclusion and exclusion criteria, excluding irrelevant papers. Ultimately, 61 studies were included in the final review. All literature was used to extract barriers and facilitators, with 53 studies used to extract potential interventions. The literature screening process is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Flow diagram for systematic review.

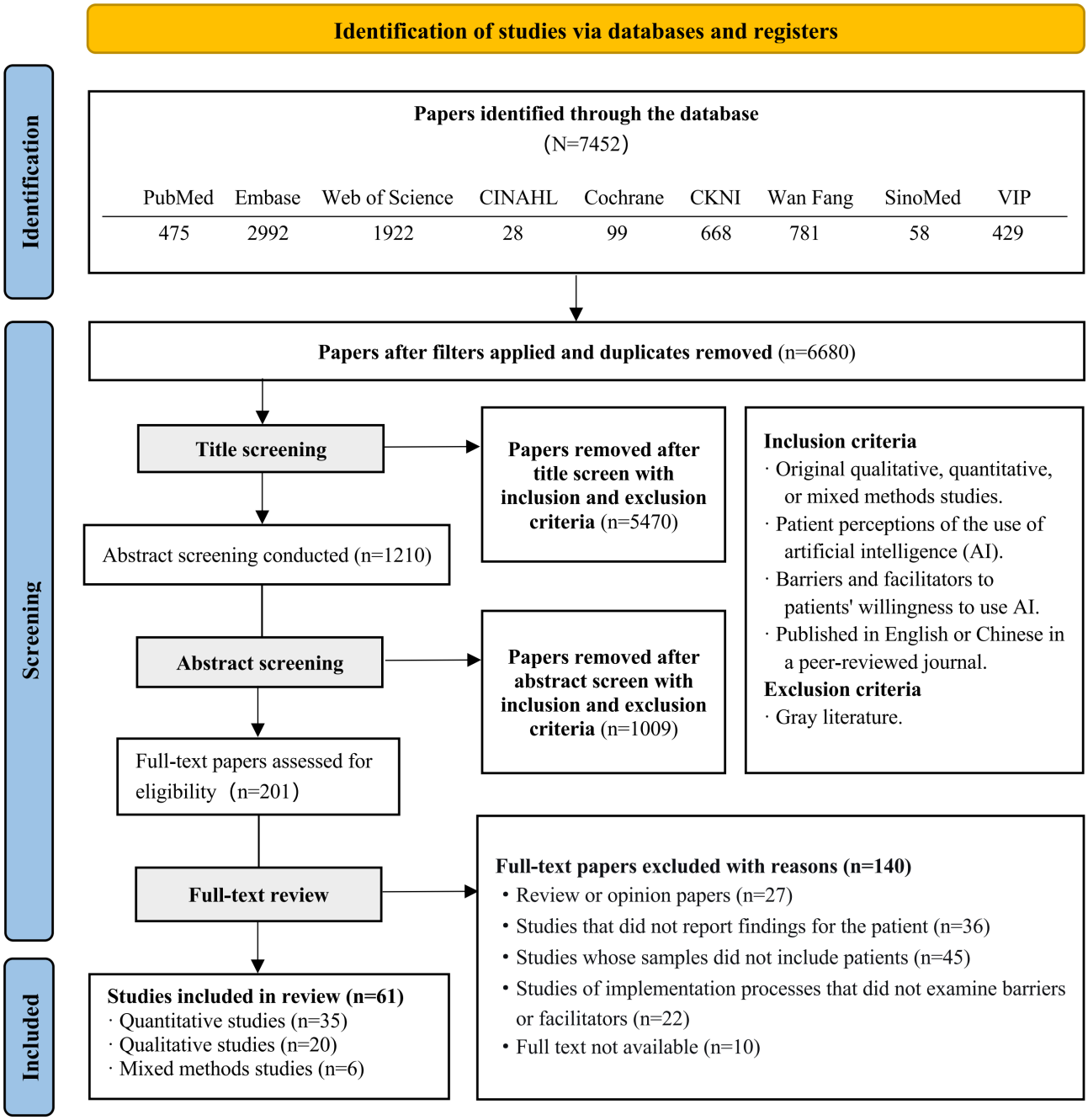

Study and Patient-Reported Experience Measure Description

This study included 61 studies categorized by research design: 20 qualitative (primarily semistructured interviews), 2 randomized controlled trials, 6 mixed methods, 6 quantitative descriptive, and 27 nonrandomized studies. Quantitative studies mainly used questionnaires, with one adapting the UTAUT2 model [36]. Sample sizes ranged from 9 to 2899 participants. Geographically, studies spanned 19 countries, with 45 from 15 developed countries (eg, 15 from the United States) and 16 from 4 developing countries (eg, 9 from China). Detailed characteristics of the included studies and participants are summarized in Table 2.

XSL•FO
RenderX

Table . Features of the studies and participants included, stratified by study design.

Authors, country, year	Study type (quality score)	Object of study	Recruitment date	Setting	Aims
Haan et al, Netherlands, 2019 [37]	Qualitative (5)	• n=20 (male=11) • Age: no restriction • Health status: no restriction	2018/07 - 2018/08	Radiology of the University Medical Center Groningen	To develop an understanding of patients' level of knowledge of AI[a] and to explore the meanings patients ascribe to key topics.
Jalil et al, Australia, 2019 [38]	Qualitative (5)	• n=9 (male=5) • Age: no restriction • Health status: T2D[c] for at least 12 months	None	None	To investigate the UX[b] of patients with T2D using a telehealth in-home monitoring device to manage T2D from home.
Adams et al, Canada, 2020 [39]	Qualitative (5)	• n=17 (male=6) • Age: no restriction • Health status: no restriction	None	None	To explore patients' initial perceptions of AI in the context of radiology and how AI tools can be developed and evaluated in ways that are consistent with patients' values.
Nelson et al, United States, 2020 [40]	Qualitative (5)	• n=48 (male=22) • Age: ≥18 years • Health status: history of melanoma, history of non-melanoma skin cancer only, and no history of skin cancer	2019/05 - 2019/08	Brigham and Women's Hospital and the melanoma clinics at the Dana-Farber Cancer Institute	To explore how patients conceptualize AI and perceive the use of AI for skin cancer screening.
Benrimoh et al, Canada, 2021 [41]	Qualitative (5)	• n=20 (male=13) • Age: no restriction • Health status: no restriction	None	Steinberg Center for Simulation and Interactive Learning	To explore the use of a simulation center environment in evaluating the usability of Aifred.
Musbahin et al, United Kingdom, 2021 [42]	Qualitative (5)	• n=28 (male=11) • Age: ≥18 years • Health status: no restriction	None	University patient involvement mailing list	To address this knowledge gap by applying the nominal group technique to explore patient public views on AI.
Zhang et al, United States, 2021 [43]	Qualitative (5)	• n=13 (male=9) • Age: no restriction • Health status: no restriction	2019/08 - 2019/09	University Institutional Review Board	To understand patients' perceptions and acceptance of using AI technology to interpret their radiology reports.
Mikkelsen et al, Denmark, 2023 [44]	Qualitative (5)	• n=10 (male=6) • Age: ≥18 years • Health status: no restriction	2019/10 - 2022/01	Denmark health care system	To uncover patient perspectives on trust regarding the patient GP[d] relationship.

XSL•FO
RenderX

Authors, country, year	Study type (quality score)	Object of study	Recruitment date	Setting	Aims
Pelayo et al, United States, 2023 [45]	Qualitative (5)	• n=20 (male=8) • Age: ≥18 years • Health status: diagnosed with type 1 or type 2 diabetes	2022/07 - 2022/11	Federally Qualified Health Center	To understand the perspectives of Latin patients with diabetes on ophthalmology, AI-based image interpretation, and general virtual care to prevent avoidable blindness in this population.
Pelly et al, Australia, 2023 [46]	Qualitative (5)	• n=38 (male=14) • Age: ≥18 years • Health status: people with a history of myocardial infarction	2020/11 - 2021/12	Cardiac Health Center	To explore the opinions of PHMI[e] and health professionals on the use of AI for secondary prevention of MI[f].
Robertson et al, United States, 2023 [47]	Qualitative (5)	• n=41 (male=20) • Age: ≥18 years • Health status: patients with type 1 or type 2 diabetes	None	None	To support and ensure safe use of AI or ML[g] technologies in health care.
Neves et al, China, 2023 [48]	Qualitative (5)	• n=12 (male=15) • Age: ≥18 years • Health status: patients with stroke confirmed by the Chinese guidelines for the prevention	2021/06 - 2022/03	Neurology department of a general hospital	To investigate the experience of robot-assisted gait training experience in patients with stroke.
Godoy Junior et al, Italy, 2024 [49]	Qualitative (5)	• n=27 (male=16) • Age: no restriction • Health status: adult persons with Parkinson disease with a minimum of 2 years with PD[i] diagnosis	2021/09 - 2022/08	None	To explore patient and neurologist perspectives on AI-assisted RMS[h].
Maris et al, Germany, 2024 [50]	Qualitative (5)	• n=24 (male=10) • Age: no restriction • Health status: those who had either an ICD[j] and/or a heart condition with increased risk of SCD[k]	None	Dutch ICD Patient Association	To explore perspectives of patients on the ethical use of AI, particularly in clinical decision-making regarding the implantation of an ICD.
Sachdeva et al, Cameroon, 2024 [51]	Qualitative (5)	• n=32 (male=8) • Age: 30 - 49 years • Health status: who had previously participated in the CC[l] screening program	2022/08	DSchang Regional Annex Hospital	To explore the acceptability and perspectives of women in DSchang regarding the usage of a screening tool for CC relying on AI.
Tursynbek et al, Philippines, 2024 [52]	Qualitative (5)		2024/02 - 2024/03	Westhead Hospital	

XSL•FO
RenderX

Authors, country, year	Study type (quality score)	Object of study	Recruitment date	Setting	Aims
		• n=30 (male=10) • Age: no restriction • Health status: no restriction			To describe patients' perceptions of medical AI and its integration into the health care system in Kazakhstan.
Gundlack et al, Germany, 2025 a [53]	Qualitative (5)	• n=35 (male=22) • Age: no restriction • Health status: no restriction	2022/06 - 2023/03	A district hospital for psychological illnesses	To explore how patients perceive AI in medical care, focusing on relationships with physicians and ethical aspects.
Gundlack et al, Germany, 2025 b [12]	Qualitative (5)	• n=35 (male=22) • Age: no restriction • Health status: no restriction	2022/06 - 2023/03	Martin Luther University Halle-Wittenberg	To explore perceptions toward acceptance, challenges of implementation, and potential applications of AI in medical care.
Krisnan et al, Malaysia, 2025 [13]	Qualitative (5)	• n=17 (male=14) • Age: no restriction • Health status: no restriction	2023/08	Hospital Tengku Ampuan Rahimah	To examine patient perceptions and attitudes toward the use of AI in diabetes care.
Trivedi et al, Australia, 2025 [54]	Qualitative (5)	• n=30 (male=15) • Age: no restriction • Health status: no restriction	2021/08 - 2022/05	Westhead Hospital	To explore patient perspectives on receiving support from a conversational AI support program.
Sebastian et al, United States, 2023 [55]	RCT (3)	• n=150 (male=68) • Age: no restriction • Health status: no restriction	None	None	To examine whether communication strategies are more successful in overcoming factors that hinder AI product adoption among patients.
Zhou et al, China, 2022 [56]	RCT (5)	• n=776 (male=39) • Age: no restriction • Health status: no restriction	None	None	To explore the influence of the degree of AI involvement on patients' WTA[m] and to reveal whether the reason is caused by patients' perceived threat to outside groups.
Yang et al, China, 2019 [57]	Cross-sectional (4)	• n=527 (male=199) • Age: ≥18 years • Health status: Chinese patients with cancer, who were informed of their cancer diagnosis	None	Cancer Center of Sichuan University	To assess the attitudes of Chinese patients with cancer toward the clinical use of AIM[n], and to analyze the possible influencing factors.
Jutzi et al, Germany, 2020 [58]	Cross-sectional (4)	• n=298 (male=80) • Age: no restriction • Health status: no restriction	2019/11 - 2020/01	Dermatological University Hospital	To evaluate the patients' view of AI in melanoma diagnostics in Germany, with a particular focus on patients with a history of melanoma.

XSL•FO
RenderX

Authors, country, year	Study type (quality score)	Object of study	Recruitment date	Setting	Aims
Meyer et al, United States, 2020 [59]	Cross-sectional (3)	• n=329 (male=74) • Age: no restriction • Health status: no restriction	2018/03 - 2018/05	The Isabel Symptom Checker	To examine patients' experiences using an AI-assisted online symptom checker.
Ongena et al, Netherlands, 2020 [60]	Cross-sectional (5)	• n=155 (male=86) • Age: ≥18 years • Health status: no restriction	2018/11 - 2019/03	University Medical Center Groningen	To develop and validate a standardized patient questionnaire on the implementation of AI in radiology.
Aggarwal et al, United Kingdom, 2021 [61]	Cross-sectional (5)	• n=408 (male=173) • Age: ≥16 years • Health status: no restriction	2018/06 - 2018/09	University Teaching Hospital	To identify current awareness regarding health data research, and to obtain their opinions and views on data sharing for AI research purposes.
Esmaeilzadeh et al, United States, 2021 [62]	Cross-sectional (5)	• n=634 (male=355) • Age: no restriction • Health status: no restriction	2022/05	Institutional Review Board of Florida International University	To examine how potential users (patients) perceive the benefits, risks, and use of AI clinical applications for their health care purposes.
Liu et al, China, 2021 [63]	Cross-sectional (2)	• n=428 (male=206) • Age: 18 - 85 years • Health status: no restriction	None	First Affiliated Hospital of Jinan University	To visualize and measure patients' heterogeneous preferences from various angles of AI diagnosis.
Lennartz et al, Germany, 2021 [64]	Cross-sectional (3)	• n=229 (male=99) • Age: no restriction • Health status: no restriction	2020/01 - 2020/05	Institute for Diagnostic and Intervention Radiology	To investigate patients' opinions on the use of AI in different aspects of the medical workflow.
Armero et al, United States, 2022 [65]	Cross-sectional (4)	• n=349 (male=176) • Age: ≥18 years • Health status: enrolled patients presenting to the labor and delivery unit at a tertiary care	2019/11 - 2020/06	Labor and Delivery Unit of Brigham and Women's Hospital	To evaluate and understand pregnant patients' perspectives on the implementation of AI in clinical care.
Yulan et al, China, 2022 [66]	Cross-sectional (5)	• n=446 (male=244) • Age: no restriction • Health status: not target-specific disease populations	2020/02 - 2020/04	West China Hospital of Sichuan University	To understand older patients' willingness to use AI robots and the influencing factors.
Khullar et al, United States, 2022 [67]	Cross-sectional (4)	• n=926 (male=455) • Age: no restriction • Health status: no restriction	2019/12	None	To understand public perceptions of the use of AI in diagnosis and treatment.

XSL•FO
RenderX

Authors, country, year	Study type (quality score)	Object of study	Recruitment date	Setting	Aims
Kosan et al, Germany, 2022 [68]	Cross-sectional (5)	• n=140 (male=69) • Age: ≥18 years • Health status: no restriction	2021/02 - 2021/03	Charité–Universitätsmedizin Berlin Dental Clinic	To assess patients' perspectives on AI in dentistry scientifically for radiography caries detection, and the impact of AI-based diagnosis on patients' trust.
Ayad et al, Germany, 2023 [9]	Cross-sectional (4)	• n=330 (male=165) • Age: ≥18 years • Health status: no restriction	2021/12 - 2022/03	University Hospital Munster	To assess patients' knowledge and perceptions of AI in dentistry.
Kawsar et al, United Kingdom, 2023 [69]	Cross-sectional (3)	• n=268 (male=114) • Age: ≥18 years • Health status: at least one skin lesion that could be photographed	2020/02 - 2021/08	Chelsea and Westminster Hospital	To explore patients' perspectives on the use of AI as part of their skin cancer management pathway.
Mahlknecht et al, Italy, 2023 [70]	Cross-sectional (4)	• n=116 (male=52) • Age: no restriction • Health status: no restriction	2021/09 - 2021/11	The Institute of General Practice and Public Health in Bolzano	To evaluate both patients' and physicians' attitudes toward these tools in Italian general practice settings.
Parry et al, United States, 2023 [71]	Cross-sectional (4)	• n=397 (male=141) • Age: ≥18 years • Health status: no restriction	2020/03 - 2021/08	A large urban academic center and a rural health system.	To investigate the patient opinion on the use of AI in orthopedics.
Temple et al, United Kingdom, 2023 [72]	Cross-sectional (5)	• n=95 (male=68) • Age: no restriction • Health status: no restriction	2021/11 - 2022/03	A specialist UK cancer center.	To adapt and use a validated patient questionnaire to develop an understanding of current patient views on the use of AI in radiotherapy.
Baghdadi et al, Saudi Arabia, 2024 [73]	Cross-sectional (5)	• n=382 (male=109) • Age: ≥18 years • Health status: no restriction	2022/07 - 2022/12	Tertiary care hospital	To examine patients' attitudes toward the use of AI as a tool in diagnostic radiology.
Bahadir et al, Turkey, 2024 [74]	Cross-sectional (5)	• n=272 (male=107) • Age: ≥18 years • Health status: advanced occlusal caries	2023/04 - 2024/01	The clinic of restorative dentistry	To evaluate patients' attitudes toward the use of AI in dental radiographic detection.
Esin et al, Turkey, 2024 [75]	Cross-sectional (4)	• n=725 (male=319) • Age: ≥18 years • Health status: undergone surgery	2021/05 - 2021/06	Tepee Training and Research Hospital	To evaluate the opinions of individuals regarding the use of AI and robots in the field of health care.
Fransen et al, Western European, 2024 [76]	Cross-sectional (5)		2023/01 - 2023/10	At 3 Western European medical institutes	

XSL•FO
RenderX

Authors, country, year	Study type (quality score)	Object of study	Recruitment date	Setting	Aims
		• n=212 • Age: no restriction • Health status: patients suspicious with prostate cancer undergoing prostate MRI[o]			To investigate patients' acceptance of AI for diagnosing prostate cancer on MRI scans and the factors that influenced their trust in AI diagnoses.
Huang et al, Singapore, 2024 [36]	Cross-sectional (5)	• n=100 (male=56) • Age: no restriction • Health status: no restriction	2021/09 - 2022/06	National Heart Center	To better understand factors influencing patient usage intention.
Riedl et al, Germany, 2024 [77]	Cross-sectional (5)	• n=1183 (male=535) • Age: ≥18 years • Health status: no restriction	2022	A market research company	To investigate patients' preferences for interacting with a human doctor or an AI system.
Alsanosi et al, Saudi Arabia, 2025 [11]	Cross-sectional (5)	• n=385 (male=154) • Age: no restriction • Health status: no restriction	2024/07 - 2024/12	Umm Al-Qura University	To explore knowledge, attitudes, and perceptions (KAP) regarding AI in medication adherence among chronic patients in the Makkah region.
Chan et al, Australia, 2025 [78]	Cross-sectional (4)	• n=385 (male=211) • Age: no restriction • Health status: no restriction	2024/06 - 2024/10	The Faculty of Medicine and Health	To provide a baseline understanding of patient views about the use of AI/ML in the specific context of radiotherapy to contribute toward future governance of the technology.
Ozcan et al, Texas, 2025 [79]	Cross-sectional (5)	• n=518 (male=259) • Age: no restriction • Health status: no restriction	2023/02 - 2023/08	Texas Southwestern Medical Center	To identify the independent factors associated with participants' acceptance of AI use.
Tirapelli et al, United States, 2025 [80]	Cross-sectional (5)	• n=2581 (male=1806) • Age: no restriction • Health status: no restriction	2023/01 - 2024/06	Aarhus University	To evaluate patients' perceptions of the use of AI in dental imaging diagnostics across 6 centers worldwide.
Xiangde, China, 2019 [81]	Quantitative descriptive (1)	• n=446 (male=236) • Age: no restriction • Health status: conscious outpatient and inpatient patients	2018/05 - 2018/06	The Third Hospital of Chongqing	To understand patients' cognition and trust in AI health care, and analyze the reasons.
Menghan and Weihua, China, 2019 [82]	Quantitative descriptive (4)		None	Different hospitals in Guangdong province	

XSL•FO
RenderX

Authors, country, year	Study type (quality score)	Object of study	Recruitment date	Setting	Aims
		• n=550 (male=271) • Age: no restriction • Health status: no restriction			To study the influence mechanism of the use of hospital intelligent medical systems on patients' satisfaction with medical treatment.
Siwen and Jiudi, China, 2022 [83]	Quantitative descriptive (3)	• n=200 (male=117) • Age: ≥18 years • Health status: esophageal cancer staging ranges from stage I to stage IV	2019/07 - 2020/12	Sun Yat Sen University Cancer Center	To investigate the acceptance of the P2 robot by patients, families, and medical staff for health education in patients with esophageal cancer surgery, propose targeted improvement measures, and provide a basis for the future widespread use of robots in health education.
Jin et al, China, 2023 [84]	Quantitative descriptive (5)	• n=39 (male=20) • Age: no restriction • Health status: meets the diagnostic criteria for dysphagia	2022/03 - 2022/05	Department of Neurology, Beijing Chaoyang Hospital	To design and develop a swallowing function training system suitable for patients with stroke with swallowing disorders.
Haggenmüller et al, United States, 2024 [85]	Quantitative descriptive (5)	• n=178 (male=89) • Age: no restriction • Health status: no restriction	None	American Academy of Dermatology	To investigate the criteria required for patients and dermatologists to accept AI-systems.
Macri et al, Australia, 2024 [86]	Quantitative descriptive (4)	• n=15 (male=3) • Age: ≥18 years • Health status: vitreoretinal surgery, in which postoperative face-down positioning is used	2022/12 - 2023/09	The Royal Adelaide Hospital	To examine the acceptability of an AI-generated presenter in a patient information video about face-down positioning after theoretical surgery.
Palmisciano et al, United Kingdom, 2020 [87]	Mixed methods (5)	• n=20+107 (male=45) • Age: no restriction • Health status: no restriction	Qualitative: 2019/09 and quantitative: 2019/10	Department of Neurosurgery	To evaluate attitudes of patients and their relatives regarding the use of AI in neurosurgery.
van der Zander et al, Netherlands, 2022 [88]	Mixed methods (3)	• n=377 (male=155) • Age: ≥18 years • Health status: who underwent an endoscopic procedure	2020/04 - 2021/08	Catharina Hospital Eindhoven	To investigate the perspectives on AI in health care among patients with GI[p] disorders.
Katirai et al, Japan, 2023 [89]	Mixed methods (5)	• n=11 (male=5) • Age: no restriction • Health status: no restriction	2021/02 - 2021/04	The Osaka University Graduate School of Medicine	To provide a snapshot of patient and public perceptions of AI in health care in the Japanese context.

XSL•FO
RenderX

Authors, country, year	Study type (quality score)	Object of study	Recruitment date	Setting	Aims
Robertson et al, United States, 2023 [47]	Mixed methods (5)	• n=24+2875 (male=1092) • Age: no restriction • Health status: no restriction	Qualitative: 2020/02-2020/12 and quantitative: 2021/01 - 2021/02	Clinics in Tucson, Arizona	To determine how diverse patient populations feel about the use of AI diagnostic tools.
Gonzalez et al, United States, 2024 [90]	Mixed methods (5)	• n=281 (male=140) • Age: ≥18 years • Health status: no restriction	Phase I: 2023/06 and Phase II: 2023/07 - 2023/10	Academic medical center in St. Louis	To assess public and surgical patient attitudes and perspectives on ML-CDSS[q] use in perioperative care.
Witkowski et al, United States, 2024 [91]	Mixed methods (4)	• n=600 (male=292) • Age: ≥18 years • Health status: no restriction	2023/08	A leading market research provider	To explore the extent to which patients are confident in and comfortable with the use of these technologies when it comes to their own individual care, and identify areas.

[a]AI: artificial intelligence.
[b]UX: user experience.
[c]T2D: type 2 diabetes.
[d]GP: general practitioner.
[e]PHMI: people with a history of myocardial infarction.
[f]MI: myocardial infarction.
[g]ML: machine learning.
[h]RMS: remote monitoring system.
[i]PD: Parkinson disease.
[j]ICD: implantable cardioverter-defibrillator.
[k]SCD: sudden cardiac death.
[l]CC: cervical cancer.
[m]WTA: willingness to accept.
[n]AIM: artificial intelligence in medicine.
[o]MRI: magnetic resonance imaging.
[p]GI: gastrointestinal.
[q]ML-CDSS: machine learning–clinical decision support system.

The study population includes patients using medical AI aged ≥18 years, recruited from communities, hospitals, and online surveys. Patient types varied, with a focus on patients with cancer (n=11), radiology (n=8), perioperative (n=4), diabetes (n=3), dental (n=5), and stroke (n=2). Nine studies included patient and health care professional perspectives; only patient data were analyzed to align with research objectives. The comprehensive data and full extraction details are provided in Table S6 in Multimedia Appendix 1 due to space constraints.

Quality Assessment

According to the MMAT evaluation, the average score of the included studies was 4.4. Approximately 85% (n=53) of the studies met all or 80% of the quality assessment criteria. About 11% (n=6) of the studies met 60% of the quality assessment criteria. Two studies met only 40% or less of the quality assessment criteria. In general, the methodological quality of the included studies was assessed as relatively high. The MMAT User Manual mentions the treatment of low-quality studies, which are rounded off according to the topic of the study; in this study, these 2 papers were valuable in contributing to the research topic and were therefore not excluded [25]. The MMAT scores for each included study are provided in Table 2, and detailed item-level assessments are provided in Table S7 in Multimedia Appendix 1.

Identified Barriers and Facilitators

Overview

To ensure transparency in the analytic process, we followed structured steps. First, 2 researchers independently conducted open coding and reached a consensus through discussion, forming an initial codebook. These codes were then categorized into the 7 core dimensions of UTAUT2 (eg, “concern about effectiveness” was classified under “performance expectancy”). Subsequently, codes within each UTAUT2 dimension were mapped to specific TDF domains to explain their underlying behavior change mechanisms (eg, “public misunderstanding”

XSL•FO
RenderX

corresponded to the "knowledge" domain). The model correspondence is shown in Figure 2. When overlaps in mapping occurred, the primary corresponding domain was determined through team discussion based on the most direct and dominant driving mechanism reflected in the data. Ultimately, UTAUT2 provided a clear structure of acceptance dimensions (the "what"), while TDF offered an in-depth explanation of behavioral mechanisms (the "why"), with their integration jointly supporting the presentation of the results (as shown in Table 3). Performance expectancy and effort expectancy were the most cited barriers or facilitators to patients' use of AI in medicine. The corresponding text describing barriers and facilitators is provided in Table S8 in Multimedia Appendix 1.

Figure 2. An integrated model of patient acceptance of artificial intelligence (AI) in health care: combining Theoretical Domains Framework (TDF) and Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) framework.

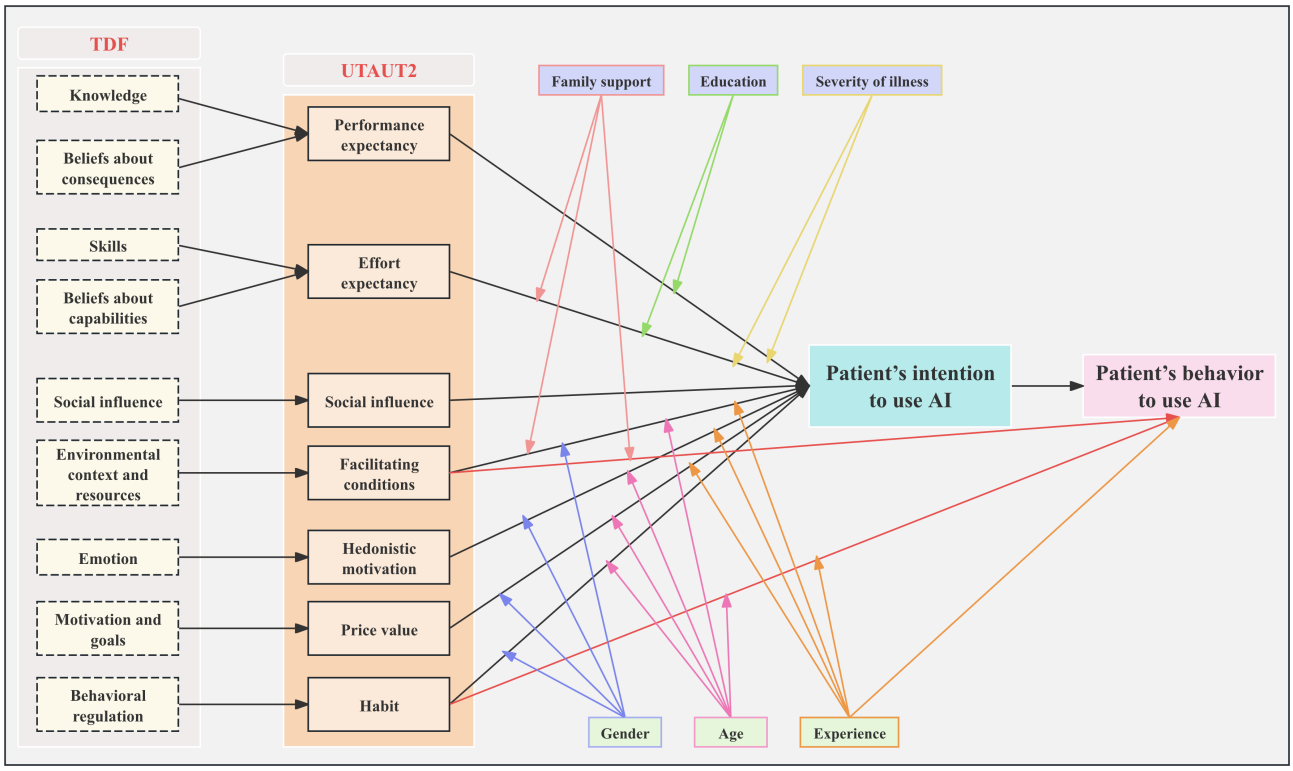

XSL•FO
RenderX

Table . Barriers and facilitators to patient acceptance of medical artificial intelligence (AI)[a], mapped to the integrated Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) and Theoretical Domains Framework (TDF).

UTAUT2[b] model construct and TDF[c] model corresponding domain	Barriers	Facilitators
Performance expectancy		
Knowledge	• Public prejudice and misunderstanding of new technology • Public fear and resistance to new technology	• Understanding the importance of AI[a] in health care
Beliefs about consequences	• Possibly worse outcomes • Lack of trust in AI[a] • Difficult to measure effectiveness	Belief in AI's faster diagnostic speed, high efficiency, and accurate diagnostic results
Effort expectancy		
Skills	• Complex technology operation • Need to invest more time in learning and adapting to new technology applications and operational processes	• Convenient operation
Beliefs about capabilities	• Uncertainty about new technology • Reduced self-efficacy	• Assisting the doctor's decision-making function • Enhanced self-efficacy
Social influence		
Social influences	• Cultural adaptability and intergroup threat • Few people around or family members use AI • Unclear medical responsibility attribution • Insufficient patient-doctor communication	• Public trust • Recommendations from trusted sources (doctors) • Improved patient-doctor communication
Facilitating conditions		
Environmental context and resources	• Insufficient patient privacy security protection • Cybersecurity not ensured • Insufficient data reliability and transparency • Technical errors and false information obtained • Lack of humanistic care in AI • Imperfect AI technology	• Protection of patient privacy (sensitive issues) • Cybersecurity ensured • Reliable data (fairness, interpretability, and transparency)
Hedonistic motivation		
Emotion	• Poor new technology usage experience • Lack of personalized experience	• Novel experience
Price value		
Motivation and goals	• High cost of new technology application • Expensive (unequal opportunities) • Fear of high trial-and-error costs	• Cost savings
Habit		
Behavioral regulation	• Unaccustomed to using AI technology • May make patients unwilling to persist in using new medical technology to manage their health.	—[d]

[a]AI: artificial intelligence.

[b]UTAUT2: Unified Theory of Acceptance and Use of Technology 2.

[c]TDF: Theoretical Domains Framework.

[d]Not available.

XSL•FO
RenderX

Performance Expectancy

Performance expectancy is the extent to which patients feel that the use of AI has contributed to the perception of illness [14], which can be further explained by the dimensions of knowledge, beliefs about outcomes, behavioral regulation, and environmental context and resources in TDF. In summary, patients' performance expectancy of medical AI is primarily shaped by a tension between widespread cognitive biases and misconceptions (n=28) and a strong belief in its efficacy when supported by empirical evidence (n=31).

Effort Expectancy

Effort expectancy reflects the effort required for patients to use AI, linked to TDF dimensions such as skills, cognition, and behavioral regulation [14]. The foremost challenges are high operational complexity and substantial learning investment (n=17), whereas facilitation hinges on interface optimization and process simplification to reduce cognitive load (n=22). This highlights that ease of use is not merely a convenience but a critical prerequisite for patient engagement with AI tools.

Social Influence

Social influence, defined as the impact of surrounding groups on AI adoption, includes subjective norms and public trust [14]. Low technology adoption, reduced clinician-patient interactions (n=26), and ethical concerns (n=27) were frequently reported as barriers. In contrast, facilitators such as endorsements from authoritative institutions and enhanced clinician-patient communication (n=9) were noted as promoting acceptance. Collectively, these findings indicate that social and relational factors, including interpersonal trust and perceived legitimacy, are consistently salient in shaping patient responses to AI. Within the TDF framework, this theme aligns with domains of interpersonal relationships and professional roles.

Facilitating Conditions

Facilitating conditions reflect organizational support for system use, aligning with the TDF dimension of environmental context and resources [14]. Key barriers included data reliability issues, privacy vulnerabilities, and cybersecurity risks, with insufficient privacy protection being a significant concern (n=22). Facilitators focused on establishing transparent data governance, enhancing privacy encryption, and reinforcing cybersecurity protocols to ensure fairness, interpretability, and transparency (n=12).

Hedonistic Motivation

This refers to the degree of pleasure or enjoyment an individual derives from using new technology [15], corresponding to the TDF dimensions of emotion, affect, motivation, and goals. The most frequently reported barrier was the mechanized nature of technology use, which led to a lack of empathy, increased patient anxiety, and a reduced sense of self-efficacy. In contrast, the primary facilitator was the design of personalized interaction models to enhance positive user experiences and integrate emotion-aware algorithms to mitigate patients' negative psychological responses.

Habit

Habit is the tendency to automate behaviors resulting from the patient's frequent use of technology [92]. The corresponding TDF dimension is behavioral regulation. Patients' reluctance to adopt new technologies was the most frequently reported barrier due to their preference for traditional methods.

Price Value

Price value reflects the balance between technology benefits and economic costs [15,93], linked to the TDF domain of environmental context and resources. A cost-benefit calculus shapes perceptions of price value. The primary barrier is the high upfront financial cost and associated risk (n=14), while facilitators focus on policy subsidies and training to alleviate economic burdens and demonstrate long-term value (n=10). This economic perspective confirms that cost-effectiveness and affordability are practical, decisive factors for many patients.

Moderating Variables

The UTAUT2 model highlights moderating variables (eg, age, gender, and experience) influencing AI adoption in health care. This study expanded the analysis to include disease severity and education level. Experience, defined as perceived performance enhancement, was linked to TDF dimensions such as knowledge and environmental context. Prior experience increased AI acceptance, while gender and age did not significantly impact most studies [47,71]. Although findings varied, disease severity and education level were identified as additional moderators. Further research is needed to clarify these relationships and develop targeted interventions for AI adoption.

BCTs and Implementation Strategies

Through group discussions, 27 factors from UTAUT2 and TDF domains were mapped to 93 potential BCTs. After consolidation, 25 distinct BCTs were assessed using the Affordability, Practicability, Effectiveness and Cost-Effectiveness, Acceptability, Side-effects or Safety, and Equity criteria and categorized by feasibility: 6 (24%) as high, 14 (56%) as medium, and 5 (20%) as low (detailed in Table S10 in Multimedia Appendix 1). High-feasibility BCTs (eg, 4.1 and 5.1) are low cost, easy to integrate, and well accepted. Medium-feasibility BCTs (eg, 6.1 and 15.1) are evidence-based but face operational constraints, while low-feasibility BCTs (eg, 9.1 and 12.1) require system- or policy-level changes beyond clinical teams. Concrete examples for these BCTs were drawn from the literature (Table S9 in Multimedia Appendix 1), resulting in 40 targeted interventions. Complete results are shown in Table 4 (the serial numbers and other details of the relevant BCT are provided in Table S3 in Multimedia Appendix 1).

XSL•FO
RenderX

Table . Behavior change techniques (BCTs) mapped from identified barriers and facilitators to inform interventions for enhancing patient artificial intelligence (AI) acceptance.

UTAUT2[a] component and relevant BCT[b]	Target level	Feasibility	Example quotes
Performance expectancy			
TDF[c] domain: knowledge			
Instructions on how to perform the behavior	Patient	High	• Provide information on the effects of using AIM[d] and educating patients about the benefit of the current health care system [76], and • participants requested that information and training be provided in a number of different ways (eg, pamphlets, in-person training, computer-guided supports, and sharing of patient experiences) [94].
Demonstration of the behavior	Patient	High	Detailed instructional content: provide a comprehensive, step-by-step guide on the functionality and operation of AI[e] systems, including visual aids such as diagrams, flowcharts, and videos to illustrate key concepts and procedures.
TDF domain: beliefs about consequences			
Information about health consequences	Patient	High	Provide information (eg, written, verbal, and visual) about AI; educate and involve patients in the future direction of this technology; demonstrating to patients that the use of AI in their own health care is safe, effective, and ethical; and establish a baseline educational level [72].
Salience of consequences	Patient	Medium	Using methods specifically designed to emphasize the consequences of performing the behavior to make them more memorable.
Anticipated regret	Patient	Medium	Enhance patients' awareness of the usefulness of AI by showcasing successful cases to patients, provide them with a preliminary understanding of the expected outcomes.
Effort expectancy			
TDF domain: skills (constructs other than interpersonal skills)			
Goal setting (behavior)	Patient	High	Devise ways to perform the behavior: provide participants with an AI device; patients need to understand all information about AI and require that all health care providers undergo technology training [47].

XSL•FO
RenderX

UTAUT2[a] component and relevant BCT[b]	Target level	Feasibility	Example quotes
Feedback on behavior	Patient	Medium	Monitor and provide feedback on exercise performance (exercise type, frequency, duration, and intensity), and monitor and show performance via mobile phone or sports bracelet.
Self-monitoring of behavior	Patient	High	Establish a method for participants to monitor and record their AI use.
Feedback on outcomes of behavior	Patient	Medium	Provide feedback about body metrics using AI devices: use sport bracelets, heart rate bands, and other devices to monitor physiological indicators; by optimizing the structure of the application and designing user-friendly apps with simple navigation and clear interfaces, usability can be improved, thereby increasing patient acceptance [84].
TDF domain: beliefs about capabilities			
Verbal persuasion about capability	Clinician	Medium	Enhance the professional quality of medical staff by improving the professional capabilities of medical staff, enhancing patients' trust in the medical staff and the AI products they recommend.
Social influences			
TDF domain: social influences			
Social support (unspecified)	All	Low	Social support (practical) formulates comprehensive laws and regulations; clarifies the liability attribution of AI in the medical field to protect the legal rights and interests of patients and medical staff [63]; strengthens publicity efforts, establishes management systems and standards, and reinforces communication and guidance.
Information about others' approval	Patient	Medium	Provide information about what other people think about AI; make participants aware that many patients have a positive attitude toward AI; and leverage primary care providers' recommendations [45].
Restructuring the social environment	All	Medium	Provide participants with a social environment suitable for using AI; change perspectives of participants' family members so that they support the participant's choice; emphasize communication between patients and medical staff. When adopting AI technology, focus on patient experience, reduce repetitive work time of medical staff, increase communication with patients, and implement the core concept of humanistic care nursing.
Facilitating conditions			

XSL•FO
RenderX

UTAUT2[a] component and relevant BCT[b]	Target level	Feasibility	Example quotes
TDF domain: environmental context and resources			
Monitoring of behavior by others without feedback	Clinician	Low	Monitoring of participants' use without feedback: monitoring of participants' use by professionals (eg, physical therapists, nursing staff, and so on) to optimize the delivery of follow-up interventions.
Credible source	System	Low	Ensure reliable and transparent data sources: safeguard the security of the network environment used by AI and ensure that patients' privacy data is not leaked.
Restructuring the physical environment	System	Low	Provide a safe and reliable data environment; formulate relevant policies and behavioral norms to ensure data security and reliability. Provide a safe place to use AI; ensure protection of data and provide patients with assurance regarding strict conditionality. Convey purpose, benefits, and potential risks in written form (brochures) for patients [51]; ensure suitable AI equipment and convenient access; protect patient privacy and increase transparency in medical data processing; enhance transparency in the decision-making process [65].
Restructuring the social environment	System	Medium	Provide a social environment that promotes AI use, ensuring AI tools preserve the integrity of the human physician-patient relationship [40].
Adding objects to the environment	System	Low	Add objects to the environment to facilitate behavior, provide personalized AIM services based on patients' specific situations to enhance their perceived value.
Hedonistic motivation			
TDF domain: emotion			
Monitoring of emotional consequences	Clinician	Medium	Provide emotional support (from professionals, exercise partners, family, or friends)
Information about emotional consequences	Clinician	Medium	Monitoring participants' feelings after use: monitor participants' feelings to avoid negative experiences and adjust subsequent tasks accordingly.
Price value			
TDF domain: motivation and goals			

XSL•FO
RenderX

UTAUT2[a] component and relevant BCT[b]	Target level	Feasibility	Example quotes
Review outcome goals	All	Medium	Empowering personalized care and patient autonomy with feedback: AI assists health care professionals in providing personalized care while safeguarding patient autonomy and decision-making ability. Meanwhile, patient usage feedback is promptly adopted to optimize AI use and enhance patient satisfaction.
Habit			
TDF domain: behavioral regulation			
Goal setting (behavior)	Patient	High	Goal setting: encourage patients to gradually adapt to AI operation through exploration and practice.
Action planning	System	Low	Tailored using programs for participants: regulatory agencies should establish normative standards and evaluation guidelines for implementing AI in health care in cooperation with health care institutions [62].
Feedback on behavior	Patient	Medium	Monitor and provide feedback on exercise performance (eg, exercise type, frequency, duration, and intensity) via mobile phone or sports bracelet.
Habit formation	Clinician	Medium	Timely adoption of patient usage feedback: regularly follow up with patients to understand their usage of AI technology and willingness to continue using it.

[a]UTAUT2: Unified Theory of Acceptance and Use of Technology 2.

[b]BCT: behavior change technique.

[c]TDF: Theoretical Domains Framework.

[d]AIM: artificial intelligence in medicine.

[e]AI: artificial intelligence.

Discussion

Principal Findings

This review aimed to explore the main barriers and enablers underlying patients' adoption of AI in hospital settings. Prior studies have examined the intention to use AI among different types of patients, but there has been no systematic synthesis of barriers and facilitators. This study used an innovative approach to synthesize patients' views on using AI in medicine, integrating barriers and facilitators of patient adoption of AI in medicine based on the UTAUT2 and TDF.

Our findings reveal that AI in medicine adoption is driven by a 3-tier mechanism: "technology acceptance, contextual modulation, and ethical security." Moreover, individual differences influence AI acceptance. For example, age moderates the perception of technological complexity, with older adults favoring simpler systems [67,72]. Education level is a cognitive mediator, affecting performance expectancy by shaping AI-related knowledge [51,60,65,71,75,84]. Additionally, disease severity exhibits an inverted U-shaped relationship with AI acceptance, where patients with moderate conditions demonstrate the highest acceptance [64,73].

Regarding the most important barriers and enablers identified in this review, performance expectancy and effort expectancy were the most cited factors influencing patients' use of AI in medicine. This finding substantiates the results of earlier studies [37,61,67]. Patients' recognition of AI's diagnostic efficiency and accuracy emerges as the primary determinant of acceptance under performance expectancy [15,30], aligning with findings from Schmitz et al [16] in cross-national telemedicine research. Additionally, within the TDF framework, the "knowledge" domain significantly impacts patient adoption of AI, highlighting

XSL•FO
RenderX

that awareness and understanding of AI technologies are crucial determinants of acceptance and use [37,61,67]. There is a need to improve AI knowledge and increase willingness to use AI, which can be achieved by enhancing patient education through outpatient health seminars and multimedia educational videos.

In addition, the operational feasibility of AI plays a key role in promoting patient use. The perceived value of AI for patients increases significantly when they are able to receive positive feedback and are effectively supported by health care professionals when using it [95]. For example, an AI robotic system that receives positive feedback from health care professionals during rehabilitation training can enhance patients' confidence and sense of accomplishment, further promoting the usage of AI [96]. However, poorer usability can be a source of fear for patients toward new technologies and, therefore, may hinder their adoption [96]. A "progressive complexity" strategy, embedding educational modules within basic applications, may facilitate gradual comprehension and trust development in AI systems [36,84].

Social networks profoundly shape patient willingness to use AI, extending beyond physician recommendations to include family members, friends, and community influence. Family acceptance acts as a critical multiplier—tech-savvy relatives often facilitate AI adoption by assisting with device operation or interpreting digital health reports, whereas households with low health literacy may resist AI due to distrust of unfamiliar technology [63]. Peer influence also plays a role: patients are more likely to embrace AI-driven tools after observing positive experiences shared by friends or online communities, particularly in managing chronic conditions such as diabetes or hypertension. Conversely, negative anecdotes within social circles (eg, stories of AI misdiagnosis) can amplify skepticism, creating "information cascades" that undermine trust [47]. These dynamics highlight the need for targeted social marketing strategies, for example, family-centered education workshops or peer advocacy programs to address intergenerational and community-specific barriers.

Cultural adaptability is primarily shaped by specific constructs within the UTAUT2 and TDF frameworks, with social influence being the most culturally sensitive dimension. This review finds that in collectivist societies (eg, East Asia), social influence often manifests as deference to family or group consensus [97], effectively shifting the adoption unit from the individual to the collective. Conversely, in individualistic cultures, it leans more on trust in professional authority or empirical evidence. Consequently, implementation strategies must be contextualized: collectivist settings may require designs that facilitate family-inclusive decision-making and leverage community opinion leaders, whereas individualistic contexts should emphasize transparent personal benefits and institutional endorsements. Systematically analyzing culture through these theoretical lenses enables more targeted and effective AI implementation across diverse health care systems.

Ethical safety concerns, particularly the "algorithmic black box," undermine trust in AI's reliability among patients and clinicians [45,58,62,64,67]. Explainable AI, by providing transparent diagnostic reasoning and interpretable reports, mitigates skepticism and strengthens technical credibility [39]. While AI may reduce patient embarrassment in sensitive consultations [46], overreliance risks eroding humanistic doctor-patient interactions, exacerbating anxiety [37,40,44,45,49,50]. Balancing privacy protection with human-centered design remains a critical research frontier.

Beyond mitigating ethical and relational barriers, our synthesis points toward the necessity of proactively rehumanizing AI-driven health care. This concept involves intentionally designing AI systems and their implementation contexts to preserve and enhance, rather than replace, essential human elements such as provider presence, empathy, and meaningful interaction. Studies, such as the work of Almokdad et al [98] in digital hospitality, underscore the critical importance of human presence in mitigating the impersonal nature of automated services. Translating this to health care, AI should be positioned as a tool that augments, rather than replaces, the clinician. To operationalize this, concrete strategies include (1) adopting "AI-as-assistant" design paradigms where clinicians remain central in communication and decision-making, (2) implementing empathy-aware algorithms to prompt providers to address patient emotional states, and (3) leveraging AI's efficiency gains to protect or increase time for direct patient-clinician interaction.

Additional psychological and economic factors include hedonic motivation, habituation, and price-value trade-offs. Positive emotional experiences with AI (eg, intuitive interfaces) boost initial acceptance [36,47,83,89,96], whereas habit-driven preference for traditional methods necessitates proactive demonstration of AI's clinical advantages [38,62,66,83,87,89]. Although AI adoption may incur higher upfront costs [9,58,63,66,67,81,89], its long-term potential to reduce unnecessary diagnostics and optimize treatment pathways could yield cost savings [88,90]. Future research should contextualize AI's economic impact within health care settings while prioritizing personalized, culturally sensitive implementations.

The BCT taxonomy (version 1) provides a standardized framework for behavioral interventions in AI implementation [34]. This study advances its practical application by introducing a feasibility lens, which transforms a theoretical list of techniques into a prioritized implementation strategy. Our assessment distinguishes high-feasibility BCTs (eg, instruction and goal setting) as immediately actionable tools for frontline adaptation, such as optimizing interfaces for older adults [99]. Medium-feasibility BCTs, including strategies such as social comparison to address privacy concerns [55], require deliberate planning and resources. Ultimately, sustainable adoption depends on low-feasibility, systemic enablers, such as transparent data governance, which establish the necessary foundation for trust. Integrating barrier mapping with this feasibility-driven prioritization creates a clear, staged model for implementation. It clarifies what clinicians can do now, what requires organizational support, and what must be advocated at the policy level, offering a pragmatic roadmap for integrating AI into health care.

In summary, the primary contribution of this study lies in its theoretical integration and methodological innovation. Distinct

XSL•FO
RenderX

from previous research that predominantly focused on technological attributes or isolated psychosocial factors, this review establishes a comprehensive analytical model that links technology acceptance with behavior change by integrating the UTAUT2 and TDF frameworks. Beyond this theoretical contribution, this integrative model deepens the understanding of barriers and facilitators to patient adoption of AI and, importantly, translates these insights into actionable evidence by mapping the findings to specific and prioritized BCTs. This evidence directly informs and empowers multiple stakeholders: it provides health care administrators with a blueprint for implementation strategies, equips clinicians with tools for patient engagement, guides AI developers toward human-centered design principles, and supports policymakers in developing frameworks for ethical evaluation and governance. Ultimately, this review bridges the gap between theoretical understanding and practical application, thereby enhancing the translational value of research on AI acceptance in health care.

Limitations and Strengths

The strength of this study lies in its systematic synthesis of qualitative and quantitative evidence, conducted through a mixed methods evaluation guided by an integrated theoretical framework. Theoretically, it innovatively combines the UTAUT2 model with the TDF to analyze barriers and facilitators influencing patients' adoption of AI in health care, thereby providing a dual perspective that incorporates both technology acceptance and behavioral science. This theory-informed approach allows for a coherent interpretation of existing evidence and establishes a robust foundation to support patient-centered AI implementation, while also informing the design of future interventions and health policy initiatives.

This review has several limitations. First, methodological constraints include potential publication bias and language restrictions (Chinese and English only), which may affect the comprehensiveness and cultural generalizability of the findings. Second, the evidence base is limited by the uneven geographical distribution of included studies, with underrepresentation from developing countries, and the retention of 2 lower-quality studies that did not alter the overall conclusions. Third, the proposed interventions are theoretically derived from TDF-BCT mapping and lack empirical validation. Finally, given the rapid evolution of AI technology and its regulatory context, the findings represent a snapshot in time and require ongoing re-evaluation as the field advances.

Conclusions

This study uses a mixed methods approach to explore key factors influencing patient acceptance of health care AI. Integrating the UTAUT2 model with the theoretical domains of the TDF systematically identifies barriers and facilitators, emphasizing the need for tailored implementation strategies. Theory-based interventions are proposed by mapping the TDF to BCTs. Findings highlight that building trust in technology and prioritizing human-centered design are essential for acceptance, while privacy and ethical concerns remain significant barriers. Further research is needed to assess the effectiveness and sustainability of patient-centered, multicomponent interventions and to establish a dynamic monitoring system for evolving patient perceptions amid technological advancements.

Acknowledgments

The declaration is submitted under collective responsibility. According to the GAIDeT taxonomy (2025), the following tasks were delegated to generative artificial intelligence (GAI) tools under full human supervision: Translation: the GAI tools used were DeepSeek and Gemini. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the outcomes.

Funding

This research was supported by grants from the Zhejiang Health Science and Technology Project (grant no 2025KY081).

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

HS was responsible for the protocol design, literature search of databases, study selection, and data extraction processes. HS and MH contributed to the selection and data extraction processes. JH and MH contributed to the quality appraisal process. HS, JH, JY, and MH were responsible for the classification of the themes. HS wrote the first draft of the manuscript. HS, JH, JY, and MH reviewed the manuscript for content analysis and discussion. JH supervised and coordinated the entire review process.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed literature search strategies, quality assessment results, and extracted raw data of barriers and facilitators.
[DOCX File, 204 KB - jmir_v28i1e80581_app1.docx]

XSL•FO
RenderX

Checklist 1
PRISMA checklist.
[DOCX File, 18 KB - jmir_v28i1e80581_app2.docx]

References

1. Aung YYM, Wong DCS, Ting DSW. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. Br Med Bull 2021 Sep 10;139(1):4-15. [doi: 10.1093/bmb/ldab016] [Medline: 34405854]
2. Liu R, Rong Y, Peng Z. A review of medical artificial intelligence. Global Health Journal 2020 Jun;4(2):42-45. [doi: 10.1016/j.glohj.2020.04.002]
3. Clark HB, Egger J, Duffy VG. AI in healthcare and medicine: a systematic literature review and reappraisal. Presented at: Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management: 15th International Conference, DHM 2024, Held as Part of the 26th HCI International Conference, HCII 2024; Jun 29 to Jul 4, 2024; Washington, DC p. 251-270. [doi: 10.1007/978-3-031-61063-9_17]
4. Krittanawong C. The rise of artificial intelligence and the uncertain future for physicians. Eur J Intern Med 2018 Feb;48:e13-e14. [doi: 10.1016/j.ejim.2017.06.017] [Medline: 28651747]
5. Ibrahim F, Münscher JC, Daseking M, Telle NT. The technology acceptance model and adopter type analysis in the context of artificial intelligence. Front Artif Intell 2024;7:1496518. [doi: 10.3389/frai.2024.1496518] [Medline: 39886023]
6. Chismar WG, Wiley-Patton S. Does the extended technology acceptance model apply to physicians. Presented at: 36th Annual Hawaii International Conference on System Sciences, 2003 Proceedings of the; Jan 6-9, 2003. [doi: 10.1109/HICSS.2003.1174354]
7. Liyanage H, Liaw ST, Jonnagaddala J, et al. Artificial intelligence in primary health care: perceptions, issues, and challenges. Yearb Med Inform 2019 Aug;28(1):41-46. [doi: 10.1055/s-0039-1677901] [Medline: 31022751]
8. Kovarik CL. Patient perspectives on the use of artificial intelligence. JAMA Dermatol 2020 May 1;156(5):493-494. [doi: 10.1001/jamadermatol.2019.5013] [Medline: 32159724]
9. Ayad N, Schwendicke F, Krois J, et al. Patients' perspectives on the use of artificial intelligence in dentistry: a regional survey. Head Face Med 2023 Jun 22;19(1):23. [doi: 10.1186/s13005-023-00368-z] [Medline: 37349791]
10. Moy S, Irannejad M, Manning SJ, et al. Patient perspectives on the use of artificial intelligence in health care: a scoping review. J Patient Cent Res Rev 2024;11(1):51-62. [doi: 10.17294/2330-0698.2029] [Medline: 38596349]
11. Alsanosi SM, Aldajani AQ, Gheliwi HA, et al. Knowledge, attitudes, and perceptions of chronic patients in Saudi Arabia regarding the use of artificial intelligence to improve medication adherence. Patient Prefer Adherence 2025;19:1781-1792. [doi: 10.2147/PPA.S519427] [Medline: 40552260]
12. Gundlack J, Negash S, Thiel C, et al. Artificial intelligence in medical care - patients' perceptions on caregiving relationships and ethics: a qualitative study. Health Expect 2025 Apr;28(2):e70216. [doi: 10.1111/hex.70216] [Medline: 40094179]
13. Krisnan L, Masrom M, Yahya Y, Krishnan M. Patient perceptions toward the application of artificial intelligence in diabetes care: a qualitative study. Asian J Soc Health Behav 2025;8(1):15-20. [doi: 10.4103/shb.shb_92_24]
14. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]
15. Schomakers EM, Lidynia C, Vervier LS, Calero Valdez A, Ziefle M. Applying an extended UTAUT2 model to explain user acceptance of lifestyle and therapy mobile health apps: survey study. JMIR Mhealth Uhealth 2022 Jan 18;10(1):e27095. [doi: 10.2196/27095] [Medline: 35040801]
16. Schmitz A, Díaz-Martín AM, Yagüe Guillén MJ. Modifying UTAUT2 for a cross-country comparison of telemedicine adoption. Comput Human Behav 2022 May;130:107183. [doi: 10.1016/j.chb.2022.107183] [Medline: 35017788]
17. Wang H, Tao D, Yu N, Qu X. Understanding consumer acceptance of healthcare wearable devices: an integrated model of UTAUT and TTF. Int J Med Inform 2020 Jul;139:104156. [doi: 10.1016/j.ijmedinf.2020.104156] [Medline: 32387819]
18. Palas JU, Sorwar G, Hoque MR, Sivabalan A. Factors influencing the elderly's adoption of mHealth: an empirical study using extended UTAUT2 model. BMC Med Inform Decis Mak 2022 Jul 24;22(1):191. [doi: 10.1186/s12911-022-01917-3] [Medline: 35871682]
19. Chang YT, Chao CM, Yu CW, Lin FC. Extending the utility of UTAUT2 for hospital patients' adoption of medical apps: moderating effects of e-health literacy. Mob Inf Syst 2021 Apr 12;2021:1-10. [doi: 10.1155/2021/8882317]
20. Cane J, O'Connor D, Michie S. Validation of the theoretical domains framework for use in behaviour change and implementation research. Implement Sci 2012 Apr 24;7:37. [doi: 10.1186/1748-5908-7-37] [Medline: 22530986]
21. Cane J, Richardson M, Johnston M, Ladha R, Michie S. From lists of behaviour change techniques (BCTs) to structured hierarchies: comparison of two methods of developing a hierarchy of BCTs. Br J Health Psychol 2015 Feb;20(1):130-150. [doi: 10.1111/bjhp.12102] [Medline: 24815766]
22. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]

XSL•FO
RenderX

23. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
24. Lockwood C, Munn Z, Porritt K. Qualitative research synthesis: methodological guidance for systematic reviewers utilizing meta-aggregation. Int J Evid Based Healthc 2015 Sep;13(3):179-187. [doi: 10.1097/XEB.0000000000000062] [Medline: 26262565]
25. Hong QN, Fàbregues S, Bartlett G, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. EFI 2018;34(4):285-291. [doi: 10.3233/EFI-180221]
26. Aromataris E, L C, Porritt K, Pilla B, Jordan Z. JBI manual for evidence synthesis. JBI. 2024. URL: https://jbi-global-wiki.refined.site/space/MANUAL [accessed 2026-04-14]
27. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008 Jul 10;8(1):45. [doi: 10.1186/1471-2288-8-45] [Medline: 18616818]
28. Popay J, Roberts H, Sowden A, et al. Guidance on the conduct of narrative synthesis in systematic reviews: a product from the ESRC Methods Programme. : Lancaster University; 2006 URL: https://doi.org/10.13140/2.1.1018.4643 [accessed 2026-04-14]
29. Camacho J, Zanoletti-Mannello M, Landis-Lewis Z, Kane-Gill SL, Boyce RD. A conceptual framework to study the implementation of clinical decision support systems (BEAR): literature review and concept mapping. J Med Internet Res 2020 Aug 6;22(8):e18388. [doi: 10.2196/18388] [Medline: 32759098]
30. Venkatesh V, Thong JYL, Xu X. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS Q 2012 Mar 1;36(1):157-178. [doi: 10.2307/41410412]
31. Di Castelnuovo A, Bonaccio M, Costanzo S, et al. Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with COVID-19: survival analysis and machine learning-based findings from the multicentre Italian CORIST Study. Nutr Metab Cardiovasc Dis 2020 Oct 30;30(11):1899-1913. [doi: 10.1016/j.numecd.2020.07.031] [Medline: 32912793]
32. Benefits and complication of peripherally inserted central catheter (PICC) versus peripherally intravenous vein lines. World Health Organization. 2012. URL: https://trialsearch.who.int/Trial2.aspx?TrialID=IRCT138903204113N2 [accessed 2026-03-27]
33. Atkins L, Francis J, Islam R, et al. A guide to using the Theoretical Domains Framework of behaviour change to investigate implementation problems. Implement Sci 2017 Jun 21;12(1):77. [doi: 10.1186/s13012-017-0605-9] [Medline: 28637486]
34. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
35. Michie S, Atkins L, Wes R. The Behaviour Change Wheel: A Guide to Developing Interventions: Silverback Publishing; 2014. URL: https://www.behaviourchangewheel.com/ [accessed 2026-04-14]
36. Huang W, Ong WC, Wong MKF, et al. Applying the UTAUT2 framework to patients' attitudes toward healthcare task shifting with artificial intelligence. BMC Health Serv Res 2024 Apr 11;24(1):455. [doi: 10.1186/s12913-024-10861-z] [Medline: 38605373]
37. Haan M, Ongena YP, Hommes S, Kwee TC, Yakar D. A qualitative study to understand patient perspective on the use of artificial intelligence in radiology. J Am Coll Radiol 2019 Oct;16(10):1416-1419. [doi: 10.1016/j.jacr.2018.12.043] [Medline: 30878311]
38. Jalil S, Myers T, Atkinson I, Soden M. Complementing a clinical trial with human-computer interaction: patients' user experience with telehealth. JMIR Hum Factors 2019 Jun 6;6(2):e9481. [doi: 10.2196/humanfactors.9481] [Medline: 31172958]
39. Adams SJ, Tang R, Babyn P. Patient perspectives and priorities regarding artificial intelligence in radiology: opportunities for patient-centered radiology. J Am Coll Radiol 2020 Aug;17(8):1034-1036. [doi: 10.1016/j.jacr.2020.01.007] [Medline: 32068006]
40. Nelson CA, Pérez-Chada LM, Creadore A, et al. Patient perspectives on the use of artificial intelligence for skin cancer screening: a qualitative study. JAMA Dermatol 2020 May 1;156(5):501-512. [doi: 10.1001/jamadermatol.2019.5014] [Medline: 32159733]
41. Benrimoh D, Tanguay-Sela M, Perlman K, et al. Using a simulation centre to evaluate preliminary acceptability and impact of an artificial intelligence-powered clinical decision support system for depression treatment on the physician-patient interaction. BJPsych Open 2021 Jan 6;7(1):e22. [doi: 10.1192/bjo.2020.127] [Medline: 33403948]
42. Musbahi O, Syed L, Le Feuvre P, Cobb J, Jones G. Public patient views of artificial intelligence in healthcare: a nominal group technique study. Digit Health 2021;7:20552076211063682. [doi: 10.1177/20552076211063682] [Medline: 34950499]
43. Zhang Z, Citardi D, Wang D, Genc Y, Shan J, Fan X. Patients' perceptions of using artificial intelligence (AI)-based technology to comprehend radiology imaging data. Health Informatics J 2021;27(2):14604582211011215. [doi: 10.1177/14604582211011215] [Medline: 33913359]

XSL•FO
RenderX

44. Mikkelsen JG, Sørensen NL, Merrild CH, Jensen MB, Thomsen JL. Patient perspectives on data sharing regarding implementing and using artificial intelligence in general practice - a qualitative study. BMC Health Serv Res 2023 Apr 4;23(1):335. [doi: 10.1186/s12913-023-09324-8] [Medline: 37016412]
45. Pelayo C, Hoang J, Mora Pinzón M, et al. Perspectives of Latinx patients with diabetes on teleophthalmology, artificial intelligence-based image interpretation, and virtual care: a qualitative study. Telemed Rep 2023;4(1):317-326. [doi: 10.1089/tmr.2023.0045] [Medline: 37908628]
46. Pelly M, Fatehi F, Liew D, Verdejo-Garcia A. Artificial intelligence for secondary prevention of myocardial infarction: a qualitative study of patient and health professional perspectives. Int J Med Inform 2023 May;173:105041. [doi: 10.1016/j.ijmedinf.2023.105041] [Medline: 36934609]
47. Robertson C, Woods A, Bergstrand K, Findley J, Balser C, Slepian MJ. Diverse patients' attitudes towards artificial intelligence (AI) in diagnosis. PLOS Digit Health 2023 May;2(5):e0000237. [doi: 10.1371/journal.pdig.0000237] [Medline: 37205713]
48. Neves MVM, Furlan L, Fregni F, et al. Robotic-assisted gait training (RAGT) in stroke rehabilitation: a pilot study. Arch Rehabil Res Clin Transl 2023;5(1) [FREE Full text]
49. Godoy Junior CA, Miele F, Mäkitie L, et al. Attitudes toward the adoption of remote patient monitoring and artificial intelligence in Parkinson's disease management: perspectives of patients and neurologists. Patient 2024 May;17(3):275-285. [doi: 10.1007/s40271-023-00669-0] [Medline: 38182935]
50. Maris MT, Koçar A, Willems DL, et al. Ethical use of artificial intelligence to prevent sudden cardiac death: an interview study of patient perspectives. BMC Med Ethics 2024 Apr 4;25(1):42. [doi: 10.1186/s12910-024-01042-y] [Medline: 38575931]
51. Sachdeva M, Datchoua AM, Yakam VF, et al. Acceptability of artificial intelligence for cervical cancer screening in Dschang, Cameroon: a qualitative study on patient perspectives. Reprod Health 2024 Jun 28;21(1):92. [doi: 10.1186/s12978-024-01828-8] [Medline: 38937771]
52. Tursynbek A, Zhaksylykova D, Cruz JP, Balay-Odao EM. Perspectives of patients regarding artificial intelligence and its application in healthcare: a qualitative study. J Clin Nurs 2024 Dec 3. [doi: 10.1111/jocn.17584] [Medline: 39628063]
53. Gundlack J, Thiel C, Negash S, et al. Patients' perceptions of artificial intelligence acceptance, challenges, and use in medical care: qualitative study. J Med Internet Res 2025 May 15;27:e70487. [doi: 10.2196/70487] [Medline: 40373300]
54. Trivedi R, Shaw T, Sheahen B, Chow CK, Laranjo L. Patient perspectives on conversational artificial intelligence for atrial fibrillation self-management: qualitative analysis. J Med Internet Res 2025 Mar 12;27:e64325. [doi: 10.2196/64325] [Medline: 40073398]
55. Sebastian G, George A, Jackson G Jr. Persuading patients using rhetoric to improve artificial intelligence adoption: experimental study. J Med Internet Res 2023 Mar 13;25:e41430. [doi: 10.2196/41430] [Medline: 36912869]
56. Zhou Y, Shi Y, Lu W, Wan F. Did artificial intelligence invade humans? The study on the mechanism of patients' willingness to accept artificial intelligence medical care: from the perspective of intergroup threat theory. Front Psychol 2022;13:866124-866124. [doi: 10.3389/fpsyg.2022.866124]
57. Yang K, Zeng Z, Peng H, Jiang Y. Attitudes of Chinese cancer patients toward the clinical use of artificial intelligence. Patient Prefer Adherence 2019;13:1867-1875. [doi: 10.2147/PPA.S225952] [Medline: 31802856]
58. Jutzi TB, Krieghoff-Henning EI, Holland-Letz T, et al. Artificial intelligence in skin cancer diagnostics: the patients' perspective. Front Med (Lausanne) 2020;7:233. [doi: 10.3389/fmed.2020.00233] [Medline: 32671078]
59. Meyer AND, Giardina TD, Spitzmueller C, Shahid U, Scott TMT, Singh H. Patient perspectives on the usefulness of an artificial intelligence-assisted symptom checker: cross-sectional survey study. J Med Internet Res 2020 Jan 30;22(1):e14679. [doi: 10.2196/14679] [Medline: 32012052]
60. Ongena YP, Haan M, Yakar D, Kwee TC. Patients' views on the implementation of artificial intelligence in radiology: development and validation of a standardized questionnaire. Eur Radiol 2020 Feb;30(2):1033-1040. [doi: 10.1007/s00330-019-06486-0] [Medline: 31705254]
61. Aggarwal R, Farag S, Martin G, Ashrafian H, Darzi A. Patient perceptions on data sharing and applying artificial intelligence to health care data: cross-sectional survey. J Med Internet Res 2021 Aug 26;23(8):e26162. [doi: 10.2196/26162] [Medline: 34236994]
62. Esmaeilzadeh P, Mirzaei T, Dharanikota S. Patients' perceptions toward human-artificial intelligence interaction in health care: experimental study. J Med Internet Res 2021 Nov 25;23(11):e25856. [doi: 10.2196/25856] [Medline: 34842535]
63. Liu T, Tsang W, Huang F, et al. Patients' preferences for artificial intelligence applications versus clinicians in disease diagnosis during the SARS-CoV-2 pandemic in China: discrete choice experiment. J Med Internet Res 2021 Feb 23;23(2):e22841. [doi: 10.2196/22841] [Medline: 33493130]
64. Lennartz S, Dratsch T, Zopfs D, et al. Use and control of artificial intelligence in patients across the medical workflow: single-center questionnaire study of patient perspectives. J Med Internet Res 2021 Feb 17;23(2):e24221. [doi: 10.2196/24221] [Medline: 33595451]
65. Armero W, Gray KJ, Fields KG, Cole NM, Bates DW, Kovacheva VP. A survey of pregnant patients' perspectives on the implementation of artificial intelligence in clinical care. J Am Med Inform Assoc 2022 Dec 13;30(1):46-53. [doi: 10.1093/jamia/ocac200] [Medline: 36250788]

XSL•FO
RenderX

66. Yulan Z, Shu X, Yang Z, Chen Q. An exploration of the current status and influencing factors of elderly patients' willingness to use artificial intelligence robots. West China Medical Journal 2022;37(9) [FREE Full text]
67. Khullar D, Casalino LP, Qian Y, Lu Y, Krumholz HM, Aneja S. Perspectives of patients about artificial intelligence in health care. JAMA Netw Open 2022 May 2;5(5):e2210309. [doi: 10.1001/jamanetworkopen.2022.10309] [Medline: 35507346]
68. Kosan E, Krois J, Wingenfeld K, Deuter CE, Gaudin R, Schwendicke F. Patients' perspectives on artificial intelligence in dentistry: a controlled study. J Clin Med 2022 Apr 12;11(8):2143. [doi: 10.3390/jcm11082143] [Medline: 35456236]
69. Kawsar A, Hussain K, Kalsi D, Kemos P, Marsden H, Thomas L. Patient perspectives of artificial intelligence as a medical device in a skin cancer pathway. Front Med (Lausanne) 2023;10:1259595. [doi: 10.3389/fmed.2023.1259595] [Medline: 38046409]
70. Mahlknecht A, Engl A, Piccoliori G, Wiedermann CJ. Supporting primary care through symptom checking artificial intelligence: a study of patient and physician attitudes in Italian general practice. BMC Prim Care 2023 Sep 4;24(1):174. [doi: 10.1186/s12875-023-02143-0] [Medline: 37661285]
71. Parry MW, Markowitz JS, Nordberg CM, Patel A, Bronson WH, DelSole EM. Patient perspectives on artificial intelligence in healthcare decision making: a multi-center comparative study. Indian J Orthop 2023 May;57(5):653-665. [doi: 10.1007/s43465-023-00845-2] [Medline: 37122674]
72. Temple S, Rowbottom C, Simpson J. Patient views on the implementation of artificial intelligence in radiotherapy. Radiography (Lond) 2023 May;29 Suppl 1:S112-S116. [doi: 10.1016/j.radi.2023.03.006] [Medline: 36964044]
73. Baghdadi LR, Mobeirek AA, Alhudaithi DR, et al. Patients' attitudes toward the use of artificial intelligence as a diagnostic tool in radiology in Saudi Arabia: cross-sectional study. JMIR Hum Factors 2024 Aug 7;11:e53108. [doi: 10.2196/53108] [Medline: 39110973]
74. Bahadir HS, Keskin NB, Çakmak E, Güneç G, Cesur Aydin K, Peker F. Patients' attitudes toward artificial intelligence in dentistry and their trust in dentists. Oral Radiol 2025 Jan;41(1):52-59. [doi: 10.1007/s11282-024-00775-1] [Medline: 39379636]
75. Esin H, Karaali C, Teker K, et al. Patients' perspectives on the use of artificial intelligence and robots in healthcare. Bratisl Lek Listy 2024;125(8):513-518. [doi: 10.4149/BLL_2024_80] [Medline: 38989754]
76. Fransen SJ, Kwee TC, Rouw D, et al. Patient perspectives on the use of artificial intelligence in prostate cancer diagnosis on MRI. Eur Radiol 2025 Feb;35(2):769-775. [doi: 10.1007/s00330-024-11012-y] [Medline: 39143247]
77. Riedl R, Hogeterp SA, Reuter M. Do patients prefer a human doctor, artificial intelligence, or a blend, and is this preference dependent on medical discipline? Empirical evidence and implications for medical practice. Front Psychol 2024;15:1422177. [doi: 10.3389/fpsyg.2024.1422177] [Medline: 39188871]
78. Chan J, Parker L, Carter S, Nickel B, Carroll S. Radiation oncology patients' perceptions of artificial intelligence and machine learning in cancer care: a multi-centre cross-sectional study. Radiother Oncol 2025 Jun;207:110891. [doi: 10.1016/j.radonc.2025.110891] [Medline: 40233873]
79. Ozcan BB, Dogan BE, Xi Y, Knippa EE. Patient perception of artificial intelligence use in interpretation of screening mammograms: a survey study. Radiol Imaging Cancer 2025 May;7(3):e240290. [doi: 10.1148/rycan.240290] [Medline: 40249272]
80. Tirapelli C, Gaêta-Araujo H, Costa ED, et al. Patient perceptions of artificial intelligence in dental imaging diagnostics: a multicentre survey. Dentomaxillofac Radiol 2025 Sep 1;54(6):427-436. [doi: 10.1093/dmfr/twaf018] [Medline: 40080713]
81. Xiangde L. Investigation on patients' cognition and trust in artificial intelligence medicine. Chinese Medical Ethics 2019;32(8):986-990.
82. Menghan W, Weihua F. Influence factors of public willingness to use internet medical service platform. J Beijing Univ Aero & Astro (Social Sciences Edition) 2020;33(3):150-156. [doi: 10.13766/j.bhsk.1008-2204.2019.0001]
83. Siwen Z, Jiudi Z. Investigation on the acceptance of health education robots among patients, family members, and medical staff in esophageal cancer surgery. Modern Nurse · Special Edition 2022;29(1). [doi: 10.13766/j.bhsk.1008-2204.2019.0001]
84. Jin FS, et al. Development and acceptance of a virtual reality system for rehabilitation training of swallowing disorders in stroke patients. Chinese Journal of Modern Nursing 2023;29(18):2421-2426 [FREE Full text] [doi: 10.3760/cma.j.cn115682-20220915-04536]
85. Haggenmüller S, Maron RC, Hekler A, et al. Patients' and dermatologists' preferences in artificial intelligence-driven skin cancer diagnostics: a prospective multicentric survey study. J Am Acad Dermatol 2024 Aug;91(2):366-370. [doi: 10.1016/j.jaad.2024.04.033] [Medline: 38670313]
86. Macri CZ, Bacchi S, Wong W, Baranage D, Sivagurunathan PD, Chan WO. A pilot survey of patient perspectives on an artificial intelligence-generated presenter in a patient information video about face-down positioning after vitreoretinal surgery. Ophthalmic Res 2024;67(1):567-572. [doi: 10.1159/000541530] [Medline: 39312898]
87. Palmisciano P, Jamjoom AAB, Taylor D, Stoyanov D, Marcus HJ. Attitudes of patients and their relatives toward artificial intelligence in neurosurgery. World Neurosurg 2020 Jun;138:e627-e633. [doi: 10.1016/j.wneu.2020.03.029] [Medline: 32179185]

XSL•FO
RenderX

88. van der Zander QEW, van der Ende-van Loon MCM, Janssen JMM, et al. Artificial intelligence in (gastrointestinal) healthcare: patients' and physicians' perspectives. Sci Rep 2022 Oct 6;12(1):16779. [doi: 10.1038/s41598-022-20958-2] [Medline: 36202957]
89. Katirai A, Yamamoto BA, Kogetsu A, Kato K. Perspectives on artificial intelligence in healthcare from a Patient and Public Involvement Panel in Japan: an exploratory study. Front Digit Health 2023;5:1229308. [doi: 10.3389/fdgth.2023.1229308] [Medline: 37781456]
90. Gonzalez XT, Steger-May K, Abraham J. Just another tool in their repertoire: uncovering insights into public and patient perspectives on clinicians' use of machine learning in perioperative care. J Am Med Inform Assoc 2025 Jan 1;32(1):150-162. [doi: 10.1093/jamia/ocae257] [Medline: 39401245]
91. Witkowski K, Dougherty RB, Neely SR. Public perceptions of artificial intelligence in healthcare: ethical concerns and opportunities for patient-centered care. BMC Med Ethics 2024 Jun 22;25(1):74. [doi: 10.1186/s12910-024-01066-4] [Medline: 38909180]
92. Limayem M, Hirt SG, Cheung CMK. How habit limits the predictive power of intention: the case of information systems continuance. MIS Q 2007 Dec 1;31(4):705-737. [doi: 10.2307/25148817]
93. Brown SA, Venkatesh V. Model of adoption of technology in households: a baseline model test and extension incorporating household life cycle. MIS Q 2005 Sep 1;29(3):399-426. [doi: 10.2307/25148690]
94. Robinson R, Liday C, Lee S, et al. Artificial intelligence in health care-understanding patient information needs and designing comprehensible transparency: qualitative study. JMIR AI 2023;2:e46487. [doi: 10.2196/46487] [Medline: 38333424]
95. Change Z, Huiling Z, Huiyin Z. Study on the influence mechanism of the use of hospital intelligent medical system on patients' satisfaction: based on the perspective of technology acceptance model. Chinese Hospital Management 2019;39(10):61-64.
96. Kai W, et al. Robot-assisted gait training experience in stroke patients: A qualitative research. Military Nursing 2023;40(2) [FREE Full text] [doi: 10.3969/j.issn.2097-1826.2023.02.004]
97. Chengfu Z, Xiangzhou W. Risk challenges of embedding artificial intelligence in public service governance. E-Government ;2023(1):37-51. [doi: 10.16582/j.cnki.dzzw.2023.01.004]
98. Almokdad E, Mouloudj K, Lee CH. Rehumanizing AI-driven service: how employee presence shapes consumer perceptions in digital hospitality settings. JTAER 2025;20(3):209. [doi: 10.3390/jtaer20030209]
99. Xiong JF, Jia TY, Li XY, et al. Identifying epidermal growth factor receptor mutation status in patients with lung adenocarcinoma by three-dimensional convolutional neural networks. Br J Radiol 2018 Dec;91(1092):20180334. [doi: 10.1259/bjr.20180334] [Medline: 30059241]

Abbreviations

AI: artificial intelligence
BCT: behavior change technique
MMAT: Mixed Methods Appraisal Tool
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Searching
TDF: Theoretical Domains Framework
UTAUT2: Unified Theory of Acceptance and Use of Technology 2

Edited by S Brini; submitted 13.Jul.2025; peer-reviewed by G Paganin, K Mouloudj; revised version received 10.Mar.2026; accepted 10.Mar.2026; published 08.May.2026.

Please cite as:
Shi H, Huang J, Yang J, Han M
Barriers and Facilitators to Patient Acceptance of Artificial Intelligence in Health Care: Systematic Review
J Med Internet Res 2026;28:e80581
URL: https://www.jmir.org/2026/1/e80581
doi:10.2196/80581

© Huiqin Shi, Jingying Huang, Jin Yang, Mengbo Han. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Efficacy of Technology-Based Interventions on the Reduction of Loneliness: Systematic Review and Meta-Analysis

Zdenek Meier, PhD; Marie Buchtova, PhD; Jan Sandora, MA; Lukas Novak, PhD; Jakub Helvich, PhD; Ondrej Buchta, MA; Jana Furstova, PhD; Klara Malinakova, PhD; Peter Tavel, DProf

Olomouc University Social Health Institute, Palacký University Olomouc, Univerzitni 244/22, Olomouc, Czech Republic

Corresponding Author:
Zdenek Meier, PhD
Olomouc University Social Health Institute, Palacký University Olomouc, Univerzitni 244/22, Olomouc, Czech Republic

Abstract

Background: Loneliness is a widespread public health concern linked to increased risks of health problems. As populations age, the demand for effective interventions to mitigate loneliness continues to grow.

Objective: This meta-analysis aimed to examine the effectiveness of technology-based interventions in reducing loneliness.

Methods: A systematic literature search was conducted in Web of Science, PsycInfo, PubMed, Scopus, Google Scholar, Embase, and the Cochrane Library in August 2024. We included randomized controlled trials that examined the effectiveness of technology-based interventions compared with any control group in reducing loneliness across all age groups. Nonrandomized studies, qualitative research, and studies lacking sufficient statistical data for effect size calculation were excluded. After screening 1089 records, 7 studies involving 580 participants met the inclusion criteria. Data were extracted by 3 independent reviewers, with discrepancies resolved by a fourth reviewer. The risk of bias was assessed using the Cochrane risk-of-bias tool. A random-effects model was used to synthesize effect estimates, with standardized mean differences as the primary effect size metric. Heterogeneity was assessed using the Q statistic and I^2 index, and a prediction interval was calculated to estimate the expected range of true effects.

Results: We found a small and statistically nonsignificant overall effect of technology-based interventions on loneliness (pooled standardized mean difference=–0.21, 95% CI –0.59 to 0.17; 95% prediction interval –1.14 to 0.63). Substantial between-study variability was present (I^2=57%; τ^2=0.08; τ=0.28), and the prediction interval indicated that true effects in future studies may range from substantial reductions to moderate increases in loneliness. Differences between intervention types could not be examined due to the limited number of eligible studies. The Egger test showed significant funnel plot asymmetry (t_5=–5.47; P=.003). However, since the trim-and-fill method identified no missing studies, the asymmetry is unlikely to be fully explained by publication bias.

Conclusions: Unlike previous reviews, which focused primarily on older populations or included nonrandomized studies, this meta-analysis provides a rigorous synthesis of only randomized controlled trials across the lifespan. Our findings do not provide evidence for a reliable reduction in loneliness following technology-based interventions. Moderate heterogeneity indicates that effects differ across studies, and the limited number of eligible trials prevented the analysis of potential moderators. By incorporating robust prediction intervals and a broader age demographic, this study offers a more comprehensive view of the variability of intervention outcomes compared with existing literature. The wide prediction interval suggests that intervention effects may vary across settings. In practice, technology-based approaches should be viewed as supportive tools that require careful tailoring rather than universal solutions to loneliness.

Trial Registration: PROSPERO CRD42024505117; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024505117

(J Med Internet Res 2026;28:e80059) doi:10.2196/80059

KEYWORDS

loneliness; technology; intervention; meta-analysis; systematic review

Introduction

Negative Impacts of Loneliness

Loneliness is a growing public health challenge linked to a wide range of adverse mental and physical health outcomes. It has gained considerable attention in recent years due to its significant impact on both mental and somatic health. It is a major risk factor for various physical health problems, influencing biological responses to stress [1,2], glucose regulation [3], cardiovascular risk [4,5], and overall health behaviors [6]. Furthermore, loneliness is a critical risk factor for mental health, as individuals experiencing loneliness exhibit

XSL•FO
RenderX

higher levels of psychological distress, depression, and poorer subjective health assessments [7]. It is associated with an increased prevalence of paranoid symptoms [8], depression [9,10], anxiety [10,11], and psychotic symptoms [8], as well as lower self-esteem and self-evaluation [12,13]. A study by Liang et al [14] identified loneliness as a risk factor for more than two-thirds of the analyzed diseases and for 13 out of 14 disease categories, particularly mental and behavioral disorders, infectious diseases, and conditions affecting the nervous, respiratory, and endocrine systems [14]. Lastly, loneliness is associated with an increased risk of mortality [15,16].

Loneliness encompasses both adaptive and maladaptive aspects. In the cognitive model, short-term loneliness can function as a social mechanism, prompting individuals to strengthen interpersonal bonds and seek social connections [17]. This adaptive process can be considered an evolutionary response to the fundamental human need for belonging [18]. However, chronic loneliness leads to negative consequences affecting both physical and mental health [17,19,20]. The model of loneliness as an adaptive mechanism with maladaptive consequences is essential not only for understanding its biological function but also for recognizing the risks associated with its long-term presence in an individual's life. Loneliness can be helpful in the short term, but chronic loneliness leads to poorer health and reduced life satisfaction [21].

Loneliness has widespread societal impacts, including economic consequences such as increased health care costs and reduced work productivity [22]. Effective solutions to loneliness lie in future research and interventions aimed at preventing its occurrence and mitigating its negative effects on both individuals and society [23-25]. As a significant public health issue, loneliness requires a multidisciplinary approach involving collaboration among experts in health care, psychology, and social sciences [24,26,27]. Key strategies for addressing loneliness are particularly found in social work and social policies, which can help reduce risk factors and promote social inclusion [26,28].

Definition

From a multidimensional perspective, loneliness can be defined as a significant psychological and social phenomenon, primarily conceptualized in the literature as the subjective perception of a discrepancy between expected and actual social relationships [29]. This discrepancy is not necessarily objectively measurable by the number of social interactions but rather by the qualitative evaluation of one's relationships [30]. Moreover, loneliness is a complex phenomenon encompassing both affective and cognitive components. The affective component relates to the emotional distress associated with feelings of isolation, whereas the cognitive component reflects the perceived inadequacy of social relationships [21].

Weiss [31] distinguished between social loneliness, which is related to a lack of social networks and a sense of social exclusion, and emotional loneliness, which is associated with the absence of close and meaningful relationships [31]. In addition to these, another critical dimension is existential loneliness, characterized by a profound sense of isolation linked to questions of life's meaning, human finiteness, and the subjective incommunicability of individual experience [32].

Based on these observations, loneliness can be explored according to its 3 dimensions: intimate loneliness, which pertains to the lack of a deep relationship with a close person; relational loneliness, which arises from the absence of quality friendships or family connections; and collective loneliness, which stems from a missing sense of belonging to a group or a broader community [33-35]. The concept of collective loneliness remains a subject of debate and requires further investigation and clarification [36].

When examining the relationship between loneliness and the quality and quantity of social relationships, particularly among older adults, it is essential to consider not only quantitative aspects, such as the number of friends and frequency of interactions, but also the perception of social support and emotional closeness, which determine the quality of relationships [30].

The Prevalence of Loneliness

Loneliness is highly prevalent across the lifespan and in diverse sociodemographic groups [7-9]. Population-based studies and meta-analyses indicate that roughly 1 quarter to one third of adolescents and older adults report at least moderate levels of loneliness, with substantial variation between countries and regions [37-48]. Although some longitudinal studies suggest a U-shaped or more complex trajectory, with elevated loneliness particularly during adolescence and in later life [37,49-51], these patterns appear to be driven by life transitions and changes in social roles rather than age per se [40,41,49]. Rates of loneliness increased further during the COVID-19 pandemic, when large-scale public health measures such as social distancing and lockdowns disrupted everyday social contact [52-56].

Chronic loneliness is especially concerning; estimates suggest that about 1 in 5 older adults experience persistent loneliness, which is most strongly associated with adverse physical and mental health outcomes [47,57]. Against this backdrop, there is an urgent need for scalable, evidence-based interventions that can be implemented across different age groups and settings.

Technology-Based Interventions for Loneliness

In recent years, digital technology has emerged as a promising avenue for delivering loneliness interventions. Technology-based interventions are approaches that use digital or communication technologies as the primary mode of delivery—for example, web-based programs, mobile apps, videoconferencing platforms, or socially assistive robots designed to enhance social connection or provide psychosocial support [58]. These interventions can extend the reach of traditional face-to-face programs, particularly for individuals who are homebound, geographically isolated, or affected by situations such as the COVID-19 pandemic, where physical distancing measures increased both loneliness and the need for remote support [58,59]. By enabling people to connect with peers, family members, or professional services via digital means, they offer a potentially scalable way to address loneliness [59,60]. However, the existing evidence base remains limited

XSL•FO
RenderX

and heterogeneous, with many small trials of variable quality and relatively few rigorous randomized controlled studies [58].

Existing systematic reviews of loneliness interventions have often focused on traditional face-to-face programs or on specific subgroups such as older adults, and several recent reviews of digital interventions report small or nonsignificant effects, substantial heterogeneity, and methodological limitations. As new trials using diverse technologies and age groups have accumulated, an updated and methodologically rigorous synthesis is needed to clarify under what conditions technology-based interventions can meaningfully reduce loneliness.

Thus, this study aims to conduct a systematic review and meta-analysis to evaluate the effectiveness of technology-based interventions compared with control conditions in reducing loneliness among individuals across the lifespan. We focus on interventions in which digital or communication technologies constitute the main delivery channel (eg, mobile apps, web-based self-help programs, videoconferencing, socially assistive robots). Specifically, we address the following research questions:

1. To what extent do technology-based interventions reduce loneliness compared with control conditions?
2. Do the effect sizes differ between distinct types of technology-based interventions (eg, app-based psychological programs, socially interactive technologies, videoconferencing-based approaches)?
3. Are intervention or sample characteristics such as participant age, intervention duration, and follow-up period associated with differences in effectiveness?

Clarifying these questions is particularly important in the context of population aging and growing shortages in health and social care workforces, where scalable digital solutions may help to complement, but not replace, traditional forms of support.

Methods

Research Design

The protocol for this study was registered with the PROSPERO (International Prospective Register of Systematic Reviews) under the registration number CRD42024505117 [61]. The systematic review and meta-analysis were conducted and reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [62,63], using the PRISMA 2020 expanded checklist to ensure comprehensive reporting. The literature search was carried out following the PRISMA-S extension for reporting search strategies [64], and the abstract was structured according to the PRISMA 2020 for Abstracts guidelines. Fully completed versions of all relevant PRISMA checklists are provided in Checklist 1.

Eligibility Criteria

To ensure methodological rigor, this meta-analysis included studies meeting the following criteria: (1) no restrictions on publication language; (2) no restrictions regarding the age of participants; (3) empirical studies using a randomized controlled trial (RCT) design; (4) use of a technology-based intervention compared with a control group; and (5) reporting of loneliness as a measured outcome (either as a primary or secondary end point), with available loneliness levels before and after the intervention or sufficient data to calculate the effect size. Furthermore, no restrictions were placed on the study setting (eg, community-dwelling or long-term care facilities) or the minimum duration of the intervention. Regarding the definition of loneliness, studies were eligible if they operationalized the construct consistent with the conceptual framework described in the *Introduction* section, assessing subjective feelings of loneliness or social isolation using quantitative measures. Studies that did not meet all of these criteria, including nonempirical studies, non-RCT designs, interventions unrelated to technology, and those lacking relevant outcome measures, were excluded.

Search Strategy

The literature search for this meta-analysis was conducted in October 2023, with a final update performed on August 21, 2024. The following databases were systematically searched from their respective inception dates: Web of Science (Clarivate), PsycInfo (EBSCOhost), PubMed (NCBI), Scopus (Elsevier), Google Scholar, Embase (Elsevier), and the Cochrane Library (Wiley). To identify relevant studies, search terms were organized into four thematic groups: (1) loneliness (eg, “loneliness,” “perceived social isolation”); (2) intervention (eg, “intervention,” “treatment”); (3) technology (eg, “technology,” “computer,” “mobile,” “app*”); and (4) study design (eg, “Randomized Controlled Trial,” “RCT”). Searches were conducted in the fields of Title, Abstract, and Keywords, except in Google Scholar, which did not support specific operators for limiting searches to these fields. Boolean operators and truncation symbols were used to ensure a comprehensive and systematic identification of studies relevant to technology-based interventions for reducing loneliness. The exact search syntaxes applied to each database are detailed in Multimedia Appendix 1. In addition to database searches, further studies were identified through citation searching in the reference lists of relevant review articles on the topic (see details below).

Selection Process

All retrieved citations were imported into Zotero reference management software [65]. Duplicate records were identified and removed using the software’s automated functions, followed by manual verification. Titles and abstracts underwent an initial screening to determine relevance. First, records that did not report published results of primary empirical research were excluded. This included books, study protocols, preprints, review articles (including meta-analyses), theoretical studies, conference contributions, master’s theses, commentaries, editorials, letters to the editor, and studies with inaccessible full texts. Initially, we also intended to exclude dissertation theses. However, we decided to retain them in the search, as we consider the dissertation review process to be sufficiently rigorous. Second, studies that were thematically irrelevant were excluded. This applied to records that did not focus on loneliness (eg, loneliness was not measured as an outcome) or did not examine the impact of technology-based approaches (eg, the intervention was not

XSL•FO
RenderX

technology-driven, or the study did not compare nontechnology vs technology-based methods). Third, studies that did not follow an RCT design were excluded. This included, for example, studies assessing the feasibility of RCTs, those using qualitative designs, or other nonexperimental approaches.

Thematically relevant reviews were used for further citation searching to help identify additional empirical studies not captured in the initial database search. The full texts of studies were reviewed based on predefined inclusion and exclusion criteria. The selection process was conducted by 3 reviewers. Each record and report was screened independently by at least 2 reviewers. Any discrepancies were resolved through consultation with a fourth reviewer. All stages of the study selection process were conducted manually; no automation tools or machine learning classifiers were used.

There were minor deviations from the original PROSPERO protocol. First, the literature search was updated in August 2024 to include the most recent evidence, extending the search period beyond the date initially planned. Second, while the protocol mentioned a dedicated search for gray literature, the final strategy relied on comprehensive database searches and citation tracking of relevant reviews. Finally, the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) approach and the calculation of prediction intervals were added during the analysis stage to provide a more robust assessment of evidence certainty and effect distribution.

Data Collection Process

A predefined data extraction form was used to collect the following information: (author, year, country), participant demographics (sample size, characteristics, age range, mean age, and gender distribution), baseline loneliness levels, measurement instruments, and intervention details (type of intervention and control, duration, and the timing of postintervention assessment). For all included studies, loneliness data were extracted using the specific quantitative measure provided by the authors. In cases where studies reported results from multiple postintervention time points, data from the assessment conducted immediately following the completion of the intervention were extracted to ensure consistency. Data extraction was conducted by 3 reviewers, with each study independently coded by 2 reviewers. Any discrepancies were resolved by involving a fourth reviewer. The process was performed manually, and no automation tools or machine learning applications were used. All relevant data were obtained through a comprehensive review of the text, tables, and figures of each report; no data extraction software for figures was required. As all included studies were published in English, no formal translation process was necessary. Only studies reporting sufficient statistical information to compute pre-post effect sizes were included. Thus, according to this criterion, 8 studies were excluded. We chose not to contact study authors for missing data because of practical constraints and the likelihood of delayed or incomplete responses. To ensure transparency and replicability, we decided that only studies with fully reported data necessary for effect size computation (including pre-post correlations) were included. This approach allows the meta-analysis to rely only on verifiable and publicly available information.

Data Synthesis

A random-effects model was used to synthesize effect estimates across studies, accounting for between-study heterogeneity. Effect sizes and their sampling variances were calculated as standardized mean differences (SMDs), specifically the difference between standardized mean changes in the intervention and control groups. This metric was chosen to accurately account for the pre-post study design and to enable the comparison of results from studies using different loneliness measurement scales [66]. The magnitude of the pooled effect size was interpreted according to Cohen [67] conventional criteria: 0.2 (small), 0.5 (medium), and 0.8 (large). All computations were performed in R using the metafor package. The summary effect size and its 95% CI were computed to estimate the overall intervention effect. A random-effects model was fitted using restricted maximum likelihood, and the Hartung-Knapp-Sidik-Jonkman [68] adjustment was applied to obtain robust confidence intervals given the small number of studies (k=7). Additionally, a prediction interval was calculated using the bootstrap confidence distribution approach proposed by Nagashima et al [69] to define the expected range of true effects in comparable populations and future studies, as the prediction interval quantifies the range within which the true effect sizes are expected to fall, providing a clearer interpretation of effect variation. Heterogeneity was assessed using the Q statistic and I^2. The Q test was performed to assess the null hypothesis of homogeneity (ie, that all studies shared a common effect size), with a significance threshold of α=.10 to account for its typically low statistical power. I^2 quantified the proportion of observed variance attributable to true heterogeneity rather than random error. However, as the I^2 index does not describe the magnitude of variation in effect sizes, we place primary emphasis on the prediction interval to interpret the distribution of true effects across different populations and settings. Given that estimates of heterogeneity based on fewer than 10 studies may be unreliable, the results should be interpreted with caution. Subgroup analysis and meta-regression were not conducted because the number of available studies was insufficient for reliable moderator testing. Egger regression test and Duval and Tweedie trim and fill procedure were conducted to examine potential dissemination bias. Potential publication bias was assessed within the same random-effects framework in R. To assess the robustness of the findings, a leave-one-out sensitivity analysis was conducted, in which the overall effect was recalculated after omitting each study in turn.

The risk of bias in the included studies was assessed using the revised Cochrane risk-of-bias 2 tool for randomized trials [70]. Three reviewers participated in the assessment process. Each study was evaluated across five domains: (1) bias arising from the randomization process, (2) bias due to deviations from intended interventions, (3) bias due to missing outcome data, (4) bias in measurement of the outcome, and (5) bias in selection of the reported result. Any discrepancies were resolved through joint discussion until a consensus was reached. The overall risk-of-bias judgment for each study followed the standard Cochrane risk-of-bias 2 algorithm, where the highest risk level

XSL•FO
RenderX

identified in any individual domain determined the overall rating. The visualization of the risk of bias assessment was produced using the *robvis* tool [71].

The certainty of evidence for the primary outcome (loneliness) was assessed using the GRADE approach [72]. Two reviewers independently performed the assessment, with any discrepancies resolved through discussion until a consensus was reached. We evaluated 5 domains: risk of bias, inconsistency, indirectness, imprecision, and publication bias. The certainty of evidence was graded as high, moderate, low, or very low, reflecting the degree of confidence that the true effect lies close to the estimate. The findings were summarized following the standard GRADE Summary of Findings table format.

Results

Study Selection

The database search yielded 1089 records. After screening titles and abstracts, full-text reports were assessed for eligibility. Detailed numbers of excluded records at each stage are presented in the PRISMA flow diagram (Figure 1). A total of 57 studies were excluded, primarily due to incomplete or insufficient data. A full list of excluded studies with reasons is provided in Multimedia Appendix 2. Notably, the study by Sun [73], which was initially considered, was excluded due to its retraction by the publisher [74].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart illustrating the study selection process for the systematic review and meta-analysis of technology-based interventions for loneliness (literature search conducted up to August 2024). RCT: randomized controlled trial.

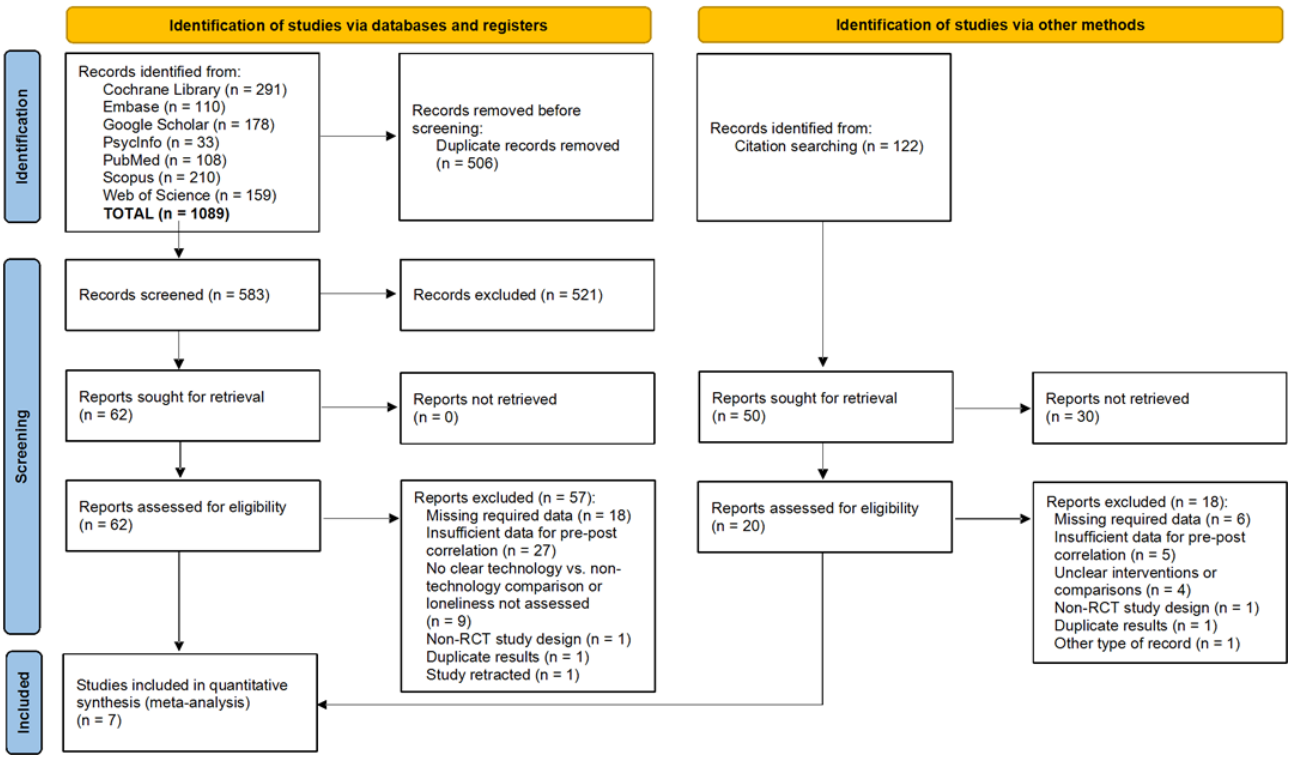

Additional studies were sought through citation searching of thematically relevant reviews (n=32). This search yielded 122 records, of which 72 duplicates were removed. Following the assessment of the remaining reports, most were excluded for reasons similar to the database search results, ultimately leaving 2 studies for inclusion (Figure 1). Moreover, to ensure the comprehensiveness of our study selection, we compared our included studies with those identified in the meta-analysis by Jin et al [75], which examined the effectiveness of technology-based interventions for reducing loneliness in older adults. Given the overlap in scope, we reviewed all 6 studies included in their meta-analysis. However, none of them met our strict inclusion criteria. Notably, although the study by Shapira et al [76] was classified as an RCT in the review by Jin et al [75], our assessment revealed that the allocation was based on participants' willingness and interest rather than randomization. Therefore, it was excluded as a non-RCT. Four other studies [77-80] were excluded due to lacking essential source data or statistical indicators needed for calculating pre-post correlations. Similarly, the study of Slegers et al [81] lacked a clearly defined comparison between a technology-based and a nontechnology-based intervention.

Importantly, the study by Schwindenhammer [82], identified through citation searching, was originally described by the author as having a quasi-experimental design. However, a detailed examination of the methodology revealed that the assignment of nursing homes to intervention or control groups was determined by a random drawing. Consequently, this study qualifies as a cluster-RCT rather than a nonrandomized quasi-experiment. Since the study satisfied the randomization requirement and provided the necessary statistical indicators for calculating pre-post correlations, it was included in the final analysis. Ultimately, 7 studies were included in the quantitative synthesis.

XSL•FO
RenderX

Study Characteristics

Tables 1 and 2 summarize the characteristics of the 7 included studies, including participant demographics, intervention details, descriptive statistics, and methodological features. The studies included in the meta-analysis were published between 2006 and 2024 and were conducted in the United States (n=3), Switzerland (n=1), Poland (n=1), New Zealand (n=1), and Israel (n=1). The target populations were diverse, ranging from high school students [83] and young adults [84] to adults [85], chronically ill rural women [86], and older adults residing in the community or nursing homes [82,87,88]. Participant ages ranged from 14 to 100 years.

Table . Characteristics of included studies: participant demographics and settings.

Study	Country	Sample characteristics	Sample N (IG[a]/CG[b])	Age range (y)	Age, mean (SD)	Gender (% female)
Brog et al [85]	Switzerland	Adults with at least mild depressive symptoms	Randomized: 107; analyzed: 95	18 - 81	40.36 (14.59)	81.3
Hill et al [86]	United States	Chronically ill rural women	Randomized: 120; analyzed: 100	35 - 65	N/A	100
Iyer et al [83]	United States	High school students	Randomized: 110; analyzed: 108	14 - 19	16.05 (1.34)[c]	81.5
Karkosz et al [84]	Poland	Young adults with subclinical depression or anxiety	Randomized: 81; analyzed: 81	18 - 35	25.7 (4.59)[c]	71.6
Robinson et al [87]	New Zealand	Older adults in hospital and rest home care	Randomized: 40; analyzed: 34	55 - 100	N/A[d]	67.5
Shapira et al [88]	Israel	Older community-dwelling adults	Randomized: 86; analyzed: 82	65 - 90	72 (5.63)	80.5
Schwindenhammer [82]	United States	Older adults in nursing homes	Cluster-randomized: 5 nursing homes (83 residents); analyzed: 80	71 - 97	86.12 (5.28)	87.5

[a]IG: intervention group.

[b]CG: control group.

[c]Computed from the dataset available, not provided in the study itself.

[d]Not applicable.

XSL•FO
RenderX

Table . Characteristics of included studies: intervention details and descriptive outcomes.

Study	Baseline		Postintervention		Measurement instrument	Intervention type	Control type	Intervention duration (wk)	Difference between baseline and follow-up measurement
	N	Mean (SD)	N	Mean (SD)					
Brog et al [85]	IG[a]: 53; CG[b]: 54	IG: 21.26 (4.82); CG: 20.37 (4.25)	IG: 43; CG: 52	IG: 19.88 (4.56); CG: 20.27 (4.04)	UCLA[c] (9-item)	Internet-based self-help program for psychological distress (CBT)[d]	Wait-list control (care as usual)	3	3 wk
Hill et al [86]	IG: 43; CG: 57	IG: 45.73 (9.99); CG: 43.97 (10.53)	IG: 43; CG: 57	IG: 43.15 (8.69); CG: 43.11 (10.94)	UCLA (20-item)	Online, asynchronous, peer-led support group and health teaching units	No intervention (usual care)	22	3 mo
Iyer et al [83]	IG: 54; CG: 54	IG: 46.54 (10.31); CG: 43.93 (10.97)	IG: 54; CG: 54	IG: 38.07 (8.44); CG: 42.89 (11.49)	UCLA (20-item)	Online, self-care program	Wait-list control	4	4 wk
Karkosz et al [84]	IG: 40; CG: 41	IG: 45.44 (12.15); CG: 42.39 (13.25)	IG: 40; CG: 41	IG: 43.06 (13.38); CG: 41.49 (13.42)	UCLA (20-item)	Web-based and mobile therapy chatbot (CBT)	Active control (self-help book)	2	2 wk
Robinson et al [87]	IG: 17; CG: 17	IG: 36.44 (9.76); CG: 31.71 (9.50)	IG: 17; CG: 17	IG: 32.23 (9.92); CG: 33.93 (8.52)	UCLA (20-item)	Companion robot	Active control (standard activities)	12	12 wk
Shapira et al [88]	IG: 68; CG: 18	IG: 5.43 (2.0); CG: 6.11 (1.9)	IG: 64; CG: 18	IG: 4.80 (1.70); CG: 6.40 (1.70)	UCLA (3-item)	Online guided group sessions via Zoom (CBT)	Wait-list control	3.5	3.5 wk
Schwindenhammer [82]	IG: 40; CG: 40	IG: 17.44 (4.63); CG: 16.45 (4.43)	IG: 40; CG: 40	IG: 14.53 (4.23); CG: 14.13 (3.88)	UCLA (10-item)	Videoconferencing sessions	No intervention (usual care)	14	14 wk

[a]IG: intervention group.

[b] CG: control group.

[c]UCLA: University of California, Los Angeles Loneliness Scale.

[d]CBT: cognitive behavioral therapy.

The interventions were categorized into 3 distinct types based on their primary mechanism of action and content: (1) digital psychological interventions (n=4), which used technology to deliver structured therapeutic content (eg, cognitive behavioral therapy [CBT], mindfulness) or skills training. This category included a web-based self-help program [85], an online self-care program [83], a therapy chatbot [84], and online guided group CBT sessions [88]. (2) Information and Communication Technology for social connection (n=2), which focused on facilitating communication with others. This included videoconferencing with family members [82] and an online peer-support group [86]. (3) Social robots (n=1), represented by a companion seal robot designed to provide companionship [87].

The duration of interventions varied considerably, ranging from 2 weeks [84] to 22 weeks [86]. All studies assessed loneliness using the University of California, Los Angeles Loneliness Scale, though different versions were used (eg, 3-item, 9-item, 10-item, or 20-item versions). Baseline loneliness scores indicated varying levels of loneliness across samples, from low to moderate to high.

Results of Syntheses

The meta-analysis demonstrated a small and statistically nonsignificant overall effect of interventions on loneliness (SMD=−0.21, 95% CI −0.59 to 0.17; P=.22), indicating that the interventions did not produce a reliable reduction in loneliness compared with control conditions (Table 3). The

XSL•FO
RenderX

random-effects model was used to account for between-study heterogeneity, which was moderate (Q=13.68, df=6; P=.03; I^2=57.44%, τ^2=0.076; τ=0.276), suggesting that a substantial proportion of the variability reflects true differences among studies rather than sampling error.

Table . Summary of random-effects meta-analysis results, showing standardized mean differences and statistical indicators for loneliness reduction in technology-based interventions vs control groups.

Study	Standardized mean difference	SE	Variance	95% CI	*Z* value	*P* value
Brog et al [85]	−0.019	0.208	0.043	−0.428 to 0.389	−0.093	.93
Hill et al [86]	−0.049	0.209	0.043	−0.457 to 0.360	−0.234	.82
Iyer et al [83]	0	0.212	0.045	−0.416 to 0.416	0.000	>.99
Karkosz et al [84]	−0.003	0.227	0.051	−0.448 to 0.441	−0.015	.99
Robinson et al [87]	−1.047	0.366	0.134	−1.764 to −0.330	−2.863	.004
Schwindenhammer [82]	0	0.231	0.053	−0.453 to 0.453	0.000	>.99
Shapira et al [88]	−0.824	0.282	0.080	−1.378 to −0.271	−2.919	.004

The estimated prediction interval ranges from –1.14 to 0.63, indicating that true effects in comparable future studies may vary from modest reductions to modest increases in loneliness (Figure 2). Given that the analysis included only 7 studies, heterogeneity estimates and the prediction interval should be interpreted with caution. Further research is needed to better identify conditions under which loneliness interventions are effective.

Figure 2. Forest plot showing study-level standardized mean differences and 95% CIs for the effect of technology-based interventions on loneliness, synthesized using a Hartung-Knapp-Sidik-Jonkman adjusted random-effects model [82-88]. HKSJ: Hartung-Knapp-Sidik-Jonkman method; I^2: *I*-squared heterogeneity statistic; Q: Cochran Q statistic; REML: restricted maximum-likelihood; SMD: standardized mean difference.

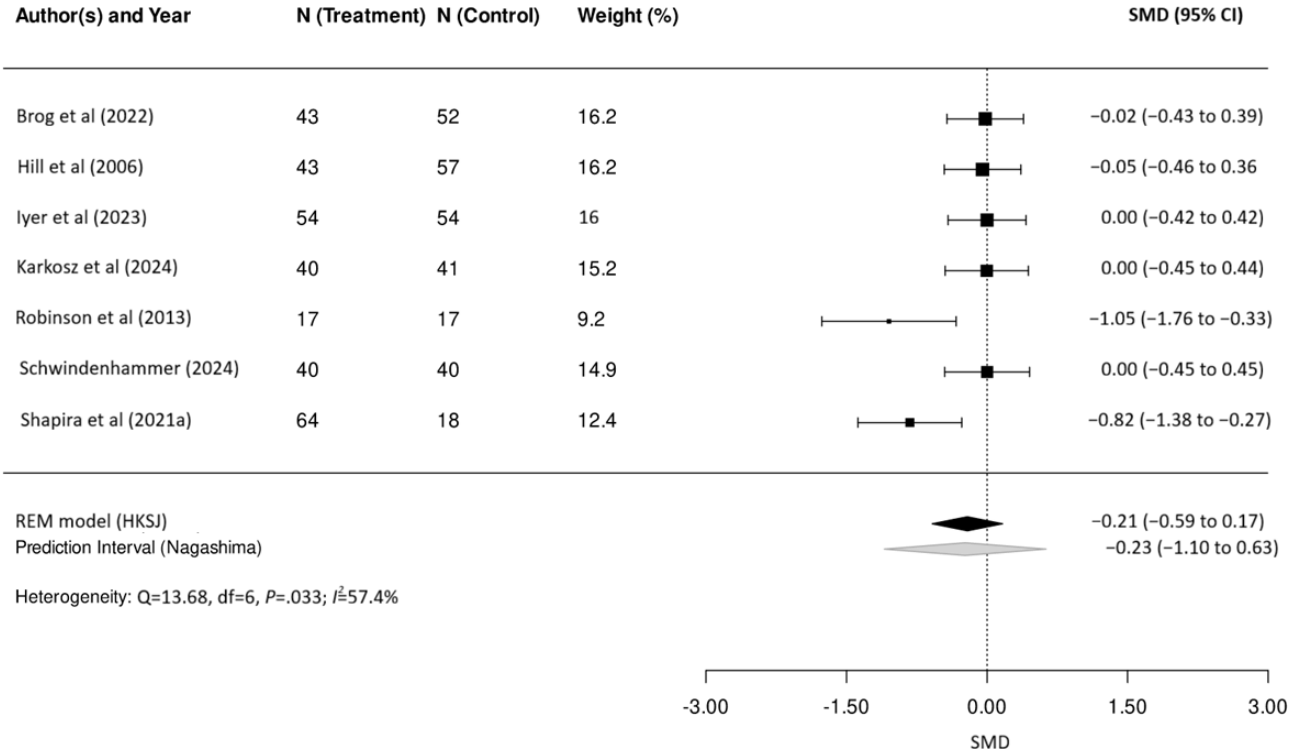

Risk of Bias

The summary of the risk of bias assessment is presented in Figure 3. Among the 7 included studies, 2 were judged to be at low risk of bias [85,88], 3 studies raised some concerns [83,84,87], and 2 studies were assessed as being at high risk of bias [82,86]. The high risk was primarily driven by domain 3 (missing outcome data) due to substantial differential attrition between the intervention and control groups [86] and domain 4 (measurement of the outcome) due to the lack of blinding of the outcome assessor during data collection [82]. Most studies raised “some concerns” in domain 2 (deviations from intended interventions) and domain 4. This reflects the inherent nature of psychosocial and technological interventions, where blinding of participants is often not feasible, and the reliance on self-reported measures may introduce bias if the participants are aware of their group assignment.

XSL•FO
RenderX

Figure 3. Risk of bias summary: review authors' judgments about each risk of bias item for each included study [82-88].

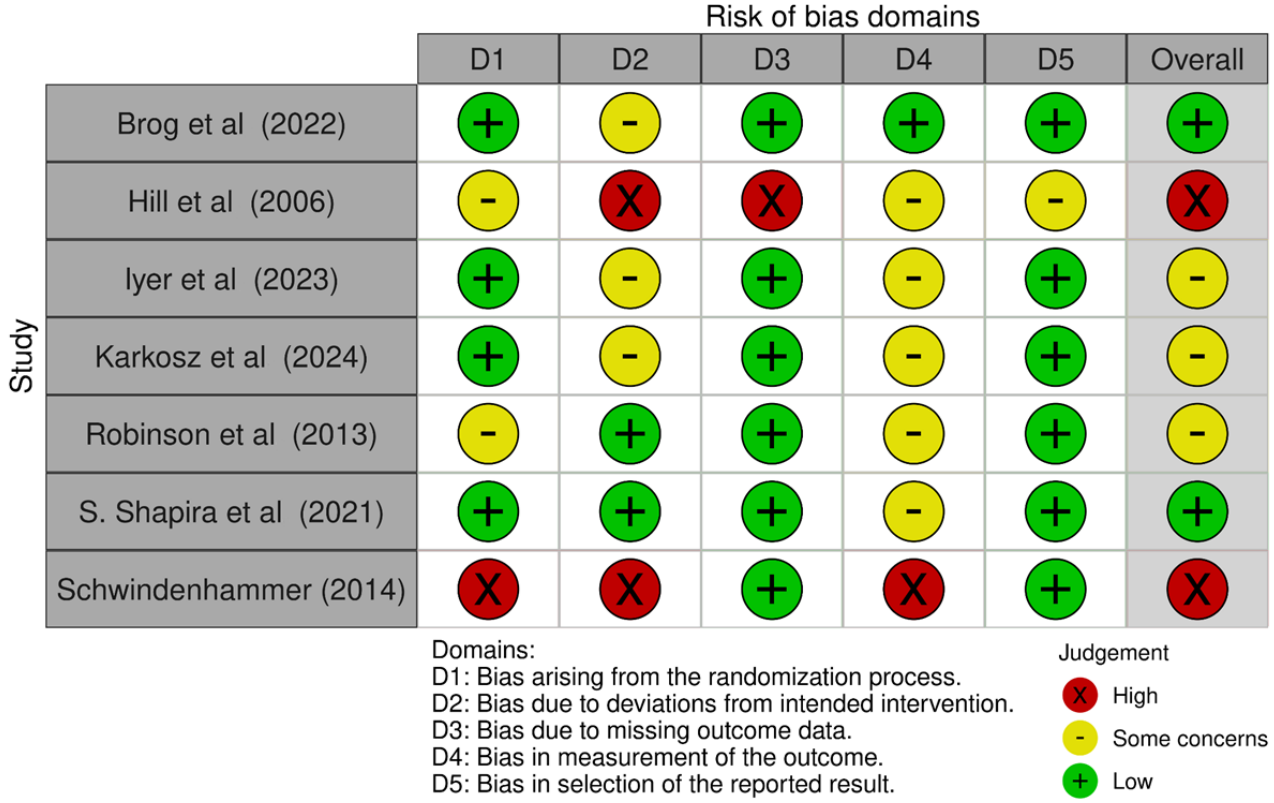

Reporting Biases

Visual inspection of the funnel plot and statistical testing were used to evaluate reporting bias. Egger regression test [89] showed statistically significant funnel plot asymmetry (t_5=–5.47, P=.003), indicating the presence of small-study effects (see Funnel Plot in Multimedia Appendix 3). However, Duval and Tweedie trim and fill method [90] did not identify any missing studies (0 missing studies), and no adjusted pooled effect was required. Accordingly, although funnel plot asymmetry was detected, the absence of imputed studies suggests that the asymmetry is likely driven by study heterogeneity (as indicated by the I^2 and wide prediction interval) rather than publication bias alone.

Sensitivity Analysis

A sensitivity analysis was conducted using a leave-one-out procedure in R, in which the pooled effect was recalculated after omitting each study in turn. The results showed that the pooled SMD remained small and nonsignificant across all iterations, and the direction of the effect did not change, indicating that no single study unduly influenced the overall findings.

Certainty of Evidence

The overall certainty of the evidence regarding the effect of technology-based interventions on loneliness was graded as very low. We downgraded the evidence by 1 level due to the risk of bias, as only 2 of the 7 included studies were rated as having a low risk of bias. We further downgraded the evidence due to inconsistencies caused by moderate heterogeneity and a wide prediction interval. Finally, we downgraded the rating due to imprecision, as the 95% CI for the pooled effect included both a noticeable benefit and no effect (crossing the 0 line). No publication bias was detected, but given the small number of studies, it could not be completely ruled out. A detailed summary of the findings and certainty assessment is provided in Multimedia Appendix 4.

Discussion

Principal Findings

The primary objective of this meta-analysis was to evaluate the effectiveness of technology-based interventions in reducing loneliness. The results showed a small and statistically nonsignificant overall effect, indicating that the available evidence does not support a reliable reduction in loneliness attributable to technology-based interventions. Importantly, these findings must be interpreted with caution, as the certainty of the evidence was graded as very low. This low confidence stems primarily from limitations in the risk of bias domain and the observed inconsistency across studies. However, the moderate heterogeneity observed in the studies suggests that the effectiveness of these interventions varies across studies. The prediction interval indicated that true effects in comparable future studies could range from substantial reductions to moderate increases in loneliness, underscoring the influence of study-specific and contextual factors.

XSL•FO
RenderX

Because the number of available studies (k=7) did not permit reliable subgroup or moderator analyses, the present meta-analysis could not empirically evaluate how different types of technology-based interventions, such as digital psychological interventions or information and communication technology for social connection, vary in their effectiveness. However, previous research provides important insights into potential sources of variability. For example, some structured approaches, such as telephone-delivered behavioral activation or mindfulness interventions, have demonstrated beneficial effects compared with less structured social-contact interventions [89], where a telephone-based behavioral activation or mindfulness intervention is effective compared with a control group that receives only a befriending intervention [89]. Moreover, classifying interventions by mechanism of action rather than just delivery mode offers insights for future research. For example, digital psychological interventions aiming to restructure maladaptive social cognition (eg, via CBT) have previously shown strong effects in face-to-face settings [28]. In our review, studies in this category [84,85,88] showed mixed results, potentially due to variations in guidance intensity. Research suggests that loneliness is a deeply personal experience shaped by social, cultural, and psychological factors [8,9]. As a result, intervention effectiveness is likely to depend on participant characteristics such as baseline loneliness, age, social resources, and psychological needs, although these factors could not be tested in the present analysis. Prior work suggests that digital mental health interventions are more effective when tailored to the needs and preferences of individuals [91], underscoring the importance of personalization rather than a uniform intervention model.

Despite the widespread interest in technology-based interventions, it is important to acknowledge the moderate proportion of heterogeneity in effects across studies. We observed an I^2 of 57%, indicating that more than half of the observed variance in effect sizes is due to true differences between studies rather than sampling error. This finding is consistent with prior observations that loneliness intervention outcomes are highly context-dependent [92-94]. However, the I^2 statistic does not reveal by how much the true effect varies across different populations. The prediction interval, in contrast, provides a direct estimate of this range. Previous reviews of loneliness interventions have also frequently reported substantial heterogeneity [28,95,96]. This variability may be due to differences in study design, with pre-post studies and nonrandomized comparative studies showing larger average effect sizes compared to randomized controlled studies [28]. The sources of heterogeneity could not be examined empirically in the present meta-analysis due to the small number of included studies; however, prior research suggests that study design, participant age, and digital literacy may influence intervention outcomes [97]. For example, younger participants often benefit more from digital programs that incorporate interactive or socially engaging elements, whereas older adults may experience additional barriers such as lower digital literacy or chronic social isolation that limit intervention impact [91,98-100]. Systematic reviews focusing specifically on older adults have reported minimal or no overall effects of digital interventions [75,101], underscoring the importance of population characteristics. Previous literature also indicates that intervention type, delivery format, and follow-up duration may influence loneliness outcomes [102-104], suggesting that program content and user engagement processes play important roles. Additionally, effects may vary by timing of assessment, with short-term improvements sometimes diminishing over longer follow-up periods [101,104]. Lastly, differences in loneliness measurement instruments and study quality cannot be overlooked. Most studies in the broader literature use the University of California, Los Angeles Loneliness Scale, but other instruments, including single-item measures, are also used [105,106]. Meta-analytic evidence indicates that effect sizes can differ by instrument, partly due to scale sensitivity and range [107-110]. The use of a random-effects model helps account for some of this variability; however, the prediction interval we calculated in the present study was wide. This highlights a critical distinction: while the confidence interval suggests a nonsignificant average effect, the prediction interval implies that in specific future settings, a technology-based intervention might show only a negligible effect on loneliness or, conversely, a somewhat larger effect than the pooled estimate suggests.

A significant challenge encountered in this meta-analysis was the lack of essential data required for the computation of the effect size. Many studies failed to report sufficient statistical details, which considerably reduced the number of studies that could ultimately be included in the meta-analysis. This limitation was particularly striking given the large number of studies initially identified as suitable during the preliminary screening phase. Furthermore, some studies provided links to external repositories, where relevant data were supposed to be available. However, upon verification, these datasets were either missing or incomplete. This issue emphasizes the critical importance of comprehensive and transparent reporting in research. Consistent and standardized reporting of results, including the provision of complete datasets, would greatly enhance the feasibility and reliability of future meta-analyses and systematic reviews [63,111]. More rigorous data availability would allow for more robust conclusions with greater explanatory value, and so improve the overall quality and impact of evidence synthesis in the field.

Our meta-analysis found a small and statistically nonsignificant overall effect of technology-based interventions on loneliness, which should be interpreted in the context of prior research on loneliness interventions. Earlier meta-analyses in the field have generally reported small-to-moderate effects. For example, the landmark review by Masi et al [28], which examined various approaches such as social skills training, enhanced social support, and opportunities for social contact, found overall small reductions in loneliness, with the strongest effects observed for interventions addressing maladaptive social cognition [28,92]. In a more recent comprehensive review, Lasgaard et al [24] synthesized 128 studies and likewise reported a moderate loneliness reduction for randomized trials (SMD on the order of ~ −0.5). These earlier findings provide important context for interpreting our more modest pooled effect, and reinforce the conclusion that loneliness interventions in general tend to yield small-to-moderate benefits. Indeed, when focusing on digital interventions for older adults, recent meta-analyses have often

XSL•FO
RenderX

found no significant effect: both Jin et al [75] and Shah et al [101] concluded that technology-based programs did not measurably reduce loneliness in older populations (pooled SMDs around 0 to −0.1; $P>.5$). An earlier analysis by Choi et al [112] had suggested that computer and internet training programs can be effective for older adults, but the evidence base at that time was limited to a few small studies. Given that the present meta-analysis spans a broader age range, its findings contribute to a more nuanced picture of how digital interventions may operate across different demographic groups. The literature has yielded mixed results depending on population characteristics and intervention modality [24,28,103], and while such comparisons could not be undertaken in the present study, prior findings highlight the value of considering demographic and contextual factors when evaluating intervention outcomes. For instance, Eccles and Qualter [23] examined interventions among children and adolescents and found that while youth loneliness can be alleviated, effects varied greatly between studies, with substantial unexplained heterogeneity. Taken together, the broader evidence base suggests that intervention effectiveness may depend on factors such as target population and intervention format, though these influences could not be directly assessed in the present meta-analysis.

Strengths and Limitations

This meta-analysis adheres to PRISMA guidelines, ensuring methodological transparency and rigor. A comprehensive search strategy across multiple databases, supplemented by citation tracking, minimizes the risk of missing relevant studies. By incorporating prediction intervals alongside standard heterogeneity assessments, the analysis provides deeper insights into the variability of intervention effects. Furthermore, it contributes to the growing evidence base on technology-based interventions for loneliness, which is particularly relevant in the context of increasing workforce shortages in health care and social services.

However, the study has several limitations. First, the number of eligible studies that provided sufficient statistical information was small, which restricted the scope of the meta-analysis and limited the ability to conduct subgroup or moderator analyses. This constrains the conclusions that can be drawn regarding factors that may influence intervention effectiveness. Furthermore, several limitations of the review process should be acknowledged. Although we did not use formal language restrictions, we used English keywords for the search. This may have limited the identification of studies published in other languages. Additionally, our search strategy relied on major databases and citation tracking. We did not conduct a separate systematic search for gray literature, so some unpublished trials might have been missed. We also decided not to contact study authors for missing data because of the practical reasons explained in the *Methods* section. This resulted in the exclusion of studies that could have been eligible if more information was available.

Implications

Technology is not a panacea for loneliness, yet it may serve as a useful tool when applied meaningfully and with consideration for the individual needs of the user. Communication technologies, such as video calls and messaging platforms, have been proposed as complements to interpersonal relationships and may help reduce feelings of loneliness in some contexts, particularly among older adults [113]. Similarly, social robots have emerged as nonpharmacological interventions capable of reducing both loneliness and depression [114]. However, a critical factor remains the manner in which these technologies are used. Active engagement in digital interactions has been shown to yield positive effects, whereas passive consumption of digital content may, conversely, exacerbate feelings of isolation [115]. Therefore, it is essential to consider users' age, technological literacy, and personal preferences when implementing technological solutions, to maximize their potential benefits and to avoid unintended increases in loneliness or social isolation.

Conclusions

This meta-analysis provides a rigorous synthesis of RCTs evaluating technology-based interventions for loneliness, including participants from adolescence to older adulthood. This study differs from existing reviews by focusing exclusively on randomized trials and the inclusion of a broader age demographic, whereas previous landmark studies often combined different study designs [28] or focused specifically on the older adult population [75]. Furthermore, the use of an RCT-only design, Hartung-Knapp adjustment, and prediction intervals offers a rigorous summary of the current evidence and highlights substantial variability in intervention outcomes. The small and nonsignificant pooled effect, together with moderate heterogeneity and a wide prediction interval, shows that results differ across studies and settings. These findings clarify current limitations in the evidence base and point to the need for better-designed and more transparently reported trials. In practice, technology-based approaches should be considered supportive tools that require careful tailoring to user characteristics and implementation contexts rather than universal solutions to loneliness.

Acknowledgments

The authors would like to express their gratitude to Marta Sykorova for her invaluable contribution to defining the search syntax, assisting in screening reports and studies, and extracting the necessary data. They also thank Kristyna Zivna for her efforts in identifying and extracting key statistical indicators essential for our analyses and meta-analytic calculations. The authors declare the use of generative artificial intelligence (AI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision:

Proofreading and editing

XSL•FO
RenderX

The GAI tool used was ChatGPT (GPT-5.1).

Responsibility for the final manuscript lies entirely with the authors.

GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Declaration submitted by: Collective responsibility

All authors declare that no other uses of large language models were made other than as stated.

Funding

The work was supported by ERDF/ESF project DigiWELL (CZ.02.01.01/00/22_008/0004583).

Data Availability

The study materials, data, and study code generated and/or analyzed during the current study are available in the Zenodo repository [116]. If any additional information regarding the data or analytic code is required, these are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: ZM, JS, PT
Data curation: MB, LN, OB
Formal analysis: JS, LN
Methodology: JS, MB, OB
Project administration: ZM, PT
Resources: JS, MB, OB
Supervision: ZM, PT
Writing – original draft: ZM, MB, JS, JH
Writing – review & editing: ZM, MB, JS, LN, JH, KM
All authors have read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search syntaxes for database search.
[DOCX File, 21 KB - jmir_v28i1e80059_app1.docx]

Multimedia Appendix 2
List of excluded studies.
[DOCX File, 34 KB - jmir_v28i1e80059_app2.docx]

Multimedia Appendix 3
Funnel plot.
[PNG File, 71 KB - jmir_v28i1e80059_app3.png]

Multimedia Appendix 4
GRADE (Grading of Recommendations Assessment, Development, and Evaluation) summary of findings: technology-based interventions compared to control conditions for reducing loneliness.
[DOCX File, 25 KB - jmir_v28i1e80059_app4.docx]

Checklist 1
PRISMA 2020 checklist, PRISMA 2020 for Abstracts checklist, and PRISMA-S checklist.
[PDF File, 367 KB - jmir_v28i1e80059_app5.pdf]

References

1. Hackett RA, Hudson JL, Chilcot J. Loneliness and type 2 diabetes incidence: findings from the English Longitudinal Study of Ageing. Diabetologia 2020 Nov;63(11):2329-2338. [doi: 10.1007/s00125-020-05258-6] [Medline: 32929525]

XSL•FO
RenderX

2. Leigh-Hunt N, Bagguley D, Bash K, et al. An overview of systematic reviews on the public health consequences of social isolation and loneliness. Public Health (Fairfax) 2017 Nov;152:157-171. [doi: 10.1016/j.puhe.2017.07.035] [Medline: 28915435]
3. Kobos E, Szewczyk A, Świątkowska T, Kryczka T, Sienkiewicz Z. Relationship between loneliness and blood glucose control in diabetes. BMC Public Health 2020 Jul 20;20(1):1140. [doi: 10.1186/s12889-020-09241-z] [Medline: 32689971]
4. Wang X, Ma H, Li X, Heianza Y, Fonseca V, Qi L. Joint association of loneliness and traditional risk factor control and incident cardiovascular disease in diabetes patients. Eur Heart J 2023 Jul 21;44(28):2583-2591. [doi: 10.1093/eurheartj/ehad306] [Medline: 37385629]
5. Zhou F, Deng L, Guo C, et al. Diabetes distress as mediators of loneliness and health promotion behaviour: a cross-sectional study. BMJ Open 2023 Dec 28;13(12):e079674. [doi: 10.1136/bmjopen-2023-079674] [Medline: 38154898]
6. Kuczmarski MF, Orsega-Smith E, Evans MK, Zonderman AB. The association of loneliness with diabetes is mediated by physical activity and BMI but not diet quality. Nutrients 2023 Nov 25;15(23):4923. [doi: 10.3390/nu15234923] [Medline: 38068781]
7. Richard A, Rohrmann S, Vandeleur CL, Schmid M, Barth J, Eichholzer M. Loneliness is adversely associated with physical and mental health and lifestyle factors: results from a Swiss national survey. PLoS ONE 2017;12(7):e0181442. [doi: 10.1371/journal.pone.0181442] [Medline: 28715478]
8. Solmi M, Veronese N, Galvano D, et al. Factors associated with loneliness: an umbrella review of observational studies. J Affect Disord 2020 Jun 15;271:131-138. [doi: 10.1016/j.jad.2020.03.075] [Medline: 32479308]
9. Beutel ME, Klein EM, Brähler E, et al. Loneliness in the general population: prevalence, determinants and relations to mental health. BMC Psychiatry 2017 Mar 20;17(1):97. [doi: 10.1186/s12888-017-1262-x] [Medline: 28320380]
10. Erzen E, Çikrikci Ö. The effect of loneliness on depression: a meta-analysis. Int J Soc Psychiatry 2018 Aug;64(5):427-435. [doi: 10.1177/0020764018776349] [Medline: 29792097]
11. Wang J, Lloyd-Evans B, Marston L, Mann F, Ma R, Johnson S. Loneliness as a predictor of outcomes in mental disorders among people who have experienced a mental health crisis: a 4-month prospective study. BMC Psychiatry 2020 May 20;20(1):249. [doi: 10.1186/s12888-020-02665-2] [Medline: 32434492]
12. Caple V, Maude P, Walter R, Ross A. An exploration of loneliness experienced by people living with mental illness and the impact on their recovery journey: an integrative review. J Psychiatr Ment Health Nurs 2023 Dec;30(6):1170-1191. [doi: 10.1111/jpm.12945] [Medline: 37382063]
13. Park C, Majeed A, Gill H, et al. The effect of loneliness on distinct health outcomes: a comprehensive review and meta-analysis. Psychiatry Res 2020 Dec;294:113514. [doi: 10.1016/j.psychres.2020.113514] [Medline: 33130511]
14. Liang YY, Zhou M, He Y, et al. Observational and genetic evidence disagree on the association between loneliness and risk of multiple diseases. Nat Hum Behav 2024 Nov;8(11):2209-2221. [doi: 10.1038/s41562-024-01970-0] [Medline: 39284978]
15. Rico-Uribe LA, Caballero FF, Martín-María N, Cabello M, Ayuso-Mateos JL, Miret M. Association of loneliness with all-cause mortality: a meta-analysis. PLoS ONE 2018;13(1):e0190033. [doi: 10.1371/journal.pone.0190033] [Medline: 29300743]
16. Zhao J, Reese JB, Han X, Yabroff R. Associations of loneliness and mortality risk among cancer survivors in the United States. J Clin Oncol 2023 Jun 1;41(16_suppl):6531-6531. [doi: 10.1200/JCO.2023.41.16_suppl.6531] [Medline: 37034189]
17. Cacioppo JT, Hughes ME, Waite LJ, Hawkley LC, Thisted RA. Loneliness as a specific risk factor for depressive symptoms: cross-sectional and longitudinal analyses. Psychol Aging 2006 Mar;21(1):140-151. [doi: 10.1037/0882-7974.21.1.140] [Medline: 16594799]
18. Baumeister RF, Leary MR. The need to belong: desire for interpersonal attachments as a fundamental human motivation. In: Interpersonal Development: Taylor & Francis; 2007:57-89. [doi: 10.4324/9781351153683-3]
19. Hawkley LC. Loneliness and health. Nat Rev Dis Primers 2022 Apr 14;8(1):22. [doi: 10.1038/s41572-022-00355-9] [Medline: 35422043]
20. Holt-Lunstad J, Smith TB, Baker M, Harris T, Stephenson D. Loneliness and social isolation as risk factors for mortality: a meta-analytic review. Perspect Psychol Sci 2015 Mar;10(2):227-237. [doi: 10.1177/1745691614568352] [Medline: 25910392]
21. Maes M, Vanhalst J. Loneliness as a double-edged sword: an adaptive function with maladaptive consequences. Eur J Dev Psychol 2025 Mar 4;22(2):188-200. [doi: 10.1080/17405629.2024.2333584]
22. Lim MH, Holt-Lunstad J, Badcock JC. Loneliness: contemporary insights into causes, correlates, and consequences. Soc Psychiatry Psychiatr Epidemiol 2020 Jul;55(7):789-791. [doi: 10.1007/s00127-020-01891-z] [Medline: 32529475]
23. Eccles AM, Qualter P. Review: alleviating loneliness in young people - a meta-analysis of interventions. Child Adolesc Ment Health 2021 Feb;26(1):17-33. [doi: 10.1111/camh.12389] [Medline: 32406165]
24. Lasgaard M, Løvschall C, Qualter P, et al. Are loneliness interventions effective in reducing loneliness? A meta-analytic review of 128 studies. Eur J Public Health 2022 Oct 21;32(Supplement_3):ckac129.266. [doi: 10.1093/eurpub/ckac129.266]
25. Osborn T, Weatherburn P, French RS. Interventions to address loneliness and social isolation in young people: a systematic review of the evidence on acceptability and effectiveness. J Adolesc 2021 Dec;93:53-79. [doi: 10.1016/j.adolescence.2021.09.007] [Medline: 34662802]

XSL•FO
RenderX

26. Crowe CL, Liu L, Bagnarol N, Fried LP. Loneliness prevention and the role of the public health system. Perspect Public Health 2024 Jan;144(1):31-38. [doi: 10.1177/17579139221106579] [Medline: 35815809]
27. Patil U, Braun KL. Interventions for loneliness in older adults: a systematic review of reviews. Front Public Health 2024;12:1427605. [doi: 10.3389/fpubh.2024.1427605] [Medline: 39091527]
28. Masi CM, Chen HY, Hawkley LC, Cacioppo JT. A meta-analysis of interventions to reduce loneliness. Pers Soc Psychol Rev 2011 Aug;15(3):219-266. [doi: 10.1177/1088868310377394] [Medline: 20716644]
29. Peplau LA, Perlman D. Loneliness: A Sourcebook of Current Theory, Research, and Therapy: John Wiley & Sons; 1982.
30. Norlin J, McKee KJ, Lennartsson C, Dahlberg L. Quantity and quality of social relationships and their associations with loneliness in older adults. Aging Ment Health 2025 Jul;29(7):1198-1208. [doi: 10.1080/13607863.2025.2460068] [Medline: 39951254]
31. Weiss RS. Loneliness: The Experience of Emotional and Social Isolation: MIT press; 1973. URL: https://psycnet.apa.org/record/1974-22306-000 [accessed 2026-04-14]
32. Bolmsjö I, Tengland PA, Rämgård M. Existential loneliness: an attempt at an analysis of the concept and the phenomenon. Nurs Ethics 2019 Aug;26(5):1310-1325. [doi: 10.1177/0969733017748480] [Medline: 29471724]
33. Cacioppo S, Grippo AJ, London S, Goossens L, Cacioppo JT. Loneliness: clinical import and interventions. Perspect Psychol Sci 2015 Mar;10(2):238-249. [doi: 10.1177/1745691615570616] [Medline: 25866548]
34. Hawkley LC, Browne MW, Cacioppo JT. How can I connect with thee? Let me count the ways. Psychol Sci 2005 Oct;16(10):798-804. [doi: 10.1111/j.1467-9280.2005.01617.x] [Medline: 16181443]
35. Hawkley LC, Gu Y, Luo YJ, Cacioppo JT. The mental representation of social connections: generalizability extended to Beijing adults. PLoS ONE 2012;7(9):e44065. [doi: 10.1371/journal.pone.0044065] [Medline: 23028486]
36. Malli MA, Ryan S, Maddison J, Kharicha K. Experiences and meaning of loneliness beyond age and group identity. Sociol Health Illn 2023 Jan;45(1):70-89. [doi: 10.1111/1467-9566.13539] [Medline: 36073866]
37. Yang K, Victor C. Age and loneliness in 25 European nations. Ageing Soc 2011 Nov;31(8):1368-1388. [doi: 10.1017/S0144686X1000139X]
38. Luhmann M, Hawkley LC. Age differences in loneliness from late adolescence to oldest old age. Dev Psychol 2016 Jun;52(6):943-959. [doi: 10.1037/dev0000117] [Medline: 27148782]
39. Mund M, Lüdtke O, Neyer FJ. Owner of a lonely heart: the stability of loneliness across the life span. J Pers Soc Psychol 2020;119(2):497-516. [doi: 10.1037/pspp0000262.supp]
40. Ong AD, Uchino BN, Wethington E. Loneliness and health in older adults: a mini-review and synthesis. Gerontology 2016;62(4):443-449. [doi: 10.1159/000441651] [Medline: 26539997]
41. von Soest T, Luhmann M, Hansen T, Gerstorf D. Development of loneliness in midlife and old age: its nature and correlates. J Pers Soc Psychol 2020 Feb;118(2):388-406. [doi: 10.1037/pspp0000219] [Medline: 30284871]
42. Igami K, Hosozawa M, Ikeda A, Bann D, Shimizu T, Iso H. Adolescent loneliness in 70 countries across Africa, America, and Asia: a comparison of prevalence and correlates. J Adolesc Health 2023 Jun;72(6):906-913. [doi: 10.1016/j.jadohealth.2022.12.029] [Medline: 36775749]
43. Su Y, Rao W, Li M, Caron G, D'Arcy C, Meng X. Prevalence of loneliness and social isolation among older adults during the COVID-19 pandemic: a systematic review and meta-analysis. Int Psychogeriatr 2023 May;35(5):229-241. [doi: 10.1017/S1041610222000199] [Medline: 35357280]
44. Anderson GO, Thayer CE. Loneliness and social connections: a national survey of adults 45 and older. : AARP Foundation; 2018 URL: https://www.icmha.org/wp-content/uploads/2020/02/AARP-740C-Loneliness-Report-v3-with-LOGO-TAG.pdf [accessed 2026-04-14]
45. Cigna U.S. loneliness index: survey of 20,000 Americans examining behaviors driving loneliness in the United States. : Cigna; 2018 URL: https://www.cigna.com/assets/docs/newsroom/loneliness-survey-2018-full-report.pdf [accessed 2026-04-14]
46. Loneliness and the workplace: 2020 U.S. report. : Cigna; 2020 URL: https://www.cigna.com/static/www-cigna-com/docs/about-us/newsroom/studies-and-reports/combatting-loneliness/cigna-2020-loneliness-report.pdf [accessed 2026-04-14]
47. Stegen H, Duppen D, Savieri P, et al. Loneliness prevalence of community-dwelling older adults and the impact of the mode of measurement, data collection, and country: a systematic review and meta-analysis. Int Psychogeriatr 2024 Sep;36(9):747-761. [doi: 10.1017/S1041610224000425] [Medline: 38525677]
48. Yang K. Loneliness: A Social Problem: Routledge; 2019. [doi: 10.4324/9781315148410]
49. Graham EK, Beck ED, Jackson K, et al. Do we become more lonely with age? A coordinated data analysis of nine longitudinal studies. Psychol Sci 2024 Jun;35(6):579-596. [doi: 10.1177/09567976241242037] [Medline: 38687352]
50. Lay-Yee R, Campbell D, Milne B. Social attitudes and activities associated with loneliness: findings from a New Zealand national survey of the adult population. Health Soc Care Community 2022 May;30(3):1120-1132. [doi: 10.1111/hsc.13351] [Medline: 33765362]
51. Nyqvist F, Victor CR, Forsman AK, Cattan M. The association between social capital and loneliness in different age groups: a population-based study in Western Finland. BMC Public Health 2016 Jul 11;16:542. [doi: 10.1186/s12889-016-3248-x] [Medline: 27400659]

XSL•FO
RenderX

52. Hickin N, Käll A, Shafran R, Sutcliffe S, Manzotti G, Langan D. The effectiveness of psychological interventions for loneliness: a systematic review and meta-analysis. Clin Psychol Rev 2021 Aug;88:102066. [doi: 10.1016/j.cpr.2021.102066] [Medline: 34339939]
53. Baarck J, d'Hombres B, Tintori G. Loneliness in Europe before and during the COVID-19 pandemic. Health Policy 2022 Nov;126(11):1124-1129. [doi: 10.1016/j.healthpol.2022.09.002] [Medline: 36182348]
54. Banerjee D, Rai M. Social isolation in Covid-19: the impact of loneliness. Int J Soc Psychiatry 2020 Sep;66(6):525-527. [doi: 10.1177/0020764020922269] [Medline: 32349580]
55. Hwang TJ, Rabheru K, Peisah C, Reichman W, Ikeda M. Loneliness and social isolation during the COVID-19 pandemic. Int Psychogeriatr 2020 Oct;32(10):1217-1220. [doi: 10.1017/S1041610220000988] [Medline: 32450943]
56. Patel SS, Clark-Ginsberg A. Incorporating issues of elderly loneliness into the Coronavirus Disease-2019 public health response. Disaster Med Public Health Prep 2020 Jun;14(3):e13-e14. [doi: 10.1017/dmp.2020.145] [Medline: 32379016]
57. Hajek A, Sutin AR, Posi G, et al. Chronic loneliness and chronic social isolation among older adults. A systematic review, meta-analysis and meta-regression. Aging Ment Health 2025 Feb;29(2):185-200. [doi: 10.1080/13607863.2024.2385448] [Medline: 39126212]
58. Shah SGS, Nogueras D, van Woerden HC, Kiparoglou V. Evaluation of the effectiveness of digital technology interventions to reduce loneliness in older adults: systematic review and meta-analysis. J Med Internet Res 2021;23(6):e24712. [doi: 10.2196/24712] [Medline: 34085942]
59. Bentley SV, Haslam C, Haslam SA, Jetten J, Larwood J, La Rue CJ. GROUPS 2 CONNECT: an online activity to maintain social connection and well-being during COVID-19. Appl Psychol Health Well Being 2022 Nov;14(4):1189-1210. [doi: 10.1111/aphw.12330] [Medline: 34958172]
60. Kenyon K, Kinakh V, Harrison J. Social virtual reality helps to reduce feelings of loneliness and social anxiety during the Covid-19 pandemic. Sci Rep 2023 Nov 7;13(1):19282. [doi: 10.1038/s41598-023-46494-1] [Medline: 37935718]
61. Sandora J, Meier Z, Tavel P, Buchtova M, Sykorova M. The effectiveness of technology-based interventions on reduction of loneliness: systematic review and meta-analysis. PROSPERO. 2024. URL: https://www.crd.york.ac.uk/PROSPERO/view/CRD42024505117 [accessed 2026-04-14]
62. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009 Aug 18;151(4):264-269. [doi: 10.7326/0003-4819-151-4-200908180-00135] [Medline: 19622511]
63. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
64. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
65. Zotero. 2024. URL: https://www.zotero.org/ [accessed 2026-04-14]
66. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-Analysis: John Wiley & Sons; 2009. [doi: 10.1002/9780470743386]
67. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Routledge; 1988. [doi: 10.4324/9780203771587]
68. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
69. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019 Jun;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]
70. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
71. McGuinness LA, Higgins JPT. Risk-of-bias VISualization (robvis): an R package and Shiny web app for visualizing risk-of-bias assessments. Res Synth Methods 2021 Jan;12(1):55-61. [doi: 10.1002/jrsm.1411] [Medline: 32336025]
72. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-926. [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
73. Sun L. Social media usage and students' social anxiety, loneliness and well-being: does digital mindfulness-based intervention effectively work? BMC Psychol 2023 Oct 31;11(1):362. [doi: 10.1186/s40359-023-01398-7] [Medline: 37904182]
74. Sun L. Retraction note: social media usage and students' social anxiety, loneliness and well-being: does digital mindfulness-based intervention effectively work? BMC Psychol 2025 May 7;13(1):480. [doi: 10.1186/s40359-025-02810-0] [Medline: 40336136]
75. Jin W, Liu Y, Yuan S, Bai R, Li X, Bai Z. The effectiveness of technology-based interventions for reducing loneliness in older adults: a systematic review and meta-analysis of randomized controlled trials. Front Psychol 2021;12:711030. [doi: 10.3389/fpsyg.2021.711030] [Medline: 34955948]
76. Shapira N, Barak A, Gal I. Promoting older adults' well-being through internet training and use. Aging Ment Health 2007 Sep;11(5):477-484. [doi: 10.1080/13607860601086546] [Medline: 17882585]

XSL•FO
RenderX

77. Hind D, Mountain G, Gossage-Worrall R, et al. Putting Life in Years (PLINY): a randomised controlled trial and mixed-methods process evaluation of a telephone friendship intervention to improve mental well-being in independently living older people. Public Health Res 2014;2(7):1-222. [doi: 10.3310/phr02070] [Medline: 25642518]
78. Tsai HH, Tsai YF, Wang HH, Chang YC, Chu HH. Videoconference program enhances social support, loneliness, and depressive status of elderly nursing home residents. Aging Ment Health 2010 Nov;14(8):947-954. [doi: 10.1080/13607863.2010.501057] [Medline: 21069600]
79. Tsai HH, Cheng CY, Shieh WY, Chang YC. Effects of a smartphone-based videoconferencing program for older nursing home residents on depression, loneliness, and quality of life: a quasi-experimental study. BMC Geriatr 2020 Jan 28;20(1):27. [doi: 10.1186/s12877-020-1426-2] [Medline: 31992217]
80. Tsai HH, Tsai YF. Changes in depressive symptoms, social support, and loneliness over 1 year after a minimum 3-month videoconference program for older nursing home residents. J Med Internet Res 2011 Nov 15;13(4):e93. [doi: 10.2196/jmir.1678] [Medline: 22086660]
81. Slegers K, van Boxtel MPJ, Jolles J. Effects of computer training and internet usage on the well-being and quality of life of older adults: a randomized, controlled study. J Gerontol B Psychol Sci Soc Sci 2008 May;63(3):176-184. [doi: 10.1093/geronb/63.3.p176] [Medline: 18559683]
82. Schwindenhammer TM. Videoconferencing intervention for depressive symptoms and loneliness in nursing home elders [Master's Thesis]. : Illinois State University; 2014 URL: https://www.proquest.com/docview/1548716414/abstract/452A58D6944F4846PQ/1 [accessed 2026-04-14]
83. Iyer RB, Vadlapudi S, Iyer L, et al. Impact of the Heartfulness program on loneliness in high schoolers: randomized survey study. Appl Psychol Health Well Being 2023 Feb;15(1):66-79. [doi: 10.1111/aphw.12360] [Medline: 35384302]
84. Karkosz S, Szymański R, Sanna K, Michałowski J. Effectiveness of a web-based and mobile therapy chatbot on anxiety and depressive symptoms in subclinical young adults: randomized controlled trial. JMIR Form Res 2024 Mar 20;8(1):e47960. [doi: 10.2196/47960] [Medline: 38506892]
85. Brog NA, Hegy JK, Berger T, Znoj H. Effects of an internet-based self-help intervention for psychological distress due to COVID-19: results of a randomized controlled trial. Internet Interv 2022 Mar;27:100492. [doi: 10.1016/j.invent.2021.100492] [Medline: 34956841]
86. Hill W, Weinert C, Cudney S. Influence of a computer intervention on the psychological status of chronically ill rural women: preliminary results. Nurs Res 2006;55(1):34-42. [doi: 10.1097/00006199-200601000-00005] [Medline: 16439927]
87. Robinson H, Macdonald B, Kerse N, Broadbent E. The psychosocial effects of a companion robot: a randomized controlled trial. J Am Med Dir Assoc 2013 Sep;14(9):661-667. [doi: 10.1016/j.jamda.2013.02.007] [Medline: 23545466]
88. Shapira S, Yeshua-Katz D, Cohn-Schwartz E, Aharonson-Daniel L, Sarid O, Clarfield AM. A pilot randomized controlled trial of a group intervention via Zoom to relieve loneliness and depressive symptoms among older persons during the COVID-19 outbreak. Internet Interv 2021 Apr;24:100368. [doi: 10.1016/j.invent.2021.100368] [Medline: 33527072]
89. Jiang D, Yeung D, Choi N, et al. Effectiveness of teledelivered behavioral activation and mindfulness interventions in reducing loneliness. Innov Aging 2023 Dec 21;7(Supplement_1):844. [doi: 10.1093/geroni/igad104.2722]
90. Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 2000 Jun;56(2):455-463. [doi: 10.1111/j.0006-341x.2000.00455.x] [Medline: 10877304]
91. Burke L. Interventions to reduce loneliness in children and adolescents: a systematic review and meta-analysis. Eur J Public Health 2024 Nov 1;34(Supplement_3):ckae144. [doi: 10.1093/eurpub/ckae144.056]
92. Beckers A, Buecker S, Casabianca EJ, Nurminen M. Effectiveness of interventions tackling loneliness. : European Commission; 2022 URL: https://www.researchgate.net/publication/366055118_Effectiveness_of_interventions_tackling_loneliness [accessed 2026-04-14]
93. Cruwys T, Haslam C, Haslam SA, Dingle GA. Misery loves company: predictors of treatment response to a loneliness intervention. Psychother Res 2023 Jun;33(5):608-624. [doi: 10.1080/10503307.2022.2143300] [Medline: 36404283]
94. Ratcliffe J, Matcham F, Molteni E, et al. Conceptualising the context and mechanisms for tackling loneliness in older adults through interventions: a critical interpretive synthesis. medRxiv. Preprint posted online on Jan 6, 2025. [doi: 10.1101/2025.01.06.25320003]
95. Fu Z, Yan M, Meng C. The effectiveness of remote delivered intervention for loneliness reduction in older adults: a systematic review and meta-analysis. Front Psychol 2022;13:935544. [doi: 10.3389/fpsyg.2022.935544] [Medline: 35967719]
96. Grillich L, Titscher V, Klingenstein P, et al. The effectiveness of interventions to prevent loneliness and social isolation in the community-dwelling and old population: an overview of systematic reviews and meta-analysis. Eur J Public Health 2023 Apr 1;33(2):235-241. [doi: 10.1093/eurpub/ckad006] [Medline: 36893335]
97. van der Vaart R, van Driel D, Pronk K, et al. The role of age, education, and digital health literacy in the usability of internet-based cognitive behavioral therapy for chronic pain: mixed methods study. JMIR Form Res 2019 Nov 21;3(4):e12883. [doi: 10.2196/12883] [Medline: 31750839]
98. Cohen-Mansfield J, Perach R. Interventions for alleviating loneliness among older persons: a critical review. Am J Health Promot 2015;29(3):e109-e125. [doi: 10.4278/ajhp.130418-LIT-182] [Medline: 24575725]

XSL•FO
RenderX

99. Eager S, Johnson S, Pitman A, Uribe M, Qualter P, Pearce E. Young people's views on the acceptability and feasibility of loneliness interventions for their age group. BMC Psychiatry 2024 Apr 23;24(1):308. [doi: 10.1186/s12888-024-05751-x] [Medline: 38654301]
100. Shekelle PG, Miake-Lye IM, Begashaw MM, et al. Interventions to reduce loneliness in community-living older adults: a systematic review and meta-analysis. J Gen Intern Med 2024 May;39(6):1015-1028. [doi: 10.1007/s11606-023-08517-5] [Medline: 38200279]
101. Shah SGS, Nogueras D, van Woerden H, Kiparoglou V. Effectiveness of digital technology interventions to reduce loneliness in adults: a protocol for a systematic review and meta-analysis. BMJ Open 2019 Sep 27;9(9):e032455. [doi: 10.1136/bmjopen-2019-032455] [Medline: 31562164]
102. Centre for Reviews and Dissemination. University of York. 2014. URL: https://www.york.ac.uk/crd [accessed 2026-04-14]
103. Morrish N, Choudhury S, Medina-Lara A. What works in interventions targeting loneliness: a systematic review of intervention characteristics. BMC Public Health 2023 Nov 9;23(1):2214. [doi: 10.1186/s12889-023-17097-2] [Medline: 37946186]
104. Blodgett JM, Tiley K, Harkness F, Musella M. What works to reduce loneliness: a rapid systematic review of 101 interventions. J Public Health Policy 2025 Jun;46(2):245-268. [doi: 10.1057/s41271-025-00561-1] [Medline: 40050496]
105. Das A, Padala KP, Crawford CG, et al. A systematic review of loneliness and social isolation scales used in epidemics and pandemics. Psychiatry Res 2021 Dec;306:114217. [doi: 10.1016/j.psychres.2021.114217] [Medline: 34644661]
106. Maes M, Qualter P, Lodder GMA, Mund M. How (not) to measure loneliness: a review of the eight most commonly used scales. Int J Environ Res Public Health 2022 Aug 30;19(17):10816. [doi: 10.3390/ijerph191710816] [Medline: 36078540]
107. Hudiyana J, Lincoln TM, Hartanto S, et al. How universal is a construct of loneliness? Measurement invariance of the UCLA loneliness scale in Indonesia, Germany, and the United States. Assessment 2022 Dec;29(8):1795-1805. [doi: 10.1177/10731911211034564] [Medline: 34301150]
108. Manera KE, Smith BJ, Owen KB, Phongsavan P, Lim MH. Psychometric assessment of scales for measuring loneliness and social isolation: an analysis of the household, income and labour dynamics in Australia (HILDA) survey. Health Qual Life Outcomes 2022 Mar 5;20(1):40. [doi: 10.1186/s12955-022-01946-6] [Medline: 35248075]
109. Schoenmakers EC, Lasgaard M, McHugh Power J. Guidelines for evaluating and reporting social isolation and loneliness interventions. J Health Psychol 2025 Feb;30(2):338-352. [doi: 10.1177/13591053241238127] [Medline: 38527950]
110. Valtorta NK, Kanaan M, Gilbody S, Hanratty B. Loneliness, social isolation and social relationships: what are we measuring? A novel framework for classifying and comparing tools. BMJ Open 2016 Apr 18;6(4):e010799. [doi: 10.1136/bmjopen-2015-010799] [Medline: 27091822]
111. Simera I, Moher D, Hirst A, Hoey J, Schulz KF, Altman DG. Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. BMC Med 2010 Apr 26;8(1):24. [doi: 10.1186/1741-7015-8-24] [Medline: 20420659]
112. Choi M, Kong S, Jung D. Computer and internet interventions for loneliness and depression in older adults: a meta-analysis. Healthc Inform Res 2012 Sep;18(3):191-198. [doi: 10.4258/hir.2012.18.3.191] [Medline: 23115742]
113. Chen YRR, Schulz PJ. The effect of information communication technology interventions on reducing social isolation in the elderly: a systematic review. J Med Internet Res 2016 Jan 28;18(1):e18. [doi: 10.2196/jmir.4596] [Medline: 26822073]
114. Yen HY, Huang CW, Chiu HL, Jin G. The effect of social robots on depression and loneliness for older residents in long-term care facilities: a meta-analysis of randomized controlled trials. J Am Med Dir Assoc 2024 Jun;25(6):104979. [doi: 10.1016/j.jamda.2024.02.017] [Medline: 38614134]
115. Nowland R, Necka EA, Cacioppo JT. Loneliness and social internet use: pathways to reconnection in a digital world? Perspect Psychol Sci 2018 Jan;13(1):70-87. [doi: 10.1177/1745691617713052] [Medline: 28937910]
116. Lukas-novak/the-effectiveness-of-technology-based-interventions-on-the-reduction-of-loneliness-meta-analysis: fix of paths. Zenodo. URL: https://zenodo.org/records/18061382 [accessed 2026-04-16]

Abbreviations

CBT: cognitive behavioral therapy
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses-Search extension
PROSPERO: International Prospective Register of Systematic Reviews
RCT: randomized controlled trial
SMD: standardized mean difference

XSL•FO
RenderX

Edited by S Brini; submitted 03.Jul.2025; peer-reviewed by H Napierala, Z Liu; revised version received 03.Feb.2026; accepted 13.Feb.2026; published 08.May.2026.

Please cite as:
Meier Z, Buchtova M, Sandora J, Novak L, Helvich J, Buchta O, Furstova J, Malinakova K, Tavel P
Efficacy of Technology-Based Interventions on the Reduction of Loneliness: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e80059
URL: https://www.jmir.org/2026/1/e80059
doi:10.2196/80059

© Zdenek Meier, Marie Buchtova, Jan Sandora, Lukas Novak, Jakub Helvich, Ondrej Buchta, Jana Furstova, Klara Malinakova, Peter Tavel. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effectiveness of Digital Health Interventions in Older Adults With Frailty and Sarcopenia: Systematic Review and Meta Analysis of Randomized Controlled Trials

Ting Dai[1,2], PhD; Changsheng Guo[3], PhD; Lingli Gao[1], PhD; Jinyu Huang[4], PhD; Yan Chen[1], PhD; Yujie Zhang[1], BA; Jing Gao[1], PhD; Xiaodong Feng[1], PhD

[1]Rehabilitation Center, The First Affiliated Hospital of Henan University of Traditional Chinese Medicine, No.19 Renmin Road, Zhengzhou, Henan, China

[2]The First Clinical Medical College, Henan University of Chinese Medicine, Zhengzhou, China

[3]Rehabilitation Center, The Second Affiliated Hospital of Heilongjiang University of Chinese Medicine, Harbin, Heilongjiang, China

[4]The First Affiliated Hospital of Henan University of CM, Heart Center/National Regional (Traditional Chinese Medicine) Cardiovascular Diagnosis and Treatment Center, Zhengzhou, Henan, China

Corresponding Author:
Xiaodong Feng, PhD
Rehabilitation Center, The First Affiliated Hospital of Henan University of Traditional Chinese Medicine, No.19 Renmin Road, Zhengzhou, Henan, China

Abstract

Background: Frailty and sarcopenia represent substantial global health challenges, frequently diminishing patients' quality of life through impaired muscle function and physical performance. Digital health interventions (DHIs) have shown promise in mitigating these conditions among older adults. However, outcomes of such interventions in this demographic are inconsistent, and a thorough synthesis of existing evidence is lacking.

Objective: This study aimed to evaluate the effectiveness of DHIs in older adults with frailty and sarcopenia.

Methods: A comprehensive search of PubMed, Web of Science, MEDLINE, Embase, and Cochrane Library was conducted from their inception until January 2026 to identify randomized controlled trials. Meta-analyses were performed using R software (R Foundation for Statistical Computing). Study quality was evaluated using the revised Cochrane Risk of Bias Tool 2.0 (Cochrane Collaboration), and evidence certainty was assessed using GRADE (Grading of Recommendations, Assessment, Development, and Evaluation).

Results: From 3506 records, 16 studies were included. DHIs significantly improved total skeletal muscle mass (weighted mean difference [WMD] 1.01, 95% CI 0.08-1.94, 95% prediction interval [PI] −0.95 to 2.96), gait speed (WMD 0.09, 95% CI 0.03-0.15, 95% PI −0.1 to 0.26), Timed Up and Go Test (TUGT: WMD −0.52, 95% CI −1.02 to −0.03, 95% PI −1.93 to 0.85), 30-second Chair Stand Test (30CST: WMD 2.19, 95% CI 0.89-3.48, 95% PI −1.59 to 5.66), balance (standardized mean difference [SMD] 0.61, 95% CI 0-1.21, 95% PI −0.94 to 2.13), and quality of life (SMD 0.16, 95% CI 0.05-0.27, 95% PI 0.04-0.28). No significant improvements were observed in Appendicular Skeletal Muscle Mass Index (ASMI), grip strength, 6-minute walk test (6MWT), 2-minute walk test (2MWT), Short Physical Performance Battery (SPPB), or BMI. Although the pooled effect was favorable, the wide 95% PI suggests substantial between-study heterogeneity. Subgroup analysis stratified by intervention duration revealed significant intersubgroup differences in ASMI (χ^2_1 =9.93; P=.0016), indicating interventions lasting ≥12 weeks were more effective for improving ASMI (WMD 0.28, 95% CI 0.06-0.50, 95% PI −0.30 to 0.83). Subgroup analysis stratified by intervention type showed significant intersubgroup differences in balance (χ^2_3 =9.89; P=.0195), with exergame-based interventions showing significant effects (SMD 0.83, 95% CI 0.26-1.40).

Conclusions: This systematic review is the first to quantify the disease-specific efficacy of DHIs in improving muscle function, physical performance, and quality of life among older adults with frailty and sarcopenia, demonstrating their unique value as a scalable complementary approach. By overcoming geographical and resource constraints, DHIs support underserved populations. However, low evidence quality and heterogeneity warrant cautious interpretation. The 95% PIs indicate that actual effects may vary with population characteristics and implementation contexts. Nonetheless, DHIs represent a promising and cost-effective strategy for service expansion. Future high-quality studies are needed to better understand their effectiveness and implementation across settings.

Trial Registration: PROSPERO CRD420251135302; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251135302

XSL•FO
RenderX

(J Med Internet Res 2026;28:e88374) doi:10.2196/88374

KEYWORDS

digital health intervention; sarcopenia; frailty; older adults; meta-analysis

Introduction

With the accelerating trend of global population aging, frailty and sarcopenia have emerged as significant public health challenges that are closely associated with geriatric syndromes [1]. Frailty is a multidimensional syndrome characterized by decreased functional reserve and increased vulnerability to stressors [2]. Sarcopenia, defined as the loss of skeletal muscle mass and function, typically precedes frailty and is a core component of frailty [2,3]. A shared pathophysiological hallmark and a key clinical link between these 2 conditions is the presence of muscle atrophy and the consequent decline in physical performance [4,5]. Systematic reviews and meta-analyses indicate that the prevalence of frailty among people aged 60 years and older is 16%, while the prevalence of prefrailty is as high as 47% [6]. The prevalence of sarcopenia ranges from 10% to 27% [7]. Critically, both of these highly prevalent conditions are important predictors of serious adverse health outcomes [8-10], profoundly impacting older adults' function, quality of life, and survival of older adults [5,11]. Despite sufficient evidence indicating that moderate-to-high intensity resistance training effectively enhances muscle function and physical performance in older adults [12], the face-to-face implementation of this intervention is constrained by practical factors such as inadequate health system resources [13], heavy burden of travel for medical care [14], and age-related physical decline in older adults themselves [15]. These factors have significantly reduced patient adherence and motivation [16], highlighting an urgent need for innovative management strategies.

In recent years, the rapid advancement of digital technologies in the health care field has brought transformative opportunities for optimizing health interventions and strengthening health systems [17].

Digital health interventions (DHIs), leveraging technologies such as telemedicine, mobile apps, wearable devices, exergames, and virtual reality, can provide support for comprehensive health management in older adults [18,19]. They enable health monitoring, personalized health education, remote interventions, and continuous surveillance [19-21], holding great promise for overcoming economic, geographical, temporal, and physical barriers. Notably, the COVID-19 pandemic has accelerated the integration and adoption of digital processes within health care globally [22].

Multiple clinical studies have demonstrated that DHIs, such as wearable devices [23], exergaming [24], and telephone reminders [25], show potential in improving functional outcomes among older adults with frailty and sarcopenia. However, their implementation and scaling face dual challenges, as follows: (1) at the system level, structural disparities in health care resource allocation, uneven regional distribution of medical device networks, and insufficient digital transformation capacity may hinder the dissemination and adoption of advanced technologies, potentially exacerbating geographic inequities in health care access [26]; and (2) at the evidence level, significant heterogeneity in study designs and methodologies has led to inconsistent conclusions regarding efficacy and a lack of consensus, thereby further limiting the clinical translation of these evidence-based findings.

Although multiple reviews have systematically summarized the effectiveness of DHIs in older adults with frailty or sarcopenia [27-32], to the best of our knowledge, considerable methodological and content gaps still exist in the existing literature. First, at the study population level, current reviews focus on either frail or sarcopenic individuals as a single group, conducting pooled effect analyses for only a single disease state, and thus fail to adequately account for the fact that frailty and sarcopenia are distinct but interrelated conditions [33], resulting in conclusions that inadequately cover this high-risk population and lack clinical applicability. Second, regarding the consistency of evidence, inconsistent findings have been reported; for instance, one review demonstrated that digital technology interventions significantly improved grip strength [27], whereas another study observed no such beneficial effect [32]. Meanwhile, the narrow range of intervention types [28,31], the absence of long-term follow-up assessments [27-32], and the limited inclusion of core outcomes covering muscle function, physical function, and quality of life further restrict the comprehensive evaluation of intervention effects [28,30,31,34]. In addition, the moderating effects of key intervention parameters (eg, intervention duration and delivery mode) on efficacy have not been systematically analyzed.

These evidence gaps hinder the development of rehabilitation strategies for patients with frailty and sarcopenia. Therefore, this meta-analysis aims to (1) systematically evaluate the effectiveness of DHIs on muscle function (muscle mass and strength), physical function, and quality of life in older adults with frailty and sarcopenia; and (2) through subgroup analyses, explore the influences of potential moderators including intervention type, duration, control group design, delivery and synchronization modes, and follow-up effects, with the goal of providing clearer evidence-based guidance for clinical practice and future research.

Methods

Protocol and Registration

This systematic review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [35], using the PRISMA 2020 checklist (Checklist 1). Additionally, the study protocol was registered with the International Prospective Register of Systematic Reviews (CRD420251135302).

XSL•FO
RenderX

Data Sources and Search Strategy

This systematic search was conducted in strict accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses search extension) guidelines [36]. A systematic search was performed across 5 major databases: PubMed (NCBI), Embase, Cochrane Library (Wiley), Web of Science, and Ovid MEDLINE. The search encompassed the period from the inception of each database to July 10, 2025, with no restrictions imposed on language or publication date. To ensure the inclusion of the most recent evidence, the search was updated on January 9, 2026. This updated search featured an optimized combination of MeSH (Medical Subject Headings) and keywords, which was informed by previous studies [27,29,30] and tailored to the characteristics of each database, thereby yielding search strategies that differed from those used initially. It should be noted that the search strategies used have not been formally peer-reviewed.

To enhance the comprehensiveness of the retrieval process, the reference lists of pertinent systematic reviews and their included studies were meticulously reviewed. In instances where essential data were absent from the included studies, we reached out to the corresponding authors via email to request the unreported yet crucial original data. The literature search was restricted to the methodologies outlined above, and gray literature, unpublished studies, and preprints were not systematically explored. All database searches were completed on January 11, 2026. Comprehensive search strategies for each database are documented in Multimedia Appendix 1.

Eligibility Criteria

Two authors independently conducted the screening and selection of studies according to predefined criteria. Any discrepancies encountered were resolved through discussion or, if necessary, by consulting a third reviewer. The determination of study eligibility was guided by the PICOS (Population, Intervention, Comparison, Outcomes, and Study Design) framework.

The inclusion criteria were as follows: (1) participants were aged ≥60 years and carried a diagnosis of frailty or sarcopenia, established according to specified diagnostic criteria. (2) The intervention implemented was a DHI delivered via digital platforms or technologies, through one or more modalities, including but not limited to telephone, video, mobile apps, or wearable devices. (3) Control groups received nondigital interventions, including but not limited to usual care, waitlist, active control, or blank control. (4) The outcomes associated with frailty and sarcopenia were evaluated through comprehensive assessments of muscle strength, muscle mass, physical performance, BMI, and quality of life. Specifically, muscle mass measurements included skeletal muscle mass and the Appendicular Skeletal Muscle Mass Index (ASMI). Physical performance was assessed using various performance-based tests, including gait speed, the 6-Minute Walk Test (6MWT), the 2-Minute Walk Test (2MWT), the Five Times Sit to Stand Test (FTSST), the 30-Second Chair Stand Test (30CST), the Timed Up and Go Test (TUGT), the Short Physical Performance Battery (SPPB), and balance assessments. These outcomes were systematically measured at both baseline and the study end point using objective and subjective instruments, and they were classified as either primary or secondary outcomes. (5) The study design was randomized controlled trials (RCTs). To enhance the inclusiveness and generalizability of the review findings, we removed the original restriction of a minimum of 10 participants per intervention group, thereby avoiding unnecessary exclusion of small but methodologically rigorous RCTs that can provide valuable evidence for the evidence synthesis.

The exclusion criteria were as follows: (1) the study population was not explicitly defined as frail (using established criteria such as the Clinical Frailty Scale [37], Fried Phenotype [2], etc) or sarcopenic (using diagnostic guidelines such as EWGSOP (European Working Group on Sarcopenia in Older People) [5], AWGS (Asian Working Group for Sarcopenia) [38], etc), and participants were aged younger than 60 years. (2) Studies using digital devices only for passive monitoring without a structured intervention. (3) The control group received DHIs. (4) No quantitative outcomes related to frailty or sarcopenia (eg, muscle strength, muscle mass, physical performance, and activities of daily living) were reported. (5) The data were incomplete or were presented in a format (eg, medians without IQRs) that made it impossible to calculate effect sizes. (6) Reviews, commentaries, news, conference abstracts, case reports, conference papers, or unpublished manuscripts

Study Selection and Data Extraction

The literature screening was conducted using EndNote software (Clarivate). Two researchers initially screened articles by reviewing titles and abstracts, removing duplicates. If relevance could not be determined from the title or abstract, the full text was assessed. All disagreements were resolved through discussion, with the corresponding author making the final decision when consensus was not reached.

We strictly adhered to the guidelines outlined in the *Cochrane Handbook for Systematic Reviews of Interventions*, designed and optimized a structured data extraction form, and validated its reliability and reproducibility through preextraction in a subset of included studies. Subsequently, 2 independent researchers extracted the following information from all included studies: basic study characteristics (first author, year of publication, country or countries in which the studies were conducted, and study design), disease and diagnostic details (patient population and diagnostic criteria), participant characteristics (sample size, age range, and gender distribution), intervention and control specifics (type of digital health technology, control intervention, and duration of intervention), as well as outcome measures and intention-to-treat (ITT) analysis. Upon completion of data extraction, a third researcher verified the extracted data.

Data extraction followed the predefined criteria below: (1) for studies with multiple publications, the most recent version with the most comprehensive methodological descriptions and outcome reports was prioritized. (2) For multiarm RCTs, data from each intervention arm were extracted independently. (3) In case of missing key outcome data, the corresponding author of the original study was first contacted via email to attempt data acquisition. If data remained unavailable, the study was

XSL•FO
RenderX

excluded from subsequent meta-analyses. (4) Immediate data from the first assessment after intervention completion were prioritized for extraction, with supplementary follow-up data retrieved to support comparative analyses. (5) For studies reporting both ITT analysis and per-protocol analysis, the former was used as the primary data source. (6) When the original literature did not directly report means and SDs, we used meta-analytical tools for data estimation and conversion to perform the necessary transformations or calculations [39]. All extracted data underwent final review by a third researcher.

Additionally, to systematically present the diversity of digital health interventions, 2 researchers classified the interventions using an inductive approach based on the full-text content of the included studies. Discrepancies arising during the classification process were resolved through consensus after discussions involving a third researcher.

Risk of Bias Assessment

Risk of bias was assessed using the revised Cochrane Risk of Bias tool (RoB 2.0), which covers 5 domains: randomization, deviations from interventions, missing data, outcome measurement, and result selection. Two reviewers independently assessed studies, resolving disagreements via discussion. Overall risk of bias was rated as low, some concerns, or high: low if all domains were low risk; some concerns if at least one domain had concerns and none were high risk; high if any domain was high risk or multiple domains had concerns.

Quality of Evidence Evaluation

The quality of evidence for each meta-analysis was appraised using the Grading of Recommendations, Assessment, Development and Evaluations (GRADE) approach [40]. Evidence quality was evaluated based on 5 downgrading factors: risk of bias, inconsistency, indirectness, imprecision, and publication bias, as well as 3 upgrading factors: effect size, impact of plausible confounding, and dose-response relationship. Evidence quality was categorized into 4 levels: high, moderate, low, and very low. Two authors independently performed the GRADE assessment with the GRADEpro GDT tool (GRADE Working Group), with discrepancies resolved by a third author.

Data Synthesis and Analysis

Although the original protocol planned to use Review Manager 5.3 software (Cochrane Collaboration) and model selection based on I^2 statistic and Q test P values, the final data analysis was performed in RStudio (Posit, PBC). A conservative random-effects model, grounded in conceptual assumptions regarding between-study heterogeneity, was used. Effect sizes were pooled using the Hartung-Knapp-Sidik-Jonkman (HKSJ) method [41]. To strengthen methodological rigor and clinical interpretability, additional analytical methods beyond the original protocol were implemented during the analytical refinement phase, including 95% prediction intervals (PIs) and heterogeneity quantification using τ^2.

This systematic review calculated pooled effect sizes using the changes in means and SDs from pre- to postintervention, with the SD of change scores estimated using the equations specified in the *Cochrane Handbook for Systematic Reviews of Interventions*. The calculations were performed as follows:

$$\text{Meanchange}=\text{Meanpost}-\text{Meanpre}$$

$$\text{SDchange}=\text{SDpre2}+\text{SDpost2}-2\times r\times\text{SDpre}\times\text{SDpost}$$

Here, r represents the correlation coefficient between pre- and postintervention values. If this correlation coefficient was not reported in the original studies, a value of r=0.5 was assumed, consistent with the Handbook's recommendations. When studies provided only between-group differences, the corresponding authors were contacted to obtain within-group change data. In instances where SEs were reported instead of SD, the SD was calculated using the formula:

$$\text{SD}=\text{SE}\times n$$

When data were reported as median (IQR), the mean and SD were derived using the sample size–based method described by Hozo et al [42]. Studies that did not report SD, SE, or 95% CI and for which the corresponding authors did not provide these values upon request were excluded from the meta-analysis. Additionally, when outcomes were stratified by sex, age, or body weight status, they were analyzed separately by subgroup.

For outcome measures assessed with identical instruments and units, the weighted mean difference (WMD) was applied to pool effect estimates; for those evaluated using different instruments or scales, the standardized mean difference (SMD) was used to ensure comparability of effect sizes. Given the predominantly small sample sizes across the included studies, Hedges g was selected as the estimator for SMD to correct for small sample bias. Following the Cochrane Handbook, SMD values of 0.2, 0.5, and 0.8 were interpreted as representing small, medium, and large effects, respectively. For multiarm trials, the shared control group was split into approximately equal subgroups, with one subgroup assigned to each intervention arm for comparison, thereby avoiding double counting of participants and preventing unit-of-analysis errors [43,44]. For cluster RCTs, we first evaluated whether the original studies used appropriate statistical methods to account for intracluster correlation [45]. Where such appropriate adjustments were absent from the original reports, an inflated SE approach was applied to approximate a corrected analysis.

To quantify the real-world variability of the intervention effect across different populations and application scenarios, the study also reports PIs [46]. Because conventional PIs perform poorly when the number of studies is small or heterogeneity is low, we used the bootstrap method proposed by Nagashima et al [47] to calculate PIs. This approach draws samples of the heterogeneity parameter τ^2 from the exact distribution of the Cochran Q statistic and combines them with the Hartung-Knapp SE to construct a predictive distribution, thereby improving the robustness of PI estimation and inference even when the number of studies is very limited (K<5) or heterogeneity is moderate. In addition, the Minimal Clinically Important Difference (MCID) reported in previous literature in the field was referenced to evaluate the clinical relevance of the estimated effects.

Given the noticeable differences among the included studies in participant characteristics, intervention protocols, and adherence,

XSL•FO
RenderX

a conservative random effects model was chosen based on the conceptual assumption of a distribution of true effects [48]. For effect size pooling, following the recommendation of the Cochrane Collaboration, we used the HKSJ method for random-effects meta-analysis. By adopting a distribution and adjusting the SE with the Q statistic, the HKSJ approach yields more robust CIs for the effect size compared with the conventional DerSimonian-Laird method, especially when the number of included studies is small and statistical power is limited; it thus better accounts for the inherent uncertainty and imprecision in effect estimation [41]. Restricted maximum likelihood estimation was used to estimate the between-study variance τ^2 so as to more accurately reflect the impact of heterogeneity on the effect estimates. Heterogeneity was assessed using the Q-test, I^2 statistic, and τ^2 metric [49]. When 3 or more studies are included, we also report 95% PIs to indicate the expected distribution range of effect sizes in new settings. The I^2 statistic is reported for comparability with other studies but is interpreted with caution [49,50]. All statistical analyses were performed using R software (version 4.5.2; R Core Team) within the RStudio environment.

When 10 or more studies were available, small-study effects were assessed using funnel plots and the Egger test, which were treated as indicators of such effects rather than sole evidence of publication bias [51,52]. To systematically investigate potential effect modifiers, subgroup analyses were conducted for intervention duration (≤8 wk vs ≥12 wk), control type (exercise vs nonexercise), intervention type, implementation mode (home or remote vs center-based), and synchronization mode (synchronous vs asynchronous). A subgroup was not analyzed if it contained fewer than 4 included studies. All subgroup findings are exploratory and do not imply causality. These subgroup analyses were not specified in the preregistered protocol. All outcomes reported align fully with the preregistered protocol. Sensitivity analysis was performed using the leave-one-out method to evaluate the robustness of the pooled effect estimates. When 10 or more studies were included, funnel plots and the Egger test were used to assess small-study effects; it should be noted that these methods are designed primarily to detect small-study effects rather than to serve as conclusive evidence of publication bias [53,54].

Results

Search Results and Study Selection

In the initial systematic database search, a total of 3506 studies were identified. After the removal of 1895 duplicate entries, 1611 studies remained. Following title and abstract screening, 97 articles underwent full-text review. After full-text assessment, 81 studies were excluded for the following reasons: unavailable full text (n=31), lack of focus on sarcopenia or frailty (n=18), inconsistent intervention measures (n=15), differing outcome indicators (n=8), nonrandomized study design (n=5), and unobtainable or nonconvertible data (n=4). Ultimately, 16 studies met the inclusion criteria for the final analysis [55-70]. The detailed screening process is shown in Figure 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the literature search and study selection.

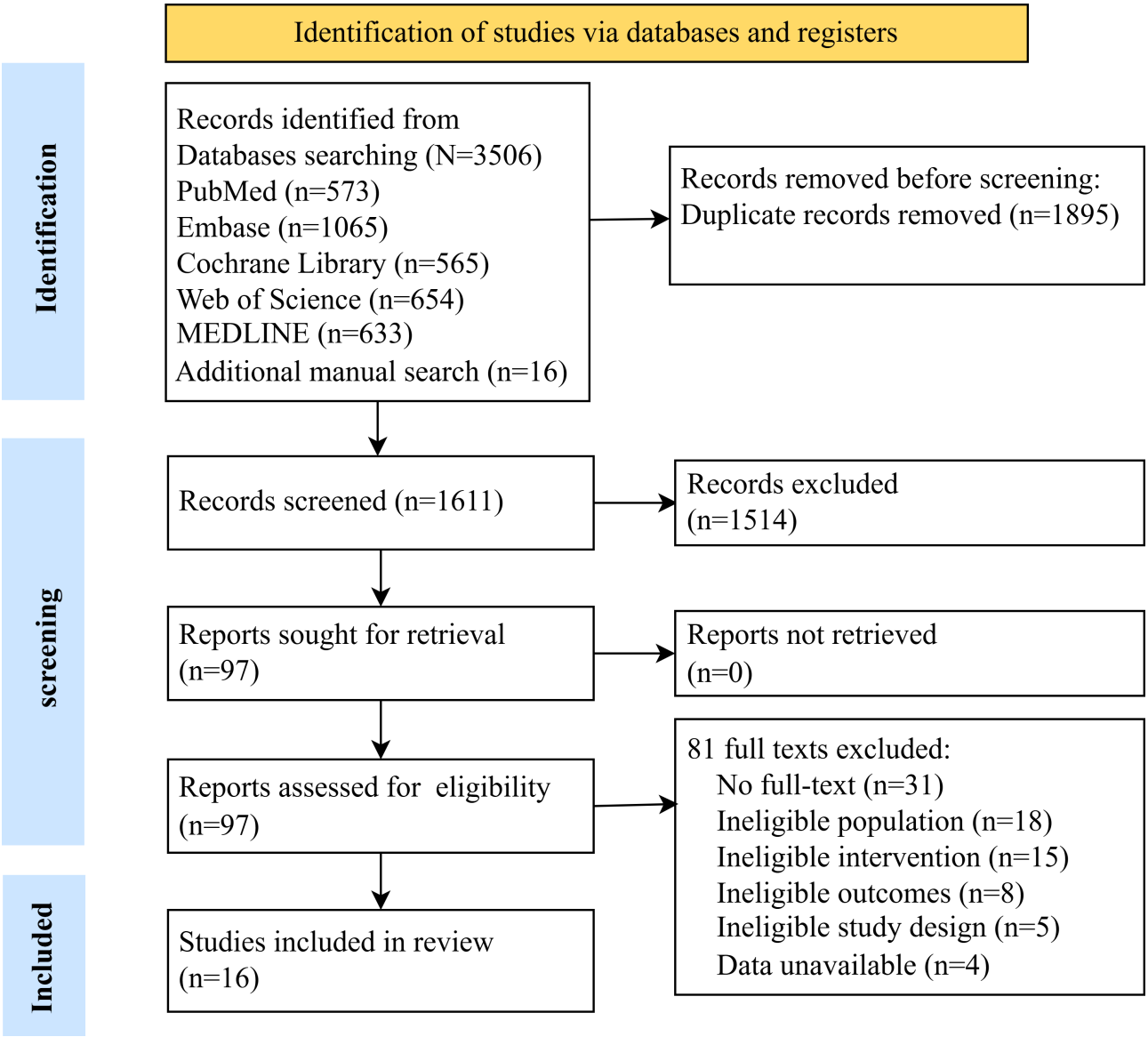

Study Characteristics

This analysis included 16 studies [55-70] involving 1109 participants from 3 geographic regions, with sample sizes ranging from 15 to 201. The geographical distribution of the studies was as follows: 9 studies were conducted in China [55,57,59-62,64,65,70], 2 studies in Korea [63,69], and one each in the United States [56], England [67], and Thailand and Spain [58]; one additional study was a multicenter trial conducted across Italy, Germany, Austria, Spain, England, Belgium, Sweden, Japan, Korea, and Australia [66]. All were RCTs completed between 2012 and 2025. Four trials used a 3-arm design [55,56,63,65], one a 4-arm design [57], and the rest a 2-arm design [58-62,64,66-70]. Seven studies focused on frailty [56,58,60,61,63,66,70], 8 on sarcopenia [55,57,59,64,65,67-69], and 1 on both [62]. Most compared DHIs with control conditions such as usual care, health education, or nondigital exercise.

DHIs mainly included apps, artificial intelligence, virtual reality, exergames (such as those using Kinect and Nintendo Wii), health intervention platforms, wearable devices, and hybrid DHIs. These were implemented in diverse settings such as hospitals, community health centers, and senior day care centers. The duration of interventions across studies varied considerably, ranging from as short as 4 weeks to as long as 12 months. The intervention frequency was most often 2 - 3 sessions per week, each lasting 20 - 60 minutes. Outcome measures included muscle mass, muscle strength, physical performance, and quality of life. Detailed study and population characteristics are in Table 1.

XSL•FO
RenderX

Table . Characteristics of the included studies.[a]

Study (year)	Country	Study design	Disease	Diagnostic criteria	Population (sample size [n, male or female], ages, years [mean, SD])	Intervention (brief introduction)	Intervention format	Control	Intervention frequency, duration, length of follow-up	Outcomes (measures)	ITT[b]
Wei et al (2025) [55]	China	RCT[c]	Sarcopenia	AWGS[d]	EG1[e]: 27 (14/13) Age: 71.57 (7.24) EG2: 25 (13/12) Age: 72.56 (7.76) CG:[f] 24 (14/10) Age: 70.77, (8.27)	EG 1: Home or remote-based individual exercise intervention with real-time AI[g] supervision and guidance EG 2: Center-based group intervention with real-time instructor supervision and feedback	EG 1: AI EG 2: Videoconferencing platform	Exercise	Frequency: 3 times a week Duration: 3 months Follow-up: NA[h]	Muscle mass: ASMI[i] Muscle strength: grip strength Physical performance: gait speed, TUGT[j] QoL[k]: SF-36[l]	NA
Daniel et al (2012) [56]	United States	RCT	Prefrailty	Frailty phenotype	EG1: 8 (3/5) Age: 80, (3.37) EG2: 8 (3/5) Age: 78.13, (5.5) CG: 7 (3/4) Age: 72.6 (4.6)	EG: Center-based group with real-time supervision and guidance by study staff during sessions	Exergame (Nintendo Wii)	Exercise	Frequency: 3 times a week Duration: 15 weeks Follow-up: NA	Physical performance: 30CST[m], TUGT, gait speed, 6MWT[n]	NA

XSL•FO
RenderX

Study (year)	Country	Study design	Disease	Diagnostic criteria	Population (sample size [n, male or female], ages, years [mean, SD])	Intervention (brief introduction)	Intervention format	Control	Intervention frequency, duration, length of follow-up	Outcomes (measures)	ITT[b]
Wang et al (2022) [57]	China	RCT	Sarcopenia	AWGS and EWGSOP[o]	EG1: 50 (8/42) Age: 70.16 (4.32) EG2: 50 (9/41) Age: 68.18 (3.93) EG3: 50 (10/40) Age: 69.72 (3.60) CG: 51 (7/44) Age: 69.88 (3.29)	EG1: Home or remote individual intervention with combined nutrition and exercise, no real-time feedback or supervision EG 2: Home or remote individual intervention with nutrition, no real-time feedback or supervision. EG 3: Home or remote individual intervention with exercise, no real-time feedback or supervision	Mobile app	Education	Frequency: 3 times a week or more Duration: 12 weeks Follow-up: NA	Muscle mass: ASMI, skeletal muscle mass Physical performance: gait speed, timing sitting and standing test, balance Body composition: BMI	NA
Mollà-Casanova et al (2023) [58]	Spain	RCT	Prefrailty and frailty	Frailty phenotype	EG: n=19 (2/17) Age: 71.68 (4.4) CG: 19 (3/16) Age: 72.26 (3.33)	EG: Center-based group exercise with real-time instructor supervision and feedback	Virtual running	Placebo	Frequency: 3 times a week Duration: 8 weeks Follow-up: 4 weeks	Physical performance: 2MWT[p], FTSST[q]	NA

XSL•FO
RenderX

Study (year)	Country	Study design	Disease	Diagnostic criteria	Population (sample size [n, male or female], ages, years [mean, SD])	Intervention (brief introduction)	Intervention format	Control	Intervention frequency, duration, length of follow-up	Outcomes (measures)	ITT[b]
Ho et al (2024) [59]	China	RCT	Sarcopenia	AWGS	EG: 27 (3/24) Age: 70.26, (4.72) CG: 31 (6/25) Age: 74.26 (6.30)	EG: Home or remote individual intervention with nutritional education, no real-time supervision or guidance	Wearable activity tracker	Education	Frequency: daily Duration: 8 weeks Follow-up: NA	Muscle mass: SMI[r] Muscle strength: grip strength Physical performance: FTSST Body composition: BMI	NA
Liao et al (2019) [60]	China	RCT	Prefrailty and frailty	Frailty phenotype	EG: 27 (8/19) Age: 79.6 (8.5) CG: 25 (8/17) Age: 84.1 (5.5)	EG: Center-based group exercise with real-time feedback, supervised by a physical therapist	Exergame	Exercise	Frequency: 3 times a week Duration: 12 weeks Follow-up: none	Muscle strength: grip strength Physical performance: 30CST, single leg stance, TUGT, gait speed	NA
Liu et al (2021) [61]	China	RCT	Frailty	Fried Frailty Index	EG: 22 (5/17) Age: 72.1 (3.7) CG: 18 (1/17) Age: 80.4 (6.83)	EG: Home or remote group exercise intervention with no real-time supervision or feedback	Wearable activity tracker	Exercise	Frequency: 3 times a week Duration: 12 weeks Follow-up: none	• Physical performance: 30CST, TUGT, 2MWT	NA
Tuan et al (2024) [62]	China	RCT	Frailty and sarcopenia	AWGS, osteoporotic fractures index	EG: 30 (10/20) Age: 78.83 (7.71) CG: 30 (11/19) Age: 78.73 (6.82)	EG: Center-based group exercise with real-time supervision and feedback	Exergame	Standard care	Frequency: 2 times a week Duration: 12 weeks Follow-up: NA	Muscle mass: skeletal muscle mass, ASMI Muscle strength: grip strength Physical performance: gait speed QoL: SF-36	YES

XSL•FO
RenderX

Study (year)	Country	Study design	Disease	Diagnostic criteria	Population (sample size [n, male or female], ages, years [mean, SD])	Intervention (brief introduction)	Intervention format	Control	Intervention frequency, duration, length of follow-up	Outcomes (measures)	ITT[b]
Lee (2025) [63]	Korea	RCT	Prefrailty	Frailty phenotype	EG1: 33 (17/13) Age: 84.90 (2.85) EG2: 30 (15/15) Age: 85.18 (3.14) CG: 31 (16/14) Age: 84.81 (2.71)	EG 1: Home or remote group exercise with real-time supervision and feedback provided by physical therapists EG 2: Center-based group exercise with real-time supervision and feedback provided by physical therapists	Videoconferencing app	Education	Frequency: 2 times a week Duration: 8 weeks Follow-up: NA	Physical performance: FTSTS, 30CST, gait speed, TUGT, BBS[s]	NA
Zhang et al (2025) [64]	China	RCT	Sarcopenia	AWGS	EG: 24 (6/18) Age: 70.47 (6.05) CG: 27 (7/20) Age: 69.81 (5.76)	EG: Home or remote individual exercise, with no real-time supervision or feedback	Mobile app	Exercises	Frequency: 3 times a week Duration: 4 weeks Follow-up: NA	Muscle mass: ASMI, skeletal muscle mass Muscle strength: grip strength Physical performance: TUGT, 30CST, 6MWT, Balance	NA

XSL•FO
RenderX

Study (year)	Country	Study design	Disease	Diagnostic criteria	Population (sample size [n, male or female], ages, years [mean, SD])	Intervention (brief introduction)	Intervention format	Control	Intervention frequency, duration, length of follow-up	Outcomes (measures)	ITT[b]
He et al (2024) [65]	China	RCT	Sarcopenia	AWGS	EG1: 24 (NA) Age: 73.67 (4.77) EG2: 23 (NA) Age: 72.26 (4.43) CG: 23 (NA) Age: 70.91 (3.94)	EG 1: Home or remote-based individual exercise intervention with real-time AI supervision and guidance EG 2: Center-based group intervention with real-time instructor supervision and feedback	EG 1: AI EG 2: Videoconferencing app	Exercise	Frequency: 3 times a week Duration: 3 months Follow-up: NA	Muscle mass: ASMI Muscle strength: grip strength Physical performance: gait speed, TUGT QoL: SF-36	NA
Rainero et al (2021) [66]	Italy, Germany, Austria, Spain, England, Belgium, Sweden, Japan, Korea, Australia	RCT	Prefrailty	Frailty phenotype	EG: 101 (30/71) Age: 70.37 (6.15) CG: 100 (23/77) Age: 73.40 (6.57)	EG: Home or remote-based individual multicomponent intervention, including nutritional support, with no real-time supervision or feedback	Health intervention platform	Education	Frequency: NA Duration: 12 months Follow-up: NA	Muscle strength: grip strength Physical performance: TUGT, gait speed, balance Fear of falling: ABC[t] QoL: WHO-QoL[u]	NA
Bailey et al (2024) [67]	England	RCT	Sarcopenia	EWGSOP, Clinical Frailty Scale version 2.0	EG: 23 (NA) Age: 75 (7) CG: 26 (NA) Age: 74 (6)	EG: Home or remote-based individual exercise with no real-time supervision or feedback	Hybrid digital health interventions	Usual care	Frequency: NA Duration: 6 months Follow-up: NA	Muscle strength: grip strength Physical performance: Balance, SPPB[v] QoL: Sarqol[w] Body composition: BMI	NA

XSL•FO
RenderX

Study (year)	Country	Study design	Disease	Diagnostic criteria	Population (sample size [n, male or female], ages, years [mean, SD])	Intervention (brief introduction)	Intervention format	Control	Intervention frequency, duration, length of follow-up	Outcomes (measures)	ITT[b]
Yueny-ongchai-wat et al (2026) [68]	Thailand	RCT	Sarcopenia	AWGS	EG: 23 (4/19) Age: 69.6 (4.6) CG: 23 (5/18) Age: 71.6 (6.3)	EG: Home or remote-based individual exercise with no real-time supervision or feedback	Virtual reality	Education	Frequency: 3 times a week Duration: 12 weeks Follow-up: NA	Muscle mass: SMI Muscle strength: grip strength Physical performance: gait speed	YES
An et al (2025) [69]	Korea	RCT	Sarcopenia	AWGS	EG: 15 (0/15) Age: 76.24 (8.7) CG: 15 (0/15) Age: 74.88 (9.1)	EG: Center-based individual exercise with therapist supervision and real-time feedback provided via mixed reality devices	Mixed reality	Exercise	Frequency: 5 times a week Duration: 4 weeks Follow-up: NA	QoL: SF-12[x]	NA

XSL•FO
RenderX

Study (year)	Country	Study design	Disease	Diagnostic criteria	Population (sample size [n, male or female], ages, years [mean, SD])	Intervention (brief introduction)	Intervention format	Control	Intervention frequency, duration, length of follow-up	Outcomes (measures)	ITT[b]
Liu et al (2025) [70]	China	RCT	Prefrailty and frailty	Frailty phenotype	EG: 15 (7/8) Age: 75.80 (6.99) CG: 15 (8/7) Age: 76.13 (8.16)	EG: Center-based exercise with no real-time supervision or feedback, individual or group not specified	Exergaming (Nintendo Switch sports games)	Exercise	Frequency: 2 times a week Duration: 12 weeks Follow-up: 1 month and 3 months	Muscle mass: ASMI Muscle strength: grip strength Physical performance: SPPB, TUGT	YES

[a]The ASMI and SMI refer to the same indicator (appendicular lean mass divided by height squared). To maintain consistency, the term ASMI is adopted for all subsequent analyses in this study; the 30SSRT and 30CST are identical in their testing protocol. To ensure consistency in reporting and analysis, the metric 30CST will be used uniformly throughout the subsequent analyses of this study.

[b]ITT: intention-to-treat.

[c]RCT: randomized controlled trial.

[d]AWGS: Asian Working Group for Sarcopenia.

[e]EG: experimental group.

[f]CG: control group.

[g]AI: artificial intelligence.

[h]NA: not applicable.

[i]ASMI: Appendicular Skeletal Muscle Mass Index.

[j]TUGT: Timed Up and Go Test.

[k]QoL: Quality of Life.

[l]SF-36: 36-item short form survey.

[m]30CST: 30-second Chair Stand Test.

[n]6MWT: 6-minute walk test.

[o]EWGSOP: European Working Group on Sarcopenia in Older People.

[p]2MWT: 2-minute walk test.

[q]FTSST: Five Times Sit to Stand Test.

[r]SMI: Skeletal Muscle Index.

[s]BBS: Berg Balance Scale.

[t]ABC: activities-specific balance confidence scale.

[u]WHO-QoL: World Health Organization Quality of Life.

[v]SBBP: Short Physical Performance Battery.

[w]SarQol: Sarcopenia Quality of Life.

[x]SF-12: 12-Item short form health survey.

Risk of Bias and Quality of Evidence

The results of the risk of bias assessment are shown in Figure 2. Six were rated as high risk of bias [55-57,65-67], 6 as some concerns [58-60,63,64,68], and 4 as low risk of bias [61,62,69,70]. In Domain 1 (randomization process), 1 study used sequential enrollment for allocation [66] and 2 studies did not provide detailed descriptions of the randomization implementation process, with unbalanced baseline data observed [56,57], resulting in a high risk of randomization bias. In Domain 2 (deviations from intended interventions), 11 studies did not report the use of ITT analysis or modified ITT analysis [55-58,60,63-68], thus indicating some concerns regarding bias. In Domain 3 (missing outcome data), one study exhibited a high attrition rate without reporting specific reasons for withdrawal [51], while another study indicated that participants' dropout might be related to their health status [55]. Both of these scenarios have raised concerns regarding the risk of bias from missing outcome data. For the remaining studies included in the meta-analysis, the number of dropouts was generally

XSL•FO
RenderX

balanced between groups, and the reasons for withdrawal were unrelated to the study outcomes. Therefore, missing data are unlikely to have distorted the true effects, indicating a low risk of bias due to missing outcome data. In Domain 4 (measurement of the outcome), a total of 6 studies were assessed to have some degree of risk of bias [55,56,59,65-67]. Among them, 5 studies were rated as being at high risk of bias because they did not report whether outcome assessors were blinded, and all their outcome measures incorporated subjective assessment items [55,56,65-67]. Another study was judged to have some concerns regarding risk of bias; although it also failed to report the implementation of blinding, all its outcome measures were objective assessments [59]. In Domain 5 (selection of the reported result), all but one study [57] was consistent with their prepublished protocols.

Figure 2. Risk of bias of included randomized controlled trials [55-70].

In GRADE assessments, overall certainty across all comparisons and outcomes is very low to low due to combinations of risk of bias, inconsistency, and imprecision (Multimedia Appendix 2).

Effectiveness of DHIs on Outcomes

Effects of DHIs on Muscle Strength

Ten studies [55,59,60,62,64-68,70] assessed grip strength, one of which [70] included follow-up data. After splitting the sample size of shared control groups [55,65] in multiarm trials, 12 study arms were included in the analysis. Although the heterogeneity statistic I^2=0%, a random-effects model was applied to provide a more conservative and generalizable estimate, given the variability in participant characteristics and intervention content across studies. Compared with controls, DHIs did not significantly improve grip strength (WMD 0.49, 95% CI –0.43 to 1.41, 95% PI –1.43 to 2.42; τ^2=0.16; P=.50; Figure S1 in Multimedia Appendix 3). The 95% CI includes zero, indicating a statistically nonsignificant average effect. The 95% PI also spans zero, reflecting considerable uncertainty in potential future effects. During follow-up, no significant improvement in grip strength was observed with DHIs (WMD 0.49, 95% CI –0.43 to 1.41, 95% PI –1.43 to 2.42; I^2=0%; τ^2=0.17; P=.50; Figure S2 in Multimedia Appendix 3). Again, both the 95% CI and the 95% PI include zero, suggesting statistical nonsignificance and substantial variability in possible future outcomes. However, the reliability of this 95% PI should be interpreted cautiously due to the limited number of available studies. In view of the risks of methodological bias and imprecision, the overall quality of evidence for the effect of DHIs on muscle strength was rated as “low” (Multimedia Appendix 2).

Effects of DHIs on Muscle Mass

Three studies [57,62,64] evaluated skeletal muscle mass. By dividing the sample size of the shared control group in a multiarm trial [57], 5 study arms were ultimately included for analysis. Although the heterogeneity statistic being I^2=0%, a random-effects model was used to provide a more conservative and generalizable effect estimate, given the variations in participant characteristics and intervention content across studies. In the analysis, the DHIs group showed a significant improvement in skeletal muscle mass compared to the control group (WMD 1.01, 95% CI 0.08-1.94, 95% PI –0.95 to 2.96;

XSL•FO
RenderX

τ^2=0; *P*=.47; Figure S3 in Multimedia Appendix 3). While the 95% CI confirms the statistical significance of the average effect, it should be noted that the 95% PI includes zero. This indicates that the beneficial effect may not be replicable in future studies or different populations. The interpretation of this PI, however, is limited by the small number of included studies. Overall, the quality of evidence supporting the effect of DHIs on skeletal muscle mass is rated as "very low" (Multimedia Appendix 2), due to risks of bias, heterogeneity in results, and data imprecision.

Eight studies [55,57,59,62,64,65,68,70] evaluated ASMI, with one study [70] providing longitudinal data. After dividing shared control group samples in multiarm trials [55,57,65], 12 study arms were included in the analysis. Pooled results showed that DHIs did not significantly improve ASMI compared with the control group (WMD 0.18, 95% CI −0.04 to 0.40, 95% PI −0.46 to 0.82; τ^2=0.06; *P*=.03; Figure S4 in Multimedia Appendix 3). Heterogeneity was moderate (I^2=48.2%). The 95% CI included zero, indicating a nonsignificant average effect; similarly, the 95% PI spanned zero, suggesting that DHIs may show no clear benefit or could even be unfavorable for ASMI in future studies or different populations. The limited number of included studies may affect the reliability of this 95% PI. During follow-up, DHIs also showed no significant improvement in ASMI (WMD 0.18, 95% CI −0.04 to 0.4, 95% PI −0.46 to 0.82; I^2=48.4%; τ^2=0.06; *P*=.03; Figure S5 in Multimedia Appendix 3). The 95% CI crosses zero, indicating that the average effect was not statistically significant. The 95% PI also spans zero, reflecting high uncertainty in future effects. Similarly, the reliability of this PI should be interpreted with caution, particularly given the limited number of follow-up studies. Due to risk of bias, inconsistency, and imprecision, the quality of evidence for the effect of DHIs on ASMI was rated as "very low" (Multimedia Appendix 2).

Subgroup analyses revealed that the improvement in ASMI was more pronounced in the ≧12 weeks subgroup (WMD 0.28, 95% CI 0.06-0.50, 95% PI −0.30 to 0.83; I^2=32.6%; τ^2=0.03; *P*=.15) compared with the ≤8 weeks subgroup (SMD −0.2, 95% CI −1.65 to 1.26; I^2=0; τ^2=0; *P*=.48; Multimedia Appendix 4). A test for subgroup differences confirmed a statistically significant distinction between the 2 subgroups (χ^2_1 =9.93; *P*=0.0016). It should be noted that this subgroup analysis was conducted post hoc and lacks the randomization of the original RCTs, thereby precluding strict control for confounding factors. Consequently, the observed difference in ASMI improvement is observational in nature and does not constitute causal evidence that intervention duration drives such a difference.

Effects of DHIs on Physical Performance

Gait Speed

Eight studies [55,57,60,62,63,65,66,68] measured gait speed. By splitting the sample size of shared control groups in multiarm trials [57,63,65], 13 study arms were ultimately included for analysis. The pooled results indicated that, compared with the control group, DHIs significantly improved gait speed (WMD 0.09, 95% CI 0.03-0.15, 95% PI −0.1 to 0.26; τ^2=0; *P*=.07; Figure S6 in Multimedia Appendix 3), with moderate heterogeneity across studies (I^2=39.1%). The 95% CI confirmed the statistical significance of the average effect; however, the 95% PI crossed zero, indicating that DHIs might not improve gait speed in future studies or different populations. The reliability of this 95% PI may be limited by the small number of included studies. Due to risks of bias and imprecision, the quality of evidence for the effect of DHIs on gait speed was rated as "low" Multimedia Appendix 2). The impact of DHIs on gait speed appeared modest.

2MWT

Two studies [58,61] reported data on 2MWT, with one study [58] providing follow-up information. Due to inconsistent measurement units across the 2 studies, SMD was used as the effect measure for meta-analysis. Results showed that compared with the control group, DHIs did not significantly improve the 2MWT (SMD 0.29, 95% CI −0.63 to 1.21; τ^2=0; *P*=.75; Figure S7 in Multimedia Appendix 3), with no observed heterogeneity between studies (I^2=0%). The 95% CI includes zero, indicating that the average intervention effect was not statistically significant. At follow-up, DHIs demonstrated a statistically significant positive effect on the 2MWT (SMD 0.37, 95% CI 0.24-0.50; I^2=0%; τ^2=0; *P*=.96; Figure S8 in Multimedia Appendix 3), representing a small to moderate effect size. However, all of the above analyses were based on only 2 studies [58,61], resulting in a limited evidence base, which constrains the reliability of the findings and precludes subgroup analyses and GRADE assessment.

The 6MWT

Two studies [56,64] reported data on 6MWT. Given the inconsistency in measurement units across the 2 studies, SMD was used as the effect measure to enable meta-analysis. The results indicated that, compared with the control group, DHIs did not significantly improve 6MWT (SMD 0.55, 95% CI −8.30 to 9.40; τ^2=0.78; *P*=.03; Figure S9 in Multimedia Appendix 3), with high heterogeneity observed between studies (I^2=78.4%). The 95% CI includes zero, suggesting that the average intervention effect did not reach statistical significance. Due to the lack of a significant intervention effect on the 6MWT outcome and the limited number of studies, no GRADE assessment, subgroup analysis, or sensitivity analysis was conducted.

The TUGT

Nine studies [55,56,60,61,63-66,70] assessed TUGT, with one study [70] providing follow-up data. After splitting the sample sizes of shared control groups in multiarm trials [55,63,65], 12 study arms were ultimately included in the analysis. The pooled results indicated that, compared with the control group, DHIs significantly improved TUGT (WMD –0.52, 95% CI –1.02 to –0.03, 95% PI –1.93 to 0.85; τ^2=0.35; *P*=.03; Figure S10 in Multimedia Appendix 3). Moderate heterogeneity was observed across studies (I^2=47.5%). The 95% CI confirms the statistical significance of the average effect; however, the wide 95% PI, which crosses the null line, suggests notable variability in intervention effects across future studies or populations, ranging from beneficial to null or even slightly unfavorable. In long-term follow-up, DHIs continued to show a significant improvement in TUGT (WMD −0.52, 95% CI −1.02 to –0.03, 95% PI −1.93 to 0.86; I^2=47.6%; τ^2=0.35; *P*=.03; Figure S11 in Multimedia

XSL•FO
RenderX

Appendix 3). The 95% CI still supports the statistical significance of the average effect, while the 95% PI again crosses zero, reflecting persistent uncertainty in long-term outcomes. Given the limited number of included studies, the reliability of this 95% PI should be interpreted with caution. According to the GRADE evidence assessment, the quality of evidence supporting the effect of DHIs on TUGT improvement was rated as "low," primarily due to risk of bias in some studies and imprecision in results (Multimedia Appendix 2).

The FTSST

Three studies [58,59,63] reported FTSST, with one study [58] providing follow-up data. After splitting the shared control group of a multiarm trial [63], 4 study arms were included in the meta-analysis. The pooled results indicated that, compared with the control group, DHIs did not improve FTSST (WMD −1.63, 95% CI −4.26 to 0.99, 95% PI −8.04 to 4.71; I^2=85.8%; τ^2=2.37; P<.001; Figure S12 in Multimedia Appendix 3), with substantial heterogeneity across studies. The 95% CI included zero, indicating that the mean intervention effect was not statistically significant. The 95% PI also crossed zero, suggesting that DHIs may remain ineffective or potentially unfavorable in future studies or different populations. The reliability of this 95% PI may be limited due to the small number of included studies. During follow-up, the effect direction was largely consistent with the postintervention results (WMD −1.80, 95% CI −4.37 to 0.76, 95% PI −8.19 to 4.45; I^2=85.4%; τ^2=2.28; P<.001; Figure S13 in Multimedia Appendix 3). The 95% CI still included the null value, and the 95% PI again crossed the null line, further supporting the lack of stability and statistical significance of the effect over the longer term. Similarly, given the limited number of studies, the reliability of this prediction interval should be interpreted with caution, as sparse data may reduce the accuracy of estimates of the distribution of future effects. The values of I^2 and τ^2 confirmed substantial variability in true effect sizes across studies, beyond what would be expected from random sampling error alone. Sensitivity analysis showed that removing any single study did not meaningfully alter the overall results, with no statistically significant changes observed (Figure S1 in Multimedia Appendix 5), indicating robustness of the primary meta-analysis findings. Given the nonsignificant effect of the intervention on FTSST and the limited number of studies, no GRADE assessment or subgroup analysis was performed for this outcome.

The 30CST

Five studies [56,60,61,63,64] measured the 30CST. After splitting the shared control group sample size in a multiarm trial [63], 6 study arms were included in the analysis. The pooled analysis showed that DHIs significantly improved 30CST compared with control (WMD 2.19, 95% CI 0.89-3.48, 95% PI –1.59 to 5.66; τ^2=0.09; P=.15; Figure S14 in Multimedia Appendix 3), despite moderate heterogeneity (I^2=38.1%). The 95% PI crossed zero, indicating that DHIs may not improve 30CST in future studies or different populations. The reliability of this 95% PI may be limited by the small number of included studies. Due to risk of bias, inconsistency, and imprecision, the quality of evidence for the effect of DHIs on the 30CST was rated as "very low" (Multimedia Appendix 2).

Subgroup results indicated that the improvement in the 30CST was greater in the ≤8 weeks subgroup (WMD 2.28, 95% CI 0.02-4.54, 95% PI −4.70 to 7.87; I^2=40.1%; τ^2=0.03; P=.19; Multimedia Appendix 4) than in the ≥12 weeks subgroup (WMD 2.26, 95% CI −5.02 to 9.55, 95% PI −12.30 to 16.21; I^2=55.8%; τ^2=4.27; P=.10; Multimedia Appendix 4). However, the between-subgroup difference was not statistically significant (χ^2_1 =0; P=.99). Regarding intervention type, application-based interventions showed a significant effect (WMD 2.28, 95% CI 0.02-4.54, 95% PI −4.48 to 7.93; I^2=40.1%; τ^2=0.03; P=.19; Multimedia Appendix 4). However, no significant differences in effect size were observed between different digital formats (χ^2_2 =0.14; P=.93). In summary, although varying trends in effect sizes were noted across the subgroups analyzed, no statistically significant between-subgroup differences were found in this analysis.

Balance

Six studies assessed balance [57,60,63,64,66,67]. Following adjustment for shared control groups in multiarm trials [57,63], 9 study arms were included in the analysis. The pooled results demonstrated a moderate improvement in balance with DHIs compared to controls (SMD 0.61, 95% CI 0-1.21, 95% PI –0.94 to 2.13; I^2=82.3%; τ^2=0.49; P<.001; Figure S15 in Multimedia Appendix 3). The lower limit of the 95% CI being zero indicates borderline statistical significance. Considerable heterogeneity was observed across studies. The 95% PI included zero, suggesting that the effect may vary substantially across different populations or future studies, and could potentially be null. The reliability of this prediction interval is limited by the small number of studies. Sensitivity analysis showed that the statistical significance of the intervention effect was lost upon removal of any single study, indicating that the findings lack robustness (Figure S2 in Multimedia Appendix 5). Due to risk of bias, inconsistency, and imprecision, the evidence regarding the effect of DHIs on balance was rated as "very low" (Multimedia Appendix 2).

To explore potential moderating factors affecting the improvement of balance function through DHIs, exploratory subgroup analyses were conducted based on intervention duration, comparator type, digital intervention type, delivery mode, and setting. Subgroup results indicated that only the exergaming group showed significant improvement in balance function (SMD 0.83, 95% CI 0.26-1.40; Multimedia Appendix 4); tests for subgroup differences confirmed statistically significant variation across subgroups (χ^2_3 =9.89; P=0.0195). Because subgroup analyses are observational in nature, even when derived from RCTs, they cannot establish causality but may suggest potential moderators. These findings should therefore be interpreted with caution.

The SPPB

Two studies [67,70] reported data on SPPB, with one study [70] providing follow-up information. Results showed that compared with the control group, DHIs did not significantly improve the SPPB (WMD −0.02, 95% CI −0.38 to 0.34, 95% PI −0.83 to 0.75; τ^2=0; P=.90; Figure S16 in Multimedia Appendix 3), with no observed heterogeneity between studies (I^2=0%). The 95% CI includes zero, indicating that the average intervention effect

XSL•FO
RenderX

was not statistically significant. During follow-up, the effect direction was largely consistent with the postintervention results (WMD −0.03, 95% CI −0.65 to 0.58, 95% PI −1.49 to 1.35; I^2=0%; τ^2=0.0001; *P*=.84; Figure S17 in Multimedia Appendix 3). The 95% CI still included the null value, and the 95% PI also crossed zero, indicating no significant long-term effect. However, all of the above analyses were based on only 2 studies, resulting in a limited evidence base that constrains the reliability of the findings and precludes subgroup analyses and GRADE assessment.

Effects of DHIs on Quality of Life

Six studies [55,62,65-67,69] assessed quality of life. Following the splitting of shared control group sample sizes from multiarm trials [55,65], 8 study arms were included in the analysis. Pooled results indicated a small but significant improvement in quality of life with DHIs compared to the control group (SMD 0.16, 95% CI 0.05-0.27, 95% PI 0.04-0.28; τ^2=0; *P*=.97; Figure S18 in Multimedia Appendix 3). The 95% CI did not cross zero, confirming statistical significance. No heterogeneity was observed across studies (I^2=0%). Similarly, the 95% PI did not cross zero, suggesting that the direction of the beneficial effect is likely to remain consistent in future studies or different populations. However, caution is warranted as the reliability of this 95% PI may be limited by the small number of included studies. Due to risk of bias and imprecision, the quality of evidence regarding the effect of DHIs on QoL was rated as "low" (Multimedia Appendix 2).

Effects of DHIs on BMI

Three studies [57,59,67] evaluated BMI. By dividing the sample size of the shared control group in a multiarm trial [57], 5 study arms were ultimately included for analysis. The pooled analysis showed that DHIs did not significantly improve BMI compared with control (WMD 0.09, 95% CI –0.36 to 0.53, 95% PI –0.85 to 1.04; τ^2=0.07; *P*=.16; Figure S19 in Multimedia Appendix 3), despite moderate heterogeneity (I^2=38.6%). The 95% CI included zero, indicating that the mean intervention effect was not statistically significant. The 95% PI also included the null value, indicating that DHIs could be ineffective or even harmful in future research or other populations. Owing to the small number of included studies, the robustness of this 95% PI should be viewed as restricted. Given the nonsignificant effect of the intervention on BMI and the limited number of studies, no GRADE assessment or subgroup analysis was performed for this outcome.

Small-Study Effects Assessment

We conducted small-study effect analyses using funnel plots combined with Egger regression test for outcome measures with at least 10 study arms included. Funnel plots (Figures S1-S4 in Multimedia Appendix 6) were constructed to visually explore potential asymmetry. Visual inspection revealed no marked asymmetry in the funnel plots for any outcome measure. Egger test results indicated no significant small-study effect for any outcome: grip strength (*P*=.92), ASMI (*P*=.54), gait speed (*P*=.59), or TUGT (*P*=.47). It should be noted that observed funnel plot asymmetry may not only reflect publication bias but could also arise from between-study heterogeneity, random error, or methodological differences. Therefore, the *P* values from Egger test in this study serve only as an indicator of potential small-study effects and should not be interpreted as definitive evidence of publication bias. This is especially relevant given the limited number of studies included, as these biases may remain undetected due to insufficient statistical power.

Discussion

Principal Findings

Overview

This study evaluated the effectiveness of DHIs in improving muscle function (muscle strength and muscle mass), physical function, and quality of life among patients with frailty and sarcopenia. Sixteen RCTs involving 1109 participants were included. The results indicated that DHIs significantly improved skeletal muscle mass, gait speed, TUGT, and 30CST. Additionally, a moderate benefit was seen in balance ability, alongside a slight but statistically significant enhancement in quality of life. In contrast, DHIs did not significantly improve grip strength, ASMI, FTSST, SPPB, and BMI. Subgroup analyses revealed that the effect on ASMI might be influenced by intervention duration, and the effect on balance ability could vary by intervention type. The reliability of these findings may be tempered by clinical heterogeneity across studies, including variations in intervention protocols and participant characteristics.

Muscle Function

Multiple diagnostic guidelines for frailty and sarcopenia consistently identify muscle strength, muscle mass, and physical function as core influencing factors of these conditions [5,38,71]. Muscle function primarily includes muscle strength and muscle mass, with grip strength serving as a straightforward and effective metric for assessing muscle strength [72]. Grip strength is also a critical component of the frailty and sarcopenia phenotype [73]. The comprehensive analysis of this study reveals that no significant improvement in grip strength was observed following DHIs, which contradicts previous research findings [29]. The 95% PI for muscle strength ranged from −1.43 to 2.42, indicating that in real-world clinical intervention settings, the efficacy of DHIs may vary significantly due to factors such as intervention protocols and participant characteristics, and in some instances, may exhibit no discernible effect. The lack of improvement in grip strength following DHIs may result from multiple factors. First, the training dose (intensity, frequency, and duration) in current DHIs is likely suboptimal for inducing strength gains in hand and forearm muscles [74]. Optimizing dose-response parameters, including volume, intensity, and progression, is crucial for the efficacy of future interventions. Furthermore, DHIs exhibit significant heterogeneity and frequently lack standardization, particularly in terms of incorporating targeted hand exercises. Furthermore, the majority of DHIs did not incorporate combined nutritional support, such as protein supplementation, which may constrain the anabolic response of muscles. Consequently, future research

XSL•FO
RenderX

should consider the integration of structured exercise regimens with nutritional co-interventions.

For muscle mass outcomes, the meta-analysis demonstrated a significant beneficial effect of DHIs, although the effect on ASMI did not reach statistical significance. This finding is consistent with previous meta-analyses reporting improvements in total skeletal muscle mass but differs from conclusions regarding ASMI [29]. For skeletal muscle mass, the 95% CI did not intersect the null line, indicating a consistent average benefit of DHIs. However, the 95% PI did intersect the null, suggesting uncertainty in clinical application and potential variability in benefits across different studies or settings, including the possibility of no benefit. Similarly, the 95% PI for the ASMI intersected the null, indicating potential inconsistency in real-world effects. Subgroup analysis suggested a potential differential association between intervention duration and ASMI, with interventions lasting 12 weeks or more showing a statistically significant improvement in ASMI. Nonetheless, the 95% PI in this subgroup still intersected the null, underscoring substantial variability in long-term effects and indicating that the benefits of DHIs may vary considerably across populations or contexts. This variability may also be attributed to the small sample size of the subgroup and unmeasured confounding factors. It is important to note that these subgroup comparisons are observational in nature and deviate from the randomized design of the original studies, with the limited sample size further constraining generalizability. Considering these limitations, the observed differences are inadequate to guide clinical decisions concerning the optimal duration of interventions for enhancing ASMI. Future large-scale, high-quality RCTs specifically designed to compare intervention durations are necessary to substantiate these preliminary findings.

In conclusion, the observed variation in the effects of DHIs on these 2 metrics may be attributable to 3 principal considerations. First, there exists an intrinsic difference in the nature of the metrics. Total muscle mass represents an absolute value that is readily influenced by fluctuations in body weight. ASMI is a height-standardized relative measure that more specifically reflects limb muscle status in relation to physical function [75]. Second, many existing DHI programs lack targeted stimulation of limb muscle groups and exhibit considerable interprotocol heterogeneity, which may partially explain their limited effectiveness in improving ASMI. Finally, the current body of evidence has limitations, including small sample sizes, insufficient statistical power, and potential inconsistencies in muscle mass measurement methodologies across studies. These findings highlight the potential value of considering both absolute muscle mass and standardized metrics such as ASMI in clinical assessments to more thoroughly evaluate the impact of DHIs. Future research would benefit from prioritizing large-scale, methodologically rigorous clinical trials that integrate targeted limb training within intervention protocols and standardize the reporting of both absolute and standardized muscle mass measures. Such an approach is expected to help clarify the specific effects of DHIs on different dimensions of muscle health, which could support the development of a more reliable foundation for their targeted clinical application.

Physical Performance

Physical performance is a core feature of frailty and sarcopenia [76] and a key driver of their associations with adverse health outcomes [77]. This study evaluated various physical performance metrics, such as gait speed, 6MWT, 2MWT, TUGT, 30CST, FTSST, SPPB, and balance. The results of the meta-analysis indicated that DHIs led to significant enhancements in gait speed, TUGT, 30CST, and balance among patients with frailty and sarcopenia. The observed improvement in gait speed aligns with findings from previous research [28,30]. The 95% PI for gait speed encompassed the null value, indicating that, despite evidence of a positive intervention effect, substantial uncertainty persists regarding the real-world effectiveness of DHIs. This uncertainty is attributed to factors such as population heterogeneity, diverse clinical settings, and variations in implementation. Additionally, the observed improvement in gait speed (0.09 m/s) did not meet the MCID threshold of 0.10 m/s [78]. This suggests that, although statistically significant, the change may not confer a clinically meaningful functional benefit. Therefore, future large-scale RCTs are necessary to ascertain whether DHIs can consistently achieve the MCID for gait speed and to elucidate their efficacy across various clinical contexts.

In addition to gait speed, the improvement in TUGT was also statistically significant. Notably, the 95% PI for TUGT similarly includes the null value, suggesting potential substantial heterogeneity in its real-world clinical effect. Furthermore, the magnitude of improvement in TUGT (−0.52 s) did not reach the established MCID threshold (2.1 s) [79], indicating that the observed change may not yet translate into clinically meaningful functional benefits. Future large-scale RCTs are needed to determine whether DHIs can consistently achieve the MCID for both gait speed and TUGT, and to clarify their robustness across diverse clinical settings.

For the 30CST, DHIs resulted in a statistically significant improvement of 2.19 repetitions, exceeding the prespecified MCID of 2 repetitions [80]. This suggests a potential for clinically meaningful functional benefit. However, the 95% PI included the null value, indicating substantial uncertainty in the effect's consistency across different populations or clinical settings. Subgroup analyses by intervention duration, delivery mode, and intervention type suggested numerical differences in effect sizes, though none of these comparisons were statistically significant, which does not support these factors as clear determinants of the treatment effect on the 30CST. In clinical interpretation, it is imperative to prioritize the potential clinical benefits of the average effect, while also recognizing its inherent variability. It is inadvisable to make definitive efficacy predictions solely based on current subgroup characteristics. To effectively translate this potential effect into a stable and generalizable clinical practice, future large-scale, stratified clinical trials are necessary. These trials should be designed to ascertain whether the intervention can consistently meet and surpass the clinically significant threshold across diverse populations, thereby offering conclusive evidence for precise clinical recommendations.

XSL•FO
RenderX

DHIs exhibited a notable enhancement in balance. Nonetheless, the 95% PI for balance included the null value, and no MCID for balance has been established in populations with frailty and sarcopenia. Therefore, whether this improvement translates into clinically meaningful benefit remains unclear. Future studies should focus on establishing an MCID for balance in these populations, followed by large-scale trials to determine whether DHIs can consistently achieve clinically significant improvements. Subgroup analysis indicated that exergaming is an effective DHI modality for improving balance in patients with frailty and sarcopenia, which aligns with findings from systematic reviews in healthy older adults [81]. Despite the fact that only one study within the analysis used exergaming, it underscored the potential of highly specific, interactive applications to improve balance. By merging exercise with gameplay, these applications engage patients in physical activity in an enjoyable manner, potentially enhancing mental well-being, exercise enjoyment, and ultimately balance [82]. This interactive approach is likely to increase patient motivation and acceptance, thereby improving adherence and sustained engagement. These factors may contribute to its enhanced effectiveness, though this interpretation should be approached with caution.

Quality of Life

DHIs demonstrated a statistically significant positive impact on the quality of life among patients with frailty and sarcopenia. The 95% CI for quality of life did not intersect the null value, thereby confirming the statistical significance of this positive effect. Furthermore, the 95% PI remained entirely within the positive range and excluded zero, indicating a consistent beneficial effect across various clinical settings or populations. Nonetheless, the effect size was small, necessitating cautious interpretation of its clinical relevance, which should be considered in light of individual circumstances and specific contexts. Given the multidimensional nature of quality of life, future research should aim to develop and evaluate integrated digital health programs that incorporate psychosocial and multidisciplinary components, with the goal of achieving more clinically meaningful improvements beyond the observed small effect [83].

BMI Changes

This study demonstrated that DHIs did not significantly improve BMI, primarily due to the wide range of baseline BMI values among the included populations (BMI 22.26 - 31.70 kg/m^2), which to some extent attenuated the overall effect size of the intervention. Existing evidence indicates that multicomponent exercise interventions incorporating aerobic, resistance, and functional training are more effective [84,85], suggesting that standalone DHIs are limited in weight management owing to the lack of a structured design. Meanwhile, BMI cannot distinguish between muscle and fat mass, thus limiting its evaluative value in older adults with sarcopenia [84]. Therefore, future studies should rigorously control baseline characteristics to reduce heterogeneity, develop structured multimodal intervention programs, and adopt a comprehensive assessment combining BMI with multiple indicators such as muscle mass and body composition.

Follow-Up

The TUGT was the only intervention that demonstrated sustained improvement during the follow-up period. Given the limited number of studies assessing long-term effects, additional research is necessary to elucidate its enduring efficacy [86].

Heterogeneity Across Studies

Exploratory subgroup analyses were undertaken for outcomes exhibiting moderate to high heterogeneity, which was predefined as involving more than 5 studies. The pooled analyses indicated moderate heterogeneity for the TUGT, 30CST, gait speed, and ASMI. In contrast, greater heterogeneity was observed for the 6MWT, FTSST, and balance assessments. Due to an insufficient number of studies, subgroup analyses were not conducted for the 6MWT, FTSST, SPPB, and BMI. The observed heterogeneity is likely attributable to variations in intervention protocols, such as frequency, duration, components, and specific methods, as well as differences in study population characteristics, including age and sample size.

Limitations

This study has several limitations. First, although the search strategy did not restrict language, only English-language publications were ultimately included, which may introduce language bias and reduce the applicability of the findings to non–English-speaking contexts. Second, only RCTs were selected to ensure methodological rigor, which may have excluded non-RCT evidence relevant to the real-world implementation and long-term feasibility of digital health tools. Third, most included studies were conducted in China, limiting the applicability of the results to other health care systems and cultural settings. Finally, substantial heterogeneity in outcome measures and control conditions across studies complicates result synthesis and limits the clarity of conclusions regarding DHI effectiveness. Future research should address these limitations through high-quality, multicenter, and cross-cultural clinical trials to verify the effectiveness of DHIs and clarify their optimal implementation models.

Implications for Clinical Practice

While the overall certainty of evidence remains limited, DHIs have the potential to serve as a scalable and patient-centered complement to traditional rehabilitation methods for individuals experiencing frailty and sarcopenia. To facilitate the integration of DHIs into existing health care systems, clinicians should consider providing structured support across 3 critical domains: accessibility, user engagement, and technical literacy. Regarding accessibility, DHIs should be thoughtfully integrated into clinical pathways, with particular attention given to patients facing functional limitations, mobility challenges, or residing in remote locations, thereby helping to mitigate geographical and physical barriers. Furthermore, the establishment of regular follow-up procedures may aid in addressing ongoing usability concerns and in evaluating sustained outcomes. To improve patient engagement, it is advisable to recommend high-interactivity tools, such as exergames, that align with individual patient interests. For older adults, DHIs should prioritize senior-friendly features, including simplified interfaces and voice assistants, to minimize barriers and enhance

XSL•FO
RenderX

adherence. In terms of technical literacy, clinical teams should develop proficiency in the operation of these tools and in providing patient guidance through comprehensive training programs. Initial support in digital health literacy should be extended to patients, while family members or caregivers should be encouraged to assist individuals with limited technical adaptability, thereby ensuring the sustainability of the intervention.

Conclusion

This study is the first systematic review to examine the impact of DHIs on muscle function (including muscle mass and muscle strength), physical performance, and quality of life among older adults experiencing frailty and sarcopenia. The results indicate that DHIs generally surpass control conditions, resulting in improvements in muscle mass and specific physical performance measures, such as gait speed, TUGT, 30CST, and balance. Additionally, DHIs have a positive impact on quality of life. Nevertheless, these results warrant cautious interpretation. Except for quality of life, all other outcomes exhibited wide PIs encompassing the null value, indicating significant heterogeneity across studies. This variability suggests that statistically significant improvements may not be consistently observed across different populations or clinical settings. The conclusions of this review must be viewed considering its methodological limitations. The wide prediction intervals for muscle mass, gait speed, TUGT, 30CST, and balance, especially the high heterogeneity and instability in balance outcomes, limit the findings' generalizability. Some included RCTs had methodological issues, such as blinding challenges and unclear allocation concealment, potentially leading to an overestimation of DHIs' effects. According to the GRADE framework, the evidence quality for outcomes such as muscle mass, gait speed, TUGT, 30CST, balance, and quality of life is rated very low to low due to concerns about bias, inconsistency, and imprecision, reducing confidence in the intervention's benefits.

Nevertheless, DHIs continue to be regarded as a viable strategy for improving the accessibility and cost-effectiveness of managing frailty and sarcopenia. These interventions are particularly valuable for addressing treatment gaps among populations that face barriers to adequate health care services, such as limited regional resources, time and transportation constraints, or physical functional limitations. Future research endeavors should prioritize the development of a standardized core outcome set for DHIs and the execution of robust, multicenter, large-scale, high-quality RCTs that include nutritional support components. These trials should also assess long-term health benefits and cost-effectiveness. Generating such high-quality evidence will facilitate the standardized application and effective integration of DHIs into the clinical management of frailty and sarcopenia, thereby establishing a foundation for their evolution into scalable and sustainable public health solutions to meet the global prevention and control requirements of these conditions.

Acknowledgments

We thank all original study authors included in this meta-analysis for their contributions to the literature. No generative artificial intelligence (AI) was used in any aspect of the preparation, writing, or revision of this manuscript. All content is solely the work of the authors.

Funding

This research received the support of Ministry of Science and Technology of the People's Republic of China grant 2023YFC3503705 and National Traditional Chinese Medicine Clinical Research Base Scientific Research Special Fund of Henan Provincial Health Commission grant 2024ZXZX1133. The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets used and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors contributed to the study conception and design. TD contributed to writing the original draft. XF handled writing for review and editing. CG and LG managed conceptualization. JH and YC were responsible for methodology. TD and CG conducted formal analysis and investigation. YZ managed resources. JG handled supervision. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Multimedia Appendix 1
The search strategies.
[DOCX File, 21 KB - jmir_v28i1e88374_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Grading of Recommendations, Assessment, Development, and Evaluation evidence profile.
[DOCX File, 19 KB - jmir_v28i1e88374_app2.docx]

Multimedia Appendix 3
Forest plot.
[DOCX File, 601 KB - jmir_v28i1e88374_app3.docx]

Multimedia Appendix 4
Subgroup analysis.
[DOCX File, 2198 KB - jmir_v28i1e88374_app4.docx]

Multimedia Appendix 5
Sensitivity analysis.
[DOCX File, 66 KB - jmir_v28i1e88374_app5.docx]

Multimedia Appendix 6
Funnel plots.
[DOCX File, 51 KB - jmir_v28i1e88374_app6.docx]

Checklist 1
PRISMA checklist.
[DOCX File, 31 KB - jmir_v28i1e88374_app7.docx]

References

1. Liotta G, Ussai S, Illario M, et al. Frailty as the future core business of public health: report of the activities of the A3 Action Group of the European Innovation Partnership on Active and Healthy Ageing (EIP on AHA). Int J Environ Res Public Health 2018 Dec 13;15(12):2843. [doi: 10.3390/ijerph15122843] [Medline: 30551599]
2. Clegg A, Young J, Iliffe S, Rikkert MO, Rockwood K. Frailty in elderly people. Lancet 2013 Mar 2;381(9868):752-762. [doi: 10.1016/S0140-6736(12)62167-9] [Medline: 23395245]
3. Larsson L, Degens H, Li M, et al. Sarcopenia: aging-related loss of muscle mass and function. Physiol Rev 2019 Jan 1;99(1):427-511. [doi: 10.1152/physrev.00061.2017] [Medline: 30427277]
4. Roberts S, Collins P, Identifying RM, Malnutrition M. Frailty and sarcopenia in the community: a narrative review. Nutrients 2021 Jul 5;13(7). [doi: 10.3390/nu13072316] [Medline: 34371823]
5. Cruz-Jentoft AJ, Bahat G, Bauer J, et al. Sarcopenia: revised European consensus on definition and diagnosis. Age Ageing 2019 Jul 1;48(4):601-601. [doi: 10.1093/ageing/afz046] [Medline: 30312372]
6. O'Caoimh R, Sezgin D, O'Donovan MR, et al. Prevalence of frailty in 62 countries across the world: a systematic review and meta-analysis of population-level studies. Age Ageing 2021 Jan 8;50(1):96-104. [doi: 10.1093/ageing/afaa219] [Medline: 33068107]
7. Petermann-Rocha F, Balntzi V, Gray SR, et al. Global prevalence of sarcopenia and severe sarcopenia: a systematic review and meta-analysis. J Cachexia Sarcopenia Muscle 2022 Feb;13(1):86-99. [doi: 10.1002/jcsm.12783] [Medline: 34816624]
8. Zhang X, Zhang W, Wang C, Tao W, Dou Q, Yang Y. Sarcopenia as a predictor of hospitalization among older people: a systematic review and meta-analysis. BMC Geriatr 2018 Aug 22;18(1):188. [doi: 10.1186/s12877-018-0878-0] [Medline: 30134867]
9. Tufan A, Tolu T, Senturk Durmus N, Alkac C, Can B. FRAIL Scale: an independent predictor of in-hospital mortality among older adults. Eur Rev Med Pharmacol Sci 2023 Nov;27(21):10396-10402. [doi: 10.26355/eurrev_202311_34313] [Medline: 37975362]
10. Wei K, Nyunt MSZ, Gao Q, Wee SL, Yap KB, Ng TP. Association of frailty and malnutrition with long-term functional and mortality outcomes among community-dwelling older adults: results from the Singapore Longitudinal Aging Study 1. JAMA Netw Open 2018 Jul 6;1(3):e180650. [doi: 10.1001/jamanetworkopen.2018.0650] [Medline: 30646023]
11. Woolford SJ, Sohan O, Dennison EM, Cooper C, Patel HP. Approaches to the diagnosis and prevention of frailty. Aging Clin Exp Res 2020 Sep;32(9):1629-1637. [doi: 10.1007/s40520-020-01559-3] [Medline: 32356135]
12. Chen YC, Chen WC, Liu CW, et al. Is moderate resistance training adequate for older adults with sarcopenia? A systematic review and network meta-analysis of RCTs. Eur Rev Aging Phys Act 2023 Nov 29;20(1):22. [doi: 10.1186/s11556-023-00333-4] [Medline: 38030985]

XSL•FO
RenderX

13. Xue Q, Xu DR, Cheng TC, Pan J, Yip W. The relationship between hospital ownership, in-hospital mortality, and medical expenses: an analysis of three common conditions in China. Arch Public Health 2023 Feb 10;81(1):19. [doi: 10.1186/s13690-023-01029-y] [Medline: 36765426]
14. Wang Q, Jiang Q, Yang Y, Pan J. The burden of travel for care and its influencing factors in China: an inpatient-based study of travel time. J Transp Health 2022 Jun;25:101353. [doi: 10.1016/j.jth.2022.101353]
15. Izquierdo M, de Souto Barreto P, Arai H, et al. Global consensus on optimal exercise recommendations for enhancing healthy longevity in older adults (ICFSR). J Nutr Health Aging 2025 Jan;29(1):100401. [doi: 10.1016/j.jnha.2024.100401] [Medline: 39743381]
16. Chen LK, Hsiao FY, Akishita M, et al. A focus shift from sarcopenia to muscle health in the Asian Working Group for Sarcopenia 2025 Consensus Update. Nat Aging 2025 Nov;5(11):2164-2175. [doi: 10.1038/s43587-025-01004-y] [Medline: 41188603]
17. Wang M, Lu X, Du Y, et al. Digital health governance in China by a whole-of-society approach. NPJ Digit Med 2025 Aug 2;8(1):496. [doi: 10.1038/s41746-025-01876-9] [Medline: 40750663]
18. Fatehi F, Samadbeik M, Kazemi A. What is digital health? Review of definitions. Stud Health Technol Inform 2020 Nov 23;275(67-71):67-71. [doi: 10.3233/SHTI200696] [Medline: 33227742]
19. Bhavnani SP, Narula J, Sengupta PP. Mobile technology and the digitization of healthcare. Eur Heart J 2016 May 7;37(18):1428-1438. [doi: 10.1093/eurheartj/ehv770] [Medline: 26873093]
20. Ou J, Li N, He H, He J, Zhang L, Jiang N. Detecting muscle fatigue among community-dwelling senior adults with shape features of the probability density function of sEMG. J Neuroeng Rehabil 2024 Nov 4;21(1):196. [doi: 10.1186/s12984-024-01497-5] [Medline: 39497122]
21. Chen Y, Yu W, Benali A, Lu D, Kok SY, Wang R. Towards human-like walking with biomechanical and neuromuscular control features: personalized attachment point optimization method of cable-driven exoskeleton. Front Aging Neurosci 2024;16:1327397. [doi: 10.3389/fnagi.2024.1327397] [Medline: 38371400]
22. Luo J, Ahmad SF, Alyaemeni A, et al. Role of perceived ease of use, usefulness, and financial strength on the adoption of health information systems: the moderating role of hospital size. Humanit Soc Sci Commun 2024;11(1). [doi: 10.1057/s41599-024-02976-9]
23. Wu CE, Manga YB. Impact of wearable-assisted walking on sarcopenia and body composition in older adults. BMC Geriatr 2025 Jul 2;25(1):466. [doi: 10.1186/s12877-025-06142-x] [Medline: 40604455]
24. Kannan L, Sahu U, Subramaniam S, et al. Gaming-based tele-exercise program to improve physical function in frail older adults: feasibility randomized controlled trial. J Med Internet Res 2024 Nov 27;26:e56810. [doi: 10.2196/56810] [Medline: 39602215]
25. Peterson MJ, Sloane R, Cohen HJ, Crowley GM, Pieper CF, Morey MC. Effect of telephone exercise counseling on frailty in older veterans: project LIFE. Am J Mens Health 2007 Dec;1(4):326-334. [doi: 10.1177/1557988307306153] [Medline: 19482814]
26. Blasiak A, Sapanel Y, Leitman D, et al. Omnichannel communication to boost patient engagement and behavioral change with digital health interventions. J Med Internet Res 2022 Nov 16;24(11):e41463. [doi: 10.2196/41463] [Medline: 36383427]
27. Han HW, Park SW, Kim DY, et al. e-Health interventions for older adults with frailty: a systematic review. Ann Rehabil Med 2023 Oct;47(5):348-357. [doi: 10.5535/arm.23090] [Medline: 37907226]
28. Lee YH, Lin CH, Wu WR, Chiu HY, Huang HC. Virtual reality exercise programs ameliorate frailty and fall risks in older adults: a meta-analysis. J Am Geriatr Soc 2023 Sep;71(9):2946-2955. [doi: 10.1111/jgs.18398] [Medline: 37165743]
29. Chen W, Yao J, Wang H, Jia S, Zhu X, Mao L. Effects of digital health interventions on muscle mass, muscle strength, and physical function in older adults with sarcopenia: a systematic review and meta-analysis. Front Public Health 2025;13(1711514):41383324. [doi: 10.3389/fpubh.2025.1711514]
30. Makizako H, Shiratsuchi D, Akaida S, et al. Effects of digital-based interventions on the outcomes of the eligibility criteria for sarcopenia in healthy older adults: a systematic review and meta-analysis. Ageing Res Rev 2025 Feb;104(102663):102663. [doi: 10.1016/j.arr.2025.102663] [Medline: 39814237]
31. Vavasour G, Giggins OM, Doyle J, Kelly D. How wearable sensors have been utilised to evaluate frailty in older adults: a systematic review. J Neuroeng Rehabil 2021 Jul 8;18(1):112. [doi: 10.1186/s12984-021-00909-0] [Medline: 34238323]
32. Tay T, Grailey K, Chen F, et al. The impact of digitally-enabled interventions on frailty and other age-related outcomes - systematic review and meta-analysis. Digit Health 2026;12:20552076251410997. [doi: 10.1177/20552076251410997] [Medline: 41623726]
33. Davies B, García F, Ara I, Artalejo FR, Rodriguez-Mañas L, Walter S. Relationship between sarcopenia and frailty in the toledo study of healthy aging: a population based cross-sectional study. J Am Med Dir Assoc 2018 Apr;19(4):282-286. [doi: 10.1016/j.jamda.2017.09.014] [Medline: 29079029]
34. Paravlic A, Šlosar L, Abazovic E, Marusic U. Effects of digital-based interventions on muscular strength in adults: a systematic review, meta-analysis and meta-regression of randomized controlled trials with quality of evidence assessment. Ann Med 2023 Dec;55(1):2230886. [doi: 10.1080/07853890.2023.2230886] [Medline: 37452683]
35. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009 Jul 21;339(jul21 1):b2535. [doi: 10.1136/bmj.b2535] [Medline: 19622551]

XSL•FO
RenderX

36. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc 2021 Apr 1;109(2):174-200. [doi: 10.5195/jmla.2021.962] [Medline: 34285662]
37. Walker E, Winter R, Hodgson LE. Frailty in older adults admitted to hospital: a six-year dual-centre retrospective study of over 53 000 clinical frailty scale assessments. Age Ageing 2025 May 31;54(6):afaf137. [doi: 10.1093/ageing/afaf137] [Medline: 40459342]
38. Chen LK, Woo J, Assantachai P, et al. Asian Working Group for Sarcopenia: 2019 consensus update on sarcopenia diagnosis and treatment. J Am Med Dir Assoc 2020 Mar;21(3):300-307. [doi: 10.1016/j.jamda.2019.12.012] [Medline: 32033882]
39. Chi KY, Li MY, Chen C, Kang E, Cochrane Taiwan. Ten circumstances and solutions for finding the sample mean and standard deviation for meta-analysis. Syst Rev 2023 Apr 1;12(1):62. [doi: 10.1186/s13643-023-02217-1] [Medline: 37005690]
40. Higgins JPT, Altman DG, Gøtzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011 Oct 18;343(oct18 2):d5928. [doi: 10.1136/bmj.d5928] [Medline: 22008217]
41. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
42. Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. BMC Med Res Methodol 2005 Apr 20;5(13):13. [doi: 10.1186/1471-2288-5-13] [Medline: 15840177]
43. Rücker G, Cates CJ, Schwarzer G. Methods for including information from multi-arm trials in pairwise meta-analysis. Res Synth Methods 2017 Dec;8(4):392-403. [doi: 10.1002/jrsm.1259] [Medline: 28759708]
44. da Costa BR, Juni P. Systematic reviews and meta-analyses of randomized trials: principles and pitfalls. Eur Heart J 2014 Dec 14;35(47):3336-3345. [doi: 10.1093/eurheartj/ehu424] [Medline: 25416325]
45. Whiting-O'Keefe QE, Henke C, Simborg DW. Choosing the correct unit of analysis in Medical Care experiments. Med Care 1984 Dec;22(12):1101-1114. [doi: 10.1097/00005650-198412000-00005] [Medline: 6513619]
46. IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open 2016 Jul 12;6(7):e010247. [doi: 10.1136/bmjopen-2015-010247] [Medline: 27406637]
47. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019 Jun;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]
48. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods 2010 Apr;1(2):97-111. [doi: 10.1002/jrsm.12] [Medline: 26061376]
49. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. Res Synth Methods 2017 Mar;8(1):5-18. [doi: 10.1002/jrsm.1230] [Medline: 28058794]
50. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023 Dec;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
51. Tang JL, Liu JL. Misleading funnel plot for detection of bias in meta-analysis. J Clin Epidemiol 2000 May;53(5):477-484. [doi: 10.1016/s0895-4356(99)00204-8] [Medline: 10812319]
52. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634. [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
53. Lau J, Ioannidis JPA, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. BMJ 2006 Sep 16;333(7568):597-600. [doi: 10.1136/bmj.333.7568.597] [Medline: 16974018]
54. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 6;327(7414):557-560. [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
55. Wei M, Meng D, He S, et al. Investigating the efficacy of AI-enhanced telerehabilitation in sarcopenic older individuals. Eur Geriatr Med 2025 Feb;16(1):115-123. [doi: 10.1007/s41999-024-01082-y] [Medline: 39453567]
56. Daniel K. Wii-hab for pre-frail older adults. Rehabil Nurs 2012;37(4):195-201. [doi: 10.1002/rnj.25] [Medline: 22744992]
57. Wang Z, Xu X, Gao S, et al. Effects of internet-based nutrition and exercise interventions on the prevention and treatment of sarcopenia in the elderly. Nutrients 2022 Jun 14;14(12):2458. [doi: 10.3390/nu14122458] [Medline: 35745187]
58. Mollà-Casanova S, Muñoz-Gómez E, Sempere-Rubio N, et al. Effect of virtual running with exercise on functionality in pre-frail and frail elderly people: randomized clinical trial. Aging Clin Exp Res 2023 Jul;35(7):1459-1467. [doi: 10.1007/s40520-023-02414-x] [Medline: 37188994]
59. Ho MH, Peng CY, Liao Y, Yen HY. Efficacy of a wearable activity tracker with step-by-step goal-setting on older adults' physical activity and sarcopenia indicators: clustered trial. J Med Internet Res 2024 Nov 1;26:e60183. [doi: 10.2196/60183] [Medline: 39486024]
60. Liao YY, Chen IH, Wang RY. Effects of Kinect-based exergaming on frailty status and physical performance in prefrail and frail elderly: a randomized controlled trial. Sci Rep 2019 Jun 27;9(1):9353. [doi: 10.1038/s41598-019-45767-y] [Medline: 31249332]
61. Liu JYW, Kwan RYC, Yin YH, Lee PH, Siu JYM, Bai X. Enhancing the physical activity levels of frail older adults with a wearable activity tracker-based exercise intervention: a pilot cluster randomized controlled trial. Int J Environ Res Public Health 2021 Sep 30;18(19):10344. [doi: 10.3390/ijerph181910344] [Medline: 34639644]

XSL•FO
RenderX

62. Tuan SH, Chang LH, Sun SF, Li CH, Chen GB, Tsai YJ. Assessing the clinical effectiveness of an exergame-based exercise training program using Ring Fit Adventure to prevent and postpone frailty and sarcopenia among older adults in rural long-term care facilities: randomized controlled trial. J Med Internet Res 2024 Jul 18;26:e59468. [doi: 10.2196/59468] [Medline: 39024000]
63. Lee K. Effects of remote exercise on physical function in pre-frail older adults: a randomized controlled trial. Med Sci Monit 2025 Jan 28;31:e947105. [doi: 10.12659/MSM.947105] [Medline: 39871464]
64. Zhang L, Ge Y, Zhao W, et al. A 4-week mobile app-based telerehabilitation program vs conventional In-person rehabilitation in older adults with sarcopenia: randomized controlled trial. J Med Internet Res 2025 Jan 24;27:e67846. [doi: 10.2196/67846] [Medline: 39854716]
65. He S, Meng D, Wei M, Guo H, Yang G, Wang Z. Proposal and validation of a new approach in tele-rehabilitation with 3D human posture estimation: a randomized controlled trial in older individuals with sarcopenia. BMC Geriatr 2024 Jul 8;24(1):586. [doi: 10.1186/s12877-024-05188-7] [Medline: 38977995]
66. Rainero I, Summers MJ, Monter M, et al. The My Active and Healthy Aging ICT platform prevents quality of life decline in older adults: a randomised controlled study. Age Ageing 2021 Jun 28;50(4):1261-1267. [doi: 10.1093/ageing/afaa290] [Medline: 33480986]
67. Bailey DP, Harper JH, Kilbride C, et al. The frail-LESS (LESS sitting and sarcopenia in frail older adults) remote intervention to improve sarcopenia and maintain independent living via reductions in sedentary behaviour: findings from a randomised controlled feasibility trial. BMC Geriatr 2024 Sep 9;24(1):747. [doi: 10.1186/s12877-024-05310-9] [Medline: 39251904]
68. Yuenyongchaiwat K, Chitjamnogchai C, Sermsinsaithong N, et al. Benefit of home-based virtual reality aerobic exercise combined with resistance training for reversing sarcopenia and reducing depression in community-dwelling older adults with sarcopenia: a randomized control trial. Arch Gerontol Geriatr 2026 Mar;142:106108. [doi: 10.1016/j.archger.2025.106108] [Medline: 41338160]
69. An Y, Min S, Park C. Mixed reality-based physical therapy in older adults with sarcopenia: preliminary randomized controlled trial. JMIR Serious Games 2025 Aug 1;13:e76357. [doi: 10.2196/76357] [Medline: 40749218]
70. Liu TW, Wong BYM, Yam TTT, et al. Effects of exergaming with a resistance component versus traditional resistance training on sarcopenia in pre-frail and frail nursing home residents: a pilot randomized controlled trial. Eur Geriatr Med 2025 Sep 11;PMID. [doi: 10.1007/s41999-025-01294-w] [Medline: 40952658]
71. Dent E, Morley JE, Cruz-Jentoft AJ, et al. Physical frailty: ICFSR International clinical practice guidelines for identification and management. J Nutr Health Aging 2019;23(9):771-787. [doi: 10.1007/s12603-019-1273-z] [Medline: 31641726]
72. Cruz-Jentoft AJ, Sayer AA. Sarcopenia. Lancet 2019 Jun 29;393(10191):2636-2646. [doi: 10.1016/S0140-6736(19)31138-9] [Medline: 31171417]
73. Dodds RM, Syddall HE, Cooper R, Kuh D, Cooper C, Sayer AA. Global variation in grip strength: a systematic review and meta-analysis of normative data. Age Ageing 2016 Mar;45(2):209-216. [doi: 10.1093/ageing/afv192] [Medline: 26790455]
74. Ainsworth BE, Haskell WL, Herrmann SD, et al. 2011 Compendium of Physical Activities: a second update of codes and MET values. Med Sci Sports Exerc 2011 Aug;43(8):1575-1581. [doi: 10.1249/MSS.0b013e31821ece12] [Medline: 21681120]
75. Hung CD, Chang CW, Tu HP, et al. The association between polypharmacy, low appendicular skeletal muscle mass index, and physical performance: a cross-sectional study. Gerontology 2026 Feb 26;41746861:1-15. [doi: 10.1159/000551103] [Medline: 41746861]
76. Muñoz-Redondo E, Morgado-Pérez A, Pérez-Sáez MJ, et al. New perspectives on frailty in light of the Global Leadership Initiative on Malnutrition, the Global Leadership Initiative on Sarcopenia, and the WHO's concept of intrinsic capacity: a narrative review. Maturitas 2023 Nov;177:107799. [doi: 10.1016/j.maturitas.2023.107799] [Medline: 37499428]
77. Buchman AS, Leurgans SE, Wang T, et al. Motor function is the primary driver of the associations of sarcopenia and physical frailty with adverse health outcomes in community-dwelling older adults. PLoS One 2021;16(2):e0245680. [doi: 10.1371/journal.pone.0245680] [Medline: 33529220]
78. Bohannon RW, Glenney SS. Minimal clinically important difference for change in comfortable gait speed of adults with pathology: a systematic review. J Eval Clin Pract 2014 Aug;20(4):295-300. [doi: 10.1111/jep.12158] [Medline: 24798823]
79. Shen Y, Shi Q, Nong K, et al. Exercise for sarcopenia in older people: a systematic review and network meta-analysis. J Cachexia Sarcopenia Muscle 2023 Jun;14(3):1199-1211. [doi: 10.1002/jcsm.13225] [Medline: 37057640]
80. Wright AA, Cook CE, Baxter GD, Dockerty JD, Abbott JH. A comparison of 3 methodological approaches to defining major clinically important improvement of 4 performance measures in patients with hip osteoarthritis. J Orthop Sports Phys Ther 2011 May;41(5):319-327. [doi: 10.2519/jospt.2011.3515] [Medline: 21335930]
81. Fang Q, Ghanouni P, Anderson SE, et al. Effects of exergaming on balance of healthy older adults: a systematic review and meta-analysis of randomized controlled trials. Games Health J 2020 Feb;9(1):11-23. [doi: 10.1089/g4h.2019.0016] [Medline: 31800322]
82. Lillo-Urzúa P, Ugarte-Llanten J, Carreño-Zilmann G, et al. E-ACTIVE AGING study protocol: evaluating an exergame-based and multicomponent exercise program for community-dwelling older adults at risk of falling. Front Physiol 2025;16(1691454):1691454. [doi: 10.3389/fphys.2025.1691454] [Medline: 41415308]

XSL•FO
RenderX

83. Vaarama M. Care-related quality of life in old age. Eur J Ageing 2009 Jun;6(2):113-125. [doi: 10.1007/s10433-009-0115-y] [Medline: 28798598]
84. Wang L, Chao J, Zhang N, et al. Prevalence and factors associated with sarcopenia in community-dwelling older adults: a systematic review and meta-analysis. Gerontology 2026;72(2):110-125. [doi: 10.1159/000549808] [Medline: 41343409]
85. Colleluori G, Aguirre L, Phadnis U, et al. Aerobic plus resistance exercise in obese older adults improves muscle protein synthesis and preserves myocellular quality despite weight loss. Cell Metab 2019 Aug 6;30(2):261-273. [doi: 10.1016/j.cmet.2019.06.008] [Medline: 31279675]
86. Clarkson P, Stephenson A, Grimmett C, et al. Digital tools to support the maintenance of physical activity in people with long-term conditions: a scoping review. Digit Health 2022;8:20552076221089778. [doi: 10.1177/20552076221089778] [Medline: 35433017]

Abbreviations

2MWT: 2-minute walk test
30CST: 30-second Chair Stand Test
6MWT: 6-minute walk test
ASMI: Appendicular Skeletal Muscle Mass Index
AWGS: Asian Working Group for Sarcopenia
DHI: digital health intervention
EWGSOP: European Working Group on Sarcopenia in Older People
FTSST: Five Times Sit to Stand Test
GRADE: Grading of Recommendations, Assessment, Development, and Evaluation
HKSJ: Hartung-Knapp-Sidik-Jonkman
ITT: intention-to-treat
MCID: Minimal Clinically Important Difference
MeSH: Medical Subject Headings
PI: prediction interval
PICOS: Population, Intervention, Comparison, Outcomes, and Study Design
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses search extension
RCT: randomized controlled trial
RoB 2.0: Cochrane Risk of Bias tool
SMD: standardized mean difference
SPPB: Short Physical Performance Battery
TUGT: Timed Up and Go Test
WMD: weighted mean difference

Edited by S Brini; submitted 24.Nov.2025; peer-reviewed by MZ Hossain, WG Hussain, X Ding, Z Su; revised version received 23.Mar.2026; accepted 23.Mar.2026; published 11.May.2026.

Please cite as:
Dai T, Guo C, Gao L, Huang J, Chen Y, Zhang Y, Gao J, Feng X
Effectiveness of Digital Health Interventions in Older Adults With Frailty and Sarcopenia: Systematic Review and Meta - Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e88374
URL: https://www.jmir.org/2026/1/e88374
doi:10.2196/88374

© Ting Dai, Changsheng Guo, Lingli Gao, Jinyu Huang, Yan Chen, Yujie Zhang, Jing Gao, Xiaodong Feng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI-Based Automation for Medication Reconciliation: Scoping Review

Juan Pablo Tabja Bortesi[1,2], HBSc; Maria P Becerra[1], HBSc; Jonathan Ranisau[1], MASc; Bonnie Wen[3]; Praveen Nadesan[4], HBSc; P J Devereaux[4,5,6], MD, PhD; Michael McGillion[5,7], PhD; Jeremy Petch[1,5,6,8], PhD

[1]Centre for Data Science and Digital Health, Hamilton Health Sciences, Hamilton, ON, Canada
[2]School of Epidemiology and Public Health, University of Ottawa, Ottawa, ON, Canada
[3]Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada
[4]Michael G. DeGroote School of Medicine, Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada
[5]Population Health Research Institute, Hamilton, ON, Canada
[6]Division of Cardiology, Department of Medicine, McMaster University, Hamilton, ON, Canada
[7]School of Nursing, Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada
[8]Institute for Health Policy, Management and Evaluation, University of Toronto, 155 College St 4th Floor, Toronto, ON, Canada

Corresponding Author:
Jeremy Petch, PhD
Centre for Data Science and Digital Health, Hamilton Health Sciences, Hamilton, ON, Canada

Abstract

Background: Medication reconciliation (MedRec) has the potential to improve patient safety by enhancing the continuity of medication information across settings. MedRec involves 3 core tasks: the creation of a best possible medication history, the identification of medication discrepancies among medication lists, and the resolution of medication discrepancies. While artificial intelligence (AI) has the potential to improve MedRec, existing reviews have not identified the ways in which researchers have used AI to facilitate MedRec tasks and their constituent subtasks or the level of automation achieved.

Objective: This scoping review aimed to map how previous research has applied AI to MedRec tasks and subtasks and assess the extent of automation achieved.

Methods: We searched MEDLINE, Embase, Web of Science, IEEE Xplore, and Compendex in June 2024 for studies that used AI to support a MedRec task or subtask, excluding entirely rule-based tools or studies focused on other aspects of medication management. After screening 2345 unique records, we conducted backward citation searching of studies included at the full-text stage, identifying an additional 795 unique records. We used a 4-stage model of human information processing as a structural lens to guide our considerations of automation, mapping the core tasks of MedRec onto this model.

Results: A total of 94 studies met the inclusion criteria. All studies addressed subtasks related to the creation of a best possible medication history. Only 2.1% (n=2) of the studies also addressed the identification of discrepancies. Thus, the highest stage of automated information processing achieved was information analysis, although most studies (92/94, 97.9%) only automated information acquisition steps. Most studies (67/94, 71.3%) used free-text clinical notes from the electronic health record, although a significant proportion (21/94, 22.3%) used images of pills or images of other medication-related items. Studies using text-based data used a variety of machine learning methods (eg, recurrent neural networks, conditional random fields, support vector machines, and transformers), whereas those that leveraged images typically used convolutional neural networks. Most studies (61/94, 64.9%) used publicly available data from benchmarking datasets (eg, n2c2 2022) and were strictly model development studies, with only 1.1% (n=1) being usability studies.

Conclusions: This is the first review to consider the role of AI in the automation of MedRec tasks, offering a basis for prioritizing future development efforts. Current applications of AI to automate MedRec tasks are preliminary, with most work focusing on the extraction of medication information and limited to proof-of-concept model development. Future work should consider addressing infrastructural barriers to the AI-based automation of MedRec tasks (eg, data incompleteness in sources of medication information) and exploring approaches to automate discrepancy resolution. Beyond developing models, there is also a need to implement them in tools and evaluate them in real-world contexts.

Trial Registration: OSF Registries osf.io/n64u3; https://osf.io/n64u3

(J Med Internet Res 2026;28:e86760) doi:10.2196/86760

XSL•FO
RenderX

KEYWORDS

medication reconciliation; artificial intelligence; AI; health IT; machine learning; patient safety; continuity of care

Introduction

There is a potential for error at every stage of the medication use cycle, which can often lead to serious harms for patients. This avoidable harm is substantial: across care settings, a substantial proportion of adverse drug events (ADEs) have been estimated to be preventable, although estimates vary by context and definition [1-3]. The economic impact of ADEs is considerable, with estimates exceeding hundreds of millions of dollars per year across acute and primary health care systems [4]. Given the scale of the problem, in 2017, the World Health Organization announced *Medication Without Harm* as its third Global Patient Safety Challenge, outlining strategies to reduce avoidable medication-related harm [5]. One of the chief causes of medication errors is inaccuracies in medication lists used throughout a patient's care.

Medication reconciliation (MedRec) is an important procedure for identifying such inaccuracies. MedRec generally includes the creation of a best possible medication history (BPMH) that is compared to a list of medications ordered by a health care professional (eg, a physician, nurse, or pharmacist) to address unintended discrepancies [6]. This typically occurs during care transitions, such as hospital admission and discharge, when medication errors frequently occur [7-10]. MedRec can also be performed in primary care or outpatient settings, often with patients at risk of readmission following discharge [11]. Recognizing the complexity of care transitions, organizations such as Accreditation Canada and the National Institute for Health and Care Excellence recommend enhanced MedRec to improve medication safety [12-15].

Previous research suggests that MedRec can be effective in identifying medication discrepancies, although evidence of improvements in clinical or health care use outcomes is mixed. A systematic review on the effectiveness of hospital-based, pharmacist-led MedRec found that MedRec was associated with significant reductions in all-cause hospital readmissions and all-cause emergency department visits despite heterogeneity across studies [16]. In contrast, a similar review found no significant reduction in health care use [17]. A Cochrane review evaluating the effect of MedRec on medication discrepancies, patient - related outcomes, and health care use concluded that the certainty of the evidence in favor of MedRec was very low, noting heterogeneity across studies in intervention implementation, outcome definitions, and intervention effects [18]. The complexity and need for multiple coordinated components in MedRec interventions may explain their inconsistent outcomes [19,20]. These workflows are often time- and labor-intensive [10,21], which can compromise fidelity.

There have been efforts to develop electronic MedRec tools over the last 2 decades. These tools have various functionalities that may reduce the time and labor demands of traditional, paper-based MedRec, such as interfaces that visually facilitate comparison of medication lists [22] and automated retrieval of community medication lists [23]. Other efforts have focused on methods for conducting MedRec tasks remotely, such as creating an interactive web form or leveraging video visits [24-26]. Due to the need for repetitive data entry, MedRec workflows are ideal targets for artificial intelligence (AI)–driven automation [27]. Although automation has not been the focus of previous literature reviews of MedRec tools, Marien et al [28] conducted a systematic review of electronic MedRec tools and found that only half of the identified tools provided automated discrepancy detection. Ciudad-Gutiérrez et al [29] reviewed a similar set of electronic MedRec tools, noting that approximately half could automatically retrieve medication information. While both functionalities involve important processes that could be automated, no review has yet considered the role of AI in automating MedRec. AI has recently been applied to specific subtasks within the broader workflows of MedRec (eg, extracting medications from free-text notes [30]), but previous literature reviews of MedRec tools have not explored how AI has been leveraged to facilitate MedRec tasks and subtasks. Existing reviews have also examined the use of AI for other aspects of medication use and safety, such as ADE prediction and detection [31], or more broadly for medication management [32,33], but none have focused specifically on MedRec and its constituent tasks and subtasks. As medication errors remain a leading cause of preventable patient harm, mapping the MedRec processes to which AI has been applied is an important step toward understanding its role in facilitating MedRec automation. This is the first review to examine the use of AI for the automation of MedRec processes broken down by specific tasks and subtasks.

The objective of this scoping review was to map the existing literature on the use of AI for MedRec tasks and subtasks, with a focus on assessing the types of functions automated by the models or tools.

Methods

We conducted this scoping review following guidance from the Joanna Briggs Institute [34]. This review is reported in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews; Checklist 1) [35].

Eligibility Criteria

Following our protocol [36], eligible studies described applications of AI for a task related to MedRec. A previous review of electronic tools that support MedRec described MedRec as consisting of three core tasks: (1) creation of a BPMH, (2) identification of discrepancies among medication lists, and (3) resolution of medication discrepancies [28]. On the basis of these criteria, we included studies that described the development or evaluation of an AI model or AI-based tool that facilitated any of these tasks or a subtask associated with any of these tasks. To maintain relevance, we ensured that the primary objectives of the included studies aligned with MedRec tasks or subtasks.

XSL•FO
RenderX

Previous scoping reviews have mapped the use of AI or machine learning (ML) for purposes related to medication adherence and ADEs [31,37]. Thus, to avoid duplicating existing work, we excluded studies focusing on adherence, side effects, adverse drug reactions, or allergies, narrowing our review to the core tasks of MedRec. Similarly, we excluded studies restricted to specific medication classes to ensure broader applicability. We also excluded models or tools that were solely rule-based, instead focusing on those that were ML-based or incorporated both ML and rule-based methods.

We included studies involving participants of any population, as well as those focusing on nonhuman elements such as medication lists. We excluded studies dealing specifically with animals or veterinary medicine. We considered all health care settings (eg, inpatient, outpatient, and primary care). We included journal articles and certain types of gray literature, namely, conference papers and preprints, but excluded abstract-only reports. We considered structured reviews for inclusion but excluded other secondary literature.

Information Sources, Search Strategy, and Study Selection

We undertook an initial preliminary search of MEDLINE (PubMed) to identify relevant articles. We used text words used in the titles and abstracts of these articles, as well as index terms used to describe them, to develop the full search strategy (Multimedia Appendix 1), which we adapted for each database. For the AI or ML component of our search, we adapted the AI- or ML-relevant terms from the search strategy used by Syrowatka et al [31] in their review of AI for ADE applications. The databases we searched were MEDLINE (Ovid), Embase (Ovid), Web of Science Core Collection, IEEE Xplore, and Compendex (Engineering Village). We searched the databases from inception to June 10, 2024, without any restrictions. In accordance with our eligibility criteria, we used a filter to exclude records indexed as involving animals and not humans. We also performed backward citation searching based on the articles included at the full-text stage using Citationchaser [38]. Due to resource constraints, we only considered English-language reports. Additional search details are presented in the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) [39] checklist (Checklist 2).

After completing the search, we uploaded all records retrieved into Rayyan (Qatar Computing Research Institute). One reviewer manually inspected all records flagged by Rayyan as potential duplicates for deduplication. Following deduplication, at least 2 independent reviewers performed title and abstract screening for each record. Where screeners disagreed about study inclusion, they discussed the records to reach consensus, consulting an additional reviewer if necessary. We retrieved the full texts of potentially relevant records and assessed them in the same manner.

Data Extraction

We abstracted data using an extraction tool that we had piloted with 20% of the included studies (Multimedia Appendix 2 [30,40-132]). At least 2 independent reviewers extracted data for each study. Where discrepancies in data extraction arose, the extractors resolved them through discussion to reach consensus, consulting a third reviewer when consensus could not be achieved. We extracted information such as primary objectives, data sources, participant characteristics, training and evaluation set sizes, MedRec tasks and subtasks facilitated by the model or tool, ML methods used, and performance metrics for the best-performing model.

Data Analysis

Previous research has explored the concept of automation from various perspectives. Our analytic approach was guided by the 4-stage model of human information processing by Parasuraman et al [133], which outlines system functions that can be automated, including information acquisition, information analysis, decision selection, and action implementation. We mapped the 3 core MedRec tasks of gathering medication data for the creation of a BPMH, identifying discrepancies, and resolving discrepancies (ie, evaluating discrepancies to determine actions) onto the first 3 stages of this model: information acquisition, information analysis, and decision selection, respectively (Figure 1 [28]). The fourth stage, action implementation, extends beyond the scope of core MedRec tasks as defined by Marien et al [28] but may include updating medication lists, communicating changes to the care team, or educating patients. This mapping approach provided a structured lens for assessing how AI-enabled automation could influence MedRec workflows.

XSL•FO
RenderX

Figure 1. Stages of human information processing and corresponding medication reconciliation (MedRec) tasks (in an inpatient admission context) outlined by Marien et al [28]. BPMH: best possible medication history.

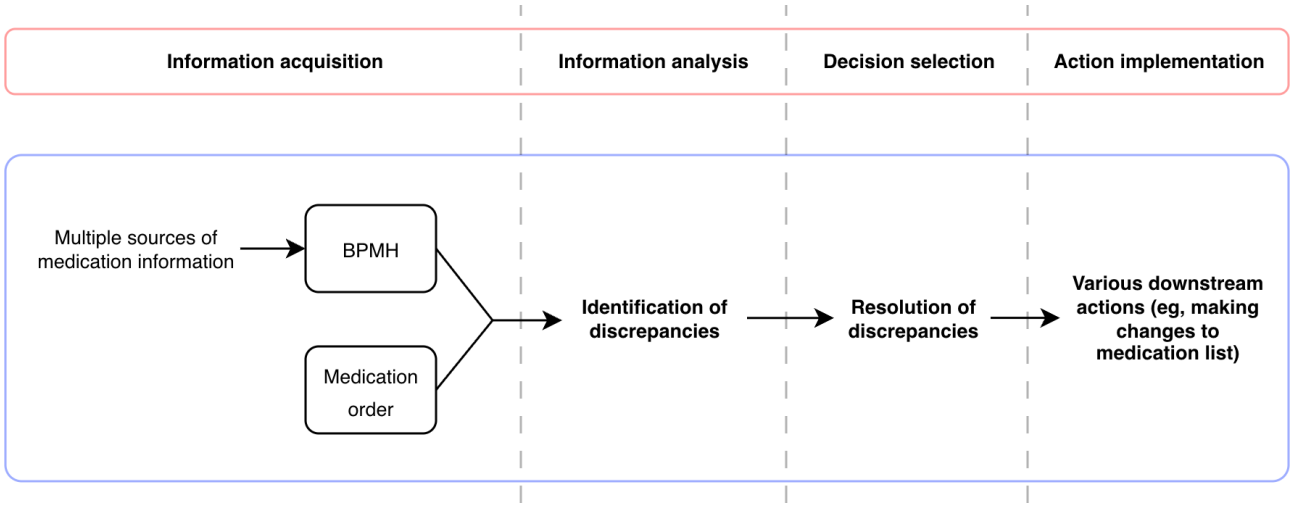

Results

Review Findings

We identified 2345 records through database searching and an additional 795 unique records from citation searching, ultimately including 94 studies (Figure 2).

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram showing the study selection process. ADE: adverse drug event; MedRec: medication reconciliation; ML: machine learning.

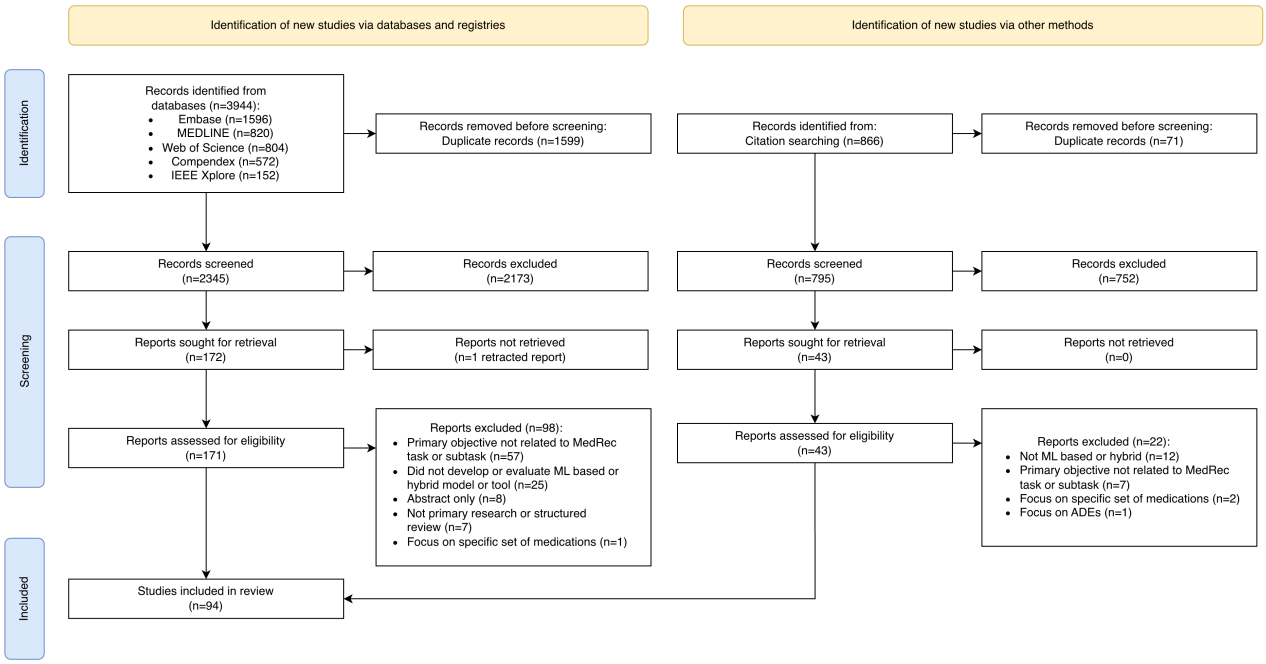

We found that the included studies were largely restricted to 1 MedRec subtask, primarily using clinical notes in the electronic health record (EHR), although they used a variety of ML methods.

MedRec Tasks and Subtasks

All the included studies (94/94, 100%) focused on subtasks related to the creation of a BPMH, whereas 2.1% (2/94) also addressed the identification of discrepancies (Table 1). In terms of BPMH creation subtasks, most studies (92/94, 97.9%) extracted medication information from textual or visual sources [30,40-130]. Only 2.1% (2/94) of the studies focused on verifying the accuracy of already compiled medication information, and both of these studies leveraged a collaborative filtering approach to identify potential omissions from pharmacy medication lists [131,132]. Of the 2 studies that addressed the identification of discrepancies, both identified medications that were consistent as well as inconsistent between medication lists. One study used a conditional random field (CRF) model to

XSL•FO
RenderX

extract medication entities from free-text clinical notes and compared them to medications extracted from structured discharge prescriptions using string matching, term co-occurrence, and drug synonyms, labeling medications from the clinical notes as “matched” or “discrepant” [80]. The other study developed a web-based application that retrieved medication lists from the EHR, allowing patients to manually confirm the extracted medications and add missing medications. Differences between the AI-extracted medications and patient-entered medications were flagged by the application [84]. No studies addressed the resolution of discrepancies between medication lists. As such, the highest stage of information processing automation achieved by the included models and tools was information analysis, with most studies (92/94, 97.9%) only automating information acquisition; none automated decision selection (Figure 3).

XSL•FO
RenderX

Table . Characteristics of the included studies (N=94).

Characteristic	Studies, n (%)
Year of publication	
2001 - 2010	14 (14.9)
2011 - 2020	45 (47.9)
2021 - 2024	35 (37.2)
Region	
Africa	1 (1.1)
Asia	19 (20.2)
Australia	4 (4.4)
Europe	15 (16)
North America	55 (60.4)
Medication reconciliation task[a]	
Creation of a BPMH[b]	94 (100)
Identification of discrepancies	2 (2.1)
Resolution of discrepancies	0 (0)
ML[c] method[a]	
Conditional random field	36 (38.3)
Recurrent neural network	34 (36.2)
Convolutional neural network	31 (33)
Transformer	25 (26.6)
Support vector machine	20 (21.3)
Maximum entropy	8 (8.5)
Random forest	6 (6.4)
K-nearest neighbor	6 (6.4)
Naive Bayes	4 (4.3)
K-means clustering	4 (4.3)
Other	21 (22.3)
Data source[a]	
Clinical note in EHR[d]	67 (71.3)
User-generated image	21 (22.3)
Pharmacy records	3 (3.2)
Patient-physician conversation transcript	2 (2.1)
Structured prescription list in the EHR	1 (1.1)
User-generated note	1 (1.1)
User-generated structured data	1 (1.1)
Medical internet forum	1 (1.1)
Data collection[a]	
Author-collected data	33 (35.1)
i2b2 2009 dataset	19 (20.2)
n2c2 2018 dataset	16 (17)
n2c2 2022 dataset	12 (12.8)
MADE 1.0 2018 dataset	7 (7.4)

XSL•FO
RenderX

Characteristic	Studies, n (%)
NIH[e] NLM[f] Pill Image Recognition dataset	4 (4.3)
Other	19 (20.2)
Stage of implementation	
Model development, retrospective	93 (98.9)
Usability study	1 (1.1)

[a]Totals greater than 100%.
[b]BPMH: best possible medication history.
[c]ML: machine learning.
[d]EHR: electronic health record.
[e]NIH: National Institutes of Health.
[f]NLM: National Library of Medicine.

Figure 3. Conceptual mapping of medication reconciliation (MedRec) tasks and subtasks identified in this review onto the stages of human information processing and frequency of included studies by year. Subtasks are listed as bullet points. BPMH: best possible medication history; NIH: National Institutes of Health; NLM: National Library of Medicine.

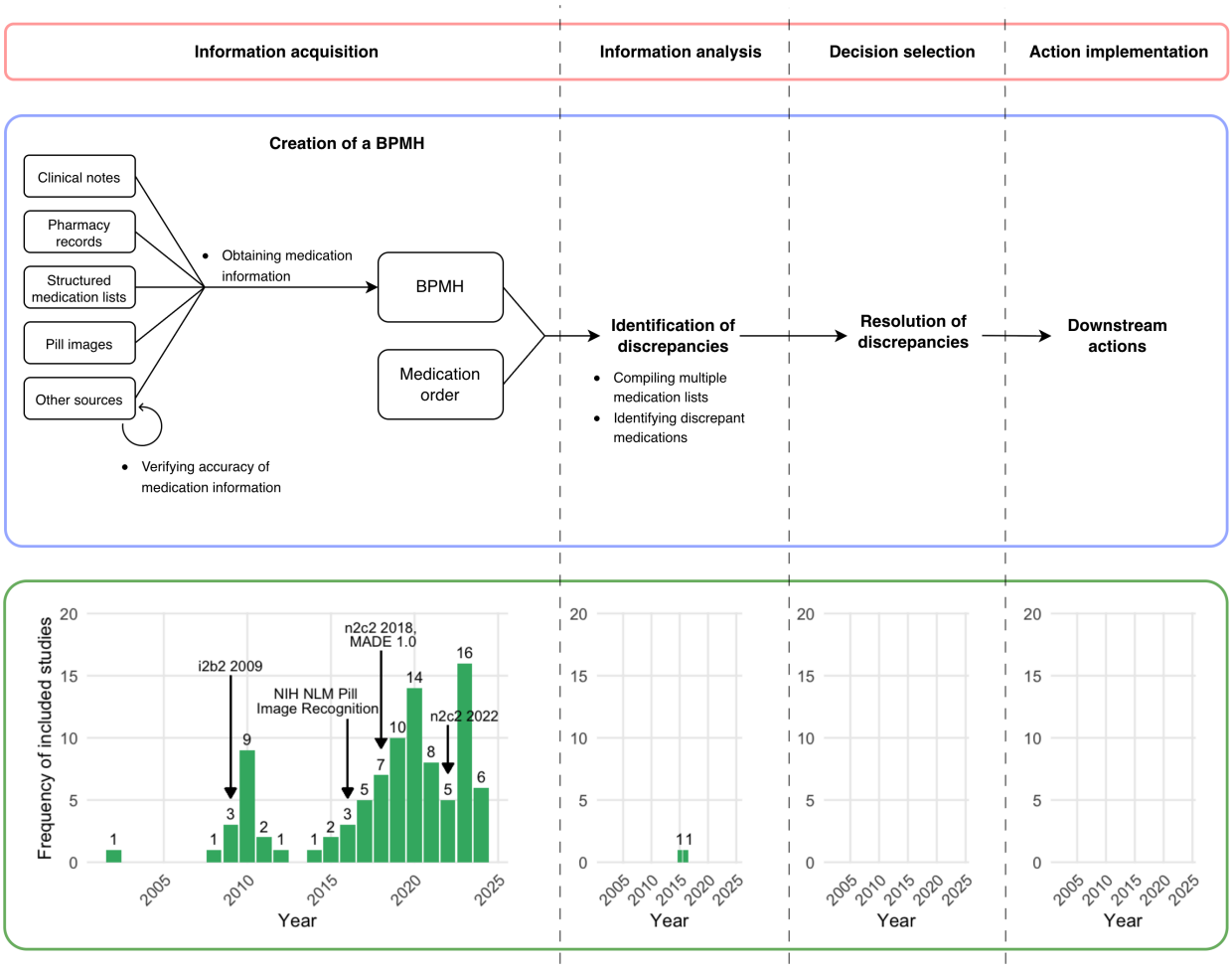

ML Methods

The included studies used a variety of ML methods, with many using multiple methods (Figure 4). The most commonly used methods included CRFs (36/94, 38.3%) [30,41,43-45,48,52,55,57,67,68,70,71,73,75,78-81,94,95,99,100,103,109,115-117,119,123,126,127], recurrent neural networks (34/94, 36.2%) [30,40,41,45,48-52,59,61,67,68,70-74,76,78,81,85,91,92,94,95,101,103,109,119,123,125-127], convolutional neural networks (31/94, 33%) [40,42,45,49,52,59,64,72,74,76,77,81,82,87,90,92,95-97,101,102,108,112,114,119,120,122-124,127,130], transformers (25/94, 26.6%) [30,46,47,57,60,74,75,85-87,92-95,101,103-107,113,118,119,121,128], and support vector machines (20/94, 21.3%) [53,55,56,59,65,66,73,79,83,86,91,99,100,111,115,123,125-127,130]. Studies using free-text data (eg, notes in the EHR) typically used CRFs, recurrent neural networks, support vector machines,

XSL•FO
RenderX

and transformers, whereas image-based studies mainly used convolutional neural networks. Owing to heterogeneity in task definitions (eg, differences in information extraction objectives and input modalities), performance metrics across studies were not directly comparable.

Figure 4. Evidence gap map illustrating the frequency of machine learning (ML) methods applied to medication reconciliation tasks by publication year. BPMH: best possible medication history; CNN: convolutional neural network; CRF: conditional random field; RNN: recurrent neural network; SVM: support vector machine.

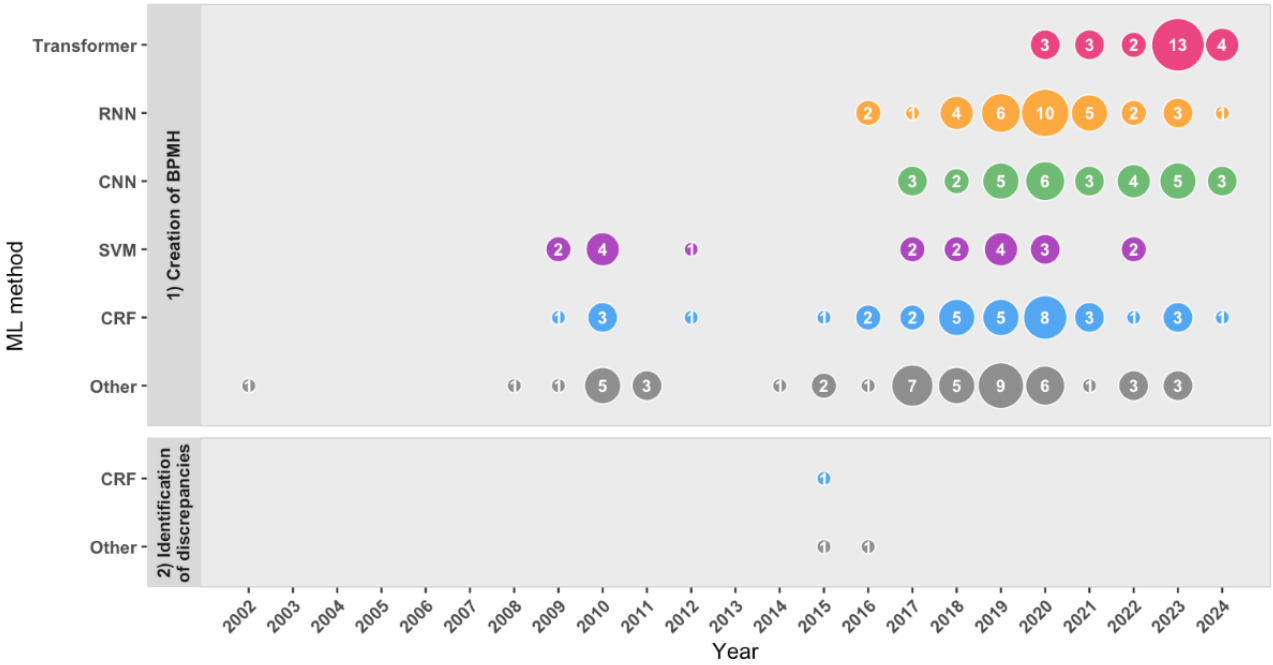

Among the studies that used free-text data sources (72/94, 76.6%), there was variation in the granularity of information extraction. Most of these studies (57/72, 79.2%) used ML-based named entity recognition techniques to extract a variety of medication-related elements, such as dosage, route of administration, frequency, and indication [30,41,43-53,55-57,59,60,63,65-71,73-75,78,81,85,91-93,99,100,103,104,106-109,113,115-119,121,123,125-128]. A smaller proportion of these studies (28/72, 38.9%) extended their approach to include relation extraction, identifying associations between these elements and specific medications (eg, linking a particular dosage to a specific medication) [30,41,43-45,47,50-52,63,73,78,79,81,87,91,93,99,100,106,107,115,116,123,125-128]. An even more limited subset (9/72, 12.5%) focused on identifying medication changes (eg, discontinuation or dosage adjustment) and the context of such changes (eg, the temporality of the change and the person responsible for the change) [46,47,57,86,103,104,106,118,121]. These studies typically assessed performance using standard evaluation metrics for natural language processing information extraction tasks, including precision, recall, and F_1-score.

Studies that used image data (21/94, 22.3%) were primarily focused on medication identification. These studies frequently analyzed images of pills [42,77,82,96,97,102,112,114,120,122,124,129,130] or blister packs [66] to identify medications or generate a ranked list of possible medications. Others applied optical character recognition to pill bottle images or handwritten prescriptions and used natural language processing methods to extract medication information [59,76,101]. Most pill identification studies (10/13, 76.9%) used single-pill images, although a few (3/13, 23.1%) addressed multi-pill image identification [77,96,97]. No studies using pill images explicitly reported using patient-generated images, although many (8/13, 61.5%) incorporated intentional variations in lighting, zooming, and background, among other aspects, to simulate photos taken in real-world conditions [77,82,97,112,120,122,124,130]. The image-based studies often used more general accuracy metrics.

Data Sources and Collection

Most included studies (67/94, 71.3%) used clinical notes in the EHR [30,41,43-46,55,56,63,65-68,70,71,73-75,78,81,83-89,91,93-95,98-100,103-106,109-111,113,115-119,121,123,125-128], and a significant proportion of studies (21/94, 22.3%) used images [40,42,59,64,66,72,76,77,82,90,96,97,101,102,112,114,120,122,124,129,130]. In most studies using images, the images featured pills (13/21, 61.9%), although 38.1% (8/21) of the studies used images of other medication-related items (eg, medication bottles and handwritten prescriptions) [40,59,64,66,72,76,90,101]. A few studies used other sources, such as pharmacy records (3/94, 3.2%) [92,131,132], patient-physician conversation transcripts (2/94, 2.1%) [58,107], structured prescription lists in the EHR (1/94, 1.1%) [80], user-generated structured data (1/94, 1.1%) [84], user-generated notes (1/94, 1.1%) [108], or a medical internet forum (1/94, 1.1%) [48].

Although many studies (33/94, 35.1%) used author-collected data [40,42,56,58,59,61,64-66,80,83,84,90,96-99,101,107,108,111-114,124,129,131,132], most studies used publicly available data from various benchmarking shared tasks, including i2b2 2009 (19/94, 20.2%) [48,53-55,61-63,74,79,88,89,93,95,100,110,115-117,119], n2c2

XSL•FO
RenderX

2018 (16/94, 17%) [30,41,45,49,52,71,73,85,87,91,93-95,123,127,128], n2c2 2022 (12/94, 12.8%) [46,47,57,60,61,75,103-106,118,121], MADE 1.0 2018 (7/94, 7.4%) [43,44,50,51,81,125,126], and the National Institutes of Health National Library of Medicine Pill Image Recognition (4/94, 4.3%) datasets [77,120,122,130]. The i2b2 2009 and n2c2 2018 datasets consisted solely of discharge summaries [100,134], whereas the n2c2 2022 and MADE 1.0 datasets were not restricted to a single note type [135,136]. The n2c2 2018 and 2022 datasets were annotated by at least 2 independent expert annotators per note. While MADE 1.0 also used 2 annotators, they were not independent. In contrast, the i2b2 2009 dataset was released without annotations, requiring researchers to perform their own labeling. The National Library of Medicine Pill Image Recognition dataset includes both standardized, high-quality pill images and more variable consumer-quality images to reflect real-world user-generated image conditions [137].

Research Maturity

Most of the included studies (93/94, 98.9%) were limited to model development, with only 1.1% (1/94) of the studies advancing to a more mature stage of implementation [84].

Discussion

Principal Findings

Overview

This scoping review aimed to characterize how researchers have used AI to facilitate MedRec tasks and subtasks. Notably, all the included studies (94/94, 100%) addressed subtasks related to the creation of a BPMH, but only a handful (2/94, 2.1%) developed models or tools that facilitated discrepancy identification, and none facilitated discrepancy resolution. While there is growing interest in applying AI to support MedRec tasks, most research remains focused on demonstrating technical feasibility rather than evaluating clinical integration or real-world impact. As such, efforts to automate MedRec tasks using AI are still preliminary, with current work largely focused on a limited set of subtasks, whereas other subtasks remain unexamined.

Creation of a BPMH

The creation of a BPMH is a critical component of MedRec. One study of inpatient MedRec found that most unintentional medication discrepancies that had the potential to cause harm were the result of medication history errors [138], highlighting the importance of this task. In addition, BPMH creation can be time-consuming [139], emphasizing the value of automating aspects of this process.

Although all studies included in this review (94/94, 100%) developed models or tools that focused on subtasks of BPMH creation—such as medication extraction or verification—only 1.1% (1/94) developed a method that could potentially be applied to automate BPMH creation in an end-to-end manner [80]. This gap suggests that the end-to-end automation of MedRec tasks is limited not only by the state of AI but also by infrastructural barriers such as limitations to the quick exchange of medication data between EHRs and pharmacies or data incompleteness [140]. Previous work examining barriers to effective MedRec has identified the inability to access reliable medication information as a recurring challenge as users often lack access to records from external settings and the information present in EHRs may be incomplete or outdated [141-143]. These challenges may explain why so few studies (1/94, 1.1%) had progressed beyond model development to more advanced stages of implementation. Given that many models target information acquisition tasks, future research should examine what information is available to clinicians at the point of care, how these tools can augment that information, and the extent to which their use reduces cognitive and operational burden. A well-integrated medication repository that consolidates comprehensive drug histories across care settings could mitigate infrastructural challenges by improving data completeness and interoperability [144]. While such systems are not yet widely adopted or fully integrated across care settings, their expansion and improvement represent an important opportunity for future MedRec work. Efforts to update and make efficient use of health care infrastructure are key to facilitating the research, development, and implementation of tools that can scale the automation of MedRec.

The BPMH subtasks that the included studies addressed were relatively broad in scope (ie, collecting medication information from various sources and verifying the accuracy of this information). While BPMH creation is central to MedRec, its subtasks overlap functionally with other medication-related processes. This functional overlap highlights the potential for cross-task model reuse and suggests that automation efforts need not be strictly siloed within this first MedRec stage. Many of these models have broader applications beyond MedRec, including pharmacovigilance, adherence monitoring, and medication review (ie, evaluation and optimization of a patient's medications [6]). More broadly, automating the creation of a complete medication list may enable downstream analytical use cases such as risk stratification (eg, age-related medication risks and fall risk from sedating medications), detection of potentially inappropriate prescribing or polypharmacy, and identification of deprescribing opportunities [145-147]. This broader relevance may partially explain the sustained research interest in developing these methods, particularly as public benchmarking initiatives such as the n2c2 2018 shared task on ADE and medication extraction [134] have incentivized the development of tools that support medication data processing across use cases.

The extent to which the included studies used data sources that reflect actual medication use is unclear, representing a key limitation in current BPMH automation research. The prevalence of patient nonadherence reported in the literature is often high [148], highlighting the importance of using sources of medication information that reflect actual use when developing a BPMH. The n2c2 2022 shared task on contextualized medication event extraction included identification of the "actor" responsible for medication changes (including patient-initiated changes) [135], suggesting that clinical notes may contain discernible indicators of actual use. Other data sources that may provide insights into actual medication use are patient-physician

XSL•FO
RenderX

and patient-nurse conversation transcripts. As with clinical notes, such tools would require contextualized approaches to distinguish between mentions of prescribed regimens and actual use. Similarly, although pill image–based models could provide a closer approximation of actual use than health care professional-documented sources, studies have yet to validate them on patient-submitted images. Future research should focus on incorporating data sources that reflect actual medication use (eg, patient self-reported data) to improve the accuracy and completeness of the BPMH.

Some tools developed may also have relevance in virtual or remote care contexts. For instance, image-based methods for medication identification could be integrated into virtual care video visits to facilitate the creation of a BPMH (eg, by acting as an additional method to verify the medications taken by patients). This functionality could serve as a supplemental verification method, which may be useful for patients with communication barriers or low health literacy or managing complex medication regimens. As health care systems continue to expand telehealth offerings, these tools may become increasingly relevant for extending MedRec capabilities outside of traditional clinical environments.

Identification of Discrepancies

Beyond gathering medication data to create a BPMH, later MedRec tasks often require tailored approaches to identify and resolve discrepancies. This review found one study that compared medications from clinical notes and discharge prescriptions and another that flagged discrepancies between AI-extracted and patient-entered medication lists. Only one of these studies incorporated patient-entered data as a source of medication information, reflecting actual medication use [84]. The other study that supported this task compared structured discharge prescription lists and medications extracted from free-text clinical notes [80], although it is unclear whether such notes reflect prescribed regimens, actual use, or both. The limited adoption of AI for this task may be partially related to the existence of non–AI-based electronic MedRec tools that facilitate discrepancy identification. For example, the Twinlist MedRec tool uses rule-based processing to detect discrepancies, using visual elements such as spatial grouping and highlighting to guide users in addressing these discrepancies [22]. Other tools have adopted similar strategies [149,150]. However, AI models hold promise for automating information analysis more comprehensively by enabling more robust discrepancy identification (eg, in cases in which medication information is not standardized in structure) or handling more complex medication regimens (eg, dose tapering). Given that existing tools for discrepancy identification may already be addressed in some settings, AI and health system researchers should engage with clinicians to identify what MedRec tasks or subtasks would benefit most from automation, ensuring that the tools meaningfully complement clinical workflows.

Resolution of Discrepancies

Discrepancy resolution represents a significant opportunity for AI innovation in MedRec as it requires interpreting and integrating contextual information that current tools do not address. Inconsistent resolution practices may contribute to the lack of clear safety benefits from MedRec, highlighting the need for standardized approaches to discrepancy resolution [151]. This process often involves determining whether a discrepancy is intentional or unintentional, a subtask that may require additional details such as the timing of a medication change. The recent n2c2 2022 shared task on contextualized medication event extraction suggests that these considerations are receiving increasing attention [135]. For example, in addition to helping discern actual medication use from clinical notes, automated contextualized extraction could facilitate the identification of health care professional-initiated medication changes as well as the nature of such changes (eg, stopping a medication), which may aid decision-making in light of discrepancies. Still, curating datasets to train models for discrepancy resolution may be particularly challenging, requiring granular annotations of clinical intent that are rarely captured in real-world records. Furthermore, other organizational barriers to MedRec, such as ambiguity in roles and lack of standardized workflows [152,153], may be less amenable to technological solutions, emphasizing the importance of evaluating tools in real-world contexts where such factors are relevant.

The automation of decision selection in MedRec requires improved data collection practices and focused model development. To mitigate this gap without overburdening clinicians, one potential strategy is to focus structured data entry efforts on high-risk or high-variability medications where the clinical stakes are greatest, such as high-alert medications [140]. For instance, given the risks of severe hemorrhage associated with improper use of antithrombic medications [154], this medication group could be a key target for AI-assisted MedRec. Focusing on high-alert medication groups such as these would allow for more targeted model training while preserving overall workflow efficiency. Moreover, to automate this MedRec task effectively, future models must be capable of analyzing multiple data types—including clinical notes, structured orders, medication timelines, and even health care professional communication patterns—to infer intent and recommend appropriate actions. This multimodal challenge, which extends beyond collecting the correct information, remains an underexplored but essential area for advancing toward fully integrated, AI-assisted MedRec workflows.

Limitations

There are limitations to this review. While the conceptual framework of MedRec tasks that we used provided a useful analytic structure, MedRec workflows vary across institutions. As such, the tasks and subtasks outlined in this review may not be universally applicable. Furthermore, although we framed the creation of a BPMH as the “information acquisition” step of MedRec, this process itself draws on medication information from multiple sources (eg, patient interview, pharmacy dispensing data, and EHR prescription lists). Hence, creating a BPMH also requires elements of information analysis and decision selection steps (eg, when sources used to create a BPMH are in conflict) such that the models and tools identified in this review are not necessarily restricted in their utility to just one MedRec task. In light of existing reviews, we excluded studies primarily focused on adherence, adverse drug reactions, side effects, or allergies; however, these aspects of medication

XSL•FO
RenderX

use are relevant to comprehensive MedRec efforts. We direct readers to the work by Bohlmann et al [37] and Syrowatka et al [31] for syntheses of the literature on AI or ML for adherence and ADE applications, which include studies relevant to the creation of a BPMH.

Finally, our search did not capture literature published since mid-2024, and therefore, our findings would not reflect more recent approaches. There may also be some commercial MedRec systems that use AI but that have not been evaluated in a research context and that, therefore, our search would not have identified.

Conclusions

AI holds potential for supporting MedRec by reducing manual workloads and improving efficiency. This work is the first to provide a comprehensive overview of the use of AI for MedRec, using an information processing framework to evaluate the degree of automation achieved across MedRec tasks. While previous reviews have explored the use of electronic tools for MedRec, none have focused on AI applications specifically. Our review highlights which MedRec tasks have received the most AI research effort and where gaps remain, providing valuable direction for future work that aims to advance the automation of MedRec. Current AI research in MedRec remains concentrated on a limited set of subtasks, primarily at the model development stage. To fully harness the potential of AI in MedRec, researchers and health care systems should approach implementation critically, emphasizing iterative evaluation and refinement. The aim is augmentation, not replacement. Rather than a simple transfer of responsibilities from human to machine, automation should be viewed as a dynamic process featuring ongoing user and AI tool interactions, as well as ongoing verification and refinement [155-157]. Researchers should balance refining AI-based models with the prevention of medication-related harms. Future research should prioritize developing models to automate discrepancy resolution, developing and evaluating AI-based MedRec tools, and addressing infrastructural barriers to broader implementation. Health systems need usable, well-integrated tools, as well as translational studies to assess their real-world effectiveness and safety. These tools will operate within complex sociotechnical systems, where success depends not only on model performance but also on usability, interoperability, and trust.

Funding

This study was funded through an award from the Population Health Research Institute at Hamilton Health Sciences. The funders had no role in the collection, analysis, and interpretation of the data; writing of this paper; or decision to submit it for publication.

Authors' Contributions

Conceptualization: JPTB, MPB, JR, PJD, MM, JP
Formal analysis: JPTB
Funding acquisition: PJD, MM
Investigation: JPTB, MPB, BW, PN
Methodology: JPTB, MPB, JR, JP
Project administration: JPTB, MPB
Supervision: PJD, MM, JP
Writing—original draft: JPTB
Writing—review and editing: MPB, JR, BW, PN, PJD, MM, JP

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full search strategies.
[DOCX File, 27 KB - jmir_v28i1e86760_app1.docx]

Multimedia Appendix 2
Extraction table.
[XLSX File, 317 KB - jmir_v28i1e86760_app2.xlsx]

Checklist 1
PRISMA-ScR checklist.
[PDF File, 743 KB - jmir_v28i1e86760_app3.pdf]

Checklist 2
PRISMA-S checklist.
[PDF File, 140 KB - jmir_v28i1e86760_app4.pdf]

XSL•FO
RenderX

Checklist 3
PRISMA 2020 for Abstracts checklist.
[PDF File, 194 KB - jmir_v28i1e86760_app5.pdf]

References

1. Hakkarainen KM, Hedna K, Petzold M, Hägg S. Percentage of patients with preventable adverse drug reactions and preventability of adverse drug reactions--a meta-analysis. PLoS One 2012;7(3):e33236. [doi: 10.1371/journal.pone.0033236] [Medline: 22438900]
2. Insani WN, Whittlesea C, Alwafi H, Man KK, Chapman S, Wei L. Prevalence of adverse drug reactions in the primary care setting: a systematic review and meta-analysis. PLoS One 2021;16(5):e0252161. [doi: 10.1371/journal.pone.0252161] [Medline: 34038474]
3. Laatikainen O, Miettunen J, Sneck S, Lehtiniemi H, Tenhunen O, Turpeinen M. The prevalence of medication-related adverse events in inpatients-a systematic review and meta-analysis. Eur J Clin Pharmacol 2017 Dec;73(12):1539-1549. [doi: 10.1007/s00228-017-2330-3] [Medline: 28871436]
4. Elliott RA, Camacho E, Jankovic D, Sculpher MJ, Faria R. Economic analysis of the prevalence and clinical and economic burden of medication error in England. BMJ Qual Saf 2021 Feb;30(2):96-105. [doi: 10.1136/bmjqs-2019-010206] [Medline: 32527980]
5. Medication without harm: WHO global patient safety challenge. World Health Organization. 2017. URL: https://iris.who.int/server/api/core/bitstreams/15520c4f-89d1-4a8f-9bab-6b771aa30acb/content [accessed 2025-03-08]
6. Penm J, Vaillancourt R, Pouliot A. Defining and identifying concepts of medication reconciliation: an international pharmacy perspective. Res Social Adm Pharm 2019 Jun;15(6):632-640. [doi: 10.1016/j.sapharm.2018.07.020] [Medline: 30100200]
7. Coleman EA, Smith JD, Raha D, Min SJ. Posthospital medication discrepancies: prevalence and contributing factors. Arch Intern Med 2005 Sep 12;165(16):1842-1847. [doi: 10.1001/archinte.165.16.1842] [Medline: 16157827]
8. Tully AP, Hammond DA, Li C, Jarrell AS, Kruer RM. Evaluation of medication errors at the transition of care from an ICU to non-ICU location. Crit Care Med 2019 Apr;47(4):543-549. [doi: 10.1097/CCM.0000000000003633] [Medline: 30855330]
9. Alqenae FA, Steinke D, Keers RN. Prevalence and nature of medication errors and medication-related harm following discharge from hospital to community settings: a systematic review. Drug Saf 2020 Jun;43(6):517-537. [doi: 10.1007/s40264-020-00918-3] [Medline: 32125666]
10. Quélennec B, Beretz L, Paya D, et al. Potential clinical impact of medication discrepancies at hospital admission. Eur J Intern Med 2013 Sep;24(6):530-535. [doi: 10.1016/j.ejim.2013.02.007] [Medline: 23514919]
11. McNab D, Bowie P, Ross A, MacWalter G, Ryan M, Morrison J. Systematic review and meta-analysis of the effectiveness of pharmacist-led medication reconciliation in the community after hospital discharge. BMJ Qual Saf 2018 Apr;27(4):308-320. [doi: 10.1136/bmjqs-2017-007087] [Medline: 29248878]
12. Required Organizational Practices 2021 Handbook: Accreditation Canada; 2021.
13. National Patient Safety Goals®: effective January 2025 for the hospital program. The Joint Commission. 2024. URL: https://digitalassets.jointcommission.org/api/public/content/9be383450fc941df806b76c5fbdd9ae6?v=3c600c3a [accessed 2026-04-29]
14. Medicines optimisation. National Institute for Health and Care Excellence. 2016. URL: https://www.nice.org.uk/guidance/qs120/resources/medicines-optimisation-pdf-75545351857861 [accessed 2026-04-29]
15. National safety and quality health service standards. Australian Commission on Safety and Quality in Health Care. 2012. URL: https://www.spkc.gov.lv/sites/spkc/files/pacientu_drosiba/standarti/australia_nsqhsstandardssept2012.pdf [accessed 2026-04-29]
16. Mekonnen AB, McLachlan AJ, Brien JA. Effectiveness of pharmacist-led medication reconciliation programmes on clinical outcomes at hospital transitions: a systematic review and meta-analysis. BMJ Open 2016 Feb 23;6(2):e010003. [doi: 10.1136/bmjopen-2015-010003] [Medline: 26908524]
17. Cheema E, Alhomoud FK, Kinsara AS, et al. The impact of pharmacists-led medicines reconciliation on healthcare outcomes in secondary care: a systematic review and meta-analysis of randomized controlled trials. PLoS One 2018;13(3):e0193510. [doi: 10.1371/journal.pone.0193510] [Medline: 29590146]
18. Redmond P, Grimes TC, McDonnell R, Boland F, Hughes C, Fahey T. Impact of medication reconciliation for improving transitions of care. Cochrane Database Syst Rev 2018 Aug 23;8(8):CD010791. [doi: 10.1002/14651858.CD010791.pub2] [Medline: 30136718]
19. Etchells E, Fernandes O. Medication reconciliation: ineffective or hard to implement? BMJ Qual Saf 2018 Dec;27(12):947-949. [doi: 10.1136/bmjqs-2018-008605] [Medline: 30355674]
20. Pevnick JM, Shane R, Schnipper JL. The problem with medication reconciliation. BMJ Qual Saf 2016 Sep;25(9):726-730. [doi: 10.1136/bmjqs-2015-004734] [Medline: 26795914]
21. Meguerditchian AN, Krotneva S, Reidel K, Huang A, Tamblyn R. Medication reconciliation at admission and discharge: a time and motion study. BMC Health Serv Res 2013 Nov 21;13:485. [doi: 10.1186/1472-6963-13-485] [Medline: 24261516]

XSL•FO
RenderX

22. Plaisant C, Chao T, Wu J, et al. Twinlist: novel user interface designs for medication reconciliation. AMIA Annu Symp Proc 2013;2013:1150-1159. [Medline: 24551399]
23. Tamblyn R, Winslade N, Lee TC, et al. Improving patient safety and efficiency of medication reconciliation through the development and adoption of a computer-assisted tool with automated electronic integration of population-based community drug data: the RightRx project. J Am Med Inform Assoc 2018 May 1;25(5):482-495. [doi: 10.1093/jamia/ocx107] [Medline: 29040609]
24. Gadallah A, McGinnis B, Nguyen B, Olson J. Assessing the impact of virtual medication history technicians on medication reconciliation discrepancies. Int J Clin Pharm 2021 Oct;43(5):1404-1411. [doi: 10.1007/s11096-021-01267-6] [Medline: 33871769]
25. Schubert CC, Penney LS, Schwartzkopf AL, et al. Expanding access to comprehensive geriatric evaluation via telehealth: development of hybrid-virtual home visits. J Gen Intern Med 2024 Feb;39(Suppl 1):36-43. [doi: 10.1007/s11606-023-08460-5] [Medline: 38227169]
26. Heyworth L, Paquin AM, Clark J, et al. Engaging patients in medication reconciliation via a patient portal following hospital discharge. J Am Med Inform Assoc 2014 Feb;21(e1):e157-e162. [doi: 10.1136/amiajnl-2013-001995] [Medline: 24036155]
27. Zayas-Cabán T, Okubo TH, Posnack S. Priorities to accelerate workflow automation in health care. J Am Med Inform Assoc 2022;30(1):195-201. [doi: 10.1093/jamia/ocac197]
28. Marien S, Krug B, Spinewine A. Electronic tools to support medication reconciliation: a systematic review. J Am Med Inform Assoc 2017 Jan;24(1):227-240. [doi: 10.1093/jamia/ocw068] [Medline: 27301747]
29. Ciudad-Gutiérrez P, Del Valle-Moreno P, Lora-Escobar SJ, Guisado-Gil AB, Alfaro-Lara ER. Electronic medication reconciliation tools aimed at healthcare professionals to support medication reconciliation: a systematic review. J Med Syst 2023 Dec 6;48(1):2. [doi: 10.1007/s10916-023-02008-0] [Medline: 38055124]
30. Christopoulou F, Tran TT, Sahu SK, Miwa M, Ananiadou S. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. J Am Med Inform Assoc 2020 Jan 1;27(1):39-46. [doi: 10.1093/jamia/ocz101] [Medline: 31390003]
31. Syrowatka A, Song W, Amato MG, et al. Key use cases for artificial intelligence to reduce the frequency of adverse drug events: a scoping review. Lancet Digit Health 2022 Feb;4(2):e137-e148. [doi: 10.1016/S2589-7500(21)00229-6] [Medline: 34836823]
32. Yao X, Rao A, Padman R. Analytical approaches for medication management for patient safety: a scoping review. NPJ Health Syst 2025;2:46. [doi: 10.1038/s44401-025-00052-1]
33. Ong JC, Chen MH, Ng N, et al. A scoping review on generative AI and large language models in mitigating medication related harm. NPJ Digit Med 2025 Mar 28;8:182. [doi: 10.1038/s41746-025-01565-7]
34. Peters MD, Godfrey C, McInerney P, Munn Z, Tricco AC, Khalil H. Scoping reviews. In: Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, editors. JBI Manual for Evidence Synthesis: JBI; 2024. [doi: 10.46658/JBIMES-24-09]
35. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
36. Tabja Bortesi J, Ranisau J. Automation in AI-based medication reconciliation tools: a scoping review protocol. OSF. Preprint posted online on Jun 10, 2024. [doi: 10.17605/OSF.IO/N64U3]
37. Bohlmann A, Mostafa J, Kumar M. Machine learning and medication adherence: scoping review. JMIRx Med 2021 Nov 24;2(4):e26993. [doi: 10.2196/26993] [Medline: 37725549]
38. Haddaway NR, Grainger MJ, Gray CT. Citationchaser: a tool for transparent and efficient forward and backward citation chasing in systematic searching. Res Synth Methods 2022 Jul;13(4):533-545. [doi: 10.1002/jrsm.1563] [Medline: 35472127]
39. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
40. Achkar R, Ghayad K, Haidar R, Saleh S, Al Hajj R. Medical handwritten prescription recognition using CRNN. Presented at: 2019 International Conference on Computer, Information and Telecommunication Systems (CITS); Aug 28-31, 2019. [doi: 10.1109/CITS.2019.8862004]
41. Alfattni G, Belousov M, Peek N, Nenadic G. Extracting drug names and associated attributes from discharge summaries: text mining study. JMIR Med Inform 2021 May 5;9(5):e24678. [doi: 10.2196/24678] [Medline: 33949962]
42. Ashraf AR, Somogyi-Végh A, Merczel S, Gyimesi N, Fittler A. Leveraging code-free deep learning for pill recognition in clinical settings: a multicenter, real-world study of performance across multiple platforms. Artif Intell Med 2024 Apr;150:102844. [doi: 10.1016/j.artmed.2024.102844] [Medline: 38553153]
43. Chapman AB, Peterson KS, Alba PR, DuVall SL, Patterson OV. Hybrid system for adverse drug event detection. In: Proceedings of the 1st International Workshop on Medication and Adverse Drug Event Detection: PMLR; 2018, Vol. 90:16-24 URL: https://proceedings.mlr.press/v90/chapman18a.html [accessed 2025-04-21]
44. Chapman AB, Peterson KS, Alba PR, DuVall SL, Patterson OV. Detecting adverse drug events with rapidly trained classification models. Drug Saf 2019 Jan;42(1):147-156. [doi: 10.1007/s40264-018-0763-y] [Medline: 30649737]
45. Chen L, Gu Y, Ji X, et al. Extracting medications and associated adverse drug events using a natural language processing system combining knowledge base and deep learning. J Am Med Inform Assoc 2020 Jan 1;27(1):56-64. [doi: 10.1093/jamia/ocz141] [Medline: 31591641]

XSL•FO
RenderX

46. Chen TY, Aquino JA, Chiu YW, Yeh WC, Chang YC. Understanding the clinical context of medication change events in clinical narratives using pre-trained clinical language models. Presented at: ICMHI '23: Proceedings of the 2023 7th International Conference on Medical and Health Informatics; May 12-14, 2023. [doi: 10.1145/3608298.3608318]
47. Chen A, Yu Z, Yang X, Guo Y, Bian J, Wu Y. Contextualized medication information extraction using transformer-based deep learning architectures. J Biomed Inform 2023 Jun;142:104370. [doi: 10.1016/j.jbi.2023.104370] [Medline: 37100106]
48. Dai X, Karimi S, Paris C. Medication and adverse event extraction from noisy text. Presented at: Proceedings of the Australasian Language Technology Association Workshop 2017; Dec 6-8, 2017 URL: https://aclanthology.org/U17-1009/ [accessed 2026-04-30]
49. Dai HJ, Su CH, Wu CS. Adverse drug event and medication extraction in electronic health records via a cascading architecture with different sequence labeling models and word embeddings. J Am Med Inform Assoc 2020 Jan 1;27(1):47-55. [doi: 10.1093/jamia/ocz120] [Medline: 31334805]
50. Dandala B, Joopudi V, Devarakonda M. IBM research system at MADE 2018: detecting adverse drug events from electronic health records. In: Proceedings of the 1st International Workshop on Medication and Adverse Drug Event Detection: PMLR; 2018, Vol. 90:39-47 URL: https://proceedings.mlr.press/v90/dandala18a.html [accessed 2025-04-21]
51. Dandala B, Joopudi V, Devarakonda M. Adverse drug events detection in clinical notes by jointly modeling entities and relations using neural networks. Drug Saf 2019 Jan;42(1):135-146. [doi: 10.1007/s40264-018-0764-x] [Medline: 30649738]
52. Dandala B, Joopudi V, Tsou CH, Liang JJ, Suryanarayanan P. Extraction of information related to drug safety surveillance from electronic health record notes: joint modeling of entities and relations using knowledge-aware neural attentive models. JMIR Med Inform 2020 Jul 10;8(7):e18417. [doi: 10.2196/18417] [Medline: 32459650]
53. Doan S, Xu H. Recognizing medication related entities in hospital discharge summaries using support vector machine. Proc Int Conf Comput Ling 2010 Aug;2010:259-266. [Medline: 26848286]
54. Doan S, Bastarache L, Klimkowski S, Denny JC, Xu H. Integrating existing natural language processing tools for medication extraction from discharge summaries. J Am Med Inform Assoc 2010;17(5):528-531. [doi: 10.1136/jamia.2010.003855] [Medline: 20819857]
55. Doan S, Collier N, Xu H, Pham HD, Tu MP. Recognition of medication information from discharge summaries using ensembles of classifiers. BMC Med Inform Decis Mak 2012 May 7;12:36. [doi: 10.1186/1472-6947-12-36] [Medline: 22564405]
56. Fan Y, Zhang R. Using natural language processing methods to classify use status of dietary supplements in clinical notes. BMC Med Inform Decis Mak 2018 Jul 23;18(Suppl 2):51. [doi: 10.1186/s12911-018-0626-6] [Medline: 30066648]
57. Gan Q, Hu M, Peterson KS, et al. A deep learning approach for medication disposition and corresponding attributes extraction. J Biomed Inform 2023 Jul;143:104391. [doi: 10.1016/j.jbi.2023.104391] [Medline: 37196988]
58. Ganoe CH, Wu W, Barr PJ, et al. Natural language processing for automated annotation of medication mentions in primary care visit conversations. JAMIA Open 2021;4(3):ooab071. [doi: 10.1093/jamiaopen/ooab071] [Medline: 34423262]
59. Gromova K, Elangovan V. Automatic extraction of medication information from cylindrically distorted pill bottle labels. Mach Learn Knowl Extr 2022;4(4):852-864. [doi: 10.3390/make4040043]
60. Guo Y, Ge Y, Sarker A. Detection of medication mentions and medication change events in clinical notes using transformer-based models. Stud Health Technol Inform 2024 Jan 25;310:685-689. [doi: 10.3233/SHTI231052] [Medline: 38269896]
61. Guzman B, Metzger I, Aphinyanaphongs Y, Grover H. Assessment of Amazon Comprehend Medical: medication information extraction. arXiv. Preprint posted online on Feb 2, 2020. [doi: 10.48550/arXiv.2002.00481]
62. Halgrim S, Xia F, Solti I, Cadag E, Uzuner Ö. Extracting medication information from discharge summaries. In: Dalianis H, Hassel M, Nilsson G, editors. Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents Association for Computational Linguistics: Association for Computational Linguistics; 2010:61-67.
63. Halgrim SR, Xia F, Solti I, Cadag E, Uzuner Ö. A cascade of classifiers for extracting medication information from discharge summaries. J Biomed Semant 2011;2(Suppl 3):S2. [doi: 10.1186/2041-1480-2-S3-S2]
64. Hassan E, Tarek H, Hazem M, Bahnacy S, Shaheen L, Elashmwai WH. Medical prescription recognition using machine learning. Presented at: 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC); Jan 27-20, 2021. [doi: 10.1109/CCWC51732.2021.9376141]
65. Iglesias JE, Rocks K, Jahanshad N, Frias-Martinez E, Andrada LP, Bui AA. Tracking medication information across medical records. AMIA Annu Symp Proc 2009 Nov 14;2009:266-270. [Medline: 20351862]
66. Itagi A, Sil R, Mohapatra S, et al. Medicine strip identification using 2-D cepstral feature extraction and multiclass classification methods. arXiv. Preprint posted online on Feb 3, 2020. [doi: 10.48550/arXiv.2003.00810]
67. Jagannatha AN, Yu H. Structured prediction models for RNN based sequence labeling in clinical text. Proc Conf Empir Methods Nat Lang Process 2016 Nov;2016:856-865. [doi: 10.18653/v1/d16-1082] [Medline: 28004040]
68. Jagannatha AN, Yu H. Bidirectional RNN for medical event detection in electronic health records. Proc Conf 2016 Jun;2016:473-482. [doi: 10.18653/v1/n16-1056] [Medline: 27885364]
69. Jagannathan V, Mullett CJ, Arbogast JG, et al. Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. Int J Med Inform 2009 Apr;78(4):284-291. [doi: 10.1016/j.ijmedinf.2008.08.006] [Medline: 18838293]

XSL•FO
RenderX

70. Jouffroy J, Feldman SF, Lerner I, Rance B, Burgun A, Neuraz A. Hybrid deep learning for medication-related information extraction from clinical texts in French: MedExt algorithm development study. JMIR Med Inform 2021 Mar 16;9(3):e17934. [doi: 10.2196/17934] [Medline: 33724196]
71. Ju M, Nguyen NT, Miwa M, Ananiadou S. An ensemble of neural models for nested adverse drug events and medication extraction with subwords. J Am Med Inform Assoc 2020 Jan 1;27(1):22-30. [doi: 10.1093/jamia/ocz075] [Medline: 31197355]
72. Kamalanaban E, Gopinath M, Premkumar S. Medicine box: doctor's prescription recognition using deep machine learning. Int J Eng Technol 2018;7(3):114. [doi: 10.14419/ijet.v7i3.34.18785]
73. Kim Y, Meystre SM. Ensemble method-based extraction of medication and related information from clinical texts. J Am Med Inform Assoc 2020 Jan 1;27(1):31-38. [doi: 10.1093/jamia/ocz100] [Medline: 31282932]
74. Kocabiyikoglu AC, Babouchkine JM, Portet F, Qader R. Neural medication extraction: a comparison of recent models in supervised and semi-supervised learning settings. Presented at: 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI); Aug 9-12, 2021. [doi: 10.1109/ICHI52183.2021.00032]
75. Koontz J, Oronoz M, Pérez A. Evaluating data augmentation for medication identification in clinical notes. In: Mitkov R, Angelova G, editors. Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing: INCOMA:578-585. [doi: 10.26615/978-954-452-092-2_063]
76. Kulathunga D, Muthukumarana C, Pasan U, Hemachandra C, Tissera M, De Silva H. PatientCare: patient assistive tool with automatic hand-written prescription reader. Presented at: 2nd International Conference on Advancements in Computing (ICAC); Dec 10-11, 2020. [doi: 10.1109/ICAC51239.2020.9357136]
77. Larios Delgado N, Usuyama N, Hall AK, et al. Fast and accurate medication identification. NPJ Digit Med 2019;2:10. [doi: 10.1038/s41746-019-0086-0] [Medline: 31304359]
78. Lerner I, Jouffroy J, Burgun A, Neuraz A. Learning the grammar of drug prescription: recurrent neural network grammars for medication information extraction in clinical texts. arXiv. Preprint posted online on Apr 22, 2022. [doi: 10.48550/arXiv.2004.11622]
79. Li Z, Liu F, Antieau L, Cao Y, Yu H. Lancet: a high precision medication event extraction system for clinical text. J Am Med Inform Assoc 2010;17(5):563-567. [doi: 10.1136/jamia.2010.004077] [Medline: 20819865]
80. Li Q, Spooner SA, Kaiser M, et al. An end-to-end hybrid algorithm for automated medication discrepancy detection. BMC Med Inform Decis Mak 2015 May 6;15:37. [doi: 10.1186/s12911-015-0160-8] [Medline: 25943550]
81. Li F, Liu W, Yu H. Extraction of information related to adverse drug events from electronic health record notes: design of an end-to-end model based on deep learning. JMIR Med Inform 2018 Nov 26;6(4):e12159. [doi: 10.2196/12159] [Medline: 30478023]
82. Ling S, Pastor A, Li J, et al. Few-shot pill recognition. Presented at: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 13-19, 2020. [doi: 10.1109/CVPR42600.2020.00981]
83. Liu F, Pradhan R, Druhl E, et al. Learning to detect and understand drug discontinuation events from clinical narratives. J Am Med Inform Assoc 2019 Oct 1;26(10):943-951. [doi: 10.1093/jamia/ocz048] [Medline: 31034028]
84. Long J, Yuan MJ, Poonawala R. An observational study to evaluate the usability and intent to adopt an artificial intelligence-powered medication reconciliation tool. Interact J Med Res 2016 May 16;5(2):e14. [doi: 10.2196/ijmr.5462] [Medline: 27185210]
85. Mahajan D, Liang JJ, Tsou CH. Extracting daily dosage from medication instructions in EHRs: an automated approach and lessons learned. arXiv. Preprint posted online on May 21, 2020. [doi: 10.48550/arXiv.2005.10899]
86. Mahajan D, Liang JJ, Tsou CH. Toward understanding clinical context of medication change events in clinical narratives. AMIA Annu Symp Proc 2022;2021:833-842. [Medline: 35308981]
87. Mahendran D, McInnes BT. Extracting adverse drug events from clinical notes. AMIA Jt Summits Transl Sci Proc 2021;2021:420-429. [Medline: 34457157]
88. Meystre SM, Thibault J, Shen S, Hurdle JF, South BR. Textractor: a hybrid system for medications and reason for their prescription extraction from clinical text documents. J Am Med Inform Assoc 2010;17(5):559-562. [doi: 10.1136/jamia.2010.004028] [Medline: 20819864]
89. Meystre SM, Thibault J, Shen S, Hurdle JF, South BR. Automatically detecting medications and the reason for their prescription in clinical narrative text documents. Stud Health Technol Inform 2010;160(Pt 2):944-948. [Medline: 20841823]
90. Miao J, Huang Y, Wang Z, Wu Z, Lv J. Image recognition of traditional Chinese medicine based on deep learning. Front Bioeng Biotechnol 2023;11:1199803. [doi: 10.3389/fbioe.2023.1199803] [Medline: 37545883]
91. Miller T, Geva A, Dligach D. Extracting adverse drug event information with minimal engineering. Proc Conf 2019 Jun;2019:22-27. [doi: 10.18653/v1/w19-1903] [Medline: 34027520]
92. Mishra RK, Roy S, Palla SK, Patel N, Patel M, Jos S. Hybrid approach combining deep learning and a rule based expert system for concept extraction from prescriptions. Annu Int Conf IEEE Eng Med Biol Soc 2023 Jul;2023:1-4. [doi: 10.1109/EMBC40787.2023.10339977] [Medline: 38082624]
93. Modersohn L, Hahn U. Influence of context in transformer-based medication relation extraction. Stud Health Technol Inform 2024 Jan 25;310:669-673. [doi: 10.3233/SHTI231049] [Medline: 38269893]

XSL•FO
RenderX

94. Narayanan S, Mannam K, Rajan SP, Rangan PV. Evaluation of transfer learning for adverse drug event (ADE) and medication entity extraction. In: Rumshisky A, Roberts K, Bethard S, Naumann T, editors. Proceedings of the 3rd Clinical Natural Language Processing Workshop: Association for Computational Linguistics; 2020:55-64. [doi: 10.18653/v1/2020.clinicalnlp-1.6]
95. Narayanan S, Mannam K, Achan P, Ramesh MV, Rangan PV, Rajan SP. A contextual multi-task neural approach to medication and adverse events identification from clinical text. J Biomed Inform 2022 Jan;125:103960. [doi: 10.1016/j.jbi.2021.103960] [Medline: 34875387]
96. Nguyen AD, Nguyen TD, Pham HH, Nguyen TH, Nguyen PL. Image-based contextual pill recognition with medical knowledge graph assistance. arXiv. Preprint posted online on Aug 4, 2022. [doi: 10.48550/arXiv.2208.02432]
97. Nguyen AD, Pham HH, Trung HT, Nguyen QV, Truong TN, Nguyen PL. High accurate and explainable multi-pill detection framework with graph neural network-assisted multimodal data fusion. PLoS One 2023;18(9):e0291865. [doi: 10.1371/journal.pone.0291865] [Medline: 37768910]
98. Pakhomov SV, Ruggieri A, Chute CG. Maximum entropy modeling for mining patient medication status from free text. Proc AMIA Symp 2002:587-591. [Medline: 12463891]
99. Patrick J, Li M. A cascade approach to extracting medication events. Presented at: Proceedings of the Australasian Language Technology Association Workshop 2009; Dec 3-4, 2009 URL: https://aclanthology.org/U09-1014/ [accessed 2026-04-30]
100. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. J Am Med Inform Assoc 2010;17(5):524-527. [doi: 10.1136/jamia.2010.003939] [Medline: 20819856]
101. Paul S, Madan G, Mishra A, Hegde N, Kumar P, Aggarwal G. Weakly supervised information extraction from inscrutable handwritten document images. arXiv. Preprint posted online on Jun 12, 2023. [doi: 10.48550/arXiv.2306.06823]
102. Pilania U, Kumar M, Singh S. Leveraging convolutional neural networks for improved drug identification. Presented at: 2023 6th International Conference on Contemporary Computing and Informatics (IC3I); Sep 14-16, 2023. [doi: 10.1109/IC3I59117.2023.10398164]
103. Ramachandran GK, Lybarger K, Liu Y, et al. Extracting medication changes in clinical narratives using pre-trained language models. J Biomed Inform 2023 Mar;139:104302. [doi: 10.1016/j.jbi.2023.104302] [Medline: 36754129]
104. Salek Faramarzi N, Patel M, Bandarupally SH, Banerjee R. Context-aware medication event extraction from unstructured text. In: Naumann T, Ben Abacha A, Bethard S, Roberts K, Rumshisky A, editors. Proceedings of the 5th Clinical Natural Language Processing Workshop: Association for Computational Linguistics; 2023:86-95. [doi: 10.18653/v1/2023.clinicalnlp-1.11]
105. Sarker S, Qian L, Dong X. Medical data augmentation via ChatGPT: a case study on medication identification and medication event classification. arXiv. Preprint posted online on Jun 10, 2023. [doi: 10.48550/arXiv.2306.07297]
106. Schäfer H, Idrissi-Yaghir A, Bewersdorff J, Frihat S, Friedrich CM, Zesch T. Medication event extraction in clinical notes: contribution of the WisPerMed team to the n2c2 2022 challenge. J Biomed Inform 2023 Jul;143:104400. [doi: 10.1016/j.jbi.2023.104400] [Medline: 37211196]
107. Selvaraj SP, Konam S. Medication regimen extraction from medical conversations. arXiv. Preprint posted online on Dec 10, 2019. [doi: 10.48550/arXiv.1912.04961]
108. Sezgin E, Hussain SA, Rust S, Huang Y. Extracting medical information from free-text and unstructured patient-generated health data using natural language processing methods: feasibility study with real-world data. JMIR Form Res 2023 Mar 7;7:e43014. [doi: 10.2196/43014] [Medline: 36881467]
109. Shah-Mohammadi F, Cui W, Finkelstein J. Comparison of ACM and CLAMP for entity extraction in clinical notes. Annu Int Conf IEEE Eng Med Biol Soc 2021;2021:1989-1992. [doi: 10.1109/EMBC46164.2021.9630611]
110. Shivade C, Cormack J, Milward D. Precise medication extraction using agile text mining. In: Velupillai S, Duneld M, Kvist M, Dalianis H, Skeppstedt M, Henriksson A, editors. Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi): Association for Computational Linguistics; 2014:75-79. [doi: 10.3115/v1/W14-1111]
111. Sohn S, Murphy SP, Masanz JJ, Kocher JP, Savova GK. Classification of medication status change in clinical narratives. AMIA Annu Symp Proc 2010 Nov 13;2010:762-766. [Medline: 21347081]
112. Srikamdee S, Suksawatchon U, Suksawatchon J, Werapan W. ClinicYA: an application for pill identification using deep learning and k-means clustering. Presented at: 2022 26th International Computer Science and Engineering Conference (ICSEC); Dec 21-23, 2022. [doi: 10.1109/ICSEC56337.2022.10049345]
113. Stevens M, Kennedy G, Churches T. Applying and improving a publicly available medication NER pipeline in a clinical cancer EMR. Stud Health Technol Inform 2024 Jan 25;310:679-684. [doi: 10.3233/SHTI231051] [Medline: 38269895]
114. Swastika W, Prilianti K, Stefanus A, et al. Preliminary study of multi convolution neural network-based model to identify pills image using classification rules. Presented at: 2019 International Seminar on Intelligent Technology and Its Applications (ISITIA); Aug 28-29, 2019. [doi: 10.1109/ISITIA.2019.8937272]
115. Tao C, Filannino M, Uzuner Ö. Prescription extraction using CRFs and word embeddings. J Biomed Inform 2017 Aug;72:60-66. [doi: 10.1016/j.jbi.2017.07.002] [Medline: 28684255]
116. Tao C, Filannino M, Uzuner Ö. FABLE: a semi-supervised prescription information extraction system. AMIA Annu Symp Proc 2018;2018:1534-1543. [Medline: 30815199]

XSL•FO
RenderX

117. Tikk D, Solt I. Improving textual medication extraction using combined conditional random fields and rule-based systems. J Am Med Inform Assoc 2010;17(5):540-544. [doi: 10.1136/jamia.2010.004119] [Medline: 20819860]
118. Tsujimura T, Yamada K, Ida R, Miwa M, Sasaki Y. Contextualized medication event extraction with striding NER and multi-turn QA. J Biomed Inform 2023 Aug;144:104416. [doi: 10.1016/j.jbi.2023.104416] [Medline: 37321443]
119. Tu H, Han L, Nenadic G. Extraction of medication and temporal relation from clinical text using neural language models. Presented at: 2023 IEEE International Conference on Big Data (BigData); Dec 15-18, 2023. [doi: 10.1109/BigData59044.2023.10386489]
120. Usuyama N, Delgado NL, Hall AK, Lundin J. ePillID dataset: a low-shot fine-grained benchmark for pill identification. Presented at: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); Jun 14-19, 2020. [doi: 10.1109/CVPRW50498.2020.00463]
121. Vasilakes J, Georgiadis P, Nguyen NT, Miwa M, Ananiadou S. Contextualized medication event extraction with levitated markers. J Biomed Inform 2023 May;141:104347. [doi: 10.1016/j.jbi.2023.104347] [Medline: 37030658]
122. Wang Y, Ribera J, Liu C, Yarlagadda S, Zhu F. Pill recognition using minimal labeled data. Presented at: 2017 IEEE Third International Conference on Multimedia Big Data (BigMM); Apr 19-21, 2017. [doi: 10.1109/BigMM.2017.61]
123. Wei Q, Ji Z, Li Z, et al. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. J Am Med Inform Assoc 2020 Jan 1;27(1):13-21. [doi: 10.1093/jamia/ocz063] [Medline: 31135882]
124. Wong YF, Ng HT, Leung KY, Chan KY, Chan SY, Loy CC. Development of fine-grained pill identification algorithm using deep convolutional network. J Biomed Inform 2017 Oct;74:130-136. [doi: 10.1016/j.jbi.2017.09.005] [Medline: 28923366]
125. Xu D, Yadav V, Bethard S. UArizona at the MADE1.0 NLP challenge. Proc Mach Learn Res 2018 May;90:57-65. [Medline: 30467557]
126. Yang X, Bian J, Gong Y, Hogan WR, Wu Y. MADEx: a system for detecting medications, adverse drug events, and their relations from clinical notes. Drug Saf 2019 Jan;42(1):123-133. [doi: 10.1007/s40264-018-0761-0] [Medline: 30600484]
127. Yang X, Bian J, Fang R, Bjarnadottir RI, Hogan WR, Wu Y. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. J Am Med Inform Assoc 2020 Jan 1;27(1):65-72. [doi: 10.1093/jamia/ocz144] [Medline: 31504605]
128. Yazdani A, Proios D, Rouhizadeh H, Teodoro D. Efficient joint learning for clinical named entity recognition and relation extraction using Fourier networks: a use case in adverse drug events. arXiv. Preprint posted online on Feb 8, 2023. [doi: 10.48550/arXiv.2302.04185]
129. Yu J, Chen Z, Kamata SI, Yang J. Accurate system for automatic pill recognition using imprint information. IET Image Process 2015 Dec;9(12):1039-1047. [doi: 10.1049/iet-ipr.2014.1007]
130. Zeng X, Cao K, Zhang M. MobileDeepPill: a small-footprint mobile deep learning system for recognizing unconstrained pill images. Presented at: MobiSys '17: Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services; Jun 19-23, 2017. [doi: 10.1145/3081333.3081336]
131. Hasan S, Duncan GT, Neill DB, Padman R. Automatic detection of omissions in medication lists. J Am Med Inform Assoc 2011;18(4):449-458. [doi: 10.1136/amiajnl-2011-000106] [Medline: 21447497]
132. Hasan S, Duncan GT, Neill DB, Padman R. Towards a collaborative filtering approach to medication reconciliation. AMIA Annu Symp Proc 2008 Nov 6;2008:288-292. [Medline: 18998834]
133. Parasuraman R, Sheridan TB, Wickens CD. A model for types and levels of human interaction with automation. IEEE Trans Syst Man Cybern A Syst Hum 2000 May;30(3):286-297. [doi: 10.1109/3468.844354]
134. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. J Am Med Inform Assoc 2020 Jan 1;27(1):3-12. [doi: 10.1093/jamia/ocz166] [Medline: 31584655]
135. Mahajan D, Liang JJ, Tsou CH, Uzuner Ö. Overview of the 2022 n2c2 shared task on contextualized medication event extraction in clinical notes. J Biomed Inform 2023 Aug;144:104432. [doi: 10.1016/j.jbi.2023.104432] [Medline: 37356640]
136. Jagannatha A, Liu F, Liu W, Yu H. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). Drug Saf 2019 Jan;42(1):99-111. [doi: 10.1007/s40264-018-0762-z] [Medline: 30649735]
137. Yaniv Z, Faruque J, Howe S, et al. The National Library of Medicine pill image recognition challenge: an initial report. IEEE Appl Imag Pattern Recognit Workshop 2016 Oct;2016. [doi: 10.1109/AIPR.2016.8010584] [Medline: 29854569]
138. Pippins JR, Gandhi TK, Hamann C, et al. Classifying and predicting errors of inpatient medication reconciliation. J Gen Intern Med 2008 Sep;23(9):1414-1422. [doi: 10.1007/s11606-008-0687-9] [Medline: 18563493]
139. Alix L, Bajeux E, Hubert J, et al. Medication reconciliation in hospital patients over the age of 65: how long does it take and how much does it cost? A time-motion study in an internal medicine ward. Eur J Intern Med 2020 Mar;73:100-102. [doi: 10.1016/j.ejim.2019.12.003] [Medline: 31859024]
140. Kashyap N, Jeffery S, Agresta T. From MedWreck to MedRec: a call to action to improve medication reconciliation. Appl Clin Inform 2024 Mar;15(2):230-233. [doi: 10.1055/a-2181-1847] [Medline: 37748724]

XSL•FO
RenderX

141. Boockvar KS, Santos SL, Kushniruk A, Johnson C, Nebeker JR. Medication reconciliation: barriers and facilitators from the perspectives of resident physicians and pharmacists. J Hosp Med 2011;6(6):329-337. [doi: 10.1002/jhm.891] [Medline: 21834114]
142. Kennelty KA, Chewning B, Wise M, Kind A, Roberts T, Kreling D. Barriers and facilitators of medication reconciliation processes for recently discharged patients from community pharmacists' perspectives. Res Social Adm Pharm 2015;11(4):517-530. [doi: 10.1016/j.sapharm.2014.10.008] [Medline: 25586885]
143. Snyder ME, Nguyen KA, Patel H, et al. Clinicians' use of health information exchange technologies for medication reconciliation in the U.S. Department of Veterans Affairs: a qualitative analysis. BMC Health Serv Res 2024 Oct 8;24(1):1194. [doi: 10.1186/s12913-024-11690-w] [Medline: 39375765]
144. Salgado-Baez E, Heidepriem R, Delucchi Danhier R, et al. Toward interoperable digital medication records on fast healthcare interoperability resources: development and technical validation of a minimal core dataset. JMIR Med Inform 2025 May 9;13:e64099. [doi: 10.2196/64099] [Medline: 40014673]
145. Scott IA, Hilmer SN, Reeve E, et al. Reducing inappropriate polypharmacy: the process of deprescribing. JAMA Intern Med 2015 May;175(5):827-834. [doi: 10.1001/jamainternmed.2015.0324] [Medline: 25798731]
146. Socrates V, Wright DS, Huang T, et al. Identifying deprescribing opportunities with large language models in older adults: retrospective cohort study. JMIR Aging 2025 Apr 11;8:e69504. [doi: 10.2196/69504] [Medline: 40215480]
147. Wabe N, Meulenbroeks I, Huang G, et al. Development and internal validation of a dynamic fall risk prediction and monitoring tool in aged care using routinely collected electronic health data: a landmarking approach. J Am Med Inform Assoc 2024 Apr 19;31(5):1113-1125. [doi: 10.1093/jamia/ocae058] [Medline: 38531675]
148. Chapman SC, Chan AH. Medication nonadherence - definition, measurement, prevalence, and causes: reflecting on the past 20 years and looking forwards. Front Pharmacol 2025;16:1465059. [doi: 10.3389/fphar.2025.1465059] [Medline: 40124783]
149. Marien S, Legrand D, Ramdoyal R, et al. A user-centered design and usability testing of a web-based medication reconciliation application integrated in an eHealth network. Int J Med Inform 2019 Jun;126:138-146. [doi: 10.1016/j.ijmedinf.2019.03.013] [Medline: 31029255]
150. Gionfriddo MR, Hu Y, Maddineni B, et al. Evaluation of a web-based medication reconciliation application within a primary care setting: cluster-randomized controlled trial. JMIR Form Res 2022 Mar 8;6(3):e33488. [doi: 10.2196/33488] [Medline: 35023836]
151. MacLaren R. Resolving medication reconciliation. Can J Hosp Pharm 2022;75(3):155-156. [doi: 10.4212/cjhp.3324] [Medline: 35847465]
152. van Sluisveld N, Zegers M, Natsch S, Wollersheim H. Medication reconciliation at hospital admission and discharge: insufficient knowledge, unclear task reallocation and lack of collaboration as major barriers to medication safety. BMC Health Serv Res 2012 Jun 21;12:170. [doi: 10.1186/1472-6963-12-170] [Medline: 22721361]
153. Gionfriddo MR, Duboski V, Middernacht A, Kern MS, Graham J, Wright EA. A mixed methods evaluation of medication reconciliation in the primary care setting. PLoS One 2021;16(12):e0260882. [doi: 10.1371/journal.pone.0260882] [Medline: 34855888]
154. High-alert medications list. ISMP Canada. URL: https://ismpcanada.ca/resource/highalertlist/ [accessed 2026-03-26]
155. Hopkins D, Schwanen T. Talking about automated vehicles: what do levels of automation do? Technol Soc 2021 Feb;64:101488. [doi: 10.1016/j.techsoc.2020.101488]
156. Mosier KL, Skitka LJ. Human decision makers and automated decision aids: made for each other. In: Parasuraman R, Mouloua M, editors. Automation and Human Performance: CRC Press; 1996.
157. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc 2017 Mar 1;24(2):423-431. [doi: 10.1093/jamia/ocw105] [Medline: 27516495]

Abbreviations

ADE: adverse drug event
AI: artificial intelligence
BPMH: best possible medication history
CRF: conditional random field
EHR: electronic health record
MedRec: medication reconciliation
ML: machine learning
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

XSL•FO
RenderX

Edited by S Brini; submitted 29.Oct.2025; peer-reviewed by J Odeniya, K Fuji, R Marshall; revised version received 17.Apr.2026; accepted 20.Apr.2026; published 11.May.2026.

Please cite as:
Tabja Bortesi JP, Becerra MP, Ranisau J, Wen B, Nadesan P, Devereaux PJ, McGillion M, Petch J
AI-Based Automation for Medication Reconciliation: Scoping Review
J Med Internet Res 2026;28:e86760
URL: https://www.jmir.org/2026/1/e86760
doi:10.2196/86760

© Juan Pablo Tabja Bortesi, Maria P Becerra, Jonathan Ranisau, Bonnie Wen, Praveen Nadesan, P J Devereaux, Michael McGillion, Jeremy Petch. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Machine Learning and Deep Learning Models for Predicting Future Falls in Community-Dwelling Older Adults: Systematic Review and Meta-Analysis of Longitudinal Evidence

Ying Gao[1,2*], MSc; Doudou Xu[1,2,3*], MM; Xinru Li[2], BSc; Jue Wang[1,2,4], MM; Linbin Wang[5], PhD; Beiwen Wu[1], PhD; Haifeng Zhao[6], PhD; Xian Qiu[1,2,4*], MM; Weiyi Zhu[1,2*], MM

[1]Department of Nursing, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, No.197, 2nd Ruijin Rd, Huangpu District, Shanghai, China

[2]School of Nursing, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[3]Department of Orthopedics, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[4]Department of Neurosurgery, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[5]Department of Psychiatry, University of Cambridge, Cambridge, United Kingdom

[6]Institute of Medical Artificial Intelligence, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Weiyi Zhu, MM
Department of Nursing, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, No.197, 2nd Ruijin Rd, Huangpu District, Shanghai, China

Abstract

Background: Machine learning (ML) and deep learning (DL) show promise for fall risk prediction, but prior reviews focused mainly on real-time fall detection, in-hospital falls, or conventional statistical models. The performance of ML-DL–based models for predicting future falls in community-dwelling older adults remains unclear.

Objective: This study aimed to review ML-DL studies for predicting future falls among community-dwelling older adults and meta-analyze discrimination where feasible.

Methods: Six databases were searched from inception to September 23, 2024, with updates on August 31, 2025, and February 28, 2026. We included longitudinal studies developing or validating ML-DL models to predict future falls in community-dwelling adults aged ≥60 years and excluded real-time detection, simulated or no fall, and inpatient studies. Risk of bias was assessed using the Prediction Model Risk of Bias Assessment Tool (PROBAST). Areas under the curve (AUCs) were meta-analyzed using Hartung-Knapp-Sidik-Jonkman random-effects models with 95% CIs. Heterogeneity, 95% prediction intervals (PIs), sensitivity analyses, and subgroup analyses were conducted.

Results: After screening 10,253 records, 28 (0.3%) studies were included; 18 (64.3%) focused on general older adults. Prediction horizons ranged from 3 months to 7 years, and fall incidence ranged from 1.6% to 46.6%. Twenty-three (82.1%) studies applied ML, and 5 (17.9%) studies used DL. Input modalities included text (n=18, 64.3%), sensor (n=5, 17.9%), image (n=1, 3.6%), and multimodal data (n=4, 14.3%). Common predictors included age, sex, fall history, depression, and basic daily activities. Only one model underwent external validation. Calibration reporting was sparse. All models were rated at high risk of bias. Ten models were meta-analyzed, yielding a pooled AUC of 0.79 (95% CI 0.69 - 0.87) with extreme heterogeneity (τ^2=0.64; τ=0.80; I^2=99.8%; Q=4128.99). The confidence-distribution bootstrap PI was 0.20 to 0.99, indicating substantial uncertainty in expected performance across new populations. Subgroup analyses indicated moderation by sample size and population type, with higher discrimination in specific populations than in general samples; however, the specific population subgroup included only 2 studies. Although all participants were community dwelling, some cohorts were recruited through clinically enriched pathways rather than general community sampling.

Conclusions: ML-DL models show potential for identifying community-dwelling older adults at elevated future fall risk; however, wide PIs, limited external validation, and high risk of bias suggest real-world performance may be optimistic. The pooled AUC should be interpreted as a summary of reported discrimination under study-specific conditions, predominantly from internally validated, high-risk-of-bias models, rather than as a robust estimate of transportable real-world performance. This review extends prior reviews by focusing on community-dwelling settings and by integrating PROBAST, Hartung-Knapp-Sidik-Jonkman meta-analysis, PIs, and modality-specific synthesis to evaluate both discrimination and uncertainty.

XSL•FO
RenderX

Findings support the use of ML-DL models for proactive fall prevention while emphasizing the need for validation and context-specific implementation.

Trial Registration: PROSPERO CRD42024580902; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024580902

(J Med Internet Res 2026;28:e84844) doi:10.2196/84844

KEYWORDS

machine learning; falls; risk prediction; community-dwelling older adults; systematic review; meta-analysis

Introduction

Rationale

Falls, defined by the World Health Organization as events in which an individual unintentionally comes to rest on the ground, floor, or another lower level [1,2], are among the most common and serious health concerns in community-dwelling older adults. Approximately 26% of older adults worldwide experience at least one fall each year [2,3]. In the United States, more than 25% of community-dwelling older adults fall annually, and 37% of these incidents result in injuries requiring medical care [4,5]. Common adverse outcomes include fractures, head injuries, functional decline, mobility limitations, and premature institutionalization [6-8], placing a heavy burden on caregivers and community health care systems. Globally, falls account for approximately 20% to 30% of moderate to severe injuries among older adults and represent the second leading cause of injury-related death [9]. Beyond physical consequences, falls are also associated with psychological distress, including depression and fear of falling, which can substantially reduce quality of life and independence [10,11]. With the rapid aging of populations worldwide, both the incidence and consequences of falls among older adults are expected to increase further [2].

Unlike institutionalized older adults, who receive structured supervision and continuous monitoring from health care providers, community-dwelling older adults often have limited access to intensive monitoring or care services [12]. Consequently, individuals at high risk of falling may remain unidentified until an adverse event occurs. Several clinical assessment tools have been developed to identify individuals at risk of falls, including the Timed Up and Go test [13], Berg Balance Scale [14], and various multifactorial fall risk assessment instruments. However, these tools typically rely on a limited set of predefined functional or clinical indicators, which are most commonly measures of mobility, balance, or physical performance, and may not fully capture the complex and multifactorial determinants of falls in community-dwelling older adults [8]. In addition, although wearable sensor technologies have been increasingly developed and can facilitate rapid responses after a fall occurs, most focus on real-time fall detection rather than prospective fall risk prediction [15]. Thus, effective early identification of high-risk individuals in community and home settings, before a fall occurs, remains a critical challenge for implementing timely and targeted prevention strategies.

Traditional statistical approaches, such as logistic regression models, have commonly been used to estimate fall risk based on predefined predictors [16]. While these models offer interpretability and clinical familiarity, they are often limited in their ability to model nonlinear relationships, high-dimensional predictors, and complex associations among risk factors [17-19]. This is particularly problematic in community populations, where fall risk is influenced by a wide range of interrelated factors, including physical function, psychological conditions, lifestyle behaviors, environmental hazards, and social support [20,21].

Recent studies have identified numerous predictors for falls, including age [22,23], history of falls [24,25], comorbidities [26,27], use of high-risk medications [22], functional performance (eg, activities of daily living [ADLs] or instrumental activities of daily living [24,28]), gait and balance deficits [29], cognitive impairment, and psychological factors such as fear of falling [30]. These findings highlight the multifactorial and context-dependent nature of falls in community settings, where diverse health, functional, and social factors jointly influence fall risk. As a result, conventional regression-based models may oversimplify these relationships and exhibit limited predictive performance when applied to heterogeneous real-world populations.

In this context, machine learning (ML) methods, including deep learning (DL), offer a promising alternative for improving fall risk prediction and supporting individualized prevention by integrating heterogeneous predictors, accommodating nonlinear relationships, and leveraging multimodal data sources [31]. ML approaches such as decision trees and random forests (RFs) learn predictive rules directly from data without requiring prespecified functional forms, and they often provide relatively transparent model structures or feature-importance measures that can facilitate clinical interpretation [32]. DL, as a subfield of ML, uses multilayer neural networks to learn hierarchical representations from raw or minimally processed inputs, which can be advantageous for high-dimensional modalities such as wearable sensor time series and imaging data [33,34]. Both ML and DL are important: ML remains a pragmatic and interpretable baseline for many structured clinical and survey datasets, whereas DL may be better suited to complex signals but typically requires larger sample sizes and more explicit strategies to ensure interpretability and generalizability in medical settings.

Although the importance of fall risk management and the potential of ML have been widely recognized, existing reviews focused on institutionalized real-time fall detection, in-hospital fall risk prediction, or conventional statistical models, rather than ML-based models designed to predict future falls for primary prevention and early identification among the community-dwelling older adults. For example, a 2025 review by Xie et al [35] focused on ML-based models for short-term

XSL•FO
RenderX

in-hospital fall outcomes, limiting their applicability to community settings where risk profiles differ substantially. A 2024 review by Dormosh et al [36] examining community-dwelling older adults primarily evaluated traditional statistical prediction models, with limited emphasis on ML-DL approaches. In addition, reviews by Usmani et al [18] and Ren et al [37] focused on real-time or near-real-time fall detection using sensor-based data in controlled or laboratory settings, rather than prospective risk prediction in real-world environments, thereby limiting their relevance for primary prevention and early risk identification in community-dwelling populations. To date, no systematic review has comprehensively evaluated the predictive performance, methodological quality, and clinical applicability of ML-based fall risk prediction models specifically designed to predict future falls among community-dwelling older adults for primary prevention and early risk identification.

Objectives

To address these gaps, this systematic review and meta-analysis aims to (1) systematically summarize ML- and DL-based prediction models for future falls in community-dwelling older adults and evaluate their predictive performance; (2) assess the methodological quality and risk of bias of included models; and (3) synthesize discrimination performance across studies using meta-analysis of area under the curve (AUC) values, and explore potential sources of heterogeneity through subgroup analyses and meta-regression, to inform the development and implementation of robust, evidence-based tools for fall prevention in real-world community settings.

Methods

Protocol and Registration

This systematic review and meta-analysis was registered in PROSPERO under the registration number CRD42024580902. Deviations from the registered protocol should be noted. First, the literature search was updated to include studies published up to February 28, 2026. Second, the final meta-analysis was restricted to studies reporting the AUC with corresponding 95% CIs, focusing on model discrimination performance. Third, because all included models were judged to be at high risk of bias, the prespecified sensitivity analysis excluding high-risk studies was not feasible; instead, leave-one-out analyses and additional sensitivity analyses addressing potential overlap among the China Health and Retirement Longitudinal Study (CHARLS)–derived cohorts were performed. Additional statistical methods were refined during the analysis to improve methodological rigor.

The review was reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement [38], and the complete PRISMA 2020 expanded checklist [38] is provided in Table S1 in Checklist 1. To strengthen transparency, the search and its reporting followed the PRISMA-S (PRISMA literature search extension) [39] (Table S2 in Checklist 1). As the review synthesized model discrimination (AUC) across studies, selected elements of the PRISMA-DTA (PRISMA Statement for Diagnostic Test Accuracy) guideline were additionally consulted to inform the reporting structure where applicable [40], and the corresponding checklist is provided in Table S3 in Checklist 1.

Eligibility Criteria

Studies were included if they met the following criteria: (1) community-dwelling older adults aged 60 years or older at baseline; (2) longitudinal studies, cohort studies, or case-control studies; (3) predictors measured at baseline; (4) ML models, including DL, for identifying participants at high risk of future falls; (5) assessment of fall-related outcomes, such as fall occurrence, fall rate, number of fallers, or fall risk; and (6) publication in English with full-text available.

Studies were excluded based on the following criteria: (1) detection of real-time falls; (2) no actual falls, such as simulated falls, fear of falling, abnormal gait, and mobility decline; (3) inpatient falls within hospital or institutional settings; and (4) reviews, meta-analyses, editorials, letters, study protocols, case reports, abstracts, preprints, or other forms of secondary research.

Information Sources

We conducted a comprehensive literature search across 6 databases, including PubMed, Embase, Web of Science Core Collection, CINAHL, Cochrane Library, and IEEE Xplore, from their inception to September 23, 2024. The search was subsequently updated on August 31, 2025, and February 28, 2026, using the same search strategies. Additionally, reference lists of included studies were examined to identify potentially eligible articles. Clinical trial registries were not searched because this review focused on published prediction model studies. No additional online resources, conference proceedings, or website browsing were systematically searched. Study authors or experts were not contacted to identify additional studies or obtain unpublished data. No additional search methods beyond electronic database searching and reference list screening were used.

Search

The search strategy combined controlled vocabulary terms (eg, MeSH and Emtree) such as “aged,” “artificial intelligence,” “machine learning,” and “accidental fall,” with relevant free-text terms. No limits or restrictions were applied during the literature search to maximize inclusivity. No published search filters were used. The detailed search strategies for all databases are provided in Multimedia Appendix 1. No previously published search strategies were adopted or adapted. The search strategies were peer reviewed by an information specialist prior to implementation.

Study Selection

Duplicate records were removed using Covidence [41] prior to screening. The same platform was used to manage the study screening process. After removing duplicate studies, 3 authors (YG, DX, and XL) independently screened the remaining articles in pairs based on titles and abstracts to assess their eligibility. Due to language accessibility, only studies published in English were included at the eligibility stage. Following the initial screening, full-text reviews were conducted for the retained articles to finalize their inclusion, and the reference

XSL•FO
RenderX

lists of included studies were then manually examined to identify any potentially relevant research. Any discrepancy among reviewers was resolved through consultation with a fourth author (XQ).

Data Collection Process

Data extraction was performed independently by 3 reviewers (YG, DX, and XL) using a standardized extraction form. All extracted data were cross-checked by 1 reviewer (YG) to ensure accuracy and consistency, and discrepancies were resolved by discussion with a fourth reviewer (XQ).

Definitions for Data Extraction

Data extraction consisted of two parts: (1) basic study characteristics, including author, country, year of publication, outcome time perspective, study design, years of data collection, data source, single-center or multicenter design, sampling method, group, sample size, participant characteristics, outcome and definition, whether the outcome was first-time fall, number of fallers, and follow-up duration; and (2) model-related information, including methods for handling missing values, data-splitting strategy, feature selection methods, modeling techniques, final predictors used in the model, calibration, AUC, and other performance metrics.

Risk of Bias and Applicability

To assess the risk of bias and applicability of the included studies, the Prediction Model Risk of Bias Assessment Tool (PROBAST) was used [42]. PROBAST consists of 20 signaling questions across 4 domains: participants, predictors, outcome, and analysis. Each question can be answered as "yes," "probably yes," "no," "probably no," or "no information," where responses of "yes" or "probably yes" indicate low risk of bias, and "no" or "probably no" indicate high risk of bias. A domain is judged to be at high risk of bias if at least one signaling question is rated as "no" or "probably no." The overall risk of bias is considered low only when all domains are judged to have low risk. Three reviewers (YG, DX, and XL) independently evaluated the risk of bias and concerns regarding applicability using the PROBAST checklist. Discrepancies between reviewers were resolved through discussion.

Prediction Performance Measures

The primary performance metric extracted from each study was the AUC, which reflects the discrimination ability of prediction models for identifying community-dwelling older adults at risk of future falls. When available, additional performance metrics, such as accuracy, sensitivity, specificity, and calibration measures, were also recorded.

Synthesis of Results

All statistical analyses were conducted in R software (version 4.5.2; R Foundation for Statistical Computing) using the metafor package (version 4.8‑0 [43]), meta package (version 8.2-1 [44]), and pimeta package (version 1.1.3 [45]).

Studies were eligible for quantitative synthesis if they reported AUC values with corresponding 95% CIs. Studies that did not report AUC values or corresponding 95% CIs were excluded. As AUC is a bounded measure (0‑1), pooling was conducted on the logit-transformed scale to stabilize variances and improve the normality assumption. For each model, SEs of the logit-transformed AUC were derived from the reported 95% CIs. Specifically, the lower and upper limits of the reported CI were first transformed using the logit function. Under the assumption that the logit-transformed AUC follows an approximately normal distribution, the SE on the logit scale was then estimated as (logit upper limit−logit lower limit)/(2×1.96). This approach derives the SE directly on the transformed scale and avoids reliance on first-order delta method approximations, thereby better accounting for the nonlinear nature of the logit transformation. Pooled estimates were subsequently back-transformed to the original AUC scale using the inverse-logit function for interpretation. In total, 9 studies comprising 10 independent cohorts and their corresponding prediction models evaluated using internal validation were included in the quantitative synthesis. The only model evaluated using external validation, along with several other models, was excluded due to insufficient information to derive SEs.

Between-study heterogeneity was quantified using I^2, τ^2, τ, and Q statistic using the metafor package. Although I^2 values of 25%, 50%, and 75% are commonly interpreted as low, moderate, and high heterogeneity [46], respectively, I^2 alone does not convey the magnitude of dispersion across different populations or settings [47]. τ^2 represents the variance of true effects across studies, and τ represents the corresponding SD. The Q statistic (Cochran Q test) was used to formally assess whether observed variability in effect estimates exceeded that expected by chance, with a significant value indicating the presence of heterogeneity.

To estimate the expected range of model performance in future populations and better quantify the real-world implications of heterogeneity, 95% prediction intervals (PIs) were additionally calculated using the pimeta package. PIs reflect the variation in true effects across settings and indicate the range of effects that may be expected in a future study [48]. Four PI estimation methods were applied, including the Higgins-Thompson-Spiegelhalter (HTS) method, the Hartung-Knapp–adjusted method (HTS-HK), the Sidik-Jonkman method (HTS-SJ), and the confidence-distribution bootstrap method proposed by Nagashima et al [49]. As HTS-type PIs rely on plug-in estimation of the heterogeneity variance and large sample approximations, their coverage may be suboptimal when the number of studies is small. In contrast, the confidence-distribution bootstrap approach more fully accounts for uncertainty in τ^2 and has been shown to achieve coverage closer to the nominal level in meta-analyses with few studies. Therefore, the confidence-distribution bootstrap PI was selected as the primary PI displayed in the main forest plots, whereas the other PI methods were retained as supplementary analyses.

As several prediction models were derived from the same large population-based database (CHARLS), particular attention was paid to the potential risk of double-counting when synthesizing model performance. Following methodological recommendations highlighting that duplicate inclusion of the same evidence may lead to overstating the precision of pooled estimates in meta-analysis, each model was examined to determine whether it represented a duplicated analysis of the

XSL•FO
RenderX

same participants or a distinct analytic cohort or subpopulation [50]. Models derived from mutually exclusive subgroups or from different analytic cohorts with distinct modeling strategies were treated as separate analytical units.

Meta-Analysis

The metafor package was used for effect estimation, while the meta package was used for visualization of forest plots. Random-effects meta-analyses were performed using the Sidik-Jonkman estimator for between-study variance (method="SJ"), together with the Hartung-Knapp adjustment (test="knha"), known as the Hartung-Knapp-Sidik-Jonkman (HKSJ) method [51]. This approach yields more accurate CIs, produces more stable pooled estimates, and reduces false-positive rates compared with the traditional DerSimonian-Laird method, particularly when heterogeneity is substantial or the number of studies is small [51].

The primary meta-analysis included the best-performing model from each study cohort, as is commonly done in prediction model syntheses. To account for studies that developed multiple models from the same cohort, a within-study weighted average meta-analysis was additionally conducted. For each study cohort, model-level AUCs were first transformed to the logit scale and then combined using inverse-variance weighting to obtain a single cohort-level average AUC. This approach mitigates the potential inflation of pooled performance that may arise when only the best-performing model from each study is considered. The resulting cohort-level averages were then used in a random-effects meta-analysis with the same HKSJ framework.

Additional Analyses

Additional analyses included sensitivity analyses, assessments of potential small-study effects and publication bias, subgroup analyses, and meta-regression. All analyses were conducted using the metafor package.

Robustness was evaluated using leave-one-out sensitivity analyses, in which each study was sequentially excluded and the pooled effect recalculated. Additional predefined sensitivity analyses were conducted to evaluate the potential impact of correlated cohorts arising from the use of the CHARLS database. These analyses included sequential models retaining only one CHARLS-derived model at a time and an analysis excluding all CHARLS-derived models. Changes in pooled AUC estimates were examined to assess whether potential participant overlap materially influenced the overall results.

Potential small-study effects and publication bias were examined using funnel plots and Egger-type regression tests; the Begg test was not used due to its even lower statistical power in small samples. A nonsignificant P value from the Egger test indicates no strong evidence of funnel plot asymmetry; however, interpretation should be made with caution because both funnel plots and regression-based asymmetry tests have limited power and may be misleading when the number of studies is small [52]. Subgroup forest plots, leave-one-out influence plots, and funnel plots were generated using the meta package for visualization. A 2-tailed P value of <.05 was considered statistically significant.

Subgroup analyses were conducted using the metafor package, according to prediction time window (≤1 y vs >1 y), sample size (≤500 vs >500 participants), model modality (text-based vs sensor-based or multimodal inputs), and population subgroup (general vs specific conditions). Subgroup analyses by validation type (internal vs external), fall outcome definition (eg, first fall, multiple falls, or falls or syncope occurrence), model type (ML vs DL), and geographic region were not performed because these categories were represented by a single study or insufficient data were available for inclusion in quantitative synthesis. Within each subgroup, random-effects meta-analyses were performed using the same HKSJ framework, and subgroup-specific pooled AUCs with corresponding 95% CIs, heterogeneity statistics (I^2, τ^2, τ, and the Q statistic), and PIs were reported.

To formally assess differences between subgroups, mixed-effects meta-regression models were fitted, with subgroup indicators specified as categorical moderators. Between-study variance was estimated using the Sidik-Jonkman method, and statistical inference was based on the Hartung-Knapp adjustment. Differences between subgroups were evaluated using omnibus tests of moderators, and the proportion of between-study heterogeneity explained by each moderator was quantified using the R^2 statistic.

Results

Study Selection

Figure 1 summarizes the search and selection process. After removing duplicates, 6865 records were identified. Following title and abstract screening, 6796 (98.6%) articles were excluded. Of the 69 (1%) full-text articles assessed, 45 (65.2%) did not meet the inclusion criteria, such as focusing on real-time fall detection or inpatient falls, and were excluded. Updated searches were conducted on August 31, 2025, and February 28, 2026, which identified an additional 3388 records. After screening and eligibility assessment, 4 additional studies met the inclusion criteria. Ultimately, a total of 28 studies were included in this review, and 9 studies, including 10 models, were included in the meta-analysis.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart illustrating the identification, screening, eligibility assessment, and inclusion of longitudinal studies evaluating machine learning and deep learning models for predicting future falls among community-dwelling older adults. Searches were performed across PubMed, Embase, Web of Science Core Collection, CINAHL, Cochrane Library, and IEEE Xplore databases from inception to September 23, 2024, and updated on August 31, 2025, and February 28, 2026. AI: artificial intelligence.

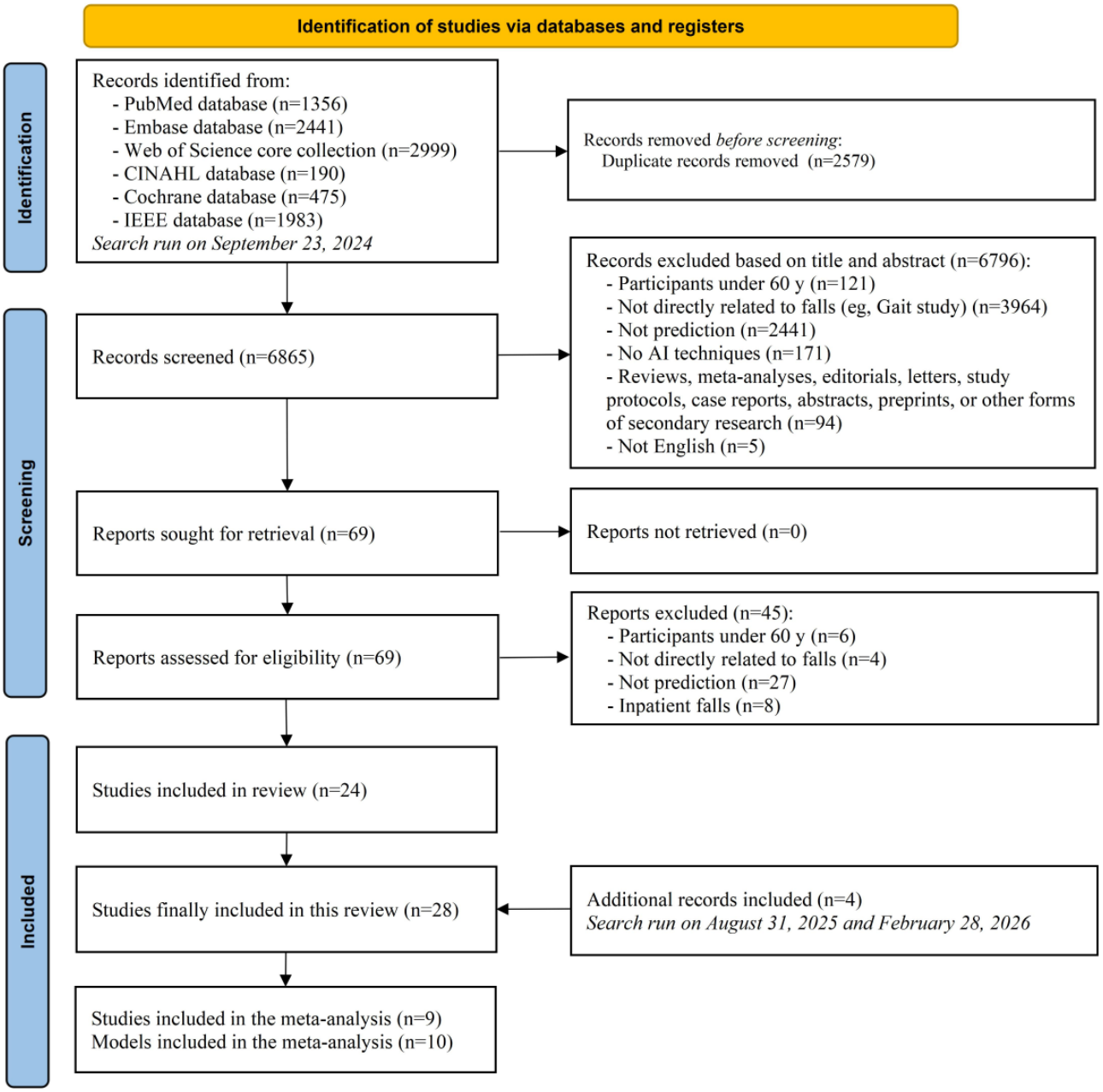

Study Characteristics

Table 1 summarizes the study design and participant characteristics of the 28 included studies. These studies were published between 2000 and 2026, with data originating from the United States (n=5, 17.9%), China (n=4, 14.3%), Japan (n=4, 14.3%), Canada (n=2, 7.1%), Mexico (n=2, 7.1%), Belgium (n=1, 3.6%), France (n=1, 3.6%), Germany (n=1, 3.6%), Brazil (n=1, 3.6%), Ireland (n=1, 3.6%), Italy (n=1, 3.6%), the Netherlands (n=1, 3.6%), Portugal (n=1, 3.6%), Sweden (n=1, 3.6%), the United Kingdom (n=1, 3.6%), and South Korea (n=1, 3.6%), concentrated in East Asia, North America, and Europe. The remaining regions are underrepresented (Figure 2).

XSL•FO
RenderX

Figure 2. Geographic distribution of included studies. This figure illustrates the geographic locations of participant recruitment in the included longitudinal studies, with participant data originating from the United States (n=5), China (n=4), Japan (n=4), Canada (n=2), Mexico (n=2), Belgium (n=1), France (n=1), Germany (n=1), Brazil (n=1), Ireland (n=1), Italy (n=1), the Netherlands (n=1), Portugal (n=1), Sweden (n=1), the United Kingdom (n=1), and South Korea (n=1), indicating that most studies were conducted in East Asia, North America, and Europe, while other global regions remain underrepresented.

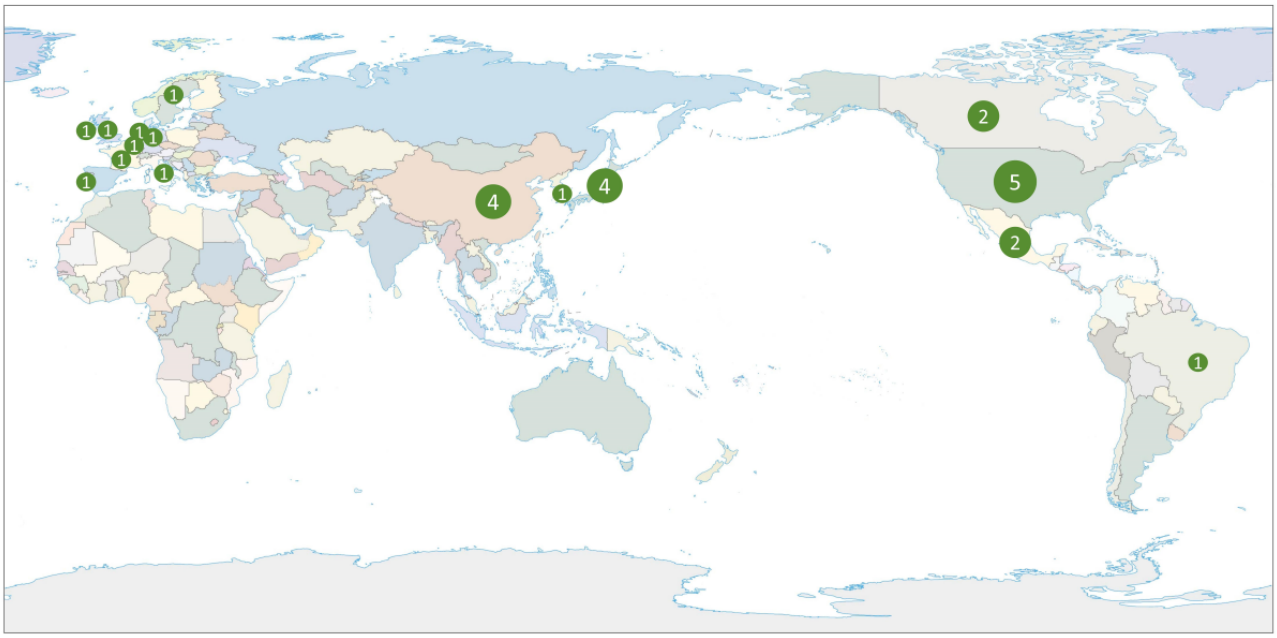

XSL•FO
RenderX

Table . Summary of study design and participant characteristics.

Author (year)	Country	Study design (collection period)	Data source	Single- or multi-center (number, if multi-center)	Sampling	Group	Sample size	Age (y)	Female (ratio), n (%)	Whether fall definition provided	Outcome (ascertainment method)	First fall	Participants (ratio) affected, n (%)	Follow-up
Wan et al (2026) [53]	China	Prospective cohort (2013 - 2018)	Public database (CHARLS[a])	Multi (28 provinces)	NR[b]	Sarcopenia	1087	71 (65 - 76)[c]	745 (68.5)	No	Fall occurrence (self-reported)	No	246 (22.6)	6 y
Lin et al (2025) [54]	China	Prospective cohort (2015 - 2018)	Public database (CHARLS)	Multi (28 provinces)	NR	—[d]	Rural: 2032; and urban: 3844	66 (63 - 72)[c]	2961 (50.4)	No	Fall occurrence (self-reported)	No	1317 (22.4)	3 y
Takeshita et al (2025) [55]	Japan	Prospective cohort (2019 - 2023)	EHR[e] and telephone survey	Multi (4 hospitals)	Hospital based	Hospitalization history	706	Fall: 84.7 (6.2)[f]; and nonfall: 83.4 (6.7)[f]	425 (60.2)	Yes	Fall occurrence (self-reported)	No	114 (16.1)	3 mo
Park et al (2025) [56]	South Korea	Prospective cohort (2016 - 2019)	Public database (KFACS[g])	Multi (NR)	NR	Cognitive frailty	443	70-84[h]	363 (81.9)	No	Fall occurrence (self-reported)	No	NR	2 y
Liu et al (2023) [57]	United States	Retrospective case-control (NR)	EHR	Single	Hospital based	Abdominal indications	9029	66.6 (9.7)[f]	5728 (63.4)	No	Fall occurrence (EHR record)	No	3535 (39.2)	6.5 y
Silveira et al (2023) [58]	Brazil	Retrospective case-control (NR)	Clinical records and fitness center assessments	Multi (2 centers)	Clinic based	CVD[i]	72	Fall: 76.5 (65-92)[c]; and nonfall: 73 (60-86)[c]	29 (40.3)	Yes	Fall occurrence (self- or proxy-reported with staff monitoring)	No	24 (33.3)	1 y
Chen et al (2023a) [24]	China	Prospective cohort (2015 - 2018)	Public database (CHARLS)	Multi (28 provinces)	NR	—	5818	67.7 (6.2)[f]	2982 (51.3)	No	Fall occurrence (self-reported)	No	1347 (23.2)	3 y
Chen et al (2023b) [28]	China	Prospective cohort (2011 - 2018)	Public database (CHARLS)	Multi (28 provinces)	NR	—	1617	67.7 (6.2)[f]	697 (43.1)	No	Fall occurrence (self-reported)	No	235 (14.5)	7 y

XSL•FO
RenderX

Author (year)	Country	Study design (collection period)	Data source	Single- or multi-center (number, if multi-center)	Sampling	Group	Sample size	Age (y)	Female (ratio), n (%)	Whether fall definition provided	Outcome (ascertainment method)	First fall	Participants (ratio) affected, n (%)	Follow-up
Dormosh et al (2023) [59]	The Netherlands	Retrospective cohort (2018 - 2019)	EHR	Multi (50 general practices)	Registry based	—	36,470	Fall: 76.6 (70.7-83.3)[c]; and nonfall: 71.4 (68-77.1)[c]	19,398 (53.2)	No	Fall occurrence (EHR record, chart review)	No	4778 (13.1)	1 y
Ramsdale et al (2023) [60]	United States	Prospective cohort (NR)	Public database (GAP[j]70+ study)	Multi (NR)	Cluster randomized	Advanced cancer	522	≥70[h]	NR	No	Fall occurrence (self-reported)	No	89 (17)	3 mo
Ikeda et al (2022) [61]	Japan	Prospective cohort (2010 - 2013)	Public database (JAGES[k])	Multi (24 municipalities)	Registry based	—	61,883	Multiple falls: 75.4 (6.1)[f]; and no or annual fall: 72.8 (5.5)[f]	33,242 (53.7)	No	Multiple falls or not (self-reported)	No	3359 (5.4)	3 y
Mishra et al (2022) [30]	United States	Retrospective cohort (NR)	EHR	Single	Facility based	—	92	86.2 (6.4)[f]	57 (62)	No	Fall occurrence (staff reported)	No	31 (33.7)	6 mo
Kelly et al (2022) [62]	Sweden	Prospective cohort (NR)	Self-collected	Single	NR	—	1705	70	817 (47.9)	Yes	Fall occurrence (self-reported)	No	255 (15)	1 y
Dasgupta et al (2022) [29]	United States	Prospective cohort (2019)	Self-collected	Single	Convenience sampling	Hospitalization history	134	68.9 (8.1)[f]	54 (40.3)	Yes	Fall occurrence (self-reported)	No	14 (10.4)	3 mo
Tang and Romero-Ortuno (2022) [27]	Ireland	Prospective cohort (2009 - 2016)	Public database (TILDA[l])	Multi (NR)	NR	—	2900	≥65[h]	NR	Yes	Falls or syncope occurrence (self-reported)	No	1351 (46.6)	6 y
Cuaya-Simbro et al (2021) [63]	Mexico	Prospective cohort (NR)	Self-collected	Single	Clinic based	Women with osteoporosis	126	74.3 (6.3)[f]	126 (100)	No	Fall occurrence (self-reported)	No	NR	2.5 y

XSL•FO
RenderX

Author (year)	Country	Study design (collection period)	Data source	Single- or multi-center (number, if multi-center)	Sampling	Group	Sample size	Age (y)	Female (ratio), n (%)	Whether fall definition provided	Outcome (ascertainment method)	First fall	Participants (ratio) affected, n (%)	Follow-up
Omae et al (2021) [64]	Japan	Prospective cohort (2017 - 2018)	Self-collected	Single	Community health check-up based	—	577	79 (77-83)[c]	304 (52.7)	No	Fall occurrence (self-reported)	No	201 (34.8)	1 y
Makino et al (2021) [65]	Japan	Prospective cohort (2011 - 2016)	Self-collected	Single	NR	—	2520	71.1 (4.7)[f]	1303 (51.7)	Yes	Fall occurrence (self-reported)	No	415 (16.5)	4 y
Cuaya-Simbro et al (2020) [66]	Mexico	Prospective cohort (2019)	Self-collected	Multi (2 centers)	NR	Women	253	≥60[h]	253 (100)	No	Fall occurrence (self-reported)	No	First 6 mo: 38 (15%); Second 6 mo: 26 (10.3%)	1 y
Cella et al (2020) [67]	Italy	Prospective cohort (NR)	Self-collected	Single	Clinic based	—	96	77.2 (6.5)[f]	62 (64.6)	Yes	Fall occurrence (diary record)	No	32 (33.3)	1 y
Silva et al (2020) [68]	Portugal	Prospective cohort (NR)	Self-collected	Single	NR	—	281	75.1 (6.9)[f]	183 (65.1)	No	Fall occurrence (self-reported)	No	74 (26.3)	1 y
Ye et al (2020) [22]	United States	Retrospective cohort (2016 - 2018)	EHR	Multi (69 facilities)	Registry based	—	265,225	≥65[h]	NR	Yes	Fall occurrence (EHR record, chart review)	No	4361 (1.6)	1 y
Kuspinar et al (2019) [69]	Canada	Prospective cohort (2002 - 2014)	Administrative data	Multi (4 provinces)	Registry based	—	126,703	77 (14)[f]	82,357 (65)	No	Fall occurrence (self-reported)	No	NR	6 mo
Gillain et al (2019) [70]	Belgium	Prospective cohort (2014 - 2017)	Self-collected	Single	Volunteer based	—	96	71.3 (5.4)[f]	48 (50)	Yes	Fall occurrence (diary record)	No	35 (36.5)	2 y
Howcroft et al (2018) [71]	Canada	Prospective cohort	Self-collected	Single	Convenience sampling	—	75	75.2 (6.6)[f]	44 (58.7)	Yes	Fall occurrence (self-reported)	No	28 (37.3)	6 mo

XSL•FO
RenderX

Author (year)	Country	Study design (collection period)	Data source	Single- or multi-center (number, if multi-center)	Sampling	Group	Sample size	Age (y)	Female (ratio), n (%)	Whether fall definition provided	Outcome (ascertainment method)	First fall	Participants (ratio) affected, n (%)	Follow-up
Deschamps et al (2016) [72]	France	Prospective cohort	Self-collected	Multi (NR)	NR	—	426	69.5 (2.6)[f]	265 (62.2)	Yes	Fall occurrence (self-reported, adjudicated by geriatric committee)	Yes	82 (19.2)	1 y
Maschke et al (2011) [73]	Germany	Prospective cohort (2007 - 2008)	Self-collected	Single	Hospital based	Hospitalization history	46	81.3[c]	37 (80.4)	Yes	Fall occurrence (self-reported)	No	19 (41.3)	1 y
Bath et al (2000) [74]	United Kingdom	Prospective cohort (1985 - 1989)	Self-collected	Single	Random sampling	—	435	≥65[h]	NR	No	Fall occurrence (self-reported)	No	114 (26.2)	4 y

[a]CHARLS: China Health and Retirement Longitudinal Study.

[b]NR: not reported.

[c]Median (IQR).

[d]Not applicable.

[e]EHR: electronic health record.

[f]Mean (SD).

[g]KFACS: Korean Frailty and Aging Cohort Study.

[h]Age range.

[i]CVD: cardiovascular disease.

[j]GAP: geriatric assessment for patients.

[k]JAGES: Japan Gerontological Evaluation Study.

[l]TILDA: The Irish Longitudinal Study on Ageing.

Among them, 24 studies (85.7%) adopted a prospective design, while 4 (14.3%) were retrospective. Two (7.1%) were case-control studies [57,58], while others (92.9%) were cohort studies. Regarding data sources, 13 (46.4%) studies relied on investigator-collected datasets, 8 (28.6%) used public databases, 6 (21.4%) used electronic health records or clinical records, and 1 (3.6%) used administrative data. Thirteen (46.4%) studies were conducted in single-center settings, and 15 (53.6%) were multicenter studies.

With regard to the study population, 18 (64.3%) studies focused on general community-dwelling older adults, with fall incidence ranging from 1.6% to 46.6%. The remaining 10 studies targeted populations with specific conditions, including hospitalization history (n=3, 10.7%) [29,55,73], sarcopenia (n=1, 3.6%) [53], cognitive frailty (n=1, 3.6%) [56], abdominal indications (n=1, 3.6%) [57], cardiovascular disease (n=1, 3.6%) [58], advanced cancer (n=1, 3.6%) [60], women with osteoporosis (n=1, 3.6%) [63], and women in general (n=1, 3.6%) [66]. Sample sizes for model development ranged from 46 [73] to 265,225 [22]. The mean or median age of participants ranged from 66 [54] to 86.2 [30] years, and the proportion of female participants ranged from 40.3% [29,58] to 100% [63,66].

Twelve (42.9%) studies explicitly reported a precise fall definition, whereas the remaining 16 (57.1%) did not. Most studies (n=26, 92.9%) used fall occurrence as the outcome, and most outcomes were ascertained exclusively via self-report (n=20, 71.4%). Only one study (3.6%) [72] specifically predicted first-time falls.

Model Development and Validation Characteristics

Table 2 summarizes the model development and validation characteristics of the included studies. Eleven (39.3%) studies reported internal validation split ratios, most commonly a 70% training and 30% test set (n=6, 54.5%). Eighteen (64.3%) studies applied cross-validation, with 10-fold cross-validation being the most frequently used approach (n=12, 66.7%). Among the

XSL•FO
RenderX

23 studies that reported feature selection methods, the most frequently applied was decision tree-based selection (n=4, 17.4%).

A total of 23 (82.1%) studies used ML techniques, while 5 (17.9%) applied dL. The most frequently used modeling approaches were decision tree (n=6, 21.4%) and RF (n=6, 21.4%).

Data modality included text, sensor, image, and multimodal sources. More than half of the studies (n=18, 64.3%) relied solely on conventional text data sources, including demographic variables, functional status, health conditions, psychological factors, medication use, behavioral factors, home environment, socioeconomic indicators, and prior fall history. The remaining studies incorporated sensor data (n=5, 17.9%), image data (n=1, 3.6%), or multimodal data (n=4, 14.3%).

XSL•FO
RenderX

Table . Summary of missing data handling methods, feature selection techniques, modeling approaches, data modalities, validation strategies, calibration assessments, and discrimination performance of included machine learning and deep learning models developed to predict future falls among community-dwelling older adults[a].

Author (year)	Missing data	Data split ratio (train:test)	Cross-validation	Feature selection	Modeling	Data modality	Calibration, 95% CI or mean (SD)	AUC[b] (95% CI)	Other discrimination metrics, 95% CI or mean (SD)
Wan et al (2026) [53]	Excluding, multiple imputation	70:30:00	10-fold	RF[c]	RF[d]	Text	NR[e]	0.97 (0.01)	Acc[f]: 92.26 (1.38); Sens[g]: 89.31 (2.85); Spec[h]: 95.19 (0.92); Precision: 94.86 (1.03); F_1: 91.98 (1.55)
Lin et al (2025) [54]	Excluding, algorithm imputation	70:30:00	10-fold	LASSO[i]	RF[d]	Text	Rural: Brier score: 0.150; urban: Brier score: 0.159	Rural: 0.732 (0.685 - 0.782); urban: 0.724 (0.687 - 0.757)	Rural: Acc: 0.667; Sens: 0.669; Spec: 0.667; F_1: 0.456; and urban: Acc: 0.754; Sens: 0.519; Spec: 0.825; F_1: 0.495
Takeshita et al (2025) [55]	KNN[j], mode	70:30:00	5-fold	Information gain	Extra trees[d]	Text	NR	0.73[d]	Acc: 0.69; Sens: 0.68; Spec: 0.69; PPV[k]: 0.29; NPV[l]: 0.92; F_1: 0.41
Park et al (2025) [56]	NR	NR	NR	RFE[m]	LR[n]	Text	NR	0.956 (0.953 - 0.958)	Acc: 0.916 (0.912 - 0.920); Sens: 0.937 (0.933 - 0.941); Spec: 0.910 (0.906 - 0.914)
Liu et al (2023) [57]	Excluding	NR	NR	NR	U-Neta	Image	NR	0.657	NR
Silveira et al (2023) [58]	Excluding	NR	5-fold	RFE-RF	RF[c]	Text	NR	NR	Acc: 0.76; Sens: 0.68
Chen et al (2023[o]) [24]	MissForest algorithm for imputation	70:30:00	10-fold	LASSO	LR[d]	Text	Brier score: 0.158	0.739 (0.690 - 0.777)	Acc: 0.671; Sens: 0.707; Spec: 0.659
Chen et al (2023[d]) [28]	Multiple imputation	70:30:00	10-fold	NR	RF[d]	Text	Brier score: 0.170 (0.163 - 0.176)	0.731 (0.721 - 0.735)	Acc: 0.830 (0.824 - 0.831); Sens: 0.545 (0.544 - 0.546); PPV: 0.667 (0.664 - 0.668)
Dormosh et al (2023) [59]	Excluding	NR	10-fold	LASSO	LR	Text	Calibration plots	0.718 (0.708 - 0.727)	NR

XSL•FO
RenderX

Author (year)	Missing data	Data split ratio (train:test)	Cross-validation	Feature selection	Modeling	Data modality	Calibration, 95% CI or mean (SD)	AUC[b] (95% CI)	Other discrimination metrics, 95% CI or mean (SD)
Ramsdale et al (2023) [60]	Excluding, Median imputation	80:20:00	NR	Ensemble feature selection	MLP[d,o,p]	Text	NR	0.75	NR
Ikeda et al (2022) [61]	Excluding, RF imputation	NR	Nested 10-fold	RF-based Boruta algorithm	XGBoost[d,q]	Text	NR	0.88±0.02	Acc: 0.88 (0.02); F_1: 0.89 (0.02)
Mishra et al (2022) [30]	Excluding	NR	5-fold	NR	SVM[d,r]	Text	Brier score: 0.15	0.80 (0.76 - 0.85)	Acc: 0.75 (0.72 - 0.79); Sens: 0.82 (0.74 - 0.89); Spec: 0.72 (0.67 - 0.76); F_1: 0.76 (0.72 - 0.79)
Kelly et al (2022) [62]	NR	75:25:00	10-fold	Automated machine learning with Bayesian optimization	SGD[d,s]	Sensor	NR	NR	Sens: 0.61 (0.49 - 0.71); Spec: 0.66 (0.61 - 0.71)
Dasgupta et al (2022) [29]	No missing	NR	NR	Wald test	HCRNN[o,t]	Sensor	NR	0.99 (0.98 - 1)	NR
Tangand Romero-Ortuno (2022) [27]	Excluding	NR	NR	NR	RF	Text	NR	NR	Precision: 0.63; Sens: 0.56; F_1: 0.59
Cuaya-Simbro et al (2021) [63]	NR	NR	10-fold	FSMC[u]	RF[d]	Sensor	NR	NR	Sens: >0.71; Spec: >0.18; PPV: >0.74; NPV: >0.66
Omae et al (2021) [64]	Multiple imputation	80:20:00	NR	RF	Decision tree	Text	NR	0.818	Acc: 0.836; Sens: 0.884; Spec: 0.766; PPV: 0.847; NPV: 0.818
Makino et al (2021) [65]	Excluding	NR	10-fold	Decision tree	Decision tree[d]	Text	NR	0.70 (0.68 - 0.72)	Acc: 0.65 (0.64 - 0.67); Sens: 0.62 (0.60 - 0.64); Spec: 0.69 (0.67 - 0.71); PPV: 0.66 (0.64 - 0.69); NPV: 0.64 (0.62 - 0.66)

XSL•FO
RenderX

Author (year)	Missing data	Data split ratio (train:test)	Cross-validation	Feature selection	Modeling	Data modality	Calibration, 95% CI or mean (SD)	AUC[b] (95% CI)	Other discrimination metrics, 95% CI or mean (SD)
Cuaya-Simbro et al (2020) [66]	NR	NR	10-fold	FSMC	DBN[d,o,v]	Sensor	NR	NR	First 6 months: Acc: 0.800; Sens: 0.737; Precision: 0.700; F1: 0.718; and second 6 months: Acc: 0.882; Sens: 0.626; Precision: 0.831; F_1: 0.714
Cella et al (2020) [67]	NR	NR	Nested 5-fold	NR	LASSO	Text, sensor	NR	0.81 (0.72 - 0.90)	Precision: 0.69; Sens: 0.78; Spec: 0.74
Silva et al (2020) [68]	NR	67:33:00	NR	PCA[w] and variance threshold method	Decision tree[d]	Text, sensor	NR	0.505 (0.043)	Acc: 0.370 (0.098); PPV: 0.262 (0.032); Sens: 0.786 (0.222); Spec: 0.225 (0.199); F_1: 0.382 (0.060)
Ye et al (2020) [22]	NR	67:33:00	10-fold	Univariate logistic regression	XGBoost[d]	Text	Isotonic regression	0.807	PPV: 0.110
Kuspinar et al (2019) [69]	NR	70:30:00	NR	Decision tree	Decision tree	Text	NR	Internal: 0.60; external: 0.55, 0.58, 0.58	NR
Gillain et al (2019) [70]	NR	NR	10-fold	Decision tree	Decision tree	Text, sensor	NR	0.84	Acc: 0.84; Sens: 0.80; Spec: 0.87; PPV: 0.78; NPV: 0.88
Howcroft et al (2018) [71]	NR	75:25:00	10-fold	Relief-F, fast correlation–based filter	SVM[d]	Sensor	NR	NR	Acc: 0.56; Sens: 0.59; Spec: 0.69; PPV: 0.56; NPV: 0.73; F_1: 0.56
Deschamps et al (2016) [72]	Algorithm	80:20:00	NR	Decision tree	Decision tree	Text	NR	0.72	Sens: 0.82; Spec: 0.61
Marschollek et al (2011) [73]	NR	NR	10×10-fold	Wrapper subset evaluator	LR[d]	Text, sensor	Brier score: 0.20	0.74	Acc: 0.72; Sens: 0.68; Spec: 0.74; PPV: 0.65; NPV: 0.77

XSL•FO
RenderX

Author (year)	Missing data	Data split ratio (train:test)	Cross-validation	Feature selection	Modeling	Data modality	Calibration, 95% CI or mean (SD)	AUC[b] (95% CI)	Other discrimination metrics, 95% CI or mean (SD)
Bath et al (2000) [74]	NR	NR	LOOCV[x]	GANN[y]	GANN[o]	Text	NR	NR	Acc: 0.76; Sens: 0.31; Spec: 0.92; PPV: 0.57; NPV: 0.79

[a]For studies discussing multiple models, only the best-performing model and its corresponding performance metrics are presented. The AUC was reported based on validation data when available. Only Kuspinar et al (2019) [69] reported external validation. Some studies reported AUC with 95% CI and others as mean (SD); these formats represent different types of variability and cannot be interconverted. Values are therefore presented as originally reported. Accuracy and F_1-score are threshold- and prevalence-dependent measures. When reported without accompanying discrimination (eg, AUC) and calibration metrics, their interpretability and comparability across studies are limited.

[b]AUC: area under the curve.

[c]RF: random forest.

[d]Studies reporting multiple models trained on the same dataset; the listed modeling method and performance correspond to the best-performing model.

[e]NR: not reported.

[f]Acc: accuracy.

[g]Sens: sensitivity.

[h]Spec: specificity.

[i]LASSO: Least Absolute Shrinkage and Selection Operator.

[j]KNN: K-nearest neighbor.

[k]PPV: positive predictive value.

[l]NPV: negative predictive value.

[m]RFE: recursive feature elimination.

[n]LR: logistic regression.

[o]Deep learning.

[p]MLP: multilayer perceptron.

[q]XGBoost: Extreme Gradient Boosting.

[r]SVM: support vector machine.

[s]SGD: stochastic gradient descent.

[t]HCRNN: hybrid-convolutional recurrent neural network.

[u]FSMC: feature selection for minority class.

[v]DBN: dynamic Bayesian network.

[w]PCA: principal component analysis.

[x]LOOCV: Leave-one-out cross-validation.

[y]GANN: genetic algorithm neural network.

Table S1 in Multimedia Appendix 1 presents the full details of the final predictors in each model. Table 3 concludes predictors used in at least three models, with the 5 most frequent being age, sex, fall history, depression, and basic activities of daily living (BADL-ADL), appearing in 15, 13, 12, 11, and 9 models, respectively. Other frequently included predictors were walking speed, chair stand test results, cognition, hypertension, fracture history, vision, and the use of cardiovascular drugs.

XSL•FO
RenderX

Table . Summary of the frequently selected predictors in included machine learning–based and deep learning–based fall risk prediction models among community-dwelling older adults.

Category and predictors[a]	Participants, n
Demographic factors	
Age	15
Sex	13
Marital status	3
Functions, gait, and balance	
BADL[b] or ADL[c,d]	9
Walking/gait speed	6
Chair stands test/rising from a chair	6
TUG[e]	5
IADL[f,g]	5
Grip or hand strength	4
Gait-related parameters (sensor)	4
Balance-related parameters (sensor)	4
Balance/Berg balance scale	3
Assistance with walking	3
Comorbidity and health status	
Cognition	6
Hypertension	6
Fracture or injury history	6
Vision or vision impairment	6
BMI	5
Diabetes	5
Arthritis or osteoarthritis	5
Incontinence	5
Memory	5
Pain	5
Cancer	4
Stroke	4
Hearing or hearing impairment	4
CVD[h]	3
Parkinson disease	3
Osteoporosis	3
Self-rated health	3
Teeth or teeth number	3
Psychological factor	
Depression	11
Lifestyles	
Sleeping duration or time asleep	5
Smoking	4
Drinking or alcohol abuse	3

XSL•FO
RenderX

Category and predictors[a]	Participants, n
Medication-related factors	
Cardiovascular drugs	6
Polypharmacy	5
Psychiatric and neurological drugs	5
Number of taken medications	4
Fall-related factors	
Fall history	12
Fear of falling	3

[a]Predictors were extracted from the final predictive model reported in each included longitudinal study. Full details of the final predictors in each model are presented in Table S1 in Multimedia Appendix 1.

[b]BADL: basic activity of daily living.

[c]ADL: activity of daily living.

[d]ADL and BADL were grouped together as they both reflect core self-care functions in the included studies.

[e]TUG: Timed Up and Go.

[f]IADL: instrumental activity of daily living.

[g]IADL was presented separately due to its distinct role in assessing instrumental daily living skills.

[h]CVD: cardiovascular disease.

Results of Individual Studies

Table 2 also summarizes the performance of the included prediction models. In terms of discrimination, 20 models (71.4%) reported AUC or c-index values, which ranged from 0.505 [68] to 0.99 [15]. Calibration was assessed in 7 (25%) studies [22,24,28,30,54,59,73], with 5 (17.9%) reporting Brier scores ranging from 0.15 [30,54] to 0.20 [73]. Only one model underwent external validation [69], while 2 models did not report either internal or external validation [27,57].

Figure S1 in Multimedia Appendix 1 shows that among studies reporting both the sample and AUC, over half of the studies (n=14/21, 66.7%) used classic ML with text data across all sample sizes, yielding median AUCs of approximately 0.73 to 0.81. DL evidence was limited, appearing mainly in nontext modalities and in studies with moderate sample sizes, with variable median AUCs (0.66 - 0.99) based on few studies.

Risk of Bias and Applicability

Table 4 presents the risk of bias and applicability assessments for the included prediction models. All models were assessed as having a high risk of bias, suggesting methodological limitations during model development and/or validation. In the participants domain, 5 models were thought to be at high risk of bias due to retrospective data sources [22,30,57-59]. In the predictors domain, 11 models were rated as high risk, primarily due to inadequate predictor definitions, measurement methods, or failure to clearly report which variables were ultimately included in the final model [22,24,27-29,53,59,60,62,63,68].

XSL•FO
RenderX

Table . Risk of bias and applicability assessment of included machine learning and deep learning models for predicting future falls in community-dwelling older adults, evaluated using the PROBAST[a].

Author (year)	Study type	ROB[b]				Applicability			Overall	
		Participants	Predictors	Outcome	Analysis	Participants	Predictors	Outcome	ROB	Applicability
Wan et al (2026) [53]	B[c]	+[d]	-[e]	-	-	-	+	+	-	-
Lin et al (2025) [54]	B	+	+	-	-	+	+	+	-	+
Takeshita et al (2025) [55]	B	+	+	-	-	-	+	+	-	-
Park et al (2025) [56]	B	+	+	-	-	-	+	+	-	-
Liu et al (2023) [57]	A[f]	-	+	-	-	-	+	+	-	-
Silveira et al (2023) [58]	B	-	+	-	-	-	+	+	-	-
Chen et al (2023a) [24]	B	+	-	-	-	+	-	+	-	-
Chen et al (2023b) [28]	B	+	-	-	-	+	-	+	-	-
Dormosh et al (2023) [59]	B	-	-	-	-	+	-	+	-	-
Ramsdale et al (2023) [60]	B	+	-	-	-	-	-	+	-	-
Ikeda et al (2022) [61]	B	+	+	-	-	+	+	-	-	-
Mishra et al (2022) [30]	B	-	+	-	-	+	+	+	-	+
Kelly et al (2022) [62]	B	+	-	-	-	+	-	+	-	-
Dasgupta et al (2022) [29]	B	+	-	-	-	-	-	+	-	-
Tang and Romero-Ortuno (2022) [27]	A	+	-	-	-	+	-	+	-	-
Cuaya-Simbro et al (2021) [63]	B	+	-	-	-	-	-	+	-	-
Omae et al (2021) [64]	B	+	+	-	-	+	+	+	-	+
Makino et al (2021) [65]	B	+	+	-	-	+	+	+	-	+
Cuaya-Simbro et al (2020) [66]	B	+	+	-	-	-	+	+	-	-

XSL•FO
RenderX

Author (year)	Study type	ROB[b]				Applicability			Overall	
		Participants	Predictors	Outcome	Analysis	Participants	Predictors	Outcome	ROB	Applicability
Cella et al (2020) [67]	B	+	+	-	-	+	+	+	-	+
Silva et al (2020) [68]	B	+	-	-	-	+	-	+	-	-
Ye et al (2020) [22]	B	-	-	?	-	+	-	+	-	-
Kuspinar et al (2019) [69]	C[g]	+	+	-	-	+	+	+	-	+
Gillain et al (2019) [70]	B	+[d]	+	-[e]	-	+	+	+	-	+
Howcroft et al (2018) [71]	B	+	+	?[h]	-	+	+	+	-	+
Deschamps et al (2016) [72]	B	+	+	-	-	+	+	+	-	+
Marschollek et al (2011) [73]	B	+	+	-	-	-	+	+	-	-
Bath et al (2000) [74]	B	+	+	-	-	+	+	+	-	+

[a]PROBAST: Prediction model Risk of Bias Assessment Tool.

[b]ROB: risk of bias.

[c]B: "development and internal or cross-validation in the same publication."

[d]+: low ROB or low concern regarding applicability.

[e]-: high ROB or high concern regarding application.

[f]A: "development only."

[g]C: "development, internal validation, and external validation in the same publication."

[h]?: unclear ROB or unclear concern regarding applicability.

In the outcome domain, all models were assessed as having either a high or unclear risk of bias, with all studies lacking blind assessments between predictors and the outcome. Sixteen (57.1%) studies only reported fall occurrence as the outcome without providing a precise definition of a fall [24,28,30,53,54,56,57,59-61,63,64,66,68,69,74], and 20 (71.4%) studies relied on self-reported or caregiver-reported outcome records, introducing potential bias [24,27-29,53-56,60-69,72,74].

Regarding the analysis domain, all models were classified as high risk of bias. Twelve (42.9%) studies had fewer than 10 events per variable, indicating insufficient sample size for reliable model estimation [28-30,53,55,58,66,67,70-72,74]. Six (21.4%) studies excluded cases with missing data [27,30,57-59,65], and 11 (39.3%) studies did not describe their methods for handling missing data [22,56,62,63,66-71,74]. Seven (25%) studies did not report the AUC or c-index [27,58,62,63,66,71,74], and 9 (32.1%) studies reported insufficient details regarding other model discrimination metrics [22,29,57-61,69,72]. Calibration assessments were absent in 21 (75%) studies [27,29,53,55-58,60-72,74], and 6 (21.4%) studies had potential risks of overfitting, underfitting, or optimism in model performance [27,53,56,57,68,69]. In terms of validation, only one model underwent external validation [69], while 2 (7.1%) models lacked both internal and external validation [27,57]. The remaining models reported internal validation only, of which 13 (26.4%) relied on random data splitting rather than more robust approaches such as cross-validation or bootstrapping.

Concerning applicability, 18 (64.3%) studies were considered to be of high concern, while 10 (35.7%) were rated as low concern. In the participants domain, 10 (35.7%) studies exhibited high applicability concerns due to the restriction of the study sample to specific subgroups, such as older adults with particular diseases, or single-sex populations [29,53,55-58,60,63,66,73]. In the predictors domain, 10 (35.7%) studies were rated as high concern due to unclear definitions, final inclusion, or measurement of predictors [22,24,27-29,59,60,62,63,68]. In the outcome domain, one study demonstrated high applicability concern due to the construction of a binary outcome variable based on multiple falls versus a combination of single fall and nonfall cases, thereby limiting real-world interpretability [61]. On the basis of the recurrent sources of bias identified by

XSL•FO
RenderX

PROBAST in the included studies, we summarize key methodological considerations as a more detailed checklist to inform future research in ML-based fall risk prediction (Table S5 in Multimedia Appendix 1).

Synthesis of Results

The data extraction sheets underlying the meta-analysis are provided in Table S6 in Multimedia Appendix 1. Ten models evaluated using internal validation were included in the primary quantitative synthesis, with the AUC used as the effect size to assess the discriminative performance of these ML-based fall risk prediction models. As only one model underwent external validation and did not provide adequate information to compute SE, it was excluded from the quantitative synthesis and instead summarized qualitatively. The pooled discriminative performance was good, with a back-transformed pooled AUC of 0.79 (95% CI 0.69 - 0.87) under the HKSJ random-effects model. Substantial between-study heterogeneity was present (τ^2=0.64; τ=0.80; I^2=99.8%; Q=4128.99; P<.001; Figure 3A [24,28-30,54,56,59,65,67]).

XSL•FO
RenderX

Figure 3. Random-effects meta-analysis of model discrimination performance using the best-performing model from each cohort. (A) Forest plot presenting individual model areas under the curve (AUCs) with corresponding 95% CIs and the pooled AUC estimated using the Hartung-Knapp-Sidik-Jonkman (HKSJ) method (back-transformed from the logit scale). Between-study heterogeneity is quantified using I^2, τ^2, τ, and Q statistic. 95% Prediction interval (PI; confidence-distribution [CD] bootstrap) is also reported in the pooled summary to illustrate the expected range of model performance in future populations. (B) Leave-one-out sensitivity analysis, in which the pooled AUC is recalculated after sequential exclusion of each model to assess the robustness of the overall estimate. (C) Funnel plot of logit-transformed AUC values to evaluate potential small-study effects [24,28-30,54,56,59,65,67].

XSL•FO
RenderX

To further characterize the expected dispersion of model performance in new populations, the main forest plot presents the 95% PI estimated using the confidence-distribution bootstrap method, which ranged from 0.20 to 0.99 (Figure 3A). PIs estimated using alternative methods were similarly wide (HTS: 0.28 - 0.98; HTS-HK: 0.40 - 0.96; and HTS-SJ: 0.40 - 0.96), leading to the same qualitative interpretation of substantial uncertainty in real-world generalizability.

Within-study averaging meta-analysis showed that the pooled discriminative performance remained comparable to the primary model-level analysis using best-performing models, with a back-transformed pooled AUC of 0.77 (95% CI 0.65 - 0.86) under the HKSJ random-effects model. Between-study heterogeneity was similarly high (τ^2=0.74; τ=0.86; I^2=99.9%; Q=6353.03; P<.001), indicating substantial true variability across study cohorts (Figure S2A in Multimedia Appendix 1). The supplementary forest plot presents the 95% PI estimated using the confidence-distribution bootstrap method, which ranged from 0.29 to 0.97 (Figure S2A in Multimedia Appendix 1). PIs estimated using alternative methods were similarly wide (HTS: 0.35 - 0.96; HTS-HK: 0.35 - 0.96; HTS-SJ: 0.35 - 0.96), supporting the same qualitative interpretation of substantial uncertainty in expected performance across new populations.

Additional Analyses

Leave-one-out sensitivity analyses for the primary analysis showed that the pooled AUC was stable (range: 0.75 - 0.80), suggesting that no single study disproportionately influenced the overall estimate (Figure 3B). As several included models were derived from the same large population-based database (CHARLS), additional sensitivity analyses were conducted to evaluate the potential impact of correlated cohorts on the pooled estimates. Sequential analyses retaining only one CHARLS-derived model at a time yielded consistent results, with pooled AUC estimates of 0.82 (95% CI 0.65 - 0.92) when retaining the rural model developed by Lin et al [54], 0.82 (95% CI 0.65 - 0.92) when retaining the urban model developed by Lin et al [54], 0.82 (95% CI 0.66 - 0.92) when retaining the model developed by Chen et al [24], and 0.82 (95% CI 0.65 - 0.92) when retaining the model developed by Chen et al [28]. An additional sensitivity analysis excluding all 4 CHARLS-derived models also produced a comparable pooled estimate (AUC 0.84, 95% CI 0.63 - 0.94), despite the reduced number of models (n=6). These findings indicate that the overall pooled discrimination estimate was robust and not materially driven by models derived from the same database.

Funnel plot inspection for the primary analysis did not reveal pronounced asymmetry, and the Egger regression test was nonsignificant (t_8=0.16; P=.88). However, these findings should be interpreted cautiously because asymmetry tests have limited power and are difficult to interpret when the number of studies is small (Figure 3C).

Similarly, for the within-study averaging meta-analysis, leave-one-out sensitivity analyses showed minimal influence of any single study cohort, with pooled AUC estimates remaining stable across iterations (range: 0.72 - 0.78; Figure S2B in Multimedia Appendix 1). Funnel plot inspection and Egger regression test likewise did not suggest strong evidence of small-study effects (t_8=1.33; P=.22), although interpretation remains limited by the small number of samples and high heterogeneity (Figure S2C in Multimedia Appendix 1).

Subgroup analyses were conducted to explore potential sources of heterogeneity. For each subgroup, pooled AUCs with corresponding 95% CIs, heterogeneity statistics (I^2, τ^2, τ, and Q statistic) are presented in Figures S3-S6 in Multimedia Appendix 1, with the 95% PI estimated using the confidence-distribution bootstrap method displayed. Prediction time window (≤1 y vs >1 y) did not significantly moderate discriminative performance (P=.95) and explained none of the observed heterogeneity. Similarly, no significant difference was observed between text-based models and sensor-based or multimodal models (P=.58); estimates for the latter subgroup were imprecise due to the small number of included studies.

Sample size (≤500 vs >500 participants) significantly moderated model discrimination. Meta-regression demonstrated a statistically significant subgroup effect (P=.02), accounting for approximately 42% of between-study heterogeneity, with higher AUCs observed in models developed from smaller samples, consistent with potential optimistic performance estimates. Population type (general vs. specific conditions) also emerged as a significant moderator (P<.001), explaining 87% of the observed heterogeneity, with higher discrimination in models developed for specific populations compared with general community-dwelling older adults; however, this finding should be interpreted cautiously given the limited number of studies in the specific population subgroup.

Discussion

Summary of Findings

Falls are a leading cause of injury, disability, and loss of independence among older adults [75-79]. Unlike institutionalized older adults under continuous professional care, community-dwelling older adults often lack intensive monitoring, making early identification of high-risk community-dwelling individuals critical for timely prevention [8]. Overall, our findings indicate that current ML-DL models show promising discriminative ability for predicting future falls. Nonetheless, their readiness for real-world implementation is limited by a lack of external validation, insufficient calibration assessment, and consistently high risk of bias. In this review, we systematically synthesized evidence from 28 studies evaluating ML-DL–based prediction models for future falls in community-dwelling older adults. These models incorporated a range of predictors, most commonly age, sex, fall history, depression, and BADL-ADL, reflecting the multifactorial nature of fall risk. Although the models generally showed acceptable discrimination, their performance varied considerably, and the available evidence was based largely on internal rather than external validation. Methodological quality assessment indicated that all included models were at high risk of bias, and meta-analyses revealed substantial between-study heterogeneity. Taken together, these findings highlight both the potential and current limitations of ML-DL–based approaches and provide an evidence base to inform the development of more robust and

XSL•FO
RenderX

clinically applicable fall risk prediction tools for proactive prevention in community settings.

Table S4 in Multimedia Appendix 1 contrasts the scope and methodological focus of key prior reviews with our systematic review. Prior reviews have emphasized fall risk management but mainly focused on cross-sectional risk classification, real-time fall detection systems, or experimental sensor-based technologies, rather than predictive models for primary prevention [18,35,37,80]. While detection systems are valuable for acute response, their capacity to inform proactive, upstream prevention strategies is limited [81,82].

Early risk identification through cost-effective prediction models is essential for implementing targeted prevention, addressing modifiable risk factors, and ultimately reducing fall incidence [83]. A recent review [36] primarily focusing on conventional statistical models, such as logistic regression and Cox regression, reported moderate to good discrimination (AUC 0.65 - 0.88) but highlighted their limited ability to capture complex interactions and integrate diverse data sources. ML approaches, in contrast, can accommodate high-dimensional data and capture nonlinear dependencies [31], making them better suited for multifactorial fall risk modeling. More recently, Xie et al [35] reviewed ML models for in-hospital fall prediction, offering insights into short-term risk management but providing limited guidance for broader, longer-term prevention in the community.

To our knowledge, no systematic review has comprehensively evaluated ML-based models for predicting future falls in older adults in community settings, where early identification and long-term prevention are most needed [18,35-37,80]. Importantly, longitudinal community-based fall risk prediction differs fundamentally from in-hospital screening or real-time detection tasks in several methodological respects. Such models rely on baseline predictors measured prior to outcome occurrence, require explicit definition of prediction horizons, and must appropriately account for censoring and loss to follow-up [84,85]. Moreover, because predicted risks are intended to guide preventive decision-making, adequate calibration across time horizons is essential, rather than discrimination alone [85,86]. Failure to recognize these distinctions risks conflating analytically and clinically distinct tasks. By restricting inclusion to longitudinal, community-based studies, this review addresses this gap and clarifies the unique methodological challenges and opportunities of ML-based approaches for proactive fall prevention.

The included studies showed substantial variation in geographic distribution and sample characteristics. Most were conducted in East Asia, North America, or Europe, with limited representation from low- and middle-income regions. This imbalance may limit generalizability, as fall risk profiles may differ substantially in underrepresented populations due to disparities in health care access, living environments, and socioeconomic factors [87]. Moreover, all included studies focused on single-ethnicity populations, further limiting our understanding of how these models perform across diverse racial or cultural groups. Sample sizes also varied considerably. Models developed from small datasets are particularly prone to overfitting [88], especially when lacking validation. Additionally, nearly half of the studies were single center, which may limit model robustness, while multicenter designs are better positioned to enhance generalizability. Importantly, although all included participants were classified as community-dwelling, recruitment sources were not uniform across studies. Some cohorts were sampled from the general community, whereas others were recruited through clinically enriched pathways. This distinction is important because community-dwelling status refers to living arrangement rather than recruitment context. For example, 1 of the 2 studies in the specific population subgroup recruited older adults being discharged home from the emergency department, representing community-living individuals with recent acute-care contact rather than a general community sample [29]. The other focused on community-dwelling older adults with cognitive frailty, a clinically defined high-risk subgroup within the broader community [56]. Consistent with this distinction, our subgroup analysis showed significantly higher discrimination in models developed for specific populations than in models developed for general community-dwelling older adults, with population subgroup explaining a large proportion of the observed heterogeneity. However, this finding should be interpreted cautiously because the specific population subgroup included only 2 studies, and these represented different forms of clinical enrichment. Nonetheless, it suggests that apparently stronger model performance may partly reflect narrower case mix and more concentrated risk profiles, rather than better transportability to the broader community-dwelling older population.

Nearly all studies predicted fall occurrence without distinguishing between first-time and recurrent events, with only one model targeting first falls [72]. This represents a critical gap, as the risk factors and prevention strategies for first falls differ from those for recurrent falls [89]. First-time falls often mark the transition into frailty and can precipitate fear of falling, reduced activity, and long-term disability. In contrast, recurrent fallers may already be known to their family or caregivers and are more likely to receive fall-related interventions [90,91]. Developing separate prediction models for these 2 groups could enhance clinical utility and enable more targeted preventive interventions. Accordingly, the first fall prediction should be treated as a distinct modeling objective with tailored populations, predictor sets, prediction horizons, and decision thresholds; related practical recommendations are provided in Box S1 in Multimedia Appendix 1.

Frequently selected predictors included age, sex, fall history, depression, and BADL. Age was the most common predictor, which is plausible given that with aging, progressive declines in neuromuscular control, sensory function, and balance increase the likelihood of instability [92]. Age is also associated with multimorbidity, polypharmacy, and sarcopenia, which further exacerbate vulnerability to falls and fall-related injuries [93,94]. Sex was another recurrent predictor, with women generally facing a higher risk, partly due to osteoporosis and postmenopausal bone loss [95], as well as differences in muscle mass and hormonal profiles [96,97].

Fall history consistently served as a strong predictor because intrinsic or extrinsic causes of previous falls often remain

XSL•FO
RenderX

unsolved and increase recurrence risk. Repeated falls may trigger fear of falling and activity avoidance, which in turn contribute to physical deconditioning and further functional decline [98]. Depression may raise fall risk both directly, through psychomotor slowing and fatigue, and indirectly, through physical inactivity and antidepressant use [99,100]. ADL-BADL reflects basic functional capacity; difficulties with tasks such as bathing or dressing often indicate underlying mobility or balance deficits and signal elevated fall risk [101,102]. Broader risk factors were also included, including motor functions, health status, lifestyle behaviors, and medication use. Prior research highlighted the importance of both motor and nonmotor factors in fall risk [103].

Our review included 23 ML-based and 5 DL-based models, reflecting the current predominance of ML in fall prediction research. Rather than demonstrating a clear superiority of one approach over the other, our findings suggest that ML and DL are each better suited to different data modalities. ML models were primarily developed using structured, text-based predictors, such as demographics, health status, medication use, lifestyle behaviors, and fall history, which are relatively inexpensive to collect and thus facilitate cost-effective fall prediction in community settings.

DL models were predominantly applied to sensor or image data, which provide high-dimensional, continuous signals capable of capturing subtle patterns not easily represented in structured variables. Although DL shows considerable potential in leveraging such complex data types, its application remains constrained by challenges in data acquisition, large sample size requirements, high computational resources, and limited interpretability [104,105]. Among the DL studies, the input data were primarily high-dimensional signals, including raw or minimally processed multivariate inertial sensor streams (eg, accelerometer or gyroscope gait time series) and imaging-derived matrices. Accordingly, the reported architectures spanned convolutional neural networks (CNNs) for local pattern extraction, recurrent units (eg, long short-term memory or gated recurrent unit) for sequential dependency modeling, and hybrid CNN-RNN designs (eg, hybrid-convolutional recurrent neural network) for end-to-end sequence learning; for imaging tasks, encoder-decoder CNNs such as U-Net support multiscale feature extraction with spatially explicit outputs. Figure S1 in Multimedia Appendix 1 suggests that DL is selectively applied to high-dimensional, nontext modalities, whereas classic ML predominates for structured data across all sample sizes. Although DL shows high median AUCs in some strata, these results are based on few studies, highlighting uncertainty in generalizability and reinforcing the need for cautious interpretation and external validation.

In our review, 9 models used sensor-based gait and balance parameters, including multimodal assessments. These models have the advantage of capturing dynamic abnormalities in movement patterns that are not easily detected through traditional clinical assessments. However, predictor definitions, measurements, and reports varied across studies, undermining the reproducibility and comparability of these models. Together, these findings underscore the need for more comprehensive and holistic modeling approaches to achieve effective prediction.

Across the 28 included studies, reported AUC values were generally acceptable but varied widely, indicating considerable variability in predictive performance across different models and study settings. However, only one model underwent external validation, but it could not be included in the quantitative synthesis due to insufficient information to derive SE. This limited use of external validation substantially constrains the ability to judge model performance in real-world settings and increases the likelihood that reported results are optimistic. All remaining models relied exclusively on internal validation approaches, such as cross-validation or random data splitting. As internal validation is known to yield optimistically biased performance estimates, these results primarily reflect within-sample discrimination and provide limited evidence regarding model performance in independent, real-world populations [106].

The meta-analysis of models with internal validation indicated apparently acceptable discrimination. However, the pooled AUC should not be interpreted as a robust estimate of real-world predictive performance. Rather, it represents a summary of reported discrimination from studies that were uniformly judged to be at high risk of bias and were based almost entirely on internal validation. In this context, an AUC of 0.79 is better understood as indicating that these models can separate higher-risk from lower-risk individuals under study-specific conditions, but that the magnitude of this performance is likely to be inflated relative to what would be expected in independent community populations. In the absence of external validation, the current evidence base primarily reflects methodological performance under controlled conditions and remains insufficient to support direct clinical implementation of ML-based fall risk prediction models.

Interpretation of this pooled estimate is further limited by extremely high between-study heterogeneity, suggesting substantial variation in study design, populations, prediction horizons, data modalities, and model inputs rather than random error alone. Accordingly, the pooled AUC should be viewed as an average across highly diverse, high-risk-of-bias studies rather than as a stable benchmark for model performance. Nevertheless, pooling remained informative by quantifying the central tendency of reported discrimination and, critically, by enabling estimation of PIs that explicitly convey the expected dispersion of model performance in new populations. Importantly, PIs provide information beyond pooled estimates by reflecting the range of model performance that may be expected in new settings, rather than only the average discrimination across studies [48,107]. This is particularly relevant for ML-based prediction models, where performance is often context-dependent and influenced by differences in populations, data sources, feature definitions, and model development strategies. In such settings, pooled AUC values alone may give an overly optimistic or simplified impression of performance, whereas PIs explicitly capture the uncertainty and variability that are likely to be encountered in real-world applications. Therefore, reporting PIs alongside pooled estimates is essential for a more realistic assessment of generalizability

XSL•FO
RenderX

and for informing decisions about model implementation in diverse community settings. Consistent pooled estimates obtained from leave-one-out sensitivity analyses and within-study averaged AUCs further suggest that no single study disproportionately influenced the overall results, but this stability should not be mistaken for low bias or strong transportability.

To systematically explore sources of heterogeneity, we conducted prespecified subgroup analyses and mixed-effects meta-regression. These analyses demonstrated that sample size and population subgroup significantly moderated discriminative performance, jointly explaining a substantial proportion of heterogeneity, whereas prediction time window and data modality did not materially affect pooled AUCs. Notably, PIs were wide across all subgroups, regardless of the estimation method, highlighting considerable uncertainty in real-world generalizability even when average discrimination appears acceptable.

Importantly, the absence of a statistically significant moderating effect of prediction time window should not be interpreted as clinical equivalence across different horizons. Models predicting short-term fall risk are conceptually aligned with transitional care and near-term preventive interventions, whereas models estimating multiyear fall risk may be more suitable for long-term prevention planning or population-level surveillance. These application scenarios differ fundamentally in terms of intervention timing, decision thresholds, update frequency, and calibration requirements. The wide variation in prediction horizons therefore reflects heterogeneity in clinical intent rather than merely methodological noise, underscoring that models developed for one time horizon should not be directly extrapolated to another without revalidation [84,85].

Beyond discrimination, additional methodological limitations further restrict clinical interpretability. All included models were rated as having a high risk of bias, reflecting common shortcomings, such as insufficient sample sizes, exclusion of missing data without clear justification, preprocessing pipelines, limited reporting of feature definitions, unclear definition of retained predictors, and reliance on self- or proxy-reported fall events. Calibration, which assesses agreement between predicted risks and observed outcomes, was rarely reported: only 5 studies provided Brier scores, of which only one reported a corresponding 95% CI, and only a single study presented a calibration plot, precluding quantitative synthesis of calibration performance. This lack of calibration assessment further limits the clinical interpretability of these models, as accurate risk estimation is essential for decision-making, threshold selection, and implementation in real-world care pathways. Consequently, most models were evaluated primarily using uncalibrated AUCs, limiting their utility for threshold-based risk stratification, shared decision-making, and intervention planning. For this reason, the pooled AUC should not be taken as evidence that these models are sufficiently reliable for deployment, but rather as a provisional summary of reported discrimination within a methodologically weak evidence base. Collectively, these methodological shortcomings compromise internal validity, hinder reproducibility, and substantially constrain the potential for real-world translation.

Unlike prior reviews [18,35-37,80], we comprehensively evaluated ML-based risk prediction models developed specifically to predict future falls among community-dwelling older adults from a preventive perspective. By including only longitudinal studies, the findings are directly relevant to real-world fall prediction rather than post hoc risk classification. This review focuses on longitudinal ML-DL models for predicting future falls in community-dwelling older adults, thereby extending beyond prior reviews centered on real-time detection, in-hospital fall prediction, or conventional statistical models. By integrating PROBAST appraisal with HKSJ meta-analysis, PIs, and modality-specific mapping of algorithm families, the review provides a structured and methodologically grounded synthesis to interpret both average discrimination and between-study variability. These findings help contextualize reported model performance and highlight key considerations for evaluating generalizability and uncertainty across different settings.

Limitations

Some limitations should be acknowledged. First, all included studies were at high risk of bias, and the predominance of models with internal validation limits confidence in generalizability. Only one model underwent external validation, and this study could not be included in the quantitative synthesis due to insufficient reporting. As a result, the pooled AUC should be interpreted as a summary of reported discrimination from a methodologically weak evidence base, rather than as a robust estimate of transportable performance in real-world settings. Second, substantial heterogeneity remained despite prespecified subgroup analyses and meta-regression, reflecting fundamental differences in study populations, designs, and modeling approaches. Some potentially relevant factors, such as algorithm type, validation strategy, outcome definition, and geographic region, could not be examined quantitatively due to data limitations. This heterogeneity further limits the extent to which any single pooled estimate can be viewed as a stable benchmark for expected model performance. In addition, the category of community dwelling was not equivalent to uniform population-based recruitment across studies. Some included cohorts were recruited through emergency care or clinically defined high-risk pathways rather than from the general community, and our subgroup analysis suggested that models developed in specific populations showed higher discrimination than those developed in general community-dwelling samples. However, because the specific population subgroup comprised only 2 studies with different types of clinical enrichment, this finding should be interpreted cautiously and cannot establish a stable or broadly generalizable subgroup effect. Third, calibration was infrequently reported, limiting assessment of clinical usability, as discrimination alone is insufficient for risk stratification and decision support. In addition, inconsistent reporting of performance metrics, limited transparency in feature selection and data preprocessing, and restricted data availability further constrain reproducibility and real-world applicability. Accordingly, even models with apparently acceptable discrimination may still provide poorly calibrated or overly optimistic risk estimates when applied outside their development settings. Fourth, several models were derived from the same

XSL•FO
RenderX

database, which may introduce partial correlation between effect estimates. Although some models were developed in distinct subpopulations or analytic cohorts, participant overlap cannot be fully excluded. However, sensitivity analyses showed that pooled AUC estimates were similar when retaining only one CHARLS-derived model or excluding all such models, suggesting minimal impact on the overall conclusions. Finally, the restriction to English-language studies and the exclusion of preprints may have introduced language and publication bias, which should be considered when interpreting the findings.

From a translation perspective, the wide PIs indicate that models with apparently acceptable average discrimination may perform substantially differently when applied to new community settings. This is particularly important given that the available pooled estimate is derived largely from internally validated, high-risk-of-bias studies and may therefore overstate real-world predictive performance. Accordingly, these tools should be considered as decision-support aids rather than stand-alone screening instruments, and implementation should be accompanied by local recalibration and context-specific threshold selection [85].

Future research should prioritize prospective, multicenter studies with rigorous external validation and standardized outcome definitions. Transparent reporting of both discrimination and calibration, as well as predictor handling and data processing, is essential; at a minimum, AUC should be reported alongside calibration measures [104,108]. In addition, impact studies, interventional trials, and silent trials are needed to determine whether these models improve patient outcomes and resource use. Public availability of feature dictionaries, preprocessing code, and calibration information will further support reproducibility and safe real-world translation. Until such evidence is available, pooled discrimination estimates should be viewed primarily as signals of potential utility rather than evidence of implementation readiness.

Conclusions

This systematic review and meta-analysis demonstrates growing interest in the use of ML-DL to predict future falls among older adults, but the current evidence should be interpreted cautiously. Although pooled discrimination appeared promising, substantial between-study heterogeneity, wide PIs, uniformly high risk of bias, limited external validation, and sparse calibration assessment indicate that the reported AUC is better understood as a summary of study-specific performance than as a robust estimate of transportable real-world accuracy. This caution is further warranted because “community dwelling” did not always correspond to recruitment from the general community, and some models were developed in clinically enriched community-living subgroups.

By focusing on longitudinal ML-DL models for predicting future falls in community-dwelling older adults, this review extends prior work centered on real-time detection, in-hospital prediction, or conventional statistical models. Through the integration of PROBAST appraisal, HKSJ meta-analysis, PIs, and modality-specific synthesis, it provides a decision-relevant framework for interpreting both discrimination and uncertainty across settings. At present, the evidence supports the potential utility of ML-DL models as decision-support tools for proactive fall prevention, but external validation, calibration assessment, local recalibration, and context-specific threshold selection remain essential before implementation.

Acknowledgments

We appreciate Dr. Ruihang Zhang from the Department of Urology, Shanghai Sixth People’s Hospital Affiliated to Shanghai Jiao Tong University School of Medicine for his support with statistical analysis techniques in this study. The authors declare the use of generative artificial intelligence (Gen-AI) in the research and writing process. According to the GAIDeT (Generative Artificial Intelligence Delegation Taxonomy, 2025), the following tasks were delegated to Gen-AI tools under full human supervision: proofreading and editing. The Gen-AI tool used was ChatGPT. Responsibility for the final manuscript lies entirely with the authors. Gen-AI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This study was supported by the 2024 Municipal Hospital Medical Treatment Technology Promotion and Optimization Management Project of Shanghai Shenkang Hospital Development Center (SHDC12024620), the Youth Development Program of Shanghai Ruijin Hospital (2024PY293), 2024 Shanghai Ruijin Hospital Nursing Research Fund (RJKH(Y)-2024-014), Shanghai Nursing Association Funding (2024MS-B13), the National Science and Technology Major Project of the Ministry of Science and Technology of China (2024ZD0536602), and Shanghai Hospital Association Nursing Funding (HLGL202509). The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

Data supporting the findings of this study are presented in the manuscript and multimedia appendix. The data extraction sheets underlying the meta-analysis are provided in Table S6 in Multimedia Appendix 1. Statistical analyses were conducted using established R packages, and the R scripts are available from the corresponding author upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

YG contributed to writing, visualization, methodology, formal analysis, data curation, and conceptualization. DX contributed to writing, methodology, formal analysis, and data curation. XL contributed to formal analysis and data curation. JW contributed to validation and formal analysis. LW contributed to manuscript revision and conceptualization. BW contributed to methodology and conceptualization. HZ contributed to methodology and manuscript revision. XQ contributed to manuscript revision, supervision, project administration, and funding acquisition. WZ contributed to manuscript revision, supervision, project administration, funding acquisition, and methodology.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional supporting materials for the systematic review and meta-analysis, including database search strategies, detailed model characteristics, and supplementary analyses.
[DOCX File, 3172 KB - jmir_v28i1e84844_app1.docx]

Checklist 1
PRISMA 2020 checklist.
[DOCX File, 166 KB - jmir_v28i1e84844_app2.docx]

References

1. Urbanek JK, Roth DL, Karas M, et al. Free-living gait cadence measured by wearable accelerometer: a promising alternative to traditional measures of mobility for assessing fall risk. J Gerontol A Biol Sci Med Sci 2023 May 11;78(5):802-810. [doi: 10.1093/gerona/glac013] [Medline: 35029661]
2. Chen Y, Dai F, Huang S, et al. Global, regional, and national burden of falls among older adults: findings from the Global Burden of Disease Study 2021 and Projections to 2040. NPJ Aging 2025 Oct 9;11(1):85. [doi: 10.1038/s41514-025-00275-4] [Medline: 41068164]
3. Salari N, Darvishi N, Ahmadipanah M, Shohaimi S, Mohammadi M. Global prevalence of falls in the older adults: a comprehensive systematic review and meta-analysis. J Orthop Surg Res 2022 Jun 28;17(1):334. [doi: 10.1186/s13018-022-03222-1] [Medline: 35765037]
4. Kakara R, Bergen G, Burns E, Stevens M. Nonfatal and fatal falls among adults aged ≥65 Years - United States, 2020-2021. MMWR Morb Mortal Wkly Rep 2023 Sep 1;72(35):938-943. [doi: 10.15585/mmwr.mm7235a1] [Medline: 37651272]
5. Older adult falls prevention. CDC. 2024. URL: https://www.cdc.gov/falls/data-research/index.html [accessed 2024-10-28]
6. Hashimoto S, Hosoi T, Yakabe M, et al. Preventative approaches to falls and frailty. Osteoporos Sarcopenia 2025 Jun;11(2 Suppl):50-53. [doi: 10.1016/j.afos.2025.04.002] [Medline: 40718351]
7. Fadem SZ. Consequences of falls. In: Understanding and Preventing Falls: A Guide to Reducing Your Risks, SZ: Springer International Publishing; 2023:61-69. [doi: 10.1007/978-3-031-39155-2_4]
8. Montero-Odasso M, van der Velde N, Martin FC, et al. World guidelines for falls prevention and management for older adults: a global initiative. Age Ageing 2022 Sep 2;51(9):afac205. [doi: 10.1093/ageing/afac205] [Medline: 36178003]
9. Kohler-Voinov LC, Sayyid ZN, Cullen KE. Multifactorial predictors of falls in older adults: a decade of data from the National Health and Aging Trends Study. BMC Geriatr 2025 Nov 25;25(1):950. [doi: 10.1186/s12877-025-06515-2] [Medline: 41291488]
10. Gambaro E, Gramaglia C, Azzolina D, Molin AD, Zeppegno P, Campani D. The complex associations between late life depression, fear of falling and risk of falls. a systematic review and meta-analysis. Ageing Res Rev 2022 Jan;73:101532. [doi: 10.1016/j.arr.2021.101532] [Medline: 34844015]
11. Harper A, Wilkinson I. Falls in older adults: causes, assessment and management. Medicine (Abingdon) 2024 Nov;52(11):677-682. [doi: 10.1016/j.mpmed.2024.08.010]
12. Yan S. Unmet community care needs and the well-being of older adults in China. Innov Aging 2024 Dec 31;8(Supplement_1):6-6. [doi: 10.1093/geroni/igae098.0017] [Medline: 38426023]
13. Podsiadlo D, Richardson S. The timed “Up & Go”: a test of basic functional mobility for frail elderly persons. J Am Geriatr Soc 1991 Feb;39(2):142-148. [doi: 10.1111/j.1532-5415.1991.tb01616.x] [Medline: 1991946]
14. Berg KO, Wood-Dauphinee SL, Williams JI, Maki B. Measuring balance in the elderly: validation of an instrument. Can J Public Health 1992;83 Suppl 2:S7-11. [Medline: 1468055]
15. Li Y, Liu P, Fang Y, et al. A decade of progress in wearable sensors for fall detection (2015–2024): a network-based visualization review. Sensors (Basel) 2025;25(7):2205. [doi: 10.3390/s25072205]
16. Hu Y, Zhang X, Slavin V, et al. Beyond comparing machine learning and logistic regression in clinical prediction modelling: shifting from model debate to data quality. J Med Internet Res 2025 Nov 5;27:e77721. [doi: 10.2196/77721] [Medline: 41191908]

XSL•FO
RenderX

17. Prakash C, Kumar R, Mittal N. Recent developments in human gait research: parameters, approaches, applications, machine learning techniques, datasets and challenges. Artif Intell Rev 2018 Jan;49(1):1-40. [doi: 10.1007/s10462-016-9514-6]
18. Usmani S, Saboor A, Haris M, Khan MA, Park H. Latest research trends in fall detection and prevention using machine learning: a systematic review. Sensors (Basel) 2021 Jul 29;21(15):5134. [doi: 10.3390/s21155134] [Medline: 34372371]
19. Riley RD, Collins GS. Stability of clinical prediction models developed using statistical or machine learning methods. Biom J 2023 Dec;65(8):e2200302. [doi: 10.1002/bimj.202200302] [Medline: 37466257]
20. Gambaro E, Gramaglia C, Azzolina D, Campani D, Molin AD, Zeppegno P. The complex associations between late life depression, fear of falling and risk of falls. A systematic review and meta-analysis. Ageing Res Rev 2022 Jan;73:101532. [doi: 10.1016/j.arr.2021.101532] [Medline: 34844015]
21. Midgley W, Mizen A, Bailey R, et al. How does the environment in and around the home affect social care and health outcomes for older people? A national longitudinal dynamic cohort study. Lancet 2023 Nov;402 Suppl 1:S69. [doi: 10.1016/S0140-6736(23)02096-2] [Medline: 37997113]
22. Ye C, Li J, Hao S, et al. Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm. Int J Med Inform 2020 May;137:104105. [doi: 10.1016/j.ijmedinf.2020.104105] [Medline: 32193089]
23. Xu Q, Ou X, Li J. The risk of falls among the aging population: a systematic review and meta-analysis. Front Public Health 2022;10:902599. [doi: 10.3389/fpubh.2022.902599] [Medline: 36324472]
24. Chen X, He L, Shi K, Wu Y, Lin S, Fang Y. Interpretable machine learning for fall prediction among older adults in China. Am J Prev Med 2023 Oct;65(4):579-586. [doi: 10.1016/j.amepre.2023.04.006] [Medline: 37087076]
25. Shao L, Shi Y, Xie XY, Wang ZA, Wang ZA, Zhang JE. Incidence and risk factors of falls among older people in nursing homes: systematic review and meta-analysis. J Am Med Dir Assoc 2023 Nov;24(11):1708-1717. [doi: 10.1016/j.jamda.2023.06.002] [Medline: 37433427]
26. Thapa R, Garikipati A, Shokouhi S, et al. Predicting falls in long-term care facilities: machine learning study. JMIR Aging 2022 Apr 1;5(2):e35373. [doi: 10.2196/35373] [Medline: 35363146]
27. Tang YT, Romero-Ortuno R. Using explainable AI (XAI) for the prediction of falls in the older population. Algorithms 2022;15(10):353. [doi: 10.3390/a15100353]
28. Chen X, Lin S, Zheng Y, He L, Fang Y. Long-term trajectories of depressive symptoms and machine learning techniques for fall prediction in older adults: evidence from the China health and retirement longitudinal study (CHARLS). Arch Gerontol Geriatr 2023 Aug;111:105012. [doi: 10.1016/j.archger.2023.105012] [Medline: 37030148]
29. Dasgupta P, Frisch A, Huber J, Sejdic E, Suffoletto B. Predicting falls within 3 months of emergency department discharge among community-dwelling older adults using self-report tools versus a brief functional assessment. Am J Emerg Med 2022 Mar;53:245-249. [doi: 10.1016/j.ajem.2021.12.071] [Medline: 35085878]
30. Mishra AK, Skubic M, Despins LA, et al. Explainable fall risk prediction in older adults using gait and geriatric assessments. Front Digit Health 2022;4:869812. [doi: 10.3389/fdgth.2022.869812] [Medline: 35601885]
31. Hawkins DM. The problem of overfitting. J Chem Inf Comput Sci 2004;44(1):1-12. [doi: 10.1021/ci0342472] [Medline: 14741005]
32. Gattani A, Dixit S, Patil M, et al. Artificial intelligence for fall detection in older adults: a comprehensive survey of machine learning, deep learning approaches, and future directions. Ageing Res Rev 2026 Jan;113:102948. [doi: 10.1016/j.arr.2025.102948] [Medline: 41265569]
33. Issaiy M, Zarei D, Kolahi S, Liebeskind DS. Machine learning and deep learning algorithms in stroke medicine: a systematic review of hemorrhagic transformation prediction models. J Neurol 2024 Dec 12;272(1):37. [doi: 10.1007/s00415-024-12810-6] [Medline: 39666168]
34. Trinh VQN, Zhang S, Kovoor J, et al. The use of natural language processing in detecting and predicting falls within the healthcare setting: a systematic review. Int J Qual Health Care 2023 Oct 17;35(4):mzad077. [doi: 10.1093/intqhc/mzad077] [Medline: 37758209]
35. Xie R, Shao L, Pei J, et al. Machine learning-based prediction models for falls in hospitalized patients: a systematic review and meta-analysis. Geriatr Nurs 2025;63:487-498. [doi: 10.1016/j.gerinurse.2025.03.059] [Medline: 40267662]
36. Dormosh N, van de Loo B, Heymans MW, et al. A systematic review of fall prediction models for community-dwelling older adults: comparison between models based on research cohorts and models based on routinely collected data. Age Ageing 2024 Jul 2;53(7):afae131. [doi: 10.1093/ageing/afae131] [Medline: 38979796]
37. Ren L, Peng Y. Research of fall detection and fall prevention technologies: a systematic review. IEEE Access 2019;7:77702-77722. [doi: 10.1109/ACCESS.2019.2922708] [Medline: 31328077]
38. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
39. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
40. McInnes MDF, Moher D, Thombs BD, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA 2018 Jan 23;319(4):388-396. [doi: 10.1001/jama.2017.19163] [Medline: 29362800]
41. The world's #1 systematic review tool. Covidence. URL: https://www.covidence.org/ [accessed 2026-03-24]

XSL•FO
RenderX

42. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019 Jan 1;170(1):51-58. [doi: 10.7326/M18-1376] [Medline: 30596875]
43. Viechtbauer W. Metafor: meta-analysis package for r. version 4.8-0. Comprehensive R Archive Network (CRAN). 2025. URL: https://CRAN.R-project.org/package=metafor [accessed 2026-04-30]
44. Schwarzer G. Meta: general package for meta-analysis. version 8.2-1. Comprehensive R Archive Network (CRAN). 2025. URL: https://CRAN.R-project.org/package=meta [accessed 2026-04-30]
45. Nagashima K, Noma H, Furukawa TA. Pimeta: prediction intervals for random-effects meta-analysis. version 1.1.3. Comprehensive R Archive Network (CRAN). 2015. URL: https://CRAN.R-project.org/package=pimeta [accessed 2026-04-30]
46. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 6;327(7414):557-560. [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
47. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023 Dec;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
48. IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open 2016 Jul 12;6(7):e010247. [doi: 10.1136/bmjopen-2015-010247] [Medline: 27406637]
49. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019 Jun;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]
50. Senn SJ. Overstating the evidence: double counting in meta-analysis and related problems. BMC Med Res Methodol 2009 Feb 13;9(1):10. [doi: 10.1186/1471-2288-9-10] [Medline: 19216779]
51. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14(1):25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
52. Lau J, Ioannidis JPA, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. BMJ 2006 Sep 16;333(7568):597-600. [doi: 10.1136/bmj.333.7568.597] [Medline: 16974018]
53. Wan R, Long D, Wang K, et al. Predicting fall risk among older adults with sarcopenia in China using machine learning models: a six-year longitudinal study from CHARLS. BMC Geriatr 2026 Jan 22;26(1):417. [doi: 10.1186/s12877-026-06977-y] [Medline: 41572187]
54. Lin L, Liu X, Cai C, Zheng Y, Li D, Hu G. Urban–rural disparities in fall risk among older Chinese adults: insights from machine learning-based predictive models. Front Public Health 2025;13:1597853. [doi: 10.3389/fpubh.2025.1597853]
55. Takeshita Y, Onishi M, Masuda H, et al. Machine learning prediction for postdischarge falls in older adults: a multicenter prospective study. J Am Med Dir Assoc 2025 Feb;26(2):105414. [doi: 10.1016/j.jamda.2024.105414] [Medline: 39701553]
56. Park C, Kim N, Kim M, Won CW, Lee BC. Advancing fall risk prediction in older adults with cognitive frailty: a machine learning approach using 2-year clinical data. PLoS ONE 2025;20(8):e0330672. [doi: 10.1371/journal.pone.0330672] [Medline: 40839571]
57. Liu D, Binkley NC, Perez A, et al. CT image-based biomarkers acquired by AI-based algorithms for the opportunistic prediction of falls. BJR Open 2023;5(1):20230014. [doi: 10.1259/bjro.20230014] [Medline: 37953870]
58. Silveira H, Lima J, Plácido J, et al. Dual-task performance, balance and aerobic capacity as predictors of falls in older adults with cardiovascular disease: a comparative study. Behav Sci (Basel) 2023 Jun 9;13(6):488. [doi: 10.3390/bs13060488] [Medline: 37366740]
59. Dormosh N, Schut MC, Heymans MW, et al. Predicting future falls in older people using natural language processing of general practitioners' clinical notes. Age Ageing 2023 Apr 1;52(4):afad046. [doi: 10.1093/ageing/afad046] [Medline: 37014000]
60. Ramsdale E, Kunduru M, Smith L, et al. Supervised learning applied to classifying fallers versus non-fallers among older adults with cancer. J Geriatr Oncol 2023 May;14(4):101498. [doi: 10.1016/j.jgo.2023.101498] [Medline: 37084629]
61. Ikeda T, Cooray U, Hariyama M, et al. An interpretable machine learning approach to predict fall risk among community-dwelling older adults: a three-year longitudinal study. J Gen Intern Med 2022 Aug;37(11):2727-2735. [doi: 10.1007/s11606-022-07394-8] [Medline: 35112279]
62. Kelly D, Condell J, Gillespie J, et al. Improved screening of fall risk using free-living based accelerometer data. J Biomed Inform 2022 Jul;131:104116. [doi: 10.1016/j.jbi.2022.104116] [Medline: 35690351]
63. Cuaya-Simbro G, Perez-Sanpablo AI, Morales EF, Quiñones Uriostegui I, Nuñez-Carrera L. Comparing machine learning methods to improve fall risk detection in elderly with osteoporosis from balance data. J Healthc Eng 2021;2021:8697805. [doi: 10.1155/2021/8697805] [Medline: 34540190]
64. Omae K, Kurita N, Takeshima T, et al. Significance of overactive bladder as a predictor of falls in community dwelling older adults: 1-year followup of the Sukagawa study. J Urol 2021 Jan;205(1):219-225. [doi: 10.1097/JU.0000000000001344] [Medline: 32856986]
65. Makino K, Lee S, Bae S, et al. Simplified decision-tree algorithm to predict falls for community-dwelling older adults. J Clin Med 2021 Nov 5;10(21):5184. [doi: 10.3390/jcm10215184] [Medline: 34768703]
66. Cuaya-Simbro G, Perez-Sanpablo AI, Muñoz-Meléndez A, Uriostegui IQ, Morales-Manzanares EF, Nuñez-Carrera L. Comparison of machine learning models to predict risk of falling in osteoporosis elderly. Found Comput Decis Sci 2020 Jun 1;45(2):66-77. [doi: 10.2478/fcds-2020-0005]

XSL•FO
RenderX

67. Cella A, De Luca A, Squeri V, et al. Development and validation of a robotic multifactorial fall-risk predictive model: a one-year prospective study in community-dwelling older adults. PLoS ONE 2020;15(6):e0234904. [doi: 10.1371/journal.pone.0234904] [Medline: 32584912]
68. Silva J, Sousa I, Cardoso JS. Fusion of clinical, self-reported, and multisensor data for predicting falls. IEEE J Biomed Health Inform 2020 Jan;24(1):50-56. [doi: 10.1109/JBHI.2019.2951230] [Medline: 31715574]
69. Kuspinar A, Hirdes JP, Berg K, McArthur C, Morris JN. Development and validation of an algorithm to assess risk of first-time falling among home care clients. BMC Geriatr 2019 Oct 14;19(1):264. [doi: 10.1186/s12877-019-1300-2] [Medline: 31610776]
70. Gillain S, Boutaayamou M, Schwartz C, et al. Using supervised learning machine algorithm to identify future fallers based on gait patterns: a two-year longitudinal study. Exp Gerontol 2019 Nov;127:110730. [doi: 10.1016/j.exger.2019.110730] [Medline: 31520696]
71. Howcroft J, Lemaire ED, Kofman J. Prospective elderly fall prediction by older-adult fall-risk modeling with feature selection. Biomed Signal Process Control 2018 May;43:320-328. [doi: 10.1016/j.bspc.2018.03.005]
72. Deschamps T, Le Goff CG, Berrut G, Cornu C, Mignardot JB. A decision model to predict the risk of the first fall onset. Exp Gerontol 2016 Aug;81:51-55. [doi: 10.1016/j.exger.2016.04.016] [Medline: 27114199]
73. Marschollek M, Rehwald A, Wolf KH, et al. Sensors vs. experts - a performance comparison of sensor-based fall risk assessment vs. conventional assessment in a sample of geriatric patients. BMC Med Inform Decis Mak 2011 Jun 28;11(1):48. [doi: 10.1186/1472-6947-11-48] [Medline: 21711504]
74. Bath PA, Pendleton N, Morgan K, Clague JE, Horan MA, Lucas SB. New approach to risk determination: development of risk profile for new falls among community-dwelling older people by use of a genetic algorithm neural network (GANN). J Gerontol A Biol Sci Med Sci 2000 Jan;55(1):M17-M21. [doi: 10.1093/gerona/55.1.m17] [Medline: 10719768]
75. Wielinski CL, Erickson-Davis C, Wichmann R, Walde-Douglas M, Parashos SA. Falls and injuries resulting from falls among patients with Parkinson's disease and other Parkinsonian syndromes. Mov Disord 2005 Apr;20(4):410-415. [doi: 10.1002/mds.20347] [Medline: 15580552]
76. Kendrick D, Kumar A, Carpenter H, et al. Exercise for reducing fear of falling in older people living in the community. Cochrane Database Syst Rev 2014 Nov 28;2014(11):CD009848. [doi: 10.1002/14651858.CD009848.pub2] [Medline: 25432016]
77. Scheffer AC, Schuurmans MJ, van Dijk N, van der Hooft T, de Rooij SE. Fear of falling: measurement strategy, prevalence, risk factors and consequences among older persons. Age Ageing 2008 Jan;37(1):19-24. [doi: 10.1093/ageing/afm169] [Medline: 18194967]
78. Scuffham P, Chaplin S, Legood R. Incidence and costs of unintentional falls in older people in the United Kingdom. J Epidemiol Community Health 2003 Sep;57(9):740-744. [doi: 10.1136/jech.57.9.740] [Medline: 12933783]
79. Reporting system (WISQARS). Centers for Disease Control and Prevention. 2016. URL: https://wisqars.cdc.gov/ [accessed 2026-04-30]
80. González-Castro A, Leirós-Rodríguez R, Prada-García C, Benítez-Andrades JA. The applications of artificial intelligence for assessing fall risk: systematic review. J Med Internet Res 2024 Apr 29;26:e54934. [doi: 10.2196/54934] [Medline: 38684088]
81. Subramaniam S, Faisal AI, Deen MJ. Wearable sensor systems for fall risk assessment: a review. Front Digit Health 2022;4:921506. [doi: 10.3389/fdgth.2022.921506] [Medline: 35911615]
82. Ishaq M, Guastella DC, Sutera G, Muscato G. A systematic review of fall detection and prediction technologies for older adults: an analysis of sensor modalities and computational models. Appl Sci (Basel) 2026;16(4):1929. [doi: 10.3390/app16041929]
83. Colón-Emeric CS, McDermott CL, Lee DS, Berry SD. Risk assessment and prevention of falls in older community-dwelling adults. JAMA 2024 Apr 23;331(16):1397-1406. [doi: 10.1001/jama.2024.1416] [Medline: 38536167]
84. Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G. Developing clinical prediction models: a step-by-step guide. BMJ 2024 Sep 3;386:e078276. [doi: 10.1136/bmj-2023-078276] [Medline: 39227063]
85. Collins GS, Dhiman P, Ma J, et al. Evaluation of clinical prediction models (part 1): from development to external validation. BMJ 2024 Jan 8;384:e074819. [doi: 10.1136/bmj-2023-074819] [Medline: 38191193]
86. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med 2019 Dec 16;17(1):230. [doi: 10.1186/s12916-019-1466-7] [Medline: 31842878]
87. McGrail K, Morgan J, Siddiqi A. Looking back and moving forward: addressing health inequities after COVID-19. Lancet Reg Health Am 2022 May;9:100232. [doi: 10.1016/j.lana.2022.100232] [Medline: 35313508]
88. Subramanian J, Simon R. Overfitting in prediction models - is it a problem only in high dimensions? Contemp Clin Trials 2013 Nov;36(2):636-641. [doi: 10.1016/j.cct.2013.06.011] [Medline: 23811117]
89. Jehu DA, Davis JC, Falck RS, et al. Risk factors for recurrent falls in older adults: a systematic review with meta-analysis. Maturitas 2021 Feb;144:23-28. [doi: 10.1016/j.maturitas.2020.10.021] [Medline: 33358204]
90. Vaishya R, Vaish A. Falls in older adults are serious. Indian J Orthop 2020 Feb;54(1):69-74. [doi: 10.1007/s43465-019-00037-x] [Medline: 32257019]

XSL•FO
RenderX

91. MacKay S, Ebert P, Harbidge C, Hogan DB. Fear of falling in older adults: a scoping review of recent literature. Can Geriatr J 2021 Dec;24(4):379-394. [doi: 10.5770/cgj.24.521] [Medline: 34912493]
92. Wen P, Sun L, Cheng L, Zhou C, Wang Z, Wang Z. Fall status and risk factors in older Chinese adults: a cross-sectional study. J Clin Nurs 2024 Jul;33(7):2609-2618. [doi: 10.1111/jocn.16965] [Medline: 38108248]
93. De Zoysa W, Mendis SA, Rathnayake N, et al. Multimorbidity, medications, and their association with falls, physical activity, and cognitive functions in older adults: multicenter study in Sri Lanka. Sci Rep 2025 Feb 20;15(1):6233. [doi: 10.1038/s41598-025-91037-5] [Medline: 39979482]
94. Prokopidis K, Giannos P, Reginster JY, et al. Sarcopenia is associated with a greater risk of polypharmacy and number of medications: a systematic review and meta-analysis. J Cachexia Sarcopenia Muscle 2023 Apr;14(2):671-683. [doi: 10.1002/jcsm.13190] [Medline: 36781175]
95. Alswat KA. Gender disparities in osteoporosis. J Clin Med Res 2017 May;9(5):382-387. [doi: 10.14740/jocmr2970w] [Medline: 28392857]
96. Okabe T, Suzuki M, Goto H, et al. Sex differences in age-related physical changes among community-dwelling adults. J Clin Med 2021 Oct 19;10(20):4800. [doi: 10.3390/jcm10204800] [Medline: 34682922]
97. Khorami F, Obaid N, Sparrey CJ. Sex differences in fall circumstances and injury biomechanics among older adults: a narrative review. Proc Inst Mech Eng H 2026 Jan;240(1):3-11. [doi: 10.1177/09544119251397581] [Medline: 41369245]
98. Step safely: strategies for preventing and managing falls across the life-course. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/978924002191-4 [accessed 2021-04-27]
99. Seppala LJ, van de Glind EMM, Daams JG, et al. Fall-risk-increasing drugs: a systematic review and meta-analysis: III. Others. J Am Med Dir Assoc 2018 Apr;19(4):372. [doi: 10.1016/j.jamda.2017.12.099] [Medline: 29402646]
100. van Poelgeest EP, Pronk AC, Rhebergen D, van der Velde N. Depression, antidepressants and fall risk: therapeutic dilemmas-a clinical review. Eur Geriatr Med 2021 Jun;12(3):585-596. [doi: 10.1007/s41999-021-00475-7] [Medline: 33721264]
101. Vermeulen J, Neyens JCL, van Rossum E, Spreeuwenberg MD, de Witte LP. Predicting ADL disability in community-dwelling elderly people using physical frailty indicators: a systematic review. BMC Geriatr 2011 Jul 1;11:33. [doi: 10.1186/1471-2318-11-33] [Medline: 21722355]
102. Ambrose AF, Paul G, Hausdorff JM. Risk factors for falls among older adults: a review of the literature. Maturitas 2013 May;75(1):51-61. [doi: 10.1016/j.maturitas.2013.02.009] [Medline: 23523272]
103. Gao Y, You H, Wang J, et al. Nonmotor symptom changes and their association with falls among Parkinson's disease patients undergoing deep brain stimulation: a 1-year cohort study. CNS Neurosci Ther 2025 Mar;31(3):e70310. [doi: 10.1111/cns.70310] [Medline: 40022502]
104. Lekadir K, Frangi AF, Porras AR, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ 2025 Feb 5;388:e081554. [doi: 10.1136/bmj-2024-081554] [Medline: 39909534]
105. Van Calster B, Collins GS, Vickers AJ, et al. Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance. Lancet Digit Health 2025 Dec;7(12):100916. [doi: 10.1016/j.landig.2025.100916] [Medline: 41391983]
106. Ho SY, Phua K, Wong L, Bin Goh WW. Extensions of the external validation for checking learned model interpretability and generalizability. Patterns (N Y) 2020 Nov 13;1(8):100129. [doi: 10.1016/j.patter.2020.100129] [Medline: 33294870]
107. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. BMJ 2011 Feb 10;342:d549. [doi: 10.1136/bmj.d549] [Medline: 21310794]
108. Debray TPA, Collins GS, Riley RD, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data (TRIPOD-Cluster): explanation and elaboration. BMJ 2023 Feb 7;380:e071058. [doi: 10.1136/bmj-2022-071058] [Medline: 36750236]

Abbreviations

ADL: activity of daily living
AUC: area under the curve
BADL: basic activity of daily living
CHARLS: China Health and Retirement Longitudinal Study
CNN: convolutional neural network
DL: deep learning
GAIDeT: Generative Artificial Intelligence Delegation Taxonomy
Gen-AI: generative artificial intelligence
HKSJ: Hartung-Knapp-Sidik-Jonkman
HTS: Higgins-Thompson-Spiegelhalter
MeSH: Medical Subject Headings
ML: machine learning
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

XSL•FO
RenderX

PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Statement for Diagnostic Test Accuracy
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
PROBAST: Prediction Model Risk of Bias Assessment Tool
RF: random forest

Edited by S Brini; submitted 26.Sep.2025; peer-reviewed by F Amakye, JE Zhang, S Jiang, W Wang; revised version received 05.Apr.2026; accepted 06.Apr.2026; published 14.May.2026.

Please cite as:
Gao Y, Xu D, Li X, Wang J, Wang L, Wu B, Zhao H, Qiu X, Zhu W
Machine Learning and Deep Learning Models for Predicting Future Falls in Community-Dwelling Older Adults: Systematic Review and Meta-Analysis of Longitudinal Evidence
J Med Internet Res 2026;28:e84844
URL: https://www.jmir.org/2026/1/e84844
doi:10.2196/84844

© Ying Gao, Doudou Xu, Xinru Li, Jue Wang, Linbin Wang, Beiwen Wu, Haifeng Zhao, Xian Qiu, Weiyi Zhu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Concerns of Using Large Language Models in Health Care Research and Practice: Umbrella Review

Feyza Yarar[1], MPH; Pauline Addis[1,2*], PhD; Megan Fairweather[1,2*], PhD; Dawn Craig[1,2], MSc; Hannah O'Keefe[1,2], MSci

[1]Population Health Sciences Institute, Faculty of Medical Sciences, Newcastle University, Framlington Place, Newcastle-Upon-Tyne, England, United Kingdom

[2]NIHR (National Institute of Health and Care Research) Innovation Observatory, Newcastle University, Newcastle-Upon-Tyne, England, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Hannah O'Keefe, MSci
Population Health Sciences Institute, Faculty of Medical Sciences, Newcastle University, Framlington Place, Newcastle-Upon-Tyne, England, United Kingdom

Abstract

Background: Large language models (LLMs), such as ChatGPT (OpenAI), are rapidly evolving, and their applications in health care are increasing. There is a growing demand for automation of routine tasks and a drive to use LLMs or similar to support research.

Objective: This umbrella review examines concerns of health care professionals and researchers related to the use of LLMs in health care research and practice. We aimed to identify common issues raised and the implications for patient care, policy, and practice.

Methods: A protocol was registered on PROSPERO (CRD420250640997). Searches were conducted in 7 databases (Ovid MEDLINE, Ovid Embase, Scopus, Web of Science, JBI Database of Systematic Reviews and Implementation Reports, Cochrane Database of Systematic Reviews, and Epistemonikos) in February 2025 and updated in February 2026. Screening was conducted in 2 stages, with independent screening by 2 reviewers. Studies published in the English language after January 2017 with at least one outcome expressing concerns of LLM or generative artificial intelligence use in health care research were included. The included studies were quality appraised for risk of bias and certainty of the evidence using AMSTAR-2 (A Measurement Tool to Assess Systematic Reviews) and GRADE (Grading of Recommendations Assessment, Development, and Evaluation), respectively. Data was extracted using a piloted form and narratively synthesized following SWiM guidelines and the PRIOR (Preferred Reporting Items for Overviews of Reviews) checklist.

Results: The search retrieved 448 systematic reviews, of which 42 met the inclusion criteria. Further, 12 distinct populations were identified, including researchers and clinicians in various medical specialties. The included reviews were assessed to be of very poor quality, and the level of overlap between primary studies could not be determined. Additionally, 15 reviews focused on ChatGPT, a further 15 on two or more LLMs, and 12 on generic artificial intelligence. Thus, 3 main themes emerged from the narrative synthesis. In order of most to least frequently discussed: (1) technical capability; (2) ethical, legal, and societal; and (3) costs.

Conclusions: To our knowledge, this is the first umbrella review to address the concerns of LLMs in health care research and practice. Thematic analyses provided insight into the complexity of different perspectives, and by using a whole population approach, it demonstrates common narratives. However, the poor quality of the included studies and potential overlap of results are substantial limitations. Data quality is at the heart of these concerns, and combative action must ensure health care professionals and researchers have the resources required to overcome these apprehensions. Ethical, legal, and societal implications of artificial intelligence use were also commonly raised. As technology accelerates and demands on health care increase, we must adapt and embrace change with equity, diversity, inclusion, and safety at the core.

Trial Registration: PROSPERO CRD420250640997; https://www.crd.york.ac.uk/PROSPERO/view/CRD420250640997

(J Med Internet Res 2026;28:e87804) doi:10.2196/87804

KEYWORDS

artificial intelligence; umbrella review; concerns; health and social care; life sciences

XSL•FO
RenderX

Introduction

Background

Currently, we live in an era where an abundance of data is being produced worldwide. While the term “big data” is generally used for predictive analytics, health care data can be considered “big data” by definition, as it is high in volume, velocity, variety, and veracity [1]. Big data is more suited to computational analysis, rather than traditional manual methods, and automating this analysis is an attractive proposition. The growing need to handle large datasets in the field of health care has led researchers to seek to leverage artificial intelligence (AI) as a means of automation [2]. The recent development of generative artificial intelligence (GenAI), particularly large language models (LLMs), has opened new frontiers in data handling [3]. In brief, GenAI uses advanced architectures, model context, and user prompts to recognize patterns in extensive data sets and generate original outputs. In the case of LLMs, this is done via transformer architectures, advanced neural networks designed to deliver next-token prediction [4].

LLMs are a versatile tool with the potential to transform health care research, but they also pose distinct challenges. As with other health care innovations, the risks and benefits of using LLMs should be weighed before implementation. Increasingly, there is a growing effort to develop strategies for the responsible use of AI. For example, leading journals do not accept papers with AI as an author, and NICE (National Institute of Health and Care Excellence) has guidelines on the use of AI in evidence generation [5,6]. Internationally, the European Union is expected to create the first AI law to be enforced in 2026, stratifying AI systems by risk level and regulating them accordingly [7]. Canada has also drafted legislation on AI [8]. Most laws focus on AI companies rather than individuals and have not yet taken effect. In addition to specific laws surrounding AI, it is crucial to comply with current unrelated but relevant laws, such as the General Data Protection Regulation (GDPR), in the interest of safety [9,10]. This will involve enhancing data security and clearly defining accountability [11,12].

The use of health care data is already bound by GDPR, and LLMs have been used in health care research applications ranging from the analysis of medical records and images to enhancing drug discovery and informing the formulation of new treatments [13-17]. Automated documentation could further assist clinical practice through use cases such as writing discharge summaries and personalized treatment or medication management plans [18,19]. This is particularly timely in the United Kingdom, where the government has pledged to embed AI throughout the National Health Service to support routine administration tasks [20]. There has also been a push toward automation methodologies in health and care research to deliver timely insights. This is particularly true in the field of evidence synthesis. However, the conversation and movement toward automation in this field have been ongoing for 2 decades with little to no progress [21-23].

As the speed of technological improvements accelerates, this can often outpace our ability to understand, assess, and mitigate concerns regarding AI [24]. Such concerns include reliability, accuracy, transparency, various ethical, security, and privacy concerns, as well as environmental concerns [25-27]. Such issues may be intrinsic to the LLM, reflecting its technical capabilities; extrinsic to the LLM, often relating to how it is used; or they may fall under both categories. When considering health care data, protection, security, and accuracy are paramount. As such, it is crucial to understand the views of individuals working in health care practice and research surrounding the use of LLMs. While research has been conducted to understand different population views, there has been no effort to cross-reference and triangulate these views. This is imperative to understand the landscape as a whole and promote multidisciplinary combative action. Thus, this umbrella review examines the concerns of health care professionals and researchers to identify areas for improvement and understand the implications for practice. It is anticipated that through the robust identification of issues, steps can be taken to mitigate concerns, instill confidence in users of AI, and that the use of AI will become more responsible.

Aims and Objectives

We aimed to map the concerns associated with the use of LLMs in health care research and practice through the following objectives: (1) identify systematic reviews that report concerns of health care professionals and health care researchers, and (2) perform qualitative analysis of the findings using inductive and deductive thematic analysis.

Methods

Study Design

Following a scoping search that confirmed the feasibility of this study, a protocol for the systematic review was developed and registered with PROSPERO (CRD420250640997) on February 24, 2025. No amendments were made to the information provided in the protocol. The umbrella review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines and the synthesis without meta-analysis guidelines, and reported using the PRIOR (Preferred Reporting Items for Overviews of Reviews) checklist, PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses - Search), and PRISMA 2020 Abstract checklist (Checklists 1-3) [28-31].

Eligibility Criteria

The SPIDER (Sample, Phenomenon of Interest, Design, Evaluation, Research Type) framework was used to outline the inclusion criteria as follows:

- Sample: health care professionals and researchers.
- Phenomenon of interest: LLMs or GenAI.
- Design: systematic reviews.
- Evaluation measures: reporting of concerns.
- Research type: qualitative research.

We included systematic reviews published since January 2017 (the year before the first LLMs being introduced publicly to ensure robust coverage of dates), where an outcome of the review was concerns surrounding the use of GenAIs or LLMs in a health and social care research context. We considered

XSL•FO
RenderX

systematic reviews as defined by the authors, provided the methodology followed a recognizable process (ie, searching, screening, data extraction, risk of bias, and synthesis). Preprints and nonpeer-reviewed papers were excluded (Textbox 1).

Textbox 1. Inclusion and exclusion criteria for the umbrella review applied to the retrieved search results.

Inclusion criteria:

- Published since January 2017
- Generative artificial intelligence (GenAI) or large language models (LLMs) used in a health care practice and research context
- Outcome for concerns surrounding the use of GenAIs or LLMs
- Systematic reviews

Exclusion criteria:

- Published before January 2017
- Research context not related to health care practice and research
- GenAIs or LLMs concerns not listed as an outcome
- Primary studies
- Letters
- Editorials reviews
- Conference abstracts
- Commentaries

Information Sources

Searches were completed on February 26, 2025, and updated on February 25, 2026, in seven databases: (1) Ovid MEDLINE (R) and Epub Ahead of Print, In-Process, In-Data-Review and Other Non-Indexed Citations, daily and versions; (2) Ovid Embase; (3) Scopus; (4) Web of Science; (5) JBI Database of Systematic Reviews and Implementation Reports; (6) Cochrane Database of Systematic Reviews; and (7) Epistemonikos.

Study registry searches, purposeful searching of gray literature sources, and citation chaining were neither performed nor were authors independently contacted for further data.

Search Strategy

A de novo search strategy was developed in MEDLINE (Ovid) using the phenomenon of interest and evaluation measures concepts of the SPIDER framework: (ethic* or concern* or raises questions or equality or equity or racial or discriminat* or EDI or (equity diversity and inclusion) or adversely or perpetuat* or persist* or bolster or pitfall* or controvers* or worry or barrier or impede or obstacle or limitation or hindrance or hurdle) AND ((LLM or large language model or GenAI or generative AI or ChatGPT or OpenAI or gpt or Gemini or DeepSeek or LlaMA or Falcon or Cohere or PaLM or Claude v1 or autoregressive language or encoder-decoder or decoder or transformer or prompt engineer) AND (research or academ*). The strategy was peer reviewed by an information specialist and translated into other databases as appropriate (Multimedia Appendix 1). A date limit of 2017 onward was applied to all searches, and results were limited to systematic reviews using built-in functions in each database. No language restrictions were applied.

Selection Process

The records from the databases were imported into Rayyan, a free online screening platform, and duplicates were removed [32]. The remaining systematic reviews were initially screened by title and abstract in duplicate (FY and PA or MF). Discrepancies were resolved by discussion, and HO provided a final judgment when a consensus could not be reached. Systematic reviews taken forward for full-text screening were independently screened by 2 reviewers (FY and PA or MF). Again, discrepancies were resolved by discussion.

Data Collection Process

The data extraction template was initially trialed on 4 systematic reviews selected at random. Further, 2 reviewers independently extracted data from 50% of the included studies (FY and PA or MF). Discrepancies were resolved via discussion or consultation with HO. The remaining studies were extracted by 1 reviewer (FY) with discussion when required.

Data Items

The following data were extracted: first author, year, title, DOI (Digital Object Identifier), journal, country of the first author, health care or research field, the LLM assessed, number of included studies, included study designs, inclusion criteria, exclusion criteria, key concerns, population raising the concern, notable quotes, statistical methods, and declared limitations.

Risk of Bias Assessment

AMSTAR-2 (A Measurement Tool to Assess Systematic Reviews) was used to assess risk of bias, including reporting bias. The 16 questions outlined in AMSTAR-2 were applied independently by 2 reviewers to each included study, following the guidelines for each question or domain included [33,34].

XSL•FO
RenderX

Synthesis Methods

The narrative synthesis was conducted by a single reviewer (FY), using a thematic analysis approach with inductive and deductive coding [35]. No statistical synthesis or meta-analysis was conducted, and quantitative effect measures were neither appropriate nor available given the qualitative nature of concerns. Concerns from the data extraction tables were analyzed and inductively synthesized into codes. Each systematic review was deductively analyzed to determine whether these codes were present in the text. The codes were organized and synthesized into main themes. The main themes were subsequently organized into overarching themes. An analysis by population was also conducted, which recorded the number of systematic reviews for a particular population that highlighted each of the coded concerns.

Certainty Assessment

The application of GRADE (Grading of Recommendations Assessment, Development, and Evaluation) to systematic reviews of qualitative research gives a measure of how well the findings reflect the phenomenon of interest and provides an indicator of certainty around the evidence. GRADE was applied to assess the depth and breadth of this study, providing an initial rating, downgrading domains (risk of bias, inconsistency, indirectness, imprecision, and publication bias), upgrading domains (large effect or dose response, if confounders would reduce the effect), and an end certainty rating. Assessment was conducted independently by 2 reviewers. Heterogeneity and sensitivity analysis were not assessed within this review.

Results

Systematic Review Selection

In total, 449 records were identified from databases, as shown in the PRISMA flowchart (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA flow diagram detailing numbers for study retrieval, screening, exclusion, and inclusion. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

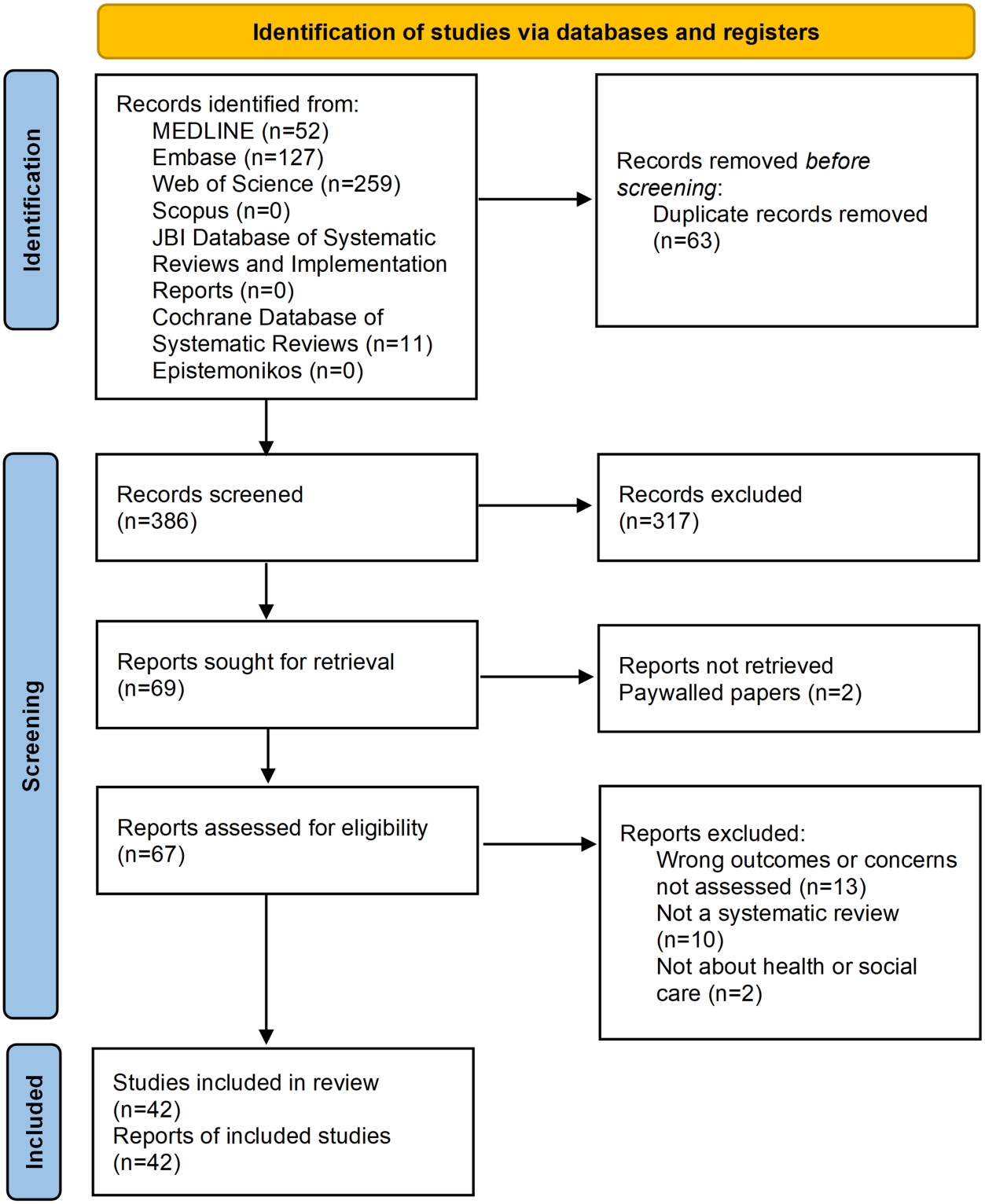

Before screening, 63 duplicates were removed. The remaining 386 systematic reviews were initially screened by title and abstract. A total of 317 reviews that did not meet the inclusion criteria were excluded. The remaining 69 systematic reviews underwent an additional full-text screening. Further, 27 reports were excluded, most commonly due to concerns not being addressed, followed by topics being unrelated to health or social care, and not being a systematic review or being inaccessible behind a paywall (Figure 1 and Multimedia Appendix 2). A final total of 42 remaining systematic reviews were included in this umbrella review.

XSL•FO
RenderX

Characteristics of Systematic Reviews

All studies identified were written in the English language. However, the country of the first author varied across the globe. Most systematic reviews originated from the United States (n=9) [10,36-43], followed by the United Kingdom (n=6) [44-49], Pakistan (n=4) [50-53], Australia (n=3) [12,54,55], Canada and Israel (n=2 each) [56-59], and 16 other countries with 1 systematic review each [9,11,60-73]. The most common year of publication was 2024, with 27 reviews, followed by 2023 with 9 reviews and 2025 with 6 reviews. No reviews were found before 2023 (Table 1).

The population raising concerns was as follows. Most commonly, it was researchers only (n=17) and other groups included clinicians (general; n=5), plastic surgeons (n=5), psychiatrists (n=4), and neurosurgeons (n=3). Less common were pediatricians, gastroenterologists, dermatologists, pathologists, cardiologists, ophthalmologists, orthopedics, and ICU nurses (n=1 each; Table 1).

The most common individual LLM was ChatGPT, with 15 systematic reviews focusing on this LLM alone [9,36,39,42,44,48,50,51,59-61,68,70-72]. Another 15 systematic reviews considered more than one LLM [10,12,37,38,40,45,46,49,54,56-58,62,65,67]. A smaller proportion of studies (n=12) examined LLMs within a broader context, such as AI in general [11,43,47,53,69,73], natural language processing [41,66], conversational agents [55,63], deep learning [52], and GenAI [64] (Table 1).

There was a positively skewed distribution of primary studies included in the systematic reviews. The IQR was between 19 and 83 (median 32; range 5 - 315) studies. Further, 3 systematic reviews were classed as outliers as they examined higher numbers of primary studies, and a further 3 reviews did not share their sample size [39,58,65]. All systematic reviews were qualitative and did not perform a meta-analysis. Only 5.1% (n=2) formally tested for agreement between reviewers using Cohen κ (Multimedia Appendix 3 for full data extraction).

XSL•FO
RenderX

Table . Characteristics of the 39 included studies in this umbrella review, including AMSTAR-2[a] and GRADE[b] ratings for each included study.

Reviews	Country	AI[c] of interest	Population raising concerns	Included studies (n)	AMSTAR-2 rating	GRADE rating
Abi-Rafeh et al [44], 2024	United Kingdom	ChatGPT (OpenAI)	Plastic surgeons	175	Critically low	Very low
Arif et al [50], 2024	Pakistan	ChatGPT	Plastic surgeons	32	Critically low	Very low
Balla et al [45], 2023	United Kingdom	AI in general - lists ChatGPT, Bard (Google LLC), and GLASS A.I. 2.0 (Glass Health)	Pediatricians	20	Critically low	Very low
Banskota et al [73], 2025	Nepal	AI (general)	Orthopedics	20	Critically low	Very low
Bečulić et al [60], 2024	Bosnia	ChatGPT	Neurosurgeons	13	Critically low	Very low
Fareed et al [53], 2025	Pakistan	LLMs[d]	Clinicians	27	Critically low	Very low
Fatima et al [51], 2024	Pakistan	ChatGPT	Clinicians	83	Critically low	Low
Garg et al [61], 2023	India	ChatGPT	Clinicians	118	Critically low	Very low
Guo et al [46], 2024	UK	LLMs: ChatGPT, BERT[e]	Researchers	40	Critically low	Low
Haltaufderheide and Ranisch [62], 2024	Germany	LLMs, ChatGPT	Researchers	53	Critically low	Low
Kiuchi et al [63], 2024	Japan	CAs[f]	Researchers	315	Critically low	Very low
Kiwan et al [47], 2024	United Kingdom	AI in general	Plastic surgeons	96	Critically low	Very low
Klang et al [36], 2023	United States	ChatGPT 3.5 (OpenAI)	Gastroenterologists	6	Critically low	Low
Kolding et al [64], 2024	Denmark	GenAI - includes ChatGPT	Psychiatrists	40	Critically low	Very low
Kucukkaya et al [9], 2024	Turkey	ChatGPT	ICU[g] nurses	5	Critically low	Very low
Kutbi [67], 2024	Saudi Arabia	AI (general), LLMs	Researchers	19	Critically low	Very low
Li and Guenier [48], 2024	United Kingdom	ChatGPT 3.5 (4) (OpenAI)	Researchers	14	Critically low	Very low
Malgaroli et al [41], 2023	United States	NLP[h]	Psychiatrists	102	Critically low	Low
Mohamed et al [65], 2024	Oman	LLMs	Researchers	N/R[i]	Critically low	Low
Moya-Salazar et al [71], 2024	Peru	ChatGPT	Researchers	14	Critically low	Very low
Nasra et al [54], 2025	Australia	AI (general), LLMs	Clinicians	22	Critically low	Very low
Omar et al [56], 2024	Israel	LLMs	Psychiatrists	16	Low	Low
Omar et al [57], 2025	Israel	LLMs	Psychiatrists	34	Low	Low
Paganelli et al [66], 2024	Italy	NLP	Dermatologists	30	Critically low	Very low

XSL•FO
RenderX

Reviews	Country	AI[c] of interest	Population raising concerns	Included studies (n)	AMSTAR-2 rating	GRADE rating
Pashangpour and Nejat [58], 2024	Canada	LLMs	Researchers	N/R	Critically low	Very low
Patil et al [37], 2024	United States	LLMs	Neurosurgeons	51	Critically low	Very low
Pressman et al [10], 2024	United States	LLMs	Plastic surgeons	53	Critically low	Very low
Pressman et al [38], 2024	United States	LLMs	Plastic surgeons	34	Critically low	Very low
Rehman et al [52], 2025	Pakistan	Deep learning (including LLMs)	Researchers	100	Critically low	Very low
Roman et al [68], 2023	United Arab Emirates	ChatGPT	Neurosurgeons	22	Critically low	Very low
Rudnicka et al [69], 2024	Poland	AI (general)	Researchers	253	Critically low	Very low
Ruksakulpiwat et al [70], 2023	Thailand	ChatGPT	Researchers	6	Critically low	Very low
Sacoransky et al [59], 2024	Canada	ChatGPT	Researchers	8	Critically low	Very low
Sallam [72], 2023	Jordan	ChatGPT	Pathologists	60	Critically low	Very low
Sanjeewa et al [55], 2024	Australia	CAs	Researchers	19	Critically low	Low
Sharma et al [42], 2024	United States	ChatGPT	Cardiologists	24	Critically low	Very low
Tangsrivimol et al [39], 2025	United States	ChatGPT	Clinicians	N/R	Critically low	Very low
Villanueva-Miranda et al [43], 2025	United States	Deep learning	Researchers	83	Critically low	Very low
Wang et al [40], 2024	United States	LLMs - ChatGPT	Researchers	65	Critically low	Very low
Wangsa et al [12], 2024	Australia	ChatGPT, Bard, Llama (Meta AI), Ernie (Baidu), and Grok (xAI)	Researchers	28	Critically low	Very low
Wong et al [49], 2024	United Kingdom	LLMs	Ophthalmologists	32	Critically low	Very low
Younis et al [11], 2024	Iraq	AI (general)	Researchers	82	Critically low	Very low

[a]AMSTAR-2: A Measurement Tool to Assess Systematic Reviews.

[b]GRADE: Grading of Recommendations Assessment, Development, and Evaluation.

[c]AI: artificial intelligence.

[d]LLM: large language model.

[e]BERT: Bidirectional Encoder Representations from Transformers.

[f]CA: conversational agent.

[g]ICU: intensive care unit.

[h]NLP: natural language processing.

[i]N/R: not reported.

Primary Study Overlap

Due to poor reporting of included primary studies within the systematic reviews, we were unable to calculate the corrected covered area to identify any overlap between primary studies. For example, some reviews listed the number of included studies but only referenced a subset in among supporting references, which meant the primary studies could not be easily identified. However, from manual inspection, it appears that primary studies that have been clearly reported were only included in a single systematic review. In either case, findings should be interpreted with caution as overinflation may be present.

XSL•FO
RenderX

Quality Assessments

Regarding risk of bias in systematic reviews, the overall rating of reviews using AMSTAR-2 was either low (n=2, 4.8% reviews) [56,57] or critically low (n=40, 95.2% reviews) [9-11,36-55,58-73] (Table 1). A total of 40 (95.2%) reviews had a valid research question (Q1) [9-12,36-41,43-65,67-73], 24 (57.1%) reviews provided descriptions of their included studies (Q3) [9,11,36,38,40-42,44,46,50,51,54-57,60-62,64-66,68,70,72], and 27 (64.2%) reviews reported no conflicts of interest (Q16) [9-12,36-44,46-60,62,63,67-73].

However, only 6 (14.2%) reviews had a protocol (Q2) [41,46,56,57,62,68], 3 (7.1%) reviews justified their choice of included studies (Q8) [41,56,65], 10 (23.8%) reviews had a literature search strategy (Q4) [41,48,55-57,61,62,64,66,68], 1 (38%) review was completely double-screened (Q5) [10,36,40,41,46,48,49,55-57,61,64,68,70,71,73], 8 (19%) reviews were completely double data-extracted (Q6) [36,53,54,56,57,61,73], 1 (2.3%) review justified their exclusions (Q7) [46], 6 (14.2%) reviews had risk of bias assessment (Q9) [36,46,55-57,62], 6 (14.2%) reviews looked for funding disclosures of constituent studies (Q10) [9,36,43,49,53,73], 6 (14.2%) reviews accounted for risk of bias when interpreting results (Q13) [36,41,55-57,65], and 5 (11.9%) reviews addressed heterogeneity (Q14) [36,41,55,57,67]. No reviews completed a meta-analysis (Q11), a risk of bias assessment for a meta-analysis (Q12), or formally assessed for publication bias (Q15; Table 2).

XSL•FO
RenderX

Table . AMSTAR-2[a] rating for each of the 39 included systematic reviews.

Reviews	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Overall rating
Abi-Rafeh et al [44], 2024	Y[b]	N[c]	N	N	N	N	N	P[d]	N	N	N/A[e]	N/A	N	N	N	Y	Critically low
Arif et al [50], 2024	Y	N	N	N	N	N	N	P	N	N	N/A	N/A	N	N	N	Y	Critically low
Balla et al [45], 2023	Y	N	N	N	N	N	N	N	N	N	N/A	N/A	N	N	N	N	Critically low
Banskota et al [73], 2025	Y	P	N	N	Y	Y	N	P	P	Y	N/A	N/A	N	N	N	Y	Critically low
Bečulić et al [60], 2024	Y	N	N	N	N	N	N	P	N	N	N/A	N/A	N	N	N	Y	Critically low
Fareed et al [53], 2025	Y	N	N	P	N	Y	N	N	N	Y	N/A	N/A	N	N	N	Y	Critically low
Fatima et al [51], 2024	Y	N	N	N	N	N	N	P	N	N	N/A	N/A	N	N	N	Y	Critically low
Garg et al [61], 2023	Y	N	N	P	Y	Y	N	P	N	N	N/A	N/A	N	N	N	N	Critically low
Guo et al [46], 2024	Y	P	N	N	Y	N	N	P	Y	N	N/A	N/A	N	N	N	Y	Critically low
Haltaufderheide and Ranisch [62], 2024	Y	P	N	P	N	N	N	P	P	N	N/A	N/A	N	N	N	Y	Critically low
Kiuchi et al [63], 2024	Y	N	N	N	N	N	N	N	N	N	N/A	N/A	N	N	N	Y	Critically low
Kiwan et al [47], 2024	Y	N	N	N	N	N	N	N	N	N	N/A	N/A	N	N	N	Y	Critically low

XSL•FO
RenderX

Reviews	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Overall rating
Klang et al [36], 2023	Y	N	N	N	Y	Y	N	P	Y	Y	N/A	N/A	Y	Y	N	Y	Critically low
Kolding et al [64], 2024	Y	N	N	P	Y	N	N	P	N	N	N/A	N/A	N	N	N	N	Critically low
Kucukkaya et al [9], 2024	Y	N	N	N	N	N	N	P	N	Y	N/A	N/A	N	N	N	Y	Critically low
Kutbi [67], 2024	Y	N	N	N	N	N	N	N	N	N	N/A	N/A	N	Y	N	Y	Critically low
Li and Guenier [48], 2024	Y	N	N	P	Y	N	N	N	N	N	N/A	N/A	N	N	N	Y	Critically low
Malgaroli et al [41], 2023	Y	P	Y	P	Y	N	N	Y	N	N	N/A	N/A	Y	Y	N	Y	Critically low
Mohamed et al [65], 2024	Y	N	Y	N	N	N	N	P	N	N	N/A	N/A	Y	N	N	N	Critically low
Moya-Salazar et al [71], 2024	Y	N	N	N	Y	N	N	N	N	N	N/A	N/A	N	N	N	Y	Critically low
Nasra et al [54], 2025	Y	N	N	N	N	N	N	P	N	N	N/A	N/A	N	N	N	Y	Critically low
Omar et al [56], 2024	Y	P	Y	P	Y	Y	N	P	Y	N	N/A	N/A	Y	Y	N	Y	Low
Omar et al [57], 2025	Y	P	N	P	Y	Y	N	P	P	N	N/A	N/A	Y	N	N	Y	Low
Paganelli et al [66], 2024	Y	N	N	P	N	Y	N	P	N	N	N/A	N/A	N	N	N	N	Critically low

XSL•FO
RenderX

Reviews	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Overall rating
Pashangpour and Nejat [58], 2024	N	N	N	N	N	N	N	N	N	N	N/A	N/A	N	N	N	Y	Critically low
Patil et al [37], 2024	Y	N	N	N	N	N	N	N	N	N	N/A	N/A	N	N	N	Y	Critically low
Pressman et al [10], 2024	Y	N	N	N	Y	N	N	N	N	N	N/A	N/A	N	N	N	Y	Critically low
Pressman et al [38], 2024	Y	N	N	N	N	N	N	P	N	N	N/A	N/A	N	N	N	Y	Critically low
Rehman et al [52], 2025	Y	N	N	N	N	N	N	N	N	N	N/A	N/A	N	N	N	Y	Critically low
Roman et al [68], 2023	Y	P	N	P	Y	N	N	P	N	Y	N/A	N/A	N	N	N	Y	Critically low
Rudnicka et al [69], 2024	Y	N	N	N	N	N	N	N	N	N	N/A	N/A	N	N	N	Y	Critically low
Ruksakulpiwat et al [70], 2023	Y	N	N	N	Y	N	N	P	N	N	N/A	N/A	N	N	N	Y	Critically low
Sacoransky et al [59], 2024	Y	N	N	N	N	N	N	N	N	N	N/A	N/A	N	N	N	Y	Critically low
Sallam [72], 2023	Y	N	N	N	N	N	N	P	N	N	N/A	N/A	N	N	N	Y	Critically low
Sanjeewa et al [55], 2024	Y	N	N	P	Y	N	N	P	Y	N	N/A	N/A	Y	Y	N	Y	Critically low

XSL•FO
RenderX

Reviews	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Overall rating
Sharma et al [42], 2024	Y	N	N	N	N	N	N	P	N	N	N/A	N/A	N	N	N	Y	Critically low
Tangsivimol et al [39], 2025	N	N	N	N	N	N	N	N	N	N	N/A	N/A	N	N	N	Y	Critically low
Villanueva-Miranda et al [43], 2025	Y	N	N	P	N	N	N	P	N	Y	N/A	N/A	N	N	N	Y	Critically low
Wang et al [40], 2024	Y	N	N	N	Y	N	N	P	N	N	N/A	N/A	N	N	N	Y	Critically low
Wangsa et al [12], 2024	Y	N	N	N	N	N	N	N	N	N	N/A	N/A	N	N	N	Y	Critically low
Wong et al [49], 2024	Y	N	N	N	Y	N	N	N	N	N	N/A	N/A	N	N	N	Y	Critically low
Younis et al [11], 2024	Y	N	N	N	N	N	N	P	N	N	N/A	N/A	N	N	N	Y	Critically low

[a]AMSTAR-2: A Measurement Tool to Assess Systematic Reviews.

[b]Y: yes.

[c]N: no.

[d]P: partial.

[e]N/A: not applicable.

Certainty of Evidence

Reviews were overall graded as either low-end certainty (n=9, 21.4%) [36,41,46,51,55-57,62,65], or very low-end certainty (n=33, 78.6%) [9-12,37-40,42-44,47-50,52-54,58-61,63,66-73]. As there were no randomized controlled trials, all reviews started as low certainty before downgrading or upgrading (initial rating).

In the downgrading domains, all studies had consistent results (inconsistency) [9-12,36-73] and most addressed the core question (indirectness) of this umbrella review (n=36, 85.7%) [9-12,36-40,42,44-46,48-51,53-62,64-68,70-73]. However, only 7 (16.6%) reviews assessed for risk of bias [36,46,52,55-57,73], 14 (33.3%) reviews directly addressed the research question (imprecision) [10,12,36,40,43,46,48,51,53,56,57,62,65,73], and 4 (9.5%) reviews considered publication bias [41,51,62,65].

In terms of upgrading domains, 8 (19%) reviews were considered as showing large effects [10,36,46,51,56,57,62,65], but only 1 (2.3%) review mentioned confounders [41]. None displayed a dose-response relationship, and as this criterion was not relevant to this review, it was not taken into account for the final scoring (Table 3).

XSL•FO
RenderX

Table . GRADE[a] assessment for each of the 39 included systematic reviews.

				Downgrading domains					Upgrading domains			
Reviews	Depth of concerns	Breadth of concerns	Initial rating	RoB[b]	Inconsistency	Indirectness	Imprecision	Publication bias	Large effect	Dose response	Confounders would reduce the effect	End certainty rating
Abi-Rafeh et al [44], 2024	Y[c]	Y	Low	N[d]	Y	Y	N	N	N	N	N	Very low
Arif et al [50], 2024	N	Y	Low	N	Y	Y	N	N	N	N	N	Very low
Balla et al [45], 2023	Y	N	Low	N	Y	Y	N	N	N	N	N	Very low
Banskota et al [73], 2025	Y	Y	Low	Y	Y	Y	Y	N	N	N	N	Very low
Bečulić et al [60], 2024	N	N	Low	N	Y	Y	N	N	N	N	N	Very low
Fareed et al [53], 2025	Y	N	Low	N	Y	Y	Y	N	N	N	N	Very low
Fatima et al [51], 2024	Y	Y	Low	N	Y	Y	Y	Y	Y	N	N	Low
Garg et al [61], 2023	N	Y	Low	N	Y	Y	N	N	N	N	N	Very low
Guo et al [46], 2024	Y	Y	Low	Y	Y	Y	Y	N	Y	N	N	Low
Haltaufderheide and Ranisch [62], 2024	Y	Y	Low	N	Y	Y	Y	Y	Y	N	N	Low
Kiuchi et al [63], 2024	Y	Y	Low	N	Y	N	N	N	N	N	N	Very low
Kiwan et al [47], 2024	N	N	Low	N	Y	N	N	N	N	N	N	Very low
Klang et al [36], 2023	Y	N	Low	Y	Y	Y	Y	N	Y	N	N	Low
Kolding et al [64], 2024	N	Y	Low	N	Y	Y	N	N	N	N	N	Very low
Kucukkaya et al [9], 2024	Y	Y	Low	N	Y	Y	N	N	N	N	N	Very low

XSL•FO
RenderX

				Downgrading domains					Upgrading domains			
Reviews	Depth of concerns	Breadth of concerns	Initial rating	RoB[b]	Inconsistency	Indirectness	Imprecision	Publication bias	Large effect	Dose response	Confounders would reduce the effect	End certainty rating
Kutbi [67], 2024	Y	Y	Low	N	Y	Y	N	N	N	N	N	Very low
Li and Guenier [48], 2024	Y	Y	Low	N	Y	Y	Y	N	N	N	N	Very low
Malgaroli et al [41], 2023	Y	Y	Low	N	Y	N	N	Y	Y	N	Y	Low
Mohamed et al [65], 2024	Y	Y	Low	N	Y	Y	Y	Y	Y	N	N	Low
Moya-Salazar et al [71], 2024	N	N	Low	N	Y	Y	N	N	N	N	N	Very low
Nasra et al [54], 2025	N	N	Low	N	Y	Y	N	N	N	N	N	Very low
Omar et al [56], 2024	Y	N	Low	Y	Y	Y	Y	N	Y	N	N	Low
Omar et al [57], 2025	N	N	Low	Y	Y	Y	Y	N	Y	N	N	Low
Paganelli et al [66], 2024	N	N	Low	N	Y	Y	N	N	N	N	N	Very low
Pashangpour and Nejat [58], 2024	Y	Y	Low	N	Y	Y	N	N	N	N	N	Very low
Patil et al [37], 2024	Y	Y	Low	N	Y	Y	N	N	N	N	N	Very low
Pressman et al [10], 2024	Y	Y	Low	N	Y	Y	Y	N	Y	N	N	Very low
Pressman et al [38], 2024	N	Y	Low	N	Y	Y	N	N	N	N	N	Very low
Rehman et al [52], 2025	N	N	Low	Y	Y	N	N	N	N	N	N	Very low
Roman et al [68], 2023	N	N	Low	N	Y	Y	N	N	N	N	N	Very low

XSL•FO
RenderX

Reviews	Depth of concerns	Breadth of concerns	Initial rating	Downgrading domains					Upgrading domains			End certainty rating
				RoB[b]	Inconsistency	Indirectness	Imprecision	Publication bias	Large effect	Dose response	Confounders would reduce the effect	
Rudnicka et al [69], 2024	Y	N	Low	N	Y	N	N	N	N	N	N	Very low
Ruksakulpiwat et al [70], 2023	N	Y	Low	N	Y	Y	N	N	N	N	N	Very low
Sacoransky et al [59], 2024	N	N	Low	N	Y	Y	N	N	N	N	N	Very low
Sallam [72], 2023	Y	Y	Low	N	Y	Y	N	N	N	N	N	Very low
Sanjeewa et al [55], 2024	N	N	Low	Y	Y	Y	N	N	N	N	N	Low
Sharma et al [42], 2024	N	N	Low	N	Y	Y	N	N	N	N	N	Very low
Tangsrivimol et al [39], 2025	Y	Y	Low	N	Y	Y	N	N	N	N	N	Very low
Villanueva-Miranda et al [43], 2025	Y	N	Low	N	Y	N	Y	N	N	N	N	Very low
Wang et al [40], 2024	Y	Y	Low	N	Y	Y	Y	N	N	N	N	Very low
Wangsa et al [12], 2024	Y	Y	Low	N	Y	Y	Y	N	N	N	N	Very low
Wong et al [49], 2024	N	Y	Low	N	Y	Y	N	N	N	N	N	Very low
Younis et al [11], 2024	Y	Y	Low	N	Y	Y	N	N	N	N	N	Very low

[a]GRADE: Grading of Recommendations Assessment, Development, and Evaluation.

[b]RoB: Risk of Bias.

[c]Y: Yes.

[d]N: No.

Interrater Reliability

κ statistics were calculated between the reviewers (FY and one of PA, MF, or HO). Significant agreement was seen for both AMSTAR-2 (0.92, indicating near-perfect agreement) and GRADE (0.75, indicating substantial agreement). The κ statistic was considered significant where it was 0.6 (substantial agreement) or higher (Table 4) [74].

XSL•FO
RenderX

Table . Interrater reliability, Cohen κ, detailing the agreement rates between reviewers when assessing risk of bias (AMSTAR-2[a]) and certainty of the evidence (GRADE[b]).

Reviewer combination	AMSTAR-2	GRADE
FY and PA	0.88[c]	0.75[d]
FY and MF	0.88[c]	0.5
FY and HO	1.00[c]	1.00[c]
Average	0.92[c] (SD 0.05)	0.75[d] (SD 0.2)

[a]AMSTAR-2: A Measurement Tool to Assess Systematic Reviews.

[b]GRADE: Grading of Recommendations Assessment, Development, and Evaluation.

[c]Shows statistical significance of near-perfect agreement between reviewers.

[d]Shows statistical significance of substantial agreement.

Synthesis of Results

Qualitative coding is a means of deriving descriptive tags to categorize data, which can then be used to generate themes. A total of 29 codes were generated from the synthesis (Multimedia Appendix 3), and seven themes emerged: (1) data quality and reliability; (2) transparency and reproducibility; (3) performance and capability; (4) technical and operational; (5) human interaction and social impact; (6) ethical, legal, and safety; and (7) costs. These could be grouped under 3 core themes: technical capability; ethical, legal, and societal; and costs. For most population groups, mentions of technical capability concerns were the greatest, followed by ethical, legal, and societal concerns, and then cost concerns (Table 5).

XSL•FO
RenderX

Table . Count of systematic reviews raising concerns by population groups under the themes derived from qualitative thematic analysis.

	Cardio[a] (1)	Derma[b] (1)	Gastro[c] (1)	GC[d] (5)	ICU[e] nurses (1)	Neuro[f] (3)	Ophtha[g] (1)	Ortho[h] (1)	Pedia[i] (1)	Patho[j] (1)	Psurg[k] (5)	Psychia[l] (4)	Resea[m] (17)
Technical capabilities													
Data quality and reliability	1	1	1	5	0	3	1	1	1	1	5	4	15
Transparency and reproducibility	1	0	0	3	0	3	0	1	1	1	3	2	8
Performance and capability	1	1	1	5	0	2	1	1	1	1	4	4	12
Technical and operations	0	0	0	1	0	1	0	1	0	0	0	1	6
Ethical, legal, and societal													
Human interaction and social impact	0	1	0	5	1	1	0	0	0	1	5	2	12
Legal, ethical, and safety	1	1	1	5	1	3	1	1	1	1	5	4	16
Costs													
All costs	0	0	0	1	0	1	0	1	0	0	0	0	4

[a]Cardio: cardiologists.
[b]Derma: dermatologists.
[c]Gastro: gastroenterologists.
[d]GC: general clinicians.
[e]ICU: intensive care unit.
[f]Neuro: neurosurgeons.
[g]Ophtha: ophthalmologists.
[h]Ortho: orthopedics.
[i]Pedia: pediatricians.
[j]Patho: pathologists.
[k]Psurg: plastic surgeons.
[l]Psychia: psychiatrists.
[m]Resea: researchers.

Data Quality and Reliability

The 42 studies reflected that bias in training algorithms and datasets mirrors what is already known on the subject, with estimates that around a quarter of studies on LLMs show bias [40]. For sensitive topics, this may be reduced to around 15% [65]. Key demographic barriers were mentioned, including sex, race, culture, language, and religion [43,55,61]. Political biases were also highlighted in closed-source algorithms [12]. Specifically, racial biases affecting individuals of black ethnicity were mentioned [62]. Sexual discrimination, bias toward female doctors was highlighted, with AI recommending fewer female doctors than male doctors. In pediatric medicine, a key issue was a limited, fragmented, or total lack of standardized training

XSL•FO
RenderX

sets for LLMs in genetic disorders [45]. Most genetic disorders are rare and disproportionately affect children, which may contribute to the incomplete training sets [75].

Furthermore, outdated and limited datasets were commonly mentioned, with the date cutoffs highlighted for various ChatGPT models. For example, ChatGPT 3.5 was pretrained until September 2021 only and did not incorporate information from the internet [11,76]. ChatGPT 4 was pretrained up until April 2023 only [46,59,73]. Obvious problems with this include a lack of recent knowledge, obsolete knowledge, or misalignment with current clinical guidelines [50,53]. Another study also acknowledged that hospital-specific protocols should be adhered to, which LLMs may not include in their outputs [9]. From clinical practice, trust guidelines may differ from general ones, particularly in the case of antimicrobials, as location affects the presence of different microbes [77].

Fabricated or fake references were highlighted in 4 of the reviews [40,50,51,54]. Constituent studies that examined image references were anatomically incorrect or fabricated in 81% of cases [9]. Model overfitting, where a model is trained too specifically so that it cannot generalize on unseen data, was explored in reviews that examine image-based medical applications. Radiologically, x-ray images may be prone to overfitting [52]. Dermatology, which relies heavily on pattern recognition, may also be prone to this phenomenon [66]. Disease management options often featured fabricated references in all versions of ChatGPT [59]. Furthermore, the potential for such references to mislead junior colleagues, including resident doctors, was mentioned [59]. Similarly, several studies mentioned that hallucinations were a cause of concern [39,43,51,53,58,64]. Some studies emphasized that it was easy to be convinced by hallucinations [51,64]. Other aspects addressed included that hallucinations were often confidently given by LLMs, especially when parameter settings that encourage more varied, random outputs are used [46]. Other factors that influenced hallucination generation were the quality of prompts, with prompts phrased as stories often causing hallucinations [58]. It was estimated in 1 review that 40% of discharge summaries written by AI have hallucinations [39].

Transparency and Reproducibility

Several studies highlight the difficulties with the transparency of LLMs, with some referring to the technology as a “black box” [37,43,51,56,69,73]. Ethical discussions are highlighted as a way to build transparency of models [10]. Profit-driven lack of transparency was only mentioned in 1 review, but another highlighted how the proprietary nature of LLMs can complicate openness and trust [53,62]. Another review emphasized that a lack of transparency can be ascribed to multiple stages, including the design and training of LLMs [65]. The same review also emphasized that care should be taken so that interventions aimed at increasing transparency do not inadvertently expose patient data. Leaked personal data can potentially also amplify bias through normalizing biased correlations as real patterns [65].

The issue of repeatability with the same or similar prompts was also isolated as a concern. Most constituent primary research papers were identified as prompt experiments [64]. However, Patil et al [37] reported that about a quarter of the research papers were found not to disclose their prompts. Another review underlined that when they were disclosed, prompts tended to be single, standalone 1-shot prompts [59]. The concept of biased prompts was raised, and prompt injection, where prompts are engineered to extract information or cause disruption for nefarious purposes, was also highlighted as a concern [36,43].

Performance and Capability

The reviews in this study broadly agreed that the performance of LLMs was comparable to or exceeded that of humans. Latest estimates of correct responses on neurology licensing examination questions revealed ChatGPT 4 had 85% accuracy compared to human performance at 73.8% [46]. Fracture detection rates with deep learning were also close to this figure at 83% [67]. Approximately 76.5% (32/42) of studies used human performance as a benchmark, and 50% (21/42) compared LLMs to human performance alone. It is essential to measure success on diverse tasks, and some reviews agreed that LLMs often struggle with complexity [11,39,45,69,72]. Lack of originality was also a commonly cited concern [11,36,44,51,61,70,72]. Most systematic reviews emphasized that LLMs needed human oversight, and several mentioned it should only be viewed as a “supplementary tool” [9,39,45,50,56,57,60].

Technical and Operational Challenges

Few studies explored technical and operational challenges. Further, 1 review cited a primary study that examined system crashes [48]. Another review highlighted slow response times and how this was a problem in robotics [58]. Similarly, a second review noted slower clinical workflows due to the verification process [53]. Aside from the speed of development outstripping research, it was noted that a lack of research was problematic [64]. Another review quoted that only 26% of studies used randomized controlled trial–type designs in assessing LLMs by users, suggesting that conclusions from research may not be completely reliable [55]. Additionally, 1 review called for rigorous validation in real-world settings, which was supported by another review raising gaps in validation as a concern [43,73].

Human Interaction and Social Impact

There was a mix of views regarding empathy and LLMs. Some reviews stressed that LLMs’ empathy was good, although it was usually assessed through technical expert reports rather than by both users and experts. Only a minority of studies (n=10) used an explicit definition of the term “empathy” [46,54,55]. Others stated there was a general lack of empathy [11,39,50,64]. Only 1 study gave a mix of such opinions [47].

Furthermore, some reviews expressed that LLMs could cause potential damage to the doctor-patient relationship [38,44,46,62]. There was a large overlap between this finding and reviews that addressed deskilling of the workforce or impacts on the job market [11,44,46,62,65]. Additionally, a lack of a sustainable relationship between new technologies and the user was emphasized in robotics due to problems with “semantics, consistency, and interactiveness” [63]. Researchers raised

XSL•FO
RenderX

concerns around challenges with linguistic complexity, such as understanding irony and sarcasm [43].

User acceptance was identified as an important factor in 2 reviews [46,67]. The importance of acceptance by various groups, including physicians, caregivers, and providers, was highlighted [63]. Potential mistrust of LLMs and humanization issues were highlighted in the mental health field [43,55]. Yet, public acceptance was viewed as very positive in 1 systematic review [40]. There was an acknowledgment that this area needs more research [48].

Further, 5 reviews emphasized that existing inequalities could become more entrenched with LLMs [40,47,62,65,70]. This is separate from concerns over bias, which could also contribute to the deepening of inequalities. Both of these events could perpetuate the other, whereby biased outputs could deepen inequalities, which in turn could lead to the introduction of further biases.

Legal, Ethical, and Safety Concerns

Almost all reviews mentioned ethical concerns as a potential problem with LLM use, with varying degrees of explanation on this topic. Accountability was addressed from a 2-fold perspective, with 4 reviews focusing on medicolegal accountability [11,44,59,72]. Others were focused on legitimacy and accountability in research [50,51]. Suggested solutions included clear guidelines on accountability [12,48,55]. Another potential solution was policy and regulations [40,65]. Related to this was the issue of academic integrity, alternatively phrased as “pedagogical risk” [9,11,12,44,62,65,68].

Most reviews raised privacy and security concerns regarding LLMs. The privacy problems highlighted by studies can be seen as a triad involving information leaks from embedded training examples, inferential disclosure, and insufficiently deidentified data [40]. There may be a trade-off between data utility and privacy [40,65]. The need to comply with existing laws, including GDPR, was emphasized by 3 systematic reviews [9,12,59]. Other studies acknowledged that further efforts were necessary, such as limiting personal data collection or conducting audits [39,40,55,73].

Only 9 reviews discussed obtaining consent to collect personal data [10,12,37,39,43,44,53,62,73]. This reflects the limited literature available on ensuring informed consent when using LLMs and highlights the fact that new technology is outpacing safety. Just 1 review sought to explain consent itself, which entails providing a full disclosure of the risks and benefits in such a way that the participant comprehends and agrees [10]. Further, 2 reviews suggested that protocols are needed to obtain consent [12,39]. Consent is crucial in anticipation of seen and unforeseen ethical issues with LLMs, and yet other systematic reviews chose to discuss LLMs as a way to streamline consent forms [37,44,62].

Safety may become a population-wide problem as well as an individual one. Mass-scale problems may become apparent with the potential of infodemics perpetuated by LLMs, according to some reviews [51,53,62,65,72]. The saturation of scientific literature with low-quality automated reviews, which may fuel infodemics, was also discussed [62]. In a pandemic context such as COVID-19, there could be even greater consequences [65].

Finally, the risk of self-propagating and uncontrolled evolution was described as “unknown” by 1 review in passing [44]. Self-propagating and uncontrolled evolution relates to the LLM’s ability to grow and change on its own. Considering that this could be very detrimental and have long-term impacts, this is important. While this possibility has been informally mentioned in technology circles for some time, there are only a small number of new reports in the literature examining this possibility, so this finding is expected.

Environmental, Processing, and Economic Costs

None of the reviews examined more than one cost aspect, and most lacked depth of answers in relation to this topic. Costs were broadly environmental related to computational processing [43,58,65]. Given the scale of climate change and its importance to health, this is an area of great interest and should be explored further [78,79]. Other costs mentioned were financial costs incurred by users [40,60].

Discussion

Principal Findings

This umbrella review aimed to narratively synthesize the concerns that health care professionals and researchers face when using AI. A wide variety of concerns are raised, which overlap and interlink, consistently affecting multiple populations. The findings indicate 3 core areas of shared concern within the health care field: technical capability of AI; ethical, legal, and societal implications for use; and associated costs.

Much of the concern surrounding technical capability lies with data quality and reproducibility. For research to be as robust as possible, we must use AI for tasks that are appropriate, judicious, and sense-checked, as well as monitoring the functions and effects of LLMs [38,39]. A good starting point is to acknowledge where, how, and when AI was used. Research integrity policies, reporting guidelines, and audits will be crucial in meeting quality standards and enabling reproducibility [9,37,40,65,73]. Most of the reviews included here advocated for guidelines, although none explored direct examples or attempted to construct a guideline. This is expected, given that international and national guidelines are only gradually emerging. However, it could be argued that research should be a proactive driver for exploring these issues rather than a reactive one.

Ethical, legal, and societal implications were varied and broad-ranging. Concerns over hallucinations, bias, inequalities, and consent provide interesting and often deeply interrelated perspectives that triangulate with the technical capabilities of AI. For instance, bias that perpetuates “hallucinations” was commonly cited [61,62,72]. Efforts for transparency may inadvertently perpetuate bias, as implicit biases in models can exist even when they show no explicit biases [80]. This gives a false sense of security surrounding systems that are lacking in objectivity. The reuse of biased datasets from such models will worsen this problem, and further research is needed to explore these perpetuating cycles. Ultimately, transparency must

XSL•FO
RenderX

be judged by humans who may reinforce biases inadvertently or fail to acknowledge problems because features are unreported or untested. Therefore, while we should strive for more transparency, particularly through legislation, guidance, and reporting, we should avoid labeling any LLM as completely "transparent."

Some biases were thought to be pervasive, for instance, "social biases entrenched in data." Specifically, female ophthalmologists were recommended less than one-third as often as their male counterparts [47]. Biases were present in topical issues; for example, "unfavorable attitudes" were described when ChatGPT was prompted to discuss topics such as climate change and Black Lives Matter [65]. Research is uncovering both implicit and overt biases toward already disadvantaged populations, which is why careful consideration of LLMs and their applications is necessary to prevent exacerbating existing health inequalities. Education surrounding considerations of equality, diversity, and inclusion when using LLMs may be helpful on an individual level. However, discriminatory biases may only be overcome by careful curation of training data underpinning the LLMs.

Some terms regarding LLMs could be revised for clarity and to destigmatize. Findings indicate the term "hallucination" for AI issues is considered unhelpful and stigmatizing for those with a psychiatric disorder, reflecting proposals in the literature for the term "AI misinformation" [64,81]. The term "hallucination" was originally used in computing science to refer to retained outputs even when artificial neural networks were pruned by removing some connections [82]. The term has evolved, first to positively describe tasks related to computer vision and improved facial recognition, and then to mean the generation of incorrect outputs in translation or object detection. Currently, it is used to mean incorrect LLM outputs produced with confidence [83]. As an overall term to describe confidently produced errors, AI "hallucination" may, on the surface, be useful. However, LLMs are not produced the same way as biological hallucination, which occurs in the absence of external stimuli [84]. By contrast, LLMs have external stimuli in the form of training data and prompts but still produce nonsensical outputs. This can be stigmatizing for those with a psychiatric disorder and is technically imprecise. AI generalization, fact fabrication, or stochastic parroting could be used as more distinct terms depending on the types of error seen [83].

Interestingly, while cost was a core theme, there was no definitive cost element that was unifying across reviews. Further explorations of cost concerns could inform cost-benefit analyses and full economic evaluations of AI use cases. We have seen such evaluations for AI-assisted health care technologies, but not regarding AI use in general health care practice and research [85].

Limitations of This Review

The constituent systematic reviews comprised primary research papers that were heterogeneous. For example, different papers used different methods of evaluation in 1 review, ranging from surveys to response ratings and interview feedback [55]. Moreover, the authors used different methods to classify the accuracy of LLMs and did not adhere to standard formal procedures for assessment [40]. The quality of the included reviews was generally poor, and the extent of publication bias was unknown. Many of the poorly reported reviews included studies, meaning we were unable to determine the overlap between primary studies using the corrected covered area. Caution should be taken that different populations' views of LLM limitations are not a complete representation of views of the broader population, and that findings may be overinflated due to the unknown overlap of primary studies.

Furthermore, thematic analyses have an element of subjectivity. Potential sources of bias include researcher bias and confirmation bias, whereby preexisting beliefs and experiences may have influenced the coding. We have attempted to limit this through group discussion of codes; however, future studies could incorporate a blinded dual coding process.

This review is also limited by the search dates of the included systematic reviews. As the field of LLMs rapidly evolves, primary studies published after the search dates may provide valuable insights into current thinking.

Conclusions

To our knowledge, this is the first umbrella review to address the concerns of LLMs in health care research and practice. Thematic analyses provided insight into the complexity of different perspectives, and by using a whole population approach, it demonstrates common narratives. However, the poor quality of the included studies is a substantial limitation, and results should be interpreted with caution. Data quality is at the heart of these concerns, and combative action must ensure health care professionals and researchers have the resources required to overcome these apprehensions if AI is to be used routinely. Ethical, legal, and societal implications of AI use were also commonly raised. As technology accelerates and demands on health care increase, we must adapt and embrace change with equity, diversity, inclusion, and safety at the core.

Acknowledgments

We give thanks to Dr Chris Marshall and Dr Malcolm Moffat for providing support and guidance. The authors declare that no artificial intelligence tools were used in the writing of this paper, nor for any element of the work, including the processing of data or creation of images or tables.

Funding

The authors are funded by the NIHR (National Institute for Health and Care Research; HSRIC-2016-10009/Innovation Observatory). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

XSL•FO
RenderX

Data Availability

All data analyzed during this study are available as supplemental materials.

Authors' Contributions

HO and DC conceptualized this study's design. FY, PA, and MF conducted this review, with support from HO and DC. The first draft of this paper was written by FY and refined by HO. All authors commented on previous versions of this paper. All authors read and approved this final paper.

Conflicts of Interest

DC is Director of the NIHR (National Institute for Health and Care Research) Innovation Observatory. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Search strategies.
[DOCX File, 17 KB - jmir_v28i1e87804_app1.docx]

Multimedia Appendix 2
Excluded studies.
[DOCX File, 59 KB - jmir_v28i1e87804_app2.docx]

Multimedia Appendix 3
Data extraction and thematic coding.
[XLSX File, 66 KB - jmir_v28i1e87804_app3.xlsx]

Checklist 1
PRIOR checklist.
[DOCX File, 22 KB - jmir_v28i1e87804_app4.docx]

Checklist 2
PRISMA-S checklist.
[DOCX File, 17 KB - jmir_v28i1e87804_app5.docx]

Checklist 3
PRISMA 2020 abstract checklist.
[DOCX File, 17 KB - jmir_v28i1e87804_app6.docx]

References

1. Batko K, Ślęzak A. The use of big data analytics in healthcare. J Big Data 2022;9(1):3. [doi: 10.1186/s40537-021-00553-4] [Medline: 35013701]
2. Rathore MM, Shah SA, Shukla D, Bentafat E, Bakiras S. The role of ai, machine learning, and big data in digital twinning: a systematic literature review, challenges, and opportunities. IEEE Access 2021;9:32030-32052. [doi: 10.1109/ACCESS.2021.3060863]
3. Goel A, Gueta A, Gilon O, et al. LLMs accelerate annotation for medical information extraction. arXiv. Preprint posted online on Dec 4, 2023. [doi: 10.48550/arXiv.2312.02296]
4. Ferraris AF, Audrito D, Caro LD, Poncibò C. The architecture of language: understanding the mechanics behind LLMs. Camb Forum AI Law Gov 2025;1:e11. [doi: 10.1017/cfl.2024.16]
5. Hufton AL. No artificial intelligence authors, for now. Patterns 2023 Apr;4(4):100731. [doi: 10.1016/j.patter.2023.100731]
6. Use of AI in evidence generation: NICE position statement. National institute of Health and Care Excellence. 2024 Aug 15. URL: https://www.nice.org.uk/corporate/ecd11 [accessed 2026-04-23]
7. EU AI act: first regulation on artificial intelligence. European Parliament. 2023 Jun 8. URL: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence [accessed 2026-04-23]
8. The Artificial Intelligence and Data Act (AIDA) – companion document. Government of Canada. URL: https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document [accessed 2026-04-23]

XSL•FO
RenderX

9. Kucukkaya A, Arikan E, Goktas P. Unlocking ChatGPT's potential and challenges in intensive care nursing education and practice: a systematic review with narrative synthesis. Nurs Outlook 2024;72(6):102287. [doi: 10.1016/j.outlook.2024.102287] [Medline: 39413564]
10. Pressman SM, Borna S, Gomez-Cabello CA, Haider SA, Haider C, Forte AJ. AI and ethics: a systematic review of the ethical considerations of large language model use in surgery research. Healthcare (Basel) 2024 Apr 13;12(8):825. [doi: 10.3390/healthcare12080825] [Medline: 38667587]
11. Younis HA, Eisa TAE, Nasser M, et al. A systematic review and meta-analysis of artificial intelligence tools in medicine and healthcare: applications, considerations, limitations, motivation and challenges. Diagnostics (Basel) 2024 Jan 4;14(1):109. [doi: 10.3390/diagnostics14010109] [Medline: 38201418]
12. Wangsa K, Karim S, Gide E, Elkhodr M. A systematic review and comprehensive analysis of pioneering AI chatbot models from education to healthcare: ChatGPT, Bard, Llama, Ernie and Grok. Future Internet 2024;16(7):219. [doi: 10.3390/fi16070219]
13. Meng X, Yan X, Zhang K, et al. The application of large language models in medicine: a scoping review. iScience 2024 May 17;27(5):109713. [doi: 10.1016/j.isci.2024.109713] [Medline: 38746668]
14. Li J, Zhou Z, Lyu H, Wang Z. Large language models-powered clinical decision support: enhancing or replacing human expertise? Intell Med 2025 Feb;5(1):1-4. [doi: 10.1016/j.imed.2025.01.001]
15. Azam M, Chen Y, Arowolo MO, Liu H, Popescu M, Xu D. A comprehensive evaluation of large language models in mining gene interactions and pathway knowledge. bioRxiv. Preprint posted online on Jan 24, 2024. [doi: 10.1101/2024.01.21.576542]
16. Ahn S. The transformative impact of large language models on medical writing and publishing: current applications, challenges and future directions. Korean J Physiol Pharmacol 2024 Sep 1;28(5):393-401. [doi: 10.4196/kjpp.2024.28.5.393] [Medline: 39198220]
17. Lin C, Kuo CF. Roles and potential of large language models in healthcare: a comprehensive review. Biomed J 2025 Oct;48(5):100868. [doi: 10.1016/j.bj.2025.100868] [Medline: 40311872]
18. Ganzinger M, Kunz N, Fuchs P, et al. Automated generation of discharge summaries: leveraging large language models with clinical data. Sci Rep 2025 May 12;15(1):16466. [doi: 10.1038/s41598-025-01618-7] [Medline: 40355506]
19. Qin H, Tong Y. Opportunities and challenges for large language models in primary health care. J Prim Care Community Health 2025;16:21501319241312571. [doi: 10.1177/21501319241312571] [Medline: 40162893]
20. Jenkins K. 10 year health plan for England. J Kidney Care 2025 Sep 2;10(5):212-213. [doi: 10.12968/jokc.2025.10.5.212]
21. Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing workload in systematic review preparation using automated citation classification. J Am Med Inform Assoc 2006;13(2):206-219. [doi: 10.1197/jamia.M1929] [Medline: 16357352]
22. Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinformatics 2010 Jan 26;11(1):55. [doi: 10.1186/1471-2105-11-55] [Medline: 20102628]
23. Eastaugh CH, Still M, Beyer FR, Wallace SA, O'Keefe H. Exploring the role of artificial intelligence in evidence synthesis: insights from the CORE Information Retrieval Forum 2025. Cochrane Evid Synth Methods 2025 Sep;3(5):e70049. [doi: 10.1002/cesm.70049] [Medline: 40933879]
24. Safety and security risks of generative artificial intelligence to 2025 (Annex B). GOV.UK. 2025 Apr 28. URL: https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper/safety-and-security-risks-of-generative-artificial-intelligence-to-2025-annex-b#conclusions [accessed 2026-04-23]
25. Tredinnick L, Laybats C. The dangers of generative artificial intelligence. Bus Inf Rev 2023 Jun;40(2):46-48. [doi: 10.1177/02663821231183756]
26. von Eschenbach WJ. Transparency and the black box problem: why we do not trust AI. Philos Technol 2021 Sep 1;34(4):1607-1622. [doi: 10.1007/s13347-021-00477-0]
27. Dhar P. The carbon impact of artificial intelligence. Nat Mach Intell 2020;2(8):423-425. [doi: 10.1038/s42256-020-0219-9]
28. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
29. Gates M, Gates A, Pieper D, et al. Reporting guideline for overviews of reviews of healthcare interventions: development of the PRIOR statement. BMJ 2022 Aug 9;378:e070849. [doi: 10.1136/bmj-2022-070849] [Medline: 35944924]
30. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
31. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ 2020 Jan 16;368:l6890. [doi: 10.1136/bmj.l6890] [Medline: 31948937]
32. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 5;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
33. Prasad M. Introduction to the GRADE tool for rating certainty in evidence and recommendations. Clin Epidemiol Glob Health 2024 Jan;25:101484. [doi: 10.1016/j.cegh.2023.101484]
34. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017 Sep 21;358:j4008. [doi: 10.1136/bmj.j4008] [Medline: 28935701]

XSL•FO
RenderX

35. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
36. Klang E, Sourosh A, Nadkarni GN, Sharif K, Lahat A. Evaluating the role of ChatGPT in gastroenterology: a comprehensive systematic review of applications, benefits, and limitations. Ther Adv Gastroenterol 2023;16:17562848231218618. [doi: 10.1177/17562848231218618] [Medline: 38149123]
37. Patil A, Serrato P, Chisvo N, Arnaout O, See PA, Huang KT. Large language models in neurosurgery: a systematic review and meta-analysis. Acta Neurochir (Wien) 2024 Nov 23;166(1):475. [doi: 10.1007/s00701-024-06372-9] [Medline: 39579215]
38. Pressman SM, Borna S, Gomez-Cabello CA, Haider SA, Haider CR, Forte AJ. Clinical and surgical applications of large language models: a systematic review. J Clin Med 2024 May 22;13(11):3041. [doi: 10.3390/jcm13113041] [Medline: 38892752]
39. Tangsrivimol JA, Darzidehkalani E, Virk HUH, et al. Benefits, limits, and risks of ChatGPT in medicine. Front Artif Intell 2025;8:1518049. [doi: 10.3389/frai.2025.1518049] [Medline: 39949509]
40. Wang L, Wan Z, Ni C, et al. Applications and concerns of ChatGPT and other conversational large language models in health care: systematic review. J Med Internet Res 2024 Nov 7;26:e22769. [doi: 10.2196/22769] [Medline: 39509695]
41. Malgaroli M, Hull TD, Zech JM, Althoff T. Natural language processing for mental health interventions: a systematic review and research framework. Transl Psychiatry 2023 Oct 6;13(1):309. [doi: 10.1038/s41398-023-02592-2] [Medline: 37798296]
42. Sharma A, Medapalli T, Alexandrou M, Brilakis E, Prasad A. Exploring the role of ChatGPT in cardiology: a systematic review of the current literature. Cureus 2024 Apr;16(4):e58936. [doi: 10.7759/cureus.58936] [Medline: 38800264]
43. Villanueva-Miranda I, Xie Y, Xiao G. Sentiment analysis in public health: a systematic review of the current state, challenges, and future directions. Front Public Health 2025;13:1609749. [doi: 10.3389/fpubh.2025.1609749] [Medline: 40620557]
44. Abi-Rafeh J, Xu HH, Kazan R, Tevlin R, Furnas H. Large language models and artificial intelligence: a primer for plastic surgeons on the demonstrated and potential applications, promises, and limitations of ChatGPT. Aesthet Surg J 2024 Feb 15;44(3):329-343. [doi: 10.1093/asj/sjad260] [Medline: 37562022]
45. Balla Y, Tirunagari S, Windridge D. Pediatrics in artificial intelligence era: a systematic review on challenges, opportunities, and explainability. Indian Pediatr 2023 Jul 15;60(7):561-569. [Medline: 37424120]
46. Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language models for mental health applications: systematic review. JMIR Ment Health 2024 Oct 18;11:e57400. [doi: 10.2196/57400] [Medline: 39423368]
47. Kiwan O, Al-Kalbani M, Rafie A, Hijazi Y. Artificial intelligence in plastic surgery, where do we stand? JPRAS Open 2024 Dec;42:234-243. [doi: 10.1016/j.jpra.2024.09.003] [Medline: 39435018]
48. Li M, Guenier AW. ChatGPT and health communication: a systematic literature review. IJEHMC 2024;15(1):1-26. [doi: 10.4018/IJEHMC.349980]
49. Wong M, Lim ZW, Pushpanathan K, et al. Review of emerging trends and projection of future developments in large language models research in ophthalmology. Br J Ophthalmol 2024 Sep 20;108(10):1362-1370. [doi: 10.1136/bjo-2023-324734] [Medline: 38164563]
50. Arif F, Safri MK, Shahzad Z, Yasmeen SF, Rahman MF, Shaikh SA. Exploring the application of CHATGPT in plastic surgery: a comprehensive systematic review. J Pak Med Assoc 2024 Apr;74(4 (Supple-4):S17-S28. [doi: 10.47391/JPMA.AKU-9S-04] [Medline: 38712405]
51. Fatima A, Shafique MA, Alam K, Fadlalla Ahmed TK, Mustafa MS. ChatGPT in medicine: a cross-disciplinary systematic review of ChatGPT's (artificial intelligence) role in research, clinical practice, education, and patient interaction. Medicine (Baltimore) 2024 Aug 9;103(32):e39250. [doi: 10.1097/MD.0000000000039250] [Medline: 39121303]
52. Rehman M, Shafi I, Ahmad J, Garcia CO, Barrera AEP, Ashraf I. Advancement in medical report generation: current practices, challenges, and future directions. Med Biol Eng Comput 2025 May;63(5):1249-1270. [doi: 10.1007/s11517-024-03265-y] [Medline: 39707049]
53. Fareed M, Fatima M, Uddin J, Ahmed A, Sattar MA. A systematic review of ethical considerations of large language models in healthcare and medicine. Front Digit Health 2025;7:1653631. [doi: 10.3389/fdgth.2025.1653631] [Medline: 41019285]
54. Nasra M, Jaffri R, Pavlin-Premrl D, et al. Can artificial intelligence improve patient educational material readability? A systematic review and narrative synthesis. Intern Med J 2025 Jan;55(1):20-34. [doi: 10.1111/imj.16607] [Medline: 39720869]
55. Sanjeewa R, Iyer R, Apputhurai P, Wickramasinghe N, Meyer D. Empathic conversational agent platform designs and their evaluation in the context of mental health: systematic review. JMIR Ment Health 2024 Sep 9;11:e58974. [doi: 10.2196/58974] [Medline: 39250799]
56. Omar M, Soffer S, Charney AW, Landi I, Nadkarni GN, Klang E. Applications of large language models in psychiatry: a systematic review. Front Psychiatry 2024;15:1422807. [doi: 10.3389/fpsyt.2024.1422807] [Medline: 38979501]
57. Omar M, Levkovich I. Exploring the efficacy and potential of large language models for depression: a systematic review. J Affect Disord 2025 Feb 15;371:234-244. [doi: 10.1016/j.jad.2024.11.052] [Medline: 39581383]
58. Pashangpour S, Nejat G. The future of intelligent healthcare: a systematic analysis and discussion on the integration and impact of robots using large language models for healthcare. Robotics 2024;13(8):112. [doi: 10.3390/robotics13080112]

XSL•FO
RenderX

59. Sacoransky E, Kwan BYM, Soboleski D. ChatGPT and assistive AI in structured radiology reporting: a systematic review. Curr Probl Diagn Radiol 2024;53(6):728-737. [doi: 10.1067/j.cpradiol.2024.07.007] [Medline: 39004580]
60. Bečulić H, Begagić E, Skomorac R, Mašović A, Selimović E, Pojskić M. ChatGPT's contributions to the evolution of neurosurgical practice and education: a systematic review of benefits, concerns and limitations. Med Glas (Zenica) 2024 Feb 1;21(1):126-131. [doi: 10.17392/1661-23] [Medline: 37950660]
61. Garg RK, Urs VL, Agarwal AA, Chaudhary SK, Paliwal V, Kar SK. Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: a systematic review. Health Promot Perspect 2023;13(3):183-191. [doi: 10.34172/hpp.2023.22] [Medline: 37808939]
62. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs). NPJ Digit Med 2024 Jul 8;7(1):183. [doi: 10.1038/s41746-024-01157-x] [Medline: 38977771]
63. Kiuchi K, Otsu K, Hayashi Y. Psychological insights into the research and practice of embodied conversational agents, chatbots and social assistive robots: a systematic meta-review. Behav Inf Technol 2024 Nov 17;43(15):3696-3736. [doi: 10.1080/0144929X.2023.2286528]
64. Kolding S, Lundin RM, Hansen L, Østergaard SD. Use of generative artificial intelligence (AI) in psychiatry and mental health care: a systematic review. Acta Neuropsychiatr 2024 Nov 11;37:e37. [doi: 10.1017/neu.2024.50] [Medline: 39523628]
65. Mohamed YA, Mohamed AHHM, Khanan A, Bashir M, Adiel MAE, Elsadig MA. Navigating the ethical terrain of AI-generated text tools: a review. IEEE Access 2024;12:197061-197120. [doi: 10.1109/ACCESS.2024.3521945]
66. Paganelli A, Spadafora M, Navarrete-Dechent C, Guida S, Pellacani G, Longo C. Natural language processing in dermatology: a systematic literature review and state of the art. J Eur Acad Dermatol Venereol 2024 Dec;38(12):2225-2234. [doi: 10.1111/jdv.20286] [Medline: 39150311]
67. Kutbi M. Artificial intelligence-based applications for bone fracture detection using medical images: a systematic review. Diagnostics (Basel) 2024 Aug 27;14(17):1879. [doi: 10.3390/diagnostics14171879] [Medline: 39272664]
68. Roman A, Al-Sharif L, AL Gharyani M. The expanding role of ChatGPT (Chat-Generative Pre-Trained Transformer) in neurosurgery: a systematic review of literature and conceptual framework. Cureus 2023 Aug 15;15(8):e43502. [doi: 10.7759/cureus.43502]
69. Rudnicka Z, Proniewska K, Perkins M, Pregowska A. Cardiac healthcare digital twins supported by artificial intelligence-based algorithms and extended reality—a systematic review. Electronics (Basel) 2024;13(5):866. [doi: 10.3390/electronics13050866]
70. Ruksakulpiwat S, Kumar A, Ajibade A. Using ChatGPT in medical research: current status and future directions. J Multidiscip Healthc 2023;16:1513-1520. [doi: 10.2147/JMDH.S413470] [Medline: 37274428]
71. Moya-Salazar J, Salazar CR, Delzo SS, Goicochea-Palomino EA, Rojas-Zumaran V. After a few months, what are the uses of OpenAI's ChatGPT in medicine? A Scopus-based systematic review. Electron J Gen Med 2024;21(2):em577. [doi: 10.29333/ejgm/14354]
72. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
73. Banskota B, Bhusal R, Yadav PK, Banskota AK. Artificial intelligence in orthopaedic education, training and research: a systematic review. BMC Med Educ 2025 Nov 13;25(1):1594. [doi: 10.1186/s12909-025-08162-y] [Medline: 41233816]
74. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282. [Medline: 23092060]
75. Facts and figures. Genetic Alliance UK. URL: https://geneticalliance.org.uk/campaigns-and-research/facts-and-figures/ [accessed 2026-05-02]
76. Wang L, Wan Z, Ni C, et al. A systematic review of chatgpt and other conversational large language models in healthcare. medRxiv. Preprint posted online on Apr 27, 2024. [doi: 10.1101/2024.04.26.24306390]
77. Confronting antimicrobial resistance 2024 to 2029. GOVUK. 2024 May 8. URL: https://www.gov.uk/government/publications/uk-5-year-action-plan-for-antimicrobial-resistance-2024-to-2029/confronting-antimicrobial-resistance-2024-to-2029 [accessed 2026-05-02]
78. Katirai A. The environmental costs of artificial intelligence for healthcare. ABR 2024 Jul;16(3):527-538. [doi: 10.1007/s41649-024-00295-4]
79. Cowls J, Tsamados A, Taddeo M, Floridi L. The AI gambit: leveraging artificial intelligence to combat climate change-opportunities, challenges, and recommendations. AI Soc 2023;38(1):283-307. [doi: 10.1007/s00146-021-01294-x] [Medline: 34690449]
80. Bai X, Wang A, Sucholutsky I, Griffiths TL. Explicitly unbiased large language models still form biased associations. Proc Natl Acad Sci U S A 2025 Feb 25;122(8):e2416228122. [doi: 10.1073/pnas.2416228122] [Medline: 39977313]
81. Hatem R, Simmons B, Thornton JE. A call to address AI "hallucinations" and how healthcare professionals can mitigate their risks. Cureus 2023 Sep;15(9):e44720. [doi: 10.7759/cureus.44720] [Medline: 37809168]
82. Thaler SL. "Virtual input" phenomena within the death of a simple pattern associator. Neural Netw 1995 Jan;8(1):55-65. [doi: 10.1016/0893-6080(94)00065-T]
83. Maleki N, Padmanabhan B, Dutta K. AI hallucinations: a misnomer worth clarifying. In: 2024 IEEE Conference on Artificial Intelligence (CAI: IEEE; 2024:133-138. [doi: 10.1109/CAI59869.2024.00033]

XSL•FO
RenderX

84. Boksa P. On the neurobiology of hallucinations. J Psychiatry Neurosci 2009 Jul;34(4):260-262. [doi: 10.1139/jpn.0934] [Medline: 19568476]
85. Wu WT, Chao YW, Lin TK, Huang CK, Hsieh PH. Economic evaluation of AI-assisted technologies in healthcare: a systematic review. J Food Drug Anal 2025 Dec 15;33(4):487-500. [doi: 10.38212/2224-6614.3570] [Medline: 41525195]

Abbreviations

AI: artificial intelligence
AMSTAR-2: A Measurement Tool to Assess Systematic Reviews
DOI: Digital Object Identifier
GDPR: General Data Protection Regulation
GenAI: generative artificial intelligence
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
LLM : large language model
NICE : National Institute of Health and Care Excellence
PRIOR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses - Search
SPIDER : Sample, Phenomenon of Interest, Design, Evaluation, Research Type

Edited by S Brini; submitted 14.Nov.2025; peer-reviewed by AA Biswas, S Sengupta; revised version received 08.Apr.2026; accepted 09.Apr.2026; published 15.May.2026.

Please cite as:
Yarar F, Addis P, Fairweather M, Craig D, O'Keefe H
Concerns of Using Large Language Models in Health Care Research and Practice: Umbrella Review
J Med Internet Res 2026;28:e87804
URL: https://www.jmir.org/2026/1/e87804
doi:10.2196/87804

© Feyza Yarar, Pauline Addis, Megan Fairweather, Dawn Craig, Hannah O'Keefe. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Understanding Remission of Long-Term Conditions Through Electronic Health Records: Scoping Review

Hilda Hounkpatin[1], PhD; Benjamin Barton[1], MSc; Margaret Ogden[1]; Rohini Mathur[2], PhD; Beth Stuart[2], PhD; Hajira Dambha-Miller[1], PhD

[1]Primary Care Research Centre, University of Southampton, Southampton, United Kingdom

[2]Wolfson Institute of Population Health, Queen Mary University of London, London, United Kingdom

Corresponding Author:
Hilda Hounkpatin, PhD
Primary Care Research Centre, University of Southampton, Southampton, United Kingdom

Abstract

Background: Multiple long-term conditions (MLTCs) require complex and prolonged treatment regimens. Remission in long-term conditions (LTCs) is important for understanding disease progression and evaluating treatment effectiveness. Electronic health records (EHRs) are increasingly used to monitor clinical outcomes, but how remission is defined within EHRs remains unclear.

Objective: This study aimed to summarize and collate the previous literature on how remission of LTCs has been defined in EHRs.

Methods: Systematic electronic searches were performed on OVID MEDLINE, Embase, CINAHL EBSCO, the Cochrane Library, and the Bielefeld Academic Search Engine for eligible studies published from inception to November 27, 2025. Quantitative studies, published in any language, on adult populations, and using EHRs to assess remission of LTCs, were eligible for inclusion. Studies that did not clearly define remission and studies on cancer remission were excluded. Data were extracted from each eligible study using a structured table. Risk of bias was not assessed, in line with scoping review methodology. A narrative approach was taken to summarize and present data from the included studies. The number and characteristics of studies were described, both overall and by condition. Findings were discussed with clinicians and data experts to ensure applicability in clinical practice.

Results: Ninety-one studies were included. Sample sizes ranged from 12 to 72.9 million adults. Studies were conducted in 18 countries, with the majority being from the United States. The majority of included studies used a cohort study design. Studies assessed how remission was defined in 12 LTCs, including inflammatory bowel disease (41/91, 45.1%), type 2 diabetes (n=15, 16.5%), depression (n=15, 16.5%), alcohol or drug misuse (n=8, 8.8%), asthma (n=3, 3.3%), multiple sclerosis (n=3, 3.3%), epilepsy (n=1, 1.1%), anemia (n=1, 1.1%), chronic kidney disease (n=1, 1.1%), autoimmune pancreatitis (n=1, 1.1%), hypertension (n=1, 1.1%), heart failure (n=1, 1.1%), and MLTC (n=1, 1.1%). Remission was typically defined using a combination of clinical codes (n=7, 7.7%), validated rating scales (n=56, 61.5%), biochemical markers (n=29, 31.9%), absence of symptoms (n=10, 11%), absence of condition-specific events (eg, hospital admissions; n=4, 4.4%), and cessation of pharmacological treatments (n=26, 28.6%). There was substantial variation in the criteria and duration of follow-up used to define remission across studies.

Conclusions: This review demonstrates that remission of LTCs can be identified and operationalized within EHRs, although remission criteria varied across studies. The review extends the literature on remission in EHRs by combining evidence synthesis and consultation with clinical and data experts to propose standardized comprehensive definitions to reliably define and implement remission of multiple LTCs in EHR-based research. This will allow cross-study comparisons and present an opportunity to advance understanding of disease trajectories and improve evaluation and monitoring of patient outcomes. Further research may apply, compare, and evaluate standardized definitions across different data sources to assess generalizability and further improve our understanding of remission of LTCs.

(J Med Internet Res 2026;28:e80796) doi:10.2196/80796

KEYWORDS

remission; resolution; long-term conditions; electronic health records; scoping review

XSL•FO
RenderX

Introduction

Multiple long-term conditions (MLTCs)—the presence of 2 or more long-term conditions within an individual [1]—are a major challenge to public health, as their management often involves prolonged treatment and care [2]. People living with MLTCs frequently face high treatment burden, reduced quality of life, and functional decline and are at increased risk of mortality and extensive health and social care use [3-8]. MLTCs are associated with worse health outcomes, with risks of mortality and hospitalization escalating sharply for individuals with 4 or more long-term conditions (LTCs) [9,10]. Effective management of MLTCs remains an unmet research priority, and there is a pressing need to identify new strategies to improve care and outcomes for individuals living with these conditions [11].

For many LTCs, remission—defined as the resolution or cessation of the disease—marks a positive clinical outcome, whether temporary or permanent. Evidence suggests that remission in various LTCs is associated with improved health outcomes [12-16]. For example, previous research using a population-based cohort in South England found that 19.1% of people living with type 2 diabetes achieved remission within a 2-year follow-up period, and those who achieved remission had a lower risk of cardiovascular outcomes and mortality [13]. Similarly, a national Swedish study found that 39.3% of people living with hypertension achieved remission during a 2-year follow-up period, and people who achieved remission had half the risk of developing cardiovascular events and mortality compared to those who did not [14]. Studies have also reported favorable clinical outcomes for people who experience remission in other conditions, such as depression, cancer, or heart disease [15,16]. These previous studies suggest that efforts to support people living with MLTCs to achieve remission could improve their health. However, this research has largely focused on specific conditions (and populations with individual conditions), making it less clear how remission of any LTC in the context of MLTCs may relate to outcomes.

Furthermore, some of these previous studies have identified remission using survey or interview studies [15,16], and it is therefore unclear how remission can be identified in electronic health records (EHRs). The study of disease and MLTCs has increasingly relied on EHRs, which have become essential tools for tracking patient health over time. EHRs provide valuable insights into disease progression and are central to the ongoing efforts to improve clinical decision-making, monitoring, and health policy [17-19]. However, there is limited research on remission in the context of MLTCs. This may be due to a lack of a consistent definition or method for defining and tracking remission of LTCs within EHRs, making the study of remission in EHRs challenging. Variation in clinical coding practices, the lack of dedicated remission fields, and differing interpretations of disease remission all further hinder comparability across studies. A consensus on the definition of remission of LTCs in EHRs is crucial to advancing MLTC research (eg, through allowing accurate estimates of MLTC prevalence and associations) and improving care (eg, tracking remission outcomes accurately), both in the United Kingdom and nationally.

As the use of EHRs in clinical research grows, it is timely to examine how remission of LTCs is defined and tracked within these systems. We conducted a scoping review, as this methodology allows for mapping the existing literature on the definition and identification of remission in EHRs and highlighting gaps in the research. For this review, we used the term remission to refer to either remission or resolution. The decision to use the term “remission” was guided by our public advisory group (people with lived experience of MLTC) who felt that many of their conditions could be controlled but not permanently resolved. Our public advisory group consisted of people from diverse backgrounds (in terms of sex, ethnicity, and socioeconomic position) and health conditions and have provided input at each stage of our study design. This scoping review aimed to collate and summarize the existing literature on how remission of LTCs that are amenable to remission has been defined and evaluated in studies using EHRs.

Methods

Review Approach

The study followed a scoping review methodology outlined by the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [20,21]. For the purpose of this review, MLTCs were defined based on a consensus process involving experts in the field and reported in detail previously [22,23]. This definition includes 56 LTCs selected for their chronic, long-term impact on health, requiring ongoing management or treatment, based on established diagnostic criteria, significant population prevalence, and known effects on morbidity and mortality. The list of conditions is provided in Table S1 in Multimedia Appendix 1.

Protocol and Registration

A review protocol for this study does not exist in any public domain.

Information Sources

Systematic searches were conducted in 5 electronic databases: OVID MEDLINE, Embase, CINAHL, the Cochrane Database of Systematic Reviews, and Bielefeld Academic Search Engine on November 27, 2025, for studies published from inception to date. These databases were selected to capture a range of relevant studies in the biomedical, nursing, and behavioral sciences literature.

Search Strategy

The search strategy is reported in line with the PRISMA-S guidelines [24]. Terms to include in our search strategy were identified based on previous similar or relevant studies, searches for MeSH (Medical Subject Headings) terms and keywords or index terms in each database, and support from a librarian at the University of Southampton. Initial searches were conducted in Ovid MEDLINE and Embase. Both databases were searched for MeSH terms and free-text keywords related to “remission,” “resolution,” and “electronic health records” or “electronic medical records,” and “chronic disease” or “long-term condition.” Searches were adjusted for both databases according

XSL•FO
RenderX

to their specific indexing. We screened the first 100 records captured in each database to identify key terms used in the title and abstracts of these records. We then refined our search strategy based on our findings, including any additional terms identified in our initial screening. We applied our refined search strategy to each of our 5 databases, adapting the search strategy as required. The search strategies for each database are available in Multimedia Appendix 2. Reference lists of included studies were not searched for additional papers, and study authors were not contacted for any additional information. Gray literature searches were conducted mainly through Bielefeld Academic Search Engine database and Google Scholar searches to identify any additional relevant references (eg, conference abstracts and theses). We did not contact authors to identify additional sources.

Eligibility Criteria

Quantitative studies were eligible for inclusion if they were conducted using EHRs, focused on an adult (aged ≥18 y) population, and assessed 1 or more of the 56 LTCs agreed by consensus work within the definition of MLTC [22,23]. In line with most existing research on MLTC, an adult population was selected as MLTC is more common in adults [25], and the criteria used to define remission may vary for adults and children. Studies were excluded if the definition of remission was not reported or unclear. Studies on remission of cancer were excluded as this is already a well-researched area [26]. Registry data and studies that did not assess remission using EHRs were excluded. Studies that reported on remission of symptoms (eg, seizures) rather than remission of an LTC of interest were also excluded. Search limits for studies on humans were applied. No language restrictions were applied.

Selection of Sources of Evidence

Search results were exported to EndNote (Clarivate) for deduplication and then imported into the Rayyan collaborative systematic review platform for screening and final study selection. Titles and abstracts of records were first screened against eligibility criteria independently by 2 reviewers (BB and HH). Discrepancies in study selection were resolved by discussion between the 2 reviewers. Full texts of potentially relevant records were sought, retrieved, and reviewed in detail by HH.

Data Charting Process and Data Items

Data were extracted from each eligible study using a structured table developed a priori by HH (following discussions with the research team). Data extracted included the LTCs assessed, author name and publication date, country, study design, population studied, study aims, remission definition (eg, clinical criteria and biomarkers), and key findings related to remission. Data were extracted by HH, and 50% of the extracted data were validated by BB. Where a single study (from the same first author) generated multiple papers, reports, or abstracts, we included the paper with the most comprehensive definition of remission. We did not contact investigators to obtain or confirm information.

Critical Appraisal of Individual Sources of Evidence

Quality of the identified studies was not assessed as this is not a requirement of scoping reviews.

Synthesis of Results

A narrative approach was taken to summarize our findings from included studies. The total number of included studies, characteristics of studies (eg, study country, study design, population studied, and setting) was described. Studies were categorized into groups of interrelated conditions, and the number of studies that focused on each specific condition and the characteristics of the studies within each group summarized. Similarities and differences in definitions and methods used to identify remission of LTCs across studies were highlighted.

Consultation With Clinicians and Data Experts

To ensure the clinical applicability of our findings, the results were discussed with a group of clinicians and data experts. These discussions were pivotal in refining the identified operational definition of remission within EHRs, incorporating insights from real-world clinical practice. Key considerations included the variability in clinical decision-making, the complexities of disease progression, and the practical limitations of current EHR systems. This collaborative process was essential for aligning our findings with the realities of patient care and the capabilities of existing health informatics systems.

We further cross-referenced any remaining conditions from our previously agreed list with expert clinical input to derive a refined list of long-term conditions amenable to remission assessment. This was an iterative process underpinned by clinical judgment to ensure both practical feasibility and clinical relevance. We then searched for relevant medical codes used to define remission (based on our agreed definitions) using Clinical Practice Research Datalink (CPRD) Aurum medical and product data dictionaries. CPRD Aurum contains routinely collected primary care data from practices across England [27]. We also searched coding systems (eg, GitHub repositories, London School of Hygiene and Tropical Medicine Data Compass) and published code lists for any relevant codes to ensure a complete code list. Where older code lists were identified, we used description terms from these code lists to identify relevant codes in our data.

Results

Selection of Sources of Evidence

Figure 1 illustrates the outcomes of the systematic search and selection process. A total of 2064 records were identified through our searches. Following deduplication (n=87/2064, 4.2%) and title or abstract screening, 252 (12.2%) records met the criteria for full-text review. Of those records, 53.2% (134/252) were excluded following full-text review. A total of 91 (36.1%) final studies were included in our review, which included 4 additional studies identified by the research team.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram illustrating search results and screening process.

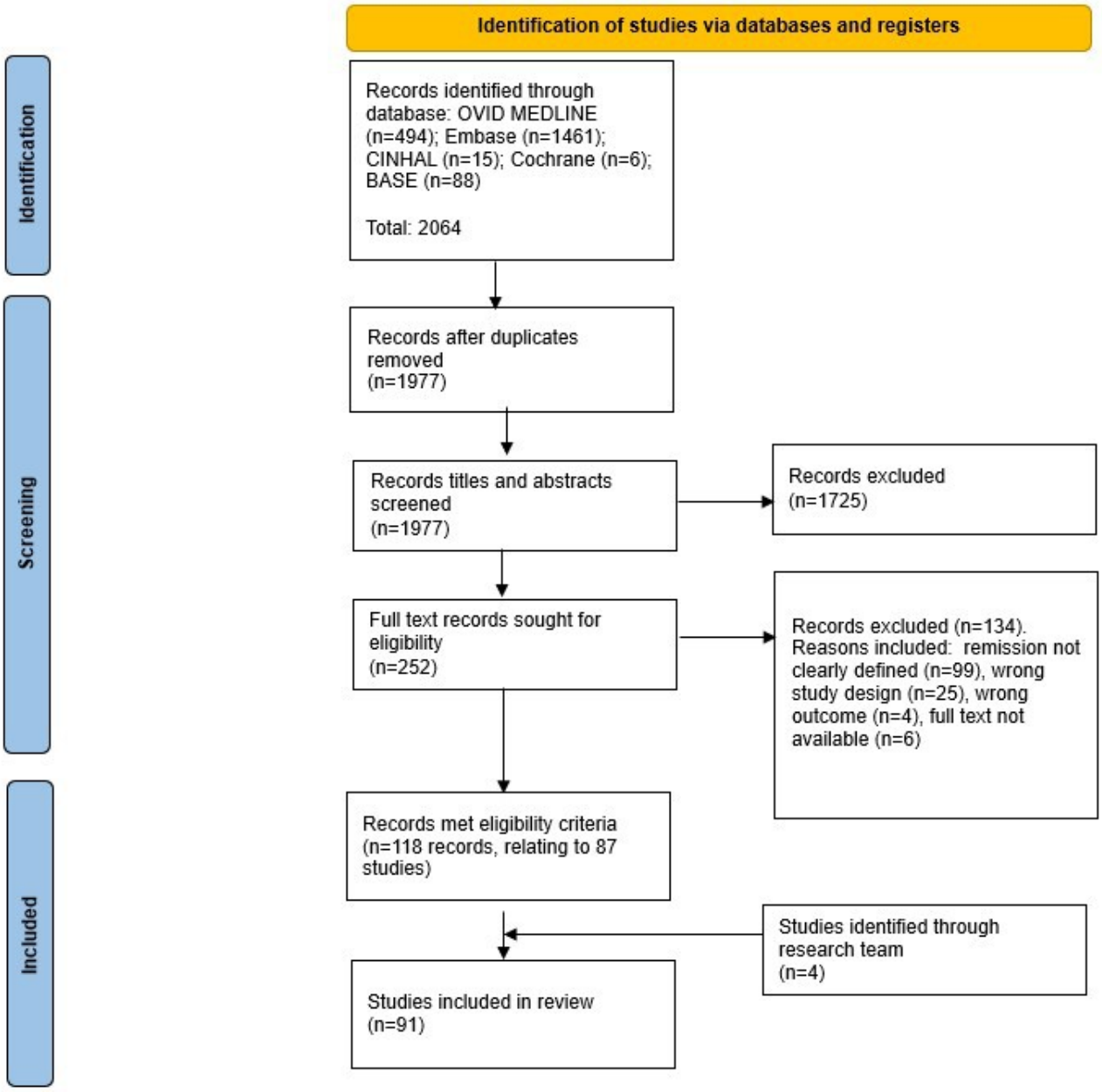

Characteristics of Sources of Evidence

Included studies were published between 2008 and 2025. The majority of the included studies were conducted in the United States (49/91, 53.8%) [28-76] followed by the United Kingdom (n=13, 14.3%) [13,77-88] and China (n=4, 4.4%) [89-92]. Three (3.3%) studies were conducted in Australia [93-95] and Saudi Arabia [96-98]. Two (2.2%) studies each were conducted in Italy [99,100], Japan [101,102], New Zealand [103,104], Spain [105,106], and Turkey [107,108]. One (1.1%) study each was conducted in Austria [109], Belgium [110], Denmark [111], Finland [112], Kuwait [113], Pakistan [114], South Korea [115], and Sweden [116]. For 1 (1.1%) study, it was unclear which country the study was based [117]. The included studies comprised cohort studies (n=70, 76.9%) [13,28-31,33,35-38,40-42,45,47-51,53-68,71,72,74-79,82-84,86,88-92,95,97,99-106,109-112,114-117], retrospective reviews (n=14, 15.4%) [32,39,43,44,46,52,70,80,81,85,93,98,108,113], cross-sectional studies (n=6, 6.6%) [34,73,88,94,96,107], and an exploratory study (n=1, 1.1%) [87]. Most studies focused on a general adult population (aged ≥18 y) or did not specify an age range, with a small number specifically investigating older populations (eg, veterans or those aged ≥60 y; n=6, 6.6%) [34,35,38,41,81,89] or specific age ranges (n=6, 6.6%) [36,63,75,96,97,112]. Sample sizes ranged from 12 patients to 72.9 million adults. Summary characteristics of the included studies are presented in Textbox 1. Full details of all included studies are available in Supplementary Table 2 in Multimedia Appendix 2.

XSL•FO
RenderX

Textbox 1. Characteristics of included studies.

- Number of publications: 91
- Publication date range: 2008 - 2025
- Country of study: United States (n=49), United Kingdom (n=13), China (n=4), Australia (n=3), Saudi Arabia (n=3), Italy (n=2), Japan (n=2), New Zealand (n=2), Spain (n=2), Turkey (n=2), Austria (n=1), Belgium (n=1), Denmark (n=1), Finland (n=1), Kuwait (n=1), Pakistan (n=1), South Korea (n=1), Sweden (n=1), and unclear (n=1)
- Study design: cohort studies (n=70), retrospective reviews (n=14), cross-sectional studies (n=6), and experimental or exploratory studies (n=1)
- Population: general or unspecified adult population (n=79), midlife population (n=6), and older populations (age ≥60 y; n=6)
- Sample size range: 12-72.9 million patients
- Condition studied: inflammatory bowel disease (n=41), type 2 diabetes (n=15), depression (n=15), alcohol or drug misuse (n=8), asthma (n=3), multiple sclerosis (n=3), anemia (n=1), chronic kidney disease (n=1), autoimmune pancreatitis (n=1), epilepsy (n=1), heart failure (n=1), hypertension (n=1), and multiple long-term conditions (n=1).

Critical Appraisal Within Sources of Evidence

The risk of bias in studies was not assessed, in line with scoping review methodology.

Results of Individual Sources of Evidence: Remission Findings

The majority of the included studies focused on remission in inflammatory bowel disease (IBD; 45.1%, 41/91) [43-61,80-85,90-95,99-101,106,109-111,113,115,117], type 2 diabetes (n=15, 16.5%) [13,62-69,86,87,102-104,114] (one additionally assessing remission of hypertension [103]), and depression (n=15, 16.5%) [32-42,76,78,105,108]. Eight (8.8%) studies assessed remission of drug or alcohol misuse [70-75,107,112], and 3 (3.3%) studies each assessed remission of asthma [29-31] and multiple sclerosis [96-98]. One (1.1%) study each assessed remission in each of anemia [28], chronic kidney disease (CKD) [77], autoimmune pancreatitis [116], epilepsy [79], and heart failure [89]. Additionally, 1 (1.1%) study investigated remission in MLTCs [88].

Results of Individual Sources of Evidence: Definitions of and Methods Used to Identify Remission in LTCs

Overview

Across studies, remission was typically defined using one or more diagnostic codes, validated rating scales, biochemical markers, or the absence of condition-specific events (eg, hospital admissions) following the discontinuation of prescribed pharmacological treatments. Figure 2 presents the distribution of remission indicators across studies for each condition. Most studies defined remission using biochemical tests or markers and validated rating scales.

Figure 2. Percentage of studies on specified conditions that used each remission indicator.

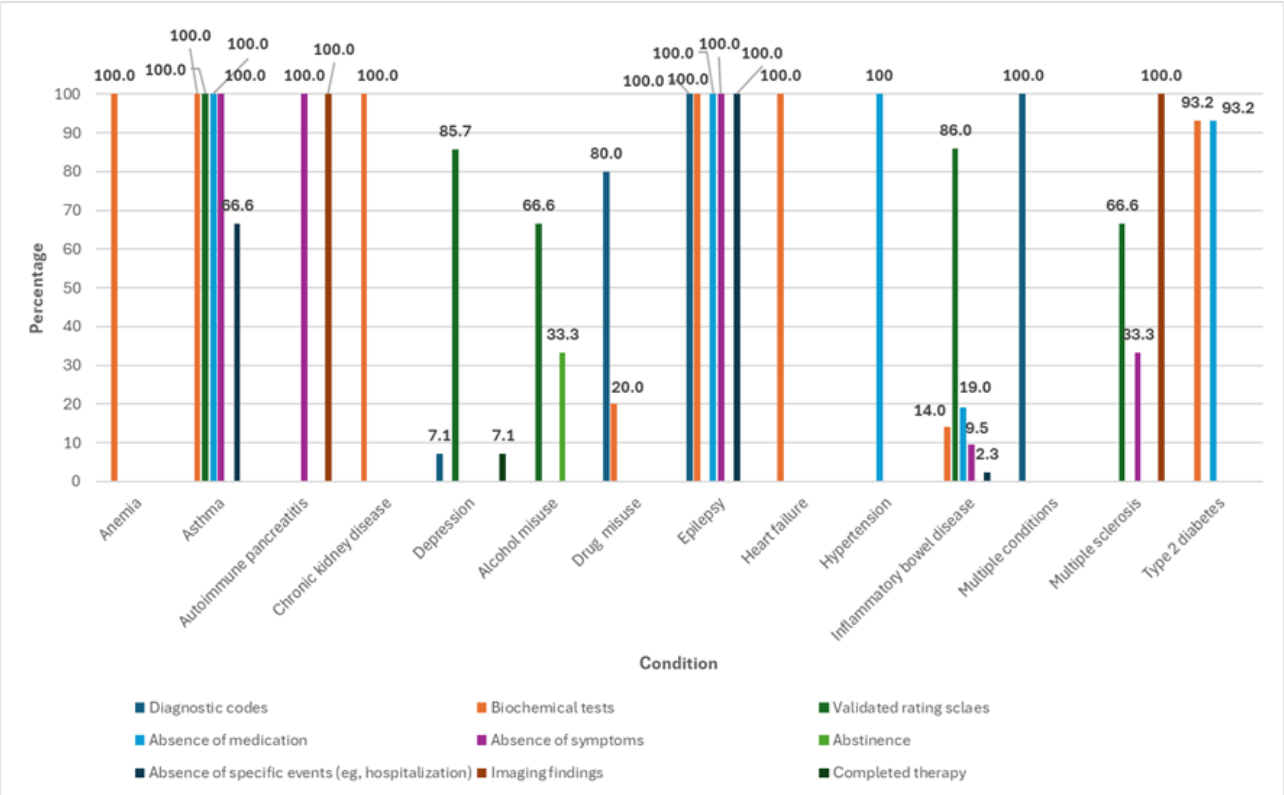

XSL•FO
RenderX

Autoimmune Conditions

Inflammatory Bowel Disease

Studies that assessed remission of IBD focused either on Crohn disease (9/41, 22%) [44,50,52,83,95,99,109,111,117] or ulcerative colitis (n=8, 19.5%) [43,55,56,58,59,90,92,115] or both (n=24, 58.5%) [45-49,51,53,54,57,60,61,80-82,84,85,91,93,94,100,101,106,110,113].

All 9 (100%) studies on remission of Crohn disease assessed clinical remission, with 2 (22.2%) specifying steroid-free clinical remission [99,117]. One (11.1%) study each additionally assessed endoscopic remission [82] and biomarker remission [109]. Most studies (n=5, 55.5%) assessed clinical remission of Crohn disease using the Harvey-Bradshaw Index; the majority of studies (n=4, 44.4%) [44,83,95,99] considered a score of <5 at follow-up as remission, while 1 (11.1%) study used a cutoff score of ≤5 [111]. One (11.1%) study defined clinical remission using Crohn Disease Activity Index score of 150. Two (22.2%) studies defined clinical remission based on stool frequency and abdominal pain, with different thresholds [52,109]. One (11.1%) study defined clinical remission as closure of all baseline fistulas [50]. Meade et al [83] additionally defined endoscopic remission using Simple Endoscopic Score for Crohn disease (SES-CD; ≤2 indicating remission). Pokryszka et al [109] additionally assessed biomarker remission based on fecal calprotectin values of ≤150 µg/g [107] and reported lower rates of clinical versus biomarker remission. Length of follow-up varied across studies and ranged from 3 to 6 months (Figure 3).

Figure 3. Range of follow-up duration across studies per condition.

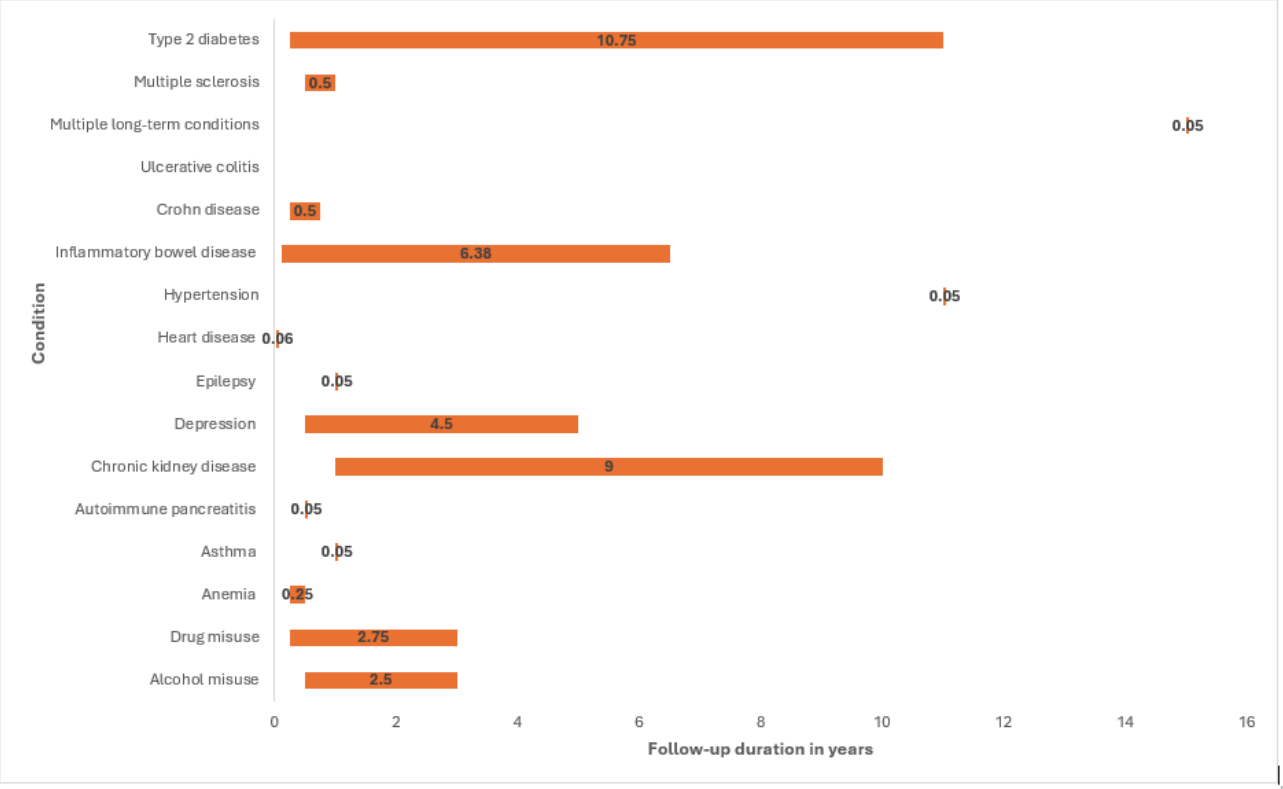

Seven (87.5%) of 8 studies on ulcerative colitis remission assessed clinical remission [43,56,58,59,90,92,115]. Two (25.0%) studies considered steroid-free clinical remission [56,58]. Clinical remission was mostly defined as a partial Mayo score <2 at follow-up [56,58,59,115]. One (12.5%) study defined clinical remission as partial Mayo score ≤2 without any subscore of >1 [90]. Two (25.0%) studies defined clinical remission using Simple Clinical Colitis Activity Index and reported a cutoff score of <2 [43] or ≤2 [92] at follow-up as clinical remission. Four (50.0%) studies assessed and defined endoscopic remission as Mayo Endoscopic Score (MES) of ≤1 [55,59,92,115], with 1 (12.5%) study considering endoscopic remission as MES of ≤1 or absence of erosions or ulcerations [59]. Zeina et al [55] additionally assessed histologic remission, defined as Simplified Geboes Score ≤0.2. Length of follow-up varied across studies and ranged from 6 weeks to 6.5 years (Figure 3).

Twenty-four (100.0%) studies assessed clinical remission in IBD (both Crohn disease and ulcerative colitis). Three (12.5%) studies specified steroid-free clinical remission [54,81,113]. Studies used multiple scales, including the Harvey-Bradshaw Index, with most (50.0%) defining remission as cutoff score of <5 [45,46,48,54,61,80,81,85,106] or <4 [47,49,93] at follow-up; partial Mayo score of <2 [48,49,51,61,80,84] or ≤2 [93,101] at follow-up; Simple Clinical Colitis Activity Index ≤2 [45,53,54,81] or <2 [49,51,80] at follow-up; Crohn Disease Activity Index <150 [91,101]; the Ulcerative Colitis Activity Index ≤4 [47] at follow-up. Two (8.3%) studies on clinical remission of IBD additionally considered absence of symptoms [46,110] and/or hospitalization [46]. Seven (29.2%) studies assessed endoscopic remission [51,53,57,91,100,110,113], defining remission as MES <2 [51,100,110,113], SES-CD of 0 [53,57], SES-CD of <3 [91], or SES-CD of <4 [100] at follow-up. Six (25.0%) studies assessed biochemical remission [80,82,84,94,100,106], defining remission as fecal calprotectin

XSL•FO
RenderX

of <250 µg/g [82,84,94,106] or fecal calprotectin of <150 µg/g [100] at follow-up. Two (8.3%) studies additionally assessed remission as C-reactive protein ≤5 mg/L [84] or C-reactive protein <10 mg/L and/or albumin >35 g/L [82]. Length of follow-up ranged from 8 weeks to 6.5 years (Figure 3).

Autoimmune Pancreatitis

One (100.0%) study assessed the outcomes of remission in autoimmune pancreatitis [116]. Remission was defined as the absence of clinical symptoms (asymptomatic, new-onset diabetes, acute pancreatitis, obstructive jaundice, weight loss, and abdominal pain) and absence of pancreatic abnormalities on imaging at 6-month follow-up [116].

Cardiometabolic Conditions

Type 2 Diabetes

Thirteen (86.7%) of the 15 studies assessing remission of type 2 diabetes defined remission as controlled glycemia for a defined period in the absence of antidiabetic medication [13,62-69,86,87,102]. Most (n=11, 73.3%) of these studies assessed glycemic control using glycated hemoglobin A_{1c} biochemical tests, although thresholds varied and included <6.5% (48 mmol/mol) [13,63-69,86,87,102] or <7% (53 mmol/mol) [62]. Duration of glycemic control in absence of medication varied across these studies and included ≥3 months [62,69,87], ≥6 months [13], ≥1 year [63-67,86,102,114], or ≥5 years [68]. One (6.7%) separate study defined type 2 diabetes remission in bariatric and metabolic surgery patients as glycemic control (hemoglobin A_{1c} level<43 mmol/mol [6.1%]) at 6 months [104]. Another study defined type 2 diabetes remission based on the absence of medication at 11 years following gastric bypass surgery [103].

Heart Failure

One (100%) study assessed remission of heart failure using blood test results [89]. Heart failure was defined as persistent elevated N-terminal pro-B-type natriuretic peptide levels within 7 and 30 days.

Hypertension

Sheikh et al [103] assessed hypertension remission based on the absence of antihypertensive medication. We did not identify other studies that assessed remission of hypertension (rather than control). However, 1 (1/91, 1.1%) of our included studies (that assessed remission in depression) assessed hypertension control using blood pressure measurements at 6-month follow-up [39]. Different cutoffs were used for different age groups (18-59 and 60-85 y) and presence of diabetes.

Related Conditions

Sheikh et al [103] assessed hypertension remission based on the absence of antihypertensive medication. We did not identify other studies that assessed remission of hypertension (rather than control). However, 1 (1/91, 1.1%) of our included studies (that assessed remission in depression) assessed hypertension control using blood pressure measurements at 6-month follow-up [39]. Different cutoffs were used for different age groups (18-59 and 60-85 y) and presence of diabetes.

Anemia

One (100%) study on patients with chronic stable heart failure assessed remission of anemia [28]. Remission of anemia was defined as normalization of hemoglobin levels (≥12 g/dL for men and ≥11 g/dL for women) with at least improvement in ≥0.5 g/dL at 6-month follow-up.

Chronic Kidney Disease

One (100%) study on CKD remission defined remission using biochemical tests (estimated glomerular filtration rate ≥60 mL/min/1.73 m^2) at 1-, 5-, and 10-year follow-up [77].

Depression

Remission of depression was mostly defined as normal scores on validated rating scales such as Patient Health Questionnaire-9 scores of <5 [32,34,36,38,39,42,78], ≤5 [33,41], <10 (in patients receiving electroconvulsive therapy) [37], or a 50% decrease from baseline in Patient Health Questionnaire-9 score (in patients aged ≥60 y) [35], Hamilton Depression Rating Scale ≤7 [108], or Geriatric Depression Scale <3 [76] at follow-up. Follow-up varied from 2 months to 1 year after intervention, with most studies assessing remission at 6 months [32,36,39,41] and/or 1 year [33,34,37,78]. One (1/15, 6.7%) study assessed remission using both the Geriatric Depression Scale and remission codes in the EHR over a median follow-up of 53 (IQR 42.1-60.5) months (4.4 y) [76]. Another (6.7%) study assessed remission using codes in the EHR [40]. One (6.7%) study considered remission of major depression as completion of 6 months of therapy [105].

Alcohol or Drug Misuse

Eight (100%) studies assessed remission of drug or alcohol misuse [70-75,107,112]. Three (37.5%) studies focused on alcohol misuse [70,71,112]. Two (25%) studies defined remission as negative screening (using the validated alcohol harm assessment tool (Alcohol Use Disorders Identification Test - Consumption score<5) [70] or the National Institute on Alcohol Abuse and Alcoholism single screening question) at follow-up. Follow-up varied across the 2 (25%) studies (1 y [70] and up to 3 y [71]). Rautiainen et al [112] assessed remission of alcohol misuse over a minimum follow-up period of 6 months and defined remission using diagnostic codes or health professional notes in the EHR.

Four (80%) of 5 studies that assessed remission of drug misuse defined remission based on clinical codes (diagnostic codes for remission or absence of codes relating to opioid positivity) [72-75]. Follow-up of studies ranged from 3 months to 3 years.

Asthma

Three (100%) studies [29-31] assessed clinical remission of asthma, with 2 (66.7%) reporting a 1-year follow-up [30,31]. For all 3 studies, remission of asthma was defined using multiple clinical indicators. These included the absence of asthma exacerbations, absence of asthma medication, ≥2 normal Asthma Control Test levels, ≥2 normal pulmonary function test values, lung function stabilization, and no absence from work or school due to asthma.

XSL•FO
RenderX

Epilepsy

One (100%) study assessed remission in epilepsy at 1-year follow-up [79]. This study defined remission as absence of seizures, absence of antiepileptic drugs, and the absence of all seizure-related health care events, such as general practitioner appointments or hospitalizations.

Multiple Sclerosis

Three (100%) [96-98] studies assessed remission of multiple sclerosis, with 2 (66.7%) reporting a follow-up of 1 year [96,97]. Remission was indicated from no evidence of disease activity, which comprised (1) no new or enlarged T2- or gadolinium-enhancing lesions, (2) no clinical relapse (ie, return of old symptoms or worsening of current multiple sclerosis symptoms), and (3) no disability progression (ie, increased walking difficulty leading to progression or an increase in Expanded Disability Status Scale of ≥1.5 between 2 time points).

Multiple Long-Term Conditions

One (100%) study assessed remission in 32 LTCs. Identification of remission was based on clinical codes (for remission) available in primary care EHRs [88].

Synthesis of Results

Figure 4 presents an evidence map of remission definition across studies included in the review and highlights potential gaps in the literature. Our review highlighted a lack of studies that examined the extent to which varying follow-up periods or the use of different remission indicators (laboratory tests vs diagnostic remission codes) influenced remission rates, except for studies on depression and IBD that reported remission rates across different follow-up durations. Some studies (eg, on IBD) did not specify the absence of medication in their definition of remission, making it less clear whether patients were truly in remission and making cross-study comparisons difficult.

Figure 4. Evidence map of included studies.

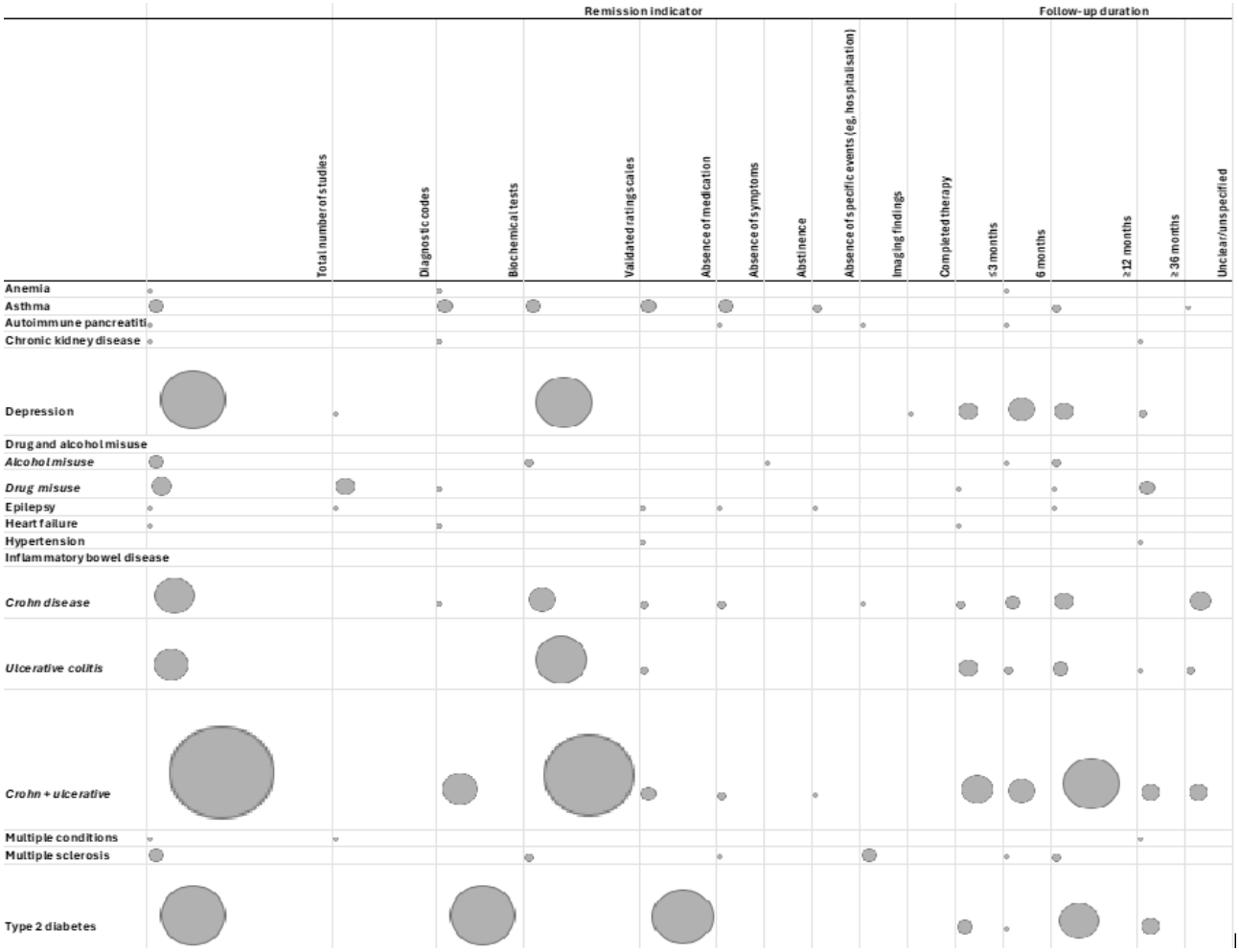

Operationalization of Remission in EHR Data

This review facilitated the identification of conditions for which remission is measurable within EHRs. Although the scoping review highlighted additional conditions with theoretical potential for remission, discussions with clinicians and CPRD data experts indicated that, in many instances, the lack of suitable EHR-derived metrics or insufficient sample sizes would preclude their meaningful inclusion. Accordingly, we restricted our final list to conditions for which remission was both plausibly measurable and supported by validated indicators within the CPRD dataset, recognizing that this pragmatic approach may omit some conditions capable of achieving remission. Table 1 presents our refined list of long-term conditions based on clinical input and availability of codes in EHRs.

XSL•FO
RenderX

Table . Recommendations for definition of remission of long-term conditions (in addition to use of diagnostic [remission] codes) based on scoping review findings and discussions with clinicians and data experts.

Long-term condition and explanation of how remission is defined within studies	Period	Clinical considerations
Anemia		
Normalization of hemoglobin levels (≥12 g/dL for men and ≥11 g/dL for women) with at least improvement in ≥0.5 g/dL for 3 mo	≥3 mo	Insufficient information on anemia type or cause hinders meaningful remission assessment
Alcohol dependency		
Screening negative (AUDIT-C[a] <5) with at least a 2-point reduction from the previous score	≥12 mo	Interpretation should ensure adequate follow-up across participants and account for the possibility that missing data may reflect follow-up occurring in specialist drug and alcohol services, which are not routinely captured in standard health records
Asthma		
No asthma-related event (defined as GP[b]) consultation or admission with a diagnosis code for asthma), no asthma exacerbation, normal lung function and asthma control, and no current use of any asthma medication	≥12 mo	It is important to consider the follow-up period and the age at diagnosis, as some conditions—such as asthma in children—may resolve over time
CKD[c]		
eGFR[d]≥60 mL/min/1.73 m^2	≥3 mo	More likely indicative of early-stage CKD, given that the condition is traditionally regarded as progressive and irreversible
Depression		
PHQ[e]-9 score<5 (following acute treatment)	≥6 mo	Completion of the PHQ and attendance at the clinic are required; therefore, the absence of a recorded score may reflect factors such as nonattendance or administrative omissions rather than clinical recovery. Furthermore, a 6-mo follow-up period may be insufficient, as conditions such as depression can relapse and remit over the course of several years or even decades
Diabetes		
HbA_{1c}[f] <6.5% without the use of any antidiabetic medication	≥6 mo	No additional comment
Endometriosis		
Symptom (chronic pelvic pain and dysmenorrhea) resolution	≥6 mo	It is also important to consider pregnancy as a potential reason for the absence of diagnostic codes over a period exceeding 6 mo
Epilepsy		
No seizure activity, having no new antiepileptic drug attempts, and the absence of all seizure-related health care events (ie, seizure-related hospitalization or seizure-related GP or outpatient visit), QOF[g] data, and medical codes indicating a seizure	≥12 mo	No additional comments
Hypertension		
Normal BP[h] in the absence of antihypertensive drugs. Normal BP varies for age groups: (1) aged 18 - 59 y: BP <140/90 mm Hg, (2) aged 60 - 85 y with diabetes: BP <140/90 mm Hg, (3) aged 60 - 85 y without diabetes: BP <150/90 mm Hg	≥6 mo	Consider follow-up period, and age-related physiological changes can contribute to an increased risk of hypotension with older age
Acute heart failure		

XSL•FO
RenderX

Long-term condition and explanation of how remission is defined within studies	Period	Clinical considerations
N-terminal pro-b-type natriuretic peptide level persistently decreased above 1.5 times the baseline level	≥6 mo	No additional comments
Inflammatory bowel disease		
For ulcerative colitis: a score of ≤3 on the Harvey-Bradshaw Index or ≤4 on the Ulcerative Colitis Activity Index; endoscopic remission (Mayo Endoscopic Score of 0 or 1) and histologic remission (Simplified Geboes Score ≤0.2); ulcerative colitis endoscopic index of severity score=0 at the last follow-up and no use of medication	≥12 mo	No additional comments
For Crohn: Harvey-Bradshaw Index or ≤4; FCP[i] remission rate (FCP <250 µg/g); Crohn Disease Activity Index <150 points and no use of medication	__[j]	No additional comments
Chronic liver disease and alcoholic liver disease		
Absence of symptoms, decrease in serum glutamate oxaloacetate transaminase level to less than twice normal and other standard liver function (serum gamma globulin and bilirubin levels) to normal and disappearance from the liver biopsy of established features of disease activity	≥6 mo	No additional comments
Thyroid disease		
Relief from symptoms, and either of the following: thyroid-stimulating hormone levels in the reference range (also known as the "normal range": 1 - 2 MIU/L) OR free thyroxine (Free T4) levels in the reference range: 77.2 - 1543.1 nmol/L OR free triiodothyronine (Free T3) levels in the reference range: 5 - 7 pmol/L	≥12 mo	Preference for clinical biomarkers over relief of symptoms

[a]AUDIT-C: Alcohol Use Disorders Identification Test - Consumption.

[b]GP: general practitioner.

[c]CKD: chronic kidney disease.

[d]eGFR: estimated glomerular filtration rate.

[e]PHQ: Patient Health Questionnaire.

[f]HbA_{1c}: hemoglobin A_{1c}.

[g]QOF: Quality and Outcomes Framework

[h]BP: blood pressure.

[i]FCP: fecal calprotectin.

[j]Not applicable.

Discussion

Summary of Evidence

This scoping review synthesized evidence on how remission of LTCs has been defined and operationalized within EHRs. We identified 91 studies [13,27-117] covering 12 LTCs, including type 2 diabetes [13,62-69,86,87,102-104,114], depression [32-42,76,78,105,108], IBD [43-61,80-85,90-95,99-101,106,109-111,113,115,117], anemia [28], epilepsy [79], CKD [77], autoimmune pancreatitis [116], alcohol and/or drug misuse [70-75,107,112], heart disease [89], asthma [29-31], hypertension [103], and multiple sclerosis [96-98]. The review improved our understanding of how remission can be identified and operationalized in EHRs as criteria used to define and identify remission in studies included diagnostic (remission) codes, validated symptom scales, biochemical markers, absence of condition-specific events, and discontinuation of treatment [13,27-117]. However, criteria and follow-up periods differed widely across studies, even within the same condition (Figures 2-4), highlighting a lack of standardized definitions for remission. Using our scoping review findings and discussions with clinicians and data experts, we propose standardized definitions for remission of LTCs that will improve identification in EHRs (Table 1).

Overall, our review indicates that remission of LTCs remains relatively underresearched as many relevant studies identified in our searches did not clearly define remission (Figure 1). Most studies were conducted in the United States [28-76] or the United Kingdom [13,77-88] and focused on IBD

XSL•FO
RenderX

[43-61,80-85,90-95,99-101,106,109-111,113,115,117] and diabetes [13,62-69,86,87,102-104,114]. The lack of studies in some countries likely reflects differences in the extent of adoption of EHRs for clinical care, particularly for low-income countries where EHR adoption may be at its earlier stages [118]. These differences may also capture differences in clinician coding, with some studies indicating clinicians in the United States, for example, spend more time actively using the EHR [119]. Furthermore, we did not identify any studies on remission of some conditions known (and deemed by clinicians we consulted) to be amenable to remission, such as peptic ulcer disease, chronic urinary tract infection, chronic liver disease, and alcoholic liver disease, peripheral vascular disease, transient ischemic attack, stroke, tuberculosis, and thyroid disease (standardized definitions provided in Table 1). Our finding that the majority of studies described remission of certain conditions may reflect differences in the clinical focus of health conditions, research priorities, and policies across countries [118,120]. For example, in the United Kingdom, the Quality and Outcomes Framework—a pay-for-performance scheme—is associated with better recording and monitoring of certain long-term conditions (such as diabetes and asthma) in clinical practice [120].

Across studies, variation in remission criteria for the same condition may partly be explained by differences in health systems, as well as differences in the availability of remission information [118,121]. For example, settings with fragmented health systems may not reliably include criteria such as absence of hospitalization in their definition of asthma remission, whereas this would be possible for integrated systems with linked primary and secondary care data [121]. Diagnostic remission codes were used for only a few of the conditions, as may be expected due to a lack of remission codes for many conditions [88]. Remission codes, where they exist, may not be commonly used due to clinician coding burden and uncertainty [122,123]. Thus, it is important to include a range of criteria in the definition of remission to allow more accurate estimates of remission. Follow-up duration for remission also varied for studies, which may impact estimates of remission rates [124]. A better understanding of how remission rates vary over time (and any factors affecting remission over time) could improve understanding of observation periods required for remission definitions.

Comparison With Existing Literature

Previous reviews on remission have examined single conditions (eg, asthma [125,126], psoriasis [127], diabetes [128,129], rheumatoid arthritis [130], psychosis [131], and systemic lupus erythematosus [132]) and did not focus on EHRs. Our findings are consistent with previous studies [125-127] in that we observed varying definitions of remission of each LTC. Moreover, these previous studies identified that individuals with MLTCs were less likely to achieve remission of specific conditions [125,127]; however, the studies did not assess remission of comorbidities [125-127]. Our review offers an opportunity to advance understanding of remission of long-term conditions by providing consensus-driven definitions of remission of multiple LTCs [Table 1]. This will allow studies (including those focused on single conditions) to explore and consider the impact of remission in other comorbidities in a consistent, rigorous way that can allow cross-study comparisons. Unlike these previous reviews on remission, our study focuses on EHRs and presents practical approaches to define remission in real-world settings where barriers such as variation in coding, fragmented health care, and health system factors make it challenging to study remission [118-123]. Our review extends previous work on remission in MLTCs [88] by presenting a comprehensive approach to defining remission in EHRs beyond the use of remission and resolved codes to better identify remission and in turn improves population-based research and clinical outcomes of LTCs.

Implications for Research and Clinical Practice

Studies on LTCs using EHRs, including those assessing prevalence and trajectories, often do not account for remission of conditions [88], which may have implications for our understanding of MLTC [133-135]. This may be partly due to a lack of consensus for identifying remission, as identified by this review. We combined findings from our review with discussions with a clinical team, data experts, and a review of existing codes for remission in EHRs. This comprehensive approach can be used to advance research in this area and improve understanding of remission, particularly in the context of MLTC [136]. Future work may validate whether remission, as identified in EHRs, reflects true remission rates [137]. As with all EHR-based studies, there is bias due to incomplete data or missing information from subgroups [138]. Future work will need to consider missing data handling techniques to address this in order to ensure remission rates are not overestimated (eg, due to patients not consulting or being tested). Further research on remission may improve understanding of which conditions are more likely to achieve remission in patients living with MLTC, as well as factors and outcomes associated with remission. This understanding can improve the evidence base on remission and allow clinicians and patients to work toward remission, as well as improve care and outcomes of MLTC [139]. Furthermore, the range of criteria used to define remission of LTCs (Figure 2) highlights a need for a comprehensive approach to clinical coding to allow accurate identification and monitoring of remission. Further research is needed to assess the reliability of EHRs in capturing remission and to explore the wider factors associated with the documentation of remission in health records.

Strengths and Limitations

This study methodology [20] offers a comprehensive and systematic examination of how the remission of LTCs has been defined and operationalized within EHRs. A key strength lies in our integrative approach, which combined evidence from the literature with expert consultation and real-world coding frameworks. This methodological triangulation enhances the relevance and applicability of our findings [136], particularly in the context of MLTCs, where the concept of remission may inform future intervention design and policy development. However, several limitations should be acknowledged. First, although we conducted thorough electronic and gray literature searches, it is possible that some relevant studies, particularly those not using MeSH terms related to EHR, may not have been

XSL•FO
RenderX

captured. Additionally, studies involving pediatric populations, cancers, or LTCs outside our predefined list for MLTCs were excluded. This decision was informed by input from our public advisory group, who emphasized the need to focus on noncancer conditions due to the relative lack of understanding around their remission. Nonetheless, this narrowed scope limits the generalizability of our findings to other potentially resolvable conditions [140], and further research will be necessary to develop comparable frameworks for these excluded areas.

Second, for some LTCs, only a small number of eligible studies were identified (n≤1), resulting in limited evidence on which to base definitions of remission. As a result, our synthesis may reflect early or preliminary interpretations for these conditions rather than consensus-based definitions. Similarly, different study characteristics, as well as varying remission indicators and duration of follow-up, made it difficult to directly compare study findings [140]. There was a lack of studies comparing whether remission rates varied based on the choice of remission indicators. Finally, our operational framework for identifying remission was derived largely from coding structures available in the UK primary care and linked datasets. Conditions for which complete or long-term follow-up is less feasible in primary care were not included. As such, the applicability of our findings may vary in international contexts or within health care systems using different data architectures [140]. Future efforts should aim to validate and refine this framework across diverse settings and data environments to enhance its robustness and global utility.

Conclusions

This review demonstrates that the remission of multiple LTCs can be identified and operationalized using EHRs, although approaches to defining remission and the duration of follow-up varied considerably across studies. While several LTCs have been studied, many others with the potential for remission remain underrepresented in the literature. Our review is innovative, as it uses both evidence synthesis and consultation with clinical and data experts to propose practical approaches to define and implement remission in EHR-based research. The review offers standardized definitions for multiple LTCs. These findings lay the groundwork for improving the identification of remission, particularly in the context of MLTCs, where a better understanding of disease trajectories is critical. Due to variation in clinical coding and missing data (eg, due to lack of test results) [137,138], it is necessary to adopt a comprehensive approach and include all these indicators (as appropriate for the specified condition) to reliably capture remission and evaluate outcomes for patients. Future research may also report on how the choice of remission indicators (eg, diagnostic codes vs biochemical biomarkers) influences findings. Further research may apply, compare, and evaluate standardized definitions across different data sources to assess generalizability and further improve our understanding of the remission of LTCs.

Funding

This study was funded by the National Institute for Health and Care Research under its Research for Patient Benefit (grant NIHR206873). The views expressed are those of the authors and not necessarily those of the National Institute for Health and Care Research or the Department of Health and Social Care.

Data Availability

The datasets used or analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of 56 conditions on which our study was based and full details of included studies.
[PDF File, 606 KB - jmir_v28i1e80796_app1.pdf]

Multimedia Appendix 2
Search strategies.
[PDF File, 176 KB - jmir_v28i1e80796_app2.pdf]

Checklist 1
PRISMA-ScR fillable checklist.
[PDF File, 725 KB - jmir_v28i1e80796_app3.pdf]

References

1. Multimorbidity: technical series on safer primary care. World Health Organization. 2016. URL: https://iris.who.int/server/api/core/bitstreams/695cc0ce-75ea-4163-8bb5-3bc9733fa63f/content [accessed 2026-04-26]
2. Multimorbidity: clinical assessment and management. National Institute of Health and Care Excellence. 2016. URL: https://www.nice.org.uk/guidance/ng56 [accessed 2026-04-26]

XSL•FO
RenderX

3. Prior A, Vestergaard CH, Vedsted P, et al. Healthcare fragmentation, multimorbidity, potentially inappropriate medication, and mortality: a Danish nationwide cohort study. BMC Med 2023 Aug 15;21(1):305. [doi: 10.1186/s12916-023-03021-3] [Medline: 37580711]
4. Morris JE, Roderick PJ, Harris S, et al. Treatment burden for patients with multimorbidity: cross-sectional study with exploration of a single-item measure. Br J Gen Pract 2021 May;71(706):e381-e390. [doi: 10.3399/BJGP.2020.0883]
5. Gebreyohannes EA, Gebresillassie BM, Mulugeta F, Dessu E, Abebe TB. Treatment burden and health-related quality of life of patients with multimorbidity: a cross-sectional study. Qual Life Res 2023 Nov;32(11):3269-3277. [doi: 10.1007/s11136-023-03473-3] [Medline: 37405663]
6. Moffat K, Mercer SW. Challenges of managing people with multimorbidity in today's healthcare systems. BMC Fam Pract 2015 Dec;16(1):129. [doi: 10.1186/s12875-015-0344-4]
7. Rodrigues LP, de Oliveira Rezende AT, Delpino FM, et al. Association between multimorbidity and hospitalization in older adults: systematic review and meta-analysis. Age Ageing 2022 Jul 1;51(7):afac155. [doi: 10.1093/ageing/afac155]
8. Islam N, Shabnam S, Khan N, et al. Combinations of multiple long term conditions and risk of hospital admission or death during winter 2021-22 in England: population based cohort study. BMJ Med 2024 Oct;3(1):e001016. [doi: 10.1136/bmjmed-2024-001016]
9. Hanlon P, Jani BD, Nicholl B, Lewsey J, McAllister DA, Mair FS. Associations between multimorbidity and adverse health outcomes in UK Biobank and the SAIL Databank: a comparison of longitudinal cohort studies. PLoS Med 2022;19(3):e1003931. [doi: 10.1371/journal.pmed.1003931]
10. Dodds RM, Bunn JG, Hillman SJ, et al. Simple approaches to characterising multiple long-term conditions (multimorbidity) and rates of emergency hospital admission: findings from 495,465 UK Biobank participants. J Intern Med 2023 Jan;293(1):100-109. [doi: 10.1111/joim.13567] [Medline: 36131375]
11. Stokes J, Bower P, Smith SM, et al. A primary care research agenda for multiple long-term conditions: a Delphi study. Br J Gen Pract 2024 Apr;74(741):e258-e263. [doi: 10.3399/BJGP.2023.0163] [Medline: 38164536]
12. Gregg EW, Chen H, Bancks MP, et al. Impact of remission from type 2 diabetes on long-term health outcomes: findings from the Look AHEAD study. Diabetologia 2024 Mar;67(3):459-469. [doi: 10.1007/s00125-023-06048-6] [Medline: 38233592]
13. Dambha-Miller H, Hounkpatin HO, Stuart B, Farmer A, Griffin S. Type 2 diabetes remission trajectories and variation in risk of diabetes complications: a population-based cohort study. PLoS One 2023;18(8):e0290791. [doi: 10.1371/journal.pone.0290791] [Medline: 37643199]
14. Stenberg E, Marsk R, Sundbom M, et al. Remission, relapse, and risk of major cardiovascular events after metabolic surgery in persons with hypertension: a Swedish nationwide registry-based cohort study. PLoS Med 2021 Nov;18(11):e1003817. [doi: 10.1371/journal.pmed.1003817] [Medline: 34723954]
15. Dennehy EB, Robinson RL, Stephenson JJ, et al. Impact of non-remission of depression on costs and resource utilization: from the COmorbidities and symptoms of DEpression (CODE) study. Curr Med Res Opin 2015 Jun;31(6):1165-1177. [doi: 10.1185/03007995.2015.1029893] [Medline: 25879140]
16. Fillies B, Stapel B, Lemke LH, et al. Remission from depression is associated with improved quality of life and preserved exercise capacity in adults with congenital heart disease. Front Cardiovasc Med 2024;11:1418342. [doi: 10.3389/fcvm.2024.1418342] [Medline: 39022619]
17. Butler JM, Gibson B, Lewis L, et al. Patient-centered care and the electronic health record: exploring functionality and gaps. JAMIA Open 2020 Oct;3(3):360-368. [doi: 10.1093/jamiaopen/ooaa044] [Medline: 33215071]
18. Manca DP. Do electronic medical records improve quality of care? Yes. Can Fam Physician 2015 Oct;61(10):846-847. [Medline: 26472786]
19. Georgiev K, McPeake J, Shenkin SD, et al. Understanding hospital activity and outcomes for people with multimorbidity using electronic health records. Sci Rep 2025;15(1):8522. [doi: 10.1038/s41598-025-92940-7]
20. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
21. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
22. Ho ISS, Azcoaga-Lorenzo A, Akbari A, et al. Measuring multimorbidity in research: Delphi consensus study. BMJ Med 2022;1(1):e000247. [doi: 10.1136/bmjmed-2022-000247] [Medline: 36936594]
23. Dambha-Miller H, Farmer A, Nirantharakumar K, et al. Artificial intelligence for multiple long-term conditions (AIM): a consensus statement from the NIHR AIM consortia. NIHR Open Research 2023. [doi: 10.3310/nihropenres.1115210.1]
24. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
25. Chowdhury SR, Chandra Das D, Sunna TC, Beyene J, Hossain A. Global and regional prevalence of multimorbidity in the adult population in community settings: a systematic review and meta-analysis. eClinicalMedicine 2023 Mar;57:101860. [doi: 10.1016/j.eclinm.2023.101860]

XSL•FO
RenderX

26. Vena JA, Copel LC. A concept analysis of remission within the domain of cancer survivorship. Nurs Forum 2021 Apr;56(2):421-428. [doi: 10.1111/nuf.12558] [Medline: 33634534]
27. Wolf A, Dedman D, Campbell J, et al. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. Int J Epidemiol 2019 Dec 1;48(6):1740-1740g. [doi: 10.1093/ije/dyz034] [Medline: 30859197]
28. Tang WHW, Tong W, Jain A, Francis GS, Harris CM, Young JB. Evaluation and long-term prognosis of new-onset, transient, and persistent anemia in ambulatory patients with chronic heart failure. J Am Coll Cardiol 2008 Feb 5;51(5):569-576. [doi: 10.1016/j.jacc.2007.07.094] [Medline: 18237687]
29. Howarth P, Lim K, Corbridge T, et al. Real-world assessment of clinical remission in asthma with biologics. Presented at: ERS Congress 2024; Sep 7-11, 2024. [doi: 10.1183/13993003.congress-2024.PA2182]
30. Carstens D, Ojeranti D, Chung Y, Chan KA, Dhopeshwarkar N. A real-world analysis of key asthma remission components in patients treated with benralizumab: the ZEPHYR-4 study. J Allergy Clin Immunol 2024 Feb;153(2):AB99. [doi: 10.1016/j.jaci.2023.11.325]
31. Lim KG, Corbridge T, Lee L, et al. Evaluation of clinical remission indicators in asthma patients treated with biologics in a real-world setting [abstract]. Am J Respir Crit Care Med 2024 May;209:A1368. [doi: 10.1164/ajrccm-conference.2024.209.1_MeetingAbstracts.A1368]
32. Angstman KB, Marcelin A, Gonzalez CA, Kaufman TK, Maxson JA, Williams MD. The impact of posttraumatic stress disorder on the 6-month outcomes in collaborative care management for depression. J Prim Care Community Health 2016 Jul;7(3):159-164. [doi: 10.1177/2150131916638329] [Medline: 26994060]
33. Sacks RM, Greene J, Hibbard JH, Overton V. How well do patient activation scores predict depression outcomes one year later? J Affect Disord 2014 Dec;169:1-6. [doi: 10.1016/j.jad.2014.07.030] [Medline: 25128858]
34. Pfoh ER, Mojtabai R, Bailey J, Weiner JP, Dy SM. Conformance to depression process measures of medicare part B beneficiaries in primary care settings. J Am Geriatr Soc 2015 Jul;63(7):1338-1345. [doi: 10.1111/jgs.13483] [Medline: 26114978]
35. Eguchi D, Matos J, Korenblatt J, Zwerling J, Kiosses D, Ceide M. The effectiveness of problem adaptation therapy (PATH) in a culture and language rich community. Am J Geriatr Psychiatry 2019 Mar;27(3):S190-S191. [doi: 10.1016/j.jagp.2019.01.104]
36. Coley RY, Boggs JM, Beck A, Hartzler AL, Simon GE. Defining success in measurement-based care for depression: a comparison of common metrics. PS 2020 Apr 1;71(4):312-318. [doi: 10.1176/appi.ps.201900295]
37. Li KJ, Slama NE, Hirschtritt ME, Anshu P, Iturralde E. Electroconvulsive therapy anesthetic choice and clinical outcomes. J ECT 2023 Jun 1;39(2):102-105. [doi: 10.1097/YCT.0000000000000895] [Medline: 36729716]
38. Liebmann EP, Resnick SG, Hoff RA, Katz IR. Outcomes for treatment of depression in the Veterans Health Administration: rates of response and remission calculated from clinical and survey data. Psychiatry Res 2023 Jun;324:115196. [doi: 10.1016/j.psychres.2023.115196]
39. Whiting A, Poolman AE, Misra A, Gordon JE, Angstman KB. Comparison of ambulatory quality measures between shared practice panels and independent practice panels. Mayo Clin Proc Innov Qual Outcomes 2023 Aug;7(4):256-261. [doi: 10.1016/j.mayocpiqo.2023.05.005] [Medline: 37388418]
40. Deng A, Espiridion ED. Multipronged electronic health record analysis of antidepressant effectiveness on depression remission in patients with concurrent depression and irritable bowel syndrome. Cureus 2024;16(7):e64968. [doi: 10.7759/cureus.64968]
41. Pfeiffer PN, Geller J, Ganoczy D, et al. Clinical outcomes of intravenous ketamine treatment for depression in the VA health system. J Clin Psychiatry 2024 Jan 8;85(1):23m14984. [doi: 10.4088/JCP.23m14984] [Medline: 38206011]
42. Benster LL, Weissman CR, Suprani F, et al. Predictive modeling of response to repetitive transcranial magnetic stimulation in treatment-resistant depression. Transl Psychiatry 2025;15(1):160. [doi: 10.1038/s41398-025-03380-w]
43. Blonski W, Buchner A, Weiner M, et al. Clinical remission is predictive of endoscopic remission and clinical disease activity is predictive of endoscopic disease activity in patients with ulcerative colitis at the time of surveillance colonoscopy. Am J Gastroenterol 2011 Oct;106:S493-S494. [doi: 10.14309/00000434-201110002-01297]
44. Kane SV, Becker BD, Neis B, et al. Su1208 Effectiveness of split-dose certolizumab pegol for the treatment of Crohn's disease: experience at a tertiary care center. Gastroenterology 2013 May;144(5):S-427. [doi: 10.1016/S0016-5085(13)61575-9]
45. Colman R, Rubin D. Efficacy of high-dose versus low-dose methotrexate in combination therapy for inflammatory bowel disease. Am J Gastroenterol 2014 Oct;109:S489-S490. [doi: 10.14309/00000434-201410002-01649]
46. Merkley SA, Beaulieu DB, Horst S, et al. Use of intravenous immunoglobulin for patients with inflammatory bowel disease with contraindications or who are unresponsive to conventional treatments. Inflamm Bowel Dis 2015 Aug;21(8):1854-1859. [doi: 10.1097/MIB.0000000000000456] [Medline: 25993689]
47. Ahsan M, Koutroumpakis F, Rivers CR, et al. S0890 Can IBD medications counteract the negative effects of a deleterious, high sugar diet? Am J Gastroenterol 2020;115(1):S459-S459. [doi: 10.14309/01.ajg.0000705608.97193.ba]
48. Vachon A, Oonk A, Fennimore B, Gerich ME, Scott FI. Sa1813 The impact of inflammatory bowel disease therapy on rates and severity of depression. Gastroenterology 2020 May;158(6):S-435 [FREE Full text]
49. Fernandez-Cano MC, Librero-Jimenez M, Lopez-Vico M, Cabello-Tapia MJ, Martin-Rodriguez MM, Sanchez-Capilla AD. Vedolizumab in real practice: experience in a tertiary hospital. UEG J 2021;9:512. [doi: 10.1002/ueg2.12144]

XSL•FO
RenderX

50. Lee T, Kamm MA, Bell S, et al. Long‐term outcomes of perianal fistulizing Crohn's disease in the biologic era. JGH Open 2021 Feb;5(2):235-241. [doi: 10.1002/jgh3.12475]
51. Cohen NA, Steinberg JM, Silfen A, et al. Endo-histologic normalization is achievable with tofacitinib and is associated with improved clinical outcomes. Dig Dis Sci 2023 Apr;68(4):1464-1472. [doi: 10.1007/s10620-022-07716-0]
52. Chugh R. Real world outcomes for upadacitinib in Crohn's disease patients at a single center. Gastroenterology 2022;162:S990 [FREE Full text]
53. Dalal R, Sharma P, Bains K, Pruce J, Allegretti JR. S911 Real world outcomes of tofacitinib for ulcerative colitis at 52 and 78 weeks. Am J Gastroenterol 2022;117(10S):e660-e661. [doi: 10.14309/01.ajg.0000860284.16240.9f]
54. Dalal RS, Pruce JC, Allegretti JR. Mo1546: One and two-year outcomes after ustekinumab dose intensification for inflammatory bowel diseases. Gastroenterology 2022 May;162(7):S-812. [doi: 10.1016/S0016-5085(22)61918-8]
55. Zeina T, Gandhi S, Mittal A, et al. Predictors and etiologies of clinical relapse among patients with ulcerative colitis in deep remission. J Clin Gastroenterol 2024 Feb 1;58(2):195-199. [doi: 10.1097/MCG.0000000000001834] [Medline: 36753459]
56. Anyanwu PA, Van Anglen LJ, Chou C, Sanchirico M, Ritter TE. S1231 Real-world effectiveness and onset of action of vedolizumab as a first-line biologic in biologic-naïve patients with ulcerative colitis. Am J Gastroenterol 2024;119(10S):S876-S877. [doi: 10.14309/01.ajg.0001034292.46400.5e]
57. Chugh R, Kattah M, Mahadevan U. Mo1866 Real world efficacy of risankizumab in Crohn's disease patients with prior exposure to ustekinumab. Gastroenterology 2024 May;166(5):S-1151. [doi: 10.1016/S0016-5085(24)03105-6]
58. Dalal RS, Kallumkal G, Cabral HJ, Barnes EL, Allegretti JR. One-year comparative effectiveness of upadacitinib vs tofacitinib for ulcerative colitis: a multicenter cohort study. Am J Gastroenterol 2024;119(8):1628-1631. [doi: 10.14309/ajg.0000000000002746]
59. Lieto S, Scalzo N, Huang K, et al. Tu1889 Real-world effectiveness and safety of ozanimod in ulcerative colitis: a multicenter study. Gastroenterology 2024 May;166(5):S-1455. [doi: 10.1016/S0016-5085(24)03790-9]
60. Scalzo N, Lieto S, Huang K(, et al. S1288 Real-world effectiveness of ozanimod for ulcerative colitis in patients with prior advanced therapy exposure: a multicenter study. Am J Gastroenterol 2024;119(10S):S921-S921. [doi: 10.14309/01.ajg.0001034520.61834.16]
61. Gold S, Freid H, Savin Shalom E, Bansal MB, Dubinsky M, Spencer EA. Glucagon-like peptde-1 agonists in inflammatory bowel disease lead to significant weight loss and improved clinical remission with minimal side effects. Gastroenterology 2024:169 [FREE Full text]
62. Arterburn D, Bogart A, Coleman KJ, et al. Comparative effectiveness of bariatric surgery vs. nonsurgical treatment of type 2 diabetes among severely obese adults. Obes Res Clin Pract 2013;7(4):e258-e268. [doi: 10.1016/j.orcp.2012.08.196] [Medline: 24306153]
63. Still CD, Wood GC, Benotti P, et al. Preoperative prediction of type 2 diabetes remission after Roux-en-Y gastric bypass surgery: a retrospective cohort study. Lancet Diabetes Endocrinol 2014 Jan;2(1):38-45. [doi: 10.1016/S2213-8587(13)70070-6] [Medline: 24579062]
64. Wood GC, Gerhard GS, Benotti P, et al. Preoperative use of incretins is associated with increased diabetes remission after RYGB surgery among patients taking insulin: a retrospective cohort analysis. Ann Surg 2015 Jan;261(1):125-128. [doi: 10.1097/SLA.0000000000000588] [Medline: 24646545]
65. Zaman JA, Shah N, Leverson GE, Greenberg JA, Funk LM. The effects of optimal perioperative glucose control on morbidly obese patients undergoing bariatric surgery. Surg Endosc 2017 Mar;31(3):1407-1413. [doi: 10.1007/s00464-016-5129-x] [Medline: 27450209]
66. Chang LS, Malmasi S, Hosomura N, et al. Metabolic surgery counseling, diabetes remission, and A1C control in patients with diabetes and obesity. Diabetes 2018 Jul 1;67(Supplement_1):2101-210P. [doi: 10.2337/db18-2101-P]
67. Still CD, Benotti P, Mirshahi T, Cook A, Wood GC. DiaRem2: incorporating duration of diabetes to improve prediction of diabetes remission after metabolic surgery. Surg Obes Relat Dis 2019 May;15(5):717-724. [doi: 10.1016/j.soard.2018.12.020]
68. Ghusn W, Ikemiya K, Al Annan K, et al. Diabetes mellitus remission in patients with BMI > 50 kg/m2 after bariatric surgeries: a real-world multi-centered study. OBES SURG 2023 Jun;33(6):1838-1845. [doi: 10.1007/s11695-023-06622-2]
69. Thapa B, Schmittdiel JA, Arterburn D, et al. Clinical and demographic characteristics associated with diabetes remission in six integrated health care systems: a retrospective cohort study. Diabetes Care 2025 Oct 1;48(10):1737-1743. [doi: 10.2337/dc25-0530] [Medline: 40734551]
70. Williams EC, Lapham GT, Bobb JF, et al. Documented brief intervention not associated with resolution of unhealthy alcohol use one year later among VA patients living with HIV. J Subst Abuse Treat 2017 Jul;78:8-14. [doi: 10.1016/j.jsat.2017.04.006]
71. Palzes VA, Kline-Simon AH, Satre DD, Sterling S, Weisner C, Chi FW. Remission from unhealthy drinking among patients with an alcohol use disorder: a longitudinal study using systematic, primary care-based alcohol screening data. J Stud Alcohol Drugs 2020 Jul;81(4):436-445. [doi: 10.15288/jsad.2020.81.436] [Medline: 32800079]
72. Zhou M, Wang Q, Zheng C, John Rush A, Volkow ND, Xu R. Drug repurposing for opioid use disorders: integration of computational prediction, clinical corroboration, and mechanism of action analyses. Mol Psychiatry 2021 Sep;26(9):5286-5296. [doi: 10.1038/s41380-020-01011-y]

XSL•FO
RenderX

73. Osterhage KP, Hser YI, Mooney LJ, et al. Identifying patients with opioid use disorder using International Classification of Diseases (ICD) codes: challenges and opportunities. Addiction 2024 Jan;119(1):160-168. [doi: 10.1111/add.16338] [Medline: 37715369]
74. Gao Z, Winhusen TJ, Gorenflo MP, et al. Artificial intelligence‐based drug repurposing with electronic health record clinical corroboration: a case for ketamine as a potential treatment for amphetamine‐type stimulant use disorder. Addiction 2025 Apr;120(4):732-744. [doi: 10.1111/add.16715]
75. Hailer A, Terebuh P, Xu R, Kaelber DC, Davis PB. 25 The association of a documented prescription of medication for opioid use disorder (MOUD) during pregnancy with maternal outcomes. J Clin Trans Sci 2025 Apr;9(s1):9-9. [doi: 10.1017/cts.2024.716]
76. Keins S, Abramson JR, Mallick A, et al. Association of depression onset and treatment with blood pressure control after intracerebral hemorrhage. Stroke 2023 Jan;54(1):105-112. [doi: 10.1161/STROKEAHA.122.040331]
77. Taal MW, Lucas B, Mcintyre N, Mcintyre C, Fluck R. #1321 Long term outcomes in people with fluctuating criteria for the diagnosis of chronic kidney disease. Nephrol Dial Transplant 2024 May 23;39(Supplement_1):gfae069-gfa0627. [doi: 10.1093/ndt/gfae069.627]
78. De Giorgi R, De Crescenzo F, Cowen PJ, Harmer CJ, Cipriani A. Real-world outcomes of concomitant antidepressant and statin use in primary care patients with depression: a population-based cohort study. BMC Med 2023 Nov 7;21(1):424. [doi: 10.1186/s12916-023-03138-5] [Medline: 37936200]
79. Powell G, Logan J, Kiri V, Borghs S. Trends in antiepileptic drug treatment and effectiveness in clinical practice in England from 2003 to 2016: a retrospective cohort study using electronic medical records. BMJ Open 2019 Dec 16;9(12):e032551. [doi: 10.1136/bmjopen-2019-032551] [Medline: 31848168]
80. Bhandare AP, Crooks B, Nigam GB, Limdi JK. P337 Switching from originator infliximab to CT-P13: single-centre experience from the UK. J Crohns Colitis 2019 Jan 25;13(Supplement_1):S270-S271. [doi: 10.1093/ecco-jcc/jjy222.461]
81. Ibraheim H, Samaan MA, Srinivasan A, et al. Effectiveness and safety of vedolizumab in inflammatory bowel disease patients aged 60 and over: an observational multicenter UK experience. Ann Gastroenterol 2020;33(2):170-177. [doi: 10.20524/aog.2020.0447] [Medline: 32127738]
82. Brownson E, Saunders J, Gerasimidis K, Karafoulidou Z, Seenan JP, Macdonald J. P504 Prospective analysis of micronutrient status and disease course in inflammatory bowel disease. J Crohns Colitis 2022 Jan 21;16(Supplement_1):i465-i466. [doi: 10.1093/ecco-jcc/jjab232.631]
83. Meade S, Routledge E, Sharma E, et al. How achievable are STRIDE-II treatment targets in real-world practice and do they predict long-term treatment outcomes? Frontline Gastroenterol 2023;14(4):312-318. [doi: 10.1136/flgastro-2022-102309] [Medline: 37409343]
84. Gros B, Ross H, Nwabueze M, et al. Long-term outcomes and predictors of vedolizumab persistence in ulcerative colitis. Ther Adv Gastroenterol 2024;17:17562848241258372. [doi: 10.1177/17562848241258372] [Medline: 39086990]
85. Colwill M, Padley J, Qazi U, et al. Risankizumab is effective following ustekinumab failure in Crohn's disease: a real-world study from a tertiary center. World J Gastrointest Pharmacol Ther 2025 Dec 5;16(4):110273. [doi: 10.4292/wjgpt.v16.i4.110273] [Medline: 41378066]
86. Gulliford MC, Booth HP, Reddy M, et al. Effect of contemporary bariatric surgical procedures on type 2 diabetes remission. a population-based matched cohort study. Obes Surg 2016 Oct;26(10):2308-2315. [doi: 10.1007/s11695-016-2103-6] [Medline: 26922184]
87. Griffiths BP, Mehar S, Blangiardo M, et al. Evaluating the effectiveness of REWIND (reducing weight with intensive dietary support) on type 2 diabetes remission. SSRN. URL: https://ssrn.com/abstract=5169449 [accessed 2026-04-28]
88. Ledwaba-Chapman L, Bisquera A, Gulliford M, et al. Applying resolved and remission codes reduced prevalence of multimorbidity in an urban multi-ethnic population. J Clin Epidemiol 2021 Dec;140:135-148. [doi: 10.1016/j.jclinepi.2021.09.005] [Medline: 34517101]
89. Liu Y, Zhang Y, Chen H, et al. Efficacy of tolvaptan on the short and mid-term prognosis in elderly patients with acute heart failure coexisting with oliguria: a retrospective cohort study. Front Cardiovasc Med 2023;9:1075631. [doi: 10.3389/fcvm.2022.1075631]
90. Xie Y, Zhuang T, Ping Y, et al. Elevated systemic immune inflammation index level is associated with disease activity in ulcerative colitis patients. Clin Chim Acta 2021 Jun;517:122-126. [doi: 10.1016/j.cca.2021.02.016] [Medline: 33662359]
91. Xie C, Qian W, Shang L, et al. Assessment of body composition-related imaging parameters indicative of sarcopenia in Chinese patients with Crohn's disease: correlation with disease severity and biologic efficacy. Am J Transl Res 2024;16(10):5427-5440. [doi: 10.62347/ZPZR8134] [Medline: 39544770]
92. Tu T, Zhuang X, Qiu Y, et al. Histological remission as a reliable predictor of clinical outcomes in patients with ulcerative colitis after conventional therapy: a 5-year analysis. BMC Gastroenterol 2025;25(1):771. [doi: 10.1186/s12876-025-04357-1]
93. McGrory L, Lynch K. An observational study of methotrexate in inflammatory bowel disease: a small but useful player. J Gastroenterology Hepatol 2022 Sep;37(S1):115-174. [doi: 10.1111/jgh.15953]
94. Alshiwanna B, Mcnamara J, Wilson W, et al. P1005 Dose escalated infliximab in inflammatory bowel disease - Crohn's colitis cure (CCC) data insights program. J Crohns Colitis 2024 Jan 24;18(Supplement_1):i1809-i1809. [doi: 10.1093/ecco-jcc/jjad212.1135]

XSL•FO
RenderX

95. Chu MKW, Day AS, Bogatic D, et al. Frequent dietetic involvement enhances adherence and clinical outcomes of exclusive enteral nutrition in adults with Crohn's disease. JGH Open 2025 Jul;9(7):e70207. [doi: 10.1002/jgh3.70207]
96. Zafar A, AlShamrani FJG. No evidence of disease activity-3 (NEDA-3) status in patients with relapsing remitting multiple sclerosis: evidence from Saudi cohort receiving mainly interferon. Mult Scler Relat Disord 2021 Jun;51:102875. [doi: 10.1016/j.msard.2021.102875]
97. Alshamrani FJG, Zafar A, Alsawad RM, et al. Achievement of no evidence of disease activity-3 with oral disease-modifying treatment in patients with relapsing-remitting multiple sclerosis. Saudi J Med Med Sci 2024;12(4):299-305. [doi: 10.4103/sjmms.sjmms_148_24] [Medline: 39539793]
98. Alharbi MA, Aldosari F, Althobaiti AH, et al. Clinical and economic evaluations of natalizumab, rituximab, and ocrelizumab for the management of relapsing-remitting multiple sclerosis in Saudi Arabia. BMC Health Serv Res 2023 May 26;23(1):552. [doi: 10.1186/s12913-023-09462-z] [Medline: 37237257]
99. Mulinacci G, Pirola L, Piazza O Sed N, et al. P1058 Safety and effectiveness of subcutaneous infliximab formulation in difficult-to-treat Crohn's disease patients with previous exposure to intravenous infliximab: a case-series. J Crohns Colitis 2024 Jan 24;18(Supplement_1):i1895-i1896. [doi: 10.1093/ecco-jcc/jjad212.1188]
100. Parigi TL, Massimino L, Carini A, et al. P710 Prevalence, characteristics and management of patients with difficult-to-treat inflammatory bowel disease: a cross-sectional study from a tertiary referral center. J Crohns Colitis 2024 Jan 24;18(Supplement_1):i1329-i1330. [doi: 10.1093/ecco-jcc/jjad212.0840]
101. Kazama T, Ando K, Ueno N, et al. Long-term effectiveness and safety of infliximab-biosimilar: a multicenter Phoenix retrospective cohort study. PLoS ONE 2023;18(9):e0288393. [doi: 10.1371/journal.pone.0288393]
102. Seki Y, Kasama K, Yokoyama R, et al. Bariatric surgery versus medical treatment in mildly obese patients with type 2 diabetes mellitus in Japan: propensity score-matched analysis on real-world data. J Diabetes Investig 2022 Jan;13(1):74-84. [doi: 10.1111/jdi.13631] [Medline: 34265175]
103. Sheikh L, Pearless LA, Booth MW. Laparoscopic Silastic Ring Mini-Gastric Bypass (SR-MGBP): up to 11-year results from a single centre. Obes Surg 2017 Sep;27(9):2229-2234. [doi: 10.1007/s11695-017-2659-9] [Medline: 28378207]
104. Witcomb Cahill H, McGuinness M, Struthers J, Bissett I, Harmston C. Provision and outcomes of publicly funded bariatric surgery in a metropolitan versus a provincial population of New Zealand. ANZ J Surg 2024 Oct;94(10):1747-1751. [doi: 10.1111/ans.19206] [Medline: 39350677]
105. Sicras-Mainar A, Navarro-Artieda R, Blanca-Tamayo M, Gimeno-de la Fuente V, Salvatella-Pasant J. Comparison of escitalopram vs. citalopram and venlafaxine in the treatment of major depression in Spain: clinical and economic consequences. Curr Med Res Opin 2010 Dec;26(12):2757-2764. [doi: 10.1185/03007995.2010.529430] [Medline: 21034375]
106. Molina Arriero G, González Furelos T, Mauriz Barreiro V, Echarri Piudo A. P531 Long-term evaluation of intensified ustekinumab in Crohn's disease: real-life results. J Crohns Colitis 2024 Jan 24;18(Supplement_1):i1039-i1039. [doi: 10.1093/ecco-jcc/jjad212.0661]
107. Danısman M, Aydin SM, Ispir GZ, Katar KS, Kurtoglu MB. Remission rates of patients with opioid use disorder who got naltrexone implantation: a retrospective study. HARCP Archives ;26:1-6. [doi: 10.62401/2531-4122-2024-54]
108. Gurel SC, Mutlu E, Başar K, Yazıcı MK. Bi-temporal electroconvulsive therapy efficacy in bipolar and unipolar depression: a retrospective comparison. Asian J Psychiatr 2021 Jan;55:102503. [doi: 10.1016/j.ajp.2020.102503] [Medline: 33296865]
109. Pokryszka J, Stadlmann M, Primas C, et al. P785 A retrospective cohort study on vedolizumab trough levels in Crohn's disease patients. J Crohns Colitis 2023 Jan 30;17(Supplement_1):i919-i919. [doi: 10.1093/ecco-jcc/jjac190.0915]
110. Jorissen C, Verstockt B, Schils N, Sabino J, Ferrante M, Vermeire S. Long-term clinical outcome after thiopurine discontinuation in elderly IBD patients. Scand J Gastroenterol 2021 Nov;56(11):1323-1327. [doi: 10.1080/00365521.2021.1965207] [Medline: 34399630]
111. Saleem R, Dubois P, Chung-Faye G, Bjarnason I, Hayee B. P482 Faecal calprotectin can be used to monitor response to therapy in anti-TNF treated patients with Crohn's disease. Journal of Crohn's and Colitis 2014 Feb;8:S268. [doi: 10.1016/S1873-9946(14)60602-4]
112. Rautiainen E, Ryynänen OP, Reissell E, Kauhanen J, Laatikainen T. Alcohol-related social and health service use patterns as predictors of death and remission in patients with AUD. J Subst Abuse Treat 2019 Jan;96:65-74. [doi: 10.1016/j.jsat.2018.10.013]
113. Shehab M, Alfadhli A, Abdullah I, Alostad W, Marei A, Alrashed F. Effectiveness of biologic therapies in achieving treatment targets in inflammatory bowel disease; real-world data from the Middle East (ENROLL study). Front Pharmacol 2024;15:1388043. [doi: 10.3389/fphar.2024.1388043]
114. Ahmadi MA, Ojla MD, Iqbal M, et al. Investigating the role of metabolic bariatric surgery in achieving remission of type 2 diabetes mellitus and evaluating the effects on glycemic control and metabolic health. Cureus 2025;17(6):e86539. [doi: 10.7759/cureus.86539]
115. Ye BD, Cheon JH, Song KH, et al. The real-world outcomes of vedolizumab in patients with ulcerative colitis in Korea: a multicenter retrospective study. Therap Adv Gastroenterol 2021 Jan;14:17562848211024769. [doi: 10.1177/17562848211024769]

XSL•FO
RenderX

116. Nikolic S, Panic N, Hintikka ES, et al. Efficacy and safety of rituximab in autoimmune pancreatitis type 1: our experiences and systematic review of the literature. Scand J Gastroenterol 2021 Nov 2;56(11):1355-1362. [doi: 10.1080/00365521.2021.1963837]
117. Han SJ, Park JJ. Analysis of prognostic factors on earlier clinical remission and long-term use for patients with Crohn's disease under vedolizumab. Gut Liver 2022;16:162.
118. Woldemariam MT, Jimma W. Adoption of electronic health record systems to enhance the quality of healthcare in low-income countries: a systematic review. BMJ Health Care Inform 2023 Jun;30(1):e100704. [doi: 10.1136/bmjhci-2022-100704] [Medline: 37308185]
119. Holmgren AJ, Downing NL, Bates DW, et al. Assessment of electronic health record use between US and non-US health systems. JAMA Intern Med 2021 Feb 1;181(2):251. [doi: 10.1001/jamainternmed.2020.7071]
120. Forbes LJ, Marchand C, Doran T, Peckham S. The role of the Quality and Outcomes Framework in the care of long-term conditions: a systematic review. Br J Gen Pract 2017 Nov;67(664):e775-e784. [doi: 10.3399/bjgp17X693077] [Medline: 28947621]
121. Goh KH, Yeow AYK, Wang L, et al. The benefits of integrating electronic medical record systems between primary and specialist care institutions: mixed methods cohort study. J Med Internet Res 2025;27:e49363. [doi: 10.2196/49363]
122. Cho H, Nguyen OT, Weaver M, et al. Electronic health record system use and documentation burden of acute and critical care nurse clinicians: a mixed-methods study. J Am Med Inform Assoc 2024 Nov 1;31(11):2540-2549. [doi: 10.1093/jamia/ocae239]
123. Davies A, Ahmed H, Thomas-Wood T, Wood F. Primary healthcare professionals' approach to clinical coding: a qualitative interview study in Wales. Br J Gen Pract 2025 Jan;75(750):e43-e49. [doi: 10.3399/BJGP.2024.0036]
124. Beaney T, Clarke J, Woodcock T, Majeed A, Barahona M, Aylin P. Effect of timeframes to define long term conditions and sociodemographic factors on prevalence of multimorbidity using disease code frequency in primary care electronic health records: retrospective study. BMJ Med 2024;3(1):e000474. [doi: 10.1136/bmjmed-2022-000474] [Medline: 38361663]
125. Shackleford A, Heaney LG, Redmond C, McDowell PJ, Busby J. Clinical remission attainment, definitions, and correlates among patients with severe asthma treated with biologics: a systematic review and meta-analysis. Lancet Respir Med 2025 Jan;13(1):23-34. [doi: 10.1016/S2213-2600(24)00293-5] [Medline: 39549709]
126. Mailhot-Larouche S, Celis-Preciado C, Heaney LG, Couillard S. Identifying super-responders: a review of the road to asthma remission. Ann Allergy Asthma Immunol 2025 Jan;134(1):31-45. [doi: 10.1016/j.anai.2024.09.023] [Medline: 39383944]
127. Balak DMW, Perez‐Chada LM, Guo LN, et al. Definitions of remission in psoriasis: a systematic literature review from the National Psoriasis Foundation. Acad Dermatol Venereol 2022 Dec;36(12):2291-2300. [doi: 10.1111/jdv.18477]
128. Kanbour S, Ageeb RA, Malik RA, Abu-Raddad LJ. Impact of bodyweight loss on type 2 diabetes remission: a systematic review and meta-regression analysis of randomised controlled trials. Lancet Diabetes Endocrinol 2025 Apr;13(4):294-306. [doi: 10.1016/S2213-8587(24)00346-2]
129. Goldenberg JZ, Day A, Brinkworth GD, et al. Efficacy and safety of low and very low carbohydrate diets for type 2 diabetes remission: systematic review and meta-analysis of published and unpublished randomized trial data. BMJ 2021;372:m4743. [doi: 10.1136/bmj.m4743]
130. Yu C, Jin S, Wang Y, et al. Remission rate and predictors of remission in patients with rheumatoid arthritis under treat-to-target strategy in real-world studies: a systematic review and meta-analysis. Clin Rheumatol 2019 Mar;38(3):727-738. [doi: 10.1007/s10067-018-4340-7]
131. Catalan A, Richter A, Salazar de Pablo G, et al. Proportion and predictors of remission and recovery in first-episode psychosis: systematic review and meta-analysis. Eur Psychiatry 2021 Nov 3;64(1):e69. [doi: 10.1192/j.eurpsy.2021.2246] [Medline: 34730080]
132. Ugarte-Gil MF, Mendoza-Pinto C, Reátegui-Sokolova C, et al. Achieving remission or low disease activity is associated with better outcomes in patients with systemic lupus erythematosus: a systematic literature review. Lupus Sci Med 2021 Sep;8(1):e000542. [doi: 10.1136/lupus-2021-000542]
133. Bisquera A, Turner EB, Ledwaba-Chapman L, et al. Inequalities in developing multimorbidity over time: a population-based cohort study from an urban, multi-ethnic borough in the United Kingdom. Lancet Reg Health Eur 2022 Jan;12:100247. [doi: 10.1016/j.lanepe.2021.100247] [Medline: 34901910]
134. Cezard G, Sullivan F, Keenan K. Understanding multimorbidity trajectories in Scotland using sequence analysis. Sci Rep 2022 Oct 1;12(1):16485. [doi: 10.1038/s41598-022-20546-4] [Medline: 36182953]
135. Head A, Birkett M, Fleming K, Kypridemos C, O'Flaherty M. Socioeconomic inequalities in accumulation of multimorbidity in England from 2019 to 2049: a microsimulation projection study. Lancet Public Health 2024 Apr;9(4):e231-e239. [doi: 10.1016/S2468-2667(24)00028-8] [Medline: 38553142]
136. Wallace J, Nwosu B, Clarke M. Barriers to the uptake of evidence from systematic reviews and meta-analyses: a systematic review of decision makers' perceptions. BMJ Open 2012;2(5):e001220. [doi: 10.1136/bmjopen-2012-001220] [Medline: 22942232]

XSL•FO
RenderX

137. Al-Sahab B, Leviton A, Loddenkemper T, Paneth N, Zhang B. Biases in electronic health records data for generating real-world evidence: an overview. J Healthc Inform Res 2024 Mar;8(1):121-139. [doi: 10.1007/s41666-023-00153-2] [Medline: 38273982]
138. Ren W, Liu Z, Wu Y, Zhang Z, Hong S, Liu H. Missing Data in Electronic Health Records (MINDER) group. Moving beyond medical statistics: a systematic review on missing data handling in electronic health records. Health Data Sci 2024;4:0176. [doi: 10.34133/hds.0176] [Medline: 39635227]
139. Owen N, Dew L, Logan S, Denegri S, Chappell LC. Research policy for people with multiple long-term conditions and their carers. J Multimorb Comorb 2022;12:26335565221104407. [doi: 10.1177/26335565221104407] [Medline: 35721799]
140. Burchett H, Umoquit M, Dobrow M. How do we know when research from one setting can be useful in another? A review of external validity, applicability and transferability frameworks. J Health Serv Res Policy 2011 Oct;16(4):238-244. [doi: 10.1258/jhsrp.2011.010124]

Abbreviations:

CKD: chronic kidney disease
CPRD: Clinical Practice Research Database
EHR: electronic health record
IBD: inflammatory bowel disease
LTC: long-term condition
MES: Mayo Endoscopic Score
MLTC: multiple long-term condition
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
SES-CD: Simple Endoscopic Score for Crohn Disease

Edited by S Brini; submitted 16.Jul.2025; peer-reviewed by F Huang, Y Chu; revised version received 09.Apr.2026; accepted 10.Apr.2026; published 19.May.2026.

Please cite as:
Hounkpatin H, Barton B, Ogden M, Mathur R, Stuart B, Dambha-Miller H
Understanding Remission of Long-Term Conditions Through Electronic Health Records: Scoping Review
J Med Internet Res 2026;28:e80796
URL: https://www.jmir.org/2026/1/e80796
doi:10.2196/80796

© Hilda Hounkpatin, Benjamin Barton, Margaret Ogden, Rohini Mathur, Beth Stuart, Hajira Dambha-Miller. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Barriers and Facilitators to the 3 Sides of Extended Reality-Rehabilitation Adoption: Scoping Review

Luuk Baltissen, BSc, MSc; Miranda Stienstra, BSc, MSc, PhD; Lien Denoo, BSc, MSc, PhD; Joris Knoben, BSc, MSc, PhD

Department Strategy and Entrepreneurship, Tilburg School of Economics and Management, Tilburg University, Warandelaan 2, Tilburg, The Netherlands

Corresponding Author:
Luuk Baltissen, BSc, MSc
Department Strategy and Entrepreneurship, Tilburg School of Economics and Management, Tilburg University, Warandelaan 2, Tilburg, The Netherlands

Abstract

Background: Rehabilitation using extended reality (XR) technologies can be used to address the growing shortage of health care staff, but the adoption level remains low. The current literature has given some first insights into what drives patients and clinicians to (not yet) adopt XR-rehabilitation tools. However, it does not sufficiently take into account that these tools will only be widely adopted if 3 types of actors collectively commit to it: developers must develop a tool, clinicians must prescribe it, and patients must use it. Because the preferences of these 3 actors may not always align, we aim to provide the first multi-actor insight into adoption.

Objective: This research aims (1) to determine what drives patients, clinicians, and developers to adopt or develop XR-rehabilitation tools and (2) to determine if and how these drivers align or misalign.

Methods: We searched PubMed, Embase, SCOPUS, and Web of Science with four search term categories: (1) types of rehabilitation care, (2) XR technologies, (3) adoption constructs, and (4) behavioral drivers. Using these search terms, we identified 1164 results, of which we included 64 in our review. All relevant empirical results within these papers were structured using the Non-adoption, Abandonment, Scale-up, Spread, and Sustainability (NASSS) framework.

Results: After exploring the adoption drivers of patients, clinicians, and developers, we identified 3 potential misalignments among these actors. The first possible misalignment is that clinicians may have much higher standards for a tool's medical efficacy. Because of this, they refuse to prescribe a medically less effective tool that would have matched the experience needs of patients and developers. The second possible misalignment is that clinicians value their work experience, while this is not a relevant factor for patients. Because using XR-rehabilitation tools can negatively impact a clinician's work experience, they may decide not to use a tool that patients and developers would have liked to use and develop. The third possible misalignment is that the patients' and clinicians' limited ability or willingness to pay may hinder the developer's economies of scale. Developers currently face high development costs, which they can recover by letting patients or clinicians pay for the tool. But these actors are not always able or willing to do so. As a result, developers may struggle to gain profitability, which limits the supply of XR tools.

Conclusions: Our scoping review provides initial evidence that differences in the behavioral drivers of patients, clinicians, and developers may lead to misalignments that hinder the adoption of XR-based rehabilitation tools. Scholars can use this review to further investigate potential misalignments between relevant stakeholders and how to resolve them. We encourage developers and regulatory institutes to collaboratively investigate the feasibility of new revenue models and product offerings to increase the adoption of XR-rehabilitation tools.

(J Med Internet Res 2026;28:e80055) doi:10.2196/80055

KEYWORDS

scoping review; extended reality; virtual reality; adoption behavior; rehabilitation care

Introduction

By 2035, a global shortage of 12.9 million health care workers is predicted [1]. Such a shortage will have tremendous negative societal consequences, such as decreased quality of care and increased staff burnout [2,3]. Rehabilitation care is particularly vulnerable to this problem, as demand in this sector is exceptionally high. A total of 1 in 3 people will need it at least once in their lifetime, and the therapy is often long and labor-intensive [4,5]. Because the workforce cannot be increased sufficiently, the sector must also look at technological innovations to reduce the workload. For instance, the Dutch Minister of Public Health stated: "With smart innovations, we can provide the care and support that is needed, thereby relieving the burden on informal caregivers and employees" [6].

XSL•FO
RenderX

One of these smart innovations is using extended reality (XR) as a tool for in-person and telerehabilitation. XR is an umbrella term encompassing virtual, augmented, and mixed reality (VR, AR, and MR) technologies that create user experiences where "real-world" elements are enhanced or replaced by "virtual" elements [7,8]. Telerehabilitation entails the delivery of rehabilitation services at a distance. Using these technologies can enhance the patient's therapy enjoyment and adherence while reducing the treatment duration [9,10]. However, despite these potential benefits and initial proof of the therapy's medical effectiveness [9], less than 5% of clinicians use it in their clinical practice [11-13]. Due to this, the potential of XR in rehabilitation care is far from realized.

Various papers have recently illustrated what drives patients and clinicians to not yet adopt XR-rehabilitation tools [14-16]. For example, the tools are deemed too expensive and often contain technical defects. Additionally, some patients need too much time to get used to the tool, and some clinicians believe that the movements patients make in XR differ too much from real-world movements.

Although these papers have provided a valuable first insight into nonadoption, they do not fully highlight the interconnectedness of both actors' adoption decisions. As one paper [16] recently stated: "Since the majority of studies have a singular perspective, there is a need to redirect the research focus towards a more comprehensive approach within the patient's environment."

Because clinicians are gatekeepers to the patient's access to care, adoption of the XR tool is inherently a collective decision by *both* actors [17,18]. To make it even more complex, the demand of these 2 actors can only be fulfilled if there are people willing to develop and sell the required XR tools (hereafter called developers).

Due to this need for supply, XR-rehabilitation tools will only be widely adopted if all 3 actors commit to it: developers must develop a tool, clinicians must prescribe it, and patients must use it. But their preferences and incentives do not always align with each other. For example, the clinician's informational advantage over their patients means that their (non)adoption decision can partially be driven by factors that patients do not consider or by clinician self-interest [19,20]. Developers are also in a buyer-supplier relationship with both patients and clinicians [21].

Because such misalignments can cause one actor to not opt in, they can prevent adoption for all. It is therefore essential to understand the drivers of all 3 actors as well as how they align and misalign. Yet, the current unilateral view on adoption does not allow for this [22]. Thus, to get the first multi-actor insight into adoption, we conducted a scoping review with the following objectives [23]:

1. Determine what drives patients, clinicians, and developers to adopt or develop XR-rehabilitation tools.
2. Determine if and how these drivers align or misalign

Methods

This work was conducted according to the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) reporting guidelines [24]. The filled-in PRISMA-ScR checklist can be found in Checklist 1. To ensure adequate and efficient database coverage for published academic papers, we searched Scopus, Embase, Web of Science, and PubMed from October 31, 2024, to January 20, 2026 [25].

We used 4 search term categories in these databases. The first category covers the main types of rehabilitation care. As Battel et al and others recently illustrated, rehabilitation is a very broad field that largely overlaps with other medical specialties like psychiatry, geriatrics, and palliative care [26-29]. To cover the broad range of relevant studies, we searched for the general term "rehab*" and included papers from all overlapping subfields, such as psychiatric and neurological rehabilitation. We also searched for the 4 therapies most commonly classified under (XR-)rehabilitation: occupational, physical, speech, and language therapy [14,16,30]. However, to avoid capturing studies that solely focus on the other specialties, we did *not* search for terms like psychiatry and neurology. The second category covers XR types. Currently, the literature views (too) many gaming devices as an XR tool, which has made it difficult to compare different studies [31-33]. To enhance comparability, we follow Milgram and Kishino's view on XR and only consider a device an XR tool if it generates a user experience where "real-world" elements are enhanced or replaced by "virtual" elements [7,8]. We therefore searched for the primary types of XR, head-mounted devices (HMDs), and exergames, but *not* for devices like the Wii, Xbox Kinect, and Virtual Reality Rehabilitation System. The third category comprises terms that are associated with adoption [34]. In this review, we define adoption as "the intention, initial decision, or action to try or employ an innovation or evidence-based practice" [34]. This term is often interchanged with similar constructs like uptake and acceptance, which are salient to all stakeholders (including developers) [34]. To capture this breadth, we included 42 different adoption-related terms in our search string [34]. Importantly, several of these terms, such as "feasibility," "integration," and "adherence," are also often used to refer to a tool's medical efficacy and technical operability [35-38]. Because these topics fall outside the scope of this review, we aimed to remove them a priori by using a fourth search term category that captures behavioral drivers [39,40]. As Nilsen [41] illustrated, the determinant framework is the only theoretical approach that investigates adoption behavior by explaining how different variables influence (ie, drive) adoption. Because this aligns with our research objective, we added all deterministic terms that Nilsen identified [41]. This includes terms such as barrier* and facilitator*. We did *not* search for terms such as obstacle* and challenge*, because they are also frequently used in other contexts. For example, patients often face health challenges due to their disorder. The full search query can be found in Multimedia Appendix 1. Multimedia Appendix 1 also outlines how the query was progressively

XSL•FO
RenderX

extended from the first to the last search date in response to reviewer and editor feedback.

Using the criteria in Textbox 1, the first author (LB) selected all resulting papers in three rounds: (1) title, (2) abstract, and (3) full text. To illustrate the reliability of the screening process, the second, third, and fourth authors (MS, LD, and JK) all blindly screened a unique and randomly selected set of at least 10% of the papers *each*. This resulted in a coverage of over 30%, which goes well beyond the typical 10%‑25% requirements for a trustworthy estimate of the interrater reliability (IRR) [42]. The first author held individual meetings with all authors after they completed each stage. In these meetings, they discussed and resolved any discrepancies that emerged. The outcomes of these meetings were directly communicated to all other authors. We measured IRR using Cohen's Kappa [43,44].

Textbox 1. All inclusion and exclusion criteria that were applied during the screening process.

Inclusion criteria

- Available in full text
- Written in English
- Peer-reviewed papers
- Research in the rehabilitation context, both in-clinic and at-home, including all subfields that overlap with other medical specialties (eg, neurological or psychological rehabilitation)
- Research on a medical application of extended reality
- Empirical data on the behavioral drivers of developers, clinicians, or patients

Exclusion criteria

- No full text available
- Not written in English
- Works that are not peer-reviewed papers (eg, conference proceedings)
- Research in a nonrehabilitation context (eg, distracting patients before surgery or providing visual enhancement)
- Research on different technologies (artificial intelligence, robotics, etc) or non–extended reality devices (Wii, Xbox Kinect, etc)
- No empirical data (eg, literature review) or data with no insight into the individual behavioral drivers (eg, only including a System Usability Score or aggregating data across actors)

A data charting form was developed by all researchers to map the descriptive variables reported in the selected papers. This entailed the actors that were treated, the type of empirical data, the adoption framework used, the usage location, the intervention type, the disorders treated, the XR-type, the device used, and the funding source.

To structure the empirical findings of the selected papers, we divided all data across 3 groups, based on the actors they apply to (ie, patients, clinicians, and developers). After this, the data for each actor group was divided into 7 subgroups based on the domains in the Non-adoption, Abandonment, Scale-up, Spread, and Sustainability (NASSS) framework (ie, 21 subdomains in total). This framework explains the Nonadoption, Abandonment and challenges to Scale-up, Spread, and Sustainability of innovations using the interaction between (1) the patient's condition, (2) the technology, (3) the value proposition, (4) the adopter system, (5) the health care organization, (6) the wider system, and (7) how these 6 domains develop over time [45,46]. We chose this specific framework because it is the adoption-related framework that best captures the complex interaction between patients, clinicians, and developers [41,46-48]. To get a more granular insight into these NASSS domains, we then used the guidelines of Colquhoun et al [49] to inductively draft a set of categories that best represented the dynamics within each domain. The first author (LB) used this structuring form to determine whether the categories were present (1) or absent (0) in each paper. To ensure reliability of this process, the second, third, and fourth authors (MS, LD, and JK) reviewed the first author's (LB) completed form for all papers they included in their screening process. They also assigned a "1" to a category if it was present in a paper, and a "0" if it was not. Here, the first author (LB) also held individual meetings to discuss any discrepancies. Because the data structuring form contains a large number of categories, there were many zeros in some papers. Because Cohen Kappa can yield deceptively low values in such cases of class imbalance, we used Gwet AC1 as the IRR metric instead [50,51]. To further substantiate our findings, we also conducted a thematic frequency count to illustrate whether and how the prevalence of themes differed between actors [52,53].

Results

The Literature Selection and Structuring Process

The search query yielded 654 unique records, of which 64 have been included in this review to analyze the behavioral drivers of patients, clinicians, and developers. The main exclusion reasons were that the work was a literature review (n=180) or that it had a different research focus, such as clinical trials or technical proofs-of-concept (n=144). Other reasons for exclusion were that the work was not a peer-reviewed article (n=139) or that it focused on non-XR devices (n=100). A total of 27 records

XSL•FO
RenderX

were removed for different reasons, mostly because they were preregistered but not yet executed (n=18). Figure 1 outlines the full selection process. MS, LD, and JK validated LB's screening and charting process for 219 of the 654 (33.5%) initially screened records. This validation process indicated that LB selected the literature nearly perfectly in both the title and abstract stage (κ=0.83) and the full paper stage (κ=0.89) [54]. The empirical findings were also charted nearly perfectly, with the average Gwet AC1 being 0.87 (range 0.63-1.00). Most studies were funded by academic, health care, and/or governmental institutions (48/64, 75%). In total, out of 64 studies, 11 (17.2%) studies did not receive external or research-specific funding, while 5 (7.8%) studies did not disclose their funding source in the paper. No studies were directly funded by firms from the life sciences industry, although some funds were provided by industry-adjacent sources like EUROSTARS. Further details can be found in Multimedia Appendix 2.

Figure 1. Study selection flow diagram.

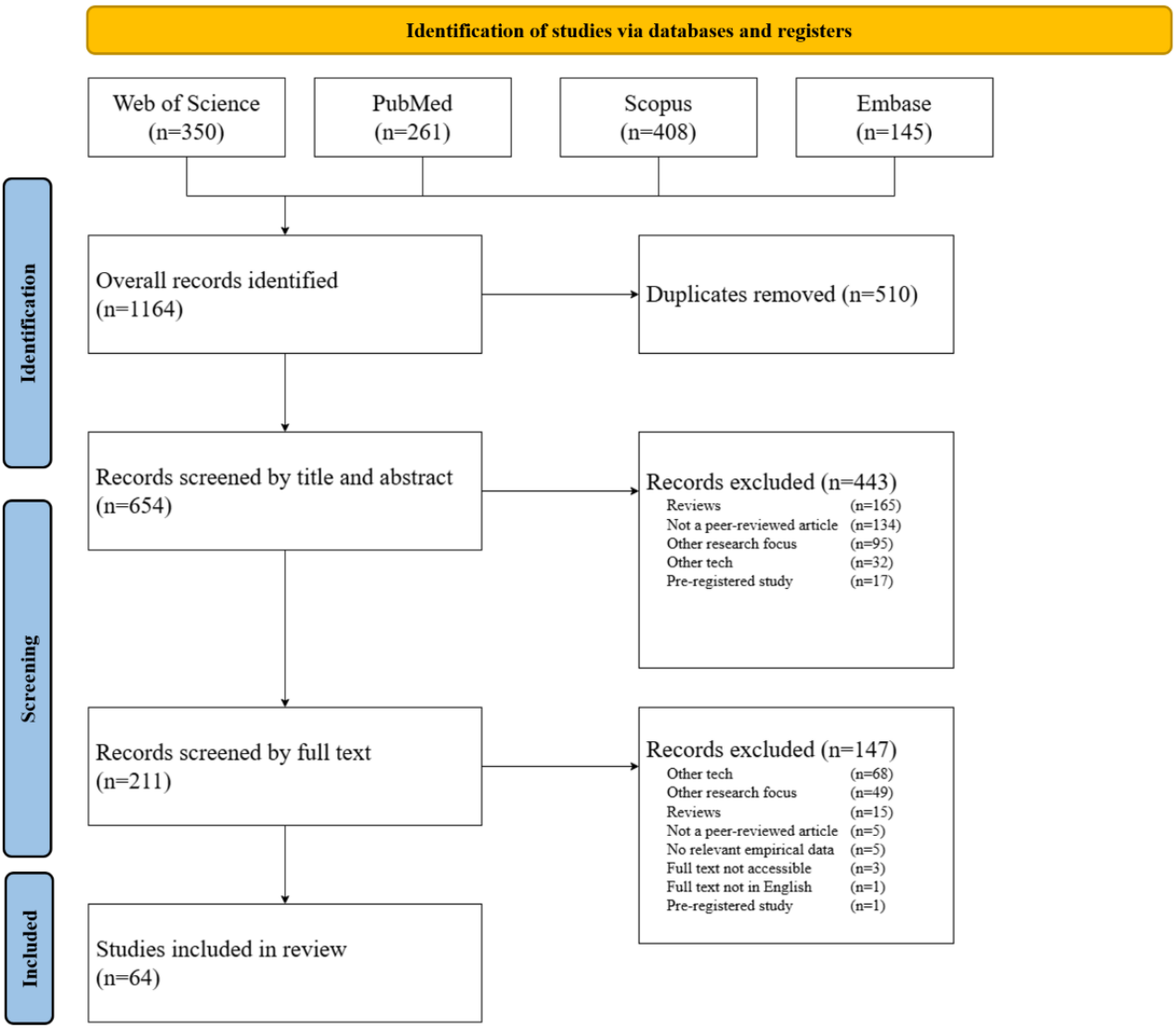

As shown in Figure 2, the developer's perspective is highly underexposed in the included papers. Patients and clinicians were treated in 62.5% (40/64) and 57.8% (37/64) of the papers, respectively, while developers were mentioned in only 6.3% (4/64). Additionally, only 25% (16/64) took a multi-actor approach, and Saidi et al [22] is the only paper that included all 3 actors. In their study, Saidi et al conducted 3 workshops and 10 interviews with patients, clinicians, and developers in stroke rehabilitation. Importantly, however, they aggregated most findings across actors and mostly referred to all 3 groups as "participants." As a result, their analysis offered limited insight into how the drivers differed across actor types. This further supports the need for a 3-sided perspective on adopting XR-rehabilitation tools. The descriptives of each included paper can be found in Multimedia Appendix 2.

XSL•FO
RenderX

Figure 2. The number of papers included per actor.

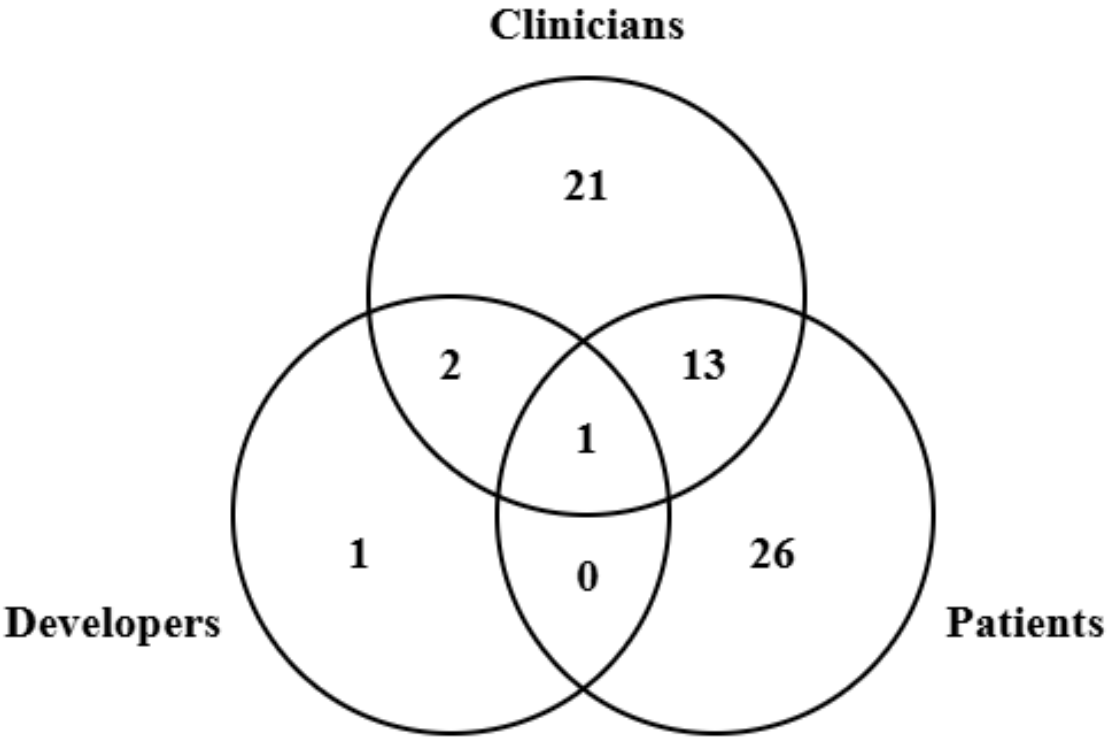

Domain 1: The Value Proposition

Overview

This domain addresses whether a new technology is worth developing in the first place, and for whom it generates value [45]. We interpret this domain as the net sum of benefits and losses. XR-rehabilitation is more likely to be adopted if the benefits of developing or adopting the tools outweigh the associated losses for *all* actors. Conversely, if at least 1 actor faces insufficient benefits, then that actor will not opt in and may prevent adoption for all. We outline the benefits and losses for all 3 actors individually.

Patients

Overall, the benefits and losses for patients can be divided into three categories: (1) care experience, (2) medical efficacy, and (3) savings or investments of time and money.

Care Experience

The most frequently mentioned category is the care experience (31/40 papers, 77.5%). Patients see numerous benefits of using XR tools. First, they make therapy more fun [31,55-83]. Currently, clinicians ask patients to do a certain set of exercises, which they must individually repeat at home. The repetitiveness makes these exercises dull, causing low adherence. At-home XR provides a way to alleviate this dullness. In XR, exercises are presented as a game that patients can play by themselves or against others. These exergames are interactive and provide feedback on their performance, which makes exercising more fun. Second, XR tools can give patients more autonomy [55,58,62,64,69,72,75,84]. Currently, patients must frequently meet their clinician for exercise instructions or supervised practice. These visits can interfere with the patient's daily activities. At-home XR tools can address this issue by providing exercise demonstrations and enabling patients to exercise at a convenient time and location. This allows the exercise schedule to better fit into the patient's daily life. Some additional gains are that the tool's biometric feedback can give patients insights into what they are capable of, can be relaxing, can give them confidence, can make them feel that they are more in touch with their caregivers, and allow them to work with novel tools [55,58,59,61,62,64,66,70,73,83,84].

Unlike the benefits, relatively few patient papers mention losses related to the care experience [55-57,62,64,70,72,74,76,80]. Only two losses were mentioned more than once: (1) frustration due to the inability to exercise effectively or to win, and (2) a feeling that therapy should be challenging rather than fun. We outline both losses below.

Frustration generally arises from poorly designed hardware and software. For example, bugs and glitches take patients out of the immersive flow and require time to resolve [57,61,72]. A poor design can also lower the patient's self-image [56,64,74,80]. Because of their impairment, some patients may find the exergames too difficult to play, may be unable to use the hardware, or may continually lose against "able-bodied" friends and family. This can make them feel limited in what they can do and possibly inferior to others. One patient excellently illustrated this point by stating that *"dealing with the frustration of using VR combined with the existing frustration of having aphasia and trying to communicate"* makes them *"feel like a total failure"* [74].

XSL•FO
RenderX

Additionally, 2 papers indicated that patients felt that the enjoyable nature of VR was inconsistent with their expectations of physiotherapy [70,76]. One patient indicated that "you have an expectation that when you go to [physical therapy] they give you exercises or manual treatment so when you leave it feels like you've done something" [76].

By replacing these regular exercises with enjoyable experiences, some patients feel they no longer receive any treatment.

Medical Efficacy

A total of 67.5% (27/40) of the patient papers illustrated that medical benefits are a major adoption driver. However, unlike the care experience, there is not one benefit that is very frequently mentioned in this category. Rather, 14 [22,55-58,60,64,66,69,70,76,80,82,85] papers mention that XR-rehabilitation tools allow patients to move more due to the motivational aspect and more opportunities to train by themselves. Additionally, 12 [55,57,58,68-70,73,76,81-84] papers mentioned that XR tools can distract patients from their pain. This is a medical benefit in itself, but it also leads to a lower fear of movement and thus better exercise execution. Other benefits were that patients can practice in a more controlled environment than the real world [59,61,74,79,86] and that they can better visualize their exercises [61,69,72,80]. Using XR tools can also reduce anxiety or stress symptoms [59,62,65], offer clinicians more (objective) insights into the patient's performance through tracking [55,72], stimulate patients cognitively [80,83], and allow them to exercise independently when they are still on the waiting list [55].

Although XR-rehabilitation is designed to provide medical aid to patients, it can also have substantial medical downsides. When using XR technology, some patients can feel nauseous, claustrophobic, or disoriented [31,55-59,61-63,66,70,71,73,74,79-83,87,88]. Others feel unsafe exercising using immersive HMDs because they cannot see their surroundings [22,58,70,71,80,89], feel that they may make wrong movements due to overexcitement or lack of supervision [55,64,80], or are afraid of becoming addicted to the exergame [59].

Savings or Investments of Time and Money

The last value proposition for patients is the savings or investments of time and money. On the one hand, XR tools can save patients time because they can rehabilitate at home, rather than having to travel to their clinician [55,57,64,74]. Conversely, using XR-rehabilitation tools often requires an upfront investment [55,57,59,61,62,64,69-74,77,78,80,83,84,88-90]. To be able to use the tool, patients must often spend time getting to understand it, setting it up, and fixing bugs when needed. They must also sometimes pay an additional fee for the hardware and software.

Clinicians

While assessing the value propositions for clinicians, it became apparent that they largely overlap with those of patients, because both actors value (1) the tool's medical efficacy, (2) the patient's care experience, and (3) investments or savings in time and money. However, despite these similarities, 2 differences appeared.

The first difference is that patients and clinicians evaluate the 3 value propositions differently. When evaluating the tool's medical efficacy, clinicians consider a broader set of factors than patients. Particularly, they are more likely to also consider the long-term risks of overcompensation, compensatory movements, infections, attachment issues, and the lower quality of tele-evaluations [10,11,13,56,66,91-101]. When evaluating the cost (savings), clinicians go beyond considering the patient's costs by also considering the costs they must bear themselves. These clinician costs include both time (eg, learning to use the tool) and money (eg, maintenance costs) [10-13,31,56,58,68,73,79,81,90-92,94-96,98-100,102-105]. When evaluating the patient's care experience, clinicians appear to value fun differently than patients do. As noted earlier, XR can be fun for patients because they do not have to do "dull" exercises, can play with friends or family, and get performance feedback. Clinicians acknowledge these aspects [10,13,56,58,81,82,92,101]. However, they seemingly perceive fun as a means to enhance therapeutic engagement and clinical outcomes, rather than a valuable end in itself. For example, one physiotherapist stated that *"for compliance – if it's fun compliance is going to go up"* [95]. Another therapist *"felt that VR helps to improve the child's engagement, which can facilitate movements, increase movement control and facilitate stability and function."* [56] One patient, on the other hand, stated that *"If [occupational or physical therapy] was pretty boring then a lot of the times I would just be like 'okay, well, let me just get through this then I'll just be able to do VR."* [68]. This suggests that while both groups recognize the benefits of enjoyable experiences, patients can also value them for their own sake. It is important to note that this is not a strict dichotomy; clinicians still appreciate the experiential benefits of XR. However, the current literature does indicate that they value it less than patients do.

The second difference is that patients do not refer to the clinician's work experience. Although clinicians also rarely mention it themselves, XR tools can make their work more enjoyable because they can work independently of time and location, feel innovative, and can offer a more varied set of treatments [10,11,13,31,84,91,92,96,97,103,105]. However, XR tools can also make the clinician's work less enjoyable, as using them can make them feel like a virtual life coach or a patient mill, can narrow their skillset, or can awkwardly interrupt the in-clinic session [13,91,97,100,103].

Developers

Of all 64 included papers, only 4 treated the developer's perspective [22,94,96,106]. Three of these papers did not truly focus on their development barriers or facilitators, but instead used their input to validate the feasibility and relevance of the clinicians' ideas on XR usage [22,94,96]. For example, in 1 paper, VR specialists "engaged in discussions with the [clinicians] about how VR tasks could be programmed to incorporate more or less challenging situations and distractions" [96]. Only the paper of Kulkov et al [106] had the developer's perspective at its core. Because of the limited number of developer-focused papers, we only found 2 indications about what the benefits and losses for developers could be.

XSL•FO
RenderX

First, 3 [94,96,106] papers indicated that developers seem to reasonably understand the benefits and losses for patients and clinicians. For example, developers mentioned that XR "provide an indoor solution to motivate physical activity, which could be achieved on a 'daily basis'" [94], "is good for building an environment that duplicates the environment people are going to be in" [96] and that "self-treatment could be offered at a convenient time and place with follow-up visits to the specialist" [106]. In 2 [96,106] papers, developers also acknowledged that XR requires high financial investments from patients and clinicians and that clinicians may feel less needed or relevant when using XR. Although developers did not mention all value propositions emphasized by patients and clinicians, they also did not mention any new or different ones. This suggests that developers have a fair understanding of the user's preferences.

The second main insight is that developers face *"costs of development [that] are still pretty high at present"* [94]. Because developers can only sustainably supply tools if they at least break even, it is plausible that these costs are a substantial barrier for them.

Domain 2: The Condition

This domain covers the patient's condition, sociocultural factors, and comorbidities. Because the developer papers did not dive into this, their perspective is not outlined below.

Patients

Of the 19 [56-62,64,66,69,71,72,74,75,79,81,87,88,90] patient papers that touch upon the condition domain, there is one major consensus: more impaired patients are less likely to adopt XR-rehabilitation. These patients generally have less confidence in their ability to use the tools. They are also less able to use them, even if they wanted to. As one stroke patient stated: "I don't think I would do anything because... like my balance still isn't wonderful" [71].

There are also indications that patients who are "too healthy" are less likely to use XR tools, because they benefit more from exercising in the real world [58,64,74]. However, because this point is sparsely mentioned, there is less widespread consensus on this aspect. Additionally, although 6 [31,60,61,66,87,90] papers included multiple disorder types, there is no definitive indication of whether patients with certain disorder types are more likely to adopt XR tools than others.

Clinicians

A total of 48.6% (18/37) of the clinician papers touch upon this domain. Twelve of these [56,58,66,81,87,92,94,96,98,99,102,103] also emphasize that clinicians will not prescribe XR tools to patients who are "too impaired," because they cannot safely use them. Additionally, while 10 [11-13,97,103,105,107-110] papers included multiple disorder types, there is again no definitive indication of which disorders clinicians are more likely to target with XR tools.

One aspect highlighted in the clinician papers, but not in the patient papers, is that clinicians require a minimum number of patients using an XR tool before they are willing to adopt it [11,92,96,99,100,103,105,107-109]. As one clinician said: "[Due to h]igh acquisition costs, the patient group for which [using XR] makes sense is too small to justify a purchase for me" [11].

Because using XR requires high financial and time investments, the tool must be used by enough patients to provide value. Consequently, clinicians who work with high-prevalence conditions are more likely to adopt XR tools, given the larger potential user base.

Domain 3: The Technology

This domain treats all features, knowledge generated or needed, and supply chain issues of the XR-rehabilitation technologies. Although the included papers treated several technologies, the HMD was the most used [10,11,22,31,55,57-65,67-70,73-83,85-88,90,94-96,98-101,111,112].

Patients

A total of 90% (36/40) of the patient papers mentioned technological factors that influence the patient's adoption decision. There is a broad consensus on the importance of 6 such factors. The most mentioned feature is content variety and customization [22,31,55-67,71,74,77,79-82,84,86,88,111]. Patients want to use tools that fit their (changing) needs, and they want to enjoy a wide range of games to stay entertained. The second-most mentioned feature was the presence of a support staff [55,57,59,60,62-66,68,69,71,73,74,76-80,84,87]. Even when exercising at home, patients want to stay in touch with clinicians for medical advice and supervision. They also want to be able to contact someone for technical support. Comfort and ease of use were the third and fourth-most mentioned features [31,55-59,61,62,66,69,72-74,76-80,82,84-88]. Devices that are uncomfortable to wear or difficult to use can cause physical strain, enhance the patient's sense of impairment, and possibly prevent patients from continuing their therapy. The 2 other frequently mentioned features are technological reliability and the availability of tutorials or guidelines [31,55-57,59-62,64-67,71,72,74,79-81,84,87,88]. Tools that are buggy are less enjoyable and stimulate patients to stop using the tool. Additionally, because most patients are new to XR, they want to be educated on how to use it.

There were also 2 features on which the findings were inconsistent. The first one is the level of immersion. Some patients prefer using an HMD with immersive VR because they are more engaging [31,55,57,68,73,74,79,81]. Others prefer nonimmersive HMDs because their inability to see the outside world makes them claustrophobic or afraid of hitting something in their physical surroundings [57,58,70,73,74,79]. It is unclear whether the benefits of immersion outweigh the risks and how these potentially differ between patients. The second inconsistent feature is the exercise type. Some patients find full-body movements with free movement more engaging than "simple" movement repetitions [68,71,74,76,89]. However, these full-body movements come at a higher risk of falling or tripping over objects. Therefore, some patients prefer seated exercises with "simple" movement repetitions [57,71,74,89]. In total, 6 papers illustrate the patient's preference for at least 1 of the 2 exercise types [57,68,71,74,76,89], with 3 of those mentioning both [71,74,89]. However, none of the works truly indicate

XSL•FO
RenderX

which exercises patients prefer, or whether preferences differ between patients.

Clinicians

A total of 97.3% (36/37) of the clinician papers mentioned XR features that influence whether clinicians adopt the XR tool. When comparing these to the patient's technology factors, it became apparent that both patients and clinicians value training or guidelines [10,11,13,56,58,66,68,81,82,84,85,87,91-93,95,96,98,99,101-105,107-110], software variety or customization [10,11,13,22,56,66,68,73,79,81,82,85,91,92,94,96,98,99,101-103,107-110], ease of use [11,12,31,56,66,73,81,82,84,85,91,93,96,97,99,101-103,105,107-109], support staff [11,13,22,31,56,68,73,82,87,92,96,98,99,102,103,105,107-109], tool reliability [10,11,31,56,66,81,82,84,92,94,97,99,101,104], and comfort [11,56,66,68,82,91,94,95,98,99,101]. There is, however, one main difference between patients and clinicians.

Namely, clinicians have a much stronger preference for technologies with low (medical) risks. Three aspects stand out in particular. First, clinicians mention guidelines and medical evidence much more than patients. Guidelines were mentioned 28 times by clinicians versus 13 times by patients. Medical evidence was mentioned 18 times by clinicians [10,11,13,56,58,91,92,96-100,103,105,107-110], and only once by patients [74]. Second, clinicians expect more depth in their guidelines and evidence. Of the 13 papers that treat both patients and clinicians [31,56,58,66,68,73,79,81,82,84,85,87,90], 4 explicitly referred to their guideline preferences [56,66,84,87]. Patients are generally satisfied with an explanation of the tool's potential benefits and practical tutorials on how to use it. For example, one patient stated that they want *"a field training module [with] the basics of how to connect to the Internet, Bluetooth, troubleshooting"* [56]. However, clinicians want more than this "everyday" information, because they see more risks and feel responsible for the patient's well-being. Hence, they also want information on treatment dosages, side effects, etc. For example, one work [56] illustrated that clinicians *"would like just some guidance on what do with [the patient]and how far to push them"* and that they *"would want to always ensure that [they]'ve demonstrated and that [the patient] is using it in the correct way"*. The third indication of the clinician's risk-aversion is that all clinician papers prefer seated exercises because they are safer than full-body exercises [22,94-96,99]. Some patients, however, have the exact opposite preference, because full-body movements are more fun to do.

Developers

Only 1 [96] paper gave an insight into the developer's perspective on XR technology. Their findings indicate that developers seem willing to adapt the technology to meet client needs, whether by adding requested features or modifying the tool to accommodate the patients' medical limitations. However, they also seem bound by what is technically feasible. As one developer explained: "Decisions around hardware are enormous...what you can do is gonna be limited to your hardware choices" [96].

As it stands, these constraints disallow developers from meeting all the needs of patients and clinicians. This possibly discourages them from developing a tool at all.

Domain 4: The Adopter System

This domain entails patient and clinician characteristics that are nonclinical in nature. Because the developer papers did not dive into this, their perspective is not outlined below.

Patients

Besides their condition (Domain 1), there are 2 main patient characteristics that influence their willingness to adopt an XR tool.

The most mentioned one is their experience with technology [55-60,62-64,66,67,69,70,72-74,89,112], which can have both positive and negative impacts. Generally, patients who are more experienced with (non-)XR technology have more confidence in their ability to use the tool, find the experience more enjoyable, and need less time to understand it. However, previous technology experience also increases the patient's expectations of the tool, particularly if they have played highly advanced video games. A negative experience in the past can also demotivate patients from trying another tool again. The second most mentioned patient characteristic is age, mostly because younger patients have more tech experience [56-58,69,74,79,89].

Clinicians

While the previous section did not indicate that patient adoption is influenced by clinician characteristics, the opposite is true for clinicians. Rather, 40.5% (15/37) of the clinician papers mentioned how patient characteristics influence the clinician's decision to adopt.

The 2 main patient characteristics mentioned by clinicians were also the ones mentioned by patients themselves: tech experience and age. As previously indicated, patients with more tech experience have more confidence in using the tool, find the experience more enjoyable, and need less time to get used to it [10,56,84,92,94-96,98,102], and as younger patients generally have more tech experience, they are also more likely to adopt [11,12,56,58,81,84,91,94-96,98,102,110]. As mentioned earlier, clinicians require a minimum number of patients before they can retrieve the financial and time investments associated with adopting XR tools. As a result, clinicians who treat younger or more tech-experienced patients may be more inclined to adopt XR-rehabilitation tools.

A total of 51.4% (19/37) of the papers also mentioned clinician characteristics that impact their likelihood to adopt. Of these characteristics, only experience with technology and age were repeatedly mentioned. As for patients, younger clinicians generally have more tech experience, meaning they better understand the tool's benefits and need less time to start working with them [10,11,13,56,58,66,79,91,92,96,99-103,107-109]. Characteristics like work experience, education type, and care specialization were also mentioned [10,11,13,100,103,108-110]. However, because these factors were sparsely mentioned, their true impact on adoption remains unclear.

XSL•FO
RenderX

Domain 5: The Organization

This domain entails how the adoption or development decision is influenced by the health care organization's capacity and readiness to adopt, the board's support, and the extent to which established work routines will be disrupted by the XR tools.

Patients

None of the patient papers illustrated how the formal health care organization influences the patient's adoption decision. Hence, it remains unclear whether and how this influences their adoption decision.

However, 15 [55,56,58,64,66,68,71,72,75,77,80,84,87-89] papers did stress the importance of informal caregivers, such as friends and family. The role of these caregivers is twofold. First, they can help patients perform their exercises when they are not able to fully do so themselves, such as by resolving bugs or putting on the HMD. Second, they can motivate patients to try the tool and continue using it by being supportive or even training with them. As an example of how caregivers can *demotivate* patients, one paper [89] illustrated that a participant felt discouraged when her daughter said, "Mom, what are you going to do with a smartphone?"

Clinicians

A total of 67.6% (25/37) of the papers illustrated how the health organization impacts the clinician's adoption decision. The results of these papers can be divided into three categories: (1) resource/time sufficiency, (2) supportiveness, and (3) institute type.

The first 2 characteristics are supported by several sources. First, 22 [11,13,56,58,66,68,73,87,91-93,96,98-100,102,103,105,107-110] papers mentioned that clinicians working in an institute where they get dedicated time and resources are more likely to adopt. As illustrated in the value proposition domain, using XR tools requires both time and money. Naturally, clinicians are unable to adopt the tools if they have insufficient access to either of the 2. Second, 18 [12,13,56,58,87,91-93,96,98-100,102,103,105,107-109] papers illustrated that clinicians working in a supportive environment are more likely to adopt. This has both an administrative component in the form of board support (ie, more resources or time) and a cultural component in which colleagues, lead users, and the board stimulate the clinician to be innovative.

The third characteristic is the institute type. Although it was only mentioned in 4 [107-110] papers, combining it with findings in other domains does indicate that institute type supports XR-adoption in 2 ways. The first way is through private funding. One paper [110] found that clinicians working in private practices have significantly more access to XR tools than those working in public practices. The authors do not explain why this relationship exists. However, as we will illustrate in the following subchapter (the broader domain), public institutions are more dependent on insurance coverage and public funding. Because both funds are sparse, it is likely that private practices have more access to XR tools. The second way is by operating on a sufficiently large scale. Contrary to the aforementioned findings, 2 papers found that most of the XR users work in (public) hospitals and rehabilitation centers [107,109]. These papers also do not explain why this is the case. But, as explained in the clinician's condition domain, it is plausibly caused by the fact that these institutes operate on a larger scale, meaning they can divide their costs over more patients. This proposition is also supported by 1 [110] paper that found that clinicians working in institutes with more than 100 children per week have significantly more access to XR tools than those working in institutes with less than 100 children per week.

Developers

Although the developer papers did not extensively examine the role of the health care organization, 2 studies still gave an indication of their value to developers. One [106] paper indicated that developers could benefit from the endorsement of key health care organizations. Recommendations for safety and success seem important in the market for XR-rehabilitation. Hence, the endorsement of reputable institutes could stimulate other organizations to also use their tools. Another paper [94] indicated that developers seem to not always fully understand the needs of patients and clinicians a priori. By collaborating with health care organizations in the development process, developers can get *"feedback [they] can work with"* [94]. By using this feedback, developers can develop a product with a better value proposition, which plausibly results in more sales.

Domain 6: The Broader Domain

This domain entails the wider institutional and sociocultural context, such as health policy, insurance reimbursement, and regulatory aspects of patient-facing technological development.

Patients

Only 12.5% (5/40) of the papers touched upon the broader domain. A total of 3 of those mentioned that a good internet connection is a necessity for at-home rehabilitation [66,84,88]. This indicates that people living in more technologically advanced regions are better able to use XR-rehabilitation tools. The other 2 [64,89] papers mentioned that some patients are not happy with the societal transition to complex technology. This requires them to continually invest time and money to be able to use new tools, while they are often still satisfied with their older forms of technology.

Clinicians

Unlike the patients, there are quite a few clinician papers that touch upon their perspective on the broader domain (18/37 papers, 48.6%). These papers illustrate 3 ways in which the broader domain impacts the clinician's adoption decision.

The first factor is again the technological advancement of the clinician's region [31,56,66,84,91,96,97,99,102,104,107,108,110]. The more advanced a region is, the more value an XR tool can provide, for example, because it has a more stable Wi-Fi connection. Clinicians living in more technologically advanced regions are also more exposed to innovations and technology. This means they generally have more tech experience and more opportunities to start using XR-rehabilitation. Mensah-Gourmel

XSL•FO
RenderX

et al [110] support this proposition by illustrating that clinicians in Western Europe generally have more access to XR-rehabilitation than clinicians in Northern, Eastern, and Southern Europe.

The second factor is insurance coverage [10,11,91,95,97]. Because new XR hardware can be expensive, clinicians are more likely to adopt a tool if it is reimbursed by the health care insurer. They also want to be covered against insurance claims if injuries were to result from XR-interventions.

The third factor is funding opportunities. Because XR-tools often require a large user base to become financially viable, they currently lack a sustainable revenue model for many clinicians. As a result, their adoption decision heavily depends on the availability of external (public) funding. When such funding is limited or unavailable, clinicians are unlikely to adopt the tools [11,13,92].

Developers

Only 1 [106] paper provided insights into the developer's perspective on the broader domain. Nevertheless, its focus on institutional barriers within the health care industry still yielded some preliminary insights.

First, developers seem to face major barriers in terms of both costs and revenue. When developers develop a tool, they must adhere to strict data management regulations. This potentially contributes to the high development costs mentioned earlier in the developer's value proposition domain. Even when developers overcome this hurdle, other medical regulations seem to prevent them from scaling up the use of at-home care, because clinicians must evaluate the patient's condition in the clinic. Additionally, the authors indicate that public clinics "*have no chance of using new solutions*" because health care insurers are hesitant to pay compensation for novel treatment procedures [106]. This arguably narrows the pool of potential customers to private practices, making it less attractive to enter the market.

However, there are also 2 ways by which the broader domain could aid developers. The first one is through collaborations that provide (1) financial support via public funding or (2) recommendations from clinics, informal caregivers, and opinion leaders. The second possible source of support is through the entrance of Big Tech firms. Currently, the market is mostly occupied by start-ups. While typically having less scale and resources than Big Tech firms, these smaller firms are not necessarily opposed to the market entry of these major companies. Rather, they expect that their reputation could change the attitudes of users and could establish a new business ecosystem for current developers. Whether these expectations are true remains to be determined.

Domain 7: Embedding and Adaptation Over Time

This domain entails how the aforementioned domains evolve over time. Although literature is limited, it still offers some preliminary insights at both the micro and macro levels. At the micro level, the most prevalent insight is that patient adherence tends to decline as sessions progress, but that major differences exist between individuals. Ploderer et al [72] excellently illustrate this by distinguishing three equally sized groups: (1) "failing" patients who stop early, (2) "temporary" patients whose usage fluctuates, and (3) "continuous" patients whose use remains relatively stable over time. Four other papers touch upon this topic by indicating that adherence either decreases [60,84,88] or stays consistent over time [75]. There are indications of what differentiates the 3 use groups, such as the patient's time and resource availability, but the evidence is currently too sparse to draw robust conclusions. Other micro-level changes are that session efficiency (actual time spent using XR-game/duration of XR-therapy session) increases over time [60] and that clinicians seem to get more confident with XR if they use it more frequently [103]. On the macro level, some actors have indicated what they expect might happen in the future. For example, clinicians expect that they must look for alternatives to address the growing demand for rehabilitation [91], and developers expect that XR's value as an alternative will continue to grow [106]. Although insightful, it is important to note that these macro-level insights are mostly based on expectations about the future rather than directly observed changes over time.

Discussion

Principal Findings

By reviewing the 7 NASSS domains for patients, clinicians, and developers, we provided a first insight into the factors that drive them to use or develop XR-rehabilitation tools. A comparison of the actors shows that their drivers are frequently aligned. For example, all 3 actors embrace tools that motivate patients to exercise more or that allow for home-based therapy with the support of relatives. However, the comparison also reveals 3 instances in which their drivers may not be fully aligned. Because these misalignments represent potential barriers that are underexplored in the current unilateral view on adoption, we discuss them in more detail below. A summary of the main drivers and the resulting misalignments can be found in Multimedia Appendix 3.

Medical Efficacy

The potential match in drivers between patients and developers seems promising. Some patients do not value medical benefits as much as they value other factors, such as fun. Some patients indicated a preference for full-body exercises over seated exercises, for example, despite the risk of falling or tripping over a cord. Patients also mentioned the importance of a medical evidence base only once and rarely mentioned the risks of overcompensation and addiction.

These drivers are likely to appeal to developers, who are more than able to develop engaging and motivational games. The low need for an evidence base reduces their costs and increases their profit potential. If patients truly attach little value to the medical benefits or evidence, this would be like buying a commercially available exergame like Beat Saber and Gorilla Tag [113]. Hence, a match between the 2 is plausible.

The clinician may not accept this trade-off, however, as it offers insufficient medical benefits. In their work, clinicians are bound by strict medical regulations that prevent them from prescribing tools that do not meet the necessary medical standards [106].

XSL•FO
RenderX

Beyond these regulatory requirements, clinicians also seem to have their own high standards. Overall, 83.8% (31/37) of the clinician papers mentioned that clinicians value the medical efficacy and risks of a tool. A total of 48.6% (18/37) of the clinician papers mentioned that they see the medical evidence base as an essential feature driving their adoption.

The importance of medical efficacy is intuitive, as it is the clinician's professional and ethical responsibility to provide the best care possible [114]. Therefore, we do not want to put this driver in a negative spotlight. But it does seem plausible that it drives clinicians to *not* prescribe a tool to patients who would have wanted to use it. This even goes for tools that provide medical benefit, but that do not meet the clinician's medical efficacy standards. Currently, there is no XR-specific research that supports this proposition. But there are indications that it happens for other medical tools, even if they have been approved by medical authorities [115,116].

The Clinician's Work Experience

As indicated in the value proposition domain, one benefit of XR-rehabilitation is that it offers autonomy to patients. Instead of having to visit the clinic, they can exercise at a time and location that suits them best. Developers can offer them these tools. It can even be argued that at-home tools are more attractive for developers because economies of scale are easier to achieve when all patients use them at home rather than share one in the clinic.

Various clinicians indicated that they also see enhanced patient autonomy as a benefit, meaning that their interests align with those of patients and developers in this case. However, there is also a set of clinicians for whom treating patients remotely makes their work less enjoyable. They can feel like a virtual coach or a patient mill, rather than a clinician. Using XR can also narrow their skillset to a niche. These losses can be quite substantial. But they are only incurred by clinicians, and not by patients and developers. Because of this, it is theoretically possible that clinicians may decide not to adopt XR, partially because doing so would impede their work experience.

However, in practice this is unlikely to happen as clinicians are highly bound by regulatory frameworks and ethical principles [117]. A more likely consequence is that clinicians will quit their jobs if they experience too much work displeasure [118]. This will not only prevent the patient from using XR but will also further increase staffing problems. Because of this, the impact of XR usage on the clinician's work experience should be an important consideration for both practitioners and researchers.

Limited Ability or Willingness to Pay Hinders the Developer's Economies of Scale

One paper illustrated that developers face "costs of development [that] are still pretty high at present" [94]. One sustainable way to recover these costs is to charge the patients and/or clinicians for the tool. However, these payments are a major adoption barrier for both actors. Naturally, patients prefer to pay as little as possible for the tool. Interestingly, clinicians share this perspective because high prices can increase the digital divide among patients [96]. However, clinicians are not always able or willing to carry the full financial burden by themselves. As a result, developers have fewer opportunities to reach sufficient economies of scale to recover their development costs. This plausibly prevents potential developers from entering the market or might cause existing developers to go out of business.

Theoretical and Practical Contributions

While previous reviews have been valuable in outlining what drives individual actors to not adopt XR-rehabilitation tools, none of them have investigated how their drivers align or misalign. By taking a more holistic point of view and combining these drivers, our findings suggest that a multi-actor view on adoption can provide additional insight into why XR-rehabilitation tools remain sparsely adopted.

We believe that this perspective may apply to other contexts as well. Adoption is also low for other rehabilitation tools, like the Wii, and for XR tools in other medical fields like palliative and mental health care [119-121]. Naturally, the specific reasons for (non-)adoption will vary across settings. For instance, autonomy benefits do not apply to an in-clinic setting, and using a personal phone requires no hardware acquisition. But the core dynamic remains consistent: tools can only be adopted on a large scale if all 3 actors opt in to it. As a result, we believe that a multi-actor perspective is valuable beyond the scope of this paper as well.

Additionally, both governments and entrepreneurs can leverage the concepts and findings in this paper to help reduce pressure on the health care system. For instance, they could consider developing new revenue models. Currently, most developers charge clinicians a 1-time purchase fee or a yearly subscription fee. These revenue models only allow clinicians with a sufficiently large user base to adopt XR tools. But there are other revenue models, such as performance-based risk-sharing arrangements [122] and managed entry agreements [123] that would allow clinicians to only pay for the tool when they use and/or benefit from it. This would also allow small-scale clinics to adopt XR tools. To date, the use of these models has been limited by barriers related to administration, evidence generation, and a lack of transparency [123]. Nonetheless, initiatives like Germany's new reimbursement process for digital health tools continue to show a positive impact on the adoption of these technologies [124]. Governments could also extend existing orphan drug legislation to relevant digital interventions [125,126]. Our paper also highlights ways in which XR developers can meet the needs of both patients and clinicians. But developers could also innovate beyond that by developing a (sharing) platform, like Netflix and Uber [127]. Such a centralized platform would increase game variety, reduce search time, and reduce per-patient hardware costs by enabling device sharing. To ensure that the development process of these innovations addresses the needs of all 3 actors, we strongly recommend adopting a collaborative approach. Cocreation practices seem especially suitable in this multi-actor context [22,128], and many methods are available. We therefore encourage stakeholders to consult the Health CASCADE database of collaborative methods to determine which one best suits their needs [129].

XSL•FO
RenderX

Limitations

Although XR-rehabilitation is a novel therapeutic approach, it has been studied across various contexts. To balance the richness and comparability of the included papers, our search query included a wide range of rehabilitation-, device-, and adoption-related search terms, but was also stringent by having a search block that captures behavioral drivers (see the Methods section for further details). Within this methodological trade-off, 2 limitations warrant consideration.

The first limitation is that it resulted in a set of papers with highly diverse research specifications in terms of the disorders treated, device types, intervention types, and usage location (see Multimedia Appendix 2). For example, Wray and Emery [100] conducted a survey where clinicians used an Oculus Quest 2 for drug or alcohol counseling, while Ploderer et al [72] conducted a mixed-methods study in patients with diabetes using their phone to track their foot ulcers at home. As most included studies were single-actor studies (n=49), the misalignments that we identified may potentially stem from these research differences, rather than true discrepancies between actors. At the same time, the low number of studies including all actors (n=1) also further corroborates the importance of future studies taking a multi-actor perspective.

The second limitation concerns the small number of developer papers. Because our review only included 4 such papers [22,94,96,106], the developer findings should be regarded as exploratory rather than conclusive. Although the overall body of developer-focused literature appears limited, the inclusion of the driver-related search terms may have contributed to the low number of included papers.

On top of these 2 query-related limitations, a third limitation is that only 31.3% (20/64) of the included papers used a theoretical framework to investigate adoption behavior. Of those 20 papers, no more than 7 (35%) used the same one (Multimedia Appendix 2). Although using an inductive research approach is valuable for providing new theoretical insights [130], the limited systematic use of concepts and definitions impedes the comparability across papers, thereby making it more difficult to unify them into a coherent conceptual overview.

A fourth limitation concerns the validation of the screening process. As illustrated in the results section, MS, LD, and JK validated the screening process for 219 of the 654 of the initially screened records (33.5%). In this validation sample, they identified 2 false exclusions and 1 false inclusion in the final set of included papers. An extrapolation of this to the nonvalidated set of papers implies that 4 papers may have been falsely excluded and 2 papers may have been falsely included. Although the screening process was conducted nearly perfectly in both the title/abstract stage (κ=0.83) and the full paper stage (κ=0.89), it is important to keep this potential limitation in mind.

The last limitation is that most papers were stringent in their respondent selection criteria by, for example, only selecting patients who had relatively little impairment. Because these people are more likely to adopt XR tools, this paper's findings might be overly optimistic about the gains of XR, thus minimizing potential losses.

Future Research Directions

As illustrated in the Results section, the current literature on the adoption of XR-rehabilitation tools contains various opportunities for future research. After reviewing these opportunities, we distilled them into 2 broad directions.

Investigating the Developer

As illustrated earlier, very little is known about the developer's role in the diffusion of XR-rehabilitation tools. To overcome this issue, we suggest that future studies focus on why, when, and how developers can(not) profitably develop and sell XR tools. Because health-tech developers face a wide array of stakeholders [131], we suggest that these papers take a holistic approach and do not solely focus on the interplay with patients and clinicians. Because questions of market entry and profitability also occur beyond health care, we encourage scholars to apply relevant frameworks from other research fields like strategic management. Some exemplary options are the 5 forces model [132], the Political, Economic, Social, Technological, Environmental, and Legal (PESTEL) framework [133], or the resource-based view [134]. Due to the currently limited insights, we also suggest initially using qualitative research methods to get a first understanding of these phenomena.

Investigating the 3 Actors Collectively

This work provides a first insight into the conflicts between patients, clinicians, and developers. However, these conflicts remain purely theoretical. Hence, we suggest that future scholars further investigate these 3 actors collectively. These works should focus on empirically validating the 3 conflicts in this paper, discovering other potential conflicts, and finding ways to overcome them.

There are numerous options to validate the 3 conflicts mentioned in this paper. Scholars could take a broad approach, where they determine whether patients and clinicians indeed have different evaluations of the medical benefits, fun, and the clinician's work experience. But they could also focus on specific features like price and immersion level. Due to the current technological constraints, scholars should also determine which value propositions and features developers are willing and able to develop. Otherwise, the demand cannot be fulfilled by the supply. Considering the current prevalence of qualitative research (see Multimedia Appendix 2), we recommend using more quantitative and experimental research methods for validation research [135,136].

Beyond validating the identified conflicts, future research should also explore other possible conflicts. For example, the developer's profit goals might incentivize them to make their games as engaging (and addictive) as possible [137]. Some patients might prefer this because they can have more fun. However, clinicians might not prescribe it to patients because they see risks associated with overtraining or misusing the therapy for sole enjoyment purposes. Conflicts may also arise when researchers incorporate additional stakeholders in the analysis. For example, insurance reimbursement is a key adoption driver for clinicians because it removes their payments for hardware and software. But a health care insurer's desire

XSL•FO
RenderX

for efficiency [138] may also pressure clinicians to have fewer physical meetings with patients, which hampers their work experience. Similarly, the high regulatory need to demonstrate safety and medical efficacy when requesting market authorization [139] likely demotivates developers from launching a product. We suggest using qualitative and collaborative methods to get an initial understanding of these theoretically possible but as of yet unexplored misalignments [128,129]. Given the limited insight into how adoption dynamics evolve over time (see Results; Domain 7: Embedding and Adaptation Over Time), we also encourage future research to use longitudinal research designs.

Conclusions

This review is the first to indicate that the limited adoption of XR-rehabilitation tools can likely not be fully explained by investigating patients, clinicians, and developers in isolation. Rather, we show that bringing these 3 actors together leads to a set of potential misalignments that help explain the low adoption level of these tools. Scholars can use this review as a basis to further investigate the conflicts between these 3 actors or to further investigate the developers' drivers to invest in XR-rehabilitation tools. Practitioners can also use these findings to create new revenue models, value propositions, and reimbursement regulations.

Acknowledgments

We would like to express our gratitude to Marcel Bogers, Maria Jacobs, Vincent Peters, Keith Murphy, Dian de Ridder, Pepijn Haverkorn van Rijsewijk, and attendants of the XR4REHAB and EHMA conferences in 2025 for their feedback and help in drafting earlier versions of this manuscript. We also declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing. The GAI tool used was: ChatGPT (GPT-4 and GPT-5 series). Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was funded by the Interreg North-West Europe program NWE0100082 'Scale-Up4Rehab' [https://scale-up4rehab.nweurope.eu]. The funder had no role in the study design, data collection and analyses, decision to publish, or preparation of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full search query.
[DOCX File, 24 KB - jmir_v28i1e80055_app1.docx]

Multimedia Appendix 2
Descriptives of all included papers.
[DOCX File, 67 KB - jmir_v28i1e80055_app2.docx]

Multimedia Appendix 3
A summary of the main drivers and the resulting misalignments.
[DOCX File, 25 KB - jmir_v28i1e80055_app3.docx]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 82 KB - jmir_v28i1e80055_app4.docx]

References

1. A universal truth: no health without a workforce. World Health Organization. 2014. URL: https://www.who.int/publications/m/item/hrh_universal_truth [accessed 2024-08-01]
2. Global strategy on human resources for health: workforce 2030. World Health Organization. 2016. URL: https://iris.who.int/handle/10665/250368 [accessed 2025-07-02]
3. Murray MK. The nursing shortage. Past, present, and future. J Nurs Adm 2002 Feb;32(2):79-84. [doi: 10.1097/00005110-200202000-00005] [Medline: 11984233]
4. Rehabilitation in health systems. World Health Organization. 2017. URL: https://iris.who.int/handle/10665/254506 [accessed 2024-03-04]

XSL•FO
RenderX

5. Rehabilitation in health systems: guide for action. World Health Organization. 2019. URL: https://iris.who.int/handle/10665/325607 [accessed 2024-03-04]
6. Het kabinet stelt tot € 108 miljoen beschikbaar voor meer technologie in zorg en welzijn. Rijksoverheid. 2024. URL: https://www.rijksoverheid.nl/actueel/nieuws/2024/04/18/het-kabinet-stelt-tot-€108-miljoen-beschikbaar-voor-meer-technologie-in-zorg--en-welzijn [accessed 2024-10-07]
7. Rauschnabel PA, Felix R, Hinsch C, Shahab H, Alt F. What is XR? towards a framework for augmented and virtual reality. Comput Human Behav 2022 Aug;133:107289. [doi: 10.1016/j.chb.2022.107289]
8. Milgram P, Kishino F. A taxonomy of mixed reality visual displays. IEICE Trans Inf Syst 1994;77(12):1321-1329 [FREE Full text]
9. Laver KE, Lange B, George S, et al. Virtual reality for stroke rehabilitation. Cochrane Database Syst Rev 2025 Jun 20;6(6):CD008349. [doi: 10.1002/14651858.CD008349.pub5] [Medline: 40537150]
10. Slatman S, Ostelo R, Goor HV, Satink T, Staal JB, Knoop J. “A new tool for my toolbox”: physiotherapists’ perceptions on therapeutic virtual reality for chronic low back pain - a qualitative descriptive study. Musculoskelet Sci Pract 2026 Feb;81:103475. [doi: 10.1016/j.msksp.2025.103475] [Medline: 41468807]
11. Felnhofer A, Pfannerstill F, Gänsler L, et al. Barriers to adopting therapeutic virtual reality: the perspective of clinical psychologists and psychotherapists. Front Psychiatry 2025;16:1549090. [doi: 10.3389/fpsyt.2025.1549090] [Medline: 40171310]
12. Felsberg DT, McGuirt JT, Ross SE, Raisbeck LD, Howard CK, Rhea CK. Clinician perspectives on virtual reality use in physical therapy practice in the United States. PLoS One 2025;20(4):e0320215. [doi: 10.1371/journal.pone.0320215] [Medline: 40173138]
13. Shiner CT, Croker G, McGhee J, Faux SG. Perspectives on the use of virtual reality within a public hospital setting: surveying knowledge, attitudes, and perceived utility among health care professionals. BMC Digit Health 2024;2(1):18. [doi: 10.1186/s44247-024-00076-x]
14. Glegg SMN, Levac DE. Barriers, facilitators and interventions to support virtual reality implementation in rehabilitation: a scoping review. PM R 2018 Nov;10(11):1237-1251. [doi: 10.1016/j.pmrj.2018.07.004] [Medline: 30503231]
15. Brepohl PCA, Leite H. Virtual reality applied to physiotherapy: a review of current knowledge. Virtual Real 2023 Mar;27(1):71-95. [doi: 10.1007/s10055-022-00654-2]
16. Bulle-Smid L, van den Heuvel R, Keuning W, Hakvoort G, Daniëls R, Verhoeven F. The use of extended reality in rehabilitation for patients with acquired brain injury: a scoping review. Stud Health Technol Inform 2025 Feb;37(1):3-44. [doi: 10.1177/10554181241290963]
17. Arrow KJ. Uncertainty and the welfare economics of medical care. In: Uncertainty in Economics: Elsevier; 1978:345-375. [doi: 10.1016/B978-0-12-214850-7.50028-0]
18. Scott A. Economics of general practice. In: Handbook of Health Economics: Elsevier, Vol. 2000:1175-1200. [doi: 10.1016/S1574-0064(00)80035-9]
19. Gottschalk F, Mimra W, Waibel C. Health services as credence goods: a field experiment. The Economic Journal 2020 Jul 1;130(629):1346-1383. [doi: 10.1093/ej/ueaa024]
20. Smith PC, Stepan A, Valdmanis V, Verheyen P. Principal-agent problems in health care systems: an international perspective. Health Policy 1997 Jul;41(1):37-60. [doi: 10.1016/s0168-8510(97)00012-2] [Medline: 10169061]
21. Smith A. The Wealth of Nations: W. Strahan and T. Cadell; 2012.
22. Saidi T, Hestetun-Mandrup AM, Pavel N, et al. Challenges and benefits of virtual reality in home-based poststroke rehabilitation: co-design qualitative study. JMIR Hum Factors 2026 Jan 15;13:e78798. [doi: 10.2196/78798] [Medline: 41539665]
23. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018 Nov 19;18(1):143. [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
24. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
25. Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. Syst Rev 2017 Dec 6;6(1):245. [doi: 10.1186/s13643-017-0644-y] [Medline: 29208034]
26. Wade DT. What is rehabilitation? an empirical investigation leading to an evidence-based description. Clin Rehabil 2020 May;34(5):571-583. [doi: 10.1177/0269215520905112] [Medline: 32037876]
27. Battel I, Arienti C, Levack W, Kiekens C, Negrini S. The new rehabilitation definition for research purposes could improve rehabilitation description in Cochrane Systematic Reviews. Cochrane Evid Synth Methods 2024 Sep;2(9):e70000. [doi: 10.1002/cesm.70000] [Medline: 40474874]
28. Negrini S, Selb M, Kiekens C, et al. Rehabilitation definition for research purposes: a global stakeholders’ initiative by Cochrane rehabilitation. Am J Phys Med Rehabil 2022 Jul 1;101(7):e100-e107. [doi: 10.1097/PHM.0000000000002031] [Medline: 35583514]

XSL•FO
RenderX

29. Wade DT. Defining rehabilitation: an exploration of why it is attempted, and why it will always fail. Clin Rehabil 2021 Dec;35(12):1650-1656. [doi: 10.1177/02692155211028018] [Medline: 34182808]
30. Aderinto N, Olatunji G, Abdulbasit MO, Edun M, Aboderin G, Egbunu E. Exploring the efficacy of virtual reality-based rehabilitation in stroke: a narrative review of current evidence. Ann Med 2023;55(2):2285907. [doi: 10.1080/07853890.2023.2285907] [Medline: 38010358]
31. Haghedooren E, Haghedooren R, Langer D, Gosselink R. Feasibility and safety of interactive virtual reality upper limb rehabilitation in patients with prolonged critical illness. Aust Crit Care 2024 Nov;37(6):949-956. [doi: 10.1016/j.aucc.2024.06.004] [Medline: 39054204]
32. Høeg ER, Povlsen TM, Bruun-Pedersen JR, et al. System immersion in virtual reality-based rehabilitation of motor function in older adults: a systematic review and meta-analysis. Frontiers in Virtual Reality 2021;2:647993. [doi: 10.3389/frvir.2021.647993]
33. Opara Zupančič M, Šarabon N. The current state of virtual reality in the management of musculoskeletal conditions and associated chronic pain: terminology, technology, and associations. Appl Sci (Basel) 2025;15(5):2564. [doi: 10.3390/app15052564]
34. Proctor E, Silmere H, Raghavan R, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011 Mar;38(2):65-76. [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
35. Markus ML. Paradigm shifts - E-business and business/systems integration. Communications of the Association for Information Systems 2000;4. [doi: 10.17705/1CAIS.00410]
36. Cramer JA, Roy A, Burrell A, et al. Medication compliance and persistence: terminology and definitions. Value Health 2008;11(1):44-47. [doi: 10.1111/j.1524-4733.2007.00213.x] [Medline: 18237359]
37. Eldridge SM, Lancaster GA, Campbell MJ, et al. Defining Feasibility and Pilot Studies in Preparation for Randomised Controlled Trials: Development of a Conceptual Framework. PLoS ONE 2016;11(3):e0150205. [doi: 10.1371/journal.pone.0150205] [Medline: 26978655]
38. Ambler SW. Process Patterns: Building Large-Scale Systems Using Object Technology: Cambridge University Press; 1998.
39. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5(1):69. [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
40. Davis K, Drey N, Gould D. What are scoping studies? a review of the nursing literature. Int J Nurs Stud 2009 Oct;46(10):1386-1400. [doi: 10.1016/j.ijnurstu.2009.02.010] [Medline: 19328488]
41. Nilsen P. Making sense of implementation theories, models and frameworks. Implement Sci 2015 Apr 21;10(1):53. [doi: 10.1186/s13012-015-0242-0] [Medline: 25895742]
42. O'Connor C, Joffe H. Intercoder reliability in qualitative research: debates and practical guidelines. Int J Qual Methods 2020 Jan 1;19:1609406919899220. [doi: 10.1177/1609406919899220]
43. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012;22(3):276-282. [doi: 10.11613/BM.2012.031]
44. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr;20(1):37-46. [doi: 10.1177/001316446002000104]
45. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
46. Greenhalgh T, Abimbola S. The NASSS framework - a synthesis of multiple theories of technology implementation. Stud Health Technol Inform 2019 Jul 30;263:193-204. [doi: 10.3233/SHTI190123] [Medline: 31411163]
47. Nilsen P. Evaluation of implementation: a taxonomy of implementation science theories, models, and frameworks. In: Kaehne A, Feather J, editors. Handbook of Health Services Evaluation Theories, Methods and Innovative Practices: Springer Nature; 2025:978-973. [doi: 10.1007/978-3-031-87869-5_9]
48. Gray CS, Shen N, Pham Q, Wiljer D. Evaluating digitally enabled health services. In: Kaehne A, Feather J, editors. Handbook of Health Services Evaluation Theories, Methods and Innovative Practices: Springer Nature; 2025:309-330. [doi: 10.1007/978-3-031-87869-5_17]
49. Colquhoun HL, Levac D, O'Brien KK, et al. Scoping reviews: time for clarity in definition, methods, and reporting. J Clin Epidemiol 2014 Dec;67(12):1291-1294. [doi: 10.1016/j.jclinepi.2014.03.013] [Medline: 25034198]
50. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 2008 May;61(Pt 1):29-48. [doi: 10.1348/000711006X126600] [Medline: 18482474]
51. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol 2013 Apr 29;13(1):61. [doi: 10.1186/1471-2288-13-61] [Medline: 23627889]
52. Namey E, Guest G, Thairu L. Data reduction techniques for large qualitative data sets. In: Guest G, MacQueen K, editors. Handbook for Team-Based Qualitative Research: AltaMira Press; 2008:137-161. [doi: 10.5040/9798216409670]
53. Sandelowski M. Real qualitative researchers do not count: the use of numbers in qualitative research. Res Nurs Health 2001 Jun;24(3):230-240. [doi: 10.1002/nur.1025] [Medline: 11526621]

XSL•FO
RenderX

54. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]
55. Luc A, Lambricht N, Aujoulat I, Detrembleur C, Pitance L. Experiences of people with persistent nonspecific neck pain who used immersive virtual reality serious games in the home setting: a qualitative study. Phys Ther 2025 Mar 3;105(3):pzae149. [doi: 10.1093/ptj/pzae149] [Medline: 39388230]
56. Rathinam C, Farr W, Ray D, Gupta R. Factors influencing virtual reality use in paediatric acquired brain injury upper limb rehabilitation: a qualitative study. BMJ Open 2025 Jan;15(1):e083120. [doi: 10.1136/bmjopen-2023-083120]
57. Moeinzadeh AM, Calder A, Petersen C, Hoermann S. Views of people with MS regarding VR-exergaming to improve physical function and cognition: a qualitative study. Disabil Rehabil 2025 Oct;47(20):5348-5357. [doi: 10.1080/09638288.2025.2461265] [Medline: 39976253]
58. Jehl NM, Hess CW, Choate ES, Nguyen HT, Yang Y, Simons LE. Navigating virtual realities: identifying barriers and facilitators to implementing VR-enhanced PT for youth with chronic pain. J Pediatr Psychol 2025 Jan 1;50(1):76-85. [doi: 10.1093/jpepsy/jsae056]
59. Dupont L, Raucher-Chéné D, Mercier N, et al. Digital travel using virtual reality in inpatient psychiatric care: focus group exploration of perspectives from individuals with lived experience. Psychiatr Rehabil J 2026 Mar;49(1):116-124. [doi: 10.1037/prj0000646] [Medline: 40310207]
60. de Vries M, Beumeler LFE, van der Meulen J, Bethlehem C, den Otter R, Boerma EC. The feasibility of virtual reality therapy for upper extremity mobilization during and after intensive care unit admission. PeerJ 2025;13:e18461. [doi: 10.7717/peerj.18461] [Medline: 39802193]
61. Stam D, Watkins M, Galyen B, Olson M, Vomhof K. Feasibility, acceptability, and tolerability of fully-immersive virtual reality use among individuals with neurological conditions: a mixed methods study. Games Health J 2025 Dec;14(6):484-492. [doi: 10.1177/2161783X251361941] [Medline: 40711833]
62. Robbemond LM, Nijland JM, Alma M, Veling W, van Driel CM. Exploring the barriers to and facilitators of using virtual reality relaxation for patients with psychiatric problems: qualitative focus group study. J Med Internet Res 2025 Jun 11;27:e65308. [doi: 10.2196/65308] [Medline: 40497635]
63. Miegel F, Lindner P, Schultz J, Lohse L, Yassari AH, Jelinek L. Breaking barriers? Targeting willingness to improve engagement in exposure therapy through a disorder-nonspecific exposure session in virtual reality. Curr Psychol 2025 Mar;44(5):3210-3226. [doi: 10.1007/s12144-024-07194-9]
64. Hardeman LES, van Benten E, Hoogendoorn EM, et al. Home-based augmented reality exercise for people with Parkinson disease: qualitative acceptability study. JMIR Rehabil Assist Technol 2025 Oct 6;12:e70802. [doi: 10.2196/70802] [Medline: 41052434]
65. McDaniel S, Pines R, He EM, et al. Virtual reality therapeutic art for adult hospitalized patients and caregivers: a mixed-methods project. J Patient Exp 2025;12:23743735251330467. [doi: 10.1177/23743735251330467] [Medline: 40144534]
66. Ramos A, Boisvert M, Traverse E, et al. Bridging needs and expectations of individuals with physical disabilities and community services stakeholders for the cocreation of an adapted physical activity platform in virtual reality: qualitative study. JMIR Serious Games 2025 Jan 17;13:e59704. [doi: 10.2196/59704] [Medline: 39829130]
67. Hewko M, Gagnon Shaigetz V, Smith MS, et al. Considering theory-based gamification in the co-design and development of a virtual reality cognitive remediation intervention for depression (bWell-D): mixed methods study. JMIR Serious Games 2025 Mar 31;13:e59514. [doi: 10.2196/59514] [Medline: 40163852]
68. King CD, Olbrecht VA, Crowley SL, et al. “Virtual reality fixed me”: a case report of the use of virtual reality during intensive interdisciplinary pain treatment. J Pediatr Rehabil Med 2024 Sep;17(3):376-384. [doi: 10.3233/PRM-230059] [Medline: 39150838]
69. Groenveld TD, Smits IG, Scholten N, de Vries M, van Goor H, Stirler VM. Pulmonary and physical virtual reality exercises for patients with blunt chest trauma: randomized clinical trial. JMIR Serious Games 2024 Dec 9;12:e54389. [doi: 10.2196/54389] [Medline: 39652866]
70. Wu SC, Chuang CW, Liao WC, Li CF, Shih HH. Using virtual reality in a rehabilitation program for patients with breast cancer: phenomenological study. JMIR Serious Games 2024 Apr 16;12:e44025. [doi: 10.2196/44025] [Medline: 38634461]
71. Krishnan S, Mandala MA, Wolf SL, Howard A, Kesar TM. Perceptions of stroke survivors regarding factors affecting adoption of technology and exergames for rehabilitation. PM R 2023 Nov;15(11):1403-1410. [doi: 10.1002/pmrj.12963] [Medline: 36787167]
72. Ploderer B, Clark D, Brown R, Harman J, Lazzarini PA, Van Netten JJ. Self-monitoring diabetes-related foot ulcers with the MyFootCare app: a mixed methods study. Sensors (Basel) 2023 Feb 24;23(5):2547. [doi: 10.3390/s23052547] [Medline: 36904750]
73. Brown L, DiCenso-Fleming T, Ensign T, Boyd AJ, Monaghan G, Binder DS. Chronic pain education delivered with a virtual reality headset in outpatient physical therapy clinics: a multi-site exploratory trial. Am J Transl Res 2023;15(5):3500-3510. [Medline: 37303618]

XSL•FO
RenderX

74. Vaezipour A, Aldridge D, Koenig S, et al. Rehabilitation supported by immersive virtual reality for adults with communication disorders: semistructured interviews and usability survey study. JMIR Rehabil Assist Technol 2023 Oct 31;10:e46959. [doi: 10.2196/46959] [Medline: 37906228]
75. Cioeta M, Pournajaf S, Goffredo M, Giovannico G, Franceschini M. Improving adherence to a home rehabilitation plan for chronic neck pain through immersive virtual reality: a case report. J Clin Med 2023 Feb 28;12(5):1926. [doi: 10.3390/jcm12051926] [Medline: 36902713]
76. Tuck N, Pollard C, Good C, et al. Active virtual reality for chronic primary pain: mixed methods randomized pilot study. JMIR Form Res 2022 Jul 13;6(7):e38366. [doi: 10.2196/38366] [Medline: 35830224]
77. Cerdán de Las Heras J, Tulppo M, Kiviniemi AM, et al. Augmented reality glasses as a new tele-rehabilitation tool for home use: patients' perception and expectations. Disabil Rehabil Assist Technol 2022 May;17(4):480-486. [doi: 10.1080/17483107.2020.1800111] [Medline: 32750254]
78. Reilly CA, Greeley AB, Jevsevar DS, Gitajn IL. Virtual reality-based physical therapy for patients with lower extremity injuries: feasibility and acceptability. OTA Int 2021 Jun;4(2):e132. [doi: 10.1097/OI9.0000000000000132] [Medline: 34746664]
79. Mayer G, Gronewold N, Polte K, et al. Experiences of patients and therapists testing a virtual reality exposure app for symptoms of claustrophobia: mixed methods study. JMIR Ment Health 2022 Dec 5;9(12):e40056. [doi: 10.2196/40056] [Medline: 36469413]
80. Sánchez-Herrera-Baeza P, Cano-de-la-Cuerda R, Oña-Simbaña ED, et al. The impact of a novel immersive virtual reality technology associated with serious games in Parkinson's disease patients on upper limb rehabilitation: a mixed methods intervention study. Sensors (Basel) 2020 Apr 11;20(8):2168. [doi: 10.3390/s20082168] [Medline: 32290517]
81. Tennant M, McGillivray J, Youssef GJ, McCarthy MC, Clark TJ. Feasibility, acceptability, and clinical implementation of an immersive virtual reality intervention to address psychological well-being in children and adolescents with cancer. J Pediatr Oncol Nurs 2020;37(4):265-277. [doi: 10.1177/1043454220917859] [Medline: 32536320]
82. MacIntyre E, Saredakis D, Stanford T, et al. VRCycle: the clinical translation of a virtual reality cycling program to enhance exercise engagement in people with knee osteoarthritis. Psychol Sport Exerc 2026 Jan;82:103009. [doi: 10.1016/j.psychsport.2025.103009] [Medline: 41151634]
83. Linge AD, Øyeflaten I, Johansen T, et al. Virtual reality gaming in occupational rehabilitation - user experiences and perceptions. J Multidiscip Healthc 2024;17:5139-5149. [doi: 10.2147/JMDH.S478052] [Medline: 39558923]
84. Rothgangel A, Braun S, Smeets R, Beurskens A. Feasibility of a traditional and teletreatment approach to mirror therapy in patients with phantom limb pain: a process evaluation performed alongside a randomized controlled trial. Clin Rehabil 2019 Oct;33(10):1649-1660. [doi: 10.1177/0269215519846539] [Medline: 31066315]
85. Kenea CD, Abessa TG, Lamba D, Bonnechère B. AdaptRehab VR: development of an immersive virtual reality system for upper limb stroke rehabilitation designed for low- and middle-income countries using a participatory co-creation approach. Bioengineering (Basel) 2025 May 28;12(6):581. [doi: 10.3390/bioengineering12060581] [Medline: 40564398]
86. Álvarez-Aguado I, Vega Córdova V, Muñoz La Rivera F, et al. Exploring technology use among older adults with intellectual disabilities: barriers, opportunities, and the role of advanced technologies. Disabil Rehabil Assist Technol 2025 Oct;20(7):2249-2260. [doi: 10.1080/17483107.2025.2498566] [Medline: 40293959]
87. Twamley J, Hamer O, Hill J, et al. Exploring the perceptions of former ICU patients and clinical staff on barriers and facilitators to the implementation of virtual reality exposure therapy: a qualitative study. Nurs Crit Care 2024 Mar;29(2):313-324. [doi: 10.1111/nicc.12868] [Medline: 36458458]
88. Choukou MA, He E, Moslenko K. Feasibility of a virtual-reality-enabled at-home telerehabilitation program for stroke survivors: a case study. J Pers Med 2023 Aug 3;13(8):1230. [doi: 10.3390/jpm13081230] [Medline: 37623481]
89. Seaborn K, Pennefather P, Fels DI. "Learn what we're going through": attitudes of older powered chair users towards mixed reality games that involve power mobility. Univ Access Inf Soc 2016 Nov;15(4):699-711. [doi: 10.1007/s10209-015-0450-z]
90. Kühne Escolà J, Demirdas R, Schulze M, et al. Virtual reality-guided therapy on a stroke unit: a feasibility study. Neurol Res Pract 2024;6(1):60. [doi: 10.1186/s42466-024-00357-w]
91. Schreiter M, Hennrich J, Wolf AL, Eymann T. The influence of previous experience on virtual reality adoption in medical rehabilitation and overcoming knowledge gaps among health care professionals: qualitative interview study. J Med Internet Res 2025 Apr 30;27:e62649. [doi: 10.2196/62649] [Medline: 40305102]
92. Mills J, Duffy O. Speech and language therapists' perspectives of virtual reality as a clinical tool for autism: cross-sectional survey. JMIR Rehabil Assist Technol 2025 Feb 27;12:e63235. [doi: 10.2196/63235] [Medline: 40014826]
93. Roy I, Salles J, Neveu E, et al. Exploring the perspectives of health care professionals on digital health technologies in pediatric care and rehabilitation. J Neuroeng Rehabil 2024 Sep 12;21(1):156. [doi: 10.1186/s12984-024-01431-9] [Medline: 39261920]
94. Bradwell HL, Cooper L, Edwards KJ, et al. Staff perceptions towards virtual reality-motivated treadmill exercise for care home residents: a qualitative feedback study with key stakeholders and follow-up interview with technology developer. BMJ Open 2023 Nov 23;13(11):e073307. [doi: 10.1136/bmjopen-2023-073307] [Medline: 37996232]

XSL•FO
RenderX

95. Brady N, Dejaco B, Lewis J, McCreesh K, McVeigh JG. Physiotherapist beliefs and perspectives on virtual reality supported rehabilitation for the management of musculoskeletal shoulder pain: a focus group study. PLoS One 2023;18(4):e0284445. [doi: 10.1371/journal.pone.0284445] [Medline: 37058507]
96. Brassel S, Brunner M, Power E, Campbell A, Togher L. Speech-language pathologists' views of using virtual reality for managing cognitive-communication disorders following traumatic brain injury. Am J Speech Lang Pathol 2023 Mar 23;32(2S):907-923. [doi: 10.1044/2022_AJSLP-22-00077] [Medline: 36580534]
97. Elor A, Conde S, Powell M, Robbins A, Chen NN, Kurniawan S. Physical therapist impressions of telehealth and virtual reality needs amidst a pandemic. Frontiers in Virtual Reality 2022;3:915332. [doi: 10.3389/frvir.2022.915332]
98. Chung OS, Robinson T, Johnson AM, et al. Implementation of therapeutic virtual reality into psychiatric care: clinicians' and service managers' perspectives. Front Psychiatry 2022;12:791123. [doi: 10.3389/fpsyt.2021.791123] [Medline: 35058823]
99. Vaezipour A, Aldridge D, Koenig S, Theodoros D, Russell T. "It's really exciting to think where it could go": a mixed-method investigation of clinician acceptance, barriers and enablers of virtual reality technology in communication rehabilitation. Disabil Rehabil 2022 Jul 17;44(15):3946-3958. [doi: 10.1080/09638288.2021.1895333]
100. Wray TB, Emery NN. Feasibility, appropriateness, and willingness to use virtual reality as an adjunct to counseling among addictions counselors. Subst Use Misuse 2022;57(9):1470-1477. [doi: 10.1080/10826084.2022.2092148] [Medline: 35754378]
101. Greenhalgh M, Fitzpatrick C, Rodabaugh T, et al. Assessment of task demand and usability of a virtual reality-based rehabilitation protocol for combat related traumatic brain injury from the perspective of veterans affairs healthcare providers: a pilot study. Frontiers in Virtual Reality 2021;2:741578. [doi: 10.3389/frvir.2021.741578]
102. Ogourtsova T, Archambault PS, Lamontagne A. Exploring barriers and facilitators to the clinical use of virtual reality for post-stroke unilateral spatial neglect assessment. Disabil Rehabil 2019 Feb;41(3):284-292. [doi: 10.1080/09638288.2017.1387292] [Medline: 29111838]
103. Pearce LMN, Hassett L, Sherrington C, Pryor J. Human interactions remain at the heart of rehabilitation with advanced technology: a practice-embedded longitudinal qualitative study with allied health clinicians. J Neuroeng Rehabil 2025 Mar 7;22(1):52. [doi: 10.1186/s12984-025-01576-1] [Medline: 40050924]
104. Aloyuni S, Alharbi R, Kashoo F, et al. Knowledge, attitude, and barriers to telerehabilitation-based physical therapy practice in Saudi Arabia. Healthcare (Basel) 2020 Nov 4;8(4):460. [doi: 10.3390/healthcare8040460] [Medline: 33158298]
105. Nicora G, Parimbelli E, Mauro MC, et al. Healthcare practitioners and robotic-assisted rehabilitation: understanding needs and barriers. J NeuroEngineering Rehabil 2025;22(1):78. [doi: 10.1186/s12984-025-01593-0]
106. Kulkov I, Tsvetkova A, Ivanova-Gongne M. Identifying institutional barriers when implementing new technologies in the healthcare industry. EJIM 2023 Jun 8;26(4):909-932. [doi: 10.1108/EJIM-02-2021-0093]
107. Alrashidi M, Tomlinson RJ, Buckingham G, Williams CA. Virtual reality current use, facilitators and barriers to implementation in paediatric physiotherapy: cross-sectional online survey of UK paediatric physiotherapists. Disabil Rehabil Assist Technol 2025 Apr;20(3):585-591. [doi: 10.1080/17483107.2024.2393695] [Medline: 39180393]
108. Cho HM, Kim H, Jang J, et al. Attitude toward virtual rehabilitation and active video games among therapists in Korea: a nationwide survey. Brain Neurorehabil 2024 Mar;17(1):e4. [doi: 10.12786/bn.2024.17.e4] [Medline: 38585030]
109. Levac DE, Glegg S, Colquhoun H, Miller P, Noubary F. Virtual reality and active videogame-based practice, learning needs, and preferences: a cross-Canada survey of physical therapists and occupational therapists. Games Health J 2017 Aug;6(4):217-228. [doi: 10.1089/g4h.2016.0089] [Medline: 28816511]
110. Mensah-Gourmel J, Bekteshi S, Brochard S, et al. Digital technologies for pediatric rehabilitation: current access and use in the European Rehatech4child survey. Eur J Phys Rehabil Med 2024 Dec;60(6):970-979. [doi: 10.23736/S1973-9087.24.08559-9] [Medline: 39387853]
111. Polechoński J, Langer A, Akbaş A, Zwierzchowska A. Application of immersive virtual reality in the training of wheelchair boxers: evaluation of exercise intensity and users experience additional load– a pilot exploratory study. BMC Sports Sci Med Rehabil 2024;16(1):80. [doi: 10.1186/s13102-024-00878-6]
112. Hosseini M, Thomas R, Pilutti L, Fallavollita P, Jutai JW. Assessing virtual reality acceptance in long-term care facilities: a quantitative study with older adults. Disabil Rehabil Assist Technol 2024 Oct;19(7):2602-2614. [doi: 10.1080/17483107.2023.2295946] [Medline: 38146956]
113. Boots B, Berg D, Hewitt E, Naugle K, Naugle K. Physical activity and enjoyment in active virtual reality games in youth: comparative analysis of gorilla tag and beat saber. JMIR Serious Games 2025 Apr 1;13:e66593. [doi: 10.2196/66593] [Medline: 40168642]
114. Beauchamp TL. Methods and principles in biomedical ethics. J Med Ethics 2003 Oct;29(5):269-274. [doi: 10.1136/jme.29.5.269] [Medline: 14519835]
115. Sekimoto M, Imanaka Y, Kitano N, Ishizaki T, Takahashi O. Why are physicians not persuaded by scientific evidence? a grounded theory interview study. BMC Health Serv Res 2006 Jul 27;6(1):92. [doi: 10.1186/1472-6963-6-92] [Medline: 16872522]

XSL•FO
RenderX

116. Steenhuysen J. Alzheimer's drug adoption in US slowed by doctors' skepticism. Reuters. 2024. URL: https://www.reuters.com/business/healthcare-pharmaceuticals/alzheimers-drug-adoption-us-slowed-by-doctors-skepticism-2024-04-23/ [accessed 2025-04-30]
117. Randall DM, Gibson AM. Ethical decision making in the medical profession: an application of the theory of planned behavior. J Bus Ethics 1991 Feb;10(2):111-122. [doi: 10.1007/BF00383614]
118. Irvine DM, Evans MG. Job satisfaction and turnover among nurses. Nurs Res 1995 Jul;44(4):246. [doi: 10.1097/00006199-199507000-00010]
119. Bin S, Alrashdi DH, Whitehead T, Riches S, Drini E. Mental health professionals' attitudes towards virtual reality therapies: a systematic review. Journal of Technology in Behavioral Science 2025;11(1):146-169. [doi: 10.1007/s41347-025-00499-x]
120. Huang Y, Deng C, Peng M, Hao Y. Experiences and perceptions of palliative care patients receiving virtual reality therapy: a meta-synthesis of qualitative studies. BMC Palliat Care 2024;23(1):182. [doi: 10.1186/s12904-024-01520-5]
121. Sawa R, Saitoh M, Morisawa T, et al. The potential application of commercially available active video games to cardiac rehabilitation: scoping review. JMIR Serious Games 2022 Mar 18;10(1):e31974. [doi: 10.2196/31974] [Medline: 35302503]
122. Carlson JJ, Chen S, Garrison LP. Performance-based risk-sharing arrangements: an updated international review. Pharmacoeconomics 2017 Oct;35(10):1063-1072. [doi: 10.1007/s40273-017-0535-z] [Medline: 28695544]
123. Greco A, Frederix GWJ, Hooft L, Ten Ham RMT. A systematic review of challenges and opportunities in the implementation of managed entry agreements for advanced therapy medicinal products. Clin Ther 2025 Feb;47(2):e16-e26. [doi: 10.1016/j.clinthera.2024.11.019] [Medline: 39706763]
124. Goeldner M, Gehder S. Digital health applications (DiGAs) on a fast track: insights from a data-driven analysis of prescribable digital therapeutics in Germany from 2020 to mid-2024. J Med Internet Res 2024 Aug 29;26:e59013. [doi: 10.2196/59013] [Medline: 39208415]
125. Drummond MF, Wilson DA, Kanavos P, Ubel P, Rovira J. Assessing the economic challenges posed by orphan drugs. Int J Technol Assess Health Care 2007;23(1):36-42. [doi: 10.1017/S0266462307051550] [Medline: 17234015]
126. Herder M. What Is the purpose of the Orphan Drug Act? PLoS Med 2017 Jan;14(1):e1002191. [doi: 10.1371/journal.pmed.1002191] [Medline: 28045908]
127. Schiavone F, Mancini D, Leone D, Lavorato D. Digital business models and ridesharing for value co-creation in healthcare: a multi-stakeholder ecosystem analysis. Technological Forecasting and Social Change 2021 May;166:120647. [doi: 10.1016/j.techfore.2021.120647]
128. Frow P, McColl-Kennedy JR, Payne A. Co-creation practices: their role in shaping a health care ecosystem. Ind Mark Manag 2016 Jul;56:24-39. [doi: 10.1016/j.indmarman.2016.03.007]
129. Agnello DM, Loisel QEA, An Q, et al. Establishing a health CASCADE–curated open-access database to consolidate knowledge about co-creation: novel artificial intelligence–assisted methodology based on systematic reviews. J Med Internet Res 2023 Jul 18;25:e45059. [doi: 10.2196/45059] [Medline: 37463024]
130. Glaser BG, Strauss AL. Discovery of Grounded Theory; Strategies for Qualitative Research: Routledge; 2017. [doi: 10.4324/9780203793206]
131. Singh N, Alshibani SM, Misra P, Nawaz R, Gupta B. Unravelling barriers in high-tech technology start-ups: practical insights and solutions for healthcare enterprises. Journal of Enterprise Information Management 2023 Nov 8;36(6):1708-1726. [doi: 10.1108/JEIM-06-2023-0318]
132. Porter ME. How competitive forces shape strategy. Harv Bus Rev 1979;57(2):137-145 [FREE Full text]
133. Aguilar FJ. Scanning the Business Environment: MacMillan; 1967.
134. Wernerfelt B. A resource - based view of the firm. Strateg Manag J 1984 Apr;5(2):171-180 [FREE Full text] [doi: 10.1002/smj.4250050207]
135. Green PE, Srinivasan V. Conjoint analysis in marketing: new developments with implications for research and practice. J Mark 1990 Oct;54(4):3-19. [doi: 10.1177/002224299005400402]
136. Harrison GW, List JA. Field experiments. J Econ Lit 2004 Nov 1;42(4):1009-1055. [doi: 10.1257/0022051043004577]
137. Balakrishnan J, Griffiths MD. Loyalty towards online games, gaming addiction, and purchase intention towards online mobile in-game features. Comput Human Behav 2018 Oct;87:238-246. [doi: 10.1016/j.chb.2018.06.002]
138. Schut FT, Henschke C, Or Z. Changing roles of health insurers in France, Germany, and the Netherlands: any lessons to learn from Bismarckian systems? Health Econ Policy Law 2023 Oct;18(4):362-376. [doi: 10.1017/S1744133123000191] [Medline: 37675507]
139. Iqbal JD, Biller-Andorno N. The regulatory gap in digital health and alternative pathways to bridge it. Health Policy Technol 2022 Sep;11(3):100663. [doi: 10.1016/j.hlpt.2022.100663]

Abbreviations

AR: augmented reality
HMD: head-mounted device
MR: mixed reality

XSL•FO
RenderX

NASSS framework: Nonadoption, Abandonment, and challenges to the Scale-up, Spread, and Sustainability framework
PESTEL framework: Political, Economic, Social, Technological, Environmental, and Legal framework
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Scoping Reviews
VR: virtual reality
XR: extended reality

Edited by A Coristine; submitted 03.Jul.2025; peer-reviewed by J Thrift, L Sabbe; revised version received 10.Apr.2026; accepted 10.Apr.2026; published 20.May.2026.

Please cite as:
Baltissen L, Stienstra M, Denoo L, Knoben J
Barriers and Facilitators to the 3 Sides of Extended Reality-Rehabilitation Adoption: Scoping Review
J Med Internet Res 2026;28:e80055
URL: https://www.jmir.org/2026/1/e80055
doi:10.2196/80055

© Luuk Baltissen, Miranda Stienstra, Lien Denoo, Joris Knoben. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Explainable AI in Cancer Imaging: Scoping Review of Methods, Modalities, and Clinical Integration

Dimitris Fotopoulos[1], MSc; Ioannis Ladakis[1], MSc; Dimitrios Filos[1], PhD; Pedro A Moreno-Sánchez[2,3], PhD; Mark van Gils[2], PhD; Ioanna Chouvarda[1], PhD

[1]Laboratory of Computing, Medical Informatics and Biomedical - Imaging Technologies, School of Medicine, Aristotle University of Thessaloniki, AUTh Campus, Thessaloniki, Greece

[2]Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland

[3]Basque Research & Technology Alliance (BRTA), TECNALIA, Derio, Basque Country, Spain

Corresponding Author:
Dimitris Fotopoulos, MSc
Laboratory of Computing, Medical Informatics and Biomedical - Imaging Technologies, School of Medicine, Aristotle University of Thessaloniki, AUTh Campus, Thessaloniki, Greece

Abstract

Background: Explainable artificial intelligence (xAI) is increasingly used in medical imaging to enhance transparency, clinical interpretability, and trust in artificial intelligence (AI)–assisted diagnostics, particularly in oncology. Evidence on how explainability is implemented, validated, and reported in cancer imaging remains fragmented.

Objective: This scoping review aimed to systematically map research applying xAI methods to radiologic cancer imaging, summarize methodological and clinical trends, and identify persistent gaps in validation and integration.

Methods: We conducted a structured search of PubMed and Scopus (final search executed on October 20, 2025), covering studies published from 2017 to December 2024. Eligible peer-reviewed articles using machine learning or deep learning were analyzed with a focus on xAI components. Data from 371 studies were extracted into predefined categories covering cancer type, imaging modality, AI model, xAI method, terminology, validation, code availability, and decision support system integration.

Results: Studies focused primarily on breast (112/371, 30.2%), lung (87/371, 23.5%), and brain (56/371, 15.1%) cancers, with prostate, thyroid, and liver cancers also represented. The primary imaging modalities were computed tomography (139/371, 37.5%) and magnetic resonance imaging (104/371, 28%). Deep learning was used in 70.1% (260/371) of studies, classical machine learning in 18.1% (67/371), hybrid pipeline methods for 10% (37/371), and emerging concept-, prototype-, or causal-based approaches accounted for 1.9% (7/371) of studies. Post hoc xAI methods were dominant (305/371, 82.2%), with visualization (163/371, 53.4%), and feature relevance (111/371, 36.4%) as the most common subcategories. Hybrid post hoc or inherent approaches comprised 12.1% (45/371) and intrinsically interpretable methods 5.7% (21/371). Data sources were mostly public (149/371, 40.2%) or mixed (100/371, 26.9%); 22.9% (85/371) used private institutional datasets, and 7.8% (29/371) did not report data sources. Among validated studies, expert or user-based validation was most common (104/193, 53.9%), followed by mixed methods (74/193, 38.3%), while quantitative metrics (10/193, 5.2%) and clinical knowledge–based (8/193, 4.1%) validation remained rare. Only 17.5% (65/371) of studies provided code and 12.1% (45/371) reported decision support system integration, with few achieving actual clinical deployment.

Conclusions: This scoping review maps xAI implementation across multiple cancer imaging modalities, revealing methodological inconsistency and insufficient validation. Most research emphasizes visualization over quantitative interpretability, and few models are reproducible or clinically implemented. These findings provide an evidence base for researchers, clinicians, and regulators to prioritize standardization of xAI reporting, quantitative validation, and user-centered frameworks to advance trustworthy AI in oncology imaging.

(J Med Internet Res 2026;28:e80645) doi:10.2196/80645

KEYWORDS

cancer imaging; explainable AI; machine learning; medical imaging; trustworthiness

XSL•FO
RenderX

Introduction

Background on Cancer and Artificial Intelligence

Cancer remains one of the leading causes of mortality worldwide, with lung, colorectal, breast, prostate, liver, as well as leukemia among the most prevalent and costly forms [1,2]. Artificial intelligence (AI)–based systems can improve diagnostic accuracy, reduce analysis time, and identify patterns that might be overlooked by human experts. Such advances indicate a future transformation in the delivery of cancer care [3,4].

A study shows that using deep learning (DL) architecture and radiomics from routine computed tomography (CT) scans can predict clinically significant tumor characteristics, such as staging, without invasive procedures, supporting, in this manner, preoperative risk assessment and the development of treatment plans for patients with cancer [5]. This illustrates both AI's diagnostic potential and a core tension; the use of less intuitive deep representations, along with hand-crafted features to represent tumors, highlights the need for interpretable results and clinical validation to promote trust and the safe, ethical application of AI in cancer imaging.

The rapid development of AI-based systems is approaching a turning point. A systematic review found that more than 900 AI-enabled medical devices have received Food and Drug Administration marketing approval, with 76% (n=717) of them to be radiology devices—yet among those with available documentation, only 5% (33/717) underwent prospective testing, and 8% (56/717) included a human-in-the-loop [6]. Despite this, it has been argued that even technologically sound and highly accurate AI systems may still face limited clinical adoption if they cannot demonstrate how they reach their conclusions [7]. The recent EU Artificial Intelligence Act (AI Act) of 2024 reflects this concern, classifying AI-based systems in digital medical products intended for cancer diagnosis as "high-risk." As a result, predictive performance alone will be insufficient, and providers will be required to establish a lifecycle risk management system and adhere to requirements for transparency, human oversight, and robustness [8,9]. In other words, the operation of a high-risk system must be sufficiently transparent so that the deployer can interpret the system's output and use it appropriately, with clear user instructions and oversight mechanisms. Whether the current generation of AI-based systems—many of which operate as opaque systems—can satisfy these transparency requirements remains an open challenge [10].

Role of Explainable AI in Clinical Practice

Modern clinical practice emphasizes "human-in-the-loop" collaboration, where AI-based systems augment rather than replace medical expertise, supporting clinical decision-making [11].

However, implementing these AI systems in hospitals and clinics is still in its early stages. The primary challenge stems from the black-box nature of AI models (ie, the reasoning process is hidden), combined with cancer's complexity as a disease, and the heterogeneity of data [12,13]. This lack of transparency—the difficulty in understanding how AI models reach their conclusions—hinders trust [14] and limits adoption of these AI-based systems in clinical practice [15,16].

Explainable artificial intelligence (xAI) is being developed to address these challenges by making AI decision-making processes transparent and understandable to humans—a quality known as interpretability. Studies demonstrate that this interpretability is essential for establishing clinical confidence and accountability, as well as for integrating AI into clinical practice [17].

However, achieving this interpretability in real clinical settings presents practical challenges. Real-world use of AI in radiology shows that clinical adoption depends on more than just accuracy—it requires workflow integration, staff training, and validation in specific clinical settings to build trust among clinicians and radiologists [18]. AI in cancer imaging encounters similar challenges and requires systems that medical experts can understand and integrate into their workflows to support safe adoption.

xAI implementation in oncology is being pursued through various technical approaches, ranging from visualization tools to advanced computational methods. For example, recent research has developed AI-based systems that convert clinical data into visual formats that the AI can analyze for patterns related to patient outcomes [19]. These systems then generate visual maps highlighting which factors influenced their predictions, allowing clinicians to assess the system's reasoning. While these approaches show promise for improving clinician acceptance of AI-based systems, extensive validation in clinical settings is needed before widespread adoption.

Research Gap and Significance of the Review

While several reviews have examined xAI in health care and oncology, most of them focus on specific imaging modalities [20,21] or isolated aspects of interpretability [22,23]. A critical gap remains; we lack a comprehensive understanding of how xAI methods are actually being implemented, validated, and prepared for clinical use, specifically in cancer imaging.

A recent analysis of radiology AI-based systems reveals a concerning pattern [24]. Although xAI methods are increasingly included in DL-based computer-aided diagnosis, they remain mainly limited to simple visualization methods. More importantly, these explanations are rarely evaluated quantitatively or validated with actual clinicians [25]. This may be due to the lack of standardized frameworks for evaluating the quality of explanations generated, which presents an additional challenge for researchers attempting to evaluate and compare various methodologies in terms of their ability to produce clinically acceptable results [26].

Furthermore, a comparative evaluation across different imaging modalities and clinical applications has also shown that the accuracy and stability of explanations provided by each methodology can differ substantially depending on the application, and several methodologies underperform significantly with even slight changes to input data [27]. This suggests that selecting an appropriate xAI method for a particular clinical use case requires careful consideration of the

XSL•FO
RenderX

task and the type of medical images, to choose a method that provides the most robust results.

A study examining technical and clinical perspectives on AI validation identified a misalignment regarding what is necessary for informed decision-making, comparing the technical transparency features integrated by developers with the actual informational requirements of clinicians [28]. If there is a chasm between the information clinicians need to make informed decisions and the information provided by system developers due to documentation required for high-risk regulatory classification, it raises the question of whether current systems are developed to meet clinician needs or simply to fulfill documentation requirements for regulatory purposes.

Taken together, these concerns prompt a fundamental question—do these xAI methods genuinely support clinical decision-making, or do they merely provide technical justification after the fact [29]? Adding to this concern, limited code and data sharing practices hinder reproducibility, making it difficult to independently verify explainability claims or assess real-world clinical utility [30]. Without systematic validation and clinical evaluation, xAI risks becoming a "checkbox" feature rather than a tool that improves patient care.

This scoping review addresses these gaps by systematically examining how xAI methods are implemented, validated, and prepared for clinical use in cancer imaging. Large-scale research initiatives, such as the European Cancer Imaging Initiative [31], are developing infrastructures to support AI training and evaluation at scale, making it increasingly important to understand how explainability is integrated into these systems.

The analysis will reveal whether technical capabilities in explainability are matched by rigorous validation and real-world integration—identifying the disconnect between what xAI methods can do and whether they actually improve clinical decision-making. By mapping the current landscape, this review provides an evidence base for improving xAI research and practice. The findings will help researchers identify methodological gaps, guide clinicians in understanding which explanation methods have undergone rigorous evaluation, and inform institutions and regulators about the current readiness of xAI-enabled cancer imaging systems for clinical deployment.

Objectives of This Review

To address these gaps, this scoping review examines how explainability is implemented, validated, and reported across cancer imaging AI research. Unlike previous reviews that focus on single imaging modalities or isolated technical aspects, this review takes a comprehensive approach, covering multiple imaging modalities and both technical and clinical dimensions of xAI.

Specifically, this scoping review aims to address the following research questions:

1. How is explainability used in cancer imaging AI (post hoc vs intrinsically interpretable and model-specific vs model-agnostic), and how has this distribution evolved until 2024?
2. To what extent are explainability methods validated and reported in a standardized manner through user studies, expert evaluation, quantitative metrics, or domain-specific assessment?
3. What is the current state of reproducibility and clinical readiness of xAI in cancer imaging, including dataset availability, code sharing, and integration into decision support systems (DSS), and which barriers hinder the real-world adoption the most?

Methods

Framework and Reporting Standards

The conduction of the research and the writing of this scoping review were done according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) framework [32]. In addition, the reporting of the search strategy follows the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Search Extension) guidelines to ensure transparency in documenting information sources, full search strategies, search limits, and deduplication procedures [33]. The completed PRISMA-ScR checklist is provided in Checklist 1.

Information Sources

To comprehensively review the literature, we developed specific search queries for PubMed and Scopus. The database interfaces used were PubMed (NCBI) and Scopus (Elsevier), as recommended for transparent reporting under PRISMA-S guidelines. PubMed was selected for its comprehensive coverage of biomedical literature, making it ideal for capturing studies on AI in medicine and cancer. Scopus was chosen due to its broad coverage of scientific literature across multiple domains, including engineering and medicine. Scopus also indexes many journals not listed in Web of Science. We prioritized peer-reviewed journal articles and excluded earlier materials and nonjournal categories, such as conference abstracts or book reviews.

Search Strategy

To identify relevant literature, we created comprehensive search queries targeting studies on xAI in the context of cancer treatment. The search was conducted in PubMed and Scopus using keywords and MeSH (Medical Subject Headings) on the topics of explainability, interpretability, accountability, trust, AI, and cancer. These terms were selected following an initial analysis of key publications, which allowed us to identify the most commonly used concepts and terminology related to explainability in AI research.

Search reporting elements—including database platforms, complete search strings, search dates, applied limits, and the deduplication approach—are documented in accordance with PRISMA-S recommendations. Full search syntaxes for each database are provided in Multimedia Appendix 1.

The PubMed query combined free text and MeSH terms with Boolean operators and field markers (eg, [tiab] and [MeSH Terms]). The Scopus query was structured with TITLE-ABS-KEY operators to search titles, abstracts, and

XSL•FO
RenderX

keywords. In Scopus, certain terms (eg, *accountab* and *trust**) were restricted to the TITLE-ABS fields to minimize retrieval of nonrelevant records that often arise when these terms appear only in keywords. Both queries excluded reviews, meta-analyses, and non–English language articles. Consistent with PRISMA-S recommendations, we documented all applied limits, including restriction to English-language records and the exclusion of reviews, meta-analyses, systematic reviews, and non–peer-reviewed materials, such as conference abstracts or book chapters.

The initial search was executed in January 2024, covering records from 2017 to 2023. No restrictions were placed on data modalities, ensuring broad inclusion of studies from different subfields of AI research related to cancer. Following peer review, the search was updated on October 20, 2025, to include studies published through December 2024. Exact search dates for each database are reported in Multimedia Appendix 1. The full query syntaxes for both databases can be found in Multimedia Appendix 1.

Duplicate records across databases were removed using a custom Python script that identified duplicates based on matching DOI fields, followed by manual verification to ensure accuracy.

Eligibility Criteria

Medical imaging modalities such as CT, positron emission tomography (PET), x-ray, mammography, magnetic resonance imaging (MRI), and ultrasound were included in the review because of their central role in cancer diagnosis, treatment planning, and frequent use in AI research. These radiologic imaging techniques provide noninvasive, clinically relevant information, and are widely used for monitoring cancer progression. In contrast, imaging modalities primarily used in other biomedical fields, such as histopathology, immunofluorescence, and dermoscopy, were excluded to maintain focus on radiologic data.

We included articles published from 2017 to December 2024. Inclusion criteria were:

1. Introduces or evaluates an AI system applied to cancer care with a focus on explainability, interpretability, or trustworthiness;
2. Published in peer-reviewed journals or as full, peer-reviewed conference papers presented at conferences (conference abstracts were excluded);
3. Written in English;
4. Not a review, opinion, editorial, commentary, or meta-analysis article.

Duplicates were removed. Studies that only mentioned explainability or interpretability superficially without methodological detail were also excluded. The search strategy and its reporting adhere to PRISMA-S guidance, ensuring transparent documentation of all search components, including database interfaces, search terms, restrictions, and handling of duplicates.

Data Extraction and Information Categories

The data extracted are organized into the following categories:

- Study identification: title, authors, year, and source of publication.
- Study characteristics: cancer type, primary objective, dataset description, and sample size.
- Data and methods: data modalities, preprocessing steps, features used, feature selection, and machine learning (ML) or DL algorithms.
- Clinical aspects: validation approach and clinical performance metrics.
- Explainability aspects: xAI methods used, their purpose, terminology, and user involvement.
- Results and limitations: key findings and reported limitations.

The detailed data extraction schema, including all the field names and descriptions, is available in Multimedia Appendix 2.

D Fotopoulos conducted the initial search and deduplication. D Fotopoulos and D Filos screened titles and abstracts. Studies were randomly distributed among the reviewers—D Fotopoulos, D Filos, and IL—and they proceeded to perform the full-text reviews. To ensure consistency across reviewers, PAMS independently reviewed a random sample of 10 studies from each batch of full-text reviews. If substantial discrepancies were observed, PAMS would increase the audit sample; conflicting evaluations were resolved by discussion among all reviewers, ensuring consensus before final inclusion decisions.

Synthesis of Results

Extracted data were managed in Microsoft Excel and analyzed using descriptive mapping with frequency counts, consistent with Joanna Briggs Institute scoping review methodology. Study characteristics (cancer type, imaging modality, dataset availability, code availability, and integration into a DSS) were coded at the time of extraction using predetermined categories established before the extraction process began. The explainability methods used in each study were categorized according to the framework described in the next section, based on a taxonomy developed by Arrieta et al [34], including the xAI strategy (inherent, post hoc, or hybrid), post hoc method family (text, visual, local, example-based, simplification, or feature relevance), model dependence (model-specific or model-agnostic), and explanation scope (local or global). Categories of validation approaches were not established a priori; however, through an inductive approach during data extraction based on the strategies reported by the included studies, 4 categories (expert or user-based, performance-based, domain or clinical knowledge-based, and mixed methods) emerged. Results will be presented as frequencies and percentages, with temporal trends evaluated when applicable. No meta-analysis or interpretative qualitative synthesis was conducted, consistent with the descriptive mapping goals of this scoping review.

Background and Theoretical Framework

Overview

This section defines the terminology and classification system for xAI methods reviewed in cancer imaging studies. The objectives are (1) to provide a standard way of referring to

XSL•FO
RenderX

"interpretability," "explainability," and "transparency" in the context of this review; (2) to group xAI techniques commonly reported in studies into reproducible families of methods; and (3) to identify how each study documented its intended application of explainability and whether any evaluations of that explainability were conducted.

The concepts and taxonomic distinctions referenced in this section are based on previous research related to the field (eg, Arrieta et al [34], Holzinger et al [35], and Roscher et al [36]). In other words, our goal in mapping each included study to this established landscape is to use the existing framework for reference.

Scope of This Framework

This review analyzed 371 studies (Multimedia Appendix 3) that used various terms to describe "interpretability, "explainability," and "transparency" for presenting their research findings with a wide range of xAI techniques. For each study included, we mapped its explainability approach onto 4 dimensions. First, we recorded the xAI strategy, distinguishing whether the approach was inherently interpretable, post hoc, or a hybrid of the two. Second, for studies that used post hoc xAI, we assigned one or more post hoc method categories, drawing on the families defined by Arrieta et al [34], which include text explanations, visual explanations, local explanations, explanations by example, explanations by simplification, and feature relevance explanations. Third, we documented the model dependence of each technique, classifying methods as either model-specific or model-agnostic. Finally, we captured the scope of the explanation, noting whether it was intended to be local (instance-level), global (model- or dataset-level), or a combination of both.

Key Concepts

Explainability and interpretability are 2 terms that can be found used interchangeably in some cases, although some literature suggests they should be distinguished, as they convey different meanings [37,38]. Both support transparency, which refers to clarity regarding the model's structure, data, and assumptions, a cornerstone of trustworthy AI [39]. We use the following working definitions to prevent ambiguity throughout this review.

Interpretability refers to an intrinsic (passive) property of a model: the extent to which a human can directly understand how inputs are mapped to outputs without requiring an additional explanatory mechanism. Roscher et al [36] describe interpretability as the inherent comprehensibility of a model. Linear regressions or shallow decision trees (DTs) are interpretable by design because their parameters map directly to explicit features, and Arrieta et al [34] relate this notion to transparency.

We use explainability to refer to an active property: the ability of a system to generate reasons, evidence, or visual and linguistic justifications for its outputs in a form intended to be understandable to a target audience. Explainability may be inherent to the architecture (eg, a model that surfaces explicit prototypes or attention maps as part of its forward pass), or it may be added after training via post hoc analysis of an otherwise opaque model.

Transparency involves openness in model structures, data, and design assumptions. Transparency—informally the opposite of opacity—varies by context and user and can be analyzed across 5 dimensions—human involvement, data, the model, inferencing, and algorithmic presence [40]. Also, according to Arrieta et al [34], which is the one that guides this review, can be divided into 3 levels:

- Simulatability: a human can mentally reproduce the model's reasoning,
- Decomposability: each component (inputs, parameters, and computations) is interpretable, and
- Algorithmic transparency: the clarity of the learning procedure itself. Linear and logistic regression models typically meet all 3.

These distinctions matter because the terms "interpretable," "explainable," and "transparent" are not used consistently. In several instances across the literature, authors describe their approach as "transparent" primarily because the underlying model class is simple (eg, a shallow DT or sparse linear model). Other studies provide only post hoc saliency maps for a deep neural network and still refer to the solution as "transparent."

In our review, we do not assign a transparency rating to each research study to measure the levels of simulatability, decomposability, or algorithmic transparency. Instead, we record (1) which model architecture was used and (2) whether interpretability is claimed to be inherent or is provided through a separate post hoc explanation method. This allows us to report, in the Results section, how often explainability is treated as a built-in design property versus an add-on.

Taxonomy of xAI Methods

Overview

To compare explainability strategies across studies, we group reported methods along these axes widely discussed in xAI research and directly aligned with our extraction fields: xAI *method*, *type of xAI*, and *aim of interpretability or explainability*.

A central distinction in this taxonomy is whether a study uses an interpretable-by-design model, a post hoc explainability technique, or a combination of both. Interpretable-by-design approaches construct the model so that its internal decision process is directly understandable. Examples include linear models built over predefined features, shallow DTs, rule-based systems, and prototype-based architectures in which predictions are explicitly linked to reference exemplars; in these cases, the explanation is embedded within the model itself. Post hoc explainability methods, by contrast, are applied after training to elucidate the behavior of otherwise opaque models, typically through outputs such as saliency maps, pixel or feature importance scores, or simplified local surrogate models. Hybrid approaches appear in studies that use an inherently interpretable model while also incorporating an additional post hoc xAI method.

During extraction, we map each study to one or more of these categories, based on how they describe their approach in *ML or DL method used*, *type of xAI*, and *xAI method* fields.

XSL•FO
RenderX

Model-Specific vs Model-Agnostic Methods

We then classified explainability techniques based on whether they depend on the internal structure of a particular model class. Model-specific methods rely on architecture-specific information, such as gradients, activation maps, attention scores, or layer relevance. Examples include gradient-weighted class activation mapping (Grad-CAM) and its variants for convolutional neural networks, attention-weight visualization in transformer architectures, and layer-wise relevance propagation. Because they incorporate internal model characteristics, these approaches generally cannot be applied unchanged to other model families. In contrast, model-agnostic methods treat the model as a black box and infer explanations by probing input-output behavior. Representative techniques include local interpretable model-agnostic explanations (LIME) [41], Shapley Additive Explanations (SHAP) [42], perturbation-based feature importance, occlusion sensitivity analyses, and locally fitted surrogate models for individual predictions. Since they do not depend on internal model layers, these methods can, in principle, be applied across a wide range of classifiers or regressors, including ensemble architectures.

Local vs Global Explanations

We have separated our methods into local explanations (explanations for a single prediction, such as heatmaps of a suspected malignancy region in 1 patient) and global explanations (overview of model performance across the entire dataset, such as top-ranked lists of the most influential features). For each method, we have documented how it describes its own scope as local, global, or both. We note that "local explanation" appears in xAI categorization by Arrieta et al [34], where "local" refers to a family of methods focused on sample-by-sample justification, typically using surrogate models of local data.

Families of Explainability Methods

Finally, for studies that use post hoc explainability, we assign each reported method to one or more method families, which follow the taxonomy presented in the study by Arrieta et al [34].

1. *Text explanations:* Methods that generate a verbal, symbolic, or natural language rationale for prediction. In imaging, this includes approaches that describe suspicious morphology (eg, spiculation and enhancement patterns) or provide a rule-like statement in words.
2. *Visual explanations*: Methods that produce visual proofs, intended to show what the model relied on. In the context of medical imaging, this could be a heatmap, an attention map, or activation overlay on the original image series to indicate regions of interest.
3. *Local explanations:* Methods that aim to explain an individual prediction in its local neighborhood, typically by constructing a simplified, human-readable surrogate just for that case (eg, a local linear model or a shallow local DT), or by providing a per-instance strategy.
4. *Explanations by example*: Methods that justify a prediction by retrieving or referencing similar prior cases, exemplars, or learned prototypes ("this lesion is malignant because it resembles these malignant prototypes"). This follows medical experts' (eg, radiologists) reasoning process.
5. *Explanations by simplification:* Methods that approximate a complex model with a simpler surrogate; for example, extracting a rule set, fitting a shallow DT, or learning a sparse linear model that mimics the original model locally or globally. The goal is to communicate decision logic in a more interpretable form.
6. *Feature relevance explanations:* Methods that assign importance scores to features, pixels, or regions to indicate which inputs most influenced the model's prediction.

Individual studies may report more than one category (eg, both saliency maps and a local surrogate model). We record all categories. In the Results section, we report the prevalence of each category across cancer types and imaging modalities.

Results

Overview of Included Studies

The literature search was conducted in the Scopus and PubMed databases, covering publications from 2017 to 2024. A total of 2715 records were identified (1843 from Scopus and 872 from PubMed). After removing 372 duplicates and 19 reviews, 2324 records remained for title and abstract screening. Of these, 1743 records were excluded for not meeting the initial criteria.

Subsequently, 581 reports were sought for retrieval, of which 16 could not be retrieved. A total of 565 reports were assessed for eligibility. Upon full-text review, 194 studies were excluded as they did not meet the specific inclusion criteria, primarily due to a lack of focus on cancer diagnosis, the absence of demonstrable xAI methods, or the use of nonradiological data. Finally, 371 studies were included in the review (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram illustrating the selection process for the studies investigated. The search identified peer-reviewed studies published between 2017 and 2024 in Scopus and PubMed. After duplicate removal and exclusion of reviews, 2324 records were screened and 371 studies met the inclusion criteria. xAI: explainable artificial intelligence.

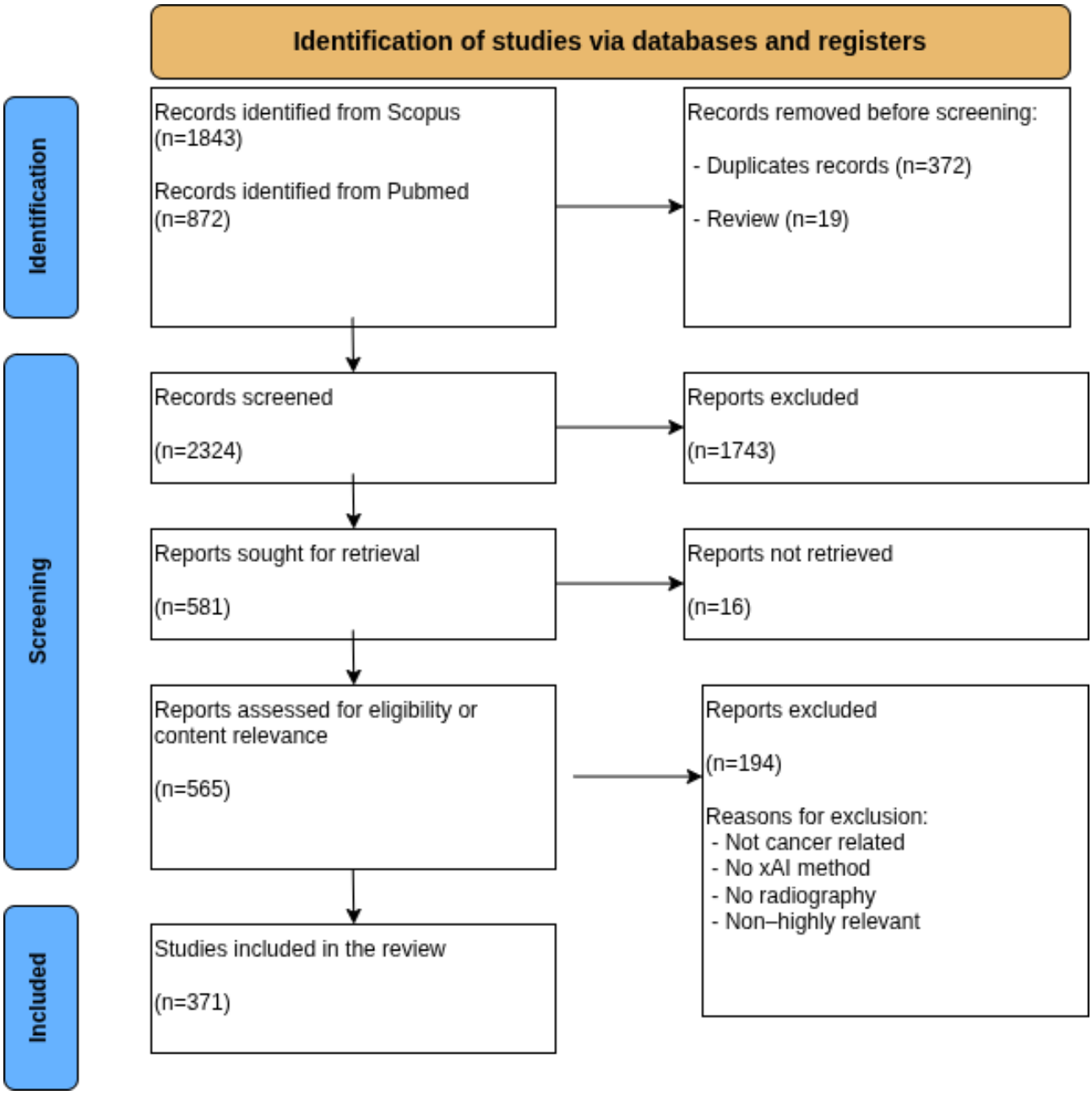

Summary of Study Characteristics

To provide a comprehensive overview of the included literature, Tables 1 and 2 summarize the key characteristics and methodological approaches of all 371 studies. These tables present the collected data on temporal distribution, cancer types, imaging modalities, dataset accessibility, xAI implementation approaches, and validation practices, offering a systematic mapping of the current landscape of xAI in cancer imaging research.

XSL•FO
RenderX

Table . Study characteristics of included studies (N=371). Temporal distribution, cancer types, imaging modalities, and dataset accessibility among the 371 included studies. Studies spanned 2017 - 2024, with most published in 2024 (n=200, 53.9%).

Characteristic and category	Studies (N=371), n (%)
Publication period	
2017 - 2023	171 (46.1)
2024	200 (53.9)
Cancer type	
Breast	112 (30.2)
Lung	87 (23.5)
Brain or CNS[a]	56 (15.1)
Prostate	22 (5.9)
Thyroid	19 (5.1)
Head and Neck	16 (4.3)
Liver (Hepatocellular)	14 (3.8)
Pancreatic	7 (1.9)
Colorectal	5 (1.3)
Renal or kidney	5 (1.3)
Cervical	4 (1.1)
Other[b]	24 (6.5)
Imaging modality	
CT[c]	139 (37.5)
MRI[d]	104 (28)
Ultrasound	63 (17)
Mammography	46 (12.4)
X-ray	8 (2.2)
PET[e]-CT	8 (2.2)
Other or multiple[f]	3 (0.8)
Dataset accessibility	
Open	149 (40.2)
Mixed or combined	100 (26.9)
Not open	85 (22.9)
Uncertain	29 (7.8)
Synthetic or custom	8 (2.2)

[a]CNS: central nervous system.

[b]Includes rare malignancies and multidisease studies (each <1%): bone, stomach, ovary, esophagus, hematologic, skin, endometrial, bladder.

[c]CT: computed tomography.

[d]MRI: magnetic resonance imaging.

[e]PET: positron emission tomography.

[f]Some studies used multiple imaging modalities.

XSL•FO
RenderX

Table . xAI[a] methods, validation, and reproducibility indicators (N=371). Distribution of xAI approaches, modeling methods, validation practices, code availability, and DSS[b] integration.

Characteristic and category	N=371, n (%[c])
xAI approach type	
Post hoc	305 (82.2)
Hybrid	45 (12.1)
Intrinsically interpretable	21 (5.7)
Post hoc method subcategories[d]	
Visualization-based	163 (53.4)
Feature relevance	111 (36.4)
Text explanations	7 (2.3)
Explanation by example	6 (2)
Simplification or surrogate	5 (1.6)
Combination (multimodal)	13 (4.3)
Primary modeling approach	
Deep learning	260 (70.1)
Classical ML[e]	67 (18.1)
Hybrid	37 (10)
Emerging or other	7 (1.9)
xAI validation status	
Validation reported[f]	193 (52)[g]
Expert or user-based	104 (28)[g]
Mixed methods	74 (38.3)[g]
Quantitative metrics	10 (5.2)[g]
Domain or clinical knowledge	8 (4.1)[g]
No validation reported	178 (48)
Code availability	
Code shared	65 (17.5)
Code not shared	280 (75.5)
Not specified	26 (7)
DSS integration	
DSS integration reported	45 (12)
No DSS integration	321 (86.5)
Not specified	5 (1.3)

[a]xAI: explainable artificial intelligence.

[b]DSS: decision support system.

[c]Percentages calculated as proportion of 305 post-hoc studies.

[d]Post hoc subcategories are mutually exclusive; “Combination” indicates use of ≥2 post hoc methods

[e]ML: machine learning.

[f]For validation subcategories, percentages in parentheses represent the proportion of the 193 validated studies; percentages can exceed 100% as categories are not mutually exclusive.

[g]Of validated.

Table 1 presents the distribution of studies by publication period, cancer type, imaging modality, and dataset accessibility. Most studies were published in 2024 (n=200, 53.9%). Breast cancer was the most frequently studied malignancy (n=112, 30.2%),

XSL•FO
RenderX

followed by lung cancer (n=87, 23.5%), and brain or central nervous system (CNS) tumors (n=56, 15.1%). CT (n=139, 37.5%) and MRI (n=104, 28%) were the most commonly used primary imaging modalities. Regarding data accessibility, 149 studies (40.2%) used open datasets, while 85 (22.9%) relied on not-open institutional or private sources.

Table 2 summarizes xAI implementation approaches and validation practices. Post hoc explainability methods dominated (n=305, 82.2%), followed by hybrid approaches (n=45, 12.1%), and intrinsically interpretable models (n=21, 5.7%). Approximately half of the studies reported some form of xAI validation (n=193, 52%), most commonly expert or user-based evaluation (n=104, 28% of all studies); validation categories were not mutually exclusive, and totals may exceed 193. Code sharing remained limited (n=65, 17.5%), and integration into clinical DSS was reported in 45 studies (12.1%).

Study Characteristics

Cancer Types

xAI methods have been applied to a wide range of cancer types, demonstrating their versatility in oncologic imaging. Breast cancer was the most extensively studied (n=112/371, 30.2%), followed by lung cancer (n=87/371, 23.5%) and brain or CNS tumors (n=56/371, 15.1%). Together, these three categories accounted for nearly 70% (255/371) of all reviewed studies, consistent with their global incidence and mortality burden and with previous bibliometric analyses of AI in oncology [1,43].

Moderately represented cancer types included prostate (22/371, 5.9%), thyroid (19/371, 5.1%), head and neck (16/371, 4.3%), and liver (hepatocellular and related; 14/371, 3.8%). Smaller groups addressed pancreatic (7/371, 1.9%), colorectal (5/371, 1.3%), renal or kidney (5/371, 1.3%), and cervical (4/371, 1.1%) malignancies. Other sites, such as stomach, bone, ovary, esophagus, hematologic, skin, endometrial, and bladder cancers, were examined in 5 or fewer studies (<1% each).

Within the *"Other"* category (n=24), 6 studies (6/371, 1.6%) addressed multidisease or nonorgan-specific contexts, including salivary gland tumors, metastasis or lymph node prediction, gastrointestinal stromal tumors, and mixed cancer cohorts that combined, for example, breast, brain, and cardiovascular imaging. Collectively, these studies highlight ongoing efforts to develop generalizable and cross-site explainable AI frameworks that extend beyond single-organ applications.

Imaging Modalities

Across all included studies, CT (n=139, 37.5%) and MRI (n=104, 28%) were the most common primary modalities, followed by ultrasound (n=63, 17%) and mammography (n=46, 12.4%). Multimodal designs were present but were assigned to a primary modality (or "Other or multiple," n=3, 0.8%) for counting in Table 1. X-ray and PET-CT were less common (n=8 each, 2.2%) but remain relevant for specific clinical contexts, particularly thoracic and metastatic imaging (Figure 2).

Figure 2. Distribution of imaging modalities used in explainable artificial intelligence studies for cancer imaging (2017 - 2024; N=371). Stacked bars indicate the number of studies per publication year by modality. CT: computed tomography; MRI: magnetic resonance imaging; PET: positron emission tomography.

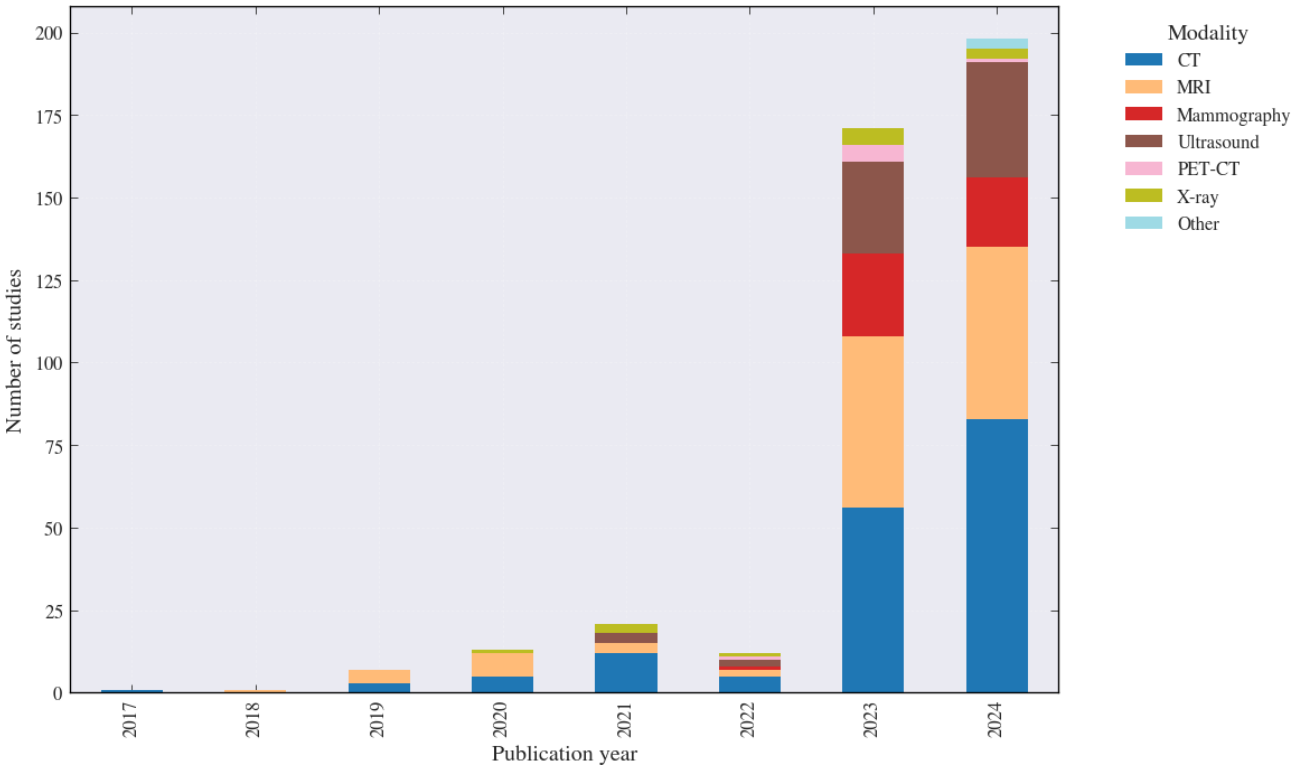

Over time, CT remained the most common modality, particularly in 2023 (56/171, 32.7%) and 2024 (83/200, 41.5%; Figure 2). MRI remained the second most common modality across the period. Ultrasound and mammography contributed a substantial minority of studies, particularly in later years, reflecting continued interest in breast-focused imaging and

XSL•FO
RenderX

multimodality-based workflows. PET-CT and x-ray appeared less frequently but remained present in clinically relevant contexts.

Dataset Availability and Composition

The datasets used in the 371 reviewed studies came from a wide range of sources, including clinical repositories, public challenges, and institutional imaging archives (Figure S1 in Multimedia Appendix 4; Table 3). This diversity demonstrates the broad scope of AI applications in oncology and reflects varying levels of accessibility and reproducibility across studies. Dataset availability was categorized into 5 groups: open (publicly accessible repositories), not open (institutional or private sources), mixed or combined (both open and private sources), synthetic or custom (purpose-built datasets), and uncertain (insufficient information).

Table . Accessibility of datasets used in studies applying artificial intelligence to cancer imaging (2017-2024; N=371). The corpus includes major cancers (breast, lung, brain, liver, and prostate) from multiple continents.

Dataset accessibility category	Studies, n (%)
Open	149 (40.2)
Mixed or combined	100 (26.9)
Not open	85 (22.9)
Uncertain	29 (7.8)
Synthetic or custom	8 (2.2)
Total	371 (100)

Most studies relied on open or mixed data sources. Open datasets accounted for 149 of 371 (40.2%) studies, making up the largest single group. These included established repositories such as LIDC-LDRI, the Cancer Imaging Archive (TCIA), the Cancer Genome Atlas (TCGA), CBISDDSM, and public competition datasets like PROSTATEx and AAPM 2017, all directly cited in the reviewed works. Openly available datasets are essential for ensuring reproducibility, benchmarking, and collaborative validation across institutions.

In contrast, not-open datasets comprised 85 of 371 (22.9%) studies, typically based on patient data from hospitals and medical centers (eg, Harbin Medical University, Ruijin Hospital, and Asan Medical Center). These datasets are inaccessible to external researchers, limiting reproducibility despite their clinical authenticity.

A substantial proportion, 100 of 371 (26.9%) studies, used a mixed or combined approach, integrating public datasets with proprietary clinical data to improve external validity and increase sample diversity. For example, some studies combined TCGA or TCIA datasets with internal cohorts for independent testing.

A smaller subset, 8 of 371 (2.2%) studies, relied on synthetic or custom datasets created specifically for research purposes, such as simulated imaging phantoms or institution-specific multimodal collections (eg, Credence Cartridge Radiomics phantom). These datasets often integrate imaging, clinical, and demographic data tailored to specific research questions.

Finally, 29 of 371 (7.8%) studies provided unclear or incomplete information about their data sources, highlighting ongoing challenges in dataset transparency and reporting standards.

Modeling Approaches

Methodological Note

Because some studies used multiple models, we assigned a single main modeling approach to each study. Categories are mutually exclusive: ML, DL, hybrid, or “atypical.” We labeled a study as hybrid when prediction relied on both DL and classical ML (eg, a Convolutional Neural Network [CNN] feature extractor with an extreme gradient boosting [XGBoost] for inference) or when the architecture explicitly integrated multiple families; otherwise, we labeled by the dominant predictive component (eg, end-to-end CNN=DL, radiomics plus gradient boosting=classical ML). Counts, therefore, sum to n=371 and refer to studies, not model instances.

Overview

In the majority of studied articles (n=260/371, 70.1%), DL architectures are used (Figure S2 in Multimedia Appendix 4 and Table 4). In most cases, they serve as the main analytical components of the analyzed studies; in other cases, they are part of hybrid pipelines where classical ML algorithms are used for secondary classification or feature aggregation. Within DL, CNN-based families (eg, Residual Network, Visual Geometry Group, and Densely Connected Convolutional Network) remain dominant for image-level tasks, U-Net variants are the default for segmentation, and transformer or attention architectures rose notably in 2023, aligning with self-attention maps as native interpretability; CNNs still constitute the largest DL subgroup (149/260, 57.3%). In hybrid CNN-to-ML setups, the DL block extracts spatial or contextual features while the ML block (eg, support vector machine [SVM], random forest [RF], and XGBoost) provides interpretable decision rules or feature rankings.

XSL•FO
RenderX

Table . Distribution of primary modeling approaches among examined studies (2017-2024; N=371).

Primary modeling approach	Studies, n (%)
Deep learning	260 (70.1)
Classical machine learning	67 (18.1)
Hybrid approaches	37 (10)
Emerging or other	7 (1.9)
Total	371 (100)

Classical ML algorithms (67/371, 18.1%) are still used frequently, mostly in analyses focused on radiomics or as comparative baselines. Models commonly selected for interpretability include RFs and SVMs (feature importance and margin or kernel effects), while since 2023 gradient-boosting families (Categorical Boosting, Light Gradient Boosting Machine, and XGBoost) are frequently used in ensemble or multimodal fusion pipelines because they connect directly to structured radiomics or clinical variables and expose feature attributions.

The number of hybrid approaches (37/371, 10%), including CNN-ML ensembles, multitask frameworks, and Concept Bottleneck Model architectures has increased between 2023 and 2024. Hybrid approaches integrate complementary model families or data sources to balance performance with interpretability. Typical patterns include CNN and ML pipelines, where a deep encoder produces compact features that a transparent learner (eg, SVM, RF, and gradient boosting) turns into decision rules or feature rankings. Overall, these designs emphasize embedding interpretability within the training pipeline rather than relying solely on post hoc explanations.

A small portion of studies (7/371, 1.9%) use modeling approaches that are less common in current practice. Notably, concept and prototype-based designs (eg, Concept Bottleneck Models, ProtoPNet, and Breast Imaging Reporting and Data System [BI-RADS]-NET) and architectures with explicit causal mechanisms (eg, Causeg-Net and BDStable-Net) aim to anchor model learning to clinically meaningful factors and human-understandable concepts, providing built-in transparency.

Explainability Methods in Cancer Imaging

Terminology Landscape in the Reviewed Studies

A qualitative analysis of the corpus identified 6 recurring categories in terminology (Figure S3 in Multimedia Appendix 4). Most usage is concentrated in the categories of interpretable or interpretability and explainable or explainability. Trustworthy and transparent also appear, but as secondary themes. Other labels, such as reliable or visualization, form a “long tail.” Because the key terms for retrieval were both explainable and interpretable, these observations are descriptive rather than unbiased frequency counts. The takeaway is that xAI vocabulary is diverse and often used interchangeably in literature. Therefore, we report this mapping to support replication and facilitate future searches.

Distribution of Explainability Types

Analysis across 371 studies published from 2017 to 2024 showed that post hoc xAI methods were dominant, used in 82.2% (n=305) of all studies (Figure S4 in Multimedia Appendix 4 and Table 5). Most of these studies used post hoc as the primary means to add transparency to their predictive models. Studies using inherently transparent models, such as those with built-in structures (eg, DTs and linear coefficients) that enable transparency by design, accounted only for 5.7% (n=21) of studies. Hybrid methods accounted for 12.1% (n=45) of studies, combining inherently interpretable models with post hoc techniques to provide more robust or complementary explanations.

Table . Distribution of explainable artificial intelligence methods across included studies.

Category	Studies, n (%)
Post hoc	305 (82.2)
Visualization-based	163 (53.4)
Feature relevance	111 (36.4)
Combination	13 (4.3)
Text explanation	7 (2.3)
Explanation by example	6 (2)
Explanation by simplification	5 (1.6)
Hybrid	45 (12.1)
Ante-hoc	21 (5.7)
Total	371 (100)

XSL•FO
RenderX

Post hoc approaches dominated (n=305, 82.2%), followed by hybrid (n=45, 12.1%) and ante hoc (intrinsic) methods (n=21, 5.7%). Within the post hoc group, visualization-based explanations, such as saliency maps and activation heatmaps, were the most prevalent (n=163, 53.4%), followed by feature relevance approaches including SHAP and layer-wise relevance propagation (n=111, 36.4%). Combination (multimodal) post hoc strategies accounted for 4.3% (n=13). Other categories, such as text explanations (n=7, 2.3%), explanations by example (n=6, 2%), and explanations by simplification (n=5, 1.6%), were comparatively infrequent, underscoring the dominance of visual and feature-based interpretability techniques in current xAI research.

The temporal evolution of explainability methods from 2017 to 2024 reveals a clear and notable trend. Post hoc methods have consistently remained the dominant approach; however, the composition of the methods used has changed. Whereas visualization-based methods, such as Grad-CAM, were previously most common, they are now increasingly rivaled by more recently available feature-perturbation methods, such as SHAP, which have even become dominant in some years.

The extent to which the study authors clearly defined the goals for explainability varied throughout the study period. Early studies were more likely to define explainability goals broadly (eg, “to improve interpretability” and “to provide explanations”) without specifying the particular clinical applications or potential user groups. In contrast, more recent studies increasingly define specific objectives (eg, to foster clinician confidence, to make model output consistent with diagnostic standards, such as BI-RADS, and to enable visual validation of lesion location).

Inherently Interpretable or Transparent Models

In only 21 of 371 studies (approximately 5.7%), researchers used models designed to be interpretable from the beginning, rather than explaining them retrospectively, indicating that fully intrinsic interpretability remains rare.

The largest group of studies (9/21) achieved interpretability by imposing structural or architectural constraints within deep models. Examples include the hierarchical semantic CNN or hierarchical semantic network for lung nodule malignancy [44], Thyroid Imaging Reporting and Data System–based multitask networks for thyroid nodules [45], prototype-based CNNs for mammography [46], and lesion-level risk aggregation models for metastatic lung cancer prognosis [47]. In these cases, the model provides clinically relevant information, such as nodule shape or margin [48], assigns specific per-lesion risk scores [47], or presents prototypical patterns as part of its standard inference process [46].

In a second subset (3/21), studies used generalized additive or sparse linear models, such as Explainable Boosting Machines and least absolute shrinkage and selection operator–regularized linear or logistic regression, for prognosis of treatment response (local failure in head and neck cancer [49]; 3-year overall survival in early-stage non–small cell lung cancer [50]) and prediction of tumor markers in kidney cancer [51]. Feature-wise contribution functions and sparse coefficients on clinically meaningful features serve as the built-in explanations of these models.

In total, 5 studies (5/21) used transparent, feature-engineered pipelines, for example, for prostate cancer segmentation [52], or liver cancer treatment monitoring [53], and tumor classification in breast MRI [54], focusing on explicit feature rankings. Furthermore, 2 studies defined their interpretability in terms of biologically grounded phenotypes, such as radiomic factors linked to CD8+ T lymphocyte infiltration and tumor microenvironment phenotypes, using these as model inputs instead of opaque latent features [55,56].

Moreover, 2 other studies (2/21) were based on probabilistic and fuzzy models (fuzzy classification for radiotherapy toxicity and Bayesian multitask learning regression), providing global interpretability through human-readable fuzzy rules or task-specific linear mappings between predictors and clinical targets. Rule-based and symbolic learners were also present in this xAI category (2/21; rule-fit for chemoradiation PET response in non–small cell lung cancer [57] and grammatical evolution classifiers for breast cancer [58]). They generated explicit human-readable rules and formulas.

Studies in this category generally adopted transparent models to meet clinical interpretability requirements, such as providing explicit feature importance rankings from radiomics features or feature-outcome correlation plots, rather than relying on visual explanations.

Post Hoc Explainability Approaches

Overview

Within the post hoc subset (n=305), visualization-based explanations were the most common (163/305, 53.4%), followed by feature relevance methods, such as SHAP and LIME (111/305, 36.4%). Text-based explanations (7/305, 2.3%), explanation-by-example approaches (6/305, 2%), and simplification or surrogate-based methods (5/305, 1.6%) were comparatively rare. A subset of studies used combination (multimodal) post hoc strategies (13/305, 4.3%), integrating 2 or more post hoc explanation types to provide complementary perspectives on model behavior.

Visualization-Based Methods

The most common class includes visualization techniques (163/305, 53.4%). Representative methods include Grad-CAM, saliency maps, and attention heat maps. These techniques are well-suited to CNN-based architectures, such as Res-Net, Dense-Net, U-Net, and their attention-enhanced variants, and are commonly used in radiology due to their intuitive and spatially localized results [59,60].

The main motivation for these methods is to highlight image regions that influence the networks’ decisions and align them with the radiologist’s reasoning, thus promoting clinical confidence. They are especially common in CNN-based classification and segmentation studies, where visual inspection provides intuitive correspondence with tumor or lesion regions. For example, studies have used Grad-CAM to highlight tumor subregions in breast [61-63], brain [64-66], prostate [67], and thyroid imaging [68,69].

XSL•FO
RenderX

Grad-CAM or CAM approaches represent the majority (73/163, 44.8%), followed by saliency or heatmap techniques, and other visualization-based strategies. An important limitation reported in the investigated studies [70] is the subjective nature of heatmap interpretation, which underscores the need for standardized metrics. Nevertheless, the emergence of variants, such as Full-Grad or attention-based overlays, reflects the ongoing refinement of visualization-based xAI approaches [71].

Feature-Relevance Methods

The second large category includes feature attribution techniques (111/305, 36.4%), in particular SHAP. For reporting, we merged in this category the studies that fall under the category of *local explanations* (eg, LIME; n=3). Although Arrieta et al [34] define them as distinct subcategories, in practice, there is overlap in their techniques and objectives. Therefore, for clarity and alignment with the methods used in the reported results, we combined them into a single *feature relevance* category.

These approaches assign importance scores to individual input features, such as radiomic markers [72,73], clinical attributes, or embedded feature vectors [74,75], and are particularly useful when DL models are applied to multimodal or structured data [76-78], including combinations of imaging and clinical information [79]. Most studies use CNN-based and feature-enhanced architectures (eg, Res-Net, VGG16, and U-Net), often in combination with attention modules or radiomic pipelines for structured inputs.

Simplification or Surrogate Models

A small group of studies (5/305, 1.6%) use simplification techniques that aim to improve interpretability by creating surrogate or human-understandable models that approximate the behavior of complex DL systems. These methods (eg, rule extraction, fuzzy inference systems, and contour-based symbolic mappings) translate model logic into structured formats that clinicians can interpret more easily. The underlying architectures are usually CNN-based classifiers. Although methodologically sound, their adoption remains limited, mostly conceptual or in proof-of-concept stages due to concerns over surrogate fidelity and reproducibility [80].

Examples include a study by Contreras et al [81], where the authors derive rule sets to mimic a black box model and explain human papillomavirus diagnosis. In [82], the model integrated a Takagi-Sugeno-Kang fuzzy system to convert feature activations into interpretable fuzzy rules. In another study [83], mathematical curve modeling was used to approximate region of interest boundaries in a clinically meaningful way. Studies reported that these simplification methods can improve interpretability while achieving comparable or higher predictive performance than the black box alternatives [84,85].

Text Explanations

Text explanation category methods were used in only a few studies (n=7), typically through post hoc textual outputs designed to complement predictions with human-interpretable reasoning. In several cases, explanations were based on clinical descriptors. For instance, semantic outputs aligned with BI-RADS criteria were used in breast cancer classification models to enhance interpretability for clinicians [86,87]. A graphical user interface–driven explanation system providing diagnostic and treatment suggestions was implemented for airway disease classification using CNN models [88].

With the advent of large language models, studies that integrate them in xAI pipelines have emerged. In one, GPT-4 (OpenAI) was evaluated for thyroid ultrasound analysis, which decomposed the reasoning behind each diagnostic output [89]. Furthermore, 2 studies linked imaging features to biological information—one associated radiomics with long noncoding RNA expression [90], while another correlated CT-derived features with biological pathways using gene set variation analysis [91].

Explanation by Example

In total, 6 studies used example-based explanations through prototype reasoning or counterfactual generation. In the first category, we have studies that used prototype-based explainers, such as the study by Yang et al [92], which justify the model's prediction by showing image regions that match learned "prototypes" of lesions (ie, the nearest positive examples aligned with clinical cues). The other category, using counterfactuals [93], includes studies that answer the question "what minimal change would change the model's outcome?," thus generating contrastive examples in order to reduce spurious cues and focus attention on the lesion. Another example in the counterfactuals category [94] altered slightly imaging, clinical, and molecular variables to identify drivers for treatment response. A common thing in this category is that all methods were model-specific and local, that is, the predictions were generated per instance.

Multimodal Explanation Strategies

A total of 13 studies used a combination of post hoc xAI methods from different subcategories, most commonly feature relevance and visualization. Feature relevance methods, such as SHAP and LIME, were used to assign importance scores to input features, while visualization techniques, like Grad-CAM and attention maps, highlighted spatial or contextual cues within medical images. Typically, visualization-based methods, such as Grad-CAM, are model-specific; however, SHAP and LIME are model-agnostic, allowing for flexible integration of these techniques across various architectures.

The use of both types of techniques was common, and they were often applied together to provide global and local insights. For example, SHAP and Grad-CAM were used in multimodal or ensemble frameworks to interpret diagnostic and prognostic models [95-98]. On the other hand, attention maps and risk heatmaps were used alongside feature attribution techniques in transformer and graph-based networks [99,100]. Overall, these studies indicate a growing trend towards hybrid xAI approaches that use multiple perspectives to promote transparency.

Explainability Practices by Imaging Modality

CT and MRI were the most common modalities, with 139 (37.5%) and 104 (28.0%) studies, respectively, followed by ultrasound (n=63, 17%) and mammography (n=46, 12.4%). Validation practices varied by modality. Studies using ultrasound and mammography reported higher rates of xAI component validation (72.2% and 76.9%, respectively) compared with CT and MRI (44.8% and 43.1%, respectively).

XSL•FO
RenderX

Expert-based validation showed a similar trend—ultrasound (61.1%) and mammography (61.5%) exceeded CT (32.1%) and MRI (25.5%).

Preferences for explainability methods also differed by modality. Most mammography studies used Grad-CAM (19/46, 42.3%), while SHAP was used in approximately 7% (3/46). In contrast, CT studies used SHAP more often (50/139, ~35.8%) than Grad-CAM (29/139, 20.9%). MRI studies showed a more balanced use, with Grad-CAM (30/104, ~28.4%) and SHAP (22/104, ~22.5%).

Validation of Explainability

Overview

In this review, we examined how studies validated their explainability components, the metrics and strategies adopted, and the extent of user involvement. Among all included studies, 193 (52%) reported at least 1 form of xAI validation. Because some studies used multiple validation strategies, categories overlap, and totals can exceed 193. *Expert or user-based validation* was reported in 104 studies (28% of all studies; 53.9% of validated studies). *Mixed methods validation* was reported in 74 studies (19.9% of all studies; 38.3% of validated studies). *Quantitative metrics* were reported in 10 studies (2.7% of all studies; 5.2% of validated studies), and *domain or clinical-knowledge validation* in 8 studies (2.2% of all studies; 4.1% of validated studies).

Validation From Experts and Users

The majority of studies (n=104/193, 53%) based their validation primarily on expert feedback obtained from doctors, radiologists, or clinicians. They assessed whether the explanations aligned with medical experts' reasoning during diagnosis. This was achieved through interviews with them, surveys of readers, or usability tests to determine if the AI explanations were logical and understandable. Even when researchers did not explicitly label their studies as "validation of xAI," many included some form of evaluation of the usability and reasoning of the AI's explanations, contributing to the overall objective of validating the AI's explanatory ability.

Quantitative Measurement and Model Comparison

Fewer studies (10/193, approximately 5%) relied on indirect assessment of explanation quality using performance metrics or model comparisons. In particular, studies reported using various metrics to measure the quality of AI-generated explanations. Among the most common metrics were area under the curve (AUC), precision-recall AUC, and feature ranking consistencies. However, most of the time, these metrics assess the quality of the model's predictions rather than the explanations themselves.

Validation Related to Specific Domains or Clinical Applications

Some studies validated explanations by checking their consistency with current clinical and/or biological knowledge (8/193, 4%). For example, they verified agreement between the model's predictions and histopathological characteristics; compared the model's explanatory features with radiologists' accepted standards to ensure consistency; or tested whether the explanatory features remained consistent across all data partitions. All these methods help demonstrate the validity of the clinical application of the xAI system in medicine.

Mixed Methods Validation

In addition to the above types of validation, some studies (74/193, 37.7%) combined 2 or more methods to validate the quality of AI explanations. Most often, this mixed methods approach involved combining expert feedback with one of the 3 types of quantitative validation described previously. By doing so, the authors aimed to balance the need for objectivity in explanation quality with the need to understand experts' perspectives on trustworthiness and usability.

For example, studies such as the one by Nowakowska et al [101] used SHAP-based feature importance analysis alongside expert radiologist review to validate the clinical coherence of model explanations. Finally, another study showed that AUC-based model performance can be triangulated with reader-study feedback to validate the alignment of visual explanations with diagnostic reasoning [48,102].

Fairness and Bias

Beyond validation, trustworthy explainable AI in health care also depends on fairness and bias mitigation. While explainability improves interpretability, it does not inherently guarantee equitable or unbiased outcomes. Only a limited number of studies explicitly assessed bias within their explainability frameworks or examined how explanations might vary across patient subgroups [51,102,103]. The intersection of explainability and fairness remains underexplored.

Code Availability and Decision-Support Integration

Code Availability

Based on the review of 371 studies, most of them did not share their code (Figure 3). Of the 371 reviewed studies, only 17.5% (65/371) reported sharing their full or partial implementation via a platform (such as GitHub) or upon request. In contrast, 82.5% (306/371) provided no access to their source code, a figure that includes studies where no code availability information was reported, which were confirmed through subset sampling to represent a lack of shared implementation. The failure to make source code available is a major hindrance to reproducing any findings, preventing peers from verifying the work and limiting advancements through collaboration in the field.

Despite the overall low transparency, a longitudinal analysis reveals a positive shift in open science practices. From 2017 to 2023, only 15.8% (27/171) of studies provided public or partial access to their code. However, among studies published in 2024, this proportion rose to 19% (38/200). While the volume of "no" results is increasing in absolute terms due to the explosion of xAI research, the relative rate of code sharing is showing a steady upward trend.

XSL•FO
RenderX

Figure 3. Time trend of code availability among included explainable artificial intelligence cancer imaging studies (2017 - 2024; N=371).

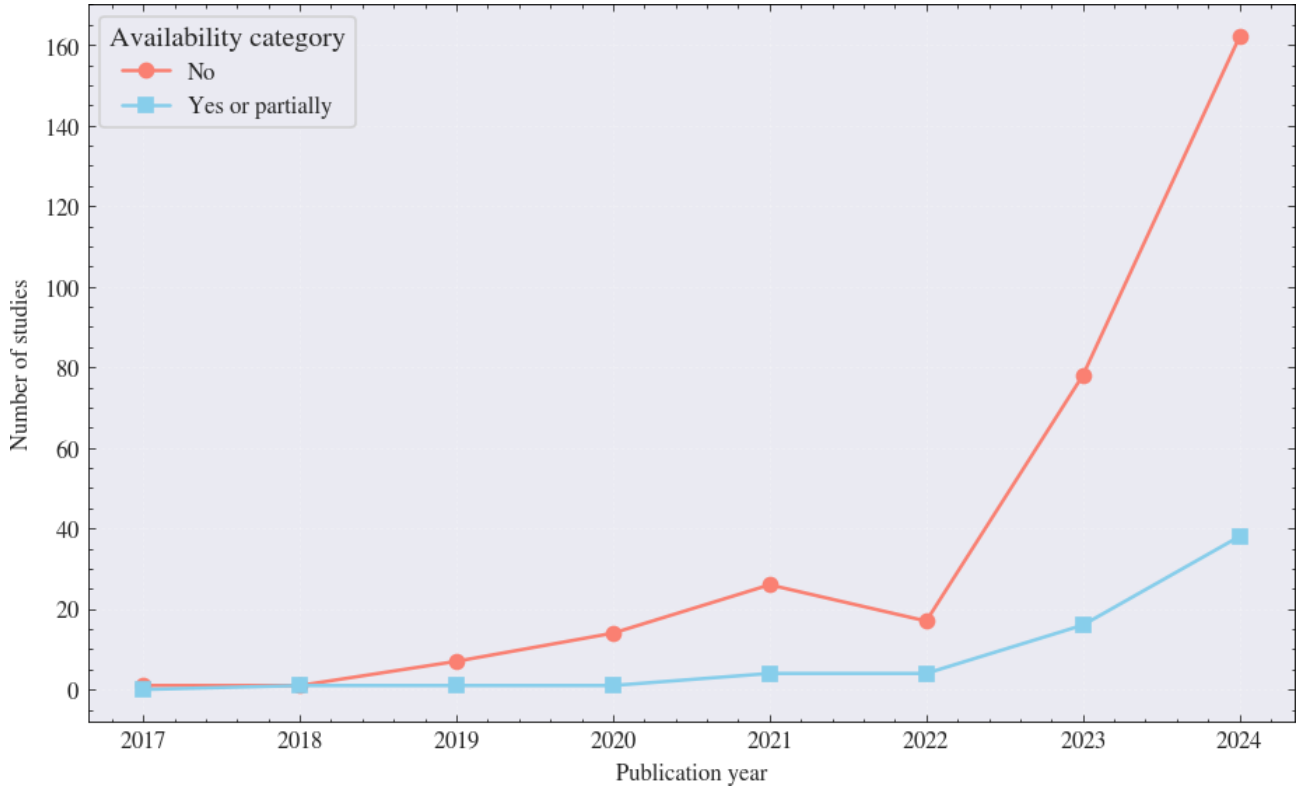

DSS Integration

Out of 371 studies, only 45 (12.1%) reported full or partial DSS integration, while 321 (86.5%) did not, and 5 (1.3%) were unspecified (Figure 4). A temporal comparison shows a modest increase in reported integration; only 11 out of 171 studies (6.4%) from 2017 to 2023 mentioned DSS use, compared with 34 out of 200 (17%) in 2024. Among all the 2024 studies, the majority discussed potential or planned integration (69/200, 34.5%), often in the form of prototypes, user interfaces, or frameworks intended for future translation, while fewer describe actual deployment within clinical workflows (34/200, 17%).

Figure 4. Integration of Explainable artificial intelligence models into decision support systems among included studies (2017 - 2024; N=371).

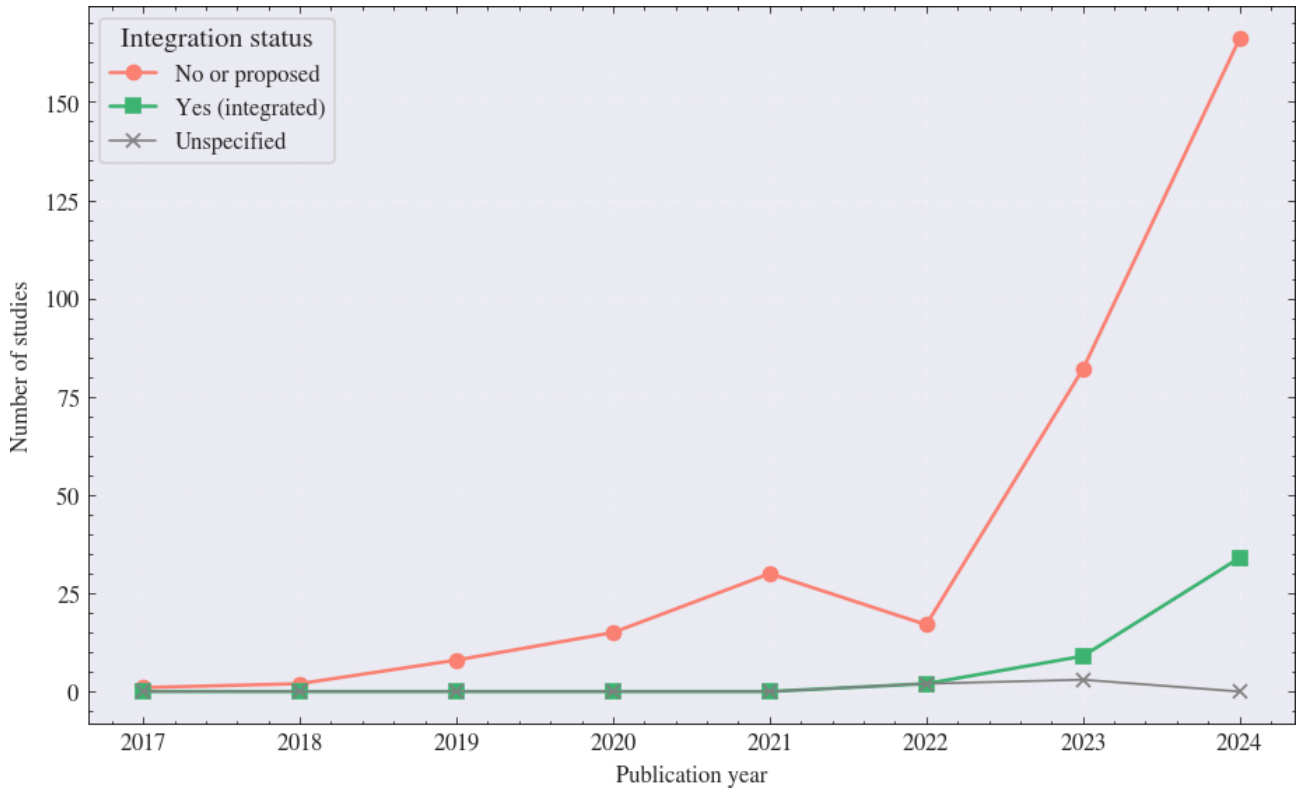

XSL•FO
RenderX

Summary of Evidence Patterns

Across the 371 included studies, research activity was concentrated in breast, lung, and brain or CNS cancers (together 255/371, approximately 70% of studies). DL was the dominant primary modeling approach (260/371, 70.1%), and post hoc explainability methods were used in most studies (305/371, 82.2%), whereas ante hoc (intrinsic) explainability approaches were uncommon (21/371, 5.7%). Explainability validation was reported in 52% (193/371) of studies, most frequently expert or user-based evaluation. Reproducibility and translation indicators were limited; 17.5% (65/371) of studies reported code availability, and 12% (45/371) reported DSS integration; an additional 18.6% (69/371) proposed clinical integration without reporting deployment.

Discussion

Summary of Findings

In this review, we examined how explainability is integrated in cancer imaging AI studies, how explanation techniques are assessed and reported, and the extent to which current reporting facilitates reproducibility and clinical use. Overall, explainability in AI studies is primarily implemented through post hoc methods, while intrinsically interpretable and hybrid approaches remain relatively uncommon. Thus, in much of the literature, transparency is treated as an add-on to predictive modeling rather than as a fundamental aspect of model design.

Furthermore, we identified significant variability in the appraisal of xAI methods. Many studies used qualitative or plausibility-based assessments, and expert or user feedback was the most reported method for evaluating explainability. However, these evaluations were generally conducted without more rigorous methods for testing the fidelity and consistency of explanations against existing domain knowledge. Conversely, the use of formal metrics for measuring explanation quality varied significantly between studies.

Our results show limited support for indicators for reproducibility and translation to clinical settings. Code sharing was infrequent, data access was often restricted, and only a small number of studies demonstrated integration of their models into clinical decision-making systems. These findings suggest that although there is an expanding body of evidence related to xAI, further development is needed to standardize the evaluation of explainability and to establish stronger clinical relevance.

Interpretation and Comparison With Previous Literature

Explainability Trends and Methodological Implications

Consistent with previous reviews in medical imaging, the dominance of post hoc explainability methods, including Grad-CAM and SHAP, reflects a trend similar to that observed in earlier literature [104,105]. This dominance is largely due to the relative ease of implementation and versatility of visual attribution methods, as well as their compatibility with a wide range of model architectures, particularly CNNs [106,107]. However, these methods generally provide qualitative explanations, which can contain a subjective side and be significantly affected by perturbation [104,108].

Although the majority of studies reporting xAI validation used some form of expert feedback, these assessments were typically limited to plausibility judgments, asking whether explanations appeared reasonable, rather than a systematic evaluation of whether model attributions corresponded to established clinical or biological knowledge.

The widespread use of SHAP across studies, driven by its ease of access and model-agnostic applicability, should be interpreted cautiously when evaluating clinical confidence. Fewer than one-third of the studies that employed SHAP included expert-based validation to assess alignment with clinical reasoning. This is notable given evidence that SHAP-derived attributions may have limited fidelity to actual model behavior [104], suggesting that technical interpretability should not be automatically equated with clinical transparency or alignment with radiological cognition.

As noted in our results, there appears to be limited adoption of inherently interpretable models, such as Explainable Boosting Machine or generalized additive model, which may be attributed to the perception—and empirical documentation—of an accuracy-interpretability tradeoff, with interpretable models showing a consistent 5% - 7% AUC penalty compared with black box approaches [104]. At the same time, the emergence of hybrid models that combine DL with traditional ML classifiers suggests growing interest in architectures that offer improved transparency while maintaining high model performance. This progression reflects trends observed in other recent surveys of explainable medical AI [109].

The stated purposes of explainability in the reviewed studies became more specific over time, shifting from general descriptions to articulating concrete clinical objectives. This trend may reflect the field's maturation from demonstrating technical feasibility to addressing how explanations should be designed for clinical utility.

Modality-Specific Considerations

We identified CT and MRI as the two most widely used modalities in our study, although the rate at which xAI methods are validated and their application varies greatly depending on the modality. In the modality-specific subset analysis, we found that ultrasound and mammography studies reported validation of xAI more frequently than CT and MRI studies. There were also differences in method preference among the modalities. For example, mammography-based studies used Grad-CAM more often, while CT-based studies used SHAP more frequently. These differences in modality-specific application suggest that xAI may not be universally applied. Instead, xAI design and validation procedures will likely need to be tailored to each specific modality and workflow [110,111].

Validation Practices and Their Limitations

While approximately half of the reviewed studies reported some form of explainability validation, the approaches and rigor varied considerably. Expert-based validation predominated, typically involving radiologists or clinicians assessing whether

XSL•FO
RenderX

explanations appeared plausible or aligned with their diagnostic reasoning. However, very few studies implemented deeper feedback loops—whether from expert critique or from xAI evaluation outcomes—to iteratively refine model features or architectures. This pattern suggests that validation often serves as post hoc verification rather than as an integral component of model development.

Quantitative validation methods, although less common, were typically indirect—relying on performance metrics (AUC and accuracy) that assess prediction quality rather than explanation fidelity. Few studies used formal metrics specifically designed to evaluate explanation quality, consistency, or clinical utility. This methodological gap, combined with the lack of standardized frameworks for assessing explanations, leaves most validation efforts qualitative and subjective [110].

Several limitations emerged across the reviewed literature. Many studies acknowledged that their models lacked multicenter or external validation, limiting confidence in both predictions and explanations. Training on small or imbalanced datasets has been associated with instability in both model predictions [112,113] and attribution patterns [104]. Researchers also noted that heatmap interpretations remain subjective and may not correspond to clinical reasoning, risking misinterpretation by users lacking appropriate context [104]. Additionally, the complexity or opacity of some model architectures can render generated explanations less actionable in clinical settings.

Importantly, validation results were rarely used to modify model design. Explainability was typically incorporated after model development and validation were complete, a pattern consistent with the predominance of post hoc rather than in-model approaches observed across the broader medical AI literature [114,115]. This post hoc approach limits opportunities for iterative refinement and may contribute to overestimation of clinical readiness.

Clinical Translation, Trustworthiness, and Reproducibility

Clinical translation of xAI remains limited due to multiple interconnected barriers. Although approximately half of the studies included some form of expert-based validation, this rarely extended beyond plausibility assessment to active clinical integration or workflow testing. This gap is consistent with previous criticisms that explanation methods do not necessarily reflect clinical reasoning and may lead to user misinterpretation [116]. Few studies included clinicians as active participants in validation, and even fewer incorporated clinical feedback to assess whether model explanations aligned with actual decision-making processes or improved workflow integration. Structured frameworks offer pathways for human-in-the-loop validation and user-centered evaluation, yet their adoption remains limited. The low percentage of reported DSS system deployments, along with the limited number of documented clinical applications, clearly shows that most systems are in the early stages of proof-of-concept models—disconnected from real-world clinical environments. However, the increase in studies conducted during the last year of this review that focus on clinical applicability indicates that the field is at a developmental stage where translation to clinical practice will be feasible.

Additionally, there are significant barriers to reproducibility. Code sharing was rare, and reliance on private or mixed datasets limits independent replication of findings. These patterns mirror challenges in the broader medical AI literature, where transparency, data governance, and external validation continue to be significant barriers to progress [117,118]. Similarly, fairness and subgroup analysis were rarely reported, leaving critical questions about algorithmic equity unanswered. In summary, while xAI methods are being widely adopted in research settings, limited validation rigor, minimal integration of clinical feedback, and insufficient attention to fairness may collectively contribute to an overestimation of clinical readiness relative to what current evidence supports.

Limitations of This Review

There are 4 main limitations of this study. First, the studies we examined showed considerable variation in how explainability was defined and implemented, making the classification as either post hoc or intrinsically interpretable subjective, even though we used standardized definitions. Second, the search was limited to PubMed and Scopus and to studies published in English, which may have led to missing relevant studies, particularly those published in computer science journals. Third, because we examined only radiologic imaging, the generalizability of our findings to other biomedical imaging disciplines, such as histopathology or dermoscopy, is uncertain. Fourth, as the field of xAI is rapidly advancing, we did not include studies published after December 2024, and the observed growth in xAI-related publications may partly reflect changes in indexing practices or reporting standards rather than a true increase in research activity.

Conclusions

This scoping review provides a comprehensive mapping of xAI methods across multiple cancer imaging modalities, addressing a gap left by previous reviews that focused on single modalities or isolated aspects of interpretability. By quantifying key indicators, including validation rates (52%), code availability (17.5%), and DSS integration (12.1%), this study offers an evidence base for researchers, clinicians, and regulators to understand the current state of xAI readiness for clinical use. The main opportunities to advance xAI in cancer imaging include validating current explainability methods—given that only 52% (193/371) of studies reported any form of explainability validation—improving data access, increasing the use of shared code, and fostering a culture of openness and trust. These elements of reproducibility and transparency are essential for developing reliable clinical AI models. Reporting guidelines, such as the Checklist for Artificial Intelligence in Medical Imaging [119], and broader frameworks (including Fairness, Universality, Traceability, Usability, Robustness, and Explainability [-AI] [118], and Developmental and Exploratory Clinical Investigations of Decision Support Systems [120]) structured requirements for transparency and trustworthy AI development; however, their full potential can only be realized through widespread adoption by the medical imaging community.

XSL•FO
RenderX

Progress in applying xAI to cancer imaging will also depend on the level of collaboration among developers, clinicians, and domain experts. Meaningful and clinically actionable explanations require clinically-informed design, early stakeholder involvement, and user-centered evaluation aligned with specific tasks and users (eg, radiologists, oncologists, or patients). Establishing common benchmark standards and standardized reporting will further support study reproducibility and comparison, and clinical translation. For regulators and institutions, these findings highlight that most xAI systems remain at the prototype stage. With standardized validation, collaborative development, and a multidisciplinary design, xAI systems can evolve from technically interpretable prototypes to clinically trustworthy tools that genuinely support decision-making in oncologic imaging.

Acknowledgments

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: code generation, proofreading and editing, and translation. The GenAI tool used was Claude 4 (Anthropic). Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the outcomes. The code was used to produce the figures in this manuscript.

Funding

This research was funded by the European Union program (European Cancer Imaging Initiative; grant 101100633).

Data Availability

The dataset generated and analyzed during this review consists of study-level information extracted from published articles. Due to the inclusion of citation details and unstandardized formatting, the data are not publicly available at this time. However, the spreadsheet can be obtained from the corresponding author upon reasonable request.

Authors' Contributions

The literature review and data extraction were performed by D Fotopoulos, D Filos, and IL. PAMS contributed by cross-checking a subset of articles screened by the other authors. D Fotopoulos conducted the analysis and drafted the manuscript. PAMS, MvG, and IC provided supervision and critical revision of manuscript drafts. All authors approved the final version for submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full search queries.
[DOCX File, 25 KB - jmir_v28i1e80645_app1.docx]

Multimedia Appendix 2
Data extraction schema.
[DOCX File, 30 KB - jmir_v28i1e80645_app2.docx]

Multimedia Appendix 3
Additional studies.
[DOCX File, 51 KB - jmir_v28i1e80645_app3.docx]

Multimedia Appendix 4
Additional figures.
[DOCX File, 1953 KB - jmir_v28i1e80645_app4.docx]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 129 KB - jmir_v28i1e80645_app5.docx]

References

XSL•FO
RenderX

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021 May;71(3):209-249. [doi: 10.3322/caac.21660] [Medline: 33538338]
2. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74(3):229-263. [doi: 10.3322/caac.21834] [Medline: 38572751]
3. Salinas MP, Sepúlveda J, Hidalgo L, et al. A systematic review and meta-analysis of artificial intelligence versus clinicians for skin cancer diagnosis. NPJ Digit Med 2024 May 14;7(1):125. [doi: 10.1038/s41746-024-01103-x] [Medline: 38744955]
4. Al-Obeidat F, Hafez W, Gador M, et al. Diagnostic performance of AI-based models versus physicians among patients with hepatocellular carcinoma: a systematic review and meta-analysis. Front Artif Intell 2024;7:1398205. [doi: 10.3389/frai.2024.1398205] [Medline: 39224209]
5. Tao J, Liu D, Hu FB, et al. Development and validation of a computed tomography-based model for noninvasive prediction of the T Stage in gastric cancer: multicenter retrospective study. J Med Internet Res 2024 Oct 9;26:e56851. [doi: 10.2196/56851] [Medline: 39382960]
6. Sivakumar R, Lue B, Kundu S. FDA approval of artificial intelligence and machine learning devices in radiology: a systematic review. JAMA Netw Open 2025 Nov 3;8(11):e2542338. [doi: 10.1001/jamanetworkopen.2025.42338] [Medline: 41201805]
7. Amann J, Vetter D, Blomberg SN, et al. To explain or not to explain?-Artificial intelligence explainability in clinical decision support systems. PLOS Digit Health 2022 Feb;1(2):e0000016. [doi: 10.1371/journal.pdig.0000016] [Medline: 36812545]
8. Aboy M, Minssen T, Vayena E. Navigating the EU AI act: implications for regulated digital medical products. NPJ Digit Med 2024 Sep 6;7(1):237. [doi: 10.1038/s41746-024-01232-3] [Medline: 39242831]
9. Artificial intelligence and amending regulations. European Union. 2024 Jun 13. URL: http://data.europa.eu/eli/reg/2024/1689/oj [accessed 2026-01-20]
10. Vardas EP, Marketou M, Vardas PE. Medicine, healthcare and the AI act: gaps, challenges and future implications. Eur Heart J Digit Health 2025 Jul;6(4):833-839. [doi: 10.1093/ehjdh/ztaf041] [Medline: 40703136]
11. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. JAMA 2018 Dec 4;320(21):2199-2200. [doi: 10.1001/jama.2018.17163] [Medline: 30398550]
12. Kann BH, Hosny A, Aerts H. Artificial intelligence for clinical oncology. Cancer Cell 2021 Jul 12;39(7):916-927. [doi: 10.1016/j.ccell.2021.04.002] [Medline: 33930310]
13. Bi WL, Hosny A, Schabath MB, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. CA Cancer J Clin 2019 Mar;69(2):127-157. [doi: 10.3322/caac.21552] [Medline: 30720861]
14. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov 2019;9(4):e1312. [doi: 10.1002/widm.1312] [Medline: 32089788]
15. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019 Oct 29;17(1):195. [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
16. Ennab M, Mcheick H. Designing an interpretability-based model to explain the artificial intelligence algorithms in healthcare. Diagnostics (Basel) 2022 Jun 26;12(7):1557. [doi: 10.3390/diagnostics12071557] [Medline: 35885463]
17. Toumaj S, Heidari A, Jafari Navimipour N. Leveraging explainable artificial intelligence for transparent and trustworthy cancer detection systems. Artif Intell Med 2025 Nov;169:103243. [doi: 10.1016/j.artmed.2025.103243] [Medline: 40839960]
18. Nilsen P, Svedberg P, Larsson I, et al. Radiology staff experiences with integrating artificial intelligence into radiology practice in a swedish hospital: qualitative case study. JMIR Form Res 2025 Dec 22;9:e77843. [doi: 10.2196/77843] [Medline: 41428898]
19. Oh SH, Lee Y, Baek JH, Sunwoo W. Deep learning and image generator health tabular data (IGHT) for predicting overall survival in patients with colorectal cancer: retrospective study. JMIR Med Inform 2025 Aug 19;13:e75022. [doi: 10.2196/75022] [Medline: 40828572]
20. Wyatt LS, van Karnenbeek LM, Wijkhuizen M, Geldof F, Dashtbozorg B. Explainable artificial intelligence (XAI) for oncological ultrasound image analysis: a systematic review. Appl Sci (Basel) 2024;14(18):8108. [doi: 10.3390/app14188108]
21. de Vries BM, Zwezerijnen GJC, Burchell GL, van Velden FHP, Menke-van der Houven van Oordt CW, Boellaard R. Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review. Front Med (Lausanne) 2023;10:1180773. [doi: 10.3389/fmed.2023.1180773] [Medline: 37250654]
22. Borys K, Schmitt YA, Nauta M, et al. Explainable AI in medical imaging: an overview for clinical practitioners - saliency-based XAI approaches. Eur J Radiol 2023 May;162:110787. [doi: 10.1016/j.ejrad.2023.110787] [Medline: 37001254]
23. Farahani FV, Fiok K, Lahijanian B, Karwowski W, Douglas PK. Explainable AI: a review of applications to neuroimaging data. Front Neurosci 2022;16:906290. [doi: 10.3389/fnins.2022.906290] [Medline: 36583102]
24. Reyes M, Meier R, Pereira S, et al. On the interpretability of artificial intelligence in radiology: challenges and opportunities. Radiol Artif Intell 2020 May 27;2(3):e190043. [doi: 10.1148/ryai.2020190043] [Medline: 32510054]

XSL•FO
RenderX

25. Koutoulakis E, Trivizakis E, Markodimitrakis E, Agelaki S, Tsiknakis M, Marias K. A critical review of explainable deep learning in lung cancer diagnosis. Artif Intell Rev 2025;59(1):28. [doi: 10.1007/s10462-025-11445-x]
26. Shifa N, Saleh M, Akbari Y, Al Maadeed S. A review of explainable AI techniques and their evaluation in mammography for breast cancer screening. Clin Imaging 2025 Jul;123:110492. [doi: 10.1016/j.clinimag.2025.110492] [Medline: 40378639]
27. Skliarov M, Shawi RE, Dhaoui C, Ahmed N. A comparative evaluation of explainability techniques for image data. Sci Rep 2025 Nov 25;15(1):41898. [doi: 10.1038/s41598-025-25839-y] [Medline: 41290895]
28. Chouvarda I, Colantonio S, Verde ASC, et al. Differences in technical and clinical perspectives on AI validation in cancer imaging: mind the gap!. Eur Radiol Exp 2025 Jan 15;9(1):7. [doi: 10.1186/s41747-024-00543-0] [Medline: 39812924]
29. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst 2021 Nov;32(11):4793-4813. [doi: 10.1109/TNNLS.2020.3027314] [Medline: 33079674]
30. Venkatesh K, Santomartino SM, Sulam J, Yi PH. Code and data sharing practices in the radiology artificial intelligence literature: a meta-research study. Radiol Artif Intell 2022 Sep;4(5):e220081. [doi: 10.1148/ryai.220081] [Medline: 36204536]
31. Martí-Bonmatí L, Blanquer I, Tsiknakis M, et al. Empowering cancer research in Europe: the EUCAIM cancer imaging infrastructure. In: Blázquez J, Pedraza S, López-Rueda A, Leo G, Roussakis Y, Sala E, et al, editors. Insights Imaging 2025 Feb 24;16(1):47. [doi: 10.1186/s13244-025-01913-x] [Medline: 39992532]
32. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev 2021 Mar 29;10(1):89. [doi: 10.1186/s13643-021-01626-4] [Medline: 33781348]
33. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
34. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 2020 Jun;58:82-115. [doi: 10.1016/j.inffus.2019.12.012]
35. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? arXiv. Preprint posted online on Dec 28, 2017. [doi: 10.48550/ARXIV.1712.09923]
36. Roscher R, Bohn B, Duarte MF, Garcke J. Explainable machine learning for scientific insights and discoveries. IEEE Access 2020;8:42200-42216. [doi: 10.1109/ACCESS.2020.2976199]
37. Lipton ZC. The mythos of model interpretability. Commun ACM 2018 Sep 26;61(10):36-43. [doi: 10.1145/3233231]
38. Arya V, Bellamy RKE, Chen PY, et al. One explanation does not fit all: a toolkit and taxonomy of AI explainability techniques. . Preprint posted online on 2019. [doi: 10.48550/ARXIV.1909.03012]
39. High-level expert group on artificial intelligence. : European Commission; 2019 URL: https://www.aepd.es/sites/default/files/2019-12/ai-definition.pdf [accessed 2026-05-01]
40. Diakopoulos N. Accountability in algorithmic decision making. Commun ACM 2016 Jan 25;59(2):56-62. [doi: 10.1145/2844110]
41. Ribeiro MT, Singh S, Guestrin C. Why should i trust you?": explaining the predictions of any classifier. 2016 Presented at: KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ACM; Aug 13-17, 2016 p. 1135-1144. [doi: 10.1145/2939672.2939778]
42. Lundberg S, Lee SI. A unified approach to interpreting model predictions. 2017 Nov 25 Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; Dec 4-9, 2017 URL: https://dl.acm.org/doi/10.5555/3295222.3295230 [accessed 2025-05-01]
43. Wang SH, Chen G, Zhong X, et al. Global development of artificial intelligence in cancer field: a bibliometric analysis range from 1983 to 2022. Front Oncol 2023;13:1215729. [doi: 10.3389/fonc.2023.1215729] [Medline: 37519796]
44. Hung SC, Wang YT, Tseng MH. An interpretable three-dimensional artificial intelligence model for computer-aided diagnosis of lung nodules in computed tomography images. Cancers (Basel) 2023 Sep 21;15(18):4655. [doi: 10.3390/cancers15184655] [Medline: 37760624]
45. Han X, Chang L, Song K, Cheng L, Li M, Wei X. Multitask network for thyroid nodule diagnosis based on TI‐RADS. Med Phys Mex Symp Med Phys 2022 Aug;49(8):5064-5080 [FREE Full text] [doi: 10.1002/mp.15724]
46. Pathak S, Schlötterer J, Veltman J, Geerdink J, Keulen M, Seifert C. Prototype-based interpretable breast cancer prediction models: analysis and challenges. : Springer; 2024 Jul 19. [doi: 10.1007/978-3-031-63787-2_2]
47. Yang F, Jemaa S, Bengtsson T, Ghaoui LE. Metastatic lung cancer prognosis via deep image-based lesion prioritization. 2024 Presented at: 2024 IEEE International Symposium on Biomedical Imaging (ISBI); May 27-30, 2024; Athens, Greece p. 1-5. [doi: 10.1109/ISBI56570.2024.10635163]
48. Shen S, Han SX, Aberle DR, Bui AA, Hsu W. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. Expert Syst Appl 2019 Aug 15;128:84-95. [doi: 10.1016/j.eswa.2019.01.048] [Medline: 31296975]
49. Morgan HE, Wang K, Dohopolski M, et al. Exploratory ensemble interpretable model for predicting local failure in head and neck cancer: the additive benefit of CT and intra-treatment cone-beam computed tomography features. Quant Imaging Med Surg 2021 Dec;11(12):4781-4796. [doi: 10.21037/qims-21-274] [Medline: 34888189]
50. Zhang R, Zhu H, Chen M, et al. A dual-radiomics model for overall survival prediction in early-stage NSCLC patient using pre-treatment CT images. Front Oncol 2024;14:1419621. [doi: 10.3389/fonc.2024.1419621] [Medline: 39206157]

XSL•FO
RenderX

51. Orton MR, Hann E, Doran SJ, et al. Interpretability of radiomics models is improved when using feature group selection strategies for predicting molecular and clinical targets in clear-cell renal cell carcinoma: insights from the TRACERx renal study. Cancer Imaging 2023 Aug 14;23(1):76. [doi: 10.1186/s40644-023-00594-3] [Medline: 37580840]
52. Magoulianitis V, Yang J, Yang Y, et al. PCa-RadHop: a transparent and lightweight feed-forward method for clinically significant prostate cancer segmentation. Comput Med Imaging Graph 2024 Sep;116:102408. [doi: 10.1016/j.compmedimag.2024.102408] [Medline: 38908295]
53. Xiong Y, Zheng Y, Long W, et al. Study on microwave ablation temperature prediction model based on grayscale ultrasound texture and machine learning. PLoS ONE 2024;19(9):e0308968. [doi: 10.1371/journal.pone.0308968] [Medline: 39321182]
54. Prinzi F, Orlando A, Gaglio S, Vitabile S. Breast cancer classification through multivariate radiomic time series analysis in DCE-MRI sequences. Expert Syst Appl 2024 Sep;249:123557. [doi: 10.1016/j.eswa.2024.123557]
55. Rifi AL, Raets C, Dufait I, De Ridder M, Barbé K. Integrating radiomics and immunology: non-invasive assessment of CD8+ T cell levels. 2024 Presented at: 2024 IEEE International Instrumentation and Measurement Technology Conference (I2MTC); May 20-23, 2024; Glasgow, United Kingdom p. 1-4. [doi: 10.1109/I2MTC60896.2024.10560686]
56. Lv T, Hong X, Liu Y, et al. AI-powered interpretable imaging phenotypes noninvasively characterize tumor microenvironment associated with diverse molecular signatures and survival in breast cancer. Comput Methods Programs Biomed 2024 Jan;243:107857. [doi: 10.1016/j.cmpb.2023.107857] [Medline: 37865058]
57. Duan C, Liu Q, Wang J, et al. GWO+RuleFit: rule-based explainable machine-learning combined with heuristics to predict mid-treatment FDG PET response to chemoradiation for locally advanced non-small cell lung cancer. Phys Med Biol 2024 Jul 23;69(15). [doi: 10.1088/1361-6560/ad6118] [Medline: 38981590]
58. Hasan Y, De LA, Amerehi F, De BD, Healy P, Ryan C. Interpretable solutions for breast cancer diagnosis with grammatical evolution and data augmentation. In: Smith S, Correia J, Cintrano C, editors. Appl Evol Comput (2014): Springer Nature; 2024:224-239. [doi: 10.1007/978-3-031-56852-7_15]
59. Kaushik S, Lamba AK, Kansal I, Khullar V, Sharma P. Explainable deep learning for lung cancer detection: comparing CNN and DenseNet201 with Grad-CAM. 2024 Presented at: 2024 2nd International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES); Dec 19-21, 2024. [doi: 10.1109/SCOPES64467.2024.10990463]
60. Zhukov A, Benois-Pineau J, Giot R. Evaluation of explanation methods of AI - CNNs in image classification tasks with reference-based and no-reference Metrics. AAIML 2023;03(1):620-646 [FREE Full text] [doi: 10.54364/AAIML.2023.1143]
61. Yao MMS, Du H, Hartman M, Chan WP, Feng M. End-to-end calcification distribution pattern recognition for mammograms: an interpretable approach with GNN. Diagnostics (Basel) 2022 Jun 2;12(6):1376. [doi: 10.3390/diagnostics12061376] [Medline: 35741186]
62. Fan M, Huang G, Lou J, Gao X, Zeng T, Li L. Cross-parametric generative adversarial network-based magnetic resonance image feature synthesis for breast lesion classification. IEEE J Biomed Health Inform 2023 Nov;27(11):5495-5505. [doi: 10.1109/JBHI.2023.3311021] [Medline: 37656652]
63. Addala V. BREAST AI: low cost, explainable artificial intelligence based app for efficient diagnosis of BREAST cancer in developing areas. 2023 Presented at: 2023 IEEE 3rd International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB); Apr 14-16, 2023; Taichung, Taiwan p. 164-167. [doi: 10.1109/ICEIB57887.2023.10170357]
64. Alshuhail A, Thakur A, Chandramma R, et al. Refining neural network algorithms for accurate brain tumor classification in MRI imagery. BMC Med Imaging 2024 May 21;24(1):118. [doi: 10.1186/s12880-024-01285-6] [Medline: 38773391]
65. M MM, T R M, V VK, Guluwadi S. Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50. BMC Med Imaging 2024 May 11;24(1):107. [doi: 10.1186/s12880-024-01292-7] [Medline: 38734629]
66. Sarker S. Transfer learning and explainable AI for brain tumor classification: a study using MRI data from Bangladesh. 2024 Presented at: 2024 6th International Conference on Sustainable Technologies for Industry 50 (STI); Dec 14-15, 2024. [doi: 10.1109/STI64222.2024.10951092]
67. Carloni G, Pachetti E, Colantonio S. Causality-driven one-shot learning for prostate cancer grading from MRI. 2023 Sep 19 Presented at: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW); Oct 2-6, 2023. [doi: 10.1109/ICCVW60793.2023.00276]
68. Kunapinun A, Songsaeng D, Buathong S, Dailey MN, Keatmanee C, Ekpanyapong M. Explainable automated TI-RADS evaluation of thyroid nodules. Sensors (Basel) 2023 Aug 21;23(16):7289. [doi: 10.3390/s23167289] [Medline: 37631825]
69. Liu Y, Feng Y, Qian L, Wang Z, Hu X. Deep learning diagnostic performance and visual insights in differentiating benign and malignant thyroid nodules on ultrasound images. Exp Biol Med (Maywood) 2023 Dec;248(24):2538-2546. [doi: 10.1177/15353702231220664] [Medline: 38279511]
70. Burgos D, Morshed A, Rashid MM, Mandala S. A comparison of machine learning models to deep learning models for cancer image classification and explainability of classification. 2024 Presented at: 2024 International Conference on Data Science and Its Applications (ICoDSA); Jul 10-11, 2024; Kuta, Bali, Indonesia p. 386-390. [doi: 10.1109/ICoDSA62899.2024.10651790]
71. Kumar K, Jyoti K. Recent advancements in Grad-CAM and variants: enhancing brain tumor detection, segmentation, and classification. Research Square. Preprint posted online on 2024 URL: https://europepmc.org/article/ppr/ppr949703 [accessed 2026-05-01] [doi: 10.21203/rs.3.rs-5485128/v1]

XSL•FO
RenderX

72. Lee IC, Huang JY, Chen TC, et al. Evolutionary learning-derived clinical-radiomic models for predicting early recurrence of hepatocellular carcinoma after resection. Liver Cancer 2021 Nov;10(6):572-582. [doi: 10.1159/000518728] [Medline: 34950180]
73. Smedley NF, Aberle DR, Hsu W. Using deep neural networks and interpretability methods to identify gene expression patterns that predict radiomic features and histology in non-small cell lung cancer. J Med Imaging (Bellingham) 2021 May;8(3):031906. [doi: 10.1117/1.JMI.8.3.031906] [Medline: 33977113]
74. Wang CJ, Hamm CA, Savic LJ, et al. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. Eur Radiol 2019 Jul;29(7):3348-3357. [doi: 10.1007/s00330-019-06214-8] [Medline: 31093705]
75. Visonà G, Spiller LM, Hahn S, et al. Machine-learning-aided prediction of brain metastases development in non-small-cell lung cancers. Clin Lung Cancer 2023 Dec;24(8):e311-e322. [doi: 10.1016/j.cllc.2023.08.002] [Medline: 37689579]
76. Ma X, Ning F, Xu X, et al. Survival prediction for non-small cell lung cancer based on multimodal fusion and deep learning. IEEE Access 2024;12:123236-123249. [doi: 10.1109/ACCESS.2024.3453930] [Medline: 39726803]
77. Liu W, Zhou M, Mo H, Zhao S. A multimodal breast cancer diagnosis approach based on efficientnet and hybrid feature selection. 2024 Presented at: 2024 China Automation Congress (CAC); Qingdao, China p. 7339-7343. [doi: 10.1109/CAC63892.2024.10865739]
78. Surkov YI, Serebryakova IA, Kuzinova YK, et al. Multimodal method for differentiating various clinical forms of basal cell carcinoma and benign neoplasms in vivo. Diagnostics (Basel) 2024 Jan 17;14(2):202. [doi: 10.3390/diagnostics14020202] [Medline: 38248078]
79. Wei L, Owen D, Rosen B, et al. A deep survival interpretable radiomics model of hepatocellular carcinoma patients. Phys Med 2021 Feb;82:295-305. [doi: 10.1016/j.ejmp.2021.02.013] [Medline: 33714190]
80. Charalampakos F, Tsouparopoulos T, Koutsopoulos I. Joint explainability-performance optimization with surrogate models for AI-driven edge services. 2025 Presented at: 2025 IEEE International Conference on Machine Learning for Communication and Networking (ICMLCN); May 26-29, 2025; Barcelona, Spain p. 1-6. [doi: 10.1109/ICMLCN64995.2025.11140577]
81. Contreras V, Bagante A, Marini N, Schumacher M, Andrearczyk V, Calvaresi D. Explanation Generation via Decompositional Rules Extraction for Head and Neck Cancer Classification: Springer Nature; 2023.
82. Murad NY, Hasan MH, Azam MH, Yousuf N, Khalique SA. Explaining deep learning decisions via fuzzy inference system on medical images. 2023 Presented at: 2023 IEEE 21st Student Conference on Research and Development (SCOReD); Dec 13-14, 2023. [doi: 10.1109/SCOReD60679.2023.10563415]
83. Peng T, Wang C, Hu T, Mao H, Di G, Zhang L. Interactive ultrasound prostate cancer segmentation using deep learning with principal curve-based fine-tuning. 2023 Presented at: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 5-8, 2023. [doi: 10.1109/BIBM58861.2023.10385872]
84. Zhang S, Liu Q. A lung nodule labelling and recognition method based on variational auto encoder. 2024 Presented at: 2024 IEEE International Conference on Medical Artificial Intelligence (MedAI); Nov 15-17, 2024; Chongqing, China p. 67-74. [doi: 10.1109/MedAI62885.2024.00016]
85. Wang D, Wang Y, Wang Y, Liu L, Li J, Huang Q. Fusion of human cognitive knowledge and machine inference for breast cancer detection. 2023 Presented at: 2023 International Conference on Advanced Robotics and Mechatronics (ICARM); Jul 8-10, 2023; Sanya, China p. 179-184. [doi: 10.1109/ICARM58088.2023.10218759]
86. Zhang B, Vakanski A, Xian M. BI-RADS-Net: an explainable multitask learning approach for cancer diagnosis in breast ultrasound images. IEEE Int Workshop Mach Learn Signal Process 2021 Oct;2021. [doi: 10.1109/mlsp52302.2021.9596314] [Medline: 35509454]
87. Zhang B, Vakanski A, Xian M. BI-RADS-NET-V2: a composite multi-task neural network for computer-aided diagnosis of breast cancer in ultrasound images with semantic and quantitative explanations. IEEE Access 2023;11:79480-79494. [doi: 10.1109/access.2023.3298569] [Medline: 37608804]
88. Koul A, Bawa RK, Kumar Y. Enhancing the detection of airway disease by applying deep learning and explainable artificial intelligence. Multimed Tools Appl 2024;83(31):76773-76805. [doi: 10.1007/s11042-024-18381-y]
89. Wang Z, Zhang Z, Traverso A, Dekker A, Qian L, Sun P. Assessing the role of GPT-4 in thyroid ultrasound diagnosis and treatment recommendations: enhancing interpretability with a chain of thought approach. Quant Imaging Med Surg 2024 Feb;14(2):1602-1615. [doi: 10.21037/qims-23-1180]
90. Yu Y, Ren W, He Z, et al. Machine learning radiomics of magnetic resonance imaging predicts recurrence-free survival after surgery and correlation of LncRNAs in patients with breast cancer: a multicenter cohort study. Breast Cancer Res 2023 Nov 1;25(1):132. [doi: 10.1186/s13058-023-01688-3] [Medline: 37915093]
91. Wang J, Dai G, Ren X, et al. Assessing the reproducibility, stability, and biological interpretability of multimodal computed tomography image features for prognosis in advanced non‐small cell lung cancer. iRADIOLOGY 2024 Feb;2(1):3-16 [FREE Full text] [doi: 10.1002/ird3.56]
92. Yang J, Barnett AJ, Donnelly J, et al. FPN-IAIA-BL: a multi-scale interpretable deep learning model for classification of mass margins in digital mammography. 2024 Presented at: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); Jun 17-18, 2024 URL: https://ieeexplore.ieee.org/document/10678415 [accessed 2026-05-01] [doi: 10.1109/CVPRW63382.2024.00506]

XSL•FO
RenderX

93. Cardoso M, Santiago C, Nascimento JC. Using counterfactual information for breast classification diagnosis. 2024 Presented at: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); Jun 17-18, 2024 URL: https://ieeexplore.ieee.org/document/10677955 [accessed 2026-05-01] [doi: 10.1109/CVPRW63382.2024.00505]
94. Zhou S, Islam UJ, Pfeiffer N, Banerjee I, Patel BK, Iquebal AS. SCGAN: Sparse CounterGAN for counterfactual explanations in breast cancer prediction. IEEE Trans Autom Sci Eng 2024 Jul;21(3):2264-2275. [doi: 10.1109/tase.2023.3333788] [Medline: 40734689]
95. Cui H, Zhao Y, Xiong S, et al. Diagnosing solid lesions in the pancreas with multimodal artificial intelligence: a randomized crossover trial. JAMA Netw Open 2024 Jul 1;7(7):e2422454. [doi: 10.1001/jamanetworkopen.2024.22454] [Medline: 39028670]
96. Zhang X, Zhang G, Qiu X, et al. Exploring non-invasive precision treatment in non-small cell lung cancer patients through deep learning radiomics across imaging features and molecular phenotypes. Biomark Res 2024 Jan 25;12(1):12. [doi: 10.1186/s40364-024-00561-5] [Medline: 38273398]
97. Xiang H, Xiao Y, Li F, et al. Development and validation of an interpretable model integrating multimodal information for improving ovarian cancer diagnosis. Nat Commun 2024 Mar 27;15(1):2681. [doi: 10.1038/s41467-024-46700-2] [Medline: 38538600]
98. Ahmed M, Bibi T, Khan RA, Nasir S. Enhancing breast cancer diagnosis in mammography: evaluation and integration of convolutional neural networks and explainable AI. 2024 Presented at: 2024 26th International Multitopic Conference (INMIC); Karachi, Pakistan p. 1-6. [doi: 10.1109/INMIC64792.2024.11004362]
99. Aly M, Ghallab A, Fathi IS. ViT-GRU: advanced brain tumor diagnosis framework: vision transformer and GRU integration for improved MRI analysis: a case study of Egypt. IEEE Access 2024:1-1. [doi: 10.1109/ACCESS.2024.3513235]
100. Yao S, Shen P, Dai F, et al. Thyroid cancer central lymph node metastasis risk stratification based on homogeneous positioning deep learning. Research (Wash D C) 2024;7:0432. [doi: 10.34133/research.0432] [Medline: 39165637]
101. Nowakowska S, Borkowski K, Ruppert C, et al. Explainable precision medicine in breast MRI: a combined radiomics and deep learning approach for the classification of contrast agent uptake. Bioengineering (Basel) 2024 May 31;11(6):556. [doi: 10.3390/bioengineering11060556] [Medline: 38927793]
102. Saeed N, Ridzuan M, Alasmawi H, Sobirov I, Yaqub M. MGMT promoter methylation status prediction using MRI scans? An extensive experimental evaluation of deep learning models. Med Image Anal 2023 Dec;90:102989. [doi: 10.1016/j.media.2023.102989] [Medline: 37827111]
103. Sluckin TC, Hekhuis M, Kol SQ, et al. A deep learning framework with explainability for the prediction of lateral locoregional recurrences in rectal cancer patients with suspicious lateral lymph nodes. Diagnostics (Basel) 2023 Sep 29;13(19):3099. [doi: 10.3390/diagnostics13193099] [Medline: 37835842]
104. Singh Y, Hathaway QA, Keishing V, et al. Beyond post hoc explanations: a comprehensive framework for accountable AI in medical imaging through transparency, interpretability, and explainability. Bioengineering (Basel) 2025 Aug 15;12(8):879. [doi: 10.3390/bioengineering12080879] [Medline: 40868392]
105. Tang D, Chen J, Ren L, Wang X, Li D, Zhang H. Reviewing CAM-based deep explainable methods in healthcare. Appl Sci (Basel) 2024 May 13;14(10):4124. [doi: 10.3390/app14104124]
106. Houssein EH, Gamal AM, Younis EMG, Mohamed E. Explainable artificial intelligence for medical imaging systems using deep learning: a comprehensive review. Cluster Comput 2025 Sep;28(7). [doi: 10.1007/s10586-025-05281-5]
107. Huff DT, Weisman AJ, Jeraj R. Interpretation and visualization techniques for deep learning models in medical imaging. Phys Med Biol 2021 Feb 2;66(4):04TR01. [doi: 10.1088/1361-6560/abcd17] [Medline: 33227719]
108. Pesapane F, Sardanelli F. Keeping AI in medicine and radiology within the framework of scientific method: measuring to close the epistemic gap. Insights Imaging 2025 Dec 22;16(1):287. [doi: 10.1186/s13244-025-02171-7] [Medline: 41427974]
109. Saw SN, Yan YY, Ng KH. Current status and future directions of explainable artificial intelligence in medical imaging. Eur J Radiol 2025 Feb;183:111884. [doi: 10.1016/j.ejrad.2024.111884] [Medline: 39667118]
110. Chen H, Gomez C, Huang CM, Unberath M. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. NPJ Digit Med 2022 Oct 19;5(1):156. [doi: 10.1038/s41746-022-00699-2] [Medline: 36261476]
111. Räz T, Pahud De Mortanges A, Reyes M. Explainable AI in medicine: challenges of integrating XAI into the future clinical routine. Front Radiol 2025;5:1627169. [doi: 10.3389/fradi.2025.1627169] [Medline: 40896521]
112. An C, Park YW, Ahn SS, Han K, Kim H, Lee SK. Radiomics machine learning study with a small sample size: single random training-test set split may lead to unreliable results. PLoS ONE 2021;16(8):e0256152. [doi: 10.1371/journal.pone.0256152] [Medline: 34383858]
113. Vrudhula A, Kwan AC, Ouyang D, Cheng S. Machine learning and bias in medical imaging: opportunities and challenges. Circ Cardiovasc Imaging 2024 Feb;17(2):e015495. [doi: 10.1161/CIRCIMAGING.123.015495] [Medline: 38377237]
114. Di Martino F, Delmastro F. Explainable AI for clinical and remote health applications: a survey on tabular and time series data. Artif Intell Rev 2023;56(6):5261-5315. [doi: 10.1007/s10462-022-10304-3] [Medline: 36320613]
115. Gambetti A, Han Q, Shen H, Soares C. A survey on human-centered evaluation of explainable AI methods in clinical decision support systems. arXiv. Preprint posted online on Nov 11, 2025 URL: http://arxiv.org/abs/2502.09849 [accessed 2026-05-01]

XSL•FO
RenderX

116. Rosenbacke R, Melhus Å, McKee M, Stuckler D. How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: systematic review. JMIR AI 2024 Oct 30;3:e53207. [doi: 10.2196/53207] [Medline: 39476365]
117. Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. Radiol Artif Intell 2022 May;4(3):e210064. [doi: 10.1148/ryai.210064] [Medline: 35652114]
118. Lekadir K, Frangi AF, Porras AR, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ 2025 Feb 5;388:e081554. [doi: 10.1136/bmj-2024-081554] [Medline: 39909534]
119. Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020 Mar;2(2):e200029. [doi: 10.1148/ryai.2020200029] [Medline: 33937821]
120. Vasey B, Novak A, Ather S, Ibrahim M, McCulloch P. DECIDE-AI: a new reporting guideline and its relevance to artificial intelligence studies in radiology. Clin Radiol 2023 Feb;78(2):130-136. [doi: 10.1016/j.crad.2022.09.131] [Medline: 36639172]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
BI-RADS: Breast Imaging Reporting and Data System
CNN: Convolutional Neural Network
CNS: central nervous system
CT: computed tomography
DL: deep learning
DSS: decision support system
DT: decision tree
Grad-CAM: gradient-weighted class activation mapping
LIME: local interpretable model-agnostic explanations
MeSH: Medical Subject Headings
ML: machine learning
MRI: magnetic resonance imaging
PET: positron emission tomography
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Search Extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RF: random forest
SHAP: Shapley Additive Explanations
SVM: support vector machine
TCGA: The Cancer Genome Atlas
TCIA: The Cancer Imaging Archive
xAI: explainable artificial intelligence
XGBoost: extreme gradient boosting

Edited by S Brini; submitted 17.Jul.2025; peer-reviewed by A Billis, W Li; revised version received 04.Feb.2026; accepted 13.Feb.2026; published 20.May.2026.

Please cite as:
Fotopoulos D, Ladakis I, Filos D, Moreno-Sánchez PA, van Gils M, Chouvarda I
Explainable AI in Cancer Imaging: Scoping Review of Methods, Modalities, and Clinical Integration
J Med Internet Res 2026;28:e80645
URL: https://www.jmir.org/2026/1/e80645
doi:10.2196/80645

© Dimitris Fotopoulos, Ioannis Ladakis, Dimitrios Filos, Pedro A Moreno-Sánchez, Mark van Gils, Ioanna Chouvarda. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Receipt of Medicines Information From the Internet and Other Information Sources Among Adult Medicine Users in Developed Economies, 2010-2025: Systematic Review

Niina Mononen, PhD; Milla-Maria Salo, MSc; Daria Timonen, Bsc; Marika Pohjanoksa-Mäntylä, PhD; Marja Airaksinen, PhD

Clinical Pharmacy Group, Division of Pharmacology and Pharmacotherapy, Faculty of Pharmacy, University of Helsinki, Viikinkaari 5, Helsinki, Finland

Corresponding Author:
Niina Mononen, PhD
Clinical Pharmacy Group, Division of Pharmacology and Pharmacotherapy, Faculty of Pharmacy, University of Helsinki, Viikinkaari 5, Helsinki, Finland

Abstract

Background: The internet, social media, and digital health tools have transformed access and receipt of medicines information (MI), complementing or replacing traditional sources, such as physicians, pharmacists, and package leaflets.

Objective: This study aimed to examine the prevalence and patterns of adult medicine users' receipt of MI from electronic sources (eg, internet and social media) compared with traditional sources, and to assess trends in the receipt of MI in developed economies since 2010.

Methods: A systematic search was conducted in CINAHL, the Cochrane Library, ProQuest, Scopus (including Embase and MEDLINE), and Web of Science databases for peer-reviewed studies conducted and published between January 1, 2010, and December 31, 2025. Studies including adults (≥18 years) in developed economies were eligible. The review followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines. Methodological quality and risk of bias were assessed using the Mixed Methods Appraisal Tool. Due to substantial heterogeneity, findings were synthesized narratively without meta-analysis.

Results: Twenty-six studies from 10 countries, involving 19,891 medicine users, were included. Most studies were cross-sectional surveys (n=19), with only one being a national long-term trend study. Physicians, pharmacists, and package leaflets consistently remained the most common MI sources, irrespective of patient group, country, or study design. No clear temporal trends in MI receipt were observed. Variations in the frequency of MI receipt were evident across research methods, medicine user groups, and types of medicines. Individuals with greater familiarity with the internet were more likely to receive MI from electronic sources, such as websites and social media. In most studies, MI obtained from social media was not distinguished from other internet-based sources. Half of the studies focused on heterogeneous medicine user groups (n=14), while the remainder examined specific medicine user groups (n=12). Although the methodological quality of the studies was generally acceptable, only 3 studies reported the use of an explicit theoretical framework.

Conclusions: Traditional MI sources remain central for adult medicine users despite the growing role of electronic platforms. While the receipt of MI from electronic sources appears more common among internet-experienced users, no significant temporal trends in using these sources were identified. Further research is needed to better distinguish between different digital MI sources, including artificial intelligence–based MI sources, and to explore their evolving roles, implications for health-related decision-making, and associations with user characteristics and health contexts in the digital era. This review is innovative in systematically comparing traditional and electronic MI sources across developed economies and in highlighting methodological and conceptual gaps in existing research. By synthesizing cross-national evidence and identifying the need for clearer differentiation of digital MI subtypes, the study contributes to the field and informs the development of more targeted, reliable, and user-centered MI strategies in real-world health care settings.

(J Med Internet Res 2026;28:e71984) doi:10.2196/71984

KEYWORDS

medicines information; medicines information source; receipt of medicines information; internet; medicine user; systematic review

XSL•FO
RenderX

Introduction

Medicines information (MI) has become increasingly accessible to consumers and medicine users through multiple modalities, reflecting decades of regulatory, technological, and professional developments [1-3]. Major drug safety crises, particularly the thalidomide catastrophe in the early 1960s, demonstrated the public health consequences of limited availability of MI, leading to major regulatory reforms [1,2,4-14]. The reforms included the introduction of statutory written medicines information (WMI) to consumers, such as standardized package leaflets (PLs) in the European Union (EU) [2,8,10-12]. The thalidomide crisis, followed by other drug safety crises, was a driving force for risk management measures, including a global shift toward empowerment and patient self-management [3,15,16]. This positioned physicians, pharmacists, and other health care professionals (HCPs) on the frontline as MI providers [2,16,17]

Since the 1990s, the digital transformation of society has profoundly reshaped access to health information and MI [2,18-20]. Online sources now range from official regulatory databases and health care provider websites to commercial platforms, patient communities, social media, and mobile health apps [21-23]. Over the past decade, and particularly following the COVID-19 pandemic, digital health services have expanded rapidly, including telehealth consultations, remote medication counseling, chat-based pharmacy services, and electronically provided PLs [16]. More recently, artificial intelligence (AI)-based tools and generative language models have emerged as novel channels for health- and medicines-related information seeking. These developments have enabled more immediate, interactive, and personalized forms of receiving MI, potentially supporting patient empowerment in self-management of chronic conditions and mitigating minor symptoms [24]. At the same time, concerns have been raised regarding information quality, misinformation, digital health literacy, and unequal access across population groups [25].

A growing body of research has examined online health information-seeking behaviors and digital health interventions [18,19,26]. Fewer studies have focused specifically on MI as a distinct domain. Existing research mainly addresses single information channels in isolation, such as internet use or social media, without systematically comparing electronic sources with traditional sources, including physicians, pharmacists, and PLs. Furthermore, it remains unclear whether the receipt of internet-based MI has increased over time relative to traditional modalities. Therefore, a comprehensive synthesis examining patterns of MI receipt across electronic and traditional sources is warranted. Understanding how adult medicine users receive MI is of considerable public health and clinical importance. In the context of rapid digital transformation, synthesizing evidence on the relative use of electronic and traditional MI sources is essential to inform policy, health care practice, and future research.

This systematic review aimed to examine the prevalence and patterns of the receipt of MI from electronic and traditional sources among adult medicine users in developed economies since 2010. The specific objectives were (1) to compare the frequency of receipt of MI from electronic sources (eg, internet and social media) and traditional sources (eg, physicians, pharmacists, and PLs) and (2) to assess whether the receipt of internet-based MI has increased over time. In addition, the review sought to identify methodological and conceptual gaps in the existing literature to guide future research and policy development in the rapidly evolving digital health context.

Methods

Design

This study was conducted as a systematic review in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines [27]. The PRISMA 2020 flow diagram is presented in the Results section, and the completed PRISMA 2020 and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search extension) checklists are provided in Checklists 1-3 [27,28]. The Covidence systematic review software (Veritas Health Innovation) was used to manage study selection, screening, and data extraction throughout the review process [29].

Eligibility Criteria

Studies were eligible for inclusion if they met predefined criteria regarding the study population, interventions, outcomes, context, study design, and time, as summarized in Table 1. Eligible studies included adults aged 18 years or older who were current medicine users and examined the receipt of MI from any information source. Studies were required to be conducted in developed economies according to the United Nations classification [30], published in peer-reviewed international scientific journals in English, and published between January 1, 2010, and December 31, 2025. Both quantitative and qualitative original research studies, as well as systematic reviews and meta-analyses, were considered for inclusion. Conference abstracts, editorials, narrative reviews, doctoral and master's theses, and gray literature were excluded.

The choice of 2010 as the cut-off year was informed by the transition of MI practices from predominantly paper-based formats to electronic formats, which became more pronounced during the 2010s [2]. In addition, the need for improved EU-level coordination of MI provision to consumers was identified in 2005, followed by the issuance of recommendations on quality and accessibility by the European Commission in 2008 [31,32]. Consequently, MI practices began to be addressed more systemically in several member states; for example, a national MI strategy was introduced in Finland in 2012 [1,2]. Furthermore, earlier studies indicated that the literature published before 2010 on medicine users' receipt of MI was limited [2].

XSL•FO
RenderX

Table . Inclusion and exclusion criteria for the literature search based on the PICOT (Population, Intervention, Context, Outcomes, and Time) framework.

Component	Inclusion criteria	Exclusion criteria
Population (P)	• Participants aged ≥18 years • Current medicine users • Outpatients, inpatients, or hospital-discharged patients • Native population • Laymen	• Participants aged <18 years • People not currently using medicines • People using illegal substances or people with substance use (including doping use). • People using only herbal medicines, complementary and alternative medicines (CAM), or dietary supplements. • Immigrant population • Health care professionals or health care students
Intervention (I)	• MI[a] sources actually used.	The following topics were excluded: • Receipt or seeking of information about health, health care services, travel health advice, food, and nutrition information. • Receipt or seeking information about someone else's medicines (eg, children, spouse, and friend). • Usability of online MI or health apps. • Receipt of MI has not been clearly separated from health information. • Research results of medicine users and nonusers have been processed together and not separated from each other. • Receipt of MI among individuals who are illiterate or have low or limited literacy skills. • Research focused on the usage or usability of only one MI source (eg, certain web pages, mobile phone apps, medical literature platforms, and so on). • Use of national electronic health record systems (eg, My Kanta).
Context (C)	• Study was conducted in developed economies[b]. • All types of research methods. • Peer-reviewed full paper. • Published in English.	• Study was conducted in developing economies[b]. • Conference abstracts, editorials, narrative reviews, doctoral and master's theses, and gray literature.
Outcomes (O)	• Receipt of MI among the study population from any information sources (eg, physicians, pharmacists, PLs[c], the internet, and so on).	—[e]
Time (T)	• Study was conducted and published between January 1, 2010 and December 31, 2025[d].	• Study was conducted and published before 2010 or after December 31, 2025.

[a]MI: medicines information.

[b]Country selection was based on the classification by the United Nations of developed economies (ie, Australia, Austria, Belgium, Bulgaria, Canada, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Japan, Latvia, Lithuania, Luxembourg, Malta, the Netherlands, New Zealand, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, Switzerland, the United Kingdom, and the United States [30]).

[c]PL: package leaflet.

[d]Papers that did not mention the year of the study or data collection were included if the study was published in 2010 or later.

[e]Not available.

Information Sources

A comprehensive literature search was conducted using 5 international bibliographic databases: CINAHL, the Cochrane Library, ProQuest, Scopus (including Embase and MEDLINE), and Web of Science. In addition to database searches, the reference lists of all included studies were screened manually to identify any additional eligible studies. The most recent search was conducted on January 15, 2026, to ensure inclusion of the most up-to-date evidence. Full search strings and database filters for each database are provided in Multimedia Appendix 1. All

XSL•FO
RenderX

searches were rerun before final analysis to ensure inclusion of the most recent studies.

Search Strategy

A systematic literature search was conducted using predefined inclusion and exclusion criteria, Boolean operators, and database-specific filters (Table 1 and Multimedia Appendix 1). One researcher (NM) performed the database searches in collaboration with an information specialist from the University of Helsinki, and the search strategy was reviewed by the research team before execution. The following five international bibliographic databases were searched: CINAHL, the Cochrane Library, ProQuest, Scopus (including Embase and MEDLINE), and Web of Science. The search strategy incorporated the following key concepts and terms: "medicines information" or "drug information"; "receipt" or "receive"; "search" or "seek"; "medicines information source or channel" or "drug information source or channel"; "consumer," "interviewee," "medicine user," "respondent," or "patient"; and "women" or "men." Full search strings and database-specific filters are provided in Multimedia Appendix 1.

Iterative pilot searches were conducted to assess the sensitivity and relevance of the search strategy. These pilots indicated that some relevant studies referred to participants exclusively as "women" or "men," necessitating the inclusion of these terms to ensure comprehensive identification of eligible studies. The pilot searches further demonstrated the need to refine and narrow the eligibility criteria applied to focus the review specifically on the receipt of MI among adult medicine users, which led to the exclusion of a substantial number of studies that did not meet this criterion.

In this review, "receipt of MI" was defined as the active acquisition and obtaining of information regarding medicines, whereas "seeking MI" refers solely to the act of searching for information and does not necessarily result in obtaining it.

Selection Process

Study selection was managed using Covidence systematic review software [29], which enabled continuous monitoring of selection consistency throughout the process. All identified records were imported into Covidence. After duplicate removal, at least two researchers (NM, DT, MMS, and MA) independently screened titles and abstracts. Full texts of potentially eligible studies were independently assessed by two reviewers. Disagreements were resolved through discussion and consensus within the research team. Final inclusion decisions were made collectively by the study group, and consensus was achieved for all included studies. According to the Covidence documentation, agreement at the final stage of study selection was 100%. In addition, reference lists of all included papers were screened to identify further potentially relevant studies. All additional records identified through reference list screening that met the eligibility criteria were subsequently included in the review.

Data Collection Process

Data extraction was conducted using a standardized extraction table developed specifically for this review. Two researchers independently extracted data, which were then reviewed and verified by the remaining members of the research team. Any discrepancies were resolved through discussion. Data management and preliminary organization were conducted using Microsoft Word and Microsoft Excel (Windows 10).

Study Risk of Bias Assessment

The methodological quality of the included studies was assessed using the Mixed Methods Appraisal Tool (MMAT; version 2018) (Multimedia Appendix 2) [33,34]. The MMAT was selected for its capacity to evaluate a wide range of study designs, including quantitative, qualitative, and mixed methods studies. The tool comprises design-specific criteria to evaluate methodological rigor and internal validity within each study category. For example, in quantitative descriptive studies, it explicitly assesses the risk of nonresponse bias [34,35].

At least 2 reviewers (NM, MMS, and MA) independently appraised each study using the MMAT checklist. Discrepancies were resolved through discussion, and a third reviewer was consulted when necessary. Each criterion was rated as "Yes" (1 point), "No" (0 points), or "Can't tell" (0 points), and the sum of the scores indicates the total quality of each study. Screening questions (S1 and S2) were excluded from the total scores. For mixed methods studies, the overall quality score was determined by the lowest score of the individual component, as the overall quality of a combination cannot exceed the quality of its weakest component [33,35]. Risk of bias assessments informed the interpretation of findings but were not used as exclusion criteria. Final quality ratings were determined by consensus among the reviewers.

Data Extraction

Data were extracted using a standardized extraction table developed specifically to address the objectives of this systematic review. Studies were content analyzed and categorized, with quantitative and qualitative studies analyzed and reported separately. Two researchers independently performed data extraction, and the results were subsequently reviewed and approved by the remaining members of the research team. Discrepancies were resolved through discussion to ensure consistency and accuracy. Data management and organization were conducted using Microsoft Word and Microsoft Excel (Windows 10), with the analysis proceeding in sequential stages.

Extracted study characteristics included authorship, year of publication, country, aim of the study, study design, study setting, study population, medicine user group, MI sources investigated, and key findings related to the receipt of MI. Where reported, additional methodological and contextual information, such as recruitment and sampling methods, definitions and measurement of the receipt of MI, outcome measures, data collection period, analytical approaches, theoretical frameworks, and quality and risk-of-bias assessments, were also collected to support transparent reporting and facilitate interpretation of heterogeneity across studies.

Tables 2 and 3 summarize the core characteristics of the included studies, while Multimedia Appendix 3 provides the full data extraction framework, developed in accordance with

XSL•FO
RenderX

the Cochrane Handbook and PRISMA 2020 recommendations [27,36]. This detailed extraction table promotes transparency, enables systematic comparison across heterogeneous study designs, and facilitates editorial and peer review, although not all extracted items were directly used in the synthesis.

XSL•FO
RenderX

Table . Summary of basic characteristics of the included studies on the receipt of medicines information (MI) among heterogeneous medicine user groups (n=14).

Authors, years, and country	Aim of the study	Setting	Design	Population	MMAT[a] score
Surveys, national (n=4)					
Mononen et al [37], 2019, Finland	To explore long-term trends in the receipt of MI[b] in the general population.	National	Cross-sectional repeated postal survey	Medicine users (n=8822), outpatients, and females (63% - 64%).	5
Hämeen-Anttila et al [38], 2018, Finland	To explore the receipt of MI from the internet among frequent internet users.	National	Cross-sectional online survey	Prescription and OTC[c] medicine users (n=2489), outpatients, and females (85%).	4
O'Donovan et al [39], 2018, United Kingdom	To explore the receipt of MI on ADRs[d] in the general population.	National	Cross-sectional on-site survey	Prescription medicine users (n=230), outpatients, and females (61%).	3
Santos et al [40], 2022, Switzerland	To explore the perceptions of conflicting information on chronic medications in the general population.	National	Cross-sectional online or on-site survey	Prescription medicine users (n=405), inpatients and outpatients, and females (57%).	3
Surveys, local (n=6)					
Krska and Morecroft [41], 2013, United Kingdom	To explore the experiences of receiving MI among hospital patients.	Hospitals (n=6) in North West England	Cross-sectional on-site survey	Hospital patients using prescription medicines (n=1218), inpatients, and females (51%).	4
Krska and Morecroft [42], 2013, United Kingdom	To assess the receipt of side-effect information in relation to suspected ADRs among hospital patients.	Hospitals (n=6) in North West England	Cross-sectional on-site survey	Hospital patients using prescription medicines (n=1218), inpatients, and females (51%).	4
Cooper and Garrett [43], 2014, Australia	To explore the experiences and preferences of receiving MI among hospital patients.	Hospitals (n=2) in New South Wales	Cross-sectional on-site survey	Hospital patients using prescription medicines (n=292), hospital-discharged patients, and females (61%).	3
DeLorme et al [44], 2011, United States	To explore the receipt of MI on prescription medicines and source selection among adult medicine users.	Southeastern metropolitan area	Cross-sectional computer-assisted telephone survey	Prescription medicine users (n=234), outpatients, and females (71%).	2
Perry et al [45], 2020, Australia	To explore the receipt of MI and supply behaviors among elite and developing athletes.	State-based sporting institute in Perth	Cross-sectional online survey	Athletes using prescription or nonprescription medicines (n=90) and outpatients.	2
Bergmo et al [46], 2025, Norway	To explore the receipt of MI from the internet among pharmacy customers.	Community pharmacies (n=11) in Tromsø and social media	Cross-sectional online and paper-based survey	Prescription and OTC medicines users (n=303), outpatients, and females (65%).	1
Qualitative interviews (n=2)					
Haverhals et al [47], 2011, United States	To explore medication management processes among adults with multimorbidity.	Metropolitan areas in Denver and Boulder, Colorado	Semistructured individual and group interviews	Prescription medicine users (n=32), outpatients, and females (10%).	5

XSL•FO
RenderX

Authors, years, and country	Aim of the study	Setting	Design	Population	MMAT[a] score
Tong et al [48], 2018, Australia and the United Kingdom	To explore the receipt of spoken MI and WMI[e] among adult medicine users.	Australian and United Kingdom consumers	Semistructured individual interviews	OTC medicine users (n=37 in Australia and n=39 in the United Kingdom), outpatients, and females (50%).	5
Mixed methods (n=2)					
Mackridge et al [49], 2018, United Kingdom	To assess the receipt of information about medication changes during admission and post-discharge support among hospital patients.	Hospitals (n=6) in North West England	Study 1: cross-sectional face-to-face survey and study 2: telephone follow-up survey.	Hospital patients using prescription medicines (study 1: n=444 and study 2: n=99), hospital-discharged patients, and females (53% in study 1 and 45% in study 2).	4
Bastholm-Rahmner et al [50], 2018, Sweden	To assess the receipt of MI and knowledge of the list of national recommended essential medicines (Wise List) among hospital patients.	Primary health care centers (n=4) in Stockholm	Study 1: on-site survey (included) and study 2: focus group discussions (not included).	Medicine users (study 1: n=312), outpatients, and females (59%).	3

[a]MMAT: Mixed Methods Appraisal Tool.

[b]MI: medicines information.

[c]OTC: over-the-counter.

[d]ADR: adverse drug reaction.

[e]WMI: written medicines information.

XSL•FO
RenderX

Table . Summary of basic characteristics of the included studies on the receipt of medicines information (MI) among adults using specific medicines (n=12).

Authors, year, and country	Aim of the study	Setting	Design	Population	MMAT[a] score
Surveys, national (n=3)					
Amundsen et al [51], 2016, Norway	To explore the use of antimigraine medications and information needs during pregnancy and breastfeeding.	National	Cross-sectional online survey	Pregnant and breastfeeding women using antimigraine medicines (n=401), outpatients, and females (100%)	5
Leonardo et al [52], 2020, Australia	To explore the receipt of MI[b] on methotrexate treatment among patients with rheumatoid arthritis.	National	Cross-sectional online survey	Patients with rheumatoid arthritis using or having used methotrexate (n=742), outpatients, and females (76%)	5
Otón et al [53], 2023, Spain	To assess the information needs and expectations in clinical care among methotrexate users.	National	Cross-sectional online survey	Patients with immune-mediated disease using methotrexate (n=283), outpatients, and females (82%)	2
Surveys, local (n=6)					
Bults et al [54], 2012, The Netherlands	To explore the receipt of MI on Q fever vaccination among patients with cardiovascular conditions.	Municipal public health service	Questionnaire-based individual interviews	Cardiovascular medicine users (n=413), outpatients, and females (39%)	5
Houser et al [55], 2016, United States	To assess sociodemographic differences in the receipt of prescription NSAID[c] risk information among adult medicine users.	Physician practices (n=39) in Alabama	Cross-sectional computer-assisted telephone survey	Prescription NSAID or OTC[d] medicine (ibuprofen or naproxen) users (n=220), outpatients, and females (75%)	5
de Toro et al [56], 2019, Spain	To assess the receipt of MI about biological medicines among patients with arthritis.	Hospitals (n=50)	Cross-sectional on-site survey	Arthritis medicine (s.c. biological) users (n=592), outpatients, and females (58%)	4
Geryk et al [57], 2016, United States	To explore the receipt of MI among patients with arthritis.	Participants were recruited in several ways	Cross-sectional online survey	Arthritis medicine users (n=328), outpatients, and females (79%)	4
Carpenter et al [58], 2011, United States	To explore the receipt of MI and the perceived credibility of MI sources among patients with vasculitis.	Participants were recruited in several ways	Cross-sectional online survey	Vasculitis medicine users (n=232), outpatients, and females (69%)	3
Carpenter et al [59], 2014, United States	To explore the sources of conflicting MI among patients with arthritis.	Participants were recruited in several ways	Cross-sectional online survey	Arthritis medicine users (n=328), outpatients, and females (79%)	3
Qualitative interviews (n=3)					
Bergsholm et al [60], 2023, Norway	To explore the receipt of MI among antibiotic users.	Central urban areas in the Norwegian South-Eastern health region	Focus group interviews (n=2)	Antibiotic medicine users (n=8), outpatients, and gender not reported	5
Hayden et al [61], 2015, United Kingdom	To explore the receipt of MI and adherence-related beliefs about methotrexate among patients with arthritis.	University Hospitals of Leicester NHS[e] Trust	Semistructured individual interviews	Patients with arthritis using oral methotrexate (n=15), outpatients, and females (73%)	5

XSL•FO
RenderX

Authors, year, and country	Aim of the study	Setting	Design	Population	MMAT[a] score
Wakob et al [62], 2022, Germany	To assess the benefits and risks of medication and the receipt of MI among patients with cardiovascular conditions.	University Hospital (n=1) offering tertiary care	Semistructured individual interviews	Cardiovascular medicine users (n=102), inpatients, and females (35%)	5

[a]MMAT: Mixed Methods Appraisal Tool.

[b]MI: medicines information.

[c]NSAID: nonsteroidal anti-inflammatory drug.

[d]OTC: over-the-counter

[e]NHS: National Health Service.

Data Synthesis and Analysis

Before selecting the analytical approach, the feasibility of conducting a meta-analysis was carefully evaluated in accordance with established methodological guidance on evidence synthesis [63]. Quantitative pooling was deemed inappropriate because the included studies were not sufficiently comparable to support a meaningful pooled estimate. The studies differed substantially in their target populations, settings (eg, hospitals, outpatient care, and community pharmacies), types of medicines, and in how the receipt of MI and MI sources were defined and operationalized. Outcome measures also varied considerably across studies.

Several methodological limitations further reduced confidence in the robustness and interpretability of any potential summary effect. These included low or unreported response rates, the use of convenience samples or online surveys with uncertain denominators, and variable study quality across designs. Pooling effect estimates from such studies would risk producing a summary effect with questionable validity and misleading precision. Moreover, although meta-analysis can be used to explore heterogeneity, the limited number of studies within meaningful subgroups, combined with diverse outcome measures and variable study quality, would make any statistical exploration of heterogeneity unreliable and potentially misleading. Taken together, these factors indicated that quantitative synthesis would not yield a valid or interpretable summary estimate. No sensitivity analyses were conducted due to the absence of quantitative pooling.

Consequently, a narrative synthesis was conducted as the primary analytic approach, allowing transparent interpretation of patterns and trends in the evidence without implying a summary effect that could not be meaningfully estimated. Given the heterogeneity of study designs, outcome measures, and the operationalization of MI-related constructs, the synthesis followed a structured qualitative analytic process. In this process, the MI sources reported in each study were examined, definitions and distinctions of internet-based MI sources were reviewed, and any theoretical frameworks were identified.

In qualitative studies, formal coding procedures were not used. Instead, key findings relevant to the research objectives were systematically identified and summarized. This approach enabled qualitative evidence to be integrated into the narrative synthesis in a manner comparable to the treatment of quantitative findings, while maintaining transparency regarding how qualitative insights informed the overall conclusions.

To enhance analytical clarity, the studies were further classified into two main categories: (1) studies examining the receipt of MI among heterogeneous adult medicine user groups and (2) studies examining the receipt of MI among adults using specific medicines. Results were reported within these categories. Finally, a summary table and a visual representation of the key findings were developed to provide a comprehensive overview of the available evidence.

Ethical Considerations

No ethics board review was required for this study, as it involved the analysis of publicly available, anonymized data only, in accordance with the Finnish National Board on Research Integrity (TENK, 2019).

Results

Characteristics of Included Studies

A total of 26 studies met the inclusion criteria (Figure 1) [37-62]. An overview of the core characteristics of the included studies is presented in Table 2, while the full data extraction, including detailed methodological and contextual information, is provided in Multimedia Appendix 3, representing 19,891 medicine users from 10 different countries. No systematic reviews or meta-analyses were identified among the included studies.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram for the literature screening covering the time period 2010 - 2025. MI: medicines information.

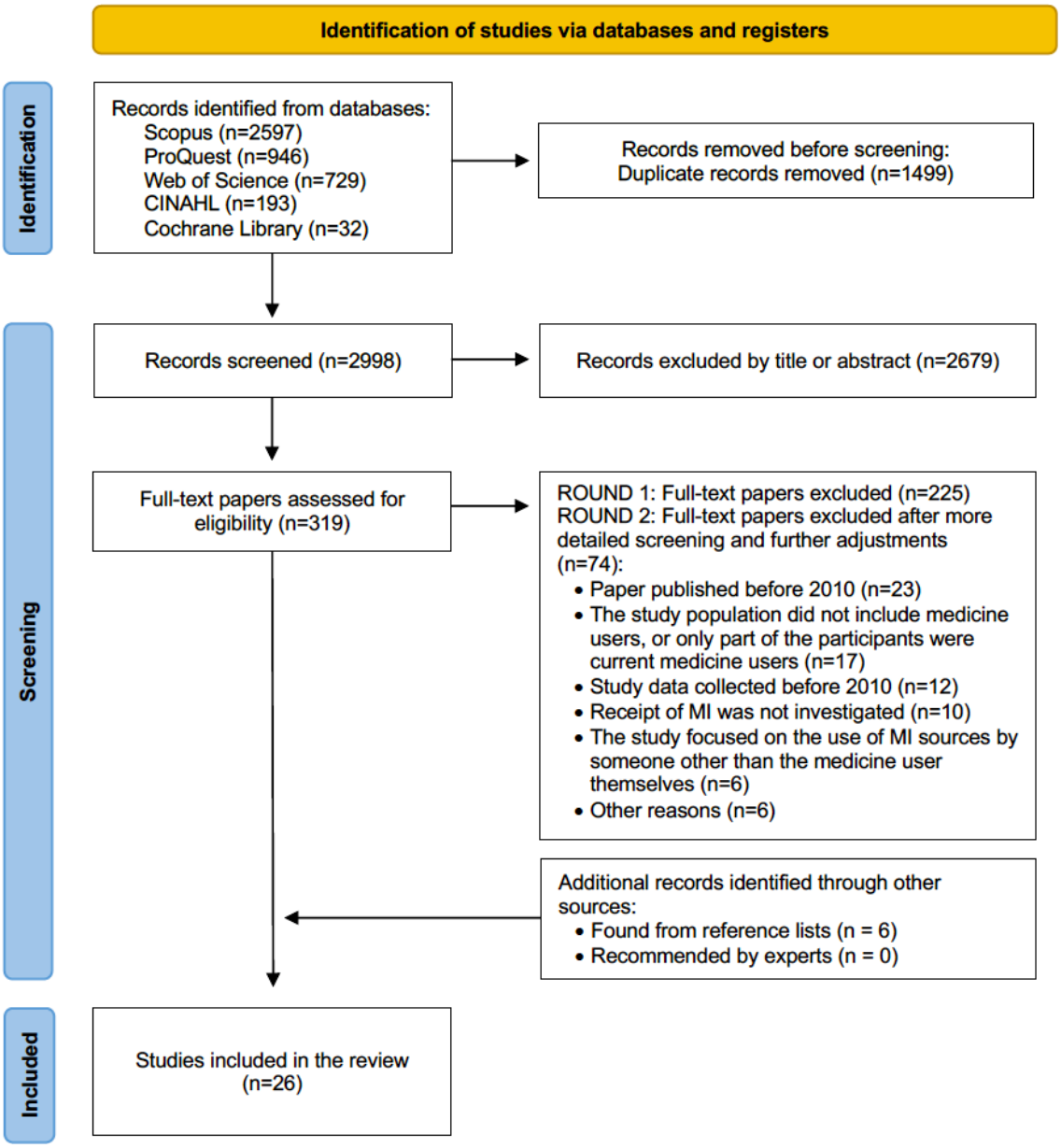

More than half of the studies (n=16, 62%) were carried out in European countries (Tables 2 and 4) [37-42,46,49-51,53,54,56,60-62]. Six (23%) studies were conducted in the United States [44,47,55,57-59] and 3 (12%) in Australia [43,45,52]. One study used a multinational design and was conducted in both Australia and the United Kingdom [48]. The predominant study designs were cross-sectional surveys (n=19, 73%), performed by mail, online, telephone, or questionnaire-based interviews [37-46,51,53-59,62] and qualitative interviews (n=5, 19%) [47,48,60-62]. Most studies were conducted among outpatients (n=20, 77%) [37-39,44-48,50-61], while 3 studies focused on inpatients [41,42,62]. Two studies examined patients at the time of hospital discharge [43,49], and 1 study included both inpatients and outpatients [40]. Approximately half of the studies (n=14, 54%) investigated heterogeneous medicine user groups [37-50], while the remaining studies (n=12, 46%) focused on specific medicine user groups [51-62]. Across the included studies, the most commonly examined MI sources were HCPs (eg, physicians, pharmacists, nurses, and midwives), the internet and social media, PLs, and family and friends (Table 2 and Multimedia Appendix 3).

XSL•FO
RenderX

Table . Characteristics of the included studies (n=26).

Category and subcategory	Frequency
Study country	
United States [44,47,55,57-59]	6
United Kingdom [39,41,42,49,61]	5
Australia [43,45,52]	3
Norway [46,51,60]	3
Finland [37,38]	2
Spain [53,56]	2
Australia and the United Kingdom [48]	1
Germany [62]	1
The Netherlands [54]	1
Sweden [50]	1
Switzerland [40]	1
Study design	
Survey, local, and cross-sectional [41-46,54-59]	12
Survey, national, and cross-sectional [37-40,51-53]	7
Qualitative interview [47,48,60-62]	5
Mixed methods [49,50]	2
Study population	
Heterogeneous medicine user groups	14
• Prescription medicine users [39,40,42,44,47,49]	6
• Not specified [37,38,41,43,50]	5
• Prescription and OTC[a] medicine users [45,46]	2
• OTC medicine users [48]	1
Specific medicine user groups	12
• Arthritis medicine users [55-57,59]	4
• Methotrexate users [52,53,61]	3
• Cardiovascular medicine users [54,62]	2
• Antibiotic medicine users [60]	1
• Antimigraine medicine users [51]	1
• Vasculitis medicine users [58]	1
Study context	
Outpatients [37-39,44-48,50-61]	20
Inpatients [41,42,62]	3
Hospital-discharged patients [43,49]	2
Inpatients and outpatients [40]	1
MI[b] sources studied	
Physician and specialist [37-41,44-47,49-62]	23
Pharmacist [37-41,43-49,51,52,55-60,62]	21

XSL•FO
RenderX

Category and subcategory	Frequency
Internet and social media [37,38,40,42,44-47,50-52,55,57-62]	18
Package leaflet (PL) [37-39,41,42,44,46-49,51,53,57-59,61,62]	17
Family[c], relatives, and friends [37,42,44,46,47,52-59,62]	14
Nurse and midwife [37-39,41,47,49,51,55-58]	12
Books, magazines, and newspapers [37,42,44,57-59]	6
Radio, television, and media [37,44,56,59,62]	5
Advertisements [37,46,57,59]	4
Health care professional (HCP), not specified [42,49,55,59]	4
Patient brochures and pamphlets [56-59]	4
Written medicines information (WMI), not specified [48,53,55,60]	4
Newsletters [57-59]	3
Support groups [57-59]	3
Medical books, drug handbooks, and health journals [44,47]	2
Podcasts [57,59]	2
WMI from HCPs [41,44]	2
Health food stores [37]	1
Hospitals [45]	1
Other patients [56]	1
Patient associations [56]	1
Pharmaceutical companies [44]	1
Telephone services and MI centers [37]	1

[a]OTC: over-the-counter.

[b]MI: medicines information.

[c]Including spouse and partner.

Methodological Quality

Among the included studies (n=26), 10 (38%) studies met all MMAT methodological quality criteria (5 points) (Table 5 and Multimedia Appendix 3) [37,47,48,51,52,54,55,60-62]. Six (23%) studies fulfilled 4 of the 5 criteria (4 points) [38,41,42,56-58], 6 (23%) studies met 3 criteria (3 points) [39,40,43,49,50,59], 3 (12%) studies met 2 criteria (2 points) [44,45,53], and 1 (4%) study met 1 criterion (1 point) [46] (Multimedia Appendix 1 and Table 4). All qualitative (n=5) studies fulfilled all MMAT criteria (5 points) [47,48,60-62]. The 2 mixed methods studies fulfilled 3-4 criteria (3 - 4 points) [49,50]. Among quantitative (n=19) studies, 5 met all criteria (5 points) [37,51,52,54,55], 5 fulfilled 4 criteria (4 points) [38,41,42,56,57], 5 fulfilled 3 criteria (3 points) [39,40,43,58,59,62], and 3 fulfilled 2 criteria (2 points) [44,45,53]. One quantitative study fulfilled only 1 criterion (1 point) [46].

The methodological characteristics that best met the quality criteria in quantitative studies (n=19) were the following: statistical analysis was appropriate to answer the research questions (18 studies met the criteria) [37-45,47,51-59]; the measurements were appropriate (18 studies) [37-46,51-55,57-59]; and a relevant sampling strategy to address the research question was applied (14 studies) (Table 5) [37,38,40-43,51,52,54-59]. The most common limitation in quantitative studies was that the risk of nonresponse bias was either rated as low without sufficient justification (9 studies) [41,42,45,49,50,53,57-59] or was not reported at all (6 studies) [38-40,43,44,46]. Only 3 studies incorporated theoretical frameworks: (1) Protection Motivation Theory and the Health Belief Model [54], (2) Comprehensive Model of Information Seeking [44], and (3) Necessity-Concerns Framework and Cognitive Dissonance Theory [61].

The heterogeneity observed across study designs, populations, outcomes, and analytical approaches precluded quantitative synthesis. Consequently, a qualitative descriptive analysis was conducted.

XSL•FO
RenderX

Table . Methodological quality assessment of the included studies (n=26) using the Mixed Methods Appraisal Tool (MMAT), organized according to MMAT quality score. Randomized controlled trials (methodological group 2) and nonrandomized studies (methodological group 3) are not presented, as none of the included studies used these study designs.

Study	Screening questions		Subquestions															MMAT[b] score[a]
			Qualitative studies					Quantitative descriptive studies					Mixed method studies					
	S1	S2	1.1	1.2	1.3	1.4	1.5	4.1	4.2	4.3	4.4	4.5	5.1	5.2	5.3	5.4	5.5	
Amundsen et al [51]	Yes	Yes						Yes	Yes	Yes	Yes	Yes						5
Bagdam et al [60]	Yes	Yes	Yes	Yes	Yes	Yes	Yes											5
Bults et al [54]	Yes	Yes						Yes	Yes	Yes	Yes	Yes						5
Haverhals et al [47]	Yes	Yes	Yes	Yes	Yes	Yes	Yes											5
Hayden et al [61]	Yes	Yes	Yes	Yes	Yes	Yes	Yes											5
Houser et al [55]	Yes	Yes						Yes	Yes	Yes	Yes	Yes						5
Leonardo et al [52]	Yes	Yes						Yes	Yes	Yes	Yes	Yes						5
Mononen et al [37]	Yes	Yes						Yes	Yes	Yes	Yes	Yes						5
Tong et al [48]	Yes	Yes	Yes	Yes	Yes	Yes	Yes											5
Wakobo et al [62]	Yes	Yes	Yes	Yes	Yes	Yes	Yes											5
de Toro et al [56]	Yes	Yes						Yes	Yes	No	Yes	Yes						4
Geryk et al [57]	Yes	Yes						Yes	Yes	Yes	Cannot tell	Yes						4
Himmen-Antti-la et al [38]	Yes	Yes						Yes	Yes	Yes	No	Yes						4
Krska and Morecroft [41]	Yes	Yes						Yes	Yes	Yes	Cannot tell	Yes						4

XSL•FO
RenderX

Study	Screening questions		Subquestions															MMAT[b] score[a]
			Qualitative studies					Quantitative descriptive studies					Mixed method studies					
	S1	S2	1.1	1.2	1.3	1.4	1.5	4.1	4.2	4.3	4.4	4.5	5.1	5.2	5.3	5.4	5.5	
Krska and Morecroft [42]	Yes	Yes						Yes	Yes	Yes	Cannot tell	Yes						4
Mackridge et al [49]	Yes	Yes						Yes	Yes	Yes	Cannot tell	Yes						4
Baldwin-Rahmner et al [50]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Cannot tell	Yes	Yes	Yes	Yes	Yes	No	3
Carpenter et al [58]	Yes	Yes						Yes	No	Yes	Cannot tell	Yes						3
Carpenter et al [59]	Yes	Yes						Yes	No	Yes	Cannot tell	Yes						3
Cooper and Garrett [43]	Yes	Yes						Yes	No	Yes	No	Yes						3
O'Donovan et al [39]	Yes	Yes						No	Yes	Yes	No	Yes						3
Santos et al [40]	Yes	Yes						Yes	No	Yes	No	Yes						3
DeLorme et al [44]	Yes	Yes						No	No	Yes	No	Yes						2
Otón et al [53]	Yes	Yes						No	No	Yes	Cannot tell	Yes						2
Perry et al [45]	Yes	Yes						No	No	Yes	Cannot tell	Yes						2
Bagmo et al [46]	Yes	Yes						No	No	Yes	No	No						1

[a]MMAT score: 5=100% compliance with all quality evaluation criteria (high confidence), 4=80% of quality evaluation criteria met (moderate confidence), 3=60% of quality evaluation criteria met (low confidence), 2=40% of quality evaluation criteria met (very low confidence), 1=20% of quality evaluation criteria met, and 0=0% of quality evaluation criteria met.

[b]MMAT: Mixed Methods Appraisal Tool.

XSL•FO
RenderX

Receipt of MI Among Heterogeneous Medicine User Groups

A total of 14 studies examined the receipt of MI among heterogeneous medicine user groups (Table 2 and Multimedia Appendix 3) [37-50]. Of these, 8 studies focused on users of unspecified prescription medicines [39-44,47,49], 3 on users of both prescription and over-the-counter (OTC) medicines [38,45,46], 2 on general medicine users [37,50], and 1 on OTC medicine users exclusively [48]. The studies were performed in the United Kingdom [39,41,42,49], Australia [43,45], Finland [37,38], Norway [46], Switzerland [40], the United States [44,47], and Sweden [50], with 1 multinational study carried out in Australia and the United Kingdom [48]. Most studies were conducted in the early 2010s, as 7 studies between 2010 and 2016 [37,38,41,42,48-50], 3 between 2017 and 2025 [40,45,46], while 4 studies did not report the year of data collection [39,43,44,47]. Cross-sectional surveys were the most frequently used study design (n=10) [37-46]. Qualitative interviews were used in 2 studies [44,49], and mixed methods were applied in 2 studies [49,50]. Most studies were conducted among outpatients (n=9) [37-39,44-48,50], while 2 studies focused on hospital inpatients [41,42], 2 on hospital-discharged patients [43,49], and 1 included both inpatients and outpatients [40].

Across the heterogeneous medicine user (n=14) studies, a consistent finding was that physicians, community pharmacists, and PLs were the most commonly reported MI sources, regardless of study design, study period, population, or country (Figure 2 and Multimedia Appendix 3) [37-50]. However, notable differences were observed between prescription and OTC medicine users. Physicians, pharmacists, and PLs were the primary information sources related to prescription medicines [39,40,42-47,49]. In contrast, pharmacists were the most common information sources for OTC medicines, particularly in situations involving symptom-based requests, first-time purchases, or when consumers actively sought advice from pharmacists [48]. Nevertheless, 38% of medicine users reported not receiving any MI from HCPs regarding their medicines [37].

The receipt of MI varied substantially by research method, patient group, and type of medicine use (Figure 2 and Multimedia Appendix 3). No clear temporal trends in the receipt of MI were identified. National and local online surveys conducted in Europe indicated that 63% - 68% of prescription and OTC medicine users received MI from the internet [38,45]. In contrast, substantially lower proportions (14% - 41%) were reported in studies using other research methods, such as paper-based surveys, computer-assisted surveys, and focus group discussions conducted in Europe and North America (ie, the United States) [37,39,42,44,46,50]. According to national online and on-site surveys from Europe, 69% - 72% of adult prescription and OTC medicine users received MI from general practitioners [38,39]. Lower proportions (21% - 51%) were reported in national postal and local computer-assisted surveys conducted in Europe and North America [37,44]. Pharmacists were reported as MI sources by 81% of hospital-discharged patients in Australia based on an on-site survey [43]. Among European outpatients, 71% - 83% received MI from pharmacists according to national and local online surveys [38,45], whereas lower proportions (27% - 49%) were reported in national surveys using other data collection methods [37,39,40]. In North America, a local computer-assisted survey indicated that 21% of outpatients received MI from pharmacists [44]. Although the PLs are among the most commonly reported MI sources, not all medicine users read them [37-39,41,42,44]. In a national online survey from Finland, 90% of prescription and OTC medicine users reported reading PLs [38]. In comparison, 12% - 67% of medicine users reported reading PLs in national and local surveys using nononline methods in Europe and North America [37,39,41,44]. Among hospital patients in the United Kingdom, 74% reported usually reading PLs, whereas 19% never read them, and 7% read them only when unexpected adverse drug reactions occurred [42].

Qualitative interviews from the United States indicated that older medicine users (≥65 years) most commonly received MI from pharmacists and physicians, followed by PLs and the internet (Figure 2 and Multimedia Appendix 3) [47]. Qualitative interviews from Australia and the United Kingdom further suggested that consumers often did not receive active spoken MI when purchasing OTC medicines [48]. OTC medicine users tended not to read labels or PLs when they were familiar with the product, and when WMI was read, it was commonly read at home rather than at the point of purchase. In the Australian data, many respondents reported not receiving a PL with their OTC medicines [48]. In contrast, respondents from the United Kingdom more frequently reported receiving PLs, with some noting that PLs were typically included in OTC medicine packages [47,48].

XSL•FO
RenderX

Figure 2. Summary of findings from the studies (n=14) on the receipt of medicines information (MI) via the internet and commonly used information sources among heterogeneous medicine user groups in 2010 - 2025. The figure contains only the results reported in the included studies. OTC: over-the-counter; PL: package leaflet; WMI: written medicines information.

Information source	Study year		Country			Patients				Medicine use		
	2010-2016	2017-2025	Europe	United States	Australia	Outpatients	Inpatients and outpatients	Hospital inpatients	Hospital-discharged patients	Prescription medicines	Prescription and OTC medicines	Not specified
Surveys, national (n=4)												
Internet	68% [39] 14%-17% [38]	38% [56]	68% [39] 38% [56] 14%-17% [38]			68% [39] 38% [56] 14%-17% [38]				38% [56]	68% [39]	14%-17% [38]
Physician (general)	72% [39] 47%-51% [38]	82% [61] 69% [56]	82% [61] 72% [39] 69% [56] 47%-51% [38]			72% [39] 69% [56] 47%-51% [38]	82% [61]			82% [61] 69% [56]	72% [39]	47%-51% [38]
Physician (specialist)		74% [61]	74% [61]				74% [61]			74% [61]		
Pharmacist	83% [39] 43%-47% [38]	49% [61] 27% [56]	83% [39] 49% [61] 43%-47% [38] 27% [56]			83% [39] 43%-47% [38] 27% [56]	49% [61]			49% [61] 27% [56]	83% [39]	43%-47% [38]
Package leaflet	90% [39] 33%-35% [38]	67% [56]	90% [39] 67% [56] 33%-35% [38]			90% [39] 67% [56] 33%-35% [38]				67% [56]	90% [39]	33%-35% [38]
Surveys, local (n=6)												
Internet	29% [53] 18% [46]	63% [57] 41% [62]	63% [57] 41% [62] 29% [53]	18% [46]		63% [57] 41% [62] 18% [46]		29% [53]		29% [53] 18% [46]	63% [57] 41% [62]	
Physician (general)	21% [46]	72% [62]	72% [62]	21% [46]		72% [62] 21% [46]				21% [46]	72% [62]	
Physician (specialist)												
Pharmacist	81% [45] 21% [46]	71%-74% [57] 53% [62]	71%-74% [57] 53% [62]	21% [46]	81% [45]	71%-74% [57] 53% [62] 21% [46]			81% [45]	81% [45] 21% [46]	71%-74% [57] 53% [62]	
Package leaflet	42%-74% [53] 47% [62] 12% [46] 12% [52]		42%-72% [53] 47% [62] 12% [52]	12% [46]		47% [62] 12% [46]		42%-74% [53] 12% [52]		42%-74% [53] 12% [46] 12% [52]	47% [62]	
Mixed methods (n=2)												
Internet	31% [41]		31% [41]			31% [41]						31% [41]
Physician (general)	36% [41]		36% [41]			36% [41]						36% [41]
Physician (specialist)												
Pharmacist	5% [55]		5% [55]						5% [55]	5% [55]		
Package leaflet												

Qualitative interviews (n=2)

- Pharmacists, physicians, PLs, and the internet were the most common MI sources among older (≥65 years) prescription medicine users in the United States [49].
- The majority of Australian and United Kingdom consumers reported that no spoken MI was actively provided when they purchased their OTC medicine. Instructions for the use of medicine were the most often received information. OTC medicine users tended not to read labels or PLs if they were familiar with the medicine. When WMI was read, instructions for using medicine were commonly read at home. The majority of Australian consumers reported that they did not receive a PL with their OTC medicine. In contrast, the overwhelming majority of United Kingdom consumers noted receiving a PL. Specifically, consumers of the United Kingdom commented that PL was standard with OTC medicine packages [58].

Receipt of MI Among Specific Medicine User Groups

Altogether, 12 studies examined the receipt of MI among adults using specific medicines for various conditions, including arthritis [55-57,59], methotrexate therapy [52,53,61], cardiovascular medications [54,62], antibiotic medications [60], antimigraine treatments [51], and vasculitis medications [58] (Table 2, Multimedia Appendix 3). The majority of the studies were carried out in the early 2010s, with 7 studies performed between 2010 and 2016 [51,54-57,59,61], 4 studies between 2017 and 2025 [52,53,60,62], and 1 study in which the year of data collection was not reported [58]. The studies were conducted in several countries, including the United States [55,57-59], Spain [53,56], Germany [62], the Netherlands [54], Norway [60], and Australia [52]. A range of research methods was used, with cross-sectional surveys being the most common design (n=8) [51-53,55-59]. Of these, 3 studies used national online surveys [51-53]. In addition, 3 studies used qualitative interviews [60-62], and 1 study used questionnaire-based interviews [54]. Most studies (n=11) were conducted among outpatients [51-61], while 1 study focused on hospital inpatients [62].

Consistent with findings from studies involving heterogeneous medicine user groups, physicians, pharmacists, and PLs were the most frequently reported MI sources across specific medicine user groups, irrespective of study design, study period, or country in which the research was conducted (Figure 3 and Multimedia Appendix 3) [51-62]. No clear temporal trends in the receipt of MI were observed over time. However, the frequency of MI receipt varied by research method, patient group, and type of medicine. For instance, among Australian

XSL•FO
RenderX

outpatients using methotrexate (n=742), specialist physicians (98%), PLs (80%), and general physicians (55%) were the most commonly reported MI sources in a national online survey [52]. In contrast, a national online survey of pregnant and breastfeeding women using antimigraine medicines (n=294) identified general physicians (87%) as the primary MI source, followed by the internet (47%) and PLs (42%) [51]. In local surveys, general physicians were the leading MI sources for cardiovascular medicine users in the Netherlands (60%, n=413) [54], whereas specialist physicians (75%) and general physicians (22% - 57%) predominated among arthritis medicine and nonsteroidal anti-inflammatory drug users in Spain and the United States [55,56,59].

Figure 3. Summary of findings from the studies (n=12) on the receipt of medicines information (MI) via the internet and commonly used information sources among specific medicine user groups in 2010 - 2025. The figure contains only the results reported in the included studies. PL: package leaflet; WMI: written medicines information.

Information source	Study year		Country			Patients		Medicine use			
	2010-2016	2017-2025	Europe	United States	Australia	Outpatients	Hospital inpatients	Arthritis medicines/ NSAID	Metho-trexate	Cardio-vascular medicines	Anti-migraine medicines
Surveys, national (n=3)											
Internet	47% [40]	39% [54]	47% [40]		39% [54]	47% [40] 39% [54]			39% [54]		47% [40]
Physician (general)	87% [40]	55% [54]	87% [40]		55% [54]	87% [40] 55% [54]			55% [54]		87% [40]
Physician (specialist)	98% [54]				98% [54]	98% [54]			98% [54]		
Pharmacist	19% [40]	37% [54]	19% [40]		37% [54]	37% [54] 19% [40]			37% [54]		19% [40]
Package leaflet	42% [40]	80% [59]	80% [59] 42% [40]			80% [59] 42% [40]			80% [59]		42% [40]
Surveys, local (n=6)											
Internet	21% [44] 17% [51]			21% [44] 17% [51]		21% [44] 17% [51]		21% [44] 17% [51]			
Physician (general)	60% [42] 57% [51] 27% [44] 22% [47]		60% [42] 22% [47]	57% [51] 27% [44]		60% [42] 57% [51] 27% [44] 22% [47]		57% [51] 27% [44] 22% [47]		60% [42]	
Physician (specialist)	75% [47] 13% [42]		75% [47] 13% [42]			75% [47] 13% [42]		75% [47]		13% [42]	
Pharmacist	16% [51]			16% [51]		16% [51]		16% [51]			
Package leaflet											

Qualitative interviews (n=3)

- All patients with arthritis (n=15) reported discussions with physicians about their diagnosis and medication, and each patient recalled being given WMI. WMI was perceived as helping a little in reinforcing expectations of benefit or alleviating concerns. Patients with arthritis reported that PLs lacked detail of how the medicine works, and specifically, the nature of the benefits that they could reasonably expect. Verbal information from physicians was highly valued but focused on disease progression rather than medication efficacy or effectiveness. Receipt of MI from the internet often helped little to allay concerns and resolve dissonance [50].
- Among hospital inpatients with cardiovascular conditions (n=102), physicians (n=94, 92%), pharmacists (n=28, 27%), PLs (n=27, 26%), and the internet (n=19, 19%) were the most common MI sources on the benefits of medicines. Physicians (n=46, 45%), PLs (n=37, 36%), relatives and friends (n=30, 29%), pharmacists (n=14, 14%), and the internet (n=14, 14%) were the most common MI sources on the risks of medicines among patients with cardiovascular conditions [60].
- Among patients using antibiotic medication (n=8), MI was received from digital platforms, printed materials, and face-to-face encounters with physicians and community pharmacists [63].

In a qualitative interview study from the United Kingdom, arthritis outpatients (n=15) reported discussing their diagnosis and medication with physicians, with each patient recalling receipt of WMI [61]. While WMI provided some reassurance about benefits and alleviated concern, patients felt it lacked detail on how the medicines worked and the specific benefits they could expect. Though highly valued, verbal information from physicians focused more on disease progression than on medication efficacy. MI from the internet rarely helped resolve concerns or clarify discrepancies. Among cardiovascular hospital inpatients in Germany (n=102), physicians (92%), pharmacists (27%), and PLs (26%) were the most common MI sources regarding medication benefits, whereas physicians (45%) and PLs (36%) were the primary information sources about medication risks [62]. A Norwegian qualitative focus group study among patients using antibiotic medication (n=8) reported that patients received MI from 3 main sources: digital platforms, printed materials, and face-to-face encounters with physicians and community pharmacists [60].

Internet as a Source of MI

Receipt of MI from the internet varied substantially by population, methodology, and country (Figures 2 and 3; Multimedia Appendix 3) [37,38,40,42,44-47,50-52,55,57-62]. Receipt of MI from the internet was more common among women, individuals younger than 65 years, those with polytechnic, college, or university degrees, and frequent internet users [37,38,40,46,58]. Based on the long-term population-based trend survey from Finland [37], the interview study from Germany [62], and the local cross-sectional surveys from Sweden [50], the United Kingdom [42], and the United States [44,55], 14% - 31% of adult medicine users reported receiving MI online. However, online surveys from Australia [45,52], Finland [38], and Norway [51], and a paper- and web-based survey from Norway [46] showed higher usage, with 37% - 68% of medicine users receiving MI from the internet. Certain medicine user groups were more likely to receive MI online; for example, 47% of pregnant and breastfeeding women using antimigraine medicines [51], and 39% of patients with rheumatoid arthritis using or having used methotrexate [52] had received MI from the internet. In comparison, 19% of patients

XSL•FO
RenderX

with cardiovascular conditions [62] and 17% of nonsteroidal anti-inflammatory drug users [55] did so.

Medicine users received MI most commonly through health portals, search engines, pharmacy websites, patient organizations' websites, and medicine authorities' websites [38,46,47,52,60,61]. Patients with chronic illness often turned to patient organizations' websites (33%), while patients with mental illness (32%) and patients with thyroid disease (26%) used discussion forums [38]. Younger individuals and those who found advertising-based internet sources useful were more likely to use these platforms [44].

The internet was typically used for quick, supplementary information, often in conjunction with other MI sources [42,47,60]. However, some patients reported feeling overwhelmed by the sheer volume of MI available online [47], with most receiving MI from an average of 1 to 3 different websites [38,46]. While some medicine users were skeptical about the reliability of online sources, others found that receiving MI exacerbated concerns or dissonance [46,47]. The receipt of MI from the internet often did little to help allay concerns and resolve dissonance, but on occasion, it was reported to exacerbate it by raising further concerns [60,61]. However, some patients resorted to avoidance strategies to limit exposure to conflicting MI or disinformation or to refrain from reading official patient information to reduce anxiety about side effects.

Medicine Users Encountering Conflicting MI

Among adult prescription medicine users, 47% had sometimes or often received conflicting MI regarding their medication, primarily regarding side effects (22%) and duration of treatment (18%) [40]. Conflicting MI was most commonly received from general practitioners (82%), specialists (74%), and pharmacists (49%). Younger medicine users (18 - 50 years) were more likely to encounter conflicting MI from the internet and social media than older users (>50 years) [40,59]. Factors associated with more conflicting MI included being prescribed more than 3 medicines, higher perceived disease severity, poor medication adherence, younger age (<50 years), higher education, non-White race, and using more MI sources [40,59]. However, gender, language, nationality, and duration of chronic diseases showed no significant differences in perceptions of conflicting MI [40]. Some medicine user groups have experienced conflicting MI more than others, such as patients with arthritis (80%) [59], pregnant women (45%) [51], and patients with cardiovascular conditions (37%) [40]. In response to conflicting MI, 45% of users sought a follow-up appointment with the same physician, and 30% sought additional medical advice [40]. Furthermore, 34% - 39% of medicine users either discontinued the medication or became nonadherent following conflicting MI [40,51].

Number of MI Sources Used

Most adult medicine users received MI on their medication from multiple information sources [37,38,44,46,51,52]. Commonly, younger and more educated, as well as those with greater medical–related concerns, tended to receive and seek MI from a greater number of sources than others [46,52]. Typically, MI was received from 2 to 4 sources [37,44,46,51,52]. Internet users, on average, consulted 1-3 different websites for MI [38,46]. On the other hand, a population-based survey indicated that 28% of adult medicine users reported not receiving information about their medication from any source [37].

Discussion

Principal Findings

Despite increasing attention to the quality and accessibility of MI to medicine users in recent decades, both nationally and globally, only a limited number of studies have focused on the receipt of MI among adult medicine users over the past 15 years (2010 - 2025). Concurrently, the variety of MI sources available to consumers and medicine users has expanded considerably, particularly with the growth and diversification of electronic information channels. Nearly all studies included physicians, pharmacists, PLs, and the internet as MI sources, with considerable variation in other sources across studies. Only 1 national long-term trend study was identified [2,37], while most studies were descriptive cross-sectional designs, and only 3 applied theory-driven measures [44,54,61].

The limited and heterogeneous use of theoretical frameworks highlights a persistent gap in the current MI research. Of the included studies, three explicitly applied theoretical approaches: (1) "Protection Motivation Theory and the Health Belief Model" [54], (2) the "Comprehensive Model of Information Seeking" [44], and (3) the "Necessity-Concerns Framework" combined with "cognitive dissonance theory" [61]. Although these frameworks provide valuable insights into health behavior, information seeking, and medication-related beliefs, they were developed more than a decade ago and predate many conceptual advances that have since shaped contemporary health and digital communication research. They were formulated before the widespread adoption of social media, mobile health technologies, and algorithmically curated content, and therefore may not fully capture emerging phenomena, such as digital trust, online credibility assessment, and the interactive nature of current MI environments.

Recent theoretical developments, such as updated technology acceptance models, empowerment-oriented frameworks, and expanded conceptualizations of health literacy, medication literacy, communication processes, and adherence models, offer more comprehensive tools for understanding how adults seek, receive, interpret, evaluate, and act upon MI in increasingly digitalized contexts. Integration of these frameworks can enhance methodological rigor by guiding study design, clarifying key constructs, and supporting consistent interpretation of findings across heterogeneous studies.

The growing role of AI-based tools further underscores the need for theoretically grounded research. AI-driven platforms introduce new dimensions related to algorithmic mediation, personalization, user-system interaction, and evolving trust dynamics, effectively repositioning MI sources within a rapidly transforming digital ecosystem [64]. Future MI research would benefit from more explicit theory-driven approaches and conceptual models that reflect contemporary digital health

XSL•FO
RenderX

communication processes and the increasing influence of AI-based information systems.

The majority of studies originated from Europe (17/26 studies), followed by the United States (6/26) and Australia (4/26). In recent decades, the EU Pharmaceutical Forum, established in 2002, has played a key role in shaping and coordinating MI practices across Europe [2,31,32]. This high-level forum aimed to develop alternative strategies for enhancing consumer access to quality MI, particularly after the EU‘s ban on direct-to-consumer advertising of prescription medicines. The forum’s recommendations were published in 2008, with special emphasis on national coordination of MI, increasing the availability and use of electronic MI, and enhancing cooperation between public and private MI providers [32]. This framework has likely influenced research activities and the implementation of strategies for reliable MI throughout the EU [1,2], as can be seen in our systematic review. For instance, Finland established its ongoing National MI Strategy in 2012, inspired by the EU Pharmaceutical Forum’s recommendations [1,2,65-68].

The introduction of statutory PLs marked a significant milestone in the EU’s MI policies in the late 1990s [11,69]. Europe remains the only continent where PLs are mandatory on all pharmaceutical packaging [70]. The implementation of PLs was planned in the United States as early as the 1970s, but the mandate was only partially realized [70]. Although PLs may not encompass the full spectrum of MI in terms of structure and content [71-77], they provide an essential, easily accessible source of basic, product-specific information for medicine users. The importance of PLs is also evidenced by our systematic review, which found that PLs were among the top 3 MI sources in all the included studies, regardless of the research method, study year, population, or country of origin. Since the 1970s, the active development of PLs and related research has been particularly prominent within the EU, leading to the establishment of EU-approved readability guidelines and mandatory user testing for PLs [11,12,70-73,78]. Today, in addition to the PLs inserted in medicinal packages, PLs are made available electronically, in spoken formats, and in Braille, further enhancing access to essential MI for medicine users with special needs [11,69].

Finland stands out as the only country where we found long-term, population-level research on consumers’ and medicine users’ receipt of MI, covering the period from 1999 to 2014 [37]. Initially, this trend study was conducted annually by the National Public Health Institute as part of the national health behavior postal survey, which represented the adult population aged 15 - 64 years (conducted during 1978 - 2014). In 2015, the Finnish Medicines Agency established a new national biennial population survey, the Medicines Barometer, focusing on timely medicine-related topics important for planning national policies and strategies promoting rational pharmacotherapy. The Medicines Barometer has included a question concerning the receipt of MI among adult Finns in 2017 and 2023 [79,80]. Since 2019, this barometer survey has been conducted exclusively as an online survey. The question about the receipt of MI has been repeated in a comparable format since 1999. Such national trend studies would also be valuable in other countries, as they provide essential data to inform national decision-making in health policy and planning. A similar national trend study on the receipt of WMI was used in the United States by the Food and Drug Administration (FDA) during the 1970s to improve consumers’ access to MI [81]. Another period of national follow-up studies on the receipt of MI in the United States was conducted in the 1990s [77].

The internet as an MI source has been mainly studied at the general level, with various studies examining its role and prevalence of use across different contexts. The internet as an MI source has rarely been studied in its subtypes, such as open-source and open-access MI databases, social media–sharing networks, community forums, blogs, consumer review networks, and different apps. The existing studies have mainly used “rough” measures, overlooking the diverse range of health- and MI-related content available online. A notable trend has been the integration of health and MI into health portals and other electronic sources, including those provided by authorities or commercial entities such as the pharmaceutical industry [2].

This systematic review indicates that although the internet has primarily been studied as a general electronic source of MI, the ongoing digitalization and proliferation of electronic information channels underscore the need to investigate other specific digital MI sources and care pathways. These may include official regulatory websites, online discussion forums, and real-time chat platforms. Such sources should be examined as distinct categories in research, rather than being aggregated under the general label of “internet as a source of MI.” It is crucial to determine where users predominantly access MI online and which websites they consider reliable and credible for health and MI. Future research should also increasingly address the role and impact of misinformation and disinformation on medicine users’ perceptions of health, decision-making processes, and health-related behaviors in the digital environment [17]. These factors may have growing implications for treatment outcomes and the effective implementation of therapeutic interventions.

Our findings indicate that the receipt of MI among adult medicine users has diversified during the 2010s. Medicine users increasingly receive MI from multiple sources, as the internet makes it easy and fast to access various types of information. They also seem to commonly receive MI from traditional sources, such as HCPs (eg, physicians, pharmacists, and nurses), media (eg, television, radio, and newspapers), the internet, specific patient-directed information sources (eg, PLs and telephone services), and personal contacts (eg, family, friends, and peers). Future research should focus on investigating the extent, reliability, and impact of commercial information as an MI source, use more theory-driven study designs and measures to gain a more comprehensive understanding of the receipt of MI among medicine users, and address the need for the development of existing MI sources from the medicine user perspective. It is also important to plan more strategic MI research so that it can cover patients and medicine users from a much larger range of conditions, including self-medication, than is the case in current research.

XSL•FO
RenderX

Strengths and Limitations

This systematic review provides a comprehensive overview of the receipt of MI among adult medicine users across diverse settings in developed countries. A further strength is the systematic examination of how MI sources were defined and operationalized across studies, enabling transparent comparison of patterns despite differences in study designs.

A key finding was the substantial heterogeneity across the included studies, spanning study designs, populations, settings, operationalizations of MI, and analytical approaches. As detailed in Multimedia Appendix 3, studies varied markedly in recruitment strategies, definitions and measurement of MI sources, and outcome measures. This multidimensional heterogeneity, particularly in methodology, population, medication in use, and MI sources, precluded conducting a meta-analysis and supports the use of a narrative synthesis to interpret patterns in the receipt of MI. Importantly, pooling effect estimates from studies with these limitations would have risked producing a summary effect with questionable validity and misleading precision. In addition, although meta-analysis can be used to explore heterogeneity, the limited number of studies within meaningful subgroups would have made such analyses difficult. Combined with diverse outcome measures and variable study quality, any statistical exploration of heterogeneity would have been unreliable and potentially misleading.

The restriction to English-language, peer-reviewed full-text papers may have introduced both language and publication bias. In addition, differences in health care systems and service delivery contexts (eg, hospital, outpatient care, and community pharmacy) likely influenced access to and use of MI sources, limiting cross-study comparability. The studies were conducted in diverse settings, which may have substantially impacted the results. Nearly half of the included studies (17/26, 65%) originated from a variety of European countries [37-42,46,49-51,53,54,56,60-62], which may affect generalizability to other regions.

Methodological limitations in the quantitative studies further reduced confidence in the evidence base. Notably, nonresponse bias was often insufficiently assessed or not reported, and was sometimes rated as low or reported without adequate justification despite low or uncountable response rates (eg, convenience samples and online surveys), reducing confidence in the robustness and generalizability of the evidence. Given the broad scope of the topic, it was necessary to refine and expand the predetermined exclusion criteria during the screening process. Applying these stringent criteria ensured that the review remained focused on the receipt of MI among adult medicine users. However, this inevitably resulted in the exclusion of a substantial number of studies not directly aligned with the objective of this review.

Although a range of research methods was used, the majority of studies were cross-sectional, using a single cross-sectional design (18/26, 69%) [38-46,51-53,55-60], and many relied on online surveys, potentially overrepresenting respondents who are more easily reached through digital platforms. Only 1 study analyzed long-term trends in the receipt of MI more than 16 years (from 1999 to 2014) [37], highlighting the critical need for further long-term, population-based analyses using diverse MI sources among medicine users. It is essential to recognize that none of the reviewed studies used a longitudinal design. The studies also did not separate internet sources into different subcategories, limiting the ability to assess the internet as a multidimensional MI source. Furthermore, only 3 studies used a theoretical framework [44,54,61], highlighting significant gaps to be addressed in future research.

A major limitation of the reviewed literature is the lack of population-level studies assessing the receipt of internet-based MI. Existing national studies from Finland indicate that the prevalence estimates are substantially lower in representative population samples than in online surveys or selected specific medicine user groups [2,37]. This pattern is corroborated by recent national monitoring studies in Finland, which show that, at the population level, the internet has not yet reached the same prevalence as traditional MI sources (eg, physicians, community pharmacists, and PLs) [79,80].

Because the review intentionally included only studies from developed countries, the findings predominantly apply to this context and do not reflect the situation in developing countries. Consequently, generalization should consider differences in MI practices, including accessibility and use of WMI and internet sources across countries and continents. Future research should investigate the receipt of MI in middle- and low-income countries, as well as among immigrant populations. Furthermore, it would be beneficial to monitor the receipt of MI on both national and international levels using tools such as barometers or automated research methods (eg, robotic surveys) with comparable measures and study designs. Tracking the evolution of consumers' access to MI will facilitate more targeted, patient-centered approaches, ultimately guiding consumers and medicine users toward more reliable and accurate MI sources. The incorporation of structured futures methodologies into MI research may further strengthen strategic planning and support systematic anticipation of emerging developments in digital health communication, AI-mediated information environments, and evolving medicine user behaviors [82].

Finally, heterogeneity in sample composition limited interpretability: half of the studies (14/26, 54%) examined heterogeneous medicine user groups, making it difficult to identify the specific medicines used [37-50,60]. This variability also hindered direct comparisons with studies focused on patients with specific medications or chronic conditions. Furthermore, the studies included in the review covered only a small proportion of the overall medicine user population. Five studies did not report the year of data collection [39,43,44,47,58], which may limit the temporal relevance of some findings.

Conclusions

Despite the increasing role of the internet and social media, traditional sources of MI, such as physicians, pharmacists, and PLs, remain the most commonly used among adult medicine users. Receipt of MI from electronic sources appears more frequent among individuals familiar with the internet, although no clear or consistent temporal increase was identified.

XSL•FO
RenderX

This review is innovative in systematically comparing traditional and electronic MI sources across developed economies during a period of rapid digital transformation. In contrast to much of the existing literature, which has focused on single modalities or treated the internet as a homogeneous source, this study highlights the limited differentiation of digital MI subtypes and the lack of consistent theoretical grounding in current research.

By synthesizing cross-national evidence and identifying methodological and conceptual gaps, this review advances the field by clarifying how traditional and emerging digital MI sources coexist in contemporary health care environments. The findings underscore the need to distinguish between websites, social media platforms, online forums, and AI-mediated tools when examining the receipt of MI and its effects on medication-related perceptions, decisions, and behaviors.

From a practical perspective, these results inform HCPs, policymakers, and digital health developers seeking to optimize reliable, user-centered MI strategies. Addressing digital literacy, misinformation, and equitable access will be increasingly important as algorithmically mediated and AI-based MI tools continue to reshape how medicine users access and act upon MI in real-world settings.

Acknowledgments

The authors wish to acknowledge Professor Katri Hämeen-Anttila for her contribution to the early phase of this research. The information content of the paper, including figures and tables, was created by the researchers without the help of generative artificial intelligence (AI). AI-based tools, specifically Grammarly (Grammarly Inc.) and ChatGPT (OpenAI), were used for final language editing and proofreading.

Funding

The authors declare that no financial support was received for this work.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request. The data (ie, published papers included in the systematic review) are available in Multimedia Appendix 3.

Authors' Contributions

NM contributed to the conceptualization and design of the work, data collection, selection, analysis, visualization, writing the original draft, reviewing and editing the manuscript, critical review, and final approval. MMS contributed to data collection, selection, analysis, and final approval. DT contributed to data collection, selection, and final approval. MPM contributed to the conceptualization and design of the work, data analysis, writing the original draft, reviewing and editing the manuscript, critical review, and final approval. MA contributed to the conceptualization and design of the work, data collection, selection, analysis, visualization, writing the original draft, reviewing and editing the paper, critical review, and final approval.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies and databases used in the literature review.
[PDF File, 69 KB - jmir_v28i1e71984_app1.pdf]

Multimedia Appendix 2
Criteria of the Mixed Methods Appraisal Tool (MMAT; version 2018).
[PDF File, 76 KB - jmir_v28i1e71984_app2.pdf]

Multimedia Appendix 3
Detailed data extraction table of included studies (n=26).
[PDF File, 682 KB - jmir_v28i1e71984_app3.pdf]

Checklist 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist for systematic reviews.
[PDF File, 215 KB - jmir_v28i1e71984_app4.pdf]

Checklist 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist for abstracts.

XSL•FO
RenderX

[PDF File, 155 KB - jmir_v28i1e71984_app5.pdf]

Checklist 3
PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search extension) checklist for reporting literature searches in systematic reviews.
[PDF File, 126 KB - jmir_v28i1e71984_app6.pdf]

References

1. Medicines information: strategic development. International Pharmaceutical Federation. 2017. URL: https://www.fip.org/files/fip/publications/2017-01-Medicines-information-strategic-development.pdf [accessed 2025-01-13]
2. Mononen N. From paper to cyber: medicines information as a strategic goal in Finland and the European Union. : University of Helsinki; 2020 URL: http://hdl.handle.net/10138/313665 [accessed 2025-01-13]
3. Hollingsworth K, Toscani M. Risk evaluation mitigation strategies: the evolution of risk management policy. Popul Health Manag 2013 Apr;16(2):132-137. [doi: 10.1089/pop.2012.0021] [Medline: 23113627]
4. Emanuel M, Rawlins M, Duff G, Breckenridge A. Thalidomide and its sequelae. Lancet 2012 Sep 1;380(9844):781-783. [doi: 10.1016/S0140-6736(12)60468-1] [Medline: 22939670]
5. Kim JH, Scialli AR. Thalidomide: the tragedy of birth defects and the effective treatment of disease. Toxicol Sci 2011 Jul;122(1):1-6. [doi: 10.1093/toxsci/kfr088] [Medline: 21507989]
6. Vargesson N. Thalidomide - induced teratogenesis: history and mechanisms. Birth Defects Research Pt C 2015 Jun;105(2):140-156 [FREE Full text] [doi: 10.1002/bdrc.21096]
7. The importance of pharmacovigilance: safety monitoring of medicinal products. World Health Organization. 2002. URL: https://www.who.int/publications/i/item/10665-42493 [accessed 2025-01-13]
8. Wahlroos H. Control of medicines in the European Union and drug information. Pharmaceuticals Policy and Law 2005 Feb;6(1):59-67. [doi: 10.3233/PPL-2005-00083]
9. History of FDA patient package insert requirements. Am J Health Syst Pharm 1980 Dec 1;37(12):1660-1662. [doi: 10.1093/ajhp/37.12.1660]
10. Always read the leaflet: getting the best information with every medicine. : Medicines and Healthcare products Regulatory Agency; 2005 URL: https://assets.publishing.service.gov.uk/media/5a748da340f0b61938c7e5e9/Always_Read_the_Leaflet___getting_the_best_information_with_every_medicine.pdf [accessed 2026-04-23]
11. Council Directive 92/27/EEC of 31 March 1992 on the labelling of medicinal products for human use and on package leaflets. EUR-Lex. 1992. URL: https://eur-lex.europa.eu/eli/dir/1992/27/oj [accessed 2025-01-13]
12. Directive 2004/27/EC of the European Parliament and of the Council of 31 March 2004 amending Directive 2001/83/EC on the Community code relating to medicinal products for human use (Text with EEA relevance). EUR-Lex. 2004. URL: http://data.europa.eu/eli/dir/2004/27/oj [accessed 2025-02-12]
13. United States Pharmacopeia. USP medication counseling behavior guideline. In: USP DI: Drug Information for the Health Care Professional Volumes I and II: Rockville, MD: United States Pharmacopeia Convention; 1997:664-675.
14. Weigmann K. Consumer reporting of adverse drug reactions: systems that allow patients to report side effects of the drugs they are taking have yielded valuable information for improving drugs safety and health care. EMBO Rep 2016 Jul;17(7):949-952. [doi: 10.15252/embr.201642616] [Medline: 27198546]
15. Inácio P. The value of patient reporting of adverse drug reactions to pharmacovigilance systems. : University of Helsinki; 2018 URL: http://hdl.handle.net/10138/242574 [accessed 2026-02-24]
16. Mesko B, deBronkart D, Dhunnoo P, Arvai N, Katonai G, Riggare S. The evolution of patient empowerment and its impact on health care's future. J Med Internet Res 2025 May 1;27:e60562. [doi: 10.2196/60562] [Medline: 40311140]
17. Kbaier D, Kane A, McJury M, Kenny I. Prevalence of health misinformation on social media-challenges and mitigation before, during, and beyond the COVID-19 pandemic: scoping literature review. J Med Internet Res 2024 Aug 19;26:e38786. [doi: 10.2196/38786] [Medline: 39159456]
18. Stifjell K, Sandanger TM, Wien C. Exploring online health information-seeking behavior among young adults: scoping review. J Med Internet Res 2025 Sep 9;27:e70379. [doi: 10.2196/70379] [Medline: 40925001]
19. Zhao YC, Zhao M, Song S. Online health information seeking behaviors among older adults: systematic scoping review. J Med Internet Res 2022 Feb 16;24(2):e34790. [doi: 10.2196/34790] [Medline: 35171099]
20. Zhang C, Mohamad E, Azlan AA, Wu A, Ma Y, Qi Y. Social media and eHealth literacy among older adults: systematic literature review. J Med Internet Res 2025 Mar 26;27:e66058. [doi: 10.2196/66058] [Medline: 40138684]
21. Cao W, Milks MW, Liu X, et al. mHealth interventions for self-management of hypertension: framework and systematic review on engagement, interactivity, and tailoring. JMIR Mhealth Uhealth 2022 Mar 2;10(3):e29415. [doi: 10.2196/29415] [Medline: 35234655]
22. Petkovic J, Duench S, Trawin J, et al. Behavioural interventions delivered through interactive social media for health behaviour change, health outcomes, and health equity in the adult population. Cochrane Database Syst Rev 2021;2021(6):CD012932. [doi: 10.1002/14651858.CD012932.pub2]

XSL•FO
RenderX

23. Allida S, Du H, Xu X, et al. mHealth education interventions in heart failure. Cochrane Database Syst Rev 2020;2020(7):CD011845. [doi: 10.1002/14651858.CD011845.pub2]
24. Song M, Elson J, Bastola D. Digital age transformation in patient-physician communication: 25-year narrative review (1999-2023). J Med Internet Res 2025 Jan 16;27:e60512. [doi: 10.2196/60512] [Medline: 39819592]
25. Zaghloul H, Fanous K, Ahmed L, et al. Digital health literacy in patients with common chronic diseases: systematic review and meta-analysis. J Med Internet Res 2025 Aug 25;27:e56231. [doi: 10.2196/56231] [Medline: 40854231]
26. Berhanu D, Leal Rato M, Canhoto AI, Vieira da Cunha J, Geraldes R. Online information search by people with multiple sclerosis: a systematic review. Mult Scler Relat Disord 2023 Nov;79:105032. [doi: 10.1016/j.msard.2023.105032] [Medline: 37801957]
27. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. [doi: 10.1136/bmj.n71]
28. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. jmla 2021;109(2):174-200. [doi: 10.5195/jmla.2021.962]
29. The world's #1 systematic review tool. Covidence. URL: https://www.covidence.org [accessed 2022-10-05]
30. World Economic Situation and Prospects 2018. : United Nations; 2018 URL: https://www.un.org/development/desa/dpad/wp-content/uploads/sites/45/publication/WESP2018_Full_Web.pdf [accessed 2026-04-23]
31. High Level Group on innovation and provision of medicines in the European Union: recommendations for action. : European Commission; 2002 URL: http://ec.europa.eu/health/ph_overview/Documents/key08_en.pdf [accessed 2025-11-17]
32. High Level Pharmaceutical Forum 2005-2008: conclusions and recommendations. : European Commission; 2008. [doi: 10.2772/22720]
33. Hong QN, Gonzalez-Reyes A, Pluye P. Improving the usefulness of a tool for appraising the quality of qualitative, quantitative and mixed methods studies, the Mixed Methods Appraisal Tool (MMAT). J Eval Clin Pract 2018 Jun;24(3):459-467. [doi: 10.1111/jep.12884] [Medline: 29464873]
34. Pace R, Pluye P, Bartlett G, et al. Testing the reliability and efficiency of the pilot Mixed Methods Appraisal Tool (MMAT) for systematic mixed studies review. Int J Nurs Stud 2012 Jan;49(1):47-53. [doi: 10.1016/j.ijnurstu.2011.07.002] [Medline: 21835406]
35. Hong QN, Pluye P, Fàbregues S, et al. Improving the content validity of the mixed methods appraisal tool: a modified e-Delphi study. J Clin Epidemiol 2019 Jul;111:49-59. [doi: 10.1016/j.jclinepi.2019.03.008] [Medline: 30905698]
36. Higgins JPT, Thomas J, Chandler J, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 65: Cochrane; 2024. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook [accessed 2026-04-23]
37. Mononen N, Airaksinen MSA, Hämeen-Anttila K, Helakorpi S, Pohjanoksa-Mäntylä M. Trends in the receipt of medicines information among Finnish adults in 1999-2014: a nationwide repeated cross-sectional survey. BMJ Open 2019 Jun 14;9(6):e026377. [doi: 10.1136/bmjopen-2018-026377] [Medline: 31203239]
38. Hämeen-Anttila K, Pietilä K, Pylkkänen L, Pohjanoksa-Mäntylä M. Internet as a source of medicines information (MI) among frequent internet users. Research in Social and Administrative Pharmacy 2018 Aug;14(8):758-764. [doi: 10.1016/j.sapharm.2017.09.007]
39. O'Donovan B, Rodgers RM, Cox AR, Krska J. Use of information sources regarding medicine side effects among the general population: a cross-sectional survey. Prim Health Care Res Dev 2019 Dec 10;20:e153. [doi: 10.1017/S1463423619000574] [Medline: 31818345]
40. Santos B, Blondon KS, Van Gessel E, et al. Patients' perceptions of conflicting information on chronic medications: a prospective survey in Switzerland. BMJ Open 2022 Nov 3;12(11):e060083. [doi: 10.1136/bmjopen-2021-060083] [Medline: 36328384]
41. Krska J, Morecroft CW. Informing patients about medicines--a hospital in-patient survey in England. Patient Educ Couns 2013 Feb;90(2):276-278. [doi: 10.1016/j.pec.2012.09.011] [Medline: 23083903]
42. Krska J, Morecroft CW. Patients' use of information about medicine side effects in relation to experiences of suspected adverse drug reactions: a cross-sectional survey in medical in-patients. Drug Saf 2013 Aug;36(8):673-680. [doi: 10.1007/s40264-013-0065-3] [Medline: 23700262]
43. Cooper JM, Garrett T. Providing medicines information and education to hospital in-patients: patients' experiences and preferences. J Pharm Pract Res 2014 Dec;44(4):213-219. [doi: 10.1002/jppr.1035]
44. DeLorme DE, Huh J, Reid LN. Source selection in prescription drug information seeking and influencing factors: applying the comprehensive model of information seeking in an American context. J Health Commun 2011 Aug;16(7):766-787. [doi: 10.1080/10810730.2011.561914] [Medline: 21614720]
45. Perry D, Librizzi B, Ngu L, et al. Medication information and supply behaviours in elite and developing athletes. J Sci Med Sport 2020 Jun;23(6):548-553. [doi: 10.1016/j.jsams.2019.12.026] [Medline: 32044233]
46. Bergmo TS, Sandsdalen V, Manskow US, Småbrekke L, Waaseth M. Internet use for obtaining medicine information: cross-sectional survey. JMIR Form Res 2023 Feb 2;7:e40466. [doi: 10.2196/40466] [Medline: 36729577]
47. Haverhals LM, Lee CA, Siek KA, et al. Older adults with multi-morbidity: medication management processes and design implications for personal health applications. J Med Internet Res 2011 Jun 29;13(2):e44. [doi: 10.2196/jmir.1813] [Medline: 21715286]

XSL•FO
RenderX

48. Tong V, Raynor DK, Aslani P. Receipt and use of spoken and written over-the-counter medicine information: insights into Australian and UK consumers' experiences. Int J Pharm Pract 2018 Apr;26(2):129-137. [doi: 10.1111/ijpp.12382] [Medline: 28544204]
49. Mackridge AJ, Rodgers R, Lee D, Morecroft CW, Krska J. Cross-sectional survey of patients' need for information and support with medicines after discharge from hospital. Int J Pharm Pract 2018 Oct;26(5):433-441. [doi: 10.1111/ijpp.12411] [Medline: 29152842]
50. Bastholm-Rahmner P, Gustafsson LL, Aggefors K, et al. Patients' knowledge and attitudes to the Wise List - a drug formulary from the Stockholm Drug and Therapeutic committee. BMC Health Serv Res 2018 Mar 12;18(1):176. [doi: 10.1186/s12913-018-2968-2] [Medline: 29530028]
51. Amundsen S, Øvrebø TG, Amble NMS, Poole AC, Nordeng H. Use of antimigraine medications and information needs during pregnancy and breastfeeding: a cross-sectional study among 401 Norwegian women. Eur J Clin Pharmacol 2016 Dec;72(12):1525-1535. [doi: 10.1007/s00228-016-2127-9] [Medline: 27624901]
52. Leonardo N, Lester S, Graham M, et al. Selection and perception of methotrexate treatment information in people with rheumatoid arthritis. Int J Rheum Dis 2020 Jun;23(6):805-812. [doi: 10.1111/1756-185X.13833] [Medline: 32537893]
53. Otón T, Carmona L, Andreu JL. What do patients on methotrexate need and expect at the clinic? An online patient survey. Rheumatol Int 2023 Apr;43(4):735-741. [doi: 10.1007/s00296-022-05249-3] [Medline: 36436083]
54. Bults M, Beaujean D, Wijkmans CJ, Timen A, Richardus JH, Voeten HACM. Why did patients with cardiovascular disease in the Netherlands accept Q fever vaccination? Vaccine (Auckl) 2012 May 14;30(23):3369-3375. [doi: 10.1016/j.vaccine.2012.03.056] [Medline: 22475862]
55. Houser SH, Au DW, Miller MJ, et al. Socio-demographic differences in risk information seeking sources for non-steroidal anti-inflammatory drugs (NSAIDS). Int J Med Inform 2016 Oct;94:222-227. [doi: 10.1016/j.ijmedinf.2016.07.017] [Medline: 27573330]
56. de Toro J, Cea-Calvo L, Battle E, et al. Perceptions of patients with rheumatic diseases treated with subcutaneous biologicals on their level of information: RHEU-LIFE survey. Reumatol Clin (Engl Ed) 2019;15(6):343-349. [doi: 10.1016/j.reuma.2017.11.006] [Medline: 29277578]
57. Geryk LL, Blalock S, DeVellis RF, Morella K, Carpenter DM. Associations between patient characteristics and the amount of arthritis medication information patients receive. J Health Commun 2016 Oct;21(10):1122-1130. [doi: 10.1080/10810730.2016.1222036] [Medline: 27668523]
58. Carpenter DM, DeVellis RF, Hogan SL, Fisher EB, DeVellis BM, Jordan JM. Use and perceived credibility of medication information sources for patients with a rare illness: differences by gender. J Health Commun 2011 Jul;16(6):629-642. [doi: 10.1080/10810730.2011.551995] [Medline: 21476166]
59. Carpenter DM, Elstad EA, Blalock SJ, DeVellis RF. Conflicting medication information: prevalence, sources, and relationship to medication adherence. J Health Commun 2014;19(1):67-81. [doi: 10.1080/10810730.2013.798380] [Medline: 24015878]
60. Bergsholm YKR, Feiring M, Charnock C, Holm LB, Krogstad T. Exploring patients' adherence to antibiotics by understanding their health knowledge and relational communication in encounters with pharmacists and physicians. Explor Res Clin Soc Pharm 2023 Dec;12:100372. [doi: 10.1016/j.rcsop.2023.100372] [Medline: 38089697]
61. Hayden C, Neame R, Tarrant C. Patients' adherence-related beliefs about methotrexate: a qualitative study of the role of written patient information. BMJ Open 2015 May 20;5(5):e006918. [doi: 10.1136/bmjopen-2014-006918] [Medline: 25995237]
62. Wakob I, Wintsche I, Frisch A, et al. Assessment of patients' views on drug benefits and risks: an interview study with cardiovascular patients. Int J Clin Pract 2022;2022:6585271. [doi: 10.1155/2022/6585271] [Medline: 36474547]
63. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-Analysis: John Wiley & Sons; 2009. URL: https://doi.org/10.1002/9780470743386 [accessed 2026-04-23]
64. Ayo-Ajibola O, Davis RJ, Lin ME, Riddell J, Kravitz RL. Characterizing the adoption and experiences of users of artificial intelligence-generated health information in the United States: cross-sectional questionnaire study. J Med Internet Res 2024 Aug 14;26:e55138. [doi: 10.2196/55138] [Medline: 39141910]
65. Hämeen-Anttila K, Luhtanen S, Airaksinen M, Pohjanoksa-Mäntylä M. Developing a national medicines information strategy in Finland--a stakeholders' perspective on the strengths, challenges and opportunities in medicines information. Health Policy 2013 Jul;111(2):200-205. [doi: 10.1016/j.healthpol.2013.04.005] [Medline: 23683473]
66. Mononen N, Pohjanoksa-Mäntylä M, Airaksinen MS, Hämeen-Anttila K. How far are we from a medication use process aiming at well-informed adherent patients with long-term medications in Finland? Qualitative study. BMJ Open 2020 Jun 21;10(6):e036526. [doi: 10.1136/bmjopen-2019-036526] [Medline: 32565471]
67. Knowledge for the rational use of medicines: current status and strategy of drug information activities until 2020. : Finnish Medicines Agency Fimea; 2012 URL: https://urn.fi/URN:ISBN:978-952-5624-20-5 [accessed 2025-12-02]
68. Kiviranta P, Hämeen-Anttila K. The medicine user at the centre of medicines information: National Medicines Information Strategy 2021-2026. : Finnish Medicines Agency Fimea; 2019 URL: https://urn.fi/URN:ISBN:978-952-7299-16-6 [accessed 2025-12-02]
69. Directive 2001/83/EC of the european parliament and of the council of 6 november 2001 on the community code relating to medicinal products for human use. EUR-Lex. 2001. URL: https://eur-lex.europa.eu/eli/dir/2001/83/oj [accessed 2025-01-13]

XSL•FO
RenderX

70. Raynor DK, Blenkinsopp A, Knapp P, et al. A systematic review of quantitative and qualitative research on the role and effectiveness of written information available to patients about individual medicines. Health Technol Assess 2007 Feb;11(5):iii. [doi: 10.3310/hta11050] [Medline: 17280623]
71. Nicolson D, Knapp P, Raynor DKT, Spoor P. Written information about individual medicines for consumers. Cochrane Database Syst Rev 2009 Apr 15;2009(2):CD002104. [doi: 10.1002/14651858.CD002104.pub3] [Medline: 19370575]
72. Raynor DK, Knapp P, Silcock J, Parkinson B, Feeney K. “User-testing” as a method for testing the fitness-for-purpose of written medicine information. Patient Educ Couns 2011 Jun;83(3):404-410. [doi: 10.1016/j.pec.2011.03.016] [Medline: 21530140]
73. van Dijk L, Monteiro SP, Vervloet M, de Bie J, Raynor DKT. Study on the package leaflets and the summaries of product characteristics of medicinal products for human use. : Netherlands Institute for Health Services Research; 2014 URL: https://www.nivel.nl/sites/default/files/bestanden/1002395.pdf [accessed 2025-12-02]
74. Dickinson D, Raynor DK, Duman M. Patient information leaflets for medicines: using consumer testing to determine the most effective design. Patient Educ Couns 2001 May;43(2):147-159. [doi: 10.1016/s0738-3991(00)00156-7] [Medline: 11369148]
75. Pander Maat H, Lentz L, Raynor DK. How to test mandatory text templates: the European patient information leaflet. PLoS ONE 2015;10(10):e0139250. [doi: 10.1371/journal.pone.0139250] [Medline: 26445004]
76. Raynor DK. User testing in developing patient medication information in Europe. Res Social Adm Pharm 2013;9(5):640-645. [doi: 10.1016/j.sapharm.2013.02.007] [Medline: 23583081]
77. Raynor DKT, Svarstad B, Knapp P, et al. Consumer medication information in the United States, Europe, and Australia: a comparative evaluation. J Am Pharm Assoc (2003) 2007;47(6):717-724. [doi: 10.1331/JAPhA.2007.06141] [Medline: 18032134]
78. Guideline on the readability of the label and package leaflet of medicinal products for human use. : European Commission; 2009 URL: https://health.ec.europa.eu/system/files/2016-11/2009_01_12_readability_guideline_final_en_0.pdf [accessed 2025-02-12]
79. Miettinen H, Mikkola H, Jyrkkä J. Medicine barometer 2023: prerequisites for drug users to implement drug treatment in everyday life. : Finnish Medicines Agency (Fimea); 2023 URL: https://urn.fi/URN:ISBN:978-952-7299-62-3 [accessed 2026-04-23]
80. Jauhonen HM, Merikoski M, Jyrkkä J, Hämeen-Anttila K. Pharmaceutical barometer 2017: conditions for partnership in pharmaceutical care. : Finnish Medicines Agency (Fimea); 2018 URL: https://urn.fi/URN:ISBN:978-952-5624-87-8 [accessed 2026-04-23]
81. Morris LA. A survey of patients’ receipt of prescription drug information. Med Care 1982 Jun;20(6):596-605. [doi: 10.1097/00005650-198206000-00005] [Medline: 7109742]
82. Mesko B, Kristóf T, Dhunnoo P, Árvai N, Katonai G. A practical guide to using futures methods in health care: approaches, applications, and case studies. J Med Internet Res 2025 Sep 16;27:e82820. [doi: 10.2196/82820] [Medline: 40956602]

Abbreviations

AI: artificial intelligence
EU: European Union
FDA: Food and Drug Administration
HCP: health care professional
MI: medicines information
MMAT: Mixed Methods Appraisal Tool
OTC: over-the-counter
PL: package leaflet
PRISMA: Preferred Reporting Items for Systematic Review and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search extension
WMI: written medicines information

Edited by S Brini; submitted 14.Feb.2025; peer-reviewed by M Ali, YY Fang; revised version received 25.Feb.2026; accepted 03.Mar.2026; published 20.May.2026.

Please cite as:
Mononen N, Salo MM, Timonen D, Pohjanoksa-Mäntylä M, Airaksinen M
Receipt of Medicines Information From the Internet and Other Information Sources Among Adult Medicine Users in Developed Economies, 2010-2025: Systematic Review
J Med Internet Res 2026;28:e71984
URL: https://www.jmir.org/2026/1/e71984
doi:10.2196/71984

XSL•FO
RenderX

© Niina Mononen, Milla-Maria Salo, Daria Timonen, Marika Pohjanoksa-Mäntylä, Marja Airaksinen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Efficacy of Internet-Based Self-Help Interventions for Irritable Bowel Syndrome: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Xiaohong Xu[1*], MSN; Ting Liu[1*], MSN; Fang Wang[2], MD; Hongyue He[1], MSN; Jiao Chen[1], MSN; Jingxian Liu[1], MSN

[1]College of Nursing, Chengdu University of Traditional Chinese Medicine, Chengdu, Sichuan, China

[2]Dean's Office, Guang'an Hospital of Traditional Chinese Medicine, No. 1 Cuiping Road, Guang'an District, Guang'an City, Guang'an, Sichuan, China

[*]these authors contributed equally

Corresponding Author:
Fang Wang, MD
Dean's Office, Guang'an Hospital of Traditional Chinese Medicine, No. 1 Cuiping Road, Guang'an District, Guang'an City, Guang'an, Sichuan, China

Abstract

Background: Irritable bowel syndrome (IBS) is a common functional gastrointestinal disorder that reduces quality of life and causes a heavy medical burden. Internet-based self-help interventions are flexible and scalable, showing potential for IBS symptom improvement, but relevant evidence is fragmented and lacks systematic review.

Objective: This systematic review aimed to comprehensively evaluate the effects of internet-based self-help interventions on symptom severity, quality of life, and visceral sensitivity, as well as comorbid depressive and anxiety symptoms, in individuals with IBS.

Methods: Only randomized controlled trials evaluating internet-based self-help interventions for individuals with IBS were included. A literature search was conducted across PubMed, Embase, Web of Science, CINAHL Complete, PsycINFO, the Cochrane Library, and 4 Chinese databases on June 25, 2025, with an updated search on March 9, 2026. Risk of bias was assessed using the Cochrane Risk of Bias 2.0 tool. Meta-analyses were performed with the Hartung-Knapp-Sidik-Jonkman-adjusted random-effects model. Effect sizes were reported as standardized mean differences (SMDs) with 95% CIs, and evidence certainty was evaluated using GRADE (Grading of Recommendations Assessment, Development, and Evaluation) criteria.

Results: A total of 17 randomized controlled trials from 7 countries were included, involving 2289 participants (predominantly female). Compared with control groups, internet-based self-help interventions were associated with a statistically significant improvement in IBS symptom severity (SMD −0.52, 95% CI −0.78 to −0.26, 95% prediction interval [PI] −1.46 to 0.42), quality of life (SMD 0.57, 95% CI 0.23 to 0.90, 95% PI −0.41 to 1.54), visceral sensitivity (SMD −0.55, 95% CI −0.89 to −0.21, 95% PI −1.43 to 0.33), and depressive symptoms (SMD −0.14, 95% CI −0.27 to −0.01, 95% PI −0.27 to −0.01). However, no statistically significant improvement was observed in anxiety symptoms (SMD −0.03, 95% CI −0.34 to 0.27, 95% PI –0.78 to 0.71). The certainty of evidence was rated as moderate to very low for all outcomes.

Conclusions: This review synthesizes the latest evidence on internet-based self-help interventions for individuals with IBS disease-specific and comorbid psychological symptoms. It stands out by encompassing a diverse range of such interventions and incorporating visceral sensitivity as a key outcome. In doing so, it establishes a more comprehensive multi-outcome evidence base for IBS digital interventions, advancing the field by clarifying the potential of these interventions as viable alternatives to conventional treatments. For real-world practice, these findings can inform targeted strategies for primary care and telemedicine platforms, especially in resource-limited regions. However, this review is limited by moderate bias risk, high heterogeneity, and moderate to very low GRADE evidence certainty. A wide 95% PI suggests that effect variability is linked to contextual and population factors, so findings should be interpreted cautiously. Future research should prioritize technical support, patients' digital health literacy, and standardized intervention protocols to further validate clinical utility.

Trial Registration: PROSPERO CRD420251080716; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251080716

(J Med Internet Res 2026;28:e87216) doi:10.2196/87216

KEYWORDS

digital health; irritable bowel syndrome; self-help behavior; cognitive behavioral therapy; meta-analysis

XSL•FO
RenderX

Introduction

Irritable bowel syndrome (IBS) is a common chronic gastrointestinal disorder affecting 5% - 11% of adults globally [1]. Its hallmark features include recurrent abdominal pain, bloating, altered bowel habits, and stool trait changes, often with urgency or incomplete evacuation [2]. Beyond physical discomfort, IBS is strongly linked to psychological conditions like anxiety, depression, hypochondriasis, and somatization [3,4]. A meta-analysis shows individuals with IBS have nearly 9 times the risk of depression compared with healthy individuals [5], with 28.8% experiencing depression and 39.1% anxiety [6]. The severity of these psychological effects is underscored by approximately 38% of individuals contemplating suicide [7].

These mental health challenges are linked to increased mortality rates and heightened susceptibility to disease [8]. Moreover, IBS significantly impairs patients' quality of life and social functioning, imposing heavy societal burdens [1]. A 2015 survey by the American College of Gastroenterology found that IBS symptoms cause an average of 9 days of reduced productivity and 2 days of work absence monthly due to illness [9]. In the United States alone, annual indirect economic losses from IBS, due to work absenteeism and reduced productivity, are estimated at up to US $20 billion [10]. Thus, effective management to alleviate IBS symptoms, especially by addressing psychological issues, is critical.

A large proportion of individuals with IBS experience significant psychological distress or comorbid mental health disorders [11]. Mounting evidence points to a bidirectional relationship between psychological comorbidities and gastrointestinal function [12-14]. Though the exact pathogenesis of IBS is not fully understood, it can be framed via a bio psycho social model: psychosocial factors regulate physiological functions through the brain-gut axis, inducing somatization symptoms [15]; conversely, gastrointestinal pathological changes affect the central nervous system, worsening psychological symptoms [16].

Presently, psychological interventions like cognitive behavioral therapy (CBT), gut-directed hypnotherapy, stress management, and mental health education are proven to positively influence intestinal symptoms and the psychological state of individuals with IBS [17]. UK National Institute for Health and Care Excellence guidelines recommend considering referral to psychological interventions for those with persistent symptoms unresponsive to 12-month pharmacological treatment [18]. However, such therapies face challenges in clinical application. For example, psychological interventions involve high costs and poor patient compliance [19]. Moreover, reliance on professionals leads to limited accessibility [20], which is particularly notable in remote areas with scarce resources. Additionally, individuals with IBS with recurrent symptoms and psychological comorbidities need ongoing nursing support, but clinical care is constrained by human resources and follow-up cycles, making it hard to provide standardized, high-frequency symptom management guidance, resulting in insufficient patient self-management capacity.

Given the limitations of traditional intervention models and individuals with IBS's need for self-management, symptom self-help strategies have become a key focus in clinical treatment. Self-help intervention, a self-driven approach targeting psychological issues, is part of standardized psychological therapy [21]. It uses diverse therapeutic resources with or without therapist guidance and enables individuals to learn independently for in-depth understanding and effective management of psychological issues [22]. Research by Dorn et al [23] shows that various self-help approaches benefit individuals with IBS symptoms and psychosocial status. As a nonpharmacological therapy, it helps individuals manage their condition and reduce health care costs through existing resources [23].

The delivery formats for self-help interventions have diversified, including internet-based platforms, mobile messages, apps, and printed materials [24]. Among these, internet-based formats have gained significant attention for their wide reach and accessibility. The global mHealth market hit a valuation of US $85.2 billion worldwide in 2024, and is expected to grow to US $284.83 billion by 2033 [25], driving the advancement of internet-delivered self-help interventions. These models fall into 2 categories: fully self-directed (providing automated standardized feedback and monitoring) [26,27] and low-intensity therapist contact (involving web-based encouragement and guidance) [28,29]. They boast advantages like low cost, privacy protection, easy accessibility, and convenience [30]. Internet-based intervention has been recognized as one of the most effective tools for IBS self-help [31].

However, the effectiveness of internet-based self-help interventions for IBS, particularly in improving symptoms and mental health, remains insufficiently assessed. Existing meta-analyses on self-help interventions include both face-to-face and internet formats, with few studies and limited reliability [32]. Though recent studies have increased, debates persist regarding impacts on symptom severity [28,33], quality of life [34,35], depression [29,34], and anxiety [34,36]. Additionally, small sample sizes [26,27] and short durations (eg, 3 wk) [34] restrict generalizability. To date, the overall effectiveness of internet-based self-help interventions on IBS symptoms, psychological status, and quality of life remains unquantified. Therefore, this systematic review aims to quantitatively assess the clinical outcomes of internet-based self-help interventions compared with control groups (active control, usual care, or waitlist) in improving symptoms, psychological status, and quality of life in individuals with IBS, and to explore the effects of guidance format, intervention content, and duration on efficacy, providing evidence-based support for clinical practice.

Methods

Overview

This protocol for systematic review and meta-analysis adheres to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (). The protocol was registered subsequent to determining methodological details via existing literature evaluation, and prior to the initiation of

XSL•FO
RenderX

data extraction. During the later stages of data analysis, we made 3 methodological adjustments to the preregistered protocol to enhance methodological rigor and clinical interpretability. The language restrictions in the search strategy were removed to expand the evidence pool; the effect model selection approach was revised based on a priori conceptual hypotheses of between-study heterogeneity instead of I^2 statistics and Q-test *P* values; and the GRADE (Grading of Recommendations Assessment, Development and Evaluation) framework was incorporated for systematic evidence quality assessment. All adjustments only optimized the analysis method and did not change the core content of the original plan.

Search Strategy

This systematic review conducted the literature search in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses–search extension) guidelines [37]. A comprehensive search was performed across the following databases from their respective inception dates to June 25, 2025: PubMed (National Center for Biotechnology Information), Embase (Embase.com), Web of Science (Clarivate Analytics), Cochrane Library (Wiley), PsycINFO (Ovid), CINAHL (EBSCOhost), WANGFANG (Wanfang Data), CNKI (China National Knowledge Infrastructure), VIP (VIP Information), and CBM (China Biology Medicine Disc; CBMdisc Platform). A supplementary search was then implemented on March 9, 2026, after optimizing the search strategy. Notably, no multiple databases were accessed via the same platform during the search process. The search was executed without language or other restrictions, and no relevant filters were applied. No intentional search was conducted on trial registries, online resources, or print publications; nor were any alternative information sources or search methods used. The search strategy for this systematic review was developed based on approaches reported in prior studies [3,38], though it was not subjected to peer review. To identify additional potential studies, the reference lists of previous meta-analyses and included articles were manually screened. We will contact corresponding authors, domain experts, and manufacturers to obtain supplementary relevant information. Detailed search strategies for each database are presented in Multimedia Appendix 1.

Inclusion and Exclusion Criteria

Inclusion criteria were as follows: (1) participants aged older than 16 years with IBS diagnosed by ROME criteria or professional medical personnel; (2) internet-delivered self-help interventions, including pure self-help interventions without therapist contact and guided interventions; (3) including active controls (placebo or other treatments) and inactive controls (routine treatment, waiting list); (4) outcomes include IBS symptom severity, quality of life, visceral sensitivity, anxiety, and depression, with IBS symptom severity and quality of life as primary outcomes, and others as secondary; (5) only include peer-reviewed randomized controlled trials (RCTs); and (6) no restrictions on language or publication year.

Exclusion criteria were as follows: (1) unavailable full texts; (2) reviews, conference abstracts, study protocols, and non-peer-reviewed studies; and (3) duplicate data or data that cannot be converted or obtained.

Study Selection and Data Extraction

Initially, EndNote's duplicate detection tool was used to identify overlapping references, which were subsequently removed following manual verification. Subsequently, 2 investigators (XX and TL) independently screened the titles and abstracts of the remaining references against the predefined inclusion and exclusion criteria, followed by full-text assessments to determine study eligibility. Any discrepancies arising during the screening or assessment process were resolved through discussion with a third investigator (FW). Data extraction will be double-checked, encompassing the following information: first author and year of publication; sociodemographic characteristics of the included studies (country, age, proportion of female participants, and sample size); diagnostic criteria for IBS; study location; intervention details (intervention content, intervention duration, and follow-up duration); control group conditions; and outcome measures.

Methodological Quality Assessment

Two reviewers (XX and TL) assessed methodological quality using the Cochrane Risk of Bias 2.0, with discrepancies resolved by a third reviewer. The tool evaluates 5 key biases: randomization process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported result, each rated as "low risk," "high risk," or "some concerns."

Statistical Analysis

This systematic review pooled effect sizes using the standardized mean difference (SMD) based on the mean values and SDs of outcome measures in the intervention and control groups, in accordance with methods recommended in the Cochrane Handbook. Given the inconsistent outcome measurement tools across studies, SMD with 95% CIs was used as the effect size metric. Hedges *g*, a small-sample bias-corrected SMD, was the actual effect size calculated in this study, with interpretation criteria analogous to Cohen *d* (0.2=small effect, 0.5=moderate effect, 0.8=large effect) [39]. Results of this systematic review were interpreted as follows: inverse scores indicated improvements in IBS symptom severity, visceral sensitivity, and psychological status; positive scores reflected better quality of life.

Quantitative analyses for this systematic review were performed using R (version 4.5.0; R Core Team) software in the RStudio environment, with core analyses conducted using the following R packages: tidyverse, meta, and dmetar. Meta-analyses were carried out in accordance with the Hartung-Knapp-Sidik-Jonkman method recommended by the Cochrane Collaboration. Compared with the standard DerSimonian-Laird method [40], the Hartung-Knapp-Sidik-Jonkman method was preferred for its ability to estimate the weighted pooled mean with higher precision while reducing the false positive rate. Model selection was based on theoretical assumptions regarding the distribution of true effects, rather than statistical considerations [41,42]. Given the variability in participant characteristics and study

XSL•FO
RenderX

protocols across included studies, a random-effects model was used for all analyses. Heterogeneity in this systematic review was quantified using the Q-test, I^2 statistic, and tau-squared (τ^2). The I^2 statistic serves only as a relative quantitative measure of heterogeneity and does not directly reflect the magnitude of variability in true effects across different populations or settings, thus having limited clinical utility [43]. Thus, this study also reports the prediction interval (PI) to quantify the real-world implications of heterogeneity, providing a more clinically relevant range of effects for clinical practice [43].

To ensure the statistical power of subgroup analyses and meta-regression, only results with significant heterogeneity from meta-analyses comprising 10 or more studies were analyzed [44,45]. Subgroup analyses of postintervention data were performed based on prespecified intervention criteria to assess differences in effect sizes across subgroups. Categorical variables included the mode of health care delivery, control group type, intervention duration, and measuring tool. Meta-regression analyses were conducted for the primary outcome, with age and proportion of female participants as moderator variables. Meta-regression analyses were based on the random-effects framework to account for residual heterogeneity not explained by moderators. Subgroup analyses and meta-regression are inherently observational analytical approaches. Even when leveraging data from RCTs, they cannot establish causal relationships; instead, their core purpose is to explore potential moderator variables. The final reported outcomes in this review were fully consistent with the prespecified outcomes. In addition, sensitivity analyses were performed by sequentially excluding each individual study and recomputing the pooled effect size to assess the robustness of the results, thereby identifying studies with a substantial impact on the pooled effect. Visual inspection via funnel plots and Egger's linear regression test was used to evaluate small-study effects [46]. Per methodological guidelines, reliable interpretation of funnel plot asymmetry requires at least 10 included studies; otherwise, the reliability of inferences is limited [47]. Meanwhile, when the number of studies is small, the statistical power of Egger's test is significantly reduced, which may increase the risk of false-negative results [48]. If it is confirmed that there is a small research effect, then the trim-and-fill method will be used to correct the combined effect size. Statistical significance was set at $P<.050$.

Evaluation of Evidence Quality

To assess the certainty of evidence, we performed a GRADE evidence quality assessment for the outcomes (ie, IBS symptom severity, quality of life, visceral sensitivity, depressive symptoms, and anxiety) based on the findings of this systematic review, in accordance with guidelines developed by the GRADE Working Group. Two reviewers (XX and TL) independently evaluated the risks of bias, inconsistency, indirectness, imprecision, and publication bias. Any discrepancies were resolved through discussion with a third reviewer (FW) to reach a consensus.

Results

Search Results

The study selection process is shown in Figure 1. Initial database searches yielded 2813 studies. After removing 782 duplicates, 2031 remained. Title and abstract screening based on inclusion criteria led to full-text evaluation of 1874 studies, with 140 excluded. Ultimately, 17 studies were incorporated into the meta-analysis [26-30,33-36,49-56].

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the literature search and study selection.

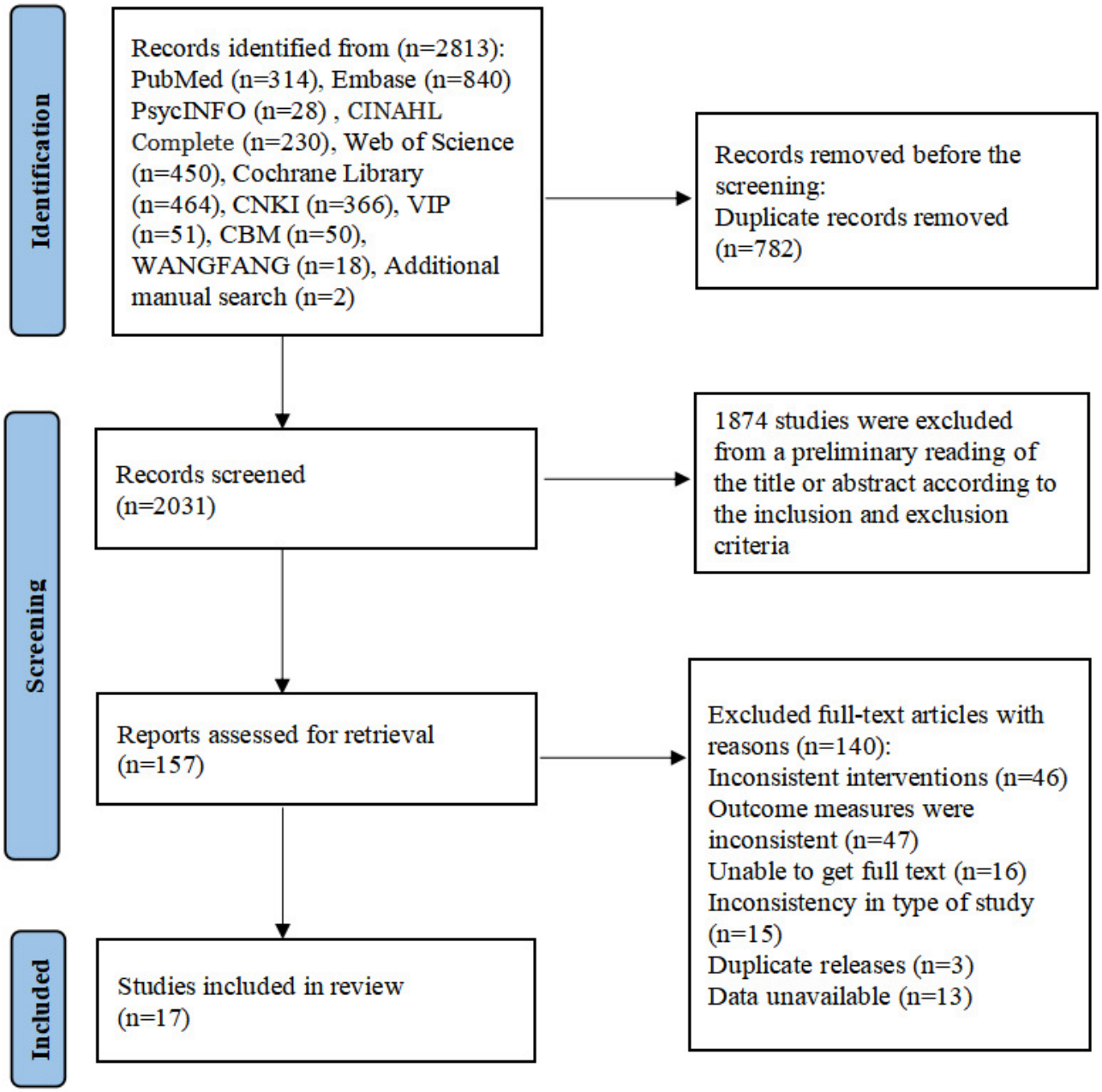

Characteristics of Included Studies

The basic characteristics of the included studies are summarized in Multimedia Appendix 2 [26-30,33-36,49-56]. A total of 2289 participants were enrolled, with a mean age ranging from 18.53 to 45.4 years and a female predominance (56.7% to 100%). For the diagnosis of IBS, the criteria adopted were primarily Rome III (n=10) and Rome IV (n=6), with 1 study using the Rome II criteria. The majority of studies were conducted in North America (n=7), followed by Europe (n=6), the Asia-Pacific region (n=3), and the Middle East (n=1). More than half of the included studies (n=9) were single-center trials, while 8 were multicenter trials.

All interventions were delivered via the Internet. The predominant intervention was CBT (n=11); the remaining interventions included yoga (n=2), progressive muscle relaxation (n=1), eHealth self-management (n=1), gut-directed hypnotherapy combined with CBT components (n=1), and online education (n=1). Most studies (n=15) adopted a guided delivery mode. For the control groups, 9 were inactive controls (usual care or waitlist), whereas the remaining 8 were active controls encompassing placebo, written expression, face-to-face education, psychoeducation, active control app, Internet-delivered stress management, CBT without exposure, and probiotics alone. The intervention lasted 3 to 10 weeks. Most studies (n=12) included a follow-up period, but only 8 provided complete follow-up data.

Quality Assessment

Risk of bias assessments are presented in Figure 2. Among these RCTs, 12 studies (70.6%) were judged to have some concerns, while 5 studies (29.4%) were assessed as having a high risk of bias.

XSL•FO
RenderX

Figure 2. ROB 2.0 traffic light plot of included randomized controlled trials [26-30,33-36,49-56].

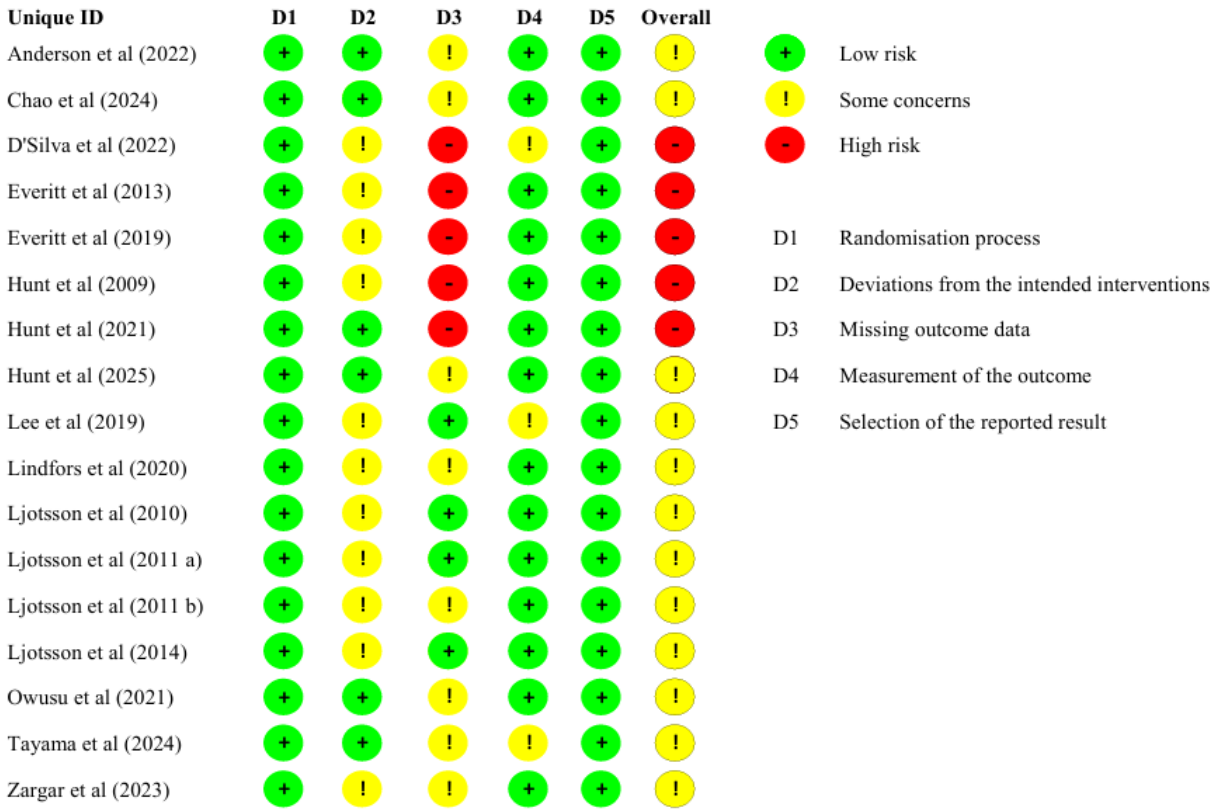

Effectiveness of Internet-Based Self-Help Interventions

Severity of IBS Symptoms

16 studies assessed IBS symptom severity, of which 8 provided follow-up data. Compared with control groups, internet-based self-help interventions were associated with a moderate, statistically significant reduction in postintervention IBS symptom severity (SMD −0.52, 95% CI −0.78 to −0.26, 95% PI −1.46 to 0.42; I^2=78.1%, τ^2=0.1793, *P*<.001) (Figure 3). Note that the 95% CI reflects the statistical certainty around this pooled average effect. While the wide 95% PI, which spanned both clinically meaningful benefit and potential harm, reflects substantial variability across populations and clinical settings, indicating both meaningful benefit and no/little effect in future studies. Given the high heterogeneity of this outcome and the low certainty of evidence according to the GRADE assessment (Table 1), the generalizability of this average effect should be viewed with caution. In addition, no statistically significant effect was observed at follow-up (SMD −0.35, 95% CI −0.80 to 0.10, 95% PI −1.52 to 0.82; I^2=77.3%, τ^2=0.2115, *P*<.001) (Figure S1 in Multimedia Appendix 3 [26-30,33-36,49-56]).

XSL•FO
RenderX

Figure 3. Forest plot of the effect on the severity of irritable bowel syndrome symptoms after internet-based self-help interventions [26-30,33-36,49-54,56].

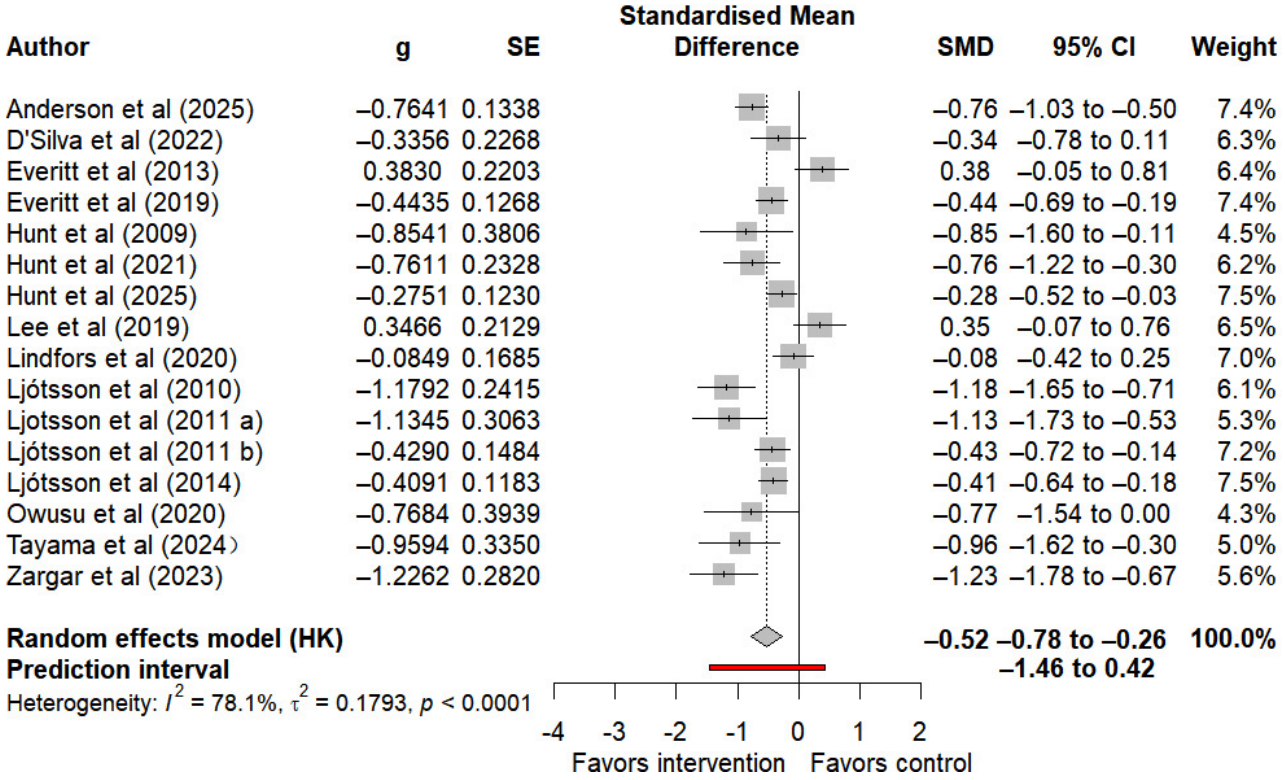

XSL•FO
RenderX

Table . GRADE (Grading of Recommendations Assessment, Development and Evaluation) evidence quality assessment.

Outcome indicators	Certainty assessment							Number of patients		Effect, absolute (95% CI)	Certainty	Importance
	Number of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Intervention	Comparison			
Severity of IBS[a] symptoms	16	Randomized trials	Serious[b]	Serious[c]	Not serious	Not serious	None	1006	1002	SMD[d] 0.52 SD lower (95% CI −0.78 to −0.26)	Low[b,c]	CRITICAL
Quality of life	13	Randomized trials	Serious[b]	Serious[c]	Not serious	Not serious	Publication bias strongly suspected[e]	793	804	SMD 0.58 SD higher (95% CI 0.23 to 0.90)	Very low[e]	CRITICAL
Visceral sensitivity	8	Randomized trials	Not serious	Serious[c]	Not serious	Not serious	Publication bias strongly suspected[e]	572	580	SMD 0.55 SD lower (95% CI −0.89 to −0.21)	Low[e]	CRITICAL
Depressive symptoms	9	Randomized trials	Serious[b]	Not serious	Not serious	Not serious	None	665	680	SMD 0.14 SD lower (95% CI −0.27 to −0.01)	Moderate[b]	IMPORTANT
Anxiety	7	Randomized trials	Not serious	Not serious	Not serious	Serious[f]	Publication bias strongly suspected[e]	560	552	SMD 0.03 SD lower (95% CI −0.34 to 0.27)	Low[f]	IMPORTANT

[a]IBS: irritable bowel syndrome.

[b]Most studies did not explicitly report blinding of participants, researchers, and outcome assessors, and the majority of included trials were judged to have unclear or high risk of bias. All outcome measures relied on subjective scales, which may have overestimated treatment effects. Consequently, the certainty of the evidence was downgraded by one level due to concerns about the risk of bias.

[c]A wider prediction interval indicates a larger potential variability in effect values, encompassing both clinically significant benefits and the possibility of no effect or even potential harm. Significant heterogeneity between studies reflects notable differences in effect estimates across individual studies. Therefore, based on the assessment criteria for inconsistency of evidence, this systematic review has downgraded the certainty of the evidence by one level.

[d]SMD: standardized mean difference.

[e]The Egger intercept for quality of life was significant (P=.02), using Duval and Tweedie's trim and fill method. The results added 4 studies, but the effect of the intervention still did not change.

[f]The 95% CI is wide and crosses the line of no effect, suggesting uncertainty in the true effect estimate. Therefore, the certainty of evidence was downgraded by one level for imprecision.

We explored sources of heterogeneity in postintervention outcomes using subgroup analyses and meta-regression. Subgroup analysis findings (Table 2 and Figures S2-S6 in Multimedia Appendix 3 [26-30,33-36,49-56]) showed that between-group differences for control group type (P=.78), intervention content (P=.49), intervention duration (P=.30), and the type of measuring tool (P=.051) were not statistically significant, suggesting these factors were not primary sources of heterogeneity. In contrast, health care contact (P=.04) emerged as one of the key contributors to high heterogeneity. Meta-regression analyses were also conducted to examine the association between study characteristics (age and female ratio) and heterogeneity (Table 3; Figures S7 and S8 in Multimedia Appendix 3 [26-30,33-36,49-56]). However, neither age (P=.77) nor female ratio (P=.27) was statistically significant, indicating they were not major sources of heterogeneity.

XSL•FO
RenderX

Table . Subgroup analysis results of internet-based self-help interventions on irritable bowel syndrome (IBS) symptom severity and quality of life.

Outcome indicator and moderate variable	SMD[a] (95% CI)	I^2 (%)	*P* value
IBS symptom severity			
Control group type			.78
Active control	−0.48 (−0.82 to −0.15)	70.2	.003
Inactive control	−0.55 (−1.02 to −0.09)	83.4	<.001
Intervention content			.49
CBT[b]	−0.47 (−0.81 to −0.12)	78.7	<.001
Other	−0.64 (−1.21 to −0.07)	78	.001
Intervention duration			.30
≤6 weeks	−0.33 (−1.03 to 0.37)	88.9	<.001
>6 weeks	−0.63 (−0.87 to −0.39)	56.5	.01
Health care contact			.04
Guided	−0.49 (−0.78 to −0.20)	80.2	<.001
Self-help interventions	−0.76 (−0.81 to −0.72)	0	.99
Measuring tool			.051
IBS-SSS[c]	−0.30 (−0.71 to 0.12)	81.9	<.001
GSRS-IBS[d]	−0.73 (−1.06 to −0.41)	71.7	<.001
Quality of life			
Control group type			.14
Active control	0.38 (−0.13 to 0.89)	74.7	<.001
Inactive control	0.79 (0.31 to 1.26)	69.2	.006
Intervention content			.85
CBT	0.55 (0.19 to 0.92)	74.6	<.001
Other	0.63 (−0.37 to 1.62)	83.4	<.001
Intervention duration			.87
≤6 weeks	0.54 (−0.50 to 1.57)	84.3	<.001
>6 weeks	0.60 (0.26 to 0.94)	73	<.001

[a]SMD: standardized mean difference.
[b]CBT: cognitive behavioral therapy.
[c]IBS-SSS: Irritable Bowel Syndrome Symptom Severity Scale.
[d]GSRS-IBS: Gastrointestinal Symptom Rating Scale-Irritable Bowel Syndrome.

Table . Meta-regression results of internet-based self-help interventions on irritable bowel syndrome (IBS) symptom severity and quality of life.

Outcome indicator and co-variate	β coefficient (SE)	95% CI	*T* test	*P* value
IBS symptom severity				
Age (years)	0.01 (0.02)	−0.03 to 0.04	0.29	.77
Female ratio	0.01 (0.01)	−0.01 to 0.04	1.16	.27
Quality of life				
Age (years)	−0.04 (0.02)	−0.10 to 0.01	−1.85	.09
Female ratio	−0.00 (0.02)	−0.05 to 0.04	−0.15	.88

XSL•FO
RenderX

Quality of Life

13 studies assessed quality of life in individuals with IBS, of which 6 provided follow-up data. Compared with control groups, internet-based self-help interventions were associated with a moderate, statistically significant improvement in postintervention QoL (SMD 0.57, 95% CI 0.23 to 0.90, 95% PI −0.41 to 1.54; I^2=76.8%, τ^2=0.18, *P*<.001) (Figure 4). Here, the 95% CI denotes the statistical certainty of this average improvement effect, while the 95% PI spanned both beneficial and null effects, indicating substantial variability in intervention outcomes across different populations, clinical settings, and implementation strategies. Given the high heterogeneity of this outcome and the very low certainty of evidence according to GRADE (Multimedia Appendix 3 [26-30,33-36,49-56]), the credibility and applicability of this conclusion should be assessed with caution. Furthermore, no statistically significant effect was observed at follow-up (SMD 0.12, 95% CI −0.02 to 0.25, 95% PI −0.07 to 0.31; I^2=0%, τ^2=0, *P*=.80) (Figure S9 in Multimedia Appendix 3 [26-30,33-36,49-56]).

Figure 4. Forest plot of the impact of internet-based self-intervention on quality of life in irritable bowel syndrome [27-30,34-36,49,50,52,53,55,56].

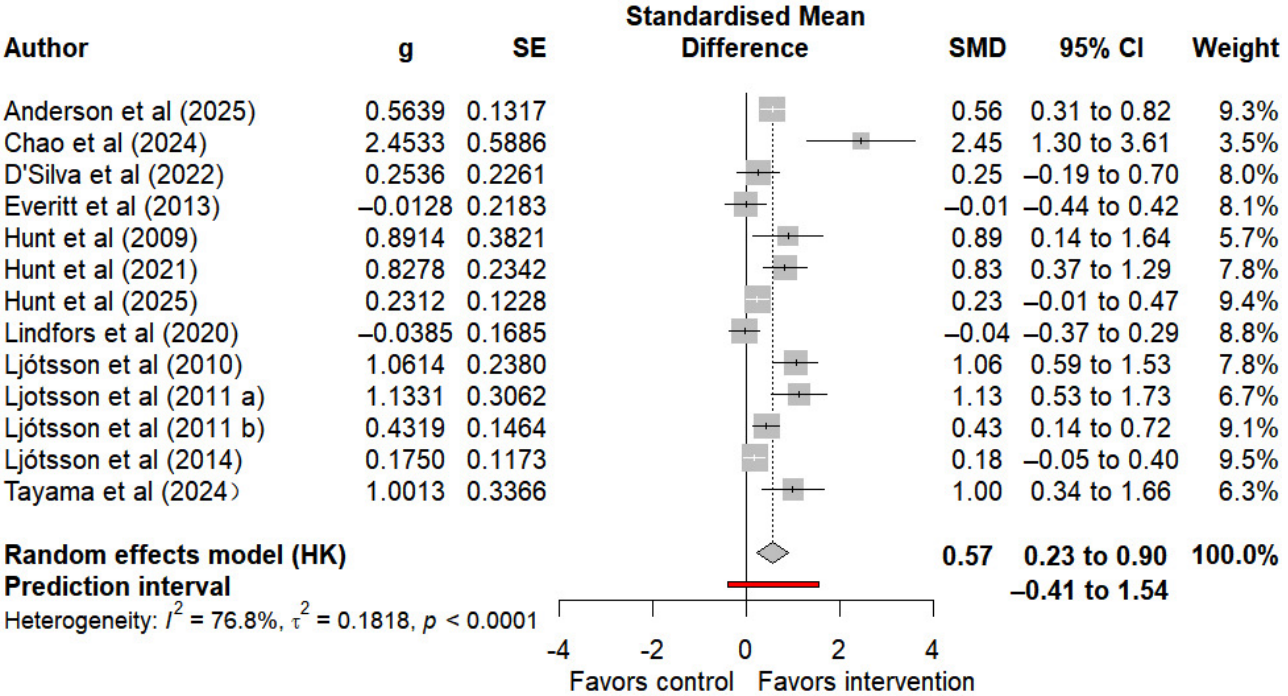

Subgroup analyses of postintervention data (Table 2 and Figures S10-S12 in Multimedia Appendix 3 [26-30,33-36,49-56]) revealed that the control group type (*P*=.14), intervention content (*P*=.85), and intervention duration (*P*=.87) were not the key influencing factors for high heterogeneity. Due to the limited number of subgroup studies, health care contact was not included as a moderating variable. Meta-regression analyses (Table 3 and Figures S13 and S14 in Multimedia Appendix 3 [26-30,33-36,49-56]) demonstrated that neither age (*P*=.09) nor the female ratio (*P*=.88) reached statistical significance, indicating they were not major sources of heterogeneity.

Visceral Sensitivity

The visceral sensitivity was used to assess gastrointestinal symptom-specific anxiety in individuals with IBS. A total of 8 studies demonstrated that internet-based self-help interventions yielded a statistically significant improvement in visceral sensitivity among these patients, with a moderate to large effect size (SMD −0.55, 95% CI −0.89 to −0.21, 95% PI −1.43 to 0.33; I^2=74.4%, τ^2=0.1198, *P*<.001) (Figure 5). Notably, the wide span of the 95% PI reflects the uncertainty surrounding the magnitude of intervention effects in future studies. The certainty of the GRADE evidence for this outcome is low, and there is high heterogeneity (Multimedia Appendix 3 [26-30,33-36,49-56]), suggesting that the effect of this intervention on the improvement of visceral sensitivity varies considerably across different studies, and the conclusions need to be interpreted in the context of specific scenarios.

XSL•FO
RenderX

Figure 5. Forest plot of the effect of internet-based self-intervention on visceral sensitivity in irritable bowel syndrome [29,30,34,35,49,52-54].

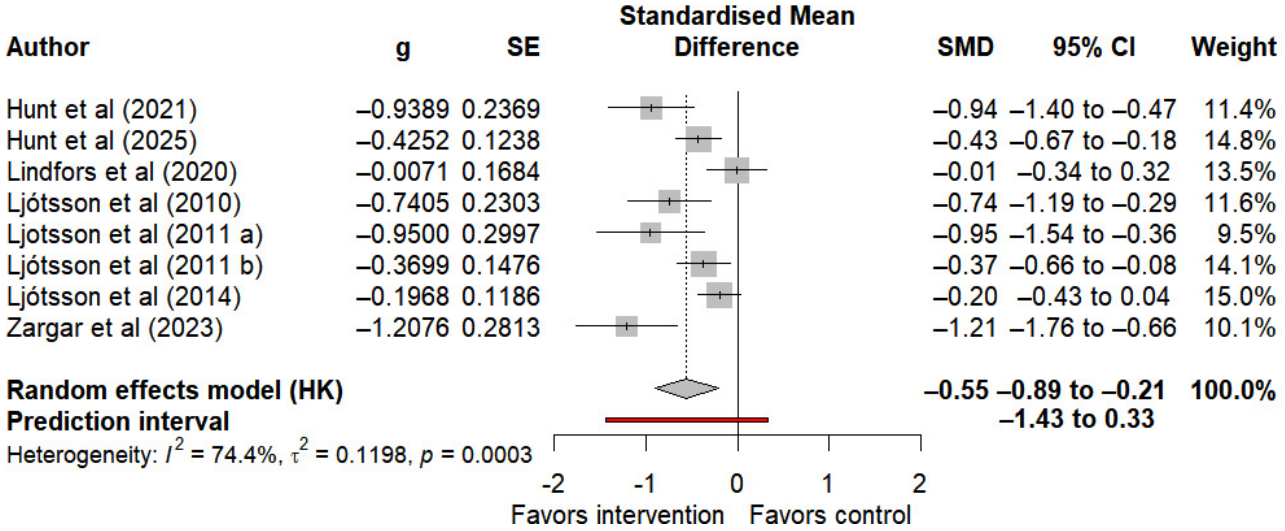

Depressive Symptoms

9 studies reported on depressive symptoms in individuals with IBS. Compared with control groups, internet-based self-help interventions were associated with statistically significant improvement in depressive symptoms (SMD −0.14, 95% CI −0.27 to −0.01, 95% PI −0.27 to −0.01; I²=0.9%, τ²<0.001, *P*=.43) (Figure 6). The certainty of the evidence according to GRADE is moderate (Multimedia Appendix 3 [26-30,33-36,49-56]). Currently, it is possible to confirm the impact of internet-based self-help interventions on depressive symptoms in individuals with IBS.

Figure 6. Forest plot of internet-based self-help interventions for depression in irritable bowel syndrome [29,30,34,36,49-53].

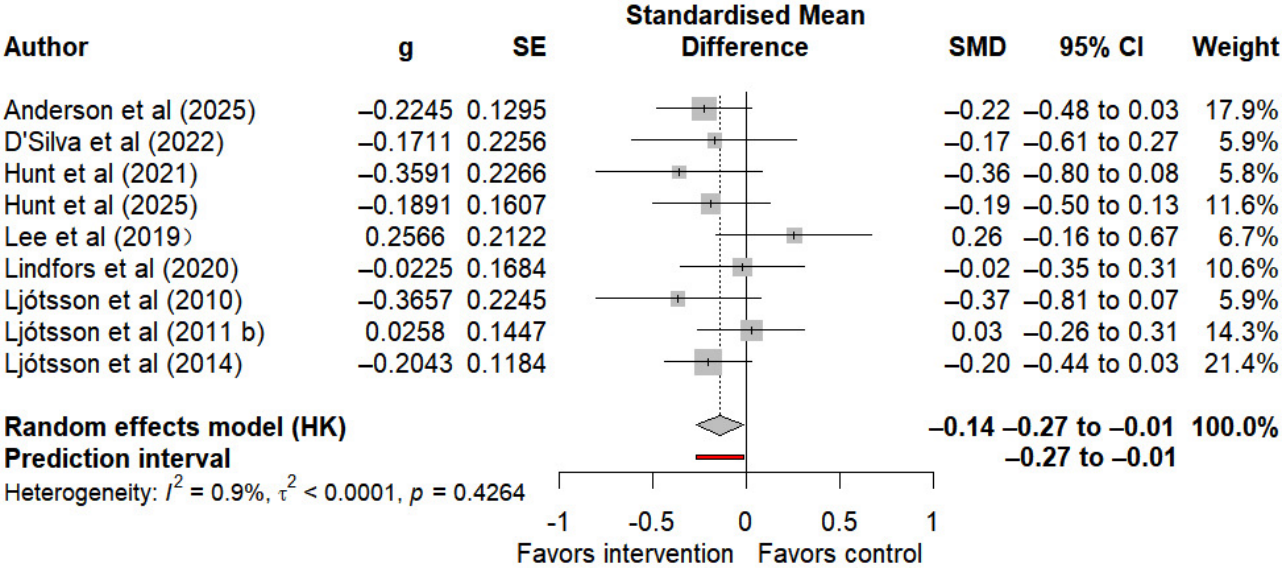

Anxiety

7 studies reported on anxiety symptoms in individuals with IBS. Meta-analysis results indicated that internet-based self-help interventions were not associated with a statistically significant effect on anxiety (SMD −0.03, 95% CI −0.34 to 0.27, 95% PI −0.78 to 0.71; I²=73.2%, τ²=0.08, *P*=.001) (Figure 7). Considering the low certainty of evidence for this outcome (Multimedia Appendix 3 [26-30,33-36,49-56]), current evidence is insufficient to support the effectiveness of internet-based self-help interventions in improving anxiety symptoms in individuals with IBS.

XSL•FO
RenderX

Figure 7. Forest plot of internet-based self-help interventions for anxiety in irritable bowel syndrome [28,34,36,49-52].

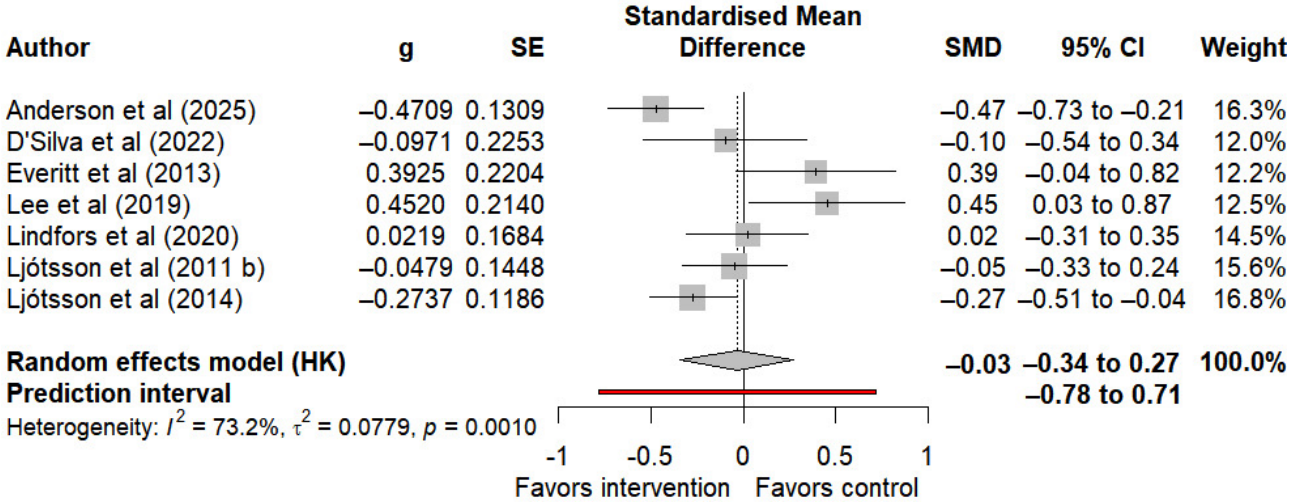

Sensitivity Analysis

A leave-one-out sensitivity analysis was conducted on postintervention results for all outcome measures to verify the robustness of pooled estimates across different outcomes after excluding each individual study. No substantial changes were observed in the pooled effect sizes or their corresponding directions in this systematic review following the exclusion of any single study, thus confirming the good overall robustness of the results (Figures S15-S19 in Multimedia Appendix 3 [26-30,33-36,49-56]).

Small-Study Effects

Small-study effects were assessed using funnel plots and Egger's test. Funnel plots were generated for the outcomes with more than 10 included studies, namely IBS symptom severity (Figure S20 in Multimedia Appendix 3 [26-30,33-36,49-56]) and quality of life (Figure S21 in Multimedia Appendix 3 [26-30,33-36,49-56]). These funnel plots exhibited asymmetry, which may indicate potential publication bias, as well as reflect heterogeneity and random chance across studies. Egger's regression intercept was applied to all outcomes to evaluate bias. The nonsignificant Egger's intercepts for IBS symptom severity (P=.27) and depressive symptoms (P=.90) suggested the absence of small-study effects. However, Egger's intercept for quality of life (P=.01), visceral sensitivity (P=.02), and anxiety (P=.03) was statistically significant, prompting the application of the Duval and Tweedie trim-and-fill method. For quality of life, 4 additional studies were included (SMD 0.33, 95% CI −0.08 to 0.75); while Visceral Sensitivity added 3 studies (SMD −0.32, 95% CI −0.71 to 0.06). The results of the meta-analysis indicated that the intervention effects of these 2 outcomes significantly changed after the supplementary studies were added, suggesting that the robustness of the relevant evidence was poor. When interpreting the conclusions, caution is required. For anxiety, after adding 2 supplementary studies, the combined effect did not significantly change (SMD −0.20, 95% CI −0.55 to 0.16), suggesting that the influence of small studies on the combined results of this outcome was relatively small.

GRADE Assessment

The GRADE evidence certainty assessment indicated that most outcomes were rated as having very low to moderate certainty (Table 1). Most studies did not explicitly report blinding procedures for participants, interventionists, and outcome assessors, and all outcomes were subjective, undermining the reliability, so we downgraded the evidence certainty for severity of IBS symptoms, quality of life, and depressive symptoms by one level due to risk of bias. Furthermore, substantial heterogeneity in IBS symptom severity, quality of life, and visceral sensitivity contributed to variability in effect estimates. Additionally, the 95% CIs and PIs were wide, encompassing not only clinically meaningful benefits but also no effect or potential harm. We therefore further downgraded evidence certainty for IBS symptom severity, quality of life, and visceral sensitivity by one level due to inconsistency. For quality of life, visceral sensitivity, and anxiety outcomes, Egger's test confirmed small-study effects. Although the trim-and-fill adjustment did not alter the intervention effect, it still indicated potential bias, prompting an additional one-level downgrade for this outcome. For anxiety outcomes, the 95% CIs were wide and crossed the null effect line, signaling uncertainty in estimating the true effect. We thus downgraded evidence certainty by one level due to imprecision. Overall, GRADE evidence certainty findings substantially impact the credibility of the pooled effect estimates.

Discussion

Principal Findings

This study extends beyond prior similar reviews focused on internet-delivered CBT by expanding the scope of intervention modalities evaluated and quantifying the effects of key implementation factors, including intervention duration, guidance mode, and core intervention content, on treatment efficacy, thereby providing more actionable evidence for clinical practice [3]. Quantitative synthesis revealed that internet-based self-help interventions led to significant postintervention improvements in IBS symptom severity, quality of life, visceral sensitivity, and depressive symptoms for individuals with IBS, with small-to-moderate effect sizes. These short-term benefits,

XSL•FO
RenderX

however, were not sustained at follow-up, reflecting the limited durability of the intervention's long-term effects. Notably, no significant improvements in anxiety symptoms were observed with these interventions, and high heterogeneity was present across all primary outcomes. Given that GRADE assessments rated the certainty of evidence as very low to moderate, these findings must be interpreted cautiously in the context of the study's methodological limitations.

Symptom Severity

IBS symptom severity serves as a primary outcome measure for evaluating the efficacy of internet-based self-help interventions. Quantitative findings demonstrated that these interventions yielded significant short-term improvements in IBS symptom severity with superior efficacy relative to control conditions, which is consistent with the conclusions of previous similar studies and further validates the value of internet-based self-help interventions in the symptom management of IBS [3]. However, the 95% PIs had a wide range, indicating that intervention effects may exhibit substantial variability across different clinical settings. Individuals with IBS who are proficient in internet use and have high health care accessibility can better complete structured intervention modules and achieve more pronounced symptom improvements; whereas those in remote areas or with low digital health literacy may experience reduced adherence due to operational difficulties and lack of guidance, yielding minimal or even no significant benefits from the interventions.

Subgroup analyses further indicate that variations in health care contact constitute a key source of high heterogeneity. We found that guided self-help interventions may yield greater benefits, which is consistent with the findings of a meta-analysis indicating that increased therapist contact can enhance the efficacy of self-help interventions [57]. This may be attributed to the fact that low-intensity guidance from health care professionals enhances intrinsic motivation and treatment adherence among individuals with IBS and fosters positive therapeutic relationships, thereby sustaining the improvements in their quality of life [28,58]. Different scales have fundamental differences in the construction definition and evaluation dimensions of symptom severity, which may have certain impacts on the presentation of effect values. However, the subgroup analysis results show that this scale difference is not the main source of heterogeneity in this study. Specifically, studies using the Gastrointestinal Symptom Rating Scale-Irritable Bowel Syndrome as the assessment tool reported larger effect sizes. This scale is a patient-reported outcome measure that focuses on the daily fluctuations of core IBS symptoms and can more sensitively capture subtle improvements in individuals' subjective perceptions [59]. In contrast, the effect value of the Irritable Bowel Syndrome-Symptom Severity Scale is relatively small and lacks statistical significance. As an assessment tool commonly used in clinical practice, the Irritable Bowel Syndrome-Symptom Severity Scale evaluation criteria place more emphasis on objective symptom indicators and may, to some extent, overlook the changes in individual subjective perception [60,61]. Meta-regression analyses demonstrated that neither mean age nor the proportion of female individuals exerted a significant effect on effect sizes, ruling out the possibility that demographic characteristics constitute a primary source of heterogeneity.

Due to the risk of bias, high heterogeneity stemming from health care contact variability, and wide 95% PIs, the GRADE certainty of evidence for this outcome was rated as low, which was downgraded for risk of bias and insufficient consistency. This low certainty of evidence does not negate the inherent value of these interventions, but rather reflects a gap between the current evidence and real-world practice. This indicates that while internet-based self-help interventions have a well-established short-term beneficial effect on IBS symptom severity, caution is warranted when generalizing this conclusion. In real-world implementation, a comprehensive consideration of factors including intervention plan, digital health literacy among individuals with IBS, and local health care service capacity is required to maximize intervention efficacy and enhance the credibility of evidence.

Quality of Life

Compared with the general population, individuals with IBS experience a significant impairment in their quality of life [62]. Thus, improving quality of life is one of the core objectives of self-management for IBS. Meta-analysis findings demonstrated that internet-based self-help interventions produced a moderate short-term improvement in the quality of life among individuals with IBS, which is consistent with the conclusions of a systematic review focusing on digital health self-management for IBS [63], providing more robust evidence to support improvements in quality of life among individuals with IBS. However, wide 95% PIs indicate that these beneficial effects exhibit substantial variability across different clinical settings, which necessitates a comprehensive evaluation based on specific implementation conditions.

Given the high heterogeneity of the results, subgroup analyses were conducted based on potential moderating variables such as the type of control group, intervention period, and intervention content. It was found that none of these factors was the main source of heterogeneity. This result indicates that although there are differences in the types of controls, intervention periods, and intervention contents among different studies, these differences did not significantly affect the overall heterogeneity among the studies. The subgroup results showed that, compared with the active control group, the effect size in the inactive control group was higher. Inactive control groups lack targeted intervention resources, resulting in a larger efficacy gap relative to internet-based self-help interventions and thus making the value of these interventions more prominent. Second, interventions with a duration of more than 6 weeks yielded superior improvements in quality of life, which aligns with the conclusion by Liegl et al [32] that self-help interventions require sufficient time to consolidate behavioral changes. Melchior et al [64] noted that quality of life in individuals with IBS is jointly influenced by multiple factors, including gastrointestinal symptoms, psychological distress, and coping strategies, and improvements in these factors depend on the long-term consolidation of health behaviors such as dietary modifications and stress management. Extending the intervention duration enables individuals with IBS to engage in repeated practice of

XSL•FO
RenderX

relaxation training and trigger identification skills within online modules, which facilitates the gradual reduction of excessive attention to gastrointestinal symptoms and thus results in a sustained improvement in quality of life. The subgroup analysis results of this study confirmed the aforementioned theoretical expectations.

However, the subgroup interaction test results of this study showed that the differences in effect sizes of these factors did not reach a statistically significant level, and thus were not identified as the main source of heterogeneity. This might be due to sample size limitations or other unidentified confounding factors resulting in insufficient statistical test power, preventing the detection of the significant contributions of these factors to heterogeneity. This suggests that the improvement in the quality of life of IBS patients may be more dependent on the core attributes of internet self-help interventions, such as convenience, autonomy, and targeted content provision, rather than the differences in specific implementation conditions, or the influence of these implementation conditions may be masked by other unidentified factors. Meta-regression analyses demonstrated that mean age and the proportion of female individuals were not a source of heterogeneity.

Due to the presence of bias risks in most studies, high unexplained heterogeneity, and a wide 95% CI range, the GRADE certainty of the evidence was rated as very low; moreover, the existence of small sample effects further downgraded the quality of the evidence, making the effect estimates highly susceptible to potential biases.

Based on these findings, in future research, when designing Internet-based self-help intervention programs for IBS, one can flexibly choose the core content of the intervention, reasonably plan the intervention period, and set the type of control according to the research purpose. More importantly, it is necessary to focus on exploring other potential sources of heterogeneity while increasing the sample size to reduce the bias of small samples, and further optimize the targeted and stable nature of the intervention program, providing more solid evidence to support the improvement of the long-term quality of life of IBS patients.

Visceral Sensitivity

Visceral sensitivity centers on gastrointestinal symptom-specific anxiety and denotes the cognitive, emotional, and behavioral responses triggered by fear of gastrointestinal symptoms. It is a key factor in the disease progression of IBS and also a well-established therapeutic target [57]. This systematic review found that internet-based self-help interventions exert a large short-term effect on visceral hypersensitivity, which is consistent with the biopsychosocial model of brain-gut axis dysfunction in IBS [58]. As Wolitzky-Taylor et al [65] note, anxiety toward visceral sensations in individuals with IBS triggers avoidant behaviors and hypervigilance. In contrast, through structured learning within internet-based self-help interventions, these individuals can adopt strategies including relaxation training and exposure exercises to reduce catastrophic symptom-related thinking, alleviate gastrointestinal hypersensitivity, and food-related fear, and thus achieve symptomatic improvements [30]. Furthermore, as a mediating variable, visceral hypersensitivity facilitates the indirect improvement of core IBS symptoms and quality of life among individuals with IBS [30,65].

It is important to note that wide 95% PIs indicate that this effect may exhibit substantial variability in future research, which warrants cautious interpretation in the context of clinical practice. The GRADE certainty of evidence for this outcome was rated as low. Its rating was downgraded because the high heterogeneity led to inconsistent effect estimation results and the presence of small study effects, which indicates that the significant short-term effect of this outcome still requires validation from additional homogeneous studies.

Psychological Distress

Psychological distress, including anxiety and depression, constitutes a common symptom cluster among individuals with IBS [66]. Psychosocial moderators in IBS are widely emphasized, as psychological factors influence IBS symptoms, modify illness-related behaviors, and contribute to adverse health outcomes [67]. However, findings from this systematic review in the domain of psychological symptoms show differences from those of previous research [3,68], whereby internet-based self-help interventions statistically improve depression, but no significant effect has been observed in the improvement of anxiety.

Findings from the meta-analysis showed a slight but statistically significant improvement in the depressive symptoms of patients with IBS. The 95% PI was consistent with the CI, not crossing the zero-effect line, suggesting that the improvement effect of this intervention on depression was stable and could show consistent minor benefits in different research scenarios. At the same time, the heterogeneity of this outcome was extremely low, further confirming the consistency of the results among studies. This conclusion is different from the previous research conclusion that such interventions had no significant improvement effect on depression. The occurrence of this positive result may be due to the targeted improvement of IBS physical symptoms by the intervention, which led to indirect psychological benefits. Although the studies included were mainly aimed at improving core physical symptoms such as abdominal pain and bloating, the relief of physical symptoms effectively reduced the patients' sense of disease burden caused by chronic diseases, thereby alleviating the associated depressive emotions [69]. In addition, most of the included studies provided psychological intervention modules, integrating relevant content of CBT, such as stress management, cognitive regulation, etc. These standardized modules may still exert a slight antidepressant effect, and the extremely low research heterogeneity also indicates that this type of internet self-help intervention has stability in improving depression. The evidence GRADE rating for the depression improvement effect of this study is moderate. It was slightly downgraded due to the certain risk of bias in the included studies, but this confirms that the conclusion has high credibility.

On the contrary, the self-help intervention based on the Internet has not shown statistically significant improvement in the anxiety symptoms of patients with IBS. This result is consistent with the conclusions of previous studies, and its evidence has

XSL•FO
RenderX

a GRADE rating of low, indicating weak evidence strength. The difference in the intervention effect between depression and anxiety may stem from the fundamental differences in the pathogenesis and influencing factors of the two in the IBS population. The anxiety of patients with IBS is often directly related to visceral hypersensitivity and fear of symptom recurrence [70]. Such anxiety symptoms require more targeted and interactive professional psychological intervention, such as exposure therapy for symptom-related fears, one-on-one cognitive restructuring guidance, etc. However, the current low-intensity, noninteractive Internet self-help intervention model included in the study may not solve this specific type of anxiety problem. Additionally, there are differences in measurement tools, and the design characteristics of different questionnaires may directly affect the accuracy of results and effect estimation [71]. The sample size of the research on anxiety is limited, and the statistical test power is insufficient, which directly reduces the credibility of the conclusion. The results of the correlation analysis of anxiety symptoms have considerable inaccuracy, and the effect estimation is prone to being interfered with by potential biases, making it impossible to rule out the possibility of the existence of the real effect; at the same time, there is a small study effect for this outcome, and the certainty of the evidence is thus further reduced. Ultimately, it is rated as low.

For individuals with IBS, self-help interventions based on the Internet can not only improve the physical symptoms, but also exert a mild antidepressant effect. This is a cost-effective, comprehensive intervention method. However, such interventions have no clear effect on the improvement of anxiety symptoms. When clinicians apply Internet self-help interventions to alleviate the physical symptoms of individuals, they need to supplement more targeted psychological support measures for individuals with IBS who have co-occurring anxiety symptoms, so as to achieve comprehensive management of somatic and psychological symptoms in individuals with IBS.

Implications for Clinical Practice

From the perspective of clinical practice, in response to the socioeconomic burden imposed by IBS, it is imperative to both establish the clinical effectiveness of intervention strategies and develop innovative service models directly accessible to individuals with IBS, thereby enhancing their access to appropriate care. Internet-based self-help interventions leverage digital platforms to transcend temporal and spatial constraints, thereby facilitating autonomous access to intervention resources for individuals with IBS. Through web-based delivery modes, these individuals can independently acquire basic IBS-related knowledge and coping skills [23], like dietary guidance, relaxation training, and symptom management [56].

This intervention model preserves the structured nature of interventions, enhances the flexibility and accessibility of participation for individuals with IBS, and reduces health care costs, making it particularly suitable for individuals with IBS who seek frequent medical care or those living in regions with limited health care resources. Despite uncertainties regarding long-term efficacy and the low quality of available evidence, internet-based self-help interventions still exhibit certain clinical value for symptom management in individuals with IBS and may therefore be considered a component of comprehensive symptom management plans for IBS.

Limitations

This study has several limitations. First, in the literature search phase, we did not access relevant online resources or clinical trial registries and only included studies published in Chinese and English, potentially introducing selection and language bias. Second, regarding study design and outcome assessment, most included studies did not explicitly report blinding of participants, interventionists, and outcome assessors. Additionally, all outcome measures relied on self-reported scales, introducing subjective bias that may undermine the accurate assessment of the clinical efficacy of these internet-based self-help interventions.

Third, measurement tool heterogeneity was evident: the same outcome constructs were evaluated using different assessment scales. Although a methodologically sound SMD was applied for data pooling, different scales demonstrate inherent differences in the construct definitions, assessment dimensions, and scoring approaches of core outcomes. This can result in ambiguous construct validity, exacerbate measurement heterogeneity, and indirectly reduce the accuracy and interpretability of pooled effect sizes. Fourth, high heterogeneity existed across all included studies. Although we conducted subgroup analyses and meta-regression per prespecified criteria, unidentified potential moderators likely remain, leading to an incomplete characterization of heterogeneity sources. Additionally, some RCTs included for specific outcomes had small sample sizes, resulting in insufficient statistical power and an inability to accurately evaluate true intervention effects.

Given the aforementioned limitations, the findings of this study should be interpreted cautiously. Future research should conduct large-sample, multicenter RCTs and prioritize multidimensional outcome measures. Additionally, it is imperative to identify the core active components of standardized intervention protocols, including intervention duration, frequency, and guidance intensity, and harmonize assessment tools for core constructs. This will mitigate the impact of measurement heterogeneity on study outcomes and further validate the true efficacy of internet-based self-help interventions for individuals with IBS.

Conclusions

This meta-analysis provides a comprehensive assessment of the latest evidence for internet-based self-help interventions in the management of individuals with IBS. Findings demonstrate these interventions produce significant short-term, small-to-moderate effect size improvements in disease-specific symptom severity, quality of life, visceral sensitivity, and depressive symptoms among this population, thus providing multi-outcome evidence to support their use as an adjunct to conventional care. It also broadens the scope of applicable intervention modalities, includes visceral sensitivity as a core outcome measure, and quantifies how key implementation factors influence intervention effectiveness, ultimately building a more comprehensive evidence base for digital interventions in the care of individuals with IBS. In real-world clinical

XSL•FO
RenderX

practice, internet-based self-help interventions offer high accessibility, convenience, and cost-effectiveness, making them particularly suitable for individuals with IBS in resource-limited health care settings or those who seek frequent clinical care. They can act as a key component of comprehensive disease management plans and inform the development of intervention strategies for primary care and telemedicine platforms.

However, these findings are subject to substantial uncertainty. Notably, the short-term benefits of these interventions are not sustained at follow-up; no significant improvements in anxiety symptoms are observed among individuals with IBS, and high heterogeneity exists across primary outcomes. Additionally, GRADE evidence certainty was rated as very low to moderate for all outcomes; moderate risk of bias, alongside effect variability associated with wide 95% PIs, further weakens the credibility of the evidence base. This highlights the need for extreme caution when translating these findings into clinical practice. In clinical practice, the implementation of internet-based self-help interventions should consider individual factors (eg, digital health literacy) and contextual factors (eg, local health care service capacity) to optimize their clinical value in a targeted manner.

Acknowledgments

There was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

All data supporting the findings of this systematic review are included in the main text and its supplementary information. For further details or additional datasets, please contact the corresponding author.

Authors' Contributions

XX contributed to conceptualization, data curation, formal analysis, investigation, methodology, and writing of the original draft, as well as review and editing of the manuscript. TL contributed to formal analysis, investigation, methodology, visualization, and writing of the original draft, as well as review and editing. FW contributed to conceptualization, supervision, and review and editing of the manuscript. HH contributed to the methodology and writing of the original draft. JC and JL contributed to formal analysis and review and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File, 15 KB - jmir_v28i1e87216_app1.docx]

Multimedia Appendix 2
Table of basic characteristics of studies on internet-based self-help interventions for individuals with Irritable Bowel Syndrome.
[DOCX File, 56 KB - jmir_v28i1e87216_app2.docx]

Multimedia Appendix 3
Subgroup analysis of all results, meta-regression bubble plots, sensitivity analysis plots, and funnel plots.
[DOCX File, 7919 KB - jmir_v28i1e87216_app3.docx]

Checklist 1
PRISMA 2020 checklist.
[DOCX File, 274 KB - jmir_v28i1e87216_app4.docx]

References

1. Spiller R, Aziz Q, Creed F, et al. Guidelines on the irritable bowel syndrome: mechanisms and practical management. Gut 2007 Dec;56(12):1770-1798. [doi: 10.1136/gut.2007.119446] [Medline: 17488783]

XSL•FO
RenderX

2. Longstreth GF, Thompson WG, Chey WD, Houghton LA, Mearin F, Spiller RC. Functional bowel disorders. A multicenter comparison of health status and development of illness severity index. Gastroenterology 2006;130(5):1480-1491. [doi: 10.1007/BF02064187] [Medline: 7729289]
3. Kim H, Oh Y, Chang SJ. Internet-delivered cognitive behavioral therapy in patients with irritable bowel syndrome: systematic review and meta-analysis. J Med Internet Res 2022 Jun 10;24(6):e35260. [doi: 10.2196/35260] [Medline: 35687389]
4. North CS, Downs D, Clouse RE, et al. The presentation of irritable bowel syndrome in the context of somatization disorder. Clin Gastroenterol Hepatol 2004 Sep;2(9):787-795. [doi: 10.1016/s1542-3565(04)00350-7] [Medline: 15354279]
5. Zhang QE, Wang F, Qin G, et al. Depressive symptoms in patients with irritable bowel syndrome: a meta-analysis of comparative studies. Int J Biol Sci 2018;14(11):1504-1512. [doi: 10.7150/ijbs.25001] [Medline: 30263003]
6. Zamani M, Alizadeh-Tabari S, Zamani V. Systematic review with meta-analysis: the prevalence of anxiety and depression in patients with irritable bowel syndrome. Aliment Pharmacol Ther 2019 Jul;50(2):132-143. [doi: 10.1111/apt.15325] [Medline: 31157418]
7. Miller V, Hopkins L, Whorwell PJ. Suicidal ideation in patients with irritable bowel syndrome. Clin Gastroenterol Hepatol 2004 Dec;2(12):1064-1068. [doi: 10.1016/s1542-3565(04)00545-2] [Medline: 15625650]
8. Owens DM, Nelson DK, Talley NJ. The irritable bowel syndrome: long-term prognosis and the physician-patient interaction. Ann Intern Med 1995 Jan 15;122(2):107-112. [doi: 10.7326/0003-4819-122-2-199501150-00005] [Medline: 7992984]
9. AGA survey shows significant burden of IBS, need for improved patient-physician communication. Healio Gastroenterology. 2015. URL: https://www.healio.com/news/gastroenterology/20151216/aga-survey-shows-significant-burden-of-ibs-need-for-improved-patientphysician-communication [accessed 2026-01-28]
10. Hulisz D. The burden of illness of irritable bowel syndrome: current challenges and hope for the future. J Manag Care Pharm 2004;10(4):299-309. [doi: 10.18553/jmcp.2004.10.4.299] [Medline: 15298528]
11. Black CJ, Yiannakou Y, Houghton LA, Ford AC. Epidemiological, clinical, and psychological characteristics of individuals with self-reported irritable bowel syndrome based on the Rome IV vs Rome III criteria. Clin Gastroenterol Hepatol 2020 Feb;18(2):392-398. [doi: 10.1016/j.cgh.2019.05.037] [Medline: 31154027]
12. Polster A, Van Oudenhove L, Jones M, Öhman L, Törnblom H, Simrén M. Mixture model analysis identifies irritable bowel syndrome subgroups characterised by specific profiles of gastrointestinal, extraintestinal somatic and psychological symptoms. Aliment Pharmacol Ther 2017 Sep;46(5):529-539. [doi: 10.1111/apt.14207] [Medline: 28671338]
13. Whitehead WE, Palsson O, Jones KR. Systematic review of the comorbidity of irritable bowel syndrome with other disorders: what are the causes and implications? Gastroenterology 2002 Apr;122(4):1140-1156. [doi: 10.1053/gast.2002.32392] [Medline: 11910364]
14. Osterberg E, Blomquist L, Krakau I, Weinryb RM, Asberg M, Hultcrantz R. A population study on irritable bowel syndrome and mental health. Scand J Gastroenterol 2000 Mar;35(3):264-268. [doi: 10.1080/003655200750024128] [Medline: 10766319]
15. Jones M, Koloski N, Boyce P, Talley NJ. Pathways connecting cognitive behavioral therapy and change in bowel symptoms of IBS. J Psychosom Res 2011 Mar;70(3):278-285. [doi: 10.1016/j.jpsychores.2010.10.004] [Medline: 21334499]
16. Koloski NA, Jones M, Kalantar J, Weltman M, Zaguirre J, Talley NJ. The brain--gut pathway in functional gastrointestinal disorders is bidirectional: a 12-year prospective population-based study. Gut 2012 Sep;61(9):1284-1290. [doi: 10.1136/gutjnl-2011-300474] [Medline: 22234979]
17. Black CJ, Thakur ER, Houghton LA, Quigley EMM, Moayyedi P, Ford AC. Efficacy of psychological therapies for irritable bowel syndrome: systematic review and network meta-analysis. Gut 2020 Aug;69(8):1441-1451. [doi: 10.1136/gutjnl-2020-321191] [Medline: 32276950]
18. Hookway C, Buckner S, Crosland P, Longson D. Irritable bowel syndrome in adults in primary care: summary of updated NICE guidance. BMJ 2015 Feb 25;350:h701. [doi: 10.1136/bmj.h701] [Medline: 25716701]
19. Cuijpers P, van Straten A, Andersson G, van Oppen P. Psychotherapy for depression in adults: a meta-analysis of comparative outcome studies. J Consult Clin Psychol 2008 Dec;76(6):909-922. [doi: 10.1037/a0013075] [Medline: 19045960]
20. Berry K, Haddock G. The implementation of the NICE guidelines for schizophrenia: barriers to the implementation of psychological interventions and recommendations for the future. Psychol Psychother 2008 Dec;81(Pt 4):419-436. [doi: 10.1348/147608308X329540] [Medline: 18652741]
21. van Straten A, Cuijpers P, Smits N. Effectiveness of a web-based self-help intervention for symptoms of depression, anxiety, and stress: randomized controlled trial. J Med Internet Res 2008 Mar 25;10(1):e7. [doi: 10.2196/jmir.954] [Medline: 18364344]
22. Wicaksana AL, Apriliyasari RW, Tsai PS. Effect of self-help interventions on psychological, glycemic, and behavioral outcomes in patients with diabetes: a meta-analysis of randomized controlled trials. Int J Nurs Stud 2024 Jan;149:104626. [doi: 10.1016/j.ijnurstu.2023.104626] [Medline: 37979371]
23. Dorn SD. Systematic review: self-management support interventions for irritable bowel syndrome. Aliment Pharmacol Ther 2010 Aug;32(4):513-521. [doi: 10.1111/j.1365-2036.2010.04374.x] [Medline: 20497137]
24. Hartmann-Boyce J, Jebb SA, Fletcher BR, Aveyard P. Self-help for weight loss in overweight and obese adults: systematic review and meta-analysis. Am J Public Health 2015 Mar;105(3):e43-e57. [doi: 10.2105/AJPH.2014.302389] [Medline: 25602873]

XSL•FO
RenderX

25. mHealth market report. Market Data Forecast. 2025. URL: https://www.marketdataforecast.com/market-reports/mobile-health-market [accessed 2026-01-28]
26. Owusu JT, Sibelli A, Moss-Morris R, van Tilburg MAL, Levy RL, Oser M. A pilot feasibility study of an unguided, internet-delivered cognitive behavioral therapy program for irritable bowel syndrome. Neurogastroenterol Motil 2021 Nov;33(11):e14108. [doi: 10.1111/nmo.14108] [Medline: 33745228]
27. Tayama J, Hamaguchi T, Koizumi K, et al. Efficacy of an eHealth self-management program in reducing irritable bowel syndrome symptom severity: a randomized controlled trial. Sci Rep 2024 Jan 3;14(1):4. [doi: 10.1038/s41598-023-50293-z] [Medline: 38172498]
28. Everitt H, Moss-Morris R, Sibelli A, et al. Management of irritable bowel syndrome in primary care: the results of an exploratory randomised controlled trial of mebeverine, methylcellulose, placebo and a self-management website. BMC Gastroenterol 2013 Apr 21;13:68. [doi: 10.1186/1471-230X-13-68] [Medline: 23602047]
29. Ljótsson B, Falk L, Vesterlund AW, et al. Internet-delivered exposure and mindfulness based therapy for irritable bowel syndrome--a randomized controlled trial. Behav Res Ther 2010 Jun;48(6):531-539. [doi: 10.1016/j.brat.2010.03.003] [Medline: 20362976]
30. Hunt M, Miguez S, Dukas B, Onwude O, White S. Efficacy of Zemedy, a mobile digital therapeutic for the self-management of irritable bowel syndrome: crossover randomized controlled trial. JMIR Mhealth Uhealth 2021 May 20;9(5):e26152. [doi: 10.2196/26152] [Medline: 33872182]
31. Pedersen N. eHealth: self-management in inflammatory bowel disease and in irritable bowel syndrome using novel constant-care web applications. eHealth by constant-care in IBD and IBS. Dan Med J 2015 Dec;62(12):B5168. [Medline: 26621403]
32. Liegl G, Plessen CY, Leitner A, Boeckle M, Pieh C. Guided self-help interventions for irritable bowel syndrome: a systematic review and meta-analysis. Eur J Gastroenterol Hepatol 2015 Oct;27(10):1209-1221. [doi: 10.1097/MEG.0000000000000428] [Medline: 26164395]
33. Everitt HA, Landau S, O'Reilly G, et al. Assessing telephone-delivered cognitive-behavioural therapy (CBT) and web-delivered CBT versus treatment as usual in irritable bowel syndrome (ACTIB): a multicentre randomised trial. Gut 2019 Sep;68(9):1613-1623. [doi: 10.1136/gutjnl-2018-317805] [Medline: 30971419]
34. Lindfors P, Axelsson E, Engstrand K, et al. Online education is non-inferior to group education for irritable bowel syndrome: a randomized trial and patient preference trial. Clin Gastroenterol Hepatol 2021 Apr;19(4):743-751. [doi: 10.1016/j.cgh.2020.04.005] [Medline: 32289541]
35. Ljótsson B, Andersson G, Andersson E, et al. Acceptability, effectiveness, and cost-effectiveness of internet-based exposure treatment for irritable bowel syndrome in a clinical sample: a randomized controlled trial. BMC Gastroenterol 2011 Oct 12;11:WOS. [doi: 10.1186/1471-230X-11-110] [Medline: 21992655]
36. D'Silva A, Marshall DA, Vallance JK, et al. Meditation and yoga for irritable bowel syndrome: a randomized clinical trial. Am J Gastroenterol 2023 Feb 1;118(2):329-337. [doi: 10.14309/ajg.0000000000002052] [Medline: 36422517]
37. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
38. Li L, Xiong L, Zhang S, Yu Q, Chen M. Cognitive-behavioral therapy for irritable bowel syndrome: a meta-analysis. J Psychosom Res 2014 Jul;77(1):1-12. [doi: 10.1016/j.jpsychores.2014.03.006] [Medline: 24913335]
39. Cohen J. A power primer. Psychol Bull 1992;112(1):155-159. [doi: 10.1037/0033-2909.112.1.155] [Medline: 19565683]
40. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
41. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods 2010 Apr;1(2):97-111. [doi: 10.1002/jrsm.12] [Medline: 26061376]
42. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. Res Synth Methods 2017 Mar;8(1):5-18. [doi: 10.1002/jrsm.1230] [Medline: 28058794]
43. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023 Dec;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
44. Schmid CH, Stark PC, Berlin JA, Landais P, Lau J. Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. J Clin Epidemiol 2004 Jul;57(7):683-697. [doi: 10.1016/j.jclinepi.2003.12.001] [Medline: 15358396]
45. Cumpston M, Li T, Page MJ, et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane handbook for systematic reviews of interventions. Cochrane Database Syst Rev 2019 Oct 3;10(10):ED000142. [doi: 10.1002/14651858.ED000142] [Medline: 31643080]
46. Sterne JAC, Sutton AJ, Ioannidis JPA, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011 Jul 22;343:d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
47. Lau J, Ioannidis JPA, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. BMJ 2006 Sep 16;333(7568):597-600. [doi: 10.1136/bmj.333.7568.597] [Medline: 16974018]

XSL•FO
RenderX

48. Afonso J, Ramirez-Campillo R, Clemente FM, Büttner FC, Andrade R. The perils of misinterpreting and misusing "Publication Bias" in meta-analyses: an education review on funnel plot-based methods. Sports Med 2024 Feb;54(2):257-269. [doi: 10.1007/s40279-023-01927-9] [Medline: 37684502]
49. Ljótsson B, Hedman E, Andersson E, et al. Internet-delivered exposure-based treatment vs. stress management for irritable bowel syndrome: a randomized trial. Am J Gastroenterol 2011 Aug;106(8):1481-1491. [doi: 10.1038/ajg.2011.139] [Medline: 21537360]
50. Anderson EJ, Peters SL, Gibson PR, Halmos EP. Comparison of digitally delivered gut-directed hypnotherapy program with an active control for irritable bowel syndrome. Am J Gastroenterol 2025 Feb 1;120(2):440-448. [doi: 10.14309/ajg.0000000000002921] [Medline: 38940439]
51. Lee TY, Hsieh TC, Sung HC, Chen WL. Internet-delivered cognitive behavior therapy for young Taiwanese female nursing students with irritable bowel syndrome-a cluster randomized controlled trial. Int J Environ Res Public Health 2019 Feb 27;16(5):708. [doi: 10.3390/ijerph16050708] [Medline: 30818837]
52. Ljótsson B, Hesser H, Andersson E, et al. Provoking symptoms to relieve symptoms: a randomized controlled dismantling study of exposure therapy in irritable bowel syndrome. Behav Res Ther 2014 Apr;55:27-39. [doi: 10.1016/j.brat.2014.01.007] [Medline: 24584055]
53. Hunt M, Dalvie A, Ipek S, Glinski S, Macks R. Efficacy of a CBT self-help app (Zemedy) versus an education, relaxation, and mindfulness app for IBS: results from post-treatment, 3-month, and 6-month follow-up. J Clin Gastroenterol 2025 Mar 12. [doi: 10.1097/MCG.0000000000002164] [Medline: 40071822]
54. Zargar F, Fahim A, Nikgoftar N, Tarrahi MJ. Comparing the effect of internet-delivered short-term progressive muscle relaxation and psychoeducation on mindful ability, visceral hypersensitivity and symptoms of patients with irritable bowel syndrome. J Educ Health Promot 2023;12:259. [doi: 10.4103/jehp.jehp_1734_22] [Medline: 37727438]
55. Chao WC, Huang JC, Young SL, et al. Interplay of yoga, physical activity, and probiotics in irritable bowel syndrome management: a double-blind randomized study. Complement Ther Clin Pract 2024 Nov;57:101892. [doi: 10.1016/j.ctcp.2024.101892] [Medline: 39126817]
56. Hunt MG, Moshier S, Milonova M. Brief cognitive-behavioral internet therapy for irritable bowel syndrome. Behav Res Ther 2009 Sep;47(9):797-802. [doi: 10.1016/j.brat.2009.05.002] [Medline: 19570525]
57. Trieschmann K, Chang L, Park S, et al. The visceral sensitivity index: a novel tool for measuring GI-symptom-specific anxiety in inflammatory bowel disease. Neurogastroenterol Motil 2022 Sep;34(9):e14384. [doi: 10.1111/nmo.14384] [Medline: 35478469]
58. Makarova AA, Diukova GM, Ruchkina IN, et al. Visceral sensitivity in diagnostics and treatment of severe irritated bowel syndrome. Ter Arkh 2022 Sep 5;94(2S):356-361. [doi: 10.26442/00403660.2022.02.201394] [Medline: 36468983]
59. Wiklund IK, Fullerton S, Hawkey CJ, et al. An irritable bowel syndrome-specific symptom questionnaire: development and validation. Scand J Gastroenterol 2003 Sep;38(9):947-954. [doi: 10.1080/00365520310004209] [Medline: 14531531]
60. Francis CY, Morris J, Whorwell PJ. The irritable bowel severity scoring system: a simple method of monitoring irritable bowel syndrome and its progress. Aliment Pharmacol Ther 1997 Apr;11(2):395-402. [doi: 10.1046/j.1365-2036.1997.142318000.x] [Medline: 9146781]
61. Schäfer SK, Weidner KJ, Hoppner J, et al. Design and validation of a German version of the GSRS-IBS - an analysis of its psychometric quality and factorial structure. BMC Gastroenterol 2017 Dec 4;17(1):139. [doi: 10.1186/s12876-017-0684-8] [Medline: 29202711]
62. Gralnek IM, Hays RD, Kilbourne A, Naliboff B, Mayer EA. The impact of irritable bowel syndrome on health-related quality of life. Gastroenterology 2000 Sep;119(3):654-660. [doi: 10.1053/gast.2000.16484] [Medline: 10982758]
63. D'Silva A, Hua N, Modayil MV, Seidel J, Marshall DA. Digital health interventions are effective for irritable bowel syndrome self-management: a systematic review. Dig Dis Sci 2025 Feb;70(2):644-664. [doi: 10.1007/s10620-024-08672-7] [Medline: 39402208]
64. Melchior C, Colomier E, Trindade IA, et al. Irritable bowel syndrome: factors of importance for disease-specific quality of life. United European Gastroenterol J 2022 Sep;10(7):754-764. [doi: 10.1002/ueg2.12277] [Medline: 35830193]
65. Wolitzky-Taylor K, Craske MG, Labus JS, Mayer EA, Naliboff BD. Visceral sensitivity as a mediator of outcome in the treatment of irritable bowel syndrome. Behav Res Ther 2012 Oct;50(10):647-650. [doi: 10.1016/j.brat.2012.05.010] [Medline: 22877888]
66. Van Oudenhove L, Levy RL, Crowell MD, et al. Biopsychosocial aspects of functional gastrointestinal disorders: how central and environmental processes contribute to the development and expression of functional gastrointestinal disorders. Gastroenterology 2016 May;150(6):1355-1367. [doi: 10.1053/j.gastro.2016.02.027]
67. Drossman DA. Functional gastrointestinal disorders: history, pathophysiology, clinical features and Rome IV. Gastroenterology 2016 Feb 19:S0016-5085(16)00223-7. [doi: 10.1053/j.gastro.2016.02.032] [Medline: 27144617]
68. Hanlon I, Hewitt C, Bell K, Phillips A, Mikocka-Walus A. Systematic review with meta-analysis: online psychological interventions for mental and physical health outcomes in gastrointestinal disorders including irritable bowel syndrome and inflammatory bowel disease. Aliment Pharmacol Ther 2018 Aug;48(3):244-259. [doi: 10.1111/apt.14840] [Medline: 29901820]

XSL•FO
RenderX

69. Hod K, Ringel Y, van Tilburg MAL, Ringel-Kulka T. Bloating in irritable bowel syndrome is associated with symptoms severity, psychological factors, and comorbidities. Dig Dis Sci 2019 May;64(5):1288-1295. [doi: 10.1007/s10620-018-5352-5] [Medline: 30565010]
70. Black CJ, Yiannakou Y, Houghton LA, et al. Anxiety-related factors associated with symptom severity in irritable bowel syndrome. Neurogastroenterol Motil 2020 Aug;32(8):e13872. [doi: 10.1111/nmo.13872] [Medline: 32396247]
71. Lee C, Doo E, Choi JM, et al. The increased level of depression and anxiety in irritable bowel syndrome patients compared with healthy controls: systematic review and meta-analysis. J Neurogastroenterol Motil 2017 Jul 30;23(3):349-362. [doi: 10.5056/jnm16220] [Medline: 28672433]

Abbreviations

CBM: China Biology Medicine Disc
CBT: cognitive behavioral therapy
CNKI: China National Knowledge Infrastructure
GRADE: Grading of Recommendations Assessment, Development and Evaluation
IBS: irritable bowel syndrome
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SMD: standardized mean difference

Edited by S Brini; submitted 05.Nov.2025; peer-reviewed by FJ Yang, P Ebrahimi; revised version received 08.Apr.2026; accepted 09.Apr.2026; published 21.May.2026.

Please cite as:
Xu X, Liu T, Wang F, He H, Chen J, Liu J
Efficacy of Internet-Based Self-Help Interventions for Irritable Bowel Syndrome: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e87216
URL: https://www.jmir.org/2026/1/e87216
doi:10.2196/87216

© Xiaohong Xu, Ting Liu, Fang Wang, Hongyue He, Jiao Chen, Jingxian Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Toward Participatory Precision Health With Co-Designed Recommendations: Systematic Review of Just-in-Time Adaptive Interventions in Adolescents and Young Adults

Kathleen W Guan[1], MRes; Mohammed Amara[2]; Viviana Cortiana[3*]; Imran Khan[4*], MS; Sıla Gürbüz[5*], MD; Carmine Iorio[6*], PhD; Rayyan Ali[7*]; Christopher Adlung[1*], MS, MPS; Crystal R Smit[8], PhD; Annabel Vreeker[8,9], PhD; Eva Thalassinou[10*], MSc; Eeske van Roekel[11], PhD; Loes Keijsers[8], PhD; Mark de Reuver[1], PhD; Caroline A Figueroa[1,12], MD, PhD

[1]Faculty of Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, Delft, The Netherlands
[10]Research and Development, Gro-up Youth Work, Berkel en Rodenrijs, The Netherlands
[11]Department of Developmental Psychology, Tilburg University, Tilburg, The Netherlands
[12]School of Social Welfare, University of California, Berkeley, Berkeley, CA, United States
[2]Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom
[3]Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy
[4]Arkansas College of Osteopathic Medicine, Fort Smith, AR, United States
[5]Faculty of Medicine, Karadeniz Technical University, Trabzon, Turkey
[6]Department of Philosophy, Social Sciences and Education, University of Perugia, Perugia, Italy
[7]School of Medicine, Imperial College London, London, United Kingdom
[8]School of Social and Behavioural Sciences, Erasmus University Rotterdam, Rotterdam, The Netherlands
[9]Department of Child and Adolescent Psychiatry/Psychology, Erasmus University Medical Center, Rotterdam, The Netherlands
[*]these authors contributed equally

Corresponding Author:
Kathleen W Guan, MRes
Faculty of Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, Delft, The Netherlands

Abstract

Background: The transition from adolescence to young adulthood (age 10 - 25 years) constitutes a sensitive developmental period marked by rapid biological, psychological, and social change, during which preventive health interventions can shape long-term outcomes. Mobile health tools offer accessible opportunities for tailored support for this population, but often adapt poorly to dynamic contexts, resulting in inconsistent engagement and effects. Just-in-time adaptive interventions (JITAIs), which tailor support in real time using ongoing data, are increasingly explored as precision health strategies. However, how these mechanisms are designed, implemented, and evaluated for adolescents and young adults (AYAs) has not been systematically reviewed.

Objective: This review aimed to synthesize the evidence on JITAIs developed for AYAs, examine how their adaptive mechanisms have been designed to support specific health goals and changing AYA contexts, and assess methodological reporting quality to inform future precision health intervention development.

Methods: We conducted a systematic review in accordance with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) and SWiM (Synthesis Without Meta-Analysis) reporting guidelines. Twelve databases were searched for peer-reviewed studies published from 2013 to 2025. Eligible studies focused on participants aged 10 to 25 years and reported real-time adaptive mobile health interventions consistent with JITAI design principles. Two reviewers independently conducted screening, data extraction, and methodological quality appraisal using the Joanna Briggs Institute checklists. AYA coauthors contributed to all phases. Due to substantial heterogeneity in study populations, intervention content, adaptive mechanisms, comparators, and outcome measurements, findings were synthesized narratively, and no meta-analysis was conducted.

Results: A total of 61 unique interventions were included. JITAIs for AYAs addressed substance use (n=24, 39.3%), mental health (n=23, 37.7%), and physical health or chronic conditions (n=14, 23%). JITAI tailoring mechanisms relied predominantly on self-reported behavioral data. Decision rules were typically symptom threshold–based, and decision points were commonly daily or event-triggered. Methodological concerns with reporting on intervention administration, participant selection, and outcome measurement reliability were pervasive across all studies, limiting the interpretability of observed effects and cross-study

XSL•FO
RenderX

comparisons. Ethical considerations, including researcher positioning and reflexivity, alongside the depth of reporting around participatory AYA engagement in design and implementation, were also inconsistent.

Conclusions: This review contributes a novel perspective to AYA digital health by moving beyond intervention outcomes to examine how core adaptive mechanisms are operationalized for AYAs across multiple health domains, while also integrating AYA perspectives into the interpretation of findings and recommendations. Unlike prior reviews focused primarily on adults or specific conditions, it identifies broader contextual, methodological, and ethical considerations relevant to AYA precision health. These findings highlight the need for more transparent, contextually responsive, and youth-centered adaptive interventions, alongside more rigorous designs for evaluating adaptive intervention components in daily life contexts.

Trial Registration: PROSPERO CRD42023473117; https://www.crd.york.ac.uk/PROSPERO/view/CRD42023473117

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2024-083870

(J Med Internet Res 2026;28:e84422) doi:10.2196/84422

KEYWORDS

adolescents; young adults; just-in-time adaptive interventions; mobile health; health behaviors; precision health

Introduction

Adolescence is a formative period for preventive health interventions. As they transition into young adulthood, adolescents experience continued brain and physical development, alongside heightened vulnerability to risk factors for poor health [1,2]. For example, partaking in behaviors that compromise health, such as substance use, poor diet, and insufficient physical activity, in early life can increase the risk of chronic diseases such as cancer [3], and disrupt key neurodevelopmental processes related to stress regulation and executive functioning [4]. Further, the transition to adulthood is a complex developmental stage, shaped by rapidly shifting and interrelated biological, psychological, and social changes that are unique to each individual [5]. Behavioral patterns during this period emerge dynamically in response to fluctuating contexts and interactions, with significant implications for long-term health. For example, food insecurity may predispose adolescents to poor dietary habits, which in turn can affect their physical health, cognitive development, and emotional resilience over time [6]. In light of these vulnerabilities, early and precisely targeted interventions during this period are critical for mitigating both immediate and long-term health risks to promote healthier developmental pathways. Given the record number of adolescents and young adults (AYAs; aged 10-25 years) globally [7], investing in tailored health interventions during adolescence can significantly improve population health outcomes while alleviating pressures on health care systems [8].

Precision health represents a significant paradigm shift from generalized medical approaches, prioritizing adaptive strategies that tailor health support to optimize diverse outcomes unique to each individual [9]. Although much of the initial application of precision health has occurred within adult medical contexts [8], there is growing recognition that such approaches may be particularly suited to addressing the multifaceted challenges during the transition to adulthood [10]. Importantly, precision health approaches during this unique period can enhance both the relevance and effectiveness of AYA health interventions. In particular, addressing behaviors and states calls for interventions that precisely tailor support to individual risk factors and contextual influences shaping AYAs' lives.

Mobile health (mHealth) technologies are a key avenue for delivering tailored health support to AYAs, whose pervasive use of digital platforms throughout their daily lives and affinity for mobile communication provide opportunities for accessible intervention delivery. Such tools have demonstrated some potential to support and promote positive health behaviors, such as increasing physical activity, improving nutrition, and reducing substance use among adolescents [11]. However, mHealth interventions often produce inconsistent effects due to challenges in delivering engaging content that is precisely relevant to AYAs' evolving health contexts over time [12,13]. Extant literature has suggested that the use of digital tools to support health behaviors may decline when interventions fail to resonate with or adapt to changing circumstances [14,15]. Demographics (eg, age and gender), psychosocial characteristics (eg, mood), and contextual factors (eg, setting of delivery) play a significant role in individual engagement with mHealth interventions [13]. Thus, it is necessary to identify key considerations for designing mobile interventions that are not only tailored but also dynamically and temporally responsive to AYAs' shifting needs.

One particularly promising precision health approach that has emerged is the just-in-time adaptive intervention (JITAI). JITAIs use ongoing data from mobile sensing technology with the aim of delivering tailored, contextually relevant support precisely when it is most needed for each individual [16]. JITAIs can provide a flexible strategy for promoting positive health behaviors across individually salient moments and contexts by dynamically adjusting content in response to individual changes. The temporal adaptability of JITAIs is particularly important given the variability in AYA development, as their health risks and support needs can rapidly shift based on environmental and intrapersonal factors such as mood, stress levels, and social interactions from moment to moment [17,18]. JITAIs can crucially enhance the responsiveness of precision health strategies by delivering support attuned to ongoing changes in AYA contexts, particularly when individuals can be receptive to this support, which can help ensure tailored intervention content remains relevant and effective in addressing evolving needs.

Despite their potential, most prior analyses of JITAIs have focused on older adults or specific health conditions [19-24].

XSL•FO
RenderX

As such, there is a limited understanding of how JITAIs can support the dynamic health needs of AYAs as they transition into early adulthood. Given the rapidly changing and diverse contexts that characterize this formative developmental stage, understanding how JITAIs adapt to this period (10-25 years) is critical.

This review thus had 3 specific objectives. First, it aimed to map and synthesize the evidence based on JITAIs developed for AYA health contexts. Second, it aimed to characterize how key adaptive components—namely, tailoring variables, intervention options, decision rules, and decision points [16]—were designed and applied in relation to targeted AYA health outcomes. Third, it aimed to critically appraise the quality of the included studies to identify methodological and evidence gaps. Overall, the findings are intended to inform the design of more effective, personalized, and contextually responsive mHealth interventions to support healthier trajectories from adolescence into adulthood and beyond.

Methods

Overview of Search Strategy

This systematic review followed reporting guidelines from PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) and SWiM (Synthesis Without Meta-Analysis), respectively [25,26]. The review was preregistered in the International Prospective Register of Systematic Reviews (PROSPERO) database under registration number CRD42023473117. The review protocol was also published after peer review [27]. No ethical approval was required as this review uses only publicly available literature.

The search strategy for the review was developed in consultation with an information specialist, and the rationale for the search strategy, including scope of dates and search terms, is reported in the additional material according to PRISMA–Search extenion (PRISMA-S) [28]. The search strategy used a combination of keywords and database-specific subject headings (eg, Medical Subject Headings terms) to capture peer-reviewed literature related to JITAIs, AYAs, digital health technologies, and health outcomes and was peer-reviewed [27]. The literature search took place across multiple electronic databases to identify relevant literature from 2013 to 2025; these databases included PubMed, Scopus, Web of Science, PsycINFO, Embase, Cumulative Index to Nursing and Allied Health Literature, Institute of Electrical and Electronics Engineers Xplore, Association for Computing Machinery Digital Library, Cochrane Central, Dimensions, World Health Organization Global Index Medicus, and Google Scholar.

The individual database searches were conducted in July 2025, with additional keyword-based searching in Google Scholar until December 2025. Reference list screening of extracted studies was used to identify further eligible studies. Further, studies included in prior systematic reviews of JITAIs were consulted [19-24], which enabled inclusion of all relevant studies published before (as the literature search strategy only started in 2013, the first explicit mention of JITAI and AYAs in the literature [29,30]). These steps helped ensure comprehensive coverage of past and emerging studies. It was not deemed necessary to contact any study authors.

Screening and Data Extraction

All literature search records were imported into Covidence, a systematic review management platform developed by Veritas Health Innovation, with duplicates automatically removed. Each title and its accompanying abstract were screened by 2 reviewers (either KWG, CA, CI, IK, and/or RA, each independently), following a pilot screening phase to ensure consistency in eligibility decisions. Full texts were also screened by 2 reviewers each. All discrepancies during screening or assessment of full texts were resolved through discussion.

Study inclusion was primarily determined by whether studies provided sufficient detail on JITAI components to inform the review's focus on how these interventions adapt to AYA contexts and health needs. Studies were eligible for inclusion if they were peer-reviewed, focused on AYAs aged 10-25 years, reported mHealth interventions that adapted in real time to user data, and were consistent with the defining features of JITAIs. Other mHealth studies were out of scope, including JITAI studies that involved participants beyond the AYA age range (determined by reported sample age range and mean age, where available) and did not report these findings separately [31]. We selected the age range of 10 to 25 years, given its increasing recognition as its own heterogeneous yet distinct developmental period among scholars [10,32-34]. We did not restrict by study type or design, such as control conditions. We also excluded just-in-time simulation studies that did not involve evaluating an intervention [35].

Data from each study were also extracted by 2 reviewers (either KWG, SG, VC, and/or RA, each independently) using a piloted, standardized data extraction form to ensure consistency. The data extraction form followed standardized conceptualizations of JITAI in the literature [16,27], which outline key components guiding interventions to be adapted to individual needs: intervention options or the different forms of support or methods for delivering that health support, which may be considered based on tailoring variables; tailoring variables, referring to the data collected about an individual, which inform how and when an intervention should be adapted; and decision rules or the guidelines that use these tailoring variables to determine the selection, timing, and delivery of intervention options at specific decision points or specific moments, such as particular times of the day or contextual triggers when these decisions are made [16,27]. These components together shape how JITAIs dynamically adjust to ensure interventions remain relevant across shifting adolescent contexts [16]. By structuring the form around these components, the method was aligned with the review's focus on mechanisms of adaptability in JITAIs.

Further, distal and proximal outcomes were also extracted, as they indicate the health objectives toward which tailoring mechanisms are directed. Specifically, distal outcomes represent the overarching, long-term goals of a JITAI, while proximal outcomes are short-term health objectives that often serve as mediators that lead toward these distal targets [16]. Thus, data extracted from each study included participant characteristics, intervention details, tailoring mechanisms, study design, and

XSL•FO
RenderX

outcomes reflecting the goals of tailoring mechanisms (both proximal and distal). Refer to Multimedia Appendices 1-3 for the full extraction data for each study.

No formal amendments were made to the registered or published protocol [27]. Several additional outcomes of interest, including user engagement and ethical considerations, were identified in the protocol as potentially informative for interpreting effectiveness; however, inconsistent measurement and incomplete reporting meant they could not be systematically synthesized or graded. While information on AYA engagement and ethical considerations was documented during data extraction for each study where reported, reporting was sparse and also lacked sufficient detail for structured synthesis.

Risk of Bias and Synthesis

Two reviewers (KWG and MA) independently assessed methodological quality, including risk of bias, using the Joanna Briggs Institute checklists for distinct study designs, with discrepancies resolved through discussion [36]. The Joanna Briggs Institute checklists for each study included critical appraisal according to temporal precedence, selection and allocation, confounding factors, intervention administration, outcome measurement (time points, method, and reliability), participant retention, and statistical conclusion validity, which we report in the Results section.

These appraisals informed consideration of the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) framework [37], which was prespecified in the protocol to assess certainty of evidence across studies. However, application of GRADE across studies requires sufficient comparability in outcome definitions, measures, and reporting [38], and this was not present within any outcome domain in this review. We therefore did not undertake a formal assessment of evidence certainty or reporting bias. Further, although studies could be grouped broadly by outcome domain, substantial heterogeneity remained in participant populations and settings, intervention content and delivery, tailoring mechanisms, comparator conditions, and outcome assessment. We therefore also did not conduct a meta-analysis because we did not deem the included studies sufficiently similar to support a meaningful pooled estimate. Findings were instead synthesized narratively and organized according to key JITAI tailoring components and outcome domains, including intervention aims, tailoring variables, decision rules, and decision points. All included studies were considered in the narrative synthesis.

Engagement of AYAs

To ground implications of this review in lived experience, we directly engaged 12 AYAs, including health activists from the World Health Organization European Region's Youth4Health network, a platform for AYA participation in health policymaking. These AYAs, including young entrepreneurs, medical students, and policy advocates, provided critical perspectives that informed study interpretations and implications. Seven AYAs were directly involved as coauthors in screening studies, extracting data, and conducting critical appraisals, including the lead author. AYA perspectives were further elicited through a dedicated virtual consultative session with AYAs who were not part of the coauthorship team. During this session, 5 AYAs reviewed the data extraction tables from the review and engaged in semistructured critical dialogue regarding intervention acceptability, design relevance, and the alignment of findings with lived adolescent experiences. These discussions were documented via meeting minutes and synthesized by KWG and MA to identify recurring interpretive themes. Given several of the AYA participants' roles as coauthors in the review alongside the limited sample, these AYA contributions are reported in the Discussion section as explicitly interpretive observations to contextualize preliminary recommendations for future work, rather than as additional empirical data derived from the included studies.

Results

Overview

A total of 3626 unique records were identified via the literature search and screened in the title and abstract phase (Figure 1). Of these, 3374 records were excluded as they were clearly focused on adult populations or not intervention studies, such as studies collecting ecological momentary assessments only. Following this exclusion, a total of 252 full-text papers were assessed for relevance. Papers that were excluded at the full-text stage did not develop mobile interventions that were tailored to dynamic adolescent contexts, for example, those offering only prescheduled digital health support without ongoing adaptation, and therefore did not meet the definition of a JITAI [16]. This resulted in a total of 69 studies (representing 61 unique interventions) developing JITAIs for AYAs included in the findings synthesis that follows.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for adolescents and young adults (AYAs) just-in-time adaptive intervention (JITAI) study selection.

XSL•FO
RenderX

Risk of Bias and Quality of Evidence

The risk of bias and critical appraisal matrix (Figure 2) shows that every study had at least one methodological concern, indicating a moderate to high risk of bias present across the evidence base. We organize these by study design (randomized controlled trial [RCT], noncontrolled experimental studies, and qualitative studies), as critical appraisal is conducted based on the specific design of a study [36]. Importantly, many "unclear" ratings are likely attributable to incomplete reporting rather than the absence of methodological safeguards, reflecting variability in reporting practices and the lack of standardized conventions in this field. These issues significantly reduce confidence in reports of effectiveness and complicate comparisons across studies.

Figure 2. Risk of bias and methodological quality appraisal of randomized controlled trials, noncontrolled experimental studies, and qualitative studies of just-in-time adaptive interventions (JITAIs) for adolescents and young adults (AYAs). RCT: randomized controlled trial [39-104].

Authors and Year	Temporal precedence	Selection and allocation	Confounding factors	Intervention administration	Outcome measurement (timepoints)	Outcome measurement (method)	Outcome measurement (reliability)	Participant retention	Statistical conclusion validity
Rodgers et al. 2005	RCT	Yes	Yes	Unclear	Yes	Yes	Yes	Unclear	Yes
Weitzel et al. 2007	RCT	Unclear	Unclear	No	Yes	Yes	Unclear	Yes	Unclear
Suffoletto et al. 2012	RCT	Yes	Yes	No	Yes	Yes	Yes	Yes	Unclear
Nollen et al. 2014	RCT	Unclear	Yes	No	Yes	Yes	Yes	Yes	Unclear
Witkiewitz et al. 2014	RCT	Unclear	Unclear	No	Yes	Yes	Yes	Yes	Yes
Suffoletto et al. 2015	RCT	Yes	Yes	Unclear	Yes	Yes	Yes	Unclear	Yes
Perry et al. 2017	RCT	Unclear	Yes	No	Yes	Yes	Yes	Unclear	Unclear
Pyky et al. 2017	RCT	Yes	Yes	Unclear	Yes	Yes	Yes	Unclear	Yes
Shrier et al. 2018	RCT	Unclear	Yes	No	Yes	Yes	Yes	Unclear	Yes
Simons et al. 2018	RCT	Unclear	Yes	No	Yes	Yes	Yes	Yes	Yes
Wright et al. 2018	RCT	Unclear	Yes	No	Yes	Yes	Unclear	No	Yes
Kennard et al. 2018	RCT	Yes	Unclear	Unclear	Yes	Yes	Yes	Yes	Yes
O'Donnell et al. 2019	RCT	Yes	Yes	Unclear	Yes	Yes	Unclear	Yes	Yes
Suffoletto et al. 2019	RCT	Yes	Unclear	Unclear	Yes	Yes	Unclear	Unclear	Yes
Hunter et al. 2020	RCT	Unclear	Yes	No	Yes	Yes	Yes	Yes	Unclear
van Aubel et al. 2020	RCT	Unclear	Yes	Unclear	Yes	Yes	Unclear	Yes	Yes
Fedele et al. 2021	RCT	Yes	Yes	No	Yes	Yes	Yes	Yes	Unclear
King-Dowling et al. 2021	RCT	Unclear	Yes	No	Yes	Yes	Unclear	Yes	Unclear
Santa Maria et al. 2021	RCT	Unclear	Yes	No	Yes	Yes	Unclear	Unclear	Yes
Suffoletto et al. 2021	RCT	Unclear	Unclear	Unclear	Yes	Yes	Yes	Yes	Unclear
Paetzold et al. 2022	RCT	Unclear	Unclear	Unclear	Yes	Yes	Unclear	Yes	Unclear
Schenkel et al. 2022	RCT	Unclear	Unclear	No	Yes	Yes	Unclear	Unclear	Yes
Bell et al. 2023	RCT	Yes	Yes	No	Yes	Yes	Yes	Yes	Unclear
Lucas-Thompson et al. 2023	RCT	Unclear	Unclear	No	Yes	Yes	Yes	Unclear	Unclear
Marciniak et al. 2023	RCT	Yes	Yes	Unclear	Yes	Yes	Yes	Yes	Yes
Reininghaus et al. 2023	RCT	Yes	Yes	Unclear	Yes	Yes	Yes	Yes	Yes
Reininghaus et al. 2024	RCT	Yes	Yes	Unclear	Yes	Yes	Yes	Yes	Yes
Orzikulova et al. 2024	MRT	Unclear	Yes	Unclear	Yes	Yes	Yes	Unclear	Yes
Engelskirchen et al. 2025	RCT	Unclear	Yes	No	Yes	Yes	Yes	Unclear	Yes
Lee et al. 2025	RCT	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Riley et al. 2008	Yes	Unclear	Unclear	Yes	No	Yes	Unclear	Unclear	Unclear
Shrier et al. 2014a	Yes	No	No	Yes	Yes	Yes	Unclear	No	Unclear
Spruijt-Metz et al. 2015	Yes	No	No	Yes	Yes	Yes	Unclear	Yes	Unclear
Ding et al. 2016	Yes	Yes	Unclear	Yes	Yes	Yes	Unclear	Unclear	Unclear
Fortier et al. 2016	Yes	No	No	Yes	Yes	Yes	Yes	Unclear	Unclear
Leonard et al. 2017	Yes	No	No	Yes	Yes	Yes	Unclear	Yes	Unclear
Bracisjewski et al. 2018	Yes	No	No	Yes	Yes	Yes	Unclear	No	Unclear
Leonard et al. 2018	Yes	Unclear	No	Yes	Yes	Yes	Unclear	Unclear	Unclear
Suffoletto et al. 2018	Yes	No	No	Yes	Yes	Yes	Unclear	Unclear	Unclear
Neal-Barnett et al. 2019	Yes	No	No	Yes	Yes	Yes	Yes	Unclear	Unclear
Kazemi et al. 2020	Yes	Yes	Unclear	Yes	No	Yes	Yes	No	Unclear
Shrier et al. 2020	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Unclear
Silk et al. 2020	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Unclear
Stevenson et al. 2020	Yes	No	No	Yes	Yes	Yes	Unclear	No	Unclear
Blevins et al. 2021	Yes	No	No	Yes	Yes	Yes	Yes	Unclear	Unclear
Coughlin et al. 2021	Yes	No	Unclear	Unclear	Yes	Yes	Yes	Unclear	Yes
Cushing et al. 2021	Yes	Unclear	Unclear	Yes	Yes	Yes	Yes	Yes	Yes
Rauschenberg et al. 2021	Yes	No	No	Yes	Yes	Yes	Unclear	Yes	Unclear
Caon et al. 2022	Yes	Yes	Unclear	Yes	Yes	Yes	Unclear	Unclear	Unclear
Domin et al. 2022	Yes	No	No	Yes	Yes	Yes	Unclear	Unclear	Unclear
Wu et al. 2024	Yes	No	Yes	Yes	Yes	Yes	Unclear	Yes	No
Cushing et al. 2025	Yes	Unclear	Unclear	Yes	Yes	Yes	Yes	Unclear	Yes
van Asselt et al. 2025	Yes	Yes	Unclear	Yes	Yes	Yes	Yes	Unclear	Yes

	Philosophical–methodological congruity	Researcher positioning and reflexivity	Participant representation	Ethics and approval	Data–conclusion consistency
Chand et al. 2006	Yes	No	Yes	Unclear	Yes
Shrier et al. 2014b	Unclear	No	Yes	Yes	Yes
Wright et al. 2016	Yes	Unclear	Yes	Yes	Yes
Cerrada et al. 2017	Unclear	No	Yes	Yes	Yes
Leonard et al. 2017	Yes	No	Yes	Yes	Yes
Shrier et al. 2017	Unclear	No	Yes	Yes	Yes
Kazemi et al. 2018	Unclear	No	No	Unclear	Yes
Lucas-Thompson et al. 2020	Unclear	No	No	Yes	Yes
Martin et al. 2020	Yes	Unclear	Yes	Yes	Yes
Acorda et al. 2021	Yes	Unclear	Yes	Yes	Yes
Ranney et al. 2021	Yes	No	Yes	Yes	Yes
Van Keulen et al. 2022	Yes	No	No	Yes	Yes
Postma et al. 2024	Yes	Yes	Yes	Yes	Yes
Hiller et al. 2025	Yes	Yes	Yes	Yes	Yes
van Asselt and Roke 2025	Yes	Yes	Unclear	Yes	Yes

Figure Legend	
Yes	Authors describe in detail
Unclear	Authors describe but not ifully
No	Authors do not describe

Intervention administration was the most prominent limitation unique to RCTs. Across studies, reporting was limited on whether treatment groups were treated identically apart from the intervention of interest. Importantly, intervention administration is inherently difficult to deliver in a blinded and standardized manner in JITAIs because these interventions are designed to deliver tailored and timely support based on momentary context, vulnerability, and receptivity, rather than static content delivered uniformly across participants [16]. As a result, intervention groups often received more personalized prompts, summaries, feedback, adaptive logic, or supportive features than comparator groups, making performance bias difficult to avoid.

Meanwhile, noncontrolled studies frequently showed weaknesses in selection procedures, as comparison groups were often absent and allocation methods lacked detailed justification. This limits confidence that observed effects can be attributed to the JITAIs being evaluated rather than to pre-existing

XSL•FO
RenderX

participant differences or co-occurring influences. Again, this limitation may be especially important for JITAIs, because their effects are hypothesized to depend on timing, context, and adaptive delivery, making it necessary to isolate the contribution of specific intervention components [16]. Further, both randomized and noncontrolled studies in this study demonstrated consistent weaknesses in outcome measurement reliability, participant retention, and statistical conclusion validity. Few studies provided sufficient study-specific evidence on measurement reliability, and loss to follow-up was often incompletely reported. Further, comprehensive details on statistical power and power assumptions were frequently absent or reported with caveats, consistent with the predominance of pilot samples across the included studies.

Within qualitative studies, the most consistent limitation was researcher positioning and reflexivity, which was rarely reported or addressed. This was reflected in the limited reporting of whether the researcher's cultural or theoretical orientation was declared. Further, the potential influence of the researcher on study participants, or vice versa, was rarely acknowledged.

Health Outcomes Addressed by JITAIs for AYAs

Overview

Among JITAIs for AYAs, distal outcomes and their related proximal outcomes demonstrate some consistent priorities as well as emerging trends, with diverse conditions as the target of tailored support. We organize key outcomes targeted in JITAIs for AYAs into 3 domains: substance use, mental health, and physical health (Table 1). While substance use typically spans both mental and physical health, we present it separately here as it has been the most frequent and distinct focus of JITAIs for AYAs. Further, while we highlight some notable observed effects as well as inconsistencies, all results we report below should be interpreted with caution given the aforementioned methodological limitations and risks of bias detected across the evidence base.

Table . Contexts of just-in-time adaptive interventions (JITAIs) for adolescents and young adults (AYAs) in substance use, mental health, and physical health.

Domain	Contexts for just-in-time adaptive intervention
Substance use (n=24)	University students with alcohol-related risk or use (n=6, 25%), emergency department patients with hazardous drinking (n=4, 16.7%), general AYA[a] population with alcohol-related risk or use (n=3, 12.5%), homeless AYAs with alcohol-related risk or use (n=2, 8.3%), primary care clinic patients with cannabis use (n=2, 8.3%), general adolescent and young adult population who smoke (n=1, 4.2%), college or university students who smoke (n=1, 4.2%), university students with alcohol and smoking co-use (n=1, 4.2%), patients receiving psychiatric partial hospitalization services who drink to cope with anxiety or depression (n=1, 4.2%), foster care transition youth with alcohol-related risk or use (n=1, 4.2%), Korean American daily smokers (n=1, 4.2%), and apprenticeship or vocational students with alcohol-related risk or use (n=1, 4.2%).
Mental health (n=23)	Clinical mental health care patients with depression, anxiety, and/or psychosis-related problems (n=5, 21.7%), university students with nonclinical mental health or stress-related problems (n=3, 13%), AYAs with mixed or nonspecific mental health care problems (n=2, 8.7%), adolescents with mental health or well-being-related needs (n=2, 8.7%), autistic AYAs in outpatient mental health services (n=2, 8.7%), AYA female patients at community clinics with sexual behavior risks (n=2, 8.7%), general population with digital behavior risks (n=1, 4.3%), homeless adolescent mothers in transitional living with psychosocial vulnerability (n=1, 4.3%), psychiatric inpatient adolescents with suicidality (n=1, 4.3%), AYAs with generalized anxiety (n=1, 4.3%), AYAs with low self-esteem and childhood adversity (n=1, 4.3%), apprenticeship or vocational students (unspecified mental health context; n=1, 4.3%), and general healthy AYAs (n=1, 4.3%).
Physical health (n=14)	Asthma patients and their parents (n=2, 14.3%), healthy adolescents (n=2, 14.3%), healthy university students (n=2, 14.3%), secondary school students (n=1, 7.1%), patients with asthma (n=1, 7.1%), AYA cancer survivors (n=1, 7.1%), adolescents undergoing cancer treatment (n=1, 7.1%), low-income racial or ethnic minority females (n=1, 7.1%), lower-educated working young adults (n=1, 7.1%), overweight Hispanic adolescents (n=1, 7.1%), and young men preparing for military conscription (n=1, 7.1%).

[a]AYA: adolescent and young adult.

JITAIs Addressing AYA Substance Use

Substance use reduction has been a consistently central focus of JITAIs for AYAs, with interventions (39.3%, n=24/61) targeting reduction in areas such as smoking cessation, alcohol moderation, and marijuana use [39-61,105]. The most common intervention contexts were for college or university students with alcohol-related risk or use (n=6, 25%) and emergency department patients with hazardous drinking (n=4, 16.7%). Novel population-specific pilot studies included interventions with homeless young adults and AYAs in foster care,

XSL•FO
RenderX

respectively, with alcohol-related risks, as well as clinic patients with cannabis use [45,53,60].

JITAIs addressing substance use among AYAs typically emphasize regulating urges and reinforcing commitment to abstinence, often using momentary tailoring based on triggers and urge states [39,45,49,50,54,60]. AYA JITAIs targeting smoking cessation demonstrated positive outcomes in several studies, including higher engagement linked to reduced smoking and a lower likelihood of smoking reported in one study [44]. Evidence includes higher abstinence at 6 weeks in an RCT (relative risk 2.20, 95% CI 1.79 - 2.70) [39]. Similarly, marijuana use reduction interventions also yielded encouraging results. Another pilot randomized trial found reduced odds of marijuana use after contextual or behavioral triggers (odds ratio 0.54, 95% CI 0.31 - 0.95), alongside a greater decline in momentary marijuana desire [45]. Additionally, interventions aimed at reducing drug use showed some promise, including in an RCT with AYAs experiencing homelessness with lower odds of reported drug use over time [49,60].

Alcohol-related outcomes were more mixed. Some studies reported reductions, including reduced heavy drinking days and drinks per drinking day. One emergency department sample found longer-term benefits, where interactive text messages reduced binge drinking and alcohol-related injury, with one third fewer binge drinking days at 9 months compared with controls (interrater reliability 0.69, 95% CI 0.60 - 0.79) [47,62,63]. In contrast, a small but nonsignificant increases in consumption during heavy drinking occasions were reported elsewhere [46]. Other studies found mostly limited or nonsignificant between-group differences in alcohol outcomes [46,105]. Another intervention increased protective behavioral strategies even when drinking outcomes did not differ [48].

JITAIs Addressing AYA Mental Health

JITAIs with a specific focus on AYA mental health and well-being emerged in 2017 and have been steadily growing since (37.7%, n=23/61). These include interventions that target anxiety, depression, psychological distress, and suicide prevention among AYAs [64-79,106]. Further, JITAIs that address AYAs' well-being more generally (as opposed to clinical symptoms) focus on enhancing emotional resilience and coping strategies [80-85], including after experiencing cyber-victimization [86]. A development since 2024 includes 2 studies focused on promoting healthy digital behaviors, specifically to address the overuse of digital devices [87,88]. Intervention contexts were highly heterogeneous, with some JITAIs designed for clinical depression, anxiety, or psychosis-related problems (n=5, 21.7%). Pilot studies included interventions conducted with autistic AYAs in outpatient services, female community clinic patients with sexual risk behaviors, and homeless adolescent mothers in transitional living programs [76,82,85]. Further, JITAIs for AYA mental health were supplements to in-person counseling or support group programs in several studies (n=9, 39.1%), although the structure of these programs (eg, individual vs group sessions and therapeutic framework) differed greatly.

Several AYA JITAIs targeting anxiety and depression symptoms demonstrated promising improvements across follow-up assessments, indicating potential effectiveness in alleviating specific clinical symptoms. Two RCTs reported promising mental health effects, with one trial observing significant reductions in depressive symptoms ($\beta=-0.34$; $P=.004$), anxiety symptoms ($\beta=-0.10$; $P=.01$), and perceived stress ($\beta=-0.18$; $P=.04$), while another found greater reductions in anxiety than in the control group (General Anxiety Disorder 7-item [GAD-7] scale: $F_{1,51}=9.06$; $P=.004$; $d=0.61$), with a nonsignificant trend for depression (Patient Health Questionnaire 8-item [PHQ-8] scale: $F_{1,51}=3.21$; $P=.08$; $d=0.50$) [70,79]. However, less definitive results were observed in other studies. In one suicide prevention trial, no significant differences were found between intervention and control groups in suicide attempts or suicidal ideation, although effects were in the hypothesized direction [65,67]. Meanwhile, findings for general mental well-being outcomes were mixed: some interventions reported reduced perceived stress or improved self-esteem and confidence, whereas others found no significant effects on stress outcomes [72,79,80,82].

JITAIs Addressing AYA Physical Health

AYA physical health JITAIs (22.9%, n=14/61) often addressed obesity risk through the promotion of physical activity and healthy diets in distinct populations (eg, racial minority groups) [89-98], although some studies have also focused on the management of chronic conditions, such as pain in cancer management and asthma [99-102]. Physical health strategies in these JITAIs often focused on incremental shifts, such as reducing sedentary behavior and promoting daily walking [90,92,93,96,97]. AYA JITAIs focused on nutrition encouraged immediate healthy choices to mitigate obesity risks [89,94,95,98]. Meanwhile, AYA JITAIs for chronic condition management promoted adherence behaviors (eg, medication adherence) as well as pain management support, incorporating hybrid "clinical escalation" logic, such as nurse practitioner alerts triggered by diary thresholds and symptom ratings in pediatric oncology pain support, as well as caregiver and clinical staff alerts triggered by safety rules in an asthma action-plan app [99-101].

Physical activity promotion outcomes were variable. Modest increases in daily physical activity levels were observed, although these changes did not always extend to reductions in sedentary time [96,97]. For example, one study showed an initial decrease in sedentary minutes following intervention ($t_{17}=-1.79$; $P=.04$), though this effect attenuated over time and became nonsignificant by the second and third weeks [97]. Interventions targeting obesity-related risks also demonstrated limited effectiveness, with no significant changes in key indicators such as screen time and BMI. However, some evidence indicated improvements in diet such as increased fruit and vegetable consumption and decreased sugar-sweetened beverage intake [89]. Finally, in a study specifically targeting cancer pain management, the intervention group reported fewer instances of moderate to severe pain compared to control ($t_{2145}=2.67$; 95% CI 0.0036 - 0.0235; $P=.01$), although there were no significant differences in average daily pain reduction [100].

XSL•FO
RenderX

Theoretical Frameworks Informing Intervention Options

Table 2 presents a domain-level synthesis of the theoretical frameworks that informed (distinct) intervention designs, alongside the primary intervention options, tailoring variables, and decision points. Note that counts overlap as studies applied multiple components simultaneously. Across all JITAIs targeting AYA health, researchers commonly used tailored digital messages, which were delivered via text or app-based messaging tools, while leveraging diverse theoretical frameworks to inform intervention content.

Table . Summary of theoretical frameworks, intervention options, tailoring variables, and decision points in just-in-time adaptive interventions (JITAIs) for adolescents and young adults (AYAs) in substance use, mental health, and physical health.

Domain	Substance use (n=24)	Mental health (n=23)	Physical health (n=14)
Theoretical frameworks	Motivational interviewing and related approaches (n=8, 33.3%), cognitive behavioral and related approaches (n=6, 25%), transtheoretical model (n=4, 16.7%), self-regulation (n=3, 12.5%), information-motivation-behavioral skills model (n=2, 8.3%), harm reduction (n=2, 8.3%), and not reported (n=3, 12.5%).	Cognitive behavioral and related approaches (n=8, 34.8%), emotion regulation approaches (n=3, 13%), motivational interviewing (n=3, 13%), mindfulness-based approaches (n=3, 13%), transdiagnostic approaches (n=3, 13%), and compassion-focused approaches (n=2, 8.7%).	Self-regulation and control-based approaches (n=6, 42.9%), Behavior Change Wheel (n=3, 21.4%), positive psychology (n=2, 14.3%), self-determination theory (n=2, 14.3%), and not reported (n=2, 14.3%).
Intervention options	Psychoeducation modules (n=16, 66.7%), supportive messages (n=15, 62.5%), self-monitoring assessment (n=14, 58.3%), goal-setting or planning support (n=8, 33.3%), interactive activities (n=8, 33.3%), protective strategies or coping support (n=8, 33.3%), and reminders (n=5, 20.8%).	Coping and emotional support messages (n=15, 65.2%), psychoeducation modules (n=15, 65.2%), therapeutic skills practice (n=11, 47.8%), self-monitoring check-ins (n=9, 39.1%), interactive activities (n=7, 30.4%), feedback/progress summaries (n=6, 26.1%), and reminders (n=5, 21.7%).	Interactive activities (n=10, 71.4%), psychoeducation modules (n=10, 71.4%), suggestions/feedback (n=10, 71.4%), goal setting (n=8, 57.1%), reminders (n=8, 57.1%), and self-monitoring/tracking (n=7, 50%).
Tailoring variables	Self-reports of behavior (n=16, 66.7%), motivation or readiness (n=15, 62.5%), triggers or high-risk situations (n=11, 45.8%), emotion or affect (n=8, 33.3%), goal attainment (n=8, 33.3%), urges (n=7, 29.2%), and self-efficacy level (n=5, 20.8%). One study collected electrodermal activity [52].	Self-reports of emotion or affect (n=15, 65.2%), behavior/recent activity (n=8, 34.8%), self-efficacy (n=5, 21.7%), cognitive state (n=5, 21.7%), goals progress (n=4, 17.4%). Passive collection included geolocation (n=4, 17.4%), device usage (n=2, 8.7%), and electrodermal activity [52].	Self-reported behavior (n=10, 71.4%), goals progress (n=7, 50%), clinical symptoms (n=5, 35.7%), triggers or high-risk situations (n=4, 28.6%). Passive collection included sedentary time (n=4, 28.6%), step count (n=4, 28.6%), geolocation (n=3, 21.4%), and food purchase history (n=1) [98].
Decision points	Daily (n=14, 58.3%), event-triggered (n=12, 50%), fixed scheduled time (n=9, 37.5%), weekly (n=8, 33.3%), and on-demand/user-initiated (n=6, 25%).	Daily (n=19, 82.6%), fixed scheduled time (n=15, 65.2%), event-triggered (n=11, 47.8%), on-demand/user-initiated (n=9, 39.1%), and weekly (n=7, 30.4%).	Daily (n=10, 71.4%), fixed scheduled time (n=7, 50%), weekly (n=6, 42.9%), event-triggered (n=5, 35.7%), and on-demand/user-initiated (n=2, 14.3%).

In AYA JITAIs targeting substance use, motivational approaches were most frequently applied (n=8, 33.3%), informing intervention components on self-reflection and dynamic goal-setting [43,45,47,50,52,53,57,59,61]. Cognitive-behavioral approaches were also frequently applied (n=6, 25%), such as informing coping prompts for managing substance cravings and related emotional distress that may trigger use [44,52,54,55,105]. Several JITAIs for AYA substance use also adopted the transtheoretical model (n=4, 16.7%), which guided the tailoring of feedback specific to an individual's stage of behavior change [40,53,56,59]. Among JITAIs focused on AYA mental health, most (n=8, 34.8%) also leveraged cognitive behavioral approaches to inform intervention content around cognitive restructuring and stress-related coping strategies [65,70,75,76,78,85,86]. JITAIs for AYA mental health also consisted of supportive messages based on emotion regulation [67,84,85], motivational interviewing [67,76,86], mindfulness-based strategies [71,77,85], and transdiagnostic approaches [69,70,72] during heightened stress events (n=3, 13% for each respective approach). Similarly, during individually challenging moments, compassion-focused JITAIs (n=2, 8.7% within this domain) delivered content aimed at promoting emotional resilience and self-kindness [64,66,69].

For JITAIs promoting AYA physical health, self-regulation and control-based approaches were most frequently adopted (n=6, 42.9%) to tailor motivational prompts that emphasize autonomy and self-efficacy to encourage sustained health behaviors [89,91,95-97]. Around 21.4% (n=3) of JITAIs for AYA physical health also adopted the Behavior Change Wheel to guide structured prompts and reinforcement strategies, such as reminders to reduce sedentary behavior and promote physical activity, including supplementation with gamified reward systems [94,95,97]. Among JITAIs focused specifically on managing physical symptoms of chronic health conditions, frameworks such as self-regulation theory were used to structure feedback mechanisms to address adherence barriers, such as for asthma management [102].

XSL•FO
RenderX

There is an overall trend of basing adaptive content on theories related to promoting AYAs' levels of motivation and self-efficacy, cognitive and behavioral change states, although the specific application of these principles varies by context. For example, cognitive-behavioral approaches can be broadly applied to aid in managing cravings in substance use in addition to cognitive restructuring in mental health, while also supporting adherence behaviors in chronic pain management. Behavioral intent and goal-setting also inform intervention content, with JITAIs commonly adapting their support based on AYA-defined goals and progress [63,89,91,96,97,103]. This significant variability underscores the need to understand how the adaptive selection of these dynamic interventions, such as timing, content, and delivery, align with ongoing AYA contexts, which is the focus of the next section.

Tailoring Variables, Decision Rules, and Decision Points

Tailoring mechanisms within AYA JITAIs are primarily informed by behavioral, psychological, and other contextual data (Table 2). Further, the tailoring variables were aligned with the proximal and distal outcomes of the interventions. For example, behavioral data, such as self-reported nutritional consumption and physical activity levels, were used to prompt personalized feedback and inform the content of messages [91,95,96]. Meanwhile, psychological variables, such as fluctuations in mood and self-efficacy, informed JITAIs aimed at improving AYAs' emotional regulation and reducing negative affect [65,67,69,70,76]. Across these approaches, contextual awareness is emphasized, with JITAIs tailoring content based on individual and environmental factors.

Among AYA JITAIs addressing substance use and mental health, most tailoring variables rely on self-reported data, such as mood, stress, and alcohol consumption. Substance use interventions predominantly tailored content to self-reports of substance use (eg, frequency, quantity, cravings, and triggers; n=16, 66.7%); additional tailoring drew upon motivational states (n=15, 62.5%) as well as high-risk social or situational contexts (eg, location, time of week, and social setting; n=11, 45.8%). One study collected electrodermal activity to determine high-risk stress states requiring intervention to prevent drinking [52]. Mental health–focused JITAIs are tailored primarily to self-reported affective states (n=15, 65.2%), such as mood, stress, and negative thinking. Some studies also tailored support based on self-reported self-efficacy and coping needs (n=5, 21.7%, respectively). Passive collection included geolocation (n=4, 17.4%), device usage (n=2, 8.7%), and electrodermal activity in one study to determine states of stress [52].

In contrast to JITAIs for AYA substance use or mental health, over half of JITAIs for AYA physical health tailored intervention delivery based on passive activity data. These include tailoring based on step counts (n=4, 28.6%) and sedentary time (n=4, 28.6%), including via sensing data using Bluetooth geofencing (n=3, 21.4%). This reflects a trend toward real-time, event-triggered decision points in JITAIs, characterized by interventions that are activated based on immediate sensor data, such as detecting physical inactivity [92,93,97]. Notwithstanding, self-reported data were also frequently leveraged in JITAIs for AYA physical health, including for monitoring behavior (n=10, 71.4%), chronic disease symptoms (n=5, 35.7%), and goal progress (n=7, 50%).

In terms of decision points, JITAI delivery is characterized by the use of daily or multiple times daily surveys to collect these tailoring variables [89,96,99-102]. These decision points enabled AYAs to prompt immediate interventions adapted based on the self-reported data. Across all AYA JITAIs, these mechanisms can also be broadly characterized as following threshold-based triggers, where responses are activated if predefined conditions are met, such as reaching a specific stress level or failing to meet activity goals [43,52,54,55,64,65,67,69,71,79,97,99,101]. This approach allowed for adaptive interventions to be activated after consistent patterns of inactivity, poor mental health, or unmet goals. Such mechanisms were further enabled through the incorporation of high-frequency, scheduled prompts, including multiple daily check-ins or morning and evening notifications [48,50,68,71,79,96,99,102,105].

Some studies also allowed for AYA-initiated check-ins to promote flexibility in need-based and real-time support [40,52,58,68,69,74,86]. Meanwhile, in studies that used behavioral monitoring, interventions were characterized by fixed-schedule decision points, which were based on predefined intervals [53,65,79,81,89,96,100,102,105]. Studies also tailored intervention schedules based on participant-defined or behaviorally relevant times, such as following peak drinking days [43,47,105]. Time of day and spatial cues have also been leveraged to refine the timing and delivery of interventions, with methods including geofencing used to automatically identify high-risk locations and detect sedentary behavior to prompt tailored notifications [46,74,97,105,106].

Since 2024, JITAIs for AYAs have leveraged artificial intelligence to enhance the adaptability of interventions. One study developed a JITAI to prevent smartphone overuse, introducing a friction-based intervention to disrupt access to blacklisted apps when overuse was detected. The authors used a machine learning model, which was updated daily with passive sensing data such as app usage, activity, social context, location, and temporal patterns to predict overuse, triggering interventions upon app launches and task-switching moments [87]. Another recent study leveraged large language models (LLMs) for dynamic adaptability, also with the goal of reducing problematic smartphone use [88]. Similarly, when AYA users opened blacklisted apps, the LLM generated persuasive messages using 4 communication strategies of understanding, comforting, evoking, and scaffolding. LLM messages were also aligned with tailoring variables, which included real-time usage, context, mental states, and user-defined goals related to app use.

Ethical Considerations in JITAIs for AYAs

A critical gap emerges in the lack of detailed reporting on AYA engagement across studies, limiting formal analysis with the tailoring mechanisms identified in prior sections. AYA's engagement in the development of JITAI components demonstrates considerable variation in both depth and rigor. Certain studies integrated the participants directly into the design process [57,59,95,102], while others consulted AYAs with demographic profiles similar to the intended intervention

XSL•FO
RenderX

audience [39,77,83]. Such engagement is frequently framed as a strategy to enhance intervention relevance and acceptability by involving adolescents in consulting on the relevance of content. However, the extent of AYA participation in these studies is also inconsistently reported, particularly within RCTs, which often prioritize the evaluation of predeveloped interventions over documentation of co-design processes. In contrast, qualitative studies focused on describing processes of adolescent engagement in JITAI component development explicitly.

While unreported involvement in these interventions is also possible, the absence of such documentation limits our ability to assess their impact on JITAI tailoring mechanisms. Moreover, among studies claiming AYA involvement, the depth of engagement was frequently superficial. It was unclear whether AYAs had substantive influence over core design decisions regarding decision rules, except for among the small sample of qualitative studies. For example, one study engaged AYAs exiting foster care in co-designing a JITAI for substance use reduction [53]. Through focus groups with these AYAs, participants provided iterative feedback on intervention structure, language, message frequency, and content relevance, which led to key modifications in JITAI content, including adjustments to message tone, content specificity, and delivery timing. Similarly, one study involved Dutch lower-educated AYAs in the iterative development of JITAI content, including informing modifications to text clarity, visual elements, navigation, and notification settings [61]. Another study involved ethnic minority adolescent females in developing a Build Your Own Theme Song app [75]. Adolescent females contributed to JITAI design by creating personal theme songs through rewriting and recording lyrics based on their experiences. Subsequently, the app would prompt adolescents to engage with their theme song during moments of negative thinking, providing a personalized coping mechanism based on their own input.

Ethical considerations were rarely addressed in detail for specific intervention components or data collection mechanisms. A minority described explicit strategies that were primarily limited to study design, such as providing devices or data plans [44,52,85], using data encryption [99], or advising participants on secure message deletion [65]. However, reporting on the ethical implications of tailoring mechanisms was notably absent. No studies discussed mechanisms for data sovereignty or the granularity of consent regarding how distinct personal data might inform algorithmic and intervention decision-making. Despite the increasing use of passive sensing and machine learning [87], there was a lack of transparency regarding how these automated decisions were explained to adolescent users. This obscures potential risks related to privacy and autonomy, highlighting an urgent need for standardized ethical reporting guidelines in AYA digital health.

Discussion

Overview of Findings

This systematic review identified key patterns in how JITAIs for AYAs use dynamic tailoring mechanisms, decision rules, and adaptive data sources to personalize health interventions for improving substance use, mental health, and physical health (including chronic disease) outcomes. Most JITAIs relied on self-reported data, with growing trends toward integrating passive and sensor-based data, such as geofencing, app usage, and physiological monitoring [65,74,85,87,89]. Decision rules were frequently based on threshold triggers or scheduled intervals, though emerging studies have begun to explore LLMs and machine learning models for real-time, adaptive interventions [87,88].

The dominance of self-report behavioral data in JITAIs also suggests limited consideration of broader contextual and social determinants of health in their tailoring mechanisms, even when studies were piloted specifically with underserved communities [49,60,61,75,85,89]. Moreover, while our risk of bias appraisal assessed participant retention across studies, JITAI-specific data missingness also includes ecological momentary assessment nonresponse, sensor data loss, or selective absence of data at clinically important moments [107,108], which likely affects outcome measurement quality as well as intervention delivery. It is crucial for future work to investigate the incorporation of both social and missingness factors in JITAI tailoring to enhance contextual relevance and better address barriers to equitable health support among diverse individuals [27,109,110].

Further, JITAI mechanisms were highly heterogeneous, and confidence in the evidence base was constrained by inconsistent reporting and heterogeneous study designs. Further, while specific outcomes have yielded positive effects, most outcomes exhibit variable or limited findings [44,46,80,90,91,105]. Taken together, these patterns suggest the need for further investigation into the methodological determinants of intervention success, particularly regarding how JITAI and study design influence observed effectiveness.

Limitations of Existing JITAI Study Designs

It is crucial to interpret current observations of effectiveness in this review, including limited effects, with significant caution. The majority of current evidence for AYA JITAIs comes from small, pilot studies. Comparison groups were often absent or weak, and attrition or response burden was common. Several authors also explicitly noted that observed improvements could not be disentangled from self-monitoring effects, baseline reporting, treatment as usual, in-person counseling, or other concurrent influences [50,55,64]. Meanwhile, while a substantial number of studies were randomized, the risk of bias assessment highlighted recurrent challenges such as limited blinding, unclear allocation concealment, and nonidentical intervention administration. Heterogeneity in comparators also reduces confidence that the current observed effects can be attributed solely to the JITAIs being tested [111].

Given these limitations, the literature may benefit less from further RCTs but more from optimization-oriented trials, particularly microrandomized trials (MRTs), which can evaluate which intervention components work, at what moments, and for what specific AYAs' daily life contexts. MRTs, in contrast to conventional RCTs, randomize intervention options at multiple and repeated decision points to help precisely determine the effectiveness of individual adaptive components across

XSL•FO
RenderX

diverse individual contexts [112]. Such an approach can enable "a holistic comprehension of intervention dynamics over various timescales" [109]. We posit that even randomized JITAI study designs may be unable to sufficiently detect within-person variability or isolate momentary effects over time and may therefore underestimate or misrepresent the effectiveness of JITAI. To deliver precise JITAIs to AYAs, it may be important to assess diverse contextual factors that may influence the effectiveness of different adaptive components, as this can help understand "when and in what contexts a particular intervention option should be delivered"—a key differentiator of JITAIs from other digital health interventions [112]. In summary, MRTs can more precisely enable the measurement of the specific tailoring variables, decision points, and intervention options that are optimal for different individuals and health outcomes [109]. MRTs can also help to estimate proximal effects even under variable engagement and contextual missingness.

Further, for qualitative studies, the risk of bias assessment highlighted that many authors did not sufficiently consider researcher positioning and reflexivity [113]. AYA's participation in JITAI development was often limited to feedback on usability or content refinement rather than shared involvement in intervention design. We also observed inconsistent considerations of privacy, consent, and ethical safeguards in AYA JITAI studies. These findings highlight critical gaps in ensuring that JITAIs align meaningfully with AYA contexts and needs alongside critical ethical standards [114].

JITAIs for AYAs: Preliminary Recommendations for Future Directions

Overview

To address limitations of existing JITAI study designs, we sought interpretive feedback from AYAs regarding our current evidence base (refer to Methods section for details of the process). Their contributions provide a complementary perspective on the evidence, highlighting gaps and priorities for future JITAI research from their perspectives as target users.

Despite current JITAIs incorporating frequent rule-based adaptations, AYAs we consulted showed a preference for simplicity and user-controlled customization of intervention options. They highlighted how current JITAI tailoring mechanisms are overly complex or opaque. They also mentioned the lack of modular designs, whereby users can select which features they wish to engage with to enhance contextual relevance. Ongoing notifications, which characterize many of the JITAIs in this review [39,53,68,79,89,96,97], were criticized for potentially contributing to notification fatigue, which may limit engagement over time. For example, our AYA consultants discussed how excessive prompts can lead to desensitization, frustration, and eventual disregard of JITAIs. Subtle, passive engagement strategies, such as visual cues or widgets that update in the background of devices and can be checked flexibly, were highlighted as a potential alternative [115]. This raises the possibility that high-frequency prompting may undermine engagement or perceived usefulness, particularly among AYAs already experiencing digital overload and other digital determinants of poor health [116,117]. Our AYA consultants also raised concerns about data privacy and lack of control over what information is collected, advocating for opt-in tailoring mechanisms that allow users to choose which data they are comfortable sharing [118].

This feedback from our AYA consultants regarding our current evidence base highlights a potential tension within current JITAI development and trends toward passive sensing and multimodal integration with artificial intelligence [119]. In an effort to enhance engagement and precision, these mechanisms may also risk reducing engagement or acceptability for some AYAs [120]. Our AYA co-authors described this phenomenon as an "engagement paradox"—a concept emerging from the divergence between the high-frequency adaptability reported in the literature and the desire for autonomy expressed by our AYA consultants. This may indicate that the success of adaptive interventions depends not only on the incorporation of multiple levels of tailoring but also on how well tailoring mechanisms align overall with AYAs' ongoing preferences and needs for autonomy [121,122]. Despite automation's role in enhancing adaptation, our AYA consultants suggested that excessive automation may reduce AYAs' engagement if the tailoring mechanisms lack transparency and explainability. As such, adaptive systems in which AYAs can define their preferred tailoring mechanisms may be a worthwhile direction for future precision health interventions. Transparent data practices, which can promote participant control over tailoring mechanisms, may also be essential to privacy considerations as well as intervention relevance [123,124].

Notwithstanding, both the AYAs we consulted and engaged in this review as co-authors were limited in number and may not represent other AYAs' perspectives. Substantial further participatory research is necessary to improve the precision and relevance of JITAIs for diverse AYA lived experiences [109]. To provide a foundation for such work, we present preliminary recommendations for future directions in investigations of AYA JITAIs.

Investigate Mechanisms of Meaningful Engagement in Real-World Contexts

Understanding the mechanisms by which JITAIs influence behavioral outcomes is critical for advancing the field [16,109]. The review highlighted uncertainties regarding how adaptive mechanisms operate and influence engagement and outcomes, alongside key limitations of existing JITAI study designs. To address these challenges, future research can use optimization trial designs [112], such as microrandomized or sequential multiple assignment trials, to systematically test how specific tailoring variables, such as timing, content format, and data type, as well as theoretical frameworks, may impact engagement and behavioral outcomes. This can enable a more precise understanding of how different intervention components and decision rules perform and help clarify the tailoring mechanisms that are most effective for specific AYA contexts [109].

Further, more meaningful co-design in creating relevant, effective interventions is a crucial next step [125]. The review revealed limited AYA involvement in the design of many JITAIs, resulting in interventions that may not fully align with users' lived experiences or preferences [110]. Insights from our

XSL•FO
RenderX

consultation with AYAs reinforced this, highlighting the value of relevance and understandability in design. There is a need to embed precision health designs within real-life contexts to ensure interventions are adaptable and impactful for ongoing AYA health needs [109,110,126]. Participatory workshops can facilitate meaningful input, enabling AYAs to contribute ideas about preferred tailoring mechanisms. For example, studies can engage adolescent participants in informing the proximal outcomes, intervention options, tailoring variables, decision points, as well as the overall user experience of JITAIs [53]. Integration of participatory design components to additionally inform, as well as interpret, individual outcomes in ongoing MRTs would enable a more comprehensive understanding of what works, for whom, and why [127]. Such approaches would help acknowledge AYAs as experts by experience, resulting in interventions that are more grounded in real-life contexts [128]. In this manner, AYAs can directly participate in shaping mechanisms in precision health interventions.

Continuous Adaptation Directed by Adolescents and Young Adults

Adaptation is a cornerstone of effective JITAI design [16], yet the review findings identified notable gaps and inconsistencies in how such adaptation is actually implemented. Further, our AYA consultants, in reviewing our evidence base, expressed alternative preferences for systems that allowed for greater control and customization rather than opaque, automated adaptations. To address these concerns, future JITAIs can incorporate modular, user-directed adaptation features that empower adolescents to adjust the level and type of adaptation as their preferences and needs evolve [123,129]. For instance, intervention component dashboards could be developed, allowing AYAs to select their preferred notification frequencies, data sources for tailoring, and engagement formats, such as text, video, or visual cues [53,61,130,131]. Such affordances for ongoing participatory adaptation may help promote AYA agency while fostering enhanced trust and relevance, ensuring that JITAI mechanisms are both transparent and AYA-centered. However, we acknowledge that such open customization introduces heterogeneity in intervention exposure, potentially complicating the assessment of internal validity [132]. To address this, we again recommend that future research use optimization trial designs, such as MRTs, which are uniquely capable of handling within-person variability [133]. By treating user-defined preferences as dynamic tailoring variables within the statistical model, researchers may be able to better understand the mechanism of such autonomy, separate from the effects of the content itself.

Minimizing Engagement Fatigue While Promoting More Meaningful Engagement

The systematic review identified inconsistent outcomes and effects among current JITAIs for AYAs, which may be in part attributable to the methodological heterogeneity or lack of consistent theoretical grounding across studies. Further, based on the interpretive feedback from AYAs we consulted, we posit that low user engagement may also be driven by "notification fatigue" [134]—a phenomenon they identified as a critical barrier to sustained use. While the empirical studies in our review rarely measured fatigue explicitly, AYAs we consulted criticized frequent and intrusive notifications for contributing to disengagement. To address these challenges, future intervention development might explore alternative, less intrusive engagement strategies [135]. For example, AYA JITAIs might explore replacing frequent text notifications with visual nudges or contextual prompts [61]. Interventions could also explore integrating prompts into commonly used platforms, such as social media feeds, to ensure engagement occurs within familiar and comfortable digital spaces to enhance relevance to AYAs [136,137].

Strengthening Ethical Standards for Transparent Personalization and Digital Self-Determination

The review identified significant gaps in ethical reporting, with most studies neglecting to mention ethical considerations in JITAI design mechanisms. Such considerations are particularly critical given JITAIs, which rely on sensitive data collection and passive monitoring [138]. Such an approach also aligns with principles of digital self-determination, which emphasize the reduction of information asymmetries by ensuring adolescents have control over their data and tailoring mechanisms [130]. Our AYA consultants also expressed discomfort with unclear or opaque data practices of current JITAIs and emphasized a strong preference for opt-in mechanisms that allowed them to control the type and extent of data they shared. In this context, a key consideration in future JITAI development may include the need to promote AYAs' individual agency over their digital data and enable participatory tailoring mechanisms [123,139].

To address these concerns, dynamic consent processes within ongoing JITAI tailoring mechanisms may be one promising approach. Dynamic consent refers to when individuals are continuously informed and able to manage their data-sharing preferences over time [104]. Such a practice can enhance data sovereignty, defined as the need for individuals to have control over the use of their shared data, without which JITAI usage may be significantly hindered [104]. These processes can provide clear, accessible information about what data are collected, how the data are used, and the purpose of each data point in tailoring mechanisms. AYAs can be further empowered to opt in or opt out of specific data types and be given the opportunity to modify data points over time. This can be operationalized through user-friendly privacy dashboards which enable real-time adjustment of intervention-tailoring settings [104,140].

Dynamic consent in AYA digital health remains very limited. Recent studies highlight the approach's alignment with technology users' desire for transparency and control [104]. Studies have also shown how dynamic consent preferences are highly individual based on demographic and social contexts [131], making the approach particularly salient for application in precision adolescent and digital health contexts. Based on our AYA consultation, we posit that increased emphasis on self-determination and sovereignty in JITAI design may help improve their alignment and relevance with adolescent health needs and expectations.

XSL•FO
RenderX

Strengths and Limitations

This review is the first to focus on JITAIs across a wide range of health conditions within a specific developmental period, highlighting the unique needs and intervention contexts for adolescents, who face distinct physical, mental, and social health challenges that differ from those of children or adults. A key strength of this review is the meaningful involvement of AYAs in screening, data extraction, and critical interpretation, which enhances the relevance and actionability of the findings and future recommendations. The current synthesis of common theoretical frameworks and tailoring mechanisms also represents an important first step toward a more coherent body of evidence for the application of JITAIs in AYA health. However, due to substantial heterogeneity in intervention designs, high risk of bias (particularly with outcome reliability and statistical validity), and inconsistent outcome effects observed, meta-analysis was not feasible. Additionally, despite a broad search strategy, some relevant studies may have been inadvertently missed, as most included interventions did not explicitly describe their mechanisms as just-in-time or adaptive.

Conclusion

This review provides the first comprehensive synthesis of how JITAI mechanisms are being designed and studied for AYAs across multiple health domains, while also integrating AYA perspectives into the interpretation of findings and future directions. Beyond showing where JITAIs are being applied, it highlights important limitations in the current evidence base, including heavy reliance on self-reported behavioral data and limited attention to broader social determinants in tailoring mechanisms. The review also highlights substantial methodological heterogeneity among studies evaluating similar outcomes, as well as insufficient ethical transparency, including limited participatory engagement to ensure intervention designs are grounded in real-life AYA contexts. Our findings suggest that progress in AYA precision health may depend not only on developing more adaptive interventions but on measuring and designing them in ways that are more transparent and participatory. Limitations of traditional study designs, including measurement of JITAI administration in RCTs, point to the need for more robust designs such as MRTs as a key methodological next step for identifying which adaptive components work, at what moments, and in which daily life contexts. Based on AYA coauthor and consultant input, we further outline preliminary recommendations for future research to help ensure JITAIs better align precision with AYA agencies, including through co-designed tailoring mechanisms, reduced notification burden, and clearer control over personal data and automated decision-making.

Acknowledgments

We thank Stella Goeschl, Anastasia Kluge, and Benedetta Vallese for their valuable input, which informed the development of the paper. We also thank Dirk Jan Ligtenbelt, who provided guidance on the search strategy. Parts of this work were presented at the Karolinska Institutet–United Nations Children's Fund Joint Conference on Global Child and Adolescent Mental Health 2024, the 19th European Association for Research on Adolescence Conference, the Society for Research on Adolescence 2026 Biennial Meeting, and the Social Data Science research group seminar at the Alan Turing Institute. We thank attendees for their thoughtful reception and feedback on earlier versions of this work. The lead author used ChatGPT solely for limited editorial support during final paper preparation, specifically grammar correction and table formatting. ChatGPT was not used for review design, data analysis, interpretation, ideation, reference generation, or any substantive writing. No artificial intelligence (AI)–generated content was included in the final paper. All outputs were reviewed and verified by the authors, who remain fully accountable for all aspects of the paper.

Funding

Authors KWG, CAF, LK, CRS, AV, and EvR are collaborators in the Convergence Flagship program, Proactive Technology-Supported Prevention and Mental Health in Adolescence (PROTECt ME), which is funded by the Health and Technology Flagship program as part of Convergence, the research alliance between Erasmus University Medical Center, Erasmus University Rotterdam, and Delft University of Technology. Additionally, KWG, CAF, and MdR are supported through the "High Tech for a Sustainable Future" capacity building program of the 4TU Federation in the Netherlands. The 4TU Federation is not involved in the development of this review or interpretation of its findings.

Data Availability

All data extracted and reported during this review are included in this published paper and its additional information files.

Authors' Contributions

KWG is the review guarantor and contributed to conceptualization, project administration, methodology, investigation, formal analysis, writing—original draft, and writing—review and editing. MA contributed to investigation, formal analysis, writing—original draft, and writing—review and editing. VC, IK, SG, CI, RA, CA, and ET each contributed equally to investigation and writing—review and editing. CRS, AV, and EvR contributed to methodology and writing—review and editing. LK contributed to funding acquisition, methodology, and writing—review and editing. MdR contributed to funding acquisition, supervision,

XSL•FO
RenderX

methodology, and writing—review and editing. CAF contributed to funding acquisition, supervision, project administration, methodology, and writing—review and editing.

Conflicts of Interest

KWG, MA, SG, VC, and CI are members of the World Health Organization Regional Office for Europe's Youth4Health network. However, their views do not represent the World Health Organization or other network members. ET is employed by Gro-up, a nonprofit organization in the Netherlands promoting the well-being of children and youth and their families. Gro-up has no involvement in the development of this review or interpretation of its findings.

Multimedia Appendix 1
Data extraction for just-in-time adaptive interventions (JITAIs) for adolescent and young adult (AYA) substance use.
[PDF File, 277 KB - jmir_v28i1e84422_app1.pdf]

Multimedia Appendix 2
Data extraction for just-in-time adaptive interventions (JITAIs) for adolescent and young adult (AYA) mental health.
[PDF File, 279 KB - jmir_v28i1e84422_app2.pdf]

Multimedia Appendix 3
Data extraction for just-in-time adaptive interventions (JITAIs) for adolescent and young adult (AYA) physical health.
[PDF File, 224 KB - jmir_v28i1e84422_app3.pdf]

Checklist 1
PRISMA-S checklist.
[PDF File, 99 KB - jmir_v28i1e84422_app4.pdf]

Checklist 2
SWiM checklist.
[PDF File, 170 KB - jmir_v28i1e84422_app5.pdf]

References

1. Chahal R, Gotlib IH, Guyer AE. Research review: brain network connectivity and the heterogeneity of depression in adolescence - a precision mental health perspective. J Child Psychol Psychiatry 2020 Dec;61(12):1282-1298. [doi: 10.1111/jcpp.13250] [Medline: 32458453]
2. Ryder JR, Kaizer AM, Jenkins TM, Kelly AS, Inge TH, Shaibi GQ. Heterogeneity in response to treatment of adolescents with severe obesity: the need for precision obesity medicine. Obesity (Silver Spring) 2019 Feb;27(2):288-294. [doi: 10.1002/oby.22369] [Medline: 30677258]
3. Kerr J, Anderson C, Lippman SM. Physical activity, sedentary behaviour, diet, and cancer: an update and emerging new evidence. Lancet Oncol 2017 Aug;18(8):e457-e471. [doi: 10.1016/S1470-2045(17)30411-4] [Medline: 28759385]
4. Fleary SA, Joseph P, Pappagianopoulos JE. Adolescent health literacy and health behaviors: a systematic review. J Adolesc 2018 Jan;62:116-127. [doi: 10.1016/j.adolescence.2017.11.010] [Medline: 29179126]
5. Cicchetti D, Rogosch FA. A developmental psychopathology perspective on adolescence. J Consult Clin Psychol 2002;70(1):6-20. [doi: 10.1037/0022-006X.70.1.6]
6. Paquin V, Muckle G, Bolanis D, et al. Longitudinal trajectories of food insecurity in childhood and their associations with mental health and functioning in adolescence. JAMA Netw Open 2021 Dec 1;4(12):e2140085. [doi: 10.1001/jamanetworkopen.2021.40085] [Medline: 34928352]
7. Adolescents. UNICEF. URL: https://data.unicef.org/topic/adolescents/overview/ [accessed 2025-03-25]
8. Johnson MH, Fearon P, Pickles A, Jones EJH. Editorial perspective: the paradox of precision health in early development – building large samples to yield individual‐level measures. Child Psychology Psychiatry 2024 Jul;65(7):991-994. [doi: 10.1111/jcpp.13974]
9. Shen MD, Chen SB, Ding XD. The effectiveness of digital twins in promoting precision health across the entire population: a systematic review. NPJ Digit Med 2024 Jun 3;7(1):145. [doi: 10.1038/s41746-024-01146-0] [Medline: 38831093]
10. Sawyer SM, Azzopardi PS, Wickremarathne D, Patton GC. The age of adolescence. Lancet Child Adolesc Health 2018 Mar;2(3):223-228. [doi: 10.1016/S2352-4642(18)30022-1]
11. Wyatt TH, Bayless AK, Krauskopf P, Gaylord N. Using mHealth applications to promote self-managed health behaviors among teens. J Pediatr Nurs 2021;59:164-172. [doi: 10.1016/j.pedn.2021.04.025] [Medline: 33932646]
12. Hightow-Weidman LB, Horvath KJ, Scott H, Hill-Rorie J, Bauermeister JA. Engaging youth in mHealth: what works and how can we be sure? Mhealth 2021;7:23-23. [doi: 10.21037/mhealth-20-48]

XSL•FO
RenderX

13. Psihogios AM, King-Dowling S, O'Hagan B, et al. Contextual predictors of engagement in a tailored mHealth intervention for adolescent and young adult cancer survivors. Ann Behav Med 2021 Nov 18;55(12):1220-1230. [doi: 10.1093/abm/kaab008] [Medline: 33674863]
14. Kozelka EE, Acquilano SC, Al-Abdulmunem M, et al. Digital mental health and its discontents: assumptions about technology that create barriers to equitable access. Psych Serv 2024 Mar 1;75(3):299-302. [doi: 10.1176/appi.ps.20230238]
15. Villegas Mejía C, Remmerswaal D, Engels RCME, Ludden GDS, Boffo M. Macro-engagement in mHealth: exploring user engagement beyond the screen. Digit Health 2024;10:20552076231225591. [doi: 10.1177/20552076231225591] [Medline: 38269371]
16. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462. [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
17. Crone EA, Dahl RE. Understanding adolescence as a period of social–affective engagement and goal flexibility. Nat Rev Neurosci 2012 Sep;13(9):636-650. [doi: 10.1038/nrn3313]
18. Hurrelmann K, Richter M. Risk behaviour in adolescence: the relationship between developmental and health problems. J Public Health 2006 Feb;14(1):20-28. [doi: 10.1007/s10389-005-0005-5]
19. Hardeman W, Houghton J, Lane K, Jones A, Naughton F. A systematic review of just-in-time adaptive interventions (JITAIs) to promote physical activity. Int J Behav Nutr Phys Act 2019 Apr 3;16(1):31. [doi: 10.1186/s12966-019-0792-7] [Medline: 30943983]
20. Oikonomidi T, Ravaud P, LeBeau J, Tran VT. A systematic scoping review of just-in-time, adaptive interventions finds limited automation and incomplete reporting. J Clin Epidemiol 2023 Feb;154:108-116. [doi: 10.1016/j.jclinepi.2022.12.006] [Medline: 36521653]
21. Perski O, Hébert ET, Naughton F, Hekler EB, Brown J, Businelle MS. Technology-mediated just-in-time adaptive interventions (JITAIs) to reduce harmful substance use: a systematic review. Addiction 2022 May;117(5):1220-1241. [doi: 10.1111/add.15687] [Medline: 34514668]
22. Teepe GW, Da Fonseca A, Kleim B, et al. Just-in-time adaptive mechanisms of popular mobile applications for individuals with depression: systematic app search and literature review. J Med Internet Res 2021 Sep 28;23(9):e29412. [doi: 10.2196/29412] [Medline: 34309569]
23. Wang L, Miller LC. Just-in-the-moment adaptive interventions (JITAI): a meta-analytical review. Health Commun 2020 Nov;35(12):1531-1544. [doi: 10.1080/10410236.2019.1652388] [Medline: 31488002]
24. Koike C, Nestor BA, Gelmez PS, Fladger A, Shrier LA, Kossowsky J. Just-in-time adaptive interventions for adolescent and young adult substance use: a systematic review. J Adolesc Health 2025 Dec;77(6):1038-1049. [doi: 10.1016/j.jadohealth.2025.07.027] [Medline: 41065631]
25. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ 2020 Jan 16;368:l6890. [doi: 10.1136/bmj.l6890] [Medline: 31948937]
26. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ :n71. [doi: 10.1136/bmj.n71]
27. Guan KW, Adlung C, Keijsers L, et al. Just-in-time adaptive interventions for adolescent and young adult health and well-being: protocol for a systematic review. BMJ Open 2024 Jul;14(7):e083870. [doi: 10.1136/bmjopen-2024-083870]
28. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021;10(1). [doi: 10.1186/s13643-020-01542-z]
29. Timms KP, Rivera DE, Collins LM, Piper ME. Control systems engineering for understanding and optimizing smoking cessation interventions. Proc Am Control Conf 2013:1964-1969. [doi: 10.1109/acc.2013.6580123] [Medline: 24362946]
30. Kumar S, Nilsen W, Pavel M, Srivastava M. Mobile health: revolutionizing healthcare through transdisciplinary research. Computer (Long Beach Calif) 2013 Jan;46(1):28-35. [doi: 10.1109/MC.2012.392]
31. Chow PI, Smith J, Saini R, et al. A novel just-in-time intervention for promoting safer drinking among college students: app testing across 2 independent pre-post trials. JMIR Hum Factors 2025 Apr 10;12:e69873. [doi: 10.2196/69873] [Medline: 40209208]
32. Blum RW, Bastos F, Kabiru CW, Le LC. Adolescent health in the 21st century. Lancet 2012 Apr 28;379(9826):1567-1568. [doi: 10.1016/S0140-6736(12)60407-3] [Medline: 22538177]
33. Jaworska N, MacQueen G. Adolescence as a unique developmental period. J Psychiatry Neurosci 2015 Sep;40(5):291-293. [doi: 10.1503/jpn.150268] [Medline: 26290063]
34. Arnone JM. Adolescents may be older than we think: today 25 is the new 18, or is it? Int J Celiac Dis 2014 Jan 5;2(2):47-48. [doi: 10.12691/ijcd-2-2-4]
35. Majmundar A, Cerrada C, Fang W, Huh J. In lieu of smoking: are smokers more likely to enact certain types of implementation intention plans during a jit smoking cessation? IntJ Behav Med 2020 Oct;27(5):556-564. [doi: 10.1007/s12529-020-09893-6]
36. Barker TH, Stone JC, Sears K, et al. Revising the JBI quantitative critical appraisal tools to improve their applicability: an overview of methods and the development process. JBI Evid Synth 2023 Mar 1;21(3):478-493. [doi: 10.11124/JBIES-22-00125] [Medline: 36121230]

XSL•FO
RenderX

37. Prasad M. Introduction to the GRADE tool for rating certainty in evidence and recommendations. Clin Epidemiol Glob Health 2024 Jan;25:101484. [doi: 10.1016/j.cegh.2023.101484]
38. Murad MH, Mustafa RA, Schünemann HJ, Sultan S, Santesso N. Rating the certainty in evidence in the absence of a single estimate of effect. Evid Based Med 2017 Jun;22(3):85-87. [doi: 10.1136/ebmed-2017-110668]
39. Rodgers A, Corbett T, Bramley D, et al. Do u smoke after txt? Results of a randomised trial of smoking cessation using mobile phone text messaging. Tob Control 2005 Aug;14(4):255-261. [doi: 10.1136/tc.2005.011577] [Medline: 16046689]
40. Riley W, Obermayer J, Jean-Mary J. Internet and mobile phone text messaging intervention for college smokers. J Am Coll Health 2008;57(2):245-248. [doi: 10.3200/JACH.57.2.245-248] [Medline: 18809542]
41. Coughlin LN, Nahum-Shani I, Philyaw-Kotov ML, et al. Developing an adaptive mobile intervention to address risky substance use among adolescents and emerging adults: usability study. JMIR Mhealth Uhealth 2021 Jan 15;9(1):e24424. [doi: 10.2196/24424] [Medline: 33448931]
42. Weitzel JA, Bernhardt JM, Usdan S, Mays D, Glanz K. Using wireless handheld computers and tailored text messaging to reduce negative consequences of drinking alcohol. J Stud Alcohol Drugs 2007 Jul;68(4):534-537. [doi: 10.15288/jsad.2007.68.534] [Medline: 17568957]
43. Suffoletto B, Callaway C, Kristan J, Kraemer K, Clark DB. Text - message - based drinking assessments and brief interventions for young adults discharged from the emergency department. Alcoholism Clin Exp Res 2012 Mar;36(3):552-560. [doi: 10.1111/j.1530-0277.2011.01646.x]
44. Witkiewitz K, Desai SA, Bowen S, Leigh BC, Kirouac M, Larimer ME. Development and evaluation of a mobile intervention for heavy drinking and smoking among college students. Psychol Addict Behav 2014 Sep;28(3):639-650. [doi: 10.1037/a0034747] [Medline: 25000269]
45. Shrier LA, Burke PJ, Kells M, et al. Pilot randomized trial of MOMENT, a motivational counseling-plus-ecological momentary intervention to reduce marijuana use in youth. Mhealth 2018;4(7):29. [doi: 10.21037/mhealth.2018.07.04] [Medline: 30148142]
46. Wright C, Dietze PM, Agius PA, et al. Mobile phone-based ecological momentary intervention to reduce young adults' alcohol use in the event: a three-armed randomized controlled trial. JMIR Mhealth Uhealth 2018 Jul 20;6(7):e149. [doi: 10.2196/mhealth.9324] [Medline: 30030211]
47. Suffoletto B, Kirisci L, Clark DB, Chung T. Which behavior change techniques help young adults reduce binge drinking? A pilot randomized clinical trial of 5 text message interventions. Addict Behav 2019 May;92:161-167. [doi: 10.1016/j.addbeh.2019.01.006]
48. O'Donnell R, Richardson B, Fuller-Tyszkiewicz M, et al. Ecological momentary assessment of drinking in young adults: an investigation into social context, affect and motives. Addict Behav 2019 Nov;98:106019. [doi: 10.1016/j.addbeh.2019.06.008] [Medline: 31247534]
49. Santa Maria D, Padhye N, Businelle M, et al. Efficacy of a just-in-time adaptive intervention to promote HIV risk reduction behaviors among young adults experiencing homelessness: pilot randomized controlled trial. J Med Internet Res 2021 Jul 6;23(7):e26704. [doi: 10.2196/26704] [Medline: 34255679]
50. Shrier LA, Rhoads A, Burke P, Walls C, Blood EA. Real-time, contextual intervention using mobile technology to reduce marijuana use among youth: a pilot study. Addict Behav 2014 Jan;39(1):173-180. [doi: 10.1016/j.addbeh.2013.09.028] [Medline: 24139665]
51. Shrier LA, Rhoads AM, Fredette ME, Burke PJ. "Counselor in Your Pocket": youth and provider perspectives on a mobile motivational intervention for marijuana use. Subst Use Misuse 2014 Jan 1;49(1-2):134-144. [doi: 10.3109/10826084.2013.824470] [Medline: 24000892]
52. Leonard NR, Silverman M, Sherpa DP, et al. Mobile health technology using a wearable sensorband for female college students with problem drinking: an acceptability and feasibility study. JMIR Mhealth Uhealth 2017 Jul 7;5(7):e90. [doi: 10.2196/mhealth.7399] [Medline: 28687533]
53. Braciszewski JM, Tzilos Wernette GK, Moore RS, et al. Developing a tailored substance use intervention for youth exiting foster care. Child Abuse Neglect 2018 Mar;77:211-221. [doi: 10.1016/j.chiabu.2018.01.013]
54. Stevenson BL, Blevins CE, Marsh E, Feltus S, Stein M, Abrantes AM. An ecological momentary assessment of mood, coping and alcohol use among emerging adults in psychiatric treatment. Am J Drug Alcohol Abuse 2020 Sep 2;46(5):651-658. [doi: 10.1080/00952990.2020.1783672]
55. Blevins CE, Marsh EL, Stein MD, Schatten HT, Abrantes AM. Project CHOICE: choosing healthy options in coping with emotions, an EMA/EMI plus in-person intervention for alcohol use. Subst Abuse 2021 Oct;42(4):569-576. [doi: 10.1080/08897077.2020.1806182]
56. Kazemi DM, Borsari B, Levine MJ, et al. Effectiveness of a theory-based mHealth intervention for high-risk drinking in college students. Subst Use Misuse 2020 May;55(10):1667-1676. [doi: 10.1080/10826084.2020.1756851] [Medline: 32394772]
57. Wright CJC, Dietze PM, Crockett B, Lim MSC. Participatory development of MIDY (Mobile Intervention for Drinking in Young people). BMC Public Health 2016 Feb 24;16(1):184. [doi: 10.1186/s12889-016-2876-5] [Medline: 26911299]

XSL•FO
RenderX

58. Cerrada CJ, Dzubur E, Blackman KCA, Mays V, Shoptaw S, Huh J. Development of a just-in-time adaptive intervention for smoking cessation among Korean American emerging adults. Int J Behav Med 2017 Oct;24(5):665-672. [doi: 10.1007/s12529-016-9628-x] [Medline: 28070868]
59. Kazemi DM, Borsari B, Levine MJ, Lamberson KA, Dooley B. REMIT: development of a mHealth theory-based intervention to decrease heavy episodic drinking among college students. Addict Res Theory 2018;26(5):377-385. [doi: 10.1080/16066359.2017.1420783] [Medline: 32694964]
60. Acorda DE, Jones JT, Maria DS. Acceptability of a just-in-time adaptive intervention for HIV prevention among youth experiencing homelessness: a qualitative analysis. J Health Care Poor Underserved 2021 May;32(2S):25-42. [doi: 10.1353/hpu.2021.0049]
61. van Keulen H, Voogt C, Kleinjan M, Kramer J, Andree R, van Empelen P. Development of a dynamically tailored mHealth intervention (What Do You Drink) to reduce excessive drinking among Dutch lower-educated students: user-centered design approach. JMIR Form Res 2022 Aug 11;6(8):e36969. [doi: 10.2196/36969] [Medline: 35969428]
62. Suffoletto B, Kristan J, Chung T, et al. An interactive text message intervention to reduce binge drinking in young adults: a randomized controlled trial with 9-month outcomes. PLoS One 2015;10(11):e0142877. [doi: 10.1371/journal.pone.0142877] [Medline: 26580802]
63. Suffoletto B, Chung T, Muench F, Monti P, Clark DB. A text message intervention with adaptive goal support to reduce alcohol consumption among non-treatment-seeking young adults: non-randomized clinical trial with voluntary length of enrollment. JMIR Mhealth Uhealth 2018 Feb 16;6(2):e35. [doi: 10.2196/mhealth.8530] [Medline: 29453191]
64. Rauschenberg C, Boecking B, Paetzold I, et al. A compassion-focused ecological momentary intervention for enhancing resilience in help-seeking youth: uncontrolled pilot study. JMIR Ment Health 2021 Aug 5;8(8):e25650. [doi: 10.2196/25650] [Medline: 34383687]
65. Suffoletto B, Goldstein T, Gotkiewicz D, Gotkiewicz E, George B, Brent D. Acceptability, engagement, and effects of a mobile digital intervention to support mental health for young adults transitioning to college: pilot randomized controlled trial. JMIR Form Res 2021 Oct 14;5(10):e32271. [doi: 10.2196/32271] [Medline: 34647893]
66. Reininghaus U, Paetzold I, Rauschenberg C, et al. Effects of a novel, transdiagnostic ecological momentary intervention for prevention, and early intervention of severe mental disorder in youth (EMIcompass): findings from an exploratory randomized controlled trial. Schizophr Bull 2023 May 3;49(3):592-604. [doi: 10.1093/schbul/sbac212] [Medline: 36738168]
67. Kennard BD, Goldstein T, Foxwell AA, et al. As Safe as Possible (ASAP): a brief app-supported inpatient intervention to prevent postdischarge suicidal behavior in hospitalized, suicidal adolescents. AJP 2018 Sep 1;175(9):864-872. [doi: 10.1176/appi.ajp.2018.17101151]
68. van Aubel E, Bakker JM, Batink T, et al. Blended care in the treatment of subthreshold symptoms of depression and psychosis in emerging adults: a randomised controlled trial of Acceptance and Commitment Therapy in Daily-Life (ACT-DL). Behav Res Ther 2020 May;128:103592. [doi: 10.1016/j.brat.2020.103592] [Medline: 32146218]
69. Paetzold I, Schick A, Rauschenberg C, et al. A hybrid ecological momentary compassion–focused intervention for enhancing resilience in help-seeking young people: prospective study of baseline characteristics in the EMIcompass trial. JMIR Form Res 2022 Nov 4;6(11):e39511. [doi: 10.2196/39511] [Medline: 36331526]
70. Bell I, Arnold C, Gilbertson T, et al. A personalized, transdiagnostic smartphone intervention (Mello) targeting repetitive negative thinking in young people with depression and anxiety: pilot randomized controlled trial. J Med Internet Res 2023 Dec 13;25:e47860. [doi: 10.2196/47860] [Medline: 38090786]
71. Lucas-Thompson RG, Prince MA, Rigsby BA, et al. Preliminary evaluation of learning to BREATHE PLUS for university students: does a multi-modal adaptive supplement strengthen effects of a mindfulness-based intervention? Mindfulness 2023 Nov;14(11):2699-2715. [doi: 10.1007/s12671-023-02229-x]
72. Reininghaus U, Daemen M, Postma MR, et al. Transdiagnostic ecological momentary intervention for improving self-esteem in youth exposed to childhood adversity. JAMA Psychiatry 2024 Mar 1;81(3):227-239. [doi: 10.1001/jamapsychiatry.2023.4590] [Medline: 38019495]
73. Postma MR, Vrancken S, Daemen M, et al. Working mechanisms of the use and acceptability of ecological momentary interventions: a realist evaluation of a guided self-help ecological momentary intervention targeting self-esteem. BMC Public Health 2024 Jun 19;24(1):1633. [doi: 10.1186/s12889-024-19143-z] [Medline: 38898412]
74. Silk JS, Pramana G, Sequeira SL, et al. Using a smartphone app and clinician portal to enhance brief cognitive behavioral therapy for childhood anxiety disorders. Behav Ther 2020 Jan;51(1):69-84. [doi: 10.1016/j.beth.2019.05.002] [Medline: 32005341]
75. Neal-Barnett A, Stadulis R, Ellzey D, et al. Evaluation of the effectiveness of a musical cognitive restructuring app for black inner-city girls: survey, usage, and focus group evaluation. JMIR Mhealth Uhealth 2019 Jun 27;7(6):e11310. [doi: 10.2196/11310] [Medline: 31188130]
76. Shrier LA, Burke PJ, Parker S, et al. Development and pilot testing of a counseling-plus-mHealth intervention to reduce risk for pregnancy and sexually transmitted infection in young women with depression. Mhealth 2020;6:17. [doi: 10.21037/mhealth.2019.11.05] [Medline: 32270009]

XSL•FO
RenderX

77. Lucas-Thompson RG, Rayburn S, Seiter NS, et al. Learning to BREATHE “Plus”: a multi-modal adaptive supplement to an evidence-based mindfulness intervention for adolescents. Front Public Health 2020;8:579556. [doi: 10.3389/fpubh.2020.579556] [Medline: 33282814]
78. Shrier LA, Spalding A. “Just Take a Moment and Breathe and Think”: young women with depression talk about the development of an ecological momentary intervention to reduce their sexual risk. J Pediatr Adolesc Gynecol 2017 Feb;30(1):116-122. [doi: 10.1016/j.jpag.2016.08.009] [Medline: 27575408]
79. Marciniak MA, Shanahan L, Myin‐Germeys I, et al. Imager-a mobile health mental imagery‐based ecological momentary intervention targeting reward sensitivity: a randomized controlled trial. Applied Psych Health Well 2024 May;16(2):576-596. [doi: 10.1111/aphw.12505] [Medline: 37942875]
80. Schenkel K, Haug S, Castro RP, et al. One SMS a day keeps the stress away? A just‐in‐time planning intervention to reduce occupational stress among apprentices. Applied Psych Health Well 2022 Nov;14(4):1389-1407. [doi: 10.1111/aphw.12340]
81. van Asselt A, Roke Y. Co-creating the Stress Autism Mate (SAM): development of an mHealth application for autistic adults using a design thinking approach. Health Psychol Behav Med 2025;13(1):2597607. [doi: 10.1080/21642850.2025.2597607] [Medline: 41378280]
82. van Asselt A, Hoeberichts K, Demirel S, Scheeren A, Roke Y. Effectiveness of the mobile Stress Autism Mate Junior application in reducing stress and improving quality of life in adolescents with autism: a pilot study. Front Psychiatry 2024;15:1469257. [doi: 10.3389/fpsyt.2024.1469257] [Medline: 39845363]
83. Hiller S, Götzl C, Rauschenberg C, et al. Health-promoting effects and everyday experiences with a mental health app using ecological momentary assessments and AI-based ecological momentary interventions among young people: qualitative interview and focus group study. JMIR Mhealth Uhealth 2025 Apr 29;13:e65106. [doi: 10.2196/65106] [Medline: 40300160]
84. Engelskirchen L, Asbrand J, Tuschen-Caffier B. Feasibility and potential efficacy of an ecological momentary intervention for training acceptance and problem solving in young people. Cogn Ther Res 2025 Sep 9. [doi: 10.1007/s10608-025-10643-9]
85. Leonard NR, Casarjian B, Fletcher RR, et al. Theoretically-based emotion regulation strategies using a mobile app and wearable sensor among homeless adolescent mothers: acceptability and feasibility study. JMIR Pediatr Parent 2018;1(1):e1. [doi: 10.2196/pediatrics.9037] [Medline: 30637376]
86. Ranney ML, Pittman SK, Moseley I, et al. Cyberbullying prevention for adolescents: iterative qualitative methods for mobile intervention design. JMIR Form Res 2021 Aug 27;5(8):e25900. [doi: 10.2196/25900] [Medline: 34448702]
87. Orzikulova A, Xiao H, Li Z, et al. Time2Stop: adaptive and explainable human-AI loop for smartphone overuse intervention. Presented at: CHI ’24: CHI Conference on Human Factors in Computing Systems; May 11-16, 2024; Honolulu, HI p. 1-20. [doi: 10.1145/3613904.3642747]
88. Wu R, Yu C, Pan X, et al. MindShift: leveraging large language models for mental-states-based problematic smartphone use intervention. 2024 Presented at: CHI ’24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems; May 11-16, 2024. [doi: 10.1145/3613904.3642790]
89. Nollen NL, Mayo MS, Carlson SE, Rapoff MA, Goggin KJ, Ellerbeck EF. Mobile technology for obesity prevention. Am J Prev Med 2014 Apr;46(4):404-408. [doi: 10.1016/j.amepre.2013.12.011]
90. Pyky R, Koivumaa-Honkanen H, Leinonen AM, et al. Effect of tailored, gamified, mobile physical activity intervention on life satisfaction and self-rated health in young adolescent men: a population-based, randomized controlled trial (MOPO study). Comput Human Behav 2017 Jul;72:13-22. [doi: 10.1016/j.chb.2017.02.032]
91. Simons D, De Bourdeaudhuij I, Clarys P, De Cocker K, Vandelanotte C, Deforche B. A smartphone app to promote an active lifestyle in lower-educated working young adults: development, usability, acceptability, and feasibility study. JMIR Mhealth Uhealth 2018 Feb 20;6(2):e44. [doi: 10.2196/mhealth.8287] [Medline: 29463491]
92. Spruijt-Metz D, Wen CKF, O’Reilly G, et al. Innovations in the use of interactive technology to support weight management. Curr Obes Rep 2015 Dec;4(4):510-519. [doi: 10.1007/s13679-015-0183-6] [Medline: 26364308]
93. Ding X, Xu J, Wang H, Chen G, Thind H, Zhang Y. WalkMore: promoting walking with just-in-time context-aware prompts. Presented at: 2016 IEEE Wireless Health; Oct 25-27, 2016; Bethesda, MD p. 1-8. [doi: 10.1109/WH.2016.7764558]
94. Caon M, Prinelli F, Angelini L, et al. PEGASO e-Diary: user engagement and dietary behavior change of a mobile food record for adolescents. Front Nutr 2022;9:727480. [doi: 10.3389/fnut.2022.727480] [Medline: 35369096]
95. Martin A, Caon M, Adorni F, et al. A mobile phone intervention to improve obesity-related health behaviors of adolescents across Europe: iterative co-design and feasibility study. JMIR Mhealth Uhealth 2020 Mar 2;8(3):e14118. [doi: 10.2196/14118] [Medline: 32130179]
96. Cushing CC, Bejarano CM, Ortega A, Sayre N, Fedele DA, Smyth JM. Adaptive mHealth intervention for adolescent physical activity promotion. J Pediatr Psychol 2021 Jun 3;46(5):536-546. [doi: 10.1093/jpepsy/jsaa125] [Medline: 33484137]
97. Domin A, Uslu A, Schulz A, Ouzzahra Y, Vögele C. A theory-informed, personalized mHealth intervention for adolescents (mobile app for physical activity): development and pilot study. JMIR Form Res 2022 Jun 10;6(6):e35118. [doi: 10.2196/35118] [Medline: 35687409]
98. Chand A, Gonzalez M, Missig J, Phanichphant P, Sun PF. Balance pass: service design for a healthy college lifestyle. Presented at: CHI06: CHI 2006 Conference on Human Factors in Computing Systems; Apr 22-27, 2006; Montréal, Canada p. 1813-1818. [doi: 10.1145/1125451.1125795]

XSL•FO
RenderX

99. Fortier MA, Chung WW, Martinez A, Gago-Masague S, Sender L. Pain buddy: a novel use of m-health in the management of children's cancer pain. Comput Biol Med 2016 Sep 1;76:202-214. [doi: 10.1016/j.compbiomed.2016.07.012] [Medline: 27479493]
100. Hunter JF, Acevedo AM, Gago-Masague S, et al. A pilot study of the preliminary efficacy of Pain Buddy: a novel intervention for the management of children's cancer‐related pain. Pediatr Blood Cancer 2020 Oct;67(10):e28278. [doi: 10.1002/pbc.28278] [Medline: 32743950]
101. Perry TT, Marshall A, Berlinski A, et al. Smartphone-based vs paper-based asthma action plans for adolescents. Ann Allergy Asthma Immunol 2017 Mar;118(3):298-303. [doi: 10.1016/j.anai.2016.11.028]
102. Fedele DA, Cushing CC, Koskela-Staples N, et al. Adaptive mobile health intervention for adolescents with asthma: iterative user-centered development. JMIR Mhealth Uhealth 2020 May 6;8(5):e18400. [doi: 10.2196/18400] [Medline: 32374273]
103. King‐Dowling S, Psihogios AM, Hill‐Kayser C, et al. Acceptability and feasibility of survivorship care plans and an accompanying mobile health intervention for adolescent and young adult survivors of childhood cancer. Pediatric Blood Cancer 2021 Mar;68(3):e28884. [doi: 10.1002/pbc.28884]
104. Lee AR, Koo D, Kim IK, Lee E, Yoo S, Lee HY. Opportunities and challenges of a dynamic consent-based application: personalized options for personal health data sharing and utilization. BMC Med Ethics 2024 Aug 31;25(1):92. [doi: 10.1186/s12910-024-01091-3] [Medline: 39217356]
105. Lee CM, Cadigan JM, Calhoun BH, et al. A just‐in‐time adaptive personalized mobile app intervention produces limited short‐term reductions in heavy episodic drinking frequency in college students: a randomized control trial during the COVID-19 pandemic. Alcohol Clin Exp Res 2025 Jun;49(6):1321-1336. [doi: 10.1111/acer.70056]
106. Pramana G, Parmanto B, Lomas J, Lindhiem O, Kendall PC, Silk J. Using mobile health gamification to facilitate cognitive behavioral therapy skills practice in child anxiety treatment: open clinical trial. JMIR Serious Games 2018 May 10;6(2):e9. [doi: 10.2196/games.8902] [Medline: 29748165]
107. Cho YW, Chow SM, Li J, et al. Within- and between-individual compliance in mobile health: joint modeling approach to nonrandom missingness in an intensive longitudinal observational study. JMIR Mhealth Uhealth 2025 Oct 30;13:e65350. [doi: 10.2196/65350] [Medline: 41167252]
108. Seewald NJ, Smith SN, Lee AJ, Klasnja P, Murphy SA. Practical considerations for data collection and management in mobile health micro-randomized trials. Stat Biosci 2019 Jul;11(2):355-370. [doi: 10.1007/s12561-018-09228-w] [Medline: 31462937]
109. Zainal NH, Liu X, Leong U, Yan X, Chakraborty B. Bridging innovation and equity: advancing public health through just-in-time adaptive interventions. Annu Rev Public Health 2025 Apr;46(1):43-68. [doi: 10.1146/annurev-publhealth-071723-103909] [Medline: 39656954]
110. Ramanadhan S, Alemán R, Bradley CD, et al. Using participatory implementation science to advance health equity. Annu Rev Public Health 2024 May;45(1):47-67. [doi: 10.1146/annurev-publhealth-060722-024251] [Medline: 38109515]
111. Barker TH, Stone JC, Sears K, et al. The revised JBI critical appraisal tool for the assessment of risk of bias for randomized controlled trials. JBI Evidence Synthesis 2023 Mar;21(3):494-506. [doi: 10.11124/JBIES-22-00430]
112. Klasnja P, Hekler EB, Shiffman S, et al. Microrandomized trials: an experimental design for developing just-in-time adaptive interventions. Health Psychol 2015 Dec;34S:1220-1228. [doi: 10.1037/hea0000305] [Medline: 26651463]
113. Lockwood C, Munn Z, Porritt K. Qualitative research synthesis: methodological guidance for systematic reviewers utilizing meta-aggregation. Int J Evid Based Healthc 2015 Sep;13(3):179-187. [doi: 10.1097/XEB.0000000000000062] [Medline: 26262565]
114. Wies B, Landers C, Ienca M. Digital mental health for young people: a scoping review of ethical promises and challenges. Front Digit Health 2021;3:697072. [doi: 10.3389/fdgth.2021.697072] [Medline: 34713173]
115. Murnane EL, Jiang X, Kong A, et al. Designing ambient narrative-based interfaces to reflect and motivate physical activity. 2020 Apr 21 Presented at: CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; Apr 25-30, 2020; Honolulu, HI p. 1-14. [doi: 10.1145/3313831.3376478]
116. Steele RG, Hall JA, Christofferson JL. Conceptualizing digital stress in adolescents and young adults: toward the development of an empirically based model. Clin Child Fam Psychol Rev 2020 Mar;23(1):15-26. [doi: 10.1007/s10567-019-00300-5] [Medline: 31392451]
117. Holly L, Demaio S, Kickbusch I. Public health interventions to address digital determinants of children's health and wellbeing. Lancet Public Health 2024 Sep;9(9):e700-e704. [doi: 10.1016/S2468-2667(24)00180-4]
118. Biasiotto R, Viberg Johansson J, Alemu MB, et al. Public preferences for digital health data sharing: discrete choice experiment study in 12 European countries. J Med Internet Res 2023 Nov 23;25:e47066. [doi: 10.2196/47066] [Medline: 37995125]
119. Timmons AC, Tutul AA, Avramidis K, et al. Developing personalized algorithms for sensing mental health symptoms in daily life. npj Mental Health Res 2025 Aug 6;4(1):34. [doi: 10.1038/s44184-025-00147-5]
120. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. NPJ Digit Med 2019;2(1):88. [doi: 10.1038/s41746-019-0166-1] [Medline: 31508498]

XSL•FO
RenderX

121. Zhu S, Wang Y, Hu Y. Facilitators and barriers to digital mental health interventions for depression, anxiety, and stress in adolescents and young adults: scoping review. J Med Internet Res 2025 Mar 24;27:e62870. [doi: 10.2196/62870] [Medline: 40127430]
122. Wilson KL, Flores S, Apata BO, Tasnim S, Garney WR, Ajayi KV. Understanding healthcare autonomy among adolescents and young adults in the United States: a scoping review. Front Health Serv 2026 Jan 12;5:1720972. [doi: 10.3389/frhs.2025.1720972]
123. Wanniarachchi VU, Greenhalgh C, Choi A, Warren JR. Personalization variables in digital mental health interventions for depression and anxiety in adolescents and youth: a scoping review. Front Digit Health 2025 May 15;7:1500220. [doi: 10.3389/fdgth.2025.1500220]
124. Baines R, Stevens S, Austin D, et al. Patient and public willingness to share personal health data for third-party or secondary uses: systematic review. J Med Internet Res 2024 Mar 5;26:e50421. [doi: 10.2196/50421] [Medline: 38441944]
125. McGovern Ó, Glennon S, Walsh I, Gallagher P, McCashin D. The use of co-design with young people for digital mental health support development: a systematic review. Internet Interv 2025 Sep;41:100835. [doi: 10.1016/j.invent.2025.100835] [Medline: 40521225]
126. Hekler E, Tiro JA, Hunter CM, Nebeker C. Precision health: the role of the social and behavioral sciences in advancing the vision. Ann Behav Med 2020 Nov 1;54(11):805-826. [doi: 10.1093/abm/kaaa018] [Medline: 32338719]
127. Leong U, Chakraborty B. Participant engagement in microrandomized trials of mHealth interventions: scoping review. JMIR Mhealth Uhealth 2023 May 22;11:e44685. [doi: 10.2196/44685] [Medline: 37213178]
128. Fusar-Poli P, Estradé A, Esposito CM, et al. The lived experience of mental disorders in adolescents: a bottom-up review co-designed, co-conducted and co-written by experts by experience and academics. World Psychiatry 2024 Jun;23(2):191-208. [doi: 10.1002/wps.21189] [Medline: 38727047]
129. Vogel C, Kraft R, Antunes Lima R, Idrees AR, Steiger P, Pryss R. Extending a highly configurable EMA and JITAI mobile app framework with passive sensing, gamification, and AI features for a large-scale physical activity and nutrition study. In: Arabnia HR, Deligiannidis L, Shenavarmasouleh F, Amirian S, Ghareh Mohammadi F, editors. Computational Science and Computational Intelligence: Springer Nature; 2025:26-41. [doi: 10.1007/978-3-031-94950-0_3]
130. Verhulst SG. Operationalizing digital self-determination. Data & Policy 2023;5:e14. [doi: 10.1017/dap.2023.11]
131. Kassam I, Ilkina D, Kemp J, Roble H, Carter-Langford A, Shen N. Patient perspectives and preferences for consent in the digital health context: state-of-the-art literature review. J Med Internet Res 2023 Feb 10;25(1):e42507. [doi: 10.2196/42507] [Medline: 36763409]
132. Bolinger E, Tyl B. Key considerations for designing clinical studies to evaluate digital health solutions. J Med Internet Res 2024 Jun 17;26:e54518. [doi: 10.2196/54518] [Medline: 38885020]
133. Liu X, Deliu N, Chakraborty B. Microrandomized trials: developing just-in-time adaptive interventions for better public health. Am J Public Health 2023 Jan;113(1):60-69. [doi: 10.2105/AJPH.2022.307150] [Medline: 36413704]
134. Kim J, Park S. Empowering individual preferences in mobile notifications: a balanced approach to cognitive load and information needs. IEEE Access 2025;13:44936-44950. [doi: 10.1109/ACCESS.2025.3549033]
135. Wang S, Zhang C, Kröse B, van Hoof H. Optimizing adaptive notifications in mobile health interventions systems: reinforcement learning from a data-driven behavioral simulator. J Med Syst 2021 Oct 18;45(12):102. [doi: 10.1007/s10916-021-01773-0] [Medline: 34664120]
136. Roberts C, Sage A, Geryk L, Sleath B, Carpenter D. Adolescent preferences and design recommendations for an asthma self-management app: mixed-methods study. JMIR Formativ Res 2018 Sep 13;2(2):e10055. [doi: 10.2196/10055]
137. Zimmermann E, Tomczyk S. The ways of using social media for health promotion among adolescents: qualitative interview and focus group study. J Med Internet Res 2025 Jun 9;27:e71510. [doi: 10.2196/71510] [Medline: 40489771]
138. Martinez-Martin N, Insel TR, Dagum P, Greely HT, Cho MK. Data mining for health: staking out the ethical territory of digital phenotyping. npj Digital Med 2018 Dec 19;1(1):68. [doi: 10.1038/s41746-018-0075-8]
139. Poulsen A, Hickie IB, Chong MK, LaMonica HM, Turner A, Iorfino F. Personalised care, youth mental health, and digital technology: a value sensitive design perspective and framework. Ethics Inf Technol 2025;27(4):61. [doi: 10.1007/s10676-025-09866-x] [Medline: 41142874]
140. Brückner S, Dridi A, Deshmukh A, et al. A user-driven consent platform for health data sharing in digital health applications. npj Digit Med 2025;8(1):699. [doi: 10.1038/s41746-025-02147-3]

Abbreviations

AYA: adolescent and young adult
GAD-7: General Anxiety Disorder 7-item
GRADE: Grading of Recommendations, Assessment, Development, and Evaluation
JITAI: just-in-time-adaptive intervention
LLM: large language model
mHealth: mobile health
MRT: microrandomized trial

XSL•FO
RenderX

PHQ-8: Patient Health Questionnaire 8-item
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search extenion
PROSPERO: International Prospective Register of Systematic Reviews
RCT: randomized controlled trial
SWiM: Synthesis Without Meta-Analysis

Edited by S Brini; submitted 19.Sep.2025; peer-reviewed by A Ortega, O Akinpeloye; revised version received 27.Mar.2026; accepted 01.Apr.2026; published 21.May.2026.

Please cite as:
Guan KW, Amara M, Cortiana V, Khan I, Gürbüz S, Iorio C, Ali R, Adlung C, Smit CR, Vreeker A, Thalassinou E, van Roekel E, Keijsers L, de Reuver M, Figueroa CA
Toward Participatory Precision Health With Co-Designed Recommendations: Systematic Review of Just-in-Time Adaptive Interventions in Adolescents and Young Adults
J Med Internet Res 2026;28:e84422
URL: https://www.jmir.org/2026/1/e84422
doi:10.2196/84422

© Kathleen W Guan, Mohammed Amara, Sıla Gürbüz, Imran Khan, Christopher Adlung, Viviana Cortiana, Rayyan Ali, Carmine Iorio, Eva Thalassinou, Crystal R Smit, Annabel Vreeker, Eeske van Roekel, Loes Keijsers, Mark de Reuver, Caroline A Figueroa. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Large Language Models in Colorectal Cancer Care and Clinical Decision Support: Systematic Review

Jinglei Tian[1], BS; Qifeng Lou[2], MSc; Xue Wang[2], MSc; Hangying Xu[1], BS; Huiting Mei[1], BS; Yanli Yu[1], BS

[1]Zhejiang Chinese Medical University, Zhejiang Chinese Medical University, Hangzhou, China, Hangzhou, Zhejiang, China

[2]Department of Gastroenterology, Hangzhou First People's Hospital, Hangzhou, China, Hangzhou, Zhejiang, China

Corresponding Author:
Qifeng Lou, MSc
Department of Gastroenterology, Hangzhou First People's Hospital, Hangzhou, China, Hangzhou, Zhejiang, China

Abstract

Background: Colorectal cancer (CRC) is a leading cause of cancer morbidity and mortality worldwide. The complexity of guideline-concordant care and unstructured clinical data has driven demand for decision-support tools. Large language models (LLMs) show promise for processing clinical data and patient–provider communication, yet evidence is fragmented, and a CRC-specific synthesis across the full care continuum is lacking.

Objective: This systematic review evaluates the current applications, performance determinants, and clinical implications of LLMs across the continuum of CRC care.

Methods: Following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), we searched 6 databases (PubMed, Embase, Web of Science, Scopus, CINAHL, Cochrane) through April 1, 2026. Eligible studies were peer-reviewed original investigations of LLMs on CRC tasks with extractable outcomes; reviews, editorials, and abstracts were excluded. Two reviewers assessed quality with QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2), PROBAST (prediction model risk of bias assessment tool), and ROBINS-I (Risk of Bias in Nonrandomized Studies - of Interventions). Data on model types, applications, prompts, input/output formats, and outcomes were analyzed descriptively, with narrative synthesis per synthesis without meta-analysis (SWiM) guidelines.

Results: Of 8880 records, 37 studies met inclusion criteria (2023‑2026), mostly from China and the United States, with GPT series most frequently evaluated. Overall risk of bias was low in 10/37 studies (27.0%), moderate in 14/37 (37.8%), unclear in 7/37 (18.9%), and high or serious in 6/37 (16.2%). Problematic domains included outcome measurement, intervention classification, patient selection, and lack of blinded assessment. LLMs showed utility in automating data extraction from clinical texts, supporting patient education, aiding diagnosis, and assisting clinical decision-making, with emerging visual interpretation and multimodal capacities. Domain-specific and multimodal models showed advantages over general-purpose models in certain tasks. Performance was significantly influenced by prompt design, from zero-shot queries to fine-tuning. Despite efficiency and outcome benefits, challenges persist regarding methodological quality, data privacy, and generalizability.

Conclusions: This review provides an integrative framework synthesizing evidence across study designs and LLM categories in CRC care. Unlike prior reviews addressing gastroenterology broadly or limited to one design, it covers the full CRC continuum and, for the first time, comparatively evaluates general-purpose, domain-specific, and multimodal LLMs, clarifying how prompt engineering and heterogeneous metrics shape outcomes. Although findings support LLMs' clinical potential, results must be interpreted cautiously, given low overall evidence quality. Most studies lacked safeguards against bias—blinded assessment, confounder adjustment, or prospective multicenter validation. Substantial heterogeneity across tasks, LLM types, prompts, reference standards, and outcomes means reported advantages cannot be generalized. Future work should prioritize real-world integration via prospective multicenter validation, robust privacy frameworks, and rigorous human oversight. Amid rising global CRC burden and health care disparities, this review informs clinical translation, equitable scaling, and policy on LLM deployment.

Trial Registration: PROSPERO CRD420251248261; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251248261

(J Med Internet Res 2026;28:e89862) doi:10.2196/89862

KEYWORDS

artificial intelligence; large language models; colorectal cancer; gastroenterology; systematic review; PRISMA

XSL•FO
RenderX

Introduction

Colorectal cancer (CRC) is the third most commonly diagnosed malignancy and the second leading cause of cancer-related mortality worldwide, with incidence projected to rise substantially through 2050 [1]. Contemporary CRC care spans a long continuum: risk stratification, screening, endoscopic and histopathological diagnosis, multidisciplinary treatment, and long-term surveillance, in which each stage generates dense, largely unstructured clinical text and requires time-sensitive, guideline-concordant decisions [2]. This labor-intensive process is time-consuming and error-prone due to visual fatigue and information gaps inherent in voluminous clinical notes, pulling clinicians from direct patient care and straining both providers and institutional resources [3,4]. Within this context, large language models (LLMs) built on the Transformer architecture have emerged as a candidate interface between complex clinical text and decision support [5]. Compared with conventional clinical decision-support and patient education modalities, LLMs offer several distinct advantages: automated extraction and processing of large-scale clinical follow-up records [6], real-time responses to patient inquiries regarding CRC symptoms and prevention [7], guidance for geographically tailored screening strategies [8], and enhanced adherence to clinical quality improvement initiatives [5], less constrained by outpatient scheduling or geographic disparities in health care resource distribution [9]. This approach conserves clinician time and reduces operational costs while simultaneously improving the accessibility, flexibility, and scalability of CRC-related health information for patients [10].

Against this backdrop, research on LLMs in CRC has expanded rapidly between 2024 and 2026, spanning the entire care continuum. In screening and early detection, GPT-4 and its successors have been evaluated for risk-stratified counseling and family-history triage for hereditary CRC syndromes [11,12], while multiple studies have also explored the clinical utility of LLM-based tools, notably ChatGPT (OpenAI), for preoperative screening consultations and postoperative surveillance monitoring in CRC patients [13,14]. In endoscopy, LLMs have been applied to automate colonoscopy report generation [15,16]. In pathology, LLMs have been used to extract tumor–node–metastasis (TNM) descriptors and microsatellite instability status [17,18]. Therapeutic decision support has emerged as a particularly active area, with LLM recommendations benchmarked against multidisciplinary tumor board consensus [19,20]. The accelerating volume of these publications makes a focused, structured synthesis both timely and necessary.

Nevertheless, digital health models are not without limitations, including technically inaccurate outputs attributable to hallucinations [21], quality assurance concerns in complex diagnostic and therapeutic recommendations [22], and challenges related to model bias, limited generalizability, and the absence of physician empathy [23]. The emerging literature also reflects substantial heterogeneity, with findings that vary across studies. Model selection is one key factor [24]. Published studies have compared various general-purpose and medically fine-tuned models, with consistent reports distinguishing the performance of GPT-4-class and domain-tuned models from that of earlier or smaller backbones in oncology evaluations, while open-source models offer data-privacy advantages but display variable accuracy across CRC tasks [24,25]. Equally consequential is the choice of prompt engineering strategy: zero-shot prompting, few-shot prompting, chain-of-thought reasoning, retrieval-augmented generation (RAG), and guideline-grounded prompting yield markedly different accuracy on identical CRC questions [14,25-28], with several studies reporting accuracy gains when few-shot or RAG approaches replace naive zero-shot baselines [27,29,30]. Additional sources of heterogeneity include differences in evaluation rubrics, question framing, and prompt language [31,32]. Consequently, 2 studies addressing apparently similar questions can reach opposing conclusions. Amini et al [13] assessed the clinical utility of freely available LLMs for colonoscopy surveillance interval recommendations across diverse settings, finding insufficient accuracy and notable limitations. In contrast, Chang et al [14], using the more capable GPT-4 model and a guideline-anchored expert panel as reference, concluded that ChatGPT-4 exhibited accuracy comparable to professional gastroenterologists.

Within the gastroenterological domain, several reviews have mapped LLM applications. Gong conducted a systematic review of LLMs in gastroenterology and gastrointestinal endoscopy, categorizing applications into knowledge-based response evaluation and document automation, with most studies focusing on GPT-series models [9]. Omar et al [15] reviewed 57 natural language processing (NLP) and LLM studies in gastroenterology and hepatology, confirming improved data extraction from electronic health records (EHRs) but noting persistent challenges in integrating these tools into routine clinical practice. Furthermore, a recent systematic review in lung cancer identified critical methodological limitations in primary LLM studies, notably a reliance on retrospective data and unclear risk of bias [33]. Given the fundamental differences in oncology protocols, the specific, multi-stage clinical trajectory of CRC, spanning distinct endoscopic, pathological, and surgical phases, necessitates an isolated, disease-specific appraisal to objectively evaluate LLM viability. However, a conspicuous gap remains: no systematic review has comprehensively evaluated the evidence for LLM applications specifically within the CRC domain. In particular, the information quality of LLM outputs across the full CRC care continuum has been insufficiently addressed in prior systematic reviews. Compounding this limitation, although recent studies have demonstrated that LLMs can achieve clinician-level performance in specific clinical tasks, substantial heterogeneity in model selection, prompt engineering strategies, and evaluation metrics precludes generalizable conclusions [34,35].

Accordingly, this systematic review aims to evaluate the performance of different LLM categories across the full CRC care continuum, address evidence gaps arising from fragmented research practices, and provide a foundation for future research and clinical translation, covering use cases, model types, optimization strategies, limitations, and future directions. Specifically, this review seeks to (1) map LLM applications across the principal clinical domains of CRC management; (2)

XSL•FO
RenderX

compare general-purpose, domain-specific, and multimodal LLMs under different prompt engineering and fine-tuning strategies; (3) classify included studies according to their research design and apply corresponding quality appraisal tools to appraise the credibility of individual studies.

Methods

Eligibility Criteria

The eligibility criteria for this review were established according to the PICOS (Population, Intervention, Comparison, Outcome, Study design) framework, as detailed in Table 1.

Table . PICOS (Population, Intervention, Comparison, Outcome, Study design) eligibility criteria.

Criteria	Definition
Participants	General population or patients with CRC.
Intervention	Artificial Intelligence, specifically LLM[a] applied in CRC[b] management. These may be applications used by patients or health care providers for auxiliary diagnosis, information extraction, knowledge-based question answering, treatment decision-making, predictive modeling, or scientific research. LLMs are advanced AI[c] systems designed to process complex clinical data, support decision-making, and enable effective communication.
Control	Control (applicable exclusively to comparative study designs): Standard clinical evaluation by health care professionals or conventional non-LLM computational algorithms. Studies without a control group were eligible for inclusion if the other criteria were met.
Outcomes	Outcome measures included: Clinical and performance effectiveness (eg, Accuracy, F_1-score, area under the curve, sensitivity, concordance rate) and qualitative/utility measures (eg, response completeness, clarity, comprehensiveness, guideline adherence).
Study types	All study types were considered (eg, exploratory or comparative designs) so long as the original research concept was implemented and tested regarding LLMs and CRC. Nonoriginal research such as books, book chapters, letters, reviews, and conference proceedings were excluded.
Other	Studies were restricted to English language only articles.

[a]LLM: large language model.

[b]CRC: colorectal cancer.

[c]AI: artificial intelligence.

Discrepancies were resolved by discussion, with arbitration by a third reviewer. This review was conducted following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 [36], with search reporting per PRISMA-S [37] and narrative synthesis per SWiM guidelines [38].

Information Sources

Relevant studies were identified by systematically searching 6 electronic databases: PubMed, Web of Science, Embase, Cochrane Library, CINAHL, and Scopus (search cutoff date: April 1, 2026). Each database was searched individually; no multi-database searching on a single platform was performed. No published search filters (eg, validated study design filters) were applied to any database search.

Search Strategy

The search strategy combined Medical Subject Headings (MeSH and EMTREE) and free-text keywords related to CRC and LLMs. These terms were adapted for each database to maximize retrieval sensitivity. Key terms included: “colonic neoplasms,” “colorectal cancer*,” “large language models,” “artificial intelligence,” “LLM,” “GPT,” “ChatGPT,” “Claude,” “Gemini,” and “LLaMA.” The search process followed the PRISMA Search Strategy Extension [20]. The complete search strategy, including specific search queries, applied limits, and the number of records retrieved from each database, is provided in Multimedia Appendix 1. The initial search was established and updated through April 1, 2026, to capture the most recent publications prior to data synthesis.

Regarding the PRISMA-S checklist, certain items were not applicable to our methodology: study registries and regulatory databases were not searched, as research on LLMs in CRC is generally not registered as clinical trials; gray literature, institutional websites, conference proceedings, and preprint servers were not searched; aside from manually screening reference lists, no citation searching tools were used; no additional search methods such as PubMed Related Articles, personal reference libraries, or other database-embedded related-article recommendation features were employed; and stakeholders or content experts were not contacted to identify additional studies, as the designed search was considered sufficiently comprehensive through database coverage alone. Although corresponding authors were contacted via email regarding missing or ambiguous data during the data extraction process, no authors, experts, manufacturers, or other parties were specifically contacted to identify additional studies or unpublished data for inclusion in this review. The search strategy

XSL•FO
RenderX

did not undergo formal external peer review (such as the PRESS checklist process) but was cross-checked and finalized by investigators within the research team. A complete PRISMA-S checklist is provided in Checklist 1.

Selection Process

EndNote X9.3.3 (Clarivate Analytics, US) was used for reference management and automated deduplication, followed by manual verification. Two reviewers (JL and HT) independently screened titles and abstracts, then full texts against eligibility criteria. Discrepancies were resolved by discussion, with arbitration by a third reviewer (QF). Interrater agreement was assessed using the Kappa statistic.

Data Collection Process

Two reviewers (JL and WX) independently extracted data using a predesigned form (WPS Office Excel). Extracted items included: title, first author, year, study design, LLM model, model modality, application scenario, prompt engineering approach, input/output formats, and outcome measures. Interreviewer consistency was calculated; disagreements were resolved by a third reviewer (QF). For missing or ambiguous data, corresponding authors were contacted via email; if unavailable after 2 weeks, items were recorded as "not reported" and excluded from descriptive analyses. No imputation was applied. For studies reporting multiple outcomes, we gave preference to the primary outcome defined by the authors; if none was specified, we selected the metric most central to the study's objective through consensus between 2 reviewers. For other types of outcomes, we extracted the reported values without modification.

Data Items

To manage the inherent overlap between technical tasks, studies were categorized based on their primary terminal clinical objective. For instance, studies employing information extraction specifically to enable automated TNM staging were classified under "Auxiliary Diagnosis" rather than "Information Extraction" to prioritize clinical utility over technical subprocesses.

Study Risk of Bias Assessment

Following Omar and Levkovich [39], the included studies were classified and evaluated based on the assessment design and outcome indicators of the studies rather than their clinical application fields. QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2) [40] was applied for diagnostic accuracy studies validating LLM performance against histopathological diagnosis, endoscopist consensus, or clinical guidelines. PROBAST (prediction model risk of bias assessment tool) [41] was applied for prediction model studies focusing on the development and validation of LLM-based predictive models. ROBINS-I (Risk of Bias in Nonrandomized Studies - of Interventions) [42] was applied for nonrandomized intervention studies evaluating the LLM application effect or clinical value, including information extraction and knowledge-based tasks. Study classifications and corresponding tools are detailed in Multimedia Appendix 2.

Given that LLM studies differ from conventional clinical trials, 2 oncology experts (QF) made minor framework-preserving adaptations to each tool; specific adaptations are documented in Multimedia Appendix 3. Assessment was conducted independently by 2 researchers (JL and HT), with a third (WX) resolving disagreements. Final results were reviewed by an expert (QF). Interrater agreement was evaluated using the Kappa statistic. Overall evidence strength was evaluated considering study quality, consistency of findings, and methodological limitations.

Synthesis Methods

Given the anticipated heterogeneity in clinical tasks, study designs, and outcome constructs, narrative synthesis following SWiM reporting guidelines [38] was planned a priori rather than quantitative meta-analysis. Meta-analysis was not conducted for four reasons: (1) substantial heterogeneity across fundamentally different clinical tasks, rendering pooled estimates uninterpretable; (2) a high proportion of studies rated at moderate, serious, or high risk of bias; (3) fewer than 5 studies within any subgroup sharing comparable task definitions, input modalities, and reference standards; and (4) marked inconsistency in outcome measures precluding standardized effect size extraction.

Reporting Bias Assessment

This systematic review employed a narrative synthesis and did not perform statistical tests for publication bias. Given the absence of a quantitative meta-analysis and the substantial heterogeneity in study design and outcome reporting across included studies, methods such as funnel plots were considered inapplicable. During evidence synthesis and result interpretation, the research team conducted a qualitative assessment of potential reporting bias. By comparing the consistency between study objectives, methods, and reported outcomes, and by incorporating study registration information (where available) and author explanations, the team cautiously discussed the potential impact of missing results on study conclusions.

Results

Study Selection

A total of 8880 records were retrieved (PubMed: 4047; Embase: 1423; Web of Science: 3061; Cochrane Library: 43; Scopus: 43; CINAHL: 263). After automated and manual deduplication using EndNote X9.3.3, 6260 unique records were identified. Following title/abstract screening, 2533 full-text articles were assessed, and 37 studies met the inclusion criteria. The screening-stage Kappa was 0.85. The screening process is presented in Figure 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram of the study selection process for examining the role of LLMs in colorectal cancer. CRC: colorectal cancer; LLM: large language model.

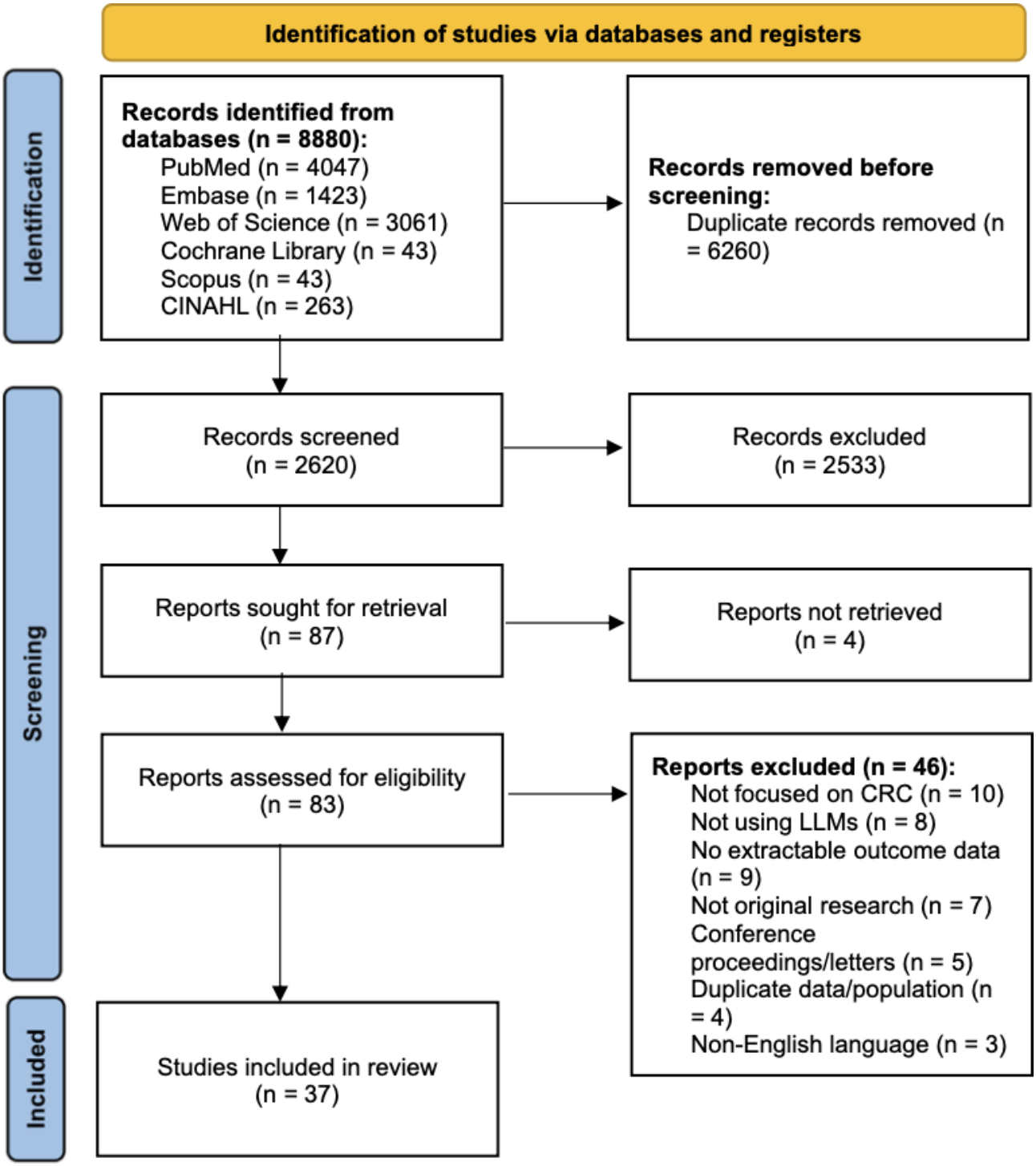

Study Characteristics

The data extraction consistency rate was 0.97. All 37 studies were published between 2023 and 2026, 2 in 2023 [22,43], 11 in 2024 [7,14,25,28,44-50], 22 in 2025 [8,10,13,17,51-67], and 2 in 2026 [20,68]. Studies primarily originated from China [7,20,25,49,50,54-57,59,64,65] and the United States [13,14,17,45,47,53,58,60,66,68], with others from Italy [62,63], Germany [46,61], Singapore [28,51], Israel [43], Switzerland [67], Spain [52], Turkey [48], the United Kingdom [44], South Korea [69], and multinational collaborations [8,10,22]. Application domains included auxiliary diagnosis [14,17,44,46,49,50,62,65], information extraction [10,17,44,52,57,69], knowledge-based question answering [7,22,25,28,45,48,50,63,64], treatment decision-making [8,20,47,53,56,61,67,68], predictive modeling [51,54], scientific research [58,66], and aided nursing [60].

The LLMs used varied widely, with the most frequent being OpenAI's GPT series. Other models included Google's Gemini, Anthropic's Claude, Meta's LLaMA series, as well as DeepSeek, GLM, and Qwen, among others. The best-performing models identified in comparative studies are summarized in Table 2.

XSL•FO
RenderX

The results suggest that models such as GPT-4, GPT-4o, and Claude 2.1 showed relatively favorable performance in some tasks [14,25,28,68,69]; o3-mini reportedly showed comparatively higher intra-model stability and expert concordance among reasoning-oriented models for multidisciplinary team decision simulation [20]. However, for specific tasks, lightweight models or domain-specialized models may also perform optimally [17,51,52]. A summary of these details is provided in Table 2.

XSL•FO
RenderX

Table . Summary of included sources. Pure LLM: text-only language model processing textual inputs exclusively. Multimodal VLM: vision-language model capable of processing both textual and visual inputs (eg, GPT-4V, GPT-4o with image input).

Study	Country	LLMs[a] used	Model type	Application domain	Best performer
Zeng, 2025 [8]	Multi-national	ChatGPT-4.5	Pure LLM	Treatment Decision	—[b]
Zeng, 2025 [56]	China	ChatGPT-4o, DeepSeek	Pure LLM	Treatment Decision	—
Schmutz, 2025 [61]	Germany	ChatGPT 4.0	Pure LLM	Treatment Decision	ChatGPT 4.0
Chatziisaak, 2025 [67]	Switzerland	ChatGPT-4	Pure LLM	Treatment Decision	—
Horesh, 2025 [53]	United States	ChatGP -3.5	Pure LLM	Treatment Decision	—
Kaiser, 2024 [47]	United States	ChatGPT-3.5, Microsoft Copilot	Pure LLM	Treatment Decision	—
Garg, 2026 [68]	United States	ChatGPT-4o	Pure LLM	Treatment Decision	GPT-4o
Qu, 2026 [20]	China	ChatGPT-o3-mini, DeepSeek-R1, Qwen qwq-plus	Pure LLM	Treatment Decision	o3-mini
Diaz, 2025 [66]	United States	AI-HOPE (LLaMA 3-based)	Pure LLM	Scientific Research	—
Yang, 2025 [58]	United States	LLaMA 3	Pure LLM	Scientific Research	—
Yang, 2025 [54]	China	BGE-M3, XGBoost	Pure LLM	Predictive Modeling	XGBoost
Kim, 2025 [51]	Singapore	BioBERT-Large, RadImageNet, 3D ResNet	Multimodal VLM	Predictive Modeling	BioBERT-Large
Lim, 2024 [28]	Singapore	GPT-4	Pure LLM	Knowledge QA	GPT-4
Hu, 2025 [64]	China	ChatGPT-4.5	Pure LLM	Knowledge QA	ChatGPT-4.5
Peng, 2024 [7]	China	ChatGPT-3.5	Pure LLM	Knowledge QA	—
Wang, 2024 [50]	China	GPT-3.5-turbo	Pure LLM	Knowledge QA	—
Zhang, 2025 [55]	China	ChatGPT-4o, Claude 3.5, DeepSeek	Pure LLM	Knowledge QA	ChatGPT-4o
Zhou, 2024 [25]	China	ChatGPT, Doctor GPT, Llama-2-70B, Mixtral-8 × 7B, Bard, Claude 2.1	Pure LLM	Knowledge QA	Claude 2.1
Gorelik, 2023 [43]	Israel	ChatGPT-4	Pure LLM	Knowledge QA	—
Maida, 2025 [63]	Italy	ChatGPT-4o	Pure LLM	Knowledge QA	—
Maida, 2025 [10]	Multi-national	ChatGPT-4	Pure LLM	Knowledge QA	—
Emile, 2023 [22]	Multi-national	ChatGPT-3.5	Pure LLM	Knowledge QA	—
Kepez, 2024 [48]	Turkey	ChatGPT-4	Pure LLM	Knowledge QA	ChatGPT-4
Atarere, 2024 [45]	United States	ChatGPT, BingChat, YouChat	Pure LLM	Knowledge QA	ChatGPT, YouChat
Yu, 2025 [57]	China	Gemini, GPT-4, GPT-4o, Claude, Llama, DeepSeek, GLM, Qwen	Pure LLM	Information Extraction	GPT-4
Chizhikova, 2025 [52]	Spain	RoBERTa	Pure LLM	Information Extraction	Task-specific models
Alzaid, 2024 [44]	UK	ChatGPT-4 Turbo, GPT-4V	Multimodal VLM	Information Extraction	—
Johnson, 2025 [17]	United States	Gemma-2-9B-It-SPPO, Llama-3-8B-Instruct	Pure LLM	Information Extraction	Gemma-2
Kim, 2025 [69]	South Korea	GPT-4	Pure LLM	Information Extraction	GPT-4
Ding, 2025 [65]	China	ChatGPT-4	Multimodal VLM	Auxiliary Diagnosis	—

XSL•FO
RenderX

Study	Country	LLMs[a] used	Model type	Application domain	Best performer
Liu, 2024 [49]	China	ChatGPT-3.5, ChatGPT-4.0	Pure LLM	Auxiliary Diagnosis	GPT-4.0
Wang, 2025 [59]	China	ChatGPT, Claude, ERNie, SAM	Multimodal VLM	Auxiliary Diagnosis	—
Ferber, 2024 [46]	Germany	ChatGPT-4V	Multimodal VLM	Auxiliary Diagnosis	GPT-4V
Massimi, 2025 [62]	Italy	ChatGPT-4o	Multimodal VLM	Auxiliary Diagnosis	GPT-4o
Amini, 2025 [13]	United States	GPT-3.5-turbo, Bard (PaLM 2)	Pure LLM	Auxiliary Diagnosis	ChatGPT-3.5
Chang, 2024 [14]	United States	ChatGPT-4	Pure LLM	Auxiliary Diagnosis	ChatGPT-4
Sehgal, 2025 [60]	United States	ChatGPT-4.1	Pure LLM	Aided Nursing	—

[a]LLM: large language model.

[b]No intermodel comparison was performed or the metric is not applicable.

Prompt Engineering and Model Training

The data extraction consistency rate was 0.97. We synthesized prompt engineering strategies, model inputs/outputs, and evaluation metrics (Table 3). Five studies [10,22,45,47,63] did not explicitly describe prompting strategies, employing basic queries primarily for educational purposes. Thirty-two studies described distinct methods, including instruction templates and instructional prompts [8,13,14,17,20,28,43,44,48,55,58-62,64-69], zero-shot learning [7,25,49,50,53,57,64], few-shot learning [46,56], fine-tuning [51,52,54], and hybrid approaches [57,60,68]. Training data were text-based in 33 studies [7,8,10,13,14,17,20,22,25,28,43-45,47-50,52-61,63,64,66-69], image-based in 2 studies [46,62], and multimodal in 2 studies [51,65]. Common outcome metrics included accuracy [7,8,17,20,25,28,44,46,48-50,52,53,55-57,59,60,62,64,65,68], F_1-score [17,51,52,57,65], area under the curve [51,54,59], sensitivity [53,57,65], and concordance rate [13,14,17,20,28,43,44,48,56,59-61,68]. A categorized summary is provided in Multimedia Appendix 4.

XSL•FO
RenderX

Table . Prompt engineering and model training.

Study	Prompt method or content	Model input	Model output	Outcome indicators
Zeng, 2025 [8]	Instruction template	Standardized patient cases	Screening and monitoring recommendations	Correct/partially correct/incorrect proportions; descriptive statistics
Amini, 2025 [13]	Instruction template	Colonoscopy reports, pathology, history, family history	Colonoscopy interval recommendation	Agreement percentage, Fleiss' kappa, McNemar test
Chang, 2024 [14]	Instruction template	Deidentified clinical data, colonoscopy reports, pathology reports	Follow-up colonoscopy interval suggestions	Agreement rate, Fleiss kappa
Johnson, 2025 [17]	Instruction template	Pathology report text	Yes/no answer	F_1-score, PPV[a], NPV[b], sensitivity, specificity, MCC[c]
Lim, 2024 [28]	Instruction template	Patient scenario descriptions	Colonoscopy interval recommendations	Correct interval percentage, hallucination rate
Gorelik, 2023 [43]	Instruction template	Structured endoscopy reports & free-text clinical notes	Guideline-based next-step recommendations; Patient result explanation letters	Guideline adherence, accuracy, Fleiss' kappa
Alzaid, 2024 [44]	Instruction template	Unstructured pathology reports	Structured JSON report with confidence	Accuracy, Kappa, AUROC[d]
Kepez, 2024 [48]	Instruction template	20 common questions on colon cancer	Answer text for each question	DISCERN, GQS, JAMA criteria, Flesch-Kincaid readability, SAM, HITS, VPI, HONcode
Zhang, 2025 [55]	Instruction template	Chinese Society of Clinical Oncology guideline standards / instructions	Colorectal cancer screening educational text	Accuracy, clarity, rigor scores
Yang, 2025 [58]	Instruction template	Natural language queries on clinical genomic data	Mutation profiles, survival curves, odds ratios	*P* values, hazard ratios, odds ratios
Wang, 2025 [59]	Instruction template	Free-text colonoscopy reports	Report-level labels	Accuracy, average precision, dice similarity coefficient, AUC[e]
Sehgal, 2025 [60]	Instruction template	Self-reported demographics	AI-generated personalized messages or chatbot dialogues	Intent score change, Cohen *d*, *P* values, OR[f], Flesch-Kincaid readability
Schmutz, 2025 [61]	Instruction template	Clinical patient summaries and pathology reports	Treatment/diagnostic recommendations	Recommendation type, information density, consistency, level of evidence, time efficiency
Massimi, 2025 [62]	Instruction template	Colonoscopy video frames	Paris classification	Accuracy, sensitivity, specificity, Fleiss' kappa
Ding, 2025 [65]	Instruction template	Pathology images and text prompts	Tissue origin, lesion classification, diagnosis	Accuracy, sensitivity, specificity, PPV, NPV, F_1-score, Kappa, ICC[g]
Diaz, 2025 [66]	Instruction template	Natural language queries for scanning and validating clinical genomic datasets	Survival analysis results, mutation frequency comparisons, statistical significance	*P* values, odds ratios, survival rates
Chatziisaak, 2025 [67]	Instruction template	Patient clinical data	Treatment recommendation	Consistency, chi-square test
Qu, 2026 [20]	Instruction template; Multi-role prompting	Structured variables and free-text summaries from clinical records	Four-category treatment classification code	Intra-model agreement; expert-model concordance, Cohen κ
Garg, 2026 [68]	Instruction template; Role prompting; Few-shot; Chain-of-thought; JSON schema enforcement	Colonoscopy reports, pathology reports, patient family history and preoperative diagnoses	Structured clinical entities and 2020 USMSTF-based surveillance interval recommendations; 2024 ACG/ASGE quality indicators	Case-level accuracy, Cohen κ; Fleiss' κ; ADR, SSLDR, cecal intubation rate, bowel prep adequacy

XSL•FO
RenderX

Study	Prompt method or content	Model input	Model output	Outcome indicators
Kim, 2025 [69]	Instruction template; Role prompting	Unstructured preoperative abdominal CT / rectal MRI reports	Lesion location and cTNM stage and reasoning	Lesion location accuracy
Kim, 2025 [51]	Fine-tuning	CT images and radiology report texts	Binary NAR score classification	AUC
Chizhikova, 2025 [52]	Fine-tuning	Spanish colon MRI report texts, numerical features, categorical features	TNM[h] staging	Accuracy, macro F_1-score, precision, recall
Yang, 2025 [54]	Fine-tuning	Clinical EHR[i] data	Binary colorectal adenoma risk	AUC, sensitivity, specificity, F_1-score, PPV, NPV, mean lead time
Ferber, 2024 [46]	Few-shot	Cancer pathology images	Image classification labels	Accuracy, confidence interval, recall
Zeng, 2025 [56]	Few-shot; Role prompting; Context learning	Real-world pathology report text	Recommendation on need for additional surgery	Accuracy; guideline consistency proportion
Peng, 2024 [7]	Zero-shot	Medical questions from books	Colorectal cancer-related answers	Accuracy, comprehensiveness scores
Zhou, 2024 [25]	Zero-shot	150 CRC-related[j] closed-ended questions	Yes/no answers	Accuracy
Liu, 2024 [49]	Zero-shot	Colorectal cancer case report texts	Primary/secondary diagnoses	Accuracy
Wang, 2024 [50]	Zero-shot	Pathology report text and related questions	Answers to pathology questions	7-point Likert scale
Horesh, 2025 [53]	Zero-shot	Clinical patient summaries	Next best management recommendation	Consistency with multidisciplinary team decisions, reasonableness score, interrater reliability
Yu, 2025 [57]	Zero-shot; Chain-of-thought	Endoscopy/colonoscopy report texts	Structured JSON including lesion location, features, layer structure, distribution, diagnosis	Precision, recall, F_1-score, accuracy
Hu, 2025 [64]	Zero-shot	Patient question texts	Answer texts	Accuracy, completeness, clarity scores
Maida, 2025 [10]	— [k]	15 questions on colorectal cancer screening	Text answers to questions	Accuracy, completeness, clarity scores
Emile, 2023 [22]	—	38 common questions on CRC prevention, diagnosis, management	Text answers	Expert consensus; consistency with guidelines
Atarere, 2024 [45]	—	15 questions on CRC screening concepts and 5 experience-based questions	Response appropriateness	Appropriateness rating
Kaiser, 2024 [47]	—	Clinical scenario questions on next management	Text recommendations for clinical questions	Accuracy score, consistency, verbosity

XSL•FO
RenderX

Study	Prompt method or content	Model input	Model output	Outcome indicators
Maida, 2025 [63]	—	Patient queries	ChatGPT-generated answers	Expert scores, patient scores

[a]PPV: positive predictive value.
[b]NPV: negative predictive value.
[c]MCC: Matthews correlation coefficient.
[d]AUROC: area under the receiver operating characteristic curve.
[e]AUC: area under the curve.
[f]OR: odds ratio.
[g]ICC: intraclass correlation coefficient.
[h]TNM: tumor–node–metastasis.
[i]EHR: electronic health record.
[j]CRC: colorectal cancer.
[k]Prompt method was not explicitly reported.

Risk of Bias in Studies

The included studies were categorized by research objective, and quality was assessed using the corresponding appraisal tool. The kappa value between the 2 reviewers was 0.95. Two predictive modeling studies [51,54] were evaluated using PROBAST (Figure 2A); both showed low risk of bias across the participants, predictors, and outcome domains, but one exhibited high risk of bias in the analysis domain. Eighteen diagnostic studies [13,14,17,25,28,43-46,49,52,56,59,62,65,67,68] were assessed using QUADAS-2 (Figure 2B); while most demonstrated acceptable applicability, risk of bias in the patient selection domain was frequently unclear or high. Seventeen intervention studies [7,8,10,20,22,47,48,53,55,58,60,61,63,64,66,69] were appraised using ROBINS-I (Figure 2C); risk of bias was predominantly low for participant selection, deviations from intended interventions, and missing data, but moderate to serious for outcome measurement and classification of interventions.

XSL•FO
RenderX

Figure 2. (A) The quality appraisal for 2 predictive studies with PROBAST (prediction model risk of bias assessment tool). (B) The quality appraisal for 18 diagnostic studies with QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2). (C) The quality appraisal for 17 intervention trials with ROBINS-I (Risk of Bias in Nonrandomized Studies - of Interventions).

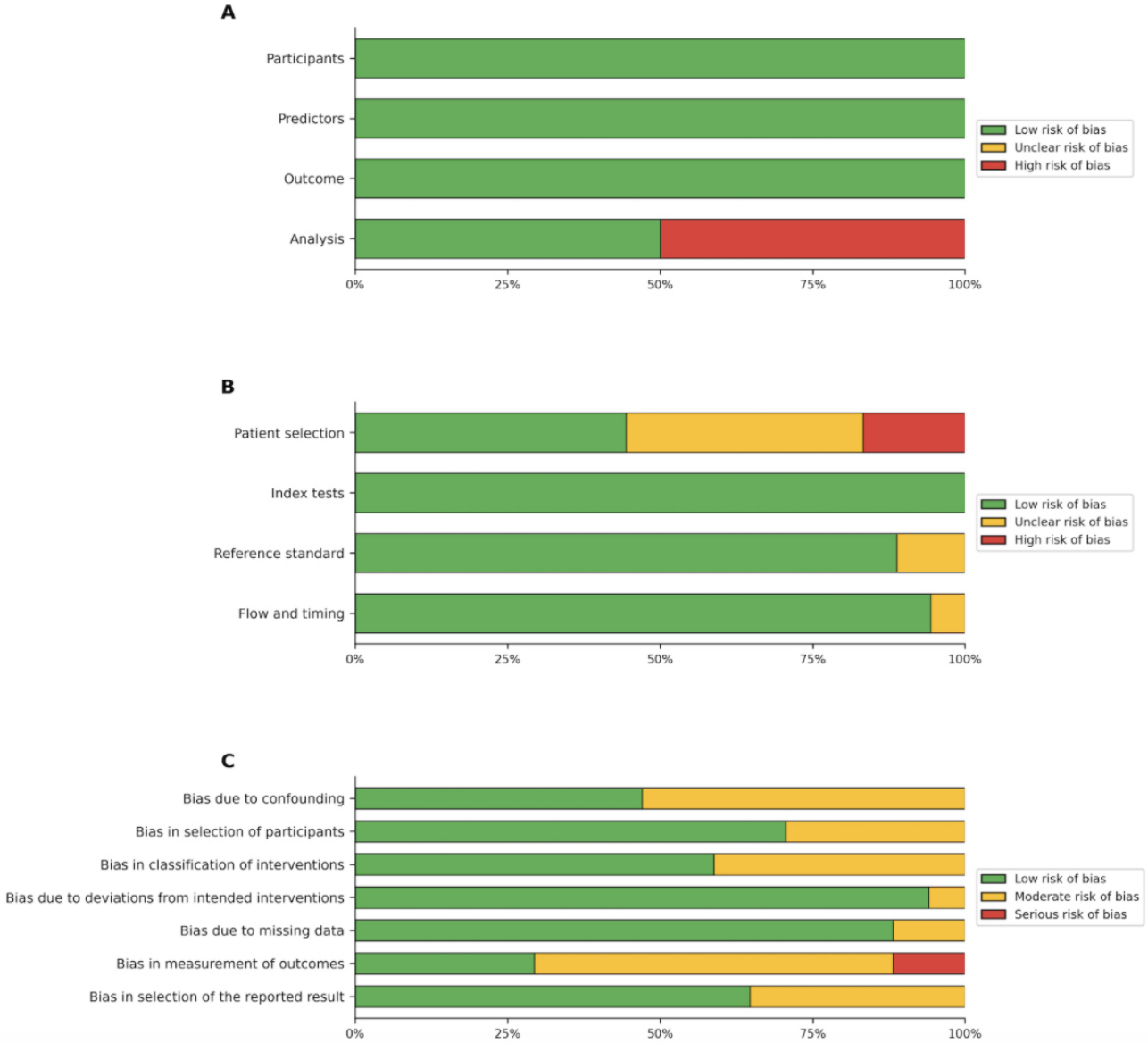

Overall, 27 of the 37 included studies were rated above low risk of bias: 6 as high or serious [28,44-46,57,59], 7 as unclear [43,47,50,52,53,62,65], and 14 as moderate [7,8,10,13,20,22,48,55,58,60,61,63,64,66], while 10 were rated as low [14,17,25,49,51,54,56,67-69]. The most problematic domains across tools were outcome measurement [8,10,13,20,22,43,46-48,50,52,53,55,58,60-62,64,66,68] and patient selection [25,28,44-46,52,56,59,62,65]. Given these recurring concerns, particularly regarding blinding [43,46,53,60,61], outcome measurement [8,10,13,20,22,43,46-48,50,52,53,55,58,60-62,64,66,68], and confounding [17,20,44-46,53,58,60-62,66,68], and the considerable heterogeneity in clinical tasks, LLM models, and outcome metrics, the overall certainty of evidence was judged as moderate to low. Quantitative meta-analysis was not feasible; even within the largest subgroup, fewer than 5 studies were sufficiently aligned in task definition, input modality, and reference standard to permit reliable pooling. A narrative synthesis was therefore adopted, and the findings should be interpreted with caution.

Discussion

Principal Findings

Through a comprehensive analysis of 37 studies, we identified 5 primary application domains of LLMs in CRC diagnosis and treatment: auxiliary diagnosis, information extraction, knowledge-based question-answering and patient education, treatment decision support, and scientific research and predictive modeling (Table 2). These domains are often interconnected in clinical practice. For instance, information extraction frequently provides structured data to support diagnostic processes [17,57], while knowledge-based question-answering is widely applied in scientific communication and patient education [7,45,48,63,64].

XSL•FO
RenderX

Applications of LLMs in CRC

LLMs enable the automated extraction of clinical features through NLP [70]. Multiple studies have utilized LLMs to extract key information from EHRs [17], endoscopy reports [27], radiology reports [25], and pathology reports [17,52]. This capability assists not only in clinical staging and histological classification [71] but also in predicting disease progression and treatment response [17]. For instance, lymph node metastasis assessment based on MRI reports [51] and tumor progression prediction from radiology reports [72] have shown promising accuracy. These advancements underscore the significant value of LLMs in early CRC screening. Early diagnosis can effectively improve survival rates [25], and mass screening achieves a high detection rate for early-stage lesions [73]. Wang leveraged LLMs to automatically extract knowledge from colonoscopy image-text records, enabling polyp detection and segmentation without manual annotation, thereby offering a novel approach to screening automation [59]. A systematic review of LLMs in gastroenterology similarly demonstrated the potential applications of LLMs in gastrointestinal endoscopy and precancerous lesion screening [74]. Despite challenges such as insufficient extraction performance for complex tasks and hallucinations reporting a lower accuracy of 55% for LLMs in classifying pedunculated polyps, indicating they cannot yet fully replace endoscopic experts [25,62], we remain optimistic about their future performance in assisting CRC diagnosis and early screening. This optimism is fueled by ongoing advancements in multimodal integration [75], the development of domain-specific models [46,76], and the continuous optimization of training data [34].

Leveraging their strong interactive capabilities and extensive knowledge, LLMs are widely evaluated for CRC medical question-answering and patient education [7,22]. Furthermore, advancing multimodal models now enable LLMs to jointly analyze medical images and text, offering CRC diagnostic and therapeutic suggestions in controlled settings [6]. Gong has recently emphasized that multimodal fusion has emerged as the dominant next-generation development trend for gastrointestinal artificial intelligence [9]; however, this important technological milestone has not yet received adequate attention in available systematic reviews. Ferber demonstrated that multimodal LLMs applying in-context learning achieved near-pathologist-level classification of cancer pathology images [46], and Kim [51] showed that combined LLM and vision deep learning architectures outperformed either modality alone for neoadjuvant rectal score prediction, which preliminarily suggests the potential of multimodal LLMs, as they can reach a level close to that of pathologists when processing pathological image classification and clinical prediction tasks, and outperform single-modality models. Despite this progress, the diagnostic accuracy of current multimodal models on morphologically complex tasks remains constrained [34,62]. This reinforces the prevailing clinical consensus that current LLMs must be deployed strictly as decision-support adjuncts rather than autonomous diagnostic agents, thereby mitigating the significant clinical risks associated with automation bias and diagnostic delay [23,49]. Furthermore, extraction performance varied markedly, dictated by underlying model architecture and optimization strategy. GPT-4, augmented with multi-strategy prompting, appeared to outperform zero-shot baselines for colonoscopy report extraction [57], while biomedical pretrained RoBERTa showed better performance than general-purpose GPT models for TNM staging in the available evidence from Spanish-language reports [52]. This discrepancy unequivocally indicates that domain-adaptive and language-specific pretraining confers fundamental structural semantic advantages that advanced prompt engineering alone cannot replicate [44], consistent with recent evaluations where specialized models exhibited superior performance within data-constrained clinical settings [77]. Nonetheless, the majority of this evidence is derived from retrospective analyses and single-center validations, with a notable paucity of prospective, multicenter clinical trials to confirm generalizability and real-world efficacy [47,55,66].

The NLP and named entity recognition capabilities of LLM extend their utility beyond direct clinical support for practitioners and patients, substantially improving the efficacy of medical research workflows [78]. In the domain of data extraction and analysis, Johnson leveraged the Gemma-2 model to accurately identify and extract key pathological diagnostic entities—such as dysplasia, high-grade dysplasia/adenocarcinoma, and invasive carcinoma—from unstructured pathology reports [17]. This high accuracy aligns robustly with Chen et al [6], who demonstrated comparable reliability in extracting oncological variables from EHRs, confirming automated information extraction as one of the most mature LLM applications. Beyond data retrieval, LLMs are increasingly serving as active engines for hypothesis generation [79,80]. Their probabilistic structure allows them to synthesize vast, disparate datasets and infer latent correlations that traditional algorithms might overlook [72]. In translational medicine, Yang developed AI-HOPE-TP53, a LLaMA 3-based conversational agent that facilitates pathway-centric analysis of clinical genomic data in early-onset CRC [58]. By rapidly generating statistical outputs like survival curves and hazard ratios, this system accelerates hypothesis-driven research in precision oncology [81]. The viability of this paradigm shift is further corroborated by Abdel-Rehim, who experimentally validated that LLM-driven pipelines can successfully identify novel, laboratory-verifiable synergistic drug combinations [80]. Furthermore, hybrid LLM architectures are democratizing access to complex analytical tools in routine practice. Yang et al [54] developed an early-stage CRC adenoma risk prediction model combining BGE-M3 semantic vector encoding with XGBoost algorithms. By enabling clinicians without specialized computational expertise to perform sophisticated risk stratification based on LLM-processed outputs, such models substantially reduce the administrative burden and facilitate a more patient-centered clinical workflow [6]. The research-supportive functions of LLMs have also expanded into foundational scholarly activities, including knowledge synthesis and the drafting of study protocols, ethics materials, and preliminary manuscript sections [43,59]. However, the originality and factual accuracy of such artificial intelligence-generated scholarly content necessitate rigorous human oversight to ensure scientific integrity [82].

XSL•FO
RenderX

Limitations of LLMs and Future Directions

Current research on LLMs in the field of CRC predominantly focuses on textual data processing [83]; investigations into other modalities, including CT images [51,67], histopathological slides [58], and bioinformatics data [58], remain in their nascent stages, demonstrating suboptimal output precision and task stability. General-purpose LLMs (eg, ChatGPT and the LLaMA series), predominantly pretrained on public databases, frequently manifest deficiencies such as delayed knowledge base updates, insufficient coverage of CRC subspecialty knowledge, a propensity for hallucinations, and an absence of authoritative evidence-based support for pivotal clinical content [8,76,84]. Conversely, although existing medical-domain-specific LLMs (eg, Med-PaLM 2, BioBERT, and ClinicalBERT) possess certain advantages in general medical tasks, their comprehensive performance in complex subspecialty tasks, such as the precision treatment of CRC, still lags behind that of large-parameter general-purpose models [57]. More critically, the reliability and generalizability of currently well-developed diagnostic and decision-support tools are severely hindered by methodological flaws; existing evidence relies disproportionately on retrospective, single-center datasets lacking temporal or geographic stratification [77]. This evaluative paradigm renders models highly susceptible to overfitting and training data leakage, thereby precipitating a drastic degradation in performance within real-world clinical environments [85]. There remains a critical paucity of rigorous prospective, multicenter clinical validation data within this domain [86,87].

The risk-of-bias assessment revealed several recurring methodological weaknesses across study designs. Among diagnostic accuracy studies evaluated with QUADAS-2, the patient selection domain was the most common source of concern, with ratings of “unclear” or “high” largely attributable to unreported sampling procedures and potentially inappropriate exclusion criteria [32,34,39,42,57]. For nonrandomized intervention studies assessed with ROBINS-I, the principal limitations were inadequate adjustment for confounding variables [40,46-48,52,58,61,62] and the absence of blinded outcome assessment, both of which may bias effect estimates [15,19,37,41,44,47,51,52,54,55,58,62]. Prediction model studies appraised with PROBAST generally performed well in the participants, predictors, and outcome domains but showed weaknesses in the analysis domain, including limited sample size, unexplained participant attrition, and insufficiently described handling of missing data [43,45]. Collectively, these methodological limitations reduce the reliability of the current evidence base and constrain its translational applicability.

Beyond data-related constraints, the intrinsic technical vulnerabilities and compliance risks of LLMs pose substantial threats to clinical safety [23]. The profound sensitivity of models to version iterations and prompt variations results in exceedingly poor reproducibility of outputs across multi-institutional settings [77]. In the absence of specific instructional constraints, models are not only prone to hallucinations but may also exacerbate negative societal biases and stereotypes [88]. Uncritical acceptance of these recommendations by clinicians may engender bias, subsequently precipitating critical diagnostic delays or inappropriate clinical interventions [89]. Furthermore, constrained by the heterogeneity of patient requirements and the stringent governance of sensitive data, applications pertaining to patient follow-up and supportive care remain the most underdeveloped [60]. Moreover, the pervasive absence of data privacy and information security protocols during the cloud-based deployment of open-source LLMs further impedes their clinical translation and real-world implementation [6,9].

To address current technical bottlenecks, it is imperative to enhance model precision and reliability through future technological advancements [90]. Multimodal integration is recognized as the predominant trajectory for next-generation technological development in this domain, offering the potential to transcend the limitations of unimodal text processing [91]. Regarding optimization strategies, RAG technology emerges as an optimal solution for tailoring general-purpose models to subspecialty clinical scenarios [9]. By interfacing with independent, verifiable, and authoritative subspecialty knowledge bases, RAG facilitates real-time knowledge updates, effectively enhances the concordance between model outputs and authoritative guidelines, substantially mitigates hallucinations, and endows models with robust interpretability [92,93]. Concurrently, prompt engineering (eg, instruction templates, few-shot learning, and chain-of-thought prompting) can rapidly augment the performance of general-purpose models in specific tasks, including pathological data extraction, treatment regimen recommendation, and follow-up protocol formulation, without altering underlying model weights [94,95].

Regarding clinical integration and ethical governance, future research priorities must pivot toward achieving real-world validity and safety [96]. Primarily, prospective, multicenter clinical validations must be conducted for diagnostic and treatment planning applications, while patient follow-up and supportive care systems must be specifically developed to rectify deficiencies in full-cycle management [97]. More crucially, LLMs cannot supplant medical professionals; their responsible clinical application must be strictly predicated on the establishment of a robust ethical governance framework [98]. This necessitates the strict enforcement of their adjunctive role under continuous human supervision, concurrent with the resolution of data privacy issues and the assurance of foundational data quality [99]. Ultimately, cross-disciplinary collaboration is imperative to delineate accountability, ensuring the synchronous evolution of governance frameworks and cutting-edge technologies [82].

Limitations of This Systematic Review

This review has several limitations. First, the majority of included studies were retrospective and single-center in design, and no prospective multicenter clinical trials establishing real-world LLM effectiveness in CRC care were identified. Only a minority conducted independent external validation, precluding confirmation of generalizability across diverse populations and institutions. Second, the rapid publication pace of LLM research means some recent developments may not have been captured despite the April 1, 2026 search cutoff. Third, restriction to English-language publications may introduce geographic bias. Fourth, several included studies evaluated proprietary commercial models such as GPT-4 and

XSL•FO
RenderX

Claude, whose architectures and training data are not fully disclosed, introducing additional transparency and reproducibility concerns. No included study reported direct industry sponsorship for LLM evaluation. Finally, the search strategy was only cross-checked internally without formal external peer review, potentially leading to omission of a few unpublished or noncore journal studies. Inherent subjectivity in quality appraisal was mitigated through independent dual assessment, third-reviewer arbitration, and expert validation [100,101].

Conclusions

This review establishes an integrative framework that synthesizes evidence across diverse study designs and LLM categories to compare their respective strengths and limitations in CRC care. Distinct from prior reviews that have addressed gastroenterology broadly or have been confined to a single study design, our work focuses specifically on the full-cycle CRC care continuum and, for the first time, comparatively evaluates general-purpose, domain-specific, and multimodal LLMs, thereby elucidating how prompt engineering and heterogeneous evaluation metrics shape reported outcomes. While our findings substantiate the clinical potential of LLMs, these results should be interpreted with caution, given the overall low quality of the available evidence. Most included studies failed to report key safeguards against bias—such as blinding of outcome assessors, adequate adjustment for confounders, or the use of prospective, multicenter designs to validate model generalizability. Moreover, the substantial heterogeneity we observed across task types, LLM categories, prompt engineering strategies, reference standards, and outcome measures indicates that the performance advantages reported for any specific LLM are confined to the corresponding tasks and clinical scenarios and cannot be generalized. Future efforts should therefore prioritize the integration of LLMs into real-world clinical practice, which will require prospective, multicenter validation, a robust privacy-protection framework, and rigorous human oversight to mitigate bias. Against the backdrop of a rising global CRC burden and persistent disparities in health care resource allocation, this review provides an evidence base to inform the clinical translation, equitable scaling, and policy formulation surrounding LLM deployment in CRC care.

Registration and Protocol

This systematic review was prospectively registered in the International Prospective Register of Systematic Reviews (PROSPERO) under registration number CRD420251248261. The review protocol is publicly accessible through the PROSPERO database. No separate protocol manuscript was published.

One amendment was made to the registered protocol: the literature search cutoff date was extended from November 1, 2025 to April 1, 2026, to capture the most recent publications prior to data synthesis. This amendment was implemented after the initial search had been completed and did not alter the review's eligibility criteria, synthesis methodology, or any other prespecified procedures. The narrative synthesis approach (SWiM), quality assessment tools (QUADAS-2, PROBAST, ROBINS-I), eligibility criteria, database selection, screening processes, and data extraction methods were all carried out as prespecified in the registered protocol. No other amendments were made.

Acknowledgments

The authors sincerely thank Zhejiang Chinese Medical University and Hangzhou First People's Hospital for providing the academic research platform, professional literature resource support, and methodological guidance for the completion of this systematic review. This manuscript was originally drafted in Chinese and subsequently translated into English. During the preparation and translation process, the authors further used ChatGPT (OpenAI) to assist with English-language polishing. All AI-generated outputs were critically reviewed and manually edited by the authors, who take full responsibility for the accuracy and integrity of the final content. The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. In accordance with the GAIDeT taxonomy (2025), GAI tools were used under full human supervision for idea generation, proofreading and editing, and translation. The GAI tools used were Gemini 3 and DeepSeek. Responsibility for the content and integrity of the final manuscript rests entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. This declaration is submitted under the collective responsibility of the authors.

Funding

This work was supported by the Clinical Research Application Project of Zhejiang Provincial Medical and Health Science and Technology Program (grant number 2024KY190), the Hangzhou Municipal Medical and Health Science and Technology Program (grant number A20241859), and the Hangzhou Municipal Biomedical Special Project (grant number 2023WJC120).

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

JL conceived the study, designed the methodology, and wrote the original draft. QF and WX contributed equally to data collection, analysis, and manuscript revision. HY, HT, and YL participated in data curation, validation, and discussion. All authors reviewed and approved the final manuscript. QF is the corresponding author. JL and WX contributed equally to this work.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed literature search strategies.
[PDF File, 104 KB - jmir_v28i1e89862_app1.pdf]

Multimedia Appendix 2
Methodological classification, appraisal tools, and evaluation metrics of the included studies.
[PDF File, 108 KB - jmir_v28i1e89862_app2.pdf]

Multimedia Appendix 3
Documentation of framework-preserving adaptations to quality appraisal tools.
[PDF File, 145 KB - jmir_v28i1e89862_app3.pdf]

Multimedia Appendix 4
Prompt engineering strategies and application scenarios in the included studies.
[PDF File, 87 KB - jmir_v28i1e89862_app4.pdf]

Checklist 1
PRISMA - S checklist.
[PDF File, 129 KB - jmir_v28i1e89862_app5.pdf]

References

1. Wu S, Zhang Y, Lin Z, Wei M. Global burden of colorectal cancer in 2022 and projections to 2050: incidence and mortality estimates from GLOBOCAN. BMC Cancer 2025 Nov 14;25(1):1770. [doi: 10.1186/s12885-025-15138-0] [Medline: 41239247]
2. Eng C, Yoshino T, Ruíz-García E, et al. Colorectal cancer. The Lancet 2024 Jul;404(10449):294-310. [doi: 10.1016/S0140-6736(24)00360-X] [Medline: 38278165]
3. Sloss EA, Abdul S, Aboagyewah MA, et al. Toward alleviating clinician documentation burden: a scoping review of burden reduction efforts. Appl Clin Inform 2024 May;15(3):446-455. [doi: 10.1055/s-0044-1787007] [Medline: 38839063]
4. Holmgren AJ, Apathy NC, Crews J, Shanafelt T. National trends in oncology specialists' EHR inbox work, 2019-2022. J Natl Cancer Inst 2025 Jun 1;117(6):1253-1259. [doi: 10.1093/jnci/djaf052] [Medline: 40037649]
5. Wong EYT, Verlingue L, Aldea M, et al. ESMO guidance on the use of large language models in clinical practice (ELCAP). Ann Oncol 2025 Dec;36(12):1447-1457. [doi: 10.1016/j.annonc.2025.09.001] [Medline: 41111032]
6. Chen D, Alnassar SA, Avison KE, Huang RS, Raman S. Large language model applications for health information extraction in oncology: scoping review. JMIR Cancer 2025 Mar 28;11(1):e65984. [doi: 10.2196/65984] [Medline: 40153782]
7. Peng W, feng Y, Yao C, et al. Evaluating AI in medicine: a comparative analysis of expert and ChatGPT responses to colorectal cancer questions. Sci Rep 2024;14(1):2840. [doi: 10.1038/s41598-024-52853-3]
8. Zeng A, Steinke J, Bocse HF, De Pastena M. Dr. LLM will see you now: the ability of ChatGPT to provide geographically tailored colorectal cancer screening and surveillance recommendations. J Clin Med 2025 Jul 18;14(14):5101. [doi: 10.3390/jcm14145101] [Medline: 40725794]
9. Gong EJ, Bang CS, Lee JJ, et al. Large language models in gastroenterology: systematic review. J Med Internet Res 2024 Dec 20;26:e66648. [doi: 10.2196/66648] [Medline: 39705703]
10. Maida M, Ramai D, Mori Y, et al. The role of generative language systems in increasing patient awareness of colon cancer screening. Endoscopy 2025 Mar;57(3):262-268. [doi: 10.1055/a-2388-6084] [Medline: 39142348]
11. Yang EW, Waldrup B, Velazquez-Villarreal E. Conversational artificial intelligence for integrating social determinants, genomics, and clinical data in precision medicine: development and implementation study of the AI-HOPE-PM system. JMIR Bioinform Biotechnol 2025 Oct 10;6:e76553. [doi: 10.2196/76553] [Medline: 41342165]
12. Pereyra L, Schlottmann F, Steinberg L, Lasa J. Colorectal cancer prevention: is chat generative pretrained transformer (Chat GPT) ready to assist physicians in determining appropriate screening and surveillance recommendations? J Clin Gastroenterol 2024;58(10):1022-1027. [doi: 10.1097/MCG.0000000000001979] [Medline: 38319619]
13. Amini M, Chang PW, Davis RO, et al. Comparing ChatGPT3.5 and Bard recommendations for colonoscopy intervals: bridging the gap in healthcare settings. Endosc Int Open 2025;13(CP):a25865912. [doi: 10.1055/a-2586-5912] [Medline: 40611845]

XSL•FO
RenderX

14. Chang PW, Amini MM, Davis RO, et al. ChatGPT4 outperforms endoscopists for determination of postcolonoscopy rescreening and surveillance recommendations. Clin Gastroenterol Hepatol 2024 Sep;22(9):1917-1925. [doi: 10.1016/j.cgh.2024.04.022] [Medline: 38729387]
15. Omar M, Nassar S, SharIf K, Glicksberg BS, Nadkarni GN, Klang E. Emerging applications of NLP and large language models in gastroenterology and hepatology: a systematic review. Front Med (Lausanne) 2024;11:1512824. [doi: 10.3389/fmed.2024.1512824] [Medline: 39917263]
16. Naito T, Nosaka T, Tanaka T, et al. Usefulness of an artificial intelligence-based colonoscopy report generation support system. Clin Endosc 2025 Mar;58(2):327-330. [doi: 10.5946/ce.2024.213] [Medline: 40010702]
17. Johnson B, Bath T, Huang X, et al. Large language models for extracting histopathologic diagnoses of colorectal cancer and dysplasia from electronic health records. BMJ Open Gastroenterol 2025 Sep 18;12(1):e001896. [doi: 10.1136/bmjgast-2025-001896] [Medline: 40973184]
18. Bräutigam K, Baker AM, Koelzer VH, Kather JN, Graham TA. Integrating artificial intelligence (AI) into colorectal cancer reporting. J Pathol 2026 Apr;268(4):367-382. [doi: 10.1002/path.70029] [Medline: 41588707]
19. Yılmaz M, Abbaslı N, Tuna S, et al. Comparison of artificial intelligence and multidisciplinary team recommendations in the management of colorectal cancer liver metastases. Sci Rep 2026;16(1):7278. [doi: 10.1038/s41598-026-38449-z]
20. Qu B, Cao L, Wu C, et al. Comparison of large language models and expert multidisciplinary team decisions in colorectal cancer. BMJ Health Care Inform 2026 Mar 10;33(1):e101780. [doi: 10.1136/bmjhci-2025-101780] [Medline: 41806973]
21. Biesheuvel LA, Workum JD, Reuland M, et al. Large language models in critical care. J Intensive Med 2025 Apr;5(2):113-118. [doi: 10.1016/j.jointm.2024.12.001] [Medline: 40241839]
22. Emile SH, Horesh N, Freund M, et al. How appropriate are answers of online chat-based artificial intelligence (ChatGPT) to common questions on colon cancer? Surgery 2023 Nov;174(5):1273-1275. [doi: 10.1016/j.surg.2023.06.005] [Medline: 37482439]
23. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs). NPJ Digit Med 2024 Jul 8;7(1):183. [doi: 10.1038/s41746-024-01157-x] [Medline: 38977771]
24. Wang Q, Zou H, Zhang H, Huang Y, Tian J, Cheng W. A survey on medical competence evaluation benchmarks for large language models. Health Care Sci 2026 Feb;5(1):4-18. [doi: 10.1002/hcs2.70050] [Medline: 41767169]
25. Zhou S, Luo X, Chen C, et al. The performance of large language model-powered chatbots compared to oncology physicians on colorectal cancer queries. Int J Surg 2024 Oct 1;110(10):6509-6517. [doi: 10.1097/JS9.0000000000001850] [Medline: 38935100]
26. Jeon S, Kim HG. A comparative evaluation of chain-of-thought-based prompt engineering techniques for medical question answering. Comput Biol Med 2025 Sep;196(Pt A):110614. [doi: 10.1016/j.compbiomed.2025.110614] [Medline: 40602316]
27. Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. JMIR Med Inform 2024 Apr 8;12:e55318. [doi: 10.2196/55318] [Medline: 38587879]
28. Lim DYZ, Tan YB, Koh JTE, et al. ChatGPT on guidelines: providing contextual knowledge to GPT allows it to provide advice on appropriate colonoscopy intervals. J Gastroenterol Hepatol 2024 Jan;39(1):81-106. [doi: 10.1111/jgh.16375] [Medline: 37855067]
29. Amugongo LM, Mascheroni P, Brooks S, Doering S, Seidel J. Retrieval augmented generation for large language models in healthcare: a systematic review. PLOS Digit Health 2025 Jun;4(6):e0000877. [doi: 10.1371/journal.pdig.0000877] [Medline: 40498738]
30. Yang Y, Jin Q, Huang F, Lu Z. Adversarial prompt and fine-tuning attacks threaten medical large language models. Nat Commun 2025 Oct 9;16(1):9011. [doi: 10.1038/s41467-025-64062-1] [Medline: 41068092]
31. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digit Med 2024 Jan 24;7(1):20. [doi: 10.1038/s41746-024-01010-1] [Medline: 38267608]
32. Williams CYK, Miao BY, Kornblith AE, Butte AJ. Evaluating the use of large language models to provide clinical recommendations in the emergency department. Nat Commun 2024 Oct 8;15(1):8236. [doi: 10.1038/s41467-024-52415-1] [Medline: 39379357]
33. Zhong R, Chen S, Li Z, et al. Large language models in lung cancer: systematic review. J Med Internet Res 2025 Sep 30;27:e74177. [doi: 10.2196/74177] [Medline: 41026980]
34. Hao Y, Qiu Z, Holmes J, et al. Large language model integrations in cancer decision-making: a systematic review and meta-analysis. NPJ Digit Med 2025 Jul 17;8(1):450. [doi: 10.1038/s41746-025-01824-7] [Medline: 40676129]
35. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. Commun Med (Lond) 2023 Oct 10;3(1):141. [doi: 10.1038/s43856-023-00370-1] [Medline: 37816837]
36. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
37. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]

XSL•FO
RenderX

38. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ 2020 Jan 16;368:l6890. [doi: 10.1136/bmj.l6890] [Medline: 31948937]
39. Omar M, Levkovich I. Exploring the efficacy and potential of large language models for depression: a systematic review. J Affect Disord 2025 Feb 15;371:234-244. [doi: 10.1016/j.jad.2024.11.052] [Medline: 39581383]
40. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
41. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019 Jan 1;170(1):W1-W33. [doi: 10.7326/M18-1377] [Medline: 30596876]
42. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016 Oct 12;355:i4919. [doi: 10.1136/bmj.i4919] [Medline: 27733354]
43. Gorelik Y, Ghersin I, Maza I, Klein A. Harnessing language models for streamlined postcolonoscopy patient management: a novel approach. Gastrointest Endosc 2023 Oct;98(4):639-641. [doi: 10.1016/j.gie.2023.06.025] [Medline: 37385548]
44. Alzaid E, Pergola G, Evans H, Snead D, Minhas F. Large multimodal model-based standardisation of pathology reports with confidence and its prognostic significance. J Pathol Clin Res 2024 Nov;10(6):e70010. [doi: 10.1002/2056-4538.70010] [Medline: 39545631]
45. Atarere J, Naqvi H, Haas C, et al. Applicability of online chat-based artificial intelligence models to colorectal cancer screening. Dig Dis Sci 2024 Mar;69(3):791-797. [doi: 10.1007/s10620-024-08274-3] [Medline: 38267726]
46. Ferber D, Wölflein G, Wiest IC, et al. In-context learning enables multimodal large language models to classify cancer pathology images. Nat Commun 2024 Nov 21;15(1):10104. [doi: 10.1038/s41467-024-51465-9] [Medline: 39572531]
47. Kaiser KN, Hughes AJ, Yang AD, et al. Accuracy and consistency of publicly available large language models as clinical decision support tools for the management of colon cancer. J Surg Oncol 2024 Oct;130(5):1104-1110. [doi: 10.1002/jso.27821] [Medline: 39155667]
48. Kepez MS, Ugur F. Comparative evaluation of information quality on colon cancer for patients: a study of ChatGPT-4 and Google. Cureus 2024 Nov;16(11):e73989. [doi: 10.7759/cureus.73989] [Medline: 39703246]
49. Liu J, Liang X, Fang D, et al. The diagnostic ability of GPT-3.5 and GPT-4.0 in surgery: comparative analysis. J Med Internet Res 2024 Sep 10;26:e54985. [doi: 10.2196/54985] [Medline: 39255016]
50. Wang A, Zhou J, Zhang P, et al. Large language model answers medical questions about standard pathology reports. Front Med (Lausanne) 2024;11:1402457. [doi: 10.3389/fmed.2024.1402457] [Medline: 39359921]
51. Kim HB, Tan HQ, Nei WL, Tan YCRS, Cai Y, Wang F. Impact of large language models and vision deep learning models in predicting neoadjuvant rectal score for rectal cancer treated with neoadjuvant chemoradiation. BMC Med Imaging 2025 Jul 31;25(1):306. [doi: 10.1186/s12880-025-01844-5] [Medline: 40745280]
52. Chizhikova M, López-Úbeda P, Martín-Noguerol T, et al. Automatic TNM staging of colorectal cancer radiology reports using pre-trained language models. Comput Methods Programs Biomed 2025 Feb;259:108515. [doi: 10.1016/j.cmpb.2024.108515] [Medline: 39602989]
53. Horesh N, Emile SH, Gupta S, et al. Comparing the management recommendations of large language model and colorectal cancer multidisciplinary team: a pilot study. Dis Colon Rectum 2025 Jan 1;68(1):41-47. [doi: 10.1097/DCR.0000000000003504] [Medline: 39679608]
54. Yang X, Xu J, Ji H, Li J, Yang B, Wang L. Early prediction of colorectal adenoma risk: leveraging large-language model for clinical electronic medical record data. Front Oncol 2025;15:1508455. [doi: 10.3389/fonc.2025.1508455] [Medline: 40444092]
55. Zhang Z, Zhang ZC, Zhang SP, et al. Comparative analysis of artificial intelligence tools for the dissemination of colorectal cancer screening guidelines: a novel perspective on early screening education. Int J Surg 2025 Nov 1;111(11):8616-8620. [doi: 10.1097/JS9.0000000000002951] [Medline: 40607944]
56. Zeng L, Cao Q, Deng J, Hu J, Pang M, Liu F. Guideline adherence in surgical decisions for T1 colorectal cancer after endoscopic resection: large language models vs clinicians. Int J Surg 2026 Jan 1;112(1):1886-1890. [doi: 10.1097/JS9.0000000000003492] [Medline: 40928382]
57. Yu Z, Fang L, Ding Y, et al. Evaluating large language models for information extraction from gastroscopy and colonoscopy reports through multi-strategy prompting. J Biomed Inform 2025 Aug;168:104844. [doi: 10.1016/j.jbi.2025.104844] [Medline: 40505790]
58. Yang EW, Waldrup B, Velazquez-Villarreal E. Conversational AI agent for precision oncology: AI-HOPE-WNT integrates clinical and genomic data to investigate WNT pathway dysregulation in colorectal cancer. Front Artif Intell 2025;8:1624797. [doi: 10.3389/frai.2025.1624797] [Medline: 40860720]
59. Wang S, Zhu Y, Yang Z, et al. Leveraging large language and vision models for knowledge extraction from large-scale image-text colonoscopy records. Nat Biomed Eng 2025 Sep 16. [doi: 10.1038/s41551-025-01500-x] [Medline: 40958005]
60. Sehgal NKR, Tonneau M, Tan A, et al. Effect of static vs. conversational AI-generated messages on colorectal cancer screening intent: a randomized controlled trial. arXiv. Preprint posted online on Jul 10, 2025. [doi: 10.48550/arXiv.2507.08211]

XSL•FO
RenderX

61. Schmutz M, Sommer S, Sander J, et al. Large language model processing capabilities of ChatGPT 4.0 to generate molecular tumor board recommendations-a critical evaluation on real world data. Oncologist 2025 Oct 1;30(10):oyaf293. [doi: 10.1093/oncolo/oyaf293] [Medline: 40973166]
62. Massimi D, Carlini L, Mori Y, et al. Large language model for interpreting the Paris classification of colorectal polyps. Endosc Int Open 2025;13(CP):a27030209. [doi: 10.1055/a-2703-0209] [Medline: 41079216]
63. Maida M, Mori Y, Fuccio L, et al. Exploring ChatGPT effectiveness in addressing direct patient queries on colorectal cancer screening. Endosc Int Open 2025;13(CP):a25689416. [doi: 10.1055/a-2568-9416] [Medline: 40376022]
64. Hu Y, Wang S, Cai P, Artificial Intelligence Colorectal Cancer Research (AI-CORE) Working Group. Multidimensional assessment of ChatGPT in colorectal cancer postoperative consultations: analysing response variations across critical clinical domains. Digit Health 2025;11:20552076251393297. [doi: 10.1177/20552076251393297] [Medline: 41181549]
65. Ding L, Fan L, Shen M, et al. Evaluating ChatGPT's diagnostic potential for pathology images. Front Med 2025;11:1507203. [doi: 10.3389/fmed.2024.1507203] [Medline: 41405831]
66. Diaz FC, Waldrup B, Carranza FG, Manjarrez S, Velazquez-Villarreal E. Artificial intelligence-enhanced precision medicine reveals prognostic impact of TGF-beta pathway alterations in FOLFOX-treated early-onset colorectal cancer among disproportionately affected populations. Int J Mol Sci 2025 Sep 17;26(18):9067. [doi: 10.3390/ijms26189067] [Medline: 41009631]
67. Chatziisaak D, Burri P, Sparn M, Hahnloser D, Steffen T, Bischofberger S. Concordance of ChatGPT artificial intelligence decision-making in colorectal cancer multidisciplinary meetings: retrospective study. BJS Open 2025 May 7;9(3):zraf040. [doi: 10.1093/bjsopen/zraf040] [Medline: 40331891]
68. Garg SK, Mau B, Hubers J, et al. Colon-Pilot: a generative AI tool for automated colonoscopy surveillance recommendations and 2024 ACG/ASGE quality benchmarking. Am J Gastroenterol 2026 Apr 1;121(4):964-973. [doi: 10.14309/ajg.0000000000003946] [Medline: 41665228]
69. Kim JS, Baek SJ, Ryu HS, et al. Using large language models for clinical staging of colorectal cancer from imaging reports: a pilot study. Ann Surg Treat Res 2025 Nov;109(5):318-327. [doi: 10.4174/astr.2025.109.5.318] [Medline: 41255477]
70. Wang L, Ma Y, Bi W, Lv H, Li Y. An entity extraction pipeline for medical text records using large language models: analytical study. J Med Internet Res 2024 Mar 29;26:e54580. [doi: 10.2196/54580] [Medline: 38551633]
71. Chen RJ, Ding T, Lu MY, et al. Towards a general-purpose foundation model for computational pathology. Nat Med 2024 Mar;30(3):850-862. [doi: 10.1038/s41591-024-02857-3] [Medline: 38504018]
72. Zhu M, Lin H, Jiang J, et al. Large language model trained on clinical oncology data predicts cancer progression. NPJ Digit Med 2025 Jul 2;8(1):397. [doi: 10.1038/s41746-025-01780-2] [Medline: 40604229]
73. Tariq R, Malik S, Khanna S. Evolving landscape of large language models: an evaluation of ChatGPT and Bard in answering patient queries on colonoscopy. Gastroenterology 2024 Jan;166(1):220-221. [doi: 10.1053/j.gastro.2023.08.033] [Medline: 37634736]
74. Maida M, Celsa C, Lau LHS, et al. The application of large language models in gastroenterology: a review of the literature. Cancers (Basel) 2024 Sep 28;16(19):3328. [doi: 10.3390/cancers16193328] [Medline: 39409948]
75. Jonnagaddala J, Shulajkovska M, Gradišek A, et al. Multimodal analysis of whole slide images in colorectal cancer. NPJ Digit Med 2025 Nov 24;8(1):719. [doi: 10.1038/s41746-025-02095-y] [Medline: 41286436]
76. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature New Biol 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
77. Giuffrè M, Kresevic S, Pugliese N, You K, Shung DL. Optimizing large language models in digestive disease: strategies and challenges to improve clinical outcomes. Liver Int 2024 Sep;44(9):2114-2124. [doi: 10.1111/liv.15974] [Medline: 38819632]
78. Fraile Navarro D, Ijaz K, Rezazadegan D, et al. Clinical named entity recognition and relation extraction using natural language processing of medical free text: a systematic review. Int J Med Inform 2023 Sep;177:105122. [doi: 10.1016/j.ijmedinf.2023.105122] [Medline: 37295138]
79. Guo S, Shariatmadari AH, Xiong G, Zhang A. Embracing foundation models for advancing scientific discovery. Presented at: 2024 IEEE International Conference on Big Data (BigData); Dec 15-18, 2024; Washington, DC, USA p. 1746-1755. [doi: 10.1109/BigData62323.2024.10825618]
80. Abdel-Rehim A, Zenil H, Orhobor O, et al. Scientific hypothesis generation by large language models: laboratory validation in breast cancer treatment. J R Soc Interface 2025 Jun;22(227):20240674. [doi: 10.1098/rsif.2024.0674] [Medline: 40462712]
81. Sun D, Hadjiiski L, Gormley J, et al. Outcome prediction using multi-modal information: integrating large language model-extracted clinical information and image analysis. Cancers (Basel) 2024 Jun 29;16(13):2402. [doi: 10.3390/cancers16132402] [Medline: 39001463]
82. Kocak Z. Publication ethics in the era of artificial intelligence. J Korean Med Sci 2024 Aug 26;39(33):e249. [doi: 10.3346/jkms.2024.39.e249] [Medline: 39189714]
83. Chen D, Parsa R, Swanson K, et al. Large language models in oncology: a review. BMJ Oncol 2025;4(1):e000759. [doi: 10.1136/bmjonc-2025-000759] [Medline: 40519217]
84. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med 2024 Sep;30(9):2613-2622. [doi: 10.1038/s41591-024-03097-1] [Medline: 38965432]

XSL•FO
RenderX

85. Huang J, Yang DM, Rong R, et al. A critical assessment of using ChatGPT for extracting structured data from clinical notes. NPJ Digit Med 2024 May 1;7(1):106. [doi: 10.1038/s41746-024-01079-8] [Medline: 38693429]
86. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 2020 Sep;26(9):1364-1374. [doi: 10.1038/s41591-020-1034-x] [Medline: 32908283]
87. Group S. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. Nat Med 2021 Feb;27(2):186-187. [doi: 10.1038/s41591-021-01229-5] [Medline: 33526932]
88. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health 2024 Jan;6(1):e12-e22. [doi: 10.1016/S2589-7500(23)00225-X] [Medline: 38123252]
89. Tschandl P, Rinner C, Apalla Z, et al. Human-computer collaboration for skin cancer recognition. Nat Med 2020 Aug;26(8):1229-1234. [doi: 10.1038/s41591-020-0942-0] [Medline: 32572267]
90. Wornow M, Xu Y, Thapa R, et al. The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit Med 2023 Jul 29;6(1):135. [doi: 10.1038/s41746-023-00879-8] [Medline: 37516790]
91. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature New Biol 2023 Apr;616(7956):259-265. [doi: 10.1038/s41586-023-05881-4] [Medline: 37045921]
92. Zakka C, Shad R, Chaurasia A, et al. Almanac - retrieval-augmented language models for clinical medicine. NEJM AI 2024 Feb;1(2). [doi: 10.1056/aioa2300068] [Medline: 38343631]
93. Artsi Y, Sorin V, Glicksberg BS, Korfiatis P, Nadkarni GN, Klang E. Large language models in real-world clinical workflows: a systematic review of applications and implementation. Front Digit Health 2025;7:1659134. [doi: 10.3389/fdgth.2025.1659134] [Medline: 41098649]
94. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv. Preprint posted online on Mar 20, 2023. [doi: 10.48550/arXiv.2303.13375]
95. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv. Preprint posted online on Jan 10, 2023. [doi: 10.48550/arXiv.2201.11903]
96. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A review of challenges and opportunities in machine learning for health. AMIA Jt Summits Transl Sci Proc 2020;2020:191-200. [Medline: 32477638]
97. Cervantes A, Adam R, Roselló S, et al. Metastatic colorectal cancer: ESMO clinical practice guideline for diagnosis, treatment and follow-up. Ann Oncol 2023 Jan;34(1):10-32. [doi: 10.1016/j.annonc.2022.10.003] [Medline: 36307056]
98. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med 2018 Mar 15;378(11):981-983. [doi: 10.1056/NEJMp1714229] [Medline: 29539284]
99. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1):310. [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
100. Ahmed M, Whicher D, Israni ST. Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril: National Academy of Medicine; 2023. URL: http://www.ncbi.nlm.nih.gov/books/NBK605955 [accessed 2026-03-18]
101. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019 Jun;6(2):94-98. [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]

Abbreviations

CRC: colorectal cancer
EHR: electronic health record
LLM: large language model
NLP: natural language processing
PICOS: Population, Intervention, Comparison, Outcome, Study design
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
PROBAST: prediction model risk of bias assessment tool
PROSPERO: prospective register of systematic reviews
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies-2
RAG: retrieval-augmented generation
ROBINS-I: Risk of Bias in Nonrandomized Studies - of Interventions
SWiM: synthesis without meta-analysis
TNM: tumor–node–metastasis

XSL•FO
RenderX

Edited by S Brini; submitted 22.Dec.2025; peer-reviewed by A Sagkriotis, I Tosun; revised version received 26.Apr.2026; accepted 27.Apr.2026; published 21.May.2026.

Please cite as:
Tian J, Lou Q, Wang X, Xu H, Mei H, Yu Y
Large Language Models in Colorectal Cancer Care and Clinical Decision Support: Systematic Review
J Med Internet Res 2026;28:e89862
URL: https://www.jmir.org/2026/1/e89862
doi:10.2196/89862

© Jinglei Tian, Qifeng Lou, Xue Wang, Hangying Xv, Huiting Mei, Yanli Yu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effectiveness of mHealth Interventions Aimed at Promoting Physical Activity and Reducing Sedentary Behavior on Work-Related Outcomes Among Workers: Systematic Review

Takako Miki[1], MPH, PhD; Michiko Nohara[1], MD, MWH, PhD; Nami Homma[1], BA; Daiki Nagamine[1], PhD; Shinji Yamaguchi[1], PhD; Kyoko Nomura[2], MD, MPH, PhD

[1]Division of Public Health, Department of Hygiene and Public Health, School of Medicine, Tokyo Women's Medical University, 8-1 Kawada-cho, Shinjuku-ku, Tokyo, Japan

[2]Department of Environmental Health Science and Public Health, Akita University Graduate School of Medicine, Akita, Japan

Corresponding Author:
Takako Miki, MPH, PhD
Division of Public Health, Department of Hygiene and Public Health, School of Medicine, Tokyo Women's Medical University, 8-1 Kawada-cho, Shinjuku-ku, Tokyo, Japan

Abstract

Background: Mobile health (mHealth) technologies have gained popularity and may play a key role in promoting health-related behaviors. However, the impact of mHealth interventions on work-related outcomes remains unclear.

Objective: We aimed to provide an overview of the effectiveness of mHealth interventions designed to encourage physical activity and decrease sedentary behavior on work-related outcomes.

Methods: A search was conducted in MEDLINE, Web of Science Core Collection, the Cochrane Central Register of Controlled Trials and Cochrane Database of Systematic Reviews, and Ichushi-Web for all publications up to September 12, 2025, without language or date restrictions. We included studies which (1) investigated the impact of mHealth interventions promoting physical activity and reducing sedentary behavior on work-related outcomes such as absenteeism, presenteeism, work productivity, work performance, and workability; (2) were designed as a randomized controlled trial or nonrandomized study of interventions; (3) were conducted among workers; and (4) were published as full-text original articles in Japanese or English. Two researchers independently conducted the screening of titles and abstracts, followed by a full-text review to confirm study eligibility and extract the data. We assessed the risk of bias using the Cochrane Collaboration Risk of Bias tool for randomized trials (RoB 2) and the Risk Of Bias In Nonrandomized Studies of Interventions (ROBINS-I). Findings were narratively synthesized.

Results: In total, 4022 records were identified from databases and other sources. After removing duplicates, 3631 studies were screened, and 17 studies (12 randomized controlled trials and 5 nonrandomized studies of interventions) met the inclusion criteria. Of the 17 eligible studies with 18,805 participants, 12 showed a favorable effect direction on ≥1 work-related outcome, but the majority exhibited a high or critical risk of bias, raising concerns about the quality of the research. There was considerable heterogeneity across studies.

Conclusions: Scientific evidence suggests that health promotion efforts that incorporate mHealth components to encourage physical activity and reduce sedentary behavior may have a favorable impact on work-related outcomes. However, the low methodological quality of the studies and their high heterogeneity due to the participant characteristics, intervention content, use of diverse work-related assessment tools, and different follow-up periods make it difficult to draw definitive conclusions about the effectiveness of the interventions. This review is the first to comprehensively synthesize evidence on the impact of mHealth interventions on work-related outcomes, extending prior reviews that primarily focused on traditional workplace-based interventions without mHealth components. This review also contributes to the field by offering guidance for future studies, including the potential value of megastudy designs that evaluate multiple interventions within the same population over the same period on the same outcomes. As a practical implication, these findings suggest that mHealth technologies may serve as a complementary strategy for supporting occupational health and productivity.

Trial Registration: UMIN Clinical Trials Registry UMIN000052290; https://center6.umin.ac.jp/cgi-open-bin/ctr_e/ctr_view.cgi?recptno=R000059690

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2023-080240

(J Med Internet Res 2026;28:e80540) doi:10.2196/80540

XSL•FO
RenderX

KEYWORDS

mHealth; physical activity; productivity; sedentary behavior; workers; mobile health

Introduction

Physical inactivity, defined as failure to meet established physical activity guidelines [1], is a major global public health issue in contemporary society with more than 5 million attributed deaths per year [2]. In addition to physical inactivity, sedentary behavior—defined as postures (eg, sitting, reclining, or lying) with characteristic energy expenditure of ≤1.5 metabolic equivalents of task while awake [3]—is also a serious health concern, due to its wide range of adverse health effects, independent of physical activity [4,5]. In particular, working populations are a priority population for these issues, given that a large proportion of their working time is sedentary [6]. Changes in industrial structure have transformed the workplace from one requiring physical labor to one dominated by sedentary work, thereby exposing workers to a range of health risks [7]. Moreover, remote working arrangements have expanded in recent years, and evidence suggests that working from home or teleworking is associated with increased sedentary behavior and reduced physical activity compared to onsite work [8]. These contextual changes underscore the need for scalable and adaptable strategies to reach workers across diverse work settings.

Due to the widespread availability and ubiquity of portable devices, health interventions using mobile health (mHealth) technologies [9] have gained attention for their potential to support lifestyle behaviors [10-12]. mHealth technology, characterized by the apps of portable and wireless systems intended to enhance wellness outcomes, bolster health care delivery, and facilitate medical research [13], offers numerous benefits for promoting health, including accessibility, individualization, real-time feedback, and potential scalability [14]. Several umbrella reviews have reported favorable effects of digital and mHealth interventions on health-related outcomes, such as physical activity and sedentary behavior [10-12]. In addition, a growing body of evidence suggests that higher levels of physical activity and lower levels of sedentary behavior are associated with favorable physical health (eg, musculoskeletal pain [15], cancer, and cardiometabolic diseases [2,4]) as well as mental health (eg, depression) [16,17]. These health conditions are commonly associated with productivity loss [18]. Consequently, strategies targeting these behaviors may in turn contribute to improved work-related outcomes. Consistently, evidence from systematic reviews [19,20] suggests potential beneficial effects of workplace interventions targeting physical activity and/or sedentary behavior on work-related outcomes; however, existing evidence remains limited and precludes drawing definitive conclusions. Moreover, these previous reviews have primarily focused on workplace interventions and have not incorporated mHealth technologies as a key component, thereby limiting the generalizability of their findings. Common barriers to workplace health promotion initiatives include financial constraints and shortages of personnel or expertise [21]. Such challenges can be mitigated through the use of mHealth technologies, which have the potential to provide scalable solutions [14]. To our knowledge, however, systematic reviews or meta-analyses have not assessed the impact of mHealth interventions aimed at promoting physical activity and decreasing sedentary behavior on work-related outcomes.

Given the priority that employers and stakeholders assign to work-related productivity and performance concerns [22], a systematic review of these interventions may provide valuable insights for occupational health management. Considering these issues, the objective of this systematic review is to evaluate the impact of mHealth interventions designed to promote physical activity and reduce sedentary behavior on work-related outcomes, including absenteeism, presenteeism, work productivity, work performance, and workability among workers in comparison with any type of control or comparison condition.

Methods

Protocol and Registration

The study protocol was registered in the UMIN Clinical Trials Registry (ID UMIN000052290) and was published elsewhere [23]. This systematic review adhered to the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) statement [24]. The PRISMA 2020 checklist is provided in Checklist 1. In addition, the search strategy and reporting follow the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) [25] to ensure full transparency and reproducibility of search methods (Checklist 2).

Search Strategy and Databases

As described in the protocol [23], the first systematic search of articles was conducted on September 23, 2023, to retrieve articles published from inception until that date. MEDLINE (via PubMed), Web of Science Core Collection (via Web of Science), the Cochrane Library (including Cochrane Central Register of Controlled Trials [CENTRAL] and Cochrane Database of Systematic Reviews), and Ichushi-Web (Japan Medical Abstracts Society) were searched.

ClinicalTrials.gov and the World Health Organization (WHO) International Clinical Trials Registry Platform (ICTRP) were reviewed via the CENTRAL. Proceedings papers indexed in Web of Science Core Collection and conference proceedings included in CENTRAL were reviewed. Study registries and conference proceedings were reviewed to determine whether full reports of eligible studies had been published. No published or validated search filters were used; the search strategies were developed specifically for this review. The search strategy was adapted from the protocol of a prior scoping review of digital health tools [26] and adjusted appropriately for the current systematic review. The search strategy was a combination of search strings that covered the areas of (1) mHealth technology, (2) physical activity and sedentary behavior, (3) workers (employees), and (4) work-related outcomes. The search strings were first developed within PubMed and then modified for use in the other databases. Consistency across databases was

XSL•FO
RenderX

maintained where possible. Terms were searched within multiple fields where available (eg, MeSH [Medical Subject Headings], keywords, title, and abstract). The overall search terms were developed by 2 researchers (TM and KN) and medical librarians at Tokyo Women's Medical University, who reviewed them to confirm that the required search terms were included and that the overall search approach was appropriate. No date or study design restrictions were applied in the database searches; studies were retrieved from database inception to the search dates. Language restrictions were not applied at the search stage, although studies written in languages other than English and Japanese were excluded during the screening process. In the Japan Medical Abstracts Society database, conference proceedings were excluded to focus on full-text original articles. Snowball searching was performed using the included studies to identify additional relevant literature. The full search strategies for all databases, exactly as executed, are provided in Multimedia Appendix 1. The searches were updated by rerunning the original strategies in each database using the same terms and procedures. All databases were last searched on September 12, 2025. To identify additional studies not captured by electronic database searches, we conducted backward citation searching by hand-screening the reference lists of included studies and relevant reviews. No forward citation searching (eg, using citation indexes or alerts) was performed. No additional information sources or search methods were used beyond the electronic database searches and backward citation searching described above.

Eligibility Criteria

Population, Intervention, Comparison, and Outcome (PICO) of studies in this systematic review were defined as follows: (P) inclusion of all workers (employees) aged 18 years or older regardless of race and ethnicity; (I) interventions using an mHealth intervention aimed at addressing one or both of the lifestyle factors of physical activity and sedentary behavior as a predominant component; (C) any type of control or comparator condition for comparison (eg, usual care, waitlist group, active control group, or pre- and post-data); and (O) work-associated outcomes (eg, absenteeism, presenteeism, work productivity, work performance, and workability). Health-related behaviors such as physical activity and sedentary behavior were also reported, although they were not definite inclusion criteria. All behavioral measurement types and units were accepted (eg, self-report and steps per day, and daily or weekly moderate-to-vigorous physical activity time).

Inclusion Criteria

Studies were included based on the following criteria:

1. Studies targeting all workers without consideration of employment status (full-time or part-time), job type, or shift type. No restrictions were placed on current diseases or risk factors.
2. Studies including mHealth interventions encouraging physical activity and decreasing sedentary behavior, for disease prevention and treatment or management of health conditions (eg, back pain or diabetes) aimed at workers. Workers with these conditions were included due to the substantial burden of lost work productivity involved [18]. Intervention setting was limited to nonclinical settings (eg, workplace, home, or outdoors).
3. Studies including mHealth interventions via mobile devices: mobile phones, smartphones, personal digital assistants, tablets, wearable activity monitors or trackers, mobile apps, SMS, and other wireless devices. Consistent with a previous systematic review [27], standard pedometers (unable to transmit data to consumer interface) were not considered mHealth components due to the lack of an electronic communication function with mobile interfaces and the internet [28].
4. Studies using mHealth technologies to promote physical activity or reduce sedentary behavior alone or with other behaviors (eg, diet). Multicomponent strategies were included where mHealth technology was one of several intervention components (eg, face-to-face counseling), as a prior review reported many mHealth interventions for physical activity and sedentary behavior were delivered as part of multicomponent interventions [29]. Inclusion of multicomponent strategies broadened the range of eligible studies and review scope. Although multicomponent interventions may increase heterogeneity and make determining whether effects are attributable to mHealth or other components difficult, mHealth interventions alone are often associated with low adherence and high attrition rates [30,31]. Given that the National Institute for Health and Care Excellence (NICE) committee recommended considering digital and mHealth interventions alongside existing services [32], multicomponent interventions were included.
5. Studies with any type of control or comparator condition for comparison.
6. Studies assessing work-related outcomes: absenteeism, presenteeism, work productivity, work performance, and workability.
7. Studies published as full-text original articles.
8. Studies conducted as randomized controlled trials (RCTs) or nonrandomized studies of interventions (NRSIs) with predefined control or comparator group (eg, nonrandomized controlled trials, cohort studies, and pre-post studies) [33]. NRSI design was included based on a scientific overview of occupational health interventions, encouraging researchers to include RCT alternatives when conducting systematic reviews [34]. This methodology provides broad evidence to guide real-world implementation where RCT would be difficult or inappropriate. Current guidance considers NRSI evidence prominent in decision-making [33]. Pilot and feasibility trials were included if they met the inclusion criteria. Quantitative and mixed methods studies were included to ensure all relevant quantitative results were obtained; for mixed methods, only quantitative results were extracted.
9. Studies published in Japanese or English. In addition to English studies, Japanese studies were also included to capture domestic research trends. No translation procedures were required for these languages in this review. To minimize language bias, no language restrictions were applied during literature searching, but studies were limited to Japanese and English during study selection.

XSL•FO
RenderX

Exclusion Criteria

Studies were excluded based on the following criteria:

1. Studies reporting results from samples mixing workers with other groups (eg, full-time students or retired older adults), unless results are reported independently for each group.
2. Non–full-text original articles (eg, editorials, reviews, protocols, letters to the editor, published conference abstracts, dissertations, books, and case reports or case series). Conference proceedings, trial registry records, and protocol papers identified during screening were not included as eligible studies in this review. However, their details were reviewed to confirm whether any original articles relevant to this review had been published.
3. Studies having an inappropriate design (eg, without a control or comparator group, accuracy or validation studies, descriptive articles, exclusively qualitative research, and mixed methods with quantitative results unsuitable for inclusion).

Study Selection Process

Records identified from the searches were managed in Microsoft Excel. Duplicate records were identified and removed manually by TM through comparison of titles and bibliographic information (eg, DOI). After duplicate removal, 2 researchers (TM and either NH, DN, or SY) independently performed initial screening by reading the titles and abstracts, then obtained the full text of retained studies for independent full-text screening. At the first screening, the kappa statistic of each pair of researchers was 0.77. The corresponding value at the full-text review stage was 0.76. Inconsistent assessments were resolved through discussion among the authors. The reasons for excluding particular studies are recorded in Multimedia Appendix 2. During the screening and selection process, TM contacted the corresponding authors of potentially eligible studies to clarify unclear information, request additional data when necessary, and confirm eligibility for inclusion in the review.

Data Extraction

Information from each of the included studies was extracted by 2 researchers (TM and either NH, DN, or SY) using a standardized data extraction form. The data to be extracted were predetermined, and their details were described in the protocol [23]. Briefly, information included author, year of publication, country, study design, participant information, intervention details, control or comparison group, work-related outcomes, summary results for each study, adverse events, adherence to the intervention, and source of funding. After extraction, information was confirmed through discussion among all authors until consensus in data collection was reached. If the studies did not include this information and/or contained unclear information, TM contacted the corresponding authors for clarification. All data extracted from each individual study included in the results of this systematic review are provided in a data extraction table (extracted data in Multimedia Appendix 3).

Quality Assessment

Bias risk for RCTs was assessed using the Revised Cochrane Collaboration Risk-of-Bias tool (RoB 2) [35], which assesses five risk domains, namely (1) randomization process; (2) deviations from the intended intervention; (3) missing outcome data; (4) outcome measurement; and (5) selection of reported results. For cluster RCTs, we assessed the 5 risk domains above and an additional domain for cluster RCTs [35]. Risk of bias (RoB) for NRSIs was assessed using the Cochrane ROBINS-I (Risk Of Bias In Nonrandomized Studies of Interventions) tool [36], which assesses bias based on (1) confounding factors, (2) participant selection, (3) classification of interventions, (4) deviations from intended interventions, (5) missing outcome data, (6) outcome measurement, and (7) selection of reported results. The RoB in each included study was independently evaluated by researchers (NH, DN, and SY) and confirmed by TM. Interrater agreement for RoB assessments was 73.9% (17/23 outcomes). Disagreements were resolved through discussion until a consensus was reached.

Data Synthesis

Due to substantial heterogeneity among the included studies in participant characteristics, intervention content, work-related outcome measures, and follow-up periods, quantitative synthesis through meta-analysis could not be conducted. Instead, a narrative synthesis approach was used. The certainty in the body of evidence (eg, using the Grading of Recommendations Assessment, Development and Evaluation [GRADE] system) was not formally assessed because the review adopted a narrative synthesis approach. The substantial heterogeneity among the included studies precluded the meaningful application of quantitative certainty grading systems. A narrative synthesis of the findings was conducted in line with the Synthesis Without Meta-Analysis (SWiM) guideline (Multimedia Appendix 4) [37]. We narratively synthesized the direction of results of individual studies using the vote-counting system [38] and visualized them using the modified effect direction plot [38,39]. The revised effect direction plot provides a comprehensive summary of the impact of mHealth interventions on work-related outcomes and health-related outcomes (Table S1 in Multimedia Appendix 3). Studies were ordered first by intervention category based on mHealth tools (mobile apps, web-based components, combinations of mobile apps and wearable activity monitors or trackers) and then chronologically by year of publication within each category. In addition, studies were further categorized based on the primary behavioral target of the intervention (physical activity, sedentary behavior, or both) to explore whether intervention focus influenced work-related outcomes (Table S2 in Multimedia Appendix 3). Heterogeneity was explored by comparing the direction of effects across these subgroups. Patterns of consistency or variation in effect direction were examined narratively and visually using the modified effect direction plot. All eligible studies, regardless of RoB or study design, contributed to the vote-counting synthesis. However, the overall conclusions were based on a comprehensive assessment, including methodological quality (RoB) and heterogeneity across studies. Following recommendations in the Cochrane Handbook for Systematic Reviews of Interventions, vote counting based on direction of effect was used for the synthesis of results [38,40]. In line with the guidance [38,40], we did not take statistical significance or magnitude of effect into account within this synthesis. We did

XSL•FO
RenderX

not consider the statistical significance of effects as attention to this feature may lead to the exclusion of underpowered studies [40]. Owing to the highly heterogeneous study characteristics (such as the lack of a consistent effect measure), we relied on the direction of effects as a standardized binary metric. Effect estimates were categorized as favorable (favoring the intervention group) or unfavorable (favoring the control or comparison group) based on the direction of effect. For work-related outcomes, increases in work productivity, work performance, and workability were considered favorable, whereas decreases in absenteeism and presenteeism were also considered favorable. For health-related outcomes, increases in physical activity were considered favorable, whereas decreases in sedentary behavior were considered favorable. To implement vote counting based on the direction of effect, we followed the study-level step-by-step procedures detailed by Boon and Thomson [38]:

- 1. We grouped studies for synthesis according to predefined categories of work-related outcomes (ie, absenteeism, presenteeism, work productivity, work performance, and work ability) and health-related outcomes (ie, physical activity and sedentary behavior). This grouping was chosen because these outcomes represent conceptually distinct domains and are commonly evaluated separately in prior occupational health research [19]. No changes were made to the categorization of these outcomes from those prespecified in the study protocol [23].
- 2. We counted favorable or unfavorable effect estimates for each type of work-related outcomes and health-related outcomes.
- 3. We determined the overall effect direction for each domain using the following algorithm. If multiple outcomes within a study reported effects in the same direction, that direction was assigned to the domain. If the direction of effect varied across outcomes, the domain was assigned the direction reported by ≥70% (ie, a clear majority) of outcomes. If fewer than 70% of outcomes showed a consistent direction of effect, or if effect estimates were identical between the intervention and control or comparison groups, or were not reported, the domain was classified as having no clear effect direction. Furthermore, we systematically tabulated the key characteristics and results of the included studies to facilitate comparison and interpretation. Data extracted from the included studies were ordered chronologically by publication date (extracted data in Multimedia Appendix 3). Individual study findings were reported using the effect measures as presented in the original articles, without reexpressing them using a different effect measure.

Results

Overview

The PRISMA 2020 flow diagram [24] of the study selection process is shown in Figure 1. The search of the 4 databases identified a total of 4018 records. After removing 391 duplicates, 3627 records were included in the screening on the basis of title and abstract, during which 3345 records were excluded, and 282 records were selected to proceed to full-text screening. Subsequently, we excluded 269 studies that did not meet the eligibility criteria for participants (n=121), intervention (n=78), outcomes (n=63), or publication type (n=7). Four additional studies identified through sources other than the database search met the eligibility criteria. Finally, 17 studies [41-57] were included in this review.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) 2020 diagram.

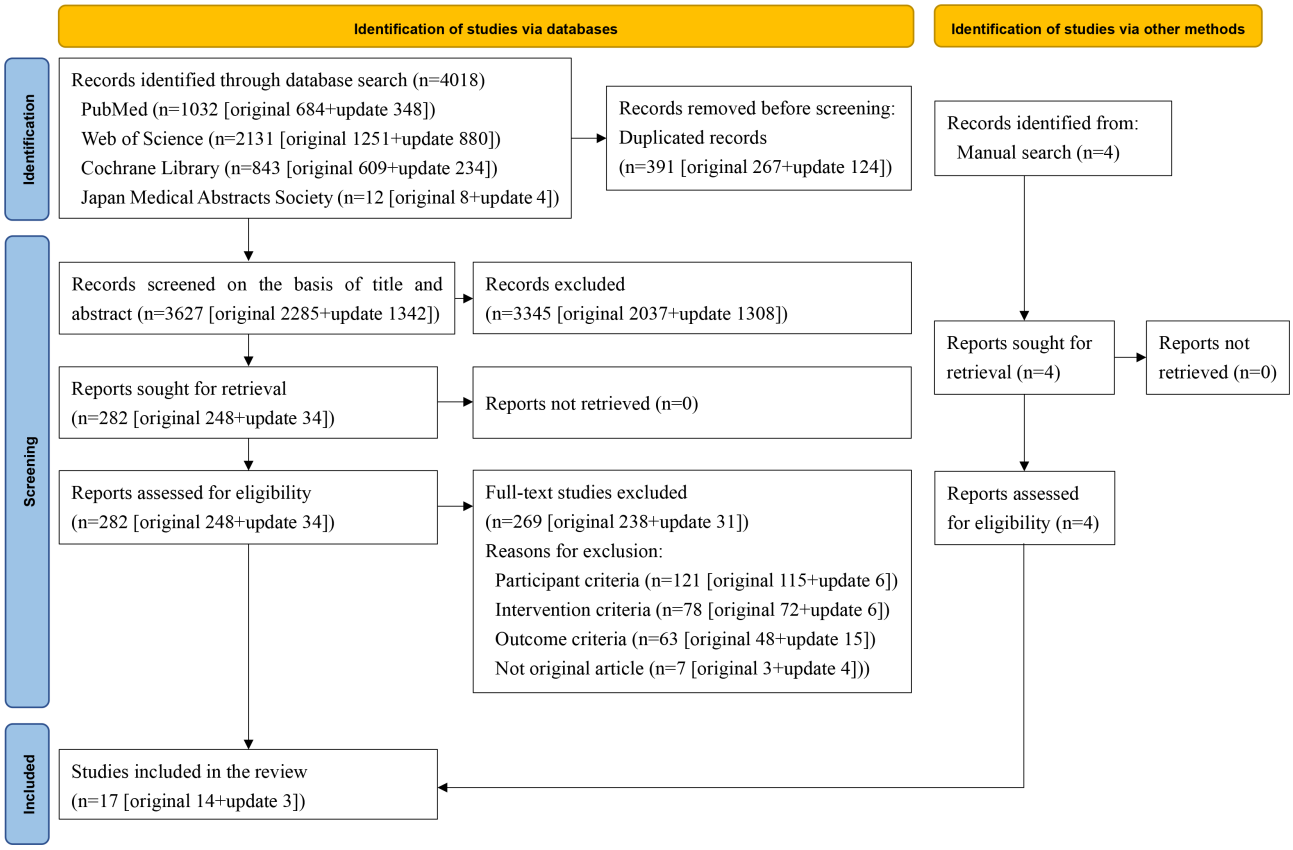

Study Characteristics

An overview of the study characteristics is given in Table 1. An extended version of the table is available in Multimedia Appendix 3 (extracted data). Out of the 17 included studies [41-57], 12 (70.6%) [41-52] were RCTs (including 4 cluster RCTs [49-52]) and 5 (29.4%) used NRSIs [53-57]. These NRSIs consisted of 3 retrospective cohort studies [53-55], 1 prospective cohort study [57], and 1 single-group, pre-post, mixed methods pilot study [56]. All included studies except one [44] (16/17, 94.1%) were conducted from 2016 onward [41-43,45-57]. Out of the 17 included studies [41-57], 7 (41.2%) were conducted in European countries: 2 each in England [52,56], Spain [46,51], and Finland [44,48], and 1 in Germany [42]. Of the remaining studies, 5 (29.4%) were conducted in North America: 3 in the United States [54,55,57], 2 in Canada [43,45], and 3 (17.6%) were conducted in Asia: 2 in Japan [41,53] and 1 in Thailand [49]. Other countries included Australia [50] and Brazil [47].

XSL•FO
RenderX

Table . The characteristics of the included studies (N=17).

Characteristics	Values
Study design, n (%)	
RCTs[a]	12 (70.6)
NRSIs[b]	5 (29.4)
Publication year, n (%)	
<2016	1 (5.9)
≥2016	16 (94.1)
Country, n (%)	
Europe	7 (41.2)
North America	5 (29.4)
Asia	3 (17.6)
Others	2 (11.8)
Total sample size, n (%)	
<100	4 (23.5)
101 - 300	8 (47.1)
301 - 1000	2 (11.8)
>1000	3 (17.6)
Target participants, n (%)	
Apparently healthy workers	10 (58.8)
Workers with certain health conditions	7 (41.2)
Percentage of females, n (%)	
<60%	9 (52.9)
≥60%	8 (47.1)
Mean age categories, n (%)	
<40 years old	4 (23.5)
≥40 years to <50 years old	11 (64.7)
≥50 years old	2 (11.8)
Duration of intervention, n (%)	
<12 weeks	3 (17.6)
≥12 weeks to <12 months	9 (52.9)
12 months	3 (17.6)
Not clearly described	2 (11.8)
Intervention focus, n (%)	
Targeted physical activity	11 (64.7)
Targeted sedentary behavior	3 (17.6)
Targeted both physical activity and sedentary behavior	3 (17.6)
Type of mHealth[c] tools, n (%)	
Mobile apps	6 (35.3)
Combinations of mobile apps and wearable activity monitors or trackers	8 (47.1)
Web-based components	3 (17.6)
Intervention strategies, n (%)	
mHealth-only	4 (23.5)

XSL•FO
RenderX

Characteristics	Values
Blended	13 (76.5)
Adherence to the intervention, n (%)	
Described	10 (58.8)
Not described	7 (41.2)
Control or comparison for work-related outcomes, n (%)	
Passive control	8 (47.1)
Active control	4 (23.5)
Pre-post comparison	5 (29.4)
Work-related outcomes, n	
Absenteeism	5
Presenteeism	2
Work productivity	7
Work performance	4
Workability	3
Health-related outcomes, n	
Physical activity	10
Sedentary behavior	7
Not described	7
Adverse event, n (%)	
Described	8 (47.1)
Not described	9 (52.9)

[a]RCT: randomized controlled trial.

[b]NRSI: nonrandomized studies of intervention.

[c]mHealth: mobile health.

Participant Characteristics

Total sample sizes ranged from <100 (4/17, 23.5%) [41,43,46,47], 101 - 300 (8/17, 47.1%) [45,48-52,55,56], 301 - 1000 (2/17, 11.8%) [42,44], and >1000 (3/17, 17.6%) [53,54,57]. Ten studies (10/17, 58.8%) focused on apparently healthy workers; the majority recruited employees who were identified as office workers [43,44,49-52], while one study each was performed in health care workers [45], police officers and staff [56], workers registered with an internet survey company [53], and military workers [48]. The other 7 studies (7/17, 41.2%) targeted workers with certain health conditions, namely metabolic syndrome [42]; off-work and undergoing rehabilitation for a wrist, hand, and/or finger injury [46]; chronic low back pain [41,47]; knee or low back pain [54]; prediabetes [55]; and pelvic floor disorders [57]. Out of the 17 included studies [41-57], 8 studies (47.1%) had a higher proportion of female participants (≥60%) [43-45,49,51,52,55,57]. The mean age of participants in the included studies ranged from 38.6 (SD 10.4) to 54.3 (SD 6.5) years. Most included studies (11/17, 64.7%) [41-46,48,51-54] reported mean participant ages between 40 and 49 years.

Intervention Characteristics

Intervention Duration and Target Behavior

The duration of the interventions ranged from 6 weeks to 12 months, with most studies reporting durations between 12 weeks and less than 12 months (9/17, 52.9%) [41,42,45,48,49,51,54-56]. Of the 17 studies reviewed [41-57], 11 (64.7%) primarily focused on promoting physical activity as their main intervention strategy [41,42,44-46,48,53-55,57], 3 (17.6%) focused on reducing sedentary behavior [43,50,52], and 3 (17.6%) targeted both physical activity and sedentary behavior [49,51,56].

Type of mHealth Tools

The main mHealth tools were combinations of mobile apps and wearable activity monitors or trackers (8/17, 47.1%) [42,48-50,54-57], followed by mobile apps (6/17, 35.3%) [41,45-47,52,53]. The remaining 3 studies (3/17, 17.6%) used web-based interventions [43,44,51].

Intervention Strategies

When interventions were defined as either mHealth-only or blended strategies (mHealth combined with other components such as personal counseling), 4 studies [45,47,51,53] were classified as mHealth-only and the remaining 13

XSL•FO
RenderX

[41-44,46,48-50,52,54-57] as blended interventions. The majority of studies (13/17, 76.5%) included components other than mHealth technology in their interventions.

Adherence

With regard to adherence to the intervention (ie, the extent to which users followed the program as designed), 10 studies (10/17, 58.8%) [41,42,44,45,49,50,52,54,56,57] reported information on adherence whereas the remaining 7 (7/17, 41.2%) [43,46-48,51,53,55] did not provide detailed information on adherence. In the absence of detailed information on adherence reported within the publication, additional data [58,59] were obtained through direct correspondence with the study authors. Due to the differences in defining and measuring adherence, as well as variations in the duration and content of the intervention across studies, comparing adherence rates between included studies was challenging. Overall, adherence to the intervention tended to decrease over time. Among mHealth devices, however, Fitbit devices tended to maintain high adherence rates [49,56,58]. One study reported high Fitbit adherence, with a mean wear time of 6.5 (SD 1.1) days per week and 21.4 (SD 4.1) hours per day at 8-month follow-up [56], whereas another study reported typical wear times of 10 - 15 hours per day at baseline, decreasing slightly to 8 - 14 hours per day after 2 - 3 months of the intervention (SDs for both time points were not reported in the original study) [49,58].

Control or Comparison Characteristics

We defined passive controls as those comprising no intervention, usual practice, or a waitlist, and active controls as comprising components intended to induce changes in one or more outcomes. Of the 17 included studies [41-57], 8 (47.1%) used passive controls [41-45,48,52,53], 4 (23.5%) used active controls [46,47,49,51], and 5 (29.4%) used pre-post comparison within the same group (pre-post design) for target work-related outcomes [50,54-57].

Work-Related Outcomes

The included studies were evaluated for work-related outcomes such as absenteeism, presenteeism, work productivity, work performance, and workability. The most common work-related outcome was work productivity in 7 studies [41,44,49,51,54,56,57], followed in order by absenteeism in 5 [44-46,52,55], work performance in 4 [43,50,52,53], workability in 3 [42,47,48], and presenteeism in 2 [51,52]. Only 4 work-related outcomes were assessed as primary outcomes [41,44,46]; the majority of work-related outcomes were measured as secondary [41-43,45,47,48,50-57] or tertiary outcomes [49]. No overall certainty assessment for each work-related outcome was conducted due to the narrative synthesis approach.

Work Productivity Outcomes

Results summarizing the impact of mHealth interventions on work-related outcomes and health-related outcomes are presented in Table S1 in Multimedia Appendix 3. Out of the 7 studies [41,44,49,51,54,56,57], 5 studies (71.4%) [41,49,51,54,57] reported favorable effect directions with work productivity outcomes in the intervention group compared with the control or comparison groups. Of the 5 studies that showed favorable effect directions [41,49,51,54,57], 3 [41,49,51] were RCTs (including 2 cluster RCTs [49,51]) and 2 [54,57] were NRSIs. Among these studies, 3 targeted physical activity [41,54,57], whereas 2 targeted both physical activity and sedentary behavior [49,51]. Regarding the mHealth components, one study [41] used mobile apps as part of the intervention, whereas another [51] conducted a web-based mHealth intervention. The remaining 3 studies [49,54,57] implemented mHealth tools combining mobile apps with wearable activity monitors or activity trackers. The remaining 2 studies (2/7, 28.6%) [44,56] showed no clear direction of effect for this outcome.

Impact of mHealth Interventions Primarily Targeting Physical Activity on Work Productivity Outcomes

Results summarizing the impact of mHealth interventions on work-related outcomes and health-related outcomes by targeted intervention types are shown in Table S2 in Multimedia Appendix 3. Out of the 4 studies [41,44,54,57] primarily aimed at increasing physical activity, 3 (75%) [41,54,57] reported favorable effect directions with work productivity outcomes in the intervention group compared with the control or comparison groups. Of the 3 studies that showed favorable effect directions [41,54,57], one [41] was an RCT and the remaining 2 were NRSIs [54,57]. In detail, a Japanese study with an RCT design [41] explored the effects of a 12-week patient education and strengthening exercise program delivered using a mobile messaging app on work productivity among patients receiving pharmacological treatment for chronic low back pain. The conventional group received only routine medical care. The between-group differences for work productivity outcomes favored the intervention group but did not reach statistical significance. Regarding the NRSI studies, the retrospective cohort study [54] among workers with either knee or low back pain in the United States evaluated the efficacy of a remote digital care program on self-reported work productivity as assessed by the Work Productivity and Activity Impairment (WPAI) questionnaire. The digital care program was delivered through a mobile app that incorporated education, sensor-guided exercise therapy, and behavioral health support given 1-on-1 by a remote health coach. The results demonstrated statistically significant improvements in WPAI-assessed work productivity from baseline to the 12-week follow-up. Additionally, in another NRSI study using a prospective cohort design [57], a digital pelvic program was conducted for postmenopausal women with pelvic floor disorders. This program combined pelvic floor muscle training (PFMT) and education, using a combination of mobile apps and wearable trackers. During the PFMT sessions, pelvic floor muscle activity was assessed with a tracker, and real-time feedback was provided through the mobile app. Work productivity impairment was reduced by around half at the intervention end (mean difference −15.08, 95% CI −17.52 to −12.64; $P<.001$). In contrast, an RCT with employees at a Finnish insurance company [44] showed no clear between-group difference in work productivity, with adjusted mean differences in the quantity and quality of work index of 1.3 (−2.0 to 4.7) and −1.1 (−4.9 to 2.8) at 6 and 12 months, respectively. This study used web-based components [44] as an mHealth technology in the intervention. Participants in the intervention

XSL•FO
RenderX

group monitored their daily physical activity, set goals, accessed a web-based tracking service, and received counseling via telephone or web messages over a 12-month period. Meanwhile, those in the control group were informed of the results of a fitness test and an information leaflet on physical activity with continued occupational health care as usual.

Impact of mHealth Interventions Primarily Targeting Sedentary Behavior on Work Productivity Outcomes

No included studies examined the impact of mHealth interventions mainly targeting sedentary behavior on work productivity outcomes.

Impact of mHealth Interventions Primarily Targeting Both Physical Activity and Sedentary Behavior on Work Productivity Outcomes

Out of the 3 studies [49,51,56] targeted both physical activity and sedentary behavior, 2 (66.7%) [49,51] reported favorable effect directions with work productivity outcomes in the intervention group compared with the control group. A cluster RCT among employees at 6 Spanish university campuses [51] examined a 19-week workplace web-based intervention that aimed to encourage office employees to increasingly "sit less and move more" during workdays. In this study, loss of performance, as assessed by the Work Limitation Questionnaire (WLQ), showed a statistically significant increase in both the intervention group (workplace web-based intervention) and active control group (pedometer, paper diary, and self-reported sitting time) across all study time points. However, after baseline, work performance was consistently better in the intervention group compared with the active comparison group. Likewise, in another cluster RCT [49] among office workers in Thailand, the intervention showed a slightly favorable effect direction for work productivity, as assessed by the WPAI. The adjusted between-group difference for percentage reduced work productivity was $\beta=-.37$ (95% CI −9.20 to 8.47), although the difference was not statistically significant. This study investigated the impact of a 6-month multicomponent short-break intervention aimed at decreasing sitting time and increasing physical activity, which included real-time feedback on activities that could be viewed via a smartwatch or using the Fitbit smartphone app (Fitbit Inc). Additionally, both this intervention group and an active control group received an information booklet on the benefits and consequences of physical activity and 7 simple exercises for promoting an active lifestyle in the office environment. In contrast, a single-group, pre-post mixed methods pilot study among police officers and staff in the United Kingdom [56] evaluated a 12-week mHealth intervention (Fitbit activity monitor and the "Bupa Boost" smartphone app; developed by Bupa). The study reported no statistically significant changes in perceived productivity, as assessed by the World Health Organization Health and Work Performance Questionnaire (WHO-HPQ), from baseline to week 6 (midintervention) or from baseline to week 12 (postintervention). Since the effect estimates of these changes were not provided, the effect direction was considered unclear.

Absenteeism Outcomes

Out of the 5 included studies [44-46,52,55] that evaluated absenteeism outcomes, 3 (60%)—2 RCTs [45,46] and a retrospective cohort study [55] targeting physical activity—demonstrated favorable effect directions with absenteeism outcomes in the intervention group compared with the control or comparison groups. Of the 3 studies that showed favorable effect directions [45,46,55], 2 used mobile apps as the mHealth tools for the intervention [45,46], whereas one used mobile apps and wearable activity trackers as mHealth technologies as part of the intervention [55]. The remaining 2 studies (2/5, 40%) [44,52] showed no clear direction of effect for this outcome.

Impact of mHealth Interventions Primarily Targeting Physical Activity on Absenteeism Outcomes

Out of the 4 studies [44-46,55] primarily targeting physical activity, 3 (75%) [45,46,55] reported favorable effect directions for absenteeism outcomes in the intervention group compared with the control or comparison groups. In one RCT [45] performed among health care workers in Canada, participants assigned to the exercise condition were instructed to install the Down Dog suite of apps (Yoga Buddhi Co), which includes body weight interval training, yoga, running, and barre. They were asked to undertake four 20-minute sessions each week (totaling 80 min/wk) over a 12-week period, either at home or at any convenient location. In contrast, participants in the waitlist control condition were instructed to continue their current level of physical activity. The results showed that participants in the exercise intervention group had fewer weeks with reported sick days than those in the control condition. In another RCT [46] conducted among workers with bone and soft-tissue injuries affecting the wrist, hand, or fingers in Spain, participants in the intervention group received a feedback-guided home exercise program delivered via a tablet-based app. Participants receiving the intervention experienced a hastened return to work compared with those in the control group who received the conventional paper-based program (mean between-group difference –18 d, 95% CI –33 to –3). Further, a retrospective cohort study [55] provided prediabetic adults with a novel personalized digital diabetes prevention program through a smartphone app, which incorporated remote monitoring, interactive mobile computing, behavior tracking tools (Fitbit Flex2 [Fitbit Inc]), an evidence-based curriculum, online peer support, and health coaching to prevent or delay type 2 diabetes onset. Results showed that work absenteeism decreased by nearly half a day per month, with a statistically significant difference in WHO-HPQ–measured absenteeism scores between baseline (mean 0.9, SD 1.2 d) and 4 months (mean 0.5, SD 1.1 d; $P<.001$). In contrast, one RCT [44] described in detail in the section on work productivity did not report a clear difference between the web-based intervention and control groups in accumulated sickness absence days during 12 months (adjusted mean difference 0.0 d; 95% CI −1.2 to 0.9).

Impact of mHealth Interventions Primarily Targeting Sedentary Behavior on Absenteeism Outcomes

In a cluster RCT performed among desk-based workers [52], participants in the intervention group were provided with a

XSL•FO
RenderX

height-adjustable workstation, attended a short seminar and supporting leaflet, received workstation instructions as well as sitting and standing targets, posters, a booklet on action planning and goal-setting, feedback on physical and sitting activity, a self-monitoring prompting tool (the DARMA cushion, which can sync data with a mobile app using Bluetooth), and coaching sessions (month 1 and quarterly), while participants in the control group received the results of health measures (eg, weight and blood pressure) and continued with their usual practice for the 12-month study period. Based on sickness absence data obtained from self-reports and organizational records, no clear difference was observed between the intervention and passive control (usual practice) groups. For self-reported data, the adjusted mean differences were 0.02 (95% CI −0.01 to 0.05) at 3 months, −0.02 (95% CI −0.05 to 0.02) at 6 months, and −0.02 (95% CI −0.10 to 0.06) at 12 months. Organizational records also showed adjusted mean differences of 1.32 days missed (95% CI −9.99 to 12.63) and 0.24 episodes (95% CI −0.35 to 0.82) from pre-intervention to 12 months.

Impact of mHealth Interventions Primarily Targeting Both Physical Activity and Sedentary Behavior on Absenteeism Outcomes

No included studies examined the impact of mHealth interventions mainly targeting both physical activity and sedentary behavior on absenteeism outcomes.

Work Performance Outcomes

Out of the 4 studies [43,50,52,53], 2 studies (50%)—an RCT [43] and a cluster RCT [52] targeting sedentary behavior—reported favorable effect directions on work performance outcomes in the intervention group compared with the control group. These studies used either mobile apps [52] or web-based components [43] as mHealth technologies in their interventions. However, one retrospective cohort study (25%) [53] that used a mobile app showed a slightly unfavorable effect direction for work performance outcomes. The remaining study (1/4, 25%) [50] showed no clear direction of effect for this outcome.

Impact of mHealth Interventions Primarily Targeting Physical Activity on Work Performance Outcomes

A retrospective cohort study [53] conducted among full-time workers registered with an internet survey company in Japan investigated the association between Pokémon GO (Niantic) use and work performance measured using the WHO-HPQ. Pokémon GO is a smartphone game app that uses location-based data and augmented reality to allow players to find, catch, and train Pokémon in real-world settings, as well as collect in-game items. The app also features “Pokémon eggs,” which can only be hatched after walking a certain distance. As a result, it encourages players to move around. The study indicated an unfavorable effect direction, with job performance decreasing to a slightly greater extent among Pokémon GO players than nonplayers, although the difference was not statistically significant (group × time $P_{interaction}$=.38; Cohen d=−0.07; 95% CI −0.21 to 0.06).

Impact of mHealth Interventions Primarily Targeting Sedentary Behavior on Work Performance Outcomes

Out of the 3 studies targeting sedentary behavior [43,50,52], 2 studies (66.7%)—1 RCT [43] and 1 cluster RCT [52]—showed favorable effect directions on work performance in the intervention group compared with the control group. In detail, the cluster RCT [52], which is described in detail in the absenteeism section, found group differences in favor of the intervention group compared with control for work performance at 3 months (β coefficient=0.24, 95% CI −0.12 to 0.61), 6 months (β coefficient=0.41, 95% CI 0.05-0.77), and 12 months (β coefficient=0.53, 95% CI 0.20-0.86). In addition, a secondary analysis of an RCT among office workers in Canada [43] examined the effect of a 6-week theory-based planning and web-based text message intervention targeting workplace sitting duration. The intervention group reported greater improvements in work performance compared to the control group, although the interaction effect was not statistically significant ($F_{1,}$ =1.67; P=.20, ηp^2=.03). In contrast, in a cluster RCT among desk-based office workers in Australia [50], the effect direction on work performance assessed using a 9-item self-rated scale was considered unclear because no group comparisons were conducted between the organizational support alone group (Group ORG) and the organizational support combined with the LUMOback activity tracker group (Group ORG + Tracker). Neither group showed significant changes in job performance at 3 or 12 months compared with baseline.

Impact of mHealth Interventions Primarily Targeting Both Physical Activity and Sedentary Behavior on Work Performance Outcomes

No included studies examined the impact of mHealth interventions mainly targeting both physical activity and sedentary behavior on work performance outcomes.

Workability Outcomes

Out of the 3 RCTs targeting physical activity [42,47,48], 2 studies (66.7%) [42,48] reported favorable effect directions on workability outcomes in the intervention group compared with the control group. These 2 studies used mobile apps and wearable activity monitors as mHealth technologies as part of the intervention [42,48]. The remaining study (1/3, 33.3%) [47] showed no clear direction of effect for this outcome.

Impact of mHealth Interventions Primarily Targeting Physical Activity on Workability Outcomes

Of the 3 studies [42,47,48] targeting physical activity on workability outcomes, 2 [42,48] reported favorable effect directions in the intervention group compared with the control group. In one RCT [42], participants in the exercise group were provided with individual recommendations for exercise at in-person meetings and using a smartphone app, Rebirth Active (d.velop AG; with the Garmin activity monitor), with the aim of doing 150 minutes of physical activity per week. This study observed that workability improved at 6 months of the exercise intervention compared to the waitlist control group. Total Work Ability Index score increased in the exercise group but not in the control group (waitlist), with a statistically significant difference of 1.50 points (95% CI 0.75-2.25; P<.001) between

XSL•FO
RenderX

groups. In another RCT [48] performed among military employees in Finland, participants in the intervention group were provided with an interactive accelerometer (ExSed Movesense and Suunto) linked to a mobile app (ExSed [UKK, Terveyspalvelut Oy]) and a cloud service, along with telephone counseling and encouragement to exercise. The control group continued their usual physical activity without the accelerometer or feedback. Workability was measured on a 0 - 10 scale comparing current workability with the highest level during one's career, with higher scores indicating better workability. The study reported that workability scores decreased in both the intervention and control groups during the 1-year follow-up, but the decline was smaller in the intervention group, with no statistically significant group × time interaction (t=1.020; P=.31). In contrast, in the remaining RCT study [47] conducted among public safety workers with chronic low back pain, no clear effect directions on workability were observed because between-group comparisons were not performed.

Impact of mHealth Interventions Primarily Targeting Sedentary Behavior on Workability Outcomes

No included studies examined the impact of mHealth interventions mainly targeting sedentary behavior on workability outcomes.

Impact of mHealth Interventions Primarily Targeting Both Physical Activity and Sedentary Behavior on Workability Outcomes

No included studies examined the impact of mHealth interventions mainly targeting physical activity and sedentary behavior on workability outcomes.

Presenteeism Outcomes

Both 2 included cluster RCTs (2/2, 100%) [51,52] demonstrated favorable effect directions on presenteeism outcomes in the intervention group compared with the control group. Among these, one study [52] targeted sedentary behavior, while the other [51] targeted both physical activity and sedentary behavior. Regarding the mHealth components, one [52] used mobile apps as part of the intervention, whereas the other [51] conducted a web-based mHealth intervention.

Impact of mHealth Interventions Primarily Targeting Physical Activity on Presenteeism Outcomes

No included studies examined the impact of mHealth interventions mainly targeting physical activity on presenteeism outcomes.

Impact of mHealth Interventions Primarily Targeting Sedentary Behavior on Presenteeism Outcomes

The cluster RCT [52], described in detail in the absenteeism section, also assessed presenteeism using the WLQ questionnaire. A favorable effect direction for overall work-related sickness presenteeism (higher scores indicate less sickness presenteeism) was observed (1/1, 100%), with group differences favoring the intervention group at 3 months (β coefficient=0.25, 95% CI 0.01-0.49; P=.04), 6 months (β coefficient=0.10, 95% CI −0.15 to 0.36; P=.42), and 12 months (β coefficient=0.25, 95% CI −0.08 to 0.58; P=.14) [52]. This study used mobile apps as mHealth technologies as part of the intervention.

Impact of mHealth Interventions Primarily Targeting Both Physical Activity and Sedentary Behavior on Presenteeism Outcomes

The cluster RCT [51], described in detail in the work productivity section, assessed presenteeism and found a favorable effect on presenteeism in the web-based intervention group compared with the active control group (1/1, 100%). This study assessed presenteeism using 3 domains for presenteeism in the WLQ, namely difficulty in meeting scheduling demands (time), performance of cognitive and interpersonal tasks (mental-interpersonal), and decreases in fulfilling the expected quantity, quality, and timeliness of completed work (output). Results showed a statistically significant 2 (group)×2 (program time points) interaction for the time ($P_{interaction}$=.005), mental-interpersonal ($P_{interaction}$=.019), and output scales for presenteeism ($P_{interaction}$=.036). In both the intervention group and active comparison group, presenteeism was statistically significantly increased across the program, showing a universal increase in the difficulty of meeting scheduling demands (time scale), doing cognitive tasks and interacting with others (mental-interpersonal scale), and in meeting requirements for the quantity and quality of completed work (output scale). However, the intervention group consistently showed lower presenteeism impairment compared to the active comparison group across time points.

Health-Related Outcomes

Ten included studies reported results related to physical activity [42-44,48-52,55,56] (Table S1 in Multimedia Appendix 3). Among these, 8 studies (80%) [42,43,49-52,55,56] demonstrated favorable effect directions on physical activity–related outcomes, whereas one study (10%) [48] showed no clear effect direction and another (10%) [44] showed an unfavorable effect direction. The majority of studies with favorable effect directions on physical activity–related outcomes (5/8, 62.5%) [42,49,50,55,56] used mHealth interventions combining mobile apps with wearable activity monitors or trackers. Of the remaining 3 studies, 1 [52] used mobile apps and 2 [43,51] used web-based components as mHealth tools for the intervention. Regarding sedentary behavior, 7 included studies reported relevant outcomes [43,48-52,56]. Of these, 6 studies (85.7%) [43,48,49,51,52,56] demonstrated favorable effect directions for sedentary behavior outcomes, whereas one study (14.3%) [50] showed no clear effect. Of these 6 studies [43,48,49,51,52,56], 1 [52] used mobile apps, 2 [43,51] used web-based components, and 3 [48,49,56] used combinations of mobile apps and wearable activity monitors or trackers as mHealth tools in the intervention. Of the 9 studies showing favorable effect directions on physical activity and/or sedentary behavior–related outcomes [42,43,48-52,55,56], all except one [51] implemented blended intervention strategies that included intervention components beyond mHealth technologies or tools.

RoB Assessment

Our RoB assessment for the included studies is shown in Figures 2-4. RoB was evaluated for each work-related outcome. In cases

XSL•FO
RenderX

where a single study reported multiple work-related outcomes (eg, work absence evaluated using both organizational records and a self-reported questionnaire), the RoB for each outcome is presented individually. In the 12 RCT studies [41-52], 18 work-related outcomes were identified (work productivity [n=5], absenteeism [n=5], work performance [n=3], workability [n=3], and presenteeism [n=2]). Regarding the overall RoB in the included RCT studies (Figures 2 and 3), 15 of 18 outcomes were rated as having a high RoB [41-45,47-52] and 3 as raising some concerns [44,46,52]. Of the 18 work-related outcomes, 14 were evaluated as either secondary [41-43,45,47,48,50-52] or tertiary outcomes [49], and only 4 were assessed as primary outcomes [41,44,46]. Consequently, the main domains of high risk of bias in many RCT studies were missing outcome data, outcome measurement, and selection of the reported result. In addition, given the nature of mHealth interventions, the included studies were necessarily conducted as open-label trials, raising concerns about the RoB due to deviations from the intended intervention. Domain 2 of the RoB 2, which assesses bias due to deviations from the intended intervention, was assessed as high risk or of some concerns in all RCT studies [41-52]. The overall risk of bias assessment for the 5 studies with an NRSI design [53-57] is summarized in Figure 4. In these 5 studies [53-57], a total of 5 work-related outcomes were identified (ie, work productivity [n=3], absenteeism [n=1], and work performance [n=1]). Regarding the overall risk of bias in these 5 studies [53-57], 3 outcomes were rated as at critical risk of bias [54-56] and 2 as at serious risk of bias [53,57]. The main domain of critical risk of bias was bias due to confounding.

Figure 2. Risk of bias assessment for randomized controlled trials at the outcome level [41-48]. QQ method: quantity and quality method, WAI: Work Ability Index, WPAI: Work Productivity and Activity Impairment.

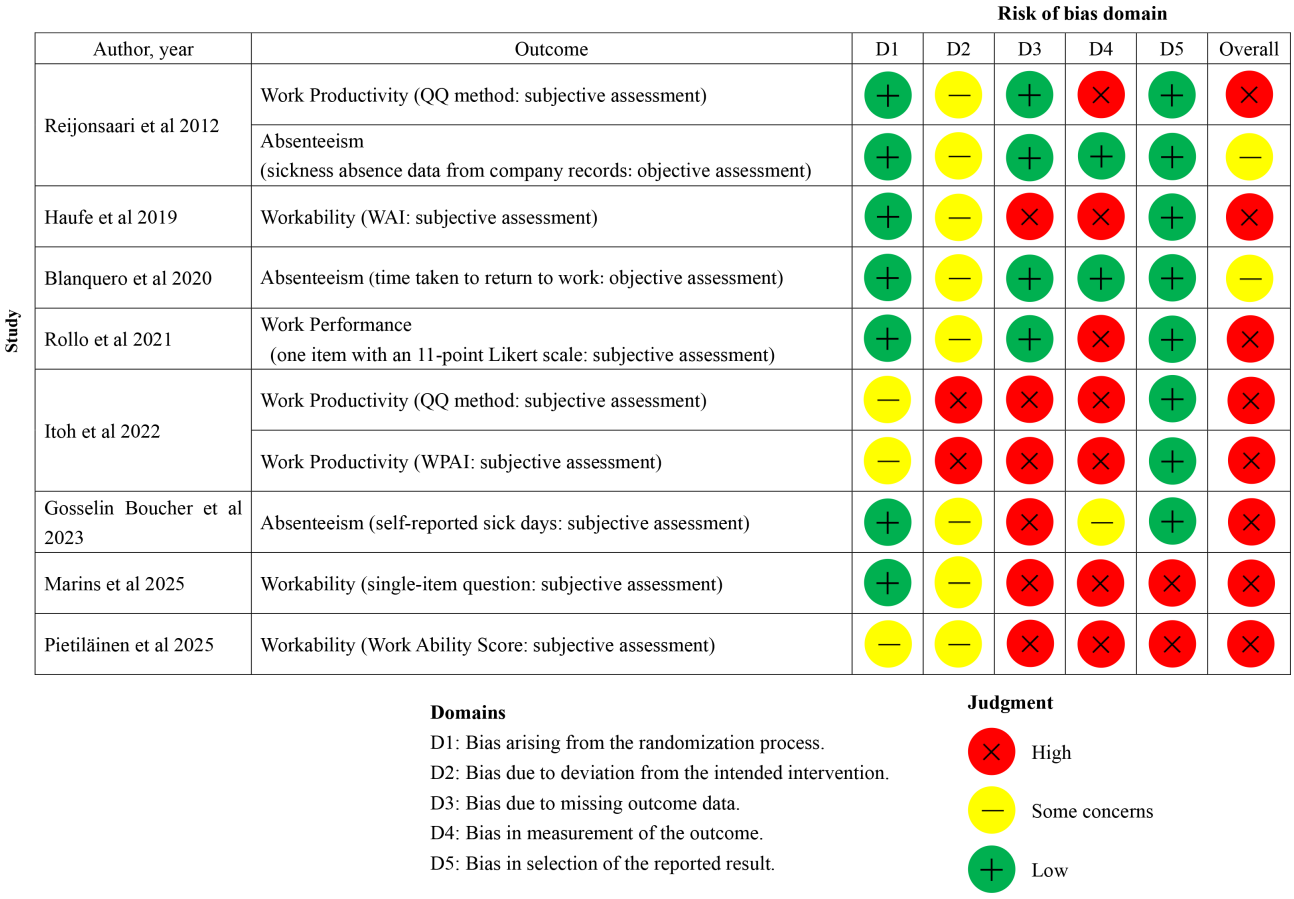

XSL•FO
RenderX

Figure 3. Risk of bias assessment for cluster randomized controlled trials at the outcome level [49-52]. WLQ: Work Limitations Questionnaire, WPAI: Work Productivity and Activity Impairment.

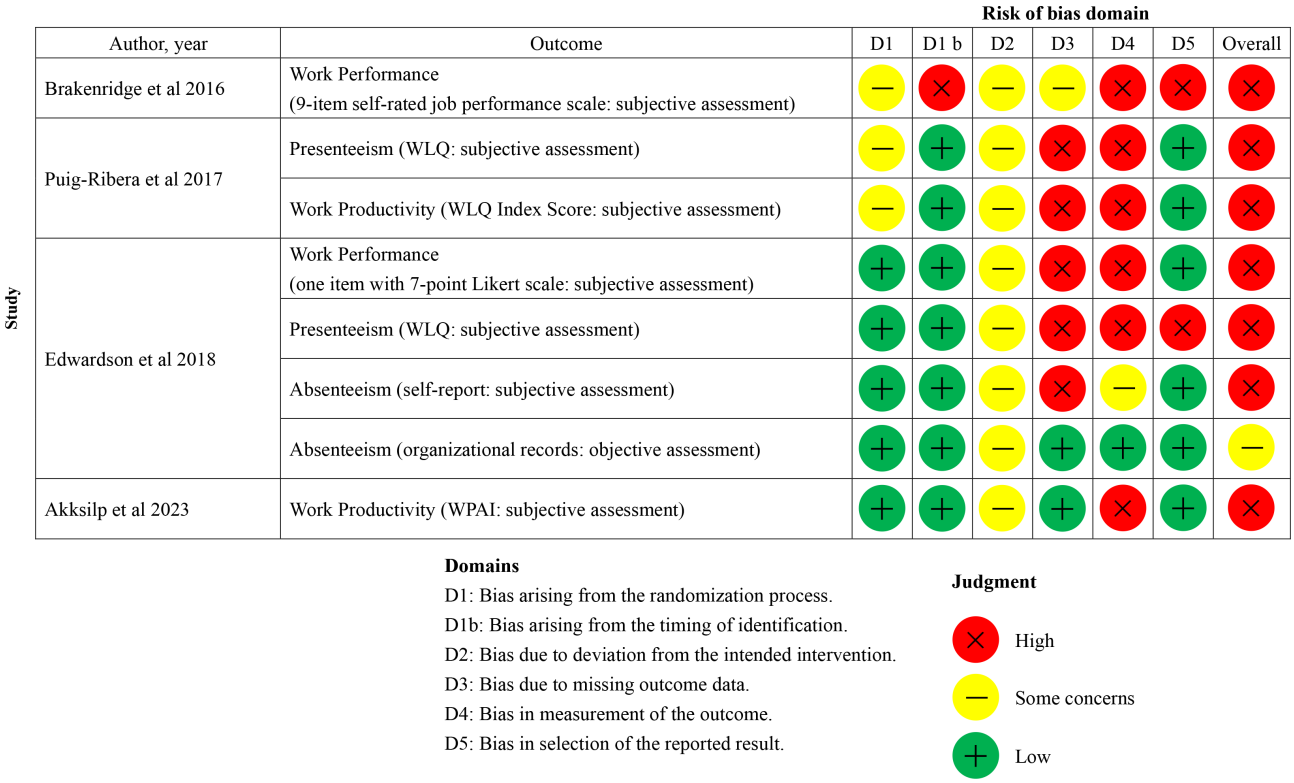

Figure 4. Risk of bias assessment for nonrandomized studies of interventions at the outcome level [53-57]. WHO-HPQ: World Health Organization Health and Work Performance Questionnaire; WPAI: Work Productivity and Activity Impairment.

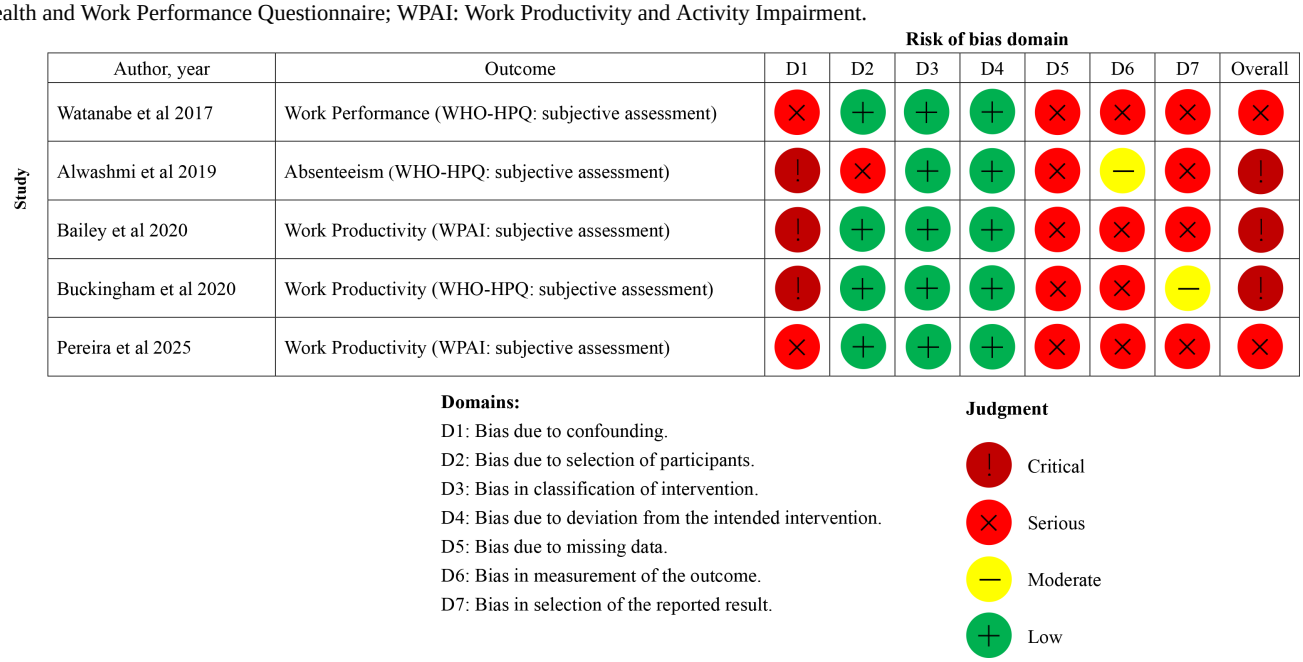

Adverse Events

Of the 17 included studies [41-57], adverse events were reported in 8 studies (47.1%; in one study [49], adverse event information was obtained from the corresponding author) [42,44,47,49,50,54,56,57], whereas no information on adverse events was provided in the remaining 9 studies (52.9%) [41,43,45,46,48,51-53,55]. The most frequent adverse events included skin reactions (including rashes, itching, and mild irritation) associated with the use of activity monitors and trackers, with the majority of adverse events being relatively minor. There were no reports of serious adverse events attributable to the intervention.

Discussion

Main Findings

This systematic review is, to our knowledge, the first to comprehensively examine the impact of mHealth interventions designed to encourage physical activity and decrease sedentary

XSL•FO
RenderX

behavior on a wide range of work-related outcomes, including absenteeism, presenteeism, work productivity, work performance, and workability in working populations. A total of 17 studies (12 RCTs and 5 NRSIs) [41-57] met the eligibility criteria. Of these, 12 studies (70.6%) [41-43,45,46,48,49,51,52,54,55,57] reported favorable effect directions on at least one of the target work-related outcomes relative to control or comparison groups. However, the majority exhibited a high or critical risk of bias, raising concerns about the quality of the research. The considerable heterogeneity and overall low methodological quality of the included studies suggest that confidence in the reported effects of mHealth interventions aimed at promoting physical activity and reducing sedentary behavior on work-related outcomes remains limited at this stage. Considering these issues, it was not possible to reach definitive conclusions about the effectiveness of mHealth interventions on work-related outcomes.

Summary of Findings on Work-Related Outcomes

Overall, 12 studies reported favorable effect directions for at least one targeted work-related outcome [41-43,45,46,48,49,51,52,54,55,57], 4 reported no clear effect direction [44,47,50,56], and 1 reported an unfavorable effect direction [53]. One study [53] observed that work performance decreased to a slightly greater extent among Pokémon GO players than among nonplayers, although the difference was not statistically significant. However, the majority of included studies indicated that mHealth intervention strategies targeting physical activity and/or sedentary behavior showed a favorable effect direction on work-related outcomes compared with control or comparison groups. Generally similar findings were observed across mHealth intervention types (ie, physical activity, sedentary behavior, or both). These findings suggest that health promotion efforts that incorporate mHealth components to encourage physical activity and/or decrease sedentary behavior are unlikely to have a substantial negative impact on work-related outcomes and have the potential to positively impact these outcomes. This finding is important as it provides useful evidence for stakeholders faced with decision-making about health promotion efforts incorporating mHealth components which encourage physical activity and decrease sedentary behavior among workers, in consideration of aspects of work-related outcomes.

Presenteeism, Work Productivity, Work Performance, and Workability

In addition to the above issues, the use of various questionnaires to assess work-related outcomes such as presenteeism, work productivity, work performance, and workability has made it difficult to compare results across studies [60-62], and thereby hindered the ability to draw clear conclusions. Most of the included studies did not use the same questionnaires, and in some cases, identical questionnaires were used to assess different outcomes. For example, the WHO-HPQ was implemented to measure work productivity, presenteeism, and work performance [19]. Work productivity was classified as the efficiency and effectiveness of employees in performing their work, in consideration of the time (presenteeism) or resources (performance) required to perform a specific task, and the problems that required resolution in achieving their objectives [63]. Accordingly, the work productivity outcomes included both performance and presenteeism variables as a result of the two variables expressing a similar concept, albeit with the use of different tools [63,64]. In the context of workability, three included studies [42,47,48] assessed this outcome, but each used a different measurement tool. The use of different tools to measure these work-related outcomes makes it difficult to compare study results and draw definitive conclusions about the impact of mHealth intervention strategies.

Absenteeism

Of the five included studies [44-46,52,55] that evaluated absenteeism outcomes, three studies targeting physical activity (two RCTs [45,46] and one retrospective cohort study [55]) demonstrated a favorable effect direction for absenteeism compared to the control or comparison groups. Absenteeism is defined as a period during which employees are absent from work due to illness or disability and is measured as the number of work hours or days missed [64]. Thus, this outcome can be objectively evaluated with materials such as workplace personnel records, making it generally less heterogeneous and more consistent in evaluation compared to other work-related outcomes. In fact, the risk of bias was less severe in those included studies that objectively evaluated absenteeism (eg, using organizational records) [44,46,52] compared to those that assessed it subjectively (eg, using self-reported measures) [45,52,55]. Among the three studies reporting a favorable direction of effect [45,46,55], only one [46] objectively assessed absenteeism, with the risk of bias assessed as being of some concerns. In contrast, the other two, which relied on subjective assessments, were judged to have a high [45] and a critical [55] risk of bias. In sum, mHealth interventions that promote physical activity may have the potential to decrease absenteeism; however, existing evidence on this issue is limited and raises concerns about bias risk. Therefore, further high-quality research is required before any definitive conclusions can be reached.

Comparison With Previous Evidence

Since this systematic review is the first to explore the impact of mHealth interventions targeting physical activity promotion and sedentary behavior reduction with work-related outcomes, our results cannot be directly compared with those of previous systematic reviews. Nevertheless, the results are supported by evidence from prior systematic reviews on related topics, which were not restricted to mHealth-based intervention strategies. Consistent with this review, a prior systematic review investigating the effect of workplace nutrition and physical activity interventions on work-related outcomes suggested their potential effectiveness, and 14 studies showed statistically significant favorable changes in absenteeism (n=7), work performance (n=2), workability (n=3), work productivity (n=1), and both workability as well as work productivity (n=1) [19]. Likewise, another systematic review confirmed that a number of studies have demonstrated the effectiveness of worksite programs based on physical activity in improving work-related outcomes, such as workability (n=5), absenteeism (n=3), and work productivity (n=5) [20]. The consistency between our findings and those of previous studies supports the reliability

XSL•FO
RenderX

of the present results and strengthens the notion that health promotion efforts incorporating mHealth technologies to promote physical activity and reduce sedentary behavior may positively impact work-related outcomes. Notably, earlier systematic reviews [19,20] primarily focused on traditional workplace-based interventions and did not consider mHealth technologies as key components of intervention. Implementation barriers commonly reported for workplace health promotion programs include financial constraints, limited personnel or expertise, and insufficient physical space [21]. mHealth technologies may help address these challenges by reducing reliance on on-site resources and enabling scalable, flexible, and context-independent intervention delivery beyond conventional workplace boundaries. The importance of scalability for broad implementation has been emphasized in health promotion research [14]. By synthesizing recent evidence, this review contributes to the existing literature by expanding the scope of intervention settings (eg, not only workplace settings but also home and outdoor contexts) and by emphasizing studies in which mHealth technologies constituted key components of the interventions, thereby enhancing their scalability and potential for broader implementation.

mHealth Strategies and Components

The included studies covered a wide range of mHealth strategies. The most frequently reported mHealth technology was mobile apps, used either alone or in combination with wearable activity monitors or trackers (14/17, 82.4%) [41,42,45-50,52-57]. Even when categorized under the same type of mHealth technology, the content and design of each mHealth intervention varied substantially. In addition, many intervention strategies included not only mHealth components but also other components, such as human support (eg, personal counseling; 13/17, 76.5%) [41-44,46,48-50,52,54-57], which further contributed to heterogeneity across studies. This makes it difficult to attribute the effects of the interventions solely to the mHealth technologies. It has been suggested that adherence to interventions is low when mHealth is used as a standalone methodology [30]. In contrast, personal communication with and integration of health care professionals has been reported to positively influence adherence to mHealth interventions focusing on physical activity [30]. In fact, in one included study using an mHealth-only intervention [45], app use declined over the 12-week intervention, with only 33 participants (33/142, 23%) continuing to use the apps regularly by the end of the trial. In the included studies, no information on adherence—defined as the extent to which users followed the program as designed—was reported for mHealth-only interventions, except for one study [45]. Consequently, adherence could not be compared between mHealth-only and blended interventions, and the details remain unclear. Nevertheless, the potentially low adherence to mHealth-only interventions may partly explain why many studies incorporated additional components rather than relying solely on mHealth technologies, which is consistent with previous literature [30]. Likewise, the NICE committee recommended considering digital and mHealth interventions as options to increase physical activity alongside existing services, while noting that their effectiveness is variable [32]. Taken together, these results indicate that mHealth interventions should be viewed as a complement to existing services, such as human support, rather than a replacement for them, both in future research and in the deployment of mHealth technologies in real-world settings.

Strengths

This systematic review has several strengths. First, it used a comprehensive search strategy across major electronic databases, facilitating an evidence-based approach to the literature search. Second, it included literature on studies using not only RCT but also NRSI designs. Incorporating non-RCT study designs provides broader evidence to guide the implementation of interventions in real-world settings, particularly in situations where conducting an RCT is difficult or inappropriate. Third, each study included in this review was selected and assessed in a comprehensive manner, and its data were extracted and assessed for quality by 2 researchers acting independently in order to remove potential biases in the review process. Additionally, we used 2 methodological quality assessment tools recommended by the Cochrane Collaboration to assess the risk of bias in the included RCTs and NRSIs.

Limitations

We also note a number of limitations of this review. First, a major limitation of this review is the substantial heterogeneity across study populations, interventions, and outcomes, which precluded meta-analysis and prevented a reliable assessment of the overall certainty of the evidence using the GRADE approach. Instead, this review relied on a narrative synthesis based on vote counting of effect directions across studies. While this approach can indicate effect direction, it cannot estimate the overall effect size across studies. This limitation may reduce the strength of the conclusions and impede the ability to identify which interventions are most effective for work-related outcomes, and the conditions under which they are most beneficial. Future studies using standardized outcome measures and comparable interventions would enable quantitative synthesis of the evidence through meta-analysis, and contribute to addressing this limitation. Second, the review was limited to papers written in English and Japanese, introducing a degree of language bias. However, we confirmed that there were no studies within the screened literature other than those in Japanese and English that met the inclusion criteria for this systematic review. Additionally, some studies that were potentially both useful and relevant may have been excluded because they were not full-text original articles, such as conference proceedings, and unpublished manuscripts. To minimize this possibility, we reviewed the content of non–full-text original articles identified during screening, such as conference proceedings, trial registry records, and protocol papers, to confirm whether any corresponding full-text original studies relevant to this review had been published. Moreover, only 2 of the included studies [41,47] used the CONSORT-EHEALTH checklist [65] or the reporting guidelines for mHealth interventions [66], which may hinder the successful implementation of effective interventions in real-world settings. Finally, many of the included studies in this review investigated relatively short-term effects, with the duration of the interventions ranging from 6 weeks to 12 months,

XSL•FO
RenderX

and the long-term impact on work-related outcomes therefore remains unclear.

Implications for Future Research

What future research is necessary to advance this field beyond the results of this systematic review? First, a key priority is to promote further high-quality research that ensures a low possibility of bias risk. Many of the included studies were evaluated as having a high or critical risk of bias, including not only NRSIs [54-56] but also RCTs [41-45,47-52]. Among the included studies, the majority of work-related outcomes were evaluated as either secondary [41-43,45,47,48,50-57] or tertiary outcomes [49], and only 4 outcomes were evaluated as primary outcomes [41,44,46]. This may have partly contributed to a high risk of bias across the domains of missing outcome data, outcome measurement, and selection of the reported results. A high proportion of missing data may reduce the reliability and generalizability of the findings. In addition, selective reporting of outcomes as well as outcome measurement based on self-reports could lead to an overestimation of intervention effects. Future research that assesses work-related outcomes as a primary outcome may help address these bias-related concerns. From other perspectives, efforts to increase adherence will likely help reduce the risk of bias due to deviation from the intended intervention. The degree to which the study intervention was implemented as intended is one of the most challenging issues in the field of mHealth interventions [31]. Addressing low adherence rates is essential, as low adherence can reduce statistical power and introduce bias, particularly by making a potentially efficacious intervention appear less effective, thereby leading to misleading recommendations in mHealth interventions [31]. While low adherence rates in mHealth interventions have recently been widely recognized, strategies that may address these difficulties are emerging [30,31]. Examples that have been shown to positively influence adherence to mHealth interventions focusing on physical activity include customizable push notifications, gamification, social features (eg, competition, social comparison, and challenges), app personalization and customization, personal communication with, and integration of, health care professionals [30]. Future studies incorporating this emerging evidence on adherence to mHealth tools may help establish effective intervention strategies that encourage physical activity and decrease sedentary behavior, thereby contributing to improvements in work-related outcomes. Finally, in addition to the challenges mentioned above, a major limitation of the current evidence is the substantial heterogeneity among the included studies, which prevents quantitative synthesis through meta-analysis. Addressing such heterogeneity might require innovative approaches such as mega-trials, where various interventions are tested simultaneously in identical populations with uniform measurement periods and consistent outcome assessments [67]. Alternatively, promoting future research that adheres to frameworks such as the CONSORT-EHEALTH checklist [65] and the reporting guidelines for mHealth interventions [66] may contribute to standardization across studies, and thereby help overcome heterogeneity among them. As an additional benefit, compliance with these checklists and guidelines could enhance research quality and, furthermore, ensure the reproducibility of the intervention, which is itself essential for the successful implementation of intervention strategies in real-world settings. It will also be valuable to observe which apps developed in research settings successfully transition to the market, while evaluating their impact and quality using tools such as the Mobile App Rating Scale (MARS) [68]. In one included study [56], a clear association was observed between engagement (defined by usage) and perceived usability and usefulness, suggesting that incorporating user feedback is essential for developing effective interventions in real-world settings.

Conclusions

In conclusion, this systematic review suggests that health promotion initiatives incorporating mHealth components to promote physical activity and reduce sedentary behavior may have a positive impact on work-related outcomes. However, the current evidence is limited and generally of low quality, due to a lack of high-quality studies and substantial heterogeneity in study populations, intervention content, outcome measures, and study designs. To draw more reliable conclusions, future research should include large-scale, high-quality studies with long-term follow-up and objective or validated outcome assessments. This review is the first to comprehensively synthesize evidence on work-related outcomes, extending prior reviews that primarily focused on traditional workplace-based interventions without mHealth components. The findings of this review may provide guidance for future studies, including the potential value of megastudy designs that evaluate multiple interventions within the same population over the same period on the same outcomes. In real-world settings, these findings suggest that mHealth technologies may serve as a complementary strategy for supporting occupational health and productivity.

Acknowledgments

The authors express their appreciation for the support and assistance of the professional medical librarians at Tokyo Women's Medical University, Tokyo, Japan, for their support in the design of the search strategy. Additionally, we would like to thank Ms. Shoko Okamoto of Tokyo Women's Medical University, Tokyo, Japan, for her support in conducting this research. The authors attest that no generative artificial intelligence was used in the writing of this manuscript.

Funding

This study was supported by AMED under grant number JP24rea522007. The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

XSL•FO
RenderX

Data Availability

The data used in this systematic review were extracted from published studies, and we do not possess specific datasets to share. All data generated during this study are included in this published article and its supplementary information files.

Authors' Contributions

TM, MN, and KN conceptualized, designed, and organized the study. TM, NH, DN, and SY conducted the title and abstract screening, full-text screening, data extraction, assessment of study quality, and critical appraisal. TM wrote the first draft of the manuscript. All authors reviewed drafts and contributed to the final version of the submitted manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search terms used in electronic database searches.
[DOCX File, 37 KB - jmir_v28i1e80540_app1.docx]

Multimedia Appendix 2
List of excluded studies along with reasons for exclusion.
[DOCX File, 97 KB - jmir_v28i1e80540_app2.docx]

Multimedia Appendix 3
Extracted data and supplementary tables S1 and S2.
[XLSX File, 45 KB - jmir_v28i1e80540_app3.xlsx]

Multimedia Appendix 4
Synthesis Without Meta-Analysis (SWiM) in systematic reviews: reporting guidelines checklist.
[DOCX File, 27 KB - jmir_v28i1e80540_app4.docx]

Checklist 1
PRISMA 2020 checklist.
[DOCX File, 37 KB - jmir_v28i1e80540_app5.docx]

Checklist 2
PRISMA-S checklist for reporting literature searches.
[DOCX File, 20 KB - jmir_v28i1e80540_app6.docx]

References

1. Thivel D, Tremblay A, Genin PM, Panahi S, Rivière D, Duclos M. Physical activity, inactivity, and sedentary behaviors: definitions and implications in occupational health. Front Public Health 2018;6(288):288. [doi: 10.3389/fpubh.2018.00288] [Medline: 30345266]
2. Salvo D, Crochemore-Silva I, Wendt A, et al. Physical activity for public health in the 21st century. Nat Med 2026 Mar 9. [doi: 10.1038/s41591-026-04237-5] [Medline: 41803342]
3. Tremblay MS, Aubert S, Barnes JD, et al. Sedentary Behavior Research Network (SBRN) - Terminology Consensus Project process and outcome. Int J Behav Nutr Phys Act 2017 Jun 10;14(1):75. [doi: 10.1186/s12966-017-0525-8] [Medline: 28599680]
4. Borodulin K, Anderssen S. Physical activity: associations with health and summary of guidelines. Food Nutr Res 2023;67. [doi: 10.29219/fnr.v67.9719] [Medline: 37441512]
5. Park JH, Moon JH, Kim HJ, Kong MH, Oh YH. Sedentary lifestyle: overview of updated evidence of potential health risks. Korean J Fam Med 2020 Nov;41(6):365-373. [doi: 10.4082/kjfm.20.0165] [Medline: 33242381]
6. Park S, Lee S, Woo S, et al. Sedentary behaviors and physical activity of the working population measured by accelerometry: a systematic review and meta-analysis. BMC Public Health 2024 Aug 6;24(1):2123. [doi: 10.1186/s12889-024-19449-y] [Medline: 39107699]
7. Church TS, Thomas DM, Tudor-Locke C, et al. Trends over 5 decades in U.S. occupation-related physical activity and their associations with obesity. PLoS One 2011;6(5):e19657. [doi: 10.1371/journal.pone.0019657] [Medline: 21647427]

XSL•FO
RenderX

8. Schöne C, Sauter M, Backé EM, Prigge M, Brendler C, Hegewald J. The impact of working from home on sedentary behaviour and physical activity compared to onsite work in the working population: a systematic review and meta-analysis. BMC Public Health 2025 Nov 17;25(1):3963. [doi: 10.1186/s12889-025-24960-x] [Medline: 41249967]
9. WHO Global Observatory for eHealth. MHealth: new horizons for health through mobile technologies: second global survey on ehealth. World Health Organization. 2011. URL: https://iris.who.int/handle/10665/44607 [accessed 2026-03-23]
10. Singh B, Ahmed M, Staiano AE, et al. A systematic umbrella review and meta-meta-analysis of eHealth and mHealth interventions for improving lifestyle behaviours. NPJ Digit Med 2024 Jul 5;7(1):179. [doi: 10.1038/s41746-024-01172-y] [Medline: 38969775]
11. Iwakura M, Ozeki C, Jung S, et al. An umbrella review of efficacy of digital health interventions for workers. NPJ Digit Med 2025 Apr 14;8(1):207. [doi: 10.1038/s41746-025-01578-2] [Medline: 40229460]
12. Mair JL, Salamanca-Sanabria A, Augsburger M, et al. Effective behavior change techniques in digital health interventions for the prevention or management of noncommunicable diseases: an umbrella review. Ann Behav Med 2023 Sep 13;57(10):817-835. [doi: 10.1093/abm/kaad041] [Medline: 37625030]
13. Mobile health: technology and outcomes in low and middle income countries (mhealth). National Institutes of Health. 2025. URL: https://fic.nih.gov/Programs/Pages/mHealth.aspx [accessed 2026-03-23]
14. Milat AJ, King L, Bauman AE, Redman S. The concept of scalability: increasing the scale and potential adoption of health promotion interventions into policy and practice. Health Promot Int 2013 Sep;28(3):285-298. [doi: 10.1093/heapro/dar097] [Medline: 22241853]
15. Dzakpasu FQS, Carver A, Brakenridge CJ, et al. Musculoskeletal pain and sedentary behaviour in occupational and non-occupational settings: a systematic review with meta-analysis. Int J Behav Nutr Phys Act 2021 Dec 13;18(1):159. [doi: 10.1186/s12966-021-01191-y] [Medline: 34895248]
16. Hu D, Xie P, Xia L, et al. The associations between sedentary behavior and risk of depression: a systematic review and dose response meta-analysis. BMC Public Health 2025 Jul 2;25(1):2254. [doi: 10.1186/s12889-025-23418-4] [Medline: 40604616]
17. Pearce M, Garcia L, Abbas A, et al. Association between physical activity and risk of depression: a systematic review and meta-analysis. JAMA Psychiatry 2022 Jun 1;79(6):550-559. [doi: 10.1001/jamapsychiatry.2022.0609] [Medline: 35416941]
18. Rojanasarot S, Bhattacharyya SK, Edwards N. Productivity loss and productivity loss costs to United States employers due to priority conditions: a systematic review. J Med Econ 2023;26(1):262-270. [doi: 10.1080/13696998.2023.2172282] [Medline: 36695516]
19. Grimani A, Aboagye E, Kwak L. The effectiveness of workplace nutrition and physical activity interventions in improving productivity, work performance and workability: a systematic review. BMC Public Health 2019 Dec 12;19(1):1676. [doi: 10.1186/s12889-019-8033-1] [Medline: 31830955]
20. Marin-Farrona M, Wipfli B, Thosar SS, et al. Effectiveness of worksite wellness programs based on physical activity to improve workers' health and productivity: a systematic review. Syst Rev 2023 May 24;12(1):87. [doi: 10.1186/s13643-023-02258-6] [Medline: 37226273]
21. Weinstein M, Cheddie K. Adoption and implementation barriers for worksite health programs in the United States. Int J Environ Res Public Health 2021 Nov 16;18(22):12030. [doi: 10.3390/ijerph182212030] [Medline: 34831782]
22. Stephenson A, McDonough SM, Murphy MH, Nugent CD, Wilson IM, Mair JL. Exploring the views of desk-based office workers and their employers' beliefs regarding strategies to reduce occupational sitting time, with an emphasis on technology-supported approaches. J Occup Environ Med 2020 Feb;62(2):149-155. [doi: 10.1097/JOM.0000000000001777] [Medline: 31934910]
23. Miki T, Nohara M, Nomura K. Effectiveness of mHealth interventions to promote physical activity and reduce sedentary behaviours on work-related productivity and performance: a systematic review protocol. BMJ Open 2024 Mar 4;14(3):e080240. [doi: 10.1136/bmjopen-2023-080240] [Medline: 38443086]
24. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
25. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
26. Karim JL, Talhouk A. Person-generated health data in women's health: protocol for a scoping review. JMIR Res Protoc 2021 May 28;10(5):e26110. [doi: 10.2196/26110] [Medline: 34047708]
27. Buckingham SA, Williams AJ, Morrissey K, Price L, Harrison J. Mobile health interventions to promote physical activity and reduce sedentary behaviour in the workplace: a systematic review. Digit Health 2019;5:2055207619839883. [doi: 10.1177/2055207619839883] [Medline: 30944728]
28. Ganesan AN, Louise J, Horsfall M, et al. International mobile-health intervention on physical activity, sitting, and weight: the Stepathlon Cardiovascular Health Study. J Am Coll Cardiol 2016 May 31;67(21):2453-2463. [doi: 10.1016/j.jacc.2016.03.472] [Medline: 27050185]

XSL•FO
RenderX

29. Schoeppe S, Alley S, Van Lippevelde W, et al. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review. Int J Behav Nutr Phys Act 2016 Dec 7;13(1):127. [doi: 10.1186/s12966-016-0454-y] [Medline: 27927218]
30. Jakob R, Harperink S, Rudolf AM, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371. [doi: 10.2196/35371] [Medline: 35612886]
31. Lopez-Alcalde J, Wieland LS, Yan Y, et al. Methodological challenges in randomized controlled trials of mHealth interventions: cross-sectional survey study and consensus-based recommendations. J Med Internet Res 2024 Dec 19;26:e53187. [doi: 10.2196/53187] [Medline: 39700488]
32. Behaviour change: digital and mobile health interventions. National Institute for Health and Care Excellence. 2020. URL: https://www.nice.org.uk/guidance/ng183/chapter/Recommendations#diet-and-physical-activity [accessed 2026-03-23]
33. Saldanha IJ, Skelly AC, Ley KV, et al. Inclusion of Nonrandomized Studies of Interventions in Systematic Reviews of Intervention Effectiveness: An Update: Rockville (MD): Agency for Healthcare Research and Quality (US); Report No.: 22-EHC033; 2022. [doi: 10.23970/AHRQEPCMETHODSGUIDENRSI] [Medline: 36153934]
34. Schelvis RMC, Oude Hengel KM, Burdorf A, Blatter BM, Strijk JE, van der Beek AJ. Evaluation of occupational health interventions using a randomized controlled trial: challenges and alternative research designs. Scand J Work Environ Health 2015 Sep 1;41(5):491-503. [doi: 10.5271/sjweh.3505] [Medline: 26030719]
35. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
36. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016 Oct 12;355:i4919. [doi: 10.1136/bmj.i4919] [Medline: 27733354]
37. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ 2020 Jan 16;368:l6890. [doi: 10.1136/bmj.l6890] [Medline: 31948937]
38. Boon MH, Thomson H. The effect direction plot revisited: application of the 2019 Cochrane Handbook guidance on alternative synthesis methods. Res Synth Methods 2021 Jan;12(1):29-33. [doi: 10.1002/jrsm.1458] [Medline: 32979023]
39. Thomson HJ, Thomas S. The effect direction plot: visual display of non-standardised effects across multiple outcome domains. Res Synth Methods 2013 Mar;4(1):95-101. [doi: 10.1002/jrsm.1060] [Medline: 23795209]
40. Brennan SE, Mckenzie JE. Chapter 12: synthesizing and presenting findings using other methods. In: Higgins JPT, Thomas J, Chandler J, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 65: Cochrane; 2024. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-12 [accessed 2026-03-23]
41. Itoh N, Mishima H, Yoshida Y, Yoshida M, Oka H, Matsudaira K. Evaluation of the effect of patient education and strengthening exercise therapy using a mobile messaging app on work productivity in Japanese patients with chronic low back pain: open-label, randomized, parallel-group trial. JMIR Mhealth Uhealth 2022 May 16;10(5):e35867. [doi: 10.2196/35867] [Medline: 35576560]
42. Haufe S, Kerling A, Protte G, et al. Telemonitoring-supported exercise training, metabolic syndrome severity, and work ability in company employees: a randomised controlled trial. Lancet Public Health 2019 Jul;4(7):e343-e352. [doi: 10.1016/S2468-2667(19)30075-1] [Medline: 31204284]
43. Rollo S, Prapavessis H. A combined health action process approach and mHealth intervention to reduce workplace sitting time in office-working adults: a secondary analysis examining health-related quality of life and work performance outcomes. Psychol Health 2021 Oct;36(10):1200-1216. [doi: 10.1080/08870446.2020.1838522] [Medline: 33108910]
44. Reijonsaari K, Vehtari A, Kahilakoski OP, van Mechelen W, Aro T, Taimela S. The effectiveness of physical activity monitoring and distance counseling in an occupational setting - results from a randomized controlled trial (CoAct). BMC Public Health 2012 May 11;12:344. [doi: 10.1186/1471-2458-12-344] [Medline: 22578104]
45. Boucher VG, Haight BL, Hives BA, et al. Effects of 12 weeks of at-home, application-based exercise on health care workers' depressive symptoms, burnout, and absenteeism: a randomized clinical trial. JAMA Psychiatry 2023 Nov 1;80(11):1101-1109. [doi: 10.1001/jamapsychiatry.2023.2706] [Medline: 37556150]
46. Blanquero J, Cortés-Vega MD, Rodríguez-Sánchez-Laulhé P, et al. Feedback-guided exercises performed on a tablet touchscreen improve return to work, function, strength and healthcare usage more than an exercise program prescribed on paper for people with wrist, hand or finger injuries: a randomised trial. J Physiother 2020 Oct;66(4):236-242. [doi: 10.1016/j.jphys.2020.09.012] [Medline: 33069608]
47. Marins EF, Primo TT, Vasconcelos BB, et al. Effectiveness of an app-delivered, self-management exercise program in public safety workers with chronic low back pain: a randomized controlled trial. Braz J Phys Ther 2025;29(5):101232. [doi: 10.1016/j.bjpt.2025.101232] [Medline: 40651078]
48. Pietiläinen E, Kyröläinen H, Parkkola K, Luukkaala T, Vasankari T. A randomized controlled trial utilizing an interactive accelerometer linked to a smartphone application for enhancing physical activity and health among military employees. Front Sports Act Living 2025;7:1549980. [doi: 10.3389/fspor.2025.1549980] [Medline: 40917622]
49. Akksilp K, Koh JJE, Tan V, et al. The physical activity at work (PAW) study: a cluster randomised trial of a multicomponent short-break intervention to reduce sitting time and increase physical activity among office workers in Thailand. Lancet Reg Health Southeast Asia 2023 Jan;8:100086. [doi: 10.1016/j.lansea.2022.100086] [Medline: 37384135]

XSL•FO
RenderX

50. Brakenridge CL, Fjeldsoe BS, Young DC, et al. Evaluating the effectiveness of organisational-level strategies with or without an activity tracker to reduce office workers' sitting time: a cluster-randomised trial. Int J Behav Nutr Phys Act 2016 Nov 4;13(1):115. [doi: 10.1186/s12966-016-0441-3] [Medline: 27814738]
51. Puig-Ribera A, Bort-Roig J, Giné-Garriga M, et al. Impact of a workplace "sit less, move more" program on efficiency-related outcomes of office employees. BMC Public Health 2017 May 16;17(1):455. [doi: 10.1186/s12889-017-4367-8] [Medline: 28511642]
52. Edwardson CL, Yates T, Biddle SJH, et al. Effectiveness of the Stand More AT (SMArT) Work intervention: cluster randomised controlled trial. BMJ 2018 Oct 10;363:k3870. [doi: 10.1136/bmj.k3870] [Medline: 30305278]
53. Watanabe K, Kawakami N, Imamura K, et al. Pokémon GO and psychological distress, physical complaints, and work performance among adult workers: a retrospective cohort study. Sci Rep 2017 Sep 7;7(1):10758. [doi: 10.1038/s41598-017-11176-2] [Medline: 28883633]
54. Bailey JF, Agarwal V, Zheng P, et al. Digital care for chronic musculoskeletal pain: 10,000 participant longitudinal cohort study. J Med Internet Res 2020 May 11;22(5):e18250. [doi: 10.2196/18250] [Medline: 32208358]
55. Alwashmi MF, Mugford G, Abu-Ashour W, Nuccio M. A digital diabetes prevention program (Transform) for adults with prediabetes: secondary analysis. JMIR Diabetes 2019 Jul 26;4(3):e13904. [doi: 10.2196/13904] [Medline: 31350833]
56. Buckingham SA, Morrissey K, Williams AJ, Price L, Harrison J. The Physical Activity Wearables in the Police Force (PAW-Force) study: acceptability and impact. BMC Public Health 2020 Nov 3;20(1):1645. [doi: 10.1186/s12889-020-09776-1] [Medline: 33143665]
57. Pereira AP, Janela D, Areias AC, et al. Innovating care for postmenopausal women using a digital approach for pelvic floor dysfunctions: prospective longitudinal cohort study. JMIR Mhealth Uhealth 2025 Apr 2;13:e68242. [doi: 10.2196/68242] [Medline: 40173388]
58. Akksilp K, Rouyard T, Isaranuwatchai W, et al. The Physical Activity at Work (PAW) program in Thai office workers: mixed methods process evaluation study. JMIR Form Res 2025 Jan 2;9:e57604. [doi: 10.2196/57604] [Medline: 39746195]
59. Biddle SJH, O'Connell SE, Davies MJ, et al. Reducing sitting at work: process evaluation of the SMArT Work (Stand More At Work) intervention. Trials 2020 May 13;21(1):403. [doi: 10.1186/s13063-020-04300-7] [Medline: 32404181]
60. Brown HE, Burton N, Gilson ND, Brown W. Measuring presenteeism: which questionnaire to use in physical activity research? J Phys Act Health 2014 Feb;11(2):241-248. [doi: 10.1123/jpah.2011-0307] [Medline: 23364284]
61. Zhang W, Tocher P, L'Heureux J, Sou J, Sun H. Measuring, analyzing, and presenting work productivity loss in randomized controlled trials: a scoping review. Value Health 2023 Jan;26(1):123-137. [doi: 10.1016/j.jval.2022.06.015] [Medline: 35961865]
62. Cadiz DM, Brady G, Rineer JR, Truxillo DM. A review and synthesis of the work ability literature. Work, Aging and Retirement 2019 Jan 18;5(1):114-138. [doi: 10.1093/workar/way010]
63. Stepanek M, Jahanshahi K, Millard F. Individual, workplace, and combined effects modeling of employee productivity loss. J Occup Environ Med 2019 Jun;61(6):469-478. [doi: 10.1097/JOM.0000000000001573] [Medline: 30985409]
64. Schultz AB, Chen CY, Edington DW. The cost and impact of health conditions on presenteeism to employers: a review of the literature. Pharmacoeconomics 2009;27(5):365-378. [doi: 10.2165/00019053-200927050-00002] [Medline: 19586075]
65. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
66. Agarwal S, LeFevre AE, Lee J, et al. Guidelines for reporting of health interventions using mobile phones: mobile health (mHealth) evidence reporting and assessment (mERA) checklist. BMJ 2016 Mar 17;352:i1174. [doi: 10.1136/bmj.i1174] [Medline: 26988021]
67. Milkman KL, Gromet D, Ho H, et al. Megastudies improve the impact of applied behavioural science. Nature New Biol 2021 Dec;600(7889):478-483. [doi: 10.1038/s41586-021-04128-4] [Medline: 34880497]
68. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 2015 Mar 11;3(1):e27. [doi: 10.2196/mhealth.3422] [Medline: 25760773]

Abbreviations

CENTRAL: Cochrane Central Register of Controlled Trials
GRADE: Grading of Recommendations Assessment, Development and Evaluation
ICTRP: International Clinical Trials Registry Platform
MARS: Mobile App Rating Scale
MeSH: Medical Subject Headings
mHealth: mobile health
NICE: National Institute for Health and Care Excellence
NRSI: nonrandomized study of interventions
PFMT: pelvic floor muscle training

XSL•FO
RenderX

PICO: Population, Intervention, Comparison, and Outcome
PRISMA: Preferred Reporting Items for Systematic reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
RCT: randomized controlled trial
RoB: risk of bias
ROBINS-I: Risk Of Bias In Nonrandomized Studies of Interventions
SWiM: Synthesis Without Meta-Analysis
WHO: World Health Organization
WHO-HPQ: World Health Organization Health and Work Performance Questionnaire
WLQ: Work Limitation Questionnaire
WPAI: Work Productivity and Activity Impairment

Edited by S Brini; submitted 12.Jul.2025; peer-reviewed by E Opoku-Agyemang, R Salari; revised version received 23.Mar.2026; accepted 24.Mar.2026; published 07.May.2026.

Please cite as:
Miki T, Nohara M, Homma N, Nagamine D, Yamaguchi S, Nomura K
Effectiveness of mHealth Interventions Aimed at Promoting Physical Activity and Reducing Sedentary Behavior on Work-Related Outcomes Among Workers: Systematic Review
J Med Internet Res 2026;28:e80540
URL: https://www.jmir.org/2026/1/e80540
doi:10.2196/80540

© Takako Miki, Michiko Nohara, Nami Homma, Daiki Nagamine, Shinji Yamaguchi, Kyoko Nomura. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

GenAI-Supported Virtual Patients in Health Care Education: Systematic Review

Juming Jiang, PhD; Megan Zichen Ye, MSc; Tyrone Tai-On Kwok, PhD; Janet Yuen Ha Wong, PhD

School of Nursing and Health Sciences, Jockey Club Institute of Healthcare, Hong Kong Metropolitan University, 11th Floor, 1 Sheung Shing Street, Homantin, Kowloon, Hong Kong, China (Hong Kong)

Corresponding Author:
Janet Yuen Ha Wong, PhD
School of Nursing and Health Sciences, Jockey Club Institute of Healthcare, Hong Kong Metropolitan University, 11th Floor, 1 Sheung Shing Street, Homantin, Kowloon, Hong Kong, China (Hong Kong)

Abstract

Background: Generative artificial intelligence (GenAI) is enhancing virtual patient simulations in health care education by enabling dynamic, adaptive interactions, reshaping how clinical skills are taught. A synthesis of the current evidence is needed to guide implementation and future research, given the pace of technological advancement.

Objective: This systematic review aims to synthesize empirical research on the design, implementation, and educational impact of GenAI-supported virtual patients in health care education.

Methods: A systematic search was conducted across 5 databases (CINAHL, Medline, Embase, Scopus, and Web of Science) from their inception to March 19, 2026. Reference lists of included studies and relevant systematic reviews were also screened. Peer-reviewed studies in English that evaluated GenAI-supported virtual patients using quantitative or mixed methods were included. Two reviewers independently screened studies and extracted data. Study quality and risk of bias were assessed critically using JBI (Joanna Briggs Institute) checklists, with disagreements resolved by consensus.

Results: A total of 15 studies met the inclusion criteria (total participants N=645), spanning health care disciplines, including nursing, medicine, pharmacy, radiography, and medical first-responder training. The virtual patients varied in design; input modalities included text (9 studies), voice (5 studies), or hybrid (1 study); output was text (9 studies), speech (5 studies), or both (1 study); 6 studies used 3D-embodied avatars, while 9 used nonembodied interfaces. A total of 13 studies used OpenAI GPT models (eg, ChatGPT), 1 used a fine-tuned model from a different provider, and 1 evaluated multiple model families (Claude, GPT, and open-source). Further, 6 studies used controlled experimental designs, including 3 randomized controlled trials (RCTs); the remainder were cross-sectional or prepost evaluations. Primary outcomes included user perceptions (14 studies), communication skills (4 studies), clinical reasoning (3 studies), and performance (7 studies). In controlled comparisons, GenAI-supported virtual patients consistently improved outcomes relative to control conditions: for example, enhanced clinical decision-making (RCT, n=21), ophthalmology history-taking skills (RCT, n=26), and medical history-taking performance (crossover RCT, n=20). The evidence base is characterized by brief intervention durations, a predominant reliance on single-session interactions, and a general lack of underpinning educational theory. No meta-analysis was performed due to the limited number of studies and significant heterogeneity in designs, interventions, and outcome measures.

Conclusions: The evidence supports the feasibility and acceptability of GenAI-supported virtual patients, with positive learner perceptions and promising outcomes for skills development. However, critical limitations remain in emotional-behavioral complexity, simulation adaptability, and research design rigor (eg, limited use of control groups and validated instruments). The review offers educators, instructional designers, and policymakers actionable insights for integrating dynamic, artificial intelligence–driven simulations while identifying crucial gaps—such as the need for theoretical grounding, longitudinal studies, and standardized design protocols—that must be addressed for safe and effective implementation.

Trial Registration: Open Science Framework (OSF) q8b5n; https://osf.io/q8b5n/files/mysz3

(J Med Internet Res 2026;28:e82756) doi:10.2196/82756

KEYWORDS

systematic review; generative AI; virtual patient; health care education; PRISMA; generative artificial intelligence; Preferred Reporting Items for Systematic Reviews and Meta-Analyses

XSL•FO
RenderX

Introduction

Background

Virtual patients are sophisticated educational simulations designed to replicate authentic clinical scenarios, enabling health care trainees to practice skills in a safe, controlled environment without risk to real patients [1,2]. They are defined as "a representation of an actual patient," which can include various forms such as software-based simulators or manikins, and specifically as "a computer program that simulates real-life clinical scenarios in which the learner acts as a health care provider," making clinical decisions [3]. They serve as a cornerstone of modern clinical education, aiming to enhance clinical reasoning, communication proficiency, and decision-making abilities [4]. The recent and rapid integration of generative artificial intelligence (GenAI)—a subset of artificial intelligence (AI) powered by large language models (LLMs) and natural language processing—is fundamentally transforming this educational tool [5,6]. Unlike traditional scripted simulations, GenAI-supported virtual patients can generate dynamic, adaptive, and contextually relevant responses in real-time, allowing for more realistic conversations, emotional responsiveness, and personalized learning experiences [7,8]. This shift marks a significant evolution in simulation-based learning, moving from static, preprogrammed cases toward interactive, intelligent patient encounters.

Rationale

The evolution of virtual patients demonstrates a clear trajectory toward greater interactivity and realism. From early, static computer-based cases on systems such as PLATO [9], modern iterations now incorporate multimedia, branching narratives, and data-driven models to create more engaging and realistic clinical encounters [10,11]. This evolution has yielded significant educational benefits. Empirical studies show that virtual patients can improve clinical decision-making, diagnostic accuracy, and foundational skills such as screening and referral across various health disciplines [11,12]. For instance, they have been successfully implemented as standardized, unfolding simulations to replace scarce pediatric clinical hours while maintaining clinical competency in nursing education [13]. Furthermore, a pilot study grounded in Experiential Learning Theory demonstrated that virtual patient simulation led to statistically significant improvements in clinical reasoning and communication skills among prelicensure nursing students [14]. They provide a scalable, consistent training environment that addresses limitations inherent to human standardized patients, such as fatigue and variability [7,15].

Despite these advances, a critical constraint remains: the predominant reliance on prescripted, linear scenarios. This often results in predictable interactions that fail to fully replicate the dynamic, adaptive, and complex nature of real patient encounters, potentially limiting learner engagement and the ability to respond to individualized student input [16,17]. Consequently, there is a recognized need for more dynamic and adaptable virtual patient models to better prepare learners for clinical practice [18].

GenAI presents a transformative solution to this limitation. A subset of AI powered by LLMs, GenAI can generate context-aware, realistic text, dialog, and even simulate human behaviors in real-time [19,20]. In health care education, integrating GenAI allows virtual patients to move beyond scripts, generating unique, adaptive responses based on learner inquiries and simulating a wider range of patient behaviors and emotional states [7,8]. This capability promises to significantly enhance the authenticity, personalization, and educational value of simulation-based training.

However, the implementation of this novel technology introduces new complexities regarding its design, pedagogical integration, and evaluation. While systematic reviews have established the effectiveness of traditional, scripted virtual patients [21,22] and scoped the broad potential of GenAI in education [5], a critical gap remains. Existing syntheses are unequipped to address the unique research questions (RQs) generated by their convergence. Specifically, the literature lacks evidence-based guidance on how the adaptive, nondeterministic nature of GenAI alters optimal instructional design, what novel evaluation frameworks are required to measure its impact on dynamic clinical reasoning, and how the practical challenges of implementation differ from those of static simulations. This dedicated synthesis is absent. Consequently, without a focused systematic review, evidence regarding the optimal design features, verifiable educational impact, and practical challenges of these advanced tools remains fragmented, hindering evidence-based adoption and focused research development [23,24].

Given these factors, it is essential to conduct a systematic review that not only summarizes the characteristics of GenAI-supported virtual patients but also provides comprehensive guidance for future research in this emerging field. Such a review will help educators, researchers, and policymakers understand the potential benefits and limitations of GenAI-supported virtual patients, thereby facilitating their integration into health care education curricula.

Objectives

The primary objective of this systematic review is to synthesize the current empirical evidence on the design, implementation, and educational impact of GenAI-supported virtual patients in health care education. Specifically, the review addresses the following RQs:

RQ1. What are the design choices, technological architecture, and educational strategies embedded within GenAI-supported virtual patients?

RQ2. What are the evaluation and educational impact, including benefits, outcomes, and related limitations of GenAI-supported virtual patients?

Methods

Study Design

This review was conducted and reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement and its expanded checklist

XSL•FO
RenderX

(Checklist 1) [25], as well as synthesis without meta-analysis (Multimedia Appendix 1) [26]. The protocol was registered and published in the Open Science Framework repository [27].

Eligibility Criteria

The following inclusion criteria were used: (1) original studies published in peer-reviewed journals, (2) focused on evaluating how GenAI-supported virtual patients affect education and training in health care–related disciplines, (3) included measurements of user experience or user outcomes, (4) conducted using quantitative or mixed methods research design, and (5) written in English. The exclusion criteria were the following: (1) book chapters, editorials, short communications, letters, and review literature; and (2) studies focused solely on the technical development of virtual patients without educational evaluation.

Information Sources

A systematic search was performed across 5 databases from their inception to March 19, 2026: CINAHL (via EBSCOhost), MEDLINE (via EBSCOhost), Embase (via Elsevier), Scopus (via Elsevier), and Web of Science Core Collection (via Clarivate). These databases were chosen for their comprehensive coverage of biomedical, nursing, allied health, and interdisciplinary literature. Furthermore, the reference lists of all included full-text papers were manually reviewed to identify additional relevant studies.

Search Strategy

The search strategy was developed iteratively for each database to ensure comprehensive coverage of the core concepts of "virtual patient" and "generative artificial intelligence." For each database, the strategy combined controlled vocabulary (where available, eg, MeSH [Medical Subject Headings] in MEDLINE and CINAHL Subject Headings) with extensive free-text terms, using appropriate field-specific syntax (eg, MH and XB in CINAHL and MEDLINE, :ti,ab in Embase, TITLE-ABS-KEY in Scopus, and TS= in Web of Science). Synonyms for virtual patients and GenAI models were grouped logically, and all free-text terms were searched in the title, abstract, and keyword fields. Searches were executed using the advanced search interface of each platform. The search was limited to records published in English and, where available, to peer-reviewed papers. The search was updated and rerun on March 19, 2026, after incorporating expanded terms and controlled vocabulary as suggested during peer review. The full search strategy is reported in Multimedia Appendix 2 and in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) checklist [28].

Several elements of the PRISMA-S checklist [28] did not apply to our methodology: we did not search databases simultaneously on a single platform (item 2); we did not search study registries (item 3); we did not browse online or print sources (item 4); we did not perform citation searching beyond checking reference lists of included studies (item 5); we did not contact authors, experts, or manufacturers to identify additional studies (item 6); we did not use any other information sources or methods beyond those described (item 7); we did not use published search filters (item 10); we did not adapt search strategies from prior reviews (item 11); we did not set up email alerts or automated updates, but we did manually rerun the search after refining the strategy (item 12); and the search was not formally peer reviewed (item 14). These items are explicitly noted as not applicable in this paper.

Selection Process

All records identified from the database searches were imported into Zotero reference management software for deduplication. The selection process was conducted in 2 phases. First, 2 reviewers (JJ and MZY) independently screened the titles and abstracts of all records against the eligibility criteria. Second, the full texts of potentially eligible studies were retrieved and independently assessed for inclusion by the same 2 reviewers. Any disagreements at either stage were resolved through discussion between the reviewers until consensus was reached. The interrater reliability was calculated using Cohen κ, resulting in κ=0.7, which indicates substantial agreement [29].

Data Collection Process

A standardized data extraction form was developed in Google Sheets (Google LLC). Further, 2 reviewers (JJ and MZY) independently extracted data from each included study. The extracted data were then cross-checked, and any discrepancies were resolved through discussion. The interrater reliability was κ=0.6, which indicates moderate agreement [29].

Data Items

Google Sheets (Google LLC) were used for data extraction. The following information was extracted from each of the papers that met the inclusion criteria: (1) publication characteristics: authors, publication year, and source; (2) study characteristics: study design and sample size; (3) intervention characteristics: description of the GenAI-supported virtual patient, including input or output modalities, use of an avatar, duration of interaction, technological details (eg, GenAI model and prompt engineering), and integration with educational strategies or theories; (4) outcomes: all reported outcome measures, including primary and secondary outcomes related to user perceptions (eg, usability, satisfaction, and perceived learning), skills (eg, communication and clinical reasoning), and performance; and (5) key results: main quantitative and qualitative findings as reported by this study's authors.

Study Risk of Bias Assessment

To critically appraise the methodological quality and risk of bias of the included studies, a formal assessment was conducted in accordance with PRISMA guidelines (item 11). The JBI (Joanna Briggs Institute) critical appraisal checklists, appropriate to each study design, were used as standardized tools [30]. Specifically, the JBI Checklist for Randomized Controlled Trials was used for randomized controlled trials (RCTs) [31], the JBI Checklist for Quasi-Experimental Studies was used for nonrandomized comparative studies [32], and the JBI Checklist for Analytical Cross-Sectional Studies was used for cross-sectional evaluations [32]. Further, 2 reviewers independently assessed each study. Any discrepancies in appraisal judgments were resolved through discussion to reach consensus. The interrater reliability was κ=0.8, which indicates

XSL•FO
RenderX

substantial agreement [29]. The results of this assessment are synthesized narratively and were considered when interpreting the overall strength and validity of the evidence presented in this review.

Effect Measures

Current review synthesized findings narratively, thus no common effect measures were pooled across studies due to significant heterogeneity in study designs, interventions, and outcome measures.

Synthesis Methods

A meta-analysis was not feasible due to the limited number of studies and substantial heterogeneity in interventions, populations, and outcome measures. Therefore, a narrative synthesis was conducted. The extracted data were summarized and organized to address the review's RQs. Findings were structured thematically to describe: (1) the implementation characteristics (design, technology, and educational strategies) of GenAI-supported virtual patients, and (2) their evaluation and educational impact (benefits, outcomes, methodological approaches, and limitations). The synthesis also integrates a discussion of the methodological quality and risk of bias of the included studies.

Reporting Bias Assessment

No formal statistical assessment of publication bias (eg, funnel plot) was performed due to the narrative synthesis approach and the small number of included studies.

Certainty Assessment

A formal assessment of the certainty of the body of evidence (eg, GRADE [Grading of Recommendations, Assessment, Development, and Evaluation]) was not conducted for this narrative review, as its primary aim was to map and characterize an emerging field rather than to estimate a pooled treatment effect.

Results

Study Selection

The initial literature search yielded 2860 studies. After screening the abstracts, 107 papers were selected for further evaluation. Further, 2 authors screened full texts independently, and the screening was then discussed as a group to ensure consensus and make the final selection decision. Ultimately, 15 papers were included in the final analysis after reading the full-text. Figure 1 provides a detailed overview of the systematic search procedure.

XSL•FO
RenderX

Figure 1. Summary of the selection of publications. GenAI: generative artificial intelligence.

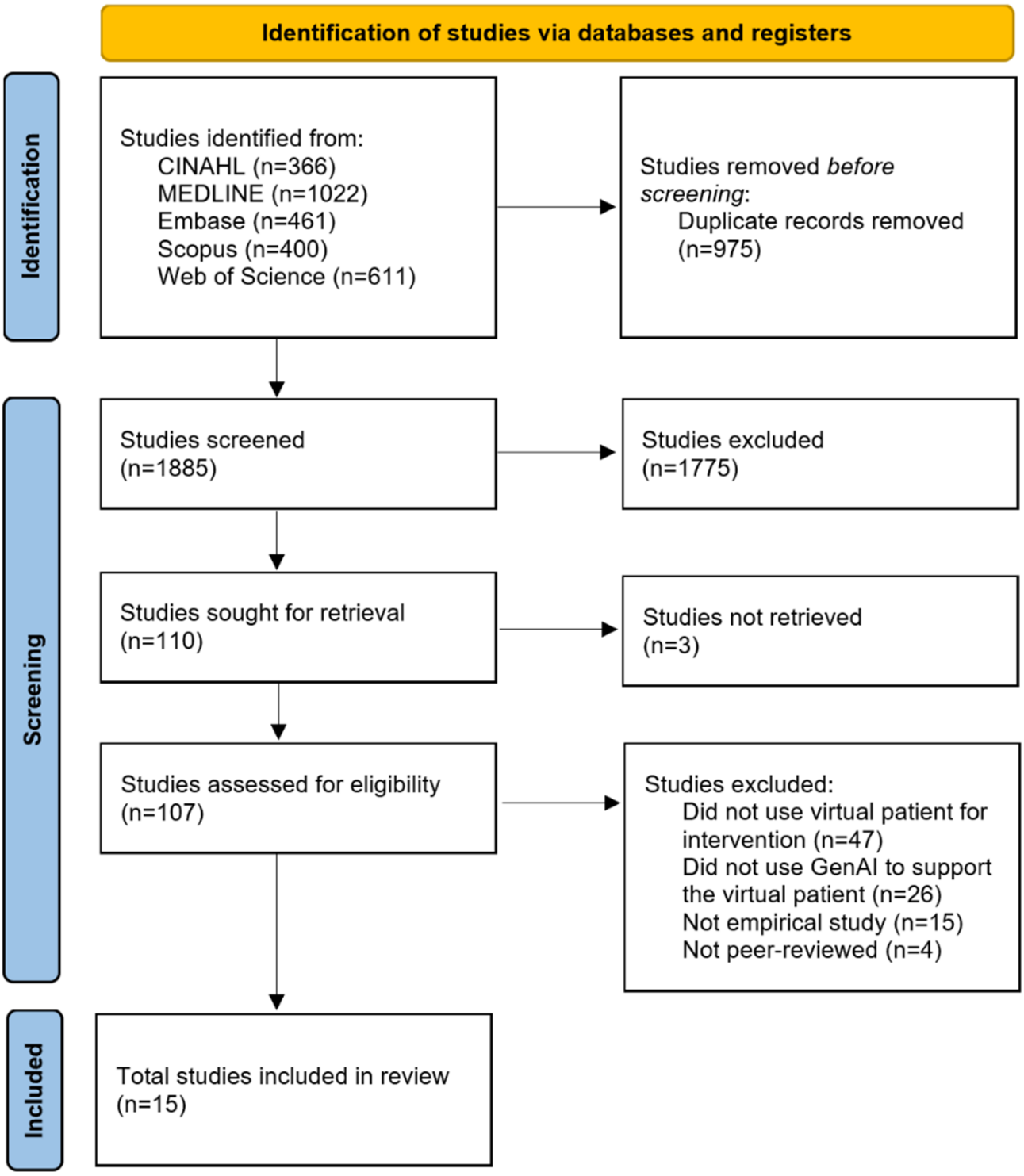

Study Characteristics

All 15 studies included in this review were published in peer-reviewed journals. Research in this area was prominently featured in the JMIR portfolio, which published 6 of the included papers: 3 papers in *JMIR Medical Education* [33-35], 2 papers in the *Journal of Medical Internet Research* [36,37], and 1 paper in *JMIR Formative Research* [38]. Temporally, 8 studies were published in 2024 [33-35,38-42], marking a peak of initial investigative activity, followed by 7 more studies in 2025 and 2026 [36,37,43-47], indicating sustained scholarly interest. The included studies used a diverse range of methodologies. The 2024 cohort predominantly featured cross-sectional (n=5) [33,34,38,39,42] and quasi-experimental (n=2) [35,40] designs, alongside 1 RCT [41]. The 2025‑2026 publications demonstrated a continued diversification of methods, including 2 additional RCTs [45,47], 4 quasi-experimental studies [36,43,44,46], and 1 proof-of-concept observational study [37]. Participant sample sizes varied widely, from smaller feasibility studies (eg, n=6 [40]) to larger-scale evaluations (eg, n=145 [35,43]), and a paired crossover study with 20 medical students [47]. The duration and frequency of interventions with the

XSL•FO
RenderX

GenAI-powered virtual patients also differed, ranging from brief, single interactions of approximately 6 - 10 minutes [38] to more extended or repeated practice sessions over several weeks [43]. This heterogeneity in publication venues, design, scale, and intervention format reflects the exploratory and rapidly evolving nature of research in this domain.

Moreover, among the 15 included studies, 13 of them explicitly stated obtaining ethics approval from an institutional review board and described informed consent procedures [33-39,41,43-45]. Specific data privacy or security measures were reported less consistently, detailed in only 8 (54%) studies [33-35,37-39,43,46]. Notably, 2 studies [40,42] documented that formal ethics approval was not required for their projects, citing their design as a quality assurance activity and a survey of professionals, respectively.

Risk of Bias in Included Studies

The methodological quality of the 15 included studies, as assessed by the JBI critical appraisal checklists, demonstrated a spectrum of risk of bias (Table 1). Further, 3 RCTs were judged to have a low to low-moderate risk of bias [41,45,47], benefiting from strong methodologies including randomization, allocation concealment, and blinded outcome assessment. However, the nature of the interventions precluded participant blinding, a common limitation in educational technology studies. In contrast, the body of quasi-experimental [35,36,40,43,44,46] and cross-sectional studies [33,34,37-39,42] presented a higher risk of bias. Common methodological weaknesses across these designs included the use of small, nonrepresentative samples, a lack of control groups, the absence of preintervention or longitudinal outcome measurements, and frequent reliance on unvalidated or self-reported outcome measures. Consequently, while the evidence from RCTs provides a stronger foundation for causal inference regarding the impact of GenAI-supported virtual patients, the overall findings of this review—particularly those related to user perceptions and feasibility—are drawn from a predominantly moderate-risk evidence base. This necessitates a cautious interpretation of the results, as positive outcomes may be influenced by study design limitations and enthusiasm for a novel technology.

XSL•FO
RenderX

Table . Risk of bias assessment of included studies.

Study type and reference citation	Overall risk of bias or appraisal judgment	Key concerns or limitations
RCTs[a]		
[41]	Low to moderate risk	Unclear allocation concealment; potential lack of blinding for treatment deliverers.
[45]	Low risk	Lack of participant blinding (acknowledged as an expected limitation for the intervention type).
[47]	Low risk	Well-designed crossover RCT; lack of participant blinding is an inherent limitation.
Quasi-experimental studies		
[40]	Moderate to high risk	Intentional selection of dissimilar participants; no control group; no preintervention measurement; unclear reliability of qualitative coding.
[35]	Moderate risk	Nonrandomized historical control group; no pretest for primary outcome; intervention was supplementary to the standard curriculum.
[36]	Moderate risk	Lack of a preintervention baseline measure for the primary outcome.
[43]	Moderate risk	No control group; no pretest measurement; reliance on self-reported data; unclear reliability of survey instrument.
[44]	Moderate risk	No control group; only partial prepost measurement for some outcomes; small convenience sample.
[46]	Moderate risk	Between-groups design with nonrandomized assignment; small sample size; single-site study.
Analytical cross-sectional studies		
[39]	High risk	Very small sample; lack of validated measures; no control for confounders; limited generalizability.
[33]	Moderate risk	Lack of confounding factor consideration; use of a single case scenario.
[34]	Moderate risk	Lack of consideration for confounding factors; single case or model; potential selection bias.
[42]	Moderate risk	Small sample; lack of objective outcome measures; no control group; potential for selection or response bias.
[38]	Moderate risk	Subjective outcome measures; no objective performance assessment; no control group; small sample.
[37]	Moderate risk	No control group; small single-site sample; reliance on self-reported benefits; no long-term skill retention measures.

[a]RCT: randomized controlled trial.

Results of Syntheses

To address the RQ1 concerning the design and implementation characteristics of GenAI-supported virtual patients, the following sections analyze the extracted data, which are comprehensively tabulated in Table 2.

XSL•FO
RenderX

Table . Characteristics of the virtual patients and the GenAI[a] models used in the 15 included studies. Detailing input or output modalities, avatar design, the AI's[b] tasked role, and using educational frameworks.

Reference citation	Input (eg, text and voice)	Output (eg, text and voice)	Avatar (2D/3D)	Movement	Emotion expression	Type or name of AI	Overview of the task for GenAI based on the designated prompt	Educational theories or models
[39]	Text	• Text	N/A[c]	N/A	N/A	• ChatGPT (version not specified)	• Simulate a realistic clinical interaction focusing on assessment, communication, and nursing care for respiratory distress	• N/A
[40]	Text	• Text	N/A	N/A	Using case scenarios such as patients with claustrophobia undergoing an MRI[d] scan	• OpenAI ChatGPT 3.5 and ChatGPT 4	• A role-play designed to simulate a conversation between a radiology technician and a patient with claustrophobia during an MRI examination. The AI takes the role of the patient, while the user plays the technician.	• N/A
[33]	Text	• Text	N/A	N/A	N/A	• OpenAI GPT-4	• Two prompts were developed: one for providing the interactive history-taking dialog, and the other for giving feedback	• N/A

XSL•FO
RenderX

Reference citation	Input (eg, text and voice)	Output (eg, text and voice)	Avatar (2D/3D)	Movement	Emotion expression	Type or name of AI	Overview of the task for GenAI based on the designated prompt	Educational theories or models
[34]	Text	• Text	N/A	N/A	N/A	• OpenAI GPT-3.5	• GPT acts as a simulated patient. The prompts were designed to guide GPT's behavior and ensure it provided medically accurate and relevant responses.	• N/A
[35]	Text	• Text	N/A	N/A	The emotional parameters were set from 1 to 10 for 8 emotions: joy, sadness, anticipation, surprise, fear, disgust, trust, and anger	• GPT-4 Turbo		• N/A

XSL•FO
RenderX

Reference citation	Input (eg, text and voice)	Output (eg, text and voice)	Avatar (2D/3D)	Movement	Emotion expression	Type or name of AI	Overview of the task for GenAI based on the designated prompt	Educational theories or models
							• The prompt is designed to simulate a chatbot role-playing as a medical patient with dynamic emotional behavior. It consists of two major phases: (1) role-play phase (simulated patient behavior): governs how the chatbot behaves during the medical consultation, and (2) feedback phase (interaction evaluation): after the roleplay ends, the chatbot switches to feedback mode and evaluates the user's performance.	
[41]	Text	• Text	N/A	N/A	N/A	• OpenAI GPT-3.5		• N/A

XSL•FO
RenderX

Reference citation	Input (eg, text and voice)	Output (eg, text and voice)	Avatar (2D/3D)	Movement	Emotion expression	Type or name of AI	Overview of the task for GenAI based on the designated prompt	Educational theories or models
							• Control group (AI simulation only): a virtual patient scenario crafted for emergency and neurological assessment training. The AI simulates a patient experiencing a traumatic brain injury. • Feedback group (AI simulation+ AI feedback): the AI first simulates the patient with the same setting in the control group and then provides diagnostic feedback assessment for participants who play as the doctor after their interaction.	
[42]	Text, voice	• Text, voice	3D	N/A	N/A			• N/A

XSL•FO
RenderX

Reference citation	Input (eg, text and voice)	Output (eg, text and voice)	Avatar (2D/3D)	Movement	Emotion expression	Type or name of AI	Overview of the task for GenAI based on the designated prompt	Educational theories or models
						• Generative conversational AI, specifically using the platform Convai (Convai Technologies Inc) and incorporating ChatGPT (version was not mentioned) for text generation	• Provide comprehensive content on ORIF[e] surgery suitable for training a large language model, which is then subsequently further expanded.	
[38]	Voice	• Voice	3D	Vive Trackers (version 3.0) were placed on the head, hands, feet, and groin of the manikin and mapped to the corresponding parts of the VP's[f] avatar. This allowed the MFRs[g] to freely move the manikin and thus the VP.	Emotion is expressed through the voice, including stuttering, groans, and cries, as well as statements reflecting fear and pain	• OpenAI GPT-3.5-Turbo	• This simulation helps practice trauma-informed care, nonverbal cues, and managing patients in acute distress. All responses are limited to a maximum of 8 stuttered words.	• N/A
[36]	Voice (student's spoken questions, converted to text via speech-to-text)	• Voice (synthesized speech via text-to-speech) • Visual (facial expressions projected onto the robot)	3D (Furhat social robot with animated face back-projected onto a translucent mask)	Natural head movements (neck with 3 degrees of freedom)	Facial expressions (eg, sad, happy, or surprised) were generated and synchronized with speech	• OpenAI GPT-3.5-turbo		• N/A

XSL•FO
RenderX

Reference citation	Input (eg, text and voice)	Output (eg, text and voice)	Avatar (2D/3D)	Movement	Emotion expression	Type or name of AI	Overview of the task for GenAI based on the designated prompt	Educational theories or models
							• Dialog generation: to generate the next patient dialog line. The prompt includes the patient case description, the last 10 dialog turns, and instructions to respond as the patient. • Expression generation: to select appropriate facial expressions (from a predefined set) at anchor points within the generated dialog text to reflect the patient's emotional state.	
[43]	Text	• Text	N/A	N/A	Textual descriptions of emotional state within dialog (eg, expressing stress or motivation). No visual or auditory emotion simulation	• OpenAI ChatGPT 3.5		• The 5As framework for smoking cessation counseling

XSL•FO
RenderX

Reference citation	Input (eg, text and voice)	Output (eg, text and voice)	Avatar (2D/3D)	Movement	Emotion expression	Type or name of AI	Overview of the task for GenAI based on the designated prompt	Educational theories or models
							• Patient simulation: to act as a smoker seeking to quit, responding in character to student-led counseling based on a predefined case scenario that includes demographics, smoking habits, and motivations. • Performance feedback: after the counseling session, to evaluate the student's performance based on a structured rubric (the 5As[h] framework, empathy, communication skills, etc) and provide detailed textual feedback on strengths and areas for improvement.	
[37]	Text	• Text	N/A	N/A		• OpenAI GPT-4		• N/A

XSL•FO
RenderX

Reference citation	Input (eg, text and voice)	Output (eg, text and voice)	Avatar (2D/3D)	Movement	Emotion expression	Type or name of AI	Overview of the task for GenAI based on the designated prompt	Educational theories or models
					Textual descriptions of the patient's emotional state within dialog (eg, expressing anxiety or distress). No visual or auditory emotion simulation			

XSL•FO
RenderX

Reference citation	Input (eg, text and voice)	Output (eg, text and voice)	Avatar (2D/3D)	Movement	Emotion expression	Type or name of AI	Overview of the task for GenAI based on the designated prompt	Educational theories or models
							• Patient simulation: to act as a simulated patient (eg, a man aged 28 years with depression or a woman aged 46 years with agoraphobia) and respond in character to physician-led text-based dialog. • Real-time feedback generation: to analyze the physician's incoming text messages in real-time, identify the use of specific communication techniques (eg, open questions, reflections, empathy, or validation), and provide immediate formative textual feedback within the chat to confirm and encourage tech-	

XSL•FO
RenderX

Reference citation	Input (eg, text and voice)	Output (eg, text and voice)	Avatar (2D/3D)	Movement	Emotion expression	Type or name of AI	Overview of the task for GenAI based on the designated prompt	Educational theories or models
							nique use. • Summative feedback generation: after the chat, to provide summarized feedback on the frequency of technique use, highlight underused techniques, and give examples for future application.	
[44]	Voice (learner's spoken questions to the virtual patient via HMD[i], processed by a speech-to-text model)	• Voice (virtual patient's spoken responses generated by the AI, delivered via HMD with an AI-generated voice)	3D	Limited (the virtual human avatar is present but has limited physical interaction; cannot perform actions such as raising clothes or turning over as requested)	Limited (primarily text-based emotional cues within dialog. This study notes limitations such as unnatural voice expressiveness and an absence of emotional sentiment	• OpenAI GPT-4o		• Instructional design model for AI education • Technology acceptance model (for evaluation)

XSL•FO
RenderX

Reference citation	Input (eg, text and voice)	Output (eg, text and voice)	Avatar (2D/3D)	Movement	Emotion expression	Type or name of AI	Overview of the task for GenAI based on the designated prompt	Educational theories or models
							• Patient simulation: to act as a virtual patient (a male aged 28 years with acute appendicitis) and engage in real-time, natural language dialog with nursing students for health assessment and communication training. The AI is provided with a detailed patient script and vital sign data to generate contextually relevant and medically accurate responses to the learner's verbal questions.	
[45]	Voice	• Voice	Visual avatar (image of patient's eye; 2D/3D not specified)	N/A	Expressed through voice (eg, anxiety or emotional responses)	• Fine-tuned Baichuan-13B-Chat (a large language model)		• Kolb's experiential learning cycle • Calgary-Cambridge communication framework

XSL•FO
RenderX

Reference citation	Input (eg, text and voice)	Output (eg, text and voice)	Avatar (2D/3D)	Movement	Emotion expression	Type or name of AI	Overview of the task for GenAI based on the designated prompt	Educational theories or models
							• To simulate a digital ophthalmology patient for medical history-taking practice. The AI acts as the patient, responding in character to students' verbal inquiries based on a detailed knowledge base derived from electronic health records. The system provides real-time interaction and, after the session, generates automated feedback and scores based on the comprehensiveness of the history taken.	
[46]	Voice	• Voice	3D	Not specified (avatar is static in the examination room; no description of physical movement)		• OpenAI GPT-4o (via API[k] calls)		• N/A

XSL•FO
RenderX

Reference citation	Input (eg, text and voice)	Output (eg, text and voice)	Avatar (2D/3D)	Movement	Emotion expression	Type or name of AI	Overview of the task for GenAI based on the designated prompt	Educational theories or models
					Facial expressions rendered in real-time by the D-ID[j] platform to create a full audiovisual experience; tone of voice (eg, irritable or rapport-building) also conveys emotion			

XSL•FO
RenderX

Reference citation	Input (eg, text and voice)	Output (eg, text and voice)	Avatar (2D/3D)	Movement	Emotion expression	Type or name of AI	Overview of the task for GenAI based on the designated prompt	Educational theories or models
							• Patient simulation: to act as a virtual patient ("Randy Rhodes," a man aged 54 years with type 2 diabetes) for medical students to interview via voice-to-voice interaction. Custom agent instructions are informed by faculty-generated case materials. "Guardrails" are placed to optimize educational value (eg, preventing the AI from revealing the diagnosis directly or ensuring accurate presentation of pertinent positive findings). Within these limits, the AI is allowed to respond adaptively to students' questions to main-	

XSL•FO
RenderX

Reference citation	Input (eg, text and voice)	Output (eg, text and voice)	Avatar (2D/3D)	Movement	Emotion expression	Type or name of AI	Overview of the task for GenAI based on the designated prompt	Educational theories or models
							tain realism.	
[47]	Text	• Text	N/A	N/A	Textual descriptions of patient emotional state are generated based on integrated personality profiles (eg, the Big Five framework) to simulate emotional realism (eg, pain and anxiety). No visual or auditory emotion simulation	Claude models (5): • Claude3 Haiku • Claude-3-Sonnet • Claude-3 - 5 Sonnet • Claude-4-Sonnet • Claude-4-Opus GPT-family models (3): • GPT-4 Turbo • GPT-4o • GPT-3.5 Turbo Open-source models (3): • DeepSeek-V3 671B • Qwen3-32B • LLaMa-3 70B	• To act as a simulated patient ("AIPatient") based on real EHR[l] data from the MIMIC-III[m] database. The AI engages in text-based dialog with medical students for history-taking practice. Its task is to provide accurate, readable, and consistent responses to clinical questions while incorporating diverse personality traits to simulate realistic patient behavior, including emotional expressions.	• N/A

[a]GenAI: generative artificial intelligence.

[b]AI: artificial intelligence.

[c]N/A: not applicable.

[d]MRI: magnetic resonance imaging.

[e]ORIF: open reduction and internal fixation.

[f]VP: virtual patient.

[g]MFR: medical first responder.

[h]5A: Ask, Advise, Assess, Assist, Arrange.

XSL•FO
RenderX

[i]HMD: head-mounted display.

[j]D-ID: deidentification.

[k]API: application programming interface.

[l]EHR: electronic health record.

[m]MIMIC-III: Medical Information Mart for Intensive Care III.

Design Choices

The designs of GenAI-supported virtual patients can be classified into three distinct categories: (1) input, (2) output, and (3) avatar. First, in terms of input methods, 11 of 15 studies reported that participants interacted with the virtual patient by entering text [33-35,37,39-41,43,47]. Meanwhile, 5 studies allowed participants to communicate verbally using voice input [36,38,44-46]. Additionally, 1 study incorporated a hybrid approach, enabling participants to use either speech-to-text functionality via a microphone or direct text input through a designated field [42]. Next, the output modalities of the virtual patient corresponded closely to the input mechanisms. In 9 studies, the virtual patient responded to participants via text-based communication [33-35,37,39-41,43,47]. In contrast, 5 studies featured virtual patients capable of generating human-like voice responses [36,38,44-46]. Notably, 1 study highlighted a more versatile approach, where the virtual patient was designed to provide responses to both text and synthesized speech [42].

Regarding the avatar design of the virtual patient in the included studies, 6 studies used a 3D-embodied virtual patient to enhance realism and immersion for participants [36,38,42,44-46]. For instance, a study [38] integrated a mixed reality tool, allowing participants not only to visually perceive the virtual patient within a digital environment but also to physically interact with a corresponding manikin. This setup enabled the virtual patient to display various injuries, movements, and facial expressions aligned with speech production, respiration patterns, and pain-related vocalizations. Similarly, research by Borg et al [36] implemented a virtual patient embedded within a robotic system. Their robot featured a 3-degree-of-freedom neck and an animated face, facilitating flexible head movements and expressive emotional displays. Mool et al [46] also used a 3D avatar, though it was noted to be largely static within the examination room environment. In contrast, the remaining 9 studies used virtual patients without avatars, relying solely on alternative interaction modalities [33-37,39-41,43,47].

In the 15 included studies, interventions involving participants interacting with GenAI-supported virtual patients lasted less than 10 minutes in 5 cases [33,38-41]. Three studies reported intervention durations exceeding 20 minutes [43-45], while the remaining 5 did not specify the duration [34-37,42]. Moreover, 7 studies featured only a single session, regardless of the intervention length [33,35,38,39,42,44,45].

Technological Architecture

A total of 13 studies explicitly stated that they used an OpenAI GPT model for generating the virtual patient [33-44,46]. Further, 1 study used a fine-tuned model from a different provider [45]. Yu et al [47] evaluated a broader range of models for their AIPatient system, including Claude models (Haiku, Sonnet variants), GPT-family models (GPT-4 Turbo, GPT-4o, and GPT-3.5 Turbo), and open-source models (DeepSeekV3 671B, Qwen3-32B, and LLaMa-3 70B). All studies specified the foundational GenAI model used. A total of 14 studies provided detailed patient case information as part of the prompt to enhance the virtual patient's responses [33-35,38,40-47]. For instance, in a study [36], the prompt consisted of a structured patient case description, the previous 10 turns of dialog, and an instruction to generate the next line of conversation. Yu et al [47] took a sophisticated approach, integrating real electronic health record data from the MIMIC-III (Medical Information Mart for Intensive Care III) database and incorporating personality profiles based on the Big Five framework to simulate diverse and realistic patient behaviors. In contrast, 1 study used a simpler prompt, wherein the GenAI was instructed merely to assume the role of a virtual patient with a specified condition (eg, respiratory distress) and engage in dialog with participants acting as nurses, without requiring additional case-specific details [39].

Educational Strategies

A total of 3 studies explicitly referenced established educational frameworks to inform their design. The study by Kim et al [44] applied an instructional design model for AI education and the technology acceptance model. The study by Luo et al [45] used Kolb's experiential learning cycle and the Calgary-Cambridge communication framework, while the study by Chinwong et al [43] was grounded in the 5As framework for smoking cessation counseling. The study by Kim et al [44] emphasized the critical role of patient case design, noting that such cases serve as the foundational structure of the curriculum, functioning as learning triggers and providing a platform for students to engage in cognitive processes reflective of physicians' workplace reasoning. Given the high-fidelity patient simulation and the level of control afforded by GenAI, this innovative approach has the potential to enhance medical education curricula, offering valuable benefits for both students and educators. In contrast, the remaining 12 studies did not specify the application of any educational theory to inform the design of the virtual patient or the overall study methodology [33-42,46,47].

To address the RQ2 concerning the educational effectiveness and learner outcomes of GenAI-supported virtual patients, the following sections analyze the extracted data, which are comprehensively tabulated in Table 3.

XSL•FO
RenderX

Table . Study design, educational purpose, intervention details, measured outcomes, and primary results of the 15 included studies.

Reference citation	Study design (participants, n)	Educational purpose	Session (duration), n	Outcomes	Validity or reliability test	Results of intervention
[39]	Cross-sectional study (n=12)	Patient communication	1 (10 min)	• Ease of use of ChatGPT • Learning engagement with ChatGPT • Recognition of the usefulness of ChatGPT in clinical education • Performance in virtual patient interaction	• N/A[a]	• Students responded positively to ChatGPT, finding it accessible, engaging, and valuable as a training tool. Those with stronger interaction skills tended to perform better overall. Key communication attributes such as clarity, relevance, and usefulness were linked to stronger outcomes.
[40]	Quasi-experimental study within design (n=6)	Radiographers' communication skills with patients with claustrophobia	10 (2 min)	• Simulation success rate • Radiographers' communication skills	• N/A	• A total of 60 simulations were conducted, achieving a success rate of 96.7% (58/60). ChatGPT-3.5 exhibited errors in 40% (12/30) of the simulations, while ChatGPT-4 showed no errors. The simulation of clinical scenarios via ChatGPT proves valuable in assessing and testing radiographers' communication skills, especially in managing patients with claustrophobia during MRI.[b]

XSL•FO
RenderX

Reference citation	Study design (participants, n)	Educational purpose	Session (duration), n	Outcomes	Validity or reliability test	Results of intervention
[33]	Cross-sectional study (n=106)	Patient history taking	1 (8 minutes)	• Quality of OpenAI GPT-4's role-play capability • Completeness of history taking	• Interrater reliability, measured by Cohen κ	• OpenAI GPT-4 demonstrated highly realistic medical responses, with over 99% deemed plausible. Its evaluations closely matched human ratings overall, though some feedback categories showed weaker agreement where OpenAI GPT-4's assessments were more detailed or differed from human perspectives.
[34]	Cross-sectional study (n=28)	Patient history taking for medical students	N/A	• The performance of OpenAI GPT as a simulated patient • Chatbot's usability	• N/A	• When questions were explicitly covered by the script (n=502, 60.3%), the GPT-provided answers were mostly based on explicit script information (n=471, 94.4%).
[35]	Quasi-experimental study (intervention group n=35, control group n=110)	Medical students' interview skills	1 (N/A)	• The scores related to medical interviewing in the pre-CC[c] OSCE[d] • Simulation-based training quality	• N/A	• Students in the AI[e]-supported group performed better in medical interviews compared to those in the control group. An inverse relationship was noted between their self-reported confidence scores and earlier examination results. Importantly, no safety issues were identified throughout the study.
[41]			4 (6 min)		• N/A	

XSL•FO
RenderX

Reference citation	Study design (participants, n)	Educational purpose	Session (duration), n	Outcomes	Validity or reliability test	Results of intervention
	Randomized controlled trial (control group n=11, feedback group n=10)	Clinical decision-making in medical students		• The performance of the participants • Clinical reasoning ability		• Medical students showed notable improvement when provided feedback. Initially, both the feedback and control groups performed similarly, confirming balanced assignment. By the end, the feedback group scored significantly higher overall, particularly in creating context and gathering information during clinical decision-making. However, there was no marked progress in their question-focusing skills.
[42]	Cross-sectional study (n=15)	Anesthesia training	1 (N/A)	Students' perception of the virtual patient: • Intuitive • User-friendly • Accuracy • Usability (use the model comfortably) • Feasibility	• N/A	

XSL•FO
RenderX

Reference citation	Study design (participants, n)	Educational purpose	Session (duration), n	Outcomes	Validity or reliability test	Results of intervention
						• The survey of 15 anesthetists revealed that the tool was generally well received. It had a median rating of 9 out of 10 for how intuitive and user-friendly it was, and a score of 8 out of 10 for simulating realistic patient responses and behaviors. Furthermore, 87% of the participants reported feeling comfortable using the model, suggesting strong confidence in its design and functionality. It seems the tool succeeded in both usability and clinical accuracy.
[38]	Cross-sectional study (n=24)	Communication training in an emergency (ie, car accident)	1 (6 - 10 min)	• Perception of voice quality • Usability of voice interactions	• N/A	

XSL•FO
RenderX

Reference citation	Study design (participants, n)	Educational purpose	Session (duration), n	Outcomes	Validity or reliability test	Results of intervention
						• The usability assessment of the virtual patient yielded moderately positive feedback, with particularly favorable scores in habitability and likeability. However, the roughly 3-second delay in response time detracted from the fluidity of interactions. MFRs[f] found it natural to evaluate the virtual patient's physiological state through verbal questions, but they also noted limitations in the dialog flow, especially the virtual patient's inability to initiate conversation. A key insight emerged around the potential of using domain-specific prompt engineering to guide responders more effectively during training.
[36]	Quasi-experimental study within design (n=15)	CR[g] training in rheumatology, comparing a social robotic VP[h] platform (LLM[i]-enhanced) to a conventional computer-based platform	1 VP case per platform (order counterbalanced). Duration not specified	• Virtual patient evaluation • Qualitative experiences of clinical reasoning, communication, and emotional skill training	• N/A	

XSL•FO
RenderX

Reference citation	Study design (participants, n)	Educational purpose	Session (duration), n	Outcomes	Validity or reliability test	Results of intervention
						• Quantitative: the social robotic platform was rated significantly higher for authenticity (mean 4.5 vs 3.9, *P*=.04) and overall learning effect (mean 4.4 vs 4.1, *P*=.01). • Qualitative: students found the robot superior for training CR, communication, and emotional skills, despite noting technical limitations.
[43]	Single group quasi-experimental, pre-post (n=145)	To practice smoking cessation counseling using the 5As[j] framework with an AI-simulated patient	Practice over 3 weeks (unrestricted frequency or duration), followed by a 2-hour classroom discussion session	• Student satisfaction • Perceived learning impact • Perceived benefits • Perceived difficulties	• N/A	• 66% of students were satisfied. Further, 84.4% reported improved understanding. Key benefits included self-assessment and adaptability. Major challenges were technical issues (88.3%) and a lack of AI understanding (58.6%).
[37]	Proof-of-concept observational study (n=28)	To train communication techniques (eg, empathy and motivational interviewing) for mental health encounters using an AI chatbot with real-time feedback	2 chat sessions (20 minutes each) with 2 different AI-simulated patients	• Accuracy of AI-generated feedback (expert-evaluated) • Participant perception of feedback • Change in frequency of communication techniques • Perceived benefit for clinical practice	• N/A	

XSL•FO
RenderX

Reference citation	Study design (participants, n)	Educational purpose	Session (duration), n	Outcomes	Validity or reliability test	Results of intervention
						• 85.38% of real-time feedback was partially or totally correct. Further, 87.27% of participants found the feedback helpful. A significant increase in the use of targeted techniques was observed from chat 1 to chat 2 (Poisson regression, *P*<.001). Over 80% agreed that the training helped them practice and apply new techniques.
[44]	Single group quasi-experimental, pre-post (n=28)	To train health assessment and therapeutic communication skills for patients with acute appendicitis using a GPT-based VP in VR[k]	1 session (1 hour total for interaction and practice)	Quantitative: • Usability • Perceived virtual learning environment (immersion, usefulness, etc) • Self-efficacy in communication Qualitative: • Training experiences • Accuracy of AI dialogs • Safety of AI dialogs • Relevance of AI dialogs • Readability of AI dialogs	• Reliability tested via Cronbach α for all scales: usability (α=.85), perceived virtual learning environment (α=.95), self-efficacy of communication (α=.88).	• Self-efficacy in communication increased significantly (pre: 61.57, post: 64.32, *P*=.009). The highest scores were for immersion and function accessibility. Qualitative themes highlighted educational benefits and technical limitations. AI dialog scored highest on readability and lowest on accuracy.
[45]	RCT[l] (LLMDP[m] group n=13, control group n=13)	To enhance ophthalmology medical history-taking skills using an LLM-based digital patient system	1 (1 h)	• Medical history-taking ability • Empathy • Student attitudes or satisfaction	• N/A	

XSL•FO
RenderX

Reference citation	Study design (participants, n)	Educational purpose	Session (duration), n	Outcomes	Validity or reliability test	Results of intervention
						• The LLMDP group showed significantly higher MH-TA[n] scores (mean 64.62, SD 9.52) vs control (54.12, SD 8.80), mean difference 10.50 points (95% CI 4.66 - 16.33, *P*<.001). The intervention group also demonstrated better empathy. High student satisfaction was reported, highlighting benefits for confidence and cost or time savings.
[46]	Between-groups, mixed methods study (GenAI[o] groups n=13, ePBLM[p] group n=13)	To explore what happens when a GenAI-enabled virtual patient is introduced within a PBL[q] tutorial for history-taking, compared to a legacy multimedia database system (ePBLM)	1 (GenAI groups: 55 - 65 min; ePBLM groups: 36 - 39 min)		Validity: • Survey items were judged by a behavioral scientist to be typical of those in other simulated patient studies. • Quiz items were verified by 2 faculty members for consistency with the case and readability. Reliability: • Quiz grading: first 5 quizzes graded independently by 2 coinvestigators to establish consistency; remaining 52 quizzes graded by 1 investigator. • Statistical tests: linear regression (OLS[r] estimation), 2-tailed with α=.05.	

XSL•FO
RenderX

Reference citation	Study design (participants, n)	Educational purpose	Session (duration), n	Outcomes	Validity or reliability test	Results of intervention
				• Primary (observational): characterization of student interactions with the patient modality and with each other during history-taking (via descriptive observation of audio-recordings), secondary (survey or quiz) • Learner perceptions: 8-item survey (5-point Likert) assessing perceptions of PBL tutorial quality (clinical accuracy, enjoyability, teamwork, etc) • Patient history recall: 11-item short-answer quiz assessing recall of patient case information (immediate and 2-week delayed)		

XSL•FO
RenderX

Reference citation	Study design (participants, n)	Educational purpose	Session (duration), n	Outcomes	Validity or reliability test	Results of intervention
						Observational findings: • GenAI presented essential case content accurately but occasionally deviated on nonessential content (eg, embellished responses, inconsistent headache history, and unreported marijuana use). • GenAI groups took ≈10 minutes longer on history-taking, partly due to collaborative troubleshooting of AI interaction. • Students treated the avatar like a sophisticated "question base," using closed-ended questions and jargon, not realistic patient interviewing. • One GenAI group showed more experimental, anthropomorphizing engagement (eg, using the patient's name and inferring attitude). Survey results: • GenAI students rated their experience significantly higher than their prior ePBLM experiences (mean total score 34.38 vs 28.38 pretutorial, *P*=.003). • Largest gains were in "simulates clinical experiences accurately"

XSL•FO
RenderX

Reference citation	Study design (participants, n)	Educational purpose	Session (duration), n	Outcomes	Validity or reliability test	Results of intervention
						(mean increase of 1.6 points). Quiz results: • Immediate recall was near ceiling in both groups (GenAI: 10.10/11; ePBLM: 9.40/11). • Delayed recall (2 weeks) decreased significantly in both groups (GenAI: 8.63; ePBLM: 7.94), but the rate of forgetting did not differ by condition (P=.052 for condition effect).
[47]	Paired crossover study (n=20 medical students)	To evaluate the fidelity, usability, and educational effectiveness of the AIPatient[s] system compared to H-SPs[t] in medical history-taking	4 interactions per student (2 cases×2 modalities: AIPatient and H-SP). Duration not explicitly specified		• Intercoder reliability: F_1-score (0.79) for NER gold-standard labels; Cohen κ (0.92) for QA accuracy ratings	

XSL•FO
RenderX

Reference citation	Study design (participants, n)	Educational purpose	Session (duration), n	Outcomes	Validity or reliability test	Results of intervention
				Primary (system performance): • Knowledge-base validity: NER[u] F_1-score • QA[v] accuracy: percentage correct in EHR[w]-based QA • Readability: Flesch Reading Ease, Flesch-Kincaid Grade Level • Robustness: accuracy variance with paraphrased questions (ANOVA) • Stability: accuracy variance with 32 personality types (ANOVA and data loss percentage) Secondary (user study): • Fidelity, usability, and educational effectiveness: measured via 5-point Likert-scale questionnaire • Clinical information gathering: OSCE-style checklist • Qualitative feedback: semistructured interviews		• AIPatient matched or exceeded H-SPs across most metrics. • Significant advantages: emotional realism (4.37 vs 3.74, $P<.01$), technical reliability (4.39 vs 3.79, $P<.01$), improving clinical reasoning skills (4.41 vs 3.97, $P<.05$). • OSCE checklist: AIPatient performed comparably or better in supporting clinical reasoning and information elicitation. • Qualitative: students found AIPatient emotionally expressive, pedagogically valuable, efficient, consistent, and usable. Identified areas for improvement included verbosity and handling of nonstandard queries.

[a]N/A: not applicable.
[b]MRI: magnetic resonance imaging
[c]Pre-CC: preclinical clerkship.
[d]OSCE: objective structured clinical examination.
[e]AI: artificial intelligence.
[f]MFR: medical first responder.
[g]CR: clinical reasoning.
[h]VP: virtual patient.
[i]LLM: large language model.
[j]5A: Ask, Advise, Assess, Assist, Arrange.
[k]VR: virtual reality.
[l]RCT: randomized controlled trial.
[m]LLMDP: large language model-based digital patient.

XSL•FO
RenderX

[n]MHTA: mental health therapy aide.
[o]GenAI: generative artificial intelligence.
[p]ePBLM: electronic problem-based learning.
[q]PBL: problem-based learning.
[r]OLS: ordinary least squares.
[s]AIPatient: artificial intelligence patient.
[t]H-SP: human-simulated patient.
[u]NER: named entity recognition.
[v]QA: question answering.
[w]EHR: electronic health record.

Educational Benefits and Learner Outcomes

Among all 15 included studies, 14 assessed participants' perceptions of the GenAI-supported virtual patient, examining factors such as usefulness [39], accuracy [42,44], and the authenticity of the patient encounter [36,37,46,47]. Additionally, 11 studies investigated the impact of GenAI-supported virtual patients on participants' learning outcomes, including performance [33-35,39,41,45,47], communication skills [38,40,44,46], and clinical reasoning ability [36,41,47]. Across multiple studies, GenAI-supported virtual patients demonstrated substantial benefits in health care education. Students rated the tool as accessible, engaging, and pedagogically valuable [39,42,46,47], with advanced models such as ChatGPT-4 achieving high scenario completion rates and error-free performance [33,40]. The simulations yielded highly realistic clinical responses and plausible feedback [33,35], and emotionally expressive interactions were found to be appropriate and contextually accurate [36,47]. Further studies confirmed strong user experiences and authentic communication [38], with improved outcomes in medical interview performance [35,45] and decision-making when feedback mechanisms were integrated [37,41]. The tool also earned high marks for intuitiveness and user comfort [42], enhanced perceptions of authenticity and learning effectiveness over conventional approaches [36,47], and received positive feedback on self-efficacy and the learning environment [44]. Further, 1 study reported that 84.4% of participants perceived an improved understanding of the subject matter [43].

Evaluation Designs and Assessment Strategies

A total of 6 studies used an experimental design incorporating a control group to assess the effects of GenAI-supported virtual patients [35,36,41,45-47]. Within these 6 studies, 5 of them compared the GenAI virtual patient with traditional pedagogical methods or a non-AI control [35,36,45-47]. The remaining 1 study used GenAI virtual patients in both the intervention and control groups, but with variations in functionality (conversation-only vs conversation+feedback) [41]. The other 9 studies did not include comparative analyses between GenAI and alternative conditions or conduct prepost intervention comparisons [33,34,37-40,42-44]. None of the studies in this review used a longitudinal design. Regarding data collection approaches, 9 studies relied exclusively on quantitative data [34,35,38-42,45,47], while the remaining 5 studies incorporated both quantitative and qualitative methods [36,37,43,44,46].

Design Limitations

Despite their promise, the GenAI-supported virtual patients reviewed exhibit several inherent limitations. Critically, behavioral and emotional complexity remains underdeveloped. While some avatars [42,44,46] incorporated basic movements or facial expressions, these were often simplistic and failed to fully replicate the nuanced nonverbal cues (eg, subtle pain indicators, authentic gaze patterns, and culturally specific gestures) essential for holistic clinical assessment and empathy training. For instance, Mool et al [46] noted that their 3D avatar was largely static within the examination room environment, with no description of physical movement. Emotional responsiveness was largely superficial, relying on prescripted ranges or simplistic vocalizations [38,42], rather than dynamically adapting emotional states based on learner interaction or physiological parameters. Furthermore, adaptability beyond dialog was constrained; virtual patients struggled to simulate evolving physical symptoms (eg, changing breath sounds and deteriorating vital signs) or accurately respond to physical examination maneuvers performed by learners within simulations. Memory and longitudinal consistency across interactions were absent, preventing virtual patients from recalling prior conversations or learner actions to build continuity. Finally, the underlying GenAI models (predominantly ChatGPT variants) introduced inherent biases and factual inaccuracies in medical content, alongside potential cultural insensitivities, raising concerns about the reliability and safety of the clinical scenarios portrayed. These limitations collectively restrict the virtual patients' ability to fully mirror the dynamism and unpredictability of real patient encounters.

In terms of limitations in research methodology, 13 of the 15 studies did not report the evaluation of the reliability or validity of their assessment instruments [33-43,45,46]. Further, 2 studies explicitly tested reliability [44,47]. Furthermore, 9 studies used designs without a control group and did not measure the same variables before and after the intervention [33,34,36-40,42,45,47], making it unclear whether the reported outcomes resulted from the intervention itself or differed across other conditions.

Discussion

Principal Findings

The current systematic review provides a comprehensive synthesis of the emerging evidence on GenAI-supported virtual patients in health care education, focusing on implementation characteristics, educational impact, and methodological

XSL•FO
RenderX

considerations. Building on the descriptive findings presented in the Results section, the following sections offer analytical comparisons across study designs, AI modalities, and educational purposes, and map identified limitations to established learning theories to guide future development.

Synthesis of Findings by Study Design, AI Modality, and Educational Purpose

When examining the evidence base by study design, a clear pattern emerges regarding the strength of causal inferences that can be drawn. The 3 RCTs provide the most robust evidence, demonstrating significant improvements in clinical decision-making [41], ophthalmology history-taking skills [45], and medical history-taking performance [47] attributable to GenAI-supported virtual patient interventions. In contrast, the cross-sectional and quasi-experimental studies, while consistently reporting positive user perceptions and self-reported skill gains [33,34,38,39,42,46], are more susceptible to enthusiasm bias and cannot establish definitive causal links between the intervention and learning outcomes. This methodological gradient underscores the need for more rigorous, controlled trials to move the field beyond proof-of-concept toward evidence-based practice.

Critically, the overall evidence base is characterized by small sample sizes, brief intervention durations (often single sessions under 10 minutes), and a near-complete absence of longitudinal follow-up. Consequently, while the studies demonstrate that GenAI-supported virtual patients are feasible and acceptable to learners, claims regarding their educational effectiveness must remain preliminary. The current findings support feasibility and acceptability more strongly than demonstrable learning gains or transfer to clinical practice.

Analysis by AI modality reveals differential alignment with educational objectives. Studies using embodied, voice-based virtual patients [36,38,44-46] predominantly targeted communication skills and emotional realism as primary outcomes. For example, Borg et al [36] found that a social robotic platform with 3D embodiment was rated significantly higher for authenticity and emotional skill training compared to a conventional computer-based platform, suggesting that physical presence and nonverbal cues may be particularly valuable for interpersonal skill development. In contrast, text-based virtual patient studies [33-35,37,39-41,43,47] more frequently targeted clinical reasoning, diagnostic accuracy, and history-taking performance. Yu et al [47] directly compared their text-based AI patient to human-simulated patients and found comparable or superior support for clinical reasoning, indicating that for cognitive skills such as information gathering and diagnostic thinking, sophisticated text-based interactions may be as effective as, or more effective than, human simulations. This modality-outcome alignment has practical implications for instructional design: educators should select or design GenAI virtual patient platforms based on the specific competencies they aim to develop.

When considering educational purpose, studies targeting communication skills [38,40,44,46] consistently reported improvements in learner empathy, interviewing technique, and patient interaction quality. For instance, Maquilón et al [38] found that medical first responders rated the virtual patient's voice interactions as natural for assessing physiological state, while Mool et al [46] observed that some students anthropomorphized the avatar, using the patient's name and inferring attitudes—behaviors indicative of authentic communication practice. Studies focused on clinical reasoning [36,41,47] demonstrated gains in diagnostic accuracy, information gathering, and decision-making, with Brügge et al [41] showing that feedback-enhanced interactions led to significantly higher clinical reasoning scores. This pattern suggests that GenAI-supported virtual patients can be effectively tailored to specific educational objectives, with modality and design features aligning with targeted learning outcomes.

Theoretical Implications of Design Limitations

From the perspective of experiential learning theory [48], the lack of longitudinal consistency and evolving patient states prevents learners from engaging in the full cycle of concrete experience, reflective observation, abstract conceptualization, and active experimentation across repeated encounters. Without memory across sessions, students cannot build upon prior interactions or observe the consequences of their clinical decisions over time—a core component of developing clinical expertise. This limitation suggests that future GenAI virtual patient designs should incorporate persistent patient states and longitudinal case progression to support complete experiential learning cycles.

Cognitive load theory [49,50] provides another lens for interpreting current limitations. The technical glitches, inconsistent responses, and need to troubleshoot AI interactions observed in several studies [38,43,46] impose extraneous cognitive load, diverting mental resources away from the germane load essential for schema construction and clinical reasoning development. Reducing technical unpredictability through more robust prompt engineering and model selection, as demonstrated by Yu et al [47], could minimize extraneous load and optimize cognitive resources for learning.

Furthermore, the absence of adaptive emotional and behavioral complexity limits opportunities for the sustained, feedback-driven practice necessary to refine complex interpersonal skills. Well-established principles of skill acquisition emphasize the importance of repeated engagement with authentic tasks, immediate feedback, and progressive challenge—elements that current GenAI virtual patients only partially provide. Incorporating dynamically adjusting emotional states based on learner interactions, informed by frameworks such as the Big Five personality model used by Yu et al [47], could better support the deliberate practice of communication and empathy.

Implications in Practice and Research

The implementation of GenAI-supported virtual patients in health care education carries significant practical implications. Educators and curriculum designers must carefully consider the technological infrastructure required to support diverse interaction modalities—text, voice, and hybrid formats—to cater to varied learning preferences and replicate realistic clinical scenarios. The integration of dynamic, AI-generated responses

XSL•FO
RenderX

demands ongoing technical oversight and periodic updates to ensure the content remains accurate and relevant. Furthermore, the relatively brief interaction sessions observed in this study suggest that more extended, immersive simulations are needed to mirror the complexity of real-world clinical practice. This calls for interdisciplinary collaborations between educators, clinicians, and technology developers to ensure that practical implementations not only leverage the technical advantages of GenAI but also align with educational objectives and clinical standards.

The finding that only 3 of the 15 included studies were explicitly grounded in established educational frameworks highlights an important area for future research. The limited adoption of frameworks indicates that many current approaches are primarily driven by technological innovation rather than pedagogical insight. By integrating robust theoretical perspectives, the design of GenAI-supported virtual patients can be enhanced to facilitate richer, more targeted learning experiences that promote the development of clinical reasoning and decision-making skills. Furthermore, theoretical models can help in formulating clear hypotheses about how these advanced simulations influence learner outcomes, guiding both the design of interventions and the evaluation of their effectiveness [51]. Encouraging further interdisciplinary studies that explicitly link technology-driven interventions with established learning theories will be essential in building a more comprehensive understanding of how these innovations can transform health care education.

For research design, only 5 of the 15 studies incorporated experimental designs with control groups, allowing for direct comparisons between traditional simulation methods and novel AI-driven interventions [35,36,41,45,47]. This experimental rigor has been crucial in establishing a baseline understanding of the potential advantages of GenAI-enhanced simulations, especially in relation to improving communication skills and clinical reasoning. The remaining 10 studies typically used within-group designs with mixed methods or solely quantitative approaches, providing insights into user perceptions and immediate learning outcomes [33,34,37-40,42-44,46]. As highlighted in the synthesis above, the current evidence base supports feasibility and acceptability more strongly than conclusive educational effectiveness. Future research must therefore prioritize larger, rigorous trials with long-term follow-up and objective measures of skill transfer to clinical practice.

To provide a more balanced perspective, the potential benefits of GenAI-supported virtual patients must be considered alongside their well-documented risks and ethical challenges. Our findings regarding positive learner perceptions and skill development exist within a context of significant technological limitations. Key concerns include algorithmic bias inherent in the training data of LLMs, which could reinforce stereotypes or lead to culturally insensitive patient portrayals, thereby misinforming clinical empathy and communication [52,53] Data privacy and security present another critical challenge, as sensitive health information used in prompt engineering or generated during simulated dialogs requires robust safeguards to comply with regulations such as HIPAA (Health Insurance Portability and Accountability Act) or GDPR (General Data Protection Regulation) [54,55]. Most critically, the tendency of GenAI models to generate plausible but incorrect or fabricated information—known as “hallucinations”—poses a direct threat to clinical accuracy [56]. In a health care education setting, an AI virtual patient providing factually wrong symptoms, pathophysiology, or treatment responses could seriously compromise foundational medical knowledge and patient safety. Therefore, the implementation of these tools necessitates rigorous validation frameworks, ongoing human oversight, and clear institutional guidelines to mitigate these risks, ensuring that innovation in simulation does not come at the cost of pedagogical integrity or ethical responsibility.

Integration of Methodological Quality

The risk of bias assessment conducted for this review underscores a critical consideration when interpreting its findings. While 3 RCTs demonstrated relatively strong methodological rigor (low to low-moderate risk) [57], most of the evidence derives from quasi-experimental and cross-sectional studies with moderate to high risk of bias [58]. This methodological landscape indicates that the current evidence base is still in a formative, proof-of-concept stage [59]. The prevalent limitations—such as small sample sizes, absence of control groups, lack of blinding, and reliance on unvalidated or self-reported outcomes—suggest that the positive results regarding user acceptance, perceived learning, and skill improvement should be viewed as promising preliminary signals rather than conclusive evidence of efficacy [60-63]. These design weaknesses increase the risk of overestimating positive effects due to confounding, measurement bias, or participant enthusiasm for novel technology [64,65]. Therefore, the synthesized findings, particularly those related to educational impact, must be interpreted with appropriate caution. This appraisal directly informs the primary recommendation of this review: future research must prioritize methodological robustness, including larger-scale randomized designs with active control groups [66], longitudinal follow-up [67], and the use of validated, objective outcome measures [65] to establish a more definitive evidence base for the educational effectiveness of GenAI-supported virtual patients.

Furthermore, our analysis identified variable reporting of ethical considerations—such as institutional review board approval and data security measures—across the included studies. While systematic reviews themselves do not require ethical approval, transparency in primary research is a cornerstone of methodological rigor and trustworthiness [68]. Moving forward, consistent and explicit ethical reporting should be considered a standard in this domain, especially when research involves simulated patient interactions and learner data, to ensure the credibility and safe translation of findings into educational practice [53,54].

Limitations

This review has several limitations that should be addressed when interpreting its findings. First, while our systematic search was comprehensive across several databases, the inclusion criteria limited the review to English-language studies, potentially excluding relevant research published in other languages or in alternative repositories. Second, many of the

XSL•FO
RenderX

primary studies included in this review are constrained by methodological limitations, such as small sample sizes, short simulation durations, and an overreliance on self-reported outcomes, which restrict the generalizability of the findings. Third, the heterogeneity in study designs and assessment tools across the interventions complicates direct comparisons and synthesis of outcomes. Furthermore, our search was limited to published, peer-reviewed literature and did not include gray literature sources such as clinical trial registries or preprint servers. While this decision aligns with standard systematic review practices, it is particularly consequential in a rapidly evolving field where early innovations often first appear as preprints or technical reports. Consequently, the review may not capture the most recent developments or emerging design approaches that have not yet undergone formal peer review. Future updates to this review should consider expanding the search to include gray literature as the evidence base matures. Lastly, the current review only included 15 papers; this limited number of studies may not capture the full spectrum of innovative practices and outcomes in this rapidly evolving field, thereby constraining the robustness of the conclusions drawn.

Conclusions

In summary, this review confirms that GenAI-supported virtual patients offer notable advances in adaptability and interactivity. This study is innovative as it constitutes the first dedicated synthesis of this specific technological application in health care education. It differs from prior reviews of virtual patients or GenAI by focusing exclusively on their intersection, thereby isolating the unique capabilities and challenges introduced by GenAI. The review brings to the field a foundational framework that classifies key design dimensions, evaluates educational impact, and identifies critical gaps, setting a clear agenda for subsequent research. The real-world implications are significant: for educators and technologists, it provides an evidence-based roadmap for developing more effective, theory-informed simulations; for institutions, it highlights the practical considerations and potential transformative value of integrating these tools to modernize clinical skills training and address scalability in health care education.

Acknowledgments

This study would not have been possible without the support of the Hong Kong Metropolitan University. The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT (Generative Artificial Intelligence Delegation Taxonomy; 2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing. The GenAI tool used was Grammarly (Superhuman Platform). Responsibility for this final paper lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the outcomes.

Funding

No funding was received for conducting this study.

Data Availability

Data is presented in Tables 2 and 3.

Authors' Contributions

Conceptualization: JJ, TTOK, JYHW
Formal analysis: MZY
Investigation: MZY
Methodology: JJ
Supervision: JYHW
Writing – original draft: JJ
Writing – review & editing: JJ, MZY, TTOK, JYHW
JJ and MZY are the cofirst authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Synthesis Without Meta-analysis (SWiM) reporting items.
[DOCX File, 20 KB - jmir_v28i1e82756_app1.docx]

Multimedia Appendix 2
Search strategy.
[DOCX File, 26 KB - jmir_v28i1e82756_app2.docx]

XSL•FO
RenderX

Checklist 1
PRISMA 2020 expanded checklist.
[DOCX File, 53 KB - jmir_v28i1e82756_app3.docx]

References

1. Kononowicz AA, Woodham LA, Edelbring S, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Jul 2;21(7):e14676. [doi: 10.2196/14676] [Medline: 31267981]
2. Cook DA, Erwin PJ, Triola MM. Computerized virtual patients in health professions education: a systematic review and meta-analysis. Acad Med 2010 Oct;85(10):1589-1602. [doi: 10.1097/ACM.0b013e3181edfe13] [Medline: 20703150]
3. Lioce L, Lopreiato J, Anderson M, et al. Healthcare simulation dictionary–third edition. : Agency for Healthcare Research and Quality; 2024 URL: https://www.ssih.org/sites/default/files/2025-03/Healthcare-Simulation-Dictionary-3.pdf [accessed 2026-04-21]
4. Dávidovics A, Dávidovics K, Hillebrand P, Rendeki S, Németh T. Virtual patient simulation to enhance medical students' clinical communication and decision-making skills: a pilot study. BMC Med Educ 2026;26(1):171. [doi: 10.1186/s12909-025-08507-7]
5. Furey P, Town A, Sumera K, Webster CA. Approaches for integrating generative artificial intelligence in emergency healthcare education within higher education: a scoping review. Crit Care Innov 2024;7(2):34-54 [FREE Full text]
6. Rodriguez DV, Lawrence K, Gonzalez J, et al. Leveraging generative AI tools to support the development of digital solutions in health care research: case study. JMIR Hum Factors 2023;11:e52885. [doi: 10.2196/52885]
7. Potter L, Jefferies C. Enhancing communication and clinical reasoning in medical education: building virtual patients with generative AI. Future Healthcare J 2024 Apr;11:100043. [doi: 10.1016/j.fhj.2024.100043]
8. Yaqoob A, Verma NK, Aziz RM. Optimizing gene selection and cancer classification with hybrid sine cosine and cuckoo search algorithm. J Med Syst 2024 Jan 9;48(1):10. [doi: 10.1007/s10916-023-02031-1] [Medline: 38193948]
9. Talbot TB, Sagae K, John B, Rizzo AA. Designing useful virtual standardized patient encounters. I/ITSEC Proceedings 2012:1-11 [FREE Full text]
10. Reed T, Pirotte M, McHugh M, et al. Simulation-based mastery learning improves medical student performance and retention of core clinical skills. Sim Healthcare 2016;11(3):173-180. [doi: 10.1097/SIH.0000000000000154]
11. Elendu C, Amaechi DC, Okatta AU, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore) 2024 Jul 5;103(27):e38813. [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
12. Padilha JM, Machado PP, Ribeiro A, Ramos J, Costa P. Clinical virtual simulation in nursing education: randomized controlled trial. J Med Internet Res 2019 Mar 18;21(3):e11529. [doi: 10.2196/11529] [Medline: 30882355]
13. Kubin L, Fogg N, Trinka M. Alternative clinical learning experiences for nursing education using virtual individual patients. Nurs Educ Perspect 2023;44(4):259-260. [doi: 10.1097/01.NEP.0000000000001066] [Medline: 36240018]
14. Williams R, Helmer B, Elliott A, Robinson D, Jimenez FA, Faragher ME. Navigating the virtual frontier: a virtual patient simulation pilot study in prelicensure baccalaureate nursing education. Clin Simul Nurs 2024 Sep;94:101589. [doi: 10.1016/j.ecns.2024.101589]
15. Hamilton A, Molzahn A, McLemore K. The evolution from standardized to virtual patients in medical education. Cureus 2024 Oct 10;16(10):e71224. [doi: 10.7759/cureus.71224] [Medline: 39525234]
16. He H, Xu X, Li S, Bueno-Vesga JA, Duan Y, Gu Y. Training nursing skills in a generative artificial intelligence-enhanced virtual reality patient encounter simulation: a qualitative study from a student perspective. Int J Educ Technol High Educ 2026;23(1):12. [doi: 10.1186/s41239-026-00587-9]
17. McCoy L, Pettit RK, Lewis JH, et al. Developing technology-enhanced active learning for medical education: challenges, solutions, and future directions. Acad Med 2015 Apr 1;115(4):202-211. [doi: 10.7556/jaoa.2015.042]
18. Andrade R. An exploration of virtual standardized patients and their effect on clinical readiness in pharmacy education. Curr Pharm Teach Learn 2026 May;18(5):102599. [doi: 10.1016/j.cptl.2026.102599] [Medline: 41713008]
19. Sengar SS, Hasan AB, Kumar S, Carroll F. Generative artificial intelligence: a systematic review and applications. Multimed Tools Appl 2025;84(21):23661-23700. [doi: 10.1007/s11042-024-20016-1]
20. Wang Y, Wang L, Siau KL. Human-centered interaction in virtual worlds: a new era of generative artificial intelligence and metaverse. Int J Hum-Comput Interact 2025 Jan 17;41(2):1459-1501. [doi: 10.1080/10447318.2024.2316376] [Medline: 38784821]
21. García-Torres D, Vicente Ripoll MA, Fernández Peris C, Mira Solves JJ. Enhancing clinical reasoning with virtual patients: a hybrid systematic review combining human reviewers and ChatGPT. Healthcare (Basel) 2024;12(22):2241. [doi: 10.3390/healthcare12222241]
22. Jay R, Sandars J, Patel R, et al. The use of virtual patients to provide feedback on clinical reasoning: a systematic review. Acad Med 2025 Feb 1;100(2):229-238. [doi: 10.1097/ACM.0000000000005908] [Medline: 39485118]
23. Neo NWS, Gunawan J, Levett-Jones T, Khoo ET, Chua WL, Liaw SY. Generative artificial intelligence in healthcare simulation-based education: a scoping review. Clin Simul Nurs 2025 Nov;108:101819. [doi: 10.1016/j.ecns.2025.101819]

XSL•FO
RenderX

24. Janssen E, McLagan R, Habeck J, Chung SY, McArthur EC, Anderson P. Barriers to breakthroughs: a scoping review of generative AI in healthcare simulation. Clin Simul Nurs 2025 Oct;107:101791. [doi: 10.1016/j.ecns.2025.101791] [Medline: 40339523]
25. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
26. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ 2020 Jan 16;368:l6890. [doi: 10.1136/bmj.l6890] [Medline: 31948937]
27. Systematic review of GenAI supported virtual patient in healthcare education. OSF. URL: https://osf.io/uxcwr/overview?view_only=9c97eb9c2df64afc97719b314a0af93f [accessed 2026-04-21]
28. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
29. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 1977 Jun;33(2):363-374. [doi: 10.2307/2529786] [Medline: 884196]
30. JBI critical appraisal tools. JBI. URL: https://jbi.global/critical-appraisal-tools [accessed 2026-04-21]
31. Barker TH, Stone JC, Sears K, et al. The revised JBI critical appraisal tool for the assessment of risk of bias for randomized controlled trials. JBI Evidence Synthesis 2023;21(3):494-506. [doi: 10.11124/JBIES-22-00430]
32. Barker TH, Habibi N, Aromataris E, et al. The revised JBI critical appraisal tool for the assessment of risk of bias for quasi-experimental studies. JBI Evidence Synthesis 2024;22(3):378-388. [doi: 10.11124/JBIES-23-00268]
33. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, et al. A language model–powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ 2024;10(1):e59213. [doi: 10.2196/59213]
34. Holderried F, Stegemann–Philipps C, Herschbach L, et al. A generative pretrained transformer (GPT)–powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ 2024;10(1):e53961. [doi: 10.2196/53961]
35. Yamamoto A, Koda M, Ogawa H, et al. Enhancing medical interview skills through AI-simulated patient interactions: nonrandomized controlled trial. JMIR Med Educ 2024 Sep 23;10(1):e58753. [doi: 10.2196/58753] [Medline: 39312284]
36. Borg A, Georg C, Jobs B, et al. Virtual patient simulations using social robotics combined with large language models for clinical reasoning training in medical education: mixed methods study. J Med Internet Res 2025;27:e63312. [doi: 10.2196/63312]
37. Herschbach L, Festl-Wietek T, Stegemann-Philipps C, et al. Evaluation of an AI-based chatbot providing real-time feedback in communication training for mental health care professionals: proof-of-concept observational study. J Med Internet Res 2025 Nov 28;27:e82818. [doi: 10.2196/82818] [Medline: 41314643]
38. Maquilón RG, Uhl J, Schrom-Feiertag H, Tscheligi M. Integrating GPT-based AI into virtual patients to facilitate communication training among medical first responders: usability study of mixed reality simulation. JMIR Form Res 2024 Dec 11;8:e58623. [doi: 10.2196/58623] [Medline: 39661979]
39. Benfatah M, Marfak A, Saad E, Hilali A, Nejjari C, Youlyouz-Marfak I. Assessing the efficacy of ChatGPT as a virtual patient in nursing simulation training: a study on nursing students' experience. Teach Learn Nurs 2024 Jul;19(3):e486-e493. [doi: 10.1016/j.teln.2024.02.005]
40. Bonfitto GR, Roletto A, Savardi M, Fasulo SV, Catania D, Signoroni A. Harnessing ChatGPT dialogues to address claustrophobia in MRI - a radiographers' education perspective. Radiography (Lond) 2024 May;30(3):737-744. [doi: 10.1016/j.radi.2024.02.015]
41. Brügge E, Ricchizzi S, Arenbeck M, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. BMC Med Educ 2024 Nov 28;24(1):1391. [doi: 10.1186/s12909-024-06399-7] [Medline: 39609823]
42. Sardesai N, Russo P, Martin J, Sardesai A. Utilizing generative conversational artificial intelligence to create simulated patient encounters: a pilot study for anaesthesia training. Postgrad Med J 2024 Mar 18;100(1182):237-241. [doi: 10.1093/postmj/qgad137] [Medline: 38240054]
43. Chinwong D, Penthinapong T, Chinwong S. Integrating ChatGPT for smoking cessation counseling practice in pharmacy education: a single group quasi-experimental study. Tob Induc Dis 2025;23(November):1-12. [doi: 10.18332/tid/211706]
44. Kim J, Won J, Lee Y. Use of a generative pre-trained transformer-based virtual patient for health assessment and communication training in nursing education: a mixed-methods study. Nurse Educ Pract 2025 Oct;88:104536. [doi: 10.1016/j.nepr.2025.104536]
45. Luo MJ, Bi S, Pang J, et al. A large language model digital patient system enhances ophthalmology history taking skills. npj Digit Med 2025;8(1):502. [doi: 10.1038/s41746-025-01841-6]
46. Mool A, Schmid J, Johnston T, et al. Using generative AI to simulate patient history-taking in a problem-based learning tutorial: a mixed-methods study. Tech Know Learn 2026:1-18. [doi: 10.1007/s10758-025-09929-4]
47. Yu H, Zhou J, Li L, et al. Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. Commun Med (Lond) 2025 Dec 19;6(1):27. [doi: 10.1038/s43856-025-01283-x] [Medline: 41420084]

XSL•FO
RenderX

48. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development: FT Press; 2014.
49. Sweller J. Cognitive load theory. In: Psychology of Learning and Motivation: Academic Press; 2011, Vol. 55:37-76. [doi: 10.1016/B978-0-12-387691-1.00002-8]
50. Fraser KL, Ayres P, Sweller J. Cognitive load theory for the design of medical simulations. Simul Healthcare 2015;10(5):295-307. [doi: 10.1097/SIH.0000000000000097]
51. Thomas G. What's the use of theory? Harvard Educ Rev 1997 Jan 1;67(1):75-105. [doi: 10.17763/haer.67.1.1x807532771w5u48]
52. Joseph J. Algorithmic bias in public health AI: a silent threat to equity in low-resource settings. Front Public Health 2025;13:1643180. [doi: 10.3389/fpubh.2025.1643180] [Medline: 40771228]
53. Panch T, Mattie H, Celi LA. The "inconvenient truth" about AI in healthcare. npj Digit Med 2019;2(1):77. [doi: 10.1038/s41746-019-0155-4] [Medline: 31453372]
54. Al-kfairy M, Mustafa D, Kshetri N, Insiew M, Alfandi O. Ethical challenges and solutions of generative AI: an interdisciplinary perspective. Informatics (MDPI) 2024;11(3):58. [doi: 10.3390/informatics11030058]
55. Duffourc MN, Gerke S, Kollnig K. Privacy of personal data in the generative AI data lifecycle. NYU J Intell Prop Ent L 2024;13(2):219-268 [FREE Full text]
56. Tran C, Hryciw BN, Moore SW, Chaput A, Seely AJE. Perceptions and use of generative artificial intelligence in medical students: a multicenter survey. J Med Educ Curric Dev 2025;12:23821205251391969. [doi: 10.1177/23821205251391969] [Medline: 41181167]
57. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
58. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016 Oct 12;355:i4919. [doi: 10.1136/bmj.i4919] [Medline: 27733354]
59. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-926. [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
60. Ioannidis JPA. Why most published research findings are false. PLoS Med 2005 Aug;2(8):e124. [doi: 10.1371/journal.pmed.0020124] [Medline: 16060722]
61. Button KS, Ioannidis JPA, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 2013 May;14(5):365-376. [doi: 10.1038/nrn3475]
62. Hróbjartsson A, Gøtzsche PC. Placebo interventions for all clinical conditions. Cochrane Database Syst Rev 2010 Jan 20;2010(1):CD003974. [doi: 10.1002/14651858.CD003974.pub3] [Medline: 20091554]
63. Rosenman R, Tennekoon V, Hill LG. Measuring bias in self-reported data. Int J Behav Healthc Res 2011 Oct;2(4):320-332. [doi: 10.1504/IJBHR.2011.043414] [Medline: 25383095]
64. Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. Lancet 2005 Apr;365(9467):1348-1353. [doi: 10.1016/S0140-6736(05)61034-3]
65. DeVellis RF, Thorpe CT. Scale Development: Theory and Applications, 5th edition: SAGE Publications, Inc; 2021.
66. Campbell DT, Stanley JC. Experimental and quasi-experimental designs for research. : Houghton Mifflin Company; 1963 URL: https://www.sfu.ca/~palys/Campbell&Stanley-1959-Exptl&QuasiExptlDesignsForResearch.pdf [accessed 2026-04-21]
67. Cook TD, Campbell DT. Quasi-Experimentation: Design and Analysis Issues for Field Settings: Houghton Mifflin Company; 1979.
68. Resnik DB. What is ethics in research & why is it important? National Institute of Environmental Health Sciences. 2020 Dec 23. URL: https://www.niehs.nih.gov/research/resources/bioethics/whatis [accessed 2026-04-21]

Abbreviations

AI: artificial intelligence
GDPR: General Data Protection Regulation
GenAI: generative artificial intelligence
GRADE: Grading of Recommendations, Assessment, Development, and Evaluation
HIPAA: Health Insurance Portability and Accountability Act
JBI: Joanna Briggs Institute
LLM: large language model
MeSH: Medical Subject Headings
MIMIC-III: Medical Information Mart for Intensive Care III
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
RCT: randomized controlled trial
RQ: research question

XSL•FO
RenderX

Edited by S Brini; submitted 21.Aug.2025; peer-reviewed by D Cunha, S Funghetto; revised version received 27.Mar.2026; accepted 31.Mar.2026; published 07.May.2026.

Please cite as:
Jiang J, Ye MZ, Kwok TTO, Wong JYH
GenAI-Supported Virtual Patients in Health Care Education: Systematic Review
J Med Internet Res 2026;28:e82756
URL: https://www.jmir.org/2026/1/e82756
doi:10.2196/82756

© Juming Jiang, Megan Zichen Ye, Tyrone Tai-On Kwok, Janet Yuen Ha Wong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Applications of Natural Language Processing and Large Language Models for Social Determinants of Health: Systematic Review

Swati Rajwal[1], MS; Avinash Kumar Pandey[2], MS; Ziyuan Zhang[3], MS; Yankai Chen[4], PhD; Michael X Liu[4], MBA; Sudeshna Das[1], PhD; Hannah Rogers[5], MS; Abeed Sarker[1], PhD; Yunyu Xiao[4], PhD

[1]Department of Biomedical Informatics, School of Medicine, Emory University, 101 Woodruff Circle, Atlanta, GA, United States

[2]Laney Graduate School, Emory University, Atlanta, GA, United States

[3]Department of Epidemiology, University of California, Berkeley, Berkeley, CA, United States

[4]Department of Population Health Sciences, Cornell University, New York, NY, United States

[5]Woodruff Health Sciences Center Library, Emory University, Atlanta, GA, United States

Corresponding Author:
Swati Rajwal, MS
Department of Biomedical Informatics, School of Medicine, Emory University, 101 Woodruff Circle, Atlanta, GA, United States

Abstract

Background: Social determinants of health (SDOH) are the social, economic, and environmental conditions that influence health outcomes. SDOH information is often embedded in unstructured text, such as notes in electronic health records and social media posts. Advances in natural language processing (NLP), including emergent large language models (LLMs), offer opportunities to extract, analyze, and interpret SDOH expressions from free text for inclusion in downstream analyses. Existing literature on NLP applications for SDOH is dispersed across disciplines and characterized by methodological heterogeneity and variability in study quality and scope, complicating synthesis and cross-study comparison.

Objective: This study aimed to examine the use of NLP, including LLMs, in SDOH research, and highlight gaps and future research directions.

Methods: We conducted a systematic review following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, searching 7 major databases for publications between 2014 and November 2025. We included journal and conference proceedings papers that applied NLP methods to identify, classify, extract, or predict SDOH from text. Three reviewers independently screened studies and extracted data; conflicts were resolved by two senior reviewers. We abstracted study metadata, dataset characteristics, NLP approaches, SDOH domains addressed, and NLP performance metrics. We also conducted risk-of-bias analyses and identified influential studies based on relative citation counts.

Results: 142 studies met the inclusion criteria. Nearly two-thirds (89/142, 62.7%) were published between 2023 and 2025, reflecting rapid recent growth. Most studies relied on electronic health records (93/142, 65.5%) and private datasets (81/142, 57.0%), while only 20.4% (29/142) used publicly available data. Commonly studied SDOH domains were housing instability (72/142, 50.7%), employment (65/142, 45.8%), and financial conditions (63/142, 44.4%); structural factors, such as immigration status (5/142, 3.5%), were rarely examined. Of studies that reported evaluation metrics, most focused on classification (26/83, 31.32%) or extraction (38/83, 45.7%), and used cross-sectional designs. Reported model performances were typically strong, with median F_1-scores ranging roughly from 0.75 to 0.85 across model categories. Only 49 studies shared code, and fewer than half clearly described model interpretability or reproducibility practices. LLMs (including encoder-decoder models) appeared in 19.7% (28/142) of studies, highlighting emerging interest but also raising new concerns around transparency and governance.

Conclusions: This review provides a timely synthesis of NLP and LLM applications across the SDOH research spectrum, addressing an important gap in a topic receiving increasing research attention. By comparing task formulations, data sources, and performance patterns, the review clarifies the research readiness of current approaches and reveals critical gaps. Our findings advance the field by highlighting the absence of a unified SDOH framework, uneven availability of public benchmarks, and limited evaluation of real-world deployment. Addressing these gaps through transparent, inclusive dataset development and implementation-focused evaluation is essential for translating NLP advances into equitable, real-world health impact.

Trial Registration: PROSPERO CRD42024578082; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024578082

International Registered Report Identifier (IRRID): RR2-10.2196/66094

(J Med Internet Res 2026;28:e83793) doi:10.2196/83793

XSL•FO
RenderX

KEYWORDS

social determinants of health; SDOH; natural language processing; large language models; systematic review; PRISMA

Introduction

Social determinants of health (SDOH) refer to nonmedical factors such as the social, economic, and environmental conditions that shape where people live, work, and age. SDOH are among the most powerful drivers of health outcomes and disparities worldwide [1,2], with influences on health [3] and well-being at the individual and population levels [4]. Contemporary estimates suggest that medical care accounts for only 10% - 20% of the modifiable contributors to healthy outcomes, while SDOH-related factors drive the remaining 80% - 90% [3,5,6]. Consequently, SDOH factors, such as housing instability, food insecurity, and structural racism, have become central to understanding the persistence of disease, quality of life, and mortality across diverse populations [7,8]. Accurately capturing SDOH and incorporating them into health care is crucial for clinicians, health systems, and policymakers aiming to address structural inequities and enhance care delivery. Clinicians, for example, can work with patients in social prescribing based on affordability or the availability of transportation to the relevant pharmacy. Similarly, with an in-depth understanding of structural SDOH, including place-based contextual metrics of economic, educational, health, and environmental conditions, policymakers may effectively guide health policies [9-11]. While health systems increasingly recognize that addressing social needs is essential for value-based care, the mechanisms to systematically identify and act upon these factors are often underdeveloped [12].

Several frameworks have been proposed to characterize SDOH into groups. The Healthy People 2030 initiative [2] groups SDOH into 5 domains: Economic Stability, Education Access and Quality, Health Care Access and Quality, Neighborhood and Built Environment, and Social and Community Context. The World Health Organization conceptualizes SDOH into 2 interacting groups: Structural and Intermediary, with the former having causal priority in influencing health outcomes. Due to their importance in influencing health, SDOH codes were also introduced into the International Classification of Diseases, 10th Revision, Clinical Modification in 2015. These are categorized as Z-codes (Z55-Z65) and are used to document socioeconomic and psychosocial circumstances that affect a person's health. Yet, the documentation of Z-codes has been low in health systems [13]. Thus, SDOH information is often buried in unstructured text, such as notes in electronic health records (EHRs) and social media postings, where it may be expressed through nuanced or implicit language, limiting information accessibility through conventional methods and requiring the development of customized text mining approaches [14].

Natural language processing (NLP) is a subfield of artificial intelligence and computer science that enables computers to process, understand, interpret, and generate human language in meaningful and useful ways. NLP encompasses a wide range of tasks, including but not limited to text classification, information extraction/entity recognition, text summarization, and language translation. NLP offers a promising tool to systematically analyze vast amounts of unstructured text data, EHRs, social media [15,16], public health reports, and other sources [17]. While biomedical domain NLP has been widely applied to extract clinical concepts such as diagnoses, medications, and procedures, much less work has focused on nonclinical aspects, such as SDOH and their influence on health outcomes.

In the context of SDOH, NLP can assist in parsing and understanding context, co-references, and the relationships between text parts. Unlike manual review (time-consuming, labor-intensive, and prone to human error), NLP enables the rapid and scalable analysis of large datasets with greater accuracy and consistency [18,19] when models are appropriately adapted to the domain and data, although performance may vary across populations and SDOH categories. A common NLP application is the identification and extraction of SDOH factors from EHRs and other text-based data, enabling the systematic classification of social needs, behavioral drivers, and environmental conditions that influence patient outcomes [20-24]. The versatility of NLP is evident in its application across diverse clinical domains, ranging from pediatric populations [25,26] to patients managing chronic conditions such as Alzheimer disease [22] and lower back pain [27]. Beyond traditional medical notes, NLP workflows have been adapted to mine insights from specialized documentation, including clinical social work notes [28] and emergency medical services records [29], as well as nonclinical data sources like social media, where it has been used to assess the impact of external crises (eg, COVID-19) on marginalized communities [30]. Underpinning these diverse applications is the rapid evolution of model architectures; specifically, the shift toward transformer-based models and generative pretrained transformers has significantly enhanced the precision of SDOH extraction from free-text data [31,32].

Despite these technological advancements, the literature remains fragmented in the application of NLP to characterize SDOH, due to disparate disciplines, data modalities, and methodological frameworks. A comprehensive NLP workflow for SDOH analysis necessitates a rigorous pipeline: (1) defining the target SDOH elements or categories; (2) selecting the appropriate NLP modeling strategy (eg, classification vs extraction); (3) curating gold-standard annotated datasets for supervision and validation; and (4) deploying the optimal strategy. Currently, no comprehensive synthesis exists that integrates these myriad applications of NLP and large language models (LLMs) for SDOH. Consequently, there is a critical knowledge gap regarding the most effective approaches to modeling SDOH, limiting the translation of these technical capabilities into standardized public health and clinical practice.

To address this gap, we conducted a systematic review of peer-reviewed studies published between 2014 and 2025. This period is marked by the introduction of Word2Vec [33], transformer models [34] such as BERT [35], and generative models like the generative pretrained transformer series [36].

XSL•FO
RenderX

Our review consolidates evidence from health, informatics, and computer science literature to assess the state of the science in applying NLP to SDOH. We broadly categorize SDOH into individual and structural factors, a simplification of more comprehensive frameworks such as Healthy People 2030 and the World Health Organization Commission on SDOH. Individual-level SDOH refer to personal circumstances such as income, education, employment status, housing conditions, and access to health care, which directly influence an individual's health outcomes. Structural SDOH encompass the broader systemic and institutional contexts (such as policies, social norms, economic systems, and structural racism) that shape and constrain the distribution of individual-level resources and opportunities [10]. This systematic review has two major objectives: (1) to characterize NLP techniques, including LLMs, used to analyze SDOH in unstructured individual and public data, including annotation, prediction, detection, and classification; and (2) to assess the effectiveness of such techniques or models, identify potential knowledge gaps, and reveal research questions relevant for future studies.

Methods

Registration and Protocol

The systematic review was registered in PROSPERO (CRD42024578082), and the protocol was published in JMIR Research Protocol [37]. The review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [38], and the search strategy was reported in accordance with PRISMA-S [39]. We conducted a narrative synthesis without meta-analysis following the Synthesis Without Meta-analysis guidelines [40] due to substantial methodological heterogeneity across studies. Formal subgroup, heterogeneity, sensitivity analyses, reporting bias analysis, or certainty of evidence assessment were not conducted as the review did not involve meta-analysis and instead examined differences across studies descriptively. Synthesis focused on identifying patterns and trends in NLP methods, SDOH domains, data sources, and reported performance rather than pooling quantitative estimates. The interdisciplinary review team consisted of a librarian and researchers with expertise in health informatics, public health, and social sciences.

Ethical Considerations

This work does not involve human participants and does not require approval from the IRB.

Eligibility Criteria

We included peer-reviewed literature (journal articles and full conference papers) published between January 2014 and November 2025. The selection of this timeframe was deliberate, as it strategically captures the entire modern deep learning era in NLP, from the rise of word embeddings (eg, Word2Vec) around 2014 to the development of the transformer architecture in 2017 and the subsequent explosion of LLMs. Only studies in English were included, although the underlying NLP data could be in any language. Eligible studies met the following specific criteria: (1) addressed a research question involving the design, development, and application of NLP in health data analysis for SDOH; (2) used text-based data; and (3) used NLP techniques or LLMs (open-source or commercial) related to SDOH. Studies using NLP in conjunction with other methods were eligible, provided NLP was a core component. Preprints (arXiv/bioRxiv), forewords, prefaces, table of contents, programs, schedules, indexes, call for papers or participation, lists of reviewers, lists of tutorial abstracts, invited talks, appendices, session information, obituaries, book reviews, newsletters, lists of proceedings, lifetime achievement awards, erratum, systematic reviews, scoping reviews, and notes were excluded.

Database and Search Strategy

We systematically searched 7 databases: PubMed, Scopus, Web of Science, PsycINFO, Health Source: Nursing/Academic, ACL Anthology, and IEEE Xplore. Searches were conducted independently in each database rather than through a multi-database platform. No study registries, manual browsing of web-based resources, citation chaining, or direct contact with authors or experts were undertaken as part of the search methodology. The search strategies were developed specifically for this review and were not adapted from previous reviews. No additional information sources or search methods were used.

The search spanned publications from January 1, 2014, to November 2, 2025. A health sciences librarian (HR) developed and iteratively refined the search strategy in consultation with the research team, using both controlled vocabulary and free-text terms. The final PubMed search strategy, including filters, is detailed in Textbox 1. Comprehensive search strategies for all information sources are provided in Multimedia Appendix 1.

XSL•FO
RenderX

Textbox 1. Search query for PubMed. A comprehensive list of all search queries customized for each database is available in Multimedia Appendix 1.

Query:

- (“Natural Language Processing”[Mesh] OR “natural language”[tw] OR NLP[tw] OR “large LM*”[tw] OR LLM[tw] OR LLMs[tw] OR “large language model*”[tw] or ChatGPT*[tw] OR “Chat GPT*”[tw] OR GPT4*[tw] OR GPT-4*[tw] OR GPT3*[tw] OR GPT-3*[tw] OR “Generative Pre-trained Transformer*”[tw] OR LLAMA[tw] OR “Claude 3”[tw] OR Mistral[tw] OR MedPaLM*[tw] OR Med-PaLM*[tw] OR “text mining”[tw] OR “text process*”[tw] OR “information retrieval”[tw] OR “information extract*”[tw]) AND (“Social Determinants of Health”[Mesh] OR SDOH[tw] OR SDH[tw] OR SBDH*[tw] OR “determinants of health”[tw] OR “health determina*”[tw] OR “life events”[tw] OR “social determinant*”[tw] OR “socioeconomic determinant*”[tw] OR “socioeconomic factor*”[tw] OR “social determinate*”[tw] OR “social factor*”[tw] OR “social need*”[tw] OR “social prescribing”[tw] OR “social determining factor*”[tw] OR “social risk*”[tw])

Filters:

- Language: English
- Years: 2014 - 2025
- Exclude: Preprints

Selection Process

Three reviewers (SR, ZZ, and YC) independently screened each study for eligibility by marking it as a “yes” (for inclusion), “no” (for exclusion), or “maybe” (in case of uncertainty about relevance) on the Covidence platform [41]. Two senior reviewers (AS and YX) resolved potential discrepancies during any screening step. Blinded voting ensured that reviewers did not view others’ decisions during the screening process. The reviewers retrieved eligible studies for second-stage review (full-text) using the same inclusion criteria and removed those that did not meet them. The final set of studies to include was approved by consensus of all reviewers.

Data Extraction

Before formal data extraction, 3 reviewers (SR, ZZ, and YC) piloted a structured data extraction form using 5 sample studies to ensure clarity and consistency. Six independent reviewers (SR, ZZ, YC, AKP, ML, and SD) then conducted the full data extraction in teams of 2, with each member independently extracting half of the assigned studies. Ambiguities were resolved within the teams to maintain accuracy and consistency. The final data captured study metadata (eg, year of publication and type), dataset characteristics (eg, source, sample size, and type), NLP approaches (eg, models used and task type), SDOH domains addressed, and performance metrics (eg, precision, F_1-score, and recall). We treated these performance metrics as effect measures for outcomes of interest (ie, model performance). Due to heterogeneity in study design, NLP tasks, datasets, and evaluation frameworks, no single standardized effect size metric was applicable, and reported metrics were extracted and summarized descriptively without statistical transformation. Model performance metrics reported by individual studies were extracted as reported and treated descriptively. We summarized study characteristics in structured tables and used visualizations, including a PRISMA flow diagram, frequency plots, heatmaps, and bubble charts, to display study selection, SDOH domains, and NLP model performance patterns. We assessed risk of bias using a short checklist adapted from the Joanna Briggs Institute [42] and covered 6 items: reporting of population demographics (Q1), clarity of study aim (Q2), relevance to SDOH and NLP (Q3), use of a reference standard (Q4), reporting of evaluation measures (Q5), and acknowledgment of study limits (Q6). Each item was rated as yes, no, or unclear. Studies were classified as low (≥5 yes), moderate (3 - 4 yes), or high (≤3 yes) risk of bias.

Influential Studies

To assess visibility and identify recurring methodological, collaborative, and reporting features, we calculated a normalized citation ratio (NCR) for each study:

$$NCR = \frac{\text{Paper`s citations}}{\text{Average of its publication year`s cohort}}$$

This approach adjusts for citation differences based on publication year.

Results

Study Selection and Characteristics

Figure 1 shows the number of articles included in each phase of the screening process, resulting in the eventual inclusion of 142 studies (Multimedia Appendix 1 shows the search strategy results across databases). Figure 2 presents publication trends over the years, illustrating a growing research interest over time, with the majority of studies being published between 2023 and 2025 (89/142, 62.7%). The majority of the studies were published in journals (109/142, 76.7%), while the remaining were in conference or workshop proceedings (Figure 2C). Also, as shown in Figure 2B, single-site studies were more common overall, but there has been a recent rise in multi-site studies (data from multiple institutes). Most studies were rated as low risk of bias (116/142), 26 as moderate, and none as high risk (Table 1). Table S5 in Multimedia Appendix 1 presents a summary of studies. Also, because this review used descriptive narrative synthesis across highly heterogeneous studies, formal heterogeneity analyses, sensitivity analyses, reporting-bias assessment, and certainty-of-evidence grading were not conducted.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram outlining the inclusion and screening processes for study selection. ACL: Association for Computational Linguistics; NLP: natural language processing; SDOH: social determinants of health.

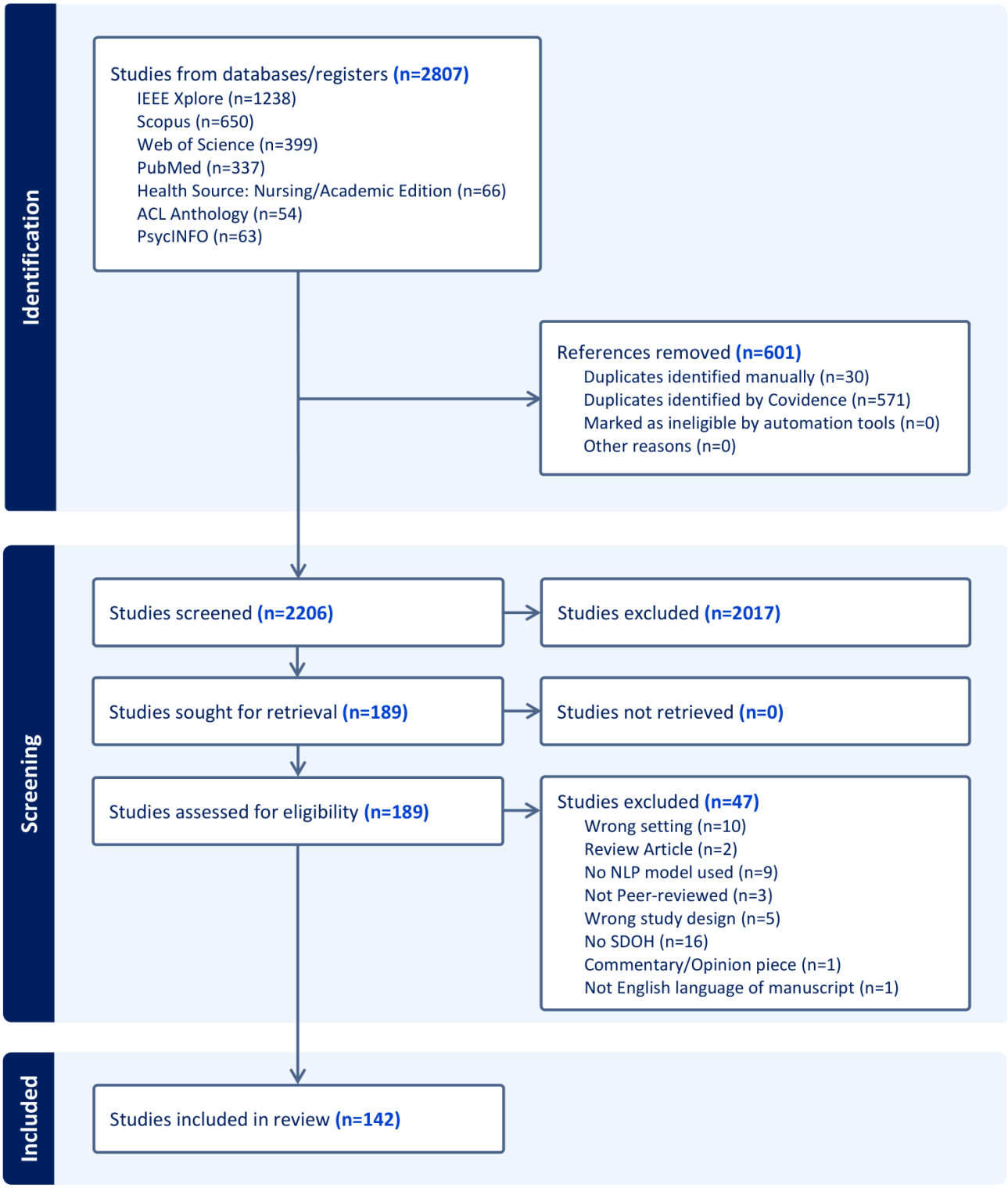

XSL•FO
RenderX

Figure 2. Yearly publication trends in natural language processing for SDOH research illustrate increasing interest over time. (A) SDOH type distribution. (B) Single- versus multi-institution dataset usage. (C) Publication venue distribution. Multi-institution datasets and journal publications increased substantially after 2022. Note that only partial data was available for 2025 at the time of review. SDOH: social determinants of health.

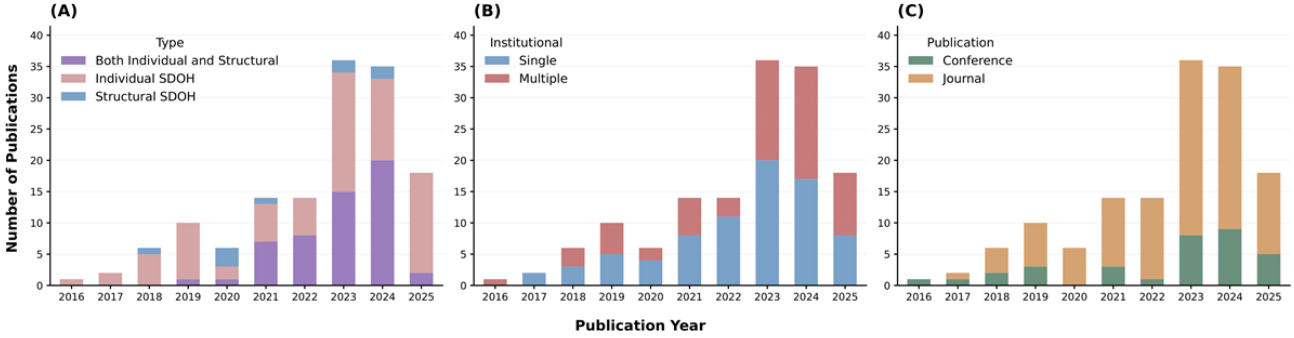

XSL•FO
RenderX

Table . Risk of bias assessment for included studies.

Reference	Q1	Q2	Q3	Q4	Q5	Q6	Total "Y"	Risk category
[43]	✓[a]	✓	✓	✓	✓	✓	6	↓[b]
[44]	✓	✓	✓	✓	•[c]	✓	5	↓
[45]	✓	✓	✓	✓	•	✓	5	↓
[46]	✓	✓	✓	✓	•	✓	5	↓
[47]	✓	✓	✓	✓	•	✓	5	↓
[48]	✓	✓	✓	×[d]	×	✓	4	≈[e]
[49]	✓	✓	✓	✓	•	✓	5	↓
[50]	×	✓	✓	✓	•	✓	4	≈
[51]	✓	✓	✓	×	•	✓	4	≈
[26]	×	✓	✓	✓	✓	✓	5	↓
[28]	✓	✓	✓	×	×	✓	4	≈
[52]	✓	✓	✓	✓	✓	✓	6	↓
[53]	×	✓	✓	✓	✓	✓	5	↓
[54]	×	✓	✓	×	•	✓	3	≈
[55]	×	✓	✓	✓	✓	✓	5	↓
[56]	✓	✓	✓	✓	✓	✓	6	↓
[57]	✓	✓	✓	✓	✓	✓	6	↓
[30]	✓	✓	✓	✓	✓	✓	6	↓
[58]	×	✓	✓	✓	✓	✓	5	↓
[59]	×	✓	✓	✓	•	✓	4	≈
[22]	×	✓	✓	✓	✓	✓	5	↓
[60]	✓	✓	✓	✓	•	✓	5	↓
[61]	✓	✓	✓	✓	✓	✓	6	↓
[62]	✓	✓	✓	✓	✓	✓	6	↓
[63]	✓	✓	✓	✓	•	✓	5	↓
[23]	✓	✓	✓	✓	✓	✓	6	↓
[64]	✓	✓	✓	✓	•	✓	5	↓
[65]	✓	✓	✓	✓	•	✓	5	↓
[66]	✓	✓	✓	✓	•	✓	5	↓
[67]	✓	✓	✓	✓	•	✓	5	↓
[29]	✓	✓	✓	✓	•	✓	5	↓
[68]	✓	✓	✓	✓	✓	✓	6	↓
[69]	✓	✓	✓	✓	✓	✓	6	↓
[70]	×	✓	✓	✓	•	✓	4	≈
[31]	×	✓	✓	✓	•	✓	4	≈
[71]	✓	✓	✓	✓	✓	✓	6	↓
[72]	✓	✓	✓	✓	•	✓	5	↓
[73]	✓	✓	✓	✓	•	✓	5	↓
[74]	×	✓	✓	✓	✓	✓	5	↓
[75]	✓	✓	✓	✓	×	✓	5	↓
[76]	×	✓	✓	✓	✓	✓	5	↓

XSL•FO
RenderX

Reference	Q1	Q2	Q3	Q4	Q5	Q6	Total "Y"	Risk category
[77]	×	✓	✓	✓	✓	✓	5	↓
[78]	×	✓	✓	✓	✓	✓	5	↓
[79]	✓	✓	✓	✓	•	✓	5	↓
[80]	×	✓	✓	✓	✓	✓	5	↓
[81]	×	✓	✓	✓	✓	✓	5	↓
[82]	✓	✓	✓	✓	✓	✓	6	↓
[83]	✓	✓	✓	✓	•	✓	5	↓
[32]	×	✓	✓	✓	✓	✓	5	↓
[84]	×	✓	✓	✓	✓	✓	5	↓
[85]	×	✓	✓	✓	✓	✓	5	↓
[86]	×	✓	✓	✓	✓	✓	5	↓
[87]	×	✓	✓	✓	✓	✓	5	↓
[88]	✓	✓	✓	×	•	✓	4	≈
[27]	×	✓	✓	✓	✓	✓	5	↓
[89]	✓	✓	✓	✓	•	✓	5	↓
[25]	✓	✓	✓	✓	•	✓	5	↓
[90]	×	✓	✓	✓	✓	✓	5	↓
[91]	✓	✓	✓	✓	•	✓	5	↓
[92]	×	✓	✓	✓	✓	✓	5	↓
[93]	✓	✓	✓	✓	✓	✓	6	↓
[94]	✓	✓	✓	✓	✓	✓	6	↓
[95]	✓	✓	✓	✓	×	✓	5	↓
[96]	✓	✓	✓	✓	×	✓	5	↓
[97]	✓	✓	✓	✓	•	✓	5	↓
[98]	✓	✓	✓	✓	✓	✓	6	↓
[99]	×	✓	✓	✓	✓	✓	5	↓
[100]	×	✓	✓	✓	✓	✓	5	↓
[101]	×	✓	✓	✓	✓	✓	5	↓
[102]	✓	✓	✓	✓	✓	✓	6	↓
[103]	✓	✓	✓	×	×	✓	4	≈
[104]	×	✓	✓	✓	•	✓	4	≈
[20]	✓	✓	✓	✓	•	✓	5	↓
[105]	✓	✓	✓	✓	×	✓	5	↓
[106]	✓	✓	✓	✓	✓	✓	6	↓
[21]	✓	✓	✓	✓	•	✓	5	↓
[107]	✓	✓	✓	✓	•	✓	5	↓
[108]	✓	✓	✓	✓	✓	✓	6	↓
[109]	×	✓	✓	✓	×	✓	4	≈
[110]	×	✓	✓	✓	•	✓	4	≈
[111]	×	✓	✓	✓	✓	✓	5	↓
[112]	×	✓	✓	×	×	✓	3	≈
[113]	×	✓	✓		×	✓	3	≈
[114]	✓	✓	✓	✓	✓	✓	6	↓

XSL•FO
RenderX

Reference	Q1	Q2	Q3	Q4	Q5	Q6	Total "Y"	Risk category
[115]	×	✓	✓	✓	•	✓	4	≈
[116]	×	✓	✓	✓	✓	✓	5	↓
[117]	✓	✓	✓	✓	•	✓	5	↓
[118]	×	✓	✓	✓	✓	✓	5	↓
[12]	×	✓	✓	✓	✓	✓	5	↓
[119]	✓	✓	✓	✓	✓	✓	6	↓
[120]	×	✓	✓	✓	•	✓	4	≈
[121]	✓	✓	✓	✓	•	✓	5	↓
[122]	×	✓	✓	✓	✓	✓	5	↓
[123]	×	✓	✓	✓	✓	✓	5	↓
[124]	×	✓	✓	✓	×	✓	4	≈
[125]	×	✓	✓	✓	✓	✓	5	↓
[126]	×	✓	✓	✓	✓	✓	5	↓
[127]	×	✓	✓	✓	•	✓	4	≈
[128]	×	✓	✓	✓	✓	✓	5	↓
[129]	×	✓	✓	×	•	✓	3	≈
[130]	✓	✓	✓	✓	✓	✓	6	↓
[131]	×	✓	✓	✓	•	✓	4	≈
[132]	✓	✓	✓	✓	✓	✓	6	↓
[133]	✓	✓	✓	✓	•	✓	5	↓
[134]	✓	✓	✓	✓	✓	✓	6	↓
[135]	×	✓	✓	✓	✓	✓	5	↓
[136]	✓	✓	✓	✓	•	✓	5	↓
[137]	×	✓	✓	✓	✓	✓	5	↓
[138]	×	✓	✓	✓	•	✓	4	≈
[139]	×	✓	✓	✓	✓	✓	5	↓
[140]	×	✓	✓	✓	•	✓	4	≈
[141]	✓	✓	✓	✓	✓	✓	6	↓
[142]	✓	✓	✓	✓	✓	✓	6	↓
[143]	×	✓	✓	✓	✓	✓	5	↓
[144]	×	✓	✓	✓	✓	✓	5	↓
[145]	×	✓	✓	✓	✓	✓	5	↓
[146]	✓	✓	✓	✓	•	✓	5	↓
[147]	×	✓	✓	✓	✓	✓	5	↓
[148]	✓	✓	✓	✓	✓	✓	6	↓
[149]	×	✓	✓	✓	✓	✓	5	↓
[150]	✓	✓	✓	✓	✓	✓	6	↓
[151]	✓	✓	✓	✓	✓	✓	6	↓
[152]	×	✓	✓	✓	✓	✓	5	↓
[153]	✓	✓	✓	✓	✓	✓	6	↓
[154]	×	✓	✓	✓	•	✓	4	≈
[155]	×	✓	✓	✓	✓	✓	5	↓
[156]	×	✓	✓	✓	✓	✓	5	↓

XSL•FO
RenderX

Reference	Q1	Q2	Q3	Q4	Q5	Q6	Total "Y"	Risk category
[157]	✓	✓	✓	✓	✓	✓	6	↓
[158]	✓	✓	✓	✓	✓	✓	6	↓
[24]	✓	✓	✓	✓	✓	✓	6	↓
[159]	×	✓	✓	✓	✓	✓	5	↓
[160]	×	✓	✓	✓	✓	✓	5	↓
[161]	×	✓	✓	✓	✓	✓	5	↓
[162]	×	✓	✓	✓	✓	✓	5	↓
[163]	✓	✓	✓	✓	✓	✓	6	↓
[164]	✓	✓	✓	✓	✓	✓	6	↓
[165]	✓	✓	✓	✓	✓	✓	6	↓
[166]	✓	✓	✓	✓	•	✓	5	↓
[167]	✓	✓	✓	✓	✓	✓	6	↓
[168]	✓	✓	✓	✓	✓	✓	6	↓
[169]	✓	✓	✓	✓	•	✓	5	↓
[170]	✓	✓	✓	✓	•	✓	5	↓

[a]✓: Yes.

[b]↓: Low risk of bias.

[c]•: Unclear.

[d]×: No.

[e]≈: Moderate risk of bias.

Data Sources and Characteristics

Across the 142 studies reviewed, over half used private datasets (81/142, 57.04%), while publicly accessible datasets were used in 29 studies (20.42%). A smaller number of studies used datasets that were available via specific data use agreements (7/142, 4.93%), accessible only for shared tasks (5/142, 3.52%), or available in partial form (subset available, 5/142, 3.52%). The most common data type was EHRs, reported in 93 studies. Eight studies used national datasets (Primary Land Use Tax Lot Output [PLUTO], CalEnviroScreen 4.0, etc). Social media sources, including Twitter, Reddit, and Facebook, were leveraged in 9 studies (5.1%). Other, less frequently reported types included research abstracts, web content, and interviews.

The majority of private datasets included institution-specific EHRs (eg, Columbia University Irving Medical Center; University of California, San Francisco; Johns Hopkins; Eskenazi; Medical University of South Carolina; UF Health; UNC; Kaiser Permanente Southern California) and clinical data warehouses tied to academic or health systems. The Veterans Affairs health system was particularly prominent, covering datasets such as: Veterans Affairs Corporate Data Warehouse [46,47,75,83,89,90,94,135], Veteran Health Administration notes [114], Supportive Services for Veteran Families [75], administrative data [136], Veterans Aging Cohort Study [141]. Several datasets incorporated augmented sources to enrich available information, such as LexisNexis, geospatial SDOH (from diverse government sources [47]), and patient-reported tools (eg, Timeline Follow-Back [83]). Publicly available datasets included Medical Information Mart for Intensive Care III (MIMIC-III) [31,32,55,57,58,74,77,78,82,85,86,95,97,104,118,125,128,130], n2c2 2018/2022 [55,58,80,85,128], SemEval-2015 [128], LGBTQ+ Minority Stress on Social Media (MiSSoM+) dataset [92], and others. MIMIC-III and its derivatives (eg, social and behavioral determinants of health-MIMIC) were frequently used for training [85,86,95], validation [82], or annotation purposes [32].

SDOH Factors

Figure 3 illustrates the most commonly studied SDOH factors across the years. We show factors that appeared in at least 15 papers. Table 2 shows the SDOH type (individual vs structural) studied in our cohort of papers. The most studied SDOH factors were housing instability (50.7%) and employment (45.8%), appearing in nearly half of the 142 papers, followed by financial context (44.4%), substance use (37.3%), and social isolation (34.5%). Education and living circumstances each appeared in 27.5% and 28.9% of studies, respectively. At the lower end, justice system involvement and language literacy each appeared in approximately 8% of studies. The least studied factors included immigration status (3.5%) and 4 factors that appeared in less than 3 papers each: access to lethal means, acculturation, digital divide, and military sexual trauma. Collectively, these results show that individual SDOH (substance use, social connection/isolation, etc) received more research attention than structural SDOH (transportation, insurance, etc) over the years. The uneven focus on certain SDOH could be due to data and methodological constraints rather than importance alone. Individual-level factors such as housing and employment are more explicitly documented and easier to annotate, whereas structural determinants (eg, immigration status or the digital

XSL•FO
RenderX

divide) are often inconsistently recorded or absent from clinical text.

Figure 3. Distribution of the most studied social determinants of health over the years. Bubble size is proportional to the number of studies. Individual-level determinants are frequently studied compared with structural determinants. Table S2 in Multimedia Appendix 1 for the category dictionary. *partial data for 2025.

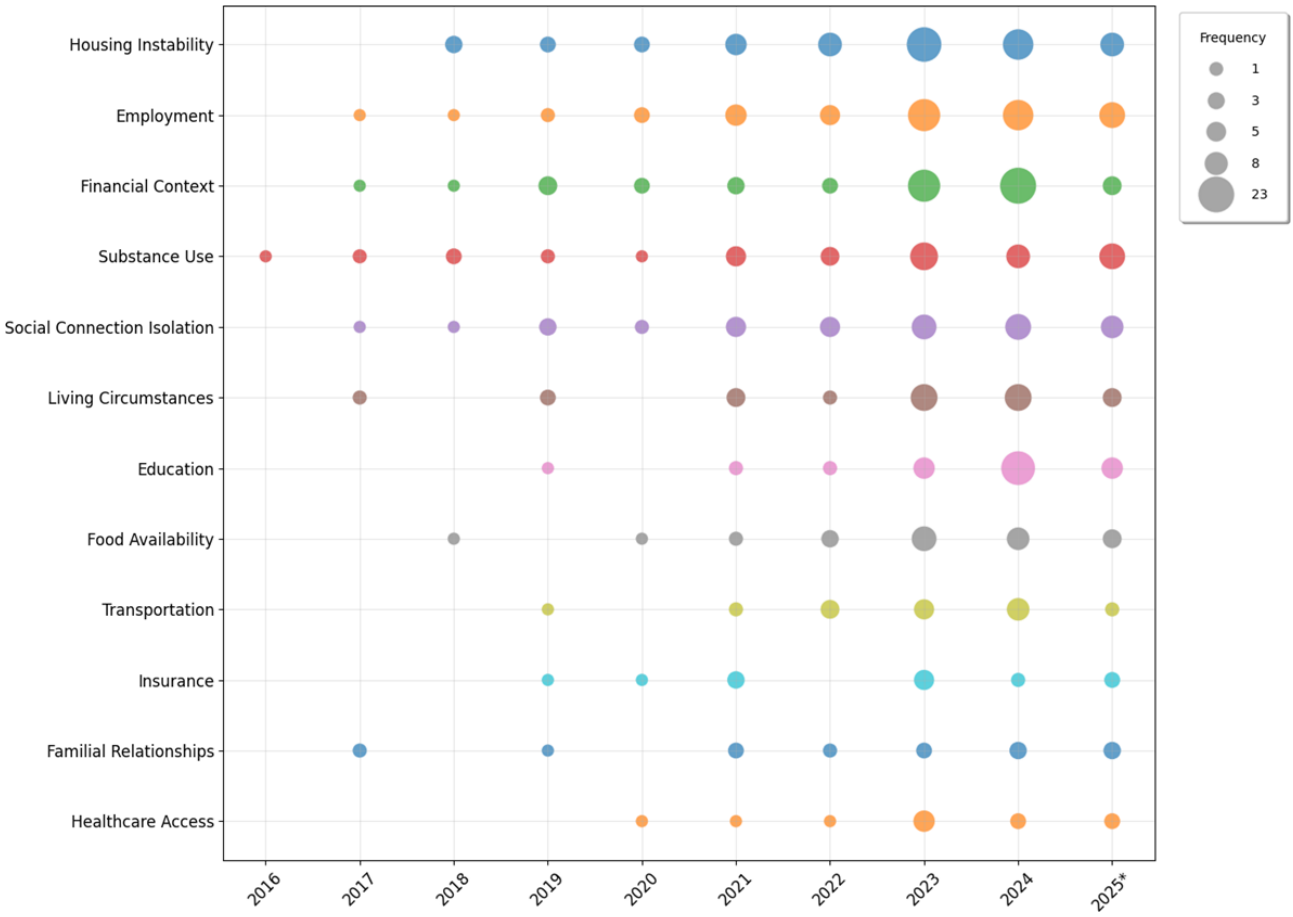

Table . Social Determinants of Health (SDOH) type distribution.

SDOH type	Reference
Individual (n=79)	[43,111,115-117,119,122-125,127,138,143,144], [12,20,87,93,94,98,99,101,114,118,121,130,136,141,142], [22,23,29,31,45,47,55,57,58,62,70,71,73,74,78,80-83,85,86,90,132-134,137,145-147,149,150,152], [24,135,148,155-166,168-170]
Structural (n=9)	[48,53,59,88,110,112,120,129,126]
Both (n=54)	[21,44,46,95,96,100,102-109,113], [25,27,32,52,56,63,64,67,75-77,79,84,89,91,92,97], [26,28,30,49-51,54,60,61,65,66,68,69,72,128,131,139,140,151,153,154,167]

NLP Methods and Performance

Figure 4A presents a heatmap illustrating the temporal distribution and frequency of the dominant NLP methods identified in this review. Among the included studies, the majority (n=83) quantified model performance using standard metrics, including F_1-score, precision, and recall. Figure 4B delineates the distribution of F_1-scores across the 5 categories, highlighting significant performance variations between rule-based, traditional machine learning, and deep learning approaches. Complementing this, Figure 5 provides a bubble chart depicting the mean precision and recall values for each model category. Collectively, the visualized performance trends show the general capabilities of different modeling approaches for SDOH tasks, though performance may vary based on task complexity, dataset characteristics, and annotation quality.

XSL•FO
RenderX

Figure 4. (A) Frequency heatmap representative of the best-performing NLP methodologies across years. Interestingly, transformer- and large language model–based studies surged over the last 3 years. (B) Box plot showing F_1-score distribution by model category. Models with fewer than 20 samples are excluded, and extreme values are removed for readability. Table S3 in Multimedia Appendix 1 details the category dictionary. *partial data for 2025. NLP: natural language processing.

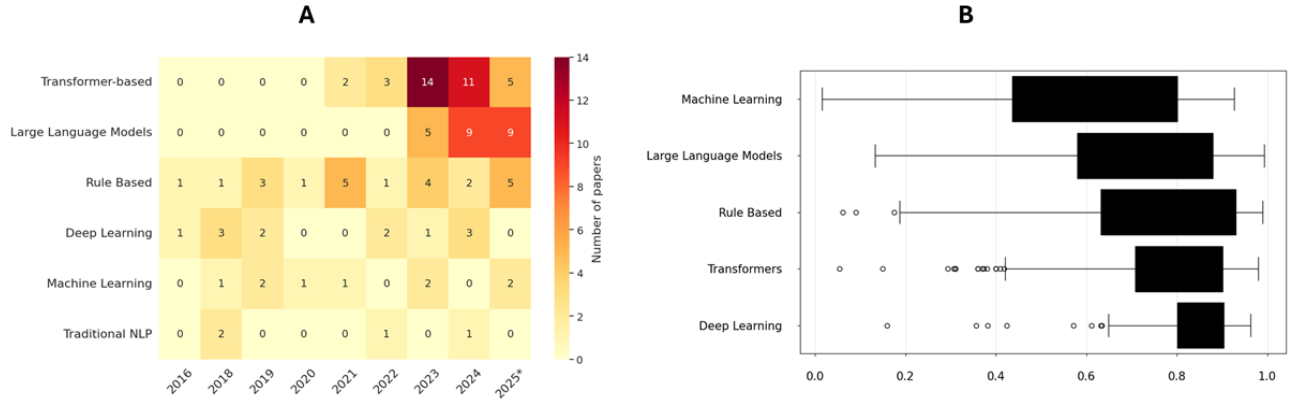

Figure 5. Bubble chart representing mean values of precision and recall for different NLP model categories. The dashed line represents the equal-precision and recall boundary. Bubble size is proportional to the number of models; only categories with ≥ 7 observations are shown. Performance metrics are from heterogeneous evaluation contexts, representative of aggregate trends, and are not individually comparable across studies. Table S4 in Multimedia Appendix 1 details NLP method categorization. LSTM: long short-term memory; ML: machine learning; NLP: natural language processing; RNN: recurrent neural network.

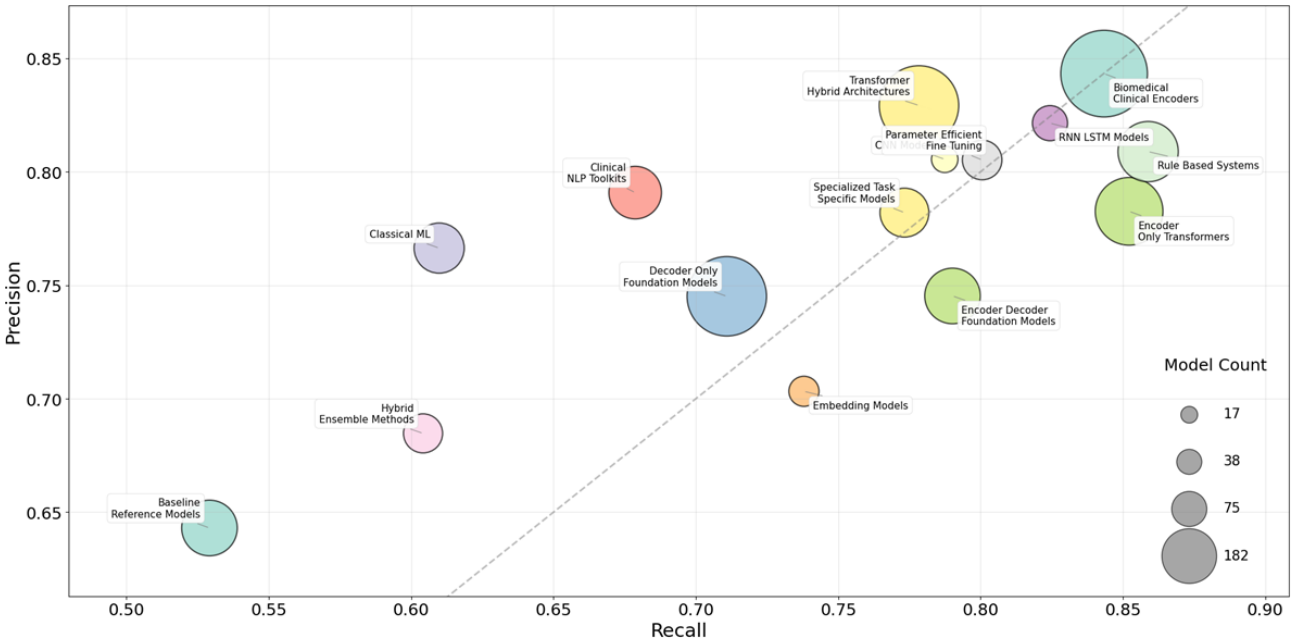

Overall, the results highlight that transformer-based encoder-only models such as BERT and generative models (encoder-decoders) trained on specialized clinical tasks currently represent the state of the art in terms of SDOH classification and extraction performance. LLMs (19.7% of reviewed studies, including encoder-decoders) and their capabilities represent an emerging area of research. The limited adoption of LLMs likely reflects practical and ethical concerns rather than a lack of capability. Challenges related to hallucinations, limited reproducibility, fairness and bias risks, and uncertainty around data leakage and governance (especially in clinical settings) may constrain their broader use despite strong performance in controlled evaluations. The majority of studies included in this review (≈90%) use a cross-sectional design, focusing on extracting or classifying SDOH information from a single point in time, providing a valuable snapshot. While useful, such approaches cannot capture the changes in SDOH factors among study subjects. Only a handful of studies [66,75,83] adopt a longitudinal approach, following individuals' records (like EHRs) over time to track changes in SDOH and their impact on health outcomes.

Model Interpretation Techniques

Model interpretation is essential for understanding not just how well a model performs, but also the basis for its decisions. A total of 82 studies reported on using interpretability techniques to gain insights into model performance. These included qualitative content analysis [29,59,131], Shapley additive explanations [47,71,163], Local Interpretable Model-Agnostic Explanations [23], attention visualization or neuron activation analysis [54], and ablation studies [57,87]. One common approach found across studies was error analysis, which plays a key role in understanding how and why models make specific mistakes. It involves examining misclassifications or incorrect

XSL•FO
RenderX

outputs to identify patterns of failure. Several studies in our review performed manual (domain expert-based) error analysis to identify model weaknesses. For instance, one paper found that their model struggled to distinguish between general and specific substance-related terms by incorrectly extracting the generic verb "smokes" as the key information when the actual target was the more specific noun "cigarettes" [77]. Another study showed that soft prompting (using trainable continuous vectors instead of manually crafted text instructions) improved the extraction of overlapping or nested SDOH concepts, especially when prompt length was tuned [58]. Error analyses were particularly prominent in studies using LLMs such as ChatGPT, Llama, and Gemini [26,31,57,59,61,155-157,159] because these models (being black-box) can produce outputs that appear coherent and convincing even when they contain mistakes. In addition, confusion matrices were used to visualize class-level misclassification patterns and highlight systematic errors [62].

Influential Studies

Analysis of the 20 influential studies published between 2018 and 2025 reveals a methodological landscape dominated by advanced NLP architectures, including BERT [60,78,80,118,130] and LLMs [57], alongside some rule-based [22,127] and hybrid approaches [27,109,168]. Notably, three-quarters (15/20) of these studies relied on manual annotation or clinician input to establish ground truth, highlighting a continued reliance on human expertise to ensure data quality and study robustness. The prevalence of multi-institutional collaborations in these studies (n=11) suggests that diverse teams and broader data sources are key drivers of research impact, likely facilitating increased generalizability and resource access. Regarding performance, the majority of studies (13/20) reported strong F_1-scores, though results varied by task complexity. High benchmarks included concept extraction scores up to 0.9118, and relation extraction at 0.8332 [58], with Seq2Seq models achieving approximately 0.889 [80] and other systems reaching 0.86 [78]. Performance heterogeneity was evident in class-specific tasks; for example, one study reported attribute performance (eg, substance use and employment) ranging from 0.81 to 0.93 [118], while another reported a wider spread from 0.491 for non-SDOH factors to 0.774 for the best-performing class (occupation) using the same model [130]. Beyond standard metrics, one study involving EHR data reported that NLP-based methodology captured SDOH in 80.03% of cases compared with 38.17% using structured fields only [89]. Surprisingly, only 9 out of these 20 studies provided publicly accessible code. Although this represents a higher proportion than the overall set (49/142, 34.5% papers shared their code), public code availability remains limited even among the most influential studies.

Publishing Venues and Funding Sources

The most frequent publication venue was the Journal of the American Medical Informatics Association (12/142, 8.5%), followed by Journal of the American Medical Informatics Association Open (4.9%), Journal of Medical Internet Research (4.2%), and Journal of Biomedical Informatics (4.2%). Conference contributions were most often from the IEEE and AMIA Annual Symposium Proceedings. A majority of studies (110/142, 77.46%) acknowledged funding. The National Institutes of Health and its affiliated institutes (National Library of Medicine, National Center for Advancing Translational Sciences, National Cancer Institute, National Institute on Aging, and NIDA) were the leading funders (64/110, 58.2% of funded studies). Other key funders included the National Science Foundation (n=10), Agency for Healthcare Research and Quality (n=6), and international agencies such as the National Institute for Health Research and Canadian Institutes of Health Research. There is an increase in the number of studies reporting federal funding for SDOH-related NLP research in recent years. The percentage of funded studies more than doubled from 33% to 67% during 2018 - 2020, to 78% - 83% during 2021 - 2025.

Discussion

Principal Findings

This systematic review synthesizes evidence from 142 studies, addressing 2 primary objectives: to characterize NLP techniques, including LLMs, used to analyze SDOH in text-based data, and to assess their reported effectiveness, gaps, and future research needs. Overall, the review reveals rapid methodological advancement and improvement in NLP performance in recent years, marked by a shift toward transformer-based architectures and emerging use of LLMs, with most applications focused on extracting or classifying individual-level SDOH from clinical and social text. Despite technical progress, the review shows that challenges related to data availability, performance reproducibility, model interpretability, and translation into real-world clinical or public health practice remain. Key findings are elaborated in the following paragraphs.

Across studies, model performance in identifying individual SDOH from unstructured text has generally improved over time, reflecting advances in representation learning and modern NLP systems' ability to capture abstract and context-dependent social concepts [171]. These advances are likely driven by both increased recognition from the research community of the role of SDOH in shaping health outcomes [172] and by the maturation of transformer-based models that perform well on complex linguistic tasks [173]. Strong reported performance, however, does not consistently translate into reproducible findings or application-ready systems [174]. Many studies relied on private, institution-specific EHR datasets, which limit independent replication and cross-site validation [175]. Although manual and clinician-led annotation was common across the reviewed studies, detailed documentation of annotation guidelines and procedures was often lacking, making it difficult to assess consistency in how SDOH concepts were defined and applied across studies. Limited transparency around data and annotation practices also has important equity implications [176]. Reliance on proprietary EHR data favors well-resourced institutions and restricts broader participation in SDOH-focused NLP research [177].

This review also sheds light on the gap between technical development and practical implementation, and limitations on model interpretability. While many studies report promising model performance, few detail pathways for integrating

XSL•FO
RenderX

NLP-derived SDOH insights into clinical workflows or public health interventions, which may be due to systemic, technical, and regulatory barriers [178]. As a result, the potential of these methods to inform equity-focused decision-making remains largely unrealized, a finding supported by prior research [176]. NLP/LLM model interpretability has received more research attention in recent years [179]. A little over half of all reviewed studies used some variant of interpretability or explainability methods, but their use was inconsistent and often inadequately discussed. This gap is relatively more critical for SDOH applications, where language may reflect complex social, cultural, and structural nuances, with uninterpretable errors disproportionately affecting marginalized groups [180]. Limited interpretability also undermines trust in systems' ability to inform care or policy, both within and outside the sphere of SDOH [181].

Future Research Directions

Based on the review, we outline several priorities for advancing NLP and LLM applications in the study of SDOH. Most of the existing work used cross-sectional designs, offering only static snapshots of social context. Future research should adopt longitudinal designs capable of capturing the dynamic nature of social risk factors, their accumulation over time, and their interaction with health trajectories [182]. Such longitudinal approaches can include continuous monitoring as well as repeated snapshot or point-in-time measures collected at predetermined intervals [183], for tracking meaningful changes in SDOH. Future work may combine unstructured narratives with structured EHR variables, imaging, and community-level data to produce more holistic and actionable models [184,185]. Effective multimodal approaches must account for heterogeneity in SDOH-containing data [186] to maximize information capture and analytical utility. Advancing the field will also require stronger commitments to transparency and reproducibility. Establishing open benchmarks, shared tasks, and community challenges has driven NLP method development on many topics [80,128], and should be the focus of future work to promote transparency and reproducibility in SDOH-related NLP. Greater attention is also needed for structural determinants of health. Fewer than one-fifth of reviewed studies focused on structural SDOH factors such as policy, racism, or socioeconomic stratification [187]. Capturing structural determinants presents unique methodological challenges, such as patient-reported data on systemic factors being limited by awareness and subjectivity [188]. Linking individual-level health narratives with community- and policy-level data may offer a more comprehensive approach to understanding how structural forces shape health outcomes in future research [189]. Bias and ethical considerations also remain underexplored. Few studies systematically examined algorithmic bias, despite well-documented risks of inequities in automated decision-making [190,191]. Future systems must embed safeguards for fairness, inclusivity, and accountability to avoid reinforcing disparities. Recent work has proposed actionable frameworks and techniques for mitigating bias in artificial intelligence systems [192-194]. Adopting such approaches will be essential for ensuring SDOH-focused models promote rather than perpetuate health inequities. In addition, while clinical concepts are supported by structured hierarchies such as the Unified Medical Language System [195], no equivalent framework exists for social (or nonclinical) determinants. Establishing such frameworks would reduce inconsistencies, enable interoperability, and support the creation of benchmark datasets. Finally, improving model generalizability remains critical, as models trained on single-institution EHRs often underperform in external settings, reflecting the lexical, demographic, and contextual variability of clinical narratives [174,196]. Future work should therefore move beyond single-site training by explicitly benchmarking models in cross-institution evaluations [197]. Training on pooled multi-site data and testing on held-out institutions, with performance reported by SDOH category and population subgroup, may provide actionable evidence of real-world generalizability.

Limitations

This review has several limitations. Determining whether studies were directly relevant to SDOH required subjective judgment, as many did not explicitly frame their research in SDOH terms. This introduces potential selection bias, despite a structured review process. To balance comprehensiveness and feasibility, we adhered to predefined selection criteria, focusing primarily on studies with a clear connection to SDOH. Although this approach helped maintain relevance, it may have led to the exclusion of studies with indirect but meaningful contributions to the field. Future work could explore more systematic methods, such as automated screening tools or expert consensus frameworks, to enhance the consistency and reproducibility of the selection process. We also acknowledge that our synthesis is constrained by heterogeneity across study designs, SDOH definitions, annotation schemes, datasets, and evaluation metrics, which limits direct comparability of reported performance across studies.

Conclusions

In this systematic review, we provide a comprehensive synthesis of NLP and LLM applications for SDOH by systematically surveying methodologies, data sources, evaluation practices, and translational readiness across 142 studies. In contrast to prior reviews that covered subdomains of SDOH, clinical contexts, or individual modeling techniques, ours provides a comparison of tasks, architectures, performance patterns, interpretability practices, and reproducibility considerations across the research spectrum, spanning both individual and structural SDOH. By discussing NLP/LLM technical performance in the broader context of data accessibility, bias, generalizability, and implementation, this review advances the field beyond proof-of-concept model development toward a clearer understanding of what is required for real-world deployment. The findings highlight that, while NLP-based SDOH extraction is technically mature for certain use cases, its impact remains limited by fragmented SDOH representations, a paucity of public benchmarks, and insufficient evaluation in clinical and public health workflows. In light of our findings, future work should prioritize public benchmarks, clearer reporting, and more inclusive datasets to advance this emerging field. Addressing these gaps has direct real-world implications

XSL•FO
RenderX

by facilitating inclusive data practices, interpretable and reproducible methods, and implementation-focused evaluations.

Acknowledgments

Grammarly and Claude were used for grammar checking and Overleaf LaTeX formatting, respectively. We thank the anonymous reviewers and editor for their constructive suggestions during the review period that improved the quality of this manuscript.

Funding

YX is supported by the National Institute of Mental Health (RF1MH134649), the American Foundation for Suicide Prevention (YIG-2-133-22), and Google. AS is supported by the National Institute on Drug Abuse (R01DA057599). The content in the paper does not necessarily represent the official views of any of the institutions mentioned. The sponsors had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

Data for this study were sourced from reviewed articles referenced in this manuscript. Literature search string queries are available in Multimedia Appendix 1.

Authors' Contributions

SR contributed to the conceptualization, methodology, investigation, data curation, and writing of the original draft, as well as review and editing of the manuscript. AKP contributed to data curation, visualization, and review and editing. ZZ and YC contributed to data curation and review and editing of the manuscript. ML and SD contributed to data curation, while HR contributed to resources and data curation. AS and YX contributed to conceptualization, supervision, and project administration. All authors participated in the review and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search queries, results, and categorization; summary of studies.
[PDF File, 403 KB - jmir_v28i1e83793_app1.pdf]

Checklist 1
PRISMA checklist.
[PDF File, 123 KB - jmir_v28i1e83793_app2.pdf]

References

1. Social determinants of health. World Health Organization. 2025. URL: https://www.who.int/health-topics/social-determinants-of-health [accessed 2026-03-03]
2. Social determinants of health (healthy people 2030). US Department of Health and Human Services, Office of Disease Prevention and Health Promotion. 2025. URL: https://odphp.health.gov/healthypeople/priority-areas/social-determinants-health [accessed 2026-03-03]
3. Hood CM, Gennuso KP, Swain GR, Catlin BB. County health rankings: relationships between determinant factors and health outcomes. Am J Prev Med 2016 Feb;50(2):129-135. [doi: 10.1016/j.amepre.2015.08.024] [Medline: 26526164]
4. Singh GK, Daus GP, Allender M, et al. Social determinants of health in the United States: addressing major health inequality trends for the nation, 1935-2016. Int J MCH AIDS 2017;6(2):139-164. [doi: 10.21106/ijma.236] [Medline: 29367890]
5. Magnan S. Social determinants of health 101 for health care: five plus five. NAM perspectives discussion paper.: National Academy of Medicine; 2017. URL: https://nam.edu/social-determinants-of-health-101-for-health-care-five-plus-five/ [accessed 2026-03-03]
6. Artiga S, Hinton E. Beyond health care: the role of social determinants in promoting health and health equity. Kaiser Family Foundation. 2018. URL: https://www.kff.org/racial-equity-and-health-policy/beyond-health-care-the-role-of-social-determinants-in-promoting-health-and-health-equity/ [accessed 2026-04-08]
7. Marmot M, Allen J, Boyce T, Goldblatt P, Morrison J. Health equity in England: the marmot review 10 years on. The Health Foundation. 2020. URL: https://www.health.org.uk/reports-and-analysis/reports/health-equity-in-england-the-marmot-review-10-years-on-0 [accessed 2026-04-08]

XSL•FO
RenderX

8. Bundy JD, Mills KT, He H, et al. Social determinants of health and premature death among adults in the USA from 1999 to 2018: a national cohort study. Lancet Public Health 2023 Jun;8(6):e422-e431. [doi: 10.1016/S2468-2667(23)00081-6] [Medline: 37244672]
9. Xiao Y, Mann JJ, Chow JCC, et al. Patterns of social determinants of health and child mental health, cognition, and physical health. JAMA Pediatr 2023 Dec 1;177(12):1294-1305. [doi: 10.1001/jamapediatrics.2023.4218] [Medline: 37843837]
10. Xiao Y, Yip PSF, Pathak J, Mann JJ. Association of social determinants of health and vaccinations with child mental health during the COVID-19 pandemic in the US. JAMA Psychiatry 2022 Jun 1;79(6):610-621. [doi: 10.1001/jamapsychiatry.2022.0818] [Medline: 35475851]
11. Xiao Y, Brown TT, Snowden LR, Chow JCC, Mann JJ. COVID-19 policies, pandemic disruptions, and changes in child mental health and sleep in the United States. JAMA Netw Open 2023 Mar 1;6(3):e232716. [doi: 10.1001/jamanetworkopen.2023.2716] [Medline: 36912834]
12. Feller DJ, Bear Don't Walk Iv OJ, Zucker J, Yin MT, Gordon P, Elhadad N. Detecting social and behavioral determinants of health with structured and free-text clinical data. Appl Clin Inform 2020 Jan;11(1):172-181. [doi: 10.1055/s-0040-1702214] [Medline: 32131117]
13. Truong HP, Luke AA, Hammond G, Wadhera RK, Reidhead M, Joynt Maddox KE. Utilization of social determinants of health ICD-10 Z-codes among hospitalized patients in the United States, 2016-2017. Med Care 2020 Dec;58(12):1037-1043. [doi: 10.1097/MLR.0000000000001418] [Medline: 32925453]
14. Park Y, Mulligan N, Gleize M, Kristiansen M, Bettencourt-Silva JH. Discovering associations between social determinants and health outcomes: merging knowledge graphs from literature and electronic health data. AMIA Annu Symp Proc 2021;2021:940-949. [Medline: 35308956]
15. Rajwal S, Pandey AK, Han Z, Sarker A. Unveiling voices: identification of concerns in a social media breast cancer cohort via natural language processing. Presented at: Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024; May 20, 2024 URL: https://aclanthology.org/2024.cl4health-1.32/ [accessed 2026-03-03]
16. Guo Y, Rajwal S, Lakamana S, et al. Generalizable natural language processing framework for migraine reporting from social media. AMIA Jt Summits Transl Sci Proc 2023;2023:261-270. [Medline: 37350878]
17. Gonzalez-Hernandez G, Sarker A, O'Connor K, Savova G. Capturing the patient's perspective: a review of advances in natural language processing of health-related text. Yearb Med Inform 2017 Aug;26(1):214-227. [doi: 10.15265/IY-2017-029] [Medline: 29063568]
18. Gauthier MP, Law JH, Le LW, et al. Automating access to real-world evidence. JTO Clin Res Rep 2022 Jun;3(6):100340. [doi: 10.1016/j.jtocrr.2022.100340] [Medline: 35719866]
19. Yu S, Le A, Feld E, et al. A natural language processing-assisted extraction system for Gleason scores: development and usability study. JMIR Cancer 2021 Jul 2;7(3):e27970. [doi: 10.2196/27970] [Medline: 34255641]
20. Rouillard CJ, Nasser MA, Hu H, Roblin DW. Evaluation of a natural language processing approach to identify social determinants of health in electronic health records in a diverse community cohort. Med Care 2022 Mar 1;60(3):248-255. [doi: 10.1097/MLR.0000000000001683] [Medline: 34984989]
21. Stemerman R, Arguello J, Brice J, Krishnamurthy A, Houston M, Kitzmiller R. Identification of social determinants of health using multi-label classification of electronic health record clinical notes. JAMIA Open 2021 Jul;4(3):ooaa069. [doi: 10.1093/jamiaopen/ooaa069] [Medline: 34514351]
22. Wu W, Holkeboer KJ, Kolawole TO, Carbone L, Mahmoudi E. Natural language processing to identify social determinants of health in Alzheimer's disease and related dementia from electronic health records. Health Serv Res 2023 Dec;58(6):1292-1302. [doi: 10.1111/1475-6773.14210] [Medline: 37534741]
23. Gray GM, Zirikly A, Ahumada LM, et al. Application of natural language processing to identify social needs from patient medical notes: development and assessment of a scalable, performant, and rule-based model in an integrated healthcare delivery system. JAMIA Open 2023 Dec;6(4):ooad085. [doi: 10.1093/jamiaopen/ooad085] [Medline: 37799347]
24. Soley N, Bentil M, Shah J, Rouhizadeh M, Taylor CO. Unveiling social determinants of health impact on adverse pregnancy outcomes through natural language processing. Sci Rep 2025 Aug 9;15(1):29183. [doi: 10.1038/s41598-025-13542-x] [Medline: 40783439]
25. Lowery B, D'Acunto S, Crowe RP, Fishe JN. Using natural language processing to examine social determinants of health in prehospital pediatric encounters and associations with EMS transport decisions. Prehosp Emerg Care 2023;27(2):246-251. [doi: 10.1080/10903127.2022.2072984] [Medline: 35500212]
26. Fu Y, Ramachandran GK, Dobbins NJ, et al. Extracting social determinants of health from pediatric patient notes using large language models: novel corpus and methods. In: Calzolari N, Kan MY, Hoste V, Lenci A, Sakti S, Xue N, editors. Presented at: Joint International Conference on Computational Linguistics, Language Resources and Evaluation; May 20-25, 2024; Torino, Italy p. 7045-7056 URL: https://aclanthology.org/2024.lrec-main.618/ [accessed 2026-03-03]
27. Lituiev DS, Lacar B, Pak S, Abramowitsch PL, De Marchis EH, Peterson TA. Automatic extraction of social determinants of health from medical notes of chronic lower back pain patients. J Am Med Inform Assoc 2023 Jul 19;30(8):1438-1447. [doi: 10.1093/jamia/ocad054] [Medline: 37080559]

XSL•FO
RenderX

28. Sun S, Zack T, Williams CYK, Sushil M, Butte AJ. Topic modeling on clinical social work notes for exploring social determinants of health factors. JAMIA Open 2024 Apr;7(1):ooad112. [doi: 10.1093/jamiaopen/ooad112] [Medline: 38223407]
29. Burnett SJ, Stemerman R, Innes JC, Kaisler MC, Crowe RP, Clemency BM. Social determinants of health in EMS records: a mixed-methods analysis using natural language processing and qualitative content analysis. West J Emerg Med 2023 Sep;24(5):878-887. [doi: 10.5811/westjem.59070] [Medline: 37788028]
30. Whitfield C, Liu Y, Anwar M. Impact of COVID-19 pandemic on social determinants of health issues of marginalized Black and Asian communities: a social media analysis empowered by natural language processing. J Racial Ethn Health Disparities 2025 Jun;12(3):1641-1656. [doi: 10.1007/s40615-024-01996-0] [Medline: 38625665]
31. Shah-Mohammadi F, Finkelstein J. AI-powered social determinants of health extraction from patient records: a GPT-based investigation. Presented at: 2024 IEEE First International Conference on Artificial Intelligence for Medicine, Health and Care (AIMHC); Feb 5-7, 2024; Laguna Hills, CA, USA p. 109-112. [doi: 10.1109/AIMHC59811.2024.00028]
32. Wang X, Gupta D, Killian M, He Z. Benchmarking transformer-based models for identifying social determinants of health in clinical notes. Proc IEEE Int Conf Healthc Inform 2023 Jun;2023:570-574. [doi: 10.1109/ichi57859.2023.00102] [Medline: 38239824]
33. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. . Preprint posted online on Jan 16, 2013 URL: https://arxiv.org/abs/1301.3781v3 [accessed 2026-03-03]
34. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. . Preprint posted online on Aug 2, 2023 URL: http://arxiv.org/abs/1706.03762 [accessed 2026-03-03]
35. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, editors. Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Jun 2-7, 2019. [doi: 10.18653/v1/N19-1423]
36. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. Semantic Scholar. 2018. URL: https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035 [accessed 2026-04-08]
37. Rajwal S, Zhang Z, Chen Y, Rogers H, Sarker A, Xiao Y. Applications of natural language processing and large language models for social determinants of health: protocol for a systematic review. JMIR Res Protoc 2025 Jan 21;14:e66094. [doi: 10.2196/66094] [Medline: 39836952]
38. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. In: Tricco AC, Welch VA, Whiting P, Moher D, editors. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
39. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
40. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ 2020 Jan 16;368:l6890. [doi: 10.1136/bmj.l6890] [Medline: 31948937]
41. The world's #1 systematic review tool. Covidence. URL: https://www.covidence.org/ [accessed 2026-03-03]
42. Aromataris E, Fernandez R, Godfrey CM, Holly C, Khalil H, Tungpunkom P. Summarizing systematic reviews: methodological development, conduct and reporting of an umbrella review approach. Int J Evid Based Healthc 2015 Sep;13(3):132-140. [doi: 10.1097/XEB.0000000000000055] [Medline: 26360830]
43. Feller DJ, Zucker J, Yin MT, Gordon P, Elhadad N. Using clinical notes and natural language processing for automated HIV risk assessment. JAIDS 2018;77(2):160-166. [doi: 10.1097/QAI.0000000000001580]
44. Bako AT, Walter-McCabe H, Kasthurirathne SN, Halverson PK, Vest JR. Reasons for social work referrals in an urban safety-net population: a natural language processing and market basket analysis approach. J Soc Serv Res 2021 May 4;47(3):414-425. [doi: 10.1080/01488376.2020.1817834]
45. Zhu J, Yalamanchi N, Jin R, Kenne DR, Phan N. Investigating COVID-19's impact on mental health: trend and thematic analysis of reddit users' discourse. J Med Internet Res 2023 Jul 12;25:e46867. [doi: 10.2196/46867] [Medline: 37436793]
46. Morrow D, Zamora-Resendiz R, Beckham JC, et al. A case for developing domain-specific vocabularies for extracting suicide factors from healthcare notes. J Psychiatr Res 2022 Jul;151:328-338. [doi: 10.1016/j.jpsychires.2022.04.009] [Medline: 35533516]
47. Kessler RC, Bauer MS, Bishop TM, et al. Evaluation of a model to target high-risk psychiatric inpatients for an intensive postdischarge suicide prevention intervention. JAMA Psychiatry 2023 Mar 1;80(3):230-240. [doi: 10.1001/jamapsychiatry.2022.4634] [Medline: 36652267]
48. Wang T, Codling D, Bhugra D, et al. Unraveling ethnic disparities in antipsychotic prescribing among patients with psychosis: a retrospective cohort study based on electronic clinical records. Schizophr Res 2023 Oct;260:168-179. [doi: 10.1016/j.schres.2023.08.024] [Medline: 37669576]
49. Chilman N, Laporte D, Dorrington S, et al. Understanding social and clinical associations with unemployment for people with schizophrenia and bipolar disorders: large-scale health records study. Soc Psychiatry Psychiatr Epidemiol 2024 Oct;59(10):1709-1719. [doi: 10.1007/s00127-024-02620-6] [Medline: 38378812]

XSL•FO
RenderX

50. Wang L, Guo Y, Yin X, Wang Y, Tong R. Exploring the determinants of health-promoting behaviors among miners: a text mining and meta-analysis. Appl Psychol Health Well Being 2024 Feb;16(1):3-24. [doi: 10.1111/aphw.12465] [Medline: 37339782]
51. Goldstein EV, Bailey EV, Wilson FA. Poverty and suicidal ideation among hispanic mental health care patients leading up to the COVID-19 pandemic. Hisp Health Care Int 2024 Mar;22(1):6-10. [doi: 10.1177/15404153231181110]
52. Xie F, Wang S, Viveros L, et al. Using natural language processing to identify the status of homelessness and housing instability among serious illness patients from clinical notes in an integrated healthcare system. JAMIA Open 2023 Jul 4;6(3). [doi: 10.1093/jamiaopen/ooad082]
53. Harris DR, Fu S, Wen A, et al. The ENACT network is acting on housing instability and the unhoused using the open health natural language processing toolkit. J Clin Transl Sci 2024;8(1):e98. [doi: 10.1017/cts.2024.543] [Medline: 39655040]
54. Ashayeri M, Abbasabadi N. Unraveling energy justice in NYC urban buildings through social media sentiment analysis and transformer deep learning. Energy Build 2024 Mar;306:113914. [doi: 10.1016/j.enbuild.2024.113914]
55. Peng C, Yang X, Yu Z, Bian J, Hogan WR, Wu Y. Clinical concept and relation extraction using prompt-based machine reading comprehension. J Am Med Inform Assoc 2023 Aug 18;30(9):1486-1493. [doi: 10.1093/jamia/ocad107]
56. Hobensack M, Song J, Oh S, et al. Social risk factors are associated with risk for hospitalization in home health care: a natural language processing study. J Am Med Dir Assoc 2023 Dec;24(12):1874-1880. [doi: 10.1016/j.jamda.2023.06.031]
57. Guevara M, Chen S, Thomas S, et al. Large language models to identify social determinants of health in electronic health records. NPJ Digit Med 2024 Jan 11;7(1):6. [doi: 10.1038/s41746-023-00970-0] [Medline: 38200151]
58. Peng C, Yang X, Smith KE, et al. Model tuning or prompt tuning? a study of large language models for clinical concept and relation extraction. J Biomed Inform 2024 May;153:104630. [doi: 10.1016/j.jbi.2024.104630] [Medline: 38548007]
59. Hu Z, Zhang Y, Rossi R, Yu T, Kim S, Pan S. Are Large Language Models Capable of Causal Reasoning for Sensing Data Analysis: Association for Computing Machinery, Inc; 2024:24-29. [doi: 10.1145/3662006.3662064]
60. Savcisens G, Eliassi-Rad T, Hansen LK, et al. Using sequences of life-events to predict human lives. Nat Comput Sci 2024 Jan;4(1):43-56. [doi: 10.1038/s43588-023-00573-5] [Medline: 38177491]
61. Luo X, Tahabi FM, Marc T, Haunert LA, Storey S. Zero-shot learning to extract assessment criteria and medical services from the preventive healthcare guidelines using large language models. J Am Med Inform Assoc 2024 Aug 1;31(8):1743-1753. [doi: 10.1093/jamia/ocae145]
62. Kwon S, Wang X, Liu W, et al. ODD: a benchmark dataset for the natural language processing based opioid related aberrant behavior detection. Presented at: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics; Jun 16-21, 2024; Mexico City, Mexico p. 4338-4359. [doi: 10.18653/v1/2024.naacl-long.244]
63. Vest JR, Mazurenko O. Non-response bias in social risk factor screening among adult emergency department patients. J Med Syst 2023 Jul 22;47(1):78. [doi: 10.1007/s10916-023-01975-8] [Medline: 37480515]
64. Harris DR, Anthony N, Quesinberry D, Delcher C. Evidence of housing instability identified by addresses, clinical notes, and diagnostic codes in a real-world population with substance use disorders. J Clin Transl Sci 2023;7(1):e196. [doi: 10.1017/cts.2023.626] [Medline: 37771412]
65. Cybulski L, Chilman N, Jewell A, et al. Improving our understanding of the social determinants of mental health: a data linkage study of mental health records and the 2011 UK census. BMJ Open 2024 Jan 29;14(1):e073582. [doi: 10.1136/bmjopen-2023-073582] [Medline: 38286672]
66. Xie K, Ojemann WKS, Gallagher RS, et al. Disparities in seizure outcomes revealed by large language models. J Am Med Inform Assoc 2024 May 20;31(6):1348-1355. [doi: 10.1093/jamia/ocae047]
67. Shafer PR, Davis A, Clark JA. Finding social need-les in a haystack: ascertaining social needs of medicare patients recorded in the notes of care managers. BMC Health Serv Res 2023 Dec 12;23(1):1400. [doi: 10.1186/s12913-023-10446-2] [Medline: 38087286]
68. Miranda O, Fan P, Qi X, et al. DeepBiomarker2: prediction of alcohol and substance use disorder risk in post-traumatic stress disorder patients using electronic medical records and multiple social determinants of health. J Pers Med 2024 Jan 14;14(1):2024. [doi: 10.3390/jpm14010094] [Medline: 38248795]
69. Yu Z, Peng C, Yang X, et al. Identifying social determinants of health from clinical narratives: a study of performance, documentation ratio, and potential bias. J Biomed Inform 2024 May;153:104642. [doi: 10.1016/j.jbi.2024.104642] [Medline: 38621641]
70. Magoc T, Allen KS, McDonnell C, et al. Generalizability and portability of natural language processing system to extract individual social risk factors. Int J Med Inform 2023 Sep;177:105115. [doi: 10.1016/j.ijmedinf.2023.105115] [Medline: 37302362]
71. Huang T, Socrates V, Gilson A, et al. Identifying incarceration status in the electronic health record using large language models in emergency department settings. J Clin Transl Sci 2024;8(1):e53. [doi: 10.1017/cts.2024.496] [Medline: 38544748]
72. Bhanot K, Erickson JS, Bennett KP. MortalityMinder: visualization and AI interpretations of social determinants of premature mortality in the United States. Information 2024;15(5):254. [doi: 10.3390/info15050254]
73. Mehta S, Lyles CR, Rubinsky AD, et al. Social determinants of health documentation in structured and unstructured clinical data of patients with diabetes: comparative analysis. JMIR Med Inform 2023 Aug 22;11:e46159. [doi: 10.2196/46159] [Medline: 37621203]

XSL•FO
RenderX

74. Richie R, Ruiz VM, Han S, Shi L, Tsui FR. Extracting social determinants of health events with transformer-based multitask, multilabel named entity recognition. J Am Med Inform Assoc 2023 Jul 19;30(8):1379-1388. [doi: 10.1093/jamia/ocad046] [Medline: 37002953]
75. Chapman AB, Scharfstein DO, Montgomery AE, et al. Using natural language processing to study homelessness longitudinally with electronic health record data subject to irregular observations. AMIA Annu Symp Proc 2023;2023:894-903. [Medline: 38222404]
76. Ravichandran U, Jungst D, Kwan E. Implementing an NLP Tool to Address SDOH Needs: Institute of Electrical and Electronics Engineers Inc; 2023. [doi: 10.1109/ICHI57859.2023.00091]
77. Zhao X, Rios A. A marker-based neural network system for extracting social determinants of health. J Am Med Inform Assoc 2023 Jul 19;30(8):1398-1407. [doi: 10.1093/jamia/ocad041] [Medline: 37011635]
78. Lybarger K, Dobbins NJ, Long R, et al. Leveraging natural language processing to augment structured social determinants of health data in the electronic health record. J Am Med Inform Assoc 2023 Jul 19;30(8):1389-1397. [doi: 10.1093/jamia/ocad073] [Medline: 37130345]
79. Stewart de Ramirez S, Shallat J, McClure K, Foulger R, Barenblat L. Screening for social determinants of health: active and passive information retrieval methods. Popul Health Manag 2022 Dec;25(6):781-788. [doi: 10.1089/pop.2022.0228]
80. Lybarger K, Yetisgen M, Uzuner Ö. The 2022 n2c2/UW shared task on extracting social determinants of health. J Am Med Inform Assoc 2023 Jul 19;30(8):1367-1378. [doi: 10.1093/jamia/ocad012] [Medline: 36795066]
81. Wang S, Dang Y, Sun Z, et al. An NLP approach to identify SDoH-related circumstance and suicide crisis from death investigation narratives. J Am Med Inform Assoc 2023 Jul 19;30(8):1408-1417. [doi: 10.1093/jamia/ocad068]
82. Sajdeya R, Mardini MT, Tighe PJ, et al. Developing and validating a natural language processing algorithm to extract preoperative cannabis use status documentation from unstructured narrative clinical notes. J Am Med Inform Assoc 2023 Jul 19;30(8):1418-1428. [doi: 10.1093/jamia/ocad080] [Medline: 37178155]
83. Chapman AB, Cordasco K, Chassman S, et al. Assessing longitudinal housing status using electronic health record data: a comparison of natural language processing, structured data, and patient-reported history. Front Artif Intell 2023;6. [doi: 10.3389/frai.2023.1187501]
84. Lelkes AD, Loreaux E, Schuster T, Chen MJ, Rajkomar A. SDOH-NLI: a dataset for inferring social determinants of health from clinical notes. Presented at: Findings of the Association for Computational Linguistics; Dec 6-10, 2023; Singapore p. 4789-4798. [doi: 10.18653/v1/2023.findings-emnlp.317]
85. Romanowski B, Ben Abacha A, Fan Y. Extracting social determinants of health from clinical note text with classification and sequence-to-sequence approaches. J Am Med Inform Assoc 2023 Jul 19;30(8):1448-1455. [doi: 10.1093/jamia/ocad071] [Medline: 37100768]
86. Ramachandran GK, Fu Y, Han B, et al. Prompt-based extraction of social determinants of health using few-shot learning. Presented at: Proceedings of the 5th Clinical Natural Language Processing Workshop; Jul 14, 2023. [doi: 10.18653/v1/2023.clinicalnlp-1.41]
87. Bashir SR, Raza S, Kocaman V, Qamar U. Clinical application of detecting COVID-19 risks: a natural language processing approach. Viruses 2022 Dec 11;14(12):2761. [doi: 10.3390/v14122761] [Medline: 36560764]
88. Lee K, Han S, Suh HS. Early impact of COVID-19 social distancing on social determinants of health and their effects on mental health and quality of life of Korean undergraduate students. Front Public Health 2023;11. [doi: 10.3389/fpubh.2023.1197143]
89. Mitra A, Pradhan R, Melamed RD, et al. Associations between natural language processing-enriched social determinants of health and suicide death among US veterans. JAMA Netw Open 2023 Mar 1;6(3):e233079. [doi: 10.1001/jamanetworkopen.2023.3079] [Medline: 36920391]
90. Yao Z, Tsai J, Liu W, et al. Automated identification of eviction status from electronic health record notes. J Am Med Inform Assoc 2023 Jul 19;30(8):1429-1437. [doi: 10.1093/jamia/ocad081] [Medline: 37203429]
91. Allen KS, Hood DR, Cummins J, Kasturi S, Mendonca EA, Vest JR. Natural language processing-driven state machines to extract social factors from unstructured clinical documentation. JAMIA Open 2023 Apr 6;6(2). [doi: 10.1093/jamiaopen/ooad024]
92. Cascalheira CJ, Chapagain S, Flinn RE, et al. Predicting linguistically sophisticated social determinants of health disparities with neural networks: the case of LGBTQ+ minority stress. Presented at: 2023 IEEE International Conference on Big Data (BigData); Dec 15-18, 2023. [doi: 10.1109/BigData59044.2023.10386882]
93. Chilman N, Song X, Roberts A, et al. Text mining occupations from the mental health electronic health record: a natural language processing approach using records from the clinical record interactive search (CRIS) platform in south London, UK. BMJ Open 2021 Mar;11(3):e042274. [doi: 10.1136/bmjopen-2020-042274]
94. Chapman AB, Jones A, Kelley AT, et al. ReHouSED: a novel measurement of veteran housing stability using natural language processing. J Biomed Inform 2021 Oct;122:103903. [doi: 10.1016/j.jbi.2021.103903] [Medline: 34474188]
95. Mitra A, Ahsan H, Li W, et al. Risk factors associated with nonfatal opioid overdose leading to intensive care unit admission: a cross-sectional study. JMIR Med Inform 2021;9(11):e32851. [doi: 10.2196/32851]

XSL•FO
RenderX

96. Sy B, Wassil M, Connelly H, Hassan A. Behavioral predictive analytics towards personalization for self-management: a use case on linking health-related social needs. SN Comput Sci 2022;3(3):237. [doi: 10.1007/s42979-022-01092-2] [Medline: 35493988]
97. Teng A, Wilcox A. Simplified data science approach to extract social and behavioural determinants: a retrospective chart review. BMJ Open 2022 Jan 18;12(1):e048397. [doi: 10.1136/bmjopen-2020-048397] [Medline: 35042703]
98. Hatef E, Singh Deol G, Rouhizadeh M, et al. Measuring the value of a practical text mining approach to identify patients with housing issues in the free-text notes in electronic health record: findings of a retrospective cohort study. Front Public Health 2021;9:697501. [doi: 10.3389/fpubh.2021.697501] [Medline: 34513783]
99. Hatef E, Rouhizadeh M, Nau C, et al. Development and assessment of a natural language processing model to identify residential instability in electronic health records' unstructured data: a comparison of 3 integrated healthcare delivery systems. JAMIA Open 2022 Apr;5(1):ooac006. [doi: 10.1093/jamiaopen/ooac006] [Medline: 35224458]
100. Raza S, Schwartz B. Detecting biomedical named entities in covid-19 texts. Presented at: Proceedings of the 1st Workshop on Healthcare AI and COVID-19; Jul 22, 2022; Hawaii p. 117-126 URL: https://proceedings.mlr.press/v184/raza22a.html
101. Datar S, Lindemann EA, Silverman G, et al. Identifying Mentions of Life Stressors in Clinical Notes: Institute of Electrical and Electronics Engineers Inc; 2021:153-160. [doi: 10.1109/ICHI52183.2021.00033]
102. Reeves RM, Christensen L, Brown JR, et al. Adaptation of an NLP system to a new healthcare environment to identify social determinants of health. J Biomed Inform 2021 Aug;120:103851. [doi: 10.1016/j.jbi.2021.103851] [Medline: 34174396]
103. Olusanya OA, Ammar N, Davis RL, Bednarczyk RA, Shaban-Nejad A. A digital personal health library for enabling precision health promotion to prevent human papilloma virus-associated cancers. Front Digit Health 2021;3:683161. [doi: 10.3389/fdgth.2021.683161] [Medline: 34713154]
104. Poulsen MN, Freda PJ, Troiani V, Davoudi A, Mowery DL. Classifying characteristics of opioid use disorder from hospital discharge summaries using natural language processing. Front Public Health 2022;10:850619. [doi: 10.3389/fpubh.2022.850619] [Medline: 35615042]
105. Dorr DA, Quiñones AR, King T, Wei MY, White K, Bejan CA. Prediction of future health care utilization through note-extracted psychosocial factors. Med Care 2022 Aug 1;60(8):570-578. [doi: 10.1097/MLR.0000000000001742] [Medline: 35658116]
106. Yu Z, Yang X, Dang C, et al. A study of social and behavioral determinants of health in lung cancer patients using transformers-based natural language processing models. AMIA Annu Symp Proc 2021;2021:1225-1233. [Medline: 35309014]
107. Tsui FR, Shi L, Ruiz V, et al. Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. JAMIA Open 2021 Jan;4(1):ooab011. [doi: 10.1093/jamiaopen/ooab011] [Medline: 33758800]
108. Yu Z, Yang X, Guo Y, Bian J, Wu Y. Assessing the documentation of social determinants of health for lung cancer patients in clinical narratives. Front Public Health 2022;10:778463. [doi: 10.3389/fpubh.2022.778463] [Medline: 35419333]
109. Hatef E, Rouhizadeh M, Tia I, et al. Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: a retrospective analysis of a multilevel health care system. JMIR Med Inform 2019 Aug 2;7(3):e13802. [doi: 10.2196/13802] [Medline: 31376277]
110. Bettencourt-Silva JH, Mulligan N, Sbodio M, et al. Discovering New Social Determinants of Health Concepts from Unstructured Data: Framework and Evaluation: IOS Press; 2020:173-177. [doi: 10.3233/SHTI200145]
111. Bucher BT, Shi J, Pettit RJ, Ferraro J, Chapman WW, Gundlapalli A. Determination of marital status of patients from structured and unstructured electronic healthcare data. AMIA Annu Symp Proc 2019;2019:267-274. [Medline: 32308819]
112. Sebestyén V, Domokos E, Abonyi J. Focal points for sustainable development strategies-text mining-based comparative analysis of voluntary national reviews. J Environ Manage 2020 Jun 1;263:110414. [doi: 10.1016/j.jenvman.2020.110414] [Medline: 32174539]
113. Cho SM, Park CU, Song M. The evolution of social health research topics: a data-driven analysis. Soc Sci Med 2020 Nov;265:113299. [doi: 10.1016/j.socscimed.2020.113299] [Medline: 32905964]
114. Conway M, Keyhani S, Christensen L, et al. Moonstone: a novel natural language processing system for inferring social risk from clinical narratives. J Biomed Semant 2019 Dec;10(1):2019. [doi: 10.1186/s13326-019-0198-0]
115. Dorr D, Bejan CA, Pizzimenti C, Singh S, Storer M, Quinones A. Identifying Patients with Significant Problems Related to Social Determinants of Health with Natural Language Processing: IOS Press; 2019:1456-1457. [doi: 10.3233/SHTI190482]
116. Du X, Bian J, Prosperi M. An Operational Deep Learning Pipeline for Classifying Life Events from Individual Tweets: Springer Verlag; 2019:54-66. [doi: 10.1007/978-3-030-11680-4_7]
117. DuBay DA, Su Z, Morinelli TA, et al. Development and future deployment of a 5 years allograft survival model for kidney transplantation. Nephrology (Carlton) 2019 Aug;24(8):855-862. [doi: 10.1111/nep.13488] [Medline: 30198104]
118. Lybarger K, Ostendorf M, Yetisgen M. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. J Biomed Inform 2021 Jan;113:103631. [doi: 10.1016/j.jbi.2020.103631]
119. Zhu VJ, Lenert LA, Bunnell BE, Obeid JS, Jefferson M, Halbert CH. Automatically identifying social isolation from clinical narratives for patients with prostate cancer. BMC Med Inform Decis Mak 2019 Mar 14;19(1):43. [doi: 10.1186/s12911-019-0795-y] [Medline: 30871518]
120. Bettencourt-Silva JH, Mulligan N, Jochim C, et al. Exploring the Social Drivers of Health during a Pandemic: Leveraging Knowledge Graphs and Population Trends in COVID-19: IOS Press BV; 2020:6-11. [doi: 10.3233/SHTI200684]

XSL•FO
RenderX

121. Zhang X, Bellolio MF, Medrano-Gracia P, Werys K, Yang S, Mahajan P. Use of natural language processing to improve predictive models for imaging utilization in children presenting to the emergency department. BMC Med Inform Decis Mak 2019 Dec 30;19(1):287. [doi: 10.1186/s12911-019-1006-6] [Medline: 31888609]
122. Feller DJ, Zucker J, Don't Walk OB 4th, et al. Towards the inference of social and behavioral determinants of sexual health: development of a gold-standard corpus with semi-supervised learning. AMIA Annu Symp Proc 2018;2018:422-429. [Medline: 30815082]
123. Bejan CA, Angiolillo J, Conway D, et al. Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. J Am Med Inform Assoc 2018 Jan 1;25(1):61-71. [doi: 10.1093/jamia/ocx059]
124. Lindemann EA, Chen ES, Wang Y, Skube SJ, Melton GB. Representation of social history factors across age groups: a topic analysis of free-text social documentation. AMIA Annu Symp Proc 2017;2017:1169-1178. [Medline: 29854185]
125. Velupillai S, Mowery DL, Conway M, Hurdle J, Kious B. Vocabulary development to support information extraction of substance abuse from psychiatry notes. Presented at: Proceedings of the 15th Workshop on Biomedical Natural Language Processing; Aug 12, 2016. [doi: 10.18653/v1/W16-2912]
126. Relia K, Akbari M, Duncan D, Chunara R. Socio-spatial self-organizing maps: using social media to assess relevant geographies for exposure to social processes. Proc ACM Hum Comput Interact 2018 Nov;2:145. [doi: 10.1145/3274414] [Medline: 30957076]
127. Navathe AS, Zhong F, Lei VJ, et al. Hospital readmission and social risk factors identified from physician notes. Health Serv Res 2018 Apr;53(2):1110-1136. [doi: 10.1111/1475-6773.12670] [Medline: 28295260]
128. Peng C, Yang X, Chen A, et al. Generative large language models are all-purpose text analytics engines: text-to-text learning is all your need. J Am Med Inform Assoc 2024 Sep 1;31(9):1892-1903. [doi: 10.1093/jamia/ocae078] [Medline: 38630580]
129. Gleize M, Mulligan N, Bettencourt-Silva JH. Social Determinant Trends of COVID-19: An Analysis Using Knowledge Graphs from Published Evidence and Online Trends 2021:744-748. [doi: 10.3233/SHTI210271]
130. Han S, Zhang RF, Shi L, et al. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. J Biomed Inform 2022 Mar;127:103984. [doi: 10.1016/j.jbi.2021.103984]
131. Suda-King C, Winch L, Tucker JM, Zuehlke AD, Hunter C, Simmons JM. Representation of social determinants of health terminology in medical subject headings: impact of added terms. J Am Med Inform Assoc 2024 Nov 1;31(11):2595-2604. [doi: 10.1093/jamia/ocae191] [Medline: 39047296]
132. Bejan CA, Ripperger M, Wilimitis D, et al. Improving ascertainment of suicidal ideation and suicide attempt with natural language processing. Sci Rep 2022 Sep 7;12(1):15146. [doi: 10.1038/s41598-022-19358-3] [Medline: 36071081]
133. Chandler MT, Cai TR, Santacroce L, Ulysse S, Liao KP, Feldman CH. Classifying individuals with rheumatic conditions as financially insecure using electronic health record data and natural language processing: algorithm derivation and validation. ACR OPEN Rheumatol 2024 Aug;6(8):481-488. [doi: 10.1002/acr2.11675] [Medline: 38747148]
134. Boch S, Hussain SA, Bambach S, DeShetler C, Chisolm D, Linwood S. Locating youth exposed to parental justice involvement in the electronic health record: development of a natural language processing model. JMIR Pediatr Parent 2022 Mar 21;5(1):e33614. [doi: 10.2196/33614] [Medline: 35311681]
135. Mahbub M, Goethert I, Danciu I, et al. Question-answering system extracts information on injection drug use from clinical notes. Commun Med 2024;4(1). [doi: 10.1038/s43856-024-00470-6]
136. Wray CM, Vali M, Walter LC, et al. Examining the interfacility variation of social determinants of health in the veterans health administration. Fed Pract 2021 Jan;38(1):15-19. [doi: 10.12788/fp.0080] [Medline: 33574644]
137. Bhate NJ, Mittal A, He Z, Luo X. Zero-shot learning with minimum instruction to extract social determinants and family history from clinical notes using GPT model. Proc IEEE Int Conf Big Data 2023 Dec;2023:1476-1480. [doi: 10.1109/BigData59044.2023.10386811] [Medline: 39101057]
138. Dai HJ, Su ECY, Uddin M, Jonnagaddala J, Wu CS, Syed-Abdul S. Exploring associations of clinical and social parameters with violent behaviors among psychiatric patients. J Biomed Inform 2017 Nov;75S:S149-S159. [doi: 10.1016/j.jbi.2017.08.009] [Medline: 28822857]
139. Purpura A, Mulligan N, Kartoun U, Koski E, Anand V, Bettencourt-Silva J. Investigating cross-domain binary relation classification in biomedical natural language processing. AMIA Jt Summits Transl Sci Proc 2024;2024:384-390. [Medline: 38827064]
140. Krishnamoorthy R, Nagarajan V, Pour H, et al. Voice-enabled response analysis agent (VERAA): leveraging large language models to map voice responses in SDoH survey. AMIA Jt Summits Transl Sci Proc 2024;2024:258-265. [Medline: 38827075]
141. Wang EA, Long JB, McGinnis KA, et al. Measuring exposure to incarceration using the electronic health record. Med Care 2019 Jun;57 Suppl 6 Suppl 2(Suppl 6 2):S157-S163. [doi: 10.1097/MLR.0000000000001049] [Medline: 31095055]
142. Zhu V, Lenert L, Bunnell B, Obeid J, Jefferson M, Halbert CH. Automatically identifying financial stress information from clinical notes for patients with prostate cancer. Cancer Res Rep 2020;1(1):102. [doi: 10.61545/crr-1-102] [Medline: 38317775]
143. Subramani S, Wang H, Vu HQ, Li G. Domestic violence crisis identification from facebook posts based on deep learning. IEEE Access 2018;6:54075-54085. [doi: 10.1109/ACCESS.2018.2871446]

XSL•FO
RenderX

144. Subramani S, Michalska S, Wang H, Du J, Zhang Y, Shakeel H. Deep learning for multi-class identification from domestic violence online posts. IEEE Access 2019;7:46210-46224. [doi: 10.1109/ACCESS.2019.2908827]
145. Liyanage C, Garg M, Mago V, Sohn S. Augmenting Reddit Posts to Determine Wellness Dimensions Impacting Mental Health: Association for Computational Linguistics; 2023:306-312. [doi: 10.18653/v1/2023.bionlp-1.27]
146. Martin EA, D'Souza AG, Saini V, Tang K, Quan H, Eastwood CA. Extracting social determinants of health from inpatient electronic medical records using natural language processing. J Epidemiol Popul Health 2024 Dec;72(6):202791. [doi: 10.1016/j.jeph.2024.202791] [Medline: 39546940]
147. Gabriel RA, Litake O, Simpson S, Burton BN, Waterman RS, Macias AA. On the development and validation of large language model-based classifiers for identifying social determinants of health. Proc Natl Acad Sci USA 2024 Sep 24;121(39). [doi: 10.1073/pnas.2320716121]
148. Ralevski A, Taiyab N, Nossal M, Mico L, Piekos S, Hadlock J. Using large language models to abstract complex social determinants of health from original and deidentified medical notes: development and validation study. J Med Internet Res 2024;26:e63445. [doi: 10.2196/63445]
149. Gong L, Bresnick J, Zhang A, Wu C, Jha K. Boosting social determinants of health extraction with semantic knowledge augmented large language model. AMIA Annu Symp Proc 2024;2024:453-462. [Medline: 40417469]
150. Moon S, Wu Y, Doughty JB, et al. Automated identification of patients' unmet social needs in clinical text using natural language processing. Mayo Clin Proc Digit Health 2024 Sep;2(3):411-420. [doi: 10.1016/j.mcpdig.2024.06.008] [Medline: 39324128]
151. Roy S, Morrell S, Zhao L, Homayouni R. Large-scale identification of social and behavioral determinants of health from clinical notes: comparison of Latent Semantic Indexing and Generative Pretrained Transformer (GPT) models. BMC Med Inform Decis Mak 2024 Oct 10;24(1):296. [doi: 10.1186/s12911-024-02705-x] [Medline: 39390479]
152. Gong L, Shor A, Zhang A, Jha K. Context-specific feature augmentation for improving social determinants of health extraction. Presented at: 2024 IEEE International Conference on Big Data (BigData); Dec 15-18, 2024; Washington, DC, USA p. 1736-1745. [doi: 10.1109/BigData62323.2024.10825225]
153. Roosan D, Chok J, Li Y, Khou T. Utilizing quantum computing-based large language transformer models to identify social determinants of health from electronic health records. Presented at: 2024 International Conference on Electrical, Computer and Energy Technologies (ICECET); Jul 25-27, 2024. [doi: 10.1109/ICECET61485.2024.10698600]
154. Shang T, Yang S, He W, et al. Leveraging social determinants of health in Alzheimer's research using LLM-augmented literature mining and knowledge graphs. AMIA Jt Summits Transl Sci Proc 2025;2025:491-500. [Medline: 40502260]
155. Keloth VK, Selek S, Chen Q, et al. Social determinants of health extraction from clinical notes across institutions using large language models. NPJ Digit Med 2025 May 17;8(1):287. [doi: 10.1038/s41746-025-01645-8] [Medline: 40379919]
156. Consoli B, Wang H, Wu X, et al. SDoH-GPT: using large language models to extract social determinants of health. J Am Med Inform Assoc 2026 Jan 1;33(1):67-78. [doi: 10.1093/jamia/ocaf094] [Medline: 40493530]
157. Gu Z, He L, Naeem A, et al. SBDH-Reader: a large language model-powered method for extracting social and behavioral determinants of health from clinical notes. J Am Med Inform Assoc 2025 Oct 1;32(10):1570-1580. [doi: 10.1093/jamia/ocaf124] [Medline: 40720632]
158. Chen Z, Lasserre P, Lin A, Rajapakshe R. Extraction of social determinants of health from electronic health records using natural language processing. JCO Clin Cancer Inform 2025 Jul;9:e2400317. [doi: 10.1200/CCI-24-00317] [Medline: 40700678]
159. Gu B, Shao V, Liao Z, et al. Scalable information extraction from free text electronic health records using large language models. BMC Med Res Methodol 2025;25(1). [doi: 10.1186/s12874-025-02470-z]
160. Wang Y, Hilsman J, Li C, et al. Development and validation of natural language processing algorithms in the national ENACT network. J Clin Trans Sci 2025;9(1). [doi: 10.1017/cts.2025.10116]
161. Peng C, Yu Z, Smith KE, Lo-Ciganic WH, Bian J, Wu Y. Enhancing cross-domain generalizability in social determinants of health extraction with prompt-tuning large language models. AMIA Jt Summits Transl Sci Proc 2025;2025:432-440. [Medline: 40502248]
162. Shao M, Kang Y, Hu X, Kwak HG, Yang C, Lu J. Mining social determinants of health for heart failure patient 30-day readmission via large language model. Stud Health Technol Inform 2025 Aug 7;329:1902-1903. [doi: 10.3233/SHTI251272] [Medline: 40776289]
163. Zaribafzadeh H, Henson JB, Chan NW, et al. Development of a natural language processing algorithm to extract social determinants of health from clinician notes. Am J Transplant 2025 Jun;25(6):1306-1318. [doi: 10.1016/j.ajt.2025.02.019]
164. Wang S, Wei Y, Ma H, et al. A multi-stage large language model framework for extracting suicide-related social determinants of health. Commun Med 2025;5(1). [doi: 10.1038/s43856-025-01114-z]
165. Chapman AB, Panadero T, Dalrymple R, et al. Studying Veteran food insecurity longitudinally using electronic health record data and natural language processing. AMIA Jt Summits Transl Sci Proc 2025 Jan 16:124-133. [Medline: 40502216]
166. Scherbakov D, Heider PM, Wehbe R, Alekseyenko AV, Lenert LA, Obeid JS. Using large language models for extracting stressful life events to assess their impact on preventive colon cancer screening adherence. BMC Public Health 2025;25(1). [doi: 10.1186/s12889-024-21123-2]

XSL•FO
RenderX

167. Wasser LM, Liang HW, Li C, et al. Identifying transportation needs in ophthalmology clinic notes using natural language processing: retrospective, cross-sectional study. JMIR Med Inform 2025 Sep 5;13:e69216. [doi: 10.2196/69216] [Medline: 40913248]
168. Patra BG, Lepow LA, Kasi Reddy Jagadeesh Kumar P, et al. Extracting social support and social isolation information from clinical psychiatry notes: comparing a rule-based natural language processing system and a large language model. J Am Med Inform Assoc 2025 Jan 1;32(1):218-226. [doi: 10.1093/jamia/ocae260] [Medline: 39423850]
169. Kim MH, Miramontes S, Mehta S, et al. Extracting housing and food insecurity information from clinical notes using cTAKES. Health Serv Res 2025 May;60 Suppl 3(Suppl 3):e14440. [doi: 10.1111/1475-6773.14440] [Medline: 39871689]
170. Sakib FA, Zhu Z, Grace KT, Yetisgen M, Uzuner O. Spurious Correlations and Beyond: Understanding and Mitigating Shortcut Learning in SDOH Extraction with Large Language Models: Association for Computational Linguistics; 2025:1097-1106. [doi: 10.18653/v1/2025.acl-short.86]
171. Xu N, Zhang Q, Du C, et al. Revealing emergent human-like conceptual representations from language prediction. Proc Natl Acad Sci U S A 2025 Nov 4;122(44):e2512514122. [doi: 10.1073/pnas.2512514122] [Medline: 41171846]
172. Li C, Guo J, Bian J, Becich MJ. Advancing social determinants of health research and practice: data, tools, and implementation. J Clin Transl Sci 2025;9(1):e58. [doi: 10.1017/cts.2024.657] [Medline: 40201637]
173. Islam S, Elmekki H, Elsebai A, et al. A comprehensive survey on applications of transformers for deep learning tasks. Expert Syst Appl 2024 May;241:122666. [doi: 10.1016/j.eswa.2023.122666]
174. Gong EJ, Bang CS, Lee JJ, Baik GH. Knowledge-practice performance gap in clinical large language models: systematic review of 39 benchmarks. J Med Internet Res 2025 Dec 1;27:e84120. [doi: 10.2196/84120] [Medline: 41325597]
175. Goldstein ND, Olivieri-Mui B, Burstyn I. Are aggregated electronic health record datasets good for research? J Gen Intern Med 2025 Nov;40(15):3743-3749. [doi: 10.1007/s11606-025-09808-9] [Medline: 40794368]
176. Li C, Mowery DL, Ma X, et al. Realizing the potential of social determinants data in EHR systems: a scoping review of approaches for screening, linkage, extraction, analysis, and interventions. J Clin Transl Sci 2024;8(1):e147. [doi: 10.1017/cts.2024.571] [Medline: 39478779]
177. Alberto IRI, Alberto NRI, Ghosh AK, et al. The impact of commercial health datasets on medical research and health-care algorithms. Lancet Digit Health 2023 May;5(5):e288-e294. [doi: 10.1016/S2589-7500(23)00025-0] [Medline: 37100543]
178. Artsi Y, Sorin V, Glicksberg BS, Korfiatis P, Nadkarni GN, Klang E. Large language models in real-world clinical workflows: a systematic review of applications and implementation. Front Digit Health 2025;7:1659134. [doi: 10.3389/fdgth.2025.1659134] [Medline: 41098649]
179. Calderon N, Reichart R. On behalf of the stakeholders: trends in nlp model interpretability in the era of llms. Presented at: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics; Apr 29 to May 4, 2025; Albuquerque, New Mexico p. 656-693 URL: https://aclanthology.org/2025.naacl-long [doi: 10.18653/v1/2025.naacl-long.29]
180. Senghor AS, Bright TJ, Kakim S, et al. A community-based approach to ethical decision-making in artificial intelligence for health care. JAMIA Open 2025 Aug;8(4):ooaf076. [doi: 10.1093/jamiaopen/ooaf076] [Medline: 40799930]
181. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform 2021 Jan;113:103655. [doi: 10.1016/j.jbi.2020.103655] [Medline: 33309898]
182. Feller DJ, Zucker J, Yin MT, Gordon P, Elhadad N, Walk OBD. Longitudinal analysis of social and behavioral determinants of health in the ehr: exploring the impact of patient trajectories and documentation practices. Presented at: Proceedings of the AMIA Annual Symposium; Mar 4, 2020; Washington, DC p. 399-407. [Medline: 32308833]
183. Hernán MA, McAdams M, McGrath N, Lanoy E, Costagliola D. Observation plans in longitudinal studies with time-varying treatments. Stat Methods Med Res 2009 Feb;18(1):27-52. [doi: 10.1177/0962280208092345] [Medline: 19036915]
184. Chiu CC, Wu CM, Chien TN, Kao LJ, Li C, Chu CM. Integrating structured and unstructured EHR data for predicting mortality by machine learning and latent dirichlet allocation method. Int J Environ Res Public Health 2023 Feb 28;20(5):4340. [doi: 10.3390/ijerph20054340] [Medline: 36901354]
185. Simon BD, Ozyoruk KB, Gelikman DG, Harmon SA, Türkbey B. The future of multimodal artificial intelligence models for integrating imaging and clinical metadata: a narrative review. Diagn Interv Radiol 2025 Jul 8;31(4):303-312. [doi: 10.4274/dir.2024.242631] [Medline: 39354728]
186. Berg K, Doktorchik C, Quan H, Saini V. Automating data collection methods in electronic health record systems: a social determinant of health (SDOH) viewpoint. Health Syst (Basingstoke) 2023;12(4):472-480. [doi: 10.1080/20476965.2022.2075796] [Medline: 38235302]
187. Egede LE, Walker RJ, Williams JS. Addressing structural inequalities, structural racism, and social determinants of health: a vision for the future. J Gen Intern Med 2024 Feb;39(3):487-491. [doi: 10.1007/s11606-023-08426-7] [Medline: 37740168]
188. Yelton B, Rumthao JR, Sakhuja M, et al. Assessment and documentation of social determinants of health among health care providers: qualitative study. JMIR Form Res 2023 Jul 3;7:e47461. [doi: 10.2196/47461] [Medline: 37399064]
189. Eisinger-Mathason TSK, Leshin J, Lahoti V, Fridsma DB, Mucaj V, Kho AN. Data linkage multiplies research insights across diverse healthcare sectors. Commun Med (Lond) 2025 Mar 4;5(1):58. [doi: 10.1038/s43856-025-00769-y] [Medline: 40038513]

XSL•FO
RenderX

190. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: implications for clinical decision-making. PLOS Digit Health 2024 Nov;3(11):e0000651. [doi: 10.1371/journal.pdig.0000651] [Medline: 39509461]
191. Chen Y, Clayton EW, Novak LL, Anders S, Malin B. Human-centered design to address biases in artificial intelligence. J Med Internet Res 2023 Mar 24;25:e43251. [doi: 10.2196/43251] [Medline: 36961506]
192. Sasseville M, Ouellet S, Rhéaume C, et al. Bias mitigation in primary health care artificial intelligence models: scoping review. J Med Internet Res 2025 Jan 7;27:e60269. [doi: 10.2196/60269] [Medline: 39773888]
193. Hasanzadeh F, Josephson CB, Waters G, Adedinsewo D, Azizi Z, White JA. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. NPJ Digit Med 2025 Mar 11;8(1):154. [doi: 10.1038/s41746-025-01503-7] [Medline: 40069303]
194. Rajwal S, Garg S, Abdel-Salam R, Zayed A. Do biased models have biased thoughts. Presented at: Second Conference on Language Modeling; Oct 7-10, 2025 URL: https://openreview.net/forum?id=vDr0RV3590 [accessed 2026-03-03]
195. Humphreys BL, Del Fiol G, Xu H. The UMLS knowledge sources at 30: indispensable to current research and applications in biomedical informatics. J Am Med Inform Assoc 2020 Oct 1;27(10):1499-1501. [doi: 10.1093/jamia/ocaa208] [Medline: 33059366]
196. Miller K, Moon S, Fu S, Liu H. Contextual variation of clinical notes induced by EHR migration. AMIA Annu Symp Proc 2023;2023:1155-1164. [Medline: 38222426]
197. Rockenschaub P, Hilbert A, Kossen T, et al. The impact of multi-institution datasets on the generalizability of machine learning prediction models in the ICU. Crit Care Med 2024;52(11):1710-1721. [doi: 10.1097/CCM.0000000000006359]

Abbreviations

EHR: electronic health record
LLM: large language model
MIMIC: Medical Information Mart for Intensive Care
MiSSoM+: Minority Stress on Social Media
NCR: normalized citation ratio
NLP: natural language processing
PLUTO: Primary Land Use Tax Lot Output
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SDOH: social determinants of health

Edited by S Brini; submitted 08.Sep.2025; peer-reviewed by LP Gorrepati, O Akinsola, R Marshall, R Abi; revised version received 09.Mar.2026; accepted 09.Mar.2026; published 28.Apr.2026.

Please cite as:
Rajwal S, Pandey AK, Zhang Z, Chen Y, Liu MX, Das S, Rogers H, Sarker A, Xiao Y
Applications of Natural Language Processing and Large Language Models for Social Determinants of Health: Systematic Review
J Med Internet Res 2026;28:e83793
URL: https://www.jmir.org/2026/1/e83793
doi:10.2196/83793

© Swati Rajwal, Avinash Kumar Pandey, Ziyuan Zhang, Yankai Chen, Michael X Liu, Sudeshna Das, Hannah Rogers, Abeed Sarker, Yunyu Xiao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Machine Learning in HIV Care and Antiretroviral Therapy: Systematic Review

Thamina Boudra[1*], VetD, MSc; Arafate Idrissou[2*], MSc; Oussama Barakat[2], PhD; Siamak Davani[3], PhD, MD; Marie-Blanche Valnet Rabier[1*], PhD, PharmD; Jennifer Lagoutte-Renosi[3*], PhD, PharmD

[1]Université Marie et Louis Pasteur, CHU Besançon, SINERGIES (UR 4662), Centre Régional de Pharmacovigilance, Besançon, France

[2]Université Marie et Louis Pasteur, SINERGIES (UR 4662), Besançon, France

[3]Université Marie et Louis Pasteur, CHU Besançon, SINERGIES (UR 4662), 19 rue Ambroise Paré, Besançon, France

[*]these authors contributed equally

Corresponding Author:
Jennifer Lagoutte-Renosi, PhD, PharmD
Université Marie et Louis Pasteur, CHU Besançon, SINERGIES (UR 4662), 19 rue Ambroise Paré, Besançon, France

Abstract

Background: Artificial intelligence (AI) is expanding across various medical fields, with machine learning (ML) being increasingly used to enhance patient management in diagnosis, prevention, and therapeutic care.

Objective: This study aims to provide an overview of ML applications in HIV care, focusing on real clinical data to improve health care for people living with HIV and on antiretroviral therapy, while highlighting unexplored areas.

Methods: Following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 reporting guidelines, we analyzed four databases: PubMed, Embase, IEEE, and Web of Science until August 31, 2024. The keywords used were: “Machine Learning,” “HIV,” and “Antiretroviral Therapy.” We excluded from this review studies (1) that were not directly focused on HIV or those that did not apply ML to real clinical data, (2) that focused on pre-exposure prophylaxis, (3) studies involving in silico antiretroviral drug development, and (4) studies on the biological mechanisms underlying HIV diagnosis. Three experts (TB, MBVR, and JLR) screened each article independently.

Results: Overall, 476 studies were identified, and after eligibility assessment, 98 were finally analyzed in detail. Three experts (TB, MBVR, and JLR) identified 6 major categories of ML applications used in the clinical field of HIV: consideration of comorbidities for people living with HIV, predicting drug resistance of the virus, monitoring HIV infection itself, predicting treatment outcomes for people living with HIV, treatment adherence for people living with HIV, and treatment recommendation for clinicians. Random forests emerged as the most used algorithm with 17.49% (43/247), proving effective in identifying biomarkers of metabolic syndrome, genetic features of the HIV envelope, and predicting neurocognitive impairment. Random forests model has several advantages: (1) handle linear, nonlinear data, and missing data, (2) reduce overfitting compared to single trees, (3) robust to noise and outliers, (4) provide feature importance measures, and (5) good generalization ability. Support vector machines demonstrated strong abilities in analyzing the associations between HIV-1 genotypes and resistance phenotypes, predicting virological response to therapy based on HIV genotype, detecting mutations associated with HIV drug resistance , and enhancing computational predictions of resistance from genotype data. Logistic regression appears to be most powerful in predicting various treatment outcomes, including virological failure, adverse events, immune changes in people living with HIV receiving antiretrovirals, and biomarkers of mitochondrial toxicity.

Conclusions: Depending on the field of application, some ML methods are more suitable and adapt better to certain HIV concerns. However, some areas, such as treatment recommendations, treatment adherence, and treatment optimization, still lack AI algorithms and need further exploration, such as therapeutical optimization. The development of new clinical decision-support systems for people living with HIV is the new challenge for the years ahead, and AI represents one of the most promising tools to address it.

(J Med Internet Res 2026;28:e79219) doi:10.2196/79219

KEYWORDS

machine learning; artificial intelligence; HIV; antiretroviral therapy; precision medicine; clinical decision support tools

XSL•FO
RenderX

Introduction

Artificial intelligence (AI) has recently gained significant attention due to its growing applications across various domains and the substantial improvements it has brought. Therefore, AI appears to be essential for improving health care [1,2]. A key discipline within AI is machine learning (ML). While AI is defined as the ability of a computer to perform tasks requiring human-like intelligence, ML focuses on developing statistical methods and algorithms that recognize patterns in datasets, improving prediction and classification accuracy when applied to new data. ML's increasing use in health care is justified mainly by the vast amounts of data generated and the ongoing digitalization of the industry, hospitals, and private practices, particularly through the adoption and exploration of electronic health records (EHRs) [3-6]. However, EHRs are not being fully exploited in the European Union due to technical and regulatory obstacles [7]. In response, many research projects were established to enhance the accessibility of EHR data for citizens and researchers in the European Union and facilitate better use [8,9].

Since the emergence of HIV in the 1980s, this disease has remained a major public health challenge, demanding considerable attention from both the scientific community and health care professionals. With advances in antiretroviral treatment, life expectancy for people living with HIV has increased considerably since the 1980s [10].

The management of people living with HIV remains a complex approach and still generates many inquiries from them and health care professionals. Although new therapeutic families have recently been developed, along with new forms such as intramuscular and subcutaneous injections, transforming the management of HIV infection, this only applies to certain eligible people living with HIV. Challenges remain. In addition, people living with HIV now have a near-normal life expectancy, comparable to that of the general population. The management of HIV infection, therefore, continues to pose challenges, particularly in older patients, who often experience age-related comorbidities leading to complex polypharmacy and increased risk of drug-drug interactions (DDI) and adverse drug reactions [11]. Thus, several therapeutic lines could be prescribed across the medical history of people living with HIV. Moreover, with the increasing volume and diversity of data routinely collected as part of people living with HIV care, managing the HIV landscape has become a complex task. This complexity is exacerbated by the heterogeneity of the information, which ranges from clinical data on types of treatment to biological data related to immune status and viral load to viral genetic signatures and socio-demographic factors, all of which require closer examination to improve care for this population.

Thus, comprehensive HIV care could improve both quality of life and long-term health outcomes. In recent years, ML has demonstrated its use in optimizing and revolutionizing various aspects of health care and research. For example, 2 pivotal applications of AI for clinical purposes are rule-based expert systems and clinical decision support systems. Clinical decision support systems combine clinicians' medical expertise with recent AI advancements to enhance clinical decision-making processes. These systems leverage extensive medical knowledge derived from medical literature, complex algorithms, and patients' EHR data to assist health care professionals in the overall quality of care while ensuring patient safety. These types of applications are time-consuming and rely on expert consensus. At first glance, it appears that there is a lack of use of ML tools in the field of HIV care. This review aims to identify the main applications of ML in HIV data management and to understand current trends. Specialized vocabulary for nonexperts in AI is summarized in Multimedia Appendix 1.

Methods

Search Strategy

An extensive search was conducted across 4 major databases, following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 reporting guidelines [12], until August 31, 2024, using a combination of keywords based on a controlled vocabulary thesaurus for indexing articles in each database: "Machine Learning," "HIV" and "Antiretroviral Therapy."

Eligibility Criteria

Studies were included in the current review if they met all the following criteria: titles and abstracts that mentioned explicitly searching terms and using real clinical data. Noninclusion criteria were studies on in silico drug design of antiretroviral therapy (ART), studies on pre-exposure prophylaxis, research focused on the pathophysiology of HIV infection, and studies discussing advancements in HIV clinical diagnostics (Table 1).

XSL•FO
RenderX

Table . Inclusion and noninclusion criteria.

Category	Inclusion criteria	Noninclusion criteria
Article type	• Title and abstract containing search terms: "Machine Learning" and "HIV" and "Antiretroviral Therapy"	• Studies on pre-exposure prophylaxis • Research focused on the pathophysiology of HIV infection • Studies discussing advancements in HIV clinical diagnostics
Language	• English	• Other languages than English
Population	• People living with HIV	__[a]
Study design	• Real-world data	• In silico drug design of ART[b]

[a]Not available.

[b]ART: antiretroviral therapy.

Study Selection

During the initial screening, 331 records were excluded due to not meeting the research criteria, such as a lack of relevance to HIV or the absence of ML applied to clinical data. Out of 145 articles that advanced to the eligibility phase, 16 were excluded because they were literature reviews or meta-analyses. After successfully retrieving 129 articles, 31 duplicates were eliminated, resulting in 98 articles [6,13-109] that fully met the guidelines and inclusion criteria for this review.

In the clinical field of care for people living with HIV, there are several areas: disease progression, response to treatment, treatment failure, comorbidities, and treatment management. Based on these elements and their reading of the various articles identified, 3 experts (TB, MBVR, and JLR) jointly proposed grouping them into 6 key categories, which are as follows: monitoring HIV infection, predicting treatment outcomes, predicting drug resistance, comorbidities, treatment adherence, and treatment recommendation. Three experts (TB, MBVR, and JLR) screened each article independently. In the event of disagreement, a joint proofreading was carried out to reach a consensus. To facilitate data analysis, we compiled a database that includes the essential details of each study: study title, year of publication, authors, geographical location, and key thematic category, such as comorbidities, prediction of resistance, or treatment adherence, with subcategories for further classification. In addition, we have extracted data on the types of information analyzed (eg, clinical and biological) and the data format (numerical, categorical, and textual). The database also documents study objectives, main results, used AI algorithms and methods, sample size, and authors' countries of origin. For data trends analysis, concerning the origin of the articles, the origin of the first author was chosen.

Data Synthesis and Analysis

The data from the included studies were processed and harmonized using customized pipelines. This step involved grouping categories within categorical variables, standardizing formats, and creating additional variables to facilitate the analysis. After data cleaning and preparation, descriptive analyses were conducted to identify key trends and distributions, which were visualized using graphs and figures to provide a clear overview of the collected data. All data processing and analysis were performed in Python using libraries such as pandas, numpy, and matplotlib.

Results

Overall, 476 studies were collected: 148 from PubMed, 200 from Embase, 16 from IEEE, and 112 from Web of Science (Figure 1).

XSL•FO
RenderX

Figure 1. Flow chart process for study selection using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 reporting guidelines. N is the number of occurrences.

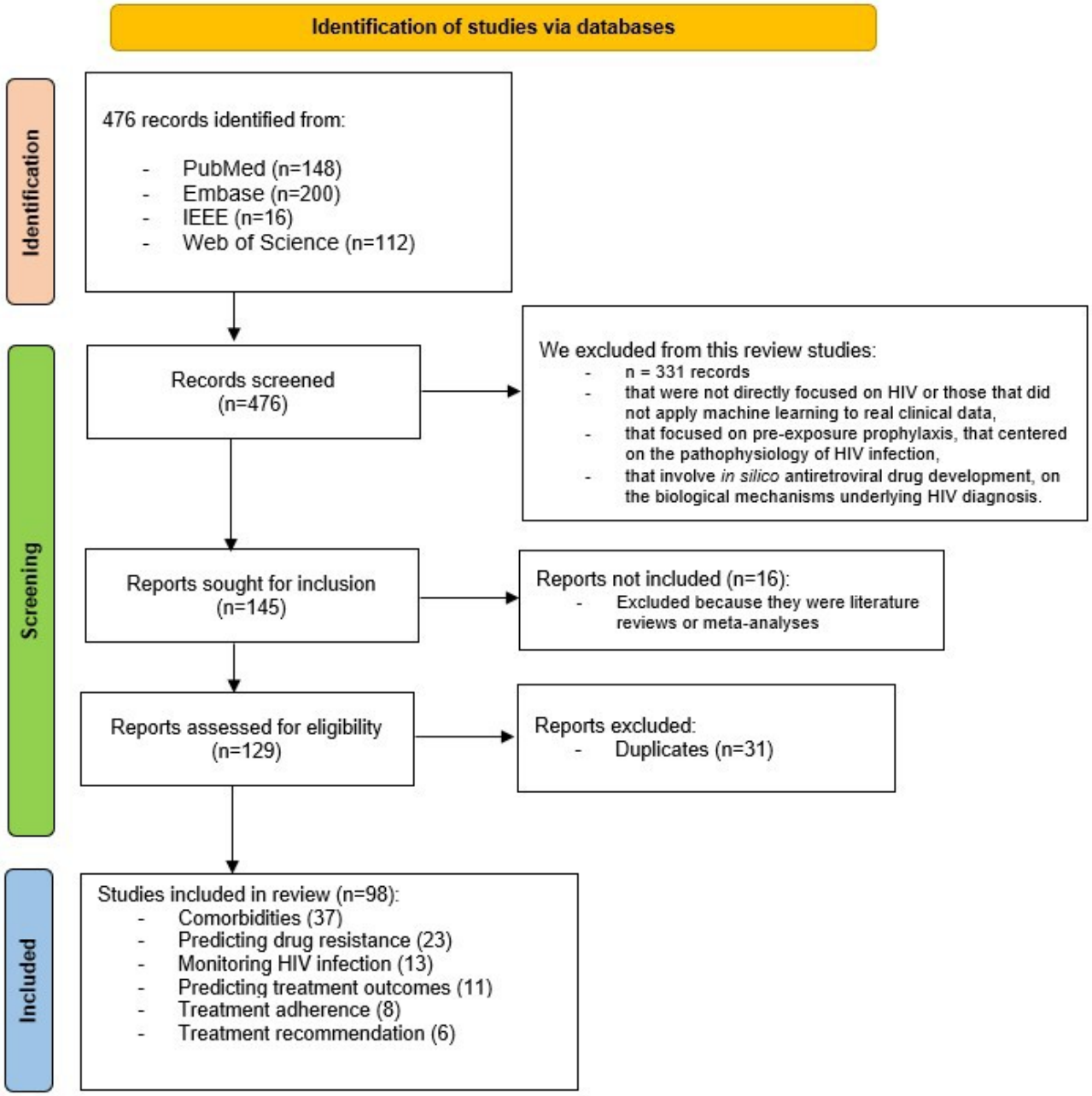

We started our analysis with the literature trends by doing a statistical analysis of publication years, which revealed that, on average, articles were published in 2018, with a mean of 2017.7 (SD 5.63) years. The quartiles provide additional insight: the first quartile indicates that 25% (24/98) of articles were published before 2015. The median year of publication is 2020 (IQR 6.75-7), meaning that half the articles were published after this date. Finally, 25% (25/98) of articles were published after 2020, highlighting a clear trend in favor of more recent studies. These results reflect a strong temporal coverage of our literature review. Regarding the distribution of papers over time, the earliest study on HIV treatment using AI was published in 2002. Although publications remained sporadic until 2005, a noticeable uptick occurred afterwards (Figure 2).

XSL•FO
RenderX

Figure 2. Distributions per year of publications relating to HIV and uses of machine learning over the past twenty years.

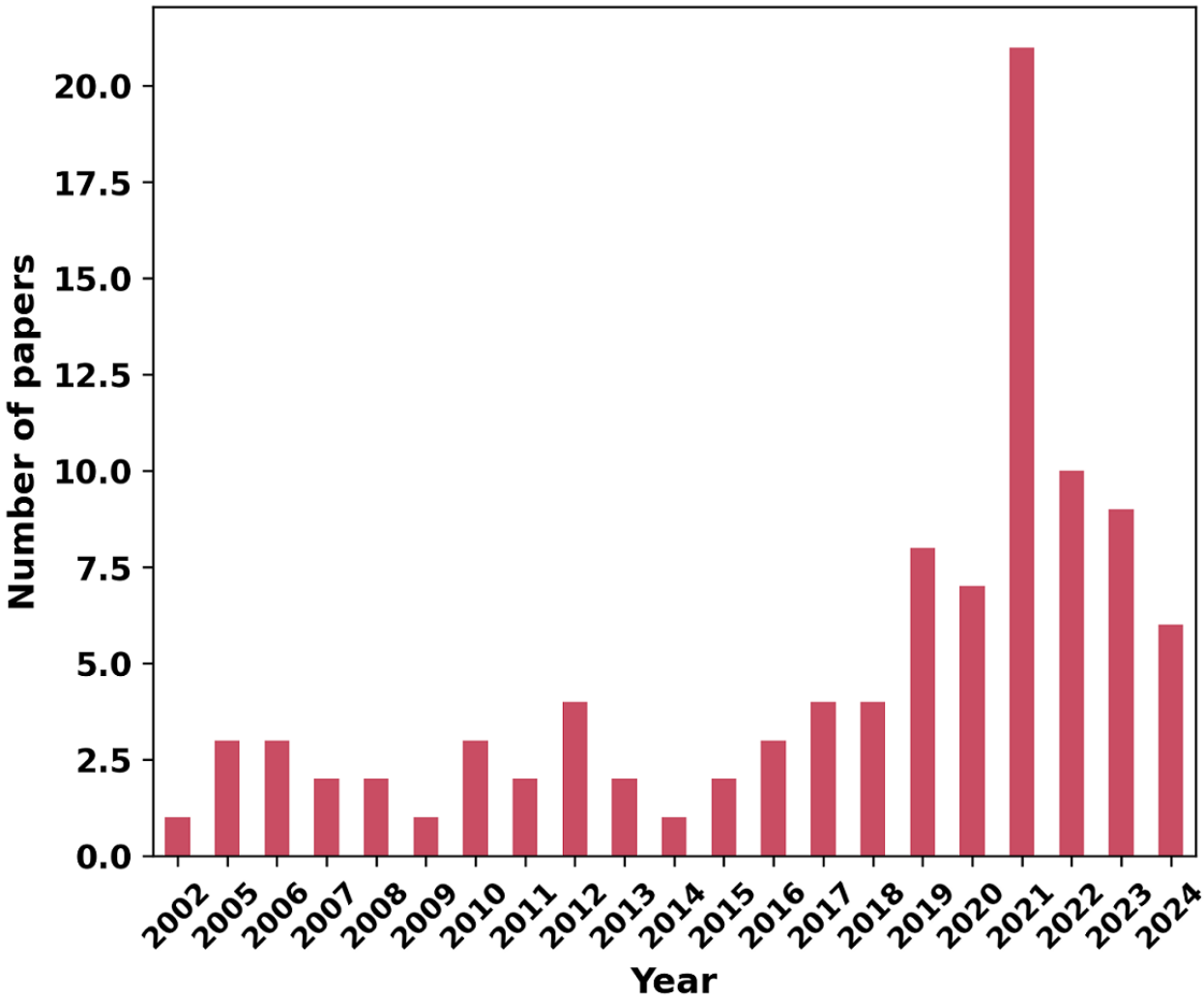

Figure 3 illustrates the number of articles published in each category since 2002. For the comorbidities category, the first article was published in 2007. Notably, there were no publications in 2009, 2011, 2015, and 2016, with a peak of 14 articles published in 2021, indicating growing interest and progress in this area. In the HIV infection monitoring category, the first article appeared in 2005, followed by a second one after a 10-year gap. In the category of predicting drug resistance, the first article was published in 2002, marking the earliest of the selected articles. It took 4 years for the second publication, with 3 articles appearing in 2016 and 2023. The first article on predicting treatment outcomes was published in 2010. In the treatment adherence category, the first article was in 2005, with a 20-year gap before the second publication. Finally, in the treatment recommendation category, 4 articles were published between 2005 and 2008. This trend saw a significant surge post-2015, culminating in a peak in 2021. The 2021 peak may have been influenced by the COVID-19 pandemic, which intensified global research efforts and resulted in a higher volume of publications (Table S6 in Multimedia Appendix 2).

XSL•FO
RenderX

Figure 3. Annual distribution of publications by category.

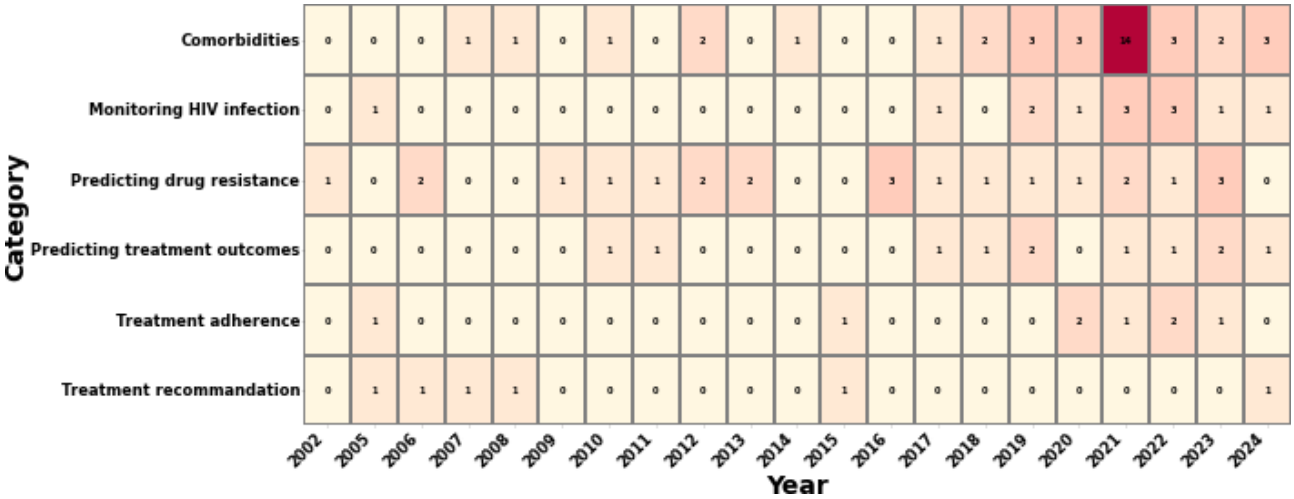

Among the 6 broad categories described above, the number of publications devoted to comorbidities amounts to 38% (n=37), closely followed by studies on the prediction of drug resistance at 23% (n=23), then monitoring HIV infection itself at 13% (n=13), predicting treatment outcomes at 11% (n=11), treatment adherence at 8% (n=8), and treatment recommendation at 6% (n=6).

More than half of the data used in the included studies came from America (more than 50% [53/98]), followed by Europe (20/98, 20%), Africa (18/98, 18%), and Asia (9/98, 9%). This imbalance revealed a strong dominance of data from Western regions, while data from Africa and Asia remained underrepresented. Overall, 52 studies were conducted in America [6,13-63], 20 in Europe [64-83], 11 in Africa [84-94], and 7 in Asia [95-101]. In addition, 8 studies are classified as collaborative [102-109], as the authors' affiliations span multiple continents. The number of first authors is 56 in America [6,13-63,102,103,107,108], 24 in Europe [64-83,104-106,109], 11 in Africa [84-94], and 7 in Asia [95-101] (Figure 4). Interestingly, the density of articles by region showed a slightly different distribution (Figure 5).

Figure 4. Sources of datasets and number of studies per continent, with number of first authors across regions. For each continent, a mini bar chart details the distribution of studies into six categories (comorbidities, prediction of drug resistance, HIV infection monitoring, prediction of treatment outcomes, treatment adherence, and treatment recommendations).

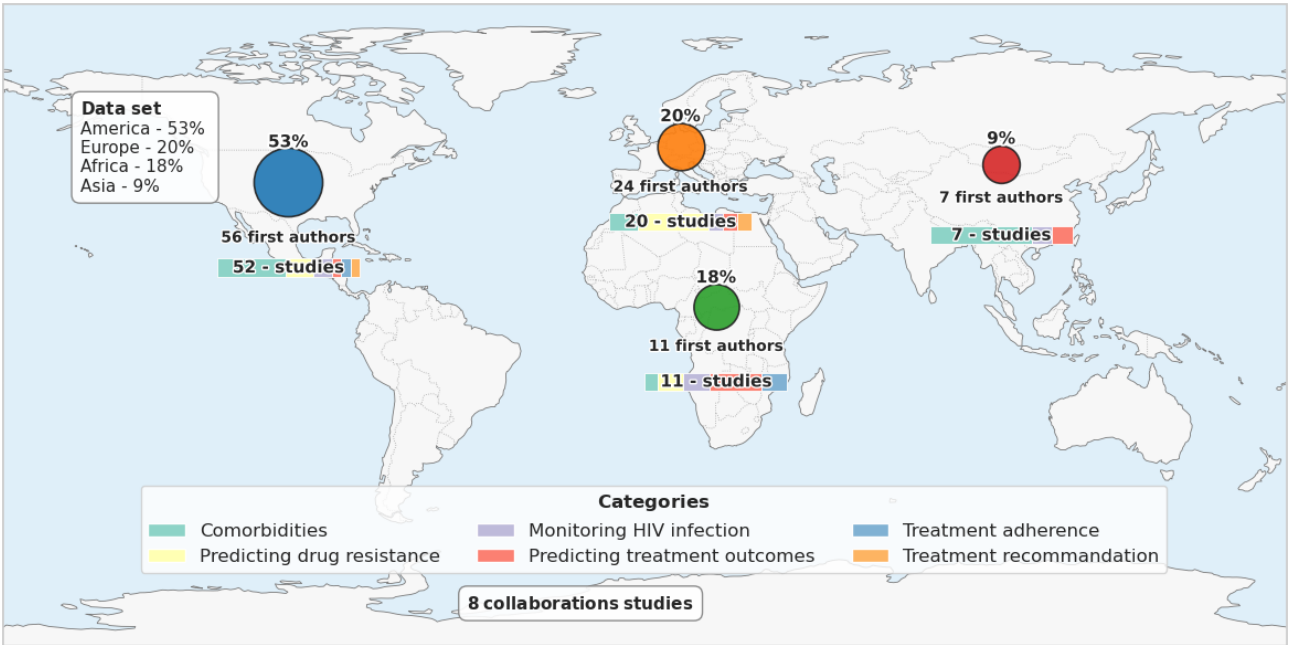

XSL•FO
RenderX

Figure 5. Density estimation of articles per year. Articles are assigned to a continent based on the affiliation of the first author. Europe (red), America (brown), Africa (blue), and Asia (purple).

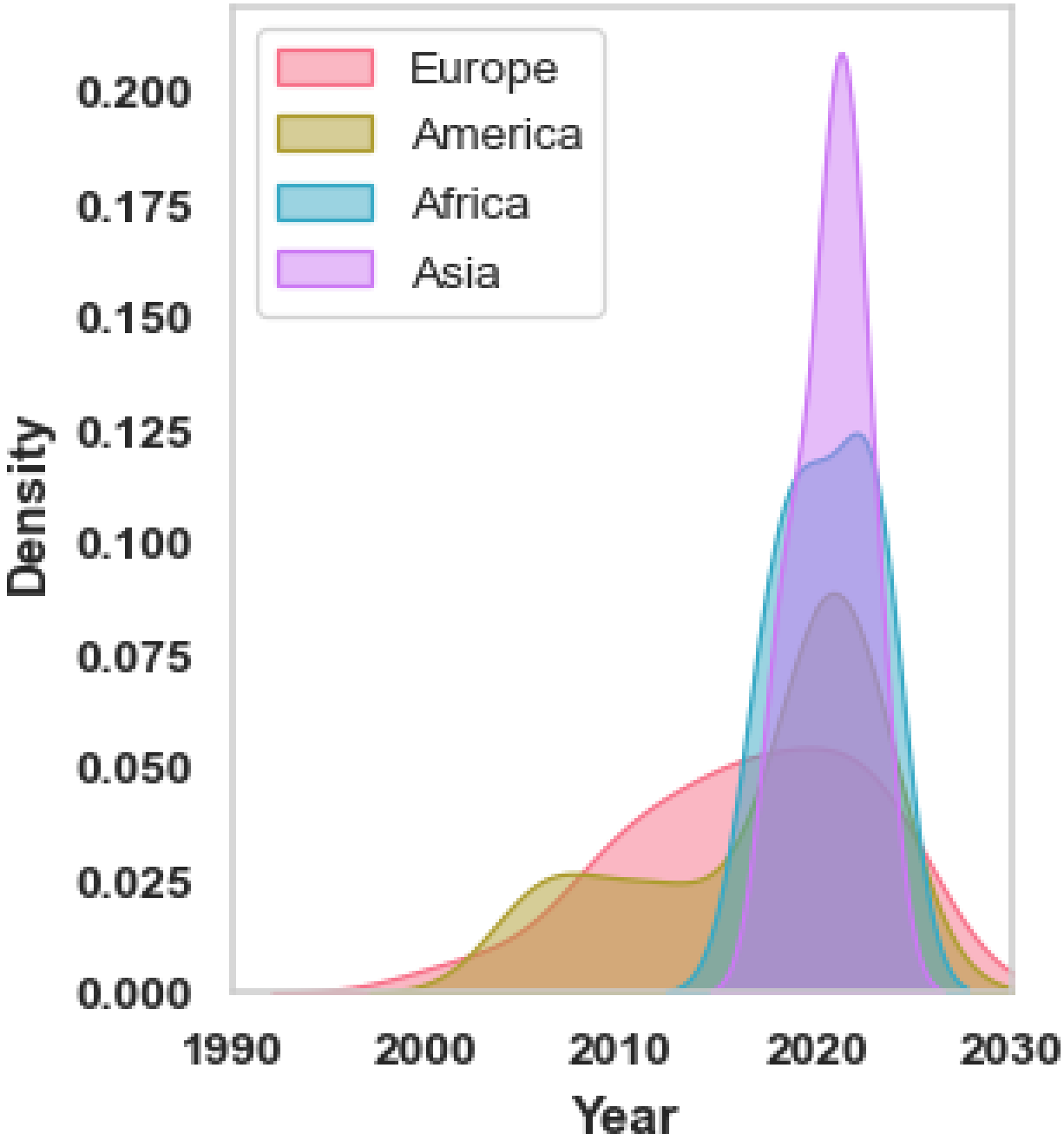

Fairness-aware causal paths decomposition around 2020, Africa and Asia presented the highest concentration of research activity on this topic, with Asia reaching the most significant peak, followed by Africa. Europe and America also make notable contributions, but with less intensity. The datasets used in these studies come from a variety of countries and are highly heterogeneous. To deepen the analysis, the various data types used in the reviewed articles are summarized in Figure 6.

XSL•FO
RenderX

Figure 6. Most used algorithms and statistical methods in publications relating to HIV and uses of machine learning. This figure presents the methods used in the different included articles, along with their frequency of occurrences across each of the six categories. Additionally, it highlights their respective affiliations. AI: artificial intelligence; FFSM: fuzzy finite state machine models; FACTS: fairness-aware causal paths decomposition.

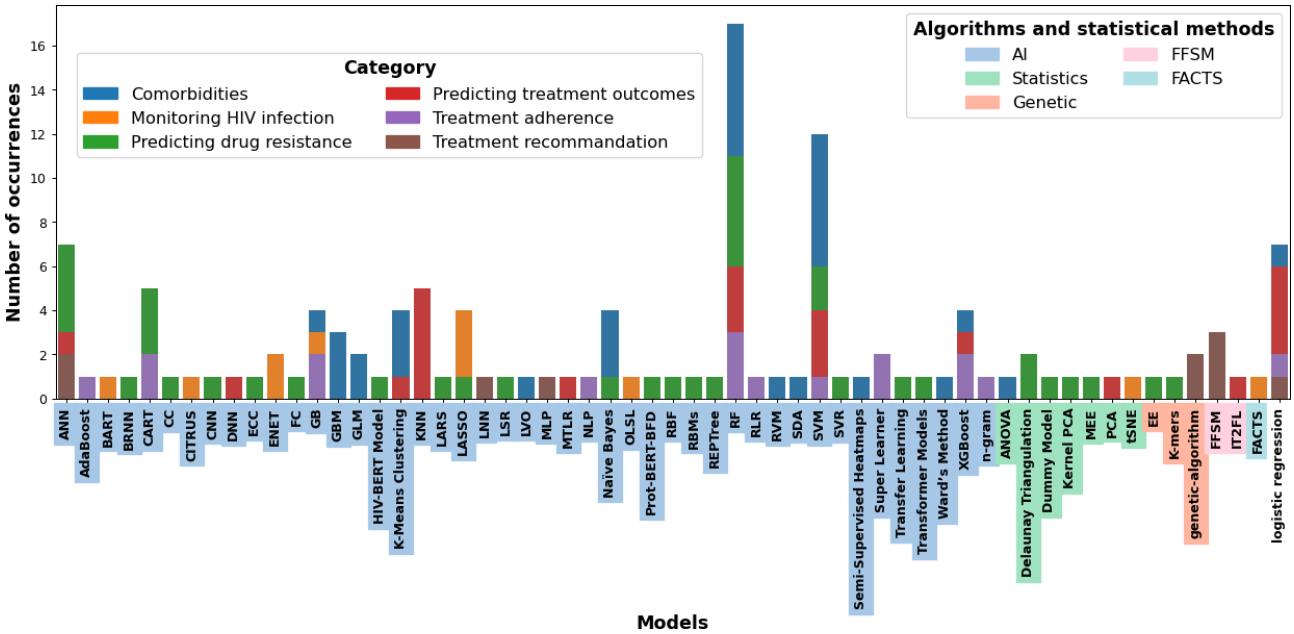

The three main data types are first numerical (n=75), then categorical (n=69), and finally textual (n=3). The 5 most used databases are cohort data (n=70), the Stanford database (n=15), the EuResist database (n=4), the Los Alamos National Laboratory database (n=3), and the Akwa Ibom database (n=2) (from a southern Nigerian state). Of these, HIV cohort data are the most widely used, particularly in studies of comorbidities, predicting drug resistance, predicting treatment outcomes, monitoring HIV infection, treatment adherence, and treatment recommendations. The Stanford database, which specializes in HIV drug resistance, comes second and is mainly used for research into resistance mutations. Finally, the EuResist database, the fruit of collaboration between several European sources, is used in 4 separate studies. A wide range of ML statistical methods were used in the studies we reviewed. The most used methods are random forest (RF), support vector machine (SVM), and logistic regression, each playing a key role in different aspects of HIV care research (Figure 7). Regarding method and algorithm trends, RF emerges as the most widely used algorithm, with 17.49% (43/247), reflecting its growing popularity in recent years. RF appears around 2010, with increasing use and peaking significantly in 2021. Artificial neural network (ANN), SVM, and logistic regression appeared after 2005, with increased use since 2017.

XSL•FO
RenderX

Figure 7. Algorithms and statistical method temporal trend (between 2000 and 2020) relating to HIV and uses of machine learning. Artificial neural network (blue), classification and regression trees (orange), random forest (green), support vector machine (red), and logistic regression (purple). ANN: artificial neural network; CART: classification and regression trees; RF: random forest; SVM: support vector machine.

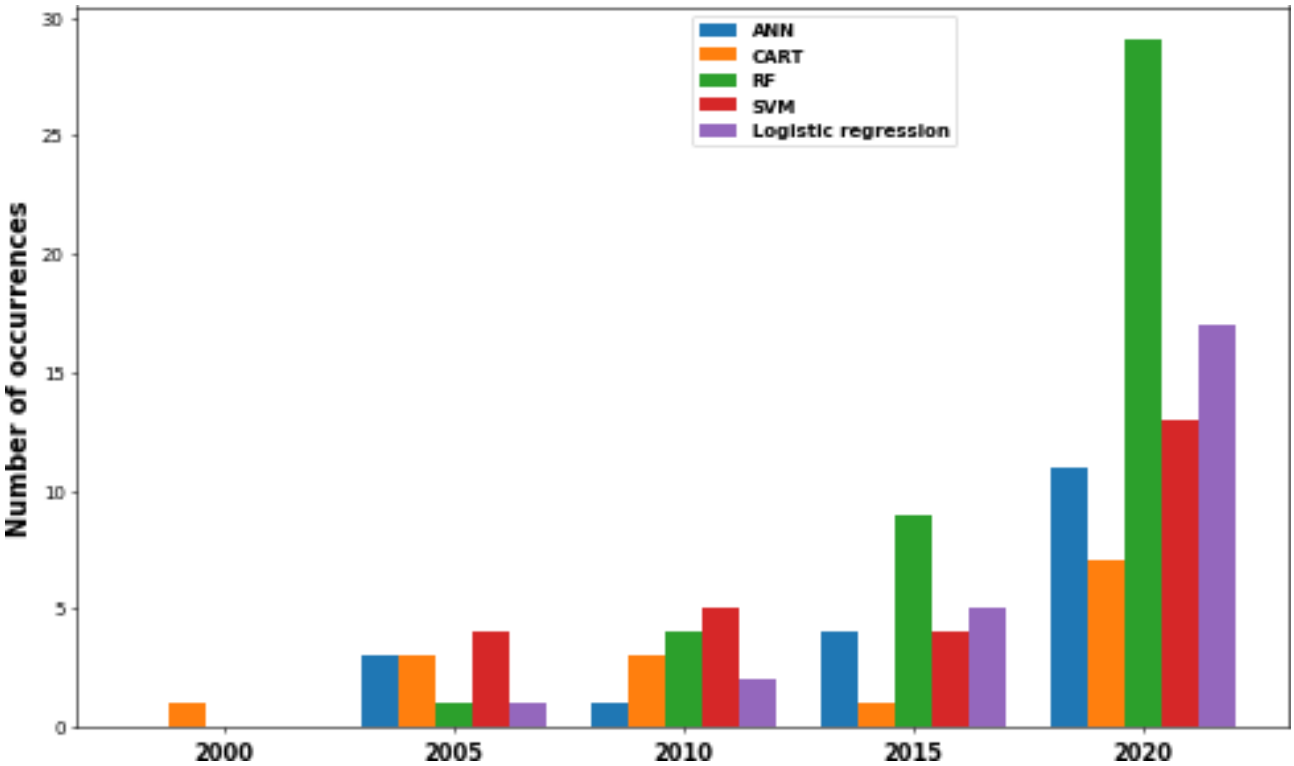

An in-depth analysis of the most used ML methods in the 6 categories mentioned above (Figure 8), as well as the tasks for which they are most effective, is detailed in this section. Logistic regression appears in all categories, whereas the ANN model was not used in monitoring HIV infection, and the other 3 methods were not used in the treatment recommendation category. We can observe that RF is the most used algorithm in each category where it appears. SVM is the second most widely used algorithm within the comorbidities category, ANN is used in predicting drug resistance, and logistic regression is used in predicting treatment outcomes and monitoring HIV infection.

Figure 8. Five most used algorithms and statistical methods relating to HIV and uses of machine learning for each category. Artificial neural network (blue), classification and regression trees (orange), random forest (green), support vector machine (red), and logistic regression (purple). ANN: artificial neural network; CART: classification and regression trees; RF: random forest; SVM: support vector machine.

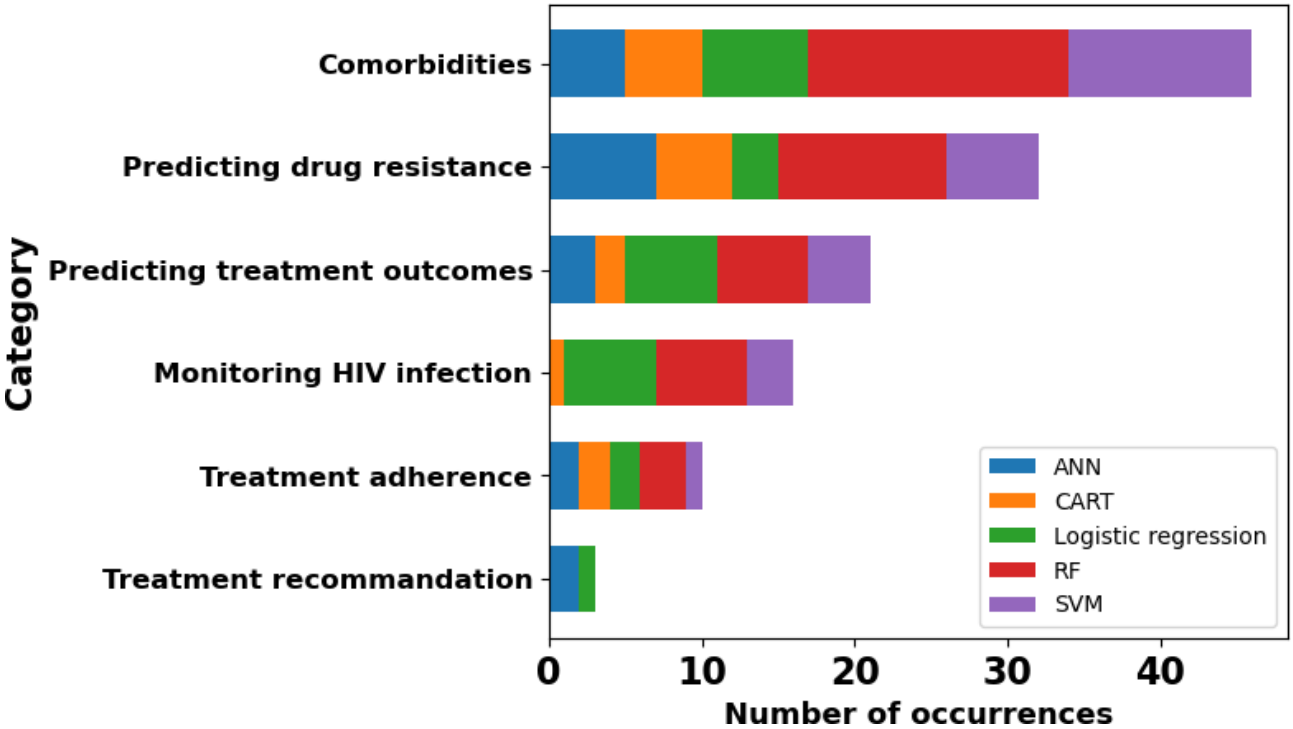

XSL•FO
RenderX

In the comorbidity category [6,13-35,61,64-67,84,95-99,102,103] (Figure S1 in Multimedia Appendix 3), which constitutes the largest part of our study data (37/98, 38%; Figure 4), ML methods were used for various tasks, including identifying risk factors for cardiac complications, predicting vaccine responses in the context of HIV co-infection, testing for retinal damage in people living with HIV that may lead to subtle visual field defects, and predicting increasing comorbidity risks using EHR data [6]. Cardinal [13] conducted a study that identified significant associations between carotid artery plaques and selected features in a population of HIV-positive and HIV-negative individuals with low to intermediate cardiovascular risk. The RF algorithm effectively distinguished between individuals with and without carotid artery plaques by combining traditional cardiovascular risk factors with strain elastography features, using 5 most discriminant combinations of features achieved area under the receiver operating characteristic (AUROC) between 0.76 and 0.80 as classification performance. For predicting vaccine response in the context of coinfection, nonlinear models estimated by regression tree and RF were more accurate than generalized linear models in predicting humoral responses to SARS-CoV-2 mRNA vaccination in people living with HIV (classification and regression trees: R^2=0.795; root mean square error=0.451 and RF: R^2=0.845; root mean square error=0.412) [64]. This is linked to the fact that the different feature selection strategies used in linear models often tend to exclude important variables which, taken in isolation, might have a more significant predictive role [110]. Kozak et al [14] and Goldbaum et al [15,61] investigated whether people living with HIV had retinal lesions resulting in subtle visual field defects, and whether ML classification methods could distinguish these defects from those of HIV-negative individuals. The results confirm that HIV has an effect on the retina, and the eyes of people living with HIV with low cluster of differentiation 4 (CD4) counts show visual field defects and retinal lesions, while those of people living with HIV with high CD4 counts can appear normal. The SVM method was able to distinguish the visual fields of people living with HIV from those of HIV-negative people, even for people living with HIV with high CD4 counts, a task that is more challenging for a human expert. The model achieved an AUROC of 0.843 for the subset of people living with HIV with low CD4 counts, and an AUROC of 0.695 for those with high CD4 counts. Another study was also conducted on predicting the risk of multidrug-resistant enterobacterial (MDR-E) infections among people living with HIV [16]. Among 4734 study participants, MDR-E was isolated from 1.6% (95% CI 1.2% - 2.1%). In unadjusted analyses, MDR-E was strongly associated with nadir CD4 cell count ≤200 cells/mm^3 (prevalence ratios [PR], 4.0; 95% CI, 2.3 - 7.4), history of an AIDS-defining clinical condition (PR, 3.7; 95% CI 2.3 - 6.2), and hospital admission in the prior 12 months (PR, 5; 95% CI 3.2 - 7.9). Searches were also made to identify biomarkers and assess the patterns of T lymphocyte cell activation associated with the development of tuberculosis following the initiation of ART in people living with HIV displaying high CD4 counts [95,102]. Furthermore, ML methods were able to identify biomarkers associated with metabolic syndrome and HIV-associated neurocognitive disorders. They also pinpointed genetic features on the HIV envelope, accurately predicted neurocognitive impairment, and recognized molecular signatures linked to HIV-associated neurocognitive disorders diagnosis [17-23,25,29-31,33-35,65,84,96,97,103].

Another example of application: a multivariable logistic regression model was more suitable to predict the risk of serious falls among an older population of people living with HIV. However, including ART classes taken by these individuals did not improve the algorithm's prediction, with the C-statistic increasing only slightly from 0.725 to 0.732 after the inclusion of ART classes [32]. The most frequently used method in this category is RF, as seen in (Figure S1 in Multimedia Appendix 3), valued for its robustness and use of the bootstrap principle.

For predicting or classifying drug resistance to antiretrovirals and identifying genetic signatures, 23% (23/98) of the articles focused on this area (Figure 4; Table S2 in Multimedia Appendix 2; Figure S2 in Multimedia Appendix 3) [36-43,62,63,68-77,85,86,104]. The classification and regression trees algorithm was able to predict drug resistance phenotypes from HIV-1 genotypes with good accuracy; some key sequence positions that are associated with drug resistance were identified using a mutual information analysis [68]. RF algorithm shows strongest correlation between predicted and actual virological responses, outperforming SVM and ANN. The mean absolute error ranged in [0.494-0.644] for RF, in [0.50-0.790] for SVM, and in [0.677-0.903] for ANN. Combining these methods further enhanced prediction accuracy, indicating that ensemble approaches are particularly effective in forecasting virological response [62]. ML methods were also used to analyze correlations between HIV-1 genotypes and resistance phenotypes, to predict virological responses to therapy based on HIV genotype, to detect mutations associated with HIV drug resistance without requiring expert knowledge, and to improve computational prediction of resistance from genotype data [36-43,63,69-77,85,86,104]. Both RF and SVM classified reverse transcriptase mutants with known resistance or sensitivity to nevirapine, achieving high accuracy in predicting susceptibility to the drug.

In the category of monitoring HIV infection, 13% (13/98) of (Table S3 in Multimedia Appendix 2; Figure S3 in Multimedia Appendix 3) studies involved applications of ML for tracking viral load and infection status [44-50,78,79,87,88,100,105]. For example, some studies identify biomarkers of viral rebound in people living with HIV from independent cohorts' pretreatment interruption [47] or tend to determine the viral compartmentalization dynamics to document a specific genetic signature of virus [44]. Mahto and Sood [49] evaluate various ML methods for predicting HIV progression and patient outcomes, identifying RF and extreme gradient boosting as the most effective, with accuracy rates of 0.88 and 0.89, respectively. They highlight the potential for further optimization through hyperparameter fine-tuning, feature engineering, and incorporating additional data sources to enhance algorithm robustness and generalizability. Kagendi and Mwau [88] developed an RF algorithm to predict HIV viral load hotspots in Kenya as an early warning system, achieving an accuracy of 0.78 and correctly identifying 434 viral load

XSL•FO
RenderX

hotspots in December 2019, demonstrating its potential to optimize ART programs by supporting proactive resource allocation. A study found that the risk of HIV infection was higher for African Americans compared to non-African Americans. By using the fairness-aware causal paths decomposition method, researchers identified several social determinants of health that contribute to this racial disparity, including education, income, violent crime, drinking, smoking, and rurality. Performance evaluation of predictive models demonstrated comparable results across methods: boosted logistic regression achieved an AUROC of 0.79, decision tree-based methods yielded an AUROC of 0.77, and RF outperformed with an AUROC of 0.80 [50]. The SVM was the top-performing algorithm for predicting CD4/CD8 ratios in patients with baseline CD4 counts below 200 cells/ml. The RF followed with R^2 values of 0.438 and 0.519 for SVM [100]. ML methods also monitored ART adherence and retention in care for people living with HIV.

For predicting treatment outcomes, 11% (11/98) of the articles explored this topic (Table S4 in Multimedia Appendix 2; Figure S4 in Multimedia Appendix 3) [51-53,80,81,89-92,101,106]. Among 7 classifiers, the RF algorithm achieved high accuracy in predicting CD4 count changes, with precision, sensitivity, and recall values close to 0.99 (sensitivity=1.00, precision=0.987, F_1-score=0.993, AUC=0.998), and it outperformed other algorithms in predicting virological failure and identifying key predictors [89]. The logistic regression excelled in predicting virological suppression [90]. ML methods were also able to predict adverse events [53], immune changes in ART recipients [52,54], biomarkers linked to mitochondrial toxicity [51,52], viral load [81,92], and highly active antiretroviral therapy (HAART) regimens recommendations [80].

For treatment adherence and predicting retention in care, 8% (8/98) of the articles examined ML applications (Table S5 in Multimedia Appendix 2; Figure S5 in Multimedia Appendix 3) [54-57,93,94,107,108]. A study conducted in Nigeria by Ogbechie et al [94] was able to predict interruptions in ART treatment at 30 days among people living with HIV, using routine program data. The method used integrated boosting tree and extreme gradient boosting techniques and demonstrated a strong performance, with a sensitivity of 0.81, specificity of 0.88, and a positive predictive value of 0.83. After its integration into the national electronic medical records system, the proportion of people living with HIV interrupted ART treatment cases decreased from 58.6% to 14.2%. Health workers reported that the model enabled proactive interventions, underscoring the potential of ML to enhance HIV treatment adherence and improve patient outcomes. Stockman et al [108] used ML methods to predict loss to follow-up among people living with HIV on ART treatment in Nigeria and Mozambique, using data from health facilities, geospatial sources, and satellite imagery. For the Mozambique dataset, the RF was the best method with an area under the precision-recall curve of 0.65, while in Nigeria, the boosted tree method achieved an area under the precision-recall curve of 0.52. Lu et al [57] demonstrated that SVM effectively models factors influencing adherence to HAART in people living with HIV, outperforming neural networks.

In total, 6 articles (6%) [58-60,82,83,109] discussed the use of ML for treatment recommendations in people living with HIV (Table S6 in Multimedia Appendix 2; Figure S6 in Multimedia Appendix 3. Fuzzy discrete event system-based HIV regimen selection system demonstrated high self-learning accuracy in predicting treatments, with more than 80% agreement with actual prescribed regimens for 35 patients under practical conditions. The system outperformed neural networks in transparency and interpretability and showed an accuracy of 0.84 to 1 in predicting new treatment regimens. Additionally, it achieved 82.9% agreement with non-expert physicians and 100% agreement with expert physicians. By considering clinical, demographic, and genotypic data, the combined system provided the best prediction performance for HAART therapy recommendations based on genotype (Figure S6 in Multimedia Appendix 3) [58-60,82,109]. Another approach was to assess the relevance of clinical drug interaction (DDI) with ART. Two DDI prediction algorithms: DeepARV-ChemBERT and DeepARV-sim were developed by Pham et al [83] for predicting DDIs between ARVs and comedications. Those algorithms achieved a weighted accuracy of 0.729 and 0.776, respectively.

Discussion

Principal Findings

Through our scoping review, we identified 6 major categories of ML applications used in the clinical field of HIV: consideration of comorbidities for people living with HIV, predicting drug resistance of the virus, monitoring HIV infection itself, predicting treatment outcomes for people living with HIV, treatment adherence for people living with HIV, and treatment recommendation for clinicians. In terms of models, RF emerged as the most used algorithm in all of the studies cited in this review.

Our results indicate that certain ML models are better suited to certain tasks in the field of HIV care. Indeed, model performance depends on the type and complexity of the data. In studies predicting the drug resistance of the virus, the data comes from the 2 main sources, namely the Stanford HIV Drug Resistance Database and EuResist DataBase. Thus, we observed that some models work better with linear relationships, while others, such as RFs, can handle both linear and nonlinear data, which may explain their widespread use in many of the articles included in this review.

Limitations

This scoping review presents a few limitations, relative to the review process itself. First, non-English articles have been excluded, which may omit some ML applications, but English remains the major language of scientific publications, and therefore, this bias appears to be limited. Second, this literature analysis did not assess in a comprehensive manner the methodological quality of the included studies, which represents a possible bias in the interpretation. To address these limitations, next research must consider in detail the robustness of the different approaches used in those studies in order to propose

XSL•FO
RenderX

best practices and specific limitations of AI models used in the field of HIV. However, our main goal was to map the existing publications using AI in the field of HIV. In Table 2, we discuss the strengths and limitations of the various ML tools identified in this literature analysis on people living with HIV care.

XSL•FO
RenderX

Table . Strengths and limitations of the most widely used machine learning algorithms.

Algorithms	Applications	Strengths	Limitations
Random forest (RF)	• Effectively distinguished between individuals with and without carotid artery plaques • Predicting vaccine response in the context of coinfection • Predicting (forecasting) virological response • Classified reverse transcriptase mutants with known resistance or sensitivity to nevirapine • Predict HIV progression and patient outcomes • Predict HIV viral load hotspots • Identified several social determinants of health contributing to this racial disparity • Predict CD4[a] count changes • Predict humoral responses	• Can handle linear, nonlinear data, and missing data • Reduce overfitting compared to single trees • Robust to noise and outliers • Provide feature importance measures • Good generalization ability	• Less interpretable than a single tree • Can be computationally heavy with large datasets • Struggles with very high-dimensional sparse data
Support vector machines (SVM)	• Able to distinguish the visual fields of people living with HIV from those of HIV-negative people • Predict risk of MDR-E[b] infections among people living with HIV • Classified reverse transcriptase mutants with known resistance or sensitivity to Nevirapine • Predict (forecasting) virological response • Effectively models factors influencing adherence to HAART[c]	• Effective in high-dimensional spaces • Works well with clear margin of separation. • Kernel trick allows nonlinear classification	• Not scalable to very large datasets. • Requires careful kernel and parameter choice • Less interpretable than simple models
Decision trees (CART)[d]	• Predict drug resistance phenotypes from HIV-1 genotypes with good accuracy • Identified several social determinants of health contributing to this racial disparity • Effectively models factors influencing adherence to HAART and identifying genetic signatures	• Easy to visualize and interpret • Handles nonlinear relationships and interactions naturally • Can work with categorical and numerical data • No need for feature scaling	• Unstable with small data changes • Prone to overfitting • Not as accurate as ensemble methods
Artificial neural networks (ANN)	• Predict drug resistance phenotypes from HIV-1 genotypes with good accuracy • Identified several social determinants of health contributing to this racial disparity • Effectively models factors influencing adherence to HAART and identifying genetic signatures	• Capture highly complex and nonlinear patterns • Can automatically learn feature representations	• Require massive amounts of data • High computational • Hard to interpret (black box) • Risk of overfitting without regularization • Training can be unstable

XSL•FO
RenderX

Algorithms	Applications	Strengths	Limitations
Logistic regression	• Predict drug-drug interaction • Predict drug resistance • Predict virological suppression • Predict the risk of serious falls among an older population of people living with HIV	• Simple, fast, and interpretable • Works well with small datasets • Easy to implement and regularize	• Assumes linear relationships between predictors and outcome • Struggles with complex and nonlinear data • Sensitive to multicollinearity and outliers • Limited predictive power compared to advanced models

[a] CD4: cluster of differentiation 4.

[b]MDR-E: multidrug-resistant enterobacterial.

[c]HAART: highly active antiretroviral therapy.

[d]CART: classification and regression trees.

We can also consider certain limitations inherent in the results. Concerning the type of AI models, a lot of articles focus on ML models, while other AI approaches, such as natural language processing and large language models have been less explored or remain underexplored. Moreover, data availability is limited: many studies rely on small or localized datasets, with frequent reuse of the same databases, reducing both novelty and representativeness. Databases dealing with the issue of AI applied to HIV are heterogeneous and require the definition of a federated database oriented by clinical field. This highlights the need to develop large-scale, interconnected, multicenter, and openly accessible datasets in order to reduce barriers to progress, for example, to improve the performance of causal inference methods.

We could similarly address as limitations the complexity of this disease. For example, another chronic disease such as type 2 diabetes mellitus [111] revealed common difficulties, in particular in the lack of model validation across different clinical contexts and populations. Indeed, López et al [111] also expose the necessity to build highly specialized and effective AI models. In a similar way, algorithms trained on specific subpopulations of people living with HIV may not provide the same performance score with other different populations of people living with HIV due to genetic variations, lifestyles, and socioeconomic factors. That is why, as in clinical trials, the characteristics of individuals are crucial, as they determine whether or not the results can be extrapolated.

The use of AI in health care raises important ethical questions, particularly with regard to patient data privacy and informed consent. Regulations governing personal data are not standardized, and the data currently available is not structured in the same way and is often incomplete, which can limit its use with AI methods and tools. The establishment of clear, globally standardized regulatory frameworks and training in data ethics for health care professionals and students is essential to ensure the safe, ethical adoption, trust of AI, and the development of sustainable solutions for the care of people living with HIV.

Future Directions

We observed through this scoping review that many applications of ML are not yet explored. Among these is the field of public health, where identifying clusters through social media monitoring would enable contact tracing. AI methods and tools could also be used for individual monitoring, disease progression, and real-time support for people living with HIV. Notably, this review did not identify any studies on predicting the onset of adverse effects and therapeutic optimization in the context of drug resistance, DDI, and iatrogenic events, as well as other associated chronic diseases that must incorporate the concept of aging. Among these areas of research, monitoring adverse effects is becoming an important field. The field of AI in pharmacovigilance is gaining momentum in health care [112-114]. Recently, several studies have focused on predicting HIV-related mortality and comorbidities, or predicting admission to intensive care [115-118].

Another point to consider for the future is the use of algorithms based on real clinical data that could be used serving a teacher and assessor in medical education, with the creation of fictitious patients, for example, to improve management by future clinicians. This should help to educate practitioners about AI.

Conclusions

In summary, our review identified the most suitable ML methods across the 6 categories chosen. RF emerged as the most used algorithm due to its versatility and suitability for both classification and prediction tasks. Following closely was the SVM, both of which are supervised ML algorithms widely applied to various aspects of HIV infection challenges. The diversity of databases is a critical element to consider. However, key challenges remain, including limited data availability, quality, and accessibility, which continue to hinder the broader adoption of evidence-based medicine in HIV care. Regular updates of ML tools are crucial to ensure that evidence-based medicine tools provide accurate, up-to-date information for clinical decision-making. Although there has been an increase in publications across various aspects of HIV care, there remains a critical lack of research addressing treatment safety and optimization. Future efforts should focus on monitoring adverse events and developing decision-support tools based on the risk of drug-related adverse effects. Addressing this gap will be the focus for future studies, where the aim to leverage AI is to develop more personalized, effective ART regimens.

XSL•FO
RenderX

Acknowledgments

The authors would like to express our sincere gratitude to Kokou Laris EDJINEDJA, Omar ELFAHIM (from Université Marie & Louis Pasteur), and the entire OCTAVIA project team Hervé Cardot, Xavier Dupuis, and Catherine Labruère-Chazal from the Université de Bourgogne Europe for their unwavering dedication and constant availability. We thank also Camelia Goga for careful reading of the manuscript. Generative artificial intelligence was not used in any portion of the manuscript writing.

Conflicts of Interest

JLR reports receiving travel grants from Gilead Sciences and ViiV Healthcare to attend international conferences, as well as personal fees from Gilead Sciences and MSD outside of this work. The other authors declare no conflicts of interest.

Multimedia Appendix 1
Artificial intelligence vocabulary - some basic definitions.
[DOCX File, 18 KB - jmir_v28i1e79219_app1.docx]

Multimedia Appendix 2
Selected studies with their summaries for each category.
[DOCX File, 82 KB - jmir_v28i1e79219_app2.docx]

Multimedia Appendix 3
Top five algorithms and statistical methods for all categories.
[DOCX File, 193 KB - jmir_v28i1e79219_app3.docx]

Checklist 1
PRISMA 2020 checklist.
[PDF File, 270 KB - jmir_v28i1e79219_app4.pdf]

References

1. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec;2(4):230-243. [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
2. Yang J, Mwangi AW, Kantor R, et al. Tree-based subgroup discovery using electronic health record data: heterogeneity of treatment effects for DTG-containing therapies. Biostatistics 2024 Apr 15;25(2):323-335. [doi: 10.1093/biostatistics/kxad014] [Medline: 37475638]
3. Birkhead GS, Klompas M, Shah NR. Uses of electronic health records for public health surveillance to advance public health. Annu Rev Public Health 2015 Mar 18;36(1):345-359. [doi: 10.1146/annurev-publhealth-031914-122747] [Medline: 25581157]
4. Marcus JL, Hurley LB, Krakower DS, Alexeeff S, Silverberg MJ, Volk JE. Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: a modelling study. Lancet HIV 2019 Oct;6(10):e688-e695. [doi: 10.1016/S2352-3018(19)30137-7] [Medline: 31285183]
5. Davis SE, Zabotka L, Desai RJ, et al. Use of electronic health record data for drug safety signal identification: a scoping review. Drug Saf 2023 Aug;46(8):725-742. [doi: 10.1007/s40264-023-01325-0] [Medline: 37340238]
6. Yang X, Zhang J, Chen S, Weissman S, Olatosi B, Li X. Utilizing electronic health record data to understand comorbidity burden among people living with HIV: a machine learning approach. AIDS 2021 May 1;35(Suppl 1):S39-S51. [doi: 10.1097/QAD.0000000000002736] [Medline: 33867488]
7. Kaipio J, Lääveri T, Hyppönen H, et al. Usability problems do not heal by themselves: national survey on physicians' experiences with EHRs in Finland. Int J Med Inform 2017 Jan;97:266-281. [doi: 10.1016/j.ijmedinf.2016.10.010] [Medline: 27919385]
8. Auffray C, Balling R, Barroso I, et al. Making sense of big data in health research: towards an EU action plan. Genome Med 2016 Jun 23;8(1):71. [doi: 10.1186/s13073-016-0323-y] [Medline: 27338147]
9. Kiourtis A, Mavrogiorgou A, Mavrogiorgos K, et al. Electronic health records at people's hands across Europe: the InteropEHRate protocols. Stud Health Technol Inform 2022 Nov 3;299:145-150. [doi: 10.3233/SHTI220973] [Medline: 36325855]
10. Kaplan-Lewis E, Aberg JA, Lee M. Aging with HIV in the ART era. Semin Diagn Pathol 2017 Jul;34(4):384-397. [doi: 10.1053/j.semdp.2017.04.002] [Medline: 28552209]
11. Back D, Marzolini C. The challenge of HIV treatment in an era of polypharmacy. J Int AIDS Soc 2020 Feb;23(2):e25449. [doi: 10.1002/jia2.25449] [Medline: 32011104]

XSL•FO
RenderX

12. Lee SW, Koo MJ. PRISMA 2020 statement and guidelines for systematic review and meta-analysis articles, and their underlying mathematics: life cycle committee recommendations. Life Cycle 2022;2. [doi: 10.54724/lc.2022.e9]
13. Cardinal MHR, Durand M, Chartrand-Lefebvre C, Soulez G, Tremblay C, Cloutier G. Association between early carotid artery plaque presence, vascular strain imaging features and traditional cardiovascular risk factors in HIV infected individuals. Presented at: 2021 IEEE International Ultrasonics Symposium (IUS); Sep 11-16, 2021; Xi'an, China p. 1-3. [doi: 10.1109/IUS52206.2021.9593389]
14. Kozak I, Sample PA, Hao J, et al. Machine learning classifiers detect subtle field defects in eyes of HIV individuals. Trans Am Ophthalmol Soc 2007;105:111-118. [Medline: 18427600]
15. Goldbaum MH, Kozak I, Hao J, et al. Pattern recognition can detect subtle field defects in eyes of HIV individuals without retinitis under HAART. Graefes Arch Clin Exp Ophthalmol 2011 Apr;249(4):491-498. [doi: 10.1007/s00417-010-1511-x] [Medline: 20865422]
16. Henderson HI, Napravnik S, Kosorok MR, et al. Predicting risk of multidrug-resistant enterobacterales infections among people with HIV. Open Forum Infect Dis 2022 Oct;9(10):ofac487. [doi: 10.1093/ofid/ofac487] [Medline: 36225740]
17. Holman AG, Gabuzda D. A machine learning approach for identifying amino acid signatures in the HIV env gene predictive of dementia. PLoS ONE 2012;7(11):e49538. [doi: 10.1371/journal.pone.0049538] [Medline: 23166702]
18. Choi JY, Hightower GK, Wong JK, et al. Genetic features of cerebrospinal fluid-derived subtype B HIV-1 tat. J Neurovirol 2012 Apr;18(2):81-90. [doi: 10.1007/s13365-011-0059-9] [Medline: 22528397]
19. Cassol E, Misra V, Dutta A, Morgello S, Gabuzda D. Cerebrospinal fluid metabolomics reveals altered waste clearance and accelerated aging in HIV patients with neurocognitive impairment. AIDS 2014 Jul 17;28(11):1579-1591. [doi: 10.1097/QAD.0000000000000303] [Medline: 24752083]
20. Underwood J, Cole JH, Leech R, Sharp DJ, Winston A, CHARTER group. Multivariate pattern analysis of volumetric neuroimaging data and its relationship with cognitive function in treated HIV disease. J Acquir Immune Defic Syndr 2018 Aug 1;78(4):429-436. [doi: 10.1097/QAI.0000000000001687] [Medline: 29608444]
21. Luckett P, Paul RH, Navid J, et al. Deep learning analysis of cerebral blood flow to identify cognitive impairment and frailty in persons living with HIV. J Acquir Immune Defic Syndr 2019 Dec 15;82(5):496-502. [doi: 10.1097/QAI.0000000000002181] [Medline: 31714429]
22. Dastgheyb RM, Sacktor N, Franklin D, et al. Cognitive trajectory phenotypes in human immunodeficiency virus–infected patients. J Acquir Immune Defic Syndr 2019 Sep 1;82(1):61-70. [doi: 10.1097/QAI.0000000000002093] [Medline: 31107302]
23. Rubin LH, Sundermann EE, Dastgheyb R, et al. Sex differences in the patterns and predictors of cognitive function in HIV. Front Neurol 2020;11:551921. [doi: 10.3389/fneur.2020.551921] [Medline: 33329301]
24. Shu C, Justice AC, Zhang X, et al. DNA methylation biomarker selected by an ensemble machine learning approach predicts mortality risk in an HIV-positive veteran population. Epigenetics 2021;16(7):741-753. [doi: 10.1080/15592294.2020.1824097] [Medline: 33092459]
25. Paul RH, Cho KS, Belden AC, et al. Machine-learning classification of neurocognitive performance in children with perinatal HIV initiating de novo antiretroviral therapy. AIDS 2020 Apr 1;34(5):737-748. [doi: 10.1097/QAD.0000000000002471] [Medline: 31895148]
26. Luckett PH, Paul RH, Hannon K, et al. Modeling the effects of HIV and aging on resting-state networks using machine learning. J Acquir Immune Defic Syndr 2021 Dec 1;88(4):414-419. [doi: 10.1097/QAI.0000000000002783] [Medline: 34406983]
27. Paul R, Tsuei T, Cho K, et al. Ensemble machine learning classification of daily living abilities among older people with HIV. EClinicalMedicine 2021 May;35:100845. [doi: 10.1016/j.eclinm.2021.100845] [Medline: 34027327]
28. Solomon IH, Chettimada S, Misra V, et al. White matter abnormalities linked to interferon, stress response, and energy metabolism gene expression changes in older HIV-positive patients on antiretroviral therapy. Mol Neurobiol 2020 Feb;57(2):1115-1130. [doi: 10.1007/s12035-019-01795-3] [Medline: 31691183]
29. Finkelstein A, Faiyaz A, Weber MT, et al. Fixel-based analysis and free water corrected DTI evaluation of HIV-associated neurocognitive disorders. Front Neurol 2021;12:725059. [doi: 10.3389/fneur.2021.725059] [Medline: 34803875]
30. Oliveira NL, Kennedy EH, Tibshirani R, et al. Longitudinal 5-year prediction of cognitive impairment among men with HIV disease. AIDS 2021 May 1;35(6):889-898. [doi: 10.1097/QAD.0000000000002827] [Medline: 33534203]
31. Pinheiro L, Pereira MLD, Andrade ED, et al. An intelligent multicriteria model for diagnosing dementia in people infected with human immunodeficiency virus. Appl Sci (Basel) 2021;11(21):10457. [doi: 10.3390/app112110457]
32. Womack JA, Murphy TE, Leo-Summers L, et al. Predictive risk model for serious falls among older persons living with HIV. J Acquir Immune Defic Syndr 2022 Oct 1;91(2):168-174. [doi: 10.1097/QAI.0000000000003030] [Medline: 36094483]
33. Mohammadzadeh N, Zhang N, Branton WG, et al. The HIV restriction factor profile in the brain is associated with the clinical status and viral quantities. Viruses 2023 Jan 23;15(2):316. [doi: 10.3390/v15020316] [Medline: 36851531]
34. Murdoch DM, Barfield R, Chan C, et al. Neuroimaging and immunological features of neurocognitive function related to substance use in people with HIV. J Neurovirol 2023 Feb;29(1):78-93. [doi: 10.1007/s13365-022-01102-2] [Medline: 36348233]

XSL•FO
RenderX

35. Marques de Menezes EG, Bowler SA, Shikuma CM, Ndhlovu LC, Norris PJ. Circulating plasma-derived extracellular vesicles expressing bone and kidney markers are associated with neurocognitive impairment in people living with HIV. Front Neurol 2024;15:1383227. [doi: 10.3389/fneur.2024.1383227] [Medline: 38725641]
36. Rhee SY, Taylor J, Wadhera G, Ben-Hur A, Brutlag DL, Shafer RW. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. Proc Natl Acad Sci U S A 2006 Nov 14;103(46):17355-17360. [doi: 10.1073/pnas.0607274103] [Medline: 17065321]
37. Ravich VL, Masso M, Vaisman II. A combined sequence-structure approach for predicting resistance to the non-nucleoside HIV-1 reverse transcriptase inhibitor Nevirapine. Biophys Chem 2011 Jan;153(2-3):168-172. [doi: 10.1016/j.bpc.2010.11.004] [Medline: 21146283]
38. Masso M. Prediction of human immunodeficiency virus type 1 drug resistance: representation of target sequence mutational patterns via an n-grams approach. 2012 Presented at: 2012 IEEE International Conference on Bioinformatics and Biomedicine; Oct 4-7, 2012; Philadelphia, PA, USA p. 1-6. [doi: 10.1109/BIBM.2012.6392665]
39. Yu X, Weber IT, Harrison RW. Prediction of HIV drug resistance from genotype with encoded three-dimensional protein structure. BMC Genomics 2014;15 Suppl 5(Suppl 5):1-13. [doi: 10.1186/1471-2164-15-S5-S1] [Medline: 25081370]
40. Shen C, Yu X, Harrison RW, Weber IT. Automated prediction of HIV drug resistance from genotype data. BMC Bioinformatics 2016 Aug 31;17 Suppl 8(Suppl 8):278. [doi: 10.1186/s12859-016-1114-6] [Medline: 27586700]
41. Pawar SD, Freas C, Weber IT, Harrison RW. Analysis of drug resistance in HIV protease. BMC Bioinformatics 2018 Oct 22;19(Suppl 11):362. [doi: 10.1186/s12859-018-2331-y] [Medline: 30343664]
42. Steiner MC, Gibson KM, Crandall KA. Drug resistance prediction using deep learning techniques on HIV-1 sequence data. Viruses 2020 May 19;12(5):560. [doi: 10.3390/v12050560] [Medline: 32438586]
43. Dampier W, Link RW, Earl JP, et al. HIV- bidirectional encoder representations from transformers: a set of pretrained transformers for accelerating HIV deep learning tasks. FrontVirol 2022;2:880618. [doi: 10.3389/fviro.2022.880618]
44. Pillai SK, Good B, Pond SK, et al. Semen-specific genetic characteristics of human immunodeficiency virus type 1 env. J Virol 2005 Feb;79(3):1734-1742. [doi: 10.1128/JVI.79.3.1734-1742.2005] [Medline: 15650198]
45. Pohlmeyer CW, Gonzalez VD, Irrinki A, et al. Identification of NK cell subpopulations that differentiate HIV-infected subject cohorts with diverse levels of virus control. J Virol 2019 Apr 1;93(7):10-1128. [doi: 10.1128/JVI.01790-18] [Medline: 30700608]
46. Weissman S, Yang X, Zhang J, Chen S, Olatosi B, Li X. Using a machine learning approach to explore predictors of healthcare visits as missed opportunities for HIV diagnosis. AIDS 2021 May 1;35(Suppl 1):S7-S18. [doi: 10.1097/QAD.0000000000002735] [Medline: 33867485]
47. Giron LB, Palmer CS, Liu Q, et al. Non-invasive plasma glycomic and metabolic biomarkers of post-treatment control of HIV. Nat Commun 2021 Jun 29;12(1):3922. [doi: 10.1038/s41467-021-24077-w] [Medline: 34188039]
48. Murnane PM, Ayieko J, Vittinghoff E, et al. Machine learning algorithms using routinely collected data do not adequately predict viremia to inform targeted services in postpartum women living with HIV. J Acquir Immune Defic Syndr 2021 Dec 15;88(5):439-447. [doi: 10.1097/QAI.0000000000002800] [Medline: 34520443]
49. Mahto R, Sood K. HIV progression and outcome prediction to enhance patient matching for clinical trials. Presented at: 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC); Jan 8-10, 2024; Las Vegas, NV, USA p. 0278-0284. [doi: 10.1109/CCWC60891.2024.10427778]
50. Prosperi M, Xu J, Guo JS, et al. Identification of social and racial disparities in risk of HIV infection in Florida using causal AI methods. Presented at: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 6-8, 2022; Las Vegas, NV, USA p. 2934-2939. [doi: 10.1109/BIBM55620.2022.9995662]
51. Lee JS, Paintsil E, Gopalakrishnan V, Ghebremichael M. A comparison of machine learning techniques for classification of HIV patients with antiretroviral therapy-induced mitochondrial toxicity from those without mitochondrial toxicity. BMC Med Res Methodol 2019 Nov 27;19(1):216. [doi: 10.1186/s12874-019-0848-z] [Medline: 31775643]
52. Bose E, Paintsil E, Ghebremichael M. Minimum redundancy maximal relevance gene selection of apoptosis pathway genes in peripheral blood mononuclear cells of HIV-infected patients with antiretroviral therapy-associated mitochondrial toxicity. BMC Med Genomics 2021 Dec 1;14(1):285. [doi: 10.1186/s12920-021-01136-1] [Medline: 34852799]
53. Premeaux TA, Bowler S, Friday CM, et al. Machine learning models based on fluid immunoproteins that predict non-AIDS adverse events in people with HIV. iScience 2024 Jun 21;27(6):109945. [doi: 10.1016/j.isci.2024.109945] [Medline: 38812553]
54. Oliwa T, Furner B, Schmitt J, Schneider J, Ridgway JP. Development of a predictive model for retention in HIV care using natural language processing of clinical notes. J Am Med Inform Assoc 2021 Jan 15;28(1):104-112. [doi: 10.1093/jamia/ocaa220] [Medline: 33150369]
55. Ramachandran A, Kumar A, Koenig H, et al. Predictive analytics for retention in care in an urban HIV clinic. Sci Rep 2020 Apr 14;10(1):6421. [doi: 10.1038/s41598-020-62729-x] [Medline: 32286333]
56. Petersen ML, LeDell E, Schwab J, et al. Super learner analysis of electronic adherence data improves viral prediction and may provide strategies for selective HIV RNA monitoring. JAIDS Journal of Acquired Immune Deficiency Syndromes 2015;69(1):109-118. [doi: 10.1097/QAI.0000000000000548]

XSL•FO
RenderX

57. Lu Z, Ying H, Lin F, Neufeld S, Luborsky M, Brawn DM. Multi-class support vector machines for modeling HIV/AIDS treatment adherence using patient data. 2005 Presented at: International Joint Conference on Neural Networks 2005; Jul 31 to Aug 4, 2005; Montreal, Que, Canada p. 2417-2422. [doi: 10.1109/IJCNN.2005.1556281]
58. Ying H, Lin F, Luan X, et al. A fuzzy discrete event system with self-learning capability for HIV/AIDS treatment regimen selection. 2005 Presented at: NAFIPS 2005 - 2005 Annual Meeting of the North American Fuzzy Information Processing Society; Jun 26 to Feb 28, 2005; Detroit, MI, USA p. 820-824. [doi: 10.1109/NAFIPS.2005.1548646]
59. Ying H, Lin F, MacArthur RD, et al. A fuzzy discrete event system approach to determining optimal HIV/AIDS treatment regimens. IEEE Trans Inf Technol Biomed 2006 Oct;10(4):663-676. [doi: 10.1109/titb.2006.874200] [Medline: 17044400]
60. Ying H, Lin F, MacArthur RD, et al. A self-learning fuzzy discrete event system for HIV/AIDS treatment regimen selection. IEEE Trans Syst Man Cybern B Cybern 2007 Aug;37(4):966-979. [doi: 10.1109/tsmcb.2007.895360] [Medline: 17702293]
61. Goldbaum MH, Falkenstein I, Kozak I, et al. Analysis with support vector machine shows HIV-positive subjects without infectious retinitis have mfERG deficiencies compared to normal eyes. Trans Am Ophthalmol Soc 2008;106:196-204. [Medline: 19277235]
62. Cheng BY, Carbonell JG. Automatic detection of HIV drug resistance-associated mutations. : IEEE; 2010 Presented at: 2010 Ninth International Conference on Machine Learning and Application; Dec 12-14, 2010; NW Washington, DC, United States p. 528-533. [doi: 10.1109/ICMLA.2010.83]
63. Pillai SK, Pond SLK, Liu Y, et al. Genetic attributes of cerebrospinal fluid-derived HIV-1 env. Brain (Bacau) 2006 Jul;129(7):1872-1883. [doi: 10.1093/brain/awl136] [Medline: 16735456]
64. Montesi G, Augello M, Polvere J, Marchetti G, Medaglini D, Ciabattini A. Predicting humoral responses to primary and booster SARS-CoV-2 mRNA vaccination in people living with HIV: a machine learning approach. J Transl Med 2024 May 7;22(1):432. [doi: 10.1186/s12967-024-05147-1] [Medline: 38715088]
65. Underwood J, Cole JH, Caan M, et al. Gray and white matter abnormalities in treated human immunodeficiency virus disease and their relationship to cognitive function. Clin Infect Dis 2017 Aug 1;65(3):422-432. [doi: 10.1093/cid/cix301] [Medline: 28387814]
66. Gelpi M, Mikaeloff F, Knudsen AD, et al. The central role of the glutamate metabolism in long-term antiretroviral treated HIV-infected individuals with metabolic syndrome. Aging (Milano) 2021 Oct 15;13(19):22732-22751. [doi: 10.18632/aging.203622]
67. Olund Villumsen S, Benfeitas R, Knudsen AD, et al. Integrative lipidomics and metabolomics for system-level understanding of the metabolic syndrome in long-term treated HIV-infected individuals. Front Immunol 2021;12:742736. [doi: 10.3389/fimmu.2021.742736] [Medline: 35095835]
68. Beerenwinkel N, Schmidt B, Walter H, et al. Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. Proc Natl Acad Sci U S A 2002 Jun 11;99(12):8271-8276. [doi: 10.1073/pnas.112177799] [Medline: 12060770]
69. Wang D, Larder B, Revell A, et al. A comparison of three computational modelling methods for the prediction of virological response to combination HIV therapy. Artif Intell Med 2009 Sep;47(1):63-74. [doi: 10.1016/j.artmed.2009.05.002] [Medline: 19524413]
70. Zazzi M, Incardona F, Rosen-Zvi M, et al. Predicting response to antiretroviral treatment by machine learning: the EuResist project. Intervirology 2012;55(2):123-127. [doi: 10.1159/000332008] [Medline: 22286881]
71. Heider D, Senge R, Cheng W, Hüllermeier E. Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction. Bioinformatics 2013 Aug 15;29(16):1946-1952. [doi: 10.1093/bioinformatics/btt331] [Medline: 23793752]
72. Riemenschneider M, Senge R, Neumann U, Hüllermeier E, Heider D. Exploiting HIV-1 protease and reverse transcriptase cross-resistance information for improved drug resistance prediction by means of multi-label classification. BioData Min 2016;9:10. [doi: 10.1186/s13040-016-0089-1] [Medline: 26933450]
73. Riemenschneider M, Hummel T, Heider D. SHIVA - a web application for drug resistance and tropism testing in HIV. BMC Bioinformatics 2016 Aug 22;17(1):314. [doi: 10.1186/s12859-016-1179-2] [Medline: 27549230]
74. Ramon E, Belanche-Muñoz L, Pérez-Enciso M. HIV drug resistance prediction with weighted categorical kernel functions. BMC Bioinformatics 2019 Jul 30;20(1):410. [doi: 10.1186/s12859-019-2991-2] [Medline: 31362714]
75. Tunc H, Sari M, Kotil S. Machine learning aided multiscale modelling of the HIV-1 infection in the presence of NRTI therapy. PeerJ 2023;11:e15033. [doi: 10.7717/peerj.15033] [Medline: 37020854]
76. Tunc H, Dogan B, Darendeli Kiraz BN, Sari M, Durdagi S, Kotil S. Prediction of HIV-1 protease resistance using genotypic, phenotypic, and molecular information with artificial neural networks. PeerJ 2023;11:e14987. [doi: 10.7717/peerj.14987] [Medline: 36967989]
77. Paremskaia A, Rudik AV, Filimonov DA, Lagunin AA, Poroikov VV, Tarasova OA. Web service for HIV drug resistance prediction based on analysis of amino acid substitutions in main drug targets. Viruses 2023 Nov 11;15(11):2245. [doi: 10.3390/v15112245] [Medline: 38005921]
78. Rojas Sánchez P, Cobos A, Navaro M, Ramos JT, Pagán I, Holguín Á. Impact of clinical parameters in the intrahost evolution of HIV-1 subtype B in pediatric patients: a machine learning approach. Genome Biol Evol 2017 Oct 1;9(10):2715-2726. [doi: 10.1093/gbe/evx193] [Medline: 29044435]

XSL•FO
RenderX

79. Ahlström MG, Ronit A, Omland LH, Vedel S, Obel N. Algorithmic prediction of HIV status using nation-wide electronic registry data. EClinicalMedicine 2019 Dec;17:100203. [doi: 10.1016/j.eclinm.2019.10.016] [Medline: 31891137]
80. Bogojeska J, Bickel S, Altmann A, Lengauer T. Dealing with sparse data in predicting outcomes of HIV combination therapies. Bioinformatics 2010 Sep 1;26(17):2085-2092. [doi: 10.1093/bioinformatics/btq361] [Medline: 20624779]
81. Zazzi M, Kaiser R, Sönnerborg A, et al. Prediction of response to antiretroviral therapy by human experts and by the EuResist data-driven expert system (the EVE study). HIV Med 2011 Apr;12(4):211-218. [doi: 10.1111/j.1468-1293.2010.00871.x] [Medline: 20731728]
82. Rosen-Zvi M, Altmann A, Prosperi M, et al. Selecting anti-HIV therapies based on a variety of genomic and clinical factors. Bioinformatics 2008 Jul 1;24(13):i399-i406. [doi: 10.1093/bioinformatics/btn141] [Medline: 18586740]
83. Pham T, Ghafoor M, Grañana-Castillo S, et al. DeepARV: ensemble deep learning to predict drug-drug interaction of clinical relevance with antiretroviral therapy. NPJ Syst Biol Appl 2024 May 6;10(1):48. [doi: 10.1038/s41540-024-00374-0] [Medline: 38710671]
84. Niemczak CE, Montagnese B, Levy J, et al. Machine learning for predicting cognitive deficits using auditory and demographic factors. PLoS ONE 2024;19(5):e0302902. [doi: 10.1371/journal.pone.0302902] [Medline: 38743715]
85. Singh Y. Machine learning to improve the effectiveness of ANRS in predicting HIV drug resistance. Healthc Inform Res 2017 Oct;23(4):271-276. [doi: 10.4258/hir.2017.23.4.271] [Medline: 29181236]
86. Ekpenyong ME, Edoho ME, Udo IJ, et al. A transfer learning approach to drug resistance classification in mixed HIV dataset. Informatics in Medicine Unlocked 2021;24:100568. [doi: 10.1016/j.imu.2021.100568]
87. Kimaina A, Dick J, DeLong A, Chrysanthopoulou SA, Kantor R, Hogan JW. Comparison of machine learning methods for predicting viral failure: a case study using electronic health record data. Stat Commun Infect Dis 2020 Sep 1;12(Suppl1):20190017. [doi: 10.1515/scid-2019-0017] [Medline: 37288469]
88. Kagendi N, Mwau M. A machine learning approach to predict HIV viral load hotspots in Kenya using real-world data. Health Data Sci 2023;3:0019. [doi: 10.34133/hds.0019] [Medline: 38487196]
89. Mamo DN, Yilma TM, Tewelgne MF, et al. Machine learning to predict virological failure among HIV patients on antiretroviral therapy in the University of Gondar Comprehensive and Specialized Hospital, in Amhara Region, Ethiopia, 2022. BMC Med Inform Decis Mak 2022;23(1):75. [doi: 10.1186/s12911-023-02167-7]
90. Bisaso KR, Karungi SA, Kiragga A, Mukonzo JK, Castelnuovo B. A comparative study of logistic regression based machine learning techniques for prediction of early virological suppression in antiretroviral initiating HIV patients. BMC Med Inform Decis Mak 2018 Sep 4;18(1):77. [doi: 10.1186/s12911-018-0659-x] [Medline: 30180893]
91. Kebede M, Zegeye DT, Zeleke BM. Predicting CD4 count changes among patients on antiretroviral treatment: application of data mining techniques. Comput Methods Programs Biomed 2017 Dec;152:149-157. [doi: 10.1016/j.cmpb.2017.09.017] [Medline: 29054255]
92. Ekpenyong ME, Etebong PI, Jackson TC. Fuzzy-multidimensional deep learning for efficient prediction of patient response to antiretroviral therapy. Heliyon 2019 Jul;5(7):e02080. [doi: 10.1016/j.heliyon.2019.e02080] [Medline: 31372545]
93. Maskew M, Sharpey-Schafer K, De Voux L, et al. Applying machine learning and predictive modeling to retention and viral suppression in South African HIV treatment cohorts. Sci Rep 2022 Jul 26;12(1):12715. [doi: 10.1038/s41598-022-16062-0] [Medline: 35882962]
94. Ogbechie MD, Fischer Walker C, Lee MT, et al. Predicting treatment interruption among people living with HIV in Nigeria: machine learning approach. JMIR AI 2023 May 12;2:e44432. [doi: 10.2196/44432] [Medline: 38875546]
95. Tibúrcio R, Barreto-Duarte B, Naredren G, et al. Dynamics of T-lymphocyte activation related to paradoxical tuberculosis-associated immune reconstitution inflammatory syndrome in persons with advanced HIV. Front Immunol 2021;12:757843. [doi: 10.3389/fimmu.2021.757843] [Medline: 34691079]
96. Ogishi M, Yotsuyanagi H. Prediction of HIV-associated neurocognitive disorder (HAND) from three genetic features of envelope gp120 glycoprotein. Retrovirology (Auckl) 2018 Jan 27;15(1):12. [doi: 10.1186/s12977-018-0401-x] [Medline: 29374475]
97. Babu H, Sperk M, Ambikan AT, et al. Plasma metabolic signature and abnormalities in HIV-infected individuals on long-term successful antiretroviral therapy. Metabolites 2019 Sep 30;9(10):210. [doi: 10.3390/metabo9100210] [Medline: 31574898]
98. Yin Y, Xue M, Shi L, et al. A noninvasive prediction model for Hepatitis B virus disease in patients with HIV: based on the population of Jiangsu, China. Biomed Res Int 2021;2021(1):6696041. [doi: 10.1155/2021/6696041] [Medline: 33860053]
99. Zhao J, Ma Z, Chen F, et al. Human immune deficiency virus-related structural alterations in the brain are dependent on age. Hum Brain Mapp 2021 Jul;42(10):3131-3140. [doi: 10.1002/hbm.25423] [Medline: 33755269]
100. Li B, Li M, Song Y, et al. Construction of machine learning models to predict changes in immune function using clinical monitoring indices in HIV/AIDS patients after 9.9-years of antiretroviral therapy in Yunnan, China. Front Cell Infect Microbiol 2022;12:867737. [doi: 10.3389/fcimb.2022.867737] [Medline: 35646738]
101. Mulyadi WJ, Qomariyah NN. Using machine learning to analyse the effect of antiretroviral therapy (ART) on people with HIV. Presented at: 2023 10th International Conference on ICT for Smart Society (ICISS); Sep 6-7, 2023; Bandung, Indonesia p. 1-5. [doi: 10.1109/ICISS59129.2023.10291717]

XSL•FO
RenderX

102. Jaganath D, Rajan J, Yoon C, et al. Evaluation of multi-antigen serological screening for active tuberculosis among people living with HIV. PLoS ONE 2020;15(6):e0234130. [doi: 10.1371/journal.pone.0234130] [Medline: 32497095]
103. Paul RH, Cho KS, Luckett P, et al. Machine learning analysis reveals novel neuroimaging and clinical signatures of frailty in HIV. J Acquir Immune Defic Syndr 2020 Aug 1;84(4):414-421. [doi: 10.1097/QAI.0000000000002360] [Medline: 32251142]
104. Blassel L, Tostevin A, Villabona-Arenas CJ, et al. Using machine learning and big data to explore the drug resistance landscape in HIV. PLOS Comput Biol 2021 Aug;17(8):e1008873. [doi: 10.1371/journal.pcbi.1008873] [Medline: 34437532]
105. Ragonnet-Cronin M, Golubchik T, Moyo S, et al. Human immunodeficiency virus (HIV) genetic diversity informs stage of HIV-1 infection among patients receiving antiretroviral therapy in Botswana. J Infect Dis 2022 Apr 19;225(8):1330-1338. [doi: 10.1093/infdis/jiab293] [Medline: 34077517]
106. Streeck H, Maestri A, Habermann D, et al. Dissecting drivers of immune activation in chronic HIV-1 infection. EBioMedicine 2022 Sep;83:104182. [doi: 10.1016/j.ebiom.2022.104182] [Medline: 35905559]
107. Benitez AE, Musinguzi N, Bangsberg DR, et al. Super learner analysis of real-time electronically monitored adherence to antiretroviral therapy under constrained optimization and comparison to non-differentiated care approaches for persons living with HIV in rural Uganda. J Int AIDS Soc 2020 Mar;23(3):e25467. [doi: 10.1002/jia2.25467] [Medline: 32202067]
108. Stockman J, Friedman J, Sundberg J, Harris E, Bailey L. Predictive analytics using machine learning to identify ART clients at health system level at greatest risk of treatment interruption in Mozambique and Nigeria. JAIDS Journal of Acquired Immune Deficiency Syndromes 2022;90(2):154-160. [doi: 10.1097/QAI.0000000000002947]
109. Herrera-Ibatá DM, Pazos A, Orbegozo-Medina RA, Romero-Durán FJ, González-Díaz H. Mapping chemical structure-activity information of HAART-drug cocktails over complex networks of AIDS epidemiology and socioeconomic data of U.S. counties. BioSystems 2015 Jun;132-133:20-34. [doi: 10.1016/j.biosystems.2015.04.007] [Medline: 25916548]
110. Michael Olusegun A. Identifying the limitation of stepwise selection for variable selection in regression analysis. AJTAS 2015;4(5):414. [doi: 10.11648/j.ajtas.20150405.22]
111. Camacho López PA, Latorre-Arevalo MG, Camacho-Naranjo P, Villabona-Florez SJ. Global research trends in artificial intelligence and type 2 diabetes mellitus: a bibliometric perspective. Cureus 2025 Jul;17(7):e88114. [doi: 10.7759/cureus.88114] [Medline: 40831862]
112. Algarvio RC, Conceição J, Rodrigues PP, Ribeiro I, Ferreira-da-Silva R. Artificial intelligence in pharmacovigilance: a narrative review and practical experience with an expert-defined Bayesian network tool. Int J Clin Pharm 2025 Aug;47(4):932-944. [doi: 10.1007/s11096-025-01975-3] [Medline: 40736603]
113. Dsouza VS. Artificial Intelli- Gence in Pharmacovigilance: A Systematic Review on Predicting Adverse Drug Reactions in Hospitalized Patients Research in Social and Admin- Istrative Pharmacy 2025. [doi: 10.1016/j.sapharm.2025.02.008]
114. Rudnisky E, Paudel K, Paudel KR. Pharmacovigilance in the era of artificial intelligence: advancements, challenges, and considerations. Cureus 2025 Jun;17(6):e86972. [doi: 10.7759/cureus.86972] [Medline: 40734859]
115. Lai S, Wei W, Yang S, et al. Machine learning-driven in-hospital mortality prediction in HIV/AIDS patients with cytomegalovirus infection: a single-centred retrospective study. J Med Microbiol 2024 Nov;73(11). [doi: 10.1099/jmm.0.001935] [Medline: 39606806]
116. Xie Z, Hu H, Kadota JL, et al. Prevention of adverse HIV treatment outcomes: machine learning to enable proactive support of people at risk of HIV care disengagement in Tanzania. BMJ Open 2024 Sep 24;14(9):e088782. [doi: 10.1136/bmjopen-2024-088782] [Medline: 39317499]
117. Chen Y, Pan K, Lu X, Maimaiti E, Wubuli M. Machine learning-based prediction of mortality risk in AIDS patients with comorbid common AIDS-related diseases or symptoms. Front Public Health 2025;13:1544351. [doi: 10.3389/fpubh.2025.1544351]
118. Li J, Ding Y, Hao Y, et al. Machine learning algorithms to predict the risk of admission to intensive care units in HIV-infected individuals: a single-centre study. Virol J 2025 Aug 5;22(1):267. [doi: 10.1186/s12985-025-02900-w] [Medline: 40764558]

Abbreviations

AI: artificial intelligence
ANN: artificial neural network
ART: antiretroviral therapy
AUROC: area under the receiver operating characteristic curve
CD4: cluster of differentiation 4
DDI: drug-drug interaction
EHR: electronic health record
HAART: highly active antiretroviral therapy
MDR-E: multidrug-resistant enterobacterial
ML: machine learning
PR: prevalence ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

XSL•FO
RenderX

RF: random forest
SVM: support vector machine

Edited by A Stone; submitted 17.Jun.2025; peer-reviewed by D Liu, J Ong; revised version received 06.Nov.2025; accepted 11.Nov.2025; published 28.Apr.2026.

Please cite as:
Boudra T, Idrissou A, Barakat O, Davani S, Valnet Rabier MB, Lagoutte-Renosi J
Machine Learning in HIV Care and Antiretroviral Therapy: Systematic Review
J Med Internet Res 2026;28:e79219
URL: https://www.jmir.org/2026/1/e79219
doi:10.2196/79219

© Thamina Boudra, Arafate Idrissou, Oussama Barakat, Siamak Davani, Marie-Blanche Valnet Rabier, Jennifer Lagoutte-Renosi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Patient Concerns Regarding Artificial Intelligence Applications in Health Care: Systematic Review and Meta-Synthesis Based on Social Ecological Theory

Jiayu Hou[1,2*], MSc; Zhiqiao Zhang[1,2*], MSc; Xuan Cheng[1], MSc; Weihong Wang[1,3], PhD

[1]School of Nursing, Hunan Normal University, No. 371, Tongzipo Road, Yuelu District, Changsha City, Hunan Province, China

[2]Kiang Wu Nursing College of Macau, Macao, China

[3]Medical Humanities Research Center, Hunan Normal University, Changsha, China

[*]these authors contributed equally

Corresponding Author:
Weihong Wang, PhD
School of Nursing, Hunan Normal University, No. 371, Tongzipo Road, Yuelu District, Changsha City, Hunan Province, China

Abstract

Background: The use of artificial intelligence (AI) in health care is growing quickly, but there is not enough research that looks at patient concerns from a multilevel perspective. Existing reviews predominantly summarize patient attitudes descriptively, lacking theoretical frameworks to explain the underlying mechanisms of these concerns.

Objective: This systematic review and meta-synthesis aimed to identify and analyze patient concerns regarding health care AI applications, using social ecological theory to reveal the multilevel interactive mechanisms of concern at the individual, interpersonal, organizational, and societal levels.

Methods: Following the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) guidelines, databases including PubMed, Embase, Web of Science, CINAHL, and Scopus were searched on March 1, 2026. Qualitative studies exploring patient perceptions of clinical AI applications were included, excluding those involving only healthy populations, technical performance, or nonclinical settings. Two researchers independently screened the literature and assessed methodological quality using the JBI-QARI (Joanna Briggs Institute Qualitative Assessment and Review Instrument) checklist. Confidence in synthesized findings was assessed using the GRADE-CERQual (Confidence in the Evidence from Reviews of Qualitative Research) approach.

Results: A total of 25 qualitative studies involving 528 participants from diverse patient groups across multiple countries were included. Six themes emerged: (1) microlevel worries about privacy and data security, including data breaches and loss of control over personal health information; (2) worries about the limits and reliability of technology, especially AI diagnostic accuracy and “black box” decision-making; (3) mesolevel effects on physician-patient relationships, including reduced face-to-face interaction and empathy; (4) trust and accountability issues, including unclear responsibility attribution and institutional oversight problems; (5) macrolevel ethical and equity issues, including algorithmic bias and health care access inequalities; and (6) worries about technology diffusion and possible replacement of health care workers.

Conclusions: This review represents the first meta-synthesis applying social ecological theory to construct patient concerns regarding medical AI. Unlike previous descriptive reviews, it reveals the interconnected “ecological imbalance” mechanisms at micro-, meso-, and macrolevels when AI is embedded in health care systems. The findings suggest that many patient concerns are based on facts rather than just misunderstandings, indicating that systemic rather than isolated interventions are needed. Practical implications include explainable algorithm design at the microlevel, improved physician-patient communication, and institutional accountability at the mesolevel, and coordinated global ethical norms and equity-promoting policies at the macrolevel. Limitations include the inclusion of studies primarily from developed regions, significant heterogeneity in AI application scenarios, and constraints inherent to secondary research. Nevertheless, addressing these multilevel concerns remains crucial for balancing technological advancement with patient-centered care and enabling sustainable AI integration.

Trial Registration: Trial Registration: PROSPERO CRD420251156502; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251156502

(J Med Internet Res 2026;28:e85663) doi:10.2196/85663

XSL•FO
RenderX

KEYWORDS

artificial intelligence; medical ethics; patient concerns; data privacy; physician-patient relationship; health care equity; meta-integration; social ecological theory

Introduction

The application of artificial intelligence (AI) technology in health care is rapidly advancing, particularly across multiple domains, including disease diagnosis, treatment decision support, personalized health management, image analysis, and drug discovery [1]. Through techniques such as deep learning and natural language processing, AI can extract critical insights from vast datasets, empowering clinicians to make more precise decisions. With breakthroughs in generative AI and large language models, AI capabilities have expanded into complex clinical decision support and physician-patient interactions, enhancing communication and improving patient outcomes through more tailored and responsive care [2]. As of 2024, the US Food and Drug Administration has approved over 950 AI or machine learning medical devices, with radiology accounting for 76% of these approvals [3]. In 2024, AI adoption among US physicians reached 66%, nearly doubling from 38% in 2023 [4]. AI implementation has significantly enhanced health care efficiency, optimized resource allocation, and delivered more personalized and precise treatment plans for patients [5]. For instance, AI can identify minute lesions during early screening and match genes with therapies in precision medicine, substantially improving cure rates and patient quality of life [6]. However, as AI technology becomes more widespread and deeply integrated into health care, patient concerns regarding ethics, privacy, and security have increasingly come to the fore [7]. These concerns not only impact patient acceptance of AI technology but also directly affect the smooth promotion and implementation of AI in medical practice [8].

When encountering AI-based medical applications, one of the patients' primary concerns is privacy protection and data security [9]. With the digitization of medical data and widespread AI adoption, patient health information is being collected, stored, and analyzed at an unprecedented scale. Many patients fear potential misuse or leakage of these sensitive data, particularly when data protection measures remain inadequate [10]. A 2023 national survey revealed that nearly half of US adults express low trust in the health care system's responsible use of AI [11]. Additionally, the transparency and explainability of AI decisions have become major patient concerns. Due to the "black box" nature of AI algorithms, patients often cannot understand how AI arrives at diagnostic and treatment recommendations. This opacity fuels skepticism about the reliability of AI systems [12] and hinders clinicians' ability to explain treatment decisions to patients [13]. Another core issue is the potential loss of patient decision autonomy, particularly in urgent situations such as emergency rooms. AI's "automated" decisions may deprive patients of choice and diminish their active role in the treatment process [14]. Furthermore, as AI gradually assumes certain medical tasks, patient concerns about the erosion of humanistic care are growing. Many patients believe that while AI can provide efficient diagnostic support, it cannot replace a physician's empathy and care. The resulting distancing in the physician-patient relationship may impact treatment outcomes and the overall patient experience [15]. A mixed methods survey of 600 US adults revealed that 30% of respondents expressed concern about AI's lack of a physician's "human touch," while 84.2% preferred AI for tasks unrelated to the physician-patient relationship, such as appointment scheduling [16].

Therefore, the application of AI in health care involves not only technological advancement but also profound shifts in physician-patient relationships, ethical principles, and cultural values. During the development of new technologies, it is important to fully consider and address patients' ethical concerns, need for privacy protection, and expectations for more humanized medical care. Existing qualitative syntheses primarily adopt a broad public perspective, empirically summarizing the advantages, risks, and recommendations of AI in health care [17], yet theoretical frameworks remain relatively underdeveloped. Given this context, this study aimed to synthesize existing literature systematically on patients' ethical concerns regarding AI through the Social Ecological Model framework. Specific research objectives are as follows:

1. To identify fundamental ethical issues (eg, privacy, trust, and humanistic care) that patients encounter with AI health care applications;
2. To analyze how these concerns influence patient acceptance of AI technology;
3. To reveal tension mechanisms at 3 levels—individual cognition, physician-patient interaction, and macrolevel institutional systems—to address potential "ecological imbalance risks."

Methods

Study Registration and Reporting Framework

This study strictly followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Checklist 1) [18] and the ENTREQ (Enhancing Transparency in Reporting the Synthesis of Qualitative Research; Checklist 2) [19]. This review aimed to explore patients' values, attitudes, and experiences regarding the application of AI in health care. The Cochrane Handbook recognizes qualitative evidence synthesis as the appropriate methodology for such questions [20]. As all 25 included studies used qualitative designs yielding narrative findings rather than quantitative effect estimates amenable to statistical pooling [21], statistical meta-analysis was not applicable to this review. We therefore used the Joanna Briggs Institute meta-aggregation approach [22]. The review was prospectively registered in PROSPERO (CRD420251156502). The registered protocol was subsequently updated to clarify the qualitative study design and the use of JBI-QARI (Joanna Briggs Institute Qualitative Assessment and Review Instrument) for quality assessment. In addition, the following methodological enhancements were made during the conduct of this review: (1) CINAHL was added as a fifth

XSL•FO
RenderX

database to improve coverage of nursing and allied health literature, (2) the GRADE-CERQual (Confidence in the Evidence from Reviews of Qualitative Research) approach was adopted to assess confidence in the synthesized findings, and (3) social ecological theory (SET) was used as a theoretical lens to interpret and discuss the findings.

Search Strategy

This study's search strategy followed the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) guideline (Checklist 3) [23]. The initial search was conducted on September 28, 2025. Following iterative refinement of the search strategy, updated searches were performed on January 4, 2026, and March 1, 2026, to capture the most recent literature. Five databases were searched from inception to the date of each search: PubMed (via National Library of Medicine), Embase (via Embase.com), Web of Science (via Clarivate, encompassing the Core Collection, KCI-Korean Journal Database, MEDLINE, ProQuest Dissertations & Theses Citation Index, SciELO Citation Index, and the Grants Index), CINAHL (via EBSCOhost), and Scopus (via Scopus.com); all other databases were searched independently through their respective platforms. The detailed search strategies for all databases are provided in Multimedia Appendix 1. A 3-step search strategy was used. First, an initial search was conducted in PubMed, and the titles, abstracts, and index terms of relevant records were analyzed to identify key search terms. Second, a comprehensive search using all identified keywords and index terms was undertaken across all databases. Third, the reference lists of included studies were hand-searched to identify any additional relevant studies. The search strategy was developed de novo for this review, and its effectiveness was validated by confirming the retrieval of known relevant studies. No formal peer review of the search strategy was conducted using standardized appraisal tools. This review focused on published qualitative research; clinical trial registries were not searched. Beyond the systematic database searches and reference list screening, no supplementary search methods were used, such as contacting authors, browsing conference proceedings, or setting up citation alerts. Only studies published in English or Chinese were included. The PubMed search strategy is presented below:

- #1 ("Artificial Intelligence"[MeSH]) OR ("Machine Learning"[MeSH]) OR ("Decision Support Systems, Clinical"[MeSH]) OR ("artificial intelligence"[tiab]) OR ("machine learning"[tiab]) OR ("deep learning"[tiab]) OR ("AI-based"[tiab]) OR ("AI-assisted"[tiab]) OR ("AI-driven"[tiab]) OR ("ChatGPT"[tiab]) OR ("large language model"[tiab]) OR ("clinical decision support"[tiab]) OR ("CDSS"[tiab]) OR ("generative AI"[tiab]) OR ("algorithm*"[tiab])
- #2 ("Patients"[MeSH]) OR (patient*[tiab]) OR (stakeholder*[tiab])
- #3 (healthcare[tiab]) OR ("health care"[tiab]) OR (medical[tiab]) OR (clinical[tiab])
- #4 ("Attitude to Health"[MeSH]) OR (concern[tiab]) OR (perception[tiab]) OR (perspective[tiab]) OR (attitude[tiab]) OR (trust[tiab]) OR (acceptance[tiab]) OR (barrier[tiab]) OR (ethical[tiab]) OR (privacy[tiab]) OR (view[tiab]) OR (opinion[tiab]) OR (experience[tiab]) OR (feeling[tiab]) OR (worry[tiab])
- #5 ("Qualitative Research"[MeSH]) OR (qualitative[tiab]) OR ("focus group"[tiab]) OR (interview[tiab]) OR ("thematic analysis"[tiab]) OR ("content analysis"[tiab]) OR ("grounded theory"[tiab]) OR (phenomenolog*[tiab]) OR (ethnograph*[tiab]) OR ("lived experience"[tiab]) OR ("narrative analysis"[tiab]) OR ("in-depth interview"[tiab]) OR ("qualitative study"[tiab]) OR ("qualitative research"[tiab])
- #6 #1 AND #2 AND #3 AND #4 AND #5

Inclusion and Exclusion Criteria

This review focused on patient concerns arising from the use of AI in clinical practice. The inclusion and exclusion criteria are presented in Table 1. A total of 19,090 records were retrieved and imported into EndNote (version 21; Clarivate Analytics). Through both automated and manual deduplication in EndNote 21, a total of 7132 (37.4%) duplicate records were identified, leaving 11,958 (62.6%) records for assessment based on the relevance of titles and abstracts. At the title and abstract screening stage, 11,874 (99.3%) records were excluded for the following reasons: not addressing AI applications in clinical health care settings (n=5818, 49%), not involving patients as participants (n=3871, 32.6%), ineligible study design (n=2030, 17.1%), and not published in English or Chinese (n=155, 1.3%). A total of 84 (0.7%) records met the selection criteria, and their full texts were retrieved for further evaluation. Following full-text assessment, 25 studies met the criteria for quality appraisal. Two authors independently conducted the screening process, and any disagreements regarding inclusion were resolved through consultation with a third author.

XSL•FO
RenderX

Table . Inclusion and exclusion criteria for qualitative studies on patient concerns regarding artificial intelligence (AI) applications in health care based on the PICOS framework.

Components	Inclusion criteria	Exclusion criteria
Participants	• Patients who receive medical services supported by AI technology and stakeholders.	• Stakeholders without patients and healthy people.
Phenomenon of interest	• The focus of the research is on the application of AI technology in health care, such as AI-assisted diagnosis, AI-assisted surgeries, AI health monitoring, and virtual nursing assistants. • The research focuses on patients' perception of AI applications, their psychological reactions, concerns, and emotions, such as concerns about privacy, fears of physician substitution, and trust in AI.	• It does not involve the application of AI technology in medical fields, or does not focus on the research of patients' psychological and emotional responses. • Only focusing on the technical aspects such as the performance and algorithm optimization of AI technology, without considering the patients' perception and emotions.
Context	• The research background focuses on AI application scenarios in the medical environment, including hospitals, clinics, telemedicine platforms, and public health monitoring.	• Research that is not conducted in actual medical settings or that is limited to technological development research in laboratory environments. • The research environment has no relation to the actual medical experience of patients and thus cannot provide in-depth information about patients' perceptions and reactions.
Study design	• The data collection method focuses on the subjective experiences of patients, paying attention to their emotional, cognitive, and worrying psychological reactions during their medical intervention with AI technology.	• Not a qualitative study. • The overly simplistic design, such as relying solely on short questionnaires or tool tests, fails to thoroughly explore the psychological and emotional responses of patients.

Quality Assessment

The JBI-QARI was used to evaluate the methodological rigor of each published study [24]. Questions answered "yes" received 1 point, and studies scoring 5 points or lower were deemed low quality and excluded from the synthesis. Two reviewers independently conducted rigorous assessments of the selected research reviews. Disagreements were resolved through discussion or consultation with a third reviewer within the team. Table 2 presents the quality assessment results for the studies included in this review.

XSL•FO
RenderX

Table . Methodological quality assessment of 25 included qualitative studies using the Joanna Briggs Institute Qualitative Assessment and Review Instrument critical appraisal checklist.

Studies	Q1[a]	Q2[b]	Q3[c]	Q4[d]	Q5[e]	Q6[f]	Q7[g]	Q8[h]	Q9[i]	Q10[j]	Total
Hurley et al [25]	Y[k]	Y	Y	Y	Y	Y	U	Y	Y	Y	9
Annamalai [26]	U	Y	Y	Y	Y	U	N[l]	Y	Y	Y	7
Čartolovni et al [27]	Y	Y	Y	Y	Y	U[m]	N	Y	Y	Y	8
Sujan et al [28]	Y	Y	Y	Y	Y	U	N	Y	Y	Y	8
Kostick-Quenet et al [29]	Y	Y	Y	Y	Y	Y	U	Y	Y	Y	9
Hesjedal et al [30]	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	9
McCradden et al [31]	Y	Y	Y	Y	Y	U	U	Y	Y	Y	8
Al-Anezi [32]	Y	Y	Y	Y	Y	U	U	Y	Y	Y	8
Freeman et al [33]	Y	Y	Y	Y	Y	U	U	Y	Y	Y	8
Viberg Johansson et al [34]	Y	Y	Y	Y	Y	N	U	Y	Y	Y	8
Jeyakumar et al [35]	Y	Y	Y	Y	Y	N	U	Y	Y	Y	8
Haan et al [36]	U	Y	Y	Y	Y	U	U	Y	Y	Y	7
Berger et al [37]	Y	Y	Y	Y	Y	U	U	Y	Y	Y	8
Khairat et al [38]	Y	Y	Y	Y	Y	U	U	Y	Y	Y	8
Litchfield et al [39]	U	Y	Y	Y	Y	U	U	Y	Y	Y	7
Omori et al [40]	Y	Y	Y	Y	Y	U	U	Y	U	Y	7
Funer et al [41]	U	Y	Y	Y	Y	N	U	Y	Y	Y	7
Giebel et al [42]	U	Y	Y	Y	Y	Y	Y	Y	Y	Y	9
Ly et al [43]	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	10
Richardson et al [44]	N	Y	Y	Y	Y	U	U	Y	Y	Y	7
Foresman et al [45]	Y	Y	Y	Y	Y	U	U	Y	Y	Y	8
Schneider et al [46]	N	Y	Y	Y	Y	N	N	Y	Y	Y	7
Zhang et al [47]	U	Y	Y	Y	Y	N	U	Y	Y	Y	7

XSL•FO
RenderX

Studies	Q1[a]	Q2[b]	Q3[c]	Q4[d]	Q5[e]	Q6[f]	Q7[g]	Q8[h]	Q9[i]	Q10[j]	Total
Gundlack et al [48]	U	Y	Y	Y	Y	N	U	Y	Y	Y	7
Steerling et al [49]	U	Y	Y	Y	Y	U	Y	Y	Y	Y	8

[a]Q1. Is there congruity between the stated philosophical perspective and the research methodology?

[b]Q2. Is there congruity between the research methodology and the research question or objectives?

[c]Q3. Is there congruity between the research methodology and the methods used to collect data?

[d]Q4. Is there congruity between the research methodology and the representation and analysis of data?

[e]Q5. Is there congruity between the research methodology and the interpretation of results?

[f]Q6. Is there a statement locating the researcher culturally or theoretically?

[g]Q7. Is the influence of the researcher on the research, and vice-versa, addressed?

[h]Q8. Are participants, and their voices, adequately represented?

[i]Q9. Is the research ethical according to current criteria or, for recent studies, and is there evidence of ethical approval by an appropriate body?

[j]Q10. Do the conclusions drawn in the research report flow from the analysis, or interpretation, of the data? the conclusions drawn in the research report flow from the analysis, or interpretation, of the data?

[k]Y: yes.

[l]N: no.

[m]U: unclear.

Data Extraction and Synthesis

Data were extracted from the studies included in this review using the JBI-QARI standardized data extraction tool. The first author extracted data from these 25 studies, including the first author and publication year, study population and sample size, research topic, study type, and primary outcomes, as detailed in Table 3. This study used a meta-synthesis approach [22]. Two researchers repeatedly read and interpreted the original studies, analyzed and interpreted the implications of the findings, grouped similar results into new categories, and then synthesized these categories into integrated outcomes to form new perspectives or interpretations. When disagreements arose between the 2 coders, a third coder was consulted.

The confidence in each synthesized finding was assessed using the GRADE-CERQual approach [50]. CERQual assesses confidence based on four components: (1) methodological limitations of the included studies, informed by the JBI-QARI critical appraisal results; (2) coherence of the finding across contributing studies; (3) adequacy of data supporting the finding, considering both the number of studies and the richness of data; and (4) relevance, defined as the extent to which the contexts of contributing studies are applicable to the review question. Each finding was assigned an overall confidence level of high, moderate, low, or very low. Two reviewers independently assessed each component and resolved disagreements through discussion. The CERQual assessment results are presented in Table 4. Heterogeneity across included studies was explored narratively by examining differences in AI application types, patient populations, and geographic contexts, as reported in Table 3 and discussed in the *Limitations* section.

XSL•FO
RenderX

Table . Characteristics of 25 included qualitative studies on patient concerns regarding artificial intelligence (AI) in health care (2019 - 2025).

Studies	Year	Country	Setting	Data collection	Research method	Participants	Interested topics	Main results
Hurley et al [25]	2024	United States	Not specified	Semistructured interviews	Phenomenological research	Stakeholders (including 20 patients)	Ethical considerations in integrating multimodal computer perception and neurotechnology in clinical care	Three main themes: 1. Perceived invasiveness of passive and continuous data collection 2. Data protection and security concerns 3. Ethical issues related to patients' awareness of data collection
Annamalai [26]	2020	India	Telepsychiatry	Semistructured interviews	Grounded theory approach	Stakeholders (including 14 patients)	Exploring challenges of AI-enabled telepsychiatry for clinical practice among urban Indian stakeholders	Four themes: 1. Ethical, legal, accountability, and regulatory issues 2. Challenges related to data 3. Health system infrastructure 4. AI related

XSL•FO
RenderX

Studies	Year	Country	Setting	Data collection	Research method	Participants	Interested topics	Main results
Čartolovni et al [27]	2023	Croatia	Hospital	Semistructured interviews	Phenomenological research	Stakeholders (including 15 patients)	Exploring multistakeholder (patients, physicians, and health care managers) insights into AI's impact on the patient-physician relationship	Four themes: 1. The current state of health care and the patient-physician relationship 2. Expectations of AI a synergetic effect between physicians 3. AI the future of health care 4. The patient-physician relationship
Sujan et al [28]	2022	United Kingdom	Hospital	Semistructured interviews	Phenomenological research	Stakeholders (including 4 patients)	Exploring perceptions of safety and safety assurance of health care AI (using AI-based ICU[a] infusion pumps as a case) among patients, hospital staff, technology developers, and regulators in the United Kingdom	Four themes: 1. The potential impact of health care AI 2. Requirements for human-AI interaction 3. Safety assurance practices and regulatory frameworks for AI and the gaps that exist 4. How incidents involving AI should be managed
Kostick-Quenet et al [29]	2024	United States	Cardiology	Semistructured interviews	Phenomenological research	Stakeholders (including 18 patients)		

XSL•FO
RenderX

Studies	Year	Country	Setting	Data collection	Research method	Participants	Interested topics	Main results
							Exploring trust criteria for an AI/ML[b]-based survival prediction algorithm (for LVAD[c] therapy in advanced heart failure) among patients, nurse coordinators, and physicians, focusing on normative and epistemic considerations	Three themes: 1. Epistemic trust considerations 2. Relational trust considerations 3. Personal belief-based trust considerations
Hesjedal et al [30]	2024	Norway	Prostate cancer diagnostics	Participant observation, semistructured interviews, and focus groups	Phenomenological research	Stakeholders (including 38 patients)	Investigating how scientists, MDs, and patients with PCa[d] relate ethical challenges of AI decision-making tools in PCa diagnostics to their understanding of "good health care," focusing on registers of valuing	• The main finding of the study is that medical physicians, scientists, and patients perceive ethical challenges with AI in prostate cancer diagnostics differently, shaped by their roles and experiences, which influence their understanding of good health care.
McCradden et al [31]	2020	Canada	Hospital	Semistructured interviews	Phenomenological research	30 patients and caregivers	Investigate perspectives on ethical issues surrounding AI in health care research among Canadian patients with meningioma, their caregivers, and health care providers	

XSL•FO
RenderX

Studies	Year	Country	Setting	Data collection	Research method	Participants	Interested topics	Main results
								Eight themes: 1. Consent 2. Privacy 3. Confidentiality 4. Responsibility 5. Accountability 6. Unintended consequences 7. Trust 8. Public engagement
Al-Anezi [32]	2024	Saudi Arabia	Chronic disease management	Semistructured interviews	Phenomenological research	29 patients	Analyze the feasibility of ChatGPT (free version 3.5) as a virtual health coach for chronic disease management, focusing on its ability to promote health literacy and support patients' self-management	• The main finding of this study is that ChatGPT shows potential as a virtual health coach by providing accessible lifestyle advice and motivation, but it faces limitations in accuracy, personalization, and trust that must be addressed before clinical use.
Freeman et al [33]	2024	Australia	Emergency department	Semistructured interviews	Grounded theory approach	28 patients	Health consumers' ethical concerns toward the use of AI in Australian emergency departments.	

XSL•FO
RenderX

Studies	Year	Country	Setting	Data collection	Research method	Participants	Interested topics	Main results
								Six themes: 1. Health consumer autonomy and consent 2. Physician decision-making and autonomy 3. Automation bias and reliance 4. Data use and privacy 5. Bias and discrimination 6. Regulation
Viberg Johansson et al [34]	2024	Sweden	Breast cancer screening	Semistructured interviews	Phenomenological research	16 patients	Swedish women's perceptions and attitudes toward the use of AI in mammography (as part of the national breast cancer screening program)	Three themes: 1. Perceived differences between AI and human assessment 2. Attitudes when implementing AI in mammography 3. Requirements when using AI in mammography
Jeyakumar et al [35]	2023	Canada	Acute or long-term medical centers	Semistructured interviews	Phenomenological research	12 patients	How patients view AI in health care and emphasize the need for trust, engagement, and strong data governance to ensure ethical and effective integration	Three themes: 1. Cultivating patients' trust fostering 2. Patient engagement establishing data governance 3. Validation of AI technologies
Haan et al [36]	2019	Netherlands			Grounded theory approach	20 patients		

XSL•FO
RenderX

Studies	Year	Country	Setting	Data collection	Research method	Participants	Interested topics	Main results
			Department of Radiology at a tertiary care academic institution	Semistructured interviews			Patient perspective on the use of AI in radiology, including awareness, uncertainties, and expectations	Six key domains: 1. Proof of technology (need for evidence on efficacy and reliability of AI) 2. Procedural knowledge (understanding how AI is implemented in practice) 3. Competence (capability of AI to produce reliable results) 4. Efficiency (faster scanning and reduced waiting times) 5. Personal interaction (importance of human communication for results) 6. Accountability (responsibility for errors made by AI)
Berger et al [37]	2025	Norway	Prostate cancer diagnostics (AI supported)	Semistructured interviews	Phenomenological research	18 patients	Patient perspectives on trust in AI-powered tools in prostate cancer diagnostics	

XSL•FO
RenderX

Studies	Year	Country	Setting	Data collection	Research method	Participants	Interested topics	Main results
								Three main dimensions of trust were identified: 1. Foundational trust 2. Interpersonal trust as mediator 3. Need for human oversight additional finding: participants were more forgiving of human errors than those made by AI, highlighting the relational and moral dimensions of trust in health care.
Khairat et al [38]	2025	United States	Cancer survivor organizations (recruitment); virtual/zoom (interviews)	Semistructured interviews	Phenomenological research	21 patients	Cancer survivors' experiences, facilitators, and barriers regarding the use of AI-based conversational tools (chatbots)	

XSL•FO
RenderX

Studies	Year	Country	Setting	Data collection	Research method	Participants	Interested topics	Main results
								Three overarching themes: 1. Preference for humans: participants strongly preferred interacting with health care professionals over chatbots. 2. Lack of empathy: chatbots were perceived as lacking the necessary empathy and emotional support. 3. Information and privacy concerns: concerns about information overload (generic and nonspecific responses) and data privacy.
Litchfield et al [39]	2025	United Kingdom	Primary care	Semistructured interviews	Phenomenological research	7 patients	Acceptability of "AmarDoctor," an AI-enabled translation and symptom-checking tool, among underserved Bangladeshi populations	Three overarching themes: 1. Enhanced accessibility and inclusivity 2. Anonymity for sensitive issues 3. Trust and safety concerns
Omori et al [40]	2025	Australia	Screening program	Semistructured interviews and focus groups	Grounded theory approach	26 patients		

XSL•FO
RenderX

Studies	Year	Country	Setting	Data collection	Research method	Participants	Interested topics	Main results
							Development of a typology of women's attitudes toward the use of AI in breast cancer screening	Four attitude types identified: 1. Enthusiasts 2. Practicalists 3. Traditionalists 4. Guardians
Funer et al [41]	2024	Germany	Surgery, nephrology, intensive home care	Semistructured interviews and focus groups	Phenomenological research	Stakeholders (including 18 patients)	Impacts of CDSSs[e] on the relationship, communication, and shared decision-making	Three overarching themes: 1. Impact on professional roles 2. Impact on relationship 3. Transparency needs
Giebel et al [42]	2025	Germany	Hospital	Semistructured interviews	Phenomenological research	Stakeholders (including 4 patients)	Opportunities to optimize AI-based CDSS and their integration into health care	Three overarching themes: 1. System optimization 2. User competence and support 3. Environmental framework
Ly et al [43]	2025	Australia	Ophthalmology	Semistructured interviews	Phenomenological research	Stakeholders (including 8 patients)	Stakeholder experiences, attitudes, enablers, barriers, and possible futures of digital diagnosis using AI for age-related macular degeneration in Australia	Three overarching themes: 1. Technological preferences 2. Divergent stakeholder priorities 3. Systemic and ethical considerations
Richardson et al [44]	2021	United States	Primary care	Focus groups	Phenomenological research	87 patients	Patient apprehensions and perspectives regarding the use of AI in health care	

XSL•FO
RenderX

Studies	Year	Country	Setting	Data collection	Research method	Participants	Interested topics	Main results
								Major themes identified: 1. Safety and human oversight 2. Preservation of choice 3. Cost, equity, and data concerns: cost, bias, security
Foresman et al [45]	2025	United States	Primary care, radiology, telehealth	Focus groups	Phenomenological research	17 patients	Patient perspectives on AI use in health care, specifically in diagnostic processes and communication	Five cross-cutting themes emerged: 1. Validation 2. Usability 3. Transparency 4. Opportunities, privacy, additional finding: comfort levels varied by interaction type; participants were most comfortable with "Digital Scribe" (low interaction) and least comfortable with "virtual human" (high interaction).
Schneider et al [46]	2025	Germany	Nephrology, surgery, home-ventilated care	Focus groups	Phenomenological research	18 patients	Patient perspectives on AI-based AI-CDSS, specifically focusing on trust, responsibility, and self-determination	

XSL•FO
RenderX

Studies	Year	Country	Setting	Data collection	Research method	Participants	Interested topics	Main results
								Three overarching observations: 1. Indecision and uncertainty 2. Shift in trust and responsibility 3. AI as support, not replacement
Zhang et al [47]	2021	United States	Radiology	Semistructured interviews	Phenomenological research	13 patients	Patients' perceptions and acceptance of using AI-based technology to interpret and comprehend radiology reports or imaging data	Four main findings: 1. Positive attitude and utility 2. Concerns 3. Trust and design requirements 4. Designers must ensure systems deliver concerning health results in an empathetic manner to optimize user experience.
Gundlack et al [48]	2025	Germany	Primary care and psychiatry	Semistructured focus groups	Phenomenological research	35 patients	Patients' perceptions of AI in medical care regarding caregiving relationships and ethics	

XSL•FO
RenderX

Studies	Year	Country	Setting	Data collection	Research method	Participants	Interested topics	Main results
								Four main themes: 1. AI perceived as beneficial for efficiency, data processing, and patient safety. 2. Key concerns include impersonality, data security, and overreliance on AI by medical staff. 3. Human interaction and emotional understanding deemed irreplaceable by AI. 4. Physicians considered primarily responsible for AI-related decisions; data transparency and privacy protection are essential for implementation.
Steerling et al [49]	2025	Sweden	Primary care	Semistructured interviews	Phenomenological research	14 health care professionals and 12 patients	Influences on trust in the use of AI-based triage in primary care	

XSL•FO
RenderX

Studies	Year	Country	Setting	Data collection	Research method	Participants	Interested topics	Main results
								Three main influences on trust: 1. Provision of accurate patient information (patients' capability and willingness). 2. Alignment with clinical expertise (standardized reasoning and experience-based knowledge). 3. Supervision of patients' health and safety (professionalism and guidance in information use). Both groups emphasized constructive dialogue and clear instructions for the use and storage of information.

[a]ICU: intensive care unit.

[b]ML: machine language.

[c]LVAD: left ventricular assist device.

[d]PCa: prostate cancer.

[e]CDSS: clinical decision support system.

XSL•FO
RenderX

Table . GRADE-CERQual summary of qualitative findings on patient concerns regarding artificial intelligence in health care.

Review finding	Contributing studies	Methodological limitations	Coherence	Adequacy	Relevance	Overall confidence
Theme 1 (3.1): Privacy and data security	13 studies	Minor concerns	No concerns	No concerns	No concerns	High
Theme 2 (3.2): Technical limitations and reliability	11 studies	Minor concerns	No concerns	No concerns	No concerns	High
Theme 3 (3.3): Impact on physician-patient relationship	11 studies	Minor concerns	No concerns	No concerns	No concerns	High
Theme 4 (3.4): Trust and accountability	13 studies	Minor concerns	No concerns	No concerns	No concerns	High
Theme 5 (3.5): Ethics and equity	8 studies	Minor concerns	No concerns	Minor concerns	Moderate concerns	Moderate
Theme 6 (3.6): Future outlook	10 studies	Minor concerns	No concerns	No concerns	No concerns	High

Results

Overview

This paper provides a systematic review and qualitative synthesis of ethical issues surrounding the application of AI in health care. A total of 25 qualitative studies involving 528 patients were included. Details of the search and screening process are presented in the PRISMA flow diagram (Figure 1). The findings were categorized into 14 themes and further consolidated into 6 key themes (Figure 2), elaborated as follows.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of study selection process (5 databases, inception to March 2026).

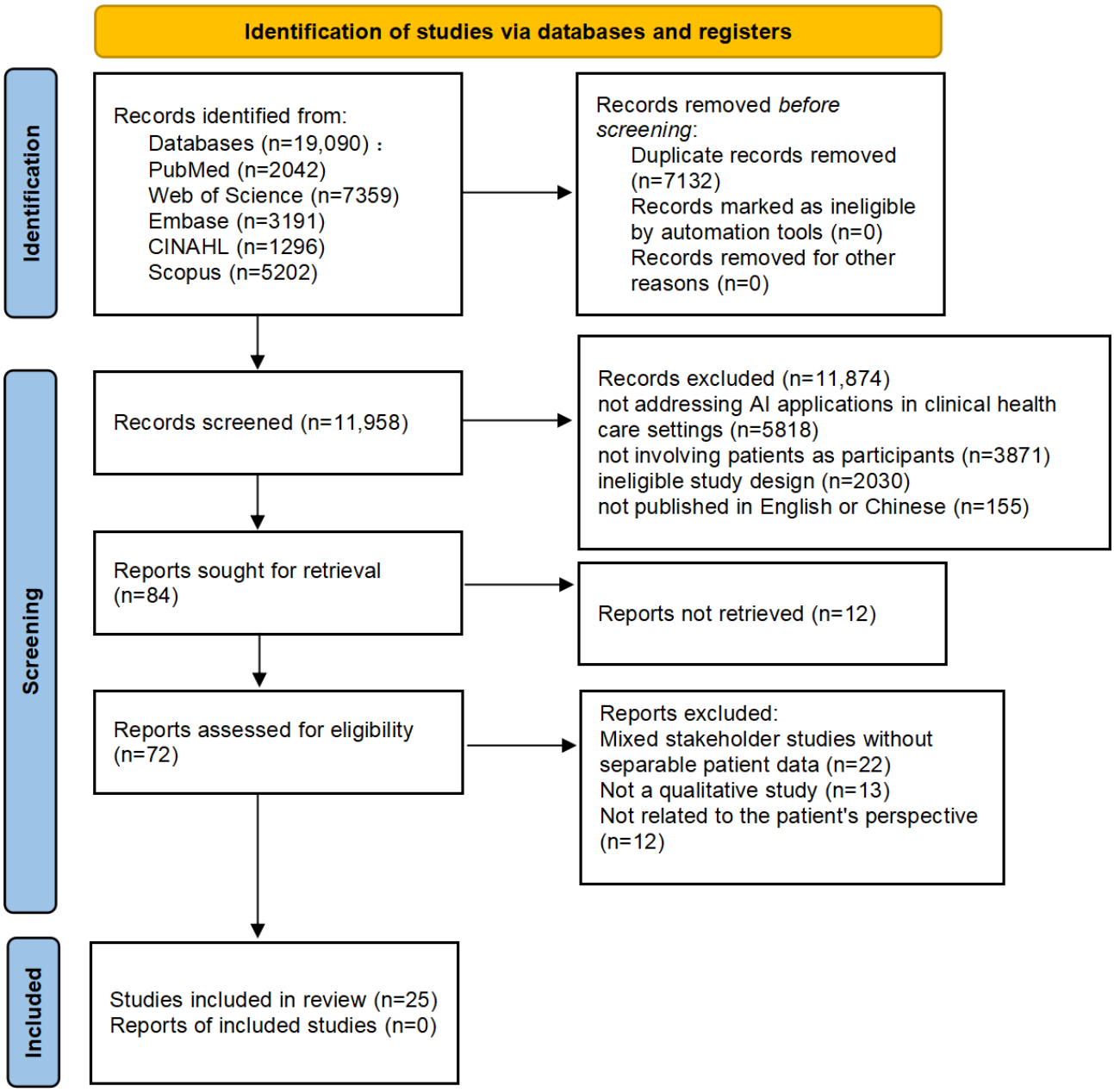

XSL•FO
RenderX

Figure 2. Social ecological theory of patient concerns regarding artificial intelligence (AI) in health care: thematic structure and multilevel interactions (25 studies, n=528).

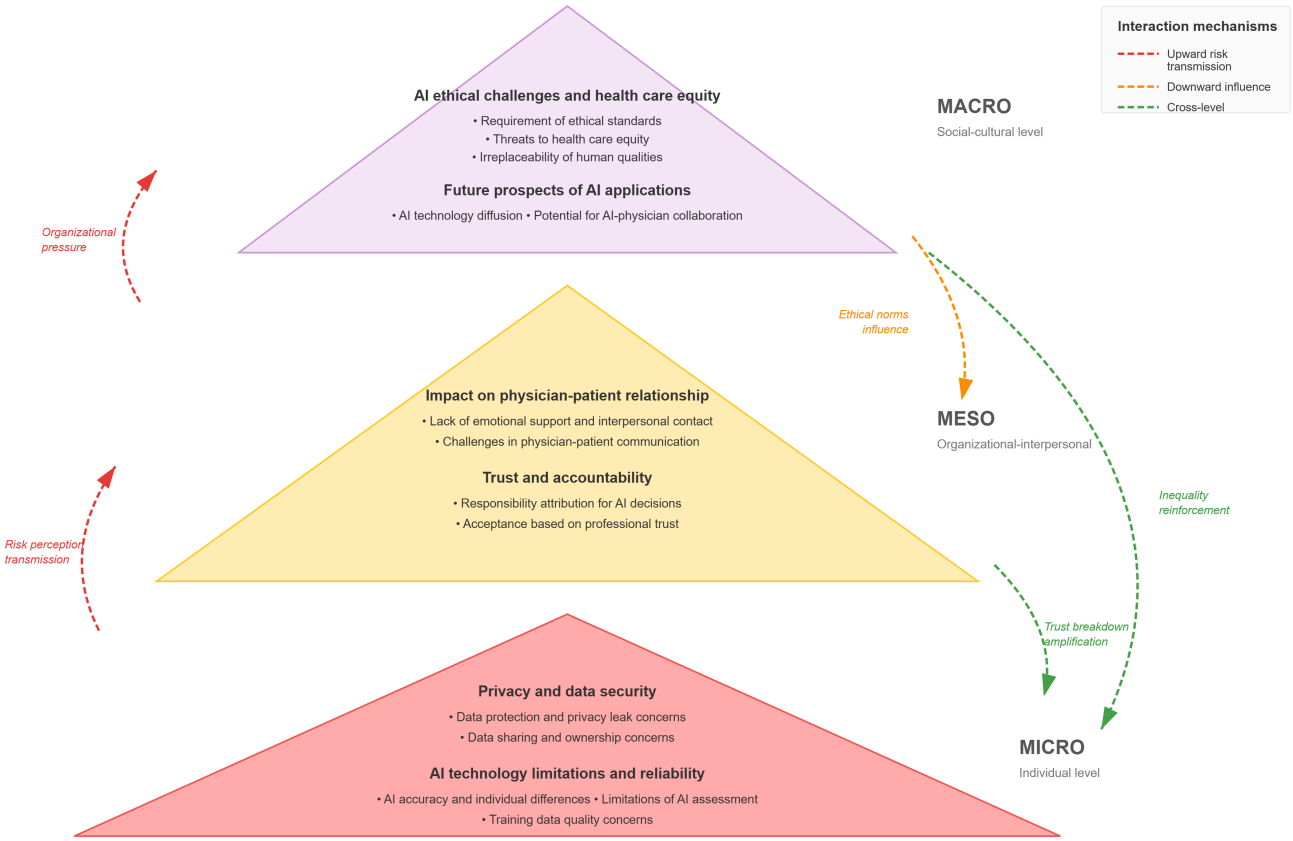

Privacy and Data Security

Concerns Over Data Protection and Privacy Breaches

Patients express concerns about the potential misuse of collected medical data, particularly in the absence of effective safeguards. Many believe that once data are uploaded to the internet, they cannot be completely erased and may even be exploited for commercial purposes. Some patients state they consent to data use only for medical purposes, provided they are explicitly informed of its scope:

> *They're going to use that [data] for whatever, and [it's like] the Internet, when you [post] something [and] it's stuck forever [there].* [P_05 25]
>
> *I am not confident enough in sharing my personal health data with ChatGPT, as it is based on AI, and there is a high possibility that the data could be misused.* [Unspecified 32]

Concerns Regarding Data Sharing and Ownership

Patients expressed confusion about data ownership, particularly when they were unclear about how their data were shared or used. They desire explicit prior notification before data use and fear unauthorized third-party access. Some patients demanded assurances that data use would not exceed the scope of their informed consent:

> *I would like to know beforehand so that they don't just send it off without informing me. It doesn't feel entirely right.* [P_08 34]
>
> *I'm worried that the data might be used by others without my permission, especially when I don't know the full extent of how it will be used.* [P_14 25]

Limitations and Reliability of AI Technology

AI Accuracy and Individual Variability

Patients generally believe that while AI algorithms can analyze vast amounts of data, they cannot fully adapt to each individual's unique health condition. A patient's health status is influenced by factors such as genetics, lifestyle, and environment—variables that algorithms often cannot fully predict. Particularly in the early stages of disease or complex scenarios, AI may fail to capture these individual differences, resulting in assessments that do not precisely align with each patient's specific needs. Therefore, patients view AI primarily as a reference tool, asserting that final diagnostic and treatment decisions should remain with experienced physicians:

> *The accuracy would be kind of iffy, because everybody has their own risk factor...* [P_09 29]
>
> *I'm fully aware that everybody's situation is unique to their own individual set of circumstances...* [P_03 29]

Limitations of AI Assessment

Although AI can process vast amounts of information and provide data-driven analysis, patients believe it cannot comprehensively account for all factors potentially affecting health. For instance, AI struggles to anticipate the impact of nonroutine variables such as unexpected events, environmental

XSL•FO
RenderX

changes, or sudden health issues. Patients noted that AI assessments are typically built on historical data, failing to dynamically reflect current health status or psychological state. This limitation frequently constrains AI assessments to broad contexts. Consequently, patients view AI assessments as supplementary tools rather than complete replacements for traditional physician judgment:

> *There are so many variables that you don't understand or can't predict...* [P_11 29]
>
> *A lot of these assessments aren't taking things into consideration like diabetes, exercise levels...* [P_04 29]

Concerns Over Training Data Quality

Patients' doubts about AI reliability extend beyond the algorithms themselves to the quality of their training data. They worry that medical AI systems' performance heavily depends on the data they learn from, yet real-world electronic health records often suffer from incomplete documentation, outdated information, or human input errors. If AI is trained on such flawed data, even the most sophisticated algorithms struggle to produce accurate, reliable outputs. Patients are further unsettled by the invisible risk of "garbage in, garbage out"—they cannot know what data the AI has learned from, nor whether it accurately reflects patients' actual health conditions:

> *So I've had a lot of different things in my medical chart that are inaccurate, very inaccurate, so if they're training [AI on that]...* [Unspecified 44]
>
> *I don't understand how it learns. If it learns wrong, who corrects it?* [Participant 11 37]

The Impact of Physician-Patient Relationships

The Lack of Emotional Support and Human Interaction

While AI can provide more precise diagnoses and treatment recommendations on a technical level, patients emphasize that a physician's emotional support and human interaction remain indispensable parts of the medical process. Especially when facing serious illnesses, patients require not only medical treatment but also emotional support from their physicians. AI cannot comprehend patients' emotional fluctuations, anxieties, or unease. This lack of emotional resonance unsettles many patients. They believe that while AI can assist in treatment, it cannot replace the emotional care provided by health care professionals—a crucial element in the therapeutic process.

> *People will still want human contact. Machines might be able to do a damned good job, but I think people still need that little human contact...* [P_01 28]
>
> *Excessive digitalization reduces personal contact between people, and that, in turn, reduces communication and connection...*
>
> *[P_04] [27]*

Challenges in Physician-Patient Communication

Patients widely agree that effective communication is fundamental to building trust between physicians and patients. Face-to-face interaction helps physicians better understand patients' conditions while making patients feel respected and cared for. With AI integration, patients worry physicians may overrely on technology, reducing face-to-face interactions. They emphasize that while AI improves efficiency, it risks neglecting deep patient engagement, which is essential for building trust and understanding in the physician-patient relationship. Especially during complex disease treatments, patients hope physicians will listen more to their needs and feelings rather than solely relying on algorithmic decisions.

> *It is crucial that he listen to me and that I listen to him. That's really the most important thing... [27]*
>
> *I wouldn't only rely on Cronko [AI]; I would use Cronko alongside a physician...* [P_13 27]

Trust and Accountability

Responsibility for AI Decision-Making

Patients widely express concern about unclear accountability when AI errors occur. AI decisions often lack the experience and emotional judgment of human physicians, making it difficult for patients to determine responsibility in the event of medical incidents. While acknowledging AI's potential to enhance efficiency and accuracy, patients indicate that the absence of a clear accountability mechanism fosters significant distrust toward AI. Particularly when AI systems malfunction, patients are uncertain whether responsibility lies with developers, hospitals, or health care providers, creating unease about adopting AI-driven medical technologies. Additionally, patients emphasize the right to question and correct AI judgments. They argue that when AI makes decisions based on flawed data or assumptions, patients should have avenues to raise objections and demand corrections, rather than passively accepting machine conclusions:

> *I believe the doctor always has the responsibility to be checking for you, and you're his responsibility, you know? The AI is not responsible; that's just a tool.* [Unspecified 44]
>
> *So I'd rather know what they're observing, and if it's [AI] wrong, I would [want to be able to] correct it rather than have them just collect data and make assumptions.* [Unspecified 44]

Acceptance Based on Professional Trust

Despite reservations about AI, patients accept it when recommended by their medical team due to trust in their physician or health care institution. Patients believe that the medical team's professional endorsement and experience enhance trust in AI systems. Many patients indicated that they would be more receptive if the medical team confirmed that the AI technology is thoroughly validated and offers benefits. Patients view physician recommendations and oversight as the primary basis for trusting AI technology. The professional team's opinion serves as a safety net for patients, particularly when emerging technologies remain unproven.

> *I trust it pretty much because I did ask how long they [the clinical team] have been doing LVAD, and I was surprised to find out that this had been going on longer than I thought it was... [29]*

XSL•FO
RenderX

Ethical Challenges of AI and Health Care Equity

Requirement of Ethical Standards

Patients believe that the application of AI in health care requires strict ethical standards, particularly when handling sensitive data. The absence of reasonable regulations may compromise patient rights. Many patients worry that AI research may prioritize speed over ethical standards, leading to data misuse or privacy violations. Additionally, patients emphasize the right to informed consent regarding AI interventions. Some patients believe health care institutions should disclose AI use before consultations, not during them, allowing time to process information and prepare mentally. Therefore, AI technology must operate under rigorous ethical review and legal frameworks to safeguard patients' fundamental rights:

> *I am afraid that [...] those who do research would like to reach a result...[but] the result should be so well grounded that you can actually vouch for it all the way.* [M30 30]

> *I want to know about it [AI involvement] before I get to the doctor... I'd rather have the opportunity to think about it and review it beforehand.* [Participant 45]

Threats to Health Care Equity

AI's impact on health care equity presents a complex duality. Some patients fear AI may exacerbate existing inequalities—low-income groups and the technologically disadvantaged may struggle to access AI-enabled health care equitably, thereby widening health disparities. However, evidence from marginalized communities reveals another possibility: minority patients view AI as a "safe space" enabling them to discuss sensitive issues without cultural shame, while multilingual capabilities compensate for dialect service gaps and bypass human "gatekeepers" perceived as biased.

> *At that moment, I didn't want to speak to anyone... If I had this tool, I would use it first before speaking to anyone else.* [P01, male, aged 28 y 39]

> *I feel that patients with fewer resources or access to technology will fall behind in terms of benefiting from AI in health care.* [Unspecified 26]

The Irreplaceability of Human Care

Patients believe that while AI can provide technical support, it cannot replace the human care and ethical judgment physicians offer in treatment. AI cannot comprehend patients' emotional needs or life contexts, factors crucial to treatment decisions. Patients emphasized that AI should only serve as an auxiliary tool, unable to substitute for physicians' roles in complex medical decisions—particularly regarding emotional support and personalized treatment.

> *Humans can express emotions, empathy, help, and give hope for a better tomorrow better than any machine.* [P_04 27]

Future Outlook for AI Applications

AI Technology Diffusion

Patients believe that the advancement of AI in health care is inevitable. Despite concerns, they acknowledge AI's potential to enhance efficiency and reduce errors, particularly in areas such as diagnostics, treatment planning, and patient management. As AI becomes widely adopted across global industries, patients recognize that it will become a vital component of future health care—although privacy and ethical issues arising from the technology require further regulation to ensure that patient data are protected and that AI systems are used responsibly in clinical settings.

> *It's becoming more and more the case that robots are controlling much of everyday life in various professions, and it's happening everywhere.* [M12 30]

Potential for AI-Physician Collaboration

Patients hold a positive view of AI collaborating with physicians, believing it can alleviate physicians' workload and improve efficiency, but it cannot replace the physician's decision-making role. Patients hope AI serves as an auxiliary tool to assist physicians rather than dominate the treatment process. Especially in complex conditions, the physician's clinical judgment and humanistic care remain crucial.

> *I think the whole system makes sense, but it can't function independently.* [P_13 27]

Discussion

Principal Findings

This study aimed to conduct an in-depth analysis of patients' ethical concerns regarding AI medical applications, explore their impact on patient acceptance and trust, and use SET to reveal multilevel tension mechanisms. SET, which is based on Bronfenbrenner's ecological systems theory [51], was later improved by researchers such as McLeroy and has become a well-known way to study public health and health behavior [52,53]. The theory's core assumption posits that individual behavior does not exist in isolation but is embedded within a multilevel, interacting environmental system. The microlevel focuses on individual knowledge, attitudes, and psychological characteristics; the mesolevel involves interpersonal relationships and organizational environments; and the macrolevel encompasses sociocultural norms and policy systems [52,54]. These 3 levels form a complex ecosystem through dynamic bidirectional influences, where changes in any one level may propagate to others.

Through a meta-synthesis of 25 qualitative studies, we identified 6 core themes: privacy and data security, technological reliability, impacts on physician-patient relationships, trust and accountability, ethical challenges and health care equity, and future perspectives. Socioecological analysis reveals that these concerns create mutually reinforcing "ecological imbalances" across micro (technological cognitive biases and data control anxieties), meso (broken physician-patient trust and institutional accountability gaps), and macro (health inequities and lagging

XSL•FO
RenderX

ethical standards) levels, thereby hindering patient acceptance and trust in AI technologies. Using this framework, the following sections will examine these findings in detail.

Microlevel: The Vicious Cycle of Technological Cognitive Bias and Data Control Anxiety

At the individual level, patients' concerns about AI technology primarily manifest as cognitive uncertainty about the technology and anxiety over data control. SET emphasizes the interaction between individual behavior and multilevel environments, where individuals' cognitive and emotional attitudes are influenced not only by internal factors but also closely tied to their social and cultural contexts [55]. During the application of AI medical technologies, individuals' cognitive biases about the technology are often accompanied by intense concerns about its transparency and controllability, which can lead to hesitance in adopting these technologies for their health care needs.

Research indicates that patients' understanding of AI technology is closely linked to their acceptance of it [56]. However, the complexity of AI technology and the black box nature of algorithms make it difficult for patients to comprehend its decision-making processes. This uncertainty directly leads to patients questioning the technology's accuracy [57]. For instance, patients worry that AI cannot account for their individualized health variations—such as differences in underlying conditions or lifestyle habits—potentially compromising the effectiveness of medical interventions [58]. Furthermore, patients' doubts about AI accuracy have extended from the algorithmic level to the data level. Drawing from experiences reviewing their own electronic health records, they have identified numerous errors in the documentation. Consequently, they fear that even well-designed algorithms trained on such flawed data may struggle to produce reliable outputs. This technological cognitive bias does not exist in isolation; it is intertwined with individuals' anxieties over data control. Patients' concerns about data control cannot be reduced to instrumental worries about privacy breaches but should be understood as existential threats to their self-integrity and narrative sovereignty [59]. Patients' anxiety over data control extends beyond privacy leaks to a profound fear of losing self-determination [60]. Medical data encapsulate patients' life experiences, health statuses, and identity information. When processed by algorithms for unforeseen purposes, patients lose not only informational control but also dominion over their own health narratives [61-63].

SET posits that individual behavior is shaped by multiple interrelated factors [55]. At the microlevel, patients' technological cognitive biases and data control anxieties do not exist in isolation but form a vicious cycle through the "cognition-anxiety" interaction [64]. Distrust of AI technology not only heightens patients' concerns about data security but also deepens their skepticism regarding technological transparency and controllability [65]. This vicious cycle may lead to patient resistance toward AI technology, subsequently affecting their overall attitude toward health care services. Similar perspectives are reflected in other studies. Research indicates that uncertainty about information technology and privacy concerns often amplify individual resistance, thereby influencing their acceptance of technology [66].

Mesolevel: Resonance Between Fractured Physician-Patient Trust and Institutional Accountability Deficits

At the mesolevel, patient concerns primarily manifest as fractures in physician-patient trust and inadequacies in health care organizational governance. Interpersonal relationships and organizational culture often constrain individual behavioral changes, according to SET [55]. Patients' apprehensions toward AI medical technologies stem not only from individual perceptions of the technology but are also closely intertwined with the quality of physician-patient relationships and the accountability of health care organizations.

Extensive research indicates that emotional empathy and interpersonal interaction between physicians and patients are crucial for building trust [67-69]. However, the introduction of AI has undermined this foundation to some extent. Studies reveal that when physicians overrely on AI technology, patients experience reduced interaction time and emotional support from their physicians, leading to diminished trust [70]. The core of the physician-patient relationship lies in physicians viewing patients as whole individuals, not machines requiring repair [71]. This relationship is based on listening, understanding, and being there for each other. Physicians know how much their patients are hurting and help them by talking to them [72-74]. AI intervention threatens the quality of this interpersonal engagement. When physicians' attention shifts to screens, when diagnoses rely on algorithmic outputs, and when communication is replaced by standardized processes, patients transform from "people receiving care" into "objects undergoing testing" [70,75].

Conversely, the absence of accountability mechanisms within health care organizations is another significant source of patient concern. Existing research indicates that many institutions lack clear AI liability frameworks, leaving patients unable to identify responsible parties when AI systems malfunction [76]. This lack of accountability not only makes patients less trusting of AI technology but also makes health care organizations look less credible [77]. SET emphasizes that deficiencies in internal organizational governance structures amplify individual-level concerns by eroding patients' trust in the health care system [78]. The absence of accountability in AI decision-making within health care organizations leaves patients without effective avenues for redress when encountering medical issues, thereby creating a mutually reinforcing negative cycle of "accountability-trust" [79]. Similar "accountability vacuum" issues have been explored in other studies, with research indicating that patients' trust in medical decisions is often severely compromised when health care organizations lack clear responsibility frameworks [80].

Macrolevel: Intensifying Issues of Health Care Equity and Lagging Ethical Standards

At the macrolevel, patient concerns regarding AI technology primarily center on health care equity and the lag in ethical regulations. SET asserts that sociocultural inequities and the

XSL•FO
RenderX

lack of ethical standards directly impact individual behavioral choices [81]. During the implementation of AI health care technologies, widespread adoption encounters obstacles arising from socioeconomic disparities. Low-income groups and patients with limited technological literacy often struggle to access AI medical services equitably. This gap exacerbates health care inequalities, fueling patient resistance toward AI technologies [82,83].

Not all socioeconomic strata equally benefit from the application of AI health care technologies, especially in regions with slower economic development or scarce resources [84]. High-end medical facilities and advanced technological resources are concentrated in major cities and economically developed regions, while low-income communities and remote areas still face significant gaps in AI health care adoption [85]. Furthermore, the high technical support and maintenance requirements of AI health care technologies make them unaffordable for many resource-constrained medical institutions, further exacerbating the unequal distribution of health care resources [86]. In this context, the "technological access barriers" experienced by patients are not merely technical difficulties but deep-seated social problems stemming from unequal socioeconomic structures and resource allocation [87].

Although the World Health Organization (WHO) has issued global AI ethics principles (WHO, 2021), implementation and regulatory rigor vary significantly across countries [88]. The lag in ethical standards manifests not only at the policy level but also creates gaps in practical implementation. For instance, some countries may lack cross-regional ethical collaboration mechanisms, leading certain health care institutions to prioritize AI services for economically advantaged groups over low-income populations due to cost considerations [89]. Such practices exacerbate health inequalities, deepening patients' concerns that AI serves only select groups rather than benefiting the broader public [90].

Interaction Mechanisms and Mutual Influence Across Levels

According to SET, concerns at the microlevel, mesolevel, and macrolevel do not exist in isolation but form mutually reinforcing chain reactions through "risk perception transmission" [55]. Research indicates that individual cognition and anxiety at the microlevel generate reverse effects at interpersonal and organizational levels, subsequently impacting broader social structures [91]. At the microlevel, individuals' cognitive biases regarding AI technology and anxieties over data control rights are transmitted to the mesolevel through interactions between individuals and health care organizations, increasing communication pressures for these organizations when deploying AI technologies [92,93]. Patients' concerns about data privacy and security influence individual behavioral decisions and prompt health care institutions to reevaluate the boundaries of AI applications, thereby driving societal-level attention to AI ethical norms [94].

The absence of accountability mechanisms for medical organizations at the mesolevel exacerbates microlevel cognitive biases and anxieties, leading to increased distrust among patients and further complicating the integration of AI technologies in health care [93,95]. Extensive research indicates that the breakdown of social support networks directly impacts individual health decisions and attitudes, with trust deficits further intensifying emotional distress and technological apprehension [96-98]. The breakdown of physician-patient trust not only reduces patient acceptance of AI technology but also leads patients to rely more on personal emotional judgments when facing medical decisions, overlooking the potential of AI technology [99].

Social equity issues at the macrolevel further permeate the microlevel, particularly the rejection of AI health care services by low-income groups. This affects their acceptance of AI technology and reinforces fears of technological displacement through sociocultural perceptions [17,100]. SET indicates that sociocultural beliefs not only shape patients' perceptions of AI technology through individual behavior but also influence organizational behavior via mesolevel cultural diffusion, thereby amplifying implementation barriers in unequal societies [101,102].

Policy Recommendations: Specific Pathways and Response Strategies

Reflecting on the findings of this study, patients' concerns are far from unfounded. Across the 25 studies included in this review, patients from diverse clinical contexts and cultural backgrounds consistently expressed anxieties about data security, skepticism toward algorithmic opacity, and questions about accountability attribution. These concerns point to objective limitations of AI medical technologies at their current stage of development and institutional gaps that remain unaddressed. Recent studies have independently corroborated the reality of these issues from various perspectives: data bias and privacy risks in clinical AI systems have been extensively documented [103], the erosion of physician-patient trust caused by algorithmic inexplicability has attracted sustained attention [104], and accountability attribution in AI-assisted medical decision-making still lacks clear legal delineation [105]. The WHO, in its 2021 guidance, also identified transparency, accountability, and equity as core principles for the ethical governance of AI in health care [106]. In other words, the concerns articulated by patients based on their lived experiences correspond precisely with the risks identified through systematic analyses by the academic community and international organizations. This implies that the goal of policy intervention should not be to "correct patient misconceptions," but rather to substantively address these well-founded and evidence-based concerns. The following recommendations are organized across the microlevel, mesolevel, and macrolevel while acknowledging the implementation challenges inherent in each pathway.

Microlevel: Incremental Transparency and Patient Empowerment

Given that the field of Explainable Artificial Intelligence remains in its developmental stages, fully visualizing algorithmic decision-making logic is not realistic in the short term; an incremental transparency strategy should therefore be adopted [107]. Specifically, drawing on the access rights for data subjects under the European Union's General Data Protection Regulation [108], health care institutions should be required to establish

XSL•FO
RenderX

data access logging systems that enable patients to query the access records and intended purposes of their health data [109]. To address patients' concerns regarding the quality of training data, accessible mechanisms for medical record review and correction should be established, safeguarding patients' rights to audit and amend their own medical records [110]. Furthermore, informed consent regarding AI involvement should be front-loaded—patients should be notified before their clinical encounter—and channels for questioning AI judgments and requesting human review should be established [111].

Mesolevel: Rebuilding Trust and Clarifying Accountability

The principle of "human oversight, AI assistance" should be institutionally safeguarded. Rather than imposing rigid communication time benchmarks, mandatory "physician confirmation checkpoints" should be embedded within AI-assisted diagnostic and treatment workflows, ensuring that critical decisions are subject to physician review and explained to patients before implementation [112]. Concurrently, performance evaluation systems should be adjusted to prevent efficiency-driven metrics from encroaching upon the space for physician-patient communication [113]. Regarding accountability attribution, a tiered liability framework should be constructed that differentiates the responsible parties for data errors, algorithmic defects, and clinical misjudgments [114], while acknowledging that existing legal frameworks contain gaps in the attribution of responsibility for AI-driven decisions, necessitating legislative follow-up [115]. Additionally, clinical scenarios in which AI cannot substitute for human practitioners should be explicitly delineated, particularly those highly dependent on emotional support and clinical judgment, such as end-of-life care, mental health treatment, and complex decision-making situations that require empathy and nuanced understanding of patient needs [116].

Macrolevel: Advancing Equitable Access and Ethical Governance

The WHO has set out global ethical standards for AI in health care, which can be used to help create an international ethical framework. National governments should build upon this foundation to develop regulatory frameworks aligned with their respective health care systems and cultural traditions, rather than pursuing an unrealistic goal of globally uniform standards. At the level of resource allocation, investment should be increased for low-income populations and remote areas, and age-friendly and low-cost AI health care tools should be developed to narrow the gap in technological accessibility [117]. However, it must be acknowledged that the barriers facing these regions extend beyond equipment scarcity to include insufficient technical maintenance capacity and digital health literacy, necessitating complementary capacity building and educational support [118]. Furthermore, equity assessments should be incorporated into the market approval review of AI health care products, requiring developers to submit impact assessment reports for diverse population groups [119].

Finally, it must be acknowledged that structural tensions exist in the advancement of AI in health care: technology developers' pursuit of algorithmic efficiency and commercial returns may conflict with patients' safety needs; health care institutions under cost pressures may find it difficult to reconcile these with adequate physician-patient communication. Confronting rather than evading these tensions is a prerequisite for formulating pragmatic policies [120].

Research Limitations

This study has the following limitations. First, the 25 included studies were primarily conducted in developed regions, such as North America, Europe, and Australia, with insufficient representation of patients from developing countries. This limits the generalizability of findings to resource-constrained settings. Second, the included studies covered a wide range of AI applications—including imaging diagnostics, clinical decision support, and virtual health assistants—where patient concerns may vary across different AI types. The thematic synthesis in this study may obscure such context-specific differences. Third, as a secondary analysis, the quality of the meta-synthesis depends on the reporting depth of the original studies. Some studies provided insufficient contextual information for patient quotes, hindering more nuanced interpretation. Finally, AI medical technology is undergoing rapid iteration. The literature included in this study reflects patient perceptions within a specific time window. As technological transparency and regulatory frameworks evolve, patient attitudes may change, potentially leading to increased trust in AI medical technology and greater acceptance of its use in health care. Future research should continuously track this dynamic process.

Conclusions

This research, through a systematic synthesis of 25 qualitative studies, identified 6 primary patient concerns regarding AI health care applications: privacy and data security, technological reliability, the impact on physician-patient relationships, trust and accountability, ethics and fairness, and ambivalent attitudes toward future developments. Unlike previous reviews focusing on the general public, this study centers on patients as core stakeholders. It pioneers the application of SET to this field, revealing a "disrupted ecological equilibrium" mechanism that propagates across microlevel, mesolevel, and macrolevel. This provides an explanatory framework—transcending descriptive induction—for understanding the deep-seated reasons behind patient resistance to AI medical technologies. The findings offer direct implications for practice: clinical institutions should establish a "human-led, AI-assisted" diagnostic model; policymakers should accelerate liability legislation and prioritize equitable technology access; and developers should adopt incremental transparency strategies while providing patients with avenues for questioning and review. Future research may explore the following directions: incorporating perspectives from patients in more developing countries and resource-constrained regions to test cross-cultural applicability; conducting comparative studies across different AI application scenarios to explore the context-specific nature of concerns; and using longitudinal designs to track the dynamic evolution of patient attitudes as technology advances and regulatory frameworks mature.

XSL•FO
RenderX

Acknowledgments

Generative artificial intelligence was not used in any portion of the manuscript writing.

Funding

This study was supported by the Hunan Provincial Department of Education's Postgraduate Education Innovation and Professional Ability Improvement Project (grant Xiangjiaotong [2020] 172-26). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

Data sharing is not applicable to this study as no datasets were generated or analyzed during this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search stage.
[PDF File, 1500 KB - jmir_v28i1e85663_app1.pdf]

Checklist 1
PRISMA 2020 checklist.
[PDF File, 163 KB - jmir_v28i1e85663_app2.pdf]

Checklist 2
ENTREQ checklist.
[PDF File, 95 KB - jmir_v28i1e85663_app3.pdf]

Checklist 3
PRISMA-S checklist.
[PDF File, 73 KB - jmir_v28i1e85663_app4.pdf]

References

1. Zeb S, Fnu N, Abbasi N, Fahad M. AI in healthcare: revolutionizing diagnosis and therapy. ijmdsa 2024;3(3):118-128. [doi: 10.47709/ijmdsa.v3i3.4546]
2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
3. Joshi G, Jain A, Araveeti SR, Adhikari S, Garg H, Bhandari M. FDA-approved artificial intelligence and machine learning (AI/ML)-enabled medical devices: an updated landscape. Electronics (Basel) 2024;13(3):498. [doi: 10.3390/electronics13030498]
4. Maslej N, Fattorini L, Perrault R, Gil Y, Parli V, Kariuki N, et al. Artificial Intelligence Index Report 2025. arXiv. 2025 Apr 8. [doi: 10.48550/arXiv.2504.07139]
5. Juang WC, Hsu MH, Cai ZX, Chen CM. Developing an AI-assisted clinical decision support system to enhance in-patient holistic health care. PLOS ONE 2022;17(10):e0276501. [doi: 10.1371/journal.pone.0276501] [Medline: 36315554]
6. Spencer KR, Ulrickson M, Grierson P, et al. BASECAMP-1: an approach to improve patient screening efficiency and to provide large datasets for precision medicine clinical trials. JCO 2025 Jun;43(16_suppl). [doi: 10.1200/JCO.2025.43.16_suppl.e23020]
7. Prakash S, Balaji JN, Joshi A, Surapaneni KM. Ethical conundrums in the application of artificial intelligence (AI) in healthcare-a scoping review of reviews. J Pers Med 2022 Nov 16;12(11):1914. [doi: 10.3390/jpm12111914] [Medline: 36422090]
8. Albashrawi MA. AI adoption in healthcare: trust and privacy concerns. Presented at: Proceedings of the International Conferences on Applied Computing 2022; Nov 8-10, 2022 URL: https://www.computing-conf.org/wp-content/uploads/2022/11/3_ICWI2022_R_017.pdf [accessed 2026-04-13]
9. Gawankar S, Nair S, Pawar V, Vhatkar A, Chavan P. Patient privacy and data security in the era of AI-driven healthcare. Presented at: 2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA); Aug 23, 2024; Pune, India p. 1-6. [doi: 10.1109/ICCUBEA61740.2024.10775004]

XSL•FO
RenderX

10. Williamson SM, Prybutok VR. Balancing privacy and progress: a review of privacy challenges, systemic oversight, and patient perceptions in AI-driven healthcare. Appl Sci (Basel) 2024;14(2):675. [doi: 10.3390/app14020675]
11. Nong P, Platt J. Patients' trust in health systems to use artificial intelligence. JAMA Netw Open 2025 Feb 3;8(2):e2460628. [doi: 10.1001/jamanetworkopen.2024.60628] [Medline: 39951270]
12. Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J Med Ethics 2021 Mar 18;47(5):329-335. [doi: 10.1136/medethics-2020-106820] [Medline: 33737318]
13. Byrnes J, Robinson M. Transparency and authority concerns with using AI to make ethical recommendations in clinical settings. Nurs Ethics 2025 Sep;32(6):1749-1760. [doi: 10.1177/09697330241307317] [Medline: 39710586]
14. Kim D, Vegt N, Visch V, Bos-De Vos M. How much decision power should (A)I have?: investigating patients' preferences towards AI autonomy in healthcare decision making. Presented at: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems; May 11, 2024. [doi: 10.1145/3613904.3642883]
15. Montemayor C, Halpern J, Fairweather A. In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. AI Soc 2022;37(4):1353-1359. [doi: 10.1007/s00146-021-01230-z] [Medline: 34054228]
16. Witkowski K, Dougherty RB, Neely SR. Public perceptions of artificial intelligence in healthcare: ethical concerns and opportunities for patient-centered care. BMC Med Ethics 2024 Jun 22;25(1):74. [doi: 10.1186/s12910-024-01066-4] [Medline: 38909180]
17. Gao S, He L, Chen Y, Li D, Lai K. Public perception of artificial intelligence in medical care: content analysis of social media. J Med Internet Res 2020 Jul 13;22(7):e16649. [doi: 10.2196/16649] [Medline: 32673231]
18. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
19. Tong A, Flemming K, McInnes E, Oliver S, Craig J. Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ. BMC Med Res Methodol 2012 Nov 27;12(1):181. [doi: 10.1186/1471-2288-12-181] [Medline: 23185978]
20. Noyes J, Booth A, Cargo M, Flemming K, Harden A, Harris J, et al. Qualitative evidence. In: Cochrane Handbook for Systematic Reviews of Interventions: John Wiley & Sons; 2019:525-545. [doi: 10.1002/9781119536604]
21. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. Introduction to Meta-Analysis: John Wiley &Amp; Sons; 2021.
22. Lockwood C, Munn Z, Porritt K. Qualitative research synthesis: methodological guidance for systematic reviewers utilizing meta-aggregation. Int J Evid Based Healthc 2015 Sep;13(3):179-187. [doi: 10.1097/XEB.0000000000000062] [Medline: 26262565]
23. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
24. Hannes K, Lockwood CS, Pearson AS. A comparative analysis of three online appraisal instruments' ability to assess validity in qualitative research. Qual Health Res 2010 Dec;20(12):1736-1743. [doi: 10.1177/1049732310378656] [Medline: 20671302]
25. Hurley ME, Sonig A, Herrington J, et al. Ethical considerations for integrating multimodal computer perception and neurotechnology. Front Hum Neurosci 2024;18:1332451. [doi: 10.3389/fnhum.2024.1332451] [Medline: 38435745]
26. Annamalai A. Challenges of building, deploying, and using AI-enabled telepsychiatry platforms for clinical practice among urban Indians: a qualitative study. Indian J Psychol Med 2021 Jul;43(4):336-342. [doi: 10.1177/0253717620973414] [Medline: 34385728]
27. Čartolovni A, Malešević A, Poslon L. Critical analysis of the AI impact on the patient-physician relationship: a multi-stakeholder qualitative study. Digit Health 2023;9:20552076231220833. [doi: 10.1177/20552076231220833] [Medline: 38130798]
28. Sujan MA, White S, Habli I, Reynolds N. Stakeholder perceptions of the safety and assurance of artificial intelligence in healthcare. SSRN Journal 2022. [doi: 10.2139/ssrn.4000675]
29. Kostick-Quenet KM, Lang BH, Smith JN, Hurley ME, Blumenthal-Barby J. Trust criteria for artificial intelligence in health: normative and epistemic considerations. J Med Ethics 2024 Aug;50(8):544-551. [doi: 10.1136/jme-2023-109338]
30. Hesjedal MB, Lysø EH, Solbjør M, Skolbekken JA. Valuing good health care: how medical doctors, scientists and patients relate ethical challenges with artificial intelligence decision-making support tools in prostate cancer diagnostics to good health care. Sociol Health Illn 2024 Nov;46(8):1808-1827. [doi: 10.1111/1467-9566.13818] [Medline: 39037701]
31. McCradden MD, Baba A, Saha A, et al. Ethical concerns around use of artificial intelligence in health care research from the perspective of patients with meningioma, caregivers and health care providers: a qualitative study. CMAJ Open 2020;8(1):E90-E95. [doi: 10.9778/cmajo.20190151] [Medline: 32071143]
32. Al-Anezi FM. Exploring the use of ChatGPT as a virtual health coach for chronic disease management. Learn Health Syst 2024 Jul;8(3):e10406. [doi: 10.1002/lrh2.10406] [Medline: 39036525]
33. Freeman S, Stewart J, Kaard R, et al. Health consumers' ethical concerns towards artificial intelligence in Australian emergency departments. Emerg Medicine Australasia 2024 Oct;36(5):768-776. [doi: 10.1111/1742-6723.14449]
34. Viberg Johansson J, Dembrower K, Strand F, Grauman Å. Women's perceptions and attitudes towards the use of AI in mammography in Sweden: a qualitative interview study. BMJ Open 2024 Feb 14;14(2):e084014. [doi: 10.1136/bmjopen-2024-084014] [Medline: 38355190]

XSL•FO
RenderX

35. Jeyakumar T, Younus S, Zhang M, et al. Preparing for an artificial intelligence-enabled future: patient perspectives on engagement and health care professional training for adopting artificial intelligence technologies in health care settings. JMIR AI 2023 Mar 2;2:e40973. [doi: 10.2196/40973] [Medline: 38875561]
36. Haan M, Ongena YP, Hommes S, Kwee TC, Yakar D. A qualitative study to understand patient perspective on the use of artificial intelligence in radiology. J Am Coll Radiol 2019 Oct;16(10):1416-1419. [doi: 10.1016/j.jacr.2018.12.043] [Medline: 30878311]
37. Berger SA, Håland E, Solbjør M. Patient perspectives on trust in artificial intelligence-powered tools in prostate cancer diagnostics. Qual Health Res 2026 Mar;36(2-3):276-288. [doi: 10.1177/10497323251387545] [Medline: 41251145]
38. Khairat S, Mehraby H, Masoumi S, et al. Evaluation of cancer survivors' experience of using AI-based conversational tools: qualitative study. JMIR Cancer 2025 Nov 14;11:e77390. [doi: 10.2196/77390] [Medline: 41237331]
39. Litchfield IJ, Delanerolle G, Harper L, Dunning S. Understanding acceptability of AI triage tools amongst underserved populations: lessons from the early phases of co-production wIth Bangladeshi communities in Birmingham, UK. Health Expect 2025 Dec;28(6):e70523. [doi: 10.1111/hex.70523] [Medline: 41342429]
40. Omori M, Basnayake P, Keogh L, Frazer HML, Kunicki KM, Lippey JF. Developing a typology of women's attitudes towards AI use in the BreastScreen Programme-a qualitative study with BreastScreen Victoria clients. Health Expect 2025 Oct;28(5):e70415. [doi: 10.1111/hex.70415] [Medline: 40884416]
41. Funer F, Schneider D, Heyen NB, et al. Impacts of clinical decision support systems on the relationship, communication, and shared decision-making between health care professionals and patients: multistakeholder interview study. J Med Internet Res 2023. [doi: 10.2196/55717] [Medline: 39178023]
42. Giebel GD, Raszke P, Nowak H, et al. Improving AI-based clinical decision support systems and their integration into care from the perspective of experts: interview study among different stakeholders. JMIR Med Inform. 2025 Jul 7 p. e69688. [doi: 10.2196/69688] [Medline: 40623684]
43. Ly A, Herse S, Williams MA, Stapleton F. Artificial intelligence for age‐related macular degeneration diagnosis in Australia: a novel qualitative interview study. Ophthalmic Physiologic Optic 2025 Sep;45(6):1282-1292. [doi: 10.1111/opo.13542]
44. Richardson JP, Smith C, Curtis S, et al. Patient apprehensions about the use of artificial intelligence in healthcare. NPJ Digit Med 2021 Sep 21;4(1):140. [doi: 10.1038/s41746-021-00509-1] [Medline: 34548621]
45. Foresman G, Biro J, Tran A, et al. Patient perspectives on artificial intelligence in health care: focus group study for diagnostic communication and tool implementation. J Particip Med 2024;17:e69564-e69564. [doi: 10.2196/69564]
46. Schneider D, Liedtke W, Klausen AD, et al. Indecision on the use of artificial intelligence in healthcare-a qualitative study of patient perspectives on trust, responsibility and self-determination using AI-CDSS. Digit Health 2025;11:20552076251339522. [doi: 10.1177/20552076251339522] [Medline: 40469779]
47. Zhang Z, Citardi D, Wang D, Genc Y, Shan J, Fan X. Patients' perceptions of using artificial intelligence (AI)-based technology to comprehend radiology imaging data. Health Informatics J 2021;27(2):14604582211011215. [doi: 10.1177/14604582211011215] [Medline: 33913359]
48. Gundlack J, Negash S, Thiel C, et al. Artificial intelligence in medical care - patients' perceptions on caregiving relationships and ethics: a qualitative study. Health Expect 2025 Apr;28(2):e70216. [doi: 10.1111/hex.70216] [Medline: 40094179]
49. Steerling E, Svedberg P, Nilsen P, Siira E, Nygren J. Influences on trust in the use of AI-based triage-an interview study with primary healthcare professionals and patients in Sweden. Front Digit Health 2025;7:1565080. [doi: 10.3389/fdgth.2025.1565080] [Medline: 40463579]
50. Lewin S, Bohren M, Rashidian A, et al. Applying GRADE-CERQual to qualitative evidence synthesis findings-paper 2: how to make an overall CERQual assessment of confidence and create a Summary of Qualitative Findings table. Implement Sci 2018 Jan 25;13(Suppl 1):10. [doi: 10.1186/s13012-017-0689-2] [Medline: 29384082]
51. Bronfenbrenner U. Toward an experimental ecology of human development. American Psychologist 1977;32(7):513-531. [doi: 10.1037/0003-066X.32.7.513]
52. McLeroy KR, Bibeau D, Steckler A, Glanz K. An ecological perspective on health promotion programs. Health Educ Q 1988 Dec;15(4):351-377. [doi: 10.1177/109019818801500401]
53. Thaivalappil A, Young I, Papadopoulos A. More than a framework? Applying socioecological models within mental health and community psychology inquiries. Canadian Journal of Community Mental Health 2024 Sep 1;43(3):37-45. [doi: 10.7870/cjcmh-2024-020]
54. Bronfenbrenner U, editor. The Ecology of Human Development: Experiments by Nature and Design: Harvard University Press; 1979. [doi: 10.4159/9780674028845]
55. Roura M. The social ecology of power in participatory health research. Qual Health Res 2021 Mar;31(4):778-788. [doi: 10.1177/1049732320979187] [Medline: 33349170]
56. Wu J, Xu L, Yu F, Peng K. Acceptance of medical treatment regimens provided by AI vs. human. Appl Sci (Basel) 2021;12(1):110. [doi: 10.3390/app12010110] [Medline: 34386269]
57. Hennrich J, Ritz E, Hofmann P, Urbach N. Capturing artificial intelligence applications' value proposition in healthcare - a qualitative research study. BMC Health Serv Res 2024 Apr 3;24(1):420. [doi: 10.1186/s12913-024-10894-4] [Medline: 38570809]

XSL•FO
RenderX

58. Babel A, Taneja R, Mondello Malvestiti F, Monaco A, Donde S. Artificial intelligence solutions to increase medication adherence in patients with non-communicable diseases. Front Digit Health 2021;3:669869. [doi: 10.3389/fdgth.2021.669869] [Medline: 34713142]
59. Lalova-Spinks T, Saesen R, Silva M, et al. Patients' knowledge, preferences, and perspectives about data protection and data control: an exploratory survey. Front Pharmacol 2023;14:1280173. [doi: 10.3389/fphar.2023.1280173] [Medline: 38445168]
60. Narkhede MR, Wankhede NI, Kamble AM. Enhancing patient autonomy in data ownership: privacy models and consent frameworks for healthcare. JDH 2025:1-23. [doi: 10.55976/jdh.4202513361-23]
61. Hahn E. Das "Recht auf Nichtwissen" des Patienten bei algorithmengesteuerter Auswertung von Big Data. MedR 2019 Mar;37(3):197-202. [doi: 10.1007/s00350-019-5176-0]
62. Kalkman S, van Delden J, Banerjee A, Tyl B, Mostert M, van Thiel G. Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence. J Med Ethics 2022 Jan;48(1):3-13. [doi: 10.1136/medethics-2019-105651] [Medline: 31719155]
63. Zolnoori M, Balls-Berry JE, Brockman TA, Patten CA, Huang M, Yao L. A systematic framework for analyzing patient-generated narrative data: protocol for a content analysis. JMIR Res Protoc 2019 Aug 26;8(8):13914. [doi: 10.2196/13914] [Medline: 31452524]
64. Norris S, Salgado F, Murray S, Amen D, Keator DB. The role of negativity bias in emotional and cognitive dysregulation: a neuroimaging study in anxiety disorders. Depress Anxiety 2025 Jan;2025(1). [doi: 10.1155/da/2739947]
65. Clements W, Thong LP, Zia A, Moriarty HK, Goh GS. A prospective study assessing patient perception of the use of artificial intelligence in radiology. APJHM 2022. [doi: 10.24083/apjhm.v17i1.861]
66. Dhagarra D, Goswami M, Kumar G. Impact of trust and privacy concerns on technology acceptance in healthcare: an Indian perspective. Int J Med Inform 2020 Sep;141(104164):104164. [doi: 10.1016/j.ijmedinf.2020.104164] [Medline: 32593847]
67. Arshad M, Sriram S, Khan S, Gollapalli PK, Albadrani MS. Mediating role of physician's empathy between physician's communication and patient's satisfaction. J Family Med Prim Care 2024 Apr;13(4):1530-1534. [doi: 10.4103/jfmpc.jfmpc_1615_23] [Medline: 38827677]
68. Karakaya Ç, Sümer FE, Demir ZG. In: Durmaz E, editor. Is Communication A Receipt For Doctors? The Relationship Between Communication Competence And Trust: A Research On Breast Patients: Marmara University; 2021. [doi: 10.17829/turcom.861934]
69. Li L, Zhai F, Liu H, Wan C, Tan J, Zeng L. Construction of trust relationship between doctors and patients: a social psychological analysis. Environ Soc Psychol 2022;6(2). [doi: 10.18063/esp.v6.i2.1395]
70. Esmaeilzadeh P, Mirzaei T, Dharanikota S. Patients' perceptions toward human-artificial intelligence interaction in health care: experimental study. J Med Internet Res 2021 Nov 25;23(11):e25856. [doi: 10.2196/25856] [Medline: 34842535]
71. Ventres WB. Looking within: intentions of practice for person-centered care. Ann Fam Med 2017 Mar;15(2):171-174. [doi: 10.1370/afm.2037] [Medline: 28289119]
72. Fuehrer S, Weil AB, Osterberg LG, Zulman DM, Meunier MR, Schwartz R. Building authentic connection in the patient-physician relationship. J Prim Care Community Health 2024;15:21501319231225996. [doi: 10.1177/21501319231225996] [Medline: 38281122]
73. Kishton R, Patel H, Saini D, Millstein J, Levy A. Listening as medicine: a thematic analysis. Patient Exp J 2023;10(1):64-71. [doi: 10.35680/2372-0247.1728]
74. Sakumoto, MD M, Joshi, MD A. Digital Empathy 2.0: connecting with patients using the written word. THMT 2023;8(5). [doi: 10.30953/thmt.v8.433]
75. Perivolaris A, Adams-McGavin C, Madan Y, et al. Quality of interaction between clinicians and artificial intelligence systems. A systematic review. Future Healthc J 2024 Sep;11(3):100172. [doi: 10.1016/j.fhj.2024.100172] [Medline: 39281326]
76. Cestonaro C, Delicati A, Marcante B, Caenazzo L, Tozzo P. Defining medical liability when artificial intelligence is applied on diagnostic algorithms: a systematic review. Front Med (Lausanne) 2023;10:1305756. [doi: 10.3389/fmed.2023.1305756] [Medline: 38089864]
77. Zhang J, Zhang ZM. Ethics and governance of trustworthy medical artificial intelligence. BMC Med Inform Decis Mak 2023 Jan 13;23(1):7. [doi: 10.1186/s12911-023-02103-9] [Medline: 36639799]
78. Platt JE, Nong P. An ecosystem approach to earning and sustaining trust in health care-too big to care. JAMA Health Forum 2023 Jan 6;4(1):e224882. [doi: 10.1001/jamahealthforum.2022.4882] [Medline: 36637813]
79. Lang BH, Nyholm S, Blumenthal-Barby J. Responsibility gaps and black box healthcare AI: shared responsibilization as a solution. Digit Soc 2023 Dec;2(3):52. [doi: 10.1007/s44206-023-00073-z] [Medline: 38596344]
80. Wang B, Asan O, Mansouri M. Perspectives of patients with chronic diseases on future acceptance of AI-based home care systems: cross-sectional web-based survey study. JMIR Hum Factors 2023 Nov 6;10:e49788. [doi: 10.2196/49788] [Medline: 37930780]
81. Nirmalasary TN, Rejo A, Windusari Y. Analysis model of socio-cultural environment on criminality rate: a rural case study. J of Law and Sust Develop 2024;12(5):e3670. [doi: 10.55908/sdgs.v12i5.3670]

XSL•FO
RenderX

82. Kambala M, Elevance Health. AI-powered healthcare: transforming patient outcomes with machine learning. JMSCR 2024;12(8):34-47. [doi: 10.18535/jmscr/v12i08.07]
83. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. J Am Med Inform Assoc 2018 Aug 1;25(8):1080-1088. [doi: 10.1093/jamia/ocy052] [Medline: 29788380]
84. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? BMJ Glob Health 2018;3(4):e000798. [doi: 10.1136/bmjgh-2018-000798] [Medline: 30233828]
85. Bu Q, Wang Z, Qi L, He C, Liren S, Xin M. Supervision and application analysis of AI medical devices. Social Medicine and Health Management 2023;4(2). [doi: 10.23977/socmhm.2023.040209]
86. Li DM, Parikh S, Costa A. A critical look into artificial intelligence and healthcare disparities. Front Artif Intell 2025;8:1545869. [doi: 10.3389/frai.2025.1545869] [Medline: 40115119]
87. Weiss D, Rydland HT, Øversveen E, Jensen MR, Solhaug S, Krokstad S. Innovative technologies and social inequalities in health: a scoping review of the literature. PLoS ONE 2018;13(4):e0195447. [doi: 10.1371/journal.pone.0195447] [Medline: 29614114]
88. Zhou K, Gattinger G. The evolving regulatory paradigm of AI in MedTech: a review of perspectives and where we are today. Ther Innov Regul Sci 2024 May;58(3):456-464. [doi: 10.1007/s43441-024-00628-3] [Medline: 38528278]
89. Jobin A, Ienca M, Vayena E. Artificial intelligence: the global landscape of ethics guidelines. Nat Mach Intell 2019;1:389-399. [doi: 10.1038/s42256-019-0088-2]
90. d'Elia A, Gabbay M, Rodgers S, et al. Artificial intelligence and health inequities in primary care: a systematic scoping review and framework. Fam Med Com Health 2022 Nov;10(Suppl 1):e001670. [doi: 10.1136/fmch-2022-001670]
91. Zhou Q, Yang L, Tang Y, et al. The mediation of trust on artificial intelligence anxiety and continuous adoption of artificial intelligence technology among primacy nurses: a cross-sectional study. BMC Nurs 2025 Jul 1;24(1):724. [doi: 10.1186/s12912-025-03406-0] [Medline: 40597194]
92. Frenkenberg A, Hochman G. It's scary to use it, it's scary to refuse it: the psychological dimensions of AI Adoption—anxiety, motives, and dependency. Systems 2025;13(2):82. [doi: 10.3390/systems13020082]
93. Huo W, Yuan X, Li X, Luo W, Xie J, Shi B. Increasing acceptance of medical AI: the role of medical staff participation in AI development. Int J Med Inform 2023 Jul;175:105073. [doi: 10.1016/j.ijmedinf.2023.105073] [Medline: 37119693]
94. Rony MKK, Numan SM, Akter K, et al. Nurses' perspectives on privacy and ethical concerns regarding artificial intelligence adoption in healthcare. Heliyon 2024 Sep 15;10(17):e36702. [doi: 10.1016/j.heliyon.2024.e36702] [Medline: 39281626]
95. Flugelman MY, Jaffe R, Luria G, Yagil D. Trust in the referring physician reduces anxiety in an integrated community-to-hospital care system. Isr J Health Policy Res 2020 May 11;9(1):7. [doi: 10.1186/s13584-020-00365-6] [Medline: 32393391]
96. Coffey M, Hannigan B, Meudell A, Jones M, Hunt J, Fitzsimmons D. Quality of life, recovery and decision-making: a mixed methods study of mental health recovery in social care. Soc Psychiatry Psychiatr Epidemiol 2019 Jun;54(6):715-723. [doi: 10.1007/s00127-018-1635-6] [Medline: 30470882]
97. Li Y, Wang X. Seeking health information on social media: a perspective of trust, self-determination, and social support. J Organ End User Comput 2018;30:1-22. [doi: 10.4018/JOEUC.2018010101]
98. Zeb R, Arzeen S, Mumtaz M, Haroon M. The perils of technology: social networking addiction and anxiety. PJMHS 2023;17(5):507-512. [doi: 10.53350/pjmhs2023175507]
99. Yokoi R, Eguchi Y, Fujita T, Nakayachi K. Artificial intelligence is trusted less than a doctor in medical treatment decisions: influence of perceived care and value similarity. International Journal of Human–Computer Interaction 2021 Jun 15;37(10):981-990. [doi: 10.1080/10447318.2020.1861763]
100. Wu C, Xu H, Bai D, Chen X, Gao J, Jiang X. Public perceptions on the application of artificial intelligence in healthcare: a qualitative meta-synthesis. BMJ Open 2023 Jan 4;13(1):e066322. [doi: 10.1136/bmjopen-2022-066322] [Medline: 36599634]
101. Buolamwini J, Gebru T, editors. Gender shades: intersectional accuracy disparities in commercial gender classification. : Proceedings of Machine Learning Research; 2018 Presented at: Proceedings of the 1st Conference on Fairness, Accountability and Transparency; New York, NY p. 77-91 URL: https://proceedings.mlr.press/v81/buolamwini18a.html [accessed 2026-04-13]
102. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
103. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
104. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform 2021 Jan;113:103655. [doi: 10.1016/j.jbi.2020.103655] [Medline: 33309898]
105. Smith H, Fotheringham K. Artificial intelligence in clinical decision-making: rethinking liability. Med Law Int 2020 Jun;20(2):131-154. [doi: 10.1177/0968533220945766]
106. Guidance W. Ethics and Governance of Artificial Intelligence for Health: World Health Organization; 2021:1-165.

XSL•FO
RenderX

107. Hulsen T. Explainable artificial intelligence (XAI): concepts and challenges in healthcare. AI 2023;4(3):652-666. [doi: 10.3390/ai4030034]
108. Bovenberg J, Peloquin D, Bierer B, Barnes M, Knoppers BM. How to fix the GDPR's frustration of global biomedical research. Science 2020 Oct 2;370(6512):40-42. [doi: 10.1126/science.abd2499] [Medline: 33004505]
109. Kiseleva A, Kotzinos D, De Hert P. Transparency of AI in healthcare as a multilayered system of accountabilities: between legal requirements and technical limitations. Front Artif Intell 2022;5:879603. [doi: 10.3389/frai.2022.879603] [Medline: 35707765]
110. Hanauer DA, Preib R, Zheng K, Choi SW. Patient-initiated electronic health record amendment requests. J Am Med Inform Assoc 2014;21(6):992-1000. [doi: 10.1136/amiajnl-2013-002574] [Medline: 24863430]
111. Park HJ. Patient perspectives on informed consent for medical AI: a web-based experiment. Digit HEALTH 2024;10:20552076241247938. [doi: 10.1177/20552076241247938] [Medline: 38698829]
112. Bakken S. AI in health: keeping the human in the loop. J Am Med Inform Assoc 2023 Jun 20;30(7):1225-1226. [doi: 10.1093/jamia/ocad091] [Medline: 37337923]
113. Nouis SCE, Uren V, Jariwala S. Evaluating accountability, transparency, and bias in AI-assisted healthcare decision-making: a qualitative study of healthcare professionals' perspectives in the UK. BMC Med Ethics 2025 Jul 8;26(1):89. [doi: 10.1186/s12910-025-01243-z] [Medline: 40629303]
114. Maliha G, Gerke S, Cohen IG, Parikh RB. Artificial intelligence and liability in medicine: balancing safety and innovation. Milbank Q 2021 Sep;99(3):629-647. [doi: 10.1111/1468-0009.12504] [Medline: 33822422]
115. Fotheringham K, Smith H. Accidental injustice: healthcare AI legal responsibility must be prospectively planned prior to its adoption. Future Healthc J 2024 Sep;11(3):100181. [doi: 10.1016/j.fhj.2024.100181] [Medline: 39371533]
116. Iserson KV. Informed consent for artificial intelligence in emergency medicine: a practical guide. Am J Emerg Med 2024 Feb;76:225-230. [doi: 10.1016/j.ajem.2023.11.022] [Medline: 38128163]
117. Osonuga A, Osonuga AA, Fidelis SC, Osonuga GC, Juckes J, Olawade DB. Bridging the digital divide: artificial intelligence as a catalyst for health equity in primary care settings. Int J Med Inform 2025 Dec;204:106051. [doi: 10.1016/j.ijmedinf.2025.106051] [Medline: 40706198]
118. Yang R, Gao S, Jiang Y. Digital divide as a determinant of health in the U.S. older adults: prevalence, trends, and risk factors. BMC Geriatr 2024;24(1):1027. [doi: 10.1186/s12877-024-05612-y]
119. Pham T. Ethical and legal considerations in healthcare AI: innovation and policy for safe and fair use. R Soc Open Sci 2025 May;12(5). [doi: 10.1098/rsos.241873]
120. Cestonaro C, Delicati A, Marcante B, Caenazzo L, Tozzo P. Defining medical liability when artificial intelligence is applied on diagnostic algorithms: a systematic review. Front Med 2023 Nov 27;10:2023. [doi: 10.3389/fmed.2023.1305756]

Abbreviations

AI: artificial intelligence
ENTREQ: Enhancing Transparency in Reporting the Synthesis of Qualitative Research
GRADE-CERQual: Confidence in the Evidence from Reviews of Qualitative Research
JBI-QARI: Joanna Briggs Institute Qualitative Assessment and Review Instrument
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
SET: social ecological theory
WHO: World Health Organization

Edited by S Brini; submitted 11.Oct.2025; peer-reviewed by N Heyen, S Jackson; revised version received 12.Mar.2026; accepted 12.Mar.2026; published 28.Apr.2026.

Please cite as:
Hou J, Zhang Z, Cheng X, Wang W
Patient Concerns Regarding Artificial Intelligence Applications in Health Care: Systematic Review and Meta-Synthesis Based on Social Ecological Theory
J Med Internet Res 2026;28:e85663
URL: https://www.jmir.org/2026/1/e85663
doi:10.2196/85663

© Jiayu Hou, Zhiqiao Zhang, Xuan Cheng, Weihong Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any

XSL•FO
RenderX

medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Online Community Support for Stroke Survivors and Caregivers: Scoping Review

Putu Wuri Handayani[1]; Kamila Alifia[1]; Juliana Sutanto[2]; Erlin Erlina[3]; Sharyn Graham Davies[4]; Arthur HP Mawuntu[5]; Muhammad Isman Jusuf[6]; Narelle Warren[7]

[1]Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

[2]Department of Human Centred Computing, Faculty of Information Technology, Monash University, Clayton, Victoria, Australia

[3]Department of Health Behaviour, Environment and Social Medicine, Faculty of Medicine, Public Health and Nursing, Universitas Gadjah Mada, Yogyakarta, Indonesia

[4]School of Languages, Literature, Cultures and Linguistics, Faculty of Arts, Monash Univertisty, Melbourne, Victoria, Australia

[5]Department of Neurology, Faculty of Medicine, Universitas Sam Ratulangi, Manado, Indonesia

[6]Department of Neurology, Faculty of Medicine, Universitas Negeri Gorontalo, Gorontalo, Indonesia

[7]School of Social Sciences, Faculty of Arts, Monash University, Wellington Rd, Melbourne, Victoria, Australia

Corresponding Author:
Juliana Sutanto
Department of Human Centred Computing, Faculty of Information Technology, Monash University, Clayton, Victoria, Australia

Abstract

Background: Previous studies found that online communities are critical in supporting stroke survivors and caregivers for stroke recovery. However, it is unclear how such online communities are designed, or could be designed.

Objective: This review aims to identify the key design elements of an online community to support stroke survivors and caregivers, that is, the actors, types of community support, and supporting technologies.

Methods: We used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) guidelines. We included journal articles related to online community support for stroke. Editorials, registers, opinion pieces, letters, and conference papers were excluded. Online databases PubMed/MEDLINE, Scopus, Web of Science, ScienceDirect, and ProQuest were searched for articles published from January 2015 to June 2025. Articles were screened based on the title, abstract, and full text using Covidence software. After screening and full-text review, we read the selected articles in detail to analyze and synthesize information on key actors, types of support, and technologies used to support stroke survivors and caregivers.

Results: A total of 77 articles were included. These articles discussed digital support technologies (52 articles), community functions and roles (18 articles), online stroke community systems (6 articles), and the stroke ecosystem (1 article). Our review found that the online community of support for stroke survivors and caregivers includes the caregivers themselves (46 articles), health workers (24 articles), and the local community/society (14 articles). Online communities mainly provide informational support, including giving advice (28 articles) and tangible aids (29 articles), followed by social support to create a sense of belonging (26 articles). Technologies discussed included mobile health (25 articles), web-based systems (12 articles), virtual/augmented reality (8 articles), sensors/wearable technology (8 articles), video-guided exercise apps (4 articles), and telehealth/telerehabilitation/teleconsultation/telestroke (4 articles). Only one examined how cultural differences influence technology.

Conclusions: Although technologies are essential in online communities of support for stroke survivors and caregivers, this review shows a lack of studies that analyze the use and role of technologies in such online communities. This could be because the key actors of the online communities are the caregivers, who mainly seek social support and therefore do not require sophisticated technology. Nevertheless, technologies such as telerehabilitation and video-guided exercise apps could be important for other actors, including the local community and health workers, to enable them to support stroke survivors and their caregivers.

(J Med Internet Res 2026;28:e71190) doi:10.2196/71190

KEYWORDS

scoping review; stroke; online community; online community assistance; digital technology

XSL•FO
RenderX

Introduction

Background

Globally, stroke is the second leading cause of death and third leading cause of disability [1]. Caused by impaired perfusion through the blood vessels to the brain, stroke requires high quality, fast, and precise management to prevent and avoid disability and death [1-3]. Furthermore, stroke causes chronic effects, including mobility problems, cognitive impairment, and depression, that can become a burden on health care systems [4]. Although the time to discharge varies considerably between and within countries, most stroke survivors are discharged home with differing degrees of residual physical impairments [5], leading to an increased reliance on their caregivers. Caregivers are often a spouse or partner, family member, friend, or significant other who provides physical, practical, transportation, and emotional help [6-8].

Regular rehabilitation therapy has been demonstrated to improve physical and cognitive capacity and quality of life for stroke survivors [9-11]. Post discharge from the acute care setting, stroke survivors should ideally undergo regular reviews of their health and social status by their primary care service provider, typically at 6 weeks, 6 months, and then annually [12]. However, the extent to which this can be realized varies for multiple reasons. Challenges encountered by stroke survivors in seeking to access appropriate health facilities for treatment or rehabilitation include: complex pathways to care, limited physical mobility, travel distances, limited transportation options, long waiting times, and unaffordable out-of-pocket expenses [10,13].

Technologies such as teleconsultation and telerehabilitation offer stroke survivors and their caregivers' ways to address these barriers [14], offering ways to receive treatment and education materials from doctors or therapists remotely. However, effective use of such technologies requires digital literacy and access: stroke survivors and caregivers who are not technology savvy may avoid using the technologies and instead seek help from others in their community [15,16], if at all. Digital literacy and access to technologies are especially important for stroke survivors living in rural or regional areas in low- and middle-income countries (LMICs) [17].

Community—including groups of families, individuals, professionals, organizations, or other types of networks and social circles—can play an essential role in stroke recovery and rehabilitation [10,18-22]. Wright et al [23] found that long-term stroke survivors' engagement with the community was key to addressing their multidimensional support needs (Figure 1). Moreover, Intamas et al [21] suggested that developing a stroke care model by involving the community could improve the quality of life of stroke survivors and caregivers.

Figure 1. Multidimensional support needs of community-based stroke survivors and their caregivers.

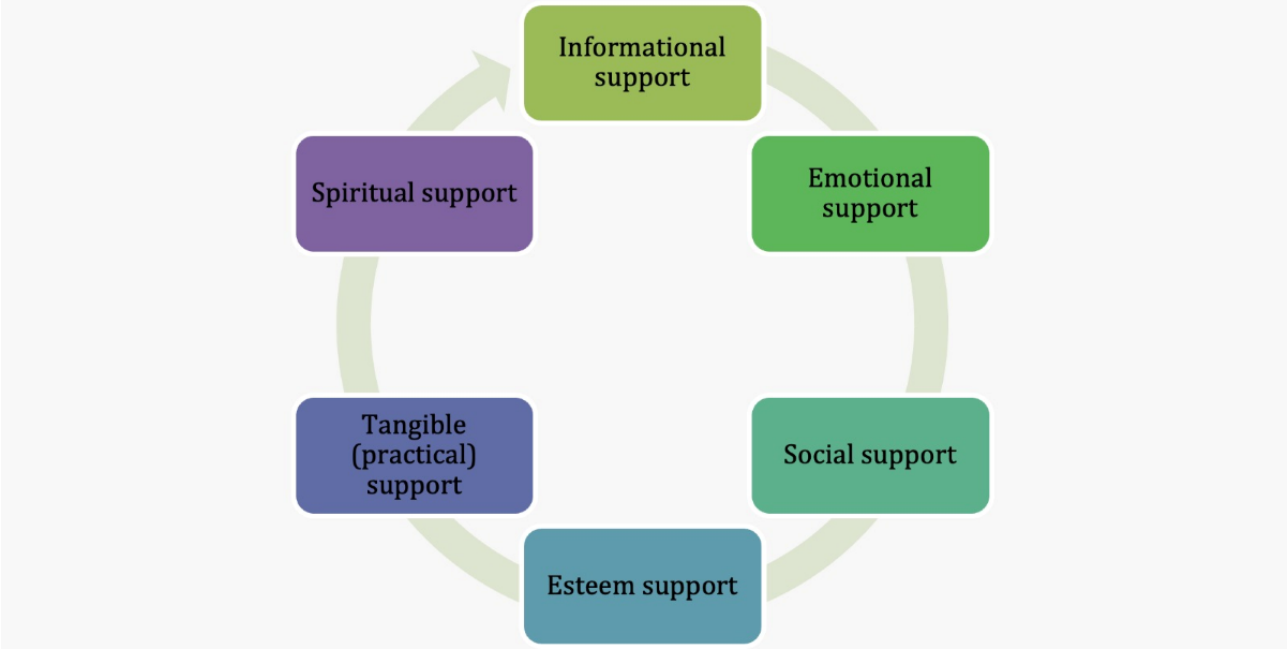

Informational support includes education provision (including on stroke symptoms and improved stroke identification) [24,25], sharing resources for raising awareness, and linking people to services and treatment options [26]. Emotional support addresses psychological concerns about the impact of stroke, recovery progress, and addressing worries about the future (including risk of future stroke) [27,28], including reassurance, encouragement, affirmations, and messages of care, concern, and empathy [29,30]. Social or network support seeks to address loneliness or social isolation that can follow a stroke by promoting a sense of belonging and connection. Social support may also include support to engage in social activities (eg, clubs or events), physical activity or exercise programs [29-31], while esteem support focuses on a person's intrinsic value and abilities to reinforce their personhood and social identity [32,33]. These aspects are complemented by sensitive tangible (practical) support, which focuses on assistance in activities of daily life [34]. Spiritual support may also play an important role at the community level; for example, Smith et al [8] found that an online community could give "prayer support." This prayer or

XSL•FO
RenderX

spiritual support appeals to people's religious or spiritual beliefs to create a sense of connection to cope with their illness [8,35-38].

Research in all contexts has highlighted the significance of community-based interventions in facilitating poststroke recovery and reintegration [16,39]. Health professionals rely on community-based initiatives to assist and monitor stroke survivors [8], share tasks associated with stroke care [17], and promote stroke prevention [40]. Community-based initiatives are especially important in LMIC, where health workforce and infrastructure issues limit the formal provision of stroke care, leaving the bulk of stroke survivorship care and management to those located at the household or community level [17,41,42].

Objectives

Despite the importance of community-based support for stroke survivors and their caregivers, there remain challenges to its delivery in many settings. Online support strategies offer much promise in addressing stroke survivors' and caregivers' needs [43,44], but how an online health community of support should be designed and what components it should include remains unclear. The aim of this scoping review is therefore to (1) identify the key actors for an online health community, (2) define the types of support to be incorporated in an online platform, and (3) determine the technological aspects of an online health community of support. These three key elements (who, what, and how) are the pillars in designing an online community of support for stroke survivors and caregivers.

Methods

Study Design

This scoping review used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) to guide its planning, designing, and reporting [45]. The PRISMA-ScR approach was adopted in order to provide an understanding of the relevant core concepts and key items, as well as to guide the synthesis of the search results, and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [46] guidelines were followed throughout (as detailed in Checklist 1).

Eligibility Criteria

We used the population, concept, and context framework to define our eligibility criteria, as shown in Textbox 1. Articles that did not focus on online community support for stroke were excluded from this study. Primary studies and review articles were included. Review articles were important in our review for an overview of the prevailing concept of online community for stroke support and the technology needed for an online health support community. We excluded editorials, registers, letters, and conference papers. Only English-language articles were included, as English is the dominant language of scientific communication, and in order to ensure feasibility, consistency, and accuracy in data extraction. We searched for articles published between January 2015 and June 2025; 2015 marks an increased use of technologies for managing stroke.

Textbox 1. Eligibility criteria.

Inclusion criteria

Population: Actors that engage in an online stroke community

Concept: An online community providing stroke support, with a focus on technology-enabled health solutions

Context: Health study or health technologies on stroke in the online community

Study Design: Primary study with any design and review study

Study Type: Published in a peer-reviewed journal from January 2015 to June 2025

Language: English

Access: Full-text available

Exclusion criteria

Population: Actors not engaged in an online stroke community

Concept: Not stroke-related studies and not focused on an online community of support for stroke survivors

Context: Not health study or health technologies on stroke in the online community

Study Design: Editorials, registers, letters, and conference papers

Study Type: Conference papers, dissertations, books

Language: Language other than English

Access: Full-text not available

Information Sources and Search Strategy

Online databases PubMed/MEDLINE, Scopus, Web of Science, ScienceDirect, and ProQuest were searched due to their comprehensive coverage of health research journals. Based on the study aims and eligibility criteria, the following search (keywords) terms were used for all database searches: online health community, online health support community, health support community, health ecosystem, technology, and stroke.

Multimedia Appendix 1 describes the step-by-step search strategy for each online database and outlines the slightly

XSL•FO
RenderX

different search mechanisms used. For example, ScienceDirect does not permit the use of wildcards (the * extension), and we used the "tiab" (title and abstract) topic search mechanism for PubMed and ScienceDirect, while ProQuest uses the "noft" (anywhere except full text) tag for topic searches. We used the "OR" Boolean operator to create search terms among the free-text keywords related to stroke, online community, and technology, as well as medical subject heading terms. Individual search terms (Multimedia Appendix 1) were then combined using the "AND" Boolean operator. We used medical subject heading terms related to the concept of stroke (ie, stroke, poststroke, cerebrovascular accident, and stroke survivor), online community (ie, online community, virtual community, support group, peer support, social support, online forum, self-help groups, social environment, caregiver), and technology (ie, eHealth, mobile health [mHealth], web-based, digital, virtual, online, software, internet-based intervention, virtual reality [VR], apps).

The difference between a community and an online community lies only in the use of online channels or digital technology used in carrying out activities. Thus, online community support includes all kinds of support via online channels or digital technology.

Search strings were applied one by one in the selected online databases. PWH and NW searched PubMed/MEDLINE and Web of Science as well as ProQuest, and KAI searched Scopus. All articles were uploaded into Covidence software by PWH, KAI, and NW. PWH, KAI, JS, NW, SD, AM, and MIJ screened articles by using the Covidence software, which was used for title and abstract screening and for full-text review. Covidence allowed identification of duplicate articles [47]. Title and abstract screening were done by PWH and KAI and verified by JS and NW. Screening for full-text articles was then undertaken by all authors independently, and we discussed the results for final consensus. Where any conflicts existed in decision-making, JS and NW made a final determination through discussion.

Data Extraction and Synthesis

A data chart was created using Microsoft Excel to identify which variables to extract. Two reviewers (PWH and KAI) independently and iteratively charted the data, wrote, and discussed the results. Additional sources, such as supplementary material, were analyzed to support the data charting.

Data were extracted based on the aims and objectives of each article, as well as data that fit with the scoping review objectives. Additional data on article demographics were extracted to capture: (1) the article's central topic, (2) country or countries where the study was conducted, (3) the methodology used, (4) the actors in the community (if relevant), (5) the type and nature of community support for stroke survivors and caregivers, and (6) the supporting technology developed or used. The full data chart is provided in Multimedia Appendix 2 [3,4,7-10,12,14,15,23,39,48-113], and the cited journals are summarized in Multimedia Appendix 3. Descriptive analyses were conducted to identify key research themes before grouping the articles based on these defined themes.

Results

Study Selection

Overview

In this review, 2496 articles were imported from the literature searches into Covidence for screening. We removed 818 duplicates; 1457 studies were deemed irrelevant, as they did not focus on stroke. The remaining 221 articles were assessed for full-text eligibility. A total of 144 articles were excluded because they did not solely focus on stroke, were protocol papers, or the topics were unrelated to our study objectives. Figure 2 shows the PRISMA flow diagram, where a total of 77 articles were selected.

XSL•FO
RenderX

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

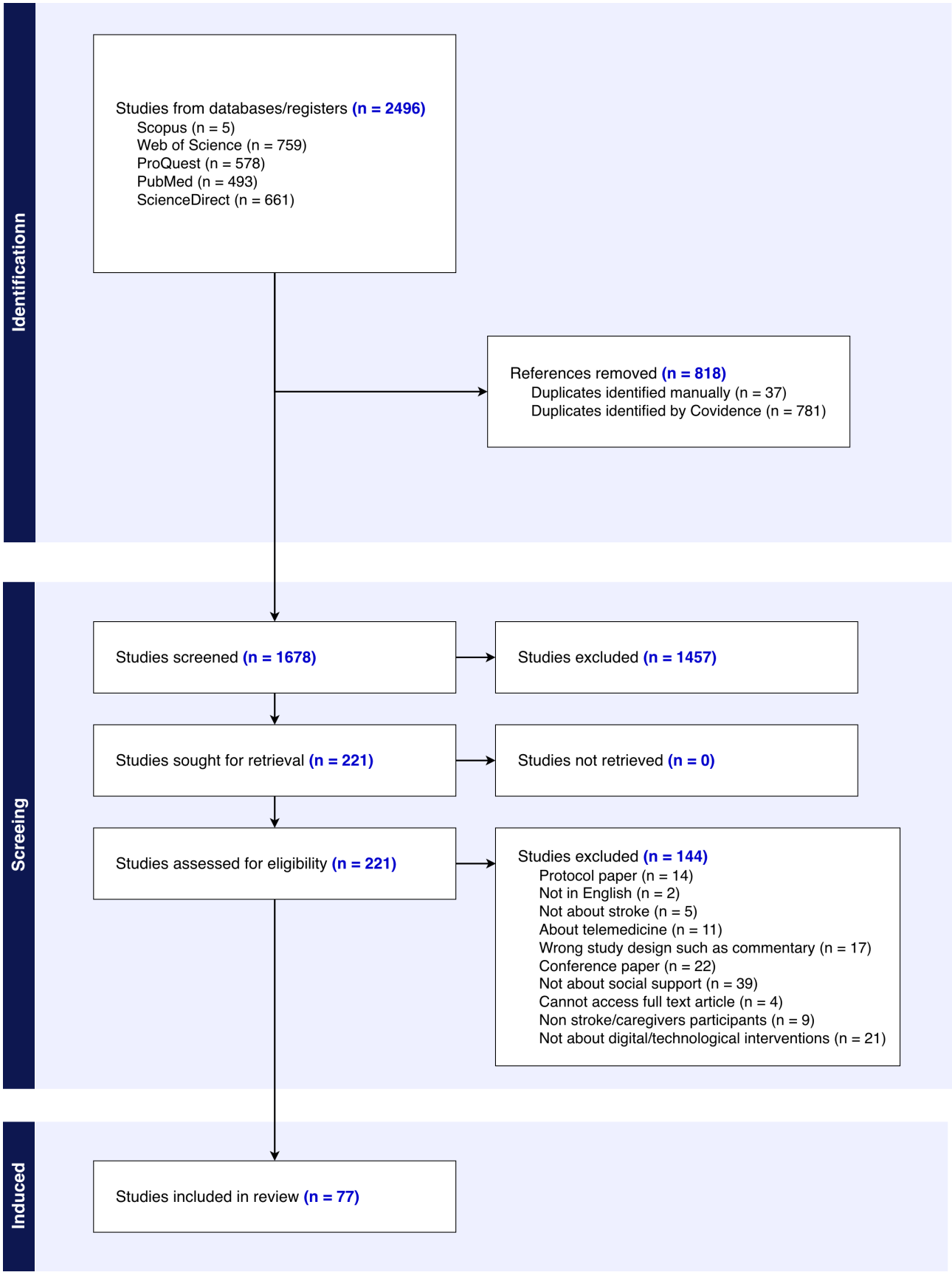

XSL•FO
RenderX

Article Demographic Summaries: Topic, Location, and Methodologies Used

Data on the main themes were extracted. The 77 articles addressed a range of topics, including: (1) digital technologies used to support people with stroke (52 articles), (2) the function and roles of community in stroke recovery (18 articles), (3) online stroke community systems (6 articles), and (4) ecosystems of stroke (1 article), that involve relevant actors, such as online communities, organizations, and implementation partners to manage, implement, and disseminate tangible actions to reduce stroke burden. Table 1 identifies the distribution of relevant articles (by reference list number) by topic. Each article has one topic group.

Table . Distribution of articles across topic areas.

Topic	Reference list number	Number of papers (%)
Digital support technologies	Thompson et al [4], Scrivener et al [48], Camicia et al [49], Demers et al [50], Freund et al [51], Johnson et al [52], Kerr et al [53], Krishnan et al [54], Lobo et al [55], Luo et al [56], Peters et al [57], Sarfo et al [58], Saywell et al [59], Silvera-Tawil et al [60], Sun et al [61], Xu et al [62], Lee et al [63], Mainali et al [64], Reeves et al [65], Requena et al [66], Cooray et al [67], Blanton et al [68], Caunca et al [69], Sureshkumar et al [70], Zhou et al [71], Vloothuis et al [72], English et al [73], Givon et al [74], Giachero et al [75], Lo et al [76], Andrades-González et al [77]; Demir and Gozum [78], De Simoni et al [79], Eriksson et al [80], Favilla et al [81], Firdaus et al [82], Firmawati et al [83], Gong et al [84], Kamwesiga et al [85], Lobo et al [86], Kechik et al [87], Pereira et al [88], Newland et al [89], Nichols et al [90], Siegel et al [91], Tsang et al [92], Lee et al [93], Olafsdottir et al [94], Olafsdottir et al [95], Juengst et al [96], Lam et al [97], Lo et al [98]	52 (67)
Community functions and roles	Leonardi and Fheodoroff [3], Lobo et al [7], Gooch et al [9], Smythe et al [10], Sauvé-Schenk et al [12], Sauerzopf et al [15], Wright et al [23], Magwood et al [39], Aravind et al [99], Cameron and Gignac [100], Fu et al [101], Pindus et al [102], White et al [103], Norlander et al [104], Deutschbein et al [105], Millar et al [106], Berkeley et al [107], Murakami et al [108]	18 (23)
Online stroke community system	Smith et al [8], Smith et al [109], Reszel et al [110], Cruickshank et al [111], Thomas et al [112], Paul et al [113]	6 (8)
Stroke ecosystem	Feigin et al [14]	1(1)

Studies were conducted in a range of settings, as shown in Table 2. Eight articles, mostly review articles, did not focus on a specific country; of the other articles, most drew on stroke research conducted in high-income (developed) countries.

XSL•FO
RenderX

Table . Location of studies reported in the reviewed articles.

Study location	Reference list number	Number of papers (%)
Australia / Oceania	Lobo et al [7][a], Scrivener et al [48], Johnson et al [52], Saywell et al [59], Silvera-Tawil et al [60], English et al [73], White et al [103]	7 (9)
North America	Reszel et al [110], Demers et al [50], Krishnan et al [54], Mainali et al [64], Reeves et al [65], Blanton et al [68], Caunca et al [69], Favilla et al [81], Siegel et al [91], Aravind et al [99], Cameron and Gignac [100], Berkeley et al [107], Lee et al [93]	13 (17)
East Asia	Lee et al [63], Zhou et al [71], Lo et al [76], Gong et al [84], Tsang et al [92], Fu et al [101], Murakami et al [108], Lam et al [97], Lo et al [98]	9 (12)
South-East Asia	Luo et al [56], Xu et al [62], Firdaus et al [82], Kechik et al [87]	4 (5)
West Asia	Givon et al [74], Demir and Gozum [78]	2 (3)
South Asia	Sureshkumar et al [70]	1 (1)
Northern Europe	Lobo et al [7][a], Cooray et al [67], Eriksson et al [80], Norlander et al [104], Olafsdottir et al [94], Olafsdottir et al [95]	6 (8)
Western Europe	Sauerzopf et al [15], Wright et al [23], Smith et al [109], Kerr et al [53], Vloothuis et al [72], De Simoni et al [79], Deutschbein et al [105], Millar et al [106], Thomas et al [112], Paul et al [113]	10 (13)
Southern Europe	Requena et al [66], Giachero et al [75], Andrades-González et al [77], Pereira et al [88]	4 (5)
East Africa	Kamwesiga et al [85]	1 (1)
West Africa	Sarfo et al [58], Nichols et al [90]	2 (3)
South Africa	Smythe et al [10]	1 (1)
Worldwide/ not specified	Leonardi and Fheodoroff [3], Thompson et al [4], Smith et al [8], Gooch et al [9], Sauvé-Schenk et al [12], Feigin et al [14], Magwood et al [39], Camicia et al [49], Freund et al [51], Lobo et al [55], Peters et al [57], Sun et al [61], Firmawati et al [83], Lobo et al [86], Newland et al [89], Pindus et al [102], Cruickshank et al [111], Juengst et al [96]	18 (23)

[a]This study was conducted in more than one country.

Table 3 shows the methodological approaches reported in each article. The 19 review articles aside, empirical articles varied in their approaches: descriptive qualitative methods (including a range of approaches such as content analysis, grounded theory, and thematic analysis); randomized control trials; interventions; feasibility/usability (quantitative) studies; one prospective cohort quantitative study; and mixed methods studies. We also found that 3 articles reported co-design or consensus studies.

XSL•FO
RenderX

Table . Methodological approaches.

Study methodology used	References	Number of papers (%)
Review article	Leonardi and Fheodoroff [3], Thompson et al [4], Lobo et al [7], Gooch et al [9], Sauvé-Schenk et al [12], Feigin et al [14], Magwood et al [39], Camicia et al [49], Freund et al [51], Lobo et al [55], Peters et al [57], Sun et al [61], Firmawati et al [83], Lobo et al [86], Kechik et al [87], Newland et al [89], Cameron and Gignac [100], Pindus et al [102], Juengst et al [96]	19 (25)
Randomized control trial	Johnson et al [52], Saywell et al [59], Lee et al [63], Reeves et al [65], Zhou et al [71], English et al [73], Givon et al [74], Giachero et al [75], Lo et al [76], Favilla et al [81], Aravind et al [99], Fu et al [101], Lam et al [97]	13 (17)
Cluster Randomized Trial	Berkeley et al [107][a]	1 (1)
Clinical trial	Scrivener et al [48]	1 (1)
Feasibility/usability (quantitative) study	Mainali et al [64], Caunca et al [69], Demir and Gozum [78]	3 (4)
Intervention study	Luo et al [56], Requena et al [66], Cooray et al [67], Vloothuis et al [72], Deutschbein et al [105], Olafsdottir et al [95], Paul et al [113]	7 (9)
Observational study (quantitative)	Demers et al [50], Sarfo et al [58], Firdaus et al [82], Siegel et al [91], Berkeley et al [107][a], Murakami et al [108]	6 (8)
Qualitative study	Smythe et al [10], Sauerzopf et al [15], Wright et al [23], Smith et al [109], Reszel et al [110], Krishnan et al [54], Xu et al [62], Andrades-González et al [77], De Simoni et al [79], Eriksson et al [80], Kamwesiga et al [85], Tsang et al [92], Norlander et al [104], Cruickshank et al [111], Thomas et al [112], Lo et al [98]	16 (20)
Mixed methods	Smith et al [8], Blanton et al [68], Sureshkumar et al [70], Gong et al [84], Nichols et al [90], Millar et al [106], Lee et al [93], Olafsdottir et al [94]	8 (10)
Co-design	Silvera-Tawil et al [60], Pereira et al [88]	2 (3)
Consensus method	Kerr et al [53]	1 (1)
Prospective Cohort (quantitative)	White et al [103]	1 (1)

[a]Article involves a secondary observational analysis of a cluster-randomized trial, so is listed twice.

Data Synthesis

The reviewed articles were concerned with a range of different study respondents (Table 4a). Almost all articles focused on stroke survivors. Most were concerned with informal caregivers, often family members or friends; a similar but slightly smaller number of articles were concerned with community or social support. Health workers were the focus of approximately one-third of the reviewed articles, while a much smaller number of articles concerned policymakers and funding bodies. Some articles addressed multiple study populations.

XSL•FO
RenderX

Table . Summary of study respondents in reviewed articles.

Respondents	References	Number of papers (%)
Stroke caregivers (ie, family and friends)	Thompson et al [4], Lobo et al [7][a], Smith et al [8], Gooch et al [9][a], Smythe et al [10][a], Sauvé-Schenk et al [12][a], Sauerzopf et al [15][a], Wright et al [23][a], Magwood et al [39], Smith et al [109][a], Camicia et al [49][a], Freund et al [51], Kerr et al [53][a], Lobo et al [55], Sun et al [61][a], Xu et al [62], Cooray et al [67], Blanton et al [68][a], Caunca et al [69], Sureshkumar et al [70], Zhou et al [71][a], Vloothuis et al [72][a], English et al [73], Andrades-González et al [77], Demir and Gozum [78], De Simoni et al [79], Eriksson et al [80], Favilla et al [81], Firdaus et al [82], Firmawati et al [83], Gong et al [84], Kamwesiga et al [85], Lobo et al [86], Kechik et al [87], Pereira et al [88][a], Nichols et al [90][a], Tsang et al [92], Aravind et al [99], Cameron and Gignac [100], Fu et al [101], Pindus et al [102], Norlander et al [104], Thomas et al [112], Olafsdottir et al [94], Olafsdottir et al [95][a], Lo et al [98]	46 (60)
Health workers (ie, therapist and nurse)	Leonardi and Fheodoroff [3], Thompson et al [4], Lobo et al [7][a], Gooch et al [9][a], Smythe et al [10][a], Feigin et al [14][a], Sauerzopf et al [15][a], Magwood et al [39], Reszel et al [110], Camicia et al [49][a], Kerr et al [53][a], Silvera-Tawil et al [60], Sun et al [61][a], Mainali et al [64], Requena et al [66], Blanton et al [68][a], Zhou et al [71][a], Vloothuis et al [72][a], Pereira et al [88][a], Newland et al [89], Nichols et al [90], Millar et al [106], Lee et al [93], Olafsdottir et al [95][a]	24 (31)
Community/society	Gooch et al [9], Smythe et al [10][a], Sauvé-Schenk et al [12][a], Sauerzopf et al [15][a], Wright et al [23][a], Smith et al [109][a], Kerr et al [53][a], Luo et al [56], Sarfo et al [58], Sun et al [61][a], Nichols et al [90][a], Murakami et al [108], Lee et al [93], Paul et al [113]	14 (18)
Funders	Feigin et al [14]	1 (1)
Policymakers	Feigin et al [14], Kerr et al [53][a]	2 (3)

[a]These studies involved more than one respondent categories.

The information presented in Table 4a highlights the different human actors involved in the provision of stroke-related health care and support. This stroke health ecosystem (Figure 3) consists of (informal) stroke caregivers, health workers, community members, or the broader society, funders, and policymakers.

XSL•FO
RenderX

Figure 3. Actors who comprise the stroke health ecosystem.

Stroke caregivers are often a spouse or partner, family member, or friend who provides required support on a regular basis [7,8]. They are often untrained—although they may draw upon training they have received for other roles—and they are most often unpaid, although they may receive a small stipend for their care provision. Active engagement of stroke caregivers could reduce health care costs, reduce burnout, improve care processes, and improve stroke survivors' outcomes [7]. Health workers offer separate and overlapping services designed to create a connection between stroke survivors and health care systems and facilitate the navigation of services to optimize stroke recovery, manage care transitions, limit social isolation, and help determine eligibility for services vital to stroke care [7]. They have received specific training relevant to their discipline (such as medicine, nursing, or allied health) or for the specific care tasks undertaken (such as completing a certificate or diploma in community care). They often receive a salary commensurate with their labor provisions. Community/society includes groups of individuals, organizations, or other types of networks and social circles [8,10]. Community/society may include neighbors or small business holders. Funders are agencies or philanthropists, including health insurance organizations, that provide funding for community-based stroke program design, development, and implementation [14]. Policymakers are health regulators who seek to translate the insights from funders and grassroots care providers into formalized regional, national, and international regulations [14].

Community Support

All but 2 articles discussed some aspects of community support. "Community support" is an umbrella term that captures a range of support activities and processes (Table 5).

XSL•FO
RenderX

Table . Types of community support included in the reviewed articles.

Types of community support	References	Number of papers (%)
Informational support	Smith et al [8], Feigin et al [14][a], Reszel et al [110][a], Camicia et al [49][a], Johnson et al [52][a], Lobo et al [55][a], Reeves et al [65][a], Cooray et al [67], Blanton et al [68][a], Caunca et al [69], Sureshkumar et al [70], Lo et al [76][a], Andrades-González et al [77][a], Demir and Gozum [78], De Simoni et al [79][a], Eriksson et al [80][a], Favilla et al [81], Firdaus et al [82], Firmawati et al [83][a], Lobo et al [86], Kechik et al [87], Nichols et al [90], Siegel et al [91], Tsang et al [92], Cruickshank et al [111][a], Thomas et al [112][a], Olafsdottir et al [94][a], Lo et al [98][a]	28 (36)
Emotional support	Smith et al [8], Smith et al [109], Blanton et al [68][a], Firmawati et al [83][a], Thomas et al [112][a], Olafsdottir et al [94][a], Olafsdottir et al [95][a], Juengst et al [96][a]	8 (10)
Social support	Leonardi and Fheodoroff [3], Smith et al [8], Feigin et al [14][a], Reszel et al [110][a], Camicia et al [49][a], Freund et al [51][a], Johnson et al [52][a], Peters et al [57][a], Sarfo et al [58][a], Saywell et al [59], Silvera-Tawil et al [60][a], Sun et al [61][a], Xu et al [62], Lee et al [63][a], Lo et al [76][a], Andrades-González et al [77][a], De Simoni et al [79][a], Eriksson et al [80][a], Kamwesiga et al [85], Pereira et al [88], Cruickshank et al [111][a], Thomas et al [112][a], Olafsdottir et al [94][a], Olafsdottir et al [95][a], Lo et al [98][a], Paul et al [113]	26 (33)
Esteem support	Scrivener et al [48], Olafsdottir et al [94][a]	2 (3)
Spiritual support	Smith et al [8]	1 (1)
Advice or knowledge	Smith et al [8], Feigin et al [14][a], Reszel et al [110][a], Camicia et al [49][a], Freund et al [51][a], Johnson et al [52][a], Lobo et al [55][a], Lee et al [63][a], Lo et al [76][a], Gong et al [84], Newland et al [89], Cruickshank et al [111][a], Olafsdottir et al [94][a], Olafsdottir et al [95][a]	14 (18)
Tangible aid	Smith et al [8], Feigin et al [14][a], Scrivener et al [48][a], Camicia et al [49][a], Johnson et al [52][a], Kerr et al [53], Krishnan et al [54], Lobo et al [55][a], Luo et al [56], Peters et al [57][a], Sarfo et al [58][a], Silvera-Tawil et al [60][a], Sun et al [61][a], Lee et al [63][a], Mainali et al [64], Reeves et al [65][a], Requena et al [66], Zhou et al [71], Vloothuis et al [72], English et al [73], Givon et al [74], Giachero et al [75], Lo et al [76][a], Deutschbein et al [105], Lee et al [93], Olafsdottir et al [94][a], Olafsdottir et al [95][a], Juengst et al [96][a], Lam et al [97]	29 (38)
Involvement in decision-making	Camicia et al [49][a], Blanton et al [68][a]	2 (3)
Not mention the type of community support	Thompson et al [4], Demers et al [50]	2 (3)

XSL•FO
RenderX

[a]These studies mentioned more than one type of community support.

Our review highlights that there are sometimes different meanings given to each type of support; thus, we defined each type of support by searching for similar meanings across studies. The reviewed articles indicate that the supports most needed by stroke survivors and caregivers are informational support, network/social support, and tangible aid (Multimedia Appendix 4 [3,4,8,14,48-61,63-98,105,109-113]).

Informational support has an educative element, consisting of ideas or suggestions for action, or providing sources of information or help. This information can help stroke survivors and their caregivers to reassess or redefine their situation, while also offering detailed information, facts, or news [23]. Emotional support is a nurturing support that helps individuals cope with the emotional consequences of a stressor [23,109]. Esteem support consists of compliments, provides agreement with the views of the recipient, and alleviates any feelings of guilt the recipient has about the situation [23,109]. Network or social support provides the stroke survivors or caregivers with access to people outside of their "inner circle" and reminds them that there are others who share similar experiences and are available for support [23,109]. Spiritual support is concerned with how people's religious or spiritual beliefs can create a sense of connection to cope with their diseases [8,34]. In contrast, tangible aid involves directly observable and agentic support and can include offers to do a direct task or to take over a task from the stroke survivors while they are stressed, to join the stroke survivors in an activity, and offers or expressions of willingness to help; it may also include lending money to the stroke survivors and caregivers [23]. Tangible aid can also involve the provision of meals, undertaking cleaning, or completing other tasks [109]. Stroke survivors and caregivers also need involvement in decision-making on the model of care, such as treatment that should be done to ensure optimal recovery for the stroke survivors [7,102].

Technological Interventions

A total of 60 of the 77 articles included in this review were concerned with a diversity of technological interventions aimed at supporting people after stroke. These included: (1) communication technologies (including emails or online videoconferencing); (2) community building, including the use of online forums; (3) well-being-related; for the purposes of health monitoring and management (such as sensors or wearable technology); or for the purposes of engaging with the world (such as via VR). Table 6 outlines these results.

XSL•FO
RenderX

Table . Technological interventions explored in the reviewed articles.

Technological intervention	References	Number of papers (%)
Mobile health (mHealth)	Thompson et al [4], Feigin et al [14][a], Camicia et al [49][a], Kerr et al [53], Lobo et al [55][a], Sarfo et al [58], Silvera-Tawil et al [60], Requena et al [66], Cooray et al [67], Zhou et al [71], Vloothuis et al [72], Andrades-González et al [77], Eriksson et al [80], Firdaus et al [82], Firmawati et al [83], Gong et al [84], Lobo et al [86], Kechik et al [87], Pereira et al [88], Nichols et al [90], Siegel et al [91], Olafsdottir et al [94][a], Olafsdottir et al [95][a], Juengst et al [96], Paul et al [113]	25 (32)
Electronic health record (EHR) or electronic medical record (EMR)	Deutschbein et al [105]	1 (1)
Electronic mail (email)	Cruickshank et al [111][a]	1 (1)
Web-based system	Leonardi and Fheodoroff [3], Feigin et al [14][a], Reszel et al [110], Camicia et al [49][a], Freund et al [51], Lobo et al [55][a], Reeves et al [65], Blanton et al [68], Caunca et al [69], Sureshkumar et al [70], Demir and Gozum [78], Favilla et al [81]	12 (16)
Online forum	Smith et al [8], Smith et al [109], De Simoni et al [79], Thomas et al [112]	4 (5)
Telehealth or telerehabilitation or teleconsultation or telestroke	Saywell et al [59], Mainali et al [64], Lo et al [76], Lam et al [97]	4 (5)
Video-guided exercise app	Scrivener et al [48], Camicia et al [49][a], Givon et al [74][a], Newland et al [89]	4 (5)
Communication technology such as video conferencing (ie, Zoom) and instant messaging (IM) app	Tsang et al [92], Cruickshank et al [111][a]	2 (2)
Virtual reality (VR) or augmented reality (AR)	Johnson et al [52], Krishnan et al [54], Luo et al [56], Sun et al [61], Lee et al [63], Givon et al [74][a], Giachero et al [75], Lo et al [98]	8 (10)
Sensors or wearable technology	Demers et al [50], Peters et al [57], Sun et al [61][a], English et al [73], Kamwesiga et al [85], Lee et al [93], Olafsdottir et al [94][a], Olafsdottir et al [95][a]	8 (10)
General platform use	Lobo et al [7], Gooch et al [9], Smythe et al [10], Sauvé-Schenk et al [12], Sauerzopf et al [15], Wright et al [23], Magwood et al [39], Xu et al [62], Aravind et al [99], Cameron and Gignac [100], Fu et al [101], Pindus et al [102], White et al [103], Norlander et al [104], Millar et al [106], Berkeley et al [107], Murakami et al [108]	17 (22)

[a]These studies explored more than one technological intervention.

Key Actors in Providing Technology-Mediated Support

Findings on community support and technological interventions were then compared across the different population and cohort groups. This often involved actors—individuals, groups/communities, and agencies who played an important role—beyond those who were study respondents, as can be seen in Table 5. Rather than summarizing Tables 4-6, Table 7 illustrates that different cohorts held different priorities in terms of support needs and the digital interventions used. Importantly, stroke caregivers were both providers and recipients, as is reflected in Table 5 and discussed below. Most technologies are developed to provide support from stroke caregivers, online communities/society, and health workers. A matrix approach provides more granular insights into the elements and interrelationship between elements of a stroke community ecosystem.

XSL•FO
RenderX

Table . Mapping actors, type of community support and supporting technology.

	Informational support	Emotional support	Social support	Esteem support	Spiritual support	Advice or knowledge	Tangible aid	Involvement in decision-making	Does not mention the type of online community support
Stroke caregivers (ie, family, friends)	Online forum [8,79,112], mHealth[a] [14,49,57,78,83,86,89,94], web-based system [14,68-70,81,110], VR/AR[c,d] gaming [98], IM[e] app [92]	Online forum [8,109,112], web-based system [68], mHealth [83,94-96]	Online forum [8,79,112], VR/AR [63,98], mHealth [14,49,77,80,88], mobile phone [85], web-based system [3,51,110]	Video-guided exercise app [48], mHealth [94]	Online forum [8]	Online forum [8]	VR/AR [54,61,74,75], mHealth [53,55,71,72,94-96], sensors or wearable technology [73,94,95], telehealth [76,97], video-guided exercise app [74], online forum [8], general platform use [23,107]	Web-based system [49,68]	mHealth [4], sensors or wearable technology [50]
Health workers (ie, therapist, nurse)	Online forum [112], VR/AR [52], telehealth [76], web-based system [65,68,78], mHealth [90,91], general platform use [10,29,101,102,105,106]	General platform use [101,106]	Online forum [112], wearable technology [85], mHealth [88], general platform use [7,10,15,99,106]	—[f]	—	VR/AR [52,63], mHealth [14,49,55,84,94,95], sensors or wearable technology [94,95], telehealth [76], video-guided app [49,89], web-based system [51,110], general platform use [101]	VR/AR [52,63,74,75], mHealth [45,53,66,71,72,94,95], video-guided exercise app [48,74], sensors or wearable technology [57,93-95], web-based system [65], telehealth [64,76,97], general platform use [107]	General platform use [7,102,106]	—
Community/society	Email [111], video conferencing [111], online forum [8,112], web-based system [65,68,69,110], mHealth [90], general platform use [7,9,10,23]	Online forum [8,109,112], general platform use [23,102]	Video conferencing [111], online forum [8,112], VR/AR [52], mHealth [58,60,94,95,113], sensors or wearable technology [57,61,85,94,95], telehealth [59,76], general platform use [9,10,23,30,104,108]	General platform use [12,23]	Online forum [8]	Email [111], video conferencing [111], online forum [8]	mHealth [53,58,60], telehealth [97], VR/AR [56,74,75], web-based system [55,65], video-guided exercise app [74], sensors or wearable technology [61], general platform use [23,107]	General platform use [102,106]	Sensors or wearable technology [50]
Funders	—	—	—	—	—	—	EHR/EMR[g,h] [105] mHealth [14]	—	—

XSL•FO
RenderX

	Informational support	Emotional support	Social support	Esteem support	Spiritual support	Advice or knowledge	Tangible aid	Involvement in decision-making	Does not mention the type of online community support
Policymakers	—	—	—	—	—	—	mHealth [14,53], EHR/EMR [105]	—	—

[b]mHealth: mobile health.

[c]VR: virtual reality.

[d]AR: augmented reality.

[e]IM: instant messaging.

[f]Not available.

[g]EMR: electronic medical record.

[h]EHR: electronic health record.

Stroke Survivors' and Caregivers' Support Requirements in the Online Community

Informational Support

Thomas et al [112] found that there is still a knowledge gap between health workers and stroke survivors and caregivers, which is a major contributor to burden for stroke survivors and caregivers [69]. As a result, they search for information from online communities to provide informal explanations and reassurance beyond that provided by the health workers [112].

The information sought is wide-ranging. Stroke survivors and caregivers require information on life after stroke, including about the underlying diseases, activities of daily living, healthy behaviors, medical information, treatment options, home-based exercises, functional skills training, and symptoms, as well as risk factors for stroke, such as ambulation and fall risk [14,23,69,70,78,79,82,83,86,87,90]. Blanton et al [68] found that providing information via short videos that engaged stroke survivors and caregivers in practicing physical tasks in the home environment would be welcomed.

The informational needs of stroke survivors and caregivers extended to mental health and well-being issues, including stress management, self-care, and practical fatigue management strategies [69,112]. These findings echo those of Lo et al [76] and Firmawati et al [83], who found that survivors and caregivers need information related to stroke care: nutrition, exercise, medications, community services, peer-sharing, and expert advice. They saw videos as appropriate for enhancing caring skills and confidence, and offering support to caregivers [81,83]. Narrating the story of stroke was seen as significant in prompting different forms of information sharing. Wright et al [23] and Smith et al [109] argued that offering stroke survivors an opportunity to share their rehabilitation journey story acted as a way to request (and offer) support from online communities.

Informational needs extend beyond the stroke survivor and their caregiver to other actors in their community. Clinical information is required by health workers for monitoring the health of stroke survivors. While this can occur through clinical and follow-up assessments or surveillance data, these can be complicated to achieve in some settings [14,67]. The source of required information should not only come from related actors such as health workers or health facilities but should also be integrated with medical record data [67]. Information principles should be well-organized, complete, and necessary, with helpful information provided in an accessible format, and including adequate resources [69].

Emotional Support

The reviewed research indicated that while stroke was seen as a physical illness, stroke survivors' emotions were affected alongside their physical function [106], thus emphasizing the need for emotional support for stroke survivors and caregivers. However, as found by Pindus et al [102], emotional support was lacking for stroke survivors and caregivers. For stroke caregivers, the required emotional support included acceptance of caregiving situations, depressive symptoms, and stress management [69,83,109]. Behavioral and cognitive issues—especially memory problems and resistance to accept help—experienced by the stroke survivors were of concern to the stroke caregivers as well as survivors themselves [69].

Through their analysis of an online community for pediatric stroke survivors, Wright et al [23] defined emotional support as including relationships, physical affection, confidentiality, sympathy, understanding or empathy, and encouragement. The importance of closeness was emphasized [23], including through offering physical affection (including via hugs and kisses) that leads to positive emotions [23,101,112]. Emotional support that facilitates individual positive emotions was noted to elevate one's overall health level and reduce negative emotions such as anxiety, fatigue, and depression [101,112]. Also important was the need for all support actors to keep stroke survivors' problems or difficulties confidential [23] and was especially important for online support communities. Gender may moderate these priorities as reported by Smith et al [8], who noted that women are likely to seek both emotional and informational support compared with men.

Online communities could also provide a platform to express sorrow or regret for the situation, and to seek reassurance or strategies for dealing with ongoing symptoms. They also allowed stroke survivors and caregivers to express their struggles in attempting to find a cause or express their difficulties in

XSL•FO
RenderX

understanding the situation [23]. At the same time, by providing a platform to disclose and acknowledge similar experiences in a way that conveys understanding, online communities provided stroke survivors and caregivers with hope and confidence [23]. Empathy for stroke survivors was often given by online communities, with emphasis placed on loneliness after stroke, strategies to fight the urge to give up on recovery, recognition of the symptoms experienced, and articulation of practical difficulties [23].

Esteem Support

Esteem support was defined as distinct from emotional support by Wright et al [23] as a form of recognition (for example, of the stroke survivor's contribution to a community) as well as to validate people's recovery processes. Sauvé-Schenk et al [12] found that identifying social service and community resource needs was important for tailoring stroke recovery interventions to stroke survivors' and caregivers' individual situations.

Esteem support also includes having the support to build esteem in others. While home-based exercise can lead to good functional results, esteem support can enhance these further. Olafsdottir et al [94,95] found that providing caregivers with information and knowledge on how to motivate the stroke survivor can empower them to become more willingly involved in the rehabilitation process at home and thus to support and motivate their stroke survivor to exercise and be more physically active. As reported by Olafsdottir et al [95] and Scrivener et al [48], community-dwelling stroke survivors found home-based exercise and physical activities more fun and less tedious with more variety in exercise and training options.

The categories of esteem support include compliments, validation, and relief of blame. Relief of blame functioned as support by reassuring a stroke survivor that they were not a burden and everyone needs support at some point [23]. Wright et al [23] also found that compliments were given to motivate stroke caregivers, and to congratulate people for joining the community. Validation was given when stroke survivors had been through bad experiences such as missed or delayed diagnoses, poor awareness of stroke, isolation, feeling tired, and wanting live chat rooms [23].

Social Support

The reviewed articles emphasized that it was not only stroke survivors who faced challenges or burdens after stroke, with many caregivers describing similar conditions. Most caregivers were unprepared for their caregiving role, particularly given the immediate need to provide care following their family member's stroke [7]. Norlander et al [104] found that people's abilities to overcome challenges and to adapt their behavior and attitude played a critical role in social participation and integration after stroke. Social participation refers to a person's engagement in activities with family, friends, peers, or community members [88]. Social integration, in contrast, is specifically concerned with the linkages between stroke survivors and caregivers with other members of their community [88]. Providing network or social support is a long-term process that constantly needs to be balanced against other priorities in life. Thus, Norlander et al [104] urged that network involvement need not only emphasize supporting each other, but also should emphasize people's motivations to be engaged in the network.

Social support was generally present in forums where messages expressed a sense of camaraderie, reiterating that members of the forum were there to support each other [23]. Social support activities could also be engaged in for pleasure, relaxation, or other emotional satisfaction, thus increasing stroke survivors' and caregivers' well-being [104]. At the same time, these social support activities were found to reduce depressive symptoms and improve functional recovery and survival after stroke [104]. By using social support in the community, stroke survivors and caregivers could also get information, knowledge, and resources related to stroke [88,102,111].

Social support often operated alongside esteem support, as could be seen where peer connection provided motivation to seek out and engage with different health workers [111]. Social support could also assist stroke survivors to use tools for physical activities or exercise in order to become familiar, comfortable, safe, and encouraged to finish their activities or exercises [94]. Olafsdottir et al [94] described examples of using tools on the proper use of a wheelchair in various scenarios—such as boarding a bus, parking a wheelchair, exiting a bus, using an elevator, and general maneuvering—as especially useful for stroke survivors and caregivers, thus linking with tangible aid (discussed below).

Spiritual Support

Our search identified limited studies [8,62] regarding prayer support, which we then combined with spiritual support to create a single category. Smith et al [8] define prayer support as a subcategory of spiritual support to improve the well-being, spiritual, or energetic state of the stroke survivor; however, prayer support does not have a material outcome (thus contrasts with action-facilitating forms of support, as in tangible aid). Stroke survivors wanted to be close with their religious community, such as the prayer group to which they belonged. Prayer or spiritual support includes providing and seeking prayers, spiritual blessings, positive karma, and warm thoughts in the community [8], and was given by writing comments for the online community [8]. Xu et al [62] described spiritual or prayer support as a form of healing, and others felt the need to retain a positive approach and not become a burden.

Advice or Knowledge

Stroke survivors and caregivers required advice or knowledge about stroke rehabilitation, specifically that provided by health care workers in the community on physical activity or exercise [14,94]. They also appreciated advice based on the stroke experiences of peers in the community [76]. Stroke caregivers expressed their need for education on how they could provide support with exercise and physical activity for the stroke survivors [94]. Stroke survivors and caregivers also needed advice on knowledge repositories or management tools, such as internet search engines, to be able to interact in the online community [55].

Advice and knowledge needed to be localized. Aside from text messages, Gong et al [84] found that health workers' voice messages using local dialects related to stroke advice or

XSL•FO
RenderX

knowledge could prevent stroke among people with low education levels in resource-limited settings. Factors including voice message structure, language, complexity and relevance, and repetition all influenced stroke survivors' and caregivers' acceptance of using voice-based apps [84]. Short video apps could be used by nurses to provide stroke education [89]. Important factors in providing stroke education videos related to the videos' frequencies, the ideal length of video delivery, and the need for repetitive opportunities to review educational stroke information [89].

Tangible Aid

Tangible aid for stroke survivors and caregivers was shown to consist of financial support (particularly through loans), and offering to take over tasks [23]. Multiple articles also identified that tangible aid could include the technologies and technical devices needed by stroke survivors and caregivers to improve their self-management [48,52,54,63,74,75,94]. Tangible aid was also facilitated via the provision of technology-based rehabilitation programs, which could provide a unique platform for stroke survivors who were unable to obtain rehabilitation services at professional institutions due to insufficient medical resources, economic disadvantage, and access issues (including inconvenient traffic) [75].

Involvement in Decision-Making

Five of the reviewed articles identified that the limited knowledge of stroke survivors and caregivers meant that community involvement in decision-making was sometimes necessary for certain actions such as taking certain medicine or treatment [7,49,68,102,106]. For example, Pindus et al [102] described that stroke survivors need help with their treatment and exercise planning, which enabled timely decisions for increasing their quality of life. Online communities facilitated the provision of informational support and advice, which could help stroke survivors and caregivers to improve their problem-solving and decision-making [102], such as about which health care disciplines to engage for particular issues.

Decision-making regarding treatments and exercises was identified by Millar et al [106] as needing to be shared equally by stroke survivors and their caregivers with health workers, who could advise on how to deal with issues or concerns as they arose. Nurses, for example, were not necessarily the most appropriate health workers to consult for medication management. In addition, the provision of prehospital screening and the development of stroke severity tools by online platforms could enhance stroke survivors' decision-making to ensure they receive the most appropriate level of care [49].

Outlining an Online Community Ecosystem for Stroke Survivors and Caregivers

Figure 4 describes a community ecosystem for stroke survivors that we have developed based on the reviewed literature. Of particular note, each layer of the ecosystem includes forms of support that are also provided through the lower layers.

XSL•FO
RenderX

Figure 4. The online community ecosystem for stroke survivors.

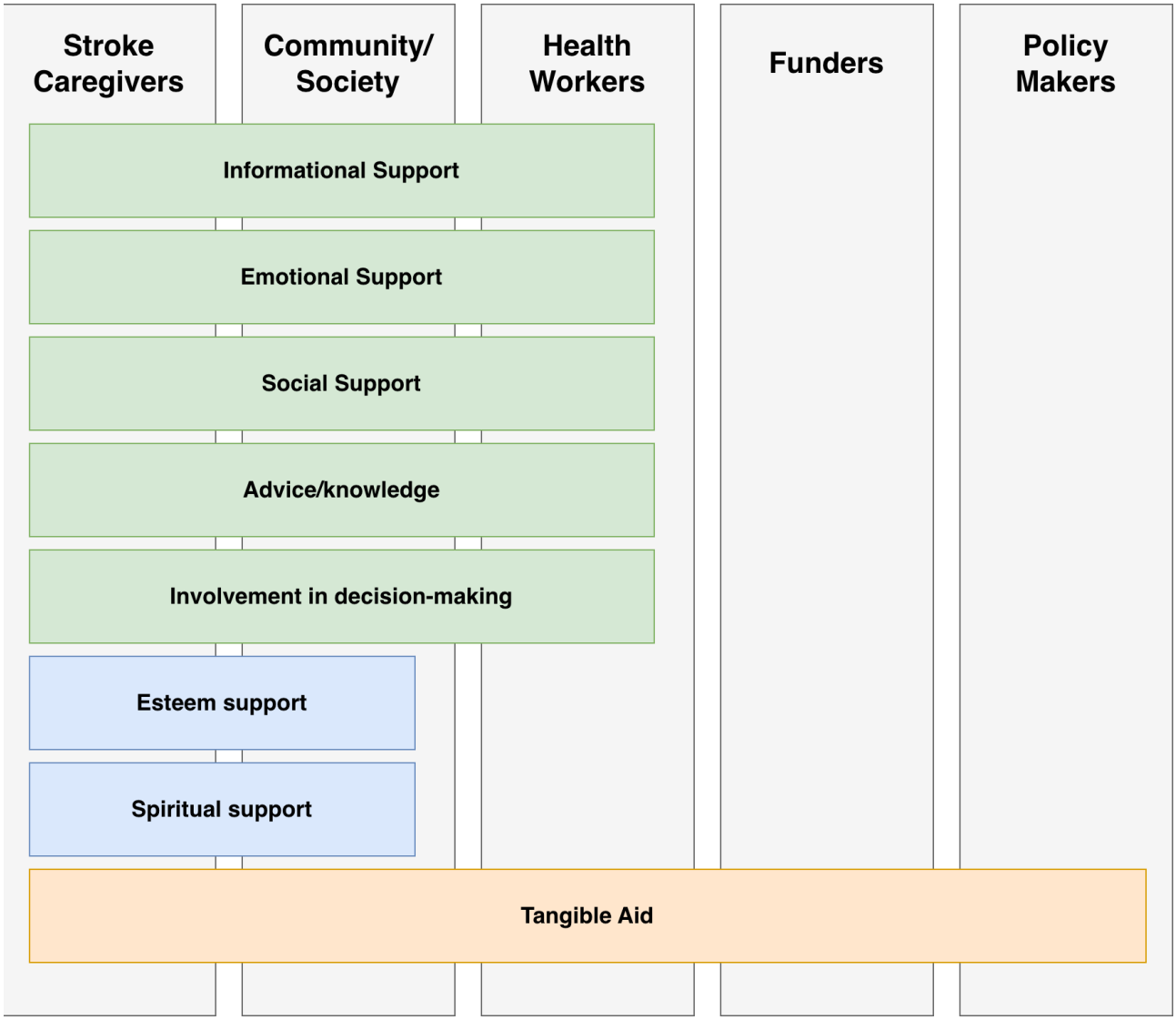

Our review findings indicate that stroke caregivers and community provide all types of support, and thus comprise the core support team for most stroke survivors. Health workers are also important, providing informational support, emotional support, social support, advice/knowledge, tangible aid, and involvement in decision-making. Health workers—such as nurses, physical therapists, occupational therapists, and speech-language pathologists—were commonly involved in community-based interventions to facilitate the necessary support to optimize stroke recovery [7,107]. Among health workers, nurses often served as coordinators of care and offered direct delivery of a range of interventions in the home and community settings [7,49]. This, combined with their capacity for continuous care, positioned nurses as critical health worker actors who are uniquely positioned to improve stroke caregivers' knowledge and assist stroke survivors [7,49].

Nurses play a key role both in acute care settings and postdischarge. Camicia et al [49], for example, emphasized the involvement of community-based health workers in providing longer-term support for stroke survivors following their transition home after a stroke. The community operates both as a destination of the recovery pathway and also as a partner that can help address the ongoing and often-unmet needs experienced postdischarge [49]. Partnerships between health services and the voluntary sector can offer innovative solutions to improve stroke survivors' recovery care [49]. Partnerships are especially important in LMIC due to the lack of community rehabilitation services; in such settings, stroke prevention services can be provided by wider community networks [14].

Online health communities can contribute to supporting stroke survivors and caregivers by providing informational support, emotional support, and discussing their conditions in an online forum [23,112]. Based on Smith et al [8], stroke survivors who received prayer-related words in online health communities experienced lower negative emotions and better well-being, possibly because religious coping mechanisms allowed them to experience less fear of death, place their trust in God, and view their illness experiences with more positivity. However, our review highlights that there remain limited studies that analyze esteem support, spiritual support, and involvement in decision-making. Most studies focused on stroke survivors' and caregivers' provision and receipt of informational support, advice/knowledge, and tangible aid for stroke physical recovery.

XSL•FO
RenderX

Technologies Used to Support Stroke Survivors and Caregivers in the Online Community

Technology-based interventions in the online community have been recommended to better meet the needs of stroke survivors and caregivers [69]. Promising technologies to meet these needs and support stroke survivors and caregivers identified in this review (Table 4c) provide advantages such as increasing technological literacy, asynchronous communication, use of multimedia platforms, and the potential to reduce disparities by increasing the reach of resources [69]. Most features used in technology-based interventions relate to rehabilitation and self-management for supporting changes in health behaviors, such as increasing physical activity [113]. Kerr et al [53] found the priorities for stroke rehabilitation systems are broad and include: access to technologies, ease of use, awareness of available technologies, technologies focused on function, supported self-management, user training, evidence of effectiveness, value for money, knowledgeable staff, and performance feedback.

Health workers were also shown to use similar technologies as stroke caregivers [61]. Our review showed that stroke caregivers used most technologies to provide support, except electronic health record/electronic medical record, email, digital health technology, and communication technologies such as video conferencing and IM apps. They also used technologies to search for information and tangible aid. mHealth, online forums, and web-based systems were the most used technologies by health workers, and were used by health workers to provide information, advice/knowledge, and tangible aid for stroke survivors' rehabilitation processes [94,95,112]. Through those technologies, health workers could also easily monitor the progress status of stroke survivors' health [48]. The community could provide all types of support to caregivers using similar technologies as health workers.

Since online forums could be used for 2-way communications, stroke caregivers also provided informational, emotional, and social support [79,109,112]. The online forum provided an open space to ask questions and share thoughts, as well as expressing comfort at reading the stories of other stroke survivors or caregivers, whilst at the same time stating their gratitude for receiving valuable assistance in a time of need [109].

mHealth and web-based systems were used to get and provide informational support, tangible aid, and involvement in decision-making for stroke caregivers. Camicia et al [49] argued that mHealth and telehealth or telerehabilitation could be used with other supporting tools to enhance the provision of care for stroke survivors. For example, clinicians could send digital education, community resources, tasks to complete (eg, home assessment), and surveys (eg, caregiver assessment) directly to the caregiver's mHealth for timely evaluation and provision of treatment. Telerehabilitation refers to the use of any telecommunication modality (eg, telephone, videophone, audio-video conference) for the delivery of rehabilitation services at a distance [97].

Providing tangible aid to stroke caregivers and survivors offered a very clear purpose for technologies. Gooch et al [9] and Olafsdottir [95] described how digital health technologies—such as VR, robotics, sensors, messaging platforms, or audio-video platforms—could be used for physical stroke rehabilitation. Video-guided physical rehabilitation—largely via exercise programs with a self-management approach—was identified as a feasible and accepted set of interventions for people with stroke where health workers could provide advice, education about physical activity, and activities to enhance the skills required for self-management of physical activity [48]. Video-guided exercise could also provide esteem support to stroke caregivers and survivors by providing ongoing motivation and tips to keep their rehabilitation interesting [48]. VR/augmented reality and sensors could also be used for supporting the physical rehabilitation process for stroke survivors [52,63]. VR/augmented reality and sensors could be connected to an online platform for collecting stroke survivors' information and providing timely feedback to stroke survivors' and therapists [52].

Articles relating to social support from the community focused on potential improvements for existing and future technologies. Smith et al [8] argued that stroke technologies' functionalities should always give stroke survivors and caregivers autonomy and agency, including the ability to approve or decline help offers from the community. Community-based health workers could provide training for stroke survivors and caregivers by using text messages, educational videos, and a workbook [14,49,68]. Criteria for health workers to use mHealth in assessing the appropriateness of mHealth apps were design, information/content, usability, functionality, ethics, security, privacy, and user-perceived value [96]. A key challenge for self-assessment and self-management using mHealth is keeping the stroke survivor and caregivers sufficiently engaged in the processes that are often repetitive, like answering questions about their status or completing home exercise programs; thus, reminders, notifications, and gamification features are deemed important to be provided in the mHealth [96].

Leonardi and Fheodoroff [3] found that internet-based or web-based systems could allow better exchange of information or stroke surveillance by involving communities, thus allowing data in real time and improving data quality [14]. Reszel et al [110] found that an internet-based system could be used to plan community-based exercise programs for people with stroke, such as conducting a thorough community assessment and developing referral pathways from the community. However, Sureshkumar et al [70] found that a major concern of internet-based systems was internet connectivity, especially in countries that do not have a stable internet connection.

Most online health communities provide support toward healing rather than health promotion [8]. These online communities help stroke survivors and caregivers through the rehabilitation process with the goal of improving functioning and prolonging the lives of stroke survivors [14].

Participatory design or co-design is often used when interactive technologies are being developed [95]. Smith et al [8] and Kerr [53] advocated for the use of participatory design that could engage all related actors, including technology developers and users, to design compassionate technology for facilitating social support for patients with life-threatening illnesses. This

XSL•FO
RenderX

participatory approach also includes consideration of trust as a factor that affects sensitive disclosures and use of technology during health crises; Smith et al [8] found that stroke survivors and caregivers already trust app-based transportation and financial services, which could be leveraged to generate better health outcomes.

Communication technologies such as telephone calls, short message services, or WhatsApp messages could also be integrated into the online community for easier exchange of information and notification of new information/discussion in the community [78,80,85,92]. Cruickshank et al [111] explained that video conferencing and emails could be used in virtual community-based stroke programs to optimize participant experiences and outcomes. A reliable sender, an accepted form of the message dispatch, an optimal timing based on stroke survivors' daily routine, and simple but relevant key messages were essential for the acceptance of message-based interventions [84]. Specifically, for voice messages, the most important factors that should be considered are the speed of audio playback, the pattern of repetition, and the use of dialect that may also improve audiences' understanding of the message contents [84]. Based on Kamwesiga et al [85], follow-up calls, text messages, and reminders can be used to support the rehabilitation process after stroke.

Innovative and emerging technologies, such as generative artificial intelligence (AI), could also be explored to help online communities get more information and support. While Jiang et al [114] found that AI is mostly still under exploration in research on stroke early detection and diagnosis, treatment and outcome prediction, and prognosis evaluation, an AI-driven chatbot is available and can be embedded into virtual community-based stroke programs.

Funders (ie, health regulators, nongovernment organizations, and private organizations) and policymakers (ie, health regulators) provide tangible aid to stroke survivors in the community [14]. Policymakers used mHealth to support stroke survivors' rehabilitation processes [53] by promoting the use of technologies in stroke rehabilitation. Policymakers are responsible for defining policies regarding the use of technology that can support online communities in supporting stroke survivors and caregivers [14,53]. They should engage with online communities, especially in providing health information and education related to stroke [14]. Policymakers also need to create a health ecosystem for managing stroke in a comprehensive and sustainable manner, where the online community is one of the actors involved in this ecosystem [14,53].

Discussion

Overview

Stroke recovery is a process that should be monitored and maintained for survivors and caregivers. This scoping review found that the key actors supporting stroke survivors are caregivers, the local community/society, and health workers. However, as stroke survivors face many difficulties during the rehabilitation process, it is important to identify innovative and scalable support solutions. Online platforms offer promising ways forward in the development of such solutions, and thus, the reviewed research emphasized the importance of online communities for stroke survivors and caregivers. The functions of online communities are to provide informational support, emotional support, esteem support, social support, spiritual support, advice/knowledge, tangible aid, and involvement in decision-making.

Even though technology has been shown to effectively connect stroke survivors and caregivers to online communities, there is a lack of integrated, available technology. As this review illustrates, stroke research related to the community and technology is underdeveloped. Therefore, this study provides insights for health regulators, health providers, health application developers, stroke survivors, and caregivers, as well as the community, to guide the provision of digital health technologies that could improve health outcomes. Our findings emphasize the importance of involving the online community in developing digital health apps for stroke survivors and caregivers to best support stroke recovery.

Principal Findings

Our review identified only 3 studies [53,60,88] that actually engage with stroke survivors and their caregivers to develop the technologies used in the stroke community, yet one of these [53] included stroke survivors and caregivers as part of a broader "user community's" priority setting activity. However, the importance of engagement is highlighted by a number of studies. From their user experience workshops with stroke survivors and health workers, Silvera-Tawil et al [60] found that there is a need for consumer-facing apps to integrate wearable activity trackers and a clinician web portal to support secondary prevention of stroke (ie, monitoring health and lifestyle measures). Pereira et al [88] conducted a workshop with stroke survivors, caregivers, and health workers to (1) explore the needs, concerns, motivations, and strategies for self-management support after stroke, (2) define the design principles based on social cognitive theory, and (3) develop and test a prototype using questionnaires to analyze usability. They found the need for personalized support for stroke management.

The design outcome from user-centered design should focus not only on the technology, but also on the health care outcomes. Lobo et al [55] and Lobo et al [86] suggested using user-centered design or participatory design to better understand the needs and respond to issues defined by users within the online community. In addition, Lobo et al [86], Kechik et al [87], Pereira et al [88], Siegel et al [91], and Juengst et al [96] suggested engaging stroke survivors and their caregivers in building apps to best meet their needs. In user-centered design, iterative processes—such as those generated by conducting workshops, focus group discussions, storyboards, games, etc—are required to better create technologies for intended users to support their daily activities [55,88]. Furthermore, these iterative processes may also help researchers or software developers to identify possible usability issues that may impact the interactions between the user and the technology [55,68,86].

Participatory design includes 2 principal values—participation and democracy—by involving several individuals with diverse

XSL•FO
RenderX

knowledge and experiences [55]. Kerr et al [53] and Blanton et al [68] proposed the principles that should be implemented for stroke technologies, such as in the rehabilitation process, namely ease of use, evidence of effectiveness, access, and value for money to enhance the technology adoption by users (ie, therapists, stroke survivors, and caregivers). Stroke digital apps should be easy to use for end users, where they could use the apps with one hand and without too much trouble [53,55,68,82]. Any app can provide value for money if the end users feel the benefits of using the app [53].

Shaw [115] found that users' technology confidence was positively associated with system usability. Shaw [115] described that digital inclusion, access to technology, and digital literacy were necessary to ensure universal access to and meaningful engagement with technology. Although technology infrastructure availability and accessibility are nationwide initiatives and policies, the online communities and health workers play a vital role in promoting digital equity, serving as educators, advocates, and digital navigators, such as guiding stroke survivors and caregivers through the complexities of the technology [115]. Multimedia Appendix 4 [3,4,8,14,48-61,63-98,105,109-113] summarizes technologies used to support stroke survivors in the community in the reviewed articles.

Limitations

Several factors shaped this review, which only includes English-language journal articles. We focused specifically on stroke, and by excluding articles that included stroke as part of a broader focus on "disability or health," we may have missed articles. Another limitation was related to terminology—an ever-expanding field. We focused on a subset of terminologies on the forms of technology; it is possible that some newer or less common technologies were not captured in our literature search. Similarly, online community support was also used as a broad umbrella term; some articles did not explicitly map the type of support for each actor, which meant that we inferred the types of support from descriptions given in the full-text article screening and extraction.

Future Directions

Future research could analyze the required functionality of technologies that should be implemented in the health ecosystem. Advanced functionalities that could be integrated with emerging technologies, such as AI, could also be enhanced for stroke digital health technologies. Providing user-centered design and apps for stroke survivors and caregivers could increase the adoption of digital health technologies specifically for the recovery process of stroke survivors (ie, rehabilitation and self-management) [53,55]. Information through digital technologies should be personalized and delivered at appropriate times (such as medication delivery, self-monitoring, and so on) [55].

Cultural differences and socioeconomic factors remain underexamined in the reviewed articles. Future research should examine how these factors could influence the community and technology engagement, as well as how technologies supporting the online community are shaped. Smith et al's [8] finding that gender may also influence health information-seeking behavior provides an important future direction by highlighting the importance of exploring stroke survivors' and caregivers' demographics to analyze in detail their information-seeking behaviors.

This review did not include social media as a specific source of support, although this remains a site for further exploring informal support, particularly in low-resource settings, where other technologies may be limited or unavailable. Future research should seek to understand how stroke survivors and caregivers access support in such settings, where a lack of health workers influences stroke recovery [14]. Finally, there is still limited research that explains in detail the role of funders and policymakers in the online stroke community; thus, future work could explore the strategic actors that could manage the sustainability of the stroke digital health apps in the online stroke community.

Conclusions

This review found that the online community plays an important role for stroke survivors and caregivers in their rehabilitation and recovery process and should be considered as part of a stroke ecosystem, in addition to stroke survivors, caregivers, health workers, community/society, funders, and policymakers (Figure 5).

XSL•FO
RenderX

Figure 5. An expanded stroke ecosystem.

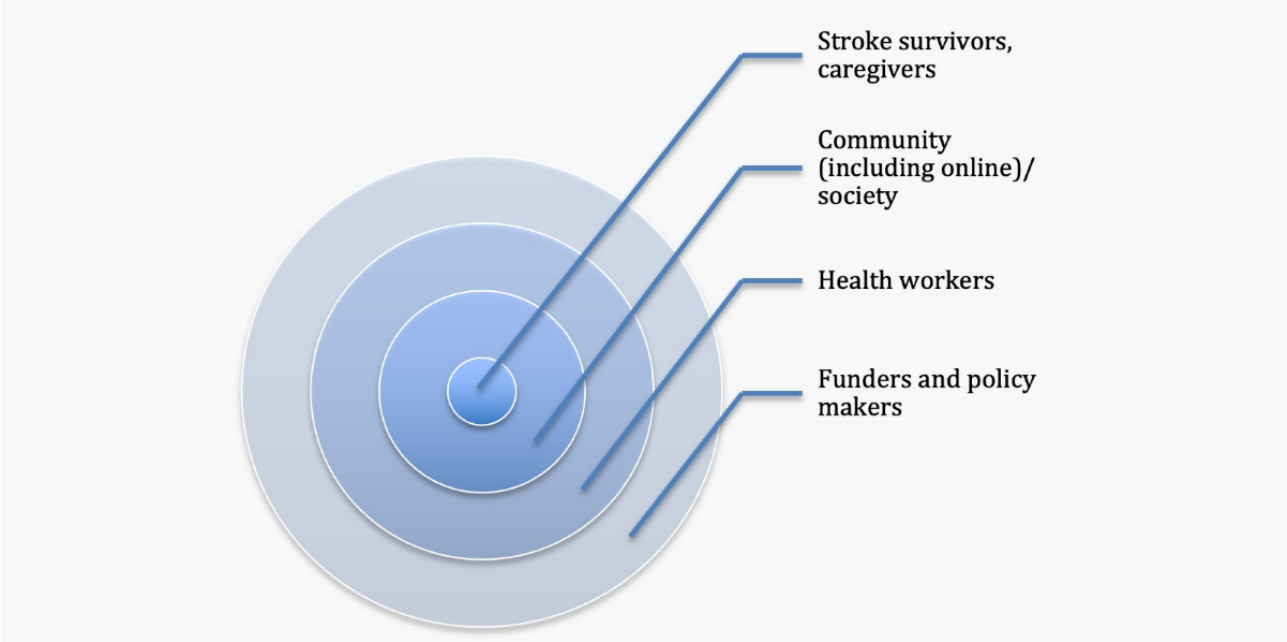

Online communities provide diverse forms of support, including informational support, emotional support, esteem support, social support, spiritual support, advice or knowledge, tangible aid, and involvement in decision-making. They therefore play an integral role in this new, expanded stroke ecosystem and offer the potential to connect to and communicate with stroke survivors and caregivers. As technologies are constantly evolving, emerging technologies such as AI may extend this ecosystem in the future by addressing stroke early detection and diagnosis, treatment, and outcome prediction and prognosis evaluation.

Funding

This research was supported by a Digital Transformations grant from Koneksi Knowledge Partnership Platform Australia Indonesia, Project 1447/CRG/2024/50-MOU.

Data Availability

All data supporting this study are available upon request.

Authors' Contributions

PWH conceptualized the study; developed the methodology; selected the articles; extracted, processed, and analyzed the articles; visualized the results; wrote the first draft of the manuscript; and reviewed and edited the manuscript. KAI conceptualized the study; developed the methodology; selected the articles; extracted, processed, and analyzed the articles; and visualized the results. JS conceptualized the study; developed the methodology; selected the articles; extracted, processed, and analyzed the articles; wrote the first draft of the manuscript; and reviewed and edited the manuscript. NW conceptualized the study; developed the methodology; selected the articles; extracted, processed, and analyzed the articles; visualized the results; and reviewed and edited the manuscript. SD, AM, MIJ, and EE reviewed and edited the manuscript. All authors approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy result.
[DOCX File, 3541 KB - jmir_v28i1e71190_app1.docx]

Multimedia Appendix 2
Data chart item (rev 8).
[DOCX File, 4028 KB - jmir_v28i1e71190_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Journals cited (rev 8).
[DOCX File, 2975 KB - jmir_v28i1e71190_app3.docx]

Multimedia Appendix 4
Technologies used to support stroke survivors in the online community.
[DOCX File, 3430 KB - jmir_v28i1e71190_app4.docx]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 4023 KB - jmir_v28i1e71190_app5.docx]

References

1. Feigin VL. Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Neurol 2021 Oct;20(10):795-820. [doi: 10.1016/S1474-4422(21)00252-0] [Medline: 34487721]
2. Venketasubramanian N, Yudiarto FL, Tugasworo D. Stroke burden and stroke services in Indonesia. Cerebrovasc Dis Extra 2022;12(1):53-57. [doi: 10.1159/000524161] [Medline: 35313314]
3. Leonardi M, Fheodoroff K. Goal setting with ICF (international classification of functioning, disability and health) and multidisciplinary team approach in stroke rehabilitation. In: Platz T, editor. Clinical Pathways in Stroke Rehabilitation: Evidence-Based Clinical Practice Recommendations: Springer; 2021:35-56. [doi: 10.1007/978-3-030-58505-1_3]
4. Thompson AN, Dawson DR, Legasto-Mulvale JM, et al. Mobile technology-based interventions for stroke self-management support: scoping review. JMIR Mhealth Uhealth 2023 Dec 6;11:e46558. [doi: 10.2196/46558] [Medline: 38055318]
5. Tyagi S, Koh GCH, Luo N, et al. Dyadic approach to supervised community rehabilitation participation in an Asian setting post-stroke: exploring the role of caregiver and patient characteristics in a prospective cohort study. BMJ Open 2020 Apr 23;10(4):e036631. [doi: 10.1136/bmjopen-2019-036631] [Medline: 32332008]
6. Warren N, Sakellariou D. Neurodegeneration and the intersubjectivities of care. Med Anthropol 2020 Jan;39(1):1-15. [doi: 10.1080/01459740.2019.1570189] [Medline: 30707041]
7. Lobo EH, Abdelrazek M, Grundy J, et al. Caregiver engagement in stroke care: opportunities and challenges in Australia and Denmark. Front Public Health 2021;9:758808. [doi: 10.3389/fpubh.2021.758808] [Medline: 34900907]
8. Smith CE, Levonian Z, Ma H, et al. I Cannot Do All of This Alone": exploring instrumental and prayer support in online health communities. ACM Trans Comput-Hum Interact 2020;27(5):1-41. [doi: 10.1145/3402855]
9. Gooch HJ, Jarvis KA, Stockley RC. Behavior change approaches in digital technology-based physical rehabilitation interventions following stroke: scoping review. J Med Internet Res 2024 Apr 24;26:e48725. [doi: 10.2196/48725] [Medline: 38656777]
10. Smythe T, Inglis-Jassiem G, Conradie T, et al. Access to health care for people with stroke in South Africa: a qualitative study of community perspectives. BMC Health Serv Res 2022 Apr 9;22(1):464. [doi: 10.1186/s12913-022-07903-9] [Medline: 35395847]
11. National service model for an integrated community stroke service. National Health Services. 2022. URL: https://www.england.nhs.uk/wp-content/uploads/2022/02/stroke-integrated-community-service-february-2022.pdf [accessed 2024-11-06]
12. Sauvé-Schenk K, Duong P, Savard J, Durand F. A systematic review of social service and community resource interventions following stroke. Disabil Rehabil 2022 Jun;44(13):2948-2957. [doi: 10.1080/09638288.2020.1851780] [Medline: 33280453]
13. Selamat SNS, Che Me R, Ahmad Ainuddin H, Salim MSF, Ramli HR, Romli MH. The application of technological intervention for stroke rehabilitation in southeast Asia: a scoping review with stakeholders' consultation. Front Public Health 2021;9:783565. [doi: 10.3389/fpubh.2021.783565] [Medline: 35198531]
14. Feigin VL, Owolabi MO, Feigin VL. Pragmatic solutions to reduce the global burden of stroke: a World Stroke Organization-Lancet Neurology Commission. Lancet Neurol 2023 Dec;22(12):1160-1206. [doi: 10.1016/S1474-4422(23)00277-6] [Medline: 37827183]
15. Sauerzopf L, Luft A, Maeusli V, Klamroth-Marganska V, Sy M, Spiess MR. Technology Use for Home-Based Stroke Rehabilitation in Switzerland From the Perspectives of Persons Living With Stroke, Informal Caregivers, and Therapists: Qualitative Interview and Focus Group Study. JMIR Rehabil Assist Technol 2024 Jul 18;11:e59781. [doi: 10.2196/59781] [Medline: 39024576]
16. Mohammed T, Nyante GG, Mothabeng JD. The impact of rehabilitation on the community life of stroke survivors in Accra, Ghana. S Afr J Physiother 2023;79(1):1839. [doi: 10.4102/sajp.v79i1.1839] [Medline: 36873961]
17. Pandian JD, Kalkonde Y, Sebastian IA, Felix C, Urimubenshi G, Bosch J. Stroke systems of care in low-income and middle-income countries: challenges and opportunities. Lancet 2020 Oct 31;396(10260):1443-1451. [doi: 10.1016/S0140-6736(20)31374-X] [Medline: 33129395]

XSL•FO
RenderX

18. Haldane V, Chuah FLH, Srivastava A, et al. Community participation in health services development, implementation, and evaluation: a systematic review of empowerment, health, community, and process outcomes. PLoS ONE 2019;14(5):e0216112. [doi: 10.1371/journal.pone.0216112] [Medline: 31075120]
19. Sacks E, Morrow M, Story WT, et al. Beyond the building blocks: integrating community roles into health systems frameworks to achieve health for all. BMJ Glob Health 2018;3(Suppl 3):e001384. [doi: 10.1136/bmjgh-2018-001384] [Medline: 31297243]
20. Guediri A, Mandigout S, Agbohessou KG, Daviet JC, Compagnat M. Impact of stroke impairment and phases in daily steps of stroke survivors living in community: a systematic review with meta-analysis. Am J Phys Med Rehabil 2025 May 1;104(5):e66-e74. [doi: 10.1097/PHM.0000000000002649] [Medline: 39642298]
21. Intamas U, Rawiworrakul T, Amnatsatsue K, Nanthamongkolchai S, Palmer MH. Care of stroke survivors in community: a case study of rural Thai community. J Health Res 2020 Sep 3;35(1):77-87. [doi: 10.1108/JHR-07-2019-0172]
22. Nguyen J, Arnaout S, Levison B, et al. Secondary stroke prevention trends in low-income community-dwelling older individuals. Sr Care Pharm 2025 May 1;40(5):223-229. [doi: 10.4140/TCP.n.2025.223] [Medline: 40296242]
23. Wright WJA, Howdle C, Coulson NS, De Simoni A. Exploring the types of social support exchanged by survivors of pediatric stroke and their families in an online peer support community: qualitative thematic analysis. J Med Internet Res 2024 Mar 15;26:e49440. [doi: 10.2196/49440] [Medline: 38488858]
24. Weltermann BM, Homann J, Rogalewski A, Brach S, Voss S, Ringelstein EB. Stroke knowledge among stroke support group members. Stroke 2000 Jun;31(6):1230-1233. [doi: 10.1161/01.str.31.6.1230] [Medline: 10835437]
25. Kessler D, Egan M, Kubina LA. Peer support for stroke survivors: a case study. BMC Health Serv Res 2014 Jun 16;14(1):256. [doi: 10.1186/1472-6963-14-256] [Medline: 24935460]
26. Reed MC, Wood V, Harrington R, Paterson J. Developing stroke rehabilitation and community services: a meta-synthesis of qualitative literature. Disabil Rehabil 2012;34(7):553-563. [doi: 10.3109/09638288.2011.613511] [Medline: 21985186]
27. Walsh ME, Galvin R, Loughnane C, Macey C, Horgan NF. Factors associated with community reintegration in the first year after stroke: a qualitative meta-synthesis. Disabil Rehabil 2015;37(18):1599-1608. [doi: 10.3109/09638288.2014.974834] [Medline: 25382215]
28. Gurr B. Emotional support for stroke survivors: share your story group. Int J Ther Rehabil 2009 Oct;16(10):564-572. [doi: 10.12968/ijtr.2009.16.10.44566]
29. Barger SD. Perceived emotional support and frequent social contacts are associated with greater knowledge of stroke warning signs: evidence from two cross-sectional us population surveys. J Health Psychol 2012 Mar;17(2):169-178. [doi: 10.1177/1359105311412837] [Medline: 21788305]
30. Harrison M, Ryan T, Gardiner C, Jones A. Psychological and emotional needs, assessment, and support post-stroke: a multi-perspective qualitative study. Top Stroke Rehabil 2017 Mar;24(2):119-125. [doi: 10.1080/10749357.2016.1196908] [Medline: 27309492]
31. Erler KS, Sullivan V, Mckinnon S, Inzana R. Social support as a predictor of community participation after stroke. Front Neurol 2019;10:1013. [doi: 10.3389/fneur.2019.01013] [Medline: 31616364]
32. Anderson S, Whitfield K. Social identity and stroke: “they don’t make me feel like, there’s something wrong with me”. Scand J Caring Sci 2013 Dec;27(4):820-830. [doi: 10.1111/j.1471-6712.2012.01086.x] [Medline: 23121474]
33. Lapadatu I, Morris R. The relationship between stroke survivors’ perceived identity and mood, self-esteem and quality of life. Neuropsychol Rehabil 2019 Mar;29(2):199-213. [doi: 10.1080/09602011.2016.1272468] [Medline: 28075218]
34. Northcott S, Hilari K. “I’ve got somebody there, someone cares”: what support is most valued following a stroke? Disabil Rehabil 2018 Oct;40(20):2439-2448. [doi: 10.1080/09638288.2017.1337242] [Medline: 28627938]
35. Oshvandi K, Torabi M, Khazaei M, Khazaei S, Yousofvand V. Impact of hope on stroke patients receiving a spiritual care program in Iran: a randomized controlled trial. J Relig Health 2024 Feb;63(1):356-369. [doi: 10.1007/s10943-022-01696-1] [Medline: 36435926]
36. Azar NS, Radfar M, Baghaei R. Spiritual self-care in stroke survivors: a qualitative study. J Relig Health 2022 Feb;61(1):493-506. [doi: 10.1007/s10943-020-01030-7] [Medline: 32445043]
37. Lamb M, Buchanan D, Godfrey CM, Harrison MB, Oakley P. The psychosocial spiritual experience of elderly individuals recovering from stroke: a systematic review. JBI Libr Syst Rev 2008;6(12):432-483. [doi: 10.11124/01938924-200806120-00001] [Medline: 27820476]
38. Chow EOW, Nelson-Becker H. Spiritual distress to spiritual transformation: stroke survivor narratives from Hong Kong. J Aging Stud 2010 Dec;24(4):313-324. [doi: 10.1016/j.jaging.2010.06.001]
39. Magwood GS, Nichols M, Jenkins C, et al. Community-based interventions for stroke provided by nurses and community health workers: a review of the literature. J Neurosci Nurs 2020 Aug;52(4):152-159. [doi: 10.1097/JNN.0000000000000512] [Medline: 32341258]
40. Forster A, Young J. Specialist nurse support for patients with stroke in the community: a randomised controlled trial. BMJ 1996 Jun 29;312(7047):1642-1646. [doi: 10.1136/bmj.312.7047.1642] [Medline: 8664717]
41. Yan LL, Li C, Chen J, et al. Prevention, management, and rehabilitation of stroke in low- and middle-income countries. eNeurologicalSci 2016 Mar;2:21-30. [doi: 10.1016/j.ensci.2016.02.011] [Medline: 29473058]

XSL•FO
RenderX

42. Yap KH, Warren N, Allotey P, Reidpath DD. Understandings of stroke in rural Malaysia: ethnographic insights. Disabil Rehabil 2021 Feb;43(3):345-353. [doi: 10.1080/09638288.2019.1624841] [Medline: 31169419]
43. Newell JM, Lyons R, Martin-Misener R, Shearer CL. Creating a supportive environment for living with stroke in rural areas: two low-cost community-based interventions. Top Stroke Rehabil 2009;16(2):147-156. [doi: 10.1310/tsr1602-147] [Medline: 19581201]
44. Denham AMJ, Wynne O, Baker AL, Spratt NJ, Bonevski B. The unmet needs of carers of stroke survivors: an evaluation of Google search results. Health Informatics J 2020 Jun;26(2):934-944. [doi: 10.1177/1460458219852530] [Medline: 31213117]
45. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
46. PRISMA for scoping reviews (PRISMA-scr). PRISMA. 2024. URL: https://www.prisma-statement.org/scoping [accessed 2024-11-06]
47. Covidence - better systematic review management. Covidence.: Covidence URL: https://www.covidence.org/ [accessed 2025-07-29]
48. Scrivener K, Sewastenko J, Bouvier-Farrell A, et al. Feasibility of a self-managed, video-guided exercise program for community-dwelling people with stroke. Stroke Res Treat 2021;2021:5598100. [doi: 10.1155/2021/5598100] [Medline: 34035890]
49. Camicia M, Lutz B, Summers D, Klassman L, Vaughn S. Nursing's role in successful stroke care transitions across the continuum: from acute care into the community. Stroke 2021 Dec;52(12):e794-e805. [doi: 10.1161/STROKEAHA.121.033938] [Medline: 34727736]
50. Demers M, Bishop L, Cain A, et al. Wearable technology to capture arm use of people with stroke in home and community settings: feasibility and early insights on motor performance. Phys Ther 2024 Feb 1;104(2):zad172. [doi: 10.1093/ptj/pzad172] [Medline: 38166199]
51. Freund M, Carey M, Dilworth S, et al. Effectiveness of information and communications technology interventions for stroke survivors and their support people: a systematic review. Disabil Rehabil 2022 Aug;44(17):4563-4578. [doi: 10.1080/09638288.2021.1913245] [Medline: 33905279]
52. Johnson L, Bird ML, Muthalib M, Teo WP. An innovative stroke interactive virtual therapy (STRIVE) online platform for community-dwelling stroke survivors: a randomized controlled trial. Arch Phys Med Rehabil 2020 Jul;101(7):1131-1137. [doi: 10.1016/j.apmr.2020.03.011] [Medline: 32283048]
53. Kerr A, Smith M, Reid L, Baillie L. Adoption of stroke rehabilitation technologies by the user community: qualitative study. JMIR Rehabil Assist Technol 2018 Aug 17;5(2):e15. [doi: 10.2196/rehab.9219] [Medline: 30120086]
54. Krishnan S, Mandala MA, Wolf SL, Howard A, Kesar TM. Perceptions of stroke survivors regarding factors affecting adoption of technology and exergames for rehabilitation. PM R 2023 Nov;15(11):1403-1410. [doi: 10.1002/pmrj.12963] [Medline: 36787167]
55. Lobo EH, Frølich A, Rasmussen LJ, et al. Understanding the methodological issues and solutions in the research design of stroke caregiving technology. Front Public Health 2021;9:647249. [doi: 10.3389/fpubh.2021.647249] [Medline: 33937175]
56. Luo Z, Lim AEP, Durairaj P, Tan KK, Verawaty V. Development of a compensation-aware virtual rehabilitation system for upper extremity rehabilitation in community-dwelling older adults with stroke. J Neuroeng Rehabil 2023 May 1;20(1):56. [doi: 10.1186/s12984-023-01183-y] [Medline: 37127574]
57. Peters DM, O'Brien ES, Kamrud KE, et al. Utilization of wearable technology to assess gait and mobility post-stroke: a systematic review. J Neuroeng Rehabil 2021 Apr 21;18(1):67. [doi: 10.1186/s12984-021-00863-x] [Medline: 33882948]
58. Sarfo FS, Obiako R, Nichols M, et al. Knowledge and perspectives of community members on risk assessment for stroke prevention using mobile health approaches in Nigeria. J Stroke Cerebrovasc Dis 2023 Sep;32(9):107265. [doi: 10.1016/j.jstrokecerebrovasdis.2023.107265] [Medline: 37487320]
59. Saywell NL, Vandal AC, Mudge S, et al. Telerehabilitation after stroke using readily available technology: a randomized controlled trial. Neurorehabil Neural Repair 2021 Jan;35(1):88-97. [doi: 10.1177/1545968320971765] [Medline: 33190615]
60. Silvera-Tawil D, Cameron J, Li J, et al. Multicomponent support program for secondary prevention of stroke using digital health technology: co-design study with people living with stroke or transient ischemic attack. J Med Internet Res 2024 Aug 22;26:e54604. [doi: 10.2196/54604] [Medline: 39172512]
61. Sun X, Ding J, Dong Y, et al. A survey of technologies facilitating home and community-based stroke rehabilitation. Int J Hum Comput Interact 2023 Mar 16;39(5):1016-1042. [doi: 10.1080/10447318.2022.2050545]
62. Xu T, O'Loughlin K, Clemson L, Lannin NA, Dean C, Koh G. Developing a falls prevention program for community-dwelling stroke survivors in Singapore: client and caregiver perspectives. Disabil Rehabil 2019 May;41(9):1044-1054. [doi: 10.1080/09638288.2017.1419293] [Medline: 29277114]
63. Lee M, Son J, Kim J, Pyun SB, Eun SD, Yoon BC. Comparison of individualized virtual reality- and group-based rehabilitation in older adults with chronic stroke in community settings: a pilot randomized controlled trial. Eur J Integr Med 2016 Oct;8(5):738-746. [doi: 10.1016/j.eujim.2016.08.166]
64. Mainali S, Stutzman S, Sengupta S, et al. Feasibility and efficacy of nurse-driven acute stroke care. J Stroke Cerebrovasc Dis 2017 May;26(5):987-991. [doi: 10.1016/j.jstrokecerebrovasdis.2016.11.007] [Medline: 28012837]

XSL•FO
RenderX

65. Reeves MJ, Fritz MC, Woodward AT, et al. Michigan stroke transitions trial. Circ Cardiovasc Qual Outcomes 2019 Jul;12(7):e005493. [doi: 10.1161/CIRCOUTCOMES.119.005493] [Medline: 31296043]
66. Requena M, Montiel E, Baladas M. Farmalarm: Application for Mobile Devices Improves Risk Factor Control After Stroke. Stroke 2019;50(7):1819-1824. [doi: 10.1161/STROKEAHA.118.024355]
67. Cooray C, Matusevicius M, Wahlgren N, Ahmed N. Mobile phone-based questionnaire for assessing 3 months modified rankin score after acute stroke: a pilot study. Circ Cardiovasc Qual Outcomes 2015 Oct;8(6 Suppl 3):S125-S130. [doi: 10.1161/CIRCOUTCOMES.115.002055] [Medline: 26515200]
68. Blanton S, Dunbar S, Clark PC. Content validity and satisfaction with a caregiver-integrated web-based rehabilitation intervention for persons with stroke. Top Stroke Rehabil 2018 Apr;25(3):168-173. [doi: 10.1080/10749357.2017.1419618] [Medline: 29334344]
69. Caunca MR, Simonetto M, Hartley G, Wright CB, Czaja SJ. Design and usability testing of the stroke caregiver support system: a mobile-friendly website to reduce stroke caregiver burden. Rehabil Nurs 2020;45(3):166-177. [doi: 10.1097/RNJ.0000000000000196] [Medline: 30418319]
70. Sureshkumar K, Murthy G, Natarajan S, Naveen C, Goenka S, Kuper H. Evaluation of the feasibility and acceptability of the “Care for Stroke” intervention in India, a smartphone-enabled, carer-supported, educational intervention for management of disability following stroke. BMJ Open 2016 Feb 2;6(2):e009243. [doi: 10.1136/bmjopen-2015-009243] [Medline: 26839011]
71. Zhou B, Zhang J, Zhao Y, et al. Caregiver-delivered stroke rehabilitation in rural China. Stroke 2019 Jul;50(7):1825-1830. [doi: 10.1161/STROKEAHA.118.021558] [Medline: 31177978]
72. Vloothuis J, de Bruin J, Mulder M, Nijland R, Kwakkel G, van Wegen EEH. Description of the CARE4STROKE programme: a caregiver-mediated exercises intervention with e-health support for stroke patients. Physiother Res Int 2018 Jul;23(3):e1719. [doi: 10.1002/pri.1719] [Medline: 29797740]
73. English C, Healy GN, Olds T, et al. Reducing sitting time after stroke: a phase II safety and feasibility randomized controlled trial. Arch Phys Med Rehabil 2016 Feb;97(2):273-280. [doi: 10.1016/j.apmr.2015.10.094] [Medline: 26529471]
74. Givon N, Zeilig G, Weingarden H, Rand D. Video-games used in a group setting is feasible and effective to improve indicators of physical activity in individuals with chronic stroke: a randomized controlled trial. Clin Rehabil 2016 Apr;30(4):383-392. [doi: 10.1177/0269215515584382] [Medline: 25952586]
75. Giachero A, Calati M, Pia L, et al. Conversational therapy through semi-immersive virtual reality environments for language recovery and psychological well-being in post stroke aphasia. Behav Neurol 2020;2020:2846046. [doi: 10.1155/2020/2846046] [Medline: 32831969]
76. Lo SHS, Chau JPC, Lau AYL, et al. Virtual multidisciplinary stroke care clinic for community-dwelling stroke survivors: a randomized controlled trial. Stroke 2023 Oct;54(10):2482-2490. [doi: 10.1161/STROKEAHA.123.043605] [Medline: 37551588]
77. Andrades-González I, Rodríguez-Estrabot N, Magdaleno-Moya R, Molina-Mula J. Perceptions and attitudes of informal caregivers of stroke patients regarding the Stroke-CareApp: a phenomenological study. Healthcare (Basel) 2025 Aug 22;13(17):2082. [doi: 10.3390/healthcare13172082] [Medline: 40941434]
78. Demir Y, Gozum S. Evaluation of quality, content, and use of the web site prepared for family members giving care to stroke patients. Comput Inform Nurs 2015 Sep;33(9):396-403. [doi: 10.1097/CIN.0000000000000165] [Medline: 26125455]
79. De Simoni A, Shanks A, Balasooriya-Smeekens C, Mant J. Stroke survivors and their families receive information and support on an individual basis from an online forum: descriptive analysis of a population of 2348 patients and qualitative study of a sample of participants. BMJ Open 2016 Apr 6;6(4):e010501. [doi: 10.1136/bmjopen-2015-010501] [Medline: 27053271]
80. Eriksson G, Söderhielm K, Erneby M, Guidetti S. Family members’ experiences of a person-centered information and communication technology-supported intervention for stroke rehabilitation (F@ce 2.0): qualitative analysis. JMIR Rehabil Assist Technol 2025 May 2;12:e69878. [doi: 10.2196/69878] [Medline: 40315427]
81. Favilla CG, Reehal N, Cummings SR, et al. Personalized video-based educational platform to improve stroke knowledge: a randomized clinical trial. J Am Heart Assoc 2024 Aug 6;13(15):e035176. [doi: 10.1161/JAHA.124.035176] [Medline: 39082426]
82. Firdaus A, Noor AAA, Wan AWZ. Explorative survey on the usage and needs of mobile health applications (mHealth) amongst caregivers in taking care of stroke survivors. MH 2022;17(2):85-100. [doi: 10.17576/MH.2022.1702.07]
83. Firmawati E, Setyopanoto I, Pangastuti HS. Mobile health application to support family caregivers in recurrent stroke prevention: scoping review. Open Access Maced J Med Sci 2022;9(T5):142-151. [doi: 10.3889/oamjms.2021.7859]
84. Gong E, Gu W, Luo E, et al. Development and local contextualization of mobile health messages for enhancing disease management among community-dwelling stroke patients in rural China: multimethod study. JMIR Mhealth Uhealth 2019 Dec 17;7(12):e15758. [doi: 10.2196/15758] [Medline: 31845901]
85. Kamwesiga JT, Tham K, Guidetti S. Experiences of using mobile phones in everyday life among persons with stroke and their families in Uganda - a qualitative study. Disabil Rehabil 2017 Mar;39(5):438-449. [doi: 10.3109/09638288.2016.1146354] [Medline: 26939597]

XSL•FO
RenderX

86. Lobo EH, Frølich A, Kensing F, et al. mHealth applications to support caregiver needs and engagement during stroke recovery: a content review. Res Nurs Health 2021 Feb;44(1):213-225. [doi: 10.1002/nur.22096] [Medline: 33341958]
87. Long Tuan Kechik TSM, Musa KI, Abdullah JM, et al. A narrative review on mobile health (mHealth) app for stroke care and rehabilitation intervention for Malaysia. Malays J Med Sci 2025 Jun 30;32(3):49. [doi: 10.21315/mjms-03-2025-154] [Medline: 40703226]
88. Mendes Pereira C, Matos M, Carvalho D, et al. Building bridges between people with stroke, families, and health professionals: development of a blended care program for self-management. J Clin Med 2024 Jan 4;13(1):300. [doi: 10.3390/jcm13010300] [Medline: 38202307]
89. Newland P, Sargent R, Van Aman MN, Hamilton K, Hendricks-Ferguson VL. Use of video education with hospitalized acute stroke patients: a literature review. MSJ 2023;32(2):106. [doi: 10.62116/MSJ.2023.32.2.106]
90. Nichols M, Sarfo FS, Singh A, et al. Assessing mobile health capacity and task shifting strategies to improve hypertension among Ghanaian stroke survivors. Am J Med Sci 2017 Dec;354(6):573-580. [doi: 10.1016/j.amjms.2017.08.005] [Medline: 29208254]
91. Siegel J, Edwards E, Mooney L, et al. A feasibility pilot using a mobile personal health assistant (PHA) app to assist stroke patient and caregiver communication after hospital discharge. Mhealth 2016;2:31. [doi: 10.21037/mhealth.2016.08.02] [Medline: 28293604]
92. Tsang WN, Lee JJ, Yang SC, Poon JCY, Lau EYY. Stroke caregivers' perception on instant messaging application use for psychological intervention: a qualitative study. Psychol Health Med 2024 Aug;29(7):1208-1221. [doi: 10.1080/13548506.2024.2332925] [Medline: 38528430]
93. Lee SI, Adans-Dester CP, Grimaldi M, et al. Enabling stroke rehabilitation in home and community settings: a wearable sensor-based approach for upper-limb motor training. IEEE J Transl Eng Health Med 2018;6:2100411. [doi: 10.1109/JTEHM.2018.2829208] [Medline: 29795772]
94. Olafsdottir SA, Jonsdottir H, Bjartmarz I, et al. Feasibility of ActivABLES to promote home-based exercise and physical activity of community-dwelling stroke survivors with support from caregivers: a mixed methods study. BMC Health Serv Res 2020 Jun 22;20(1):562. [doi: 10.1186/s12913-020-05432-x] [Medline: 32571316]
95. Olafsdottir SA, Jonsdottir H, Magnusson C, et al. Developing ActivABLES for community-dwelling stroke survivors using the medical research council framework for complex interventions. BMC Health Serv Res 2020 May 25;20(1):463. [doi: 10.1186/s12913-020-05198-2] [Medline: 32450854]
96. Juengst SB, Terhorst L, Nabasny A, et al. Use of mHealth technology for patient-reported outcomes in community-dwelling adults with acquired brain injuries: a scoping review. Int J Environ Res Public Health 2021 Feb 23;18(4):2173. [doi: 10.3390/ijerph18042173] [Medline: 33672183]
97. Lam SKY, Chau JPC, Lo SHS, et al. Evaluation of cost-effectiveness of a virtual multidisciplinary stroke care clinic for community-dwelling survivors of stroke. J Am Heart Assoc 2024 Sep 3;13(17):e035367. [doi: 10.1161/JAHA.124.035367] [Medline: 39189616]
98. Lo SHS, Chau JPC, Choi KC, et al. Stroke survivor and caregiver experiences of virtual reality gaming to promote social participation: a qualitative study. PLoS ONE 2024;19(12):e0315826. [doi: 10.1371/journal.pone.0315826] [Medline: 39693334]
99. Aravind G, Bashir K, Cameron JI, et al. Community-based exercise programs incorporating healthcare-community partnerships to improve function post-stroke: feasibility of a 2-group randomized controlled trial. Pilot Feasibility Stud 2022 Apr 22;8(1):88. [doi: 10.1186/s40814-022-01037-9] [Medline: 35459194]
100. Cameron JI, Gignac MAM. "Timing It Right": a conceptual framework for addressing the support needs of family caregivers to stroke survivors from the hospital to the home. Patient Educ Couns 2008 Mar;70(3):305-314. [doi: 10.1016/j.pec.2007.10.020] [Medline: 18155388]
101. Fu B, Mei Y, Lin B, et al. Effects of a benefit-finding intervention in stroke caregivers in communities. Clin Gerontol 2022;45(5):1317-1329. [doi: 10.1080/07317115.2020.1765062] [Medline: 32496892]
102. Pindus DM, Mullis R, Lim L, et al. Correction: stroke survivors' and informal caregivers' experiences of primary care and community healthcare services - a systematic review and meta-ethnography. PLoS ONE 2018;13(4):e0196185. [doi: 10.1371/journal.pone.0196185] [Medline: 29668761]
103. White J, Magin P, Attia J, Sturm J, McElduff P, Carter G. Predictors of health-related quality of life in community-dwelling stroke survivors: a cohort study. Fam Pract 2016 Aug;33(4):382-387. [doi: 10.1093/fampra/cmw011] [Medline: 26980854]
104. Norlander A, Iwarsson S, Jönsson AC, Lindgren A, Månsson Lexell E. Participation in social and leisure activities while re-constructing the self: understanding strategies used by stroke survivors from a long-term perspective. Disabil Rehabil 2022 Aug;44(16):4284-4292 [FREE Full text] [doi: 10.1080/09638288.2021.1900418] [Medline: 33779458]
105. Deutschbein J, Grittner U, Schneider A, Schenk L. Community care coordination for stroke survivors: results of a complex intervention study. BMC Health Serv Res 2020 Dec 19;20(1):1143. [doi: 10.1186/s12913-020-05993-x] [Medline: 33341112]
106. Duncan Millar J, Mason H, Kidd L, IMPETUS Research Team. What is important in supporting self-management in community stroke rehabilitation? a Q methodology study. Disabil Rehabil 2023 Jul;45(14):2307-2315. [doi: 10.1080/09638288.2022.2087766] [Medline: 35757835]

XSL•FO
RenderX

107. Jones Berkeley SB, Johnson AM, Mormer ER, et al. Referral to community-based rehabilitation following acute stroke: findings from the COMPASS pragmatic trial. Circ Cardiovasc Qual Outcomes 2024 Jan;17(1):e010026. [doi: 10.1161/CIRCOUTCOMES.123.010026] [Medline: 38189125]
108. Murakami K, Tsubota-Utsugi M, Satoh M, et al. Impaired higher-level functional capacity as a predictor of stroke in community-dwelling older adults: the Ohasama study. Stroke 2016 Feb;47(2):323-328. [doi: 10.1161/STROKEAHA.115.011131] [Medline: 26732573]
109. Smith FE, Jones C, Gracey F, Mullis R, Coulson NS, De Simoni A. Emotional adjustment post-stroke: a qualitative study of an online stroke community. Neuropsychol Rehabil 2021 Apr;31(3):414-431. [doi: 10.1080/09602011.2019.1702561] [Medline: 31855105]
110. Reszel J, van den Hoek J, Nguyen T, et al. How community-based teams use the stroke recovery in motion implementation planner: longitudinal qualitative field test study. JMIR Form Res 2022 Jul 29;6(7):e37243. [doi: 10.2196/37243] [Medline: 35904855]
111. Cruickshank A, Brooks ED, Sperling C, Nelson ML, Singh H. Exploring the experiences of adults with stroke in virtual community-based stroke programs: a qualitative descriptive study. BMC Health Serv Res 2024;24(1):600. [doi: 10.1186/s12913-024-11043-7] [Medline: 38715062]
112. Thomas K, Gamlin C, De Simoni A, Mullis R, Mant J. How is poststroke fatigue understood by stroke survivors and carers? a thematic analysis of an online discussion forum. BMJ Open 2019 Jul 9;9(7):e028958. [doi: 10.1136/bmjopen-2019-028958] [Medline: 31289087]
113. Paul L, Wyke S, Brewster S, et al. Increasing physical activity in stroke survivors using STARFISH, an interactive mobile phone application: a pilot study. Top Stroke Rehabil 2016 Jun;23(3):170-177. [doi: 10.1080/10749357.2015.1122266] [Medline: 27077973]
114. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec;2(4):230-243. [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
115. Shaw RJ. Access to technology and digital literacy as determinants of health and health care. Creat Nurs 2023 Aug;29(3):258-263. [doi: 10.1177/10784535231211682] [Medline: 37909069]

Abbreviations

AI: artificial intelligence
LMIC: low- and middle-income country
mHealth: mobile health
PRISMA: Preferred reporting items for systematic reviews and meta-analyses
PRISMA-ScR: Preferred reporting items for systematic reviews and meta-analyses extension for Scoping Reviews
VR : virtual reality

Edited by A Coristine; submitted 12.Jan.2025; peer-reviewed by F Opia, JP Engkasan, NE Suarez, OLG Atigossou, T Bakas; revised version received 27.Feb.2026; accepted 27.Feb.2026; published 29.Apr.2026.

Please cite as:
Handayani PW, Alifia K, Sutanto J, Erlina E, Davies SG, Mawuntu AHP, Jusuf MI, Warren N
Online Community Support for Stroke Survivors and Caregivers: Scoping Review
J Med Internet Res 2026;28:e71190
URL: https://www.jmir.org/2026/1/e71190
doi:10.2196/71190

© Putu Wuri Handayani, Kamila Alifia, Juliana Sutanto, Erlin Erlina, Sharyn Graham Davies, Arthur HP Mawuntu, Muhammad Isman Jusuf, Narelle Warren. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Image-Based Deep Learning for Cataract Diagnosis: Systematic Review and Meta-Analysis

Ruixi Li[1*], MM; Hongyi Li[1*], MM; Chong Li[2*], BM; Shuo Li[3], MM; Linhong Lei[4], MM; Dan Tao[5], MD

[1]Department of Ophthalmology, The Second Affiliated Hospital of Kunming Medical University, Kunming, Yunnan Province, China

[2]Department of Ophthalmology, People's Hospital of Dongchuan District, Kunming, Yunnan Province, China

[3]Department of Ophthalmology, Kunming Children's Hospital (The affiliated Children's Hospital of Kunming Medical University), Kunming, China

[4]Department of Ophthalmology, Qilin District People's Hospital of Qujing City, Qujing, Yunnan Province, China

[5]Department of Ophthalmology, Yunnan Provincial Maternal and Child Health Hospital, No. 200 Gulou Road, Wuhua District, Kunming, Yunnan Province, China

[*]these authors contributed equally

Corresponding Author:
Dan Tao, MD
Department of Ophthalmology, Yunnan Provincial Maternal and Child Health Hospital, No. 200 Gulou Road, Wuhua District, Kunming, Yunnan Province, China

Abstract

Background: Cataracts are an eye condition characterized by high prevalence and blindness-inducing potential, and effective approaches are required for their early diagnosis, underscoring the clinical significance of this study.

Objective: This study aims to evaluate the performance of deep learning (DL) in cataract diagnosis and assess its potential as an effective tool for automated diagnosis, and compare the diagnostic accuracy of DL versus both machine learning and human experts.

Methods: A systematic search was conducted in Web of Science, Embase, IEEE Xplore, PubMed, and Cochrane Library until April 1, 2025, for studies on image-based DL for cataract detection or clinical subtype classification. The included studies were assessed for the risk of bias (RoB) using Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2). Bivariate mixed effects models were used for data analyses, and publication bias was assessed by Deeks' funnel plots.

Results: Sixty-three studies were finally included. The quality assessment indicated a high or unclear RoB in the patient selection (34 studies) and index test (44 studies) domains. Meanwhile, in the reference standard domain, the risk of bias was high or unclear in only 2 studies. Image-based DL achieved a sensitivity of 96% (95% CI 0.95 - 0.97) and a specificity of 98% (0.96 - 0.98) for cataract detection, with an area under the ROC curve (AUC) of 0.99 (0.98 - 1.00). For cataract classification, the sensitivity and specificity of image-based DL were 94% (0.93 - 0.96) and 97% (0.96 - 0.98), respectively, with an AUC of 0.99 (0.98 - 0.99). Despite the strong overall performance, the model's generalization capability was challenged by its lower performance observed on independent external datasets (detection: sensitivity 87%, specificity 93%; classification: sensitivity 89%, specificity 90%), potentially attributable to domain shift between the training and validation data.

Conclusions: Image-based DL has demonstrated high precision in the detection and classification of cataracts, showing potential advantages over traditional machine learning methods, though validation remains limited. Its performance falls within the range of reported accuracy of human experts, highlighting the high feasibility of automated diagnosis. However, validation data limitations, coupled with moderate-quality evidence and high heterogeneity, constrain the utility of DL in auxiliary diagnosis. The model's sensitivity dropped to 87% in external validation, restricting its generalization capability, so caution should be exercised in broad clinical implementation.

(J Med Internet Res 2026;28:e78869) doi:10.2196/78869

KEYWORDS

image-based; deep learning; meta-analysis; cataract; systematic review

Introduction

The lens is a biconvex, transparent structure in the anterior segment of the eye, which focuses light to project images of objects at varying distances onto the retina. A cataract, defined as clouding of the lens, is primarily an age-related degenerative disease, and congenital and pediatric cataracts also occur [1]. Early cataracts are asymptomatic, but progressive clouding can lead to visual impairment, greatly reducing quality of life and productivity [2]. Cataracts are the major cause of visual loss

XSL•FO
RenderX

worldwide. In 2020, the estimated number of the blind aged 50 years and older was 15 million, and moderate to severe visual disorders due to cataracts affected an estimated 79 million people aged 50 years and older [2]. These figures suggest the number of patients with blindness and moderate to severe visual disorders rose by 30% and 93%, respectively, versus the year 2000 [3]. Blinding cataracts refer to cataracts causing severe visual impairment (visual acuity ≤3/60) or blindness (visual acuity <3/60) according to the WHO (World Health Organization) and *ICD-11* (*International Classification of Diseases, 11th Revision)* [4]. About 94% of blinding cataracts occur in low- and middle-income countries (LMICs), and cataract-related visual disorders are causally related to poverty in low-resource settings (LRS) [1]. As the conventional diagnosis of cataracts relies on ophthalmologists and complex equipment, it is subjective and requires health care resources, making it difficult to satisfy the need for large-scale screening. Moreover, due to the limitations of infrastructure and the lack of trained personnel in LMICs, inequalities are present in early cataract detection and severity classification.

Deep learning (DL) can process pixel-level information that cannot be recognized by the human eye to greatly contribute to the analysis of medical images, assist doctors in clinical decision-making, and enhance screening efficiency. As a branch of artificial intelligence (AI), DL models are inspired by the brain and specialized in pattern recognition [5]. Therefore, DL has been rapidly evolving with broad prospects in the field of medical image analysis involving disease detection, classification, segmentation, and image registration [6]. For example, convolutional neural networks (CNNs), as the primary technique of DL for image learning, perform excellently in image classification and feature extraction, making it a cornerstone in medical imaging [7]. Residual network (ResNet), one of the landmark architectures of CNNs, has contributed to the development of DL, especially with outstanding performance in benchmarking in image recognition and classification [8]. The superiority of the ResNet model in the medical field lies in its capability to efficiently train the deep network and raise the accuracy of image recognition [9].

The strategic importance of AI is to raise the quality of care and potentially reduce costs in high-income economies and to address critical health care issues and staffing shortages and provide access to specialized skills in LMICs [10]. Nowadays, DL is developed mostly based on data from high-income countries and regions and relies on high-resolution images and advanced electronic devices. In remote regions, however, fragmented health care systems are generally characterized by insufficient infrastructure, a shortage of professionals, and a lack of health care resources, so it is difficult to guarantee the quality of screening and diagnosis. The following problems are present in the available DL models: (1) lack of compatibility: mismatch with the low-cost equipment used in LRS [10]; (2) lack of generalization capability: significant decline in the model performance in real-world scenarios, especially under different lighting conditions or in different patient populations [11]; (3) lack of clinical validation: a recent systematic review of studies assessing the use of AI algorithms for medical image analysis found that only 6% of the included studies (n=516) conducted external validation, and the assessment of the diagnostic efficiency was lacking [12]; and (4) training data bias and lack of diversity in data (race, age, and disease subtypes) weaken the model's generalization capability [13].

In the past 5 years, breakthroughs have been made in the use of DL for ophthalmic image analysis in diabetic retinopathy (DR), retinopathy of prematurity, and glaucoma. By automating the massive processing of ophthalmic images, DL can achieve a more accurate and rapid diagnosis of cataracts, reducing the doctors' subjectivity and errors of conventional methods. Moreover, DL can fuse multiple image data (eg, slit lamp, fundus, and optical coherence tomography [OCT] images) to make a comprehensive and accurate diagnosis through multimodal image analysis. In addition, diagnostic results can be obtained quickly from the DL model with real-time diagnostic capability, greatly improving work efficiency and achieving large-scale screening and early diagnosis, especially in primary health care institutions and LRS [14]. Many DL-based diagnostic tools have been approved by the US Food and Drug Administration (FDA), but they require further evaluation and independent quality review [15]. For example, IDx-DR [16,17], the first AI system approved by the FDA, is a CNN-based system for automated screening for DR, with high sensitivity and specificity, which can contribute to the early diagnosis and lower the risk of visual loss in clinical practice, especially in LRS.

Image-based DL has exhibited greater potential for automatic cataract detection and classification using fundus and slit lamp images. However, the existing findings are still heterogeneous, and a systematic review of DL algorithms for cataract image analysis is still lacking. Therefore, this study was conducted to systematically assess the performance of different DL models in cataract detection and classification from sensitivity, specificity, and the area under the ROC curve (AUC), thereby revealing the methodological and reporting quality and contributing to the clinical translation of DL.

Methods

Registration and Study Design

This study adhered to the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-analyses of Diagnostic Test Accuracy Studies) [18], and the study protocol was registered on PROSPERO (International Prospective Register of Systematic Reviews; CRD420251030230). We acknowledge that the registration was completed on April 10, 2025, following the initial literature search conducted on April 1, 2025; thus, this is a retrospective registration.

Search Strategy, Eligibility Criteria, and Data Extraction

Based on the predefined criteria, 2 investigators (RXL and HYL) independently searched Web of Science, IEEE Xplore, Embase, PubMed, and Cochrane Library up to April 1, 2025, for studies published from 2019 to 2025.

The retrieved studies were first imported into EndNote to eliminate duplicate publications. Then the title and abstract were

XSL•FO
RenderX

read, and the full text of clearly or potentially eligible studies was examined. No restrictions were imposed on the geographical location or study setting. The following study types were excluded, including letters, non–peer-reviewed reports, narrative reviews, animal studies, and conference abstracts. The search strategy is shown in Multimedia Appendix 1, with specific search phrases, Boolean operations, and field restrictions. The following studies were excluded during full-text screening: studies failing to report key values such as true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN), making it impossible to make a contingency table; studies reporting only composite metrics like sensitivity, specificity, or AUC without providing the underlying data; and studies that reported data ambiguously or in a manner that prevented accurate construction of a 2×2 contingency table. The following is a table of inclusion and exclusion criteria (Table 1). Discrepancies were resolved through discussion with a third investigator (SL or LHL) if needed.

Table . Eligibility criteria table, which outlines the following aspects: study design, language, data extraction, disease type, and intervention.

Variable	Inclusion criteria	Exclusion criteria
Study design	Randomized controlled trials, prospective observational study, retrospective diagnostic accuracy study, cross-sectional diagnostic study	• Letters, non–peer-reviewed reports, narrative reviews, animal studies, and conference abstracts.
Language	Full-text in English	• Non-English publications
Data extraction	• Studies can report key values such as true positive (TP), false positive (FP), false negative (FN), and true negative (TN), and can create a contingency table. • Studies report on comprehensive indicators such as sensitivity, specificity, or AUC, and provide basic data; • Studies that report clear data and can accurately construct 2×2 contingency tables.	• Studies failing to report key values such as true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN), making it impossible to make a contingency table; • Studies reporting only composite metrics like sensitivity, specificity, or AUC without providing the underlying data; • Studies that reported data ambiguously or in a manner that prevented accurate construction of a 2×2 contingency table.
Disease type	Cataracts	• Noncataract
Intervention	Cataract diagnosis (detection and/or classification)	• Noncataract diagnosis

Only studies on the performance of image-based DL algorithms in cataract detection and classification were included following the eligibility criteria (Table 2).

Two investigators (RXL and HYL) independently extracted the following data using standardized data extraction tables: basic characteristics (country, publication year, study site, and type), dataset characteristics (nature, number of images, and presence or absence of external validation), and performance metrics (sensitivity and specificity). Discrepancies were settled by discussion with a third investigator. Contingency tables were used to directly extract the data on binary diagnostic accuracy, including TP, FP, TN, and FN. During data extraction, double-checking was performed, and the original authors were contacted to obtain supplementary data. These data were then used to calculate pooled sensitivity, specificity, and other metrics. If a study provided multiple contingency tables for the same or different DL algorithms, they were assumed to be independent of each other.

XSL•FO
RenderX

Table . Outlines of the study design and basic demographic characteristics, including authorship details, participant information (number of participants and mean or median age), study design type, inclusion criteria, and exclusion criteria (n=number of participants).

Studies	Participants			N	Median age (range)	Study design
	Inclusion criteria	Exclusion criteria	Labels			
Lin et al (2019) [19]	Patients aged less than 14 years, with or without eye symptoms, and with no history of eye surgery. All participants were required to undergo slit-lamp photography, and sedatives such as chloral hydrate when necessary.	Patients who already had a definitive diagnosis of cataract, other ocular abnormalities, or ocular trauma.	Noncataract or cataract	350	6.58 (6.13-7.03)	Multicenter randomized controlled trial (RCT)
Deepak and Bhat (2024) [20]	NR[a]	NR	Cataract or glaucoma or normal.	NR	NR	Cross-sectional diagnostic study
Zhao et al (2024) [21]	NR	NR	Noncataract or mild cataract or moderate cataract or severe cataract.	NR	NR	Cross-sectional diagnostic study
Zia et al (2023) [22]	NR	NR	Cataract or glaucoma or diabetic retinopathy or neutral.	NR	NR	Retrospective diagnostic accuracy study
Zhang et al (2023) [23]	NR	NR	Normal or low-grade or high-grade.	543	NR	Prospective diagnostic accuracy study
Zeboulon et al (2022) [24]	Patients of either one of the following clinical categories: clear lens or cataract. Patients with clear lens had no history of refractive surgery and had a best corrected visual acuity (BCVA) of at least 20/20. Patients with cataract had significant visual discomfort and were scheduled for surgery. All types of cataracts were included, and 4 experienced cataract and refractive surgeons of the department performed the patient inclusions (authors PZ, CP, WG, and DG).	Patients with any corneal disease.	Normal or cataract or background.	157	NR	Retrospective diagnostic accuracy study
Zhang et al (2024) [25]	NR	NR	NC[b] severity level: normal or mild or severe.	530	NR	Retrospective diagnostic accuracy study
Zhang et al (2022) [26]	NR	NR	NC severity level: normal or mild or severe.	543	NR	Retrospective diagnostic accuracy study

XSL•FO
RenderX

Studies	Participants			N	Median age (range)	Study design
	Inclusion criteria	Exclusion criteria	Labels			
Xie et al (2023) [27]	Patients with cataract whose best corrected distance visual acuity (BCDVA) was good (>0.6) within 1 month after cataract surgery, and patients without cataract without refractive media opacities. The fundus images were captured without mydriasis before surgery.	Traumatic cataracts, congenital cataracts and lens dislocation, corneal diseases, asteroid hyalosis, vitreous hemorrhage, and severe retinal and optic nerve diseases. Poor quality and unreadable images were also excluded: images out of focus; images underexposed; images overexposed; incomplete images with more than 1/3 peripheral halo.	Noncataract or mild cataracts/ or visually impaired cataracts.	5245	NR	Retrospective diagnostic accuracy study
Wu et al (2022) [28]	NR	Patients with congenital cataract, intraocular lens, aphakic eye, severe eye trauma, or corneal opacity.	Cataract or noncataract with normal-quality images or noncataract with poor-quality images.	30,668	NR	Retrospective diagnostic accuracy study
Vasan et al (2023) [29]	New patients of both the paid and free service facilities who were aged 40 years and older with the BCVA less than 20/40 in either eye. Participants were recruited immediately after the vision examination before further investigation and ophthalmologist examination.	Patients with trauma or vulnerabilities, patients who were unwilling to participate in the study or with dilated pupil.	Negative or positive or “can’t say” or “not asked.”	1407	NR	Prospective diagnostic study
Hassan et al (2024) [30]	NR	NR	Normal or cataract or glaucoma or diabetic or uveitis.	NR	NR	Retrospective diagnostic accuracy study
Ueno et al (2024) [31]	NR	NR	Normal or cataract or infectious keratitis or immunological keratitis or corneal scar or corneal deposits or bullous keratopathy or ocular surface tumor or primary angle-closure glaucoma.	NR	NR	Retrospective diagnostic accuracy study
Singh et al (2024) [32]	NR	NR	Cataract or glaucoma or diabetic retinopathy.	NR	NR	Retrospective diagnostic accuracy study

XSL•FO
RenderX

Studies	Participants			N	Median age (range)	Study design
	Inclusion criteria	Exclusion criteria	Labels			
Shafiq et al (2024) [33]	NR	NR	Glaucoma or cataracts or diabetic retinopathy or myopia or macular degeneration.	NR	NR	Retrospective diagnostic accuracy study
Santone et al (2024) [34]	NR	NR	Normal or cataract.	4785	NR	Retrospective diagnostic accuracy study
Jawad et al (2024) [35]	NR	NR	Normal or glaucoma or cataract or myopia or others.	NR	NR	Retrospective diagnostic accuracy study
Janti et al (2024) [36]	Patients who were aged 45 years and older, and participants of both genders, including males and females.	Patients found to be critically ill after the examination, and those who were not willing to participate in the study.	Cataract positive (mature or immature) or cataract negative (normal and intraocular lens).	495	61.2 (NR-NR)	Prospective, observational diagnostic accuracy study
Emir and Colak (2024) [37]	NR	NR	Healthy or diabetic retinopathy or glaucoma or cataract or age-related macular degeneration or hypertension/myopia or others.	NR	NR	Retrospective diagnostic accuracy study
Ogundokun et al (2024) [38]	NR	NR	AMD[c] or cataract or diabetes or glaucoma or hypertension myopia or normal.	NR	NR	Retrospective diagnostic accuracy study
Nguyen and Lin (2024) [39]	NR	NR	Cataract or normal.	NR	NR	Retrospective diagnostic accuracy study
Mai et al (2024) [40]	NR	NR	Control (no cataract present), without posterior polar cataract (PPC; cataract present without PPC), with PPC (cataract present with PPC).	103	NR	Retrospective diagnostic accuracy study
Raveenthini et al (2024) [41]	NR	NR	AMD or cataract or diabetic retinopathy or glaucoma or normal.	NR	NR	Retrospective diagnostic accuracy study
Rafay et al (2023) [42]	NR	NR	Cataract or diabetic retinopathy or glaucoma or normal.	NR	NR	Cross-sectional diagnostic study
Abbas et al (2023) [43]	NR	NR	Glaucoma or diabetic retinopathy or cataract or normal.	300	NR	Cross-sectional diagnostic study
Uyar et al (2024) [44]	NR	NR	Cataract or DR or glaucoma or normal.	NR	NR	Retrospective diagnostic accuracy study
Serwaa et al (2024) [45]	NR	NR	Glaucoma-positive or glaucoma-negative or cataracts-positive or cataract-negative.	NR	NR	Retrospective diagnostic accuracy study

XSL•FO
RenderX

Studies	Participants			N	Median age (range)	Study design
	Inclusion criteria	Exclusion criteria	Labels			
Zhang et al (2022) [46]	NR	NR	NC[b] severity level: normal, mild, or severe.	543	61.30 (42.65-79.95)	Retrospective diagnostic accuracy study
Glaret Subin and Muthukannan (2022) [47]	NR	NR	AMD, diabetic retinopathy, cataract, or glaucoma.	5000	NR	Retrospective diagnostic accuracy study
Xiao et al (2024) [48]	NR	NR	Cortical cataract (CC) severity level: normal, mild, or severe.	469	NR	Retrospective diagnostic accuracy study
Wang et al (2024) [49]	NR	NR	2 coarse-grained types (noncataract No C and posterior subcapsular cataract [PSC]), 7 fine-grained types (NC II, NC III, ≥ NC IV, CC I, CC II, CC III, and CC IV).	NR	NR	Retrospective diagnostic accuracy study
Kumari and Saxena (2024) [50]	NR	NR	Diseased class or normal class.	NR	NR	Retrospective diagnostic accuracy study
Devaraj et al (2024) [51]	Immature cataract, mature cataract, no cataract, and prior cataract operation with intraocular lens (IOL) inserted.	NR	Cataract or non-cataract.	7726	50 (NR-NR）	Prospective observational study
Al-Saadi et al (2024) [52]	NR	NR	Normal, early, moderate, or severe.	NR	NR	Retrospective diagnostic accuracy study
Elsawy et al (2023) [53]	Phakic eyes.	Pseudophakic and aphakic eyes.	NC, cortical cataract (cortical lens opacity), or PSC.	2573	69.84 (62.12-77.56)	Retrospective diagnostic accuracy study

XSL•FO
RenderX

Studies	Participants			N	Median age (range)	Study design
	Inclusion criteria	Exclusion criteria	Labels			
Akram and Debnath, 2020 [54]	The symptoms of selected eye diseases include several visual abnormalities in the eye region, particularly blurred, clouded, or yellowing lens, gray or white spots on the cornea, red or bloodshot eyes, yellow or greenish-yellow coatings on eyes, foamy white spots in sclera, swollen eyes, eyelid deformity such as the length of the lower eyelid being turned out from the eye, or reddish bumps on the edge of an inner eyelid depending on specific diseases, and symptoms are different for each disease.	NR	Bitot's spot of vitamin A deficiency, cataracts, conjunctivitis, corneal ulcer, ectropion, healthy, or periorbital cellulitis or trachoma.	NR	NR	Cross-sectional diagnostic study
Jiang et al (2021) [55]	NR	NR	Opacity area: Limited or Extensive	NR	1.58 (0.7-2.46)	Retrospective diagnostic accuracy study
Yadav and Yadav (2023) [56]	NR	NR	Severity: no, mild, moderate, or severe.	NR	NR	Retrospective diagnostic accuracy study
Yadav and Yadav (2023) [57]	NR	NR	Severity: no, mild, moderate, or severe.	NR	NR	Retrospective diagnostic accuracy study
Subin and Kannan (2022) [58]	NR	NR	Cataract, diabetic retinopathy, glaucoma, or normal.	NR	NR	Cross-sectional diagnostic study
Pratap and Kokil (2019) [59]	NR	NR	Normal, mild, moderate, or severe.	NR	NR	Cross-sectional diagnostic study
Luo et al (2021) [60]	NR	NR	Normal, glaucoma, cataract, or AMD.	NR	NR	Retrospective diagnostic accuracy study
Imran et al (2020) [61]	NR	NR	Normal, mild, moderate, or severe.	NR	NR	Retrospective diagnostic accuracy study
Imran et al (2021) [62]	NR	NR	Normal, mild, moderate, or severe.	NR	NR	Retrospective diagnostic accuracy study
Acar et al (2021) [63]	NR	NR	Noncataract or cataract.	NR	NR	Retrospective diagnostic accuracy study
Olaniyan et al (2024) [64]	NR	NR	Normal or cataract.	NR	NR	Retrospective diagnostic accuracy study

XSL•FO
RenderX

Studies	Participants			N	Median age (range)	Study design
	Inclusion criteria	Exclusion criteria	Labels			
Ganokratanaa et al (2023) [65]	NR	NR	Normal or cataract.	NR	NR	Retrospective diagnostic accuracy study
Gan et al (2023) [66]	NR	NR	Four stages of cataract: incipient stage, intumescent stage, mature stage, or hypermature stage.	NR	NR	Cross-sectional diagnostic study
Tham et al (2022) [67]	NR	Visual impairment caused by other pathologies, incomplete or missing data on cataract grading, or BCVA.	Normal or visually significant cataract.	13,482	NR	Retrospective diagnostic accuracy study
Siddique (2022) [68]	NR	NR	Cataract, chalazion, normal, or squint.	NR	NR	Retrospective diagnostic accuracy study
Sirajudeen et al (2022) [69]	NR	NR	Cataract or normal.	200	NR	Cross-sectional diagnostic study
Junayed et al (2021) [70]	NR	NR	Cataract or non-cataract.	NR	NR	Cross-sectional diagnostic study
Hu et al (2020) [71]	NR	NR	Pronounced cataract (2) or early cataract (1) or normal (0).	NR	NR	Retrospective diagnostic accuracy study
Hu et al (2021) [72]	NR	NR	Cataract or normal.	38	58 (NR-NR)	Retrospective diagnostic accuracy study
Lai et al (2022) [73]	NR	NR	Cataract or normal.	NR	NR	Retrospective diagnostic accuracy study
Askarian et al (2021) [74]	NR	NR	Cataract or healthy.	NR	NR	Retrospective diagnostic accuracy study

XSL•FO
RenderX

Studies	Participants			N	Median age (range)	Study design
	Inclusion criteria	Exclusion criteria	Labels			
Son et al, 2022 [75]	Patients had available anterior segment photograph data.	Patients with pathologic features of the cornea, anterior chamber, lens, or iris that interfere with the detection of lens images (eg, corneal opacity or edema, uveitis, and iris defects including aniridia, coloboma, and iridocorneal endothelial syndrome) and a medical history of previous ophthalmic surgery (eg, keratoplasty, implantable Collamer lens, and cataract surgery); patients with retinal and vitreal diseases involving visual pathways that could interfere with visual acuity and final management plan.	Cortical opacity; nuclear color; nuclear opalescence; PSC: normal or mild or moderate or severe.	NR	NR	Cross-sectional diagnostic study
Saju and Rajesh, 2022 [76]	NR	NR	Five types of cataracts: cortical or hyper mature or mature or nuclear or posterior.	NR	NR	Cross-sectional diagnostic study
Chellaswamy et al (2022) [77]	NR	NR	Cataract or diabetic retinopathy or glaucoma or normal or AMD.	NR	NR	Retrospective diagnostic accuracy study
Lu et al (2022) [78]	NR	Eyes with corneal opacity or other corneal disease that might significantly interfere with lens observation and blurred region of interest due to poor fixation or eyes with small pupils that prevented manual cataract evaluation.	Nuclear cataract or cortical cataract or posterior subcapsular cataract.	NR	NR	Retrospective diagnostic accuracy study
Al-Naji et al (2024) [79]	NR	NR	Normal or cataract or foreign body or glaucoma or subconjunctival hemorrhage or viral conjunctivitis.	645	NR	Prospective observational study
Elloumi (2022) [80]	NR	NR	Healthy or mild or moderate or severe cataract.	NR	NR	Retrospective diagnostic accuracy study

XSL•FO
RenderX

Studies	Participants			N	Median age (range)	Study design
	Inclusion criteria	Exclusion criteria	Labels			
Zannah et al (2024) [81]	NR	NR	Cataract or diabetic retinopathy or glaucoma or normal.	NR	NR	Retrospective diagnostic accuracy study

[a]NR: not reported.

[b]NC: nuclear cataract.

[c]AMD: age-related macular degeneration.

Primary and Secondary Outcomes

Primary and secondary outcomes were defined to assess the performance of DL in cataract detection and classification. The primary outcomes included sensitivity, specificity, and positive and negative likelihood ratios.

The secondary outcomes were used to assess the accuracy of DL versus machine learning (ML) algorithms in cataract diagnosis through subgroup analyses and compare DL algorithms with human experts in studies using identical datasets. The datasets in each study were also investigated for the reference standard used to determine whether transfer learning was applied, the methods for model testing and validation, and the sources and characteristics of the datasets (Table 3).

XSL•FO
RenderX

Table . Summary of indicators, algorithms, and data sources (n=number of images).

Studies	Indicator definition		Algorithm			Data source			
	Device	Exclusion of poor-quality cases	Algorithm architecture	ML[a] or DL[b]	Transfer learning applied	Source of data	Number of cases for training or test or internal or external validation	Data range	Open access data
Lin et al (2019) [19]	Slit-lamp photography	Yes	CC-Cruiser	DL	No	ZOC, located in Guangzhou in southern China. The other 4 eye clinics are affiliated with Shenzhen Eye Hospital, the Central Hospital of Wuhan, the Second Affiliated Hospital of Fujian Medical University, and Kaifeng Eye Hospital	NR[c]/350/NR/NR	August 9, 2017-May 25, 2018	NR
Deepak and Bhat (2024) [20]	Retinal fundus camera	No	Darknet-53	DL	Yes	Ocular Disease Intelligent Recognition (ODIR)	4000/1000/NR/NR	NR	ODIR
Zhao et al (2024) [21]	Slit-lamp photography	No	NCME-Net	DL	No	Shenzhen Eye Hospital and the Eye Hospital of Nanjing Medical University	553/100/139/NR	NR	NR
Zia et al (2023) [22]	Retinal fundus camera	No	Improved SqueezeNet model	DL	No	ODIR-IMAGE, Kaggle dataset	1500/400/NR/NR	NR	ODIR-IMAGE, Kaggle dataset
Zhang et al (2023) [23]	CASIA2 AS-OCT ophthalmology device (TOMEY Inc, Japan)	Yes	Ensemble Logistic Regression (EMLR) framework	ML	No	AS-OCT-NC2 dataset	7831/3611/NR/NR	NR	NR
Zeboulon et al (2022) [24]	Swept Source Optical Coherence Tomography (SS-OCT)	No	U-Net model	DL	No	Anterion (Heidelberg)	Development set/validation set: 504/1326	NR	NR
Zhang et al (2024) [25]	CASIA2 AS-OCT ophthalmology device (TOMEY Inc)	No	RCRNets	DL	No	AS-OCT-NC2 dataset	9394/3903/100/NR	NR	NR

XSL•FO
RenderX

Studies	Indicator definition		Algorithm			Data source			
	Device	Exclusion of poor-quality cases	Algorithm architecture	ML[a] or DL[b]	Transfer learning applied	Source of data	Number of cases for training or test or internal or external validation	Data range	Open access data
Zhang et al (2022) [26]	CASIA2 AS-OCT ophthalmology device (TOMEY Inc)	Yes	RIR-Net-2‐34	DL	No	The local hospital	7831/3611/NR/NR	NR	UCSD dataset, Heidelberg OCT dataset
Xie et al (2023) [27]	Retinal fundus camera	Yes	DenseNet121[d]	DL	No	Zhejiang Eye Hospital at Wenzhou (ZEHWZ)	4901/1048/1048/1338	September 2020-March 2021	NR
Wu et al (2022) [28]	Retinal fundus camera	Yes	Anti-interference model (convolutional neural network [CNN])	DL	No	The Chinese PLA (People's Liberation Army) General Hospital	14401/7765/1800/NR	September 2018-May 2021	NR
Vasan et al (2023) [29]	Smartphone camera	Yes	E-Paarvai App (CNN）	DL	No	A large eye care hospital in South India	1400/NR/2619/NR	January 2022-April 2022	NR
Hassan et al (2024) [30]	Retinal fundus camera	No	OcularNET	DL	Yes	The Kaggle machine-learning platform	4000/2200/NR/NR	NR	The Kaggle machine-learning platform.
Ueno et al (2024) [31]	Smartphone camera	Yes	YOLO V.5	DL	No	23 tertiary eye centers in Japan	5270/836/NR/NR	2019‐2020	NR
Singh et al (2024) [32]	Retinal fundus camera	No	A novel ensembled deep learning CNN model	DL	Yes	The Kaggle database	4217/NR/NR/NR	NR	The Kaggle database
Shafiq et al, (2024) [33]	OCT, retinal fundus camera	No	The Dual-Eye-FeatureNet model	DL	No	Structured analysis of the retina (STARE), DRIVE, high-resolution fundus (HRF)	NR/483/NR/NR	NR	STARE, DRIVE, HRF
Santone et al (2024) [34]	Retinal fundus camera	No	The STANDARD_CNN model	DL	No	the ODIR 5K dataset	6987/957/1627/NR	NR	the ODIR 5K dataset
Jawad et al (2024) [35]	Retinal fundus camera	Yes	Swin Transformer models (Swin-T)	DL	No	ODIR, the Retina dataset, available on Kaggle	7000/3400/NR/300	NR	ODIR, the Retina dataset, available on Kaggle

XSL•FO
RenderX

Studies	Indicator definition		Algorithm			Data source			
	Device	Exclusion of poor-quality cases	Algorithm architecture	ML[a] or DL[b]	Transfer learning applied	Source of data	Number of cases for training or test or internal or external validation	Data range	Open access data
Janti et al (2024) [36]	Smartphone camera	Yes	Smartphone-based cataract screening application	DL	No	The AIIMS (All India Institute of Medical Sciences) Bibinagar, Hyderabad, Telangana, India	NR/990/NR/NR	April 2024-July 2024	NR
Emir and Colak (2024) [37]	Retinal fundus camera	Yes	The residual neural network (ResNet) 50	DL	No	The ODIR dataset	3198/930/471/NR	NR	The ODIR dataset
Ogundokun et al (2024) [38]	Retinal fundus camera	No	MobileNet[e]V2-SVM (support vector machine)	DL	No	Ocular dataset from the Kaggle repository	1629/202/203/NR	NR	The Kaggle database
Nguyen and Lin (2024) [39]	Retinal fundus camera	No	Hybrid CNN Approach	DL	Yes	The Kaggle database	888/278/222/NR	NR	The Kaggle database
Mai et al (2024) [40]	A Zeiss OPMI Lumera T surgical microscope	No	ConvNeXt-Tiny model	DL	No	Department of Ophthalmology, Far Eastern Memorial Hospital, New Taipei, Taiwan.	NR/NR/103/NR	January 1, 2018-December 31, 2021	NR
M et al (2024) [41]	Retinal fundus camera	No	XGB classifier model	ML	No	HRF, DR HAGIS, DIARET DB0, DRISHTI, KAGGLE, E-OPTHA, RIM ONE, ORIGA, ACRIMA, DRIONS-DB, STARE, ARIA, IDRID, ICHALLENGE AMD, ODIR, RFMID, KAGGLE CATARACT, HARVARD V1, DERBI DATA, ICHALLENGE, GLAUCOMA	1044/261/2/NR/NR	NR	HRF, DR HAGIS, DIARET DB0, DRISHTI, KAGGLE, E-OPTHA, RIM ONE, ORIGA, ACRIMA, DRIONS-DB, STARE, ARIA, IDRID, ICHALLENGE AMD, ODIR, RFMID, KAGGLE CATARACT, HARVARD V1, DERBI DATA, ICHALLENGE, GLAUCOMA

XSL•FO
RenderX

Studies	Indicator definition		Algorithm			Data source			
	Device	Exclusion of poor-quality cases	Algorithm architecture	ML[a] or DL[b]	Transfer learning applied	Source of data	Number of cases for training or test or internal or external validation	Data range	Open access data
Rafay et al (2023) [42]	Retinal fundus camera	No	EfficientNet[f] B3	DL	Yes	The Kaggle database	2949/1268/NR/NR	NR	The Kaggle database
Abbas et al (2023) [43]	Retinal fundus camera	No	Deep-ocular mode	DL	Yes	The retinal fundus multi-disease image dataset (RFMiD) and ODIR	1222/521/NR/NR	NR	RFMiD and ODIR
Uyar et al (2024) [44]	Retinal fundus camera	No	ABC-based weighted ensemble model	DL	Yes	The Eye Disease Dataset (EDD), from the Kaggle	3372/426/419/NR	NR	EDD, from the Kaggle
Serwaa et al (2024) [45]	Retinal fundus camera	No	LBPSCN: Local Binary Pattern Scaled Capsule Network	DL	No	The Kaggle database	NR	NR	The Kaggle database
Zhang et al (2022) [46]	CASIA2 AS-OCT ophthalmology device (TOMEY Inc)	Yes	Clinical-awareness attention network (CCA-Net)	DL	No	A local health physical center	9619/3141/3441/NR	NR	The ACRIMA dataset, the UCSD dataset
Glaret Subin and Muthukannan (2022) [47]	Fundus camera	No	FPOA-CNN	DL	No	Various medical centers in China collected by the Shanggong Medical Technology Co., Ltd.	NR	NR	ODIR database
Xiao et al (2024) [48]	CASIA2 AS-OCT ophthalmology device (TOMEY Inc)	No	ResNet34-MSSA	DL	No	CASIA2 AS-OCT dataset	3969/1271/1305/NR	NR	The LAG dataset, the SD-OCT dataset
Wang et al, 2024 [49]	Slit-lamp photography	No	MGCNet	DL	No	The Cataract Center of Beijing Tongren Hospital (BTH; Beijing, China)	2912/970/970/NR	NR	The APTOS2019 dataset, The HAM10000 dataset
Kumari and Saxena (2024) [50]	Retinal fundus camera	No	RINet	DL	No	Multiple repositories	NR	NR	NR

XSL•FO
RenderX

Studies	Indicator definition		Algorithm			Data source			
	Device	Exclusion of poor-quality cases	Algorithm architecture	ML[a] or DL[b]	Transfer learning applied	Source of data	Number of cases for training or test or internal or external validation	Data range	Open access data
Devaraj et al (2024) [51]	Smartphone camera	No	EfficientNet-v2 Model	DL	No	The ophthalmology departments at King George's Medical University (KGMU) and Balrampur Hospital, Lucknow	1708/753/275/NR	October 29, 2022-September 23, 2023	NR
Al-Saadi et al (2024) [52]	Retinal fundus camera	No	An Automated Wavelet Scattering Network	DL	No	The ODIR database	357/155/NR/NR	NR	The ODIR database
Elsawy et al (2023) [53]	Color fundus photography (CFP)	No	Deep Opacity Net	DL	Yes	The AREDS2 dataset, the Singapore Epidemiology of Eye Diseases study (SEED)	122/75/41/73/1708	2006 - 2008	NR
Akram and Debnath (2020) [54]	Digital camera	No	A deep convolution neural network (DCNN) model	ML	No	International Center for Eye Health, clinical images for symptoms on faces from the University of Rochester, UCSD School of Medicine and VA Medical Center, the Primary Care Dermatology Society, and other different resources	140/250/NR/NR	NR	NR
Jiang et al (2021) [55]	Slit-lamp photography	No	CCNN-Ensemble	DL	Yes	Zhongshan Ophthalmic Center of Sun Yat-sen University	470 (Training and validation)/132/79	June 2015-February 2020	NR

XSL•FO
RenderX

Studies	Indicator definition		Algorithm			Data source			
	Device	Exclusion of poor-quality cases	Algorithm architecture	ML[a] or DL[b]	Transfer learning applied	Source of data	Number of cases for training or test or internal or external validation	Data range	Open access data
Yadav and Yadav (2023) [56]	Retinal fundus camera	Yes	CNN with 2D DFT	DL	No	HRF, STARE, MESSIDOR, DRIVE, DRIONS_DB, and IDRiD datasets, as well as images obtained from the internet	NR	NR	HRF, STARE, MESSIDOR, DRIVE, DRIONS_DB, and IDRiD datasets
Yadav and Yadav (2023) [57]	Retinal fundus camera	No	CNN with ensemble of SVM, NB, RF	DL	No	HRF, STARE, MESSIDOR, DRIVE, DRIONS_DB, and IDRiD databases, as well as other images collected from the internet	NR	NR	HRF, STARE, MESSIDOR, DRIVE, DRIONS_DB, and IDRiD databases
Subin and Kannan (2022) [58]	Retinal fundus camera	No	AMSO-RNN (recurrent neural network) Model	DL	No	ODIR database	224×960×NR×NR	NR	ODIR database

XSL•FO
RenderX

Studies	Indicator definition		Algorithm			Data source			
	Device	Exclusion of poor-quality cases	Algorithm architecture	ML[a] or DL[b]	Transfer learning applied	Source of data	Number of cases for training or test or internal or external validation	Data range	Open access data
Pratap and Kokil (2019) [59]	Retinal fundus camera	Yes	Pre-trained CNN	TL	Yes	HRF image database, STARE, standard diabetic retinopathy database (DIARETDB0), e-ophtha: a color fundus image database, MESSIDOR database, DRIVE database, fundus image registration (FIRE) dataset, digital retinal images for optic nerve segmentation database (DRIONS-DB), IDRiD, available datasets from Dr Hossein Rabbani, and other internet resources	400/400/NR/NR	NR	HRF image database, STARE, DIARETDB0, e-ophtha: a color fundus image database, MESSIDOR database, DRIVE database, FIRE dataset, DRIONS-DB, IDRiD, available datasets from Dr Hossein Rabbani, and other internet resources
Luo et al (2021) [60]	Retinal fundus camera	No	FCL-EfficientNet-B3	DL	No	Shanggong Medical Technology Co, Ltd. OIA-ODIR dataset	10000/274/NR/NR	NR	OIA-ODIR dataset
Imran et al (2020) [61]	Retinal fundus camera	No	The combination of DL models (AlexNet, ResNet, and VGG[g]Net) and SVM	DL	Yes	The Tongren Hospital, China	6424/1607/NR/NR	NR	NR

XSL•FO
RenderX

Studies	Indicator definition		Algorithm			Data source			
	Device	Exclusion of poor-quality cases	Algorithm architecture	ML[a] or DL[b]	Transfer learning applied	Source of data	Number of cases for training or test or internal or external validation	Data range	Open access data
Imran et al (2021) [62]	High-resolution fundus camera Canon-EOS-40D with additional settings such as 72 DPI resolution, no-flash, manual exposure, and auto-white balance	No	A novel hybrid method, namely CRNN, based on CNN and RNN	DL	Yes	The Tongren Hospital, China	6424/1606/NR/NR	NR	NR
Acar et al (2021) [63]	Retinal fundus camera	No	VGGNet	DL	Yes	The Kaggle Ocular Disease Recognition database	3891/1216/973/NR	NR	The Kaggle Ocular Disease Recognition database
Olaniyan et al (2024) [64]	Slit-lamp photography	No	Hybrid Siamese-VGG16 model	DL	No	Kaggle's public repository	NR/121/NR/NR	NR	Kaggle's public repository
Ganokratanaa et al (2023) [65]	Slit-lamp photography	No	LeNet-CNN	DL	No	NR	560/140/NR/NR	NR	NR
Gan et al (2023) [66]	Slit-lamp photography	No	Automatic segmentation DTL platform	DL	No	Department of Ophthalmology, Jiangxi Provincial People's Hospital	517/130/nr/nr	NR	NR
Tham et al (2022) [67]	Retinal fundus camera	Yes	The ResNet-50	DL	No	Singapore Eye Research Institute	8045/1692/NR/1605	NR	NR
Siddique (2022) [68]	Photos from phones and internet	No	MobileNet	DL	No	4 hospitals from Bangladesh	1762/439/NR/NR	NR	NR
Sirajudeen et al (2022) [69]	Retinal fundus camera	No	Novel Kernel-based CNN	DL	No	The Kaggle database	320/80/NR/NR	NR	The Kaggle database
Junayed et al (2021) [70]	Retinal fundus camera	No	CataractNet	DL	No	Multiple databases	904/226/NR/NR	NR	Multiple databases
Hu et al (2020) [71]	Smartphone with slit-lamp	No	UDFA (Faster-RCNN)[h]	DL	No	Marked Slit Lamp Picture Project (MSLPP)	1127/281/9483/NR	NR	MSLPP

XSL•FO
RenderX

Studies	Indicator definition		Algorithm			Data source			
	Device	Exclusion of poor-quality cases	Algorithm architecture	ML[a] or DL[b]	Transfer learning applied	Source of data	Number of cases for training or test or internal or external validation	Data range	Open access data
Hu et al (2021) [72]	The iSpector-mini mobile phone slit lamp developed by Shenyang EyeROBO Intelligent Technology Co, Ltd.	No	ACCV	DL	No	A cooperation hospital	1064/304/152/NR	NR	NR
Lai et al (2022) [73]	Digital camera	No	CNNDCI	DL	No	GitHub.com	7735/1938/9/NR	NR	GitHub.com
Askarian et al (2021) [74]	Smartphone camera	No	SVM	ML	No	NR	63/30/7/NR	NR	NR
Son et al (2022) [75]	Slit-lamp photography	No	An ensemble of 3 AI algorithms: ResNet18, WideResNet50-2, and ResNext50	DL	Yes	Local outpatient clinic	2706/792/446/NR	January 2017-December 2020	NR
Saju and Rajesh (2022) [76]	Slit-lamp photography	No	Dense CNN+EfficientNet classification model	DL	No	The DRIMDB dataset, various hospitals	NR/264/NR/NR	NR	The DRIMDB dataset
Chellaswamy et al (2022) [77]	Retinal fundus camera	No	WODCNN method	DL	No	KAGGLE, MESSIDOR, ORIGA, DRIVE, STARE datasets	1661/414/NR/NR	NR	KAGGLE, MESSIDOR, ORIGA, DRIVE, STARE datasets
Lu et al (2022) [78]	Slit-lamp photography	Yes	Faster R-CNN and ResNet	DL	No	An internal dataset from the EENT Hospital of Fudan University and an external dataset from the Pujiang Eye Study	964/214/156/NR	An internal dataset of slit lamp photographs of the anterior segment of cataract-affected eyes taken between 2018 and 2020. Another external dataset of slit lamp photographs taken between March 2018 and August 2019.	NR

XSL•FO
RenderX

Studies	Indicator definition		Algorithm			Data source			
	Device	Exclusion of poor-quality cases	Algorithm architecture	ML[a] or DL[b]	Transfer learning applied	Source of data	Number of cases for training or test or internal or external validation	Data range	Open access data
Al-Naji et al (2024) [79]	Retinal fundus camera	No	InceptionResNetV2	DL	No	The Balad Ruz General Hospital and Ibn Al-Haitham Teaching Eye Hospital	453/194/NR/NR	January 2, 2023-July 7, 2023	NR
Elloumi (2022) [80]	Retinal fundus camera	No	Ensemble Learning (InceptionV3, MobileNetV2, and NasNet-Mobile)	DL	Yes	the Kaggle platform	354/118/118/NR	NR	The "Cataract Dataset," "Ocular Disease Recognition (ODiR)"
Zannah et al (2024) [81]	Retinal fundus camera	No	BayeSVM500 model	ML	Yes	Cataract dataset, Glaucoma dataset, High-Resolution Fundus (HRF) Image Database, Kaggle, IEEE-Dataport, and Pattern Recognition Lab	4144/1037/NR/NR	NR	The Kaggle database, High-Resolution Fundus (HRF) Image Database

[a]ML: machine learning.

[b]DL: deep learning.

[c]NR: not reported.

[d]DenseNet: dense convolutional network.

[e]MobileNet: efficient convolutional neural networks for mobile vision apps.

[f]EfficientNet: rethinking model scaling for convolutional neural networks.

[g]VGG: visual geometry group.

[h]Faster-R CNN: fast region–based convolutional network.

Quality Assessment

Three investigators independently assessed the risk of bias (RoB) and applicability concerns in the clinical context in the included studies by using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) [82]. Cross-validation was implemented to enhance interrater agreement, and the methodology for consistency assessment was explicitly reported. Deeks' funnel plots were drawn to evaluate the publication bias if more than 10 studies were included. Statistical significance was set at $P<.05$. Deeks' funnel plot asymmetry test was performed using the Deeks command within the *MIDAS* package in STATA 18.0 (StataCorp LLC).

Statistical Analysis

STATA 18.0 and RevMan 5.4 (Review Manager; The Cochrane Collaboration) were used for data analyses. The summary receiver operating characteristic (SROC) curve plotting, assessment of heterogeneity, and analysis of publication bias were performed using STATA 18.0 to enhance transparency and reproducibility. The heterogeneity was assessed by the Cochran Q test and I^2 statistic. The threshold of the I^2 statistic for quantifying heterogeneity proposed by Higgins et al [83] was adopted: $I^2 \leq 25\%$: low, $I^2 \approx 50\%$: moderate, $I^2 \geq 75\%$: substantial. This strategy was specifically designed to identify whether inconsistencies in study results stemmed from random factors or reflected substantive discrepancies. When significant heterogeneity ($P<.05$ or $I^2>50\%$) was identified [84], a bivariate mixed effects model was adopted. The bivariate random-effects model was fitted using the *MIDAS* package in STATA 18.0.

XSL•FO
RenderX

This method demonstrated particular efficacy in synthesizing pooled estimates of sensitivity, specificity, and AUC across studies. Its advantage resides in the capability to systematically address metric variability while preserving the intrinsic correlation.

To evaluate the accuracy of the DL algorithm, a hierarchical SROC curve was fitted. We calculated corresponding 95% CIs for sensitivity, specificity, and AUC using the Delta method, which linearized the nonlinear relation of the log-transformed sensitivity and specificity by a first-order Taylor expansion. Then the variance-covariance matrix of the parameter estimates was propagated, whereas prediction intervals (PIs) incorporated between-study heterogeneity by modeling the covariance structure of sensitivity and specificity [84,85]. Sensitivity analyses were conducted on all DL algorithms to be evaluated, rather than only the one with the highest accuracy. A random-effects model was used to explain potential between-study variability.

Ethical Considerations

This study required no informed consent or ethical approval. Data previously collected from human subjects in ethical/institutional review board–approved studies were used. All studies included adhered to the Declaration of Helsinki.

Results

Search Results

Initially, 2235 studies were retrieved, of which 492 duplicates were excluded. Following study screening, 1680 studies were excluded from quantitative synthesis (meta-analysis), including 1617 studies involving animal research, nondisease studies, surgical technique investigations, reviews, and conference reports; 10 studies not using deep learning algorithms; 46 studies lacking sufficient data for constructing 2×2 contingency tables or reporting data in formats incompatible with pooling (eg, AUC only); 4 studies not focusing on cataract diagnostic models; and 3 studies that did not address cataract disease (Figure 1; Checklist 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of study selection. The PRISMA flow of search methodology and literature selection process is displayed.

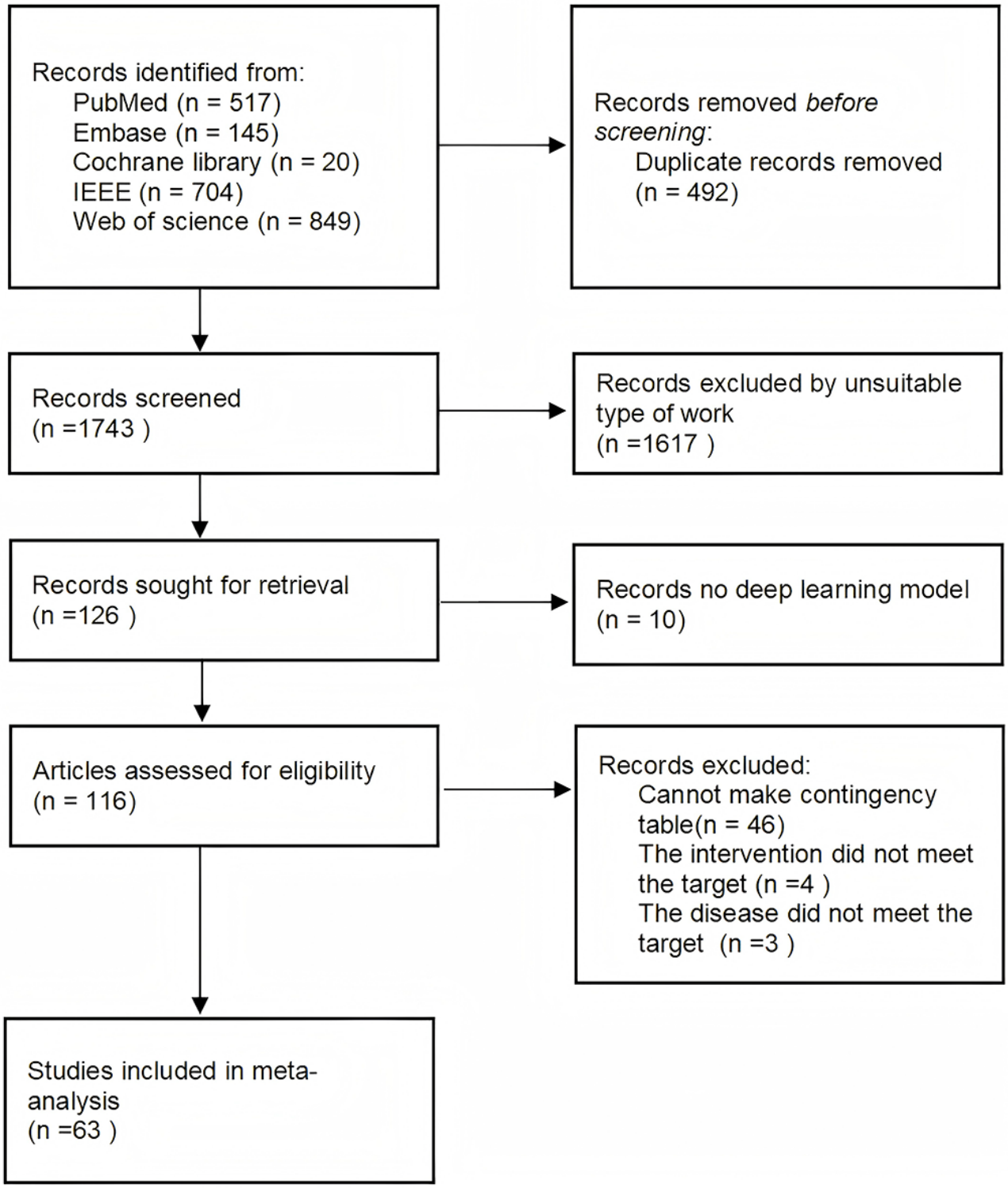

Study Characteristics

As shown in Tables 2-4, all the included studies were published in 2019 - 2024, involving 171,416 images. Retrospective data were used in 46 studies, prospective data in only 5 studies, and cross-sectional data in 12 studies. The data came from open-access sources in 31 studies. The sample size was prespecified in one study, and low-quality images were excluded in 15 studies. External validation was conducted in 12 studies, while the remainder performed internal validation. Six studies compared DL with traditional ML models using the same dataset, while another 6 compared DL models with human experts. Forty-four studies focused on cataract detection, 17 on cataract classification, and 2 on both detection and classification. Cataract detection was categorized as binary detection (presence vs absence of cataracts, n=21) and multidisease detection (n=25). Moreover, cataracts were classified into mild (n=18), moderate (n=11), and severe (n=19) types. Additionally, 49 studies did

XSL•FO
RenderX

not describe cataract classification; among the other studies, clinical subtypes included posterior subcapsular cataract (PSC; n=4), pediatric cataract (n=2), posterior polar cataract (PPC; n=1), nuclear cataract (NC; n=10), and cortical cataract (CC; n=6).

XSL•FO
RenderX

Table . Author information, reference standards, types of internal validation, and whether external validation is applied.

Studies	Reference standard	Type of internal validation	External validation
Lin et al (2019) [19]	Expert consensus, consistent label	Hold-out cross-validation	No
Deepak and Bhat (2024) [20]	Clinical diagnosis by ophthalmologists; validated by trained readers	5-fold cross-validation	No
Zhao et al (2024) [21]	AREDS No 4 guidelines	5-fold cross-validation	No
Zia et al (2023) [22]	Clinical diagnosis by ophthalmologists; validated by trained readers	5-fold cross-validation	No
Zhang et al (2023) [23]	The LOCS III[a] system	Hold-out cross-validation	No
Zeboulon et al (2022) [24]	The cataract fraction (CF)	5-fold cross-validation	Yes
Zhang et al (2024) [25]	The LOCS III system	Hold-out cross-validation	Yes
Zhang et al (2022) [26]	The LOCS III system	Hold-out cross-validation	Yes
Xie et al (2023) [27]	Cataract specialists	Hold-out cross-validation	Yes
Wu et al (2022) [28]	Cataract specialists	Hold-out cross-validation	Yes
Vasan et al (2023) [29]	Slit lamp diagnosis with dilated eyes by an ophthalmologist	Hold-out cross-validation	No
Hassan et al (2024) [30]	Clinical diagnosis by ophthalmologists; validated by trained readers	Hold-out cross-validation	No
Ueno et al (2024) [31]	Corneal specialists	Hold-out cross-validation	No
Singh et al (2024) [32]	Clinical diagnosis by ophthalmologists; validated by trained readers	Hold-out cross-validation	No
Shafiq et al (2024) [33]	Clinical diagnosis by ophthalmologists	NR[b]	No
Santone et al (2024) [34]	Clinical diagnosis by ophthalmologists (based on electronic medical records and verified by trained readers)	Hold-out cross-validation	No
Jawad et al (2024) [35]	Clinical diagnosis by ophthalmologists; validated by trained readers	Hold-out cross-validation	Yes
Janti et al (2024) [36]	The LOCS III system	NR	No
Emir and Colak (2024) [37]	Clinical diagnosis by ophthalmologists	Hold-out cross-validation	No
Ogundokun et al (2024) [38]	Clinical diagnosis by ophthalmologists; validated by trained readers	Hold-out cross-validation	No
Nguyen and Lin (2024) [39]	Clinical diagnosis by ophthalmologists; validated by trained readers	5-fold cross-validation	No
Mai et al (2024) [40]	Visual inspection of the surgical view of the cataract in the surgery video	5-fold cross-validation	No
Raveenthini et al (2024) [41]	Clinical diagnosis by ophthalmologists	ten-fold cross-validation	No
Rafay et al (2023) [42]	Clinical diagnosis by ophthalmologists; validated by trained readers	Hold-out cross-validation	No
Abbas et al (2023) [43]	Clinical diagnosis by ophthalmologists	Hold-out cross-validation	No
Uyar et al (2024) [44]	Clinical diagnosis by ophthalmologists; validated by trained readers	10-fold cross-validation	No
Serwaa et al (2024) [45]	Clinical diagnosis by ophthalmologists; validated by trained readers	NR	No
Zhang et al (2022) [46]	Experienced ophthalmologists	NR	Yes

XSL•FO
RenderX

Studies	Reference standard	Type of internal validation	External validation
Glaret Subin and Muthukannan (2022) [47]	Clinical diagnosis by ophthalmologists; validated by trained readers	10-fold cross-validation	No
Xiao et al (2024) [48]	Clinical grading by ophthalmologists	NR	No
Wang et al (2024) [49]	Clinical diagnosis by ophthalmologists	Hold-out cross-validation	No
Kumari and Saxena (2024) [50]	Clinical diagnosis by ophthalmologists	5-fold cross-validation	No
Devaraj et al (2024) [51]	Clinical diagnosis by ophthalmologists	Hold-out cross-validation	No
Al-Saadi et al (2024) [52]	Clinical diagnosis by ophthalmologists; validated by trained readers	Hold-out cross-validation	No
Elsawy et al (2023) [53]	The Wisconsin Cataract Grading System, the AREDS2 NS severity scale.	Hold-out cross-validation	Yes
Akram and Debnath (2020) [54]	Clinical grading by ophthalmologists	10-fold cross-validation	No
Jiang et al (2021) [55]	Three senior ophthalmologists	5-fold cross-validation	Yes
Yadav and Yadav (2023) [56]	A professional ophthalmologist	NR	No
Yadav and Yadav (2023) [57]	Clinical diagnosis by ophthalmologists	NR	No
Subin and Kannan (2022) [58]	Clinical diagnosis by ophthalmologists; validated by trained readers	Hold-out cross-validation	No
Pratap and Kokil (2019) [59]	Clinical diagnosis by ophthalmologists; validated by trained readers	Hold-out cross-validation	No
Luo et al (2021) [60]	Trained ophthalmologists	5-fold cross-validation	No
Imran et al (2020) [61]	Two retinal experts	5-fold cross-validation	No
Imran et al (2021) [62]	Clinical diagnosis by ophthalmologists	5-fold cross-validation	No
Acar et al (2021) [63]	Clinical diagnosis by ophthalmologists	Monte-Carlo cross-validation	No
Olaniyan et al (2024) [64]	Clinical diagnosis by ophthalmologists; validated by trained readers	Hold-out cross-validation	No
Ganokratanaa et al (2023) [65]	NR	5-fold cross-validation	No
Gan et al (2023) [66]	Experienced ophthalmologists	5-fold cross-validation	No
Tham et al (2022) [67]	Wisconsin cataract grading system or AREDS system	5-fold cross-validation	Yes
Siddique (2022) [68]	Clinical grading by ophthalmologists	5-fold cross-validation	No
Sirajudeen et al (2022) [69]	Clinical diagnosis by ophthalmologists; validated by trained readers	5-fold cross-validation	No
Junayed et al (2021) [70]	Composite reference standard (Expert clinical diagnosis or grading from original source datasets)	5-fold cross-validation	No
Hu et al (2020) [71]	Ophthalmologists with more than 5 years of clinical experience	Hold-out cross-validation	No
Hu et al (2021) [72]	The LOCS III system	Hold-out cross-validation	No
Lai et al (2022) [73]	The LOCS III system	5-fold cross-validation	Yes
Askarian et al (2021) [74]	NR	10-fold cross-validation	No
Son et al (2022) [75]	The LOCS III system	Hold-out cross-validation	No

XSL•FO
RenderX

Studies	Reference standard	Type of internal validation	External validation
Saju and Rajesh (2022) [76]	Clinical diagnosis by ophthalmologists	Hold-out cross-validation	No
Chellaswamy et al (2022) [77]	Clinical diagnosis by ophthalmologists	Hold-out cross-validation	No
Lu et al (2022) [78]	The LOCS III system	Hold-out cross-validation	Yes
Al-Naji et al (2024) [79]	Clinical grading by ophthalmologists	Hold-out cross-validation	No
Elloumi (2022) [80]	Clinical diagnosis by ophthalmologists; validated by trained readers	5-fold cross-validation	No
Zannah et al (2024) [81]	Clinical diagnosis by ophthalmologists; validated by trained readers	5-fold cross-validation	No

[a]LOCS III: Lens Opacities Classification System III.

[b]NR: not reported.

Pooled Performance of DL Algorithms

Finally, 63 studies [19-81] with sufficient data (97 contingency tables) were included for the assessment of DL performance in cataract detection and classification [86]. Hierarchical SROC curves for cataract detection (45 contingency tables) and classification (52 contingency tables) are provided in Figures 2A and 3A, respectively. The classification task involved multiclassification (eg, mild, moderate, and severe cataracts), and separate SROC curves were generated for each category. For cataract detection, the pooled sensitivity and specificity of DL were 96% (0.95 - 0.97) and 98% (0.97 - 0.99), respectively, with an AUC of 0.99 (0.98 - 1.00). For cataract classification, DL had pooled sensitivity and specificity of 94% (0.93 - 0.96) and 97% (0.96 - 0.98), respectively, with an AUC of 0.99 (0.98 - 1.00). Great heterogeneity and inconsistency were observed across cataract severity (mild: I^2=99%, moderate: I^2=96%, severe: I^2=99%; P<.001), suggesting substantial variability in diagnostic or methodological methods across studies. 25 contingency tables for mild cataracts were used in 18 studies, 16 contingency tables for moderate cataracts in 11 studies, and 28 contingency tables for severe cataracts in 19 studies, which may introduce classification imbalance, potentially influencing the performance assessment.

Figure 2. Pooled overall performance of deep learning (DL) algorithms for cataract detection. (A) Receiver operating characteristic (ROC) curves of all studies included in the meta-analysis (36 studies with 45 tables). (B) ROC curves of studies reporting the highest accuracy (36 studies with 36 tables). The cataract detection is divided into binary detection and multidisease detection. AUC: area under the ROC curve; SROC: summary receiver operating characteristic.

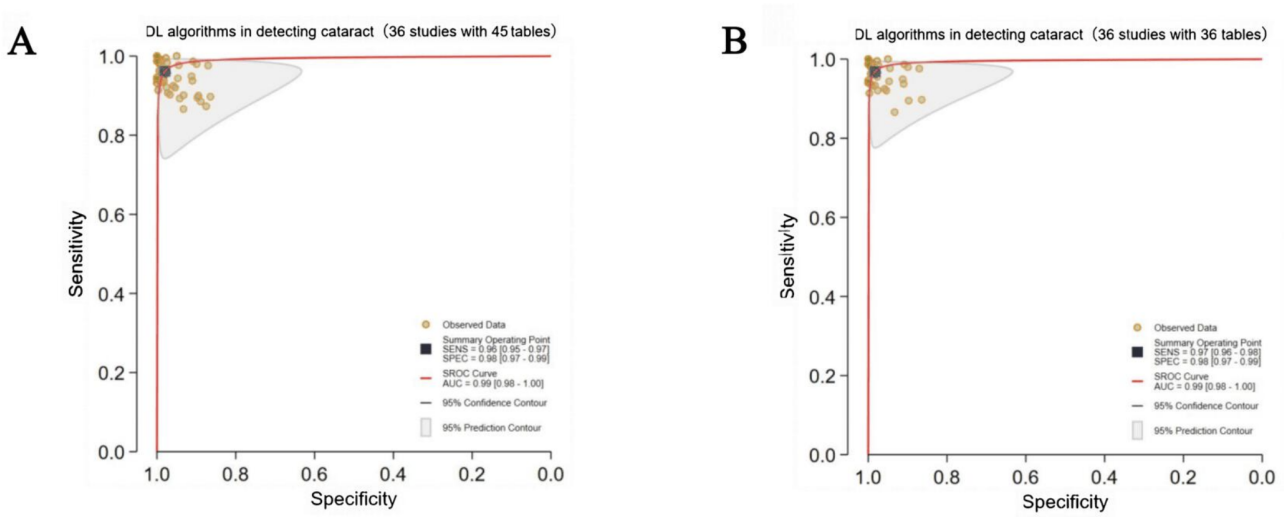

XSL•FO
RenderX

Figure 3. Pooled overall performance of deep learning (DL) algorithms for cataract classification. (A) Receiver operating characteristic (ROC) curves of all studies included in the meta-analysis (17 studies with 52 tables); (B) ROC curves of studies reporting the highest accuracy (17 studies with 17 tables). The cataract classification is divided by severity (mild, moderate, and severe) and clinical subtypes. AUC: area under the ROC curve; DL: deep learning; SROC: summary receiver operating characteristic.

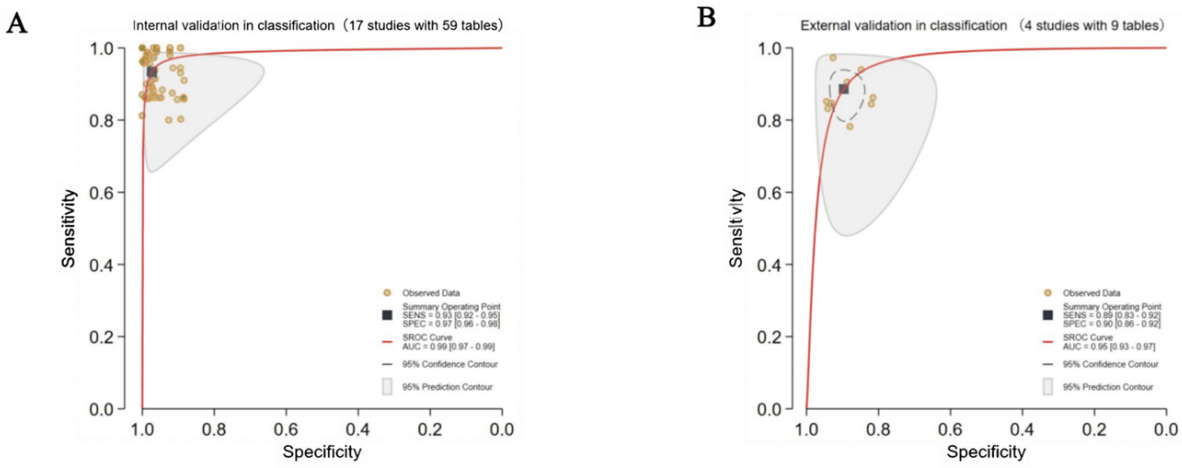

One or more DL algorithms were reported in most studies, and DL with the highest accuracy was selected, ultimately obtaining 53 contingency tables. For cataract detection, DL had pooled sensitivity and specificity of 97% (96% - 98%) and 98% (97% - 99%), respectively, with an AUC of 0.99 (0.98 - 1.00) (Figure 2B). For cataract classification, the pooled sensitivity and specificity were 95% (0.92 - 0.97) and 98% (0.96 - 0.99), respectively, with an AUC of 0.99 (0.98 - 1.00) (Figure 3B). Threshold analyses were conducted using STATA 18.0 to investigate threshold effects. For diagnostic models in primary studies that did not prespecify a threshold, sensitivity and specificity corresponding to the optimal threshold reported in the study were extracted. If a primary study reported results for multiple thresholds, we prioritized extracting and analyzing data corresponding to the threshold associated with the reported primary endpoint or optimal operating point. The spatial distribution of classification thresholds in ROC curves was analyzed; each threshold corresponded to a unique point on the ROC curve, and systematic evaluation of these points could help us quantify performance variability across thresholds and detect threshold-driven instability (defined as significant performance fluctuations within narrow threshold ranges). The SROC curve displayed no “shoulder-arm” distribution (Figures 2 and 3).

However, the model’s performance in independent external datasets (detection: sensitivity 87%, specificity 93%; classification: sensitivity 89%, specificity 90%) was lower than the overall estimates. Notably, the lower performance observed in external validation datasets compromised the generalization capability of the model, potentially attributable to domain shift, warranting caution when applied to new populations or settings.

Subgroup Analyses

Overview

Traditional ML or DL algorithms were reported in the included studies. These studies varied in primary objectives (detection, classification, or both detection and classification). Due to overlapping objectives among studies, the sum of the number of studies on traditional ML and DL algorithms did not match the total number of included studies. According to the Lens Opacities Classification System III (LOCS III) [87] and the methods of Mackenbrock et al [88] and Gali et al [89], we classified cataracts into mild, moderate, and severe types and further categorized cataracts into PSC, pediatric cataract, PPC, NC, and CC.

Detection

Algorithm Types

DL algorithms were described in 36 studies (45 contingency tables). The pooled sensitivity and specificity of DL were 96% (95% - 97%) and 98% (97% - 99%), respectively, with an AUC of 0.99 (0.98 - 1.00; Figure 2A). Additionally, traditional ML algorithms were described in 5 studies (13 contingency tables). The pooled sensitivity and specificity of ML were 90% (87% - 91%) and 94% (91% - 96%), respectively, with an AUC of 0.95 (0.93 - 0.97; Figure 4A).

XSL•FO
RenderX

Figure 4. Pooled overall performance of (no-DL) machine learning (ML) algorithms. (A) Receiver operating characteristic (ROC) curves of studies using ML algorithms for cataract detection (5 studies with 13 tables); (B) ROC curves of studies using ML algorithms to classify cataracts (6 studies with 15 tables). AUC: area under the ROC curve; ML: machine learning; SROC: summary receiver operating characteristic.

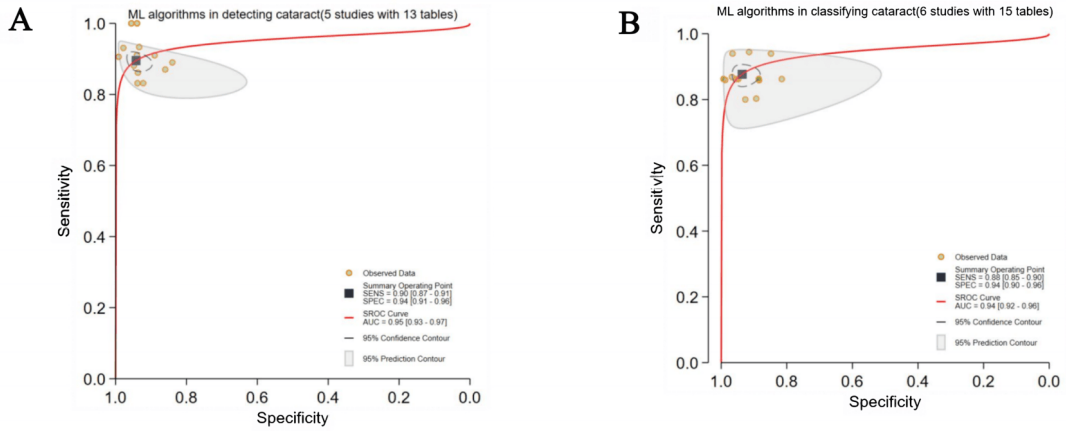

Disease Types

Around 37 studies (57 contingency tables) focused on nonspecific or general cataracts; the pooled sensitivity, specificity, and AUC were 96% (94% - 97%), 98% (97%-98%), and 0.99 (0.98 - 1.00). The Cochran Q test revealed a statistically significant result (Q=199.023; P<.001) and high inconsistency (I^2=99%; 95% CI 98 - 99), indicating significant between-study variability. For cataract detection (CC: one study with one contingency table, pediatric cataract: 2 studies with eight contingency tables, NC: 2 studies with 3 contingency tables, PSC: one study with one contingency table, PPC: one study with one contingency table), the pooled sensitivity and specificity were 93% (90% - 95%) and 96% (92% - 98%), respectively, and the AUC was 0.97 (0.95 - 0.98) in 4 studies (13 contingency tables). Due to the extremely limited sample sizes across clinical subtypes, with most subtypes reported by only one study, the meta-analysis results were less robust. Consequently, only a pooled heterogeneity assessment could be conducted. The Cochran Q test revealed a statistically significant result (Q=9.167; P=.005) and high inconsistency (I^2=78%; 95% CI 53 - 100), indicating significant between-study variability (Figures 5A and 5B).

Figure 5. Pooled performance based on disease types in cataract detection. (A) Receiver operating characteristic (ROC) curves of studies detecting unclassified cataracts (37 studies with 57 tables); (B) ROC curves of studies detecting subtypes of cataract (4 studies with 13 tables). AUC: area under the ROC curve; SROC: summary receiver operating characteristic.

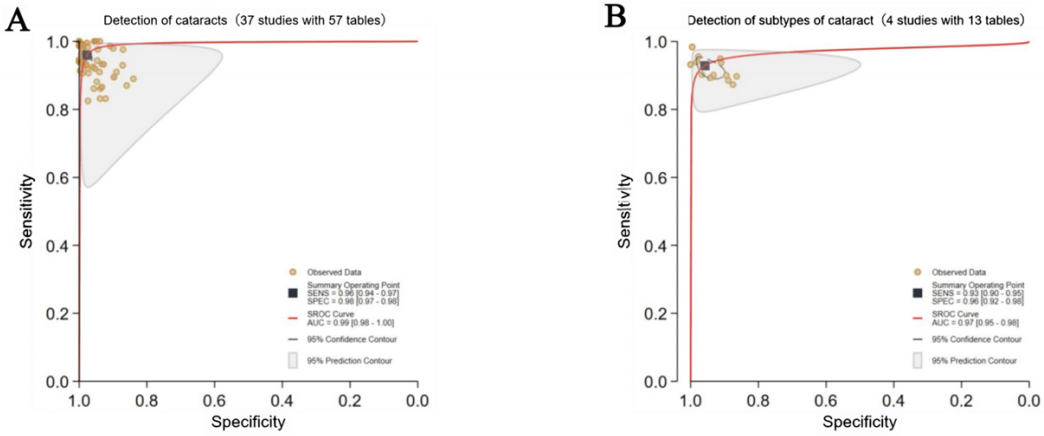

XSL•FO
RenderX

Validation Types

Great heterogeneity and inconsistency were observed in the validation methods for cataract detection (I^2=99% for both internal and external validation; P<.001), suggesting substantial variability in diagnostic or methodological methods across studies. The Deeks' funnel plots revealed no great publication bias in the internal validation (P=.68) or the external validation (P=.23). The internal validation was conducted in 38 studies (54 contingency tables), which showed a pooled sensitivity of 96% (95% - 97%) and a pooled specificity of 98% (97%-98%), with an AUC of 0.99 (0.98 - 1.00). The external validation was performed in only 8 studies (15 contingency tables), which showed a pooled sensitivity of 87% (81% - 92%) and a pooled specificity of 93% (86% - 96%), respectively, with an AUC of 0.95 (0.93 - 0.97; Figures 6A and 6B).

Figure 6. Pooled performance based on validation types in cataract detection. (A) Receiver operating characteristic (ROC) curves of studies with internal validations (38 studies with 54 tables); (B) ROC curves of studies with external validations (8 studies with 15 tables). AUC: area under the ROC curve; SROC: summary receiver operating characteristic.

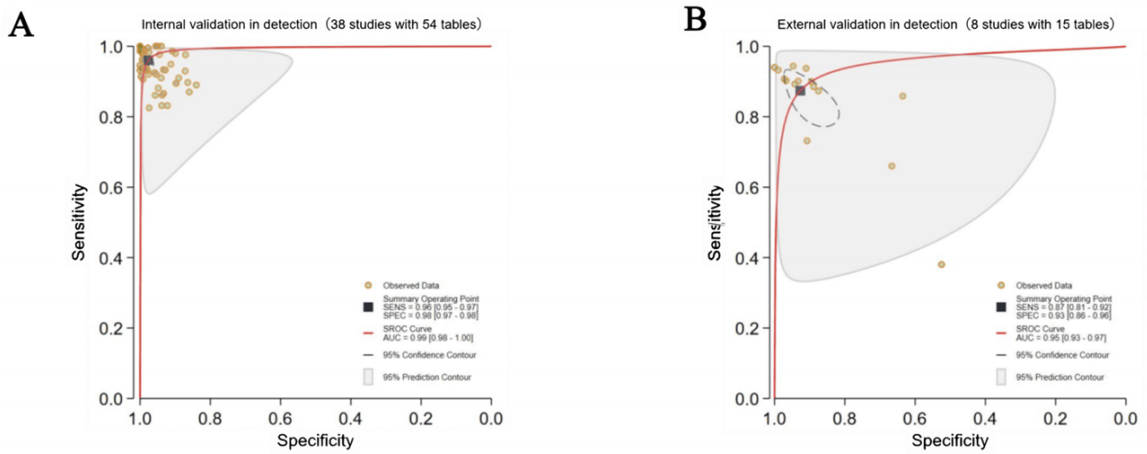

Classification

Algorithm Types

DL algorithms were described in 17 studies (52 contingency tables); the pooled sensitivity and specificity of DL were 94% (93% - 96%) and 97% (96% - 98%), respectively, with an AUC of 0.99 (0.98 - 1.00; Figure 3A). Additionally, traditional ML algorithms were described in 6 studies (15 contingency tables); the pooled sensitivity and specificity of ML were 88% (85% - 90%) and 94% (90% - 96%), respectively, with an AUC of 0.94 (0.92 - 0.96; Figure 4B).

Disease Types

For mild cataracts, 18 studies (25 contingency tables) had a pooled sensitivity of 92% (89 - 94%) and a pooled specificity of 96% (94% - 97%), with an AUC of 0.98 (0.96 - 0.99). For moderate cataracts, 11 studies (16 contingency tables) found comparable performance: SE=94% (90% - 96%), specificity=97% (95% - 98%), and AUC=0.99 (0.97 - 0.99). For severe cataracts, 19 studies (28 contingency tables) had a sensitivity of 93% (90% - 95%) and a specificity of 98% (96% - 99%), with an AUC of 0.98 (0.97 - 0.99). Great heterogeneity and inconsistency were observed across cataract severity (mild: I^2=99%, moderate: I^2=96%, severe: I^2=99%; P<.001), suggesting substantial variability in diagnostic or methodological methods across studies. Furthermore, eight studies with 28 contingency tables focused on nonspecific or general cataracts, which had a pooled sensitivity of 95% (92% - 96%) and a pooled specificity of 98% (97% - 99%), with an AUC of 0.99 (0.98 - 1.00). Great heterogeneity and inconsistency were observed across cataract severity (nonspecific or general cataract: I^2=96%, NC: 100%; P<.001), suggesting substantial variability in diagnostic or methodological methods across studies. In the NC subgroup (9 studies with 27 contingency tables), extremely high heterogeneity was observed (I^2=100%). Consequently, to maintain statistical consistency, we did not calculate a pooled estimate for this subgroup. Instead, the descriptive analysis showed a sensitivity of 89% - 93% and a specificity of 93% - 97% across these studies. Four studies with 15 contingency tables considered other clinical subtypes (PSC: one study with 3 contingency tables; CC: 4 studies with 12 contingency tables), with pooled sensitivity, specificity, and AUC of 91% (84% - 95%), 96% (94% - 97%), and 0.98 (0.96 - 0.99). Due to the extremely limited sample sizes of other clinical subtypes, the meta-analysis results were less robust. Consequently, only a pooled heterogeneity assessment could be conducted. The Cochran Q test revealed a statistically significant result (Q=120.355; P<.001) and high inconsistency (I^2=98%; 95% CI 97 - 99), indicating significant between-study variability (Figures 7A-7E).

XSL•FO
RenderX

Figure 7. Pooled performance based on disease types in cataract classification. (A) receiver operating characteristic (ROC) curves of studies on classification of mild cataract (18 studies with 25 tables); (B) ROC curves of studies on classification of moderate cataract (11 studies with 16 tables); (C) ROC curves of studies on classification of severe cataract (19 studies with 28 tables); (D) ROC curves of studies on classification of cataracts (8 studies with 28 tables); (E) ROC curves of studies on classification of other subtypes of cataract (4 studies with 15 tables). AUC: area under the ROC curve; SROC: summary receiver operating characteristic.

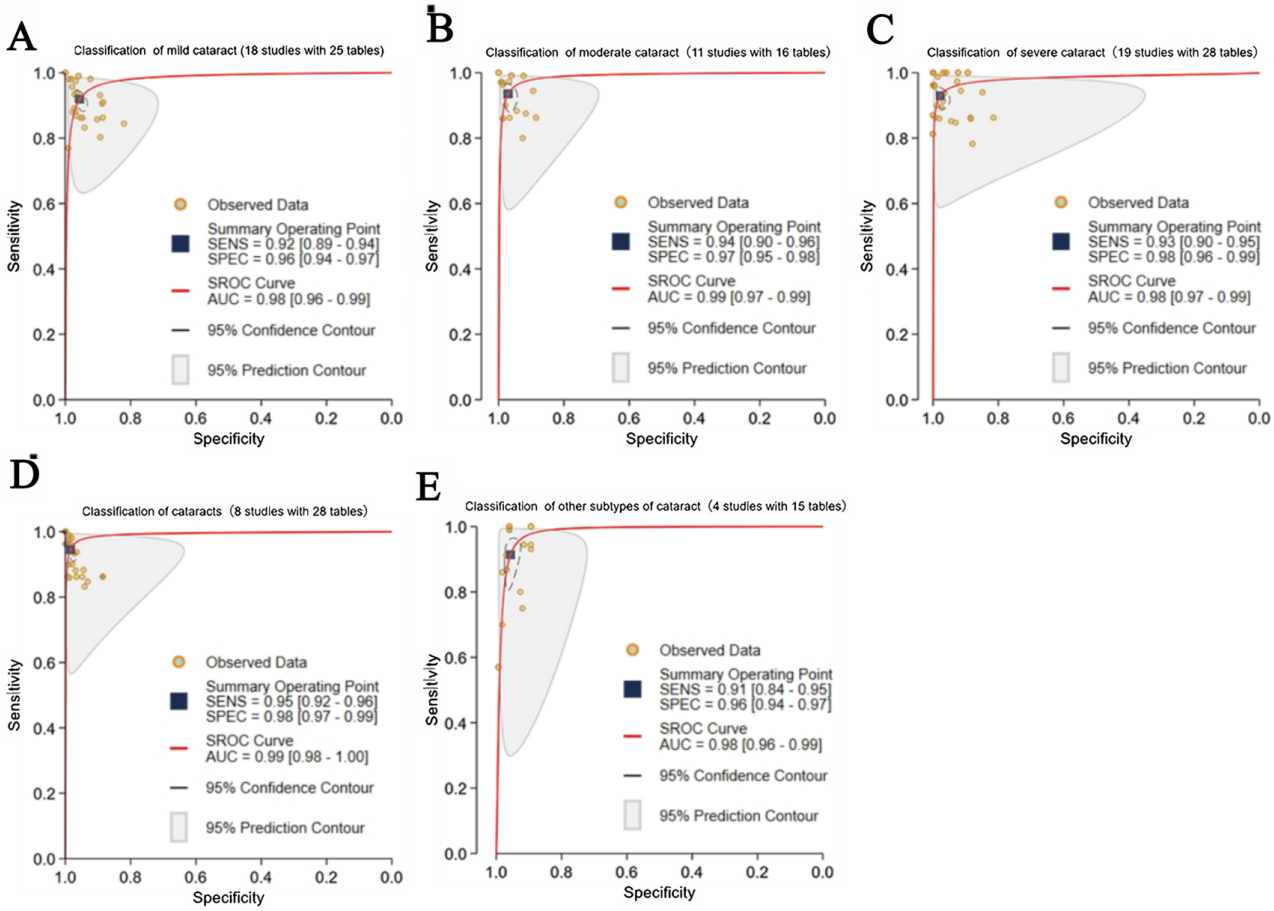

Validation Types

The internal validation was adopted in 17 studies (59 contingency tables), which showed a pooled sensitivity of 93% (92%-95%) and a pooled specificity of 97% (96% - 98%), with an AUC of 0.99 (0.97 - 0.99). The external validation was adopted in only 4 studies (9 contingency tables), which showed a pooled sensitivity of 89% (83% - 92%) and a pooled specificity of 90% (86% - 92%), with an AUC of 0.95 (0.93 - 0.97; Figure 8).

Figure 8. Pooled performance based on validation types in cataract classification. (A) Receiver operating characteristic (ROC) curves of studies with internal validations (17 studies with 59 tables); (B) ROC curves of studies with external validations (4 studies with 9 tables). AUC: area under the ROC curve; SROC: summary receiver operating characteristic.

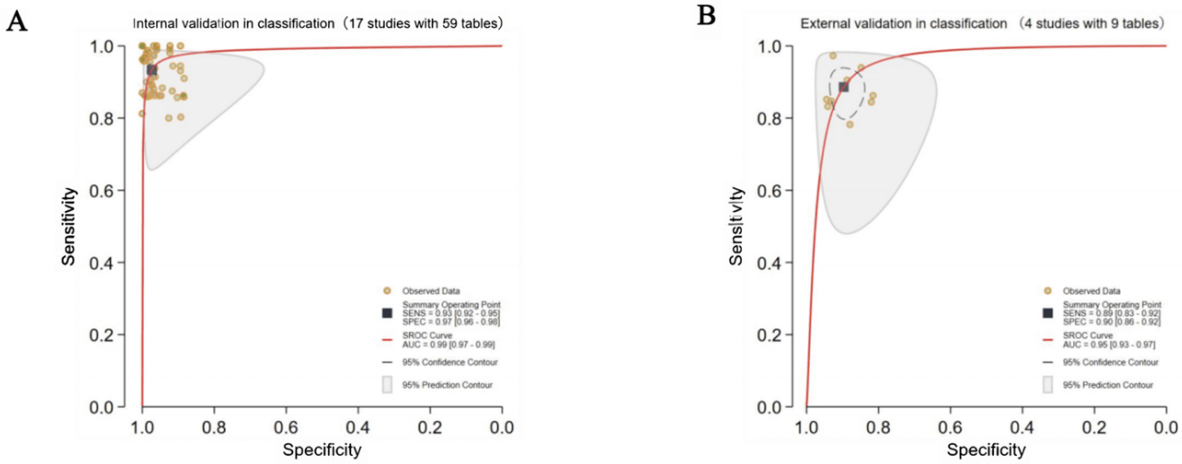

XSL•FO
RenderX

Comparison Between DL and ML Algorithms

Detection

Two studies compared DL (8 contingency tables) and ML algorithms (7 contingency tables) using the same dataset. The pooled sensitivity was 94% (92% - 96%) for DL and 91% (85% - 95%) for ML algorithms. The pooled specificity was 99% (97% - 100%) for DL and 90% (86% - 93%) for ML algorithms. The AUC was 0.97 (0.95 - 0.98) for DL and 0.96 (0.94 - 0.97) for ML algorithms. For DL algorithms, the Cochran Q test revealed a statistically significant result (Q=9.675; *P*=.004) and high inconsistency (I^2=79%; 95% CI 55 - 100). For ML algorithms, the Cochran Q test also revealed a statistically significant result (Q=5.853; *P*=0.03) and high inconsistency (I^2=66%; 95% CI 23 - 100). Therefore, both DL and ML exhibited significant heterogeneity. However, due to overlapping CIs and no direct between-group comparisons, the statistical significance of the between-group heterogeneity could not be conclusively established based on these findings. The comparison between DL and ML algorithms for cataract detection was constrained at the time of this writing by minimal validation evidence, with only 2 studies available for direct benchmarking. As a result, it severely limited statistical power, compromised generalization capability, and inflated performance estimates for DL models (Figures 9A and 9B).

Figure 9. Pooled subgroup performance of deep learning (DL) versus machine learning (ML) algorithms. AUC: area under the ROC curve; SROC: summary receiver operating characteristic.

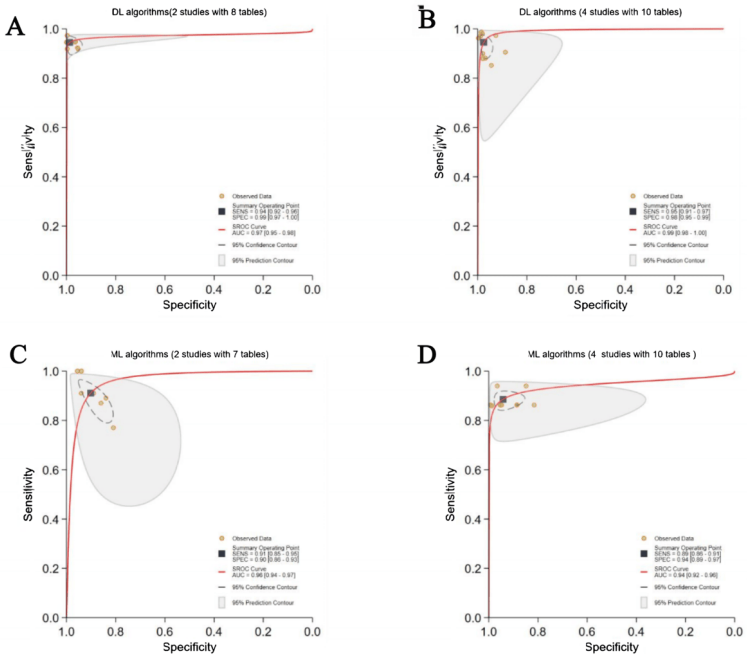

Classification

Four studies compared DL (10 contingency tables) and ML algorithms (10 contingency tables) using the same dataset. The pooled sensitivity was 95% (91% - 97%) for DL and 89% (86% - 91%) for ML algorithms. The pooled specificity was 98% (95% - 99%) for DL and 94% (89% - 97%) for ML algorithms. The AUC was 0.99 (0.98 - 1.00) for DL and 0.94 (0.92 - 0.96) for ML algorithms (Figures 9C and 9D).

DL Algorithms Versus Human Experts

Direct comparisons between DL algorithms (7 contingency tables) and human experts (10 contingency tables) were performed across 7 studies using the same datasets. For DL algorithms, quantitative pooling was not conducted due to extreme heterogeneity (Q=933.852; *P*<.001; I^2=100%). Instead, a descriptive analysis revealed highly variable performance of DL algorithms, with sensitivity estimates ranging from 72% to 93% and specificity ranging from 64% to 99%. In contrast, human experts demonstrated moderate heterogeneity (Q=5.811; *P*=0.03; I^2=66%), allowing for a pooled analysis. The pooled sensitivity and pooled specificity for human experts were 93% (95% CI 77% - 98%) and 95% (95% CI 79% - 99%), respectively, with an AUC of 0.98 (95% CI 0.97 - 0.99; Figure 10). These results highlight that DL algorithms exhibited significantly higher heterogeneity than human experts. The Deeks' funnel plot indicated no significant publication bias for human experts (*P*=.16), but potential borderline publication bias for DL algorithms (*P*=.05). The extreme heterogeneity (I^2=100%) observed in the DL algorithms underscores that their efficacy is highly dependent on specific study conditions and architectures, precluding a uniform performance metric.

XSL•FO
RenderX

Figure 10. Pooled performance of human experts. AUC: area under the ROC curve; SROC: summary receiver operating characteristic.

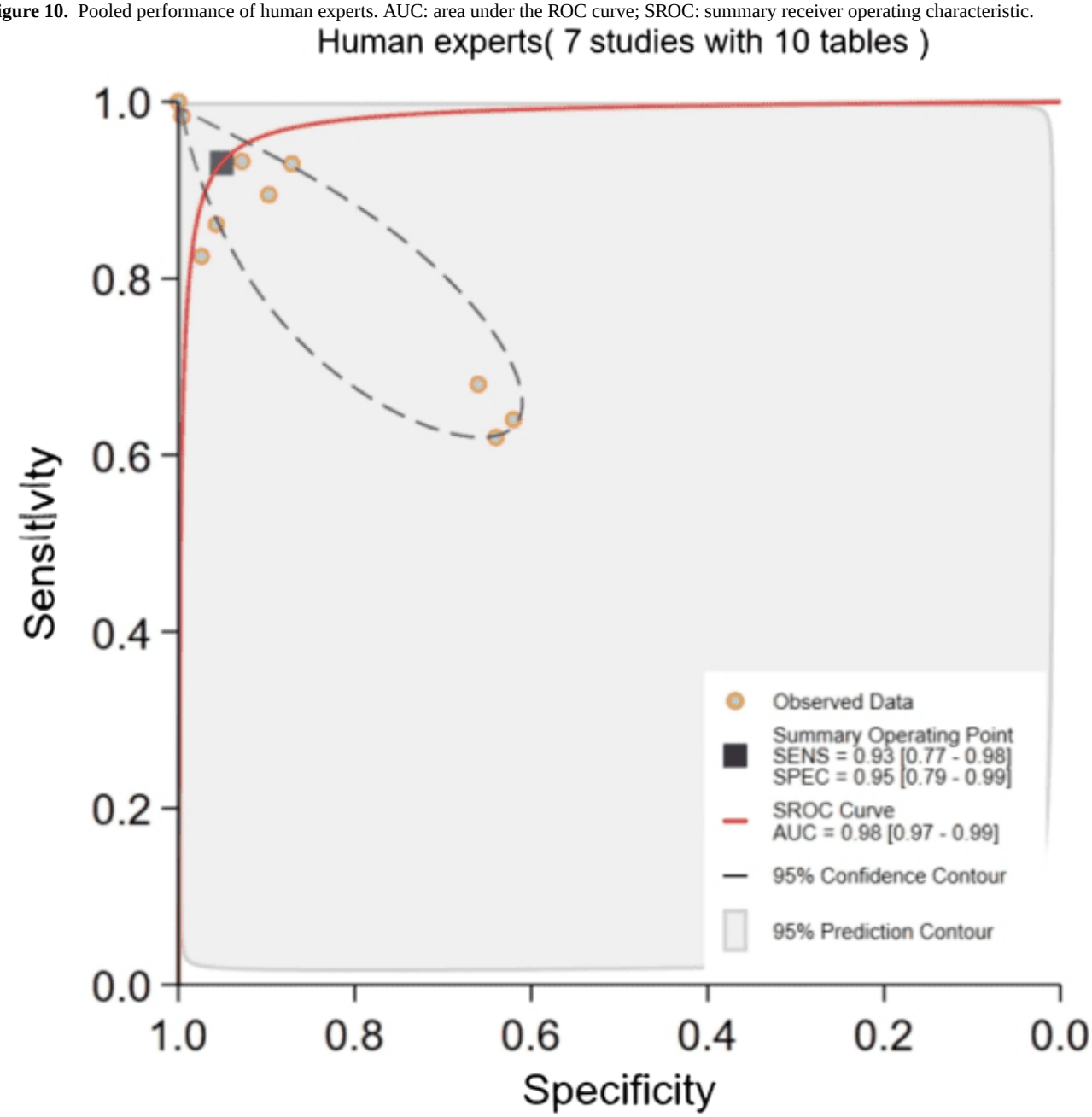

Publication Bias and Heterogeneity

The Deeks' funnel plot revealed no significant publication bias for cataract detection (*P*=.48). However, potential borderline publication bias was detected for cataract classification (*P*=.05). However, the wide distribution of studies near the regression line in the plot should be further considered (Figures 11A and 11B).

XSL•FO
RenderX

Figure 11. Summary estimate of pooled performance using funnel plots. (A) Deeks' funnel plot asymmetry test of studies detecting cataracts; (B) Deeks' funnel plot asymmetry test of studies classifying cataracts.

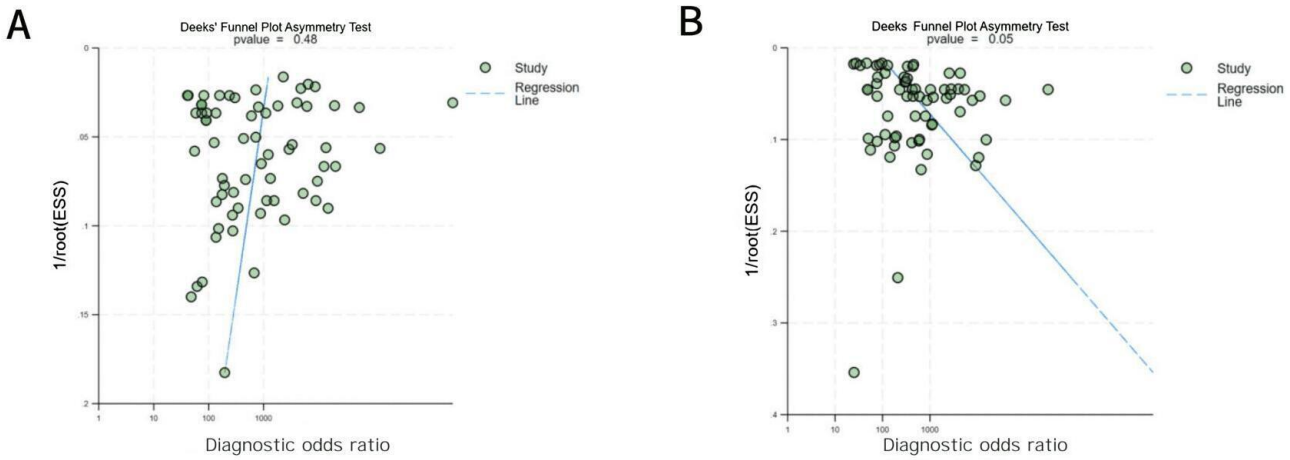

The included studies showed substantial heterogeneity. The high I^2 values (sensitivity: 95.00% and specificity: 97.11% for cataract detection; sensitivity: 95.94% and specificity: 98.55% for classification) suggested considerable between-study variability (P<.001), which was further explored by subgroup analyses (Figures 12 and 13).

Figure 12. Summary estimate of pooled performance using forest plots. The forest plot of studies for the cataract detection (41 studies).

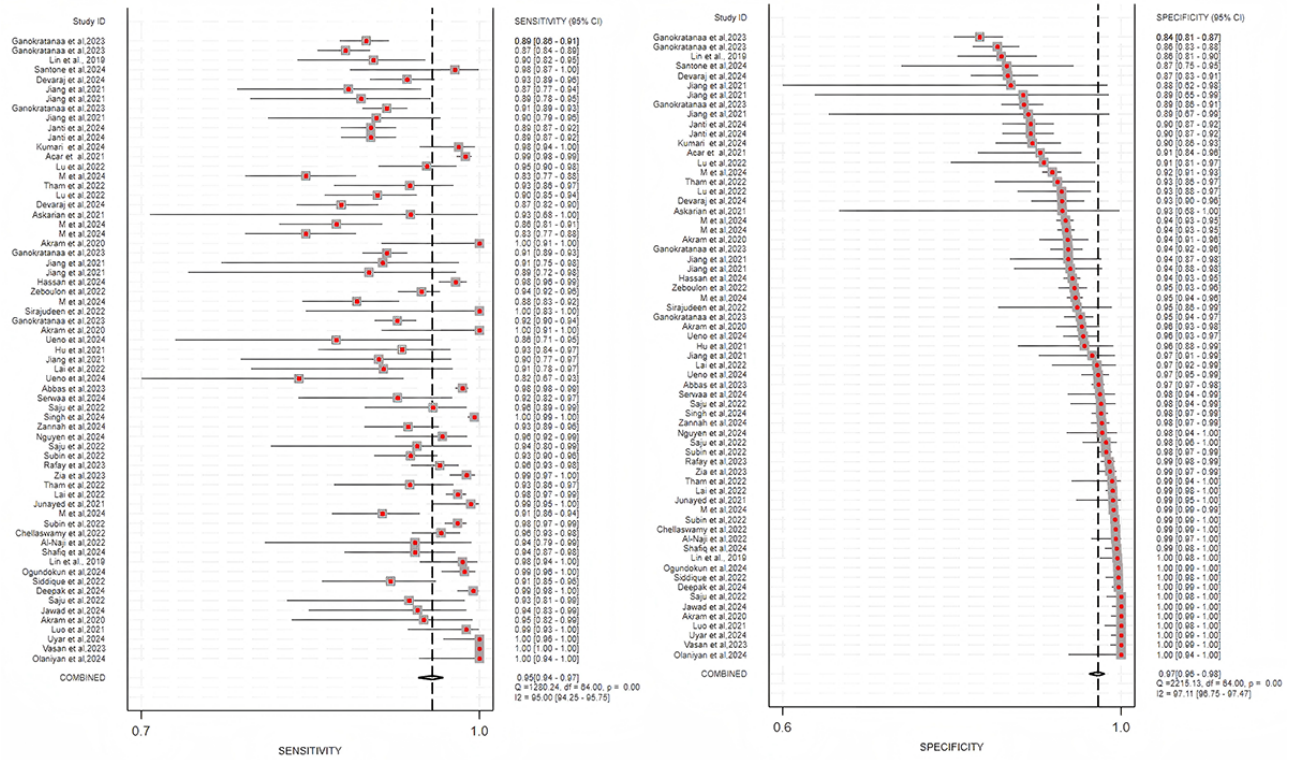

XSL•FO
RenderX

Figure 13. Summary estimate of pooled performance using forest plots. The forest plot of studies for the cataract classification (19 studies).

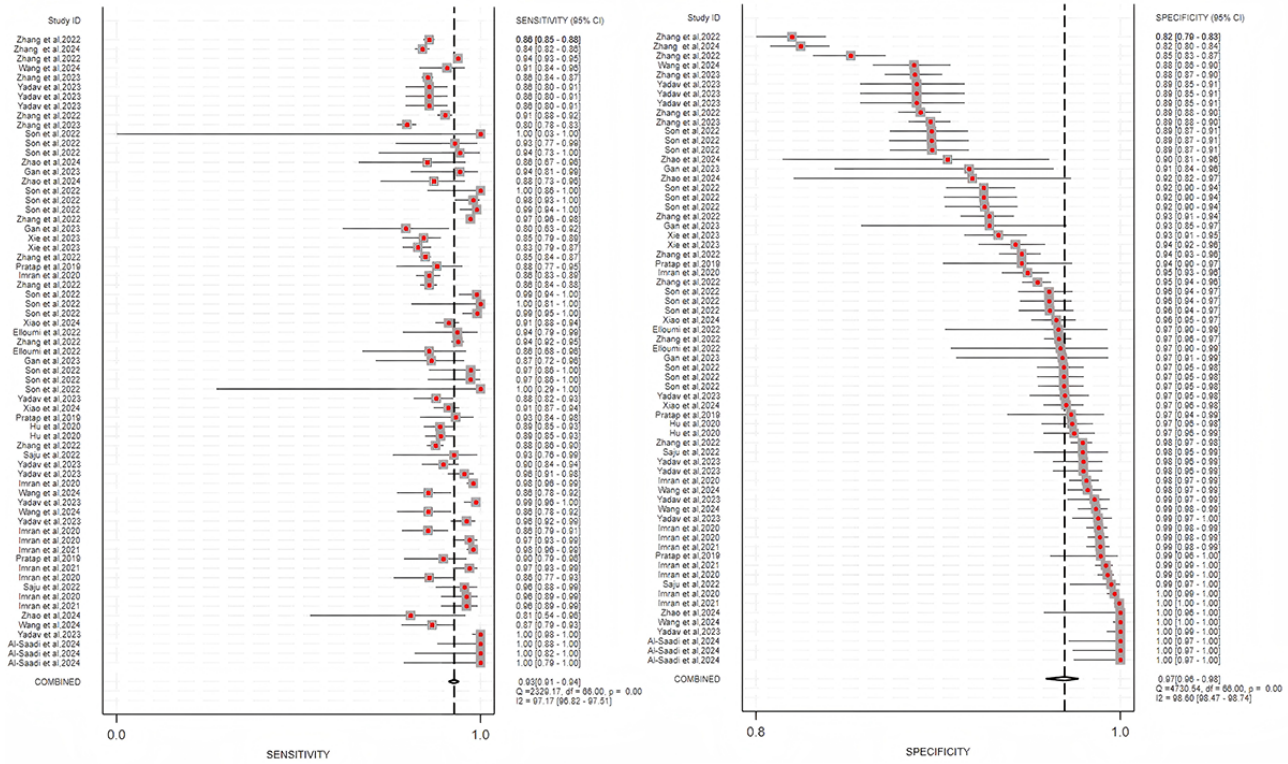

RoB

The quality of included studies was assessed using QUADAS-2, and a summary of findings is displayed in Figure S1 in Multimedia Appendix 1 (summary of methodological quality of 63 studies included). A detailed assessment for each item of RoB and applicability concern is also provided in Figure S2 in Multimedia Appendix 1 (diagrams of methodological quality of 63 studies included). High RoB or selection bias was detected in 34 studies due to a lack of randomization or eligibility criteria in the patient selection domain; RoB was high or unclear in 44 studies due to no predefined threshold in the index test domain.

Due to inconsistencies in reference standard (no reporting of whether blinding was implemented and the presence or absence of a predefined threshold), RoB was high or unclear in 2 studies. RoB was high or unclear in 34 studies in the domain of flow and timing due to no mention of whether the same gold standard was used or the presence or absence of an appropriate time gap.

High or unclear applicability in the domain of patient selection was detected in one study and unclear applicability in the domain of index test in 5 studies; 2 studies showed applicability concerns in the domain of reference standard.

Discussion

This meta-analysis included 63 studies and rigorously assessed the study quality using the QUADAS-2 to summarize the evidence to date on the performance of image-based DL in cataract diagnosis. The results revealed that DL algorithms might offer higher accuracy than traditional ML algorithms and fall within the range of reported accuracy of human experts in the detection of cataracts, demonstrating potential as tools for automated diagnosis. However, given the moderate quality and high heterogeneity of the current evidence base, these DL algorithms are considered primarily as adjuncts to cataract diagnosis.

DL has made tremendous progress in automated image analysis [6]. In clinical practice, a severe imbalance between supply and demand is present in ophthalmologic diagnosis. As AI advances, DL is expected to raise diagnostic efficiency and thus help alleviate health care resource inequality.

Four relevant systematic reviews and meta-analyses were identified. (1) Cheung et al [90] found high sensitivity, specificity, and reproducibility and great heterogeneity of the ML model for cataract diagnosis in children and adults, but the strength of evidence was limited since only 11 studies with 13 contingency tables were included. (2) Liu et al [91] found comparable performance of DL and experts in medical imaging, but only 2 out of 18 ophthalmology studies involved cataracts, necessitating in-depth research on DL in cataract diagnosis. (3) Aggarwal et al [92] included 82 ophthalmology studies, which did not involve cataracts, and verified that DL possesses good sensitivity, specificity, and AUC for the feature identification of other eye diseases, but the algorithmic evaluation criteria remain to be standardized due to high heterogeneity. (4) Islam et al [93] demonstrated that DL can be applied to retinal vessel segmentation, so it can be popularized in LMICs. Therefore, methodological limitations should be critically assessed during the clinical translation of DL to further improve its reliability.

This meta-analysis systematically evaluated the effectiveness of DL versus traditional ML in cataract detection and classification to provide a basis for clinical decision-making. For cataract detection, DL had a pooled sensitivity of 96% (95% CI 95% - 97%) and a pooled specificity of 98% (97% - 99%), with an AUC of 0.99 (0.98 - 1.00); traditional ML had a pooled

XSL•FO
RenderX

sensitivity of 90% (87% - 91%) and a pooled specificity of 94% (91% - 96%), with an AUC of 0.95 (0.93 - 0.97). For cataract classification, the pooled sensitivity and specificity of DL were 94% (93%-96%) and 97% (96% - 98%), with an AUC of 0.99 (0.98 - 1.00); the pooled sensitivity and specificity of traditional ML were 88% (85% - 90%) and 94% (90% - 96%), with an AUC of 0.94 (0.92 - 0.96). Available evidence suggests that DL models exhibit high sensitivity and specificity in automated cataract diagnosis and that medical image-based DL still demonstrates superior robustness to ML despite the RoB in most studies. However, our conclusions were made partly based on studies of low quality due to a lack of external validation and nonstandardized reporting of performance metrics, which may overestimate algorithmic accuracy. The use of overlapping public repositories introduced potential clustering bias. To ensure robust generalizability, future research must prioritize validation on independent, multicenter, and nonpublic datasets. Furthermore, the lack of a sensitivity analysis excluding high-RoB studies implied that the high overall performance of DL models could be partially influenced by methodologically weaker studies, warranting validation in future high-quality trials. Finally, due to the limited number of eligible studies, we pooled smartphone-based data from varying modalities, including diffuse photography and slit-lamp adapters. While this aligned with our assessment of general mobile health accessibility, it introduced potential optical heterogeneity. Consequently, statistical quantification of performance differences between optical sectioning and diffuse lighting was not feasible, warranting future separate investigations.

We identified the potential of DL for clinical use from the included studies, but the number of studies directly comparing the performance of DL and ML in cataract diagnosis was limited, with only 2 in detection and 4 in classification. Traditional ML (eg, logistic regression, random forest, and support vector machine) is dependent on manual feature engineering and classifiers, whose performance is limited by feature quality and domain knowledge [94]. By contrast, DL (eg, CNN and transformer-based vision models) has higher accuracy in image recognition by automatic feature extraction but requires higher data volume and stronger arithmetic support [95].

The number of included studies was limited, so image segmentation performance was not meta-analyzed. Image segmentation is essentially a pixel-level prediction task, which relies on the ability to characterize discriminative features. However, it is often difficult for traditional ML to effectively capture such features. In contrast, DL demonstrates significantly better performance by automated learning of complex patterns in highly heterogeneous data [6]. The property of DL models is such that they contain millions of parameters that need to be optimized by data-driven training, while traditional ML models display higher stability in small-sample multiclassification [95]. These findings suggest that traditional ML models may have the potential to raise the accuracy in cataract classification, while DL is more advantageous in reducing misdiagnosis rates. A combination of the 2 may facilitate the clinical translation of AI in the future.

Despite the available evidence on the clinical translational potential of DL, only 7 studies compared the efficacy of DL and human experts in cataract detection, and no studies compared their performance in cataract classification. Notably, significant interrater variability was present in expert performance due to heterogeneity in the cumulative clinical experience and health care resource allocation, highlighting the need for more comprehensive comparative studies. These studies generally report positive results of DL, but an optimism bias in algorithm performance may be produced due to insufficient sample sizes and methodological heterogeneity. Therefore, it is urgently needed to adopt a standardized study design and transparent reporting norms for improving the quality of evidence and clinical translational value of DL. DL has been successfully applied to retinal fundus image analysis [14] for automated diagnosis of DR [96] and glaucoma [97]. Considering the technical feasibility of DL in ophthalmic disease screening, it is recommended that DL serve as an adjunct to clinical diagnosis to optimize the treatment process by human-machine collaboration. Additionally, no great publication bias was detected, but the findings need to be interpreted with caution. It is recommended that a clinician's diagnostic efficiency be used as a core assessment metric and that a real-world validation be incorporated in subsequent studies.

In the field of AI methodology, several standardized guidelines have been recently issued [98,99]. However, no unified consensus has been reached on AI for cataract diagnosis. At the time of this writing, computer-aided diagnostic techniques mainly use a combination of medical image processing and AI, but research focuses mostly on eye diseases of the posterior segment (eg, DR, and age-related macular degeneration) and less on the anterior segment (eg, cataracts). Nowadays, cataract diagnosis relies on slit-lamp microscopy of lens morphology, and cataract classification is based on the LOCS III [87], the Oxford Clinical Cataract Classification and Grading System [100], and the American Cooperative Cataract Research Group method [101]. DL has made a preliminary breakthrough in automated cataract classification, such as the turbidity-density-location assessment system developed by Lin et al [19] based on anterior segment slit lamp images and the DL classification model developed by Zhou et al [102] using retinal fundus images, but a standardized severity scale adapted to DL is urgently needed for its clinical use. Therefore, it is recommended that future studies standardize the criteria for performance assessment of DL algorithms, with a focus on improving the transparency of methodology reporting and repeatability validation process.

Data scarcity and lack of generalization capability are key scientific challenges for DL. In this systematic review, retrospective designs were adopted in most of the included studies, while prospective designs and multicenter clinical trials accounted for less than 5%, and their annotation criteria were not optimized for the need of DL, restricting methodological rigor. We should recognize that most of the studies used double-blind, annotated, and quality-controlled image data for model training, which effectively improved diagnostic accuracy and reduced the RoB while balancing data size and quality. The included studies generally used data enhancement techniques,

XSL•FO
RenderX

indicating the lack of high-quality annotated datasets and prospective validation studies.

The acquisition of representative data for clinical validation remains the major bottleneck. The model's clinical generalization capability is severely restricted due to the long time consumption of pixel-level annotation of fundus images, domain shift resulting from cross-device and cross-race variations, and a lack of fine-grained pathology characterization. In the future, research should focus on multimodal data fusion for modeling (eg, OCT plus fundus images), deployment of edge computing architectures, and embedding of causal inference modules, thereby establishing intelligent diagnosis and treatment systems with both clinical credibility and engineering practicality.

In this study, great methodological heterogeneity was found in the reporting of DL performance metrics, including incomplete reporting of sensitivity or specificity, a lack of 2×2 contingency tables, and general overreliance on aggregated metrics such as AUC-ROC and F_1-score. Of particular note is that high AUC values (>0.90) may mask the risk of misjudgment of key positive events (eg, progressive cataract) in clinical scenarios with a severe imbalance in category distribution. Based on these findings, it is recommended that the confusion matrix serve as a core reporting metric. The above-mentioned problems can be gradually settled by high-quality studies.

Another key obstacle is no consensus on the interpretability of DL decision-making mechanisms. It is difficult to intuitively understand the decision-making logic of DL models due to their complex network structure and nonlinear feature extraction, which, as a "black-box" characteristic, has been deemed a challenge for clinical use. Recently, researchers have gradually revealed the intrinsic mechanism of DL models by gradient-weighted class activation mapping, adversarial testing, and causal inference [95]. For example, Chang et al [103] used adversarial samples to analyze the decision-making basis of DL models in glaucoma detection. Araújo et al [104] located the key region of DR in fundus images by multiple-instance learning. Explainable AI (XAI) is breaking through the limitations of traditional AI by synchronizing decision results with attributional explanations [105]. Abràmoff et al [106] systematically reviewed the interpretability framework for DL models in the medical field, laying a theoretical foundation for their clinical translation. Future research needs to further explore the innovative methods of XAI in medical image analysis to enhance clinical credibility.

This study showed that DL was applied primarily to the screening of eye diseases such as DR for which mature diagnostic guidelines have been established. Notably, differences in health care resource allocation should be considered when popularizing DL. In subsequent studies, the clinical effect and health economic benefits should be assessed across different DL algorithms, and the "black-box" problem of DL algorithms should be solved using interpretability methods to enhance clinical acceptance.

Acknowledgments

This work was supported by the Regional Science Fund Project of the National Natural Science Foundation of China (grant No. 82360203). The sponsors or funding organizations had no role in the design or conduct of this research.

Funding

The authors declared no financial support was received for this work.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Writing - original draft preparation: RL, HL
Writing - review and editing: RL, HL, DT
Conceptualization: RL, HL, SL, LL, DT, CL
Methodology: RL, HL, SL, LL, DT, CL
Formal analysis and investigation: RL, HL, SL, LL
Funding acquisition: DT
Resources: DT, CL
Supervision: DT, CL
And all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Methodological quality summary and risk of bias graph of the studies included in the meta-analysis.

XSL•FO
RenderX

[DOC File, 1647 KB - jmir_v28i1e78869_app1.doc]

Checklist 1
PRISMA checklist.
[PDF File, 183 KB - jmir_v28i1e78869_app2.pdf]

References

1. Cicinelli MV, Buchan JC, Nicholson M, Varadaraj V, Khanna RC. Cataracts. Lancet 2023 Feb 4;401(10374):377-389. [doi: 10.1016/S0140-6736(22)01839-6] [Medline: 36565712]
2. Steinmetz JD, Bourne RRA, Briant PS, et al. Causes of blindness and VISION impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study. Lancet Glob Health 2021 Feb;9(2):e144-e160. [doi: 10.1016/S2214-109X(20)30489-7]
3. Bourne R, Steinmetz JD, Flaxman S, et al. Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study. Lancet Glob Health 2021 Feb;9(2):e130-e143. [doi: 10.1016/S2214-109X(20)30425-3]
4. Dawson-Squibb JJ, Davids EL, Viljoen M, Rice K, Stein DJ. The WHO international classification of diseases 11th revision (ICD-11). In: Matson JL, editor. Handbook of Clinical Child Psychology: Integrating Theory and Research into Practice: Springer International Publishing; 2023:53-78. [doi: 10.1007/978-3-031-24926-6_4]
5. Ming S, Yao X, Guo X, et al. Performance of ChatGPT in ophthalmic registration and clinical diagnosis: cross-sectional study. J Med Internet Res 2024 Nov 14;26:e60226. [doi: 10.2196/60226] [Medline: 39541581]
6. Chen X, Wang X, Zhang K, et al. Recent advances and clinical applications of deep learning in medical image analysis. Med Image Anal 2022 Jul;79:102444. [doi: 10.1016/j.media.2022.102444] [Medline: 35472844]
7. Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK. Medical image analysis using convolutional neural networks: a review. J Med Syst 2018 Oct 8;42(11):226. [doi: 10.1007/s10916-018-1088-1] [Medline: 30298337]
8. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jun 27-30, 2016. [doi: 10.1109/CVPR.2016.90]
9. Hasanah SA, Pravitasari AA, Abdullah AS, Yulita IN, Asnawi MH. A deep learning review of ResNet Architecture for lung disease identification in CXR image. Appl Sci (Basel) 2023;13(24):13111. [doi: 10.3390/app132413111]
10. Williams D, Hornung H, Nadimpalli A, Peery A. Deep learning and its application for healthcare delivery in low and middle income countries. Front Artif Intell 2021;4:553987. [doi: 10.3389/frai.2021.553987] [Medline: 33997772]
11. Zhou K, Liu Z, Qiao Y, Xiang T, Loy CC. Domain generalization: a survey. IEEE Trans Pattern Anal Mach Intell 2023 Apr;45(4):4396-4415. [doi: 10.1109/TPAMI.2022.3195549] [Medline: 35914036]
12. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J Radiol 2019 Mar;20(3):405-410. [doi: 10.3348/kjr.2019.0025] [Medline: 30799571]
13. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019 Oct 29;17(1):195. [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
14. Li T, Bo W, Hu C, et al. Applications of deep learning in fundus images: a review. Med Image Anal 2021 Apr;69:101971. [doi: 10.1016/j.media.2021.101971] [Medline: 33524824]
15. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 2020;3:118. [doi: 10.1038/s41746-020-00324-0] [Medline: 32984550]
16. Khan Z, Gaidhane AM, Singh M, et al. Diagnostic accuracy of IDX-DR for detecting diabetic retinopathy: a systematic review and meta-analysis. Am J Ophthalmol 2025 May;273:192-204. [doi: 10.1016/j.ajo.2025.02.022] [Medline: 39986640]
17. Verbraak FD, Abramoff MD, Bausch GCF, et al. Diagnostic accuracy of a device for the automated detection of diabetic retinopathy in a primary care setting. Diabetes Care 2019 Apr;42(4):651-656. [doi: 10.2337/dc18-0148] [Medline: 30765436]
18. McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: the PRISMA-DTA statement. JAMA 2018 Jan 23;319(4):388-396. [doi: 10.1001/jama.2017.19163] [Medline: 29362800]
19. Lin H, Li R, Liu Z, et al. Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: a multicentre randomized controlled trial. EClinicalMedicine 2019 Mar;9:52-59. [doi: 10.1016/j.eclinm.2019.03.001] [Medline: 31143882]
20. Deepak GD, Bhat SK. Deep learning-based CNN for multiclassification of ocular diseases using transfer learning. Comput Methods Biomech Biomed Eng Imaging Vis 2024 Dec 31;12(1). [doi: 10.1080/21681163.2024.2335959]
21. Zhao J, Wan C, Li J, Zhang Z, Yang W, Li K. NCME-Net: nuclear cataract mask encoder network for intelligent grading using self-supervised learning from anterior segment photographs. Heliyon 2024 Jul;10(14):e34726. [doi: 10.1016/j.heliyon.2024.e34726]
22. Zia A, Mahum R, Ahmad N, Awais M, Alshamrani AM. Eye diseases detection using deep learning with BAM attention module. Multimed Tools Appl 2024;83(20):59061-59084. [doi: 10.1007/s11042-023-17839-9]

XSL•FO
RenderX

23. Zhang X, Xiao Z, Wu X, et al. Nuclear cataract classification in anterior segment OCT based on clinical global–local features. Complex Intell Syst 2023 Apr;9(2):1479-1493. [doi: 10.1007/s40747-022-00869-5]
24. Zéboulon P, Panthier C, Rouger H, Bijon J, Ghazal W, Gatinel D. Development and validation of a pixel wise deep learning model to detect cataract on swept-source optical coherence tomography images. J Optom 2022;15 Suppl 1(Suppl 1):S43-S49. [doi: 10.1016/j.optom.2022.08.003] [Medline: 36229338]
25. Zhang X, Xiao Z, Yang B, Wu X, Higashita R, Liu J. Regional context-based recalibration network for cataract recognition in AS-OCT. Pattern Recognit DAGM 2024 Mar;147:110069. [doi: 10.1016/j.patcog.2023.110069]
26. Zhang X, Xiao Z, Fu H, et al. Attention to region: region-based integration-and-recalibration networks for nuclear cataract classification using AS-OCT images. Med Image Anal 2022 Aug;80:102499. [doi: 10.1016/j.media.2022.102499] [Medline: 35704990]
27. Xie H, Li Z, Wu C, et al. Deep learning for detecting visually impaired cataracts using fundus images. Front Cell Dev Biol 2023;11:1197239. [doi: 10.3389/fcell.2023.1197239] [Medline: 37576595]
28. Wu X, Xu D, Ma T, et al. Artificial intelligence model for antiinterference cataract automatic diagnosis: a diagnostic accuracy study. Front Cell Dev Biol 2022;10:906042. [doi: 10.3389/fcell.2022.906042] [Medline: 35938155]
29. Vasan CS, Gupta S, Shekhar M, et al. Accuracy of an artificial intelligence-based mobile application for detecting cataracts: results from a field study. Indian J Ophthalmol 2023 Aug;71(8):2984-2989. [doi: 10.4103/IJO.IJO_3372_22] [Medline: 37530269]
30. Ul Hassan M, Al-Awady AA, Ahmed N, et al. A transfer learning enabled approach for ocular disease detection and classification. Health Inf Sci Syst 2024 Dec;12(1):36. [doi: 10.1007/s13755-024-00293-8] [Medline: 38868156]
31. Ueno Y, Oda M, Yamaguchi T, et al. Deep learning model for extensive smartphone-based diagnosis and triage of cataracts and multiple corneal diseases. Br J Ophthalmol 2024 Sep 20;108(10):1406-1413. [doi: 10.1136/bjo-2023-324488] [Medline: 38242700]
32. Singh S, Banoub R, Sanghvi HA, et al. An artificial intelligence driven approach for classification of ophthalmic images using convolutional neural network: an experimental study. Curr Med Imaging 2024;20:e15734056286918. [doi: 10.2174/0115734056286918240419100058] [Medline: 38721793]
33. Shafiq M, Fan Q, Alghamedy FH, Obidallah WJ. DualEye-FeatureNet: a dual-stream feature transfer framework for multi-modal ophthalmic image classification. IEEE Access 2024;12:143985-144008. [doi: 10.1109/ACCESS.2024.3469244]
34. Santone A, Cesarelli M, Colasuonno E, Bevilacqua V, Mercaldo F. A method for ocular disease diagnosis through visual prediction explainability. Electronics (Basel) 2024 Jul;13(14):2706. [doi: 10.3390/electronics13142706]
35. Jawad MA, Khursheed F, Nawaz S, Mir AH. Towards improved fundus disease detection using Swin Transformers. Multimed Tools Appl 2024;83(32):78125-78159. [doi: 10.1007/s11042-024-18627-9]
36. Janti SS, Saluja R, Tiwari N, et al. Evaluation of the clinical impact of a smartphone application for cataract detection. Cureus 2024 Oct;16(10):e71467. [doi: 10.7759/cureus.71467] [Medline: 39539903]
37. Emir B, Colak E. Performance analysis of pretrained convolutional neural network models for ophthalmological disease classification. Arq Bras Oftalmol 2023;87(5):e20220124. [doi: 10.5935/0004-2749.2022-0124] [Medline: 39298728]
38. Ogundokun RO, Awotunde JB, Akande HB, Lee CC, Imoize AL. Deep transfer learning models for mobile-based ocular disorder identification on retinal images. CMC 2024;80(1):139-161. [doi: 10.32604/cmc.2024.052153]
39. Nguyen VV, Lin CL. Enhancing cataract detection through hybrid CNN approach and image quadration: a solution for precise diagnosis and improved patient care. Electronics (Basel) 2024;13(12):2344. [doi: 10.3390/electronics13122344]
40. Mai ELC, Chen BH, Su TY. Innovative utilization of ultra-wide field fundus images and deep learning algorithms for screening high-risk posterior polar cataract. J Cataract Refract Surg 2024 Jun 1;50(6):618-623. [doi: 10.1097/j.jcrs.0000000000001419] [Medline: 38350234]
41. Raveenthini M, Lavanya R, Benitez R. Interpretable diagnostic system for multiocular diseases based on hybrid meta-heuristic feature selection. Comput Biol Med 2025 Jan;184:109486. [doi: 10.1016/j.compbiomed.2024.109486] [Medline: 39615233]
42. Rafay A, Asghar Z, Manzoor H, Hussain W. EyeCNN: exploring the potential of convolutional neural networks for identification of multiple eye diseases through retinal imagery. Int Ophthalmol 2023 Oct;43(10):3569-3586. [doi: 10.1007/s10792-023-02764-5] [Medline: 37291412]
43. Abbas Q, Albathan M, Altameem A, Almakki RS, Hussain A. Deep-ocular: improved transfer learning architecture using self-attention and dense layers for recognition of ocular diseases. Diagnostics (Basel) 2023 Oct 10;13(20):3165. [doi: 10.3390/diagnostics13203165] [Medline: 37891986]
44. Uyar K, Yurdakul M, Taşdemir Ş. Abc-based weighted voting deep ensemble learning model for multiple eye disease detection. Biomed Signal Process Control 2024 Oct;96:106617. [doi: 10.1016/j.bspc.2024.106617]
45. Serwaa M, Mensah PK, Adekoya AF, Ayidzoe MA. LBPSCN: local binary pattern scaled capsule network for the recognition of ocular diseases. Int J Adv Comput Sci Appl 2024 Jan;15(6). [doi: 10.14569/IJACSA.2024.01506155]
46. Zhang X, Xiao Z, Hu L, et al. CCA-Net: clinical-awareness attention network for nuclear cataract classification in AS-OCT. Knowl Based Syst 2022 Aug;250:109109. [doi: 10.1016/j.knosys.2022.109109]
47. Glaret Subin P, Muthukannan P. Optimized convolution neural network based multiple eye disease detection. Comput Biol Med 2022 Jul;146:105648. [doi: 10.1016/j.compbiomed.2022.105648] [Medline: 35751184]

XSL•FO
RenderX

48. Xiao Z, Zhang X, Zheng B, Guo Y, Higashita R, Liu J. Multi-style spatial attention module for cortical cataract classification in AS-OCT image with supervised contrastive learning. Comput Methods Programs Biomed 2024 Feb;244:107958. [doi: 10.1016/j.cmpb.2023.107958] [Medline: 38070390]
49. Wang Q, Sun H, Feng Y, Dong Z, Bai C. MGCNet: multi-granularity cataract classification using denoising diffusion probabilistic model. Displays 2024 Jul;83:102716. [doi: 10.1016/j.displa.2024.102716]
50. Kumari P, Saxena P. Cataract detection and visualization based on multi-scale deep features by RINet tuned with cyclic learning rate hyperparameter. Biomed Signal Process Control 2024 Jan;87:105452. [doi: 10.1016/j.bspc.2023.105452]
51. Devaraj M, Namasivayam V, Srichandan SS, et al. Development and testing of artificial intelligence-based mobile application to achieve cataract backlog-free status in Uttar Pradesh, India. Asia Pac J Ophthalmol (Phila) 2024;13(5):100094. [doi: 10.1016/j.apjo.2024.100094] [Medline: 39187013]
52. Al-Saadi EH, Khdiar AN, Al-Saadi LH. An automated wavelet scattering network classification using three stages of cataract disease. Baghdad SciJ 2024 Feb;21(9):3044 [FREE Full text] [doi: 10.21123/bsj.2024.8995]
53. Elsawy A, Keenan TDL, Chen Q, et al. A deep network DeepOpacityNet for detection of cataracts from color fundus photographs. Commun Med (Lond) 2023 Dec 16;3(1):184. [doi: 10.1038/s43856-023-00410-w] [Medline: 38104223]
54. Akram A, Debnath R. An automated eye disease recognition system from visual content of facial imagesusing machine learning techniques. Turk J Elec Eng Comp Sci 2020 Apr;28(2):917-932. [doi: 10.3906/elk-1905-42]
55. Jiang J, Wang L, Fu H, et al. Automatic classification of heterogeneous slit-illumination images using an ensemble of cost-sensitive convolutional neural networks. Ann Transl Med 2021 Apr;9(7):550. [doi: 10.21037/atm-20-6635] [Medline: 33987248]
56. Yadav S, Yadav J. Automatic cataract severity detection and grading using deep learning. J Sens 2023 Jan;2023(1) [FREE Full text] [doi: 10.1155/2023/2973836]
57. Yadav S, Yadav J. Enhancing cataract detection precision: a deep learning approach. Trait Signal 2023 Aug 31;40(4):1413-1424 [FREE Full text] [doi: 10.18280/ts.400410]
58. Subin PG, Kannan PM. Multiple eye disease detection using hybrid adaptive mutation swarm optimization and RNN. Int J Adv Comput Sci Appl 2022;13(9). [doi: 10.14569/IJACSA.2022.0130946]
59. Pratap T, Kokil P. Computer-aided diagnosis of cataract using deep transfer learning. Biomed Signal Process Control 2019 Aug;53:101533. [doi: 10.1016/j.bspc.2019.04.010]
60. Luo X, Li J, Chen M, Yang X, Li X. Ophthalmic disease detection via deep learning with a novel mixture loss function. IEEE J Biomed Health Inform 2021 Sep;25(9):3332-3339. [doi: 10.1109/JBHI.2021.3083605] [Medline: 34033552]
61. Imran A, Li J, Pei Y, Akhtar F, Yang JJ, Dang Y. Automated identification of cataract severity using retinal fundus images. Comput Methods Biomech Biomed Eng Imaging Vis 2020 Nov 1;8(6):691-698. [doi: 10.1080/21681163.2020.1806733]
62. Imran A, Li J, Pei Y, Akhtar F, Mahmood T, Zhang L. Fundus image-based cataract classification using a hybrid convolutional and recurrent neural network. Vis Comput 2021 Aug;37(8):2407-2417. [doi: 10.1007/s00371-020-01994-3]
63. Acar E, Türk Ö, Ertuğrul Ö, Aldemi ̇r E. Employing deep learning architectures for image-based automatic cataract diagnosis. Turk J Elec Eng Comp Sci 2021 Sep;29(SI-1):2649-2662. [doi: 10.3906/elk-2103-77]
64. Olaniyan J, Olaniyan D, Obagbuwa IC, Esiefarienrhe BM, Odighi M. Transformative transparent hybrid deep learning framework for accurate cataract detection. Appl Sci (Basel) 2024;14(21):10041. [doi: 10.3390/app142110041]
65. Ganokratanaa T, Ketcham M, Pramkeaw P. Advancements in cataract detection: the systematic development of LeNet-convolutional neural network models. J Imaging 2023 Sep 26;9(10):197. [doi: 10.3390/jimaging9100197] [Medline: 37888304]
66. Gan F, Liu H, Qin WG, Zhou SL. Application of artificial intelligence for automatic cataract staging based on anterior segment images: comparing automatic segmentation approaches to manual segmentation. Front Neurosci 2023;17:1182388. [doi: 10.3389/fnins.2023.1182388] [Medline: 37152605]
67. Tham YC, Goh JHL, Anees A, et al. Detecting visually significant cataract using retinal photograph-based deep learning. Nat Aging 2022 Mar;2(3):264-271. [doi: 10.1038/s43587-022-00171-6] [Medline: 37118370]
68. Siddique M. Convolutional neural network modeling for eye disease recognition. Int J Onl Eng 2022 Jul;18(9):115-130. [doi: 10.3991/ijoe.v18i09.29847]
69. Sirajudeen A, Balasubramaniam A, Karthikeyan S. Novel angular binary pattern (NABP) and kernel based convolutional neural networks classifier for cataract detection. Multimed Tools Appl 2022 Nov;81(27):38485-38512. [doi: 10.1007/s11042-022-13092-8]
70. Junayed MS, Islam MB, Sadeghzadeh A, Rahman S. CataractNet: an automated cataract detection system using deep learning for fundus images. IEEE Access 2021;9:128799-128808. [doi: 10.1109/ACCESS.2021.3112938]
71. Hu S, Wang X, Wu H, et al. Unified diagnosis framework for automated nuclear cataract grading based on smartphone slit-lamp images. IEEE Access 2020;8:174169-174178. [doi: 10.1109/ACCESS.2020.3025346] [Medline: 33747677]
72. Hu S, Luan X, Wu H, et al. ACCV: automatic classification algorithm of cataract video based on deep learning. Biomed Eng Online 2021 Aug 5;20(1):78. [doi: 10.1186/s12938-021-00906-3] [Medline: 34353324]
73. Lai CJ, Pai PF, Marvin M, Hung HH, Wang SH, Chen DN. The use of convolutional neural networks and digital camera images in cataract detection. Electronics (Basel) 2022;11(6):887. [doi: 10.3390/electronics11060887]

XSL•FO
RenderX

74. Askarian B, Ho P, Chong JW. Detecting cataract using smartphones. IEEE J Transl Eng Health Med 2021;9:3800110. [doi: 10.1109/JTEHM.2021.3074597] [Medline: 34786216]
75. Son KY, Ko J, Kim E, et al. Deep learning-based cataract detection and grading from slit-lamp and retro-illumination photographs: model development and validation study. Ophthalmol Sci 2022 Jun;2(2):100147. [doi: 10.1016/j.xops.2022.100147] [Medline: 36249697]
76. Saju B, Rajesh R. Eye-vision net: cataract detection and classification in retinal and slit lamp images using deep network. Int J Adv Comput Sci Appl 2022;13(12). [doi: 10.14569/IJACSA.2022.0131227]
77. Chellaswamy C, Geetha TS, Ramasubramanian B, Abirami R, Archana B, Divya Bharathi A. Optimized convolutional neural network based multiple eye disease detection and information sharing system. Presented at: 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS); May 25-27, 2022. [doi: 10.1109/ICICCS53718.2022.9788334]
78. Lu Q, Wei L, He W, et al. Lens opacities classification system III-based artificial intelligence program for automatic cataract grading. J Cataract Refract Surg 2022 May 1;48(5):528-534. [doi: 10.1097/j.jcrs.0000000000000790] [Medline: 34433780]
79. Al‐Naji A, Khalid GA, Mahmood MF, Chahl JS. Computer vision for eye diseases detection using pre‐trained deep learning techniques and raspberry Pi. J Eng 2024 Jul;2024(7) [FREE Full text] [doi: 10.1049/tje2.12410]
80. Elloumi Y. Cataract grading method based on deep convolutional neural networks and stacking ensemble learning. Int J Imaging Syst Tech 2022 May;32(3):798-814 [FREE Full text] [doi: 10.1002/ima.22722]
81. Zannah TB, Abdulla-Hil-Kafi M, Sheakh M, et al. Bayesian optimized machine learning model for automated eye disease classification from fundus images. Computation 2024;12(9):190. [doi: 10.3390/computation12090190]
82. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
83. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 6;327(7414):557-560. [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
84. Reitsma JB, Glas AS, Rutjes AWS, Scholten R, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005 Oct;58(10):982-990. [doi: 10.1016/j.jclinepi.2005.02.022] [Medline: 16168343]
85. Dwamena B. MIDAS: Stata module for meta-analytical integration of diagnostic test accuracy studies. Statistical Software Components 2009 Feb;14 [FREE Full text]
86. Chu Y, Hu S, Li Z, et al. Image analysis-based machine learning for the diagnosis of retinopathy of prematurity: a meta-analysis and systematic review. Ophthalmol Retina 2024 Jul;8(7):678-687. [doi: 10.1016/j.oret.2024.01.013] [Medline: 38237772]
87. Chylack LT, Wolfe JK, Singer DM, Leske MC, Bullimore MA, Bailey IL. The Lens Opacities Classification System III. Arch Ophthalmol 1993 Jun 1;111(6):831. [doi: 10.1001/archopht.1993.01090060119035]
88. Mackenbrock LHB, Labuz G, Baur ID, Yildirim TM, Auffarth GU, Khoramnia R. Cataract classification systems: a review. Klin Monbl Augenheilkd 2024 Jan;241(1):75-83. [doi: 10.1055/a-2003-2369] [Medline: 38242135]
89. Gali HE, Sella R, Afshari NA. Cataract grading systems: a review of past and present. Curr Opin Ophthalmol 2019 Jan;30(1):13-18. [doi: 10.1097/ICU.0000000000000542] [Medline: 30489359]
90. Cheung R, So S, Malvankar-Mehta MS. Diagnostic accuracy of machine learning classifiers for cataracts: a systematic review and meta-analysis. Expert Rev Ophthalmol 2022 Nov 2;17(6):427-437. [doi: 10.1080/17469899.2022.2142120]
91. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019 Oct;1(6):e271-e297. [doi: 10.1016/S2589-7500(19)30123-2] [Medline: 33323251]
92. Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med 2021 Apr 7;4(1):65. [doi: 10.1038/s41746-021-00438-z] [Medline: 33828217]
93. Islam MM, Poly TN, Walther BA, Yang HC, Li YCJ. Artificial intelligence in ophthalmology: a meta-analysis of deep learning models for retinal vessels segmentation. J Clin Med 2020 Apr 3;9(4):1018. [doi: 10.3390/jcm9041018] [Medline: 32260311]
94. Shen D, Wu G, Zhang D, Suzuki K, Wang F, Yan P. Machine learning in medical imaging. Comput Med Imaging Graph 2015 Apr;41:1-2. [doi: 10.1016/j.compmedimag.2015.02.001] [Medline: 25727143]
95. Castiglioni I, Rundo L, Codari M, et al. AI applications to medical images: from machine learning to deep learning. Phys Med 2021 Mar;83:9-24. [doi: 10.1016/j.ejmp.2021.02.006] [Medline: 33662856]
96. Islam MM, Yang HC, Poly TN, Jian WS, Jack Li YC. Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: a systematic review and meta-analysis. Comput Methods Programs Biomed 2020 Jul;191:105320. [doi: 10.1016/j.cmpb.2020.105320] [Medline: 32088490]
97. Buisson M, Navel V, Labbé A, et al. Deep learning versus ophthalmologists for screening for glaucoma on fundus examination: a systematic review and meta-analysis. Clin Exp Ophthalmol 2021 Dec;49(9):1027-1038. [doi: 10.1111/ceo.14000] [Medline: 34506041]

XSL•FO
RenderX

98. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. BMJ 2020 Sep 9;370:m3210. [doi: 10.1136/bmj.m3210] [Medline: 32907797]
99. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ 2020 Sep 9;370:m3164. [doi: 10.1136/bmj.m3164] [Medline: 32909959]
100. Sparrow JM, Bron AJ, Brown NAP, Ayliffe W, Hill AR. The Oxford clinical cataract classification and grading system. Int Ophthalmol 1986 Dec;9(4):207-225. [doi: 10.1007/BF00137534] [Medline: 3793374]
101. Chylack LT, Lee MR, Tung WH, Cheng HM. Classification of human senile cataractous changes by the American Cooperative Cataract Research Group (CCRG) method. I. Instrumentation and technique. Invest Ophthalmol Vis Sci 1983 Apr;24(4):424-431. [Medline: 6832915]
102. Zhou Y, Li G, Li H. Automatic cataract classification using deep neural network with discrete state transition. IEEE Trans Med Imaging 2020 Feb;39(2):436-446. [doi: 10.1109/TMI.2019.2928229] [Medline: 31295110]
103. Chang J, Lee J, Ha A, et al. Explaining the rationale of deep learning glaucoma decisions with adversarial examples. Ophthalmology 2021 Jan;128(1):78-88. [doi: 10.1016/j.ophtha.2020.06.036] [Medline: 32598951]
104. Araújo T, Aresta G, Mendonça L, et al. DR|GRADUATE: uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. Med Image Anal 2020 Jul;63:101715. [doi: 10.1016/j.media.2020.101715] [Medline: 32434128]
105. van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal 2022 Jul;79:102470. [doi: 10.1016/j.media.2022.102470] [Medline: 35576821]
106. Abràmoff MD, Cunningham B, Patel B, et al. Foundational considerations for artificial intelligence using ophthalmic images. Ophthalmology 2022 Feb;129(2):e14-e32. [doi: 10.1016/j.ophtha.2021.08.023] [Medline: 34478784]

Abbreviations

AI: artificial intelligence
AUC: area under the receiver operating characteristic curve
CC: cortical cataract
CNN: convolutional neural network
DL: deep learning
DR: diabetic retinopathy
FDA: the US Food and Drug Administration
FN: false negative
FP: false positive
***ICD-11*:** *International Classification of Diseases, 11th Revision*
LMIC: low- and middle-income country
LOCS III: lens opacities classification system III
LRS: low-resource settings
ML: machine learning
NC: nuclear cataract
OCT: optical coherence tomography
PI: prediction interval
PPC: posterior polar cataract
PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-analyses of Diagnostic Test Accuracy
PROSPERO: International Prospective Register of Systematic Reviews
PSC: posterior subcapsular cataract
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies-2
ResNet: residual network
RevMan: Review Manager
RoB: risk of bias
SROC: summary receiver operating characteristic curve
TN: true negative
TP: true positive
WHO: World Health Organization
XAI: explainable artificial intelligence

XSL•FO
RenderX

Edited by A Coristine; submitted 11.Jun.2025; peer-reviewed by GL Legault, H Tanabe, X Wang, Y Lu; revised version received 14.Feb.2026; accepted 18.Feb.2026; published 29.Apr.2026.

Please cite as:
Li R, Li H, Li C, Li S, Lei L, Tao D
Image-Based Deep Learning for Cataract Diagnosis: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e78869
URL: https://www.jmir.org/2026/1/e78869
doi:10.2196/78869

© Ruixi Li, Hongyi Li, Chong Li, Shuo Li, Linhong Lei, Dan Tao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Hybrid Care Modifications in the Delivery of Nonpandemic Care During the COVID-19 Pandemic: Scoping Review

Natalia Sanchez Villalobos[1], MD; Tessa van Loenen[1], PhD; Lem Ngongalah[2], PhD; Máire A Connolly[2], MD, PhD; Mart L Stein[3], PhD; Chantal P Rovers[4], MD, PhD; Aura Timen[1], MD, PhD

[1]Department of Primary and Community Care, Radboud University Medical Center, Geert Grooteplein Noord 21, Nijmegen, The Netherlands
[2]School of Health Sciences, University of Galway, Galway, Ireland
[3]Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, The Netherlands
[4]Department of Internal Medicine, Radboud University Medical Center, Nijmegen, The Netherlands

Corresponding Author:
Natalia Sanchez Villalobos, MD
Department of Primary and Community Care, Radboud University Medical Center, Geert Grooteplein Noord 21, Nijmegen, The Netherlands

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2026/1/e101728/

Abstract

Background: The COVID-19 pandemic had an unprecedented impact on the delivery of health care, with digital interventions accelerating more than ever before. However, evidence of how hybrid care models, combining digital health interventions with in-person care, were implemented during the pandemic remains scattered. Understanding hybrid care models is imperative to build resilient health systems that can ensure access to care during crisis situations.

Objective: The study aimed to examine the implementation of hybrid care modifications to support the delivery of nonpandemic health care services in Europe during the COVID-19 pandemic.

Methods: A scoping review was conducted following PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines. Systematic searches were conducted in PubMed or MEDLINE, Embase, CINAHL, Web of Science, and PsycINFO on May 22, 2024, and updated on January 14, 2026. Studies were eligible if they included primary data on the use of digital care modifications implemented or scaled up during the COVID-19 pandemic for the delivery of nonpandemic health care services in Europe. Non–peer-reviewed publications and studies with a primary focus on mental health or pediatric care were excluded. Quality appraisal was conducted using the Mixed Methods Appraisal Tool. Descriptions of digital care modifications were inductively analyzed and used to create digital flows, combining telehealth systems, digital interventions, and care functions. Digital care modifications were categorized according to their hybrid care implementation (digital-only or hybrid). Study evaluations were extracted using the Kirkpatrick model.

Results: A total of 189 studies were included for analysis. Studies covered evidence from 2020 to 2024, a total of 23 countries, and 37 health care disciplines. Hybrid care implementation was reported in over 60% (115/189) of the studies, describing various forms of digital and in-person care. Care modifications incorporating in-person and digital care components were more commonly described in specialty care contexts. A total of 68 distinct digital flows were identified, with a limited number of telehealth systems allowing substantial variety in both interventions and care functions. Prominent digital flows included the use of online platforms to support video and messaging for follow-up care. Over half of the studies did not describe any kind of evaluation.

Conclusions: This review has shown how few telehealth systems were able to support a variety of care functions in the delivery of nonpandemic care throughout the COVID-19 pandemic, underscoring their practical versatility. Integrating digital health as part of hybrid care models is essential in designing care pathways that can adapt to different contexts, including future health crises. Although a comprehensive search was conducted, the heterogeneous reporting of care modifications may have influenced the interpretation of the findings. In the future, research may expand the application of hybrid care models to innovative strategies for effective crisis management.

(J Med Internet Res 2026;28:e84756) doi:10.2196/84756

KEYWORDS

hybrid care; digital health; telemedicine; COVID-19 pandemic; health care delivery

XSL•FO
RenderX

Introduction

The COVID-19 pandemic had an unprecedented impact on the delivery of health care, leading to modifications that sought to compensate for service disruptions. By 2022, the World Health Organization reported that over 90% of countries worldwide were still facing disruptions in health service delivery [1]. As health systems continue to recover post pandemic, it is critical to understand the modifications implemented to support the delivery of nonpandemic health care (ie, health care provided for conditions other than COVID-19 disease). Among the various modifications implemented, the rapid surge in the use of digital health played a critical role in responding to the challenges posed by the pandemic, which restricted resources and impeded the safe delivery of in-person care [2]. Understanding modifications to care, particularly those involving digital health, is crucial for building resilient health systems that can ensure access to care, including during emergency situations.

Digital health refers to the use of technologies, including those for communication and information exchange, to support health systems [3]. Previous literature has covered the adoption and use of digital health during the COVID-19 pandemic, showing its potential benefits for operational and clinical applications such as data services, population-level responses, decision-support systems, and direct patient care [4,5]. For instance, the value of digital health interventions during the pandemic has been described for patients with conditions such as cancer and heart failure [6,7]. Although it has been recognized that digital health should be examined beyond the implementation of individual tools, research to date has tended to focus on standalone digital health interventions. Rather than studying these interventions as silos, they can be better examined as part of broader care pathways [5], highlighting the critical need for integration into existing health care systems [8].

Alongside digital health interventions, hybrid approaches have emerged as integrated alternatives to ensure the continuity of essential health services. Hybrid care can be defined as the delivery of health care services involving in-person (face-to-face) care and digital health components. Hybrid care provision has been studied previously in various contexts [9-12], including explorations of its potential in contrast to standard models of care for patients with chronic conditions [13,14]. In the context of the COVID-19 pandemic, previous research has shown a positive impact of the use of hybrid care models in specific clinical contexts [15]. Additionally, research on the effectiveness of telehealth compared to in-person care has highlighted the importance of further researching the integration between these types of care [16].

Given the increased uptake, hybrid care is anticipated to become the new norm for health care delivery in the coming years, offering novel avenues for unlocking access to care during regular and crisis situations. However, little is still known about how, and for which purposes, these models of care have been implemented and evaluated in previous crisis contexts. A scoping review was chosen to systematically map available evidence regarding the different types of modifications to care.

This scoping review aimed to examine the implementation of hybrid care modifications to support the delivery of nonpandemic health care services in Europe during the COVID-19 pandemic. Specifically, it aimed to characterize modifications of care involving digital health, including the telehealth systems used and their functionalities, and how they were integrated into hybrid models of care. Additionally, it aimed to provide an overview of the maturity and extent of evaluations conducted on the implemented care modifications.

Methods

Study Design

A scoping review was conducted to examine reports on health delivery modifications incorporating digital health components, which were introduced due to the COVID-19 pandemic to deliver nonpandemic health care services. The review was conducted as part of the Horizon Europe RAPIDE (Regular and Unplanned Care Adaptive Dashboard for Cross-Border Emergencies) project and followed the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines [17] (Checklist 1). The research protocol was registered on the Open Science Framework Registries [18].

Eligibility Criteria

The main outcome of interest for included studies was modifications to health delivery involving a digital health intervention, implemented for the provision of nonpandemic care in the context of the COVID-19 pandemic in Europe. Evidence sources included peer-reviewed primary research studies. Aligned with this outcome and context, the following specific inclusion criteria were applied: (1) studies reporting primary data on the use of digital care modifications (ie, health care delivery modifications with digital health components), alongside descriptions of their implementation, intervention characteristics (eg, specific setting, patient groups, and clinical contexts), telehealth systems used, and their care functions; (2) studies describing digital care modifications introduced or scaled up in the context of the COVID-19 pandemic and aimed for nonpandemic care services; (3) studies focused on care for adult populations; and (4) studies reporting data from countries within the European Union, the European Economic Area, the Schengen area, or the United Kingdom.

In addition, the following exclusion criteria were considered: (1) non–peer-reviewed publications; (2) studies that did not separately report outcomes for eligible countries; (3) publications in a language other than English, Dutch, French, or Spanish; and (4) studies with a primary focus on mental health or pediatric health care. Studies with a focus on mental and pediatric care were excluded due to the distinct organization and service delivery models used in these types of care, as these studies would have introduced additional variability beyond the scope of this review.

Information Sources

A comprehensive search was initially conducted on May 22, 2024, and later updated on January 14, 2026. The following electronic databases were searched, without applying

XSL•FO
RenderX

multidatabase searching: PubMed or MEDLINE (Ovid), Embase (Ovid), CINAHL (EBSCOhost), Web of Science (Clarivate), and PsycINFO (Ovid). No other registries or online resources were consulted, and no additional data were sought through contacts. No additional information sources or search methods were used.

Search

The search aimed to cover a broad range of literature, for which the concepts "delivery of health care," "nonpandemic care," and "COVID-19" were used to build the search strategy across all databases, including a combination of keywords and indexed vocabulary. No further limits, restrictions, or search filters were applied. No existing search strategies from other reviews were adapted or reused. The search was developed and reviewed by all authors in consultation with an information specialist, and no other external experts peer-reviewed the search. The original search strategy was reproduced for updating the search, which was consistent with the original search in all databases. The reproducible searches for all databases are available in Multimedia Appendix 1. Duplicates were removed by the researchers using EndNote (Clarivate) and Rayyan.ai (Qatar Computing Research Institute) deduplication features. All information related to the search used is reported in accordance to the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) [19] (Checklist 2).

Selection of Sources of Evidence

A broad search was conducted with the aim of capturing the diversity of service delivery modifications. First, all studies describing service delivery modifications (both digital and nondigital) and meeting all other eligibility criteria were preliminarily included. Next, studies were categorized according to the types of service delivery modifications, and those applying digital care modifications were selected for inclusion. This approach ensured the selection of studies that addressed digital health as part of broader service modification, even if it was not the primary focus.

As part of the screening, a pilot was first conducted for title and abstract screening to ensure clarity among researchers regarding eligibility criteria. Title and abstract screening was conducted by 3 researchers (NSV, TVL, and LN), and 10.4% (798/7652) of the total records were screened by at least two researchers independently. Full-text screening was conducted by 2 researchers (NSV and LN), with 5% (32/641) of the records screened by both researchers independently. The remaining records were screened separately due to strong agreement in the pilot screening. Citation searching was conducted for review studies that met the initial inclusion criteria through a manual search of reference lists; cited studies were then individually screened for inclusion. All discrepancies were resolved by consensus. The web application Rayyan.ai was used to support the screening process.

Data Charting Process

A standardized data extraction form was developed and piloted among 3 researchers (NSV, TVL, and Dilek Yildirim), until agreement and consistency were reached. Studies were then split among the researchers who conducted the data extraction (NSV, TVL, and Dilek Yildirim). The online platform Airtable (Formagrid Inc) was used for data extraction, and Microsoft Excel was used to support additional data collation for reporting.

Data Items

Data extracted included the year of publication, study design, country, health care discipline, level of care, pandemic period of study data, presence and type of evaluation, and characteristics of the digital care modifications (including telehealth systems, digital interventions, care functions, and in-person care components). Definitions for the uses of digital technology for health, including systems and interventions, were based on World Health Organization [3] and were used to inform both the terminology and categorization of digital components during data extraction and synthesis.

Critical Appraisal of Individual Sources of Evidence

The methodological quality of the studies was appraised by 2 researchers (NSV and TVL) using the Mixed Methods Appraisal Tool (MMAT) [20]. The researchers independently rated a subset of 5 studies and aligned on the interpretation of criteria. The remaining studies were divided and separately appraised by the 2 researchers using a standardized form. Percentages of quality criteria met were calculated per study, and no studies were excluded based on methodological quality.

Synthesis of Results

Descriptions of digital care modifications were inductively analyzed and used to create distinct *digital flows*, defined as the combination of 1 telehealth system, 1 digital intervention, and 1 care function (Table 1). Digital flows were constructed by the researchers based on descriptions of digital health modifications in each study. For studies describing several systems, interventions, or care functions, multiple digital flows were recorded. Digital care modifications were further categorized according to their hybrid care implementation, registering whether digital health was implemented as an individual approach (digital-only) or in combination with the provision of in-person care components (hybrid). Study evaluations of reported modifications were recorded following the Kirkpatrick model of evaluation, allowing multiple levels of evaluation, if applicable [21]. The Kirkpatrick model covers 4 distinct levels of evaluation and has been widely applied across health disciplines, allowing its pragmatic application across diverse outcomes as well as its straightforward interpretation [21].

XSL•FO
RenderX

Table . Definitions used to construct digital flows for mapping hybrid care modifications in the delivery of nonpandemic care.

Concept	Definition
Telehealth systems	Information and telecommunication technologies used to provide clinical care [3]
Digital health interventions	Technology capabilities designed to address a specific health objective [3]
Care function	Function or purpose of the digital health intervention: includes triage (prioritization of patients' needs based on the urgency of their condition), first consultation (initial contact between a patient and a health care provider to evaluate a novel health concern), follow-up (subsequent care contacts to evaluate previously diagnosed conditions), among others

Results

Selection of Sources of Evidence

In total, 7652 distinct studies were identified, of which 641 full texts were screened for eligibility. Studies included within identified reviews (n=18) were further screened for eligibility, resulting in 38 additional included studies. Finally, 189 studies met eligibility criteria and were included in the analysis (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram depicting the study selection process. [a]Article retracted. [b]Outcome defined as digital health modifications meeting inclusion criteria. [c]Includes opinion pieces, editorials, letters to the editor, correspondence, commentaries, position papers, and other non–peer-reviewed publications. [d]Review articles snowballed for additional records meeting inclusion criteria, with 38 additional records identified for inclusion.

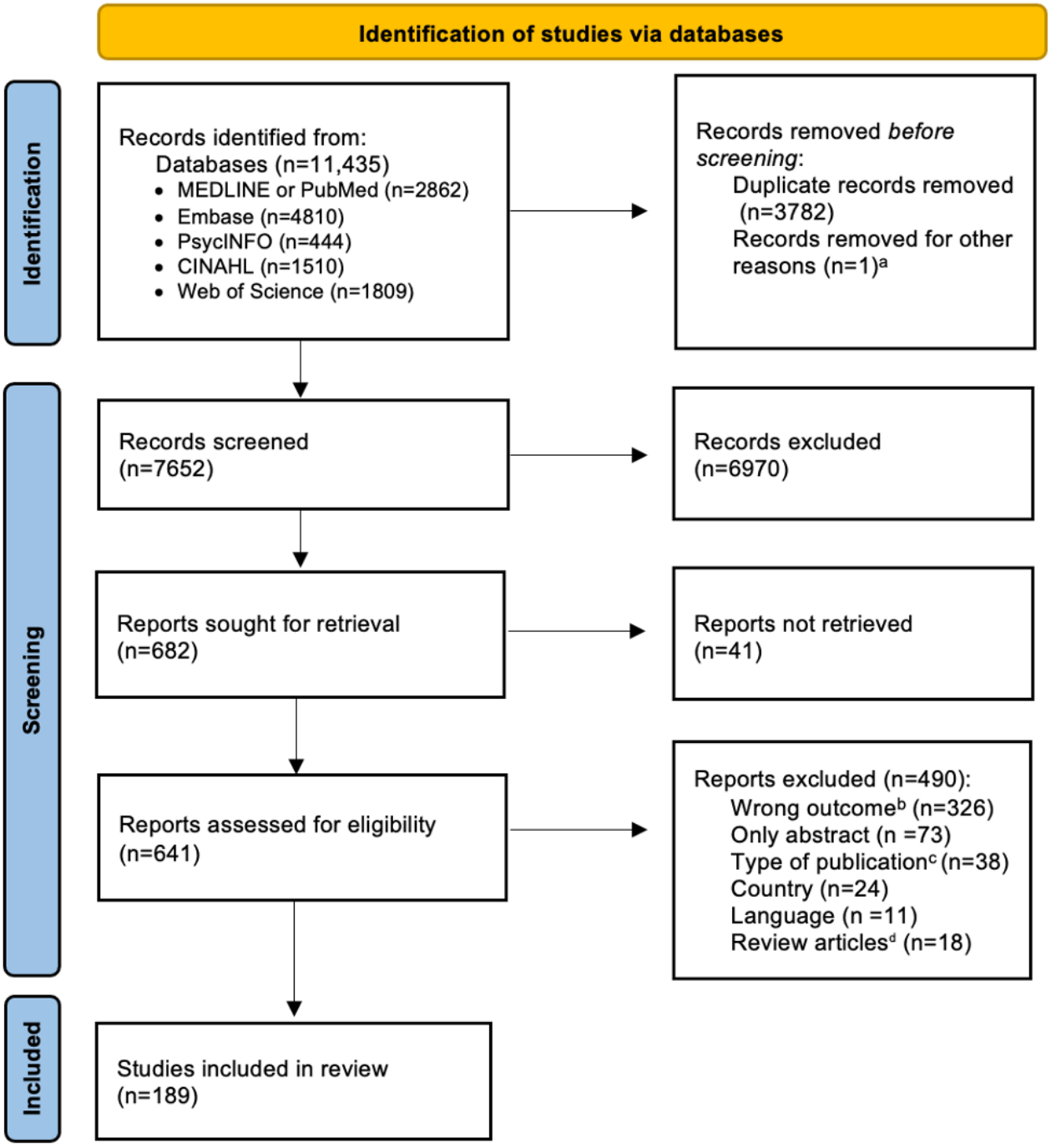

Characteristics of Sources of Evidence

Included studies were published between 2020 and 2025. Most studies were written in English (183/189, 96.9%), and several had full text available in Spanish (n=4) [22-25] and French (n=2) [26,27]. Over 80% (n=157) of the studies were either observational (n=79) or descriptive (n=78). The studies covered 23 distinct countries and 37 distinct health care disciplines across different settings, with a strong dominance of secondary and tertiary care (n=139) and nonsurgical disciplines (26/37, 70.3%). Data derived from the year 2020 were most frequently reported, and 114 studies exclusively covered data from February to June 2020. The studies' characteristics are found in Table 2 (additional details provided in Multimedia Appendix 2).

XSL•FO
RenderX

Table . Characteristics of included studies in the scoping review.

Characteristics	Studies
Year of publication, n (%)	
2020	65 (34.4)
2021	57 (30.2)
2022	27 (14.3)
2023	17 (9)
2024	15 (7.9)
2025	8 (4.2)
Countries[a], n	
United Kingdom	52
Italy	45
Spain	18
France	12
The Netherlands	8
Ireland	7
Portugal	6
Germany	5
Poland	5
Sweden	4
Austria	3
Belgium	3
Denmark	3
Norway	3
Romania	3
Switzerland	3
Bulgaria	1
Croatia	1
Czech Republic	1
Finland	1
Hungary	1
Malta	1
Slovenia	1
Studies focusing on 5 or more countries	6
Study design, n (%)	
Observational	79 (41.8)
Descriptive	78 (41.3)
Qualitative	14 (7.4)
Mixed methods	13 (6.9)
Experimental	5 (2.6)
Levels of care[b], n	
Secondary or tertiary care	139
Primary care	57

XSL•FO
RenderX

Characteristics	Studies
Community care	5
Home care or residential care	1
Alternative medicine	1
Not specified	1
Health care disciplines[c], n	
General practice or family medicine	46
Oncology	32
Cardiology	14
Neurology	14
Orthopedics and trauma	14
Endocrinology	8
Gastrointestinal medicine	8
Gynecology or obstetrics, including reproductive health	7
Dentistry or oral surgery	6
Dermatology	4
Midwifery	4
Ophthalmology	4
Plastic surgery and burn unit	4
Rheumatology	4
Urology	4
Immunology	3
Physiotherapy	3
Pharmacy services	3
Other[d]	26
Multiple (>5)	5
Not specified[e]	3
Pandemic period of reported data[f], n	
February-June 2020	170
July-December 2020	59
January-June 2021	39
July-December 2021	26
January-December 2022	8
2023 - 2024	8
Not specified	6

[a]For studies reporting more than 1 country (but <5), all countries were counted separately.

[b]14 studies reported more than 1 level of care.

[c]For studies reporting more than 1 health care discipline (but <5), all disciplines were counted separately.

[d]Other health care disciplines, less frequently reported. Details provided in Multimedia Appendix 2.

[e]Specific areas of care not indicated.

[f]Refers to the period corresponding to the data reported in studies (not the date of publication). Studies may report data from several periods. The number of studies in each period is reported and therefore studies may be represented in different time periods.

XSL•FO
RenderX

Critical Appraisal Within Sources of Evidence

Quality appraisal was conducted for 84.1% (159/189) of the studies, while 15.9% (30/189) comprised narrative accounts of care modifications and did not meet the MMAT screening criteria to be appraised. Nearly 60% (94/159) of the appraised studies met 80% or more of the quality criteria, and only a small number of studies (9/159) met 40% or less of the MMAT criteria.

Synthesis of Results

Digital and Hybrid Care Modifications

Hybrid care modifications (incorporating both in-person and digital care components) were reported by 60.8% (115/189) of the studies, while digital-only care modifications (without in-person care components) were reported by 36% (68/189) of the studies. In 3.2% (6/189) of the studies, the extent of hybrid care use was not reported, and broad descriptions of digital care use did not allow classification into a specific category; although these studies could not be classified, they were still incorporated in the descriptive analysis of digital care modifications.

Care modifications incorporating hybrid modalities integrated digital and in-person components in a variety of ways. Hybrid care was used either for a determined subset of patients, for example, offering both in-person and digital care for the same patient but for different care purposes, or for different patient populations, for example, selecting either digital care or in-person care. Service delivery modifications also ranged from being predominantly focused on digital care while allowing in-person care in certain contexts [28-59] to approaches that incorporated a digital care component while predominantly maintaining in-person care [60-63]. For instance, several studies described the dominance of remote services, with in-person visits used only for specific clinical indications [28,36,48] or care purposes such as therapy administration [34] or clinical examination [41]. Most modifications, however, described the incorporation of substantial elements of both in-person and digital care.

Care modifications incorporating in-person and digital care components were commonly described in specialty care contexts. Organized care pathways, involving structured models for hybrid care provision, were found in areas, including orthopedic surgery [64,65], oncology [66,67], plastic surgery [68], and endocrinology [69]. Set clinical criteria, definition of patient groups, and applications of triage were common considerations reported to define the extent of application of hybrid care, both in primary and specialty care contexts.

Among modifications incorporating substantial in-person and digital care components, in-person care was commonly offered to patients requiring nondeferrable care [24,65,68,70-80], often decided after remote assessments [65,71,73,77]. Several studies reported in-person care only in clinically necessary cases [81-84]. Specific clinical criteria were often used to select in-person care contacts, including cases of severe disease or suspicion of malignancy [66,78,85-88]. Several studies focusing on cancer patients prioritized in-person consultations for new diagnoses or initial assessments [26,72,74,76,82]. Other patient factors were considered to select in-person versus remote care, including patient preference [82,87,89], patient age [69,90], and ease of access to health care facilities [91]. Conversely, remote consultations were more commonly reported for known patients [81,92] and scheduled follow-up consultations [83,93-99].

Studies commonly reported multiple in-person care components taking place within hybrid care provision, for either the same group of patients or for different patients within a health service. In general, in-person consultations and surgical interventions were the most frequently reported in-person care components involved in hybrid care use. In several cases, interventions such as physical examinations and functional tests were explicitly reported separately from overall consultations (Table 3).

Table . In-person care components reported in modifications involving hybrid care use for the delivery of nonpandemic care throughout the COVID-19 pandemic in Europe.

In-person care interventions	Studies, n
Consultation	85
Surgical intervention	23
Therapy administration	18
Minor procedures or interventions	17
Imaging	16
Laboratory tests	15
Physical examination	13
Inpatient interventions	7
Functional tests	5
Endoscopic procedures	3
Vital signs measurement	3
Not specified	6

In addition, hybrid and digital care modifications were implemented to varying degrees throughout the pandemic. Hybrid care (including in-person and digital, predominantly digital, and predominantly in-person modifications) was more

XSL•FO
RenderX

frequently reported than digital-only care during the early stages of the pandemic (February to June 2020), while digital-only care was reported more often between July 2020 and June 2021. In 2022, hybrid care variations remained more common than digital-only modifications, although overall reporting decreased. Studies reporting care modifications in the postpandemic period (2023 and 2024) were scarce (Figure 2).

Figure 2. Variations in the report of hybrid care modifications for the delivery of nonpandemic care across different periods of the COVID-19 pandemic in Europe. Bubbles depict number of studies reporting different types of hybrid care modifications. Unspecified values are excluded.

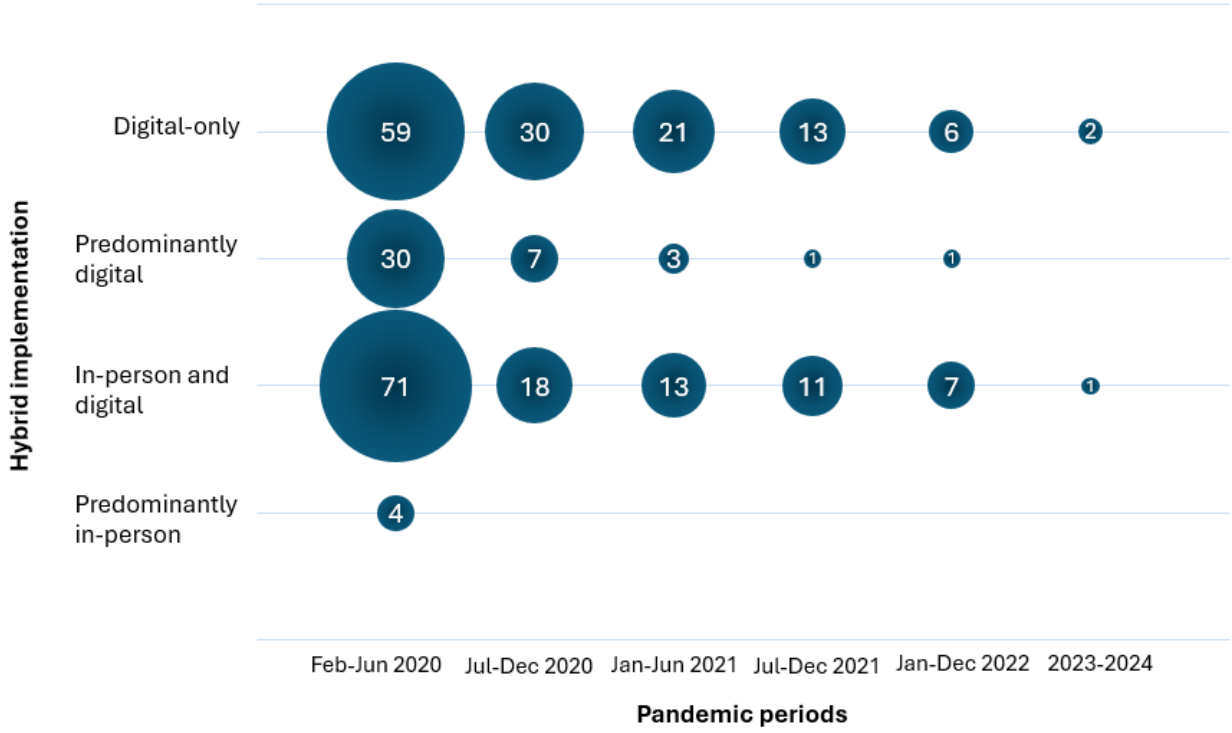

Digital Flows

Digital care modifications identified are illustrated through the creation of digital flows, composed of 3 levels that indicate telehealth systems, digital interventions, and care functions (Figure 3). A total of 433 individual digital care elements, comprising the 3 levels of telehealth systems, digital interventions, and care functions, were extracted, creating 68 distinct digital flows. Substantial variety was found in both interventions and care functions, while the diversity of telehealth systems was more limited. Distinct combinations of systems, interventions, and care functions formed several prominent flows.

XSL•FO
RenderX

Figure 3. Digital flows illustrating digital care modifications implemented during the COVID-19 pandemic for the delivery of nonpandemic care in Europe. Digital flows are composed of 3 levels that indicate telehealth systems, digital interventions, and care functions. All telephone calls were assumed to have occurred via telephone service, if not indicated otherwise. Telehealth system for video calls was not indicated in 25 cases, for which the use of "website or platform" was assumed. Digital interventions were classified following the World Health Organization classification of digital interventions, services, and applications in health [3] whenever possible, with multiple interventions recorded when applicable.

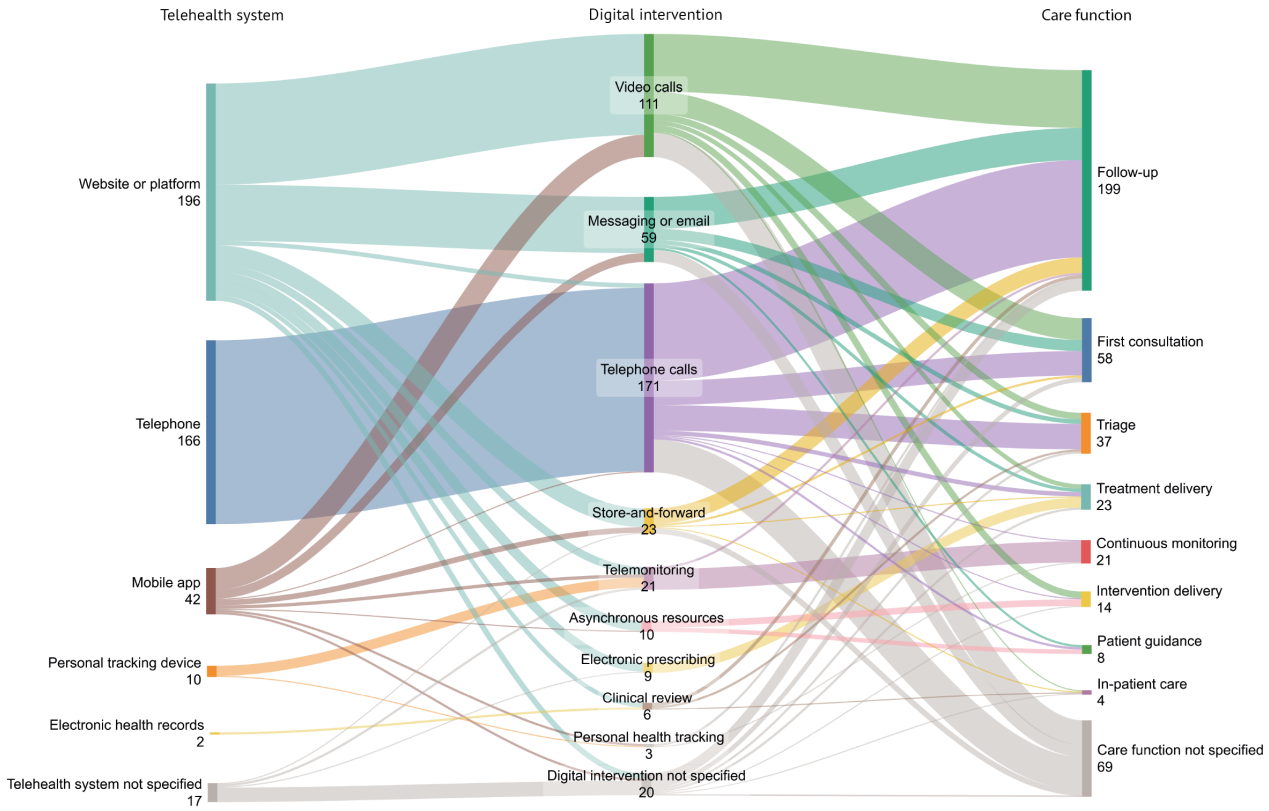

One of the most prominent digital flows was the use of websites or platforms to conduct video calls, for the provision of follow-up care. This flow was commonly described in outpatient specialty care contexts, including neurology [22,23,29,100-102] and orthopedics [43,55,103,104]. Follow-up consultations in these cases often involved chronic patient care and postsurgical assessments, respectively. While specialty care was more common, video calls for follow-up purposes were also found at the primary and community care levels [51,95,105-111]. Additionally, the use of mobile technologies to support video calls was frequently linked to the follow-up care of patients in specialty contexts [35,43,49,82,94,100,103,112-114].

In addition to video calls, the role of messaging and email stood out. The application of these interventions was more evenly divided among primary and secondary or tertiary levels of care. At the primary care level, text-based interventions involved triage, first consultations, follow-up, and patient guidance as care functions. At the secondary or tertiary care level, interventions entailed follow-up of patients, including in cardiology [46,49,62,114-116] and oncology [32,34,117,118] services. Additionally, a number of text-based interventions highlighted mobile apps as the telehealth systems used [49,109,111,114,118-120].

Another key digital flow involved the use of telephone calls, commonly with the functions of follow-up, first consultations, and triage. The use of telephone calls was extensively reported at the secondary or tertiary care level across specialties including oncology [32,34,45,50,54,59,63,76,78,79,121-124], orthopedics [30,37,38,55,64,65,75,104,125-127], and neurology [22,29,53,100,101,128-132]. At the primary care level, telephone use often involved follow-up care, but first consultations and triage also took a substantial role.

Moreover, telemonitoring as a digital intervention was supported by personal tracking devices, websites or online platforms, and mobile apps and was most commonly used for continuous patient monitoring. Telemonitoring included live monitoring of patients' clinical status [133], periodical monitoring to assess treatment adherence and adjust management plans [49], and a combination with other interventions such as telephone follow-up consultations [56,120]. In one instance, it also included the use of artificial intelligence (AI) for instant patient feedback [120].

Evaluation of Modifications

Over half of the studies (96/189, 50.8%) did not describe any kind of evaluation. Studies without evaluation were most

XSL•FO
RenderX

commonly descriptive (n=55) and often exclusively considered data from February to June 2020 (n=64). Studies that evaluated modifications (93/189, 49.2%) were mostly limited to the year 2020 (n=66). Studies involving the use of certain telehealth systems more commonly reported evaluations, including 73.8% (31/42) of studies reporting the use of mobile apps and 61.4% (105/171) of studies reporting the use of websites or platforms.

Among all studies with evaluations, evaluations at a reaction level (n=73) were the most common, followed by evaluations at the results level (n=29) and the behavior level (n=9; Figure 4). No studies included evaluations at a learning level. At the reaction level, methodologies for evaluation ranged from the use of surveys and questionnaires [23,25,28,29,32,34,35,42,43,47,49,52,53,55,56,58,67,78,94,100-104,110,127-154] to qualitative participant accounts, often collected through interviews and focus group discussions [49,52,67,81,100,105-110,119,122,133,142,149,153,155-159].

Figure 4. Levels of evaluation (Kirkpatrick model) and associated evaluated elements found in studies describing hybrid care modifications for the delivery of nonpandemic care during the COVID-19 pandemic in Europe. Each level size is represented according to the number of studies with evaluations at each level. A total of 18 studies reported multiple levels of evaluation. HCP: health care provider [23,25,28,31-34,37,40-42,45,46,48,51-58,66,67,71,77-80,93,95,99-109,112-115,119,121,126-171].

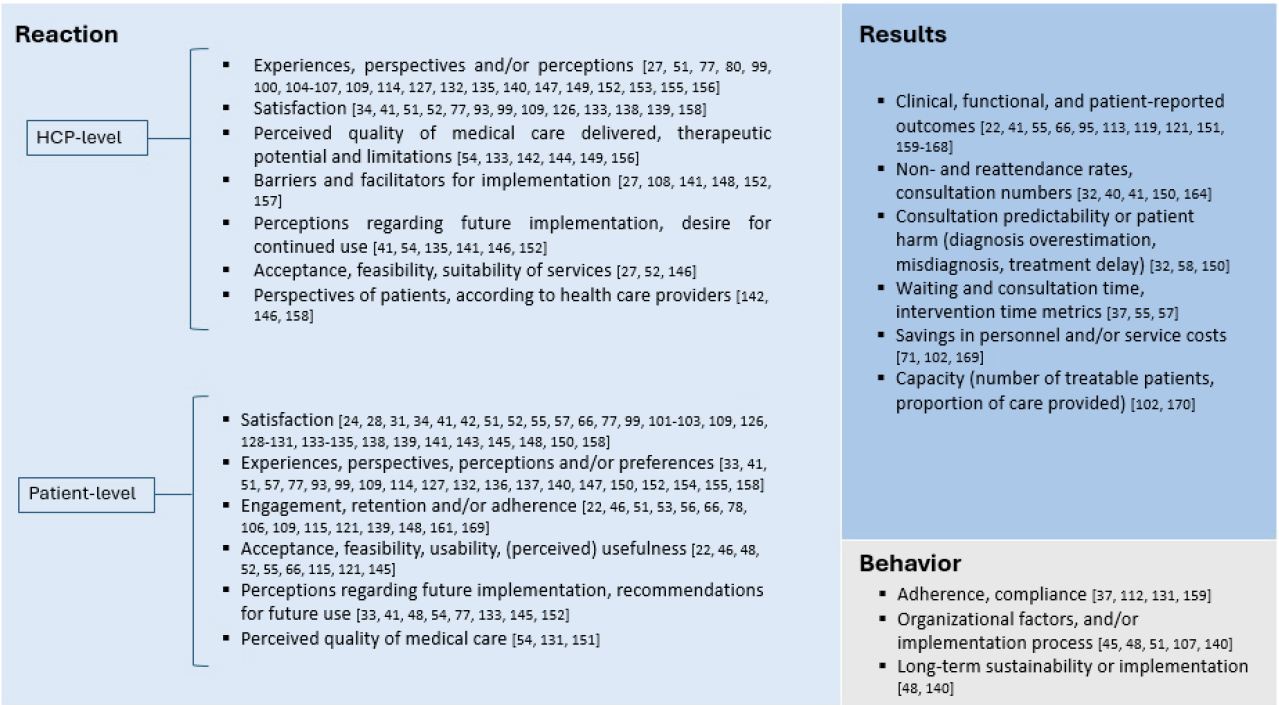

Studies evaluated at the reaction level commonly assessed satisfaction and experiences and perceptions of health care providers and patients. These studies generally reported high levels of satisfaction, as well as overall positive experiences. Concerning health care providers' perspectives, some studies also reported disadvantages such as increased time investment and workload [42,81,156], increased cognitive demand [52], and concerns regarding contact and communication with patients [28,115,133]. From patients' perspectives, studies highlighted benefits, including economic and time travel–related costs [78,115,130]. Additionally, some studies reported concerns related to limited uptake, impeded access, or lower satisfaction among specific patient groups, including the older population [25,35,43,131,134,136,159].

Multiple studies evaluated patients' and providers' perspectives on future implementation of digital health, in some cases proposing new hybrid models of care to be incorporated into new care standards [138]. Other studies used evaluations to build guidance for overcoming barriers for long-term implementation [49] and to join wider initiatives for adapting care services beyond the pandemic period [46]. Additionally, several studies explored the experiences of other stakeholders, including staff involved in the organization of care (eg, receptionists, practice managers, program managers, IT staff, medical advisors, health insurance representatives) [105,106,119,149,153,158] and caregivers [129,130].

In relation to evaluations at the results level, studies mainly focused on the measurement of clinical outcomes (including adverse events and laboratory measurements) [42,56,96,114,120,151,152,160-166], functional outcomes (including physical function measures) [23,67,122,167,168], and patient-reported outcomes (including quality of life and general health status) [42,56,67,122,169].

Discussion

Summary of Evidence

This scoping review has provided a comprehensive examination of modifications involving hybrid care implemented during the COVID-19 pandemic to support nonpandemic care services. To our knowledge, this is the first review focusing on hybrid care modifications implemented throughout the entire pandemic period. During the COVID-19 pandemic, digital health was considerably scaled up to support the provision of nonpandemic

XSL•FO
RenderX

care, which could be observed through frequent reports of digital health as a novel or upscaled application. By mapping digital flows, we were able to examine the diverse ways digital care modifications were implemented in practice, highlighting the importance of analyzing the different components involved in digital health implementation. Linked to digital flows, the limited number of telehealth systems supporting a variety of care functions underscores their practical versatility. This review not only examines the variety of telehealth systems, interventions, and care functions used in different health care settings but also expands the understanding of their implementation as part of hybrid care pathways.

Our findings show that digital health was commonly described as part of hybrid care provision, rather than standalone digital approaches. Central factors for implementation included the selection of patient groups, prioritization based on clinical condition and urgency, and availability of telehealth systems, which often defined the distribution and use cases of in-person and digital interventions. The diverse ways in which in-person and digital care were combined reflect the constant need for adaptation of care models throughout the pandemic and across health care settings. Previous experiences with specific health care settings and patient groups may point to necessary context-specific adaptations [170-172]. Also linked to the constant need for adaptation, the number of publications describing hybrid care modifications decreased throughout the years. This decrease may be partly reflective of the return to prepandemic care delivery but may also be related to a decrease in the reporting of novel care delivery modifications.

The creation of digital flows revealed the various ways digital health technologies were implemented during the pandemic. Notably, a few telehealth systems facilitated the application of numerous interventions and care functions. Online websites and platforms were more commonly reported than mobile technologies, which may have been linked to unclear reporting of which specific systems were used and whether overlaps with other systems were present. Similarly, reports of mobile technology application may have been limited to specific demographic groups, such as the younger population. The concomitant use of multiple telehealth systems may reflect efforts to target distinct care functions and different types of users. Evidence on disparities of technology adoption for different users, including older people and ethnic minorities [173,174], underscores the need for flexibility, adaptation, and consideration of access requirements in the implementation of models of care during emergency situations.

Emerging technologies using AI, such as virtual health assistants, were nearly absent in our findings. Previous studies indicate the use of these technologies during the pandemic was concentrated in areas other than direct nonpandemic care, such as diagnosis and detection of COVID-19 [175], which may also be related to their early stage of development, as well as regulatory barriers hindering wider adoption. As technology advances, it is expected that these types of interventions might take on a more prominent role in the future.

In contrast to telehealth systems, a wider variety of digital interventions was identified. The frequent use of telephones is in line with findings from previous studies [176,177] and may be associated with their familiarity, low cost, and ease of access for providers and patients. The common report of the use of video calls during the pandemic has also been previously reported in multiple health care contexts [177-179]. While text-based interventions have been examined in previous studies [176], more research may be needed to better understand their scope, especially in relation to other interventions such as those involving store-and-forward technologies.

Interventions were most commonly linked to follow-up care, aligning with studies from the first months of the pandemic [176,177,180]. Previous research has pointed out that remote care may yield more positive outcomes when patients have a pre-existing relationship with providers [181]. Distinguishing between new patients and follow-up or known patients may also be of value for assessing satisfaction with remote consultations [55,131]. The broad application of different interventions for follow-up care indicates great potential for enabling continuity of care for patients with chronic conditions, both during normal and crisis situations.

Additionally, our findings show that evaluations of modifications in the delivery of nonpandemic care were limited in scope. Over half of the included studies did not report any form of evaluation, and most of those that did focus on short-term user experiences rather than behavioral, clinical, or long-term implementation outcomes. Many of these care modifications were introduced rapidly as emergency responses to an evolving crisis, often without the time or infrastructure to plan for formal evaluation. Time pressure and resource scarcity during the pandemic, alongside the frequent ad hoc nature of modifications, may have contributed to the low number of evaluations.

Evaluations of patients' and providers' experiences and satisfaction align with previous studies, which have similarly centered on user acceptability and adoption [182] and have found high levels of satisfaction with digital health during the pandemic [183,184]. Behavioral and organizational outcomes also align with evaluations conducted before the pandemic, which have evaluated aspects such as behavior change, cost-effectiveness, and clinical and patient-centered outcomes [185]. Regarding clinical outcomes, previous evidence has shown clinical efficacy in the use of digital health interventions for a variety of patient groups [186,187], highlighting its relevance for remote care provision. Although less frequently found, evaluations conducted with the aim of exploring long-term implementation underscore the critical need for addressing sustainability in digital health implementation [188].

Limitations

This review has applied a comprehensive and systematic methodological approach. However, some limitations are worth noting. Although the inclusion of solely peer-reviewed publications helped ensure rigor in the synthesized data, evidence reported through other means has not been included. Additionally, a publication bias may exist toward positive experiences with digital health, and less successful practices may be underrepresented. The variation in study designs, alongside the diversity of health care contexts and patient

XSL•FO
RenderX

groups, may have influenced the interpretation of the findings. Future research may provide additional insights into which specific patient groups may benefit more from the implementation of hybrid care.

This review limited its scope to modifications directly related to service provision; however, other types of interventions, such as those for interprofessional collaboration or patient education, may also be valuable for improving timely access to high-quality care. Strategies to improve remote communication between primary and secondary care providers, as well as the implementation of remote multidisciplinary meetings, have shown potential to improve the organization of care and may be a focus for future research [34,42,75,80,99,121].

Moreover, differences in the use of terminology for digital health greatly contribute to diverse reporting of systems, interventions, and functions and may have influenced the reporting of results. Future research can benefit from using common terminology frameworks, increasing consistency in reporting and helping improve future understanding of the implementation and evaluation of digital and hybrid care. Such common frameworks may also be essential for the development of legal and regulatory procedures [180].

Conclusions

This review synthesizes findings essential for understanding the scope of hybrid care modifications used in the delivery of nonpandemic care and implemented throughout the COVID-19 pandemic. It provides a unique view as it analyzes digital health interventions as part of broader care pathways, which combine different types of care delivery implemented to sustain access to care. Integrating digital health as part of hybrid care models is essential in designing care pathways that can adapt to different contexts, including future health emergency situations.

The use of digital health and emerging technologies such as AI will continue to shape the future of health care delivery. As evidence regarding its applications advances, research should identify which combinations of digital interventions and care functions offer the greatest accessibility and cost-effectiveness. In the future, research may expand the application of hybrid care models to innovative strategies for the effective management of any crisis involving disrupted infrastructure. Policymakers, researchers, and implementers should also prioritize strategies to reduce inequities in access to digital and in-person care, particularly for underserved populations. Finally, evaluations should move beyond short-term user experiences and assess long-term implementation, clinical outcomes, and sustainability of hybrid care models.

Acknowledgments

The authors wish to thank Dilek Yildirim for her assistance with data extraction and Alice Tillema for her assistance with the development of the search strategy. Preliminary findings from this study were previously presented as abstracts at the 14th European Congress on Tropical Medicine and International Health 2025 and at the 18th European Public Health Conference 2025 [189,190]. These abstracts included preliminary results and analyses concerning a subset of the reviewed articles, which differ from the final results and analyses described in this study. No generative AI was used in the preparation of this manuscript.

Funding

This study was funded by the European Union. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. The funder had no involvement in the study design, data collection, analysis, interpretation, or writing of the manuscript.

Data Availability

Data generated during this study are included in this published article and its supplementary information files. Additional datasets used for analysis during this study are available from the corresponding author on reasonable request.

Authors' Contributions

AT, CPR, MAC, MLS, TVL conceptualized the review. NSV developed the search strategy, searched and screened studies, extracted and analyzed data, and wrote the first draft of the manuscript. TVL screened studies, extracted data, and analyzed data. LN assisted in developing the search strategy and screened studies. AT, MLS, TVL, and CPR supervised the review process and assisted in preparation of the draft for submission. All authors contributed to the interpretation of results and revisions of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOC File, 86 KB - jmir_v28i1e84756_app1.doc]

XSL•FO
RenderX

Multimedia Appendix 2
Extracted data of included studies.
[DOCX File, 99 KB - jmir_v28i1e84756_app2.docx]

Checklist 1
PRISMA-ScR checklist.
[DOC File, 65 KB - jmir_v28i1e84756_app3.doc]

Checklist 2
PRISMA-S checklist.
[DOCX File, 17 KB - jmir_v28i1e84756_app4.docx]

References

1. Third round of the global pulse survey on continuity of essential health services during the COVID-19 pandemic. : World Health Organization; 2022 URL: https://iris.who.int/server/api/core/bitstreams/057fc218-13ce-4f50-8103-eab552b8e02c/content [accessed 2026-04-01]
2. The COVID-19 pandemic and the future of telemedicine. : OECD Publishing; 2023 URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/01/the-covid-19-pandemic-and-the-future-of-telemedicine_1c878192/ac8b0a27-en.pdf [accessed 2026-04-01]
3. Classification of digital interventions, services and applications in health: a shared language to describe the uses of digital technology for health. : World Health Organization; 2023 URL: https://iris.who.int/server/api/core/bitstreams/bb20354e-ebee-44aa-8a58-7b1da85c6f95/content [accessed 2026-04-01]
4. Golinelli D, Boetto E, Carullo G, Nuzzolese AG, Landini MP, Fantini MP. Adoption of digital technologies in health care during the COVID-19 pandemic: systematic review of early scientific literature. J Med Internet Res 2020 Nov 6;22(11):e22280. [doi: 10.2196/22280] [Medline: 33079693]
5. Gunasekeran DV, Tseng RMWW, Tham YC, Wong TY. Applications of digital health for public health responses to COVID-19: a systematic scoping review of artificial intelligence, telehealth and related technologies. NPJ Digit Med 2021 Feb 26;4(1):40. [doi: 10.1038/s41746-021-00412-9] [Medline: 33637833]
6. Garavand A, Khodaveisi T, Aslani N, Hosseiniravandi M, Shams R, Behmanesh A. Telemedicine in cancer care during COVID-19 pandemic: a systematic mapping study. Health Technol 2023 Jul;13(4):665-678. [doi: 10.1007/s12553-023-00762-2]
7. Severino P, Prosperi S, D'Amato A, et al. Telemedicine: an effective and low-cost lesson from the COVID-19 pandemic for the management of heart failure patients. Curr Heart Fail Rep 2023 Oct;20(5):382-389. [doi: 10.1007/s11897-023-00624-y] [Medline: 37665424]
8. Tossaint-Schoenmakers R, Versluis A, Chavannes N, Talboom-Kamp E, Kasteleyn M. The challenge of integrating ehealth into health care: systematic literature review of the Donabedian model of structure, process, and outcome. J Med Internet Res 2021 May 10;23(5):e27180. [doi: 10.2196/27180] [Medline: 33970123]
9. Pilosof NP, Barrett M, Oborn E, Barkai G, Zimlichman E, Segal G. Designing for flexibility in hybrid care services: lessons learned from a pilot in an internal medicine unit. Front Med Technol 2023;5:1223002. [doi: 10.3389/fmedt.2023.1223002] [Medline: 38053662]
10. van Steenkiste J, Verberk-Jonkers I, de Koning S, Voss-de Haan J, de Jong-Verhagen B, Dohmen D. Patient engagement in a hybrid care pathway for hypertension: not one size fits all. J Patient Exp 2024;11:23743735241297626. [doi: 10.1177/23743735241297626] [Medline: 39654654]
11. Tossaint-Schoenmakers R, Kasteleyn MJ, Rauwerdink A, Chavannes N, Willems S, Talboom-Kamp EPWA. Development of a quality management model and self-assessment questionnaire for hybrid health care: concept mapping study. JMIR Form Res 2022 Jul 7;6(7):e38683. [doi: 10.2196/38683] [Medline: 35797097]
12. Walter KH, Glassman LH, Levine JA, et al. Telebehavioral health, in-person, and hybrid modalities of treatment delivery among US service members: longitudinal observational study. JMIR Ment Health 2026 Jan 5;13:e83809. [doi: 10.2196/83809] [Medline: 41490370]
13. Bola H, Rai A, Penumaka R, Ulucay E, Levin E, Maron D. Cardiac rehabilitation for coronary artery disease: gaps, digital models, and the future of personalized prevention. Am J Cardiol 2026 Mar 1;262:16-27. [doi: 10.1016/j.amjcard.2025.12.013] [Medline: 41468986]
14. Zakaria H, Paul G, Debs J, et al. Impact of virtual patient engagement on glycemic control in type 2 diabetes: a retrospective observational study from GluCare hybrid care model. Front Endocrinol 2025;16. [doi: 10.3389/fendo.2025.1695381]
15. Fan S, Imsirovic H, Cooper C. Telemedicine and hybrid engagement models facilitate hepatitis C cascade of care during periods of healthcare disruption. Ann Hepatol 2025 Dec 24;31(1):102178. [doi: 10.1016/j.aohep.2025.102178] [Medline: 41453643]

XSL•FO
RenderX

16. Hatef E, Wilson RF, Zhang A, et al. Effectiveness of telehealth versus in-person care during the COVID-19 pandemic: a systematic review. NPJ Digit Med 2024 Jun 15;7(1):157. [doi: 10.1038/s41746-024-01152-2] [Medline: 38879682]
17. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
18. Sanchez Villalobos N, van Loenen T, Ngongalah L, et al. Hybrid care modifications in the delivery of non-pandemic care during the COVID-19 pandemic: a scoping review protocol. OSF. URL: https://doi.org/10.17605/OSF.IO/MDRAS [accessed 2026-04-24]
19. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
20. Hong QN, Pluye P, Fàbregues S, et al. Mixed Methods Appraisal Tool (MMAT). PBWorks. 2018. URL: http://mixedmethodsappraisaltoolpublic.pbworks.com/ [accessed 2026-04-09]
21. Kirkpatrick D, Kirkpatrick J. Evaluating Training Programs: The Four Levels, 3rd edition: Berrett-Koehler Publishers, Inc; 2006. URL: https://books.google.co.in/books/about/Evaluating_Training_Programs.html?id=BJ4QCmvP5rcC&redir_esc=y [accessed 2026-04-01]
22. Fuentes B, Alonso de Leciñana M, Calleja-Castaño P, et al. Impact of the COVID-19 pandemic on the organisation of stroke care. Madrid Stroke Care Plan. Neurologia (Engl Ed) 2020;35(6):363-371. [doi: 10.1016/j.nrl.2020.05.007] [Medline: 32563566]
23. Noé E, Navarro MD, Amorós D, et al. Effectiveness, adherence and usability of a teleneurorehabilitation programme to ensure continuity of care for patients with acquired brain injury during the COVID-19 pandemic [Article in Spanish]. Rev Neurol 2021 Nov 16;73(10):345-350. [doi: 10.33588/rn.7310.2021275] [Medline: 34755887]
24. Soler MJ, Macia Heras M, Ortiz A, del Pino y Pino MD, Salgueira Lazo M. Impacto de la pandemia COVID-19 en los servicios de Nefrología españoles [Article in Spanish]. Nefrología 2020 Nov;40(6):579-584. [doi: 10.1016/j.nefro.2020.08.002]
25. Gomes-de Almeida S, Marabujo T, do Carmo-Gonçalves M. Telemedicine satisfaction of primary care patients during COVID-19 pandemics [Article in Spanish]. Semergen 2021;47(4):248-255. [doi: 10.1016/j.semerg.2021.01.005] [Medline: 33781673]
26. Ohnleiter T, Piot L, Rogenmuser A, Noirclerc M, Hamlaoui R, Grandgirard A. Management of a radiotherapy center during the COVID-19 outbreak: the experience of the Mulhouse Hospital Centre (France) [Article in French]. Cancer Radiother 2020 Jun;24(3):188-193. [doi: 10.1016/j.canrad.2020.04.002] [Medline: 32334905]
27. Viegas L, Dupie I, Rigal L, et al. Triage of patients and remote consultations in primary care facilities during the COVID-19 pandemic in France (PRICOV-19 study) [Article in French]. Sante Publique 2023 Dec 11;35(4):393-403. [doi: 10.3917/spub.234.0393] [Medline: 38078634]
28. Bos WH, van Tubergen A, Vonkeman HE. Telemedicine for patients with rheumatic and musculoskeletal diseases during the COVID-19 pandemic; a positive experience in the Netherlands. Rheumatol Int 2021 Mar;41(3):565-573. [doi: 10.1007/s00296-020-04771-6] [Medline: 33449162]
29. De Marchi F, Sarnelli MF, Serioli M, et al. Telehealth approach for amyotrophic lateral sclerosis patients: the experience during COVID-19 pandemic. Acta Neurol Scand 2021 May;143(5):489-496. [doi: 10.1111/ane.13373] [Medline: 33185886]
30. Faria G, Onubogu IK, Tadros BJ, Relwani J. Change in practice due to COVID-19 - early experiences of a United Kingdom district general hospital in trauma & orthopaedics. J Orthop 2020;22:288-290. [doi: 10.1016/j.jor.2020.06.004] [Medline: 32565644]
31. Kilduff CL, Thomas AA, Dugdill J, et al. Creating the Moorfields' virtual eye casualty: video consultations to provide emergency teleophthalmology care during and beyond the COVID-19 pandemic. BMJ Health Care Inform 2020 Aug;27(3):e100179. [doi: 10.1136/bmjhci-2020-100179] [Medline: 32796085]
32. Ambrosini F, Di Stasio A, Mantica G, Cavallone B, Serao A. COVID-19 pandemic and uro-oncology follow-up: a "virtual" multidisciplinary team strategy and patients' satisfaction assessment. Arch Ital Urol Androl 2020 Jun 23;92(2). [doi: 10.4081/aiua.2020.2.78] [Medline: 32597103]
33. Bourdon H, Jaillant R, Ballino A, et al. Teleconsultation in primary ophthalmic emergencies during the COVID-19 lockdown in Paris: experience with 500 patients in March and April 2020. J Fr Ophtalmol 2020 Sep;43(7):577-585. [doi: 10.1016/j.jfo.2020.05.005] [Medline: 32564983]
34. Rodler S, Apfelbeck M, Schulz GB, et al. Telehealth in uro-oncology beyond the pandemic: toll or lifesaver? Eur Urol Focus 2020 Sep 15;6(5):1097-1103. [doi: 10.1016/j.euf.2020.05.010] [Medline: 32534969]
35. Barsom EZ, Meijer HAW, Blom J, Schuuring MJ, Bemelman WA, Schijven MP. Emergency upscaling of video consultation during the COVID-19 pandemic: contrasting user experience with data insights from the electronic health record in a large academic hospital. Int J Med Inform 2021 Jun;150(104463):104463. [doi: 10.1016/j.ijmedinf.2021.104463] [Medline: 33872824]
36. Brunasso AMG, Massone C. Teledermatologic monitoring for chronic cutaneous autoimmune diseases with smartworking during COVID-19 emergency in a tertiary center in Italy. Dermatol Ther 2020 Jul;33(4):e13495. [doi: 10.1111/dth.13695] [Medline: 32458588]

XSL•FO
RenderX

37. Gabriel C, Mathiot A, Boumediane M, et al. Organization of outpatient consultations at a hand surgery department in a French university hospital during the COVID-19 lockdown. Hand Surg Rehabil 2021 Feb;40(1):17-24. [doi: 10.1016/j.hansur.2020.10.006] [Medline: 33130022]
38. Geerdink TH, Salentijn DA, de Vries KA, et al. Optimizing orthopedic trauma care delivery during the COVID-19 pandemic. a closed-loop audit of implementing a virtual fracture clinic and fast-track pathway in a Dutch level 2 trauma center. Trauma Surg Acute Care Open 2021;6(1):e000691. [doi: 10.1136/tsaco-2021-000691] [Medline: 34632079]
39. Jácome C, Pereira AM, Amaral R, et al. The use of remote care during the coronavirus disease 2019 pandemic: a perspective of Portuguese and Spanish physicians. Eur Ann Allergy Clin Immunol 2022 Jan;54(1):25-29. [doi: 10.23822/EurAnnACI.1764-1489.184] [Medline: 33354962]
40. Legrottaglie EF, Balia L, Camesasca FI, et al. Management of an ophthalmology department during COVID-19 pandemic in Milan, Italy. Eur J Ophthalmol 2021 Sep;31(5):2259-2267. [doi: 10.1177/1120672120960334] [Medline: 32962415]
41. Mugnai G, Volpiana A, Cavedon S, Paolini C, Perrone C, Bilato C. Boosting telemedicine through remote monitoring of cardiac electronic devices during the Italian COVID-19 outbreak. Cardiol J 2021;28(2):336-338. [doi: 10.5603/CJ.a2021.0011] [Medline: 33634839]
42. O'Donovan M, Buckley C, Benson J, et al. Telehealth for delivery of haemophilia comprehensive care during the COVID-19 pandemic. Haemophilia 2020 Nov;26(6):984-990. [doi: 10.1111/hae.14156] [Medline: 32997849]
43. Perrin A, Mainard N, Limousin M, et al. Satisfaction and feasibility of videoconsultation (VC) in orthopaedic and trauma surgery in the context of the COVID-19 pandemic: prospective study of 783 patients. Orthop Traumatol Surg Res 2023 Feb;109(1):103345. [doi: 10.1016/j.otsr.2022.103345] [Medline: 35671985]
44. Ramaswami U, D'Amore S, Finnegan N, Hughes D, Kazemi M, Lysosomal Disorders Team, Royal Free London NHS Foundation Trust. Impact of SARS-CoV-2 (COVID-19) pandemic on patients with lysosomal storage disorders and restoration of services: experience from a specialist centre. Intern Med J 2021 Oct;51(10):1580-1593. [doi: 10.1111/imj.15473] [Medline: 34487419]
45. Rossi B, Zoccali C, Baldi J, et al. Reorganization tips from a sarcoma unit at time of the COVID-19 pandemic in Italy: early experience from a Regional Referral Oncologic Center. J Clin Med 2020 Jun 15;9(6):1868. [doi: 10.3390/jcm9061868] [Medline: 32549298]
46. Sacchi C, Andersson K, Roczniewska M, et al. Mind the gap: analysis of two pilot projects of a home telehealth service for persons with complex conditions in a Swedish hospital. BMC Health Serv Res 2023 May 9;23(1):463. [doi: 10.1186/s12913-023-09409-4] [Medline: 37161458]
47. Schnoor K, Versluis A, Chavannes NH, Talboom-Kamp EPWA. The usability of Homelab, a digital self-service at a Dutch general practice, for diagnostic tests: pilot study with a questionnaire. JMIR Form Res 2023 Jan 26;7:e42151. [doi: 10.2196/42151] [Medline: 36701183]
48. Świerad M, Dyrbuś K, Szkodziński J, Zembala MO, Kalarus Z, Gąsior M. Telehealth visits in a tertiary cardiovascular center as a response of the healthcare system to the Severe Acute Respiratory Syndrome Coronavirus 2 pandemic in Poland. Pol Arch Intern Med 2020 Aug 27;130(7-8):700-703. [doi: 10.20452/pamw.15370] [Medline: 32426953]
49. Testa S, Mayora-Ibarra O, Piras EM, et al. Implementation of tele visit healthcare services triggered by the COVID-19 emergency: the Trentino Province experience. Z Gesundh Wiss 2022;30(1):77-92. [doi: 10.1007/s10389-021-01609-8] [Medline: 34150467]
50. van de Haar J, Hoes LR, Coles CE, et al. Caring for patients with cancer in the COVID-19 era. Nat Med 2020 May;26(5):665-671. [doi: 10.1038/s41591-020-0874-8] [Medline: 32405058]
51. Wanat M, Hoste M, Gobat N, et al. Transformation of primary care during the COVID-19 pandemic: experiences of healthcare professionals in eight European countries. Br J Gen Pract 2021 Aug;71(709):e634-e642. [doi: 10.3399/BJGP.2020.1112] [Medline: 33979303]
52. Wherton J, Greenhalgh T, Shaw SE. Expanding video consultation services at pace and scale in Scotland during the COVID-19 pandemic: national mixed methods case study. J Med Internet Res 2021 Oct 7;23(10):e31374. [doi: 10.2196/31374] [Medline: 34516389]
53. Willems LM, Balcik Y, Noda AH, et al. SARS-CoV-2-related rapid reorganization of an epilepsy outpatient clinic from personal appointments to telemedicine services: a German single-center experience. Epilepsy Behav 2020 Nov;112:107483. [doi: 10.1016/j.yebeh.2020.107483] [Medline: 33181898]
54. Klain M, Nappi C, Maurea S, et al. Management of differentiated thyroid cancer through nuclear medicine facilities during COVID-19 emergency: the telemedicine challenge. Eur J Nucl Med Mol Imaging 2021 Mar;48(3):831-836. [doi: 10.1007/s00259-020-05041-0] [Medline: 32965559]
55. Luengo-Alonso G, Pérez-Tabernero FGS, Tovar-Bazaga M, Arguello-Cuenca JM, Calvo E. Critical adjustments in a department of orthopaedics through the COVID-19 pandemic. Int Orthop 2020 Aug;44(8):1557-1564. [doi: 10.1007/s00264-020-04647-1] [Medline: 32474718]
56. Piro A, Magnocavallo M, Della Rocca DG, et al. Management of cardiac implantable electronic device follow-up in COVID-19 pandemic: lessons learned during Italian lockdown. J Cardiovasc Electrophysiol 2020 Nov;31(11):2814-2823. [doi: 10.1111/jce.14755] [Medline: 32954600]

XSL•FO
RenderX

57. Omboni S, Ballatore T, Rizzi F, Tomassini F, Panzeri E, Campolo L. Telehealth at scale can improve chronic disease management in the community during a pandemic: an experience at the time of COVID-19. PLoS ONE 2021;16(9):e0258015. [doi: 10.1371/journal.pone.0258015] [Medline: 34587198]
58. Muthiah S, Craig FE, Sinclair S, et al. Rapid expansion of a teledermatology web application for digital dermatology assessment necessitated by the COVID-19 pandemic: retrospective evaluation. JMIR Dermatol 2023 Jul 26;6:e36307. [doi: 10.2196/36307] [Medline: 37632929]
59. Hardman JC, Tikka T, Paleri V, ENT UK, BAHNO and INTEGRATE (The UK ENT Trainee Research Network). Remote triage incorporating symptom-based risk stratification for suspected head and neck cancer referrals: a prospective population-based study. Cancer 2021 Nov 15;127(22):4177-4189. [doi: 10.1002/cncr.33800] [Medline: 34411287]
60. Chana M, Muse S, Ball S, Bennett R, McCarthy R. Critical limb ischaemia in the time of COVID-19: establishing ambulatory service provision. Ann R Coll Surg Engl 2022 Nov;104(9):673-677. [doi: 10.1308/rcsann.2021.0294] [Medline: 34941433]
61. Bonalumi G, Giambuzzi I, Barbone A, et al. A call to action becomes practice: cardiac and vascular surgery during the COVID-19 pandemic based on the Lombardy emergency guidelines. Eur J Cardiothorac Surg 2020 Aug 1;58(2):319-327. [doi: 10.1093/ejcts/ezaa204] [Medline: 32584978]
62. Mazzone P, Peretto G, Radinovic A, et al. The COVID-19 challenge to cardiac electrophysiologists: optimizing resources at a referral center. J Interv Card Electrophysiol 2020;59(2):321-327. [doi: 10.1007/s10840-020-00761-7] [Medline: 32425656]
63. Montesi G, Di Biase S, Chierchini S, et al. Radiotherapy during COVID-19 pandemic. How to create a no fly zone: a Northern Italy experience. Radiol Med 2020 Jun;125(6):600-603. [doi: 10.1007/s11547-020-01217-8] [Medline: 32415473]
64. Mathai NJ, Venkatesan AS, Key T, Wilson C, Mohanty K. COVID-19 and orthopaedic surgery: evolving strategies and early experience. Bone Jt Open 2020 May;1(5):160-166. [doi: 10.1302/2633-1462.15.BJO-2020-0021.R1] [Medline: 33241227]
65. Morgan C, Ahluwalia AK, Aframian A, Li L, Sun SNM. The impact of the novel coronavirus on trauma and orthopaedics in the UK. Br J Hosp Med 2020 Apr 2;81(4):1-6. [doi: 10.12968/hmed.2020.0137]
66. Butler D, Davies-Husband C, Dhanda J, et al. Head and neck oncological ablation and reconstruction in the COVID-19 era - our experience to date. Br J Oral Maxillofac Surg 2020 Oct;58(8):1008-1013. [doi: 10.1016/j.bjoms.2020.06.011] [Medline: 32576467]
67. Denti M, Pecorari A, Accogli MA, et al. Facing the COVID-19 pandemic: an Italian feasibility study of a mixed in-person/telerehabilitation intervention for cancer patients. Cancer Med 2024 Aug;13(15):e70022. [doi: 10.1002/cam4.70022] [Medline: 39095954]
68. Pignatti M, Pinto V, Miralles MEL, Giorgini FA, Cannamela G, Cipriani R. How the COVID-19 pandemic changed the plastic surgery activity in a regional referral center in Northern Italy. J Plast Reconstr Aesthet Surg 2020 Jul;73(7):1348-1356. [doi: 10.1016/j.bjps.2020.05.002] [Medline: 32499187]
69. Ceccato F, Voltan G, Sabbadin C, et al. Tele-medicine versus face-to-face consultation in endocrine outpatients clinic during COVID-19 outbreak: a single-center experience during the lockdown period. J Endocrinol Invest 2021 Aug;44(8):1689-1698. [doi: 10.1007/s40618-020-01476-2] [Medline: 33355915]
70. Caravatta L, Rosa C, Di Sciascio MB, et al. COVID-19 and radiation oncology: the experience of a two-phase plan within a single institution in central Italy. Radiat Oncol 2020 Sep 29;15(1):226. [doi: 10.1186/s13014-020-01670-9] [Medline: 32993690]
71. Carter E, Currie CC, Asuni A, et al. The first six weeks - setting up a UK urgent dental care centre during the COVID-19 pandemic. Br Dent J 2020 Jun;228(11):842-848. [doi: 10.1038/s41415-020-1708-2] [Medline: 32541745]
72. Collins PM, Madden A, O'Connell C, et al. Urological service provision during the COVID-19 period: the experience from an Irish tertiary centre. Ir J Med Sci 2021 May;190(2):455-460. [doi: 10.1007/s11845-020-02352-x] [Medline: 32856269]
73. Crusz SM, Hall PE, Earwicker K, et al. Providing an acute oncology service during the COVID-19 pandemic. Clin Med (Lond) 2021 Sep;21(5):e548-e551. [doi: 10.7861/clinmed.2020-0693] [Medline: 34385297]
74. Kenis I, Theys S, Hermie E, Foulon V, Van Hecke A. Impact of COVID-19 on the organization of cancer care in Belgium: lessons learned for the (post-)pandemic future. Int J Environ Res Public Health 2022 Sep 30;19(19):12456. [doi: 10.3390/ijerph191912456] [Medline: 36231756]
75. Nuñez JH, Porcel JA, Pijoan J, et al. Rethinking trauma hospital services in one of Spain's largest university hospitals during the COVID-19 pandemic. How can we organize and help? Our experience. Injury 2020 Dec;51(12):2827-2833. [doi: 10.1016/j.injury.2020.09.055] [Medline: 33004206]
76. Rajasekaran RB, Whitwell D, Cosker TDA, Gibbons CLMH. Service delivery during the COVID-19 pandemic: experience from the Oxford Bone Tumour and Soft Tissue Sarcoma service. J Clin Orthop Trauma 2020 Jul;11(Suppl 4):S419-S422. [doi: 10.1016/j.jcot.2020.05.035] [Medline: 32774004]
77. Shah A, Bryant C, Patel J, Tagar H, Akintola D, Obisesan O. COVID-19: establishing an oral surgery-led urgent dental care hub. Br Dent J 2020 Jun;228(12):957-963. [doi: 10.1038/s41415-020-1713-5] [Medline: 32591712]
78. Smrke A, Younger E, Wilson R, et al. Telemedicine during the COVID-19 pandemic: impact on care for rare cancers. JCO Glob Oncol 2020 Jul;6(6):1046-1051. [doi: 10.1200/GO.20.00220] [Medline: 32639877]

XSL•FO
RenderX

79. Paleri V, Hardman J, Tikka T, Bradley P, Pracy P, Kerawala C. Rapid implementation of an evidence-based remote triaging system for assessment of suspected referrals and patients with head and neck cancer on follow-up after treatment during the COVID-19 pandemic: model for international collaboration. Head Neck 2020 Jul;42(7):1674-1680. [doi: 10.1002/hed.26219] [Medline: 32374942]
80. Scaldaferri F, Pugliese D, Privitera G, et al. Impact of COVID-19 pandemic on the daily management of biotechnological therapy in inflammatory bowel disease patients: reorganisational response in a high-volume Italian inflammatory bowel disease centre. United European Gastroenterol J 2020 Aug;8(7):775-781. [doi: 10.1177/2050640620929133] [Medline: 32438878]
81. Anderson H, Scantlebury A, Galdas P, Adamson J. Remote and technology-mediated working during the COVID-19 pandemic: a qualitative exploration of the experiences of nurses working in general practice (the GenCo Study). J Adv Nurs 2024 Apr;80(4):1592-1606. [doi: 10.1111/jan.15921] [Medline: 37909600]
82. Miceli L, Dal Mas F, Biancuzzi H, et al. Doctor@Home: through a telemedicine co-production and co-learning journey. J Cancer Educ 2022 Aug;37(4):1236-1238. [doi: 10.1007/s13187-020-01945-5] [Medline: 33442862]
83. Rayo MN, Fernández-Buhigas I, Ferrer E, Arrébola M, Gil MM, Santacruz B. Application of a new protocol for providing obstetric care in an outpatient service during the COVID-19 pandemic in a public hospital in Madrid, Spain. Front Med (Lausanne) 2022;9(902640):902640. [doi: 10.3389/fmed.2022.902640] [Medline: 35991653]
84. Vassallo C, Scerri Harney A, Abela S. An audit on the provision of telemedicine in primary care in Malta during the COVID-19 pandemic. Malta Med J 2024;36(1):34-44 [FREE Full text]
85. Del Hoyo J, Millán M, Garrido-Marín A, et al. Changes in the management of IBD patients since the onset of COVID-19 pandemic. A path toward the implementation of telemedicine in Spain? [Article in English, Spanish]. Gastroenterol Hepatol 2022 Nov;45(9):697-705. [doi: 10.1016/j.gastrohep.2021.08.006] [Medline: 34508808]
86. Joy M, McGagh D, Jones N, et al. Reorganisation of primary care for older adults during COVID-19: a cross-sectional database study in the UK. Br J Gen Pract 2020 Aug;70(697):e540-e547. [doi: 10.3399/bjgp20X710933] [Medline: 32661009]
87. Pascual J, Mazuecos A, Sánchez-Antolín G, et al. Best practices during COVID-19 pandemic in solid organ transplant programs in Spain. Transplant Rev (Orlando) 2023 Jan;37(1):100749. [doi: 10.1016/j.trre.2023.100749] [Medline: 36889117]
88. Luciani LG, Mattevi D, Cai T, Giusti G, Proietti S, Malossini G. Teleurology in the time of COVID-19 pandemic: here to stay? Urology 2020 Jun;140:4-6. [doi: 10.1016/j.urology.2020.04.004] [Medline: 32298686]
89. Stansfield J, Dobbs S, Harrison R, et al. Management of ENT emergencies during the coronavirus disease 2019 pandemic. J Laryngol Otol 2021 Feb;135(2):117-124. [doi: 10.1017/S0022215121000530] [Medline: 33612142]
90. Nebsbjerg MA, Vestergaard CH, Bomholt KB, Christensen MB, Huibers L. Use of video in telephone triage in out-of-hours primary care: register-based study. JMIR Med Inform 2024 Apr 4;12:e47039. [doi: 10.2196/47039] [Medline: 38596835]
91. Minniti A, Maglione W, Pignataro F, Cappadona C, Caporali R, Del Papa N. Taking care of systemic sclerosis patients during COVID-19 pandemic: rethink the clinical activity. Clin Rheumatol 2020 Jul;39(7):2063-2065. [doi: 10.1007/s10067-020-05191-4] [Medline: 32462423]
92. Chappell P, Dias A, Bakhai M, Ledger J, Clarke GM. How is primary care access changing? A retrospective, repeated cross-sectional study of patient-initiated demand at general practices in England using a modern access model, 2019-2022. BMJ Open 2023 Aug 17;13(8):e072944. [doi: 10.1136/bmjopen-2023-072944] [Medline: 37591638]
93. Saibeni S, Scucchi L, Dragoni G, et al. Activities related to inflammatory bowel disease management during and after the coronavirus disease 2019 lockdown in Italy: how to maintain standards of care. United European Gastroenterol J 2020 Dec;8(10):1228-1235. [doi: 10.1177/2050640620964132] [Medline: 33070758]
94. Fieux M, Duret S, Bawazeer N, Denoix L, Zaouche S, Tringali S. Telemedicine for ENT: effect on quality of care during COVID-19 pandemic. Eur Ann Otorhinolaryngol Head Neck Dis 2020 Sep;137(4):257-261. [doi: 10.1016/j.anorl.2020.06.014] [Medline: 32624390]
95. Crowley D, Delargy I. A national model of remote care for assessing and providing opioid agonist treatment during the COVID-19 pandemic: a report. Harm Reduct J 2020 Jul 17;17(1):49. [doi: 10.1186/s12954-020-00394-z] [Medline: 32680520]
96. El Moazen G, Pfeifer B, Loid A, Kastner P, Ciardi C. The effectiveness of telemedical monitoring program DiabCare Tirol for patients with gestational diabetes mellitus. Stud Health Technol Inform 2021 Oct 27;285(205-10):205-210. [doi: 10.3233/SHTI210599] [Medline: 34734875]
97. Casella D, Fusario D, Cassetti D, et al. The patient's pathway for breast cancer in the COVID-19 era: an Italian single-center experience. Breast J 2020 Aug;26(8):1589-1592. [doi: 10.1111/tbj.13958] [Medline: 32596965]
98. Sheil O, McAuliffe FM. Reorganisation of obstetric services during the COVID pandemic – experience from National Maternity Hospital Dublin Ireland. Best Pract Res Clin Obstet Gynaecol 2021 Jun;73:104-112. [doi: 10.1016/j.bpobgyn.2021.03.013]
99. Murris F, Huchon C, Zilberman S, et al. Impact of the first lockdown for coronavirus 19 on breast cancer management in France: a multicentre survey. J Gynecol Obstet Hum Reprod 2021 Nov;50(9):102166. [doi: 10.1016/j.jogoh.2021.102166] [Medline: 34033966]

XSL•FO
RenderX

100. Altmann P, Leutmezer F, Ponleitner M, et al. Remote visits for people with multiple sclerosis during the COVID-19 pandemic in Austria: the TELE MS randomized controlled trial. Digit Health 2022;8:20552076221112154. [doi: 10.1177/20552076221112154] [Medline: 35847524]
101. Benaque A, Gurruchaga MJ, Abdelnour C, et al. Dementia care in times of COVID-19: experience at Fundació ACE in Barcelona, Spain. J Alzheimers Dis 2020;76(1):33-40. [doi: 10.3233/JAD-200547] [Medline: 32538856]
102. Pareyson D, Pantaleoni C, Eleopra R, et al. Neuro-telehealth for fragile patients in a tertiary referral neurological institute during the COVID-19 pandemic in Milan, Lombardy. Neurol Sci 2021 Jul;42(7):2637-2644. [doi: 10.1007/s10072-021-05252-9] [Medline: 33929645]
103. Muschol J, Heinrich M, Heiss C, et al. Digitization of follow-up care in orthopedic and trauma surgery with video consultations: health economic evaluation study from a health provider's perspective. J Med Internet Res 2023 Dec 25;25:e46714. [doi: 10.2196/46714] [Medline: 38145481]
104. Dunkerley S, Thelwall C, Omiawele J, Smith A, Deo S, Lowdon I. Patient care modifications and hospital regulations during the COVID-19 crisis created inequality and functional hazard for patients with orthopaedic trauma. Int Orthop 2020 Dec;44(12):2481-2485. [doi: 10.1007/s00264-020-04764-x] [Medline: 32767088]
105. Papoutsi C, Shaw S, Greenhalgh T. Implementing video group consultations in general practice during COVID-19: a qualitative study. Br J Gen Pract 2022 Jul;72(720):e483-e491. [doi: 10.3399/BJGP.2021.0673] [Medline: 35636969]
106. Silsand L, Severinsen GH, Berntsen G. Preservation of person-centered care through videoconferencing for patient follow-up during the COVID-19 pandemic: case study of a multidisciplinary care team. JMIR Form Res 2021 Mar 5;5(3):e25220. [doi: 10.2196/25220] [Medline: 33646965]
107. Lapão LV, Peyroteo M, Maia M, et al. Implementation of digital monitoring services during the COVID-19 pandemic for patients with chronic diseases: design science approach. J Med Internet Res 2021 Aug 26;23(8):e24181. [doi: 10.2196/24181] [Medline: 34313591]
108. Stewart J, McCorry N, Reid H, Hart N, Kee F. Implementation of remote asthma consulting in general practice in response to the COVID-19 pandemic: an evaluation using extended Normalisation Process Theory. BJGP Open 2022 Mar;6(1):BJGPO.2021.0189. [doi: 10.3399/BJGPO.2021.0189] [Medline: 34862167]
109. Murphy M, Scott LJ, Salisbury C, et al. Implementation of remote consulting in UK primary care following the COVID-19 pandemic: a mixed-methods longitudinal study. Br J Gen Pract 2021;71(704):e166-e177. [doi: 10.3399/BJGP.2020.0948] [Medline: 33558332]
110. Békási S, Girasek E, Győrffy Z. Telemedicine in community shelters: possibilities to improve chronic care among people experiencing homelessness in Hungary. Int J Equity Health 2022 Dec 17;21(1):181. [doi: 10.1186/s12939-022-01803-4] [Medline: 36528777]
111. Asomugha AU, Pakai A. Trends and shifts in Swedish telemedicine consultations during the pre-COVID-19, COVID-19, and post-COVID-19 periods: retrospective observational study. JMIR Form Res 2025 May 16;9:e60294. [doi: 10.2196/60294] [Medline: 40378415]
112. Corea F, Ciotti S, Cometa A, et al. Telemedicine during the coronavirus disease (COVID-19) pandemic: a multiple sclerosis (MS) outpatients service perspective. Neurol Int 2021 Jan 18;13(1):25-31. [doi: 10.3390/neurolint13010003] [Medline: 33477432]
113. Giudice A, Barone S, Muraca D, et al. Can teledentistry improve the monitoring of patients during the COVID-19 dissemination? A descriptive pilot study. Int J Environ Res Public Health 2020 May 13;17(10):3399. [doi: 10.3390/ijerph17103399] [Medline: 32414126]
114. Salzano A, D'Assante R, Stagnaro FM, et al. Heart failure management during the COVID-19 outbreak in Italy: a telemedicine experience from a heart failure university tertiary referral centre. Eur J Heart Fail 2020 Jun;22(6):1048-1050. [doi: 10.1002/ejhf.1911] [Medline: 32463534]
115. Peretto G, De Luca G, Campochiaro C, et al. Telemedicine in myocarditis: evolution of a mutidisciplinary "disease unit" at the time of COVID-19 pandemic. Am Heart J 2020 Nov;229:121-126. [doi: 10.1016/j.ahj.2020.07.015] [Medline: 32957030]
116. Russo V, Cassini R, Caso V, et al. Nursing teleconsultation for the outpatient management of patients with cardiovascular disease during COVID-19 pandemic. Int J Environ Res Public Health 2021 Feb 21;18(4):2087. [doi: 10.3390/ijerph18042087] [Medline: 33669951]
117. Perrone AM, Dondi G, Giunchi S, et al. COVID-19 free oncologic surgical hub: the experience of reallocation of a gynecologic oncology unit during pandemic outbreak. Gynecol Oncol 2021 Apr;161(1):89-96. [doi: 10.1016/j.ygyno.2020.09.030] [Medline: 33223219]
118. Gebbia V, Piazza D, Valerio MR, Borsellino N, Firenze A. Patients with cancer and COVID-19: a WhatsApp Messenger-based survey of patients' queries, needs, fears, and actions taken. JCO Glob Oncol 2020 May;6:722-729. [doi: 10.1200/GO.20.00118] [Medline: 32412811]
119. Solberg Carlsson K, Øvretveit J, Ohrling M. Rapid implementation of remote digital primary care in Stockholm and implications for further system-wide implementation: practitioner's and manager's experience of the Always Open mobile application. Scand J Prim Health Care 2023 Sep;41(3):232-246. [doi: 10.1080/02813432.2023.2229387] [Medline: 37470469]

XSL•FO
RenderX

120. Albert L, Capel I, García-Sáez G, Martín-Redondo P, Hernando ME, Rigla M. Managing gestational diabetes mellitus using a smartphone application with artificial intelligence (SineDie) during the COVID-19 pandemic: much more than just telemedicine. Diabetes Res Clin Pract 2020 Nov;169:108396. [doi: 10.1016/j.diabres.2020.108396] [Medline: 32890548]
121. Brandes AA, Ardizzoni A, Artioli F, et al. Fighting cancer in coronavirus disease era: organization of work in medical oncology departments in Emilia Romagna region of Italy. Future Oncol 2020 Jul;16(20):1433-1439. [doi: 10.2217/fon-2020-0358] [Medline: 32437242]
122. Wu F, Rotimi O, Laza-Cagigas R, Rampal T. The feasibility and effects of a telehealth-delivered home-based prehabilitation program for cancer patients during the pandemic. Curr Oncol 2021 Jun 17;28(3):2248-2259. [doi: 10.3390/curroncol28030207]
123. Forrester M, Breitenfeld L, Castelo-Branco M, Aperta J. The effects of the COVID-19 pandemic in oncology patient management. Int J Environ Res Public Health 2022 Jul 25;19(15):9041. [doi: 10.3390/ijerph19159041] [Medline: 35897414]
124. Kaddour H, Jama GM, Stagnell S, Kaddour S, Guner K, Kumar G. Remote triaging of urgent suspected head and neck cancer referrals: our experience during the first wave of the COVID-19 pandemic. Eur Arch Otorhinolaryngol 2022 Feb;279(2):1111-1115. [doi: 10.1007/s00405-021-07135-3] [Medline: 34661717]
125. Miguela Álvarez SM, Bartra Ylla A, Salvador Carreño J, Castillón P, García Cardona C, Anglès Crespo F. Telephone consultation service in orthopedics during COVID-19 pandemic [Article in Spanish]. Revista Española de Cirugía Ortopédica y Traumatología (English Edition) 2021 May;65(3):167-171. [doi: 10.1016/j.recote.2020.07.007]
126. Chesser TJS, Handley R, Kloos J, et al. International trauma care: initial European approaches during the COVID 19 pandemic. OTA Int 2021 Mar;4(1 Suppl):e112. [doi: 10.1097/OI9.0000000000000112] [Medline: 38630066]
127. Gilbert AW, Billany JCT, Adam R, et al. Rapid implementation of virtual clinics due to COVID-19: report and early evaluation of a quality improvement initiative. BMJ Open Qual 2020 May;9(2):e000985. [doi: 10.1136/bmjoq-2020-000985] [Medline: 32439740]
128. Banks J, Corrigan D, Grogan R, et al. LoVE in a time of COVID: clinician and patient experience using telemedicine for chronic epilepsy management. Epilepsy Behav 2021 Feb;115:107675. [doi: 10.1016/j.yebeh.2020.107675] [Medline: 33342712]
129. Vasta R, Moglia C, D'Ovidio F, et al. Telemedicine for patients with amyotrophic lateral sclerosis during COVID-19 pandemic: an Italian ALS referral center experience. Amyotroph Lateral Scler Frontotemporal Degener 2021 May;22(3-4):308-311. [doi: 10.1080/21678421.2020.1820043] [Medline: 32924624]
130. Capozzo R, Zoccolella S, Musio M, Barone R, Accogli M, Logroscino G. Telemedicine is a useful tool to deliver care to patients with amyotrophic lateral sclerosis during COVID-19 pandemic: results from Southern Italy. Amyotroph Lateral Scler Frontotemporal Degener 2020 Nov;21(7-8):542-548. [doi: 10.1080/21678421.2020.1773502] [Medline: 32530314]
131. McKenna MC, Al-Hinai M, Bradley D, et al. Patients' experiences of remote neurology consultations during the COVID-19 pandemic. Eur Neurol 2020;83(6):622-625. [doi: 10.1159/000511900] [Medline: 33147591]
132. Motolese F, Magliozzi A, Puttini F, et al. Parkinson's disease remote patient monitoring during the COVID-19 lockdown. Front Neurol 2020;11:567413. [doi: 10.3389/fneur.2020.567413] [Medline: 33117262]
133. van de Vijver S, Hummel D, van Dijk AH, et al. Evaluation of a digital self-management platform for patients with chronic illness in primary care: qualitative study of stakeholders' perspectives. JMIR Form Res 2022 Aug 3;6(8):e38424. [doi: 10.2196/38424] [Medline: 35921145]
134. Hartl L, Semmler G, Hofer BS, et al. COVID-19-related downscaling of in-hospital liver care decreased patient satisfaction and increased liver-related mortality. Hepatol Commun 2021 Oct;5(10):1660-1675. [doi: 10.1002/hep4.1758] [Medline: 34222742]
135. Barth J, Canella C, Oehler M, Witt CM. Digital consultations during COVID-19: a multiperspective mixed-methods study in an integrative medicine setting in Switzerland. J Altern Complement Med 2021 Jul;27(7):569-578. [doi: 10.1089/acm.2020.0539] [Medline: 33960805]
136. Doica IP, Florescu DN, Oancea CN, et al. Telemedicine chronic viral hepatitis C treatment during the lockdown period in Romania: a pilot study. Int J Environ Res Public Health 2021 Apr 1;18(7):3694. [doi: 10.3390/ijerph18073694] [Medline: 33916226]
137. Quinn LM, Olajide O, Green M, Sayed H, Ansar H. Patient and professional experiences with virtual antenatal clinics during the COVID-19 pandemic in a UK tertiary obstetric hospital: questionnaire study. J Med Internet Res 2021 Aug 31;23(8):e25549. [doi: 10.2196/25549] [Medline: 34254940]
138. Thomas I, Siew LQC, Rutkowski K. Synchronous telemedicine in allergy: lessons learned and transformation of care during the COVID-19 pandemic. J ALLERGY Clin Immunol Pract 2021 Jan;9(1):170-176. [doi: 10.1016/j.jaip.2020.10.013] [Medline: 33091636]
139. Tyler JMB, Pratt AC, Wooster J, Vasilakis C, Wood RM. The impact of increased outpatient telehealth during COVID-19: retrospective analysis of patient survey and routine activity data from a major healthcare system in England. Int J Health Plan Manag 2021 Jul;36(4):1338-1345. [doi: 10.1002/hpm.3185]
140. Byrne E, Watkinson S. Patient and clinician satisfaction with video consultations during the COVID-19 pandemic: an opportunity for a new way of working. J Orthod 2021 Mar;48(1):64-73. [doi: 10.1177/1465312520973677] [Medline: 33251951]

XSL•FO
RenderX

141. Pinar U, Anract J, Perrot O, et al. Preliminary assessment of patient and physician satisfaction with the use of teleconsultation in urology during the COVID-19 pandemic. World J Urol 2021 Jun;39(6):1991-1996. [doi: 10.1007/s00345-020-03432-4] [Medline: 32909174]
142. Shaw SE, Hughes G, Wherton J, et al. Achieving spread, scale up and sustainability of video consulting services during the COVID-19 pandemic? Findings from a comparative case study of policy implementation in England, Wales, Scotland and Northern Ireland. Front Digit Health 2021;3:754319. [doi: 10.3389/fdgth.2021.754319] [Medline: 34988546]
143. Eger K, Paroczai D, Bacon A, et al. The effect of the COVID-19 pandemic on severe asthma care in Europe: will care change for good? ERJ Open Res 2022 Apr;8(2):00065-2022. [doi: 10.1183/23120541.00065-2022] [Medline: 35582679]
144. Florea M, Lazea C, Gaga R, et al. Lights and shadows of the perception of the use of telemedicine by Romanian family doctors during the COVID-19 pandemic. Int J Gen Med 2021;14:1575-1587. [doi: 10.2147/IJGM.S309519] [Medline: 33953605]
145. Gesuete FP, Molle M, Gubitosi A, et al. Telemonitoring wound recovery with smartphone: an Italian experience during pandemic period. Plast Reconstr Surg Glob Open 2023 May;11(5):e5076. [doi: 10.1097/GOX.0000000000005076] [Medline: 37250826]
146. Gleeson LL, Ludlow A, Wallace E, et al. Changes to primary care delivery during the COVID-19 pandemic and perceived impact on medication safety: a survey study. Explor Res Clin Soc Pharm 2022 Jun;6:100143. [doi: 10.1016/j.rcsop.2022.100143] [Medline: 35702683]
147. Johnsen TM, Norberg BL, Kristiansen E, et al. Suitability of video consultations during the COVID-19 pandemic lockdown: cross-sectional survey among Norwegian general practitioners. J Med Internet Res 2021 Feb 8;23(2):e26433. [doi: 10.2196/26433] [Medline: 33465037]
148. Stepaniuk A, Pawlukianiec C, Krawiel M, Lewoc M, Baran A, Flisiak I. Great hopes or disappointment - a survey-based study on patients' and doctors' perception of telemedicine during the COVID-19 pandemic in Poland. Postepy Dermatol Alergol 2022 Apr;39(2):384-391. [doi: 10.5114/ada.2022.113827] [Medline: 35645687]
149. Rzewuska Díaz M, Locock L, Keen A, Melvin M, Myhill A, Ramsay C. Implementation of a web-based outpatient asynchronous consultation service: mixed methods study. J Med Internet Res 2024 Jun 4;26:e48092. [doi: 10.2196/48092] [Medline: 38833695]
150. Tensen E, Kuziemsky C, Jaspers MW, Peute LW. General practitioners' perspectives about remote dermatology care during the COVID-19 pandemic in the Netherlands: questionnaire-based study. JMIR Dermatol 2023 Jun 13;6:e46682. [doi: 10.2196/46682] [Medline: 37632975]
151. Li JPO, Thomas AAP, Kilduff CLS, et al. Safety of video-based telemedicine compared to in-person triage in emergency ophthalmology during COVID-19. EClinicalMedicine 2021 Apr;34:100818. [doi: 10.1016/j.eclinm.2021.100818] [Medline: 33842860]
152. Wilk M, Surowiec P, Matejko B, et al. Diabetes management delivery and pregnancy outcomes in women with gestational diabetes mellitus during the first wave of the 2020 COVID-19 pandemic: a single-reference center report. J Diabetes Res 2021;2021:5515902. [doi: 10.1155/2021/5515902] [Medline: 34307689]
153. Angelovská O, Dobiášová K, Těšinová JK. Pandemic COVID-19 as a challenge for telemedicine in the Czech Republic. Int J Health Plann Manage 2025 Jan;40(1):271-286. [doi: 10.1002/hpm.3863] [Medline: 39497023]
154. Sanders MW, Field J, Newell D, Osborne N. Use of remote consultations by chiropractors in the United Kingdom during the COVID-19 pandemic: a cross-sectional survey. J Chiropr Med 2025;24(1-4):106-118. [doi: 10.1016/j.jcm.2025.07.003]
155. Boydell N, Reynolds-Wright JJ, Cameron ST, Harden J. Women's experiences of a telemedicine abortion service (up to 12 weeks) implemented during the coronavirus (COVID-19) pandemic: a qualitative evaluation. BJOG 2021 Oct;128(11):1752-1761. [doi: 10.1111/1471-0528.16813] [Medline: 34138505]
156. Deml MJ, Minnema J, Dubois J, et al. The impact of the COVID-19 pandemic on the continuity of care for at-risk patients in Swiss primary care settings: a mixed-methods study. Soc Sci Med 2022 Apr;298:114858. [doi: 10.1016/j.socscimed.2022.114858] [Medline: 35247784]
157. Ochieng L, Salehi M, Ochieng R, et al. Augmented video consultations in care homes during the COVID-19 pandemic: a qualitative study. BJGP Open 2022 Dec;6(4):BJGPO.2022.0073. [doi: 10.3399/BJGPO.2022.0073] [Medline: 35764408]
158. Riboli-Sasco E, El-Osta A, El Asmar ML, et al. Investigating barriers & facilitators for the successful implementation of the BP@home initiative in London: primary care perspectives. PLoS ONE 2024;19(2):e0298898. [doi: 10.1371/journal.pone.0298898] [Medline: 38422101]
159. Kollmann J, Sana S, Magnée T, et al. Patients' and professionals' experiences with remote care during COVID-19: a qualitative study in general practices in low-income neighborhoods. Prim Health Care Res Dev 2024 Jun 3;25:e32. [doi: 10.1017/S1463423624000240] [Medline: 38826073]
160. Cervantes-Torres L, Romero-Blanco C. Longitudinal study of the flash glucose monitoring system in type 1 diabetics: an mHealth ally in times of COVID-19. J Clin Nurs 2023 Jul;32(13-14):3840-3851. [doi: 10.1111/jocn.16523] [Medline: 36071646]
161. Severino P, D'Amato A, Prosperi S, et al. Clinical support through telemedicine in heart failure outpatients during the COVID-19 pandemic period: results of a 12-months follow up. J Clin Med 2022 May 16;11(10):2790. [doi: 10.3390/jcm11102790] [Medline: 35628916]

XSL•FO
RenderX

162. Abedin N, Kilbinger C, Queck A, et al. Telemedicine hybrid care models in gastroenterology outpatient care: results from a German tertiary center. J Clin Med 2025 Apr 4;14(7):2471. [doi: 10.3390/jcm14072471] [Medline: 40217919]
163. Cuevas Fernández FJ, Quintana AE, López OTG, Galeote JCG, de León AC, Aguirre-Jaime A. The role of telemedicine in the monitoring and control of patients with hypertension during the COVID−19 pandemic: a multicenter study in primary care. Fam Med Prim Care Rev 2024;26(4):438-443. [doi: 10.5114/fmpcr.2024.144911]
164. Zondag AGM, Haitjema S, de Groot MCH, et al. Comparison of outpatient attendance, cardiovascular risk management and cardiovascular health across preCOVID-19, during and postCOVID-19 periods: a prospective cohort study. BMJ Open 2025 Jul 16;15(7):e092374. [doi: 10.1136/bmjopen-2024-092374] [Medline: 40669897]
165. Beauquis J, Petit AE, Michaux V, Sagué V, Henrard S, Leprince JG. Dental emergencies management in COVID-19 pandemic peak: a cohort study. J Dent Res 2021 Apr;100(4):352-360. [doi: 10.1177/0022034521990314] [Medline: 33541180]
166. Munda A, Indihar B, Okanovič G, Zorko K, Steblovnik L, Barlovič DP. Maternal and perinatal outcomes during the COVID-19 epidemic in pregnancies complicated by gestational diabetes. Zdr Varst 2023 Mar;62(1):22-29. [doi: 10.2478/sjph-2023-0004] [Medline: 36694793]
167. Bauwens PH, Fayard JM, Tatar M, et al. Evaluation of a smartphone application for self-rehabilitation after anterior cruciate ligament reconstruction during a COVID-19 lockdown. Orthop.Traumatol Surg Res 2023 Feb;109(1):103342. [doi: 10.1016/j.otsr.2022.103342]
168. Grobe-Einsler M, Taheri Amin A, Faber J, et al. Development of SARAhome, a new video-based tool for the assessment of ataxia at home. Mov Disord 2021 May;36(5):1242-1246. [doi: 10.1002/mds.28478] [Medline: 33433030]
169. Pulvirenti F, Cinetto F, Milito C, et al. Health-related quality of life in common variable immunodeficiency Italian patients switched to remote assistance during the COVID-19 pandemic. J Allergy Clin Immunol Pract 2020 Jun;8(6):1894-1899. [doi: 10.1016/j.jaip.2020.04.003] [Medline: 32278865]
170. Bhadola S, Tang C, Marks A, et al. Disparate healthcare access and telehealth-based hybrid consultations during the COVID-19 pandemic. WORK 2022;73(2):377-382. [doi: 10.3233/WOR-211463]
171. Banbury A, Smith AC, Taylor ML, et al. Cancer care and management during COVID-19: a comparison of in-person, video and telephone consultations. J Telemed Telecare 2022 Dec;28(10):733-739. [doi: 10.1177/1357633X221123409] [Medline: 36346931]
172. Ashley C, Williams A, Dennis S, et al. Telehealth's future in Australian primary health care: a qualitative study exploring lessons learnt from the COVID-19 pandemic. BJGP Open 2023 Jun;7(2):BJGPO.2022.0117. [doi: 10.3399/BJGPO.2022.0117] [Medline: 36750374]
173. Sachs JW, Graven P, Gold JA, Kassakian SZ. Disparities in telephone and video telehealth engagement during the COVID-19 pandemic. JAMIA Open 2021 Jul;4(3):ooab056. [doi: 10.1093/jamiaopen/ooab056] [Medline: 34632322]
174. Weber E, Miller SJ, Shroff N, Beyrouty M, Calman N. Recent telehealth utilization at a large federally qualified health center system: evidence of disparities even within telehealth modalities. Telemed J E Health 2023 Nov;29(11):1601-1612. [doi: 10.1089/tmj.2022.0511] [Medline: 36961396]
175. Singh K, Kaur N, Prabhu A. Combating COVID-19 crisis using artificial intelligence (AI) based approach: systematic review. Curr Top Med Chem 2024;24(8):737-753. [doi: 10.2174/0115680266282179240124072121] [Medline: 38318824]
176. Abd-Alrazaq A, Hassan A, Abuelezz I, et al. Overview of technologies implemented during the first wave of the COVID-19 pandemic: scoping review. J Med Internet Res 2021 Sep 14;23(9):e29136. [doi: 10.2196/29136] [Medline: 34406962]
177. Ganjali R, Jajroudi M, Kheirdoust A, Darroudi A, Alnattah A. Telemedicine solutions for clinical care delivery during COVID-19 pandemic: a scoping review. Front Public Health 2022;10:937207. [doi: 10.3389/fpubh.2022.937207] [Medline: 35937265]
178. Li HL, Chan YC, Huang JX, Cheng SW. Pilot study using telemedicine video consultation for vascular patients' care during the COVID-19 period. Ann Vasc Surg 2020 Oct;68:76-82. [doi: 10.1016/j.avsg.2020.06.023] [Medline: 32562832]
179. Manski-Nankervis JA, Davidson S, Hiscock H, et al. Primary care consumers' experiences and opinions of a telehealth consultation delivered via video during the COVID-19 pandemic. Aust J Prim Health 2022 Jun;28(3):224-231. [doi: 10.1071/PY21193] [Medline: 35287793]
180. Doraiswamy S, Abraham A, Mamtani R, Cheema S. Use of telehealth during the COVID-19 pandemic: scoping review. J Med Internet Res 2020 Dec 1;22(12):e24087. [doi: 10.2196/24087] [Medline: 33147166]
181. Ramachandran M, Brinton C, Wiljer D, Upshur R, Gray CS. The impact of eHealth on relationships and trust in primary care: a review of reviews. BMC Prim Care 2023 Nov 3;24(1):228. [doi: 10.1186/s12875-023-02176-5] [Medline: 37919688]
182. Pong C, Tseng RMWW, Tham YC, Lum E. Current implementation of digital health in chronic disease management: scoping review. J Med Internet Res 2024 Dec 12;26(26):e53576. [doi: 10.2196/53576] [Medline: 39666972]
183. Andrews E, Berghofer K, Long J, Prescott A, Caboral-Stevens M. Satisfaction with the use of telehealth during COVID-19: an integrative review. Int J Nurs Stud Adv 2020 Nov;2:100008. [doi: 10.1016/j.ijnsa.2020.100008] [Medline: 33083791]
184. Pogorzelska K, Chlabicz S. Patient satisfaction with telemedicine during the COVID-19 pandemic-a systematic review. Int J Environ Res Public Health 2022 May 17;19(10):6113. [doi: 10.3390/ijerph19106113] [Medline: 35627650]

XSL•FO
RenderX

185. Bashi N, Fatehi F, Mosadeghi-Nik M, Askari MS, Karunanithi M. Digital health interventions for chronic diseases: a scoping review of evaluation frameworks. BMJ Health Care Inform 2020 Mar;27(1):e100066. [doi: 10.1136/bmjhci-2019-100066] [Medline: 32156751]
186. Snoswell CL, Chelberg G, De Guzman KR, et al. The clinical effectiveness of telehealth: a systematic review of meta-analyses from 2010 to 2019. J Telemed Telecare 2023 Oct;29(9):669-684. [doi: 10.1177/1357633X211022907] [Medline: 34184580]
187. Kuan PX, Chan WK, Fern Ying DK, et al. Efficacy of telemedicine for the management of cardiovascular disease: a systematic review and meta-analysis. Lancet Digit Health 2022 Sep;4(9):e676-e691. [doi: 10.1016/S2589-7500(22)00124-8] [Medline: 36028290]
188. Valdes D, Shanker A, Hijazi G, et al. Global evidence on the sustainability of telemedicine in outpatient and primary care during the first 2 years of the COVID-19 pandemic: scoping review using the Nonadoption, Abandonment, Scale-Up, Spread, and Sustainability (NASSS) framework. Interact J Med Res 2025 Feb 28;14:e45367. [doi: 10.2196/45367] [Medline: 40053716]
189. Bernhard-Nocht-Institut for Tropical Medicine (BNITM) Federation of European Societies for Tropical Medicine and International Health (FESTMIH) German Society for Tropical Medicine TMaGHD. Strengthening pandemic preparedness – enabling access through hybrid care models – project RAPIDE. In: 14th European Congress on Tropical Medicine and International Health (ECTMIH 2025); 2025: Frontiers; 2025. URL: https://www.frontiersin.org/books/14th_European_Congress_on_Tropical_Medicine_and_International_Health_ECTMIH_2025/14193 [accessed 2026-03-03]
190. Sanchez Villalobos N, vanLoenen T, Ngongalah L, Connolly M, Timen A. Pathways for hybrid care: modifications in non-pandemic care delivery during the COVID-19 pandemic. Eur J Public Health 2025 Oct 1;35(Supplement_4). [doi: 10.1093/eurpub/ckaf161.101]

Abbreviations

AI: artificial intelligence
MMAT: Mixed Methods Appraisal Tool
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
RAPIDE: Regular and Unplanned Care Adaptive Dashboard for Cross-Border Emergencies

Edited by S Brini; submitted 24.Sep.2025; peer-reviewed by AL Bigal, N Namazee; revised version received 12.Mar.2026; accepted 12.Mar.2026; published 30.Apr.2026.

Please cite as:
Sanchez Villalobos N, van Loenen T, Ngongalah L, Connolly MA, Stein ML, Rovers CP, Timen A
Hybrid Care Modifications in the Delivery of Nonpandemic Care During the COVID-19 Pandemic: Scoping Review
J Med Internet Res 2026;28:e84756
URL: https://www.jmir.org/2026/1/e84756
doi:10.2196/84756

© Natalia Sanchez Villalobos, Tessa van Loenen, Lem Ngongalah, Máire A Connolly, Mart L Stein, Chantal P Rovers, Aura Timen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Use of Mobile Technologies for Smoking Cessation Among Smokers and Former Smokers: Systematic Review

Meret Lakeberg[1,2], MSc; Falk Hoffmann[1,2], Prof Dr; Jannik Ohmes[1], BSc; Birte Berger-Höger[2,3], Prof Dr; Jasmin Helbach[1], Dr rer medic

[1]Department of Health Services Research, School of Medicine and Health Sciences, Carl von Ossietzky Universität Oldenburg, Ammerländer Heerstraße 114-118, Oldenburg, Germany

[2]Leibniz Science Campus, Digital Public Health Bremen, Bremen, Germany

[3]Faculty 11 Human and Health Sciences, University of Bremen, Institute for Public Health and Nursing Research, Bremen, Germany

Corresponding Author:
Jasmin Helbach, Dr rer medic
Department of Health Services Research, School of Medicine and Health Sciences, Carl von Ossietzky Universität Oldenburg, Ammerländer Heerstraße 114-118, Oldenburg, Germany

Abstract

Background: In an era of widespread mobile phone usage, digital public health interventions offer a new cost-effective way of improving public health. In the context of smoking cessation, studies indicate that mobile technologies have the potential to support individuals to quit smoking. However, there is no systematic synthesis of how often they are used by smokers and former smokers.

Objective: The aim of this study is to assess the prevalence of mobile technology use for smoking cessation among smokers and former smokers and to examine their intention to use.

Methods: MEDLINE via PubMed, Embase, and PsycInfo were searched from inception to February 13, 2025. Studies were eligible if they reported how often smokers and former smokers in high-income countries used mobile technologies for smoking or vaping cessation. Study quality was assessed using the Joanna Briggs Institute tool for prevalence studies. Data synthesis was conducted narratively.

Results: Twenty-seven cross-sectional studies were included, 25 on smoking and 2 on vaping cessation. The 25 studies on smoking cessation collected data between 2005 and 2024 and comprised 117 to 27,323 participants (mean age 19.9 - 50.3 years; n=8). Lifetime prevalences of mobile technology use for smoking cessation ranged between 2.5% and 35.9% (n=8), depending on technology type and population. Period prevalences (0% - 12%; n=4) and point prevalences (1.1% - 10.9%; n=11) were generally lower. Regardless of the prevalence type, the internet was the most frequently used technology (0.8% - 35.9%; n=14). Intention to use mobile technologies for smoking cessation ranged from 19.5% for Twitter to 46.7% for websites (n=2). Of the 2 studies on vaping cessation, 1 presented lifetime prevalence (1.1% - 17.3%), while the other presented period prevalence (5.5% - 6.3%). The intention to use mobile technologies for vaping cessation ranged from 9.7% for web-based programs to 34.6% for apps (n=1). Based on the risk of bias assessment, study quality was heterogeneous, with frequent limitations in sampling procedures, reporting, and reliance on self-reported measures.

Conclusions: This review provides novel insights into the role of mobile technologies in smoking cessation. Evidence indicates that the prevalence of mobile technology use for smoking cessation is low and that disparities in access and engagement exist. However, there is a high intention to use such tools. Therefore, efforts should focus on delivering existing evidence-based tools rather than developing new ones. Included studies were characterized by high methodological variability and poor reporting, so the results must be interpreted with caution. Overall, despite the widespread availability of mobile technologies to support smoking cessation, research on their utilization remains limited.

Trial Registration: PROSPERO CRD42025647875; https://www.crd.york.ac.uk/PROSPERO/view/CRD42025647875

(J Med Internet Res 2026;28:e83072) doi:10.2196/83072

KEYWORDS

smoking cessation; mHealth; digital health; quitting smoking; tobacco

XSL•FO
RenderX

Introduction

Tobacco consumption remains the leading preventable cause of disease, disability, and death worldwide, placing a substantial burden on public health systems [1]. Despite decades of prevention efforts, smoking tobacco continues to claim millions of lives each year. The *Global Burden of Disease Study 2019* reported 7.69 million smoking-attributable deaths and over 200 million disability-adjusted life years, with 86.9% of these deaths occurring among current smokers [2]. According to the World Health Organization (WHO) *Tobacco Trends Report 2025*, approximately 1 in 5 adults globally still use tobacco [3]. While smoking prevalence has steadily declined since the 1960s, particularly in high-income countries [4,5], the absolute number of smokers remains high. In Europe, the recent smoking prevalence ranges from 8% in Sweden up to 36% in Greece [6]. Furthermore, the emergence of new tobacco products and technologies, such as electronic cigarettes (e-cigarettes), has led to a new era of nicotine dependency [1]. In 2025, the WHO published its first estimate of the global prevalence of e-cigarettes, indicating that more than 100 million individuals worldwide were currently vaping [3].

Smoking cessation interventions play a key role in the prevention of smoking-related diseases. In 2022, approximately two-thirds of smokers in the United States expressed an intention to quit, and about half of all smokers had made at least 1 quit attempt [7]. Unaided quit attempts resulted in long-term abstinence (6 - 12 mo) in approximately 3% to 5% of cases, whereas the use of evidence-based support—such as behavioral counseling combined with nicotine replacement therapy, bupropion, or varenicline—can increase success rates by an additional 7% to 9% [8]. In Germany, data from the DEBRA study (Deutsche Befragung zum Rauchverhalten) show that smoking cessation attempts are common, but only 13% of individuals trying to quit in 2016 to 2019 used at least one evidence-based method [9]. This underlines a persistent gap between available evidence-based interventions and their real-world uptake.

In an era characterized by the widespread use of mobile phones among the global population, with over 82% of individuals above the age of 10 years owning a mobile phone, and a proportion exceeding 95% in high-income countries [10], digital public health interventions offer a new scalable, user-friendly, and cost-effective approach to enhancing public health [11,12], particularly in the domains of prevention and health promotion. In recent years, there has been an increasing integration of digital intervention modalities, encompassing mobile apps, SMS text messaging services, and other web-based interventions, into public health strategies [13-15], including strategies focusing on smoking cessation [16]. The effectiveness of digital smoking cessation interventions has been examined in numerous systematic reviews, which have indicated a promising role for mobile technologies in promoting smoking cessation and abstinence [17-22]. A systematic review examining the cessation of vaping has also demonstrated the efficacy of digital interventions [23]. However, there are no systematic reviews examining the prevalence of the utilization of such aids by smokers and former smokers. Nevertheless, existing evidence shows substantial variability in the prevalence of mobile technology use for smoking cessation across technology types and countries [24-26]. In an analysis of 8 European countries, Papadakis et al [25] reported mobile app use ranging from 0% to 6.2% and use of internet-based resources from 0.75% to 10.2%. Furthermore, user characteristics, including age, sex, and socioeconomic status, appear to be associated with differences in the uptake and engagement with mobile smoking cessation technologies [26-30]. In order to fully exploit the potential of mobile smoking cessation interventions to support behavior change, it is first necessary to systematically synthesize all existing evidence on how often smokers and former smokers use these tools.

Therefore, the primary aim of this systematic review is to provide information on the prevalence of the use of mobile technologies for smoking cessation among smokers and former smokers in high-income countries. A secondary aim is to examine the intention to use these technologies.

Methods

Overall study design

This systematic review was conducted in accordance with the Joanna Briggs Institute guideline for systematic reviews of prevalence and incidence [31]. The reporting follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [32], and its extension for searches (PRISMA-S; Checklist 1). A study protocol was registered in advance in the PROSPERO (Prospective Register of Systematic Reviews; CRD42025647875).

Literature Search

MEDLINE (via PubMed), Embase (via Elsevier), and PsycInfo (via EBSCO) were searched from inception to February 13, 2025. No restrictions were applied with regard to either language, study design, or publication date, and no search filters were used. The search strategy was developed by 1 researcher (ML) and reviewed by 2 other researchers (JH and FH). The final search strategies for all databases can be found in Table S1 in Multimedia Appendix 1. To complement the database searches, a forward (citing) and backward (cited) citation analyses of the included studies were performed on May 5, 2025, using Web of Science Core Collection following the TARCiS statement [33]. This analysis was repeated on newly included references until no additional studies could be identified, resulting in a second iteration on May 20, 2025, and a third iteration on May 22, 2025. Registries, online resources, and browsing were not searched, nor were other information sources (eg, contacting authors, experts, and manufacturers) used, as the focus of this review was exclusively on identifying published peer-reviewed studies.

Eligibility Criteria

We included published observational studies, including cross-sectional studies and cohort studies with data analyzed in a cross-sectional manner. We excluded PhD theses and conference abstracts as well as interventional studies, including randomized controlled trials. Exceptions were made for reports

XSL•FO
RenderX

of randomized controlled trials with a cross-sectional analysis of baseline data. No language restrictions were applied.

The study eligibility criteria were defined using the CoCoPop (condition, context, population) approach, which is recommended for reviews of prevalence and incidence data [34].

Condition

To be included, studies had to report the use of mobile technologies for smoking cessation, addressing smoking of any tobacco product. We included mobile technologies that function as stand-alone tools. The types of mobile technologies assessed in this review included mobile phone apps, SMS text messaging interventions, websites, web-based programs, and social media platforms accessed via mobile devices that provide informational or motivational content related to smoking cessation. Studies that referred to internet use as a smoking cessation resource, without specifying the access device, were included under the category of mobile technologies, acknowledging that a substantial proportion of internet use now occurs via mobile devices [35].

Studies examining the use of interventions that relied exclusively on videoconferencing, telephone calls, or online forums and social media used solely for direct interaction between individuals were excluded.

Context

For inclusion, the studies had to be conducted in high-income countries according to the OECD (Organisation for Economic Co-operation and Development)–Development Assistance Committee to ensure greater transferability to the German context [36].

Population

Studies were required to include data on smokers and former smokers as defined by the study authors, with at least 100 individuals in these categories. Studies focusing on specific populations, such as patients with multiple sclerosis or students, were also considered if relevant. To minimize the risk of selection bias, studies exclusively involving individuals who had previously used a specific cessation technology were excluded.

Study Selection

The search results were imported into EndNote (Version 20, Clarivate), where duplicates were removed. For screening, the results were transferred to Rayyan (Qatar Computing Research Institute). For calibration, all reviewers (ML, JH, and JO) independently screened 100 articles based on title and abstract for inclusion or exclusion. After consensus on these records, ML and one other reviewer (JH or JO) independently screened all remaining titles and abstracts. The full texts of all articles that met the inclusion criteria were also assessed independently by ML and one other reviewer (JH or JO). Any discrepancies were resolved by consensus or, if necessary, by consultation with the third reviewer.

Data Extraction

The data were extracted directly into the results tables by ML and were subsequently verified by the second reviewer (JO). Discrepancies were resolved through discussion or by the third reviewer (JH). The following items were extracted: study characteristics, characteristics of the study population, definition of smokers, prevalences of mobile technology use for smoking cessation, age- and sex-stratified prevalences of mobile technology use for smoking cessation, types of mobile technology, and prevalences of the intention to use mobile technologies.

Quality Assessment

The quality of the included studies was independently assessed by two reviewers (ML and JO) using the Joanna Briggs Institute's critical appraisal checklist for studies reporting prevalence data [37], which was adapted for our research question (Table S2 in Multimedia Appendix 1). The quality assessment was piloted to calibrate the reviewers using 2 studies. Any disagreement was resolved by discussion or by the third reviewer (JH).

Data Synthesis

The data were summarized descriptively using a narrative synthesis. Due to substantial variability in the clinical and methodological characteristics of the studies, which would result in noncomparable outcomes and populations, it was deemed inappropriate to conduct a meta-analysis by pooling prevalence across studies [37,38]. Instead, this systematic review aimed to describe this variability and present prevalence estimates at the study level, as well as overall prevalence ranges categorized by prevalence type and mobile technologies. Subgroup analyses were conducted on the basis of sex and age, with the prevalence ranges being described in a descriptive manner.

For studies reporting multiple time points, only the most recent data were included to reflect the most current usage patterns. If multiple articles were based on the same study population, only the most recent publication was included. The quality assessment was carried out for each report.

Deviation From the Protocol

During the peer-review process, it was recommended that studies assessing populations solely using noncombustible products, such as e-cigarettes or vapers, should not be included. In accordance with this recommendation, the primary focus of the review has now been narrowed to populations smoking tobacco products. However, given that the search and initial review also incorporated a systematic identification of studies focusing on the use of mobile technologies in the context of vaping cessation, we decided that these findings should also be presented briefly, but separated from studies on smoking cessation.

Results

Study Selection

The database search identified 7110 records (Figure 1). After screening the title and abstract, 105 articles were read, and 21 articles reporting on 20 studies were included. A

XSL•FO
RenderX

forward-backward citation search was performed on those articles, resulting in the inclusion of 8 additional articles. Based on those 8 articles, a second round of citation analysis was conducted, resulting in 3 additional reports for inclusion. The last round of citation analysis conducted with those 3 reports could not identify any new articles. Overall, 27 studies documented in 32 reports were included in this review. Of these, 25 focused on smoking cessation, while 2 studies focused on vaping cessation. All articles were written in English. The excluded articles are listed in Table S3 in Multimedia Appendix 1 with reasons for exclusion.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) study selection diagram. Citation searching was conducted and reported following the TARCIS Statement. The assignment of the reasons for exclusion was done hierarchically in the order presented.

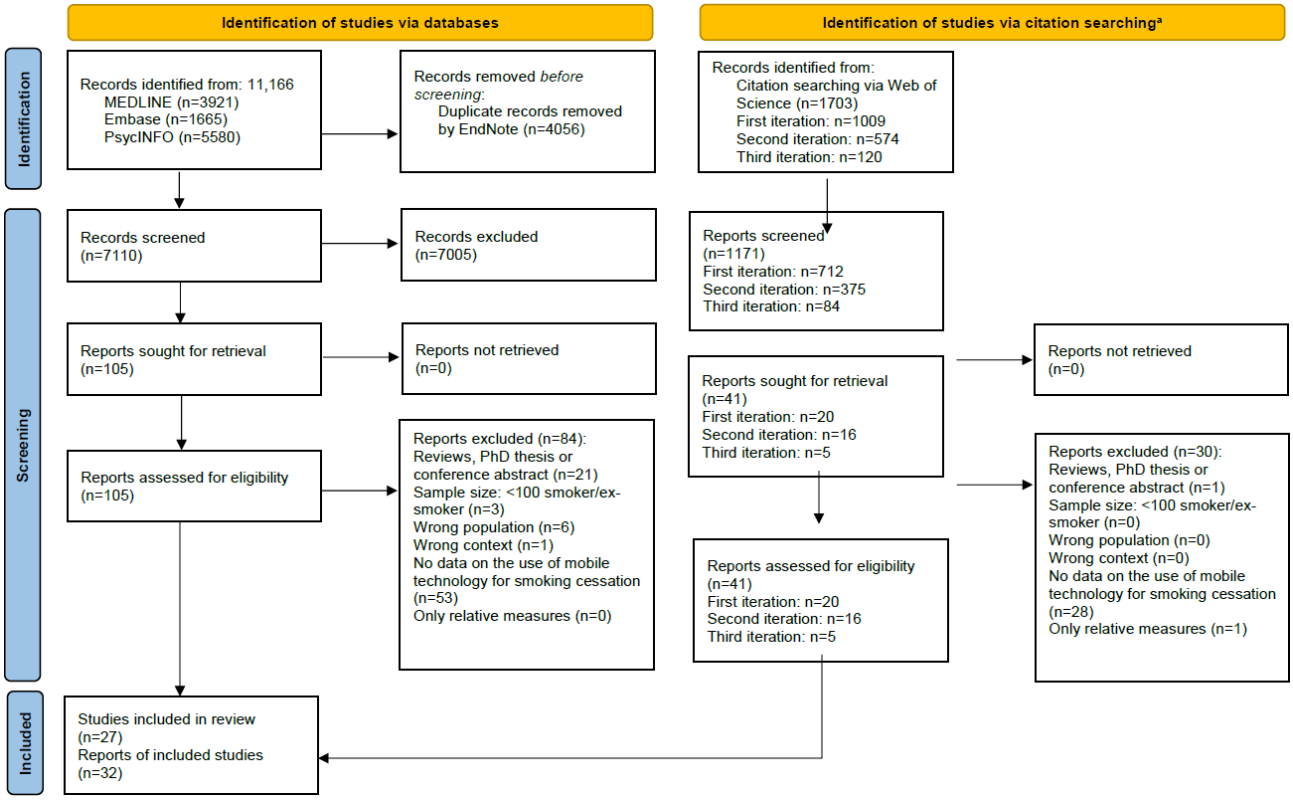

Study and Participant Characteristics of the Studies on Smoking Cessation

The characteristics of the 25 studies focusing on smoking cessation are shown in Table 1. A total of 9 studies were conducted in the United States [39-47], 4 studies were from the United Kingdom [30,48-50], 2 each from Norway [51,52], Canada [53,54], and Australia [55,56], and 1 each from the Netherlands [57] and from Greece [58]. Four studies included several countries in their scope [24-26,59]. All included studies used a cross-sectional design. In terms of data collection methods, 11 studies used online surveys, 4 relied on personal interviews, 4 applied mixed methods, 3 utilized paper-pencil surveys, 2 conducted telephone interviews, and 1 study did not report its data collection method. The types of mobile technologies investigated differed between studies and included websites or general internet use (n=16), smartphone apps (n=8), SMS text messaging programs (n=2), and social media (n=2). Six studies asked for the use of several tools without differentiation, which included web-based programs as well.

The number of participants per study ranged from 117 to 27,323. Although several studies reported the age distribution of their overall study population or subgroups, only 8 out of 25 provided the mean age specifically for smokers, with values ranging from 19.9 to 50.3 years. The proportion of female participants varied between 0% and 77% (n=18). Participants were categorized as either current or former smokers, with the definition of “smoker” varying between studies (Table S4 in Multimedia Appendix 1). Over a third of the included studies (n=10) focused exclusively on smokers who had recently attempted to quit.

XSL•FO
RenderX

Table . General characteristics of the included studies on smoking cessation (n=25) and vaping cessation (n=2).

Authors and year	Country	Data collection	Year of data	Sample size, n	Population	Mean age (years); female sex (%)
Smoking cessation						
Borrelli et al [24] (2015)	United States/United Kingdom	Online survey	2014	1000	Smokers aged 18 years and older	43.9 years; 55%
Bottorff et al [53] (2016)	Canada	Online survey	2014	117	Smokers, only men[a] column	39.8 years; 0%
Caraballo et al [39] (2017)	United States	Online survey	2014 - 2016	15,943	Smokers aged 18 years and older who made a quit attempt in the last 3 months	N/R[b]; 48.7%
Chevalking et al [57] (2018)	Netherlands	Online survey	2015 - 2016	802	Smokers and former smokers[a]	N/R
Curry et al [40] (2007) and companion report: Coups et al [60] (2009)	United States	Personal interview	2005	2747	Smokers aged 18 years and older who made a quit attempt in the past year	N/R
Graham and Amato [61] (2019)	United States	Online survey	2017	1736	Smokers aged 18 years and older	N/R
Gravely et al [59] (2021)	Australia, Canada, England, and United States	Online survey	2020	3614	Smokers, vapers, and former smokers aged 18 years and older who made a quit attempt or quit smoking in the last 2 years	N/R; 53%
Hummel et al [26] (2018)	England, Germany, Greece, Hungary, Netherlands, Poland, Romania, and Spain	Personal interview or online survey	2016	10,683	Smokers aged 18 years and older who made a quit attempt in the past year	N/R
Jackson et al [30] (2019)	United Kingdom	Personal interview	2006 - 2018	18,929	Smokers aged 16 years and older who made at least 1 quit attempt in the past year	N/R; 52%
Jackson et al [48] (2022)	United Kingdom	Personal interview	2020 - 2021	1550	Smokers and former smokers (smoking during preceding year) aged 18 years and older who made at least 1 quit attempt in the past year	N/R; 50%
Jackson et al [49] (2025)	United Kingdom	Telephone interview	2023 - 2024	1642	Smokers and former smokers (smoking during preceding year) aged 16 years and older who made at least 1 quit attempt in the past year	N/R
Jayakumar et al [54] (2019)	Canada	Online survey	2013 - 2018	2773	Smokers and former smokers (had quit within last 3 years) aged 18 years and older	N/R; 64%

XSL•FO
RenderX

Authors and year	Country	Data collection	Year of data	Sample size, n	Population	Mean age (years); female sex (%)
Kostagiolas et al [58] (2023)	Greece	Paper-pencil survey	2019	150	Smokers aged 18 years and older	N/R; 58%
Li et al [56] (2025)	Australia	Online survey	2017 - 2020	1244	Smokers aged 18 years and older from rural/remote locations	43.6 years; 77%
Lund and Lund [52] (2022)	Norway	Telephone interview	2017 - 2020	874	Smokers and former smokers aged 20 years and older	50.3 years; 47%
Lund and Lund [51] (2023)	Norway	Online survey	2020	1111	Smokers and former smokers[a]	N/R; 52.8%[c]
Moon et al [42] (2020)	United States	Paper-pencil survey	2018 - 2019	227	Smokers with Korean background[a]	N/R; 12.8%
Oliver et al [43] (2018)	United States	Online survey or paper-pencil survey	2013 - 2017	180	Smokers aged 18-65 years	N/R
Papadakis et al [25] (2020)	England, Germany, Greece, Hungary, Netherlands, Poland, Romania, and Spain	N/R	2017 - 2018	2699	Smokers[a]	N/R
Patterson et al [44] (2021)	United States	Personal interview and telephone interview	2018 - 2019	659	Smokers aged 18 years and older	N/R; 43%[d]
Perski et al [50] (2022) and companion reports: Beard et al [62] (2016) and Perski et al [63] (2019)	United Kingdom	Personal interview	2014 - 2020	5892	Smokers aged 16 years and older who made a quit attempt in the past year	N/R; 50%
Ramo et al [45] (2015)	United States	Online survey	2010 - 2011	570	Smokers aged 18-25 years	19.9 years; 30%
Thrul et al [46] (2021) and companion reports: Mojtabai et al [64](2020) and Soulakova and Crockett [65] (2018)	United States	Personal interview or telephone interview	2010 - 2011	27,323	Smokers aged 18 years and older	42.7 years; 46%
Tofighi et al [47](2019)	United States	Paper-pencil survey	2015	157	Smokers enrolled as patients in inpatient detoxification treatment and post discharge (for alcohol and/or heroin)[a]	43.7 years; 9%
Twyman et al [55] (2018)	Australia	Online survey	2012 - 2013	646	Smokers and former smokers aged 18 years and older who are financially disadvantaged and made a quit attempt before	39 years; 52%
Vaping cessation						

XSL•FO
RenderX

Authors and year	Country	Data collection	Year of data	Sample size, n	Population	Mean age (years); female sex (%)
Dai et al [66] (2023)	United States	Online survey	2021	889	e-Cigarette[e] users, school grade 6 - 12 years, who made quit attempts in the past year[a]	N/R; 51%
Jones et al [67] (2023)	United States	Online survey	N/R	185	Vapers and former vapers aged 14-19 years	16.9 years; 52%

[a]Included age not displayed.

[b]N/R: not reported.

[c]Own calculation based on the numbers provided by the authors in Table 2 (587/1111×100=52.8%).

[d]Own calculation based on the numbers provided by the authors in Table 2 (280/659×100=42.5%).

[e]e-Cigarette: electronic cigarette.

Prevalence of Use of Mobile Technologies for Smoking Cessation

The data collection periods from the 25 studies on smoking cessation span from 2005 to 2024. A noticeable clustering of studies was observed between 2014 and 2020 (Figure 2), most of which focused on the use of apps and internet-based interventions. After 2021, fewer studies reporting on the use of mobile technologies for smoking cessation were identified.

A total of 8 studies reported the lifetime prevalence of using mobile technology for smoking cessation, which ranged from 2.5% to 35.9% (see Table 2). Four of these studies examined the usage of websites or the internet for smoking cessation, with prevalences ranging from 7.1% up to 35.9%. Mobile app usage ranged from 2.5% to 15.2% (n=4). SMS text messaging programs showed lifetime prevalences between 2.7% and 13.2% (n=2). Two studies presented data on social media use. While 1 study reported a combined prevalence of 22% for social media, another study differentiated between Twitter and Facebook and reported prevalences between 7.6% and 14.8%.

Period prevalence, covering usage within the past several months to the last 2 years, was assessed in 4 studies and ranged from 0% to 12%. Three studies examining the use of websites or the internet for smoking cessation were identified, with 1 study reporting a prevalence of 12%, while the other 2 observed prevalences of less than 2%. One study presented a combined prevalence for apps or SMS text messaging, showing that these tools were never used.

Point prevalence, referring to use during the most recent quit attempt, was reported in 11 studies, with prevalence ranging from 1.1% to 10.9%. For websites and internet use, the prevalences of use ranged from 0% to 10.9% (n=8). Four studies examined smartphone app use, which ranged between 0% and 10.7%.

Two studies reported more than 1 type of prevalence, and 1 study did not specify the type of prevalence assessed.

XSL•FO
RenderX

Figure 2. Use of mobile technologies for smoking cessation among smokers and former smokers over time, by prevalence type and technology (n=23; Curry et al [40] and Ramo et al [45] were not included in the figure as no prevalence type was stated or only stratified results were presented).

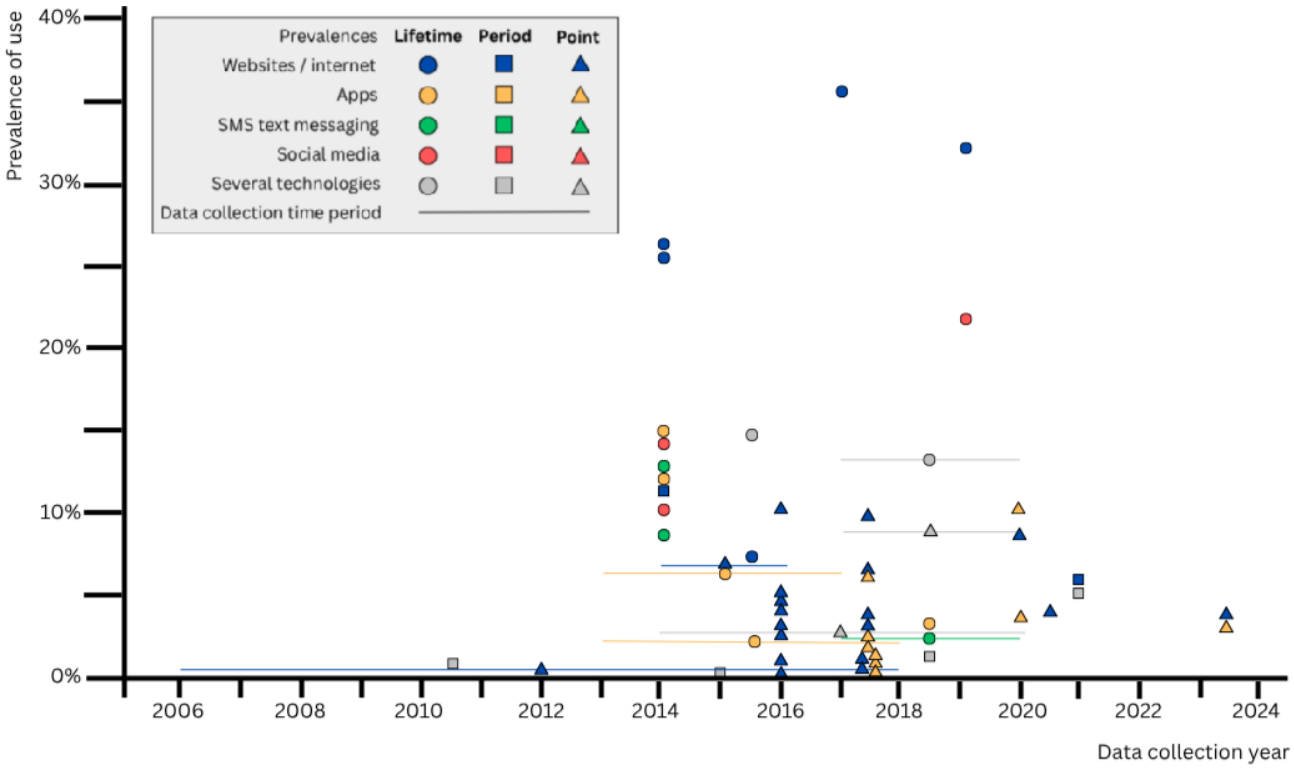

XSL•FO
RenderX

Table . Prevalence of the use of different mobile technology tools for smoking cessation among smokers and former smokers, categorized by lifetime, period, and point prevalence.

Authors and year	Country (Sample size, n)	Year of data	Type and prevalence of mobile technology use for smoking cessation, %
Lifetime prevalences			

Borrelli et al [24] (2015)	• United States (n=500)
	• United Kingdom (n=500)
		2014	• United States
			  • Websites: 26.2%
			  • Text messaging program: 13.2%
			  • Apps: 15.2%
			  • Twitter: 12.2%
			  • Facebook: 14.8%
			• United Kingdom
			  • Website: 25.6%
			  • Text messaging program: 8.8%
			  • Apps: 12.8%
			  • Twitter: 7.6%
			  • Facebook: 10.4%

Chevalking et al [57] (2018)	• Netherlands (n=802)	2015 - 2016	• Web-based program or apps: 15.5%[a] • Websites: 7.1%[b]
Graham and Amato [61] (2019)	• United States (n=1736)	2017	• Websites: 35.9%[c]
Jayakumar et al [54] (2020)	• Canada (n=2773)	2013 - 2018	• Apps: 2.5%
Kostagiolas et al [58] (2023)	• Greece (n=150)	2019	• Internet search engines: 33.3%[d] • Social media (Facebook, Twitter, Blogs): 22.0%[e] • Public institutions websites: 18.7%[f] • Medical websites: 23.3%[g]
Li et al [56] (2025)	• Australia (n=1244)	2017 - 2020	• Online services: 13.5% • Text messaging program: 2.7%
Moon et al [42] (2020)	• United States (n=227)	2018 - 2019	• Apps: 3.5%
Oliver et al [43] (2018)	• United States (n=180)	2013 - 2017	• Apps: 6.1%
Period prevalences			
Bottorff et al [53] (2016) (Period prevalence: last 3 months)	• Canada (n=117)	2014	• Websites: 12%
Patterson et al [44] (2021) (Period prevalence: last 2 years)	• United States (n=659)	2018 - 2019	• Internet/web-based program: 1.7%[h] • Videos: 0.8%[i]
Thrul et al [46] (2021) (Period prevalence: past year)	• United States (n=27,323)	2010 - 2011	• Internet/websites/web-based programs: 0.8%[c]
Tofighi et al [47] (2019) (Period prevalence: past year)	• United States (n=157)	2015	• Apps or text messaging tool: 0%
Point prevalences (for last quit attempt)			
Caraballo et al [39] (2017)	• United States (n=15,943)	2014 - 2016	• Websites: 7.1%[c]

XSL•FO
RenderX

Authors and year	Country (Sample size, n)	Year of data	Type and prevalence of mobile technology use for smoking cessation, %
Curry et al [40] (2007)	• United States (n=2747)	2005	• Internet: N/R[j]
Gravely et al [59] (2021)	• Australia, Canada, England, United States (n=3614)	2020	• Apps: 4.6%[c] • Websites: 8.5%[c]

Hummel et al [26] (2018)	• England (n=3536)
	• Germany (n=1003)
	• Greece (n=1000)
	• Hungary (n=1000)
	• Netherlands (n=1136)
	• Poland (n=1006)
	• Romania (n=1001)
	• Spain (n=1001)
		2016	• Internet
			  • England: 10.9%[c]
			  • Germany: 4.8%[c]
			  • Greece: 0.4%[c]
			  • Hungary: 2.7%[c]
			  • Netherlands: 5.7%[c] (period prevalence: last 6 months)
			  • Poland: 5.7%[c]
			  • Romania: 3.2%[c]
			  • Spain: 0%[c]

Jackson et al [30] (2019)	• United Kingdom (n=18,929)	2006 - 2018	• Websites: 1.1%
Jackson et al [48] (2022)	• United Kingdom (n=1550)	2020 - 2021	• Websites: 4.1%
Jackson et al [49] (2025)	• United Kingdom (n=1642)	2023 - 2024	• Websites: 4.6%[c] • Apps: 3.6%[c]
Lund and Lund [52] (2022)	• Norway (n=874)	2017 - 2020	• Web/mobile: 9.2%
Lund and Lund [51] (2023)	• Norway (n=1111)	2020	• Apps: 10.7%

Papadakis et al [25] (2020)	• England (n=1482)
	• Germany (n=185)
	• Greece (n=157)
	• Hungary (n=127)
	• Poland (n=179)
	• Romania (n=284)
	• Spain (n=285)
		2017 - 2018	• Apps/Internet:
			  • England: 6.2%[c]/10.2[c]
			  • Germany: 4.2%[c]/6.7[c]
			  • Greece: 0.0%[c]/0.7[c]
			  • Hungary: 1.3%[c]/ 2.6[c]
			  • Poland: 0.8%[c]/1.8[c]
			  • Romania: 0.4%[c]/1.7[c]
			  • Spain: 0.9%[c]/0.3[c]

Perski et al [50] (2022)	• United Kingdom (n=5892)	2014 - 2020	• Digital aid (websites, smartphone apps, and other): 2.5%

Prevalence type not reported or more than one prevalence type

XSL•FO
RenderX

Authors and year	Country (Sample size, n)	Year of data	Type and prevalence of mobile technology use for smoking cessation, %
Ramo et al [45] (2015) (Prevalence type not reported)	• USA (n=570)	2010 - 2011	• Internet: 2%
Twyman et al [55] (2018) (Lifetime prevalence in current smokers, point prevalence in former smokers)	• Australia (n=646)	2012 - 2013	• Internet: 1%

[a]Own calculation based on the numbers provided by Chevalking et al [57] on page 4 (124/802 × 100 = 15.5%).

[b]Own calculation based on the numbers provided by Chevalking et al [57] on page 4 (57/802 × 100 = 7.1%).

[c]Weighted results.

[d]Own calculation based on the numbers provided by Kostagiolas et al [58] in Multimedia Appendix 1 [(26 + 24)/150 × 100 = 33.3%].

[e]Own calculation based on the numbers provided by Kostagiolas et al [58] in Multimedia Appendix 1 [(20 +13)/150 ×100 = 22.0%].

[f]Own calculation based on the numbers provided by Kostagiolas et al [58] in Multimedia Appendix 1 [(13 + 15)/150 × 100 = 18.7%].

[g]Own calculation based on the numbers provided by Kostagiolas et al [58] in Multimedia Appendix 1 [(15 + 20)/150 × 100 = 23.3%].

[h]Own calculation based on the numbers provided by Patterson et al [44] in Table 2 [(3 + 8)/659 × 100 = 1.7%].

[i]Own calculation based on the numbers provided by Patterson et al [44] in Table 2 [(1 + 4)/659 × 100 = 0.8%].

[j]N/R: not reported.

Prevalence of Use of Mobile Technologies for Smoking Cessation, Stratified by Sex and Age

A total of 7 studies reported sex-stratified prevalences of mobile technology use for smoking cessation (Table S5 in Multimedia Appendix 1). Of these, 1 study focused on lifetime prevalences, 2 on period prevalences, and 5 on point prevalences. The study on lifetime prevalence and both studies on period prevalences found comparable usage patterns between males and females. Of the 5 studies that reported point prevalences, 3 did not demonstrate any sex-specific differences in utilization. One study presented higher prevalences of app utilization in females, and 1 study identified significant sex differences in internet utilization that varied across different countries. While males reported higher internet use for smoking cessation in Germany (6.6% males vs 2.9% females), England (14.4% vs 7.3%), Hungary (5.3% vs 0%), and Poland (7.7% vs 3.3%), the opposite was observed in the Netherlands (5.3% vs 6.3%) and Norway (for app use; 5.9% vs 15%), where females reported higher usage.

A total of 5 studies (1 on period prevalence and 4 on point prevalences) reported age-stratified data. Four of the studies assessed the utilization of websites or the internet, and one study assessed the utilization of mobile apps. Regardless of the prevalence type and mobile technology, younger individuals have been found to demonstrate higher utilization of mobile technologies.

Intention to Use Mobile Technology for Smoking Cessation

Two studies examined the intention to use mobile technologies for smoking cessation (Table 3). One of these studies reported a high intention to use websites (46.7%) and smartphone apps (42.7%). The intention to use SMS text messaging programs was found to be 28.5%, whereas the use of social media, including Twitter and Facebook, was found to be 18.95% and 28.1%, respectively. The other study concentrated on social media in general, reporting an intention to use of 31%.

Table . The prevalence of smokers' intentions to use mobile technologies to help them quit smoking in the future.

Authors and year	Country (Sample size, n)	Outcome	Prevalence of intention to use mobile technologies in the future, %
Borrelli et al [24] (2015)	United States and United Kingdom (n=1000)	Future intentions to use technology to quit smoking	• Websites: 46.7% • SMS text messaging program: 28.5%[a] • Smartphone app: 42.7% • Twitter: 19.5%[b] • Facebook: 28.1%[c]
Ramo et al [45] (2015)	United States (n=570)	Intention to use strategies to cut down/quit smoking	• Social media: 31%

[a]Own calculation based on the numbers provided by the authors in Table 5 [(157+128)/1000=28.5%].

[b]Own calculation based on the numbers provided by the authors in Table 5 [(113+82)/1000=19.5%].

[c]Own calculation based on the numbers provided by the authors in Table 5 [(152+129)/1000=28.1%].

XSL•FO
RenderX

Prevalence of Use and Intention to Use Mobile Technologies for Vaping Cessation

A total of 2 studies were identified that focus on vaping cessation [66,67]. The characteristics of these studies are presented in Table 2. Both studies were conducted in the United States, employed a cross-sectional design, and used an online survey for data collection. One study presented lifetime prevalences of the use of SMS text messaging programs (17.3%), apps (12.4%), web-based programs (9.7%), and different social media platforms (1.1% to 12.4%; Table S6 in Multimedia Appendix 1). The other study presented period prevalences of the use of the internet (6.3%) and apps or SMS text messaging programs (5.5%). The intention to use mobile technologies for vaping cessation was presented in 1 study (Table S6 in Multimedia Appendix 1), with the highest intention to use mobile apps (34.6%) being followed by SMS text messaging programs (30.8%) and social media (22.8%).

Quality Assessment of Studies on Smoking Cessation

Of the 25 studies on smoking cessation, approximately half employed an appropriate sampling frame to address the target population (n=1; Table 4) and used appropriate sampling methods (n=13). A suitable sample size was reported in 80% (20/25) of the studies, while 40% (10/25) provided sufficient detail on study methodology. Forty percent (10/25) of the studies demonstrated unclear attrition, while 56% (14/25) reported sufficient sample coverage. All studies relied on self-reported data. The outcome was measured in a standard and reliable manner in 88% (22/25) of the studies, and 80% (20/25) of the studies applied appropriate statistical analyses. Furthermore, 72% (18/25) of the studies did not report the response rate, and 1 study exhibited an inadequate response rate.

The quality of studies varies according to the prevalence types. While only 13% (1/8) of studies reporting lifetime prevalences employed an appropriate sample frame, 25% (1/4) of studies reporting period prevalences and 81% (9/11) of those on point prevalences did so. Furthermore, appropriate sampling methods were employed in 13% (1/8) of studies reporting lifetime prevalences compared to 50% (2/4) of studies reporting period prevalences and 81% (9/11) of studies reporting point prevalences. The sample size was deemed to be adequate in 63% (5/8) of studies reporting lifetime prevalences, 50% (2/4) of studies reporting period prevalences, and 100% (11/11) of studies reporting point prevalences. With regard to the reporting of study methodology, 9% (1/11) of studies on point prevalences were less adequately reported than 50% (4/8) of studies on lifetime prevalences and 75% (3/4) of studies on period prevalences. For all other categories, there is only minor variability in study quality.

XSL•FO
RenderX

Table . Overview of the quality of the included studies[a].

	Was the sample frame appropriate to address the target population?	Were study participants sampled in an appropriate way?	Was the sample size adequate?	Were the study subjects and the setting described in detail?	Was the data analysis conducted with sufficient coverage of the identified sample?	Were valid methods used for the identification of the condition?	Was the condition measured in a standard, reliable way for all participants?	Was there appropriate statistical analysis?	Was the response rate adequate, and if not, was the low response rate managed appropriately?
Smoking cessation									
Lifetime prevalence									
Borrelli et al [24] (2015)	No	No	Yes	Yes	No	No	Yes	Yes	Unclear
Chevalking et al [57] (2018)	No	No	Yes	Yes	Unclear	No	Yes	Yes	Unclear
Graham and Amato [61] (2019)	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Unclear
Jayakumar et al [54] (2019)	No	No	Yes	No	Unclear	No	Yes	Yes	Unclear
Kostagiolas et al [58] (2023)	Unclear	Unclear	No	No	Yes	No	Yes	Yes	Yes
Li et al [56] (2025)	No	No	Yes	Yes	Unclear	No	Yes	Yes	Unclear
Moon et al [42] (2020)	No	No	No	No	Yes	No	Yes	Yes	Unclear
Oliver et al [43] (2018)	No	No	No	Yes	Yes	No	No	No	Unclear
Period prevalence									
Bottorff et al [53] (2016)	No	No	No	Yes	Unclear	No	Yes	Yes	Unclear
Patterson et al [44] (2021)	No	Yes	Yes	No	Unclear	No	Yes	Yes	Unclear
Thrul et al [46] (2021)	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Unclear
Mojtabai et al [64] (2020)	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Unclear
Soulakova and Crockett [65] (2016)	Yes	Yes	Yes	No	Yes	No	No	Yes	Unclear
Tofighi et al [47] (2019)	No	No	No	Yes	Unclear	No	Yes	Yes	Unclear
Point prevalence									
Caballo et al [39] (2017)	Yes	Yes	Yes	No	Unclear	No	Yes	No	Yes

XSL•FO
RenderX

	Was the sample frame appropriate to address the target population?	Were study participants sampled in an appropriate way?	Was the sample size adequate?	Were the study subjects and the setting described in detail?	Was the data analysis conducted with sufficient coverage of the identified sample?	Were valid methods used for the identification of the condition?	Was the condition measured in a standard, reliable way for all participants?	Was there appropriate statistical analysis?	Was the response rate adequate, and if not, was the low response rate managed appropriately?
Curry et al [40] (2007)	Yes	Yes	Yes	No	Yes	No	Yes	No	Yes
Coups et al [60] (2009)	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes
Gravely et al [59] (2021)	No	No	Yes	No	Yes	No	Yes	Yes	Unclear
Hummel et al [26] (2018)	Yes	Yes	Yes	No	Unclear	No	No	Yes	Yes
Jackson et al [30] (2019)	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Unclear
Jackson et al [48] (2022)	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Unclear
Jackson et al [49] (2025)	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Unclear
Lund and Lund [52] (2022)	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No
Lund and Lund [51] (2023)	No	No	Yes	No	Unclear	No	Yes	No	Unclear
Papadakis et al [25] (2020)	Yes	Yes	Yes	No	Unclear	No	Yes	Yes	Yes
Perski et al [50] (2022)	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Unclear
Beard et al [62] (2016)	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Unclear
Perski et al [63] (2019)	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Unclear
Prevalence type not reported or more than one prevalence type									
Ramo et al [45] (2015)	No	No	Yes	Yes	Yes	No	Yes	No	Unclear
Twyman et al [55] (2018)	No	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Vaping cessation									
Lifetime prevalence									

XSL•FO
RenderX

	Was the sample frame appropriate to address the target population?	Were study participants sampled in an appropriate way?	Was the sample size adequate?	Were the study subjects and the setting described in detail?	Was the data analysis conducted with sufficient coverage of the identified sample?	Were valid methods used for the identification of the condition?	Was the condition measured in a standard, reliable way for all participants?	Was there appropriate statistical analysis?	Was the response rate adequate, and if not, was the low response rate managed appropriately?
Dai et al [66] (2023)	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes
Period prevalence									
Jones et al [67] (2023)	No	No	No	Yes	Yes	No	Yes	Yes	Unclear

[a]Further information on the quality assessment criteria can be found in Table S2 in Multimedia Appendix 1.

Quality Assessment of Studies on Vaping Cessation

Of the 2 studies on vaping cessation, 1 study reporting lifetime prevalences employed an appropriate sampling frame as well as appropriate sampling methods and a suitable sample size, whereas the study reporting period prevalences did not. The study methodology was only described in sufficient detail by the study on period prevalence. Both studies had sufficient sample coverage, measured the outcomes in a standard and reliable manner, and employed appropriate statistical analyses but relied on self-reported data.

Discussion

Principal Findings

This systematic review identified 25 studies examining smoking and 2 studies focusing on vaping cessation. Studies on smoking cessation have revealed a wide variability in the reported prevalences of mobile technology use for smoking cessation, ranging from 0% to 35.9% within the data collection period spanning from 2005 to 2024. The clinical and methodological variability observed was influenced by the type of prevalence measured, the specific mobile technology examined, the year of data collection, and the population under study. The lifetime prevalences of mobile technology use for smoking cessation (2.5% - 35.9 %; n=8) were found to be higher than the period prevalences (0% - 12%; n=4) and point prevalences (1.1% - 10.9%; n=11). Regardless of the prevalence type, the internet was the most frequently used technology (0.8% - 35.9%; n=14). In some countries, sex differences were observed, with males usually showing higher prevalences than females. In addition, younger people consistently showed a higher usage of mobile tools. The intention to use was much higher than the actual usage, with prevalences up to 46.7% for websites. In the 2 studies on vaping cessation, the prevalence ranged from 1.1% to 17.2%, and the intention to use mobile technologies ranged from 9.7% for web-based programs to 34.6% for apps.

Our results indicate a peak in research activity between 2014 and 2020, which may be influenced by the digital transformation of health care and the growing number of smoking cessation apps and other digital tools, which are often described as promising [11,18]. Digital tools are increasingly incorporated into public health strategies, including those of the WHO [15,16], and there is a discernible social shift in the popularity of digital platforms. However, since 2021, there has been a noticeable decline in available data, which may be due to delays in publication, shifting research priorities, or the disruptive impact of the COVID-19 pandemic on data collection and health research more broadly.

Although we were able to include more than 27 studies, those are characterized by high methodological variability, encompassing variations in target populations, settings and countries, study aims, sampling frames, outcome definitions, recall periods, and the types of mobile technology assessed. Furthermore, the quality of the majority of the included studies was found to be suboptimal, with particular concerns regarding the adequacy of the sample frame, the sampling strategy employed, and the methods used to assess the utilization of smoking cessation tools. A substantial number of studies have categorized a range of digital tools within broad terms (eg, web/mobile [52], digital aids [50], and online services [56]), without clear differentiation, which has limited the interpretability of specific usage patterns. Furthermore, there is often an absence of reporting on clinical data (eg, comorbidities or tobacco dependence), general smoking patterns (eg, the number of cigarettes smoked per day or number of recent quit attempts), and data on individuals' technical affinities (eg, mobile phone user patterns or use of other eHealth apps). These factors, however, are crucial in predicting the success of mobile technology in facilitating smoking cessation [68]. Only 2 studies focusing on vaping cessation were identified, which is somewhat surprising given that trends in e-cigarette use have shown significant increases in recent years [68], and digital interventions have been shown to be effective for vaping cessation [23].

Overall, the prevalences of mobile technology use for smoking cessation ranged from 0% to 35.9%, with the majority of studies presenting prevalences from less than 15%. Regardless of the prevalence type, the internet was the most frequently used (up to 35.9%). However, in many studies [25,26,40,44-46,55,58], "internet use" for smoking cessation was poorly defined and could include a variety of digital services that may not be

XSL•FO
RenderX

considered stand-alone smoking cessation interventions, such as static informational websites, blogs, or online forums. Furthermore, it is often unclear whether access to web-only interventions occurred via mobile devices or desktop computers. However, it is evident that in the current era, over 90% of internet users utilize smartphones to access the internet [35]. Nevertheless, it is important to recognize that this may not have been the case in previous years, and that a significant proportion of users, estimated at over 50%, still use laptops or desktop computers to access the internet [35]. Surprisingly, the use of apps to support smoking cessation remains relatively low (lifetime prevalences 2.5% - 15.2%), despite the steadily increasing number of available smoking cessation apps [69,70]. By 2017, 177 unique apps relevant to smoking cessation were identified in the iPhone App Store, 139 in Google Play Store, 70 in the BlackBerry App Store, and 55 in the Windows Phone Store [70]. Nevertheless, only 4% of the top applications recommended by leading app stores were found to have any scientific support [71], which makes it challenging for users to identify an appropriate tool [72]. Besides smoking cessation apps, social media platforms represent another promising approach, with a lifetime prevalence of use reaching up to 22.0%. Notwithstanding, only 2 studies [24,58] were identified that examined the prevalence of social media use for smoking cessation, so these numbers must be interpreted with caution. It is evident that social media platforms, which boast over 5 billion users worldwide and high penetration rates in Europe, ranging from 70.4% in Eastern Europe to 81.7% in Northern Europe [73], have the capacity to reach a vast audience. Moreover, reviews have emphasized the feasibility, acceptability, and preliminary effectiveness of social media–based interventions for smoking cessation [74,75]. However, retention rates in such interventions have been reported as low [75], suggesting that social media often functions more as a by-product than as a sustained support tool. A substantial body of evidence supports the notion of digital interventions as a complementary treatment modality for smoking cessation [17-19,70,71,76]. Nevertheless, there remains a paucity of high-quality data regarding the reach of such interventions, and the actual uptake remains low.

The use of mobile technologies for smoking cessation is influenced by a variety of factors, including cultural context, individual age, and socioeconomic status. In 2 of the included studies [25,26], which were based on data from the Europe-wide EUREST-PLUS ITC (European Regulatory Science on Tobacco—International Tobacco Control Project) Study on smoking behavior. Significant differences between countries were identified, and in both studies, the lowest reported use of mobile cessation technologies was found in Greece, where smoking prevalence remains the highest among European countries at 42%. Conversely, the higher prevalence of use was observed in countries with lower smoking prevalence, such as the United Kingdom and the Netherlands (both 12%) [6]. This finding suggests that variations in national smoking norms and tobacco control policies, digital infrastructure, access to health care, and cultural norms may influence both the demand for and the implementation of digital tools. Socioeconomic status and age are also likely to be relevant. Younger persons <40 years of age, compared to older people, showed a consistently higher use of apps, social media, and SMS text messaging programs [26,30,40,46,51]—likely due to their high digital affinity [77]. These patterns may be linked to differences in digital literacy, and especially digital health literacy [78,79], which has been described as a fundamental factor for enabling a successful digital transformation of health care systems and supporting informed decision-making in health contexts [80]. This is particularly relevant for older populations, in which levels of digital competence tend to be lower [81]. Particularly, low use was observed in vulnerable populations, such as financially disadvantaged groups (1%) or people with substance use disorders (0%) [47,55]—possibly reflecting lower motivation to quit or additional barriers to accessing support. Overall, the intention to use mobile technologies for smoking cessation was up to 46.7%, which is somewhat higher than the actual use, indicating a general openness to digital smoking cessation tools. Consequently, efforts should focus less on developing new technologies and more on effectively delivering existing evidence-based tools to populations in need. Addressing disparities in access and engagement is essential to ensure equitable benefits and avoid reinforcing health inequalities. Continuous evaluation of user needs across diverse groups will help guide targeted strategies to enhance uptake and impact.

Strengths and Limitations

To the best of our knowledge, this is the first systematic review to examine the prevalence of mobile technology use for smoking cessation. While it is possible that some studies were not identified due to the titles and abstracts of these studies not explicitly indicating the inclusion of mobile technologies, the comprehensive search strategy, which was supplemented by repeated forward and backward citation analyses, proved to be highly effective in identifying additional relevant studies.

However, it must be recognized that the present review has certain limitations. First, following the recommendations made during the peer-review process, it was decided to deviate from the established protocol and to narrow the primary focus to populations that smoke tobacco products. Studies assessing populations solely using noncombustible products were now reported only briefly as additional findings. Second, internet use was included as a form of mobile technology. However, a significant number of studies have not provided a detailed definition of internet use, which has resulted in difficulties in classifying it as a stand-alone smoking cessation tool. Moreover, although mobile devices, including smartphones, tablets, and laptops, have become prevalent instruments for accessing the internet in recent years, this has not always been the case. Consequently, the relevance of the internet as a digital smoking cessation aid may have been overestimated. Third, for better comparability, the review focused on high-income countries, which share broadly similar digital infrastructure, access to technology, tobacco control policies, and social norms around smoking; as a result, countries outside this category—such as China, which has a high research output—were not included. Furthermore, we exclusively extracted stratified prevalences on age and sex; yet, other diversity dimensions may also be important. No meta-analysis was conducted due to substantial heterogeneity across studies; yet, the findings provide a broad and current overview of existing prevalence data.

XSL•FO
RenderX

Conclusion

The prevalence of mobile technology use for smoking cessation is relatively low, and there appears to be disparities in access and engagement, with older and socially disadvantaged groups showing lower prevalences of use. However, the intention to use mobile technologies for smoking cessation is relatively high, indicating a general openness toward such tools. The included studies, however, were characterized by high methodological variability and poor reporting, so the results must be interpreted with caution. It is recommended that future research employ more standardized definitions. Despite the pervasive availability of mobile technologies designed to facilitate smoking cessation and the existence of evidence regarding the efficacy of specific tools, this review indicates that the overall real-world impact of these tools remains limited. Consequently, efforts should concentrate more on evaluating user needs and delivering existing evidence-based tools to populations in need instead of developing new technologies.

Acknowledgments

There was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript.

Funding

The authors gratefully acknowledge the support of the Leibniz ScienceCampus Bremen Digital Public health, which is jointly funded by the Leibniz Association (W72/2022), the Federal State of Bremen, and the Leibniz Institute for Prevention Research and Epidemiology–BIPS. The funder did not influence the content of the study.

Data Availability

All data generated or analyzed during this study are included in this published article and its multimedia appendices.

Authors' Contributions

Conceptualization: ML (lead), FH (supporting), BB-H (supporting), JH (supporting)
Methodology: JH (lead), FH (equal), ML (supporting)
Formal analysis: ML (lead), JH (supporting)
Data curation: ML (lead), JO (supporting)
Investigation: ML (lead), JO (supporting), JH (supporting)
Writing—original draft: ML (lead), FH (supporting), JH (supporting)
Writing—review and editing: ML (lead), FH (supporting), JH (supporting), BB-H (supporting), JO (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy, JBI Critical Appraisal Checklist, excluded articles with reason of exclusion, studies definition of smokers, and results stratified by sex/age.
[DOCX File, 53 KB - jmir_v28i1e83072_app1.docx]

Checklist 1
PRISMA-S checklist.
[PDF File, 121 KB - jmir_v28i1e83072_app2.pdf]

References

1. WHO report on the global tobacco epidemic, 2025: warning about the dangers of tobacco. : World Health Organization; 2025 URL: https://www.who.int/publications/i/item/9789240112063 [accessed 2026-03-06]
2. GBD 2019 Tobacco Collaborators. Spatial, temporal, and demographic patterns in prevalence of smoking tobacco use and attributable disease burden in 204 countries and territories, 1990-2019: a systematic analysis from the Global Burden of Disease Study 2019. Lancet 2021 Jun 19;397(10292):2337-2360. [doi: 10.1016/S0140-6736(21)01169-7] [Medline: 34051883]
3. WHO global report on trends in prevalence of tobacco use 2000–2024 and projections 2025–2030. : World Health Organization; 2025 URL: https://www.who.int/publications/i/item/9789240116276 [accessed 2026-03-06]
4. Dai X, Gakidou E, Lopez AD. Evolution of the global smoking epidemic over the past half century: strengthening the evidence base for policy action. Tob Control 2022 Mar;31(2):129-137. [doi: 10.1136/tobaccocontrol-2021-056535] [Medline: 35241576]

XSL•FO
RenderX

5. U.S. Department of Health and Human Services. The health consequences of smoking—50 years of progress: a report of the surgeon general. : U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health; 2014 URL: https://www.ncbi.nlm.nih.gov/books/NBK179276/pdf/Bookshelf_NBK179276.pdf [accessed 2026-03-27]
6. Attitudes of Europeans towards tobacco and related products. : European Commission; 2024 URL: https://europa.eu/eurobarometer/surveys/detail/2995 [accessed 2026-03-06]
7. VanFrank B, Malarcher A, Cornelius ME, Schecter A, Jamal A, Tynan M. Adult smoking cessation - United States, 2022. MMWR Morb Mortal Wkly Rep 2024 Jul 25;73(29):633-641. [doi: 10.15585/mmwr.mm7329a1] [Medline: 39052529]
8. Taylor AH, Ussher M. Physical activity as an aid in smoking cessation. In: Routledge Handbook of Physical Activity and Mental Health: Routledge; 2013:451-464. [doi: 10.4324/9781003423430-41]
9. Kotz D, Batra A, Kastaun S. Smoking cessation attempts and common strategies employed. Dtsch Arztebl Int 2020 Jan 6;117(1-2):7-13. [doi: 10.3238/arztebl.2020.0007] [Medline: 32008606]
10. Measuring digital development facts and figures 2025. : International Telecommunication Union (ITU); 2025 URL: https://www.itu.int/dms_pub/itu-d/opb/ind/d-ind-ict_mdd-2025-3-pdf-e.pdf [accessed 2026-03-06]
11. Whittaker R, McRobbie H, Bullen C, Rodgers A, Gu Y. Mobile phone-based interventions for smoking cessation. Cochrane Database Syst Rev 2016 Apr 10;4(4):CD006611. [doi: 10.1002/14651858.CD006611.pub4] [Medline: 27060875]
12. Wienert J, Jahnel T, Maaß L. What are digital public health interventions? First steps toward a definition and an intervention classification framework. J Med Internet Res 2022 Jun 28;24(6):e31921. [doi: 10.2196/31921] [Medline: 35763320]
13. Liu S, Ma J, Sun M, Zhang C, Gao Y, Xu J. Mapping the landscape of digital health intervention strategies: 25-year synthesis. J Med Internet Res 2025 Jan 13;27:e59027. [doi: 10.2196/59027] [Medline: 39804697]
14. Alsaqqa HH, Alwawi A. Digital intervention for public health: searching for implementing characteristics, concepts and recommendations: scoping review. Front Public Health 2023;11:1142443. [doi: 10.3389/fpubh.2023.1142443] [Medline: 37790710]
15. Global strategy on digital health 2020-2025. : World Health Organization; 2021 URL: https://iris.who.int/server/api/core/bitstreams/1f4d4a08-b20d-4c36-9148-a59429ac3477/content [accessed 2026-03-06]
16. WHO clinical treatment guideline for tobacco cessation in adults. : World Health Organization; 2024 URL: https://iris.who.int/server/api/core/bitstreams/2deb01bc-1be9-4e9c-b113-5efbb67780c7/content [accessed 2026-03-06]
17. Fang YE, Zhang Z, Wang R, et al. Effectiveness of eHealth smoking cessation interventions: systematic review and meta-analysis. J Med Internet Res 2023 Jul 28;25:e45111. [doi: 10.2196/45111] [Medline: 37505802]
18. Rushender R, Logaraj M, Krishnamoorthy Y. Effectiveness of mobile phone applications for tobacco cessation: an umbrella review. Drug Alcohol Depend 2024 Oct 1;263:112425. [doi: 10.1016/j.drugalcdep.2024.112425] [Medline: 39216199]
19. Guo YQ, Chen Y, Dabbs AD, Wu Y. The effectiveness of smartphone app-based interventions for assisting smoking cessation: systematic review and meta-analysis. J Med Internet Res 2023 Apr 20;25:e43242. [doi: 10.2196/43242] [Medline: 37079352]
20. Amin S, Billah MS, Jahra Jarin E. A systematic review of digital intervention of smoking abstinence during the COVID-19 era. Int J Health Sci Res 2024;14(3):35-42. [doi: 10.52403/ijhsr.20240307]
21. Li S, Li Y, Xu C, et al. Efficacy of digital interventions for smoking cessation by type and method: a systematic review and network meta-analysis. Nat Hum Behav 2025 Oct;9(10):2054-2065. [doi: 10.1038/s41562-025-02295-2] [Medline: 40897803]
22. Di Palo MP, Di Spirito F, Garofano M, et al. Effectiveness and adherence of standalone digital tobacco cessation modalities: a systematic review of systematic reviews. Healthcare (Basel) 2025 Aug 26;13(17):2125. [doi: 10.3390/healthcare13172125] [Medline: 40941478]
23. Bhadania M, Yadav S, Mathur A, Acharya A, Mehta V. Mapping the effectiveness of digital health interventions in promoting electronic cigarette cessation: a systematic review and meta-analysis of randomized controlled trials. BMC Public Health 2025 Oct 28;25(1):3645. [doi: 10.1186/s12889-025-24990-5] [Medline: 41152828]
24. Borrelli B, Bartlett YK, Tooley E, Armitage CJ, Wearden A. Prevalence and frequency of mHealth and eHealth use among US and UK smokers and differences by motivation to quit. J Med Internet Res 2015 Jul 4;17(7):e164. [doi: 10.2196/jmir.4420] [Medline: 26149323]
25. Papadakis S, Katsaounou P, Kyriakos CN, et al. Quitting behaviours and cessation methods used in eight European Countries in 2018: findings from the EUREST-PLUS ITC Europe Surveys. Eur J Public Health 2020 Jul 1;30(Suppl_3):iii26-iii33. [doi: 10.1093/eurpub/ckaa082] [Medline: 32918825]
26. Hummel K, Nagelhout GE, Fong GT, et al. Quitting activity and use of cessation assistance reported by smokers in eight European countries: findings from the EUREST-PLUS ITC Europe Surveys. Tob Induc Dis 2018;16(Suppl 2):1-17. [doi: 10.18332/tid/98912] [Medline: 31363422]
27. Strecher VJ, McClure J, Alexander G, et al. The role of engagement in a tailored web-based smoking cessation program: randomized controlled trial. J Med Internet Res 2008 Nov 4;10(5):e36. [doi: 10.2196/jmir.1002] [Medline: 18984557]
28. Zeng EY, Vilardaga R, Heffner JL, Mull KE, Bricker JB. Predictors of utilization of a novel smoking cessation smartphone app. Telemed J E Health 2015 Dec;21(12):998-1004. [doi: 10.1089/tmj.2014.0232] [Medline: 26171733]

XSL•FO
RenderX

29. Wangberg SC, Bergmo TS, Johnsen JAK. Adherence in internet-based interventions. Patient Prefer Adherence 2008 Feb 2;2:57-65. [Medline: 19920945]
30. Jackson SE, Kotz D, West R, Brown J. Moderators of real-world effectiveness of smoking cessation aids: a population study. Addiction 2019 Sep;114(9):1627-1638. [doi: 10.1111/add.14656] [Medline: 31117151]
31. Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z. JBI Manual for Evidence Synthesis: Joanna Briggs Institute; 2024. URL: https://jbi-global-wiki.refined.site/space/MANUAL [accessed 2026-03-27]
32. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
33. Hirt J, Nordhausen T, Fuerst T, Ewald H, Appenzeller-Herzog C, TARCiS study group. Guidance on terminology, application, and reporting of citation searching: the TARCiS statement. BMJ 2024 May 9;385:e078384. [doi: 10.1136/bmj-2023-078384] [Medline: 38724089]
34. Munn Z, Stern C, Aromataris E, Lockwood C, Jordan Z. What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. BMC Med Res Methodol 2018 Jan 10;18(1):5. [doi: 10.1186/s12874-017-0468-4] [Medline: 29316881]
35. Digital 2026 global overview report. We Are Social. URL: https://wearesocial.com/uk/blog/2025/10/digital-2026-global-overview-report/ [accessed 2026-03-06]
36. ODA recipients: countries, territories, and international organisations. Organisation for Economic Co-operation and Development. URL: https://www.oecd.org/en/topics/sub-issues/oda-eligibility-and-conditions/dac-list-of-oda-recipients.html [accessed 2026-03-06]
37. Munn Z, Moola S, Lisy K, Riitano D, Tufanaru C. Methodological guidance for systematic reviews of observational epidemiological studies reporting prevalence and cumulative incidence data. Int J Evid Based Healthc 2015 Sep;13(3):147-153. [doi: 10.1097/XEB.0000000000000054] [Medline: 26317388]
38. Hoffmann F, Eggers D, Pieper D, Zeeb H, Allers K. An observational study found large methodological heterogeneity in systematic reviews addressing prevalence and cumulative incidence. J Clin Epidemiol 2020 Mar;119:92-99. [doi: 10.1016/j.jclinepi.2019.12.003] [Medline: 31809847]
39. Caraballo RS, Shafer PR, Patel D, Davis KC, McAfee TA. Quit methods used by US adult cigarette smokers, 2014-2016. Prev Chronic Dis 2017 Apr 13;14:E32. [doi: 10.5888/pcd14.160600] [Medline: 28409740]
40. Curry SJ, Sporer AK, Pugach O, Campbell RT, Emery S. Use of tobacco cessation treatments among young adult smokers: 2005 National Health Interview Survey. Am J Public Health 2007 Aug;97(8):1464-1469. [doi: 10.2105/AJPH.2006.103788] [Medline: 17600243]
41. Graham EL, Koralnik IJ, Liotta EM. Therapeutic approaches to the neurologic manifestations of COVID-19. Neurotherapeutics 2022 Sep;19(5):1435-1466. [doi: 10.1007/s13311-022-01267-y] [Medline: 35861926]
42. Moon J, Rigg JS, Smith JE. Korean American smokers' perspectives on mobile smoking cessation applications. Tob Use Insights 2020;13:1179173X20972384. [doi: 10.1177/1179173X20972384] [Medline: 33402856]
43. Oliver JA, Hallyburton MB, Pacek LR, et al. What do smokers want in a smartphone-based cessation application? Nicotine Tob Res 2018 Nov 15;20(12):1507-1514. [doi: 10.1093/ntr/ntx171] [Medline: 29065202]
44. Patterson JG, Hinton A, Cooper SE, Wewers ME. Differences in quit attempts, successful quits, methods, and motivations in a longitudinal cohort of adult tobacco users by sexual orientation. Nicotine Tob Res 2021 Oct 7;23(11):1952-1957. [doi: 10.1093/ntr/ntab116] [Medline: 34060633]
45. Ramo DE, Liu H, Prochaska JJ. A mixed-methods study of young adults' receptivity to using Facebook for smoking cessation: if you build it, will they come? Am J Health Promot 2015;29(4):e126-e135. [doi: 10.4278/ajhp.130326-QUAL-128] [Medline: 24575728]
46. Thrul J, Riehm KE, Cohen JE, Alexander GC, Vernick JS, Mojtabai R. Tobacco control policies and smoking cessation treatment utilization: a moderated mediation analysis. PLoS ONE 2021;16(8):e0241512. [doi: 10.1371/journal.pone.0241512] [Medline: 34460821]
47. Tofighi B, Lee JD, Sherman S, Schatz D, El-Shahawy O. Smoking patterns and preferences for technology assisted smoking cessation interventions among adults with opioid and alcohol use disorders. J Subst Use 2019;24(6):660-665. [doi: 10.1080/14659891.2019.1642407] [Medline: 32952442]
48. Jackson SE, Cox S, Shahab L, Brown J. Prevalence of use and real-world effectiveness of smoking cessation aids during the COVID-19 pandemic: a representative study of smokers in England. Addiction 2022 Sep;117(9):2504-2514. [doi: 10.1111/add.15903] [Medline: 35491743]
49. Jackson SE, Brown J, Buss V, Shahab L. Prevalence of popular smoking cessation aids in England and associations with quit success. JAMA Netw Open 2025 Jan 2;8(1):e2454962. [doi: 10.1001/jamanetworkopen.2024.54962] [Medline: 39821398]
50. Perski O, West R, Brown J. Estimated failure to report unsuccessful quit attempts by type of cessation aid: a population survey of smokers in England. J Smok Cessat 2022;2022:5572480. [doi: 10.1155/2022/5572480] [Medline: 36568905]
51. Lund I, Lund M. Quit smoking: methods and outcomes for Norwegian adults. Discov Soc Sci Health 2023;3:12. [doi: 10.1007/s44155-023-00043-3]

XSL•FO
RenderX

52. Lund M, Lund I. Smoking cessation aids and strategies: a population-based survey of former and current smokers in Norway. BMC Public Health 2022 Mar 31;22(1):631. [doi: 10.1186/s12889-022-13032-z] [Medline: 35361172]
53. Bottorff JL, Oliffe JL, Sarbit G, et al. Evaluation of quitnow men: an online, men-centered smoking cessation intervention. J Med Internet Res 2016 Apr 20;18(4):e83. [doi: 10.2196/jmir.5076] [Medline: 27097991]
54. Jayakumar N, Chaiton M, Zhang B, Selby P, Schwartz R. Sex differences in use of smoking cessation services and resources: a real-world study. Tob Use Insights 2020;13:1179173X20901500. [doi: 10.1177/1179173X20901500] [Medline: 32030067]
55. Twyman L, Bonevski B, Paul C, et al. What factors are associated with abstinence amongst socioeconomically disadvantaged smokers? A cross-sectional survey of use of cessation aids and quitting approach. Drug Alcohol Rev 2018 Feb;37(2):170-179. [doi: 10.1111/dar.12561] [Medline: 28616900]
56. Li W, Paul CL, Baker AL, et al. Quit attempts, use of smoking cessation treatments and quitting intention among treatment seekers in rural areas. J Subst Use Addict Treat 2025 Feb;169:209603. [doi: 10.1016/j.josat.2024.209603] [Medline: 39672339]
57. Chevalking SL, Ben Allouch S, Brusse-Keizer M, Postel MG, Pieterse ME. Identification of users for a smoking cessation mobile app: quantitative study. J Med Internet Res 2018;20(4):e118. [doi: 10.2196/jmir.7606]
58. Kostagiolas P, Parnavela S, Theodorou P. The impact of smokers' information-seeking behavior on smoking cessation. Adv Exp Med Biol 2023;1425:645-662. [doi: 10.1007/978-3-031-31986-0_63] [Medline: 37581838]
59. Gravely S, Cummings KM, Hammond D, et al. Self-reported quit aids and assistance used by smokers at their most recent quit attempt: findings from the 2020 International Tobacco Control Four Country Smoking and Vaping Survey. Nicotine Tob Res 2021 Aug 29;23(10):1699-1707. [doi: 10.1093/ntr/ntab068] [Medline: 33837435]
60. Coups EJ, Dhingra LK, Heckman CJ, Manne SL. Receipt of provider advice for smoking cessation and use of smoking cessation treatments among cancer survivors. J Gen Intern Med 2009 Nov;24 Suppl 2(Suppl 2):S480-S486. [doi: 10.1007/s11606-009-0978-9] [Medline: 19838854]
61. Graham AL, Amato MS. Twelve million smokers look online for smoking cessation help annually: Health Information National Trends Survey Data, 2005-2017. Nicotine Tob Res 2019 Jan 4;21(2):249-252. [doi: 10.1093/ntr/nty043] [Medline: 29660037]
62. Beard E, Brown J, Michie S, Kaner E, Meier P, West R. Use of aids for smoking cessation and alcohol reduction: a population survey of adults in England. BMC Public Health 2016 Dec 8;16(1):1237. [doi: 10.1186/s12889-016-3862-7] [Medline: 27931202]
63. Perski O, Jackson SE, Garnett C, West R, Brown J. Trends in and factors associated with the adoption of digital aids for smoking cessation and alcohol reduction: a population survey in England. Drug Alcohol Depend 2019 Dec 1;205:107653. [doi: 10.1016/j.drugalcdep.2019.107653] [Medline: 31675544]
64. Mojtabai R, Riehm KE, Cohen JE, Alexander GC, Vernick JS, Thrul J. Cigarette excise taxes, clean indoor air laws, and use of smoking cessation treatments: a mediation analysis. Prev Med 2020 Jul;136:106098. [doi: 10.1016/j.ypmed.2020.106098] [Medline: 32333928]
65. Soulakova JN, Crockett LJ. Unassisted quitting and smoking cessation methods used in the United States: analyses of 2010-2011 Tobacco Use Supplement to the Current Population Survey Data. Nicotine Tob Res 2017 Dec 13;20(1):30-39. [doi: 10.1093/ntr/ntw273] [Medline: 27798084]
66. Dai HD, Hanh P, Guenzel N, Morgan M, Kerns E, Winickoff JP. Adoption of vaping cessation methods by US adolescent e-cigarette users. Pediatrics 2023 Nov 1;152(5):e2023062948. [doi: 10.1542/peds.2023-062948] [Medline: 37781732]
67. Jones E, Endrighi R, Weinstein D, Jankowski A, Quintiliani LM, Borrelli B. Methods used to quit vaping among adolescents and associations with perceived risk, addiction, and socio-economic status. Addict Behav 2023 Dec;147:107835. [doi: 10.1016/j.addbeh.2023.107835] [Medline: 37598643]
68. Mattingly DT, Hart JL. Trends in current electronic cigarette use among youths by age, sex, and race and ethnicity. JAMA Netw Open 2024 Feb 5;7(2):e2354872. [doi: 10.1001/jamanetworkopen.2023.54872] [Medline: 38315486]
69. Kalgotra P, Raja U, Sharda R. Growth in the development of health and fitness mobile apps amid COVID-19 pandemic. Digit Health 2022;8:20552076221129070. [doi: 10.1177/20552076221129070] [Medline: 36211794]
70. Cobos-Campos R, Cordero-Guevara JA, Apiñaniz A, et al. The impact of digital health on smoking cessation. Interact J Med Res 2023 Mar 15;12:e41182. [doi: 10.2196/41182] [Medline: 36920468]
71. Haskins BL, Lesperance D, Gibbons P, Boudreaux ED. A systematic review of smartphone applications for smoking cessation. Transl Behav Med 2017 Jun;7(2):292-299. [doi: 10.1007/s13142-017-0492-2] [Medline: 28527027]
72. Bold KW, Garrison KA, DeLucia A, et al. Smartphone apps for smoking cessation: systematic framework for app review and analysis. J Med Internet Res 2023 Jul 13;25:e45183. [doi: 10.2196/45183] [Medline: 37440305]
73. Digital 2024: 5 billion social media users. We Are Social. URL: https://wearesocial.com/uk/blog/2024/01/digital-2024-5-billion-social-media-users/ [accessed 2026-03-06]
74. Luo T, Li MS, Williams D, et al. Using social media for smoking cessation interventions: a systematic review. Perspect Public Health 2021 Jan;141(1):50-63. [doi: 10.1177/1757913920906845]
75. Naslund JA, Kim SJ, Aschbrenner KA, et al. Systematic review of social media interventions for smoking cessation. Addict Behav 2017 Oct;73:81-93. [doi: 10.1016/j.addbeh.2017.05.002] [Medline: 28499259]

XSL•FO
RenderX

76. Gliwska E, Mańczuk M. Digital interventions in smoking cessation – a brief overview of systematic reviews and meta-analyses. Nowotwory 2024;74(1):66-71. [doi: 10.5603/njo.98341]
77. Young people - digital world. Eurostat. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Young_people_-_digital_world [accessed 2026-03-06]
78. Papp-Zipernovszky O, Horváth MD, Schulz PJ, Csabai M. Generation gaps in digital health literacy and their impact on health information seeking behavior and health empowerment in Hungary. Front Public Health 2021;9(May):635943. [doi: 10.3389/fpubh.2021.635943] [Medline: 34055714]
79. Hsu WC. The effect of age on electronic health literacy: mixed-method study. JMIR Hum Factors 2019 Apr 21;6(2):e11480. [doi: 10.2196/11480] [Medline: 31066696]
80. Dratva J, Schaeffer D, Zeeb H. Digital health literacy in Germany: current status, concepts, and challenges. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2024 Mar;67(3):277-284. [doi: 10.1007/s00103-024-03841-5] [Medline: 38315221]
81. Estrela M, Semedo G, Roque F, Ferreira PL, Herdeiro MT. Sociodemographic determinants of digital health literacy: a systematic review and meta-analysis. Int J Med Inform 2023 Sep;177(May):105124. [doi: 10.1016/j.ijmedinf.2023.105124] [Medline: 37329766]

Abbreviations

CoCoPop: condition, context, population
DEBRA : Deutsche Befragung zum Rauchverhalten
e-cigarette: electronic cigarette
EUREST-PLUS ITC: European Regulatory Science on Tobacco—International Tobacco Control Project
OECD: Organisation for Economic Co-operation and Development
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
WHO: World Health Organization

Edited by S Brini; submitted 29.Aug.2025; peer-reviewed by S Cha, S McIntosh; revised version received 26.Jan.2026; accepted 27.Jan.2026; published 30.Apr.2026.

Please cite as:
Lakeberg M, Hoffmann F, Ohmes J, Berger-Höger B, Helbach J
Use of Mobile Technologies for Smoking Cessation Among Smokers and Former Smokers: Systematic Review
J Med Internet Res 2026;28:e83072
URL: https://www.jmir.org/2026/1/e83072
doi:10.2196/83072

© Meret Lakeberg, Falk Hoffmann, Jannik Ohmes, Birte Berger-Höger, Jasmin Helbach. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effects of Stand-Alone Digital Lifestyle Interventions on Weight-Related Outcomes in Adults With Overweight or Obesity: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Si-An Lee[1], MSc; Jin-Hyuck Park[2], PhD

[1]Department of ICT convergence, The Graduate School, Soonchunhyang University, Asan, Republic of Korea

[2]Department of Occupational Therapy, College of Medical Science, Soonchunhyang University, Room 1401, Medical Science, 22, Soonchunhayng-ro, Shinchang-myeon, Asan, Republic of Korea

Corresponding Author:
Jin-Hyuck Park, PhD
Department of Occupational Therapy, College of Medical Science, Soonchunhyang University, Room 1401, Medical Science, 22, Soonchunhayng-ro, Shinchang-myeon, Asan, Republic of Korea

Abstract

Background: Obesity is a major global health concern, and scalable digital solutions are urgently needed. While digital lifestyle interventions (DLSIs) have shown promise, prior meta-analyses often included hybrid formats with human support, limiting insights into the effectiveness of fully digital interventions.

Objective: This study aimed to evaluate the independent effects of standalone DLSIs—defined as interventions delivered exclusively via digital platforms without in-person or adjunctive support—on anthropometric and dietary outcomes in adults with overweight or obesity.

Methods: We searched MEDLINE, Embase, PsycINFO, Web of Science, and the Cochrane Library from inception through March 4, 2026. Eligible studies were randomized controlled trials (RCTs) evaluating stand-alone DLSIs in adults with overweight or obesity. Interventions were included if they targeted diet or physical activity exclusively through digital platforms. We included fully automated, asynchronous, or one-to-many synchronous systems without individualized support. Studies involving hybrid interventions, including one-to-one synchronous human interaction, nonadult populations, or non-RCT designs, were excluded. Two independent reviewers performed study selection and data extraction. Risk of bias was assessed using the Cochrane Risk of Bias 2.0 tool (Cochrane Bias Methods Group). Meta-analysis used a random-effects model with the Hartung-Knapp-Sidik-Jonkman method, and heterogeneity was assessed using I^2 statistics. The certainty of evidence was evaluated using the Grading of Recommendations, Assessment, Development, and Evaluation approach.

Results: A total of 19 RCTs involving 3556 participants were included. Stand-alone DLSIs significantly improved anthropometric outcomes compared to controls (standardized mean difference 0.26, 95% CI 0.14 - 0.38; 95% prediction interval [PI] −0.16 to 0.68; P<.001; 19 studies; n=3556; I^2=56.1%), corresponding to an additional weight loss ranging from 2.62 kg to 6.55 kg, depending on the baseline body weight. Significant improvements were also found in dietary outcomes (standardized mean difference 0.26, 95% CI 0.04 - 0.48; 95% PI −0.29 to 0.81; P=.008; 8 studies; n=1365; I^2=57.5%). Subgroup analyses for anthropometric outcomes revealed significant differences only by control group type (P<.001), with waitlist controls showing the largest effect. For dietary outcomes, no significant subgroup differences were found (P>.05). While most studies showed a low risk of bias, substantial statistical heterogeneity was observed in some outcomes. Consequently, the certainty of evidence for both outcomes was rated as moderate.

Conclusions: This review is innovative as it is the first to isolate the pure efficacy of stand-alone DLSIs by excluding synchronous human support. Our findings provide moderate-certainty evidence that these tools are effective for weight management and dietary improvement without human intervention. While stand-alone DLSIs offer a highly scalable, cost-effective first-step intervention, the PIs included zero, and substantial heterogeneity was observed, suggesting that benefits may vary across settings. Future research should identify user characteristics that maximize engagement with unguided digital tools.

Trial Registration: PROSPERO CRD420251053974; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251053974

(J Med Internet Res 2026;28:e81070) doi:10.2196/81070

XSL•FO
RenderX

KEYWORDS

digital intervention; digital device; lifestyle; weight; diet

Introduction

Obesity remains a critical global health challenge, with far-reaching implications for morbidity, mortality, and health care systems. According to the latest data from the World Health Organization (WHO), the global prevalence of obesity among adults aged 18 years and older has more than doubled, rising from 7% in 1990 to 16% in 2022 [1]. Recent evidence from 2023 further indicates that if current trends persist, over 4 billion people, more than half of the world's population, will be living with overweight or obesity by 2035, placing an unprecedented economic burden on global health care infrastructures [2]. This escalating trend is projected to exert a staggering economic impact, with the global economic burden estimated to reach US $4.32 trillion annually by 2035, a figure comparable to the impact of the COVID-19 pandemic [2]. Overweight or obesity is a major risk factor for numerous noncommunicable diseases—including cardiovascular disease, type 2 diabetes, musculoskeletal disorders, and certain cancers—and contributes substantially to the global burden of disease and premature death [3]. This growing epidemic necessitates accessible, sustainable, and scalable strategies for weight management.

Lifestyle interventions, which typically integrate dietary modification, physical activity enhancement, and behavioral change techniques, have emerged as cornerstone strategies in the clinical management of overweight or obesity [4]. Traditionally, these interventions have been delivered through face-to-face counseling or group-based sessions and have demonstrated efficacy in reducing body weight and improving cardiometabolic outcomes [5]. Structured behavioral components, such as goal setting, self-monitoring, and problem solving, are widely recognized and incorporated into clinical practice guidelines for overweight or obesity treatment [6].

However, a key limitation of traditional lifestyle interventions is their reliance on in-person delivery, which poses considerable barriers to access. The requirement for physical attendance may hinder participation among individuals in rural or underserved communities, those with mobility or scheduling constraints, and those facing socioeconomic disadvantages [7]. These accessibility challenges were further exacerbated during the COVID-19 pandemic, which disrupted in-person care worldwide and underscored the urgent need for scalable alternatives [8]. In the postpandemic era, there has been a significant shift toward digital-first health care models, driven by the need for cost-effective solutions that can bypass the limitations of human-led medical resources [9].

Digital health technologies offer a compelling solution to these limitations. Digital lifestyle interventions (DLSIs)—including web-based platforms, mobile apps, and SMS text messaging–based systems—enable the delivery of personalized, interactive, and cost-effective behavioral support at scale [10]. These interventions can promote engagement through real-time feedback, self-monitoring tools, and flexible access, making them particularly well-suited for wide implementation across diverse populations [11]. Recent advancements in artificial intelligence (AI) and automated tailoring have further enhanced the potential of these platforms to provide high-quality, personalized support without the need for constant human supervision [10,12].

Prior systematic reviews have reported that digital interventions yield modest benefits over conventional offline approaches in facilitating weight loss [12,13]. However, these conclusions are often constrained by the inclusion of hybrid interventions that combine digital components with adjunctive modalities, such as web-based platforms supplemented by therapist support or face-to-face counseling. Such designs make it difficult to isolate the independent effect of digital platforms. Further complicating the evidence base, one meta-analysis included online interventions delivered in group formats—such as virtual group meetings—which, while not involving in-person contact, may introduce behavioral confounders, such as peer accountability and social facilitation [14]. Additionally, another review focused exclusively on web-based interventions, excluding alternative digital modalities such as mobile apps or SMS text messaging–based tools, and thereby failed to capture the broader spectrum of modern digital delivery systems [15].

Given the limitations of prior evidence syntheses—many of which have included hybrid or group-mediated interventions—there remains a critical need for a focused meta-analysis that isolates the independent impact of digital delivery. While hybrid models integrating in-person coaching have shown promise, their high operational costs and lack of scalability limit their utility for large-scale public health programs [16]. This creates a critical knowledge gap regarding the stand-alone efficacy of digital technology. By isolating digital interventions from individualized human-facilitated components, this systematic review and meta-analysis provides a clearer estimation of digital intervention effects. This methodological refinement eliminates confounding from human-facilitated components, allowing for a clearer assessment of the technology's intrinsic impact. Furthermore, by synthesizing data across a wide range of delivery formats, including mobile apps, web platforms, and SMS text messaging–based systems, this study captures the contemporary landscape of digital health tools. This methodological focus allows for a more precise and high-certainty assessment of the independent effects of stand-alone DLSIs on weight management in adults with overweight or obesity.

Therefore, the primary objective of this study was to evaluate the efficacy of stand-alone DLSIs on anthropometric and dietary outcomes in adults with overweight or obesity. Specifically, we aimed to (1) quantify the pooled effect size of these interventions compared to various control conditions, (2) identify potential moderators of effectiveness through subgroup analyses, and (3) assess the certainty of the evidence to provide clinical and policy recommendations for scalable weight management.

XSL•FO
RenderX

Methods

Overview

This systematic review and meta-analysis follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines and was prospectively registered with PROSPERO (CRD420251053974; Checklist 1). While the study was conducted according to the preregistered protocol, we additionally used the Hartung-Knapp-Sidik-Jonkman (HKSJ) method for the random-effects model and calculated 95% prediction intervals (PIs) to provide a more conservative and robust estimation of the evidence, which were not explicitly detailed in the initial protocol.

Eligibility Criteria

Studies were included if they met the following criteria: (1) adults (≥18 years) with overweight or obesity (BMI≥23 kg/m^2); (2) the intervention involved lifestyle interventions aimed at weight management, delivered exclusively via digital platforms or technologies, including mobile apps, computer software, websites, or SMS text messaging; (3) the intervention was delivered independently without any in-person human support. We defined stand-alone DLSIs as those where the core therapeutic components—including feedback, goal setting, and progress monitoring—were driven exclusively by digital platforms. Specifically, we included all digital interventions that did not require physical, face-to-face clinical contact, such as automated or periodic app notifications, SMS text messaging reminders, and preprogrammed, asynchronous feedback (eg, batch-sent educational emails or system-generated responses triggered by user-entered data), as these represent scalable alternatives to traditional resource-intensive programs. For studies using potentially ambiguous terminology (eg, “coaching,” “messaging,” or “peer support”), we verified the non–in-person nature of these components by cross-referencing study protocols or technical implementation sections. This verification ensured that any human involvement was neither in-person nor one-to-one support; (4) a non-DLSI control group was used for comparison, such as active controls, nonspecific active controls, or waitlist controls; (5) outcomes included anthropometric measures (eg, body weight, BMI, or waist circumference) or dietary measures (eg, caloric intake); and (6) the study design was a randomized controlled trial (RCT). Exclusion criteria were as follows: (1) participants were not primarily selected based on overweight or obesity status; (2) participants were pregnant women; (3) any intervention incorporating in-person human support (eg, face-to-face counseling or physical clinical visits) where the independent effect of the digital technology could not be isolated; (4) interventions involving resource-intensive, one-to-one digital interaction with human providers. Specifically, we excluded labor-intensive, real-time interaction such as personalized phone calls or individual video conferencing, as these do not align with the scalable nature of stand-alone DLSIs; (5) studies included DLSIs in the control group; and (6) studies lacking sufficient data for meta-analysis were excluded. For potentially eligible studies with incomplete or ambiguous data, we attempted to contact the corresponding authors to obtain the missing information. If no response was received or the provided data remained inadequate for calculating effect sizes, the studies were excluded from both the systematic review and the meta-analysis. Both peer-reviewed papers and conference abstracts were included in the main analysis, although conference abstracts were required to report RCT findings explicitly.

Information Sources

The electronic databases were searched by the corresponding author using MEDLINE, Embase, PsycINFO, Web of Science, and the Cochrane Library to search for peer-reviewed papers in English, with no start date and no language restrictions. To ensure the specific indexing and search capabilities of each database were fully used, all databases were searched individually through their respective official interfaces rather than using a multidatabase search platform. We did not search study registries (eg, the WHO International Clinical Trials Registry Platform) for unpublished or ongoing trials, as our search was focused on peer-reviewed, published literature.

Search Strategy

We conducted and reported our literature search in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search Extension) guidelines to ensure transparency and reproducibility [17] (Checklist 2). The search strategies were developed independently for this study and did not adopt or reuse search strings from prior literature reviews. The strategies were constructed by combining 3 main domains (digital interventions, obesity or overweight, and weight management) using a comprehensive set of Medical Subject Headings and relevant keywords tailored to each database’s indexing structure [15]. To ensure transparency and minimize the risk of missing studies, we followed the Cochrane Handbook guidelines for search strategy development. The full, unabridged search strategy for each database is provided in Multimedia Appendix 1. However, the search strategy did not undergo a formal peer review process by an external librarian.

The reference lists of all included studies and relevant studies were searched manually, and the gray literature was also searched by entering the same keywords into Google on May 18, 2025. During the preparation and revision of the meta-analysis, the search was repeated 4 times to identify newly published trials. The last search was conducted on March 4, 2026, following the previous update on June 30, 2025.

Selection Process

After eliminating duplicate records, a 2-phase screening procedure was implemented: initial screening of titles and abstracts, followed by a full-text review. Two independent reviewers (SAL and JHP) assessed the eligibility of each record at both stages using Covidence (Veritas Health Innovation). To maintain consistency across the process, a random 10% subset of entries was jointly reviewed during each screening phase to calculate interrater agreement. Any disagreements were resolved through discussion and consensus between the 2 reviewers until a mutual agreement was reached. Interrater reliability, quantified using Cohen kappa, indicated near-perfect agreement at both the title and abstract screening stage (0.92) and full-text review

XSL•FO
RenderX

(0.94). Reviewers remained blinded to each other's assessments throughout to ensure objectivity. After completing the literature search, 2 independent reviewers screened the titles and abstracts of all retrieved records to assess eligibility. Full-text papers were subsequently obtained for trials deemed potentially relevant and reviewed in detail for inclusion. In addition, the reference lists of all included papers and related review papers were manually examined to identify any supplementary studies.

Data Collection

Data were extracted from each included study by 2 independent reviewers (SAL and JHP) using a standardized, pilot-tested data extraction form to ensure accuracy and consistency. Any discrepancies in the extracted data were resolved through detailed discussion and internal cross-checking between the 2 reviewers. If essential data were missing or unclear in the original reports, we attempted to contact the primary authors via email to obtain the necessary information.

Data Items

From each included study, the following information was extracted when available: (1) bibliographical data (ie, first author and year); (2) sample characteristics (ie, sample size, mean age, and population); (3) intervention characteristics (ie, details of intervention and control conditions, target lifestyle, theoretical framework, delivery format, duration, and outcome domains); (4) participant retention rates, defined as the proportion of participants in the intervention group who completed the postintervention assessment relative to the total number of participants initially randomized; and (5) data required to calculate within-group or between-group effect sizes for RCTs. To calculate effect sizes, both baseline (preintervention) and postintervention data were extracted for all outcomes. However, follow-up data were excluded from the meta-analysis to maintain consistency across studies and avoid bias resulting from varying follow-up durations.

Eligible interventions were defined as those aiming to change diet and/or physical activity behaviors using established behavior change techniques (BCTs; eg, goal setting and self-monitoring). Interventions were included if at least 50% of their content targeted behavior change in either domain and incorporated one or more BCTs listed in the BCT Taxonomy v1 [18]. Detailed BCT classification for each included study is provided in Multimedia Appendix 2. The primary outcome was anthropometric measures, including body weight, BMI, and waist circumference. When a study reported multiple anthropometric variables, we applied a predefined hierarchy to select a single outcome for the primary analysis: (1) body weight, (2) BMI, and (3) waist circumference. Secondary outcomes included dietary measures, with total energy intake (kcal/d) prioritized as the primary dietary metric.

Study Risk of Bias Assessment

Risk of bias in individual studies was appraised independently by 2 independent reviewers using the Cochrane Collaboration's Risk of Bias 2 tool guidelines in 5 domains: randomization process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported result [19]. A study was categorized as high risk if at least one domain was rated "high risk." If there were no high-risk domains but at least one domain raised concerns, the study was rated as "some concerns." Only studies with low risk in all domains were designated as "low risk."

Effect Measures

The primary outcomes (anthropometric measures) and secondary outcomes (dietary measures) were continuous variables. To account for the diversity in measurement scales across studies, standardized mean differences (SMDs) with 95% CIs were used as the summary effect measure. SMD estimates of 0.20, 0.50, and 0.80 were interpreted as small, moderate, and large effect sizes, respectively [20]. For outcomes where a lower value indicates clinical improvement (eg, body weight, BMI, and waist circumference), the direction of the effect size was aligned during data coding in the Comprehensive Meta-Analysis software (Biostat) so that positive SMDs consistently represent a greater reduction in the intervention group. To avoid double-counting in the meta-analysis and ensure that each study contributed independent participants, we only included one pair of baseline and postintervention data per study in the primary analysis, selected according to the aforementioned hierarchy. SMDs were calculated using the change from baseline to postintervention for both the intervention and control groups. In cases where a study reported outcomes for multiple intervention arms compared to a single control group, we combined the intervention groups or split the control group to maintain the independence of observations. When change scores were not directly reported, they were calculated using baseline and postintervention means and SDs, assuming a correlation coefficient of 0.5 where necessary.

Synthesis Methods

Quantitative synthesis was performed using a random-effects model, as clinical and methodological heterogeneity was expected across the included stand-alone digital interventions. To improve the precision of the pooled estimates and minimize the risk of false positives, the HKSJ method was applied.

Statistical heterogeneity was assessed using the Cochran Q statistic and the I^2 index, with I^2>50% indicating substantial heterogeneity [21]. To further quantify the dispersion of effects in real-world settings, 95% PIs were calculated alongside the pooled estimates. Subgroup analyses were prespecified based on the type of control group (active vs waitlist), intervention duration (≤12 weeks vs >12 weeks), delivery type (app vs website), and target lifestyle (diet vs physical activity vs combined). All statistical analyses were conducted using Comprehensive Meta-Analysis (version 3) software (Biostat) and R (version 4.5.1; R Foundation for Statistical Computing) with the "*meta*" package.

Reporting Bias Assessment

To assess potential reporting biases, including publication bias and small-study effects, we used both visual and statistical methods. Funnel plot asymmetry was visually inspected. Additionally, the Egger regression test was performed to statistically evaluate small-study effects, with a P<.10 indicating significant asymmetry [22]. Furthermore, a trim-and-fill sensitivity analysis was conducted to estimate the potential

XSL•FO
RenderX

impact of missing studies on the pooled effect size [23]. We also considered whether the observed asymmetry could be attributed to factors other than publication bias, such as differences in study quality or substantial interstudy heterogeneity.

Certainty Assessment

The Grading of Recommendations, Assessment, Development, and Evaluation framework was used to evaluate the certainty of evidence across 5 domains: risk of bias, imprecision, inconsistency, indirectness, and publication bias. Depending on the quality assessment within these domains, the certainty of evidence could be either downgraded or upgraded. The final classification for each outcome was assigned to 1 of 4 levels: "very low," "low," "moderate," or "high" [24]. Two independent reviewers (SAL and JHP) assessed the evidence across 5 domains. Any discrepancies in the grading were resolved through consensus between the 2 reviewers.

Ethical Considerations

This study is a systematic review and meta-analysis of previously published literature and did not involve direct interaction with human participants or the collection of identifiable private information. According to institutional and national guidelines (eg, the Bioethics and Safety Act of the Republic of Korea), this type of research is exempt from Institutional Review Board review as it uses publicly available data and carries no risk to individuals.

Results

Study Selection

A total of 34,304 literature records were identified from the databases, and 12 additional records were identified through other sources. After removing 17,327 duplicate records, 16,977 records were screened based on titles and abstracts, of which 16,325 were excluded. The primary reasons for exclusion at this stage were (1) clearly irrelevant topics (eg, nondigital clinical treatments or pharmacological interventions), (2) ineligible study types (eg, study protocols, review papers, or editorials), and (3) ineligible populations (eg, pediatric or adolescent samples) identifiable directly from titles or abstracts. Subsequently, 652 full-text papers were assessed for eligibility. Of these, 633 papers were excluded for the following reasons: ineligible combined interventions (n=335), ineligible control groups including digital interventions (n=229), lack of weight-related outcome measures (n=68), and ineligible participants (n=1). Finally, 19 papers were included in this systematic review and meta-analysis (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of included studies. The diagram displays the number of records identified, screened, assessed for eligibility, and included in the systematic review and meta-analysis.

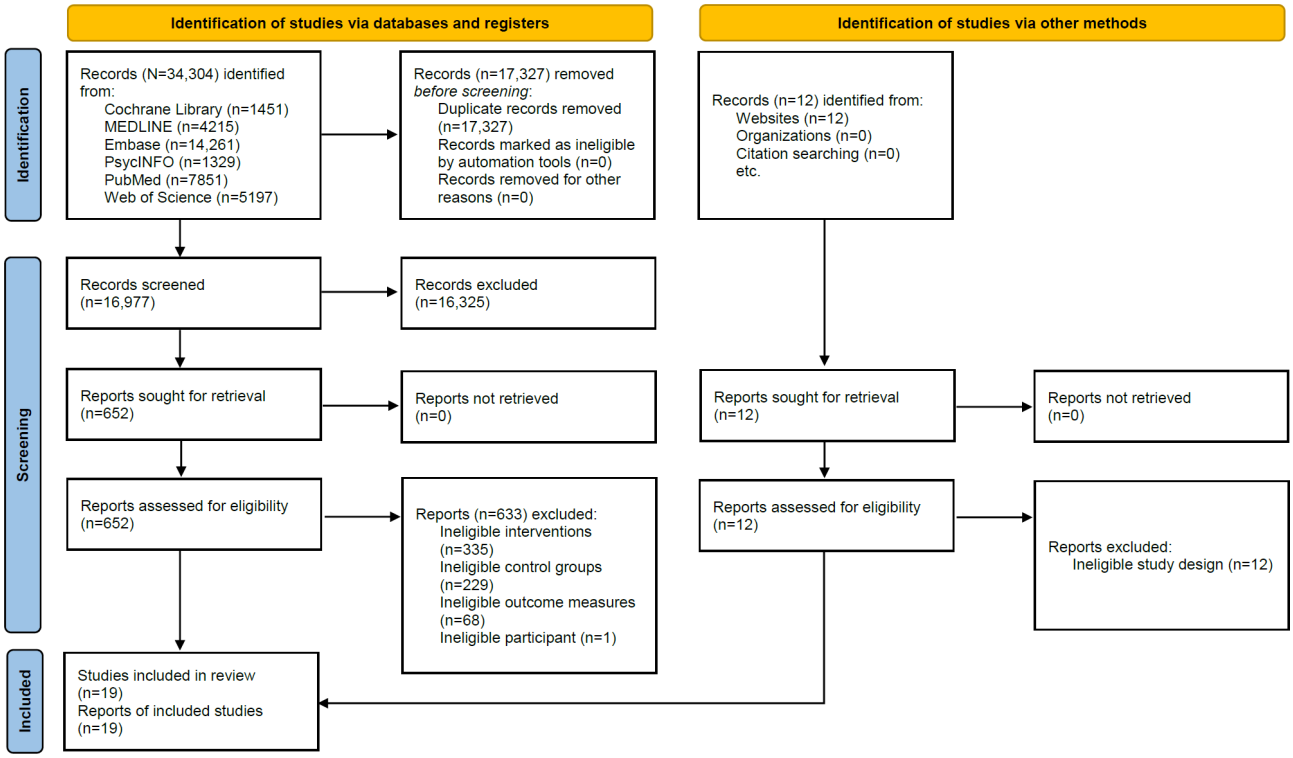

Study Characteristics

Table 1 provides characteristics of the 19 included studies. All studies were published between 2007 and 2024. A total of 3556 participants (71% female) were included in the 19 trials, and the total sample sizes per study ranged from 31 to 650 participants. The mean participant age ranged from 29.4 to 51.2 years. All participants were classified as having overweight or obesity (BMI≥23 kg/m^2), with varying inclusion thresholds across studies.

XSL•FO
RenderX

Table . Characteristics of the included studies in the meta-analysis.

Authors	Sample size	Mean age (SD) years	Female participants, n (%)	Population (BMI)	Target lifestyle	Behavioral strategy	Delivery format	Comparison	Duration
Allen et al [25]	E[a]=17 and C[b]=18	43.9 (12.6)	29 (82.8)	28 - 42	Diet and physical activity	Self-monitoring with calorie targets and behavioral counseling for diet and physical activity.	App	Active control: in-person nutritional counseling	12 weeks
Brame et al [26]	E=78 and C=75	50.0 (10.8)	70 (45.7)	27.5 - 34.9	Diet	Self-monitoring and personalized goal setting for diet and weight with structured coaching and feedback.	Website	Nonspecific active control (usual care)	12 weeks
Carter et al [27]	E1=43, E2=43, and C=43	41.8 (9.1)	99 (76.7)	≥27	Diet	Self-monitoring of diet and activity with calorie goals, feedback, and tailored text support.	E1: App E2: Website	Active control (food diary)	6 months
Chung et al [28]	E=19, C1=16, and C2=19	Not reported	34 (62.9)	≥25	Diet	Self-monitoring of dietary intake with personalized feedback based on nutrient analysis.	Website	C1: active control (food diary) and C2: waitlist control	12 weeks
Collins et al [29]	E=99 and C=104	41.8 (10.1)	118 (58.1)	25 - 40	Diet and physical activity	Self-monitoring with tailored calorie targets, behavioral feedback for diet, and physical activity.	Website	Waitlist control	12 weeks
Dunn et al [30]	E=42 and C=38	47.9 (10.3)	Not reported	≥25	Diet and physical activity	Structured goal tracking and coaching for diet and physical activity.	Website	Waitlist control	15 weeks
Hurkmans et al [31]	E=30 and C=22	44.5 (11.3)	30 (57.6)	29 - 34	Diet and physical activity	Self-monitoring and education for diet and physical activity.	App	Waitlist control	12 weeks

XSL•FO
RenderX

Authors	Sample size	Mean age (SD) years	Female participants, n (%)	Population (BMI)	Target lifestyle	Behavioral strategy	Delivery format	Comparison	Duration
Kohl et al [32]	E=78 and C=75	48.9 (11.1)	109 (71.2)	27.5 - 34.9	Diet	Self-monitoring with personalized diet logging, energy-density feedback, and interactive behavior-change activities.	Website	Nonspecific active control (education)	12 weeks
Kraschnewski et al [33]	E=50 and C=50	50.3 (10.9)	69 (69.0)	≥25	Diet and physical activity	Self-monitoring with goal setting and tailored behavior modeling for diet, activity, and weight control.	Website	Waitlist control	12 weeks
Krukowski et al [34]	E=161 and C=157	46.6 (10.0)	295 (92.7)	25 - 50	Diet and physical activity	Self-monitoring with behavioral counseling for diet, physical activity, and weight control.	Website	Active control (in-person lifestyle intervention)	6 months
Lim et al [35]	E=99 and C=105	51.2 (9.7)	72 (35.2)	≥23	Diet and physical activity	Self-monitoring with feedback, goal setting, and dietitian messaging for diet, activity, and weight loss.	App	Waitlist control	6 months
Lugones-Sanchez et al [36]	E=318 and C=332	48.3 (9.6)	445 (68.4)	27.5 - 40	Diet and physical activity	Self-monitoring with calorie feedback, activity tracking, and behavioral goal setting for diet and physical activity.	App	Active control (in-person counseling)	12 weeks
McConnon et al [37]	E=111 and C=110	47.7 (not reported)	170 (77.0)	≥30	Diet and physical activity	Self-monitoring with personalized advice, motivational feedback, and behavioral tools for diet and physical activity.	Website	Nonspecific active control (usual care)	12 months

XSL•FO
RenderX

Authors	Sample size	Mean age (SD) years	Female participants, n (%)	Population (BMI)	Target lifestyle	Behavioral strategy	Delivery format	Comparison	Duration
Moravcova et al [38]	E=50 and C=50	43.3 (9.5)	71 (71.0)	≥30	Diet and physical activity	Self-monitoring with personalized daily goals, education, and dietitian support for diet, activity, and lifestyle behaviors.	App	Active control (in-person lifestyle intervention)	12 weeks
Padwal et al [39]	E=225, C1=215, and C2=211	40.4 (9.8)	540 (82.9)	≥35	Diet and physical activity	Self-guided behavior change program with modular content for diet and physical activity.	Website	C1: active control (in-person) and C2: nonspecific active control (education)	12 weeks
Steinberg et al [40]	E=47 and C=44	43.8 (11.0)	68 (74.7)	25 - 40	Diet and physical activity	Self-monitoring with smart scale, weight tracking, and weekly tailored feedback with behavioral lessons.	Website	Waitlist control	6 months
Svetkey et al [41]	E=120 and C=123	29.4 (4.3)	169 (69.5)	≥25	Diet and physical activity	Self-monitoring with goal setting, behavioral prompts, and peer support for diet, activity, and weight management.	App	Waitlist control	24 months
Vaz et al [42]	E=15 and C=16	43.2 (3.5)	Not reported	25 - 42	Diet and physical activity	Self-monitoring with wearable tracking, food photo logging, and coaching and peer feedback for diet, activity, and weight loss.	App	Waitlist control	6 months
Yardley et al [43]	E=45 and C=43	50.5 (13.5)	60 (68.1)	≥30	Diet and physical activity	Self-monitoring with goal setting, dietary choice, and cognitive-behavioral tools for weight management.	Website	Nonspecific active control (usual care)	6 months

[a]E: experimental group.

[b]C: control group.

XSL•FO
RenderX

The interventions targeted either diet alone or a combination of diet and physical activity [25-43]; notably, no study targeted physical activity alone. Common intervention strategies included self-monitoring, personalized goal setting, and tailored feedback. Delivery platforms included mobile apps (n=7) [25,31,35,36,38,41,42], websites (n=11) [26,28-30,32-34,37,39,40,43], or both (n=1) [27]. Comparison groups included active controls (eg, in-person nutritional counseling and food diaries; n=7) [25,27,28,34,36,38,39], nonspecific active controls (eg, education or usual care; n=5) [26,32,37,39,43], and waitlist controls (n=9) [28-31,33,35,40-42]. The duration of interventions varied from 12 weeks to 24 months, with 12-week programs being the most common (n=9; Table 1) [25,26,28,29,31-33,36,38,39]. To ensure the stand-alone nature of the included interventions, we meticulously verified the delivery mechanisms of all digital interactions. While some studies used terms such as coaching, feedback, or peer support, these were confirmed to be delivered through automated, system-generated algorithms or asynchronous, preprogrammed modules without real-time human clinical labor (Multimedia Appendix 3).

Body weight and BMI were the most frequently reported as primary outcomes. Eight studies also assessed dietary outcomes such as caloric intake and macronutrient composition [25,30-33,35,36,40]. Retention rates, as reported in Table 2, ranged from 44.1% to 95%, with most studies achieving rates above 70%. Mobile app interventions tended to show slightly higher retention compared to web-only platforms, although variation existed (Table 2).

XSL•FO
RenderX

Table . Adherence and measures characteristics. Retention rate (%) = (number of participants in the intervention group who completed the final follow-up / number of participants initially randomized to the intervention group) ×100.

Author, year	Retention rate, n/N (%)	Anthropometric measures	Dietary measures	Timing of measures
Allen et al [25]	10/17 (58.8%)	Body weight (kg), BMI (kg/m^2), and waist circumference (cm)	Dietary, calories, fruit and vegetable, and sodium intake (Kcal/d)	Baseline and 12 weeks
Brame et al [26]	39/78 (50%)	Body weight (kg), BMI (kg/m^2), and waist circumference (cm)	Not reported	Baseline and 12 weeks
Carter et al [27]	App: 40/43 (93%) Website: 19/43 (44.1%)	Body weight (kg), BMI (kg/m^2), and body fat (%)	Not reported	Baseline, 6 weeks, and 6 months
Chung et al [28]	19/20 (95%)	Body weight (kg), BMI (kg/m^2), body fat (%), and waist-to-hip ratio (%)	Not reported	Baseline, 6 weeks, and 12 weeks
Collins et al [29]	74/99 (74.7%)	Body weight (kg), BMI (kg/m^2), and waist circumference (cm)	Not reported	Baseline and 12 weeks
Dunn et al [30]	(28/44) (66.6%)	Body weight (kg) and BMI (kg/m^2)	Eating confidence (5-point Likert scale)	Baseline and 15 weeks
Hurkmans et al [31]	24/30 (80%)	BMI (kg/m^2)	Energy intake (Kcal/d)	Baseline and 12 weeks
Kohl et al [32]	55/78 (70.5%)	Body weight (kg), fat mass (kg), fat-free mass (kg), and waist circumference (cm)	Energy and protein intake (Kcal/d), carbohydrate, fat, alcohol, and fiber intake (g/d)	Baseline, 12 weeks, 6 months, and 12 months
Kraschnewski et al [33]	43/50 (86%)	Body weight (kg) and BMI (kg/m^2)	Caloric intake (Kcal/d)	Baseline and 12 weeks
Krukowski et al [34]	153/161 (95%)	BMI (kg/m^2)	Not reported	Baseline and 6 months
Lim et al [35]	94/99 (94.9%)	Body weight (kg) and BMI (kg/m^2)	Caloric intake (Kcal/d)	Baseline, 3 months, and 6 months
Lugones-Sanchez et al [36]	218/318 (67.9%)	Body weight (kg), BMI (kg/m^2), waist circumference (cm), and hip circumference (cm)	Energy intake (Kcal/d)	Baseline, 3 months, and 12 months
McConnon et al [37]	54/111 (48.6%)	Body weight (kg)	Not reported	Baseline, 6 months, and 12 months
Moravcova et al [38]	32/50 (64%)	Body weight (kg), BMI (kg/m^2), waist circumference (cm), body fat (%), and muscle mass (kg)	Not reported	Baseline, 3 months, and 6 months
Padwal et al [39]	166/225 (73.7%)	Body weight (kg) and BMI (kg/m^2)	Not reported	Baseline and 12 weeks
Steinberg et al [40]	45/47 (95.7%)	Body weight (kg)	Caloric intake (Kcal/d)	Baseline, 3 months, and 6 months
Svetkey et al [41]	104/120 (86.6%)	Body weight (kg)	Not reported	Baseline, 6 months, 12 months, and 24 months
Vaz et al [42]	13/15 (86.6%)	Body weight (kg)	Not reported	Baseline and 6 months
Yardley et al [43]	39/45 (86.6%)	Body weight (kg)	Not reported	Baseline, 6 months, and 12 months

Risk of Bias in Studies

Among the 19 studies, 3 studies were classified to have a high risk of bias, primarily due to issues in missing outcome data and selection of the reported results. The remaining 16 studies were judged as either low risk (n=7) or raising some concerns (n=9), with the most common concerns arising from deviations

XSL•FO
RenderX

from intended interventions or lack of information on prespecified analysis plans (Figure 2 [25-43]).

Figure 2. Risk of bias summary for included studies. This figure displays the overall risk of bias assessments across all included studies using the Cochrane Risk of Bias tool (version 2) [25-43].

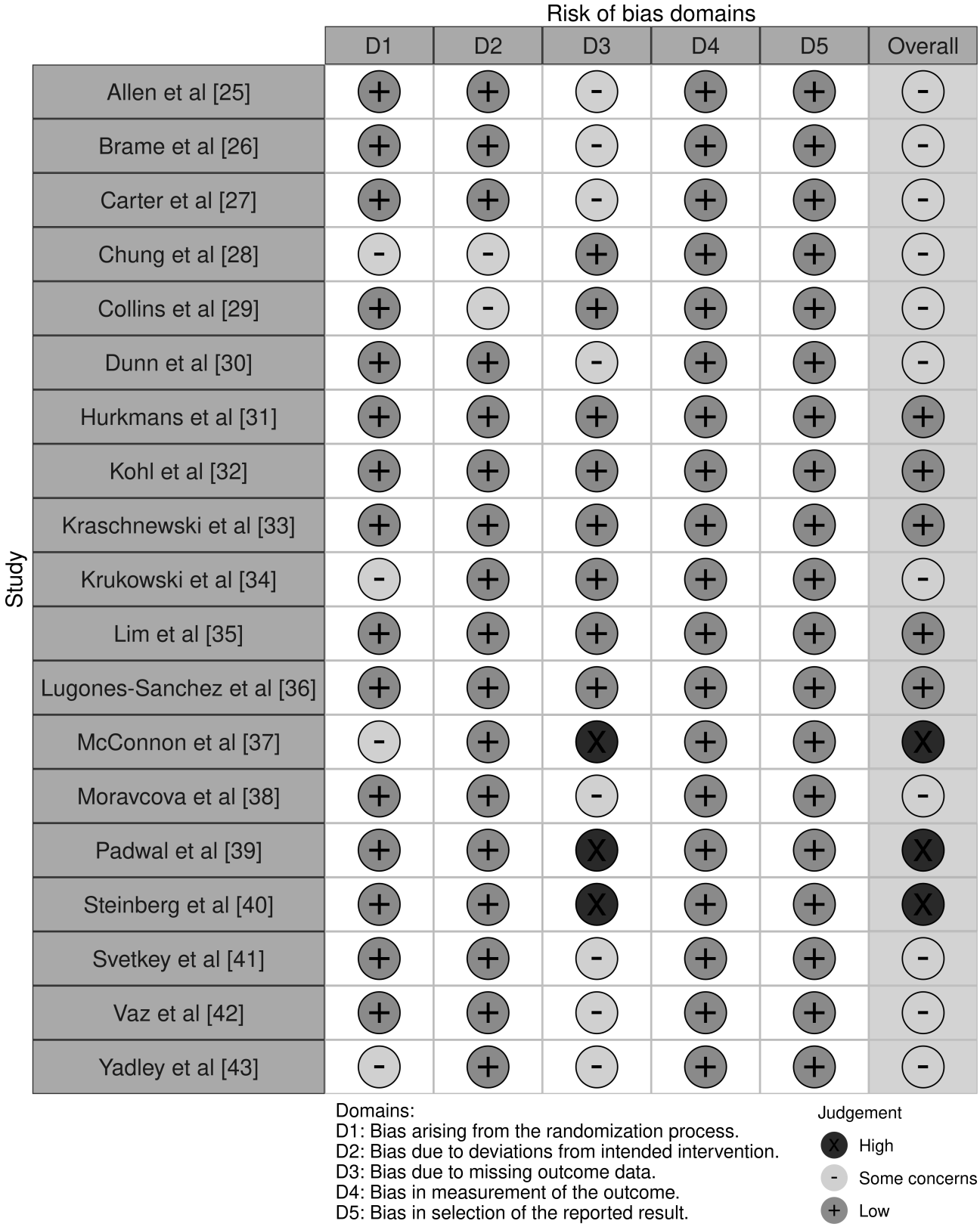

Results of Individual Studies

For all 19 included studies (N=3556), the individual effect sizes for anthropometric outcomes ranged from SMD −0.04 to 0.72. Similarly, for the 8 studies reporting dietary outcomes (n=1365), individual SMDs showed a range of −0.10 to 0.62. The summary data for each study, including means, SDs, and sample sizes for both intervention and control groups at baseline and postintervention, are detailed in Table 2. The forest plots (Figure

XSL•FO
RenderX

3 [25-43] and Figure 4 [25,30-32,35,36,40]) provide a visual representation of the individual study effects alongside their 95% CIs.

Figure 3. Effect of stand-alone digital lifestyle interventions (DLSIs) on anthropometric outcomes. Forest plot displaying the effect sizes of digital lifestyle interventions on anthropometric outcomes. Gray squares represent the standardized mean difference (SMD) for individual studies, and horizontal gray lines indicate the 95% CIs. The black diamond at the bottom represents the pooled effect size, calculated using the Hartung-Knapp-Sidik-Jonkman (HKSJ) method. A positive SMD favors stand-alone DLSIs [25-43].

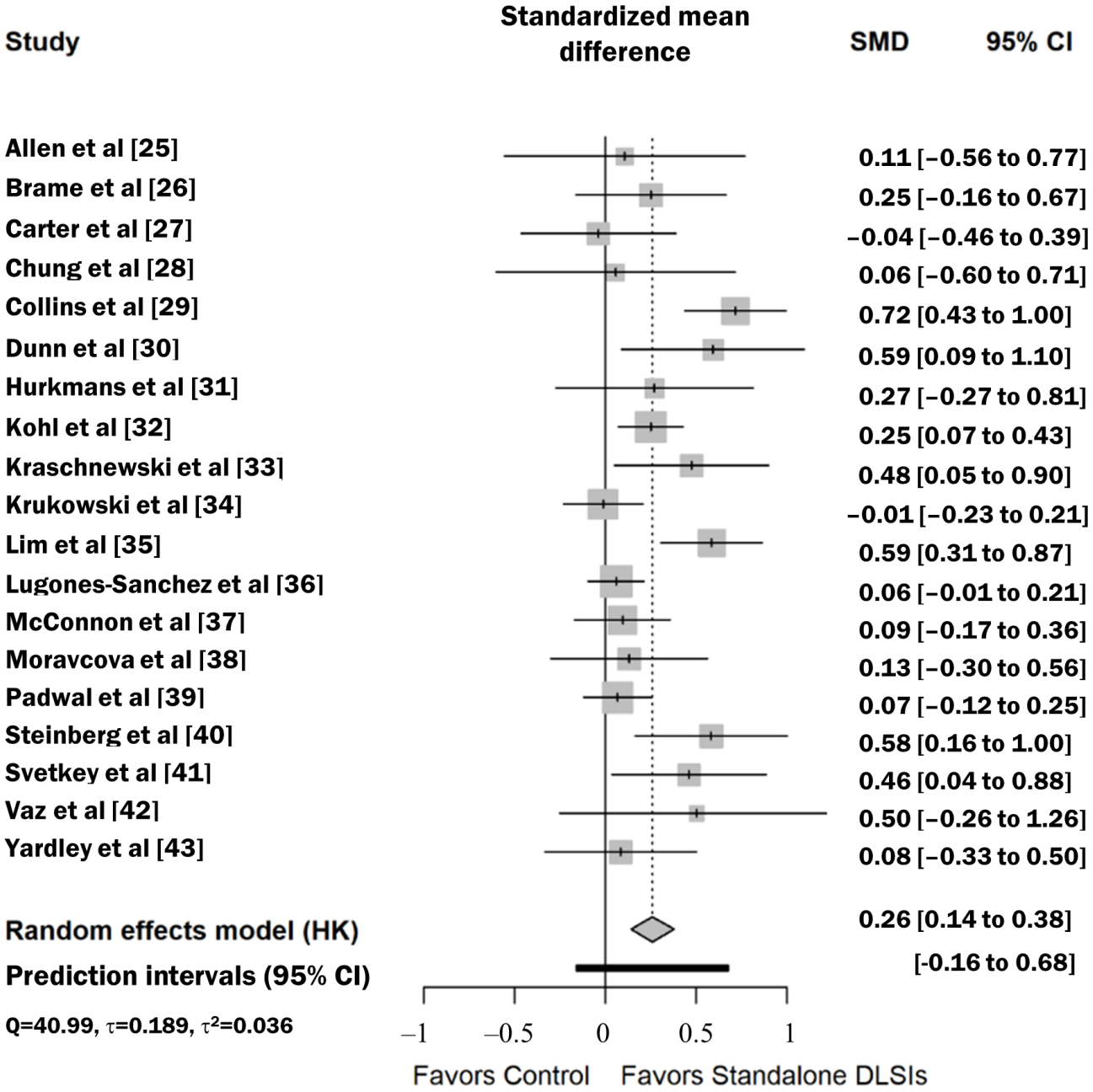

XSL•FO
RenderX

Figure 4. Effect of stand-alone digital lifestyle interventions (DLSIs) on dietary outcomes. Forest plot displaying the effect sizes of digital lifestyle interventions on dietary outcomes. Gray squares represent the standardized mean difference (SMD) for individual studies, and horizontal gray lines indicate the 95% CIs. The black diamond at the bottom represents the pooled effect size, calculated using the Hartung-Knapp-Sidik-Jonkman (HKSJ) method. A positive SMD favors stand-alone DLSIs [25,30-32,35,36,40].

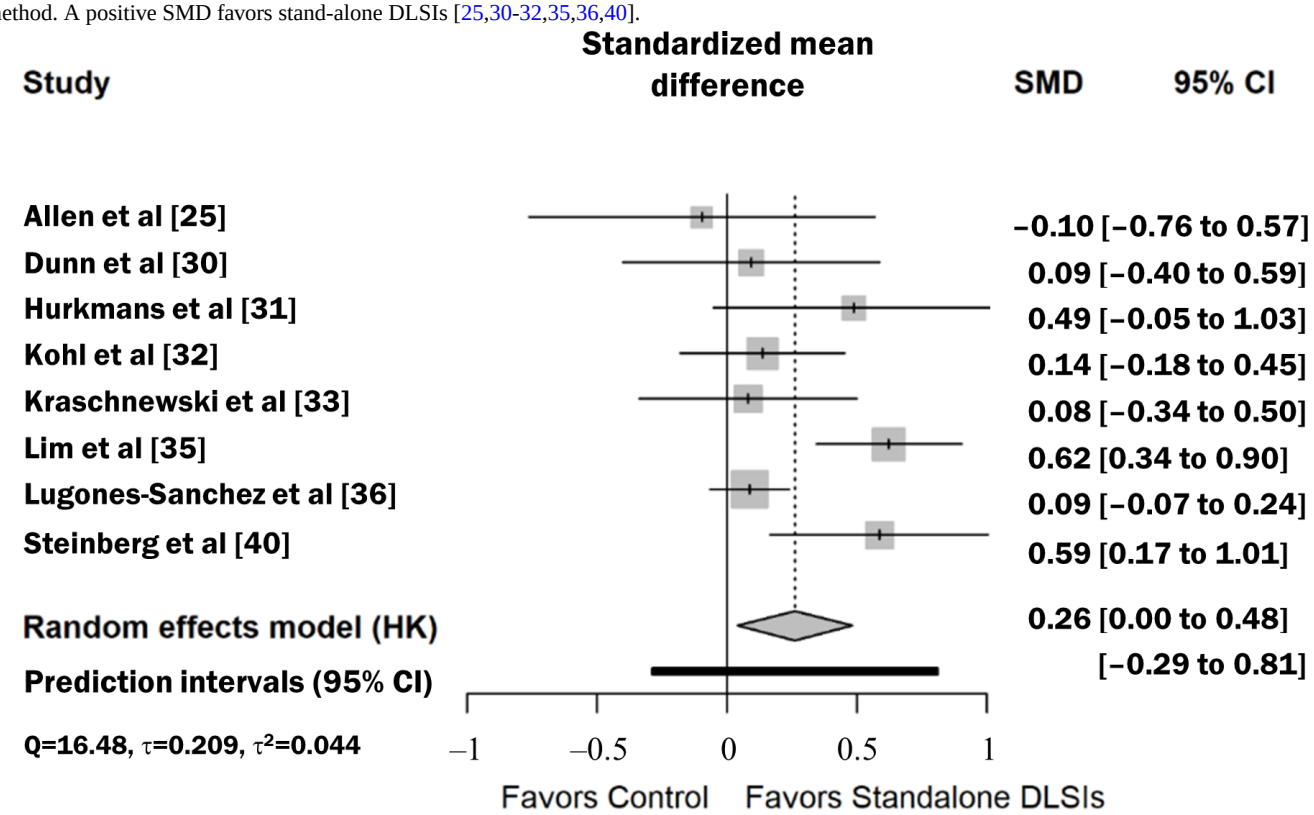

Results of Syntheses

The meta-analysis using a random-effects model with the HKSJ adjustment demonstrated that stand-alone DLSIs had a statistically significant effect on improving anthropometric outcomes compared to controls (SMD 0.26, 95% CI 0.14 - 0.38; P<.001; Table 3). To aid clinical interpretation, the pooled SMD was translated into kilograms based on a pooled baseline SD of 16.7 kg, derived strictly from the included studies that reported body weight in kilograms. This corresponds to an additional weight loss of approximately 4.34 kg. Considering the variability in baseline SDs across the included studies (ranging from 10.1 kg to 25.2 kg), the estimated weight loss could plausibly range from 2.62 kg to 6.55 kg, depending on the population's baseline variance. While the I^2 statistic (56.1%; P<.001) indicated substantial heterogeneity (Q=40.99; P<.001; τ=0.189; τ^2=0.036), the 95% PI was calculated as −0.16 to 0.68. This indicates that while the average effect is significant, the outcome in specific future settings is expected to vary, and there is a possibility that the intervention may show no benefit or even be less effective than the control in certain contexts. For dietary outcomes, stand-alone DLSIs also showed a significant effect (SMD 0.26, 95% CI 0.04 - 0.48; P=.008; Table 3). Substantial heterogeneity (I^2=57.5%; P=.01) was observed (Q=16.48; P=.01; τ=0.209; τ^2=0.044), and the 95% PI ranged from −0.29 to 0.81, indicating that while the average effect is positive, the impact of these interventions in a new specific setting could vary and may not always achieve a positive result.

XSL•FO
RenderX

Table . Effects of stand-alone digital lifestyle interventions on anthropometric and dietary outcomes. In cases where a study reported outcomes for multiple intervention arms compared to a single control group, we combined the intervention groups or split the control group to maintain the independence of observations.

Outcomes	Anthropometric outcomes							Dietary outcomes						
Main analysis	*k*	SMD[a] (95% CI)	95% PI[b]	I^2	Q	τ	τ^2	*k*	SMD (95% CI)	95% CI	I^2	Q	τ	τ^2
Overall effects	19	0.26 (0.14 to 0.38)[c]	−0.16 to 0.68	56.1[c]	40.99	0.189	0.036	8	0.26 (0.04 to 0.48)[d]	−0.29 to 0.81	57.5[e]	16.48[e]	0.209	0.044
Subgroup analysis														
Target lifestyle							$P_{subgroup}$=.27							—[f]
Diet and physical activity	15	0.27 (0.12 to 0.42)[c]	−0.04 to 0.58	67.5[c]	36.92[c]	0.000	0.000	7	0.30 (0.08 to 0.53)[*]	−0.16 to 0.76	62.6[d]	18.72[d]	0.228	0.052
Diet	4	0.14 (−0.05 to 0.32)	−0.07 to 0.35	0.0	4.31	0.205	0.042	Not applicable as only one study remained	Not applicable as only one study remained	Not applicable as only one study remained	Not applicable as only one study remained	Not applicable as only one study remained	Not applicable as only one study remained	Not applicable as only one study remained
Duration							$P_{subgroup}$=.71							$P_{subgroup}$=.44
≤12 weeks	10	0.23 (0.07 to 0.40)[d]	−0.17 to 0.65	56.1[d]	20.48[d]	0.165	0.027	5	0.12 (−0.06 to 0.30)	−0.30 to 0.54	0	2.38	0.119	0.014
>12 weeks	9	0.28 (0.06 to 0.50)[e]	−0.27 to 0.84	60.4[d]	20.21[d]	0.221	0.9	3	0.47 (−0.21 to 1.16)	−0.70 to 1.65	42.6	3.48	0.215	0.046
Control group							$P_{subgroup}$<.001							—
Active controls	8	0.04 (−0.08 to 0.16)	−0.08 to 0.16	0.00	4.61	0.000	0.000	3	0.23 (−0.23 to 0.70)	−0.46 to 0.92	64.4	5.62	0.324	0.105
Nonspecific active controls	5	0.13 (0.00 to 0.26)[e]	0.00 to 0.26	0.00	2.15	0.000	0.000	Not applicable as only one study remained	Not applicable as only one study remained	Not applicable as only one study remained	Not applicable as only one study remained	Not applicable as only one study remained	Not applicable as only one study remained	Not applicable as only one study remained
Waitlist controls	9	0.57 (0.42 to 0.73)[c]	0.42 to 0.72	0.00	7.91	0.000	0.000	5	0.36 (0.11 to 0.62)[d]	0.11 to 0.61	44.2	7.17	0.145	0.021
Delivery type							$P_{subgroup}$=.90							$P_{subgroup}$=.61

XSL•FO
RenderX

Outcomes	Anthropometric outcomes							Dietary outcomes						
Main analysis	k	SMD[a] (95% CI)	95% PI[b]	I^2	Q	τ	τ^2	k	SMD (95% CI)	95% CI	I^2	Q	τ	τ^2
App	8	0.24 (−0.00 to 0.49)	−0.20 to 0.38	62.3[d]	15.92[e]	0.219	0.048	5	0.33 (0.02 to 0.64)[e]	−0.21 to 0.87	73.6[d]	15.15[d]	0.266	0.071
Website	15	0.23 (0.08 to 0.37)[d]	−0.03 to 0.49	60.2[d]	27.64[d]	0.167	0.028	4	0.22 (−0.04 to 0.48)	−0.04 to 0.48	22.7	3.88	0.089	0.008
Sensitivity analysis														
Aggregated (study-level)	19	0.27 (0.14 to 0.41)[c]	−0.15 to 0.70	56.7[d]	34.65[d]	0.188	0.035	8	0.22 (0.02 to 0.42)[e]	−0.32 to 0.75	54.7[e]	13.25	0.192	0.037
Excluding high risk of bias	16	0.27 (0.14 to 0.41)[c]	−0.21 to 0.74	88.1[c]	126.12[c]	0.213	0.045	7	0.20 (0.00 to 0.39)[e]	−0.37 to 0.77	64.2[e]	16.74	0.211	0.044
≤6-month duration	15	0.26 (0.13 to 0.39)[c]	−0.22 to 0.74	88.1[c]	117.41	0.214	0.046	7	0.26 (0.03 to 0.48)[e]	−0.34 to 0.86	64.3[e]	16.81	0.219	0.048

[a]SMD: standardized mean difference.

[b]PI: prediction interval.

[c]P<.001.

[d]P<.01.

[e]P<.05.

[f]Not available.

Subgroup analyses were conducted to explore potential moderators of intervention effects. First, analysis by target behavior revealed that interventions focusing on both diet and physical activity demonstrated a small significant effect on anthropometric outcomes (SMD 0.27, 95% CI 0.12 - 0.42; 95% PI −0.04 to 0.58; P<.001; I^2=67.5%), while diet-only interventions did not reach statistical significance (SMD 0.14, 95% CI −0.05 to 0.32; 95% PI −0.07 to 0.35; P=.12; I^2=0%). For dietary outcomes, combined-target interventions yielded a significant effect (SMD 0.30, 95% CI 0.08 - 0.53; 95% PI −0.16 to 0.76; P=.008; I^2=62.6%; Table 3).

Second, intervention duration was examined. Stand-alone DLSIs ≤12 weeks showed a small, significant effect on anthropometric outcomes (SMD 0.23, 95% CI 0.07 - 0.40; 95% PI −0.17 to 0.65; P=.02; I^2=56.1%), comparable to those >12 weeks (SMD 0.28, 95% CI 0.06 - 0.50; 95% PI −0.27 to 0.84; P=.01; I^2=60.4%). The test for subgroup differences confirmed no significant disparity between short- and long-term durations (P=.71). For dietary outcomes, longer interventions did not show a significant effect (SMD 0.47, 95% CI −0.21 to 1.16; 95% PI −0.70 to 1.65; P>.05; I^2=42.6%), although no significant heterogeneity was observed. Similarly, shorter ones did not reach significance (SMD 0.12, 95% CI −0.06 - 0.30; 95% PI −0.30 to 0.54; P>.05; I^2=0%), despite having no heterogeneity (Table 3). The formal test for subgroup differences indicated that the difference between these 2 duration categories was not statistically significant (P=.44). This suggests that intervention duration does not significantly moderate the effects on dietary outcomes, and stand-alone DLSIs did not yield a consistent significant impact on diet regardless of the program length.

Third, by control type, stand-alone DLSIs showed no significant effects vs active controls on anthropometric outcomes (SMD 0.04, 95% CI −0.08 to 0.16; 95% PI −0.08 to 0.16; P=.52; I^2=0.0%) and diet (SMD 0.23, 95% CI −0.23 - 0.70; 95% PI −0.46 to 0.92; P=.33; I^2=64.4%). Compared to nonspecific active controls, the effect on anthropometric outcomes was small but significant (SMD 0.13, 95% CI 0.00 - 0.26; 95% PI 0.00 to 0.26; P=.05; I^2=0%). The strongest effects were observed vs waitlist controls, with a moderate effect on anthropometric outcomes (SMD 0.57, 95% CI 0.42 - 0.73; 95% PI 0.42 to 0.72; P<.001; I^2=0%) and a significant effect on diet (SMD 0.36, 95% CI 0.11 - 0.62; 95% PI 0.11 to 0.61; P=.006; I^2=44.2%). A significant subgroup difference was observed for

XSL•FO
RenderX

anthropometric outcomes (P<.001), indicating control type as a significant moderator (Table 3).

Fourth, by delivery type, interventions delivered via mobile apps showed no significant effect on anthropometric outcomes (SMD 0.24, 95% CI 0.00 - 0.49; 95% PI −0.20 to 0.68; P=.05; I^2=62.3%), whereas a significant effect was observed for dietary outcomes (SMD 0.33, 95% CI 0.02 - 0.64; 95% PI −0.21 to 0.87; P=.03; I^2=73.6%). Website-based interventions also demonstrated a significant effect on anthropometric outcomes (SMD 0.23, 95% CI 0.08 - 0.37; 95% PI −0.03 to 0.49; P=.004; I^2=60.2%), but did not reach statistical significance for dietary outcomes (SMD 0.22, 95% CI −0.04 - 0.48; 95% PI −0.04 to 0.48; P=.10; I^2=22.7%; Table 3). No significant subgroup differences were found between app and website delivery for either anthropometric (P=.90) or dietary outcomes (P=.61; Table 3).

Sensitivity analyses were conducted by aggregating multiple effect sizes within the same study using their arithmetic average to evaluate the robustness of the primary findings against potential unit-of-analysis errors. The significant effects of stand-alone DLSIs on both anthropometric and dietary outcomes remained robust. For anthropometric outcomes, the pooled effect size slightly increased (SMD 0.27, 95% CI 0.14 - 0.41; P<.001; 95% PI −0.15 to 0.70; I^2=56.7%), while for dietary outcomes, the effect size decreased modestly (SMD 0.22, 95% CI 0.02 - 0.42; P=.03; 95% PI −0.32 to 0.75; I^2=54.7%; Table 3). In addition, a secondary sensitivity analysis was performed by excluding 3 studies identified as having a high risk of bias while maintaining the study-level aggregation of effect sizes. For anthropometric outcomes, the pooled effect size remained statistically significant (SMD 0.27, 95% CI 0.14 - 0.39; P<.001; 95% PI −0.21 to 0.74; I^2=88.1%). Similarly, for dietary outcomes, the effect remained significant even after excluding lower-quality evidence (SMD 0.20, 95% CI 0.01 - 0.39; P=.04; 95% PI −0.37 to 0.77; I^2=64.2%).

Furthermore, to ensure the findings were not skewed by the inclusion of long-term trials, a third sensitivity analysis was restricted to trials with a duration of ≤6 months. For anthropometric outcomes, the effect remained highly consistent (SMD 0.26, 95% CI 0.13 - 0.39; P<.001; 95% PI −0.22 to 0.74; I^2=88.1%). Similarly, for dietary outcomes, the significant benefit was maintained (SMD 0.26, 95% CI 0.04 - 0.48; P=.03; 95% PI −0.34 to 0.86; I^2=64.3%; Table 3). These cumulative results indicate that the overall conclusions were not sensitive to the presence of multiple outcomes per study, the inclusion of studies with a high risk of bias, or variations in trial duration, further confirming the stability and reliability of the main analysis.

Reporting Biases

Small-study effects were assessed for significant results in the main analysis using funnel plots and the Egger test. For anthropometric outcomes, the Egger test yielded an intercept of 1.35 (95% CI –1.54 to 4.24; t_{17}=0.946; P=.35), and for dietary outcomes, the intercept was 0.28 (95% CI –1.19 to 1.75; t_6=0.108; P=.92), indicating no statistical evidence of asymmetry in either case (Figures 5 and 6). However, specifically for anthropometric outcomes, visual inspection suggested potential asymmetry in the lower-left quadrant, possibly due to the heavy reliance on waitlist-controlled trials and the risk of unpublished small-scale trials. To address this, a trim-and-fill sensitivity analysis was performed (Figure 5), which identified 2 potentially missing studies on the left side of the funnel. After imputing these studies, the adjusted pooled effect size remained statistically significant (adjusted SMD 0.21, 95% CI: 0.09 - 0.34; P<.001). The heterogeneity for the adjusted model was observed (Q_{20}=120.54; P<.001; τ=0.294; τ^2=0.086), confirming the robustness of the primary findings. Consequently, the observed effect sizes for both outcomes are unlikely to be driven primarily by publication bias.

XSL•FO
RenderX

Figure 5. Trim-and-fill funnel plot for anthropometric outcomes. Open circles indicate 2 imputed studies identified via trim-and-fill analysis to address visual asymmetry. The model uses the Hartung-Knapp-Sidik-Jonkman (HKSJ) adjustment. The pooled effect remains statistically significant after imputation, confirming the robustness of the anthropometric results against potential publication bias.

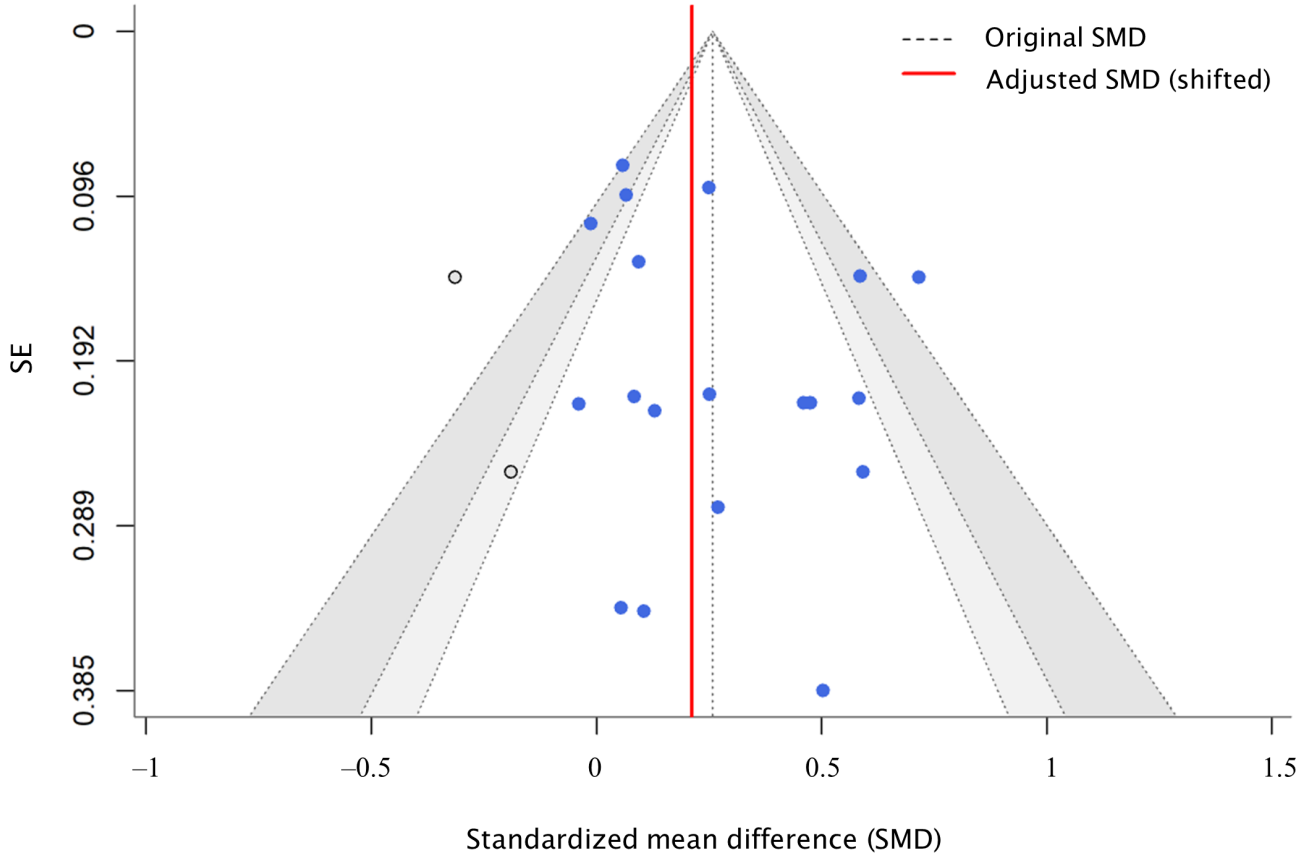

XSL•FO
RenderX

Figure 6. Funnel plot for dietary outcomes. The funnel plot illustrates the relationship between standardized mean differences (SMDs) and their corresponding SEs for dietary outcomes. The plot was generated using the Hartung-Knapp-Sidik-Jonkman (HKSJ) adjustment to account for study heterogeneity.

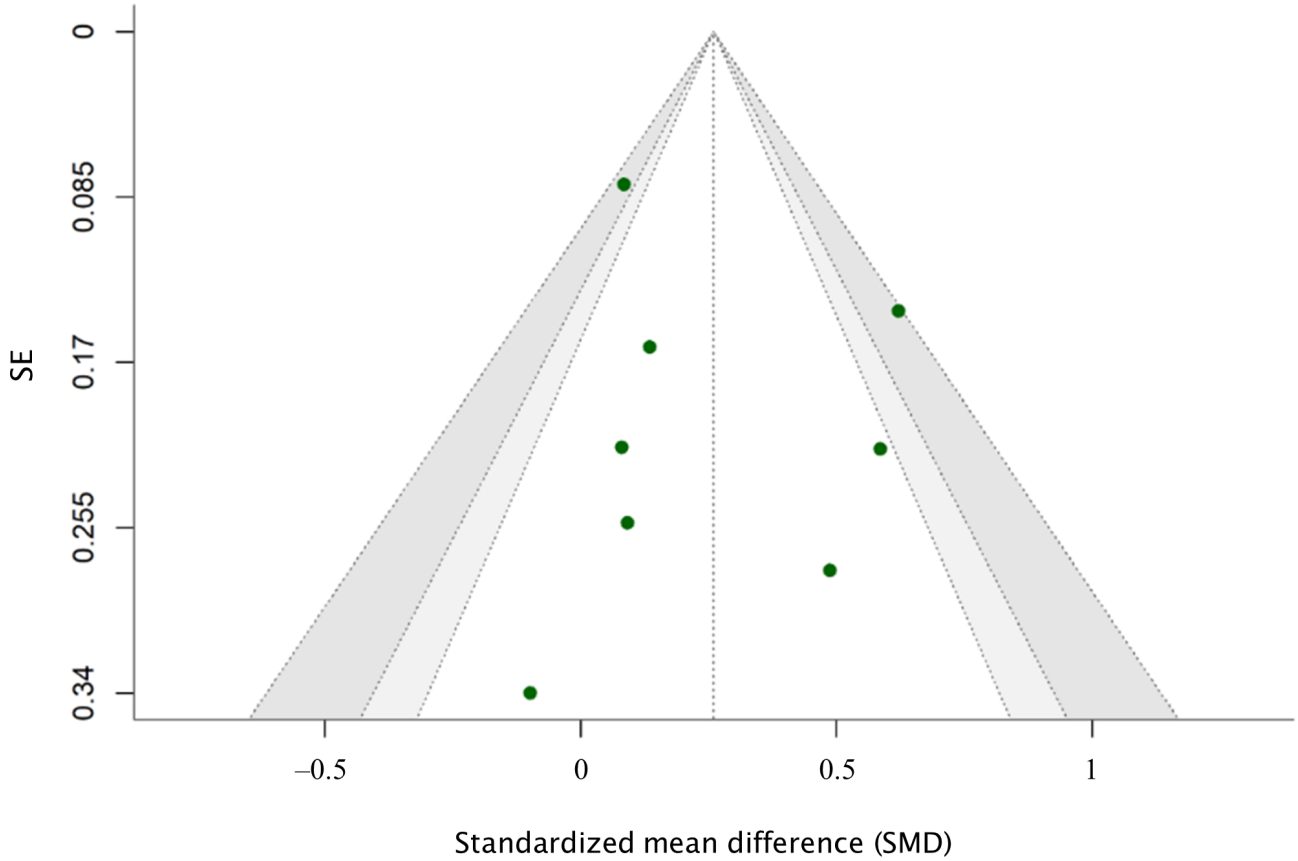

Certainty of Evidence

Based on the Grading of Recommendations, Assessment, Development, and Evaluation assessment of each significant effect estimate, this study concludes that stand-alone DLSIs provide a moderate-certainty benefit for both anthropometric and dietary outcomes. The certainty of evidence for anthropometric outcomes was downgraded by one level due to serious concerns regarding inconsistency. This decision was made because of the substantial statistical heterogeneity. Furthermore, our subgroup analysis revealed that the type of control group was a key source of heterogeneity, indicating that the observed dispersion was largely explained by variations in study design rather than inconsistent intervention effects. Similarly, the certainty of evidence for dietary outcomes was downgraded by one level due to serious inconsistency among studies, despite a consistent direction of effects (Table 4).

Table . Summary of Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) assessment for each outcome.

Outcomes	No of studies	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	SMD[a] (95% CI)	Quality
Anthropometric outcomes	19 (3556 participants)	Not serious	Serious inconsistency	Not serious	Not serious	Not serious	0.26 (0.14 - 0.38) (equivalent to 2.62-6.55 kg reduction)	⊕⊕⊕◯ Moderate
Dietary outcomes	8 (1365 participants)	Not serious	Serious inconsistency	Not serious	Not serious	Not serious	0.26 (0.04 - 0.48)	⊕⊕⊕◯ Moderate

[a]SMD: standardized mean difference.

Meta-Regression

Meta-regression analysis showed that retention rates impact both anthropometric (β=0.005, 95% CI 0.0001 - 0.0100; *P*=.05) and dietary outcomes (β=0.0175, 95% CI 0.0103 - 0.0246; *P*<.001), indicating that higher participant retention was associated with greater intervention effectiveness (Figures 7 and 8). This suggests that an increase in participant retention was associated with a corresponding increase in the SMD, highlighting the importance of strategies to minimize dropout in digital interventions.

XSL•FO
RenderX

Figure 7. Meta-regression of participant retention rates on anthropometric outcomes. This bubble plot shows the association between the retention rate (%) and the standardized mean difference (SMD) of anthropometric outcomes. Each bubble represents a single study, with the size of the bubble corresponding to the inverse of the variance (study weight). The solid line represents the fitted regression line, showing a significant positive slope.

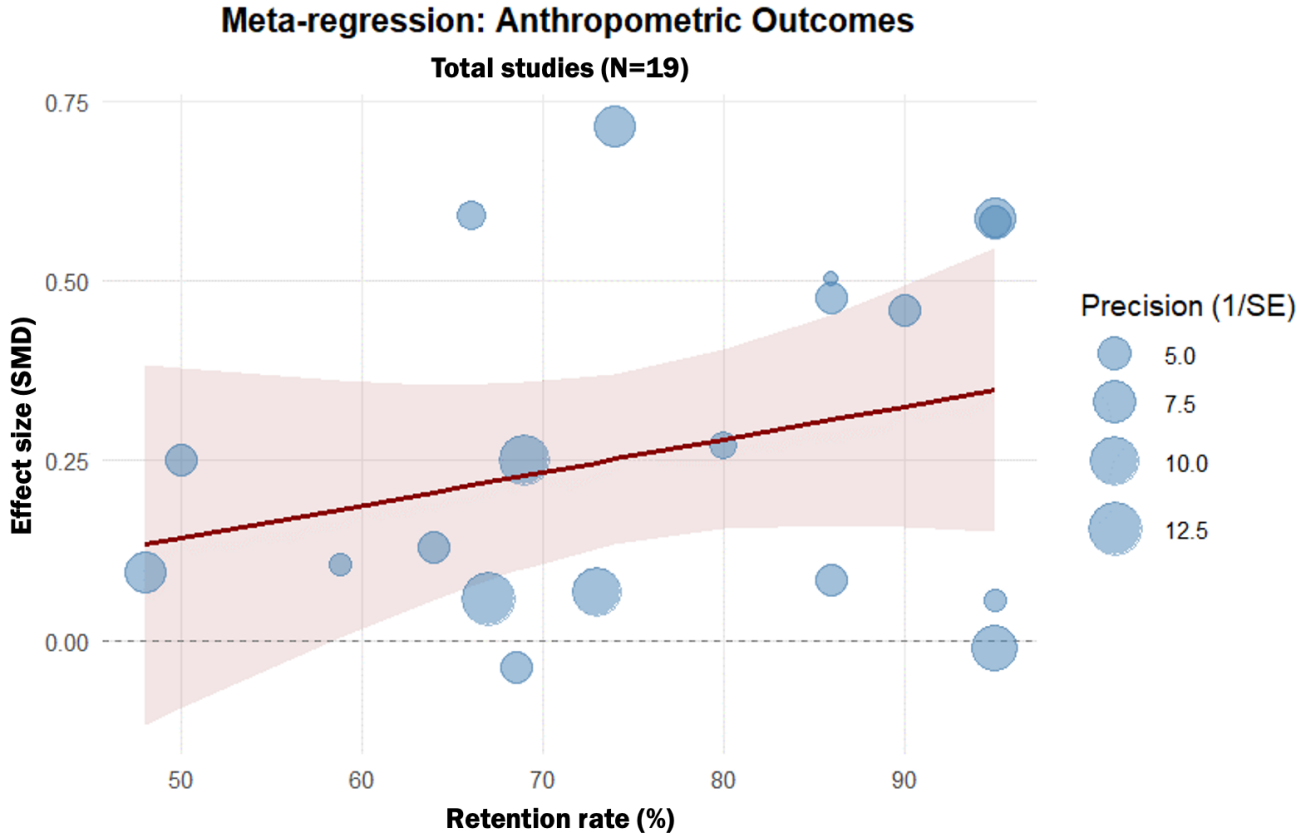

Figure 8. Meta-regression of participant retention rates on dietary outcomes. This bubble plot illustrates the relationship between retention rate (%) and the standardized mean difference (SMD) for dietary outcomes. Each bubble represents a single study, with the size of the bubble corresponding to the inverse of the variance (study weight). The solid line represents the fitted regression line, showing a significant positive slope.

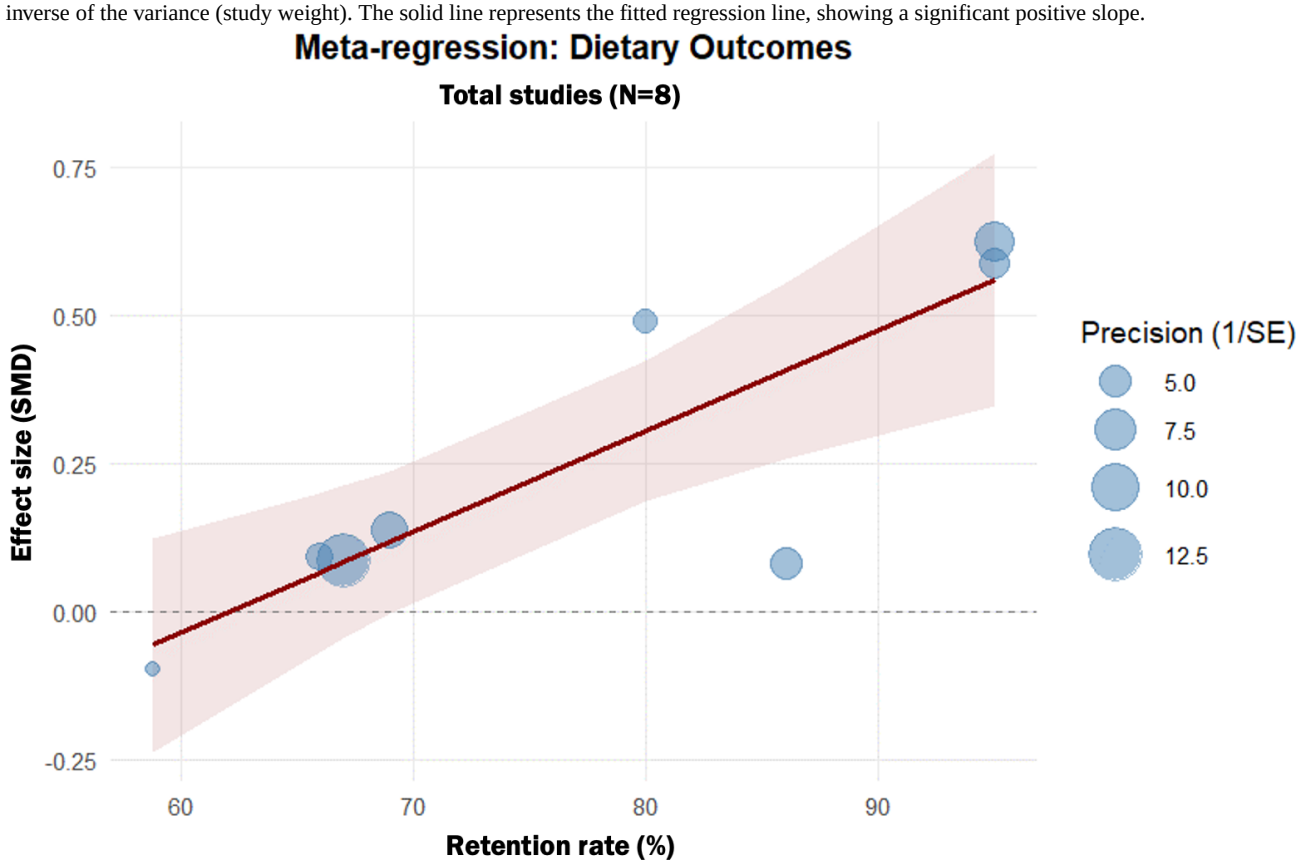

XSL•FO
RenderX

Discussion

Interpretation

This systematic review and meta-analysis evaluated the efficacy of stand-alone DLSIs on anthropometric and dietary outcomes among adults with overweight or obesity, with a rigorous focus on isolating the independent therapeutic potential of digital technology. Our findings demonstrate that stand-alone DLSIs can achieve statistically significant improvements in both anthropometric and dietary outcomes, supporting their role as independent interventions. Notably, the observed magnitude of weight loss exceeds the conventional threshold for a minimal clinically important difference in obesity management, which is typically cited as 2 - 3 kg [6]. This suggests that stand-alone digital platforms can achieve clinically meaningful outcomes comparable to more resource-intensive traditional programs, reinforcing their viability as a scalable public health tool, especially in resource-constrained settings [44].

To our knowledge, this study is innovative as it is the first to strictly delineate the effect of DLSIs by eliminating adjunctive human interaction. Distinct from previous meta-analyses that often included multimodal or hybrid interventions [14,15], this study clarifies the specific therapeutic potential of digital technology alone. Prior research incorporating structured group formats or peer interactions often reported numerically larger effect sizes [14,15]. However, our results confirm that stand-alone DLSIs still yield statistically significant benefits, suggesting that the digital component itself is a robust driver of change. This reinforces the utility of these interventions in environments where clinical labor or real-time moderation is unavailable [45].

To ensure a balanced interpretation of these results, several analytical factors must be considered. First, the observed statistical heterogeneity reflects the inherent diversity in digital health implementation [46]. Notably, our subgroup analyses clarified that this variance was primarily driven by comparator intensity, as heterogeneity was fully resolved in the active-control subgroup. Interventions targeting both diet and physical activity demonstrated synergistic effects, likely due to more comprehensive behavioral self-regulation [47,48]. Regarding intervention duration, an interesting discrepancy emerged: stand-alone DLSIs demonstrated consistent and significant benefits for anthropometric outcomes regardless of program length, indicating a durable impact on physical parameters. In contrast, dietary outcomes did not show statistically significant improvements in either short-term or long-term subgroups. The lack of significant effect across all timeframes suggests that stand-alone DLSIs may face challenges in independently facilitating meaningful dietary changes, regardless of the duration provided [49,50]. This discrepancy between anthropometric success and dietary stagnation may imply that while digital tools effectively promote weight loss—potentially through enhanced self-monitoring or physical activity tracking—they may require additional integrated support, such as environmental modifications, to overcome the complexities of dietary habit transformation [15]. Furthermore, effectiveness followed a gradient based on comparator type, being most effective against waitlist controls and diminishing against structured active comparators, likely reflecting nonspecific intervention effects such as attention or structure [51,52]. Meta-regression analyses further revealed that higher retention rates were significantly associated with greater improvements in both outcomes, underscoring that sustained user engagement is a critical moderator of effectiveness [51,52].

Furthermore, the risk of bias within the evidence base must be noted, as approximately 63.2% of the included studies were rated as having high risk or some concerns. Sensitivity analyses, however, supported the robustness of these findings, as excluding studies with a high risk of bias preserved the significance of the main results while leading to a marked reduction in heterogeneity, thereby reinforcing the reliability of the observed effects across higher-quality trials.

The credibility of these results is further reflected in the certainty of evidence. The certainty for both anthropometric and dietary outcomes was rated moderate and was primarily downgraded due to serious inconsistency. This moderate certainty indicates that while the current evidence is promising, further high-quality research will be valuable to increase the precision of these estimates and confirm the stability of the observed effects across diverse intervention strategies.

Importantly, the interpretation of these findings requires a clear distinction between the average effect and the distribution of effects across different settings. While the 95% CIs confirm a significant average effect, the calculated 95% PIs for both outcomes cross the null line. This critical distinction suggests that while stand-alone DLSIs are effective on average, the expected effect in a specific future clinical setting may be less certain, with the potential for null or even negative outcomes depending on the implementation context [53].

This study offers several methodological and practical strengths. First, it is the first meta-analysis to isolate the effects of DLSIs, thereby eliminating potential confounding from adjunctive intervention formats. Second, by evaluating both anthropometric and dietary outcomes across diverse delivery modes and comparator types, this study provides a comprehensive and scalable understanding of stand-alone DLSI effectiveness [6]. Third, the inclusion of detailed subgroup and meta-regression analyses helped clarify the conditions under which stand-alone DLSIs are most effective—particularly in relation to targeted behaviors, intervention duration, and comparator group. Fourth, the exclusive inclusion of RCTs, combined with rigorous quality assessments and sensitivity analyses, enhances the methodological credibility of the findings. Finally, the moderate certainty of evidence for both anthropometric and dietary outcomes further underscores the reliability of the observed effects, reinforcing the potential of stand-alone DLSIs as evidence-based tools for obesity management [24].

Limitations

Nevertheless, several limitations should be acknowledged when interpreting these findings. First, most included trials featured short intervention periods (typically ≤12 weeks), which restrict conclusions regarding the long-term maintenance of weight-related changes. Second, substantial heterogeneity was

XSL•FO
RenderX

not consistently explained by device type or behavioral change techniques, suggesting that the variance may stem from unmeasured variables such as individual-level user interaction patterns. Third, the lack of standardized reporting on intervention intensity and cost-effectiveness hinders direct translation into scalable health policy. Fourth, maintaining user engagement remains an inherent challenge in unsupervised digital programs [51]. Fifth, adherence and user engagement—critical determinants of digital intervention success—were underreported or inconsistently measured across trials, precluding formal analysis of these as moderators. Sixth, the presence of potential publication bias for certain outcomes suggests that the observed effect sizes might be slightly overestimated, necessitating a cautious interpretation [22]. Finally, the absence of formal cost-effectiveness evaluations within the included studies limits the translation of findings into scalable health policy recommendations.

To address these limitations, future research should prioritize long-term RCTs that can evaluate the sustainability of DLSI-induced weight changes beyond the typical short-term window. Additionally, more granular analyses are needed to isolate the active components and delivery mechanisms—such as platform type or specific behavioral change techniques—that contribute most significantly to intervention efficacy. Special attention should be given to developing and testing personalized features, such as AI-driven tailoring that adapts to real-time user data, and the cultural adaptation of content to meet the specific needs of diverse ethnic and linguistic populations [54]. Given the underreporting of adherence and engagement metrics in current trials, future studies should integrate standardized measures of user interaction and test the impact of adaptive strategies, including AI-driven personalization, on maintaining user retention and enhancing outcomes. Finally, incorporating cost-effectiveness analyses will provide critical insight for policymakers considering the adoption of stand-alone DLSIs in routine care and public health programming.

Implications

This systematic review and meta-analysis provides several critical implications for the field of digital health and obesity management. First, in terms of innovation and distinctiveness, our study is unique in that, unlike previous reviews that often conflated purely digital tools with hybrid models involving in-person coaching, it isolated the stand-alone effect of DLSIs without individualized clinical support. This methodological focus distinguishes our review from existing reviews by strictly excluding human-contact confounding factors to demonstrate that digital technology alone can achieve statistically significant weight loss and dietary improvements [14,15]. Second, regarding its contribution to the field, our findings highlight that while the overall effect size of stand-alone interventions may be smaller than that of human-supported ones, they offer a highly scalable and cost-effective solution for public health. This study specifically identifies that stand-alone DLSIs are most effective when compared to minimal-intervention or waitlist controls, providing a clear benchmark for future digital health tool development. Finally, for real-world implications, these results suggest that stand-alone DLSIs can serve as an accessible first-step intervention in a stepped-care model for obesity [55]. For clinicians and policymakers, this means that stand-alone digital platforms can be widely deployed to populations with limited access to intensive in-person counseling, offering a viable strategy to reduce the global burden of obesity-related diseases [1,2].

Conclusions

Our study validates stand-alone DLSIs as effective primary tools for obesity management, offering moderate-certainty evidence free from direct human-contact confounding. This review contributes to the field by validating technology-driven interventions as effective primary tools. In the real world, these results imply that stand-alone DLSIs offer a scalable first step in stepped-care models, providing an accessible strategy for clinicians and policymakers to address the global obesity burden. However, given the substantial heterogeneity and 95% PIs crossing the null, these average effects should not be viewed as guaranteed benefits. The context-dependent efficacy of stand-alone DLSIs suggests that while they are a viable public health alternative, their implementation requires careful monitoring for consistency across different settings.

Acknowledgments

We thank all contacted authors for providing additional information and data for this meta-analysis. The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading, editing, and reformatting text. The GAI tool used was Gemini (version 3.0). Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the outcomes. This declaration was submitted by JHP.

Funding

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (grant no 2021R1I1A3041487), and the Soonchunhyang University Research Fund. The funders had no involvement in the study design, data collection, analysis, interpretation of results, decision to publish, or preparation of the manuscript.

Data Availability

The datasets generated and/or analyzed during this study, including the extracted study characteristics and outcome data used for the meta-analyses, are available from the corresponding author upon reasonable request.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File, 26 KB - jmir_v28i1e81070_app1.docx]

Multimedia Appendix 2
Behavior change techniques (BCTs) applied in the included studies.
[DOCX File, 15 KB - jmir_v28i1e81070_app2.docx]

Multimedia Appendix 3
Verification of digital-led delivery mechanisms and intervention components in the included trials.
[DOCX File, 19 KB - jmir_v28i1e81070_app3.docx]

Checklist 1
PRISMA 2020 checklist.
[DOCX File, 26 KB - jmir_v28i1e81070_app4.docx]

Checklist 2
PRISMA-S checklist.
[DOCX File, 18 KB - jmir_v28i1e81070_app5.docx]

References

1. Obesity and overweight. World Health Organization. 2025. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [accessed 2026-03-01]
2. Boutari C, Mantzoros CS. A 2022 update on the epidemiology of obesity and a call to action: as its twin COVID-19 pandemic appears to be receding, the obesity and dysmetabolism pandemic continues to rage on. Metab Clin Exp 2022 Aug;133:155217. [doi: 10.1016/j.metabol.2022.155217] [Medline: 35584732]
3. Ng M, Fleming T, Robinson M, et al. Global, regional, and national prevalence of overweight and obesity in children and adults during 1980-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet 2014 Aug 30;384(9945):766-781. [doi: 10.1016/S0140-6736(14)60460-8] [Medline: 24880830]
4. Bray GA, Frühbeck G, Ryan DH, Wilding JPH. Management of obesity. Lancet 2016 May 7;387(10031):1947-1956. [doi: 10.1016/S0140-6736(16)00271-3] [Medline: 26868660]
5. Franz MJ, VanWormer JJ, Crain AL, et al. Weight-loss outcomes: a systematic review and meta-analysis of weight-loss clinical trials with a minimum 1-year follow-up. J Am Diet Assoc 2007 Oct;107(10):1755-1767. [doi: 10.1016/j.jada.2007.07.017] [Medline: 17904936]
6. Jensen MD, Ryan DH, Apovian CM, et al. 2013 AHA/ACC/TOS guideline for the management of overweight and obesity in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and The Obesity Society. J Am Coll Cardiol 2014 Jul 1;63(25 Pt B):2985-3023. [doi: 10.1016/j.jacc.2013.11.004] [Medline: 24239920]
7. Wadden TA, Butryn ML, Hong PS, Tsai AG. Behavioral treatment of obesity in patients encountered in primary care settings: a systematic review. JAMA 2014 Nov 5;312(17):1779-1791. [doi: 10.1001/jama.2014.14173] [Medline: 25369490]
8. Almandoz JP, Xie L, Schellinger JN, et al. Impact of COVID-19 stay-at-home orders on weight-related behaviours among patients with obesity. Clin Obes 2020 Oct;10(5):e12386. [doi: 10.1111/cob.12386] [Medline: 32515555]
9. Naik N, Hameed BMZ, Sooriyaperakasam N, et al. Transforming healthcare through a digital revolution: a review of digital healthcare technologies and solutions. Front Digit Health 2022;4:919985. [doi: 10.3389/fdgth.2022.919985] [Medline: 35990014]
10. Thomas JG, Wing RR. Health-e-call, a smartphone-assisted behavioral obesity treatment: pilot study. JMIR Mhealth Uhealth 2013 Apr 17;1(1):e3. [doi: 10.2196/mhealth.2164] [Medline: 25100672]
11. Spring B, Duncan JM, Janke EA, et al. Integrating technology into standard weight loss treatment: a randomized controlled trial. JAMA Intern Med 2013 Jan 28;173(2):105-111. [doi: 10.1001/jamainternmed.2013.1221] [Medline: 23229890]
12. Arem H, Irwin M. A review of web-based weight loss interventions in adults. Obes Rev 2011 May;12(5):e236-e243. [doi: 10.1111/j.1467-789X.2010.00787.x] [Medline: 20804523]
13. Hutchesson MJ, Rollo ME, Krukowski R, et al. eHealth interventions for the prevention and treatment of overweight and obesity in adults: a systematic review with meta-analysis. Obes Rev 2015 May;16(5):376-392. [doi: 10.1111/obr.12268] [Medline: 25753009]

XSL•FO
RenderX

14. Milne-Ives M, Burns L, Swancutt D, et al. The effectiveness and usability of online, group-based interventions for people with severe obesity: a systematic review and meta-analysis. Int J Obes 2025 Apr;49(4):564-577. [doi: 10.1038/s41366-024-01669-2]
15. Beleigoli AM, Andrade AQ, Cançado AG, Paulo MNL, Diniz MDFH, Ribeiro AL. Web-based digital health interventions for weight loss and lifestyle habit changes in overweight and obese adults: systematic review and meta-analysis. J Med Internet Res 2019 Jan 8;21(1):e298. [doi: 10.2196/jmir.9609] [Medline: 30622090]
16. Gentili A, Failla G, Melnyk A, et al. The cost-effectiveness of digital health interventions: a systematic review of the literature. Front Public Health 2022;10:787135. [doi: 10.3389/fpubh.2022.787135] [Medline: 36033812]
17. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc 2021 Apr 1;109(2):174-200. [doi: 10.5195/jmla.2021.962] [Medline: 34285662]
18. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
19. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
20. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Routledge Academic; 1998.
21. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 6;327(7414):557-560. [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
22. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634. [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
23. Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 2000 Jun;56(2):455-463. [doi: 10.1111/j.0006-341x.2000.00455.x] [Medline: 10877304]
24. Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol 2011 Apr;64(4):380-382. [doi: 10.1016/j.jclinepi.2010.09.011] [Medline: 21185693]
25. Allen JK, Stephens J, Dennison Himmelfarb CR, Stewart KJ, Hauck S. Randomized controlled pilot study testing use of smartphone technology for obesity treatment. J Obes 2013;2013:151597. [doi: 10.1155/2013/151597] [Medline: 24392223]
26. Brame J, Centner C, Berg N, Bartlam M, Gollhofer A, König D. Effects of a 12-week web-based weight loss program for adults with overweight and obesity on COVIDAge and lifestyle-related cardiometabolic risk factors: a randomized controlled trial. Front Public Health 2022;10:868255. [doi: 10.3389/fpubh.2022.868255] [Medline: 35669738]
27. Carter MC, Burley VJ, Nykjaer C, Cade JE. Adherence to a smartphone application for weight loss compared to website and paper diary: pilot randomized controlled trial. J Med Internet Res 2013 Apr 15;15(4):e32. [doi: 10.2196/jmir.2283] [Medline: 23587561]
28. Chung LMY, Law QPS, Fong SSM, Chung JWY. Teledietetics improves weight reduction by modifying eating behavior: a randomized controlled trial. Telemed J E Health 2014 Jan;20(1):55-62. [doi: 10.1089/tmj.2013.0104] [Medline: 24205807]
29. Collins CE, Morgan PJ, Jones P, et al. A 12-week commercial web-based weight-loss program for overweight and obese adults: randomized controlled trial comparing basic versus enhanced features. J Med Internet Res 2012 Apr 25;14(2):e57. [doi: 10.2196/jmir.1980] [Medline: 22555246]
30. Dunn C, Olabode-Dada O, Whetstone L, et al. Using synchronous distance education to deliver a weight loss intervention: a randomized trial. Obesity (Silver Spring) 2016 Jan;24(1):44-50. [doi: 10.1002/oby.21315] [Medline: 26637964]
31. Hurkmans E, Matthys C, Bogaerts A, Scheys L, Devloo K, Seghers J. Face-to-face versus mobile versus blended weight loss program: randomized clinical trial. JMIR Mhealth Uhealth 2018 Jan 11;6(1):e14. [doi: 10.2196/mhealth.7713] [Medline: 29326093]
32. Kohl J, Brame J, Centner C, et al. Effects of a web-based lifestyle intervention on weight loss and cardiometabolic risk factors in adults with overweight and obesity: randomized controlled clinical trial. J Med Internet Res 2023 Jun 27;25:e43426. [doi: 10.2196/43426] [Medline: 37368484]
33. Kraschnewski JL, Stuckey HL, Rovniak LS, et al. Efficacy of a weight-loss website based on positive deviance: a randomized trial. Am J Prev Med 2011 Dec;41(6):610-614. [doi: 10.1016/j.amepre.2011.08.012] [Medline: 22099238]
34. Krukowski RA, Tilford JM, Harvey-Berino J, West DS. Comparing behavioral weight loss modalities: incremental cost-effectiveness of an internet-based versus an in-person condition. Obesity (Silver Spring) 2011 Aug;19(8):1629-1635. [doi: 10.1038/oby.2010.341] [Medline: 21253001]
35. Lim SL, Ong KW, Johal J, et al. Effect of a smartphone app on weight change and metabolic outcomes in Asian adults with type 2 diabetes: a randomized clinical trial. JAMA Netw Open 2021 Jun 1;4(6):e2112417. [doi: 10.1001/jamanetworkopen.2021.12417] [Medline: 34081137]
36. Lugones-Sanchez C, Recio-Rodriguez JI, Agudo-Conde C, et al. Long-term effectiveness of a smartphone app combined with a smart band on weight loss, physical activity, and caloric intake in a population with overweight and obesity (EVIDENT 3 Study): randomized controlled trial. J Med Internet Res 2022 Feb 1;24(2):e30416. [doi: 10.2196/30416] [Medline: 35103609]

XSL•FO
RenderX

37. McConnon A, Kirk SF, Cockroft JE, et al. The internet for weight control in an obese sample: results of a randomised controlled trial. BMC Health Serv Res 2007 Dec 19;7:206. [doi: 10.1186/1472-6963-7-206] [Medline: 18093289]
38. Moravcová K, Sovová M, Ožana J, et al. Comparing the efficacy of digital and in-person weight loss interventions for patients with obesity and glycemic disorders: evidence from a randomized non-inferiority trial. Nutrients 2024 May 16;16(10):1510. [doi: 10.3390/nu16101510] [Medline: 38794747]
39. Padwal RS, Klarenbach S, Sharma AM, et al. The evaluating self-management and educational support in severely obese patients awaiting multidisciplinary bariatric care (EVOLUTION) trial: principal results. BMC Med 2017 Mar 2;15(1):46. [doi: 10.1186/s12916-017-0808-6] [Medline: 28249576]
40. Steinberg DM, Tate DF, Bennett GG, Ennett S, Samuel-Hodge C, Ward DS. The efficacy of a daily self-weighing weight loss intervention using smart scales and e-mail. Obesity (Silver Spring) 2013 Sep;21(9):1789-1797. [doi: 10.1002/oby.20396] [Medline: 23512320]
41. Svetkey LP, Batch BC, Lin PH, et al. Cell phone intervention for you (CITY): a randomized, controlled trial of behavioral weight loss intervention for young adults using mobile technology. Obesity (Silver Spring) 2015 Nov;23(11):2133-2141. [doi: 10.1002/oby.21226] [Medline: 26530929]
42. Vaz CL, Carnes N, Pousti B, Zhao H, Williams KJ. A randomized controlled trial of an innovative, user-friendly, interactive smartphone app-based lifestyle intervention for weight loss. Obes Sci Pract 2021 Oct;7(5):555-568. [doi: 10.1002/osp4.503] [Medline: 34631134]
43. Yardley L, Ware LJ, Smith ER, et al. Randomised controlled feasibility trial of a web-based weight management intervention with nurse support for obese patients in primary care. Int J Behav Nutr Phys Act 2014 May 21;11:67. [doi: 10.1186/1479-5868-11-67] [Medline: 24886516]
44. Matthay EC, Hagan E, Gottlieb LM, et al. Powering population health research: considerations for plausible and actionable effect sizes. SSM Popul Health 2021 Jun;14:100789. [doi: 10.1016/j.ssmph.2021.100789] [Medline: 33898730]
45. Recommendations on digital interventions for health system strengthening. World Health Organization. 2019. URL: https://www.who.int/publications/i/item/9789241550505 [accessed 2026-03-01]
46. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implementation Sci 2011 Dec;6(1):42. [doi: 10.1186/1748-5908-6-42]
47. Biddle SJH, Asare M. Physical activity and mental health in children and adolescents: a review of reviews. Br J Sports Med 2011 Sep;45(11):886-895. [doi: 10.1136/bjsports-2011-090185] [Medline: 21807669]
48. Yardley L, Choudhury T, Patrick K, Michie S. Current issues and future directions for research into digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):814-815. [doi: 10.1016/j.amepre.2016.07.019] [Medline: 27745680]
49. Bates S, Norman P, Breeze P, Brennan A, Ahern AL. Mechanisms of action in a behavioral weight-management program: latent growth curve analysis. Ann Behav Med 2022 Jan 1;56(1):64-77. [doi: 10.1093/abm/kaab019] [Medline: 33821937]
50. Johns DJ, Hartmann-Boyce J, Jebb SA, Aveyard P, Behavioural Weight Management Review Group. Diet or exercise interventions vs combined behavioral weight management programs: a systematic review and meta-analysis of direct comparisons. J Acad Nutr Diet 2014 Oct;114(10):1557-1568. [doi: 10.1016/j.jand.2014.07.005] [Medline: 25257365]
51. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017 Jun;7(2):254-267. [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
52. Mohr DC, Cuijpers P, Lehman KA. Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions. J Med Internet Res 2011 Mar 10;13(1):e30. [doi: 10.2196/jmir.1602] [Medline: 21393123]
53. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. BMJ 2011 Feb 10;342:d549. [doi: 10.1136/bmj.d549] [Medline: 21310794]
54. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462. [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
55. Carels RA, Wott CB, Young KM, et al. Successful weight loss with self-help: a stepped-care approach. J Behav Med 2009;32:503-509. [doi: 10.1007/s10865-009-9218-3]

Abbreviations

AI: artificial intelligence
BCT: behavior change technique
DLSI: digital lifestyle intervention
HKSJ: Hartung-Knapp-Sidik-Jonkman
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search Extension
RCT: randomized controlled trial
SMD: standardized mean difference

XSL•FO
RenderX

WHO: World Health Organization

Edited by S Brini; submitted 22.Jul.2025; peer-reviewed by N Anwar, WJ He; revised version received 22.Mar.2026; accepted 24.Mar.2026; published 04.May.2026.

Please cite as:
Lee SA, Park JH
Effects of Stand-Alone Digital Lifestyle Interventions on Weight-Related Outcomes in Adults With Overweight or Obesity: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e81070
URL: https://www.jmir.org/2026/1/e81070
doi:10.2196/81070

© Si-An Lee, Jin-Hyuck Park. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 4.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Deep Learning Algorithms Versus Radiologists in Digital Breast Tomosynthesis for Breast Cancer Detection: Systematic Review and Meta-Analysis

Shewen Lyu, PhD; Zepeng Wang, PhD; Yujing Mu, PhD; Luyao Wang, PhD; Xiaohua Pei, MD

Beijing University of Chinese Medicine Third Affiliated Hospital, 51 Xiaoguan Street, Andingmenwai, Chaoyang District, Beijing, China

Corresponding Author:
Xiaohua Pei, MD
Beijing University of Chinese Medicine Third Affiliated Hospital, 51 Xiaoguan Street, Andingmenwai, Chaoyang District, Beijing, China

Abstract

Background: Deep learning (DL) algorithms for digital breast tomosynthesis (DBT) have proliferated, demonstrating emerging potential in enhancing lesion detection and classification.

Objective: This study aimed to compare the diagnostic performance of DL algorithms for DBT with that of radiologists of varying experience and assess the clinical impact of DL assistance.

Methods: A systematic search of PubMed, Embase, Web of Science, and the Cochrane Library was conducted up to November 8, 2025. Included studies compared the performance of stand-alone DL algorithms for DBT, radiologist interpretation alone, and DL-assisted diagnosis. Study quality was assessed using the Prediction Model Risk of Bias Assessment Tool+Artificial Intelligence (PROBAST+AI). Performance metrics were pooled using bivariate random effects and generalized linear mixed models.

Results: A total of 13 studies with 38,565 patients were included in the final analysis. Stand-alone DL algorithms achieved a pooled sensitivity of 0.88 (95% CI 0.80-0.93), specificity of 0.74 (95% CI 0.59-0.85), and area under the receiver operating characteristic curve (AUC) of 0.89 (95% CI 0.86-0.92). While DL performance showed no statistically significant difference compared to all radiologists (AUC=0.89 vs 0.88; *P*=.64) or senior radiologists (AUC=0.89 vs 0.90; *P*=.48), DL demonstrated significantly superior sensitivity compared to junior radiologists (0.88 vs 0.76; *P*=.03). Notably, DL assistance did not statistically improve diagnostic metrics for radiologists across any experience level. Meta-regression identified validation methods as a significant source of heterogeneity.

Conclusions: DL algorithms for DBT exhibited strong diagnostic proficiency and showed higher sensitivity than junior radiologists, suggesting their potential utility as adjunctive tools to help reduce oversight in less experienced settings. However, given that DL assistance did not significantly elevate overall human performance, current models act primarily as supplementary aids rather than definitive clinical tools. Future prospective multimodal studies are warranted to validate these findings and optimize clinical integration.

Trial Registration: PROSPERO CRD420251242858; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251242858

(J Med Internet Res 2026;28:e91659) doi:10.2196/91659

KEYWORDS

digital breast tomosynthesis; deep learning; breast neoplasms; diagnostic accuracy; meta-analysis; artificial intelligence; AI

Introduction

Breast cancer is the most commonly diagnosed cancer and the leading cause of cancer deaths among women worldwide, with an estimated 2.3 million new cases and 666,000 deaths occurring worldwide in 2022. This profound global burden underscores the urgency for early and accurate detection to improve prognosis and reduce the burden of invasive treatments [1,2]. While imaging serves as the cornerstone of screening, the transition from digital mammography to digital breast tomosynthesis (DBT) has revolutionized clinical practice. By providing quasi-3D volumetric data, DBT significantly alleviates the issue of tissue superposition, thereby increasing cancer detection rates and reducing unnecessary recall rates [3].

Despite these advantages, DBT introduces new challenges. The substantial increase in image volume significantly prolongs interpretation time, contributing to radiologist fatigue and potential cognitive overload [4,5]. Furthermore, despite overall improvements in detection rates, challenges regarding false positives (FPs) and false negatives (FNs) persist, with specific subtypes such as invasive lobular carcinoma prone to being overlooked [4,6]. The interpretation of DBT imagery is also

XSL•FO
RenderX

heavily dependent on the radiologist's experience, which increases subjectivity and the risk of misdiagnosis [7].

To address these diagnostic bottlenecks, deep learning (DL) algorithms for DBT have been increasingly developed, demonstrating emerging potential in enhancing lesion detection and classification. These algorithms aim to assist radiologists by extracting complex feature representations that may be imperceptible to the human eye [7]. However, results across the literature are inconsistent. While some pivotal studies suggest DL superiority [8], others indicate that algorithms may struggle with FPs or lack generalizability across different vendors and populations [9,10]. Moreover, the comparative performance of DL against radiologists of varying expertise (eg, junior and senior radiologists) remains a subject of ongoing debate [9,10]. This review includes data from 38,565 patients across the included studies.

Given the rapid accumulation of new evidence and the heterogeneity of study designs, a rigorous synthesis of current data is warranted. Therefore, the purpose of this systematic review and meta-analysis was to comprehensively compare the relative diagnostic performance and added value of DL algorithms vs radiologists of varying experience levels and evaluate potential factors influencing the diagnostic performance of these algorithms.

Methods

The meta-analysis was carried out in full compliance with the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Diagnostic Test Accuracy) guidelines [11]. Furthermore, the protocol for this study was registered with PROSPERO (CRD420251242858).

Ethical Considerations

This was a systematic review and meta-analysis, so ethics approval and consent to participate are not applicable.

Search Strategy

A comprehensive literature search was conducted up to November 8, 2025, across 4 electronic databases: PubMed, Embase, Web of Science, and Cochrane Library. The search strategy, designed and executed to maximize sensitivity, used a combination of free-text terms and controlled vocabulary (eg, MeSH [Medical Subject Headings] terms in PubMed). Key concepts included three domains: (1) artificial intelligence (AI; eg, "Deep Learning"), (2) the target disease (eg, "Breast Neoplasms"), and (3) the imaging modality (eg, "Digital Breast Tomosynthesis"). No restrictions were placed on language, publication date, or study type. Two independent reviewers (SL and YM) performed initial title and abstract screening followed by full-text assessment of potentially eligible studies. To ensure literature saturation, the reference lists of all included articles were manually screened. The full, detailed search syntax for each database is provided in Table S1 in Multimedia Appendix 1.

Inclusion and Exclusion Criteria

Studies were selected according to the participants, index test, target condition, reference standard, outcomes, and setting framework: women undergoing breast cancer screening via DBT (participants); evaluated the performance of stand-alone DL algorithms for DBT, independent radiologist interpretation, or DL-assisted radiologist diagnosis (index test); breast cancer, confirmed via histopathology for positive cases (target condition); final diagnosis based on histopathology (for positive cases) or clinical imaging follow-up (for negative cases; reference standard); primary outcomes, including diagnostic performance measures (sensitivity, specificity, and area under the receiver operating characteristic [ROC] curve [AUC]), and secondary outcomes, comprising clinical impact metrics (detection rate, positive predictive value [PPV], and recall rate; outcomes); and original studies using retrospective or prospective cohorts from screening programs or clinical databases (setting).

In addition, we systematically excluded studies whose titles and abstracts were clearly irrelevant, as well as noneligible publication types, including reviews, case reports, conference abstracts, meta-analyses, and letters to the editor. Furthermore, studies that did not involve DL for DBT; without algorithm comparison; and with true positive (TP), FP, true negative (TN), and FN data not available were also excluded. The screening process was conducted in duplicate by 2 independent reviewers (SL and ZW), with any disagreements resolved through consultation with a third reviewer (XP).

Quality Assessment and Certainty of Evidence

We used the updated Prediction Model Risk of Bias Assessment Tool+AI (PROBAST+AI) quality assessment tool [12], which replaces the original 2019 Prediction Model Risk of Bias Assessment Tool instrument. This tool adopts a 2-phase structure comprising model development and model evaluation. Each phase includes 7 domains addressing participants, data sources, predictors, outcome assessment, and analytical approaches. For each domain, the risk-of-bias judgment is categorized as low, high, or unclear based on responses to predefined signaling questions. The full set of signaling questions and evaluation templates can be found in Tables S2 and S3 in Multimedia Appendix 1. To ensure objectivity and accuracy in the evaluation process, 2 reviewers (SL and LW) independently assessed the risk of bias in the included studies. The certainty of evidence was evaluated using the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) framework [13]. Detailed evaluation items, decision rules, and domain-specific judgments can be found in Table S4 in Multimedia Appendix 1.

Data Extraction

Two reviewers (SL and YM) independently extracted data from the included full-text articles, and disagreements were resolved through consultation with a third reviewer (XP). Extracted items included patient characteristics, details of DL methods, DBT imaging parameters, stand-alone DL performance, radiologist performance, and radiologist performance with DL assistance. Because most studies did not report full contingency tables, we used two strategies to derive TP, FP, FN, and TN values: (1) back calculation using reported sensitivity, specificity, number of positive cases, and total sample size; and (2) redigitizing ROC curves using the GetData software to obtain optimal

XSL•FO
RenderX

sensitivity and specificity based on the Youden index. For DL algorithms, only validation set results were collected. When multiple DL models or radiologists were reported with overlapping cohorts, only the best-performing result (highest AUC) was extracted to avoid duplication.

Outcome Measures

The primary outcome measures included the sensitivity, specificity, and AUC of DL algorithms, radiologists, and radiologists assisted by DL, as well as detection rate, PPV, and recall rate. Secondary outcomes focused on diagnostic performance stratified by radiologist experience both with and without DL assistance. Sensitivity (TP rate) reflected the ability to correctly identify cancer cases and was calculated as TP/(TP + FN) × 100%. Specificity (TN rate) represented the ability to correctly identify noncancer cases, calculated as TN/(TN + FP) × 100%. AUC summarizes overall discriminative ability. Detection rate was defined as TP/N × 100%, representing the proportion of cancers correctly detected in the screening population. PPV was calculated as TP/(TP + FP) × 100%, and recall rate was calculated as (TP + FP)/N × 100%. Radiologists with less than 5 years of experience were classified as junior, those with 5 or more years of experience were classified as senior, and studies without explicit experience data were categorized as unspecified.

Statistical Analysis

Given the expected methodological heterogeneity across studies, we applied a bivariate random-effects model [14] to pool sensitivity, specificity, and AUC estimates. For PPV, recall rate, and detection rate, log-transformation was performed prior to synthesis using a random-effects generalized linear mixed model framework. Differences in pooled diagnostic performance were assessed using a mean *Z*-test, with statistical significance defined as a *P* value of less than .05. Heterogeneity was quantified using the Higgins I^2 statistic [15]. For substantial heterogeneity (I^2>50%), bivariate box plots were used to explore potential sources, and multivariable meta-regression was conducted for DL algorithms to evaluate the impact of validation strategy, study design, region of interest, and data splitting method. Temporal changes in DL performance were examined using bubble plots, whereas violin plots were used to visualize differences in radiologist performance before and after DL assistance. Fagan nomograms were generated to assess the clinical implications for patients. The assessment of publication bias was conducted using the Deeks funnel plot [16]. Analyses were executed using Stata (version 15.1; StataCorp) with the *midas* and *metadta* commands, as well as R (version 4.5.1; R Foundation for Statistical Computing) using the *ggplot2* and *tidyverse* packages.

Results

Literature Search and Study Selection

A total of 1076 potentially relevant records were identified through the initial database search. Of these 1076 records, after removing 422 (39.2%) duplicates, 654 (60.8%) proceeded to title and abstract screening. During this stage, of the 654 remaining articles, 634 (96.9%) were excluded due to clear irrelevance or noneligible publication types. The remaining 20 articles were assessed in full text. Following detailed evaluation, of these 20 articles, 2 (10%) [17,18] were excluded because the data required to construct contingency tables (TP, FP, TN, and FN) were unavailable, 4 (20%) [19-22] were excluded because they were not research on DL for DBT, and 3 (15%) [23-25] lacked direct algorithm comparisons. In addition, 2 eligible records were identified from nondatabase sources. Ultimately, 13 studies met all the inclusion criteria and were incorporated into the meta-analysis [8-10,26-35]. The study selection process followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines and is illustrated in Figure 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram illustrating the study selection process. DBT: digital breast tomosynthesis; DL: deep learning; FN: false negative; FP: false positive; TN: true negative; TP: true positive.

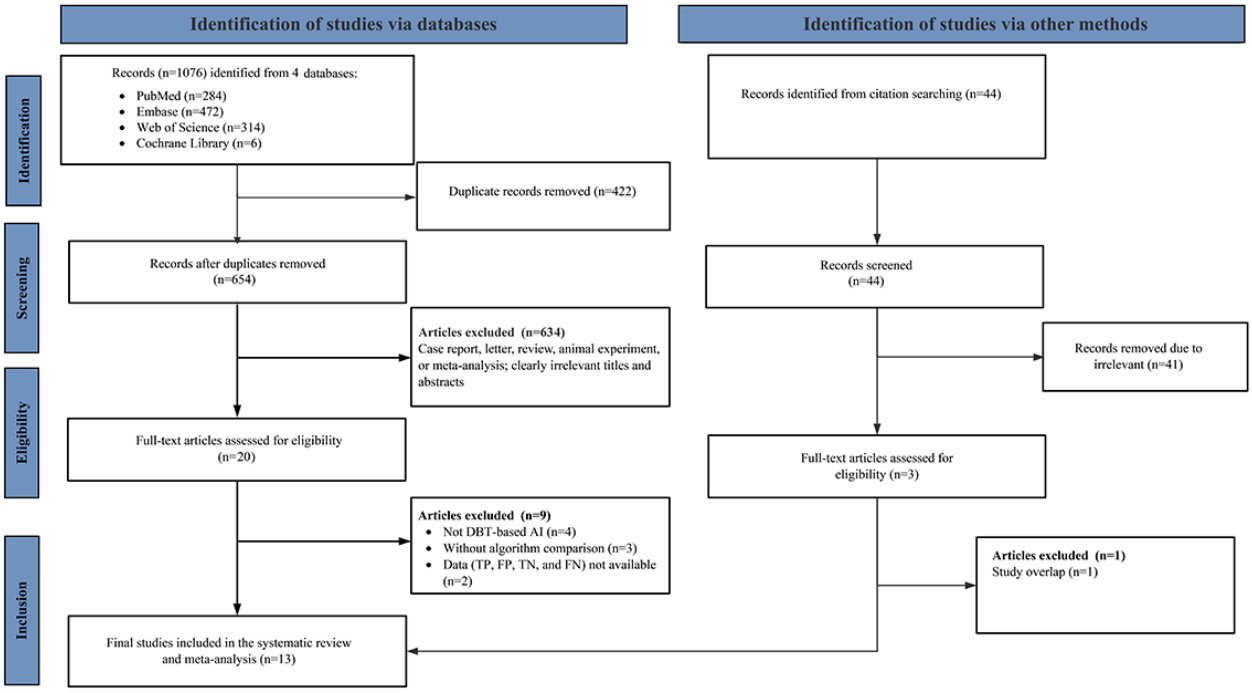

Study Description and Quality Assessment

A total of 13 studies met the eligibility criteria. Among these, 46.2% (n=6) [8,9,28,31,32,35] reported internal validation cohorts (n=25,885 patients), and 61.5% (n=8) [9,10,26,27,29,30,33,34] reported external validation cohorts (n=12,680 patients; n=1, 7.7% of the studies reported both). In total, 61.5% (n=8) of the studies [9,10,26,27,29,30,33,34] evaluated stand-alone DL algorithms, and 84.6% (n=11) [8-10,26,28-33,35] assessed changes in radiologist performance before and after DL assistance. All studies (n=13, 100%) [8-10,26-35] reported outcomes for the overall radiologist group, whereas 15.4% (n=2) [10,35] included junior radiologists, 30.8% (n=4) [10,27,31,35] included senior radiologists, and 69.2% (n=9) [8,9,26,28-30,32-34] involved radiologists with unspecified experience levels. The included studies were published between 2017 and 2025. A total of 92.3% (n=12) [8-10,27-35] of the studies were retrospective in design, and 7.7% (n=1) [26] were prospective; all used pathological biopsy as the reference standard. Detailed characteristics are summarized in Tables S5 to S11 in Multimedia Appendix 1.

The quality assessment using the PROBAST+AI tool is shown in Figure 2 and Tables S2 and S3 in Multimedia Appendix 1. For model development, no study was rated as high risk in terms of either overall risk of bias or applicability concerns. For model evaluation, 50% (4/8) of the studies were judged to have high overall risk of bias, whereas none showed high applicability concerns. Overall, although the risk of bias in the evaluation phase was notable, the applicability of the included studies was generally acceptable. According to the GRADE framework, the certainty of the evidence ranged from low to moderate, primarily downgraded due to risk of bias and imprecision, as detailed in Table S4 in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 2. Risk of bias and applicability concerns of the included studies using the Prediction Model Risk of Bias Assessment Tool+Artificial Intelligence (PROBAST+AI) tool: (A) summary of PROBAST+AI assessment for model development and (B) summary of PROBAST+AI assessment for model evaluation.

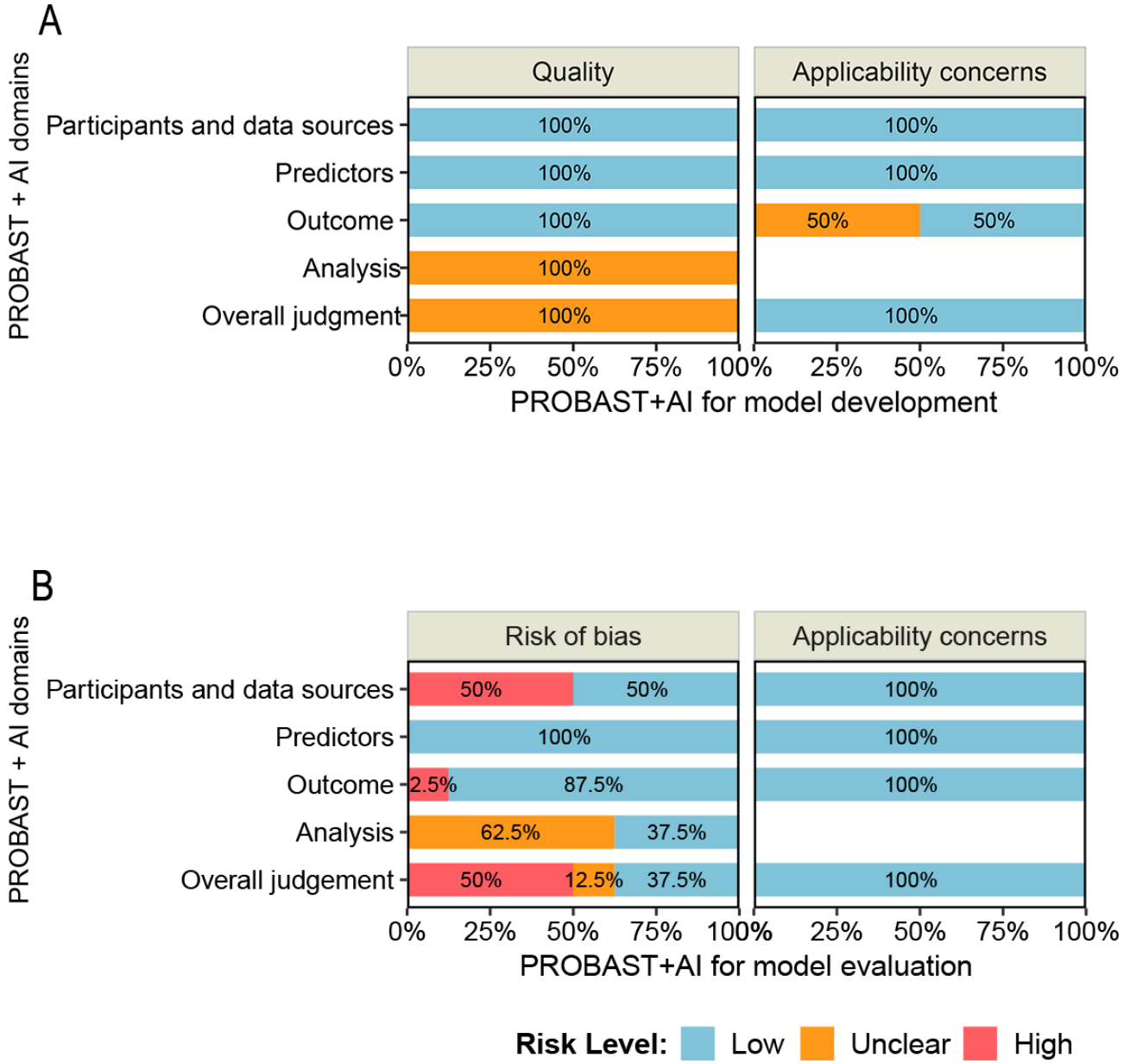

DL Algorithms

The sensitivity of DL algorithms was found to be 0.88 (95% CI 0.80-0.93; I^2=96.22%; low certainty), and the specificity was 0.74 (95% CI 0.59-0.85; I^2=99.63%; low certainty). The AUC was 0.89 (95% CI 0.86-0.92; low certainty), whereas the detection rate was 0.14 (95% CI 0.06-0.29; I^2=98.8%; low certainty). The PPV was 0.41 (95% CI 0.18-0.70; I^2=99.1%; low certainty), and the recall rate was 0.39 (95% CI 0.27-0.53; I^2=99.7%; low certainty). As shown in Tables 1 and 2 and Figures S1 to S5 in Multimedia Appendix 1, the bubble chart indicates that the AUC values remained relatively stable from 2021 to 2024, as shown in Figure 3A.

XSL•FO
RenderX

Table . Deep learning (DL) algorithms for digital breast tomosynthesis (DBT) vs radiologists of different experience levels in terms of diagnostic performance outcomes for breast cancer diagnosis.

Subgroup	Sample size, n	Validation datasets, n	Sensitivity (95% CI)	Difference in sensitivity[a]		Specificity (95% CI)	Difference in specificity[a]		AUC[b] (95% CI)	Difference in AUC[a]	
				Z-score	*P* value		*Z*-score	*P* value		*Z*-score	*P* value
Overall DL for DBT	12,555	8	0.88 (0.80-0.93)	__[c]	—	0.74 (0.59 - 0.85)	—	—	0.89 (0.86-0.92)	—	—
DL for DBT internal validation	5182	1	0.96 (0.90-1.00)	2.63	.009	0.40 (0.03-0.76)	1.95	.05	0.84 (0.78-0.89)	1.62	.11
DL for DBT external validation	7373	7	0.85 (0.77-0.90)	—	—	0.78 (0.65-0.87)	—	—	0.89 (0.86-0.91)	—	—
DL for DBT vs all radiologists				1.33	.18		1.51	.13		0.46	.64
DL for DBT	12,555	8	0.88 (0.80-0.93)			0.74 (0.59-0.85)			0.89 (0.86-0.92)		
All radiologists	36,245	14	0.83 (0.79-0.86)			0.86 (0.75-0.92)			0.88 (0.85-0.91)		
DL for DBT vs junior radiologists				2.12	.03		1.35	.18		—	—
DL for DBT	12,555	8	0.88 (0.80-0.93)			0.74 (0.59-0.85)			0.89 (0.86-0.92)		
Junior radiologists	310	2	0.76 (0.66-0.84)			0.94 (0.48-1.00)			—		
DL for DBT vs senior radiologists				0.50	.62		1.28	.20		.50	.62
DL for DBT	12,555	8	0.88 (0.80-0.93)			0.74 (0.59-0.85)			0.89 (0.86-0.92)		
Senior radiologists	961	4	0.86 (0.81-0.90)			0.93 (0.48-1.00)			0.90 (0.87-0.92)		

[a]On the basis of 2-sided *Z*-test.

[b]AUC: area under the receiver operating characteristic curve.

[c]Not applicable.

XSL•FO
RenderX

Table . Deep learning (DL) for digital breast tomosynthesis (DBT) vs radiologists of different experience levels in terms of clinical impact outcomes for breast cancer diagnosis.

Subgroup	Sample size n	Validation datasets, n	Detection rate (95% CI)	Difference in detection rate[a]		PPV[b] (95% CI)	Difference in PPV[a]		Recall rate (95% CI)	Difference in recall rate[a]	
				Z-score	*P* value		*Z*-score	*P* value		*Z*-score	*P* value
Overall DL for DBT	12,555	8	0.14 (0.06-0.29)	__[c]	—	0.41 (0.18-0.70)	—	—	0.39 (0.27-0.53)	—	—
DL for DBT internal validation	5182	1	0.09 (0.08-0.09)	0.87	.38	0.13 (0.12-0.15)	2.33	.02	0.63 (0.62-0.65)	4.04	<.001
DL for DBT external validation	7373	7	0.15 (0.06-0.33)	—	—	0.47 (0.19-0.76)	—	—	0.36 (0.24-0.50)	—	—
DL for DBT vs all radiologists				0.18	.91		1.27	.21		0.96	.34
DL for DBT	12,555	8	0.14 (0.06-0.29)			0.41 (0.18-0.70)			0.39 (0.27-0.53)		
All radiologists	36,245	14	0.15 (0.06-0.30)			0.62 (0.40-0.79)			0.30 (0.19-0.45)		
DL for DBT vs junior radiologists				5.07	>.99		3.41	<.001		2.05	.04
DL for DBT	12,555	8	0.14 (0.06-0.29)			0.41 (0.18-0.70)			0.39 (0.27-0.53)		
Junior radiologists	310	2	0.47 (0.42-0.53)			0.87 (0.81-0.91)			0.55 (0.47-0.63)		
DL for DBT vs senior radiologists				1.94	.05		3.28	.001		0.25	.81
DL for DBT	12,555	8	0.14 (0.06-0.29)			0.41 (0.18-0.70)			0.39 (0.27-0.53)		
Senior radiologists	961	4	0.36 (0.19-0.57)			0.90 (0.70-0.97)			0.43 (0.22-0.68)		

[a]On the basis of 2-sided *Z*-test.

[b]PPV: positive predictive value.

[c]Not applicable.

XSL•FO
RenderX

Figure 3. Bubble plot and violin plots: (A) bubble plot of temporal trends in area under the receiver operating characteristic curve (AUC) performance of different deep learning (DL) models, (B) violin plots of diagnostic outcomes for all radiologists before and after DL assistance, (C) violin plots of diagnostic outcomes for junior radiologists before and after DL assistance, and (D) violin plots of diagnostic outcomes for senior radiologists before and after DL assistance. AI: artificial intelligence; CNN: convolutional neural network; NA: not available; PPV: positive predictive value; ResNet: residual neural network.

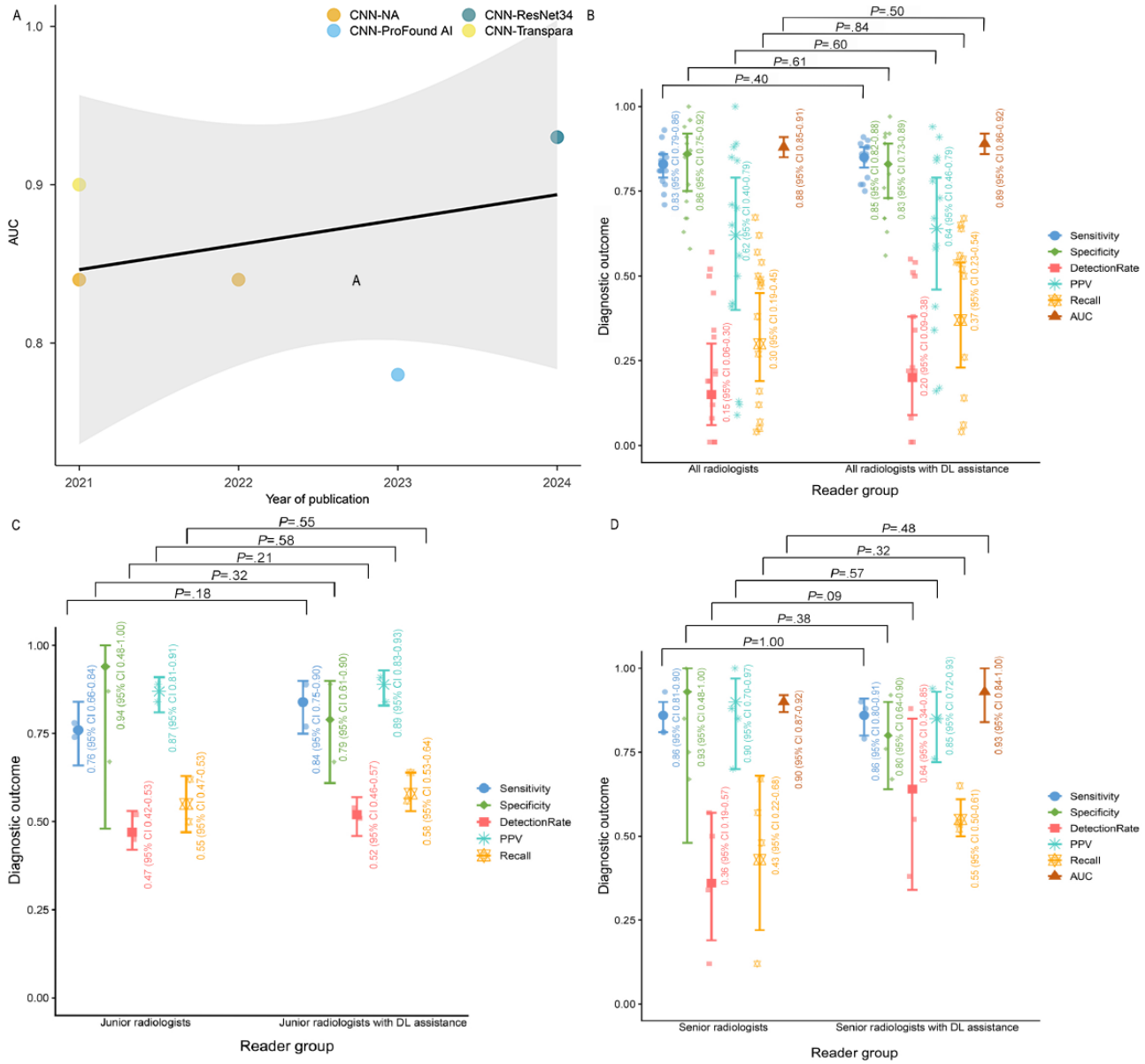

All Radiologists

The sensitivity of all radiologists was found to be 0.83 (95% CI 0.79-0.86; I^2=70.78%; low certainty), and the specificity was 0.86 (95% CI 0.75-0.92; I^2=99.35%; low certainty). The AUC was 0.88 (95% CI 0.85-0.91; low certainty), whereas the detection rate was 0.15 (95% CI 0.06-0.30; I^2=99.4%; moderate certainty). The PPV was 0.62 (95% CI 0.40-0.79; I^2=98%; low certainty), and the recall rate was 0.30 (95% CI 0.19-0.45; I^2=99.5%; low certainty). As shown in Table 1 and 2 and Figures S6 to S10 in Multimedia Appendix 1, there were no statistically significant differences between the performance of DL algorithms and that of all radiologists across all metrics.

Radiologists With Different Levels of Experience

DL algorithms for DBT achieved markedly higher sensitivity than junior radiologists (0.88 vs 0.76; Z=2.12; P=.03). However, DL algorithms for DBT exhibited significantly lower PPV than junior radiologists (0.41 vs 0.87; Z=3.41; P<.001) and significantly lower recall rate than junior radiologists (0.39 vs 0.55; Z=2.05; P=.04). Additionally, DL algorithms for DBT had a significantly lower PPV than senior radiologists (0.41 vs 0.90; Z=3.28; P=.001). These findings are illustrated in Tables 1 and 2.

Changes Before and After DL Assistance

With DL assistance, there were no significant improvements in any of the outcome measures for all radiologists. Similarly, there were no significant improvements in any of the outcome

XSL•FO
RenderX

measures for junior or senior radiologists. These findings are shown in Tables S12 and S13 in Multimedia Appendix 1.

Heterogeneity Testing: Bivariate Box Plots and Meta-Regression

The bivariate box plots suggested that the studies by Shoshan et al [9] and Pinto et al [29] might have contributed to the heterogeneity of DL algorithms, whereas those by Bassi et al [10] and Resch et al [27] might have been sources of heterogeneity among all radiologists, as shown in Figures 4A and 4B [9,10,27,29,35]. Meta-regression indicated that this heterogeneity primarily arose from differences in validation methods (internal validation vs external validation; specificity *P*=.05) and geographic regions (Europe vs North America; sensitivity *P*<.001), as shown in Table S14 in Multimedia Appendix 1.

Figure 4. Bivariate box plots and Deeks funnel plots: (A) bivariate box plot for stand-alone deep learning (DL) for digital breast tomosynthesis (DBT) reading, (B) bivariate box plot for radiologist reading, (C) Deeks funnel plot for stand-alone DL for DBT reading, and (D) Deeks funnel plot for radiologist reading [9,10,27,29,35]. ESS: effective sample size.

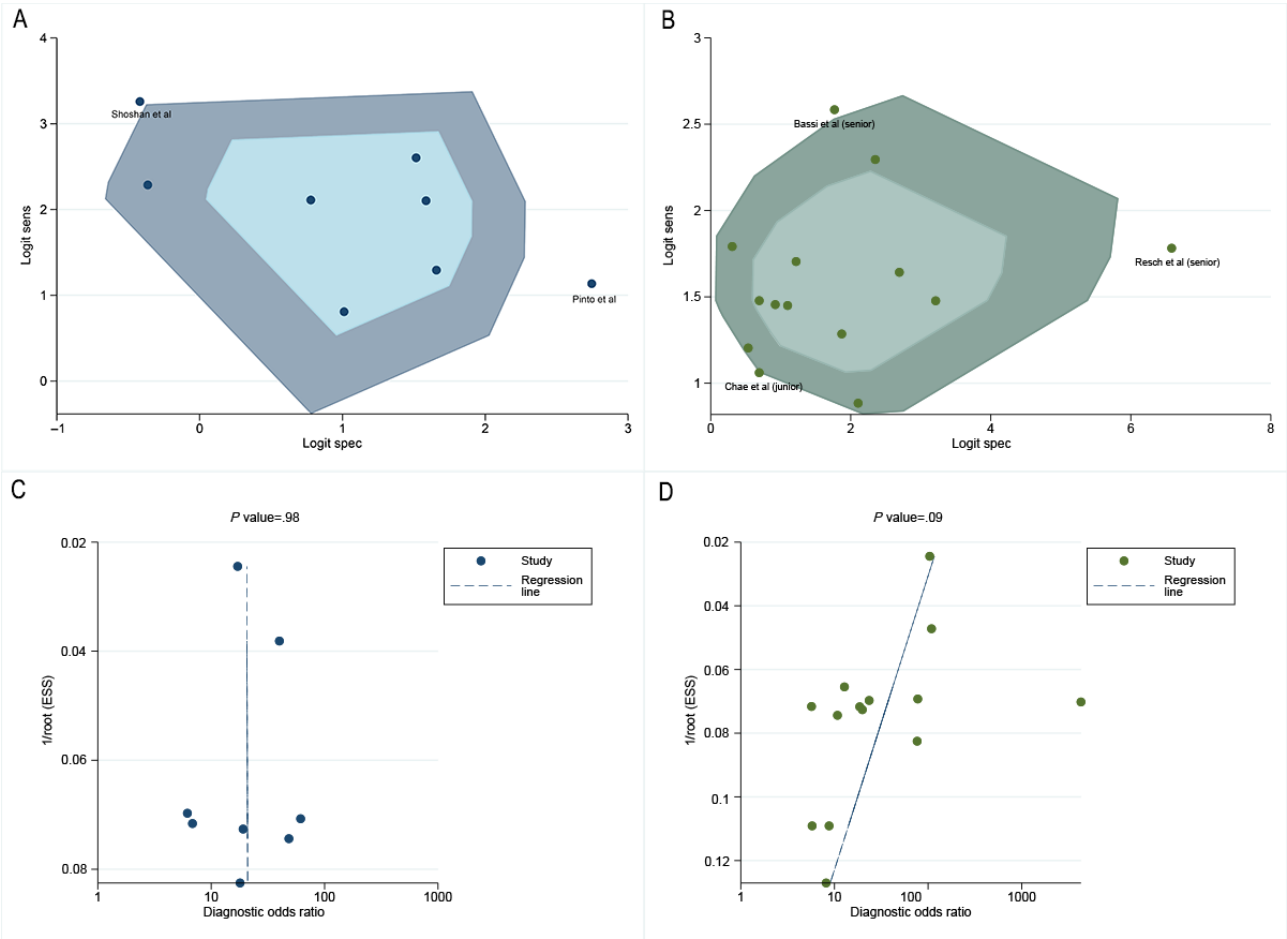

Sensitivity Analysis

To assess the robustness of the primary findings, 2 sensitivity analyses were conducted. First, after excluding outliers identified by the bivariate box plot graphical diagnostic method, the analysis yielded a sensitivity of 0.86 (95% CI 0.78 - 0.91), specificity of 0.74 (95% CI 0.62 - 0.83), AUC of 0.88 (95% CI 0.85 - 0.91), detection rate of 0.13 (95% CI 0.05 - 0.33), PPV of 0.38 (95% CI 0.14 - 0.69), and recall of 0.36 (95% CI 0.23 - 0.53). Second, after excluding studies assessed as having a high risk of bias using the PROBAST+AI tool in the validation set, the recalculated pooled effect sizes were as follows: sensitivity of 0.86 (95% CI 0.77 - 0.91), specificity of 0.80 (95% CI 0.57 - 0.93), AUC of 0.90 (95% CI 0.87 - 0.92), detection rate of 0.21 (95% CI 0.15 - 0.27), PPV of 0.61 (95% CI 0.37 - 0.81), and recall of 0.40 (95% CI 0.26 - 0.56). These results were consistent with those of the primary analysis, indicating that the overall conclusions regarding diagnostic performance were robust, neither unduly influenced by high-risk studies nor biased by individual outliers. These findings are shown in Table S15 in Multimedia Appendix 1.

Clinical Application Value and Publication Bias

The Deeks funnel plot asymmetry test showed no evidence of publication bias (0.98 vs 0.09), as illustrated in Figures 4C and 4D [9,10,27,29,35]. Using the median prevalence from the included studies as the prior probability, the Fagan nomogram for DL algorithms (median prevalence 26%, IQR 13.84%–38.95%) indicated a positive posttest probability of 54%, while that for all radiologists (median prevalence 26%, IQR 12.60%–55.01%) indicated a positive posttest probability of 67%, as shown in Figure 5.

XSL•FO
RenderX

Figure 5. Fagan nomograms illustrating the posttest probability of breast cancer with digital breast tomosynthesis (DBT) reading: (A) stand-alone deep learning (DL) for DBT reading and (B) radiologist reading without DL assistance. LR: likelihood ratio.

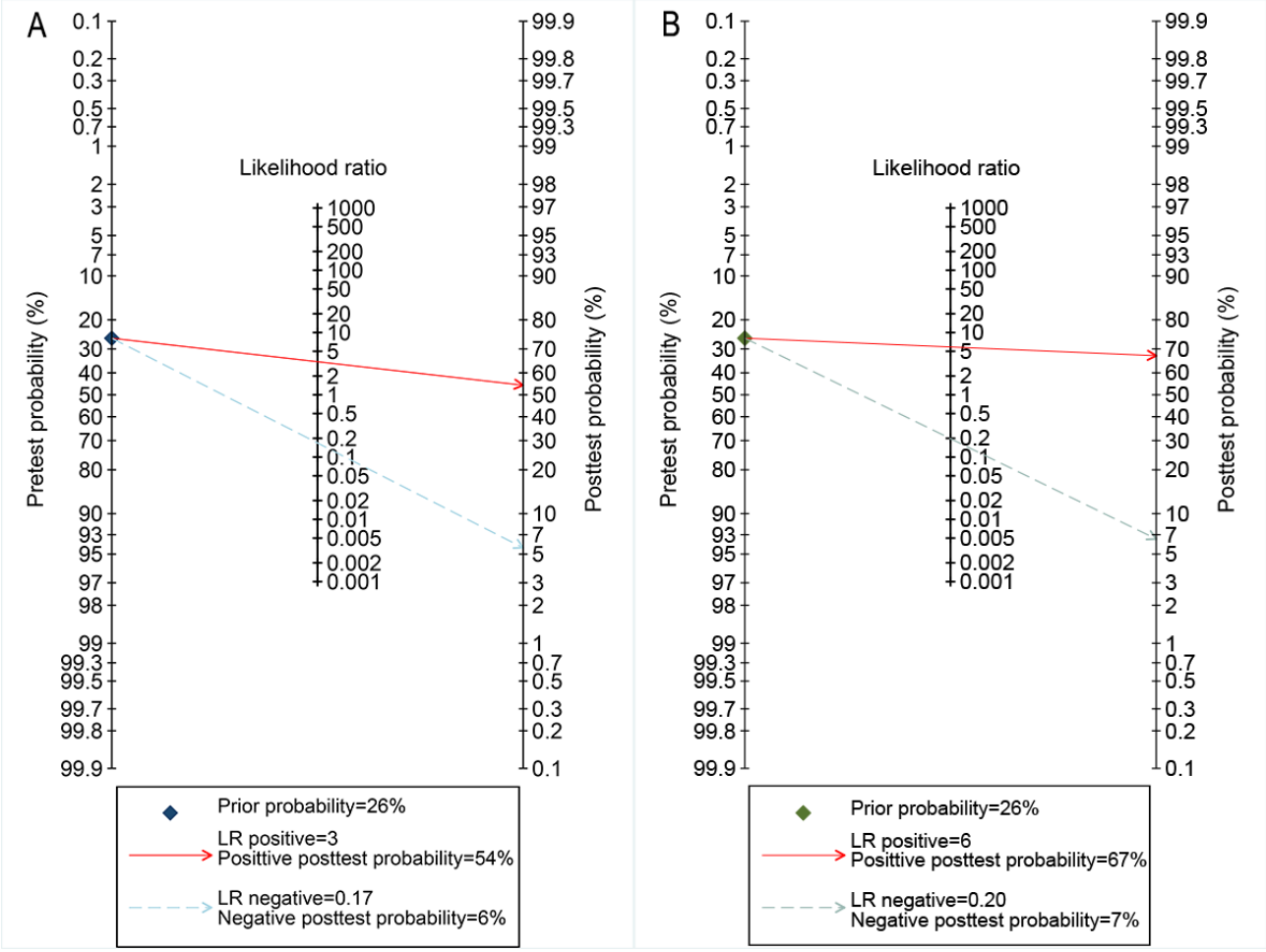

Discussion

Principal Findings

Our systematic review and meta-analysis demonstrated that DL algorithms using DBT achieved diagnostic proficiency comparable to the aggregate performance of radiologists and senior experts while exhibiting significantly superior sensitivity compared to junior radiologists. The observed parity between DL and senior radiologists suggests that current computational models have attained a level of pattern recognition equivalent to seasoned clinical judgment. However, the fact that DL has not significantly superseded senior radiologists implies that, while algorithms facilitate standardization, they currently lack the nuance required to outperform complex human decision-making [36-38]. Conversely, the significant performance gap regarding junior practitioners likely stems from the algorithms' robust capacity to systematically analyze vast volumes of annotated data, enabling the detection of subtle or occult lesions that may be obscured by the subjectivity or limited pattern recognition characteristic of early-career experience [39]. By minimizing the specific oversight errors associated with inexperience [40], DL models effectively function as a "safety net" enhancing diagnostic sensitivity. Consequently, these findings support the integration of DL not as a stand-alone replacement for expert review but as a vital adjunctive tool to augment the capabilities of less experienced radiologists and bridge diagnostic disparities in resource-constrained environments.

Regarding clinical impact metrics, our initial analysis showed that DL algorithms had lower PPV and recall rates than junior radiologists. However, these findings should be interpreted with caution as they likely reflect spectrum bias and the limited number of studies available for the junior subgroup (2/13, 15.4%). Specifically, the junior radiologist data were derived from highly enriched cohorts, with a malignancy prevalence of 70% in the study by Chae et al [35] and 57% positive findings in the study by Bassi et al [10]. As PPV is a prevalence-dependent metric, the exceptionally high disease burden in these datasets artificially inflated the pooled estimates for the junior subgroup, reflecting fundamental differences in the underlying test populations rather than intrinsic radiologist superiority.

More importantly, an essential clinical counterpoint to the sensitivity advantage of DL algorithms is the significantly lower pooled PPV of stand-alone DL algorithms (0.41) compared to all radiologists (0.62) and senior radiologists (0.90). A PPV of 0.41 implies that less than half of the cases flagged by the DL algorithms are truly malignant, corresponding to a substantial FP burden. This finding is particularly contradictory as one of the primary, well-documented advantages of DBT over traditional mammography is its ability to reduce unnecessary

XSL•FO
RenderX

recall rates. Using low-PPV DL algorithms as a primary or concurrent reading tool could erode this advantage, leading to a cascade of adverse consequences. These include unnecessary biopsies, heightened patient psychological distress and anxiety, and the overconsumption of health care resources—costs that are thoroughly documented in breast screening literature. Therefore, we urge a cautious approach to claims advocating for current DL algorithms as universal standardization tools. Their integration should be highly context dependent: in settings experiencing a severe shortage of experienced radiologists, the sensitivity benefits may offset the FP costs; however, in fully resourced screening environments with available senior radiologists, the net clinical benefit of stand-alone DL remains uncertain. Before recommending these models as reliable standardization tools, future deployment frameworks must incorporate well-defined, acceptable safety thresholds for FP rates and prospectively evaluate downstream clinical and economic outcomes.

Interestingly, our analysis comparing radiologist performance with and without DL assistance revealed no statistically significant incremental benefit. This lack of synergistic enhancement likely results from a "ceiling effect," where experienced radiologists and high-performing algorithms achieve similar diagnostic performance, leaving limited room for improvement [41,42]. Furthermore, the clinical utility of DL assistants is constrained by "automation bias" and the opaque nature of "black box" algorithms [7,43]. When radiologist confidence in the algorithm is low—or when the AI serves merely as a concurrent reader without explainable features—the translation of algorithmic output into improved decision-making is diminished [44]. This suggests that, for DL to provide substantive added value, future systems must move beyond binary classification to provide interpretability and context-aware insights.

In comparison with a previous meta-analysis published by Yoon et al [45] in 2023, which reported that stand-alone AI significantly outperformed radiologists (AUC=0.90 vs 0.79) for breast cancer diagnosis, our study presents more conservative findings (AUC=0.89 vs 0.88; no significant difference). This discrepancy is attributable to the substantial expansion of the evidence base in our review. By incorporating approximately 3 times the number of studies, our analysis mitigates small-study effects that may have overestimated AI superiority in previous reviews. Our analysis suggests that the diagnostic performance of DL algorithms is currently comparable with the aggregate performance of the overall radiologist population in DBT interpretation but not demonstrably superior.

Beyond prior analyses, we assessed the specific impact of the DL assistant by comparing diagnostic performance without and with the DL assistant, offering critical insights into the human-AI interaction. We also incorporated essential clinical impact metrics such as detection rate, PPV, and recall rate to assess practical utility beyond discrimination accuracy. Furthermore, the methodological rigor was strengthened through the application of the PROBAST+AI tool for quality assessment and the GRADE approach to evaluate the certainty of evidence. These enhancements and stratified analyses collectively supply new, high-quality evidence that clarifies the current application of DL in breast imaging.

The extreme heterogeneity observed in several pooled metrics (I^2>98%) warrants careful interpretation. Although meta-regression identified the validation strategy (internal vs external) as a significant contributor to heterogeneity in specificity (P=.05), this single factor cannot account for the near-complete variance in metrics such as PPV and recall rate. We posit that the residual heterogeneity reflects fundamental differences inherent to the DL algorithms themselves. First, algorithm design thresholds vary considerably across vendors: some systems are deliberately calibrated toward high sensitivity for triage purposes [27], whereas others prioritize specificity to minimize FP recalls, resulting in divergent operating points along the ROC curve. Second, training set composition, including differences in malignancy prevalence, patient demographics, geographic regions, mammographic density distribution, and scanner hardware, introduces domain-specific biases that cannot be homogenized through statistical pooling alone [28]. Third, variability in FP tolerance thresholds across health care systems and screening programs further compounds interstudy differences [9]. Given this landscape, future studies should prioritize standardized reporting of operating thresholds and algorithm configurations to facilitate more meaningful comparisons.

Additionally, bivariate box plots pinpointed the studies by Shoshan et al [9] and Pinto et al [29] as distinct outliers. This heterogeneity is mechanistically explainable: Shoshan et al [9] optimized their algorithm for a "triage" workflow, deliberately sacrificing specificity to maximize sensitivity for ruling out normal cases, resulting in a skewed performance profile compared to standard diagnostic models. Meanwhile, the divergence in the study by Pinto et al [29] is attributable to a small sample size (N=190), which introduces substantial statistical instability and random variation into the results.

Looking ahead, most of the DL algorithms included in our study were confined to single-modality DBT data, lacking the contextual depth provided by supplementary imaging and clinical history. Future iterations must prioritize multimodal integration to synthesize comprehensive diagnostic insights by correlating findings with the patient's clinical background [46]. Implementation is further constrained by technical and systemic barriers, including data scarcity, regulatory challenges, and limited generalizability. While emerging techniques such as few-shot learning and self-supervised models may address these gaps, sustained multidisciplinary efforts are essential to optimize AI safety and deliver comprehensive solutions that genuinely augment radiological practice [47]. Currently, the opacity of decision-making processes remains a critical hurdle, highlighting the urgent need for future AI systems to prioritize transparency and interpretability [48].

Some limitations of this meta-analysis should be considered when interpreting the results. First, the preponderance of retrospective designs introduces potential selection bias, necessitating validation through large-scale prospective trials. Second, to distinguish between independent datasets and mitigate the risk of patient overlap, we extracted data exclusively

XSL•FO
RenderX

from the highest-performing DL algorithm in each study context. This methodological necessity may introduce an optimism bias, potentially leading to an overestimation of the average algorithmic performance. Third, the limited number of studies stratified by radiologist experience (particularly for juniors) constrains the statistical robustness of subgroup analyses. Fourth, because most included studies did not report complete contingency tables, we used the GetData software to redigitize published ROC curves and derived operating points using the Youden index. This approach introduces two layers of methodological limitation: (1) the manual redigitization process could introduce subjective measurement error, potentially introducing small but nonnegligible inaccuracies in the extracted sensitivity and specificity values; and (2) more critically, the Youden index–derived operating point represents a theoretical optimum rather than the actual clinical threshold used in each study's real-world setting. Clinically deployed algorithms may operate at thresholds deliberately chosen to balance sensitivity and specificity according to institutional recall policies or regulatory requirements. Consequently, our extracted performance values may not faithfully reflect the true clinical performance of these algorithms as implemented, and this methodological limitation should be considered when interpreting the pooled estimates.

Conclusions

In conclusion, DL algorithms for DBT demonstrated strong diagnostic performance. Although initial subgroup analyses indicated potentially higher sensitivity than junior radiologists, these findings are based on limited studies and require extensive validation to confirm their reliability. The current lack of a significant incremental benefit in human-AI collaborative workflows suggests that AI implementation should be approached cautiously—not as a stand-alone replacement or an automatic performance booster but rather as a supplementary "second opinion." Future research must prioritize the development of explainable AI and prospective multimodal studies to better define the true synergistic potential of human-machine collaboration in breast cancer screening.

Acknowledgments

During the preparation of this work, the authors used DeepSeek-V3.2 to improve readability and language quality. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

This research was funded by the National Natural Science Foundation of China (grant 8237153504).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: SL
Data curation: XP
Formal analysis: SL, ZW, YM, LW, XP
Investigation: ZW, YM, LW, XP
Methodology: ZW, YM, LW, XP
Software: SL
Writing—original draft: SL
Writing—review and editing: LW, XP
All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables and figures for the systematic review and meta-analysis.
[DOCX File, 3202 KB - jmir_v28i1e91659_app1.docx]

Checklist 1
PRISMA checklist.
[DOCX File, 20 KB - jmir_v28i1e91659_app2.docx]

References

1. Giaquinto AN, Sung H, Miller KD, et al. Breast cancer statistics, 2022. CA Cancer J Clin 2022 Nov;72(6):524-541. [doi: 10.3322/caac.21754] [Medline: 36190501]

XSL•FO
RenderX

2. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74(3):229-263. [doi: 10.3322/caac.21834] [Medline: 38572751]
3. Chong A, Weinstein SP, McDonald ES, Conant EF. Digital breast tomosynthesis: concepts and clinical practice. Radiology 2019 Jul;292(1):1-14. [doi: 10.1148/radiol.2019180760] [Medline: 31084476]
4. Gao Y, Moy L, Heller SL. Digital breast tomosynthesis: update on technology, evidence, and clinical practice. Radiographics 2021;41(2):321-337. [doi: 10.1148/rg.2021200101] [Medline: 33544665]
5. Raichand S, Blaya-Novakova V, Berber S, Livingstone A, Noguchi N, Houssami N. Digital breast tomosynthesis for breast cancer diagnosis in women with dense breasts and additional breast cancer risk factors: a systematic review. Breast 2024 Oct;77:103767. [doi: 10.1016/j.breast.2024.103767] [Medline: 38996609]
6. Wang R, Chen F, Chen H, et al. Deep learning in digital breast tomosynthesis: current status, challenges, and future trends. MedComm (2020) 2025 Jun;6(6):e70247. [doi: 10.1002/mco2.70247] [Medline: 40491967]
7. Bai J, Posner R, Wang T, Yang C, Nabavi S. Applying deep learning in digital breast tomosynthesis for automatic breast cancer detection: a review. Med Image Anal 2021 Jul;71:102049. [doi: 10.1016/j.media.2021.102049] [Medline: 33901993]
8. Balleyguier C, Arfi-Rouche J, Levy L, et al. Improving digital breast tomosynthesis reading time: a pilot multi-reader, multi-case study using concurrent Computer-Aided Detection (CAD). Eur J Radiol 2017 Dec;97:83-89. [doi: 10.1016/j.ejrad.2017.10.014] [Medline: 29153373]
9. Shoshan Y, Bakalo R, Gilboa-Solomon F, et al. Artificial intelligence for reducing workload in breast cancer screening with digital breast tomosynthesis. Radiology 2022 Apr;303(1):69-77. [doi: 10.1148/radiol.211105] [Medline: 35040677]
10. Bassi E, Russo A, Oliboni E, et al. The role of an artificial intelligence software in clinical senology: a mammography multi-reader study. Radiol Med 2024 Feb;129(2):202-210. [doi: 10.1007/s11547-023-01751-1] [Medline: 38082194]
11. Salameh JP, Bossuyt PM, McGrath TA, et al. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. BMJ 2020 Aug 14;370:m2632. [doi: 10.1136/bmj.m2632] [Medline: 32816740]
12. Moons KG, Damen JA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ 2025 Mar 24;388:e082505. [doi: 10.1136/bmj-2024-082505] [Medline: 40127903]
13. Gopalakrishna G, Mustafa RA, Davenport C, et al. Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but doable. J Clin Epidemiol 2014 Jul;67(7):760-768. [doi: 10.1016/j.jclinepi.2014.01.006] [Medline: 24725643]
14. Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MG, Stijnen T. Bivariate random effects meta-analysis of ROC curves. Med Decis Making 2008;28(5):621-638. [doi: 10.1177/0272989X08319957] [Medline: 18591542]
15. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002 Jun 15;21(11):1539-1558. [doi: 10.1002/sim.1186] [Medline: 12111919]
16. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol 2005 Sep;58(9):882-893. [doi: 10.1016/j.jclinepi.2005.01.016] [Medline: 16085191]
17. Morrell S, Hutel M, Lucena O, et al. Dynamic AI-assisted ipsilateral tissue matching for digital breast tomosynthesis. Eur J Radiol 2025 Nov;192:112356. [doi: 10.1016/j.ejrad.2025.112356] [Medline: 40925265]
18. Letter H, Peratikos M, Toledano A, et al. Use of artificial intelligence for digital breast tomosynthesis screening: a preliminary real-world experience. J Breast Imaging 2023 May 22;5(3):258-266. [doi: 10.1093/jbi/wbad015] [Medline: 38416890]
19. Uematsu T, Nakashima K, Harada TL, Nasu H, Igarashi T. Artificial intelligence computer-aided detection enhances synthesized mammograms: comparison with original digital mammograms alone and in combination with tomosynthesis images in an experimental setting. Breast Cancer 2023 Jan;30(1):46-55. [doi: 10.1007/s12282-022-01396-4] [Medline: 36001270]
20. Dahlblom V, Andersson I, Lång K, Tingberg A, Zackrisson S, Dustler M. Artificial intelligence detection of missed cancers at digital mammography that were detected at digital breast tomosynthesis. Radiol Artif Intell 2021;3(6):e200299. [doi: 10.1148/ryai.2021200299] [Medline: 34870215]
21. Uematsu T, Nakashima K, Harada TL, Nasu H, Igarashi T. Comparisons between artificial intelligence computer-aided detection synthesized mammograms and digital mammograms when used alone and in combination with tomosynthesis images in a virtual screening setting. Jpn J Radiol 2023 Jan;41(1):63-70. [doi: 10.1007/s11604-022-01327-5] [Medline: 36068450]
22. Larsen M, Aglen CF, Lee CI, et al. Artificial intelligence evaluation of 122 969 mammography examinations from a population-based screening program. Radiology 2022 Jun;303(3):502-511. [doi: 10.1148/radiol.212381] [Medline: 35348377]
23. Plimpton SR, Milch H, Sears C, et al. External validation of a commercial artificial intelligence algorithm on a diverse population for detection of false negative breast cancers. J Breast Imaging 2025 Jan 25;7(1):16-26. [doi: 10.1093/jbi/wbae058] [Medline: 39401322]

XSL•FO
RenderX

24. Dahlblom V, Dustler M, Zackrisson S, Tingberg A. Workload reduction of digital breast tomosynthesis screening using artificial intelligence and synthetic mammography: a simulation study. J Med Imaging (Bellingham) 2025 Nov;12(Suppl 2):S22005. [doi: 10.1117/1.JMI.12.S2.S22005] [Medline: 40313361]
25. Berti A, Scapicchio C, Iacconi C, et al. An explainable-by-design end-to-end AI framework based on prototypical part learning for lesion detection and classification in Digital Breast Tomosynthesis images. Comput Struct Biotechnol J 2025;27:2649-2660. [doi: 10.1016/j.csbj.2025.06.008] [Medline: 40599244]
26. van Winkel SL, Rodríguez-Ruiz A, Appelman L, et al. Impact of artificial intelligence support on accuracy and reading time in breast tomosynthesis image interpretation: a multi-reader multi-case study. Eur Radiol 2021 Nov;31(11):8682-8691. [doi: 10.1007/s00330-021-07992-w] [Medline: 33948701]
27. Resch D, Lo Gullo R, Teuwen J, et al. AI-enhanced mammography with digital breast tomosynthesis for breast cancer detection: clinical value and comparison with human performance. Radiol Imaging Cancer 2024 Jul;6(4):e230149. [doi: 10.1148/rycan.230149] [Medline: 38995172]
28. Raya-Povedano JL, Romero-Martín S, Elías-Cabot E, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M. AI-based strategies to reduce workload in breast cancer screening with mammography and tomosynthesis: a retrospective evaluation. Radiology 2021 Jul;300(1):57-65. [doi: 10.1148/radiol.2021203555] [Medline: 33944627]
29. Pinto MC, Rodriguez-Ruiz A, Pedersen K, et al. Impact of artificial intelligence decision support using deep learning on breast cancer screening interpretation with single-view wide-angle digital breast tomosynthesis. Radiology 2021 Sep;300(3):529-536. [doi: 10.1148/radiol.2021204432] [Medline: 34227882]
30. Park EK, Kwak SY, Lee W, Choi JS, Kooi T, Kim EK. Impact of AI for digital breast tomosynthesis on breast cancer detection and interpretation time. Radiol Artif Intell 2024 May;6(3):e230318. [doi: 10.1148/ryai.230318] [Medline: 38568095]
31. Kim JG, Haslam B, Diab AR, et al. Impact of a categorical AI system for digital breast tomosynthesis on breast cancer interpretation by both general radiologists and breast imaging specialists. Radiol Artif Intell 2024 Mar;6(2):e230137. [doi: 10.1148/ryai.230137] [Medline: 38323914]
32. Elías-Cabot E, Romero-Martín S, Raya-Povedano JL, Brehl AK, Álvarez-Benito M. Impact of real-life use of artificial intelligence as support for human reading in a population-based breast cancer screening program with mammography and tomosynthesis. Eur Radiol 2024 Jun;34(6):3958-3966. [doi: 10.1007/s00330-023-10426-4] [Medline: 37975920]
33. Conant EF, Toledano AY, Periaswamy S, et al. Improving accuracy and efficiency with concurrent use of artificial intelligence for digital breast tomosynthesis. Radiol Artif Intell 2019 Jul 31;1(4):e180096. [doi: 10.1148/ryai.2019180096] [Medline: 32076660]
34. Chen IE, Joines MM, Capiro N, et al. Commercial artificial intelligence versus radiologists: NPV and recall rate in large population-based digital mammography and tomosynthesis screening mammography cohorts. AJR Am J Roentgenol 2025 Dec;225(6):e2532889. [doi: 10.2214/AJR.25.32889] [Medline: 40899676]
35. Chae EY, Kim HH, Jeong JW, Chae SH, Lee S, Choi YW. Decrease in interpretation time for both novice and experienced readers using a concurrent computer-aided detection system for digital breast tomosynthesis. Eur Radiol 2019 May;29(5):2518-2525. [doi: 10.1007/s00330-018-5886-0] [Medline: 30547203]
36. Balkenende L, Teuwen J, Mann RM. Application of deep learning in breast cancer imaging. Semin Nucl Med 2022 Sep;52(5):584-596. [doi: 10.1053/j.semnuclmed.2022.02.003] [Medline: 35339259]
37. Abdullah KA, Marziali S, Nanaa M, Escudero Sánchez L, Payne NR, Gilbert FJ. Deep learning-based breast cancer diagnosis in breast MRI: systematic review and meta-analysis. Eur Radiol 2025 Aug;35(8):4474-4489. [doi: 10.1007/s00330-025-11406-6] [Medline: 39907762]
38. Witowski J, Heacock L, Reig B, et al. Improving breast cancer diagnostics with deep learning for MRI. Sci Transl Med 2022 Sep 28;14(664):eabo4802. [doi: 10.1126/scitranslmed.abo4802] [Medline: 36170446]
39. Gu Y, Xu W, Lin B, et al. Deep learning based on ultrasound images assists breast lesion diagnosis in China: a multicenter diagnostic study. Insights Imaging 2022 Jul 28;13(1):124. [doi: 10.1186/s13244-022-01259-8] [Medline: 35900608]
40. Lotter W, Diab AR, Haslam B, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. Nat Med 2021 Feb;27(2):244-249. [doi: 10.1038/s41591-020-01174-9] [Medline: 33432172]
41. Xue P, Wang J, Qin D, et al. Deep learning in image-based breast and cervical cancer detection: a systematic review and meta-analysis. NPJ Digit Med 2022 Feb 15;5(1):19. [doi: 10.1038/s41746-022-00559-z] [Medline: 35169217]
42. Tan H, Wu Q, Wu Y, et al. Mammography-based artificial intelligence for breast cancer detection, diagnosis, and BI-RADS categorization using multi-view and multi-level convolutional neural networks. Insights Imaging 2025 May 21;16(1):109. [doi: 10.1186/s13244-025-01983-x] [Medline: 40397242]
43. Galati F, Maroncelli R, De Nardo C, et al. Deep learning with transfer learning on digital breast tomosynthesis: a radiomics-based model for predicting breast cancer risk. Diagnostics (Basel) 2025 Jun 26;15(13):1631. [doi: 10.3390/diagnostics15131631] [Medline: 40647630]
44. Dan Q, Xu Z, Burrows H, Bissram J, Stringer JS, Li Y. Diagnostic performance of deep learning in ultrasound diagnosis of breast cancer: a systematic review. NPJ Precis Oncol 2024 Jan 27;8:21. [doi: 10.1038/s41698-024-00514-z]

XSL•FO
RenderX

45. Yoon JH, Strand F, Baltzer PAT, et al. Standalone AI for breast cancer detection at screening digital mammography and digital breast tomosynthesis: a systematic review and meta-analysis. Radiology 2023 Jun;307(5):e222639. [doi: 10.1148/radiol.222639] [Medline: 37219445]
46. Rao VM, Hla M, Moor M, et al. Multimodal generative AI for medical image interpretation. Nature 2025 Mar;639(8056):888-896. [doi: 10.1038/s41586-025-08675-y] [Medline: 40140592]
47. Sosna J, Joskowicz L, Saban M. Navigating the AI landscape in medical imaging: a critical analysis of technologies, implementation, and implications. Radiology 2025 Jun;315(3):e240982. [doi: 10.1148/radiol.240982] [Medline: 40552997]
48. Champendal M, Müller H, Prior JO, Dos Reis CS. A scoping review of interpretability and explainability concerning artificial intelligence methods in medical imaging. Eur J Radiol 2023 Dec;169:111159. [doi: 10.1016/j.ejrad.2023.111159] [Medline: 37976760]

Abbreviations

AI: artificial intelligence
AUC: area under the receiver operating characteristic curve
DBT: digital breast tomosynthesis
DL: deep learning
FN: false negative
FP: false positive
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
MeSH: Medical Subject Headings
PPV: positive predictive value
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Diagnostic Test Accuracy
PROBAST+AI: Prediction Model Risk of Bias Assessment Tool+Artificial Intelligence
ROC: receiver operating characteristic
TN: true negative
TP: true positive

Edited by M Balcarras; submitted 18.Jan.2026; peer-reviewed by F Sheida, Z Yu; revised version received 26.Mar.2026; accepted 26.Mar.2026; published 06.May.2026.

Please cite as:
Lyu S, Wang Z, Mu Y, Wang L, Pei X
Deep Learning Algorithms Versus Radiologists in Digital Breast Tomosynthesis for Breast Cancer Detection: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e91659
URL: https://www.jmir.org/2026/1/e91659
doi:10.2196/91659

© Shewen Lyu, Zepeng Wang, Yujing Mu, Luyao Wang, Xiaohua Pei. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Multidimensional Effects of Telemedicine on Patients With Spinal Cord Injury: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Wen Zhong[1,2*], MM; Liyi Huang[1,2*], PhD; Xin Sun[1,2], MSc; Rui Liu[1,2], MSc; Lu Wang[1,2], MD; Quan Wei[1,2], MD, PhD

[1]Rehabilitation Medicine Center and Institute of Rehabilitation Medicine, West China Hospital, Sichuan University, Wuhou District, No.37 Guoxue Alley, Chengdu, China

[2]Key Laboratory of Rehabilitation Medicine in Sichuan Province, West China Hospital, Sichuan University, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Quan Wei, MD, PhD
Rehabilitation Medicine Center and Institute of Rehabilitation Medicine, West China Hospital, Sichuan University, Wuhou District, No.37 Guoxue Alley, Chengdu, China

Abstract

Background: Spinal cord injury (SCI) causes persistent physical and psychological impairments and is associated with reduced quality of life. Telemedicine may improve rehabilitation access and follow-up care, but its effectiveness across multiple outcome domains in SCI remains uncertain.

Objective: This study aimed to evaluate the effects of telemedicine interventions on psychological health, quality of life, sleep, functional independence, and participation, and pain intensity in individuals with SCI.

Methods: We searched PubMed, Web of Science, Embase, Ovid MEDLINE, and Cochrane CENTRAL until 17 February 2026. We included English-language randomized controlled trials (RCTs) of telemedicine interventions in individuals with SCI. Two reviewers independently screened studies, extracted data, and assessed risk of bias using the Risk of Bias 2 (RoB 2; Cochrane) tool. Random-effects meta-analyses used the Hartung-Knapp-Sidik-Jonkman method with restricted maximum likelihood estimation of between-study variance. Effects were summarized as standardized mean differences (SMD) or mean differences (MD) with 95% CIs. For main meta-analyses, 95% prediction intervals were reported when at least 5 studies were available, but not for analyses with fewer than 5 studies or for subgroup meta-analyses. Certainty of evidence was assessed using GRADE (Grading of Recommendations Assessment, Development, and Evaluation).

Results: We included 33 studies (35 reports). Telemedicine improved the World Health Organization Quality of Life-BREF (WHOQOL-BREF) social domain (MD 3.27, 95% CI 0.64 to 5.89; P=.03) and sleep quality at 3 months (MD −2.24, 95% CI −3.82 to −0.67; P=.04). Depressive symptoms also improved in the >3-≤6 months follow-up subgroup (SMD −0.31, 95% CI −0.57 to −0.04; P=.03). Overall effects for depressive symptoms were not significant (SMD −0.11, 95% CI −0.26 to 0.05; prediction interval −0.37 to 0.15; P=.16; I^2=36.3%), while findings for anxiety, other WHOQOL-BREF domains, sleep quality at 1 month, functional outcomes, and pain intensity generally favored telemedicine but did not reach statistical significance. Approximately half of the studies were rated as low risk overall on RoB 2, with most remaining studies rated as having some concerns and a smaller subset rated as high risk. GRADE certainty was high for the >3-≤6-month depressive-symptoms subgroup, moderate for the WHOQOL-BREF social domain, Pittsburgh Sleep Quality Index (PSQI), and Spinal Cord Independence Measure (SCIM), and low for depressive symptoms overall, anxiety, and pain intensity.

Conclusions: Telemedicine may improve selected outcomes in SCI, with the most consistent evidence for social aspects of quality of life, sleep after sustained intervention exposure, and a more favorable effect on depressive symptoms in midterm follow-up subgroup analyses. These results suggest telemedicine as a practical adjunct for extending SCI rehabilitation access and continuity. Further trials should focus on optimizing intervention components, intensity, and patient targeting.

Trial Registration: PROSPERO CRD420251149977; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251149977

(J Med Internet Res 2026;28:e87088) doi:10.2196/87088

KEYWORDS

spinal cord injury; telemedicine; meta-analysis; randomized controlled trials; mental health; quality of life; sleep; functional independence; pain

XSL•FO
RenderX

Introduction

Spinal cord injury (SCI) refers to damage to the spinal cord resulting from various causes, leading to temporary or permanent neurological dysfunction [1,2]. The etiology of SCI can be broadly classified into traumatic and nontraumatic categories. Traumatic SCI is typically caused by external physical impacts that induce acute injury, whereas nontraumatic SCI results from acute or chronic diseases that compromise the structure and function of the spinal cord. In traumatic SCI, both primary injury and subsequent secondary injury profoundly alter spinal cord tissue structure, ultimately leading to irreversible neurological deficits. According to the 2021 Global Burden of Disease study, there were an estimated 15,400,682 prevalent cases of SCI worldwide, with the majority occurring among working-age individuals (20 - 69 years) [3]. The global incidence is estimated at 23.77 cases per million population, with traumatic SCI accounting for 26.48 cases per million and nontraumatic SCI for 17.93 cases per million [4]. SCI not only results in severe motor and sensory impairments but is also frequently complicated by pulmonary infections, deep vein thrombosis, pressure ulcers, and joint contractures, imposing devastating effects on patients' physical health, social participation, and occupational capacity [5]. Consequently, patients with SCI often experience varying degrees of loss of independence and face an increased risk of mortality.

Telemedicine has emerged as one of the most significant innovations in modern health care, offering new opportunities for the long-term management and rehabilitation of patients with SCI. Given the challenges of limited mobility and restricted access to specialized rehabilitation services, telemedicine is particularly well-suited to this population. Telemedicine services can be delivered via telephone, videoconferencing, computer-based platforms, and robotic-assisted audio or video communication. Based on the target of service delivery, telemedicine can be classified into provider-to-provider (health care professionals to other clinicians or caregivers), direct-to-patient (health care professionals to patients), store-and-forward (transfer of health information to recipients), web-based educational interventions, and interactive home monitoring [6]. Evidence suggests that telemedicine may provide substantial benefits in self-management, psychological well-being, complication prevention, pain control, and functional recovery among patients with SCI. However, the heterogeneity in study objectives, intervention modalities, and outcome measures has limited the identification of optimal strategies and robust conclusions, thereby constraining the broader application of telemedicine in SCI care.

Several systematic reviews have attempted to evaluate the effectiveness of telemedicine in SCI. For instance, a meta-analysis published in 2022 demonstrated that telemedicine interventions significantly reduced the incidence of pressure injuries and improved Pressure Ulcer Scale for Healing (PUSH) scores among community-dwelling patients with SCI. Furthermore, network meta-analysis indicated that hybrid comprehensive telemedicine interventions were more effective than partial telemedicine, no telemedicine, or single-modality telemedicine. Nonetheless, as only four randomized controlled trials (RCTs) could be pooled for analyses of healing rates and PUSH scores, the robustness of these findings remains limited [7]. More recently, a meta-analysis published in 2025 reported that telemedicine significantly improved the quality of life and reduced depressive symptoms in patients with SCI. However, the quality-of-life analysis was characterized by high heterogeneity; for depression, although subgroup analyses were performed by follow-up duration and intervention type, no subgroup analysis was conducted based on different depression assessment scales. Additionally, data from multiple time points of the same trial were included as independent entries and counted toward the total sample size, which may have introduced bias in the overall estimates [8].

SCI not only causes profound neurological impairments but also gives rise to multiple complications that substantially affect patients' functional recovery, mental health, quality of life, and sleep. With an increasing number of RCTs investigating the potential benefits of telemedicine in SCI, it is crucial to clarify the extent to which patients may benefit from such interventions, thereby promoting their broader application in clinical practice. Therefore, the objective of this study was to systematically review and meta-analyze the effects of telemedicine interventions on psychological health, health-related quality of life, sleep-related functioning, functional independence and rehabilitation outcomes, and pain intensity in patients with SCI, aiming to provide robust evidence for clinical decision-making and to guide the optimization of tele-rehabilitation strategies in this population.

Methods

Protocol and Registration

The protocol for this systematic review was registered with the International Prospective Register of Systematic Reviews (PROSPERO; registration number: CRD420251149977).

Eligibility Criteria

Eligibility criteria were established based on the PICOS (Population, Intervention, Comparison, Outcome, and Study Design) framework. The study population included patients diagnosed with SCI, with no restrictions on sex, age, injury level, or severity. Eligible interventions were defined as telemedicine-related modalities, including video consultations, mobile health apps, remote monitoring, and telephone follow-up. Comparators were not restricted and could include usual care or no tele-intervention. Outcomes were prespecified in five domains, including psychological health (eg, depression and/or anxiety), health-related quality of life, sleep-related functioning, functional independence, and rehabilitation outcomes, and pain intensity. Trials contributed to meta-analysis only when outcome data for at least one prespecified domain were available and extractable; otherwise, they were included in the descriptive synthesis. Only RCTs were eligible. To ensure that comparisons reflected allocation to mode of care delivery rather than self-selection, we included only trials in which participants were randomized to receive a telemedicine-delivered intervention versus a comparator condition (eg, usual care, standard follow-up, or in-person care). Studies in which telemedicine exposure was determined primarily by participant preference,

XSL•FO
RenderX

acceptance of an offer, or other nonrandom allocation mechanisms were excluded. In line with the search strategy, studies were restricted to English-language full-text reports.

The exclusion criteria were as follows: (1) nonprimary research articles, including systematic reviews, narrative commentaries, conference abstracts, and case reports, and other non-RCT designs; (2) interventions not related to telemedicine; and (3) duplicate publications or studies with overlapping datasets. These exclusions were applied to ensure methodological rigor and extractable data.

Meta-analysis was performed when quantitative data were extractable; otherwise, findings were synthesized narratively. Trials not reporting any prespecified outcome domain were included in the descriptive summary (Table 1) but were not included in the quantitative synthesis. When outcomes were measured but insufficiently reported for extraction, authors were contacted where feasible, and missing results were documented.

XSL•FO
RenderX

Table . Baseline features and design elements of included studies.

Study	Country	Study design	Sample size	Sex (male), n (%)	Age (year)	Interventions	Duration	Outcomes	Follow-up
Arora et al (2017) [9] / Arora et al (2017) [10] (same RCT[a]; 2 reports)	India and Bangladesh	RCT[a]	INT[b]: 60; CON[c]: 60	INT:52 (87); CON: 54 (90)	INT: mean 35 (SD 11); CON: mean 36 (SD 12)	INT: weekly phone calls from a trained health-care professional; CON: usual care	12 weeks	Size of the pressure ulcer; PUSH[d]; Depth of PU[e]; Undermining distance of PU; Braden scale; HADS[f]; WHODAS[g]; EQ-5D-5L[h]; Self-rated health EQ-5D-VAS[i]; Participants' impression of PU status; Participants' confidence to manage PU; Clinician's impression of PU status; Participants' satisfaction; Self-report time for PU resolution; Reduction in the size of pressure ulcer; EQ-5D-5L; Cost of delivering the intervention; Productivity cost	12 weeks
Badr et al (2024) [11]	United States	RCT	INT: 32; CON: 31	INT: 29 (91); CON: 29 (94)	INT: mean 61.6 (SD 10.0); CON: mean 59.8 (SD 10.4)	INT: education about sleep-disordered breathing (SDB) with positive airway pressure (PAP), goal setting, troubleshooting, and motivational enhancement; CON: nondirective sleep education only	3 months	PAP use; PSQI[j]; CHART[k]; WHOQOL[l]; PHQ-9[m]; FFS[n]; ESS	1, 3, and 6 months

XSL•FO
RenderX

Study	Country	Study design	Sample size	Sex (male), n (%)	Age (year)	Interventions	Duration	Outcomes	Follow-up
Bombardier et al (2021) [12]	United States	Pilot RCT	INT: 7; CON: 8	INT: 5 (71); CON: 6 (75)	INT: mean 56 (SD 13); CON: mean 49 (SD 13)	INT: 16-session Diabetes Prevention Program delivered via telephone; CON: usual care	24 weeks	$VO_{2\ peak}$[o]; BMI; DXA; ISI[p]; PARA-SCI[q]; Brief Pain Inventory intensity and interference scales; Wheelchair User's Shoulder Pain Index; PHQ-9; SF-12[r]; WHOQOL-BREF[s]; Exercise self-efficacy	6 months
Bombardier et al (2023) [13]	United States	RCT	INT: 89; CON: 85	INT: 63 (71); CON: 70 (82)	INT: mean 47.8 (SD 15); CON: mean 47.5 (SD 14.8)	INT: assessment, medical care coordination, adherence support, outcome monitoring, and decision support, along with brief psychological interventions to the patients via up to 12 in-person or telephone sessions; CON: usual care	4 months	WHOQOL-BREF; NRS[t]; PHQ-9, HSCL-20B; LT-PAQ-SCI[u]	4 and 8 months
Burke et al (2019) [14]	Ireland	RCT	INT: 35; CON: 34	INT: 25 (71%); CON: 27 (79%)	INT: mean 50 (SD 12.3); CON: mean 52 (SD 13.8)	INT: 6 module cognitive behavioral therapy pain management programs (CBT - PMPs) weekly; CON: usual care	6 weeks	WHOQOL - BREF; ISCIPBDS[v]; DN4[w]; CPAQ - 8[x]; BPI[y]; HADS; PSQI	6 weeks and 3 months

XSL•FO
RenderX

Study	Country	Study design	Sample size	Sex (male), n (%)	Age (year)	Interventions	Duration	Outcomes	Follow-up
Carlson et al (2017) [15]	United States	RCT	INT: 83; CON: 87	INT: 72 (84.1); CON: 69 (85.7)	INT: mean 41.7 (SD 12.9); CON: mean 42.5 (SD 12.2)	INT: 12-month lifestyle-based treatment administered by health care professionals, largely via in-home visits and phone contacts; CON: usual care	12 months	MSPrIs[z]; SF-36[aa]; SWLS[ab]; Quasi-Adaptive Short Form for the Patient Reported Outcomes Measurement Information System (PROMIS) Version 1 Depression Item Bank; ISEL[ac]; Pressure Ulcer Knowledge Test; Adapted Moorong Self Efficacy Scale; Adapted Cut Down, Annoyed, Guilty, and Eye Opener (CAGE) Questionnaire	12 and 24 months
Chantanachai et al (2025) [16]	Thailand	RCT	INT[b]: 15; CON[c]: 15	INT: 10 (66.7) CON: 9 (60)	INT: mean 42.7 (SD 11.6); CON: mean 41.2 (SD 12.8)	INT: 20 min of active-tDCS, followed by a 1 h exercise program; CON: 20 min of sham-tDCS, followed by a 1 h exercise program, 12 sessions (3 times a week)	4 weeks	ISNCSCI[ad]; SCIM[ae]; TAI[af]; WHOQOL-BREF–Thai; H-reflex; m-MAS[ag]; Upper limb muscle strength test	1 month
Chemtob et al (2019) [17]	Canada	Pilot RCT	INT: 10; CON: 12	All: 16 (72.7)	All: mean 51.64 (SD 12.13)	INT: 1 h counseling session per week via an online video-chat platform; CON: regular routine	8 weeks	SDT[ah]; LT-PA[ai]; MV-PA[aj]; Life Satisfaction Questionnaire-11; PHQ-9; Meaning Questionnaire	6 and 10 weeks

XSL•FO
RenderX

Study	Country	Study design	Sample size	Sex (male), n (%)	Age (year)	Interventions	Duration	Outcomes	Follow-up
Coulter et al (2017) [18]	United Kingdom	Pilot RCT	INT: 15; CON: 8	INT: 9 (56); CON: 5 (63)	INT: mean 51.5 (SD 13.0); CON: mean 48.1 (SD 10.6)	INT: web-based physiotherapy twice per week; CON: usual care	8 weeks	6MPT[ak], 6MWT[al]; work HR–resting HR; Borg RPE; HADS; WHO-QOL BREF	8 weeks
Dallolio et al (2008) [19]	Belgium, Italy, and United Kingdom	RCT	INT: 69; CON: 68	All: 107 (84.25)	All: median 40 (IQR 18-85)	INT: telemedicine weekly sessions in the first 2 months, followed by biweekly telemedicine sessions for 4 months; CON: standard care	6 months	FIM[am]; SCIM; clinical outcomes; patient satisfaction questionnaire	2 and 6 months
Dorstyn et al (2012) [20]	Australia	RCT	INT: 20; CON: 19	INT: 13 (65); CON: 14 (74)	INT: mean 53.8 (SD 16.3); CON: mean 53.1 (SD 20.0)	INT: Seven telecounseling sessions were delivered over 12 weeks; CON: standard care	12 weeks	DASS21[an]; M.I.N.I.[ao]; SCL CSQ[ap]; SCL EWB[aq]; MDSS[ar]; Cost-effectiveness	12 weeks and 3 months
Dorstyn et al (2019) [21]	Australia	Pilot RCT	INT: 25; CON: 23	INT: 12 (48); CON: 15 (65.2)	INT: mean 43.0 (SD 10.9); CON: mean 40.7 (SD 11.0)	INT: provided job-searching and career planning information through text, videos, and interactive activities; CON: wait-list control	4 weeks	JSES[as]; LOT-R[at]; PHQ-9	4 weeks
Hearn et al (2018) [22]	United Kingdom	RCT	INT: 36; CON: 31	INT: 17 (47); CON: 14 (45)	INT: mean 43.8 (SD 8.7); CON: mean 45.2 (SD 12.2)	INT: online mindfulness intervention; CON: internet-delivered psychoeducation	8 weeks	HADS; WHOQoL-BREF; FFMQ25[au]; NRS; PCS[av], Retention rates	8 weeks and 3 months

XSL•FO
RenderX

Study	Country	Study design	Sample size	Sex (male), n (%)	Age (year)	Interventions	Duration	Outcomes	Follow-up
Hossain et al (2016) [23]	Bangladesh	Pilot RCT	INT: 15; CON: 15	INT: 13 (86.7); CON: 13 (86.7)	INT: median 29 (IQR 24-35); CON: median 34 (IQR 23-36)	INT: regular telephone contact and 3 home visits over 2 years; CON: usual care	2 years	all-cause mortality; SCI-SCS[aw]; Presence of pressure ulcers; CES Depression Scale[ax]; SF12; PUSH; SCIM; WHODAS	2 years
Hossain et al (2021) [24]	Bangladesh	RCT	INT: 204; CON: 206	INT: 181 (89); CON: 188 (91)	INT: median 31.4 (IQR 24.5 - 41.0); CON: median 33.4 (IQR 25.7 - 45.0)	INT: 36 phone calls and 3 home visits over the first 2 years following discharge; CON: usual care	2 years	Death; SCI-SCS; PUSH; CESD-R[ay]; WHODAS; SF12; SCIM	2 years
Houlihan et al (2017) [25]	USA	RCT	INT: 42; CON: 42	INT: 30 (71.4); CON: 32 (76.2)	INT: mean 44.0 (SD 13.4); CON: mean 47.5 (SD 14.4)	INT: 42 experimental subjects on a tapered call schedule; CON: usual care	6 months	PAM[az]; Social and Role Activities Limitations; Quality of life; Communication With Physicians Scale; Patient Satisfaction Scale; Global rating of change	2, 4, and 6 months
Irgens et al (2022) [26]/ Irgens et al 2024 [27] (same RCT; reports)	Norway	RCT	INT: 28; CON: 28	INT: 24 (86); CON: 21 (78)	INT: mean 58 (SD 14); CON: mean 58 (SD 13)	INT: additional videoconference (VCG); CON: regular care	1 year	QALYs[ba]; HRQoL[bb]; Cost measurement; environmental analysis; Michelin Travel's Route Planner; PI healing[bc], Interaction, Satisfaction, and Safety	1 year

XSL•FO
RenderX

Study	Country	Study design	Sample size	Sex (male), n (%)	Age (year)	Interventions	Duration	Outcomes	Follow-up
Kowalczewski et al (2011) [28]	Canada	RCT (crossover design)	INT: 7; CON: 6	All: 7 (54)	All: mean 35.92 (SD 11.96)	INT: "Re-Joyce ET" comprised FES-ET on a workstation, tele-supervised 1 h/d, 5 d/week; CON: conventional exercise therapy, tele-supervised 1 h/d, 5 d/week	6 weeks	ARAT[bd]; grasp and pinch forces; RAHFT[be]	6, 16, and 30 weeks
Kryger et al (2019) [29]	United States	RCT	INT: 19; CON: 19	INT: 13 (68); CON: 12 (63)	INT: mean 37.9 (SD 13.4); CON: mean 44.1 (SD 15.3)	INT: 30 min of training to use the Interactive Mobile Health and Rehabilitation (iMHere) app; CON: standard care	9 months	UTIs[bf]; Number of pressure injuries; ED visits[bg]; Number of hospitalizations; COPM[bh]; Adolescent Self-Management and Independence Scale; BDI-II[bi]; Patient Assessment of Chronic Illness Care; WHOQOL; Craig Handicap Assessment and Reporting Technique Short Form	3, 6, and 9 months
Lawrason et al (2023) [30]	Canada	RCT	INT: 12; CON: 9	INT: 7 (58.3); CON: 5 (55.6)	INT: mean 37.15 (SD 10.12); CON: mean 43.37 (SD 12.67)	INT: SCI Step Together, features weekly programming; CON: continue normal daily activity	8 weeks	LTPA; Quality participation	4 and 8 weeks
Li et al (2021) [31]	China	RCT	INT: 39; CON: 39	All: 40 (51.3)	All: mean 47.12 (SD 10.90)	INT: continuous care via the WeChat platform; CON: regular care	6 months	FAD[bj]; HPLPII[bk]	3 and 6 months
Li et al (2021) [32]	China	RCT	INT: 40; CON: 40	INT:28 (70); CON: 26 (65)	INT: median 57.8 (IQR 23 - 72); CON: median 59.1 (IQR 17 - 69)	INT: online home nursing care; CON: routine discharge guidance	12 months	Complications; ODI[bl]; SF-36	6 and 12 months

XSL•FO
RenderX

Study	Country	Study design	Sample size	Sex (male), n (%)	Age (year)	Interventions	Duration	Outcomes	Follow-up
Liu et al (2021) [33]	China	RCT	INT: 49; CON: 49	INT: 41 (84); CON: 40 (82)	INT: mean 40.37 (SD 12.18); CON: mean 43.06 (SD 12.06)	INT: 5 follow-ups conducted by trained nurses through the app, which had 4 core functions; CON: a routine telephone follow-up	12 weeks	MSES[bm]; SF-36	12 and 24 weeks
Liu et al (2023) [34]	China	RCT	INT: 49; CON: 49	INT: 41 (83.7); CON: 40 (81.6)	INT: mean 40.37 (SD 12.18); CON: mean 43.06 (SD 12.06)	INT: 5 sessions nurse-led multidisciplinary team via APP; CON: one routine telephone counseling	12 weeks	BDI-II	12 and 24 weeks
Mackelprang et al (2016) [35]	United States	RCT	INT: 85; CON: 83	INT: 70 (82); CON: 63 (76)	INT: mean 40.4 (SD 15.7); CON: mean 42.0 (SD 16.0)	INT: eleven 30-45-minute scheduled telephone calls to provide education, resources, and support; CON: usual care	10 months	health care use and medical complications composite; PHQ-9; Delighted-Terrible Scale; EuroQol thermometer; Subjective health using a single item from the SF-36; CHART-SF[bn]	3, 6, 9, and 12 months
Mercier et al (2015) [36]	United States	Pilot RCT	INT: 53; CON: 53	INT: 42 (79.2); CON: 34 (64.2)	INT: mean 45.8 (SD 12.1); CON: mean 45.0 (SD 14.0)	INT: CareCall intervention of regular automated phone calls that delivered educational content and peer and clinical expert perspectives; CON: usual care and the resource book	6 months	PHQ-9; PUSH; Health care use; CHART-SF; Engaging with the CareCall system	6 months
Migliorini et al (2016) [37]	Australia	RCT	INT: 34; CON: 25	INT: 25 (74); CON: 17 (68)	INT: mean 47.5 (SD 12.2); CON: mean 52.8 (SD 12.9)	INT: Electronic Personal Administration of Cognitive Therapy (ePACT); CON: wait-list control telephone interviews	10 - 12 weeks	DASS21; Personal Well-being Index–Adult; SCL EWL[bo];	3 and 6 months

XSL•FO
RenderX

Study	Country	Study design	Sample size	Sex (male), n (%)	Age (year)	Interventions	Duration	Outcomes	Follow-up
Pain et al (2007) [38]	UK, Italy, and Belgium	RCT	INT: 38; CON: 39	All: 64 (82.8)	All: median 42 (IQR 19-86)	INT: Internet-based video-link technology; CON: standard support	6 months	WHO QoL-100, HADS	2 and 6 months
Rintala et al (2008) [39]	United States	RCT	INT 1: 20; CON 1: 11; CON 2: 10	INT: 20 (100); CON 1: 11 (100); CON 2: 10 (100)	INT 1: mean 54.82 (SD 10.85); CON 1: mean 50.64 (SD 8.83); CON 2: mean 53.86 (SD 13.63)	INT 1: pressure ulcer education and monthly structured telephone follow-up; CON 1: monthly mail or telephone follow-up without educational content; CON 2: quarterly mail or telephone follow-up without educational content	24 months	Time to pressure ulcer recurrence;	24 months
Swarnakar et al (2023) [40]	India	RCT	INT: 15; CON: 15	INT: 12 (80); CON: 14 (93.3)	INT: mean 28.2 (SD 6.9); CON: mean 26.3 (SD 7.7)	INT: biweekly telerehabilitation sessions; CON: usual care	8 weeks	SCIM; CAS[bp]	4 and 8 weeks
Worobey et al (2018) [41]	United States	RCT	INT 1: 10; INT 2: 39; CON: 22	INT 1: 6 (60); INT 2: 31 (79.5); CON: 21 (95.5)	INT 1: median 53.5 (IQR 47.0 - 55.0); INT 2: median 54.0 (IQR 45.0 - 61.0) CON: median 56.0 (IQR 54.0 - 58.0)	INT 1: Individualized, in-person transfer training session; INT 2: a web-based transfer training module; CON: web training at their follow-up visit	6 months	TAI	2, 6, and 12 months

XSL•FO
RenderX

Study	Country	Study design	Sample size	Sex (male), n (%)	Age (year)	Interventions	Duration	Outcomes	Follow-up
Jaglal et al (2026) [42]	Canada	Pilot RCT	INT: 31; CON: 34	INT: 17 (57); CON: 13 (43)	INT: mean 49.6 (SD 11.1); CON: mean 48.7 (SD 14.1)	INT: web-based peer health-coaching self-management program: up to 14×1 h online video coaching sessions, goal setting, and action planning, and sortable resource library; CON: usual care	6 months	self-management skills (Health Education Impact Questionnaire et al); rehospitalization days; secondary conditions, self-efficacy; HRQoL; social and role limitations etc	6 and 12 months

XSL•FO
RenderX

Study	Country	Study design	Sample size	Sex (male), n (%)	Age (year)	Interventions	Duration	Outcomes	Follow-up
Young et al (2025) [43]	United States	RCT	INT 1: 12; INT 2: 12; CON (AC): 12	INT 1: 6 (50.0); INT 2: 7 (58.3); CON: 4 (33.3)	INT 1: mean 42.0 (SD 11.6); INT 2: mean 47.7 (SD 11.8); CON: mean 46.4 (SD 12.0)	INT 1: movement-to-music teleexercise via prerecorded videos 3×/week; INT 2: standard exercise training prerecorded videos 3×/week; CON: weekly educational articles (attention control)	8 weeks	feasibility (recruitment and retention), usability ([bq]SUS, [br]Health-ITUES + interviews), acceptability ([bs]PACES-8 + interviews); preliminary outcomes: [bt]LTPAQ-SCI, [bu]PROMIS short forms	8, 12, and 16 weeks

[a]RCT: randomized clinical trial.
[b]INT: intervention.
[c]CON: control.
[d]PUSH: Pressure Ulcer Scale for Healing.
[e]PU: pressure ulcer.
[f]HADS: Hospital Anxiety and Depression Scale.
[g]WHODAS: World Health Organization Disability Assessment Schedule.
[h]EQ-5D-5L: Euro Quality of Life 5-dimensional 5-level.
[i]VAS: Visual Analog Scale.
[j]PSQI: Pittsburgh Sleep Quality Index.
[k]CHART: Craig Handicap Assessment and Reporting Technique.
[l]WHOQOL: World Health Organization Quality-of-Life Scale
[m]PHQ-9: Patient Health Questionnaire-9.
[n]FFS: Flinders Fatigue Scale.
[o]$VO_{2\ peak}$: peak oxygen consumption.
[p]ISI: Insulin sensitivity index.
[q]PARA-SCI: Physical Activity Recall Assessment for SCI.
[r]SF-12: Medical Outcomes Study Short Form-12.
[s]WHOQOL-BREF: World Health Organization Quality of Life-BREF Scale.
[t]NRS: numerical rating scales.
[u]LTPAQ-SCI: Leisure Time Physical Activity Questionnaire for SCI.
[v]ISCIPBDS: International Spinal Cord Injury Pain Basic Data Set.
[w]DN4: Douleur Neuropathique en 4 Questions.
[x]CPAQ - 8: Chronic Pain Acceptance Questionnaire - 8.
[y]BPI: Brief Pain Inventory.
[z]MSPrIs: Medically serious pressure injuries.
[aa]SF-36: 36-item short-form health survey.
[ab]SWLS: Satisfaction with Life Scale.
[ac]ISEL: Interpersonal Support Evaluation List.
[ad]ISNCSCI: Neurological Classification of Spinal Cord Injury.
[ae]SCIM: spinal cord independence measure.
[af]TAI: transfer assessment instrument version 4.0.
[ag]m-MAS: modified-Modified Ashworth Score.
[ah]SDT: self-determination theory.
[ai]LTPA: leisure-time physical activity.
[aj]MVPA: moderate to vigorous LTPA.
[ak]6MPT: 6-minute Push Test.

XSL•FO
RenderX

[al]6MWT: 6-minute Walk Test.
[am]FIM: Functional Independence Measure.
[an]DASS-21: Depression Anxiety Stress Scale-21.
[ao]MINI: Mini International Neuropsychiatric Interview.
[ap]SCL CSQ: Spinal Cord Lesion Coping Strategies Questionnaires.
[aq]SCL EWB: Spinal Cord Lesion Emotional Well-Being Questionnaire.
[ar]MDSS: Multidimensional Measure of Social Support.
[as]JSES: Job Procurement Self-Efficacy Scale.
[at]LOT-R: Life Orientation Test-Revised.
[au]FFMQ25: Five Facet Mindfulness Questionnaire.
[av]PCS: The Pain Catastrophizing Scale.
[aw]SCI-SCS: Spinal Cord Injury Secondary Conditions Scale.
[ax]CES Depression Scale: Center for Epidemiologic Studies Depression Scale.
[ay]CESD-R: Center for Epidemiologic Studies Depression Scale revised version.
[az]PAM: Patient Activation Measure.
[ba]QALYs: quality-adjusted life years.
[bb]HRQoL: health-related quality of life.
[bc]PI healing: Pressure Injuries healing.
[bd]ARAT: Action Research Arm Test.
[be]RAHFT: ReJoyce automated hand function test.
[bf]UTIs: urinary tract infections.
[bg]ED visit: emergency department visit.
[bh]COPM: Canadian Occupational Performance Measure.
[bi]BDI-II: Beck Depression Inventory-II.
[bj]FAD: Family Assessment Device.
[bk]HPLPII: Health Promoting Lifestyle Profile.
[bl]ODI: Oswestry Disability Index.
[bm]MSES: Moorong Self-Efficacy Scale.
[bn]CHART-SF: Craig Handicap Assessment and Reporting Technique Short Form.
[bo]SCL EWL: Spinal Cord Lesion Emotional Well-being Questionnaire version 1 Australia.
[bp]CAS: Coronavirus Anxiety Scale.
[bq]SUS: System Usability Scale.
[br]Health-ITUES: Health Information Technology Usability Evaluation Scale.
[bs]PACES-8: Physical Activity Enjoyment Scale–8 item version.
[bt]LTPAQ-SCI: Leisure Time Physical Activity Questionnaire for People with Spinal Cord Injury.
[bu]PROMIS: Patient-Reported Outcomes Measurement Information System.
[bv]ESS: Epworth Sleepiness Scale.

Information Sources

To ensure comprehensive identification of relevant studies, we searched PubMed (via the NLM interface), Web of Science Core Collection (via Clarivate), Embase (via Ovid), MEDLINE (via Ovid), and the Cochrane Central Register of Controlled Trials (CENTRAL; via the Cochrane Library). The final search update was performed on February 17, 2026. In addition to database searching, we also searched trial registries and conducted supplementary searches (citation searching, websites and search engines, and hand searching). We additionally performed backward and forward citation searching using the reference lists of included studies and relevant reviews, and the "cited reference" function in Web of Science. Searches were restricted to English-language full-text reports. The search methods and reporting followed the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search) extension [44].

Search Strategy

A comprehensive strategy was developed by combining controlled vocabulary (eg, MeSH and Emtree terms) with free-text terms and adapting syntax to each database. Search concepts included (1) spinal cord injury and related terms, (2) telemedicine-related modalities (including telerehabilitation, telerehab, and closely related telehealth and telemedicine terms), and (3) an RCT component. To avoid overly broad retrieval, broad digital-delivery terms, such as online, telephone, app-based, web-based, and internet-based terms, were combined with rehabilitation- or therapy-related terms. An RCT filter was applied where appropriate, and animal-only records were excluded. The full, reproducible search strategies for all databases, copied exactly as run (including filters and limits), are provided in Table S1 in Multimedia Appendix 1. No natural language processing or text-mining tools were used to generate or refine search terms. No automated search translation tools were used; search strings were adapted manually to each

XSL•FO
RenderX

database. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram has been updated accordingly (Checklist 1). All records were imported into EndNote for record management and deduplication before screening. The electronic search strategy underwent internal checking for syntax and completeness. When outcome data were measured but incompletely reported, study authors were contacted when feasible and necessary to obtain missing information.

Selection Process

Two reviewers (WZ and LH) conducted the study selection process in a 2-phase approach, initial screening of titles and abstracts to assess relevance, followed by full-text evaluation to determine final eligibility. Reference management software (EndNote) was used to eliminate duplicate records, and all screening decisions and reasons for exclusion were systematically documented. Any discrepancies between reviewers were resolved through consensus discussions involving a third reviewer. No additional information was sought from investigators, and no automation tools were used in the selection process. No translation was required because searches were restricted to English-language full-text reports.

Data Collection Process

The data extraction process was independently conducted by 2 reviewers (WZ and LH) using a predesigned Microsoft Excel spreadsheet. Extracted information included basic study details (author, year of publication, and country), study design, sample size, participant characteristics (eg, sex and age), type of intervention and control group settings, outcomes, and follow-up duration. Any discrepancies in extracted data were resolved through discussion, with a third reviewer adjudicating when necessary. In cases of missing or unclear data, attempts were made to contact the original authors for clarification. All extracted data were cross-verified to ensure accuracy and consistency. No automation tools were used for data extraction, and no translation was required because inclusion was restricted to English-language full-text reports. No software was used to extract data from figures. When multiple reports corresponded to the same study, data were collated and extracted as a single study; where overlapping or inconsistent information was identified, the most complete report and/or the longest follow-up was prioritized.

Data Items

We focused on five core outcome domains:

1. Psychological health, including depression and anxiety, was assessed using validated instruments such as the Patient Health Questionnaire-9 (PHQ-9), Beck Depression Inventory (BDI), Center for Epidemiologic Studies Depression Scale (CES-D), and Hospital Anxiety and Depression Scale (HADS).
2. Health-related quality of life, measured using the World Health Organization Quality of Life-BREF (WHOQOL-BREF), EuroQol Five Dimensions Questionnaire (EQ-5D), and Short Form Health Survey-36 (SF-36).
3. Sleep-related functioning was evaluated using the Pittsburgh Sleep Quality Index (PSQI) and Epworth Sleepiness Scale (ESS).
4. Functional independence and rehabilitation outcomes were assessed using the Spinal Cord Independence Measure (SCIM), Functional Independence Measure (FIM), and Craig Handicap Assessment and Reporting Technique (CHART).
5. Pain intensity, measured using the Numerical Rating Scale (NRS).

For each outcome domain, we extracted all relevant measurement instruments, assessment time points, and available analytical results. For meta-analysis, we prioritized postintervention values. For meta-analysis, we prioritized postintervention values at the end of the intervention; when multiple postintervention time points were available, we prioritized the end-of-intervention time point and summarized other time points narratively and explored them in sensitivity analyses when appropriate. Effect estimates were calculated respecting each instrument's original scale without recoding. No changes were made to outcome domain definitions or selection rules after protocol registration. When required information was missing or unclear and could not be confirmed with the study authors, it was recorded as not reported, and no unverifiable assumptions were made.

Study Risk of Bias Assessment

The risk of bias was evaluated using the Cochrane Risk of Bias 2 (version 2; RoB 2; 22 August 2019 guidance), implemented using the RoB 2 Excel tool (ROB2_IRPG beta v9), a structured framework designed to assess the internal validity of trial outcome results. RoB 2 addresses five distinct domains: (1) bias related to the randomization procedure; (2) bias resulting from deviations from the intended intervention; (3) bias associated with incomplete outcome data; (4) bias in the assessment of outcomes; and (5) bias stemming from selective reporting of results. Each domain is appraised through a standardized decision algorithm, culminating in an overall judgment for each outcome, classified as "low risk," "some concerns," or "high risk." Two reviewers independently assessed risk of bias for all included trials, with disagreements resolved by discussion and third-reviewer adjudication when needed. No adaptations were made to the RoB 2 tool, no additional information was sought from study investigators for risk-of-bias assessment, and no automation tools were used. Risk-of-bias judgments informed sensitivity analyses and the certainty of evidence appraisal.

Data Analysis

Effect Measures

All outcomes synthesized quantitatively were continuous. We used mean difference (MD) when outcomes were measured using the same instrument and standardized mean difference (SMD) when different instruments were used within the same outcome domain. Results were reported with corresponding 95% CIs, which were used to quantify the uncertainty around the average pooled effect. No prespecified thresholds were used to categorize effect sizes, and synthesized results were not re-expressed into alternative effect measures.

XSL•FO
RenderX

Synthesis Methods

Studies were grouped for synthesis by the 5 predefined outcome domains. Within each domain, studies were considered eligible for meta-analysis when interventions and outcome constructs were sufficiently comparable and postintervention data were extractable; otherwise, findings were synthesized narratively. We addressed unit-of-analysis issues by ensuring that each participant sample contributed only once to any given meta-analysis and by linking multiple reports of the same trial to a single study, thereby avoiding double-counting and correlated data [45]. Data were prepared for synthesis by prioritizing end-of-intervention postintervention values when multiple follow-up time points were reported and by harmonizing effect direction across instruments; when required, summary statistics were missing, authors were contacted where feasible and otherwise studies were not pooled for that outcome. Given anticipated clinical and methodological heterogeneity across trials, all meta-analyses used a random-effects model [46]. Random-effects inference was based on the Hartung-Knapp-Sidik-Jonkman method, and between-study variance was estimated using restricted maximum likelihood [47]. Heterogeneity was quantified using Cochran Q and the between-study variance (τ^2) and its square root (τ); I^2 was reported as a descriptive statistic and considered in interpretation, but not used to choose the meta-analysis model. Confidence intervals were used to quantify uncertainty around the average pooled effect. To aid interpretation of how true effects might vary across settings, 95% prediction intervals were additionally reported for main meta-analyses, but not for subgroup meta-analyses. Prediction intervals were reported when at least 5 studies were available; for analyses including 5 - 9 studies, they were interpreted cautiously, whereas for analyses with fewer than 5 studies they were not reported [48]. We additionally conducted exploratory subgroup analyses for depressive symptoms, stratified by intervention duration, depression measurement instrument, and follow-up time point, to explore potential sources of between-study heterogeneity. Analyses were implemented in R (*meta package*; R Foundation for Statistical Computing).

Reporting Bias Assessment

Potential reporting biases and small-study effects were assessed using funnel plot asymmetry when at least 10 studies were available for an outcome meta-analysis. When fewer than 10 studies were available, formal assessment of funnel plot asymmetry was not performed due to limited power and this limitation was noted. Two reviewers (WZ and RL) assessed funnel plot asymmetry independently, with a third reviewer adjudicating when needed. The Egger test was not performed. No additional information was sought from study investigators specifically for this assessment, no automation tools were used, and no dedicated tool for bias due to missing results was applied. Funnel plot asymmetry was interpreted as evidence of small-study effects, for which publication bias is only one of several possible explanations [49].

Certainty Assessment

The certainty of evidence for each key outcome was assessed using the GRADE approach (Grading of Recommendations Assessment, Development, and Evaluation). Two reviewers (WZ and LW) independently rated the certainty of evidence as high, moderate, low, or very low, considering risk of bias, inconsistency, indirectness, imprecision, and publication bias. Disagreements were resolved through discussion with third-reviewer adjudication. GRADE judgment and certainty ratings were reported alongside pooled results in the Results section.

Ethical Considerations

Ethical approval is not applicable. This study is a systematic review and meta-analysis of previously published studies and did not involve the collection of individual-level data.

Results

Study Selection

A total of 940 records were identified through comprehensive database searches, including PubMed (n=101), Web of Science (n=249), EMBASE and Ovid MEDLINE (n=418), and the Cochrane Central Register of Controlled Trials (n=168), with an additional 4 records identified from prior systematic reviews (n=4). After deduplication using reference management software, 531 unique records remained for title and abstract screening, of which 414 were excluded (reviews or meta-analyses, n=75; irrelevant interventions, n=19; other records not meeting the predefined inclusion criteria, n=320). A total of 117 reports were sought for full-text retrieval, and all were retrieved (n=0). After full-text assessment, 82 reports were excluded due to study protocols without results (n=54) or lack of an RCT design (n=28). Ultimately, we included 33 studies (35 reports) [9-43]. Two studies had multiple reports (Arora et al 2017 [9]/2017 [10]; Irgens et al 2022 [26]/2024 [27]). The study selection process is shown in Figure 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.

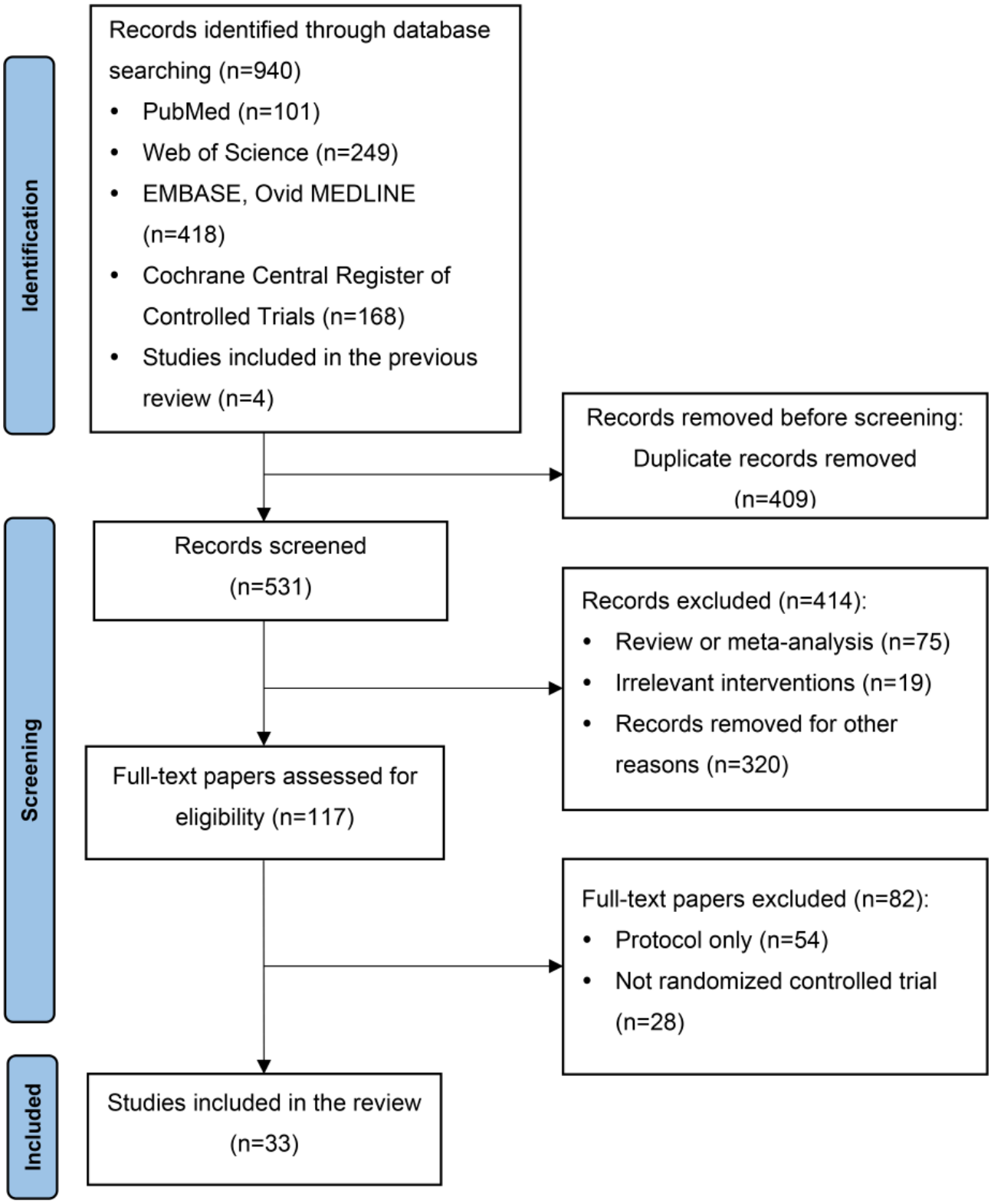

Study Characteristics

We included 33 studies (35 reports) [9-43]; 2 studies had multiple reports (Arora et al 2017/2017 [9,10] clinical and economic evaluations; Irgens et al 2022/2024 [26,27] clinical and cost-utility and environmental analyses), while all other studies were single reports. Sample sizes ranged from 13 participants to >400, with participants predominantly male, and reported mean or median ages generally spanning adulthood to older age. Intervention duration varied from single-session training to 8-12 weeks, to 3-6 months, and up to 2 years in community follow-up programs. Follow-up commonly occurred at the end of treatment and at short-term time points such as 1-3 months. Some studies extended follow-up to 6-12 months or longer. Full study characteristics are presented in Table 1.

The included studies used a diverse array of remote or hybrid intervention modalities. Telephone-based interventions were

XSL•FO
RenderX

the most frequently used modality, accounting for 27.3% of studies (9/33) [9,11-13,20,35,36,39]. These interventions typically involved scheduled calls by health care professionals to deliver health consultations, psychological counseling, or disease-specific education. In some cases, telephone contact was supplemented with home visits to enhance personalization and continuity of care (9.1%, 3/33) [15,23,24]. Web-based or mobile app and social-platform interventions accounted for 18.2% of the included studies (6/33) [29-34]. These platforms, such as WeChat (Tencent), mobile apps, and dedicated websites, facilitated ongoing care, rehabilitation guidance, interactive education, and multidisciplinary support. These interventions were typically led by nurses or rehabilitation teams. Videoconferencing-based interventions were used in 15.2% of studies (5/33) [17,19,26,27,38,42], offering real-time interaction and visual feedback for psychological support, rehabilitation, and educational guidance. Tele-supervised rehabilitation or teleexercise programs also accounted for 15.2% of studies (5/33) [16,18,28,40,43] and were delivered through remote supervision and periodic reassessment, sometimes combined with adjunct technologies. A minority of studies investigated novel intervention modalities, such as transcranial direct current stimulation (tDCS) and functional electrical stimulation exercise training (FES-ET), reflecting the evolving landscape of remote therapeutic strategies (eg, Chantanachai et al [16] and Kowalczewski et al [28]). Psychological and behavioral interventions were widely represented, including cognitive behavioral therapy (CBT), mindfulness-based training, vocational planning, and electronic cognitive therapy (eg, Burke et al [14]; Hearn et al [22]; and Migliorini et al [37]). Full details of intervention modalities for each study are provided in Table 1.

Risk of Bias Assessment

The risk of bias across included studies was assessed using the Cochrane RoB 2 tool, and the results are summarized in Figure S1 in Multimedia Appendix 1. Multimedia Appendix 1 presents study-level judgments across the 5 RoB 2 domains (randomization process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported result) together with an overall judgment. Overall, a substantial proportion of included studies were rated as having some concerns, and a smaller subset was judged to be at high risk of bias, while approximately half were rated as low risk overall. Figure S1B Multimedia Appendix 1 shows the distribution of risk of bias judgments by domain. Missing outcome data had the highest proportion of low-risk judgments (approximately 94%), indicating that incomplete outcome data were generally well addressed in most trials. The randomization process domain also showed a high proportion of low-risk judgments (approximately 71%), suggesting that sequence generation and allocation procedures were relatively robust in many studies. In contrast, deviations from intended interventions and selection of the reported result were more frequently rated as some concerns, reflecting limitations related to intervention adherence, analytic approach, or reporting transparency. Measurement of the outcome showed a more mixed pattern, with a notable proportion of studies judged to be at high risk (approximately 23%), consistent with challenges related to blinding, outcome ascertainment, or reliance on self-reported measures. Taken together, these results indicate that concerns were driven primarily by issues related to deviations, outcome measurement, and selective reporting rather than missing data.

Results of Individual Studies and Syntheses

Psychological Health

This systematic review synthesized evidence from 16 studies [9,11,13,14,18,20,21,22,23-34,35,36,37,42] evaluating the efficacy of telehealth interventions in mitigating depressive symptoms among individuals with SCI. For each study, Figure 2A presents group-level summary statistics for the intervention and control groups, including sample size, mean, and SD, together with the study-specific effect estimate and its 95% CI. Due to heterogeneity in depression assessment instruments across studies, SMD was used as the effect size metric, and pooled estimates were calculated using a random-effects model. Individual study effects varied in direction and magnitude, with several trials suggesting reductions in depressive symptoms and others showing little or no difference between groups. The pooled random-effects estimate suggested no clear average benefit (SMD=−0.11, 95% CI −0.26 to 0.05; P=.16), based on 740 participants in the intervention groups and 730 in the control groups. Heterogeneity was moderate (Q_{15}=23.54; P=.07; tau=0.106; tau²=0.0113; I^2=36.3%). The prediction interval was −0.37 to 0.15, suggesting that the true effect may vary across settings and may not be consistently beneficial in all contexts.

XSL•FO
RenderX

Figure 2. Forest plots show (A) depression and (B) anxiety. Each study reports n, mean, and SD for intervention and control groups and the study effect with 95% CI. Depression is summarized as standardized mean difference (SMD). Hedges *g* and anxiety as standardized mean difference (MD). Random-effects models were used. Heterogeneity is reported as Q, tau, tau², and I^2. Negative values favor the intervention [9,11,13,14,18,20,21,22,23-34,35,36,37,42].

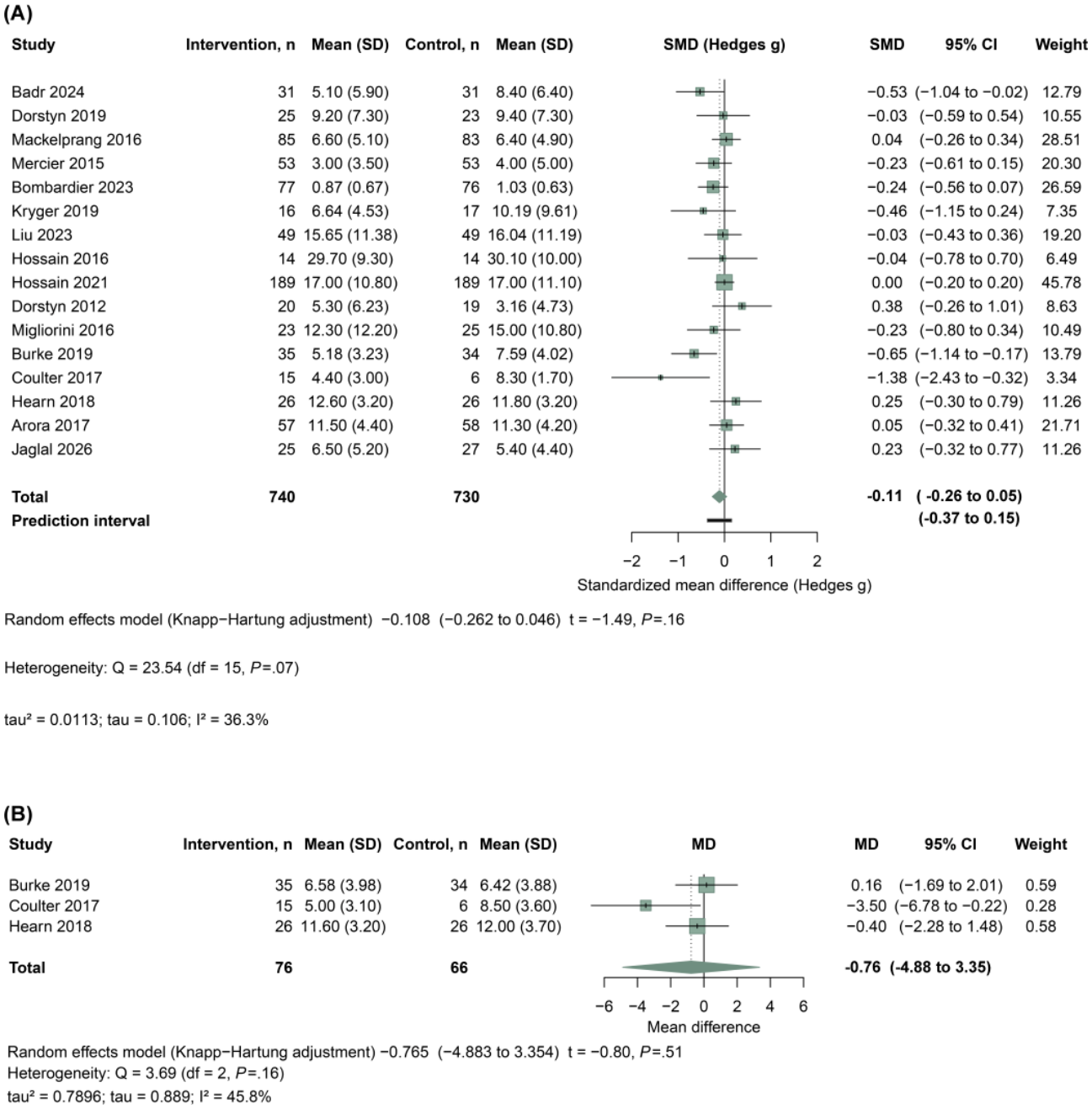

To explore potential sources of between-study heterogeneity, prespecified subgroup analyses were conducted according to intervention duration, depression instrument, and follow-up time point, with results presented in Figures S2-S4 in Multimedia Appendix 1. When stratified by intervention duration (Figure S2 in Multimedia Appendix 1), the pooled effect for interventions lasting 5 - 12 weeks was SMD=−0.19 (95% CI −0.59 to 0.20), whereas interventions longer than 12 weeks showed a smaller pooled effect (SMD=−0.06, 95% CI −0.21 to 0.09). The test for subgroup differences did not indicate statistically significant effect modification by duration. Subgrouping by depression measure (Figure S3 in Multimedia Appendix 1) similarly yielded no evidence of differential effects across instruments (test for subgroup differences, *P*=.43), with pooled estimates spanning from SMD=−0.33 (HADS-D) to approximately null effects for CES-D. Finally, stratification by follow-up time point (Figure S4 in Multimedia Appendix 1) suggested that effects varied across assessment windows: estimates were small at 1-≤3 months (SMD= −0.21, 95% CI −0.52 to 0.09), but were more pronounced and statistically significant at >3-≤6 months (SMD= −0.31, 95% CI −0.57 to −0.04; *P*=.03). At >6 months, effects were again closer to null (SMD= −0.07, 95% CI −0.22 to 0.09). Overall, these subgroup analyses did not provide strong evidence that intervention duration or outcome instrument accounted for heterogeneity, whereas follow-up stratification suggested a potential time-dependent pattern, with a more favorable effect at >3-≤6 months than at earlier or later assessment windows.

XSL•FO
RenderX

Additionally, 3 studies (Burke et al [14], Coulter et al [18], and Hearn et al [22]) used the HADS-Anxiety subscale to evaluate anxiety outcomes (Figure 2B). The pooled random-effects estimate was imprecise and did not show clear evidence of benefit (MD=−0.76, 95% CI −4.88 to 3.35; *P*=.51), based on 76 participants in the intervention groups and 66 in the control groups. Heterogeneity was moderate (Q_2=3.69; *P*=.16; tau=0.889; tau²=0.7896; I^2=45.8%).

Health-Related Quality of Life

Four studies [11,14,22,29] reported WHOQOL-BREF domain scores and were included in the meta-analyses of health-related quality of life, covering the physical, psychological, social, and environmental domains (Figure 3). For each study, Figure 3 presents group-level summary statistics for the intervention and control groups (sample size, mean, and SD), together with the study-specific effect estimate and its 95% CI. Pooled effects were summarized as mean differences because all studies used the same instrument and domain scoring. Overall, telehealth interventions did not show clear evidence of benefit in the physical domain (MD=1.06, 95% CI −4.20 to 6.32; *P*=.57; Q_3=3.78; *P*=.29; tau=4.6312; tau²=2.152; I^2=20.5%) or the psychological domain (MD=4.46, 95% CI −2.87 to 11.79; *P*=.15; Q_3=7.80; *P*=.05; tau=3.893; tau²=15.1525; I^2=61.5%), whereas a statistically significant improvement was observed in the social domain (MD=3.27, 95% CI 0.64 to 5.89; *P*=.03; Q_3=0.63; *P*=.89; tau=0.000; tau²=0.0000; I^2=0.0%). Effects in the environmental domain were imprecise and did not show clear evidence of benefit (MD=3.34, 95% CI −1.42 to 8.10; *P*=.11; Q_3=2.83; *P*=.42; tau=0.001; tau²=0.0000; I^2=0.0%).

XSL•FO
RenderX

Figure 3. Meta-analysis of telehealth interventions on quality of life outcomes in individuals with spinal cord injury (SCI). Pooled effect sizes are presented for the 4 World Health Organization Quality of Life-BREF (WHOQOL-BREF) domains: physical, emotional, social, and environmental [11,14,22,29].

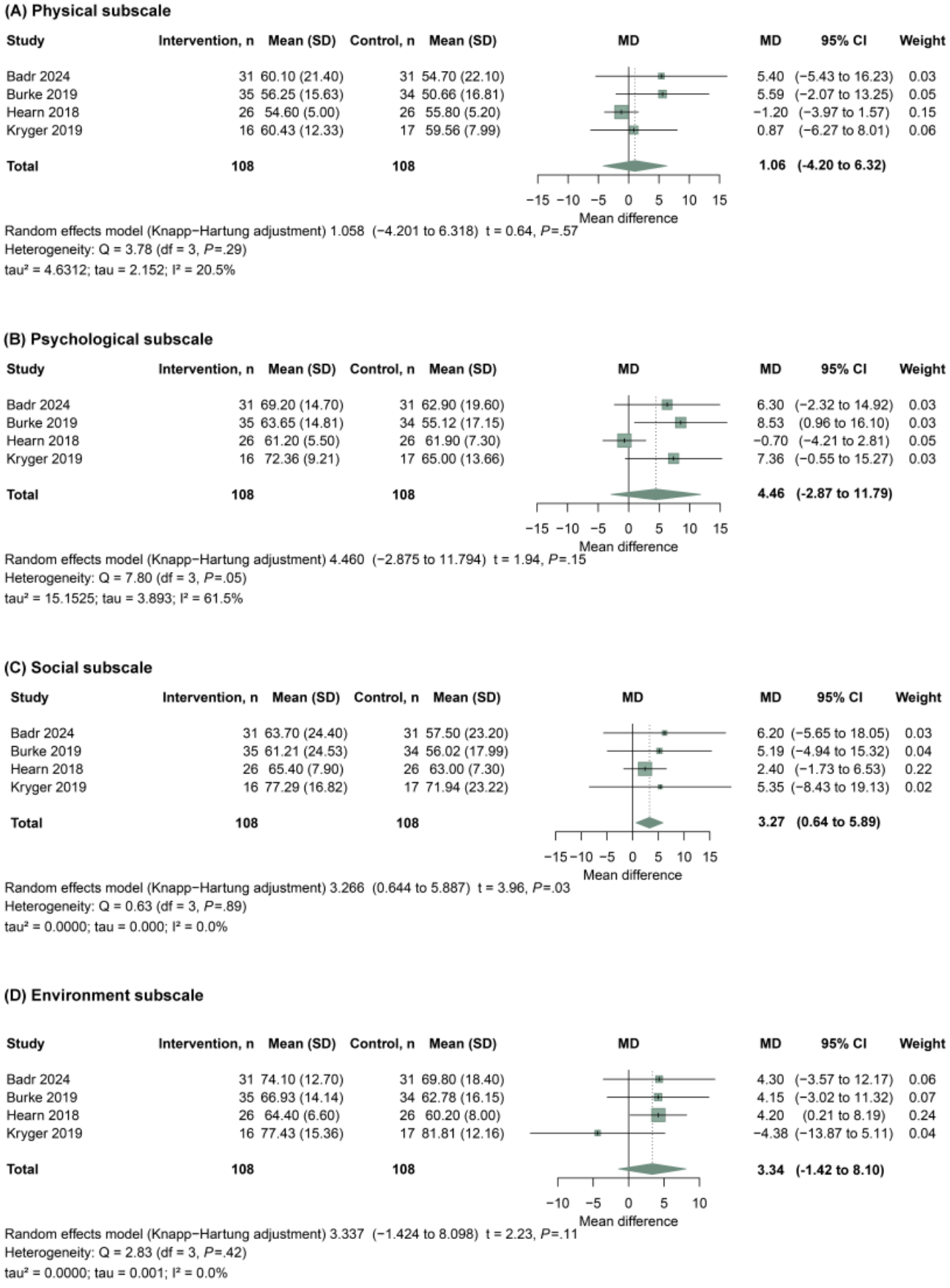

XSL•FO
RenderX

Sleep-Related Functioning

This meta-analysis evaluated the impact of telehealth interventions on sleep-related functioning, using PSQI as the outcome measure (Figure 4). Studies were stratified into 2 subgroups based on follow-up duration. At 1 month (Figure 4A), the pooled random-effects estimate did not show clear evidence of benefit (MD= −1.36, 95% CI −3.12 to 0.39; *P*=.06), based on 65 participants in the intervention groups and 61 in the control groups, with no observed heterogeneity (Q_1=0.03; *P*=.85; tau=0.000; tau²=0.0000; *I*²=0.0%); however, at 3 months (Figure 4B), telehealth interventions were associated with improved PSQI scores (MD= −2.24, 95% CI −3.82 to −0.67; *P*=.04), based on 65 participants in the intervention groups and 64 in the control groups, again with no observed heterogeneity (Q_1=0.03; *P*=.87; tau=0.000; tau²=0.0000; *I*²=0.0%), indicating a statistically significant improvement in sleep-related functioning among patients receiving telehealth interventions.

Figure 4. Meta-analysis of telehealth interventions on sleep quality in individuals with spinal cord injury (SCI). Sleep outcomes were assessed using the Pittsburgh Sleep Quality Index (PSQI). Studies were stratified by follow-up duration into short-term (1 month) and longer-term (3 months) subgroups [11,14].

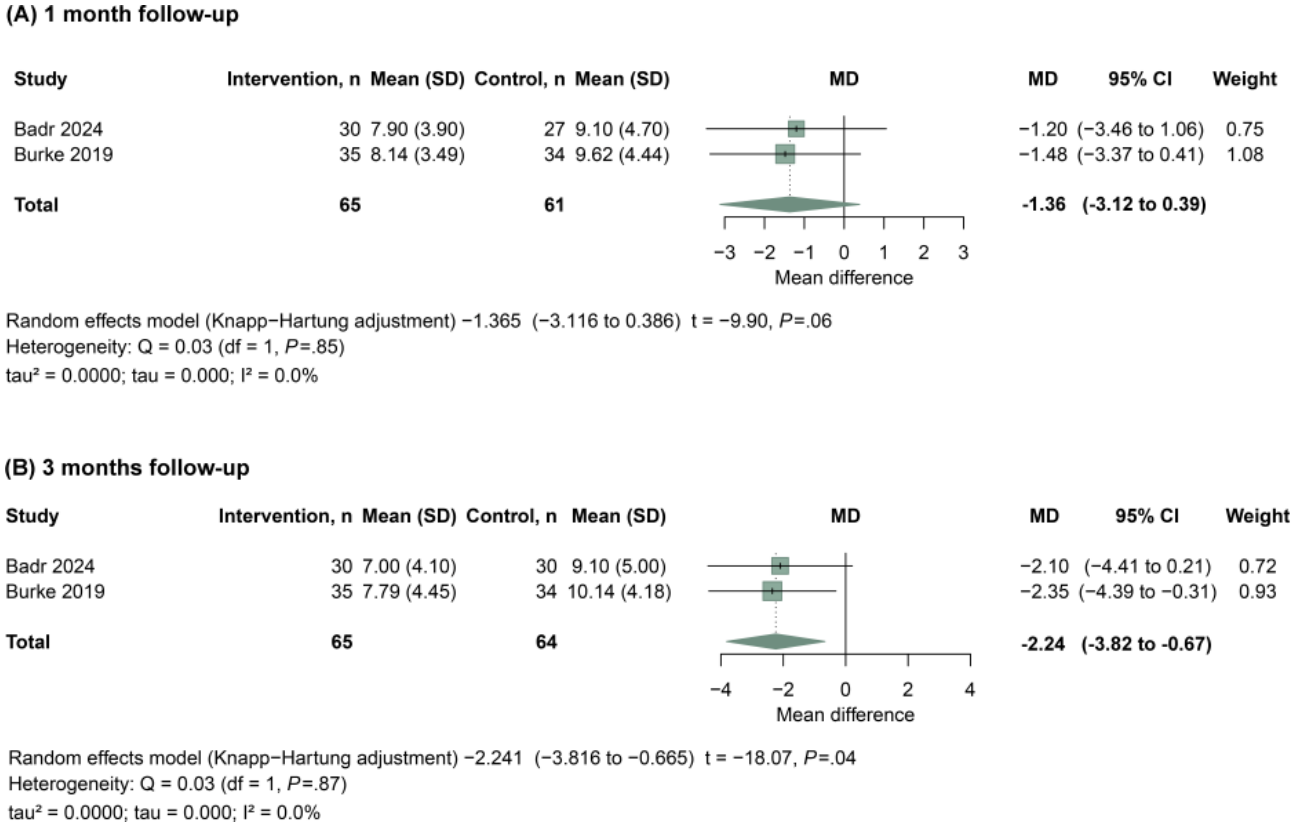

Functional Independence and Rehabilitation Outcomes

This meta-analysis included 4 studies [19,23,24,40] evaluating the impact of telehealth interventions on functional recovery in patients with SCI, using the SCIM scores as the outcome metric (Figure 5A). The pooled MD was 2.09 (95% CI −1.20 to 5.39; *P*=.14), indicating no clear average benefit. Between-study heterogeneity was negligible (Q_3=1.68; *P*=.64; tau=0.00; tau²=0.00; *I*²=0.0%), suggesting that study findings were similar in direction and magnitude despite the absence of a statistically significant pooled effect. Between-study heterogeneity was negligible (Q_3=1.68; *P*=.64; tau=0.00; tau²=0.00; *I*²=0.0%). To further explore domain-specific effects, functional outcomes were analyzed using CHART (Figure 5B), which includes physical independence, mobility, occupation, and social integration dimensions. In physical independence, telehealth interventions showed no statistically significant benefit (MD=−1.80, 95% CI −8.70 to 5.10; *P*=.38), with no observed heterogeneity (I^2=0%). In the mobility domain, the effect was nonsignificant (MD=–0.54, 95% CI –9.91 to 8.83; *P*=.83). The pooled results across occupation and social integration dimensions also did not reach statistical significance (occupation, MD=–1.16, 95% CI −21.75 to 19.43; *P*=.83; social integration, MD=2.0, 95% CI −19.26 to 23.27, *P*=.72)

XSL•FO
RenderX

Figure 5. Meta-analysis of telehealth interventions on functional recovery in individuals with spinal cord injury (SCI). (A) Pooled analysis of 4 studies using the Spinal Cord Independence Measure (SCIM) to assess functional independence [19,23,24,40]. (B) Subgroup analysis using the Craig Handicap Assessment and Reporting Technique (CHART), evaluating 4 domains: physical independence, mobility, occupation, and social integration [11,35,36].

XSL•FO
RenderX

Pain Intensity

This meta-analysis included 3 studies [13,14,22] that evaluated the effect of telehealth interventions on pain relief in patients with SCI, using the NRS as the outcome measure (Figure 6). The pooled random-effects estimate did not show clear evidence of benefit (MD=−0.52, 95% CI −1.69 to 0.66; *P*=.20), based on 138 participants in the intervention groups and 136 in the control groups. Between-study heterogeneity was low (Q_2=2.39; *P*=.30; tau=0.231; tau²=0.0532; *I*²=16.2%).

Figure 6. Meta-analysis of telehealth interventions on pain relief in individuals with spinal cord injury (SCI). Pain outcomes were assessed using the NRS [13,14,22].

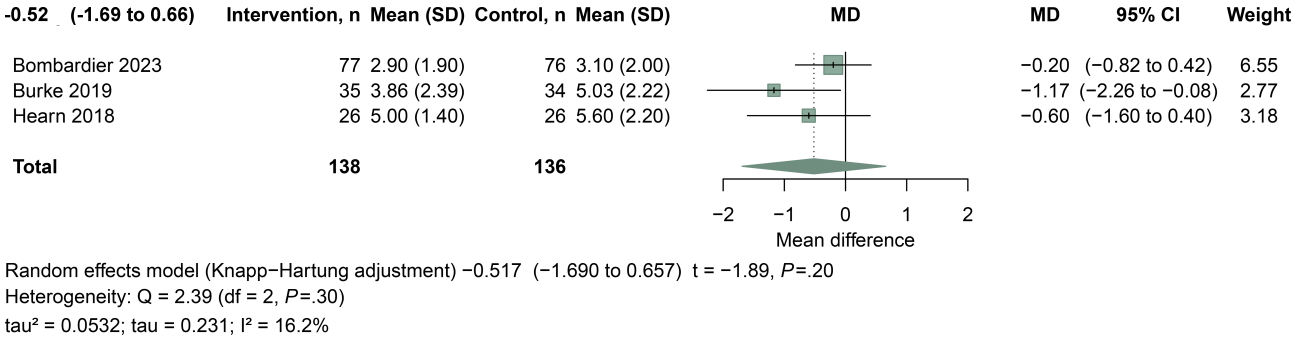

Certainty of Evidence

We assessed certainty of evidence for each critical outcome using GRADE and summarized results in a Summary of Findings table (Table S2 in Multimedia Appendix 1). Overall certainty ranged from high to low, with downgrading most commonly due to risk of bias and imprecision (small samples and wide confidence intervals), with detailed rationale provided in the Summary of Findings table footnotes. Evidence for depressive symptoms was of low certainty overall, although certainty varied by follow-up, including a high-certainty estimate for the >3-≤6 months subgroup and moderate or low certainty at other time windows. Evidence for anxiety was of low certainty, mainly limited by imprecision. For WHOQOL-BREF, certainty was moderate for the social relationships domain but low for the physical, psychological, and environment domains. For sleep quality assessed by PSQI, evidence at 1 month and 3 months was moderate certainty, downgraded primarily for imprecision because few trials reported sleep outcomes. For functional independence and participation, evidence was moderate for SCIM, while CHART domains ranged from moderate to low. For pain intensity (NRS), evidence was of low certainty.

Discussion

Principal Findings

Consistent with our objective to systematically review and meta-analyze the effects of telemedicine interventions on psychological health, health-related quality of life, sleep-related functioning, functional independence and rehabilitation outcomes, and pain intensity in patients with SCI, this review suggests that telemedicine may offer benefits for selected outcomes, but that these effects are domain-specific and time-dependent. The clearest signals of benefit were observed for the WHOQOL-BREF social domain and sleep quality at 3 months, findings that were supported by low between-study heterogeneity and moderate-certainty evidence. A more favorable effect on depressive symptoms was also observed in the >3-≤6 months follow-up subgroup, and although the certainty of evidence for this subgroup was high, this finding appears to reflect a time-specific benefit rather than a consistent overall effect across follow-up periods. By contrast, the overall pooled effect for depressive symptoms was not statistically significant, the prediction interval crossed the null, and certainty was low, suggesting that any overall benefit may be uncertain and may vary across settings. Evidence for anxiety, functional recovery, and pain intensity was also limited by imprecision, small sample sizes, lower-certainty evidence, and risk-of-bias concerns in a proportion of included trials, particularly related to deviations from intended interventions, outcome measurement, and selective reporting. Taken together, these findings suggest that telemedicine may serve as a useful complement to standard rehabilitation pathways in SCI, but that its effectiveness is likely to depend on the outcome targeted, intervention characteristics, follow-up window, and the strength and consistency of the underlying evidence across domains.

Telemedicine interventions may influence outcomes through two related but conceptually distinct pathways. First, remote delivery may be particularly suitable for individuals living with SCI by reducing barriers related to mobility, transportation, and geographic distance, thereby supporting continuity and timeliness of follow-up care [50-52]. Second, many telemedicine models embed enhanced care exposure, such as more frequent contacts, proactive monitoring, structured coaching, and facilitated self-management [53,54], which may improve outcomes irrespective of whether care is delivered remotely or in person. Because intervention dose and comparator intensity were variably reported and often differed across trials, the current evidence does not allow us to disentangle modality effects from care-intensity effects, and this should be considered when interpreting domain-specific benefits.

Psychological complications such as depression and anxiety are common complications following SCI [55]. A pooled analysis estimated the average prevalence of depression to be approximately 22% (95% CI 19% - 26%) [56]. Recent evidence syntheses highlight a substantial and persistent mental health burden after traumatic SCI, with depression being the most frequently studied condition [57]. Beyond prevalence, depressive symptoms have important clinical consequences. In a prospective follow-up study, higher depressive symptom burden

XSL•FO
RenderX

was identified as a determinant of lower participation in inpatient rehabilitation [58]. Consistent with this, a large SCI Model Systems cohort study found that depression during inpatient rehabilitation was associated with poorer postdischarge outcomes at 1 year, including greater rehospitalization burden, worse pain severity, lower participation, and reduced life satisfaction, supporting the clinical value of early screening and timely intervention [59]. Positive affect is a crucial driver of self-management, whereas negative affect constitutes a significant barrier to self-care [60]. Our analysis further revealed that the most significant reduction in depressive symptoms occurred during the 3‑6 month follow-up period, whereas no significant effects were observed at 1‑3 months or beyond 6 months. We hypothesize that this may be attributed to the difficulty patients experience in accepting the reality of disability during the early stages of SCI, coupled with unstable treatment adherence. Over time, as patients gradually adapt and engage more actively with therapeutic measures, the benefits of telemedicine interventions become more evident during mid-term follow-up. Moreover, with the trend toward shorter hospital stays, patients with SCI now have fewer opportunities for inpatient rehabilitation and education, and higher adherence in the early postdischarge period may partially explain the absence of significant long-term differences [61,62]. Interestingly, one previous meta-analysis suggested no significant effects of telemedicine on depression at 3‑6 months (heterogeneity 59%) [8], whereas our findings demonstrated a significant benefit during the same period (heterogeneity 0%). This discrepancy highlights the need for additional high-quality studies to further clarify the actual impact of telemedicine on post-SCI depression. In addition, the contrast between the significant >3 and ≤6 months subgroup finding and the nonsignificant overall pooled effect indicates that any benefit for depressive symptoms should be interpreted as time-specific rather than as a consistent overall effect across follow-up periods, especially given the low certainty of the overall evidence. It is also important to note that different depression scales were used across the included studies, such as PHQ-9, HSCL-20B, BDI-II, CESD, DASS21-Depression, and HADS. In subgroup analyses restricted to HADS-based studies, the pooled estimate trended in the direction of improvement but remained statistically uncertain, highlighting limited power and the need for standardized assessment and reporting. Given that HADS has been widely used in medically ill populations and the depression subscale has supporting psychometric evidence [63], future trials using harmonized measures and clinically interpretable thresholds may help clarify whether the observed trends translate into consistent and clinically meaningful benefits.

With advances in medical care, the survival rate of patients with traumatic SCI has steadily increased [64]; however, this has been accompanied by long-term challenges such as mobility impairments, polypharmacy, bladder and bowel dysfunction, and restrictions in activities of daily living, all of which markedly reduce quality of life [65]. Observational studies using the WHOQOL-BREF consistently show substantially lower scores across domains in people with SCI compared with reference populations. In some cohorts, the psychological domain is among the lowest-scoring domains, underscoring the multidimensional and persistent burden of reduced quality of life in this population [66-68]. In our synthesis, telemedicine showed the most consistent benefit in the WHOQOL-BREF social domain at short-term follow-up, whereas evidence for the physical, psychological, and environmental domains remained uncertain and was supported by lower-certainty evidence. The social dimension encompasses interpersonal relationships, sexual life, and social support from friends, where telemedicine also showed favorable effects. In line with our findings, a previous meta-analysis reported that telemedicine interventions improved quality of life at 3 months, whereas the between-group difference disappeared at 12 months. However, given the limited sample size (only two studies), the representativeness of these results remains debatable [8]. Future trials should therefore report domain-level trajectories with adequate follow-up, document intervention intensity and engagement, and include clinically interpretable benchmarks to clarify whether early domain-specific gains translate into sustained improvements in broader quality-of-life outcomes.

Sleep disturbances are common after SCI and can exacerbate comorbidities that directly shape health-related quality of life, particularly mood, pain, and daytime functioning [69,70]. Recent syntheses indicate a substantial burden of insomnia symptoms in SCI and emphasize the need for routine sleep assessment across the continuum of care [71]. Among the studies included in our review, only two systematically assessed the effects of telemedicine using the PSQI. Our findings suggest that telemedicine had a limited impact at 1 month but more favorable effects at 3 months, although this interpretation should remain cautious because sleep outcomes were informed by only a small number of trials despite moderate-certainty evidence. The observed pattern of minimal short-term change with more favorable effects after sustained intervention is clinically plausible, as sleep improvements often require iterative adjustment and adherence over time. Consistent with this, a 12-week mHealth program incorporating multiple behavior-change strategies showed a favorable trend toward improved PSQI scores [72], and an earlier systematic review reported that telecounseling may improve sleep difficulties as part of broader comorbidity management, although the evidence base was small [54]. Until a larger and more methodologically consistent evidence base is available, conclusions regarding sleep benefits should remain cautious, and emphasis should be placed on standardizing definitions, follow-up windows, and reporting across trials.

Concerning functional independence, we did not observe significant improvements in pooled SCIM or CHART outcomes, and the overall evidence remained uncertain despite negligible heterogeneity for SCIM and moderate certainty for that outcome. Nonetheless, several individual trials provide signals that may help explain which intervention components are more likely to support functional gains. Nonetheless, individual trials provide signals that warrant interpretation. Swarnakar et al [40] reported substantial improvements in SCIM self-care and mobility domains after 8 weeks of telemedicine, with greater overall gains than controls. Similarly, Dallolio et al [19] found that 6 months of intervention significantly improved Functional Independence Measure (FIM) scores, particularly in grooming,

XSL•FO
RenderX

dressing, and transfers, consistent with trends observed in SCIM. In contrast, Hossain et al [24] did not observe improvements in SCIM, which may reflect the intervention's focus on complications such as pressure ulcers, urinary tract infections, incontinence, depression, and respiratory conditions rather than targeted functional training. Moreover, the absence of patient selection based on risk profiles or proactive help-seeking behaviors may have diluted the intervention's impact [35]. Given that rehabilitation services in specialized centers are time-limited, telemedicine offers a valuable opportunity to extend long-term rehabilitation and professional guidance, particularly during the COVID-19 pandemic, when maintaining continuity of care remained crucial [40]. Taken together, these findings suggest that functional gains from telemedicine are more likely when interventions include clearly specified functional training components with sufficient intensity and when they are delivered to patients most likely to benefit.

Chronic pain is another common and debilitating consequence of SCI, with profound effects on daily functioning, mental health, and quality of life [73,74]. Oral pharmacological treatment for neuropathic pain after SCI is often inadequate, commonly resulting in only about a 20% - 30% reduction in pain intensity [75], while neuromodulation approaches (eg, transcranial direct current stimulation, transcranial magnetic stimulation) face limitations related to uncertain efficacy and accessibility [76]. Cognitive behavioral therapy–based pain management programs (CBT-PMP) have shown efficacy in improving mood, pain, and daily activities among individuals with SCI [77]. However, their broader implementation is constrained by mobility challenges and the need for specialized expertise [14]. In this context, telemedicine offers a practical route to extend access to structured psychological pain management. A Cochrane review of remotely delivered psychological therapies for chronic pain concluded that internet-delivered CBT can produce small improvements in pain intensity and functional disability, although benefits were not consistently maintained at follow-up, and evidence certainty varied across outcomes [78]. Telemedicine, particularly when integrated with CBT-PMP, is a promising approach to overcoming these barriers. Previous meta-analyses have demonstrated that internet-based psychological interventions can significantly alleviate symptoms in patients with chronic pain [79] and chronic diseases [80]. Although our review did not find significant improvements in SCI-related pain, the available trials suggested a possible benefit, but the evidence remained limited by small sample sizes, low certainty, and imprecision. Larger, multicenter RCTs with sufficient power are needed to clarify the true efficacy of telemedicine for pain management in SCI.

Limitations

Several limitations should be considered when interpreting these findings. First, the number of included studies was relatively small, and some outcomes, such as sleep and pain, were reported in only a few trials, limiting precision and generalizability. Second, clinical and methodological heterogeneity was substantial, including variation in intervention modality, content, intensity, follow-up duration, and outcome measures, which likely contributed to imprecision and limited inference about which telemedicine approaches work best for specific outcomes. Third, many trials were underpowered and had relatively short follow-up, which constrains the assessment of durability and may underestimate longer-term effects that depend on sustained engagement. In addition, even within RCTs, differential uptake and engagement may occur after randomization. Participants with greater digital literacy, technology access, or comfort with telehealth may be more likely to adhere to and benefit from telemedicine interventions, whereas those with lower access or skills may disengage, which could introduce bias through deviations from intended interventions and differential attrition. Few trials reported baseline digital literacy or technology access, limiting our ability to assess effect modification by these factors and potentially constraining generalizability to settings with limited digital resources. Finally, subgroup analyses by delivery modality were limited, as most interventions relied primarily on telephone contacts, whereas emerging approaches such as virtual coaching, synchronous video, and app-supported monitoring may differ in effectiveness across domains.

Conclusion

This systematic review and meta-analysis offer an up-to-date, domain-based synthesis of randomized evidence on telemedicine interventions for SCI. Compared with earlier reviews, it incorporates additional recent trials and evaluates five prespecified outcome domains, enabling a more granular assessment of where evidence is most consistent and where uncertainty remains. Overall, telemedicine may be associated with improvements in selected patient-important outcomes, with the most consistent signals observed for social aspects of health-related quality of life and sleep-related functioning after sustained intervention exposure, whereas effects on depression, functional independence, and pain are less consistent and remain uncertain in several analyses. In real-world SCI care, these findings support telemedicine as a pragmatic adjunct to standard rehabilitation and follow-up, particularly for extending continuity and access for individuals facing mobility, geographic, or resource barriers. Future adequately powered randomized trials should standardize outcomes and follow-up, report intervention dose and engagement, and evaluate which telemedicine modalities and components work best for specific outcomes and patient subgroups to inform domain-specific, personalized tele-rehabilitation strategies.

Acknowledgments

The authors attest that no generative artificial intelligence technology was used in the preparation of the text, figures, or any other content included in this manuscript.

XSL•FO
RenderX

Funding

The work was supported by the National Key R&D Program of China (Grant No. 2023YFC3603800 and No.2023YFC3603801), and National Natural Science Foundation of China (Grant No. 82372574, 82172534, 82202793, and 82302867), and China Postdoctoral Science Foundation (Grant No. 2023M732443), and the National Key R&D Program of China (Grant No.2024ZD0526005).

Data Availability

All data generated or analyzed during this study are included in this published article. The datasets used in the meta-analysis are derived from previously published randomized controlled trials, which are cited in the reference list. Data extracted from the included studies are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: WZ, LH
Methodology: WZ, LH
Literature search: WZ, LH
Data curation: WZ, LH, XS, RL, QW
Risk of bias assessment: WZ, LH, XS, RL, QW
Formal analysis (statistical analysis): WZ, LH, XS, RL, QW
Data synthesis: WZ, LH, LW, QW
Sensitivity analysis: WZ, LH, LW, QW
Visualization: WZ, LH, LW, QW
Supervision: QW
Study design oversight: QW
Writing–review & editing: WZ, LH, XS, RL, LW, QW
All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Database search strategies and summary of findings for telemedicine outcomes in spinal cord injury.
[DOCX File, 7178 KB - jmir_v28i1e87088_app1.docx]

Checklist 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 expanded checklist.
[PDF File, 277 KB - jmir_v28i1e87088_app2.pdf]

References

1. Ahuja CS, Wilson JR, Nori S, et al. Traumatic spinal cord injury. Nat Rev Dis Primers 2017 Apr 27;3(1):17018. [doi: 10.1038/nrdp.2017.18] [Medline: 28447605]
2. McDonald JW, Sadowsky C. Spinal-cord injury. Lancet 2002 Feb 2;359(9304):417-425. [doi: 10.1016/S0140-6736(02)07603-1] [Medline: 11844532]
3. Lu Y, Shang Z, Zhang W, et al. Global, regional, and national burden of spinal cord injury from 1990 to 2021 and projections for 2050: A systematic analysis for the Global Burden of Disease 2021 study. Ageing Res Rev 2025 Jan;103:102598. [doi: 10.1016/j.arr.2024.102598] [Medline: 39603465]
4. Lu Y, Shang Z, Zhang W, et al. Global incidence and characteristics of spinal cord injury since 2000-2021: a systematic review and meta-analysis. BMC Med 2024 Jul 8;22(1):285. [doi: 10.1186/s12916-024-03514-9] [Medline: 38972971]
5. Hu X, Xu W, Ren Y, et al. Spinal cord injury: molecular mechanisms and therapeutic interventions. Signal Transduct Target Ther 2023 Jun 26;8(1):245. [doi: 10.1038/s41392-023-01477-6] [Medline: 37357239]
6. Irgens I, Rekand T, Arora M, et al. Telehealth for people with spinal cord injury: a narrative review. Spinal Cord 2018 Jul;56(7):643-655. [doi: 10.1038/s41393-017-0033-3] [Medline: 29515211]
7. Chen G, Wang T, Zhong L, et al. Telemedicine for Preventing and Treating Pressure Injury After Spinal Cord Injury: Systematic Review and Meta-analysis. J Med Internet Res 2022 Sep 7;24(9):e37618. [doi: 10.2196/37618] [Medline: 36069842]
8. Tan Y, Liang X, Ming W, et al. The efficacy of telemedicine interventions on quality of life and depression in individuals with spinal cord injury: a systematic review and meta-analysis. Front Psychiatry 2025;16:1434376. [doi: 10.3389/fpsyt.2025.1434376] [Medline: 39916741]

XSL•FO
RenderX

9. Arora M, Harvey LA, Glinsky JV, et al. Cost-effectiveness analysis of telephone-based support for the management of pressure ulcers in people with spinal cord injury in India and Bangladesh. Spinal Cord 2017 Dec;55(12):1071-1078. [doi: 10.1038/sc.2017.87] [Medline: 28809389]
10. Arora M, Harvey LA, Glinsky JV, et al. Telephone-based management of pressure ulcers in people with spinal cord injury in low- and middle-income countries: a randomised controlled trial. Spinal Cord 2017 Feb;55(2):141-147. [doi: 10.1038/sc.2016.163] [Medline: 27995939]
11. Badr MS, Martin JL, Sankari A, et al. Intensive support does not improve positive-airway pressure use in spinal cord injury/disease: a randomized clinical trial. Sleep 2024 May 10;47(5):zsae044. [doi: 10.1093/sleep/zsae044] [Medline: 38422375]
12. Bombardier CH, Dyer JR, Burns P, et al. A tele-health intervention to increase physical fitness in people with spinal cord injury and cardiometabolic disease or risk factors: a pilot randomized controlled trial. Spinal Cord 2021 Jan;59(1):63-73. [doi: 10.1038/s41393-020-0523-6] [Medline: 32694748]
13. Bombardier CH, Fann JR, Ehde DM, et al. Collaborative CARE versus usual CARE to improve quality of life, pain, depression, and physical activity in outpatients with spinal cord injury: The SCI-CARE randomized controlled clinical trial. J Neurotrauma 2023 Dec;40(23-24):2667-2679. [doi: 10.1089/neu.2023.0200] [Medline: 37597201]
14. Burke D, Lennon O, Blake C, et al. An internet-delivered cognitive behavioural therapy pain management programme for spinal cord injury pain: a randomized controlled trial. Eur J Pain 2019 Aug;23(7):1264-1282. [doi: 10.1002/ejp.1402] [Medline: 31002442]
15. Carlson M, Vigen CLP, Rubayi S, et al. Lifestyle intervention for adults with spinal cord injury: results of the USC-RLANRC pressure ulcer prevention study. J Spinal Cord Med 2019 Jan;42(1):2-19. [doi: 10.1080/10790268.2017.1313931] [Medline: 28414254]
16. Chantanachai T, Apiworajirawit I, Klamruen P, et al. Tele-rehabilitation using transcranial direct current stimulation combined with exercise in people with spinal cord injury: a randomized controlled trial. J Rehabil Med 2025 May 7;57:jrm42353. [doi: 10.2340/jrm.v57.42353] [Medline: 40331505]
17. Chemtob K, Rocchi M, Arbour-Nicitopoulos K, Kairy D, Fillion B, Sweet SN. Using tele-health to enhance motivation, leisure time physical activity, and quality of life in adults with spinal cord injury: a self-determination theory-based pilot randomized control trial. Psychol Sport Exerc 2019 Jul;43:243-252. [doi: 10.1016/j.psychsport.2019.03.008]
18. Coulter EH, McLean AN, Hasler JP, Allan DB, McFadyen A, Paul L. The effectiveness and satisfaction of web-based physiotherapy in people with spinal cord injury: a pilot randomised controlled trial. Spinal Cord 2017 Apr;55(4):383-389. [doi: 10.1038/sc.2016.125] [Medline: 27596027]
19. Dallolio L, Menarini M, China S, et al. Functional and clinical outcomes of telemedicine in patients with spinal cord injury. Arch Phys Med Rehabil 2008 Dec;89(12):2332-2341. [doi: 10.1016/j.apmr.2008.06.012] [Medline: 19061746]
20. Dorstyn D, Mathias J, Denson L, Robertson M. Effectiveness of telephone counseling in managing psychological outcomes after spinal cord injury: a preliminary study. Arch Phys Med Rehabil 2012 Nov;93(11):2100-2108. [doi: 10.1016/j.apmr.2012.06.002] [Medline: 22705237]
21. Dorstyn D, Roberts R, Murphy G, et al. Work and SCI: a pilot randomized controlled study of an online resource for job-seekers with spinal cord dysfunction. Spinal Cord 2019 Mar;57(3):221-228. [doi: 10.1038/s41393-018-0200-1] [Medline: 30262878]
22. Hearn JH, Finlay KA. Internet-delivered mindfulness for people with depression and chronic pain following spinal cord injury: a randomized, controlled feasibility trial. Spinal Cord 2018 Aug;56(8):750-761. [doi: 10.1038/s41393-018-0090-2] [Medline: 29581519]
23. Hossain MS, Harvey LA, Rahman MA, et al. A pilot randomised trial of community-based care following discharge from hospital with a recent spinal cord injury in Bangladesh. Clin Rehabil 2017 Jun;31(6):781-789. [doi: 10.1177/0269215516654207] [Medline: 27311454]
24. Hossain MS, Harvey LA, Islam MS, et al. A community-based intervention to prevent serious complications and death 2 years after discharge in people with spinal cord injury in Bangladesh (CIVIC): a randomised trial. Spinal Cord 2021 Jun;59(6):649-658. [doi: 10.1038/s41393-020-00546-9] [Medline: 32917948]
25. Houlihan BV, Brody M, Everhart-Skeels S, et al. Randomized trial of a peer-led, telephone-based empowerment intervention for persons with chronic spinal cord injury improves health self-management. Arch Phys Med Rehabil 2017 Jun;98(6):1067-1076. [doi: 10.1016/j.apmr.2017.02.005] [Medline: 28284835]
26. Irgens I, Midelfart-Hoff J, Jelnes R, et al. Videoconferencing in pressure injury: randomized controlled telemedicine trial in patients with spinal cord injury. JMIR Form Res 2022 Apr 19;6(4):e27692. [doi: 10.2196/27692] [Medline: 35438645]
27. Irgens I, Kleven L, Midelfart-Hoff J, et al. Cost-utility analysis and impact on the environment of videoconference in pressure injury. A randomized controlled trial in individuals with spinal cord injury. Spinal Cord Ser Cases 2024 Mar 8;10(1):10. [doi: 10.1038/s41394-024-00621-w] [Medline: 38459049]
28. Kowalczewski J, Chong SL, Galea M, Prochazka A. In-home tele-rehabilitation improves tetraplegic hand function. Neurorehabil Neural Repair 2011 Jun;25(5):412-422. [doi: 10.1177/1545968310394869] [Medline: 21372246]

XSL•FO
RenderX

29. Kryger MA, Crytzer TM, Fairman A, et al. The effect of the interactive mobile health and rehabilitation system on health and psychosocial outcomes in spinal cord injury: randomized controlled trial. J Med Internet Res 2019 Aug 28;21(8):e14305. [doi: 10.2196/14305] [Medline: 31464189]
30. Lawrason SVC, Martin Ginis KA. Evaluating the feasibility, acceptability, and engagement of an mHealth physical activity intervention for adults with spinal cord injury who walk: a randomized controlled trial. J Sport Exerc Psychol 2023 Apr 1;45(2):61-76. [doi: 10.1123/jsep.2022-0087] [Medline: 36870347]
31. Li J, Li QP, Yang BH. Participatory continuous nursing using the WeChat platform for patients with spinal cord injuries. J Int Med Res 2021 May;49(5):3000605211016145. [doi: 10.1177/03000605211016145] [Medline: 34038208]
32. Li QP, Li J, Pan HY. Effects of online home nursing care model application on patients with traumatic spinal cord injury. Risk Manag Healthc Policy 2021;14:1703-1709. [doi: 10.2147/RMHP.S301874] [Medline: 33935524]
33. Liu T, Xie S, Wang Y, et al. Effects of app-based transitional care on the self-efficacy and quality of life of patients with spinal cord injury in China: randomized controlled trial. JMIR Mhealth Uhealth 2021 Apr 1;9(4):e22960. [doi: 10.2196/22960] [Medline: 33792555]
34. Liu Y, Hasimu M, Jia M, et al. The effects of APP-based intervention for depression among community-dwelling individuals with spinal cord injury: a randomized controlled trial. Arch Phys Med Rehabil 2023 Feb;104(2):195-202. [doi: 10.1016/j.apmr.2022.10.005] [Medline: 36332676]
35. Mackelprang JL, Hoffman JM, Garbaccio C, Bombardier CH. Outcomes and lessons learned from a randomized controlled trial to reduce health care utilization during the first year after spinal cord injury rehabilitation: telephone counseling versus usual care. Arch Phys Med Rehabil 2016 Oct;97(10):1793-1796. [doi: 10.1016/j.apmr.2016.03.002] [Medline: 27039058]
36. Mercier HW, Ni P, Houlihan BV, Jette AM. Differential impact and use of a telehealth intervention by persons with MS or SCI. Am J Phys Med Rehabil 2015 Nov;94(11):987-999. [doi: 10.1097/PHM.0000000000000291] [Medline: 25888652]
37. Migliorini C, Sinclair A, Brown D, Tonge B, New P. A randomised control trial of an Internet-based cognitive behaviour treatment for mood disorder in adults with chronic spinal cord injury. Spinal Cord 2016 Sep;54(9):695-701. [doi: 10.1038/sc.2015.221] [Medline: 26690861]
38. Pain H, Soopramanien A, Dallolio L, et al. Outcomes from a randomized controlled trial of telerehabilitation for people with spinal cord injuries. J Telemed Telecare 2007 Jul;13(1_suppl):46-48. [doi: 10.1258/135763307781645095]
39. Rintala DH, Garber SL, Friedman JD, Holmes SA. Preventing recurrent pressure ulcers in veterans with spinal cord injury: impact of a structured education and follow-up intervention. Arch Phys Med Rehabil 2008 Aug;89(8):1429-1441. [doi: 10.1016/j.apmr.2008.01.015] [Medline: 18674978]
40. Swarnakar R, Yadav S, Wadhwa S, Venkataraman S. Effectiveness of telerehabilitation in persons with spinal cord injury during the COVID-19 pandemic (TELE-SCOPE): a single-center, double-blind, randomized controlled trial. Cureus 2023 Jul;15(7):e41513. [doi: 10.7759/cureus.41513] [Medline: 37551233]
41. Worobey LA, Rigot SK, Hogaboom NS, Venus C, Boninger ML. Investigating the efficacy of web-based transfer training on independent wheelchair transfers through randomized controlled trials. Arch Phys Med Rehabil 2018 Jan;99(1):9-16. [doi: 10.1016/j.apmr.2017.06.025] [Medline: 28782541]
42. Jaglal SB, Allin SJ, Craven BC, et al. A pilot randomised controlled trial of the Spinal Cord Injury and You (SCI&U) online peer health coaching self-management program. Pilot Feasibility Stud 2026 Jan 28;12(1):29. [doi: 10.1186/s40814-026-01769-y] [Medline: 41593819]
43. Young HJ, Mohanraj S, Malone LA, et al. Feasibility, usability, and acceptability of a randomized controlled trial evaluating teleexercise interventions for individuals with spinal cord injury: interim analysis of the Spinal Cord Injury Program in Exercise (SCIPE) Study. Arch Rehabil Res Clin Transl 2025 Sep;7(3):100495. [doi: 10.1016/j.arrct.2025.100495] [Medline: 40980516]
44. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc 2021 Apr 1;109(2):174-200. [doi: 10.5195/jmla.2021.962] [Medline: 34285662]
45. Senn SJ. Overstating the evidence: double counting in meta-analysis and related problems. BMC Med Res Methodol 2009 Feb 13;9:10. [doi: 10.1186/1471-2288-9-10] [Medline: 19216779]
46. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods 2010 Apr;1(2):97-111. [doi: 10.1002/jrsm.12] [Medline: 26061376]
47. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
48. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023 Dec;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
49. Sterne JAC, Sutton AJ, Ioannidis JPA, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011 Jul 22;343(jul22 1):d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
50. Woo C, Guihan M, Frick C, Gill CM, Ho CH. What's happening now! Telehealth management of spinal cord injury/disorders. J Spinal Cord Med 2011;34(3):322-331. [doi: 10.1179/2045772311Y.0000000003] [Medline: 21756573]

XSL•FO
RenderX

51. Touchett H, Apodaca C, Siddiqui S, et al. Current approaches in telehealth and telerehabilitation for Spinal Cord Injury (TeleSCI). Curr Phys Med Rehabil Rep 2022;10(2):77-88. [doi: 10.1007/s40141-022-00348-5] [Medline: 35493027]
52. Phillips VL, Vesmarovich S, Hauber R, Wiggers E, Egner A. Telehealth: reaching out to newly injured spinal cord patients. Public Health Rep 2001;116 Suppl 1(Suppl 1):94-102. [doi: 10.1093/phr/116.S1.94] [Medline: 11889278]
53. Mirbaha S, Morgan A, Tang A, Smith-Turchyn J, Richardson J. Models of telehealth service delivery in adults with spinal cord injuries: scoping review. JMIR Rehabil Assist Technol 2023 Jun 29;10:e41186. [doi: 10.2196/41186] [Medline: 37384377]
54. Dorstyn D, Mathias J, Denson L. Applications of telecounselling in spinal cord injury rehabilitation: a systematic review with effect sizes. Clin Rehabil 2013 Dec;27(12):1072-1083. [doi: 10.1177/0269215513488001] [Medline: 23823709]
55. Migliorini C, Tonge B, Taleporos G. Spinal cord injury and mental health. Aust N Z J Psychiatry 2008 Apr;42(4):309-314. [doi: 10.1080/00048670801886080] [Medline: 18330773]
56. Williams R, Murray A. Prevalence of depression after spinal cord injury: a meta-analysis. Arch Phys Med Rehabil 2015 Jan;96(1):133-140. [doi: 10.1016/j.apmr.2014.08.016] [Medline: 25220943]
57. Muthu S, Viswanathan VK, Palani B MS, et al. Mental health problems in traumatic spinal cord injury patients - an umbrella systematic review. Global Spine J 2025 Dec 8;16(4):21925682251407983. [doi: 10.1177/21925682251407983] [Medline: 41359011]
58. Civelek GM, Tulunoglu S, Ozisler Z. Factors affecting rehabilitation participation in patients with spinal cord injury. Neurosciences (Riyadh) 2025 Jul;30(3):201-208. [doi: 10.17712/nsj.2025.3.20240085] [Medline: 40670067]
59. Redepenning DH, Hoffman JM, McKernan GP, Worobey LA. Relationship between depression during inpatient rehabilitation and health, participation, and satisfaction with life at 1-year postinjury for individuals with spinal cord injury: findings from the spinal cord injury model systems database. Arch Phys Med Rehabil 2026 Mar;107(3):416-424. [doi: 10.1016/j.apmr.2025.06.014] [Medline: 40619117]
60. Munce SEP, Straus SE, Fehlings MG, et al. Impact of psychological characteristics in self-management in individuals with traumatic spinal cord injury. Spinal Cord 2016 Jan;54(1):29-33. [doi: 10.1038/sc.2015.91] [Medline: 26055818]
61. Burns AS, Santos A, Cheng CL, et al. Understanding length of stay after spinal cord injury: insights and limitations from the access to care and timing project. J Neurotrauma 2017 Oct 15;34(20):2910-2916. [doi: 10.1089/neu.2016.4935] [Medline: 28245734]
62. Eriks-Hoogland IE, Müller L, Hirsch BDN, Studer L, Gemperli A, Anderson CE. Non-adherence to follow-up care in persons with spinal cord injury within 10 years after initial rehabilitation. J Rehabil Med 2024 Aug 26;56:jrm41083. [doi: 10.2340/jrm.v56.41083] [Medline: 39185548]
63. Sakakibara BM, Miller WC, Orenczuk SG, Wolfe DL, SCIRE Research Team. A systematic review of depression and anxiety measures used with individuals with spinal cord injury. Spinal Cord 2009 Dec;47(12):841-851. [doi: 10.1038/sc.2009.93] [Medline: 19621021]
64. Zadra A, Bruni S, DE Tanti A, et al. Life expectancy and long-term survival after traumatic spinal cord injury: a systematic review. Eur J Phys Rehabil Med 2024 Oct;60(5):822-831. [doi: 10.23736/S1973-9087.24.08462-4] [Medline: 38842067]
65. Eriks-Hoogland I, Jordan X, Baumberger M, et al. Recommendations for long-term follow-up care of secondary health conditions in spinal cord injury/disorder: a systematic review. Front Rehabil Sci 2024;5:1371553. [doi: 10.3389/fresc.2024.1371553] [Medline: 39463853]
66. Jindal R, Bansal P, Gupta S, Garg SK. Quality of life after traumatic thoracolumbar spinal cord injury: a North Indian perspective. Spinal Cord 2023 Jul;61(7):374-382. [doi: 10.1038/s41393-023-00900-7] [Medline: 37161055]
67. Altahla R, Alshorman J, Ali-Shah SZ, Nasb M, Tao X. Quality of life and subjective well-being comparison between traumatic, nontraumatic chronic spinal cord injury, and healthy individuals in China. PeerJ 2024;12:e18709. [doi: 10.7717/peerj.18709] [Medline: 39726753]
68. Chang F, Xie H, Zhang Q, et al. Quality of life of adults with chronic spinal cord injury in mainland China: a cross-sectional study. J Rehabil Med 2020 May 20;52(5):jrm00058. [doi: 10.2340/16501977-2689] [Medline: 32399579]
69. Sankari A, Badr MS, Martin JL, Ayas NT, Berlowitz DJ. Impact of spinal cord injury on sleep: current perspectives. Nat Sci Sleep 2019;11(219-29):219-229. [doi: 10.2147/NSS.S197375] [Medline: 31686935]
70. Norrbrink Budh C, Hultling C, Lundeberg T. Quality of sleep in individuals with spinal cord injury: a comparison between patients with and without pain. Spinal Cord 2005 Feb;43(2):85-95. [doi: 10.1038/sj.sc.3101680] [Medline: 15570322]
71. Morrison A, Dorstyn DS. Insomnia in spinal cord injury: a meta-analysis of observational studies. Sleep Med Rev 2025 Dec;84(102195):102195. [doi: 10.1016/j.smrv.2025.102195] [Medline: 41202519]
72. Hoevenaars D, Holla JFM, Te Loo L, et al. Mobile App (WHEELS) to promote a healthy lifestyle in wheelchair users with spinal cord injury or lower limb amputation: usability and feasibility study. JMIR Form Res 2021 Aug 9;5(8):e24909. [doi: 10.2196/24909] [Medline: 34379056]
73. Ataoğlu E, Tiftik T, Kara M, Tunç H, Ersöz M, Akkuş S. Effects of chronic pain on quality of life and depression in patients with spinal cord injury. Spinal Cord 2013 Jan;51(1):23-26. [doi: 10.1038/sc.2012.51] [Medline: 22547044]
74. Wollaars MM, Post MWM, van Asbeck FWA, Brand N. Spinal cord injury pain: the influence of psychologic factors and impact on quality of life. Clin J Pain 2007 Jun;23(5):383-391. [doi: 10.1097/AJP.0b013e31804463e5] [Medline: 17515736]

XSL•FO
RenderX

75. Baastrup C, Finnerup NB. Pharmacological management of neuropathic pain following spinal cord injury. CNS Drugs 2008;22(6):455-475. [doi: 10.2165/00023210-200822060-00002] [Medline: 18484790]
76. Li L, Huang H, Yu Y, et al. Non-invasive brain stimulation for neuropathic pain after spinal cord injury: a systematic review and network meta-analysis. Front Neurosci 2021;15(800560):800560. [doi: 10.3389/fnins.2021.800560] [Medline: 35221889]
77. Heutink M, Post MWM, Bongers-Janssen HMH, et al. The CONECSI trial: results of a randomized controlled trial of a multidisciplinary cognitive behavioral program for coping with chronic neuropathic pain after spinal cord injury. Pain 2012 Jan;153(1):120-128. [doi: 10.1016/j.pain.2011.09.029] [Medline: 22100355]
78. Rosser BA, Fisher E, Janjua S, Eccleston C, Keogh E, Duggan G. Psychological therapies delivered remotely for the management of chronic pain (excluding headache) in adults. Cochrane Database Syst Rev 2023 Aug 29;8(8):CD013863. [doi: 10.1002/14651858.CD013863.pub2] [Medline: 37643992]
79. Eccleston C, Fisher E, Craig L, Duggan GB, Rosser BA, Keogh E. Psychological therapies (Internet-delivered) for the management of chronic pain in adults. Cochrane Database Syst Rev 2014 Feb 26;2014(2):CD010152. [doi: 10.1002/14651858.CD010152.pub2] [Medline: 24574082]
80. Mehta S, Peynenburg VA, Hadjistavropoulos HD. Internet-delivered cognitive behaviour therapy for chronic health conditions: a systematic review and meta-analysis. J Behav Med 2019 Apr;42(2):169-187. [doi: 10.1007/s10865-018-9984-x] [Medline: 30387008]

abbreviations

BDI: Beck Depression Inventory
CBT-PMP: Cognitive behavioral therapy–based pain management program
CENTRAL: Cochrane Central Register of Controlled Trials
CES-D: Center for Epidemiologic Studies Depression Scale
CHART: Craig Handicap Assessment and Reporting Technique
EQ-5D: EuroQol Five Dimensions Questionnaire
ESS: Epworth Sleepiness Scale
FIM: Functional Independence Measure
GRADE: Grading of Recommendations Assessment, Development and Evaluation
HADS: Hospital Anxiety and Depression Scale
MD: mean difference
NRS: Numerical Rating Scale
PHQ-9: Patient Health Questionnaire-9
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search
PSQI: Pittsburgh Sleep Quality Index
PUSH: Pressure Ulcer Scale for Healing
RCT: randomized controlled trial
RoB 2: Risk of Bias 2
SCI: spinal cord injury
SCIM: Spinal Cord Independence Measure
SF-36: Short Form Health Survey-36
SMD: standardized mean difference
WHOQOL-BREF: World Health Organization Quality of Life-BREF

Edited by S Brini; submitted 04.Nov.2025; peer-reviewed by EP Greenup, L Xie; revised version received 10.Mar.2026; accepted 10.Mar.2026; published 06.May.2026.

Please cite as:
Zhong W, Huang L, Sun X, Liu R, Wang L, Wei Q
Multidimensional Effects of Telemedicine on Patients With Spinal Cord Injury: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e87088
URL: https://www.jmir.org/2026/1/e87088
doi:10.2196/87088

© Wen Zhong, Liyi Huang, Xin Sun, Rui Liu, Lu Wang, Quan Wei. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.May.2026. This is an open-access article distributed under the terms of the Creative Commons

XSL•FO
RenderX

Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Comparing Pulmonary Telerehabilitation and Center-Based Pulmonary Rehabilitation for Effectiveness and Adherence in Chronic Obstructive Pulmonary Disease: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Ya Li[1,2*], MSc; Hailong Zhang[1,2*], PhD; Guixiang Zhao[1,2], PhD; Jiansheng Li[1,2], PhD; Haoxuan Huang[1,2], BS; Longyu Wang[1,2], MSc; Peilin Jia[1,2], MSc

[1]Lung Disease Diagnosis and Treatment Center, National Medical Center, The First Affiliated Hospital of Henan University of Chinese Medicine, 19 Renmin Road, Zhengzhou, Henan, China

[2]Collaborative Innovation Center for Chinese Medicine and Respiratory Diseases Co-constructed by Henan Province & Ministry of Education of P.R. China, Henan Key Laboratory of Chinese Medicine for Respiratory Diseases, Henan University of Chinese Medicine, Zhengzhou, Henan, China

[*]these authors contributed equally

Corresponding Author:
Hailong Zhang, PhD
Lung Disease Diagnosis and Treatment Center, National Medical Center, The First Affiliated Hospital of Henan University of Chinese Medicine, 19 Renmin Road, Zhengzhou, Henan, China

Abstract

Background: Pulmonary rehabilitation (PR) is a cornerstone of chronic obstructive pulmonary disease (COPD) management; however, access to traditional center-based PR (CBPR) remains limited. Digital and remote models, collectively termed pulmonary telerehabilitation (Tele-PR), have increasingly been used, but their heterogeneity in technology use, supervision, and interaction mode may influence effectiveness and sustainability.

Objective: This systematic review and meta-analysis aimed to compare the effectiveness and adherence of Tele-PR with those of CBPR in adults with COPD while systematically evaluating the impacts of supervision intensity and delivery models on key clinical outcomes.

Methods: This review followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 and PRISMA-S (Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension) guidelines. PubMed, Embase, the Cochrane Library, and the Web of Science were searched from inception to December 10, 2025, to identify randomized controlled trials comparing Tele-PR or home-based PR (HBPR) with CBPR in adults with COPD. Random effects meta-analyses were conducted using the Hartung-Knapp-Sidik-Jonkman method. Between-study heterogeneity was assessed using τ^2, I^2, and 95% prediction intervals. Risk of bias was evaluated with the Cochrane Risk of Bias 2 tool, and certainty of evidence was graded using the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) approach.

Results: Seventeen randomized controlled trials involving 1658 participants were included. After intervention, Tele-PR and CBPR showed comparable average effects on exercise capacity by 6-minute walk distance (k=9; n=950, 57.3%; mean difference –5.37 m, 95% CI –15.68 to 4.95; P=.26; τ^2=103.97; I^2=28.2%; 95% prediction intervals=–32.73 to 22.27). Although pooled effects were not statistically significant, substantial heterogeneity was observed across remote delivery models. Subgroup analyses linked digitally supported, synchronously supervised Tele-PR to less between-study variance across several outcomes, indicating greater consistency in treatment effects across different settings while revealing that low-technology HBPR yielded more variable outcomes, particularly in symptom burden. At long-term follow-up (≥6 mo), between-group differences in functional and symptom outcomes diminished, and short-term gains in exercise capacity did not consistently translate into increased daily physical activity. Certainty of evidence ranged from moderate to very low, mainly downgraded for performance bias, inconsistency across intervention models, and imprecision.

Conclusions: Tele-PR may achieve short-term clinical outcomes comparable to CBPR. Distinct from prior reviews, we stratified remote programs by delivery models and supervision, identifying digitally supported Tele-PR and low-technology HBPR as 2 clinically distinct paradigms with differing consistency of effects. We further propose a structured "supervision gradient" to interpret model-dependent variability in effects across Tele-PR approaches, providing a context-sensitive framework for evidence-informed, model-specific implementation. Future remote rehabilitation should integrate real-time professional supervision and long-term behavioral maintenance to sustain benefits. Tele-PR may be particularly valuable for expanding PR access, while CBPR remains essential for patients requiring close in-person supervision or complex multidisciplinary care.

XSL•FO
RenderX

Trial Registration: PROSPERO CRD42025633252; https://www.crd.york.ac.uk/PROSPERO/view/CRD42025633252

(J Med Internet Res 2026;28:e80500) doi:10.2196/80500

KEYWORDS

chronic obstructive pulmonary disease; Tele-PR; pulmonary rehabilitation; digital health; meta-analysis

Introduction

Chronic obstructive pulmonary disease (COPD) is a progressive respiratory disorder characterized by persistent airflow limitations, commonly associated with chronic systemic inflammation, small-airway dysfunction, and emphysema [1]. Owing to its high prevalence, mortality, and disability burden, COPD imposes a substantial public health burden and is recognized as a major chronic condition threatening global health. It is projected that, by 2030, COPD will account for approximately 4.5 million deaths annually, making it the third leading cause of death worldwide. As the disease progresses, severe dyspnea and reduced exercise tolerance may emerge, significantly impairing activities of daily living [2]. Pulmonary rehabilitation (PR) is a comprehensive, individualized intervention based on a thorough patient assessment. PR has been widely incorporated into clinical management guidelines for COPD as the most effective nonpharmacological approach for alleviating dyspnea, improving health status, and enhancing exercise capacity [3,4].

PR is typically delivered by a multidisciplinary team. Current guidelines recommend that, over a 6‑ to 8-week program, patients should attend at least two individualized assessment and training sessions per week. To optimize outcomes, programs should include endurance training, resistance training, and targeted exercises for upper and lower limbs [5]. However, due to high medical costs, health system constraints, and geographic barriers, many patients fail to achieve the expected benefits of PR [6]. Globally, fewer than 2% of eligible patients are estimated to have access to PR services, highlighting the urgent need for effective interventions to improve PR uptake [7]. PR can be implemented in various settings. Pulmonary telerehabilitation (Tele-PR) delivers exercise training, education, and self-management support through synchronous videoconferencing or asynchronous platforms and has the potential to overcome geographic and transportation barriers, thereby enhancing access to rehabilitation services [8-10].

Randomized controlled trials (RCTs) have shown that, when frequency and intensity are comparable, community- and home-based rehabilitation programs achieve outcomes similar to those of hospital-based programs [11]. The current evidence base is no longer limited to COPD. Recent studies in the broader field of respiratory rehabilitation have provided strong support for the effectiveness of remote interventions. In an RCT, Sánchez-Romero et al [12] reported that an evidence-based eHealth educational tool combined with rehabilitation training yielded significant improvements in musculoskeletal symptoms and patient adherence. In addition, a network meta-analysis by Martínez-Pozas et al [13,14] confirmed that Tele-PR was noninferior to face-to-face rehabilitation in improving physical function and quality of life and produced comparable effects in reducing dyspnea and fatigue. These successful applications in complex respiratory conditions provide supportive evidence for the feasibility and potential effectiveness of implementing Tele-PR in COPD.

Nevertheless, several key uncertainties remain in the current evidence regarding Tele-PR in COPD. First, prior systematic reviews have frequently pooled remote interventions with varying levels of technology, supervision intensity, and interaction modalities, thereby implicitly assuming that Tele-PR is relatively homogeneous—a presumption that may obscure important differences across implementation models. Second, evidence on the long-term maintenance effects of Tele-PR remains limited, particularly regarding whether postintervention improvements in exercise capacity translate into sustained changes in daily physical activity, for which consistent conclusions are still lacking. Third, the potential influence of different supervision modalities on safety, consistency of treatment effects, and patient behavioral responses has not been systematically elucidated.

To address these gaps, this systematic review and meta-analysis was designed to comprehensively compare Tele-PR with center-based pulmonary rehabilitation (CBPR) in patients with COPD. The objectives were to (1) evaluate the comparative effectiveness of Tele-PR versus CBPR on exercise capacity, symptom burden, and health-related quality of life; (2) stratify remote rehabilitation by supervision intensity and technological features to compare synchronously supervised, digitally supported Tele-PR with traditional low-technology home-based PR (HBPR); and (3) examine outcome trajectories at the end of the intervention and during long-term follow-up to inform evidence-based implementation and resource allocation.

Methods

Registration

This systematic review and meta-analysis was conducted and reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement [15] and the PRISMA-S extension for literature searches. An expanded PRISMA 2020 checklist, the PRISMA 2020 checklist for abstracts, and the PRISMA-S checklist [16] are provided in Checklist 1. The structure of the Methods section follows the corresponding PRISMA 2020 headings to enhance transparency, reproducibility, and suitability for editorial review. Protocol changes are detailed in Multimedia Appendix 1.

Search Strategy

A comprehensive literature search was performed in PubMed (National Library of Medicine), Embase (Elsevier), the Cochrane Library (Wiley), and the Web of Science (Clarivate) from database inception to October 15, 2024, with an updated search

XSL•FO
RenderX

conducted on December 10, 2025. The search strategy was independently developed de novo and refined by 2 reviewers to identify RCTs comparing digitally supported Tele-PR or low-technology HBPR with CBPR in adults with COPD. Controlled vocabulary terms (eg, MeSH and Emtree) and free-text keywords were combined using Boolean operators. Searches were limited to articles published in English, but no restrictions were applied based on publication date or publication status (Multimedia Appendix 2). Although specific gray literature databases and clinical trial registries were not searched, we manually screened the reference lists of included studies and relevant systematic reviews to identify additional eligible trials. Furthermore, corresponding authors were contacted to request missing or unpublished data where necessary; however, no additional data were obtained. The reasons for exclusion at the full-text screening stage are reported in Multimedia Appendix 3.

Eligibility Criteria

Study eligibility was defined according to the PICOS (population, intervention, comparator, outcomes, and study design) framework. Eligible studies were RCTs enrolling adult participants (aged ≥18 y) with COPD diagnosed according to the Global Initiative for Obstructive Lung Disease criteria (postbronchodilator forced expiratory volume in 1 s [FEV_1]/forced vital capacity of <0.70), regardless of Global Initiative for Obstructive Lung Disease stage (I to IV) or exacerbation history. Tele-PR interventions were eligible if they could be classified a priori into one of 2 conceptually distinct models based on technological infrastructure and supervision intensity. Digitally supported Tele-PR was defined as any comprehensive PR program delivered primarily in the home setting and supported by information and communication technology, enabling real-time or asynchronous interaction with health professionals (eg, synchronous videoconferencing, mobile apps, web-based platforms, or wearable-supported monitoring with structured feedback), while HBPR encompassed lower-technology home programs relying primarily on structured telephone coaching and/or printed training materials, provided that systematic professional guidance or supervision was delivered (ie, not fully unsupervised self-training). To harmonize heterogeneous Tele-PR implementations, we used a prespecified hierarchical taxonomy: delivery models (digitally supported Tele-PR vs low-technology HBPR), supervision modality (synchronous real-time supervised sessions vs asynchronous delayed feedback or periodic contacts), and supervision intensity (high: ≥2 structured contacts per week; medium: ~1 per week; and low or minimal: <1 per week or instruction only). These categories were defined based on trial descriptions and were used for subgroup and exploratory analyses. Meanwhile, the comparator was CBPR, which was defined as any standardized PR program delivered in hospital or community settings with direct, in-person professional supervision. Studies were required to report at least one prespecified outcome (eg, 6-min walk distance [6MWD], COPD Assessment Test [CAT], St. George's Respiratory Questionnaire [SGRQ], Chronic Respiratory Questionnaire, modified Medical Research Council Dyspnea scale, Hospital Anxiety and Depression Scale, daily step count, self-efficacy, dropout, adverse events, or exacerbations). Outcomes reported by only a small number of studies were synthesized descriptively or exploratorily.

Studies were excluded if they (1) used nonrandomized designs (eg, observational studies); (2) were reviews, case series, editorials, protocols, or conference abstracts without accessible full texts; (3) enrolled participants with contraindications to exercise due to severe cardiovascular, musculoskeletal, or other limiting comorbidities; or (4) evaluated interventions not primarily designed for COPD rehabilitation or involved mixed programs in which the specific effects of Tele-PR or HBPR could not be isolated.

Study Selection and Data Extraction

Study selection and data extraction were performed independently by 2 teams using a standardized workflow. Duplicates were removed in EndNote 21 (Clarivate), after which titles or abstracts and full texts were screened in duplicate. Interrater agreement was assessed using Cohen κ [17], indicating substantial agreement for title or abstract screening (κ=0.75) and almost perfect agreement for full-text eligibility assessment (κ=0.92). Disagreements were resolved through discussion with a third reviewer. Backward citation tracking of included studies was conducted as a supplementary search step after full-text screening. Key study characteristics (eg, first author, year, sample size, design, participant characteristics, intervention details, and outcomes) were extracted in duplicate; missing data were requested from corresponding authors when necessary, and discrepancies were resolved through discussion with a third researcher.

Data Analysis

We pooled dichotomous outcomes using risk ratios and continuous outcomes using mean differences (MDs), as outcomes were measured on consistent scales. Meta-analyses were conducted in R (version 4.3.1; R Foundation for Statistical Computing) using the meta package (version 7.0 - 0). Given the anticipated clinical and methodological heterogeneity across interventions, populations, and delivery models, all meta-analyses were conducted using random effects models, irrespective of the magnitude of statistical heterogeneity. This choice was conceptual rather than data-driven and reflects the assumption that true effects may vary across settings. Pooled effect estimates were calculated using the Hartung-Knapp-Sidik-Jonkman (HKSJ) method, which provides more robust CIs and reduces the risk of false-positive findings, particularly when the number of included studies is small [18]. Between-study variance (τ^2) was estimated using the Sidik-Jonkman estimator. When outcomes were reported at multiple time points, separate meta-analyses were conducted for the end-of-intervention and long-term follow-up data (≥6 mo), as these time points address distinct clinical questions regarding short-term efficacy and durability of effects. For the primary meta-analyses, 95% PIs were calculated using the confidence distribution approach proposed by Nagashima et al [19] via the pimeta package. This approach accounts for uncertainty in the estimation of between-study variance and improves coverage probability compared with conventional normal-based methods [20]. While I^2 values are reported descriptively, interpretation of heterogeneity primarily

XSL•FO
RenderX

emphasized PIs, given their greater relevance for real-world applicability. Meta-analyses were not conducted when fewer than 3 studies were available, as estimation of between-study heterogeneity and prediction intervals (PIs) would be statistically unreliable; such outcomes were synthesized narratively. To improve numerical stability for PI estimation, step counts were rescaled (per 100 steps per day) without changing standardized interpretation. Where appropriate (k≥10 studies), we assessed small-study effects using contour-enhanced funnel plots, Egger test (for continuous outcomes), and Peters linear regression test (for binary outcomes), interpreting these as indicators of small-study effects rather than sole proof of publication bias [21,22]. To explore potential sources of heterogeneity, meta-regressions for continuous moderators were visualized using bubble plots, whereas subgroup analyses for categorical moderators were summarized in tabular format to provide detailed group-specific estimates. These analyses are inherently observational and explore potential moderators rather than establish causality [23]. Leave-one-out sensitivity analyses were conducted for all primary outcomes to assess the robustness of pooled estimates and to identify influential studies contributing to between-study heterogeneity. Multiple reports from the same trial were treated as a single study unit, and no study contributed more than one effect size to the same meta-analysis. For multiarm trials, only one eligible comparison per study was included, preventing the duplicate use of shared control groups [24].

Quality Assessment

Methodological quality was assessed using the Cochrane Risk of Bias 2 tool [25]. Two reviewers independently evaluated 5 domains: randomization, deviations from intended interventions, missing outcome data, outcomes measurement, and selective reporting. Each domain was rated as low risk, some concerns, or high risk. Disagreements were resolved through discussion or consultation with a third reviewer. The certainty of evidence for primary outcomes was graded using the GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) approach [26], considering risk of bias, inconsistency, indirectness, imprecision, and publication bias.

Ethical Considerations

As this study is a systematic review and meta-analysis of previously published data, it did not involve the recruitment of human participants or access to identifiable private patient information. Consequently, institutional review board or research ethics board approval was not required. Ethical approval and informed consent were obtained by the authors of the primary studies included in this review.

Results

Study Characteristics

A total of 8847 records were retrieved in the initial search. After stepwise screening, 17 studies were included. The selection process is presented in a PRISMA flow diagram (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram illustrating the study selection process, identifying 17 randomized controlled trials (RCTs) comparing pulmonary telerehabilitation with center-based pulmonary rehabilitation (PR) in patients with chronic obstructive pulmonary disease (COPD).

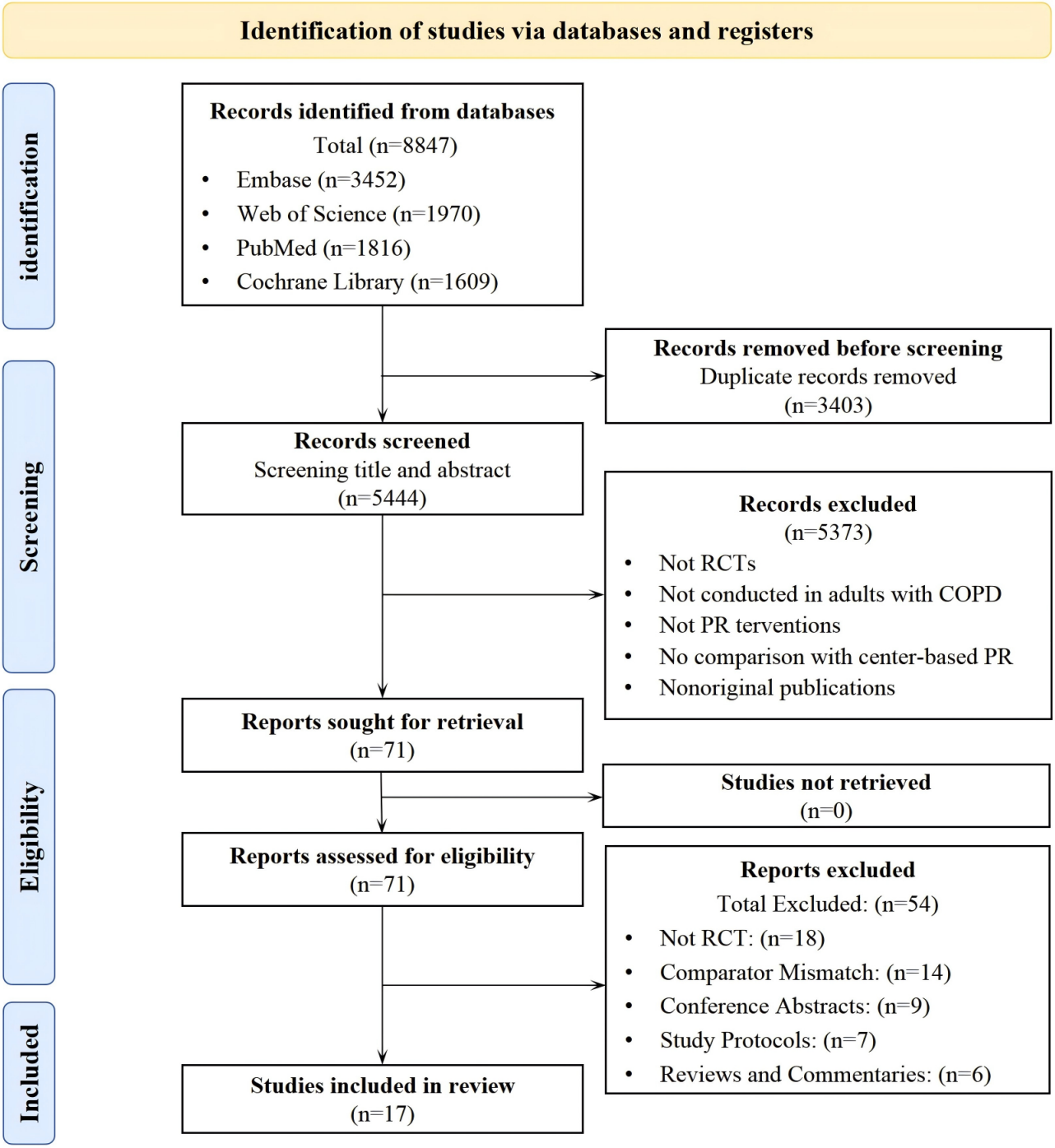

A total of 17 trials [27-43] were identified and included in the quantitative synthesis. To ensure statistical independence, multiple reports originating from the same trial were linked via trial registration numbers, and data were extracted from a single source per outcome to avoid participant double-counting (Multimedia Appendix 4 for a detailed audit of study independence). The included trials were published between 2008 [36] and 2025 [43]. The trials were conducted across multiple regions, including Europe (the United Kingdom [27,34,35,39], Spain [29,37,43], Denmark [30,31], and Greece [40]), Australia [28,33,38,41], Asia (China [32]), North America (Canada [36]), and South America (Brazil [42]). Sample sizes ranged from 54 participants [29] to more than 150 participants [39]. All participants were adults with a confirmed diagnosis of COPD. One trial [33] enrolled patients with chronic respiratory diseases; however, in this meta-analysis, only data from its COPD subgroup were extracted. The most studies included a higher proportion of male participants. Regarding disease severity, participants predominantly had moderate-to-severe airflow limitations. Tele-PR interventions were classified into 3 distinct models: synchronous video supervision, in which direct remote supervision was provided by physiotherapists via real-time videoconferencing [30,31,33]; asynchronous digital platform support, in which structured

XSL•FO
RenderX

rehabilitation content was delivered through web-based platforms [27,34,40], mobile apps, or wearable sensors [29,32], with supervision primarily provided as data-driven asynchronous feedback; and low-technology and telephone support [28,35,36,38,39,42,43], which relied primarily on printed materials for self-management guidance and was supplemented by telephone follow-up to enhance adherence and motivation. A minimal supervised model was also reported [37]. All trials compared Tele-PR with CBPR. The intervention duration was typically 8 to 10 weeks [29,31,33,36,38,41], with some studies lasting 7 weeks [27,28,39] or extending to 12 to 13 weeks [42,43]. Several trials reported long-term follow-up data at 6 to 12 months after the intervention [30,32,33,40,41,43]. Detailed characteristics of the included RCTs are summarized in Table 1. Intervention components and supervision models are detailed in Multimedia Appendix 5.

XSL•FO
RenderX

Table . Characteristics of the included randomized controlled trials.

Study	Country	Sample size (Tele-PR[a]/control)	Age (years), mean (SD)	Male, % (n/N)	FEV_1[b] (% predicted), mean (SD)	Tele-PR intervention (mode)	Supervision (type and frequency)	Control group (setting)	Duration
Chaplin et al (2017) [27]	UK	51/52	• I[c]: 66.4 (10.1) • C[d]: 66.1 (8.1)	• I: 74% (38/51) • C: 63% (33/52)	• I: 59 (29) • C: 55 (21)	Web based (interactive website)	Asynchronous: weekly emails or calls for remote support	Hospital or community PR[e]	7 weeks
Holland et al (2017) [41]	Australia	80/86	• I: 69 (13) • C: 69 (10)	• I: 60.0% (48/80) • C: 59.3% (51/86)	• I: 52 (19) • C: 49 (19)	Telephone based (health coaching)	Minimal: 1 home visit+7 weekly telephone calls	Hospital outpatient PR	8 weeks
Horton et al (2018) [39]	UK	145/142	• I: 68 (9) • C: 67 (8)	• I: 64.1% (93/145) • C: 66.2% (94/142)	• I: 48 (19) • C: 49 (17)	Manual based ("SPACE for COPD")	Minimal: self-managed with manual+2 calls (weeks 2 and 4)	Hospital outpatient PR	7 weeks
Hansen et al (2020) [31]	Denmark	67/67	• I: 68.4 (8.7) • C: 68.2 (9.4)	• I: 47.8% (32/67) • C: 41.8% (28/67)	• I: 33 (10) • C: 34 (8)	Video based (real-time videoconferencing)	Synchronous: real-time supervision by physio via tablet	Hospital outpatient PR	10 weeks
Maltais et al (2008) [36]	Canada	126/126	• I: 66 (9) • C: 66 (9)	• I: 54% (68/126) • C: 57% (72/126)	• I: 46 (13) • C: 43 (13)	Home based (exercise equipment provided)	Remote: initial home visit+weekly telephone calls	Hospital outpatient PR	8 weeks
Vasilopoulou et al (2017) [40]	Greece	50/50	• I: 66.9 (9.6) • C: 66.7 (7.3)	• I: 93% (47/50) • C: 76% (38/50)	• I: 50 (22) • C: 52 (17)	Tablet/web (platform monitoring)	Asynchronous: data upload 3-4 times per week+ weekly feedback calls	Hospital outpatient PR	12 months
Sacristán-Galisteo et al (2025) [43]	Spain	40/40	• I: 67.9 (8.8) • C: 68.9 (8.2)	• I: 72.5% (29/40) • C: 72.5% (29/40)	• I: 53 (16) • C: 55 (16)	Telephone based (home exercise diary)	Remote: weekly telephone calls (15-20 min) for motivation	Health center (face-to-face)	13 weeks
Li et al (2022) [32]	China	50/49	• I: 65.9 (8.9) • C: 65.6 (8.8)	• I: 82.0% (41/50) • C: 76.0% (38/50)	• I: 49 (11) • C: 50 (11)	App/WeChat (smartphone uploads)	Remote: daily interaction via WeChat group (uploads and feedback)	Hospital outpatient PR	12 months
Cox et al (2022) [33]	Australia	47/50	• I: 68 (9) • C: 67 (9)	• I: 42.3% (30/71) • C: 50.7% (36/71)	• I: 59 (25) • C: 63 (26)	Video based (virtual group video)	Synchronous: real-time group videoconferencing via tablet	Hospital outpatient PR	8 weeks

XSL•FO
RenderX

Study	Country	Sample size (Tele-PR[a]/control)	Age (years), mean (SD)	Male, % (n/N)	FEV_1[b] (% predicted), mean (SD)	Tele-PR intervention (mode)	Supervision (type and frequency)	Control group (setting)	Duration
Cerdán-de-las-Heras et al (2021) [29]	Spain	27/27	• I: 67.4 (10) • C: 72.5 (7)	• I: 59.3% (16/27) • C: 55.6% (15/27)	• I: 36 (14) • C: 33 (9)	App/sensor ("VAPA" app+chest sensor)	Hybrid: autonomous AI supervision+remote therapist checks	Hospital standard PR	8 weeks
Burge et al (2021) [28]	Australia	70/71	69 (10)[f]	41.8% (59/141)[g]	50 (37 - 63)[h]	Telephone based (home visit+calls)	Minimal: 1 home visit+7 weekly telephone calls	Hospital outpatient PR	7 weeks
Chaplin et al (2022) [34]	UK	51/52	• I: 68.3 (6.5) • C: 67.4 (8.6)	• I: 90.0% (18/20) • C: 55.9% (19/34)	• I: 54 (27) • C: 56 (19)	Web based (online progress monitoring)	Asynchronous: online tracking+weekly remote contact	Hospital or community PR	8 weeks
Hansen et al (2023) [30]	Denmark	67/67	• 68 (9)[f]	• 44.8% (60/134)[g]	• 33.1 (9.4)	Home based (long-term follow-up)	Supervised: (see [31])	Hospital outpatient PR	62 weeks
Güell et al (2008) [37]	Spain	28/29	• I: 66 (6) • C: 63 (7)	• 100.0% (57/57)[g]	• I: 39 (8) • C: 38 (7)	Home based (unsupervised after instruction)	Minimal: instructions at start, then self-managed	Hospital PR	9 weeks
Lahham et al (2019) [38]	Australia	20/25	• I: 67 (7) • C: 68 (9)	• I: 45.0% (9/20) • C: 56.0% (14/25)	• I: 52 (19) • C: 54 (19)	Telephone based (home visit+calls)	Indirect: initial supervision then weekly calls	Supervised group training	8 weeks
Mendes de Oliveira et al (2010) [42]	Brazil	42/46	• I: 66.4 (10) • C: 71.3 (7)	• I: 81.8% (27/33) • C: 82.6% (19/23)	• I: 48 (23) • C: 52 (24)	Home based (unsupervised after training)	Minimal: initial training at clinic, then self-monitored	Clinic-based PR	12 weeks
Horton et al (2021) [35]	UK	26/25	• I: 67 (9) • C: 67 (7)	• I: 68.3% (43/63) • C: 67.3% (37/55)	• I: 47 (18) • C: 51 (19)	Manual based (similar to [39])	Minimal: same as [39] (manual+2 calls)	Center-based PR	7 weeks

[a]Tele-PR: pulmonary telerehabilitation.

[b]FEV_1: forced expiratory volume in 1 second.

[c]I: intervention group (pulmonary telerehabilitation).

[d]C: control group (center-based PR).

[e]PR: pulmonary rehabilitation

[f]Overall.

[g]All.

[h]Median (IQR).

XSL•FO
RenderX

Methodological Quality of Included Studies

The methodological quality of the included RCTs varied considerably (Figure 2). A total of 4 RCTs were rated as low risk of bias [28,33,36,41], while the remaining studies were classified as high risk or as having some concerns, mainly due to the analysis strategy they used (domain 2) or missing outcomes data (domain 3). For domain 1 (randomization process), risk was generally low, as most studies reported adequate sequence generation and allocation concealment; however, concerns were raised in a study by Cerdán-de-las-Heras et al [29] due to substantial baseline imbalance in a small sample. For domain 2 (deviations from intended interventions), risk depended on the type of analysis because blinding was difficult to implement. Studies using a strict intention-to-treat analysis were rated low risk [28,29,31,33,36,40,41,43], whereas studies relying on per-protocol analyses or completer-only analyses were judged to exhibit high risk [34,35,37-39,42]. Domain 3 (missing outcomes data) was the primary source of bias. High-risk ratings resulted from excessive dropout rates [27,39] or differential dropout between groups [31,37]. Long-term data in a study by Hansen et al [31] and Sacristán-Galisteo et al [43] were affected by substantial missingness, whereas only Cox et al [33] and Vasilopoulou et al [40] showed minimal missing data. For domain 4 (measurement of the outcome), risk was predominantly low, as blinded assessors were used for objective outcomes in most studies; Li et al [32] was an exception, as the absence of assessor blinding may have introduced detection bias. For domain 5 (selection of the reported result), risk was very low, and, in nearly all studies, reported outcomes were consistent with preregistered protocols or trial registry records. The risk of bias assessment of the included RCTs is provided in Multimedia Appendix 6.

Figure 2. Risk of bias summary of randomized controlled trials [27-43].

Meta-Analysis Results

Six-Minute Walk Distance

Nine trials [29,31,33,36,37,40-43] provided analyzable end-of-intervention data for 6MWD. In the primary random effects meta-analysis with HKSJ adjustment, no statistically significant difference in functional capacity was observed between Tele-PR and CBPR (k=9; n=950, 57.3%; mean difference [MD] –5.37 m, 95% CI –15.68 to 4.95; P=.26; τ^2=103.97; I^2=28.2%; PI=–32.73 to 22.27; Figure 3).

Potential sources of heterogeneity were explored using subgroup analyses and meta-regression. Meta-regression identified intervention duration as a significant source of heterogeneity (β=−5.22; P=.01; Table 2 and Figure 4). No significant associations were observed for economic status, publication year, training intensity, or baseline FEV_1 (all P>.05; Table 2 and Multimedia Appendix 7).

XSL•FO
RenderX

Figure 3. Forest plot of randomized controlled trials evaluating the effect of interventions on 6-minute walk distance in patients with chronic obstructive pulmonary disease at the end of the intervention [29,31,33,36,37,40-43]. The analysis was performed using a random effects model. CBPR: center-based pulmonary rehabilitation; HKSJ: Hartung-Knapp-Sidik-Jonkman; MD: mean difference; Tele-PR: pulmonary telerehabilitation.

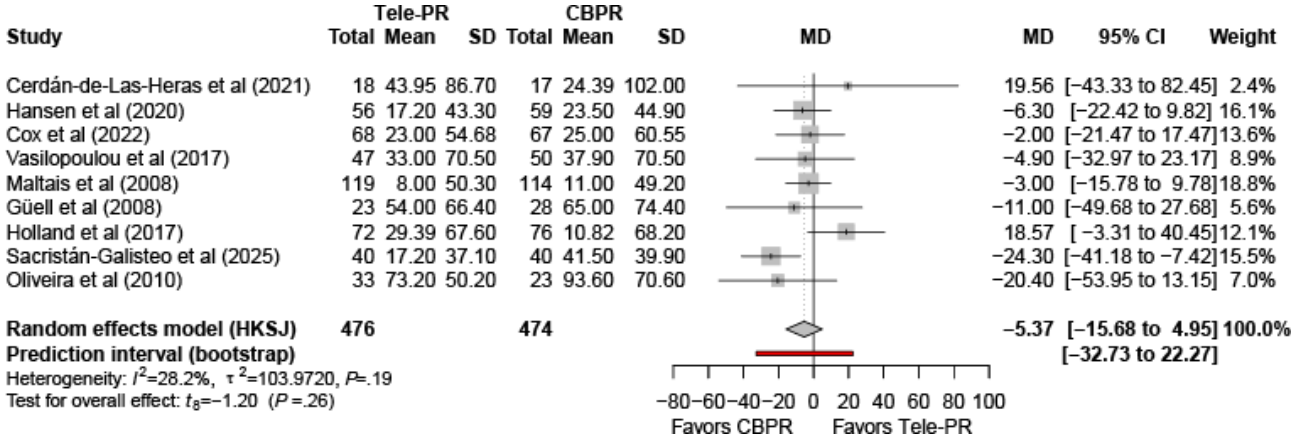

Table . Meta-regression results of examining covariates influencing short-term 6-min walk distance.

Covariate	β coefficient (SE)	*t* (df)	*P* value
Economic status (developing vs developed)	−16.16 (17.78)	−0.91 (7)	.39
Intervention duration (per week)	−5.22 (1.42)	−3.68 (7)	.01
Publication year	−0.33 (0.78)	−0.42 (7)	.69
Training intensity (medium vs high)	−17.94 (10.59)	−1.69 (7)	.13
Baseline FEV_1[a] (% predicted)	−0.01 (0.58)	−0.02 (7)	.98

[a]FEV_1: forced expiratory volume in 1 second.

XSL•FO
RenderX

Figure 4. Meta-regression bubble plot of the association between intervention duration (weeks) and the mean difference in 6-min walk distance (6MWD) [29,31,33,36,37,40,41,42,43]. The size of each bubble is proportional to the weight of the study in the random effects model. The solid line represents the predicted regression line, and the shaded area indicates the 95% CI.

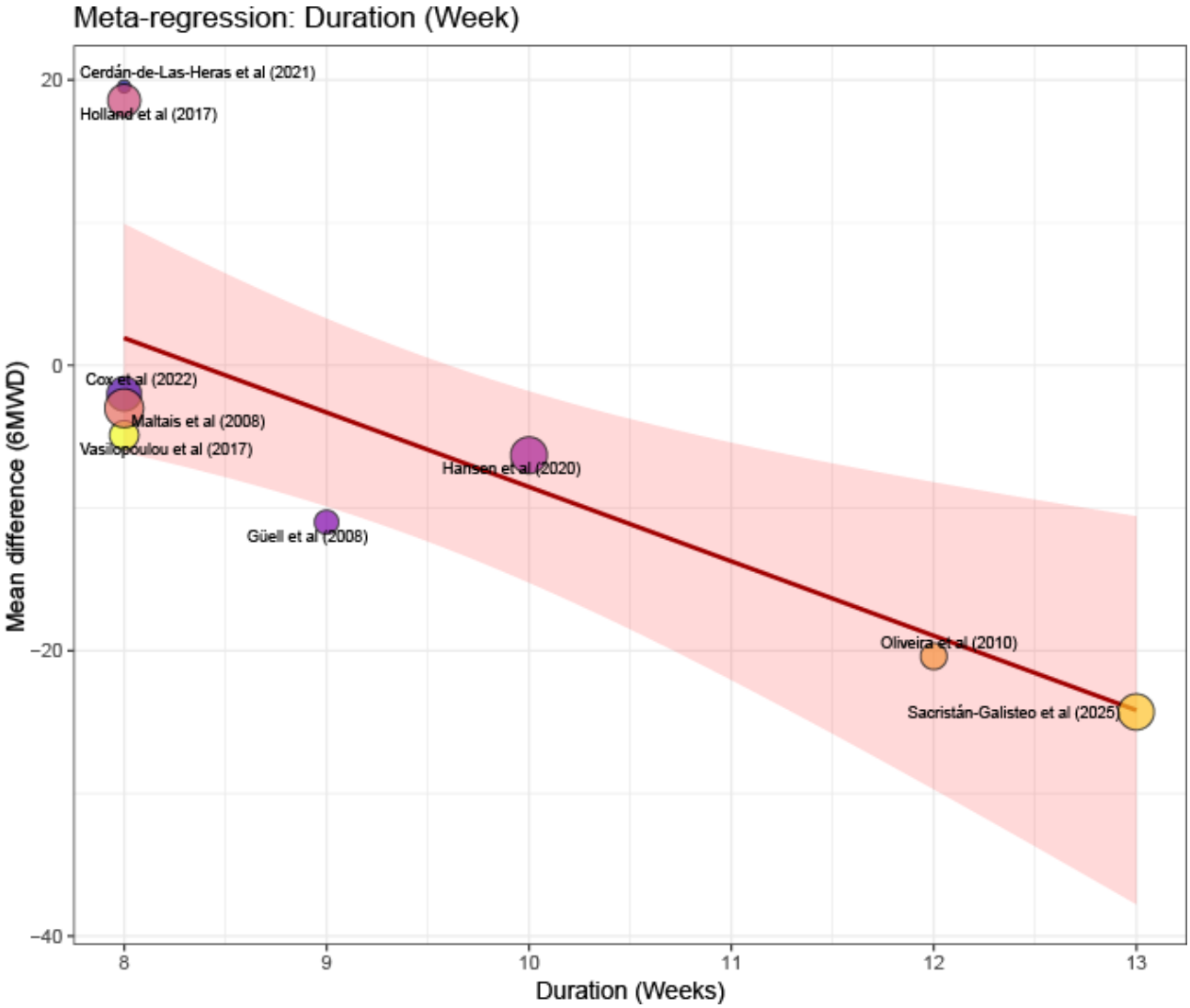

At long-term follow-up, pooled analysis of 9 studies [29,31-33,36,37,40,41,43] also showed no statistically significant difference in 6MWD between groups (k=9; n=948, 57.2%; MD 2.97 m, 95% CI −12.22 to 18.17; P=.66; τ^2=319.32; I^2=44.5%; PI=−41.89 to 47.84; Figure 5).

Figure 5. Forest plot of randomized controlled trials assessing 6-minute walk distance in patients with chronic obstructive pulmonary disease. The analysis includes outcomes at long-term follow-up (≥6 mo) [29,31-33,36,37,40,41,43]. CBPR: center-based pulmonary rehabilitation; HKSJ: Hartung-Knapp-Sidik-Jonkman; MD: mean difference; Tele-PR: pulmonary telerehabilitation.

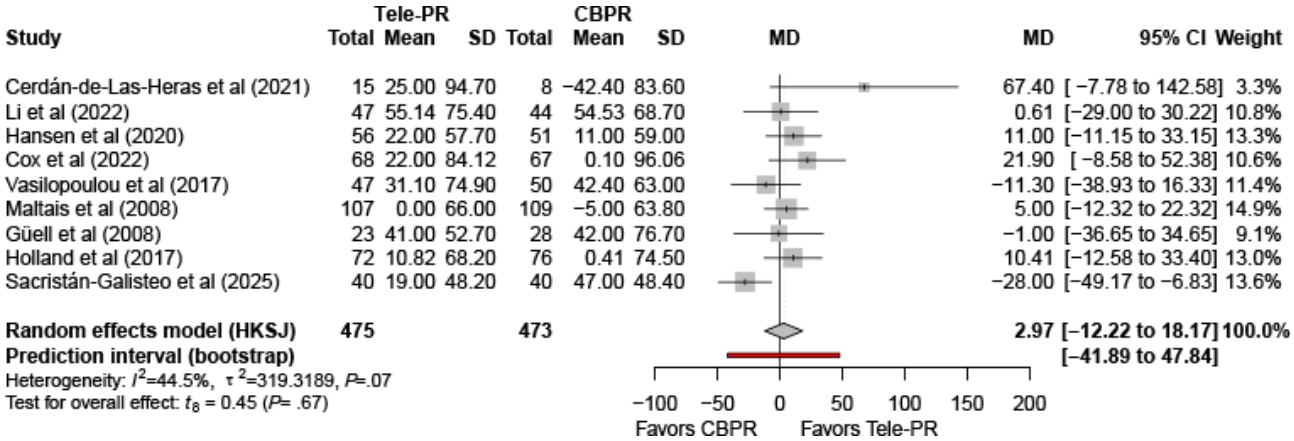

XSL•FO
RenderX

Meta-regression analyses at long-term follow-up included 5 covariates: economic status, intervention duration, training intensity, publication year, and baseline FEV_1. None of the covariates reached statistical significance at the .05 level (all $P>.05$; Table 3 and Multimedia Appendix 7).

Table . Meta-regression results of examining covariates influencing long-term 6-min walk distance.

Covariate	β coefficient (SE)	*t* (df)	*P* value
Economic status (developing vs developed)	−2.84 (22.81)	−0.12 (7)	.90
Publication year	−0.34 (1.20)	−0.28 (7)	.79
Intervention duration (weeks)	−6.47 (3.15)	−2.06 (7)	.08
Training intensity (medium vs high)	−3.19 (17.10)	−0.19 (7)	.86
Baseline FEV_1[a] (% predicted)	−0.61 (0.87)	−0.70 (7)	.51

[a]FEV_1: forced expiratory volume in 1 second.

Daily Steps

Five trials [29,31,34,35,41] reported daily step counts at the end of the intervention. The pooled analysis showed no statistically significant difference between Tele-PR and CBPR (k=5; n=253, 15.3%; MD 4.97 per 100 steps per day, 95% CI –1.84 to 11.78; P=.11; τ^2=18.05; I^2=23.8%; PI=–8.19 to 18.76; Figure 6).

At long-term follow-up, pooled analysis of 3 studies [31,35,41] showed no statistically significant overall difference (Figure 7).

Figure 6. Forest plot of randomized controlled trials assessing daily step counts in patients with chronic obstructive pulmonary disease at the end of the intervention [29,31,34,35,41]. CBPR: center-based pulmonary rehabilitation; HKSJ: Hartung-Knapp-Sidik-Jonkman; MD: mean difference; Tele-PR: pulmonary telerehabilitation.

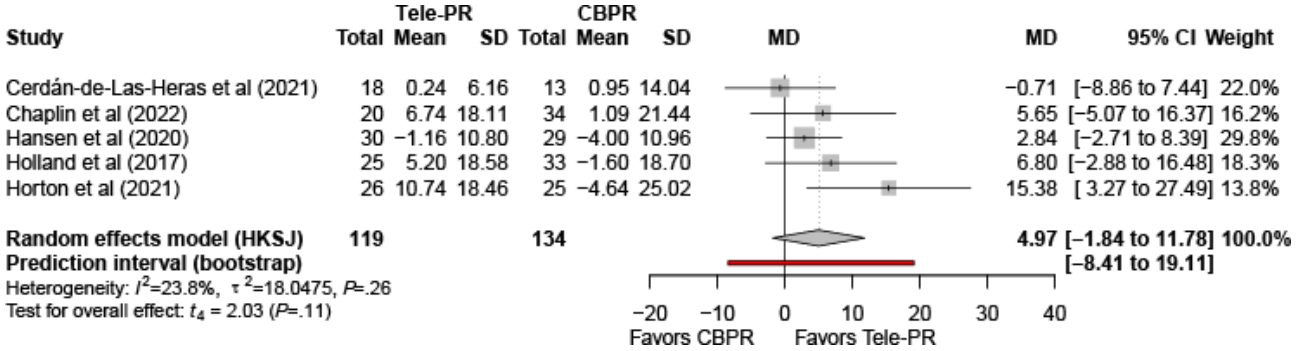

Figure 7. Forest plot of randomized controlled trials evaluating the effect of interventions on daily step counts in patients with chronic obstructive pulmonary disease at long-term follow-up (≥6 mo) [31,35,41]. The analysis was performed using a random effects model. CBPR: center-based pulmonary rehabilitation; HKSJ: Hartung-Knapp-Sidik-Jonkman; MD: mean difference; Tele-PR: pulmonary telerehabilitation.

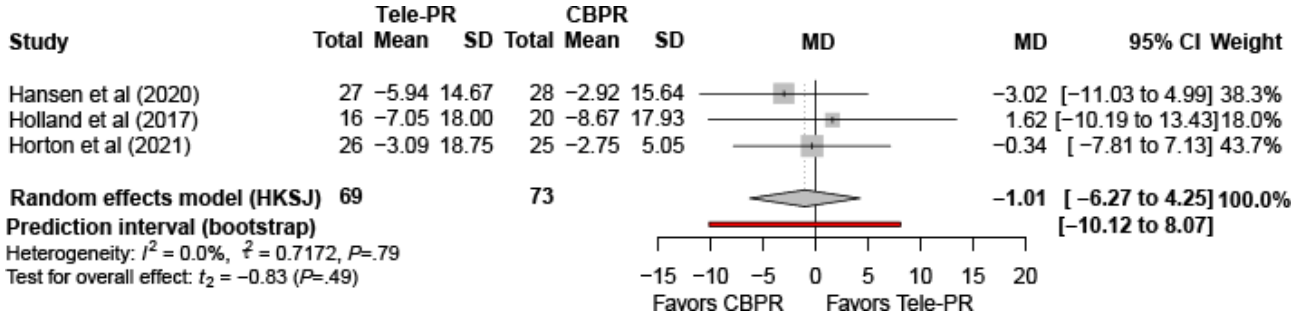

St. George's Respiratory Questionnaire

Three trials [29,36,40] reported SGRQ outcomes at the end of the intervention. Pooled analysis showed no statistically significant difference between Tele-PR and CBPR (k=3; n=362, 21.8%; MD 0.45, 95% CI –7.11 to 8.00; P=.82; τ^2=3.84; I^2=18.8%; PI=–14.65 to 17.49; Figure 8).

At long-term follow-up [29,36,40], pooled analysis likewise showed no statistically significant difference between groups (k=3; n=336, 20.3%; MD –0.01, 95% CI –5.08 to 5.05; $P>.99$; τ^2=1.51; I^2=0%; PI=–10.48 to 12.39; Figure 9).

XSL•FO
RenderX

Figure 8. Forest plot of randomized controlled trials evaluating the effect of interventions on St. George's Respiratory Questionnaire in patients with chronic obstructive pulmonary disease at the end of the intervention [29,36,40]. The analysis was performed using a random effects model. CBPR: center-based pulmonary rehabilitation; HKSJ: Hartung-Knapp-Sidik-Jonkman; MD: mean difference; Tele-PR: pulmonary telerehabilitation.

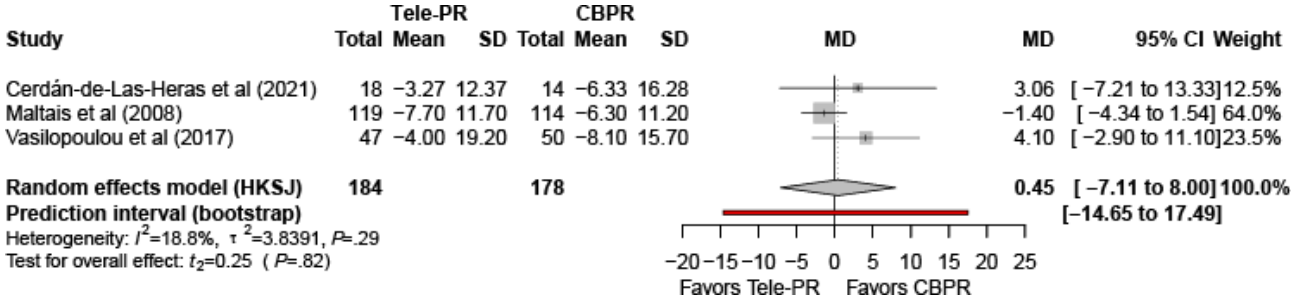

Figure 9. Forest plot of randomized controlled trials evaluating the effect of interventions on St. George's Respiratory Questionnaire in patients with chronic obstructive pulmonary disease at long-term follow-up (≥6 mo) [29,36,40]. The analysis was performed using a random effects model. CBPR: center-based pulmonary rehabilitation; HKSJ: Hartung-Knapp-Sidik-Jonkman; MD: mean difference; Tele-PR: pulmonary telerehabilitation.

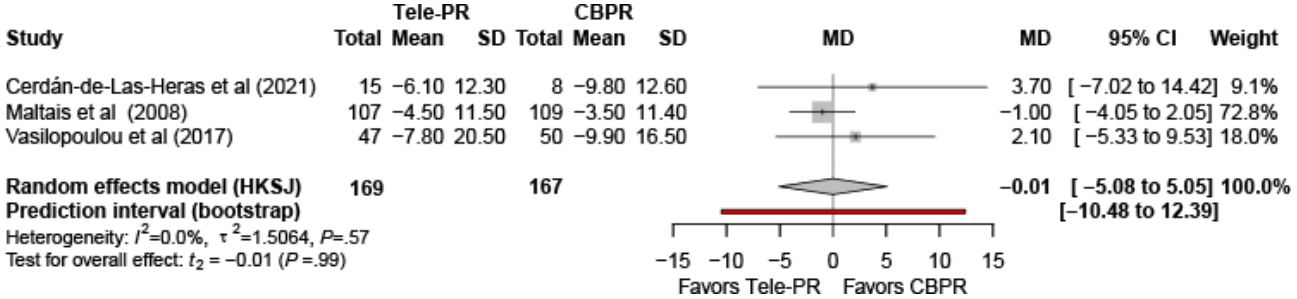

COPD Assessment Test

Three trials [31,40,43] reported CAT (chronic obstructive pulmonary disease assessment test) scores at the end of the intervention. Pooled analysis showed no statistically significant difference between Tele-PR and CBPR (k=3; n=298, 18.0%; MD 0.24, 95% CI –8.44 to 8.92; P=.91; τ^2=10.42; I^2=89.1%; PI=–16.98 to 17.50; Figure 10).

At long-term follow-up [31,32,40,43], pooled analysis showed no statistically significant overall difference (k=4; n=374, 22.6%; MD 1.07, 95% CI –2.89 to 5.20; P=.45; τ^2=4.54; I^2=75.4%; PI=–7.63 to 9.54; Figure 11).

Figure 10. Forest plot of randomized controlled trials evaluating the effect of interventions on chronic obstructive pulmonary disease (COPD) assessment test in patients with COPD at the end of the intervention [31,40,43]. The analysis was performed using a random effects model. CBPR: center-based pulmonary rehabilitation; HKSJ: Hartung-Knapp-Sidik-Jonkman; MD: mean difference; Tele-PR: pulmonary telerehabilitation.

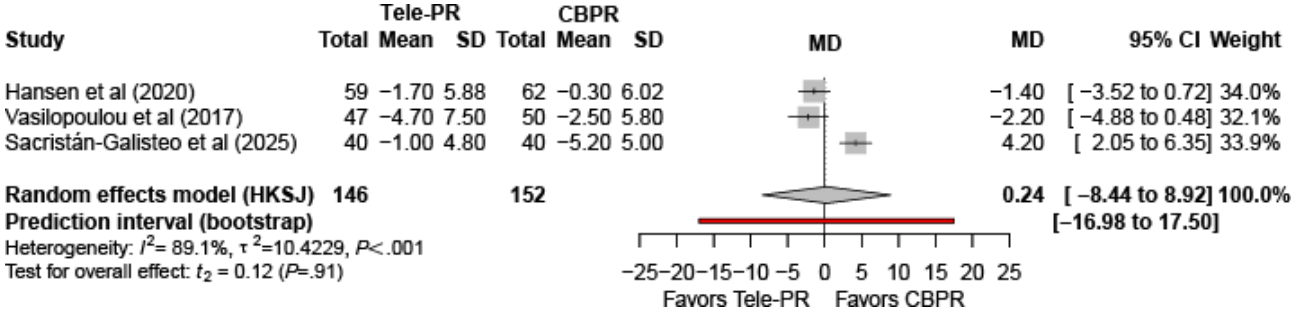

Figure 11. Forest plot of randomized controlled trials evaluating the effect of interventions on chronic obstructive pulmonary disease assessment test in patients with COPD at long-term follow-up (≥6 mo) [31,32,40,43]. The analysis was performed using a random effects model. CBPR: center-based pulmonary rehabilitation; HKSJ: Hartung-Knapp-Sidik-Jonkman; MD: mean difference; Tele-PR: pulmonary telerehabilitation.

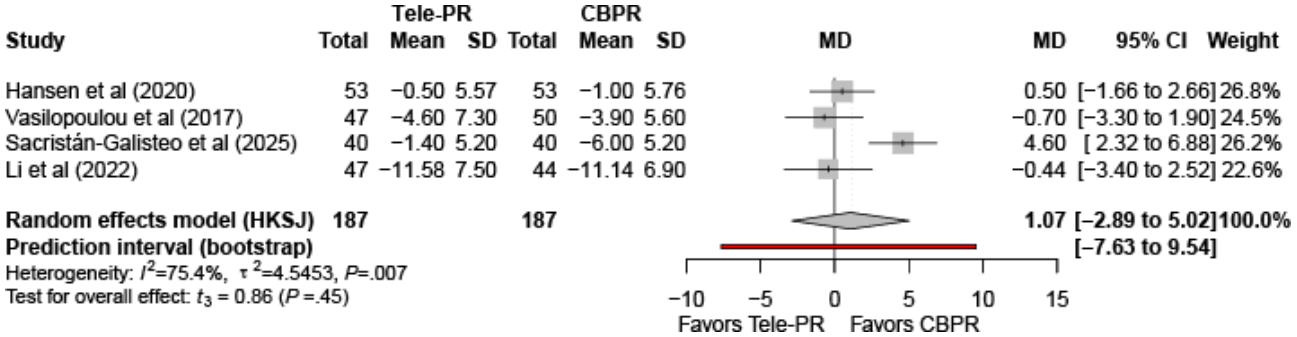

XSL•FO
RenderX

Chronic Respiratory Questionnaire–Dyspnea

Six trials [27,33,36,37,39,41] reported Chronic Respiratory Questionnaire–Dyspnea outcomes at the end of the intervention. Pooled analysis showed no statistically significant difference between Tele-PR and CBPR (k=6; n=791, 47.7%; MD 0.10, 95% CI –0.39 to 0.60; P=.62; τ^2=0.22; I^2=24.8%; PI=–0.75 to 0.97; Figure 12).

At long-term follow-up [33,36,37,39,41], pooled analysis also showed no statistically significant difference (k=5; n=686, 41.4%; MD 0.15, 95% CI –0.25 to 0.54; P=.34; τ^2=0.07; I^2=0%; PI=-0.79 to 0.96; Figure 13).

Figure 12. Forest plot of randomized controlled trials evaluating the effect of interventions on Chronic Respiratory Questionnaire–Dyspnea domain in patients with chronic obstructive pulmonary disease at the end of the intervention [27,33,36,37,39,41]. The analysis was performed using a random effects model. CBPR: center-based pulmonary rehabilitation; HKSJ: Hartung-Knapp-Sidik-Jonkman; MD: mean difference; Tele-PR: pulmonary telerehabilitation.

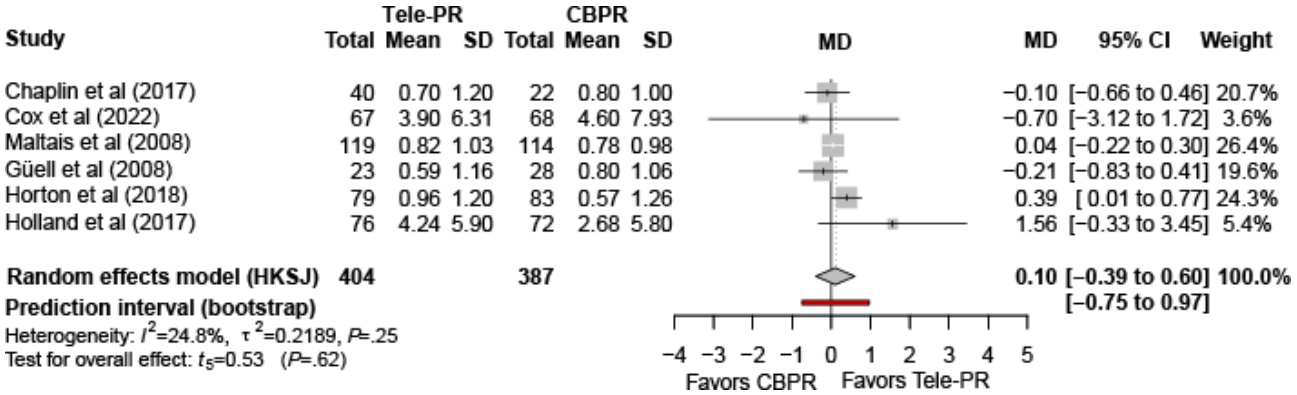

Figure 13. Forest plot of randomized controlled trials evaluating the effect of interventions on Chronic Respiratory Questionnaire–Dyspnea domain in patients with chronic obstructive pulmonary disease at long-term follow-up (≥6 mo) [33,36,37,39,41]. The analysis was performed using a random effects model. CBPR: center-based pulmonary rehabilitation; HKSJ: Hartung-Knapp-Sidik-Jonkman; MD: mean difference; Tele-PR: pulmonary telerehabilitation.

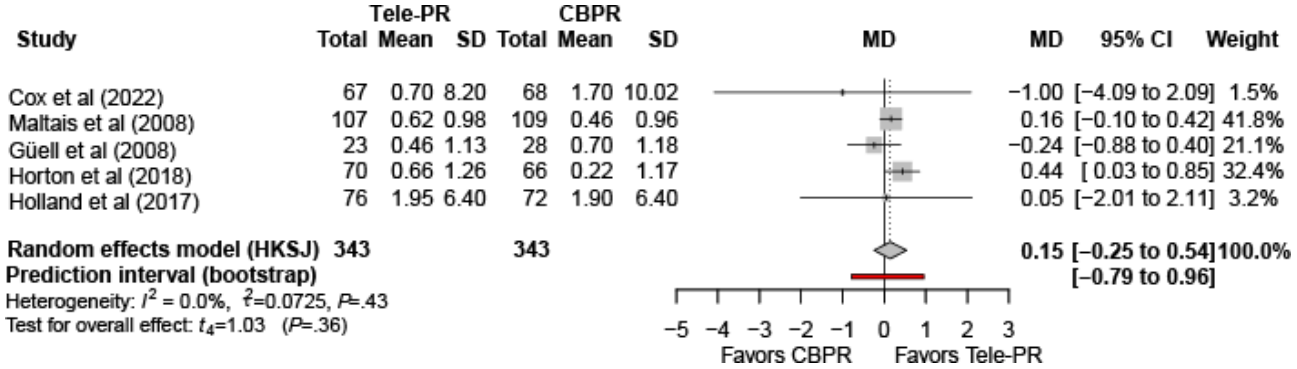

Hospital Anxiety and Depression Scale–Anxiety

Three trials [31,33,43] reported Hospital Anxiety and Depression Scale–Anxiety outcomes at the end of the intervention. Pooled analysis showed no statistically significant difference between Tele-PR and CBPR (k=3; n=325, 19.6%; MD –0.43, 95% CI –2.41 to 1.55; P=.45; τ^2=0.29; I^2=38.3%; PI=–4.83 to 3.93; Figure 14).

At long-term follow-up [31,33,43], pooled analysis showed no statistically significant overall difference (Figure 15).

Figure 14. Forest plot of randomized controlled trials evaluating the effect of interventions on Hospital Anxiety and Depression Scale–Anxiety subscale in patients with chronic obstructive pulmonary disease at the end of intervention [31,33,43]. The analysis was performed using a random effects model. CBPR: center-based pulmonary rehabilitation; HKSJ: Hartung-Knapp-Sidik-Jonkman; MD: mean difference; Tele-PR: pulmonary telerehabilitation.

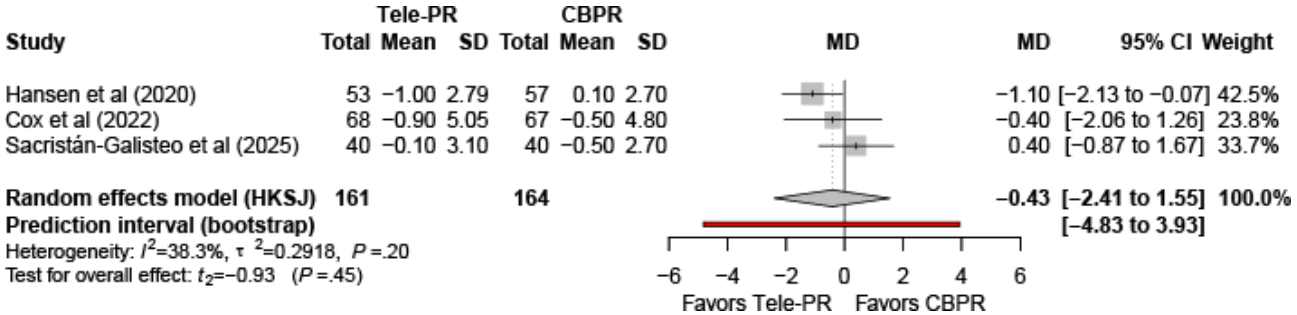

XSL•FO
RenderX

Figure 15. Forest plot of randomized controlled trials evaluating the effect of interventions on Hospital Anxiety and Depression Scale–Anxiety subscale in patients with chronic obstructive pulmonary disease at long-term follow-up (≥6 mo) [31,33,43]. The analysis was performed using a random effects model. CBPR: center-based pulmonary rehabilitation; HKSJ: Hartung-Knapp-Sidik-Jonkman; MD: mean difference; Tele-PR: pulmonary telerehabilitation.

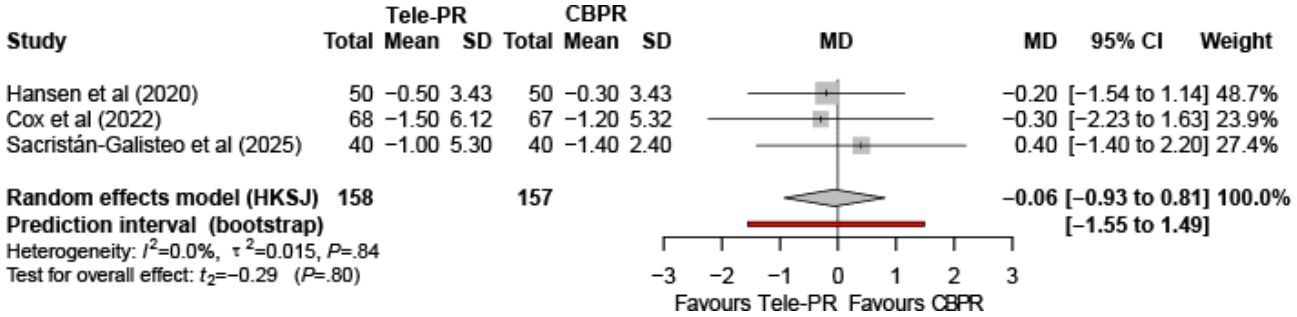

Hospital Anxiety and Depression Scale–Depression

Three trials [31,33,43] reported Hospital Anxiety and Depression Scale–Depression outcomes at the end of the intervention. Pooled analysis showed no statistically significant difference between Tele-PR and CBPR (k=3; n=325, 19.6%; MD 0.03, 95% CI –1.92 to 1.97; P=.96; τ^2=0.30; I^2=39.5%; PI =–4.01 to 4.13; Figure 16).

Figure 16. Forest plot of randomized controlled trials evaluating the effect of interventions on Hospital Anxiety and Depression Scale–Depression subscale in patients with chronic obstructive pulmonary disease at the end of intervention [31,33,43]. The analysis was performed using a random effects model. CBPR: center-based pulmonary rehabilitation; HKSJ: Hartung-Knapp-Sidik-Jonkman; MD: mean difference; Tele-PR: pulmonary telerehabilitation.

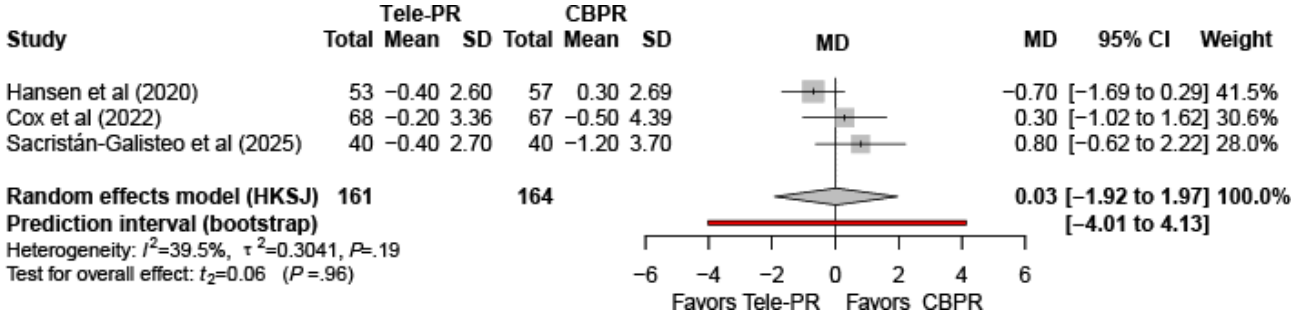

At long-term follow-up [31,33,43], pooled analysis likewise showed no statistically significant overall difference (Figure 17).

Figure 17. Forest plot of randomized controlled trials evaluating the effect of interventions on Hospital Anxiety and Depression Scale–Depression subscale in patients with chronic obstructive pulmonary disease at long-term follow-up (≥6 mo) [31,33,43]. The analysis was performed using a random effects model. CBPR: center-based pulmonary rehabilitation; HKSJ: Hartung-Knapp-Sidik-Jonkman; MD: mean difference; Tele-PR: pulmonary telerehabilitation.

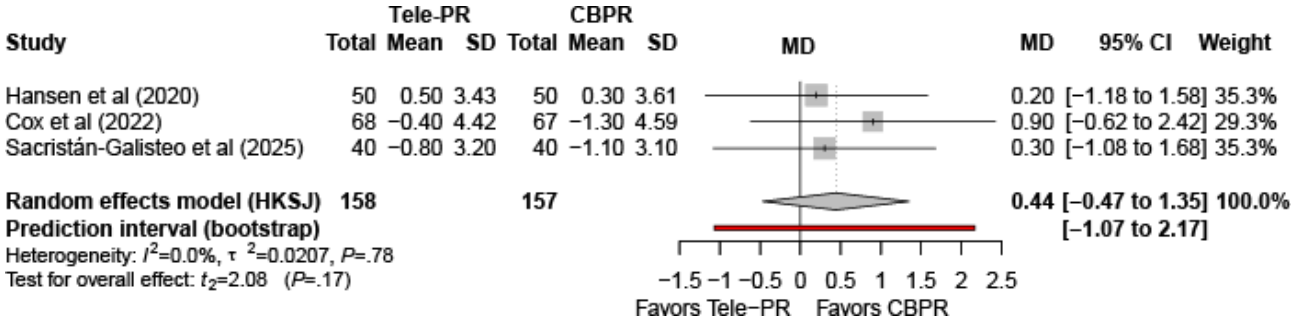

Modified Medical Research Council Questionnaire

Three trials [33,40,41] reported modified Medical Research Council questionnaire outcomes at the end of intervention. Pooled analysis showed no statistically significant difference between Tele-PR and CBPR (k=3; n=380, 22.9%; MD 0.09, 95% CI –0.78 to 0.95; P=.71; τ^2=0.09; I^2=74.8%; PI=–1.76 to 1.88; Figure 18).

At long-term follow-up [32,33,40,41], pooled analysis showed no statistically significant overall difference (Figure 19).

XSL•FO
RenderX

Figure 18. Forest plot of randomized controlled trials evaluating the effect of interventions on the modified Medical Research Council in patients with chronic obstructive pulmonary disease at the end of intervention [33,40,41]. The analysis was performed using a random effects model. CBPR: center-based pulmonary rehabilitation; HKSJ, Hartung-Knapp-Sidik-Jonkman; MD: mean difference; Tele-PR: pulmonary telerehabilitation.

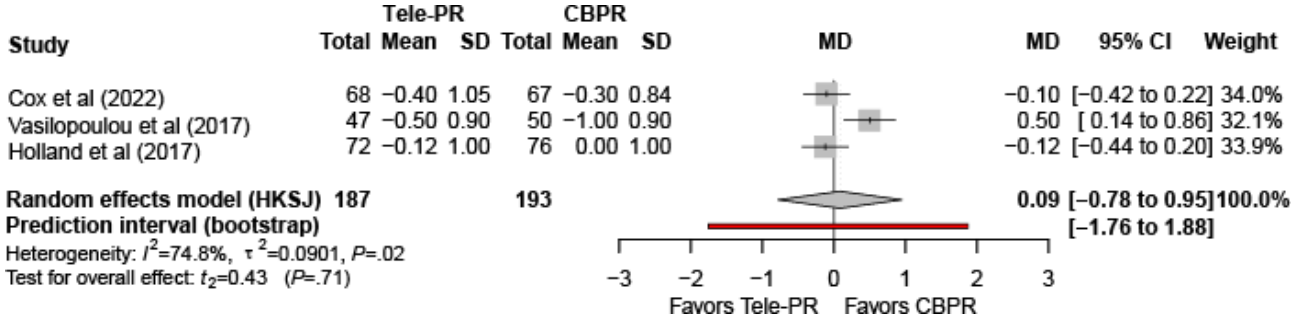

Figure 19. Forest plot of randomized controlled trials evaluating the effect of interventions on the modified Medical Research Council in patients with chronic obstructive pulmonary disease at long-term follow-up (≥6 mo) [32,33,40,41]. The analysis was performed using a random effects model. CBPR: center-based pulmonary rehabilitation; HKSJ: Hartung-Knapp-Sidik-Jonkman; MD: mean difference; Tele-PR: pulmonary telerehabilitation.

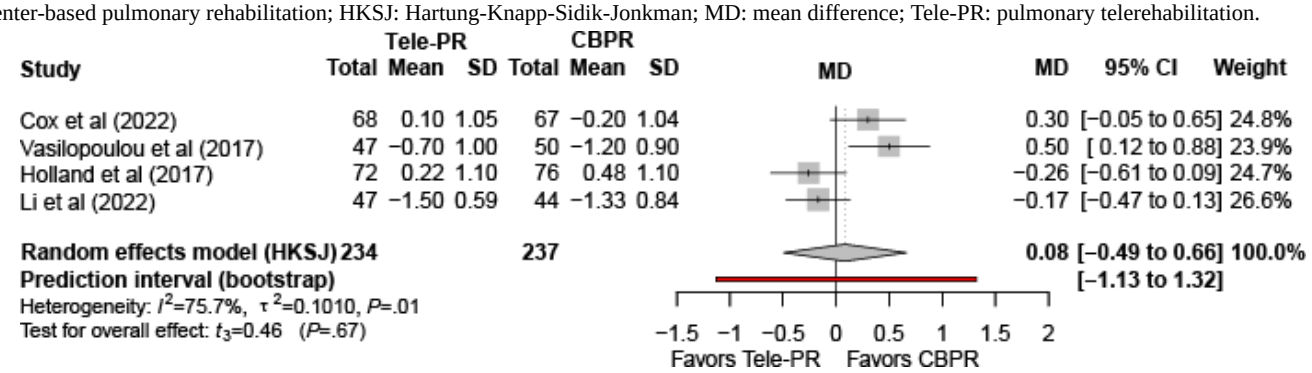

Pulmonary Rehabilitation Adapted Index of Self-Efficacy

Three trials [33,39,41] reported Pulmonary Rehabilitation Adapted Index of Self-Efficacy outcomes at the end of the intervention. Pooled analysis showed no statistically significant difference between Tele-PR and CBPR (k=3; n=448, 27.0%; MD –0.83, 95% CI –9.29 to 7.64; P=.71; τ^2=9.74; I^2=86.7%; PI=–16.96 to 15.44; Figure 20).

At long-term follow-up [33,39,41], pooled analysis likewise showed no statistically significant overall difference (Figure 21).

Figure 20. Forest plot of randomized controlled trials evaluating the effect of interventions on Pulmonary Rehabilitation Adapted Index of Self-Efficacy in patients with COPD at the end of intervention [33,39,41]. The analysis was performed using a random effects model. CBPR: center-based pulmonary rehabilitation; HKSJ: Hartung-Knapp-Sidik-Jonkman; MD: mean difference; Tele-PR: pulmonary telerehabilitation.

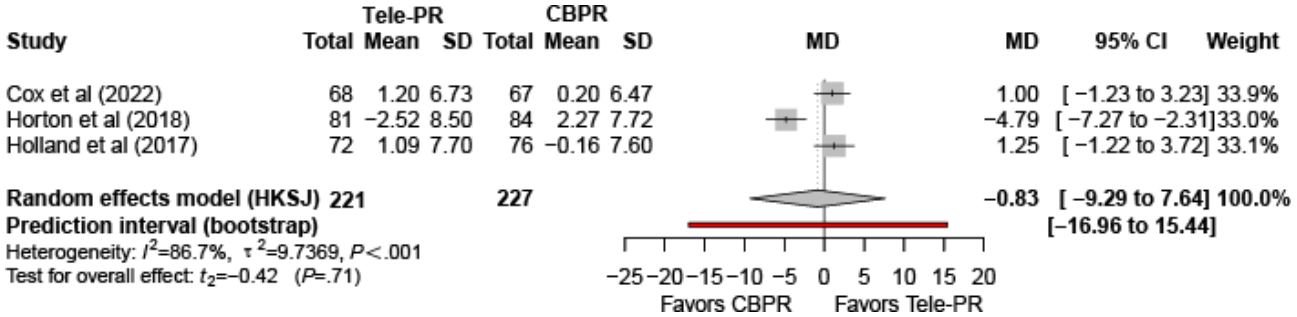

XSL•FO
RenderX

Figure 21. Forest plot of randomized controlled trials evaluating the effect of interventions on Pulmonary Rehabilitation Adapted Index of Self-Efficacy in patients with chronic obstructive pulmonary disease at long-term follow-up (≥6 mo) [33,39,41]. The analysis was performed using a random effects model. CBPR: center-based pulmonary rehabilitation; HKSJ: Hartung-Knapp-Sidik-Jonkman; MD: mean difference; Tele-PR: pulmonary telerehabilitation.

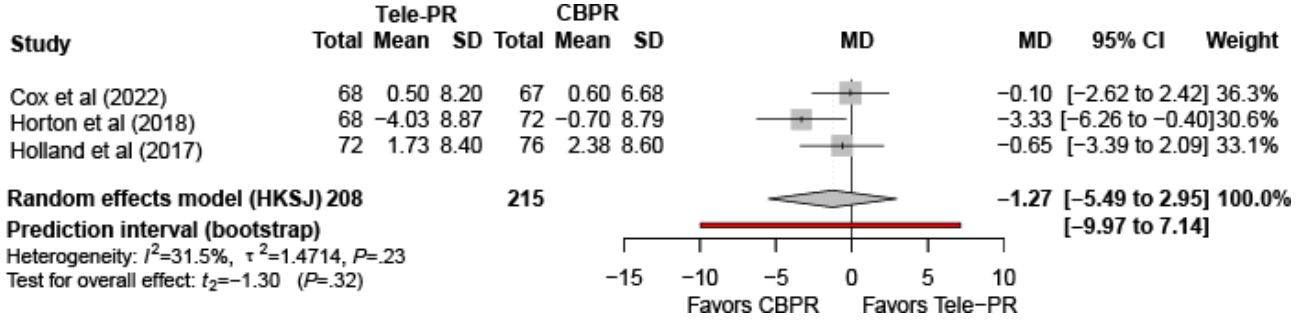

Dropout Rate

Ten trials [27,29,31-33,36,39,41-43] reported end of intervention dropout data. Pooled random effects meta-analysis with HKSJ adjustment showed no statistically significant difference in dropout rate between Tele-PR and CBPR (k=10; n=1404, 84.7%; risk ratio 0.66, 95% CI 0.40-1.07; P=.08; τ^2=0.33; I^2=76.4%; PI=0.15-3.01; Figure 22).

Meta-regression analyses indicated that intervention duration, training intensity, publication year, baseline FEV_1 (%), and economic status were not statistically significantly associated with dropout rate (all P>.05; Table 4 and Multimedia Appendix 7).

At long-term follow-up, pooled analysis of 4 studies [32,36,40,43] showed no statistically significant difference in dropout rate (Figure 23).

Figure 22. Forest plot of randomized controlled trials evaluating the effect of interventions assessing dropout rates in patients with chronic obstructive pulmonary disease at the end of intervention [27,29,31-33,36,39,41-43]. The analysis was performed using a random effects model. CBPR: center-based pulmonary rehabilitation; HKSJ: Hartung-Knapp-Sidik-Jonkman; RR: risk ratio; Tele-PR: pulmonary telerehabilitation.

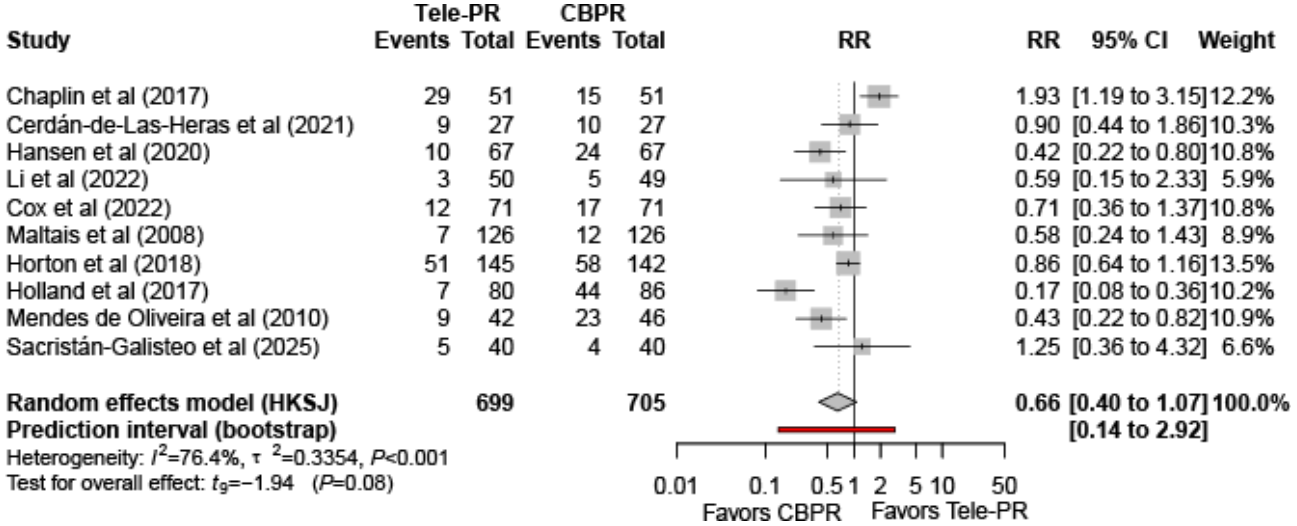

Table . Meta-regression results of examining covariates influencing end of intervention dropout rate.

Covariate	β coefficient (SE)	t (df)	P value
Intervention duration (weeks)	0.08 (0.12)	0.68 (8)	.52
Training intensity (medium vs high)	0.04 (0.55)	0.07 (8)	.94
Economic status (developing vs developed)	−0.38 (0.60)	−0.64 (8)	.54
Publication year	0.03 (0.05)	0.76 (8)	.47
Baseline FEV_1[a] (%)	0.02 (0.03)	0.85 (8)	.42

[a]FEV_1: forced expiratory volume in 1 second.

XSL•FO
RenderX

Figure 23. Forest plot of randomized controlled trials evaluating the effect of interventions assessing dropout rates in patients with chronic obstructive pulmonary disease at long-term follow-up (≥6 mo) [32,36,40,43]. The analysis was performed using a random effects model. CBPR: center-based pulmonary rehabilitation; HKSJ: Hartung-Knapp-Sidik-Jonkman; RR: risk ratio; Tele-PR: pulmonary telerehabilitation.

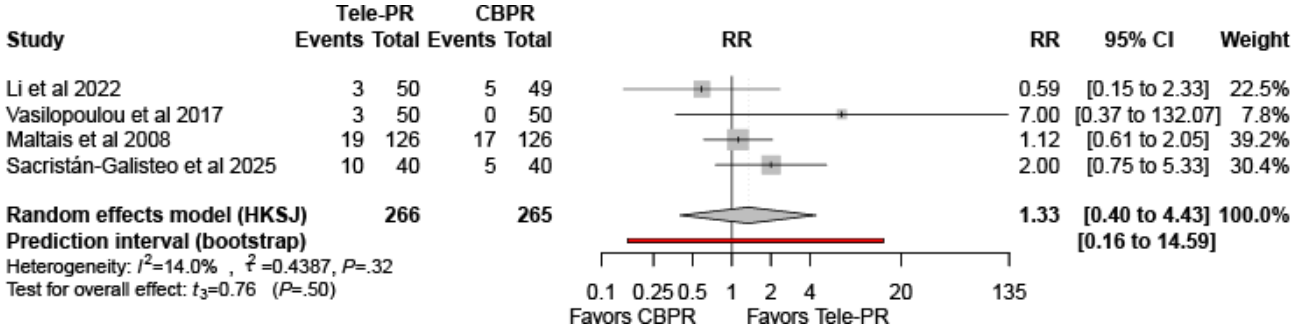

Endurance Shuttle Walk Test

Two trials [27,39] Endurance Shuttle Walk Test outcomes. In a study by Horton et al [39], both home-based and center-based PR demonstrated clinically meaningful improvements exceeding the minimal clinically important difference (mean change 212 s vs 353 s, respectively). The between-group difference was −141 seconds (95% CI −252 to −31; *P*=.01). In a study by Chaplin et al [27], significant within-group improvements were observed in both the web-based (mean change 189 s) and conventional PR groups (mean change 185 s), with no significant between-group difference.

Willingness to Pay and Direct Costs

Willingness to pay did not differ meaningfully between Tele-PR and CBPR, with reported means of approximately US $175 to $176 and medians of US $83 to $100, and no statistically significant difference was detected (*P*=.98) [28]. In 1 study [41], the estimated cost of Tele-PR was AUD 298 (US $224), which was slightly lower than that of CBPR at AUD 312 (US $234), corresponding to a difference of AUD 14 (US $11).

Adverse Events

Eight (47.1%) [29,31,33,36,39-41,43] of 17 included trials reported adverse events or clinical safety outcomes, including serious adverse events, hospitalizations, acute exacerbations of COPD, or deaths. Among studies that explicitly monitored serious adverse events, no intervention-related serious adverse events were identified [31,33,36,39,41]. Event rates, where reported, were generally similar between telerehabilitation and center-based rehabilitation [31,33,36,40]. However, harms were not consistently prespecified or systematically reported across trials, and the scope and definitions of adverse events varied substantially.

Subgroup Analysis

Subgroup analyses were conducted according to delivery models, supervision intensity, and supervision modality (Multimedia Appendix 7). For 6MWD at both the end of intervention and long-term follow-up (≥6 mo), no statistically significant subgroup differences were observed across delivery models, supervision intensity, or supervision modality (all *P*>.05), although moderate heterogeneity was noted in some low-technology HBPR strata. For daily steps and Chronic Respiratory Questionnaire–Dyspnea (CRQ-D) domain, no significant subgroup effects by delivery models were identified at either time point (all *P*>.05). In contrast, for the CAT, delivery models showed significant subgroup differences at both end of intervention and long-term follow-up (both *P*<.001), with digitally supported Tele-PR demonstrating small, nonsignificant changes, whereas low-technology HBPR showed worsening CAT scores in single-study strata. For the dropout rate, no significant subgroup differences were detected by delivery models or supervision modality (all *P*>.05), although heterogeneity was substantial in several subgroups. Overall, subgroup effects were generally limited, except for CAT, and several strata were constrained by small numbers of studies.

Sensitivity Analysis and Publication Bias Test

Leave-one-out sensitivity analyses were conducted for all primary outcomes to assess the robustness of pooled estimates and to identify influential studies contributing to between-study heterogeneity (Multimedia Appendix 8). Overall, pooled estimates for several outcomes showed sensitivity to the exclusion of individual studies, indicating that some findings were influenced by specific trial characteristics.

For 6MWD, exclusion of a single telephone-based study [40] resulted in a shift of the pooled estimate toward CBPR (MD –8.63 m; *P*=.045), indicating that the overall effect estimate was sensitive to this study. For daily step counts, the marginally significant pooled effect was driven primarily by 1 study [35]; removal of this study rendered the pooled effect nonsignificant (*P*=.12) and substantially reduced heterogeneity.

For patient-reported outcomes, substantial heterogeneity in health status measured by the CAT was attributable to a single study [43]; exclusion of this study eliminated heterogeneity and attenuated the pooled effect. Similarly, heterogeneity in self-efficacy outcomes was largely driven by 1 study [39], and removal of this study reduced between-study variance and altered the pooled estimate.

In contrast, pooled estimates for modified Medical Research Council Dyspnea score, depressive symptoms, and dropout rate were robust to leave-one-out analyses. Although individual studies [34,41] showed effects in opposite directions for dropout, the overall conclusion of no statistically significant difference between groups remained unchanged across all sensitivity analyses.

Visual inspection of the funnel plot for dropout rate revealed a relatively symmetrical distribution of studies (Figure 24). This was confirmed by Peters test, which showed no evidence of

XSL•FO
RenderX

significant publication bias or small-study effects ($t(8)$=.43; P=.68). However, given the limited number of included studies, the results of publication bias tests should be interpreted cautiously.

Figure 24. Contour-enhanced funnel plot for the assessment of publication bias in dropout rates [27,29,31-33,36,39,41-43]. Each dot represents an individual study included in the meta-analysis. The vertical axis represents the SE (a measure of study precision), and the horizontal axis represents the risk ratio (log scale). The shaded regions indicate statistical significance levels (P<.10, P<.05, and P<.01), while the white area corresponds to nonsignificant results (P>.10).

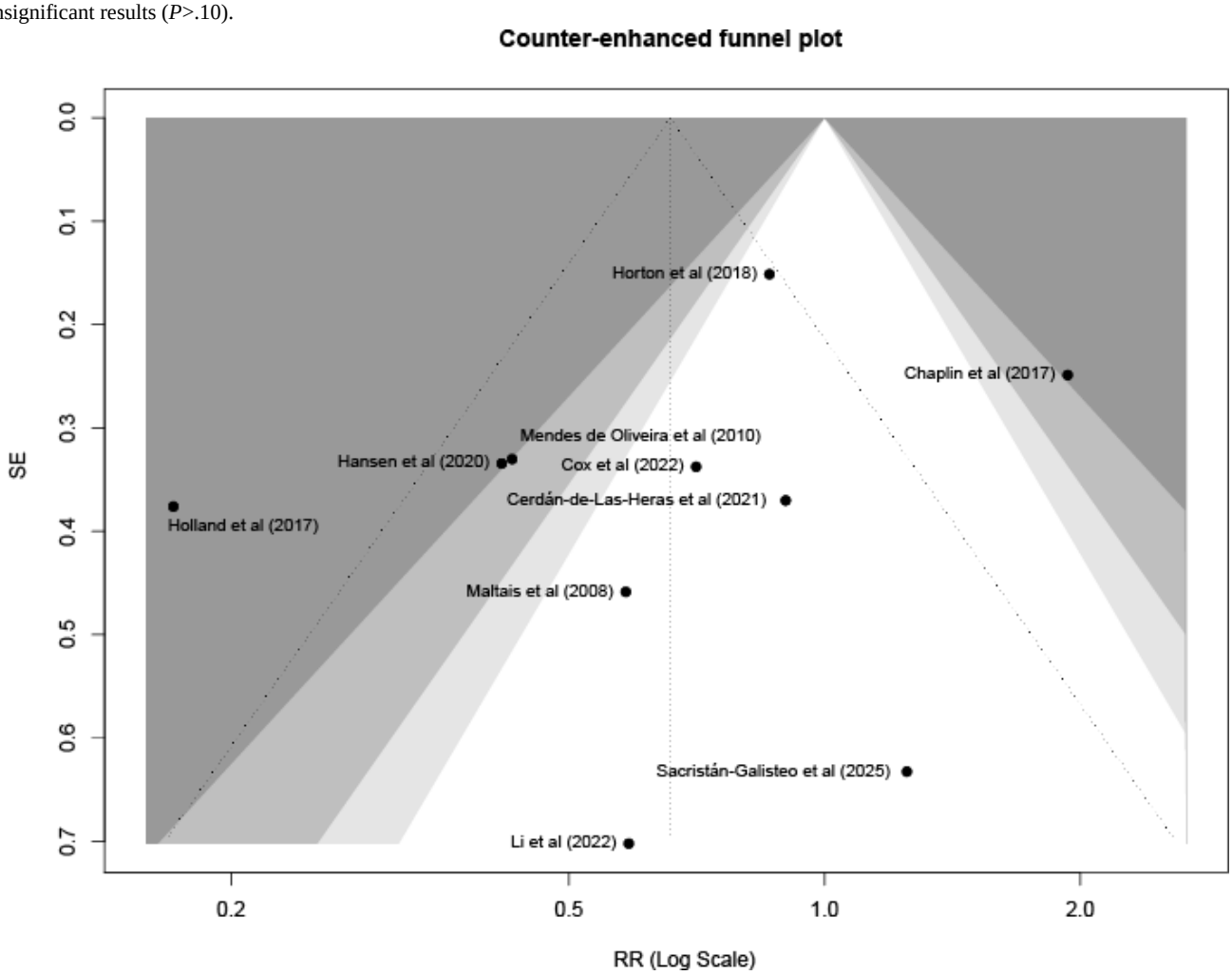

GRADE Assessment

The certainty of evidence ranged from moderate to very low across outcomes. Evidence for 6MWD was rated as low certainty and was downgraded by 1 level for risk of bias because blinding was not feasible and attrition occurred in several trials. Evidence for dropout rate was rated as very low certainty and was downgraded for risk of bias, substantial inconsistency, and imprecision, reflected by wide CIs and PIs. Health status outcomes (CAT, SGRQ, and Chronic Respiratory Questionnaire–Dyspnea) were rated as low certainty and were downgraded for inconsistency, likely related to heterogeneous intervention modalities, and for imprecision. These certainty ratings should temper interpretation of pooled estimates and underscore variability in real-world effects. The certainty of evidence assessed using the GRADE approach is summarized in Table 5.

XSL•FO
RenderX

Table . Summary of findings and certainty of evidence (GRADE) for tele-pulmonary rehabilitation compared with center-based pulmonary rehabilitation.

Certainty assessment							Patients, n		Effect		Certainty	Importance
Studies (n)	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Intervention	Comparison	Relative (95% CI)	Absolute (95% CI)		
6-Minute walk distance (end of intervention)												
9	Randomized trials	Serious[a]	Serious[b]	Not serious	Not serious	None	476	474	—[c]	MD[d] −5.37 m (−15.68 to 4.95)	Low[a,b]	Critical
6-Minute walk distance (long-term follow-up [≥6 mo])												
9	Randomized trials	Serious[a]	Serious[e]	Not serious	Serious[f]	None	475	473	—	MD 2.97 m (−12.22 to 18.17)	Very low[a,e]	Critical
Dropout rate (end of intervention)												
10	Randomized trials	Serious[a]	Serious[g]	Not serious	Serious[h]	None	142/699 (20.3%)	212/705 (30.1%)	RR[i] 0.66 (0.40 to 1.07)	102 fewer per 1000 (from −180 to 21)	Very low[a,g,h]	Critical
Dropout rate (long-term follow-up [≥6 mo])												
4	Randomized trials	Serious[a]	Not serious	Not serious	Very serious[j]	None	35/266 (13.2%)	27/265 (10.2%)	RR 1.33 (0.40 to 4.43)	34 more per 1000 (from 61 to 349)	Very low[a,j]	Critical
COPD assessment test (end of intervention)												
3	Randomized trials	Serious[a]	Serious[k]	Not serious	Serious[l]	None	146	152	—	MD 0.24 (−8.44 to 8.92)	Very low[a,k,l]	Important
COPD assessment test (long-term follow-up [≥6 mo])												
4	Randomized trials	Serious[a]	Serious[m]	Not serious	Serious[n]	None	187	187	—	MD 1.07 (−2.89 to 5.2)	Very low[a,m,n]	Important
Daily steps (end of intervention)												
5	Randomized trials	Serious[a]	Serious[o]	Not serious	Serious[p]	None	119	134	—	MD 4.97 (−1.84 to 11.78)	Very low[a,o,p]	Important
Daily steps (long-term follow-up [≥6 mo])												
3	Randomized trials	Serious[a]	Not serious	Not serious	Very serious[q]	None	69	73	—	MD −1.01 (−6.27 to 4.2)	Very low[a,q]	Important
St. George's Respiratory Questionnaire (end of intervention)												
3	Randomized trials	Serious[a]	Serious[r]	Not serious	Serious[s]	None	184	178	—	MD 0.45 (−7.11 to 8)	Very low[a,r,s]	Important
St. George's Respiratory Questionnaire (long-term follow-up [≥6 mo])												
3	Randomized trials	Serious[a]	Not serious	Not serious	Serious[t]	None	169	167	—	MD −0.01 (−5.08 to 5.05)	Low[a,t]	Important
Chronic Respiratory Questionnaire–Dyspnea (end of intervention)												

XSL•FO
RenderX

Certainty assessment							Patients, n		Effect		Certainty	Importance
Studies (n)	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Intervention	Comparison	Relative (95% CI)	Absolute (95% CI)		
6	Randomized trials	Serious[a]	Not serious	Not serious	Not serious	None	404	387	—	MD 0.1 (−0.39 to 0.6)	Moderate[a]	Important
Chronic Respiratory Questionnaire–Dyspnea (long-term follow-up [≥6 mo])												
5	Randomized trials	Serious[a]	Not serious	Not serious	Serious[u]	None	343	343	—	MD 0.15 (−0.25 to 0.54)	Low[a,u]	Important
Hospital Anxiety and Depression Scale–Anxiety (end of intervention)												
3	Randomized trials	Serious[a]	Not serious	Not serious	Serious[v]	None	161	164	—	MD −0.43 (−2.41 to 1.55)	Low[a,v]	Important
Hospital Anxiety and Depression Scale–Anxiety (long-term follow-up [≥6 mo])												
3	Randomized trials	Serious[a]	Not serious	Not serious	Serious[w]	None	158	157	—	MD −0.06 (−0.93 to 0.81)	Low[a,w]	Important
Hospital Anxiety and Depression Scale–Depression (end of intervention)												
3	Randomized trials	Serious[a]	Not serious	Not serious	Serious[v]	None	161	164	—	MD 0.03 (−1.92 to 1.97)	Low[a,v]	Important
Hospital Anxiety and Depression Scale–Depression (long-term follow-up [≥6 mo])												
3	Randomized trials	Serious[a]	Not serious	Not serious	Serious[x]	None	158	157	—	MD **0.44** (−0.47 to 1.35)	Low[a,x]	Important
Modified Medical Research Council Questionnaire (end of intervention)												
3	Randomized trials	Serious[a]	Serious[y]	Not serious	Not serious	None	187	193	—	MD 0.09 (−0.78 to 0.95)	Low[a,y]	Important
Modified Medical Research Council Questionnaire (long-term follow-up [≥6 mo])												
4	Randomized trials	Serious[a]	Serious[z]	Not serious	Serious[aa]	None	234	237	—	MD 0.08 (−0.49 to 0.6)	Very low[a,z,aa]	Important
PR Adapted Index of Self-Efficacy (end of intervention)												
3	Randomized trials	Serious[a]	Serious[ab]	Not serious	Serious[ac]	None	221	227	—	MD −0.83 (−9.29 to 7.6)	Very low[a,ab,ac]	Important
PR Adapted Index of Self-Efficacy (long-term follow-up [≥6 mo])												

XSL•FO
RenderX

Certainty assessment							Patients, n		Effect		Certainty	Importance
Studies (n)	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Intervention	Comparison	Relative (95% CI)	Absolute (95% CI)		
3	Randomized trials	Serious[a]	Not serious	Not serious	Serious[ad]	None	206	215	—	MD −1.27 (−5.49 to 2.95)	Low[a,ad]	Important

[a]Most studies had a high risk of performance bias due to lack of blinding of participants and personnel, which is inherent to exercise interventions.

[b]Downgraded for inconsistency due to model-level variability and prediction intervals spanning clinically important benefit and harm across settings.

[c]Not applicable.

[d]MD: mean difference.

[e]Downgraded 1 level due to moderate heterogeneity (I^2=44.5%) and variability in long-term adherence across studies.

[f]Although the CI crosses the line of no effect, it excludes the minimal clinically important difference of 30 m, suggesting no clinically meaningful difference.

[g]Downgraded 1 level due to extreme heterogeneity (I^2=86%) with conflicting directions of effect.

[h]Downgraded 1 level due to wide CIs (0.40-1.07) crossing the line of no effect.

[i]RR: risk ratio.

[j]Downgraded 2 levels due to very wide CIs that cross the line of no effect and include potential for both substantial benefit and substantial harm.

[k]Substantial heterogeneity was observed across studies, with inconsistent effect estimates across intervention models.

[l]CIs include both benefit and no effect and do not consistently exceed the minimal clinically important difference.

[m]Downgraded 1 level due to high heterogeneity (I^2=75%) with conflicting directions of effect across studies.

[n]Downgraded 1 level due to wide CIs that cross the line of no effect and include the minimal clinically important difference of 2 points for potential harm.

[o]Although statistical heterogeneity was low, effect estimates were sensitive to individual studies and varied across intervention models.

[p]CIs include both increased and decreased daily step counts and do not exclude no effect.

[q]Downgraded 2 levels due to very small sample size (n=142, below the optimal information size) and wide CIs that cross the line of no effect.

[r]inconsistent effect estimates across different intervention models.

[s]CIs include both benefit and no effect and do not consistently exceed the minimal clinically important difference.

[t]Downgraded 1 level due to CIs crossing the line of no effect and including the minimal clinically important difference of 4 units in both directions.

[u]Downgraded 1 level due to CIs crossing the line of no effect and including the minimal clinically important difference of 0.5 points in both directions.

[v]CIs included both benefit and no effect, and the total sample size was limited, resulting in imprecision.

[w]Downgraded 1 level due to limited sample size (n=315, below optimal information size) and CIs crossing the line of no effect.

[x]Downgraded 1 level due to limited sample size (n=315, below optimal information size) and CIs crossing the line of no effect.

[y]Substantial heterogeneity was observed across studies, with inconsistent effect estimates.

[z]Downgraded 1 level due to high heterogeneity with conflicting directions of effect.

[aa]Downgraded 1 level due to wide CIs that cross the line of no effect and include the minimal clinically important difference of 0.5 points in both directions.

[ab]Substantial heterogeneity was observed across studies, with highly variable effect estimates depending on intervention modality.

[ac]CIs were wide and crossed the line of no effect, and no established minimal clinically important difference is available for Pulmonary Rehabilitation Adapted Index of Self-Efficacy.

[ad]Downgraded 1 level due to wide CIs crossing the line of no effect.

Discussion

Principal Findings

This systematic review and meta-analysis evaluated the relative effectiveness of Tele-PR compared to CBPR in patients with COPD. Overall, no statistically significant differences were observed at the end of the intervention across key outcomes, including functional capacity, dyspnea, and health-related quality of life. These findings indicate that, on average, Tele-PR may achieve short-term clinical effects similar to those of CBPR when implemented under structured and well-defined conditions. However, subsequent subgroup analyses and investigations of heterogeneity suggest that Tele-PR should not be regarded as a single homogeneous intervention [33]. Substantial variation was observed in effect consistency and stability across different remote delivery models, particularly with respect to supervision intensity and interaction modality. Accordingly, interpretations of “equivalence” between Tele-PR and CBPR based solely on pooled average effects should be made with caution, as such summaries may obscure meaningful differences across implementation contexts [19].

Long-Term Efficacy and Maintenance Rehabilitation

In this analysis, differences in outcomes between rehabilitation modalities diminished during postintervention follow-up, indicating a time-dependent convergence. This pattern suggests that, regardless of delivery format, remote and center-based PR

XSL•FO
RenderX

programs face challenges in sustaining long-term benefits [3]. Several factors may contribute to this attenuation of effects, although these potential explanatory variables were not directly assessed in this meta-analysis. First, adherence to regular exercise training often declines after structured supervision ends, particularly when ongoing support is limited [44,45]. Second, longer follow-up periods may include seasonal variations, acute exacerbations, or intercurrent illnesses, which can disrupt training continuity and negatively impact functional outcomes [46,47]. Third, many existing Tele-PR programs primarily focus on exercise training, while systematic reinforcement of relapse prevention and long-term self-management skills remains relatively underdeveloped [48]. Importantly, this convergence of outcomes does not imply that early rehabilitation effects lack clinical significance. Rather, it underscores the need to conceptualize PR—whether delivered remotely or in person—as an ongoing process that requires sustained behavioral and educational support, rather than as a time-limited intervention. This perspective aligns with the concept of maintenance rehabilitation emphasized in recent clinical guidelines and supports the future integration of educational components, behavioral maintenance strategies, and digital platforms to improve long-term effectiveness [49-51].

Physiological Mechanisms and Safety Profile

The included studies predominantly enrolled patients with moderate-to-severe COPD, a population in whom home-based training has traditionally raised safety concerns [3]. Evidence suggests that, under appropriate screening and standardized implementation, Tele-PR was not associated with increased short-term serious adverse events and demonstrated a safety profile comparable to that of CBPR [31,33,36,39,41]. Functional improvements likely extend beyond aerobic enhancement, as PR improves respiratory mechanics and ventilatory efficiency [14,52-54]. Although physiological markers were not directly synthesized, dyspnea improvements in Tele-PR suggest overlapping mechanisms with conventional PR [55]. Emerging evidence of improved gas exchange in hypercapnic patients undergoing PR further supports integrating refined physiological monitoring in higher-risk subgroups [56].

Supervision Modalities and Psychological Safety

An important observation from this review is that supervision intensity may contribute to heterogeneity in the effectiveness of remote rehabilitation. Subgroup analyses [31,33] indicated that synchronous, video-supervised programs demonstrated lower statistical heterogeneity in functional capacity outcomes, whereas low-supervision or asynchronous approaches exhibited greater variability. These findings raise the possibility that, beyond training dose alone, real-time interaction may support psychological safety, perceived support, and confidence during exercise, which in turn may influence engagement and adherence [48,57]. Conversely, insufficient supervision in home-based programs may heighten concerns regarding exercise-related risk and limit participation [58,59]. However, these interpretations should be approached cautiously. Given the exploratory nature of subgroup analyses and the limited number of studies within specific supervision categories, the independent contributions of psychological factors, adherence, and training intensity cannot be clearly disentangled. Future trials should prospectively incorporate measures of psychological safety, self-efficacy, and fear of exercise to clarify the pathways through which supervision modalities influence outcomes.

Comparison With Standard Care

The relative effectiveness of Tele-PR was closely linked to the characteristics of the comparator intervention. When compared to lower-intensity programs or usual care, Tele-PR was more likely to demonstrate a favorable relative effect [60]; however, when compared to high-intensity, face-to-face supervised rehabilitation, these advantages were substantially attenuated [43]. These findings suggest that conclusions regarding the equivalence of Tele-PR should be interpreted as context dependent rather than as intrinsic or universally applicable [61]. Accordingly, Tele-PR is best viewed as complementary to, rather than as a universal substitute for, high-quality center-based rehabilitation, with optimal use determined by patient characteristics, resource availability, and service capacity [62].

Barriers to Implementation and Digital Literacy

A notable gap was observed between improvements in functional capacity and changes in daily physical activity. Although gains in exercise capacity were observed under certain remote rehabilitation models, corresponding increases in daily step counts were not consistently demonstrated and were sensitive to the influence of individual studies [35]. This pattern suggests that functional improvement alone may be insufficient to produce sustained behavioral change [63]. Future Tele-PR strategies may therefore need to incorporate structured behavioral interventions alongside monitoring and feedback mechanisms. Digital literacy and technology access remain implementation barriers [64,65], and simpler, low-technology models with strong human support may achieve higher adherence in certain populations [41]. Platforms should prioritize usability and minimize cognitive burden to avoid undermining effectiveness [66,67].

Cost-Effectiveness and Economic Considerations

Although this review primarily focused on clinical outcomes, economic considerations are critical for the sustainable implementation of remote rehabilitation. Available evidence [32,40,68,69] suggests that Tele-PR is broadly comparable to CBPR in terms of direct program costs and may offer potential long-term economic benefits by improving accessibility and intervention completion rates, consistent with findings from recent cost-utility evaluations of digital therapeutics for PR. However, cost-effectiveness evidence remains heterogeneous [70], underscoring the need for standardized economic evaluations in future trials.

Innovation and Contribution

Previous systematic reviews of PR and telerehabilitation in COPD have largely focused on estimating average between-group differences, reporting minimal clinically important difference achievement, or comparing remote and center-based delivery formats [71,72], while foundational Cochrane work established the overall efficacy of PR compared with usual care [73]. Few syntheses have systematically

XSL•FO
RenderX

examined supervision-related heterogeneity or incorporated PIs to estimate expected effect ranges across settings. We observed substantial variation in program structure, supervision models, and comparator intensity, complicating the interpretation of pooled averages. By integrating supervision-based stratification with PI-informed interpretation, this review provides a context-sensitive framework for evaluating Tele-PR relative to CBPR. To our knowledge, it is among the first to frame Tele-PR equivalence as context dependent rather than model intrinsic, supporting tailored implementation in real-world health care systems.

Limitations

Several limitations should be acknowledged. First, most included studies were nonblinded, and implementation bias could not be fully avoided. Second, the number of long-term follow-up studies was limited, resulting in insufficient statistical power for certain long-term outcome analyses. Third, definitions of "usual care" and rehabilitation intensity varied across studies, potentially introducing residual confounding. In addition, several subgroup analyses were based on a small number of studies and should therefore be interpreted as exploratory rather than definitive, particularly where a subgroup was informed by a single trial. The term "noninferiority" as used in this study refers to the overall distribution of effect directions and their 95% CIs, rather than a prespecified noninferiority margin; thus, it should be interpreted as indicating clinical equivalence rather than constituting a strict statistical test of noninferiority. Consistent with the PRISMA-S reporting framework, the omission of trial registry and gray literature searches may have resulted in the underrepresentation of ongoing, unpublished, or null-result trials, potentially contributing to residual publication bias despite comprehensive database coverage. Furthermore, the small number of studies available for several outcomes precluded formal risk of bias–based sensitivity analyses, which should be taken into account when assessing the robustness of the findings. The clinical implications outlined earlier are primarily based on evidence of moderate to low certainty and should be applied cautiously in real-world settings, with due consideration of local resources, patient characteristics, and operational conditions. Subgroup and meta-regression analyses were exploratory and intended to generate hypotheses rather than confirm model superiority.

Conclusions

Distinct from prior reviews that treated Tele-PR as a homogeneous intervention, we stratified remote programs by supervision intensity and delivery models and proposed a structured "supervision gradient" framework to interpret model-dependent consistency of effects, with practical implications for designing Tele-PR as a supervision-structured service incorporating real-time oversight and long-term behavioral maintenance support. Tele-PR is a feasible option for delivering PR to adults with COPD and, on average, yields short-term clinical outcomes comparable to those of CBPR. As effects varied across studies and the certainty of the evidence was low to very low, implementation should be context aware and model specific. Programs incorporating structured real-time professional supervision appear to produce more consistent outcomes, whereas low-technology HBPR may require enhanced behavioral support to achieve stable symptom control and improvements in daily activity levels. Longer-term maintenance remains a challenge, indicating that remote rehabilitation should be designed as an ongoing service rather than a time-limited intervention. Accordingly, Tele-PR may be particularly valuable for expanding access to PR and addressing existing accessibility gaps, while CBPR remains essential for patients who require close in-person supervision or complex multidisciplinary care.

Acknowledgments

The authors gratefully acknowledge all patients who participated in the studies included in this meta-analysis and the study authors who provided original data for their help with this study. The authors declare the use of generative artificial intelligence (Gen-AI) in the research and writing process. According to the GAIDeT (Generative Artificial Intelligence Delegation Taxonomy 2025) [74], summarizing text was delegated to Gen-AI tools under full human supervision. The Gen-AI tool used was ChatGPT (OpenAI, GPT-4-level large language model). Responsibility for the final manuscript lies entirely with the authors. Gen-AI tools are not listed as authors and do not bear responsibility for the final outcomes. The tool was used to improve clarity, language, and structure of the manuscript. All scientific decisions, data analysis, interpretation of results, and conclusions were made by the authors.

Funding

Funding was received from the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2023ZD0506700, 2023ZD0506704) and the 2023 Henan Special Research Project of TCM on "Double First-Class" Construction (grant HSRP-DFCTCM-2023-3-12).

Data Availability

All data generated or analyzed during this study are included in this published study and its supplementary information files.

Authors' Contributions

HZ, YL, and GZ formulated the research question and conceived the study. LW and PJ contributed to the methodology. YL, LW, PJ, and HH designed and wrote the review protocol with crucial input from JL and HZ. YL and LW developed the search strategy and conducted the literature search. YL and HH independently screened all records and determined eligibility. PJ oversaw the

XSL•FO
RenderX

study selection process. All authors contributed to conflict resolution. YL, PJ, GZ, and HH were responsible for data curation. YL and LW critically appraised the included studies with crucial input from GZ and PJ. HH and PJ independently addressed the quality of effect estimates. YL and GZ were responsible for data visualization. YL supervised the study. JSL and HZ were responsible for project administration. YL and GZ wrote the first draft of the manuscript. PLJ, HH, LW, JSL, and HZ were involved in the writing and revision of the manuscript. All authors critically reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Changes to preregistered protocol.
[DOCX File, 15 KB - jmir_v28i1e80500_app1.docx]

Multimedia Appendix 2
Full electronic search strategies for all databases.
[DOCX File, 22 KB - jmir_v28i1e80500_app2.docx]

Multimedia Appendix 3
Reasons for exclusion of full-text articles.
[DOCX File, 20 KB - jmir_v28i1e80500_app3.docx]

Multimedia Appendix 4
Audit of study independence linkage of multiple reports via trial registration numbers and strategy.
[DOCX File, 14 KB - jmir_v28i1e80500_app4.docx]

Multimedia Appendix 5
Key intervention strategies and supervision models.
[DOCX File, 19 KB - jmir_v28i1e80500_app5.docx]

Multimedia Appendix 6
Risk of bias assessment of included randomized controlled trials.
[DOCX File, 20 KB - jmir_v28i1e80500_app6.docx]

Multimedia Appendix 7
Subgroup analyses and meta-regression plots.
[DOCX File, 776 KB - jmir_v28i1e80500_app7.docx]

Multimedia Appendix 8
Leave-one-out sensitivity analyses for all primary outcomes.
[DOCX File, 28 KB - jmir_v28i1e80500_app8.docx]

Checklist 1
PRISMA 2020, PRISMA 2020 for abstracts, and PRISMA-S checklists.
[PDF File, 280 KB - jmir_v28i1e80500_app9.pdf]

References

1. GBD 2016 Occupational Chronic Respiratory Risk Factors Collaborators. Global and regional burden of chronic respiratory disease in 2016 arising from non-infectious airborne occupational exposures: a systematic analysis for the Global Burden of Disease Study 2016. Occup Environ Med 2020 Mar;77(3):142-150. [doi: 10.1136/oemed-2019-106013]
2. Rothnie KJ, Müllerová H, Smeeth L, Quint JK. Natural history of chronic obstructive pulmonary disease exacerbations in a general practice-based population with chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2018 Aug 15;198(4):464-471. [doi: 10.1164/rccm.201710-2029OC] [Medline: 29474094]
3. Spruit MA, Singh SJ, Garvey C, et al. An official American Thoracic Society/European Respiratory Society statement: key concepts and advances in pulmonary rehabilitation. Am J Respir Crit Care Med 2013 Oct 15;188(8):e13-e64. [doi: 10.1164/rccm.201309-1634ST] [Medline: 24127811]

XSL•FO
RenderX

4. Rochester CL, Alison JA, Carlin B, et al. Pulmonary rehabilitation for adults with chronic respiratory disease: an Official American Thoracic Society Clinical Practice Guideline. Am J Respir Crit Care Med 2023 Aug 15;208(4):e7-e26. [doi: 10.1164/rccm.202306-1066ST] [Medline: 37581410]
5. Alison JA, McKeough ZJ, Johnston K, et al. Australian and New Zealand Pulmonary Rehabilitation Guidelines. Respirology 2017 May;22(4):800-819. [doi: 10.1111/resp.13025] [Medline: 28339144]
6. Spitzer KA, Stefan MS, Priya A, et al. Participation in pulmonary rehabilitation after hospitalization for chronic obstructive pulmonary disease among Medicare beneficiaries. Ann Am Thorac Soc 2019 Jan;16(1):99-106. [doi: 10.1513/AnnalsATS.201805-332OC] [Medline: 30417670]
7. Desveaux L, Janaudis-Ferreira T, Goldstein R, Brooks D. An international comparison of pulmonary rehabilitation: a systematic review. COPD 2015 Apr;12(2):144-153. [doi: 10.3109/15412555.2014.922066] [Medline: 24984085]
8. Sharma A, Harrington RA, McClellan MB, et al. Using digital health technology to better generate evidence and deliver evidence-based care. J Am Coll Cardiol 2018 Jun;71(23):2680-2690. [doi: 10.1016/j.jacc.2018.03.523]
9. Bhavnani SP, Parakh K, Atreja A, et al. 2017 Roadmap for Innovation-ACC Health Policy Statement on healthcare transformation in the era of digital health, big data, and precision health: a report of the American College of Cardiology Task Force on health policy statements and systems of care. J Am Coll Cardiol 2017 Nov 28;70(21):2696-2718. [doi: 10.1016/j.jacc.2017.10.018] [Medline: 29169478]
10. Zhuang M, Hassan II, W Ahmad WMA, et al. Effectiveness of digital health interventions for chronic obstructive pulmonary disease: systematic review and meta-analysis. J Med Internet Res 2025 May 26;27:e76323. [doi: 10.2196/76323]
11. Cox NS, McDonald C, Burge AT, Hill CJ, Bondarenko J, Holland AE. Comparison of clinically meaningful improvements after center-based and home-based telerehabilitation in people with COPD. Chest 2025 Apr;167(4):1003-1011. [doi: 10.1016/j.chest.2024.11.001] [Medline: 39522594]
12. Sánchez-Romero EA, García-Barredo-Restegui T, Martínez-Rolando L, et al. Addressing post-COVID-19 musculoskeletal symptoms through pulmonary rehabilitation with an evidence-based eHealth education tool: preliminary results from a pilot randomized controlled clinical trial. Medicine (Abingdon) 2025;104(10):e41583. [doi: 10.1097/MD.0000000000041583]
13. Martínez-Pozas O, Corbellini C, Cuenca-Zaldívar JN, Meléndez-Oliva É, Sinatti P, Sánchez Romero EA. Effectiveness of telerehabilitation versus face-to-face pulmonary rehabilitation on physical function and quality of life in people with post COVID-19 condition: a systematic review and network meta-analysis. Eur J Phys Rehabil Med 2024 Oct;60(5):868-877. [doi: 10.23736/S1973-9087.24.08540-X] [Medline: 39235257]
14. Martínez-Pozas O, Meléndez-Oliva E, Rolando LM, Rico JAQ, Corbellini C, Sánchez Romero EA. The pulmonary rehabilitation effect on long covid-19 syndrome: a systematic review and meta-analysis. Physiother Res Int 2024 Apr;29(2):e2077. [doi: 10.1002/pri.2077] [Medline: 38389299]
15. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71]
16. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc 2021 Apr 1;109(2):174-200. [doi: 10.5195/jmla.2021.962] [Medline: 34285662]
17. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282. [Medline: 23092060]
18. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
19. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019 Jun;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]
20. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023 Dec;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
21. Sterne JAC, Sutton AJ, Ioannidis JPA, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011 Jul 22;343(jul22 1):d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
22. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of two methods to detect publication bias in meta-analysis. JAMA 2006 Feb 8;295(6):676-680. [doi: 10.1001/jama.295.6.676] [Medline: 16467236]
23. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Stat Med 2002 Jun 15;21(11):1559-1573. [doi: 10.1002/sim.1187] [Medline: 12111920]
24. Senn SJ. Overstating the evidence: double counting in meta-analysis and related problems. BMC Med Res Methodol 2009 Feb 13;9:10. [doi: 10.1186/1471-2288-9-10] [Medline: 19216779]
25. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898]
26. What is GRADE? GRADE. URL: https://www.gradeworkinggroup.org/ [accessed 2026-02-01]
27. Chaplin E, Hewitt S, Apps L, et al. Interactive web-based pulmonary rehabilitation programme: a randomised controlled feasibility trial. BMJ Open 2017 Mar 31;7(3):e013682. [doi: 10.1136/bmjopen-2016-013682] [Medline: 28363923]

XSL•FO
RenderX

28. Burge AT, Holland AE, McDonald CF, et al. “Willingness to pay”: the value attributed to program location by pulmonary rehabilitation participants. COPD: Journal of Chronic Obstructive Pulmonary Disease 2021 May 4;18(3):281-287. [doi: 10.1080/15412555.2021.1924127]
29. Cerdán-de-Las-Heras J, Balbino F, Løkke A, Catalán-Matamoros D, Hilberg O, Bendstrup E. Effect of a new tele-rehabilitation program versus standard rehabilitation in patients with chronic obstructive pulmonary disease. J Clin Med 2021 Dec 21;11(1):11. [doi: 10.3390/jcm11010011] [Medline: 35011755]
30. Hansen H, Torre A, Kallemose T, Ulrik CS, Godtfredsen NS. Pulmonary telerehabilitation vs. conventional pulmonary rehabilitation - a secondary responder analysis. Thorax 2023 Oct;78(10):1039-1042. [doi: 10.1136/thorax-2023-220065] [Medline: 37451863]
31. Hansen H, Bieler T, Beyer N, et al. Supervised pulmonary tele-rehabilitation versus pulmonary rehabilitation in severe COPD: a randomised multicentre trial. Thorax 2020 May;75(5):413-421. [doi: 10.1136/thoraxjnl-2019-214246] [Medline: 32229541]
32. Li Y, Qian H, Yu K, Huang Y. The long-term maintenance effect of remote pulmonary rehabilitation via social media in copd: a randomized controlled trial. Int J Chron Obstruct Pulmon Dis 2022;17:1131-1142. [doi: 10.2147/COPD.S360125] [Medline: 35586118]
33. Cox NS, McDonald CF, Mahal A, et al. Telerehabilitation for chronic respiratory disease: a randomised controlled equivalence trial. Thorax 2022 Jul;77(7):643-651. [doi: 10.1136/thoraxjnl-2021-216934] [Medline: 34650004]
34. Chaplin E, Barnes A, Newby C, Houchen-Wolloff L, Singh SJ. Comparison of the impact of conventional and web-based pulmonary rehabilitation on physical activity in patients with chronic obstructive pulmonary disease: exploratory feasibility study. JMIR Rehabil Assist Technol 2022 Mar 10;9(1):e28875. [doi: 10.2196/28875] [Medline: 35266871]
35. Horton EJ, Ruksenaite J, Mitchell K, Sewell L, Newby C, Singh SJ. A comparison of physical activity between home-based and centre-based pulmonary rehabilitation: a randomised controlled secondary analysis. Front Rehabil Sci 2021;2:743441. [doi: 10.3389/fresc.2021.743441] [Medline: 36188808]
36. Maltais F, Bourbeau J, Shapiro S, et al. Effects of home-based pulmonary rehabilitation in patients with chronic obstructive pulmonary disease: a randomized trial. Ann Intern Med 2008 Dec 16;149(12):869-878. [doi: 10.7326/0003-4819-149-12-200812160-00006] [Medline: 19075206]
37. Güell MR, de Lucas P, Gáldiz JB, et al. Home vs hospital-based pulmonary rehabilitation for patients with chronic obstructive pulmonary disease: a Spanish multicenter trial. Arch Bronconeumol 2008 Oct;44(10):512-518. [Medline: 19006630]
38. Lahham A, McDonald CF, Mahal A, et al. Participation in physical activity during center and home-based pulmonary rehabilitation for people with COPD: a secondary analysis of a randomized controlled trial. J Cardiopulm Rehabil Prev 2019 Mar;39(2):E1-E4. [doi: 10.1097/HCR.0000000000000373] [Medline: 30688793]
39. Horton EJ, Mitchell KE, Johnson-Warrington V, et al. Comparison of a structured home-based rehabilitation programme with conventional supervised pulmonary rehabilitation: a randomised non-inferiority trial. Thorax 2018 Jan;73(1):29-36. [doi: 10.1136/thoraxjnl-2016-208506] [Medline: 28756402]
40. Vasilopoulou M, Papaioannou AI, Kaltsakas G, et al. Home-based maintenance tele-rehabilitation reduces the risk for acute exacerbations of COPD, hospitalisations and emergency department visits. Eur Respir J 2017 May;49(5):1602129. [doi: 10.1183/13993003.02129-2016] [Medline: 28546268]
41. Holland AE, Mahal A, Hill CJ, et al. Home-based rehabilitation for COPD using minimal resources: a randomised, controlled equivalence trial. Thorax 2017 Jan;72(1):57-65. [doi: 10.1136/thoraxjnl-2016-208514] [Medline: 27672116]
42. Mendes de Oliveira JC, Studart Leitão Filho FS, Malosa Sampaio LM, et al. Outpatient vs. home-based pulmonary rehabilitation in COPD: a randomized controlled trial. Multidiscip Respir Med 2010 Dec 20;5(6):401-408. [doi: 10.1186/2049-6958-5-6-401] [Medline: 22958267]
43. Sacristán-Galisteo C, Del Corral T, Gómez-Pesquera C, et al. Long-term effects of face-to-face supervision versus telephone supervision during a community-based pulmonary rehabilitation programme in people with COPD using minimal equipment: a randomized controlled trial. BMC Prim Care 2025 Aug 27;26(1):268. [doi: 10.1186/s12875-025-02962-3] [Medline: 40866835]
44. Jansons PS, Haines TP, O’Brien L. Interventions to achieve ongoing exercise adherence for adults with chronic health conditions who have completed a supervised exercise program: systematic review and meta-analysis. Clin Rehabil 2017 Apr;31(4):465-477. [doi: 10.1177/0269215516653995] [Medline: 27301799]
45. Duregon F, Quinto G, Vecchiato M, et al. From theory into practice: insights from a real-world implementation model for tailored exercise prescription in chronic diseases. BMC Sports Sci Med Rehabil 2025 Dec 29;17(1):373. [doi: 10.1186/s13102-025-01419-5] [Medline: 41462295]
46. Josa-Culleré A, Basagaña X, Koch S, et al. Short-term effects of air pollution and weather on physical activity in patients with chronic obstructive pulmonary disease (COPD). Environ Res 2024 Apr 15;247:118195. [doi: 10.1016/j.envres.2024.118195] [Medline: 38237751]
47. Donaldson GC, Goldring JJ, Wedzicha JA. Influence of season on exacerbation characteristics in patients with COPD. Chest 2012 Jan;141(1):94-100. [doi: 10.1378/chest.11-0281] [Medline: 21799024]

XSL•FO
RenderX

48. Olsen M, Nielsen C, Emme C, Godtfredsen NS, Hansen H. Why we keep going: a qualitative longitudinal study of the motivation and engagement among patients with COPD during a long-term Danish tele-rehabilitation program. Patient Prefer Adherence 2025;19:3573-3587. [doi: 10.2147/PPA.S549332] [Medline: 41257245]
49. Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease 2025 report. : Global Initiative for Chronic Obstructive Lung Disease; 2025 URL: https://goldcopd.org/2025-gold-report/ [accessed 2026-02-10]
50. Holland AE, Cox NS, Houchen-Wolloff L, et al. Defining modern pulmonary rehabilitation. An official American Thoracic Society Workshop Report. Ann Am Thorac Soc 2021 May;18(5):e12-e29. [doi: 10.1513/AnnalsATS.202102-146ST] [Medline: 33929307]
51. Rochester CL, Vogiatzis I, Holland AE, et al. An official American Thoracic Society/European Respiratory Society Policy Statement: enhancing implementation, use, and delivery of pulmonary rehabilitation. Am J Respir Crit Care Med 2015 Dec 1;192(11):1373-1386. [doi: 10.1164/rccm.201510-1966ST] [Medline: 26623686]
52. Nacarato D, Sardeli AV, Mariano LO, Chacon-Mikahil MPT. Cardiovascular telerehabilitation improves functional capacity, cardiorespiratory fitness and quality of life in older adults: a systematic review and meta-analysis. J Telemed Telecare 2024 Sep;30(8):1238-1248. [doi: 10.1177/1357633X221137626] [Medline: 36469017]
53. Richardson CR, Franklin B, Moy ML, Jackson EA. Advances in rehabilitation for chronic diseases: improving health outcomes and function. BMJ 2019 Jun 17;365:l2191. [doi: 10.1136/bmj.l2191] [Medline: 31208954]
54. Corbellini C, Rossino E, Massaccesi R, et al. Improvements in perimeter thoracic mobility on patients with COPD after pulmonary rehabilitation: a case series. ELECTRON J GEN MED 2022 Feb;19(3):em361. [doi: 10.29333/ejgm/11671]
55. Chung C, Kim AR, Kang DY, et al. Clinical efficacy of smartphone app-based pulmonary rehabilitation in chronic respiratory diseases: randomized controlled and feasibility trials. J Med Internet Res 2025 Nov 28;27:e76801. [doi: 10.2196/76801] [Medline: 41313804]
56. Corbellini C, Tavella S, Gugliotta E, Zampese S, Sanchez-Romero EA, Villafane J. Hypercapnia and functional improvements during pulmonary rehabilitation. Presented at: ERS International Congress 2021 Abstracts; Sep 5-8, 2021. [doi: 10.1183/13993003]
57. Chan SWC. Chronic disease management, self-efficacy and quality of life. J Nurs Res 2021 Feb 1;29(1):e129. [doi: 10.1097/JNR.0000000000000422] [Medline: 33427791]
58. Fukuta Y, Arizono S, Tanaka S, et al. Effects of real-time remote cardiac rehabilitation on exercise capacity and quality of life: a quasi-randomised controlled trial. BMC Geriatr 2023 Jun 24;23(1):388. [doi: 10.1186/s12877-023-04113-8] [Medline: 37353730]
59. Saka S, Gurses HN, Bayram M. Effect of inspiratory muscle training on dyspnea-related kinesiophobia in chronic obstructive pulmonary disease: a randomized controlled trial. Complement Ther Clin Pract 2021 Aug;44(101418):101418. [doi: 10.1016/j.ctcp.2021.101418] [Medline: 34034036]
60. Sakunrag I, Boontha N, Boonpattharatthiti K, Dhippayom T. Impact of tele-pulmonary rehabilitation in patients with chronic obstructive disease: a systematic review and network meta-analysis. Telemed J E Health 2025 Apr;31(4):441-450. [doi: 10.1089/tmj.2024.0476] [Medline: 39630505]
61. Burge AT, Cox NS, Abramson MJ, Holland AE. Interventions for promoting physical activity in people with chronic obstructive pulmonary disease (COPD). Cochrane Database Syst Rev 2020 Apr 16;4(4):CD012626. [doi: 10.1002/14651858.CD012626.pub2] [Medline: 32297320]
62. Ward TJC, Latimer L, Daynes E, et al. Impact of pulmonary rehabilitation programme design on effectiveness in COPD: a systematic review and component network meta-analysis. EClinicalMedicine 2025 Sep;87:103433. [doi: 10.1016/j.eclinm.2025.103433] [Medline: 40896469]
63. Spruit MA, Pitta F, McAuley E, ZuWallack RL, Nici L. Pulmonary rehabilitation and physical activity in patients with chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2015 Oct 15;192(8):924-933. [doi: 10.1164/rccm.201505-0929CI] [Medline: 26161676]
64. Choi NG, Dinitto DM. The digital divide among low-income homebound older adults: internet use patterns, eHealth literacy, and attitudes toward computer/Internet use. J Med Internet Res 2013 May 2;15(5):e93. [doi: 10.2196/jmir.2645] [Medline: 23639979]
65. Wilson J, Heinsch M, Betts D, Booth D, Kay-Lambkin F. Barriers and facilitators to the use of e-health by older adults: a scoping review. BMC Public Health 2021 Aug 17;21(1):1556. [doi: 10.1186/s12889-021-11623-w] [Medline: 34399716]
66. Bentley CL, Powell L, Potter S, et al. The use of a smartphone app and an activity tracker to promote physical activity in the management of chronic obstructive pulmonary disease: randomized controlled feasibility study. JMIR Mhealth Uhealth 2020 Jun 3;8(6):e16203. [doi: 10.2196/16203] [Medline: 32490838]
67. Chengyu Z, Xueyan H, Ying F. Research on disease management of chronic disease patients based on digital therapeutics: a scoping review. Digit Health 2024;10:20552076241297064. [doi: 10.1177/20552076241297064] [Medline: 39525556]
68. Zanaboni P, Dinesen B, Hoaas H, et al. Long-term telerehabilitation or unsupervised training at home for patients with chronic obstructive pulmonary disease: a randomized controlled trial. Am J Respir Crit Care Med 2023 Apr 1;207(7):865-875. [doi: 10.1164/rccm.202204-0643OC] [Medline: 36480957]

XSL•FO
RenderX

69. Park H, Jang M, Kim D, et al. Cost-utility analysis and value-based pricing of digital therapeutics for pulmonary rehabilitation in chronic respiratory disease: economic evaluation based on a randomized controlled trial. J Med Internet Res 2025 Dec 15;27:e73739. [doi: 10.2196/73739] [Medline: 41397244]
70. Liu S, Zhao Q, Li W, Zhao X, Li K. The cost-effectiveness of pulmonary rehabilitation for COPD in different settings: a systematic review. Appl Health Econ Health Policy 2021 May;19(3):313-324. [doi: 10.1007/s40258-020-00613-5] [Medline: 33079374]
71. Bondarenko J, Dal Corso S, Dillon MP, et al. Clinically important changes and adverse events with centre-based or home-based pulmonary rehabilitation in chronic respiratory disease: a systematic review and meta-analysis. Chron Respir Dis 2024;21:14799731241277808. [doi: 10.1177/14799731241277808] [Medline: 39187265]
72. Zhu J, Kikuchi Y, Nishimura S, Shibata K. Physiological and psychological therapeutic effects of telerehabilitation treatment on patients with chronic obstructive pulmonary disease: a systematic review and meta-analysis. J Phys Ther Sci 2025 Nov;37(11):572-578. [doi: 10.1589/jpts.37.572] [Medline: 41209595]
73. McCarthy B, Casey D, Devane D, Murphy K, Murphy E, Lacasse Y. Pulmonary rehabilitation for chronic obstructive pulmonary disease. Cochrane Database Syst Rev 2015 Feb 23;2015(2):CD003793. [doi: 10.1002/14651858.CD003793.pub3] [Medline: 25705944]
74. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (Generative AI Delegation Taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res 2026 Apr;33(3):2544331. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]

Abbreviations

CAT: chronic obstructive pulmonary disease assessment test
CBPR: center-based pulmonary rehabilitation
COPD: chronic obstructive pulmonary disease
GAIDeT: Generative Artificial Intelligence Delegation Taxonomy
Gen-AI: generative artificial intelligence
GRADE: Grading of Recommendations, Assessment, Development, and Evaluation
HBPR: home-based pulmonary rehabilitation
HKSJ: Hartung-Knapp-Sidik-Jonkman
MD: mean difference
PI: prediction interval
PICOS: population, intervention, comparator, outcomes, and study design
PR: pulmonary rehabilitation
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SGRQ: St George's Respiratory Questionnaire
Tele-PR: pulmonary telerehabilitation

Edited by S Brini; submitted 11.Jul.2025; peer-reviewed by EA Sanchez-Romero, H Chen; accepted 02.Mar.2026; published 17.Apr.2026.

Please cite as:
Li Y, Zhang H, Zhao G, Li J, Huang H, Wang L, Jia P
Comparing Pulmonary Telerehabilitation and Center-Based Pulmonary Rehabilitation for Effectiveness and Adherence in Chronic Obstructive Pulmonary Disease: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e80500
URL: https://www.jmir.org/2026/1/e80500
doi:10.2196/80500

© Ya Li, Hailong Zhang, Guixiang Zhao, Jiansheng Li, Haoxuan Huang, Longyu Wang, Peilin Jia. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Clinical Effectiveness of Immersive Virtual Reality Exercise Interventions: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Riley C C Brown, PhD; Megan H Ross, PhD; Trevor G Russell, PhD

RECOVER Injury Research Centre, The University of Queensland, Level 7, Surgical Treatment and Rehabilitation Services (STARS), Herston, Queensland, Australia

Corresponding Author:
Riley C C Brown, PhD
RECOVER Injury Research Centre, The University of Queensland, Level 7, Surgical Treatment and Rehabilitation Services (STARS), Herston, Queensland, Australia

Abstract

Background: Physical inactivity remains a global health concern, with only one in 5 adults meeting combined aerobic and muscle-strengthening guidelines. Exercise interventions delivered through immersive virtual reality (IVR) offer a novel mode of delivery. Little is known about the clinical effectiveness or feasibility of exercise via IVR across population groups. A detailed understanding of clinical effectiveness and feasibility is required for clinicians to decide whether to include IVR in exercise practice.

Objective: The objective of this systematic review was to assess the clinical effectiveness of IVR interventions using aerobic or anaerobic exercise.

Methods: A systematic review incorporating meta-analyses was conducted. Searches were conducted across PubMed, Embase, Web of Science, and CINAHL from inception until January 6, 2026. Randomized controlled trials including participants with an acute health condition, chronic disease, history of reconstructive or restorative surgery, and older adults implementing IVR exercise and reporting clinical effectiveness outcomes were included. Random effects meta-analyses were conducted for between-group comparisons for clinical effectiveness outcomes, grouped according to comparator group activity (exercising/nonexercising). Risk of Bias was assessed using the Cochrane Risk of Bias 2 tool and the certainty of evidence with Grading of Recommendations, Assessment, Development, and Evaluation.

Results: Twenty-six trials with 846 total participants were included in this review, with 23 progressing to meta-analyses. Pooled analyses revealed a general trend for IVR, but no statistical differences with comparator intervention (exercising or nonexercising) for mobility and functional balance (exercising: standardized mean difference [SMD] −0.345, 95% CI −1.095 to 0.406; *P*=.29; nonexercising SMD −0.322, 95% CI −0.931 to 0.288; *P*=.22), functional leg strength (exercising: SMD −0.161, 95% CI −0.573 to 0.250; *P*=.33; nonexercising: SMD −0.351, 95% CI −1.750 to 1.049; *P*=.48), quality of life (exercising: SMD 0.036, 95% CI −0.444 to 0.516; *P*=.84; nonexercising: SMD −0.053, 95% CI −0.839 to 0.728; *P*=.80) or other outcome domains. Eighty percent of outcomes assessed were rated as “some concerns” (n=16) or at “high” (n=21) risk of overall bias. Grading of Recommendations, Assessment, Development, and Evaluation certainty grading was deemed to be “low” or “very low” for all outcomes.

Conclusions: This systematic review incorporating meta-analyses provides initial evidence for the clinical effectiveness of IVR exercise interventions. This review differs from previous literature by systematically collecting and appraising evidence exclusively from IVR aerobic/anaerobic exercise interventions from across a variety of populations and settings, and including a broad range of clinical effectiveness outcomes. Initial evidence may suggest that IVR exercise does not seem to statistically differ from comparators for clinical effectiveness outcomes. However, high heterogeneity, substantial risk of bias among trials, and “low” to “very low” certainty in evidence reduce overall confidence in the findings. While these results indicate that IVR may be a viable option for the delivery of exercise, a more robust methodology in future trials is needed to properly verify findings and improve certainty. This will help to determine the real-world applicability of IVR exercise interventions for the improvement of health-related measures.

Trial Registration: PROSPERO CRD420250650110; https://www.crd.york.ac.uk/PROSPERO/view/CRD420250650110

(J Med Internet Res 2026;28:e87542) doi:10.2196/87542

KEYWORDS

virtual reality; immersive virtual reality; physical activity; effectiveness; feasibility; exercise

XSL•FO
RenderX

Introduction

Physical inactivity is a substantial public health concern directly responsible for more than 7% of all-cause and cardiovascular disease deaths globally [1]. In 2022, the global age-standardized prevalence of adults not meeting the recommended guidelines of at least 150 minutes of moderate-intensity aerobic physical activity was 31% [2], with projections forecasting an increase to 35% by 2030 [3]. When expanded to account for muscle-strengthening recommendations (at least 2 sessions per week), one in 5 adults meets the combined guidelines [4]. Exercise interventions incorporating aerobic and anaerobic (eg, muscle strengthening activities) have been shown to improve disease control and reduce cardiovascular risk for a variety of cohorts [5]. However, traditional exercise programs often have low rates of adherence, particularly in chronic disease cohorts (≈50%) [6]. This highlights a persistent implementation gap and underscores the need for novel approaches to exercise delivery to promote sustained physical activity and exercise adherence. Digital health technologies may represent an effective and accessible form of physical activity intervention in clinical cohorts [7].

Immersive virtual reality (IVR) is an advanced form of digital health and computer-human interaction that immerses users in a 3D, computer-generated environment using a head-mounted display (HMD) [8]. Early research suggests that IVR may serve as an acceptable and effective strategy for delivering cognitive and psychological interventions across multiple populations [9-14]. Contemporary research also suggests that physical rehabilitation (eg, balance and range of motion training) incorporating IVR is comparable to conventional approaches for stroke motor recovery [15-17] and musculoskeletal pain improvement [18,19]. Furthermore, a 2024 systematic review suggests that IVR can be used as a tool to support physical activity promotion and psychological well-being through reduced perceived exertion, increased enjoyment, and improved self-efficacy [20]. Although these interventions appear to positively influence physical activity and exercise adherence in different populations [20-22], the extent to which IVR translates into meaningful improvements in health-related outcomes remains uncertain.

While research is limited compared with cognitive or physical rehabilitation, IVR exercise interventions for healthy adults have shown promise for improving physical activity level [20,21,23], neuromuscular strength [24], physical performance [24], and reducing exercise-induced pain [25]. Exercise delivered via IVR may have the potential to improve adherence, enjoyment, and sustainability of exercise for clinical populations [26-29]. This may be in part due to heightened immersion leading to improved intrinsic motivation, enjoyment, and higher psychological engagement during exercise activity [23,30-32]. However, much of the previous review literature in this field has either combined immersive and nonimmersive technologies (eg, mobile apps) or primarily focused on behavioral and perceptual outcomes. Research findings regarding the clinical effectiveness of IVR exercise interventions remain scarce. A robust and broad investigation into the clinical effectiveness of these intervention types across clinical population groups is warranted. A detailed understanding of effectiveness is required to determine whether IVR could be promoted as an alternate or adjunct evidence-based exercise intervention by clinicians to improve health-related outcomes. Therefore, the objective of this systematic review was to assess the clinical effectiveness of IVR interventions using aerobic or anaerobic exercise approaches for health-related outcomes. Additionally, many feasibility considerations related to IVR exercise are not investigated systematically in previous literature. Feasibility information is vital to assess the potential for the implementation of the IVR exercise by clinicians. Therefore, this review also aimed to assess the feasibility of delivering these interventions to clinical cohorts through analysis of exercise session attendance and adherence, technological issues, motion sickness, safety, and participant experiences.

Methods

Overview

This systematic review incorporating meta-analysis is reported according to the expanded 2020 PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) checklist [33]. This review and protocol were registered under review number CRD420250650110 through PROSPERO International prospective register of systematic reviews on February 24, 2025.

Terminology

IVR is a form of computer-human interface that allows the user to become immersed in a 3D, computer-generated environment while wearing an HMD [8].

Aerobic exercise is a structured and planned physical activity in which the body's large muscles move in a rhythmic manner for a sustained period of time [34].

Anaerobic exercise is a structured and planned physical activity consisting of brief high-intensity bursts of exercise (such as strength training or fast boxing), where oxygen demand surpasses oxygen supply [34].

Information Sources, Search Strategy, and Selection Process

The search strategy used in this systematic review is reported according to the PRISMA-S (PRISMA–Search) statement for reporting literature searches in systematic reviews [35]. Systematic searches were conducted by one reviewer (RCCB) across 4 databases from inception until January 14, 2025, and re-run on January 6, 2026. The databases searched included PubMed, Embase (Elsevier), Web of Science (Clarivate), and CINAHL Complete (EBSCOhost). Study registries were not searched for article identification purposes. Key search terms included "immersive virtual reality," "immersive virtual reality exercise," and "physical activity." Search strategies for each database were cocreated by all members of the review team and are available in Multimedia Appendix 1. Search strategies from previous systematic reviews were not reused. Recursive and forward searching of reference lists for all included studies and other systematic reviews (those published in 2024 through Google Scholar) was completed. Covidence software (Veritas Health Innovation) was used to deduplicate records from the

XSL•FO
RenderX

databases and complete screening procedures. Google Translate was used to translate reports from other languages to English. Only randomized controlled trials (RCTs) were included in this review. The non-RCT automation tool within Covidence software was applied to assist with the identification of ineligible study types prior to study screening. No additional information sources or search methods were used. Two reviewers (RCCB and MHR) contributed to study screening and selection. All disagreements in screening were resolved through consensus discussion.

Eligibility Criteria

Participants

Included trials were not restricted by sex. Trials were included if participants had any of the following: (1) an acute health condition (eg, post–COVID-19 sequelae), (2) long-lasting chronic disease with persistent effects (eg, musculoskeletal or neurological conditions), (3) history of reconstructive or restorative surgery (eg, total knee arthroplasty), or (4) older adults. Studies including children or adolescents were not included.

Interventions

Participants must have undergone an IVR intervention incorporating aerobic or anaerobic exercise (supervised or unsupervised). Intervention duration must have been at least 3 weeks to allow for potential changes in effectiveness outcome measures. Trials were excluded if they used augmented reality (eg, smart glasses), non-IVR (eg, computer screen), semi-IVR (eg, driving simulator), an exergaming intervention not using an HMD (eg, Xbox Kinect), or were a nonexercise or balance-focused intervention.

Comparators

Exercising and nonexercising comparator groups were included. Single-arm trials were excluded.

Outcomes

Trials were included if they reported changes in continuous clinical effectiveness outcome measures. Authors were contacted for further information if eligible outcomes were measured but not reported. Eligible questionnaires presenting multiple domain scores were averaged for inclusion into meta-analyses (eg, quality of life questionnaires). Secondary feasibility outcomes were safety metrics (eg, adverse events), session attendance rates, exercise adherence, technological issues, motion sickness, and participant experiences.

Study Design

Completed RCTs were included. All other study designs (eg, nonrandomized trials, reviews, protocols, unpublished manuscripts, conference abstracts) were excluded. Studies were not restricted by language.

Data Items and Collection Process

Data describing participant and trial characteristics, eligibility criteria, interventions, trial funding sources, clinical effectiveness outcomes, and feasibility outcomes were first extracted by 2 independent reviewers (RCCB and MHR) for 30% of trials (n=8) using Covidence and Microsoft Excel (Microsoft Corp) and tabulated into descriptive tables. After it was ascertained that there was 100% agreement between reviewers, one independent reviewer extracted data for the remainder of the trials (n=18; RCCB). Where further information was needed (n=3), authors were contacted via email by one reviewer (RCCB). Clinical effectiveness data were obtained for all reported timeframes in each individual trial. For studies presenting multiple assessments for the same clinical effectiveness domain, one was chosen for entry into the meta-analysis. All data in this review that were not reported in the original manuscripts were obtained via author correspondence.

Risk of Bias and Certainty in Cumulative Estimates

The revised Cochrane Risk of Bias 2 tool (RoB2) [36] was used for quality appraisal for included trials. Study bias was assessed as “high,” “some concerns,” or “low” across 5 domains (randomization process, deviations from intended interventions, missing outcome data, measurement of the outcome, selection of the reported result) and facilitated by 2 independent reviewers (RCCB and MHR) for 28% of outcomes (n=15). Journal articles, protocol papers, and trial registrations were used in quality appraisal. Once consensus was achieved, the remaining outcomes were assessed by one reviewer (n=38; RCCB). An overall risk of bias for each outcome was determined using the RoB2 method [36]. Quality appraisal was based on published manuscripts. The certainty of evidence for clinical effectiveness outcomes was evaluated using the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) approach [37,38]. This was facilitated using GRADEPro GDT software [39], and assessed risk of bias, indirectness, inconsistency, imprecision, and other factors at the outcome level. The certainty of evidence was categorized as “high,” “moderate,” “low,” or “very low.” Outcome domains were downgraded if ≥50% of included trials presented issues in each criterion. Given the effect of exercise per se on clinical effectiveness outcomes, GRADE was facilitated according to the comparator group activity.

Synthesis Methods and Meta-Analysis

Meta-analyses were completed using Comprehensive Meta-Analysis Software (version 4; Biostat) [40]. Clinical effectiveness domains reported in 3 or more trials with comparable control groups (ie, exercising or nonexercising) progressed to meta-analyses. Random effects meta-analyses with the Hartung-Knapp-Sidik-Jonkman adjustment allow for differences in treatment effect to be present and accounted for throughout the included trials with greater precision [41,42] and were conducted for the effects of the IVR exercise interventions for mobility and functional balance, condition severity (eg, WOMAC), quality of life, pain intensity, composite static and dynamic balance assessments, aerobic physical activity (device-measured), and static balance. Eligibility of outcome inclusion into individual meta-analyses was completed as part of the data extraction process. Double-counting of data for individual trials in meta-analyses did not occur. Where applicable, data were converted from 95% CIs into SDs for inclusion into the meta-analysis. For the meta-analyses,

XSL•FO
RenderX

preintervention and postintervention means and SDs, and sample size per group were used. A within-group Cohen *d* effect size was calculated to estimate the change from baseline for each group. We used a plausible and conservative pre-post correlation of .5 measured within each comparison group [43,44]. For the effect size difference between groups, standardized mean differences (SMDs) were used. This was due to the differing forms of measurement for the above-listed clinical effectiveness domains. Prespecified levels of magnitude for SMD were set at .2 for small, .5 for moderate, and .8 for large [45]. The SMD and 95% CIs were calculated using random-effect meta-analyses with the inverse of variance. Prediction intervals were calculated for meta-analyses with 5 or more included studies. Statistical heterogeneity was assessed via the Q-test and prediction intervals, and between-study variability was calculated using the I^2 statistic (0%‑25%=low, 26%‑74%=moderate, ≥75%=high) [46]. Prediction intervals were not calculated where statistical heterogeneity was low [47]. Data are presented in tables as per population group due to substantial heterogeneity identified. Small-study effects were assessed using the Egger test and visual inspection of funnel plots [48]. Additionally, the risk of bias due to missing results in meta-analyses was assessed through Egger test, funnel plots, and domains 3‑5 of the RoB2 tool. Analyses were grouped according to comparator group activity (exercising/nonexercising). Sensitivity analyses conducted included (1) removing individual trial results from the models to ascertain the effects on the overall results, and (2) running pre-post correlation at levels of 0.6, 0.7, 0.8, and 0.9 to assess influence on the overall result in line with Cochrane guidelines [43,44]. Subgroup analyses based on population were not completed due to the scarcity of common data.

Ethical Considerations

This systematic review uses data that are publicly available through previous literature and did not require research ethical approval. The review was registered through PROSPERO International prospective register of systematic reviews on February 24, 2025 (CRD420250650110).

Results

Overview

Figure 1 displays that 5241 records were identified in the initial search. A total of 26 trials were included, with 23 progressing to meta-analyses.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) flow diagram.

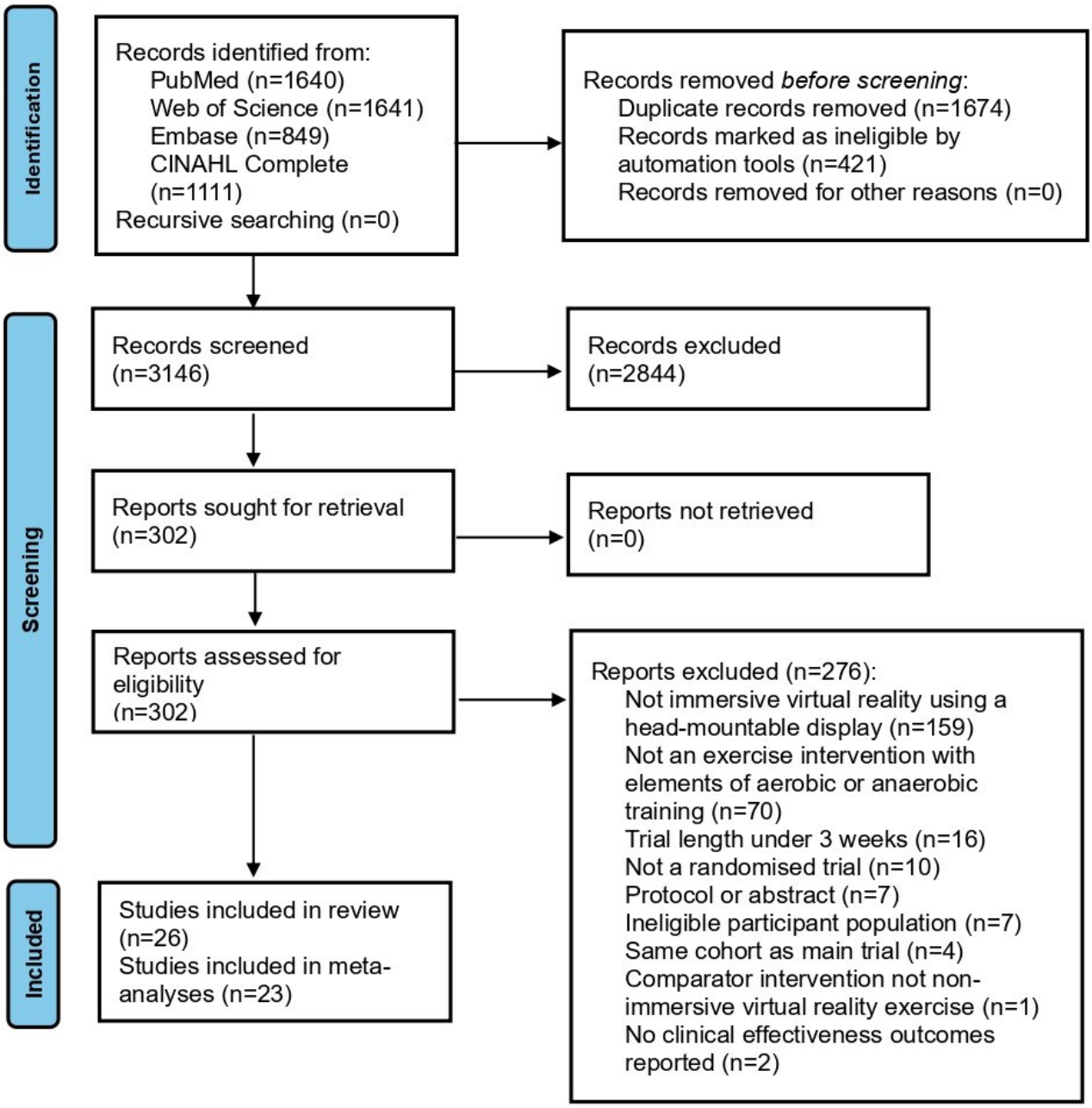

Trial Characteristics

Trial characteristics are displayed in Table 1. Trials were conducted from 2021 to 2025 in Spain [28,49-51] (n=4), China [29,52,53] (n=3), Germany [54-56] (n=3), India [57-59] (n=3), and other countries [60-72] (n=13). A total of 846 people participated across all trials. Average sample size across trials was 33.5 (SD 18.0) participants, and mean age was 60.7 (SD 15.8) years. Trials included participants with neurological disorders [49-51,59,70,71], older adults [28,52,56,60,61,63,64,67-69], musculoskeletal conditions [29,55,57,58,62,66], cardiopulmonary diseases [53,65], cancer [54], and metabolic conditions [72]. The most common population group was older adults, accounting for 349 (41%) of all participants.

XSL•FO
RenderX

Table . Trial characteristics, including design, population, setting, provider, follow-up points, and funding source.

Reference, year, country, population	Trial design	Participants (women/men)		Age (years), mean (SD)		Setting	Provider	Follow up points	Trial funding source
		Intervention	Control	Intervention	Control				
Older adults									
Barsasella et al, 2021, Taiwan, Older adults [60]	RCT[a]	29 (25/4)	31 (21/10)	71.7 (7.4)	70.7 (7.4)	Community	Researcher	6 weeks	Ministry of Science and Technology, Taiwan; Taipei Medical University; Wanfang hospital; Ministry of Education
Campo-Prieto et al, 2022, Spain, Older adults [28]	Feasibility RCT	13 (11/2)	11 (10/1)	85.08 (8.48)	84.82 (8.10)	Community	NR[b]	10 weeks	Galician Government (Xunta de Galicia) Predoctoral fellowship
Drazich et al, 2023, United States of America, Older adults [61]	Pilot RCT	10 (9/1)	9 (6/3)	76.4 (7.6)	71.4 (3.9)	Community	Researcher	8 weeks	Sigma Theta Tau international honor society of nurses; Southern Nursing Research Society
Kershner et al, 2024, United States of America, Older adults [63]	Pilot RCT	5 (3/2)	4 (3/1)	66.00 (63.60 - 68.40)[c,d]	65.50 (63.25 - 67.75)[d]	Community	Behavioral coach	4 weeks	Wells Fargo Faculty Scholar Award
Kwan et al, 2021, China, Older adults [52]	RCT	9 (8/1)	8 (7/1)	73 (7.5)[c]	77.5 (15.3)	Community	Researcher	8 weeks	Innovation and Technology Fund for Better Living from The Hong Kong Polytechnic University
Lima Rêbelo et al, 2021, Brazil, Older adults [64]	RCT	20 (16/4)	17 (15/2)	69.25 (5.67)	71.41 (5.94)	Community	Researcher	8 weeks	No funding source
Parmak et al, 2025, Northern Cyprus, Older adults [69]	RCT	22 (18/4)	22 (18/4)	71.14 (4.83)	75.36 (9.16)	Community	Physiotherapist	8 weeks	No funding source
Vorwerg-Gall et al, 2024, Germany, Older adults [56]	Pilot RCT	23 (14/9)	12 (10/2)	69 (4.7)	66 (5.2)	Community	Physiotherapist	6 weeks	Federal Ministry of Education and Research Germany

XSL•FO
RenderX

Reference, year, country, population	Trial design	Participants (women/men)		Age (years), mean (SD)		Setting	Provider	Follow up points	Trial funding source
		Intervention	Control	Intervention	Control				
Yalfani et al, 2024, Iran, Older adults [67]	RCT	12 (12/0)	12 (12/0)	68.25 (2.92)	67.08 (2.9)	Laboratory	NR	8 weeks	No funding source
Zak et al, 2024, Poland, Older women [68]	RCT	40 (40/0)	40 (40/0)	76.7 (2.21)	76.75 (1.77)	NR	Physiotherapist	6 and 9 weeks	Chancellor's Grant: Jan Kochanowski University of Kielce, Poland
Neurological									
An and Park, 2022, South Korea, Incomplete spinal cord injury [71]	RCT	20 (8/12)	20 (9/11)	42.27 (5.51)	43.00 (7.15)	Community	Physiotherapist	4 weeks	NR
Peláez-Vélez et al, 2023, Spain, Stroke [49]	Pilot RCT	12 (3/9)	12 (5/7)	51.91 (18.58)	59.58 (15.97)	Hospital	Physiotherapist	6 weeks	No funding source
Ramos et al, 2025, Australia, Developmental disability [70]	RCT	13 (3/10)	8 (3/5)	36.0 (12.0)	36.0 (13.0)	Community	Exercise Physiologist	8 weeks	University of South Australia
Rodriguez-Fuentes et al, 2024, Spain (a), Parkinson disease [50]	RCT	30 (11/19)	22 (11/11)	70.87 (6.67)	70.59 (6.67)	NR	NR	12 weeks	Intramural Call for Biomedical Research Projects 2022, from the Galicia Sur Health Research Institute
Rodriguez-Fuentes et al, 2024, Spain (b), Multiple sclerosis [51]	Feasibility RCT	8 (6/2)	10 (7/3)	41.13 (4.88)	48.2 (5.40)	NR	Physiotherapist	8 weeks	Colexio Oficial de Fisioterapeutas de Galicia (Official Physical Therapy Council of Galician, Spain), section Research Grant 2023
Vishnuram et al, 2024, India, Stroke [59]	Pilot RCT	4 (NR)	4 (NR)	NR	NR	NR	Physiotherapist	4 weeks	No funding source
Musculoskeletal									

XSL•FO
RenderX

Reference, year, country, population	Trial design	Participants (women/men)		Age (years), mean (SD)		Setting	Provider	Follow up points	Trial funding source
		Intervention	Control	Intervention	Control				
Gsangaga et al, 2023, Malaysia, Postanterior cruciate ligament reconstruction [62]	RCT	15 (5/10)	15 (2/13)	28.6 (NR)	25.1 (NR)	Hospital	Physiotherapist	26 weeks	Dana Fundamental PPUKM
Lo et al, 2024, China, Knee osteoarthritis [29]	Pilot RCT	15 (10/5)	15 (13/2)	63 (60-67)[c,d]	64 (62-65)[d]	Community	Physiotherapist	12 weeks	Hong Kong Jockey Club Charities Trust
Naqvi et al, 2022, India, Distal radius fracture [57]	Pilot RCT	10 (NR)	10 (NR)	NR	NR	Hospital	NR	2 and 4 weeks	No funding source
Nishitha et al, 2024, India, Total knee arthroplasty [58]	RCT	18 (NR)	18 (NR)	NR	NR	NR	Physiotherapist	2, 3, 4, 9, and 12 weeks	NR
Stamm et al, 2022, Germany, Chronic low back pain [55]	Pilot RCT	11 (8/3)	11 (6/5)	75 (5.8)	75.5 (4.39)	Laboratory	Physiotherapist	4 weeks	German Federal Ministry of Education and Research
Tuck et al (2022), New Zealand, Chronic musculoskeletal pain [66][d]	Mixed methods RCT	10 (8/2)	10 (5/5)	41.3 (17.7)	38.7 (15.3)	Hospital	Physiotherapist	6 weeks	Auckland University of Technology
Cardiopulmonary									
Rutkowski et al, 2022, Poland, Post COVID-19 [65]	RCT	16 (NR)	16 (NR)	NR	NR	Hospital	Physiotherapist	3 weeks	The Polish National Agency for Academic Exchange (Urgency Grant)
Wang et al, 2023, China, Coronary heart disease [53]	RCT	18 (9/9)	18 (8/10)	72.50 (6.16)	72.61 (5.45)	Hospital	Physiotherapist	12 weeks	NR
Cancer									
Schrempf et al, 2023, Germany, Colorectal cancer [54]	Pilot RCT	31 (12/19)	31 (13/18)	60.4 (9.5)	60.9 (9.7)	Hospital	Hospital staff or doctoral students	1 and 4 weeks	University of Augsburg

XSL•FO
RenderX

Reference, year, country, population	Trial design	Participants (women/men)		Age (years), mean (SD)		Setting	Provider	Follow up points	Trial funding source
		Intervention	Control	Intervention	Control				
Metabolic									
Seo et al, 2023, South Korea, Overweight middle-aged women[e] [72]	RCT	23 (23/0)	23 (23/0)	47.74 (5.50)	48.26 (7.56)	Hospital	Researcher	4 and 8 weeks	National Research Foundation of Korea grant (Korean Government)

[a]RCT: randomized controlled trial.

[b]NR: not reported.

[c]The third group consists of the same participants as the control and is not included in this review.

[d]Median (IQR).

[e]TG of participants is 24 (24/0) and age is 49.00 (6.77).

Intervention Characteristics

Details of the virtual reality interventions are summarized in Table 2. Aerobic interventions were used in 9 trials [50,52,54,57,60,61,63,65,72], anaerobic interventions in 4 [29,59,68,71], and a mix of aerobic/anaerobic in 13 [28,49,51,53,55,56,58,62,64,66,67,69,70]. Exercise intensity metrics were reported in 9 trials [50,53-56,63,65,70,72]. Dedicated IVR systems were used in 19 trials [28,49-57,60-67,69], 4 used smartphones in combination with a head mountable stand [29,59,68,72], and 3 did not specify [58,70,71]. Physiotherapists were the most common intervention provider (14/26, 54%). Intervention duration ranged from 30 days to 26 weeks, with 6 - 8 weeks being the most common (14/26, 54%). Most participants in virtual reality groups were supervised in person (19/26, 73%).

XSL•FO
RenderX

Table . Intervention characteristics, including type, immersive virtual reality system used, game type, additional equipment, sessions per week, intensity, session length, length of intervention, location, and cointerventions.

Reference	Intervention	IVR system	Game type	Additional equipment	Sessions/wk	Intensity	Session length (mins)	Length (wks)	Location	Cointerventions
Older adults										
Barsasella et al, 2021 [60]	I[a]: Aerobic C[b]: NA[c]	HTC Vive	Commercial	None	I: 2 C: NA	I: NR[d] C: NA	I: 15 C: NA	I & C: 6	I: In-person (S[e]) C: NA	I & C: NA
Campo-Prieto et al, 2022 [28]	I: Aerobic & Anaerobic (boxing) C: NA	HTC Vive Pro	Commercial	None	I: 3 C: NA	I: NR C: NA	I: 6 C: NA	I & C: 10	I & C: In-person (S)	I: IVR[g] given in addition to UC[h] C: NA
Drazich et al, 2023 [61]	I: Aerobic C: NA	Oculus Quest 2	Commercial	Nordic Track R35 Recumbent Cycle Ergometer	I: 2 C: NA	I: Patient-led (no metric) C: NA	I: 40 C: NA	I & C: 8	I: In-person (S) C: NA	I: Received control intervention in addition C: Education program
Kershner et al, 2024 [63]	I & C: Aerobic	Meta Quest 2	Commercial	None	I & C: Patient-led	I: MVPA[i] (>70% max HR[j]) C: MVPA (>70% max HR)	I & C: Pt-led	I & C: 4	I: Home (S) C: Home (US[f])	I & C: 8 x individual and 4 x group meetings with behavior coach to discuss MVPA achievement
Kwan et al, 2021 [52]	I & C: Aerobic	HTC Vive Focus Plus	Bespoke	Under-desk ergometer	I & C: 2	I & C: Patient-led (no metric)	I & C: 15	I & C: 8	I & C: In-person (S)	I & C: Additional cognitive training
Lima Rêbelo et al, 2021 [64]	I: Aerobic & Anaerobic (boxing & strength) C: Balance	Oculus Rift	Commercial	None	I & C: 2	I: NR C: Progressive (no metric)	I & C: 50	I & C: 8	I & C: In-person (S)	I & C: NA
Parmak et al, 2025 [69]	I: Aerobic & Anaerobic (boxing) C: Balance, Anaerobic (strength)	Oculus Meta Quest 2	Commercial	None	I & C: 3	I: Patient-led (no metric)	I & C: 35	I & C: 8	I: In-Person (S) C: Home (US)	I & C: NA
Vorwerg-Gall et al, 2024 [56]	I & C: Aerobic & Anaerobic (strength)	HTC Vive Pro	Bespoke	Dumbbells with hand location tracking	I & C: 2	I & C: Aerobic: 40% - 60% HRR[k] Anaerobic (strength): 2 - 3 sets for 20 - 30 reps[l]	I & C: 30	I & C: 6	I & C: In-person (S)	I & C: Breathing exercises

XSL•FO
RenderX

Reference	Intervention	IVR system	Game type	Additional equipment	Sessions/wk	Intensity	Session length (mins)	Length (wks)	Location	Cointerventions
Yalfani et al, 2024 [67]	I: Aerobic & Anaerobic (boxing) C: NA	HTC Vive	Commercial	None	I: 3 C: NA	I: NR C: NA	I: 30 C: NA	I & C: 8	I: In-person (S) C: NA	I: NA C: No rehabilitation intervention received
Zak et al, 2024 [68]	I & C: Anaerobic (strength)	VR ONE plus-ZEISS (smartphone)	Commercial	None	I & C: 3	I & C: NR	I & C: 60	I & C: 6	I & C: In-person (S)	I: Given control intervention in addition C: NA
Neurological										
An and Park, 2022 [71]	I & C: Anaerobic (strength)	NR	Bespoke	Lower limb sensors	I & C: 3	I & C: NR	I & C: 30	I & C: 4	I & C: NR	I & C: NA
Peláez-Vélez et al, 2023 [49]	I & C: Aerobic & Anaerobic (strength)	Oculus Quest 2	Bespoke	None	I: 5 (+3 IVR) C: 5	I & C: NR	I: 60 (+30 for IVR) C: 60	I & C: 6	I & C: In-person (S)	I: Received control intervention C: NA
Ramos et al, 2025 [70]	I: Aerobic & Anaerobic (boxing) C: NA	NR	Commercial	None	I: 3 C: NA	I: *Aerobic & Anaerobic (boxing)*: Borg RPE scale (NR) C: NA	I: 60 C: NA	I & C: 8	I: In-person (S) C: NA	I & C: NA
Rodriguez-Fuentes et al, 2024 (a) [50]	I & C: Aerobic	Meta Quest 2	Commercial	Cycle ergometer with IVR connectivity	I & C: 2	I & C: 80 - 90 RPM[m]; 70% - 80% of Max HR, Borg RPE[n] 8-9/10	I & C: 25	I & C: 12	I & C: In-person (S)	I & C: NA
Rodriguez-Fuentes et al, 2024 (b) [51]	I: Aerobic & Anaerobic (boxing) C: NR (“Exercising”)	Oculus Quest 3	Commercial	None	I: 2 (+ C) C: NR	I & C: NR	I: 6 C: NR	I & C: 8	I & C: In-person (S)	I: Received control intervention in addition to VR C: Usual activities conducted in-center (NR)
Vishnuram et al, 2024 [59]	I & C: Anaerobic (strength)	JioDive 360[0] (smartphone)	NR	None	I: 14 (+ NR IVR frequency) C: 14	I & C: NR	I: 30 (+30 - 90 for IVR) C: 30	I & C: 4	I & C: NR	I: Received C intervention in addition + PNM[o]
Musculoskeletal										
Gsangaga et al, 2023 [62]	I & C: Aerobic & Anaerobic (boxing)	PlayStation® VR	Commercial	None	I: 14 (+ IVR every 2) C: 14	I & C: NR	I: NR (+30 IVR) C: NR	I: 26 (first 12 C) C: 26	I & C: In-person (S) & Home (US)	I: Received C in addition to IVR C: NA

XSL•FO
RenderX

Reference	Intervention	IVR system	Game type	Additional equipment	Sessions/wk	Intensity	Session length (mins)	Length (wks)	Location	Cointerventions
Lo et al, 2024 [29]	I & C: Anaerobic (strength)	VR Shinecon 5.0 (smartphone)	Bespoke	Lower limb sensors with IVR	I & C: 5	I & C: NR	I & C: 30	I & C: 12	I & C: Home (US)	I & C: NA
Naqvi et al, 2022 [57]	I: Aerobic C: NR	Oculus Quest	Commercial	NR	I & C: 5	I & C: NR	I & C: 60	I & C: 4	I & C: NR (S)	I & C: NA
Nishitha et al, 2024 [58]	I: Aerobic & Anaerobic (strength) C: Anaerobic (strength)	NR	NR	None	I & C: 4	I: NR C: "High" (metric NR)	I: 40 C: NR	I & C: 12	I & C: In-person (S)	I: Nonimmersive VR exercise program included C: NA
Stamm et al, 2022 [55]	I: Aerobic & Anaerobic (strength) C: Anaerobic (strength)	HTC Vive	Bespoke	None	I & C: 3	I: 75% HRR C: NR	I & C: 30	I & C: 4	I & C: In-person (S)	I & C: Education sessions
Tuck et al, 2022 [66]	I & C: Aerobic, anaerobic (strength & boxing) TG: NA	HTC Vive	Commercial	None	I: 2 C: NR TG[p]: NA	I & C: NR TG: NA	I & C: NR TG: NA	I, C & TG: 6	I: In-person (S) C: Home (US) TG: NA	I & C: Education sessions TG: Waitlist control
Cardiopulmonary										
Rutkowski et al, 2022 [65]	I & C: Aerobic	VR TierOne	Bespoke	COSMED cycle ergometer with IVR connectivity	I & C: 7	I & C: 20% - 80% APHRM[q]	I & C: NR	I & C: 3	I & C: In-person (S)	I: Additional relaxation exercises in IVR C: NA
Wang et al, 2023 [53]	I: Aerobic & Anaerobic (boxing) C: NA	Pico Neo 3	Commercial	None	I: 7 C: NA	I: "low-medium" intensity (Borg RPE 12/20; 60% - 70% APHRM) C: NA	I: 20 - 30 C: NA	I & C: 12	I: In-person (S) C: NA	I: Breathing exercises C: Education and follow-ups
Cancer										
Schrempf et al, 2023 [54]	I: Aerobic C: NA	Oculus Quest 2	Commercial	None	I & C: 7	I: "Moderate" 50% - 70% APHRM C: NA	I & C: 30	I & C: 30 days	I: In-person (S & US) C: In-person (S)	I: NA C: Nonexercise physiotherapy treatment
Metabolic										

XSL•FO
RenderX

Reference	Intervention	IVR system	Game type	Additional equipment	Sessions/wk	Intensity	Session length (mins)	Length (wks)	Location	Cointerventions
Seo et al, 2023 [72]	I & TG: Aerobic C: NA	Smartphone HMD[r]	Commercial	Cycle ergometer with sensors for IVR	I & TG: 3 - 5 C: NA	I & TG: "Low" intensity (talk test) C: NA	I & TG: 50 C: NA	I, C & TG: 8	I & TG: Home (US) C: NA	I & TG: NA C: Daily activities with no intervention

[a]I: intervention.
[b]C: control.
[c]NA: not applicable.
[d]NR: not reported.
[e]S: supervised.
[f]US: unsupervised.
[g]IVR: immersive virtual reality.
[h]UC: usual care.
[i]MVPA: moderate-vigorous physical activity.
[j]HR: heart rate.
[k]HRR: heart rate reserve.
[l]Reps: repetitions.
[m]RPM: revolutions per minute.
[n]RPE: rating of perceived exertion.
[o]PNM: peripheral nerve mobilization.
[p]TG: third group.
[q]APHRM: age-predicated heart rate maximum.
[r]HMD: head-mounted display.

Outcomes

Data on all 30 clinical effectiveness outcome domains identified in this review are presented in Table S1 in Multimedia Appendix 1. Clinical effectiveness outcome domains progressing to meta-analyses included mobility and functional balance, condition severity, quality of life, functional leg strength, pain intensity, composite static and dynamic balance assessments, aerobic physical activity (device-measured), and static balance. Outcome domains included in meta-analyses are summarized in Table 3.

XSL•FO
RenderX

Table . Outcomes included in meta-analyses from included trials.

Reference	Intervention, mean (SD)		Control, mean (SD)	
	Baseline	Postintervention	Baseline	Postintervention
Older adults				
Barsasella et al, 2021 [60]				
Functional leg strength: 30 s Sit to Stand Test (repetitions)	21.6 (9.0)	22.0 (7.8)	19.8 (7.3)	19.8 (7.2)
Mobility and functional balance: 8-Foot Up and Go Tests	7.5 (2.7)	6.5 (2.5)[d]	7.0 (2.7)	5.7 (1.7)[d]
Static balance: Single Leg Stances	16.6 (9.6)	15.8 (10.1)	14.5 (10.4)	15.1 (9.7)
Campo-Prieto et al, 2022 [28]				
Functional leg strength: 5 × Sit to Stand Tests	15.6 (4.5)	13.8 (3.5)[c]	21.2 (12.6)	25.6 (14.2)[d]
Mobility and functional balance: Timed Up and Go Tests	17.9 (6.4)	19.0 (6.6)	23.2 (9.3)	26.3 (11.8)[d]
Quality of life: Short Form-12 Health Survey (SF12): Mental Component Summary (score)	48.8 (8.7)	53.4 (8.7)	50.4 (10.5)	55.2 (9.5)
Quality of life: Short Form-12 Health Survey (SF12): Physical Component Summary (score)	48.8 (9.5)	48.5 (9.0)[c]	42.1 (12.7)	44.4 (9.9)
Kershner et al, 2024 [63]				
Aerobic physical activity (device-measured): Weekly Total Physical Activity (min)	773.8 (243.0)	807.77 (189.6)	322.0 (137.5)	705.5 (244.1)
Lima Rêbelo et al, 2021 [64]				
Mobility and functional balance: Timed Up and Go Test (s)[b]	NR	−1.71 (2.3)[c]	NR[e]	−1.22 (3.6)
Static balance: Clinical Test of Sensory Interaction and Balance C.1 (score)[b]	NR	1.0 (6.8)	NR	1.5 (6.9)
Static balance: Clinical Test of Sensory Interaction and Balance C.2 (score)[b]	NR	2.6 (8.3)	NR	4.1 (10.6)[c]
Static balance: Clinical Test of Sensory Interaction and Balance C.3 (score)[b]	NR	3.9 (11.6)	NR	7.3 (12.1)
Static balance: Clinical Test of Sensory Interaction and Balance C.4 (score)[b]	NR	8.0 (11.8)[d]	NR	13.2 (14.6)[d]
Parmak et al, 2025 [69]				
Composite static and dynamic balance: Fullerton Advanced Balance Scale (score)	28.8 (4.7)	34.5 (3.8)[c]	27.3 (6.6)	30.2 (6.1)

XSL•FO
RenderX

Reference	Intervention, mean (SD)		Control, mean (SD)	
	Baseline	Postintervention	Baseline	Postintervention
Functional leg strength: 30 s Sit to Stand test (repetitions)	12.1 (3.2)	13.9 (2.4)[c]	10.7 (2.9)	11.8 (2.7)
Mobility and functional balance: Eight-Step Walk Tests	5.9 (0.9)	5.6 (0.9)	6.1 (1.1)	5.7 (1.1)
Quality of Life: World Health Organization Quality of Life Instrument: Older Adults (score)	80.3 (8.2)	83.9 (9.8)	74.4 (10.0)	77.6 (11.0)
Vorwerg-Gall et al, 2024 [56]				
Functional leg strength: 5 × Sit to Stand Tests[b]	NR	−1.0 (1.4)[d]	NR	−1.5 (1.9)[d]
Yalfani et al, 2024 [67]				
Functional leg strength: 30 s Sit to Stand Test (repetitions)	10.3 (1.8)	12.1 (1.6)[c,d]	10.1 (1.5)	9.3 (2.1)
Static balance: Center of Pressure Anteroposterior Sway (N/cm^2)[f]	18.5 (7.5)	11.3 (4.3)[c,d]	12.3 (2.64)	14.3 (3.7)
Static balance: Center of Pressure Mediolateral Sway (N/cm^2)	8.9 (2.3)	6.6 (0.5)[c,d]	7.7 (2.4)	8.3 (2.4)
Mobility and functional balance: Timed Up and Go Tests	12.6 (2.2)	9.6 (1.6)[c]	11.7 (2.1)	11.3 (2.2)
Zak et al, 2024 [68]				
Mobility and functional balance: Timed Up and Go Tests	14.0 (1.7)	12.6 (1.3)[d]	13.5 (1.1)	11.4 (1.1)[c,d]
Composite static and dynamic balance: Berg Balance Scale (score)	39.7 (1.3)	41.9 (2.4)[d]	39.3 (0.9)	42.6 (2.8)[c,d]
Neurological				
An and Park, 2022 [71]				
Functional leg strength: 5 × Sit to Stand Tests	54.1 (6.7)	35.2 (7.8)[c,d]	52.6 (6.5)	37.5 (10.0)[c]
Mobility and functional balance: Timed Up and Go Tests	23.2 (4.1)	14.5 (4.4)[c,d]	22.3 (3.2)	15.3 (4.7)[d]
Peláez-Vélez et al, 2023 [49]				
Composite static and dynamic balance: Tinetti Test Gait (score)	5.4 (3.0)	9.2 (2.9)[d]	3.6 (4.5)	5.8 (4.0)
Composite static and dynamic balance: Tinetti Test Balance (score)	8.6 (4.1)	13.6 (3.1)[d]	6.5 (6.1)	9.1 (5.7)
Composite static and dynamic balance: Berg Balance Scale (score)	27.0 (15.9)	46.0 (13.1)[d]	21.3 (22.8)	28.9 (20.4)

XSL•FO
RenderX

Reference	Intervention, mean (SD)		Control, mean (SD)	
	Baseline	Postintervention	Baseline	Postintervention
Condition severity: Motricity Index (score)	67.1 (31.7)	84.0 (23.1)[d]	72.7 (37.7)	75.7 (36.7)
Ramos et al, 2025 [70]				
Mobility and functional balance: Timed Up and Go Tests	8.5 (1.9)	7.9 (1.6)	8.7 (1.9)	8.9 (2.3)
Functional leg strength: 30 s Sit to Stand Test (repetitions)	11.0 (3.0)	11.0 (3.0)	9.0 (2.0)	11.0 (3.0)
Rodriguez-Fuentes et al, 2024 (a) [50]				
Functional leg strength: 5 × Sit to Stand Tests	14.3 (5.4)	13.8 (6.7)	13.1 (3.0)	13.2 (4.5)
Mobility and functional balance: Timed Up and Go Tests	14.1 (18.3)	10.2 (6.4)[b]	12.5 (9.2)	13.7 (4.5)
Composite static and dynamic balance: Tinetti Test (score)	22.4 (4.9)	26.4 (3.5)[c,d]	25.0 (3.9)	25.5 (3.6)
Quality of life: Parkinson disease Questionnaire (score)	7.0 (5.6)	6.4 (4.9)[c]	10.8 (7.5)	9.7 (6.5)
Condition severity: Movement Disorders Society Modified Unified Parkinson Disease Rating Scale (score)	21.9 (18.8)	15.1 (9.8)[d]	42.7 (31.3)	34.8 (23.4)
Rodriguez-Fuentes et al, 2024 (b) [51]				
Composite static and dynamic balance: Tinetti Test (score)	26.1 (2.1)	26.6 (1.5)	25.2 (2.9)	25.4 (2.3)
Functional leg strength: 5 × Sit to Stand Tests	14.4 (4.5)	12.8 (3.8)[d]	12.7 (4.7)	12.1 (2.8)
Mobility and functional balance: Timed Up and Go Tests	10.7 (5.6)	9.2 (4.7)[d]	8.0 (1.9)	8.7 (2.3)
Musculoskeletal				
Gsangaya et al, 2023 [62]				
Pain intensity: Numerical Pain Rating Scale (score)	4.6 (0.8)	0.4 (0.5)[c,d]	4.8 (0.9)	0.9 (0.7)[d]
Condition severity: International Knee Documentation Committee Score (score)	82.9 (3.2)	93.9 (1.0)[b,d]	85.0 (3.8)	92.9 (1.2)[d]
Lo et al, 2024 [29]				
Pain intensity: Numerical Pain Rating Scale (score)	5.9 (1.9)	4.8 (1.8)	4.8 (1.5)	4.6 (2.4)
Condition severity: Western Ontario and McMaster Universities Osteoarthritis Index (score)	902.9 (454.8)	791.3 (425.1)	752.0 (549.3)	742.4 (510.5)

XSL•FO
RenderX

Reference	Intervention, mean (SD)		Control, mean (SD)	
	Baseline	Postintervention	Baseline	Postintervention
Quality of life: EuroQol Visual Analog Scale (score)	62.0 (19.6)	71.6 (12.8)	67.3 (17.2)	67.7 (17.4)
Aerobic physical activity (device-measured): MET/d (score)	34.2 (1.2)	34.4 (1.3)	35.0 (1.4)	34.9 (2.0)
Naqvi et al, 2022 [57]				
Pain intensity: Visual Analog Scale (score)	7.0 (0.7)	1.77 (0.4)[c,d]	7.5 (0.5)	4.2 (0.3)[d]
Condition severity: Disabilities of the Arm, Shoulder, and Hand Questionnaire (score)	80.2 (1.7)	13.0 (1.7)[c,d]	80.7 (2.8)	28.1 (6.6)[d]
Nishitha et al, 2024 [58]				
Pain intensity: Numerical Pain Rating Scale (score)	3.4 (0.5)	1.5 (0.5)[b]	3.8 (0.4)	2.3 (0.5)
Condition severity: Western Ontario and McMaster Universities Osteoarthritis Index (score)	51.2 (1.3)	14.9 (1.0)[b]	62.1 (1.6)	19.0 (2.0)
Mobility and functional balance: Timed Up and Go Tests	23.2 (1.6)	8.6 (1.1)[b]	28.7 (1.3)	12.1 (0.9)
Stamm et al, 2022 [55]				
Pain intensity: Numerical Pain Rating Scale (score)	3.6 (2.4)	2.9 (2.0)	2.9 (2.4)	1.6 (1.5)
Condition severity: Hannover Functional Ability Questionnaire (score)	73.1 (10.6)	81.8 (11.2)[d]	69.8 (16.8)	72.7 (15.7)
Quality of life: Short Form-12 Health Survey (SF12): Mental Component Summary (score)	46.4 (10.6)	48.4 (7.1)	50.3 (7.7)	56.2 (4.8)[d]
Quality of life: Short Form-12 Health Survey (SF12): Physical Component Summary (score)	41.0 (7.8)	39.3 (8.0)	35.9 (7.9)	37.8 (7.3)
Tuck et al, 2022 [66]				
Pain intensity: Brief Pain Inventory (score)[a,b]	8.4 (1.8)	−1.0 (0.9)	8.1 (1.2)	−0.2 (2.3)
Aerobic physical activity (device-measured): Daily Active Minutes (mins)[a,b]	NA	19.5 (64.5)	NA	−21.1 (91.5)
Cardiopulmonary				
Rutkowski et al, 2022 [65]				
Quality of life: World Health Organization Quality of Life Scale (score)	61.6 (16.2)	62.8 (14.1)	59.7 (17.9)	62.9 (16.8)[d]
Wang et al, 2023 [53]				
Quality of life: 36-Item Short Form Survey (score)	66.6 (12.0)	79.7 (11.6)[d]	66.3 (11.5)	76.2 (11.9)

XSL•FO
RenderX

Reference	Intervention, mean (SD)		Control, mean (SD)	
	Baseline	Postintervention	Baseline	Postintervention
Cancer				
Schrempf et al, 2023 [54]				
Quality of life: EQ-5D-5L (score)	0.9 (0.1)	0.8 (0.2)	0.9 (0.1)	0.8 (0.2)

[a]Exercising control group.

[b]Change from baseline.

[c]Between-group difference.

[d]Within-group difference.

[e]NR: Not reported.

[f]N/cm^2: Newtons per square centimeter.

Meta Analyses

The between-group analyses for IVR versus comparator, grouped by comparator group activity (exercising/nonexercising) are presented in Figure 2 [28,50,51,58,60,64,67-71] for mobility and functional balance (exercising: SMD −0.345, 95% CI −1.095 to 0.406; *P*=.29; nonexercising: SMD −0.322, 95% CI −0.931 to 0.288; *P*=.22), Figure 3 [28,50,51,56,60,67,69-71] for functional leg strength (exercising: SMD −0.161, 95% CI −0.573 to 0.250; *P*=.34; nonexercising: SMD −0.351, 95% CI −1.750 to 1.049; *P*=.48), Figure 4 [28,29,50,53-55,65,69] for quality of life (exercising: SMD 0.036, 95% CI −0.444 to 0.516; *P*=.85; nonexercising: SMD −0.053, 95% CI −0.839 to 0.728; *P*=.80), Figure 5 [29,49,50,55,57,58,62] for condition severity (SMD −0.153, 95% CI −2.227 to 1.921; *P*=.86), Figure 6 [29,55,57,58,62,66] for pain intensity (SMD −0.783, 95% CI −2.069 to 0.502; *P*=.18), Figure 7 [49-51,68,69] for composite static and dynamic balance assessments (SMD −0.310, 95% CI −0.870 to 0.249; P=.39), Figure 8 [60,64,67] for static balance (SMD −0.189, 95% CI −2.020 to 1.642; *P*=.70) and Figure 9 [29,63,66] for aerobic physical activity (device-measured; SMD 0.145, 95% CI −2.427 to 2.717; *P*=.83).

Figure 2. Between-group random effects meta-analysis of mobility and functional balance assessments, grouped by comparator group activity [28,50,51,58,60,64,67-71]. D1: Domain 1; D2: Domain 2; D3: Domain 3; D4: Domain 4; D5: Domain 5; DD: Developmental Disability; HKSJ: Hartung-Knapp-Sidik-Jonkman; iSCI: Incomplete Spinal Cord Injury; MS: Multiple Sclerosis; O: Overall; PD: Parkinson disease; SMD: standardized mean difference; TKR: Total Knee Replacement.

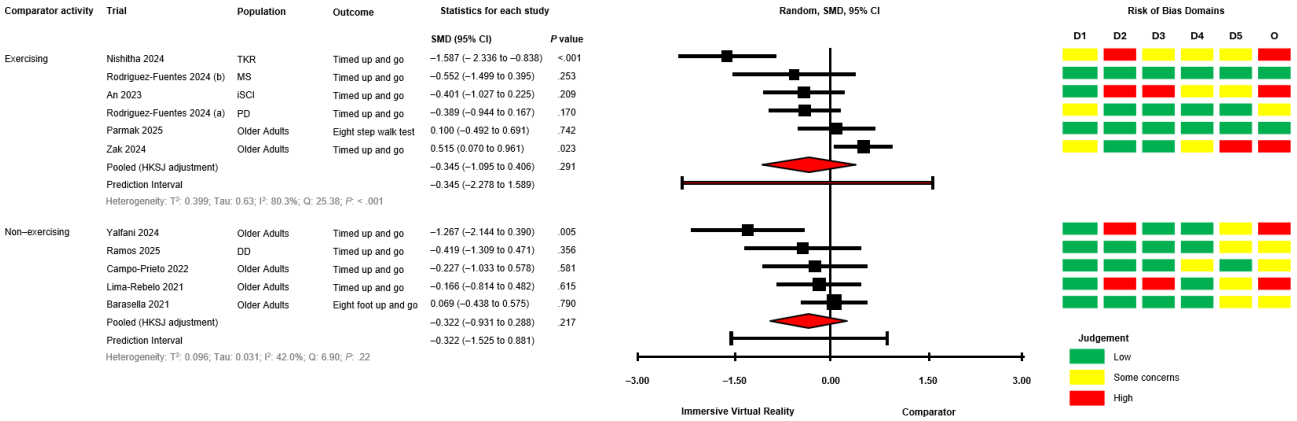

XSL•FO
RenderX

Figure 3. Between-group random effects meta-analysis of functional leg strength assessments, grouped by comparator group activity [28,50,51,56,60,67,69-71]. 30 s STS: 30-seconds Sit to Stand Test; 5 × STS: 5-times Sit to Stand Test; D1: Domain 1; D2: Domain 2; D3: Domain 3; D4: Domain 4; D5: Domain 5; DD: Developmental Disability; HKSJ: Hartung-Knapp-Sidik-Jonkman; iSCI: Incomplete Spinal Cord Injury; MS: Multiple Sclerosis; O: Overall; PD: Parkinson disease; SMD: standardized mean difference.

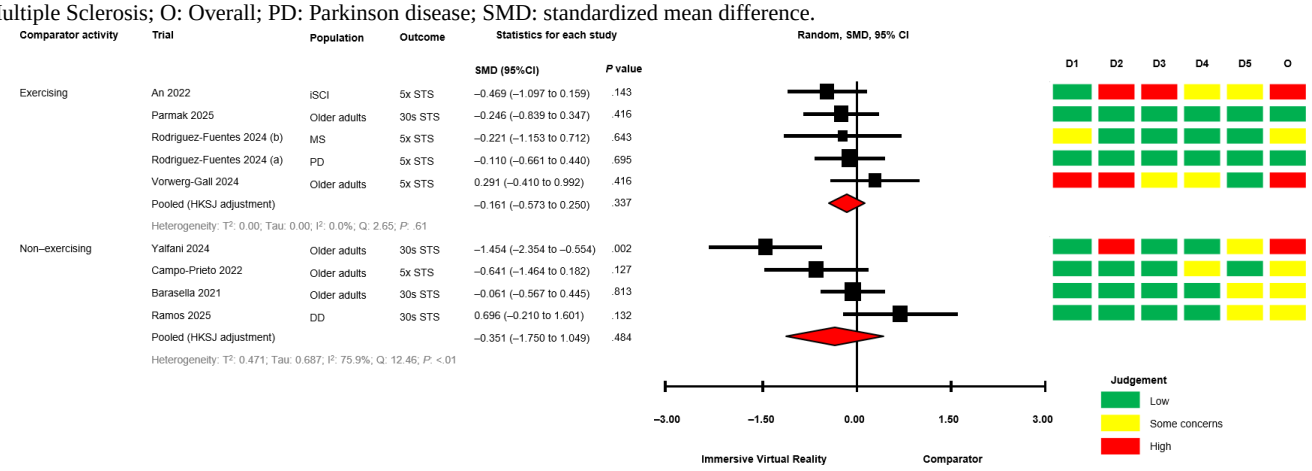

Figure 4. Between-group random effects meta-analysis of quality of life assessments, grouped by comparator group activity [28,29,50,53-55,65,69]. Ca: Cancer; CBP: Chronic Back Pain; CHD: Coronary Heart Disease; D1: Domain 1; D2: Domain 2; D3: Domain 3; D4: Domain 4; D5: Domain 5; EQ-VAS: EQ–Visual Analog Scale; HKSJ: Hartung-Knapp-Sidik-Jonkman; O: Overall; OA: Osteoarthritis; PD: Parkinson disease; PDQ-39: Parkinson disease questionnaire; SF-12: Short Form-12 Health Survey; SF-36: 36-Item Short Form Survey; SMD: standardized mean difference; WHOQOL-BREF: World Health Organization quality of life scale; WHOQOL-OA: World Health Organization quality of life instrument: Older adults.

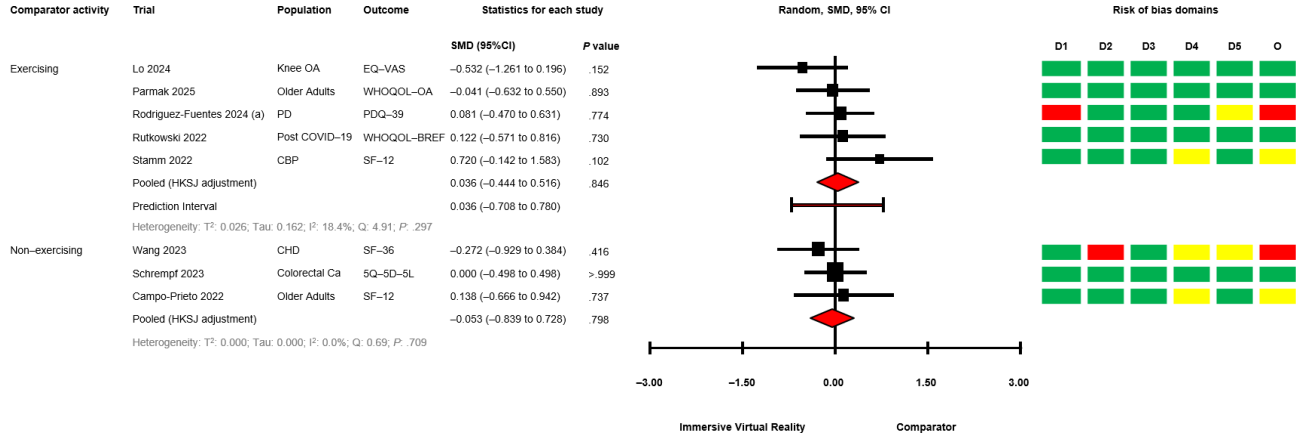

Figure 5. Between-group random effects meta-analysis of condition severity assessments for trials with an exercising comparator group [29,49,50,55,57,58,62]. ACLR: Anterior Cruciate Ligament Reconstruction; CLBP: Chronic Lower Back Pain; D1: Domain 1; D2: Domain 2; D3: Domain 3; D4: Domain 4; D5: Domain 5; DASH: Disability of the Arm, Shoulder, and Hand Questionnaire; Ffb-H-R: Hannover Functional Ability Questionnaire; HKSJ: Hartung-Knapp-Sidik-Jonkman; IKDC: International Knee Documentation Committee Score; MDS_UPDRS: Movement Disorders Society Modified Unified Parkinson Disease Rating Scale; O: Overall; OA: Osteoarthritis; PD: Parkinson disease; SMD: standardized mean difference; TKA: Total Knee Arthroplasty; WOMAC: Western Ontario and McMaster Universities Osteoarthritis Index.

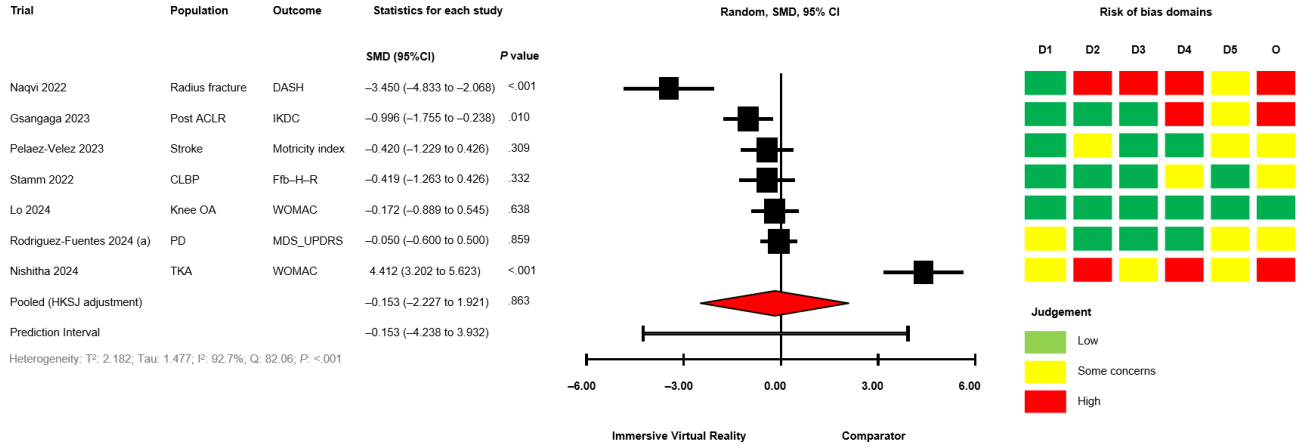

XSL•FO
RenderX

Figure 6. Between-group random effects meta-analysis of pain intensity assessments for trials with an exercising comparator group [29,55,57,58,62,66]. ACLR: Anterior Cruciate Ligament Reconstruction; BPI: Brief Pain Inventory; CLBP: Chronic Lower Back Pain; D1: Domain 1; D2: Domain 2; D3: Domain 3; D4: Domain 4; D5: Domain 5; HKSJ: Hartung-Knapp-Sidik-Jonkman; Musc: Musculoskeletal; NPRS: Numerical Pain Rating Scale; O: Overall; OA: Osteoarthritis; SMD: standardized mean difference; TKA: Total Knee Arthroplasty; VAS: Visual Analog Scale.

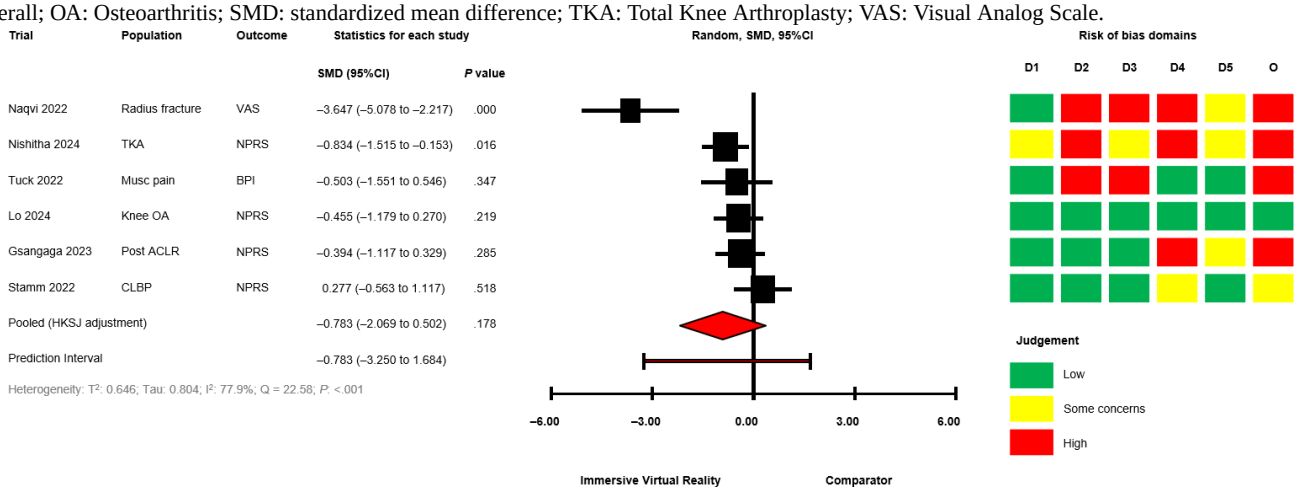

Figure 7. Between-group random effects meta-analysis of composite static and dynamic balance assessments for trials with an exercising comparator group [49-51,68,69]. BBS: Berg Balance Scale; D1: Domain 1; D2: Domain 2; D3: Domain 3; D4: Domain 4; D5: Domain 5; FABT: Fullerton Advanced Balance Scale; HKSJ: Hartung-Knapp-Sidik-Jonkman; MS: Multiple Sclerosis; O: Overall; PD: Parkinson disease; SMD: standardized mean difference.

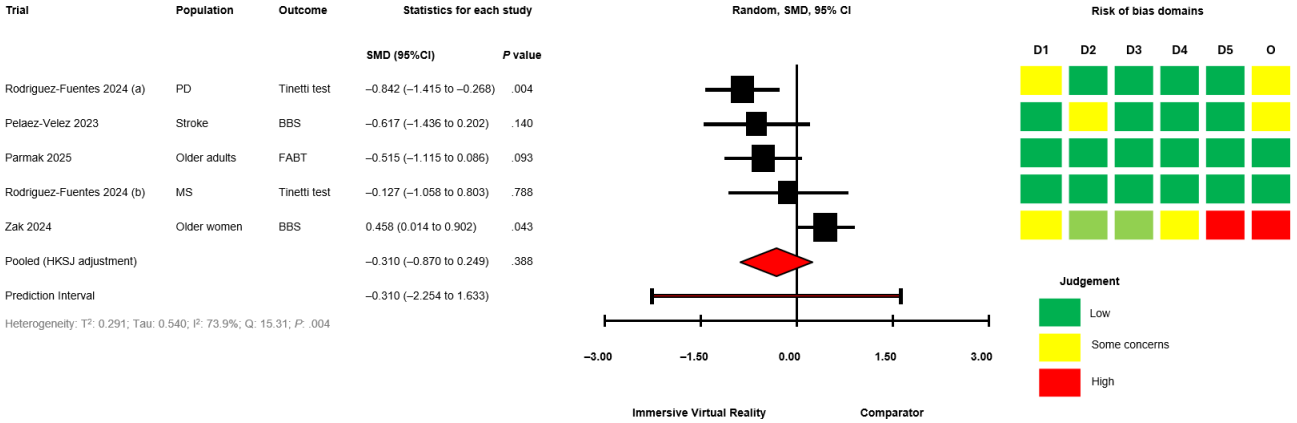

Figure 8. Between-group random effects meta-analysis of static balance assessments for trials with a nonexercising comparator group [60,64,67]. CoP: Center of Pressure; CTSIB: Clinical Test of Sensory Interaction and Balance Test; D1: Domain 1; D2: Domain 2; D3: Domain 3; D4: Domain 4; D5: Domain 5; HKSJ: Hartung-Knapp-Sidik-Jonkman; O: Overall; SLS: Single Leg Stand test; SMD: standardized mean difference.

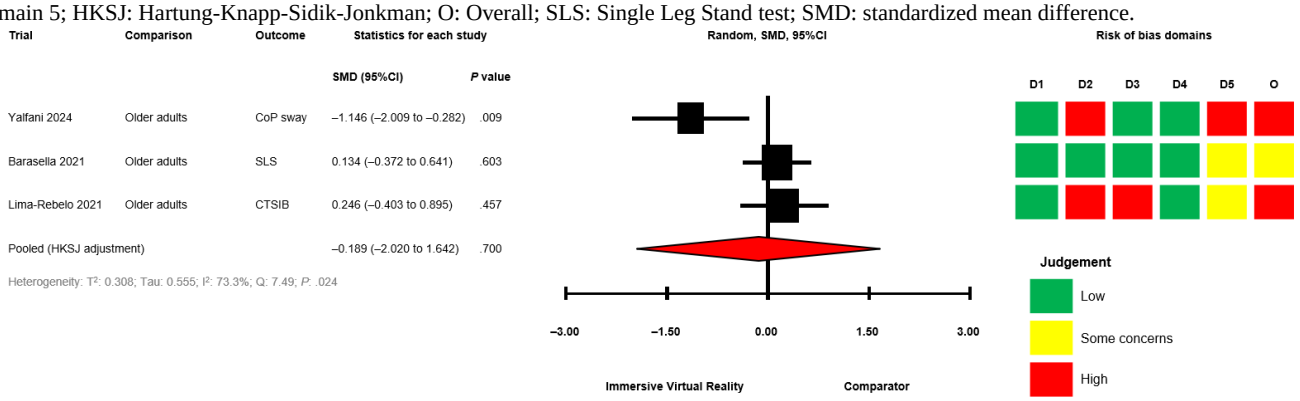

XSL•FO
RenderX

Figure 9. Between-group random effects meta-analysis of aerobic physical activity (device-measured) assessments for trials with an exercising comparator group [29,63,66]. D1: Domain 1; D2: Domain 2; D3: Domain 3; D4: Domain 4; D5: Domain 5; HKSJ: Hartung-Knapp-Sidik-Jonkman; Met: Metabolic Equivalent; Musc: Musculoskeletal; O: Overall; OA: Osteoarthritis; PA: Physical Activity; SMD: standardized mean difference.

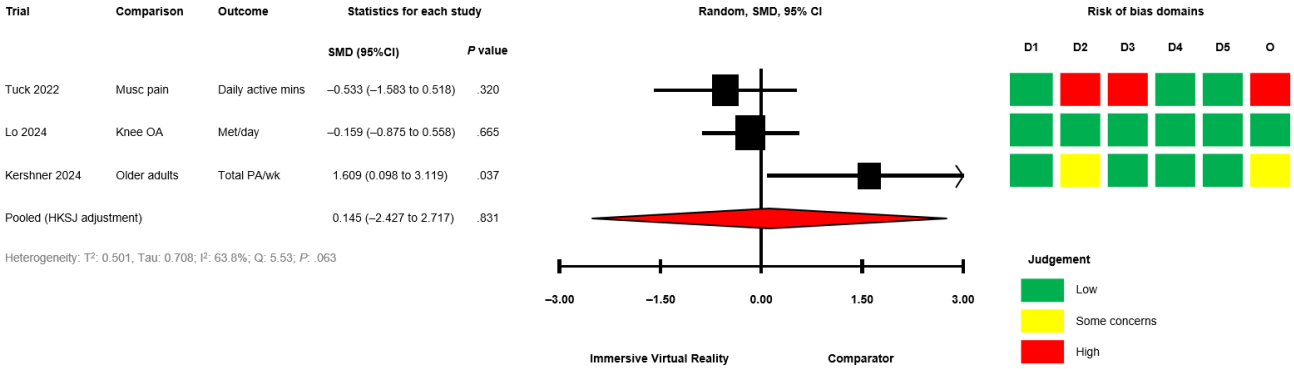

No significant effects were observed compared with exercising and nonexercising comparators for any outcome domain. Substantial heterogeneity was identified among most exercising (mobility and functional balance: T^2=0.399, Tau=0.630, I^2=80.3%, Q=25.38, P<.01; condition severity: T^2=2.182, Tau=1.477, I^2=92.7%, Q=82.06, P<.01; pain intensity: T^2=0.646, Tau=0.804, I^2=77.9%, Q=22.58, P<.01; composite static and dynamic balance: T^2=0.291, Tau=0.540, I^2=73.9%, Q=15.31, P<.01) and nonexercising (functional leg strength: T^2=0.471, Tau=0.687, I^2=75.9%, Q=12.46, P<.01; static balance: T^2=0.308, Tau=0.555, I^2=73.3%, Q=7.49, P=.02) comparator meta-analyses. Statistical heterogeneity was not identified for functional leg strength (T^2=0.00, Tau=0.00, I^2=0.0%, Q=2.65, P=.61), quality of life (T^2=0.026, Tau=0.162, I^2=18.4%, Q=4.91, P=.30) or aerobic physical activity (device-measured; T^2=0.501, Tau=0.708, I^2=63.8%, Q=5.53, P=.06) exercising meta-analyses, and mobility and functional balance (T^2=0.096, Tau=0.031, I^2: 42.0%, Q=6.90, P=.22) and quality of life (T^2=0.000, Tau=0.000, I^2=0.0%, Q=0.69, P=.71) nonexercising meta-analysis. Calculated prediction intervals were often substantially wider than the 95% CI, indicating significant heterogeneity for most outcome domains (Figures 2-9). Visual inspection of funnel plots did not suggest the presence of small-study effects for all outcome domains (Figures S1-S8 in Multimedia Appendix 1). However, the Egger statistic indicated evidence for a small-study effect for mobility and functional balance outcomes (Figure 2; P=.01). Sensitivity analyses identified single study removal of Zak et al [68] (functional balance, composite static and dynamic balance), Nishitha et al [58] (condition severity), Naqvi et al [57] (pain intensity), Stamm et al [55] (pain intensity), and Barsasella et al [60] (functional leg strength) resulted in altered statistical significance in meta-analytical models. Manipulation of pre-post correlation values did not result in altered statistical significance for any meta-analysis.

Primary Clinical Effectiveness Outcome Domains

Overview

The following section is a narrative description of outcomes under each of the 8 primary clinical effectiveness domains. These include outcomes that were unable to be included in meta-analyses due to a lack of pooled comparative data (ie, less than 3 comparable control groups across trials; Table 3, Figures S9-S16 in Multimedia Appendix 1).

Mobility and Functional Balance

Mobility and functional balance assessments were reported in 12 trials (46% of total) [28,50-52,58,60,64,67-71]. Between-group differences favoring IVR (P<.05) were reported in 4 trials (3 with exercising comparators) [50,58,67,71]. One trial reported a between-group difference favoring an exercising comparator group [68]. Between-group differences were either not statistically significant or unreported in 7 trials [28,51,52,60,64,69,70]. Within-group improvements (P<.05) for IVR groups were reported in 5 trials [51,58,60,64,71], and comparator interventions in 4 trials (3 exercising comparators) [52,60,68,71]. One trial reported a within-group deterioration in a nonexercising comparator group [28].

Condition Severity

Assessments of change in individual condition severity were reported in 8 trials (31% of total) [29,49,50,52,55,57,58,62]. Between-group differences favoring IVR (P<.05) were reported in 3 trials (3 with exercising comparators) [57,58,62]. Between-group differences were either not statistically significant or unreported in 5 trials [29,49,50,52,55]. Within-group improvements (P<.05) for IVR were reported in 4 trials [49,57,58,62], and exercising comparators in 2 trials [57,62].

Quality of Life

Assessments of change in participant quality of life were reported in 8 trials (31% of total) [28,29,50,53-55,65,69]. Between-group differences favoring IVR (P<.05) were reported in one trial with a nonexercising comparator [28], and one with an exercising comparator [50]. Between-group differences were either not significant or unreported in 6 trials [29,53-55,65,69]. Within-group improvement (P<.05) for a virtual reality group was reported in one trial [55] and for the mental composite score of the Short Form-12 Health Survey for an exercising comparator [55]. One trial reported a deterioration of the environmental subscale of the World Health Organization Quality of Life Scale questionnaire in an exercising comparator [65].

XSL•FO
RenderX

Functional Leg Strength

Measurements of functional leg strength were reported in 9 trials (35% of total) [28,50,51,56,60,67,69-71]. Between-group differences ($P<.05$) favoring IVR were reported in 4 trials (2 with exercising comparators) [28,67,69,71]. Between-group differences were either not significant or unreported in 5 trials (3 exercising comparators) [50,51,56,60,70]. Within-group improvements ($P<.05$) for virtual reality groups were reported in 3 trials [51,56,67,71], and 2 exercising comparators [56,71]. One trial reported significant within-group deterioration in a nonexercising comparator [28].

Pain Intensity

Pain intensity outcomes were reported in 6 trials (23% of total) [29,55,57,58,62,66]. Between-group differences ($P<.05$) favoring IVR were reported in 3 trials with exercising comparators [57,58,62]. Between-group differences were either not significant or unreported in 3 trials with exercising comparators [29,55,66]. Within-group improvements ($P<.05$) for IVR groups were reported in 2 trials and their exercising comparators [57,62].

Composite Static and Dynamic Balance Assessments

Composite static and dynamic balance assessments were described in 6 trials (23% of total) [28,49-51,68,69]. Three trials reported between-group differences ($P<.05$) favoring IVR (2 with exercising comparators) [28,50,69]. One trial reported a between-group difference favoring an exercising comparator [68]. Between-group differences were either not significant or unreported in 2 trials with exercising comparators [49,51]. Within-group improvements for IVR groups were reported in 4 trials [28,49,50,68], and one exercising comparator [73].

Aerobic Physical Activity (Device-Measured)

Physical activity was assessed using device-based measures in 3 trials (12% of total) [29,63,66]. No trials reported any between- or within-group differences ($P<.05$) for any measure.

Static Balance

Static balance outcomes were reported in 4 trials (15% of total) [60,64,67,68]. A between-group difference ($P<.05$) favoring IVR was reported for one trial with a nonexercising comparator [67]. Within-group improvement ($P<.05$) favoring IVR was reported in 3 trials [64,67,68], and 2 comparators (one exercising) [64,68].

Secondary Feasibility Outcome Measures

Feasibility outcomes are reported in Table 4.

XSL•FO
RenderX

Table . Session attendance, exercise adherence, technological issues, safety, simulator sickness, and participant satisfaction for included trials.

Reference	Session attendance	Exercise adherence	Technological issues (IVR[b])	Motion sickness (IVR)	Safety	Participant experiences (IVR)
Older adults						
Barsasella et al, 2021 [60]	I[c] & C[d]: NR[e]	I & C: NR	NR	Simulator sickness description No SS[f] reported	Adverse events I & C: None observed	NR
Campo-Prieto et al, 2022 [28]	I & C: NR	Percentage of participants completing all sessions I: 92.3% completed all sessions C: NR	NR	SSQ[g] No SS reported due to VR	Adverse events I & C: None observed	SUS[h] (total score) SUS: 73.96 (SD 16.77), GEQ PE (domain score) 2.87 (SD 0.86), GEQ NE (domain score) 0.22 (SD 0.21), GEQ T (domain score) 0.37 (SD 0.43), GEQ RR (domain score) 0.19 (SD 0.33), Ad hoc questionnaire Generally positive responses regarding IVR
Drazich et al, 2023 [61]	Mean sessions attended I: 15/16 C: NA	I & C: NR	NR	SSQ No SS reported due to IVR	I & C: NR	AIM[n]: Acceptability (% of participants) 100% found the intervention acceptable; 90% found the intervention highly acceptable AIM[o]: Appropriateness (% of participants) 100% found the intervention appropriate; 70% found the intervention highly appropriate
Kershner et al, 2024 [63]	Total coaching sessions attended I: 59/60 (98%) C: 46/48 (96%)	IVR sessions completed I: 5.5 [0.5][a] C: NA	NR	VRSQ[p] (total score) 2.0 [3.0[a]] (slight or no SS) 44% indicated "no feelings of sickness or discomfort"	Adverse events I & C: None observed	Enjoyment scale (−5 to +5) 5.0 [.0][a] Participant survey responses Positive responses in the intervention group for IVR enjoyment and immersiveness
Kwan et al, 2021 [52]	Attendance rate for completers I & C: 100%	NR	NR	VRSQ (total score) 4.63 [18.33[a]]	Study withdrawals I: 1 (experienced mild SS – VRSQ 18.3/100) C: 2 (moderate pain in lower limbs)	NR
Lima Rêbelo et al, 2021 [64]	I & C: NR	Sessions completed I: 14.2 (SD 2.3) 89% C: 14.5 (SD 2.1) 91%	NR	NR	Serious adverse events I & C: None observed	NR
Parmak et al, 2025 [69]	Participation rate I & C: 100%	I & C: NR	NR	VRSQ (total score) 0.09 (SD 0.29)	Adverse events I & C: None observed	VAS satisfaction, I: 9.32 (SD 1.09) C: 7.45 (SD 1.54)
Vorwerg-Gall et al, 2024 [56]	I & C: NR	I & C: NR	NR	NR	I & C: NR	NR

XSL•FO
RenderX

Reference	Session attendance	Exercise adherence	Technological issues (IVR[b])	Motion sickness (IVR)	Safety	Participant experiences (IVR)
Yalfani et al, 2024 [67]	I & C: NR	I & C: NR	NR	NR	I & C: NR	NR
Zak et al, 2024 [68]	I & C: NR	I & C: NR	NR	NR	Adverse events I & C: None observed	NR
Neurological						
An and Park, 2022 [71]	I & C: NR	I & C: NR	NR	NR	I & C: NR	NR
Peláez-Vélez et al, 2023 [49]	I & C: NR	I & C: NR	NR	NR	I & C: NR	NR
Ramos et al, 2025 [70]	Sessions attended I: 92.7% C: NA	I: NR C: NA	NR	NR	Adverse events I: None observed C: NA[m]	NR
Rodriguez-Fuentes et al, 2024 (a) [50]	Average sessions attended I & C: 42 [range 40 - 48 (whole sample)]	I & C: NR	NR	SSQ (Nausea) 7% of intervention group SSQ (Oculomotor) 2.93% of intervention group SSQ (Disorientation) 6.29% of intervention group	Adverse events I & C: None observed	SUS (total score) 82.9/100 GEQ PE[i] (domain score) 2.92/4 GEQ NE[j] (domain score) .04/4 GEQ T[k] (domain score) .23/4 GEQ RR[l] (domain score) .15/4
Rodriguez-Fuentes et al, 2024 (b) [51]	I & C: NR	Program adherence I: 100% C: NR	NR	SSQ (total score) 1.37/48 (range: 0 - 6)	Adverse events I: None observed C: NR	SUS (total score) 90.31/100 (range 72.5 - 100)
Vishnuram et al, 2024 [59]	NR	NR	NR	NR	I & C: NR	NR
Musculoskeletal						
Gsangaya et al, 2023 [62]	Session attendance I: 100% C: 100%	Program adherence I: 100% C: 100%	NR	NR	Adverse events I & C: None observed	NR
Lo et al, 2024 [29]	NR	Exercise adherence I: 77.2% [36.8 - 104%] C: 62.1% [40.4 - 166.0%]	Participant qualitative interviews Qualitative themes show technological challenges and inconvenience of HMD[q] during exercise	Number of participants reporting SS after exercise 5 Participant qualitative interviews 2 (13%) of participants reported SS throughout the intervention General themes of SS	Adverse events I & C: None observed	Participant qualitative interviews Qualitative outcomes indicate a mix of positive and negative experiences regarding satisfaction and acceptability using the HMD for exercise
Naqvi et al, 2022 [57]	NR	NR	NR	NR	I & C: NR	NR
Nishitha et al, 2024 [58]	NR	NR	NR	NR	I & C: NR	NR

XSL•FO
RenderX

Reference	Session attendance	Exercise adherence	Technological issues (IVR[b])	Motion sickness (IVR)	Safety	Participant experiences (IVR)
Stamm et al, 2022 [55]	NR	Training units completed, I: 9.18 (SD 1.47) (77%), C: 10.81 (SD 1.6) (90%)	Observed technical issues Connection issues with base stations and internet	NR	Serious adverse events I & C: None observed	Qualitative description of UEQ[r] IVR was rated as at least "above average" on each individual subscale. "Excellent" rating for attractiveness and perspicuity. "Good" rating for efficiency, dependability, and stimulation "Above average" for originality
Tuck et al, 2022 [66]	NR	Session completion I: All participants completed ≥7 sessions. 30% (3/10) completed 12. C & TG: NR	NR	NR	Adverse events I: None observed C, TG: NR	Participant qualitative interviews Generally positive attitudes towards VR for enjoyment and perceived benefits
Cardiopulmonary						
Rutkowski et al, 2022 [65]	NR	NR	NR	NR	NR	NR
Wang et al, 2023 [53]	NR	NR	NR	NR	MACE[s] incidence I: 11.11% (1x angina, 1 x arrhythmia) C: 16.67% (1x arrhythmia, 2 x angina, 1 x restenosis)	NR
Cancer						
Schrempf et al, 2023 [54]	NR	Sessions performed as planned I: 75.6% C: NR	NR	NR	Adverse events I: None observed C: NR	EORTC[t] (total score) I: 72.6
Metabolic						

XSL•FO
RenderX

Reference	Session attendance	Exercise adherence	Technological issues (IVR[b])	Motion sickness (IVR)	Safety	Participant experiences (IVR)
Seo et al, 2023 [72]	NR	NR	NR	NR	NR	”Exercise fun” 4-point Likert scale I: 32.82 (SD 4.61/40)

[a]Median [IQR].
[b]IVR: immersive virtual reality.
[c]I: intervention.
[d]C: control.
[e]NR: not reported.
[f]SS: simulator sickness.
[g]SSQ: simulator sickness questionnaire.
[h]SUS: system usability scale.
[i]GEQ PE: game experience questionnaire - positive experience.
[j]GEQ NE: game experience questionnaire - negative experience.
[k]GEQ T: game experience questionnaire - tiredness.
[l]GEQ RR: game experience questionnaire - returning to reality.
[m]NA: not applicable.
[n]AIM: acceptability of intervention measure
[o]AIM: appropriateness of intervention measure.
[p]VRSQ: virtual reality sickness questionnaire.
[q]HMD: head-mounted display.
[r]UEQ: user experience questionnaire.
[s]MACE: major adverse cardiovascular event.
[t]EORTC: European Organization for Research and Treatment of Cancer inpatient satisfaction 32-item questionnaire.

Safety Metrics

Safety metrics were reported in 16 trials (62% of included studies) [28,29,50-55,60,62-64,66,68-70]. Fifteen studies reported adverse events incidence [28,29,50,51,53-55,60,62-64,66,68-70], and one trial reported safety-related study withdrawals for the IVR and exercising comparator groups [52]. No study-related adverse events were reported. No trials reported an increased number of adverse events in the IVR groups compared with comparators. Fourteen trials (88% of trials reporting safety metrics) reported no adverse events in any study group [28,29,50,51,54,55,60,62-64,66,68-70].

Session Attendance/Exercise Adherence

Session attendance was reported in 7 trials (27% of total) [50,52,61-63,69,70]. For trials differentiating between groups, attendance was 98% for IVR exercise sessions and 99% for exercising comparators [52,61-63,69]. Adherence to exercise prescription was reported in 9 trials (35% of total). For trials providing a calculable percentage of session completion by participants, IVR groups completed an average 87% of sessions (n=7) [28,29,51,54,55,62,64], and comparator interventions completed 94% (n=3) [55,62,64].

Motion Sickness

Motion sickness in response to IVR exposure was reported in 9 trials (32% of total) [28,29,50-52,60,61,63,69]. Generally, low levels of motion sickness were reported after IVR exercise interventions via quantitative assessments. One trial described low levels of motion sickness through participant qualitative interviews [29]. One trial terminated the intervention for one participant due to persistent motion sickness symptoms [29].

Technical Issues

Two trials (8% of total) [29,55] reported technical issues for IVR interventions. Common technical issues reported were primarily due to disturbance to the Internet connection and HMD inconvenience.

Participant Experiences

Eleven trials (42% of total) [28,29,50,51,54,55,61,63,66,69,72] reported measures related to participant experiences with IVR exercise post intervention. For qualitative assessments, responses generally indicated positive user experiences and high usability for IVR, with participants citing overall enjoyment, motivation, and beneficial perceived benefits [29,66]. Negative experiences with IVR primarily surrounded the discomfort and inconvenience of using an HMD during exercise sessions [29].

Risk of Bias in Trials and Confidence in Cumulative Outcomes

Risk of bias assessments for individual trials progressing to meta-analyses are presented alongside forest plots in (Figures 2-9). Risk of bias proportions for separate outcome domains are presented in Table S2 in Multimedia Appendix 1. Many trials exhibited issues with measurement and selection of reported results across all outcome domains (domains 4 and 5 on the Cochrane RoB2 tool). In turn, 75% of outcomes included across all meta-analyses were rated as “some concerns” (n=18) or at “high” risk of overall bias (n=21). GRADE assessments for outcome domains progressing to meta-analysis are presented

XSL•FO
RenderX

in Table S2 in Multimedia Appendix 1. All outcome domains were deemed to have "low" or "very low" certainty levels. Risk of bias, indirectness of intervention, and imprecision of effects had the main impacts on the level of certainty.

Discussion

Principal Findings

This systematic review with meta-analyses assessed the clinical effectiveness of IVR interventions using aerobic or anaerobic exercise. The analysis combined 26 RCTs, including 846 participants. Population groups included adults with neurological disorders, older adults, musculoskeletal conditions, cardiopulmonary diseases, metabolic conditions, and cancer. Pooled data indicated no differences compared with the comparator for clinical effectiveness, with high levels of statistical and clinical heterogeneity and low certainty ratings leading to decreased confidence in findings. While scarcely reported among included trials, findings were generally positive regarding the feasibility of IVR exercise. This has important ramifications for both practitioners who are seeking to adopt IVR as a novel alternative for exercise delivery and researchers conducting future trials.

Clinical Effectiveness of Immersive Virtual Reality Exercise Interventions

For all clinical effectiveness outcome domains progressing to meta-analysis, certainty in effect was classified as "very low" or "low." Many included trials exhibited issues with the measurement and selection of reported results, leading to increased risk of bias. Limited comparative data, small sample sizes, substantial heterogeneity, and increased risk of bias across all outcomes decreased certainty in all findings. This outlines a need for robust clinical trials to be conducted to improve confidence in observed findings.

Pooled data seem to suggest that aerobic or anaerobic exercise via IVR may be comparable to exercising controls. This observation seems to be common across supervised and unsupervised comparator groups across populations. Recent evidence suggests that IVR may be effective for mitigating short-term pain intensity during and post exercise (≈30% reduction) [74], although evidence on longer-term pain is unclear [75]. Attentional distraction and altered pain modulation may be potential mechanisms enhanced in IVR [76]. This review identified a nonsignificant effect favoring IVR for pain intensity compared with exercising controls, marred by high statistical and clinical heterogeneity across population groups. The relationship between IVR exercise and pain intensity outcomes should be emphasized in future robust clinical trials.

No significant differences were observed between IVR and nonexercising comparators for all meta-analysis outcome domains. This observation may be in part due to a low number of trials with nonexercising comparators progressing to meta-analyses. This may have led to a potential lack of power to detect any differences in the meta-analytical models. Exercise prescription, supplementary equipment, and additional intervention components received by the IVR groups varied substantially across all trials. While it cannot be expected that all weekly physical activity be completed in IVR, most included trials either did not meet the World Health Organization physical activity [77] or American College of Sports Medicine resistance exercise guidelines [78] (inclusive of other intervention elements) or did not report exercise prescription in sufficient detail. This indicates a potential issue with appropriate intervention dosage and a lack of reporting for activities being completed outside of the trial context. This lack of clarity in dosage or reporting could have contributed to the nonimprovement of relevant outcomes and may help to partly explain meta-analysis findings.

Inclusion criteria for this review mandated the inclusion of aerobic or anaerobic exercise as part of the IVR intervention. As such, it is surprising that common exercise-related outcome measures were not regularly assessed (eg, muscular fitness and cardiorespiratory fitness). A 2024 systematic review identified that IVR improved physical activity enjoyment, intrinsic motivation, and exercise intention compared with traditional and nonimmersive physical activity interventions [20]. Muscular and cardiorespiratory fitness outcomes were not often assessed across trials. Given the clinical importance of these outcomes for overall health [79,80], future research should prioritize their collection. A 2022 RCT in apparently healthy adults used a resistance training-based IVR intervention compared with a self-directed exercise approach [24]. The trial found that IVR exercise improved pertinent cardiometabolic and physical fitness measures (eg, muscular strength, systolic blood pressure) significantly more than control, possibly due to gamification elements improving exercise adherence [24]. Future research should assess whether resistance training-based IVR has a similar effect on cardiometabolic outcomes for clinical populations. Given the broad population inclusion criteria in the current review, there was a lack of literature from certain chronic condition populations where exercise interventions may be used for disease control (eg, type 2 diabetes) [81]. In addition to the populations identified in this review, future work should attempt to assess clinical effectiveness outcomes in these groups.

Feasibility of IVR Exercise Interventions

Determining the feasibility of IVR exercise for service delivery is important for widespread implementation. Safety concerns for IVR exercise may be compounded by technology-related factors (eg, immersiveness decreasing spatial awareness) contributing to potential injury risk [82]. However, regular physical activity and exercise are often critical for condition management [83-85], and IVR presents a unique opportunity to stimulate exercise enjoyment [20]. Previous literature suggests that IVR is safe for physical rehabilitation [86-88], but exercise-specific literature is lacking. While IVR may present an opportunity to deliver exercise in a novel way, individual risk analysis is imperative to address safety concerns. This is particularly true for population groups susceptible to IVR-specific side effects, such as cognitive overload and motion sickness [89]. While only just over half of the trials reported safety metrics (14/26, 56%), no trials in the current review reported an increased exercise-related adverse event rate in IVR groups for any population.

XSL•FO
RenderX

A common side effect reported from IVR interventions is motion sickness, commonly derived from eye strain, disorientation, or nausea [90]. Recent literature reports high variability concerning the incidence of motion sickness, with between 25% and 60% of participants experiencing symptoms [91]. Occurrence of motion sickness in IVR is multifaceted and linked to hardware type (eg, display type and mode, time delay), IVR content (eg, duration, controllability, realism), and individual susceptibility [90,91]. This review identified minimal-slight severity symptoms, indicating that motion sickness was not common or severe in controlled environments. However, findings regarding motion sickness were not commonly reported among trials (8/25, 32%). This is surprising given the high incidence of motion sickness symptoms in IVR and the unclear additional effects of exercise. Therefore, for clinicians using IVR exercise, we recommend an individualized approach to combat potential motion sickness symptoms. This includes (1) selecting IVR hardware and content based on patient feedback [90,91], (2) accounting for motion sickness history and IVR-readiness [90,91], and (3) using IVR interventions based on safety guidelines for individual systems [92].

Perceived technical issues have been cited as a barrier to the implementation of IVR among health care professionals [93]. This includes technical malfunctioning of IVR equipment, unstable Internet connectivity, and system usability [93]. Technical issues were poorly reported in the current review, with only 2 trials describing experiences with technical challenges. Given that a perceived lack of experience/confidence and fear of technical issues disrupting clinical sessions may dissuade clinicians from using IVR [93], future trials should prioritize their reporting and troubleshooting techniques. While trials in the current review suggest that attendance to sessions and adherence to exercise prescription were high, they were not reported commonly and lacked standardized presentation. The inability to verify whether participants attended and completed prescribed sessions makes it difficult to determine whether the IVR exercise was associated with the observed findings. This notion is supported by a 2024 systematic review of physiotherapist-led telerehabilitation interventions, which identified a lack of standardized reporting for attendance and adherence [94]. Since these are often linked with clinical effectiveness, more nuanced and comprehensive investigations into attendance and adherence are warranted [94].

Generally positive participant experiences were observed for participants using the IVR exercise in the included trials. Self-report questionnaires identified high levels of enjoyment and motivation, system usability, and overall acceptability among participants. Qualitative descriptions of experience found that perceived personal benefits and novelty of IVR may be the driving factors for positive experiences. These themes are consistent throughout the literature, with high degrees of immersion [32], distraction [32,95], and incorporation of gamification (eg, leaderboards) [32,96] often cited as factors influencing positive experiences with IVR. Sources of negative experiences related to the inconvenience of wearing an HMD for exercise. A 2022 systematic review observed that older adults reported several problems with wearing an HMD [97]. These included that the HMD was too heavy, caused general discomfort, and resulted in negative affect [97]. Many participants were willing to tolerate these discomforts to experience positive IVR features (such as immersive networking) [97]. These findings, combined with the current review, may suggest that the selection of IVR exercise should be individualized and based on participant preference.

Future Directions

Due to the advancement of IVR in recent years, the potential for integration into exercise practice is immense. This review identifies that future trials should investigate the clinical effectiveness of IVR exercise with more rigorous methodology. Future studies should focus on the measurement and selection of the reported results being derived from pre-established protocols to minimize bias, measuring common exercise-related outcome measures, sample size calculations to acquire statistical power, and intention-to-treat analysis to increase the validity of results. Future studies should emphasize the reporting of session attendance and exercise adherence, technical issues, and participant experiences to assess overall feasibility.

This systematic review has several important limitations to consider. Cumulatively, there was a considerable amount of bias among included studies resulting from the measurement and selection of reported clinical effectiveness results. In turn, “low” to “very low” certainty in the findings was identified through GRADE analysis. Both comparative (IVR vs alternate intervention) and additive (IVR + usual care vs usual care) intervention designs were included for exercising comparators, limiting the specificity and interpretability of findings. While SMD allowed for the comparison of multiple outcome measures across different scales, variability in SDs between populations may limit generalizability. This may have resulted in confounding and potentially distorted the observed effects in meta-analyses [98,99]. Additionally, small-sample bias may have been present in meta-analyses [73]. High statistical heterogeneity was observed in most meta-analyses, and sensitivity analyses revealed statistically significant differences in multiple meta-analytical models through the removal of single trials. These could not be accounted for using meta-regression due to the small number of trials included for each outcome and indicate the fragility of meta-analytical models. Additionally, a confidence distribution approach for the calculation of prediction intervals in meta-analyses with small trial numbers [100] could not be implemented through CMA software. Therefore, the reported prediction intervals may underestimate the uncertainty of the generated effects. A large amount of variety in population groups and participants is present in this review, and subgroup analyses by condition domain were not possible due to sparse outcome data. Therefore, overall findings should be observed with caution when applied to specific populations.

Conclusions

The findings of this systematic review incorporating meta-analyses provide initial evidence for the clinical effectiveness of IVR exercise interventions. The initial evidence may suggest that IVR exercise interventions do not seem to statistically differ from nonexercising or exercising comparators for changes in clinical effectiveness outcomes. However,

XSL•FO
RenderX

significant heterogeneity, high risk of potential bias, and low certainty ratings decrease confidence in the observed findings. While the results indicate that IVR may be a viable option for the delivery of exercise, future trials with more robust methodology for the monitoring and reporting of clinical effectiveness outcomes are needed to verify findings. Importantly, future trials should also emphasize reporting of session attendance and exercise adherence, safety metrics, technical issues, and participant experiences. Implementing these measures will limit the risk of bias observed and provide more accurate insights into the effectiveness and real-world applicability of IVR exercise interventions.

Acknowledgments

We wish to acknowledge and thank Dr Camilla Williams and Ms Jenny Zheng for their contributions to this review.

Funding

No external sources of funding were used to assist in the preparation of this article. Funders had no role in the review process.

Data Availability

The datasets from the current review are available from the corresponding author upon reasonable request.

Authors' Contributions

All authors contributed to the review conception, design, and generation of the search strategy. RCCB conducted the systematic search. RCCB and MHR facilitated study screening, quality assessment, and data extraction. RCCB conducted the data analysis. All authors contributed to data interpretation. RCCB was the principal writer of the manuscript. All authors contributed to manuscript writing and revisions and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Online supplementary material.
[PDF File, 657 KB - jmir_v28i1e87542_app1.pdf]

Checklist 1
PRISMA checklist.
[PDF File, 380 KB - jmir_v28i1e87542_app2.pdf]

References

1. Katzmarzyk PT, Friedenreich C, Shiroma EJ, Lee IM. Physical inactivity and non-communicable disease burden in low-income, middle-income and high-income countries. Br J Sports Med 2022 Jan;56(2):101-106. [doi: 10.1136/bjsports-2020-103640] [Medline: 33782046]
2. Strain T, Flaxman S, Guthold R, et al. National, regional, and global trends in insufficient physical activity among adults from 2000 to 2022: a pooled analysis of 507 population-based surveys with 5·7 million participants. Lancet Glob Health 2024 Aug;12(8):e1232-e1243. [doi: 10.1016/S2214-109X(24)00150-5] [Medline: 38942042]
3. Global levels of physical inactivity in adults: off track for 2030. World Health Organization. 2024. URL: https://www.who.int/publications/i/item/9789240096905 [accessed 2025-05-01]
4. Garcia-Hermoso A, López-Gil JF, Ramírez-Vélez R, Alonso-Martínez AM, Izquierdo M, Ezzatvar Y. Adherence to aerobic and muscle-strengthening activities guidelines: a systematic review and meta-analysis of 3.3 million participants across 32 countries. Br J Sports Med 2023 Feb;57(4):225-229. [doi: 10.1136/bjsports-2022-106189] [Medline: 36418149]
5. Pedersen BK, Saltin B. Exercise as medicine - evidence for prescribing exercise as therapy in 26 different chronic diseases. Scand J Med Sci Sports 2015 Dec;25 Suppl 3(1-72):1-72. [doi: 10.1111/sms.12581] [Medline: 26606383]
6. Anderson L, Taylor RS. Cardiac rehabilitation for people with heart disease: an overview of Cochrane systematic reviews. Cochrane Database Syst Rev 2014 Dec 12;2014(12):CD011273. [doi: 10.1002/14651858.CD011273.pub2] [Medline: 25503364]
7. Zangger G, Bricca A, Liaghat B, et al. Benefits and harms of digital health interventions promoting physical activity in people with chronic conditions: systematic review and meta-analysis. J Med Internet Res 2023 Jul 6;25:e46439. [doi: 10.2196/46439] [Medline: 37410534]
8. Abbas JR, O'Connor A, Ganapathy E, et al. What is virtual reality? a healthcare-focused systematic review of definitions. Health Policy Technol 2023 Jun;12(2):100741. [doi: 10.1016/j.hlpt.2023.100741]

XSL•FO
RenderX

9. Calderone A, Carta D, Cardile D, et al. Use of virtual reality in patients with acquired brain injury: a systematic review. J Clin Med 2023 Dec 14;12(24). [doi: 10.3390/jcm12247680] [Medline: 38137752]
10. Elphinston RA, Vaezipour A, Fowler JA, Russell TG, Sterling M. Psychological therapy using virtual reality for treatment of driving phobia: a systematic review. Disabil Rehabil 2023 May;45(10):1582-1594. [doi: 10.1080/09638288.2022.2069293] [Medline: 35532316]
11. Ma J, Zhao D, Xu N, Yang J. The effectiveness of immersive virtual reality (VR) based mindfulness training on improvement mental-health in adults: a narrative systematic review. Explore (NY) 2023 May;19(3):310-318. [doi: 10.1016/j.explore.2022.08.001]
12. Snoswell AJ, Snoswell CL. Immersive virtual reality in health care: systematic review of technology and disease states. JMIR Biomed Eng 2019;4(1):e15025. [doi: 10.2196/15025]
13. Tortora C, Di Crosta A, La Malva P, et al. Virtual reality and cognitive rehabilitation for older adults with mild cognitive impairment: a systematic review. Ageing Res Rev 2024 Jan;93(102146):102146. [doi: 10.1016/j.arr.2023.102146] [Medline: 38036103]
14. Vaezipour A, Aldridge D, Koenig S, et al. Rehabilitation supported by immersive virtual reality for adults with communication disorders: semistructured interviews and usability survey study. JMIR Rehabil Assist Technol 2023 Oct 31;10:e46959. [doi: 10.2196/46959] [Medline: 37906228]
15. Patsaki I, Dimitriadi N, Despoti A, et al. The effectiveness of immersive virtual reality in physical recovery of stroke patients: a systematic review. Front Syst Neurosci 2022;16:880447. [doi: 10.3389/fnsys.2022.880447] [Medline: 36211591]
16. Kiper P, Godart N, Cavalier M, et al. Effects of immersive virtual reality on upper-extremity stroke rehabilitation: a systematic review with meta-analysis. J Clin Med 2023 Dec 27;13(1):146. [doi: 10.3390/jcm13010146] [Medline: 38202153]
17. Demeco A, Zola L, Frizziero A, et al. Immersive virtual reality in post-stroke rehabilitation: a systematic review. Sensors (Basel) 2023 Feb 3;23(3):1712. [doi: 10.3390/s23031712] [Medline: 36772757]
18. Lo HHM, Zhu M, Zou Z, et al. Immersive and nonimmersive virtual reality-assisted active training in chronic musculoskeletal pain: systematic review and meta-analysis. J Med Internet Res 2024 Aug 19;26:e48787. [doi: 10.2196/48787] [Medline: 39159449]
19. Goudman L, Jansen J, Billot M, et al. Virtual reality applications in chronic pain management: systematic review and meta-analysis. JMIR Serious Games 2022 May 10;10(2):e34402. [doi: 10.2196/34402] [Medline: 35536641]
20. Mocco A, Valmaggia L, Bernardi L, Alfieri M, Tarricone I. Enhancing physical activity with immersive virtual reality: a systematic review. Cyberpsychol Behav Soc Netw 2024 May;27(5):303-317. [doi: 10.1089/cyber.2023.0394] [Medline: 38506442]
21. Ng YL, Ma F, Ho FK, Ip P, Fu KW. Effectiveness of virtual and augmented reality-enhanced exercise on physical activity, psychological outcomes, and physical performance: a systematic review and meta-analysis of randomized controlled trials. Comput Human Behav 2019 Oct;99:278-291. [doi: 10.1016/j.chb.2019.05.026]
22. García-González J, Verdejo-Herrero A, Romero-Del Rey R, et al. Perceptions of immersive virtual reality for physical activity among individuals with hypertension at risk of cardiovascular disease: a qualitative study. Healthcare (Basel) 2025 Apr 9;13(8):858. [doi: 10.3390/healthcare13080858] [Medline: 40281807]
23. Giakoni-Ramírez F, Godoy-Cumillaf A, Espoz-Lazo S, Duclos-Bastias D, Del Val Martín P. Physical activity in immersive virtual reality: a scoping review. Healthcare (Basel) 2023 May 25;11(11):1553. [doi: 10.3390/healthcare11111553] [Medline: 37297694]
24. Mologne MS, Hu J, Carrillo E, et al. The efficacy of an immersive virtual reality exergame incorporating an adaptive cable resistance system on fitness and cardiometabolic measures: a 12-week randomized controlled trial. Int J Environ Res Public Health 2022 Dec 23;20(1):210. [doi: 10.3390/ijerph20010210] [Medline: 36612530]
25. Wender CL. Immersive virtual reality to relieve exercise-induced pain caused by aerobic cycling. Pain Manag 2022 Jul;12(5):665-674. [doi: 10.2217/pmt-2021-0095] [Medline: 35174723]
26. Bond S, Laddu DR, Ozemek C, Lavie CJ, Arena R. Exergaming and virtual reality for health: implications for cardiac rehabilitation. Curr Probl Cardiol 2021 Mar;46(3):100472. [doi: 10.1016/j.cpcardiol.2019.100472] [Medline: 31606141]
27. Su Z, Zhang L, Lian X, Guan M. Virtual reality-based exercise rehabilitation in cancer-related dysfunctions: scoping review. J Med Internet Res 2024 Feb 26;26:e49312. [doi: 10.2196/49312] [Medline: 38407951]
28. Campo-Prieto P, Cancela-Carral JM, Rodríguez-Fuentes G. Feasibility and effects of an immersive virtual reality exergame program on physical functions in institutionalized older adults: a randomized clinical trial. Sensors (Basel) 2022 Sep 6;22(18):6742. [doi: 10.3390/s22186742]
29. Lo HHM, Ng M, Fong PYH, et al. Examining the feasibility, acceptability, and preliminary efficacy of an immersive virtual reality-assisted lower limb strength training for knee osteoarthritis: mixed methods pilot randomized controlled trial. JMIR Serious Games 2024 Sep 27;12:e52563. [doi: 10.2196/52563] [Medline: 39331525]
30. Touloudi E, Hassandra M, Galanis E, Goudas M, Theodorakis Y. Applicability of an immersive virtual reality exercise training system for office workers during working hours. Sports (Basel) 2022 Jun 29;10(7):104. [doi: 10.3390/sports10070104] [Medline: 35878115]
31. Harrington R. Impact of immersive virtual reality exercise on physiological and psychological outcomes in college students: a comparison with traditional cardiorespiratory exercise. JEN 2025;8(1). [doi: 10.53520/jen2025.103201]

XSL•FO
RenderX

32. Mouatt B, Smith AE, Mellow ML, Parfitt G, Smith RT, Stanton TR. The use of virtual reality to influence motivation, affect, enjoyment, and engagement during exercise: a scoping review. Front Virtual Real 2020;1:564664. [doi: 10.3389/frvir.2020.564664]
33. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
34. American College of Sports Medicine. ACSM's Guidelines for Exercise Testing and Prescription: Lippincott williams & wilkins; 2013.
35. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
36. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898]
37. Balshem H, Helfand M, Schünemann HJ, et al. GRADE guidelines: 3. rating the quality of evidence. J Clin Epidemiol 2011 Apr;64(4):401-406. [doi: 10.1016/j.jclinepi.2010.07.015]
38. Schünemann HJ, Higgins JP, Vist GE. Completing 'summary of findings' tables and grading the certainty of the evidence. In: Cochrane Handbook for Systematic Reviews of Interventions: John Wiley & Sons; 2019:375-402 URL: https://onlinelibrary.wiley.com/doi/book/10.1002/9781119536604 [doi: 10.1002/9781119536604]
39. McMaster University and Evidence Prime. GRADEpro. 2025. URL: https://www.gradepro.org/ [accessed 2026-04-11]
40. Borenstein M, Hedges L, Higgins J, Rothstein H. Comprehensive meta-analysis: a computer program for research synthesis. : Biostat; 2013.
41. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
42. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. BMJ 2011 Feb 10;342:d549. [doi: 10.1136/bmj.d549] [Medline: 21310794]
43. Borenstein M, Hedges LV, Higgins JPT. Introduction to Meta‐Analysis: John wiley & sons; 2021. URL: https://onlinelibrary.wiley.com/unavailable-obooks [doi: 10.1002/9781119558378]
44. Higgins JP, Li T, Deeks JJ. Chapter 6: choosing effect measures and computing estimates of effect. In: Cochrane Handbook for Systematic Reviews of Interventions: John Wiley & Sons; 2024.
45. Andrade C. Mean difference, standardized mean difference (SMD), and their use in meta-analysis. J Clin Psychiatry 2020;81(5):11349. [doi: 10.4088/JCP.20f13681]
46. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023 Dec;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
47. Partlett C, Riley RD. Random effects meta-analysis: coverage performance of 95% confidence and prediction intervals following REML estimation. Stat Med 2017 Jan 30;36(2):301-317. [doi: 10.1002/sim.7140] [Medline: 27714841]
48. Sterne JAC, Sutton AJ, Ioannidis JPA, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011 Jul 22;343(jul22 1):d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
49. Peláez-Vélez FJ, Eckert M, Gacto-Sánchez M, Martínez-Carrasco Á. Use of virtual reality and videogames in the physiotherapy treatment of stroke patients: a pilot randomized controlled trial. Int J Environ Res Public Health 2023 Mar 8;20(6):4747. [doi: 10.3390/ijerph20064747] [Medline: 36981652]
50. Rodríguez-Fuentes G, Campo-Prieto P, Cancela-Carral JM. Immersive virtual reality high-intensity aerobic training to slow Parkinson's disease: the ReViPark program. Appl Sci (Basel) 2024;14(11):4708. [doi: 10.3390/app14114708]
51. Rodríguez-Fuentes G, Ferreiro-Gómez E, Campo-Prieto P, Cancela-Carral JM. Exergames and immersive virtual reality as a novel therapy approach in multiple sclerosis: randomised feasibility study. J Clin Med 2024 Sep 30;13(19):5845. [doi: 10.3390/jcm13195845] [Medline: 39407905]
52. Kwan RYC, Liu JYW, Fong KNK, et al. Feasibility and effects of virtual reality motor-cognitive training in community-dwelling older people with cognitive frailty: pilot randomized controlled trial. JMIR Serious Games 2021 Aug 6;9(3):e28400. [doi: 10.2196/28400] [Medline: 34383662]
53. Wang Y, Yang SL, Luo SX, Tong H, Fang Q, Guo YZ. Effects of virtual reality in phase I cardiac rehabilitation training for elderly coronary heart disease patients after percutaneous coronary intervention. Sheng Li Xue Bao 2023 Dec 25;75(6):953-961. [doi: 10.13294/j.aps.2023.0077] [Medline: 38151357]
54. Schrempf MC, Zanker J, Arndt TT, et al. Immersive virtual reality fitness games to improve recovery after colorectal surgery: a randomized single blind controlled pilot trial. Games Health J 2023 Dec;12(6):450-458. [doi: 10.1089/g4h.2023.0004] [Medline: 37428543]
55. Stamm O, Dahms R, Reithinger N, Ruß A, Müller-Werdan U. Virtual reality exergame for supplementing multimodal pain therapy in older adults with chronic back pain: a randomized controlled pilot study. Virtual Real 2022;26(4):1291-1305. [doi: 10.1007/s10055-022-00629-3] [Medline: 35194374]
56. Vorwerg-Gall S, Stamm O, Perotti L, Müller-Werdan U. Efficacy of a 6-week VR exergame for older adults with essential hypertension: a randomized controlled pilot trial. Virtual Real 2024;29(1):1. [doi: 10.1007/s10055-024-01073-1]

XSL•FO
RenderX

57. Naqvi WM, Qureshi MI, Nimbulkar G, Umate L. Gamification for distal radius fracture rehabilitation: a randomized controlled pilot study. Cureus 2022 Sep;14(9):e29333. [doi: 10.7759/cureus.29333] [Medline: 36277562]
58. Nishitha K, Anitha A, Thaheera D. Effectiveness of virtual reality-based rehabilitation and high-intensity exercise program for total knee arthroplasty patients: a randomised controlled trial. JCDR 2024;18(11). [doi: 10.7860/JCDR/2024/71122.20263]
59. Vishnuram S, A K, Suganthirababu P, Ramalingam V, Srinivasan V, Alagesan J. Effect of peripheral nerve mobilization and VR-based gait training on gait parameters among patients with chronic ACA stroke – a pilot study. Phys Occup Ther Geriatr 2024 Oct;42(4):470-480. [doi: 10.1080/02703181.2024.2332253]
60. Barsasella D, Liu MF, Malwade S, et al. Effects of virtual reality sessions on the quality of life, happiness, and functional fitness among the older people: a randomized controlled trial from Taiwan. Comput Methods Programs Biomed 2021 Mar;200:105892. [doi: 10.1016/j.cmpb.2020.105892] [Medline: 33280934]
61. Drazich BF, Anokye D, Zhu S, et al. Motivating older adults through immersive virtual exercise (MOTIVE): a randomized pilot study. Geriatr Nurs 2023;54:229-236. [doi: 10.1016/j.gerinurse.2023.09.019] [Medline: 37844539]
62. Gsangaya MR, Htwe O, Selvi Naicker A, et al. Comparison between the effect of immersive virtual reality training versus conventional rehabilitation on limb loading and functional outcomes in patients after anterior cruciate ligament reconstruction: a prospective randomized controlled trial. Asia Pac J Sports Med Arthrosc Rehabil Technol 2023 Oct;34:28-37. [doi: 10.1016/j.asmart.2023.09.002] [Medline: 37841643]
63. Kershner K, Morton D, Robison J, N'dah KW, Fanning J. Assessing the feasibility and acceptability of virtual reality for remote group-mediated physical activity in older adults: pilot randomized controlled trial. JMIR Form Res 2024 Nov 8;8:e53156. [doi: 10.2196/53156] [Medline: 39514256]
64. Lima Rebêlo F, de Souza Silva LF, Doná F, Sales Barreto A, de Souza Siqueira Quintans J. Immersive virtual reality is effective in the rehabilitation of older adults with balance disorders: a randomized clinical trial. Exp Gerontol 2021 Jul 1;149:111308. [doi: 10.1016/j.exger.2021.111308] [Medline: 33744393]
65. Rutkowski S, Bogacz K, Czech O, Rutkowska A, Szczegielniak J. Effectiveness of an inpatient virtual reality-based pulmonary rehabilitation program among COVID-19 patients on symptoms of anxiety, depression and quality of life: preliminary results from a randomized controlled trial. Int J Environ Res Public Health 2022 Dec 17;19(24):16980. [doi: 10.3390/ijerph192416980] [Medline: 36554860]
66. Tuck N, Pollard C, Good C, et al. Active virtual reality for chronic primary pain: mixed methods randomized pilot study. JMIR Form Res 2022 Jul 13;6(7):e38366. [doi: 10.2196/38366] [Medline: 35830224]
67. Yalfani A, Abedi M, Raeisi Z, Asgarpour A. The effects of virtual reality training on postural sway and physical function performance on older women with chronic low back pain: a double-blind randomized clinical trial. J Back Musculoskelet Rehabil 2024;37(3):761-770. [doi: 10.3233/BMR-230260] [Medline: 38217579]
68. Zak M, Sikorski T, Michalska A, et al. The effects of physiotherapy programmes, aided by virtual reality solutions, on balance in older women: a randomised controlled trial. J Clin Med 2024 Oct 28;13(21):6462. [doi: 10.3390/jcm13216462] [Medline: 39518601]
69. Parmak D, Angın E, Iyigun G. Effects of immersive virtual reality on physical function, fall-related outcomes, fatigue, and quality of life in older adults: a randomized controlled trial. Healthcare (Basel) 2025 Jul 24;13(15):1800. [doi: 10.3390/healthcare13151800] [Medline: 40805833]
70. Ramos JS, Alexander J, Dalleck LC, et al. The impact of a workplace virtual reality-delivered exercise program on central hemodynamics in people with developmental disability living in Australia: a pilot randomised waitlist-controlled study. Prev Med Rep 2025 Nov;59:103256. [doi: 10.1016/j.pmedr.2025.103256] [Medline: 41113688]
71. An Y, Park C. The effects of virtual soccer game on balance, gait function, and kick speed in chronic incomplete spinal cord injury: a randomized controlled trial. Spinal Cord 2022 Jun;60(6):504-509. [doi: 10.1038/s41393-021-00745-y] [Medline: 34999726]
72. Seo EY, Kim YS, Lee YJ, Hur MH. Virtual reality exercise program effects on body mass index, depression, exercise fun and exercise immersion in overweight middle-aged women: a randomized controlled trial. Int J Environ Res Public Health 2023 Jan 4;20(2):900. [doi: 10.3390/ijerph20020900] [Medline: 36673656]
73. Lin L. Bias caused by sampling error in meta-analysis with small sample sizes. PLoS One 2018;13(9):e0204056. [doi: 10.1371/journal.pone.0204056] [Medline: 30212588]
74. MacIntyre E, Sigerseth M, Larsen TF, et al. Get your head in the game: a replicated single-case experimental design evaluating the effect of a novel virtual reality intervention in people with chronic low back pain. J Pain 2023 Aug;24(8):1449-1464. [doi: 10.1016/j.jpain.2023.03.013] [Medline: 37030584]
75. Rooney T, Sharpe L, Winiarski N, et al. A synthesis of meta-analyses of immersive virtual reality interventions in pain. Clin Psychol Rev 2025 Apr;117:102566. [doi: 10.1016/j.cpr.2025.102566] [Medline: 40058296]
76. Viderman D, Tapinova K, Dossov M, Seitenov S, Abdildin YG. Virtual reality for pain management: an umbrella review. Front Med (Lausanne) 2023;10:1203670. [doi: 10.3389/fmed.2023.1203670] [Medline: 37521355]
77. Bull FC, Al-Ansari SS, Biddle S, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. Br J Sports Med 2020 Dec;54(24):1451-1462. [doi: 10.1136/bjsports-2020-102955]

XSL•FO
RenderX

78. American College of Sports Medicine. American college of sports medicine position stand. progression models in resistance training for healthy adults. Med Sci Sports Exerc 2009 Mar;41(3):687-708. [doi: 10.1249/MSS.0b013e3181915670] [Medline: 19204579]
79. Al-Mallah MH, Sakr S, Al-Qunaibet A. Cardiorespiratory fitness and cardiovascular disease prevention: an update. Curr Atheroscler Rep 2018 Jan 16;20(1):1. [doi: 10.1007/s11883-018-0711-4] [Medline: 29340805]
80. Lopez-Jaramillo P, Lopez-Lopez JP, Tole MC, Cohen DD. Muscular strength in risk factors for cardiovascular disease and mortality: a narrative review. Anatol J Cardiol 2022 Aug;26(8):598-607. [doi: 10.5152/AnatolJCardiol.2022.1586] [Medline: 35924286]
81. Boulé NG, Haddad E, Kenny GP, Wells GA, Sigal RJ. Effects of exercise on glycemic control and body mass in type 2 diabetes mellitus: a meta-analysis of controlled clinical trials. JAMA 2001 Sep 12;286(10):1218-1227. [doi: 10.1001/jama.286.10.1218] [Medline: 11559268]
82. Proffitt R, Warren J, Lange B, Chang CY. Safety and feasibility of a first-person view, full-body interaction game for telerehabilitation post-stroke. Int J Telerehabil 2018;10(1):29-36. [doi: 10.5195/ijt.2018.6250] [Medline: 30147841]
83. Daste C, Kirren Q, Akoum J, Lefèvre-Colau MM, Rannou F, Nguyen C. Physical activity for osteoarthritis: efficiency and review of recommandations. Joint Bone Spine 2021 Dec;88(6):105207. [doi: 10.1016/j.jbspin.2021.105207] [Medline: 33962031]
84. Jadczak AD, Makwana N, Luscombe-Marsh N, Visvanathan R, Schultz TJ. Effectiveness of exercise interventions on physical function in community-dwelling frail older people: an umbrella review of systematic reviews. JBI Database System Rev Implement Rep 2018 Mar;16(3):752-775. [doi: 10.11124/JBISRIR-2017-003551] [Medline: 29521871]
85. Mcleod JC, Stokes T, Phillips SM. Resistance exercise training as a primary countermeasure to age-related chronic disease. Front Physiol 2019;10(645):645. [doi: 10.3389/fphys.2019.00645] [Medline: 31244666]
86. Chen J, Or CK, Li Z, Yeung EHK, Zhou Y, Hao T. Effectiveness, safety and patients' perceptions of an immersive virtual reality-based exercise system for poststroke upper limb motor rehabilitation: a proof-of-concept and feasibility randomized controlled trial. Digit Health 2023;9:20552076231203599. [doi: 10.1177/20552076231203599] [Medline: 37766909]
87. Saldana D, Neureither M, Schmiesing A, et al. Applications of head-mounted displays for virtual reality in adult physical rehabilitation: a scoping review. Am J Occup Ther 2020 Sep 1;74(5):7405205060p1-7405205060p15. [doi: 10.5014/ajot.2020.041442]
88. Soltani P, Andrade R. The influence of virtual reality head-mounted displays on balance outcomes and training paradigms: a systematic review. Front Sports Act Living 2020;2:531535. [doi: 10.3389/fspor.2020.531535] [Medline: 33634259]
89. Msweli NT, Phahlane M. Health-related issues of immersive technologies: a systematic literature review. In: Msweli NT, Phahlane M, editors. Informatics (MDPI) 2025;12(2):47. [doi: 10.3390/informatics12020047]
90. Chang E, Kim HT, Yoo B. Virtual reality sickness: a review of causes and measurements. Int J Hum Comput Interact 2020 Oct 20;36(17):1658-1682. [doi: 10.1080/10447318.2020.1778351]
91. Rebenitsch L, Owen C. Estimating cybersickness from virtual reality applications. Virtual Real 2021 Mar;25(1):165-174. [doi: 10.1007/s10055-020-00446-6]
92. Meta quest 2 english - EU/UK health and safety warnings. Meta. 2024. URL: https://www.meta.com/au/legal/quest/health-and-safety-warnings [accessed 2025-07-01]
93. Kouijzer M, Kip H, Bouman YHA, Kelders SM. Implementation of virtual reality in healthcare: a scoping review on the implementation process of virtual reality in various healthcare settings. Implement Sci Commun 2023 Jun 16;4(1):67. [doi: 10.1186/s43058-023-00442-2] [Medline: 37328858]
94. Simmich J, Ross MH, Russell T. Real-time video telerehabilitation shows comparable satisfaction and similar or better attendance and adherence compared with in-person physiotherapy: a systematic review. J Physiother 2024 Jul;70(3):181-192. [doi: 10.1016/j.jphys.2024.06.001] [Medline: 38879432]
95. Stewart TH, Villaneuva K, Hahn A, et al. Actual vs. perceived exertion during active virtual reality game exercise. Front Rehabil Sci 2022;3:887740. [doi: 10.3389/fresc.2022.887740] [Medline: 36189005]
96. Barbour B, Sefton L, Bruce RM, Valmaggia L, Runswick OR. Acute psychological and physiological benefits of exercising with virtual reality. PLoS One 2024;19(12):e0314331. [doi: 10.1371/journal.pone.0314331] [Medline: 39693283]
97. Healy D, Flynn A, Conlan O, McSharry J, Walsh J. Older adults' experiences and perceptions of immersive virtual reality: systematic review and thematic synthesis. JMIR Serious Games 2022 Dec 6;10(4):e35802. [doi: 10.2196/35802] [Medline: 36472894]
98. Greenland S, Maclure M, Schlesselman JJ, Poole C, Morgenstern H. Standardized regression coefficients: a further critique and review of some alternatives. Epidemiology (Sunnyvale) 1991 Sep;2(5):387-392. [Medline: 1742392]
99. Greenland S, Schlesselman JJ, Criqui MH. The fallacy of employing standardized regression coefficients and correlations as measures of effect. Am J Epidemiol 1986 Feb;123(2):203-208. [doi: 10.1093/oxfordjournals.aje.a114229] [Medline: 3946370]
100. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019 Jun;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]

XSL•FO
RenderX

Abbreviations

GRADE: Grading of Recommendations, Assessment, Development and Evaluation
HMD: head-mounted display
IVR: immersive virtual reality
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search
RCT: randomized controlled trial
RoB2: revised Cochrane Risk of Bias 2 tool
SMD: standardized mean difference

Edited by S Brini; submitted 11.Nov.2025; peer-reviewed by B Lai, K Aoki; revised version received 28.Jan.2026; accepted 19.Feb.2026; published 20.Apr.2026.

Please cite as:
Brown RCC, Ross MH, Russell TG
Clinical Effectiveness of Immersive Virtual Reality Exercise Interventions: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e87542
URL: https://www.jmir.org/2026/1/e87542
doi:10.2196/87542

© Riley C C Brown, Megan H Ross, Trevor G Russell. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Recognizing the Importance of Design, Content, and Delivery Features of Health Animations for Preventive Health Behaviors: Realist Review

Kathleen McCorry[1], MSc; Gregory Maniatopoulos[2], PhD; Linda Errington[3], MSc; Ellie Land[4], MA; Michael Craig[5], PhD; Santosh Vijaykumar[5], PhD; Till Bärnighausen[1,6,7], MD; Nicola O'Brien[5], PhD

[1]Heidelberg Institute of Global Health, Faculty of Medicine and University Hospital, Heidelberg University, Heidelberg, Germany

[2]School of Business, University of Leicester, Leicester, United Kingdom

[3]Walton Library, Newcastle University, Newcastle upon Tyne, United Kingdom

[4]School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, United Kingdom

[5]School of Psychology, Faculty of Health and Wellbeing, Northumbria University, Northumberland Building, Newcastle upon Tyne, United Kingdom

[6]Department of Global Health and Population, Harvard TH Chan School of Public Health, Harvard University, Boston, MA, United States

[7]Africa Health Research Institute, Somkhele and Durban, South Africa

Corresponding Author:
Nicola O'Brien, PhD
School of Psychology, Faculty of Health and Wellbeing, Northumbria University, Northumberland Building, Newcastle upon Tyne, United Kingdom

Abstract

Background: Animated messages designed to promote preventive health behaviors (health animations) are a prevalent form of digital health communication globally and are used across a variety of health behaviors. Health animations are visual and can be less reliant on language, helping to reduce health literacy barriers and inequities. Being easily and inexpensively shared, they are a potentially powerful tool for disease prevention. Evidence suggests that animations can be effective in health and non–health care settings. Despite a plethora of health animations in existence and their potential reach and scale, evidence underpinning their design and application, including the potential causal mechanisms at work, is often limited or unknown.

Objective: This realist review aimed to understand why, how, for whom, to what extent, and when health animations designed to promote preventive health behaviors work.

Methods: The review was conducted in accordance with the Realist and Meta-narrative Evidence Syntheses: Evolving Standards (RAMESES). Peer-reviewed publications identified through database searching from inception until April 27, 2025, as well as gray literature, were considered for inclusion. Animations designed to promote preventive health behaviors in any population and evaluations using any design were included. Animations that could not be viewed, were designed to treat illness or disease, or were part of multicomponent interventions were excluded. Data were appraised for their relevance, rigor, and richness. Data syntheses sought to produce context-mechanism-outcome configurations, contributing to a program theory of health animations. International stakeholder workshops with professionals and members of the public were used to sense check findings and refine the program theory.

Results: This review synthesized data from evaluations of 48 health animations. Within the data, design, content, and delivery constructs were identified, including audience or challenge representation, using storytelling, evoking emotion, accessibility, and other contexts. These contexts enabled the triggering of key mechanisms, such as identification, transportation into a story, attention, building self-efficacy, and cognitive processes. The evidence available for synthesis in the building of our program theory of health animations was limited by a lack of data on behavioral outcomes, meaning that the theory is largely derived from evidence of the contexts and mechanisms influencing key determinants known to affect behavior. These key determinants include behavioral intentions, skills and attitudes, and knowledge as a key factor in raising awareness of the need to change behavior.

Conclusions: This realist review advances the understanding of the impact of health animations designed to promote preventive health behaviors by providing insight into the design, content, and delivery features at work. Our program theory describes the specific contexts and mechanisms that influence evidence-based determinants of behavior and behavior change. These contexts and mechanisms, therefore, should be considered during health animation design and development processes. A set of 10 recommendations is provided to this end.

Trial Registration: Review registration: PROSPERO CRD42023447127; https://www.crd.york.ac.uk/PROSPERO/view/CRD42023447127

XSL•FO
RenderX

(J Med Internet Res 2026;28:e79769) doi:10.2196/79769

KEYWORDS

animation; health behavior; health communication; preventive medicine; realist review

Introduction

Background

Effective public health communication plays a crucial role in empowering individuals with information and skills to have more control over the decisions and behaviors that impact their health. Modifiable behavioral risk factors, such as sedentary behavior, unhealthy diets, not attending health screenings, and high-risk sexual behaviors, contribute to between a quarter and a half of the global burden on health and social care, as well as deaths [1-3]. Health communication science can critically inform the strategies required to ensure that accurate and accessible information and support are provided to individuals, communities, and populations to promote health and positive health behaviors, and reduce subsequent morbidity and mortality [4].

Effective health communication strategies need to address health literacy, defined as the knowledge and competencies needed to access, understand, and use health information [5]. Including visual stimuli within health communication is encouraged, particularly when trying to reach individuals disadvantaged by lower literacy and education levels; knowledge, understanding, attention, and recall are improved when visual images are used in health communication compared with using text alone [6]. This is supported by the picture superiority effect from cognitive psychology, which shows that visual information is more memorable than textual information. This suggests that visual information is more effectively encoded into memory, as it activates both verbal and visual memory systems, strengthening memory formation and subsequent recall [7].

Health animations, defined as simulated motion pictures depicting the movement of drawn (or simulated) objects [8] related to health, are a highly prevalent approach to visual health communication. Due to their digital nature, animations can be easily and inexpensively shared via the internet, apps, and social media, catalyzing their widespread global application as a digital health solution [9] to deliver public health information, promote health behavior change, and prevent ill health. Compared with paper-based or in-person health communication, animations have greater reach and scale, overcoming geographical barriers and exploiting the rapid rise and evolution of social media platforms (eg, TikTok, WhatsApp, and Instagram) and online mediums (ie, local and international governmental, nongovernmental, voluntary sector, charity organizations, or websites).

Emerging reviews and meta-analytical evidence of animations suggest that there are benefits to learning and cognitive load from animations as educational tools for patients [10,11], health care professionals [12], and in non–health care settings [13]. However, evidence of health animations in the promotion of public health preventive behaviors, such as lifestyle, hygiene, and preventive health appointment uptake—which can prevent and control communicable and noncommunicable diseases—remains unclear. To date, evidence relies on individual empirical studies [14-16] investigating the impact of health animations on behaviors and social cognitive determinants (eg, recall, knowledge, and intention). These individual studies exhibit substantial heterogeneity in study design, comparators, target health behaviors, target audiences, and contextual complexities in the animation design and content. Therefore, rather than attempting to synthesize evidence for the effectiveness of health animations promoting preventive health behaviors from a pool of heterogeneous studies, we adopt a realist review approach to iteratively identify and synthesize evidence to uncover the specific causal mechanisms being triggered by animations that, in turn, facilitate certain outcomes.

Aim of the Review

The aim of this realist review was to identify and synthesize evidence about health animations to understand how, why, for whom, to what extent, and in which contexts they are expected to produce their effects. We aimed to generate knowledge about the causes of outcomes in particular circumstances in terms of the context-mechanism-outcome configurations (CMOCs) of realist reviews and syntheses. By understanding how and why health animations and the processes underpinning them work (or fail to work), we aimed to translate key learnings into a set of recommendations to inform the design, development, and implementation of future public health animations.

Methods

Overview

This realist review and synthesis was conducted in accordance with the Realist and Meta-Narrative Evidence Syntheses: Evolving Standards (RAMESES) [17-19] and a published protocol [20]. Textbox 1 provides definitions of key realist concepts. The realist review methodology and structure are well-suited to the complexity and heterogeneity of health animations and the methods used to evaluate them. Health animations are used across a variety of health behaviors and therefore can produce multiple varied outcomes. While conventional systematic reviews focus on outcomes, realist reviews focus on the contexts and the mechanisms they trigger to produce those outcomes. Furthermore, realist methodology does not seek to separate and categorize the elements of the evidence being explored but instead finds causation and configurations that lead to a better understanding of how and why effects are produced [21]. This is achieved by identifying contexts and mechanisms and examining how they interact to produce outcomes; once identified, CMOCs can be developed. The following 5 iterative steps were undertaken with some elements within the steps occurring simultaneously.

XSL•FO
RenderX

Textbox 1. Definitions of key realist concepts [19,21].

Context

- A set of factors or conditions influencing how an animation operates or works
- Context triggers mechanisms

Mechanism

- Hidden, underlying, context sensitive, causal force triggered by context
- Mechanisms are behind "why" and "what" makes things happen
- They are transferable
- Can be internal cognitive or psychological processes (ie, developing understanding and the building of confidence)
- Not states of being but processes

Outcome

- Outcomes vary including intentions, behaviors, and attitudes
- Outcomes help make what is functioning as context clearer

Context-mechanism-outcome configuration (CMOC)

- A lens through which to think about how things are working
- Support the knowledge being generated through the synthesis
- Help to tell the story of what is happening

Program theory

- Underpinned by theory (explicitly or not) and reflecting what is understood about causation
- Seeks to explain how a program (ie, an intervention) is expected to work
- May include multiple CMOCs and can be expressed in diagram or narrative format

Demiregularity

- Semipredictable behavioral patterns reflecting individuals' thoughts and actions that are influenced by contextual differences

Step 1: Defining the Review Scope and Identifying Existing Theories

Theories and models across multiple disciplines were explored for insights and understanding of what may influence how health animations are perceived, interpreted, or experienced. These insights were discussed within the review team to develop initial program theories. Exploratory background searches were conducted to identify literature on the use of health animations and their impacts and effects, helping to define the review's scope.

Step 2: Search for Evidence

Initial searches of electronic databases were conducted from database inception until July 10, 2023, with ongoing searching throughout the review process (in accordance with the iterative nature of realist reviews). A final search of electronic databases was conducted on April 27, 2025. The databases searched were MEDLINE, Scopus, CINAHL, EMBASE, PsycINFO, ProQuest Social Science, Web of Science, and the Cochrane Library. The search strategy, combining relevant Medical Subject Headings (MeSH) terms and keywords for concepts of animation, health behavior, and behavior change, is available in Multimedia Appendix 1. Reference lists were checked, and websites of governmental, nongovernmental, and third-sector organizations known to promote public health were searched to identify additional health animations. Google and YouTube were also searched using keywords to identify additional health animations and organizational sources of animations.

Rayyan (Rayyan Systems, Inc) was used to manage the records, including deduplication. Title and abstract screening was conducted by 1 reviewer (KMC), with a random 10% sample screened independently by another reviewer (NO'B). Full-text screening was conducted independently by 2 reviewers (KMC and NO'B). The authors were contacted to provide animations when they were not publicly available. Textbox 2 presents the inclusion and exclusion criteria.

XSL•FO
RenderX

Textbox 2. Inclusion and exclusion criteria of publications

Inclusion criteria

- Participants: any
- Interventions: animations communicating health messages designed to promote preventive health behaviors or social cognitive determinants thereof
- Comparators: other interventions, usual care, no intervention, or no comparator
- Types of study designs: any
- Contexts: any
- Outcomes: any, including, but not limited to: health behavior, behavioral cognitions (attitude, intention, and self-efficacy), knowledge, awareness, emotions, memory, and recall.

Exclusion criteria

- Animations designed to treat illness and disease.
- Multicomponent interventions where animation effects could not be isolated from other components.
- No access to animation

Step 3: Selection and Appraisal of Evidence

Evidence was selected using the criteria outlined in step 2 of the protocol [20]. Peer-reviewed and gray literature publications were included where they could potentially contribute to building the program theory. The relevance, rigor, and richness of the data were used to assess the quality of evidence [22]. Questions asked of the data in relation to relevance included: Does the evidence help develop CMOCs and contribute to their advancement? Is the building and testing of program theory supported by this evidence? Does it speak to the research question? Questions asked in relation to richness included: Is there a sufficient level of description about the development of the intervention and the theories and concepts underpinning it? Was there discussion about factors that impacted the functioning or outcomes of the intervention? Were details provided about how the intervention was meant to work? Was this information general or specific to the intervention? Questions asked in relation to rigor included: What methods were used to generate the data? What are the credibility and trustworthiness of these methods?

Step 4: Data Extraction

Data extraction was conducted by 1 reviewer (KMC) and checked by a second reviewer (NO'B). Bespoke data extraction sheets were developed to record study and animation characteristics. Study characteristics included information about participants, outcomes, findings, health behaviors targeted, and theories influencing animation design. Animation characteristics included length, use of sound, color, language, and text, development information, composition, character profiles, use of cultural indicators, symbols, and metaphors.

Step 5: Data Synthesis and Theory Refinement

Data synthesis and theory refinement were developed through multiple viewings of the animations, discussions within the review team, and consideration of the identified theories and models. The initial synthesis and CMOC development involved KMC, NO'B, and GM reviewing the data to identify patterns (ie, demiregularities) and elements that the animations shared to define the "if" (ie, contexts) and to list the "then" (ie, outcomes). Data were then interrogated to define the "because" (ie, mechanisms), resulting in if-then-because (context-outcome-mechanism) statements [23]. The integrity of the if-then-because statements was questioned as discrete statements by considering the included data and their potential interpretations in relation to the initial program theories identified in step 1. In doing so, an emerging program theory was developed.

Four stakeholder workshops were held to sense check preliminary findings and to discuss and refine the if-then-because statements and the emerging program theory. Two international online workshops were conducted with professionals in April 2024 (n=21), allowing for the testing and refinement of the program theory with individuals who develop, commission, and/or promote animations or who work with individuals or systems that may benefit from health animations. Two in-person workshops were held with members of the public in the United Kingdom in May 2024 and June 2024 (n=18), enabling the testing and refinement of the program theory with public stakeholders targeted by and exposed to health animations. Following these workshops, feedback was used to review the if-then-because statements and translate them into CMOCs. The CMOCs were tested against the data and refined as required. They were then used to develop a set of recommendations to inform the design, development, and implementation of future public health animations. Both the CMOCs and the recommendations were presented to professional and public stakeholders for feedback during a final online workshop in September 2024 (n=17).

Ethical Considerations

Ethical approval for the stakeholder workshops was granted by the Northumbria University Research Ethics Committee (reference #6803). Stakeholders provided informed consent before the workshops, and their privacy and confidentiality were ensured. Public stakeholders were remunerated for their time and contribution with a £25 (US $33.8) shopping voucher.

XSL•FO
RenderX

Results

Overview

The final sample for this review included 48 unique single animations or series of animations; 39 were reported in 47 peer-reviewed publications [14-16,24-67] and the remaining 9 were gray literature [68-76]. Figure 1 displays the flow of the selection of animations. Categories of health behaviors targeted were communicable diseases and hygiene behaviors (n=11), vaccination (n=10), general health promotion (n=8), screening (n=5), substance use (n=5), sexual health (n=4), nutrition (n=3), and oral health (n=2). Study designs included: randomized controlled trials (n=18), randomized studies (n=12), pre-post studies (n=8), mixed methods studies (n=5), and cross-sectional studies (n=4).

Figure 1. Flow diagram of animations included in the realist review, adapted from PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses).

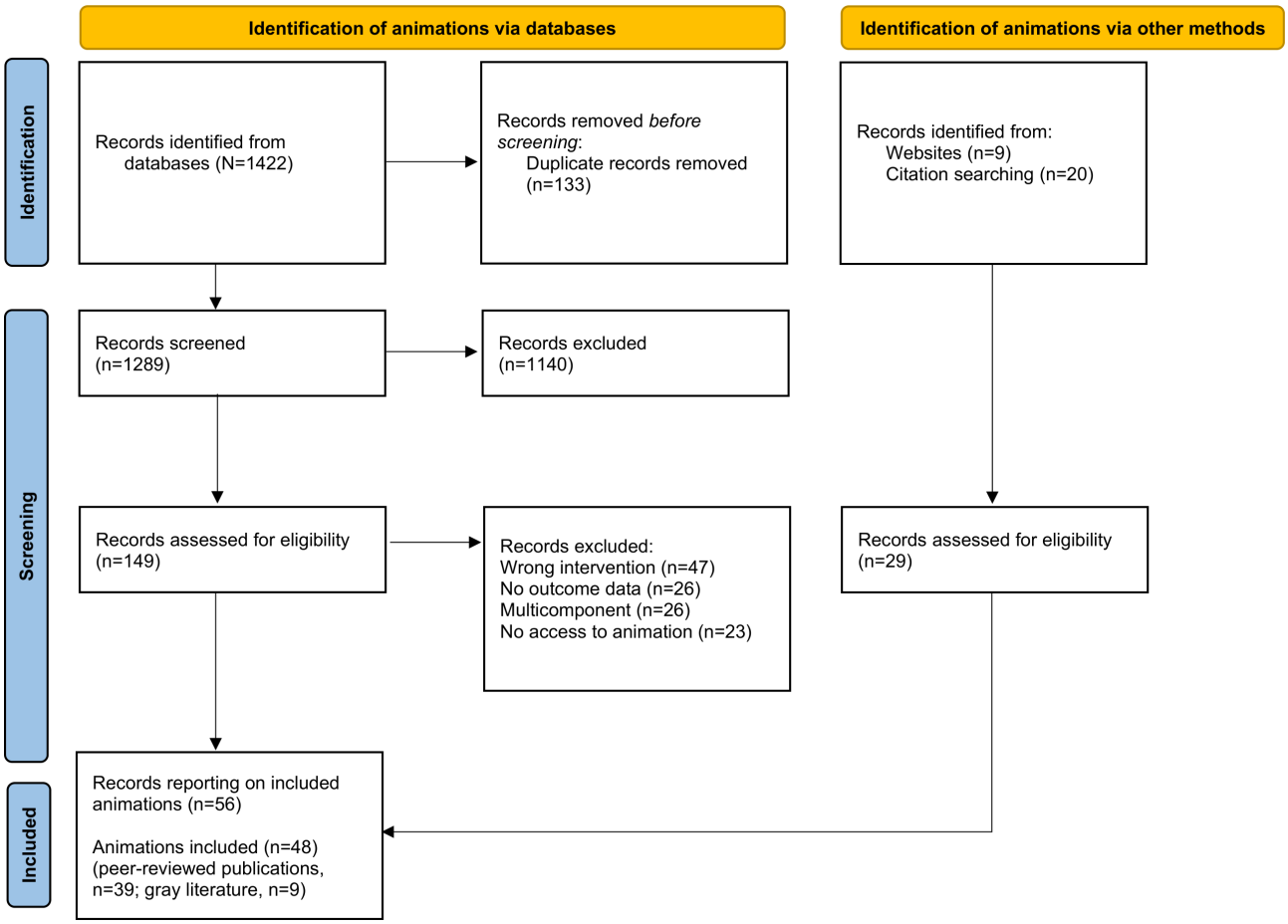

Of the 48 animations included, 28 were single animations and 20 were animation series; for the purposes of data synthesis, animation series were considered as single animations. Outcome measures included knowledge, behavioral intentions, attitudes, skills, acceptability, and viewer engagement; 8 studies also included behavioral outcomes. Many animations reported details of co-design processes, for example, in a study by Willis et al [65], and theories and models (eg, social cognitive theory [47], elaboration likelihood model [24], and theory of planned behavior [50]) that informed animation development, evaluation, and outcome measures, and provided key insight and information in the building of the program theory.

Most animations used storytelling to convey messages (n=22), while others used entertainment elements (n=18), storytelling elements (n=4), or no storytelling/entertainment elements (n=4). Textbox 3 provides an explanation of the different categories of entertainment that were used. Further study design details, including the demographic and geographical target audience of the animations, as well as the animation characteristics, such as the use of sound, language, text, and length, are presented in Multimedia Appendices 2 and 3 [14-16,24-76].

XSL•FO
RenderX

Textbox 3. Explanation of entertainment categories used in animations

Storytelling

- Characters, plot line, beginning/middle/end, and entertainment elements

Storytelling elements

- Testimonials, persona building, informational plus persona/character/situation building

Entertainment elements

- Anthropomorphism, characters in action, and so on

No storytelling or entertainment elements

- Images only function to inform

Realist Analysis

The main characteristics contributing to understanding why, how, for whom, to what extent, and in which contexts health animations are expected to promote preventive health behaviors were considered within 3 high-level constructs: design, content, and delivery. These constructs did not operate in isolation but often worked together to enhance effects. It is within these constructs that the contexts, mechanisms, and outcomes identified in the analysis were situated (Figure 2). Consequently, the following 9 CMOCs were developed and tested with stakeholders (Table 1). The final column of Tables S1 and S2 in Multimedia Appendices 2 and 3, respectively, details which CMOCs were associated with each animation included. Additional detail on the constructs, contexts, and mechanisms is provided in Multimedia Appendix 4.

Figure 2. Diagram depicting high-level constructs of animation design, content, and delivery, as well as the related contexts, mechanisms, and outcomes identified in the analysis.

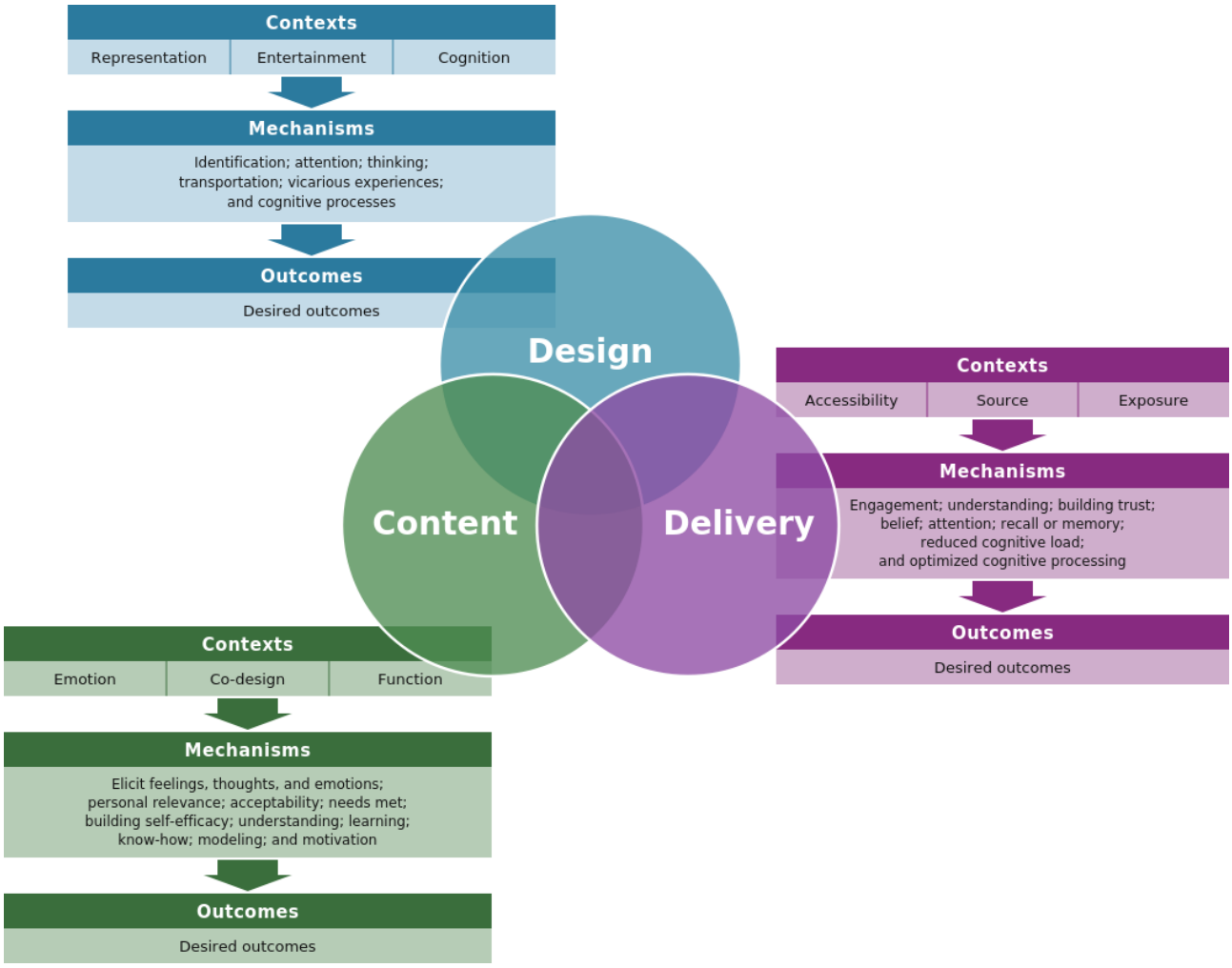

XSL•FO
RenderX

Table . Context-mechanism-outcome configuration (CMOC) descriptors and the number of animations included in the review shaping and relating to each CMOC.

CMOC	Descriptor	Number of animations
Design CMOC 1: representation	If the animation is designed to represent a particular audience (or shared challenge), desired outcomes are more likely to be achieved, because the audience being represented identified with its elements allowing them to pay greater attention and think about what they are watching.	27
Design CMOC 2: entertainment	If the animation uses storytelling, desired outcomes are more likely to be achieved, because the audience will be transported into the story.	44
Design CMOC 3: cognition	If the animation is designed to facilitate efficient cognitive processing, then the desired outcomes are more likely to be achieved because cognitive mechanisms were enhanced.	44
Content CMOC 1: emotion	If the animation presents its content in an evocative manner, then desired outcomes are more likely to be achieved, because the animation content elicited powerful feelings, thoughts, and emotions.	22
Content CMOC 2: co-design	If the content of the animation is co-designed, then desired outcomes are more likely to be achieved, because the animation is more relevant, acceptable, and meets the needs of the target audience.	14
Content CMOC 3: function	If the content of the animation presents the challenge and the solution, then desired outcomes are more likely to be achieved, because the content of the animation increased audience understanding and provided 'know-how' information that helped the audience feel empowered and believe in their capacity to deal with the challenge.	44
Delivery CMOC 1: accessibility	If the animation is delivered in an accessible manner for the target audience, then desired outcomes are more likely to be achieved, because they will be better able to understand and engage with the information.	43
Delivery CMOC 2: source	If the animation is delivered by a source that is perceived to be credible, reliable, and trusted, then desired outcomes are more likely to be achieved, because they believed and engaged with the information and therefore better attended to and recalled it.	33
Delivery CMOC 3: exposure	If the mode, channel, and/or intensity of the delivery of the animation encapsulates the purpose and needs of the target audience and/or behavior, then desired outcomes are more likely to be achieved, because the potential exposure to the animation optimizes cognitive processing and responds to the needs of the audience.	31

Design

The construct *design* refers to the ways in which health animations are constructed to express the ideas and messages they are trying to communicate. How an animation looks, feels, and sounds, as well as its length, use of symbols and metaphors, character profiles, and other features, are all products of the design.

Design CMOC 1: Representation

Twenty-seven animations (reported in 27 publications [16,24,25,32,36,37,40,41,43-45,47,49,54,56-58,62,65,68-74,76]) were identified as shaping and relating to this CMOC. Within the context of designing animations to represent audiences or specific challenges, the mechanism of identification emerged as contributing to outcomes. Once the audience identifies themselves or the challenge depicted in the animation, the

XSL•FO
RenderX

mechanisms of attention and thinking may also be triggered. Professional and public stakeholders felt that this finding resonated with their experiences, represented their understanding, and promoted inclusivity. Professional stakeholders further commented that this CMOC reduces unconscious biases and that identifying oneself is key to engagement, increasing accessibility, particularly for marginalized communities. Some stakeholders expressed that the desired outcomes and target audience should guide how representation is used.

An example of this CMOC in practice can be seen in the animated series "Meena" [73]. This animation was purposely designed to reflect the lives of girls in South Asia, with special attention paid to accurately representing the look and sound of characters, their environments, and their typical experiences.

Design CMOC 2: Entertainment

Forty-four animations (reported in 52 publications [15,16,24-33,35-45,47,49-76]) were identified as shaping and relating to this CMOC. Storytelling and entertainment include the use of characters, plot, storylines, imagery, visuals, and other elements. Additionally, anthropomorphism was used to make biological processes easier to understand and entertaining at the same time, such as representing white blood cells as soldiers [35]. Transportation is a mechanism that refers to one engaging in a way that enables greater attention and creates vicarious experiences for the viewer as they become caught up and entertained by the story. Stakeholders supported this CMOC expressing that storytelling or entertainment make messages easier to remember and more relatable. Professional stakeholders expressed that storytelling/entertainment can be facilitated without reliance on language, increasing accessibility for those with language and/or literacy barriers.

The Story of Cholera [69] is an example of an animation that fully uses storytelling. This animation tells the story of a young boy whose father contracts cholera and becomes very ill. The boy subsequently learns how to prevent cholera and teaches the people in his village hygiene practices to prevent the disease. The storytelling in this animation is supported by the use of color, music, sound effects, and narration. The entertainment context often works in conjunction with other contexts such as emotion. An example of this can be seen in an animation depicting the story of a pregnant mother's decision-making journey as she learns about the flu vaccine and considers the health implications (and experiences the associated emotions) for herself and her baby [57].

Design CMOC 3: Cognition

Forty-four animations (reported in 52 publications [14-16,24-52,54-61,63,64,66-73,75,76]) were identified as shaping and relating to this CMOC. Design features that facilitate efficient cognitive processing, such as the use of familiar characters and structures, cultural representations in terms of metaphors and symbols, composition of images, and other elements, can trigger and enhance cognitive mechanisms. These cognitive mechanisms may include memory or recall, thinking, learning, attention, and cognitive load. Stakeholders agreed with the CMOC while also cautioning against overwhelming audiences with the heavy use of text and multiple simultaneous messages.

An animated series for children, *Akili and Me*, which has been culturally adapted and was reported in 2 studies [32,47], contains many elements described in this CMOC. The studies discuss the use of characters and structures that children become familiar with as well as the use of color and music to ease cognitive load and improve learning, recall, and retention of information.

Content

The construct *content* refers to the message itself, the information it contains, and how it is expressed.

Content CMOC 1: Emotion

Twenty-two animations (reported in 27 publications [16,24,25,27,28,30,39-43,49-51,53,57,58,63-65,67-71,74,75]) were identified as shaping and relating to this CMOC. This CMOC was met with much enthusiasm by stakeholders, with remarks that emotion has the power to increase message engagement and behavioral motivation. The value, relevance, and importance of evocative messaging were recognized by all stakeholders, with some cautioning that consideration for the emotions and feelings being elicited is needed so as not to induce or trigger distress or negative thoughts, emotions, or behaviors. Professional stakeholders communicated that evocative content could be powerful enough to be impactful without the use of language.

CoVideo is an animated video that evokes emotion through storytelling, character movements, the use of colors, sounds, and other elements [16]. This short animation is an example of how emotions can be expressed and elicited in an audience without using language.

Content CMOC 2: Co-Design

Fourteen animations (reported in 15 publications [24-26,33,37,43,45,56,57,59,60,65,68,73,76]) were identified as shaping and relating to this CMOC. When the target audience of an animation co-designs the content, the animation can reflect what is important and relevant to them, triggering the mechanisms of perceived personal relevance, acceptability, and the feeling that individuals' needs are being met. Stakeholders agreed that co-design provides crucial insight into the lived experiences of the target audience, while reducing barriers to messages, increasing accessibility and identification, and thereby meeting the needs of the intended viewer.

Digital Animation in Service Improvement is an example of an animation intervention that engaged in co-design at multiple points in time and with different groups [43]. This animation was developed over a 6-month period in consultation with individuals who were representative of the target audience as well as relevant medical professionals. These groups informed the key messages, cultural appropriateness, and other elements of the design and content of the animation.

Content CMOC 3: Function

Forty-four animations (reported in 52 publications [14-16,24-30,32-39,41-45,47-64,66-76]) were identified as shaping and relating to this CMOC. Animations that present a

XSL•FO
RenderX

challenge and its solution can trigger mechanisms such as building self-efficacy, understanding, and learning. Educating audiences about behavioral challenges is largely insufficient for behavior change but showing *how* to overcome these challenges can increase motivation by providing critical know-how information. When animations present this type of content, they often provide modeling and vicarious learning opportunities, increasing self-efficacy and motivation. Stakeholders agreed with this CMOC and felt that providing solutions is vital to empower individuals to overcome health challenges, and that solution-based content may complement and increase the effectiveness of emotion-based content.

An animation providing information about menopause serves as an example of this CMOC [44]. The short videos explain what menopause is, what symptoms may be experienced, multiple potential solutions for women experiencing symptoms, and resources they can use to find more information or get assistance.

Delivery

The construct *delivery* refers to how health animations are experienced, including who is delivering the message, what platforms are being used, and how and through what modes animations are accessed. The delivery of health animations also includes the intensity of exposure.

Delivery CMOC 1: Accessibility

Forty-three animations (reported in 50 publications [14-16,24-39,41-67,69-71,73]) were identified as shaping and relating to this CMOC. This CMOC recognizes that accessibility is key to the delivery of health animations. When animations are delivered in an accessible way for the target audience, mechanisms of engagement and understanding are triggered. Accessible delivery can vary for different audiences and challenges (eg, in-person or online, viewing in a group, alone, or in the presence of a health care professional). Accessibility also relates to literacy levels and meeting audiences' needs. Professional stakeholders felt that this CMOC addressed an important consideration and made sense within their organizations. The settings where animations are shown and how they are accessed were discussed by stakeholders, including the need to address barriers to comprehension and literacy for marginalized communities.

A series of GIFs promoting COVID-19 preventive behaviors display the multiple meanings of accessibility expressed in this CMOC [56]. First, the format—GIFs—are small image files that are easily shared on mobile devices and, second, the GIFs are free of written or spoken words (ie, languageless), thereby communicating the messages through visual storytelling only and addressing potential language and literacy barriers.

Delivery CMOC 2: Source

Thirty-three animations (reported in 41 publications [14-16,24-39,41-45,47,50,51,53-60,63-66,68,73]) were identified as shaping and relating to this CMOC. Source, in relation to this CMOC, can refer to the producer, person, or platform delivering the message, as well as the setting in which it is viewed. If the source is perceived to be credible, reliable, and trusted by the viewer, it can trigger mechanisms such as belief, engagement, and building trust, thereby enhancing attention, recall, and memory. Stakeholders felt that when a source is recognized and trusted, audiences engage more quickly and easily, and would be more likely to act on the message. Many stakeholders recognized that perceptions of credibility are influenced by online and offline misinformation and disinformation, and that the digital and nondigital settings where people are exposed to animations impact the perceived trustworthiness.

Many animations displayed clear information about the source of the information contained within them and who was providing the information. One example of this is an animation promoting flu vaccines for pregnant women, where there is clear information about the source of the animation and how or where to get additional information or have questions answered [57].

Delivery CMOC 3: Exposure

Thirty-one animations (reported in 38 publications [14-16,25-32,34-39,41,42,44,45,47,49-51,53-60,62-65,68]) were identified as shaping and relating to this CMOC. This CMOC refers to the way audiences are exposed to health animations in terms of auditory and/or visual channels, cognitive pathways, and intensity of viewing (single or series, over time or all at once, and so on). If these and other delivery elements meet the needs of the target audience, then cognitive load may be decreased, thereby further optimizing cognitive processing. Stakeholders expressed that the CMOC was clear and made sense to them and that addressing audiences' needs can maximize an animation's impact.

A series of short animations, titled "Respect Your Brain," reported in 2 studies [36,37], is associated with this CMOC. These well-received short animations have a similar format throughout the series and were delivered online as a package.

These CMOCs were used to develop a set of 10 recommendations for the design, content, and delivery of future health animations in research and practice (Textbox 4).

XSL•FO
RenderX

Textbox 4. Research and practice recommendations for the design, content and delivery of health animations

- *Design* the animation to *represent the target audience* (or target behavior or challenge) so that the audience can identify with and pay greater attention to the animation elements
- *Design* the animation to be *entertaining and use storytelling* so that the audience can be transported into the story
- *Design* the animation to *facilitate efficient cognitive processing* (memory, recall, learning, and so on).
- *Content* should be presented in an *evocative manner* so that powerful feelings, thoughts, and emotions are elicited.
- *Content* should be *co-designed* so that the animation is more relevant, acceptable, and meets the needs of the target audience.
- *Content* should include *information and a solution* so that the audience understands how, and feels confident, to act on the animation's information.
- *Content and delivery* should be *accessible* (language, literacy levels, free of jargon, and so on) to the target audience so that the messages can be understood.
- *Deliver* the animation in a *setting* that the target audience can *easily access* (online health or community center, public space, and so on) so that they can engage with and understand it.
- *Deliver* the animation using a *source* (individual or organization, platform, and/or setting) that is *perceived as credible* by the target audience so that they trust and engage with the message.
- *Deliver* the animation using a *mode* (in-person, digital, or both), *channel* (auditory or visual pathways), and *intensity* (single or series, one time or repeated) that *meet the needs* of the target audience so that cognitive load is reduced.

Discussion

Principal Findings

This realist review identified and synthesized evidence from 48 health animations providing an understanding of how, why, for whom, to what extent, and in which contexts health animations designed to promote preventive health behaviors work or fail to work. With input from public and professional stakeholders, we developed a program theory comprising high-level constructs and CMOCs, which were translated into a set of recommendations to inform the design, development and implementation of future public health animations.

The review produced 4 key findings. First, there are three interrelated high-level constructs (ie, design, content, and delivery) that often work synergistically. Furthermore, when exploring these constructs and their use in health animations, they were broken down into specific contexts thought to trigger causal mechanisms.

Second, entertainment was identified as a key context used in many health animations, through storytelling and its elements. Storytelling is based on the theory that humans understand and give meaning to their life experiences by telling stories [77]. Moyer-Guse et al [78] discussed the role of the cognitive mechanisms of persuasion and transportation by a story in relation to childhood vaccination attitudes and further discussed the power of narratives and entertainment to affect attitudes and beliefs. More broadly, "entertainment education" approaches have historically been used for health communication globally and to good effect across a range of health domains [79]. In this review, the terms "storytelling" and "entertainment" were used instead of "narrative" and "didactic," as conversations with the research team and stakeholders revealed that the terms "narrative" and "didactic" had the capacity to take on multiple meanings and cause confusion.

Third, the entertainment context often operated in conjunction with other contexts identified in the review. For example, when emotionally evocative content was present, storytelling, and/or entertainment were also present. This finding is in line with evidence that stories can capture audiences' attention, educate and entertain them, and evoke emotions [80]. Concurrent with evoking emotions, is the possibility of psychological reactance, which, in the context of health communication refers to an audience's resistance to act upon a health message in the intended way due to a perceived threat to their freedom [81].

Fourth, contexts that facilitated accessibility were deemed particularly valuable to stakeholders, who frequently discussed the important challenge for health communications to address health literacy. Health literacy is a key social determinant of health [82]. Literature reviews consistently show that low levels of health literacy are associated with worse health outcomes, including higher mortality rates, more hospitalizations and emergency department visits, and poor engagement with healthy behaviors, health messages, and preventive health care services [83,84]. Educational attainment has been shown to be the most important determinant of health literacy, followed by other socioeconomic factors such as financial resources [85]. Therefore, within the current context of health animations, it is of paramount importance to consider the role of these factors in determining the impact of an animation. The contexts that facilitate accessibility to and comprehension of the information and resources provided in health animations were present across all three of the high-level constructs identified in this review and therefore equally require full consideration in efforts to enhance health literacy.

Recommendations for Future Research and Practice

Based on the findings of this review, we suggest a set of recommendations for researchers and practitioners to consider when designing and developing future health animations. The relative weight or focus that can be given to each recommendation may depend on multiple factors including the health topic, audience, time, and financial resources available to develop the health animation. For example, when developing an animation designed for a global or multilingual audience, to

XSL•FO
RenderX

maximize its potential reach without requiring additional resources to develop multiple versions in different languages, it would be of paramount importance that the animation does not rely on written or spoken language to be understood, thereby reducing language and literacy barriers. Instead, storytelling, entertainment, and emotional elements could be exploited to ensure that the animation visuals alone can work to promote the target health behaviors. It would also be important to consider the most acceptable and utilized modes for animation delivery (eg, social media, instant messaging, and screens in public places) that meet the differing needs of subgroups in the target population. Researchers and practitioners should also consult these recommendations when reviewing data from individual studies evaluating health animations to appraise the specific features of the animation design, content, and delivery.

Limitations and Strengths

Few evaluations of animations provided data on behavioral outcomes and therefore, our program theory largely reflects the contexts and mechanisms at play when producing effects on key evidence-based behavioral determinants. The behavioral determinants that the animations were found to impact included evidence-based correlates of subsequent health behavior, namely behavioral skills, attitudes, and intentions [86]. Critically, interventions targeting behavioral skills and attitudes are known to produce small effects on health behavior [87], and therefore animations we identified as evidencing impact on these determinants may also yield positive behavioral changes. Moreover, many of the animations identified in our review reported knowledge gains, and even though interventions targeting knowledge are known to produce only minimal effects on behavior [87], knowledge often plays a critical role in health behavior change in terms of raising awareness of the health problem or topic, the need to change behavior, and its link to behavioral and self-regulation skills [88,89].

Future studies are urged to collect both short- and long-term behavioral data when evaluating health animations to expand the evidence base regarding their effect on behavioral outcomes. Furthermore, future research is needed to investigate the role animations could play as part of wider behavior change interventions, potentially targeting not only the more impactful individual determinants of behavior but also sociostructural factors [86,87]. Finally, investigations into potential dose-response effects through repeated exposure to an animation (or a series) are required to further our knowledge of the use of animations, especially since they can be easily and inexpensively shared digitally [9].

Limitations in the type of data, depth, and quality of reporting in studies on the design and implementation of public health interventions within published literature are well known [90]. Consequently, the pool of available literature for this review was reduced due to studies failing to provide access to the animation. Nevertheless, increasing interest in and application of health animations mean that the scientific evidence base will continue to expand.

Despite the aforementioned limitations, this review also has the following strengths. This is the first realist review of health animations designed for preventive health behavior promotion in the general population. We strictly adhered to the standards and guidance for conducting and reporting realist reviews [18,19,22], and as such, the review team engaged in regular discussion and reflection on the evidence identified. We actively and authentically included public and professional stakeholders throughout the review process, seeking to ensure that we involved international professional stakeholders from different sectors and public stakeholders from diverse ethnic and sociodemographic backgrounds. The resultant program theory extends the knowledge base by identifying and refining thinking around the questions of why, how, for whom, to what extent, and in which contexts health animations work as a digital form of health communication. The iterative, cyclical nature of realist reviews meant that searching for new evidence continued throughout the review process; therefore, noteworthy new evidence was captured and included in the synthesis. This new evidence reinforced initial findings shared with stakeholders and contributed to the currency of the review outcomes.

Conclusions

Animations provide a compelling digital approach to support the promotion of preventive health behaviors. Their visual and entertaining nature, with reduced reliance on language, means that they can help reduce health literacy barriers in the target audience and address health inequities in populations. The findings of this realist review of health animations designed to promote preventive health behaviors, along with the program theory it has generated, provide novel insights into the specific contexts and mechanisms influencing evidence-based determinants of behavior and behavior change. It is critical that the contexts and mechanisms found to drive animation impacts in the evidence we reviewed, as well as the associated recommendations we provide, are considered by those designing and developing health animations in the future.

Acknowledgments

The authors thank all contributors of publications who provided access to the animations. No generative artificial intelligence tools were used at any stage in the preparation of the manuscript.

Funding

The authors declared no financial support was received for this work.

XSL•FO
RenderX

Data Availability

The datasets generated during the data extraction process of this review are available from the corresponding author upon request.

Authors' Contributions

NO conceptualized the idea for this review. KMC and NO led the design of the review and wrote the review protocol, with methodological advice from GM, EL, MC, SV, and TB. LE designed the search strategy and conducted the database searches. KMC conducted the gray literature searches. KMC and NO independently screened all records and determined eligibility, with GM providing consensus in cases of disagreement. KMC extracted data and assessed the quality. Data analysis was conducted by KMC, GM, and NO. KMC and NO wrote the first draft of the manuscript. GM, LE, EL, MC, SV, and TB critically reviewed and approved the content of the manuscript. All authors had full access to all the data in the study and accept responsibility for the decision to submit it for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Ovid MEDLINE(R) search strategy.
[DOCX File, 14 KB - jmir_v28i1e79769_app1.docx]

Multimedia Appendix 2
Table of characteristics for peer-reviewed publications.
[DOCX File, 42 KB - jmir_v28i1e79769_app2.docx]

Multimedia Appendix 3
Table of characteristics for gray literature.
[DOCX File, 19 KB - jmir_v28i1e79769_app3.docx]

Multimedia Appendix 4
Table of constructs, contexts, and mechanisms.
[DOCX File, 17 KB - jmir_v28i1e79769_app4.docx]

References

1. The Health Foundation—written evidence (NHS0172). : House of Lords; 2016 URL: https://www.parliament.uk/globalassets/documents/lords-committees/NHS-Sustainability/long-term-sustainability-nhs-committee-written-evidence.pdf [accessed 2022-11-11]
2. Martinez R, Lloyd-Sherlock P, Soliz P, et al. Trends in premature avertable mortality from non-communicable diseases for 195 countries and territories, 1990-2017: a population-based study. Lancet Glob Health 2020 Apr;8(4):e511-e523. [doi: 10.1016/S2214-109X(20)30035-8] [Medline: 32199120]
3. Bolnick HJ, Bui AL, Bulchis A, et al. Health-care spending attributable to modifiable risk factors in the USA: an economic attribution analysis. Lancet Public Health 2020 Oct;5(10):e525-e535. [doi: 10.1016/S2468-2667(20)30203-6] [Medline: 33007211]
4. Bernhardt JM. Communication at the core of effective public health. Am J Public Health 2004 Dec;94(12):2051-2053. [doi: 10.2105/ajph.94.12.2051] [Medline: 15569948]
5. Health literacy development for the prevention and control of noncommunicable diseases: volume 2. a globally relevant perspective. : World Health Organization; 2022 URL: https://www.who.int/publications/i/item/9789240055353 [accessed 2022-11-11]
6. Houts PS, Doak CC, Doak LG, Loscalzo MJ. The role of pictures in improving health communication: a review of research on attention, comprehension, recall, and adherence. Patient Educ Couns 2006 May;61(2):173-190. [doi: 10.1016/j.pec.2005.05.004] [Medline: 16122896]
7. Nelson DL, Reed VS, Walling JR. Pictorial superiority effect. J Exp Psychol Hum Learn 1976 Sep;2(5):523-528. [doi: 10.1037/0278-7393.2.5.523] [Medline: 1003125]
8. Mayer RE, Moreno R. Animation as an aid to multimedia learning. Educ Psychol Rev 2002 Mar;14(1):87-99. [doi: 10.1023/A:1013184611077]
9. The ongoing journey to commitment and transformation: digital health in the WHO European Region, 2023. : World Health Organization; 2023 URL: https://www.who.int/europe/publications/m/item/digital-health-in-the-who-european-region-the-ongoing-journey-to-commitment-and-transformation [accessed 2025-03-20]

XSL•FO
RenderX

10. Feeley TH, Keller M, Kayler L. Using animated videos to increase patient knowledge: a meta-analytic review. Health Educ Behav 2023 Apr;50(2):240-249. [doi: 10.1177/10901981221116791] [Medline: 35950326]
11. Moe-Byrne T, Evans E, Benhebil N, Knapp P. The effectiveness of video animations as information tools for patients and the general public: a systematic review. Front Digit Health 2022;4:1010779. [doi: 10.3389/fdgth.2022.1010779] [Medline: 36777110]
12. Knapp P, Benhebil N, Evans E, Moe-Byrne T. The effectiveness of video animations in the education of healthcare practitioners and student practitioners: a systematic review of trials. Perspect Med Educ 2022 Dec;11(6):309-315. [doi: 10.1007/s40037-022-00736-6] [Medline: 36472616]
13. Höffler TN, Leutner D. Instructional animation versus static pictures: a meta-analysis. Learn Instr 2007 Dec;17(6):722-738. [doi: 10.1016/j.learninstruc.2007.09.013]
14. Meppelink CS, van Weert JCM, Haven CJ, Smit EG. The effectiveness of health animations in audiences with different health literacy levels: an experimental study. J Med Internet Res 2015 Jan 13;17(1):e11. [doi: 10.2196/jmir.3979] [Medline: 25586711]
15. Elling JM, De Vries H. Influence of animation- versus text-based delivery of a web-based computer-tailored smoking cessation intervention on user perceptions. Eur J Health Commun 2021;2(3):1-23. [doi: 10.47368/ejhc.2021.301]
16. Vandormael A, Adam M, Greuel M, et al. The effect of a wordless, animated, social media video intervention on covid-19 prevention: online randomized controlled trial. JMIR Public Health Surveill 2021 Jul 27;7(7):e29060. [doi: 10.2196/29060] [Medline: 34174778]
17. Pawson R. Evidence-Based Policy: A Realist Perspective: Sage Publications Ltd; 2006.
18. Wong G, Greenhalgh T, Westhorp G, Buckingham J, Pawson R. RAMESES publication standards: realist syntheses. BMC Med 2013 Jan 29;11(1):21. [doi: 10.1186/1741-7015-11-21] [Medline: 23360677]
19. Wong G, Westhorp G, Pawson R, Greenhalgh T. Realist synthesis: RAMESES training materials. The RAMESES Project. 2013 URL: https://www.ramesesproject.org/media/Realist_reviews_training_materials.pdf [accessed 2023-01-07]
20. McCorry K, Maniatopoulos G, Errington L, et al. Animations to communicate public health prevention messages: a realist review protocol. BMJ Open 2024 Aug 9;14(8):e083013. [doi: 10.1136/bmjopen-2023-083013] [Medline: 39122406]
21. Duddy C, Wong G. Grand rounds in methodology: when are realist reviews useful, and what does a “good” realist review look like? BMJ Qual Saf 2023 Mar;32(3):173-180. [doi: 10.1136/bmjqs-2022-015236] [Medline: 36585019]
22. Dada S, Dalkin S, Gilmore B, Hunter R, Mukumbang FC. Applying and reporting relevance, richness and rigour in realist evidence appraisals: advancing key concepts in realist reviews. Res Synth Methods 2023 May;14(3):504-514. [doi: 10.1002/jrsm.1630] [Medline: 36872619]
23. Hunter R, Gorely T, Beattie M, Harris K. Realist review. Int Rev Sport Exerc Psychol 2022 Dec 31;15(1):242-265. [doi: 10.1080/1750984X.2021.1969674]
24. Adam M, Johnston J, Job N, et al. Evaluation of a community-based mobile video breastfeeding intervention in Khayelitsha, South Africa: the Philani MOVIE cluster-randomized controlled trial. PLOS Med 2021 Sep;18(9):e1003744. [doi: 10.1371/journal.pmed.1003744] [Medline: 34582438]
25. Adam M, Kwinda Z, Dronavalli M, et al. Effect of short, animated video storytelling on maternal knowledge and satisfaction in the perinatal period in South Africa: randomized controlled trial. J Med Internet Res 2023 Oct 13;25:e47266. [doi: 10.2196/47266] [Medline: 37831505]
26. Aeilts AM, Carpenter KM, Hovick SR, Byrne L, Shoben AB, Senter L. BRCAShare-Assessment of an animated digital message for intrafamilial communication of pathogenic variant positive test results: a feasibility study. J Genet Couns 2023 Apr;32(2):475-485. [doi: 10.1002/jgc4.1656] [Medline: 36426678]
27. Aspiras O, Lucas T, Thompson HS, Manning MA. Medical mistrust, culturally targeted message framing, and colorectal cancer screening among African Americans. J Behav Med 2023 Oct;46(5):871-881. [doi: 10.1007/s10865-023-00415-9] [Medline: 37140761]
28. Aspiras O, Lucas T, Thompson HS, et al. Culturally targeted message framing and colorectal cancer screening preferences among African Americans. Health Psychol 2023 Jan;42(1):1-4. [doi: 10.1037/hea0001246] [Medline: 36409102]
29. Bayraktar Nahir C, Coşgun A, Şahin ND, Altan H. Primary school children’s oral hygiene knowledge assessed with different educational methods: a cross-sectional study. BMC Oral Health 2025 Jan 28;25(1):148. [doi: 10.1186/s12903-025-05493-6] [Medline: 39875880]
30. Beleites F, Adam M, Favaretti C, et al. Evaluating the impact of short animated videos on COVID-19 vaccine hesitancy: an online randomized controlled trial. Internet Interv 2024 Mar;35:100694. [doi: 10.1016/j.invent.2023.100694] [Medline: 38149090]
31. Bonner C, Taba M, Fajardo MA, et al. Using health literacy principles to improve understanding of evolving evidence in health emergencies: optimisation and evaluation of a COVID-19 vaccination risk-benefit calculator. Vaccine 2024 Oct;42(24):126296. [doi: 10.1016/j.vaccine.2024.126296]
32. Borzekowski DLG, Lando AL, Olsen SH, Giffen L. The impact of an educational media intervention to support children’s early learning in Rwanda. IJEC 2019 Apr;51(1):109-126. [doi: 10.1007/s13158-019-00237-4]

XSL•FO
RenderX

33. Brijnath B, Markusevska S, Lock K, et al. Evaluating the effectiveness of a culturally adapted dementia prevention animation on ethnically diverse participants' knowledge of dementia prevention. Clin Gerontol 2025 Jan 28:1-14. [doi: 10.1080/07317115.2025.2453710] [Medline: 39873516]
34. Burapasikarin C, Manonai J, Wattanayingcharoenchai R. The effect of an educational video on long-acting reversible contraception (LARC) utilization at 6-8 weeks postpartum period: a randomized controlled trial. Arch Gynecol Obstet 2020 Dec;302(6):1503-1509. [doi: 10.1007/s00404-020-05710-3] [Medline: 32734413]
35. Crum AJ, Heathcote LC, Morrison Z, et al. Changing mindsets about side effects of the COVID-19 vaccination: a randomized controlled trial. Ann Behav Med 2023 Oct 16;57(11):901-909. [doi: 10.1093/abm/kaad020] [Medline: 37279932]
36. Debenham J, Newton N, Champion K, et al. Neuroscience literacy and substance use prevention: How well do young people understand their brain? Health Promot J Austr 2022 Apr;33(2):395-402. [doi: 10.1002/hpja.516] [Medline: 34173994]
37. Debenham J, Birrell L, Newton NC, et al. Co-design of digital public health substance use resources: a collaboration between young people and experts. Health Promot J Austr 2025 Jan;36(1):e874. [doi: 10.1002/hpja.874] [Medline: 38797528]
38. Ewald DR, Orsini MM, Strack RW. The path to good health: shifting the dialogue and promoting social ecological thinking. SSM Popul Health 2023 Jun;22:101378. [doi: 10.1016/j.ssmph.2023.101378] [Medline: 36992717]
39. Favaretti C, Vandormael A, Hachaturyan V, et al. Participant engagement and reactance to a short, animated video about added sugars: web-based randomized controlled trial. JMIR Public Health Surveill 2022 Jan 24;8(1):e29669. [doi: 10.2196/29669] [Medline: 35072639]
40. Gallegos-Jeffrey A, Barth C, Smith S, et al. Examining engagement and responses to a digital binge drinking campaign for young males in Florida. J Am Coll Health 2025 May;73(5):2213-2218. [doi: 10.1080/07448481.2023.2266038] [Medline: 37856480]
41. Grigg J, Manning V, Lockie D, et al. A brief intervention for improving alcohol literacy and reducing harmful alcohol use by women attending a breast screening service: a randomised controlled trial. Med J Aust 2023 Jun 19;218(11):511-519. [doi: 10.5694/mja2.51991] [Medline: 37247850]
42. Hachaturyan V, Adam M, Favaretti C, et al. Reactance to social authority in a sugar reduction informational video: web-based randomized controlled trial of 4013 participants. J Med Internet Res 2021 Nov 22;23(11):e29664. [doi: 10.2196/29664] [Medline: 34813490]
43. Haith-Cooper M, Stacey T, Bailey F, Broadhead-Croft S. The co-development and feasibility-testing of an innovative digital animation intervention (DAISI) to reduce the risk of maternal sepsis in the postnatal period. Matern Child Health J 2020 Jul;24(7):837-844. [doi: 10.1007/s10995-020-02932-4] [Medline: 32356131]
44. Hammarberg K, Bandyopadhyay M, Nguyen H, et al. Development and evaluation of 4 short, animated videos for women in midlife promoting positive health behaviors: survey study. Interact J Med Res 2024 Dec 2;13:e60949. [doi: 10.2196/60949] [Medline: 39621404]
45. Harrison C, Greves G, Barnard E, et al. The effect of an educational animation on knowledge of testicular health and fertility of adolescents. Hum Reprod 2023 Dec 4;38(12):2470-2477. [doi: 10.1093/humrep/dead195] [Medline: 37805989]
46. Jones ASK, Fernandez J, Grey A, Petrie KJ. The impact of 3-D models versus animations on perceptions of osteoporosis and treatment motivation: a randomised trial. Ann Behav Med 2017 Dec;51(6):899-911. [doi: 10.1007/s12160-017-9913-1] [Medline: 28474286]
47. Kauffman LE, Dura EA, Borzekowski DLG. Emotions, strategies, and health: examining the impact of an educational program on Tanzanian preschool children. Int J Environ Res Public Health 2022 May 12;19(10):5884. [doi: 10.3390/ijerph19105884] [Medline: 35627420]
48. Lazić A, Kalinova KN, Packer J, et al. Social nudges for vaccination: how communicating herd behaviour influences vaccination intentions. Br J Health Psychol 2021 Nov;26(4):1219-1237. [doi: 10.1111/bjhp.12556] [Medline: 34495566]
49. Leiner M, Handal G, Williams D. Patient communication: a multidisciplinary approach using animated cartoons. Health Educ Res 2004 Oct;19(5):591-595. [doi: 10.1093/her/cyg079] [Medline: 15150139]
50. Lucas T, Thompson HS, Blessman J, et al. Effects of culturally targeted message framing on colorectal cancer screening among African Americans. Health Psychol 2021 May;40(5):305-315. [doi: 10.1037/hea0001073] [Medline: 34152784]
51. Lucas T, Rogers CR, Aspiras O, Manning M, Dawadi A, Thompson HS. Message framing for men? Gender moderated effects of culturally targeted message framing on colorectal cancer screening receptivity among African Americans. Psychol Men Masc 2023 Apr;24(2):103-112. [doi: 10.1037/men0000418] [Medline: 37193560]
52. Maisrikrod SC, Currie M, Govan BL, et al. Design and development of an internationally applicable educational video to increase community awareness in regions with high prevalence of melioidosis and diabetes. Am J Trop Med Hyg 2023 Mar 1;108(3):503-506. [doi: 10.4269/ajtmh.22-0024] [Medline: 36646077]
53. Mansor N, Ahmad N, Md Said S, Tan KA, Sutan R. Effectiveness of a theory-based digital animated video intervention to reduce intention and willingness to sext among diploma students: cluster randomized controlled trial. J Med Internet Res 2023 Oct 20;25:e48968. [doi: 10.2196/48968] [Medline: 37862090]
54. Mohd Jaini SNB, Sinor MZ, Ahmad B, Hasan R, Sanusi SY. Effectiveness of educational videos comparing single versus multiple topics: a cluster randomized controlled trial. Cureus 2024 Jul;16(7):e64028. [doi: 10.7759/cureus.64028] [Medline: 39109139]

XSL•FO
RenderX

55. Molina-Salas Y, Romera-Guirado FJ, García-Rubio A, Pérez-Martín JJ, Zornoza-Moreno M, for FLUSCHOOLVAC Team. Influence of the Olfatín Project on the reduction of pain related to intranasal influenza vaccination, as part of a school influenza vaccination program. J Pediatr Nurs 2024;79:52-58. [doi: 10.1016/j.pedn.2024.08.022] [Medline: 39197260]
56. O'Brien N, Vijaykumar S, Craig M, et al. A pre- and post-intervention study testing the effect of exposure to languageless animated images communicating COVID-19 preventive behaviours on behavioural intentions and beliefs of Guatemalan adults. J Glob Health 2022 Jul 22;12:05018. [doi: 10.7189/jogh.12.05018] [Medline: 35862262]
57. Parsons J, Grimley C, Newby K. Effectiveness of a digital intervention in increasing flu vaccination–related risk appraisal, intention to vaccinate and vaccination behaviour among pregnant women. Health Educ Behav 2022 Dec;49(6):1033-1041. [doi: 10.1177/10901981221077935] [Medline: 35255730]
58. Ruparel M, Quaife SL, Ghimire B, et al. Impact of a lung cancer screening information film on informed decision-making: a randomized trial. Ann Am Thorac Soc 2019 Jun;16(6):744-751. [doi: 10.1513/AnnalsATS.201811-841OC] [Medline: 31082267]
59. Salmon DA, Dudley MZ, Brewer J, et al. LetsTalkShots: personalized vaccine risk communication. Front Public Health 2023;11:1195751. [doi: 10.3389/fpubh.2023.1195751] [Medline: 37457264]
60. Schuh HB, Rimal RN, Breiman RF, et al. Evaluation of online videos to engage viewers and support decision-making for COVID-19 vaccination: how narratives and race/ethnicity enhance viewer experiences. Front Public Health 2023;11:1192676. [doi: 10.3389/fpubh.2023.1192676] [Medline: 37670826]
61. Mohd Shukri NA, Norman MFH, Rostam MA, et al. Development and evaluation of educational video on healthy maternal diet to prevent child stunting. JSSM 2024;19(10):59-71. [doi: 10.46754/jssm.2024.10.006]
62. Smith J, Cvejic E, Houssami N, et al. Randomized trial of information for older women about cessation of breast cancer screening invitations. J Gen Intern Med 2024 Jun;39(8):1332-1341. [doi: 10.1007/s11606-024-08656-3] [Medline: 38409512]
63. Sumnall HR, Atkinson AM, Anderson M, McAuley A, Trayner K. How to save a life: public awareness of a national mass media take home Naloxone campaign, and effects of exposure to campaign components on overdose knowledge and responses. Int J Drug Policy 2024 Sep;131:104111. [doi: 10.1016/j.drugpo.2023.104111] [Medline: 37563039]
64. Vandormael A, Hachaturyan V, Adam M, Favaretti C, Gates J, Bärnighausen T. Effect of a story-based, animated video to reduce added sugar consumption: a web-based randomized controlled trial. J Glob Health 2021;11:04064. [doi: 10.7189/jogh.11.04064] [Medline: 34737864]
65. Willis LA, Kachur R, Castellanos TJ, et al. Developing a motion comic for HIV/STD prevention for young people ages 15-24, part 2: evaluation of a pilot intervention. Health Commun 2018 Mar;33(3):229-237. [doi: 10.1080/10410236.2016.1255841] [Medline: 28033470]
66. Witus LS, Larson E. A randomized controlled trial of a video intervention shows evidence of increasing COVID-19 vaccination intention. PLoS ONE 2022;17(5):e0267580. [doi: 10.1371/journal.pone.0267580] [Medline: 35587462]
67. Zhu P, Tatar O, Griffin-Mathieu G, et al. The efficacy of a brief, altruism-eliciting video intervention in enhancing COVID-19 vaccination intentions among a population-based sample of younger adults: randomized controlled trial. JMIR Public Health Surveill 2022 May 30;8(5):e37328. [doi: 10.2196/37328] [Medline: 35544437]
68. COVIDLatino. URL: https://covidlatino.org/ [accessed 2024-02-01]
69. The Story of Cholera. Global Health Media. URL: https://globalhealthmedia.org/our-projects/cholera-series/ [accessed 2024-02-23]
70. The Story of Coronavirus. Global Health Media. URL: https://globalhealthmedia.org/our-projects/coronavirus-series/ [accessed 2024-02-22]
71. The Story of Ebola. Global Health Media. URL: https://globalhealthmedia.org/our-projects/ebola-series/ [accessed 2024-02-23]
72. New animation videos promote health and well-being of adolescents. UNESCO. URL: https://www.unesco.org/en/articles/new-animation-videos-promote-health-and-well-being-adolescents/ [accessed 2024-02-15]
73. Meena and UNICEF. UNICEF Bangladesh. URL: https://www.unicef.org/bangladesh/en/meena-and-unicef/ [accessed 2024-03-12]
74. Cholera animation. United Nations International Children's Emergency Fund YouTube page. URL: https://www.youtube.com/watch?v=G954-nmMAzg/ [accessed 2024-04-04]
75. Cholera—you can fight it too. World Health Organization Eastern Mediterranean Region YouTube page. 2018 Jan 14. URL: https://www.youtube.com/watch?v=8-sqticNg5o/ [accessed 2024-05-25]
76. Intercultural dialogue and the mojojoy fable. Pan American Health Organization YouTube page. 2024 May 18. URL: https://www.youtube.com/watch?v=FuytMCGNHGk/ [accessed 2024-08-08]
77. Fisher WR. The narrative paradigm: in the beginning. J Commun 1985 Dec 1;35(4):74-89. [doi: 10.1111/j.1460-2466.1985.tb02974.x]
78. Moyer-Gusé E, Rader K, Lavis S. Transportation into an entertainment narrative about the MMR vaccine: an investigation of self-referencing and issue-related thoughts in narrative persuasion. J Health Commun 2022 Aug 3;27(8):585-592. [doi: 10.1080/10810730.2022.2138641] [Medline: 36314638]
79. Shen F, Han J. Effectiveness of entertainment education in communicating health information: a systematic review. Asian J Commun 2014 Nov 2;24(6):605-616. [doi: 10.1080/01292986.2014.927895]

XSL•FO
RenderX

80. Krause RJ, Rucker DD. Strategic storytelling: when narratives help versus hurt the persuasive power of facts. Pers Soc Psychol Bull 2020 Feb;46(2):216-227. [doi: 10.1177/0146167219853845] [Medline: 31179845]
81. Reynolds-Tylus T. Psychological reactance and persuasive health communication: a review of the literature. Front Commun 2019;4:56. [doi: 10.3389/fcomm.2019.00056]
82. Coughlin SS, Vernon M, Hatzigeorgiou C, George V. Health literacy, social determinants of health, and disease prevention and control. J Environ Health Sci 2020;6(1):3061. [Medline: 33604453]
83. Walters R, Leslie SJ, Polson R, Cusack T, Gorely T. Establishing the efficacy of interventions to improve health literacy and health behaviours: a systematic review. BMC Public Health 2020 Jun 30;20(1):1040. [doi: 10.1186/s12889-020-08991-0] [Medline: 32605608]
84. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med 2011 Jul 19;155(2):97-107. [doi: 10.7326/0003-4819-155-2-201107190-00005] [Medline: 21768583]
85. Stormacq C, Van den Broucke S, Wosinski J. Does health literacy mediate the relationship between socioeconomic status and health disparities? Integrative review. Health Promot Int 2019 Oct 1;34(5):e1-e17. [doi: 10.1093/heapro/day062] [Medline: 30107564]
86. Hagger MS. Psychological determinants of health behavior. Annu Rev Psychol 2025 Jan;76(1):821-850. [doi: 10.1146/annurev-psych-020124-114222] [Medline: 39413250]
87. Albarracín D, Fayaz-Farkhad B, Samayoa JAG. Determinants of behaviour and their efficacy as targets of behavioural change interventions. Nat Rev Psychol 2024 Jun;3(6):377-392. [doi: 10.1038/s44159-024-00305-0] [Medline: 40909436]
88. de Vries H. An integrated approach for understanding health behavior; the I-Change model as an example. Psychol Behav Sci Int J 2017;2(2):555585. [doi: 10.19080/PBSIJ.2017.02.555585]
89. Ryan P. Integrated Theory of Health Behavior Change: background and intervention development. Clin Nurse Spec 2009;23(3):161-170. [doi: 10.1097/NUR.0b013e3181a42373] [Medline: 19395894]
90. Armstrong R, Waters E, Moore L, et al. Improving the reporting of public health intervention research: advancing TREND and CONSORT. J Public Health (Oxf) 2008 Mar;30(1):103-109. [doi: 10.1093/pubmed/fdm082] [Medline: 18204086]

Abbreviations

CMOC: context mechanism outcome configuration
MeSH: Medical Subject Headings
RAMESES: Realist and Meta-Narrative Evidence Syntheses: Evolving Standards

Edited by S Brini; submitted 27.Jun.2025; peer-reviewed by D Azizoddin, KK Jermsittiparsert; revised version received 17.Feb.2026; accepted 19.Feb.2026; published 23.Apr.2026.

Please cite as:
McCorry K, Maniatopoulos G, Errington L, Land E, Craig M, Vijaykumar S, Bärnighausen T, O'Brien N
Recognizing the Importance of Design, Content, and Delivery Features of Health Animations for Preventive Health Behaviors: Realist Review
J Med Internet Res 2026;28:e79769
URL: https://www.jmir.org/2026/1/e79769
doi:10.2196/79769

© Kathleen McCorry, Gregory Maniatopoulos, Linda Errington, Ellie Land, Michael Craig, Santosh Vijaykumar, Till Bärnighausen, Nicola O'Brien. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effectiveness of Digital Serious Games on Knowledge and Attitudes in Public Health Education: Systematic Review and Bayesian Network Meta-Analysis of Randomized Controlled Trials

Di Huang[1], RN, MMed; Dongjun Wu[2], PhD; Rene Hexel[3], PhD; Christine Brown-Wilson[4], RN, PhD; Jing Zhou[5], RN, PhD; Wendy Moyle[1], RN, PhD

[1]School of Nursing and Midwifery, Griffith University, 170 Kessels Road, Brisbane, Queensland, Australia
[2]School of Nursing and Midwifery, Griffith University, Gold Coast, Queensland, Australia
[3]School of Information and Communication Technology, Griffith University, Brisbane, Queensland, Australia
[4]School of Nursing and Midwifery, Queen's University Belfast, Belfast, Northern Ireland, United Kingdom
[5]Department of Nursing, The Affiliated Hospital of Zunyi Medical University, Zunyi, Guizhou, China

Corresponding Author:
Di Huang, RN, MMed
School of Nursing and Midwifery, Griffith University, 170 Kessels Road, Brisbane, Queensland, Australia

Abstract

Background: Inadequate health literacy and low engagement challenge public health education. Digital serious games show potential to enhance health knowledge and attitudes. However, the comparative effectiveness of different game formats is unclear.

Objective: This study aimed to evaluate and compare the effectiveness of different digital serious game formats in improving public health knowledge and attitudes.

Methods: This systematic review and Bayesian network meta-analysis followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines. Seven databases (PubMed, CINAHL, Embase, PsycINFO, Cochrane Library, Scopus, and Web of Science) were searched from January 2000 to October 2025. An updated search in February 2026 identified no additional studies. Eligible studies were randomized controlled trials (RCTs) involving nonprofessional participants comparing digital serious games with traditional or noninteractive education. Standardized mean differences and 95% credible intervals were pooled using Bayesian network models with random effects. Subgroup analyses examined population characteristics, intervention duration, health topic, and delivery format. Risk of bias was assessed using the Cochrane risk-of-bias tool, and evidence certainty was rated using the Grading of Recommendations Assessment, Development and Evaluation.

Results: Forty randomized controlled trials from 19 countries (N=8764 participants) were included. Digital serious games significantly improved knowledge (standardized mean difference 0.66, 95% CI 0.32 - 0.99; I^2=89.1%) and attitudes (standardized mean difference 0.50, 95% CI 0.27 - 0.76; I^2=80.7%) compared with traditional education. Multisession interventions showed larger effects than single-session interventions for knowledge (0.76 vs 0.43) and attitudes (0.53 vs 0.30), with greater improvements among adolescents, nonpatient populations, and Asian studies. Network meta-analysis showed low heterogeneity (I^2=8% for knowledge; 3% for attitudes). Mobile app–based, computer-offline, and web-based games ranked highest for knowledge; computer-offline, web-based, and virtual reality games ranked highest for attitudes. Evidence certainty was moderate for knowledge and low-to-moderate for attitudes.

Conclusions: Digital serious games improve public health knowledge and attitudes across diverse contexts. Using a Bayesian network meta-analysis of randomized controlled trials, this review compares the relative effectiveness of different game formats. Mobile app–based, computer-offline, and web-based games most improved knowledge; computer-offline, web-based, and virtual reality formats most improved attitudes. Multisession interventions were more effective than single-session ones, particularly for adolescents and nonpatient populations. These findings guide scalable digital health education strategies. Future research requires adequately powered trials, longer follow-up, and standardized frameworks.

Trial Registration: PROSPERO CRD420251056704; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251056704

(J Med Internet Res 2026;28:e89281) doi:10.2196/89281

KEYWORDS

digital serious games; public health education; knowledge; attitudes; Bayesian network meta-analysis; randomized controlled trials

XSL•FO
RenderX

Introduction

Inadequate health knowledge and attitudes remain a significant barrier to achieving global health targets [1], despite more than a century of organized public health education through mass media campaigns, school-based curricula, and community programs [2,3]. Health literacy gaps are evident across settings and income levels. In the European Union, 27% to 48% of adults have inadequate health literacy [4]. In China, only 31.87% of residents were health literate in 2024 [5], with substantial disparities by residence and education. In the United States, fewer than one-third of school-aged children met grade-level reading standards, indicating persistent barriers to acquiring and applying health information [6,7]. These gaps undermine vaccination and screening, compromise chronic disease (CD) management and medication adherence [8], and contribute to inequities in health outcomes and avoidable health care costs [9,10]. Addressing them requires strategies that can sustain engagement, broaden reach, and adapt to rapidly changing information environments.

Traditional public health education has improved awareness and behaviors, but its impact is limited by low long-term engagement, unequal access, and weak adaptability to rapidly changing communication environments [11,12]. The COVID-19 pandemic magnified these weaknesses, with school closures disrupting learning for more than 1.6 billion learners worldwide and exposing the fragility of knowledge dissemination systems [13]. This interruption also accelerated the adoption of digital health interventions, which offer scalable, interactive, and adaptable complements to conventional programs, extending reach, promoting equitable access, and strengthening public health knowledge, attitudes, and behaviors [14,15].

Within digital health interventions, digital serious games have emerged as a promising strategy to address persistent gaps in participation and impact [16,17]. By combining education with interactive and immersive play, they can sustain engagement and improve knowledge retention, supporting long-term behavioral change [18,19]. Their formats have progressed from desktop programs to mobile apps and online platforms and now increasingly incorporate virtual reality, augmented reality, artificial intelligence, and wearable devices [20-23]. This evolution has enabled broad application in public education, including infectious disease preparedness, CD management, and dementia awareness [24-26]. With their promise of scalability and equitable access [27-29], digital serious games are increasingly regarded as a complement to conventional approaches and digital tools, with potential to support prevention and health promotion at the population level.

Despite this promise, evidence remains insufficient to guide large-scale implementation. A 2020 scoping review mapped digital serious games for health education across health care providers, patients, and the public, showing expansion beyond disease-specific contexts but without assessing relative effectiveness across formats or populations [30]. Other reviews have concentrated on single conditions or target groups, such as diabetes [31], upper limb rehabilitation [32], or vaccination, offering little comparative insight [33]. Existing meta-analyses are similarly constrained, relying on pairwise comparisons that cannot establish the relative effectiveness of multiple intervention types [34,35]. For policymakers and educators, the central question is no longer whether digital serious games can work, but rather which formats are most effective, for which populations, and under what circumstances. No systematic evaluation has yet addressed these comparative questions, leaving a critical gap in the evidence needed to inform equitable and scalable public health education strategies.

To our knowledge, this systematic review and Bayesian network meta-analysis is the first to synthesize and compare the effectiveness of different formats of digital serious games in improving public health knowledge and attitudes and to examine how population characteristics, intervention duration, and contextual factors may moderate their impact.

Methods

Information Sources and Search Strategy

For this systematic review and network meta-analysis, we followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 and reported the search process in accordance with PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search extension). The completed PRISMA 2020, PRISMA 2020 Expanded, PRISMA-S, and PRISMA for Abstract checklists are provided in Checklist 1. The protocol was registered in PROSPERO (International Prospective Register of Systematic Reviews; CRD420251056704) [36]. All PRISMA 2020 items were reviewed against the manuscript to ensure complete and transparent reporting. We systematically searched 7 electronic databases (PubMed, CINAHL, Embase, PsycINFO, Cochrane Library, Scopus, and Web of Science) for studies published between January 1, 2000, and October 1, 2025, to identify any newly published studies meeting the eligibility criteria. Search terms combined keywords and Medical Subject Headings terms related to serious games, digital games, video games, public health education, knowledge, attitudes, and diseases. The search strategy was refined in accordance with Chapter 4.4 of the *Cochrane Handbook* to maximize sensitivity, including the expansion of controlled vocabulary and additional free-text synonyms. An updated search was conducted in February 2026, and no additional eligible studies were identified. Full search strategies for each database are provided in Multimedia Appendix 1. Reference lists of relevant systematic reviews and meta-analyses were also screened for potentially eligible studies (Multimedia Appendix 1). Gray literature sources and clinical trial registries were not searched separately.

All records were imported into Covidence (Veritas Health Innovation, Melbourne, Australia) for deduplication, screening, and data management. Two reviewers (DH and DW) independently screened titles, abstracts, and full texts, resolving disagreements by consensus or through consultation with a third reviewer (WM). Study authors were contacted when additional clarification was required.

XSL•FO
RenderX

Eligibility Criteria

Eligibility criteria followed the population, intervention, comparator, outcomes, and study design framework (Multimedia Appendix 2) [37]. We included randomized controlled trials (RCTs), cluster RCTs, and pilot RCTs of digital serious games designed to improve health-related knowledge or attitudes in nonprofessional populations, including children, adolescents, adults, informal caregivers, patients, and the general public. Interventions were required to be delivered via digital platforms, such as web-based applications, mobile apps, computer software, virtual reality, augmented reality, or robot-assisted systems. Comparators included no intervention, conventional education, or digital nongame tools. The primary outcomes were changes in knowledge or attitudes, assessed using validated instruments when reported.

We excluded studies evaluating nondigital games; interventions targeting clinical treatment, rehabilitation, or professional training; and studies that did not report at least 1 primary outcome. Nonrandomized studies, qualitative research, reviews, commentaries, protocols, and conference abstracts were also excluded.

Data Extraction

A standardized data extraction form was initially developed (DH) and subsequently refined (DH and DW) in accordance with the *Cochrane Handbook for Systematic Reviews of Interventions* [38]. The form was pilot-tested on a subset of studies to ensure reliability and reproducibility before full implementation. For each eligible trial, 2 reviewers (DH and DW) independently extracted data and verified the results with a third reviewer (WM). Extracted variables included study characteristics (identifier, year, country, design, and setting); participant characteristics (population group, mean age, sex distribution, sample size, and patient status); intervention details (type of digital serious game, delivery platform, educational content, duration, and follow-up, if reported); comparator details (type and format); and outcome measures (assessment tools, baseline and postintervention scores, and key findings related to knowledge or attitudes).

Outcomes

The primary outcomes were changes in health-related knowledge, including understanding of diseases, prevention, and health promotion, as well as changes in health-related attitudes, beliefs, perceptions, and intentions. Both outcomes were assessed using validated instruments when reported. No secondary outcomes were prespecified.

Bias Risk Assessment

The risk of bias for each included study was assessed using the revised Cochrane risk-of-bias tool for randomized trials tool with the appropriate version applied to individually randomized and cluster-randomized trials [39]. The tool evaluates potential bias in the following domains: the randomization process, deviations from intended interventions, missing outcome data, measurement of the outcomes, and selection of the reported results. Each trial was independently appraised by 2 reviewers (DH and DW) and categorized as low risk, some concerns, or high risk of bias. Any disagreements were resolved by consensus, with persistent discrepancies adjudicated by a third reviewer (WM).

Statistical Analysis

Pairwise Meta-Analysis

Pairwise meta-analyses were first performed to estimate pooled effects. For each study, mean values and SDs of intervention and control groups were extracted. When SDs were not directly reported, they were imputed from SEs, *P* values, *t* values, or 95% CIs. Studies without sufficient information for conversion were excluded from quantitative pooling.

Given the expected heterogeneity across populations, interventions, and outcome measures, pooled standardized mean differences (SMDs) with 95% CIs were calculated using a random-effects model with Hartung-Knapp adjustment to provide more robust variance estimation under conditions of limited study numbers and substantial between-study variability [40].

Between-study heterogeneity was assessed using Cochran *Q* test and quantified using the I^2 statistic [41]. Robustness of pooled estimates in the pairwise meta-analyses was evaluated through sensitivity analyses, including sequential exclusion of individual studies, application of fixed-effect models, and removal of trials at high risk of bias. Prespecified subgroup analyses explored potential sources of heterogeneity, stratified by population group, patient status, health topic, duration, publication decade, geographical region, and sex distribution (Multimedia Appendix 3).

Bayesian Network Meta-Analysis

Bayesian network meta-analysis was conducted using random-effects models implemented via Markov chain Monte Carlo simulation [42]. Although the included studies differed in populations, health topics, and intervention formats, all interventions were digital serious games targeting health knowledge or attitudes, supporting the conceptual comparability required for network meta-analysis. The analysis followed a predefined 8-step network meta-analysis workflow, including network construction, Bayesian model estimation, convergence diagnostics, inconsistency assessment, treatment ranking, estimation of relative treatment effects, calculation of prediction intervals, and robustness analysis. Analyses were performed in R software (version 4.5.2; R Foundation for Statistical Computing) using the *gemtc* and *netmeta* packages. Four Markov chains were run in parallel with different initial values, each with 5000 burn-in iterations followed by 20,000 sampling iterations. Convergence was assessed using trace plots and the Gelman-Rubin diagnostic, with a potential scale reduction factor below 1.05 indicating adequate convergence. Noninformative priors were specified for treatment effects, and vague priors were applied to the between-study heterogeneity parameter to minimize prior influence on the model estimates. Model fit was evaluated using the deviance information criterion [43].

Pooled SMDs with corresponding 95% credible intervals (CrIs) were generated for all intervention comparisons. Prediction intervals were additionally calculated to reflect the expected range of effects in future studies. Local inconsistency was

XSL•FO
RenderX

assessed using the node-splitting method [44], and global inconsistency was evaluated by comparing the deviance information criterion between the consistency model and the unrelated mean effects model. Between-study heterogeneity was accommodated using the random-effects model and quantified using the between-study variance parameter and overall network I^2. Relative treatment rankings were estimated using the surface under the cumulative ranking curve (SUCRA), mean ranks, and rank probabilities. Sensitivity analyses were conducted by applying alternative prior distributions for the heterogeneity parameter to assess the robustness of the model estimates. Complete R scripts for both pairwise and network meta-analyses are provided in Multimedia Appendix 4.

Evaluation of Publication Bias

Publication bias and small-study effects were assessed using the Egger regression test ($P<.10$) and comparison-adjusted funnel plots within the network meta-analysis framework. Interpretation of funnel plot asymmetry was undertaken cautiously, as between-study variability and model complexity may contribute to apparent asymmetry, independent of publication bias.

Certainty of Evidence

The certainty of the evidence was appraised using the GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) framework, following the GRADE Working Group's guidance. Certainty was evaluated across the domains of study design, risk of bias, inconsistency, indirectness, imprecision, and potential publication bias [45]. All included randomized trials, including pilot and cluster trials, were initially rated as high-certainty evidence. Downgrading was applied when serious limitations were identified, including high risk of bias, substantial unexplained heterogeneity, indirectness of the evidence in relation to the review question, imprecision of effect estimates, or potential publication bias, in accordance with GRADE guidance [46]. Potential publication bias was also assessed. Final ratings were categorized as high, moderate, low, or very low certainty.

Results

Study Selection

The database search yielded 9269 records, and a further 47 records were identified through reference searches of relevant systematic reviews and meta-analyses. After the removal of 5352 duplicates, 3917 records were screened by title and abstract (Figure 1). Of these, 3816 records were excluded, and 1 report could not be retrieved because the full text was unavailable. A total of 88 full-text articles were assessed for eligibility, of which 48 were excluded for reasons summarized in Multimedia Appendix 5. In total, 40 studies were included in the systematic review and 30 in the Bayesian network meta-analysis (Figure 2).

XSL•FO
RenderX

Figure 1. Forest plots showing pooled standardized mean differences (SMDs) for the effects of digital serious games on public health knowledge and attitudes in randomized controlled trials. (A) Knowledge outcomes (27 trials; random-effects model with Knapp-Hartung adjustment: SMD 0.66, 95% CI 0.32 - 0.99; prediction interval −0.97 to 2.29; I^2=89.1%). (B) Attitude outcomes (16 trials; random-effects model with Knapp-Hartung adjustment: SMD 0.50, 95% CI 0.27 - 0.76; prediction interval −0.42 to 1.43; I^2=80.7%). Squares represent individual study effect sizes (size proportional to study weight); horizontal lines represent 95% CIs; diamonds indicate pooled summary estimates [47-75].

XSL•FO
RenderX

(A) Knowledge

Study	Experimental Total	Experimental Mean	Experimental SD	Control Total	Control Mean	Control SD	SMD	95% CI	Weight
Aljafari et al., 2017	51	60.60	6.00	54	61.10	5.10	−0.09	[−0.47 to 0.29]	4.0%
Fadda et al., 2017	48	12.80	1.90	44	10.50	2.30	1.09	[0.65 to 1.53]	3.8%
Froome et al., 2020	39	13.50	3.80	34	10.40	3.20	0.87	[0.39 to 1.35]	3.7%
Goodman et al., 2024	26	9.40	3.10	27	9.10	3.00	0.10	[−0.44 to 0.64]	3.6%
Haruna et al., 2018	40	62.50	12.30	40	55.40	11.80	0.58	[0.14 to 1.03]	3.8%
Henkemans et al., 2017	14	17.20	4.10	13	13.50	3.80	0.91	[0.11 to 1.70]	2.9%
Kumar RS et al., 2022	48	8.94	1.97	42	7.88	1.27	0.63	[0.20 to 1.05]	3.9%
Koohmareh et al., 2021	30	12.63	1.13	30	7.63	1.99	3.05	[2.29 to 3.81]	3.0%
Shegog et al., 2001	38	21.10	5.40	33	17.80	6.30	0.56	[0.08 to 1.04]	3.7%
Tang et al., 2022	50	70.09	11.58	46	57.49	16.58	0.88	[0.46 to 1.30]	3.9%
Whiteley et al., 2018	32	3.09	1.42	29	3.00	1.28	0.07	[−0.44 to 0.57]	3.7%
Zolfaghari et al., 2021	46	14.30	2.00	47	13.10	1.60	0.66	[0.24 to 1.08]	3.9%
Aljafari et al., 2022	144	57.80	6.30	139	57.50	5.10	0.05	[−0.18 to 0.29]	4.3%
Carcioppolo et al., 2022	100	6.75	2.62	102	6.13	2.22	0.25	[−0.02 to 0.53]	4.2%
Espinosa−Curiel et al., 2022	15	5.70	2.40	12	4.08	1.60	0.75	[−0.04 to 1.54]	2.9%
Ghadam et al., 2023	80	56.20	9.40	80	50.60	6.90	0.68	[0.36 to 0.99]	4.1%
Huang et al., 2025	51	6.14	1.83	51	7.06	1.82	−0.50	[−0.89 to −0.11]	3.9%
Joubert et al., 2016	37	64.80	12.30	34	63.70	11.90	0.09	[−0.38 to 0.56]	3.8%
Kato et al., 2008	164	0.66	0.20	148	0.63	0.20	0.15	[−0.07 to 0.37]	4.3%
Koniou et al., 2025	51	36.92	3.24	51	33.14	3.08	1.19	[0.76 to 1.61]	3.9%
Kumar et al., 2004	19	13.30	1.50	21	12.20	2.00	0.61	[−0.03 to 1.24]	3.3%
Liu et al., 2024	18	7.28	2.06	18	6.33	2.18	0.44	[−0.22 to 1.10]	3.3%
Mack et al., 2020	172	22.84	4.05	144	20.69	3.84	0.54	[0.32 to 0.77]	4.3%
Maddison et al., 2022	15	6.60	2.10	8	5.80	2.60	0.34	[−0.53 to 1.20]	2.8%
Nazmi et al., 2025	45	10.40	1.60	45	4.96	1.36	3.63	[2.95 to 4.31]	3.2%
Bartholomew et al., 2000	29	14.40	13.40	25	2.00	12.70	0.93	[0.37 to 1.50]	3.5%
Tan et al., 2022	178	10.00	2.30	196	10.00	2.30	0.00	[−0.20 to 0.20]	4.3%
Random effects model	**1580**			**1513**			**0.66**	**[0.32 to 0.99]**	**100.0%**
Prediction interval								**[−0.97 to 2.29]**	

Heterogeneity: I^2 = 89.1%, τ^2 = 0.60, P < .001

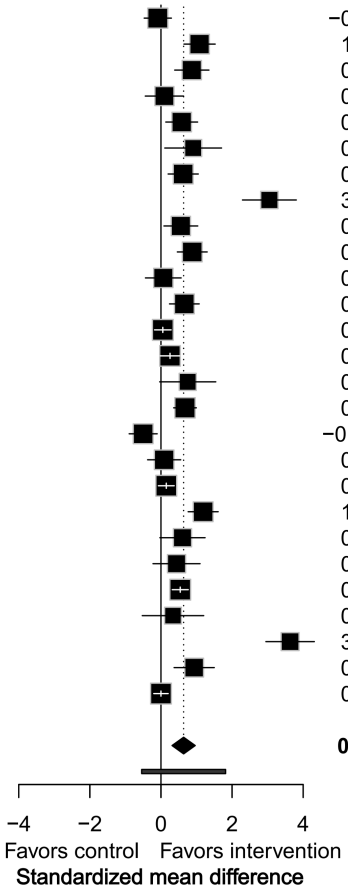

(B) Attitude

Study	Experimental Total	Experimental Mean	Experimental SD	Control Total	Control Mean	Control SD	SMD	95% CI	Weight
Fadda et al., 2017	48	3.60	0.50	44	3.30	0.60	0.54	[0.12 to 0.96]	6.5%
Pouls et al., 2022	110	2.50	4.90	111	2.30	5.20	0.04	[−0.22 to 0.30]	7.5%
Shegog et al., 2001	38	56.50	9.80	32	51.50	10.70	0.48	[0.01 to 0.96]	6.1%
Tang et al., 2022	50	82.00	23.44	46	79.62	17.94	0.11	[−0.29 to 0.51]	6.6%
Whiteley et al., 2018	32	19.83	4.68	29	19.48	5.12	0.07	[−0.43 to 0.57]	5.9%
Carcioppolo et al., 2022	100	5.89	1.02	102	5.65	1.11	0.22	[−0.05 to 0.50]	7.4%
Espinosa−Curiel et al., 2022	15	3.07	0.80	12	2.30	1.40	0.68	[−0.11 to 1.46]	4.1%
Ghadam et al., 2023	80	71.10	6.00	80	61.00	9.30	1.28	[0.94 to 1.63]	7.0%
Joubert et al., 2016	37	58.60	14.10	34	59.40	13.70	−0.06	[−0.52 to 0.41]	6.2%
Koniou et al., 2025	51	77.18	6.21	51	71.29	5.98	0.96	[0.55 to 1.37]	6.5%
Liu et al., 2024	18	5.41	0.65	18	4.91	0.71	0.72	[0.04 to 1.39]	4.8%
Maganty et al., 2018	20	4.10	0.90	20	2.40	1.00	1.75	[1.01 to 2.49]	4.4%
Nazmi et al., 2025	45	62.71	7.10	45	56.31	8.00	0.84	[0.41 to 1.27]	6.4%
Nowak et al., 2020	81	2.24	0.84	81	2.12	0.83	0.14	[−0.17 to 0.45]	7.2%
Bartholomew et al., 2000	29	64.10	7.70	25	57.80	8.80	0.75	[0.20 to 1.31]	5.5%
Tan et al., 2022	178	12.50	1.20	196	12.30	1.50	0.15	[−0.06 to 0.35]	7.8%
Random effects model	**932**			**926**			**0.50**	**[0.24 to 0.76]**	**100.0%**
Prediction interval								**[−0.42 to 1.43]**	

Heterogeneity: I^2 = 80.7%, τ^2 = 0.18, P < .001

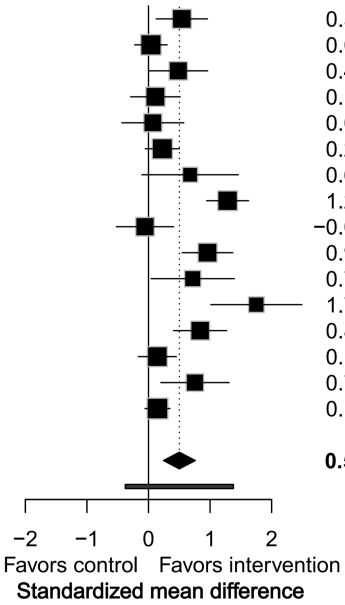

XSL•FO
RenderX

Figure 2. PRISMA flow diagram of study identification, screening, and selection for randomized controlled trials evaluating digital serious games in public health education. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

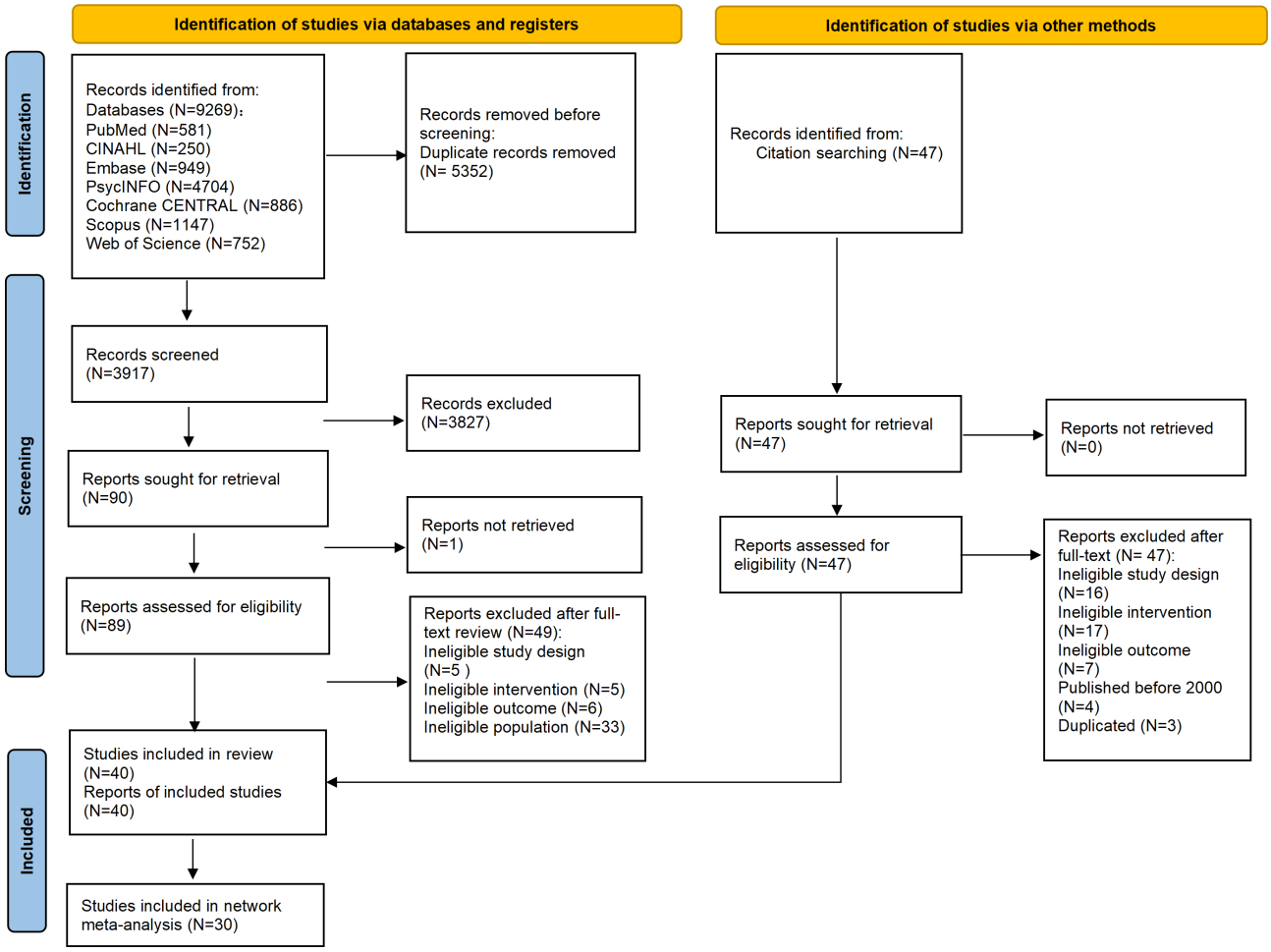

Study Characteristics

A total of 40 RCTs published between 2000 and 2025 were included (Table 1). Research output increased markedly after 2020, with 5 (12.5%) studies published before 2010 [47-50,76], 10 (22.5%) between 2010 and 2019 [51-57,77-79], and 25 (62.5%) since 2020 [58-75,80-86]. Studies were conducted across 19 countries, most frequently in the United States (n=13, 32.5%) [47-50,54,56,60,65,74,76,78,79,83], followed by China (n=4, 10%) [64,68,70,85], Iran (n=4, 10%) [58,62,67,73], and the Netherlands (n=3, 7.5%) [53,63,77]. All studies adopted a randomized controlled design, including 10 (25%) pilot trials and 5 (12.5%) cluster trials.

The design and characteristics of digital serious games are summarized in Multimedia Appendix 6. Interventions evolved with technological development. Early studies used computer-offline serious games (n=7, 17.5%) [47,49-52,80,85] and video-based serious games (n=6, 15%) [48,61,66,76,78,79]. Later studies adopted computer or web-based online games (n=8, 20%) [55,56,65,71,73,75,77,83] and mobile-app games (n=13, 32.5%) [54,57-59,62-64,67,68,72,81,82,84]. Since 2020, immersive formats such as virtual-reality serious games (n=3, 7.5%) [69,74,86] and augmented-reality serious games (n=2, 5%) [60,71] have become more common, and robot-assisted serious games (n=1, 2.5%) [53] were included from 2017. Three games, *Watch, Discover, Think, and Act* [47,50], *Re-Mission* [76,78], and *Dental Detective* [51,80], were each evaluated in 2 trials due to updated versions or applications in different populations. All iterations were therefore included in the synthesis.

A total of 8764 participants were included (intervention =4374; control =4390). Twelve studies (30%) targeted children, 11 (27.5%) adolescents, 11 (27.5%) adults, and 6 (15%) mixed populations (eg, children and adolescents or adolescents and adults). Thirteen studies (32.5%) included more women than men, 9 (n=9, 22.5%) were gender-balanced, and 18 (45%) included a higher proportion of men.

The most frequent educational topics were CD education (n=8, 20%), cancer education (n=6, 15%), sexual and reproductive health education (n=6, 15%), and nutrition and healthy lifestyle education (n=6, 15%). Other topics included vaccination and infectious-disease prevention (n=4, 10%), oral health education (n=4, 10%), medication and antimicrobial-resistance education (n=3, 7.5%), and psychological and developmental health education (n=3, 7.5%).

Among the 40 studies, 35 (87.5%) reported knowledge outcomes, of which 28 (80%) showed significant improvement and 7 (20%) found no between-group difference [51,54,60,72,75,77,80]. A total of 22 (55%) studies assessed attitude outcomes, with 15 (68.2%) showing positive changes and 7 (31.8%) showing no improvement [51,54,55,57,63,75,80].

XSL•FO
RenderX

A total of 30 (75%) studies were included in the Bayesian network meta-analysis [47-75,80], including those reporting knowledge outcomes (n=27, 67.5%) [47-55,57-62,64-73,75,80] and attitude outcomes (n=16, 40%) [47,50,54-57,63-67,69,70,73-75].

XSL•FO
RenderX

Table . Characteristics of included randomized controlled trials evaluating digital serious games for public health education.

Study and outcome	Study design	Country	Population	Age (y), mean (SD)	Female, %	Patient status	Intervention	Comparator	Sample size (IG[a]/CG[b])	Health topic	Duration	Baseline	Key findings
Aljafari et al [51], 2017													
Knowledge	RCT[c]	United Kingdom	Children	6.5 (1.55)	45	Patient	CF[d]	FE[e]	55/54	OH[f]	Single session	IG 56.0 (9.6); CG 53.4 (10.6)	Knowledge increased in both groups, with no significant differences observed between them
Attitude	NA[g]	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	No significant changes were observed in perceived susceptibility or perceived importance
Aljafari et al [80], 2022													
Knowledge	RCT	Jordan	Children	6.5 (0.5)	48	Nonpatient	CF	NI[h]	144/139	OH	Single session	IG 56.5 (5.9); CG 57.2 (4.9)	Knowledge improved in the intervention group, but no significant difference was found between groups at follow-up

XSL•FO
RenderX

Study and outcome	Study design	Country	Population	Age (y), mean (SD)	Female, %	Patient status	Intervention	Comparator	Sample size (IG[a]/CG[b])	Health topic	Duration	Baseline	Key findings
Attitude	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	Attitudes, intentions, and self-efficacy improved in the IG and were sustained for 6 mo, with no significant change in HIV testing uptake
Bartholomew et al [50], 2000													
Knowledge	Pilot RCT	United States	Children	IG 9.8 (2.1); CG 9.5 (1.9)	35	Patient	CF	NI	29/25	CD[i]	Multiple sessions	IG 76.1 (12.8); CG 78.4 (14.5)	Knowledge and self-management behaviors significantly improved in the IG compared with controls, with greater gains observed in younger children
Attitude	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	The intervention reduced emergency visits and hospitalizations among participants with moderate-to-severe asthma

XSL•FO
RenderX

Study and outcome	Study design	Country	Population	Age (y), mean (SD)	Female, %	Patient status	Intervention	Comparator	Sample size (IG[a]/CG[b])	Health topic	Duration	Baseline	Key findings
Beale et al [76], 2007													
Knowledge	RCT	United States	Adolescents	IG 15.79 (2.62); CG 16.0 (2.89)	32	Patient	VI[j]	AC[k]	191/169	Cancer	Multiple sessions	IG 59.31 (16.9); CG 59.73 (15.6)	Knowledge increased in both groups, with the Re-Mission game showing greater gains than the CG
Beaujean et al [77], 2016													
Knowledge	Cluster RCT	Netherlands	Children	NA	50	Nonpatient	CO[l]	NI	199/372	ID[m]	Single session	NA	Knowledge improved across all groups, with no significant differences between the game, leaflet, and CGs
Bloomfield et al [82], 2025													
Knowledge	Cluster RCT	Australia	Adolescents	NA	50	Nonpatient	AP[n]	FE	442/346	ID	Single session	NA	Knowledge gains in the IG were maintained at 3- and 6-mo follow-up, while the CG remained stable
Boomer et al [83], 2024													

XSL•FO
RenderX

Study and outcome	Study design	Country	Population	Age (y), mean (SD)	Female, %	Patient status	Intervention	Comparator	Sample size (IG[a]/CG[b])	Health topic	Duration	Baseline	Key findings
Knowledge	RCT	United States	Adolescents	15.4 (1.2)	48	Nonpatient	CO	OT[o]	145/142	SRH[p]	Multiple sessions	NA	Knowledge increased in the IG compared with controls and was sustained over 12 months
Carcioppolo et al [65], 2022													
Knowledge	RCT	United States	Adults	47.3 (17.6)	53	Nonpatient	CO	NI	100/102	Cancer	Single session	IG 6.75 (2.6); CG 6.13 (2.2)	The game improved melanoma identification compared with the CG
Attitude	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	IG 5.89 (1.0); CG 5.65 (1.1)	The game enhanced self-efficacy and promoted more positive prevention attitudes
Espinosa-Curiel et al [66], 2022													
Knowledge	Pilot RCT	Mexico	Children	9.9 (0.8)	59	Nonpatient	VI	FE	15/12	NHL[q]	Multiple sessions	IG 5.7 (2.4); CG 4.08 (1.6)	The game improved children's knowledge of physical activity compared with controls

XSL•FO
RenderX

Study and outcome	Study design	Country	Population	Age (y), mean (SD)	Female, %	Patient status	Intervention	Comparator	Sample size (IG[a]/CG[b])	Health topic	Duration	Baseline	Key findings
Attitude	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	IG 3.07 (0.8); CG 2.3 (1.4)	Attitudes toward sexual health improved, particularly among boys and younger adolescents
Fadda et al [57], 2017													
Knowledge	RCT	Italy	Adults	34.2 (4.7)	94.60	Nonpatient	AP	NI	48/44	Vaccination	Multiple sessions	10.3 (2.1)	Knowledge scores increased significantly in the gamified app group compared with the CG
Attitude	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	3.4 (0.6)	Vaccination intention and decision confidence improved in the gamified app group, with no significant change in vaccination attitude or recommendation intention
Fiellin et al [79], 2017													

XSL•FO
RenderX

Study and outcome	Study design	Country	Population	Age (y), mean (SD)	Female, %	Patient status	Intervention	Comparator	Sample size (IG[a]/CG[b])	Health topic	Duration	Baseline	Key findings
Knowledge	RCT	United States	Adolescent	12.5 (1.1)	50	Nonpatient	VI	AC	166/165	SRH	Multiple sessions	NA	Knowledge improved in the IG, including greater awareness of menstrual hygiene management and contraceptive methods
Froome et al [59], 2020													
Knowledge	Pilot RCT	Canada	Children	9.0 (0.8)	38	Nonpatient	AP	NI	39/34	NHL	Multiple sessions	IG 10.3 (2.9); CG 10.2 (3.1)	Knowledge scores increased significantly in the IG compared with controls, with notable gains in fruits, protein, and whole grains knowledge
Ghadam et al [67], 2023													
Knowledge	RCT	Iran	Adolescents	14.2 (0.7)	100	Nonpatient	AP	TE[r]	80/80	NHL	Multiple sessions	IG 47.9 (9.9); CG 46.6 (7.8)	Knowledge scores increased significantly in the IG compared with controls

XSL•FO
RenderX

Study and outcome	Study design	Country	Population	Age (y), mean (SD)	Female, %	Patient status	Intervention	Comparator	Sample size (IG[a]/CG[b])	Health topic	Duration	Baseline	Key findings
Attitude	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	IG 60.0 (6.7); CG 57.6 (12.9)	Attitude scores were higher in the IG compared with the CG
Goodman et al [60], 2024													
Knowledge	Pilot RCT	United States	Children	10.9 (2.9)	50	Patient	AR[s]	VE[t]	26/27	CD	Single session	IG 9.4 (3.1); CG 9.1 (3.0)	Knowledge increased in both groups, with no significant difference between game and video education
Haruna et al [52], 2018													
Knowledge	Cluster RCT	Malaysia	Children and adolescents	13.6 (1.1)	50	Nonpatient	CF	TE	40/40	SRH	Multiple sessions	IG 62.5 (12.3); CG 55.4 (11.8)	Knowledge scores were significantly higher in the game-based group compared with traditional education
Attitude	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	Attitude, motivation, and engagement were higher in the game-based learning group than in the CG
Henkemans et al [53], 2017													

XSL•FO
RenderX

Study and outcome	Study design	Country	Population	Age (y), mean (SD)	Female, %	Patient status	Intervention	Comparator	Sample size (IG[a]/CG[b])	Health topic	Duration	Baseline	Key findings
Knowledge	Pilot RCT	Netherlands	Children	10.5 (2.0)	50	Patient	RB[u]	TA	14/13	CD	Multiple sessions	IG 17.2 (4.1); CG 13.5 (3.8)	Knowledge scores increased significantly in the robot-assisted game group compared with controls
Attitude	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	Autonomy, competence, relatedness, motivation, and engagement were higher in the robot-assisted group than in the CG
Huang et al [68], 2025													
Knowledge	RCT	China	Adults	63.5 (11.7)	24.5	Patient	AP	FE	51/51	CD	Multiple sessions	IG 6.14 (1.8); CG 7.06 (1.8)	Knowledge scores improved significantly in the IG compared with the CG at post test
Huang et al [81], 2024													

XSL•FO
RenderX

Study and outcome	Study design	Country	Population	Age (y), mean (SD)	Female, %	Patient status	Intervention	Comparator	Sample size (IG[a]/CG[b])	Health topic	Duration	Baseline	Key findings
Knowledge	RCT	Singapore	Adults	36.7 (10.4)	57	Nonpatient	AP	NI	90/90	MAR[v]	Single session	IG 8.5 (1.6); CG 8.3 (1.7)	Knowledge scores were higher in the app-based serious game group compared with the CG
Joubert et al [55], 2016													
Knowledge	Pilot RCT	France	Adolescents	15.5 (1.5)	48	Patient	CO	TA[w]	37/34	CD	Multiple sessions	IG 64.8 (12.3); CG 63.7 (11.9)	Diabetes-related knowledge increased significantly in the IG compared with controls
Attitude	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	IG 58.6 (14.1); CG 59.4 (13.7)	Positive trends were observed in self-management attitudes and behaviors, though less pronounced than knowledge gains
Kato et al [48], 2008													
Knowledge	RCT	United States	Adolescents and adults	NA	32.3	Patient	VI	NI	164/148	Cancer	Multiple sessions	IG 0.59 (0.2); CG 0.63 (0.2)	Cancer-related knowledge improved significantly in the IG compared with the CG
Khalil et al [78], 2016													

XSL•FO
RenderX

Study and outcome	Study design	Country	Population	Age (y), mean (SD)	Female, %	Patient status	Intervention	Comparator	Sample size (IG[a]/CG[b])	Health topic	Duration	Baseline	Key findings
Attitude	RCT	United States	Adolescents	NA	50	Nonpatient	VI	EM[x]	166/50	Cancer	Single session	2.92 (1.2)	Participants in the IG reported higher perceived susceptibility to cancer than controls, and the effect was sustained for 20 d
Koohmareh et al [62], 2021													
Knowledge	RCT	Iran	Adults	IG 52.6 (8.4); CG 53.3 (7.9)	56.7	Patient	AP	EM	30/30	NHL	Multiple sessions	IG 7.27 (2.6); CG 7.47 (2.4)	Knowledge increased significantly in the IG compared with the CG
Koniou et al [69], 2025													
Knowledge	RCT	Greece	Adults	20.8 (2.5)	70.8	Nonpatient	VR[y]	NI	51/51	PDH[z]	Single session	IG 36.92 (3.2); CG 33.14 (3.1)	Knowledge scores improved significantly in the VR group compared with the CG
Attitude	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	IG 77.18 (6.2); CG 71.29 (6.0)	Attitudes toward autism improved significantly in the VR group compared with controls
Kumar et al [49], 2004													

XSL•FO
RenderX

Study and outcome	Study design	Country	Population	Age (y), mean (SD)	Female, %	Patient status	Intervention	Comparator	Sample size (IG[a]/CG[b])	Health topic	Duration	Baseline	Key findings
Knowledge	Pilot RCT	United States	Children and adolescents	13.6 (2.5)	45	Patient	CF	NI	19/21	CD	Multiple sessions	NA	Knowledge scores increased significantly in IG compared with CG
Kumar RS et al [61], 2022													
Knowledge	RCT	India	Adolescents	13.62 (1.37)	42	Nonpatient	VI	TE	48/42	OH	Single session	IG 7.29 (1.7); CG 7.30 (1.3)	Knowledge increased significantly in the IG compared with traditional education
Liu et al [70], 2024													
Knowledge	Pilot RCT	China	Children	7.09 (0.9)	38.9	Nonpatient	AR	WN[aa]	18/18	Cancer	Multiple sessions	IG 6.17 (2.3); CG 6.56 (2.1)	Knowledge improved significantly in the IG compared with controls
Attitude	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	IG 4.98 (0.6); CG 4.95 (0.6)	Attitudes improved significantly in the IG compared with controls
Mack et al [71], 2020													
Knowledge	Cluster RCT	Germany	Children	10.5 (0.5)	50	Nonpatient	CO	TE	172/144	NHL	Multiple sessions	IG 20.77 (4.2); CG 20.28 (3.8)	Knowledge improved significantly in the IG compared with controls after the intervention
Maddison et al [72], 2022													

XSL•FO
RenderX

Study and outcome	Study design	Country	Population	Age (y), mean (SD)	Female, %	Patient status	Intervention	Comparator	Sample size (IG[a]/CG[b])	Health topic	Duration	Baseline	Key findings
Knowledge	Pilot RCT	New Zealand	Children and adolescents	11.2 (1.8)	70	Patient	AP	WN	15/8	CD	Multiple sessions	IG 4.6 (6.6); CG 5.8 (2.6)	No significant group differences were found; the intervention was perceived as engaging, but effects were not sustained
Maganty et al [56], 2018													
Attitude	Pilot RCT	United States	Adults	59.1 (15.5)	NA	Nonpatient	CO	NI	20/20	Cancer	Single session	IG 2.3 (1.1); CG 2.2 (1.0)	Confidence in melanoma recognition improved significantly in the game group compared with no intervention
Nazmi et al [73], 2025													
Knowledge	RCT	Iran	Adolescents	12.77 (0.5)	100	Nonpatient	CO	NI	45/45	PDH	Multiple sessions	IG 5.11 (1.3); CG 5.22 (1.0)	Knowledge scores increased significantly in the IG compared with controls
Attitude	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	IG 61.9 (7.3); CG 57.6 (6.3)	Practice scores improved significantly in the IG compared with the CG
Nowak et al [74], 2020													

XSL•FO
RenderX

Study and outcome	Study design	Country	Population	Age (y), mean (SD)	Female, %	Patient status	Intervention	Comparator	Sample size (IG[a]/CG[b])	Health topic	Duration	Baseline	Key findings
Attitude	RCT	United States	Adults	NA	NA	Nonpatient	VR	NI	81/81	Vaccination	Single session	NA	The VR intervention enhanced vaccine confidence, beliefs about community immunity, and vaccination intention compared with the CG
Pouls et al [63], 2022													
Attitude	RCT	Netherlands	Adults	61.2 (11.3)	73	Patient	AP	TA	110/111	MAR	Multiple sessions	IG 5.0 (5.1); CG 5.8 (4.3)	No significant differences were found in medication beliefs or adherence between intervention and CGs
Raj et al [84], 2025													
Knowledge	RCT	India	Adolescents	16.7(0.1)	100	Nonpatient	AP	NI	769/928	SRH	Multiple sessions	NA	Knowledge improved significantly in the IG compared with the CG
Shegog et al [47], 2001													

XSL•FO
RenderX

Study and outcome	Study design	Country	Population	Age (y), mean (SD)	Female, %	Patient status	Intervention	Comparator	Sample size (IG[a]/CG[b])	Health topic	Duration	Baseline	Key findings
Knowledge	RCT	United States	Children	10.9 (1.1)	34	Patient	CF	NI	38/33	CD	Single session	IG 18.6 (6.5); CG 15.7 (5.8)	The IG showed greater improvement in asthma self-management knowledge compared with the CG
Attitude	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	IG 53.4 (9.7); CG 51.6 (9.9)	The IG demonstrated higher self-efficacy and more positive attributions regarding asthma management than the CG
Tan et al [75], 2022													
Knowledge	RCT	Singapore	Adults	35.7 (9.6)	59.6	Nonpatient	CO	WN	178/196	Vaccination	Single session	IG 10.0 (2.3); CG 10.0 (2.3)	Knowledge scores increased in both groups, with no significant difference between them
Attitude	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	IG 12.5 (1.2); CG 12.3 (1.5)	Attitude scores increased in both groups, with no significant difference between them
Tang et al [64], 2022													

XSL•FO
RenderX

Study and outcome	Study design	Country	Population	Age (y), mean (SD)	Female, %	Patient status	Intervention	Comparator	Sample size (IG[a]/CG[b])	Health topic	Duration	Baseline	Key findings
Knowledge	RCT	China	Adolescents	13.5 (0.6)	50	Nonpatient	AP	TE	50/46	SRH	Single session	IG 11.2 (3.4); CG 10.9 (3.2)	Knowledge scores increased significantly in the IG compared with controls
Attitude	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	The intervention group showed more positive attitudes toward HIV prevention than the CG
Vandeweerdt et al [86], 2022													
Attitude	RCT	Belgium	Adults	NA	NA	Nonpatient	VR	WN	208/208	Vaccination	Single session	NA	Knowledge and perceived health awareness improved significantly in the IG compared with the CG
Wang et al [85], 2025													
Knowledge	Cluster RCT	China	Children	NA	48	Nonpatient	CF	NI	40/39	NHL	Multiple sessions	IG 8.93 (2.4); CG 8.38 (2.1)	The intervention enhanced vaccination intention and collective responsibility compared with text-based education
Whiteley et al [54], 2018													

XSL•FO
RenderX

Study and outcome	Study design	Country	Population	Age (y), mean (SD)	Female, %	Patient status	Intervention	Comparator	Sample size (IG[a]/CG[b])	Health topic	Duration	Baseline	Key findings
Knowledge	RCT	United States	Adolescents and adults	22.4 (2.5)	21.3	Patient	AP	NI	32/29	SRH	Multiple sessions	IG 2.44 (1.2); CG 3.00 (1.0)	The IG showed improvements in HIV knowledge, but the difference was not significant
Attitude	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	IG 17.94 (4.0); CG 18.21 (5.5)	The IG reported slightly higher self-efficacy for ART[ab] use compared with the CG, but the difference was not significant
Zolfaghari et al [58], 2021													

XSL•FO
RenderX

Study and outcome	Study design	Country	Population	Age (y), mean (SD)	Female, %	Patient status	Intervention	Comparator	Sample size (IG[a]/CG[b])	Health topic	Duration	Baseline	Key findings
Knowledge	RCT	Iran	Adolescents and adults	36.4 (4.7)	100	Nonpatient	AP	WN	46/47	OH	Multiple sessions	IG 11.3 (1.9); CG 10.5 (2.1)	The IG had higher knowledge scores than the CG at posttest and 3-month follow-up

[a]IG: intervention group.
[b]CG: control group.
[c]RCT: randomized controlled trial.
[d]CF: computer offline serious games delivered via PC, tablet, or DVD.
[e]FE: face-to-face education.
[f]OH: oral health.
[g]NA: not available.
[h]NI: no intervention.
[i]CD: chronic diseases.
[j]VI: video-based serious games.
[k]AC: active control (non–health-related video game).
[l]CO: computer or web-based online serious games.
[m]ID: infectious diseases.
[n]AP: mobile app serious games.
[o]OT: other serious games.
[p]SRH: sexual and reproductive health.
[q]NHL: nutrition and healthy lifestyle.
[r]TE: traditional classroom or lecture-based education.
[s]AR: augmented reality serious games.
[t]VE: educational videos without gamified elements.
[u]RB: robot-assisted serious games.
[v]MAR: medication and antimicrobial resistance.
[w]TA: treatment as usual.
[x]EM: educational materials (leaflets, booklets, or pamphlets).
[y]VR: virtual reality serious games.
[z]PDH: psychological and developmental health.
[aa]WN: web-based nongame education.
[ab]ART: antiretroviral therapy.

Risk-of-Bias Results

Among the 40 included RCTs, methodological quality was generally moderate to high (Multimedia Appendix 7). For the 35 individually randomized trials, low risk of bias was most frequently observed in the randomization process (25/35, 71.4%), deviations from intended interventions (31/35, 88.6%), and measurement of outcomes (33/35, 94.3%). “Some concerns” were mainly identified in the selection of reported results (14/35, 40%) and overall bias judgment (24/35, 68.6%), mainly due to the absence of preregistered protocols or incomplete reporting of secondary outcomes. One trial was rated as high risk of bias in the domain of deviations from intended interventions because participants and facilitators were not blinded during gameplay [51], and another trial was judged as high risk for the same reason, with substantial researcher involvement potentially influencing participant responses [68]. No other study was rated as high risk in any domain. Among the 5 cluster-randomized trials, methodological quality was similarly high; all studies were rated as low risk for the randomization process and missing outcome data, with only minor concerns regarding the selection of the reported result. Taken together, 31.4% (11/35) of studies were judged as low risk, 68.6% (24/35) as having some concerns, and none as high risk.

XSL•FO
RenderX

Results of the Meta-Analyses

Across the 40 included studies, 27 reported data on knowledge outcomes, and 16 on attitude outcomes [47-75], with 21 trials contributing to both outcome categories. Compared with controls, digital serious games significantly improved public health knowledge (SMD=0.66; 95% CI 0.32 - 0.99; P<.001; I^2=89.1%) and showed a moderate positive effect on health attitudes (SMD=0.50; 95% CI 0.27 - 0.76; P<.001; I^2=80.7%) (Figure 1). Considerable heterogeneity was observed across studies (knowledge: Q=239.23; P<.001; attitude: Q=77.75; P<.001). Funnel plots showed mild asymmetry for both outcomes (Multimedia Appendix 8). Egger's regression test indicated potential small-study effects for knowledge (P=.006) but not for attitude (P=.05).

The overall certainty of evidence, assessed using the GRADE framework, was moderate for both knowledge and attitude outcomes. Although substantial heterogeneity was present, the direction of effects remained consistent across studies, and subgroup analyses explained much of the observed variation by intervention duration, population type, and health topic. Minor methodological concerns related to randomization, allocation concealment, and small-study effects contributed to downgrading from high to moderate certainty. Indirectness and imprecision did not materially affect the certainty ratings, as all included trials directly addressed the review question and yielded precise pooled estimates (Multimedia Appendix 9).

Subgroup and Moderator Analyses

Subgroup analyses were undertaken to explore potential sources of heterogeneity across intervention duration, study region, patient status, health topic, publication year, population type, and sex. Across both knowledge and attitude outcomes, multisession interventions consistently yielded larger effects than single-session exposure (knowledge: χ^2_1=4.04; P=.04; attitude: χ^2_1=4.97; P=.03), indicating that repeated game participation reinforced learning and attitude internalization. Effect sizes were also greater among nonpatient populations than among patients (knowledge: χ^2_1=7.13; P=.008; attitude: χ^2_1=9.97; P=.002), suggesting that individuals without disease burden may be more receptive to health information. Considerable variation was observed across health topics (knowledge: χ^2_6=120.32; P<.001; attitude: χ^2_6=176.14; P<.001), with cancer- and CD-focused games achieving the highest impact, whereas effects were smaller for vaccination and oral health education. Regional differences were modest but favored studies conducted in Asia ($\chi^2$4=10.18; P=.04). Publication year, age group, and sex composition did not consistently influence effect estimates (Multimedia Appendix 10).

Bayesian Network Meta-Analysis

The knowledge network comprised 14 interventions, including 7 types of digital serious games and 7 traditional or nongame comparators, forming 26 direct comparisons and 5 closed loops (Figure 3A). Between-study heterogeneity in the network meta-analysis was low (τ=2.75; 95% CrI 1.58 - 4.69; τ^2=7.57; network I^2=8%). Digital serious games produced the greatest improvements in knowledge outcomes (Figure 4). Mobile app–based games showed significantly higher effects than traditional education (mean difference 5.46; 95% CrI 2.00 - 9.39), treatment as usual (4.87; 95% CrI 1.06 - 9.37), and no intervention (2.82; 95% CrI 0.09 - 5.79). Computer-offline and web-based serious games also achieved superior gains compared with traditional education (4.87; 95% CrI 0.65 - 7.95 and 4.12; 95% CrI 0.79 - 6.05, respectively), whereas robot-assisted, virtual reality, and video-based games showed weaker comparative effects. Bayesian ranking analyses indicated that mobile app–based, computer-offline, and web-based serious games consistently ranked highest for improving knowledge outcomes (Figure 5A). Prediction intervals were calculated to reflect the expected range of treatment effects in future studies (Multimedia Appendix 11). No significant inconsistency between direct and indirect evidence was detected in node-splitting analyses (all P>.05; Multimedia Appendix 12), and the consistency and unrelated mean effects models showed nearly identical model fit (Deviance Information Criterion; DIC=109.899 vs 109.894; ΔDIC=0.005). Sensitivity analyses using alternative prior distributions produced nearly identical SUCRA values and treatment rankings, indicating robust results (Multimedia Appendix 13).

The attitude network comprised 11 interventions, including 6 types of digital serious games and 5 traditional or nongame comparators, forming 16 direct comparisons and 3 closed loops (Figure 3B). Between-study heterogeneity in the network meta-analysis was low (τ=3.19; 95% CrI 1.42 - 6.53; τ^2=10.20; network I^2=3%). Digital serious games produced greater improvements in health attitudes compared with traditional or nongame education (Figure 4). Computer-offline, web-based, and virtual reality serious games showed the largest improvements in attitude outcomes compared with traditional education (13.28; 95% CrI 3.30 - 22.92; 11.30; 95% CrI 1.53 - 21.00; and 11.61; 95% CrI 1.33 - 21.74, respectively), whereas video-based, face-to-face, and no-intervention conditions showed weaker or inconsistent effects. Bayesian ranking analyses indicated that computer-offline, web-based, and virtual reality serious games ranked highest for improving health attitudes (Figure 5B). Prediction intervals were calculated to reflect the expected range of treatment effects in future studies (Multimedia Appendix 11). No significant inconsistency between direct and indirect evidence was detected in node-splitting analyses (all P>.05; Multimedia Appendix 12), and the consistency and unrelated mean effects models showed similar model fit (DIC=58.47 vs 60.12; ΔDIC=1.65). Sensitivity analyses using alternative prior distributions produced nearly identical SUCRA values and treatment rankings, indicating that the results were robust (Multimedia Appendix 13).

XSL•FO
RenderX

Figure 3. Network structures of digital serious game interventions included in the network meta-analysis for knowledge and attitude outcomes. Panels A and B illustrate the network structures of digital serious game interventions for knowledge and attitude outcomes, respectively. Each node represents an intervention, and each connecting line indicates a direct comparison between interventions in the included randomized controlled trials. The size of each node is proportional to the number of participants receiving that intervention, and the thickness of the connecting lines reflects the number of direct comparisons. AP: mobile app serious games; AR: augmented reality serious games; CF: computer offline serious games delivered via PC, tablet, or DVD; CO: computer or web-based online serious game; EM: educational materials (leaflets, booklets, or pamphlets); FE: face-to-face education; NI: no intervention; RB: robot-assisted serious games; TA: treatment as usual; TE: traditional classroom or lecture-based education; VE: educational videos without gamified elements; VI: video-based serious games; VR: virtual reality serious games; WN: web-based nongame education.

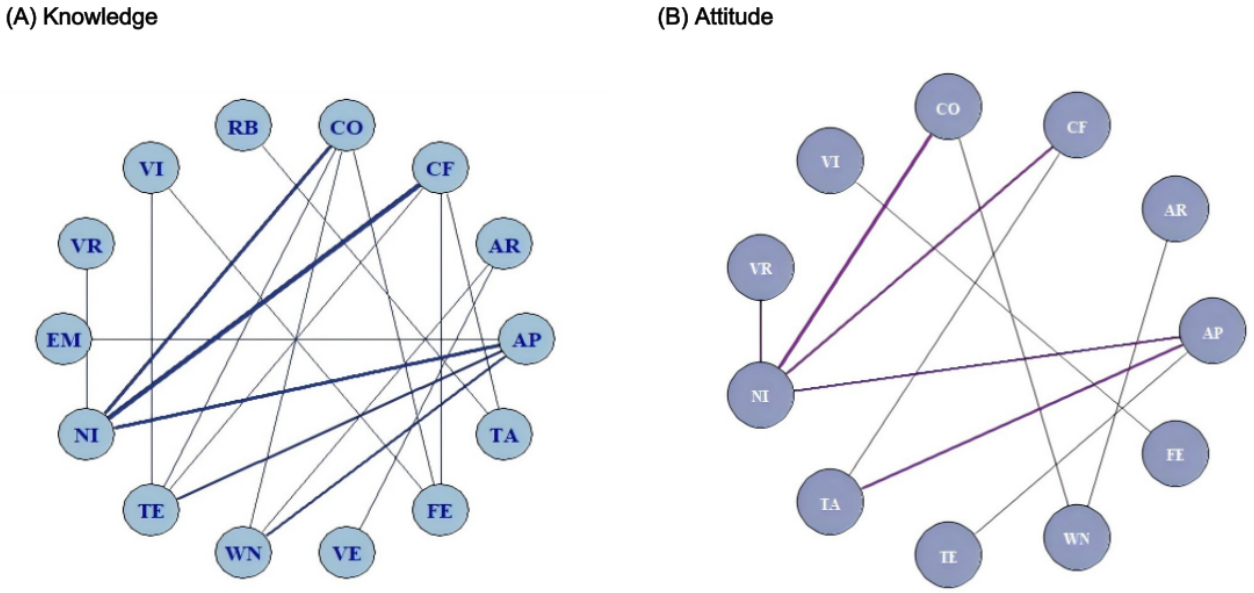

XSL•FO
RenderX

Figure 4. Heatmap of pairwise comparisons from the network meta-analysis of digital serious game interventions for knowledge and attitude outcomes. The heatmap summarizes pairwise mean differences with corresponding 95% credible intervals between interventions. Values below the diagonal represent knowledge outcomes, whereas values above the diagonal represent attitude outcomes. Positive values favor the column-defining intervention, whereas negative values favor the row-defining intervention. Darker shading represents larger absolute mean differences. Statistically significant comparisons are indicated by *(P<.05). AP: mobile app serious games; AR: augmented reality serious games; CF: computer offline serious games delivered via PC, tablet, or DVD; CO: computer or web-based online serious game; EM: educational materials (leaflets, booklets, or pamphlets); FE: face-to-face education; NI: no intervention; RB: robot-assisted serious games; TA: treatment as usual; TE: traditional classroom or lecture-based education; VE: educational videos without gamified elements; VI: video-based serious games; VR: virtual reality serious games; WN: web-based nongame education.

Knowledge | Attitude

	NI	AP	AR	CF	CO	EM	FE	RB	TA	TE	VE	VI	VR	WN
NI		1.12 (–3.32,6.12)	2.68 (–7.95,13.50)	4.37 (–0.89, 9.70)	2.37 (–1.62,6.77)		31.88 (–157.88, 195.96)		1.52 (–4.55, 8.05)	–8.93 (–17.49, 0.08)		32.76 (–156.52, 196.91)	2.68 (–2.24, 7.93)	2.14 (–5.83, 10.44)
AP	-2.82* (-5.79, -0.09)		1.56 (–10.39, 13.44)	3.26 (–3.34, 9.41)	1.25 (-5.12, 7.55)		30.73 (–158.98, 195.16)		0.36 (–4.81, 5.67)	–10.05* (–17.42, –2.53)		31.58 (–157.38, 195.73)	1.59 (–5.42, 8.56)	1.05 (–8.44, 10.59)
AR	-2.66 (-10.03, 4.31)	0.16 (-6.90, 7.17)		1.73 (–10.44, 13.74)	–0.31 (–10.17, 9.85)		28.73 (–161.22, 193.38		–1.15 (–13.78, 11.48)	–11.59 (–25.56, 2.52)		29.86 (–160.43, 194.29)	0.02 (–11.73, 11.99)	–0.48 (–7.58, 6.64)
CF	-2.23 (-5.39, 0.45)	0.60 (-3.35, 4.27)	0.43 (–7.36, 7.92)		–1.99 (–8.60, 5.04)		27.55 (–162.33, 191.58)		–2.83 (–9.39, 4.18)	–13.28* (–22.92, –3.30)		28.20 (–160.59, 192.33)	–1.70 (–8.83, 5.91)	-2.21 (–11.83, 7.75)
CO	-1.50 (-4.26, 1.27)	1.33 (–1.99, 4.89)	1.14 (–5.93, 8.46)	0.72 (–2.71, 4.66)			29.31 (–160.69, 193.28)		–0.86 (–8.62, 6.85)	–11.30* (-21.00, –1.53)		30.08 (–159.14, 194.44)	0.31 (–6.22, 6.80)	–0.21 (–7.31, 6.77)
EM	2.19 (-4.50, 8.59)	4.99 (–0.79,10.79)	4.80 (–4.27,13.85)	4.38 (–2.40,11.62)	3.67 (–3.15, 10.31)									
FE	-1.39 (-5.68, 2.75)	1.46 (–3.20, 6.24)	1.28 (–6.74, 9.32)	0.84 (–3.22, 5.30)	0.11 (–4.15, 4.15)	–3.56 (–10.98, 3.97)			–30.33 (–194.87, 159.73)	–40.89 (–204.66, 148.06)		0.79 (-6.28, 7.80)	–28.96 (–193.11, 160.20)	–29.54 (–193.48, 160.72)
RB	- 4.73 (-15.54, 5.80)	–1.88 (–13.02, 9.01)	–2.06 (–15.12,10.69)	–2.43 (–12.66,7.92)	–3.32 (–14.17,7.55)	–6.8 (–19.47, 5.53)	–3.31 (–14.59, 7.75)							
TA	-1.05 (-9.67, 7.32)	1.79 (–7.18, 10.56)	1.58 (–9.60, 12.64)	1.19 (–6.75, 9.22)	0.45 (–8.47, 9.16)	–3.18 (–13.95, 7.38)	0.35 (–8.81, 9.35)	3.66 (–2.80, 10.09)		–10.43* (–19.47, –1.32)		31.03 (-158.53, 195.59)	1.21 (–7.00, 9.23)	0.66 (–9.67, 11.00)
TE	2.63 (-0.86, 6.44)	5.46 * (2.00, 9.39)	5.32 (–2.11, 13.21)	4.87* (1.06, 9.37)	4.12* (0.65, 7.95)	0.47 (–6.20, 7.59)	4.00 (–5.15, 12.99)	7.35 (–3.48, 18.66)	3.70 (2.00, 9.39)			41.62 (-147.60, 205.98)	11.61* (1.33, 21.74)	11.10 (–1.16, 23.03)
VE	-2.35 (-11.85, 6.83)	0.45 (–8.79, 9.59)	0.27 (–5.65, 6.33)	–0.15 (–9.82, 9.78)	–0.87 (–10.31,8.44)	–4.54 (–15.42, 6.29)	–1.00 (–10.95, 8.92)	2.31 (–11.76, 16.65)	–1.28 (–13.89, 11.38)	–5.00 (–15.00, 4.34)				
VI	-0.65 (-5.72, 4.57)	2.16 (–3.08, 7.73)	2.03 (–6.41, 10.65)	1.57 (–3.44, 7.20)	0.86 (–4.28, 5.97)	–2.81 (–10.64, 5.32)	0.73 (–3.85, 5.50)	4.07 (–7.50, 15.84)	0.40 (–9.22, 10.41)	–3.27 (–8.16, 1.28)	1.72 (–8.49, 12.28)		–29.65 (–194.28, 159.29)	–30.39 (–194.59, 159.90)
VR	-3.78 (-9.62, 2.09)	–0.95 (–7.26, 5.73)	–1.11 (–10.31, 8.32)	–1.56 (–7.77, 5.33)	–2.29 (–8.73, 4.19)	–5.98 (–14.47, 2.98)	–2.38 (–9.52, 4.93)	–0.92 (–11.16, 13.21)	–2.70 (–12.99, 7.64)	–6.41 (–13.54, 0.47)	–1.49 (–12.32, 9.89)	–3.16 (–11.03, 4.66)		–0.50 (–10.18, 9.06)
WN	-1.74 (-5.90, 2.27)	1.11 (–2.45, 4.71)	0.96 (–4.98, 6.88)	0.50 (–4.13, 5.48)	–0.23 (–4.34, 3.72)	–3.86 (–10.72, 2.96)	–0.34 (–5.84, 5.10)	2.98 (–8.20, 14.46)	–0.66 (–9.94, 8.77)	–4.36 (–9.30, 0.15)	0.63 (–7.68, 9.13)	–1.09 (–7.22, 4.91)	2.09 (–5.22, 8.94)	

XSL•FO
RenderX

Figure 5. Bayesian ranking probabilities of digital serious game interventions for improving knowledge and attitude outcomes. Panels A and B present the Bayesian ranking probabilities for knowledge and attitude outcomes, respectively. Each stacked bar represents the probability that a given intervention ranks at each possible position among all interventions included in the network meta-analysis. Darker shading indicates higher ranking probabilities. Higher rankings indicate greater effectiveness. AP: mobile app serious games; AR: augmented reality serious games; CF: computer offline serious games delivered via PC, tablet, or DVD; CO: computer or web-based online serious game; EM: educational materials (leaflets, booklets, or pamphlets); FE: face-to-face education; NI: no intervention; RB: robot-assisted serious games; TA: treatment as usual; TE: traditional classroom or lecture-based education; VE: educational videos without gamified elements; VI: video-based serious games; VR: virtual reality serious games; WN: web-based nongame education.

XSL•FO
RenderX

(A) Knowledge

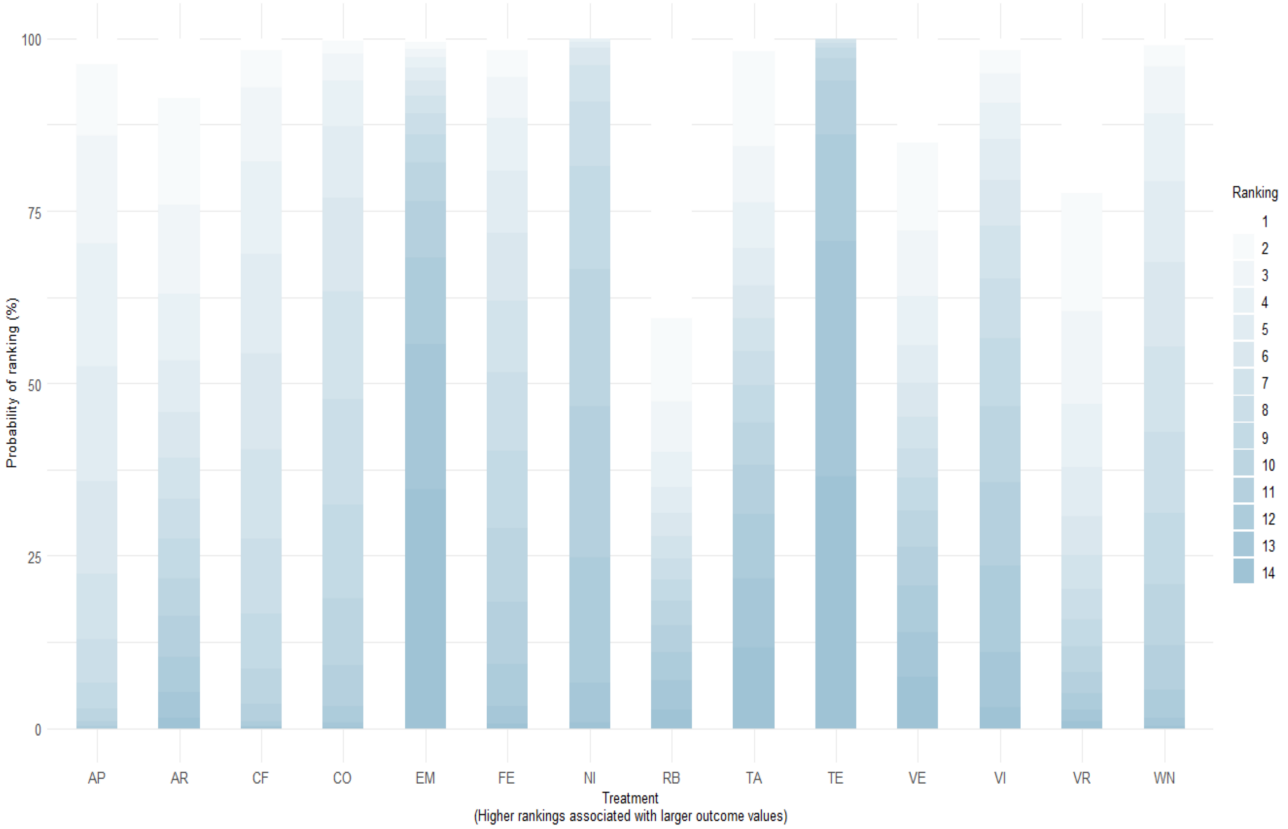

(B) Attitude

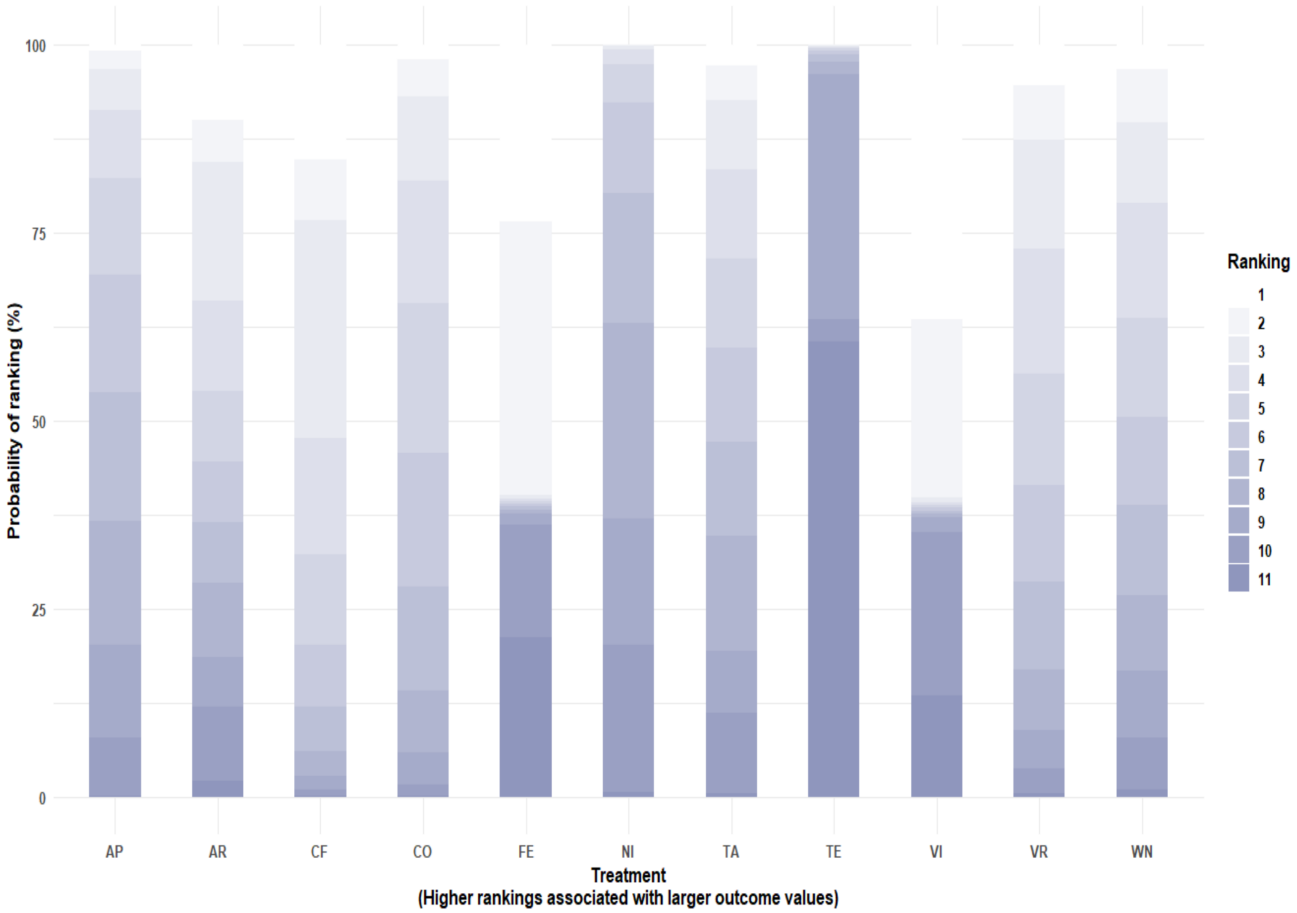

XSL•FO
RenderX

Discussion

Principal Findings

In this systematic review and Bayesian network meta-analysis of 40 RCTs involving 8764 participants, digital serious games were associated with improvements in public health knowledge and attitudes compared with traditional or noninteractive education. Greater effects were observed with multisession interventions. Subgroup analyses indicated stronger responses among adolescents and nonpatient populations, particularly in studies conducted in Asia and in interventions addressing psychological or developmental health topics. Network meta-analysis further demonstrated differences across delivery formats: mobile app–based games ranked highest for knowledge outcomes, whereas computer-offline and web-based formats showed greater relative effectiveness for attitude change, while video-based and traditional education formats consistently ranked lower. By integrating pairwise and network meta-analysis within a cross-disease framework, this study enables comparative evaluation across formats and population contexts, addressing limitations of prior reviews restricted to single conditions or pairwise comparisons [87-89].

Despite substantial heterogeneity observed in the pairwise meta-analysis, sensitivity analyses confirmed the stability of pooled estimates, indicating that the variability primarily reflects contextual and population-level differences rather than methodological bias. Differences across health topics highlight the influence of content relevance and narrative structure in shaping learning outcomes. Interventions targeting psychological or developmental health often incorporate self-management scenarios and emotionally salient components that may enhance perceived relevance and retention [90]. The advantage of repeated exposure is consistent with reinforcement and memory consolidation processes, whereas single-session interventions may offer insufficient opportunities for feedback and integration. Larger improvements among nonpatient populations may relate to lower baseline knowledge and reduced ceiling effects [91]. The heterogeneity observed, therefore, reflects meaningful contextual differentiation rather than instability of effect estimates.

These findings should also be interpreted considering methodological factors, including the risk of bias and the certainty of evidence. Risk-of-bias assessment using the revised Cochrane risk-of-bias tool for randomized trial tool indicated that several studies had methodological limitations that may have influenced effect estimates. The overall certainty of evidence assessed using the GRADE framework was rated as moderate for knowledge outcomes and low-to-moderate for attitudes. In addition, while CIs represent the average effect across studies, prediction intervals reflect the potential variation in effects across different implementation settings, indicating that intervention effects may vary depending on population characteristics and context.

Beyond these methodological considerations, the findings also suggest several interpretive mechanisms underlying the educational effects of serious games. Although the network meta-analysis suggested relatively consistent comparative effects across formats, the pooled estimates and ranking probabilities should still be interpreted with caution because the included studies varied in populations, health topics, and intervention characteristics. From a structural perspective, the network findings further suggest that the educational impact of serious games may operate along complementary cognitive and affective pathways. Interventions incorporating adaptive feedback, progressive challenge, and opportunities for repeated engagement are more likely to activate sustained cognitive processing, thereby facilitating the consolidation and integration of information [92]. In contrast, formats characterized by low interactivity and fixed content delivery may limit learner control and cognitive activation [93]. With respect to attitudinal outcomes, narrative-driven and role-playing designs appear more conducive to attitude change through mechanisms of perspective-taking and emotional engagement [94], while immersive simulations may intensify affective involvement through first-person experiential framing. Together, these findings indicate that knowledge gains are primarily supported by cognitive reinforcement processes, whereas attitudinal change is more closely linked to emotional immersion and social resonance. The integration of both pathways may, therefore, strengthen the overall educational impact of serious games in public health contexts.

Importantly, the relative balance between cognitive structure and experiential immersion is not only a theoretical distinction but also a practical one. Designs that prioritize deep affective engagement often require greater technological resources and infrastructural investment, whereas cognitively structured, feedback-oriented formats may be more feasible for large-scale dissemination [95]. This interplay between experiential intensity and implementation feasibility becomes particularly salient in population-level public health education [96].

Consistent with this structural tension, the network analysis highlights trade-offs between experiential depth and scalability. Virtual-reality and robot-assisted formats may achieve high experiential fidelity yet face barriers related to cost and accessibility, while mobile and web-based interventions enable broader reach, albeit sometimes with reduced experiential richness [97]. These findings suggest that innovation should not focus solely on increasing technical sophistication but rather on developing adaptive architectures capable of preserving feedback, learner autonomy, and emotional resonance across diverse delivery contexts [98,99]. Achieving equilibrium between structural fidelity and affective relevance may be essential for translating short-term knowledge improvements into sustained behavioral and attitudinal change in population health education.

Implications for Practice and Research

Evidence from this review suggests that digital serious games can extend the reach of public health education in settings where conventional programs face limitations in coverage or engagement. The consistent advantages of mobile and web-based formats over resource-intensive technologies indicate that scalability depends more on accessibility and design efficiency than on technical sophistication [22,100]. In practice, prioritizing adaptive, feedback-driven mobile platforms may

XSL•FO
RenderX

yield greater population impact than investing in high-fidelity but low-reach systems, such as virtual reality or robotics [21,101]. At the same time, stronger effects observed among adolescents and women highlight both the potential for targeted implementation and the need to address equity gaps among patients and older adults who demonstrate lower engagement. Inclusive design, tailored difficulty adjustment, and integration within existing community or school-based programs may help reduce digital exclusion and sustain long-term participation [102,103].

Beyond implementation considerations, the current evidence base remains fragmented, with substantial variation in populations, intervention formats, and outcome measures. Most randomized trials are small and exploratory, often limited to school-aged or student samples. Future research should prioritize adequately powered trials involving adult and older populations and incorporate medium- and long-term follow-up to determine whether gains in knowledge and attitudes translate into sustained behavioral change [30].

A structured implementation framework is also needed to define minimal effective exposure, establish evaluation benchmarks, and clarify ethical standards for educational gaming [104,105]. Standardization represents a critical next step. Although thematic diversity in public health education is expected, a unified evaluation framework should be developed to assess usability, implementation quality, and core design attributes—such as interactivity, feedback mechanisms, immersion, and accessibility—using validated instruments [97,98]. Establishing shared data infrastructures that systematically document intervention characteristics, engagement metrics, and outcome measures would further enhance comparability and enable cumulative synthesis across studies. Strengthening these methodological and infrastructural foundations will be essential for advancing serious game research from isolated trials toward a coherent and reproducible scientific field.

Limitations

The interpretation of this synthesis should take into account the methodological and conceptual variability among the included trials, which likely contributed to the substantial heterogeneity observed in the pairwise meta-analysis. Health education topics, measurement instruments, and feedback structures differed widely, complicating direct comparison of effect sizes across studies [30]. Although subgroup analyses identified several sources of heterogeneity, the diversity in outcome definitions and analytical strategies inevitably limited the precision of pooled estimates.

In several recent studies, particularly those published after 2020, technological advancements have led serious games, such as *Food Adventure Quest* and *Amoo*, to evolve beyond stand-alone formats, increasingly integrating complementary educational components, such as video segments and classroom instruction [62,85]. Although this convergence complicates the identification of the game's independent effects, it reflects the growing trend toward multimodal approaches in health education. It indicates that serious games are becoming embedded components of broader digital learning ecosystems that combine interactive games, web-based modules, and instructor-led components.

Reporting transparency also varied across studies. Protocol preregistration and complete reporting of secondary outcomes were often absent, resulting in the overall methodological quality being rated as moderate, despite generally low risks of bias in randomization, intervention delivery, and outcome measurement [106]. These limitations indicate that the main threat to internal validity stems from incomplete documentation rather than flawed trial conduct, underscoring the need for preregistered protocols and comprehensive reporting standards in future evaluations of serious games.

Conclusions

This systematic review and Bayesian network meta-analysis of RCTs provides a comprehensive evaluation of the educational impact of digital serious games in public health. Unlike earlier reviews that focused on individual interventions or specific health topics, this study compares multiple digital serious game formats within a unified analytical framework. Across 40 trials, mobile-, computer-, and web-based formats generally produced the greatest improvements in knowledge outcomes, while computer-, web-, and virtual reality–based formats showed stronger effects for attitude change. Multisession interventions sustained learning and attitudinal change more effectively than single-session exposure, highlighting the importance of reinforcement and continued engagement. The overall certainty of evidence was moderate, reflecting methodological heterogeneity across trials. These findings contribute comparative evidence to the field of digital health education and offer practical guidance for selecting scalable serious game interventions in real-world public health programs. Strengthening implementation strategies, standardizing outcome evaluation, and extending trials to underrepresented adult and older populations are important next steps.

Acknowledgments

We would like to thank Professor Wendy Moyle for funding support of this study. In addition, we thank all corresponding authors of the included studies for providing additional data and clarifications. The first author, DH, acknowledges the China Scholarship Council for scholarship support during this research. No generative AI tools were used to generate scientific content, analyses, results, or interpretations.

Funding

This study was funded by consultancy research funds awarded to Professor Wendy Moyle, Griffith University.

XSL•FO
RenderX

Data Availability

All extracted data and analytical codes used in this study are available from the corresponding author upon reasonable request. Unpublished data obtained from individual researchers will be shared only with their explicit permission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full electronic search strategies for all databases used in this systematic review, including PubMed, CINAHL, Embase, APA PsycINFO, Cochrane CENTRAL, Scopus, and Web of Science.
[DOCX File, 31 KB - jmir_v28i1e89281_app1.docx]

Multimedia Appendix 2
Detailed eligibility criteria for study inclusion and exclusion in this systematic review, including population, intervention, comparator, outcomes, study design, language, and publication type.
[DOCX File, 13 KB - jmir_v28i1e89281_app2.docx]

Multimedia Appendix 3
Subgroup classification and coding schema used for subgroup analyses in this systematic review.
[DOCX File, 13 KB - jmir_v28i1e89281_app3.docx]

Multimedia Appendix 4
R scripts used to conduct the pairwise meta-analysis and Bayesian network meta-analysis, including data preparation, model fitting, subgroup analyses, and sensitivity analyses.
[DOCX File, 16 KB - jmir_v28i1e89281_app4.docx]

Multimedia Appendix 5
Detailed characteristics of the included studies, including study design, sample size, population characteristics, intervention features, and outcome measures.
[DOCX File, 40 KB - jmir_v28i1e89281_app5.docx]

Multimedia Appendix 6
Design features and characteristics of the digital serious games included in this review, including developers, interaction mechanisms, and educational purposes.
[DOCX File, 27 KB - jmir_v28i1e89281_app6.docx]

Multimedia Appendix 7
Risk of bias assessment for included trials, including individually randomized trials and cluster randomized trials.
[DOCX File, 142 KB - jmir_v28i1e89281_app7.docx]

Multimedia Appendix 8
Funnel plots assessing potential publication bias for knowledge and attitude outcomes across the included studies.
[DOCX File, 28 KB - jmir_v28i1e89281_app8.docx]

Multimedia Appendix 9
Grading of Recommendations Assessment, Development and Evaluation summarizing the certainty of evidence for knowledge and attitude outcomes across the included studies.
[DOCX File, 14 KB - jmir_v28i1e89281_app9.docx]

Multimedia Appendix 10
Forest plots of subgroup analyses for knowledge and attitude outcomes across 7 moderators, including intervention duration, study region, and population characteristics.
[DOCX File, 784 KB - jmir_v28i1e89281_app10.docx]

Multimedia Appendix 11

XSL•FO
RenderX

Forest plots of the Bayesian network meta-analysis for knowledge and attitude outcomes showing pooled mean differences with 95% credible intervals.
[PDF File, 26 KB - jmir_v28i1e89281_app11.pdf]

Multimedia Appendix 12
Node-splitting analyses assessing inconsistency between direct and indirect evidence in the network meta-analysis for knowledge and attitude outcomes.
[PDF File, 40 KB - jmir_v28i1e89281_app12.pdf]

Multimedia Appendix 13
Comparison of surface under the cumulative ranking curve values and treatment rankings under uniform and half-normal prior distributions for knowledge and attitude outcomes.
[DOCX File, 19 KB - jmir_v28i1e89281_app13.docx]

Checklist 1
PRISMA reporting checklists used in this review, including the PRISMA 2020 checklist, PRISMA 2020 expanded checklist, PRISMA 2020 for abstracts checklist, and PRISMA-S checklist.
[PDF File, 845 KB - jmir_v28i1e89281_app14.pdf]

References

1. A global health strategy for 2025–2028: advancing equity and resilience in a turbulent world (fourteenth general programme of work). : World Health Organization; 2025 URL: https://iris.who.int/server/api/core/bitstreams/46cc7cac-e35e-451b-808e-1f0e4ad5f68c/content [accessed 2026-04-01]
2. Fazel M, Hoagwood K, Stephan S, Ford T. Mental health interventions in schools 1: mental health interventions in schools in high-income countries. Lancet Psychiatry 2014 Oct;1(5):377-387. [doi: 10.1016/S2215-0366(14)70312-8] [Medline: 26114092]
3. Dodd S, Widnall E, Russell AE, et al. School-based peer education interventions to improve health: a global systematic review of effectiveness. BMC Public Health 2022 Dec 2;22(1):2247. [doi: 10.1186/s12889-022-14688-3] [Medline: 36461024]
4. Baccolini V, Rosso A, Di Paolo C, et al. What is the prevalence of low health literacy in European Union member states? A systematic review and meta-analysis. J Gen Intern Med 2021 Mar;36(3):753-761. [doi: 10.1007/s11606-020-06407-8] [Medline: 33403622]
5. China achieves 31.87% health literacy rate in 2024. The State Council of the People's Republic of China. 2025. URL: https://english.www.gov.cn/archive/statistics/202501/10/content_WS6780e424c6d0868f4e8eeae6.html [accessed 2025-09-29]
6. Yu YS, Altares A, Leib A, et al. Sociodemographic disparities in health literacy among American adults: a national survey study. Prev Med Rep 2025 Sep;57:103179. [doi: 10.1016/j.pmedr.2025.103179] [Medline: 40727277]
7. O'Neill G. Global health literacy: Delaware and beyond!. Dela J Public Health 2025 Apr;11(1):24-25. [doi: 10.32481/djph.2025.04.06] [Medline: 40331175]
8. van der Gaag M, Heijmans M, Spoiala C, Rademakers J. The importance of health literacy for self-management: a scoping review of reviews. Chronic Illn 2022 Jun;18(2):234-254. [doi: 10.1177/17423953211035472] [Medline: 34402309]
9. Magnani JW, Mujahid MS, Aronow HD, et al. Health literacy and cardiovascular disease: fundamental relevance to primary and secondary prevention: a scientific statement from the American Heart Association. Circulation 2018 Jul 10;138(2):e48-e74. [doi: 10.1161/CIR.0000000000000579] [Medline: 29866648]
10. Vasileia E, Koulierakis G, Fouskas T, Liarigkovinou A. Health literacy and acceptance of COVID-19 preventive measures and vaccination in the European Union: a scoping review. Health Lit Res Pract 2025 Jan;9(1):e46-e55. [doi: 10.3928/24748307-20250219-01] [Medline: 40064011]
11. St Leger L. Schools, health literacy and public health: possibilities and challenges. Health Promot Int 2001 Jun;16(2):197-205. [doi: 10.1093/heapro/16.2.197] [Medline: 11356758]
12. Zhao T, Deng W, Yang Y, Jin Y, Xu F. Enhancing student satisfaction through open collaborative practical teaching reforms in public health education: a comparative study. Front Public Health 2025;13:1546962. [doi: 10.3389/fpubh.2025.1546962] [Medline: 40109428]
13. Education: from COVID-19 school closures to recovery. UNESCO. URL: https://www.unesco.org/en/covid-19/education-response [accessed 2025-09-29]
14. Dettori M, Castiglia P. COVID-19 and digital health: evolution, perspectives and opportunities. Int J Environ Res Public Health 2022 Jul 12;19(14):8519. [doi: 10.3390/ijerph19148519] [Medline: 35886366]
15. Getachew E, Adebeta T, Muzazu SGY, et al. Digital health in the era of COVID-19: reshaping the next generation of healthcare. Front Public Health 2023;11:942703. [doi: 10.3389/fpubh.2023.942703] [Medline: 36875401]

XSL•FO
RenderX

16. Kermavnar T, Visch VT, Desmet PMA. Games in times of a pandemic: structured overview of COVID-19 serious games. JMIR Serious Games 2023 Mar 7;11:e41766. [doi: 10.2196/41766] [Medline: 36634265]
17. DeSmet A, Thompson D, Baranowski T, Palmeira A, Verloigne M, De Bourdeaudhuij I. Is participatory design associated with the effectiveness of serious digital games for healthy lifestyle promotion? A meta-analysis. J Med Internet Res 2016 Apr 29;18(4):e94. [doi: 10.2196/jmir.4444] [Medline: 27129447]
18. Xu M, Luo Y, Zhang Y, Xia R, Qian H, Zou X. Game-based learning in medical education. Front Public Health 2023;11:1113682. [doi: 10.3389/fpubh.2023.1113682] [Medline: 36935696]
19. Alencar NES, Pinto MAO, Leite NT, Silva C. Serious games for sex education of adolescents and youth: integrative literature review. Cien Saude Colet 2022 Aug;27(8):3129-3138. [doi: 10.1590/1413-81232022278.00632022] [Medline: 35894324]
20. Abd-Alrazaq A, Abuelezz I, Hassan A, et al. Artificial intelligence-driven serious games in health care: scoping review. JMIR Serious Games 2022 Nov 29;10(4):e39840. [doi: 10.2196/39840] [Medline: 36445731]
21. Checa D, Miguel-Alonso I, Bustillo A. Immersive virtual-reality computer-assembly serious game to enhance autonomous learning. Virtual Real 2021 Dec 23:1-18. [doi: 10.1007/s10055-021-00607-1] [Medline: 34961808]
22. Buffel C, van Aalst J, Bangels AM, et al. A web-based serious game for health to reduce perioperative anxiety and pain in children (CliniPup): pilot randomized controlled trial. JMIR Serious Games 2019 Jun 1;7(2):e12431. [doi: 10.2196/12431] [Medline: 31199324]
23. Dietvorst E, Aukes MA, Legerstee JS, et al. A smartphone serious game for adolescents (Grow It! app): development, feasibility, and acceptance study. JMIR Form Res 2022 Mar 3;6(3):e29832. [doi: 10.2196/29832] [Medline: 35238795]
24. Aster A, Scheithauer S, Middeke AC, et al. Use of a serious game to teach infectious disease management in medical school: effectiveness and transfer to a clinical examination. Front Med 2022;9:863764. [doi: 10.3389/fmed.2022.863764]
25. Thomas TH, Sivakumar V, Babichenko D, Grieve VLB, Klem ML. Mapping behavioral health serious game interventions for adults with chronic illness: scoping review. JMIR Serious Games 2020 Jul 30;8(3):e18687. [doi: 10.2196/18687] [Medline: 32729836]
26. Craig S, Stark P, Wilson CB, Carter G, Clarke S, Mitchell G. Evaluation of a dementia awareness game for undergraduate nursing students in Northern Ireland: a pre-/post-test study. BMC Nurs 2023 May 22;22(1):177. [doi: 10.1186/s12912-023-01345-2] [Medline: 37217937]
27. Wang Y, Wang Z, Liu G, et al. Application of serious games in health care: scoping review and bibliometric analysis. Front Public Health 2022;10:896974. [doi: 10.3389/fpubh.2022.896974]
28. Haoran G, Bazakidi E, Zary N. Serious games in health professions education: review of trends and learning efficacy. Yearb Med Inform 2019 Aug;28(1):240-248. [doi: 10.1055/s-0039-1677904] [Medline: 31022747]
29. Suppan M, Stuby L, Harbarth S, et al. Nationwide deployment of a serious game designed to improve COVID-19 infection prevention practices in Switzerland: prospective web-based study. JMIR Serious Games 2021 Nov 25;9(4):e33003. [doi: 10.2196/33003] [Medline: 34635472]
30. Sharifzadeh N, Kharrazi H, Nazari E, et al. Health education serious games targeting health care providers, patients, and public health users: scoping review. JMIR Serious Games 2020 Mar 5;8(1):e13459. [doi: 10.2196/13459] [Medline: 32134391]
31. Nørlev J, Sondrup K, Derosche C, Hejlesen O, Hangaard S. Game mechanisms in serious games that teach children with type 1 diabetes how to self-manage: a systematic scoping review. J Diabetes Sci Technol 2022 Sep;16(5):1253-1269. [doi: 10.1177/19322968211018236] [Medline: 34024156]
32. Koutsiana E, Ladakis I, Fotopoulos D, Chytas A, Kilintzis V, Chouvarda I. Serious gaming technology in upper extremity rehabilitation: scoping review. JMIR Serious Games 2020 Dec 11;8(4):e19071. [doi: 10.2196/19071] [Medline: 33306029]
33. Ohannessian R, Yaghobian S, Verger P, Vanhems P. A systematic review of serious video games used for vaccination. Vaccine (Auckl) 2016 Aug;34(38):4478-4483. [doi: 10.1016/j.vaccine.2016.07.048]
34. Tavares N. The use and impact of game-based learning on the learning experience and knowledge retention of nursing undergraduate students: a systematic literature review. Nurse Educ Today 2022 Oct;117:105484. [doi: 10.1016/j.nedt.2022.105484] [Medline: 35917707]
35. Moschonis G, Siopis G, Jung J, et al. Effectiveness, reach, uptake, and feasibility of digital health interventions for adults with type 2 diabetes: a systematic review and meta-analysis of randomised controlled trials. Lancet Digit Health 2023 Mar;5(3):e125-e143. [doi: 10.1016/S2589-7500(22)00233-3] [Medline: 36828606]
36. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
37. Schardt C, Adams MB, Owens T, Keitz S, Fontelo P. Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC Med Inform Decis Mak 2007 Jun 15;7(1):16. [doi: 10.1186/1472-6947-7-16] [Medline: 17573961]
38. Higgins JP, Thomas J, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions, 2nd edition: Wiley; 2019. URL: https://onlinelibrary.wiley.com/doi/book/10.1002/9781119536604 [accessed 2026-04-01]
39. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]

XSL•FO
RenderX

40. Furukawa TA, Barbui C, Cipriani A, Brambilla P, Watanabe N. Imputing missing standard deviations in meta-analyses can provide accurate results. J Clin Epidemiol 2006 Jan;59(1):7-10. [doi: 10.1016/j.jclinepi.2005.06.006] [Medline: 16360555]
41. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 6;327(7414):557-560. [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
42. Efthimiou O, Rücker G, Schwarzer G, Higgins JPT, Egger M, Salanti G. Network meta-analysis of rare events using the Mantel-Haenszel method. Stat Med 2019 Jul 20;38(16):2992-3012. [doi: 10.1002/sim.8158] [Medline: 30997687]
43. Neupane B, Richer D, Bonner AJ, Kibret T, Beyene J. Network meta-analysis using R: a review of currently available automated packages. PLoS ONE 2014;9(12):e115065. [doi: 10.1371/journal.pone.0115065] [Medline: 25541687]
44. Higgins JPT, Jackson D, Barrett JK, Lu G, Ades AE, White IR. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. Res Synth Methods 2012 Jun;3(2):98-110. [doi: 10.1002/jrsm.1044] [Medline: 26062084]
45. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011 Apr;64(4):383-394. [doi: 10.1016/j.jclinepi.2010.04.026] [Medline: 21195583]
46. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. J Clin Epidemiol 2011 Dec;64(12):1294-1302. [doi: 10.1016/j.jclinepi.2011.03.017] [Medline: 21803546]
47. Shegog R, Bartholomew LK, Parcel GS, Sockrider MM, Mâsse L, Abramson SL. Impact of a computer-assisted education program on factors related to asthma self-management behavior. J Am Med Inform Assoc 2001;8(1):49-61. [doi: 10.1136/jamia.2001.0080049] [Medline: 11141512]
48. Kato PM, Cole SW, Bradlyn AS, Pollock BH. A video game improves behavioral outcomes in adolescents and young adults with cancer: a randomized trial. Pediatrics 2008 Aug;122(2):e305-e317. [doi: 10.1542/peds.2007-3134] [Medline: 18676516]
49. Kumar VS, Wentzell KJ, Mikkelsen T, Pentland A, Laffel LM. The DAILY (Daily Automated Intensive Log for Youth) trial: a wireless, portable system to improve adherence and glycemic control in youth with diabetes. Diabetes Technol Ther 2004 Aug;6(4):445-453. [doi: 10.1089/1520915041705893] [Medline: 15320998]
50. Bartholomew LK, Gold RS, Parcel GS, et al. Watch, discover, think, and act: evaluation of computer-assisted instruction to improve asthma self-management in inner-city children. Patient Educ Couns 2000 Feb;39(2-3):269-280. [doi: 10.1016/s0738-3991(99)00046-4] [Medline: 11040726]
51. Aljafari A, Gallagher JE, Hosey MT. Can oral health education be delivered to high-caries-risk children and their parents using a computer game? - a randomised controlled trial. Int J Paediatr Dent 2017 Nov;27(6):476-485. [doi: 10.1111/ipd.12286] [Medline: 28052471]
52. Haruna H, Hu X, Chu SKW, Mellecker RR, Gabriel G, Ndekao PS. Improving sexual health education programs for adolescent students through game-based learning and gamification. Int J Environ Res Public Health 2018 Sep 17;15(9):2027. [doi: 10.3390/ijerph15092027] [Medline: 30227642]
53. Henkemans OAB, Bierman BPB, Janssen J, et al. Design and evaluation of a personal robot playing a self-management education game with children with diabetes type 1. Int J Hum Comput Stud 2017 Oct;106:63-76. [doi: 10.1016/j.ijhcs.2017.06.001]
54. Whiteley L, Brown LK, Mena L, Craker L, Arnold T. Enhancing health among youth living with HIV using an iPhone game. AIDS Care 2018;30(sup4):21-33. [doi: 10.1080/09540121.2018.1503224] [Medline: 30626196]
55. Joubert M, Armand C, Morera J, Tokayeva L, Guillaume A, Reznik Y. Impact of a serious videogame designed for flexible insulin therapy on the knowledge and behaviors of children with type 1 diabetes: the LUDIDIAB pilot study. Diabetes Technol Ther 2016 Feb;18(2):52-58. [doi: 10.1089/dia.2015.0227] [Medline: 26440963]
56. Maganty N, Ilyas M, Zhang N, Sharma A. Online, game-based education for melanoma recognition: a pilot study. Patient Educ Couns 2018 Apr;101(4):738-742. [doi: 10.1016/j.pec.2017.11.003] [Medline: 29137837]
57. Fadda M, Galimberti E, Fiordelli M, Romanò L, Zanetti A, Schulz PJ. Effectiveness of a smartphone app to increase parents' knowledge and empowerment in the MMR vaccination decision: a randomized controlled trial. Hum Vaccin Immunother 2017 Nov 2;13(11):2512-2521. [doi: 10.1080/21645515.2017.1360456] [Medline: 29125783]
58. Zolfaghari M, Shirmohammadi M, Shahhosseini H, Mokhtaran M, Mohebbi SZ. Development and evaluation of a gamified smart phone mobile health application for oral health promotion in early childhood: a randomized controlled trial. BMC Oral Health 2021 Jan 7;21(1):18. [doi: 10.1186/s12903-020-01374-2] [Medline: 33413304]
59. Froome HM, Townson C, Rhodes S, et al. The effectiveness of the foodbot factory mobile serious game on increasing nutrition knowledge in children. Nutrients 2020 Nov 6;12(11):3413. [doi: 10.3390/nu12113413] [Medline: 33172094]
60. Goodman K, Arriaga RI, Korman R, et al. Pediatric emergency department-based asthma education tools and parent/child asthma knowledge. Allergy Asthma Clin Immunol 2024 Mar 25;20(1):24. [doi: 10.1186/s13223-024-00884-w] [Medline: 38528606]
61. K RSK, Deshpande AP, Ankola AV, et al. Effectiveness of a visual interactive game on oral hygiene knowledge, practices, and clinical parameters among adolescents: a randomized controlled trial. Children (Basel) 2022 Nov 26;9(12):1828. [doi: 10.3390/children9121828] [Medline: 36553272]

XSL•FO
RenderX

62. Koohmareh Z, Karandish M, Hadianfard AM. Effect of implementing a mobile game on improving dietary information in diabetic patients. Med J Islam Repub Iran 2021;35:68. [doi: 10.47176/mjiri.35.68] [Medline: 34277505]
63. Pouls BPH, Bekker CL, Gundogan F, et al. Gaming for adherence to medication using eHealth in rheumatoid arthritis (GAMER) study: a randomised controlled trial. RMD Open 2022 Nov;8(2):e002616. [doi: 10.1136/rmdopen-2022-002616] [Medline: 36410776]
64. Tang J, Zheng Y, Zhang D, et al. Evaluation of an AIDS educational mobile game (AIDS Fighter · Health Defense) for young students to improve AIDS-related knowledge, stigma, and attitude linked to high-risk behaviors in China: randomized controlled trial. JMIR Serious Games 2022 Jan 24;10(1):e32400. [doi: 10.2196/32400] [Medline: 34870603]
65. Carcioppolo N, Kim S, Sanchez M, et al. Evaluating a game-based randomized experiment to increase melanoma identification among adults living in the U.S. Soc Sci Med 2022 Oct;310:115281. [doi: 10.1016/j.socscimed.2022.115281] [Medline: 36001916]
66. Espinosa-Curiel IE, Pozas-Bogarin EE, Hernández-Arvizu M, et al. HelperFriend, a serious game for promoting healthy lifestyle behaviors in children: design and pilot study. JMIR Serious Games 2022 May 6;10(2):e33412. [doi: 10.2196/33412] [Medline: 35522474]
67. Ghadam OS, Sohrabi Z, Mehrabi M, et al. Evaluating the effect of digital game-based nutrition education on anemia indicators in adolescent girls: a randomized clinical trial. Food Sci Nutr 2023 Feb;11(2):863-871. [doi: 10.1002/fsn3.3120] [Medline: 36789052]
68. Huang X, Jiang Z, Dai Y, et al. Effect of gamification on improved adherence to inhaled medications in chronic obstructive pulmonary disease: randomized controlled trial. J Med Internet Res 2025 May 14;27:e65309. [doi: 10.2196/65309] [Medline: 40367505]
69. Koniou I, Douard E, Lanovaz MJ. Brief report: virtual reality to raise awareness about autism. J Autism Dev Disord 2025 Sep;55(9):3378-3386. [doi: 10.1007/s10803-023-06216-y] [Medline: 38079034]
70. Liu B, Wan X, Li X, Zhu D, Liu Z. An augmented reality serious game for children's optical science education: randomized controlled trial. JMIR Serious Games 2024 Feb 1;12:e47807-e47807. [doi: 10.2196/47807] [Medline: 38329284]
71. Mack I, Reiband N, Etges C, et al. The kids obesity prevention program: cluster randomized controlled trial to evaluate a serious game for the prevention and treatment of childhood obesity. J Med Internet Res 2020 Apr 24;22(4):e15725. [doi: 10.2196/15725] [Medline: 32329742]
72. Maddison R, Baghaei N, Calder A, et al. Feasibility of using games to improve healthy lifestyle knowledge in youth aged 9-16 years at risk for type 2 diabetes: pilot randomized controlled trial. JMIR Form Res 2022 Jun 17;6(6):e33089. [doi: 10.2196/33089] [Medline: 35713955]
73. Nazmi S, Omrani A, Behmanesh F, Nikbakht HA, Mehrabi M, Hamzehpour R. Improving pubertal health education for adolescent girls through a gamified learning approach. J Pediatr Adolesc Gynecol 2025 Jun;38(3):320-327. [doi: 10.1016/j.jpag.2025.01.003] [Medline: 39800112]
74. Nowak GJ, Evans NJ, Wojdynski BW, et al. Using immersive virtual reality to improve the beliefs and intentions of influenza vaccine avoidant 18-to-49-year-olds: considerations, effects, and lessons learned. Vaccine (Auckl) 2020 Jan 29;38(5):1225-1233. [doi: 10.1016/j.vaccine.2019.11.009] [Medline: 31806533]
75. Tan A, Koh E, Sankari U, Tang J, Goh CK, Tan NC. Effects of a serious game on knowledge, attitude and practice in vector control and dengue prevention among adults in primary care: a randomised controlled trial. Digit Health 2022;8:20552076221129099. [doi: 10.1177/20552076221129099] [Medline: 36185390]
76. Beale IL, Kato PM, Marin-Bowling VM, Guthrie N, Cole SW. Improvement in cancer-related knowledge following use of a psychoeducational video game for adolescents and young adults with cancer. J Adolesc Health 2007 Sep;41(3):263-270. [doi: 10.1016/j.jadohealth.2007.04.006] [Medline: 17707296]
77. Beaujean D, Gassner F, Wong A, Steenbergen JE, Crutzen R, Ruwaard D. Education on tick bite and Lyme borreliosis prevention, aimed at schoolchildren in the Netherlands: comparing the effects of an online educational video game versus a leaflet or no intervention. BMC Public Health 2016 Nov 16;16(1):1163. [doi: 10.1186/s12889-016-3811-5] [Medline: 27852247]
78. Khalil GE, Beale IL, Chen M, Prokhorov AV. A video game promoting cancer risk perception and information seeking behavior among young-adult college students: a randomized controlled trial. JMIR Serious Games 2016 Jul 28;4(2):e13. [doi: 10.2196/games.5793] [Medline: 27470927]
79. Fiellin LE, Hieftje KD, Pendergrass TM, et al. Video game intervention for sexual risk reduction in minority adolescents: randomized controlled trial. J Med Internet Res 2017 Sep 18;19(9):e314. [doi: 10.2196/jmir.8148] [Medline: 28923788]
80. Aljafari A, ElKarmi R, Nasser O, Atef A, Hosey MT. A video-game-based oral health intervention in primary schools-a randomised controlled trial. Dent J (Basel) 2022 May 19;10(5):90. [doi: 10.3390/dj10050090] [Medline: 35621543]
81. Huang Z, Ow JT, Tang WE, Chow A. An evidence-based serious game app for public education on antibiotic use and resistance: randomized controlled trial. JMIR Serious Games 2024 Sep 5;12:e59848. [doi: 10.2196/59848] [Medline: 39235853]
82. Bloomfield L, Boston J, Masek M, Andrew L, Barwood D, Devine A. Evaluating the efficacy of a serious game to deliver health education about invasive meningococcal disease: clustered randomized controlled equivalence trial. JMIR Serious Games 2025 Feb 11;13:e60755. [doi: 10.2196/60755] [Medline: 39932769]

XSL•FO
RenderX

83. Boomer TP, Larkin K, Duncan LR, Fernandes CSF, Fiellin LE. A serious video game targeting HIV testing and counseling: a randomized controlled trial. J Adolesc Health 2024 Feb;74(2):252-259. [doi: 10.1016/j.jadohealth.2023.08.016] [Medline: 37815773]
84. Raj A, Shankar L, Dixit A, et al. Enhancing reproductive health among adolescent girls in India: results of an individualized RCT to study the efficacy of the Go Nisha Go Mobile Game. Reprod Health 2025 Apr 7;22(1):48. [doi: 10.1186/s12978-025-02005-1] [Medline: 40197317]
85. Wang R, Yao J, Leong C, Moltchanova E, Hoermann S. Promoting learning about nutrition and healthy eating behaviors in Chinese children through an alternate reality game: a pilot study. Nutrients 2025 Mar 30;17(7):1219. [doi: 10.3390/nu17071219]
86. Vandeweerdt C, Luong T, Atchapero M, et al. Virtual reality reduces COVID-19 vaccine hesitancy in the wild: a randomized trial. Sci Rep 2022 Mar 17;12(1):4593. [doi: 10.1038/s41598-022-08120-4] [Medline: 35301359]
87. Charlier N, Zupancic N, Fieuws S, Denhaerynck K, Zaman B, Moons P. Serious games for improving knowledge and self-management in young people with chronic conditions: a systematic review and meta-analysis. J Am Med Inform Assoc 2016 Jan;23(1):230-239. [doi: 10.1093/jamia/ocv100] [Medline: 26186934]
88. DeSmet A, Shegog R, Van Ryckeghem D, Crombez G, De Bourdeaudhuij I. A systematic review and meta-analysis of interventions for sexual health promotion involving serious digital games. Games Health J 2015 Apr;4(2):78-90. [doi: 10.1089/g4h.2014.0110] [Medline: 26181801]
89. Lau HM, Smit JH, Fleming TM, Riper H. Serious games for mental health: are they accessible, feasible, and effective? A systematic review and meta-analysis. Front Psychiatry 2017;7:209. [doi: 10.3389/fpsyt.2016.00209] [Medline: 28149281]
90. Andrew L, Barwood D, Boston J, Masek M, Bloomfield L, Devine A. Serious games for health promotion in adolescents – a systematic scoping review. Educ Inf Technol 2023 May;28(5):5519-5550. [doi: 10.1007/s10639-022-11414-9]
91. Dudai Y, Karni A, Born J. The consolidation and transformation of memory. Neuron 2015 Oct 7;88(1):20-32. [doi: 10.1016/j.neuron.2015.09.004] [Medline: 26447570]
92. Mullins JK, Sabherwal R. Gamification: a cognitive-emotional view. J Bus Res 2020 Jan;106:304-314. [doi: 10.1016/j.jbusres.2018.09.023]
93. Maxim RI, Arnedo-Moreno J. Identifying key principles and commonalities in digital serious game design frameworks: scoping review. JMIR Serious Games 2025 Mar 5;13:e54075. [doi: 10.2196/54075] [Medline: 40053743]
94. Liao X, Noor NFM. Design components of serious game based on flow theories. Adv Educ Technol Psychol 2025;9(3):67-80. [doi: 10.23977/aetp.2025.090311]
95. Gómez-León MI. Serious games to support emotional regulation strategies in educational intervention programs with children and adolescents. Systematic review and meta-analysis. Heliyon 2025 Feb;11(4):e42712. [doi: 10.1016/j.heliyon.2025.e42712] [Medline: 40034297]
96. David OA, Magurean S, Tomoiagă C. Do improvements in therapeutic game-based skills transfer to real life improvements in children's emotion-regulation abilities and mental health? A pilot study that offers preliminary validity of the REThink in-game performance scoring. Front Psychiatry 2022;13:828481. [doi: 10.3389/fpsyt.2022.828481] [Medline: 35386525]
97. Kai L, Tan WH, Saari EM. Dimensions of interactive pervasive game design: systematic review. JMIR Serious Games 2023 Aug 22;11:e42878. [doi: 10.2196/42878] [Medline: 37463860]
98. Alexiou A, Schippers MC, Oshri I, Angelopoulos S. Narrative and aesthetics as antecedents of perceived learning in serious games. Inf Technol People 2022 Dec 19;35(8):142-161. [doi: 10.1108/ITP-08-2019-0435]
99. Martin F, Dennen VP, Bonk CJ. A synthesis of systematic review research on emerging learning environments and technologies. Educ Technol Res Dev 2020;68(4):1613-1633. [doi: 10.1007/s11423-020-09812-2] [Medline: 32837122]
100. Zhang X, Lai E. A web-based gaming approach to decrease HIV-related stigma: game development and mixed methods evaluation. JMIR Serious Games 2022 Dec 15;10(4):e37219. [doi: 10.2196/37219] [Medline: 36520508]
101. Bunt L, Greeff J, Taylor E. Enhancing serious game design: expert-reviewed, stakeholder-centered framework. JMIR Serious Games 2024 May 31;12:e48099. [doi: 10.2196/48099] [Medline: 38820585]
102. Aster A, Laupichler MC, Zimmer S, Raupach T. Game design elements of serious games in the education of medical and healthcare professions: a mixed-methods systematic review of underlying theories and teaching effectiveness. Adv Health Sci Educ Theory Pract 2024 Nov;29(5):1825-1848. [doi: 10.1007/s10459-024-10327-1] [Medline: 38563873]
103. Kwak M, Kim BJ, Chung JB. Serious game development for public health: participatory design approach to COVID-19 quarantine policy education. JMIR Serious Games 2024 Oct 15;12:e54968. [doi: 10.2196/54968] [Medline: 39405084]
104. Epstein DS, Zemski A, Enticott J, Barton C. Tabletop board game elements and gamification interventions for health behavior change: realist review and proposal of a game design framework. JMIR Serious Games 2021 Mar 31;9(1):e23302. [doi: 10.2196/23302] [Medline: 33787502]
105. Pistono A, dos Santos AMP, Baptista RJV, Mamede HS. Framework for adaptive serious games. Comp Applic In Engineering 2024 Jul;32(4):e22731. [doi: 10.1002/cae.22731]
106. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]

XSL•FO
RenderX

Abbreviations

CD: chronic diseases
CrI: credible interval
GRADE: Grading of Recommendations Assessment, Development and Evaluation
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search extension
RCT: randomized controlled trial
SMD: standardized mean difference
SUCRA: surface under the cumulative ranking curve

Edited by S Brini; submitted 09.Dec.2025; peer-reviewed by G Wang, S Xiong; revised version received 17.Mar.2026; accepted 17.Mar.2026; published 24.Apr.2026.

Please cite as:
Huang D, Wu D, Hexel R, Brown-Wilson C, Zhou J, Moyle W
Effectiveness of Digital Serious Games on Knowledge and Attitudes in Public Health Education: Systematic Review and Bayesian Network Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e89281
URL: https://www.jmir.org/2026/1/e89281
doi:10.2196/89281

© Di Huang, Dongjun Wu, Rene Hexel, Christine Brown-Wilson, Jing Zhou, Wendy Moyle. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Accuracy of Deep Learning for Detecting Axillary Lymph Node Metastasis in Breast Cancer: Systematic Review and Meta-Analysis

Xueying Wang[1*], MMed; Tiantian Li[2*], MMed; Xiaohang Wang[3], PhD; Deyuan Fu[1*], PhD

[1]Department of Breast Surgery, Northern Jiangsu People's Hospital Affiliated to Yangzhou University, No.98 Nantong West Road, Guangling District, Yangzhou, China

[2]Department of Ultrasonography, Northern Jiangsu People's Hospital Affiliated to Yangzhou University, Yangzhou, China

[3]Institute of Translational Medicine, Jiangsu Key Laboratory of Integrated Traditional Chinese and Western Medicine for Prevention and Treatment of Senile Diseases, Medical College, Yangzhou University, Yangzhou, China

[*]these authors contributed equally

Corresponding Author:
Deyuan Fu, PhD
Department of Breast Surgery, Northern Jiangsu People's Hospital Affiliated to Yangzhou University, No.98 Nantong West Road, Guangling District, Yangzhou, China

Abstract

Background: Axillary lymph node metastasis (ALNM) is an important factor in detecting breast cancer (BC). However, the noninvasive diagnosis of ALNM remains challenging. While some deep learning (DL) models have been developed for preoperative ALNM assessment, their performance lacks systematic evaluation.

Objective: This study aims to evaluate the effectiveness of DL in detecting ALNM, providing evidence to support clinical diagnostic tools.

Methods: Embase, Web of Science, PubMed, and Cochrane Library were searched from their inception through January 26, 2026. The Quality Assessment of Diagnostic Accuracy Studies was used to assess the risk of bias in the included studies. A bivariate mixed effects model was applied for analysis, and subgroup analyses were conducted based on different imaging modalities.

Results: This meta-analysis included 28 independent studies and pooled data from 20,811 patients with BC. Among them, 7123 cases had confirmed ALNM. The overall diagnostic performance of the DL model (bivariate mixed effects) for detecting ALNM in BC was as follows: sensitivity 0.80 (95% CI 0.76 - 0.84), specificity 0.85 (95% CI 0.80 - 0.88), diagnostic odds ratio (DOR) 22 (95% CI 16 - 30), and area under the summary receiver operating characteristic curve (AUC) 0.89 (95% CI 0.86 - 0.92). The positive likelihood ratio (LR+) was 5.2 (95% CI 4.1 - 6.5), and the negative likelihood ratio (LR−) was 0.24 (95% CI 0.19 - 0.29). For ultrasound-based DL models targeting ALNM detection, the pooled sensitivity and specificity were 0.79 (95% CI 0.72 - 0.84) and 0.86 (95% CI 0.79 - 0.91), respectively. Diagnostic performance metrics showed an LR+ of 5.5 (95% CI 3.8 - 8.1), an LR− of 0.25 (95% CI 0.19 - 0.32), a DOR of 22 (95% CI 15 - 33), and an AUC of 0.89 (95% CI 0.86 - 0.91). Regarding magnetic resonance imaging–based DL models for detecting ALNM, the pooled sensitivity was 0.78 (95% CI 0.71 - 0.83) and the pooled specificity was 0.82 (95% CI 0.76 - 0.87). Corresponding metrics included an LR+ of 4.4 (95% CI 3.3 - 5.9), an LR− of 0.27 (95% CI 0.21 - 0.35), a DOR of 16 (95% CI 11 - 25), and an AUC of 0.87 (95% CI 0.84 - 0.90). For computed tomography (CT)–based models, the sensitivity was 0.90 (95% CI 0.78 - 0.96), the specificity was 0.88 (95% CI 0.84 - 0.92), and the AUC was as high as 0.91 (95% CI 0.89 - 0.94).

Conclusions: Current DL methods for detecting ALNM in BC primarily utilize ultrasound, magnetic resonance imaging, and CT. DL models based on all 3 modalities demonstrated good diagnostic performance. CT had the highest sensitivity and AUC, while its specificity was comparable to that of ultrasound. These findings provide supportive evidence for the development or optimization of clinical diagnostic models.

Trial Registration: PROSPERO CRD42024609828; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024609828

(J Med Internet Res 2026;28:e77593) doi:10.2196/77593

KEYWORDS

breast cancer; axillary lymph node metastasis; deep learning; systematic review; ultrasound; magnetic resonance imaging; computed tomography

XSL•FO
RenderX

Introduction

Breast cancer (BC) is a common and life-threatening tumor in women. According to the latest global statistics of 2025, there were 2.3 million new BC cases (accounting for 25% of all new cancer cases among women) and 670,000 BC-related deaths (accounting for 15.5% of cancer deaths in women) worldwide in 2022. By 2050, new cases of BC are projected to rise by 38%, while BC deaths are expected to increase by 68%, with the fastest growth observed in low- and middle-income countries [1]. The incidence of BC varies notably by region. New Zealand and Australia have the highest incidence of BC, with an age-standardized incidence rate of 100.3 per 100,000, while South Asia has the lowest rate at 26.7 per 100,000 [1]. Moreover, the incidence of BC is rising at approximately 0.6% per year and is trending toward a younger age at onset [2,3]. Consequently, BC has emerged as a severe disease burden on society. BC treatment includes local and systemic therapies, with surgery, radiotherapy, and chemotherapy as the primary therapeutic approaches. Currently, immunotherapy [4], brachytherapy [5], and neoadjuvant therapy [6] have significantly improved the clinical outcomes and prolonged the survival of patients with BC. However, some patients still experience poor prognoses due to influencing factors, such as axillary lymph node metastasis (ALNM), molecular subtypes, and lymphovascular invasion. As ALNM can elevate the risk of recurrence and metastasis of BC, it is an important factor for assessing the treatment and prognosis of BC [7]. Currently, the accurate detection of ALNM before BC surgery remains challenging. Although techniques including fine-needle aspiration cytology and core needle biopsy can be used in clinical practice, these invasive procedures may lead to complications such as implantation metastasis and hematoma. Moreover, these procedures may cause false negatives, thereby affecting diagnostic accuracy and timely treatment. Thus, developing a preoperative, efficient, accurate, and safe method for detecting ALNM is of great clinical significance for improving the diagnostic and prognostic evaluation of patients with BC. Furthermore, it serves as a critical research topic that requires in-depth exploration.

With advances in machine learning (ML), many researchers have applied ML approaches to oncology. ML can integrate high-dimensional data, including clinical features, gene expression, and imaging characteristics, to construct models for diagnosing disease, assessing progression, and evaluating prognosis [8-10]. Traditional ML depends on manual coding and requires variable selection, image segmentation, and extraction of image features using specialized software (eg, 3D Slicer [Slicer Community] and ITK-SNAP [Penn Image Computing and Science Laboratory]). The features are then imported into ML for filtering, modeling, and validation. However, information may be lost during the extraction and screening process. Although deep learning (DL) can automatically train on images, it still relies on manual segmentation for model construction in current research. DL can directly train models based on segmented images. Therefore, this technique preserves maximal information, enhancing the diagnostic accuracy [11-13]. Accordingly, some researchers have developed image-based DL models [14-16] to improve the preoperative diagnosis of ALNM. Nonetheless, systematic evidence on the accuracy of DL for detecting ALNM remains limited. Therefore, this study aims to evaluate the performance of current image-based DL models for detecting ALNM in BC and provide evidence-based insights for developing or updating intelligent diagnostic tools for clinical practice.

Methods

Study Registration

This study was performed in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines (Checklist 1) and was registered in the PROSPERO (Prospective Register of Systematic Reviews) database (ID CRD42024609828).

Eligibility Criteria

The eligibility criteria were established as the basis for subsequent literature screening as shown in Table 1.

XSL•FO
RenderX

Table . Eligibility criteria for study inclusion in the systematic review and meta-analysis of DL[a] models for detecting ALNM[b] status in patients with BC[c].

PICOS	Inclusion criteria	Exclusion criteria
Participants (P)	• Studies involving patients with BC	• Studies that did not strictly differentiate BC from other tumors
Intervention (I)	• Studies that developed DL models for ALNM detection	• Studies that only applied traditional ML[d] • Studies that only performed image segmentation without constructing a DL model
Comparison (C)	• None	• None
Outcomes (O)	• Studies that reported metrics for evaluating ML model performance (eg, sensitivity, specificity, area under the summary receiver operating characteristic curve, negative likelihood ratio, positive likelihood ratio, and diagnostic odds ratio)	• Studies that did not report the aforementioned performance metrics
Study design (S)	• Case-control, cohort, or cross-sectional studies • Studies published in English	• Conference abstracts, guidelines, meta-analyses, reviews, or expert opinions • Lymph node status not confirmed by biopsy • No clear differentiation between the sentinel and axillary lymph nodes

[a]DL: deep learning.

[b]ALNM: axillary lymph node metastasis.

[c]BC: breast cancer.

[d]ML: machine learning.

Data Sources and Search Strategy

Web of Science, Cochrane Library, PubMed, and Embase were searched from their inception up to January 26, 2026. The search strategy was designed by combining free-text terms and subject headings. There were no temporal or geographical restrictions. The search strategy is illustrated in Table S1 in Multimedia Appendix 1.

Study Selection and Data Extraction

Initially retrieved studies were imported into EndNote. After removing duplicates, the titles and abstracts of the remaining studies were further reviewed. After a full-text review, eligible studies were selected. Data extraction was performed using a standardized form. The extracted information included digital object identifier, publication year, patient source, country, title, study design, first author, number of ALNM cases, total cases, number of ALNM cases within the training cohort, total cases within the training cohort, number of ALNM cases within the validation cohort, total cases within the validation cohort, methods used to generate the validation cohort, whether region of interest segmentation was performed, model type, and whether a clinical comparison was conducted. Two reviewers independently selected the studies and extracted the data. Disagreements were resolved through consultation with a third reviewer.

Risk of Bias in Included Studies

The risk of bias (RoB) of the included studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies [17]. This tool assessed overall bias and clinical applicability in 4 domains: index test, flow and timing, reference standard, and patient selection. Each domain was rated as low, high, or unclear RoB according to specific criteria. Two reviewers independently assessed the RoB and cross-checked the evaluation results. Disagreements were resolved through consultation with a third reviewer.

Synthesis Methods

Data analysis was performed using Stata 15.0 (StataCorp LLC). A bivariate mixed effects (BME) model was used to explore the nonlinear relationship between sensitivity and specificity. The model estimated the following along with their corresponding 95% CIs: pooled sensitivity, specificity, area under the summary receiver operating characteristic curve (AUC), negative likelihood ratio (LR−), positive likelihood ratio (LR+), and diagnostic odds ratio (DOR). Publication bias was evaluated using the Deeks funnel plot. Sensitivity and specificity were derived from a 2×2 contingency table. Since most studies did not report contingency tables, the pooled sensitivity, specificity, precision, and case numbers were used for calculations. In the validation phase, a meta-analysis was conducted. When multiple validation sets were available within a single investigation, all were incorporated into the analysis. Subgroup analyses were performed, stratified by image type and validation set generation method. Furthermore, the analysis aggregated the outcomes from both internal and external validations for DL models developed using ultrasound, magnetic resonance imaging (MRI), and computed tomography (CT). Moreover, for studies with multiple validation cohorts, only the cohort with the highest Youden index was retained to ensure

XSL•FO
RenderX

that a single validation cohort was retained in each study for the sensitivity analysis. The meta-analysis was then repeated with this reduced dataset. A *P* value <.05 was considered statistically significant.

Results

Study Selection

A total of 2225 studies were selected from the databases, among which 521 duplicates were removed. Following a title and abstract screening, 1659 studies were further excluded. The full texts of the remaining 45 studies were assessed, and 17 studies were removed for the following reasons: inaccessible conference abstracts (n=4), absence of outcome indicators for assessing the accuracy of DL (n=6), sole focus on image segmentation (n=2), use of positron emission tomography-CT as the diagnostic gold standard (n=1), and lack of clear differentiation between sentinel and axillary lymph nodes (n=4). Ultimately, 28 studies [14,16,18-43] were included in the analysis. The study screening process is illustrated in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram illustrating the identification and selection process of studies evaluating DL models for the preoperative diagnosis of ALNM in BC. Data sources: Web of Science, Cochrane Library, PubMed, and Embase (search date: January 26, 2026). ALNM: axillary lymph node metastasis; BC: breast cancer; PET-CT: positron emission tomography-computed tomography.

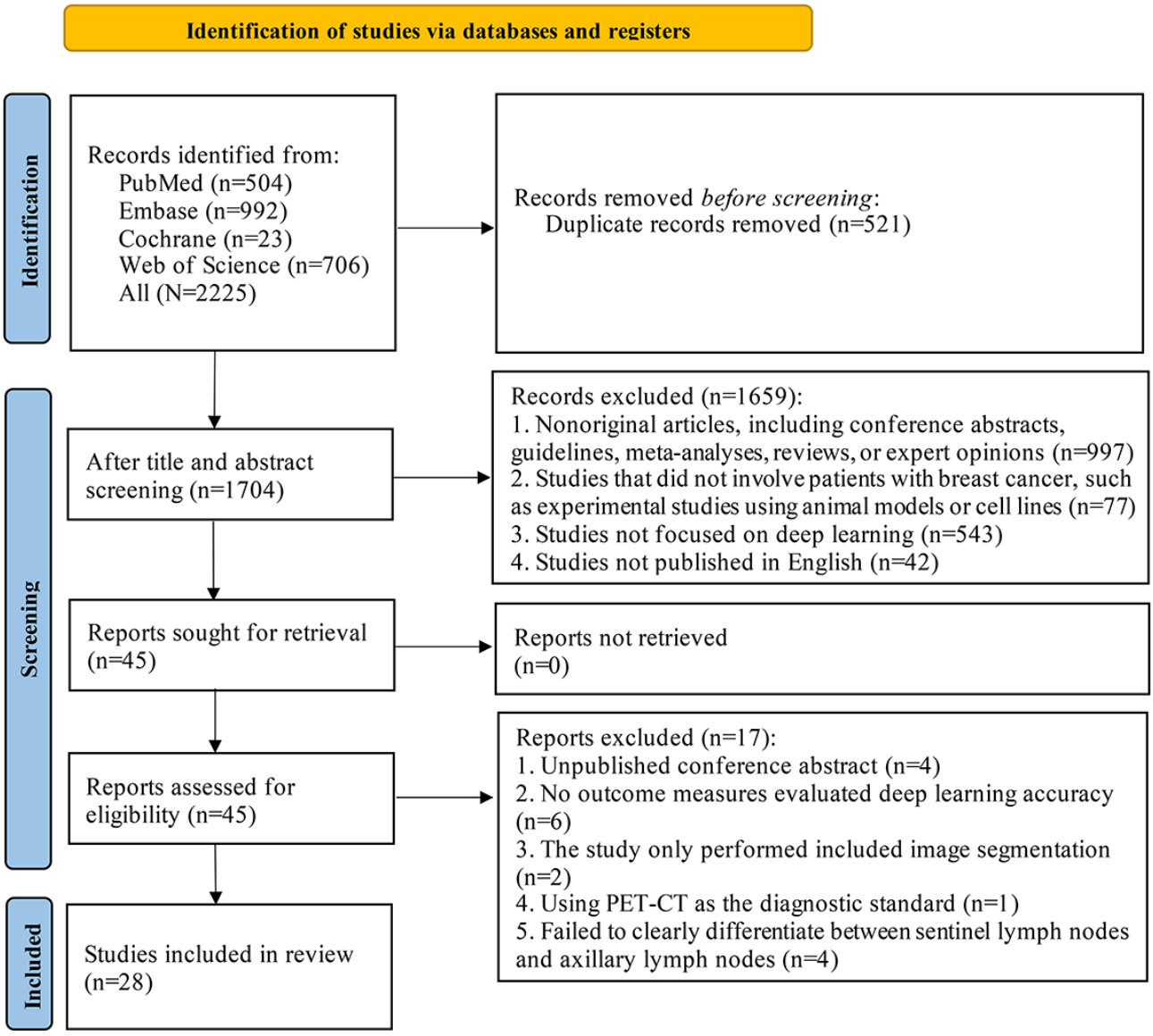

Study Characteristics

A total of 28 studies published between 2018 and 2026 were included. These studies involved 20,811 patients with BC, 7123 of whom had ALNM. Twenty-three of the studies were from China, 3 from the United States, 1 from Italy, and 1 from Korea. All of the studies used a case-control design. Seventeen of the studies were single-center, and 11 were multicenter. Regarding imaging modalities, 10 studies utilized MRI, 14 used ultrasound images, and 4 used CT scans. All 28 studies clearly described their validation methods. Twenty studies employed random sampling, 10 used external validation, and 3 utilized cross-validation. Twenty-three studies conducted manual segmentation, and 7 reported the comparisons between DL models and clinical physicians (Table 2).

XSL•FO
RenderX

Table . Characteristics of the 28 included studies on DL[a] for ALNM[b] in patients with BC[c] published between 2018 and 2026[d].

No.	Included study and year of publication	Country of the author	Study type	Patient source	Modality	Number of lymph node metastasis cases	Total number of cases	Total number of cases in the training set	Generation method of the validation set	Number of cases in the validation set	Whether region of interest segmentation is used or not	Segmentation method	Network architecture	Network input	Comparison with clinicians
1	Gong et al [18] (2025)	China	Case-control	Single center	US[e]	312	1280	921	Random sampling, temporal validation	I1[f]: 102; I2: 257	Yes	Manual	Transformer	2D	No
2	Ayekum et al [19] (2025)	China	Case-control	Multicenter	US	295	820	621	External validation	E1[g]: 112; E2: 87	Yes	Manual	ResNet50[h] and GCN[i]	2D	No
3	Li et al [20] (2025)	China	Case-control	Multicenter	US	1013	1557	1144	Random sampling, external validation	I: 273; E: 140	Yes	Manual	Transformer	2D	No
4	Gu et al [21] (2025)	China	Case-control	Multicenter	MRI[j]	201	520	287	Random sampling, external validation	I: 124; E: 109	Yes	Semi-automatic	Convolutional neural network (CNN)	3D	No
5	Dai et al [22] (2025)	China	Case-control	Multicenter	MRI	386	935	742	Random sampling	I: 83; E: 110	Yes	Manual	3D ResNet, 3D-Xception, and HR-Net[k]	2D	Yes
6	Sun et al [23] (2025)	China	Case-control	Single center	CT[l]	91	258	Not reported	Random sampling	Not reported	Yes	Manual	ResNet101	2D	No
7	Wang et al [24] (2024)	China	Case-control	Single center	US	67	266	212	Random sampling	I: 27; T: 27(3)	Yes	Automatic	DeepLabV3, ResNet-101, and CNN	2D	No
8	Polat et al [25] (2024)	America	Case-control	Single center	MRI	152	253	All cases	Cross-validation (2)	All cases	Yes	Manual	CNN	3D	No
9	Zhou et al [26] (2024)	China	Case-control	Multicenter	MRI	530	1259	801	Random sampling	I: 344; E: 114	Yes	Manual	ResNet101, ResNeXt101, and DenseNet[m]	2D	Yes

XSL•FO
RenderX

No.	Included study and year of publication	Country of the author	Study type	Patient source	Modality	Number of lymph node metastasis cases	Total number of cases	Total number of cases in the training set	Generation method of the validation set	Number of cases in the validation set	Whether region of interest segmentation is used or not	Segmentation method	Network architecture	Network input	Comparison with clinicians
10	Park et al [14] (2024)	Korea	Case-control	Single center	CT	303	523	417	Random sampling	I1: 53; I2: 53	Yes	Manual	DenseNet 121	2D axial CT slice	No
11	Liu et al [27] (2024)	China	Case-control	Multi-center	US	327	883	621	External validation	E1: 112; E2: 87; E3: 63	Yes	Manual	ResNet50	2D US image	Yes
12	Guo et al [28] (2024)	China	Case-control	Multi-center	MRI	1064	2063	1256	Random sampling, external validation	I: 539; E1: 153; E2: 115	Yes	Manual	Convolutional recurrent neural network	2D image	No
13	Wei et al [29] (2023)	China	Case-control	Single center	US	349	892	535	Random sampling	I1: 178; I2: 179	Yes	Manual	Inception_v3, ResNet101, ResNet50, VGG19[n], and Video_R(2+1)D	2D grayscale image	No
14	Li et al [16] (2023)	China	Case-control	Multi-center	US	127	320	210	External validation	I: 61; E: 49	Yes	Automatic	CNN (R2+1D, TIN[o], ResNet-3D)	Video sequence (3D)	Yes
15	Gao et al [30] (2023)	China	Case-control	Multi-center	MRI	387	941	742	Random sampling, external validation	I: 83; E: 116	Yes	Manual	3D ResNet + CBAM[p]	3D DCE-MRI[q] volume data	Yes
16	Zhang et al [31] (2022)	China	Case-control	Single center	MRI	98	252	202	Random sampling	50	Yes	Manual	ResNet50	2D	No
17	Zhang et al [32] (2022)	China	Case-control	Single center	US	394	952	902	External validation	50	No	None	CNN	2D	Yes
18	Wang et al [33] (2022)	China	Case-control	Single center	MRI	163	348	315	Random sampling	33	Yes	Manual	ResNet50	2D	No
19	Sun et al [34] (2022)	America	Case-control	Single center	US	64	169	All cases	Cross-validation	All cases	Yes	Manual	Custom CNN and ResNet-101	2D	No

XSL•FO
RenderX

No.	Included study and year of publication	Country of the author	Study type	Patient source	Modality	Number of lymph node metastasis cases	Total number of cases	Total number of cases in the training set	Generation method of the validation set	Number of cases in the validation set	Whether region of interest segmentation is used or not	Segmentation method	Network architecture	Network input	Comparison with clinicians
20	Santucci et al [35] (2022)	Italy	Case-control	Single center	MRI	27	128	All cases	Cross-validation	All cases	Yes	Manual	SFB-NET[r], VB-NET[s], 2DS-NET[t] (optimal)	2D slice and 3D volume	No
21	Li et al [36] (2022)	China	Case-control	Single center	US	489	2131	1491	Random sampling	640	Yes	Manual	ResNet-50	2D	No
22	Cattell et al [37] (2022)	America	Case-control	Single center	MRI	67	198	109	Random sampling	I1: 54; I2: 35	Yes	Manual	VGG16	2D	No
23	Zeng et al [38] (2021)	China	Case-control	Single center	CT	100	229	153	Random sampling	76	Yes	Manual	Custom CNN (decoupling convolution)	2D	No
24	Liu et al [39] (2021)	China	Case-control	Single center	CT	400	800	480	Random sampling	I1: 160; I2:160	Yes	Manual	DA-VGG19	2D	No
25	Zhou et al [40] (2020)	China	Case-control	Multi-center	US	420	834	680	Random sampling, external validation	I: 76; E: 78	No	None	Inception V3, Inception-ResNet V2, and ResNet-101	2D	Yes
26	Zheng et al [41] (2020)	China	Case-control	Single center	US	247	584	466	Random sampling	118	Yes	Manual	ResNet-50	2D	No
27	Sun et al [42] (2020)	China	Case-control	Single center	US	136	479	359	Random sampling	120	Yes	Manual	DenseNet-121	2D	No

XSL•FO
RenderX

No.	Included study and year of publication	Country of the author	Study type	Patient source	Modality	Number of lymph node metastasis cases	Total number of cases	Total number of cases in the training set	Generation method of the validation set	Number of cases in the validation set	Whether region of interest segmentation is used or not	Segmentation method	Network architecture	Network input	Comparison with clinicians
28	Guo et al [43] (2020)	China	Case-control	Multi-center	US	365	937	542	External validation	395	Yes	Manual	Custom CNN	2D	No

[a]DL: deep learning.

[b]ALNM: axillary lymph node metastasis.

[c]BC: breast cancer.

[d]The table details study design, patient source, imaging modality, and validation methods. The 8th, 19th, and 20th studies in the table employed cross-validation methodologies. In the 3 studies, all cases were used in both the training and validation sets, and the training and validation sets were not independently partitioned. The construction of 2×2 contingency tables was not possible due to missing data, and some study cohorts were not included in the statistical analysis.

[e]US: ultrasound.

[f]I: internal validation cohort.

[g]E: external validation cohort.

[h]ResNet50: Residual Network 50.

[i]GCN: graph convolutional network.

[j]MRI: magnetic resonance imaging.

[k]HRNet: high-resolution network.

[l]CT: computed tomography.

[m]DenseNet: densely connected convolutional network.

[n]VGG: Visual Geometry Group.

[o]TIN: temporal interlacing network.

[p]CBAM: convolutional block attention module.

[q]DCE: dynamic contrast-enhanced.

[r]SFB-NET: single fixed-size box network.

[s]VB-NET: variable-size box network.

[t]2DS-NET: 2-dimensional slice network.

RoB in Included Studies

Risk of Selection Bias and Outcome Assessment

All studies adopted consecutive sampling and applied image-based DL. Although these studies followed a case-control design, there was no bias in evaluating the performance of image-based DL. Therefore, the risk of selection bias was considered low as the case exclusions were reasonable. None of the studies described how the gold standard was interpreted. However, since this meta-analysis focused on artificial intelligence (AI)–based DL, the interpretation of the reference standard had minimal impact on the role of DL in assessing positive events. Thus, the RoB in outcome assessment was considered low. The time interval between the reference standard and the index test was appropriate in all 28 studies, each of which used a single, consistent gold standard. Regarding patient flow, all studies were rated as having a low RoB.

Concerns Regarding Applicability (High RoB)

However, 3 studies were rated as having a high RoB in the assessment of clinical applicability because they relied solely on cross-validation.

Summary of Assessments

The RoB assessments are illustrated in Figures 2 and 3.

XSL•FO
RenderX

Figure 2. Summary of RoB, as assessed via the QUADAS-2 tool, for the 28 included case-control studies (2018 - 2026) that evaluated the image-based DL models in detecting ALNM among patients with BC. The assessment across the 4 domains (patient selection, index test, reference standard, and flow and timing) indicates an overall low RoB. ALNM: axillary lymph node metastasis; BC: breast cancer; DL: deep learning; RoB: risk of bias; QUADAS: Quality Assessment of Diagnostic Accuracy Studies.

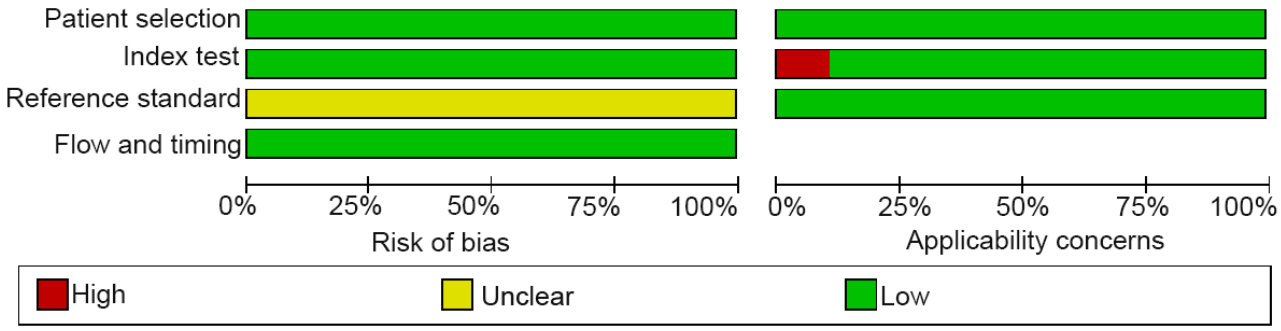

XSL•FO
RenderX

Figure 3. Overall low concerns regarding applicability identified by QUADAS-2 assessment of DL studies for ALNM diagnosis [14,16,18-24,26-33,36-43]. Among the 28 studies evaluated, 3 studies that used only cross-validation raised concerns due to their limited generalizability to broader clinical practice. ALNM: axillary lymph node metastasis; DL: deep learning; QUADAS: Quality Assessment of Diagnostic Accuracy Studies.

XSL•FO
RenderX

Overall Model

Across all included studies, a total of 40 2×2 contingency tables from validation cohorts were utilized to evaluate the accuracy of DL in detecting ALNM, with an ALNM prevalence of 40.1%. The pooled results from the BME model were as follows: a sensitivity of 0.80 (95% CI 0.76 - 0.84), a specificity of 0.85 (95% CI 0.80 - 0.88), an LR+ of 5.2 (95% CI 4.1 - 6.5), an LR− of 0.24 (95% CI 0.19 - 0.29), a DOR of 22 (95% CI 16 - 30), and the AUC of 0.89 (95% CI 0.86 - 0.92) (Figures 4 and 5). The Deeks funnel plot revealed no significant publication bias among the studies (*P*=.11) (Figure 6). Assuming a prior probability of 0.4, a positive test result corresponded to a true positive (TP) probability of 0.78 (Figure 7).

Figure 4. Forest plots showing the pooled sensitivity and specificity of DL models for diagnosing ALNM [14,16,18-43]. The data were pooled from 40 validation cohorts across 28 studies, demonstrating an overall sensitivity of 0.80 (95% CI 0.76 - 0.84) and a specificity of 0.85 (95% CI 0.80 - 0.88). ALNM: axillary lymph node metastasis; DL: deep learning.

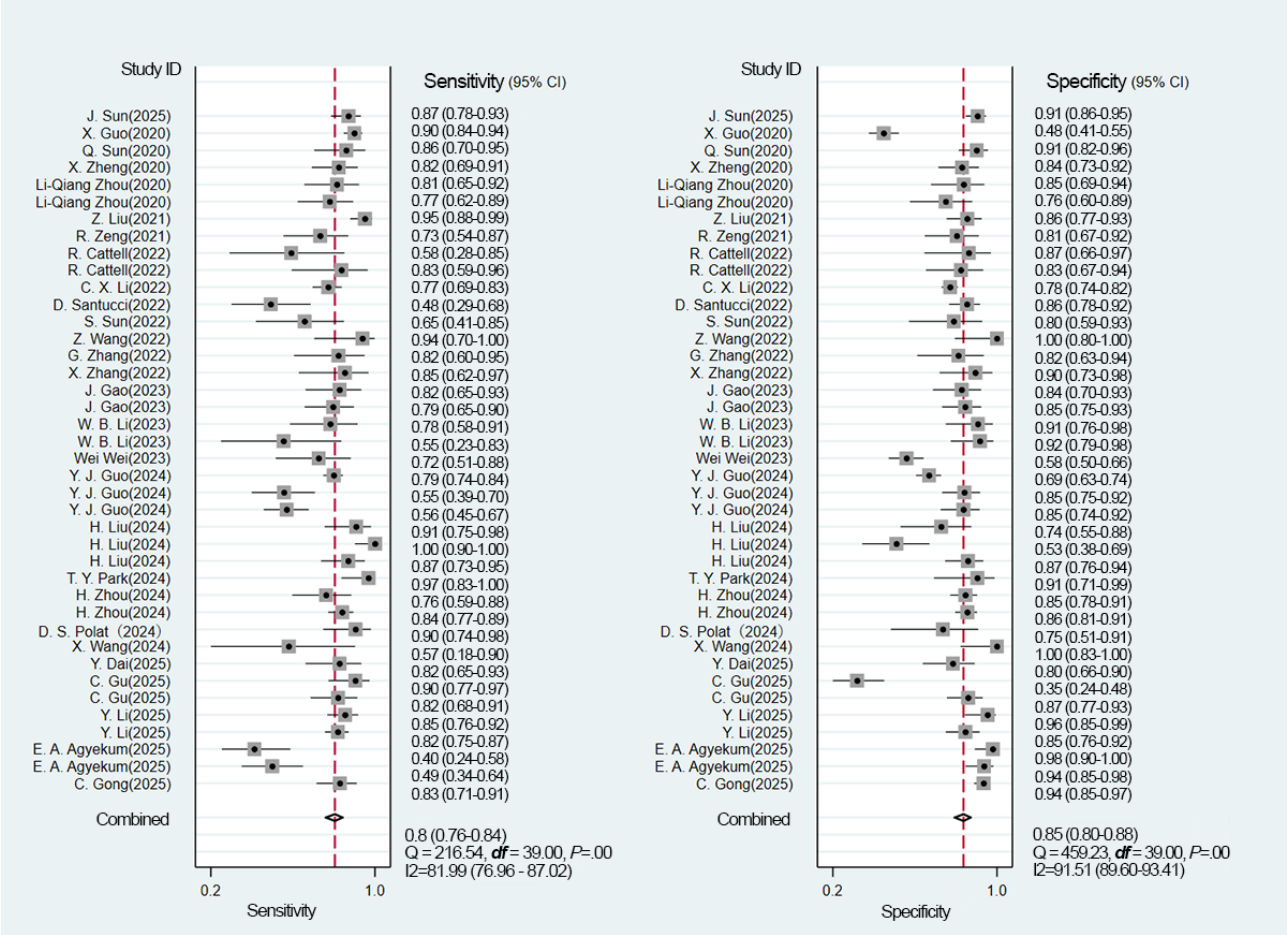

The internal validation analysis used 26 diagnostic 4-fold tables to evaluate the accuracy of a DL model in detecting ALNM. The observed prevalence of ALNM was 39%. The BME model yielded the following pooled estimates: a sensitivity of 0.82 (95% CI 0.78 - 0.85), a specificity of 0.85 (95% CI 0.82 - 0.88), an LR+ of 5.6 (95% CI 4.4 - 7.0), and an LR− of 0.21 (95% CI 0.18 - 0.26). The DOR was 26 (95% CI 18 - 38), with an AUC of 0.90 (95% CI 0.87 - 0.92) (Figures S1 and S2 in Multimedia Appendix 1). The Deeks funnel plot indicated no significant publication bias (*P*=.21) (Figure S3 in Multimedia Appendix 1). Assuming a prior probability of 0.4, a positive test result corresponded to a TP probability of 0.79 (Figure S4 in Multimedia Appendix 1).

For the external validation, 14 diagnostic 4-fold tables were assessed, with an ALNM prevalence of 43%. The BME model demonstrated a pooled sensitivity of 0.78 (95% CI 0.66 - 0.86), a specificity of 0.82 (95% CI 0.72 - 0.90), an LR+ of 4.4 (95% CI 2.9 - 6.8), an LR− of 0.27 (95% CI 0.19 - 0.39), a DOR of 16 (95% CI 10 - 26), and an AUC of 0.87 (95% CI 0.84 - 0.90; Figures S5 and S6 in Multimedia Appendix 1). The Deeks funnel plot indicated no significant publication bias (*P*=.16; Figure S7 in Multimedia Appendix 1). Assuming a prior probability of 0.4, a positive test result corresponded to a TP probability of 0.75 (Figure S8 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 5. Summary receiver operating characteristic curve (with prediction and confidence contours) from the bivariate meta-analysis evaluating the diagnostic accuracy of DL models for detecting ALNM in BC. The curve is based on 40 contingency tables derived from the validation cohorts of the included studies. The analysis yielded a summary AUC of 0.89 (95% CI 0.86 - 0.92). ALNM: axillary lymph node metastasis; AUC: area under the summary receiver operating characteristic curve; BC: breast cancer; DL: deep learning.

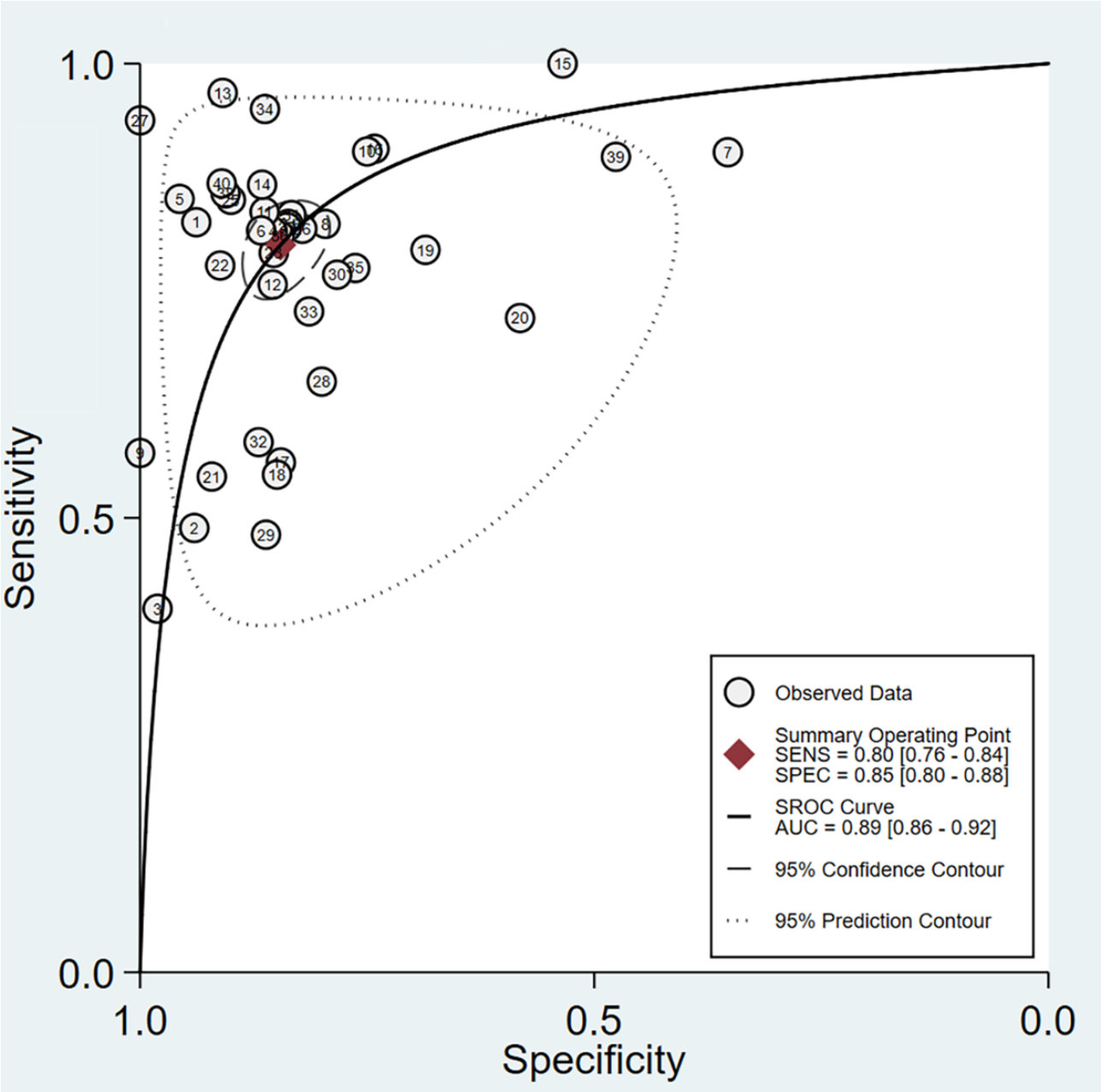

XSL•FO
RenderX

Figure 6. Assessment of potential publication bias using the Deeks funnel plot asymmetry test for the included studies on DL models in detecting ALNM in BC. The test indicated potential publication bias (*P*=.11). ALNM: axillary lymph node metastasis; BC: breast cancer; DL: deep learning; ESS: effective sample size.

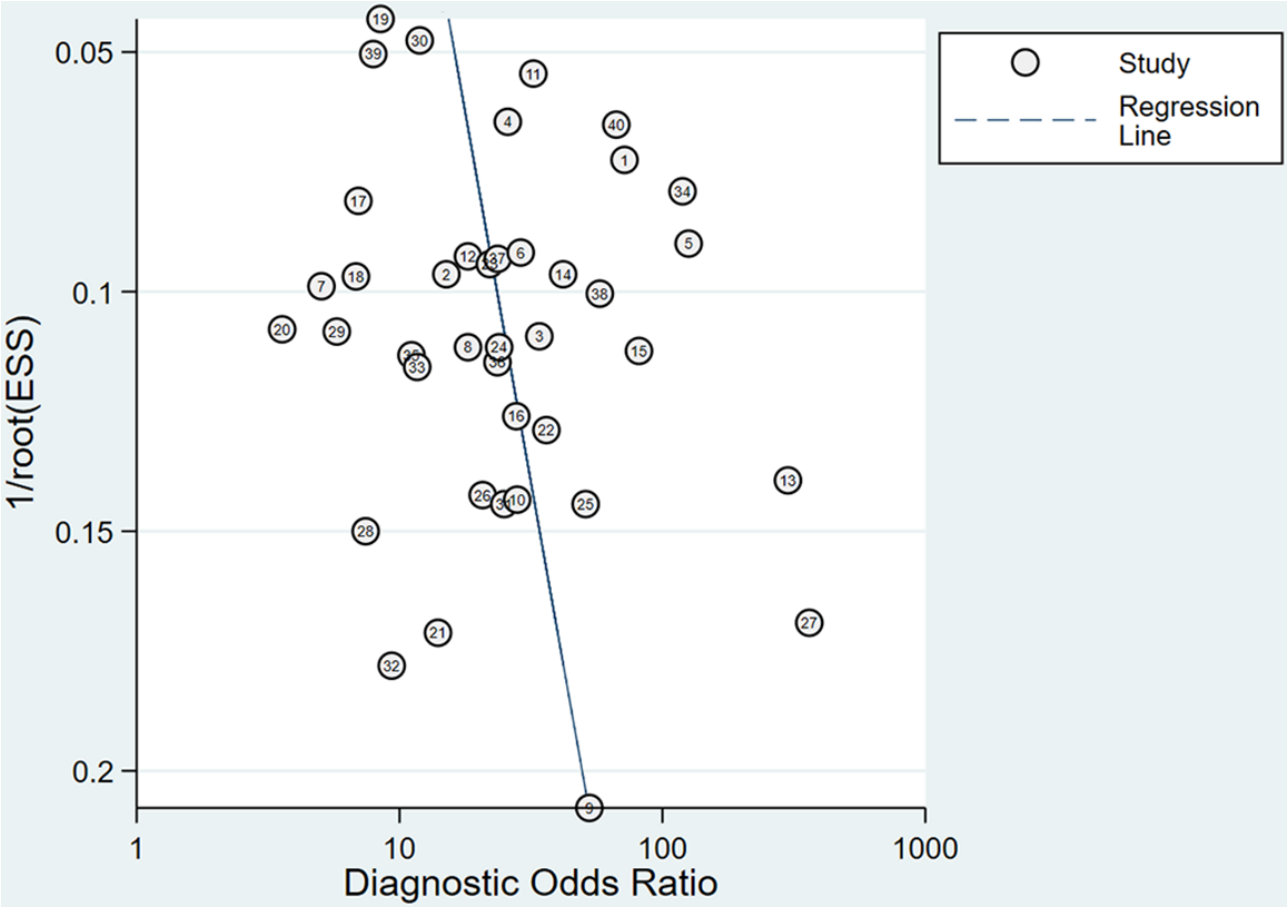

XSL•FO
RenderX

Figure 7. Fagan nomogram illustrating the 78% posttest probability of ALNM following a positive DL model result, given a 40% pretest probability. ALNM: axillary lymph node metastasis; DL: deep learning; LR: likelihood ratio.

XSL•FO
RenderX

Conventional Ultrasound

Fourteen studies utilizing conventional ultrasound were included, which provided 20 distinct validation cohorts (due to multiple cohorts from several studies) for assessing DL models, with an ALNM prevalence of 38 %. The pooled results from the BME model were as follows: a sensitivity of 0.79 (95% CI 0.72 - 0.84), a specificity of 0.86 (95% CI 0.79 - 0.91), an LR+ of 5.5 (95% CI 3.8 - 8.1), an LR− of 0.25 (95% CI 0.19 - 0.32), a DOR of 22 (95% CI 15 - 33), and an AUC of 0.89 (95% CI 0.86 - 0.91) (Figures S9 and S10 in Multimedia Appendix 1). The Deeks funnel plot demonstrated no significant publication bias (*P*=.22) (Figure S11 in Multimedia Appendix 1). Assuming a prior probability of 0.4, a positive test result corresponded to a TP probability of 0.79 (Figure S12 in Multimedia Appendix 1).

The internal validation involved 11 diagnostic 4-fold tables to evaluate the accuracy of a DL model in detecting ALNM. The observed ALNM prevalence was 33%. The BME model yielded the following pooled estimates: a sensitivity of 0.80 (95% CI 0.76 - 0.83), a specificity of 0.85 (95% CI 0.79 - 0.90), an LR+ of 5.5 (95% CI 3.6 - 8.3), an LR− of 0.24 (95% CI 0.19 - 0.29), a DOR of 23 (95% CI 13 - 41), and an AUC of 0.85 (95% CI 0.81 - 0.88) (Figures S13 and S14 in Multimedia Appendix 1). The Deeks funnel plot indicated no significant publication bias (*P*=.71) (Figure S15 in Multimedia Appendix 1). Assuming a prior probability of 0.4, a positive test result corresponded to a TP probability of 0.78 (Figure S16 in Multimedia Appendix 1).

For the external validation, 9 diagnostic 4-fold tables were analyzed, with an ALNM prevalence of 47%. The BME model demonstrated a pooled sensitivity of 0.80 (95% CI 0.65 - 0.90), a specificity of 0.85 (95% CI 0.71 - 0.93), an LR+ of 5.4 (95% CI 2.9 - 10.3), an LR− of 0.23 (95% CI 0.13 - 0.40), a DOR of 23 (95% CI 13 - 44), and an AUC of 0.90 (95% CI 0.87 - 0.92; Figures S17 and S18 in Multimedia Appendix 1). The Deeks funnel plot indicated no significant publication bias (*P*=.16; Figure S19 in Multimedia Appendix 1). Assuming a prior probability of 0.4, a positive test result corresponded to a TP probability of 0.78 (Figure S20 in Multimedia Appendix 1).

Magnetic Resonance Imaging

Ten studies evaluated DL models based on MRI, with an ALNM prevalence of 42%. The pooled results from the BME model were as follows: a sensitivity of 0.78 (95% CI 0.71 - 0.83), a specificity of 0.82 (95% CI 0.76 - 0.87), an LR+ of 4.4 (95% CI 3.3 - 5.9), an LR− of 0.27 (95% CI 0.21 - 0.35), a DOR of 16 (95% CI 11 - 25), and an AUC of 0.87 (95% CI 0.84 - 0.90; Figures S21 and S22 in Multimedia Appendix 1). The Deeks funnel plot demonstrated no significant publication bias (*P*=.22; Figure S23 in Multimedia Appendix 1). Assuming a prior probability of 0.4, a positive test result corresponded to a TP probability of 0.75 (Figure S24 in Multimedia Appendix 1).

The internal validation encompassed 11 diagnostic 4-fold tables to examine the accuracy of a DL model in detecting ALNM. The observed prevalence of ALNM was 44%. The BME model yielded the following pooled estimates: a sensitivity of 0.80 (95% CI 0.74 - 0.86), a specificity of 0.84 (95% CI 0.79 - 0.88), an LR+ of 4.9 (95% CI 3.7 - 6.6), an LR− of 0.23 (95% CI 0.17 - 0.32), a DOR of 21 (95% CI 13 - 35), and an AUC of 0.89 (95% CI 0.86 - 0.92; Figures S25 and S26 in Multimedia Appendix 1). The Deeks funnel plot indicated no significant publication bias (*P*=.16; Figure S27 in Multimedia Appendix 1). Assuming a prior probability of 0.4, a positive test result corresponded to a TP probability of 0.77 (Figure S28 in Multimedia Appendix 1).

The external validation cohorts for MRI consisted of 4 studies encompassing 5 distinct cohorts, with an ALNM prevalence of 37%. The pooled results were as follows: a sensitivity of 0.73 (95% CI 0.57 - 0.85), a specificity of 0.78 (95% CI 0.61 - 0.90), an LR+ of 3.4 (95% CI 2.0 - 5.8), an LR− of 0.35 (95% CI 0.23 - 0.52), a DOR of 10 (95% CI 6 - 17), and an AUC of 0.82 (95% CI 0.78 - 0.85; Figures S29 and S30 in Multimedia Appendix 1). The Deeks funnel plot indicated no significant publication bias (*P*=.91; Figure S31 in Multimedia Appendix 1). Assuming a prior probability of 0.4, a positive test result corresponded to a TP probability of 0.69 (Figure S32 in Multimedia Appendix 1).

Computed Tomography

Regarding CT-based approaches, 4 studies developed DL models utilizing CT images. Of these, 3 studies were conducted in China and 1 in Korea. All 4 studies employed internal validation to generate their validation cohorts, with an ALNM prevalence of 43%. The pooled results were as follows: a sensitivity of 0.90 (95% CI 0.78 - 0.96), a specificity of 0.88 (95% CI 0.84 - 0.92), an LR+ of 7.8 (95% CI 5.3 - 11.5), an LR− of 0.11 (95% CI 0.05 - 0.26), a DOR of 68 (95% CI 24 - 196), and an AUC of 0.91 (95% CI 0.89 - 0.94; Figures S33 and S34 in Multimedia Appendix 1). The Deeks funnel plot indicated no significant publication bias (*P*=.97; Figure S35 in Multimedia Appendix 1). Assuming a prior probability of 0.4, a positive test result corresponded to a TP probability of 0.84 (Figure S36 in Multimedia Appendix 1).

Chinese Population

Fifteen studies, comprising 34 diagnostic 4-fold tables, validated DL models constructed using Chinese populations, with an ALNM prevalence of 40%. The pooled results from the BME model were as follows: a sensitivity of 0.80 (95% CI 0.76 - 0.84), a specificity of 0.85 (95% CI 0.80 - 0.88), an LR+ of 5.3 (95% CI 4.1 - 6.8), an LR− of 0.23 (95% CI 0.19 - 0.28), a DOR of 23 (95% CI 16 - 31), and an AUC of 0.89 (95% CI 0.86 - 0.92; Figures S37 and S38 in Multimedia Appendix 1). The Deeks funnel plot revealed no significant publication bias (*P*=.08; Figure S39 in Multimedia Appendix 1). Assuming a prior probability of 0.4, a positive test result corresponded to a TP probability of 0.78 (Figure S40 in Multimedia Appendix 1).

Population of Other Countries

Six studies, comprising 5 diagnostic 4-fold tables, validated DL models constructed using non-Chinese populations, with an ALNM prevalence of 38%. The pooled results from the BME model were as follows: a sensitivity of 0.79 (95% CI 0.59 - 0.90), a specificity of 0.84 (95% CI 0.78 - 0.89), an LR+ of 5.0 (95% CI 3.5 - 7.3), an LR− of 0.25 (95% CI 0.12 - 0.53), a DOR of 20 (95% CI 7 - 53), and the AUC of 0.85 (95% CI 0.82 - 0.88; Figures S41 and S42 in Multimedia Appendix 1).

XSL•FO
RenderX

The Deeks funnel plot demonstrated no significant publication bias (*P*=.67; Figure S43 in Multimedia Appendix 1). Assuming a prior probability of 0.4, a positive test result corresponded to a TP probability of 0.77 (Figure S44 in Multimedia Appendix 1).

Sensitivity Analysis

A sensitivity analysis was conducted by retaining only 1 validation cohort per study. For studies with multiple cohorts, the validation cohort with the highest Youden index was selected.

Analysis Based on Conventional Ultrasound Studies

Fourteen studies utilizing conventional ultrasound were included in the assessment of DL models, with an ALNM prevalence of 34%. The pooled results from the BME model were as follows: a sensitivity of 0.79 (95% CI 0.73 - 0.84), a specificity of 0.86 (95% CI 0.78 - 0.92), an LR+ of 5.8 (95% CI 3.7 - 9.1), an LR− of 0.24 (95% CI 0.19 - 0.31), a DOR of 24 (95% CI 14 - 41), and an AUC of 0.88 (95% CI 0.85 - 0.91; Figures S45 and S46 in Multimedia Appendix 1). The Deeks funnel plot demonstrated no significant publication bias (*P*=.24; Figure S47 in Multimedia Appendix 1). Assuming a prior probability of 0.4, a positive test result corresponded to a TP probability of 0.79 (Figure S48 in Multimedia Appendix 1).

Analysis Based on MRI Studies

Ten studies evaluated DL models based on MRI, with an ALNM prevalence of 44%. The pooled results from the BME model were as follows: a sensitivity of 0.82 (95% CI 0.75 - 0.87), a specificity of 0.84 (95% CI 0.78 - 0.88), an LR+ of 5.0 (95% CI 3.6 - 6.8), an LR− of 0.22 (95% CI 0.16 - 0.30), a DOR of 22 (95% CI 13 - 38), and an AUC of 0.90 (95% CI 0.87 - 0.92; Figures S49 and S50 in Multimedia Appendix 1). The Deeks funnel plot demonstrated no significant publication bias (*P*=.11; Figure S51 in Multimedia Appendix 1). Assuming a prior probability of 0.4, a positive test result corresponded to a TP probability of 0.77 (Figure S52 in Multimedia Appendix 1).

Human Clinical Experts

Seven studies compared the diagnostic performance of human clinical experts in assessing the ALNM status based on imaging. Among them, 4 studies included 7 clinicians who differentiated ALNM using ultrasound. The pooled sensitivity was 0.65 (95% CI 0.59 - 0.71), specificity was 0.75 (95% CI 0.70 - 0.79), LR+ was 2.6 (95% CI 2.1 - 3.2), LR− was 0.46 (95% CI 0.39 - 0.56), DOR was 6 (95% CI 4 - 8), and the AUC was 0.76 (95% CI 0.20 - 0.97; Figures S53 and S54 in Multimedia Appendix 1). The Deeks funnel plot showed no significant publication bias (*P*=.80; Figure S55 in Multimedia Appendix 1). Assuming a prior probability of 0.4, a positive test result corresponded to a TP probability of 0.63 (Figure S56 in Multimedia Appendix 1). Three investigations assessed the performance of human clinicians in diagnosing ALNM via MRI. In the study by Jing Gao et al [30], 3 radiologists independently evaluated MRI scans for ALNM in the validation set, achieving sensitivities between 64.7% and 75.0% and specificities ranging from 70.6% to 83.8%. With the assistance of the attention-based DL model RCNet, diagnostic performance was markedly improved, resulting in sensitivities of 76.5% to 83.3% and specificities of 80.9% to 89.7%. In the study by Y. Dai et al, 3 human experts evaluated the diagnostic ability of MRI for lymph node metastasis in BC (a sample size of 110). The sensitivity ranged from 0.681 to 0.767, and the specificity ranged from 0.746 to 0.841. Furthermore, the study by Zhou et al [26] found that 3 clinical experts achieved sensitivities ranging from 0.660 to 0.745 and specificities from 0.731 to 0.836 in detecting ALNM. When combined with DL, sensitivity increased to 0.787 - 0.829 and specificity increased to 0.806 - 0.895. These findings underscore the clinical utility of AI as an auxiliary diagnostic tool.

Discussion

Main Findings of This Study

This meta-analysis revealed that DL approaches for detecting ALNM in BC primarily relied on MRI, ultrasound, and CT. The sensitivity and specificity for ultrasound-based DL models were 0.79 and 0.86, respectively. Meanwhile, the sensitivity and specificity of MRI-based models were 0.78 and 0.82, respectively. Furthermore, CT-based models demonstrated a sensitivity of 0.90 and a specificity of 0.88. The sensitivity of CT was superior to that of MRI and ultrasound, while no substantial differences in specificity were observed among the 3 modalities. These robust performance metrics validated the potential of DL as a complementary tool to conventional imaging assessments.

Comparison With Previous Reviews

Currently, ML has gained widespread attention in the diagnosis and treatment of BC. Previous systematic reviews have evaluated the accuracy of ML models based on various imaging modalities. For instance, a meta-analysis by Jing Zhang et al [44], which included 13 studies involving 1618 patients, demonstrated that dynamic contrast-enhanced MRI radiomics had promising diagnostic performance in detecting ALNM and sentinel lymph node metastasis in patients with BC. However, only 2 of these studies employed DL methods, and the pooled sensitivity and specificity were 0.84 (95% CI 0.53 - 0.96) and 0.65 (95% CI 0.31 - 0.89), respectively. The limited number of studies constrained the interpretation of the results. Another meta-analysis by Chen et al [45], which included 14 studies, focused on ML-based MRI for diagnosing ALNM in patients with BC, and the pooled sensitivity and specificity were 0.79 (95% CI 0.74 - 0.84) and 0.77 (95% CI 0.73 - 0.81), respectively. This study primarily summarized models, such as support vector machines, logistic regression, and linear discriminant analysis, without specifically addressing the diagnostic accuracy of DL techniques. Eldaly et al [46] systematically reviewed radiomics approaches for detecting ALNM. Their results revealed that the AUC of the included studies varied from 0.72 to 0.93. However, they did not provide a detailed summary of diagnostic performance, thus limiting the assessment of its true accuracy. Liu et al [47] investigated the feasibility of AI algorithms based on CT and MRI for detecting ALNM in BC. They found that MRI-based methods demonstrated a sensitivity of 0.85 (95% CI 0.79 - 0.90) and specificity of 0.81 (95% CI 0.66 - 0.83), while CT-based methods showed a sensitivity of 0.88 (95% CI: 0.79 - 0.94) and

XSL•FO
RenderX

specificity of 0.80 (95%CI: 0.69 - 0.88). A radiomics study [48] included MRI, ultrasound, CT, and X-ray mammography. A meta-analysis of 30 studies involving a total of 5611 patients was conducted. The pooled sensitivity and specificity of radiomics for detecting ALNM were 0.86 (95% CI 0.82 - 0.88) and 0.79 (95% CI 0.73 - 0.84), respectively, demonstrating strong overall diagnostic accuracy. Gong et al [48] comprehensively evaluated the performance of radiomics in detecting ALNM. Their results demonstrated an overall diagnostic accuracy of 23 (95% CI 16 - 33) and a sensitivity of 0.86 (95% CI 0.82 - 0.88). Radiomics serves as a promising noninvasive approach that can contribute to offering new quantitative modalities for disease diagnosis. However, their study did not strictly differentiate between types of ML, and different ML methods demonstrated varying performance in detecting positive events. Furthermore, there was no strict distinction between datasets in their study, with notable variations observed between the results in the training and validation sets.

With the advancement of AI, there has been a shift from traditional ML to DL. DL can analyze complex and diverse data, enabling personalized diagnosis and treatment that aligns with the current trend of precision medicine. Previous meta-analyses have lacked comprehensive evidence on DL models for detecting ALNM in patients with BC across various imaging modalities.

The present meta-analysis only included DL-based studies to minimize the variability of outcomes caused by different modalities. DL offers significant advantages over traditional ML with more advanced techniques for image processing, contributing to the development of diagnostic models for diseases [37]. Our findings suggested that image-based DL for detecting ALNM in BC demonstrated favorable diagnostic performance. MRI, ultrasound, and CT are commonly used techniques for detecting ALNM in BC. However, MRI is not necessary or routinely used in BC diagnosis due to its higher cost. Our findings reveal that CT-based models achieved higher sensitivity, specificity, and AUC than models based on ultrasound and MRI. However, since only 4 studies on CT were included, the results should be validated using larger datasets. Therefore, future research should develop more CT-based diagnostic models using AI to expand its role in diagnosing BC.

The trade-off between sensitivity and specificity for ultrasound and MRI highlights their complementary roles in clinical practice. The high sensitivity of ultrasound models effectively reduces false negatives by excluding metastasis, whereas the high specificity of MRI models improves confirmation of positive cases by minimizing false positives. These findings support the development of an integrated multimodal DL strategy, with ultrasound as an initial screening tool and MRI for confirming difficult cases. However, real-world validation is required.

Before developing DL-based intelligent diagnostic tools, several challenges require attention. First, the impact of different imaging protocols and preprocessing methods on DL performance must be thoroughly examined. Second, DL models typically require large image datasets. Therefore, future studies should incorporate more extensive imaging data to ensure the robustness of the DL model. Third, the impact of population heterogeneity on model performance should be carefully considered. Future research should include multicenter, multiethnic cohorts when developing DL models. Fourth, external validation is essential to enhance the credibility and generalizability of these DL models. Nevertheless, most included studies relied solely on internal validation, making it difficult to conduct external validation across multiple centers. Among the 28 studies, the majority employed single-center internal validation. Therefore, future research should incorporate more multicenter studies to confirm the real-world applicability of these models.

Study Limitations

This is the first study to systematically investigate the feasibility of DL for detecting ALNM in BC. However, several limitations should be noted. First, despite a comprehensive literature search, the number of included studies was limited, restricting in-depth analysis and the generalizability of the findings. Second, most included studies relied on random sampling validation and lacked diverse external validation, undermining the overall reliability and generalizability of the results. Third, only a few studies directly compared DL models with clinical experts, and the available data were too limited to support a robust effect size analysis. As a result, it remains uncertain whether DL surpasses clinical experts in diagnostic performance. Fourth, many studies did not provide detailed descriptions of critical aspects, such as image acquisition protocols, segmentation techniques, network architectures and inputs, and reference standards (including biopsy methods and reference intervals). Fifth, this meta-analysis relied solely on data reported in the published articles and their supplementary materials. We did not contact the corresponding authors regarding missing or incomplete data. This may have resulted in the exclusion of some potentially eligible studies, which could introduce bias if the missing data differ systematically from the included data. Sixth, most included studies were based on retrospective evidence. Therefore, future research should focus on standardizing image processing methods and conducting prospective, multicenter external validations to develop DL models with enhanced robustness and broader applicability. Finally, this meta-analysis incorporated validation cohorts from the included studies. For studies that provided multiple validation cohorts, each cohort was included as an independent data point in the analysis. This approach may violate the inherent assumption of independence in the standard BME model. While we conducted a sensitivity analysis by extracting only a single validation cohort per original study, this alternative method may also introduce a certain degree of publication bias. Consequently, although DL models show promise for detecting ALNM, the current findings should be interpreted with caution. Future prospective studies with preregistered protocols are needed to verify these results and minimize publication bias.

Conclusions

This study revealed that DL models based on medical imaging showed promising accuracy in detecting ALNM in patients with BC. These findings may provide an evidence-based foundation

XSL•FO
RenderX

for developing intelligent diagnostic tools. However, the findings were primarily based on models developed from single-center, retrospective datasets. Therefore, future studies should adopt a series of strategies, including but not limited to expanding the sample size, integrating multimodal imaging for joint modeling, and conducting prospective validation, to further enhance the performance of DL models while ensuring their applicability and clinical utility.

Acknowledgments

We confirm that no generative artificial intelligence tools were used at any stage in the preparation of this paper. All content, including references and citations, is the original work of the authors, who take full responsibility for its accuracy and integrity.

Funding

This work was supported by the National Natural Science Foundation of China (grant number 82072909) and by the Hospital-Level Research Project of Subei People's Hospital of Jiangsu Province (grant number SBQN25007).

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The complete electronic search strategies and all supplementary figures and tables for the systematic review. It includes the detailed search syntax for the PubMed, Web of Science, Embase, and Cochrane Library databases (Table S1), as well as 56 supplementary figures (Figures S1-S56) that present the supplementary data, analyses, and visualizations supporting the main findings of the review, such as sensitivity analyses, specificity analyses, the summary receiver operating characteristic curve, Deeks funnel plot, and Fagan nomogram.
[DOC File, 5206 KB - jmir_v28i1e77593_app1.doc]

Checklist 1
PRISMA checklist.
[PDF File, 86 KB - jmir_v28i1e77593_app2.pdf]

References

1. Kim J, Harper A, McCormack V, et al. Global patterns and trends in breast cancer incidence and mortality across 185 countries. Nat Med 2025 Apr;31(4):1154-1162. [doi: 10.1038/s41591-025-03502-3] [Medline: 39994475]
2. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA Cancer J Clin 2024;74(1):12-49. [doi: 10.3322/caac.21820] [Medline: 38230766]
3. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021 May;71(3):209-249. [doi: 10.3322/caac.21660] [Medline: 33538338]
4. Schmid P, Cortes J, Dent R, et al. Overall survival with pembrolizumab in early-stage triple-negative breast cancer. N Engl J Med 2024 Nov 28;391(21):1981-1991. [doi: 10.1056/NEJMoa2409932] [Medline: 39282906]
5. Strnad V, Polgár C, Ott OJ, et al. Accelerated partial breast irradiation using sole interstitial multicatheter brachytherapy compared with whole-breast irradiation with boost for early breast cancer: 10-year results of a GEC-ESTRO randomised, phase 3, non-inferiority trial. Lancet Oncol 2023 Mar;24(3):262-272. [doi: 10.1016/S1470-2045(23)00018-9] [Medline: 36738756]
6. Chen XC, Jiao DC, Qiao JH, et al. De-escalated neoadjuvant weekly nab-paclitaxel with trastuzumab and pertuzumab versus docetaxel, carboplatin, trastuzumab, and pertuzumab in patients with HER2-positive early breast cancer (HELEN-006): a multicentre, randomised, phase 3 trial. Lancet Oncol 2025 Jan;26(1):27-36. [doi: 10.1016/S1470-2045(24)00581-3] [Medline: 39612919]
7. Chang JM, Leung JWT, Moy L, Ha SM, Moon WK. Axillary nodal evaluation in breast cancer: state of the art. Radiology 2020 Jun;295(3):500-515. [doi: 10.1148/radiol.2020192534] [Medline: 32315268]
8. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol 2022 Jan;23(1):40-55. [doi: 10.1038/s41580-021-00407-0] [Medline: 34518686]
9. Elemento O, Leslie C, Lundin J, Tourassi G. Artificial intelligence in cancer research, diagnosis and therapy. Nat Rev Cancer 2021 Dec;21(12):747-752. [doi: 10.1038/s41568-021-00399-1] [Medline: 34535775]

XSL•FO
RenderX

10. Swanson K, Wu E, Zhang A, Alizadeh AA, Zou J. From patterns to patients: advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. Cell 2023 Apr 13;186(8):1772-1791. [doi: 10.1016/j.cell.2023.01.035] [Medline: 36905928]
11. Zhan T. DL 101: basic introduction to deep learning with its application in biomedical related fields. Stat Med 2022 Nov 20;41(26):5365-5378. [doi: 10.1002/sim.9564] [Medline: 36041380]
12. Jiang Y, Zhang Z, Wang W, et al. Biology-guided deep learning predicts prognosis and cancer immunotherapy response. Nat Commun 2023 Aug 23;14(1):5135. [doi: 10.1038/s41467-023-40890-x] [Medline: 37612313]
13. Zhong Y, Cai C, Chen T, et al. PET/CT based cross-modal deep learning signature to predict occult nodal metastasis in lung cancer. Nat Commun 2023 Nov 18;14(1):7513. [doi: 10.1038/s41467-023-42811-4] [Medline: 37980411]
14. Park TY, Kwon LM, Hyeon J, Cho BJ, Kim BJ. Deep learning prediction of axillary lymph node metastasis in breast cancer patients using clinical implication-applied preprocessed ct images. Curr Oncol 2024 Apr 18;31(4):2278-2288. [doi: 10.3390/curroncol31040169] [Medline: 38668072]
15. Chen Y, Wang L, Dong X, et al. Deep learning radiomics of preoperative breast MRI for prediction of axillary lymph node metastasis in breast cancer. J Digit Imaging 2023 Aug;36(4):1323-1331. [doi: 10.1007/s10278-023-00818-9]
16. Li WB, Du ZC, Liu YJ, et al. Prediction of axillary lymph node metastasis in early breast cancer patients with ultrasonic videos based deep learning. Front Oncol 2023;13:1219838. [doi: 10.3389/fonc.2023.1219838] [Medline: 37719009]
17. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
18. Gong C, Wu Y, Zhang G, et al. Computer-assisted diagnosis for axillary lymph node metastasis of early breast cancer based on transformer with dual-modal adaptive mid-term fusion using ultrasound elastography. Comput Med Imaging Graph 2025 Jan;119:102472. [doi: 10.1016/j.compmedimag.2024.102472] [Medline: 39612691]
19. Agyekum EA, Kong W, Agyekum DN, et al. Ultrasound derived deep learning features for predicting axillary lymph node metastasis in breast cancer using graph convolutional networks in a multicenter study. Sci Rep 2025 Jul 30;15(1):27796. [doi: 10.1038/s41598-025-13086-0] [Medline: 40738938]
20. Li Y, Chen Z, Ding Z, et al. Applying a multi-task and multi-instance framework to predict axillary lymph node metastases in breast cancer. NPJ Precis Oncol 2025 Jun 18;9(1):195. [doi: 10.1038/s41698-025-00971-0] [Medline: 40533499]
21. Gu C, He Y, Lin J, et al. Preoperative prediction of axillary lymph node metastasis in breast invasive ductal carcinoma patients using a deep learning model based on dynamic contrast-enhanced magnetic resonance imaging: a multicenter study. Gland Surg 2025 Nov 30;14(11):2288-2301. [doi: 10.21037/gs-2025-365] [Medline: 41377877]
22. Dai Y, Liang Y, Sun L, et al. Predicting axillary lymph node metastasis in breast cancer patients using CNN-GCN on DCE-MRI: a multicenter study. BMC Med Imaging 2025 Nov 20;25(1):481. [doi: 10.1186/s12880-025-01988-4] [Medline: 41266995]
23. Sun J, Xi X, Wang M, et al. A deep learning model based on chest CT to predict benign and malignant breast masses and axillary lymph node metastasis. Biomol Biomed 2025 Mar 17;26(1):91-99. [doi: 10.17305/bb.2025.12010] [Medline: 40100034]
24. Wang X, Nie L, Zhu Q, et al. Artificial intelligence assisted ultrasound for the non-invasive prediction of axillary lymph node metastasis in breast cancer. BMC Cancer 2024 Jul 29;24(1):910. [doi: 10.1186/s12885-024-12619-6] [Medline: 39075447]
25. Polat DS, Nguyen S, Karbasi P, et al. Machine learning prediction of lymph node metastasis in breast cancer: performance of a multi-institutional MRI-based 4D convolutional neural network. Radiol Imaging Cancer 2024 May;6(3):e230107. [doi: 10.1148/rycan.230107] [Medline: 38607282]
26. Zhou H, Hua Z, Gao J, et al. Multitask deep learning-based whole-process system for automatic diagnosis of breast lesions and axillary lymph node metastasis discrimination from dynamic contrast-enhanced-MRI: a multicenter study. J Magn Reson Imaging 2024 May;59(5):1710-1722. [doi: 10.1002/jmri.28913] [Medline: 37497811]
27. Liu H, Zou L, Xu N, et al. Deep learning radiomics based prediction of axillary lymph node metastasis in breast cancer. NPJ Breast Cancer 2024 Mar 12;10(1):22. [doi: 10.1038/s41523-024-00628-4] [Medline: 38472210]
28. Guo YJ, Yin R, Zhang Q, et al. MRI-based kinetic heterogeneity evaluation in the accurate access of axillary lymph node status in breast cancer using a hybrid CNN-RNN model. J Magn Reson Imaging 2024 Oct;60(4):1352-1364. [doi: 10.1002/jmri.29225] [Medline: 38205712]
29. Wei W, Ma Q, Feng H, et al. Deep learning radiomics for prediction of axillary lymph node metastasis in patients with clinical stage T1-2 breast cancer. Quant Imaging Med Surg 2023 Aug 1;13(8):4995-5011. [doi: 10.21037/qims-22-1257] [Medline: 37581073]
30. Gao J, Zhong X, Li W, et al. Attention-based deep learning for the preoperative differentiation of axillary lymph node metastasis in breast cancer on DCE-MRI. J Magn Reson Imaging 2023 Jun;57(6):1842-1853. [doi: 10.1002/jmri.28464] [Medline: 36219519]
31. Zhang X, Liu M, Ren W, et al. Predicting of axillary lymph node metastasis in invasive breast cancer using multiparametric MRI dataset based on CNN model. Front Oncol 2022;12:1069733. [doi: 10.3389/fonc.2022.1069733] [Medline: 36561533]

XSL•FO
RenderX

32. Zhang G, Shi Y, Yin P, et al. A machine learning model based on ultrasound image features to assess the risk of sentinel lymph node metastasis in breast cancer patients:applications of scikit-learn and SHAP. Front Oncol 2022;12:944569. [doi: 10.3389/fonc.2022.944569] [Medline: 35957890]
33. Wang Z, Sun H, Li J, et al. Preoperative prediction of axillary lymph node metastasis in breast cancer using CNN based on multiparametric MRI. J Magn Reson Imaging 2022 Sep;56(3):700-709. [doi: 10.1002/jmri.28082] [Medline: 35108415]
34. Sun S, Mutasa S, Liu MZ, et al. Deep learning prediction of axillary lymph node status using ultrasound images. Comput Biol Med 2022 Apr;143:105250. [doi: 10.1016/j.compbiomed.2022.105250] [Medline: 35114444]
35. Santucci D, Faiella E, Gravina M, et al. CNN-based approaches with different tumor bounding options for lymph node status prediction in breast DCE-MRI. Cancers (Basel) 2022 Sep 21;14(19):4574. [doi: 10.3390/cancers14194574] [Medline: 36230497]
36. Li C, Guo Y, Jia L, et al. A convolutional neural network based on ultrasound images of primary breast masses: prediction of lymph-node metastasis in collaboration with classification of benign and malignant tumors. Front Physiol 2022;13:882648. [doi: 10.3389/fphys.2022.882648] [Medline: 35721528]
37. Cattell R, Ying J, Lei L, et al. Preoperative prediction of lymph node metastasis using deep learning-based features. Vis Comput Ind Biomed Art 2022 Mar 7;5(1):8. [doi: 10.1186/s42492-022-00104-5] [Medline: 35254557]
38. Zeng R, Zhang X, Zheng C, et al. Decoupling convolution network for characterizing the metastatic lymph nodes of breast cancer patients. Med Phys 2021 Jul;48(7):3679-3690. [doi: 10.1002/mp.14876] [Medline: 33825207]
39. Liu Z, Ni S, Yang C, et al. Axillary lymph node metastasis prediction by contrast-enhanced computed tomography images for breast cancer patients based on deep learning. Comput Biol Med 2021 Sep;136:104715. [doi: 10.1016/j.compbiomed.2021.104715] [Medline: 34388460]
40. Zhou LQ, Wu XL, Huang SY, et al. Lymph node metastasis prediction from primary breast cancer US images using deep learning. Radiology 2020 Jan;294(1):19-28. [doi: 10.1148/radiol.2019190372] [Medline: 31746687]
41. Zheng X, Yao Z, Huang Y, et al. Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer. Nat Commun 2020 Mar 6;11(1):1236. [doi: 10.1038/s41467-020-15027-z] [Medline: 32144248]
42. Sun Q, Lin X, Zhao Y, et al. Deep learning vs. radiomics for predicting axillary lymph node metastasis of breast cancer using ultrasound images: don't forget the peritumoral region. Front Oncol 2020;10(53):53. [doi: 10.3389/fonc.2020.00053] [Medline: 32083007]
43. Guo X, Liu Z, Sun C, et al. Deep learning radiomics of ultrasonography: identifying the risk of axillary non-sentinel lymph node involvement in primary breast cancer. EBioMedicine 2020 Oct;60:103018. [doi: 10.1016/j.ebiom.2020.103018] [Medline: 32980697]
44. Zhang J, Li L, Zhe X, et al. The diagnostic performance of machine learning-based radiomics of DCE-MRI in predicting axillary lymph node metastasis in breast cancer: a meta-analysis. Front Oncol 2022;12:799209. [doi: 10.3389/fonc.2022.799209] [Medline: 35186739]
45. Chen C, Qin Y, Chen H, Zhu D, Gao F, Zhou X. A meta-analysis of the diagnostic performance of machine learning-based MRI in the prediction of axillary lymph node metastasis in breast cancer patients. Insights Imaging 2021 Nov 3;12(1):156. [doi: 10.1186/s13244-021-01034-1] [Medline: 34731343]
46. Eldaly AS, Avila FR, Torres-Guzman RA, et al. Radiomics and artificial intelligence in predicting axillary lymph node metastasis in breast cancer: a systematic review. Curr Med Imaging 2023;19(6):564-578. [doi: 10.2174/1573405618666220822093226] [Medline: 35996255]
47. Liu CJ, Zhang L, Sun Y, et al. Application of CT and MRI images based on an artificial intelligence algorithm for predicting lymph node metastasis in breast cancer patients: a meta-analysis. BMC Cancer 2023 Nov 22;23(1):1134. [doi: 10.1186/s12885-023-11638-z] [Medline: 37993845]
48. Gong X, Guo Y, Zhu T, Peng X, Xing D, Zhang M. Diagnostic performance of radiomics in predicting axillary lymph node metastasis in breast cancer: a systematic review and meta-analysis. Front Oncol 2022;12:1046005. [doi: 10.3389/fonc.2022.1046005] [Medline: 36518318]

Abbreviations

AI: artificial intelligence
ALNM: axillary lymph node metastasis
AUC: area under the summary receiver operating characteristic curve
BC: breast cancer
BME: bivariate mixed effects
CT: computed tomography
DL: deep learning
DOR: diagnostic odds ratio
LR+: positive likelihood ratio
LR−: negative likelihood ratio
ML: machine learning

XSL•FO
RenderX

MRI: magnetic resonance imaging
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO : Prospective Register of Systematic Reviews
RoB: risk of bias
TP: true positive

Edited by A Coristine; submitted 16.May.2025; peer-reviewed by AM Ahmadzadeh, M Beltrão; revised version received 27.Feb.2026; accepted 27.Feb.2026; published 16.Apr.2026.

Please cite as:
Wang X, Li T, Wang X, Fu D
Accuracy of Deep Learning for Detecting Axillary Lymph Node Metastasis in Breast Cancer: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e77593
URL: https://www.jmir.org/2026/1/e77593
doi:10.2196/77593

© Xueying Wang, Tiantian Li, Xiaohang Wang, Deyuan Fu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Machine Learning in Predicting the Risk of Esophagogastric Variceal Bleeding Among Patients With Liver Cirrhosis: Systematic Review and Meta-Analysis

Yuan Lian[1*], MS; Xinping Qiu[1], MBBS; Geli Wang[1], MBBS; Nannan Liu[1], MBBS; Jing Zhao[1], MBBS; Shanshan Song[1], MBBS; Shiqi Wang[1], MS; Mingjun Sun[1,2*], MD

[1]Department of Gastroenterology, Fuxin Mining General Hospital of Liaoning Health Industry Group, No.43,Health Care Street,Haizhou District, Fuxin, China

[2]Department of Gastroenterology, The First Hospital of China Medical University, Shenyang, China

[*]these authors contributed equally

Corresponding Author:
Yuan Lian, MS
Department of Gastroenterology, Fuxin Mining General Hospital of Liaoning Health Industry Group, No.43,Health Care Street,Haizhou District, Fuxin, China

Abstract

Background: Liver cirrhosis (LC) can lead to several complications. Esophageal variceal bleeding (EVB) and esophagogastric variceal bleeding (EGVB) are particularly severe, leading to a high risk of mortality. Early identification of esophageal varices and esophagogastric varices is essential. Several studies have constructed prediction models for EVB and EGVB in patients with LC. However, robust systematic evidence to prove their performance is lacking.

Objective: We included original studies that developed prediction models for esophageal or gastric variceal bleeding in patients with LC under different modeling variables. This study aimed to review the predictive performance of various models for EVB or EGVB in patients with LC, providing insights into the development or updating of simplified scoring tools in the future.

Methods: PubMed, Web of Science, Embase, and the Cochrane Library were searched up to August 21, 2024, to collect original full-text studies on machine learning (ML) in the prediction of EVB and EGVB in patients with LC. The models were evaluated using the Prediction Model Risk of Bias Assessment Tool (PROBAST). Subgroup analyses were carried out based on the modeling variables.

Results: In total, 21 studies were included, with 7011 patients with LC, among whom 1412 (20.14%) developed EVB and 733 (10.45%) developed EGVB. The meta-analysis results suggested that the pooled c-index, sensitivity, and specificity of the prediction model for predicting EVB in the validation set were 0.85 (95% CI 0.77 - 0.92), 0.93 (95% CI 0.87 - 0.96), and 0.66 (95% CI 0.46 - 0.82), respectively. The pooled c-index, sensitivity, and specificity of the prediction model for predicting EGVB in the validation set were 0.89 (95% CI 0.85 - 0.94), 0.77 (95% CI 0.66 - 0.85), and 0.81 (95% CI 0.67 - 0.90), respectively. The subgroup analysis based on modeling variables revealed that, for predicting EVB, the c-index in the validation set was 0.84 (95% CI 0.80 - 0.88) for models based on clinical features, 0.82 (95% CI 0.69 - 0.96) for radiomics-based models, 0.78 (95% CI 0.67 - 0.89) for models based on radiomics and clinical features, and 0.97 (95% CI 0.95 - 1.00) for models based on endoscopic features. Subgroup analyses based on modeling variables revealed that, for predicting EGVB, the c-index in the validation set was 0.91 (95% CI 0.86 - 0.96) for models based on clinical features and 0.85 (95% CI 0.75 - 0.96) for models based on radiomics and clinical features.

Conclusions: ML methods are feasible for predicting EVB and EGVB in patients with LC. Nevertheless, the number of included original studies is limited. In the future, more studies with larger sample sizes are needed to promote the application of ML in the early assessment of EVB and EGVB in patients with LC in clinical practice.

Trial Registration: PROSPERO CRD42024585100; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024585100

(J Med Internet Res 2026;28:e78203) doi:10.2196/78203

KEYWORDS

liver cirrhosis; hepatic; prediction model; esophageal variceal bleeding; esophagogastric variceal bleeding; radiomics; machine learning

XSL•FO
RenderX

Introduction

Liver cirrhosis (LC) often occurs after long-term inflammation, resulting in fibrotic tissue and regenerative nodules and, subsequently, portal hypertension [1]. Approximately 1 million deaths worldwide are attributable to LC every year [2]. LC is the 11th leading cause of death worldwide and the third leading cause of death among individuals aged 45 to 64 years. Together with liver cancer, it represents 3.5% of global deaths [2]. LC contributed to 2.4% of deaths worldwide in 2019 [3]. LC is a notable cause of both morbidity and mortality in patients with chronic liver diseases, leading to hepatocellular carcinoma (HCC) and liver decompensation. LC-induced portal hypertension is associated with complications such as ascites, hepatic encephalopathy, acute kidney injury, hepatorenal syndrome, and portal hypertension–related bleeding, such as esophagogastric variceal bleeding (EGVB) [1]. Thus, LC becomes a major health care burden worldwide.

The incidence of EGVB is high among the primary complications of LC. It is the second most prevalent complication in patients with LC, after ascites. EGVB, a frequent cause of bleeding, is regarded as a medical emergency. This complication, even when well managed, is associated with a mortality rate of approximately 20% within 6 weeks. If an infection occurs, the risk of mortality increases further [1]. EGVB is a critical condition and often recurs. In severe cases, it may be life threatening [4]. Currently, the treatment methods for EGVB are diverse. For stable patients after volume resuscitation, cautious transfusion is advised, with a transfusion threshold of 7 g/dL. For hemostatic treatment, vasoactive agents such as octreotide, somatostatin, or terlipressin, together with endoscopic interventions such as variceal ligation, should be used. A transjugular intrahepatic portosystemic shunt is used for resuscitation when uncontrolled bleeding occurs. It is advisable to administer prophylactic antibiotics, such as ceftriaxone, for 7 days or until the patient is discharged. For patients at high risk of recurrent bleeding, specifically those with a Child-Turcotte-Pugh B score between 8 and 9 points and active bleeding or a Child-Turcotte-Pugh C score ranging from 10 to 13 points, pre-emptive transjugular intrahepatic portosystemic shunt within 72 hours of admission could prolong their survival. For patients who are not at high risk, β-blockers combined with variceal ligation are recommended [1]. These conditions not only intensify patients' pain but also shorten their lifespan and increase their economic burden. Thus, early identification of the risk of esophagogastric varices (EGV) in patients with LC is crucial for preventing EGVB, lowering its threat to health and life and the economic burden on families.

Endoscopy, which is interventional, is the primary tool for diagnosing esophageal variceal bleeding (EVB) and EGVB. As the gold standard for detecting EVB and EGVB, endoscopy is also recommended for diagnosing and screening for EGV and assessing the risk of bleeding [5]. Nonetheless, endoscopic examination is an invasive procedure, and a portion of patients may refuse this examination due to low tolerance. Moreover, repeated procedures might diminish long-term adherence. Furthermore, this examination may also induce EGVB in low-risk patients. Thus, recent studies have focused on the development of noninvasive or minimally invasive tools to identify the risks of EGV at an early stage. However, effective tools for the early prediction of EGV remain lacking. Recently, as machine learning (ML) has been increasingly used in medical diagnostics, some studies have investigated the feasibility of using ML techniques to evaluate the risk of EGVB in patients with LC. Nevertheless, their detection performance varies notably with the modeling approach or the variables used. Currently, there is insufficient evidence to demonstrate the feasibility of using ML to identify EVB and EGVB in patients with LC. Thus, this study included original studies that developed prediction models for esophageal or gastric variceal bleeding in patients with LC under different modeling variables. The aim of this study was to review the performance of ML in the early identification of EVB and EGVB in patients with LC, providing evidence for the advancement and updating of early risk assessment tools for EGV.

Methods

Study Registration

This study was performed based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Checklist 1) and was prospectively registered with PROSPERO (ID CRD42024585100).

Eligibility Criteria

Before screening the original studies, we established comprehensive eligibility criteria (Table 1).

XSL•FO
RenderX

Table . Eligibility criteria for screening the original studies for this systematic review.

Items	Inclusion criteria	Exclusion criteria
Population	Patients with LC[a] as the study participants.	Studies that did not strictly distinguish between patients with liver cirrhosis and those with hepatocellular carcinoma.
Intervention	Studies that developed prediction models for the risk of EVB[b] or EGVB[c].	Studies that solely explored risk factors without creating a complete prediction model, studies that only centered on validated mature scales, studies on the predictive accuracy of a single factor for outcomes, and studies that did not strictly distinguish EVB or EGVB from other upper gastrointestinal bleeding.
Comparison	None	None
Outcome	Studies reporting any of the following outcome measures to evaluate the prediction model: c-index, sensitivity, specificity, accuracy, recall, precision, confusion matrix, F_1-score, and calibration curve.	None
Study design	Studies adopting case-control, cohort, and cross-sectional designs; and original studies reported in English.	Conference abstracts that were published without peer review and for studies that repeatedly trained and validated models based on the same dataset, only the most recently published study was included.

[a]LC: liver cirrhosis.

[b]EVB: esophageal variceal bleeding

[c]EGVB: esophagogastric variceal bleeding.

Data Sources and Search Strategy

PubMed, Web of Science, Embase, and the Cochrane Library were searched up to August 21, 2024. Both subject terms and free terms were used to design search terms. No restrictions regarding region or period were imposed. Multimedia Appendix 1 provides the detailed search strategy.

Study Selection and Data Extraction

Articles retrieved from the databases were imported into EndNote (Clarivate Analytics). First, duplicates were removed. By reading the titles and abstracts, potentially eligible original studies were chosen. The full texts of these studies were then downloaded and read to determine eligible articles. Subsequently, a standardized spreadsheet was developed to extract data (eg, study title, first author, publication year, and study type) from the included articles. Two investigators independently screened the studies, extracted the data, and then cross-checked the results. Any disagreements were resolved through consultation with a third investigator.

Risk of Bias Assessment

The risk of bias (ROB) in the included studies was appraised using the Prediction Model Risk of Bias Assessment Tool (PROBAST). This tool encompasses 4 domains: participant, predictor variable, outcome, and statistical analysis. Questions in each domain could be answered with 3 responses: yes or probably yes, no or probably no, and no information. A domain was deemed to have high ROB if any question was answered with no or probably no. A domain was deemed to have low ROB if all questions were answered with yes or probably yes. The overall ROB was low if all domains were considered low ROB. The overall ROB was high if at least one domain was considered to have a high ROB. Two investigators independently appraised the ROB using the PROBAST, and their results were cross-checked. Disagreements, if any, were resolved by a third investigator.

Synthesis Methods

For evaluating the overall accuracy, a meta-analysis of the c-indices of ML models was performed. In certain original studies, the 95% CI and SEs for the c-index were missing. We referred to the study by Debray et al [6] to estimate the SEs. Heterogeneity across studies was assessed using the I^2 statistic. A random-effects model was used if I^2 was greater than 50%, whereas a fixed effects model was adopted if I^2 was less than 50%. Additionally, the sensitivity and specificity, derived from the diagnostic quadrangle table, were pooled through bivariate mixed effects models. However, such a table was not available in many of the original studies. In this case, the number of cases with sensitivity, specificity, and precision was combined to compute the estimates. Sensitivity and specificity were extracted based on the optimal Youden index. This meta-analysis was performed in Stata (version 15.0; StataCorp LLC).

Results

Study Selection

In total, 1215 articles were retrieved from the databases. Following the removal of 144 (11.85%) duplicates, 1071 (88.15%) articles were left. In total, 1041(85.68%) articles were excluded due to unrelated topics, including meta-analyses, reviews, guidelines, animal experiments, case reports, letters,

XSL•FO
RenderX

and registration protocols. The full texts of the remaining 30 (2.47%) articles were then downloaded and read. We further excluded 4 (0.33%) conference abstracts that had not undergone peer review, 1 (0.08%) study that analyzed only risk factors without developing a complete ML model, 2 (0.16%) studies without an accuracy assessment of model outcomes, and 2 (0.16%) studies published based on the same dataset of another 2 studies. Finally, 21 (1.73%) articles were included [7-27] (Figure 1).

Figure 1. .PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of the literature screening process.

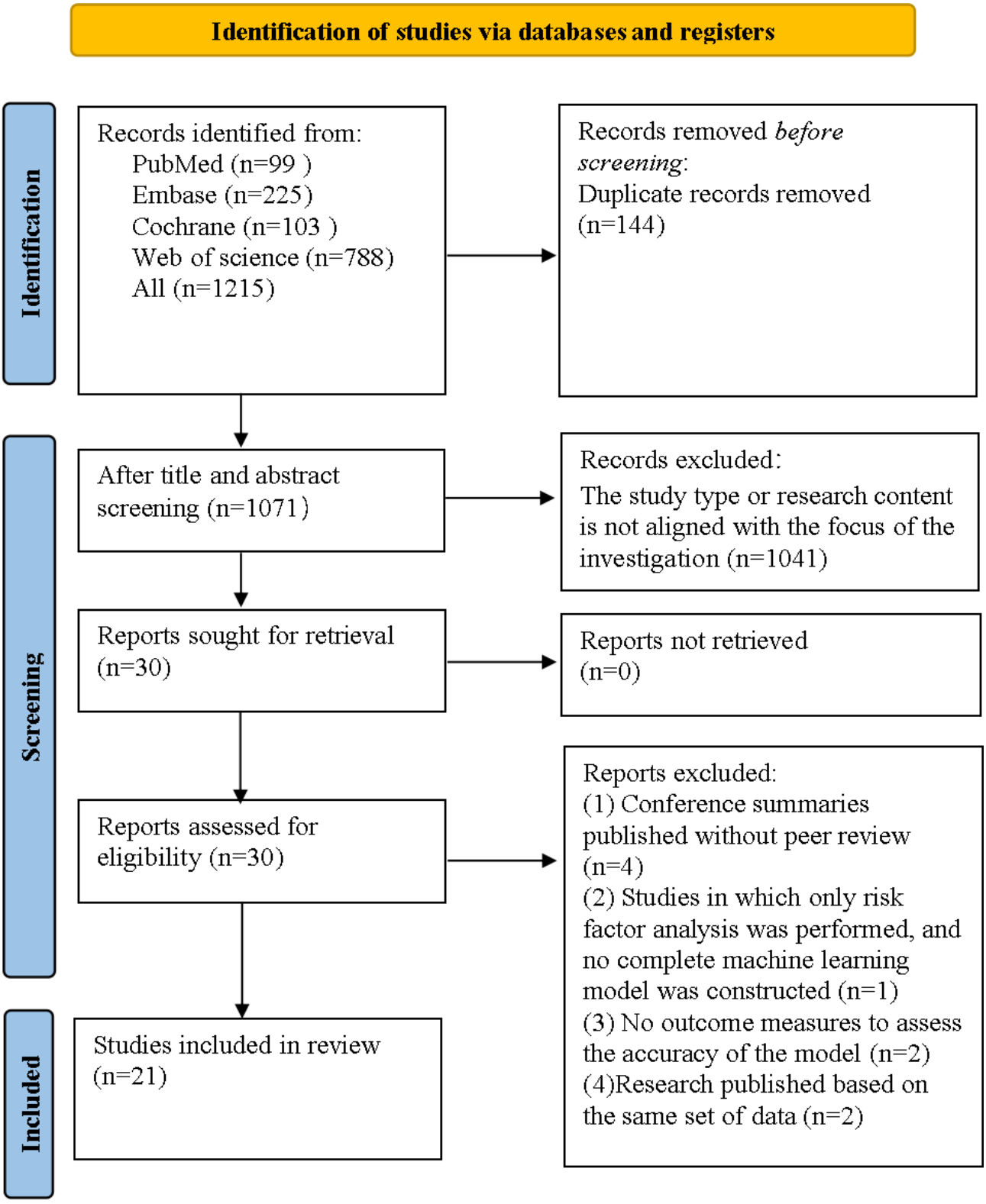

XSL•FO
RenderX

Study Characteristics

The included studies were published between 2014 and 2024, involving a total of 7011 patients with LC. Among these studies, 12 studies included 1412 cases of EVB [8,10-15,19-22,25], and 9 studies included 733 cases of EGVB [7-9,11,12,16,17,26,27]. These studies were conducted in 4 countries: 18 studies in China [7-20,23,25-27], 1 study in Egypt [21], 1 study in Iran [22], and 1 study in South Korea [24]. Among the included studies, 14 were case-control studies [7,12-16,18-20,22,24-27], 6 were cohort studies [8-11,21,23], and 1 was a cross-sectional study [15]. There were 16 single-center studies [7-9,12-17,19-23,25,27] and 5 multicenter studies [10,11,18,24,26]. Nine articles [7-9,11,12,16,17,26,27] centered on EGVB, and 12 articles focused on EVB [10,13-15,18-25]. Fourteen studies clearly outlined the method used to generate the validation set. Only 3 studies used external validation [10,18,19]. Two studies used time-based split [8,25], while the remaining studies conducted internal validation with random sampling. The models involved were mainly logistic regression models, while only a few studies constructed other types of models. Fifteen models were constructed based on clinical features [7-9,11,15-25]. Three models were based on radiomics [13,26,27]. Four models were based on radiomics and clinical features [12,14,26,27], and 1 model was based on endoscopic features in 1 study [10]. The detailed results are presented in Table 2.

XSL•FO
RenderX

Table . Basic characteristics of the included studies.

First author	Year of publication	Country of authors	Study type	Patient source	Predicted events	Event cases (n=2145), n (30.59%)	Cases (n=7011), n	Cases in training set (n=5554), n	Generation method of validation set	Cases in the validation set (n=1457), n	Model type	Modeling variables
Liu et al [7]	2022	China	Case-control study	Single center	EGVB[a]	77	330	220	Random sampling	110	LR[b]	Clinical features (emotional stimuli, improper diet, overwork, lower temperature, and increased abdominal pressure)
Zhang et al [8]	2022	China	Prospective cohort study	Single center	EGVB	52	172	128	Time-based split	44	Cox[c]	Clinical features (vPPG[d], PLT[e], albumin, and INR)[f]
Tan et al [9]	2023	China	Prospective study (cohort)	Single center	EGVB	74	230	160	Random sampling	70	LR	Clinical features (right liver lobe volume, diameters of maximum esophagogastric varices, and portal vein system)
Wang et al [10]	2023	China	Retrospective cohort study	Multicenter	EVB[g]	121	502	275	Random sampling and external validation	227	DL[h], XGBoost[i], GLM[j], GBM[k], and RF[l]	Endoscopic features (EV[m] bleeding image)

XSL•FO
RenderX

First author	Year of publication	Country of authors	Study type	Patient source	Predicted events	Event cases (n=2145), n (30.59%)	Cases (n=7011), n	Cases in training set (n=5554), n	Generation method of validation set	Cases in the validation set (n=1457), n	Model type	Modeling variables
Hou et al [11]	2023	China	Prospective cohort study	Multicenter	EGVB	131	1100	999	Random sampling	101	Cox	Clinical features (gender, alcohol and smoking history, decompensation, ascites, location and size of varicose veins, ALT[n], GGT[o], HCT[p], and NLR[q] levels, and RBC[r] count)
Luo et al [12]	2023	China	Retrospective case-control study	Single center	EGVB	88	211	149	Random sampling	62	LR	Radiomics and clinical features (albumin, fibrinogen, portal vein thrombosis, aspartate aminotransferase, and spleen thickness)
Meng et al [13]	2021	China	Retrospective case-control study	Single center	EVB	39	173	121	Random sampling	52	Cox	Radiomics
Liu et al [14]	2022	China	Retrospective case-control study	Single center	EVB	69	317	222	Random sampling	95	LR	Radiomics and clinical features (hemoglobin)

XSL•FO
RenderX

First author	Year of publication	Country of authors	Study type	Patient source	Predicted events	Event cases (n=2145), n (30.59%)	Cases (n=7011), n	Cases in training set (n=5554), n	Generation method of validation set	Cases in the validation set (n=1457), n	Model type	Modeling variables
Xu et al [15]	2024	China	Case-control study	Single center	EVB	114	241	160	Random sampling	81	LR	Clinical features (spleen length and platelet distribution width PDW[s])
Lin et al [17]	2024	China	Retrospective cross-sectional study	Single center	EGVB	70	210	210	None[t]		LR	Clinical features (LSM[u] and GBS[v] scoring)
Zhang et al [18]	2023	China	Retrospective analysis	Multicenter	EVB	77	714	527	External validation	187	Cox	Clinical features (the size of varices, red wale marks, ascites, spleen thickness, γ-glutamyltransferase, and hematocrit)
Yang et al [19]	2023	China	Retrospective case-control study	Single center	EVB	98	200	140	External validation	60	LR	Clinical features (SSM[w] and LSM)
Fu et al [20]	2023	China	Case-control study	Single center	EVB	40	78	78	None		LR	Clinical features (age, gender, platelet count, Child-Pugh classification, and normalized iodine density)

XSL•FO
RenderX

First author	Year of publication	Country of authors	Study type	Patient source	Predicted events	Event cases (n=2145), n (30.59%)	Cases (n=7011), n	Cases in training set (n=5554), n	Generation method of validation set	Cases in the validation set (n=1457), n	Model type	Modeling variables
Samy et al [21]	2022	Egypt	Prospective cohort study	Single center	EVB	42	85	85	None		LR	Clinical features (main coronary vein size, paraesophageal collateral, short gastric collateral, gastric varicosities, child class C, and size of EVs >4 mm)
Salahshour et al [22]	2020	Iran	Retrospective case-control study	Single center	EVB	50	124	124	None		LR	Clinical features (presence of SGC[x], SGC size, presence of EV, and MELD[y] score)
Li et al [16]	2022	China	Retrospective case-control study	Single center	EGVB	85	142	142	None		LR	Clinical features (spleen volume expansion rate, serum sodium level, hemoglobin level, and prothrombin time)
Xu et al [23]	2014	China	Prospective cohort study	Single center	EVB	51	416	416	None		LR	Clinical features (ΔMELD, LGVV[z], and LGVBFD)[aa]

XSL•FO
RenderX

First author	Year of publication	Country of authors	Study type	Patient source	Predicted events	Event cases (n=2145), n (30.59%)	Cases (n=7011), n	Cases in training set (n=5554), n	Generation method of validation set	Cases in the validation set (n=1457), n	Model type	Modeling variables
Heo et al [24]	2019	Korea	Retrospective case-control study	Multicenter	EVB	61	262	262	None		LR	Clinical features (ARFI[ab] velocity, spleen diameter, platelet count, ASPS[ac], Child-Pugh class, EV status, portal hypertensive gastropathy, etiology of cirrhosis, age)
Liu et al [25]	2022	China	Retrospective case-control study	Single center	EVB	650	1099	823	Time-based split	276	LR	Clinical features (PLT, Hb[ad], ALB[ae]/GLB[af], FBG[ag], Cl[ah], and CTPD[ai])
Yang et al [26]	2019	China	Retrospective case-control study	Multicenter	EGVB	91	295	236	Random sampling	59	RF	Radiomics and clinical features (sex, portal vein thrombosis, and portal hypertension)

XSL•FO
RenderX

First author	Year of publication	Country of authors	Study type	Patient source	Predicted events	Event cases (n=2145), n (30.59%)	Cases (n=7011), n	Cases in training set (n=5554), n	Generation method of validation set	Cases in the validation set (n=1457), n	Model type	Modeling variables
Zhang et al [27]	2023	China	Retrospective study	Single center	EGVB	65	110	77	Random sampling	33	LR	Radiomics and clinical features (PLT)

[a]EGVB: esophagogastric variceal bleeding.
[b]LR: logistic regression.
[c]Cox: Cox regression.
[d]vPPG: virtual portal pressure gradient.
[e]PLT: platelet.
[f]INR: international normalized ratio.
[g]EVB: esophageal variceal bleeding.
[h]DL: deep learning.
[i]XGBoost: Extreme Gradient Boosting.
[j]GLM: general linear regression.
[k]GBM: gradient boost machine.
[l]RF: random forest.
[m]EV: esophageal variceal.
[n]ALT: alanine aminotransferase.
[o]GGT: γ-glutamyl transferase.
[p]HCT: hematocrit.
[q]NLR: neutrophil-lymphocyte ratio.
[r]RBC: red blood cell.
[s]PDW: platelet distribution width.
[t]In the generation method of the validation set, “None” indicates that there is no independent validation set.
[u]LSM: liver stiffness measurement.
[v]GBS: Glasgow-Blatchford score.
[w]SSM: spleen stiffness measurement.
[x]SGC: short gastric.
[y]MELD: Model for End-stage Liver Disease.
[z]LGVV: left gastric vein blood flow velocity.
[aa]LGVBFD: left gastric vein blood flow direction.
[ab]ARFI: acoustic radiation force impulse.
[ac]ASPS: ARFI–spleen diameter-to-platelet ratio scores.
[ad]Hb: hemoglobin.
[ae]ALB: albumin.
[af]GLB: globulin.
[ag]FBG: fasting blood glucose.
[ah]Cl: serum chloride.
[ai]CTPD: computed tomography portal vein diameter.

ROB Assessment

The ROB in the included models was assessed. Fourteen models were from case-control studies, introducing a high ROB in participant selection. Fifteen models were built based on clinical features, resulting in a higher ROB in the predictive factors. Three models were based on radiomics, and thus, it was challenging to assess ROB. In terms of results, the included studies clearly outlined diagnostic criteria for EVB and EGVB, and no evidence of high ROB was noted in these diagnostic standards.

Of the models included, the number of cases was small in 9 training sets, resulting in a higher ROB. For 7 models, the number of cases in the validation set could not be estimated, leading to an unclear ROB. Regarding the handling of missing values, missing values were deleted in 6 models, leading to a high ROB. Only one study applied univariate analysis to select

XSL•FO
RenderX

variables for the models, and 7 studies did not provide details on their variable selection methods. Among the included models, only one model was cross-validated, and thus, a high ROB was identified in the study. The full assessment results are available in Figure 2.

Figure 2. Risk of bias assessment results of the included studies by the Prediction Model Risk of Bias Assessment Tool (PROBAST).

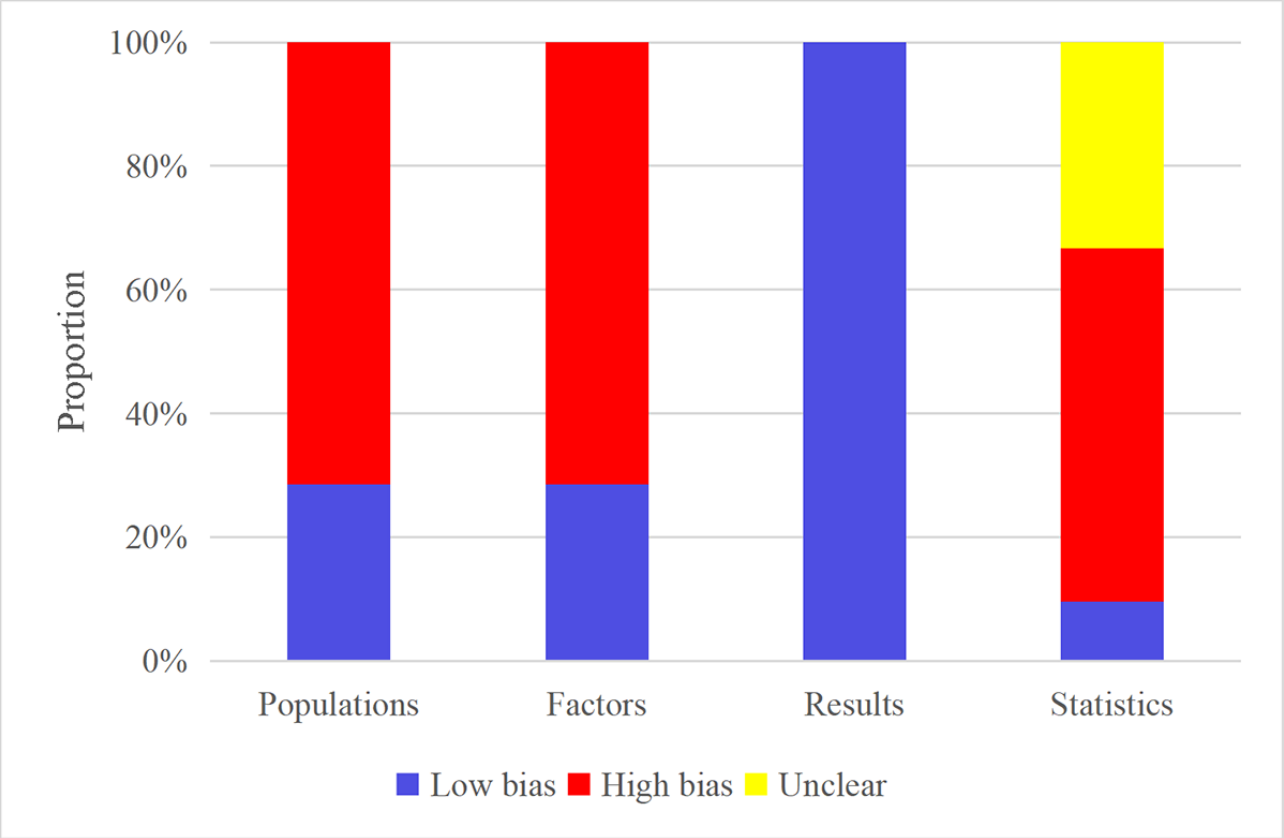

Meta-Analysis

Esophageal Variceal Bleeding

Pooled Results

Twelve models were created for predicting EVB. The meta-analysis was performed using a random-effects model. The pooled c-index, sensitivity, and specificity in the training set were 0.88 (95% CI 0.83 - 0.92), 0.93 (95% CI 0.83 - 0.97), and 0.75 (95% CI 0.65 - 0.84), respectively (Figures 3 and 4). The pooled c-index, sensitivity, and specificity in the validation set were 0.85 (95% CI 0.77 - 0.92), 0.93 (95% CI 0.87 - 0.96), and 0.66 (95% CI 0.46 - 0.82), respectively (Figures 5 and 6).

XSL•FO
RenderX

Figure 3. Forest plot for meta-analysis of c-index of machine learning for predicting esophageal variceal bleeding and esophagogastric variceal bleeding in patients with liver cirrhosis in the training set [7-9,11,12,13-16,17-19,20,21,22-24,26,27].

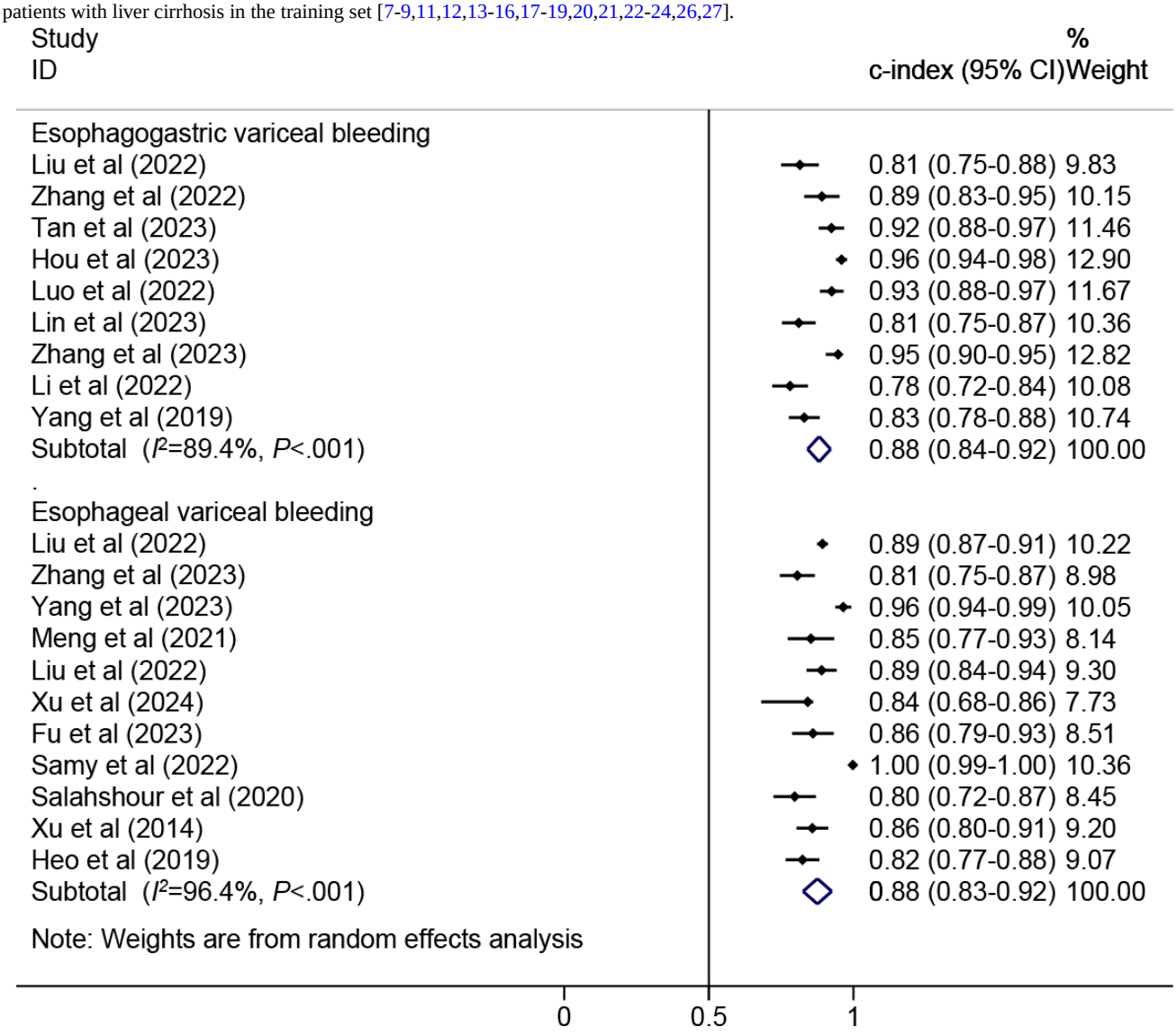

XSL•FO
RenderX

Figure 4. Forest plot for meta-analysis of sensitivity and specificity of machine learning for predicting esophageal variceal bleeding in patients with liver cirrhosis in the training set [14,15,18-25].

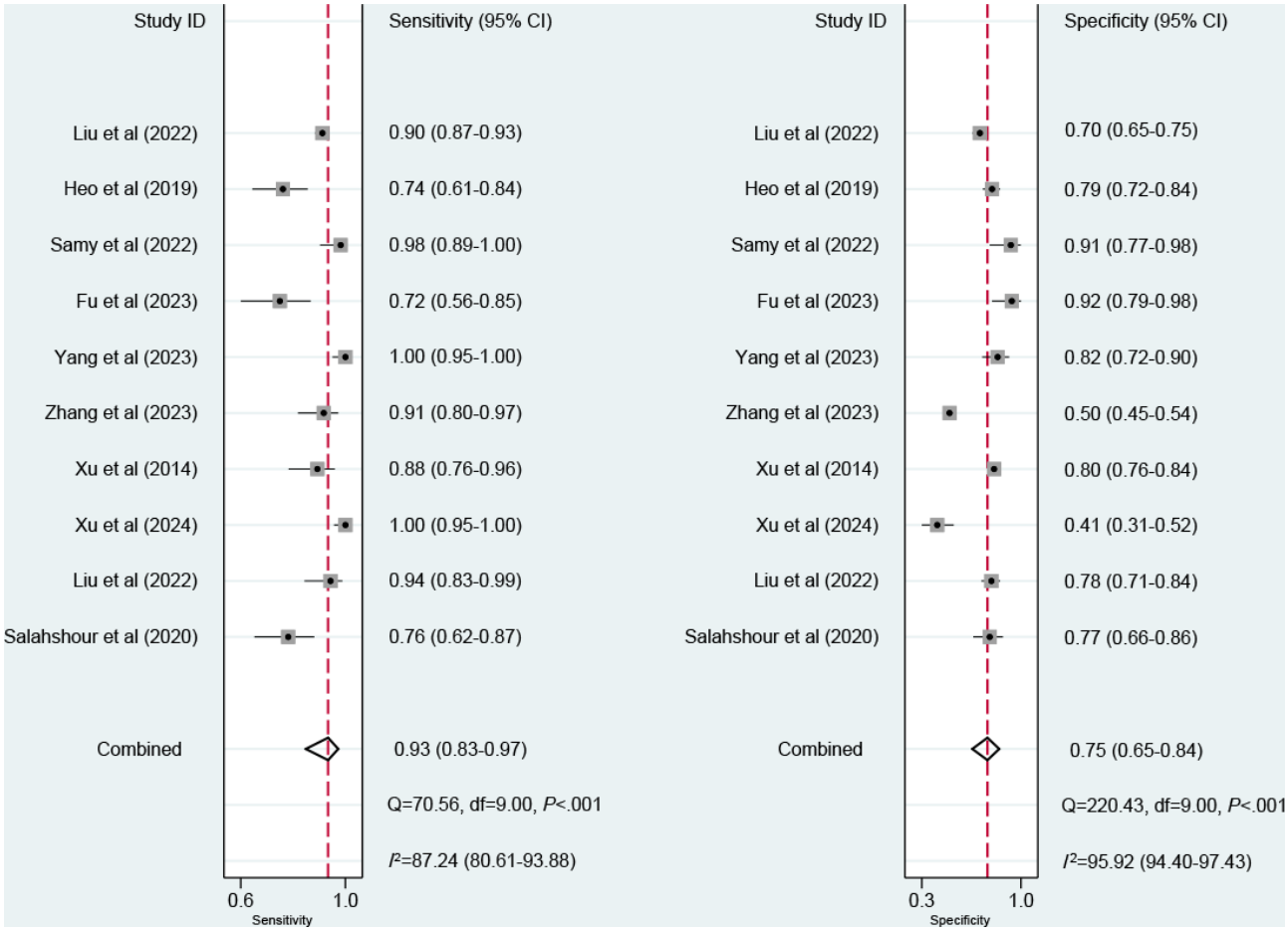

XSL•FO
RenderX

Figure 5. Forest plot for meta-analysis of c-index of machine learning for predicting esophageal variceal bleeding and esophagogastric variceal bleeding in patients with liver cirrhosis in the validation set [7-15,18,19,25-27].

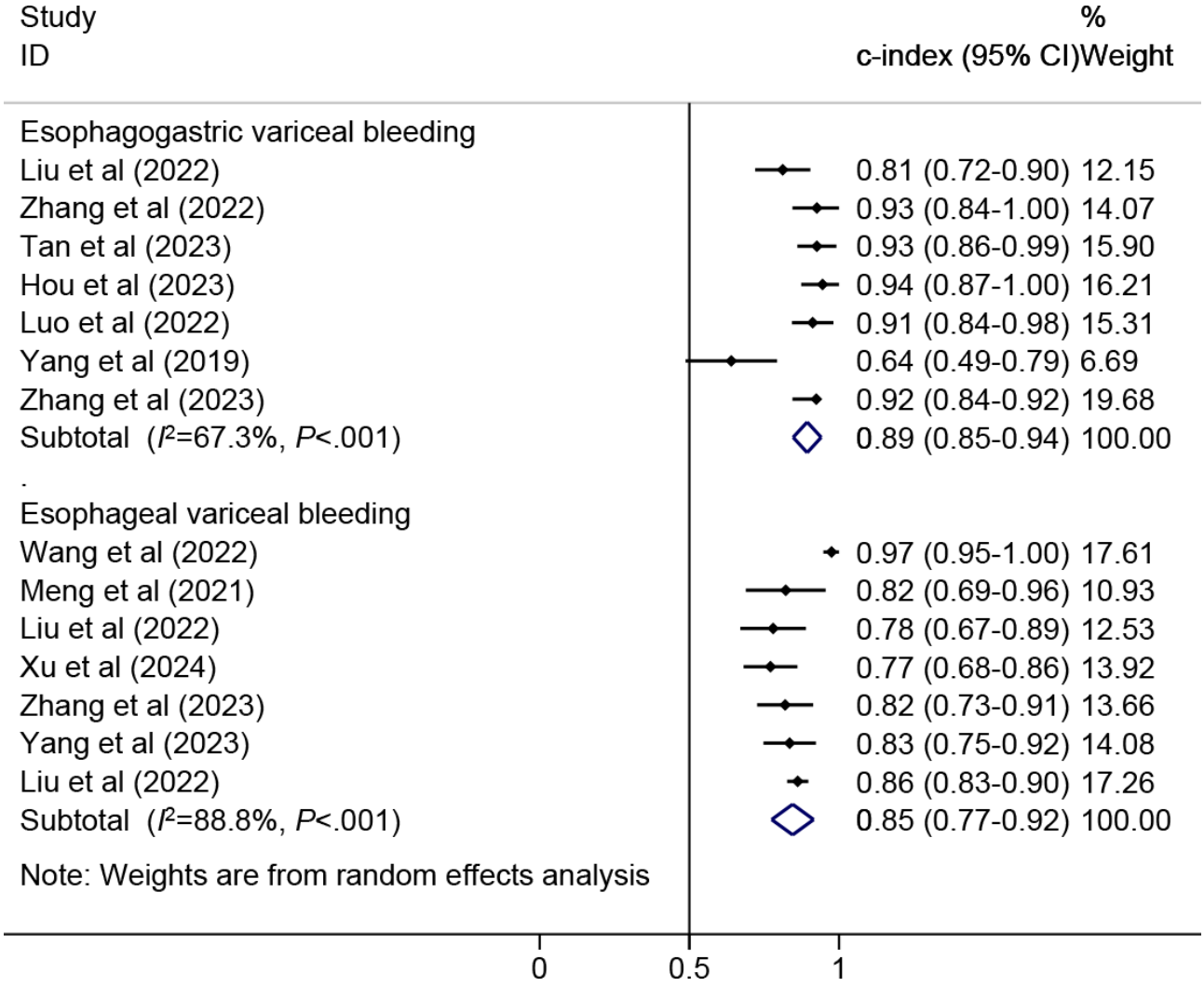

XSL•FO
RenderX

Figure 6. Forest plot for meta-analysis of sensitivity and specificity of machine learning for predicting esophageal variceal bleeding in patients with liver cirrhosis in the validation set [10,14,15,18,19].

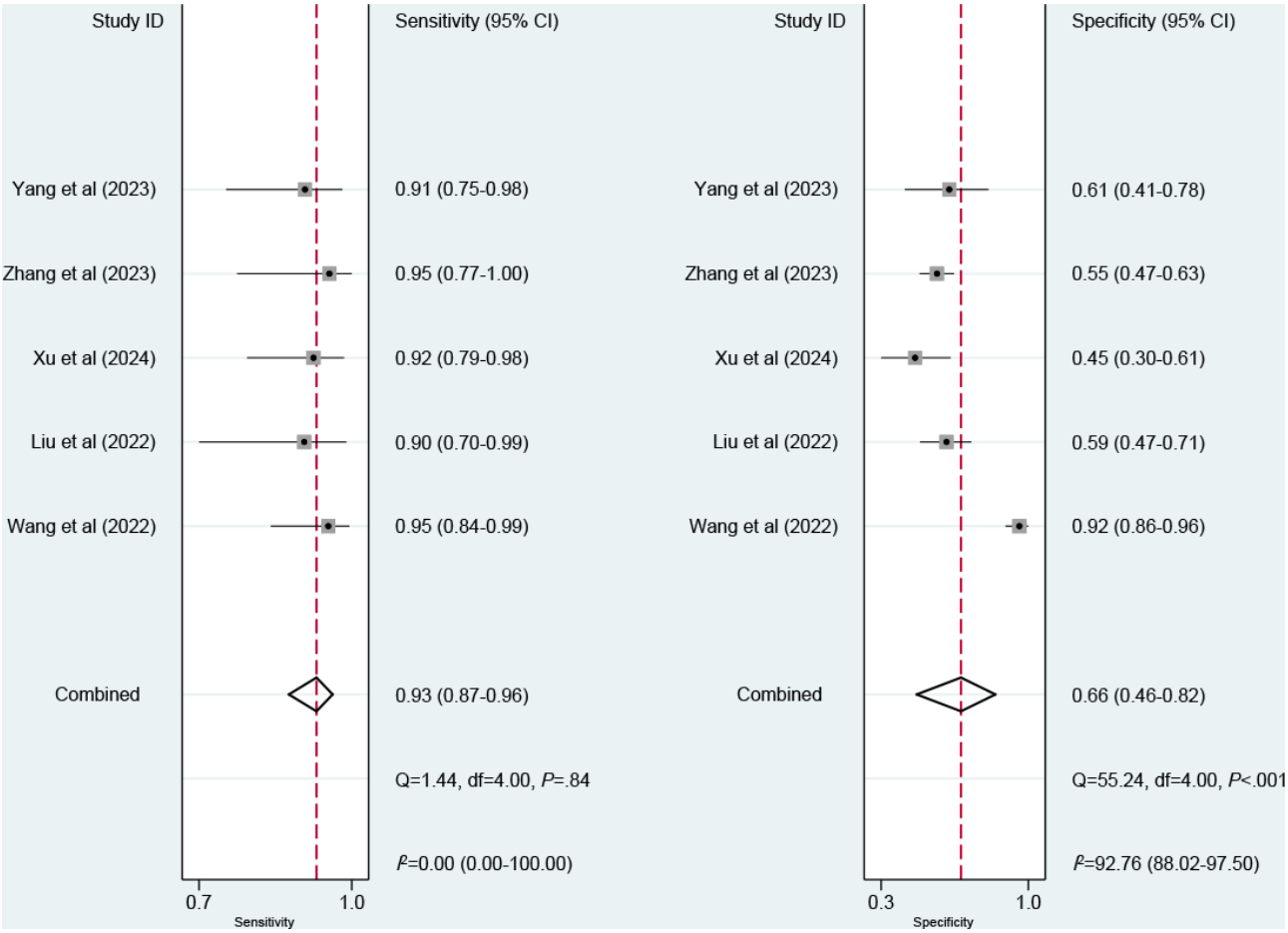

Subgroup Analysis

A subgroup analysis was performed based on modeling variables and the generation method of the validation set. In the subgroup analysis by model variables, models were grouped based on clinical features, radiomics-based models, models based on radiomics and clinical features, and models based on endoscopic features. In the training set, the c-index was 0.88 (95% CI 0.82 - 0.93) for models based on clinical features, 0.85 (95% CI 0.77 - 0.93) for radiomics-based models, and 0.89 (95% CI 0.84 - 0.94) for models based on radiomics and clinical features (Multimedia Appendix 2). In the validation set, the c-index was 0.84 (95% CI 0.80 - 0.88) for models based on clinical features, 0.82 (95% CI 0.69 - 0.96) for radiomics-based models, 0.78 (95% CI 0.67 - 0.89) for models based on radiomics and clinical features, and 0.97 (95% CI 0.95 - 1.00) for models based on endoscopic features (Multimedia Appendix 3). In the subgroup analysis by the generation method of the validation set, models were grouped as random sampling, external validation, and time-based split. The c-index was 0.78 (95% CI 0.72 - 0.85) for models based on random sampling, 0.88 (95% CI 0.77 - 1.00) for external validation, and 0.86 (95% CI 0.83 - 0.90) for time-based split (Multimedia Appendix 4).

Esophagogastric Variceal Bleeding

Pooled Results

Nine models were created for predicting EGVB. The meta-analysis was conducted using the random-effects model. The pooled results revealed that the c-index, sensitivity, and specificity in the training set were 0.88 (95% CI 0.84 - 0.92), 0.90 (95% CI 0.78 - 0.96), and 0.79 (95% CI 0.58 - 0.91), respectively (Figures 3 and 7). The pooled c-index, sensitivity, and specificity in the validation set were 0.89 (95% CI 0.85 - 0.94), 0.77 (95% CI 0.66 - 0.85), and 0.81 (95% CI 0.67 - 0.90), respectively (Figures 5 and 8).

XSL•FO
RenderX

Figure 7. Forest plot for meta-analysis of sensitivity and specificity of machine learning for predicting esophagogastric variceal bleeding in patients with liver cirrhosis in the training set [7,9,12,16,17,27].

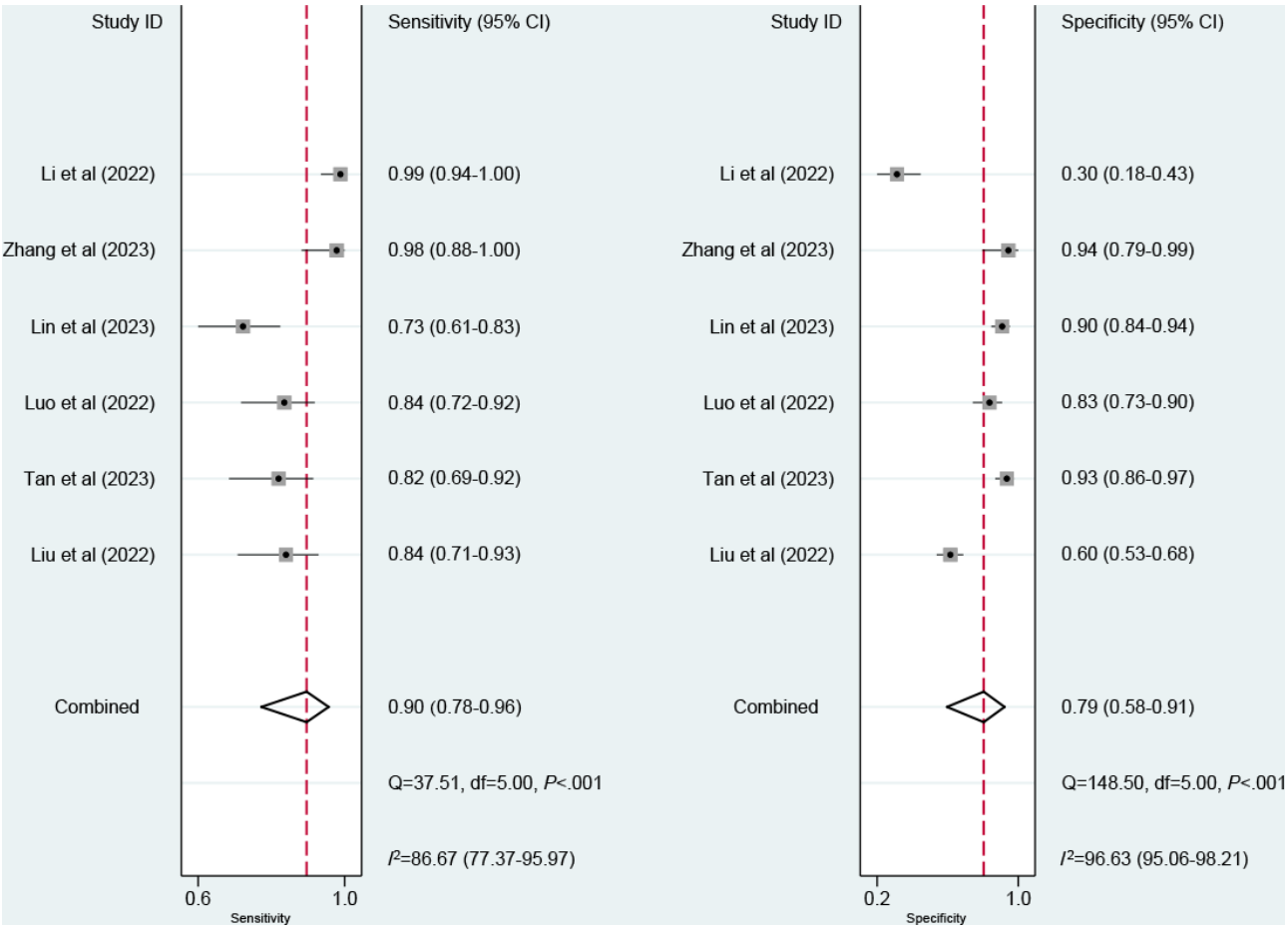

XSL•FO
RenderX

Figure 8. Forest plot for meta-analysis of sensitivity and specificity of machine learning for predicting esophagogastric variceal bleeding in patients with liver cirrhosis in the validation set [7,9,12,26,27].

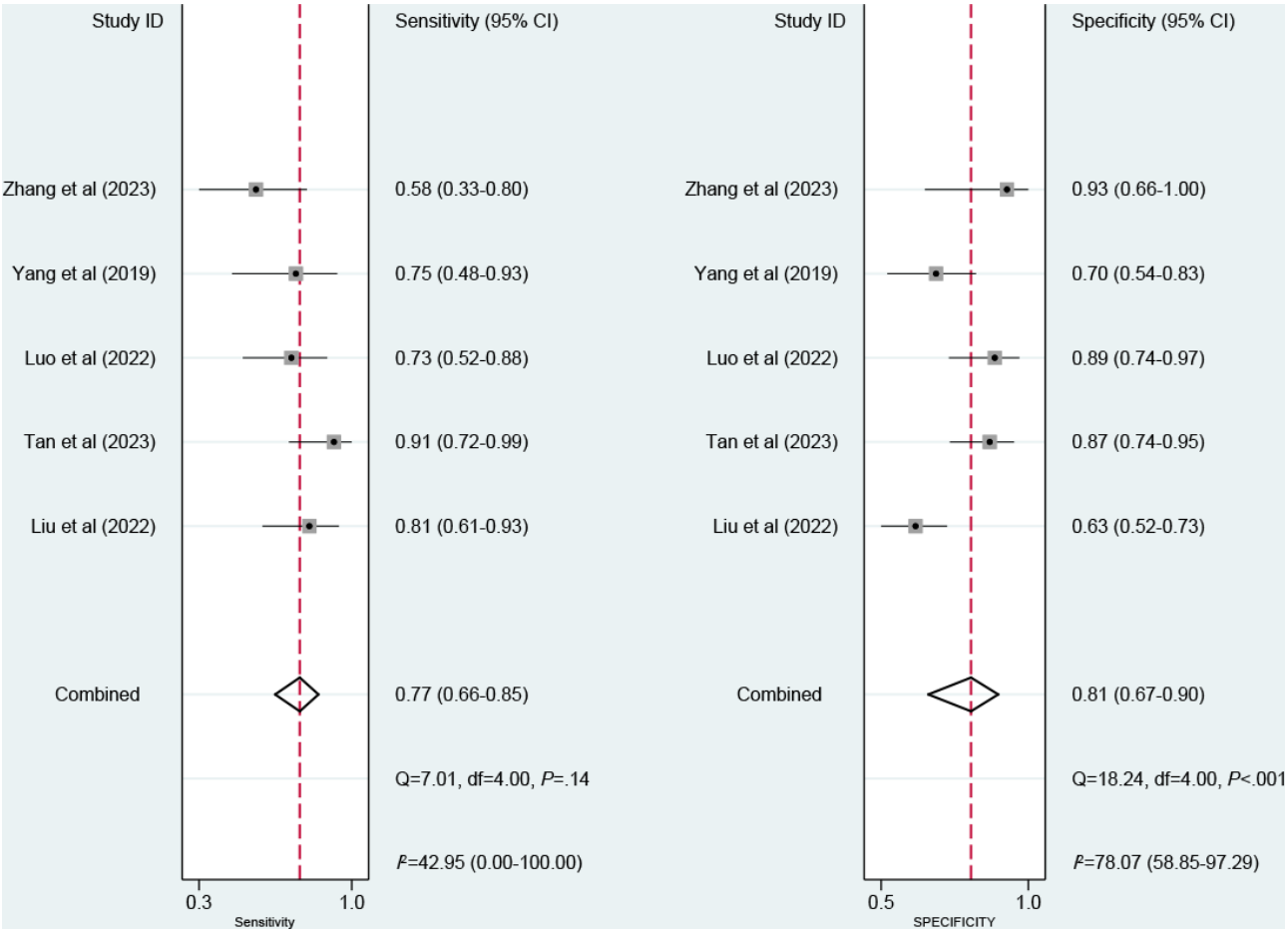

Subgroup Analysis

A subgroup analysis was performed based on modeling variables and the generation method of the validation set. In the subgroup analysis by model variables, models were grouped as models based on clinical features and models based on radiomics and clinical features. In the training set, the c-index was 0.87 (95% CI 0.80 - 0.93) for models based on clinical features and 0.90 (95% CI 0.84 - 0.97) for models based on radiomics and clinical features (Multimedia Appendix 5). In the validation set, the c-index was 0.91 (95% CI 0.86 - 0.96) for models based on clinical features, 0.85 (95% CI 0.75 - 0.96) for models based on radiomics and clinical features (Multimedia Appendix 6). In the subgroup analysis by the generation method of the validation set, models were grouped as random sampling and time-based split. The c-index was 0.89 (95% CI 0.83 - 0.94) for models based on random sampling and 0.93 (95% CI 0.85 - 1.00) for time-based split (Multimedia Appendix 7).

Discussion

Principal Findings

The risk of EVB and EGVB in patients with LC is high. Our study demonstrates that ML may be a feasible method for the early prediction of the risk of EVB and EGVB. In the validation set, the c-index, sensitivity, and specificity for ML models in the prediction of EVB were 0.85 (95% CI 0.77 - 0.92), 0.93 (95% CI 0.87 - 0.96), and 0.66 (95% CI 0.46 - 0.82), respectively. Moreover, in the validation set, the c-index, sensitivity, and specificity for ML models in the prediction of EGVB were 0.89 (95% CI 0.85 - 0.94), 0.77 (95% CI 0.66 - 0.85), and 0.81 (95% CI 0.67 - 0.90), respectively. Notably, 2 studies by Yang and Xu reported a sensitivity of 1.00 (100%) in the training set, but their specificity did not reach 1.00. Therefore, no perfect separation was observed. Furthermore, we also found that in the validation set, both studies reported sensitivities greater than 0.9 in the validation set, indicating no excessive overfitting. Therefore, this result indicates that the c-index and sensitivity of ML models are relatively high.

Comparison With Previous Studies

In patients with LC, EVB and EGVB represent a noticeable challenge that must be addressed in clinical research. Previous studies have established prediction models for EVB and EGVB. For instance, Malik et al [28] explored the use of ML for EVB or grading in patients with LC. Their results indicate that ML is highly promising for predicting esophageal varices in patients with LC and may reduce the necessity for minor surgeries. Nevertheless, their study points out that there is considerable heterogeneity and methodological limitations in their included studies, resulting in a high or unclear ROB. In the future, large-scale prospective trials are needed, and the evaluation standards for ML should be standardized to improve the practicality of these models in clinical settings.

XSL•FO
RenderX

Radiomics has recently attracted increasing attention in research on LC. In the study by Xue et al [29], transfer learning (TL) radiomics using multimodal ultrasound imaging was used for classifying the stage of liver fibrosis (LF). Their results revealed that in both the gray scale modality (GM) and elastogram modality (EM), TL exhibits superior diagnostic performance to non-TL approaches, with significantly higher areas under the receiver operating characteristic curve (AUCs; all $P<.01$). Single-modal GM and EM both outperform liver stiffness measurement (LSM) and serum indices for staging LF (all $P<.001$). The prediction performance of multimodal GM+EM is higher than GM+LSM, GM and EM alone, LSM, and biomarkers (all $P<.05$). The AUC of multimodal GM+EM for staging LF is 0.950 for S4, 0.932 for S3 or more, and 0.930 for S2 or more. Yan et al [30] used multiphase magnetic resonance imaging–based radiomics to forecast the histological grading of HCC. According to their results, 33.8% of the cases were high-grade HCCs. The levels of α-fetoprotein (odds ratio 1.89) and tumor size (>5 cm; odds ratio 2.33) are strongly associated with HCC grade. The performance of the combined model is high in predicting the grade of HCC in the test dataset (AUC=0.801), with favorable calibration and clinical utility. Park et al [31] reviewed the applications of radiomics and deep learning (DL) in liver diseases. According to their results, radiomics and DL are promising techniques for imaging-based assessment of liver diseases. Recent studies have reported the potential uses of radiomics and DL in staging LF, detecting portal hypertension, describing focal liver lesions, and predicting malignant liver tumors. Consequently, radiomics has been receiving growing attention in the clinical diagnosis and treatment of LC.

In our study, the included variables primarily consisted of interpretable clinical features and radiomic features. The accuracy of radiomics-based models for predicting EVB and EGVB in patients with LC is relatively favorable. In the included studies, 5 models were constructed based on radiomics features [12-14,26,27], among which 1 model was based on radiomics features [13] and 4 models were based on radiomics and clinical features [12,14,26,27]. The 5 studies describe the detection of bleeding in LC based on computed tomography (CT) or magnetic resonance imaging scans. Luo et al [12] developed a model based on 5 CT features from the liver and 3 CT features from the spleen. Its AUCs were 0.817 and 0.741 in the training and validation cohorts, respectively. Moreover, the clinical-radiomics model also exhibits notably predictive performance, with an AUC of 0.925 in the training cohort and 0.912 in the validation cohort. In the study by Meng et al [13], the Rad-score Liver-Spleen, based on 10 features, demonstrated strong discriminative performance. Its c-index is 0.853 (95% CI 0.776 - 0.904) in the training cohort and 0.822 (95% CI 0.749 - 0.875) in the validation cohort. According to Liu et al [14], the AUCs of their model are 0.89 (95% CI 0.84 - 0.94) and 0.78 (95% CI 0.68 - 0.87) in the training and validation datasets, respectively. These studies demonstrate that radiomics-based ML models have high predictive accuracy.

Although only a few studies on radiomic features were included in this study, their findings reveal that radiomics is essential in the diagnosis and treatment of LC. Given the small number of studies, their effect sizes were not pooled. However, the c-index of radiomics-based models was examined in the training and validation sets for both EVB and EGVB. For EVB, the c-index values were 0.85 (95% CI 0.77 - 0.93) and 0.82 (95% CI 0.69 - 0.96) for radiomics-based models in the training and validation sets, respectively. For EGVB, the c-index values were 0.90 (95% CI 0.84 - 0.97) and 0.85 (95% CI 0.75 - 0.96) for models based on radiomics and clinical features in the training and validation sets, respectively. The data indicate that radiomics appears to have a favorable performance for early prediction of EVB and EGVB. Therefore, future studies should focus on applying radiomics for the early differentiation of EVB and EGVB. However, due to the limited number of studies on radiomics, the results should be interpreted with caution. The detection performance of models constructed with traditional clinical features appears to be similar to that of models based on radiomics features. However, due to the limited number of studies on radiomics, more radiomics methods need to be explored to assess the risk of EVB and EGVB in patients with LC. Therefore, more studies need to be included in future research.

Furthermore, there was only one study based on endoscopic features. Endoscopy is broadly used in the management of EGV in patients with LC. The European Society of Gastrointestinal Endoscopy Guideline [32] offers a comprehensive overview of the role of endoscopy in the diagnosis and treatment of variceal bleeding. However, given the limited number of studies based on endoscopic images, it is impossible to assess their performance. Overall, this study demonstrates that ML has high predictive accuracy. Nonetheless, a high ROB is observed. Moreover, it is difficult to assess the ROB of models based on endoscopic features. Therefore, we believe that ML is a feasible approach to predicting EVB and EGVB in patients with LC, but it comes with a high ROB.

We conducted subgroup analyses by modeling variables, including those based on clinical characteristics and those based on imaging, to further elucidate the sources of heterogeneity. However, subgroup analyses revealed substantial heterogeneity within the subgroups. Due to the limited number of included articles, we were unable to conduct subgroup analysis by additional variables. Furthermore, subgroup analysis remains a significant challenge for meta-analyses on ML models. Given the limited number of included articles, the interpretation of differences in results across different types of ML is limited. Heterogeneity between different model types is a concern in meta-analyses of ML models. Even within the same model, differences in training rules can contribute to heterogeneity. Furthermore, the number of cases remains a potential influencing factor. Overall, this study aimed to quantitatively summarize ML models for the early prediction of esophageal or gastric variceal bleeding in LC, better demonstrating the application value of ML. However, high heterogeneity was observed. Although we attempted to address the heterogeneity from multiple perspectives, it was difficult to further quantify the impact of these differences on the results due to population differences, varying methods of assessing bleeding, and a limited number of studies. Therefore, this may impose certain

XSL•FO
RenderX

limitations on the interpretation of the results. Future research should further consider the impact of these factors on the results.

This study reveals a clear association between high AUC values and high ROB. Although several studies reported extremely high discrimination accuracy (AUC >0.90), PROBAST assessment showed that these studies generally had a high or unclear ROB. In the field of ML, near-perfect discrimination performance (AUC≈1.0) in small-sample, single-center studies is often a warning sign of methodological flaws rather than reflecting true performance. Specifically, among models with AUC ≥0.95, most were derived from single-center studies with a sample size of fewer than 200. These small sample studies with high AUC values mainly had the following problems in the PROBAST assessment: (1) the high ROB in participant selection was attributable to a retrospective design and case-control design; (2) the high ROB in analysis may be explained by the lack of appropriate validation strategy and treatment measures for overfitting; and (3) high ROB in predictors was due to insufficient transparency in variable selection and treatment. This suggests that limitations in methodology may lead to overestimation of model performance. Especially, extremely high AUC values observed in small-sample studies may reflect overfitting of the model to the training data rather than true clinical utility. This phenomenon may be due to multiple factors, such as occasional outperformance due to insufficient sample size, overfitting due to inadequate validation, and definitions of predictors and outcomes. The high ROB for the analysis domain in the PROBAST assessment (particularly the lack of measures to address overfitting) supports this explanation. Models with a c-index or AUC above 0.95 may be due to a small sample size and single-center design, with a potential risk of overfitting. Notably, high discrimination was also observed in larger, more rigorously designed multicenter trials [10].

These findings have important implications for clinical translation. First, for exceptionally high-performing models, the methodological rigor should be rigorously assessed. Second, future studies should include a sufficient sample size, perform cross-validation or external validation, and report calibration performance rather than focusing solely on discrimination. Finally, multicenter collaborations and prospective designs are crucial for generating generalizable predictive models. The discrimination performance of radiomics-based models in this study was high (AUC 0.91 - 0.93). However, given the limited number of related studies (n=5) with small sample sizes and insufficient validation, these results should be interpreted with caution.

Strengths and Limitations

This study is the first to explore ML in predicting the risk of EVB and EGVB in patients with LC, offering a strong reference for the future development or updating of intelligent tools. Nonetheless, certain limitations must be considered. First, owing to the small number of radiomics studies, this study only summarized the c-index of radiomics-based models and did not discuss their sensitivities and specificities. Hence, the results should be interpreted with caution. Second, only some of the studies explicitly described their validation sets, while others did not perform any model validation. Third, the models were primarily developed using internal validation methods, with no independent external validation, which, to some extent, restricts the interpretation of the results. Fourth, while current ML techniques demonstrate excellent sensitivity in validation datasets, the false-positive rate remains high. Future research should not only improve sensitivity but also optimize specificity. Fifth, given significant methodological and clinical heterogeneity in the current evidence, our results merely summarize the current state of application of ML. It has the potential to serve as an early predictive tool for esophageal or gastric variceal bleeding in LC. Future research should further improve the methodology and model performance to establish definitive clinical benchmarks. Sixth, although we conducted subgroup analysis based on the generation method of the validation set, there are few studies on external validation. The results of external validation need to be interpreted carefully. Seventh, there was excessive heterogeneity in the meta-analysis of sensitivity and specificity. Hence, the pooled estimates should be considered descriptive averages rather than true predictive performance. The results need to be interpreted with caution. Seventh, 18 of the 21 included studies were from China, which may limit the generalizability of the results to other regions. Therefore, the generalizability of the results still needs to be validated by more data. Finally, although our results show high AUC and c-index values, a large number of the included studies have a high ROB (particularly in the analysis domain, eg, small sample size, a lack of external validation, and potential overfitting). Hence, our results should be interpreted with caution.

Conclusions

The performance of ML-based models for predicting EVB and EGVB in patients with LC is relatively favorable, particularly those based on radiomics. However, as the number of the included original studies is limited, it is impossible to explore the impact of validation set generation methods, model types, and other factors on outcomes. Therefore, future studies should include multicenter cohorts and integrate multi-omics databases to develop ML prediction tools with improved predictive performance and broader applicability, thereby developing specialized prevention strategies to reduce the risk of EVB and EGVB in patients with LC.

Funding

The authors declare that they did not receive any funding from any source.

XSL•FO
RenderX

Authors' Contributions

YL contributed to conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing the original draft, reviewing and editing the article, visualization, supervision, and project administration. MS contributed to conceptualization, methodology, visualization, supervision, project administration, and funding acquisition. XQ contributed to conceptualization, methodology, validation, project administration, and funding acquisition; GW contributed to validation, formal analysis, and investigation. NL contributed to formal analysis, investigation, resources, and data curation. JZ contributed to resources and data curation. SS contributed to software, validation, formal analysis, investigation, writing the original draft, and reviewing and editing the article. SW contributed to software, validation, formal analysis, investigation, writing the original draft, and reviewing and editing the article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Literature search strategy.
[PNG File, 218 KB - jmir_v28i1e78203_app1.png]

Multimedia Appendix 2
Forest plot of meta-analysis of c-index of machine learning models based on different modeling variables for predicting esophageal variceal bleeding in patients with liver cirrhosis in the training set.
[PNG File, 78 KB - jmir_v28i1e78203_app2.png]

Multimedia Appendix 3
Forest plot of meta-analysis of c-index of machine learning models based on different modeling variables for predicting esophageal variceal bleeding in patients with liver cirrhosis in the validation set.
[PNG File, 80 KB - jmir_v28i1e78203_app3.png]

Multimedia Appendix 4
Forest plot of meta-analysis of c-index of machine learning models based on different methods of generating the validation set for predicting esophageal variceal bleeding patients with liver cirrhosis in the validation set.
[PNG File, 68 KB - jmir_v28i1e78203_app4.png]

Multimedia Appendix 5
Forest plot of meta-analysis of c-index of machine learning models based on different modeling variables for predicting esophagogastric variceal bleeding patients with liver cirrhosis in the training set.
[PNG File, 72 KB - jmir_v28i1e78203_app5.png]

Multimedia Appendix 6
Forest plot of meta-analysis of c-index of machine learning models based on different modeling variables for predicting esophagogastric variceal bleeding patients with liver cirrhosis in the validation set.
[PNG File, 63 KB - jmir_v28i1e78203_app6.png]

Multimedia Appendix 7
Forest plot of meta-analysis of c-index of machine learning models based on different methods of generating the validation set for predicting esophagogastric variceal bleeding patients with liver cirrhosis in the validation set.
[PNG File, 55 KB - jmir_v28i1e78203_app7.png]

Checklist 1
PRISMA checklist.
[PDF File, 613 KB - jmir_v28i1e78203_app8.pdf]

References

1. Ginès P, Krag A, Abraldes JG, Solà E, Fabrellas N, Kamath PS. Liver cirrhosis. Lancet 2021 Oct 9;398(10308):1359-1376. [doi: 10.1016/S0140-6736(21)01374-X] [Medline: 34543610]
2. Asrani SK, Devarbhavi H, Eaton J, Kamath PS. Burden of liver diseases in the world. J Hepatol 2019 Jan;70(1):151-171. [doi: 10.1016/j.jhep.2018.09.014] [Medline: 30266282]

XSL•FO
RenderX

3. Huang DQ, Terrault NA, Tacke F, et al. Global epidemiology of cirrhosis - aetiology, trends and predictions. Nat Rev Gastroenterol Hepatol 2023 Jun;20(6):388-398. [doi: 10.1038/s41575-023-00759-2] [Medline: 36977794]
4. Cherian JV, Deepak N, Ponnusamy RP, Somasundaram A, Jayanthi V. Non-invasive predictors of esophageal varices. Saudi J Gastroenterol 2011;17(1):64-68. [doi: 10.4103/1319-3767.74470] [Medline: 21196656]
5. Garcia-Tsao G, Abraldes JG, Berzigotti A, Bosch J. Portal hypertensive bleeding in cirrhosis: risk stratification, diagnosis, and management: 2016 practice guidance by the American Association for the study of liver diseases. Hepatology 2017 Jan;65(1):310-335. [doi: 10.1002/hep.28906] [Medline: 27786365]
6. Debray TP, Damen JA, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. Stat Methods Med Res 2019 Sep;28(9):2768-2786. [doi: 10.1177/0962280218785504] [Medline: 30032705]
7. Liu R, Sun Y, Xu K, Shi H, Sheng S, Kong D. A histogram model to predict the risk of bleeding from oesophageal and gastric variceal rupture in cirrhosis. J Coll Physicians Surg Pak 2022 May;32(5):586-590. [doi: 10.29271/jcpsp.2022.05.586] [Medline: 35546692]
8. Zhang S, Song W, Yang B, et al. A non-invasive model based on the virtual portal pressure gradient to predict the first variceal hemorrhage in cirrhotic patients. Hepatol Int 2022 Aug;16(4):926-935. [doi: 10.1007/s12072-022-10344-5] [Medline: 35585289]
9. Tan BG, Tang Z, Ou J, et al. An improved model based on quantitative features of right liver lobe, maximum varices, and portal vein system measured on magnetic resonance imaging to predict oesophagogastric variceal haemorrhage secondary to hepatitis B-related cirrhosis. Quant Imaging Med Surg 2023 Dec 1;13(12):7741-7752. [doi: 10.21037/qims-23-353] [Medline: 38106265]
10. Wang Y, Hong Y, Wang Y, et al. Automated multimodal machine learning for esophageal variceal bleeding prediction based on endoscopy and structured data. J Digit Imaging 2023 Feb;36(1):326-338. [doi: 10.1007/s10278-022-00724-6]
11. Hou Y, Yu H, Zhang Q, et al. Machine learning-based model for predicting the esophagogastric variceal bleeding risk in liver cirrhosis patients. Diagn Pathol 2023 Feb 23;18(1):29. [doi: 10.1186/s13000-023-01293-0] [Medline: 36823660]
12. Luo R, Gao J, Gan W, Xie WB. Clinical-radiomics nomogram for predicting esophagogastric variceal bleeding risk noninvasively in patients with cirrhosis. World J Gastroenterol 2023 Feb 14;29(6):1076-1089. [doi: 10.3748/wjg.v29.i6.1076] [Medline: 36844133]
13. Meng D, Wei Y, Feng X, et al. CT-based radiomics score can accurately predict esophageal variceal rebleeding in cirrhotic patients. Front Med (Lausanne) 2021;8:745931. [doi: 10.3389/fmed.2021.745931] [Medline: 34805214]
14. Liu H, Sun J, Liu G, Liu X, Zhou Q, Zhou J. Establishment of a non-invasive prediction model for the risk of oesophageal variceal bleeding using radiomics based on CT. Clin Radiol 2022 May;77(5):368-376. [doi: 10.1016/j.crad.2022.01.046] [Medline: 35241274]
15. Xu J, Tan L, Jiang N, et al. Assessment of nomogram model for the prediction of esophageal variceal hemorrhage in hepatitis B-induced hepatic cirrhosis. Eur J Gastroenterol Hepatol 2024 Jun 1;36(6):758-765. [doi: 10.1097/MEG.0000000000002750] [Medline: 38683192]
16. Li J, Li J, Ji Q, et al. Nomogram based on spleen volume expansion rate predicts esophagogastric varices bleeding risk in patients with hepatitis B liver cirrhosis. Front Surg 2022;9:1019952. [doi: 10.3389/fsurg.2022.1019952] [Medline: 36468077]
17. Lin LY, Zeng DW, Liu YR, Zhu YY, Huang LL. Diagnostic value of liver stiffness measurement combined with risk scores for esophagogastric variceal bleeding in patients with hepatitis B cirrhosis. Eur J Radiol 2024 Apr;173:111385. [doi: 10.1016/j.ejrad.2024.111385] [Medline: 38377895]
18. Zhang Q, Niu S, Yang L, et al. A novel prognostic model for predicting the risk of first variceal hemorrhage in patients with HBV-related cirrhosis. Front Cell Infect Microbiol 2023;13:1062172. [doi: 10.3389/fcimb.2023.1062172] [Medline: 36733609]
19. Yang LB, Gao X, Li H, et al. Non-invasive model for predicting high-risk esophageal varices based on liver and spleen stiffness. World J Gastroenterol 2023 Jul 7;29(25):4072-4084. [doi: 10.3748/wjg.v29.i25.4072]
20. Fu S, Chen D, Zhang Z, Shen R. Predictive value of spectral computed tomography parameters in esophageal variceal rupture and bleeding in cirrhosis. Turk J Gastroenterol 2023 Apr;34(4):339-345. [doi: 10.5152/tjg.2023.21908] [Medline: 37089047]
21. Samy M, Gamal D, Othman MHM, Ahmed SA. Assessment of variceal bleeding in cirrhotic patients: accuracy of multi-detector computed tomography. Egypt J Radiol Nucl Med 2022 Dec;53(1):62. [doi: 10.1186/s43055-022-00738-8]
22. Salahshour F, Mehrabinejad MM, Rashidi Shahpasandi MH, et al. Esophageal variceal hemorrhage: the role of MDCT characteristics in predicting the presence of varices and bleeding risk. Abdom Radiol (NY) 2020 Aug;45(8):2305-2314. [doi: 10.1007/s00261-020-02585-5] [Medline: 32447415]
23. Xu XD, Dai JJ, Qian JQ, Pin X, Wang WJ. New index to predict esophageal variceal bleeding in cirrhotic patients. World J Gastroenterol 2014 Jun 14;20(22):6989-6994. [doi: 10.3748/wjg.v20.i22.6989] [Medline: 24944493]
24. Heo JY, Kim BK, Park JY, et al. Multicenter retrospective risk assessment of esophageal variceal bleeding in patients with cirrhosis: an acoustic radiation force impulse elastography-based prediction model. Gut Liver 2019 Mar 15;13(2):206-214. [doi: 10.5009/gnl18292] [Medline: 30602219]

XSL•FO
RenderX

25. Liu CH, Liu S, Zhao YB, et al. Development and validation of a nomogram for esophagogastric variceal bleeding in liver cirrhosis: a cohort study in 1099 cases. J Dig Dis 2022 Oct;23(10):597-609. [doi: 10.1111/1751-2980.13145] [Medline: 36400743]
26. Yang JQ, Zeng R, Cao JM, et al. Predicting gastro-oesophageal variceal bleeding in hepatitis B-related cirrhosis by CT radiomics signature. Clin Radiol 2019 Dec;74(12):976. [doi: 10.1016/j.crad.2019.08.028] [Medline: 31604574]
27. Zhang Y, Zhao Q, Wen J. Splenic CT radiomics nomogram predicting the risk of upper gastrointestinal hemorrhage in cirrhosis. Journal of Radiation Research and Applied Sciences 2023 Mar;16(1):100486. [doi: 10.1016/j.jrras.2022.100486]
28. Malik S, Tenorio BG, Moond V, Dahiya DS, Vora R, Dbouk N. Systematic review of machine learning models in predicting the risk of bleed/grade of esophageal varices in patients with liver cirrhosis: a comprehensive methodological analysis. J Gastroenterol Hepatol 2024 Oct;39(10):2043-2059. [doi: 10.1111/jgh.16645] [Medline: 38886175]
29. Xue LY, Jiang ZY, Fu TT, et al. Transfer learning radiomics based on multimodal ultrasound imaging for staging liver fibrosis. Eur Radiol 2020 May;30(5):2973-2983. [doi: 10.1007/s00330-019-06595-w] [Medline: 31965257]
30. Yan Y, Si Z, Chun C, et al. Multiphase MRI-based radiomics for predicting histological grade of hepatocellular carcinoma. J Magn Reson Imaging 2024 Nov;60(5):2117-2127. [doi: 10.1002/jmri.29289] [Medline: 38344854]
31. Park HJ, Park B, Lee SS. Radiomics and deep learning: hepatic applications. Korean J Radiol 2020 Apr;21(4):387-401. [doi: 10.3348/kjr.2019.0752] [Medline: 32193887]
32. Gralnek IM, Camus Duboc M, Garcia-Pagan JC, et al. Endoscopic diagnosis and management of esophagogastric variceal hemorrhage: European Society of Gastrointestinal Endoscopy (ESGE) Guideline. Endoscopy 2022 Nov;54(11):1094-1120. [doi: 10.1055/a-1939-4887] [Medline: 36174643]

Abbreviations

AUC: area under the receiver operating characteristic curve
CT: computed tomography
DL: deep learning
EGVB: esophagogastric variceal bleeding
EM: elastogram modality
EVB: esophageal variceal bleeding
GM: gray scale modality
HCC: hepatocellular carcinoma
LC: liver cirrhosis
LF: liver fibrosis
LSM: liver stiffness measurement
ML: machine learning
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST: Prediction Model Risk of Bias Assessment Tool
ROB: risk of bias
TL: transfer learning

Edited by A Coristine; submitted 28.May.2025; peer-reviewed by H Lin, L Jiang, R Luo; revised version received 14.Feb.2026; accepted 17.Feb.2026; published 08.Apr.2026.

Please cite as:
Lian Y, Qiu X, Wang G, Liu N, Zhao J, Song S, Wang S, Sun M
Machine Learning in Predicting the Risk of Esophagogastric Variceal Bleeding Among Patients With Liver Cirrhosis: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e78203
URL: https://www.jmir.org/2026/1/e78203
doi:10.2196/78203

© Yuan Lian, Xinping Qiu, Geli Wang, Nannan Liu, Jing Zhao, Shanshan Song, Shiqi Wang, Mingjun Sun. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Innovations in Deaf Health Care Communication: Systematic Review of Sign Language Recognition Systems

Milena Soriano Marcolino[1,2,3], MD, PhD; Lucca Fagundes Ramos de Oliveira[2,4], MD; Lucas Rocha Valle[1,2]; Luiza Marinho Motta Santa Rosa[2,5], MD; Gabriela Teodora de Souza Sanches[1,2]; Natalia Sales Santos[6], MSc; Michelle Ralil da Costa[3], PhD; Elidea Lucia Almeida Bernardino[7], PhD; Raniere Alislan Almeida Cordeiro[7], PhD; Raquel Oliveira Prates[6], PhD; Zilma Silveira Nogueira Reis[1], MD, PhD; Mario Fernando Montenegro Campos[6], PhD

[1]Medical School, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[2]Telehealth Center, University Hospital, Universidade Federal de Minas Gerais, Avenida Professor Alfredo Balena, 190, Santa Efigênia, Belo Horizonte, Brazil

[3]Institute for Health Technology Assessment, Porto Alegre, Brazil

[4]Medical School, Universidade Federal de Ouro Preto, Ouro Preto, Brazil

[5]Faculdade de Ciências Médicas de Minas Gerais, Belo Horizonte, Brazil

[6]Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[7]Faculty of Arts and Sciences, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Corresponding Author:
Milena Soriano Marcolino, MD, PhD
Medical School, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Abstract

Background: Deaf individuals often face communication challenges when interacting with those who can hear. Within health care settings, these challenges may pose risks to their safety, potentially resulting in misdiagnoses, treatment errors, and decreased quality of care.

Objective: This study aims to systematically review the evidence on communication systems reported in the literature that use human-computer interaction techniques to support communication between deaf individuals who use sign language and hearing health professionals in health care settings. The review focuses on systems that are either currently in use or proposed for use in health care and that have been tested using human participants or videos of human users.

Methods: A comprehensive search was performed via MEDLINE, Web of Science, ACM, IEEE Xplore, Scopus, and Google Scholar in March 2025. The inclusion criteria comprised studies developing a sign language recognition system within a health care context and testing with human users. Eligible studies underwent screening by 2 independent investigators (LRV and LMMSR or LFRdO and GTdSS), with any disagreements resolved by a senior researcher (MSM).

Results: The search retrieved 21,778 publications, and screening of reference lists identified 2 additional studies, resulting in a total of 23 studies meeting the eligibility criteria. Most systems (15/23, 65.2%) were image-based, while 34.8% (8/23) relied on sensors (glove-based or depth-sensing). Applications varied across health care settings, including general hospital care (10/23, 43.5%), emergencies (8/23, 34.8%), and primary care (4/23, 17.4%). All systems were in the development and testing stage, with no data on security and psychological impacts. Accuracy ranged from 25% to 100% for image-based and 72% to 99.7% for sensor-based systems. Bidirectionality and facial expression recognition, crucial for effective communication, were largely overlooked.

Conclusions: Image-based systems were more common than sensor-based ones, though both showed wide variability in accuracy in recognizing and interpreting signs. Most systems failed to address critical aspects such as bidirectional communication and the recognition of facial expressions, essential for effective communication. None fully addresses the requirements for integration into health care settings. These findings highlight the need for further research on implementation, usability, and impact on the quality of care for deaf patients.

International Registered Report Identifier (IRRID): RR2-10.2196/55427

(J Med Internet Res 2026;28:e70417) doi:10.2196/70417

KEYWORDS

computer neural network; artificial intelligence; biomedical technology; sign language; hearing loss; deafness; communication barriers; gestures; PRISMA; Preferred Reporting Items for Systematic Reviews and Meta-Analyses

XSL•FO
RenderX

Introduction

Sign languages are a form of communication characterized by the successful coordination of gestures, body, head, and hand motions, with facial expressions. They are naturally developed and highly structured systems, governed by a set of linguistic rules, different from spoken languages with no standardized written forms [1]. They enable comprehensive and rich linguistic systems for formulating words and phrases, making them the primary mode of communication for many hearing-impaired people [2]. According to the World Health Organization, over 430 million people worldwide experience disabling hearing loss, a number expected to exceed 700 million by 2050 [3]. Despite the World Federation of the Deaf estimating that there are over 200 sign languages and over 70 million deaf individuals who use them, reliable data on sign language users remains scarce [4].

Even though sign language is the primary mode of communication for millions of deaf individuals, they still encounter significant barriers in daily interactions, particularly in critical areas such as health care, education, and public services, where most professionals do not understand or use sign language [5-7]. Specifically in health care, this communication gap can lead to misdiagnoses, inadequate treatment, and reduced quality of care [8].

In response to this challenge, and recognizing the significance of sign language in fostering social integration for the deaf community, researchers have developed sign language recognition (SLR) systems. SLR refers to the identification and interpretation of sign language gestures and movements [9]. These systems aim to facilitate real-time translation between sign and spoken languages, to build a communication bridge between deaf individuals who communicate through sign language and ordinary people, and to enhance accessibility and inclusion across multiple domains, including health care [10]. In this pursuit, machine learning and other artificial intelligence (AI) techniques have become important emerging tools to help overcome communication barriers, enhancing not only accessibility but also valuing the cultural identity of the deaf community.

Regarding automatic SLR systems, it is crucial to highlight the use of advanced algorithms and other AI techniques to interpret the complex gestures and movements inherent to sign language communication. With extensive use of advanced computer vision algorithms and machine learning techniques, these systems are becoming increasingly able to automatically recognize and translate signs into text of the target language, thereby facilitating effective communication between deaf individuals and those who do not understand sign language [11-13].

The SLR systems comprise different types of methodologies and technologies to recognize and translate sign language. In this paper, we organize them into 2 main categories: image-based and sensor-based. Even though strictly speaking, cameras are sensors, we used sensor-based to identify the approaches that use other sensor modalities. Image-based approaches use computer vision techniques coupled with deep learning models to analyze video streams from cameras that capture hand gestures, body movements, and facial expressions to accurately detect and interpret signs [11,13,14]. Sensor-based systems use wearable devices equipped with sensors, such as accelerometers and gyroscopes, to capture hand and body movements and translate signs into written or spoken language [11,12,15]. These systems can be divided into 2 subgroups: glove-based systems and depth-sensor systems. Glove-based systems use specialized gloves embedded with sensors to capture fine-grained hand movements, enabling real-time translation into spoken or written language. In contrast, depth-sensor systems rely on depth-sensing cameras, such as Kinect or light detection and ranging, to obtain 3D information about hand and body movements. By analyzing the 3D pose of different body parts, these systems allow more precise recognition and interpretation of sign language gestures [16].

Each of the aforementioned types of systems has strengths and limitations concerning accuracy in detecting and interpreting signs, portability, cost, and accessibility. Additionally, user-friendliness is an important consideration in the development and implementation of SLR systems. It is critical for ensuring accessibility and inclusion for deaf individuals. As these systems are intended to be used by people with varying levels of technical expertise, their design must prioritize simplicity, intuitiveness, and ease of use [17]. This is particularly true in health care settings, where effective communication with medical staff is essential. Furthermore, given the sensitive nature of health care interactions, it is crucial to assess how these systems address ethical concerns, such as patient data privacy, and practical challenges related to accessibility, including cost and implementation feasibility in diverse health care settings.

Therefore, this study aims to systematically review the evidence on the translation systems developed for deaf people who communicate through sign language with hearing health professionals in a health care context, which are already in use or proposed for use and have been tested with human users or videos of human users. The main research question was: What technologies have been developed and tested in real-world settings to translate sign and oral languages, facilitating communication between deaf patients who primarily use sign language and health care workers? The specific questions are as follows:

- In which context of health care have these technologies been used?
- Which languages (sign and oral) can these technologies translate?
- Which technologies are required for it to be used on-site? How were they developed?
- How were they deployed and tested?
- How has the communication between health care workers and deaf people been improved by using these technologies?
- How was the efficacy of these technologies evaluated?
- Is the system or technology “bidirectionally” interactive?
- How do these systems address ethical concerns in health care settings, such as patient privacy and data security?

XSL•FO
RenderX

Methods

Overview

The research protocol was registered in the Open Science Framework and previously published in detail [18,19]. It followed guidance from the Cochrane Guidelines and the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [20,21] (Checklist 1). A multidisciplinary team comprising researchers from health and computing domains, along with linguistic specialists in sign language, collaboratively conducted the systematic review. Two members of the team are linguistic specialists and sign language researchers, one of them is deaf.

Search Strategy

Independent researchers performed a literature search using Web of Science, MEDLINE, IEEE Xplore, ACM, Scopus, and Google Scholar. A preliminary search strategy, developed by 5 authors (MSM, LFRO, LRV, ROP, and ZSNR), incorporated MeSH and defined text words relevant to the topic (Multimedia Appendix 1). The last search was conducted on March 27, 2025.

All studies, regardless of publication date or language, from inception onward were considered. In the case of unpublished studies, the authors were contacted at least 3 times via email or other social networks used by researchers to request additional information. Reference lists of eligible studies were examined to identify additional eligible studies.

Study Selection

Prospective, retrospective, or descriptive studies that address the development of communication systems specifically designed for deaf individuals in health care encounters and that involve testing with human users or videos of human users were included. Human users could be people of any age who are deaf and whose primary communication modality is sign language.

Studies that do not address the specified research questions, do not mention testing with human users or videos of human users, or do not mention use in health care encounter contexts were excluded. Short communications, conference abstracts, and correspondences were not excluded.

Independent researchers blindly screened the studies. Titles and abstracts of identified studies were individually reviewed to assess eligibility. Full-text versions of papers that were not excluded at this initial stage were read for a thorough examination. Subsequently, potentially pertinent studies were independently evaluated to ascertain if they aligned with the inclusion criteria. Any disagreements were resolved by a senior researcher (MSM). Whenever necessary, corresponding authors were contacted to obtain data not included in the publication using email and ResearchGate.

Search data for the identified studies and information for each stage of study selection were registered in detail, following the guidelines of the PRISMA methodology [20,21].

Data Extraction

A data extraction table was custom-designed for this study and independently piloted by 2 researchers, as well as data extraction. The extraction was checked by 2 other reviewers, one researcher with a computer science background and a senior researcher with a health science background. Conflicts were resolved by consensus or by consulting a senior researcher (MSM). Details on the variables extracted were previously published [18,19].

Furthermore, the researchers responsible for each study included in the systematic review were contacted via email (with up to 3 attempts) and ResearchGate (at least 1 attempt) to obtain updates on the current status of their systems and to request any additional information regarding their application in real-world contexts.

The definitions used regarding the type of SLR system, the corpus formation, and the health context are available in Multimedia Appendix 2. To extract data from papers that met the eligibility criteria, the authors of the study developed a codebook with clear definitions for all variables to ensure consistent data collection (Multimedia Appendix 3).

Data Analysis

A qualitative synthesis was performed to analyze the data, and a narrative synthesis of the evidence was conducted to provide an overview of the results. The results are summarized according to system types: image-based and sensor-based.

Results

Search Results and Study Selection

The search retrieved 21,778 publications, 2021 of which were duplicates. Most studies were excluded after title and abstract analysis (n=19,605). In total, 6 reports were unavailable, of which 4 did not respond despite repetitive contact attempts with corresponding authors, and 2 did not make any contact details available (Multimedia Appendix 4), thus leaving 146 selected for full-text screening. Of those, 22 studies were selected for inclusion after applying the eligibility criteria. In 2 cases, publications by the same research groups were identified. The first case involved a deep learning–based system for recognizing emergency gestures in Indian Sign Language to support communication with hearing-impaired individuals [22,23]. The journal paper published in IEEE Access in 2022 represents an extended and more comprehensive version of the earlier conference paper. It includes additional models (such as 3D convolutional neural network [CNN] and You Only Look Once [YOLO] v5), a more detailed methodology, a larger dataset, and more robust performance metrics, including mean average precision, precision, and recall. Given its methodological completeness and broader evaluation, only the IEEE Access paper was included in this review, while the conference version was excluded to avoid duplication of data and analysis [22].

The second case involved 2 publications from a Brazilian research group describing a system for recognizing Brazilian Sign Language (Libras) in the health context [15,24]. The 2024 journal paper incorporates a more advanced architecture (multiple-stream versus 2-stream), enhanced performance results, and a more detailed methodology, while using the same dataset as the earlier work. Thus, only the 2024 version was included [15].

XSL•FO
RenderX

Additionally, 2 studies were obtained from reference list screening, totaling 23 studies ultimately included in the review (Figure 1) [12-15,22,25-42].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart [21].

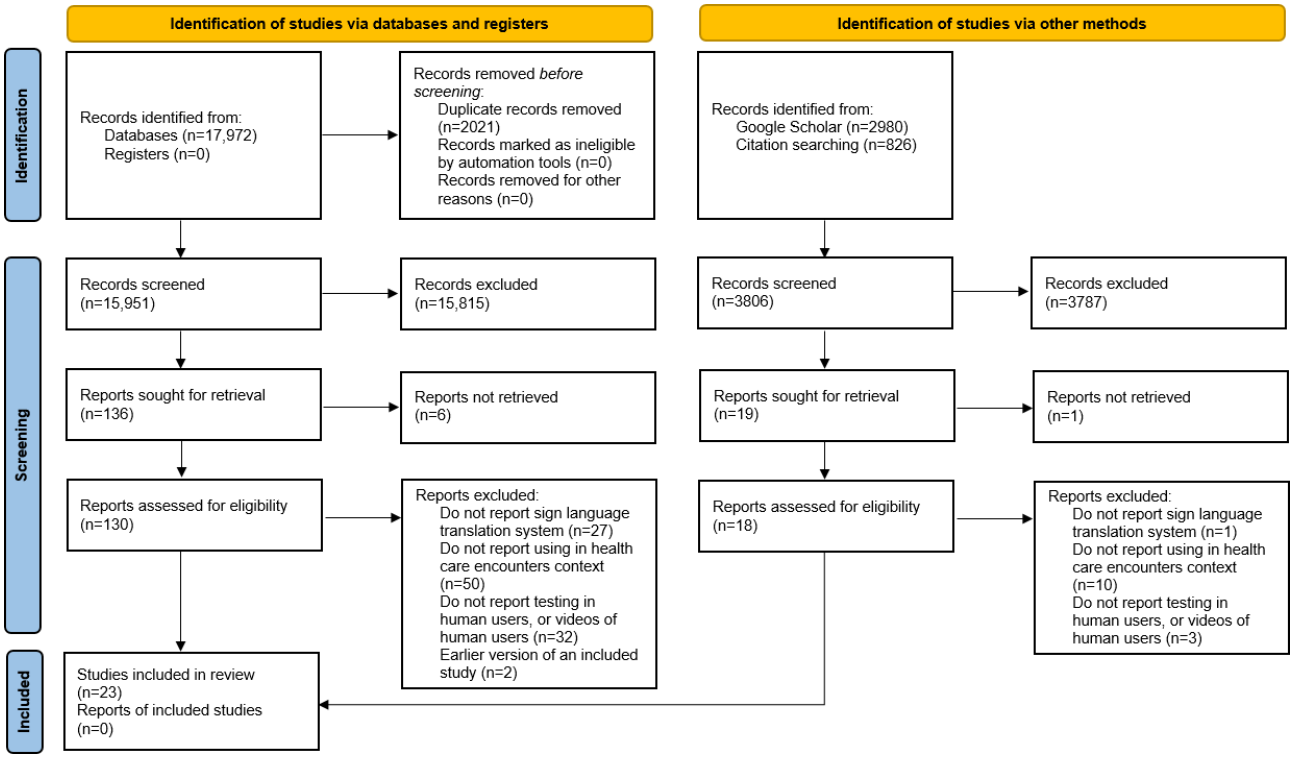

Characteristics of Included Studies

The main characteristics of the included studies are summarized in Tables 1 and 2 and Multimedia Appendix 5 [12-15,22,25-42]. They were published in 18 journals and conferences from 2015 to 2024. All of them were published in English. In 15 studies, the system was classified as image-based [13-15,22,25-28,30,31,36-39,42], and in 7, it was sensor-based. Of these, 4 were depth-sensing [29,33,35,41], and 3 were glove-based [32,34,40].

XSL•FO
RenderX

Table . Main architectural, functional, and technical characteristics of sign language recognition systems.

	Languages involved (oral and sign)	Technology needed		Infrastructure needed
		Hardware or software	Development technology[a]	
Image-based				
Xia et al (2022) [13]	Chinese and Chinese Sign Language	Kendryte K210, 8-megapixel camera, microphone arrays, operating system FreeRTOS, LabelImg, Autodesk Inventor	TensorFlow, Keras, Darknet	SLA[b] 3D printing technology and heart-speaker device
Pikoulis et al (2022) [14]	Greek and Greek Sign Language	No information	Sentence-BERT[c], Google Mediapipe	Computer and camera
da Silva et al (2024) [15]	Brazilian Portuguese and Brazilian Sign Language (Libras)	Intel Core i3 and all experiments were performed on Ubuntu 18.04 LTS[d]	LSTM[e] network, OpenPose, Keras, cuDNN[f] 8.5	Computer and smartphone HD[g] camera
Das et al (2023) [25]	English and Indian Sign Language	No information	CNN[h], BiLSTM[i]	Computer and camera
Ko et al (2019) [26]	Korean and Korean Sign Language	No information	OpenPose, PyTorch	Computer and camera
Barrientos-Villalta et al (2022) [27]	Peruvian Spanish and Peruvian Sign Language	Google Cloud, Storage, Google Cloud Run	Google Mediapipe, LSTM neural network	Mobile devices and internet
Ramírez Sánchez et al (2021) [28]	Spanish and Mexican Sign Language	Webcam	Google MediaPipe, OpenCV, CNN, HMM[j], Viterbi algorithm	Webcam and computer
Gandhi et al (2021) [30]	English and Indian Sign Language	Mobile phones with their camera resolutions	CNN using Image Stacking (VGG-16[k] and Resnet50) CNN + LSTM and LSTM with OpenPose	Mobile phone and camera
Uchil et al (2019) [31]	English and Indian Sign Language	Intel Core i5 CPU[l] running macOS Mojave and Core i3 CPU running Windows 7, smartphone	OpenPose library, OpenCV	Mobile phone camera
Areeb et al (2022) [22]	English and Indian Sign Language	No information	3D CNN, pretrained VGG-16, LSTM (RNN-LSTM)[m] scheme, YOLO[n] v5	Computer and camera
Adithya and Rajesh (2020) [36]	English and Indian Sign Language	Digital camera	LSTM network	Computer and camera
Ihsan et al (2024) [37]	English and American Sign Language	Intel Core i5, Visual Studio Code	Pretrained CNN (MobileNetV2) model, BiLSTM model, Mediapipe, TensorFlow, Keras	Computer, camera, and smartphones
Das et al (2024) [38]	English and Indian Sign Language	No information	CNNs, LSTM, long-term recurrent convolutional network model	Computer and mobile phone camera
Faisal et al (2023) [39]	Arabic and Saudi Sign Language	Robot Operating System, IClone Pro, 3DXchange, Unity	Self-developed concise 3D graph convolutional network, time delay neural network model, DTW[o] algorithm, FastSpeech2, Kaldi toolkit, MediaPipe	Portable electronic devices or computers equipped with a camera
Bellil et al (2024) [42]	Arabic and Algerian Sign Language	iPhone XS MAX	1D-CNN, MediaPipe	Computer and smartphone
Sensor-based (depth-sensing)				

XSL•FO
RenderX

	Languages involved (oral and sign)	Technology needed		Infrastructure needed
		Hardware or software	Development technology[a]	
Hisham and Hamouda (2019) [29]	Egyptian Arabic and Egyptian Arabic Sign Language	Kinect SDK, Windows 7, Visual Basic.NET, C# and C++	Bayesian network, AdaBoosting, DTW, and HMM	Computer and Microsoft Kinect sensor
Sarhan et al (2015) [33]	Arabic and Arabic Sign Language	No information	HMM	Computer and Microsoft Kinect sensor
Süzgün et al (2015) [35]	English and Turkish Sign Language	Personal computer and a touch display	DTW algorithm	Computer and Microsoft Kinect sensor
Dewasurendra et al (2020) [41]	Sinhala and Sri Lankan Sign Language	Webcam, mobile device with camera, microphone, GPS	EfficientNet-Lite0, Pillow (Python image processing library), MaryTTS Framework, CMU[p] Sphinx 4 toolkit (Sphinx4+ Sphinx-Train) and natural language processing, TensorFlow	Kinect and Leap Motion Controller
Sensor-based (glove-based)				
Deji Dere et al (2022) [32]	English and American Sign Language	Arduino Nano 33 BLE, Edge Impulse software	1D-CNN model	Inertial Measurement Unit
Guo et al (2023) [34]	English and American Sign Language	INMO AIR with Android 10 Go	Android-MediaPipe, VOSK[q] library	Mobile phone and augmented reality glasses[r]
Luqman and Mahmoud (2020) [40]	Arabic and Arabic Sign Language	No information	KenLM (statistical language model), MADAMIRA[s] (morphological analyzer and disambiguation tool)	Computer and camera
Hybrid				
Sosa-Jiménez et al (2022) [12]	Mexican Spanish and Mexican Sign Language	Intel Core i7 running Windows 7, Intel Core i5 running Windows 8.1, Microsoft Visual Studio 2015	Markov models (probabilistic) and neural networks	Computer and Microsoft Kinect sensor

[a]Artificial intelligence and imaging processing.
[b]SLA: stereolithography.
[c]BERT: bidirectional encoder representations from transformers.
[d]LTS: long-term support.
[e]LSTM: long short-term memory.
[f]cuDNN: CUDA deep neural network library.
[g]HD: high definition.
[h]CNN: convolutional neural network.
[i]BiLSTM: bidirectional long short-term memory.
[j]HMM: hidden Markov model.
[k]VGG-16: visual geometry group 16 layers.
[l]CPU: central processing unit.
[m]RNN-LSTM: recurrent neural network with a long short-term memory.
[n]YOLO: You Only Look Once.
[o]DTW: dynamic time warping.
[p]CMU: speech recognition toolkit.
[q]VOSK: offline speech recognition toolkit [43].
[r]Mobile phones and augmented reality glasses: portable electronic devices that enable interactive digital experiences.
[s]MADAMIRA: a system for morphological analysis and disambiguation of Arabic [44].

XSL•FO
RenderX

Table . Main characteristics of the recognition approaches of sign language recognition systems.

	Corpus formation			Captures facial expressions and body movement	Health context
	Type	Isolated words	Sentences		
Image-based					
Xia et al (2022) [13]	Isolated words	19	0	N/A[a]	General hospital care
Pikoulis et al (2022) [14]	Isolated words and sentences	6319 words with 1374 of them being unique (excluding repetitions)	1029 simple sentences with 945 of them being unique (excluding repetitions).	N/A	Psychiatric interviews
da Silva et al (2024) [15]	Isolated words	50	0	Yes	General hospital care
Das et al (2023) [25]	Isolated words	8	0	N/A	Emergency situations
Ko et al (2019) [26]	Isolated words and sentences	419	105	Yes	Emergency situations
Barrientos-Villalta et al (2022) [27]	Isolated words	17	0	Yes	General hospital care
Ramírez Sánchez et al (2021) [28]	Isolated words and sentences	49	20	Yes	Primary care consultations
Gandhi et al (2021) [30]	Isolated words	20	0	Yes	General hospital care
Uchil et al (2020) [31]	Isolated words	20	0	Yes	General hospital care
Areeb et al (2022) [22]	Isolated words	8	0	N/A	Emergency situations
Adithya and Rajesh (2020) [36]	Isolated words	8	0	No	Emergency situations
Ihsan et al (2024) [37]	Isolated words	30	0	N/A	General hospital care
Das et al (2024) [38]	Isolated words	6	0	N/A	Primary care consultations
Faisal et al (2023) [39]	Isolated words	293	0	N/A	Primary care consultations
Bellil et al (2024) [42]	Isolated words	10	0	No	General hospital care
Sensor-based (depth-sensing)					
Hisham and Hamouda (2019) [29]	Isolated words	42	0	No	General hospital care
Sarhan et al (2015) [33]	Isolated words	16	0	No	Emergency consultations
Süzgün et al (2015) [35]	Isolated words	33	0	N/A	General hospital care
Dewasurendra et al (2020) [41]	Isolated words	0	N/A	No	Emergency situations
Sensor-based (glove-based)					
Deji Dere et al (2022) [32]	Isolated words	5	0	No	Emergency situations
Guo et al (2023) [34]	Isolated words	More than 550	0	N/A	Emergency situations
Luqman and Mahmoud (2020) [40]	Isolated words and sentences	3327 sign words and 30,296 singular words with their plurals	600	Yes	General hospital care

XSL•FO
RenderX

	Corpus formation			Captures facial expressions and body movement	Health context
	Type	Isolated words	Sentences		
Hybrid					
Sosa-Jiménez et al (2022) [12]	Isolated words	43	0	Yes	Primary care consultations

[a]N/A: not applicable.

The system developed by Sosa-Jiménez et al [12] is the only one classified as hybrid, as it leverages Kinect's red, green, blue, and depth sensors to recognize signs through both visual input and 3D skeletal tracking. Because it is the only hybrid system, it is presented separately in Table 1 and Multimedia Appendix 5; however, in the following sections, it will be discussed alongside the sensor-based systems (depth-sensing).

Aspects Related to System Development

Languages Involved, Direction of Communication, and Capturing of Facial Expressions

Image-Based

The systems studied supported a range of spoken-sign language pairs, including Arabic (Algerian and Saudi), English, Chinese, Brazilian, Portuguese, Greek, Indian, Korean, Peruvian, and Mexican Spanish, and their respective sign languages (Table 2). Only 3 papers successfully implemented a bidirectional communication between spoken and sign languages [15,35,39], while another study attempted bidirectionality but did not achieve it [13].

Only 2 studies explicitly addressed structural and cultural differences between spoken and sign languages in their translation processes [15,28]. Additionally, 7 systems incorporated the recognition of facial expressions, facial key points, and other nonmanual features (such as audio phonemes and body posture) to enhance translation accuracy and communication effectiveness [15,26-28,30,31,39].

Sensor-Based

These systems also supported multiple spoken-sign language pairs, including combinations involving Arabic (and its Egyptian variant), English, Turkish, Sinhala, and Mexican Spanish, along with their respective national sign languages (Table 1).

In total, 3 studies reported successful implementation of bidirectional communication between spoken and sign languages [12,34,41]. One study took linguistic differences into account by applying rule-based morphological and syntactic processing to accommodate features specific to Arabic Sign Language—such as subject-initial word order, lack of inflection, and the use of spatial and visual cues—thus enabling transformation into grammatically correct Arabic sentences [40].

Two sensor-based systems also linked corpus construction to the use of facial expression and body movement capture for more precise recognition [12,32]. As anticipated, no glove-based system used such an approach.

Corpus Used to Generate the Language Database for System Development

Image-Based

Overall, most systems relied on corpora composed of isolated words or terms, except for 3 studies, which included full sentences [14,26,28]. Two studies defined their corpus based on real-world contexts [15,26], while others relied on datasets designed by the authors [13,30], consulted health professionals [14], or extracted content from dictionaries [31]. Four systems focused on essential or emergency vocabulary, often including the alphabet, numbers, and isolated signs relevant to health care, obtained through internet searches or expert input [22,25,36].

Sensor-Based

One glove-based system developed its corpus through consultations with health professionals and internet searches [32], while another was based on real emergency call transcripts [34]. A third glove-based system included an extensive set of 600 health-related sentences covering various types (nominal, verbal, and questions) with over 3000 signs [40]. A fourth one focused on basic signs only (fingerspelling the alphabet, numbers, and isolated sign words) [32].

Among depth-sensing systems, 3 built their corpora using isolated words or terms [29,33], while others developed their corpora in consultation with health care professionals [12,35] or constructed its corpus using full sentences commonly used during emergency calls [41].

Health Care Context

Image-Based

Among the 15 studies in this group, 46.7% (7/15) were based on the general hospital environment [13,15,27,30,31,37,42], 36.7% (4/15) focused on emergency services or situations [22,25,26,36], 20% (3/15) were based on primary care consultations [28,38,39], and 6.7% (1/15) on psychiatric interviews [14].

Sensor-Based

In total, 2 of the glove-based systems used emergency services or situations as a basis [32,34], and the remaining one used the general hospital environment [40], while 2 depth-sensing systems were based on care provided in a general hospital environment [29,35], 1 used the context of primary care consultations [12], and 2 used emergency services [33,41].

XSL•FO
RenderX

Approaches and Techniques for System Development

Image-Based

In the image-based systems reviewed, different datasets have been created, involving capturing images in many lighting conditions and backgrounds, recording of signs made by users fluent in sign languages, extracting keyframes from videos [15,22,25-28,33-38], and image stacking [30]. These datasets have been divided into training, validation, and test sets for developing and evaluating the proposed systems [13,14].

Data processing included methods such as hierarchical classification, feature extraction through neural networks, human key point estimation, and pose, hand, and skeleton coordinates of people in videos [14,15,25,27]. Different models were used, such as CNNs, long short-term memory, hidden Markov model (HMM), OpenPose, 3D graph convolutional network, long-term recurrent convolutional network, 1D-CNN, and bidirectional long short-term memory, both individually and in combination, for the classification and prediction of sign language signals [15,22,26,27,30,37-39,42].

Additionally, only 1 paper presents an avatar module that transforms a hearing person's text into sign language [39]. The system comprises 3 modules: a signal recognition module, a speech recognition and synthesis module, and an avatar module. Each module performs specific tasks to ensure the integrated functioning of the system.

Sensor-Based

In sensor-based systems, development methods involved capturing sign language signs with a Kinect, Leap Motion Controller, and cameras on wearable devices such as smart glasses [32-35,41]. Then, the captured data were processed and analyzed to extract relevant characteristics, such as position, movement, and shape of the hand; wrist trajectory; and other specific features of the sign [32,34].

Different techniques have been applied in the development of these systems, including machine learning models such as random forest, naïve Bayes classifier, Ada-Boosting, dynamic time warping, and HMMs [12,29,35]. Furthermore, preprocessing, segmentation, and feature extraction techniques have also been used to improve the effectiveness and accuracy of the systems in detecting and interpreting signs [32-34].

One of the systems also integrated machine translation processes by applying morphological and syntactic analysis to restructure sentence structure and ensure grammatical agreement [40].

Technologies Involved

The technologies involved in system development are shown in Multimedia Appendix 6.

The Infrastructure and Training Required for Use

Image-Based

Desktop and notebook computers, cameras, sensors, and supporting devices and equipment were needed to implement 3 proposed systems [25,27,36,42], whereas a mobile phone was required for 4 of them [27,30,31,42]. Beyond internet connectivity, 1 proposal [29] required downloading apps. In the image-based group, 2 systems explicitly mentioned using a phone's camera [30,31], and 3 mentioned using a webcam or external camera [13,36,37]. None of the studies addressed the details of the training required for the translation system. In total, 3 papers failed to address the structure and training required for use [14,22,26].

Sensor-Based

Three systems, all depth sensing [29,33,35], addressed the need for a computer connected to a power source. Only 1 depth-sensing system pointed out the need for a mobile phone with an integrated camera [41]. One glove-based [34] proposal required downloading an app, and this was the only one that described the use of a phone's camera, additionally requiring augmented reality glasses with a Bluetooth interface. Another one, from the depth-sensing type [33], mentioned the existence of an external camera device. Microsoft Kinect is a standard device for sensor-based systems that uses depth sensing [45]. None of the studies addressed the details of the training required for the translation system. One paper did not discuss the necessary structure and training for implementation [40].

Multidisciplinary Partnerships

Image-Based

Various disciplines were represented among the research teams, who reported back to university departments of informatics [14,30], applied sciences [27,42], engineering [13,14,30,37,38], technology [26,30,31,37], computer science [22,38,39,42], and biomedical physics [37]. Only 4 studies mentioned the participation of linguistic specialists [13,15,26,32] during corpus development or database recording to improve sign language representation, and only 1 mentioned having involved physicians [14] in selecting the words to create the corpus. Along with sign language interpreters, hearing-impaired native sign language speakers [11,14,26] were engaged to record the videos.

Sensor-Based

The sensor-based systems encompassed a great variety of backgrounds, including university departments of statistics [12], AI [12,32], informatics [12,41], applied sciences [12], engineering [29,33,35], technology [33,41], and computer science [32,34,40]. To contribute to the enhancement of sign language representation, some systems involved linguistic specialists [35] in the development of a corpus or database, and also deaf native sign language speakers [32,34] and sign language interpreters worked on recording the videos. Physicians [12,32,35] and deaf individuals [35] were also involved in selecting the words for the corpus.

Sample Size (Sample Size of Individuals to Develop the System, Users, and Content)

Image-Based

The sample size (number of individuals) used to develop the systems ranged from 2 [31] to 26 [22,25,36] among systems that used the same people for development and testing, and 1 study did not inform the sample size used in the development phase [38]. For systems that used different individuals for development and testing, the number used in the development

XSL•FO
RenderX

phase ranged from 14 [26] to 33 [39], and the number of individuals used in the test phase ranged from 2 [39] to 6 [26]. Four studies [27,28,30,42] did not mention whether they used the same individuals for both development and testing. The dataset's contents ranged from 21 scripts [14] to 145,035 videos (Table 2) [39].

Sensor-Based

The sample size (number of individuals) used to develop the systems ranged from 2 [32,40] to 6 [34] among systems that used the same people for development and testing. For systems that used different people for development and testing, the number of people to develop ranged from 2 [40] to 12 [12] (2 studies [33,46] did not inform the sample size in development), and the number of users to test ranged from 2 [39] to 10 [29].

The content of the datasets ranged from 10 videos (5 for the training phase and 5 for the testing phase) [32] to 1260 samples (840 samples for the training set and 420 samples for the testing set) [29]. In total, 2 studies did not inform the size of the content of their datasets (Table 2) [40,41].

Cost

Image-Based

None of the included studies provided quantitative data on the costs associated with the development of their SLR systems, either in terms of technological infrastructure or human resources.

Sensor-Based

As with image-based studies, no quantitative analysis has been reported regarding the costs for developing the SLR systems. However, one of the studies mentioned that the development cost of the device "SmartCall" was much less when compared to those of similar works (which used sensors embedded in smartwatches), suggesting that researchers in low- and middle-income countries could readily prototype their own devices [32].

Testing

Image-Based

The tests have been conducted in simulated environments involving different training networks, varying noise levels, and different scenarios [13,14]. Data splitting strategies into training and testing sets [15,22,25,36-39] and comparing different methods [18] have been used to evaluate the accuracy and effectiveness of the developed systems in detecting and interpreting signs.

Accuracy measures in detecting and interpreting signs ranged from 25% [27] to 100% [42]. Precision ranged from 91.5% [25] to 100% [42] (5 systems) [22,24,25,37,42]. Recall ranged from 90.1% [25] to 100% [42] (4 systems) [24,25,37,42]. F_1-score ranged from 90.7% [25] to 100% [42] (4 systems) [24,25,37,42]. No study presented specificity and sensitivity (Multimedia Appendix 5).

A common difficulty was the failure of some systems to predict similar hand movements, given the proximity of the key points [22,30,31]. For that reason, double-handed signs were more accurately classified than single-handed ones [22].

Sensor-Based

The tests were conducted using datasets, in which different users performed sign language signs repeatedly in different scenarios [12,32,35]. Testing also involved splitting data into training, validation, and test sets and evaluating the performance of the systems [29,32,33]. Additionally, usability and user acceptance tests have been conducted to assess the effectiveness and practicality of the systems in real-life situations [34,35].

Accuracy measures in detecting and interpreting signs ranged from 80.5% [33] to 99.8% [41] (8 systems) [12,29,32-35,40,41], being 80.5% [33] to 99.8% [41] for systems with depth-sensing sensors and 72% [32] to 92% [40] for systems with hand glove sensors. Only 1 study [12] presented precision (">90%"). Recall ranged from 90.1% [25] to 98.8% [24] (3 systems) [24,25,37]. F_1-score ranged from 85% [32] to 88.6% [12] (2 systems) [12,32]. Only 1 study [12] presented specificity and sensitivity (99.8% and 87.9%, respectively; Multimedia Appendix 5).

Evaluation or Use in Real Context

Image-Based

Among image-based studies, authors of 11 of the studies [13-15,22,25-28,31,36,38,39] did not respond to the emails sent by the research team, so it was impossible to update the current status of their systems. Authors of 2 of the studies informed that their systems have not been implemented in real scenarios or commercial settings [30,42]. One author mentioned that his team is in the process of developing a smartphone app for interpreting medical signs and that they will proceed to real-world implementation once it is ready [37].

Sensor-Based

Among sensor-based studies, the authors of 3 of the studies [29,33,41] did not respond to the emails sent by the research team, so it was not possible to update the current status of their systems. Authors of 4 other studies informed that their systems have never been deployed or commercialized [12,34,35,40]. Authors of the other study announced that their system is currently undergoing major design improvements and patent applications and that they hope to make it commercial very soon [32]. Authors of one of the studies reported that they open-sourced their dataset and their source code implementation to facilitate adoption of their system and accelerate technology transfer among research groups [32].

Emergency Responses Capability

Regarding emergency circumstances, 13 studies [12-15,22,25,26,28,30,32,33,35,36] analyzed the capabilities of the system to respond in urgent situations according to various parameters.

Image-Based

Several image-based systems assessed key performance indicators relevant to urgent health care contexts, including recognition speed for signs and phrases [12,13], the system's ability to reliably alert health care providers [12,14,34], and the clarity of communication output [12,26]. Some studies also

XSL•FO
RenderX

addressed design considerations critical for emergency use, such as error tolerance (the system's resilience to technical failures) [25], and the inclusion of medical vocabulary specific to emergency scenarios [30]. One system also incorporated structured protocols aimed at guiding interactions with deaf patients during emergencies [15]. However, a common limitation was the small size of the vocabulary recognized, which restricted the ability to handle diverse emergency dialogues [22,36].

Sensor-Based

Sensor-based systems also demonstrated potential for real-time emergency support. For instance, one system was specifically designed to improve access to emergency services through depth-sensing technology [41], while another emphasized robustness through error-tolerant functionality [12]. Similar to image-based tools, sensor-based systems faced limitations related to the restricted number of medical terms recognized, which could hinder effective communication during critical situations [32,34].

Technological Characteristics

Image-Based

None of the studies reported on critical aspects such as system reliability (graceful degradation or recovery after crashes), data security, video storage, or deletion policies after translation, or objective metrics comparing communication outcomes with and without the system. Translation time was superficially addressed in 1 paper [22]. Authors compared 3 different models (a 3D CNN; a pretrained visual geometry group 16 layers and a recurrent neural network with a long short-term memory scheme; and a model based on YOLO v5, an advanced object detection algorithm). When comparing the proposed techniques, the YOLO-based model was faster, with an image processing rate of 40 - 90 frames per second. Therefore, it could be used for emergency sign recognition with only a few milliseconds of delay.

Sensor-Based

Only 3 studies addressed translation time in sensor-based systems [29,34,41]. One study compared different machine learning classifiers and reported the following processing times: 8.09 seconds for the naïve Bayes classifier, 15.3 seconds for random forest, 17.63 seconds for dynamic time warping, and 20.8 seconds for HMM [29]. Another system reported an average latency of 0.55 seconds for translating 550 signs, with mobile translation taking 122 milliseconds and phoneme streaming 206 milliseconds to render the animation [34]. The third study reported that the system required 3 to 4 seconds from the end of the gesture to generate the output. In test scenarios, it consistently outperformed Google Cloud, with response times of 2.5, 2.6, and 2.2 seconds, compared to 4.1, 4.2, and 3.7 seconds for Google Cloud [41].

As with image-based systems, none of the sensor-based systems discussed reliability, data privacy, postprocessing video storage, or communication effectiveness metrics.

User Experience

Image-Based

User experience was not explicitly evaluated in any of the image-based studies. A few of them commented on some factors related to the usability of the developed systems, both about their facilities and difficulties. These feedbacks are listed in Multimedia Appendix 5.

Sensor-Based

The evaluation of users' experience was subjectively reported in 4 sensor-based studies [34,35,40,41]. The first one, classified as glove-based, conducted a user study on the signers' experience. The study involved 12 participants, and Quality of Experience was rated on a 5-point scale, ranging from 0=worst evaluation)to 5=best evaluation, to assess the usefulness of the system. The obtained results were 4.2 for accessibility, 4.3 for usability, and 4.6 for overall experience [34].

Additionally, a second study, classified as depth-sensing, developed a questionnaire score using the same scale for 5 statements. In total, 5 participants used the system, and afterward, responded to the questionnaire. The questionnaire results obtained for each statement question were 4.85 for "I was able to express my situation using HospiSign," 4.85 for "HospiSign is easy to use," 5.00 for "Most people could learn to use HospiSign quickly," 4.45 for "Options were sufficiently clear," and 5.00 for "HospiSign would help the hearing impaired" [35].

Another study, from the glove-based category, provided a manual evaluation performed by 3 native Arabic speakers. Each of the 600 sentences was classified by them as understandable, somehow understandable, and not understandable according to grammatical and semantic metrics. The respective results were 80%, 12% and 8%, demonstrating an acceptable translation provision in 92% of the sample [40].

Additionally, a third study, classified as depth-sensing, conducted a user satisfaction survey with 10 users, and the results were classified as high, medium, and low satisfaction. The results reported were user registration: high, text-to-call: high, call-to-sign: medium, sign-to-call: medium, and prerecorded message: high [41].

Finally, a fourth study, classified as depth-sensing as well, developed a questionnaire score using the same scale for 5 statements. In total, 5 participants used the system, and afterward, responded to the questionnaire, and the results obtained for each statement were 4.85 for "I was able to express my situation using HospiSign," 4.85 for "HospiSign is easy to use," 5.00 for "Most people could learn to use HospiSign quickly," 4.45 for "Options were sufficiently clear," and 5.00 for "HospiSign would help the hearing impaired" [35].

There were also some comments on the usability of the developed systems, concerning the user experience. These feedbacks are listed in Multimedia Appendix 5.

Ethical Issues Addressed

Ethical concerns related to the use of communication systems for deaf and hard-of-hearing individuals were largely

XSL•FO
RenderX

underexplored across the included studies. Importantly, none of the included studies discussed broader ethical implications of the systems themselves, such as data privacy and security, the handling and potential storage of video or biometric information, or the autonomy and psychological impact of using these technologies. Despite their role in sensitive health care contexts, no system reported design strategies for ensuring user dignity, minimizing potential emotional distress, or addressing users' preferences regarding data management. Additionally, none of the studies evaluated whether addressing ethical aspects influenced system usability or user trust, including transparency, control over personal data, and informed decision-making.

Discussion

Principal Findings

The development of SLR systems represents a promising step toward improving communication between deaf individuals and health care providers. Despite the vast literature in SLR systems, many computer science studies focus on creating algorithms for recognizing (and less often translating) signed content [1]. These teams often lack deaf individuals with firsthand experience of the challenges the technology aims to address, and they may not fully understand the linguistic complexities of the language. Additionally, the algorithms are typically trained on datasets that do not reflect real-world scenarios, making these 1D approaches to sign language processing of limited practical value [1]. In our analysis, the comprehensive search resulted in over 21,000 papers, but only 23 studies met our inclusion criteria. All analyzed systems are in the development and testing stage, with no real application yet, and this review highlights significant challenges that hinder the implementation of these systems in real-world settings. These challenges include variability in methodologies, limitations in dataset quality, and the underrepresentation of key communication elements, such as bidirectionality and facial expressions. Furthermore, the absence of cost analysis and the lack of multilingual integration reflect gaps in the current body of research. To address these issues, it is essential to evaluate both the technical aspects and the practical implications of SLR systems, aiming for inclusive and effective solutions.

Image-based systems accounted for the majority of the studies. This predominance highlights the versatility of image-based systems, which capture facial expressions, head movements, lip reading, and hand gestures using simple devices like mobile phone cameras [47]. In contrast, it is possible to infer that sensor-based systems, which rely on advanced technologies such as depth-sensing cameras and sensor-embedded gloves, incur higher costs [48]. However, a critical gap identified across studies was the lack of clear cost-related information. Only 1 study briefly mentioned that its device, "SmartCall," was more affordable compared to similar solutions, suggesting its feasibility for low- and middle-income countries [32].

Given the lack of sufficient cost-related data in the studies reviewed, it would be challenging to provide accurate inferences regarding the cost-effectiveness of these technologies. Furthermore, as costs vary significantly depending on the country and context, it would be difficult to generalize these findings across all regions. This is a notable limitation, and we believe that future studies should address this gap by providing more detailed and context-specific cost analyses.

Corpus Definition and Number of Samples

Corpus definition and dataset size were highly heterogeneous across the reviewed studies, reflecting diverse development contexts and objectives. Most systems relied on isolated words, limiting applicability in real-world health care communication, while some incorporated phrases or real clinical contexts, occasionally guided by health professionals or dictionaries [14,26,28,40]. A few image-based systems also integrated facial expressions and body movements, enhancing linguistic representation [15,26-28,30,31,39]. Among sensor-based systems, corpus construction lacked multimodal features, particularly in glove-based systems, which are limited in capturing facial and body expressions.

Sample sizes ranged widely from as few as 10 to over 145,000 videos, raising concerns about the adequacy of data for training reliable models. This lack of standardization in corpus design and dataset size limits the comparability, scalability, and generalizability of SLR systems in clinical settings. Establishing minimum standards for corpus development could improve system robustness and support broader implementation in health care.

Key Aspects of Effective Communication

Bidirectionality and facial expression recognition are 2 critical components for effective communication, but were not included in the majority of publications. Most systems were unidirectional, translating only from sign language to written text, impairing its applicability in real-world contexts. While such systems may enable health care practitioners to understand what a deaf patient is communicating, they fail to facilitate communication in the opposite direction, thereby undermining true interactive dialogue.

Facial expression recognition is a vital element for capturing linguistic prosody, and although face landmarks were considered in 8 papers [15,26-28,30,31,39,40], only 1 of them explicitly addressed it as an indicator of grammatical tense [28]. This omission hinders the natural and nuanced communication required in health care contexts.

Notably, only 3 studies explicitly acknowledged the cultural and structural distinctions between sign languages and oral languages [15,28,40]. This omission is significant, as sign languages are not mere gestural representations of spoken language, but complete linguistic systems with their own grammar, syntax, and cultural context. Failure to consider these differences can lead to inaccurate or overly literal translations, potentially compromising communication quality and user trust in assistive systems. The limited attention to these linguistic specificities suggests that many of the reviewed systems may have been developed from a technical standpoint, with insufficient involvement of deaf communities or sign language experts.

Given the numerous challenges involved, it is nearly impossible to develop SLR tools that achieve 100% accuracy with a large

XSL•FO
RenderX

vocabulary [49]. For example, the same sign can have large changes in shape when it is in different locations in the sentence [50]. In the studies analyzed, only 1 demonstrated perfect accuracy (100%), but this was based on an image-based system, which was tested only with isolated words [12]. Studies that involved a relevant number of sentences and words tended to perform worse, with accuracy sometimes too low to enable effective communication [14,26].

Multilingual Challenges

Sign languages are not universal, and they are not mutually intelligible, and they can present significant differences between countries, and sometimes even within the same country [51]. While some studies proposed extending their systems to multiple languages [12,28,41], none successfully implemented multilingual translation. The inherent diversity of sign languages—each with unique grammar, vocabulary, and body configurations—poses substantial challenges to creating universal systems [27,52,53]. Furthermore, there was a predominance of English corpora [25,30-32,34,36,37]. This is probably driven by high research investments, as well as the global reach and accessibility of English [54], which not only facilitates better comprehension within the medical academic community but also enables broader dissemination and open sharing of the corpus [55]. This indicates a gap in accessibility for non-English–speaking regions, emphasizing the importance of expanding datasets to include diverse languages and cultures. To address this gap, it is crucial to foster collaboration between researchers from different countries, sharing sign language corpora and visual data, so that a truly multilingual system can be developed.

Multidisciplinary Partnerships

As Bragg et al [1] described, an interdisciplinary approach is crucial to sign language processing. Representatives of the deaf community and health care practitioners must be involved in the team to better understand the needs of the community and the specific purposes for which the technology is intended. Linguistics plays a key role in identifying the structures of sign languages that the algorithms need to process. Natural language processing and machine translation offer valuable techniques for modeling, analyzing, and translating. Computer vision is necessary for recognizing signed content, while computer graphics are essential for generating it. Finally, human-computer interaction and design are vital for developing comprehensive systems that meet the needs of the community and integrate seamlessly into people's daily lives [1].

The inclusion of multidisciplinary teams was a notable feature in several studies, combining expertise from linguistics, inclusive design, AI, and health care practitioners. However, the participation of deaf community representatives was limited, potentially undermining cultural and linguistic relevance.

Ethical and Data Security Concerns

As these systems are intended to support real-time communication in clinical settings, their design must go beyond technical optimization to consider ethical issues such as privacy, autonomy, and emotional safety, dimensions largely overlooked in the reviewed studies. The reviewed literature does not adequately address the psychological impact of using these systems. Issues such as the psychological effects on users, especially those who are deaf or hard of hearing, were not discussed in any of the studies. These concerns are particularly important, as assistive technology should promote autonomy and well-being without causing additional stress or emotional burdens for users [56].

Data security is another critical concern that received little attention. Many SLR systems collect sensitive information, such as facial images, gestures, and body movements, which may contain identifiable or private data. However, few studies addressed how these data are stored, protected, or deleted. In health care contexts—where confidentiality is paramount—this lack of attention to privacy and data protection represents a significant oversight.

These omissions highlight a critical gap: as assistive communication technologies are implemented in real-world health care settings, particularly among structurally marginalized populations like the deaf community, ethical design and responsible implementation are essential. Future studies should incorporate ethical considerations from the early stages of system development and report these aspects transparently.

User Experience

Overall, few studies provided reports on the users' experience with the developed systems, which were structurally evaluated and described in only 2 of them [22,36], and only with a small number of participants (12 in Areeb et al [22] and 5 in Adithya and Rajesh [36]). The other studies focused on usability metrics of the developed systems [12,14,15,25,29]. The issues raised corroborate difficulties described in the literature. All recognition systems face the challenge of hand tracking and struggle with self-occlusion between fingers, with additional difficulties in single-hand signals [57]. Such an aspect was mentioned especially by the authors of image-based systems, since in their approaches they did not attempt 3D perception to mitigate occlusion [14,15]. Furthermore, another significant obstacle is the interpersonal variation between signers, which includes not only their physical characteristics but also the way they perform gestures. Minor differences in signaling are important constraints for AI and recognition tools, justifying the reports regardless of the system's type [12,26].

Although the studies that presented user evaluation [34,35] reported that the systems performed well, showing positive impacts on accessibility and effective communication, the evidence is limited, as only 2 studies were conducted, and both involved a small number of users. User experience goes beyond the system's accuracy and usability, and according to International Organization for Standardization [58], it refers to a "person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service," including "all the users' emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviours and accomplishments that occur before, during and after use." Thus, generating a positive user experience for the deaf users is essential for the adoption and use of the proposed systems. While studies typically focus on accuracy and usability metrics, it is equally important to evaluate other factors, such as learning

XSL•FO
RenderX

cost and response speed, which can significantly affect user adoption, especially in health care contexts, where ease of use and time efficiency are critical.

Saeed et al [59], in a review on system-based sensory gloves for SLR, reported that user comfort was a challenge, as using a sensory glove required the user to wear a bulky glove containing sensors, cables, and a circuit board, which limited the user's hand mobility. However, we believe this issue is gradually being reduced, as sensors continue to evolve, becoming smaller and more efficient. Furthermore, the ability to connect these sensors via Bluetooth has further minimized the bulkiness, offering more comfort and flexibility for the user.

The design of interfaces for automatic sign language translation systems, specifically their integration within web-based and social media applications, necessitates careful consideration of numerous factors. Research conducted by Debevc et al [46] and Kožuh et al [60] addresses methodologies for enhancing the interaction experience for deaf and hard-of-hearing users within digital communication platforms [46,60]. The authors examine methods to improve accessibility [46,60].

A recurring concern highlighted in these studies is the potential for picture-in-picture windows to disrupt the user experience for nondeaf individuals. Therefore, the implementation of picture-in-picture displays should be judiciously considered and minimized. The domain of automatic sign language translation is characterized by rapid advancements and multifaceted complexities. The techniques used to optimize the user experience for these systems, while of significant relevance, warrant a dedicated review to comprehensively assess their efficacy and implications.

Limitations and Strengths of the Review

This review is primarily qualitative, as the diversity in methodologies, technologies, and parameter calculations precluded quantitative synthesis. Additionally, there are no standardized tools to assess the risk of bias in these types of studies included. Furthermore, this review focused on systems that recognize sign language. Although these systems do not fully represent the portion of the deaf community that prefers lip reading, writing, or other methods of communication, they do include individuals who use sign language. Finally, 4 publications could not be included, as they were not retrieved despite at least 3 contact attempts regarding 2 studies and due to the unavailability of authors' contact details of the other 2 studies.

The strengths of this systematic review include prospective registration and publication of a protocol [18,19]. Furthermore, it adheres to the Cochrane Guidelines and the PRISMA statement [20,21], and it has a broad search strategy encompassing multiple languages and publication types. A multidisciplinary team, including a deaf linguist, ensured diverse perspectives and representation.

Additionally, this study is innovative, as it focuses specifically on SLR for the health context, an area in which social inclusion and understanding differences are extremely important. Although other reviews were previously published, the majority of them cannot be considered true systematic reviews [49,59,61,62]. They lack a prepublication of a protocol, a robust methodology, a broad search strategy, paired screening and data extraction, and comprehensiveness. They usually focus on technical aspects of a broader and more general analysis of sign language translation systems and use restricted publication periods and language (English only) in the search. Overall, they did not include a search in MEDLINE, which is a crucial database for health-related research, and presented methodological shortcomings, such as not mentioning the use of paired reviewers for study selection and data extraction. Furthermore, they have not reported extracting data on user experience, cost, and data security. Unlike these more general reviews, our study is unique in focusing on the health care context. Therefore, our study fills these gaps, providing a more robust, detailed, and comprehensive contribution specifically to the health field.

Final Considerations

Image-based systems dominated the field, demonstrating greater accessibility but facing challenges with lighting conditions and environmental variability. While sensor-based systems showed higher precision, their reliance on specialized equipment limits scalability. Critical aspects, such as bidirectional communication and facial expression recognition, remain underexplored, hindering the practical implementation of these systems in health care settings.

Based on the evidence reviewed, effective SRL systems for real-world health care scenarios are scarce. Future research should focus on developing adaptive systems capable of recognizing diverse sign languages and addressing the varying communication needs of the deaf community. It should prioritize inclusive and participatory design processes that respect the linguistic integrity of sign languages and address sociocultural nuances critical to effective communication. Expanding datasets, incorporating user feedback, and fostering multidisciplinary collaborations will be pivotal for creating inclusive and scalable solutions.

Additionally, future research should incorporate a comprehensive assessment of the proposed systems, addressing both technical and user-centered aspects. This assessment should consider device compatibility (especially for smartphone-based systems), required learning time, translation accuracy across different signers and accents [63], user trust in automated translation, its effect on self-expression, and privacy implications. These evaluations should not only be conducted once the system is fully implemented but also throughout the development process (ie, formative evaluation) [64].

Another significant gap identified in the literature is the lack of a standardized evaluation framework for SLR systems. The inclusion of consistent and well-defined indicators, such as user satisfaction and economic viability, would provide a more comprehensive understanding of the systems' effectiveness and their potential for widespread adoption. User satisfaction is a crucial success factor, and including economic indicators would help assess whether the systems can be feasibly adopted in different settings, especially in low- and middle-income countries. The implementation of the framework would not only enhance the comparability of findings but also provide valuable

XSL•FO
RenderX

insights for policymakers and stakeholders in the health care industry.

Furthermore, we suggest that the development of future SLR solutions involves experts in ethics, law, and data security to ensure compliance with legal and ethical standards and to protect users during the implementation of these technologies. Studies should provide more detailed information about how ethical and data security issues are managed and the potential impacts on user trust and adoption.

One notable strength identified in this review was the initiative of a research group to open-source both their dataset and system code [32]. This practice promotes transparency, reproducibility, and collaboration—key elements for the advancement of SLR systems. By allowing other researchers to access and build upon existing work, open data and open-source tools can accelerate the refinement of models and the development of more accurate and user-friendly solutions. In the long term, such collaborative practices may contribute to the creation of robust, scalable systems that are truly usable in health care settings, bridging communication gaps and improving care for deaf individuals. Encouraging the adoption of open science principles within the SLR research community could significantly enhance innovation and standardization efforts in this emerging field.

Conclusions

This review highlights the ongoing development of communication systems designed to assist deaf individuals who use sign language to improve their interaction with health care providers. There was a predominance of image-based approaches over sensor-based ones, though both have demonstrated substantial variability in accuracy in recognizing and interpreting signs. However, critical aspects such as bidirectional communication and the recognition of facial expressions, which are essential for effective communication, were notably absent in most studies. None of the systems has reported to have addressed all aspects critical to integration into health care settings. These findings underscore the need for further research, especially regarding the practical implementation of these systems, their usability, and their overall effect on the quality of care for deaf patients.

Acknowledgments

This manuscript was prepared by the authors without the assistance of artificial intelligence tools.

Funding

This study was funded in part by the Institute for Health Assessment and Translation for Highly Relevant Chronic and Neglected Diseases (IATS CARE)/National Institute of Science and Technology for Health Technology Assessment (Instituto de Avaliação de Tecnologias em Saúde) or National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico; grant 465518/2014-1), Financiadora de Estudos e Projetos (grant 01.21.0097.00), and SignumWeb Comunicação Inclusiva Ltda, as part of the project “Captar-Libras: Sistema de Comunicação por vídeos para surdos aplicado ao pré-atendimento médico.” MSM, MFMC, and LRV were supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (grants 311742/2025-4, 316841/2021-8, and 134878/2023-0, respectively). The funding sources had no role in the design of this study and will not have any role during its execution, analyses, interpretation of the data, or decision.

Data Availability

All data generated or analyzed during this study are included in this published paper and its Multimedia Appendices 1,2,3,4,5,6.

Authors' Contributions

MSM, LFRdO, LRV, ROP, and ZSNR developed a draft of the literature search, which LFRdO and LRV executed. MSM, ROP, ZSNR, and MFMC guided the construction of the protocol, and MSM is a guarantor of the review. LFRdO, LRV, LMMSR, GTdSS, NSS, and MRdC contributed to the work’s extraction, analysis, or interpretation of data. MSM, LFRdO, LRV, LMMSR, MRdC, and GTdSS drafted the manuscript; all the other authors performed critical revisions. ELAB and RAAC are professors of linguistics and experts in sign language. They reviewed the entire manuscript to ensure the language was consistent with the literature on disability and deafness. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File, 2988 KB - jmir_v28i1e70417_app1.docx]

Multimedia Appendix 2
The definitions used regarding the type of sign language recognition system, the corpus formation, and the health context.

XSL•FO
RenderX

[DOCX File, 17 KB - jmir_v28i1e70417_app2.docx]

Multimedia Appendix 3
Codebook.
[DOCX File, 2489 KB - jmir_v28i1e70417_app3.docx]

Multimedia Appendix 4
Studies not retrieved.
[DOCX File, 14 KB - jmir_v28i1e70417_app4.docx]

Multimedia Appendix 5
Development and testing of different sign language recognition systems.
[DOCX File, 30 KB - jmir_v28i1e70417_app5.docx]

Multimedia Appendix 6
Technologies necessary for sign language recognition system implementation.
[DOCX File, 16 KB - jmir_v28i1e70417_app6.docx]

Checklist 1
PRISMA checklist.
[DOCX File, 32 KB - jmir_v28i1e70417_app7.docx]

References

1. Bragg D, Koller O, Bellard M, et al. Sign language recognition, generation, and translation: an interdisciplinary perspective. Presented at: Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility; Oct 28-30, 2019. [doi: 10.1145/3308561.3353774]
2. Kendon A. Gesture Visible Action as Utterance: Cambridge University Press; 2004.
3. Deafness and hearing loss. World Health Organization. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss [accessed 2024-11-14]
4. World Federation of the Deaf. URL: https://wfdeaf.org [accessed 2025-12-05]
5. Ubido J, Huntington J, Warburton D. Inequalities in access to healthcare faced by women who are deaf. Health Soc Care Community 2002 Jul;10(4):247-253. [doi: 10.1046/j.1365-2524.2002.00365.x] [Medline: 12193168]
6. Kuenburg A, Fellinger P, Fellinger J. Health care access among deaf people. J Deaf Stud Deaf Educ 2016 Jan;21(1):1-10. [doi: 10.1093/deafed/env042] [Medline: 26405210]
7. Mansutti I, Achil I, Rosa Gastaldo C, Tomé Pires C, Palese A. Individuals with hearing impairment/deafness during the COVID-19 pandemic: a rapid review on communication challenges and strategies. J Clin Nurs 2023 Aug;32(15-16):4454-4472. [doi: 10.1111/jocn.16572] [Medline: 36320127]
8. Rannefeld J, O'Sullivan JL, Kuhlmey A, Zoellick JC. Deaf and hard-of-hearing patients are unsatisfied with and avoid German health care: results from an online survey in German Sign Language. BMC Public Health 2023 Oct 18;23(1):2026. [doi: 10.1186/s12889-023-16924-w] [Medline: 37848898]
9. Rastgoo R, Kiani K, Escalera S. Sign language recognition: a deep survey. Expert Syst Appl 2021 Feb;164:113794. [doi: 10.1016/j.eswa.2020.113794]
10. Papastratis I, Chatzikonstantinou C, Konstantinidis D, Dimitropoulos K, Daras P. Artificial intelligence technologies for sign language. Sensors (Basel) 2021 Aug 30;21(17):5843. [doi: 10.3390/s21175843] [Medline: 34502733]
11. Parton BS. Sign language recognition and translation: a multidisciplined approach from the field of artificial intelligence. J Deaf Stud Deaf Educ 2006;11(1):94-101. [doi: 10.1093/deafed/enj003] [Medline: 16192405]
12. Sosa-Jiménez CO, Ríos-Figueroa HV, Solís-González-Cosío AL. A prototype for Mexican Sign Language recognition and synthesis in support of a primary care physician. IEEE Access 2022;10:127620-127635. [doi: 10.1109/ACCESS.2022.3226696]
13. Xia K, Lu W, Fan H, Zhao Q. A sign language recognition system applied to deaf-mute medical consultation. Sensors (Basel) 2022 Nov 24;22(23):9107. [doi: 10.3390/s22239107] [Medline: 36501809]
14. Pikoulis EV, Bifis A, Trigka M, Constantinopoulos C, Kosmopoulos D. Context-aware automatic sign language video transcription in psychiatric interviews. Sensors (Basel) 2022 Mar 30;22(7):2656. [doi: 10.3390/s22072656] [Medline: 35408270]
15. da Silva DRB, de Araújo TMU, do Rêgo TG, Brandão MAC, Gonçalves LMG. A multiple stream architecture for the recognition of signs in Brazilian sign language in the context of health. Multimed Tools Appl 2024;83(7):19767-19785. [doi: 10.1007/s11042-023-16332-7]

XSL•FO
RenderX

16. S Kumar S, Wangyal T, Saboo V, Srinath R. Time series neural networks for real time sign language translation. Presented at: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA); Dec 17-20, 2018. [doi: 10.1109/ICMLA.2018.00043]
17. Alaghband M, Maghroor HR, Garibay I. A survey on sign language literature. Mach Learn Appl 2023 Dec;14:100504. [doi: 10.1016/j.mlwa.2023.100504]
18. Sign language recognition system for deaf patients: a systematic review. Open Science Framework. 2023. URL: https://osf.io/fpemr/overview [accessed 2026-01-28]
19. Marcolino MS, Oliveira LFR, Valle LR, et al. Sign language recognition system for deaf patients: protocol for a systematic review. JMIR Res Protoc 2025 Jan 23;14(1):e55427. [doi: 10.2196/55427] [Medline: 39847417]
20. Higgins JPT. Cochrane Handbook for Systematic Reviews of Interventions Version 65 (Updated August 2024): Cochrane; 2024.
21. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
22. Areeb QM, Nadeem M, Alroobaea R, Anwer F. Helping hearing-impaired in emergency situations: a deep learning-based approach. IEEE Access 2022;10:8502-8517. [doi: 10.1109/ACCESS.2022.3142918]
23. Areeb QM, Nadeem M. Deep learning based hand gesture recognition for emergency situation: a study on Indian Sign Language. Presented at: 2021 International Conference on Data Analytics for Business and Industry (ICDABI); Oct 25-26, 2021. [doi: 10.1109/ICDABI53623.2021.9655842]
24. da Silva DRB, Araujo TMU, Rêgo TD, Brandão MAC. A two-stream model based on 3D convolutional neural networks for the recognition of brazilian sign language in the health context. Presented at: WebMedia '20: Proceedings of the Brazilian Symposium on Multimedia and the Web; Nov 30 to Dec 4, 2020. [doi: 10.1145/3428658.3430980]
25. Das S, Biswas SK, Purkayastha B. Indian Sign Language recognition system for emergency words by using shape and deep features. Presented at: 2023 11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON); Feb 10-11, 2023. [doi: 10.1109/IEMECON56962.2023.10092312]
26. Ko SK, Kim CJ, Jung H, Cho C. Neural sign language translation based on human keypoint estimation. Appl Sci (Basel) 2019;9(13):2683. [doi: 10.3390/app9132683]
27. Barrientos-Villalta GF, Quiroz P, Ugarte W. Peruvian Sign Language recognition using recurrent neural networks. Presented at: International Conference on Advanced Research in Technologies, Information, Innovation and Sustainability; Sep 12-15, 2022. [doi: 10.1007/978-3-031-20319-0_34]
28. Ramírez Sánchez JE, Rodríguez AA, Mendoza MG. Real-time Mexican Sign Language interpretation using CNN and HMM. Presented at: Mexican International Conference on Artificial Intelligence; Oct 25-30, 2021. [doi: 10.1007/978-3-030-89817-5_4]
29. Hisham B, Hamouda A. Supervised learning classifiers for Arabic gestures recognition using Kinect V2. SN Appl Sci 2019 Jul;1(7):1-21. [doi: 10.1007/s42452-019-0771-2]
30. Gandhi J, Gandhi P, Gosar A, Chaudhari S. Video recognition techniques for Indian Sign Language in healthcare domain. Presented at: 2021 2nd International Conference for Emerging Technology (INCET); May 21-23, 2021. [doi: 10.1109/INCET51464.2021.9456116]
31. Uchil AP, Jha S, Sudha BG. Vision-based deep learning approach for dynamic Indian Sign Language recognition in healthcare. Presented at: Computational Vision and Bio-Inspired Computing: ICCVBIC 2019; Sep 25-26, 2019. [doi: 10.1007/978-3-030-37218-7_43]
32. Deji Dere M, Dere RO, Adesina A, Yauri AR. SmartCall: a real-time, sign language medical emergency communicator. Presented at: 2022 5th Information Technology for Education and Development (ITED); Nov 1-3, 2022. [doi: 10.1109/ITED56637.2022.10051420]
33. Sarhan NA, El-Sonbaty Y, Youssef SM. HMM-based Arabic Sign Language recognition using Kinect. Presented at: 2015 Tenth International Conference on Digital Information Management (ICDIM); Dec 23-24, 2015. [doi: 10.1109/ICDIM.2015.7381873]
34. Guo Y, Zhao J, Ding B, et al. Sign-to-911: emergency call service for sign language users with assistive AR glasses. Presented at: ACM MobiCom '23: Proceedings of the 29th Annual International Conference on Mobile Computing and Networking; Oct 2-5, 2023. [doi: 10.1145/3570361.3613260]
35. Süzgün M, Özdemir H, Camgöz N, et al. HospiSign: an interactive sign language platform for hearing impaired. J Naval Sci Eng 2015;11(3):75-92 [FREE Full text]
36. Adithya V, Rajesh R. Hand gestures for emergency situations: a video dataset based on words from Indian Sign Language. Data Brief 2020 Aug;31:106016. [doi: 10.1016/j.dib.2020.106016] [Medline: 32715044]
37. Ihsan MA, Eram AF, Nahar L, Kadir MA. MediSign: an attention-based CNN-BiLSTM approach of classifying word level signs for patient-doctor interaction in hearing impaired community. IEEE Access 2024;12:33803-33815. [doi: 10.1109/ACCESS.2024.3370684]
38. Das HV, Mohan K, Paul L, Kumaresan S, Nair CS. Transforming consulting atmosphere with Indian Sign Language translation. Multimed Tools Appl 2024;83(5):13543-13555. [doi: 10.1007/s11042-023-15214-2]

XSL•FO
RenderX

39. Faisal M, Alsulaiman M, Mekhtiche M, et al. Enabling two-way communication of deaf using Saudi Sign Language. IEEE Access 2023;11:135423-135434. [doi: 10.1109/ACCESS.2023.3337514]
40. Luqman H, Mahmoud SA. A machine translation system from Arabic Sign Language to Arabic. Univ Access Inf Soc 2020 Nov;19(4):891-904. [doi: 10.1007/s10209-019-00695-6]
41. Dewasurendra D, Kumar A, Perera I, Jayasena D, Thelijjagoda S. Emergency communication application for speech and hearing-impaired citizens. Presented at: 2020 From Innovation to Impact (FITI); Dec 15-15, 2020. [doi: 10.1109/FITI52050.2020.9424899]
42. Bellil L, Ghiri MG, Boulesnane A. Empowering deaf community in healthcare communication: 1D-CNN-based Algerian Sign Language recognition system. Presented at: 2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS); Apr 24-25, 2024. [doi: 10.1109/PAIS62114.2024.10541233]
43. Vosk: speech recognition toolkit. Alphacephei. URL: https://alphacephei.com/vosk/ [accessed 2026-03-24]
44. Arabic language disambiguation for natural language processing applications. Columbia University. URL: https://inventions.techventures.columbia.edu/technologies/arabic-language-disambiguation--cu14012 [accessed 2026-03-24]
45. Shotton J, Fitzgibbon A, Cook M, et al. Real-time human pose recognition in parts from single depth images. Presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jun 20-25, 2011. [doi: 10.1109/CVPR.2011.5995316]
46. Debevc M, Kosec P, Holzinger A. Improving multimodal web accessibility for deaf people: sign language interpreter module. Multimed Tools Appl 2011 Aug;54(1):181-199. [doi: 10.1007/s11042-010-0529-8]
47. Madhiarasan M, Roy PP. A comprehensive review of sign language recognition: different types, modalities, and datasets. arXiv. Preprint posted online on Apr 7, 2022. [doi: 10.48550/arXiv.2204.03328]
48. Haji Mohd MN, Mohd Asaari MS, Lay Ping O, Rosdi BA. Vision-based hand detection and tracking using fusion of kernelized correlation filter and single-shot detection. Appl Sci (Basel) 2023;13(13):7433. [doi: 10.3390/app13137433]
49. Al-Qurishi M, Khalid T, Souissi R. Deep learning for sign language recognition: current techniques, benchmarks, and open issues. IEEE Access 2021;9:126917-126951. [doi: 10.1109/ACCESS.2021.3110912]
50. Yang R, Sarkar S, Loeding B. Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming. IEEE Trans Pattern Anal Mach Intell 2009;32(3):462-477. [doi: 10.1109/TPAMI.2009.26]
51. Amin MS, Rizvi STH, Hossain MM. A comparative review on applications of different sensors for sign language recognition. J Imaging 2022 Apr 2;8(4):98. [doi: 10.3390/jimaging8040098] [Medline: 35448225]
52. Figg B. Deafness and other communication disorders. J Consum Health Internet 2023 Apr 3;27(2):199-209. [doi: 10.1080/15398285.2023.2204717]
53. Bihari Dash D. Significance of globalisation and English language. Int J Stud Engl Lang Lit 2022;10(5):10-16. [doi: 10.20431/2347-3134.1005002]
54. Nye B, Jessy Li J, Patel R, et al. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. Presented at: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Jul 15-20, 2018. [doi: 10.18653/v1/P18-1019]
55. Zheng L, Liang B, Jiang A. Recent advances of deep learning for sign language recognition. Presented at: 2017 International Conference on Digital Image Computing; Nov 29 to Dec 1, 2017. [doi: 10.1109/DICTA.2017.8227483]
56. Jiménez-Arberas E, Díez E. Psychosocial impact of assistive devices and other technologies on deaf and hard of hearing people. Int J Environ Res Public Health 2021 Jul 7;18(14):7259. [doi: 10.3390/ijerph18147259] [Medline: 34299710]
57. Adeyanju IA, Bello OO, Adegboye MA. Machine learning methods for sign language recognition: a critical review and analysis. Intell Syst Appl 2021 Nov;12:200056. [doi: 10.1016/j.iswa.2021.200056]
58. Ergonomics of human-system interaction: Part 210: Human-centered design for interactive systems. International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-1:v1:en [accessed 2025-04-13]
59. Saeed ZR, Zainol ZB, Zaidan BB, Alamoodi AH. A systematic review on systems-based sensory gloves for sign language pattern recognition: an update from 2017 to 2022. IEEE Access 2022;10:123358-123377. [doi: 10.1109/ACCESS.2022.3219430]
60. Kožuh I, Hintermair M, Holzinger A, Volčič Z, Debevc M. Enhancing universal access: deaf and hard of hearing people on social networking sites. Univ Access Inf Soc 2015 Nov;14(4):537-545. [doi: 10.1007/s10209-014-0354-3]
61. Boggaram A, Boggaram A, Sharma A, Srinivasa Ramanujan A, Bharathi R. Sign language translation systems. Int J Softw Sci Comput Intell 2022 Jan;14(1):1-33. [doi: 10.4018/IJSSCI.311448]
62. Waghmare PP. Techniques for generating sign language a comprehensive review. J Inst Eng India Ser B 2024 Dec;105(6):1789-1803. [doi: 10.1007/s40031-024-01118-8]
63. Rowley K, Cormier K. Accent or not? Language attitudes towards regional variation in British Sign Language. Appl Linguist Rev 2023 Jul 26;14(4):919-943. [doi: 10.1515/applirev-2020-0144]
64. Preece J, Rogers Y, Sharp H. Interaction Design: Beyond Human-Computer Interaction, 6th edition: John Wiley & Sons; 2023.

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
CNN: convolutional neural network
HMM: hidden Markov model
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SLR: sign language recognition
YOLO: You Only Look Once

Edited by Y Li; submitted 22.Dec.2024; peer-reviewed by A Holzinger, S Singh, Z Yang; revised version received 03.Aug.2025; accepted 05.Aug.2025; published 09.Apr.2026.

Please cite as:
Marcolino MS, Oliveira LFRD, Valle LR, Rosa LMMS, Sanches GTDS, Santos NS, Costa MRD, Bernardino ELA, Cordeiro RAA, Prates RO, Reis ZSN, Campos MFM
Innovations in Deaf Health Care Communication: Systematic Review of Sign Language Recognition Systems
J Med Internet Res 2026;28:e70417
URL: https://www.jmir.org/2026/1/e70417
doi:10.2196/70417

© Milena Soriano Marcolino, Lucca Fagundes Ramos de Oliveira, Lucas Rocha Valle, Luiza Marinho Motta Santa Rosa, Gabriela Teodora de Souza Sanches, Natalia Sales Santos, Michelle Ralil da Costa, Elidea Lucia Almeida Bernardino, Raniere Alislan Almeida Cordeiro, Raquel Oliveira Prates, Zilma Silveira Nogueira Reis, Mario Fernando Montenegro Campos. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Using Group Chats to Drive Behavior Change in Digital Health Interventions: Scoping Review and Realist Synthesis

Hein Thu[1], MPH; Felicia Jia Hui Chan[2], MSc; Jumana Hashim[3], PhD; Shahmir H Ali[2], PhD

[1]Burma Academy, Yangon, Myanmar

[2]Saw Swee Hock School of Public Health, National University of Singapore, Tahir Foundation Building, 12 Science Drive 2, #10-01, Singapore, Singapore

[3]Behavioural and Implementation Science Interventions, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

Corresponding Author:
Shahmir H Ali, PhD
Saw Swee Hock School of Public Health, National University of Singapore, Tahir Foundation Building, 12 Science Drive 2, #10-01, Singapore, Singapore

Abstract

Background: Group chats on platforms such as WhatsApp (Meta Platforms, Inc), WeChat (Tencent Holdings Limited), and Telegram (Telegram FZ-LLC) are central to everyday communication in many settings, including across low- and middle-income countries and among groups often overlooked by one-to-one or app-based digital health tools. Yet their roles and underlying mechanisms as intentionally designed health interventions have not been comprehensively examined.

Objective: This scoping review with realist synthesis aimed to map the characteristics and reported outcomes of group chat–based health interventions and contextual conditions shaping their outcomes. It mapped intervention characteristics and identified context-mechanism-outcome (CMO) configurations driving behavior change.

Methods: We included empirical studies in which group chats were the primary digital component of the health intervention. Searches of PubMed, Embase, MEDLINE, Web of Science, and Scopus (2005‑2026) and reference screening identified eligible studies. Data were extracted on study characteristics, health domains, settings, participants, platform, intervention features, and outcome domains. For the realist synthesis, CMO configurations were analyzed using reflexive thematic analysis informed by social cognitive theory and synthesized into a final program theory.

Results: Eighty-one studies were included. Publications increased sharply after 2020, with evidence clustering in mental health (25/81, 30.9%) and maternal and child health (20/81, 24.7%). Most studies enrolled adult participants, including pregnant or postpartum women, caregivers, and people managing chronic conditions. WhatsApp (38/81, 46.9%) and WeChat (21/81, 25.9%) were the most frequently studied platforms. Interventions were commonly delivered in clinical or community settings and typically involved mixed membership groups that included health professionals alongside participants (68/81, 84.0%). The realist synthesis identified 12 recurring CMO configurations across 5 domains: capability and actionability, confidence and motivation, modeling and norms, safe and supportive environment and access, and self-regulation and maintenance. These mechanisms were activated or suppressed by facilitation quality, group composition, cultural alignment, technological access, and social norms. The final program theory depicts group chats as dynamic systems where access, facilitation, and group structure shape how the domains amplify or dampen one another, explaining shifts from low engagement to high trust spaces sustaining behavior change.

Conclusions: This review integrates descriptive mapping of interventions with realist program theory to explain how context and facilitation activate mechanisms of change, unlike prior reviews that have focused narrowly on effectiveness of digital modalities. This dual approach provides a practical, mechanism-focused basis for designing and scaling group chat interventions in practice, particularly given their potential as a low-cost, high-reach strategy embeddable within broader digital health programs. Realizing this potential depends on treating group chats as purposefully designed social environments, with deliberate attention to facilitation quality, equity-oriented access, and alignment between group structure, norms, and communication practices.

Trial Registration: PROSPERO CRD420251082787; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251082787

(J Med Internet Res 2026;28:e88911) doi:10.2196/88911

KEYWORDS

group chats; health behavior; social cognitive theory; realist synthesis; scoping review

XSL•FO
RenderX

Introduction

The global spread of smartphones, mobile internet, and messaging apps has normalized digital communication and reshaped health promotion and care delivery [1]. Digital tools now support prevention, screening, chronic care, and adherence at scale and low cost [2-4]. Social media expanded interactive engagement, and 5.5 billion people were online in 2024, about 68% of the global population [3,4]. Many reviews still pool disparate modalities such as apps, SMS, and social media under 1 digital health label, which obscures modality-specific mechanisms and user dynamics [5]. Relatedly, artificial intelligence (AI)–enabled conversational agents (chatbots) have also emerged as a rapidly growing approach in digital behavioral interventions, with recent syntheses reporting generally modest, heterogeneous effects across outcomes and highlighting ongoing considerations around safety, privacy, and governance in health contexts [6].

Given this expanding and heterogeneous landscape, a clearer emphasis on mechanisms is needed. Behavior change theories are often invoked to guide digital design and interpretation, for example, observational learning, social reinforcement, and self-efficacy, but many studies reference theory briefly and rarely test mechanisms within digital environments [7-11]. Realist synthesis has grown as a way to explain what works, for whom, in what contexts, and how, advancing context-mechanism-outcome (CMO) explanations for digital interventions [12-14], yet most syntheses remain modality agnostic (ie, do not disaggregate by technology type, such as apps vs SMS vs group chats), focus primarily on evidence from more complex, app-based interventions (which may be less translatable in low- and middle-income country (LMIC) settings where simpler messaging tools predominate), or lack intensive exploration of digital technologies that engage more complex forms of behavior change (ie, those involving social and community processes).

Social media messaging has now become an integral part of the digital health landscape of many countries, particularly LMICs [15], where platforms such as WhatsApp (Meta Platforms, Inc), Telegram (Telegram FZ-LLC), Viber (Rakuten, Inc), Signal (Signal Technology Foundation), and WeChat (Tencent Holdings Limited) are widely used by patients and providers. More than 85% of physicians and nurses own smartphones or tablets, and 60% to 80% of clinical staff use text messaging for patient communication [16]. In LMIC settings, messaging apps have supported health education and COVID-19 response, and about 80% of households have mobile phone access [17,18]. At the same time, much recent innovation in digital behavioral interventions has focused on one-to-one formats, including conversational agents and AI-enabled chatbots that deliver tailored coaching through dyadic dialogue [19]. However, group chats are a qualitatively different intervention environment, as they are multiparty, relationship-embedded spaces where norms, privacy, and trust are coproduced in real time, so design lessons from one-to-one chatbot models may not translate directly.

Within this ecosystem, group chats on WhatsApp, Telegram, WeChat, LINE (LINE Corporation), and Signal are multiuser threads that enable real-time, small-group, conversational interaction. This format differs from one-to-one messaging (the dominant configuration for most chatbot interventions), which is dyadic and linear, and from broadcast social media, which is asymmetric and feed-based. Participants exchange messages, images, and voice notes within semiprivate family, peer, or community networks, which affords familiarity, immediacy, and intimacy [20,21]. In many LMIC settings, group chats (particularly on platforms such as WhatsApp) are a particularly important part of the digital health environment [22]. Importantly, unlike other types of digital platforms, they often remain accessible to older adults and lower literacy users because they are low bandwidth and simple to use [20,23]. Compared with app-centric self-tracking or SMS reminders, group chats foreground interaction over individual logging and mirror offline support structures such as trust, reciprocity, and collective efficacy.

Despite widespread everyday use, research on group chats has largely been observational [24,25], often focusing on understanding social support dynamics or how health information or misinformation spreads in these spaces. What is missing is a consolidated synthesis of group chats as a primary intervention platform. Evidence remains fragmented on where, how, and for whom group chats are used intentionally to deliver health content, coaching, or peer support, and on how facilitation, privacy, culture, and gender norms shape participation. This gap matters because group chats can engage populations that are often excluded by data-heavy or app-centric models and can better reflect the interpersonal and community networks that drive health behavior change. As noted in prior reviews, this is particularly important in LMIC contexts, where group chats embedded within familiar, everyday social messaging platforms may be more feasible and acceptable than standalone health apps, yet empirical evidence on their implementation and the mechanisms through which they operate remains limited and fragmented [26-28]. A focused synthesis is therefore needed to understand which design and delivery elements are transferable across diverse settings or require adaptation depending on contexts [29].

To address these gaps, this scoping review with realist synthesis aims to map and synthesize the existing literature on group chat–based health interventions and to explain how these interventions work in practice. Specifically, the scoping review describes the characteristics of group chat interventions across studies, including platform type, key design features, target populations, health domains, and geographic distribution, and summarizes the outcomes reported across behavioral, psychosocial, and clinical domains. The realist synthesis then explores the mechanisms and contextual conditions that shape engagement and behavior change, including facilitation style, group composition, trust, cultural norms, and platform features. Together, these approaches identify where evidence is concentrated, highlight key gaps, and offer an explanatory framework to inform the design and adaptation of future group chat interventions.

XSL•FO
RenderX

Methods

Study Design and Reporting Frameworks

This review combined a scoping review with a realist synthesis to generate both descriptive and explanatory insights into group chat–based health interventions. The scoping review mapped the scope, characteristics, and distribution of the evidence base, including intervention design features, platforms, target populations, health domains, and settings. It also summarized the range of outcomes reported across studies, including behavioral, psychosocial, and clinical outcomes, to clarify what has been evaluated and where evidence remains limited. The realist synthesis complemented this by examining how and why group chat interventions achieved outcomes, drawing on qualitative, mixed methods, and process data to develop and refine program theories. The scoping review component was reported in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [30]; the completed reporting checklist is provided in Checklist 1. The search was reported in accordance with PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) [31]; the full search checklist and search strings are provided in Checklist 2. The realist synthesis was reported in accordance with RAMESES II (Realist And Meta-narrative Evidence Syntheses: Evolving Standards II) standards [32]. The review was prospectively registered in PROSPERO (CRD420251082787) as a systematic review, but it was conducted as a scoping review to better reflect the mapping aims and breadth of the evidence. Quality appraisal was also planned initially using Cochrane RoB-1 and ROBINS-I. However, no formal critical appraisal or risk-of-bias assessment was conducted, as the scoping review aimed to map and characterize the evidence base rather than exclude studies based on study quality.

Development of Initial Program Theories

The Initial Program Theories (IPTs) outlined hypothesized pathways through which participation in group chats might have influenced health behavior. These IPTs were developed prior to data extraction and served as a starting framework for realist analysis, subsequently refined as evidence was synthesized. IPT development began with a scoping review of literature on digital health, peer support, and messaging-based interventions to identify recurring contexts (eg, facilitation, message cadence, and trust) and mechanisms (eg, modeling, reinforcement, accountability, and belonging). These insights were organized using social cognitive theory (SCT) [33], which was chosen as the guiding framework because it is an established and widely applied model for understanding mechanisms of behavior change in digital health interventions and for informing their design [34]. SCT links social interaction, observational learning, and reinforcement (core processes in group chat environments) to behavior change, providing a strong theoretical basis for explaining how digital communication can shape motivation, confidence, and action. Within this framework, constructs such as self-efficacy, observational learning, behavioral capability, reinforcement, environment, and self-regulation were mapped to the interpersonal and contextual dynamics of group chats. Provisional CMO propositions were then articulated to capture how design features and social processes might generate change; for example, how active facilitation could create accountability, or how peer similarity and trust might foster belonging and sustained engagement.

The IPTs were refined through iterative team discussions involving researchers with expertise in behavioral science, digital health, and qualitative methods. Refinement emphasized theoretical coherence and cross-context relevance, particularly for LMIC settings where group chats are often the most accessible digital channel. Contextual barriers and facilitators were incorporated into a matrix accompanying the IPTs to support later analysis. The final framework, summarized in Multimedia Appendices 1 and 2, identified 6 SCT-informed pathways linking group chat mechanisms to outcomes: behavioral capability, self-efficacy, environmental affordances, observational learning, reinforcement, and self-regulation.

Information Sources and Search Strategy

A comprehensive search of the peer-reviewed literature was conducted to identify empirical studies evaluating the use of group chat–based digital platforms as part of a health intervention. Initially, on September 19, 2025, searches were carried out in PubMed (via PubMed), Embase, MEDLINE (via Ovid), Web of Science (via Web of Science), and Scopus (via Scopus) for records published between January 1, 2005, and September 1, 2025, limited to English-language and human studies. Embase and MEDLINE were searched simultaneously on the Ovid platform (Wolters Kluwer) using a single search strategy. The search was rerun on February 18, 2026, to capture studies published from September 2, 2025, onward, using the same search concepts and eligibility criteria. Search strings combined controlled vocabulary and free-text terms related to group chats (eg, “group chat,” “WhatsApp group,” “Telegram group,” and “WeChat”) with health- and intervention-related terms (eg, “health promotion,” “behavior change,” “intervention,” “trial,” and “evaluation”). Filters excluded reviews, protocols, commentaries, and other nonempirical papers, and no published search filters were used. No additional restrictions were applied beyond the date range, English language, and human studies. The search strategy was not peer-reviewed and was not adapted from prior reviews. The complete search strategies for all databases, copied exactly as run (including the February 18, 2026 update), are provided in Checklist 2.

No study registries were searched, no targeted website browsing was undertaken, and no authors or experts were contacted to identify additional studies. Reference lists of included studies and relevant reviews were screened to identify additional publications, including linked articles reporting process evaluations or qualitative findings from the same intervention.

Eligibility Criteria

The studies were eligible if they evaluated a health-related intervention in which a group chat served as the core digital space for intervention delivery and interaction. Group chat was defined as a multiuser messaging thread within a messaging platform that enables interaction among participants, with or

XSL•FO
RenderX

without facilitation, via text and optionally images, audio, or video. Eligible interventions used the group chat to deliver one or more core functions such as health education, counseling or coaching, peer support, facilitated discussion, collaborative problem-solving, or care coordination, rather than logistics alone. Eligible study designs included randomized controlled trials (RCTs), quasi-experimental studies, pilot and feasibility studies, and program evaluations reporting quantitative, qualitative, or mixed methods findings. Included studies were mapped according to outcomes reported across behavioral, psychosocial, and clinical domains, as well as engagement and implementation outcomes where available. Studies reporting qualitative or process data linked to an intervention were retained where they contributed explanatory information to the realist synthesis.

Studies were excluded if they lacked a defined human population, did not implement a health intervention, or used group chats only peripherally. Peripheral use was defined as group chats used primarily for recruitment, administrative coordination, appointment scheduling, or 1-way announcements without meaningful interactive discussion or delivery of intervention content. Publications without primary data, including protocols, conceptual papers, reviews, and conference abstracts, were excluded. Only full-text, peer-reviewed journal articles were included.

Study Selection

Records from all sources were imported into Covidence (Veritas Health Innovation), where duplicates were removed using the platform's deduplication function with manual review to confirm and resolve remaining duplicates. To ensure consistent application of eligibility criteria, reviewers jointly piloted a small sample of abstracts prior to screening. Title and abstract screening and full-text screening were conducted independently by 2 reviewers. Disagreements were resolved through discussion, with a third reviewer consulted when consensus could not be reached. The study selection process is illustrated in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

Data Charting

Data charting was conducted using structured, piloted forms designed to extract and organize information for both the mapping of study characteristics and the realist synthesis components of the review. Data were extracted and entered into these forms by 1 reviewer and independently checked by a second reviewer, with discrepancies resolved through discussion and arbitration by a third reviewer when needed.

For the scoping review, charted items included bibliographic details (author and year), country or setting, health domain, study aims, study design, participant characteristics and sample size, platform type, intervention overview, and how the group chat functioned within the intervention. Where reported, we also charted key group chat features (eg, group structure, communication modalities, facilitation or moderation practices, and any stated privacy or participation rules), comparator conditions where applicable, outcome domains assessed, outcome measurement approaches and timing, and study-reported findings.

For the realist synthesis, reviewers extracted verbatim or summarized data describing contextual conditions (eg, cultural setting, group composition, and level of facilitation), mechanisms proposed or observed to drive change, and outcomes achieved. When mechanisms were not explicitly stated, they were inferred from the authors' descriptions of processes or participant experiences. CMO coding was undertaken by 2 reviewers and refined through team discussion.

Analysis and Synthesis

For the scoping review, extracted data were summarized descriptively to map study characteristics, intervention features, and outcome domains across the included studies. Studies were grouped by health domain, population, setting, and platform type, and outcomes were summarized across behavioral, psychosocial, and clinical domains, including how and when they were measured where reported. This descriptive synthesis was used to identify areas where evidence is more developed and where gaps remain.

For the realist synthesis, analysis followed a realist logic of inquiry to develop and refine program theories explaining how, why, and under what conditions group chat interventions achieved outcomes. The process began with all extracted CMO data, each describing how contextual features interacted with intervention processes to generate change. Using a reflexive approach to thematic analysis [35], the team examined CMOs to identify recurring patterns, interconnections, and configurations across studies. CMOs were coded and grouped into higher-order clusters representing common or complementary configurations observed across settings. As patterns stabilized, these clusters were organized into broader domains and compared against the IPTs. The developing domains were informed by, but not limited to, the original IPT categories. While SCT provided the conceptual scaffold, categorizations evolved where new mechanisms or linkages emerged, ensuring the final framework reflected both theoretical grounding and empirical diversity. The framework was iteratively refined through team discussion and consensus-building, with regular meetings to test interpretations, examine alternative explanations, and address negative cases. A detailed audit trail documented each iteration of coding, synthesis, and decision-making.

The final framework was distilled into a visual program theory that offered an evidence-based account of how group chat interventions operate and a heuristic for informing their design and adaptation. Relationship candidates were identified through cross-case comparison of CMO clusters and mapped only when supported by consistent empirical patterns or theoretically coherent linkages. The team iteratively refined these pathways through structured deliberation, examining rival explanations and negative cases to determine whether connections were salient enough to visualize. Directionality was assigned where temporal or causal ordering was evident in the data, while bidirectional relationships were included when reciprocal influence was repeatedly described.

XSL•FO
RenderX

Results

Included Studies

A total of 4130 records were identified through database searches (PubMed=1326; Scopus=1296; Embase=764; Web of Science=744) and manual citation screening (n=29). After removing 1467 duplicates, 2692 titles and abstracts were screened, of which 2542 were excluded. Full-text assessment was conducted for 150 articles, of which 69 were excluded because they did not meet the inclusion criteria, most commonly because group chat was not central to the intervention (n=41), no group chat component was present (n=10), full text was unavailable (n=8), or outcomes were not health related (n=1), with additional exclusions. A total of 81 studies [36-116] were included for the scoping review and realist synthesis (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram for studies included in the scoping review and realist synthesis of group chat interventions.

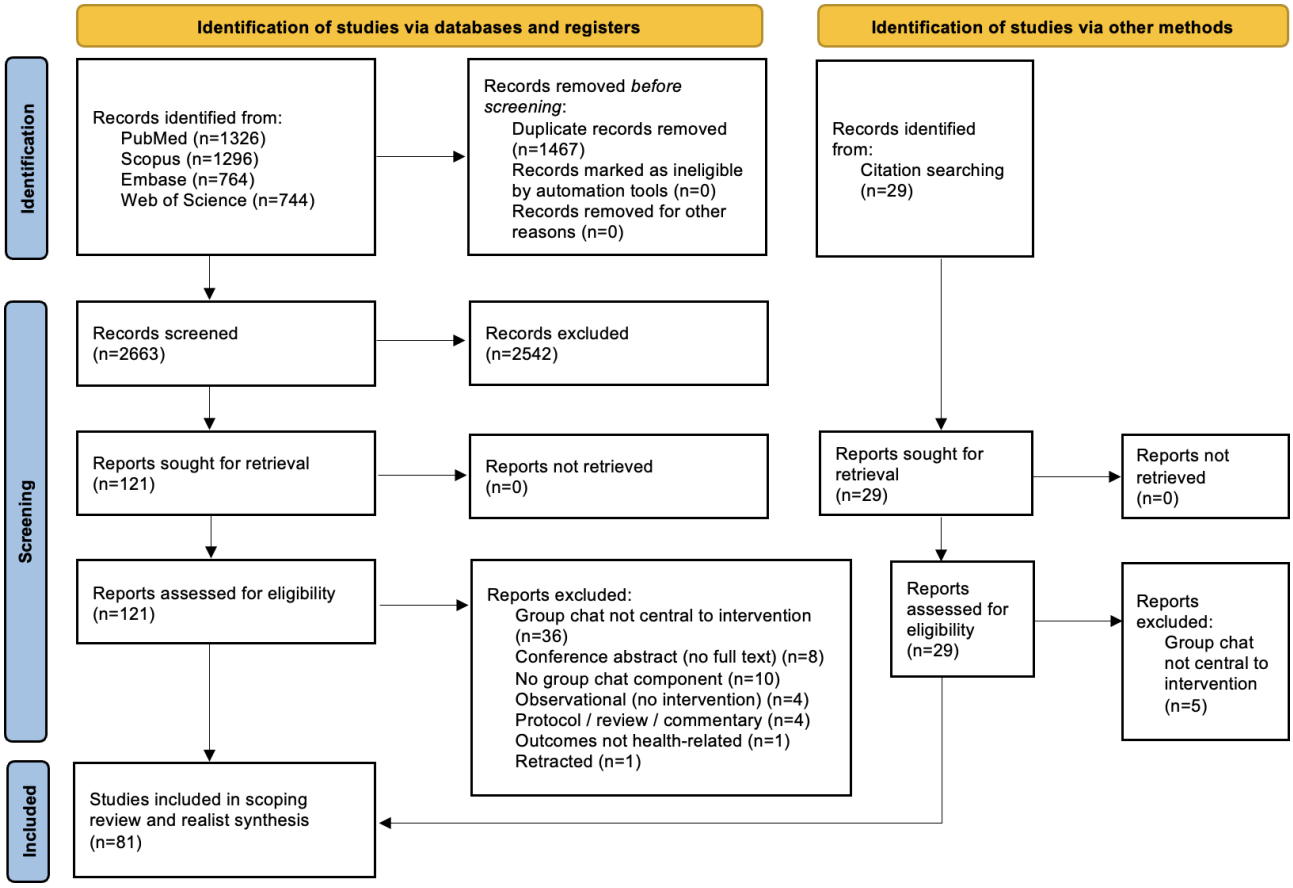

Study Characteristics

Table 1 summarizes the breadth of study designs, settings, and intervention features across the 81 included studies [36-116]. Designs spanned RCTs (30/81, 37%), quasi-experimental or single-group evaluations (15/81, 18.5%), and non-RCTs (11/81, 13.6%), alongside pilot and feasibility studies (10/81, 12.3%) and smaller numbers of qualitative (6/81, 7.4%), mixed methods (5/81, 6.2%), and program evaluations (4/81, 4.9%). Geographically, studies were distributed across 23 countries, with the largest single-country contributions from China (18/81, 22.2%) and the United States (10/81, 12.3%), followed by Indonesia (7/81, 8.6%), Hong Kong SAR and Iran (5/81 each, 6.2%), and Japan (4/81, 4.9%). Four studies were multicountry in design [41,61,78,86], and the studies spanned Sub-Saharan Africa, Latin America, the Middle East, Europe, and other parts of Asia (25/81, 30.9%).

XSL•FO
RenderX

Table . Summary of study characteristics.

Characteristics	Studies included (N=81)
Study design, n (%)	
Randomized controlled trial	30 (37.0)
Single-group or quasi-experimental study	15 (18.5)
Nonrandomized controlled trial	11 (13.6)
Pilot and feasibility study	10 (12.3)
Qualitative descriptive study	6 (7.4)
Mixed methods evaluation	5 (6.2)
Retrospective observational study	2 (2.5)
Real-world program evaluation	2 (2.5)
Country, n (%)	
China	18 (22.2)
United States	10 (12.3)
Indonesia	7 (8.6)
Hong Kong SAR	5 (6.2)
Iran	5 (6.2)
Japan	4 (4.9)
Multicountry	4 (4.9)
Germany	3 (3.7)
Other[a]	25 (30.9)
Health topic, n (%)	
Mental health	25 (30.9)
Maternal and child health	20 (24.7)
Weight management, nutrition, or eating disorders	13 (16.0)
Smoking cessation	11 (13.6)
Dementia and caregiving	9 (11.1)
Diabetes	8 (9.9)
Physical activity	7 (8.6)
Others[b]	37 (45.7)
Setting[c], n (%)	
Clinical	39 (48.1)
Community or home	34 (42.0)
Workplace or school	8 (9.9)
Type of membership, n (%)	
Mixed membership	68 (84.0)
Homogeneous peer-led	13 (16.0)
Communication modalities[d], n (%)	
Text+dynamic media (audios, videos, etc)	41 (50.6)
Text only	20 (24.7)
Text+static media (images, emojis, links, etc)	15 (18.5)
Not explicitly stated	5 (6.2)
SCT[e] constructs aligned, n (%)	

XSL•FO
RenderX

Characteristics	Studies included (N=81)
Behavioral capability	61 (75.3)
Self-efficacy	63 (77.8)
Environment	69 (85.2)
Observational learning	58 (71.6)
Reinforcements	70 (86.4)
Self-control	22 (27.2)
Theories explicitly connected, n (%)	
Cognitive behavioral therapy	4 (4.9)
Social cognitive theory	7 (8.6)
Health Belief Model	3 (3.7)
Other[f]	21 (25.9)
Importance of group chat as intervention[g], n (%)	
Core intervention	42 (51.9)
Standalone intervention	13 (16.0)
Reinforcement intervention	26 (32.1)

[a]The other countries include Kenya, Turkey, Cameroon, Uganda, Brazil, South Africa, Ghana, Nigeria, Tanzania, United Kingdom, Italy, Canada, South Korea, Saudi Arabia, Netherlands, and Philippines.

[b]The other clinical topics include cancer and oncology, chronic disease self-management, parenting and child development, immunization, sexual and reproductive health, respiratory diseases, substance use, and rehabilitation.

[c]For setting, clinical interventions or interventions between health professionals were chosen as “clinical,” public health interventions which happened at the community level were chosen as “community/home” and workplace or school-based interventions were chosen as “workplace/school.”

[d]The chat platforms included WhatsApp, WeChat, Line, Viber, KakaoTalk, and Telegram.

[e]SCT: social cognitive theory.

[f]The other theories explicitly connected include Social Learning Theory, Parenting for Lifelong Health framework, Haddon Matrix, locus of control theory, Big Five Theory, MOST (Multiphase Optimization Strategy Framework) framework, Theory of Reasoned Action, Behavior Change Techniques (BCTs) taxonomy, Theory of Gender and Power, Information-Motivation-Behavior (IMB) Model, Behavior Change Wheel (BCW), Socioecological model, Theoretical framework of Acceptability, Psychoanalytic/Small Group theory, Wagner’s Chronic Illness Care (CIC) Model, Information-Motivation-Behavioral Skills (IMB) theory, Inquiry-Based Learning (IBL), Reinforcement theory, Capability, Opportunity, Motivation–Behavior (COM-B) Theory, Planned Behavior Theory, Self-Regulation Theory, Self-Discrepancy Theory, and Peers for Progress framework.

[g]If the group chat is: major part of the whole intervention for behavioral change, it is rated as “core”; the one and only behavioral intervention, it is rated as “standalone”; and as reinforcement or supplementary to the other behavioral interventions, it is rated as “reinforcement.”

The health topics addressed were diverse and often overlapping across studies. Mental health was the most frequently targeted domain (25/81, 30.9%), encompassing interventions addressing depression, anxiety, caregiving burden, and related psychosocial conditions. Maternal and child health was the second most common focus (20/81, 24.7%), including antenatal and postnatal care, breastfeeding support, and child development programs. Other prominent topics included weight management, nutrition, or eating disorders (13/81, 16%), smoking cessation (11/81, 13.6%), dementia and caregiving (9/81, 11.1%), diabetes self-management (8/81, 9.9%), and physical activity (7/81, 8.6%). A heterogeneous range of topics was addressed (37/81, 45.7%), including HIV, cancer, cardiovascular rehabilitation, and chronic obstructive pulmonary disease self-management. As studies frequently addressed more than 1 health domain, these proportions sum to more than 100%.

Nearly half of the studies were implemented in clinical settings (39/81, 48.1%), with the remainder primarily in community or home settings (34/81, 42%) and a smaller subset in workplace or school settings (8/81, 9.9%). Most interventions used mixed membership group formats comprising health professionals alongside participants (68/81, 84%) rather than homogeneous peer-led groups (13/81, 16%). Communication modalities were most often text combined with dynamic media such as audio and video (41/81, 50.6%), followed by text-only (20/81, 24.7%) and text combined with static media such as images, emojis, and links (15/81, 18.5%). Group chats most often served as the core intervention component (42/81, 51.9%), followed by use as a reinforcement or adjunct within broader multicomponent programs (26/81, 32.1%), and as standalone interventions (13/81, 16.0%). Of the 81 studies [36-116], theoretical frameworks connected to interventions included SCT (n=7, 8.6%), cognitive behavioral therapy (n=4, 4.9%), and the Health Belief Model (n=3, 3.7%). SCT constructs were commonly identifiable across interventions, with reinforcements (70/81, 86.4%) and environment (69/81, 85.2%) most frequently operationalized, followed by self-efficacy (63/81, 77.8%), behavioral capability (61/81, 75.3%), and observational learning (58/81, 71.6%). These descriptive patterns supported later theory-informed synthesis rather than serving as causal claims within the mapping stage.

XSL•FO
RenderX

The publication timeline indicates a small number of early studies from 2007 onward, followed by accelerated growth in the early 2020s (Figure 2). More than three-quarters of the evidence base (63/81, 77.8%) was published between 2020 and 2025, with the cumulative total increasing sharply in recent years. Prior to 2020, annual publication counts remained low, with no more than 3 studies published in any single year. This figure is intended to situate the field's maturity and pace of expansion rather than to infer drivers of growth.

Figure 2. Trends of the included publications over time (N=81).

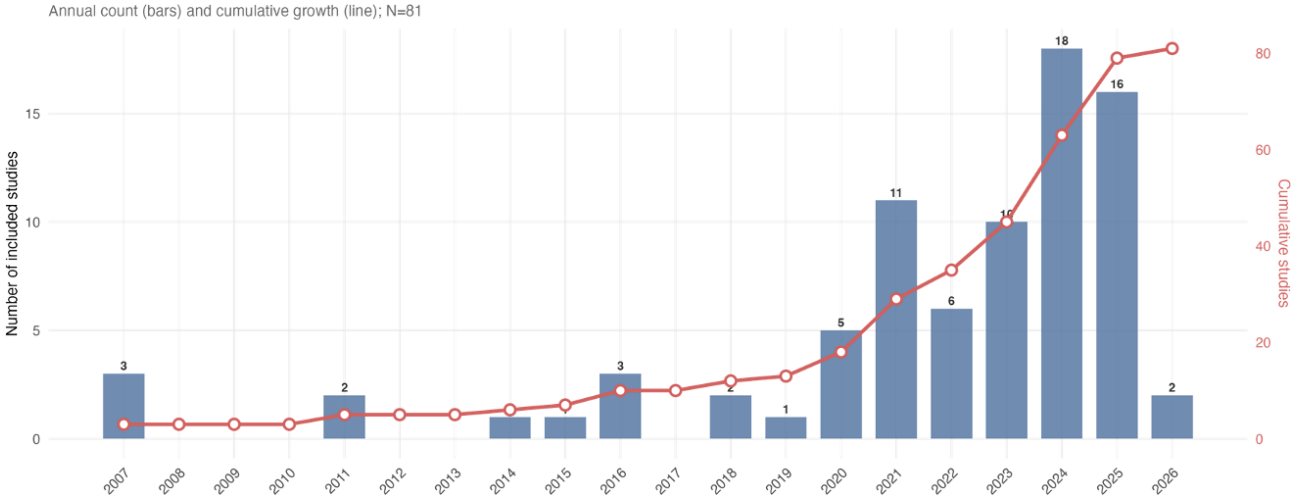

Distribution of Platforms, Settings, and Evidence Across Health Topics

Platform use varied by context (Figure 3). WhatsApp was the most frequently studied platform overall (38/81, 46.9%), with substantial representation in community settings (n=26) and additional use in clinical settings (n=11). WeChat was also commonly evaluated (21/81, 25.9%), largely in clinical contexts (n=15), reflecting its integration within China's health care system. Customized or other platforms, including LINE, Viber, KakaoTalk (Kakao Corporation), and Telegram, were used across both clinical and community settings (19/81, 23.5%). This mapping is presented to describe delivery channels across contexts and to support later interpretation of how platform affordances may shape implementation conditions.

Figure 3. Usage of group chat platforms by study settings (N=81). The other chat platforms include Line, Viber, KakaoTalk, and Telegram.

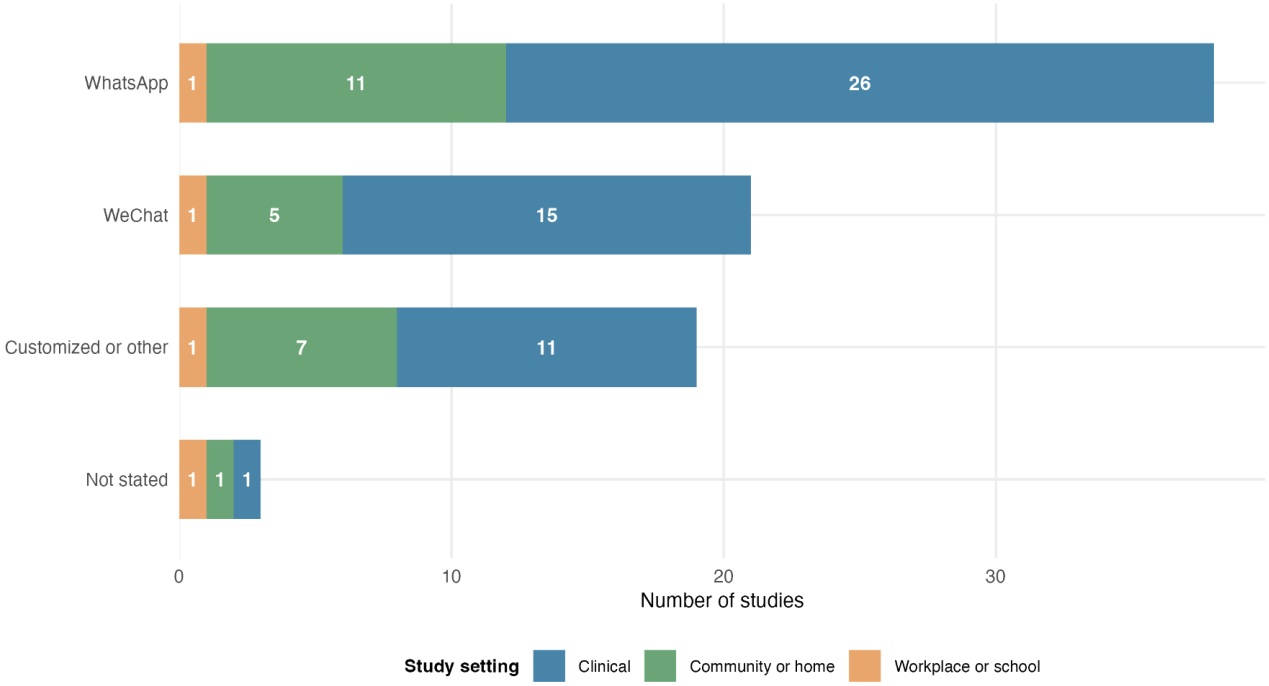

To further characterize the distribution of evidence across health topics and outcome domains, Figure 4 presents an evidence gap map of all 81 included studies [36-116]. Engagement and behavioral outcomes were the most consistently studied across nearly all health topics, while psychosocial outcomes remained notably sparse, particularly for smoking cessation, weight management, and dementia and caregiving, indicating areas where the evidence base requires further development.

XSL•FO
RenderX

Figure 4. Evidence gap map of group chat interventions based on outcome domains and clinical topics (N=81). Others in the study design include mixed methods evaluation, retrospective observational study, and real-world program evaluation. Other in the clinical topics include cancer and oncology, chronic disease self-management, parenting and child development, immunization, respiratory diseases, substance use, and rehabilitation. RCT: randomized controlled trial.

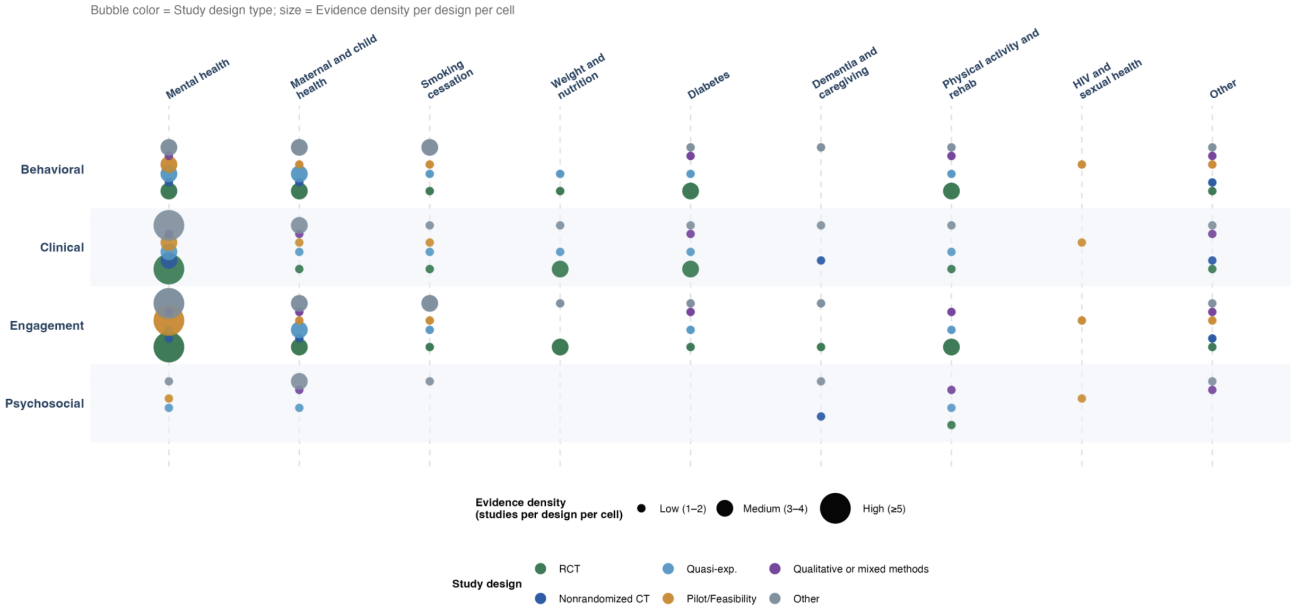

Final CMO Configurations

The analysis produced 12 CMO configurations, reflecting the most recurrent and explanatory patterns across the included studies. These configurations describe how specific contextual conditions triggered mechanisms within group chat–based interventions, shaping behavioral, psychosocial, and engagement outcomes (Multimedia Appendix 3) [36,37,39,40,42-47,49,54,58,59,61-63,65-68,71,72,74,75,77,79,81,82,84-86,88,89,94,99,99-106,108-116].

Domain 1: Capability and Actionability

Group chats enhanced participants' capacity to adopt health behaviors through several mechanisms. Timed prompts broke down complex actions into manageable steps, aiding timely task completion; for example, a WhatsApp program in Semarang used frequent prompts and multimedia guidance to support early breastfeeding, iron use, hygiene, and danger-sign monitoring [100]. Multimedia content (eg, videos and audio clips) paired with threaded Q&As helped participants develop practical skills by allowing them to view demonstrations, ask questions, and revisit materials. Chats also corrected misinformation in real time, reducing cognitive dissonance, such as in a Hong Kong WhatsApp group where mothers clarified details about influenza vaccination and gained confidence in clinic navigation [71]. These elements collectively strengthened behavioral capability across varied health domains.

Domain 2: Confidence and Motivation

Sustaining participants' belief in their capacity to change relied on personalized encouragement, visible recognition, and synchronous interaction. Tailored moderator feedback helped participants reframe setbacks constructively, boosting confidence and activating peer accountability. For example, a WhatsApp-based cognitive behavioral therapy program for overweight women in Iran used daily therapist interactions to support cognitive reframing and behavior monitoring, improving BMI [36]. Group recognition through photo sharing, comments, and public acknowledgment amplified motivation by making progress visible, fostering social validation and sustained engagement. Some interventions added brief audio or video calls, enhancing text-based support with immediacy and nonverbal cues. For instance, weekly video-based group chats that used real-time audio and visible nonverbal cues created stronger social presence and cohesion, leading to higher engagement compared with text-only formats [98]. This synchronous interaction deepened interpersonal bonds, increasing accountability and mutual support.

Domain 3: Modeling and Norms

Observational learning worked through direct peer exchange and the emergence of informal knowledge brokers. When participants shared personal experiences and outcomes, others gained concrete, relatable examples of applying health behaviors in real-life contexts. This peer demonstration normalized both success and struggle. For example, shared contraceptive experiences in digitally facilitated postabortion groups reduced anxiety and boosted confidence in effective method adoption [85]. Social comparison helped participants identify feasible tactics and adjust expectations based on peers' experiences. In some groups, participants assumed bridging roles, especially when expert input was intermittent. These emergent peer leaders relayed information, validated insights, and sustained discussion, reducing the gap between laypeople and professionals. For example, in moderated parent groups, more knowledgeable members absorbed expert input and began clarifying information and guiding peers [53]. These informal knowledge brokers multiplied learning pathways and strengthened network cohesion, enhancing the sustainability of professional guidance.

Domain 4: Safe, Supportive Environment and Access

Group chat effectiveness hinged on structural and social characteristics that fostered authentic engagement. Small, closed

XSL•FO
RenderX

groups with privacy safeguards and clear participation rules created psychological safety, encouraging participants to disclose sensitive information. For instance, anonymous online chats for adolescents with depressive symptoms reduced judgment fears and deepened emotional sharing [54]. However, privacy alone was insufficient; supportive social norms that discouraged judgment and promoted mutual respect were also key. Cultural and linguistic alignment further shaped trust and relevance. Interventions using participants' primary language, culturally familiar content, and trusted messengers enhanced understanding and engagement. For example, a WeChat cessation program for Chinese immigrant smokers used tailored messages and coach support to improve quitting knowledge and self-efficacy [65]. Such cultural fit was especially crucial for low-literacy or marginalized communities often overlooked by standard health communication.

Domain 5: Self-Regulation and Maintenance

Group chats sustained behavior change by externalizing accountability and automating behavioral cues. Public goal setting with daily status updates made commitments visible, fostering social accountability and self-monitoring. Peer visibility created gentle pressure to follow through and modeled consistent behavior, as seen in small anonymous digital groups where daily abstinence updates and peer "applause" improved motivation and self-efficacy, leading to higher smoking cessation rates at 12 weeks [102]. Frequent peer messages reinforced shared goals through consistent social prompting. Time-based reminders complemented this by serving as environmental cues, reducing reliance on fluctuating motivation and helping establish habitual routines. A WeChat program for coronary artery disease patients used thrice-weekly messages for 12 months to reinforce self-management and improve diet, exercise, medication adherence, and blood pressure monitoring [67]. Despite these insights, evidence on the persistence of these self-regulation mechanisms after interventions ended remains limited, leaving long-term maintenance uncertain.

Consolidated Program Theory of Group Chat Interventions

Building on the CMO domains in Multimedia Appendix 3 [36,37,39,40,42-47,49-54,58,59,61-63,65-68,71,72,74,75,77,79,81,82,84-86,88-90,94-96,99-106,108-116], Figure 5 synthesizes how group chat interventions were observed to operate within layered and interacting contextual conditions. Rather than functioning within single, isolated enablers, systems, technology, settings, participants, and social factors often combined in ways that shaped who participated, how safely they engaged, and how reliably groups were maintained. Across studies, conditions that supported digital access, regular facilitation, and a sense of social familiarity or safety frequently co-occurred, while technological instability, limited staff presence, or socially constrained dynamics were commonly described as barriers.

Within these contextual environments, the mechanism patterns identified in the review tended to operate in combination. Capability-building processes (eg, prompts, guidance, and corrections) were often noted alongside elements that created a psychologically safe space for questions or disclosure. Peer modeling and emerging norms helped translate guidance into everyday routines, while motivational inputs and emotional reassurance supported continued effort. Self-regulation processes, such as routine check-ins, public goals, and habit cues, were typically described in studies where other mechanisms were already active. Across interventions, these interdependencies appeared to shape how participants learned, interacted, and sustained behavioral changes over time.

The outcome layer of the model draws on the scoping review and links these mechanisms to behavioral, psychosocial, and clinical outcomes. Improvements were reported in breastfeeding technique and caregiving practices, safety and hygiene behaviors, emotional well-being, and day-to-day disease self-management. Across studies, health behavior consistency and self-management routines emerged as an important foundation for other outcomes, frequently appearing alongside or prior to changes in clinical indicators such as weight, glycemic markers, blood pressure, or disease perceptions. Examples from the review illustrate these patterns: in clinic-based WhatsApp cessation groups in Hong Kong, recent quitters who engaged more consistently benefited from peer reinforcement and achieved higher abstinence rates [57], whereas in maternal health groups where husbands felt embarrassed and participation remained low, limited psychological safety and weak norms were noted as barriers to attitudinal change [100]. Overall, the program theory depicts group chat interventions as dynamic, socially embedded systems whose effectiveness depends on how well contextual conditions, mechanisms, and outcomes remain in alignment over time.

XSL•FO
RenderX

Figure 5. Visualized context-mechanism-outcome pathways from realist synthesis and scoping review describing how group chat interventions were observed to shape health-related outcomes. The figure synthesizes cross-study context-mechanism-outcome patterns into a final program theory. Contextual conditions (system, technology, setting, participant, and social factors) list facilitators (F) and barriers (B). Colored solid arrows represent the primary pathways underpinning the 5 major domains in the Final Program Theory (Multimedia Appendix 3) [36,37,39,40,42-47,49-54,58,59,61-63,65-69,71,72,74,75,77,79,81,82,84-86,88-90,94-96,99-105,106-115,116] and their most salient context-mechanism-outcome relationships. Dotted arrows indicate additional relationships within or across context, mechanism, or outcome domains. One-way arrows depict primarily unidirectional influences, while 2-way arrows reflect bidirectional or mutually reinforcing relationships. Outcome clusters summarize the main proximal and distal effects derived from the scoping review. Arrows visualize the most prominent patterns identified, while other secondary or more distal relationships may exist. The figure highlights those most consistently supported across included studies.

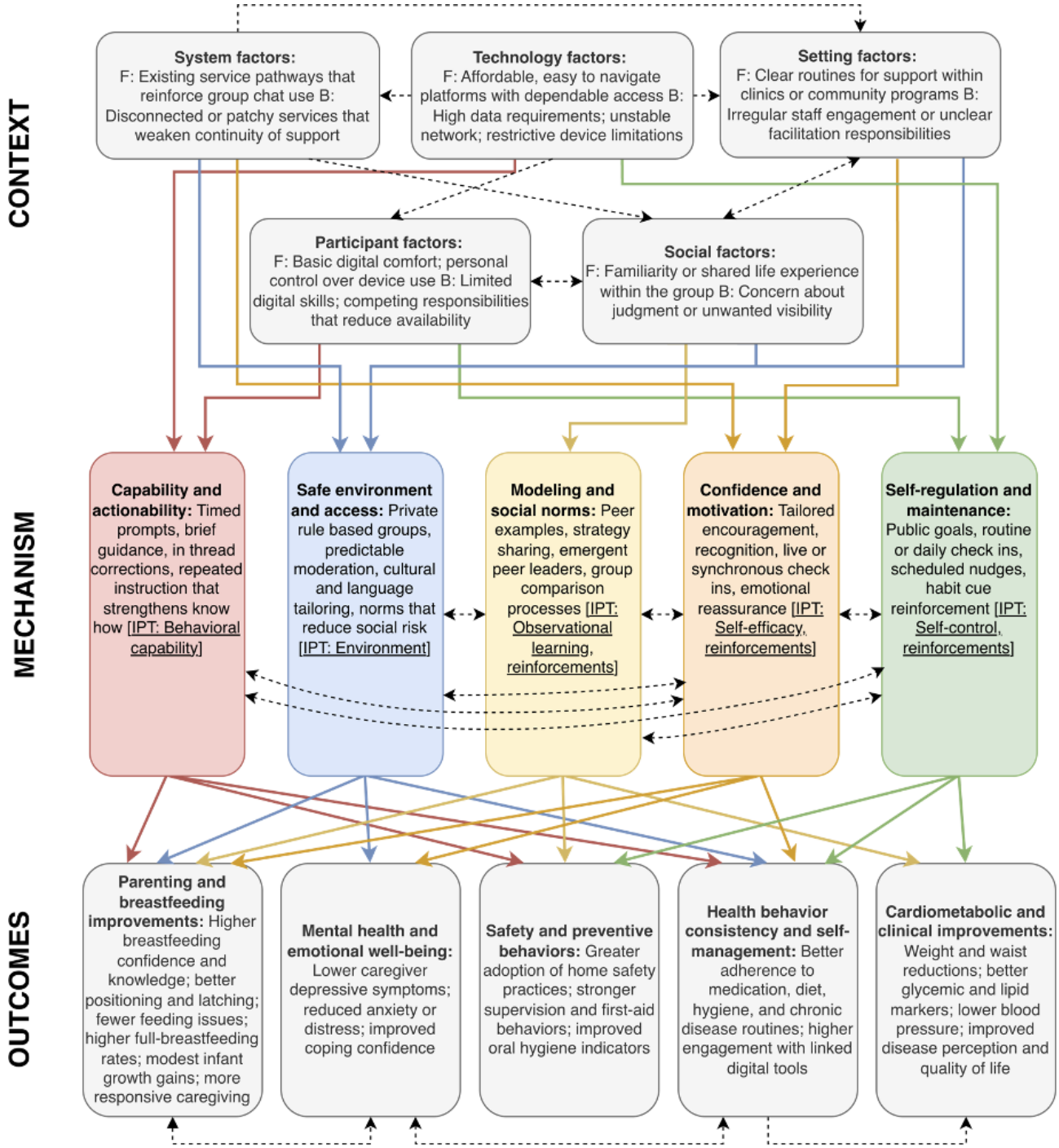

Discussion

Principal Findings

This scoping review with realist synthesis mapped and explained the evidence on group chat–based health interventions. In line with our objectives, the scoping review component described where evidence is concentrated and how interventions have been designed and evaluated, while the realist synthesis examined how contextual conditions shape engagement and behavior change. Findings highlight how the evidence base on group chat interventions has grown rapidly since 2020 and is concentrated most strongly in mental health and maternal and child health. WhatsApp and WeChat were the most commonly

XSL•FO
RenderX

used platforms across community and clinical settings, and most interventions relied on mixed membership groups that included health professionals alongside participants. The realist synthesis identified 12 recurring CMO configurations that clustered into 5 mechanism domains: capability and actionability, confidence and motivation, modeling and norms, safe supportive environments and access, and self-regulation and maintenance. Together, these findings offer a theory-informed account of how group chat interventions operate in practice and highlight key evidence gaps, including the underrepresentation of older adults and lower literacy users and the limited reporting of implementation processes.

In many digital health programs, group chats are positioned as an enabling social layer alongside other components (eg, coaching sessions, educational modules, wearables, or reminder systems), which may enhance support but also make it difficult to isolate whether observed effects are driven by peer processes, professional input, self-monitoring, or their interaction. We also excluded many studies where group chats were mentioned but not substantively implemented. In several cases, the feature was added with minimal detail on facilitation, engagement, or theoretical rationale, suggesting it was treated as a generic communication channel rather than a mechanism-based component. Taken together, these patterns indicate a broader missed opportunity: group chats are often positioned as supplementary add-ons rather than designed and evaluated as interventions with their own distinct social and behavioral mechanisms.

Previous reviews on digital health communications, including those in digital mental health, tend to emphasize adaptability, cost, and organizational policies rather than the mechanisms that drive behavior change [13]. At the same time, earlier realist reviews often pool together diverse digital health technologies [12,117], which makes it difficult to detect modality-specific dynamics. By focusing specifically on group chat interventions, our review is able to surface mechanisms such as peer modeling, emergent leadership, and collective accountability that are overlooked in broader syntheses. We further show that informal peer leaders and knowledge brokers naturally emerge within these groups [46,53,68,101], suggesting that long-term effectiveness may depend more on peer-to-peer support than on continuous professional involvement. This contrasts with Schlief et al [118] emphasis on therapeutic relationships and Shahid et al [119] focus on team-based professional care, highlighting how the peer-led potential of group chats remains underused. Moreover, the predominance of group chat interventions in LMICs, especially in Asia, was another notable finding. Rapid mobile uptake [120], collectivist social norms [121], and rising chronic disease burdens [122,123] create strong demand for low-cost, scalable tools that mirror everyday communication practices. These contexts also underscore the potential of group chats to reach users who may be excluded by app-centric or data-heavy models [124], including those with shared devices, limited data, or intermittent access. At the same time, a clear disconnect emerged: although group chats are widely used by populations often absent from more complex digital health tools (including older adults and lower literacy users), these groups were still largely missing from the intervention studies. This likely reflects recruitment patterns centered on maternal, youth, and clinic-based populations rather than the wider pool of everyday messaging app users [125,126]. A related concern is that recruitment may still be shaped by assumptions about which groups are easiest to reach digitally, even though recent participatory realist work cautions against assuming who can or cannot benefit without empirical evidence [127]. Broadening future studies to include these underrepresented groups and testing needed adaptations in facilitation, content, or onboarding will be essential for understanding how group chat interventions function across diverse ages, literacy levels, and access conditions [12].

The group chats in our realist synthesis were predominantly small, closed groups with preexisting relationships, intensive facilitation by providers or researchers, and a single, bounded communication space [36,44,46,54,58,61,69,72,74,81,89,100,101,103,114-116]. These design choices supported safety and disclosure but limited the multidirectional information flows across wider social networks that naturally occur in everyday messaging use [127]. By contrast, observational work on platforms such as WhatsApp, WeChat, and Facebook Messenger (Meta Platforms, Inc) has shown that group chats can rapidly diffuse information and misinformation during crises [128], coordinate mutual aid [129], and function as hubs for resource sharing and collective problem-solving [130]. Across intervention trials, effects also often appeared to be contingent on engagement intensity and facilitation quality, reinforcing that group chats often operate as social and motivational infrastructures rather than simple content-delivery mechanisms [131]. Our realist synthesis highlighted the centrality of peers in driving engagement and change within intervention groups [37,43,44,46,52,53,59,61,68,69,75,81,82,84,85,88-90,94,95,101-104,108-110,113-116], suggesting substantial untapped potential to design group chat interventions that position communities as active agents of change rather than passive recipients. Future work could more deliberately harness social diffusion, for example, by seeding core groups and explicitly supporting participants to adapt and carry content into their own family, neighborhood, faith-based, or workplace chats, while safeguarding norms of privacy and trust. Some emerging models already point in this direction. For instance, the “Let’s Chat” intervention trained young Vietnamese American adults as family health advocates who then initiated and led cancer screening discussions within their existing intergenerational family group chats, demonstrating how a relatively light-touch scaffold can catalyze preventive conversations across multiple related networks [46]. Designing and testing similar networked group chat architectures, including multigroup structures and clearly defined bridging roles, represents a key opportunity for innovation in digital health.

This review has several strengths, including being the first to focus specifically on group chats as a primary digital health delivery channel, integrating a scoping review of empirical studies with a realist synthesis grounded in an a priori SCT-informed program theory and conducted in line with PRISMA-ScR [30] and RAMESES II [32]. This methodological combination provides a unique contribution by mapping how group chat interventions have been designed and implemented

XSL•FO
RenderX

across settings while also explaining the mechanisms and contextual conditions through which they operate in practice. Our findings show that group messaging functions as a socially embedded environment in which capability building, modeling, reinforcement, motivation, and self-regulation unfold through everyday interaction. The realist approach was especially valuable for unpacking these processes, allowing us to identify recurring CMO patterns across diverse platforms and settings. These insights advance behavioral theory in digital health by showing how constructs such as observational learning, collective efficacy, and social reinforcement are activated within group chat ecosystems, and they offer clearer design principles for tailoring interventions to local cultures, facilitation capacities, and user needs.

Limitations

Several limitations should be noted. First, a formal critical appraisal across all included studies was not undertaken. Given the heterogeneity in study designs, populations, intervention formats, and outcome measures, applying a single quality assessment tool across the full evidence base was not feasible. Second, we limited inclusion to peer-reviewed, English-language full texts; this language restriction may have excluded relevant studies and may bias the apparent geographic distribution of evidence, particularly for platforms and settings where local language publication is common. We also excluded conference proceedings, protocols, and gray literature, which may have led to undercapture of emerging interventions and implementation learnings that are not yet published as full articles. Third, realist inference depends on the availability of contextual and process data, yet many primary studies, including RCT reports, provided limited detail on facilitation practices, engagement dynamics, and implementation adaptations. The frequent absence of rich process data constrained CMO specification in some cases and may have required cautious inference from brief descriptions, potentially leading to underspecification or misclassification of mechanisms and contexts.

Conclusion

Group chats are central to how communities communicate yet remain underused in formal digital health design. This review shows that group chats are not only delivery channels but social infrastructure that can mobilize collective action, reinforce norms, and support sustained behavior change in ways distinct from other digital modalities. By centering group chats as a distinct intervention format and integrating a scoping review with realist program theory, we explain how context and facilitation activate mechanisms of change, offering a practical roadmap for designing and scaling low-cost, high-reach interventions within broader digital initiatives. AI-enabled chatbots and facilitation tools may extend this potential through scalable tailoring, just-in-time prompts, multilingual clarification, and triage while reducing moderator burden [111], but also raise design and governance questions around trust, peer dynamics, misinformation and safety, and when human facilitation is essential for relationship-building and accountability [132]. Compared with prior reviews that pool digital modalities or focus primarily on effectiveness [20,22,24], this synthesis clarifies what makes group chats distinctive and actionable for implementation.

In practice, group chat interventions may be most useful when deliberately structured to harness peer interaction and facilitation, or when integrated with complementary modalities such as brief onboarding or coaching to establish goals and norms [133], self-guided app or web modules for skill building reinforced through chat-based practice and feedback [134], and SMS or wearable-triggered prompts that feed into timely group problem-solving and accountability [135]. The program theory generated here supports implementers and policymakers in configuring design features, facilitation approaches, and contextual conditions (system, technology, setting, participant, and social) to activate specific mechanisms, moving beyond ad hoc use of group message threads toward purpose-built group environments. For researchers, priorities include targeted evaluations of chatbot-mediated group chat interventions, testing how agents deliver core functions (prompting, feedback, personalization, and triage) within groups, and comparative studies that isolate the contribution of group chat components relative to other digital modalities. With deliberate, theory-driven development, group chat interventions can be more consistently specified, implemented, and evaluated as a scalable and equitable component of digital health systems across diverse settings.

Acknowledgments

We would like to thank Jun Wei Goh, Joy Hui En Tan, Yee Rui Khoo, and Khon Wathan for their valuable contributions during the early phases of this project. The authors confirm that no generative artificial intelligence (AI) or AI-assisted technologies were used in the creation of text, figures, or other informational content in this manuscript.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: SHA
Data curation: HT, SHA
Formal analysis: HT, SHA
Investigation: HT, SHA
Methodology: JH, SHA
Supervision: SHA
Writing – original draft: HT
Writing – review & editing: HT, FJHC, JH, SHA

Conflicts of Interest

None declared.

Multimedia Appendix 1
Initial Program Theories for group chat–based health interventions (guided by social cognitive theory).
[DOCX File, 15 KB - jmir_v28i1e88911_app1.docx]

Multimedia Appendix 2
Cross-cutting facilitators and barriers for activating Initial Program Theory mechanisms in group chat–based interventions.
[DOCX File, 15 KB - jmir_v28i1e88911_app2.docx]

Multimedia Appendix 3
Final program theory synthesizing context-mechanism-outcome configurations in group chat health interventions was analyzed.
[DOC File, 46 KB - jmir_v28i1e88911_app3.doc]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 109 KB - jmir_v28i1e88911_app4.docx]

Checklist 2
PRISMA-S checklist.
[DOCX File, 18 KB - jmir_v28i1e88911_app5.docx]

References

1. Seckman C, Van de Castle B. Understanding digital health technologies using mind maps. J Nurs Scholarsh 2021 Jan;53(1):7-15 [FREE Full text] [doi: 10.1111/jnu.12611]
2. Huang F, Blaschke S, Lucas H. Beyond pilotitis: taking digital health interventions to the national level in China and Uganda. Global Health 2017 Jul 31;13(1):49. [doi: 10.1186/s12992-017-0275-z] [Medline: 28756767]
3. Zheng H, Ling R. Drivers of social media fatigue: a systematic review. Telemat Inform 2021 Nov;64:101696. [doi: 10.1016/j.tele.2021.101696]
4. Facts and figures 2024 - internet use. ITU. URL: https://www.itu.int/itu-d/reports/statistics/2024/11/10/ff24-internet-use [accessed 2026-03-17]
5. Stark AL, Geukes C, Dockweiler C. Digital health promotion and prevention in settings: scoping review. J Med Internet Res 2022 Jan 28;24(1):e21063. [doi: 10.2196/21063] [Medline: 35089140]
6. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. J Med Internet Res 2023 Feb 24;25(1):e40789. [doi: 10.2196/40789] [Medline: 36826990]
7. Marsch LA. Digital health data-driven approaches to understand human behavior. Neuropsychopharmacol 2021 Jan;46(1):191-196. [doi: 10.1038/s41386-020-0761-5]
8. Dagum P. Digital biomarkers of cognitive function. NPJ Digit Med 2018 Mar 28;1(1):10. [doi: 10.1038/s41746-018-0018-4]
9. Vereen RN, Kurtzman R, Noar SM. Are social media interventions for health behavior change efficacious among populations with health disparities?: a meta-analytic review. Health Commun 2023 Jan;38(1):133-140. [doi: 10.1080/10410236.2021.1937830] [Medline: 34148445]
10. Rothman AJ. “Is there nothing more practical than a good theory?”: Why innovations and advances in health behavior change will arise if interventions are used to test and refine theory. Int J Behav Nutr Phys Act 2004 Jul 27;1(1):11. [doi: 10.1186/1479-5868-1-11] [Medline: 15279674]
11. Grace-Farfaglia P. Social cognitive theories and electronic health design: scoping review. JMIR Hum Factors 2019 Jul 19;6(3):e11544. [doi: 10.2196/11544] [Medline: 31325290]

XSL•FO
RenderX

12. Goswami A, Poole L, Thorlu-Bangura Z, et al. The use of digital health interventions for cardiometabolic diseases among South Asian and Black minority ethnic groups: realist review. J Med Internet Res 2023 Jan 6;25:e40630. [doi: 10.2196/40630] [Medline: 36607732]
13. Villarreal-Zegarra D, Alarcon-Ruiz CA, Melendez-Torres GJ, et al. Development of a framework for the implementation of synchronous digital mental health: realist synthesis of systematic reviews. JMIR Ment Health 2022 Mar 29;9(3):e34760. [doi: 10.2196/34760] [Medline: 35348469]
14. Peláez Zuberbuhler J, Pietrantoni L, Mazzetti G, et al. A systematic realist synthesis of digital interventions for enhancing mental health at work: contexts, mechanisms, and outcomes. Int J Ment Health Syst 2025 Jan 9;19(1):3. [doi: 10.1186/s13033-024-00655-5] [Medline: 39780192]
15. Fitzpatrick PJ. Improving health literacy using the power of digital communications to achieve better health outcomes for patients and practitioners. Front Digit Health 2023;5:1264780. [doi: 10.3389/fdgth.2023.1264780] [Medline: 38046643]
16. Kuhlmann S, Ahlers-Schmidt CR, Steinberger E. TXT@WORK: pediatric hospitalists and text messaging. Telemed J E Health 2014 Jul;20(7):647-652. [doi: 10.1089/tmj.2013.0200] [Medline: 24784021]
17. How countries are using edtech (including online learning, radio, television, texting) to support access to remote learning during the COVID-19 pandemic. EdTech Hub. 2020. URL: https://docs.edtechhub.org/lib/RGLRHG75 [accessed 2025-11-07]
18. Carvalho S, Crawfurd L. School's out: now what? Center for Global Development. URL: https://www.cgdev.org/blog/schools-out-now-what [accessed 2025-11-07]
19. He Y, Yang L, Qian C, et al. Conversational agent interventions for mental health problems: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2023 Apr 28;25(1):e43862. [doi: 10.2196/43862] [Medline: 37115595]
20. Carvalho J, Francisco R, Relvas AP. Family functioning and information and communication technologies: How do they relate? A literature review. Comput Human Behav 2015 Apr;45:99-108. [doi: 10.1016/j.chb.2014.11.037]
21. Mannell K. Plural and porous: reconceptualizing the boundaries of mobile messaging group chats. J Comput Mediat Commun 2020 Jul 1;25(4):274-290. [doi: 10.1093/jcmc/zmaa007]
22. Manji K, Hanefeld J, Vearey J, Walls H, de Gruchy T. Using WhatsApp messenger for health systems research: a scoping review of available literature. Health Policy Plan 2021 Jun 1;36(5):594-605. [doi: 10.1093/heapol/czab024] [Medline: 33860314]
23. Nakamura CA, Seward N, Peters TJ, et al. A self-help mobile messaging intervention to improve subthreshold depressive symptoms among older adults in a socioeconomically deprived region of Brazil (PRODIGITAL): a pragmatic, two-arm randomised controlled trial. Lancet Reg Health Am 2024 Nov;39:100897. [doi: 10.1016/j.lana.2024.100897] [Medline: 39435262]
24. Treadgold BM, Coulson NS, Campbell JL, Lambert J, Pitchforth E. Quality and misinformation about health conditions in online peer support groups: scoping review. J Med Internet Res 2025 May 16;27(1):e71140. [doi: 10.2196/71140] [Medline: 40377972]
25. Eysenbach G, Powell J, Englesakis M, Rizo C, Stern A. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. BMJ 2004 May 15;328(7449):1166. [doi: 10.1136/bmj.328.7449.1166] [Medline: 15142921]
26. McCool J, Dobson R, Whittaker R, Paton C. Mobile health (mHealth) in low- and middle-income countries. Annu Rev Public Health 2022 Apr 5;43(1):525-539. [doi: 10.1146/annurev-publhealth-052620-093850] [Medline: 34648368]
27. Seiler J, Libby TE, Jackson E, Lingappa JR, Evans WD. Social media-based interventions for health behavior change in low- and middle-income countries: systematic review. J Med Internet Res 2022 Apr 14;24(4):e31889. [doi: 10.2196/31889] [Medline: 35436220]
28. Tudor Car L, Dhinagaran DA, Kyaw BM, et al. Conversational agents in health care: scoping review and conceptual analysis. J Med Internet Res 2020 Aug 7;22(8):e17158. [doi: 10.2196/17158] [Medline: 32763886]
29. Hou T, Ho MH, Li H, Lin CC. The effectiveness of instant messaging-based interventions on health behavior change: a systematic review and meta-analysis. Worldviews Evid Based Nurs 2025 Aug;22(4):e70066. [doi: 10.1111/wvn.70066] [Medline: 40717021]
30. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
31. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
32. Wong G, Westhorp G, Manzano A, Greenhalgh J, Jagosh J, Greenhalgh T. RAMESES II reporting standards for realist evaluations. BMC Med 2016 Jun 24;14(1):96. [doi: 10.1186/s12916-016-0643-1] [Medline: 27342217]
33. Bandura A. Social cognitive theory: an agentic perspective. Annu Rev Psychol 2001;52(1):1-26. [doi: 10.1146/annurev.psych.52.1.1] [Medline: 11148297]
34. Morrison LG. Theory-based strategies for enhancing the impact and usage of digital health behaviour change interventions: a review. Digit Health 2015;1:2055207615595335. [doi: 10.1177/2055207615595335] [Medline: 29942544]

XSL•FO
RenderX

35. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Aug 8;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
36. Abedishargh N, Ramezani Farani A, Gharraee B, Farahani H. Effectiveness of internet-based cognitive behavioral therapy in weight loss, stress, anxiety, and depression via virtual group therapy. Iran J Psychiatry Behav Sci 2021 Jan;15(3):e113096. [doi: 10.5812/ijpbs.113096]
37. Alhazmy RS, Khalil AH, Almutary H. Effects of an instructional WhatsApp group on self-care and HbA1c among female patients with type 2 diabetes mellitus. PLoS One 2024;19(9):e0305845. [doi: 10.1371/journal.pone.0305845] [Medline: 39292676]
38. Allotey D, Flax VL, Ipadeola AF, et al. Feasibility and acceptability of integrating a multicomponent breastfeeding promotion intervention into routine health services in private health facilities in Lagos State, Nigeria: a mixed methods process evaluation. PLoS One 2024;19(4):e0301695. [doi: 10.1371/journal.pone.0301695] [Medline: 38669231]
39. Anaje CC, Okpala CI, Enechukwu NA, Ezejiofor OI, Malachy DE, Nwiyi OK. The impact of WhatsApp as a health education tool in albinism: interventional study. JMIR Dermatol 2023 Nov 21;6:e49950. [doi: 10.2196/49950] [Medline: 37988154]
40. Bauer S, Wolf M, Haug S, Kordy H. The effectiveness of internet chat groups in relapse prevention after inpatient psychotherapy. Psychother Res 2011 Mar;21(2):219-226 [FREE Full text] [doi: 10.1080/10503307.2010.547530] [Medline: 21347978]
41. Carmona S, Alayed N, Al-Ibrahim A, D'Souza R. Realizing the potential of real-time clinical collaboration in maternal-fetal and obstetric medicine through WhatsApp. Obstet Med 2018 Jun;11(2):83-89. [doi: 10.1177/1753495X18754457] [Medline: 29997691]
42. Cheung YTD, Chan CHH, Lai CKJ, et al. Using WhatsApp and Facebook online social groups for smoking relapse prevention for recent quitters: a pilot pragmatic cluster randomized controlled trial. J Med Internet Res 2015 Oct 22;17(10):e238. [doi: 10.2196/jmir.4829] [Medline: 26494159]
43. Cheung WC, Miles LM, Hawkes RE, French DP. Experiences of online group support for engaging and supporting participants in the National Health Service Digital Diabetes Prevention Programme: a qualitative interview study. J Health Serv Res Policy 2024 Apr;29(2):100-110. [doi: 10.1177/13558196231212846] [Medline: 38096783]
44. Chory A, Nyandiko W, Martin R, et al. HIV-related knowledge, attitudes, behaviors and experiences of Kenyan adolescents living with HIV revealed in WhatsApp group chats. J Int Assoc Provid AIDS Care 2021;20:2325958221999579. [doi: 10.1177/2325958221999579] [Medline: 33657911]
45. Pereira AAC, Destro JR, Picinin Bernuci M, Garcia LF, Rodrigues Lucena TF. Effects of a WhatsApp-delivered education intervention to enhance breast cancer knowledge in women: mixed-methods study. JMIR Mhealth Uhealth 2020 Jul 21;8(7):e17430. [doi: 10.2196/17430] [Medline: 32706726]
46. Duong HT, Hopfer S. Let's Chat: development of a family group chat cancer prevention intervention for Vietnamese families. Health Educ Behav 2021 Apr;48(2):208-219. [doi: 10.1177/1090198121990389] [Medline: 33586455]
47. Duong HT, Hopfer S. "Let's Chat": process evaluation of an intergenerational group chat intervention to increase cancer prevention screening among Vietnamese American families. Transl Behav Med 2021 Apr 7;11(3):891-900. [doi: 10.1093/tbm/ibaa120] [Medline: 33290557]
48. Fan HSL, Ho MY, Ko RWT, et al. Feasibility and effectiveness of WhatsApp online group on breastfeeding by peer counsellors: a single-blinded, open-label pilot randomized controlled study. Int Breastfeed J 2022 Dec 22;17(1):91. [doi: 10.1186/s13006-022-00535-z] [Medline: 36544208]
49. Faridah F, Anies A, Kartasurya MI, Widjanarko B. Online educational intervention: Improving maternal knowledge and attitudes in providing developmental stimulation for stunting toddlers. Narra J 2024 Apr;4(1):e591. [doi: 10.52225/narra.v4i1.591] [Medline: 38798876]
50. Fauziansyah A. Support group WhatsApp chat method on knowledge of heart failure patients at RSUD dr. R. Koesma Tuban. JNSU 2024 May 27;12(1):9-17. [doi: 10.21776/ub.jik.2024.012.01.02]
51. Femyliati R, Fikri AM, Andriani E. The effect of nutrition education through social media on nutritional knowledge and menu selection in food delivery application. Amerta Nutr 2023 Jun 9;7(2):248-254. [doi: 10.20473/amnt.v7i2.2023.248-254]
52. Feng Y, Ma X, Zhang Q, et al. Effectiveness of WeChat-group-based parental health education in preventing unintentional injuries among children aged 0-3: randomized controlled trial in Shanghai. BMC Public Health 2022 Nov 16;22(1):2086. [doi: 10.1186/s12889-022-14462-5] [Medline: 36380326]
53. Feng Y, Li X, Ma X, et al. Using online social networks to provide a parental health-education intervention for preventing unintentional injuries among children aged 0–3 years: a randomized controlled trial and social network analysis in Shanghai, China. Front Public Health 2023 Jan 11;10:1049851. [doi: 10.3389/fpubh.2022.1049851]
54. Gerrits RS, van der Zanden RAP, Visscher RFM, Conijn BP. Master your mood online: a preventive chat group intervention for adolescents. Aust e-J Adv Ment Health 2007 Jan;6(3):152-162. [doi: 10.5172/jamh.6.3.152]
55. Golkaramnay V, Bauer S, Haug S, Wolf M, Kordy H. The exploration of the effectiveness of group therapy through an Internet chat as aftercare: a controlled naturalistic study. Psychother Psychosom 2007;76(4):219-225. [doi: 10.1159/000101500] [Medline: 17570960]

XSL•FO
RenderX

56. Guo Z, Luk TT, Weng X, et al. Behavioral support by ex-smoking peers using instant messaging for smoking cessation: a randomized controlled trial. CHEST 2026 Jan;169(1):300-311. [doi: 10.1016/j.chest.2025.07.4110] [Medline: 40939935]
57. He WJA, Wang Q, Lam TH, et al. Effectiveness of instant messaging-based online group support for preventing smoking relapse: a pragmatic randomized controlled trial. Int J Ment Health Addiction 2025 Jun;23(3):2460-2477. [doi: 10.1007/s11469-024-01239-7]
58. Hong YA, Shen K, Han HR, et al. Two-year follow-up of dementia caregivers after a digital health intervention WECARE: a mixed-method study. Aging Ment Health 2025 Apr;29(4):631-638. [doi: 10.1080/13607863.2024.2449141] [Medline: 39757693]
59. Huang F, Zhang S, Tian Y, et al. Effect of mobile health based peripartum management of gestational diabetes mellitus on postpartum diabetes: a randomized controlled trial. Diabetes Res Clin Pract 2021 May;175:108775. [doi: 10.1016/j.diabres.2021.108775] [Medline: 33771645]
60. Huang HY, Chen CL, Lin YL, et al. Optimizing treatment for hospitalized patients with COPD: a study on the impact of a LINE app-based multidisciplinary team approach targeting pharmacological treatment, lifestyle changes, and smoking cessation. Int J Chron Obstruct Pulmon Dis 2025;20:1149-1159. [doi: 10.2147/COPD.S511869] [Medline: 40264970]
61. Hucker A, McCabe MP. A qualitative evaluation of online chat groups for women completing a psychological intervention for female sexual dysfunction. J Sex Marital Ther 2014;40(1):58-68 [FREE Full text] [doi: 10.1080/0092623X.2012.675020] [Medline: 23768099]
62. Janowski R, Cluver LD, Shenderovich Y, et al. Optimizing engagement with a smartphone app to prevent violence against adolescents: results from a cluster randomized factorial trial in Tanzania. J Med Internet Res 2025 Mar 10;27(1):e60102. [doi: 10.2196/60102] [Medline: 40063069]
63. Jia Y, Xiao Y, Chen H, et al. Effects of group communication norms on daily steps in a team-based financial incentive mobile phone intervention in Shanghai, China. Int J Behav Nutr Phys Act 2025 Jan 18;22(1):9. [doi: 10.1186/s12966-025-01707-w] [Medline: 39827159]
64. Jia N, Schensul JJ, Zhang M, Kong L, Yan Q, Dai Y. Effectiveness of a mobile breastfeeding monitoring tool among mothers in WeChat groups on breastfeeding exclusivity and self-efficacy: intention-to-treat and per-protocol analyses of a randomized controlled trial. J Med Internet Res 2025 Aug 15;27(1):e67024. [doi: 10.2196/67024] [Medline: 40815829]
65. Jiang N, Zhao A, Rogers ES, et al. Feasibility and preliminary effects of a social media–based peer-group mobile messaging smoking cessation intervention among Chinese immigrants who smoke: pilot randomized controlled trial. JMIR Mhealth Uhealth 2024 Jul 22;12:e59496. [doi: 10.2196/59496] [Medline: 39037756]
66. Jocson RM, Alampay LP, Lachman JM, et al. Feasibility and acceptability of a digital parent group chat intervention to prevent child and adolescent maltreatment in the Philippines. Int J Psychol 2024 Dec;59(6):832-840. [doi: 10.1002/ijop.13148] [Medline: 38840552]
67. Kang G, Zhang H, Zhou J, Wan D. The WeChat platform-based health education improves health outcomes among patients with stable coronary artery disease. Patient Educ Couns 2023 Jun;111:107704. [doi: 10.1016/j.pec.2023.107704] [Medline: 36906932]
68. Keramatikerman M, Varaei S, Vaezi M, Sayadi L. Peer support-based online education, burden of care and quality of life among family caregivers of patients with leukaemia: non-randomised clinical trial. BMJ Support Palliat Care 2024 Dec 19;14(e3):e2827-e2835. [doi: 10.1136/spcare-2023-004610] [Medline: 38272654]
69. Kiirya Y, Kitaka S, Kalyango J, et al. Acceptability of an online peer support group as a strategy to improve antiretroviral therapy adherence among young people in Kampala district, Uganda: qualitative findings. BMC Infect Dis 2025 Apr 3;25(1):461. [doi: 10.1186/s12879-025-10831-8] [Medline: 40181265]
70. Lee H, Lee H, Kim Y, Kim S, Lee YM. Network support using social networking services to increase exercise adherence among Korean-Chinese middle-aged migrant women: mixed methods study. JMIR Mhealth Uhealth 2020 Nov 5;8(11):e19159. [doi: 10.2196/19159] [Medline: 33151155]
71. Liao Q, Fielding R, Cheung YTD, Lian J, Yuan J, Lam WWT. Effectiveness and parental acceptability of social-networking interventions for promoting seasonal influenza vaccination among young children: a randomized control trial (Preprint). J Med Internet Res 2020 Feb 28;22(2):e16427. [doi: 10.2196/16427] [Medline: 32130136]
72. Lieberman MA. Psychological characteristics of people with Parkinson's disease who prematurely drop out of professionally led Internet chat support groups. Cyberpsychol Behav 2007 Dec;10(6):741-748. [doi: 10.1089/cpb.2007.9956] [Medline: 18085960]
73. Lin BS, Li YY, Qiao C, et al. Implementation of WeChat - based patient–doctor interaction in the management of Helicobacter pylori infection: a propensity score matching analysis. J Dig Dis 2022 May;23(5-6):280-287. [doi: 10.1111/1751-2980.13114]
74. Mendivil-Aguayo P, Rivera M, Armendariz D, et al. Zoom & WhatsApp digital information and communication technologies (ICTs) enhance community engaged research with women immigrants from Mexico. J Community Pract 2024;32(2):212-237 [FREE Full text] [doi: 10.1080/10705422.2024.2351935] [Medline: 38883275]
75. Mirshahi A, Bakitas M, Khoshavi M, et al. The impact of an integrated early palliative care telehealth intervention on the quality of life of heart failure patients: a randomized controlled feasibility study. BMC Palliat Care 2024 Jan 23;23(1):22. [doi: 10.1186/s12904-024-01348-z] [Medline: 38254058]

XSL•FO
RenderX

76. Mohebbi P, Maleki A, Ebrahimi L, Mirzaeyan H. The effect of group counseling based on positive psychology on the WhatsApp social media platform on the severity of premenstrual syndrome symptoms: a randomized clinical trial. BMC Womens Health 2024 Nov 9;24(1):600. [doi: 10.1186/s12905-024-03437-w] [Medline: 39522001]
77. Owusu GP, Dordoye SAY, Asigbey S, Dordoye EK. WhatsApp as a tool for psychiatric crisis management: lessons from a Ghanaian support group. BMC Psychiatry 2025 Jul 11;25(1):694. [doi: 10.1186/s12888-025-07068-9] [Medline: 40646470]
78. Pankiv E, Achaibar K, Hossain A, Fiadjoe JE, Matava CT. The role of WhatsAppTM in pediatric difficult airway management: a study from the PeDI Collaborative. Pediatr Anesth 2023;33:1001-1011. [doi: 10.1111/pan.14760]
79. Rakhshani T, Dolatkhah SM, Kashfi SM, Khani Jeihooni A. The effect of a self-learned virtual learning package on knowledge, attitude, and self-care behaviors of COVID-19 in people referred to health and treatment centers. BMC Public Health 2024 Jun 26;24(1):1710. [doi: 10.1186/s12889-024-19233-y] [Medline: 38926819]
80. Ricci É, Dimov T, da Silva Cassais T, Dellbrügger AP. Occupational therapy in Brazil during the COVID-19 pandemic: peer support groups as mental health intervention strategy. World Fed Occup Ther Bull 2021 Jan 2;77(1):33-35. [doi: 10.1080/14473828.2020.1840767]
81. Rolando S, Arrighetti G, Fornero E, Farucci O, Beccaria F. Telegram as a space for peer-led harm reduction communities and netreach interventions. Contemp Drug Probl 2023 Jun;50(2):190-201. [doi: 10.1177/00914509221145196]
82. Ronen K, Mugo C, Kaggiah A, et al. Facilitated WhatsApp support groups for youth living with HIV in Nairobi, Kenya: single-arm pilot intervention study. JMIR Form Res 2023 Nov 13;7:e49174. [doi: 10.2196/49174] [Medline: 37955957]
83. Saleha N, Delfina R, Nurlaili N, Ardiansyah F, Nafratilova M. Online affirmation and peer support are effective for reducing medical personnel stress in dealing with Covid-19 patients. Unnes J Public Health 2021 Jul 15;10(2):151-161 [FREE Full text] [doi: 10.15294/ujph.v10i2.42861]
84. Scholl J, Kohls E, Görges F, et al. Acceptability and feasibility of the transfer of face-to-face group therapy to online group chats in a psychiatric outpatient setting during the COVID-19 pandemic: longitudinal observational study. JMIR Form Res 2021 Jul 23;5(7):e27865. [doi: 10.2196/27865] [Medline: 34161252]
85. Shi D, Liu C, Huang L, Chen XQ. Post-abortion needs-based education via the WeChat platform to lessen fear and encourage effective contraception: a post-abortion care service intervention-controlled trial. BMC Womens Health 2024 Mar 5;24(1):159. [doi: 10.1186/s12905-024-03004-3] [Medline: 38443889]
86. Skeen S, Marlow M, du Toit S, et al. Using WhatsApp support groups to promote responsive caregiving, caregiver mental health and child development in the COVID-19 era: a randomised controlled trial of a fully digital parenting intervention. Digit Health 2023;9:20552076231203893. [doi: 10.1177/20552076231203893] [Medline: 37928327]
87. Stewart M, Masuda JR, Letourneau N, et al. Online support intervention for adolescents with asthma and allergies. J Asthma Allergy Educ 2011 Dec;2(6):306-317. [doi: 10.1177/2150129711402686]
88. Sukriani LW, Arisani LG. Effecttiveness of WhatsApp group on breastfeeding practices. Indian J Forensic Med Toxicol 2020;14:3550-3555. [doi: 10.37506/ijfmt.v14i4.12178]
89. Tabira K, Oguma Y, Yoshihara S, et al. Digital peer-supported app intervention to promote physical activity among community-dwelling older adults: nonrandomized controlled trial. JMIR Aging 2024 May 30;7:e56184. [doi: 10.2196/56184] [Medline: 38814686]
90. Tanaka K, Sasai H, Wakaba K, et al. Professional dietary coaching within a group chat using a smartphone application for weight loss: a randomized controlled trial. J Multidiscip Healthc 2018;11:339-347. [doi: 10.2147/JMDH.S165422] [Medline: 30038502]
91. Tian Y, Zhang S, Huang F, Ma L. Comparing the efficacies of telemedicine and standard prenatal care on blood glucose control in women with gestational diabetes mellitus: randomized controlled trial. JMIR Mhealth Uhealth 2021 May 25;9(5):e22881. [doi: 10.2196/22881] [Medline: 33783365]
92. Tian X, Liu Y, Zhang J, et al. Efficacy of a WeChat-based, multidisciplinary, full-course nutritional management program on the nutritional status of patients with ovarian cancer undergoing chemotherapy: randomized controlled trial. JMIR Mhealth Uhealth 2024 Nov 4;12:e56475. [doi: 10.2196/56475] [Medline: 39496160]
93. Velloza J, Poovan N, Meisner A, et al. Adaptive HIV pre-exposure prophylaxis adherence interventions for young women in Johannesburg, South Africa: a sequential multiple-assignment randomised trial. Lancet HIV 2025 Feb;12(2):e105-e117. [doi: 10.1016/S2352-3018(24)00268-6] [Medline: 39681126]
94. Wasilewski MB, Kokorelias KM, Nonoyama M, et al. The experience of family caregivers of ventilator-assisted individuals who participated in a pilot web-based peer support program: a qualitative study. Digit Health 2022;8:20552076221134964. [doi: 10.1177/20552076221134964] [Medline: 36452271]
95. Weng X, Liu K, Xia Y, Song C, Guo N, Wang MP. Combined chatbot and counselor support for smoking cessation in China: a pilot, mobile-based, randomized controlled trial. Nicotine Tob Res 2025 Dec 22:ntaf260. [doi: 10.1093/ntr/ntaf260] [Medline: 41427871]
96. Weng X, Song C, Liu K, Zhao S, Yang H, Wang MP. Mobile phone-based smoking-cessation intervention in patients with chronic diseases in China: a sequential multiple assignment randomized trial. Nicotine Tob Res 2025 Feb 24;27(3):502-510. [doi: 10.1093/ntr/ntae230] [Medline: 39324617]

XSL•FO
RenderX

97. West DS, Harvey JR, Krukowski RA, Prewitt TE, Priest J, Ashikaga T. Do individual, online motivational interviewing chat sessions enhance weight loss in a group-based, online weight control program? Obesity (Silver Spring) 2016 Nov;24(11):2334-2340. [doi: 10.1002/oby.21645] [Medline: 27616628]
98. West DS, Stansbury M, Krukowski RA, Harvey J. Enhancing group-based internet obesity treatment: a pilot RCT comparing video and text-based chat. Obes Sci Pract 2019 Dec;5(6):513-520. [doi: 10.1002/osp4.371] [Medline: 31890241]
99. Wu Q, Kue J, Zhu X, et al. Effects of nurse-led support via WeChat, a smartphone application, for breast cancer patients after surgery: a quasi-experimental study. Telemed J E Health 2020 Feb;26(2):226-234. [doi: 10.1089/tmj.2018.0293] [Medline: 31313973]
100. Wulandari R, Suwandono A, Kartasurya MI, Nugraheni SA. Postpartum care behavior improvement during COVID-19 pandemic in Indonesia using mobile-health interactive message. Ethiop J Health Sci 2022 Mar;32(2):243-254. [doi: 10.4314/ejhs.v32i2.4] [Medline: 35693576]
101. Xu J, Akezhuoli H, Zhou M, et al. Development and evaluation of a culturally adapted digital-platform integrated multifaceted intervention to promote the utilization of maternal healthcare services: a single-arm pilot study. Int J Equity Health 2023 Oct 17;22(1):217. [doi: 10.1186/s12939-023-02033-y] [Medline: 37848902]
102. Yoshihara S, Takahashi K, Uemura C, Murakami S, Harada D, Yamato H. Effectiveness of a digital peer-supported app intervention in promoting smoking cessations: nonrandomized controlled trial. JMIR Mhealth Uhealth 2025 Aug 19;13:e68638. [doi: 10.2196/68638] [Medline: 40829152]
103. Yoshihara S, Takahashi K, Uemura C, et al. Smoking cessation strategies for different types of cigarette users using a digital peer-supported app and nicotine aids: prospective study. J Med Internet Res 2025 Aug 5;27(1):e75876. [doi: 10.2196/75876] [Medline: 40764052]
104. Yuan LP, Yu ZH, Zhang XC, et al. The psychological effect of forming WeChat groups between medical staff and patients with COVID-19. Front Public Health 2021;9:586465. [doi: 10.3389/fpubh.2021.586465] [Medline: 34249824]
105. Yurtsal B, Hasdemir Ö. Effects of the WhatsApp midwife breastfeeding support line on early postpartum breastfeeding process of mothers. Health Care Women Int 2022 Dec;43(12):1433-1448. [doi: 10.1080/07399332.2021.1972303] [Medline: 35417317]
106. Zanoni BC, Archary M, Sibaya T, et al. Acceptability, feasibility and preliminary effectiveness of the mHealth intervention, InTSHA, on retention in care and viral suppression among adolescents with HIV in South Africa: a pilot randomized clinical trial. AIDS Care 2024 Jul;36(7):983-992. [doi: 10.1080/09540121.2024.2361240] [Medline: 38976571]
107. Zerwas SC, Watson HJ, Hofmeier SM, et al. CBT4BN: a randomized controlled trial of online chat and face-to-face group therapy for bulimia nervosa. Psychother Psychosom 2017;86(1):47-53. [doi: 10.1159/000449025] [Medline: 27883997]
108. Zhang X, Chen X, Kourkoumelis N, Gao R, Li G, Zhu C. A social media-promoted educational community of joint replacement patients using the WeChat App: survey study. JMIR Mhealth Uhealth 2021 Mar 18;9(3):e18763. [doi: 10.2196/18763] [Medline: 33734094]
109. Zhang B, Kalampakorn S, Powwattana A, Sillabutra J, Liu G. A transtheoretical model-based online intervention to improve medication adherence for Chinese adults newly diagnosed with type 2 diabetes: a mixed-method study. J Prim Care Community Health 2024;15:21501319241263657. [doi: 10.1177/21501319241263657] [Medline: 39077970]
110. Zotti F, Dalessandri D, Salgarello S, et al. Usefulness of an app in improving oral hygiene compliance in adolescent orthodontic patients. Angle Orthod 2016 Jan;86(1):101-107. [doi: 10.2319/010915-19.1] [Medline: 25799001]
111. Zainuddin Z, Abdullah AZ, Jafar N, et al. Health literacy brisk walking exercise on clinical outcomes of blood sugar in patients with type 2 diabetes mellitus in Indonesia. Pharmacogn J 2023 Apr 30;15(2):433-438 [FREE Full text] [doi: 10.5530/pj.2023.15.68]
112. Djouma Nembot F, Buh Nkum C, Ateudjieu D, et al. Effectiveness of a WhatsApp-based communication on improving access to antenatal care interventions in sub-Saharan Africa: a randomized control health facility trial. Digit Health 2025;11:20552076251379349. [doi: 10.1177/20552076251379349] [Medline: 41000572]
113. Xie B, Li Y, Hwang WC, et al. A culturally tailored diabetes self-management education program with mobile health integration for Chinese Americans with type 2 diabetes: development and pilot evaluation study. JMIR Form Res 2026 Feb 3;10:e77372. [doi: 10.2196/77372] [Medline: 41632517]
114. Jackson L, Drury R, Paolo Azzaro G, et al. Investigating the acceptability and feasibility of three online interventions for caregivers of infants with feeding difficulties. Inquiry 2025;62:469580251375911. [doi: 10.1177/00469580251375911] [Medline: 41108178]
115. Liu X, Liu T, Chen T, Ma R. Efficacy of a digital peer support program on weight management and mental health in university students with preobesity: randomized controlled trial. JMIR Mhealth Uhealth 2026 Jan 21;14:e78960. [doi: 10.2196/78960] [Medline: 41564335]
116. Hoffman S, Hanass-Hancock J, Harrison AD, et al. Masibambane (Let Us Work Together), Ladies Chat: pilot evaluation of an online, gender-enhanced, group-interactive workshop to promote oral PrEP to young South African women. AIDS Behav 2025 Sep;29(9):2932-2945. [doi: 10.1007/s10461-025-04747-2] [Medline: 40506646]
117. Goh K, Contractor S, Van Belle S. A realist-informed review of digital empowerment strategies for adolescents to improve their sexual and reproductive health and well-being. J Urban Health 2022 Dec;99(6):1141-1156. [doi: 10.1007/s11524-022-00678-8] [Medline: 36070044]

XSL•FO
RenderX

118. Schlief M, Saunders KRK, Appleton R, et al. Synthesis of the evidence on what works for whom in telemental health: rapid realist review. Interact J Med Res 2022 Sep 29;11(2):e38239. [doi: 10.2196/38239] [Medline: 35767691]
119. Shahid N, Parker G, Bielecki JM, Rac V, Berta W. A realist review of factors critical for the implementation of eHealth in chronic disease management. BMC Health Serv Res 2025 Apr 2;25(1):496. [doi: 10.1186/s12913-025-12361-0] [Medline: 40176012]
120. Ho CWL, Caals K, Zhang H. Heralding the digitalization of life in post-pandemic East Asian societies. J Bioeth Inq 2020 Dec;17(4):657-661. [doi: 10.1007/s11673-020-10050-7] [Medline: 33169256]
121. Shi J, Kim HK, Salmon CT, Tandoc EC, Goh ZH. Cultural tightness-looseness and normative social influence in eight Asian countries: associations of individual and collective norms with vaccination intentions. Soc Sci Med 2024 Jan;340:116431. [doi: 10.1016/j.socscimed.2023.116431] [Medline: 38000175]
122. Dans A, Ng N, Varghese C, Tai ES, Firestone R, Bonita R. The rise of chronic non-communicable diseases in southeast Asia: time for action. Lancet 2011 Feb 19;377(9766):680-689. [doi: 10.1016/S0140-6736(10)61506-1] [Medline: 21269677]
123. Pradhan J, Pai M, Dwivedi R, et al. Burden of non-communicable diseases in South Asia: a decomposition analysis. J Health Popul Nutr 2025 Apr 18;44(1):124. [doi: 10.1186/s41043-025-00827-0] [Medline: 40251654]
124. Park S, Humphry J. Exclusion by design: intersections of social, digital and data exclusion. Inf Commun Soc 2019 Jun 7;22(7):934-953. [doi: 10.1080/1369118X.2019.1606266]
125. Ng MY, Olgin JE, Marcus GM, Lyles CR, Pletcher MJ. Email-based recruitment into the health eHeart study: cohort analysis of invited eligible patients. J Med Internet Res 2023 Dec 22;25(1):e51238. [doi: 10.2196/51238] [Medline: 38133910]
126. Krukowski RA, Ross KM, Western MJ, et al. Digital health interventions for all? Examining inclusivity across all stages of the digital health intervention research process. Trials 2024 Jan 30;25(1):98. [doi: 10.1186/s13063-024-07937-w] [Medline: 38291539]
127. Burgess RA, Jennings HM, Mathias K, et al. Nurturing transformative participation: a participatory realist review of mental health interventions in low resource settings. SSM - Ment Health 2025 Jun;7:100416. [doi: 10.1016/j.ssmmh.2025.100416] [Medline: 41542295]
128. Pasitselska O, Buehling K, Gagrčin E. Chat groups as local civic infrastructure: a case study of “Solidary neighborhood help” Telegram groups during the COVID-19 pandemic in Germany. New Media Soc 2025 Jul 13. [doi: 10.1177/14614448251349426]
129. Wilson KR, Roskill OM, Mahr J. Mutual aid using digital technology: a case study of virtual community organizing during the COVID-19 pandemic. J Community Pract 2022 Jul 3;30(3):255-278. [doi: 10.1080/10705422.2022.2102101]
130. Isari D, Pontiggia A, Virili F. Working with tweets vs. working with chats: an experiment on collaborative problem solving. Comput Human Behav 2016 May;58:130-140. [doi: 10.1016/j.chb.2015.12.052]
131. Wang D, Wang H, Yu M, Ashktorab Z, Tan M. Group chat ecology in enterprise instant messaging: how employees collaborate through multi-user chat channels on slack. Proc ACM Hum-Comput Interact 2022 Mar 30;6(CSCW1):1-14 [FREE Full text] [doi: 10.1145/3512941]
132. Rafikova A, Voronin A. Human–chatbot communication: a systematic review of psychological studies. AI Soc 2025 Oct;40(7):5389-5408. [doi: 10.1007/s00146-025-02277-y]
133. Chao EC, Zhang M, Houle MA, Rataj H. Collaboratively designing an app for a more personalized, community-endorsed continuous glucose monitoring onboarding experience: an early study. J Diabetes Sci Technol 2024 Jan;18(1):14-21. [doi: 10.1177/19322968231213654] [Medline: 37978817]
134. Jenness JL, Bhattacharya A, Kientz JA, Munson SA, Nagar RR. Lessons learned from designing an asynchronous remote community approach for behavioral activation intervention for teens. Behav Res Ther 2022 Apr;151:104065. [doi: 10.1016/j.brat.2022.104065] [Medline: 35248749]
135. Wang L, Miller LC. Just-in-the-Moment Adaptive Interventions (JITAI): a meta-analytical review. Health Commun 2020 Nov;35(12):1531-1544. [doi: 10.1080/10410236.2019.1652388] [Medline: 31488002]

Abbreviations:

AI: artificial intelligence
CMO: context-mechanism-outcome
IPT: Initial Program Theory
LMIC: low- and middle-income country
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RAMESES II: Realist And Meta-narrative Evidence Syntheses: Evolving Standards II
RCT: randomized controlled trial
SCT: social cognitive theory

XSL•FO
RenderX

Edited by S Brini; submitted 03.Dec.2025; peer-reviewed by G Owusu, W Evans; accepted 25.Feb.2026; published 13.Apr.2026.

Please cite as:
Thu H, Chan FJH, Hashim J, Ali SH
Using Group Chats to Drive Behavior Change in Digital Health Interventions: Scoping Review and Realist Synthesis
J Med Internet Res 2026;28:e88911
URL: https://www.jmir.org/2026/1/e88911
doi:10.2196/88911

© Hein Thu, Felicia Jia Hui Chan, Jumana Hashim, Shahmir H Ali. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Technologies and Biomarkers for Locomotor Capacity Assessment in Older Adults: Systematic Review

Shuhan Zhou[1*], PhD; Jundan Huang[1*], PhD; Jia Yu[1], MSN; Xiaoyang Li[1], PhD; Chi Zhang[1], PhD; Yinan Zhao[1], PhD; Mingyue Hu[1], PhD; Hui Feng[1,2,3], PhD

[1]Xiangya School of Nursing, Central South University, 172 Tongzipo Road, Yuelu District, Changsha, China

[2]Oceanwide Health Management Institute, Central South University, Changsha, China

[3]Hunan Engineering Research Center for Intelligent Medical Care, Central South University, Changsha, China

[*]these authors contributed equally

Corresponding Author:
Hui Feng, PhD
Xiangya School of Nursing, Central South University, 172 Tongzipo Road, Yuelu District, Changsha, China

Abstract

Background: Locomotor capacity, encompassing endurance, balance, muscle strength, muscle function, muscle power, and joint function of the body, is a key determinant of functional ability in older adults. Assessment tools based on digital technologies for objectively assessing locomotor capacity are increasingly being developed, but their reliability, validity, and clinical potential remain underexplored.

Objective: This systematic review aims to evaluate the current state of digital technologies, assess their validity and reliability for assessing locomotor capacity, and facilitate their effective implementation in clinical settings.

Methods: Systematic literature searches were performed in 6 electronic databases from inception to March 7, 2025. Citation lists from the included studies and gray literature from Google Scholar were additionally searched. Studies focusing on the reliability and validity of digital technologies for assessing locomotor capacity in general older adults were included. Standardized forms were used to extract information on study characteristics, participant demographics, digital technology details, and validity and reliability results. Methodological quality assessment and rating of measurement properties were conducted in accordance with the COSMIN (Consensus-Based Standards for the Selection of Health Measurement Instruments) guidelines.

Results: A total of 14 studies were included, of which 13 assessed balance using inertial measurement units, smartphones, balance boards, and force plates, and 1 assessed muscle power using smartphones. Fifty-one digital biomarkers were identified, including 47 for balance and 4 for muscle power assessment. Test-retest reliability coefficients ranged from 0.016 to 0.97, and validity was context specific. Overall, 13 studies demonstrated sufficient test-retest reliability and validity, whereas 1 study was rated as insufficient for convergent validity. Methodological quality was rated as “doubtful” or “inadequate” in 11 studies.

Conclusions: This review provides a comprehensive summary of digital technologies for assessing locomotor capacity in older adults and identifies 51 digital biomarkers with generally acceptable reliability and validity. Unlike previous studies that focused on specific sensor types or disease-specific populations, this review integrates evidence across technologies within general older populations, providing insights into the clinical application potential of digital biomarkers as well as the key translational barriers limiting their real-world implementation. Specifically, existing digital technologies show considerable promise for early detection of functional decline, longitudinal monitoring, and informing personalized interventions. However, their clinical applicability remains constrained by limited assessment of certain locomotor components and by methodological shortcomings across current studies. Future research should prioritize rigorous, high-quality investigations that expand evaluation to a broader range of locomotor components in real-world settings while developing age-friendly tools with enhanced clinical interpretability.

Trial Registration: PROSPERO CRD420251074143; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251074143

(J Med Internet Res 2026;28:e83814) doi:10.2196/83814

KEYWORDS

locomotor capacity; mobility; digital technology; digital biomarkers; assessment tools

Introduction

With an aging population, physical disability among older adults is becoming a pressing public health issue [1]. Locomotor capacity is a key factor for maintaining independence in daily activities [2]. The World Health Organization defines locomotor capacity as “a state (static or dynamic over time) of the musculoskeletal system that encompasses endurance, balance,

XSL•FO
RenderX

muscle strength, muscle function, muscle power, and joint function of the body" [3]. In older adults, locomotion limitations increase the risk of adverse outcomes, such as frailty, falls, hospitalization, and death [4]. The prevalence of positive screenings for locomotion limitation varies from 2.8% to 52% depending on the setting [5].

Traditional methods for evaluating locomotor capacity, such as the Short Physical Performance Battery [6], exhibit several limitations, including the requirement for specialized assessor training, susceptibility to assessor bias, and ceiling effects [7,8]. Given these limitations, there is growing interest in developing assessment tools using digital technologies that offer more objective, quantifiable, and frequent monitoring of locomotor capacity [9-11]. Moreover, these frequent measures can capture short-term intraindividual fluctuations that may be early and sensitive indicators of negative trends and critical transitions [12]. For example, Kumar et al [13] used a tri-axial accelerometer worn by older adults over 48 hours to identify continuous walking bouts of 20, 30, 40, 50, and 60 seconds without pauses. They found that gait variability, asymmetry, and irregularity produced a sensitivity of 76.8% and specificity of 80% between robust and physical prefrail or frail individuals. By identifying these early signs from an individual's typical mobility pattern, digital biomarkers have the potential to contribute to early risk prediction models for functional decline and disability, offering insights that conventional assessments may miss [14]. Digital technologies for measuring digital biomarkers can be categorized into 3 types [15]. First, wearable devices, such as wrist-worn accelerometers and portable sensors, provide continuous monitoring of movement patterns. Second, smartphones or applications are used for assessment. Third, nonvisual technologies enable passive and unobtrusive measurement in sensor-equipped environments.

Growing empirical evidence highlights the potential of digital technologies for assessing locomotor capacity. For instance, Burq et al [16] developed smartwatch-based unsupervised active tests of motor function in Parkinson disease and established the analytical validity of associated digital measures. Zhu et al [10] explored the use of inertial measurement units (IMUs) for assessing lower limb muscle strength during sit-to-stand transfers in older adults. Smartphone-based accelerometry data measured during sit-to-stand tests were also useful to identify subtle changes in locomotor capacity [17]. Several recent reviews [18-20] have summarized the application of digital technologies for assessing locomotor capacity. However, these reviews primarily focused on specific sensor types or disease populations, thereby overlooking other viable technologies and providing limited insight into assessments conducted in general older adult populations. In addition, the lack of robust evidence regarding the reliability and validity of digital assessment tools significantly limits their clinical utility [21]. The gap between research use and clinical application remains insufficiently established.

Therefore, this systematic review aimed to summarize digital technologies and biomarkers and evaluate their performance in assessing locomotor capacity in general older adults. The objectives were to (1) describe the types and mounting locations of digital technologies or devices used to measure overall or specific aspects of locomotor capacity; (2) identify digital biomarkers and corresponding assessment tasks derived from these technologies; and (3) assess the reliability and validity of digital technologies for detecting locomotor capacity.

Methods

Review Protocol and Amendments

This systematic review (registered in PROSPERO [International Prospective Register of Systematic Reviews]: CRD420251074143) followed the guidelines outlined in the PRISMA-COSMIN (Preferred Reporting Items for Systematic Reviews and Meta-Analyses-Consensus-Based Standards for the Selection of Health Measurement Instruments) for Outcome Measures Instruments in Systematic Reviews (OMIs) (Checklist 1) [22]. The only amendments to the registered review protocol were the inclusion of additional searches from Google Scholar and citation lists to ensure a more comprehensive review of the relevant literature.

Search Strategies

The literature search was reported according to PRISMA-S to promote transparency of the search process [23]. The search terms were constructed using the PICO (Population, Intervention, Control, and Outcomes) framework, as displayed in Table S1 in Multimedia Appendix 1. Six databases, including PubMed (National Library of Medicine), Web of Science (Clarivate), Cochrane Library (Wiley), IEEE Xplore (IEEE), Scopus (Elsevier), and Embase (Elsevier), were systematically searched on March 7, 2025, without any restrictions or search filters. Major terms related to older adults, digital technology, assessment, and locomotor capacity were combined for the initial search, according to the review by Honvo et al [24]. The search strategy was developed in consultation with an experienced librarian and experts in geriatric medicine and digital health. In addition, we searched citation lists from included studies by browsing reference lists and gray literature on Google Scholar on November 17, 2025, limiting the search to the first 100 publications, as search results beyond this number rapidly lost relevance to the topic of the review [25]. No other online resources, such as conference proceedings or print journals, were specifically browsed for this review. No study registries were searched, and no additional studies were sought by contacting authors, manufacturers, or others. The search strategy is detailed in Table S2 in Multimedia Appendix 1. The search process was not updated after the initial search date.

Eligibility Criteria

Following the PICO framework, the inclusion criteria were (1) study population (P): studies enrolling older adults as research subjects or separately reporting results on older people; (2) interventions (I): studies focusing on digital biomarkers (defined as objective, quantifiable physiological and behavioral measurements acquired through portable, wearable, implantable, or ingestible digital devices [15]), derived from digital technologies; (3) comparators (C): any or none; (4) outcomes (O): studies on the assessment of locomotor capacity (ie, endurance, balance, muscle strength, muscle function, muscle

XSL•FO
RenderX

power, joint function) while examining the measurement properties of reliability and validity; and (5) studies written in English.

Reliability focused on 2 types [26,27]: (1) test - retest reliability: the extent to which scores for patients who have not changed are the same for repeated measurements over time; and (2) measurement error: the systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured.

Validity prioritized 2 categories [27]: (1) criterion validity: the degree to which the scores of a device are an adequate reflection of a "gold standard"; and (2) hypothesis testing for construct validity: the degree to which the scores of a device are consistent with hypotheses, including convergent validity (measured by relationships to scores of other instruments) and discriminative validity (assessed through differences between relevant groups).

The exclusion criteria included the following: (1) systematic reviews and literature reviews; (2) books and other non–peer-reviewed literature; (3) studies conducted on populations with specific medical conditions (eg, Parkinson disease, stroke); (4) studies focusing on digital measures during rehabilitation exercises (eg, supervised therapeutic movements, postsurgical rehabilitation protocols, physical therapy interventions, and exercise programs designed for recovery or strength training); (5) studies lacking data on measurement properties; and (6) studies without full text.

Study Selection

After the literature search, all identified studies were imported into EndNote 21, and duplicates were removed. Two authors independently screened the remaining titles and abstracts for potential eligibility and then retrieved the full-text articles of the selected studies. Any discrepancies regarding study selection were resolved through consultation with a third reviewer.

Data Extraction

Two authors independently extracted data using standardized forms (Table S3 in Multimedia Appendix 1), and discrepancies were resolved through discussion among the 3 authors. Data items comprised: (1) characteristics of studies (first author, year of publication, and study design); (2) participant demographic information (sample size, age, sex distribution, and setting); (3) digital technology- or device-related information (type, number, assessment tasks, and digital biomarkers); and (4) results of validity and reliability.

Quality Assessment

Overview

Quality assessment was independently undertaken by 2 authors based on the COSMIN guidelines [28], which comprised 3 substeps: the methodological quality of each study for a measurement property, the rated result of each study for a measurement property, and the pooled results of all available studies for a measurement property. Given the substantial heterogeneity in locomotion protocols and variables across the included studies, pooled results were not performed. In cases of disagreement during the evaluation process, a third researcher was consulted to reach a consensus.

Methodological Quality Assessment

For the methodological quality of the included studies, 4 measurement properties from the consensus-based COSMIN checklist [29] were evaluated: test-retest reliability, measurement error, criterion validity, and hypothesis testing for construct validity. As this checklist was originally developed to assess measurement properties of a patient-reported outcome measure, the extended version of COSMIN was used to replace the original boxes on test-retest reliability or measurement error [30]. The methodological quality of each study for a measurement property was categorized as very good, adequate, doubtful, or inadequate. The "worst score counts" principle was applied to determine the overall methodological quality of each measurement property. When a study reported a range of values or multiple values for a measurement property, the best value was selected for the measurement property rating [24].

Measurement Properties Quality Assessment

We first assessed each study for its measurement property quality according to the updated COSMIN criteria for good measurement properties [28]. Each property's quality was rated as sufficient (+), insufficient (−), or indeterminate (?).

When at least 2 studies used the same protocols and variables, we synthesized the results. If the ratings were consistent across studies, the results from different studies for one measurement property were qualitatively summarized or pooled using meta-analysis, ultimately assigning a rating of "+" or "−."

Results

Literature Search Result

A total of 14,680 articles were identified from 6 databases, and 4,733 duplicates were removed. Following the initial screening of titles and abstracts, 123 articles were fully evaluated, with 9 of them eligible for inclusion in the review [31-39]. An additional 5 publications were obtained from reference lists [40,41] and gray literature [42-44]. The study selection process is presented in Figure 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for the study selection process. This diagram illustrates the study selection process for the systematic review of digital measures of locomotor capacity in older adults, detailing each stage (from record identification through screening, eligibility assessment, and final inclusion) and providing explicit reasons for exclusion at each stage.

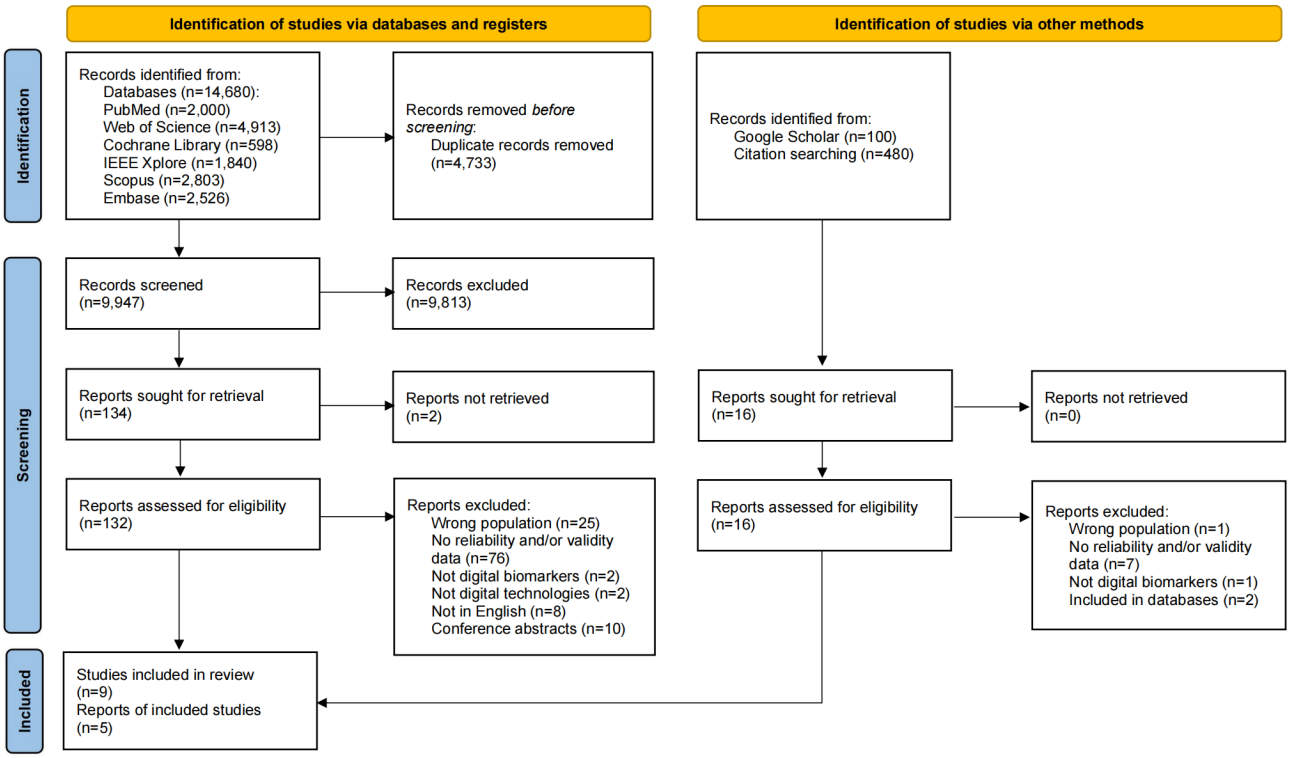

Study Characteristics

The main characteristics of the included studies and populations are summarized in Table 1. These studies were published between 2013 and 2024 and were conducted across 10 countries, primarily in the United States (n=3). Thirteen studies employed a cross-sectional design, while 1 adopted a pretest-posttest design. Across all studies, balance was the predominant assessment target (n=13), with only 1 study evaluating muscle power. No eligible studies assessed endurance, muscle strength, or joint function, and the exclusion reasons included incorrect population [45], lack of reliability and/or validity data [46], studies not involving digital biomarkers or technologies [47,48], and conference abstracts [49]. The sample sizes ranged from 12 to 248, with the majority from community-dwelling populations.

XSL•FO
RenderX

Table . Characteristics of the included studies.

Study	Country	Study design	Assessment target	Sample (count; age[a] (y); male participants)	Population sources
Cerrito et al [31] (2015)	Switzerland	Cross-sectional study	Muscle power	16; 73.5 (10.4); 16	Retirement homes and community-dwelling
Chang et al [42] (2013)	China	Pretest-posttest design	Balance	20; 67.32 (3.43); NR[b]	NR
da Silva et al [32] (2013)	Brazil	Cross-sectional study	Balance	28; 69 (5); 8	Community-dwelling
De Groote et al [33] (2021)	Belgium	Cross-sectional study	Balance	97; 50 - 90; 42	Community-dwelling
Greene et al [34] (2022)	Ireland	Cross-sectional study	Balance	168; 75.0 (7.2); 51	Community-dwelling
Harro and Garascia [40] (2019)	United States	Cross-sectional study	Balance	46; 67.7 (5.1); 22	Community-dwelling
Kuntapun et al [35] (2020)	Thailand	Cross-sectional study	Balance	12; 75.6 (5.6); 3	NR
Levy et al [36] (2018)	United States	Cross-sectional study	Balance	49; 71.3 (7.3); 23	Community-dwelling
McManus et al [37] (2022)	Ireland	Cross-sectional study	Balance	248; 74.9 (6.5); 91	Community-dwelling
Okada et al [38] (2024)	Japan	Cross-sectional study	Balance	27; 74.7 (7.1); 7	Community-dwelling
Olsen et al [43] (2023)	New Zealand	Cross-sectional study	Balance	34; 42 - 94; 14	Community-dwelling
Pooranawatthanakul and Siriphorn [44] (2023)	Thailand	Cross-sectional study	Balance	20; 70.85 (4.09); 3	Community-dwelling
Scaglioni-Solano and Aragón-Vargas [41] (2013)	Costa Rica	Cross-sectional study	Balance	37; 69 (8); NR	Exercise groups
Zhou et al [39] (2021)	United States	Cross-sectional study	Balance	15; 71.4 (5.9); 7	Community-dwelling

[a]Age values are given as mean (SD) or range.

[b]NR: not reported.

Balance Assessment

Device Type and Number

Table 2 provides an overview of the devices, tasks, variables, missing data, and data denoising methods. In terms of the assessment of balance, the most common device type was a smartphone (n=5), followed by a force plate (n=3), an IMU (n=2), a balance board (n=2), and an infrared depth sensor (n=1). All 13 studies used a single device for data collection. Two studies further assessed the reliability and validity across different sensor placement locations.

XSL•FO
RenderX

Table . Description of devices, tasks, variables, and data processing in the included studies.

Study	Devices (type; number)	Task settings	Device variables	Missing data and handling methods	Data denoising methods
Balance					
Chang et al [42] (2013)	Balance board; n=1	Three 10-s trials of stance with EO[a], stance with EC[b], and one-leg stance with EO	Center of gravity displacement	NR[c]	NR
da Silva et al [32] (2013)	Force plate; n=1	Three 30-s trials of the one-leg stance	COP[d] area, COP sway RMS[e] amplitude AP[f], COP sway RMS amplitude ML[g], mean velocity AP, mean velocity ML, mean frequency AP, and mean frequency ML	Complete data; NA[h]	Low-pass second-order Butterworth filter
De Groote et al [33] (2021)	Smartphone; n=1	Four 35-s trials of bipodal stance with EO, bipodal stance with EC, a dual task consisting of bipodal stance with EO and a cognitive task, and semi-tandem stance with EO	RMS acceleration ML, RMS acceleration AP, mean acceleration ML, and mean acceleration AP	Complete data; NA	Savitzky-Golay filter
Greene et al [34] (2022)	IMU[i]; n=1 (on thigh or sternum)	Five times sit-to-stand test	Balance model: including total time, coefficient of variation of sit-stand-sit time, mean time to stand up, spectral entropy angular velocity, mean Z-axis acceleration at stand start, and mean Z-axis acceleration at sit end	Complete data; NA	Butterworth IIR digital filter; signal validity checks
Harro and Garascia [40] (2019)	Force plate; n=1	Three standardized tests: limits of stability, motor control test, and sensory organization test	Limits of stability average end point excursion, motor control test average latency, sensory organization test composite equilibrium, and sensory organization test vestibular ratio scores	NR	NR
Kuntapun et al [35] (2020)	Smartphone; n=1 (attached horizontally at the level of the third lumbar vertebrae or rested on the right hip)	Walking 10 m at self-selected comfortable speed under three conditions: level walking, irregular surface walking, and obstacle crossing	COM[j] AP displacement, COM ML displacement, and COM SI[k] displacement	NR	Butterworth fourth-order low-pass filter
Levy et al [36] (2018)	Balance tracking system (consisting of a force plate and software); n=1	Six 20-s static standing trials of 3 with EO and 3 with EC	COP sway ML and COP sway AP	Complete data; NA	Dual-pass Butterworth filter
McManus et al [37] (2022)	IMU; n=1	Two 30-s trials of semi-tandem stance with EO and a narrow stance with EC	Balance score: including RMS acceleration magnitude, RMS acceleration ML, and RMS acceleration AP	Complete data; NA	Fourth order low-pass Butterworth filter

XSL•FO
RenderX

Study	Devices (type; number)	Task settings	Device variables	Missing data and handling methods	Data denoising methods
Okada et al [38] (2024)	Infrared depth sensor; n=1	20-s stepping-in-place test	Total movement distance, knee movement distance, maximum movement displacement, sway index, and step number	NR	NR
Olsen et al [43] (2023)	Smartphone; n=1	Four 30-s standing tasks: firm surface with EO, firm surface with EC, compliant surface with EO, and compliant surface with EC; and two 6-s walking tasks with EO, each performed four times: looking straight ahead and turning head	Postural stability, postural stability ML, postural stability AP, walking speed, mean step length, mean step time, step length variability, step time variability, step length asymmetry, and step time asymmetry	NR	NR
Pooranawatthanakul and Siriphorn [44] (2023)	Smartphone; n=1	Three standardized tests: the Modified Clinical Test of Sensory Interaction in Balance (ie, standing on a flat surface and a foam surface with EO or EC), a single-leg stance test, and a limit of stability test	RMS acceleration	Complete data; NA	Fourth-order Butterworth filter
Scaglioni-Solano and Aragón-Vargas [41] (2013)	Balance board; n=1	Four 30-s standing tasks (firm surface with EO, firm surface with EC, compliant surface with EO, and compliant surface with EC) and 10-s tandem stance test	COP displacement	Complete data; NA	Eighth-order Butterworth filter
Zhou et al [39] (2021)	Smartphone; n=1	Three 30-s trials of EO, EC, and dual task standing	Two-dimensional path length and RMS acceleration	Complete data; NA	Low-pass filter
Muscle power					
Cerrito et al [31] (2015)	Smartphone; n=1	Sit-to-stand movement	Total movement duration, peak force, rate of force development, and peak power	Complete data; NA	The signal was digitally low-pass filtered

[a]EO: eyes open.
[b]EC: eyes closed.
[c]NR: not reported.
[d]COP: center of pressure.
[e]RMS: root mean square.
[f]AP: anteroposterior.
[g]ML: mediolateral.
[h]NA: not applicable.
[i]IMU: inertial measurement unit.
[j]COM: center of mass.
[k]SI: superior-inferior.

Assessment Tasks

The included studies employed a diverse array of tasks designed to assess different aspects of balance. Nine studies [32,33,36,37,39,41-44] applied static standing tasks with eyes open and eyes closed, including bipodal, semi-tandem, narrow, dual task, and one-leg stance. Dynamic balance tasks were evaluated in 6 studies [34,35,38,40,43,44] using tasks such as

XSL•FO
RenderX

walking, the 5-time sit-to-stand test, stepping-in-place test, limits of stability, motor control test, and sensory organization test. Two studies [43,44] conducted both static and dynamic standing tasks.

Device Variables

As presented in Table 3, a total of 47 distinct variables related to balance were identified from 13 studies, categorized as temporal (8, 17%), spatial (12, 25.5%), linear acceleration (11, 23.4%), angular velocity (1, 2.1%), position (11, 23.4%), and energy (4, 8.5%). Five studies [32,34,39,40,43] simultaneously assessed multiple categories of variables, primarily temporal, linear acceleration, and position variables.

XSL•FO
RenderX

Table . Classification and definitions of digital biomarkers of balance assessment.

Category and variable	Definition (unit)	Study
Temporal		
Walking speed	The mean of the ratios of step length to step time (m/s)	Olsen et al [43] (2023)
Mean step time	The mean of the time between two consecutive initial contacts of alternative feet (s)	Olsen et al [43] (2023)
Step time variability	The RMS[a] of the SD of left step times and the SD of right step times (%)	Olsen et al [43] (2023)
Step time asymmetry	The percentage difference between left and right mean step times compared to the overall mean step time (%)	Olsen et al [43] (2023)
Total time to complete five times sit-to-stand test	The total time to complete five times sit-to-stand test (s)	Greene et al [34] (2022)
Coefficient of variation of sit-stand-sit time	The coefficient of variation of sit–stand–sit time across repetitions (%)	Greene et al [34] (2022)
Mean time to stand up	The mean time of the stand-up phase across repetitions (s)	Greene et al [34] (2022)
Motor control test average latency	The average time between the force plate translation and the individual's active force responses in each leg (ms)	Harro and Garascia [40] (2019)
Spatial		
Mean step length	The mean of the AP[b] distance between two consecutive initial contacts of alternative feet (m)	Olsen et al [43] (2023)
Step length variability	The RMS of the SD of left step lengths and the SD of right step lengths (%)	Olsen et al [43] (2023)
Step length asymmetry	The percentage difference between left and right mean step lengths compared to the overall mean step length (%)	Olsen et al [43] (2023)
Periodicity index	The step symmetry between the right and left step within a stride and the gait regularity across strides (%)	Olsen et al [43] (2023)
COM[c] AP displacement	The average of minimum to maximum displacement in the AP direction across all step cycles (cm)	Kuntapun et al [35] (2020)
COM ML[d] displacement	The average of minimum to maximum displacement in the ML direction across all step cycles (cm)	Kuntapun et al [35] (2020)
COM SI[e] displacement	The average of minimum to maximum displacement in the SI direction across all step cycles (cm)	Kuntapun et al [35] (2020)
Total movement distance	The total linear distance determined from the sum total of step-by-step changes in head position from its initial to its final position (m)	Okada et al [38] (2024)
Knee movement distance	The mean of the sum of the three-dimensional movements of each knee joint point during stepping (m)	Okada et al [38] (2024)
Maximum movement displacement	The maximum displacement of the head along the movement path from its initial position (m)	Okada et al [38] (2024)
Sway index	The ratio of total movement distance to maximum movement displacement	Okada et al [38] (2024)
Step number in 20-s stepping-in-place test	The total number of foot contacts beginning with the return of the right foot to the ground after raising the right leg to initiate the test (steps)	Okada et al [38] (2024)
Linear acceleration		

XSL•FO
RenderX

Category and variable	Definition (unit)	Study
RMS acceleration	RMS of acceleration in AP and ML directions (mm/s^2)	Pooranawatthanakul and Siriphorn [44] (2023); Zhou et al [39] (2021)
RMS acceleration magnitude	RMS of the magnitude of the 3-axis acceleration	McManus et al [37] (2022)
RMS acceleration AP	RMS of COM acceleration in AP direction (m/s^2)	De Groote et al [33] (2021); McManus et al [37] (2022)
RMS acceleration ML	RMS of COM acceleration in ML direction (m/s^2)	De Groote et al [33] (2021); McManus et al [37] (2022)
Mean acceleration AP	Average absolute value of COM acceleration in AP directions (m/s^2)	De Groote et al [33] (2021)
Mean acceleration ML	Average absolute value of COM acceleration in ML direction (m/s^2)	De Groote et al [33] (2021)
Mean Z-axis acceleration at stand-start	The average Z-axis (vertical) accelerometer value recorded at the detected stand-start time across repetitions (m/s^2)	Greene et al [34] (2022)
Mean Z-axis acceleration at sit-end	The average Z-axis (vertical) accelerometer value recorded at the detected sit-end time across repetitions (m/s^2)	Greene et al [34] (2022)
Postural stability	The negative natural logarithm of the mean of the absolute acceleration along mediolateral, anterior–posterior, and vertical axes resultant vector ($-\ln[m/s^2]$)	Olsen et al [43] (2023)
Postural stability ML	The negative natural logarithm of the absolute acceleration along ML axis resultant vector ($-\ln[m/s^2]$)	Olsen et al [43] (2023)
Postural stability AP	The negative natural logarithm of the absolute acceleration along AP axis resultant vector ($-\ln[m/s^2]$)	Olsen et al [43] (2023)
Angular velocity		
Spectral entropy angular velocity	The spectral entropy of the angular-velocity signal	Greene et al [34] (2022)
Position		
Center of gravity displacement	The average displacement of the center of gravity (cm)	Chang et al [42] (2013)
2D path length	The total length of the acceleration trajectories (mm)	Zhou et al [39] (2021)
COP[f] displacement	The distance traveled by the COP (cm)	Scaglioni-Solano and Aragón-Vargas [41] (2013)
COP sway RMS amplitude AP	RMS amplitude of COP sway in AP direction (cm)	da Silva et al [32] (2013)
COP sway RMS amplitude ML	RMS amplitude of COP sway in ML direction (cm)	da Silva et al [32] (2013)
Mean velocity AP	Mean velocity of COP in AP direction (cm/s)	da Silva et al [32] (2013)
Mean velocity ML	Mean velocity of COP in ML direction (cm/s)	da Silva et al [32] (2013)
COP area	95% confidence ellipse area of COP (cm^2)	da Silva et al [32] (2013)
Limits of stability average end point excursion	The average ratio of the initial distance traveled by the center of gravity on the primary attempt to reach each of the 8 multidirectional visual targets (%)	Harro and Garascia [40] (2019)

XSL•FO
RenderX

Category and variable	Definition (unit)	Study
Sensory organization test composite equilibrium	The average equilibrium score (a measure of anterior/posterior sway during each trial compared to a theoretical stability limit of 12.5°) from conditions with fixed surface and normal vision, and fixed surface with eyes closed, added to the equilibrium scores from conditions with fixed surface and sway referenced vision, sway-referenced surface and normal vision, sway-referenced surface and EC, and sway-referenced surface and sway-referenced vision, then divided by the total number of trials	Harro and Garascia [40] (2019)
Sensory organization test vestibular ratio scores	The ratio of the equilibrium score in the sway-referenced surface, eyes closed condition to the fixed surface, normal vision condition	Harro and Garascia [40] (2019)
Energy		
Mean frequency AP	Mean frequency of COP in AP direction (Hz)	da Silva et al [32] (2013)
Mean frequency ML	Mean frequency of COP in ML direction (Hz)	da Silva et al [32] (2013)
COP sway AP	15.5 multiplied by the result of subtracting the sum of the force sensor data from the bottom left and bottom right corners from the sum of the force sensor data from the top left and top right corners and then dividing that result by the total sum of force sensor data from all four corners	Levy et al [36] (2018)
COP sway ML	24.25 multiplied by the result of subtracting the sum of the force sensor data from the top left and bottom left corners from the sum of the force sensor data from the top right and bottom right corners and then dividing that result by the total sum of force sensor data from all four corners	Levy et al [36] (2018)

[a]RMS: root mean square.

[b]AP: anteroposterior.

[c]COM: center of mass.

[d]ML: mediolateral.

[e]SI: superior-inferior.

[f]COP: center of pressure.

The device-derived variables extracted in the reviewed studies varied and were largely determined by the employed technology and the specific balance task. Five of the 7 studies utilizing smartphones or IMUs reported linear acceleration–based variables, with 4 studies [33,37,39,44] quantifying root mean square (RMS) acceleration. Among the 5 studies using force plates or balance boards, 4 reported position-based variables, primarily center of pressure (COP) metrics [32,40-42]. Additionally, 5 of the 6 studies involving dynamic tasks extracted temporal and spatial variables. Notably, 2 studies integrated multiple device-derived variables into composite balance scores [34,37].

Data Processing

Eight studies reported complete datasets and therefore did not require missing data handling, whereas 5 studies did not report the occurrence or management of missing data. Data denoising methods were described in 9 studies and primarily involved signal-filtering techniques, including low-pass Butterworth, Savitzky-Golay, and dual-pass filters.

Reliability

Table S4 in Multimedia Appendix 1 displays the findings on reliability and validity statistics. All studies investigated the test-retest reliability of balance assessment using the intraclass correlation coefficient (ICC). The reported ICC values varied considerably, ranging from 0.016 to 0.99, depending on the device types, tasks, or variable of choice. Chang et al [42] reported the highest test-retest reliability for the average displacement of gravity derived from a balance board during stance with eyes closed. Measurement error was reported in 6 studies, with standard error of measurement (SEM) values ranging from 0.02 to 16.1 and minimal detectable change (MDC) values ranging from 0.3 to 44.6.

When examined by device type and task setting, smartphones reported in 5 studies demonstrated ICCs from 0.016 to 0.97, with higher values in static tasks (0.50 - 0.97) [33,39,43,44] and wider variability during dynamic tasks (0.016 - 0.96) [35,43,44]. IMUs showed ICC values from 0.30 to 0.95, with more consistent reliability in static tasks (0.83 - 0.95) [37] and lower, broader values in dynamic tasks (0.30 - 0.91) [34]. Across both smartphones and IMUs, static tasks produced better

XSL•FO
RenderX

and more consistent reliability than dynamic tasks. Force plates reported ICCs from 0.40 to 0.85 for static tasks [32,36] and 0.710 to 0.898 for dynamic tasks [40]. Balance boards were used exclusively for static assessments and yielded ICCs from 0.64 to 0.99 [41,42], whereas infrared depth sensors were applied only for dynamic assessments, with ICCs between 0.64 and 0.96 [38]. Regarding variable choice, temporal parameters showed ICCs from 0.30 to 0.91 [34,40,43], spatial parameters from 0.016 to 0.96 [35,38,43], linear acceleration metrics from 0.50 to 0.95 [33,34,37,39,43,44], angular velocity metrics from 0.67 to 0.81 [34], position-related metrics from 0.40 to 0.99 [32,39-42], and energy parameters from 0.72 to 0.83 [32,36]. As the reported SEM values corresponded to heterogeneous outcome variables with different units and scales, we did not perform comparisons across studies.

Validity

Of the 7 studies that investigated criterion validity, 4 used a force plate as the gold standard [33,36,39,41] and 3 used 3D motion capture [35,43,44]. In smartphone-based assessments compared with motion capture, spatial variables measured during dynamic tasks showed substantial variability, with correlations ranging from −0.51 to 0.98 [43]. In contrast, linear acceleration variables demonstrated consistently excellent correlations in static tasks, with coefficients of 0.91 to 0.98 [43,44], whereas dynamic tasks yielded a low, nonsignificant correlation of 0.37 [44]. When compared with force plates, smartphone-derived linear acceleration measures showed correlations of 0.14 to 0.82 in static tasks [33] and 0.38 to 0.60 in dynamic tasks [39]. Energy variables derived from force plates yielded Pearson correlations of 0.82 to 0.89 [36], and position variables from balance boards produced regression coefficients between 0.92 and 1.05 across static tasks [41].

Four studies [32,37,40,42] assessed convergent validity. Position variables derived from balance boards in static tasks showed significant positive correlations with smart balance master systems, ranging from 0.58 to 0.86 [42]. IMU-derived linear acceleration variables from static tasks had significant, fair correlations of 0.30 to 0.34 with Time Up and Go time [37]. In contrast, da Silva et al [32], using static tasks, and Harro et al [40], using dynamic tasks, reported generally low and nonsignificant correlations between position, energy, and temporal variables from force plates and clinical testing.

Discriminative validity was reported in 5 studies [32,34,37,38]. Two studies [32,42] compared young and older adults, and 2 [37,38] compared individuals with normal versus impaired balance, all demonstrating statistically significant between-group differences across multiple parameters, including position parameters from balance boards and force plates, and linear acceleration parameters from IMUs in static tasks, and spatial parameters from infrared depth sensors in dynamic tasks. One study [34] developed classification models using IMU data during dynamic tasks on the sternum and thigh, achieving 76.76% and 81.69% accuracy, respectively.

Muscle Power Assessment

One study assessed muscle power using a smartphone during a sit-to-stand task [31]. The derived variables, including total movement duration, peak force, rate of force development, and peak power, were measured. Complete datasets were obtained, and the raw signals were filtered before analysis. Test-retest reliability of these variables ranged from 0.43 to 0.92. Measurement error, expressed as SEM, varied between 3.1 and 26.1. Convergent validity, evaluated by comparing smartphone measures with force plate data, showed a correlation coefficient ranging from 0.69 to 0.98, with peak power demonstrating a statistically significant correlation.

Quality Assessment

Methodological Quality Assessment

The methodological quality and evaluation results of measurement properties are summarized in Table 4. For test-retest reliability, 9 of 14 studies reported “doubtful,” primarily because repeated assessments were conducted without familiar tests or adequate rest time, which may introduce variations and affect participant stability on the construct to be measured. For measurement error, 3 of 6 studies received a “doubtful” rating due to the absence of familiarization procedures, while the remaining studies were rated as very good.

XSL•FO
RenderX

Table . Methodological quality of studies and evaluation results of measurement properties based on COSMIN.

Study	Test-retest reliability		Measurement error		Criterion validity		Convergent validity		Discriminative validity	
	MQ[a]	ER[b]	MQ	ER	MQ	ER	MQ	ER	MQ	ER
Balance										
Chang et al [42] (2013)	D[c]	+[d]	NR[e]	NR	NR	NR	V[f]	+	D	+
da Silva et al [32] (2013)	V	+	V	?[g]	NR	NR	V	+	V	+
De Groote et al [33] (2021)	V	+	NR	NR	V	+	NR	NR	NR	NR
Greene et al [34] (2022)	D	+	NR	NR	NR	NR	NR	NR	D	+
Harro and Garascia [40] (2019)	V	+	V	?	NR	NR	I[h]	+	NR	NR
Kuntapun et al [35] (2020)	D	+	NR	NR	V	+	NR	NR	NR	NR
Levy et al [36] (2018)	D	+	D	?	V	+	NR	NR	NR	NR
McManus et al [37] (2022)	D	+	NR	NR	NR	NR	V	+	D	+
Okada et al [38] (2024)	D	+	NR	NR	NR	NR	NR	NR	V	+
Olsen et al [43] (2023)	D	+	D	?	V	+	NR	NR	NR	NR
Pornwatanakul and Siriphorn [44] (2023)	D	+	D	?	V	+	NR	NR	NR	NR
Scaglioni-Solano and Aragón-Vargas [41] (2013)	V	+	V	?	V	+	NR	NR	NR	NR
Zhou et al [39] (2021)	D	+	NR	NR	V	–[i]	NR	NR	NR	NR
Muscle power										
Cerrito et al [31] (2015)	V	+	V	?	NR	NR	I	+	NR	NR

[a]MQ: methodological quality.
[b]ER: evaluation results.
[c]D: doubtful.
[d]+: sufficient.
[e]NR: not reported.
[f]V: very good.
[g]?: indeterminate.
[h]I: inadequate.
[i]–: insufficient.

XSL•FO
RenderX

Four studies evaluated criterion validity, and all were assessed as "very good." Convergent validity was assessed in 5 studies, with 2 studies poorly describing the construct and measurement properties of the comparator instrument, resulting in an "inadequate" methodological quality rating. Of the 5 studies involving discriminative validity, 3 were rated "doubtful" owing to insufficient description of important subgroup characteristics.

Quality of Measurement Properties

As shown in Table 4, test-retest reliability, convergent validity, and discriminative validity were rated as sufficient (+) in 14, 5, and 5 studies, respectively. Measurement error was rated as indeterminate (?) in all 6 studies, as none defined a minimal important change (MIC). Criterion validity was rated as insufficient (−) in 1 study due to a correlation coefficient below 0.70 with the gold standard.

Discussion

Overview

This systematic review summarized digital technologies and biomarkers for assessing locomotor capacity in general older adults, including their validity and reliability. Fourteen studies were included, identifying 5 types of digital technologies and 51 digital biomarkers (47 for balance and 4 for muscle power). Twelve studies reported generally acceptable test-retest reliability and validity for balance, indicating promising potential. One study evaluated muscle power and demonstrated good reliability and validity, though further research is needed to validate these findings. However, caution is warranted when interpreting these results, as reliability and validity vary by context, and the overall strength of evidence remains limited. Future research should address other locomotor components and focus on developing assessments that simultaneously ensure suitability, reliability, and validity [50,51].

Reliability

The test-retest reliability reported across included studies was acceptable, suggesting that digital technologies can produce stable measurements of locomotor capacity among older adults. Nevertheless, there was considerable variability in reported reliability coefficients, ranging from 0.016 to 0.97. This broad range highlights the nascent stage of the field and could be attributed to differences in device modalities, task settings, and the specific digital biomarkers extracted. Specifically, force plates generally exhibited more consistent test-retest reliability than smartphone- and IMU-based systems, which is commonly attributed to their direct measurement of ground reaction forces and COP trajectories and the absence of variability introduced by wearable sensor placement and soft-tissue artifacts [52,53]. Consistent with previous literature [54], portable sensors yielded higher and more consistent reliability for static tasks than for dynamic tasks [43,44], likely because dynamic conditions introduce increased movement variability, sensor micro-slippage, and more complex signal processing requirements. Variable choice also markedly influenced reliability. For example, in a wearable sensor-based sit-to-stand assessment, linear acceleration and angular velocity variables demonstrated higher test-retest reliability than temporal variables [34], likely because they are directly derived from continuous sensor signals, whereas temporal measures rely on event detection. Overall, these findings underscore the need for standardized sensor fixation and cautious use of integrated spatial metrics when employing mobile devices and depth sensors in locomotor assessment of older adults.

Only about half of the studies provided measurement error, such as SEM and MDC, which limited the assessment of whether observed differences reflect true change or fall within the bounds of random measurement error. Among those that reported measurement error, none defined a MIC, which prevents evaluation of whether the reported MDC values correspond to clinically meaningful change [55]. Therefore, the rating results of measurement error for relevant studies remained indeterminate. Future studies should systematically report SEM, MDC, and MIC to enable robust evaluation of measurement performance and to support meaningful clinical interpretation.

Validity

Most studies showed acceptable validity, supporting the use of digital technologies to evaluate locomotor capacity. Regarding criterion validity, smartphone-derived linear acceleration measures showed strong correlations with 3D motion capture systems in static tasks but substantially weaker associations during dynamic conditions [44], highlighting the increased difficulty of accurately capturing complex movement patterns with smartphone sensors. In the studies comparing digital biomarkers with force plates, the correlation coefficients of linear acceleration variables derived from smartphones were generally lower than those of the energy variables from force plates and position variables from balance boards [33,36,41]. This could be attributed to inherent differences in the constructs measured; specifically, smartphones leveraged acceleration signals, whereas force plates and balance boards recorded ground reaction forces and COP.

As for convergent validity, the correlation between IMUs and Timed Up and Go tests was significant but low. This weak correlation may be because Timed Up and Go tests mainly assess rougher motor ability, whereas digital tools can measure balance dynamics and subtle changes more precisely [56]. Furthermore, the correlation between the force plate and clinical tests was generally low and nonsignificant in most cases. Clinical tests can measure functional balance abilities, whereas force plates can directly analyze balance related to reactive postural strategies and sensory integration [40,42].

The results of discriminative validity indicated that device-based measures were valid in differentiating balance performance between younger and older participants, which was consistent with a previous systematic review [57]. This not only reinforces the well-established physiological understanding of age-related declines in balance control, but more importantly, it positions digital biomarkers as sensitive tools for objectively quantifying these subtle yet significant changes associated with the aging process. The devices also distinguished between groups with impaired and normal balance. However, the criteria used to define balance groups varied across studies. This heterogeneity makes it difficult to compare the discriminative ability of different device-based measures.

XSL•FO
RenderX

Digital Biomarkers

In this systematic review, we identified several digital biomarkers that demonstrated excellent reliability and validity for assessing locomotor capacity in older adults. Aligning with previous studies [58,59], COP parameters, including COP displacement, mean frequency anteroposterior (AP), COP sway AP, and COP sway mediolateral (ML), measured from balance boards and force plates during static standing tasks, exhibited strong reliability and validity for assessing balance. Furthermore, RMS acceleration, derived from smartphones, was also identified as a reliable and valid indicator of static balance. It has been frequently utilized in portable sensor-based balance assessments [33,37,39]. Beyond static balance measures, dynamic balance indicators such as walking speed, mean step time, and mean step length showed great reliability and validity for assessing dynamic balance. These gait-related biomarkers were also used to enhance the discrimination of gait disturbances and disease severity in individuals with Parkinson disease [56]. Additionally, total movement distance and step number in the 20-second stepping test measured by infrared depth sensors exhibited strong test-retest reliability and differentially discriminated between high- and low-balance subgroups. These 2 variables appeared to contribute to the differentiation of disordered balance and to suggest directions for future research.

Among studies focused on muscle power, Cerrito et al [31] demonstrated that smartphone-based peak power during sit-to-stand was a robust digital biomarker for evaluating lower-body muscle power in older adults. The use of peak power as a valid indicator of muscle power was further supported by studies employing linear transducers [60]. Furthermore, sit-to-stand muscle power was found to be a stronger predictor of falls, fractures, and physical independence than grip strength [61,62].

Integrating multiple variables can address multidimensional deficits that single measures miss [63] and has shown utility in frailty classification [64]. There is a growing trend toward constructing composite indices for balance impairment. Although 2 studies [34,37] integrating IMU-derived variables reported encouraging discriminative performance, these indices remain exploratory given the absence of rigorous validation and external replication. To establish their clinical applicability in balance assessment, future studies should focus on systematic performance evaluation and external validation.

Quality of Evidence

Eleven studies were rated as having doubtful or inadequate methodological quality, leading to inflated or underestimated reliability and validity. Specifically, many studies lacked details about the familiarization process or the rest periods between repeated measurements. These omissions may introduce variation due to participant fatigue, learning effects, or other extraneous factors.

In addition, only 3 included studies reported a sample size greater than 50, and none used a sample size calculation. Small sample sizes may inflate the impact of individual variability, reduce statistical power, and result in unstable estimates of reliability and validity. In heterogeneous older adult populations, insufficient sample sizes further limit the generalizability of findings and increase uncertainty in measurement performance. Future research should prioritize adequately powered studies with predefined sample size considerations tailored to the intended measurement properties.

Only De Groote et al [33] met both good methodological quality and an adequate sample size, but the selected indicators, including RMS acceleration ML, RMS acceleration AP, mean acceleration ML, and mean acceleration AP, did not demonstrate satisfactory reliability and validity. Although many studies have reported good reliability and validity for other parameters, the methodological flaws in these studies necessitate caution when interpreting these findings.

The Gap Between Research and Clinical Application

The digital transformation of healthcare in recent years has significantly advanced digital technologies, facilitating the transition of digital biomarkers from research contexts to clinical settings. Nevertheless, incorporating digital technology into healthcare to assess locomotor capacity in older adults comes with its own set of challenges.

One major challenge lies in balancing cost-effectiveness and measurement accuracy when selecting appropriate technologies to assess locomotor capacity. Force plates that measure postural sway by calculating COP are considered the gold standard for assessing balance [65]. However, their high cost and limited portability restrict their widespread clinical application [36]. More affordable and available alternatives, such as smartphones, IMUs, balance boards, and infrared depth sensors, have been suggested to demonstrate acceptable reliability and validity for assessing balance and locomotor performance [66]. In addition, emerging technologies are providing new perspectives for monitoring locomotor function. For example, Teel et al [67] explored the VR-based balance module for detecting persistent balance deficits in clinical concussion care, reporting a sensitivity of 85.7% and a specificity of 87.8%. As the field advances, hybrid technology integration may enable the capture of different aspects of locomotor function, thereby offering a more nuanced and holistic understanding of an individual's physical health [61]. Manupibul et al [68] demonstrated that the integration of force-sensitive resistance sensors and IMUs provided complementary data, high accuracy, and enhanced performance to estimate the gait parameters.

The successful clinical translation of digital technologies for locomotor capacity requires careful consideration of user experience and usability, particularly ease of use among older adults with digital exclusion [69,70]. Tools or systems with complex interfaces or that require frequent manual adjustments may pose substantial barriers to adoption, compliance, and long-term use in older populations. Consistent findings from qualitative studies highlight ease of use as a key facilitator of technology adoption in this population [71,72]. Current research primarily involved participants performing specific, structured tasks under controlled conditions, resulting in potential missing data [33,39]. To achieve clinical effectiveness, future developments in locomotor capacity assessment tools should emphasize unobtrusive sensing, voice-based guidance, and minimal user interaction [73].

XSL•FO
RenderX

Context-specific performance also represents a key challenge for clinical application. Most of the studies included in this review collected data during the performance of predefined tasks under controlled conditions [35,40,44]. While certain digital biomarkers may demonstrate strong reliability and validity in specific tasks, their performance may vary across different tasks or settings. For example, RMS acceleration derived from smartphones in static tasks exhibited excellent criterion validity, but its performance deteriorated when applied to dynamic tasks [44]. Its performance can also be influenced by factors such as sensor placement, sampling frequency, and configured range [74]. Furthermore, the lack of standardized definitions and measurement protocols for these biomarkers complicates their reproducibility and generalization across studies.

To facilitate clinical application, interpretability represents an important consideration for future research. Interpretability can be considered from two perspectives: the clinical meaningfulness of digital biomarkers and the transparency of algorithms. The interpretability of digital biomarkers depends on how they reflect locomotor capacity, including the establishment of meaningful cut-off points that correspond to declines or improvements in mobility. Minimal clinically important difference is defined as the smallest change in a clinical outcome measure that is considered beneficial and meaningful to patients. By establishing minimal clinically important difference thresholds, clinicians can determine whether changes in digital biomarkers are clinically significant, thereby supporting decision-making [75]. With the integration of digital biomarkers, artificial intelligence-based algorithms may be employed, raising concerns about “black-box” systems, and clinical implementation should prioritize explainable artificial intelligence approaches to improve healthcare professionals’ and patients’ understanding of the measurement mechanism [76].

Limitations

This systematic review offers a comprehensive overview of digital technologies for assessing locomotor capacity in older adults, providing integrated insights into digital health and physical function assessment. This systematic review has several limitations to note. First, we included only articles published in English, which may introduce publication bias. Second, although our search strategy was intentionally broad, the included studies predominantly focused on balance assessment, with other key components of locomotor capacity remaining underrepresented, reflecting a gap in the existing literature on other locomotor capacity components. Third, the included studies exhibited considerable heterogeneity, with variable study quality and assessment protocols. This variability is due to the emerging and expanding nature of this research field, which has yet to establish optimal assessment protocols and highlights a valuable strength of this review. Fourth, this systematic review predominantly focused on general older adults and excluded populations with specific diseases, such as Parkinson disease. The scope of our findings may be limited, and the generalizability to populations with specific diseases requires further investigation. Finally, this review may not have captured all emerging technologies, such as virtual reality and artificial intelligence, which are rapidly evolving and hold considerable potential for locomotor capacity assessment.

Implications

Despite these limitations, the sufficient test-retest reliability and validity of the device-based assessment support its potential use in examining older adults’ balance and muscle power. Several recommendations need to be formulated. First, future studies should enhance methodological quality and adhere to COSMIN terminologies and recommendations [27,77]. Second, researchers should thoroughly evaluate both reliability and validity and compare the relevant measurement properties for various digital tools to facilitate the development of the optimal system. Third, alongside rigorous validation of measurement properties, future research should prioritize developing user-friendly digital tools. Current studies often employ task-specific protocols, limiting their widespread adoption in clinical practice and community settings. Fourth, variability in reliability and validity results indicated that no single device or method can be universally recommended for locomotor capacity assessment at this stage. Further studies need to standardize protocols and assess the robustness of these technologies in diverse real-world settings. Lastly, given the current diversity of technologies, the studies included in this review do not cover all existing tools, such as virtual reality and artificial intelligence applications. Future research is expected to explore the integration of these emerging technologies for a more precise and fine-grained assessment of locomotor capacity.

Conclusion

This review offers an integrated understanding of the potential and limitations of digital technologies for assessing locomotor capacity in older adults, summarizing 51 digital biomarkers with generally acceptable reliability and validity. Unlike previous studies that predominantly targeted specific sensor types or disease-specific populations, our review incorporates diverse technologies and biomarkers across general older populations, offering a novel perspective on the potential for scalable, objective, and remote monitoring of changes in locomotor capacity related to aging. Our findings pave the way for developing digital biomarkers that enable the early detection of subtle declines in locomotor capacity, which is critical for timely intervention. However, the current evidence reveals several limitations, particularly task dependency and limited investigation of key locomotor components. The methodological quality of the included studies was generally doubtful, limiting the immediate clinical applicability of digital assessments. Future studies should expand the evaluation to a broader range of locomotor components and improve methodological quality with predefined sample size calculations. Furthermore, age-friendly assessments with enhanced clinical interpretability should be explored to facilitate clinical implementation. By addressing these challenges, digital biomarkers could become essential tools for enhancing the monitoring and management of older adults’ locomotor capacity, contributing to healthy aging strategies.

XSL•FO
RenderX

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) to improve grammar and spelling, and to rephrase sentences for enhanced clarity. According to GAIDeT (Generative AI Delegation Taxonomy; 2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing. The GAI tool used was ChatGPT-5.2. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This systematic review was funded by the National Key Research & Development Program of China (grant 2023YFC3604602, 2023YFC3605204), Fundamental Research Funds for the Central Universities of Central South University (grant 2025ZZTS0171), and the Hunan Provincial Innovation Foundation for Postgraduate (grant CX20250407). The funder had no involvement in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Data Availability

Data sharing was not applicable to this systematic review as no datasets were generated or analyzed.

Authors' Contributions

Conceptualization: SZ, JH
Data curation: SZ, JH, CZ, YZ
Methodology: SZ, JH
Supervision: MH, HF
Writing – original draft: SZ, JH, JY, XL
Writing – review & editing: MH, HF
All authors read and approved the final manuscript.
HF and MH contributed equally to this work and share the co-corresponding authorship.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search terms, search strategies, data extraction form, and description of reliability and validity in the included studies.
[DOCX File, 45 KB - jmir_v28i1e83814_app1.docx]

Checklist 1
PRISMA-COSMIN checklist.
[DOCX File, 28 KB - jmir_v28i1e83814_app2.docx]

References

1. Wang S, Yang Z, Tan X, Lai F, Luo L, Ding Y. Association between standing height and physical disability among U.S. adults aged 60 years and older: findings from NHANES 2015-2018. BMC Geriatr 2024 Jun 18;24(1):529. [doi: 10.1186/s12877-024-05100-3] [Medline: 38890578]
2. Braun T, Thiel C, Peter RS, et al. Association of clinical outcome assessments of mobility capacity and incident disability in community-dwelling older adults - a systematic review and meta-analysis. Ageing Res Rev 2022 Nov;81:101704. [doi: 10.1016/j.arr.2022.101704] [Medline: 35931411]
3. Veronese N, Honvo G, Amuthavalli Thiyagarajan J, et al. Attributes and definitions of locomotor capacity in older people: a World Health Organisation (WHO) locomotor capacity working group meeting report. Aging Clin Exp Res 2022 Mar;34(3):481-483. [doi: 10.1007/s40520-022-02080-5] [Medline: 35133612]
4. Wang G, Zhou Y, Zhang L, et al. Prevalence and incidence of mobility limitation in Chinese older adults: evidence from the China Health and Retirement Longitudinal Study. J Nutr Health Aging 2024 Mar;28(3):100038. [doi: 10.1016/j.jnha.2024.100038] [Medline: 38280833]
5. de Oliveira VP, Ferriolli E, Lourenço RA, González-Bautista E, de Souto Barreto P, de Mello RGB. The sensitivity and specificity of the WHO's ICOPE screening tool, and the prevalence of loss of intrinsic capacity in older adults: a scoping review. Maturitas 2023 Nov;177:107818. [doi: 10.1016/j.maturitas.2023.107818] [Medline: 37542782]
6. López-Ortiz S, Lista S, Peñín-Grandes S, et al. Defining and assessing intrinsic capacity in older people: a systematic review and a proposed scoring system. Ageing Res Rev 2022 Aug;79(101640):101640. [doi: 10.1016/j.arr.2022.101640] [Medline: 35569785]

XSL•FO
RenderX

7. Jung HW, Roh H, Cho Y, et al. Validation of a multi-sensor-based kiosk for short physical performance battery. J Am Geriatr Soc 2019 Dec;67(12):2605-2609. [doi: 10.1111/jgs.16135] [Medline: 31441514]
8. Bergquist R, Weber M, Schwenk M, et al. Performance-based clinical tests of balance and muscle strength used in young seniors: a systematic literature review. BMC Geriatr 2019 Jan 9;19(1):9. [doi: 10.1186/s12877-018-1011-0] [Medline: 30626340]
9. Piau A, Steinmeyer Z, Cesari M, et al. Intrinsic capacitiy monitoring by digital biomarkers in integrated care for older people (ICOPE). J Frailty Aging 2021;10(2):132-138. [doi: 10.14283/jfa.2020.51] [Medline: 33575701]
10. Zhu Y, Li H, Wu X, Chen N. Accuracy validation of a sensor-based inertial measurement unit and motion capture system for assessment of lower limb muscle strength in older adults-a novel and convenient measurement approach. Sensors (Basel) 2024 Sep 18;24(18):6040. [doi: 10.3390/s24186040] [Medline: 39338786]
11. Estévez-Pedraza Á, Parra-Rodríguez L, Martínez-Méndez R, Portillo-Rodríguez O, Ronzón-Hernández Z. A novel model to quantify balance alterations in older adults based on the center of pressure (CoP) measurements with a cross-sectional study. PLoS ONE 2021;16(8):e0256129. [doi: 10.1371/journal.pone.0256129] [Medline: 34398918]
12. Piau A, Wild K, Mattek N, Kaye J. Current state of digital biomarker technologies for real-life, home-based monitoring of cognitive function for mild cognitive impairment to mild Alzheimer disease and implications for clinical care: systematic review. J Med Internet Res 2019 Aug 30;21(8):e12785. [doi: 10.2196/12785] [Medline: 31471958]
13. Pradeep Kumar D, Toosizadeh N, Mohler J, Ehsani H, Mannier C, Laksari K. Sensor-based characterization of daily walking: a new paradigm in pre-frailty/frailty assessment. BMC Geriatr 2020 May 6;20(1):164. [doi: 10.1186/s12877-020-01572-1] [Medline: 32375700]
14. Huang J, Zhou S, Xie Q, Yu J, Zhao Y, Feng H. Digital biomarkers for real-life, home-based monitoring of frailty: a systematic review and meta-analysis. Age Ageing 2025 Mar 28;54(4):afaf108. [doi: 10.1093/ageing/afaf108] [Medline: 40251836]
15. Sun YM, Wang ZY, Liang YY, Hao CW, Shi CH. Digital biomarkers for precision diagnosis and monitoring in Parkinson's disease. NPJ Digit Med 2024 Aug 21;7(1):218. [doi: 10.1038/s41746-024-01217-2] [Medline: 39169258]
16. Burq M, Rainaldi E, Ho KC, et al. Virtual exam for Parkinson's disease enables frequent and reliable remote measurements of motor function. NPJ Digit Med 2022 May 23;5(1):65. [doi: 10.1038/s41746-022-00607-8] [Medline: 35606508]
17. Hayek R, Gutman I, Baranes G, Nudelman Y, Springer S. Smartphone-based sit-to-stand analysis for mobility assessment in middle age. Innov Aging 2024;8(10):igae079. [doi: 10.1093/geroni/igae079] [Medline: 39391811]
18. Polvorinos-Fernández C, Sigcha L, Borzì L, et al. Evaluating motor symptoms in Parkinson's disease through wearable sensors: a systematic review of digital biomarkers. Appl Sci 2024;14(22):10189. [doi: 10.3390/app142210189]
19. Woelfle T, Bourguignon L, Lorscheider J, Kappos L, Naegelin Y, Jutzeler CR. Wearable sensor technologies to assess motor functions in people with multiple sclerosis: systematic scoping review and perspective. J Med Internet Res 2023 Jul 27;25:e44428. [doi: 10.2196/44428] [Medline: 37498655]
20. Ghislieri M, Gastaldi L, Pastorelli S, Tadano S, Agostini V. Wearable inertial sensors to assess standing balance: a systematic review. Sensors (Basel) 2019 Sep 20;19(19):4075. [doi: 10.3390/s19194075] [Medline: 31547181]
21. Lipsmeier F, Taylor KI, Kilchenmann T, et al. Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. Mov Disord 2018 Aug;33(8):1287-1297. [doi: 10.1002/mds.27376] [Medline: 29701258]
22. Elsman EBM, Mokkink LB, Terwee CB, et al. Guideline for reporting systematic reviews of outcome measurement instruments (OMIs): PRISMA-COSMIN for OMIs 2024. J Clin Epidemiol 2024 Sep;173:111422. [doi: 10.1016/j.jclinepi.2024.111422] [Medline: 38849061]
23. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
24. Honvo G, Sabico S, Veronese N, et al. Measures of attributes of locomotor capacity in older people: a systematic literature review following the COSMIN methodology. Age Ageing 2023 Oct 28;52(Suppl 4):iv44-iv66. [doi: 10.1093/ageing/afad139] [Medline: 37902521]
25. Mohsen F, Al-Absi HRH, Yousri NA, El Hajj N, Shah Z. A scoping review of artificial intelligence-based methods for diabetes risk prediction. NPJ Digit Med 2023 Oct 25;6(1):197. [doi: 10.1038/s41746-023-00933-5] [Medline: 37880301]
26. Johnston W, O'Reilly M, Argent R, Caulfield B. Reliability, validity and utility of inertial sensor systems for postural control assessment in sport science and medicine applications: a systematic review. Sports Med 2019 May;49(5):783-818. [doi: 10.1007/s40279-019-01095-9] [Medline: 30903440]
27. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010 Jul;63(7):737-745. [doi: 10.1016/j.jclinepi.2010.02.006] [Medline: 20494804]
28. Prinsen CAC, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res 2018 May;27(5):1147-1157. [doi: 10.1007/s11136-018-1798-3] [Medline: 29435801]
29. Mokkink LB, de Vet HCW, Prinsen CAC, et al. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. Qual Life Res 2018 May;27(5):1171-1179. [doi: 10.1007/s11136-017-1765-4] [Medline: 29260445]

XSL•FO
RenderX

30. Mokkink LB, Boers M, van der Vleuten CPM, et al. COSMIN risk of bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. BMC Med Res Methodol 2020 Dec 3;20(1):293. [doi: 10.1186/s12874-020-01179-5] [Medline: 33267819]
31. Cerrito A, Bichsel L, Radlinger L, Schmid S. Reliability and validity of a smartphone-based application for the quantification of the sit-to-stand movement in healthy seniors. Gait Posture 2015 Feb;41(2):409-413. [doi: 10.1016/j.gaitpost.2014.11.001] [Medline: 25467428]
32. da Silva RA, Bilodeau M, Parreira RB, Teixeira DC, Amorim CF. Age-related differences in time-limit performance and force platform-based balance measures during one-leg stance. J Electromyogr Kinesiol 2013 Jun;23(3):634-639. [doi: 10.1016/j.jelekin.2013.01.008] [Medline: 23403137]
33. De Groote F, Vandevyvere S, Vanhevel F, Orban de Xivry JJ. Validation of a smartphone embedded inertial measurement unit for measuring postural stability in older adults. Gait Posture 2021 Feb;84:17-23. [doi: 10.1016/j.gaitpost.2020.11.017] [Medline: 33260077]
34. Greene BR, Doheny EP, McManus K, Caulfield B. Estimating balance, cognitive function, and falls risk using wearable sensors and the sit-to-stand test. Wearable Technol 2022;3:e9. [doi: 10.1017/wtc.2022.6] [Medline: 38486905]
35. Kuntapun J, Silsupadol P, Kamnardsiri T, Lugade V. Smartphone monitoring of gait and balance during irregular surface walking and obstacle crossing. Front Sports Act Living 2020;2:560577. [doi: 10.3389/fspor.2020.560577] [Medline: 33345119]
36. Levy SS, Thralls KJ, Kviatkovsky SA. Validity and reliability of a portable balance tracking system, BTrackS, in older adults. J Geriatr Phys Ther 2018;41(2):102-107. [doi: 10.1519/JPT.0000000000000111] [Medline: 27893566]
37. McManus K, Greene BR, Ader LGM, Caulfield B. Development of data-driven metrics for balance impairment and fall risk assessment in older adults. IEEE Trans Biomed Eng 2022 Jul;69(7):2324-2332. [doi: 10.1109/TBME.2022.3142617] [Medline: 35025734]
38. Okada S, Takeshima N, Fujita E, Kohama T, Kusunoki M, Brechue WF. The stepping test, and infrared depth sensor, provide reliable measures of balance in community-dwelling older adults. J Phys Ther Sci 2024 Jan;36(1):9-20. [doi: 10.1589/jpts.36.9] [Medline: 38186969]
39. Zhou J, Jiang X, Yu W, et al. A smartphone app-based application enabling remote assessments of standing balance during the COVID-19 pandemic and beyond. IEEE Internet Things J 2021;8(21):15818-15828. [doi: 10.1109/JIOT.2021.3064442]
40. Harro CC, Garascia C. Reliability and validity of computerized force platform measures of balance function in healthy older adults. J Geriatr Phys Ther 2019;42(3):E57-E66. [doi: 10.1519/JPT.0000000000000175] [Medline: 29324510]
41. Scaglioni-Solano P, Aragón-Vargas LF. Validity and reliability of the Nintendo Wii Balance Board to assess standing balance and sensory integration in highly functional older adults. Int J Rehabil Res 2014 Jun;37(2):138-143. [doi: 10.1097/MRR.0000000000000046] [Medline: 24445863]
42. Chang WD, Chang WY, Lee CL, Feng CY. Validity and reliability of Wii Fit Balance Board for the assessment of balance of healthy young adults and the elderly. J Phys Ther Sci 2013 Oct;25(10):1251-1253. [doi: 10.1589/jpts.25.1251] [Medline: 24259769]
43. Olsen S, Rashid U, Allerby C, et al. Smartphone-based gait and balance accelerometry is sensitive to age and correlates with clinical and kinematic data. Gait Posture 2023 Feb;100:57-64. [doi: 10.1016/j.gaitpost.2022.11.014] [Medline: 36481647]
44. Pooranawatthanakul K, Siriphorn A. Testing the validity and reliability of a new Android application-based accelerometer balance assessment tool for community-dwelling older adults. Gait Posture 2023 Jul;104:103-108. [doi: 10.1016/j.gaitpost.2023.06.016] [Medline: 37379735]
45. Arai T, Obuchi S, Shiba Y, Omuro K, Nakano C, Higashi T. The feasibility of measuring joint angular velocity with a gyro-sensor. Arch Phys Med Rehabil 2008 Jan;89(1):95-99. [doi: 10.1016/j.apmr.2007.07.051] [Medline: 18164337]
46. Huang CH, Nihey F, Fukushi K, Kajitani H, Nozaki Y, Nakahara K. Constructing and testing a lightweight model of converting single stride of in-shoe-motion-sensor-measured foot motion to TUG-represented mobility. IEEE Sens Lett 2023 Aug;7(8):1-4. [doi: 10.1109/LSENS.2023.3297797]
47. Huet J, Nordez A, Sarcher A, Mathieu M, Cornu C, Boureau AS. Concordance of freehand 3D ultrasound muscle measurements with sarcopenia parameters in a geriatric rehabilitation ward. J Cachexia Sarcopenia Muscle 2025 Feb;16(1):e13648. [doi: 10.1002/jcsm.13648] [Medline: 39575643]
48. Romero Morales C, Calvo Lobo C, Rodríguez Sanz D, Sanz Corbalán I, Ruiz Ruiz B, López López D. The concurrent validity and reliability of the leg motion system for measuring ankle dorsiflexion range of motion in older adults. PeerJ 2017;5:e2820. [doi: 10.7717/peerj.2820] [Medline: 28070457]
49. Alemayoh TT, Lee JH, Okamoto S. A neural network-based lower extremity joint angle estimation from insole data. Presented at: 2023 20th International Conference on Ubiquitous Robots (UR); Jun 25-28, 2023 URL: https://ieeexplore.ieee.org/abstract/document/10202438 [accessed 2026-03-28] [doi: 10.1109/UR57808.2023.10202438]
50. Pires IM, Denysyuk HV, Villasana MV, et al. Development technologies for the monitoring of six-minute walk test: a systematic review. Sensors (Basel) 2022 Jan 12;22(2):581. [doi: 10.3390/s22020581] [Medline: 35062542]
51. Storm FA, Cesareo A, Reni G, Biffi E. Wearable inertial sensors to assess gait during the 6-minute walk test: a systematic review. Sensors (Basel) 2020 May 6;20(9):2660. [doi: 10.3390/s20092660] [Medline: 32384806]

XSL•FO
RenderX

52. Ancillao A, Tedesco S, Barton J, O'Flynn B. Indirect measurement of ground reaction forces and moments by means of wearable inertial sensors: a systematic review. Sensors (Basel) 2018 Aug 5;18(8):2564. [doi: 10.3390/s18082564] [Medline: 30081607]
53. Prisco G, Pirozzi MA, Santone A, et al. Validity of wearable inertial sensors for gait analysis: a systematic review. Diagnostics (Basel) 2024 Dec 27;15(1):39795564. [doi: 10.3390/diagnostics15010036] [Medline: 39795564]
54. van Helden JFL, Martinez-Valdes E, Strutton PH, Falla D, Chiou SY. Reliability of high-density surface electromyography for assessing characteristics of the thoracic erector spinae during static and dynamic tasks. J Electromyogr Kinesiol 2022 Dec;67:102703. [doi: 10.1016/j.jelekin.2022.102703] [Medline: 36096034]
55. de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes 2006 Aug 22;4(54):54. [doi: 10.1186/1477-7525-4-54] [Medline: 16925807]
56. Ortega-Bastidas P, Gómez B, Aqueveque P, Luarte-Martínez S, Cano-de-la-Cuerda R. Instrumented Timed Up and Go Test (iTUG)-more than assessing time to predict falls: a systematic review. Sensors (Basel) 2023 Mar 24;23(7):3426. [doi: 10.3390/s23073426] [Medline: 37050485]
57. Baker N, Gough C, Gordon SJ. Inertial sensor reliability and validity for static and dynamic balance in healthy adults: a systematic review. Sensors (Basel) 2021 Jul 30;21(15):5167. [doi: 10.3390/s21155167] [Medline: 34372404]
58. Lo PY, Su BL, You YL, Yen CW, Wang ST, Guo LY. Measuring the reliability of postural sway measurements for a static standing task: the effect of age. Front Physiol 2022;13:850707. [doi: 10.3389/fphys.2022.850707] [Medline: 35634138]
59. van Gulick DJJ, van der Leeden M, Lucas C, Stuiver MM. Reliability of static and dynamic balance assessments using the Footwork Pro pressure® platform in a clinical setting. J Sci Med Sport 2025 Dec;28(12):1034-1041. [doi: 10.1016/j.jsams.2025.08.006] [Medline: 40866129]
60. Balachandran AT, Orange ST, Wang Y, Lustin R, Vega A, Quiles N. Comparison of two popular transducers to measure sit-to-stand power in older adults. PLoS ONE 2024;19(8):e0308808. [doi: 10.1371/journal.pone.0308808] [Medline: 39133754]
61. Gao S, Chen J, Xia Y, et al. Wearable technologies for assisted mobility in the real world. Nat Commun 2025 Dec 8;16(1):10988. [doi: 10.1038/s41467-025-67126-4] [Medline: 41361181]
62. Hetherington-Rauth M, Magalhães JP, Alcazar J, et al. Relative sit-to-stand muscle power predicts an older adult's physical independence at age of 90 yrs beyond that of relative handgrip strength, physical activity, and sedentary time: a cross-sectional analysis. Am J Phys Med Rehabil 2022 Nov 1;101(11):995-1000. [doi: 10.1097/PHM.0000000000001945] [Medline: 35034060]
63. Milner T, Brown MRG, Jones C, Leung AWS, Brémault-Phillips S. Multidimensional digital biomarker phenotypes for mild cognitive impairment: considerations for early identification, diagnosis and monitoring. Front Digit Health 2024;6:1265846. [doi: 10.3389/fdgth.2024.1265846] [Medline: 38510280]
64. Fan S, Ye J, Xu Q, et al. Digital health technology combining wearable gait sensors and machine learning improve the accuracy in prediction of frailty. Front Public Health 2023;11:1169083. [doi: 10.3389/fpubh.2023.1169083] [Medline: 37546315]
65. Kelly D, Esquivel KM, Gillespie J, et al. Feasibility of sensor technology for balance assessment in home rehabilitation settings. Sensors (Basel) 2021 Jun 28;21(13):4438. [doi: 10.3390/s21134438] [Medline: 34203571]
66. Zhong R, Rau PLP. Are cost-effective technologies feasible to measure gait in older adults? A systematic review of evidence-based literature. Arch Gerontol Geriatr 2020;87:103970. [doi: 10.1016/j.archger.2019.103970] [Medline: 31743825]
67. Teel EF, Gay MR, Arnett PA, Slobounov SM. Differential sensitivity between a virtual reality balance module and clinically used concussion balance modalities. Clin J Sport Med 2016 Mar;26(2):162-166. [doi: 10.1097/JSM.0000000000000210] [Medline: 26505696]
68. Manupibul U, Tanthuwapathom R, Jarumethitanont W, Kaimuk P, Limroongreungrat W, Charoensuk W. Integration of force and IMU sensors for developing low-cost portable gait measurement system in lower extremities. Sci Rep 2023 Jun 30;13(1):10653. [doi: 10.1038/s41598-023-37761-2] [Medline: 37391570]
69. Mannheim I, Wouters EJM, Köttl H, van Boekel LC, Brankaert R, van Zaalen Y. Ageism in the discourse and practice of designing digital technology for older persons: a scoping review. Gerontologist 2023 Aug 24;63(7):1188-1200. [doi: 10.1093/geront/gnac144] [Medline: 36130318]
70. De Santis KK, Mergenthal L, Christianson L, Busskamp A, Vonstein C, Zeeb H. Digital technologies for health promotion and disease prevention in older people: scoping review. J Med Internet Res 2023 Mar 23;25:e43542. [doi: 10.2196/43542] [Medline: 36951896]
71. Ding J, Ren Y, Xu J, Hu Q, Chu T. Experience of older adults using smart devices and mHealth apps in nursing homes in Chinese megacities: a descriptive qualitative study. Digit Health 2025;11:20552076251353334. [doi: 10.1177/20552076251353334] [Medline: 40547432]
72. Garcia Reyes EP, Kelly R, Buchanan G, Waycott J. Understanding older adults' experiences with technologies for health self-management: interview study. JMIR Aging 2023 Mar 21;6:e43197. [doi: 10.2196/43197] [Medline: 36943333]
73. Sharma N, Brinke JK, Van Gemert-Pijnen JEWC, Braakman-Jansen LMA. Implementation of unobtrusive sensing systems for older adult care: scoping review. JMIR Aging 2021 Oct 6;4(4):e27862. [doi: 10.2196/27862] [Medline: 34612822]

XSL•FO
RenderX

74. Teng S, Kim JY, Jeon S, et al. Analyzing optimal wearable motion sensor placement for accurate classification of fall directions. Sensors (Basel) 2024 Oct 4;24(19):6432. [doi: 10.3390/s24196432] [Medline: 39409472]
75. Trigg A, Ratitch B, Kruesmann F, Majumder M, Parfionovas A, Krahn U. Interpretation of change in novel digital measures: a statistical review and tutorial. Digit Biomark 2025;9(1):52-66. [doi: 10.1159/000543899] [Medline: 40103940]
76. Sadeghi Z, Alizadehsani R, Cifci MA, et al. A review of explainable artificial intelligence in healthcare. Comput Electr Eng 2024 Aug;118:109370. [doi: 10.1016/j.compeleceng.2024.109370]
77. Gagnier JJ, Lai J, Mokkink LB, Terwee CB. COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. Qual Life Res 2021 Aug;30(8):2197-2218. [doi: 10.1007/s11136-021-02822-4] [Medline: 33818733]

Abbreviations

AP: anteroposterior
COP: center of pressure
COSMIN: Consensus-Based Standards for the Selection of Health Measurement Instruments
ICC: intraclass correlation coefficient
IMU: inertial measurement unit
MDC: minimal detectable change
MIC: minimal important change
ML: mediolateral
OMI: Outcome Measures Instruments in Systematic Reviews
PICO: Population, Intervention, Control, and Outcomes
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO: International Prospective Register of Systematic Reviews
RMS: root mean square
SEM: standard error of measurement

Edited by S Brini; submitted 09.Sep.2025; peer-reviewed by C Ahmadu, J Guo, W Qi; revised version received 07.Mar.2026; accepted 09.Mar.2026; published 14.Apr.2026.

Please cite as:
Zhou S, Huang J, Yu J, Li X, Zhang C, Zhao Y, Hu M, Feng H
Digital Technologies and Biomarkers for Locomotor Capacity Assessment in Older Adults: Systematic Review
J Med Internet Res 2026;28:e83814
URL: https://www.jmir.org/2026/1/e83814
doi:10.2196/83814

© Shuhan Zhou, Jundan Huang, Jia Yu, Xiaoyang Li, Chi Zhang, Yinan Zhao, Mingyue Hu, Hui Feng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Health-Related Digital Ecological Momentary Assessment in Children (Aged 5– 11 Years): Systematic Review

Sydney Charitos, MSc; Lauren Thompson, PhD; Amberly Brigden, PhD; Jon Bird, DPhil

School of Engineering Mathematics and Technology, University of Bristol, 1 Cathedral Square, Bristol, United Kingdom

Corresponding Author:
Sydney Charitos, MSc
School of Engineering Mathematics and Technology, University of Bristol, 1 Cathedral Square, Bristol, United Kingdom

Abstract

Background: Digital ecological momentary assessment (EMA) collects data on experiences as they occur in daily life, capturing dynamic, context-sensitive experiences often missed by retrospective reporting. While EMA shows promise for pediatric health research, preadolescents have distinct socioemotional and cognitive characteristics likely to affect engagement. Existing reviews have not focused on the acceptability and feasibility of EMA protocols for this age group.

Objective: This review aimed to examine digital EMA protocols used with children aged 5 - 11 years across health domains, focusing on protocol characteristics, acceptability, and feasibility. We address 3 research questions (RQs)—RQ1: What are the characteristics of these protocols? RQ2: What is the feasibility and acceptability of these protocols? RQ3: What are the characteristics of high and low response rate protocols?

Methods: We searched 10 databases (CINAHL, Embase, ACM Digital Library, IEEE Xplore, Cochrane Library, PsycINFO, Web of Science, PubMed, Scopus, and MEDLINE) for peer-reviewed studies published up to October 2025. Eligible studies used EMA with children aged 5 - 11 years to collect health data via digital devices. Two researchers independently screened studies (SC and LT); one (SC) conducted quality assessment and data extraction. Findings were narratively synthesized.

Results: We identified 17 protocols across 37 studies. Most targeted nonclinical populations, used handheld devices, spanned 3 - 28 days, and applied interval-contingent prompting (RQ1). Response rates were available or calculable for 15 of 17 protocols, ranging from 48% to 92% (RQ2). Six protocols reported response rates of ≥80%. However, key data required for pooling (eg, raw counts for planned vs completed prompts) were missing or selectively reported. This contributed to 13 of 17 protocols being rated at critical risk of bias (ROBINS-I, v2). As a result, the strength of evidence was limited by poor reporting and high risk of bias. Facilitators included uncomplicated, engaging technology, reminders, and caregiver involvement. Barriers included device burden, restricted device access, difficulty with accurate reporting, stigma, limited device awareness, and insufficient caregiver support. High-response protocols (≥80%) often involved older children or clinical groups, ≥3-week duration, fixed schedules (≥20 items per prompt, 3 or 4 times per day), timing customization, and incentives (RQ3).

Conclusions: This review provides the first systematic synthesis on preadolescents, offering insight into EMA protocol design beyond prior work treating children as a single group. By examining 17 EMA protocols, the review identifies gaps in developmental appropriateness and reporting quality, highlighting where the evidence may differ from adolescent and adult EMA research. The results suggest that digital EMA for preadolescents requires greater focus on child-centered design to increase acceptability and adherence, alongside improved reporting standards, so protocols can be meaningfully compared. With these advances, EMA could be more effectively integrated into pediatric health monitoring, tailored to the needs of different age groups.

Trial Registration: Prospero ref-CRD42022373812; https://www.crd.york.ac.uk/PROSPERO/view/CRD42022373812

(J Med Internet Res 2026;28:e79291) doi:10.2196/79291

KEYWORDS

children; systematic review; ecological momentary assessment; acceptability; feasibility; PRISMA; mobile phone

Introduction

Background

Ecological momentary assessment (EMA) is a research method used to collect data on individuals' behaviors, experiences, and physiological states as they occur in real time and in everyday settings [1]. EMA has proven particularly valuable in health research because it captures experiences that are often dynamic, personal, and shaped by context [2-4]. For example, symptoms such as pain, fatigue, or mood often fluctuate over short periods and may be influenced by daily routines, social interactions, or environmental triggers [5-7]. In contrast, traditional methods, such as retrospective reporting, may struggle to measure these experiences accurately due to recall bias and the influence of recency effects and emotional salience [1]. In recent years, the

XSL•FO
RenderX

increasing availability of smartphones and wearable devices has supported the integration of EMA into individuals' daily routines [8-10]. A typical EMA protocol uses these devices to deliver prompts, brief surveys sent to participants at scheduled times. Each prompt includes 1 or more questions, referred to as the item count, which can vary depending on the protocol's aims and design [1].

EMA may be a particularly useful method for capturing health-related experiences in preadolescent children (aged 5 - 11 years). Children at this age demonstrate metacognitive ability, reflected in their capacity to consider their own emotions and experiences, at a level similar to that of adults [11]. In contrast, traditional retrospective self-report approaches can be especially challenging for this age group [12]. At this developmental stage, children's long-term memory is still maturing, which can limit how accurately they recall past events [13]. They can also find it difficult to verbally express their health experiences, especially in unfamiliar clinical settings or when speaking with health care professionals they do not know personally [14]. EMA may help overcome these barriers by enabling in-the-moment reporting rather than requiring reflection on past experiences.

However, a central challenge in implementing an EMA protocol is achieving both acceptability, that is, how suitable, engaging, or satisfying the method is perceived to be by participants, and feasibility, that is, how practical and realistic it is to implement [15]. These dimensions are essential for ensuring sustained participation and the collection of high-quality data. As a result, researchers have explored a variety of methodological adaptations aimed at improving protocol acceptability and feasibility. For example, micro-EMA is a protocol design that uses ultrabrief prompts (within a few seconds) to reduce participant burden [16]. Researchers have also incorporated context-aware delivery methods that adjust prompts based on user behavior or timing to minimize disruption [17,18]. Such adaptations reflect the ongoing evolution of EMA as researchers refine approaches to better support acceptability and feasibility across a range of participants and contexts.

EMA has consistently demonstrated acceptability and feasibility in adult populations, with average response rates (the proportion of scheduled prompts completed by participants) frequently exceeding 80% [19], a standard of high response rate in EMA studies [20]. Although certain protocol features, such as item count and prompt frequency, can influence response rates, adults generally maintain strong response levels (above 70%) despite these variations [19]. A factorial experiment designed to isolate the individual effects of specific protocol characteristics found that common variations did not significantly affect response rates [21]. Instead, individual differences and the interactions between protocol elements were identified as having a more substantial influence on response rates. These findings indicate that EMA protocols for adult populations can be flexibly adapted to specific research requirements without substantially compromising response rates.

Compared with adults, existing evidence on response rates in EMA protocols for preadolescent children remains limited and inconclusive. A 2023 meta-analysis suggested that age may not be a strong predictor of EMA response rate; however, this analysis included only 6 protocols involving children younger than 12 years (1.2% total sample), limiting confidence in the finding for this age group [22]. Furthermore, many reviews aggregate data across preadolescent (aged 5 - 11 years) and adolescent (12 - 18 years) age groups, making it difficult to isolate trends specific to younger children [23-25]. For example, Wen et al [23] examined mobile EMA with "children and adolescents," but only one of their included studies overlaps with the present review, illustrating how younger children are rarely the focus of EMA research. Their review focused on mobile EMA in health science databases and largely adolescent samples, whereas the present review searched a wider range of interdisciplinary sources, including Human-Computer Interaction (HCI) venues, and targeted protocols specifically designed for preadolescent children (aged 5 - 11 years). Within this aggregation, there is some evidence that children and adolescents may exhibit lower response rates than adults; a systematic review reported an average compliance rate of 78.3% among pediatric samples, below the levels typically observed in adult EMA studies [23]. Despite this tendency to group age bands or assume similar patterns across age groups, systematic reviews nonetheless emphasize the importance of adapting EMA protocols to better meet the developmental and contextual needs of younger children [23,24,26]. Recommendations include restricting prompts to occur outside school hours and limiting internet access to support caregiver oversight [24]. While such recommendations are noted within existing systematic reviews, they are not typically the primary focus, underscoring the need for more targeted research on how EMA protocols can be effectively adapted for younger children.

Aims

This systematic review aimed to investigate how EMA protocols are being used with children aged 5 - 11 years, across health domains and population characteristics, focusing on their acceptability and feasibility for children. This review aimed to build on the established literature on the adherence of children to EMA protocols by investigating the acceptability and feasibility of these different protocol characteristics and understanding how current protocols attempt to improve adherence. In this review, we investigated the following research questions (RQs): (1) What are the characteristics of health-related EMA protocols being used with children aged 5-11 years? (2) What is the feasibility and acceptability of these EMA protocols with children aged 5-11 years? (3) What characteristics of EMA protocols, RQ1, are related to high and low response rate, RQ2, when using EMA with children between the ages of 5 and 11 years?

Methods

This review was registered in the PROSPERO database and follows the guidance of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [27]. See PRISMA and PRISMA-S checklists.

Search Strategy

The literature search was first conducted on March 13, 2023, and updated on October 28, 2024, and October 15, 2025, in the following databases: CINAHL, ACM Digital Library,

XSL•FO
RenderX

PsycINFO, Embase, MEDLINE, Cochrane Library, IEEE Xplore, PubMed, Scopus, and Web of Science. Each database was searched via its native platform (eg, MEDLINE via Ovid, PsycINFO via Ovid, and Scopus via Elsevier), with full platform—database mappings and exact search strings provided in PRISMA-S checklist. These were chosen to span multiple disciplines including medical and HCI research. The search included key terms for (1) children aged 5-11 years, such as "school child" or "minor" and (2) EMA, such as "Experience Sampling" or "Ambulatory Assessment" (Multimedia Appendix 1 [28,29] provides the full list of search terms used). No other limits or restrictions were placed on searches other than date restrictions for subsequent literature reviews in 2024 and 2025. The search strategy was not formally peer reviewed but was discussed with the wider research team. No study registries, conference proceedings, organizational websites, or gray literature sources were searched, as the review focused on peer-reviewed empirical studies. The search and reporting methods were documented in accordance with the PRISMA-S (PRISMA Search) checklist, with full database-specific search strategies, platforms, and adaptations reported in PRISMA-S checklist.

Inclusion and Exclusion Criteria

Studies were included if they fulfilled the following criteria: (1) peer-reviewed or empirical studies with any study design, (2) published in English, (3) the research involved the use of EMA to collect health-related information (physical, mental, or social) on a child (aged 5‑11 years) either by the child themselves or proxy-reported by a parent or a caregiver, and (4) a digital technology (eg, smartphone) was used to collect EMA data. For the purpose of this review, EMA is defined as the assessment of a phenomenon in natural settings at least twice a day [30]. Studies were excluded if they fulfilled the following criteria: (1) the study included children younger than 5 years or older than 11 years or did not report adaptations specific to the preadolescent age range (5‑11 years). Protocols designed for a broader group of children (eg, aged 5‑18 years) without age-specific considerations were excluded, as the focus of this review was on protocols intentionally tailored for preadolescents; or (2) the paper was considered a review on previously reported studies, protocol, nonempirical or nonscientific paper (eg, book chapters), or when the full text was not available.

Screening Procedures

The search results were imported into Zotero (Corporation for Digital Scholarship) and duplicates were removed. In stage 1, each title and abstract was independently screened by 2 authors (SC and LT) for relevance. If an inclusion decision could not be made based on the title and abstract, the papers were included for full-text review. Papers deemed relevant at the title- and abstract-screening stage moved onto stage 2 where they were independently double-screened against the inclusion and exclusion criteria using the data management platform Rayaan (Rayyan Systems Inc), with reasons for exclusion recorded. Disagreements at both stages were discussed and resolved in meetings by the reviewers (SC and LT), with persistent disagreements being resolved through discussion with the wider research team (AB and JB). If the full text was not available through institutional subscriptions, interlibrary loan, or author contact (2 email attempts), the study was excluded. Where the full text was available but did not contain the information needed, 2 attempts were made to contact authors by email; if the information was not provided, the study was excluded. In addition, reference lists of included studies were manually screened to identify any additional eligible papers.

Data Extraction and Synthesis

Data were extracted by the primary author (SC), following the protocol in Textbox 1, which aligns with the Checklist for Reporting EMA Studies (CREMAS) [25]. In line with our PROSPERO registration, methods were adapted during analysis: RQ3 was added to examine protocol features linked to response rates, while planned technology-based taxonomies were not applied due to insufficient reporting. A quantitative meta-analysis was not feasible, as key data required for pooling (eg, raw counts for planned vs completed prompts and corresponding variance estimates) were missing or selectively reported across protocols. This missingness contributed to 13 of 17 protocols being judged at critical risk of bias during quality assessment (see Multimedia Appendix 2 [31-60] for full assessments and Figure 1 for a summary). As a result, a narrative synthesis was undertaken [61]. Pooled response rate estimates were calculated for the subset of protocols not judged at critical risk of bias using a random-effects model, with mean response rates, 95% CIs, and a prediction interval reported.

The search identified 31 eligible papers describing 37 studies. As several papers reported on the same underlying methodology, and some studies included distinct subgroups with methodological differences, these were consolidated into 17 unique EMA protocols, which formed the unit of analysis for this review. Where protocols spanned multiple papers, these were grouped under project or funding names, with "main" or "sub" labels applied where necessary. Protocols were named based on the explicit project name when provided or if this was not available, the funding body.

XSL•FO
RenderX

Textbox 1. Data extraction categories and details extracted.

General protocol aim and participants (study-level information included in Multimedia Appendix 3 [31-60]).

- Publication details: authors, year of publication, country of publication, and publication location
- Sample size
- Demographics: participant age, health condition if specified
- Protocol aim

Research question (RQ) 1: What are the characteristics of the ecological momentary assessment (EMA) protocol?

- EMA purpose (eg, intervention or observational tool)
- Training (included in the Checklist for Reporting EMA Studies [CREMAS]): child and parent
- Technology (included in the CREMAS): reporter, device, software, response method, any additional devices (eg, accelerometer)
- Monitoring period (included in the CREMAS)
- Prompting design (included in the CREMAS): strategy (eg, interval, event-contingent), number of questions
- Prompting frequency (included in the CREMAS)
- Design features (included in the CREMAS): piloting (item and protocol level), reward for participation, customization, reminder systems, support networks, and protocol flexibility.

RQ 2: Is it feasible and acceptable to use digital devices to implement an EMA methodology with children between the ages of 5 and 11 years for health purposes?

- Feasibility details: dropout, attrition, response latency, and response rate (planned vs complete).
- Acceptability details: qualitative evidence relating to acceptability and feasibility (eg, exit interviews)
- Any information captured about acceptability and feasibility but not captured through formal qualitative or quantitative data collection, that is, authors' interpretations in the results and discussion.

RQ 3: What characteristics of EMA protocols, RQ1, are related to high and low adherence, RQ2, when using EMA with children between the ages of 5 and 11 years?

- Additional synthesis of data extracted from RQ1 and RQ2, see "Data Extraction and Synthesis" section for details.

XSL•FO
RenderX

Figure 1. Summary of risk of bias for the 17 identified ecological momentary assessment protocols, assessed using the Risk Of Bias In Nonrandomized Studies of Interventions, version 2 (ROBINS-I v2) using robviz tool [33]. The figure displays domain-level and overall ratings, showing that most protocols (13 out of 17) were judged to be at critical risk of bias [31-60,62].

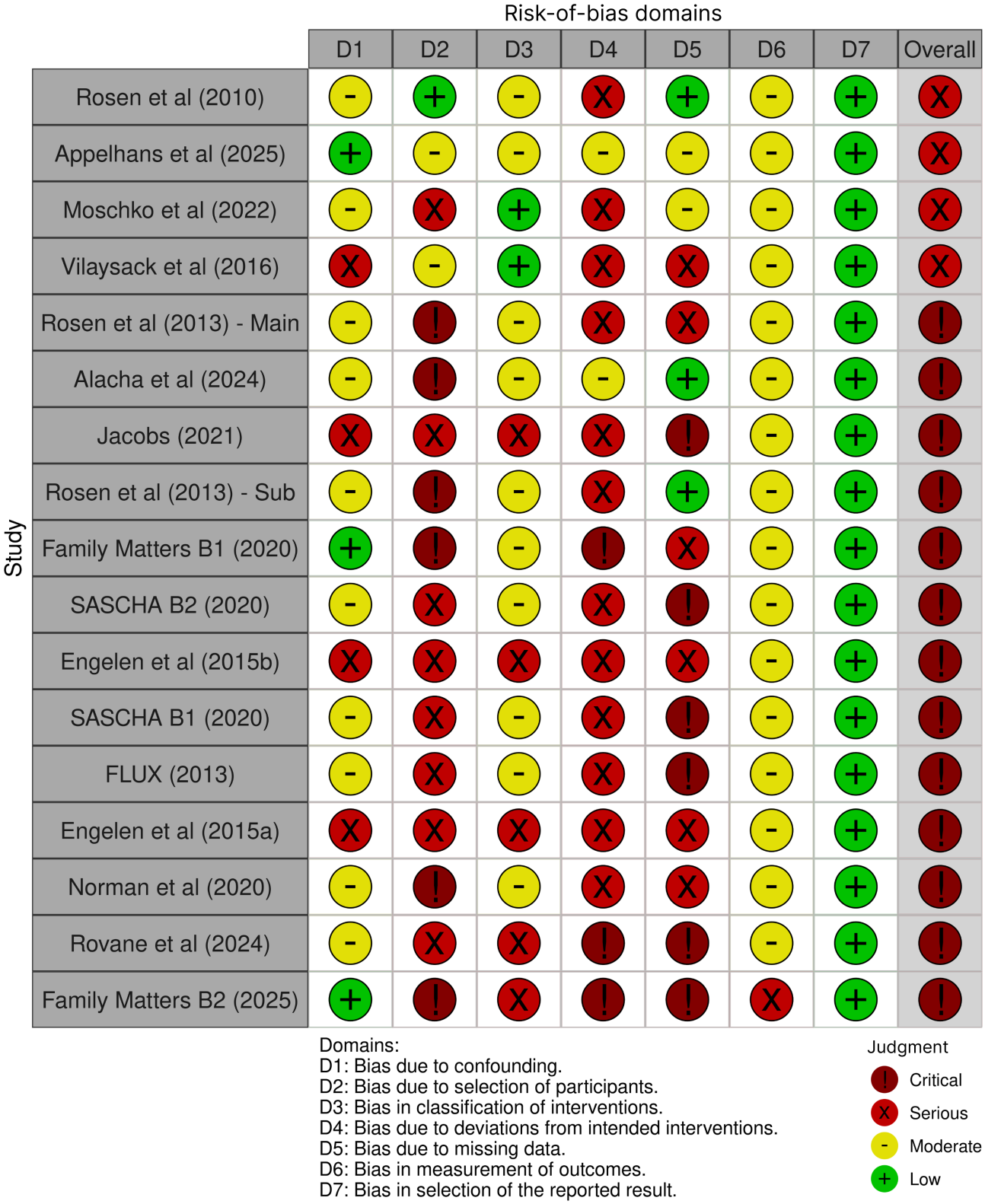

For RQ1, we extracted the frequency of each protocol feature and synthesized findings through descriptive statistics. To ensure consistency, ranges were reported when no overall metric was provided across subanalyses. Within the CREMAS design features category, 5 subcategories were identified to support consistent reporting: rewards for participation, piloting, customization, reminder systems, and support networks.

For RQ2, we extracted adherence-related metrics (including response rate, attrition, and latency) and qualitative data on acceptability and feasibility. If overall response rates were not reported, they were calculated from raw data (planned vs actual responses). Full calculations can be found in Multimedia Appendix 4 [31-60]. The planned metaethnography was not feasible, as only 2 of 17 protocols included formal qualitative

XSL•FO
RenderX

research on acceptability and feasibility. Vilaysack et al [31] provided a summary of feedback; Norman et al [32] applied manifest content analysis to identify barriers (demanding, challenging in irregular situations) and facilitators (uncomplicated, engaging). Instead, we applied a thematic synthesis approach [63], using the framework of barriers and facilitators by Norman et al as a deductive starting point and expanding it inductively to incorporate additional author observations and participant quotes.

For RQ3, which examined how protocol characteristics (RQ1) relate to response rate patterns (RQ2), no additional data were extracted. Instead, to synthesize data for this RQ, protocols were categorized as high or low response rate using an 80% threshold (Stone and Shiffman [20]). Researchers then identified protocol characteristics (identified in RQ2) found in >50% of protocols within each group.

Quality Assessment

One author conducted a quality assessment, discussing any borderline assessments with the wider team. We used 2 complementary tools: the CREMAS [25], and the Risk Of Bias In Nonrandomized Studies of Interventions, version 2 (ROBINS-I v2) [64]. The CREMAS is a 16-item tool designed to enhance the transparency and rigor of EMA reporting. For the purposes of this review, each item was assessed using a binary coding scheme (“present” or “not present”), with ambiguous cases conservatively marked as “not present.” The ROBINS-I v2 tool was used to evaluate risk of bias across 7 domains. Full details of both assessments are provided in Multimedia Appendix 2, with a summary of ROBINS-I v2 ratings shown in Figure 1.

Results

The PRISMA flowchart is represented in Figure 2. In this section, we synthesize data extracted to address our RQs as outlined in Textbox 1. Protocols are grouped by overall ROBINS-I rating (critical, serious, or moderate) and sorted within each group by response rate to facilitate comparison across quality levels.

XSL•FO
RenderX

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. EMA: ecological momentary assessment.

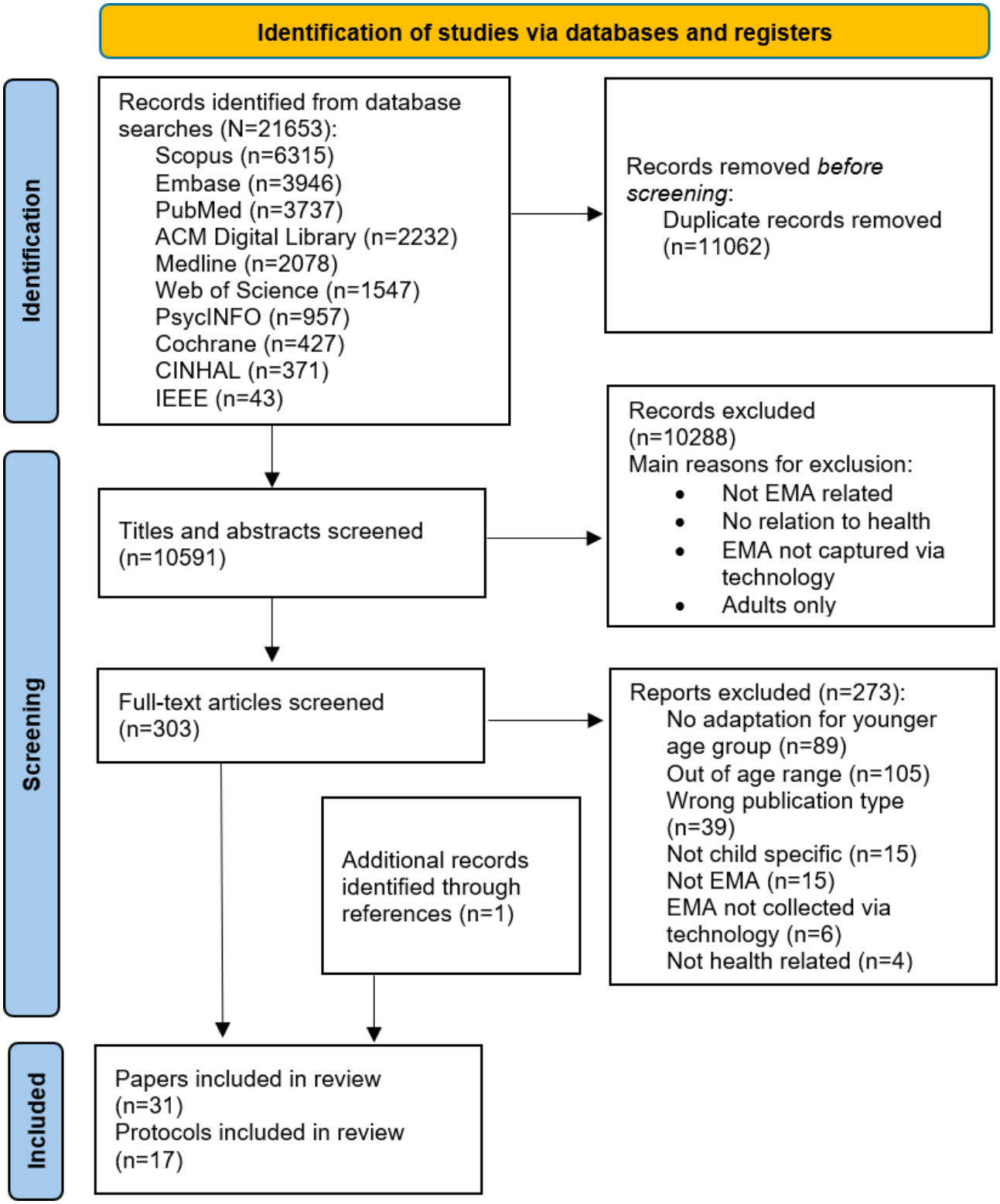

Research question 1: What are the characteristics of the EMA protocol?

For RQ1, Tables 1-3 summarize key characteristics of the included protocols. Table 1 summarizes protocol domains and participants; Table 2 outlines training and technology features; and Table 3 details prompt implementation.

XSL•FO
RenderX

Table . Overview of included ecological momentary assessment protocols, including study location, sample size, child age, condition, and primary aim (RQ1).

Protocol	Country	Sample size	Child's age (years)	Condition	Aim
Serious					
Rosen and Epstein (2010) [33]	United States	2	8-11	ADHD[a]	To assess links between parental stress, feeding practices, and child-eating behaviors via EMA[b].
Appelhans et al (2025) [34]	United States	60	5-10	None	To examine whether parent-supported recreational activities can displace discretionary eating and electronic entertainment.
Moschko et al (2022) [35]	Germany	70	9-11	ADHD	To investigate daily self-regulation in children with ADHD and parent-child interactions.
Vilaysack et al (2016) [31]	Australia	10	5-7	None	To assess EMA feasibility with typically developing children.
Critical					
Rosen et al (2013)—Main [36]	United States	11	7-11	ADHD	To examine parent EMA proxy reports of children's emotional dysregulation.
Alacha et al (2024) [37]	United States	47	9-10	None	To investigate effects of positive affect variability on homework problems in children with ADHD.
Jacobs (2021) [38-40]	Germany	84	8-11	ADHD	To explore links between emotional regulation, sleep, and well-being using ambulatory methods.
Rosen et al (2013)—Sub [36]	United States	5	5-7	None	To explore how children can use EMA to self-report on their own emotional dysregulation.
Family Matters B1 (2020) [41-47]	United States	128-150	9-11	None	To examine diets of low-income, racially and ethnically diverse families.
SASCHA B2 (2020) [40,48-50,62]	Germany	108	5-7	None	To assess how transition to secondary school affects well-being and academic success.
Engelen, Bundy, Lau et al (2015) [51]	Australia	20	9-11	None	To examine links between children's activity levels and contextual factors.

XSL•FO
RenderX

Protocol	Country	Sample size	Child's age (years)	Condition	Aim
SASCHA B1 (2020) [48,50,62]	Germany	90	8-11	None	To investigate self-esteem, peer ties, and academic functioning before secondary school.
FLUX (2013) [38,52-57]	Germany	82-110	9-11	ADHD	To examine how sleep, activity, affect, peers, and worry relate to working memory and well-being.
Engelen, Bundy, Bauman et al (2015) [58]	Australia	246	5-7	None	To assess EMA feasibility for describing after-school activity patterns.
Norman et al (2020) [32]	Sweden	20	5-7	None	To examine feasibility and validity of photo-based EMA for children's diets.
Rovane et al (2025) [59]	United States	92	5-11	ASD[c]	To explore links between parental emotion regulation, stress, and child behavior in ASD.
Family Matters B2 (2025) [60]	United States	436	5-7	None	To examine how parental stress, mood, and coping relate to children's physical activity and screen time in daily life.

[a]ADHD: attention-deficit/hyperactivity disorder.

[b]EMA: ecological momentary assessment.

[c]ASD: autism spectrum disorder.

XSL•FO
RenderX

Table . Summary of protocol training and technology features across ecological momentary assessment protocols with children aged 5-11 years (research question 1).

Protocol	Reporter	Training child, parent	Technology	Response method	Reported question domain(s)
Serious					
Rosen and Epstein (2010) [33]	Parent	N[a], Y[b]	PDA[c]	VAS[d] (11: −5 to 5), VAS (10: 1 to 10), Categorical, and Likert (5)	Affect
Appelhans et al (2025) [34]	Parent	NR[e], Y	Smartphone	Multi-Choice, Single-Choice	Activity (type), diet (food intake)
Moschko et al (2022) [35]	Child, parent	Y, NR	Child: smartphone; parent: survey (online or paper)	Likert (6)	Self-regulation, social relationships
Vilaysack et al (2016) [31]	Child	Y, Y	Smartphone	VAS (NR), Multi-Choice, Categorical	Activity
Critical					
Rosen et al (2013)—Main [36]	Parent	NR, Y	PDA	VAS (NR)	Affect
Alacha et al (2024) [37]	Child	NR, Y	Mobile phone	Likert (5)	Affect
Jacobs (2021) [38-40]	Child, parent	Y, NR	Smartphone	Likert (5), Working Memory Updating Task, and Time Picker	Sleep, affect, and working memory
Rosen et al (2013)—Sub [36]	Parent	NR, Y	PDA	VAS (NR)	Affect
Family Matters B1 (2020) [41-47]	Child	NR, Y	Tablet. optional: phone, accelerometer	Categorical, VAS (NR), Multi-Choice, and Likert (NR)	Food intake, stress, sleep, affect, and activity
SASCHA B2 (2020) [40,48-50,62]	Parent	Y, Y	Smartphone (app)	Likert (5), Working Memory Updating Task, and Time Picker	Sleep, affect, social relationships, working memory, self-esteem, achievement goals, self-regulation, and academic success
Engelen, Bundy, Lau et al (2015) [51]	Child	NR, NR	PDA, accelerometer	Multi-Choice, VAS (NR)	Activity
SASCHA B1 (2020) [48,50,62]	Child	Y, Y	Smartphone	Likert (5), Working Memory Updating Task, and Time Picker	Social relationships, affect, self-esteem, achievement goals, self-regulation, working memory, and academic success
FLUX (2013) [38,52-57]	Child, Parent	Y; Y	Smartphone, accelerometer	Likert (5), Working Memory Updating Task, and Time Picker	Affect, working memory, and sleep
Engelen, Bundy, Bauman et al (2015) [58]	Parent	NR, Y	PDA	Multi-Choice, VAS (NR)	Activity
Norman et al (2020) [32]	Child	NR, Y	Mobile phone	Photo, Free Text	Food intake
Rovane et al (2025) [59]	Parent	NR, Y	Smartphone	Likert (9), Categorical, and Likert (5)	Stress, behavior
Family Matters B2 (2025) [60]	Parent	NR, Y	Smartphone	Single-Choice, VAS (NR), Multi-Choice, and Likert (NR)	Diet, stress, sleep, affect, and activity

XSL•FO
RenderX

[a]N: Training did not happen.
[b]Y: Training is reported.
[c]PDA: personal digital assistant.
[d]VAS: visual analog scale.
[e]NR: not reported.

XSL•FO
RenderX

Table . Summary of prompt implementation details across ecological momentary assessment protocols with children aged 5-11 years (research question 1).

Protocol	Period	Strategy	Type[a]	Items, n	Frequency, n	Prompt interval
Serious						
Rosen and Epstein (2010) [33]	28 days	Interval	Fixed	29	3	Before school, after school, and evening
Appelhans et al (2025) [34]	17 days	Interval	Pseudorandom	4-5	3	Pseudorandom times during non–school hours and through entire day on weekends
Moschko et al (2022) [35]	18 days; 3× over 13 months	Interval	Random	7- 8	3 children; 1 parent	Before school, afternoon, and evening
Vilaysack et al (2016) [31]	7 days	Interval	Random	7	8	Random during waking hours (including school)
Critical						
Rosen et al (2013)—Main [36]	28 days	Interval	Fixed	1	3	Before school, after school, and evening
Alacha et al (2024) [37]	7 days	Interval	Fixed	20	3	Morning, afternoon or after school, and evening
Jacobs (2021) [38-40]	21 days	Interval	NR[b]	17-25	4	Before school (later on weekends), afternoon, and evening
Rosen et al (2013)—Sub [36]	28 days	Interval	Fixed	1 child; 2 parents	3	Before school, after school, and evening
Family Matters B1 (2020) [41-47]	8-10 days	Interval, event	Fixed, random	10-30	5 + event	Even split across waking hours
SASCHA B2 (2020) [40,48-50,62]	4 weeks	Interval	NR	28-41	4	Before school, morning (including school), afternoon, and evening
Engelen, Bundy, Lau et al (2015) [51]	4 days; 2× 13-week gap	Interval	Random	12	3	Between afternoon and evening
SASCHA B1 (2020) [48,50,62]	4 weeks	Interval	NR	28-41	4	Before school, morning (including school), afternoon, and evening.
FLUX (2013) [38,52-57]	4 weeks	Interval	Fixed	21-26	4	Morning (including school), midday (including school), afternoon, and evening
Engelen, Bundy, Bauman et al (2015) [58]	4 days; 2× 13-week gap	Interval	Random	12	3	Between afternoon and evening
Norman et al (2020) [32]	3 days	Interval, event	Fixed	1	1 + event	Evening
Rovane et al (2025) [59]	7 days	Interval	Random	3	5	Random during waking hours
Family Matters B2 (2025) [60]	7 days +	Interval	Fixed, random	10-34	4	Randomly within 3-hour window, EoD[c] available later in the day

XSL•FO
RenderX

[a]Studies typically described prompting as "random" or "pseudorandom," but few reported whether constraints (eg, minimum time between prompts) were applied. Our synthesis, therefore, reflects the terminology reported by authors, acknowledging that "random" may have been implemented differently across protocols.

[b]NR not reported.

[c]EoD: end of day.

General Protocol Domains and Participants

Table 1 summarizes protocol domains and participants (RQ1). Across the 17 EMA protocols, most (n=13/17) were designed to focus solely on either key stage (KS) 1 (ages 5 - 7 years; n=5) or KS2 (ages 8 - 11 years; n=8) [33,35,36,38-40,48-50,52-57,62] compared with KS1 [31,32,41-47,51,58]. Protocols sampled typically developing children [31,32,34,36,37,40-51,58,60,62] more frequently than those with specific conditions: attention-deficit/hyperactivity disorder (ADHD; n=5) [33,35-37] and autism spectrum disorder (n=1) [59]. Female representation varied (0% - 56%) but was between 40% and 56% in 13 protocols [31,32,34,35,37-58,60,62] and often lower among condition-related protocols, with 4 of 6 having ≤23% female participants [33,35,36,59].

Research question 1: What are the characteristics of the EMA protocol?

Reporter Training

Most protocols (n=15/17) used a single reporter: usually parent-reported in KS1 (n=7) [32,34,37,41-47,51,58-60], whereas protocols involving KS2 were split between child self-reports and parent proxy reports (n=6 child; n=7 parent) [35,36,38-40,48-50,52-57,62]. Where reported (n=5), parent respondents were predominantly female (74% - 100%) [32,33,35,41-47,59]. Child training was reported in 6 protocols and always involved verbal instruction with hands-on practice, with session durations ranging from 45 minutes to 4.5 hours when reported (n=3) [31,38-40,52-57]. Parent training was reported in 14 protocols [31-34,36,37,40-50,58-60,62]. Of these, 9 provided verbal instruction [31,33,34,37,40-50,59,60,62], 6 provided written materials [32,38,40-50,52-58,62], and 6 offered hands-on practice [31,33,34,41-47,59,60]. When reported (n=3), parent training sessions ranged from 15 to 45 minutes [31,33,59]. Three protocols provided training to both children and parents when the child was the reporter [31,40,48-50,52,53,55-57,62], but only 1 protocol included practice-based training for parents in this context [31]. In protocols where both child and parent were reporters (n=2), only 1 member of the pair was reported to have received training [35].

Technology

Nearly all protocols (n=16/17) used dedicated handheld devices, most commonly smartphones (n=11) [31,32,34,35,37-40,48-50,52-57,59,60,62] or Personal Digital Assistants (n=5) [33,36,41-47,51,58], with 1 using a tablet [41-47]. Three protocols paired EMA with accelerometers for passive tracking [38,41-47,51-57]. Twelve protocols provided devices, typically restricting functionality to the research app (n=9) [33,35,36,38-50,55-57,62], while 5 allowed personal device use [32,34,35,37,60]; this occurred only when parents were reporting.

Response Methods

The most popular response collection method was a Likert scale (n=10/17) [33,35,37-48,50,52-57,59,60,62], most commonly 5-point scale (n=7) [33,37-40,48-50,52-57,59,62]. Nearly all (n=14) combined 2 or more response types, often pairing Likert scales with visual analog scales or categorical options.

EMA Purpose and Domain

All protocols used EMA as an observational tool and not an intervention. Twelve captured multiple domains, while 5 focused on a single domain [31,32,37,51,58]. Affect was most common (n=10), followed by activity (n=6) and sleep (n=5).

Monitoring Period

Monitoring periods across the 17 protocols spanned from 3 to 28 days for a single wave: ≤7 days (n=7) [31,32,37,51,58-60], 8 - 14 days (n=1) [41-47], 15 - 21 days (n=3) [35,38-40], and ≥22 days (n=6) [33,36,38,40,48-50,52-57,62]. Only the Family Matters protocols (n=2) [41-47] included flexibility beyond their stated monitoring periods—extending participation if families did not complete a "full day" of EMA (2 of 4 interval - contingent prompts, 1 mealtime survey, and 1 end - of - day survey).

Prompt Design and Frequency

All protocols included interval-contingent prompting, typically 3-4 times per day (n=13/17) [33-40,48-58,60,62]. Prompt timing followed three main patterns: (1) 3 windows per day (before school, after school, and evening; n=5) [33,35-40], (2) 4 windows (adding a midmorning school prompt; n=4) [35,43-46,48-52], (3) or prompts spaced evenly or randomly across waking hours (n=5, one including schooltime prompts with child reporters) [31,34,41-47,59,60]. One protocol [32] added an end-of-day prompt mid-study after a poor response rate to daytime event-contingent prompts was observed. Six protocols varied item count by time of day (eg, shorter morning and longer end-of-day surveys) [35,38-50,52-57,62]. Response windows ranged from 3 minutes to 6 hours, with 2 protocols linking window length to the interval until the next prompt [40,48-50,62] and 2 adjusted it by time of day [41-47,60].

Design Features

Design features relevant to reporting bias and participant burden were categorized into 5 domains—rewards, piloting, customization, reminders, and support. These domains are summarized, with detailed protocol-level information provided in Multimedia Appendix 5 [31-60,62].

Rewards for Participation

Of the 17 protocols, 13 offered participant reimbursement. Fixed payments, meaning amounts provided regardless of response rate, were reported in 5 protocols [35,37-40,48-50,59,62]. Six protocols tied bonuses to response rate thresholds of 60% - 90%, typically offering an extra US $6 to US $12 for meeting these

XSL•FO
RenderX

targets (n=3) [38-40,48-50,59,62], or stating that the total payment was dependent on response rate (eg, up to US $100) (n=5) [34,38,41-48,50,52-57,60,62]. The maximum payment per prompt ranged from US $0.48 [37] to US $2.68 [60], with an average of US $0.95 per prompt. Six protocols supplemented cash with nonmonetary incentives (eg, iPad [41-47], activity center tickets [32], and data summary [59]). Two protocols explicitly offered no reimbursement [36].

Piloting

Only 2 of the 17 protocols reported fully piloting its EMA protocol [31,60]. In terms of item piloting, 3 protocols reported piloting individual items with children from the target age group [38-40,52-57]. Three protocols adapted existing non-EMA measures for some, but not all, EMA items, with limited detail on sources or testing [32,35,40-50,54,55,61,62].

Customization to Protocol

Of the 17 protocols, 11 allowed some customization to support adherence, which always involved tailoring prompt timing to participants' daily routines at the start of the study, with 1 protocol providing the option to update this midway through the study period [33]. Only 2 protocols offered multiple customization options including delivery method (text or email) and EMA question language [41-47], or letting parents choose between paper and digital surveys [35].

Reminder Systems

Four protocols explicitly referenced reminder systems [31,32,41-47,60]. To prompt event-contingent entries, 1 protocol used scheduled text reminders [32] and another embedded the event-contingent survey at the start of the interval-contingent prompt (n=1) [41-47]. One protocol triggered follow-up contact after 2 days of missed data [41-47]. Another adjusted prompt volume to minimize disruption—silent in the morning, vibrating during school hours, and audible in the evening [48,50,62]. Additional strategies included follow-up reminders after the main prompt [31,60], sticker tracking sheets [41-47], and regular caregiver check-ins (every 2-3 days) [31].

Support Networks

Eight protocols reported support systems via 3 main groups. Research staff often assisted (eg, hotline, in-person meeting, and scheduled calls) (n=5) [33,34,41-48,50,60,62]. Teaching staff or schools contributed to 4 protocols, more passively by allowing the protocol to occur [38-40,48-50,62] and more actively by monitoring survey completion and response times [38,52-57]. Parents were only formally integrated in an explicit support role in 1 protocol [52,55-57], where they were asked to track their child's EMA completion.

Research question 2: Is it feasible and acceptable to use digital devices to implement an EMA methodology with children between the ages of 5 and 11 years for health purposes?

Research question 2: Is it feasible and acceptable to use digital devices to implement an Ecological Momentary Assessment methodology with children between the ages of 5 and 11 for health purposes?

In total, 11 of the 17 included protocols did not report the overall response rate across all question domains [32-35,43-53,55,62,63]. As a result, the values in Table 4 were calculated from available data, where possible; further details on these calculations can be found in Multimedia Appendix 1.

XSL•FO
RenderX

Table . Summary of feasibility and acceptability outcomes across ecological momentary assessment protocols with children aged 5-11 years (research question 2).

Protocol	Dropout: reason	Exclusion from analysis: criteria	Technical issues	Average response latency in minutes (range)[a]	Child RR[b]	Parent RR
Serious						
Rosen and Epstein (2010) [33]	0%: NR[c]	0%: NR	NR	NR	N/A[d]	91%[e]
Appelhans et al (2025) [34]	NR	NR	NR	NR	N/A	70%
Moschko et al (2022) [35]	30%[f] (between waves): NR	4%: No parent data	3%	5	62%[e]	56%[e]
Vilaysack et al (2016) [31]	NR	NR	NR	1.75 (0.5-3)	48%[e]	N/A
Critical						
Rosen et al (2013)—Main [36]	NR	13%[f]: NR	13%	NR	N/A	87%
Alacha et al (2024) [37]	NR	36%[f]: <5 successive ratings	NR	NR	N/A	86%
Jacobs (2021) [38-40]	NR	NR	NR	6.5 (3-10)	85%[e]	N/A
Rosen et al (2013)—Sub [36]	NR	NR	NR	NR	84%	NR
Family Matters B1 (2020) [41-47]	NR	NR; incomplete survey set. 4%-6% (accelerometer: <4 days and <4 hours per day)	NR	3.5 (2-5)	N/A	80%
SASCHA B2 (2020) [40,48-50,62]	NR	NR	NR	NR	78%[e]	N/A
Engelen, Bundy, Lau et al (2015)[e] [51]	10% (accel): 1 misplaced, 1 NR	NR	5%, up to 20% (accelerometer)	NR	N/A	75%[e]
SASCHA B1 (2020) [48,50,62]	NR	NR	NR	NR	71%[e]	N/A
FLUX (2013) [38,52-57]	4%: NR. 17%[f] (accel): 13 NR, 5 lost, 1 stolen	0%-1%: no paired data. 9% (accel): <4 days wear	2% (accelerometer)	12.5 (10-15)	66%[e]	N/A
Engelen, Bundy, Bauman et al (2015)[f] [58]	2%: includes 1 school absence	13%[f]: “nonvalid data”	NR	NR	N/A	51%
Norman et al (2020) [32]	NR	10%: late daily surveys	NR	NR	N/A	49%[e]
Rovane et al (2024) [59]	NR	NR	NR	NR	N/A	NR
Family Matters B2 (2025) [60]	NR	31%[f]: did not “complete” study days	NR	NR	N/A	NR

[a]Values are the average time in minutes, and values within parentheses are the range, if reported.

[b]RR: response rate.

[c]NR: not reported.

[d]Not applicable.

[e]Calculated from provided study information (Multimedia Appendix 1).

XSL•FO
RenderX

[f]Protocols with >11% dropout were considered significant [22].

Quantitative Data

Exclusion from analysis was the most commonly reported source of attrition (n=8), with criteria including incomplete daily surveys [41-47], nonvalid data exclusions [51], and fewer than 5 successive ratings [37]. Dropout rates ranged from 2% to 30% (n=4) [35,38-40,51-58]. Technical issues contributed to attrition in 4 protocols [35,36,38,51-57].

Fifteen protocols reported (n=6) or enabled calculation of a response rate (n=9), defined as the percentage of planned prompts completed. Six met the high-adherence threshold (≥80%) [33,36-47] and 10 fell below it (48% - 78%). Pooled estimates could be calculated only for 4 protocols using a random-effects model (mean 64.6%, 95% CI 53.6% - 75.5%), reflecting incomplete and inconsistent reporting (Multimedia Appendix 4). Only Moschko et al [35] collected both child and parent data (62% child and 56% parent), limiting direct comparisons between self- and proxy report. Five protocols reported response latency, the time between prompt delivery and participant response, averaging 6 minutes [31,35,38-47,52-57].

Qualitative Data

Only 2 of the 17 protocols [31,32] incorporated formal qualitative interviews, one using manifest analysis of parent interviews [32] and the other summarizing joint child-parent interviews [31], so we supplemented these findings with extracted observations and author reflections from the "Results" and "Discussion" sections of the remaining 13 protocols. Using the work of Norman et al [32] as a foundation, we identified 4 facilitators and 6 barriers (Tables 5 and 6).

Table . Facilitators to acceptability and feasibility reported across protocols (research question 2).

Theme	Details
Uncomplicated (n=3) [31-33]	EMA[a] tools were easy to use or included simplifying features. Familiar tasks (eg, taking photographs) required little explanation, prompts included clear training and visuals, and participation was described as not a major inconvenience.
Engaging (n=2) [31,32]	Participants found EMA tasks enjoyable and reported positive attitudes, with reference to educational benefits for children.
Caregiver support (n=1) [31]	Parents helped children interpret ambiguous prompts, families supported participation by listening for alerts when devices were left in shared spaces and reminding children to respond, and school staff enabled access and ensured audible alerts during school hours.
Reminders (n=3) [31,32,41]	Ongoing researcher contact supported engagement. Norman et al [32] added an end-of-day survey mid-study to address low event-contingent response rates, which was reflected on positively.

[a]EMA: ecological momentary assessment.

XSL•FO
RenderX

Table . Barriers to acceptability and feasibility reported across protocols (research question 2).

Theme	Details
Demanding (n=3) [32,61]	EMA[a] prompts were seen as burdensome due to high effort or intensity, which was linked to dropout; timing (eg, during busy mornings) also contributed to perceived burden.
Challenging to report accurately (n=6) [31,32,47,54,61]	Both child self-reporters and proxy reporters struggled with precise reporting. Children exhibited extreme response bias on Likert or VAS[b] scales and often could not map prompts onto real-world activities, sometimes needing caregiver help, while proxies found it difficult to categorize nonconforming events (eg, buffet-style meals) and potentially skewed responses due to social desirability or guessing when not observing the child.
Device awareness (n=3) [31,32,38,52-57]	Devices were frequently forgotten, left uncharged, or misplaced. Norman et al [32] found that proxy-reporting parents often overlooked event-contingent prompts, compromising data completeness.
Device access (n=3) [[31,36]	Children sometimes lost physical access to devices due to extracurricular commitments, restrictive school policies, examinations, or institutional rules; caregiver control (eg, muting or storing devices) further limited engagement.
Stigma (n=1) [31]	Using EMA devices in school led to peer questioning and discomfort. In one instance, stigma resulted in a participant discontinuing EMA participation during school hours.
Lack of caregiver support (n=2) [31,36]	Participation decreased when caregivers lacked time to assist children or when teaching staff required children to mute or silence devices during lessons.

[a]EMA: ecological momentary assessment.

[b]VAS: visual analog scale.

We introduce a third category, mitigators, based on participant and author suggestions for addressing barriers. Only Vilaysack et al [31] reported participant feedback, which included improving notification sounds and providing a carrying case for the EMA collection device. The authors agreed with these suggestions, recommending louder, longer alerts and training children to adjust volume settings or automating them to suit school environments [31]. A carrying bag was also recommended for future use [31].

Three protocols recommended including additional reporters including the child (n=1) [37], primary caregivers (n=1) [38,52-57], secondary caregivers (n=1) [41-47], and teachers (n=1) [37]. Alternative strategies to improve data collection, particularly in reference to children's difficulties expressing their experiences, included combining EMA with qualitative methods (n=3) [37,41-47,60]; using sensor data to add contextual information (n=3) [35,38,52-57,60]; and extending the monitoring period to account for inconsistencies in responses (n=3) [41-47,51,58]. Authors also emphasized the need for training children and parents to express experiences effectively, particularly in mapping experiences to scales (n=1) [31], and acquiring subject knowledge (n=1) [41-47]. Two protocols recommended greater customization, including adapting data collection for participants whose first language differed from that of the researchers or protocol [41-47], offering optional custom reminders [58].

Research question 3: What characteristics of the EMA protocol (RQ1) are related to high and low response rate (RQ2) when using EMA with children between the ages of 5 and 11 years?

Of the protocols with a reported or calculated response rate (n=15/17), high- and low response rate characteristics are summarized in Table 7. However, as the majority of protocols were judged at critical risk of bias, these comparisons should be interpreted with caution. High response rate protocols (n=6) [33,36-47] predominantly recruited older KS2 children with ADHD and relied on parent report. In contrast, low response rate protocols (n=9) [31,32,35,38,40,48-58,62] were evenly split between KS1 and KS2, typically included children without identified health conditions, and more commonly used child self-report. Both groups used verbal instruction, but hands-on practice and written instruction were more common in low-response protocols. Both groups also used handheld devices, although smartphones specifically were more common in low response rate protocols [33-35,61,63,64]. Monitoring periods differed, with longer durations in high-response protocols (≥3 weeks) and shorter durations (<3 weeks) in low-response protocols.

XSL•FO
RenderX

Table . Comparison of high and low response rate ecological momentary assessment protocols with children aged 5-11 years, showing majority design features within each group.

Characteristic	High response rate	Low response rate
Demographics	• KS1[a] children (n=1/6), • KS2[b] children (n=5/6)	• KS1 children (n=5/9), • KS2 children (n=5/9)
	• ADHD[c] (n=4/6)	• No condition (n=8/9)
	• Children reporting (n=2/6) • Parents reporting (n=5/6)	• Children reporting (n=5/9) • Parents reporting (n=5/9)
Training	• Verbal instruction (n=4/6) • Written instruction (n=1/6) • Practice for reporter (n=3/6)	• Verbal instruction (n=6/9) • Written instruction (n=4/9) • Practice for reporter (n=6/9)
Technology	• Smartphone (n=2/6)	• Smartphone (n=7/9)
Monitoring period	• ≥3 weeks (n=4/6)	• ≥3 weeks (n=3/9)
	• <3 weeks (n=2/6)	• <3 weeks (n=6/9)
Prompt frequency	• 3 prompts per day (n=4/6)	• 3-4 prompts per day (n=6/9)
Prompting design	Interval-contingent (n=6/6) • Random (n=1/6) • Fixed (n=5/6)	Interval-contingent (n=9/9) • Random (n=5/9) • Fixed (n=2/9)
	• Prompts before school, after school, and evening (n=5/6)	• Prompts before school, after school, and evening (n=2/9)
	• ≤12 items per prompt (n=2/6)	• ≤12 items per prompt (n=6/9)
	• >13 items per prompt (n=4/6)	• >13 items per prompt (n=3/9)
	• Use of scales for responses (n=6/6)	• Use of scales for responses (n=7/9)
	• Multiple response types in 1 prompt (n=3/6)	• Multiple response types in 1 prompt (n=7/9)
	• ≤1 hour to respond to prompt (n=1/6)	• ≤1 hour to respond to prompt (n=6/9)
Design features	• Monetary reward (n=4/6)	• Monetary reward (n=5/9)
	• Prompt timing customization (n=6/6)	• Prompt timing customization (n=3/6)

[a]KS1: key stage 1.

[b]KS2: key stage 2.

[c]ADHD: attention-deficit/hyperactivity disorder.

All protocols used interval-contingent sampling with 3-4 prompts each day, and high response rate protocols predominantly used fixed schedules, typically avoiding school hours (n=4/6) by prompting before school, after school, and in the evening. In contrast, low response rate protocols used random schedules. Based on the maximum number of items used in a single prompt within each protocol, high-response protocols typically included 20 or more items (n=4/6), while low-response protocols more often included 12 or fewer items (n=5/8). Both groups used scale-based data collection methods, but low-response protocols more frequently combined multiple response formats within a single prompt and imposed shorter response windows of 1 hour or less. In the high response rate group, most protocols offered monetary incentives and all allowed participants to customize prompt timing; by contrast, in the low-response group, incentives were less consistently used and prompt timing customization was uncommon.

Discussion

Principal Findings

This review aimed to answer 3 RQs by identifying the characteristics of EMA protocols used with children aged 5-11 years (RQ1), the feasibility and acceptability of EMA protocols for children aged 5-11 years (RQ2), and the protocol characteristics linked to high and low response rates (RQ3).

XSL•FO
RenderX

Most protocols used handheld devices, interval-contingent prompting, and Likert scales, with training usually verbal and practice-based to collect self- or parent-reported data across monitoring periods of 3-28 days (RQ1). Feasibility and acceptability remain difficult to assess due to heterogeneous reporting and that 13 of 17 protocols were rated at critical risk of bias using the ROBINS-I tool. Key data required for pooling (eg, raw counts for planned vs completed prompts and corresponding variance estimates) were missing or selectively reported, preventing meaningful quantitative synthesis. Reported facilitators of acceptability and feasibility included the protocol being uncomplicated and engaging, the use of reminders, and caregiver support. Barriers included device access issues, reporter difficulty reporting accurately, reporting burden, stigma, lack of protocol awareness, and insufficient caregiver involvement. Suggested future mitigations included improvements to reminder systems, carrying aids, longer monitoring periods, increased customization, passive data collection, and involving additional reporters or methods (RQ2). High-response protocols (n=6, ≥80%) [20] more often involved older children (KS2), those with specific health conditions (ADHD), and featured longer monitoring periods (≥3 weeks). These protocols typically used fixed, interval-contingent prompt schedules, with greater than 20 items and prompting 3 times a day, often before school, after school, and in the evening. The majority also offered limited customization of prompt timing and provided monetary reimbursement for participation. Given the high risk of bias across most protocols, these contrasts should be interpreted as tentative (RQ3).

Comparison With Prior Work

Evidence on the feasibility and acceptability of EMA protocols for preadolescent children remains limited, and how EMA protocol design can best support these outcomes is unclear [22-26]. Despite this, there is a growing recognition that protocols should be adapted for this age group [23,24,26]. To our knowledge, this is the first review to systematically examine the feasibility and acceptability of EMA protocols for preadolescent children. While challenges such as technical issues and participant burden are common across age groups [1,19,65], this review identified additional factors that appear especially important for preadolescents and their caregivers acting as proxy reporters, including difficulties responding to prompts accurately, inconsistent access to reporting devices, and device-related stigma. To explore the impact of protocol characteristics on acceptability and feasibility, we apply the Technology Acceptance Model, which identifies perceived ease of use and perceived usefulness as key influences on technology adoption [66]. Using this framework helps distinguish between protocol features that may have supported acceptability, those that produced unexpected patterns, and those that created barriers to technology acceptability.

Several features of high response rate protocols appeared to align with the Technology Acceptance Model's concept of perceived ease of use and usefulness, potentially making participation more acceptable. Predictable prompting schedules (eg, before school, after school, and evening), opportunities to customize timing, and the avoidance of in-school prompts may have helped reduce disruption (eg, limiting negative teacher interference) and make participation feel easier to manage within family routines [31,67]. The use of familiar technologies to caregivers, uncomplicated interfaces, encouraging caregiver support in responding to prompts and simple visuals may have supported participation by reducing cognitive and knowledge barriers, thereby improving perceived ease of use [68]. Perceived usefulness may have been reinforced through external motivators such as monetary incentives and through the involvement of caregivers, including school staff, who facilitated tasks [69-72], and the framing of EMA tasks as enjoyable or educational [73].

Other features associated with high response rates, however, diverged from common EMA design practices and therefore require more cautious interpretation. Adult EMA literature typically favors shorter monitoring periods (≤2 weeks) [23,25] and brief prompts to reduce participant burden [16]. In contrast, in this review, longer periods (≥3 weeks) and extensive prompts (≥20 items per prompt) were present in the majority of high response rate protocols. One explanation may be that longer durations help embed EMA into daily routines, with the stability offsetting potential fatigue [1,22]. However, these more demanding protocols also tended to provide stronger external supports, such as closer researcher involvement or monetary incentives, suggesting that response rates may have reflected a combination of routine and additional resources rather than duration or length alone. Training showed a similar counterintuitive pattern: hands-on practice was more common in low response rate protocols, indicating that practice is more important for supporting valid and accurate reporting than for increasing response rates [74]. A further unexpected pattern was the higher response rates linked to children with ADHD. Although this might initially appear counterintuitive, children with long-term conditions may be more intrinsically motivated to engage, as they are more likely to recognize a personal need that technology could help address [75]. In contrast, children without a long-term health condition may not perceive the same relevance or potential benefit.

By contrast, several protocol features reduced perceived ease of use, reflecting both design complexity and developmental challenges. For example, nearly all protocols relied on response formats such as Likert or visual analog scales, despite evidence that alternative formats, such as item ranking and semantic differential scales, may be more developmentally appropriate for younger children [76-78]. The continued reliance on such scales highlights a broader adult-centered design approach that may overlook children's cognitive abilities and real-world reporting capacities [79]. Furthermore, many 5 - 7 year olds may struggle with reading, reinforcing the need for adult support in interpreting prompts [80,81]. Yet, reliance on caregivers may overlook the developmental trajectory of literacy: children who already have the skills may be denied opportunities to practice reporting independently, while those still acquiring literacy may be excluded altogether.

Beyond developmental challenges, logistical and practical barriers also reduced ease of use. More complex protocols, such as those combining multiple response formats or enforcing narrow response windows (<1 hour), may have placed additional cognitive and logistical demands on children and caregivers [82,83]. These challenges are especially important, given the

XSL•FO
RenderX

documented difficulty, in both this review and wider literature, that children face with extreme response bias and mapping their experiences onto structured input formats [80]. Furthermore, the lack of flexibility beyond initial prompt timing customization contrasts with adult EMA, where participants can sometimes delay prompts or choose alternative response formats to reduce burden [84,85]. For example, 1 protocol using photo-based meal tracking struggled to capture buffet-style eating, illustrating how rigid input formats can fail to reflect real-world variability [32]. While this is a known limitation for adults [1], it may be amplified in children, as caregivers may resist disrupting routines and children may see little benefit in adjusting their behavior to enable easier data capture [67,86]. Additionally, lack of device awareness contributed to issues such as forgetting to charge it or bring it with them, disrupting participation [23]. These challenges may be greater for preadolescents, who may not own devices suitable for EMA and are instead given separate ones, making routine integration more difficult [23,30]. Using wearables, which remain on the body, may help reduce this burden [16,87]; however, no protocol in this review used such an approach. Three protocols recommended incorporating sensor data to contextualize and support children's self-reports, suggesting a potential role for wearables in future protocols [88].

Perceived usefulness was often less prioritized than ease of use. While many protocols included monetary incentives, the included protocols did not report how they established personal relevance for either children or caregivers. Emphasizing the value of participation may be particularly important when EMA is used solely as a measurement tool, as was the case in all included protocols, to avoid participants feeling that they contribute data without clear benefits or autonomy [8,89]. In addition, preadolescent children may be better motivated by immediate rewards [90], yet only 1 protocol in this review included a daily reward system [41-47]. Social stigma also shaped perceived usefulness. In 1 protocol, preadolescents reported attracting unwanted attention when using a smartphone at school [31]. This highlights how adult assumptions about appropriate technologies can overlook that many preadolescents lack regular access to smartphones [91,92], which may contribute to feelings of unfamiliarity or stigma. Few protocols in this review involved children directly in their development (eg, piloting), limiting opportunities to identify usability issues and integrate children's perspectives [93,94].

Preadolescents are embedded in a shifting network of caregivers, teachers, peers, and wider family [95], yet many protocols included in this review treat them as isolated participants and do not explicitly involve caregivers in supporting roles. While caregivers were often given verbal or written instructions when the child was the reporter, protocols rarely included opportunities for them to practice using the device. School cooperation was also limited, restricting access to essential adult support. Furthermore, simultaneous tracking by children and caregivers was rare, although others recommended incorporating multiple perspectives in future work [37,38,41-47,52-57]. Such approaches may offer benefits including shared responsibility, enhanced motivation, and enriched data perspectives [96,97]. Preadolescents rely heavily on adults and their peers to interpret, manage, and respond to daily tasks [98,99], making limited involvement of these networks in protocol design both a missed opportunity and a barrier to engagement [31]. Protocols that acknowledge and actively involve this network, through co-design, piloting, and contextual adaptation, are more likely to avoid common pitfalls and create experiences that are not only feasible but also genuinely valuable to children and those who support them [100-102], although this must be balanced against the risk that proxy reporting shifts the focus away from children's voices. In KS1 protocols, for example, parents often acted as reporters, sometimes entering responses in dialogue with their child, but in other cases responding based only on their own observations. The latter approach arguably reflects EMA of parents about their children, rather than EMA of children themselves, raising questions of validity and comparability with adult EMA where self-report is standard. These issues highlight the importance and complexity of situating EMA within children's lives, emphasizing the need for EMA protocols that fit within children's everyday social and caregiving contexts, ensuring that both children and their caregivers can meaningfully engage with and benefit from these tools [70,103].

Strengths and Limitations

Digital health is an interdisciplinary field, and a key strength of this review is the interdisciplinary approach taken by searching a large number of databases (n=10), spanning both health science databases and HCI venues. The review aimed to investigate EMA protocols used with children (aged 5 - 11 years) to understand developmentally sensitive design implications. To investigate this, we looked across the broad spectrum of child health behaviors and conditions, which means our review is not limited to any specific health domain, making the findings relevant to a broader scope of researchers working with pediatric EMA protocols. However, due to heterogeneity and lack of reporting, a formal meta-analysis of quantitative data nor metaethnography of qualitative data was feasible, and we instead used a narrative review approach [61] to identify patterns that may inform future hypothesis generation. We also excluded studies that included our target age group but did not report specific protocol adaptations for younger children (eg, studies covering ages 5 - 18 years without age-specific design considerations). The use of narrative synthesis, while appropriate given the diversity of study design, has been critiqued for limited transparency [104]. Additionally, the scarcity of detailed qualitative and quantitative data restricted the depth of our analysis. This limitation highlights a need for future research exploring child and caregiver views to improve EMA protocols. A further limitation is that most protocols were conducted in Western, high-income contexts, with little reporting of socioeconomic background and inconsistent reporting of ethnicity. Where provided, samples were often predominantly White, with only a few protocols including more diverse populations. Future research should prioritize more diverse samples and clearer demographic reporting.

Conclusions

This review provides the first systematic evidence base focused exclusively on digital EMA with preadolescent children. It

XSL•FO
RenderX

examined 17 digital EMA protocols involving children aged 5 - 11 years, highlighting gaps in developmental appropriateness (eg, absence of child-focused piloting) and inconsistencies in reporting quality that limit both interpretability and comparability across studies.

In contrast to existing reviews that primarily emphasize adherence or feasibility in children of different ages as a single group and adults, these findings emphasize the importance of preadolescent acceptance and developmental considerations when EMA is used with this specific age group. Barriers included device access issues, finding it challenging to report accurately (eg, response options not matching how they want to express themselves), reporting burden, stigma, lack of protocol awareness, and insufficient caregiver involvement. Facilitators included uncomplicated, engaging technology, reminders, and caregiver involvement. Several counterintuitive patterns also emerged: protocols with longer durations and more items per prompt were linked to high response rates, further highlighting the importance of considering preadolescents distinctly.

The review contributes to the field by consolidating the evidence base and identifying protocol characteristics, while also highlighting the need for improved and more consistent reporting of feasibility, acceptability, and response metrics. Without such improvements, meaningful comparison across protocols and cumulative knowledge building remains limited.

From a practical perspective, the findings suggest that future digital EMA research should focus on perceived ease of use (eg, predictable prompting schedules, simplified response formats, and flexibility in fitting daily routines) and perceived usefulness (eg, immediate rewards, personally relevant activities, clear explanations of purpose, and addressing stigma) as part of a more child-centered design approach. Given children's dependence on caregivers and teachers, involving these adults is likely to support perceived ease of use (eg, assistance with prompts, device availability, and charging) and perceived usefulness (eg, clarifying relevance, reinforcing engagement, and managing social dynamics). With greater developmental alignment and improved reporting standards, digital EMA could be more effectively integrated into pediatric health monitoring in ways that are sensitive to the needs of different age groups.

Acknowledgments

The authors declare the use of generative artificial intelligence in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing. The GAI tool used was Grammarly, ChatGPT-4.0. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. A declaration was submitted by SC.

Funding

This work was supported by the EPSRC Digital Health and Care Centre for Doctoral Training (UKRI grant EP/S023704/1).

Data Availability

The data extracted and analyzed during this review are available in the supplementary files of this manuscript (Checklists 1 and 2 and Multimedia Appendices 1-4), with the full data extraction tables shown in Multimedia Appendix 5.

Authors' Contributions

SC conceptualized this work in collaboration with the coauthors. SC carried out the search strategy and conducted the article screening with LT. SC performed the quality assessment. SC wrote the initial manuscript. LT, AB, and JB reviewed, refined, and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full search strategy.
[DOCX File, 18 KB - jmir_v28i1e79291_app1.docx]

Multimedia Appendix 2
Quality assessment including CREMAS (Checklist for Reporting EMA Studies) and ROBINS-I v2 (Risk Of Bias In Nonrandomized Studies of Interventions, version 2).
[XLSX File, 25 KB - jmir_v28i1e79291_app2.xlsx]

Multimedia Appendix 3
Study-level overviews.
[DOCX File, 151 KB - jmir_v28i1e79291_app3.docx]

XSL•FO
RenderX

Multimedia Appendix 4
Response rate calculations.
[XLSX File, 18 KB - jmir_v28i1e79291_app4.xlsx]

Multimedia Appendix 5
Data extraction tables (full).
[XLSX File, 29 KB - jmir_v28i1e79291_app5.xlsx]

Checklist 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File, 214 KB - jmir_v28i1e79291_app6.docx]

Checklist 2
PRISMA-S (An Extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews) checklist.
[DOCX File, 12 KB - jmir_v28i1e79291_app7.docx]

References

1. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
2. Hartson KR, Huntington-Moskos L, Sears CG, et al. Use of electronic ecological momentary assessment methodologies in physical activity, sedentary behavior, and sleep research in young adults: systematic review. J Med Internet Res 2023 Jun 29;25:e46783. [doi: 10.2196/46783] [Medline: 37384367]
3. Overton M, Ward S, Swain N, et al. Are ecological momentary assessments of pain valid and reliable? A systematic review and meta-analysis. Clin J Pain 2023 Jan 1;39(1):29-40. [doi: 10.1097/AJP.0000000000001084] [Medline: 36524770]
4. Gratch I, Choo TH, Galfalvy H, et al. Detecting suicidal thoughts: the power of ecological momentary assessment. Depress Anxiety 2021 Jan;38(1):8-16. [doi: 10.1002/da.23043] [Medline: 32442349]
5. Chan G, Alslaity A, Wilson R, Orji R. Exploring variance in users' moods across times, seasons, and activities: a longitudinal analysis. Presented at: Adjunct Publication of the 24th International Conference on Human-Computer Interaction with Mobile Devices and Services; Sep 28, 2022; Vancouver BC Canada p. 1-7. [doi: 10.1145/3528575.3551444]
6. Sherriff B, Clark C, Killingback C, Newell D. Impact of contextual factors on patient outcomes following conservative low back pain treatment: systematic review. Chiropr Man Therap 2022 Apr 21;30(1):20. [doi: 10.1186/s12998-022-00430-8] [Medline: 35449074]
7. Gao L, Sun Y, Pan L, et al. Current status and influencing factors of fatigue in patients with rheumatoid arthritis: a cross-sectional study in China. Int J Nurs Pract 2022 Jun;28(3):e12996. [doi: 10.1111/ijn.12996] [Medline: 34309127]
8. Dao KP, De Cocker K, Tong HL, Kocaballi AB, Chow C, Laranjo L. Smartphone-delivered ecological momentary interventions based on ecological momentary assessments to promote health behaviors: systematic review and adapted checklist for reporting ecological momentary assessment and intervention studies. JMIR Mhealth Uhealth 2021 Nov 19;9(11):e22890. [doi: 10.2196/22890] [Medline: 34806995]
9. Huckins JF, daSilva AW, Wang W, et al. Mental health and behavior of college students during the early phases of the COVID-19 pandemic: longitudinal smartphone and ecological momentary assessment study. J Med Internet Res 2020 Jun 17;22(6):e20185. [doi: 10.2196/20185] [Medline: 32519963]
10. Doherty K, Balaskas A, Doherty G. The design of ecological momentary assessment technologies. Interact Comput 2020 May 17;32(3):257-278. [doi: 10.1093/iwcomp/iwaa019]
11. Salles A, Ais J, Semelman M, Sigman M, Calero CI. The metacognitive abilities of children and adults. Cogn Dev 2016 Oct;40:101-110. [doi: 10.1016/j.cogdev.2016.08.009]
12. Thunnissen MR, Aan Het Rot M, van den Hoofdakker BJ, Nauta MH. Youth psychopathology in daily life: systematically reviewed characteristics and potentials of ecological momentary assessment applications. Child Psychiatry Hum Dev 2022 Dec;53(6):1129-1147. [doi: 10.1007/s10578-021-01177-8] [Medline: 34075490]
13. Forsberg A, Guitard D, Adams EJ, Pattanakul D, Cowan N. Children's long-term retention is directly constrained by their working memory capacity limitations. Dev Sci 2022 Mar;25(2):e13164. [doi: 10.1111/desc.13164] [Medline: 34328244]
14. Wiljén A, Chaplin JE, Crine V, et al. The development of an mHealth tool for children with long-term illness to enable person-centered communication: user-centered design approach. JMIR Pediatr Parent 2022 Mar 8;5(1):e30364. [doi: 10.2196/30364] [Medline: 35258466]
15. McCallum C, Rooksby J, Gray CM. Evaluating the impact of physical activity apps and wearables: interdisciplinary review. JMIR Mhealth Uhealth 2018 Mar 23;6(3):e58. [doi: 10.2196/mhealth.9054] [Medline: 29572200]
16. Intille S, Haynes C, Maniar D, Ponnada A, Manjourides J. µEMA: microinteraction-based ecological momentary assessment (EMA) using a smartwatch. Proc ACM Int Conf Ubiquitous Comput 2016 Sep;2016:1124-1128. [doi: 10.1145/2971648.2971717] [Medline: 30238088]

XSL•FO
RenderX

17. Aminikhanghahi S, Schmitter-Edgecombe M, Cook DJ. Context-aware delivery of ecological momentary assessment. IEEE J Biomed Health Inform 2020 Apr;24(4):1206-1214. [doi: 10.1109/JBHI.2019.2937116] [Medline: 31443058]
18. Kumar D, Haag D, Blechert J, Niebauer J, Smeddinck JD. Feature selection for physical activity prediction using ecological momentary assessments to personalize intervention timing: longitudinal observational study. JMIR Mhealth Uhealth 2025;13:e57255-e57255. [doi: 10.2196/57255]
19. Williams MT, Lewthwaite H, Fraysse F, Gajewska A, Ignatavicius J, Ferrar K. Compliance with mobile ecological momentary assessment of self-reported health-related behaviors and psychological constructs in adults: systematic review and meta-analysis. J Med Internet Res 2021 Mar 3;23(3):e17023. [doi: 10.2196/17023] [Medline: 33656451]
20. Stone AA, Shiffman S. Capturing momentary, self-report data: a proposal for reporting guidelines. Ann Behav Med 2002;24(3):236-243. [doi: 10.1207/S15324796ABM2403_09] [Medline: 12173681]
21. Businelle MS, Hébert ET, Shi D, et al. Investigating best practices for ecological momentary assessment: nationwide factorial experiment. J Med Internet Res 2024 Aug 12;26:e50275. [doi: 10.2196/50275] [Medline: 39133915]
22. Wrzus C, Neubauer AB. Ecological momentary assessment: a meta-analysis on designs, samples, and compliance across research fields. Assessment 2023 Apr;30(3):825-846. [doi: 10.1177/10731911211067538] [Medline: 35016567]
23. Wen CKF, Schneider S, Stone AA, Spruijt-Metz D. Compliance with mobile ecological momentary assessment protocols in children and adolescents: a systematic review and meta-analysis. J Med Internet Res 2017 Apr 26;19(4):e132. [doi: 10.2196/jmir.6641] [Medline: 28446418]
24. Heron KE, Everhart RS, McHale SM, Smyth JM. Using mobile-technology-based ecological momentary assessment (EMA) methods with youth: a systematic review and recommendations. J Pediatr Psychol 2017 Nov 1;42(10):1087-1107. [doi: 10.1093/jpepsy/jsx078] [Medline: 28475765]
25. Liao Y, Skelton K, Dunton G, Bruening M. A systematic review of methods and procedures used in ecological momentary assessments of diet and physical activity research in youth: an adapted STROBE checklist for reporting EMA studies (CREMAS). J Med Internet Res 2016 Jun 21;18(6):e151. [doi: 10.2196/jmir.4954] [Medline: 27328833]
26. Mason TB, Do B, Wang S, Dunton GF. Ecological momentary assessment of eating and dietary intake behaviors in children and adolescents: a systematic review of the literature. Appetite 2020 Jan 1;144:104465. [doi: 10.1016/j.appet.2019.104465] [Medline: 31541670]
27. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021:n71. [doi: 10.1136/bmj.n71]
28. Yang YS, Ryu GW, Choi M. Methodological strategies for ecological momentary assessment to evaluate mood and stress in adult patients using mobile phones: systematic review. JMIR Mhealth Uhealth 2019 Apr 1;7(4):e11215. [doi: 10.2196/11215] [Medline: 30932866]
29. Boluyt N, Tjosvold L, Lefebvre C, Klassen TP, Offringa M. Usefulness of systematic review search strategies in finding child health systematic reviews in MEDLINE. Arch Pediatr Adolesc Med 2008 Feb;162(2):111-116. [doi: 10.1001/archpediatrics.2007.40] [Medline: 18250233]
30. de Vries LP, Baselmans BML, Bartels M. Smartphone-based ecological momentary assessment of well-being: a systematic review and recommendations for future studies. J Happiness Stud 2021;22(5):2361-2408. [doi: 10.1007/s10902-020-00324-7] [Medline: 34720691]
31. Vilaysack B, Cordier R, Doma K, Chen YW. Capturing everyday experiences of typically developing children aged five to seven years: a feasibility study of experience sampling methodology. Aust Occup Ther J 2016 Dec;63(6):424-433. [doi: 10.1111/1440-1630.12336] [Medline: 27723103]
32. Norman Å, Kjellenberg K, Torres Aréchiga D, Löf M, Patterson E. “Everyone can take photos.” Feasibility and relative validity of phone photography-based assessment of children’s diets - a mixed methods study. Nutr J 2020 May 27;19(1):50. [doi: 10.1186/s12937-020-00558-4] [Medline: 32460760]
33. Rosen PJ, Epstein JN. A pilot study of ecological momentary assessment of emotion dysregulation in children: Journal of ADHD and Related Disorders; 2010, Vol. 1:39-52 URL: https://www.researchgate.net/publication/303157300_A_pilot_study_of_ecological_momentary_assessment_of_emotion_dysregulation_in_children [accessed 2025-11-01]
34. Appelhans BM, Bleil ME, Crane MM. Leveraging recreational activities to reduce obesity-related behaviors in children from lower-income households. Appetite 2025 Oct 1;214:108171. [doi: 10.1016/j.appet.2025.108171] [Medline: 40473025]
35. Moschko T, Stadler G, Gawrilow C. Fluctuations in children’s self-regulation and parent-child interaction in everyday life: an ambulatory assessment study. J Soc Pers Relat 2023 Jan;40(1):254-276. [doi: 10.1177/02654075221116788]
36. Rosen PJ, Epstein JN, Van Orden G. I know it when I quantify it: ecological momentary assessment and recurrence quantification analysis of emotion dysregulation in children with ADHD. Atten Defic Hyperact Disord 2013 Sep;5(3):283-294. [doi: 10.1007/s12402-013-0101-2] [Medline: 23338519]
37. Alacha HF, Rosen PJ, Bufferd SJ. Positive affect variability is associated with homework management difficulties in children with ADHD. J Child Fam Stud 2024 Sep;33(9):2933-2946. [doi: 10.1007/s10826-024-02852-9]
38. Kramer AC, Neubauer AB, Leonhardt A, Brose A, Dirk J, Schmiedek F. Ambulatory assessment of rumination and worry: capturing perseverative cognitions in children’s daily life. Psychol Assess 2021 Sep;33(9):827-842. [doi: 10.1037/pas0001020] [Medline: 33970655]

XSL•FO
RenderX

39. Neubauer AB, Kramer AC, Schmiedek F. Assessing domain-general need fulfillment in children and adults: introducing the General Need Satisfaction and Frustration Scale. Psychol Assess 2022 Nov;34(11):1022-1035. [doi: 10.1037/pas0001169] [Medline: 36074610]
40. Neubauer AB, Kramer AC, Schmidt A, Könen T, Dirk J, Schmiedek F. Reciprocal relations of subjective sleep quality and affective well-being in late childhood. Dev Psychol 2021 Aug;57(8):1372-1386. [doi: 10.1037/dev0001209] [Medline: 34591579]
41. Trofholz A, Tate A, Janowiec M, et al. Ecological momentary assessment of weight-related behaviors in the home environment of children from low-income and racially and ethnically diverse households: development and usability study. JMIR Res Protoc 2021 Dec 1;10(12):e30525. [doi: 10.2196/30525] [Medline: 34855612]
42. Tate A, Telke S, Trofholz A, Miner M, Berge JM. Stressed Out! Examining family meal decisions in response to daily stressors via ecological momentary assessment in a racially/ethnically diverse population. Prev Med Rep 2020 Dec;20:101251. [doi: 10.1016/j.pmedr.2020.101251] [Medline: 33344147]
43. Berge JM, Fertig AR, Trofholz A, Neumark-Sztainer D, Rogers E, Loth K. Associations between parental stress, parent feeding practices, and child eating behaviors within the context of food insecurity. Prev Med Rep 2020 Sep;19:101146. [doi: 10.1016/j.pmedr.2020.101146] [Medline: 32642403]
44. Loth KA, Fertig A, Trofholz A, et al. Concordance of children's intake of selected food groups as reported by parents via 24-h dietary recall and ecological momentary assessment. Public Health Nutr 2021 Jan;24(1):22-33. [doi: 10.1017/S1368980020001111] [Medline: 32576304]
45. Loth KA, Tate AD, Trofholz A, et al. Ecological momentary assessment of the snacking environments of children from racially/ethnically diverse households. Appetite 2020 Feb 1;145:104497. [doi: 10.1016/j.appet.2019.104497] [Medline: 31669580]
46. de Brito JN, Loth KA, Tate A, Berge JM. Associations between parent self-reported and accelerometer-measured physical activity and sedentary time in children: ecological momentary assessment study. JMIR Mhealth Uhealth 2020 May 19;8(5):e15458. [doi: 10.2196/15458] [Medline: 32348283]
47. Wirthlin R, Linde JA, Trofholz A, Tate A, Loth K, Berge JM. Associations between parent and child physical activity and eating behaviours in a diverse sample: an ecological momentary assessment study. Public Health Nutr 2020 Oct;23(15):2728-2736. [doi: 10.1017/S136898002000052X] [Medline: 32539887]
48. Blume F, Irmer A, Dirk J, Schmiedek F. Day-to-day variation in students' academic success: the role of self-regulation, working memory, and achievement goals. Dev Sci 2022 Nov;25(6):e13301. [doi: 10.1111/desc.13301] [Medline: 35780513]
49. Galeano-Keiner EM, Neubauer AB, Irmer A, Schmiedek F. Daily fluctuations in children's working memory accuracy and precision: variability at multiple time scales and links to daily sleep behavior and fluid intelligence. Cogn Dev 2022 Oct;64:101260. [doi: 10.1016/j.cogdev.2022.101260]
50. Schmidt A, Neubauer AB, Dirk J, Schmiedek F. The bright and the dark side of peer relationships: differential effects of relatedness satisfaction and frustration at school on affective well-being in children's daily lives. Dev Psychol 2020 Aug;56(8):1532-1546. [doi: 10.1037/dev0000997] [Medline: 32525325]
51. Engelen L, Bundy AC, Lau J, et al. Understanding patterns of young children's physical activity after school--it's all about context: a cross-sectional study. J Phys Act Health 2015 Mar;12(3):335-339. [doi: 10.1123/jpah.2013-0153] [Medline: 24828415]
52. Könen T, Dirk J, Schmiedek F. Cognitive benefits of last night's sleep: daily variations in children's sleep behavior are related to working memory fluctuations. J Child Psychol Psychiatry 2015 Feb;56(2):171-182. [doi: 10.1111/jcpp.12296] [Medline: 25052368]
53. Könen T, Dirk J, Leonhardt A, Schmiedek F. The interplay between sleep behavior and affect in elementary school children's daily life. J Exp Child Psychol 2016 Oct;150:1-15. [doi: 10.1016/j.jecp.2016.04.003] [Medline: 27236036]
54. Kühnhausen J, Leonhardt A, Dirk J, Schmiedek F. Physical activity and affect in elementary school children's daily lives. Front Psychol 2013;4:456. [doi: 10.3389/fpsyg.2013.00456] [Medline: 23885246]
55. Dirk J, Schmiedek F. Variability in children's working memory is coupled with perceived disturbance: an ambulatory assessment study in the school and out-of-school context. Res Hum Dev 2017 Jul 3;14(3):200-218. [doi: 10.1080/15427609.2017.1340051]
56. Neubauer AB, Dirk J, Schmiedek F. Momentary working memory performance is coupled with different dimensions of affect for different children: a mixture model analysis of ambulatory assessment data. Dev Psychol 2019 Apr;55(4):754-766. [doi: 10.1037/dev0000668] [Medline: 30556707]
57. Schmidt A, Dirk J, Schmiedek F. The importance of peer relatedness at school for affective well‐being in children: between‐ and within‐person associations. Soc Dev 2019 Nov;28(4):873-892. [doi: 10.1111/sode.12379]
58. Engelen L, Bundy AC, Bauman A, Naughton G, Wyver S, Baur L. Young children's after-school activities - there's more to it than screen time: a cross-sectional study of young primary school children. J Phys Act Health 2015 Jan;12(1):8-12. [doi: 10.1123/jpah.2013-0075] [Medline: 24510028]
59. Rovane AK, Hock RM, Yang CH, Hills KJ. Parent facilitation of child emotion regulation in ASD: an ecological momentary assessment study. J Autism Dev Disord 2025 Dec;55(12):4197-4211. [doi: 10.1007/s10803-024-06541-w] [Medline: 39227520]

XSL•FO
RenderX

60. Kracht CL, Tate A, de Brito JN, Trofholz A, Berge JM. Association between parental stress, coping, mood, and subsequent child physical activity and screen-time: an ecological momentary assessment study. BMC Public Health 2025 Feb 22;25(1):729. [doi: 10.1186/s12889-025-21738-z] [Medline: 39987022]
61. Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, et al. Guidance on the Conduct of Narrative Synthesis in Systematic Reviews: A Product From the ESRC Methods Programme Version. 2006. [doi: 10.13140/2.1.1018.4643]
62. Schmidt A, Dirk J, Neubauer AB, Schmiedek F. Evaluating sociometer theory in children's everyday lives: inclusion, but not exclusion by peers at school is related to within-day change in self-esteem. Eur J Pers 2021 Sep;35(5):736-753. [doi: 10.1177/0890207020962328]
63. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008 Jul 10;8:45. [doi: 10.1186/1471-2288-8-45] [Medline: 18616818]
64. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016 Oct 12;355:i4919. [doi: 10.1136/bmj.i4919] [Medline: 27733354]
65. Carpenter RW, Wycoff AM, Trull TJ. Ambulatory assessment. Assessment 2016 Aug;23(4):414-424. [doi: 10.1177/1073191116632341] [Medline: 26887808]
66. Marangunić N, Granić A. Technology acceptance model: a literature review from 1986 to 2013. Univ Access Inf Soc 2015 Mar;14(1):81-95. [doi: 10.1007/s10209-014-0348-1]
67. Selman SB, Dilworth‐Bart JE. Routines and child development: a systematic review. J of Family Theo & Revie 2024 Jun;16(2):272-328 [FREE Full text] [doi: 10.1111/jftr.12549]
68. Xiong C, D'Souza A, El-Khechen-Richandi G, et al. Perceptions of Digital Technology Experiences and Development Among Family Caregivers and Technology Researchers: Qualitative Study. JMIR Form Res 2022 Jan 28;6(1):e19967. [doi: 10.2196/19967] [Medline: 35089150]
69. Abdelazeem B, Abbas KS, Amin MA, et al. The effectiveness of incentives for research participation: a systematic review and meta-analysis of randomized controlled trials. PLoS One 2022;17(4):e0267534. [doi: 10.1371/journal.pone.0267534] [Medline: 35452488]
70. Pina LR, Sien SW, Ward T, Yip JC, Munson SA, Fogarty J, et al. From personal informatics to family informatics: understanding family practices around health monitoring. Presented at: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing; Feb 25, 2017; Portland, Oregon, USA p. 2300-2315. [doi: 10.1145/2998181.2998362]
71. Lowry C, Leonard-Kane R, Gibbs B, Muller LM, Peacock A, Jani A. Teachers: the forgotten health workforce. J R Soc Med 2022 Apr;115(4):133-137. [doi: 10.1177/01410768221085692] [Medline: 35271774]
72. Paquette ET, Palac H, Bair E, et al. The importance of engaging children in research decision-making: a preliminary mixed-methods study. Ethics Hum Res 2020 May;42(3):12-20. [doi: 10.1002/eahr.500049] [Medline: 32421946]
73. Schepers S, Dreessen K, Zaman B. Fun as a user gain in participatory design processes involving children. Presented at: Proceedings of the 17th ACM Conference on Interaction Design and Children; Jun 19, 2018; Trondheim, Norway p. 396-404. [doi: 10.1145/3202185.3202763]
74. Wen CKF, Junghaenel DU, Newman DB, et al. The effect of training on participant adherence with a reporting time frame for momentary subjective experiences in ecological momentary assessment: cognitive interview study. JMIR Form Res 2021 May 26;5(5):e28007. [doi: 10.2196/28007] [Medline: 34037524]
75. Anastasi JK, Capili B, Norton M, McMahon DJ, Marder K. Recruitment and retention of clinical trial participants: understanding motivations of patients with chronic pain and other populations. Front Pain Res (Lausanne) 2023;4:1330937. [doi: 10.3389/fpain.2023.1330937] [Medline: 38606348]
76. Burton N, Burton M, Fisher C, Peña PG, Rhodes G, Ewing L. Beyond Likert ratings: improving the robustness of developmental research measurement using best-worst scaling. Behav Res Methods 2021 Oct;53(5):2273-2279. [doi: 10.3758/s13428-021-01566-w] [Medline: 33821456]
77. van Laerhoven H, van der Zaag-Loonen HJ, Derkx BHF. A comparison of Likert scale and visual analogue scales as response options in children's questionnaires. Acta Paediatr 2004 Jun;93(6):830-835. [doi: 10.1080/08035250410026572] [Medline: 15244235]
78. Rocereto JF, Puzakova M, Anderson RE, Kwak H. The role of response formats on extreme response style: a case of Likert-type vs. semantic differential scales. In: Sarstedt M, Schwaiger M, Taylor CR, editors. Emerald Group Publishing Limited 2011:53-71. [doi: 10.1108/S1474-7979(2011)0000022006]
79. Mellor D, Moore KA. The use of Likert scales with children. J Pediatr Psychol 2014 Apr;39(3):369-379. [doi: 10.1093/jpepsy/jst079] [Medline: 24163438]
80. Chambers CT, Johnston C. Developmental differences in children's use of rating scales. J Pediatr Psychol 2002;27(1):27-36. [doi: 10.1093/jpepsy/27.1.27] [Medline: 11726677]
81. Manu M, Torppa M, Vasalampi K, Lerkkanen M, Poikkeus A, Niemi P. Reading development from kindergarten to age 18: the role of gender and parental education. Read Res Q 2023 Oct;58(4):505-538. [doi: 10.1002/rrq.518]
82. Barrouillet P, Camos V. As time goes by: temporal constraints in working memory. Curr Dir Psychol Sci 2012;21:413-419. [doi: 10.1177/0963721412459513]

XSL•FO
RenderX

83. Ahmed SF, Ellis A, Ward KP, Chaku N, Davis-Kean PE. Working memory development from early childhood to adolescence using two nationally representative samples. Dev Psychol 2022 Oct;58(10):1962-1973. [doi: 10.1037/dev0001396] [Medline: 35771499]
84. Stone AA, Schneider S, Smyth JM. Evaluation of pressing issues in ecological momentary assessment. Annu Rev Clin Psychol 2023 May 9;19(1):107-131. [doi: 10.1146/annurev-clinpsy-080921-083128] [Medline: 36475718]
85. Prochnow T, Wang WL, Wang S, et al. Understanding longitudinal ecological momentary assessment completion: results from 12 months of burst sampling in the TIME Study. JMIR Mhealth Uhealth 2025 Oct 22;13:e67117. [doi: 10.2196/67117] [Medline: 41124691]
86. Kuczynski L, Burke T, Song-Choi P. Mothers' perspectives on resistance and defiance in middle childhood: promoting autonomy and social skill. Soc Sci (Basel) 2021;10(12):469. [doi: 10.3390/socsci10120469]
87. Dey AK, Wac K, Ferreira D, Tassini K, Hong JH, Ramos J. Getting closer: an empirical investigation of the proximity of user to their smart phones. Presented at: Proceedings of the 13th International Conference on Ubiquitous Computing; Sep 17, 2011; Beijing, China p. 163-172. [doi: 10.1145/2030112.2030135]
88. Thompson L, Charitos S, Bird J, Marshall P, Brigden A. Exploring the use of smartwatches and activity trackers for health-related purposes for children aged 5 to 11 years: systematic review. J Med Internet Res 2025 Jan 27;27:e62944. [doi: 10.2196/62944] [Medline: 39870369]
89. Druin A. The role of children in the design of new technology. Behav Inf Technol 2002 Jan;21(1):1-25. [doi: 10.1080/01449290110108659]
90. Steinberg L, Graham S, O'Brien L, Woolard J, Cauffman E, Banich M. Age differences in future orientation and delay discounting. Child Dev 2009;80(1):28-44. [doi: 10.1111/j.1467-8624.2008.01244.x] [Medline: 19236391]
91. Richter A, Adkins V, Selkie E. Youth perspectives on the recommended age of mobile phone adoption: survey study. JMIR Pediatr Parent 2022 Oct 31;5(4):e40704. [doi: 10.2196/40704] [Medline: 36315219]
92. Moreno MA, Kerr BR, Jenkins M, Lam E, Malik FS. Perspectives on smartphone ownership and use by early adolescents. J Adolesc Health 2019 Apr;64(4):437-442. [doi: 10.1016/j.jadohealth.2018.08.017] [Medline: 30389204]
93. Malmqvist J, Hellberg K, Möllås G, Rose R, Shevlin M. Conducting the pilot study: a neglected part of the research process? Methodological findings supporting the importance of piloting in qualitative research studies. Int J Qual Methods 2019 Jan 1;18:1609406919878341. [doi: 10.1177/1609406919878341]
94. Hassan ZA, Schattner P, Mazza D. Doing a pilot study: why is it essential? Malays Fam Physician 2006;1(2-3):70-73. [Medline: 27570591]
95. Manalel JA, Antonucci TC. Beyond the nuclear family: children's social networks and depressive symptomology. Child Dev 2020 Jul;91(4):1302-1316. [doi: 10.1111/cdev.13307] [Medline: 31573692]
96. Dzubur E, Huh J, Maher JP, Intille SS, Dunton GF. Response patterns and intra-dyadic factors related to compliance with ecological momentary assessment among mothers and children. Transl Behav Med 2018 Mar 1;8(2):233-242. [doi: 10.1093/tbm/ibx002] [Medline: 29381785]
97. Campbell R, Goodman-Williams R, Feeney H, Fehler-Cabral G. Assessing triangulation across methodologies, methods, and stakeholder groups: the joys, woes, and politics of interpreting convergent and divergent data. American Journal of Evaluation 2020 Mar;41(1):125-144. [doi: 10.1177/1098214018804195]
98. Simons-Morton B, Chen R. Peer and parent influences on school engagement among early adolescents. Youth Soc 2009 Sep 1;41(1):3-25. [doi: 10.1177/0044118X09334861] [Medline: 19888349]
99. Mascia ML, Langiu G, Bonfiglio NS, Penna MP, Cataudella S. Challenges of preadolescence in the school context: a systematic review of protective/risk factors and intervention programmes. Education Sciences 2023;13(2):130. [doi: 10.3390/educsci13020130]
100. Silva LM, Cibrian FL, Bonang C, et al. Co-designing situated displays for family co-regulation with ADHD children. Presented at: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems; May 11, 2024; Honolulu HI, USA p. 1-19. [doi: 10.1145/3613904.3642745]
101. Scheer ER, Werner NE, Coller RJ, et al. Designing for caregiving networks: a case study of primary caregivers of children with medical complexity. J Am Med Inform Assoc 2024 Apr 19;31(5):1151-1162. [doi: 10.1093/jamia/ocae026] [Medline: 38427845]
102. Warren JL, Antle AN, Kitson A, Davoodi A. A codesign study exploring needs, strategies, and opportunities for digital health platforms to address pandemic-related impacts on children and families. Int J Child Comput Interact 2023 Sep;37:100596. [doi: 10.1016/j.ijcci.2023.100596]
103. Coiera E, Yin K, Sharan RV, et al. Family informatics. J Am Med Inform Assoc 2022 Jun 14;29(7):1310-1315. [doi: 10.1093/jamia/ocac049] [Medline: 35380677]
104. Campbell M, Katikireddi SV, Sowden A, McKenzie JE, Thomson H. Improving Conduct and Reporting of Narrative Synthesis of Quantitative Data (ICONS-Quant): protocol for a mixed methods study to develop a reporting guideline. BMJ Open 2018 Feb;8(2):e020064. [doi: 10.1136/bmjopen-2017-020064]

XSL•FO
RenderX

Abbreviations

ADHD: attention deficit/hyperactivity disorder
ASD: autism spectrum disorder
CREMAS: Checklist for Reporting EMA Studies
EMA: ecological momentary assessment
HCI: Human-Computer Interaction
KS1: key stage 1
KS2: key stage 2
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
ROBINS-I v2: Risk Of Bias In Nonrandomized Studies of Interventions, version 2
RQ: research question

Edited by S Brini; submitted 19.Jun.2025; peer-reviewed by J Hepp, Y Feng; revised version received 15.Jan.2026; accepted 16.Jan.2026; published 14.Apr.2026.

Please cite as:
Charitos S, Thompson L, Brigden A, Bird J
A Health-Related Digital Ecological Momentary Assessment in Children (Aged 5– 11 Years): Systematic Review
J Med Internet Res 2026;28:e79291
URL: https://www.jmir.org/2026/1/e79291
doi:10.2196/79291

© Sydney Charitos, Lauren Thompson, Amberly Brigden, Jon Bird. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Role of Digital Biomarkers in Physiological Signal-Based Depression Assessment: Systematic Review and Meta-Analysis

Hyeongsuk Lee[1], RN, PhD; Seung-Gul Kang[2], MD, PhD; SeonHeui Lee[1], RN, PhD

[1]Research Institute of AI and Nursing Science, College of Nursing, Gachon University, 191 Hambangmoe-ro, Yeonsu-gu, Incheon, Republic of Korea
[2]Department of Psychiatry, Gil Medical Center, Gachon University College of Medicine, Incheon, Republic of Korea

Corresponding Author:
SeonHeui Lee, RN, PhD
Research Institute of AI and Nursing Science, College of Nursing, Gachon University, 191 Hambangmoe-ro, Yeonsu-gu, Incheon, Republic of Korea

Abstract

Background: Digital biomarkers are increasingly being used to support depression assessment by providing objective, continuous, and real-time physiological and behavioral data. However, most existing studies have focused on individual biomarkers, such as sleep or cardiac parameters, while integrative evaluations that capture the multidimensional nature of depression remain limited.

Objective: This systematic review evaluated digital biomarkers for depression and synthesized evidence on differences between individuals with depression and controls.

Methods: Eligible studies included observational or interventional studies examining digital biomarkers for depression with validated outcome measures. We searched major international and Korean databases, including MEDLINE, PsycINFO, CINAHL, IEEE Xplore, Web of Science, Cochrane Library, KISS, RISS, KMbase, and KoreaMed, from inception to December 28, 2025. Risk of bias was assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 tool and the Scottish Intercollegiate Guidelines Network checklist. Meta-analyses were conducted using random-effects models with the Hartung-Knapp-Sidik-Jonkman method, and other outcomes were narratively summarized.

Results: The search yielded 39,617 records, of which 132 studies involving 57,852 participants met the inclusion criteria. These studies encompassed various digital biomarkers, including sleep, physical activity, cardiac measures, smartphone-derived data, speech, GPS data, and circadian rhythms. A meta-analysis of 22 studies (6947 participants) revealed that individuals with depression had significantly longer sleep onset latency (5 studies; n=292; +4.75 min, 95% CI 2.46-7.04; *P*=.005; 95% prediction interval [PI] 0.01-10.27) and time in bed (3 studies; n=236; +31.81 min, 95% CI 18.22-45.39; *P*=.01; 95% PI 2.28-55.16). Physical activity counts were also significantly lower (5 studies; n=462; standardized mean difference −0.71, 95% CI −1.33 to −0.09; *P*=.03; 95% PI −2.18 to 0.71). Although individuals with depression showed a lower sleep efficiency, higher mean heart rate, and lower SD of normal-to-normal intervals, these differences were not statistically significant. Other digital markers yielded inconsistent results. Overall, these findings indicate that no single digital biomarker sufficiently captures depression-related changes. Instead, the results support the superiority of personalized, multimodal approaches. However, the generalizability of these findings is limited by the lack of standardized data collection protocols and high clinical heterogeneity across studies, as reflected in wide PIs.

Conclusions: Certain digital biomarkers, particularly sleep onset latency and physical activity counts, showed consistent average differences between the depression and control groups. However, wide PIs indicate substantial variability across settings, suggesting that no single marker is sufficient for reliable detection. This study advances the field by providing a comprehensive meta-analysis of multidimensional digital biomarkers, establishing a quantitative foundation for objective depression screening and monitoring. These findings support the use of personalized, multimodal digital phenotyping approaches and highlight the need for standardized, clinically interpretable frameworks for real-world depression monitoring.

Trial Registration: PROSPERO CRD42024518136; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024518136

(J Med Internet Res 2026;28:e76432) doi:10.2196/76432

KEYWORDS

digital biomarkers; depression; wearable electronic devices; sleep quality; ambulatory monitoring

XSL•FO
RenderX

Introduction

Depression is a significant mental health issue that affects over 300 million people worldwide and is one of the leading causes of disability-adjusted life years [1]. However, the current diagnostic process for depression largely relies on self-reported questionnaires and subjective clinical judgment, raising concerns regarding its accuracy and consistency [2]. These "snapshot" evaluations often fail to capture the dynamic, fluctuating nature of depressive symptoms in real-world settings, leading to delayed interventions and suboptimal treatment outcomes. To bridge this gap, digital biomarkers have emerged as a transformative objective approach, enabling the moment-by-moment quantification of individual-level human phenotypes in situ using data from personal digital devices [3,4].

Digital biomarkers, derived from smartphones, wearables, and ambient sensors, provide continuous, noninvasive, and high-frequency longitudinal data [3,5]. These markers encompass a wide range of clinical dimensions, including sleep patterns, physical activity levels, heart rate variability (HRV), vocal characteristics, and social interaction data. Recent advancements in sensor technology have significantly enhanced the precision of these metrics, offering unprecedented insights into the physiological and behavioral underpinnings of mood disorders [6,7]. These digital biomarkers possess unique properties that indicate their potential to complement or even replace traditional subjective methods of diagnosing depression. However, as the field matures, a critical challenge has surfaced regarding the inconsistency across various research findings [8,9].

While numerous studies have identified potential biomarkers, significant variability in hardware, data collection duration, and analytical pipelines has led to fragmented findings, raising concerns regarding their reproducibility and generalizability [10-12]. Although attempts have been made to conduct systematic reviews focused on digital biomarkers related to depression, these efforts have often been hindered by issues, such as the limited number of related studies or data heterogeneity, making meta-analyses infeasible [4]. Consequently, existing review reports are confined to specific biomarkers (eg, HRV or sleep data) or merely categorize and summarize findings without deeper integration [13,14]. Given that depression is a systemic disorder characterized by complex interactions between biological rhythms and behavioral shifts, isolated metrics are insufficient to capture its full multidimensionality. The lack of comprehensive systematic reviews poses a critical barrier to understanding the clinical utility and practical implementation of digital biomarkers.

There is, therefore, an urgent clinical and scientific need for a comprehensive, multimodal meta-analysis. Such an investigation is essential to distinguish "robust indicators" from "context-specific noise" and to establish the pooled effect sizes necessary for developing reliable diagnostic algorithms. By synthesizing evidence across diverse domains, including sleep, physical activity, cardiac measures, smartphone usage, speech, GPS, and circadian parameters, this systematic review aims to provide a comprehensive evaluation of the digital biomarker landscape. Specifically, this systematic review performs meta-analyses to quantify group differences between individuals with depression and controls without depression, thereby strengthening the evidence base for personalized, preemptive depression management in the digital health era.

Methods

Study Design and Registration

The protocol was prospectively registered in the PROSPERO database (CRD42024518136) and prepared following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) protocol checklist (Checklist 1).

Ethical Considerations

This systematic review was exempted by the Institutional Review Board (IRB) of Gachon University (IRB number: 1044396 - 202403 HR-052-01).

Eligibility Criteria

The inclusion criteria were as follows: (1) inclusion of participants with depression, (2) use of digital biomarkers to assess depression severity, and (3) reporting diagnostic concordance with validated assessment tools. The exclusion criteria were as follows: (1) non-English or non-Korean articles, (2) duplication, (3) inaccessible full text, and (4) reviews or qualitative studies.

Information Sources and Search Strategy

This systematic review involved a search of major academic databases, including the Cochrane Library (Wiley), MEDLINE (Ovid), PsycINFO (Ovid), CINAHL (EBSCOhost), IEEE Xplore (IEEE), Web of Science (Clarivate), and Korean academic databases, such as KISS, RISS, KMbase, and KoreaMed. The final search of all sources was conducted on December 28, 2025. Trial registries, gray literature, and author contacts were omitted as the primary search provided sufficient data. Additionally, we manually screened reference lists and removed duplicates using EndNote 20.

The search strategy combined Medical Subject Headings (MeSH) and free-text keywords related to depression and digital biomarkers. These terms were adapted for each database to maximize search sensitivity. The key terms included "depressi*," "MDD," "wearable," "application," "smartwatch," "biomarker*," "sleep*," "speech," "behavioral parameter*," "electroencephalogram," and "electrocardiogram." The search process followed the PRISMA Search Strategy (PRISMA-S) extension [15]. The full search strategy, including specific search strings, limits applied, and the number of records retrieved per database, is provided in Multimedia Appendix 1.

Selection Process

Two independent reviewers screened the titles and abstracts after removing duplicates. Potentially relevant studies and manually identified records underwent full-text assessment. Any disagreements were resolved through consensus with a third reviewer.

XSL•FO
RenderX

Data Collection Process and Data Extraction

Data were extracted by 2 independent reviewers using standardized forms. Extracted characteristics included author, year, country, study design, number of participants, age, sex, biomarker measurement device, measurement period, depression indicators, and analytical methods.

Study Risk of Bias Assessment

Quality was assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool to evaluate diagnostic accuracy and the Scottish Intercollegiate Guidelines Network (SIGN) tool for case-control studies. All eligible studies were included irrespective of their quality scores.

Effect Measures and Synthesis Methods

A meta-analysis was performed on quantitatively synthesizable digital biomarkers, and a systematic review was conducted on other biomarkers. All statistical analyses were performed in R software (version 4.5.0; R Foundation for Statistical Computing) using the “meta” and “pimeta” packages. Pooled effects, expressed as mean differences (MDs) or standardized mean differences (SMDs), were estimated using a random-effects model with the Hartung-Knapp-Sidik-Jonkman (HKSJ) method [16,17]. Restricted maximum likelihood was used for τ^2 estimation. Results have been presented with 95% CIs. To account for the distribution of true effects across different settings, 95% prediction intervals (PIs) were additionally calculated using the parametric bootstrap approach proposed by Nagashima et al [18], which is robust for small study numbers.

Parameters that could not be meta-analyzed due to reporting inconsistencies or insufficient data were narratively synthesized. These included specific sleep measures (eg, sleep fragmentation, rapid eye movement [REM] sleep, REM latency, and slow-wave sleep [SWS]); physical activity (eg, light physical activity [LPA] and energy expenditure); cardiac HRV indices (eg, the root mean square of successive differences [RMSSD], low frequency [LF], high frequency [HF], and LF/HF ratio); and parameters related to speech, GPS, and circadian rhythms.

Reporting Bias Assessment and Certainty Assessment

Reporting bias was considered qualitatively, as a formal statistical assessment (eg, funnel plot) was infeasible due to fewer than 10 studies per outcome. This was supported by comprehensive database searches and manual screening of reference lists. The certainty of evidence was assessed using the GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) approach, considering risk of bias, inconsistency, imprecision, and indirectness [19].

Results

Study Selection

The initial search yielded 39,617 studies. After removing duplicates and excluding records automatically identified as ineligible through journal-type filtering, 21,915 studies remained. Following the screening of titles and abstracts, 21,649 studies were excluded. Before conducting a full-text review, a manual search through reference lists and citation tracking of relevant systematic reviews identified 17 additional studies. Consequently, full-text reviews were conducted on 283 articles, of which 132 studies involving 57,852 participants met the inclusion criteria (Figure 1).

XSL•FO
RenderX

Figure 1. Flow diagram of the study selection process for examining the role of digital biomarkers in depression assessment.

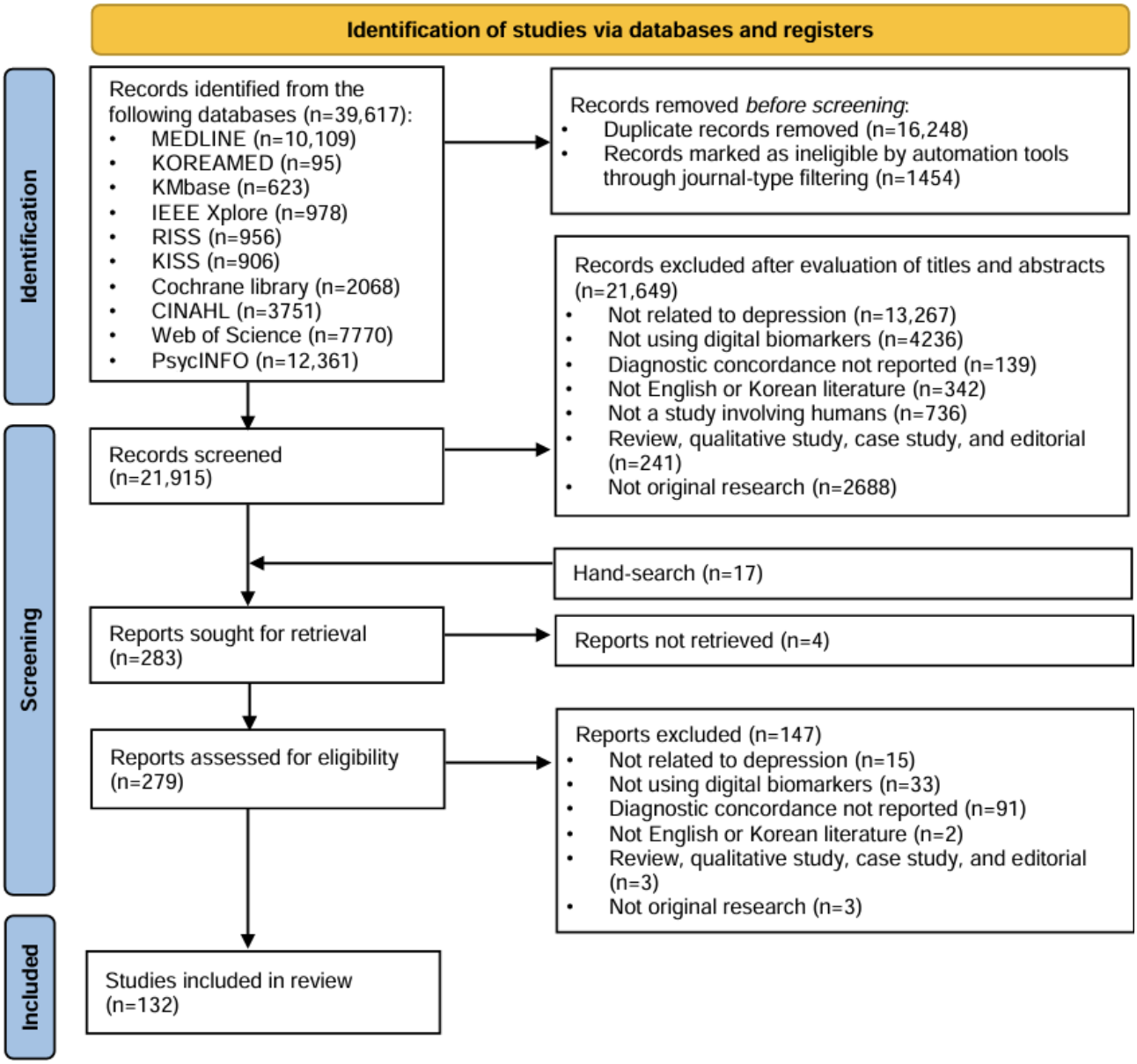

Study Characteristics

The characteristics of the included studies are summarized in Tables S1 and S2 in Multimedia Appendix 2. The included studies were categorized based on the types of digital biomarkers investigated. A total of 87 studies used single parameters, including sleep parameters (n=23) [20-42], cardiac parameters (n=19) [43-61], physical activity parameters (n=16) [62-77], smartphone-based parameters (n=9) [78-86], speech parameters (n=10) [87-96], circadian rhythm parameters (n=7) [97-103], electroencephalogram parameters (n=2) [104,105], and video-based parameters (n=1) [106]. Furthermore, 45 studies used multiple parameters to assess digital biomarkers [4,107-150]. Of the 132 studies, 63 (47.7%) used wearable devices to continuously measure biomarkers in daily life, typically for more than a week. Study designs encompassed large-scale cohorts and group comparisons (eg, depression vs control). Analytical approaches varied, including regression analyses for associations, evaluations of diagnostic accuracy, and assessments of group differences (Multimedia Appendix 3).

Validated depression assessment tools were used, with the most frequent being the Patient Health Questionnaire (PHQ), followed by the Hamilton Depression Rating Scale (HDRS), Beck Depression Inventory (BDI), Geriatric Depression Scale (GDS), Korean version of the GDS (SGDS-K), and Center for Epidemiologic Studies Depression Scale (CES-D) (Multimedia Appendix 4). Biomarker categories and their detailed features are illustrated in Multimedia Appendix 5 and Table S3 in Multimedia Appendix 2.

Risk of Bias in Studies

Based on the SIGN assessment, all 73 case-control studies demonstrated acceptable internal validity with clearly defined groups. However, participation rates for each group showed substantial variability, ranging from 20.1% to 100%, and many studies (70/73, 96%) did not report comparisons between participants and nonparticipants. Additionally, potential

XSL•FO
RenderX

confounding factors were insufficiently addressed in several studies (Table S4 in Multimedia Appendix 2).

Among diagnostic accuracy studies assessed using QUADAS-2, most studies (58/59, 98%) showed a low risk of bias in the index test, reference standard, and flow and timing domains. In contrast, great concern regarding applicability was identified because study populations often did not align with intended clinical targets, limiting the generalizability of the findings to real-world settings (Table 1).

XSL•FO
RenderX

Table . Study quality assessment using QUADAS-2[a].

First author	Year	Risk of bias[b]				Applicability concerns[b]		
		Patient selection	Index test	Reference standard	Flow and timing	Patient selection	Index test	Reference standard
Peng [21]	2023	1	1	1	1	1	1	1
Hasanzadeh [28]	2020	1	1	1	1	1	1	1
Ding [30]	2019	1	1	1	1	1	1	1
Coutts [50]	2020	2	1	1	1	2	1	1
Roh [55]	2014	1	1	1	1	2	1	1
Zhang [57]	2012	2	1	1	1	2	1	1
Espino-Salinas [63]	2022	1	1	1	1	2	1	1
Hsueh [70]	2021	1	1	1	1	1	1	1
Jakobsen [71]	2020	1	1	1	1	1	1	1
Zhao [72]	2019	1	1	1	1	2	1	1
Ku [74]	2018	1	1	1	1	1	1	1
Fadul [78]	2023	1	1	1	1	2	1	1
Auerbach [79]	2022	1	1	1	1	1	1	1
Otte Andersen [80]	2022	1	1	1	1	2	1	1
Opoku Asare [81]	2021	2	2	1	1	1	1	1
Chikersal [82]	2021	2	1	1	1	2	1	1
Zhang [83]	2021	1	1	1	1	1	1	1
Pedrelli [84]	2020	2	1	1	1	2	1	1
Mastoras [85]	2019	1	1	1	1	2	1	1
Saeb [86]	2015	2	1	1	1	2	1	1
Wiseman [87]	2025	1	1	1	1	1	1	1
Kim [90]	2023	1	1	1	1	1	1	1
Zhao [92]	2022	1	1	1	1	1	1	1
Ye [93]	2021	1	1	1	1	1	1	1
Klangpornkun [94]	2021	1	2	1	1	1	2	1
Demiroglu [95]	2020	1	1	1	1	1	1	1
Yamamoto [96]	2020	1	1	1	1	1	1	1
Choi [99]	2021	1	1	1	1	1	1	1
Anik [104]	2024	2	1	1	1	2	1	1
Tian [105]	2025	1	1	1	1	1	1	1
Islam [106]	2024	1	1	1	1	1	1	1
Makhmutova [107]	2022	1	1	1	1	1	1	1
Price [108]	2024	1	1	1	1	1	1	1
Griffiths [110]	2022	1	1	1	1	2	1	1
Rykov [113]	2021	2	1	1	1	2	1	1

XSL•FO
RenderX

First author	Year	Risk of bias[b]				Applicability concerns[b]		
		Patient selection	Index test	Reference standard	Flow and timing	Patient selection	Index test	Reference standard
Ahmed [119]	2022	2	1	1	1	2	1	1
Choi [120]	2022	2	1	1	1	2	1	1
Mahendran [121]	2019	2	1	1	1	2	1	1
Xu [122]	2019	2	1	1	1	2	1	1
Lu [123]	2018	2	1	1	1	2	1	1
Farhan [124]	2016	1	1	1	1	2	1	1
Karimi [125]	2025	1	1	1	1	1	1	1
Jacobson [126]	2020	2	1	1	1	2	1	1
Wang [127]	2018	1	1	1	1	2	1	1
Zhou [129]	2022	1	1	1	1	1	1	1
Williamson [130]	2019	1	1	1	1	2	1	1
Asare [132]	2022	1	1	1	1	2	1	1
Narziev [133]	2020	1	1	1	1	2	1	1
Kim [134]	2019	1	1	1	1	2	1	1
Di Matteo [137]	2021	1	1	1	1	1	1	1
Sverdlov [138]	2021	1	1	1	1	1	1	1
Minaeva [140]	2020	1	1	1	1	1	1	1
Dai [141]	2022	1	1	1	1	1	1	1
Tazawa [142]	2020	1	1	1	1	1	1	1
Thati [143]	2023	2	1	1	1	2	1	1
Jiang [144]	2024	1	1	1	1	1	1	1
Bai [148]	2021	2	1	1	1	2	1	1
Borelli [149]	2025	1	1	1	1	1	1	1
Chen [150]	2024	1	1	1	1	1	1	1

[a]QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies-2.

[b]1: low risk; 2: high risk.

Results of Individual Studies and Syntheses

Meta-analysis results are presented in Figure 2. A detailed overview of the parameters that could not be quantitatively synthesized and their reported associations with depression is provided in Multimedia Appendix 6.

XSL•FO
RenderX

Figure 2. Forest plots of sleep-related parameters (A: total sleep time [20,33,109,115-117,147,150]; B: sleep efficiency [20,33,39,115-117,147,150]; C: wake after sleep onset [20,33,39,115,116]; D: sleep onset latency [20,39,40,115,116]; E: time in bed [20,115,132]), physical activity–related parameters (F: step counts [69,132,147]; G: physical activity counts [20,76,111,134,150]; H: sedentary time [64,111,150]; I: moderate-to-vigorous activity [64,69,109,147]), and cardiac parameters (J: mean heart rate [43,52,54,58,61]; K: SD of normal-to-normal intervals [58,61,114,147]). Pooled estimates have been calculated using a random-effects model with the Hartung-Knapp-Sidik-Jonkman adjustment. The between-study variance (^{2}) has been estimated via the restricted maximum-likelihood method. 95% prediction intervals (PIs) have been derived using the Nagashima parametric bootstrap method.

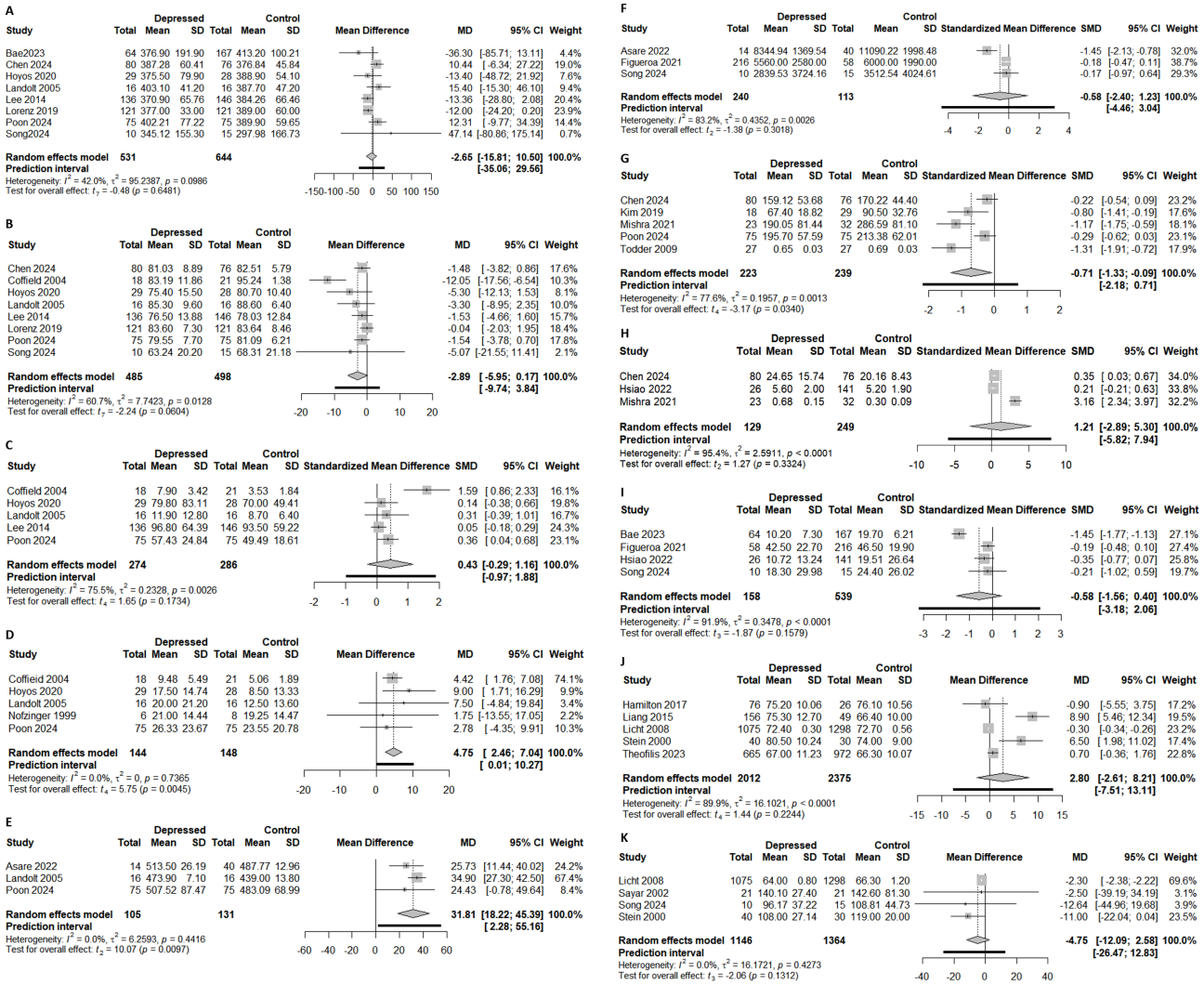

Sleep-Related Parameters and Depression

Total Sleep Time

Total sleep time (TST) showed no significant difference between the depression and control groups (−2.65 min, 95% CI −15.81 to 10.50; *P*=.65; 95% PI −35.06 to 29.56; Figure 2A). Some studies suggested an association between TST and depressive symptoms [109,112,131,135], with TST identified as the most predictive variable in 21% of patients in a predictive model [131]. Treatment studies also found a negative correlation between reduced depressive symptoms and longer TST [112]. Conversely, other studies [4,20,22,36,41,115,132,136,147,150], including those focusing on older adults and medication-treated patients [26,116], reported no significant differences.

Sleep Efficiency

Sleep efficiency (SE) was lower in the depression group but not statistically significant (−2.89%, 95% CI −5.95 to 0.17; *P*=.06; 95% PI −9.74 to 3.84; Figure 2B). Several studies reported an improvement in SE with symptom reduction [35,39] and negative correlations across diverse populations [24,26,38,148]. However, some studies found no significant differences [20,115-117,147,150] or considered SE to be a nonsignificant predictor after adjustment [36,37]. Overall, SE tended to be lower in patients with depression, although the significance varied according to the analytical method.

Wake After Sleep Onset

Wake after sleep onset (WASO) tended to be longer in the depression group, though the pooled effect was not statistically significant (SMD 0.43, 95% CI −0.29 to 1.16; *P*=.17; 95% PI −0.97 to 1.88; Figure 2C). Previous evidence has been mixed: several studies reported prolonged WASO in the depression group [20,39,132], with 1 showing posttreatment improvement [34], while others found no consistent differences [33,115,116].

Sleep Fragmentation

Sleep fragmentation and frequent awakenings were identified as predictors of depression [24,39,110].

XSL•FO
RenderX

Sleep Onset Latency

Sleep onset latency (SOL) was significantly longer in the depression group (4.75 min, 95% CI 2.46-7.04, *P*=.005; 95% PI 0.01-10.27; Figure 2D). Most studies supported this trend [35,36,116], although some reported no significant differences in the adjusted models or after symptom resolution [34,39]. Overall, a prolonged SOL was consistently associated with depressive symptoms.

Time in Bed

Time in bed (TIB) was significantly longer in the depression group (31.81 min, 95% CI 18.22-45.39; *P*=.01; 95% PI 2.28-55.16; Figure 2E). Most studies corroborated this finding, with linear mixed models consistently associating longer TIB with depressive symptoms [26,115,132,145]. Although the 95% PI remained entirely above zero, its wide range suggests that the extent of TIB prolongation may vary substantially across individuals or settings.

REM Sleep and REM Latency

Most studies on REM sleep relied on small hospital-based samples, limiting generalizability. One study found that REM sleep was significantly reduced in treated patients, particularly during the first third of the night [41]. While some research identified REM sleep as a potential predictor of depression [110], others found no group differences [115,116] or explanatory value [26,31]. Findings on REM latency were also inconsistent; 1 study showed that antidepressant-treated patients had increased REM latency [41], whereas 2 studies found no difference between the depression and control groups [33,115]. In older adults, REM latency was reported to be comparable or longer [116]. Longitudinal analyses suggested that SOL may serve as a more reliable marker than REM latency [26].

Non–Rapid Eye Movement Sleep and SWS

Evidence on non–rapid eye movement (NREM) and SWS remains limited. While 1 study found a significant association between NREM sleep and depression [26], 2 studies reported no such relationship [41,116]. Similarly, no studies to date have identified a significant association between SWS and depression [33,41,115,116].

Sleep Onset, Midpoint, and Offset

Findings regarding sleep timing (onset, midpoint, and offset) varied. Several studies, particularly those using actigraphy, reported delayed timing in individuals with depression [23,112,113,116,150]. For instance, a study of hospitalized patients showed a weak negative correlation between sleep onset time and depression (r=−0.381), indicating a delay before discharge [112]. Other research in working adults and middle-aged women found later sleep midpoints and offsets associated with higher depression scores [23,113,116]. However, these results remain inconsistent across measurement methods and study populations [26,116,127,136,148].

Physical Activity Parameters and Depression

Step Counts

The meta-analysis showed no significant difference in daily step counts between the depression and control groups (SMD −0.58, 95% CI −2.40 to 1.23; *P*=.30; 95% PI −4.46 to 3.04; Figure 2F), although the point estimate suggested fewer steps for the depression group. Longitudinal studies indicated that higher step counts were associated with reduced depression severity, with counts tending to increase during recovery [65,142]. In older adults, more daily steps were associated with a lower risk of future depressive symptoms [63]. However, some studies reported a limited predictive value of step counts [145,147].

Physical Activity Counts

Physical activity counts were lower in the depression group (SMD −0.71, 95% CI −1.33 to −0.09; *P*=.03; 95% PI −2.18 to 0.71; Figure 2G). This finding is consistent with actigraphy-based studies reporting reduced activity levels in individuals with depression [4,38,63,71,149,150]. Strong negative correlations with symptom severity were reported, particularly in hospitalized patients [112]. Longitudinal analyses showed that activity counts decreased during depressive episodes and increased with symptom improvement, supporting their relevance as state-dependent markers [4]. Notably, daytime activity levels were sensitive to clinical improvement, whereas nighttime activity measures did not reflect symptom changes [76]. Despite these promising results, the wide PI encompassing zero underscores the need for standardized approaches to account for potential inconsistencies in future observations.

Sedentary Time

Sedentary time showed no significant difference between the depression and control groups (SMD 1.21, 95% CI −2.89 to 5.30; *P*=0.33; 95% PI −5.82 to 7.94; Figure 2H). Findings regarding sedentary time were inconsistent. While some studies linked increased sedentary time to depression [66], others found no association after adjusting for moderate-to-vigorous physical activity (MVPA) or overall activity levels [74,135,149]. Longitudinal and intervention-based analyses further suggested that sedentary time alone is a relatively weak and nonspecific marker of depressive symptoms compared with mobility-related indicators such as activity counts or homestay [136]. This limitation may be particularly evident in older populations, where prolonged sedentary time may occur regardless of the depressive status [64,111].

MVPA Findings

The meta-analysis showed no significant difference in MVPA between the depression and control groups (SMD −0.58, 95% CI −1.56 to 0.40; *P*=.16; 95% PI −3.18 to 2.06; Figure 2I). However, individual studies suggested a possible link, with higher MVPA associated with reduced depression severity [68], particularly in older adults [109]. Notably, 1 study observed significant reductions in MVPA on weekends among individuals with depression [69].

LPA Findings

The results of LPA were mixed. Some studies reported that a greater time spent in LPA was associated with fewer depressive symptoms [68,74], whereas others found no significant association [110,111,123,147,149]. Overall, LPA may be more effective when interpreted alongside MVPA or step data.

XSL•FO
RenderX

Energy Expenditure

Energy expenditure, measured in kilocalories or metabolic equivalents of tasks (METs), has shown mixed findings. One study reported significantly lower energy expenditure in patients with depression than in controls [132], and a longitudinal study found increases in energy expenditure with symptom improvement [142]. Conversely, another study observed no group differences [117]. Further studies are required to assess its value as a biomarker.

Cardiac Parameters and Depression

Time Domain

Mean Heart Rate

The meta-analysis showed a higher mean heart rate (mHR) during depression, although it was not statistically significant (2.80 beats per min, 95% CI −2.61 to 8.21; *P*=.22; 95% PI −7.51 to 13.11; Figure 2J). Notably, elevated nighttime mHR was linked to depression severity, whereas daytime mHR showed no consistent association [44,113].

SD of Normal-to-Normal Intervals

The SD of normal-to-normal intervals (SDNN) was lower in the depression group, but the meta-analysis result was not significant (−4.75 ms, 95% CI −12.09 to 2.58; *P*=.13; 95% PI −26.47 to 12.83; Figure 2K). While some studies reported significance [54,61,125], others did not [49,56]. Recent wearable electrocardiogram (ECG) research suggests that the SDNN tends to be reduced in individuals at risk of depression, reflecting autonomic dysregulation; however, its standalone discriminative power may be limited compared with other short-term HRV indices [125].

RMSSD Findings

The RMSSD tended to be lower in the depression group but was not consistently significant across studies [49,56,61,114]. Although some research on first-episode patients reported a significant association [54], others found no correlation between depression scores and the RMSSD [45,47,57,145]. Recent digital phenotyping studies using wearable ECG or multimodal sensing reported reduced RMSSD-related features in high-risk individuals, though their utility was often context-dependent or enhanced when combined with other features [125,147,149].

Proportion of Normal-to-Normal Intervals

Most studies showed no significant group differences in the proportion of normal-to-normal intervals (pNN) [43,54,112], although this has been highlighted in diagnostic models [53,55]. Wearable ECG-based screening studies similarly reported that pNN-related features contributed modestly to multivariate models but showed limited standalone discriminative ability for depression [125].

Mean RR Interval

The findings for the mean RR interval (RRI) were mixed. Two studies reported shorter RRIs in patients with severe depression [45,54], while 2 other studies found no relationship [56,59]. The mean RRI tended to be shorter in patients with depression, suggesting an autonomic imbalance; however, the small number of available studies (2/4, 50%) limits the generalizability of the findings [45,54], highlighting the need for further large-scale research.

SD of Heart Rate

Studies reported mixed findings regarding the SD of heart rate (HR). While some studies found that a reduced SD of HR was associated with greater depression severity [44], others showed an opposite association or a nonsignificant association [113,131].

Frequency Domain

HF Power

HF power is generally lower in individuals with depression [43,45,49,53,56,59,61] and may predict future depressive symptoms [49,51,147]. However, some studies reported nonsignificant associations in multivariate models [54,57,61]. A recent wearable ECG study suggested that while HF-related parasympathetic indices contribute to depression classification models, they may lose independent significance when integrated with other HRV features [125].

LF Power

LF power has shown inconsistent results. While several studies reported lower LF in depression [45,49,51,59,61], others found higher values or no significant associations [54,56,57]. Wearable ECG-based studies indicated that LF-related features contribute variably to depression classification models, reflecting substantial heterogeneity and limited standalone interpretability [125].

LF/HF Ratio

Some studies reported a higher LF/HF ratio in the depression group [45,47,53,54], whereas others found no significant associations after adjustment [56,57,61]. Digital phenotyping evidence suggests that while the LF/HF ratio may reflect altered autonomic balance, its discriminative performance remains inconsistent across populations and analytic models [125,147]. Overall, the LF/HF ratio may show an increasing trend with depression but requires further validation as a reliable biomarker.

Very Low Frequency and Ultra-Low Frequency

Very low frequency (VLF) power was generally lower in individuals with depression [51,57,59,61] and was identified as a key feature in diagnostic models [57], although some studies reported nonsignificant associations after adjustment [49]. Ultra-low frequency (ULF) power also tended to decrease with greater depression severity [51,61]; however, 1 study found no significant difference between the depression and control groups [59]. This limited evidence warrants further investigation.

Total Power

Total power findings were inconsistent across studies. While 1 study reported significant effects on depression severity [54], other studies found no such differences [49,56,61]. However, its utility as a standalone biomarker is not well-supported.

Smartphone Parameters and Depression

Phone Usage Frequency

Findings regarding usage frequency were mixed. One study reported higher use in individuals with depression, particularly

XSL•FO
RenderX

students [86], while other studies found no such association [80,145]. Multimodal studies suggested that raw usage volume may not differ between groups; instead, temporal patterns, screen time, and communication regularity appear more relevant [4,81,138,149]. Furthermore, physical activity and sleep features often outperformed usage frequency in predictive models [84,133,137].

Phone Usage Duration

Results regarding phone usage duration were conflicting. While some studies observed longer usage in individuals with depression [86,122,127], others reported a shorter duration [132] or no differences [145]. Its predictive value appeared more pronounced in younger populations [122] but generally remained inferior to physiological indicators, such as HRV or physical activity [148,149]. These demographic variations and the relative inconsistency across findings limit its potential as a reliable biomarker.

Phone Calls

The frequency of phone calls was lower in some individuals with depression [138], though this effect was modest and highly dependent on age and social context [4,149]. Younger users preferred text-based communication over traditional calls [82,84,133]. Consequently, call frequency was insufficient as a biomarker but may complement indicators, such as activity or sleep [4,131,145,149].

Light Exposure

Low light exposure was shown to be linked to relapse [135] or depression in older adults [134], yet other studies reported no effects [133,137,142]. While potentially relevant for specific subgroups, this parameter requires further validation.

Number of Bluetooth-Connected Devices

Used as a proxy for social contact, this parameter was shown to correlate negatively with depression [146] and was integrated into predictive models [82,122]. While useful for assessing social activity, broader application requires more robust evidence.

Typing Patterns

Typing behavior analyzed via machine learning showed high accuracy in predicting depression [78,86], though further validation is needed to establish its reliability as a biomarker.

Speech Parameters and Depression

Speech parameters were categorized into speech flow and voice acoustic parameters.

Speech Flow Parameters

Speech Rate

Slower speech rates are consistently associated with greater depression severity [88,96], with recent evidence confirming them as robust indicators of major depressive disorder (MDD) that correlate with objective executive dysfunction [87,144]. Models incorporating speech rate and duration outperformed acoustic-only models [95], though further validation across diverse populations is needed.

Speech Duration

Shorter speaking times were linked to higher depression severity [95,127,137,139,146]. Recent automated assessments and smartphone-derived data identified reduced active speaking time as a primary predictor of depression [137,144,150], establishing speech duration as a potential digital biomarker.

Pause Time

While some associations between depression and pause duration were nonsignificant [129], the majority of studies reinforced that increased pausing effectively captures psychomotor retardation and contributes to high-accuracy multimodal detection [87,88,96,144]. These findings underscore the potential of pause-related features as a reliable objective marker.

Voice Acoustic Parameters

Mel-Frequency Cepstral Coefficients

Mel-frequency cepstral coefficients were significantly associated with the severity of depression in most studies [91-93,129]. Recent multimodal analyses further confirmed their effectiveness when integrated with facial and cardiovascular patterns [144], although a study found text-based features more predictive than mel-frequency cepstral coefficients [95].

Fundamental Frequency

Findings on fundamental frequency were mixed. Some studies reported significant differences or enhanced effectiveness through multimodal integration [92,93,129,144], whereas others found no significant associations [89,91], necessitating further clarification of its role.

Jitter

Jitter was significantly higher in some individuals with depression [89], but other studies did not identify it as a significant variable [95,129]. Recent evidence suggests that it provides discriminative value when integrated into multimodal frameworks [144], warranting further investigation.

Shimmer

Shimmer showed significant associations with depression in some studies [89,129], while others found no such link [95]. Its predictive power was notably enhanced within multimodal frameworks, contributing to more robust detection than when used alone [144].

GPS Parameters and Depression

Total Distance

Most studies found a negative correlation between depression and total distance traveled [4,79,123,132,139,141,146]. Although not always statistically significant [86,124,145], reduced mobility was a promising digital biomarker, particularly when integrated into multimodal detection models [149].

Location Variance

Lower location variance was consistently associated with higher depression severity [4,82,86,124,132,145]. While 1 study noted inconsistencies depending on smartphone types [123], location variance remains a key feature in high-accuracy multimodal models for detecting depressive symptoms [149].

XSL•FO
RenderX

Time Spent at Home

A significant positive correlation exists between depression severity and time spent at home [4,86,123,124,127,132,149]. Notably, early changes in "homestay" were identified as a critical predictor of symptom improvement [136], demonstrating strong potential for longitudinal depression monitoring.

Location Entropy and Normalized Location Entropy

Reduced location and normalized entropy were generally associated with higher depression scores [4,86,123,124,132,149], reflecting less diverse movements. However, some studies found no significant associations [127,145].

Number of Locations Visited

Fewer visited locations correlated negatively with depression in most studies [4,123,127,132,137,149], though 1 study found no significant correlation [86].

Time Spent Moving

While associations between time spent moving and depression were inconsistent [86,123,124,145], recent longitudinal data identified early changes in moving time as a key predictor of symptom improvement [136]. These features are considered essential in multimodal frameworks for depression detection [149].

Average Moving Speed

Average moving speed was identified as a key feature in a depression prediction model [4,126], although its correlation often varied by device type [123]. While its standalone predictive power may be limited, it remains a crucial component in high-accuracy multimodal frameworks [149].

Circadian Rhythm Parameters and Depression

Interdaily Stability

Most studies found no significant group differences [97,99,100,117,118,150], whereas 2 studies reported lower interdaily stability and greater depression severity [103,113]. While interdaily stability may reflect irregular daily routines, current evidence is limited.

Intradaily Variability

Generally, intradaily variability was found to be unrelated to depression [97,99,100,117,118,150], despite 1 study linking higher intradaily variability to greater severity [103]. Its utility as a standalone biomarker remains restricted.

Midline Estimating Statistic of Rhythm

Midline estimating statistic of rhythm was lower in individuals with greater depression severity [102,148] and was identified as an important predictor in other models [99,101]. Consistently, recent evidence indicates that patients with MDD exhibit significantly lower midline estimating statistic of rhythm than controls [150], potentially reflecting reduced energy levels, though some studies found no associations [97,140].

Amplitude, Acrophase, and Relative Amplitude

Lower amplitude was observed in the depression group in some studies [101,102], while others found no significant differences [97,99,140,148,150]. Regarding acrophase, most studies found no association [97,101,102,113,148], but a recent study noted a significantly later acrophase in MDD, suggesting a delayed circadian phase [150]. Relative amplitude was generally lower in individuals with depression, suggesting flatter activity cycles, although these findings often did not reach statistical significance [97,99,100,117].

Pseudo F-Statistic

The pseudo F-statistic (F-pseudo), which measures the circadian rhythm strength, was lower in individuals with depressive symptoms [101,102], suggesting weaker or irregular rhythms. However, other studies reported no significant associations [99,113].

Most Active 10-Hour Period and Least Active 5-Hour Period

The most active 10-hour period and the least active 5-hour period showed no significant differences between the depression and control groups across multiple studies [97,99,117]. These parameters currently provide no evidence of reliability as biomarkers for depression.

Reporting Bias and Certainty of Evidence

The qualitative assessment of reporting bias suggested a low likelihood of missing relevant studies, supported by comprehensive multidatabase searches and manual reference screening. According to the GRADE approach, the certainty of evidence ranged from low to very low across the key digital biomarkers (Table 2). The certainty was low for SOL, TIB, and physical activity counts, whereas it was very low for TST, SE, WASO, and mHR. The overall certainty was mainly downgraded due to inconsistency across studies and imprecision associated with wide CIs or PIs.

XSL•FO
RenderX

Table . GRADE (Grading of Recommendations, Assessment, Development and Evaluation) summary of findings for key digital biomarkers in individuals with depression.

Certainty assessment	Studies, n	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Individuals with depression, n	Controls (no depression), n	Effect, relative (95% CI)	Effect, absolute (95% CI)	Certainty	Importance
Total sleep time	8	Nonrandomized studies	Not serious	Serious[a]	Not serious	Serious[b]	None	531	644	—[c]	MD[d] 2.65 min fewer (15.81 fewer to 10.5 more)	Very low[a,b]	Critical
Sleep efficiency	8	Nonrandomized studies	Not serious	Serious[e]	Not serious	Serious[b]	None	485	498	—	MD 2.89% lower (5.95 lower to 0.17 higher)	Very low[b,e]	Critical
Wake after sleep onset	5	Nonrandomized studies	Not serious	Serious[e]	Not serious	Serious[b]	None	274	286	—	SMD[f] 0.43 SD more (0.29 fewer to 1.16 more)	Very low[b,e]	Critical
Sleep onset latency	5	Nonrandomized studies	Not serious	Not serious	Not serious	Not serious	None	144	148	—	MD 4.75 min more (2.46 more to 7.04 more)	Low	Critical
Time in bed	3	Nonrandomized studies	Not serious	Not serious	Not serious	Not serious	None	105	131	—	MD 31.81 min more (18.22 more to 45.39 more)	Low	Critical
Physical activity counts	5	Nonrandomized studies	Not serious	Not serious	Not serious	Not serious	None	223	239	—	SMD 0.71 SD lower (1.33 lower to 0.09 lower)	Low	Critical
Mean heart rate	5	Nonrandomized studies	Not serious	Serious[e]	Not serious	Serious[b]	None	2012	2375	—	MD 2.8 bpm higher (2.61 lower to 8.21 higher)	Very low[b,e]	Critical

[a]Downgraded for serious inconsistency due to conflicting findings across studies and a wide prediction interval.

[b]Downgraded for serious imprecision due to a wide CI crossing the null effect.

[c]Not applicable.

[d]MD: mean difference.

XSL•FO
RenderX

[e]Downgraded for serious inconsistency due to moderate heterogeneity and variable results across studies.

[f]SMD: standardized mean difference.

Discussion

Principal Findings

This systematic review synthesized digital biomarkers for depression across diverse domains—sleep, physical activity, cardiac parameters, smartphone usage, speech, GPS data, and circadian rhythms—to identify more consistent indicators across multiple digital signals. Our meta-analysis identified prolonged SOL, increased TIB, and reduced activity counts as the most consistent behavioral features associated with depression. These findings support the hypothesis that digital phenotyping can capture objective manifestations of depression, particularly sleep initiation difficulties and reduced energy expenditure, which are often difficult to quantify through traditional self-reports.

A key contribution of this systematic review is the application of quantitative meta-analyses in a field where such synthesis was previously considered methodologically challenging [10,11]. Unlike prior meta-analyses that focused on single domains or summarized findings narratively because of methodological heterogeneity [13,14], this review systematically evaluates digital biomarkers across diverse domains. Consistent with earlier reports, the results indicate that depression is not characterized by a single physiological signature but rather by a constellation of behavioral and biological changes [6,149].

Furthermore, unlike recent scoping reviews on general mental illness [151], our analysis is specifically tailored to MDD. Importantly, the interpretation of these findings must consider the substantial heterogeneity observed across studies. The CIs reflect the average effect across the included studies, whereas the PIs indicate the range of effects that may occur in future settings. For several outcomes, wide PIs crossing the null value suggest that effect sizes may vary considerably depending on the study population, device type, measurement protocol, and analytical approach. These findings indicate that group-level average effects may not generalize consistently across contexts, supporting the use of personalized, longitudinal multibiomarker models for effective monitoring and intervention.

Key Digital Biomarkers

Several biomarkers emerged as important indicators of depression across multiple domains. In the sleep domain, individuals with depression showed significantly longer SOL and increased TIB compared to controls without depression. These findings suggest that the sleep-wake cycle in MDD is characterized more by structural fragmentation than by simple reductions in sleep duration. Specifically, increased TIB likely reflects hallmark symptoms such as psychomotor retardation and lethargy. While polysomnography remains the gold standard [152], these results demonstrate that wearable devices can capture such patterns in ecologically valid, naturalistic settings.

The absence of significant differences in TST, SE, and WASO, which are parameters frequently associated with depression in previous reviews [4], highlights the clinical heterogeneity of MDD [153,154]. Depression may manifest as either insomnia or hypersomnia, depending on the subtype, potentially neutralizing average effects in pooled analyses [155]. The wide PIs further indicate substantial heterogeneity, likely driven by differences in study populations, medication use, and device-specific scoring. These findings underscore that absolute sleep quantity alone is an insufficient marker, and the relationship between sleep and depression is shaped by complex biopsychosocial factors [153,154]. Consequently, temporal and qualitative features, such as SOL and TIB, may serve as more clinically informative indicators than total sleep volume [9,156].

With respect to sleep architecture, evidence regarding REM-related parameters remains inconclusive. While some studies suggest that shortened REM latency may precede depression [157], results for REM duration and frequency are inconsistent [26,31,41,115,116]. Age-related changes [33] and reliance on single-night laboratory measurements further limit generalizability. These findings highlight the need for longitudinal, real-world assessments to clarify the relationship between sleep architecture and depression.

Physical activity counts emerged as a sensitive state-dependent marker, showing significant reductions in depression groups. Compared with simple step counts, activity counts better capture movement intensity and frequency [65,142,158], reflecting the energy deficits associated with MDD. Although daily step counts were not significantly different in pooled analyses, longitudinal evidence suggests that they may be sensitive to individual recovery trajectories [65,142]. Several studies also reported more pronounced reductions during daytime hours and on weekends, emphasizing the importance of temporal patterns rather than simple daily averages [158,159]. In contrast, MVPA, sedentary time, and LPA showed inconsistent results. This lack of robust significance likely reflects high heterogeneity and the limited specificity of these markers when used in isolation.

GPS-derived parameters, including total distance traveled, location entropy, time spent at home, and number of locations visited, capture behavioral and environmental changes associated with depression. Mobility patterns, such as reduced diversity in visited locations (entropy) and increased homestay durations, have emerged as promising indicators of depressive symptoms [160]. Recent longitudinal evidence suggests that early changes in homestay and mobility patterns may predict symptom improvement [136], highlighting the value of GPS-derived features in multimodal monitoring frameworks [149].

Cardiac parameters showed less consistent results. Individuals with depression tended to exhibit higher nocturnal mHR and lower HRV, although these differences were not statistically significant in pooled analyses. These trends may reflect physiological hyperarousal and autonomic dysregulation [161,162], but their effects appear to vary across populations and study conditions. Accordingly, these markers may be more informative when interpreted within multimodal models that account for age, medication use, and comorbid conditions. Despite this overall inconsistency, short-term HRV indices and

XSL•FO
RenderX

frequency-domain features (eg, HF and VLF) remain meaningful components in diagnostic and multimodal frameworks, particularly as sensitive indicators of autonomic imbalance in first-episode depression or high-risk individuals [54,125,149].

Speech parameters, including slower speech rates and longer pauses, consistently distinguished individuals with MDD from controls in several studies [87,144]. These acoustic features likely reflect psychomotor retardation.

Implications for Personalized Digital Phenotyping

Digital biomarkers offer several advantages over traditional self-reported measures by providing objective, continuous data that reduce recall bias and capture early physiological and behavioral changes [163,164]. Passive monitoring enables large-scale, low-burden data collection, supporting early detection and personalized, data-driven interventions [5,8,165].

However, the substantial variability observed across individuals highlights the need to interpret these signals within a personalized framework, as traditional group-based models may be inadequate in many contexts [166]. The predictive power of biomarkers varies with age. TIB, daily step counts, and MVPA are more predictive in older adults [11,109,167], whereas smartphone usage patterns show stronger associations in younger individuals [11,122]. High-resolution temporal data, such as nighttime HRV or weekend reductions in MVPA, may provide additional insights into symptom dynamics, underscoring the importance of temporal patterns in digital measurements [69]. Monitoring sleep-related biomarkers in relation to changes in individual symptoms may further improve the prediction and tracking of depressive episodes [6]. Together, these findings suggest that digital monitoring strategies should focus on detecting deviations from an individual's baseline rather than relying on universal thresholds.

Methodological Limitations of the Included Studies and Evidence Certainty

Despite identifying several robust biomarkers, parameters, such as LPA, sedentary time, and specific cardiac measures, showed inconsistent results. This inconsistency, together with the wide PIs observed across studies, likely reflects not only diverse device technologies and measurement protocols, but also the symptomatic variability across depression subtypes [155,168]. While wearable devices offer objective physiological data, they do not directly capture emotional states [169]. Furthermore, the predominance of cross-sectional designs limits our ability to determine the temporal or causal relationships between these digital signals and depressive symptoms [170].

Risk-of-bias assessments also indicated several methodological concerns. Although participation rates were somewhat low in certain studies [24,34,59,62,128], they were generally acceptable across the majority of the included research [4,25,27,31,32,35,38,42,45,49,52-54,56,60,67,68,76,88,99,101,111,112,114,115,117,131,145]. However, comparisons between participants and nonparticipants were often not reported, raising the possibility of selection bias. For example, individuals with more severe depressive symptoms, lower motivation, or limited familiarity with wearable technologies may have been less likely to participate or remain in the studies. In addition, adjustment for confounding variables was limited in several studies. For diagnostic accuracy studies, most domains showed low risk of bias, but concerns regarding applicability were common because study populations did not always reflect real-world clinical settings.

Consistent with these methodological limitations, the GRADE assessment indicated low to very low certainty of evidence for the key digital biomarkers. This was primarily due to inconsistency across studies and imprecision associated with wide CIs and PIs. These findings suggest that the pooled estimates should be interpreted cautiously and that further well-designed, standardized studies are needed to strengthen the evidence base.

Limitations of This Review

This systematic review has several limitations. First, substantial heterogeneity across studies limited the precision and generalizability of the pooled estimates, reflecting differences in devices, study populations, monitoring periods, and analytic methods. Second, the overall certainty of evidence was low to very low according to the GRADE framework, which reduces confidence in the pooled estimates. Third, a formal statistical assessment of reporting bias was not feasible because fewer than 10 studies were available for each meta-analysis. Finally, many studies included nonclinical or convenience samples, which may limit the generalizability of the findings to real-world clinical populations.

Future Directions for Digital Biomarker Research

Future research should move beyond the identification of individual markers toward the development of integrated, clinically actionable digital phenotyping systems. The consistent associations observed for indicators, such as SOL, TIB, and activity counts, suggest that certain behavioral signals may serve as foundational components of continuous mental health monitoring. The next phase of research should therefore focus on embedding these markers within longitudinal, multimodal frameworks that support personalized clinical decision-making and improve the precision of depression monitoring [149,150,155], with the potential to inform more proactive intervention strategies in real-world settings.

Achieving this transition will require greater methodological standardization across devices, measurement protocols, and analytic pipelines. Reducing reliance on proprietary, nontransparent algorithms and promoting device-agnostic, reproducible approaches will be essential for ensuring clinical validity and interoperability across health care systems. Integration of digital biomarker data into electronic health records may further enable real-time, context-aware decision support while minimizing additional cognitive burden on clinicians.

More broadly, digital biomarkers may support earlier detection and more adaptive treatment strategies by providing high-frequency, objective data on symptom trajectories [136,164]. At the population level, scalable signals such as physical activity and mobility patterns may also facilitate screening and risk stratification in settings with limited access to mental health professionals. Future work should therefore prioritize longitudinal, multimodal designs; representative

XSL•FO
RenderX

clinical populations; standardized measurement protocols; and transparent reporting of analytic methods.

Conclusion

This systematic review provides a comprehensive synthesis of multimodal digital biomarkers for depression. Unlike previous reviews that focused on single signals or technical feasibility, this systematic review advances the field by establishing a standardized framework for objective clinical decision-making through a rigorous meta-analysis. The findings indicate that while certain markers, particularly SOL and physical activity counts, show consistent average differences, their effects vary substantially across settings, as reflected by wide PIs. These results suggest that depression cannot be reliably characterized by a single digital biomarker. Instead, a multimodal, personalized approach that integrates physiological, behavioral, and contextual signals is likely to be more effective for real-world applications. More broadly, this systematic review demonstrates that quantitative synthesis in digital phenotyping is feasible despite substantial heterogeneity and that meaningful signals can be identified when methodological rigor and transparent reporting are applied. Establishing standardized, clinically interpretable digital biomarker frameworks will be essential for advancing objective, continuous, and personalized assessments of depression in routine care.

Funding

This study was supported by a grant (24202MFDS201) from the Ministry of Food and Drug Safety in 2024. The funder was not involved in the study design, data collection, analysis, interpretation, or writing of the manuscript.

Disclaimer

Generative artificial intelligence tools were not used in the preparation, writing, or editing of this manuscript.

Data Availability

No original individual participant data were collected for this systematic review. All the data used in this systematic review and meta-analysis were extracted from previously published studies. Therefore, individual-level datasets cannot be deposited in a public repository. The list of the included studies and extracted variables is available from the corresponding author upon reasonable request.

Authors' Contributions

SHL contributed to the study conception and design. HL and SHL assisted with the statistical analysis and provided administrative support. HL, SHL, and SGK wrote the first draft of the manuscript. All authors contributed to the writing and preparation of the final manuscript. All authors had full access to all the data in the study and had the final responsibility for the decision to submit the manuscript for publication. HL and SHL accessed and verified the data reported in this manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search terms and strategy.
[DOCX File, 63 KB - jmir_v28i1e76432_app1.docx]

Multimedia Appendix 2
Detailed study characteristics and quality assessment of the included studies.
[DOCX File, 165 KB - jmir_v28i1e76432_app2.docx]

Multimedia Appendix 3
Participants and analytical methods in the studies.
[DOCX File, 23 KB - jmir_v28i1e76432_app3.docx]

Multimedia Appendix 4
Depression measurement tools included in the studies.
[DOCX File, 34 KB - jmir_v28i1e76432_app4.docx]

Multimedia Appendix 5
Digital biomarkers included in the studies.
[DOCX File, 79 KB - jmir_v28i1e76432_app5.docx]

XSL•FO
RenderX

Multimedia Appendix 6
Digital biomarkers not included in the meta-analysis.
[DOCX File, 63 KB - jmir_v28i1e76432_app6.docx]

Checklist 1
PRISMA checklist.
[DOCX File, 278 KB - jmir_v28i1e76432_app7.docx]

References

1. Li S, Xu Y, Zheng L, et al. Sex difference in global burden of major depressive disorder: findings from the Global Burden of Disease Study 2019. Front Psychiatry 2022;13:70890. [doi: 10.3389/fpsyt.2022.789305]
2. Pu J, Yu Y, Liu Y, et al. ProMENDA: an updated resource for proteomic and metabolomic characterization in depression. Transl Psychiatry 2024 May 30;14(1):229. [doi: 10.1038/s41398-024-02948-2] [Medline: 38816410]
3. Macias Alonso AK, Hirt J, Woelfle T, Janiaud P, Hemkens LG. Definitions of digital biomarkers: a systematic mapping of the biomedical literature. BMJ Health Care Inform 2024 Apr 8;31(1):345. [doi: 10.1136/bmjhci-2023-100914] [Medline: 38589213]
4. Aledavood T, Luong N, Baryshnikov I, et al. Multimodal digital phenotyping study in patients with major depressive episodes and healthy controls (mobile monitoring of mood): observational longitudinal study. JMIR Ment Health 2025 Feb 21;12:e63622. [doi: 10.2196/63622] [Medline: 39984168]
5. Ikäheimonen A, Luong N, Baryshnikov I, et al. Predicting and monitoring symptoms in patients diagnosed with depression using smartphone data: observational study. J Med Internet Res 2024 Dec 3;26:e56874. [doi: 10.2196/56874] [Medline: 39626241]
6. Jacobson NC, Weingarden H, Wilhelm S. Digital biomarkers of mood disorders and symptom change. NPJ Digit Med 2019;2(1):3. [doi: 10.1038/s41746-019-0078-0] [Medline: 31304353]
7. Jung HW, Kim DY, Lee I, et al. Key features of digital phenotyping for monitoring mental disorders: systematic review. J Med Internet Res 2025 Nov 5;27:e77331. [doi: 10.2196/77331] [Medline: 41191793]
8. Zhuparris A, Maleki G, van Londen L, et al. A smartphone- and wearable-based biomarker for the estimation of unipolar depression severity. Sci Rep 2023 Nov 1;13(1):18844. [doi: 10.1038/s41598-023-46075-2] [Medline: 37914808]
9. Zierer C, Behrendt C, Lepach-Engelhardt AC. Digital biomarkers in depression: a systematic review and call for standardization and harmonization of feature engineering. J Affect Disord 2024 Jul 1;356:438-449. [doi: 10.1016/j.jad.2024.03.163] [Medline: 38583596]
10. De Angel V, Lewis S, White K, et al. Digital health tools for the passive monitoring of depression: a systematic review of methods. NPJ Digit Med 2022 Jan 11;5(1):3. [doi: 10.1038/s41746-021-00548-8] [Medline: 35017634]
11. Zarate D, Stavropoulos V, Ball M, de Sena Collier G, Jacobson NC. Exploring the digital footprint of depression: a PRISMA systematic literature review of the empirical evidence. BMC Psychiatry 2022 Jun 22;22(1):421. [doi: 10.1186/s12888-022-04013-y] [Medline: 35733121]
12. Leaning IE, Ikani N, Savage HS, et al. From smartphone data to clinically relevant predictions: a systematic review of digital phenotyping methods in depression. Neurosci Biobehav Rev 2024 Mar;158:105541. [doi: 10.1016/j.neubiorev.2024.105541] [Medline: 38215802]
13. Ho FYY, Poon CY, Wong VWH, et al. Actigraphic monitoring of sleep and circadian rest-activity rhythm in individuals with major depressive disorder or depressive symptoms: a meta-analysis. J Affect Disord 2024 Sep 15;361:224-244. [doi: 10.1016/j.jad.2024.05.155] [Medline: 38851435]
14. Bizzozero-Peroni B, Díaz-Goñi V, Jiménez-López E, et al. Daily step count and depression in adults: a systematic review and meta-analysis. JAMA Netw Open 2024 Dec 2;7(12):e2451208. [doi: 10.1001/jamanetworkopen.2024.51208] [Medline: 39680407]
15. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
16. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14(1):25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
17. Higgins JPT, Thomas J, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.5 (updated August 2024): Cochrane; 2024. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook [accessed 2026-03-11]
18. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019 Jun;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]

XSL•FO
RenderX

19. Schünemann H, Brożek J, Guyatt G, Oxman A. GRADE Handbook - Handbook for Grading the Quality of Evidence and the Strength of Recommendations Using the GRADE Approach: GRADE Working Group; 2013. URL: https://gdt.gradepro.org/app/handbook/handbook.html [accessed 2026-03-11]
20. Poon CY, Cheng YC, Wong VWH, et al. Directional associations among real-time activity, sleep, mood, and daytime symptoms in major depressive disorder using actigraphy and ecological momentary assessment. Behav Res Ther 2024 Feb;173:104464. [doi: 10.1016/j.brat.2023.104464] [Medline: 38159415]
21. Peng D, Liu W, Luo Y, Mao Z, Zheng WL, Lu BL. Deep depression detection with resting-state and cognitive-task EEG. Annu Int Conf IEEE Eng Med Biol Soc 2023 Jul;2023:1-4. [doi: 10.1109/EMBC40787.2023.10340667] [Medline: 38083722]
22. Comas M, Solis Flores A, Lovato N, et al. The relationship between anxiety, subjective and objective sleep, chronotype and circadian rhythms with depressive symptoms in insomnia disorder. Brain Sci 2023 Apr 4;13(4):613. [doi: 10.3390/brainsci13040613] [Medline: 37190578]
23. Bowman MA, Kline CE, Buysse DJ, et al. Longitudinal association between depressive symptoms and multidimensional sleep health: the SWAN Sleep Study. Ann Behav Med 2021 Jun 28;55(7):641-652. [doi: 10.1093/abm/kaaa107] [Medline: 33410460]
24. Smith P, Kandakatla A, Frankel CW, et al. Sleep quality, depressive symptoms, and transplant outcomes: follow-up analyses from the ADAPT prospective pilot study. Gen Hosp Psychiatry 2021;72:53-58. [doi: 10.1016/j.genhosppsych.2021.06.011] [Medline: 34298477]
25. Mohammadi Y, Moradi MH. Prediction of depression severity scores based on functional connectivity and complexity of the EEG signal. Clin EEG Neurosci 2021 Jan;52(1):52-60. [doi: 10.1177/1550059420965431] [Medline: 33040603]
26. Zhang Y, Folarin AA, Sun S, et al. Predicting depressive symptom severity through individuals' nearby Bluetooth device count data collected by mobile phones: preliminary longitudinal study. JMIR Mhealth Uhealth 2021 Jul 30;9(7):e29840. [doi: 10.2196/29840] [Medline: 34328441]
27. Chen F, Zhao L, Li B, Yang L. Depression evaluation based on prefrontal EEG signals in resting state using fuzzy measure entropy. Physiol Meas 2020 Oct 6;41(9):095007. [doi: 10.1088/1361-6579/abb144] [Medline: 33021227]
28. Hasanzadeh F, Mohebbi M, Rostami R. Graph theory analysis of directed functional brain networks in major depressive disorder based on EEG signal. J Neural Eng 2020 Mar 27;17(2):026010. [doi: 10.1088/1741-2552/ab7613] [Medline: 32053813]
29. Rajpurkar P, Yang J, Dass N, et al. Evaluation of a machine learning model based on pretreatment symptoms and electroencephalographic features to predict outcomes of antidepressant treatment in adults with depression: a prespecified secondary analysis of a randomized clinical trial. JAMA Netw Open 2020 Jun 1;3(6):e206653. [doi: 10.1001/jamanetworkopen.2020.6653] [Medline: 32568399]
30. Ding X, Yue X, Zheng R, Bi C, Li D, Yao G. Classifying major depression patients and healthy controls using EEG, eye tracking and galvanic skin response data. J Affect Disord 2019 May 15;251:156-161. [doi: 10.1016/j.jad.2019.03.058] [Medline: 30925266]
31. Gould CE, Karna R, Jordan J, et al. Subjective but not objective sleep is associated with subsyndromal anxiety and depression in community-dwelling older adults. Am J Geriatr Psychiatry 2018 Jul;26(7):806-811. [doi: 10.1016/j.jagp.2018.03.010] [Medline: 29709510]
32. Haghighi M, Ludyga S, Rahimi B, et al. In patients suffering from major depressive disorders, quantitative EEG showed favorable changes in left and right prefrontal cortex. Psychiatry Res 2017 May;251:137-141. [doi: 10.1016/j.psychres.2017.02.012] [Medline: 28199912]
33. Lee HJ, Lee JS, Kim T, Yoon IY. Relationship between sleep disturbances and cognitive impairments in older adults with depression. Sleep Medicine and Psychophysiology 2014 Jun 30;21(1):5-13. [doi: 10.14401/KASMED.2014.21.1.5]
34. Maglione JE, Ancoli-Israel S, Peters KW, et al. Subjective and objective sleep disturbance and longitudinal risk of depression in a cohort of older women. Sleep 2014 Jul 1;37(7):1179-1187. [doi: 10.5665/sleep.3834] [Medline: 25061246]
35. Pillai V, Steenburg LA, Ciesla JA, Roth T, Drake CL. A seven day actigraphy-based study of rumination and sleep disturbance among young adults with depressive symptoms. J Psychosom Res 2014 Jul;77(1):70-75. [doi: 10.1016/j.jpsychores.2014.05.004] [Medline: 24913345]
36. Maglione JE, Ancoli-Israel S, Peters KW, et al. Depressive symptoms and subjective and objective sleep in community-dwelling older women. J Am Geriatr Soc 2012 Apr;60(4):635-643. [doi: 10.1111/j.1532-5415.2012.03908.x] [Medline: 22428562]
37. Bei B, Milgrom J, Ericksen J, Trinder J. Subjective perception of sleep, but not its objective quality, is associated with immediate postpartum mood disturbances in healthy women. Sleep 2010 Apr;33(4):531-538. [doi: 10.1093/sleep/33.4.531] [Medline: 20394323]
38. Goyal D, Gay C, Lee K. Fragmented maternal sleep is more strongly correlated with depressive symptoms than infant temperament at three months postpartum. Arch Womens Ment Health 2009 Aug;12(4):229-237. [doi: 10.1007/s00737-009-0070-9] [Medline: 19396527]
39. Coffield TG, Tryon WW. Construct validation of actigraphic sleep measures in hospitalized depressed patients. Behav Sleep Med 2004;2(1):24-40. [doi: 10.1207/s15402010bsm0201_3] [Medline: 15600222]

XSL•FO
RenderX

40. Nofzinger EA, Nichols TE, Meltzer CC, et al. Changes in forebrain function from waking to REM sleep in depression: preliminary analyses [of 18F]FDG PET studies. Psychiatry Research: Neuroimaging 1999 Aug;91(2):59-78. [doi: 10.1016/S0925-4927(99)00025-6]
41. Luthringer R, Toussaint M, Schaltenbrand N, et al. A double-blind, placebo-controlled evaluation of the effects of orally administered venlafaxine on sleep in inpatients with major depression. Psychopharmacol Bull 1996;32(4):637-646. [Medline: 8993085]
42. Roemer RA, Shagass C, Dubin W, Jaffe R, Siegal L. Quantitative EEG in elderly depressives. Brain Topogr 1992;4(4):285-290. [doi: 10.1007/BF01135566] [Medline: 1510872]
43. Theofilis P, Oikonomou E, Lazaros G, et al. The association of depression with QT duration: a comparison between individuals younger or older than 65 years. Psychosom Med 2023 Jan 1;85(1):18-25. [doi: 10.1097/PSY.0000000000001152] [Medline: 36516316]
44. Siddi S, Bailon R, Giné-Vázquez I, et al. The usability of daytime and night-time heart rate dynamics as digital biomarkers of depression severity. Psychol Med 2023 Jun;53(8):3249-3260. [doi: 10.1017/S0033291723001034] [Medline: 37184076]
45. Subramaniam V, R N. Relationship between heart rate variability and major depressive disorder in young adults. Swiss Arch Neurol Psychiatr Psychother 2023;174:126-129. [doi: 10.4414/sanp.2023.03284]
46. Lee SY, Waring ME, Park CL, Blake EC. Do depressive symptoms predict blood pressure control in US veterans? J Gen Intern Med 2022 Jan;37(1):57-63. [doi: 10.1007/s11606-021-06709-5] [Medline: 33772439]
47. Chung AH, Gevirtz RN, Gharbo RS, Thiam MA, Ginsberg JPJ. Pilot study on reducing symptoms of anxiety with a heart rate variability biofeedback wearable and remote stress management coach. Appl Psychophysiol Biofeedback 2021 Dec;46(4):347-358. [doi: 10.1007/s10484-021-09519-x] [Medline: 34308526]
48. da Estrela C, McGrath J, Booij L, Gouin JP. Heart rate variability, sleep quality, and depression in the context of chronic stress. Ann Behav Med 2021 Mar 16;55(2):155-164. [doi: 10.1093/abm/kaaa039] [Medline: 32525208]
49. Lee MS, Seo YE, Mok YE, Lee SH. Heart rate variability after treatment for depression in North Korean defectors. Appl Psychophysiol Biofeedback 2021 Mar;46(1):11-18. [doi: 10.1007/s10484-020-09491-y] [Medline: 33074464]
50. Coutts LV, Plans D, Brown AW, Collomosse J. Deep learning with wearable based heart rate variability for prediction of mental and general health. J Biomed Inform 2020 Dec;112(112):103610. [doi: 10.1016/j.jbi.2020.103610] [Medline: 33137470]
51. Huang M, Shah A, Su S, et al. Association of depressive symptoms and heart rate variability in Vietnam war-era twins: a longitudinal twin difference study. JAMA Psychiatry 2018 Jul 1;75(7):705-712. [doi: 10.1001/jamapsychiatry.2018.0747] [Medline: 29799951]
52. Hamilton JL, Alloy LB. Physiological markers of interpersonal stress generation in depression. Clin Psychol Sci 2017;5(6):911-929. [doi: 10.1177/2167702617720211] [Medline: 29123970]
53. Sun G, Shinba T, Kirimoto T, Matsui T. An objective screening method for major depressive disorder using logistic regression analysis of heart rate variability data obtained in a mental task paradigm. Front Psychiatry 2016;7:180. [doi: 10.3389/fpsyt.2016.00180] [Medline: 27867364]
54. Liang CS, Lee JF, Chen CC, Chang YC. Reactive heart rate variability in male patients with first-episode major depressive disorder. Prog Neuropsychopharmacol Biol Psychiatry 2015 Jan 2;56:52-57. [doi: 10.1016/j.pnpbp.2014.08.004] [Medline: 25149628]
55. Roh T, Hong S, Yoo HJ. Wearable depression monitoring system with heart-rate variability. Annu Int Conf IEEE Eng Med Biol Soc 2014;2014:562-565. [doi: 10.1109/EMBC.2014.6943653] [Medline: 25570021]
56. Lee CK, Yoo SK. An analysis of the relationship between self-reported anxiety, depressiveness and parameters of heart rate variability based on photoplethysmography [Article in Korean]. Sci Emot Sens 2012;3:345-354 [FREE Full text]
57. Zhang ZX, Tian XW, Lim JS. Neuro-fuzzy network-based depression diagnosis algorithm using optimal features of HRV. The Journal of the Korea Contents Association 2012 Feb 28;12(2):1-9. [doi: 10.5392/JKCA.2012.12.02.001]
58. Licht CMM, de Geus EJC, Zitman FG, Hoogendijk WJG, van Dyck R, Penninx B. Association between major depressive disorder and heart rate variability in the Netherlands Study of Depression and Anxiety (NESDA). Arch Gen Psychiatry 2008 Dec;65(12):1358-1367. [doi: 10.1001/archpsyc.65.12.1358] [Medline: 19047522]
59. Kojima M, Hayano J, Fukuta H, et al. Loss of fractal heart rate dynamics in depressive hemodialysis patients. Psychosom Med 2008 Feb;70(2):177-185. [doi: 10.1097/PSY.0b013e31816477a1] [Medline: 18256338]
60. Stroup-Benham CA, Markides KS, Black SA, Goodwin JS. Relationship between low blood pressure and depressive symptomatology in older people. J Am Geriatr Soc 2000 Mar;48(3):250-255. [doi: 10.1111/j.1532-5415.2000.tb02642.x] [Medline: 10733049]
61. Stein PK, Carney RM, Freedland KE, et al. Severe depression is associated with markedly reduced heart rate variability in patients with stable coronary heart disease. J Psychosom Res 2000;48(4-5):493-500. [doi: 10.1016/s0022-3999(99)00085-9] [Medline: 10880671]
62. Wang X, Pathiravasan CH, Zhang Y, et al. Association of depressive symptom trajectory with physical activity collected by mHealth devices in the Electronic Framingham Heart Study: cohort study. JMIR Ment Health 2023 Jul 14;10:e44529. [doi: 10.2196/44529] [Medline: 37450333]

XSL•FO
RenderX

63. Espino-Salinas CH, Galván-Tejada CE, Luna-García H, et al. Two-dimensional convolutional neural network for depression episodes detection in real time using motor activity time series of Depresjon dataset. Bioengineering (Basel) 2022 Sep 9;9(9):458. [doi: 10.3390/bioengineering9090458] [Medline: 36135004]
64. Hsiao C, Hsueh MC, Liao Y. Associations between objectively measured sedentary behavior patterns and depressive symptoms in older adults: a cross sectional study. Ment Health Phys Act 2022 Oct;23:100471. [doi: 10.1016/j.mhpa.2022.100471]
65. Ramsey CM, Lynch KG, Gehrman PR, et al. Daily steps and depressive symptoms: a longitudinal evaluation of patients with major depressive disorder in the precision medicine in mental health care study. J Affect Disord 2022 Mar 1;300:334-340. [doi: 10.1016/j.jad.2021.12.116] [Medline: 34979178]
66. Hussenoeder FS, Conrad I, Pabst A, et al. Physical activity and mental health: the connection between step count and depression, anxiety and quality of sleep. Psychol Health Med 2023;28(9):2419-2429. [doi: 10.1080/13548506.2022.2159453] [Medline: 36529963]
67. Michalak J, Niemeyer H, Tschacher W, Baumann N, Chi Zhang X, Adolph D. Subjective and objective measures of activity in depressed and non-depressed individuals in everyday life. J Exp Psychopathol 2022 Apr;13(2):20438087221092582. [doi: 10.1177/20438087221092582]
68. Li YM, Konstabel K, Mõttus R, Lemola S. Temporal associations between objectively measured physical activity and depressive symptoms: an experience sampling study. Front Psychiatry 2022;13:920580. [doi: 10.3389/fpsyt.2022.920580] [Medline: 35923450]
69. Figueroa CA, Vittinghoff E, Aguilera A, Fukuoka Y. Differences in objectively measured daily physical activity patterns related to depressive symptoms in community dwelling women - mPED trial. Prev Med Rep 2021 Jun;22:101325. [doi: 10.1016/j.pmedr.2021.101325] [Medline: 33659156]
70. Hsueh MC, Stubbs B, Lai YJ, Sun CK, Chen LJ, Ku PW. A dose response relationship between accelerometer assessed daily steps and depressive symptoms in older adults: a two-year cohort study. Age Ageing 2021 Feb 26;50(2):519-526. [doi: 10.1093/ageing/afaa162] [Medline: 32980870]
71. Jakobsen P, Garcia-Ceja E, Riegler M, et al. Applying machine learning in motor activity time series of depressed bipolar and unipolar patients compared to healthy controls. PLoS ONE 2020;15(8):e0231995. [doi: 10.1371/journal.pone.0231995] [Medline: 32833958]
72. Zhao N, Zhang Z, Wang Y, et al. See your mental state from your walk: recognizing anxiety and depression through Kinect-recorded gait data. PLoS ONE 2019;14(5):e0216591. [doi: 10.1371/journal.pone.0216591] [Medline: 31116785]
73. Averill IR, Crowe M, Frampton CM, et al. Clinical response to treatment in inpatients with depression correlates with changes in activity levels and psychomotor speed. Aust N Z J Psychiatry 2018 Jul;52(7):652-659. [doi: 10.1177/0004867417753549] [Medline: 29417833]
74. Ku PW, Steptoe A, Liao Y, Sun WJ, Chen LJ. Prospective relationship between objectively measured light physical activity and depressive symptoms in later life. Int J Geriatr Psychiatry 2018 Jan;33(1):58-65. [doi: 10.1002/gps.4672] [Medline: 28181713]
75. Holmquist S, Mattsson S, Schele I, Nordström P, Nordström A. Low physical activity as a key differentiating factor in the potential high-risk profile for depressive symptoms in older adults. Depress Anxiety 2017 Sep;34(9):817-825. [doi: 10.1002/da.22638] [Medline: 28489319]
76. Todder D, Caliskan S, Baune BT. Longitudinal changes of day-time and night-time gross motor activity in clinical responders and non-responders of major depression. World J Biol Psychiatry 2009;10(4):276-284. [doi: 10.3109/15622970701403081] [Medline: 19921969]
77. Lemke MR, Wendorff T, Mieth B, Buhl K, Linnemann M. Spatiotemporal gait patterns during over ground locomotion in major depression compared with healthy controls. J Psychiatr Res 2000;34(4-5):277-283. [doi: 10.1016/s0022-3956(00)00017-0] [Medline: 11104839]
78. Fadul R, Alfalahi H, Shehhi AA, Hadjileontiadis L. Depressive disorder remote detection through touchscreen typing behaviour. Annu Int Conf IEEE Eng Med Biol Soc 2023 Jul;2023:1-4. [doi: 10.1109/EMBC40787.2023.10340393] [Medline: 38082634]
79. Auerbach RP, Srinivasan A, Kirshenbaum JS, Mann JJ, Shankman SA. Geolocation features differentiate healthy from remitted depressed adults. J Psychopathol Clin Sci 2022 May;131(4):341-349. [doi: 10.1037/abn0000742] [Medline: 35230855]
80. Otte Andersen T, Skovlund Dissing A, Rosenbek Severinsen E, et al. Predicting stress and depressive symptoms using high-resolution smartphone data and sleep behavior in Danish adults. Sleep 2022 Jun 13;45(6):zsac067. [doi: 10.1093/sleep/zsac067] [Medline: 35298650]
81. Opoku Asare K, Terhorst Y, Vega J, Peltonen E, Lagerspetz E, Ferreira D. Predicting depression from smartphone behavioral markers using machine learning methods, hyperparameter optimization, and feature importance analysis: exploratory study. JMIR Mhealth Uhealth 2021 Jul 12;9(7):e26540. [doi: 10.2196/26540] [Medline: 34255713]
82. Chikersal P, Doryab A, Tumminia M, et al. Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: a machine learning approach with robust feature selection. ACM Trans Comput Hum Interact 2021;28(1):1-41. [doi: 10.1145/3422821]

XSL•FO
RenderX

83. Zhang Y, Folarin AA, Sun S, et al. Relationship between major depression symptom severity and sleep collected using a wristband wearable device: multicenter longitudinal observational study. JMIR Mhealth Uhealth 2021 Apr 12;9(4):e24604. [doi: 10.2196/24604] [Medline: 33843591]
84. Pedrelli P, Fedor S, Ghandeharioun A, et al. Monitoring changes in depression severity using wearable and mobile sensors. Front Psychiatry 2020;11:584711. [doi: 10.3389/fpsyt.2020.584711] [Medline: 33391050]
85. Mastoras RE, Iakovakis D, Hadjidimitriou S, et al. Touchscreen typing pattern analysis for remote detection of the depressive tendency. Sci Rep 2019 Sep 16;9(1):13414. [doi: 10.1038/s41598-019-50002-9] [Medline: 31527640]
86. Saeb S, Zhang M, Karr CJ, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J Med Internet Res 2015 Jul 15;17(7):e175. [doi: 10.2196/jmir.4273] [Medline: 26180009]
87. Wiseman M, Yep R, Wood Alexander M, et al. Objective speech measures capture depressive symptoms and associated cognitive difficulties. Transl Psychiatry 2025 Nov 17;15(1):525. [doi: 10.1038/s41398-025-03728-2] [Medline: 41257835]
88. Wadle LM, Ebner-Priemer UW, Foo JC, et al. Speech features as predictors of momentary depression severity in patients with depressive disorder undergoing sleep deprivation therapy: ambulatory assessment pilot study. JMIR Ment Health 2024 Jan 18;11:e49222. [doi: 10.2196/49222] [Medline: 38236637]
89. Silva WJ, Lopes L, Galdino MKC, Almeida AA. Voice acoustic parameters as predictors of depression. J Voice 2024 Jan;38(1):77-85. [doi: 10.1016/j.jvoice.2021.06.018] [Medline: 34353686]
90. Kim AY, Jang EH, Lee SH, Choi KY, Park JG, Shin HC. Automatic depression detection using smartphone-based text-dependent speech signals: deep convolutional neural network approach. J Med Internet Res 2023 Jan 25;25:e34474. [doi: 10.2196/34474] [Medline: 36696160]
91. Wang Y, Liang L, Zhang Z, et al. Fast and accurate assessment of depression based on voice acoustic features: a cross-sectional and longitudinal study. Front Psychiatry 2023;14:1195276. [doi: 10.3389/fpsyt.2023.1195276]
92. Zhao Q, Fan HZ, Li YL, et al. Vocal acoustic features as potential biomarkers for identifying/diagnosing depression: a cross-sectional study. Front Psychiatry 2022;13:815678. [doi: 10.3389/fpsyt.2022.815678] [Medline: 35573349]
93. Ye J, Yu Y, Wang Q, et al. Multi-modal depression detection based on emotional audio and evaluation text. J Affect Disord 2021 Dec 1;295:904-913. [doi: 10.1016/j.jad.2021.08.090] [Medline: 34706461]
94. Klangpornkun N, Ruangritchai M, Munthuli A, et al. Classification of depression and other psychiatric conditions using speech features extracted from a Thai psychiatric and verbal screening test. Annu Int Conf IEEE Eng Med Biol Soc 2021 Nov;2021:651-656. [doi: 10.1109/EMBC46164.2021.9629571] [Medline: 34891377]
95. Demiroglu C, Beşirli A, Ozkanca Y, Çelik S. Depression-level assessment from multi-lingual conversational speech data using acoustic and text features. J Audio Speech Music Proc 2020 Dec;2020(1):17. [doi: 10.1186/s13636-020-00182-4]
96. Yamamoto M, Takamiya A, Sawada K, et al. Using speech recognition technology to investigate the association between timing-related speech features and depression severity. PLoS ONE 2020;15(9):e0238726. [doi: 10.1371/journal.pone.0238726] [Medline: 32915846]
97. Ali FZ, Parsey RV, Lin S, Schwartz J, DeLorenzo C. Circadian rhythm biomarker from wearable device data is related to concurrent antidepressant treatment response. NPJ Digit Med 2023 Apr 29;6(1):81. [doi: 10.1038/s41746-023-00827-6] [Medline: 37120493]
98. Braund TA, Zin MT, Boonstra TW, et al. Smartphone sensor data for identifying and monitoring symptoms of mood disorders: a longitudinal observational study. JMIR Ment Health 2022 May 4;9(5):e35549. [doi: 10.2196/35549] [Medline: 35507385]
99. Choi JG, Ko I, Han S. Depression level classification using machine learning classifiers based on actigraphy data. IEEE Access 2021;9:116622-116646. [doi: 10.1109/ACCESS.2021.3105393]
100. George SV, Kunkels YK, Booij S, Wichers M. Uncovering complexity details in actigraphy patterns to differentiate the depressed from the non-depressed. Sci Rep 2021 Jun 29;11(1):13447. [doi: 10.1038/s41598-021-92890-w] [Medline: 34188115]
101. Smagula SF, Ancoli-Israel S, Blackwell T, et al. Circadian rest-activity rhythms predict future increases in depressive symptoms among community-dwelling older men. Am J Geriatr Psychiatry 2015 May;23(5):495-505. [doi: 10.1016/j.jagp.2014.06.007] [Medline: 25066948]
102. Maglione JE, Ancoli-Israel S, Peters KW, et al. Depressive symptoms and circadian activity rhythm disturbances in community-dwelling older women. Am J Geriatr Psychiatry 2014 Apr;22(4):349-361. [doi: 10.1016/j.jagp.2012.09.003] [Medline: 23567424]
103. Luik AI, Zuurbier LA, Hofman A, Van Someren EJW, Tiemeier H. Stability and fragmentation of the activity rhythm across the sleep-wake cycle: the importance of age, lifestyle, and mental health. Chronobiol Int 2013 Dec;30(10):1223-1230. [doi: 10.3109/07420528.2013.813528] [Medline: 23971909]
104. Anik IA, Kamal AHM, Kabir MA, Uddin S, Moni MA. A robust deep-learning model to detect major depressive disorder utilizing EEG signals. IEEE Trans Artif Intell 2024;5(10):4938-4947. [doi: 10.1109/TAI.2024.3394792]
105. Tian F, Zhang H, Tan Y, et al. An on-board executable multi-feature transfer-enhanced fusion model for three-lead EEG sensor-assisted depression diagnosis. IEEE J Biomed Health Inform 2025 Jan;29(1):152-165. [doi: 10.1109/JBHI.2024.3487012] [Medline: 39466874]

XSL•FO
RenderX

106. Islam R, Bae SW. FacePsy: an open-source affective mobile sensing system - analyzing facial behavior and head gesture for depression detection in naturalistic settings. Proc ACM Hum-Comput Interact 2024 Sep 24;8(MHCI):1-32. [doi: 10.1145/3676505]
107. Makhmutova M, Kainkaryam R, Ferreira M, Min J, Jaggi M, Clay I. Predicting changes in depression severity using the PSYCHE-D (Prediction of Severity Change-Depression) model involving person-generated health data: longitudinal case-control observational study. JMIR Mhealth Uhealth 2022 Mar 25;10(3):e34148. [doi: 10.2196/34148] [Medline: 35333186]
108. Price GD, Heinz MV, Collins AC, Jacobson NC. Detecting major depressive disorder presence using passively-collected wearable movement data in a nationally-representative sample. Psychiatry Res 2024 Feb;332:115693. [doi: 10.1016/j.psychres.2023.115693] [Medline: 38194801]
109. Bae S, Jang M, Kim GM, et al. Nonlinear associations between physical function, physical activity, sleep, and depressive symptoms in older adults. J Clin Med 2023 Sep 16;12(18):6009. [doi: 10.3390/jcm12186009] [Medline: 37762948]
110. Griffiths C, da Silva KM, Leathlean C, Jiang H, Ang CS, Searle R. Investigation of physical activity, sleep, and mental health recovery in treatment resistant depression (TRD) patients receiving repetitive transcranial magnetic stimulation (rTMS) treatment. J Affect Disord Rep 2022 Apr;8:100337. [doi: 10.1016/j.jadr.2022.100337] [Medline: 35619990]
111. Mishra R, Park C, York MK, et al. Decrease in mobility during the COVID-19 pandemic and its association with increase in depression among older adults: a longitudinal remote mobility monitoring using a wearable sensor. Sensors (Basel) 2021 Apr 29;21(9):3090. [doi: 10.3390/s21093090] [Medline: 33946664]
112. Raoux N, Benoit O, Dantchev N, et al. Circadian pattern of motor activity in major depressed patients undergoing antidepressant therapy: relationship between actigraphic measures and clinical course. Psychiatry Res 1994 Apr;52(1):85-98. [doi: 10.1016/0165-1781(94)90122-8] [Medline: 8047624]
113. Rykov Y, Thach TQ, Bojic I, Christopoulos G, Car J. Digital biomarkers for depression screening with wearable devices: cross-sectional study with machine learning modeling. JMIR Mhealth Uhealth 2021 Oct 25;9(10):e24872. [doi: 10.2196/24872] [Medline: 34694233]
114. Sayar K, Güleç H, Gökçe M, Ismail AK. Heart rate variability in depressed patients. Psychiatry Clin Psychopharmacol 2002;12(3):130-133 [FREE Full text]
115. Landolt HP, Gillin JC. Similar sleep EEG topography in middle-aged depressed patients and healthy controls. Sleep 2005 Feb;28(2):239-247. [doi: 10.1093/sleep/28.2.239] [Medline: 16171249]
116. Hoyos CM, Gordon C, Terpening Z, et al. Circadian rhythm and sleep alterations in older people with lifetime depression: a case-control study. BMC Psychiatry 2020 Apr 29;20(1):192. [doi: 10.1186/s12888-020-02606-z] [Medline: 32349697]
117. Lorenz N, Spada J, Sander C, Riedel-Heller SG, Hegerl U. Circadian skin temperature rhythms, circadian activity rhythms and sleep in individuals with self-reported depressive symptoms. J Psychiatr Res 2019 Oct;117:38-44. [doi: 10.1016/j.jpsychires.2019.06.022] [Medline: 31279242]
118. Vanderlind WM, Beevers CG, Sherman SM, et al. Sleep and sadness: exploring the relation among sleep, cognitive control, and depressive symptoms in young adults. Sleep Med 2014 Jan;15(1):144-149. [doi: 10.1016/j.sleep.2013.10.006] [Medline: 24332565]
119. Ahmed A, Ramesh J, Ganguly S, Aburukba R, Sagahyroon A, Aloul F. Investigating the feasibility of assessing depression severity and valence-arousal with wearable sensors using discrete wavelet transforms and machine learning. Information 2022;13(9):406. [doi: 10.3390/info13090406]
120. Choi J, Lee S, Kim S, Kim D, Kim H. Depressed mood prediction of elderly people with a wearable band. Sensors (Basel) 2022;22(11):4174. [doi: 10.3390/s22114174]
121. Mahendran N, Vincent DR, Srinivasan K, et al. Sensor-assisted weighted average ensemble model for detecting major depressive disorder. Sensors (Basel) 2019 Nov 6;19(22):4822. [doi: 10.3390/s19224822] [Medline: 31698678]
122. Xu X, Chikersal P, Doryab A, et al. Leveraging routine behavior and contextually-filtered features for depression detection among college students. Proc ACM Interact Mob Wearable Ubiquitous Technol 2019 Sep 9;3(3):1-33. [doi: 10.1145/3351274]
123. Lu J, Shang C, Yue C, et al. Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning. Proc ACM Interact Mob Wearable Ubiquitous Technol 2018 Mar 26;2(1):1-21. [doi: 10.1145/3191753]
124. Farhan AA, Yue C, Morillo R, et al. Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data. Presented at: 2016 IEEE Wireless Health (WH); Oct 25-27, 2026. [doi: 10.1109/WH.2016.7764553]
125. Karimi S, Nateghi M, Cestero GI, et al. Prescreening depression using wearable electrocardiogram and photoplethysmogram data from a psycholinguistic experiment. Physiol Meas 2025 Aug 2;46(8):085004. [doi: 10.1088/1361-6579/adf6fe] [Medline: 40752509]
126. Jacobson NC, Chung YJ. Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones. Sensors (Basel) 2020 Jun 24;20(12):3572. [doi: 10.3390/s20123572] [Medline: 32599801]
127. Wang R, Wang W, Dasilva A, et al. Tracking depression dynamics in college students using mobile phone and wearable sensing. Proc ACM Interact Mob Wearable Ubiquitous Technol 2018 Mar;2(1):1-26. [doi: 10.1145/3191775] [Medline: 39449996]

XSL•FO
RenderX

128. Di Matteo D, Fotinos K, Lokuge S, et al. The relationship between smartphone-recorded environmental audio and symptomatology of anxiety and depression: exploratory study. JMIR Form Res 2020 Aug 13;4(8):e18751. [doi: 10.2196/18751] [Medline: 32788153]
129. Zhou Y, Yao X, Han W, Wang Y, Li Z, Li Y. Distinguishing apathy and depression in older adults with mild cognitive impairment using text, audio, and video based on multiclass classification and shapely additive explanations. Int J Geriat Psychiatry 2022 Nov;37(11):2345-2356. [doi: 10.1002/gps.5827]
130. Williamson JR, Young D, Nierenberg AA, Niemi J, Helfer BS, Quatieri TF. Tracking depression severity from audio and video based on speech articulatory coordination. Comput Speech Lang 2019 May;55:40-56. [doi: 10.1016/j.csl.2018.08.004] [Medline: 32773961]
131. Shah RV, Grennan G, Zafar-Khan M, et al. Personalized machine learning of depressed mood using wearables. Transl Psychiatry 2021 Jun 9;11(1):338. [doi: 10.1038/s41398-021-01445-0] [Medline: 34103481]
132. Opoku Asare K, Moshe I, Terhorst Y, et al. Mood ratings and digital biomarkers from smartphone and wearable data differentiates and predicts depression status: a longitudinal data analysis. Pervasive Mob Comput 2022 Jul;83:101621. [doi: 10.1016/j.pmcj.2022.101621]
133. Narziev N, Goh H, Toshnazarov K, Lee SA, Chung KM, Noh Y. STDD: short-term depression detection with passive sensing. Sensors (Basel) 2020 Mar 4;20(5):1396. [doi: 10.3390/s20051396] [Medline: 32143358]
134. Kim H, Lee S, Lee S, Hong S, Kang H, Kim N. Depression prediction by using ecological momentary assessment, Actiwatch data, and machine learning: observational study on older adults living alone. JMIR Mhealth Uhealth 2019 Oct 16;7(10):e14149. [doi: 10.2196/14149] [Medline: 31621642]
135. Kumagai N, Tajika A, Hasegawa A, et al. Predicting recurrence of depression using lifelog data: an explanatory feasibility study with a panel VAR approach. BMC Psychiatry 2019 Dec 11;19(1):391. [doi: 10.1186/s12888-019-2382-2] [Medline: 31829206]
136. Funkhouser CJ, Weiner LS, Crowley RN, et al. Early changes in passively sensed homestay predict depression symptom improvement during digital behavioral activation. Behav Res Ther 2025 Oct;193:104815. [doi: 10.1016/j.brat.2025.104815] [Medline: 40614686]
137. Di Matteo D, Fotinos K, Lokuge S, et al. Automated screening for social anxiety, generalized anxiety, and depression from objective smartphone-collected data: cross-sectional study. J Med Internet Res 2021 Aug 13;23(8):e28918. [doi: 10.2196/28918] [Medline: 34397386]
138. Sverdlov O, Curcic J, Hannesdottir K, et al. A study of novel exploratory tools, digital technologies, and central nervous system biomarkers to characterize unipolar depression. Front Psychiatry 2021;12:640741. [doi: 10.3389/fpsyt.2021.640741] [Medline: 34025472]
139. Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health. Psychiatr Rehabil J 2015 Sep;38(3):218-226. [doi: 10.1037/prj0000130] [Medline: 25844912]
140. Minaeva O, Riese H, Lamers F, Antypa N, Wichers M, Booij SH. Screening for depression in daily life: development and external validation of a prediction model based on actigraphy and experience sampling method. J Med Internet Res 2020 Dec 1;22(12):e22634. [doi: 10.2196/22634] [Medline: 33258783]
141. Dai R, Kannampallil T, Zhang J, Lv N, Ma J, Lu C. Multi-task learning for randomized controlled trials: a case study on predicting depression with wearable data. Proc ACM Interact Mob Wearable Ubiquitous Technol 2022;6(2):1-23. [doi: 10.1145/3534591]
142. Tazawa Y, Liang KC, Yoshimura M, et al. Evaluating depression with multimodal wristband-type wearable device: screening and assessing patient severity utilizing machine-learning. Heliyon 2020 Feb;6(2):e03274. [doi: 10.1016/j.heliyon.2020.e03274] [Medline: 32055728]
143. Thati RP, Dhadwal AS, Kumar P, Sainaba P. Multimodal depression detection: using fusion strategies with smart phone usage and audio-visual behavior. Int J Artif Intell Tools 2023 Mar;32(2):2340008. [doi: 10.1142/S0218213023400080]
144. Jiang Z, Seyedi S, Griner E, et al. Multimodal mental health digital biomarker analysis from remote interviews using facial, vocal, linguistic, and cardiovascular patterns. IEEE J Biomed Health Inform 2024 Mar;28(3):1680-1691. [doi: 10.1109/JBHI.2024.3352075] [Medline: 38198249]
145. Moshe I, Terhorst Y, Opoku Asare K, et al. Predicting symptoms of depression and anxiety using smartphone and wearable data. Front Psychiatry 2021;12:625247. [doi: 10.3389/fpsyt.2021.625247] [Medline: 33584388]
146. Wang R, Chen F, Chen Z, et al. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. Presented at: UbiComp '14: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing; Sep 13-17, 2014. [doi: 10.1145/2632048.2632054]
147. Song S, Seo Y, Hwang S, Kim HY, Kim J. Digital phenotyping of geriatric depression using a community-based digital mental health monitoring platform for socially vulnerable older adults and their community caregivers: 6-week living lab single-arm pilot study. JMIR Mhealth Uhealth 2024 Jun 17;12:e55842. [doi: 10.2196/55842] [Medline: 38885033]
148. Bai R, Xiao L, Guo Y, et al. Tracking and monitoring mood stability of patients with major depressive disorder by machine learning models using passive digital data: prospective naturalistic multicenter study. JMIR Mhealth Uhealth 2021 Mar 8;9(3):e24365. [doi: 10.2196/24365] [Medline: 33683207]

XSL•FO
RenderX

149. Borelli JL, Wang Y, Li FH, et al. Detection of depressive symptoms in college students using multimodal passive sensing data and light gradient boosting machine: longitudinal pilot study. JMIR Form Res 2025 Jun 3;9:e67964. [doi: 10.2196/67964] [Medline: 40460426]
150. Chen J, Chan NY, Li CT, et al. Multimodal digital assessment of depression with actigraphy and app in Hong Kong Chinese. Transl Psychiatry 2024 Mar 18;14(1):150. [doi: 10.1038/s41398-024-02873-4] [Medline: 38499546]
151. Shen S, Qi W, Zeng J, et al. Passive sensing for mental health monitoring using machine learning with wearables and smartphones: scoping review. J Med Internet Res 2025 Aug 14;27:e77066. [doi: 10.2196/77066] [Medline: 40811794]
152. Lee YJ, Lee JY, Cho JH, Kang YJ, Choi JH. Performance of consumer wrist-worn sleep tracking devices compared to polysomnography: a meta-analysis. J Clin Sleep Med 2025 Mar 1;21(3):573-582. [doi: 10.5664/jcsm.11460] [Medline: 39484805]
153. Poon K, Ho MSH, Tai APL, Leung MK, Chan MCM, Lau WKW. Mediating roles of sleep quality and resilience in the relationships between chronotypes and mental health symptoms. Sci Rep 2024 Mar 11;14(1):5874. [doi: 10.1038/s41598-024-56688-w] [Medline: 38467740]
154. Röver C, Knapp G, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. BMC Med Res Methodol 2015 Nov 14;15:99. [doi: 10.1186/s12874-015-0091-1] [Medline: 26573817]
155. Su Y, Li M, Caron J, Li D, Meng X. Differential effects of lifetime stressors on major depressive disorder severity: a longitudinal community-based cohort study. Eur Psychiatry 2024 Oct 4;67(1):e66. [doi: 10.1192/j.eurpsy.2024.1783] [Medline: 39363747]
156. Price GD, Heinz MV, Song SH, Nemesure MD, Jacobson NC. Using digital phenotyping to capture depression symptom variability: detecting naturalistic variability in depression symptoms across one year using passively collected wearable movement and sleep data. Transl Psychiatry 2023 Dec 9;13(1):381. [doi: 10.1038/s41398-023-02669-y] [Medline: 38071317]
157. Yasugaki S, Okamura H, Kaneko A, Hayashi Y. Bidirectional relationship between sleep and depression. Neurosci Res 2025 Feb;211:57-64. [doi: 10.1016/j.neures.2023.04.006] [Medline: 37116584]
158. Meng Y, Ma N, Shi Y, et al. The association of physical activity and sedentary behavior with depression in US adults: NHANES 2007-2018. Front Public Health 2024;12:1404407. [doi: 10.3389/fpubh.2024.1404407] [Medline: 38975361]
159. Hayashi M, Takeshima M, Hosoya T, Kume Y. 24-hour rest-activity rhythm in middle-aged and older persons with depression. Int J Environ Res Public Health 2023 Mar 27;20(7):5275. [doi: 10.3390/ijerph20075275] [Medline: 37047891]
160. Edler JS, Winter M, Steinmetz H, Cohrdes C, Baumeister H, Pryss R. Predicting depressive symptoms using GPS-based regional data in Germany with the CORONA HEALTH app during the COVID-19 pandemic: cross-sectional study. Interact J Med Res 2024 Dec 3;13:e53248. [doi: 10.2196/53248] [Medline: 39625745]
161. Alshanskaia EI, Zhozhikashvili NA, Polikanova IS, Martynova OV. Heart rate response to cognitive load as a marker of depression and increased anxiety. Front Psychiatry 2024;15:1355846. [doi: 10.3389/fpsyt.2024.1355846] [Medline: 39056018]
162. Xu L, Zhai X, Shi D, Zhang Y. Depression and coronary heart disease: mechanisms, interventions, and treatments. Front Psychiatry 2024;15:1328048. [doi: 10.3389/fpsyt.2024.1328048] [Medline: 38404466]
163. Monaco F, Vignapiano A, Piacente M, et al. Innova4Health: an integrated approach for prevention of recurrence and personalized treatment of major depressive disorder. Front Artif Intell 2024;7:1366055. [doi: 10.3389/frai.2024.1366055] [Medline: 38774832]
164. Choo M, Park D, Cho M, Bae S, Kim J, Han DH. Exploring a multimodal approach for utilizing digital biomarkers for childhood mental health screening. Front Psychiatry 2024;15:1348319. [doi: 10.3389/fpsyt.2024.1348319] [Medline: 38666089]
165. Mandryk RL, Birk MV, Vedress S, et al. Remote assessment of depression using digital biomarkers from cognitive tasks. Front Psychol 2021;12:767507. [doi: 10.3389/fpsyg.2021.767507] [Medline: 34975656]
166. Hurwitz E, Butzin-Dozier Z, Master H, et al. Harnessing consumer wearable digital biomarkers for individualized recognition of postpartum depression using the All of Us research program data set: cross-sectional study. JMIR Mhealth Uhealth 2024 May 2;12:e54622. [doi: 10.2196/54622] [Medline: 38696234]
167. Chen J, Lai TF, Lin LJ, Park JH, Liao Y. Is overall and timing-specific physical activity associated with depression in older adults? Front Public Health 2023;11:567123. [doi: 10.3389/fpubh.2023.1241170]
168. Hernández-Vicente A, Hernando D, Marín-Puyalto J, et al. Validity of the Polar H7 heart rate sensor for heart rate variability analysis during exercise in different age, body composition and fitness level groups. Sensors (Basel) 2021 Jan 29;21(3):902. [doi: 10.3390/s21030902] [Medline: 33572800]
169. Sato W, Murata K, Uraoka Y, Shibata K, Yoshikawa S, Furuta M. Emotional valence sensing using a wearable facial EMG device. Sci Rep 2021 Mar 11;11(1):5757. [doi: 10.1038/s41598-021-85163-z] [Medline: 33707605]
170. Caldiroli A, Capuzzi E, Affaticati LM, et al. Candidate biological markers for social anxiety disorder: a systematic review. Int J Mol Sci 2023 Jan 3;24(1):835. [doi: 10.3390/ijms24010835] [Medline: 36614278]

Abbreviations

BDI: Beck Depression Inventory

XSL•FO
RenderX

CES-D: Center for Epidemiologic Studies Depression Scale
ECG: electrocardiogram
GDS: Geriatric Depression Scale
GRADE: Grading of Recommendations, Assessment, Development, and Evaluation
HDRS: Hamilton Depression Rating Scale
HF: high frequency
HKSJ: Hartung-Knapp-Sidik-Jonkman
HR: heart rate
HRV: heart rate variability
IDAS: Inventory of Depression and Anxiety Symptoms
IDS-SR: Inventory of Depressive Symptomatology
IRB: Institutional Review Board
LF: low frequency
LPA: light physical activity
MD: mean difference
MDD: major depressive disorder
MET: metabolic equivalent of task
mHR: mean heart rate
MVPA: moderate-to-vigorous physical activity
NREM: non–rapid eye movement
PHQ: Patient Health Questionnaire
pNN: proportion of normal-to-normal intervals
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies-2
REM: rapid eye movement
RMSSD: root mean square of successive differences
RRI: RR interval
SDNN: SD of normal-to-normal intervals
SE: sleep efficiency
SGDS-K: shortened Korean version of the Geriatric Depression Scale
SIGN: Scottish Intercollegiate Guidelines Network
SMD: standardized mean difference
SOL: sleep onset latency
SWS: slow-wave sleep
TIB: time in bed
TST: total sleep time
ULF: ultra-low frequency
VLF: very low frequency
WASO: wake after sleep onset

Edited by S Brini; submitted 23.Apr.2025; peer-reviewed by S Oh, V Manea; revised version received 11.Feb.2026; accepted 13.Feb.2026; published 02.Apr.2026.

Please cite as:
Lee H, Kang SG, Lee S
The Role of Digital Biomarkers in Physiological Signal-Based Depression Assessment: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e76432
URL: https://www.jmir.org/2026/1/e76432
doi:10.2196/76432

© Hyeongsuk Lee, Seung-Gul Kang, SeonHeui Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Accuracy of Radiomics-Based Machine Learning for Predicting Risk of Recurrence in Non–Small Cell Lung Cancer: Systematic Review and Meta-Analysis

Junpei Wu[1,2], MSc; Ye Zhang[3], MSc; Jiaye Wang[4], MSc; Chengshui Chen[5], PhD; Shiyu Hu[2*], MSc; Wenyu Chen[2,5*], PhD

[1]Department of Pulmonary and Critical Care Medicine, Yiwu Central Hospital, Yiwu, China

[2]Department of Respiratory, The Affiliated Hospital of Jiaxing University, 1882 Zhonghuan South Road, Nanhu District, Jiaxing, China

[3]Health Management Center, The Affiliated Hospital of Jiaxing University, Jiaxing, China

[4]Department of Internal Medicine, Zhejiang Chinese Medical University, Hangzhou, China

[5]Department of Respiratory, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China

[*]these authors contributed equally

Corresponding Author:
Wenyu Chen, PhD
Department of Respiratory, The Affiliated Hospital of Jiaxing University, 1882 Zhonghuan South Road, Nanhu District, Jiaxing, China

Abstract

Background: During the diagnosis and treatment of non–small cell lung cancer (NSCLC), detecting the risk of its recurrence in an early phase is still challenging. Recent studies have investigated the radiomics-based machine learning (ML) models for detecting the risk of recurrence in NSCLC. However, there is still insufficient systematic evidence to prove its efficiency.

Objective: This study is designed to systematically evaluate the effectiveness of radiomics-based ML in predicting the risk of recurrence in NSCLC, aiming to provide evidence-based support for the subsequent development of scoring tools to forecast recurrence risk.

Methods: For acquiring research on radiomics-based models for forecasting the risk of recurrence in NSCLC, Cochrane Library, Web of Science, PubMed, and Embase were systematically retrieved, up to October 24, 2025. Studies on analyzing the recurrence of NSCLC using radiomics-based ML were included, while those in which only texture analysis was conducted or radiomics-based ML was not constructed were excluded. The Radiomics Quality Score (RQS) was used to appraise the eligible studies. Subgroup analyses were conducted according to the variables of the model, the background of treatment, the stage of lung cancer, and the pathological type.

Results: Ultimately, 30 eligible studies in total were included, covering 7964 patients with NSCLC. According to the meta-analysis, the c-index of radiomics-based ML models for forecasting the risk of recurrence in NSCLC was 0.850 (95% CI 0.834 - 0.866, 95% prediction interval [PI] 0.623 - 1.004) in the training set. Specifically, the pooled c-index was 0.876 (95% CI 0.853 - 0.900) among the patients receiving the stereotactic body radiation therapy and 0.825 (95% CI 0.804 - 0.848) among those who received surgeries combined with other adjuvant treatment regimens. The c-index of the radiomics-based ML models combined with clinical features for forecasting the risk of recurrence in NSCLC was 0.833 (95% CI 0.822 - 0.854, 95% PI 0.717 - 0.945) in the training set. In contrast, the c-index of radiomics-based ML models for forecasting the risk of recurrence in NSCLC was 0.878 (95% CI 0.854 - 0.902, 95% PI 0.681 - 1.000) in the validation set. The c-index of radiomics-based ML models combined with clinical features for forecasting the risk of recurrence in NSCLC was 0.854 (95% CI 0.830 - 0.878, 95% PI 0.655 - 0.992) in the validation set. The average RQS across the included studies was 27.4%, revealing methodological limitations and an absence of standardization.

Conclusions: This study is the first to confirm that radiomics-based ML models effectively predict the risk of recurrence in NSCLC. This study provides evidence-based support for the subsequent development or updating of radiomics-based ML models. However, the current methodological application of radiomics remains concerning. Therefore, in the future, research should standardize the workflow for implementing radiomics-based ML and incorporate multicenter imaging data to enhance its generalizability.

Trial Registration: PROSPERO CRD42025631191; https://www.crd.york.ac.uk/PROSPERO/view/CRD42025631191

(J Med Internet Res 2026;28:e84223) doi:10.2196/84223

XSL•FO
RenderX

KEYWORDS

non–small cell lung cancer; NSCLC; machine learning; ML; radiomics; recurrence; prognostic prediction

Introduction

As documented in the Cancer Statistics Report 2024, lung cancer serves as a major contributor to global cancer-related mortality and ranks second in the incidence of new cancer cases among male individuals [1]. With an 85% incidence among individuals with lung cancer, non–small cell lung cancer (NSCLC) emerges as the most frequently observed pathological subtype, and its incidence is gradually increasing [2]. Hence, enhancing the prevention, screening, and management of NSCLC is clinically significant.

The current treatment methods for NSCLC are constantly evolving. Early-stage NSCLC surgeries and radiotherapy, which are traditional treatment approaches, are still widely used and have helped numerous patients achieve promising therapeutic outcomes. In recent years, emerging therapies such as molecular targeted therapy, antitumor angiogenesis therapy, and immunotherapy [3-5] for NSCLC have significantly progressed. Those innovative treatment methods offer patients more options and significantly improve their survival rates and quality of life. Although those treatment methods have led to significant survival benefits for patients with NSCLC, there are still some serious adverse prognoses, posing significant challenges in the formulation of treatment plans in clinical practice [6,7]. In particular, recurrence is a serious adverse outcome in NSCLC treatment. According to the epidemiological statistics on lung cancer, the 5-year cumulative recurrence rate of NSCLC (stage I) after surgery is 25% to 30%, and it is as high as 60% to 70% for NSCLC (stage IIIA) [8,9]. Therefore, early identification of the risk of recurrence in NSCLC is clinically significant for formulating individualized therapeutic schedules as well as rehabilitation plans.

Unfortunately, up to now, there has been a lack of tools for effectively predicting the recurrence of NSCLC in clinical practice. In recent years, with artificial intelligence evolving, radiomics has received widespread attention in clinical practice. Radiomics refers to the process of converting a group of high-data-content medical images into digital ones and combining machine learning (ML) to aid in the diagnosis of diseases, prediction of progression, and prognosis of diseases [10]. Radiomics was originally proposed by Lambin et al [11] from the Netherlands in 2012. In comparison with traditional models, radiomics-based models are capable of extracting imaging features through high-throughput methods, which can more comprehensively reflect the spatial heterogeneity of tumors and are equipped with higher information dimensions and discriminative capabilities. Therefore, it has shown significant advantages in multiple studies. The study by Chen et al [12] has revealed a high c-index of radiomics-based ML models for forecasting the risk of recurrence in NSCLC (the stage of early epidermal growth factor receptor mutation). In the study by Yan et al [13], a radiomics-based ML is effective for predicting lymph node metastasis in NSCLC.

Beyond conventional ML algorithms based on radiomics, recent research has also explored more advanced deep learning techniques and AI algorithms capable of perceiving uncertainty for medical prediction. These models use dimensionality and techniques of size reduction for feature selection, maintaining robustness even in datasets with incomplete or noisy information. For instance, deep ML algorithms integrating feature elimination and dimensionality reduction have been successfully applied to the diagnosis of thyroid cancer with randomly missing data. Additionally, big data-driven supervised ML models based on similar mechanisms have demonstrated high diagnostic performance in nonalcoholic steatohepatitis [14,15]. These advances highlight the potential of incorporating high-dimensional feature learning and uncertainty modeling into frameworks of medical prediction. However, their application in forecasting the recurrence of NSCLC remains limited, and no systematic review has been conducted to synthesize current evidence.

Against this backdrop, some researchers have sought to establish radiomics-based ML models to forecast the risk of recurrence in NSCLC. However, current studies are restricted by limited sample size, heterogeneous design, and data sourced from a single center [16]. These studies are absent from systematic quantitative integration and heterogeneity analysis, thereby failing to provide high-quality and evidence-based support. This poses a challenge to the development of intelligent models for predicting the risk of recurrence in NSCLC. Therefore, this study is designed to systematically evaluate the predictive performance of radiomics-based ML models for the recurrence risk of NSCLC and their current application in this field. This study provides evidence-based support for the future standardized application of radiomics and the development or update of prediction and assessment tools based on radiomics-based ML.

Methods

Registration

This study has been prospectively registered on the International Prospective Register of Systematic Reviews (PROSPERO) and was implemented strictly following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines to guarantee transparency and repeatability (Checklist 1) [17].

Eligibility Criteria

Inclusion Criteria

The inclusion criteria were as follows: (1) individuals pathologically or clinically diagnosed with NSCLC; (2) studies on the recurrence in NSCLC using radiomics-based ML; (3) studies reporting outcome indicators of the accuracy of ML; (4) cohort studies, case-control studies, or cross-sectional studies; and (5) studies reported in English.

XSL•FO
RenderX

Exclusion Criteria

The exclusion criteria were as follows: (1) unpublished conference abstracts were included; (2) in studies, an ML model only based on clinical features was constructed for predicting the risk of recurrence in NSCLC, and a radiomics-based ML model was not established; (3) studies failed to strictly distinguish NSCLC from other types of lung cancers; and (4) in studies, outcome indicators were absent from appraising the predictive precision of an ML model, including c-index, sensitivity, specificity, accuracy, recovery rate, precision rate, confusion matrix, and F_1-score.

Data Sources and Search Strategy

To ensure transparency, the literature search was conducted in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search Extension) guidelines (Checklist 2) [18]. PubMed, Cochrane Library, Embase, and Web of Science were systematically searched for studies on predicting NSCLC recurrence based on radiomics from the inception of the databases to October 24, 2025, with no restrictions on region or year of publication. No specialized study registries were searched, nor were relevant experts contacted to obtain unpublished data. The reference lists of all included studies were manually checked (backward searching) to avoid missing potentially relevant studies. The search strategy was a combination of subject terms with relevant free terms. The subject terms comprised lung neoplasms (MeSH), machine learning (MeSH), and recurrence (MeSH). The search strategy for each database and the retrieval process are detailed in Table S2 in Multimedia Appendix 1 and Checklist 2. Prepublished search filters or search strategies for other literature reviews were employed, irrespective of language or study design.

Literature Screening and Data Collection

The retrieved studies were imported into EndNote (Clarivate). After the elimination of duplicates based on title, author, and publication year, the studies were systematically reviewed for their titles and abstracts. Ultimately, the eligible studies were determined through a full-text review. In the included studies, recurrence was defined as recurrence of tumors within or adjacent to the treatment zone, mediastinal recurrence, or distant metastasis following curative-intent treatment such as surgery, chemoradiotherapy, or stereotactic body radiotherapy. Recurrence was generally identified according to the criteria of response evaluation criteria in solid tumors or confirmed via histopathology. To ensure accuracy, in several studies, double-blind evaluation was performed independently by 2 radiologists or oncologists, with disagreements resolved through negotiation. Only studies reporting a clear definition of recurrence and an explicit criterion for follow-up were included in this meta-analysis. To address the high dimensionality of radiomics data, procedures of multistep feature selection and dimensionality reduction were used across the included studies. Commonly applied methods included reproducibility filtering via intraclass correlation coefficient, redundancy removal based on correlation, univariate statistical screening, and feature selection based on ML, such as least absolute shrinkage and selection operator, recursive feature elimination, or principal component analysis. These approaches reduced feature space, enhanced the steadiness of models, thereby ensuring that only the most robust and clinically relevant features were utilized in the construction of prediction models.

Before extracting the data, a standardized spreadsheet for data extraction was created, including title, DOI, first author, country of first author, publication year, study design, follow-up duration, source of radiomics, number of participants involved in image segmentation, image segmentation software, case number of NSCLC in the training set, case number of NSCLC in the validation set, method of generating the validation set, presence or absence of overfitting, total number of participants in the validation set, method of variable selection, method of model construction, presence or absence of a radiomics score, presence or absence of other features combined for model construction (such as clinical features and genetic features), presence or absence of assessment of clinical applicability, presence or absence of evaluation of model calibration, presence or absence of overfitting assessment of the model, c-index, sensitivity, specificity, and diagnostic contingency table.

Literature screening and data extraction were implemented independently by two investigators (J Wu and SH), followed by a cross-verification to ensure consistency. All discrepancies were resolved by a third researcher (YZ) to reach a consensus.

Evaluation for Study Quality

Two investigators used the Radiomics Quality Score (RQS) to assess the methodological quality and analyze risk of bias among the included studies, followed by a cross-verification. Any discrepancy was resolved by a third researcher to make a final decision. The RQS covered 16 items and described the entire process of radiomics, with a maximum score of 36 points.

Synthesis Methods

Meta-analysis was performed for the c-index, an index for appraising the overall precision of ML models. In some original studies, when a 95% CI and a SE of the c-index were not reported, the SEs were estimated based on the study by Debray et al [19]. Among the ML models, discrepancies were identified in the selected variables, with parameters varying. Therefore, in terms of meta-analysis of the c-index, this study privileged the random effects model (REM). Furthermore, the results were calibrated using the Hartung-Knapp-Sidik-Jonkman approach, and the prediction intervals were presented. Subsequently, the bivariate mixed effects model was additionally used to conduct a meta-analysis for sensitivity and specificity. The meta-analysis was carried out via Stata 15.0 (StataCorp LLC).

Results

Literature Selection

A total of 5423 reports were retrieved, among which 1007 duplicate studies were removed, with 4416 studies remaining. A total of 4383 studies were screened out according to the titles and abstracts, due to irrelevance to the study topic or an unmatched study design. The full texts of the remaining 33 studies that were preliminarily in line with the inclusion criteria were obtained. Through the review of the full texts, 2 unpublished conference abstracts and 1 study without outcome

XSL•FO
RenderX

indicators were excluded. Ultimately, 30 studies were included. The process of literature selection is depicted in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart for literature selection.

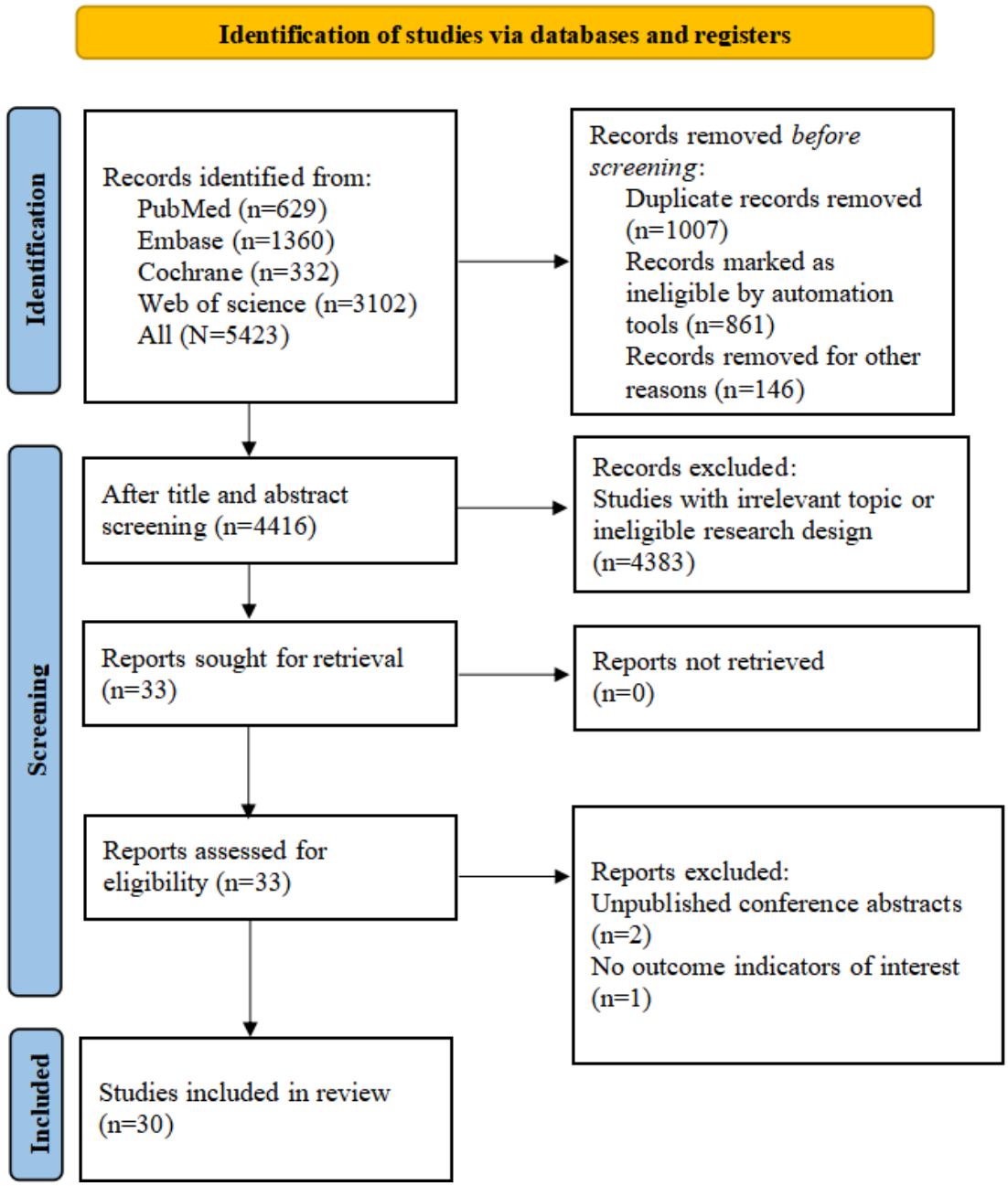

Features of the Included Studies

Those 30 selected studies were all published between 2019 and 2025, in 9 countries. The study types were all cohort studies. The study data were from multiple centers in 11 studies, from a single center in 17 studies, and from a registration database in 2 studies. A total of 8 studies clearly delineated the pathological types of lung adenocarcinoma and/or lung squamous cell carcinoma, while 22 did not delineate the pathological type of the tumor, merely describing it as NSCLC. A total of 19 studies included early-stage lung cancer (based on tumor node metastasis [TNM] staging in stages I-II), 9 included locally advanced or advanced lung cancer (based on TNM staging in stages III-IV), and 2 did not specify the tumor stage. Among the included studies, 16 discussed tumor recurrence after the stereotactic body radiation therapy (SBRT), and 22 discussed tumor recurrence after surgery and/or radiotherapy and chemotherapy. The follow-up time was mainly concentrated between 2 and 70 months. The image sources of the 26 studies mainly included computed tomography (CT) and positron emission tomography–computed tomography (PET-CT). Among them, 24 studies used CT, 2 studies used PET-CT, and 4 studies used both CT and PET-CT. The generation method of the validation set mainly involved random sampling and cross-validation. Among them, 6 studies included external validation (Table S2 in Multimedia Appendix 1).

Evaluation of the Quality of the Selected Studies

None of the included studies described the use of multiple segmentation approaches, repeated measurements at different

XSL•FO
RenderX

time points, or pretests with varying image parameters. None of them conducted prospective registration, provided a description of the degree of consistency with the "gold standard" method, or conducted cost-effectiveness analysis. Therefore, none of the included studies received any points in these categories. In 7 studies, only 1 doctor was involved in the image segmentation, and thus, they did not receive any points. Only the study by Aonpong et al [20] discussed the correlation between radiomics and tumor biology. A total of 16 studies did not report the radiomics score and therefore received a score of 0 in this category. All the included literature reported the discriminant statistics and their statistical significance, but only 15 studies applied the multiple sampling method. Accordingly, in this domain, 15 studies were awarded 2 points, while the remaining studies received 1 point. A total of 5 studies reported calibration statistics and were awarded 2 points, whereas the remaining studies received no points. Among the 30 included studies, all verified the radiomics features, with 1 study using 3 external datasets for verification, another study using 2 external datasets, and 7 studies using 1 external dataset for verification. A total of 7 studies performed decision curve analysis to evaluate the potential clinical utility of the models, each receiving 2 points. A total of 24 studies disclosed images and segmentation, earning 2 points each. One study disclosed data codes, receiving 3 points. A total of 5 studies did not disclose codes, data, or segmentation methods and therefore received no points. The final average score was 27.5% (Table S3 in Multimedia Appendix 1).

Meta-Analysis: Radiomics-Based ML

In the training set, the c-index was pooled by REM (τ^2=0.0116; τ=0.1035; I^2=88.3%), and the c-index of the radiomics-based ML for predicting the risk of recurrence in NSCLC was 0.850 (95% CI 0.834 - 0.866, 95% prediction interval [PI] 0.623 - 1.000; Figure 2).

Figure 2. Forest plot for meta-analysis of the c-index of radiomics-based machine learning (ML) for predicting risk of recurrence in non–small cell lung cancer (NSCLC) in the training set. The forest plot was analyzed using a random effects model and corrected by the Knapp-Hartung method [20-35]. DBP: diastolic blood pressure; OR: odds ratio; SBP: systolic blood pressure.

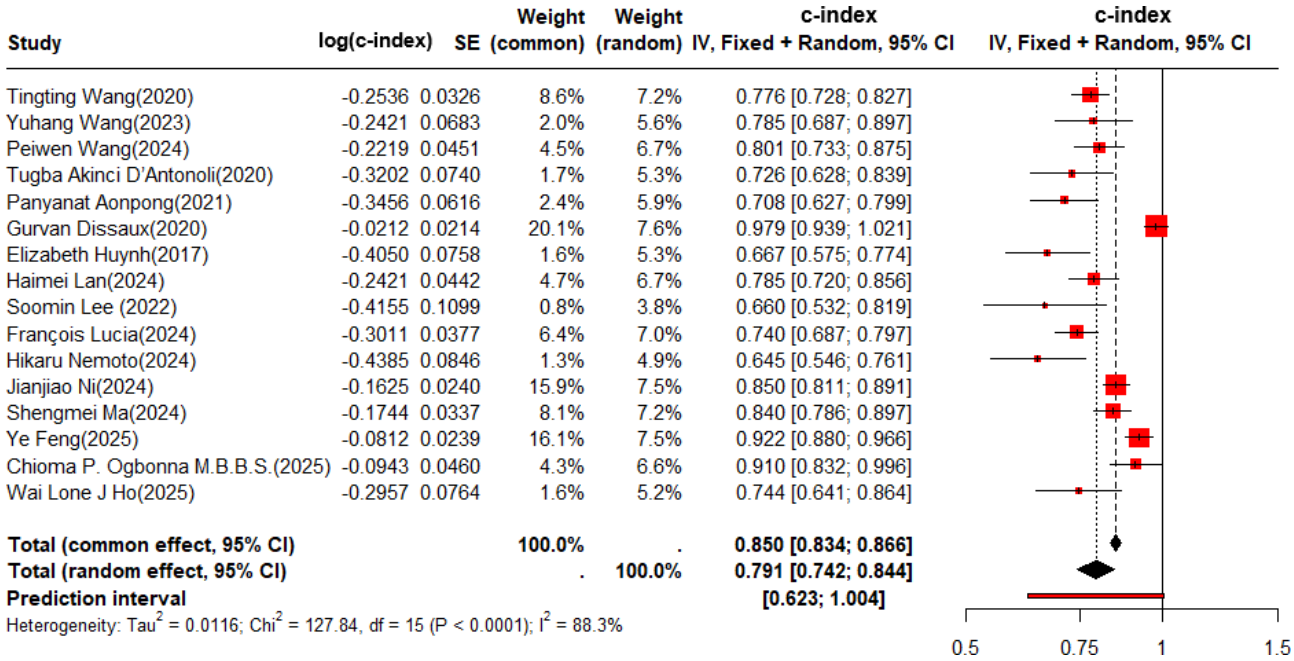

The pooled c-index was 0.876 (95% CI 0.853 - 0.900) among the patients receiving SBRT, and 0.825 (95% CI 0.804 - 0.848) among those who received surgeries combined with other adjuvant treatment regimens (Figure S1 in Multimedia Appendix 1). The pooled c-index was 0.815 (95% CI 0.776 - 0.856) among the patients with adenocarcinoma and squamous cell carcinoma as the pathological type, and 0.856 (95% CI 0.839 - 0.874) among other patients with unspecified specific pathological types (Figure S2 in Multimedia Appendix 1). The pooled c-index was 0.851 (95% CI 0.832 - 0.870) among the patients with early-stage NSCLC, and 0.848 (95% CI 0.818 - 0.879) among the patients at a locally advanced stage and an advanced stage (Figure S3 in Multimedia Appendix 1).

In the validation set, the c-index was pooled by REM (τ^2=0.0072; τ=0.0750; I^2=69.3%), and the c-index of the radiomics-based ML for predicting the risk of recurrence in NSCLC was 0.878 (95% CI: 0.854 - 0.902, 95% PI 0.681 - 1.000; Figure 3).

XSL•FO
RenderX

Figure 3. Forest plot for meta-analysis of the c-index of radiomics-based machine learning for predicting risk of recurrence in non–small cell lung cancer in the validation set. The forest plot was analyzed using a random effects model and corrected by the Knapp-Hartung method [21-30,36]. IV: independent variable.

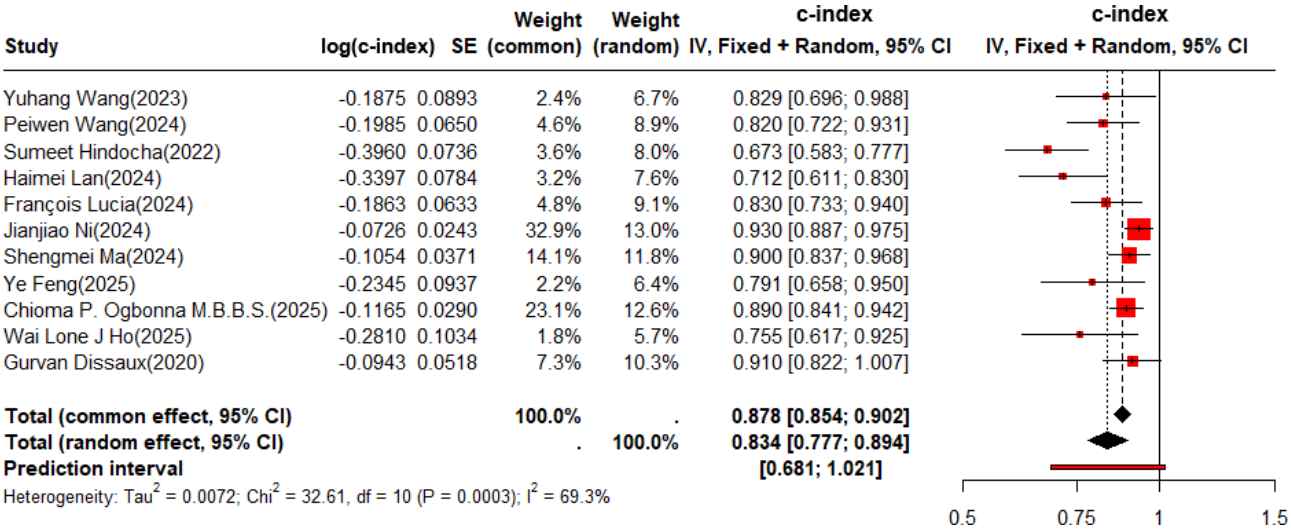

The pooled c-index was 0.907 (95% CI 0.877 - 0.937) among the patients receiving SBRT, and 0.819 (95% CI 0.780 - 0.859) among those who received surgeries combined with other adjuvant treatment regimens (Figure S4 in Multimedia Appendix 1). The pooled c-index was 0.858 (95% CI 0.807 - 0.913) among the patients with adenocarcinoma-squamous cell carcinoma as the pathological type, and 0.883 (95% CI 0.856 - 0.910) among other patients with unspecified specific pathological types (Figure S5 in Multimedia Appendix 1). Among the patients with early-stage NSCLC, the pooled c-index was 0.896 (95% CI 0.870 - 0.922), and among the patients at a locally advanced stage and an advanced stage, the pooled c-index was 0.759 (95% CI 0.702 - 0.821; Figure S6 in Multimedia Appendix 1). In the subgroup with fewer than 100 cases, the pooled c-index was 0.851 (95% CI 0.815 - 0.888), while in the subgroup with more than 100 cases, it was 0.897 (95% CI 0.865 - 0.929; Figure S7 in Multimedia Appendix 1).

ML Based on Radiomics Combined With Clinical Features

In the training set, the c-index was pooled by REM (τ^2=0.0042; τ=0.0628; I^2=74.9%), and the c-index of the ML based on radiomics combined with clinical features for predicting the risk of recurrence in NSCLC was 0.833 (95% CI 0.822 - 0.854, 95% PI 0.717 - 0.945; Figure 4).

XSL•FO
RenderX

Figure 4. Forest plot for meta-analysis of the c-index of radiomics-based machine learning combined with clinical features for predicting risk of recurrence in non–small cell lung cancer in the training set. The forest plot was analyzed using a random effects model and corrected by the Knapp-Hartung method [20-22,24-27,30-32,37-48]. IV: independent variable.

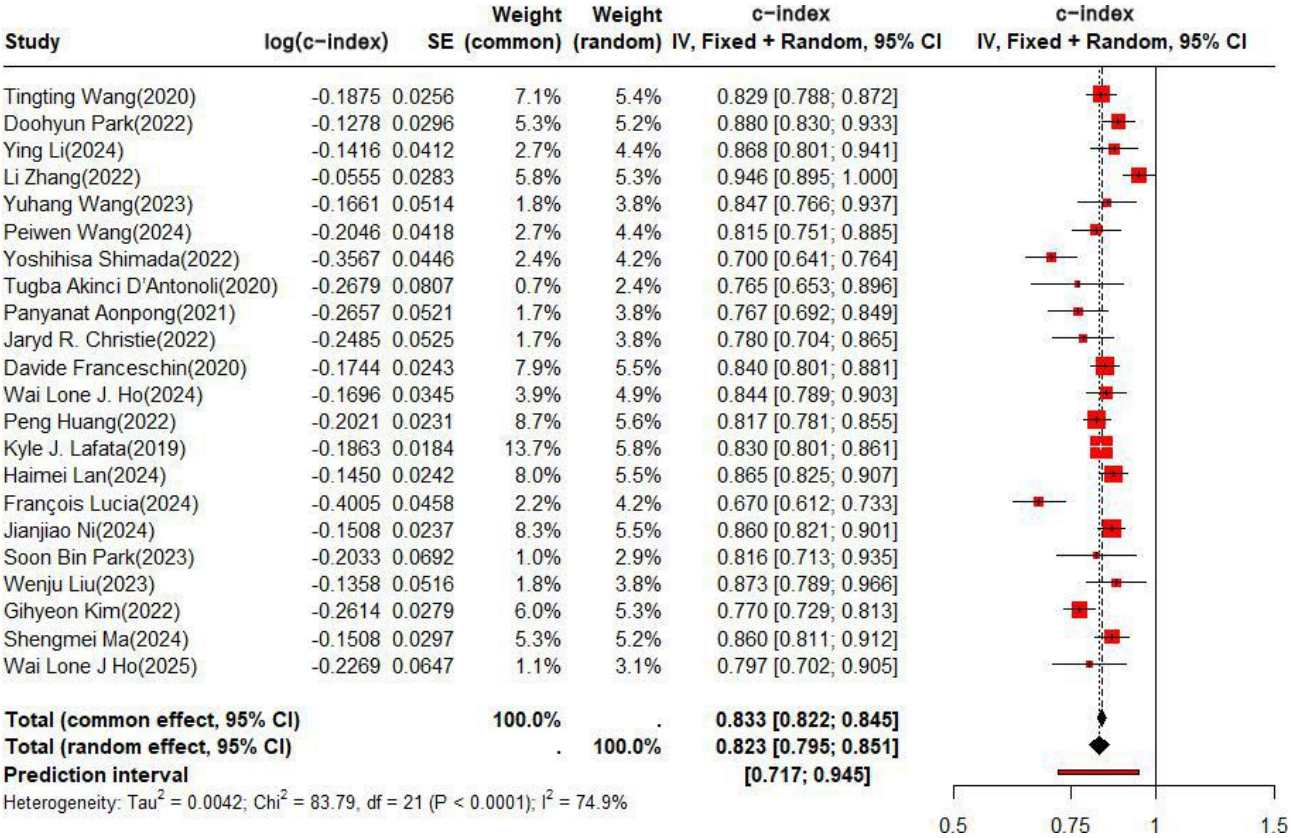

The pooled c-index was 0.828 (95% CI 0.808 - 0.848) among the patients receiving SBRT, and 0.836 (95% CI 0.823 - 0.850) among the patients who received surgeries combined with other adjuvant treatment regimens (Figure S8 in Multimedia Appendix 1). The pooled c-index was 0.882 (95% CI 0.858 - 0.907) among the patients with adenocarcinoma-squamous cell carcinoma, and 0.823 (95% CI 0.810 - 0.837) among other patients with unspecified specific pathological types (Figure S9 in Multimedia Appendix 1). Among the NSCLC patients, the pooled c-index was 0.842 (95% CI 0.829 - 0.855), and among the patients at a locally advanced stage and an advanced stage, the pooled c-index was 0.802 (95% CI 0.779 - 0.826; Figure S10 in Multimedia Appendix 1).

In the validation set, the c-index was pooled by REM (τ^2=0.0079; τ=0.1069; I^2=65.1%), and the c-index of the ML based on radiomics combined with clinical features for predicting the risk of recurrence in NSCLC was 0.854 (95% CI 0.830 - 0.878, 95% PI 0.655 - 0.992; Figure 5).

XSL•FO
RenderX

Figure 5. Forest plot for meta-analysis of the c-index of radiomics-based machine learning combined with clinical features for predicting risk of recurrence in non–small cell lung cancer in the validation set. The forest plot was analyzed using a random effects model and corrected by the Knapp-Hartung method [21,22,24-27,30,36,38,39,41,44,46]. IV: independent variable.

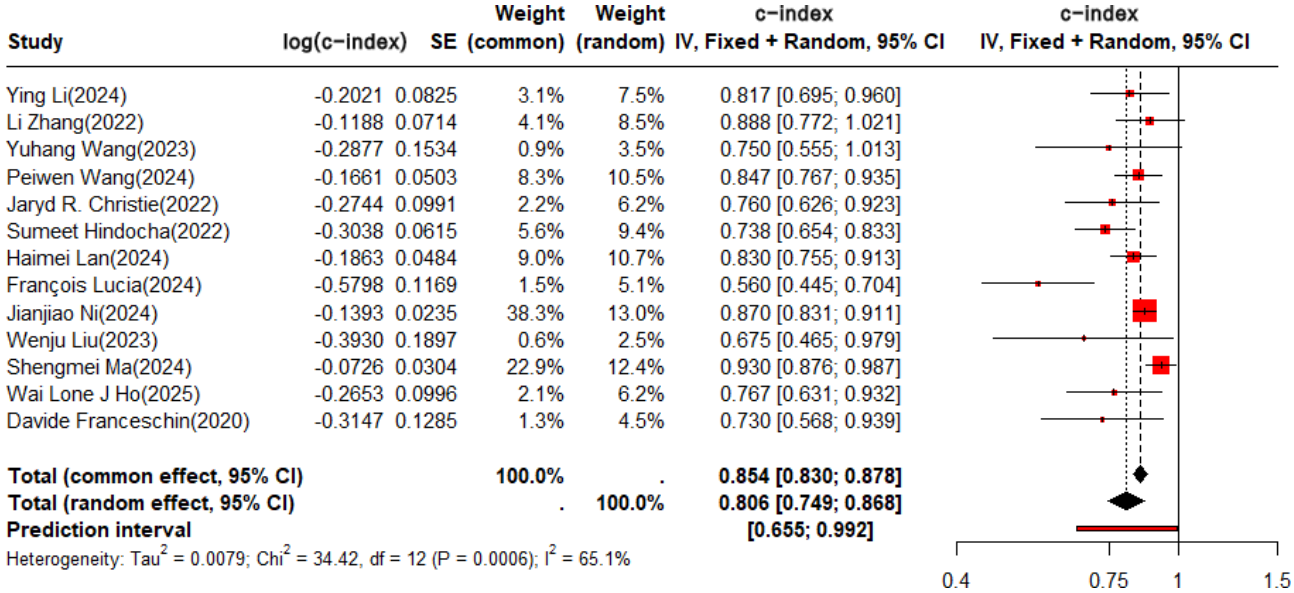

The pooled c-index was 0.851 (95% CI 0.814 - 0.890) among the patients receiving SBRT, and 0.855 (95% CI 0.8824 - 0.888) among the patients who received surgeries combined with other adjuvant treatment regimens (Figure S11 in Multimedia Appendix 1). The pooled c-index was 0.889 (95% CI 0.850 - 0.930) among the patients with adenocarcinoma-squamous cell carcinoma, and 0.831 (95% CI 0.801 - 0.862) among other patients with unspecified specific pathological types (Figure S12 in Multimedia Appendix 1). Among the patients with early-stage NSCLC, the pooled c-index was 0.870 (95% CI 0.842 - 0.897), and among the patients at a locally advanced stage and an advanced stage, the pooled c-index was 0.788 (95% CI 0.738 - 0.842; Figure S13 in Multimedia Appendix 1). The pooled c-index in the subgroup with fewer than 100 cases was 0.855 (95% CI 0.823 - 0.888), while in the subgroup with more than 100 cases, it was 0.852 (95% CI 0.816 - 0.889; Figure S14 in Multimedia Appendix 1).

Discussion

Principal Findings

That meta-analysis of 30 cohort studies demonstrated that the radiomics-based ML achieved pooled c-indexes of 0.850 in the training set and 0.878 in the validation set for predicting recurrence in NSCLC. These findings indicate that radiomics, as a noninvasive and reproducible imaging analysis technique, holds clinical potential for predicting the risk of recurrence in NSCLC. Most studies, for the purpose of improving the prediction accuracy and strengthening the interpretability of the radiomics-based ML, tend to combine clinical features with radiomics features. However, this study revealed that although these attempts indeed enhanced the c-index of the ML model, they failed to show remarkable advantages. It is plausible that radiomics features may inherently capture certain clinically relevant information associated with recurrence, while the clinical features added in the above studies vary greatly and fail to fully reflect the biological characteristics of patients. Therefore, the additional clinical features of patients lead to a limited improvement in the predictive efficiency of the model, which, to some extent, reflects that the radiomics-based model may be more stable.

To thoroughly ascertain the predictive efficiency of the radiomics-based model, subgroup analyses were conducted from multiple dimensions such as stage, treatment modality, and pathological classification. Regarding disease stage, the c-index of the model in patients with NSCLC at an early stage was 0.96, which was higher than that in patients with locally advanced NSCLC (c-index=0.759). The tumors in early-stage NSCLC are usually well-defined in terms of boundaries and relatively intact in structure, and the imaging features of the tumors are more easily recognized and quantified by the model, thereby improving the predictive stability and discriminative capacity of the models. In contrast, the tumors in patients with locally advanced NSCLC often invade adjacent areas such as the mediastinum, pericardium, and chest wall, or present lymph node metastases such as at the pulmonary hilum and mediastinum, and some patients may also suffer from complications such as obstructive pneumonia. These factors contribute to more complex imaging features and increased heterogeneity, which may interfere with the extraction and learning of the model for key features and thereby affect the prediction accuracy [49,50]. Additionally, patients with locally advanced NSCLC often undergo treatment interventions, including neoadjuvant chemotherapy and radiotherapy, which may further change the expression patterns of imaging features [51].

The subgroup analyses, particularly those with wide CIs, should be interpreted with caution. This study included 7 studies using SBRT and 18 studies utilizing surgical treatment. Although the c-index of the 2 groups was similar, the CI of the SBRT group was significantly wider, suggesting that the stability of the prediction results in this group was poorer. These findings are

XSL•FO
RenderX

exploratory and may reflect variability due to small sample sizes or heterogeneity among included studies. Therefore, the subgroup results should be considered hypothesis-generating rather than conclusive. Future research with larger datasets and consistent analytical frameworks is warranted to confirm these preliminary trends. Moreover, SBRT serves as a relatively new and highly technology-dependent treatment modality, and there are significant differences in implementation plans, dose settings, and patient selection criteria among different studies, which may also lead to increased uncertainty in the performance evaluation of the model [52]. These results imply that the stability of current radiomics-based models in patients undergoing SBRT or other nonsurgical treatments still requires further verification. In the future, more rigorous prospective investigations with larger sample sizes are needed to assess the applicability and reliability of the radiomics-based model in patients with NSCLC receiving SBRT treatment, thereby facilitating its integration into personalized treatment plans.

In terms of pathological classification, a total of 22 studies in this study only mentioned NSCLC without further subdividing pathological subtypes (squamous cell carcinoma and adenocarcinoma). The findings revealed that the radiomics-based prediction model exhibited superior predictive accuracy in studies classifying adenocarcinoma and squamous cell carcinoma than in those without specific subtype classification. Adenocarcinoma and squamous cell carcinoma are the 2 most prevalent pathological subtypes; furthermore, due to their histological heterogeneity, they may be more identifiable in imaging features, thereby enabling the model to extract more discriminative imaging features and improving the prediction accuracy [53,54]. In contrast, in studies that did not explicitly include information on pathological subtypes, the predictive performance of their models may be limited due to high sample heterogeneity and inconsistent imaging features. In training and validating subsequent radiomics-based models, explicit histological classifications should be included as much as possible to strengthen the stability and the practical applicability of models. Moreover, given the distinct biological characteristics, therapeutic responses, and prognoses across subtypes, combining histopathological information for multimodal fusion modeling may enhance the efficacy of the radiomics-based prediction models. In summary, while the subgroup analysis is supported by the meta-analysis and the sensitivity analysis, the findings should be considered preliminary and hypothesis-generating. In the future, additional prospective studies with a large sample size and based on multicenter data are warranted to validate these findings.

Radiomics has achieved considerable progress in predicting the risk of recurrence in NSCLC. However, due to the inherent complexity of radiomics and the limited reproducibility of the numerous processes involved, it still faces many challenges in clinical applications. Particularly, in image segmentation, feature extraction, and dataset analysis, the application of radiomics is still limited to academic research [55-57]. To begin with, a primary technical challenge in current radiomics research is the absence of standardized imaging protocols [58]. Different medical institutions use scanning equipment and image acquisition parameters that vary greatly. Even for the same patient, the imaging quality and image features obtained at different times or with different devices may vary, thereby affecting the stability of feature extraction and the generalization ability of models [55,59-62]. Standardizing the processing of image data across devices and centers may be the key to improving the clinical applicability of models. Currently, some studies attempt to reduce the influence of image differences through image normalization, resampling, and image domain adaptation [59], but more confirmatory research is needed. In the future, pilot studies with small samples may be conducted prior to formal model construction to evaluate the actual impact of image protocol differences on model construction. Given the subjectivity and consistency issue of the image segmentation process, the stability of models is affected, and manual segmentation is influenced by operator expertise, leading to variability [58,63,64]. Therefore, more studies attempt to apply deep learning methods for automatic or semiautomatic segmentation and to promote the application of automatic segmentation algorithms in multidisease and multimodal data to improve efficiency and reproducibility [65-67].

Moreover, radiomics usually involves high-dimensional small-sample data, so it is frequently associated with overfitting, whereby the radiomics-based model performs well on training data but underperforms on new data. This issue may be alleviated through techniques such as feature selection and data augmentation. In the previous study by Ye et al [68], the robustness of the model was improved through data augmentation and feature selection, and kernelization methods were used, and principal component analysis was performed to lower the feature dimension to effectively reduce the risk of overfitting. In summary, as an emerging tool for imaging analysis, radiomics holds potential for clinical practice. Nonetheless, before radiomics is applied in clinical practice, the above-mentioned problems should be addressed.

It should be noted that the overall methodological quality of the included studies, as reflected by the relatively low mean RQS (27.4%), indicates substantial limitations in current radiomics-based ML research. Many of the included studies were absent from external validation, prospective design, or standardized feature extraction procedures, which may introduce bias and restrict the generalizability of the pooled diagnostic accuracy. Therefore, the results of this meta-analysis should be interpreted with caution. Future studies are encouraged to adhere to standardized radiomics workflows and reporting guidelines, and to perform multicenter validations to enhance methodological rigor and clinical reliability. Furthermore, the overall RQS of the included studies was relatively low, primarily manifested in the following aspects: biological relevance, cost-benefit analysis, comparison with the gold standard, and prospective design. Biological relevance involves the link of imaging features with genetic or molecular features; however, only 1 included study reported the potential correlation of gene levels and imaging features. None of the original studies included were prospectively registered, and some were absent from external validation or solely relied on internal validation based on single-center data. These deficiencies collectively led to the low RQS.

XSL•FO
RenderX

Strengths and Limitations

Although this meta-analysis demonstrates the potential of radiomics-based ML for forecasting the risk of recurrence in NSCLC, certain limitations still need to be resolved to strengthen the reliability and clinical applicability of radiomics-based models for the prediction of the risk of recurrence in NSCLC. First, this study primarily utilizes retrospective data, potentially introducing selection bias and other issues, thereby limiting the generalizability of the research results. Second, although subgroup analyses are conducted, it is still difficult to completely eliminate heterogeneity. Finally, some studies are limited to the sample sizes, lack external validation, and lack long-term follow-up analysis of actual clinical prognosis, which restricts the assessment of the practical applicability of radiomics-based models.

Conclusion

This systematic review, for the first time, confirms that radiomics-based ML models effectively predict the risk of recurrence in NSCLC, and clinically relevant subgroups (eg, model type, treatment background, stage, and pathology) were explored. Differing from individual original studies or extensive reviews focusing on radiomics with mixed heterogeneous end points, this study concentrates on recurrence, a clinically operable outcome indicator, thereby clarifying areas where model performance is most pronounced and methodological gaps are identified. In clinical practice, robust radiomics-based ML tools may be leveraged to support individualized monitoring intensity, aid in formulating treatment regimens, and optimize risk stratification in clinical trials. Nonetheless, the current application of radiomics methodologies remains insufficient. Therefore, in the future, studies should focus on standardizing the implementation procedure of radiomics-based ML and integrating multicenter imaging data to enhance generalizability.

Funding

This work was supported by the Key Construction Disciplines of Provincial and Municipal Co construction of Zhejiang (2023-SSGJ-002); the National Oncology Clinical Key Speciality (2023-GJZK-001); the Jiaxing Key Laboratory of Precision Treatment for Lung Cancer, the Early Diagnosis and Comprehensive Treatment of Lung Cancer Innovation Team Building Project, and the Scientific Technology Plan Program for Healthcare in Zhejiang Province (2023KY329, 2023KY1196, and 24 2023RC098); the Science and Technology Project of Jiaxing (2025CGZ041); and Zhejiang Province Postdoctoral Research Project (ZJ2024156).

Data Availability

The datasets used and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: J Wu
Formal analysis and investigation: WC, SH, J Wang, CC
Funding acquisition: YZ
Investigation: WC, SH, J Wu, CC
Methodology: J Wu
Resources: J Wang
Supervision: WC
Writing – original draft: J Wu
Writing – review and editing: SH
All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.
SH (hushiyu0317@163.com) and WC (00135116@zjxu.edu.cn) are co-corresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplemental tables and figures: the detailed literature search strategies (Table S1), characteristics of the included studies (Table S2), Radiomics Quality Score assessments (Table S3), and forest plots depicting subgroup analyses for the meta-analysis of machine learning models predicting recurrence in non–small cell lung cancer (Figures S1-S14).
[7Z File, 556 KB - jmir_v28i1e84223_app1.7z]

Checklist 1
PRISMA checklist.
[DOCX File, 272 KB - jmir_v28i1e84223_app2.docx]

XSL•FO
RenderX

Checklist 2
PRISMA-S checklist.
[DOCX File, 22 KB - jmir_v28i1e84223_app3.docx]

References

1. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin 2023 Jan;73(1):17-48. [doi: 10.3322/caac.21763] [Medline: 36633525]
2. Leiter A, Veluswamy RR, Wisnivesky JP. The global burden of lung cancer: current status and future trends. Nat Rev Clin Oncol 2023 Sep;20(9):624-639. [doi: 10.1038/s41571-023-00798-3] [Medline: 37479810]
3. Liu L, Soler J, Reckamp KL, Sankar K. Emerging targets in non–small cell lung cancer. Int J Mol Sci 2024 Sep 18;25(18):10046. [doi: 10.3390/ijms251810046] [Medline: 39337530]
4. Mithoowani H, Febbraro M. Non-small-cell lung cancer in 2022: a review for general practitioners in oncology. Curr Oncol 2022 Mar 9;29(3):1828-1839. [doi: 10.3390/curroncol29030150] [Medline: 35323350]
5. Araghi M, Mannani R, Heidarnejad Maleki A, et al. Recent advances in non–small cell lung cancer targeted therapy; an update review. Cancer Cell Int 2023 Aug 11;23(1):162. [doi: 10.1186/s12935-023-02990-y] [Medline: 37568193]
6. Lin X, Xie M, Yao J, et al. Immune-related adverse events in non-small cell lung cancer: occurrence, mechanisms and therapeutic strategies. Clin Transl Med 2024 Mar;14(3):e1613. [doi: 10.1002/ctm2.1613] [Medline: 38451000]
7. Jairam V, Lee V, Park HS, et al. Treatment-related complications of systemic therapy and radiotherapy. JAMA Oncol 2019 Jul 1;5(7):1028-1035. [doi: 10.1001/jamaoncol.2019.0086] [Medline: 30946433]
8. Monsalve AF, Hoag JR, Resio BJ, et al. Variable impact of prior cancer history on the survival of lung cancer patients. Lung Cancer 2019 Jan;127:130-137. [doi: 10.1016/j.lungcan.2018.11.040] [Medline: 30642541]
9. Riely GJ, Wood DE, Ettinger DS, et al. Non–small cell lung cancer, version 4.2024, NCCN clinical practice guidelines in oncology. J Natl Compr Canc Netw 2024 May;22(4):249-274. [doi: 10.6004/jnccn.2204.0023] [Medline: 38754467]
10. Mohamed SK, Walsh B, Timilsina M, et al. On predicting recurrence in early stage non–small cell lung cancer. AMIA Annu Symp Proc 2021;2021:853-862. [Medline: 35308971]
11. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012 Mar;48(4):441-446. [doi: 10.1016/j.ejca.2011.11.036] [Medline: 22257792]
12. Chen J, Chen A, Yang S, Liu J, Xie C, Jiang H. Accuracy of machine learning in preoperative identification of genetic mutation status in lung cancer: a systematic review and meta-analysis. Radiother Oncol 2024 Jul;196:110325. [doi: 10.1016/j.radonc.2024.110325] [Medline: 38734145]
13. Yan HJ, Zhao JS, Zuo HD, et al. Dual-region computed tomography radiomics-based machine learning predicts subcarinal lymph node metastasis in patients with non–small cell lung cancer. Ann Surg Oncol 2024 Aug;31(8):5011-5020. [doi: 10.1245/s10434-024-15197-w] [Medline: 38520581]
14. Tutsoy O, Sumbul HE. A novel deep machine learning algorithm with dimensionality and size reduction approaches for feature elimination: thyroid cancer diagnoses with randomly missing data. Brief Bioinform 2024 May 23;25(4):bbae344. [doi: 10.1093/bib/bbae344] [Medline: 39007597]
15. Xia F, Wei W, Wang J, Duan Y, Wang K, Zhang C. Machine learning model for non-alcoholic steatohepatitis diagnosis based on ultrasound radiomics. BMC Med Imaging 2024 Aug 20;24(1):221. [doi: 10.1186/s12880-024-01398-y] [Medline: 39164667]
16. Libling WA, Korn R, Weiss GJ. Review of the use of radiomics to assess the risk of recurrence in early-stage non–small cell lung cancer. Transl Lung Cancer Res 2023 Jul 31;12(7):1575-1589. [doi: 10.21037/tlcr-23-5] [Medline: 37577298]
17. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
18. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
19. Debray TP, Damen JA, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. Stat Methods Med Res 2019 Sep;28(9):2768-2786. [doi: 10.1177/0962280218785504] [Medline: 30032705]
20. Aonpong P, Iwamoto Y, Han XH, Lin L, Chen YW. Improved genotype-guided deep radiomics signatures for recurrence prediction of non–small cell lung cancer. Annu Int Conf IEEE Eng Med Biol Soc 2021 Nov;2021:3561-3564. [doi: 10.1109/EMBC46164.2021.9630703] [Medline: 34892008]
21. Wang P, Luo Z, Luo C, Wang T. Application of a comprehensive model based on CT radiomics and clinical features for postoperative recurrence risk prediction in non-small cell lung cancer. Acad Radiol 2024 Jun;31(6):2579-2590. [doi: 10.1016/j.acra.2023.11.028]
22. Wang Y, Ding Y, Liu X, et al. Preoperative CT-based radiomics combined with tumour spread through air spaces can accurately predict early recurrence of stage I lung adenocarcinoma: a multicentre retrospective cohort study. Cancer Imaging 2023 Sep 7;23(1):83. [doi: 10.1186/s40644-023-00605-3] [Medline: 37679806]

XSL•FO
RenderX

23. Dissaux G, Visvikis D, Da-Ano R, et al. Pretreatment 18F-FDG PET/CT radiomics predict local recurrence in patients treated with stereotactic body radiotherapy for early-stage non-small cell lung cancer: a multicentric study. J Nucl Med 2020 Jun;61(6):814-820. [doi: 10.2967/jnumed.119.228106] [Medline: 31732678]
24. Lan H, Wei C, Xu F, et al. 2.5D peritumoural radiomics predicts postoperative recurrence in stage I lung adenocarcinoma. Front Oncol 2024;14:1382815. [doi: 10.3389/fonc.2024.1382815] [Medline: 39267836]
25. Lucia F, Louis T, Cousin F, et al. Multicentric development and evaluation of [18F]FDG PET/CT and CT radiomic models to predict regional and/or distant recurrence in early-stage non-small cell lung cancer treated by stereotactic body radiation therapy. Eur J Nucl Med Mol Imaging 2024 Mar;51(4):1097-1108. [doi: 10.1007/s00259-023-06510-y] [Medline: 37987783]
26. Ni J, Chen H, Yu L, et al. Predicting regional recurrence and prognosis in stereotactic body radiation therapy-treated clinical stage I non-small cell lung cancer using a radiomics model constructed with surgical data. Int J Radiat Oncol Biol Phys 2024 Nov 15;120(4):1096-1106. [doi: 10.1016/j.ijrobp.2024.06.010] [Medline: 38936632]
27. Ma S, Li J, Chen Y, et al. Machine learning based on clinical information and integrated CT radiomics to predict local recurrence of stage IA lung adenocarcinoma after microwave ablation. J Vasc Interv Radiol 2024 Dec;35(12):1823-1832. [doi: 10.1016/j.jvir.2024.08.018] [Medline: 39208929]
28. Feng Y, Xu Y, Wang J, et al. Prediction of early recurrence after bronchial arterial chemoembolization in non-small cell lung cancer patients using dual-energy CT: an interpretable model based on SHAP methodology. Acad Radiol 2025 Oct;32(10):6320-6329. [doi: 10.1016/j.acra.2025.07.039] [Medline: 40784844]
29. Ogbonna CP, Breen WG, Le Noach P, et al. Radiomics-based prediction of local recurrence after stereotactic body radiation therapy for early-stage non-small cell lung cancer. Ann Am Thorac Soc 2025 Aug;22(8):1236-1243. [doi: 10.1513/AnnalsATS.202410-1047OC] [Medline: 40208641]
30. Ho WLJ, Fetisov N, Hall LO, Goldgof D, Schabath MB. Utilizing clinicopathological and radiomic features for risk stratification of lung cancer recurrence. Acad Radiol 2025 Sep;32(9):5557-5566. [doi: 10.1016/j.acra.2025.04.062] [Medline: 40379589]
31. Wang T, Deng J, She Y, et al. Radiomics signature predicts the recurrence-free survival in stage I non-small cell lung cancer. Ann Thorac Surg 2020 Jun;109(6):1741-1749. [doi: 10.1016/j.athoracsur.2020.01.010] [Medline: 32087134]
32. Akinci D'Antonoli T, Farchione A, Lenkowicz J, et al. CT radiomics signature of tumor and peritumoral lung parenchyma to predict nonsmall cell lung cancer postsurgical recurrence risk. Acad Radiol 2020 Apr;27(4):497-507. [doi: 10.1016/j.acra.2019.05.019] [Medline: 31285150]
33. Nemoto H, Saito M, Satoh Y, et al. Evaluation of the performance of both machine learning models using PET and CT radiomics for predicting recurrence following lung stereotactic body radiation therapy: a single-institutional study. J Appl Clin Med Phys 2024 Jul;25(7):e14322. [doi: 10.1002/acm2.14322] [Medline: 38436611]
34. Lee S, Jung J, Hong H, Kim BS. Prediction of two-year recurrence-free survival in operable NSCLC patients using radiomic features from intra- and size-variant peri-tumoral regions on chest CT images. Diagnostics (Basel) 2022 May 25;12(6):1313. [doi: 10.3390/diagnostics12061313] [Medline: 35741123]
35. Huynh E, Coroller TP, Narayan V, et al. Associations of radiomic data extracted from static and respiratory-gated CT scans with disease recurrence in lung cancer patients treated with SBRT. PLoS ONE 2017;12(1):e0169172. [doi: 10.1371/journal.pone.0169172] [Medline: 28046060]
36. Hindocha S, Charlton TG, Linton-Reid K, et al. Gross tumour volume radiomics for prognostication of recurrence & death following radical radiotherapy for NSCLC. NPJ Precis Oncol 2022 Oct 27;6(1):77. [doi: 10.1038/s41698-022-00322-3] [Medline: 36302938]
37. Park D, Oh D, Lee M, et al. Importance of CT image normalization in radiomics analysis: prediction of 3-year recurrence-free survival in non-small cell lung cancer. Eur Radiol 2022 Dec;32(12):8716-8725. [doi: 10.1007/s00330-022-08869-2] [Medline: 35639142]
38. Li Y, Zhao J, Li R, et al. Predicting prognosis in patients with stage IA lung adenocarcinoma with a micropapillary component using a nomogram based on computed tomography radiomics and clinicopathologic factors: a retrospective analysis. Transl Lung Cancer Res 2024 Oct 31;13(10):2585-2602. [doi: 10.21037/tlcr-24-544] [Medline: 39507031]
39. Zhang L, Lv L, Li L, et al. Radiomics signature to predict prognosis in early-stage lung adenocarcinoma (≤3 cm) patients with no lymph node metastasis. Diagnostics (Basel) 2022 Aug 6;12(8):1907. [doi: 10.3390/diagnostics12081907] [Medline: 36010257]
40. Shimada Y, Kudo Y, Maehara S, et al. Radiomics with artificial intelligence for the prediction of early recurrence in patients with clinical stage IA lung cancer. Ann Surg Oncol 2022 Dec;29(13):8185-8193. [doi: 10.1245/s10434-022-12516-x] [Medline: 36070112]
41. Christie JR, Daher O, Abdelrazek M, et al. Predicting recurrence risks in lung cancer patients using multimodal radiomics and random survival forests. J Med Imaging (Bellingham) 2022 Nov;9(6):066001. [doi: 10.1117/1.JMI.9.6.066001] [Medline: 36388142]
42. Lafata KJ, Hong JC, Geng R, et al. Association of pre-treatment radiomic features with lung cancer recurrence following stereotactic body radiation therapy. Phys Med Biol 2019 Jan 8;64(2):025007. [doi: 10.1088/1361-6560/aaf5a5] [Medline: 30524018]

XSL•FO
RenderX

43. Park SB, Kim KU, Park YW, Hwang JH, Lim CH. Application of 18 F-fluorodeoxyglucose PET/CT radiomic features and machine learning to predict early recurrence of non-small cell lung cancer after curative-intent therapy. Nucl Med Commun 2023 Feb 1;44(2):161-168. [doi: 10.1097/MNM.0000000000001646] [Medline: 36458424]
44. Liu W, Qiao X, Ge H, et al. Recurrence patterns are significantly associated with the 18F-FDG PET/CT radiomic features of patients with locally advanced non-small cell lung cancer treated with chemoradiotherapy. Oncol Lett 2023 Jul;26(1):317. [doi: 10.3892/ol.2023.13903] [Medline: 37332327]
45. Kim G, Moon S, Choi JH. Deep learning with multimodal integration for predicting recurrence in patients with non-small cell lung cancer. Sensors (Basel) 2022 Aug 31;22(17):6594. [doi: 10.3390/s22176594] [Medline: 36081053]
46. Franceschini D, Cozzi L, De Rose F, et al. A radiomic approach to predicting nodal relapse and disease-specific survival in patients treated with stereotactic body radiation therapy for early-stage non-small cell lung cancer. Strahlenther Onkol 2020 Oct;196(10):922-931. [doi: 10.1007/s00066-019-01542-6] [Medline: 31722061]
47. Huang P, Illei PB, Franklin W, et al. Lung cancer recurrence risk prediction through integrated deep learning evaluation. Cancers (Basel) 2022 Aug 27;14(17):4150. [doi: 10.3390/cancers14174150] [Medline: 36077686]
48. Ho WLJ, Fetisov N, Hall LO, Goldgof D, Schabath MB. Evaluating clinical and radiomic features for predicting lung cancer recurrence pre- and post-tumor resection. Proc SPIE Int Soc Opt Eng 2024 Feb;12926:1292623. [doi: 10.1117/12.3006091] [Medline: 38993353]
49. Lou F, Sima CS, Rusch VW, Jones DR, Huang J. Differences in patterns of recurrence in early-stage versus locally advanced non–small cell lung cancer. Ann Thorac Surg 2014 Nov;98(5):1755-1760. [doi: 10.1016/j.athoracsur.2014.05.070] [Medline: 25110337]
50. Zhang N, Liang R, Gensheimer MF, et al. Early response evaluation using primary tumor and nodal imaging features to predict progression-free survival of locally advanced non–small cell lung cancer. Theranostics 2020;10(25):11707-11718. [doi: 10.7150/thno.50565] [Medline: 33052242]
51. Huber RM, Kauffmann-Guerrero D, Hoffmann H, Flentje M. New developments in locally advanced nonsmall cell lung cancer. Eur Respir Rev 2021 Jun 30;30(160):200227. [doi: 10.1183/16000617.0227-2020] [Medline: 33952600]
52. Schneider BJ, Daly ME, Kennedy EB, et al. Stereotactic body radiotherapy for early-stage non–small-cell lung cancer: American Society of Clinical Oncology endorsement of the American Society for Radiation Oncology evidence-based guideline. J Clin Oncol 2018 Mar 1;36(7):710-719. [doi: 10.1200/JCO.2017.74.9671] [Medline: 29106810]
53. Liu Y, Chang Y, Zha X, et al. A combination of radiomic features, imaging characteristics, and serum tumor biomarkers to predict the possibility of the high-grade subtypes of lung adenocarcinoma. Acad Radiol 2022 Dec;29(12):1792-1801. [doi: 10.1016/j.acra.2022.02.024] [Medline: 35351366]
54. Yang M, Shi L, Huang T, et al. Value of contrast-enhanced magnetic resonance imaging-T2WI-based radiomic features in distinguishing lung adenocarcinoma from lung squamous cell carcinoma with solid components >8 mm. J Thorac Dis 2023 Feb 28;15(2):635-648. [doi: 10.21037/jtd-23-142] [Medline: 36910079]
55. Recht MP, Dewey M, Dreyer K, et al. Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations. Eur Radiol 2020 Jun;30(6):3576-3584. [doi: 10.1007/s00330-020-06672-5] [Medline: 32064565]
56. Ponsiglione A, Stanzione A, Spadarella G, et al. Ovarian imaging radiomics quality score assessment: an EuSoMII radiomics auditing group initiative. Eur Radiol 2023 Mar;33(3):2239-2247. [doi: 10.1007/s00330-022-09180-w] [Medline: 36303093]
57. Park JE, Kim D, Kim HS, et al. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to Radiomics Quality Score and TRIPOD statement. Eur Radiol 2020 Jan;30(1):523-536. [doi: 10.1007/s00330-019-06360-z] [Medline: 31350588]
58. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. Int J Radiat Oncol Biol Phys 2018 Nov 15;102(4):1143-1158. [doi: 10.1016/j.ijrobp.2018.05.053] [Medline: 30170872]
59. Zwanenburg A, Vallières M, Abdalah MA, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology 2020 May;295(2):328-338. [doi: 10.1148/radiol.2020191145] [Medline: 32154773]
60. Berenguer R, Pastor-Juan MDR, Canales-Vázquez J, et al. Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters. Radiology 2018 Aug;288(2):407-415. [doi: 10.1148/radiol.2018172361] [Medline: 29688159]
61. Welch ML, McIntosh C, Haibe-Kains B, et al. Vulnerabilities of radiomic signature development: the need for safeguards. Radiother Oncol 2019 Jan;130:2-9. [doi: 10.1016/j.radonc.2018.10.027] [Medline: 30416044]
62. Meyer M, Ronald J, Vernuccio F, et al. Reproducibility of CT radiomic features within the same patient: influence of radiation dose and CT reconstruction settings. Radiology 2019 Dec;293(3):583-591. [doi: 10.1148/radiol.2019190928] [Medline: 31573400]
63. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology 2016 Feb;278(2):563-577. [doi: 10.1148/radiol.2015151169] [Medline: 26579733]
64. van Velden FHP, Kramer GM, Frings V, et al. Repeatability of radiomic features in non–small-cell lung cancer [(18)F]FDG-PET/CT studies: impact of reconstruction and delineation. Mol Imaging Biol 2016 Oct;18(5):788-795. [doi: 10.1007/s11307-016-0940-2] [Medline: 26920355]

XSL•FO
RenderX

65. Parmar C, Rios Velazquez E, Leijenaar R, et al. Robust radiomics feature quantification using semiautomatic volumetric segmentation. PLoS ONE 2014;9(7):e102107. [doi: 10.1371/journal.pone.0102107] [Medline: 25025374]
66. Orlhac F, Soussan M, Maisonobe JA, Garcia CA, Vanderlinden B, Buvat I. Tumor texture analysis in 18F-FDG PET: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. J Nucl Med 2014 Mar;55(3):414-422. [doi: 10.2967/jnumed.113.129858] [Medline: 24549286]
67. Nestle U, Kremp S, Schaefer-Schuler A, et al. Comparison of different methods for delineation of 18F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non–small cell lung cancer. J Nucl Med 2005 Aug;46(8):1342-1348. [Medline: 16085592]
68. Ye J, Luo J, Xu S, Wu W. One-slice CT image based kernelized radiomics model for the prediction of low/mid-grade and high-grade HNSCC. Comput Med Imaging Graph 2020 Mar;80:101675. [doi: 10.1016/j.compmedimag.2019.101675] [Medline: 31945637]

Abbreviations

CT: computed tomography
ML: machine learning
NSCLC: non–small cell lung cancer
PET-CT: positron emission tomography–computed tomography
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search Extension
REM: random effects model
RQS: Radiomics Quality Score
SBRT: stereotactic body radiation therapy
TNM: tumor node metastasis

Edited by S Brini; submitted 16.Sep.2025; peer-reviewed by M Liao, O Tutsoy; revised version received 13.Jan.2026; accepted 15.Jan.2026; published 02.Apr.2026.

Please cite as:
Wu J, Zhang Y, Wang J, Chen C, Hu S, Chen W
Accuracy of Radiomics-Based Machine Learning for Predicting Risk of Recurrence in Non–Small Cell Lung Cancer: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e84223
URL: https://www.jmir.org/2026/1/e84223
doi:10.2196/84223

© Junpei Wu, Ye Zhang, Jiaye Wang, Chengshui Chen, Shiyu Hu, Wenyu Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Multimodal AI for Alzheimer Disease Diagnosis: Systematic Review of Datasets, Models, and Modalities

Ziwen Yu[1*], MSc; Anthony Mulholland[1*], PhD; Tianyan Huang[2]; Qiang Liu[1], PhD

[1]School of Engineering Mathematics and Technology, University of Bristol, Tankard's Close, Ada Lovelace Building, Bristol, United Kingdom

[2]Medical Physics and Biomedical Engineering, University College London, London, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Qiang Liu, PhD
School of Engineering Mathematics and Technology, University of Bristol, Tankard's Close, Ada Lovelace Building, Bristol, United Kingdom

Abstract

Background: Early detection of Alzheimer disease (AD) is essential for timely intervention; yet, diagnostic performance varies widely across modalities and datasets. Recent multimodal artificial intelligence (AI) models have made significant progress, but the evidence base remains fragmented due to heterogeneous datasets, modeling frameworks, and reporting quality.

Objective: This systematic review aimed to analyze studies on multimodal AI models for AD diagnosis, prognosis, and risk prediction over 5 years. We evaluated dataset characteristics, modality combinations, modeling strategies, performance metrics, and methodological limitations. We further discuss real-world implications and translational pathways.

Methods: Following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines, we systematically searched PubMed, IEEE Xplore, Scopus, ACM Digital Library, Cochrane, and arXiv, with the final datasets last searched on November 15, 2025. Studies applying multimodal machine learning or deep learning to AD, mild cognitive impairment, and dementia outcomes were included, whereas studies using a single modality or lacking sufficient methodological detail were excluded. QUADAS-2 (Revised Quality Assessment of Diagnostic Accuracy Studies tool) assessed risk of bias. Extracted performance results were synthesized across 4 major multimodal dataset families.

Results: A total of 66 studies met the inclusion criteria. Across datasets, multimodal models consistently outperformed single-modal baselines. Alzheimer's Disease Neuroimaging Initiative–based diagnosis achieved an average accuracy of 92.5% (SD 3.8%), while mild cognitive impairment–conversion models achieved an average area under the curve (AUC) of 0.922 (SD 0.045), and several fusion architectures reported AUCs above 0.95. In contrast, UK Biobank risk-prediction studies reported an average AUC of 0.84 (SD 0.056), and this reflects performance in large, population-based datasets. DementiaBank speech-language studies achieved an average AUC of 0.813 (SD 0.042), and cross-lingual AD detection achieved an accuracy of 77% (SD 6.5%). Self-collected multimodal datasets demonstrated average accuracies around 96% (SD 2.4%), but their generalizability is limited due to small sample sizes and single-center designs.

Conclusions: This systematic review demonstrates that multimodal AI models consistently outperform single-modal models for AD diagnosis, prognosis, and risk prediction by integrating complementary biological, clinical, and behavioral information. Unlike prior reviews, this review provides a unified synthesis across heterogeneous clinical, imaging, genetic, and linguistic datasets, enabling cross-domain comparison of modeling strategies and performance. However, the generalizability of reported performance was limited due to substantial heterogeneity in dataset composition, outcome definitions, and validation, and prevalent risks of bias. By evaluating these factors, this review clarifies where current evidence is robust and where caution is warranted. The findings highlight the need for standardized multimodal benchmarks, transparent evaluation protocols, and clinically grounded model design to enable reliable real-world deployment. Overall, this work advances the field by framing multimodal AI not only as a performance-driven tool but also as a translational framework for equitable, interpretable, and scalable AD diagnosis.

Trial Registration: PROSPERO CRD420251241895; https://www.crd.york.ac.uk/PROSPE-RO/view/CRD420251241895

(J Med Internet Res 2026;28:e85414) doi:10.2196/85414

KEYWORDS

alzheimer's disease, neurodegenerative disease; multimodal dataset; machine learning; deep learning; multimodal fusion; computer-aided diagnosis; early diagnosis

XSL•FO
RenderX

Introduction

Alzheimer disease (AD) is the most prevalent neurodegenerative disorder and the leading cause of dementia worldwide [1]. With an aging global population, AD has become one of the most costly and deadly diseases of the 21st century, imposing profound emotional, financial, and caregiving burdens on patients, families, and health systems. By 2050, the number of people with AD is projected to rise from 55 million in 2020 to approximately 139 million [1].

The progression of AD includes the preclinical stage, mild cognitive impairment (MCI), and symptomatic stages, with varying degrees of symptom severity. The preclinical stage is a key window for intervention, during which neuropathological changes have commenced, but clinical symptoms remain largely undetectable [2]. Despite advances in awareness and screening, up to 75% of dementia cases remain undiagnosed worldwide, particularly in low- and middle-income countries [3]. This persistent diagnostic gap highlights the need for low-cost, scalable, and accurate early detection tools to enable timely intervention and slow disease progression [4].

Artificial intelligence (AI) has emerged as a promising approach for improving the early detection and management of AD. By systematically integrating and analyzing multimodal data, AI-based diagnostic frameworks offer powerful tools to enhance early detection accuracy and facilitate timely intervention.

Recent work has used transformer-based models to integrate imaging, genetic, and linguistic data. Multimodal transformers combining magnetic resonance imaging (MRI) or positron emission tomography (PET) with clinical features and cognitive assessments have reported improved diagnostic accuracy and interpretability [5-7]. In parallel, GPT-style architectures, BERT (Bidirectional Encoder Representations From Transformers) variants, and domain-adapted language models improve extraction of linguistic and semantic markers linked to early cognitive decline [8,9]. Self-supervised speech models also perform strongly for detecting MCI and early AD from spontaneous speech [10]. Together, these advances reflect a shift toward unified, more interpretable, and clinically translatable multimodal systems that capture both biological and behavioral aspects of AD.

Traditional machine learning (ML) [11], ensemble methods [12], deep learning [13], and reinforcement learning (RL) [14] can perform well on unimodal data, but clinical diagnosis integrates structural and behavioral information [15]. Unimodal AI can therefore diverge from clinical workflows and miss complementary signals (eg, MRI for structural change plus speech features for cognitive decline [16]), increasing the risk of modality-specific overfitting and poorer real-world performance. Accordingly, recent work has shifted toward multimodal integration for AD diagnosis, yet many studies emphasize incremental accuracy gains while underaddressing generalizability, interpretability, and cost-effectiveness needed for adoption. The literature also remains fragmented: recent reviews often cover multimodal clinical phenotyping datasets [17] and multimodal linguistic cognitive-impairment datasets [18] separately, obscuring cross-modal insights such as how imaging and speech biomarkers might jointly improve early detection.

Recent multimodal methods have substantially improved AD detection. However, a comprehensive systematic review that integrates evidence across both clinical and linguistic modalities, fusion strategies, and critically evaluates methodological quality, dataset diversity, and reporting transparency is still lacking. To address these gaps, this review investigates how multimodal models are applied to AD diagnosis, prognosis, and risk prediction and compares performance across different modality combinations and dataset families published between 2019 and 2025. We also examine modeling and fusion strategies alongside validation practices and assess methodological quality and risk of bias using QUADAS-2 (Revised Quality Assessment of Diagnostic Accuracy Studies tool). Furthermore, key multimodal combinations within public datasets are analyzed in relation to their diagnostic performance, and datasets are categorized to evaluate their suitability for AD research and clinical translation. Overall, this review provides a comprehensive synthesis of multimodal AI in AD diagnosis, bridges previously disconnected research streams, and offers practical guidance for future model development and clinical adoption.

Methods

Study Design

This review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines [19], with the search procedures reported following PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) [20] and developed using the principles outlined in the Cochrane Handbook [21]. These methods were applied to systematically identify and evaluate studies on computer-aided AD diagnosis, with a particular focus on those using multimodal clinical phenotyping datasets and multimodal linguistic-based cognitive impairment datasets.

Source of the Study and Search Criteria

We developed and internally reviewed independent search strategies (no external peer review). We manually searched multiple databases to identify AI-driven multimodal approaches for AD diagnosis, rather than using an integrated multidatabase platform. As this review targets methodological advances mainly reported in peer-reviewed computational literature, we did not search trial registries (ClinicalTrials.gov, World Health Organization's International Clinical Trials Registry Platform). We also avoided validated or published filters, instead iteratively refining customized controlled-vocabulary and free-text terms for AD or dementia, multimodal data, and AI through pilot screening to maximize sensitivity.

Searches were performed in PubMed (447 records; January 1, 2019, to November 13, 2025), Scopus (1086 records; all years through November 13, 2025, filtered to PUBYEAR > 2018), IEEE Xplore (2229 records; January 1, 2020 to November 13, 2025), ACM Digital Library (2067 records; 1 January 2020, to 15 November 2025), Cochrane Library (1061 records; all available years through November 15, 2025), and arXiv (1081

XSL•FO
RenderX

records; all available years through November 15, 2025). We included the verbatim search strings for all databases, and because arXiv does not support bulk export, an arXiv search Python (Python Software Foundation) script is provided in Multimedia Appendix 1 [11,14,22-44].

Eligibility Criteria

The inclusion criteria were if studies were considered eligible if they met all the following conditions: (1) focused on AD, MCI, or related dementias as the primary clinical outcome; (2) applied AI or ML methods for computer-aided diagnosis, classification, or prediction; (3) used multimodal data, defined as any combination of at least two distinct modalities (eg, neuroimaging, clinical phenotyping, genetics, or linguistic features); (4) reported quantitative evaluation metrics; and (5) written in English.

The exclusion criteria were if studies met any of the following conditions: (1) single-modal approaches using only a single imaging modality, cognitive test, or biomarker, without any multimodal integration; (2) works without reported performance metrics or with insufficient methodological detail; (3) works not addressing diagnosis, classification, or prediction (eg, treatment response, drug trials, and lifestyle interventions); (4) duplicate publications or overlapping datasets without providing additional methodological contribution; and (5) non-English publications.

Selection Process

The study selection process followed the PRISMA 2020 guidelines, and the protocol was registered. The final search update was conducted in November 2025. All records retrieved from the databases were first imported into Zotero, where duplicates were automatically detected and removed.

The initial search identified 7435 records. After removing 3047 duplicates, 4388 records remained for title and abstract screening. A total of 4021 records were obviously irrelevant at the title and abstract level, 252 studies were excluded for the following main reasons:

- Focused on outcomes unrelated to AD diagnosis, classification, or prediction (eg, drug trials, treatment response, lifestyle interventions; n=140).
- Used unimodal data without multimodal integration (n=46).
- No sufficient methodological details (n=47).

Finally, 66 studies were included in the systematic synthesis, and all were successfully retrieved (reports not retrieved=0).

Overview of AI-Assisted AD Diagnosis

The workflow of AI-assisted AD diagnosis involves 3 stages, as illustrated in Figure 1. The initial stage involves comprehensive data acquisition, where information is collected from multiple modalities, including neuroimaging, biomarkers, genetics, and speech or behavioral signals. The second stage involves feature extraction and model development, followed by an interpretable and explainable analysis to ensure that AI models can effectively support clinical decision-making.

XSL•FO
RenderX

Figure 1. Overview of the AI pipeline for AD diagnosis. Multimodal inputs undergo preprocessing and feature extraction before model training for classification or regression tasks. Model interpretability supports explanation and performance evaluation. AD: Alzheimer disease; AI: artificial intelligence; LIME: Local Interpretable Model-Agnostic Explanations; SHAP: Shapley Additive Explanations; XAI: explainable artificial intelligence.

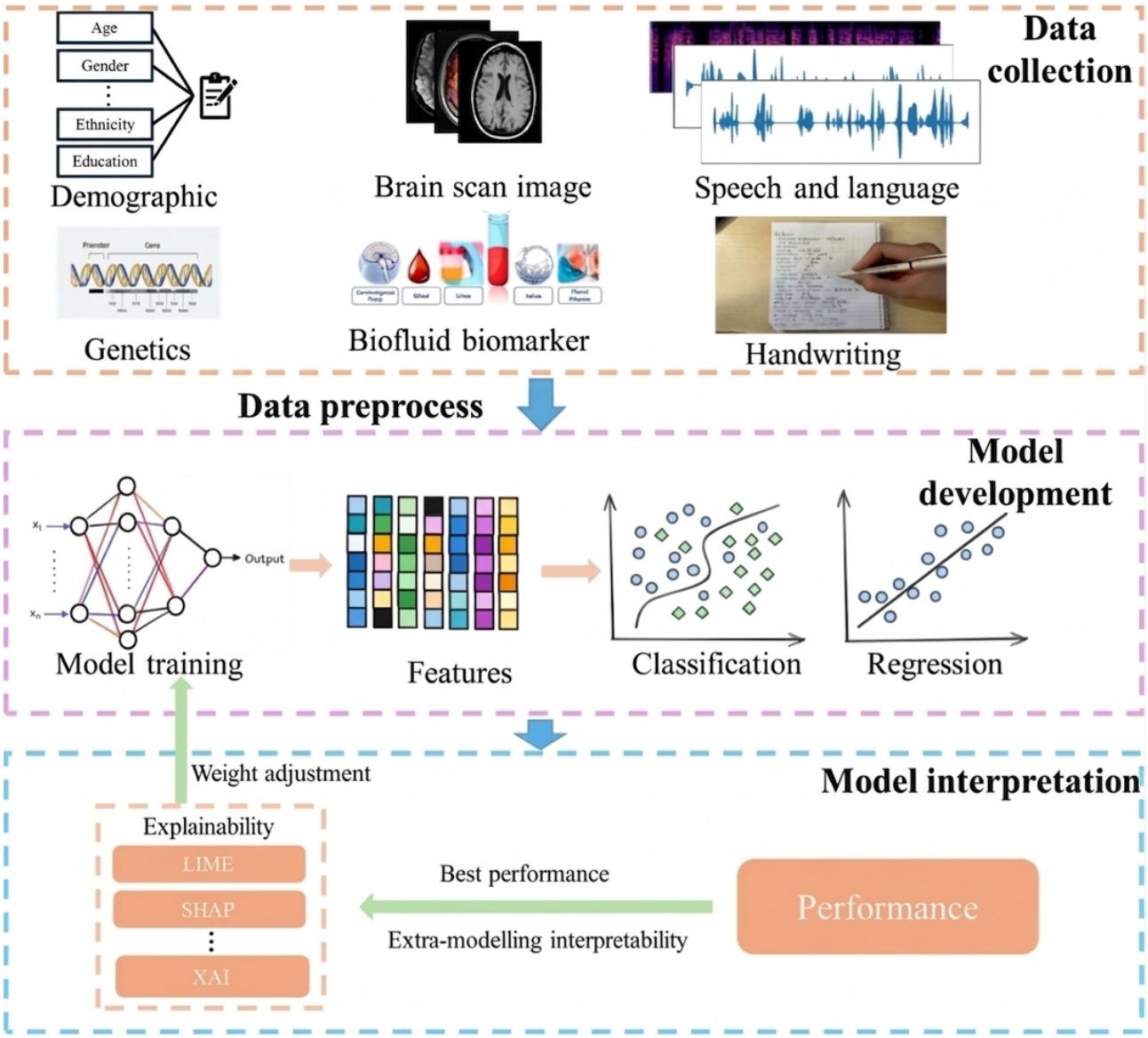

Performance Evaluation Metrics

To ensure the clinical applicability and scientific rigor of computer-aided diagnosis models for AD, it is essential to systematically evaluate their performance using a variety of quantitative metrics. We have summarized all performance evaluation metrics in Multimedia Appendix 2.

Risk of Bias and Quality Assessment

We assessed methodological quality with QUADAS-2 [45], evaluating risk of bias in 4 domains. Patient selection raised the main concern: 61% (40/66) of outcomes were high risk due to poor reporting or nonrepresentative sampling. For the index test, 76% (50/66) were unclear risk because procedures and decision thresholds were insufficiently described, and 20% (13/66) were high risk. The reference standard showed 76% (50/66) unclear risk from limited methodological detail, with no high-risk ratings. Flow and timing had the greatest uncertainty: 85% (56/66) were unclear owing to missing information on testing intervals and participant flow. Figure 2 summarizes domain-level risk distributions; Multimedia Appendix 3 reports study-level assessments.

XSL•FO
RenderX

Figure 2. Summary of the QUADAS-2 plot across the 66 included studies in the domain. QUADAS-2: Revised Quality Assessment of Diagnostic Accuracy Studies tool.

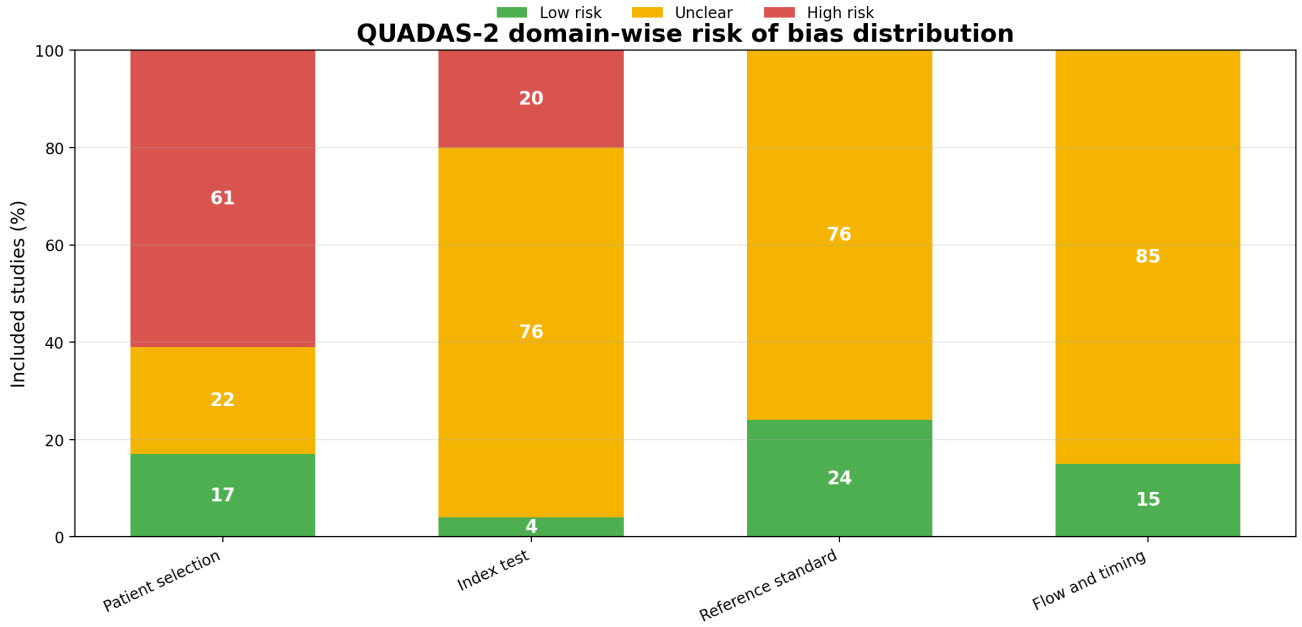

Given frequent unclear and high risk in key domains, we interpreted diagnostic performance cautiously, especially without external validation or a clearly defined reference standard. Future benchmarking should emphasize transparent reporting, prespecified thresholds, and multicenter evaluation to reduce bias and improve reproducibility.

Results

Overview

Following study selection (the complete selection process is summarized in the PRISMA 2020 flow diagram, Figure 3), we first summarize the overall profile of the included literature to contextualize the subsequent synthesis. Figure 4 provides a temporal overview (2019 - 2025) of modeling approaches across included studies, illustrating how methodological focus has shifted over time and informing interpretation of the evidence base.

XSL•FO
RenderX

Figure 3. Flow diagram of PRISMA. AD: Alzheimer disease; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

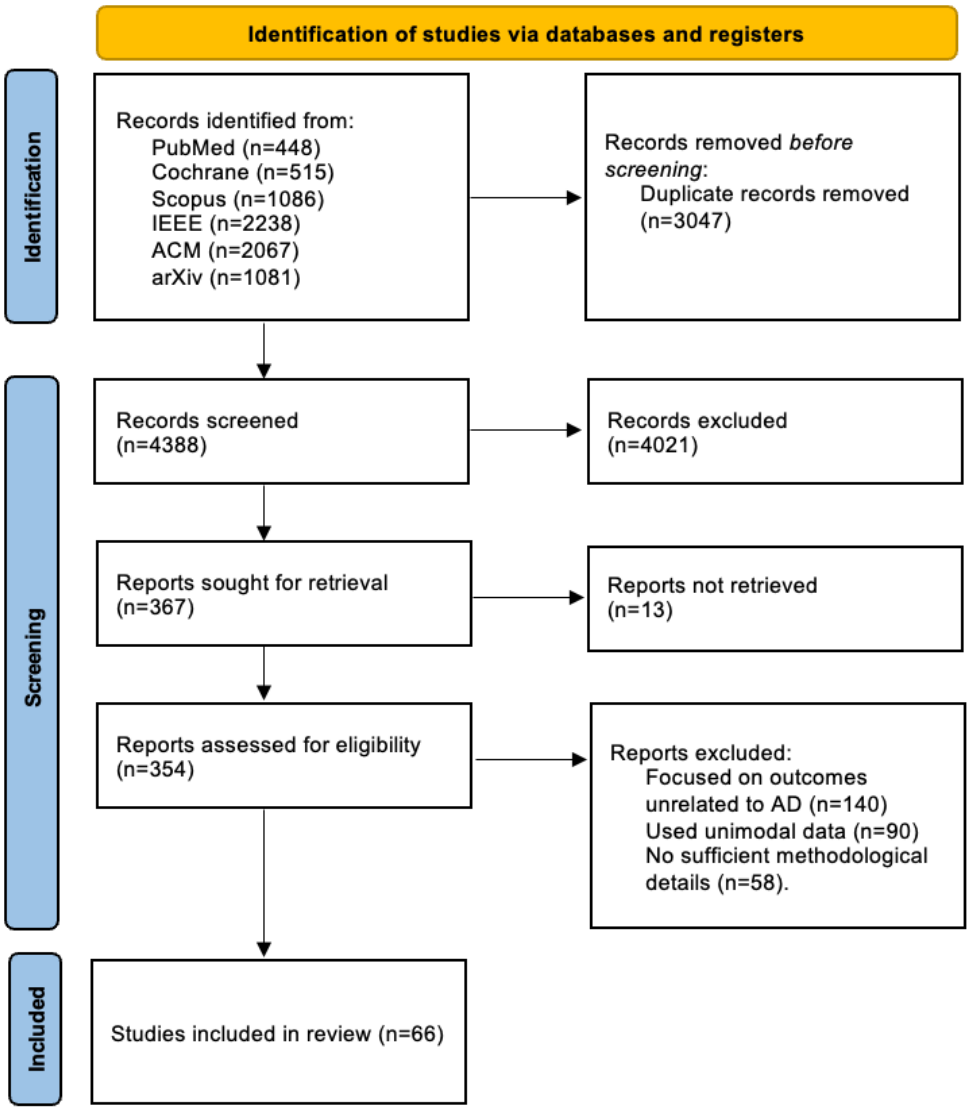

XSL•FO
RenderX

Figure 4. Temporal trends of machine-learning methods used for AD diagnosis (2019 - 2025). AD: Alzheimer disease.

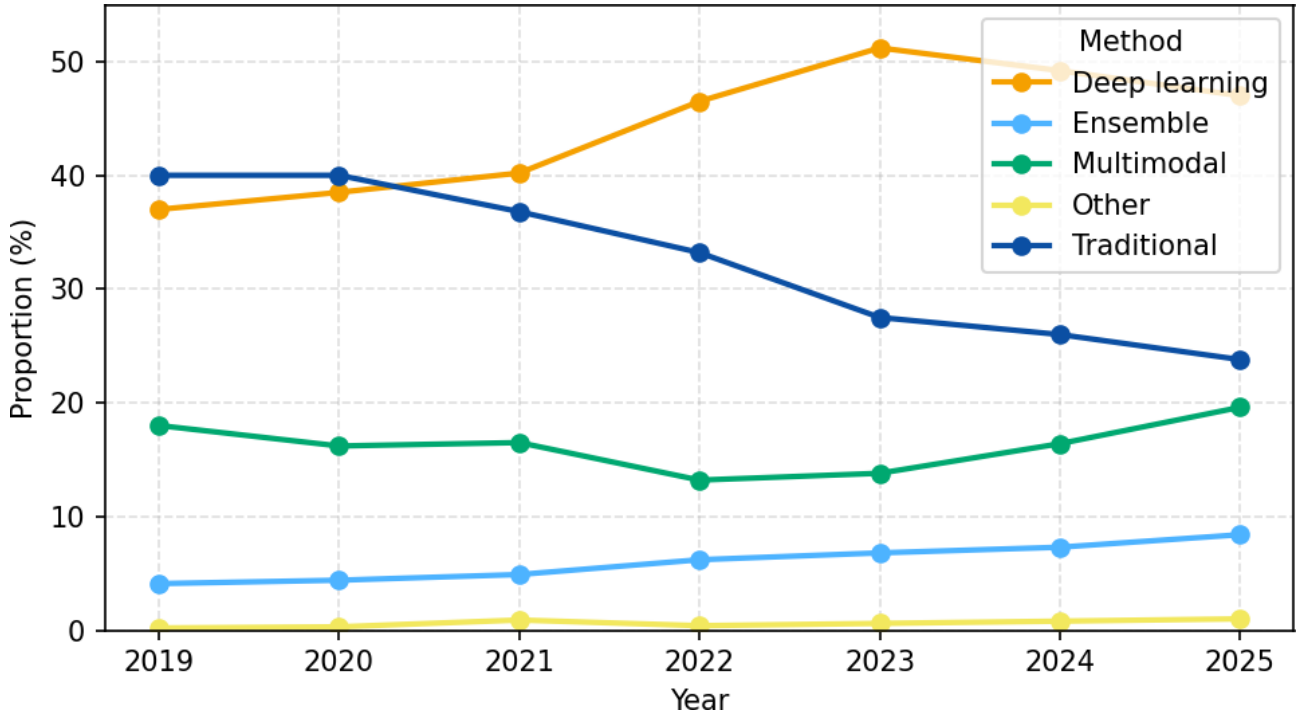

Single Modality

Overview

If readers are already familiar with traditional ML and deep learning approaches, they may wish to proceed directly to the next section, which focuses on multimodal data integration for AD diagnosis.

A concise overview of these baseline methods is provided in Multimedia Appendix 4, which also includes a summary of RL. As most RL studies address sequential decision-making tasks rather than direct diagnostic modeling, their methodological details are presented in Multimedia Appendix 4, to maintain focus on multimodal diagnostic frameworks in the main text.

Deep Learning

Compared with traditional ML, deep learning enables hierarchical feature extraction, capturing complex patterns in high-dimensional data. It is therefore widely used to process and integrate AD-related multimodal inputs, including neuroimaging, clinical scores, genetics, and speech. Key approaches and findings are summarized below.

Recurrent neural networks are effective for modeling sequential data such as longitudinal clinical records and speech signals, but they are susceptible to vanishing gradients in long sequences [46-49]. Long short-term memory networks address this limitation through gated memory mechanisms, enabling more stable training and improved capture of long-term dependencies. Consequently, long short-term memory models have been widely applied in AD research for analyzing temporal and sequential modalities [50-53].

The transformer model, which leverages attention mechanisms, dynamically assigns different weights to input features based on their relative importance. Each layer of the transformer consists of multiple attention heads, allowing the model to capture diverse feature representations by attending to various aspects. Transformer models use attention mechanisms to weight input features and capture diverse representations through multihead attention, enabling efficient and scalable training [54]. Owing to these advantages, they have been widely adopted in AD diagnosis and multimodal learning, where their encoder-decoder architecture facilitates effective integration of heterogeneous data sources [13,55-57].

Ensemble Learning

Ensemble learning improves generalization and robustness by combining multiple base models, including bagging and boosting methods such as AdaBoost (Adaptive Boosting), XGBoost (Extreme Gradient Boosting), and LightGBM (Light Gradient-Boosting Machine), and has been widely applied in AD detection and progression prediction [12,58-60]. However, ensemble models may introduce redundant features, offer limited gains on small datasets, and incur higher computational costs, which can restrict real-time or resource-constrained deployment.

Summarization for Single Modality

Traditional single-modality ML approaches can achieve high performance in AD-related tasks; however, they are constrained by several inherent limitations:

First, regarding information completeness, structural MRI alone has limited sensitivity to functional and molecular changes and cannot fully capture AD-related cognitive and behavioral alterations. Combining MRI with PET, neuropsychological tests, speech, electroencephalography (EEG), and genetic or biomarker data provides a more complete, multidimensional view of disease progression and patient heterogeneity [61].

Second, regarding model robustness, in multimodal data, residual noise in one modality may persist despite denoising, but other modalities can provide complementary signals that

XSL•FO
RenderX

improve robustness. Leveraging multisensory-style integration, multimodal models better reflect biological cognition and can yield more reliable decisions [62].

Third, in cross-modal learning, transformer architectures use cross-modal attention to learn associations between modalities. Some studies apply them in weakly supervised or cross-modal guided settings, using one modality to constrain or guide representation learning in another [63].

Fourth, in real-world decision-making, multimodal learning better matches real-world diagnosis, which integrates multiple information sources. Using diverse modalities aligns models with clinical workflows and improves translational potential in practice [64].

Therefore, this review analyzes the methodological strengths and limitations of multimodal models for computer-assisted AD diagnosis, focusing on how dataset grouping and classification choices affect evaluation. By classifying datasets, we enable model comparisons under a unified setup, allowing for more direct assessment of generalizability and cross-dataset stability.

Multimodal Data

Multimodal Dataset Overview

High-quality data plays a pivotal role in training AI models for computer-aided diagnosis and detection. Robust datasets not only enhance the generalization ability of models but also help mitigate the risk of overfitting. Commonly used datasets for AI-assisted diagnosis of AD can be broadly categorized into 2 types. The first type of multimodal clinical phenotyping dataset, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI), UK Biobank, and the Open Access Series of Imaging Studies (OASIS), focuses on neuroimaging modalities, including MRI, functional MRI, genetic data, and electronic health records. The second type—multimodal cognitive-linguistic behavioral dataset centers on sequential data modalities, including audio, video, and transcribed language data, such as the Pitt Corpus and ADReSS (Alzheimer's Dementia Recognition Through Spontaneous Speech). Each dataset type provides unique features that contribute to the comprehensive modeling of AD progression and diagnosis. The commonly used datasets, along with their population demographics and associated modalities, are summarized in Table 1.

XSL•FO
RenderX

Table . Commonly used dataset.

Dataset	Population demographics		Modalities	Link
	Male	Age (years)		
UK Biobank	23,000 (46)[a]	56.5 (8.1)[b]	MRI[c], fMRI[d], genetic, lifestyle scores, activity monitor, and EHR[e]	[65]
ADNI[f]				
ADNI-1	469 (57.3)[a]	75 (6.9)[b]	MRI, PET[g], genetic, and EHR	[66]
ADNI-GO[h] and 2	473 (53)[a]	72.5 (7.2)[b]	MRI, PET, genetic, and EHR	[66]
ADNI-3	471 (49)[a]	74.9 (8.1)[b]	MRI, PET, genetic, and EHR	[66]
OASIS[i]				
OASIS-1	177 (42.5)[a]	57 (39)[b]	MRI, PET, CT, genetic, and EHR	[67]
OASIS-2	60 (40)[a]	78 (18)[b]	MRI, PET, CT, genetic, and EHR	[67]
OASIS-3	622 (45)[a]	69 (26.56)[b]	MRI, PET, CT, genetic, and EHR	[67]
OASIS-4	663 subjects	57.5 (36.5)[b]	MRI, PET, CT, genetic, and EHR	[67]
NACC[j]	23,625 (44.62)[a]	73.3 (10.5)[b]	MRI, PET, genetics, and EHR	[68]
FHS[k]	718 (42)[a]	80.76 (8.2)[b]	MRI, genetic, and EHR	[69]
AIBL[l]	289 (43.72)[a]	73.5 (7.03)[b]	MRI, PET, genetic, and EHR	[70]
Dementia bank				
Pitt corpus	HC[m]: 104, AD[n]: 208/552	—[o]	Audio and text	[71]
ADReSS[p]	HC: 78, AD: 78/156	—	Audio and text	[71]
ADReSSo[q]	HC: 115, AD: 122/237	—	Audio and text	[71]
ADReSS-M[r]	HC: 143, AD: 148/291	—	Audio and text	[71]
TAUKADIAL[s]	106/507	—	Audio and text	[71]
Multimodal dementia corpus	HC: 10, AD: 12/816	—	Audio, typed, and hand-written	[72]
ADReFV[t]	AD: 25	66.68 (2.08)[b]	Video	[73]
GENCODE[u]	Human: 78,686	—	Genetics	[74]

[a]n (%).
[b]Mean (SD).
[c]MRI: magnetic resonance imaging.
[d]fMRI: functional magnetic resonance imaging.
[e]EHR: electronic health record.
[f]ADNI: Alzheimer's Disease Neuroimaging Initiative.
[g]PET: positron emission tomography.
[h]ADNI-GO: Alzheimer's Disease Neuroimaging Initiative – Grand Opportunity.
[i]OASIS: Open Access Series of Imaging Studies.
[j]NACC: National Alzheimer's Coordinating Centre.

XSL•FO
RenderX

[k]FHS: Framingham Heart Study.

[l]AIBL: Australian Imaging, Biomarkers and Lifestyle Study.

[m]HC: health control.

[n]AD: Alzheimer disease.

[o]Not available.

[p]ADReSS: Alzheimer's Dementia Recognition Through Spontaneous Speech.

[q]ADReSSo: Alzheimer's Dementia Recognition Through Spontaneous Speech 2021 Challenge.

[r]ADReSS-M: Multilingual Alzheimer's Dementia Recognition Through Spontaneous Speech Challenge.

[s]TAUKADIAL: Speech-Based Cognitive Assessment in Chinese and English.

[t]ADReFV: Alzheimer's Disease Recognition from Face & Voice.

[u]GENCODE: gene code.

To better understand current research trends, this paper reviewed studies from the past 5 years on multimodal AI models for AD diagnosis. The literature was categorized by dataset type. (1) Multimodal clinical phenotyping datasets: ADNI dominates this category, used in about 80% of studies, while others, such as UK Biobank, OASIS, National Alzheimer's Coordinating Centre (NACC), Framingham Heart Study, and Australian Imaging, Biomarkers and Lifestyle Study, account for the remaining 20%, mainly for supplementary analysis or external validation. (2) Multimodal cognitive-linguistic behavioral datasets: The ADReSS series is most widely used, representing around 70% of studies. The remaining 30% use datasets often for complementary analysis or benchmarking.

Multimodal Clinical Phenotyping Datasets

Multimodal clinical phenotyping datasets integrate MRI, PET, or diffusion tensor imaging, biomarkers from blood, cerebrospinal fluid, or genomics, and standardized cognitive assessments. This review summarizes representative resources, highlighting their modalities, distinguishing features, and contributions to diagnostic modeling (Table 2).

XSL•FO
RenderX

Table . Multimodal clinical phenotyping datasets related paper. The exceptionally high performances reported in some of these studies can be attributed to specific methodological factors: (1) for two studies [75] and [76], the absence of external validation likely inflated the results; (2) for another study [77], the use of a small but highly controlled dataset, extensive sample expansion, multimodal feature fusion, and pronounced disease-related electrophysiological signatures contributed to the elevated accuracy; and (3) for yet another study [78], the integration of rich gait features with optimized machine learning techniques in a controlled experimental setting facilitated unusually high performance.

Study	Datasets	Model type	Type of task	Modalities	Outcomes	Validation	Results	Limitation
Xue et al [79], 2024	NACC[a], ADNI[b], AIBL[c], FHS[d], PPMI[e], and OASIS[f]	Transformer-based multimodal model	Differential diagnosis of 10 etiologies	MRI[g] (T1, T2, FLAIR[h]), clinical, neuropsychological tests, and PET[i]	Differential diagnosis probabilities and AD[j]/MCI[k]/NC[l] classification	Internal: NACC held out; external: ADNI and FHS	Etiology classification AUROC[m] 0.96; strong alignment with PET biomarkers and neuropathology	Imbalanced etiologies, training label subjectivity, and limited racial diversity
Shi et al [80], 2018	ADNI	MM-SDPN[n]	Classification (AD vs NC, MCI vs NC)	T1 MRI and FDG-PET[o]	Classification accuracy	Comparative vs single-modality and state-of-the-art models	Outperformed single-modality DPN[p]/SDPN[q] and concatenated models	ADNI-only dataset (limits generalizability); ROI[r]-based features rather than voxel-wise
Allwright et al [81], 2023	UK Biobank	XGBoost[s]	Risk prediction (Incident AD)	Demographics, lifestyle, genetics, and medical history	Prediction of incident AD (2-10 years) and risk factor ranking	Internal: nested 3-fold cross-validation; external: held-out validation set	AUROC 0.77, APOE-ε4[t] identified as the strongest risk factor, and liver enzymes or frailty as predictors	*ICD-10*[u] under-ascertainment, healthy volunteer bias, and observational design
Gu et al [82], 2025	UK Biobank	LightGBM[v]	Risk prediction (incident dementia in ASCVD[w] patients)	Clinical, biological assays, cognitive tests, and physical measures	All-cause incident dementia, AD, and VD[x] incidence	Temporal: train (2006-2009) and test (2010 cohort)	5-year dementia AUC[y] 0.903, AD AUC 0.775, and accuracy 0.851	Sample mostly European descent, static baseline features, and potential overfitting
You et al [83], 2022	UK Biobank	LightGBM	Risk prediction (5/10-year horizon)	Demographics, lifestyle, blood biomarkers, and genetics	Incident all-cause dementia and AD prediction	Internal: 5-fold cross-validation	AUC 0.848 (all-cause), 0.862 (AD), and outperformed CAIDE[z] and DRS[aa]	Limited external validation, population predominantly White, and feature selection fully data-driven
Calvo et al [84], 2024	UK Biobank	Multivariable logistic regression	Risk association analysis	Questionnaire, ICD[ab] records, and genotypes	Odds of AD related to menopause type	Single cohort: multivariable adjustment	Early bilateral oophorectomy associated with 4-fold AD odds (OR[ac] 4.12) and HT[ad] use protective	Low case numbers in subgroups; self-reported HT use and healthy volunteer bias
Yi et al [85], 2025	UK Biobank, ADNI, PPMI, and IXI[ae]	3D-ViT[af]	BAG[ag] estimation and GWAS[ah]	T1-weighted MRI, genetics (SNP[ai] and xQTL[aj])	BAG and drug target prioritization	External: ADNI, PPMI, and IXI	MAE[ak] ≈ 2.6 and identified 7 high-confidence drug targets (eg, MAPT[al] and TNFSF12[am])	European-ancestry bias; lack of biological "ground truth" for brain age

XSL•FO
RenderX

Study	Datasets	Model type	Type of task	Modalities	Outcomes	Validation	Results	Limitation
Yousefzadeh et al [86], 2024	UK Biobank (retina cohort)	VGG-16[an] classifier + LAVA[ao] (XAI[ap])	Binary classification and explainability	Retinal fundus images	AD vs NC classification and neuron-level explanations	Internal: nested 5-fold cross-validation	Accuracy 71.4% and identified 7 latent clusters linking vascular and cognitive decline	Small AD sample size (n=100), cross-sectional design, and UK Biobank volunteer bias
Gong et al [87], 2023	UK Biobank	SuperBigFLICA (semisupervised Bayesian fusion)	Phenotype discovery	Multimodal MRI (47 modalities)	Latent components predictive of non-imaging phenotypes	Internal: train, validation, or test split	Up to 46% improvement over expert IDPs[aq] and interpretable multimodal modes	Linear modeling constraints and UK Biobank population bias
Lian et al [88], 2022	ADNI-1, ADNI-2, and AIBL	Attention-guided HybNet (3D FCN[ar] + hybrid network)	Diagnosis and prognosis	Structural T1 MRI	AD vs NC classification; pMCI[as] vs sMCI[at] prediction	External: trained ADNI-1, and validated ADNI-2 and AIBL	ADNI-2 AD vs NC accuracy 0.919 (AUC 0.965) and outperformed ROI/VBM[au] methods	Heavy preprocessing reliance, limited demographic diversity, and potential overfitting
Lian et al [89], 2022	ADNI-1 and ADNI-2	MWAN[av]	Joint regression of clinical scores	Structural T1 MRI and clinical scores	MMSE[aw], CDRSB[ax], and ADAS-Cog[ay] prediction	Cross-validation: across ADNI-1 and ADNI-2	Lower RMSE[az] and higher correlation coefficients than single-task baselines	Restricted to the ADNI cohorts and potential overfitting to the modest sample size
Li et al [90], 2019	ADNI-1, ADNI-GO/2[ba], and AIBL	3D CNN[bb] + Cox proportional hazards	Time-to-event prognosis	Hippocampal MRI patches and clinical variables	Progression from MCI to AD and risk stratification	External: trained ADNI-1, and validated ADNI-GO/2 and AIBL	C-index 0.864 (combined model) and significant risk-based stratification of MCI	Focus on the hippocampus only and potential cohort and scanner bias (1.5T vs 3T)
Qiu et al [15], 2022	NACC, ADNI, and ADCP[bc]	Multimodal deep learning (3D CNN + FCN)	Multiclass classification	Structural MRI, demographics, and neuropsychology	Diagnosis (NC, MCI, AD, and nADD[bd]) and saliency maps	External: trained NACC, and validated ADNI and independent cohorts	Performance comparable to neurologists and saliency aligned with pathology	Retrospective design and heterogeneity in protocols across cohorts
Oh et al [91], 2023	ADNI	LEAR[be] framework (CNN + RL[bf] + XAI)	Diagnosis and interpretation	Structural T1 MRI	AD vs non-AD classification and counterfactual maps	Internal: cross-validation on ADNI	Improved accuracy and generalization, and localized plausible atrophy patterns	Single-cohort (ADNI) and XAI evaluation, partly qualitative
Lian et al [92], 2020	ADNI-1 and ADNI-2	Hierarchical FCN	Diagnosis and atrophy localization	Structural T1 MRI	AD vs NC, MCI vs NC, and atrophy pattern mapping	External: trained ADNI-1 and tested ADNI-2	Improved accuracy vs conventional features and interpretable atrophy maps	ADNI-only; strong reliance on preprocessing and registration
Avsec et al [93], 2021	Genomic reference datasets	Enformer (transformer)	Genomic prediction	DNA sequence	Gene expression and chromatin state prediction	Internal: held-out chromosomes	Improved capture of long-range regulatory effects vs previous models	Limited to available cell types and assays

XSL•FO
RenderX

Study	Datasets	Model type	Type of task	Modalities	Outcomes	Validation	Results	Limitation
Yang et al [94], 2021	ADNI	Deep learning and super learner	Prognosis	MRI, cognitive, and biomarkers	Diagnostic classification and prognostic risk signature	Internal: cross-validation within ADNI	Derived signature distinguished diagnostic groups and progression risk	Limited external validation and restricted to the ADNI research cohort
Lee et al [95], 2024	ADNI and UK or Singapore Clinics	PPM[bg]	Prognosis (MCI to AD)	MRI (gray matter) and cognitive tests	Individualized prognostic index	External: independent real-world memory clinics	Accuracy ≈81.7%, AUC ≈0.84; and index predicted conversion better than atrophy alone	Heterogeneity in real-world clinical data and potential site effects
Zhu et al [96], 2021	ADNI and AIBL	DA-MIDL[bh]	Diagnosis	Structural MRI patches	AD vs NC and MCI vs NC	External: trained ADNI and tested AIBL	Higher accuracy and generalizability than baselines, and attention maps aligned with pathology	Reliance on structural MRI and potential dataset-specific overfitting
Zhang et al [97], 2024	ADNI	GCN[bi], SHAP[bj], and automatic fusion	Diagnosis	Cognitive, MRI, PET, and risk factors	AD vs non-AD diagnosis, and multimodal feature selection	Internal: two ADNI multimodal cohorts	Accuracies of 95.9% and 91.9%, and efficient selection of clinically important features	Complex model deployment and reliance on ADNI data
Velazquez and Lee [75], 2022	ADNI EMCI[bk]	Ensemble (random forest and CNN)	Prediction of conversion	DTI[bl] (ADC[bm] maps) and EHR[bn]	EMCI to AD conversion prediction	Internal: held-out test set	98.81% accuracy and feature importance explainability provided	Small converter sample size and potential overfitting
Zhang et al [76], 2024	ADNI	Multimodal learning machine (ELM[bo] ensemble)	Diagnosis	MRI features and neuropsychological tests	NC, MCI, and AD classification	Internal: cross-validation on ADNI	>98% accuracy and F_1-score, and no observed bias between MCI and AD	A single research cohort and very high accuracy require external verification
Bi et al [98], 2020	ADNI	Cluster evolutionary random forest	Diagnosis	Resting-state fMRI[bp] and SNP	AD vs control classification and biomarker identification	Comparative vs competing methods	Identified significant brain region-gene pairs and effective classification	Small multimodal sample size and complex hyperparameters
Bi et al [99], 2022	ADNI	Weighted evolutionary random forest	Pathogen detection	Resting-state fMRI and SNP	MCI identification and pathogenic factor extraction	Comparative vs state-of-the-art methods	Superior MCI identification performance, and highlighted key ROIs and SNPs	High-dimensional fusion features, small N, and overfitting risk
Hashmi and Barukab [100], 2023	OASIS	Deep RL and neural network	Staging	Structural MRI	4-class dementia staging	Internal: augmented vs baseline	RL augmentation improved accuracy by ≈6% and recall by ≈13%	Single open dataset; focus on MRI only
Wang et al [101], 2024	ADNI-1/2/3	Multimodal DL[bq] with an interaction layer	Prognosis (MCI to AD)	MRI, clinical, and genetics (SNP)	4-year conversion prediction	External: generalized to ADNI-3	AUC 0.962 (cross-validation), 0.939 (test); interaction effects improved accuracy	ADNI-only and potential overfitting despite cross-validation

XSL•FO
RenderX

Study	Datasets	Model type	Type of task	Modalities	Outcomes	Validation	Results	Limitation
Hatami et al [102], 2024	ADNI	DNN[br] and RL (data augmentation)	Classification	Structural MRI	AD vs NC classification	Comparative vs baseline augmentation approaches	Precision ≈0.95; RL-guided augmentation outperformed baselines	Single research cohort and no external clinical validation
Tabarestani et al [103], 2020	ADNI	Distributed multitask regression	Longitudinal progression	MRI, PET, CSF[bs], EEG[bt], and clinical	Prediction of longitudinal cognitive scores	Comparative vs unimodal or multimodal methods	Reduced errors, particularly in sparse or incomplete longitudinal data	Model complexity and potential sensitivity to hyperparameters
Burkhart et al [104], 2024	ADNI and Singapore Memory Clinic	Unsupervised multimodal trajectory modeling	Prognosis	Cognitive, amyloid PET, and MRI	Cognitive health clustering and progression prediction	External: real-world memory clinic data	Better stratification than standard clinical assessments and robust to missing data	Unsupervised complexity and reliance on ADNI for training
El-Sappagh et al [105], 2021	ADNI	Random forest and SHAP (multilayer)	Diagnosis and progression	11 modalities (MRI, PET, CSF, and clinical)	Multiclass diagnosis and MCI progression detection	Internal: cross-validation	Diagnosis accuracy 93.95%, progression accuracy 87.08%, and interpretable	High complexity and challenges for routine care deployment
Lee et al [106], 2024	ADNI and 4 Korean hospitals	GBM[bu]	Conversion prediction	MRI (T1, T2-FLAIR[bv]), amyloid PET, and clinical	MCI to AD conversion (4-year)	Internal: nested cross-validation with modality combinations	T1 and amyloid PET is the best combination, and T2-FLAIR did not improve prediction	Small multicenter sample and site and scanner heterogeneity
Yuan et al [107], 2021	ADNI	Multimodal cotraining (random forest)	MCI subtype classification	Structural MRI and SNP	sMCI vs pMCI classification	External: ADNI-2 independent test set	Accuracy 85.5% and cotraining outperformed single modality	Dependence on feature selection and ADNI-only
Cirincione et al [108], 2024	TADPOLE[bw] (ADNI)	Ensemble integration	Prediction	MRI, PET, clinical, and cognitive	Future dementia prediction in MCI	Internal: held-out test set	AUC 0.81, and outperformed XGBoost and deep learning baselines	Single research dataset and the complexity of multimodal integration
Cassani and Falk [109], 2020	Clinical EEG	Feature engineering and ML	Diagnosis and severity	Resting-state EEG	AD vs normal, and mild vs moderate AD classification	Internal: cross-validation	Modulation spectral features outperformed traditional EEG features	Small sample size, resting state only, and single center
Cilia et al [110], 2021	Custom (Naples)	Deep transfer learning (CNN)	Diagnosis	Online handwriting (dynamic)	Early AD detection	Internal: cross-validation	Dynamic features (color-encoded) are superior to shape-only images	Single-center dataset and task-specific protocol
Kmetzsch et al [111], 2022	PREV-DEMALS[bx]	Supervised variational autoencoder	Disease progression modeling	MRI and microRNA	Disease progression score (FTD[by]/ALS[bz])	Validation: synthetic data and cohort evaluation	Outperformed competing models in capturing progression trajectory	Small sample (rare disease) and cross-sectional data used for progression

XSL•FO
RenderX

Study	Datasets	Model type	Type of task	Modalities	Outcomes	Validation	Results	Limitation
Mengoudi et al [112], 2020	UCL[ca] and Insight 46	Self-supervised deep neural network	Diagnosis	Eye-tracking (gaze or pupil)	Dementia vs control classification	Comparative vs handcrafted features	Self-supervised features are more sensitive than handcrafted metrics	Modest sample size, mixed dementia subtypes, and specialized hardware
Tsai et al [113], 2024	Taiwan NHI[cb]	MAND[cc]	Incidence prediction	EHR (ICD codes) and demographics	Dementia incidence risk	Internal: held-out test set	AUC 0.901 and outperformed traditional CTR[cd] models	Coding errors in administrative data are specific to the Taiwan NHI
Park et al [22], 2024	Korean memory clinics	SVM[ce]	Diagnosis (MCI vs HC[cf])	VR[cg] biomarkers, MRI, and neuropsychological tests	MCI vs healthy control classification	Internal: train or test split	VR, MRI AUC 0.89, and VR biomarkers comparable to MRI alone	Small sample (n=54) and VR hardware requirement
Wu et al [114], 2022	Clinical EEG	WiGMM[ch]	Severity detection	Resting-state EEG	Unsupervised dementia degree detection	Internal: latent structure analysis	Captured latent dementia degrees matching clinical status	Unsupervised labeling requires careful interpretation
Zhang et al [115], 2025	Chinese memory clinics	FCRN[ci] and MLP[cj] (patch-based)	Diagnosis	MRI, PET, clinical, and genotype	AD vs normal and MCI vs normal classification	Internal: cross-validation	Accuracy ≈96% (AD), ≈92% (MCI), and interpretable probability maps	Single-country clinical cohorts and limited ethnic diversity
Fabietti et al [77], 2023	Mouse models	Ensemble machine learning	Early detection (animal)	LFP[ck]	AD vs control mouse classification	Internal: channel masking robustness tests	Accuracy 99.4% and robust to artifacts	Preclinical animal model results and small sample size
Seifallahi et al [78], 2022	Single center	SVM	Diagnosis	Kinect V2 (gait or TUG[cl])	AD vs healthy control classification	Internal: leave-one-out cross-validation	Accuracy 98.68% using 12 skeletal features	Small sample, case-control design may overestimate performance
Fan et al [116], 2024	CVD[cm] patients (Wuhan)	ViT[cn] (MRI) and XGBoost (clinical)	VCI[co] diagnosis	MRI (T1, T2-FLAIR) and clinical	Vascular cognitive impairment diagnosis	External: independent CVD dataset	The hybrid model has an AUC of 0.965 and is comparable to expert neurologists	CVD-specific cohort and complex ViT and XGBoost pipeline
Beebe-Wang et al [117], 2021	Aging cohort (US)	Nonlinear ML and SHAP	Imminent prediction (3 years)	Clinical, neuropsychological	Incident dementia within 3 years	Internal: cross-validation	Sparse model (4 tests) comparable to full battery	Prediction limited to a 3-year horizon and a single health system

XSL•FO
RenderX

Study	Datasets	Model type	Type of task	Modalities	Outcomes	Validation	Results	Limitation
Battineni et al [118], 2021	Public MRI dataset	Gradient boosting	Classification	MRI features and demographics	AD vs non-AD classification	Internal: cross-validation	Accuracy 97.58% (gradient boosting performed best)	Small public dataset and lack of external validation

[a]NACC: National Alzheimer's Coordinating Centre.
[b]ADNI: Alzheimer's Disease Neuroimaging Initiative.
[c]AIBL: Australian Imaging, Biomarkers, and Lifestyle Study.
[d]FHS: Framingham Heart Study.
[e]PPMI: Parkinson Progression Markers Initiative.
[f]OASIS: Open Access Series of Imaging Studies.
[g]MRI: magnetic resonance imaging.
[h]FLAIR: fluid-attenuated inversion recovery.
[i]PET: positron emission tomography.
[j]AD: Alzheimer disease.
[k]MCI: mild cognitive impairment.
[l]NC: normal control.
[m]AUROC: area under the receiver operating characteristic curve.
[n]MM-SPDN: multimodal stacked deep polynomial network.
[o]FDG-PET: fluorodeoxyglucose-positron emission tomography.
[p]DPN: deep polynomial network.
[q]SPDN: stacked deep polynomial network.
[r]ROI: region of interest.
[s]XGBoost: Extreme Gradient Boosting.
[t]APOE-ε4: apolipoprotein E epsilon 4 allele.
[u]*ICD-10*: *International Statistical Classification of Diseases, Tenth Revision.*
[v]LightGBM: Light Gradient-Boosting Machine.
[w]ASCVD: atherosclerotic cardiovascular disease.
[x]VD: vascular dementia.
[y]AUC: area under the curve.
[z]CAIDE: Cardiovascular Risk Factors, Aging, and Incidence of Dementia.
[aa]DRS: Dementia Risk Score.
[ab]*ICD*: *International Classification of Diseases.*
[ac]OR: odds ratio.
[ad]HT: hormone therapy.
[ae]IXI: Information Extraction From Images.
[af]3D-ViT: 3D vision transformer.
[ag]BAG: brain age gap.
[ah]GWAS: genome-wide association study.
[ai]SNP: single-nucleotide polymorphism.
[aj]xQTL: molecular quantitative trait locus.
[ak]MAE: mean absolute error.
[al]MAPT: microtubule-associated protein tau.
[am]TNFSF12: Tumor Necrosis Factor (Ligand) Superfamily, Member 12.
[an]VGG-16: Visual Geometry Group 16-Layer Network.
[ao]LAVA: Granular Neuron-Level Explainer.
[ap]XAI: explainable artificial intelligence.
[aq]IDP: imaging-derived phenotype.
[ar]FCN: fully convolutional network.
[as]pMCI: progressive mild cognitive impairment.
[at]sMCI: stable mild cognitive impairment.
[au]VBM: voxel-based morphometry.
[av]MWAN: multi-task weakly-supervised attention.

XSL•FO
RenderX

[aw]MMSE: Mini-Mental State Examination.
[ax]CDRSB: Clinical Dementia Rating–Sum of Boxes.
[ay]ADAS-Cog: Alzheimer Disease Assessment Scale–Cognitive Subscale.
[az]RMSE: root mean square error.
[ba]ADNI-GO/2: Alzheimer's Disease Neuroimaging Initiative – Grand Opportunity / Phase 2.
[bb]CNN: convolutional neural network.
[bc]ADPC: Alzheimer Disease Prediction Challenge.
[bd]nADD: non-Alzheimer disease dementia.
[be]LEAR: learn-explain-reinforce.
[bf]RL: reinforcement learning.
[bg]PPM: Predictive Prognostic Model.
[bh]DA-MIDL: dual attention multi-instance deep learning.
[bi]GCN: graph convolutional network.
[bj]SHAP: Shapley Additive Explanations.
[bk]EMCI: early mild cognitive impairment.
[bl]DTI: diffusion tensor imaging.
[bm]ADC: apparent diffusion coefficient.
[bn]EHR: electronic health record.
[bo]ELM: extreme learning machine.
[bp]fMRI: functional magnetic resonance imaging.
[bq]DL: deep learning.
[br]DNN: deep neural network.
[bs]CSF: cerebrospinal fluid.
[bt]EEG: electroencephalography.
[bu]GBM: Gradient Boosting Machine.
[bv]T2-FLAIR: T2-weighted fluid-attenuated inversion recovery.
[bw]TADPOLE: The Alzheimer Disease Prediction of Longitudinal Evolution.
[bx]PREV-DEMALS: Predict to Prevent Frontotemporal Lobar Degeneration and Amyotrophic Lateral Sclerosis.
[by]FTD: frontotemporal dementia.
[bz]ALS: amyotrophic lateral sclerosis.
[ca]UCL: University College London.
[cb]NHI: National Health Insurance.
[cc]MAND: Multimodal Attention Network.
[cd]CTR: clinical trial registration.
[ce]SVM: support vector machine.
[cf]HC: healthy control.
[cg]VR: virtual reality.
[ch]WiGMM: Warped Infinite Gaussian Mixture.
[ci]FCRN: fully convolutional residual network.
[cj]MLP: multilayer perceptron.
[ck]LFP: local field potentials.
[cl]TUG: Timed Up and Go.
[cm]CVD: cardiovascular disease.
[cn]ViT: vision transformer.
[co]VCI: vascular cognitive impairment.

UK Biobank Dataset

UK Biobank enables population-level association studies and early-risk modeling. It has been widely used in AD diagnosis research, including the following notable studies:

Recent UK Biobank–based studies have applied diverse multimodal ML and deep learning approaches for AD risk prediction and diagnosis, integrating neuroimaging, genetic, clinical, and lifestyle data. These models generally achieved moderate to high performance (area under the curve [AUC] ≈0.77 - 0.90) and demonstrated improved diagnostic utility compared with conventional assessment methods [79,81-83]. Several studies further emphasized the importance of genetic and hormonal factors in risk stratification [84,85]. In addition, explainable and semisupervised frameworks have enhanced model interpretability and scalability for population-level analysis, facilitating clinically relevant phenotyping and disease monitoring [86,87].

XSL•FO
RenderX

This section describes multimodal model implementation in the UK Biobank. As shown in the analysis and Table 2, UK Biobank data support AD diagnosis and risk prediction, but limitations remain: class imbalance, which may bias training, and the need for external validation to confirm generalizability beyond the UK Biobank cohort.

ADNI Dataset

ADNI provides a rich and diverse collection of demographic information, multimodal data, and clinical assessments. Owing to its comprehensive scope and longitudinal design, it has become one of the most widely adopted benchmark datasets for computer-aided diagnosis of AD. The following studies exemplify its use:

Recent ADNI-based studies have developed a wide range of multimodal and deep learning frameworks integrating neuroimaging, genetic, cognitive, and clinical data for AD diagnosis and MCI-to-AD progression prediction. Attention-based, multitask, ensemble, and time-to-event models have enabled accurate localization of disease-related regions, improved prognostic modeling, and enhanced interpretability through explainable artificial intelligence techniques such as SHAP (Shapley Additive Explanations) and counterfactual analysis [15,75,76,88-92,97]. Several approaches further incorporated RL, semisupervised learning, and data augmentation to improve robustness and generalizability in heterogeneous and imbalanced datasets [98-103,107]. These models typically achieved high diagnostic and prognostic performance (AUC up to ≈0.96), with some demonstrating strong external validation and clinical relevance [93-96,101,104-106,108]. Nevertheless, existing reviews and benchmarking studies have highlighted persistent limitations, including dataset bias, inconsistent evaluation protocols, and limited cross-center validation, underscoring the need for standardized and reproducible multimodal frameworks [119].

While ADNI provides a comprehensive and standardized multimodal resource for AD research and supports robust model performance, several limitations remain. These include class imbalance, underrepresentation of racially diverse populations, and limited external validation, which may bias model training and restrict generalizability across clinical settings.

Self-Collected Datasets

While public datasets such as ADNI provide standardized benchmarks, self-collected datasets enable more flexible acquisition of targeted modalities. Representative studies include the following.

Studies based on self-collected datasets have explored diverse multimodal fusion strategies. EEG- and local field potentials–based models, as well as hybrid MRI–PET–biomarker frameworks, demonstrated high diagnostic and staging accuracy and supported interpretable risk mapping [77,109,111,114,115]. In parallel, behavioral and digital biomarkers derived from handwriting, eye tracking, virtual reality, and motion capture have enabled noninvasive and low-cost screening with strong classification performance [22,78,110,112]. Large-scale real-world health records and hybrid deep learning models further facilitated population-level risk prediction and vascular cognitive impairment assessment, achieving robust AUC values above 0.90 [113,120]. Overall, self-collected datasets have expanded the scope of multimodal AD research by enabling flexible modality integration and novel biomarker discovery, while remaining constrained by limited sample sizes and heterogeneous acquisition protocols.

Self-collected datasets offer distinct advantages, including targeted modality acquisition, novel biomarker discovery, such as microRNA, local field potentials, and handwriting, and enhanced real-world clinical utility. However, self-collected datasets typically endure limited sample sizes, which increases susceptibility to overfitting and compromises generalizability across diverse populations.

Multimodal Linguistic-Based Cognitive Impairment Datasets

Beyond multimodal clinical phenotyping datasets, multimodal linguistic-based cognitive impairment datasets represent an equally vital research resource. These datasets offer a noninvasive and cost-effective methodology for detecting cognitive decline, particularly valuable for identifying early-stage or subtle impairments where traditional neuroimaging or biomarker data may yield inconclusive results. Capturing spontaneous or semistructured speech and language patterns pushes the development of AI in speech data. Recent work is shown in Table 3.

XSL•FO
RenderX

Table . Multimodal linguistic-based cognitive impairment datasets related papers.

Study	Datasets	Model type	Type of task	Modalities	Outcomes	Validation	Results	Limitation
Ilias et al [121], 2023	ADReSS[a] and ADReSSo[b]	Multimodal transformer (BERT[c] and DeiT[d]) with optimal transport	Dementia detection (AD[e] vs non-AD)	Audio (spectrograms) and text (transcripts)	Classification metrics and calibration	Internal: ADReSS or ADReSSo	Accuracy ≈91.25%, F_1-score ≈91.06%; improved calibration vs baselines	Small, curated datasets, English-only, and potential overfitting
Poor et al [122], 2024	I-CONECT[f]	Multimodal cross-transformer with coattention	MCI[g] prediction (MCI vs NC[h])	Audio, text, and vision (facial video)	AUC[i] scores	Internal: cross-validation	Trimodal AUC 85.3%, and outperformed unimodal (60.9%) and bimodal (76.3%) models	Single cohort (I-CONECT), cross-sectional, and complex architecture
Lin and Washington [123], 2024	DementiaBank (Pitt)	Wav2vec (audio) and Word2Vec (text)	Dementia classification	Audio, text, and timestamps	Accuracy and AUROC[j]	Internal: cross-validation	Text augmentation improved accuracy to ≈80% (AUROC 90%), and timestamps added minimal value	Single corpus: timestamps lacked resolution, and a modest sample size
Ortiz-Perez et al [124], 2023	DementiaBank (Pitt)	Multimodal ensemble (CNN[k] and transformer)	Prediction of dementia signs	Audio and text	Classification accuracy	Internal: held-out test sets	Text-only transformer best (accuracy 90.36%) and audio contributed less than text	Single English dataset, broad diagnosis category, and task constrained to picture description
Ilias and Askounis [125], 2022	ADReSS (DementiaBank)	Transformer (BERT) and Siamese Network	AD identification and severity estimation	Text (transcripts)	Accuracy and interpretability (LIME[l])	Internal: cross-validation	Single-task accuracy 87.50%, multitask accuracy 86.25%, and distinct linguistic patterns identified	Small dataset, text only, MMSE[m] treated as categorical, and no acoustic information
Wen et al [126], 2023	DementiaBank (Pitt)	Transformer and causal counterfactual XAI[n]	AD detection	Text (part-of-speech tag features)	Accuracy; F_1-score; feature importance	Internal: cross-validation	Accuracy 92.2%, F_1-score 0.955, identified 12 key part-of-speech features linked to AD	Text only (part-of-speech), reliance on tagging accuracy, and no acoustic or imaging data
Chen et al [127], 2023	DementiaBank (Pitt)	SpeechFormer + + (hierarchical transformer)	Paralinguistic AD detection	Audio (acoustic features)	Accuracy; F_1-score	Internal: held-out test sets	Outperformed standard transformers and CNN/RNN[o] baselines and SOTA[p] performance	Single corpus, complex computation, audio only, and no cross-lingual evaluation

XSL•FO
RenderX

Study	Datasets	Model type	Type of task	Modalities	Outcomes	Validation	Results	Limitation
Zheng et al [128], 2022	DementiaBank (Pitt)	N-gram, AWD-LSTM[q], or neural models	Dementia detection	Text (context words, stop words, and part-of-speech)	Classification accuracy	Internal: held-out test data	Combined model (vocabulary and grammar) accuracy 81.54%, and grammar contributes comparably to context	Specific to task or language, and moderate performance vs multimodal approaches
Nambiar et al [129], 2022	DementiaBank (Pitt)	Deep Classifiers (BERT/ALBERT[r] + BiLSTM[s])	Early dementia detection	Text (transcripts)	Accuracy; F_1-score	Internal: train and test splits	BERT + BiLSTM accuracy 0.812; ALBERT + BiLSTM F_1-score 0.81; contextual embeddings superior	Text only; reliance on manual transcripts; single dataset
Priyadarshinee et al [130], 2023	ADReSSo-2021	ML[t] classifiers (SVM[u], RF[v], and NN[w])	AD detection	Audio and text (transcripts)	Classification accuracy	Internal: held-out test set	Text features (accuracy 88.7%) outperformed audio, and file-level features were superior to frame-level	Benchmarking context, single task, and single language
Liu et al [131], 2023	ADReSS, ADReSSo, and the local Chinese dataset	Ensemble ML (VAD[x] pause and acoustic)	AD detection	Audio (acoustic and VAD pause features)	Accuracy	Internal: cross-validation; cross-lingual (Chinese)	Ensemble improved accuracy by ≈8% on ADReSS, and accuracy 80% on the local Chinese dataset	Small local dataset (n=10), handcrafted features, and ensemble complexity
Shah et al [23], 2023	ADReSS-M	Logistic regression and SVR	Cross-lingual AD detection; MMSE regression	Audio (duration, pause, and intelligibility) and metadata	Accuracy and RMSE[y]	External: Greek test set	English cross-validation accuracy 74.7%, Greek Test accuracy 69.57%, and MMSE RMSE 4.77 (Greek)	Small Greek sample, modest accuracy, and simple ML models vs deep learning
Mahajan and Baths [132], 2021	ADReSS	Bimodal framework (CNN-LSTM[z] and Speech-GRU[aa])	AD detection	Audio and text	Classification accuracy	Internal: cross-validation	Bimodal enriched model improved performance by ≈6.25% over acoustic baselines	Small dataset, potential overfitting, and single task (picture description)
Mei et al [133], 2023	ADReSS-M	Bilingual wav2vec 2.0 + XGBoost[ab]	Cross-lingual AD detection and MMSE prediction	Audio (acoustic, silence, and low-frequency bands)	Accuracy and RMSE	External: Greek test set	Accuracy 73.9% (Greek), MMSE RMSE 4.610, and low-frequency speech aided transfer	Very small Greek sample, speech-only, and challenge context
Meerza et al [134], 2022	ADReSS	FL[ac] (LSTM[ak] and feed-forward)	Privacy-preserving AD diagnosis	Audio (Mel-frequency and pause features)	Accuracy and fairness metrics	Internal: simulated FL clients	FL accuracy close to the centralized baseline, and q-FedAvg improved fairness	Simulated clients, single dataset, and relies on feature extraction

XSL•FO
RenderX

Study	Datasets	Model type	Type of task	Modalities	Outcomes	Validation	Results	Limitation
Chen et al [135], 2023	ADReSS-M	SVM or NN on pretrained features	Cross-lingual AD detection	Audio (paralinguistic and XLSR-53[ae]), and text (ASR[af]	Accuracy and RMSE	External: Greek test set	Accuracy 69.6% (Greek), RMSE 4.788, and paralinguistic features transferable	Performance below monolingual systems and reliance on ASR quality
Ilias et al [121], 2023	ADReSS	Multimodal transformer (ViT[ag], BERT, and GMU[ah])	AD detection	Audio (spectrograms) and text	Accuracy and F_1-score	Internal: cross-validation	High eighties or low nineties accuracy, ViT is best for acoustic, and fusion surpassed SOTA	Small dataset, binary classification focus, and external generalization untested
Tamm et al [136], 2023	ADReSS-M[ai]	Sequence models (transfer learning)	Cross-lingual AD detection and MMSE	Audio features and demographics	Accuracy and RMSE	External: Greek test set	Accuracy 82.6% (Greek), RMSE 4.345, and ranked second in the challenge	Small Greek sample, acoustic only, and transfer limited to English-Greek
Woszczyk et al [137], 2022	ADReSS	Transformers vs traditional ML	AD detection	Audio and text	Classification accuracy	Internal: held-out test data	Data augmentation improved performance and was comparable to SOTA	Augmentations tuned for ADReSS and a single speech task

XSL•FO
RenderX

Study	Datasets	Model type	Type of task	Modalities	Outcomes	Validation	Results	Limitation
Jin et al [138], 2023	ADReSS-M	CONSEN[aj] ensemble (acoustic and disfluency)	Multilingual AD detection and MMSE	Audio (acoustic embeddings and disfluency)	Accuracy and RMSE	External: Greek test set	First place in the challenge, accuracy 86.69% (Greek), and RMSE 3.727	Challenge dataset, ensemble complexity, and reliance on diarization quality

[a]ADReSS: Alzheimer Dementia Recognition Through Spontaneous Speech.
[b]ADReSSo: Alzheimer's Dementia Recognition Through Spontaneous Speech only.
[c]BERT: Bidirectional Encoder Representations From Transformers.
[d]DeiT: Data-Efficient Image Transformers.
[e]AD: Alzheimer disease.
[f]I-CONECT: Identifying Cognition in the Elderly Through Conversational Engagement.
[g]MCI: mild cognitive impairment.
[h]NC: normal control.
[i]AUC: area under the curve.
[j]AUROC: area under the receiver operating characteristic curve.
[k]CNN: convolutional neural network.
[l]LIME: Local Interpretable Model-Agnostic Explanations.
[m]MMSE: Mini-Mental State Examination.
[n]XAI: explainable artificial intelligence.
[o]RNN: recurrent neural network.
[p]SOTA: state of the art.
[q]AWD-LSTM: Average stochastic gradient descent weight-dropped long short-term memory
[r]ALBERT: A Lite Bidirectional Encoder Representations From Transformers.
[s]BiLSTM: bidirectional long short-term memory.
[t]ML: machine learning.
[u]SVM: support vector machine.
[v]RF: random forest.
[w]NN: neural network.
[x]VAD: voice activity detection.
[y]RMSE: root mean square error.
[z]CNN-LSTM: convolutional neural network long short-term memory.
[aa]Speech-GRU: Speech Gated Recurrent Unit.
[ab]XGBoost: Extreme Gradient Boosting.
[ac]FL: federated learning.
[ad]LSTM: long short-term memory.
[ae]XLSR-53: cross-lingual speech representation-version 53
[af]ASR: automatic speech recognition.
[ag]ViT: vision transformer.
[ah]GMU: gated multimodal unit.
[ai]ADReSS-M: Alzheimer Dementia Recognition through Spontaneous Speech – Multimodal.
[aj]CONSEN: complementary and simultaneous ensemble.

Recent studies have shown that multimodal fusion of speech and text using transformer-based architectures substantially improves AD detection performance, with F_1-scores above 0.90 on ADReSS and ADReSSo (Alzheimer's Dementia Recognition Through Spontaneous Speech 2021 Challenge) datasets [121,132,139]. Linguistic feature engineering and interpretable language models further enhanced classification accuracy, achieving up to 92.2% accuracy and F_1-scores of 0.955 using compact part-of-speech features [124-126,128,130]. Cross-lingual approaches based on language-agnostic and transfer learning methods enabled moderate generalization, with accuracies ranging from 69% to 73.9% in English-Greek transfer settings [23,127,133,136]. To support real-world deployment, lightweight and hierarchical models achieved around 80% accuracy with reduced computational cost [131,135]. In addition, data augmentation and ensemble strategies improved robustness in low-resource scenarios, yielding F_1-score gains of 5% - 7% and competitive challenge performance (accuracy 86.69%) [123,137,138].

XSL•FO
RenderX

Summarization Based on All Multimodal Datasets and Quantitative Analysis

Table 2 and Table 3 summarize the recent state-of-the-art models across the 2 major types of multimodal datasets, extracted according to the Cochrane Handbook. Full QUADAS-2 forms are available in Multimedia Appendix 5. Based on these results, the following quantitative synthesis compares performance trends across all multimodal datasets. Across the 4 major dataset categories, modality choices and model performance show clear dataset-dependent patterns as shown in Table 4. UK Biobank studies mainly combine MRI, clinical variables, and genetic features, with 2 diagnosis studies reporting an average accuracy of 71.4% (SD 5.2%) and 4 risk-prediction studies reaching an average AUC of 0.84 (SD 0.056). ADNI studies use the most comprehensive modality integrations, with 3 diagnosis studies averaging 92.5% (SD 3.8%) accuracy, 3 MCI-conversion studies achieving a mean AUC of 0.922 (SD 0.045), and risk-prediction studies reaching an average AUC of 0.81 (SD 0.06); these tasks collectively achieve the strongest results, with fusion models frequently reporting AUC values above 0.95. DementiaBank studies differ fundamentally by focusing on speech- and language-based modalities; 9 diagnosis studies report an average AUC of 0.813 (SD 0.042), and 5 cross-lingual AD-detection studies show a mean accuracy of 77% (SD 6.5%), where transformer architectures consistently outperform classical approaches, with models such as BERT + DeiT (Data-Efficient Image Transformers), BERT + ViT (vision transformer), and RoBERTa + (Robustly Optimized Bidirectional Encoder Representations From Transformers Approach) DNN (deep neural network) showing F_1-scores exceeding 0.90. Self-collected datasets are typically smaller and more heterogeneous; 3 diagnosis studies report an average accuracy of 96% (SD 2.4%), and lightweight models such as EEGNet or ViT-based hybrids demonstrate strong predictive capacity when applied to EEG or structural MRI.

XSL•FO
RenderX

Table . Summary of representative modality combinations and top-performing models in multimodal AI[a]-aided AD[b] diagnosis.

Dataset and task	Counts	Average performance	Best performance modalities	Related article
UK Biobank				
Diagnosis	2	Accuracy=71.4%	Retinal fundus images	[79,81-87,140]
Risk prediction	4	AUC[c]=84%	Clinical, biological assays, cognitive tests, and physical measures	[79,81-87,140]
Other	3	N/A[d]	Multimodal MRI[e] (T1, T2, MRI, etc)	[79,81-87,140]
ADNI[f]				
Diagnosis	3	Accuracy=92.5%	Structural MRI features and neuropsychological tests	[15,75,76,79,80,89-92,94-108,119,141]
MCI[g] conversion	3	AUC=92.2%	Structural MRI, clinical variables, and genetics (SNP[h])	[15,75,76,79,80,89-92,94-108,119,141]
MMSE[i] regression	2	No integration	Whole-brain T1-weighted MRI and clinical scores	[15,75,76,79,80,89-92,94-108,119,141]
Risk prediction	7	AUC=81%	MRI, PET[j], clinical, and cognitive	[15,75,76,79,80,89-92,94-108,119,141]
Other	13	N/A	N/A	[15,75,76,79,80,89-92,94-108,119,141]
Dementia bank				
Diagnosis	9	AUC=81.3%	Text transcripts → part-of-speech feature vectors	Table 3
Cross-lingual AD detection	5	Accuracy=77%	Multimodal acoustic fusion	Table 3
Other	6	N/A	N/A	Table 3
Self-collected datasets				
Diagnosis	3	Accuracy=96%	MRI, PET, clinical, and genotype	[22,77,78,106,109-115,117,120,142]
Other	6	No integration	Different task	[22,77,78,106,109-115,117,120,142]

[a]AI: artificial intelligence.
[b]AD: Alzheimer disease.
[c]AUC: area under the curve.
[d]N/A: not available.
[e]MRI: magnetic resonance imaging.
[f]ADNI: Alzheimer Disease Neuroimaging Initiative.
[g]MCI: mild cognitive impairment.
[h]SNP: single-nucleotide polymorphism.
[i]MMSE: Mini-Mental State Examination.
[j]PET: positron emission tomography.

To interpret these results and limit metric inflation, note that purely internal cross-validation tends to overestimate performance: AUC is typically ≈5 - 15 points higher than with external validation. Small or tightly controlled datasets also report accuracies ≈10% - 20% above those in large, heterogeneous cohorts. Severe class imbalance can further raise accuracy while lowering F_1-score or sensitivity; without correction, imbalance may inflate results by ≈5% - 12%. Cross-sectional models often score higher in single-timepoint evaluations, whereas longitudinal designs usually yield lower but more stable estimates, which are more informative for follow-up and clinical use.

These findings should be interpreted in light of substantial heterogeneity and risk of bias. Variation in sample composition, task definitions, and evaluation procedures across datasets limits direct comparison of performance metrics. QUADAS-2 also indicated frequent unclear and high-risk in-patient selection, reference standards, and flow or timing, especially in studies using only internal validation or selected samples. Reported metrics, therefore, likely represent upper-bound estimates rather than expected real-world performance, and apparent gains often

XSL•FO
RenderX

reflect dataset-specific effects rather than generalizable model superiority.

Overall, the evidence shows that modality effectiveness varies substantially across datasets, transformer models deliver the highest gains in speech-language tasks, and large clinical phenotyping datasets such as UK Biobank and ADNI still rely mainly on traditional machine-learning or custom fusion frameworks rather than modern cross-modal transformers. This gap highlights an opportunity to develop transformer-based multimodal integration approaches tailored to large, heterogeneous clinical datasets.

Multimodal Fusion Taxonomy

A structured multimodal fusion taxonomy clarifies the performance of different integration strategies across datasets (Tables 2 and 3). A total of 4 main paradigms are commonly used: early, intermediate, late, and attention- or graph-based fusion.

Early fusion concatenates low-level features and performs well for aligned modalities such as MRI + PET, often achieving AUC>0.95 in ADNI studies, but is sensitive to missing data and feature-scale heterogeneity. Intermediate fusion combines latent representations from modality-specific encoders and is effective for heterogeneous inputs such as MRI + speech or EEG + clinical data, as demonstrated by high performance in ADReSS-based models, although it may be unstable in small datasets. Late fusion aggregates model outputs and is robust to missing modalities, performing well in large datasets such as the UK Biobank, but underuses fine-grained cross-modal interactions.

Across paradigms, limited modality availability and high acquisition costs remain key challenges, underscoring the need for adaptive and clinically feasible fusion strategies.

Discussion

Principal Findings

This review synthesized multimodal AI studies for AD across diverse dataset families, including clinical phenotyping and cognitive-linguistic datasets. Multimodal fusion generally outperformed unimodal baselines, but the gain is dataset-dependent and should be interpreted cautiously. Strong performance in curated cohorts and constrained speech benchmarks may not generalize to population-based or multicenter settings. QUADAS-2 also indicated frequent risk of bias and unclear reporting across domains, likely inflating metrics and limiting comparability. Accordingly, headline accuracy and AUC should be treated as upper-bound estimates unless supported by external validation and transparent reporting.

Challenges and Future Directions

In recent years, multimodal models have demonstrated remarkable potential in computer-aided diagnosis and risk prediction for AD. While these methods have achieved significant successes, several challenges remain that warrant careful examination. In this section, this systematic review summarizes the common limitations identified in existing studies and proposes directions for future research to advance the field.

Clinical and Translational Implications

Multimodal AI could support AD diagnosis through several clinical pathways. In memory clinics, models combining MRI, cognitive scores, and blood biomarkers could triage referrals, prioritizing patients for specialist review or PET. In general practice, speech-based and routine clinical-feature models could be embedded in consultations to flag early cognitive change. In radiology, MRI-clinical fusion could act as a second reader, reducing interobserver variability and supporting less experienced clinicians. Where imaging or specialist access is limited, speech, digital questionnaires, and basic clinical data could enable telemedicine-based screening and follow-up. At the population level, these models could support risk stratification and targeted monitoring. To enable real-world deployment, research should prioritize external multicenter validation, integration with electronic health records, and evaluation of regulatory feasibility, cost-effectiveness, and clinical impact.

Ethical and Regulatory Implications

Deploying multimodal AI for AD diagnosis requires ethical and regulatory safeguards. As datasets often combine imaging, clinical records, genomics, and speech, they fall under strict privacy regimes (eg, General Data Protection Regulation in the European Union; HIPAA [Health Insurance Portability and Accountability Act] in the United States), requiring explicit consent, data minimization, and secure handling, with added complexity for sensitive modalities such as speech and genomic data. Clinical deployment is also shaped by medical-AI governance frameworks (eg, the European Union AI Act, Food and Drug Administration Software as a Medical Device guidance, and UK Medicines and Healthcare products Regulatory Agency Good Machine Learning Practice), which emphasize transparency, risk management, and postdeployment monitoring. Fairness is essential because demographic imbalance can yield uneven performance across age, ethnicity, and language groups. Interpretability (eg, imaging attention maps and linguistic saliency) supports clinical accountability and aligns with explainability expectations. Future work should incorporate privacy-preserving methods, bias audits, and regulatory-aligned validation pipelines to enable responsible clinical integration.

Data Privacy and Data-Sharing Constraints

Access to multimodal AD data remains severely restricted by privacy regulations and ethical constraints, which limit data sharing and external validation. This restricts the sharing and usage of comprehensive datasets needed for robust external validation and generalizability.

Federated learning (FL) provides a technically viable privacy-preserving solution; however, differences in data formats and institutional infrastructures still impede its large-scale deployment. For instance, Meerza et al [134] pioneered FL for AD speech diagnosis using mel-frequency cepstral coefficients and pause features, maintaining model performance while ensuring privacy through

XSL•FO
RenderX

q-FedAvg/q-FedSGD optimization. Nambiar [129] validated an ALBERT (A Lite Bidirectional Encoder Representations From Transformers) + BiLSTM (bidirectional long short-term memory) hybrid model on the ADReSS dataset, achieving strong performance without compromising data confidentiality. In parallel, multi-institutional collaborations leveraging publicly available datasets such as ADNI, UK Biobank, and OASIS have enabled richer external validation while adhering to rigorous privacy standards [15,79,88,95,100,139,140].

Despite encouraging results, FL still lacks harmonized protocols and interoperable platforms. This limits cross-center reproducibility and weakens clinical credibility. International collaboration also remains constrained by regulatory differences. Future work should prioritize unified federated frameworks with standardized protocols and privacy-preserving methods to enable secure global data collaboration [143,144].

As most datasets lack fully matched modalities per participant, multimodal fusion often relies on representation- or population-level integration rather than early fusion. Early fusion requires paired samples and is therefore infeasible across datasets. By contrast, late fusion and embedding-level integration can train unimodal models separately and combine them via meta-learners, cross-modal transformers, or probabilistic ensembles. Domain adaptation, transfer learning, and harmonization can also combine heterogeneous cohorts at the population level to improve generalizability. A standardized benchmark could further support this by defining shared preprocessing, label taxonomies, and evaluation metrics, enabling meaningful comparison or representation-stage fusion even without subject-level pairing.

Data Imbalance

Severe class imbalance remains a major obstacle, biasing training toward the majority class and inflating accuracy while masking low sensitivity to early disease. In addition, datasets such as the UK Biobank are dominated by White European ancestry, limiting generalizability across racially and ethnically diverse populations. Addressing this requires both technical mitigation and proactive recruitment of underrepresented groups so models better reflect population heterogeneity.

Researchers have applied data-level interventions such as generative adversarial network–based augmentation, diffusion models, and resampling [123,137,138,145,146]; algorithm-level solutions, including cost-sensitive, loss-focused, ensemble, and class-weighted training schemes [147-152]; and evaluation-focused remedies [153] have been developed to mitigate biases.

Current methods frequently introduce new challenges, such as overfitting or inadequate performance in minority classes. Moreover, efforts to increase diversity remain inadequate. Future directions should focus on novel adaptive resampling methods, generative methods for synthetic minority data creation, and dedicated efforts to include and characterize underrepresented populations to ensure equitable and robust clinical applicability across diverse populations.

Lack of Standardized and Longitudinal Data

Differences in acquisition protocols and diagnostic criteria across datasets limit comparability of imaging, cognitive, and biomarker outcomes. Longitudinal evidence is also constrained: even in relatively standardized resources such as ADNI, limited long-term follow-up hampers modeling the temporal dynamics of disease progression.

Future work should standardize key acquisition elements and diagnostic criteria across longitudinal studies and strengthen coordination across institutions. Building on this, a multimodal benchmark spanning imaging, clinical, biomarker, behavioral, and linguistic modalities would enable cross-dataset validation, improve comparability, and support reproducible evaluation of new models. These steps would strengthen temporal modeling and provide more reliable evidence for clinical translation.

Dataset-Specific Limitations

Data imbalance is prevalent across many AD datasets, but the nature of this issue varies substantially between cohorts. This review, therefore, outlines the dataset-specific limitations of commonly used AD cohorts and corpora.

ADNI participants are generally healthier, with fewer comorbidities and a restricted age range (55 - 90 y), limiting representativeness. Protocol differences across centers and evolving diagnostic standards introduce heterogeneity, while frequent reliance on subsets hampers comparability [154].

Of UK Biobank, dementia outcomes are derived mainly from health records, leading to potential misclassification and delayed ascertainment. Participants show strong volunteer bias, and PET or cerebrospinal fluid biomarkers are limited to a small subset, constraining multimodal analyses [155].

OASIS provides open neuroimaging data but with relatively small AD/MCI sample sizes and inconsistent modality coverage. Limited longitudinal depth and cross-scanner variability further reduce reproducibility [156].

Of NACC, data are aggregated from multiple centers with heterogeneous recruitment and diagnostic protocols, making harmonization challenging. The cohort is clinic-based rather than population-representative, and missing biomarker modalities are common [157].

Although high-quality, Australian Imaging, Biomarkers and Lifestyle Study is smaller than ADNI and NACC and is often used only for validation. Regional recruitment and protocol differences reduce ethnic diversity and cross-cohort comparability [158].

Of Pitt Corpus, this is the most widely used speech dataset, but remains small and imbalanced. Tasks are constrained, limiting ecological validity, and cross-linguistic generalizability is poor [159].

Of the ADReSS series, the ADReSS benchmarks provide standardized speech corpora but are modest in size and restricted to English. Narrow task design and small training partitions raise concerns of overfitting and limited external validity [18].

XSL•FO
RenderX

Of self-collected cohorts, locally collected datasets often involve small, single-site samples with heterogeneous acquisition protocols. Missing modalities, limited follow-up, and selection bias further restrict their generalizability [153].

Dataset challenges are compounded by unrepresentative cohorts, incomplete modalities, and poor cross-center consistency, limiting model robustness and cross-dataset generalization in AD diagnosis. Future work should improve data coordination and standardization, enable more practical sharing mechanisms, and adopt cross-cohort validation where feasible. Strengthening data quality and access is essential for translating multimodal AI methods into clinical use.

Model Interpretability and Explainability

A major limitation of multimodal ML models in clinical AD diagnosis is limited interpretability and transparency. Many high-performing models provide insufficient insight into their decision processes, which can hinder clinical adoption and reduce confidence among end users.

Efforts that have been made toward model interpretability include designing inherently transparent models. For example, some studies demonstrate emerging explainability strategies, including hybrid neuro-symbolic models [160] that generate interpretable reports and post hoc methods such as SHAP, LIME (Local Interpretable Model-Agnostic Explanations), gradient-based saliency, and graph-masking techniques [161,162], which collectively enhance transparency in multimodal AD diagnosis.

Current interpretability methods often fail to produce explanations that clinicians can use reliably. Future work should prioritize clinically grounded explainability, including interactive visualizations and concise workflow-aligned natural-language summaries. Hybrid designs that combine deep learning with structured reasoning can further improve transparency by making decision logic explicit. For deployment, models should also report prediction uncertainty and demonstrate compatibility with clinical systems and regulatory requirements.

Beyond technical advances, incorporating patient and public involvement can improve multimodal AI development for AD. Patients and caregivers can help shape evaluation and result communication, not just act as end users, aligning explanations with patient priorities and addressing transparency and fairness. Engaging patient and public involvement earlier in model design may therefore support more interpretable and clinically usable diagnostic tools.

Heterogeneous Multiview Learning Problem

Integrating data across studies is challenging because single datasets rarely cover all modalities, forcing combinations such as ADNI with UK Biobank. However, differences in cohorts, imaging protocols, and cognitive assessment frameworks create substantial heterogeneity that limits direct pooling and comparability.

This heterogeneity hinders building unified models that generalize across nonoverlapping cohorts, so single-dataset models often fail out of domain. Platform-agnostic methods that tolerate missing or inconsistent modalities are therefore needed. Proposed solutions include shared latent-space learning [163], multibranch networks [164], and mixture-of-experts architectures [165] to support partial fusion and cross-dataset adaptation, but most still assume strong cross-domain alignment or require substantial retraining under domain shift.

Despite recent progress, multimodal methods often assume strict cross-domain alignment and require extensive retraining under domain shift or missing modalities. Future work should develop robust, platform-agnostic frameworks that adapt to changing modality availability and distribution shifts with minimal performance loss and advance representation learning to derive stable joint embeddings from heterogeneous data.

Uncertainty Quantification and Clinical Applicability

Although multimodal AD models have advanced, most studies still omit uncertainty quantification (eg, confidence or prediction intervals). Models typically provide deterministic outputs without communicating reliability, despite clinicians relying on uncertainty to guide management and treatment decisions. Future work should embed uncertainty metrics into diagnostic models to better align with clinical needs and improve interpretability, reliability, and real-world adoption.

Risks of Data Leakage in Multimodal AI Modeling

Another limitation is data leakage, which can inflate performance. Common forms include subject-level leakage (samples from the same participant in both training and test sets), patch-level overlap in MRI slice and patch models, and transcript or utterance-level leakage in speech datasets when multiple segments come from 1 individual. Many studies did not report whether participant-independent splits were enforced. Clearer reporting of partitioning and rigorous participant-level cross-validation are therefore essential to ensure real-world generalizability.

Conclusions

This review synthesizes evidence on multimodal AI approaches for AD across clinical, neuroimaging, genetic, and linguistic data, systematically comparing modeling strategies, validation practices, and performance trends across heterogeneous datasets. In contrast to prior modality-specific reviews, the findings show that multimodal models generally outperform unimodal approaches, although performance varies substantially with dataset characteristics, modality availability, and cross-source alignment. High accuracies are often reported in curated or internally validated cohorts, whereas population-based and externally validated studies yield more modest but clinically realistic results, reflecting substantial heterogeneity and risk of bias.

Despite these limitations, the evidence demonstrates that multimodal AI captures complementary biological and behavioral signals relevant to AD, offering clear advantages for diagnosis and risk prediction. Transformer-based architectures and speech- or behavior-derived modalities show promise for scalable and noninvasive early detection. However, meaningful clinical translation will require harmonized benchmarking, transparent reporting, and rigorous external validation. Overall,

XSL•FO
RenderX

this review advances the field by contextualizing performance gains within their methodological constraints and by outlining practical directions for developing robust, interpretable, and generalizable multimodal AI systems. These insights support the responsible integration of AI into real-world dementia screening, risk prediction, and early intervention strategies.

Acknowledgments

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT (Generative Artificial Intelligence for Digital Twins) taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing. The GenAI tool used was ChatGPT-5.2. Responsibility for the final manuscript lies entirely with the authors. Declaration submitted by: JMIR Publications. We used ChatGPT-5.2 (OpenAI) to conduct a grammatical review of the abstract and conclusion sections.

Funding

This work received no specific financial or nonfinancial support. No funder or sponsor had any role in the design of the review; data collection, analysis, or interpretation; writing of this paper; or the decision to submit for publication.

Data Availability

This systematic review did not generate any new datasets. All data extracted and analyzed in this systematic review were obtained from publicly available publications included in the review. No additional unpublished or proprietary data were used.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full database search strategies for PubMed, Scopus, IEEE Xplore, and ACM Digital Library, including complete Boolean queries, search fields, filters, and publication date limits used for study identification.
[DOCX File, 999 KB - jmir_v28i1e85414_app1.docx]

Multimedia Appendix 2
Performance evaluation for AD diagnosis. AD: Alzheimer disease.
[DOCX File, 22 KB - jmir_v28i1e85414_app2.docx]

Multimedia Appendix 3
Complete QUADAS-2 risk-of-bias assessments for all included studies, summarizing judgments across patient selection, index test, reference standard, and flow or timing, with detailed study-level ratings. QUADAS-2: Revised Quality Assessment of Diagnostic Accuracy Studies Tool.
[DOCX File, 38 KB - jmir_v28i1e85414_app3.docx]

Multimedia Appendix 4
Overview of traditional machine-learning models applied in Alzheimer disease research, including SVM, decision trees, HMMs, KNN, logistic regression, GMMs, and foundational CNN or RL descriptions, with methodological principles and limitations. CNN: convolutional neural network; GMM: gaussian mixture models; HMM: hidden markov model ; KNN: k-nearest neighbors; RL: reinforcement learning; SVM: support vector machine.
[DOCX File, 42 KB - jmir_v28i1e85414_app4.docx]

Multimedia Appendix 5
Cochrane Handbook 5.3.3–aligned data-extraction tables summarizing study design, datasets, participants, modalities, preprocessing, model architectures, validation schemes, outcomes, and limitations for all included studies.
[DOCX File, 185 KB - jmir_v28i1e85414_app5.docx]

Checklist 1
Completed PRISMA 2020, PRISMA-S checklist, and PRISMA expanded checklist specifying reporting locations for all required items, including eligibility criteria, search methods, extraction procedures, bias assessments, and synthesis reporting. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for literature searches.
[PDF File, 11026 KB - jmir_v28i1e85414_app6.pdf]

XSL•FO
RenderX

References

1. Scheltens P, De Strooper B, Kivipelto M, et al. Alzheimer's disease. Lancet 2021 Apr 24;397(10284):1577-1590. [doi: 10.1016/S0140-6736(20)32205-4] [Medline: 33667416]
2. 2024 Alzheimer's disease facts and figures. Alzheimers Dement 2024 May;20(5):3708-3821. [doi: 10.1002/alz.13809]
3. World alzheimer report 2024. : Alzheimer's Disease International; 2024 URL: https://www.alzint.org/resource/world-alzheimer-report-2024/ [accessed 2026-02-10]
4. Kaštelan S, Gverović Antunica A, Puzović V, et al. Non-invasive retinal biomarkers for early diagnosis of Alzheimer's disease. Biomedicines 2025 Jan 24;13(2):283. [doi: 10.3390/biomedicines13020283] [Medline: 40002697]
5. Castellano G, Esposito A, Lella E, Montanaro G, Vessio G. Automated detection of Alzheimer's disease: a multi-modal approach with 3D MRI and amyloid PET. Sci Rep 2024 Mar 3;14(1):5210. [doi: 10.1038/s41598-024-56001-9] [Medline: 38433282]
6. Bi Y, Abrol A, Fu Z, Calhoun VD. A multimodal vision transformer for interpretable fusion of functional and structural neuroimaging data. Hum Brain Mapp 2024 Dec 1;45(17):e26783. [doi: 10.1002/hbm.26783] [Medline: 39600159]
7. Yu Q, Ma Q, Da L, et al. A transformer-based unified multimodal framework for Alzheimer's disease assessment. Comput Biol Med 2024 Sep;180:108979. [doi: 10.1016/j.compbiomed.2024.108979] [Medline: 39098237]
8. Leng Y, He Y, Amini S, et al. A GPT-4o-powered framework for identifying cognitive impairment stages in electronic health records. npj Digit Med 2025 Jul 3;8(1):401. [doi: 10.1038/s41746-025-01834-5]
9. Balabin H, Tamm B, Spruyt L, et al. Natural language processing-based classification of early Alzheimer's disease from connected speech. Alzheimer's Dement 2025 Feb;21(2):e14530. [doi: 10.1002/alz.14530] [Medline: 39868827]
10. Yang X, Hong K, Zhang D, Wang K. Early diagnosis of Alzheimer's disease based on multi-attention mechanism. In: Fati SM, editor. PLOS ONE 2024;19(9):e0310966. [doi: 10.1371/journal.pone.0310966]
11. Wijeratne PA, Alexander DC. Learning transition times in event sequences: the event-based hidden markov model of disease progression. Inf Process Med Imaging 2021 Jun;12729(14):583-595. [doi: 10.1007/978-3-030-78191-0_45]
12. Huh YJ, Park JH, Kim YJ, Kim KG. Ensemble learning-based Alzheimer's disease classification using electroencephalogram signals and clock drawing test images. Sensors (Basel) 2025 May 2;25(9):2881. [doi: 10.3390/s25092881] [Medline: 40363322]
13. Karasu E, Baytaş İ. Conversion-aware forecasting of Alzheimer's disease via featurewise attention. Pattern Anal Applic 2025 Jun;28(2):64. [doi: 10.1007/s10044-025-01447-4]
14. Xiao X, Li Y, Wu Q, et al. Development and validation of a novel predictive model for dementia risk in middle-aged and elderly depression individuals: a large and longitudinal machine learning cohort study. Alz Res Therapy 2025 May 13;17(1):103. [doi: 10.1186/s13195-025-01750-6]
15. Qiu S, Miller MI, Joshi PS, et al. Multimodal deep learning for Alzheimer's disease dementia assessment. Nat Commun 2022 Jun 20;13(1):3404. [doi: 10.1038/s41467-022-31037-5] [Medline: 35725739]
16. Chakravarthi BA, Shivakanth G. Integrating multimodal AI techniques and MRI preprocessing for enhanced diagnosis of Alzheimer's disease: clinical applications and research horizons. IEEE Access 2025;13:63519-63531. [doi: 10.1109/ACCESS.2025.3557533]
17. Elazab A, Wang C, Abdelaziz M, et al. Alzheimer's disease diagnosis from single and multimodal data using machine and deep learning models: achievements and future directions. Expert Syst Appl 2024 Dec;255:124780. [doi: 10.1016/j.eswa.2024.124780]
18. Ding K, Chetty M, Noori Hoshyar A, Bhattacharya T, Klein B. Speech based detection of Alzheimer's disease: a survey of AI techniques, datasets and challenges. Artif Intell Rev 2024 Oct 12;57(12):325. [doi: 10.1007/s10462-024-10961-6]
19. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
20. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
21. Li T, Higgins JP, Deeks JJ. Collecting data. In: Cochrane Handbook for Systematic Reviews of Interventions 2019:109-141. [doi: 10.1002/9781119536604]
22. Park B, Kim Y, Park J, et al. Integrating biomarkers from virtual reality and magnetic resonance imaging for the early detection of mild cognitive impairment using a multimodal learning approach: validation study. J Med Internet Res 2024 Apr 17;26:e54538. [doi: 10.2196/54538] [Medline: 38631021]
23. Shah Z, Qi SA, Wang F, et al. Exploring language-agnostic speech representations using domain knowledge for detecting Alzheimer's dementia. ICASSP 2023 - 2023 IEEE Int Conf Acoustics, Speech Signal Process (ICASSP) 2023:1-2. [doi: 10.1109/ICASSP49357.2023.10095593]
24. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273-297. [doi: 10.1007/BF00994018]
25. Sharma A, Kaur S, Memon N, Jainul Fathima A, Ray S, Bhatt MW. Alzheimer's patients detection using support vector machine (SVM) with quantitative analysis. Neurosci Inf 2021 Nov;1(3):100012. [doi: 10.1016/j.neuri.2021.100012]

XSL•FO
RenderX

26. Gao X, Liu H, Shi F, Shen D, Liu M. Brain status transferring generative adversarial network for decoding individualized atrophy in Alzheimer's disease. IEEE J Biomed Health Inform 2023 Oct;27(10):4961-4970. [doi: 10.1109/JBHI.2023.3304388] [Medline: 37607152]
27. Lazli L. Improved Alzheimer disease diagnosis with a machine learning approach and neuroimaging: case study development. JMIRx Med 2025 Apr 21;6:e60866. [doi: 10.2196/60866] [Medline: 40257754]
28. Hossain F, Halder RK, Uddin MN. An integrated machine learning based adaptive error minimization framework for Alzheimer's stage identification. Intell-Based Med 2025;11:100243. [doi: 10.1016/j.ibmed.2025.100243]
29. Fulkar B, Dhale T, Pacharaney U, Deshmukh S. Early detection of chronic diseases using machine and deep learning algorithms. 2025 4th Int Conf Sentiment Anal Deep Learn (ICSADL) 2025:1656-1661. [doi: 10.1109/ICSADL65848.2025.10933005]
30. Sathiya A, Basha CH, S V, Sharmila P JJ, S P, Indhumathi R. Enhancing Alzheimer's disease detection using optimized attribute selection and random forest classifier for improved accuracy. 2025 Int Conf Visual Anal Data Visualization (ICVADV) 2025:1174-1179. [doi: 10.1109/ICVADV63329.2025.10961844]
31. Saleh AW, Gupta G, Khan SB, Alkhaldi NA, Verma A. An Alzheimer's disease classification model using transfer learning Densenet with embedded healthcare decision support system. Decis Anal J 2023 Dec;9:100348. [doi: 10.1016/j.dajour.2023.100348]
32. Baucum M, Khojandi A, Papamarkou T. Hidden markov models as recurrent neural networks: an application to Alzheimer's disease. 2021 IEEE 21st Int Conf Bioinf Bioeng (BIBE) 2021:1-6. [doi: 10.1109/BIBE52308.2021.9635256]
33. Cai Z, Zeng D, Marder KS, Honig LS, Wang Y. Dynamic classification of latent disease progression with auxiliary surrogate labels. arXiv. Preprint posted online on Dec 11, 2024. [doi: 10.3390/math11102335]
34. Chen Y, Pham TD. Development of a brain MRI-based hidden Markov model for dementia recognition. Biomed Eng Online 2013;12 Suppl 1(Suppl 1):S2. [doi: 10.1186/1475-925X-12-S1-S2] [Medline: 24564961]
35. Vats NA, Yadavalli A, Gurugubelli K, Vuppala AK. Acoustic features, BERT model and their complementary nature for Alzheimer's dementia detection. IC3 '21 2021 Aug 5:267-272. [doi: 10.1145/3474124.3474162]
36. Xiao R, Cui X, Qiao H, et al. Early diagnosis model of Alzheimer's disease based on sparse logistic regression with the generalized elastic net. Biomed Signal Process Control 2021 Apr;66:102362. [doi: 10.1016/j.bspc.2020.102362]
37. Ablimit A, Botelho C, Abad A, Schultz T, Trancoso I. Exploring dementia detection from speech: cross corpus analysis. ICASSP 2022 - 2022 IEEE Int Conf Acoust, Speech Signal Proc (ICASSP) 2022:6472-6476. [doi: 10.1109/ICASSP43922.2022.9747167]
38. Lahmiri S. Integrating convolutional neural networks, kNN, and Bayesian optimization for efficient diagnosis of Alzheimer's disease in magnetic resonance images. Biomed Signal Process Control 2023 Feb;80:104375. [doi: 10.1016/j.bspc.2022.104375]
39. Suwalka D, Pandita D, Godse S, Patil RR, Salam Khan A, Kumar A. AI applications and simulation-based learning integrating future of nursing education. 2024 Int Conf Intell Innovative Pract Eng Manage (IIPEM) 2024:1-6. [doi: 10.1109/IIPEM62726.2024.10925639]
40. Chaudhari A, Saratkar S, Thute T. AI-enhanced imaging techniques for understanding Alzheimer's progression. 2025 Int Conf Mach Learn Auton Syst (ICMLAS) 2025:1174-1179. [doi: 10.1109/ICMLAS64557.2025.10969042]
41. Ango R, C KKR, Fatima S, Nag A. Brain connectivity analysis in Alzheimer's disease using graph convolutional network. 2024 4th Int Conf Soft Comput Secur Appl (ICSCSA) 2024:133-139. [doi: 10.1109/ICSCSA64454.2024.00028]
42. Chattopadhyay T, Joshy NA, Ozarkar SS, et al. Deep learning algorithms for Alzheimer's disease detection based on diffusion MRI: tests in Indian and North American cohorts. Alzheimer's Dementia 2024 Dec;20(S2):e089294. [doi: 10.1002/alz.089294] [Medline: 39737627]
43. Ma D, Zhang H, Wang L. Editorial: deep learning methods and applications in brain imaging for the diagnosis of neurological and psychiatric disorders. Front Neurosci 2024;18:1497417. [doi: 10.3389/fnins.2024.1497417] [Medline: 39411146]
44. Williams C, Anik FI, Hasan MM, et al. Advancing brain-computer interface closed-loop systems for neurorehabilitation: A systematic review of AI and machine learning innovations in biomedical engineering (preprint). JMIR Biomed Eng 2025 Nov 5;10:e72218. [doi: 10.2196/72218] [Medline: 41191851]
45. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
46. de Swart WK, Loog M, Krijthe JH. A comparative study of methods for dynamic survival analysis. Front Neurol 2025;16:1504535. [doi: 10.3389/fneur.2025.1504535] [Medline: 40040908]
47. Kang MK, Hong KS, Yang D, Kim HK. Multi-scale neural networks classification of mild cognitive impairment using functional near-infrared spectroscopy. Biocybern Biomed Eng 2025 Jan;45(1):11-22. [doi: 10.1016/j.bbe.2024.12.001]
48. Abir SI, et al. EEG functional connectivity and deep learning for automated diagnosis of Alzheimer's disease and schizophrenia. JCSTS 2025 Jan 26;7(1):82-99. [doi: 10.32996/jcsts.2025.7.1.7]
49. Sathish R, Muthukumar R, Dhivya K, Karthikkumar S. Deep learning and IoT-enabled framework for accurate classification and monitoring of alzheimer's disease based on eeg signal analysis. 2025 Fifth Int Conf Adv Electr, Comput, Commun Sustainable Technol (ICAECT) 2025:1-8. [doi: 10.1109/ICAECT63952.2025.10958882]

XSL•FO
RenderX

50. Dubey AK, Kapoor R, Saraswat M. Optimized machine learning for medical data analysis and disease prediction. 2024 Int Conf Artif Intell Emerging Tech (Global AI Summit 2024:1282-1286. [doi: 10.1109/GlobalAISummit62156.2024.10948001]
51. K P, Chitla VB, Aftab A, Kamath S. LSTM-based assistance for people with Alzheimer's disease. 2025 Int Conf Intell Innovative Tech Comput, Electr Electron (IITCEE) 2025:1-5. [doi: 10.1109/IITCEE64140.2025.10915385]
52. Pan J, Fan Z, Smith GE, Guo Y, Bian J, Xu J. Federated learning with multi-cohort real-world data for predicting the progression from mild cognitive impairment to Alzheimer's disease. Alzheimer's Dement 2025 Apr;21(4):e70128. [doi: 10.1002/alz.70128] [Medline: 40219846]
53. Zuo Y, Zhang B, Dong Y, et al. Glypred: lysine glycation site prediction via CCU–LightGBM–BiLSTM framework with multi-head attention mechanism. J Chem Inf Model 2024 Aug 26;64(16):6699-6711. [doi: 10.1021/acs.jcim.4c01034]
54. Zhu M, Xu Z, Zhang Q, Liu Y, Gu D, Xu SD. GCSTormer: gated swin transformer with channel weights for image denoising. Expert Syst Appl 2025 Jul;284:127924. [doi: 10.1016/j.eswa.2025.127924]
55. Han X, Xue R, Feng J, et al. Hypergraph foundation model for brain disease diagnosis. IEEE Trans Neural Netw Learning Syst 2025;36(10):17702-17716. [doi: 10.1109/TNNLS.2025.3554755]
56. Lu SY, Zhang YD, Yao YD. A regularized transformer with adaptive token fusion for Alzheimer's disease diagnosis in brain magnetic resonance images. Eng Appl Artif Intell 2025 Sep;155:111058. [doi: 10.1016/j.engappai.2025.111058]
57. Li X, Zhu W, Qiu P, Dumitrascu OM, Youssef A, Wang Y. A BERT-style self-supervised learning CNN for disease identification from retinal images. arXiv. Preprint posted online on Apr 25, 2025. [doi: 10.48550/arXiv.2504.18049]
58. Mahapatra C. Exploring advanced applications of artificial intelligence in neuropharmacology: a comprehensive overview. Biol Life Sci. Preprint posted online on May 8, 2025. [doi: 10.20944/preprints202505.0369.v1]
59. Ren H, Zheng Y, Li C, et al. Using machine learning to predict cognitive decline in older adults from the Chinese longitudinal healthy longevity survey: model development and validation study. JMIR Aging 2025 Apr 30;8:e67437. [doi: 10.2196/67437] [Medline: 40305830]
60. Shah YAR, Qureshi SM, Qureshi HA, Shah SUR, Ahmad A, Shiwlani A. Advances in artificial intelligence and machine learning for neurodegenerative disease: a literature review. WJRR 2024 Sep 5;19(3):4-18. [doi: 10.31871/WJRR.19.3.8]
61. Fatima G, Ashiquzzaman A, Kim SS, Kim YR, Kwon HS, Chung E. Vascular and glymphatic dysfunction as drivers of cognitive impairment in Alzheimer's disease: insights from computational approaches. Neurobiol Dis 2025 May;208:106877. [doi: 10.1016/j.nbd.2025.106877] [Medline: 40107629]
62. Yang X, Dang X, Cai J, Li J, Wang X, Heng P. Temporal‐multimodal consistency alignment for Alzheimer's cognitive assessment prediction. Med Phys Mex Symp Med Phys 2025 Jun;52(6):5064-5080. [doi: 10.1002/mp.17767]
63. Sadeghian R, Haider F, Fraser K, Tasaki S, Muniz-Terrera G. Editorial: methods in artificial intelligence for dementia 2024. Front Dement 2024;3:1444825. [doi: 10.3389/frdem.2024.1444825] [Medline: 39081595]
64. Kale M, Wankhede N, Pawar R, et al. AI-driven innovations in Alzheimer's disease: integrating early diagnosis, personalized treatment, and prognostic modelling. Ageing Res Rev 2024 Nov;101:102497. [doi: 10.1016/j.arr.2024.102497]
65. UKbiobank. URL: https://ukbiobank.ac.uk [accessed 2026-02-07]
66. ADNI. URL: https://adni.loni.usc.edu [accessed 2026-02-07]
67. Open access series of imaging studies (OASIS). Washington University in St Louis. URL: https://sites.wustl.edu/oasisbrains/ [accessed 2026-02-07]
68. NACC. URL: https://naccdata.org/ [accessed 2026-02-07]
69. Framingham Heart Study. URL: https://www.framinghamheartstudy.org/ [accessed 2026-02-07]
70. aibl. URL: https://aibl.csiro.au [accessed 2026-02-07]
71. TalkBank. URL: https://dementia.talkbank.org/ [accessed 2026-02-07]
72. Gkoumas D, Wang B, Tsakalidis A, et al. A longitudinal multi-modal dataset for dementia monitoring and diagnosis. Lang Resour Eval 2024;58(3):883-902. [doi: 10.1007/s10579-023-09718-4] [Medline: 39323983]
73. Xu T, Wang X, Lun X, Pan H, Wang Z. ADReFV: face video dataset based on human‐computer interaction for Alzheimer's disease recognition. Comput Animation Virtual 2023 Jan;34(1):e2127. [doi: 10.1002/cav.2127]
74. GENCODE. URL: https://www.gencodegenes.org/ [accessed 2026-02-07]
75. Velazquez M, Lee Y. Multimodal ensemble model for Alzheimer's disease conversion prediction from early mild cognitive impairment subjects. Comput Biol Med 2022 Dec;151(Pt A):106201. [doi: 10.1016/j.compbiomed.2022.106201] [Medline: 36370583]
76. Zhang M, Cui Q, Lü Y, Yu W, Li W. A multimodal learning machine framework for Alzheimer's disease diagnosis based on neuropsychological and neuroimaging data. Comput Ind Eng 2024 Nov;197:110625. [doi: 10.1016/j.cie.2024.110625]
77. Fabietti M, Mahmud M, Lotfi A, et al. Early detection of Alzheimer's disease from cortical and hippocampal local field potentials using an ensembled machine learning model. IEEE Trans Neural Syst Rehabil Eng 2023;31:2839-2848. [doi: 10.1109/TNSRE.2023.3288835]
78. Seifallahi M, Mehraban AH, Galvin JE, Ghoraani B. Alzheimer's disease detection using comprehensive analysis of Timed Up and Go Test via Kinect V.2 camera and machine learning. IEEE Trans Neural Syst Rehabil Eng 2022;30:1589-1600. [doi: 10.1109/TNSRE.2022.3181252] [Medline: 35675251]
79. Xue C, Kowshik SS, Lteif D, et al. AI-based differential diagnosis of dementia etiologies on multimodal data. Nat Med 2024 Oct;30(10):2977-2989. [doi: 10.1038/s41591-024-03118-z] [Medline: 38965435]

XSL•FO
RenderX

80. Shi J, Zheng X, Li Y, Zhang Q, Ying S. Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. IEEE J Biomed Health Inform 2018 Jan;22(1):173-183. [doi: 10.1109/JBHI.2017.2655720]
81. Allwright M, Mundell HD, McCorkindale AN, et al. Ranking the risk factors for Alzheimer's disease; findings from the UK Biobank study. Aging Brain 2023;3:100081. [doi: 10.1016/j.nbas.2023.100081] [Medline: 37384134]
82. Gu Z, Liu S, Ma H, et al. Estimation of machine learning-based models to predict dementia risk in patients with atherosclerotic cardiovascular diseases: UK Biobank study. JMIR Aging 2025 Feb 26;8:e64148. [doi: 10.2196/64148] [Medline: 40009844]
83. You J, Zhang YR, Wang HF, et al. Development of a novel dementia risk prediction model in the general population: a large, longitudinal, population-based machine-learning study. eClinicalMedicine 2022 Nov;53:101665. [doi: 10.1016/j.eclinm.2022.101665]
84. Calvo N, McFall GP, Ramana S, et al. Associated risk and resilience factors of Alzheimer's disease in women with early bilateral oophorectomy: data from the UK Biobank. J Alzheimers Dis 2024 Nov;102(1):119-128. [doi: 10.3233/JAD-240646] [Medline: 39497303]
85. Yi F, Yuan J, Somekh J, et al. Genetically supported targets and drug repurposing for brain aging: a systematic study in the UK Biobank. Sci Adv 2025 Mar 14;11(11):eadr3757. [doi: 10.1126/sciadv.adr3757] [Medline: 40073132]
86. Yousefzadeh N, Tran C, Ramirez-Zamora A, Chen J, Fang R, Thai MT. Neuron-level explainable AI for Alzheimer's disease assessment from fundus images. Sci Rep 2024 Apr 2;14(1):7710. [doi: 10.1038/s41598-024-58121-8] [Medline: 38565579]
87. Gong W, Bai S, Zheng YQ, Smith SM, Beckmann CF. Supervised phenotype discovery from multimodal brain imaging. IEEE Trans Med Imaging 2023 Mar;42(3):834-849. [doi: 10.1109/TMI.2022.3218720] [Medline: 36318559]
88. Lian C, Liu M, Pan Y, Shen D. Attention-guided hybrid network for dementia diagnosis with structural MR images. IEEE Trans Cybern 2022 Apr;52(4):1992-2003. [doi: 10.1109/TCYB.2020.3005859] [Medline: 32721906]
89. Lian C, Liu M, Wang L, Shen D. Multi-task weakly-supervised attention network for dementia status estimation with structural MRI. IEEE Trans Neural Netw Learning Syst 2022 Aug;33(8):4056-4068. [doi: 10.1109/TNNLS.2021.3055772]
90. Li H, Habes M, Wolk DA, Fan Y. Alzheimer's disease neuroimaging initiative and the Australian Imaging Biomarkers and Lifestyle Study of Aging. A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal magnetic resonance imaging data. Alzheimer's Dementia 2019 Aug;15(8):1059-1070. [doi: 10.1016/j.jalz.2019.02.007] [Medline: 31201098]
91. Oh K, Yoon JS, Suk HI. Learn-explain-reinforce: counterfactual reasoning and its guidance to reinforce an Alzheimer's disease diagnosis model. IEEE Trans Pattern Anal Mach Intell 2023 Apr;45(4):4843-4857. [doi: 10.1109/TPAMI.2022.3197845] [Medline: 35947563]
92. Lian C, Liu M, Zhang J, Shen D. Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. IEEE Trans Pattern Anal Mach Intell 2020 Apr;42(4):880-893. [doi: 10.1109/TPAMI.2018.2889096] [Medline: 30582529]
93. Avsec Ž, Agarwal V, Visentin D, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods 2021 Oct;18(10):1196-1203. [doi: 10.1038/s41592-021-01252-x]
94. Yang L, Wang X, Guo Q, et al. For the Alzheimer's disease neuroimaging initiative. deep learning based multimodal progression modeling for Alzheimer's disease. Stat Biopharm Res 2021 Jul 3;13:337-343. [doi: 10.1080/19466315.2021.1884129]
95. Lee LY, Vaghari D, Burkhart MC, et al. Robust and interpretable AI-guided marker for early dementia prediction in real-world clinical settings. eClinicalMedicine 2024 Aug;74:102725. [doi: 10.1016/j.eclinm.2024.102725]
96. Zhu W, Sun L, Huang J, Han L, Zhang D. Dual attention multi-instance deep learning for Alzheimer's disease diagnosis with structural MRI. IEEE Trans Med Imaging 2021 Sep;40(9):2354-2366. [doi: 10.1109/TMI.2021.3077079]
97. Zhang M, Cui Q, Lü Y, Li W. A feature-aware multimodal framework with auto-fusion for Alzheimer's disease diagnosis. Comput Biol Med 2024 Aug;178:108740. [doi: 10.1016/j.compbiomed.2024.108740] [Medline: 38901184]
98. Bi XA, Hu X, Wu H, Wang Y. Multimodal data analysis of Alzheimer's disease based on clustering evolutionary random forest. IEEE J Biomed Health Inform 2020 Oct;24(10):2973-2983. [doi: 10.1109/JBHI.2020.2973324] [Medline: 32071013]
99. Bi XA, Xing Z, Zhou W, Li L, Xu L. Pathogeny detection for mild cognitive impairment via weighted evolutionary random forest with brain imaging and genetic data. IEEE J Biomed Health Inform 2022 Jul;26(7):3068-3079. [doi: 10.1109/JBHI.2022.3151084] [Medline: 35157601]
100. Hashmi A, Barukab O. Dementia classification using deep reinforcement learning for early diagnosis. Appl Sci (Basel) 2023 Jan 22;13(3):1464. [doi: 10.3390/app13031464]
101. Wang Y, Gao R, Wei T, et al. Predicting long-term progression of Alzheimer's disease using a multimodal deep learning model incorporating interaction effects. J Transl Med 2024 Mar 11;22(1):265. [doi: 10.1186/s12967-024-05025-w]
102. Hatami M, Yaghmaee F, Ebrahimpour R. Investigating the potential of reinforcement learning and deep learning in improving Alzheimer's disease classification. Neurocomputing 2024 Sep;597:128119. [doi: 10.1016/j.neucom.2024.128119]
103. Tabarestani S, Aghili M, Eslami M, et al. A distributed multitask multimodal approach for the prediction of Alzheimer's disease in a longitudinal study. Neuroimage 2020 Feb 1;206:116317. [doi: 10.1016/j.neuroimage.2019.116317] [Medline: 31678502]

XSL•FO
RenderX

104. Burkhart MC, Lee LY, Vaghari D, et al. Unsupervised multimodal modeling of cognitive and brain health trajectories for early dementia prediction. Sci Rep 2024 May 10;14(1):10755. [doi: 10.1038/s41598-024-60914-w] [Medline: 38729989]
105. El-Sappagh S, Alonso JM, Islam SMR, Sultan AM, Kwak KS. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. Sci Rep 2021 Jan 29;11(1):2660. [doi: 10.1038/s41598-021-82098-3] [Medline: 33514817]
106. Lee MW, Kim HW, Choe YS, et al. A multimodal machine learning model for predicting dementia conversion in Alzheimer's disease. Sci Rep 2024 May 29;14(1):12276. [doi: 10.1038/s41598-024-60134-2]
107. Yuan S, Li H, Wu J, Sun X. Classification of mild cognitive impairment with multimodal data using both labeled and unlabeled samples. IEEE/ACM Trans Comput Biol and Bioinf 2021 Nov 1;18(6):2281-2290. [doi: 10.1109/TCBB.2021.3053061]
108. Cirincione A, Lynch K, Bennett J, et al. Prediction of future dementia among patients with mild cognitive impairment (MCI) by integrating multimodal clinical data. Heliyon 2024 Sep 15;10(17):e36728. [doi: 10.1016/j.heliyon.2024.e36728] [Medline: 39281465]
109. Cassani R, Falk TH. Alzheimer's disease diagnosis and severity level detection based on electroencephalography modulation spectral "patch" features. IEEE J Biomed Health Inform 2020 Jul;24(7):1982-1993. [doi: 10.1109/JBHI.2019.2953475] [Medline: 31725401]
110. Cilia ND, D'Alessandro T, De Stefano C, Fontanella F, Molinara M. From online handwriting to synthetic images for Alzheimer's disease detection using a deep transfer learning approach. IEEE J Biomed Health Inform 2021 Dec;25(12):4243-4254. [doi: 10.1109/JBHI.2021.3101982]
111. Kmetzsch V, Becker E, Saracino D, et al. Disease progression score estimation from multimodal imaging and MicroRNA data using supervised variational autoencoders. IEEE J Biomed Health Inform 2022 Dec;26(12):6024-6035. [doi: 10.1109/JBHI.2022.3208517]
112. Mengoudi K, Ravi D, Yong KXX, et al. Augmenting dementia cognitive assessment with instruction-less eye-tracking tests. IEEE J Biomed Health Inform 2020 Nov;24(11):3066-3075. [doi: 10.1109/JBHI.2020.3004686] [Medline: 32749977]
113. Tsai H, Yang TW, Ou KH, Su TH, Lin C, Chou CF. Multimodal attention network for dementia prediction. IEEE J Biomed Health Inform 2024 Nov;28(11):6918-6930. [doi: 10.1109/JBHI.2024.3438885]
114. Wu EQ, Peng XY, Chen SD, Zhao XY, Tang ZR. Detecting Alzheimer's dementia degree. IEEE Trans Cogn Dev Syst 2022 Mar;14(1):116-125. [doi: 10.1109/TCDS.2020.3015131]
115. Zhang H, Ni M, Yang Y, et al. Patch-based interpretable deep learning framework for Alzheimer's disease diagnosis using multimodal data. Biomed Signal Process Control 2025 Feb;100:107085. [doi: 10.1016/j.bspc.2024.107085]
116. Fan CC, Yang H, Zhang C, et al. Graph reasoning module for Alzheimer's disease diagnosis: a plug-and-play method. IEEE Trans Neural Syst Rehabil Eng 2023;31:4773-4780. [doi: 10.1109/TNSRE.2023.3337533]
117. Beebe-Wang N, Okeson A, Althoff T, Lee SI. Efficient and explainable risk assessments for imminent dementia in an aging cohort study. IEEE J Biomed Health Inform 2021 Jul;25(7):2409-2420. [doi: 10.1109/JBHI.2021.3059563]
118. Battineni G, Hossain MA, Chintalapudi N, et al. Improved Alzheimer's disease detection by MRI using multimodal machine learning algorithms. Diagnostics (Basel) 2021 Nov 13;11(11):2103. [doi: 10.3390/diagnostics11112103]
119. Nguyen H, Chu NN. An introduction to deep learning research for Alzheimer's disease. IEEE Consumer Electron Mag 2021 May 1;10(3):72-75. [doi: 10.1109/MCE.2020.3048254]
120. Fan F, Song H, Jiang J, et al. Development and validation of a multimodal deep learning framework for vascular cognitive impairment diagnosis. iScience 2024 Oct;27(10):110945. [doi: 10.1016/j.isci.2024.110945]
121. Ilias L, Askounis D, Psarras J. Detecting dementia from speech and transcripts using transformers. Comput Speech Lang 2023 Apr;79:101485. [doi: 10.1016/j.csl.2023.101485]
122. Poor FF, Dodge HH, Mahoor MH. A multimodal cross-transformer-based model to predict mild cognitive impairment using speech, language and vision. Comput Biol Med 2024 Nov;182:109199. [doi: 10.1016/j.compbiomed.2024.109199] [Medline: 39332117]
123. Lin K, Washington PY. Multimodal deep learning for dementia classification using text and audio. Sci Rep 2024 Jun 16;14(1):13887. [doi: 10.1038/s41598-024-64438-1]
124. Ortiz-Perez D, Ruiz-Ponce P, Tomás D, Garcia-Rodriguez J, Vizcaya-Moreno MF, Leo M. A deep learning-based multimodal architecture to predict signs of dementia. Neurocomputing 2023 Sep;548:126413. [doi: 10.1016/j.neucom.2023.126413]
125. Ilias L, Askounis D. Explainable identification of dementia from transcripts using transformer networks. IEEE J Biomed Health Inform 2022 Aug;26(8):4153-4164. [doi: 10.1109/JBHI.2022.3172479] [Medline: 35511841]
126. Wen B, Wang N, Subbalakshmi K, Chandramouli R. Revealing the roles of part-of-speech taggers in Alzheimer disease detection: scientific discovery using one-intervention causal explanation. JMIR Form Res 2023 May 2;7:e36590. [doi: 10.2196/36590] [Medline: 37129944]
127. Chen X, Pu Y, Li J, Zhang WQ. Cross-lingual Alzheimer's disease detection based on paralinguistic and pre-trained features. ICASSP 2023 - 2023 IEEE Int Conf Acoustics, Speech Signal Proc (ICASSP) 2023:1-2. [doi: 10.1109/ICASSP49357.2023.10095522]
128. Zheng C, Bouazizi M, Ohtsuki T. An evaluation on information composition in dementia detection based on speech. IEEE Access 2022;10:92294-92306. [doi: 10.1109/ACCESS.2022.3203068]

XSL•FO
RenderX

129. Nambiar AS, Likhita K, Pujya K, Gupta D, Vekkot S, Lalitha S. Comparative study of deep classifiers for early dementia detection using speech transcripts. 2022 IEEE 19th India Counc Int Conf (INDICON) 2022:1-6. [doi: 10.1109/INDICON56171.2022.10039705]
130. Priyadarshinee P, Clarke CJ, Melechovsky J, Lin CMY, B. T. B, Chen JM. Alzheimer's dementia speech (audio vs. text): multi-modal machine learning at high vs. low resolution. Appl Sci (Basel) 2023;13(7):4244. [doi: 10.3390/app13074244]
131. Liu J, Fu F, Li L, et al. Efficient pause extraction and encode strategy for Alzheimer's disease detection using only acoustic features from spontaneous speech. Brain Sci 2023 Mar 11;13(3):477. [doi: 10.3390/brainsci13030477] [Medline: 36979287]
132. Mahajan P, Baths V. Acoustic and language based deep learning approaches for Alzheimer's dementia detection from spontaneous speech. Front Aging Neurosci 2021;13:623607. [doi: 10.3389/fnagi.2021.623607] [Medline: 33613269]
133. Mei K, Ding X, Liu Y, et al. The USTC system for ADReSS-M challenge. ICASSP 2023 - 2023 IEEE Int Conf Acoustics, Speech Signal Proc (ICASSP) 2023:1-2. [doi: 10.1109/ICASSP49357.2023.10094714]
134. Ali Meerza SI, Li Z, Liu L, Zhang J, Liu J. Fair and privacy-preserving Alzheimer's disease diagnosis based on spontaneous speech analysis via federated learning. 2022 44th Ann Int Conf IEEE Eng Med Biol Soc (EMBC) 2022:1362-1365. [doi: 10.1109/EMBC48229.2022.9871204]
135. Chen W, Xing X, Xu X, Pang J, Du L. SpeechFormer++: a hierarchical efficient framework for paralinguistic speech processing. IEEE/ACM Trans Audio Speech Lang Process 2023;31:775-788. [doi: 10.1109/TASLP.2023.3235194]
136. Tamm B, Vandenberghe R, Van Hamme H. Cross-lingual transfer learning for alzheimer's detection from spontaneous speech. ICASSP 2023 - 2023 IEEE Int Conf Acoust, Speech Signal Process (ICASSP) 2023:1-2. [doi: 10.1109/ICASSP49357.2023.10096770]
137. Woszczyk D, Hedlikova A, Akman A, Demetriou S, Schuller B. Data augmentation for dementia detection in spoken language. Proc Interspeech 2022 2022:2858-2862. [doi: 10.21437/Interspeech.2022-10210]
138. Jin L, Oh Y, Kim H, et al. CONSEN: complementary and simultaneous ensemble for Alzheimer's disease detection and MMSE score prediction. ICASSP 2023 - 2023 IEEE Int Conf Acoustics, Speech Signal Proc (ICASSP) 2023:1-2. [doi: 10.1109/ICASSP49357.2023.10096253]
139. Ilias L, Askounis D. Context-aware attention layers coupled with optimal transport domain adaptation and multimodal fusion methods for recognizing dementia from spontaneous speech. Knowl Based Syst 2023 Oct;277:110834. [doi: 10.1016/j.knosys.2023.110834]
140. Azevedo T, Bethlehem RAI, Whiteside DJ, et al. Identifying healthy individuals with Alzheimer's disease neuroimaging phenotypes in the UK Biobank. Commun Med 2023 Jul 20;3(1):100. [doi: 10.1038/s43856-023-00313-w]
141. Liang S, Chen T, Ma J, Ren S, Lu X, Du W. Identification of mild cognitive impairment using multimodal 3D imaging data and graph convolutional networks. Phys Med Biol 2024 Dec 7;69(23):235002. [doi: 10.1088/1361-6560/ad8c94]
142. Jahan S, Abu Taher K, Kaiser MS, et al. Explainable AI-based Alzheimer's prediction and management using multimodal data. PLOS ONE 2023;18(11):e0294253. [doi: 10.1371/journal.pone.0294253] [Medline: 37972072]
143. Jahan S, Saif Adib M, Huda SM, et al. Federated explainable AI-based Alzheimer's disease prediction with multimodal data. IEEE Access 2025;13:43435-43454. [doi: 10.1109/ACCESS.2025.3547343]
144. Myrzashova R, Alsamhi SH, Shvetsov AV, Hawbani A, Guizani M, Wei X. BCFTL: blockchain-enabled multimodal federated transfer learning for decentralized Alzheimer's diagnosis. IEEE Internet Things J 2025;12(15):29656-29669. [doi: 10.1109/JIOT.2025.3569652]
145. Chen K, Weng Y, Huang Y, et al. A multi‐view learning approach with diffusion model to synthesize FDG PET from MRI T1WI for diagnosis of Alzheimer's disease. Alzheimers Dement 2025 Feb;21(2):e14421. [doi: 10.1002/alz.14421]
146. Lin W, Lin W, Chen G, et al. Bidirectional mapping of brain MRI and PET with 3D Reversible GAN for the diagnosis of Alzheimer's disease. Front Neurosci 2021;15:646013. [doi: 10.3389/fnins.2021.646013] [Medline: 33935634]
147. Gupta B, Jegannathan GK, Alam MS, et al. Multimodal lightweight neural network for Alzheimer's disease diagnosis integrating neuroimaging and cognitive scores. Neurosci Inf 2025 Sep;5(3):100218. [doi: 10.1016/j.neuri.2025.100218]
148. Chen Z, Wang Z, Zhao M, et al. A new classification network for diagnosing Alzheimer's disease in class-imbalance MRI datasets. Front Neurosci 2022 Aug 25;16:807085. [doi: 10.3389/fnins.2022.807085]
149. Sarma M, Chatterjee D. Multistage diagnosis of Alzheimer's disease from clinical data using 'deep ensemble learning'. JAIAI 2024;01(1):122-138. [doi: 10.54364/JAIAI.2024.1109]
150. Mujahid M, Rehman A, Alam T, Alamri FS, Fati SM, Saba T. An efficient ensemble approach for Alzheimer's disease detection using an adaptive synthetic technique and deep learning. Diagnostics (Basel) 2023 Jul 26;13(15):2489. [doi: 10.3390/diagnostics13152489] [Medline: 37568852]
151. Dubey Y, Bhongade A, Palsodkar P, Fulzele P. Efficient explainable models for Alzheimer's disease classification with feature selection and data balancing approach using ensemble learning. Diagnostics (Basel) 2024 Dec 10;14(24):2770. [doi: 10.3390/diagnostics14242770] [Medline: 39767131]
152. Mandawkar U, Diwan T. Hybrid cuttle Fish-Grey wolf optimization tuned weighted ensemble classifier for Alzheimer's disease classification. Biomed Signal Process Control 2024 Jun;92:106101. [doi: 10.1016/j.bspc.2024.106101]
153. Jasodanand VH, Kowshik SS, Puducheri S, et al. AI-driven fusion of multimodal data for Alzheimer's disease biomarker assessment. Nat Commun 2025 Aug 11;16(1):7407. [doi: 10.1038/s41467-025-62590-4] [Medline: 40789853]

XSL•FO
RenderX

154. Weiner MW, Kanoria S, Miller MJ, et al. Overview of Alzheimer's disease neuroimaging Initiative and future clinical trials. Alzheimer's Dement 2025 Jan;21(1):e14321. [doi: 10.1002/alz.14321] [Medline: 39711072]
155. Wilkinson T, Schnier C, Bush K, et al. Identifying dementia outcomes in UK Biobank: a validation study of primary care, hospital admissions and mortality data. Eur J Epidemiol 2019 Jun;34(6):557-565. [doi: 10.1007/s10654-019-00499-1]
156. Thulasimani V, Shanmugavadivel K, Cho J, Easwaramoorthy SV. A review of datasets, optimization strategies, and learning algorithms for analyzing Alzheimer's dementia detection. Neuropsychiatr Dis Treat 2024;20:2203-2225. [doi: 10.2147/NDT.S496307] [Medline: 39588176]
157. Chan KCG, Xia F, Kukull WA. NACC data: who is represented over time and across centers, and implications for generalizability. Alzheimer's Dement 2025 Sep;21(9):e70657. [doi: 10.1002/alz.70657] [Medline: 40968249]
158. Fowler C, Rainey-Smith SR, Bird S, et al. Fifteen years of the Australian Imaging, Biomarkers and Lifestyle (AIBL) study: progress and observations from 2,359 older adults spanning the spectrum from cognitive normality to Alzheimer's disease. J Alzheimer's Dis Rep 2021 Mar 11;5(1):443-468. [doi: 10.3233/ADR-210005]
159. Yang Q, Li X, Ding X, Xu F, Ling Z. Deep learning-based speech analysis for Alzheimer's disease detection: a literature review. Alz Res Therapy 2022 Dec 14;14(1):186. [doi: 10.1186/s13195-022-01131-3]
160. He Y, Wang Z, Zhang Y, et al. NeuroSymAD: a neuro-symbolic framework for interpretable Alzheimer's disease diagnosis. . Preprint posted online on Mar 1, 2025. [doi: 10.48550/arXiv.2503.00510]
161. Sadeghi A, Hajati F, Argha A, Lovell NH, Yang M. Interpretable graph-based models on multimodal biomedical data integration: a technical review and benchmarking. arXiv. Preprint posted online on May 3, 2025. [doi: 10.48550/arXiv.2505.01696]
162. Mahamud E, Assaduzzaman M, Islam J, Fahad N, Hossen MJ, Ramanathan TT. Enhancing Alzheimer's disease detection: an explainable machine learning approach with ensemble techniques. Intell-Based Med 2025;11:100240. [doi: 10.1016/j.ibmed.2025.100240]
163. Zhou T, Liu M, Thung KH, Shen D. Latent representation learning for Alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data. IEEE Trans Med Imaging 2019 Oct;38(10):2411-2422. [doi: 10.1109/TMI.2019.2913158]
164. Sharma R, Sibille L, Fahmi R. Multi‐branch convolutional neural network for Alzheimer's disease versus normal control classification using PET images. Alzheimer's Dementia 2023 Jun;19(S3):e061092. [doi: 10.1002/alz.061092]
165. Zhang J, Yu X, Chen T, et al. BrainNet-moe: brain-inspired mixture-of-experts learning for neurological disease identification. . Preprint posted online on Mar 5, 2025. [doi: 10.48550/ARXIV.2503.07640]

Abbreviations

AD: Alzheimer disease
AdaBoost: Adaptive Boosting
ADNI: Alzheimer's Disease Neuroimaging Initiative
ADReSS: Alzheimer's Dementia Recognition Through Spontaneous Speech
ADReSSo: Alzheimer's Dementia Recognition Through Spontaneous Speech 2021 Challenge
AI: artificial intelligence
ALBERT: A Lite Bidirectional Encoder Representations From Transformers
AUC: area under the curve
BERT: Bidirectional Encoder Representations From Transformers
BiLSTM: bidirectional long short-term memory
DeiT: Data-Efficient Image Transformers
DNN: DNN
EEG: electroencephalography
FL: federated learning
HIPAA : Health Insurance Portability and Accountability Act
LightGBM: Light Gradient-Boosting Machine
LIME: Local Interpretable Model-Agnostic Explanations
MCI: mild cognitive impairment
ML: machine learning
MRI: magnetic resonance imaging
NACC: National Alzheimer's Coordinating Centre
OASIS: Open Access Series of Imaging Studies
PET: positron emission tomography
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
QUADAS-2: Revised Quality Assessment of Diagnostic Accuracy Studies tool
RL: reinforcement learning

XSL•FO
RenderX

RoBERTa+: Robustly Optimized Bidirectional Encoder Representations From Transformers Approach
SHAP: Shapley Additive Explanations
ViT: vision transformer
XGBoost: Extreme Gradient Boosting

Edited by S Brini; submitted 07.Oct.2025; peer-reviewed by F Elkourdi, MM Sikder, O Oyetunji, S Jiang; revised version received 09.Jan.2026; accepted 09.Jan.2026; published 25.Mar.2026.

Please cite as:
Yu Z, Mulholland A, Huang T, Liu Q
Multimodal AI for Alzheimer Disease Diagnosis: Systematic Review of Datasets, Models, and Modalities
J Med Internet Res 2026;28:e85414
URL: https://www.jmir.org/2026/1/e85414
doi:10.2196/85414

© Ziwen Yu, Anthony Mulholland, Tianyan Huang, Qiang Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effectiveness of Mobile Health for Improving Medication Adherence in Patients With Cancer: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Wei Chao Liu[1], MN; Li Ping Wang[1,2], PhD; Qian Yu Zhang[1], MN; Jing Han Yang[1], MN

[1]School of Public Health and Nursing, Hangzhou Normal University, No. 2318 Yuhangtang Road, Hangzhou Normal University Cangqian Campus, Hangzhou, China

[2]Engineering Research Center of Mobile Health Management System, Ministry of Education, Hangzhou, China

Corresponding Author:
Li Ping Wang, PhD
School of Public Health and Nursing, Hangzhou Normal University, No. 2318 Yuhangtang Road, Hangzhou Normal University Cangqian Campus, Hangzhou, China

Abstract

Background: Medication adherence among patients with cancer is generally low. Mobile health (mHealth) has gradually been applied to improve this situation, but systematic evidence of its effectiveness remains lacking.

Objective: We aimed to evaluate the effect of mHealth on improving medication adherence among patients with cancer.

Methods: This systematic review included randomized controlled trials (RCTs) evaluating the impact of mHealth on medication adherence among patients with cancer. Systematic searches were conducted in PubMed, Web of Science, CINAHL, Cochrane Library, Embase, Sinomed, CNKI, Cqvip, and ClinicalTrials.gov from inception to December 31, 2025. Two researchers independently performed literature screening, data extraction, and risk of bias assessment. Effects were pooled using a random-effects model (Hartung-Knapp-Sidik-Jonkman), and standardized mean differences (SMDs) and odds ratios (ORs) with 95% CIs have been reported. Evidence quality was assessed using the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) framework.

Results: A total of 17 RCTs (1309 participants) from 8 countries published between 2016 and 2025 were included. mHealth interventions included mobile apps, websites, and text messaging services. The meta-analysis revealed that compared with controls, mHealth interventions significantly improved medication adherence rates (OR 3.47, 95% CI 1.92-6.26; *P*=.002), medication adherence scores (SMD 1.01, 95% CI 0.51-1.52; *P*=.001), self-efficacy (SMD 0.90, 95% CI 0.29-1.51; *P*=.01), and service satisfaction while reducing symptom burden (SMD −0.38, 95% CI −0.61 to −0.14; *P*=.008). However, mHealth had no significant effect on health literacy (SMD 0.51, 95% CI –1.50 to 2.52; *P*=.29). Subgroup analysis revealed that interventions lasting <3 months outperformed those lasting ≥3 months in improving adherence scores (SMD 1.37, 95% CI 0.78-1.96 vs SMD 0.49, 95% CI −0.39 to 1.37; χ^2_1=5.98; *P*=.01). Regarding intervention format, text messaging services demonstrated superior efficacy compared with mobile apps and websites (SMD 1.53, 95% CI −5.49 to 8.55 vs SMD 1.01, 95% CI 0.42-1.61 and SMD 0.11, 95% CI −0.34 to 0.56, respectively; χ^2_2=10.28; *P*=.006). Across cancer types, mHealth most significantly improved adherence scores in patients with breast cancer (SMD 1.29, 95% CI −5.25 to 7.83), outperforming the findings in patients with leukemia and other cancer types (SMD 0.28, 95% CI −0.87 to 1.42 and SMD 1.09, 95% CI 0.10-2.08, respectively; χ^2_2=8.86; *P*=.01).

Conclusions: Our findings confirm that mHealth plays a positive role in improving medication adherence, enhancing patient self-efficacy, increasing patient satisfaction with services, and alleviating symptom burden. However, these findings should be interpreted with caution owing to substantial heterogeneity, a moderate risk of bias, and a low certainty of evidence. Future research should enhance methodological quality by conducting multicenter, large-sample, high-quality RCTs and should explore the long-term effects and cost-effectiveness of mHealth across diverse health care settings and patient populations to clarify its role and value within comprehensive cancer care management systems.

Trial Registration: PROSPERO CRD420251162181; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251162181

(J Med Internet Res 2026;28:e85949) doi:10.2196/85949

KEYWORDS

mobile health; medication adherence; cancer; randomized controlled trial; systematic review; meta-analysis

XSL•FO
RenderX

Introduction

With the continuous rise in global cancer incidence, cancer has become one of the major public health threats to human health. In recent years, the rapid advancement of targeted therapy, immunotherapy, and endocrine therapy has led to the increasingly widespread clinical application of oral anticancer drugs [1]. Compared with intravenous treatment, oral therapy offers advantages, such as convenient administration and lower cost, making it a first-line treatment option in comprehensive cancer management [2]. Statistics indicate that approximately 60% of patients with cancer require 12 - 24 weeks of oral endocrine therapy, targeted therapy, immunotherapy, or chemotherapy [3].

Medication adherence refers to whether patients take their drugs as prescribed by health care professionals and whether they consistently continue the prescribed medications [4]. For oral therapies, treatment efficacy is largely dependent on patient adherence [5]. Unlike intravenous therapy, oral treatment relies primarily on patients and caregivers administering medications independently at home, lacking real-time supervision by health care providers. Compounded by factors, such as prolonged treatment duration, significant drug toxicity and side effects, heavy financial burden, and increased psychological stress, patients with cancer are prone to behaviors such as missed doses, unauthorized dose adjustment, and premature discontinuation [6,7]. Studies indicate that medication adherence among patients with cancer is below 50%, reflecting an overall low level [8].

Poor medication adherence not only reduces the effectiveness of oral therapy but may also cause a series of adverse effects for patients. First, regarding symptom control, patients who fail to take medication regularly or who arbitrarily adjust dosages are prone to fluctuations in blood drug concentrations. This makes it difficult to achieve sustained suppression of tumor growth, thereby diminishing treatment efficacy and exacerbating the burden of symptoms, such as pain, fatigue, and nausea, severely impacting the quality of life of patients [9]. Second, regarding treatment safety, poor medication adherence may increase the risk of adverse events. Failure to follow prescribed regimens not only disrupts established treatment schedules but may also result in insufficient accumulation of effective drug doses, complicating the pattern of adverse reaction occurrence. Clinicians struggle to accurately distinguish drug-induced adverse reactions from natural disease progression, impairing clinical judgment and treatment decisions. This delays optimal intervention timing and increases the likelihood of patients requiring emergency care or unplanned hospitalizations due to complications [10]. Finally, regarding long-term disease outcomes, the greatest concern with poor medication adherence lies in inducing tumor resistance and elevating the risk of recurrence and progression. When tumor cells are repeatedly exposed to drug concentrations insufficient to inhibit their growth, the development of tumor resistance becomes inevitable. This ultimately leads to treatment failure, increasing the risk of tumor recurrence and disease progression, thereby threatening the long-term survival of patients [11]. Consequently, effectively improving medication adherence among patients with cancer has become one of the most pressing challenges in contemporary oncology treatment management.

With the continuous advancement of mobile communication technology, mobile health (mHealth) has emerged as a novel health care service model that is being increasingly applied in chronic disease and cancer management. Utilizing wireless devices, such as smartphones, health monitoring wristbands, and smart pill bottles, mHealth provides services, including medication reminders, symptom monitoring, health education, and remote follow-ups, offering the advantages of cost-effectiveness, convenience, and unrestricted access regardless of time or location [12]. Recent studies indicate that mHealth interventions positively influence medication adherence among patients with cancer [13].

However, current research findings on the impact of mHealth interventions on medication adherence among patients with cancer remain inconsistent. Studies vary greatly in intervention formats and implementation durations, and their effectiveness across different cancer types remains unclear. Although a 2022 systematic review indicated that mHealth interventions can effectively improve medication adherence among chronic disease populations, including patients with cancer [14], systematic evaluations specifically targeting patients with cancer are currently lacking. Furthermore, this review did not perform subgroup analyses based on patient characteristics, intervention formats, or duration, leaving the impact of these variables on adherence improvement unclear. There is a critical need to systematically synthesize existing evidence to more accurately assess the practical efficacy of mHealth interventions in enhancing medication adherence among patients with cancer.

We aim to conduct a systematic review and meta-analysis to systematically integrate evidence on the impact of mHealth interventions on medication adherence among patients with cancer. Through subgroup analyses, we further explore the effects of intervention format, duration, and cancer type on intervention outcomes, thereby providing evidence-based guidance for clinical practice and mHealth intervention design.

Methods

Registration

This systematic review adheres to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist (Checklist 1) and has been registered in PROSPERO (CRD42024522261) [15].

Search Strategy

The initial literature search strategy for this systematic review was designed according to the preregistered protocol. However, the original protocol included only Chinese and English studies and used insufficient search terms, resulting in limitations of the search strategy. Therefore, we optimized the search strategy in the final draft to enhance comprehensiveness and reduce language bias. These modifications have been explicitly reported as post-hoc adjustments in accordance with research standards.

This systematic review's literature search has been reported according to the PRISMA-S statement [16]. A combination of

XSL•FO
RenderX

subject headings and free-text terms was used for systematic retrieval, covering the following databases and registries: PubMed (NCBI), Embase (Embase.com), Web of Science (Clarivate), Cochrane Library (Wiley), CINAHL (EBSCO), China National Knowledge Infrastructure (CNKI), CqVIP (VIP Information), Sinomed (China Biology Medicine Disc; CBMdisc Platform), and ClinicalTrials.gov (NLM). The literature search timeframe was set from database inception to June 28, 2025. To ensure the timeliness of the search results, the retrieval strategy was optimized and rerun on December 31, 2025, to incorporate newly published literature.

The search strategy for this systematic review was developed by refining and optimizing strategies from prior studies [17], and it underwent peer review by experts in library and information retrieval. Search parameters were set to include only randomized controlled trials (RCTs) and peer-reviewed articles, with no restrictions on publication year or language, and no predefined filters were applied. To maximize retrieval volume, snowball citation searching was performed. This involved screening the reference lists of the included studies (backward citation searching) as well as identifying and reviewing studies that cited the included articles (forward citation searching). It should be noted that the literature search for this review was conducted exclusively using the aforementioned method and did not include searches of printed conference proceedings or government websites. Detailed search strategies are provided in Multimedia Appendix 1.

Inclusion and Exclusion Criteria

Inclusion Criteria

The inclusion criteria were predefined using the PICOS (participants, interventions, control/comparison measures, outcomes, study type) framework [15]. The criteria were as follows: (1) participants: patients with a pathologically confirmed cancer diagnosis who were currently taking medication; (2) interventions: mHealth interventions designed to improve medication adherence among patients with cancer, including apps, websites, and text messaging services; (3) control/comparison measures: placebo controls, routine or standard interventions, and other non-mHealth interventions; (4) outcomes: primary outcomes, including medication adherence rate and mean medication adherence score, and secondary outcomes, including health literacy, self-efficacy, symptom burden, and satisfaction; and (5) study type: studies with a RCT design.

Exclusion Criteria

We excluded the following studies: (1) non–peer-reviewed research, such as research proposals, conference abstracts, and reviews, typically lacking rigorous methodological validation; (2) those for which full-text access was unavailable; and (3) those lacking accessible or convertible medication adherence data.

Definition and Classification of mHealth

The mHealth interventions included in this review varied in their implementation formats. To facilitate subsequent integration and comparison, we predefined the following classification criteria based on the primary delivery platform and user interaction model of the intervention content [18]:

1. Mobile apps: Interventions are primarily implemented through dedicated software installed on smartphones or tablets. Such apps can be accessed via mainstream app stores (eg, App Store and Google Play) or through links provided by the research team. They typically feature interactive interfaces (eg, menus, buttons, and data input fields) and support offline functionality for at least some features.
2. Websites: Interventions are primarily delivered through mobile device browsers or web applications (without requiring local software installation). Key characteristics include: access via URL, interfaces optimized for mobile screens (responsive design), and typical requirement of a persistent internet connection.
3. Text messaging services: Interventions are primarily delivered via short messaging or multimedia messaging systems, push text, or multimedia content. These services rely on cellular network messaging systems using asynchronous communication. Owing to character length limitations, the content is typically concise.

Study Selection

After importing the retrieved studies into EndNote 20.6 (Clarivate Analytics) and removing duplicates using the software, 2 researchers (JHY and QYZ) were invited to cross-review the titles and abstracts to complete the initial screening and full-text reading of the remaining literature. When disagreements arose between the 2 researchers regarding inclusion decisions, a third researcher (WCL) was consulted to participate in discussions. Where necessary, the third researcher made the final determination on whether to include a study. For studies lacking medication adherence data or where full-text access was unavailable, we first attempted to contact the corresponding author to obtain the data. If the author did not respond or the data could not be obtained, the study was excluded from the meta-analysis.

Data Extraction

For included studies, data were extracted using Microsoft Excel software, encompassing the following: (1) general information (eg, first author's name, nationality, publication year, theoretical framework, study design, and background); (2) characteristics of patients with cancer (eg, sample size, mean age/age range, and disease type); (3) intervention characteristics (eg, intervention format, intervention content, intervention duration, and intervention frequency); and (4) primary outcomes and measurement tools.

Definition and Classification of Outcome Indicators

Primary Outcome Indicators

Medication adherence was defined as the degree of alignment between a patient's actual medication-taking behavior and the prescribed treatment regimen.

To clarify how this metric was measured across studies, we documented the following three operational dimensions during data extraction:

XSL•FO
RenderX

1. Measurement tools: Categorized into 1 of 5 types: electronic monitoring devices (eg, smart pillboxes), pharmacy dispensing/dispensing records, standardized self-assessment scales (eg, Morisky scale), researcher-designed questionnaires or interviews, and pill counting methods.
2. Data format and definition: Extracted each study's specific criteria for defining "adherence" (eg, "number of missed doses in the past 7 days" or "proportion of days adhering to the prescription ≥80%") and recorded the reporting format of results (continuous variable or dichotomous variable).
3. Processing procedures: For continuous data (eg, mean adherence score), directly extract mean, SD, and sample size; for dichotomous data (eg, adherence rate), extract the number of adherence events and the total sample size.

Based on this classification, we conducted separate meta-analyses for studies reporting continuous and dichotomous variables. This approach avoids introducing additional assumptions and potential bias from forced data format conversion. When pooling continuous data, since different adherence scoring tools (with varying scales) were used across studies, we selected the standardized mean difference (SMD) as the pooled effect measure. This allowed all studies to be included in a unified statistical comparison model. It is important to emphasize that while SMD primarily addresses statistical comparability, it does not eliminate potential conceptual differences underlying varying operational definitions. Such differences may influence the clinical interpretation of results, necessitating caution in analysis. For dichotomous variables, we selected the odds ratio (OR) as the effect measure owing to its greater prevalence and interpretability in epidemiological research.

Secondary Outcome Indicators

The secondary outcome indicators were as follows:

1. Symptom burden: The overall experience of patients with cancer regarding symptoms arising from the disease and treatment in terms of frequency, severity, and functional impact [19].
2. Self-efficacy: The strength of confidence patients with cancer have in their ability to successfully perform and adhere to specific disease management and treatment-related behaviors [20].
3. Service satisfaction: The level of satisfaction with health care services among patients with cancer [21].
4. Health literacy: The ability of patients with cancer to access, understand, evaluate, and apply health information to make appropriate health decisions for maintaining and promoting their own health [22].

Risk of Bias, Level of Evidence, and Methodological Quality Assessment

Two researchers (QYZ and JHY) assessed the risk of bias in RCTs using the 7 domains of The Cochrane Collaboration's tool for risk of bias assessment (version 2.0) [23]. The Cochrane risk of bias assessment tool encompasses 5 key domains: randomization process, intervention deviation, missing data, outcome assessment, and selective reporting. Risks within each domain are categorized as low risk, high risk, or unclear risk. Studies meeting low risk in all domains are classified as grade A, those meeting criteria in some domains are classified as grade B, and those failing to meet criteria in any domain are classified as grade C. This systematic review included only grade A or B literature, and evidence quality was assessed using GRADE Pro software (Evidence Prime Inc). This tool categorizes evidence quality into 4 levels: high, moderate, low, and very low. Factors leading to downgrading encompass 5 aspects: risk of bias, imprecision, inconsistency, indirectness, and publication bias. High-quality evidence indicates very high certainty that the true effect is close to the estimated effect; moderate-quality evidence reflects moderate confidence in the effect estimate, with the true effect potentially close to but possibly substantially different from the estimate; low-quality evidence indicates limited certainty about the effect estimate, with the true effect likely substantially different from the estimate; and very low–quality evidence indicates almost no confidence in the effect estimate, with the true effect highly likely to be substantially different from the estimate [24]. When 2 researchers (QYZ and JHY) disagreed on the risk of bias or evidence grade of a study, a third researcher (WCL) was consulted. If necessary, the third researcher made the final decision on whether to include the study.

Data Analysis

A narrative synthesis approach was used to summarize study characteristics, including research design, publication year, country, participant characteristics, and intervention features. For outcome measures unsuitable for meta-analysis, a narrative synthesis approach was also applied for analysis.

Although the preregistration protocol specified the use of Stata (Version 17.0; StataCorp) and RevMan (Version 5.4; Cochrane) for statistical analysis, the final statistical analysis was conducted in RStudio (Posit) using a conservative random-effects model based on the assumption of between-study heterogeneity. Effect sizes were pooled using the Hartung-Knapp-Sidik-Jonkman (HKSJ) method [25]. To enhance methodological rigor and clinical interpretability, multiple analytical measures were introduced during the optimization phase: the Egger test was used to assess small-sample effects, 95% prediction intervals (PIs) were incorporated, τ^2 was used to quantify heterogeneity, and the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) approach was used to assess evidence quality [26].

For continuous variables, the SMD was used to calculate the pooled effect size, interpreted according to Cohen *d* criteria: 0.2, 0.5, and 0.8 represent small, moderate, and large effects, respectively. For dichotomous variables, the OR was used. Each pooled effect size is accompanied by a 95% CI. Additionally, we report PIs to quantify the variability of actual effects across different populations or settings [27].

The Cochrane Collaboration recommends the HKSJ method for random-effects meta-analysis [28]. When facing insufficient statistical power due to a small number of included studies, random-effects meta-analysis based on the HKSJ method outperforms the standard DerSimonian-Laird method [27]. By

XSL•FO
RenderX

using a t-distribution and adjusting SEs using the q statistic, the HKSJ method provides more robust effect size CI estimates. Particularly with limited studies, HKSJ-estimated CIs are more accurate as they better reflect the inherent uncertainty and imprecision of effect estimates [29]. Model selection should be based on conceptual assumptions about the true effect distribution rather than descriptive statistics like I^2 [26]. Given the presence of differences in participant characteristics, interventions, and measurement tools across the included studies, a conservative random-effects model was used for all analyses. This systematic review used restricted maximum likelihood estimation to account for between-study variance, thereby more accurately reflecting the impact of true heterogeneity on effect estimates. Heterogeneity was quantified using Q-tests, I^2, and τ^2 measures [27]. All statistical analyses were conducted in the R Studio environment using R software (version 4.5.2), utilizing packages including metafor, forestplot, and ggplot2.

Subgroup analyses of postintervention data were based on priority intervention criteria: (1) intervention duration, (2) cancer type, and (3) intervention type. Sensitivity analyses were conducted to assess result stability. Funnel plots were generated, and Egger regression intercept tests were applied to examine small-sample effects. Note that these methods aim to detect small-sample effects rather than directly assess publication bias [30,31]. Small-sample effects may stem from publication bias, selective reporting, genuine heterogeneity between studies, and methodological differences. According to methodological standards, funnel plots require at least 10 studies for reliable interpretation; otherwise, the reliability of interpretations is low [30]. Egger regression intercept analysis was applied to all outcome measures in this systematic review.

Deviations From the Registered Protocol

This systematic review strictly followed the protocol registered in advance on the PROSPERO platform (CRD42024522261). During methodological refinement and peer review, we identified the following limitations in the original protocol and made two subsequent revisions accordingly:

1. Optimization of the search strategy: The original protocol initially proposed a search strategy combining subject headings with free-text terms. Following peer and expert review, we adopted recommendations to “expand search terms to reduce the risk of omission.” This involved supplementing search terms, such as adding “online*” and “drug therapy computer assisted,” to enhance search comprehensiveness.
2. Change in the statistical analysis tool: Although the original protocol specified Stata (version 17.0) and RevMan (version 5.4) for data analysis, during implementation, we ultimately selected RStudio to perform all statistical analyses. This decision was made to leverage RStudio’s flexibility in data visualization and statistical model implementation.

Results

Literature Search Results

This literature search identified a total of 848 studies, including 846 from databases and clinical trial registries and 2 from reference lists. After deduplication, 703 studies were ultimately retained. Following preliminary screening of titles and abstracts against the inclusion criteria, 30 studies underwent full-text assessment. After full-text evaluation, 13 studies were excluded: 6 due to incomplete data, 5 due to unavailable full texts, and 2 due to differing outcome measures. Ultimately, 17 studies [19-22,32-44] were included (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

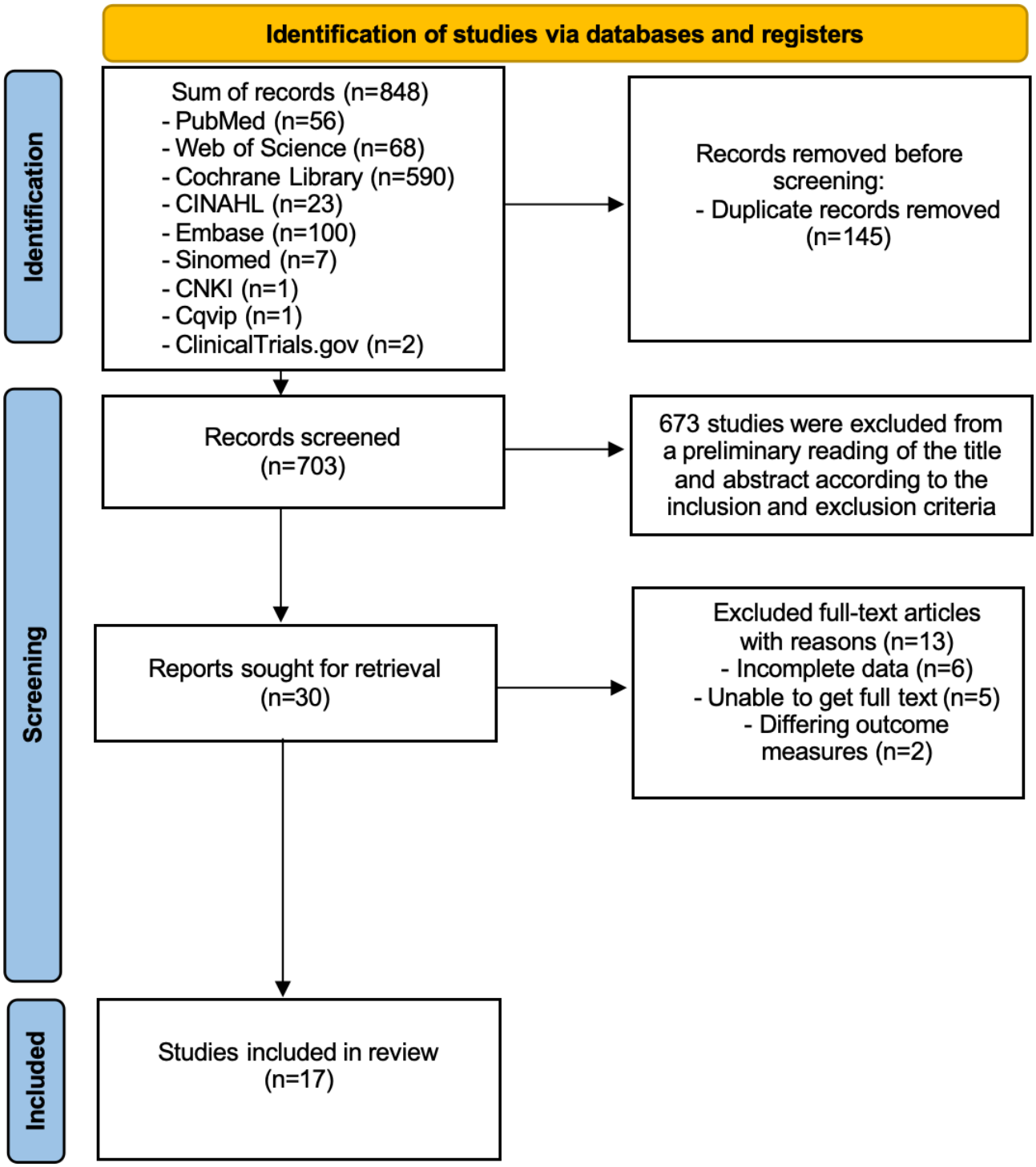

Characteristics of the Included Studies

This systematic review included 17 RCTs. Six studies were published in Simplified Chinese [20,21,36,37,42,44], while the remaining 11 were English-language publications [19,22,32-35,38-41,43], spanning publication dates from 2016 to 2025. Studies originated from 8 countries: the United States, China, Turkey, Australia, the Netherlands, South Korea, Iran, and Finland, and all were conducted in hospital and home settings.

The studies included a total of 1309 participants, with sample sizes ranging from 28 to 166 cases. Cancer types included breast cancer (3 studies) [33,39,43]; leukemia (4 studies) [22,35,40,41]; unspecified cancer types (7 studies) [19,32,34,37,38,42,44]; and thyroid cancer, pheochromocytoma, and prostate cancer (1 study each) [20,21,36].

Regarding intervention formats, the primary methods used were mobile apps (10 studies) [19,33,34,36,37,40-44], text messaging services (5 studies) [20,32,35,38,39], and websites (2 studies) [21,22]. Additionally, 2 studies utilized smart pill bottles as supplementary tools, integrated with either mobile apps or text messaging services [39,43]. Regarding intervention content, mobile apps and websites primarily enhanced medication adherence through online Q&A sessions, multiplayer online

XSL•FO
RenderX

games, and science-popularization video pushes, while text messaging services focused on delivering personalized medication reminders.

Regarding intervention providers, 7 studies explicitly reported that physicians and/or nurses delivered interventions via mHealth platforms [20,21,37,39,40,42,44], while the remaining 10 did not specify the type of health care professional [19,22,32-36,38,41,43]. The intervention duration ranged from 3 weeks to 9 months, with only 3 studies reporting frequency (daily to 3 times weekly) [20,33,40]. Regarding theoretical frameworks, only 1 study designed interventions based on social cognitive theory [32].

All studies assigned control groups to receive usual care. Twelve studies described usual care components, including health education, medication guidance, and regular follow-up [19-21,32-40], while the remaining 5 did not provide specific details [22,41-44]. Detailed study characteristics are summarized in Table 1.

XSL•FO
RenderX

Table . Characteristics of the included studies.

Author (year), country	Study design	Theory	Study setting	Cancer type	Sample, n	Age (years)	Details of the experimental group (format, content, duration, and frequency)	Control group	Main outcomes	Outcome indicator
Spoelstra et al [32] (2016), United States	RCT[a]	Social cognitive theory	Home	Breast cancer and other cancers	69	Experimental: mean 60.1 (SD 10.1); Control: mean 59.9 (SD 11.2)	Format: text messaging service; Content: text message reminders to take patient cancer pills; Duration: 3 weeks; Frequency: NA[b]	Usual care	The difference in the score between the experimental and control groups was 0.29 for adherence, and the difference was not statistically significant.	Symptom burden: MDASI-C[c]; Medication adherence: pill count
Kim et al [33] (2018), Korea	RCT	__[d]	Hospital	Breast cancer	72	Experimental: 49.8; Control: 52.1	Format: app (mobile game); Content: education for preventing the side effects of anticancer drugs; Duration: 3 weeks; Frequency: 3 times a week, 30 minutes each time	Usual care	The mobile game group showed better drug adherence (MARS[e]; mean 7.6, SD 0.7 vs mean 6.5, SD 0.5; *P*<.001).	Medication adherence: MARS; Symptom burden: BREF[f]
Verweij et al [22] (2023), Netherlands	Cluster RCT	—	Hospital and home	Leukemia	86	Range: 18 - 65	Format: website; Content: provides accurate and easy-to-understand information about the disease, medication, side effects, and the effect on daily life, and optimizes medication compliance; Duration: 6 months; Frequency: NA	Usual care	No statistical difference was found in the mean total score of patient medication compliance between groups (*P*=.922).	Health literacy: eHEALS[g]; Medication adherence: MARS; Symptom burden: EORTC QLQ-CML24[h]

XSL•FO
RenderX

Author (year), country	Study design	Theory	Study setting	Cancer type	Sample, n	Age (years)	Details of the experimental group (format, content, duration, and frequency)	Control group	Main outcomes	Outcome indicator
Greer et al [19] (2020), United States	RCT	—	Hospital and home	Hematologic cancer, lung cancer, breast cancer, glioma, sarcoma, gastric cancer, genital cancers, and melanoma	166	Range: 21-88; Experimental: mean 52.85 (SD 13.74); Control: mean 53.76 (SD 12.08)	Format: app; Content: a personalized medication dosing schedule, an adherence and symptom reporting module, educational resources for symptom management and other cancer-related topics, and Fitbit integration for tracking physical activity; Duration: 12 weeks; Frequency: NA	Usual care	Patient-reported adherence (MMAS-4[i]), symptom burden, quality of life, satisfaction with treatment, and health care utilization did not differ significantly between groups.	Medication adherence: MMAS-4; Symptom burden: MDASI-C; Service satisfaction: FACIT-TS-PS[j]
Wang et al [21] (2020), China	RCT	—	Home	Abdominal aortic aneurysm	82	Range: 36-71; Experimental: mean 51.83 (SD 4.56); Control: mean 52.53 (SD 4.23)	Format: website/app; Content: health monitoring, entering, and querying; Duration: 6 months; Frequency: NA	Usual care	The medication compliance of the observation group was significantly higher than that of the control group ($P<.05$).	Medication adherence: self-made medication adherence questionnaire; Service satisfaction: self-made follow-up satisfaction survey questionnaire

XSL•FO
RenderX

Author (year), country	Study design	Theory	Study setting	Cancer type	Sample, n	Age (years)	Details of the experimental group (format, content, duration, and frequency)	Control group	Main outcomes	Outcome indicator
Karaaslan-Eşer et al [34] (2021), Turkey	Cluster RCT	—	Hospital and home	Gastrointestinal cancers, lung cancer, breast cancer, and renal cancer	84	Experimental: mean 60.33 (SD 9.31); Control: mean 62.14 (SD 9.97)	Format: app; Content: providing information about oral anticancer agents and symptom-specific recommendations; Duration: 6 months; Frequency: NA	Usual care	The mean score of the intervention group increased over the first-, third-, and sixth-month measurements (*P*<.05).	Medication adherence: OCAS[k]; Symptom burden: MSAS[l]
Kekäl et al [35] (2016), Finland	RCT	—	Hospital	Leukemia	68	Range: 25 - 83	Format: text messaging service; Content: patients were offered daily text messages to their mobile phones based on the dosing regimen of their treatment to help them remember the time to take their medication; Duration: 9 months; Frequency: NA	Usual care	The groups were similar in adherence as assessed by MMAS class (*P*>.05).	Medication adherence: MMAS
Wang et al [36] (2022), China	RCT	—	Hospital and home	Thyroid cancer	98	Experimental: mean 48.2 (SD 8.0); Control: mean 46.6 (SD 9.4)	Format: app; Content: send medication reminders; researchers monitor patients' follow-up status in the background and provide timely reminders; Duration: 60 days; Frequency: NA	Usual care	The medication compliance of the study group was significantly better than that of the control group (*P*<.01).	Medication adherence: MMAS-8; Health literacy: Medication Belief Questionnaire

XSL•FO
RenderX

Author (year), country	Study design	Theory	Study setting	Cancer type	Sample, n	Age (years)	Details of the experimental group (format, content, duration, and frequency)	Control group	Main outcomes	Outcome indicator
Che et al [37] (2022), China	RCT	—	Hospital and home	Pheochromocytoma	76	Range: 28-65; Experimental: mean 42.65 (SD 3.79); Control: mean 43.26 (SD 3.92)	Format: app; Content: includes a disease knowledge module, medical interaction module, cutting-edge information module, and appointment follow-up module; Duration: 4 weeks; Frequency: NA	Usual care	The medication adherence score was higher in the intervention group than in the control group (*P*<.05).	Medication adherence: MMAS-8
He et al [20] (2023), China	RCT	—	Hospital and home	Prostate cancer	80	Range: 61-87; Experimental: mean 74.95 (SD 6.78); Control: mean 72.05 (SD 6.56)	Format: text messaging service; Content: at 7:00 every morning, a greeting message and medication reminder are sent to the WeChat group; Duration: 6 months; Frequency: once a day	Usual care	At 1, 3, and 6 months, medication adherence was higher in the education group than in the control group, with statistically significant differences (*P*<.05).	Medication adherence: MMAS-8
Başoğlu et al [38] (2024), Turkey	RCT	—	Home	Colorectal cancer, breast cancer, and others	48	Experimental: mean 72.00 (SD 7.86); Control: mean 70.52 (SD 5.32)	Format: text messaging service; Content: send patients 1 educational video and 4 text messages per week via WhatsApp; Duration: 4 weeks; Frequency: NA	Usual care	There was no significant difference between the mean scores of the intervention and control groups regarding the pretest application of the OCAS (*P*>.05).	Medication adherence: OCAS

XSL•FO
RenderX

Author (year), country	Study design	Theory	Study setting	Cancer type	Sample, n	Age (years)	Details of the experimental group (format, content, duration, and frequency)	Control group	Main outcomes	Outcome indicator
Graetz et al [39] (2024), United States	Cluster RCT	—	Hospital and home	Breast cancer	28	Range: 33-82; Experimental: mean 58.1 (SD 10.1); Control: mean 60.2 (SD 12.3)	Format: smart pill bottle + text messaging service; Content: text message reminders for missed or incorrect doses; smart pill bottle provides real-time medication adherence monitoring; Duration: 90 days; Frequency: NA	Usual care	Among participants who received the intervention, the average adherence rate was 79.1% for usual care participants and 89.0% for PRO[m] participants (*P*=.05).	Medication adherence: Nomi smart pill bottle system; Symptom burden: self-made symptom burden questionnaire
Dang et al [40] (2025), Australia	RCT	—	Home	Leukemia	28	Range: 18 - 65	Format: app; Content: medication-taking schedule, individualized daily medication reminders, and weekly side effect surveys; Duration: 12 weeks; Frequency: once a day	Usual care	There was no difference in the proportion of patients who had optimal adherence (medication rate of adherence ≥90%) between the intervention and control groups (100% for both groups).	Medication adherence: pharmacy refills; Symptom burden: FACT-G[n]; Self-efficacy: PRO

XSL•FO
RenderX

Author (year), country	Study design	Theory	Study setting	Cancer type	Sample, n	Age (years)	Details of the experimental group (format, content, duration, and frequency)	Control group	Main outcomes	Outcome indicator
Hosseinpour et al [41] (2025), Iran	RCT	—	Hospital	Leukemia	70	Range: 10 - 21	Format: app; Content: offers features such as updates (eg, adjusting the volume of alerts, selecting alert types, setting alarms based on dates, and adding games to encourage adolescent engagement); Duration: 8 weeks; Frequency: NA	Usual care	The adherence score in the intervention group differed significantly from that in the control group (*P*<.05).	Medication adherence: MARS
Lin et al [42] (2018), China	RCT	—	Home	Colorectal cancer, lung cancer, liver cancer, and others	97	Range: 29 - 74	Format: app; Content: the app provides medication instructions, including information on usage methods, dosage, and side effects; Duration: 3 months; Frequency: NA	Usual care	Compared with the control group, the observation group demonstrated significantly higher compliance (*P*<.05).	Medication adherence: self-made medication adherence questionnaire; Self-efficacy: CPSS[o]

XSL•FO
RenderX

Author (year), country	Study design	Theory	Study setting	Cancer type	Sample, n	Age (years)	Details of the experimental group (format, content, duration, and frequency)	Control group	Main outcomes	Outcome indicator
Park et al [43] (2022), Korea	RCT	—	Home	Breast cancer	57	Experimental: mean 52.07 (SD 9.34); Control: mean 54.74 (SD 7.87)	Format: app + smart pill bottle; Content: participants were reminded to take their medication as per the beeping and blinking function of the smart pill bottle cap, which indicated the scheduled time; they also received smartphone notifications that the time of taking the medication was recorded automatically; Duration: 4 weeks; Frequency: NA	Usual care	Mean medication adherence rates were higher in the experimental group than in the control group (*P*=.004).	Medication adherence: Pillsy mobile app; Self-efficacy: medication self-efficacy scale

XSL•FO
RenderX

Author (year), country	Study design	Theory	Study setting	Cancer type	Sample, n	Age (years)	Details of the experimental group (format, content, duration, and frequency)	Control group	Main outcomes	Outcome indicator
Zhao et al [44] (2023), China	RCT	—	Home	NA	100	Experimental: mean 60.67 (SD 6.59); Control: mean 60.82 (SD 6.76)	Format: app; Content: remind patients to take their medication on time via mobile apps or text messages, and provide guidance on dosage, frequency, and other aspects to ensure patients take their medication accurately and on schedule; Duration: 3 months; Frequency: NA	Usual care	Following the intervention, the Morisky medication adherence scores of patients in the experimental group were higher than the scores of patients in the control group, with a statistically significant difference ($P<.05$).	Self-efficacy: SCS[p]; Medication adherence: MMAS-8

[a]RCT: randomized controlled trial.
[b]NA: not applicable.
[c]MDASI-C: M.D. Anderson Symptom Inventory.
[d]Not reported.
[e]MARS: Medication Adherence Rating Scale.
[f]BREF: The World Health Organization Quality of Life-BREF Scale.
[g]eHEALS: eHealth Literacy Scale.
[h]EORTC QLQ-CML24: European Organization for Research and Treatment of Cancer Quality of Life - Chronic Myeloid Leukemia 24-item.
[i]MMAS: Morisky Medication Adherence Scale.
[j]FACIT-TS-PS: Functional Assessment of Chronic Illness Treatment-Treatment Satisfaction-Patient Satisfaction.
[k]OCAS: Oral Chemotherapy Adherence Scale.
[l]MSAS: Memorial Symptom Assessment Scale.
[m]PRO: patient-reported outcomes.
[n]FACT-G: Functional Assessment of Cancer Therapy-General.
[o]CPSS: Chronic Pain Self-Efficacy Scale.
[p]SCS: Self-Control Scale.

Risk of Bias in Studies

Ten studies reported randomization methods, while 5 described allocation concealment methods. Since improving medication adherence among patients with cancer using mHealth technology constitutes a behavioral intervention, blinding participants and researchers proved challenging. Consequently, none of the studies used blinding, resulting in a high risk of performance bias. Furthermore, as outcome measures were primarily assessed using subjective scales, the risk of measurement bias remains unclear. Only 2 studies used intention-to-treat analysis to address missing outcome data. No studies demonstrated selective reporting bias or other types of bias, resulting in an overall quality rating of grade B (Figure 2).

XSL•FO
RenderX

Figure 2. Risk of bias. (A) Risk of bias graph; (B) Risk of bias summary [19-22,32-44].

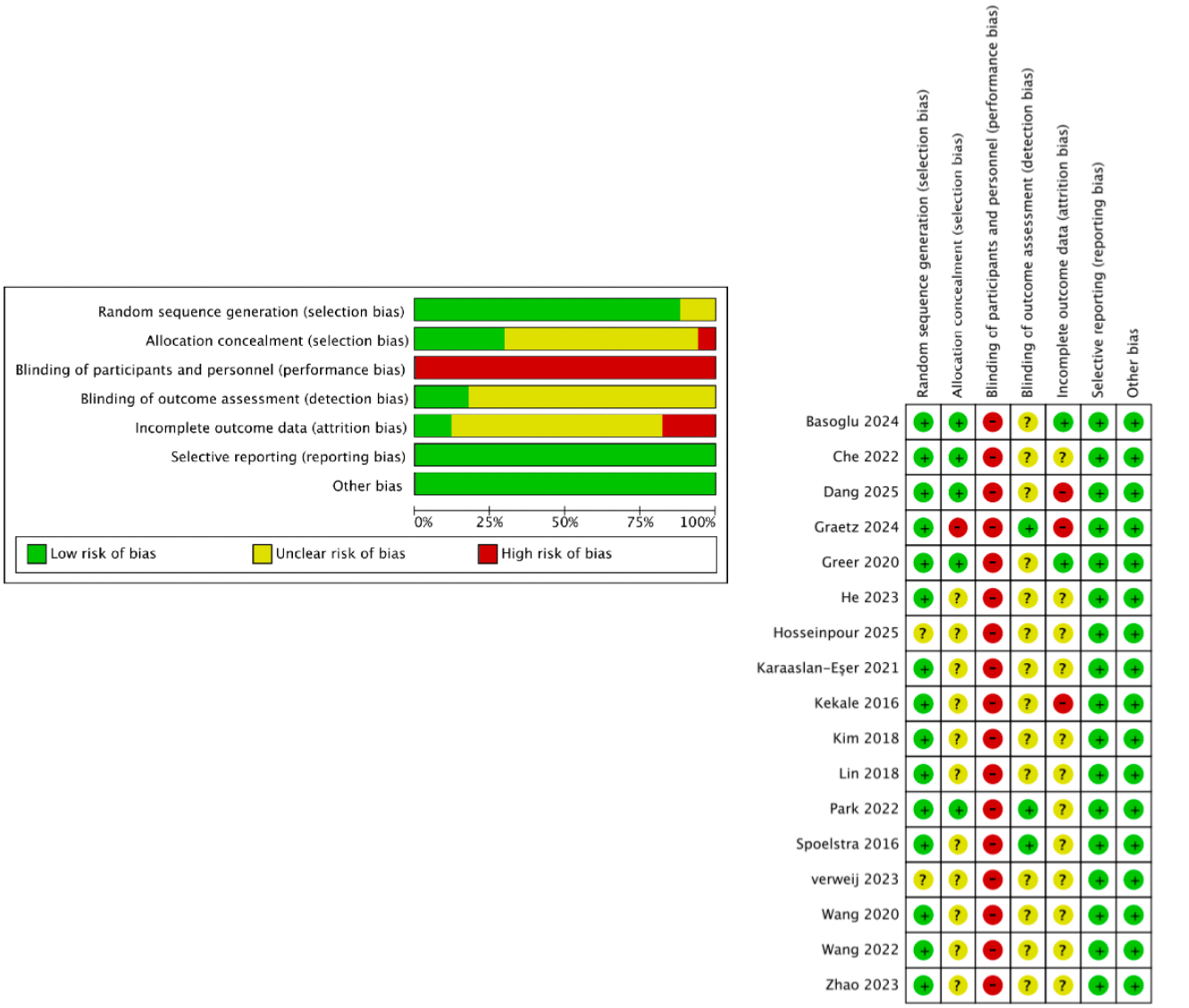

Evidence Quality Assessment

This systematic review used the GRADE evidence rating tool to assess the quality of evidence for each outcome measure. Overall, evidence quality ranged from moderate to very low. For the impact of mHealth interventions on medication adherence rates, the evidence quality was moderate, downgraded primarily due to risk of bias and imprecise data. For the effect of mHealth on the mean medication adherence score, the evidence quality was very low, downgraded due to high risk of bias, inconsistent results, and imprecise data. For the effect of mHealth on symptom burden, the evidence quality was low, downgraded due to risk of bias and imprecise data.

Regarding the effect of mHealth on self-efficacy, the evidence quality was also low, downgraded due to risk of bias and imprecision. The GRADE evidence summaries for each outcome are presented in Table 2, with detailed downgrade justifications available in Multimedia Appendix 2.

XSL•FO
RenderX

Table . GRADE (Grading of Recommendations Assessment, Development, and Evaluation) evidence.

Outcomes	Studies (design), n	Risk of bias	Inconsistency	Indirectness	Imprecision	Publication bias	Patients[a]		Effect		Certainty
							Mobile health	Usual care	Relative, OR[b] (95% CI)	Absolute, value (95% CI)	
Medication adherence rate	7 (RCT[c])	Serious	Not serious	Not serious	Serious	Undetected	269/313 (85.9)	190/295 (64.4)	3.47 (1.92-6.26)	219 more per 1000 (from 132 more to 275 more)	Moderate
Mean medication adherence scores	10 (RCT)	Serious	Serious	Not serious	Serious	Undetected	361	340	—[d]	SMD[e] 1.01 (0.51 to 1.52)	Very low
Symptom burden	7 (RCT)	Serious	Not serious	Not serious	Serious	Undetected	267	234	—	SMD −0.38 (−0.61 to −0.14)	Low
Self-efficacy	6 (RCT)	Serious	Not serious	Not serious	Serious	Undetected	214	191	—	SMD 0.98 (0.25 to 1.72)	Low

[a]Data are presented as n/N (%) or n.

[b]OR: odds ratio.

[c]RCT: randomized controlled trial.

[d]Not applicable.

[e]SMD: standardized mean difference.

Meta-Analysis Results

Effect of mHealth on the Medication Adherence Rate

Seven studies evaluated the impact of mHealth on the medication adherence rate [19,21,32,35,36,39,42]. Despite low heterogeneity (I^2=18.3%; τ^2=0.0658), we used a random-effects model to provide a more conservative and generalizable effect estimate. This decision was based on differences in adherence measures, intervention content, and measurement tools across the included studies. Pooled results demonstrated that mHealth significantly improved the medication adherence rate compared to controls (OR 3.47, 95% CI 1.92-6.26, 95% PI 1.46-8.21; I^2=18.3%; τ^2=0.0658; P=.002; Figure 3). The 95% CI confirmed the statistical significance of the average effect, with low heterogeneity among studies (I^2=18.3%; τ^2=0.0658). Furthermore, the 95% PI was entirely above 1, indicating that the benefit of mHealth on the medication adherence rate is likely reproducible in future studies. Note that the small number of included studies may render this 95% CI unreliable. The quality of evidence regarding mHealth's impact on the medication adherence rate was moderate (Table 2 and Multimedia Appendix 2). Despite downgrading due to risk of bias and imprecision, mHealth still demonstrated a significant effect on the medication adherence rate.

Figure 3. Forest plot of the effect of mobile health on the medication adherence rate [19,21,32,35,36,39,42]. OR: odds ratio.

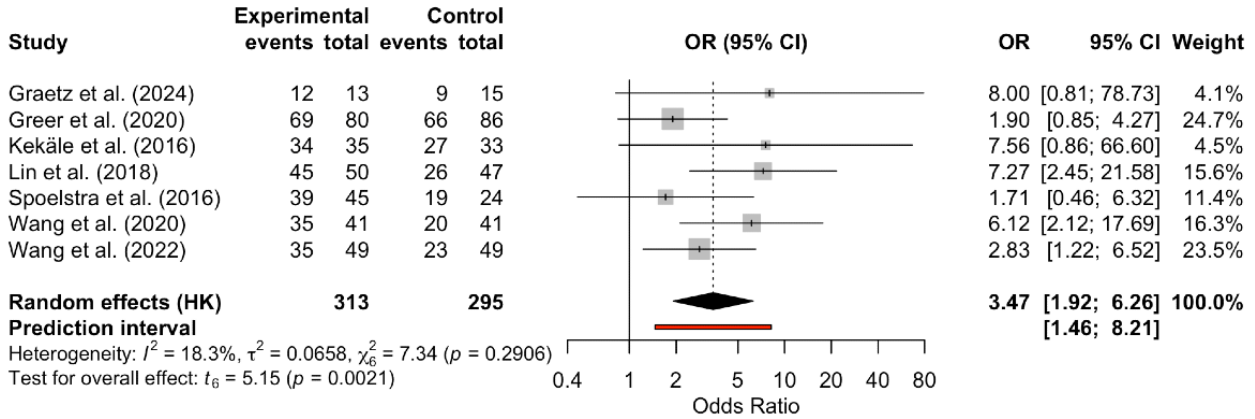

XSL•FO
RenderX

Effect of mHealth on Mean Medication Adherence Scores

Ten studies evaluated the impact of mHealth on mean medication adherence scores [20,22,33,34,37,38,40,41,43,44]. Given the high heterogeneity in the analyses (I^2=83.2%; τ^2=0.3975), we used a random-effects model for pooling to provide a more conservative and generally applicable effect estimate. The pooled analysis revealed that mHealth interventions significantly improved patients' mean medication adherence scores (SMD 1.01, 95% CI 0.51-1.52; I^2=83.2%; τ^2=0.3975; *P*=.001; Figure 4). However, the PI (95% PI −0.50 to 2.52) for this effect estimate was wide and included zero, suggesting that the true effect may vary considerably around zero in future studies. This implies that mHealth may not yield significant effects for certain populations or in different contexts.

Figure 4. Forest plot of the effect of mobile health on mean medication adherence scores [20,22,33,34,37,38,40,41,43,44]. SMD: standardized mean difference.

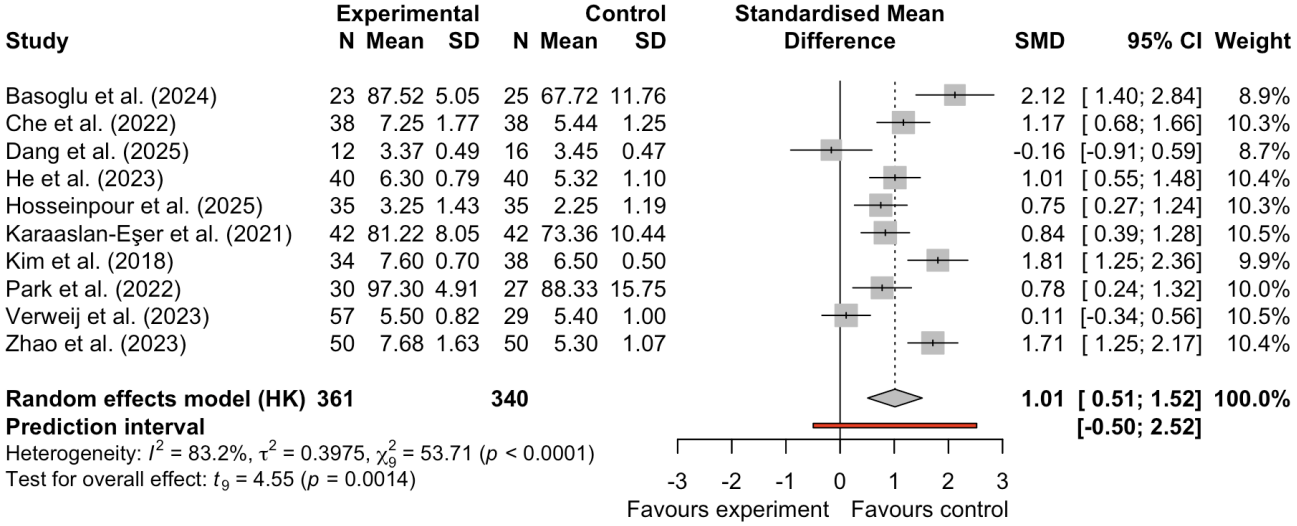

This result should be interpreted with caution. First, the extremely high heterogeneity (I^2=83.2%; τ^2=0.3975) indicates substantial differences in effect sizes across studies, potentially attributable to variations in intervention protocols, study populations, or measurement tools. Second, although the pooled effect was statistically significant, the quality of evidence for this outcome was rated as very low according to the GRADE criteria (Table 2 and Multimedia Appendix 2), primarily due to multiple downgrades for risk of bias, inconsistency, and imprecision. Therefore, while this analysis suggests that mHealth positively impacts average medication adherence scores, the certainty of this conclusion is constrained by the quality and heterogeneity of the existing evidence.

Sensitivity analyses by excluding individual studies showed that excluding the study by Dang et al [40] did not substantially alter the intervention effect on average medication adherence scores (SMD 1.12, 95% CI 0.64-1.60, 95% PI −0.24 to 2.49; I^2=81.9%; τ^2=0.3086; *P*<.001; see Figure S1 in Multimedia Appendix 3), further confirming the robustness of the original meta-analysis results.

To explore the effects of intervention duration, intervention type, and participant characteristics on mean medication adherence scores, we conducted exploratory subgroup analyses. Regarding intervention duration, the subgroup with duration <3 months (SMD 1.37, 95% CI 0.78-1.96; *P*=.001) showed a significantly greater effect than the subgroup with duration ≥3 months (SMD 0.49, 95% CI −0.39 to 1.37; *P*=.006), with a statistically significant between-group difference (χ^2_1=5.98; *P*=.01). Regarding intervention type, significant differences in effect size were observed among text messaging services (SMD 1.53, 95% CI −5.49 to 8.55; *P*=.01), apps (SMD 1.01, 95% CI 0.42-1.61; *P*<.001), and websites (SMD 0.11, 95% CI −0.34 to 0.56), and there were significant differences in effectiveness (χ^2_2=10.28, *P*=.006), with text messaging services yielding the highest point estimate but exhibiting an extremely wide CI. Subgroup differences by cancer type were also statistically significant (χ^2_2=8.86; *P*=.01). The breast cancer subgroup yielded the highest point estimate (SMD 1.29, 95% CI −5.25 to 7.83; *P*=.009), followed by other cancer types (SMD 1.09, 95% CI 0.10-2.08; *P*=.65) and leukemia (SMD 0.28, 95% CI −0.87 to 1.42; *P*=.07). It is important to emphasize that all analyses were post hoc exploratory comparisons, were not adjusted for randomization, and may be influenced by confounding factors (eg, shorter interventions potentially involving more frequent clinician-patient interactions). The observed intergroup differences represent only suggestive correlations and cannot be interpreted as causal relationships. Furthermore, limited numbers of studies and small sample sizes within some subgroups may have resulted in insufficient statistical power and unstable results. Therefore, these findings should be regarded as observational evidence for generating hypotheses and require further validation in future studies.

Effect of mHealth on Symptom Burden

Seven studies evaluated the impact of mHealth technologies on symptom burden in patients with cancer [19,22,32-34,39,40]. Despite low heterogeneity (I^2=4.8%; τ^2=0.0089), we used a random-effects model to obtain a more conservative and generalizable effect estimate. This decision was based on the variability in intervention content and measurement tools across the included studies. Compared with controls, mHealth

XSL•FO
RenderX

interventions yielded modest effects on symptom burden improvement (SMD −0.38, 95% CI −0.61 to −0.14, 95% PI −0.71 to −0.04; I^2=4.8%; τ^2=0.0089; P=.008; Figure 5). The 95% CI confirmed the statistical significance of the pooled effect, with low study heterogeneity (I^2=4.8%). Furthermore, the 95% PI did not cross zero, indicating that the effect of mHealth on symptom burden improvement is likely reproducible in future studies. Note that the small number of included studies may render this 95% PI unreliable. The quality of evidence regarding mHealth's impact on symptom burden was rated as low (Table 2 and Multimedia Appendix 2). This downgrade resulted from risk of bias and imprecision, indicating a small effect of mHealth on symptom burden.

Figure 5. Forest plot of the effect of mobile health on symptom burden [19,22,32-34,39,40]. SMD: standardized mean difference.

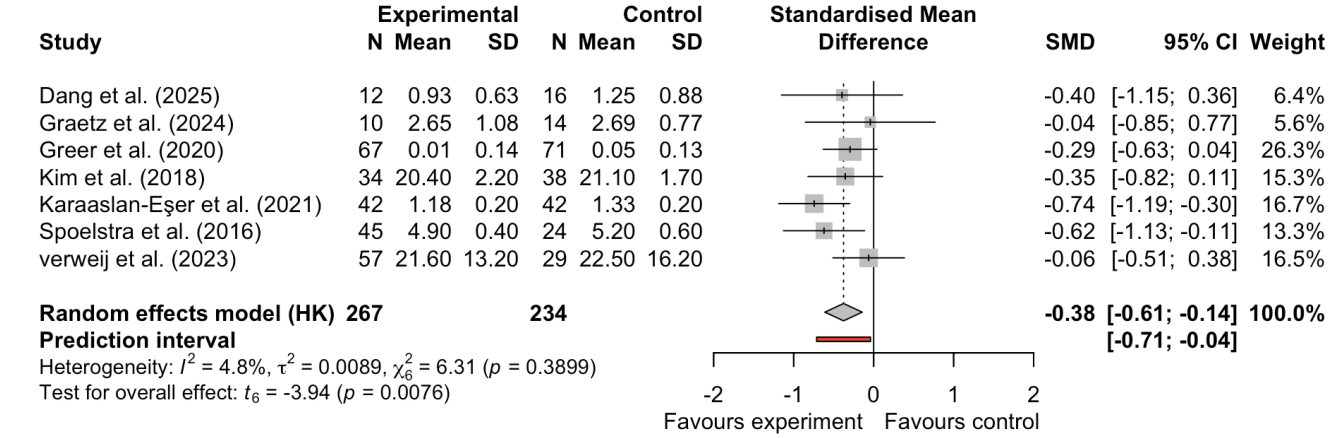

Effect of mHealth on Self-Efficacy

Six studies evaluated the impact of mHealth on self-efficacy among patients with cancer [32,38,40,42-44]. Due to high heterogeneity in this meta-analysis (I^2=75.9%; τ^2=0.2443), we used a random-effects model to obtain a more conservative and generalizable effect estimate. Pooled results demonstrated that mHealth significantly enhanced patient self-efficacy compared with controls (SMD 0.90, 95% CI 0.29-1.51, 95% PI −0.50 to 2.30; I^2=75.9%; τ^2=0.2443; P=.01; Figure 6). The 95% CI confirmed the statistical significance of the average effect, but the 95% PI spanned zero, indicating that mHealth may not improve self-efficacy in future studies or different populations. Note that the small number of included studies may render this 95% PI unreliable. The evidence quality for self-efficacy was low (Table 2 and Multimedia Appendix 2). The evidence rating was downgraded due to risk of bias and imprecision. mHealth still significantly influenced self-efficacy. Sensitivity analysis by excluding individual studies showed that removing the study by Dang et al [40] did not substantially alter the effect of mHealth interventions on mean self-efficacy scores (SMD 0.98, 95% CI 0.25-1.72, 95% PI −0.63 to 2.59; I^2=81.9%; τ^2=0.268; P=.02; Figure S5 in Multimedia Appendix 3), further confirming the robust stability of the original meta-analysis results (Figure 6).

Figure 6. Forest plot of the effect of mobile health on self-efficacy [32,38,40,42-44]. SMD: standardized mean difference.

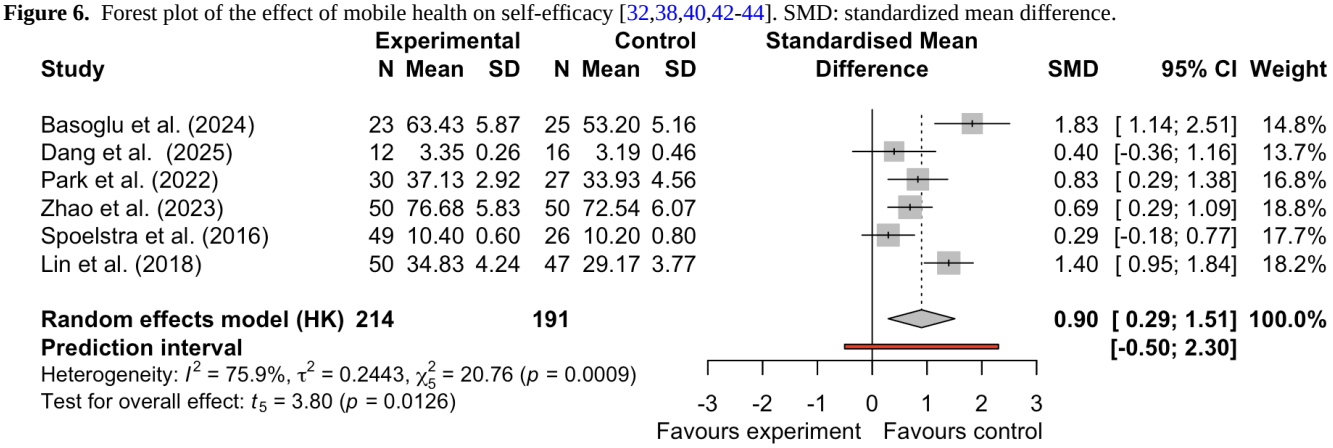

Effect of mHealth on Service Satisfaction

Two studies evaluated the impact of mHealth on service satisfaction, but neither provided the mean and SD data required for a meta-analysis of service satisfaction. After multiple attempts to contact the corresponding authors of these studies, the requested data could not be obtained. Therefore, we conducted a descriptive analysis of the available data [19,21]. Wang et al [21] reported that during the 2-month follow-up period after the intervention, patients were surveyed using a satisfaction questionnaire developed by the hospital's mHealth platform. The results showed significantly higher satisfaction scores in the intervention group compared with the control group. In contrast, Greer et al [19] used the validated Functional Assessment of Chronic Illness Treatment-Therapy Satisfaction-Patient Satisfaction (FACIT-TS-PS) scale to assess

XSL•FO
RenderX

patient satisfaction, finding no statistically significant difference between intervention and control group scores.

Effect of mHealth on Health Literacy

Two studies assessed health literacy levels [22,36]. Despite low heterogeneity (I^2=5.5%; τ^2=0.0028), given the differing measurement tools used across studies (Verweij et al [22] assessed patient health literacy using the validated eHealth Literacy Scale; Wang et al [36] used the validated Medication Beliefs Questionnaire), we used a random-effects model to obtain a more conservative and generalizable effect estimate. The meta-analysis revealed no statistically significant effect of interventions on health literacy (SMD 0.51, 95% CI −1.50 to 2.52; I^2=5.5%; τ^2=0.0028; *P*=.29; Figure 7). However, this result must be interpreted with caution for the following primary reasons. First, the inclusion of only 2 studies severely limited the statistical power of the analysis and the reliability of its conclusions. Second, the PI (95% PI −1.61 to 2.63) and heterogeneity parameter (τ^2=0.0028) calculated from this limited number of studies were highly unstable, rendering them incapable of providing meaningful estimates for future study effects or true heterogeneity.

Figure 7. Forest plot of the effect of mobile health on health literacy [22,36]. SMD: standardized mean difference.

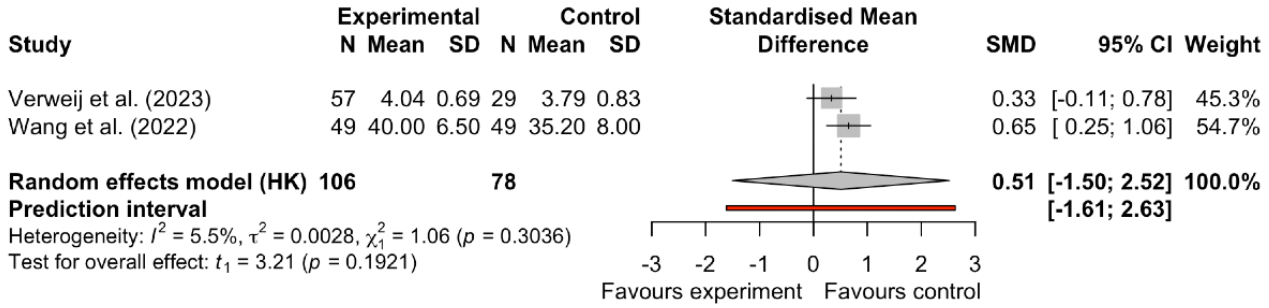

Small Study Effects Assessment

To assess the small-sample effect, we followed the methodological guidance by Sterne et al [30], combining visual inspection of funnel plots with the Egger regression intercept test to analyze 4 outcome measures: mean medication adherence scores, medication adherence, symptom burden, and self-efficacy.

It should be noted that, except for the mean medication adherence score, fewer than 10 studies were included for each of the other outcome measures, which is below the ideal threshold for reliable assessment proposed by Lau et al [45]. Nevertheless, funnel plots were constructed for each measure (Figures S6-S9 in Multimedia Appendix 3) for preliminary exploration. Visual inspection revealed no obvious asymmetry in the funnel plots; however, this visual assessment itself carries considerable uncertainty due to the limited number of studies.

Subsequent quantitative analysis using the Egger test also failed to indicate statistically significant small-sample effects (medication adherence average score: z=1.64; *P*=.75; symptom burden: z=0.18; *P*=.91; self-efficacy: z=1.29; *P*=.78; and medication adherence rate: z=1.53; *P*=.26).

However, these findings must be interpreted with caution. The primary limitation lies in the insufficient number of included studies. First, a small sample size significantly reduces the statistical power of the Egger test, increasing the risk of type II error (ie, failing to detect a genuinely existing small-sample effect) [30]. Second, neither the symmetry of the funnel plot nor the nonsignificance of the Egger test provides sufficient evidence to rule out publication bias [45]. Furthermore, funnel plot asymmetry itself may stem from clinical heterogeneity or differences in methodological quality rather than specifically indicating publication bias. Therefore, based on the current limited data, the findings of this systematic review cannot rule out the possibility of potential publication bias or other forms of selective reporting bias.

Discussion

Summary of Research Findings

This systematic review aimed to systematically evaluate the impact of mHealth interventions on medication adherence among patients with cancer. The results indicate that mHealth interventions can improve medication adherence rates and mean adherence scores among patients with cancer. They also demonstrate positive effects in reducing symptom burden, enhancing self-efficacy, and increasing service satisfaction. However, mHealth showed no significant impact on health literacy. Subgroup analysis results indicate that the effectiveness of mHealth interventions is influenced by intervention timing, intervention type, and cancer type. Furthermore, variations in the intervention protocols and assessment tools used across the included studies may have impacted the reliability of the findings.

mHealth Effectively Improves Medication Adherence

Regarding the medication adherence rate, the 95% PI following the mHealth intervention ranged from 1.46 to 8.21, indicating that the benefits of mHealth for the medication adherence rate are reproducible in future studies. In terms of the mean medication adherence score, the mHealth intervention demonstrated a significant effect, consistent with previous research findings [46]. From the perspective of specific effects and real-world effect variability, the 95% PI for the mean medication adherence score ranged from −0.50 to 2.52, encompassing zero. This indicates that while this study demonstrates a positive impact of mHealth interventions on the mean medication adherence score, the effect may not be significant in certain populations or different contexts.

XSL•FO
RenderX

Furthermore, the included studies utilized specific channels, such as mobile apps, websites, and text messaging services, to deliver knowledge and skills related to oral medication through diverse formats, including real-time online Q&A sessions, multiplayer online games, educational videos, and push notifications. These interventions also provided timely medication reminders. Both the content and methods effectively enhanced medication adherence among patients with cancer by addressing intrinsic knowledge and skill acquisition as well as extrinsic behavioral supervision. Based on this, it can be inferred that improved medication adherence contributes to optimizing the overall effectiveness of cancer treatment, such as reducing symptom burden and enhancing patient self-efficacy. Data analysis results also support this inference, showing that mHealth interventions not only significantly reduced patients' symptom burden but also improved their self-efficacy and service satisfaction. However, the impact of mHealth interventions on health literacy was not statistically significant (P=.29), contradicting previous research findings [47]. This discrepancy may stem from the inclusion of only 2 studies in the health literacy outcome analysis, severely limiting statistical power and reducing the reliability of conclusions. Finally, this systematic review used subgroup analyses to examine the effects of intervention format, timing, and cancer type on mHealth intervention outcomes.

Text Messaging Services Have the Best Intervention Outcomes

Mobile apps and websites offer health education information in diverse formats, including videos, audio, and text. Moreover, the information exchange between these platforms and patients with cancer—that is, between information providers and recipients—is often bidirectional and interactive. In contrast, text messaging services, which are somewhat outdated, provide information in a relatively monotonous format dominated by text, and the information transmission through this channel is predominantly one-way [48]. Theoretically, mobile apps and websites should yield superior intervention outcomes compared with text messages. However, our subgroup analysis revealed that among the 3 mHealth intervention formats, text messaging services demonstrated greater effectiveness than mobile apps and websites. The findings of this systematic review align with the research by Boima et al [49], which demonstrated that text messaging interventions were more effective than mobile apps and web-based interventions in reducing medication nonadherence events among patients with cardiovascular diseases.

The underlying reason for this discrepancy may be related to the age characteristics of patients with cancer. As shown in Table 1, the average age of patients included in this systematic review was relatively high, with 6 studies having an average age exceeding 55 years [20,32,34,38,39,44] and 2 studies having an average age exceeding 70 years [20,38]. Although text messaging services represent a simple, one-way communication method, their content is often concise, clear, and straightforward, making it easier to understand and implement. This characteristic reduces cognitive load during information reception and processing, potentially aligning better with the information format preferences and cognitive functional characteristics of older patients with cancer. Furthermore, compared with mobile apps and websites, text messaging services impose lower demands on mobile devices and network connectivity, enhancing their practical feasibility and applicability. These factors may explain why text messaging interventions outperformed mobile app and website interventions among patients with cancer, who are predominantly older adults.

This finding also raises the following questions: Have existing mHealth interventions delivered via mobile apps and websites sufficiently considered the cognitive characteristics and information preferences of older adults during their design and development? Could this oversight inadvertently contribute to and exacerbate the so-called digital divide?

The Optimal Intervention Duration for mHealth is Less Than 3 Months

This systematic review conducted subgroup analyses based on different intervention durations using 3 months as the cutoff point. The selection of this cutoff point was primarily based on 2 considerations. First, from the perspective of clinical treatment patterns, the duration of oral therapy for patients with cancer typically ranges from 12 to 24 weeks. Three months corresponds to the midpoint of the treatment course, marking a critical transition from the early to the later stages of therapy [50]. Second, prior studies indicate that medication adherence often declines around 3 months after initiation due to factors like drug side effects and forgetfulness [8]. Subgroup analysis revealed that mHealth interventions lasting <3 months significantly improved the mean medication adherence scores of patients with cancer compared with those lasting ≥3 months. The findings of this systematic review align with the study by Al-Arkee et al [51], which demonstrated that mHealth interventions lasting <3 months were more effective than those lasting ≥3 months in improving the average medication adherence scores of patients with cardiovascular diseases.

From a behavioral science perspective, the superior effectiveness of mHealth interventions lasting <3 months, compared with those lasting ≥3 months, may stem from the behavioral habituation cycle. Research by Lally et al [52] indicates that the median time required to establish a new behavior is 66 days. This suggests that the critical phase for behavioral acquisition occurs mainly within the first 2 to 3 months. During this window, mHealth interventions can effectively promote the formation and consolidation of medication-taking behaviors. They do so by providing timely reminders, feedback, and external frameworks. Once the target behavior is automated, continued external prompts may offer diminishing marginal benefits [53]. After habit formation, long-term interventions (≥3 months) typically offer few additional benefits. In some cases, they may even reduce effectiveness by reinforcing external dependence or causing intervention fatigue. In summary, mHealth interventions of <3 months match the natural process of habit formation. By maximizing gains during the most sensitive adaptation period, they significantly improve medication adherence.

XSL•FO
RenderX

mHealth Interventions Show the Most Effective Outcomes in Improving Medication Adherence Among Patients With Breast Cancer

This systematic review included subgroup analyses for different cancer types. Notably, the results indicate that mHealth interventions demonstrate greater efficacy in improving medication adherence among patients with breast cancer compared to patients with leukemia and those with other cancer types. However, no studies have specifically examined the impact of cancer type on the effectiveness of mHealth interventions in enhancing medication adherence. Therefore, future research should conduct in-depth analyses of the differential effects of such interventions across various cancer types.

Limitations

This systematic review has several limitations. First, it included only 17 studies, and the small sample size may lead to imprecise 95% CIs and biased estimates, thereby weakening the reliability of the results. Additionally, the limited number of studies resulted in insufficient statistical power, potentially overlooking potential small-sample effects. Although the existing evidence is insufficient to prove the existence of small-sample effects, we cannot completely rule out the influence of publication bias or selective reporting bias.

Second, among the 17 included studies, only the study by Spoelstra et al [32] used a theoretical model (social cognitive theory) to guide the design of the mHealth platform. Interventions targeting medication adherence are fundamentally interventions targeting medication-taking behavior. From a behavioral science perspective, it is necessary to leverage well-established health behavior theoretical models to explore and analyze the underlying logic and influencing factors of medication adherence among patients with cancer. This approach aims to uncover the core drivers behind medication-taking behavior, thereby enabling the development of more precise and effective medication adherence intervention strategies.

Furthermore, all 17 included studies used uniform interventions for all patients. No tailored interventions were designed or implemented based on individual patient characteristics. Future design of mHealth strategies for cancer medication adherence should incorporate methods, such as latent profile analysis, user profiling, and in-depth interviews, to thoroughly analyze patient traits and needs. Integrating patient profiles into platform design will enable personalized, precision-based scientific interventions [54,55].

Conclusions

Building upon previous systematic reviews, this systematic review systematically reviewed and comprehensively analyzed the impact of mHealth interventions on medication adherence among patients with cancer. Overall, existing evidence suggests that mHealth interventions hold potential advantages in improving medication adherence, self-efficacy, and service satisfaction, as well as reducing symptom burden. However, considerable heterogeneity exists across certain outcome measures, and the overall evidence quality remains moderate to low. Additionally, the included studies exhibited a moderate risk of bias, necessitating cautious interpretation of these findings. It is currently inconclusive to assert superiority over conventional care.

Despite these limitations, the significance of this systematic review lies not only in confirming the efficacy of mHealth interventions for improving medication adherence among patients with cancer but also in revealing their strategic value within the context of digital health care transformation. As chronic disease management shifts from hospital-centered to patient-centered models, mHealth care offers a viable pathway for achieving continuous care, personalized support, and low-cost, scalable interventions. Future research should enhance methodological rigor by conducting multicenter, large-sample, high-quality RCTs. These studies should explore the long-term efficacy and cost-effectiveness of mHealth care across diverse health care settings and patient populations to clarify its role and value within comprehensive cancer management systems.

Acknowledgments

We thank all contributors and reviewers for their insights.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this systematic review.

Data Availability

All data supporting the findings of this systematic review are included in the main text and its supplementary information. Further details or additional datasets can be obtained from the corresponding author upon reasonable request.

Authors' Contributions

WCL participated in conceptualization, data curation, formal analysis, investigation, methodology, writing – original draft, and writing – review and editing. LPW participated in conceptualization, supervision, and writing – review and editing. QYZ participated in formal analysis, investigation, methodology, visualization, and writing – original draft. JHY participated in methodology and writing – original draft.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File, 14 KB - jmir_v28i1e85949_app1.docx]

Multimedia Appendix 2
GRADE (Grading of Recommendations Assessment, Development, and Evaluation) evidence quality assessment.
[DOCX File, 15 KB - jmir_v28i1e85949_app2.docx]

Multimedia Appendix 3
Subgroup analysis, sensitivity analysis plots, and funnel plots for all results.
[DOCX File, 1597 KB - jmir_v28i1e85949_app3.docx]

Checklist 1
PRISMA checklist.
[PDF File, 507 KB - jmir_v28i1e85949_app4.pdf]

References

1. Lohi SM, Chindhalore CA, Dakhale G, Gajbhiye S. Cost variation and affordability of oral targeted cancer therapy in India: comparison of branded and janaushadhi products. Cureus 2025 Dec;17(12):e100287. [doi: 10.7759/cureus.100287] [Medline: 41613654]
2. Yan L, Xiaomeng W, Xiaoya Z, Xiaoyun H, Chunlan X. Summary of best evidence for prevention and management of oral chemotherapy drugs exposure in cancer patients at home [Article in Chinese]. Journal of Nursing Science 2025(4):92-96 [FREE Full text] [doi: 10.3870/j.issn.1001-4152.2025.04.092]
3. Yunyi C, Zifang J, Zefan H. Research progress on medication literacy of cancer patients receiving oral antineoplastic drugs [Article in Chinese]. Journal of Nursing Science 2024;39(14):126-129 [FREE Full text] [doi: 10.3870/j.issn.1001-4152.2024.14.126]
4. Cramer JA, Roy A, Burrell A, et al. Medication compliance and persistence: terminology and definitions. Value Health 2008;11(1):44-47. [doi: 10.1111/j.1524-4733.2007.00213.x] [Medline: 18237359]
5. Uyehara N, Nguyen V, Yu S, et al. The impact of patient characteristics and social drivers of health factors on oral oncolytic adherence. J Manag Care Spec Pharm 2025 Mar;31(3):306-320. [doi: 10.18553/jmcp.2025.31.3.306] [Medline: 40021464]
6. Jafari M, Shahverdian A, Sadigh G, Van Etten RA. Impact of patient personality on adherence to oral anticancer medications: an opportunity? JMIR Cancer 2024 Oct 30;10:e57199. [doi: 10.2196/57199] [Medline: 39475848]
7. van den Bemt P, Blijham MY, Ten Broek L, et al. Patient reported medication-related problems, adherence and waste of oral anticancer medication over time. Eur J Hosp Pharm 2025 Aug 21;32(5):421-426. [doi: 10.1136/ejhpharm-2024-004205] [Medline: 39019578]
8. Tongyu G, Yan S, Fang Z. Single-center evaluation of adherence to immunosuppressive drugs and assessment of influencing factors in 152 renal transplant recipients [Article in Chinese]. China Medicine and Pharmacy 2025;15(12):121-124 [FREE Full text] [doi: 10.20116/j.issn2095-0616.2025.12.25]
9. DiMatteo MR. Variations in patients' adherence to medical recommendations: a quantitative review of 50 years of research. Med Care 2004 Mar;42(3):200-209. [doi: 10.1097/01.mlr.0000114908.90348.f9] [Medline: 15076819]
10. Ashraf H, Warren E, Latner JP, et al. Effect of medication adherence and early follow-up on unplanned readmission post-myocardial infarction: quality improvement in a safety net hospital. Crit Pathw Cardiol 2021 Jun 1;20(2):71-74. [doi: 10.1097/HPC.0000000000000231] [Medline: 32657972]
11. Frezzetti D, Caridi V, Marra L, et al. The impact of inadequate exposure to epidermal growth factor receptor-tyrosine kinase inhibitors on the development of resistance in non-small-cell lung cancer cells. Int J Mol Sci 2024 Apr 29;25(9):4844. [doi: 10.3390/ijms25094844] [Medline: 38732063]
12. Xu Q, Xu Y, Liu X, Ma X. The application of mobile health in self-management among patients undergoing dialysis: scoping review. J Med Internet Res 2026 Jan 2;28:e76880. [doi: 10.2196/76880] [Medline: 41481914]
13. Fahrni ML, Saman KM, Alkhoshaiban AS, Naimat F, Ramzan F, Isa KAM. Patient-reported outcome measures to detect intentional, mixed, or unintentional non-adherence to medication: a systematic review. BMJ Open 2022 Sep 19;12(9):e057868. [doi: 10.1136/bmjopen-2021-057868] [Medline: 36123061]
14. Zhou TT, Wang R, Gu SJ, et al. Effectiveness of mobile medical apps in ensuring medication safety among patients with chronic diseases: systematic review and meta-analysis. JMIR Mhealth Uhealth 2022 Nov 22;10(11):e39819. [doi: 10.2196/39819] [Medline: 36413386]

XSL•FO
RenderX

15. Linder LA, Wu YP, Macpherson CF, et al. Oral medication adherence among adolescents and young adults with cancer before and following use of a smartphone-based medication reminder app. J Adolesc Young Adult Oncol 2019 Apr;8(2):122-130. [doi: 10.1089/jayao.2018.0072] [Medline: 30307778]
16. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
17. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
18. Xu X, Wang F, Liao S, Liu J, Xiao L. Tele-cognitive behavioral therapy for the treatment of diabetes-related distress in idividuals with diabetes mellitus: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2025 Dec 24;27:e80476. [doi: 10.2196/80476] [Medline: 41442672]
19. Greer JA, Jacobs JM, Pensak N, et al. Randomized trial of a smartphone mobile app to improve symptoms and adherence to oral therapy for cancer. J Natl Compr Canc Netw 2020 Feb;18(2):133-141. [doi: 10.6004/jnccn.2019.7354] [Medline: 32023526]
20. He Y, Wang Q, Yan S, Wang L, Fu R. Application of mobile health education in improving medication compliance of elderly patients with endocrine therapy for prostate cancer [Article in Chinese]. The Medical Forum 2023;27(24):12-15. [doi: 10.19435/j.1672-1721.2023.24.004]
21. Wang B, Wang H, Sun H, et al. The effects of application of hospital information platform of interactively medical treatment based on “Internet +” on medication compliance and quality of life during follow-up of patients with abdominal aortic aneurysm surgery [Article in Chinese]. China Medical Equipment 2020;17(9):35-39. [doi: 10.3969/J.ISSN.1672-8270.2020.09.008]
22. Verweij L, Ector G, Smit Y, et al. Effectiveness of digital care platform CMyLife for patients with chronic myeloid leukemia: results of a patient-preference trial. BMC Health Serv Res 2023 Mar 8;23(1):228. [doi: 10.1186/s12913-023-09153-9] [Medline: 36890512]
23. Higgins JPT, Altman DG, Gøtzsche PC, et al. The Cochrane Collaboration’s tool for assessing risk of bias in randomised trials. BMJ 2011 Oct 18;343(oct18 2):d5928. [doi: 10.1136/bmj.d5928] [Medline: 22008217]
24. Joanna Briggs Institute Levels of Evidence and Grades of Recommendation Working Party. Supporting document for the Joanna Briggs Institute levels of evidence and grades of recommendation. The Joanna Briggs Institute. 2014. URL: https://jbi.global/sites/default/files/2019-05/JBI%20Levels%20of%20Evidence%20Supporting%20Documents-v2.pdf [accessed 2026-03-23]
25. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
26. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023 Dec;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
27. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019 Jun;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]
28. Higgins JPT, Thomas J, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions, 2nd edition: John Wiley & Sons; 2019. [doi: 10.1002/9781119536604]
29. Fleiss JL. The statistical basis of meta-analysis. Stat Methods Med Res 1993;2(2):121-145. [doi: 10.1177/096228029300200202] [Medline: 8261254]
30. Sterne JAC, Sutton AJ, Ioannidis JPA, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011 Jul 22;343:d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
31. Shi L, Lin L. The trim-and-fill method for publication bias: practical guidelines and recommendations based on a large database of meta-analyses. Medicine (Abingdon) 2019;98(23):e15987. [doi: 10.1097/MD.0000000000015987]
32. Spoelstra SL, Given CW, Sikorskii A, et al. Proof of concept of a mobile health short message service text message intervention that promotes adherence to oral anticancer agent medications: a randomized controlled trial. Telemed J E Health 2016 Jun;22(6):497-506. [doi: 10.1089/tmj.2015.0126] [Medline: 26716365]
33. Kim HJ, Kim SM, Shin H, Jang JS, Kim YI, Han DH. A mobile game for patients with breast cancer for chemotherapy self-management and quality-of-life improvement: randomized controlled trial. J Med Internet Res 2018 Oct 29;20(10):e273. [doi: 10.2196/jmir.9559] [Medline: 30578205]
34. Karaaslan-Eşer A, Ayaz-Alkaya S. The effect of a mobile application on treatment adherence and symptom management in patients using oral anticancer agents: a randomized controlled trial. Eur J Oncol Nurs 2021 Jun;52:101969. [doi: 10.1016/j.ejon.2021.101969] [Medline: 33991868]
35. Kekäle M, Söderlund T, Koskenvesa P, Talvensaari K, Airaksinen M. Impact of tailored patient education on adherence of patients with chronic myeloid leukaemia to tyrosine kinase inhibitors: a randomized multicentre intervention study. J Adv Nurs 2016 Sep;72(9):2196-2206. [doi: 10.1111/jan.12978] [Medline: 27113362]
36. Wang M, Su W, Zhang J, et al. Effects of mobile health intervention on endocrine therapy compliance andthyroid stimulating hormone compliance rate in postoperative patients withthyroid stimulating hormone compliance rate in postoperative

XSL•FO
RenderX

patients withthyroid cancerthyroid cancer [Article in Chinese]. Oncology Progress 2022;20(9):927-930. [doi: 10.11877/j.issn.1672-1535.2022.20.09.15]
37. Che H. The effect of a mobile health-based chronic disease management system on medication adherence, negative emotions, and blood pressure improvement in pheochromocytoma patients [Article in Chinese]. Chinese General Practice Nursing 2022;20(8):1114-1115. [doi: 10.12104/j.issn.1674-4748.2022.08.030]
38. Başoğlu S, Polat Ü. The effect of education and monitoring via tele-nursing to elderly cancer patients using oral anticancer agents on self-efficacy and medication adherence: a randomized controlled trial. Semin Oncol Nurs 2024 Oct;40(5):151692. [doi: 10.1016/j.soncn.2024.151692] [Medline: 39030135]
39. Graetz I, Hernandez S, Arshad S, et al. Leveraging mobile health to improve capecitabine adherence among women with breast cancer: a pilot randomized controlled trial. JCO Oncol Pract 2024 Oct;20(10):1376-1383. [doi: 10.1200/OP.24.00031] [Medline: 38917401]
40. Dang TH, Wickramasinghe N, Jayaraman PP, et al. Digital solution to support medication adherence and self-management in patients with cancer (SAMSON): pilot randomized controlled trial. JMIR Form Res 2025 Feb 19;9:e65302. [doi: 10.2196/65302] [Medline: 39969972]
41. Hosseinpour M, Shirinabadi Farahani A, Varzeshnejad M, Nasiri M. Check the impact of mobile health on medication adherence in adolescents with leukemia. Asian Pac J Cancer Prev 2025 Mar 1;26(3):751-755. [doi: 10.31557/APJCP.2025.26.3.751] [Medline: 40156390]
42. Lin H, Zheng Z, Chen Y, et al. Practical experience in implementing health education for cancer pain patients using the “317 Care” app in the context of “Internet Plus” [Article in Chinese]. Chinese Rural Health Service Administration 2018;38(6):771-773. [doi: 10.3969/j.issn.1005-5916.2018.06.024]
43. Park HR, Kang HS, Kim SH, Singh-Carlson S. Effect of a smart pill bottle reminder intervention on medication adherence, self-efficacy, and depression in breast cancer survivors. Cancer Nurs 2022;45(6):E874-E882. [doi: 10.1097/NCC.0000000000001030] [Medline: 34661562]
44. Zhao D, Gong Y, Ye Q, et al. The application effectiveness of a mobile healthcare-based chronic disease management system in cancer patient treatment [Article in Chinese]. Journal of Frontiers of Medicine 2023;13(21):74-76. [doi: 10.3969/j.issn.2095-1752.2023.21.023]
45. Lau J, Ioannidis JPA, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. BMJ 2006 Sep 16;333(7568):597-600. [doi: 10.1136/bmj.333.7568.597] [Medline: 16974018]
46. Psihogios AM, El-Khatib A, Matos K, et al. Human-centered design of a personalized digital health intervention to improve oral chemotherapy adherence in adolescents and young adults with hematologic cancers. Pediatr Blood Cancer 2025 Aug;72(8):e31756. [doi: 10.1002/pbc.31756] [Medline: 40346736]
47. Zisopoulou S, Kazantzaki G, Adamakidou T, et al. Ehealth literacy, beliefs about medication and adherence in older adults with chronic diseases: a cross-sectional study. Adv Exp Med Biol 2026;1489:187-196. [doi: 10.1007/978-3-032-03394-9_19] [Medline: 41252006]
48. Wiljer D, Catton P. Multimedia formats for patient education and health communication: does user preference matter? J Med Internet Res 2003;5(3):e19. [doi: 10.2196/jmir.5.3.e19] [Medline: 14517110]
49. Boima V, Doku A, Agyekum F, Tuglo LS, Agyemang C. Effectiveness of digital health interventions on blood pressure control, lifestyle behaviours and adherence to medication in patients with hypertension in low-income and middle-income countries: a systematic review and meta-analysis of randomised controlled trials. EClinicalMedicine 2024 Mar;69:102432. [doi: 10.1016/j.eclinm.2024.102432] [Medline: 38333367]
50. Chiavarini M, Giacchetta I, Rosignoli P, Fabiani R. E-health and m-health in obesity management: a systematic review and meta-analysis of RCTs. Nutrients 2025 Jul 1;17(13):2200. [doi: 10.3390/nu17132200] [Medline: 40647304]
51. Al-Arkee S, Mason J, Lane DA, et al. Mobile apps to improve medication adherence in cardiovascular disease: systematic review and meta-analysis. J Med Internet Res 2021 May 25;23(5):e24190. [doi: 10.2196/24190] [Medline: 34032583]
52. Lally P, van Jaarsveld CHM, Potts HWW, Wardle J. How are habits formed: modelling habit formation in the real world. Euro J Social Psych 2010 Oct;40(6):998-1009. [doi: 10.1002/ejsp.674]
53. Pongchaiyakul C, Driessen S. Evaluating the effect of mobile applications “My A:Care” and “Smart Coach” on adherence to lipid-lowering treatment in patients with dyslipidemia: a prospective, randomized, open-label clinical study. Front Digit Health 2025;7:1502990. [doi: 10.3389/fdgth.2025.1502990] [Medline: 40692655]
54. Amico KR, Mugavero M, Krousel-Wood MA, Bosworth HB, Merlin JS. Advantages to using social-behavioral models of medication adherence in research and practice. J Gen Intern Med 2018 Feb;33(2):207-215. [doi: 10.1007/s11606-017-4197-5] [Medline: 29204968]
55. Yuan Q, Yongfeng T, Dinghua Y, et al. Research progress on the application of “patient profile” based on big data in the field of nursing [Article in Chinese]. West China Medical Journal 2025;40(8):1351-1355. [doi: 10.7507/1002-0179.202410248]

Abbreviations:

GRADE: Grading of Recommendations Assessment, Development, and Evaluation

XSL•FO
RenderX

HKSJ: Hartung-Knapp-Sidik-Jonkman
mHealth: mobile health
OR: odds ratio
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SMD: standardized mean difference

Edited by S Brini; submitted 16.Oct.2025; peer-reviewed by Q Wang, SA Kristina; revised version received 26.Feb.2026; accepted 03.Mar.2026; published 30.Mar.2026.

Please cite as:
Liu WC, Wang LP, Zhang QY, Yang JH
Effectiveness of Mobile Health for Improving Medication Adherence in Patients With Cancer: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e85949
URL: https://www.jmir.org/2026/1/e85949
doi:10.2196/85949

© Wei Chao Liu, Li Ping Wang, Qian Yu Zhang, Jing Han Yang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Artificial Intelligence Tools for Automating Evidence Synthesis: Scoping Review

Sashika Harasgama[1], BSc, MScPH, MD; Helen Pearce[1], BA; Cameron Appel[1], BAc, MSc; Liam Loftus[1], BM BCh; Helena Painter[1], DTM&H, BSc, MBChB; Isla Kuhn[2], MA, MSc; Justine Karpusheff[3], PGCE, MSc, PhD; Aji Ceesay[3], BA, MSc; John Ford[1], DTM&H, MBChB, MSc, PhD

[1]Wolfson Institute of Population Health, Queen Mary University of London, Whitechapel Campus, London, United Kingdom

[2]Medical Library, University of Cambridge, Cambridge, United Kingdom

[3]The Health Foundation, London, United Kingdom

Corresponding Author:
Sashika Harasgama, BSc, MScPH, MD
Wolfson Institute of Population Health, Queen Mary University of London, Whitechapel Campus, London, United Kingdom

Abstract

Background: Rapidly and accurately synthesizing large volumes of evidence is a time- and resource-intensive process. Once published, reviews often risk becoming outdated, limiting their usefulness for decision makers. Recent advancements in artificial intelligence (AI) have enabled researchers to automate stages of the evidence synthesis process, from literature searching and screening to data extraction and analysis. As previous reviews on this topic have been published, a significant number of tools have been further developed and evaluated. Furthermore, as generative AI increasingly automates evidence synthesis, understanding how it is studied and applied is crucial, given both its benefits and risks.

Objective: This review aimed to map the current landscape of evaluated AI tools used to automate evidence synthesis.

Methods: Following the Joanna Briggs Institute methodology for scoping reviews, we searched Ovid MEDLINE, Ovid Embase, Scopus, and Web of Science in February 2025 and conducted a gray literature search in April 2025. We included articles published in any language from January 2021 onward. Two reviewers independently screened citations using Rayyan, and data were extracted based on study design and key AI-related technical features.

Results: We identified 7841 unique citations through database searches and 19 records through gray literature searching. A total of 222 articles were included in the review. We identified 65 AI tools and 25 open-source models or machine learning (ML) algorithms that automate parts of or the whole evidence synthesis pathway. A total of 54.1% (n=120) of the studies were published in 2024, reflecting a trend toward researching general-purpose large language models (LLMs) for evidence synthesis automation. The most popular tool studied was generative pretrained transformer models, including its conversational interface ChatGPT (n=70, 31.5%). Moreover, 31.1% (n=69) studied tools automated by traditional ML algorithms. No studies compared traditional ML tools to LLM-based tools. In addition, 61.7% (n=137) and 26.1% (n=58) studied AI-assisted automation of title and abstract screening and data extraction, respectively, the 2 most intensive stages and, therefore, amenable to automation. Technical performance outcomes were the most frequently reported, with only 4.1% (n=9) of studies reporting time- or workload-specific outcomes. Few studies pragmatically evaluated AI tools in real-world evidence synthesis settings.

Conclusions: This review comprehensively captures the broad, evolving suite of AI automation tools available to support evidence synthesis, leveraged by increasingly complex AI approaches that range from traditional ML to LLMs. The notable shift toward studying general-purpose generative AI tools reflects how these technologies are actively transforming evidence synthesis practice. The lack of studies in our review comparing different AI approaches for specific automation stages or evaluating their effectiveness pragmatically represents a significant research gap. Optimal tool selection will likely depend on the review topic and methodology and researcher priorities. While they offer potential for reducing workload, ongoing evaluation to mitigate AI bias and to ensure the integrity of reviews is essential for safeguarding evidence-based decision-making.

(J Med Internet Res 2026;28:e81597) doi:10.2196/81597

KEYWORDS

artificial intelligence; machine learning; large language models; automation; evidence synthesis; systematic reviews as a topic; ChatGPT

XSL•FO
RenderX

Introduction

High-quality evidence synthesis is essential for guiding policy and practice. However, producing such evidence reviews is not only time- and resource-intensive but also challenging to keep up to date due to the volume of studies being published each year. The recent COVID-19 pandemic highlighted the challenges of having accurate, contemporaneous, and rapidly synthesized data, and the opportunities that automation presents in mitigating these [1,2]. Previous studies have estimated that an average systematic review can take approximately 67 weeks to complete [3], often too long for decision makers. The evolving language and text capabilities of artificial intelligence (AI) have increased the scope of automation within the evidence review process, with studies showing that automated tools can help complete systematic reviews in days to weeks [4,5] and significantly reduce workload [6,7].

The advent of evidence synthesis automation was driven by machine learning (ML), the technology that underpins most AI evidence synthesis tools today. ML is a subset of AI that focuses on creating algorithmic systems that can learn from data, recognize patterns, and make decisions, often with model characteristics manually selected by humans [8]. An AI discipline called natural language processing enables computers to understand, interpret, and work with human language (such as written text) and, when used in conjunction with ML, can automate certain language-related tasks [8]. Deep learning is a subset of ML that uses neural networks (computational models that consist of layers of interconnected processing units) to automatically learn more complex patterns from large datasets [8].

Time- and workload-intensive steps within the evidence synthesis pathway, particularly title and abstract screening and data extraction, have been made more efficient through automation tools using machine and deep learning. However, the introduction of generative AI (ie, AI that can autonomously produce text, speech, or other outputs) in the form of commercially available large language models (LLMs), such as ChatGPT (OpenAI) or Claude (Anthropic), has created an opportunity for more creative and complex evidence synthesis automation. Studies are currently trialing LLM-based methods to automate individual stages, such as data extraction, to entire clinical evidence synthesis pipelines [9,10].

LLMs are leveraged by deep learning models called transformers, which use ‘attention’ mechanisms to determine how important each word is in relation to others. They are trained on massive amounts of written data and can capture nuance and context in language more effectively than traditional ML models. Another popular automation approach uses transformers called Bidirectional Encoder Representations from Transformers (BERT), which reads inputs bidirectionally and is pretrained on large sets of text data and then fine-tuned to excel at domain-specific tasks. These newer AI tools may outperform current automation tools in some tasks; however, researchers are still exploring their potential for evidence synthesis, and their performance is yet to be thoroughly validated over time [11,12]. In fact, the use of advanced and generative AI approaches is not without contention and legitimate risks, with the research community fearing loss of academic integrity in the process [11,13,14].

Reviews published to date on this topic have primarily focused on ML tools and approaches used for automating systematic reviews in particular [12,15-17]. For example, Khalil et al [15] performed a scoping review of automation tools for systematic reviews up to mid-2021 and found 10 validated tools that all used ML. Jimenez et al [16] undertook a mapping review of ML tools to assist with systematic reviews and identified 63 tools at the time of publication in December 2022. Khalil et al [18] also conducted a review in 2024 on automation tools for scoping reviews. While these reviews were comprehensive, the rate at which AI technology evolves often means they themselves risk being outdated. A more recent review published by Lieberum et al [12] focused solely on the use of LLMs for performing systematic reviews and found that in half the included studies, LLMs had promising applications, particularly in screening. However, the authors noted that despite the optimism, LLMs are still not quite ready for direct integration into research practice.

The rapid development and likely increasing adoption of generative AI tools to automate evidence synthesis and inform policy and practice warrant a more contemporaneous scoping of the literature. It presents significant risks to the quality and methodological rigor of evidence synthesis, despite its opportunities. Therefore, it is essential to understand the breadth and scope of AI use in this advancing field.

In this scoping review, we aimed to systematically map AI tools available for all types of evidence synthesis, across all its stages, and to describe the current landscape by exploring the underlying automation approaches used by the tools and identifying trends in automated evidence synthesis research.

Methods

Ethical Considerations

Ethical approval was not required for this scoping review, as it involved analysis of previously published studies and did not include primary data collection.

Review Design

We used the Joanna Briggs Institute methodological guidance for scoping reviews [19]. Using the “PCC” mnemonic (population, concept, and context), our review aimed to focus on current AI tools (*P*) used to automate (*C*) the evidence synthesis pathway (*C*). The reporting of this paper was guided by the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist [20], available in Checklist 1. Our review was not registered a priori in PROSPERO.

We aimed to identify studies across steps of the evidence synthesis pathway: (1) searching for the evidence, (2) screening the evidence, (3) extracting the data, (4) assessing the quality of the evidence, (5) analyzing the data, and (6) writing the review. We also aimed to identify which AI methods or techniques were most leveraged for tool development and to

XSL•FO
RenderX

determine which parts of the evidence synthesis pathway were the most automated. Our secondary objectives included discussing outcomes typically used to assess performance.

Search Strategy

The search strategy was adapted from 2 prior reviews' search strategies [15,16] and iteratively refined through an initial search of Ovid MEDLINE, based on analysis of titles, abstracts, and keywords returned as well as cross-referencing with key articles. The final search was executed across four databases: Ovid MEDLINE, Ovid Embase, Scopus, and Web of Science. We searched for studies published from January 1, 2021, to February 14, 2025, with no language restrictions. Details of our search strategy are available in Part S1 in Multimedia Appendix 1. Our search strategy combined three groups of search terms: (1) evidence review and systematic review terms; (2) AI methods, including ML, deep learning, and other techniques; and (3) terms reflecting the act of automation, such as "support," "assist," and "perform." We also conducted a gray literature search in April 2025 after data extraction of our included studies using Google's search engine and citation snowballing methods with the tool Litmaps [21] to identify further relevant studies from our included studies.

Eligibility Criteria

Full details of the eligibility criteria are displayed in Textbox 1. All primary and secondary studies that evaluated the use of AI tools for evidence synthesis in health and care research were included. We did not exclude studies if they did not pertain to health and care research specifically; however, we excluded studies if the AI tool was deemed not transferable to that setting. Studies that did not provide sufficient evaluation of the tool, through technical or other performance-related metrics, were also excluded.

Textbox 1. Eligibility criteria using the population, intervention, comparison, and outcome (PICO) framework.

Inclusion criteria

- Population and setting
  - Primary or secondary studies that evaluate the use of artificial intelligence (AI) tools for evidence synthesis in health and care research
  - Tool developed in any country
- Intervention
  - AI tools defined as having two characteristics: adaptivity and autonomy (based on the Department for Science, Innovation and Technology guidance)
- Comparison
  - If comparison, to manual research methods
  - If comparison, to other automated tools
- Outcome
  - Any technical or performance-related outcome assessing the AI tool
  - Would be dependent on evidence synthesis task but could include outcomes such as
    - Sensitivity, specificity, precision, and area under the curve
    - Number of relevant studies identified in screening
    - Quality of data extraction
    - Accuracy of meta-analysis
    - Quality and accuracy of writing of evidence synthesis report

Exclusion criteria

- Population or setting
  - Studies do not provide any evaluation of AI tools (only descriptions)
- Intervention
  - Automation tools that are not AI or where it is unclear if it is AI

We defined AI according to the regulatory definition provided by the UK Government Department of Science, Innovation and Technology's policy paper [22]. It defines AI as having two distinct characteristics of adaptivity and autonomy, meaning that AI can continually learn and infer patterns not envisioned by human programmers and can make decisions without the intent or oversight of a human. We excluded studies that discussed automated tools but did not display these

XSL•FO
RenderX

characteristics. We also defined evidence synthesis according to the Cochrane definition, which "involves combining information from multiple studies investigating the same topic to comprehensively understand their findings" [23]. No studies were excluded based on comparators or outcomes.

Citations identified through our search were imported and deduplicated using EndNote [24], with Rayyan [25] used to identify further duplicates. Using Rayyan for the entire screening process, records were initially screened via title and abstract by 1 reviewer. Full-text articles were then uploaded and screened for eligibility by another reviewer, with queries or discrepancies resolved by discussion with a third researcher.

Data Extraction and Synthesis

Data from included studies were extracted into a Microsoft Excel template. Extracted information included baseline study characteristics, such as authors, year of publication, study type, tool name, and outcomes measured. We also extracted specific information regarding the AI method used for tool development, web links to the tools or source code if available, as well as paywall features of the tool if relevant. Consistent with Joanna Briggs Institute scoping review methodology, we did not perform a critical appraisal of the included studies. Findings were thematically synthesized and presented visually, alongside being narratively described with descriptive statistics.

Results

Overview

A total of 11,226 studies were retrieved from database searches, with 3385 citations removed as duplicates (Figure 1). A total of 7841 citations were screened by their title and abstract, and 7499 were excluded after being deemed irrelevant. The remaining 342 citations underwent full-text screening, and 208 articles were included based on our eligibility criteria. An additional 19 records were identified through snowballing methods and a gray literature search, with 14 records being included. In total, 222 articles were included in the final review.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram illustrating the study selection process, including the number of records identified, screened, assessed for eligibility, and included in the final review. AI: artificial intelligence.

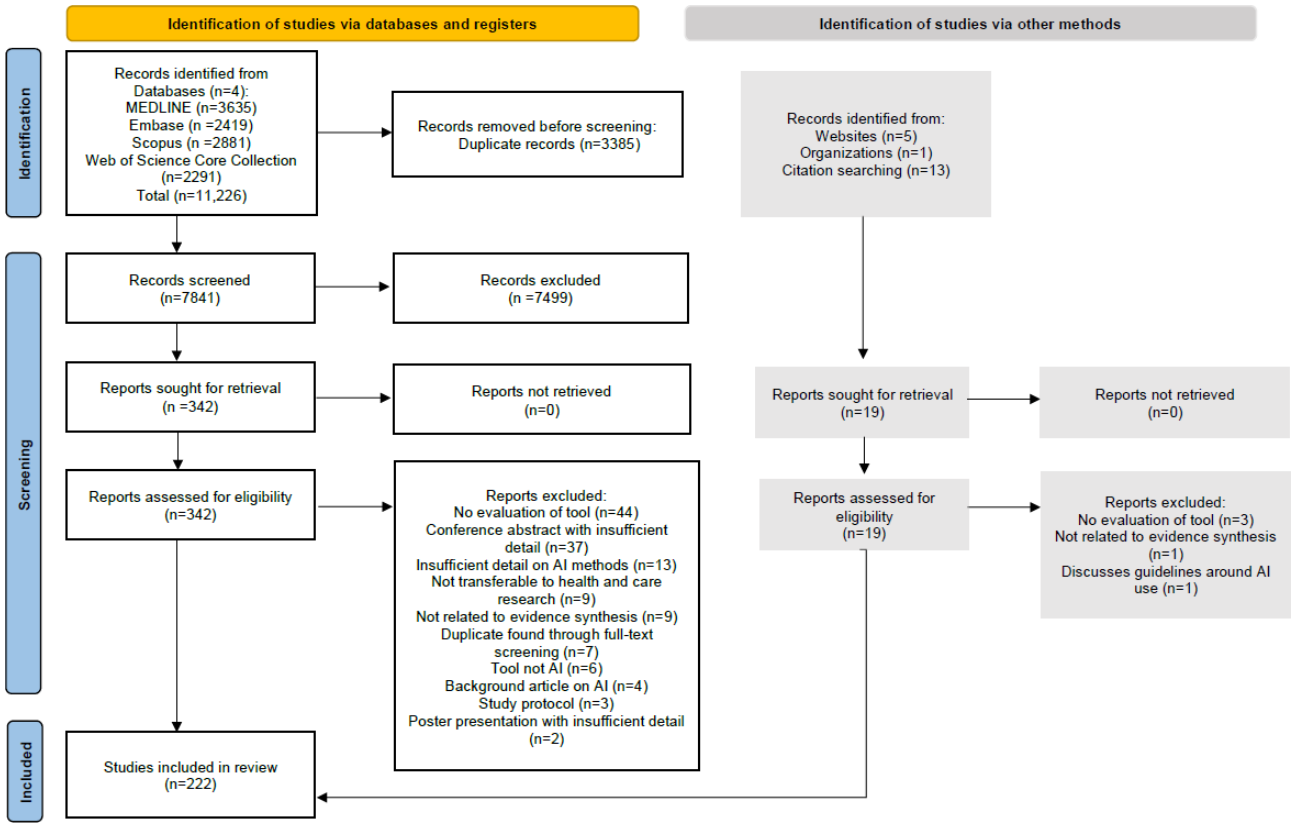

Study Characteristics

We included 206 (92.8%) primary studies and 16 (7.2%) secondary studies, with 2 (0.9%) stating they were explicitly AI assisted in their methods [26,27]. Three (1.4%) studies included were written in a language other than English [26,28,29]. A total of 11.7% (26/222) of studies were conference abstracts or posters. We identified 65 (29.3%) distinct AI tools and 25 (11.3%) open-source models or algorithms to automate the evidence synthesis process (Table S3 and S4 in Multimedia Appendix 1). Forty-two (18.9%) of the included studies had no specific tool name or data availability for their model.

Included studies could be divided into three broad categories: (1) methodological studies whereby researchers develop a novel algorithm or model and evaluate its effectiveness for automation; (2) evaluation, diagnostic accuracy, or feasibility studies of existing tools; and (3) comparative studies comparing the performance of AI tools to human researchers or similar tools executing the same task. Ten (4.5%) studies compared the performance of popular general-purpose LLMs, notably

XSL•FO
RenderX

ChatGPT, Claude, and Gemini, across various tasks, including screening and writing [30-40]. No studies compared traditional ML tools to LLM-based tools. There was only 1 (0.5%) study that evaluated an AI tool using a randomized trial study design [41].

A total of 54.1% (120/222) of included studies were published in 2024, followed by 16.7% (37/222) in 2023 and 14% (31/222) in 2022. Traditional ML, as an automation method, has remained relatively stable across the years, consistently supporting screening and extraction tasks (Figure 2). The use of BERTs has gradually increased since 2021, likely reflecting a growing interest in fine-tuned transformer models for domain-specific tasks. General-purpose LLMs saw a steep increase in 2024, likely coinciding with the widespread availability and maturity of models such as GPT-4 and Claude.

Figure 2. Year-on-year change in the study of three artificial intelligence methods for evidence synthesis automation. BERT: Bidirectional Encoder Representations from Transformers; LLM: large language model; ML: machine learning.

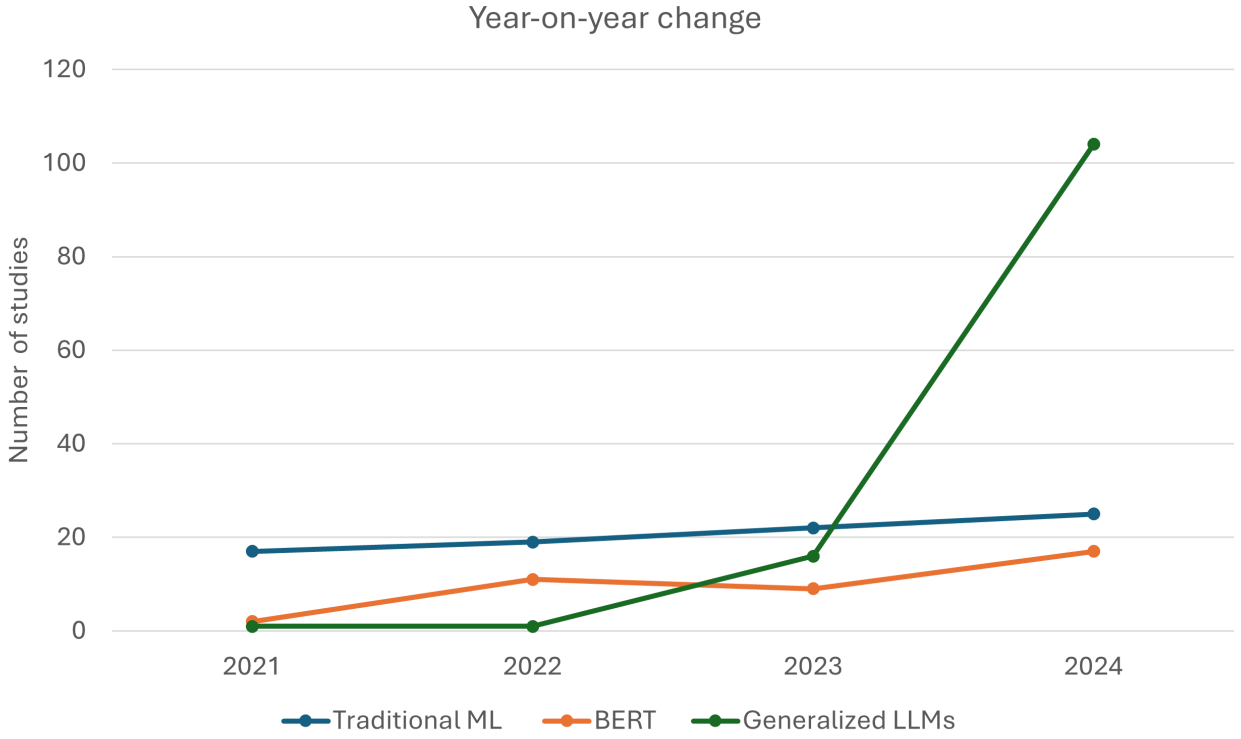

Evidence Synthesis Pathway

Many studies explored multiple areas of the evidence synthesis pathway, with most studies focusing on title and abstract screening (137/222, 61.7%), followed by data extraction (58/222, 26.1%), then literature searching (42/222, 18.9%; Figure 3). The proportion of studies per evidence synthesis stage reflects how amenable it is to automation. We categorized all 65 tools by evidence synthesis stage in Table S5 in Multimedia Appendix 1.

Figure 3. Percentage of studies examining the automation of distinct evidence synthesis stages using artificial intelligence.

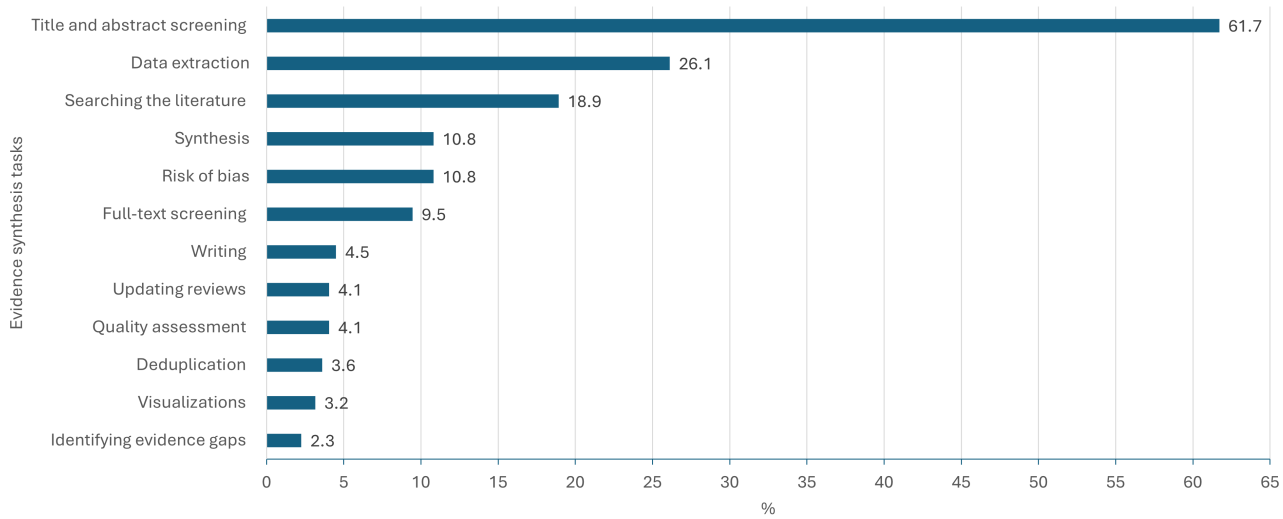

XSL•FO
RenderX

In total, 52.7% (117/222) of included studies researched tools leveraged by transformers. Some examples include general-purpose LLMs such as OpenAI's generative pretrained transformer (GPT) models or Mistral's 8×22B, as well as BERT models such as BioBERT or PubMedBERT (pretrained on predominantly biomedical text) that were fine-tuned for specific classification or screening tasks. In addition, 31.1% (69/222) evaluated tools underpinned by traditional ML algorithms alone, such as RCT Tagger, Research Screener, and Abstrackr. Some tools used an ensemble of ML and transformer-based methods, such as ASReview.

Evolution of AI Methods

A range of AI approaches were used to automate the evidence synthesis pathway, and these have evolved in computational power or complexity. The schematic in Figure 4 demonstrates two distinct categories of learning: traditional ML using classifier or clustering algorithms and deep learning leveraged by classical neural networks or transformers. Transformers are further divided into their subcategories of BERTs and general-purpose LLMs (ie, not designed specifically for evidence synthesis). Language and text processing abilities of the AI tools increase with complexity and computational power.

Figure 4. Schematic overview of artificial intelligence methods for evidence synthesis, ranging from traditional machine learning to neural networks and transformer-based large language models. AI: artificial intelligence.

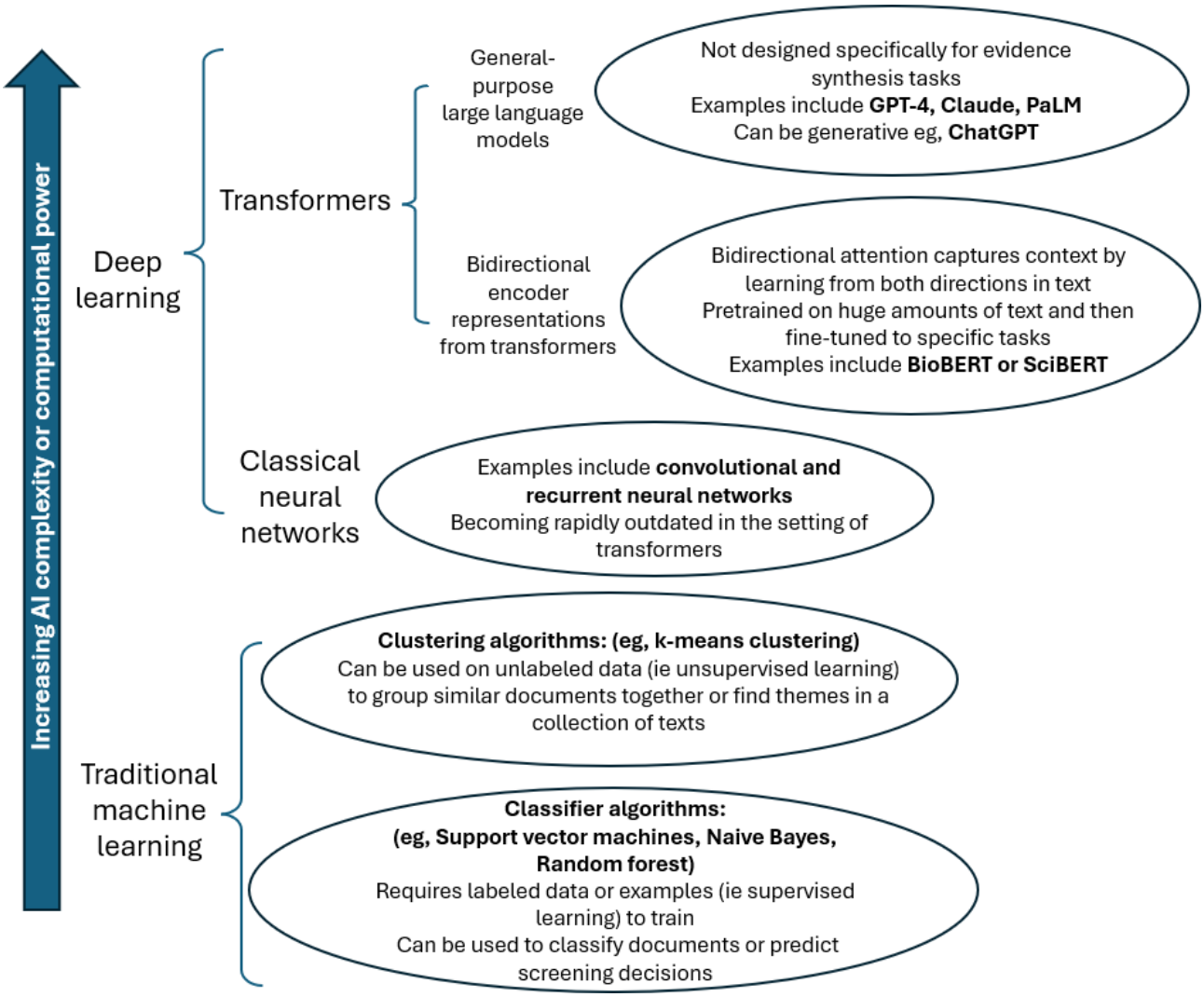

Both traditional ML and general-purpose LLMs were equally used to automate screening, with a similar distribution noted in data extraction (Figure 5). Traditional ML approaches being frequently leveraged for these stages likely highlight their ongoing suitability for evidence synthesis. Meanwhile, general-purpose LLMs can be applied across the entire evidence synthesis pathway. Generative LLMs are the only AI tool at present capable of producing text and therefore solely represent the automation of writing in evidence synthesis (Figure 5).

XSL•FO
RenderX

Figure 5. Visualization of included studies depicting the approximate number of studies by evidence synthesis stage and automation method. AI: artificial intelligence.

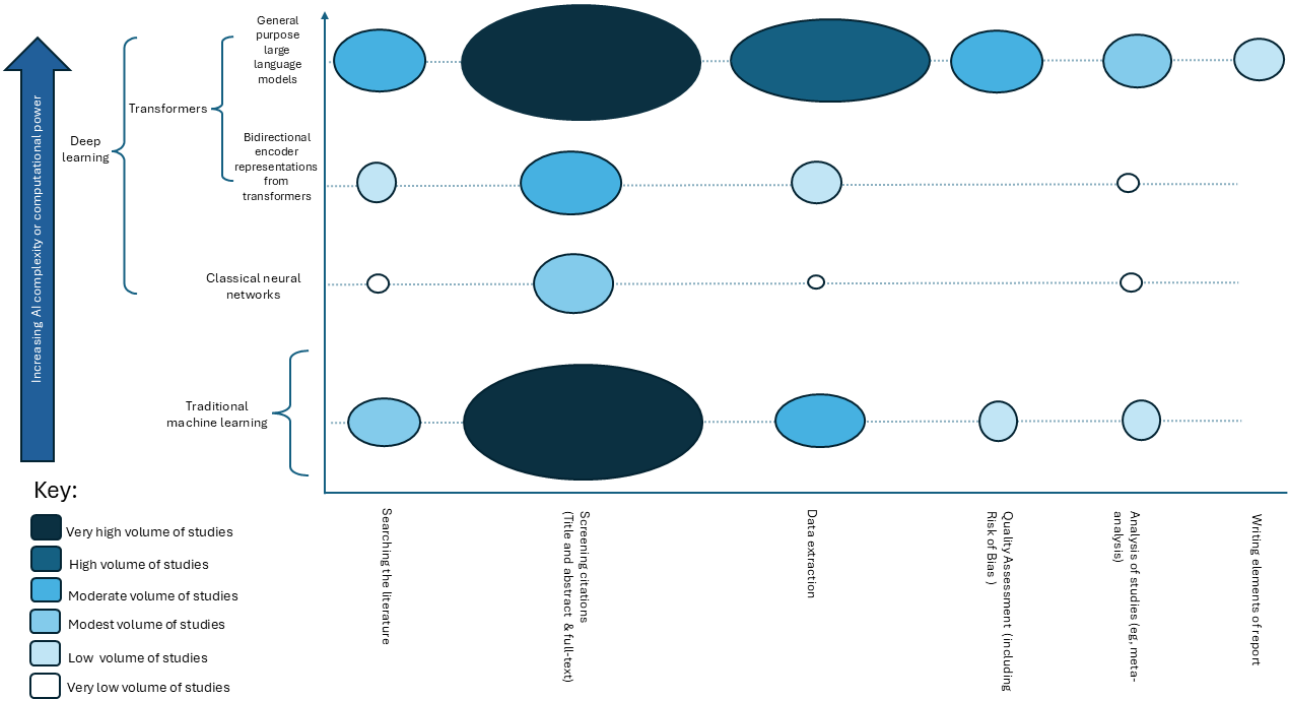

Distribution of Commonly Studied Tools

Of 222 studies, 70 (31.5%) used GPT models or ChatGPT. Other most cited tools included Anthropic's Claude (16/222, 7.2%), Rayyan (13/222, 5.9%), ASReview (12/222, 5.4%), Abstrackr (9/222, 4.1%), Google's Gemini (8/222, 3.6%), Covidence (5/222, 2.3%), DistillerSR (6/222, 2.7%), Colandr (3/222, 1.4%), and EPPI Reviewer (2/222, 0.9%). An alphabetized list of tools and their categorization across evidence synthesis tasks is available in Tables S3 and S5 in Multimedia Appendix 1. A list of all open-source models and algorithms is also available in Table S4 in Multimedia Appendix 1.

Outcomes

The most common outcomes measured were those designed to assess the technical performance of AI tools and models. Sensitivity and recall, which are the same measure, were used in 22.5% (50/222) and 23% (51/222) of studies, respectively. Other frequently reported metrics included specificity, accuracy, precision, area under the curve or area under the receiver operating characteristic curve, and F_1-score.

Using screening as an example, the area under the curve or the area under the receiver operating characteristic curve measures overall classification performance, indicating how well a model can distinguish between studies for inclusion versus exclusion. The F_1-score measures the balance between precision (how many papers selected by the model were actually relevant) and recall (how many relevant papers were successfully identified).

We also found specific outcomes designed to assess the researchers' workload or time saved. A popular outcome measure included Work Saved Over Sampling (WSS), which quantifies the percentage of manual screening work that can be avoided by using an AI tool while still maintaining the ability to identify all or most relevant studies. For example, a WSS of 30% means researchers can skip reviewing 30% of the citations while still identifying all relevant studies. Of 222 studies, 9 (4.1%) studies used WSS as a metric to assess the effectiveness of their automated screening tool [42-50].

Agreement between human researchers and AI tools, measured using metrics such as Cohen κ, was also used to assess the validity of AI decisions, although few included studies (12/222, 5.4%) used this approach.

Discussion

Principal Findings

This scoping review builds on previous reviews by comprehensively mapping the latest automation tools available across the evidence synthesis pathway and makes an important contribution to understanding how these tools are currently being researched and integrated into evidence synthesis workflows. We identified 65 distinct AI tools, some targeting specific aspects of the evidence synthesis pathway and others covering multiple stages, along with 25 open-source ML algorithms or BERT models designed or fine-tuned for evidence synthesis automation tasks. While published studies have traditionally focused on ML approaches for title and abstract screening and data extraction—the stages most readily automated—recent years have seen a notable shift toward transformer-based approaches for these same stages. Particularly in the past 2 years, there has been a rapid increase in published articles exploring the use of general-purpose LLMs, such as GPT, for evidence synthesis, despite these models not being developed specifically for it. This is notable given the risks generative AI poses due to its training methods and bias. We found that far fewer studies focused on optimizing the search for relevant studies or the analysis of the data or literature.

XSL•FO
RenderX

What the Results Mean

While transformer-based approaches are expanding the types of complex tasks that can be automated, task-specific ML methods continue to be developed and evaluated for well-defined problems. For example, priority screening using ML-based approaches has been validated thoroughly in the literature [43,51]. On the basis of the results from our review, a wide range of AI tools are being continually deployed and studied at different points across the evidence synthesis pathway, likely reflecting that no one tool or approach is superior to the other and that there is still room for improvement within evidence synthesis automation. While several studies attempted to compare tools with similar underlying AI methods, the lack of studies comparing distinct AI approaches, particularly traditional ML- to LLM-based tools, means we have no clear understanding of their comparative effectiveness.

It is likely that the choice of AI tools will be shaped by the expertise of the evidence synthesized, as well as the purpose and aim of the review itself. Extensively validated traditional ML tools or BERT models fine-tuned to specific research domains may be preferred by researchers or organizations who have a well-defined evidence synthesis task for automation, due to their reliability. For example, researchers conducting systematic reviews of drug or intervention effectiveness may find several well-evaluated tools at their disposal, as most available ML and BERT-based tools excel at interpreting and extracting data from randomized trials with clear eligibility criteria.

General-purpose LLMs, on the other hand, offer flexible language understanding and generative capabilities that can assist researchers in rapidly processing large volumes of text, identifying relevant information, and synthesizing preliminary insights across diverse topics. While these tools may allow for more nuanced interactions with complex evidence, they are likely less acceptable to traditional academic standards.

Most tools identified in this review were designed to optimize systematic reviews, predominantly in the biomedical domain. This is likely because of the very strict methodology that is applied to such reviews, which allows for replicability and, therefore, optimal comparison to the gold standard, manually performed reviews. Limited research explored AI automation in complex, multidisciplinary fields such as public health or social sciences, and few tools addressed nonsystematic review methods, such as narrative, realist, or integrative reviews. This disparity likely reflects a difference in data structure; more standardized data found in quantitative reviews is easier to process and is more machine readable than qualitative data, which tends to be heterogeneous and context dependent.

BERT models continue to be used as a method of automation for screening and data extraction-related tasks, with popular models such as SciBERT and BioBERT being fine-tuned to perform classification tasks for biomedical research questions notably. There is early emerging evidence, however, that general-purpose LLMs are equal if not superior to BERTs at most evidence synthesis tasks [37,52,53]. This potentially indicates to researchers that general-purpose LLMs, even if not designed specifically for evidence synthesis, may be a suitable option for automation if unable to invest the time and resources required to fine-tune BERT models.

Most studies included in our review used technical outcomes to assess tool performance, rather than directly comparing outcomes between AI tools and human researchers performing the same task. While sensitivity and specificity measures are important, there are implications for feasibility if studies do not assess AI tools pragmatically. Adoption of AI tools is likely going to be higher if there is a clear indication that the tool will save a researcher's time and improve workload for similar outputs.

Comparison With Previous Literature

Similar to the findings in Khalil and Jimenez's reviews [15,16], we found that the task most amenable to automation within evidence synthesis was title and abstract screening, with 62% of included studies exploring both ML and deep learning tools to automate this stage. We also found that most ML-based tools identified previously in both reviews remain active and usable today, with several having significant updates incorporating more advanced AI methods such as EPPI Reviewer, Covidence, and Rayyan. Our review was also uniquely able to identify several LLM-based tools that are built specific for evidence synthesis, such as Elicit, Scite, and Consensus. We also found in our review that general-purpose LLMs were predominantly applied to screening and searching of the literature, with GPT being the most frequently cited general-purpose LLM, consistent with Lieberum's review [12].

Unlike previous reviews, however, which studied ML- and LLM-based tools in isolation, our review demonstrates comparatively the strong shift in the evidence base to studying general-purpose LLMs for evidence synthesis tasks in the last 5 years, particularly for screening, extraction, and, uniquely, writing-related tasks. However, the evidence base for AI-generated or enhanced scientific writing is still evolving and remains a widely debated topic in the research community due to ethical and logistical issues attached to the practice [54-57]. Most journals at present ask for disclosure of AI assistance in the writing process.

Our review also uniquely sought to identify comparative studies between different AI approaches, particularly between traditional ML- and LLM-based tools, as previous reviews focused on distinct approaches in isolation. However, we were unable to identify any such studies within our search time frame, highlighting an important gap in the research. Recent domain-specific evidence has shown that LLM-based tools may outperform ML tools in evidence synthesis tasks such as screening, although these findings are limited due to a lack of generalizability [58].

Strengths and Limitations

There are several strengths to this scoping review. First, it offers a broad and comprehensive overview of the field, including over 200 studies. We clearly mapped AI tool use across the entire evidence synthesis pathway, providing complete lists of tools available for specific or multiple tasks. It also uses previous reviews' frameworks to expand on the topic of automation in evidence synthesis, adding consistently to this evidence base.

XSL•FO
RenderX

The search strategy was robust, covering multiple academic databases as well as gray literature sources, increasing the likelihood of capturing both peer-reviewed and nontraditional publications.

However, there are several limitations. Given the rapidly evolving nature of AI in evidence synthesis, it is possible that several recent publications were not captured. There is also temporal bias: general-purpose LLMs have only gained widespread adoption in the past 2 years, meaning fewer high-quality peer-reviewed studies are available compared to evaluations of older AI approaches. In addition, this review included studies that explicitly evaluated AI methods or tools to automate evidence synthesis, thereby excluding studies that may have researched AI without clear or proper disclosure of it. This potentially limited our evidence base and underestimated the actual AI usage in evidence synthesis. Additionally, a lack of granularity in reporting the underlying AI methods or technical approaches restricted the meaningful categorization and comparison of tools. Poor reporting of tool development methods also limits generalizability to health and care, particularly when models were trained within specific research domains. Publication bias may also exist, as tools developed in academic settings are more likely to appear in scholarly journals, whereas commercial tools may be underrepresented in the literature. Finally, an inherent limitation of scoping review methodology is the lack of critical appraisal of the evidence base, which limits our understanding of the quality of studies included.

Recommendations for Policy and Practice

Until we can put in place mitigations that are effective in reducing error and bias in AI tools, there will be a need for a human researcher in the loop [59]. Therefore, the tools highlighted in this paper should be used as adjuncts to humans, rather than replacements, to maintain high-quality and rigorous reviews.

Funders and researchers should continue to support efforts to build an evidence base in this field that can provide a feedback loop for developers to improve upon the tools' effectiveness, feasibility, and acceptability. In particular, assessing these tools in real-world settings and comparing traditional ML methods to newer LLM-based approaches is pertinent to understanding automation opportunities and challenges. Researchers and funders should foster knowledge sharing on the effective use of emerging AI tools, such as through the International Collaboration of the Automation of Systematic Reviews [60]. This can be through training resources, collaborative learning opportunities, and sharing best practices. This will help build collective understanding and capacity across the research community.

Almost all tools found had paid upgrades or monthly subscriptions. The lack of comprehensive tools that are free to use is concerning from both an accessibility and equity angle and is likely to influence how such tools are integrated into research practice.

Quality standards are needed to ensure ethical tool development, and researchers and stakeholders should adhere to guidelines that stipulate methods to properly and appropriately integrate automation tools into evidence synthesis, such as the recommendations and guidance for Responsible AI in Evidence Synthesis [61].

Decision makers who use AI-assisted evidence syntheses should be aware of the issues around error and bias and should exercise caution when using reviews generated by tools that claim to perform the whole process.

LLMs also have significant electricity demands due to their scale, data requirements, and computational demands [62,63]. Estimates suggest that generative AI processing, from data center operation and cooling, could consume over 8% of the United States' electricity and 5% of Europe's by 2030 [64]. Furthermore, manufacturing generative processing units poses significant environmental impacts, particularly for communities near production like in Taichung, Taiwan [64]. While some developers are exploring LLMs using renewable energy, such developments remain in their infancy [65].

In the setting of many organizations and institutions requiring consideration of sustainability in their practice, it is important that we review the environmental implications of LLMs if implemented regularly into evidence synthesis workflows. Therefore, we recommend that sustainability research be undertaken around the use of such AI tools, which is not at all considered in the current evidence base.

Conclusions

Having accurate, high-quality, and contemporaneous reviews of evidence is essential to making the most informed decisions about health and care. Many of the processes involved in evidence synthesis—particularly identifying relevant studies, screening titles and abstracts, and extracting structured data—involve repetitive, rule-based tasks that are well-suited to automation using AI. With the ongoing development of ML tools and the introduction of LLMs, researchers now have a large suite of AI-based automation tools available to them, as highlighted in our review. While some tools are yet to be comprehensively evaluated, validated, and compared, given the risks generative AI poses to evidence synthesis integrity, the evidence base highlights many potential avenues to improve manual workload. As we continue to integrate such tools into research workflows, it is vital that we continue to monitor outcomes, maintain transparency, and always have a researcher-in-the-loop to ensure that evidence reviews continue to be of a high standard.

Acknowledgments

The Health Foundation was not involved in the design and methodology of this review. They provided final comments to the paper that were incorporated into the manuscript. The authors would like to acknowledge Starry Chow (The University of Hong

XSL•FO
RenderX

Kong—Li Ka Shing Faculty of Medicine) for their input in translation during the full-text screening process. Generative artificial intelligence was not used to inform this review's design or methodology or generate any text or figures. It was only used to refine writing for grammar and clarity.

Funding

"An overview and evaluation of the use of automated tools in rapid evidence synthesis" project is supported by the Health Foundation, an independent charitable organization working to build a healthier United Kingdom (reference FR-0006738).

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during this study.

Authors' Contributions

Conceptualization: SH (equal), JF (lead)
Data curation: SH (equal), H Pearce (equal), CA (supporting), IK (supporting)
Formal analysis: SH (lead), JF (supporting)
Funding acquisition: JF
Investigation: SH (lead), H Pearce (equal), IK (supporting)
Methodology: SH (lead), JF (equal)
Project administration: SH (lead), JF (equal)
Supervision: JF
Visualization: SH (lead), H Pearce (supporting), JF (supporting)
Writing – original draft: SH (lead)
Writing – review & editing: H Pearce (supporting), H Painter (supporting), CA (supporting), LL (supporting), IK (supporting), JK (supporting), AC (supporting), JF (equal)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials including detailed database search strategies, PRISMA-S checklist, and categorized tables of AI tools and open-source models identified in the review.
[DOCX File, 49 KB - jmir_v28i1e81597_app1.docx]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 28 KB - jmir_v28i1e81597_app2.docx]

References

1. Khalil H, Tamara L, Rada G, Akl EA. Challenges of evidence synthesis during the 2020 COVID pandemic: a scoping review. J Clin Epidemiol 2022 Feb;142:10-18. [doi: 10.1016/j.jclinepi.2021.10.017] [Medline: 34718121]
2. Tercero-Hidalgo JR, Khan KS, Bueno-Cavanillas A, et al. Artificial intelligence in COVID-19 evidence syntheses was underutilized, but impactful: a methodological study. J Clin Epidemiol 2022 Aug;148:124-134. [doi: 10.1016/j.jclinepi.2022.04.027] [Medline: 35513213]
3. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open 2017 Feb 27;7(2):e012545. [doi: 10.1136/bmjopen-2016-012545] [Medline: 28242767]
4. Clark J, Glasziou P, Del Mar C, Bannach-Brown A, Stehlik P, Scott AM. A full systematic review was completed in 2 weeks using automation tools: a case study. J Clin Epidemiol 2020 May;121:81-90. [doi: 10.1016/j.jclinepi.2020.01.008] [Medline: 32004673]
5. Clark J, McFarlane C, Cleo G, Ishikawa Ramos C, Marshall S. The impact of systematic review automation tools on methodological quality and time taken to complete systematic review tasks: case study. JMIR Med Educ 2021 May 31;7(2):e24418. [doi: 10.2196/24418] [Medline: 34057072]
6. Abogunrin S, Muir JM, Zerbini C, Sarri G. How much can we save by applying artificial intelligence in evidence synthesis? Results from a pragmatic review to quantify workload efficiencies and cost savings. Front Pharmacol 2025;16:1454245. [doi: 10.3389/fphar.2025.1454245] [Medline: 39959426]

XSL•FO
RenderX

7. Rogers K, Miller A, Girgis A, Clark EC, Neil-Sztramko SE, Dobbins M. Leveraging AI to optimize maintenance of health evidence and offer a one-stop shop for quality-appraised evidence syntheses on the effectiveness of public health interventions: quality improvement project. J Med Internet Res 2025 Jul 29;27(1):e69700. [doi: 10.2196/69700] [Medline: 40729661]
8. Ofori-Boateng R, Aceves-Martins M, Wiratunga N, Moreno-Garcia CF. Towards the automation of systematic reviews using natural language processing, machine learning, and deep learning: a comprehensive review. Artif Intell Rev 2024;57(8):200. [doi: 10.1007/s10462-024-10844-w]
9. Liu J, Lai H, Zhao W, et al. AI-driven evidence synthesis: data extraction of randomized controlled trials with large language models. Int J Surg 2025;111(3):2722-2726. [doi: 10.1097/JS9.0000000000002215]
10. Wang Z, Cao L, Danek B, Jin Q, Lu Z, Sun J. Accelerating clinical evidence synthesis with large language models. NPJ Digit Med 2025 Aug 8;8(1):509. [doi: 10.1038/s41746-025-01840-7] [Medline: 40775042]
11. Clark J, Barton B, Albarqouni L, et al. Generative artificial intelligence use in evidence synthesis: a systematic review. Res synth methods 2025 Jul;16(4):601-619. [doi: 10.1017/rsm.2025.16]
12. Lieberum JL, Toews M, Metzendorf MI, et al. Large language models for conducting systematic reviews: on the rise, but not yet ready for use-a scoping review. J Clin Epidemiol 2025 May;181:111746. [doi: 10.1016/j.jclinepi.2025.111746] [Medline: 40021099]
13. Siemens W, von Elm E, Binder H, et al. Opportunities, challenges and risks of using artificial intelligence for evidence synthesis. BMJ Evid Based Med 2025 Dec 1;30(6):381-384. [doi: 10.1136/bmjebm-2024-113320] [Medline: 39788693]
14. Ong AY, Merle DA, Wagner SK, Keane PA. Exploring the dilemma of AI use in medical research and knowledge synthesis: a perspective on deep research tools. J Med Internet Res 2025 Jul 15;27(1):e75666. [doi: 10.2196/75666] [Medline: 40663724]
15. Khalil H, Ameen D, Zarnegar A. Tools to support the automation of systematic reviews: a scoping review. J Clin Epidemiol 2022 Apr;144:22-42. [doi: 10.1016/j.jclinepi.2021.12.005] [Medline: 34896236]
16. Cierco Jimenez R, Lee T, Rosillo N, et al. Machine learning computational tools to assist the performance of systematic reviews: a mapping review. BMC Med Res Methodol 2022 Dec 16;22(1):322. [doi: 10.1186/s12874-022-01805-4] [Medline: 36522637]
17. Roth S, Wermer-Colan A. Machine learning methods for systematic reviews:: a rapid scoping review. Dela J Public Health 2023 Nov;9(4):40-47. [doi: 10.32481/djph.2023.11.008] [Medline: 38173960]
18. Khalil H, Pollock D, McInerney P, et al. Automation tools to support undertaking scoping reviews. Res Synth Methods 2024 Nov;15(6):839-850. [doi: 10.1002/jrsm.1731] [Medline: 38885942]
19. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020 Oct;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
20. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
21. Litmaps. URL: https://www.litmaps.com/ [accessed 2025-05-12]
22. A pro-innovation approach to AI regulation. GOV.UK. URL: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper [accessed 2024-12-20]
23. Evidence synthesis - what is it and why do we need it? Cochrane. 2019 Sep 13. URL: https://www.cochrane.org/news/evidence-synthesis-what-it-and-why-do-we-need-it [accessed 2024-12-20]
24. EndNote. URL: https://endnote.com/ [accessed 2025-05-12]
25. Rayyan OM. AI and the evolution of journalistic practices. J Inf Stud Technol 2025 Jul 31;2. [doi: 10.5339/jist.2025.15]
26. Cardoso Sampaio R, Chagas V, Sinimbu Sanchez C. An artificial intelligence (AI)-assisted scoping review of emerging uses of AI in qualitative research and its ethical considerations. Rev Pesq Qual 2024;12(30):01-28. [doi: 10.33361/RPQ.2024.v.12.n.30.729]
27. Teperikidis E, Boulmpou A, Potoupni V, Kundu S, Singh B, Papadopoulos C. Does the long-term administration of proton pump inhibitors increase the risk of adverse cardiovascular outcomes? A ChatGPT powered umbrella review. Acta Cardiol 2023 Nov;78(9):980-988. [doi: 10.1080/00015385.2023.2231299] [Medline: 37431972]
28. Guo Y, Zhang X, Sun W, Deng H. Application of automated literature screening tools in systematic reviews. Med J Peking Union Med Coll Hosp 2024;15(4):921-926. [doi: 10.12290/xhyxzz.2023-0257]
29. Esposito C, Dell'Omo G, Di Ianni D, Di Procolo P. Human vs. ChatGPT. Is it possible obtain comparable results in the analysis of a scientific systematic review? Recenti Prog Med 2024 Sep;115(9):420-425. [doi: 10.1701/4334.43184] [Medline: 39269357]
30. AlSagri HS, Farhat F, Sohail SS, Saudagar AKJ. ChatGPT or Gemini: who makes the better scientific writing assistant? J Acad Ethics 2025 Sep;23(3):1121-1135. [doi: 10.1007/s10805-024-09549-0]
31. De Cassai A, Dost B, Karapinar YE, et al. Evaluating the utility of large language models in generating search strings for systematic reviews in anesthesiology: a comparative analysis of top-ranked journals. Reg Anesth Pain Med 2025 Jan 19;Published online January 2025:4. [doi: 10.1136/rapm-2024-106231] [Medline: 39828514]
32. Delgado-Chaves FM, Jennings MJ, Atalaia A, et al. Transforming literature screening: the emerging role of large language models in systematic reviews. Proc Natl Acad Sci U S A 2025 Jan 14;122(2):e2411962122. [doi: 10.1073/pnas.2411962122] [Medline: 39761403]

XSL•FO
RenderX

33. Konet A, Thomas I, Gartlehner G, et al. Performance of two large language models for data extraction in evidence synthesis. Res Synth Methods 2024 Sep;15(5):818-824. [doi: 10.1002/jrsm.1732] [Medline: 38895747]
34. Li M, Sun J, Tan X. Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. Syst Rev 2024 Aug 21;13(1):219. [doi: 10.1186/s13643-024-02609-x] [Medline: 39169386]
35. Omar M, Nassar S, Hijazi K, Glicksberg BS, Nadkarni GN, Klang E. Generating credible referenced medical research: a comparative study of openAI's GPT-4 and Google's gemini. Comput Biol Med 2025 Feb;185:109545. [doi: 10.1016/j.compbiomed.2024.109545] [Medline: 39667055]
36. Rathi H, Malik A, Behera DC, Kamboj G. Msr28 use of large language model (LLM) for full-text screening in systematic literature reviews: a comparative analysis. Value Health 2024 Jun;27(6):S264. [doi: 10.1016/j.jval.2024.03.1461]
37. Sanghera R, Thirunavukarasu AJ, Khoury ME, et al. High-performance automated abstract screening with large language model ensembles. arXiv. Preprint posted online on Nov 22, 2024. [doi: 10.48550/arXiv.2411.02451]
38. Šuster S, Baldwin T, Verspoor K. Zero- and few-shot prompting of generative large language models provides weak assessment of risk of bias in clinical trials. Res Synth Methods 2024 Nov;15(6):988-1000. [doi: 10.1002/jrsm.1749] [Medline: 39176994]
39. Thelwall M. Is Google Gemini better than ChatGPT at evaluating research quality? J. Data Inf. Sci. 2025 Apr 1;10(2):1-5. [doi: 10.2478/jdis-2025-0014]
40. Trad F, Charafeddine J, Chkahtoura M, Rahme M, Fuleihan GEH, Chehab A. Streamlining systematic reviews in medical research: a novel application of large language models. J Bone Miner Res 2024;39(Supplement 1). [doi: 10.1093/jbmr/zjae038] [Medline: 38477812]
41. Arno A, Thomas J, Wallace B, Marshall IJ, McKenzie JE, Elliott JH. Accuracy and efficiency of machine learning-assisted risk-of-bias assessments in "real-world" systematic reviews: a noninferiority randomized controlled trial. Ann Intern Med 2022 Jul;175(7):1001-1009. [doi: 10.7326/M22-0092] [Medline: 35635850]
42. Bravo A, Bennetts L, Atanasov P. Accelerating the early identification of relevant studies in title and abstract screening. Presented at: 2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC); Nov 12-14, 2021. [doi: 10.1109/ISCSIC54682.2021.00034]
43. Ferdinands G, Schram R, de Bruin J, et al. Performance of active learning models for screening prioritization in systematic reviews: a simulation study into the Average Time to Discover relevant records. Syst Rev 2023 Jun 20;12(1):100. [doi: 10.1186/s13643-023-02257-7] [Medline: 37340494]
44. Yao X, Kumar MV, Su E, Flores Miranda A, Saha A, Sussman J. Evaluating the efficacy of artificial intelligence tools for the automation of systematic reviews in cancer research: a systematic review. Cancer Epidemiol 2024 Feb;88:102511. [doi: 10.1016/j.canep.2023.102511] [Medline: 38071872]
45. Chai KEK, Lines RLJ, Gucciardi DF, Ng L. Research Screener: a machine learning tool to semi-automate abstract screening for systematic reviews. Syst Rev 2021 Apr 1;10(1):93. [doi: 10.1186/s13643-021-01635-3] [Medline: 33795003]
46. van Dinter R, Catal C, Tekinerdogan B. A multi-channel convolutional neural network approach to automate the citation screening process. Appl Soft Comput 2021 Nov;112:107765. [doi: 10.1016/j.asoc.2021.107765]
47. Kusa W, Hanbury A, Knoth P. Automation of citation screening for systematic literature reviews using neural networks: a replicability study. Presented at: Advances in Information Retrieval - 44th European Conference on (IR) Research, (ECIR) 2022; Apr 10-14, 2022. [doi: 10.1007/978-3-030-99736-6_39]
48. Muthu S. The efficiency of machine learning-assisted platform for article screening in systematic reviews in orthopaedics. Int Orthop 2023 Feb;47(2):551-556. [doi: 10.1007/s00264-022-05672-y] [Medline: 36562816]
49. Akinseloyin O, Jiang XR, Palade V. A question-answering framework for automated abstract screening using large language models. J Am Med Inform Assoc 2024 Sep 1;31(9):1939-1952. [doi: 10.1093/jamia/ocae166] [Medline: 39042516]
50. Feng Y, Liang S, Zhang Y, et al. Automated medical literature screening using artificial intelligence: a systematic review and meta-analysis. J Am Med Inform Assoc 2022 Jul 12;29(8):1425-1432. [doi: 10.1093/jamia/ocac066] [Medline: 35641139]
51. Tsou AY, Treadwell JR, Erinoff E, Schoelles K. Machine learning for screening prioritization in systematic reviews: comparative performance of Abstrackr and EPPI-Reviewer. Syst Rev 2020 Apr 2;9(1):73. [doi: 10.1186/s13643-020-01324-7] [Medline: 32241297]
52. Jahan I, Laskar MTR, Peng C, Huang J. Evaluation of ChatGPT on biomedical tasks: a zero-shot comparison with fine-tuned generative transformers. arXiv. Preprint posted online on Aug 24, 2023. [doi: 10.48550/arXiv.2302.10198]
53. Zhong Q, Ding L, Liu J, Du B, Tao D. Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned BERT. arXiv. Preprint posted online on Mar 2, 2023. [doi: 10.48550/arXiv.2302.10198]
54. Draxler F, Werner A, Lehmann F, et al. The AI ghostwriter effect: when users do not perceive ownership of AI-generated text but self-declare as authors. ACM Trans Comput-Hum Interact 2024 Apr 30;31(2):1-40. [doi: 10.1145/3637875]
55. Hosseini M, Gordijn B, Kaebnick GE, Holmes K. Disclosing generative AI use for writing assistance should be voluntary. Res Ethics 2025 Jun 21;21(4):728-735. [doi: 10.1177/17470161251345499] [Medline: 40950847]
56. Cohen JF, Moher D. Generative artificial intelligence and academic writing: friend or foe? J Clin Epidemiol 2025 Mar;179:111646. [doi: 10.1016/j.jclinepi.2024.111646] [Medline: 39706536]

XSL•FO
RenderX

57. Formosa P, Bankins S, Matulionyte R, Ghasemi O. Can ChatGPT be an author? Generative AI creative writing assistance and perceptions of authorship, creatorship, responsibility, and disclosure. AI & Soc 2025 Jun;40(5):3405-3417. [doi: 10.1007/s00146-024-02081-0]
58. Dai ZY, Wang FQ, Shen C, et al. Accuracy of large language models for literature screening in thoracic surgery: diagnostic study. J Med Internet Res 2025 Mar 11;27:e67488. [doi: 10.2196/67488] [Medline: 40068152]
59. Li Y, Datta S, Rastegar-Mojarad M, et al. Enhancing systematic literature reviews with generative artificial intelligence: development, applications, and performance evaluation. J Am Med Inform Assoc 2025 Apr 1;32(4):616-625. [doi: 10.1093/jamia/ocaf030] [Medline: 40036547]
60. International Collaboration for the Automation of Systematic Reviews. URL: https://icasr.github.io/ [accessed 2025-05-12]
61. Thomas J, Flemyng E, Noel-Storr A. Responsible AI in Evidence Synthesis (RAISE): guidance and recommendations. Open Science Framework. 2024 Aug 29. URL: https://osf.io/fwaud/ [accessed 2025-04-24]
62. Rillig MC, Ågerstrand M, Bi M, Gould KA, Sauerland U. Risks and benefits of large language models for the environment. Environ Sci Technol 2023 Mar 7;57(9):3464-3466. [doi: 10.1021/acs.est.3c01106] [Medline: 36821477]
63. Ji Z, Jiang M. A systematic review of electricity demand for large language models: evaluations, challenges, and solutions. Renew Sustain Energy Rev 2026 Jan;225:116159. [doi: 10.1016/j.rser.2025.116159]
64. Hosseini M, Gao P, Vivas-Valencia C. A social-environmental impact perspective of generative artificial intelligence. Environ Sci Ecotechnol 2025 Jan;23:100520. [doi: 10.1016/j.ese.2024.100520] [Medline: 39811486]
65. Singh A, Patel NP, Ehtesham A, Kumar S, Khoei TT. A survey of sustainability in large language models: applications, economics, and challenges. arXiv. Preprint posted online on Mar 29, 2023. [doi: 10.48550/arXiv.2412.04782]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
GPT: generative pretrained transformer
LLM: large language model
ML: machine learning
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
WSS: Work Saved Over Sampling

Edited by A Mavragani, S Brini; submitted 01.Aug.2025; peer-reviewed by E Ting, H Khalil; revised version received 27.Oct.2025; accepted 03.Nov.2025; published 30.Mar.2026.

Please cite as:
Harasgama S, Pearce H, Appel C, Loftus L, Painter H, Kuhn I, Karpusheff J, Ceesay A, Ford J
Artificial Intelligence Tools for Automating Evidence Synthesis: Scoping Review
J Med Internet Res 2026;28:e81597
URL: https://www.jmir.org/2026/1/e81597
doi:10.2196/81597

© Sashika Harasgama, Helen Pearce, Cameron Appel, Liam Loftus, Helena Painter, Isla Kuhn, Justine Karpusheff, Aji Ceesay, John Ford. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Smartphone-Based Digital Phenotyping Across Health Conditions: Scoping Review

Arlen Dumas[1], BS; Joanne Hokayem[2], BS; Georgia Goodman[3], BS; Krishna Venkatasubramanian[1], PhD; Peter Chai[2], MD, MMS

[1]Department of Computer Science and Statistics, University of Rhode Island, 9 Greenhouse Road, Kingston, RI, United States

[2]The Fenway Institute, Boston, MA, United States

[3]UCLA (University of California, Los Angeles) Health, Los Angeles, CA, United States

Corresponding Author:
Krishna Venkatasubramanian, PhD
Department of Computer Science and Statistics, University of Rhode Island, 9 Greenhouse Road, Kingston, RI, United States

Abstract

Background: Smartphone-based digital phenotyping uses built-in sensors and usage patterns to passively capture behavioral and environmental data relevant to health and has been applied extensively in mental health and chronic disease contexts.

Objective: This review synthesizes studies that use smartphone-based digital phenotyping, defined as approaches that rely exclusively on onboard smartphone sensors to characterize specific health conditions. To our knowledge, this work provides the most comprehensive cross-condition synthesis of smartphone-based digital phenotyping to date, spanning mental health, physical health, and substance use disorders (SUDs), and highlighting common practices, gaps, and opportunities for future research.

Methods: We conducted a scoping review of English-language, peer-reviewed papers published between 2012 and 2025 in Google Scholar, IEEE Xplore, ACM Digital Library, and PubMed using terms such as "mobile sensing" and "digital phenotyping." Eligible papers used onboard smartphone sensors to assess health and went beyond self-report. Studies that did not rely on smartphone auxiliary sensing modalities or digital phenotyping were excluded.

Results: We performed a descriptive synthesis of study characteristics, sensors, and health domains. Of 111 papers identified, 65 met inclusion criteria. Most studies were observational and relied on passive sensing. Sample sizes ranged from fewer than 10 to over 18,000 participants, with a median of 52 (IQR=26‑126). Mental health conditions were most frequently examined, including depression (n=16), bipolar disorder (n=11), stress or anxiety (n=10), and schizophrenia (n=8). Less commonly studied conditions included SUDs (n=7), Parkinson disease (n=4), and sleep apnea (n=2). Sensor streams varied widely and included diverse passive smartphone data sources capturing mobility, communication, device usage, environmental context, and user interaction patterns. Ground-truth measurements most commonly relied on validated clinical scales (eg, Patient Health Questionnaire-9, Young Mania Rating Scale [YMRS], and Pittsburgh Sleep Quality Index; n=41), followed by ecological momentary assessments (n=18), clinician-confirmed diagnoses (n=9), and physiological measures such as polysomnography (n=3). Across studies, recurring methodological limitations included incomplete or inconsistent sensor descriptions, limited reporting of data quality (eg, sampling rates and missingness), and heterogeneous validation practices. These issues limit comparability and reproducibility and underscore the need for clearer reporting standards and greater data availability.

Conclusions: This scoping review provides the first comprehensive synthesis of smartphone-only digital phenotyping studies spanning mental health, physical health, and SUDs. Unlike prior reviews, this work maps behavioral associations derived exclusively from smartphone sensors across a broad range of health domains. The primary contribution of this review lies in its consolidation of behavioral associations observed across studies, enabling researchers to correlate new findings to the existing evidence base and identify opportunities for replication, extension, or clinical translation. Collectively, these findings highlight both the promise of smartphone-based digital phenotyping in real-world settings and the need for improved standardization to support translation into clinical and public health applications.

(J Med Internet Res 2026;28:e84146) doi:10.2196/84146

KEYWORDS

digital phenotyping; smartphones; digital health; scoping review; health behaviors; mobile phone

XSL•FO
RenderX

Introduction

Since their introduction in 1994, smartphones have penetrated almost every aspect of human life worldwide [1]. These devices are ubiquitous in many cultures and have transformed access to the internet and communication with others. Like many other sectors, health care has been directly impacted by the use of smartphones [2,3]. Patients can now conduct virtual visits with their health care providers [4], have access to a wide range of sources of medical information, and can access health records on smartphones [3]. Recent survey data in the United States suggests that 62% of individuals would like to access their medical records through mobile-facing apps, and in 2022, 48% of patients accessed medical records via the internet at least once [5]. Additionally, increased use of wearable and other health-related tracking devices has positioned smartphones as a centralized hub for health data and behavior tracking [6,7]. This data suggests that individuals access and use smartphones to assess and modulate individual health, positioning these devices as important platforms to interact with individuals [6,8].

One potential strategy to leverage smartphones to understand health behavior is through digital phenotyping, a strategy that leverages data from mobile computing devices that most people now have with them (eg, smartphones, smartwatches, rings, and other wearable devices) to capture a person's behavior, and through them, understand any health conditions present [9]. More specifically, digital phenotyping is the collection and analysis of data generated through human interaction with mobile computing devices. It encompasses (1) passive input collected from sensors on a smartphone, such as continuous GPS monitoring, accelerometry data; and (2) active input, such as ecological momentary assessments (EMAs) completed by the user of the smartphone. By integrating such passive and active inputs, digital phenotyping aims to explore behaviors that correlate with specific health conditions [10].

Smartphone-based digital phenotyping is a form of digital phenotyping in which smartphones are the only devices used for collecting sensor data from individuals. Smartphones offer a unique advantage due to their widespread use, enabling researchers to collect a rich, high-volume dataset that can be used to assess behaviors (also known as phenotypes) in the context of a specific health condition in an unobtrusive way [11]. Data that has been captured via smartphones can be analyzed through a variety of methods to provide insight into a variety of mental and physical health conditions [12]. Compared to traditional approaches of monitoring and diagnosing complex health conditions (eg, schizophrenia and bipolar disorder) that require considerable human interventions, smartphone-based digital phenotyping offers an opportunity to passively track these conditions over time [13]. The other potential advantage of smartphone-based digital phenotyping is the opportunity to democratize diagnoses of various health conditions simply by assessing smartphone use parameters in the context of the user's known actions [14]. Despite its promise, multiple challenges exist to operationalize smartphone-based digital phenotyping, account for variability in data, and generalize findings across patient populations [9].

This scoping review aims to comprehensively describe the current state of smartphone-based digital phenotyping work. Our scope is more expansive than existing scoping reviews as it covers any and all health conditions that have been explored as part of smartphone-based phenotyping. By exploring the breadth of health conditions studied by smartphone-based digital phenotyping, we can (1) show the versatility of the smartphone-based digital phenotyping to address the needs of a whole variety of health conditions, and (2) obtain a broader understanding of how smartphone-based digital phenotyping can be used across different clinical contexts. Further, in this work, in keeping with the spirit of digital phenotyping, we describe the relationship between smartphone data and behaviors that are exhibited in the presence of specific medical conditions for the works reviewed. We avoid going into algorithmic or machine learning (ML) methods in this scoping review, as that can sometimes promote the discovery without understanding the mindset when looking at prior work. As most of the studies reviewed are observational, the relationships reported between smartphone-derived behaviors and health conditions should be interpreted as associational rather than causal.

This scoping review aims to synthesize the existing literature on smartphone-only digital phenotyping. Specifically, we focused on studies that exclusively use smartphone sensors to characterize health conditions. Our objectives are to consolidate evidence across health domains, describe how behavioral signals are captured and assessed using different smartphone sensor streams, and to identify recurring methodological limitations and gaps in existing literature. By taking a cross-condition perspective, this review clarifies where evidence is concentrated, where it remains sparse, and how future research can better support replication, comparability, and translation into clinical and public health settings.

Methods

We conducted a scoping review to understand the current state of smartphone-based digital phenotyping. No review protocol was registered for this study. Methods adhered to PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [15]. The search strategy was reported in accordance with PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) guidelines [16]. We systematically queried Google Scholar, IEEE Xplore, ACM Digital Library, and PubMed for peer-reviewed papers published between 2012 and October 2025. The revised search strategy yielded approximately 3700 records across the 4 databases before screening. Search terms included "mobile sensing," "digital phenotyping," "smartphone sensing," "personal sensing," and "smartphone assessment." Terms were developed through discussion among this study's team, consisting of experts in digital phenotyping, digital health technologies, behavioral medicine, and computer science. A sample PubMed search string is as follows: ("digital phenotyping" OR "smartphone sensing" OR "mobile sensing") AND (mental OR behavioral OR health OR monitoring). Full search strategies for each database are provided in Multimedia Appendix 1. In addition to database searching, we used backward reference screening

XSL•FO
RenderX

of included papers to identify additional studies; we did not search study registries or gray literature. The search strategy followed PRISMA-S guidelines and was informed by prior reviews in digital phenotyping and mobile sensing. We did not use any standardized search filters.

All records retrieved from databases were exported or saved in PDF format from their respective databases and managed manually. Before formal screening, records were manually reviewed to remove duplicate entries and studies that were clearly outside the scope of the review. A substantial proportion of retrieved records were excluded at this stage due to the use of wearable or auxiliary sensors, non–health-related applications, review papers, study protocols, or a primary focus on algorithmic development rather than health characterization. Following this initial filtering step, 111 records remained and proceeded to formal title and abstract screening. This approach is consistent with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidance and is reflected in the PRISMA flow diagram.

Screening followed a 2-stage process consistent with PRISMA-ScR [15]. Two authors (AD and JH) independently screened a total of 111 retrieved PDF records for relevance to smartphone-based sensing and health-related outcomes. A total of 43 studies were identified from PubMed, 21 from IEEE Xplore, 26 from ACM digital library, and 21 from Google Scholar (N=111). After this initial title and abstract review, the same reviewers examined the full text of the remaining studies to confirm that they met the inclusion criteria, which required the use of smartphone-based sensing and a defined health-related outcome. Screening was performed manually, and any disagreements between reviewers were resolved through discussion until consensus was reached. No formal interrater reliability statistic (eg, Cohen k) was calculated, as the screening process was exploratory and descriptive in nature. Reasons for exclusion at each stage, including the use of auxiliary sensors, EMA-only studies, or the absence of health-related outcomes, are summarized in the Results section.

We included only English-language peer-reviewed papers that (1) primarily used smartphone-based sensor data as the input for digital phenotyping studies, (2) described the use of smartphone-based digital phenotyping with relation to a predescribed health condition. We excluded (1) non-English language literature, (2) conference abstracts, white papers, or other policy papers, (3) papers that described digital phenotyping through the use of smartphones as a platform to exclusively perform EMAs or other surveys without using on-board smartphone sensor data.

To find studies for this scoping review, we first determined the eligibility criteria for our studies. For a work to be considered within our scope for this scoping review, it must meet all three of the inclusion criteria listed here: (1) the studies must use digital phenotyping approaches that exclusively use smartphone-based sensor data as input; (2) the studies must not include approaches that exclusively use smartphones as a method for manual data entry, such as EMAs, surveys, etc; and (3) the studies must use digital phenotyping explicitly in the characterization of one or more well-defined health conditions. This includes studies that looked at monitoring the general well-being of an individual.

We did not include studies in this scoping review that solely focused on algorithmic advances or enhanced methods to capture data without analyzing it in the context of specific health conditions.

Using the keywords listed above, we identified 111 studies, of which 46 did not meet the inclusion criteria. This resulted in 65 studies that form the basis of this review. Table 1 shows the frequency of the health conditions represented. The most commonly studied condition was depression (n=18), followed by bipolar disorder (n=10) and stress (n=9). The least represented conditions were substance use disorder (cannabis; n=1) and attention-deficit/hyperactivity disorder (ADHD, n=1).

Before we proceed further, a note about terminology: in the rest of the paper we use the terms study and studies interchangeably. As each of the 65 studies we looked at in this review reported on 1 study, the 2 terms are synonymous in the context of this study.

XSL•FO
RenderX

Table . This table depicts the counts of health conditions considered by the studies in this scoping review. For studies that cover multiple health conditions, each condition is counted (ie, studies that consider both anxiety and depression are present in the counts for both conditions).

Condition	Count
Depression	18
Bipolar disorder	10
Stress	9
Schizophrenia	8
Anxiety	7
SUD[a] - alcohol	5
Parkinson disease	4
Sleep apnea	3
SUD - opioids	2
Well-being	2
ADHD[b]	1
SUD - cannabis	1

[a]SUD: substance use disorder.

[b]ADHD: attention-deficit/hyperactivity disorder.

Results

Overview

Before we delve into the details of the findings of our scoping review, we quickly present an overview of prior scoping reviews in the digital phenotyping space and argue why our work is different. We then delve into the findings of our scoping review. Figure 1 depicts a PRISMA flowchart of the review process.

XSL•FO
RenderX

Figure 1. PRISMA flow diagram illustrating the screening and selection process for studies included in this scoping review. The initial search across 4 databases (Google Scholar, IEEE Xplore, ACM Digital Library, and PubMed) identified 111 records published between January 1, 2012, and October 31, 2025. After the removal of nonrelevant records during title and abstract screening, 65 full-text papers met the inclusion criteria for smartphone-based digital phenotyping using on-device sensor data and a defined health-related outcome. Studies were excluded for reasons including use of auxiliary sensors or wearables as primary data sources, EMA-only designs without sensor data, absence of a health-related condition or outcome, or non–peer-reviewed publication type. The final set of 65 studies formed the basis of the analysis presented in this review. EMA: ecological momentary assessment; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

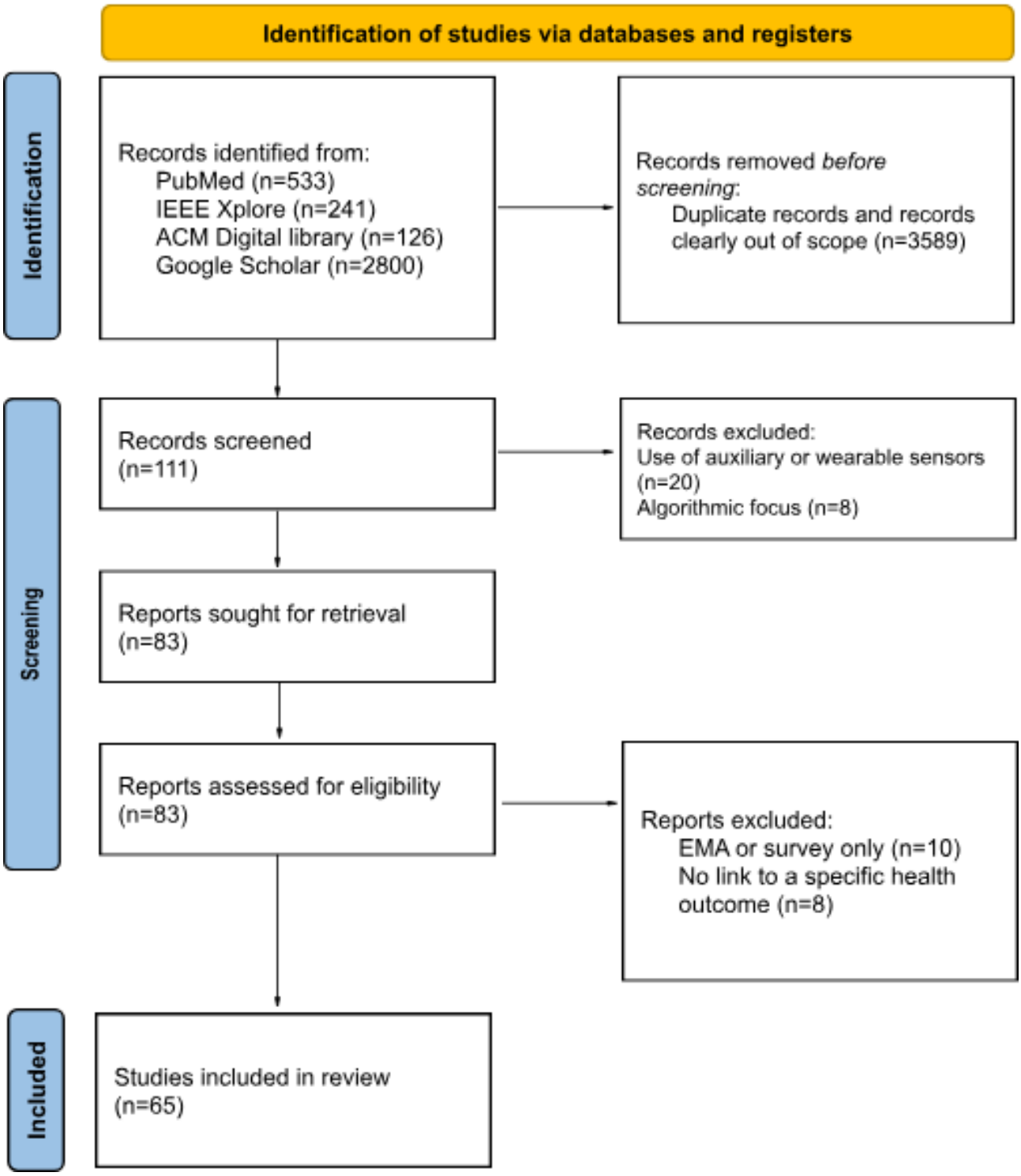

Prior Scoping Reviews

There are existing reviews that have explored the use of digital phenotyping. Our search yielded 21 prior reviews. We classify these reviews based on the types of sensing technology used: smartphones, smartwatches, and auxiliary sensors. We define auxiliary sensors to be any sensing device outside of smartphones and smartwatches, such as smart textiles or sensors worn on the body. A brief overview of the reviews reported on in this section can be found in Table 2, which we now elaborate on. These prior reviews were not part of the review process but

XSL•FO
RenderX

are only presented to show the distinction between the existing reviews and our scoping review.

Table . Summary of existing digital phenotyping reviews (n=19). This table summarizes prior reviews and reviews examining digital phenotyping across different sensing modalities and health domains. Each citation lists the number of studies covered, primary sensing technologies (eg, smartphones, smartwatches, or auxiliary sensors), publication year, and main health focus.

Citation	Studies (n)	Sensing modalities used	Year	Area of interest
[17]	119	Smartphones	2018	Health and well-being
[18]	18	Smartphones	2019	Bipolar disorder
[19]	224	Smartwatches	2020	Presurgical risk assessment, postsurgical outcome prediction
[20]	62	Smartphones, smartwatches, auxiliary sensors	2021	Bipolar disorder
[21]	92	Smartphones, smartwatches, auxiliary sensors	2021	Depression, bipolar disorder, anxiety, stress
[22]	51	Smartphones, smartwatches	2022	Depression
[23]	46	Smartphones, smartwatches, auxiliary sensors	2022	Machine learning practices
[24]	31	Smartphones, smartwatches	2022	Available datasets and smartphone apps
[25]	81	Smartphones, auxiliary sensors	2023	Parkinson disease, Alzheimer disease
[26]	11	Smartphones, smartwatches, auxiliary sensors	2023	Anxiety
[14]	40	Smartphones	2023	Schizophrenia, depression, general well-being
[27]	29	Smartphones, smartwatches	2023	Mood disorders, schizophrenia-spectrum
[28]	47	Smartphones	2023	Social behavior sensing
[29]	59	Smartphones, smartwatches	2023	Depression, anxiety, stress, schizophrenia, mood, general well-being
[30]	35	Smartphones	2024	Youth depression and anxiety
[31]	24	Smartphones	2024	Machine learning practices
[32]	40	Smartphones	2024	Depression, stress, anxiety
[33]	112	Smartphones	2025	Digital phenotyping pipeline
[34]	16	Smartphones, smartwatches, auxiliary sensors	2025	Stress monitoring
This scoping review	65	Smartphones	2025	Depression, bipolar disorder, schizophrenia, anxiety, stress, general well-being, substance use, sleep apnea, Parkinson disease

Earlier reviews predominantly examined mental health outcomes such as depression, bipolar disorder, and stress, whereas more recent work expanded to neurological, physiological, and multimodal applications. The final row (gray) highlights the present scoping review, which uniquely synthesizes 65 smartphone-only studies spanning 9 health domains. About a third of the reviews (n=7) focus exclusively on using smartphones for digital phenotyping. Three of these reviews focus on describing phenotyping in the context of specific health conditions: depression [31], bipolar disorder [18], and a combination of depression, stress, and anxiety [32]. The remaining reviews include a scoping review of 35 studies on depression and anxiety in youth [30], a review of 47 studies using smartphones for social behavior sensing [28], a large review of 112 studies characterizing digital phenotyping pipelines [33], and a systematic review of 119 studies on

XSL•FO
RenderX

smartphone-based health sensing [17]. Additionally, Bufano et al [27] report on 29 clinical studies using smartphones for psychiatric relapse monitoring.

Another third of the reviews (n=6) cover a combination of smartphone and smartwatch-based digital phenotyping. Similar to the previous category, 2 of the studies focus on phenotyping work addressing mental health conditions: depression [22] and depression, anxiety, stress, schizophrenia, mood, and general well-being [29]. One review covers a purely smartwatch-based digital phenotyping to assess presurgical risk and postsurgical care [19]. The last review in this category reports on publicly available applications and datasets for digital phenotyping that use both smartphones and smartwatches [24].

The remaining reviews (n=8) report on digital phenotyping using a combination of smartphones, smartwatches, and auxiliary sensors. Like the prior 2 sections, the reviews in this category focused on mental health conditions: Alzheimer disease [25]; depression, bipolar disorder, anxiety, and stress [21]; bipolar disorder [20]; and anxiety [26]. The studies by Dlima et al [23] and Lee et al [14] focus on the use of ML to interpret phenotyping data, while Bufano et al [27] examine combined sensing modalities in psychiatric relapse studies. Most recently, Kallio et al [34] provide a review of approximately 16 studies focused on continuous stress monitoring in cognitively demanding work environments using smartphones, smartwatches, and ambient sensors. Their review emphasizes real-world deployment, unobtrusiveness, and user acceptance, expanding the digital phenotyping literature into long-term occupational health monitoring.

Most of these reviews (n=15) cover research that uses more sensing modalities than those available on smartphones. Those that do cover purely smartphone-based digital phenotyping (n=7) are scoped in terms of specific mental health conditions, youth populations, or derived behavioral features.

Our review is different from these prior reviews for three reasons: (1) it focuses exclusively on smartphone-based digital phenotyping; (2) its scope is much more expansive, in that it covers any and all health conditions (and not just mental health conditions); and (3) it focuses on describing the relationships between behaviors captured through phenotyping in the context of specific health conditions and not on algorithms used to extract the behaviors.

Findings of This Scoping Review

Our aim is to understand the current state of smartphone-based digital phenotyping search which led to the identification of 111 records. After title or abstract screening, 65 studies met the inclusion criteria and proceeded to full-text review; all 65 were included in the synthesis. Across the 65 included studies, we conducted a descriptive synthesis to summarize the sample sizes, study durations, and participant populations. The median sample size across all studies was 52 participants (IQR=26 - 126), with a median study duration of 8 weeks (IQR 3 wk to 6 mo). Studies commonly used multiple different measurements for ground truth. Clinical rating scales (eg, Patient Health Questionnaire-9, Hamilton Depression Rating Scale, YMRS, Generalized Anxiety Disorder-7, and Positive and Negative Affect Schedule) were used in 63% (41/65) studies, EMAs in 28% (18/65) studies, clinician-confirmed diagnoses in 14% (9/65) studies, and physiological validation measures (eg, polysomnography or gait laboratory testing) in 5% (3/65) studies. Each of these studies was additionally described by the study setting. Controlled laboratory studies are conducted under precise conditions. Naturalistic “daily life” studies capture behavior in real-world environments. Hybrid study designs intentionally combine both settings within the same study [35].

One way of examining smartphone-based digital phenotyping research is by considering the conditions that studies target. These works can be grouped into 6 broad categories: mood disorders, stress-related disorders, schizophrenia, disorders from the use of psychoactive substances, nervous system disorders, and general well-being management.

In all of these efforts, the smartphone sensors were used to collect a variety of contextual information about the individual, such as call or text logs, phone or app usage, location, accelerometer, and microphone data. These were used to extract phenotypes, such as social functioning, physical activity, movement patterns, and sleep patterns, which were associated with the specific health condition being considered. These phenotypes are then analyzed using some form of ML algorithm to answer questions about specific health conditions. To build such models, knowing the ground truth of the health condition of the participants is essential. This ground truth was usually assessed using EMAs or an appropriate scale based on the condition being tracked. In our description below, we do not focus on the intricacies of the ML algorithms used, as they are largely similar across most of the studies. Rather, as much as possible, we focus on the explanations that these models offer about the relationship between sensing, phenotypes, and the health condition being considered.

For each health condition we discuss in this section, we categorize the work done in one or more of the following focus categories: monitoring, diagnosis, and intervention. Monitoring involves collecting and analyzing smartphone-based digital phenotypes to track the health condition being considered. Diagnosis uses smartphone-based digital phenotypes to identify the presence of the health condition or risks thereof. Intervention involves implementing strategies or treatments, often as a response to monitoring a health condition. Not every health condition is worked on in all 3 categories.

Mood Disorders

Mood is defined as a pervasive and sustained feeling that is experienced internally [36]. It impacts nearly all aspects of a person’s behavior externally. Mood disorders are described by marked disruptions in emotions. For example, experiencing severe lows called depression or highs called hypomania or mania [36]. Though there are a variety of mood disorders, smartphone-based digital phenotyping work in the context of mood disorders has targeted only 2 major conditions in this regard: depression and bipolar disorder, which we now describe.

Depression

Depression (otherwise known as depressive disorder and major depressive disorder) involves low moods and a loss of pleasure

XSL•FO
RenderX

or interest in activities over a prolonged period of time [37]. Over the years, studies have shown that passive sensor data collected from smartphones can be associated with depression [38-40]. Much of the digital phenotype effort to explore and characterize depression has focused on 4 types of phenotyping efforts: tracking social interactions, tracking location diversity, vocal characteristics, and tracking sleep quality. For these approaches, the ground truth was collected via EMA or self-reported using pertinent scales, such as the Personal Health Questionnaire [41], the Center for Epidemiological Studies-Depression Scale [42], the Children's Depression Rating Scale-Revised [43], and the Clinical Global Impressions-Severity Scale [44]. Phenotyping work done in the depression space was primarily focused on the monitoring of depression, with some work in the diagnosis focus category. Table 3 provides a summary of the studies reviewed in this section.

XSL•FO
RenderX

Table . Summary of smartphone-based digital phenotyping studies focused on depression (n=16). Studies examined depressive symptoms and mood-related behaviors through passive smartphone sensing. Common phenotypes included sociability, mobility, activity, and sleep patterns. Sensors most frequently used were GPS, accelerometer, screen state, and communication logs, with PHQ-9[a] as the most common ground-truth measure.

Citation	Condition	Phenotypes considered	Sensors used	Ground truth	Participants (n)	Duration	Focus	Study setting
[38]	Depression	Sociability, movement patterns	App use, gyroscope	PHQ-9	558 adults	10.7 days average	Monitoring	Naturalistic setting (app recruitment)
[39]	Depression	Movement patterns	GPS, app logs, call logs, text logs, IMU[b]	PHQ-9	38 adults	2 years	Monitoring	Naturalistic setting (Spain)
[40]	Depression	Movement	GPS, weather, light	DASS[c], PANAS-X[d]	31 adults	2 weeks	Monitoring	Naturalistic setting (United States)
[45]	Depression	Sociability	App use, call logs, GPS, battery, screen state	CES-D[e]	120 adults	1 month	Monitoring	Naturalistic setting (China)
[46]	Depression	Activity, sociability	GPS, screen state	PHQ-9	40 adults	2 weeks	Monitoring	Naturalistic setting (United States)
[47]	Depression	Sleep, activity, sociability	Call logs, accelerometer, light, GPS, screen state	PHQ-9	47 adults with diabetes	20 weeks	Monitoring	Naturalistic setting (India)
[48]	Depression	Sociability, activity, movement	Call or text logs, GPS, accelerometer, light, screen state	PHQ-9, HAM-D[f], HAM-A[g]	13 adolescents	8 weeks	Monitoring	Naturalistic setting (United States)
[49]	Depression	Sociability, mobility	GPS, gyroscope, call or text logs	CDRS-R[h]	24 adolescents	8 weeks	Monitoring	Naturalistic setting (South Korea)
[50]	Depression	Activity, movement	GPS	PHQ-9	48 college students	10 weeks	Monitoring	Naturalistic setting (United States)
[51]	Depression	Movement patterns	GPS	PHQ-8[i]	28 adults	2 months	Monitoring	Naturalistic setting (app store recruitment)
[52]	Depression	Activity, sociability	Accelerometer, GPS, WiFi, call or text logs	PHQ-9	126 adults	9 months	Monitoring	Naturalistic setting (app store recruitment)
[53]	Depression	Movement patterns	Accelerometer, GPS	PHQ-9	33 adults	11 weeks	Monitoring	Naturalistic setting (Bangladesh)
[54]	Depression	Activity, sleep disruption	GPS, accelerometer, logs, screen state	Self-reported mood	100 college students	28 days	Monitoring	Naturalistic setting (United States)
[55]	Depression	Sociability	Microphone, text, GPS	PHQ-9	302 college students	6 months	Monitoring, diagnosis	Hybrid setting (United States)
[56]	Depression	Psychological stress	Microphone	BDI[j]	1814 adults	2 years	Monitoring, diagnosis	Naturalistic setting (Japan)

XSL•FO
RenderX

Citation	Condition	Phenotypes considered	Sensors used	Ground truth	Participants (n)	Duration	Focus	Study setting
[57]	Depression	Sociability	Microphone, call or text logs, GPS	PHQ-9, GAD-7[k]	70 college students	28 weeks	Diagnosis	Hybrid setting (United States)

[a]PHQ-9: Patient Health Questionnaire-9.
[b]IMU: Inertial measurement unit.
[c]DASS: Depression Anxiety Stress Scales.
[d]PANAS-X: Positive and Negative Affect Schedule—Expanded Version.
[e]CES-D: Center for Epidemiologic Studies Depression Scale.
[f]HAM-D: Hamilton Depression Rating Scale.
[g]HAM-A: Hamilton Anxiety Rating Scale.
[h]CDRS-R: Children's Depression Rating Scale-Revised.
[i]PHQ-8: Patient Health Questionnaire-8.
[j]BDI: Beck Depression Inventory.
[k]GAD-7: Generalized Anxiety Disorder-7.

Concerning monitoring, one of the most common ways of characterizing depression and depressive episodes is by looking at changes in social interaction. The social interaction of a person can be approximated by measuring the number of incoming calls a person receives as well as the amount of time people spend on social media apps [45]. It was found that a lower volume of incoming calls and a higher frequency of social media app use were indicative of depression, indicating that individuals who were depressed were not communicating with others and were instead turning to their phones as a potential form of self-soothing [45]. Similar findings on the presence of inverse correlation between number and length of phone calls, text messages, and overall phone use patterns and depression were found in 4 different studies [46-48,55]. However, in 1 study, the authors looked at an adolescent population in South Korea (as opposed to adults in the aforementioned work) and found that those with depression received more calls compared to the control participants [49]. Though the study did find that a person spending more time on smartphones may be associated with depression, as mentioned above.

Phone call and text logs as a proxy for social connection were not the only way to infer and characterize depression. Another common approach was to look at location features (based on the GPS and inertial measurement unit). It has repeatedly been found that movement patterns may be a similarly strong proxy to mental state [50,51]. A case in point is a study that looked at individuals with spinal cord injuries. For such individuals, there was a direct correlation between a decrease in the number of locations visited (presumably due to their condition) and depression [58]. This was found to be true not just for adults but also for younger populations as well [48,59]. A similar pattern was also reported in other studies that showed that the variance, number of location clusters, and location entropy all decreased in relation to increased depression scores [50,51]. When the location data was analyzed by making assumptions about the location context, such as duration of time away from home, the information could be used to provide interventions for individuals with depression [52]. Beyond just diversity of location, in 1 study, the researchers found that physical activity in general was associated with depression, with lower physical activity (in terms of distance variance from any starting point, greater resting time, greater time using the phone in a lying position, fewer steps, less exercise, and greater staying at home) being associated with increased depression [53].

A few studies have also looked at using sleeping patterns to characterize depression. Measuring sleep quality from smartphones, unlike wearable devices (eg, smartwatches), is not straightforward. Consequently, researchers often make broad assumptions about the daily routines of a typical person and check to see if smartphone usage patterns can detect a break in this routine. For instance, in one of these studies, a higher number of instances of "screen on" states recorded from 12 AM to 6 AM was interpreted as indicating a lower sleep quality or disturbed sleep schedule in individuals [47]. This pattern of being up at night was found to be associated with the presence of depression. In a similar study on college students, it was found that higher variance in sleep patterns was associated with depression [54].

Concerning diagnosis, in recent years, an interesting and different approach to detecting the presence of depression in individuals has been explored using vocal characteristics. In 2 different studies, the authors looked at changes in vocalization and tone based on features such as vocal fry, vocal jitter, and monotony measured via unscripted and scripted audio recordings to determine the presence of depression [55,57]. A similar study also analyzed voice features during phone calls to determine the presence of depression [56]. In one of the aforementioned studies, the researchers tried to extend their vocal characteristics-based depression models to the most extreme cases where individuals are experiencing suicidal ideation [57]. They found that such models were not sufficient and required additional context to predict suicidal risk [57].

Bipolar Disorder

Bipolar disorder is a mental health condition that causes individuals to experience extreme mood swings, typically between emotional highs (known as mania or hypomania) and lows (depression) [60]. There are several types of bipolar and

XSL•FO
RenderX

related disorders, not all of which require for both manic and depressive episodes to be present, but all of which require clear changes in mood, energy, and activity levels [60]. Though very similar to the phenotypes used in characterizing depression, the way they are used for characterizing bipolar disorder is somewhat different. As discussed in the prior section, we know that it is possible to identify depression through a variety of methods, but it is important to note that depressive episodes in bipolar disorder may not be as intense or last as long [60]. In addition, mixed states (having depressive symptoms and manic symptoms at the same time) and manic episodes are just as important to recognize in individuals with bipolar disorder, meaning that while some techniques may be useful between both disorders, they are not the same disorder and cannot be treated as such. With respect to smartphone-based digital phenotyping, considerable work has also focused on exploring diseases such as bipolar disorder. These efforts have primarily looked at 4 types of phenotyping efforts: tracking phone usage, tracking location diversity and routine, and tracking sleep quality. Further, the ground truth for bipolar disorder was collected via EMA or self-reported using a combination of scales such as Hamilton Depression Rating Scale [61] and YMRS [62]. Phenotyping work in bipolar disorder was primarily focused on the monitoring category, with some work for both diagnosis and intervention. Table 4 provides a summary of the research reviewed in this section.

XSL•FO
RenderX

Table . Summary of smartphone-based digital phenotyping studies focused on bipolar disorder (n=11). Studies used passive smartphone sensors to assess behavioral and physiological markers associated with bipolar disorder. Commonly examined features included sociability, activity levels, and sleep-related changes. Ground-truth measures primarily relied on clinical scales such as HAM-D[a], YMRS[b], and QIDS-SR[c].

Citation	Condition	Phenotypes considered	Sensors used	Ground truth	Participants (n)	Duration	Focus	Study setting
[63]	BD[d]	Physical activity, sociability	GPS, cell tower pings, accelerometer, call logs, text logs	Daily self-reported mood, HAM-D, YMRS	13 adults with BD	12 months	Monitoring	Naturalistic setting (Germany)
[64]	BD	Sociability	Accelerometer, GPS, call data	HAM-D, YMRS	10 adults with BD	12 weeks	Monitoring	Naturalistic setting (Austria)
[65]	BD	Physical activity, sociability	Screen state, text logs, call logs, cell tower pings	HAM-D, YMRS	29 adults with BD	12 weeks	Monitoring	Naturalistic setting (Denmark)
[66]	BD	Sociability	Call logs, text logs	Self-reported mood, YMRS, HAM-D	26 adults with BD, 12 controls	28 days	Monitoring	Naturalistic setting (United States)
[67]	BD	Movement, activity, sociability	Accelerometer, GPS, WiFi, Bluetooth, call logs	Mood self-report, HAM-D, MSS[e], ADS[f]	10 inpatients with BD	8 weeks	Monitoring	Naturalistic setting (Austria)
[68]	BD	Activity, movement patterns	GPS, WiFi, cell tower pings	QIDS-SR	29 adults with BD, 20 controls	3 months	Monitoring	Naturalistic setting (United Kingdom)
[69]	BD	Movement, sleep disruption	GPS	ASRM[g], DASS-21[h]	159 adults with BD	1 year	Monitoring	Hybrid setting (Taiwan)
[70]	BD	Activity, sociability, movement	Accelerometer, microphone, battery, screen state, logs, calls	SRM-5[i]	7 adults with BD	4 weeks	Monitoring	Naturalistic setting (Germany)
[71]	BD	Sociability	Call or text logs, microphone, screen state, accelerometer, cell tower pings	Mood self- report	28 adults with BD	12 weeks	Monitoring, diagnosis	Naturalistic setting (United States)
[72]	BD	Sleep, activity	Accelerometer, light sensor	PSQI[j]	22 adults with BD, 23 controls	3 weeks	Intervention	Hybrid setting (Austria)
[73]	BD	Sociability, stress, sleep	Accelerometer, WiFi, logs, screen state, Bluetooth, microphone	Mood self-report, HAM-D, YMRS	18 adults with BD	5 months	Monitoring, diagnosis, intervention	Naturalistic setting (app store recruitment)

[a]HAM-D: Hamilton Depression Rating Scale.

[b]YMRS: Young Mania Rating Scale.

[c]QIDS-SR: Quick Inventory of Depressive Symptomatology—self-report.

XSL•FO
RenderX

[d]BD: bipolar disorder.

[e]MSS: Mania Self-Rating Scale.

[f]ADS: Common Depression Scale.

[g]ASRM: Altman Self-Rating Mania Scale.

[h]DASS-21: Depression Anxiety Stress Scales-21.

[i]SRM-5: Social Rhythm Metric Scale-5.

[j]PSQI: Pittsburgh Sleep Quality Index.

Concerning monitoring, like tracking depression, several efforts in recent years have focused on tracking a bipolar individual's mood over time. This information can be used to determine if a person is having a depressive or manic episode. In this regard, one of the common phenotypes is smartphone use pattern [64,65,71,73]. In 1 study, the researchers used features of phone calls themselves, such as total length, average call length, number of unique calls, total number of calls, etc, along with sounds emanating during a phone call to determine the current mood or change in moods of bipolar individuals [64]. The study indicated the presence of some kind of association between bipolar states and these phone usage features, but those exact associations were not clear. In 2 other similar studies, it was found that severe depressive symptoms of bipolar disorder were associated with longer "screen on" state per day, more received calls per day, fewer outgoing calls per day, fewer answered incoming calls per day, and less movement [65,66]. On the other hand, manic symptoms showed the exact opposite pattern as described above.

Another approach to characterizing bipolar disorder has been to look at movement or physical activity using smartphone-based GPS data [67-69]. In another study, researchers found that physical activity changes as mood states transition from one to another [67]. When transitioning from a manic state to a baseline normal state, physical activity decreases, but it increases when transitioning from a depressed state to baseline [67]. In a study focused on quantifying depressive states in those with bipolar disorder, the researchers derived a variety of features from GPS data from an individual's smartphone and were able to use them to predict acute depressive states with high accuracy [68].

Concerning diagnosis, in the context of smartphone-based bipolar disorder diagnosis, researchers looked at vocal patterns along with self-reported surveys on a variety of mood and sleep patterns, along with smartphone usage metadata from individuals with bipolar disorder to detect manic and depressive states [71]. Overall, it was found that while it was possible to classify bipolar states on voice features alone, combining different passively sensed and self-report data with voice features creates a more robust model in terms of detecting subtler manifestations of bipolar states (eg, hypomania as opposed to mania) [71]. Another potential indicator of a bipolar person's mood is in patterns of app use on the smartphone by a bipolar individual [73]. In particular, the way that bipolar individuals interact with their phones has a strong correlation to what moods they are experiencing at the time. For instance, in 1 study, the researchers found that increased social interaction is associated with a decrease in self-reported stress level, irritability, and mixed mood. This indicated that increased social interaction is a good marker for normal social functioning outside of mood swings. Further, patterns in the average number of apps being used can be used to detect the presence of manic or depressive episodes, with a higher number of apps used being positively associated with self-reported stress, and a decrease in the average number of apps used was suggested to indicate a depressive episode [73].

Concerning intervention applications, finally, sleep disturbances have been shown to be early warning signs of manic or depressive episodes in individuals with bipolar disorder [72]. As stated above, smartphone use (screen on or off, and accelerometer readings) can provide insight into lower sleep quality and satisfaction with sleep, alluding to potential sleep disturbances [73]. In a different study, the researchers showed that a smartphone-based app could be a valid tool for measuring sleep onset and wake-up time, to measure sleep disturbances. This information can then be used for the management of the depressive symptoms of bipolar disorder [72].

Stress-Related Disorders

Stress is the perception of an individual that their environmental taxes them or exceeds their adaptive capacity [74]. Stress-related disorders are maladaptive responses to short- or long-term exposures to physical or emotional stressors. Smartphone-based phenotyping work for stress-related disorders can be broadly categorized into 2 groups: anxiety disorders and stress quantification. We describe these below. Table 5 provides a summary of the reviewed work in this section.

XSL•FO
RenderX

Table . Summary of smartphone-based digital phenotyping studies focused on anxiety, stress, and schizophrenia (n=18). Studies span sociability, movement or activity, sleep, and psychological stress phenotypes using signals such as accelerometer, GPS, microphone, app logs, and communication metadata. Ground-truth measures include standardized scales, self-reports, and clinical or laboratory validations.

Citation	Condition	Phenotypes considered	Sensors used	Ground truth	Participants (n)	Duration	Focus	Study setting
[75]	Anxiety	Sociability, movement patterns	Accelerometer, call logs, text logs	SIAS[a], DASS-21[b], PANAS[c]	59 college students	2 weeks	Monitoring	Naturalistic setting (United States)
[76]	Anxiety	Sociability, physical activity	App log, power level, screen, GPS, call or text logs	SIAS, UCLA-LS[d]	127 adults	30 days	Monitoring	Naturalistic setting (China)
[77]	Anxiety	Sociability, physical activity	Light, accelerometer, gyroscope, app logs	STAI[e]	20 healthy adults	30 days	Monitoring	Hybrid setting (Japan)
[78]	Stress	Movement, activity, sociability, psychological stress	Accelerometer, GPS, microphone, WiFi, call or text logs, app logs	Mood self-report	30 employees	8 weeks	Monitoring	Naturalistic setting (Italy)
[79]	Stress	Sleep, activity, sociability	Microphone, GPS, WiFi, accelerometer, light sensor	PSS[f], PHQ-9[g], UCLA-LS	47 young adults	10 weeks	Monitoring	Naturalistic setting (United States)
[80]	Stress	Physical activity, sociability	Accelerometer	OLBI[h]	30 employees	8 weeks	Monitoring	Naturalistic setting (Italy)
[81]	Stress	Sociability	Call or text logs, Bluetooth	Stress self-report	117 graduate students	7 months	Monitoring	Naturalistic setting (United States)
[82]	Stress	Activity, sociability, sleep, movement	Accelerometer, microphone, light, GPS, Bluetooth	PHQ-9, PSS, UCLA-LS	48 students	10 weeks	Monitoring	Naturalistic setting (United States)
[83]	Stress	Activity, sociability, stress, movement	GPS, WiFi, microphone, accelerometer	GPA[i], stress or mood self-report	48 students	10 weeks	Monitoring	Naturalistic setting (United States)
[84]	Stress	Psychological stress	Keyboard pressure	Mood self-report, event context	11 adults	1 laboratory session	Diagnosis	Controlled laboratory setting (United States)
[85]	Schizophrenia	Sociability	Call logs, text logs, microphone	BPRS[j], hospital relapse records	61 adults with schizophrenia	1 year	Monitoring	Naturalistic setting (United States)
[86]	Schizophrenia	Activity, sociability	Call logs, text logs, microphone	Social function diary	5 adults with schizophrenia	5 days	Monitoring	Naturalistic setting (location not stated)

XSL•FO
RenderX

Citation	Condition	Phenotypes considered	Sensors used	Ground truth	Participants (n)	Duration	Focus	Study setting
[87]	Schizophrenia	Movement patterns	GPS	Self-reported activity or context	85 adults with schizophrenia, 58 controls	7 days	Monitoring	Naturalistic setting (United States)
[88]	Schizophrenia	Sociability	GPS, accelerometer, screen, call or text logs	PHQ-9, GAD-7[k], self-reported symptoms	45 adults with schizophrenia, 43 controls	3 months	Diagnosis	Naturalistic setting (United States)
[89]	Schizophrenia	Movement, sociability	Accelerometer, GPS, call or text logs, screen, battery	Symptom report	17 adults with schizophrenia	3 months	Diagnosis	Naturalistic setting (United States)
[90]	Schizophrenia	Movement, activity, sociability, sleep	Accelerometer, GPS, app logs, light, call or text logs	Mood self-report	21 adults with schizophrenia	8 months	Diagnosis	Naturalistic setting (United States)
[91]	Schizophrenia	Activity level	Accelerometer	Self-reported activity	50 adults with schizophrenia, 70 controls	6 days	Diagnosis	Naturalistic setting (United States)
[92]	Schizophrenia	Sleep quality	GPS, screen, call or text logs, accelerometer	PSQI[l], laboratory sleep assessment	17 schizophrenia	3 months	Diagnosis	Naturalistic setting (United States)

[a]SIAS: Social Interaction Anxiety Scale.
[b]DASS-21: Depression Anxiety Stress Scales-21.
[c]PANAS: Positive and Negative Affect Schedule.
[d]UCLA-LS: University of California, Los Angeles Loneliness Scale.
[e]STAI: Spielberger State-Trait Anxiety Inventory.
[f]PSS: Perceived Stress Scale.
[g]PHQ-9: Patient Health Questionnaire-9.
[h]OLBI: Oldenburg Burnout Inventory.
[i]GPA: Grade Point Average.
[j]BPRS: Brief Psychiatric Rating Scale.
[k]GAD-7: Generalized Anxiety Disorder-7.
[l]PSQI: Pittsburgh Sleep Quality Index.

Anxiety Disorder

Anxiety disorders usually involve a persistent feeling of anxiety or dread, which can interfere with daily life. People living with such disorders experience frequent anxiety for extended periods of time [93]. For these efforts, the ground truth was collected via EMA or self-reported using pertinent scales, such as the General Anxiety Disorder Scale (Generalized Anxiety Disorder-7) [94]. There were relatively fewer studies in this space, all of which focused on monitoring.

Concerning monitoring, in 1 study, researchers showed that social function derived based on basic statistics collected about phone calls made, received, missed, diversity of callers, etc, can be used to detect the presence of social anxiety disorder in an individual [75]. Other studies found that lack of social connection in terms of lower call use and increased social media application use [76], and use of phones in the dark [77] were associated with higher levels of anxiety disorders. Further, days of the week may also be important in recognizing anxiety, as self-reported anxiety appears to increase toward the end of the weekend and decrease toward the end of the week, possibly aligning with perceived stress or uncertainty about workloads during the week [77]. Overall, the results for characterizing anxiety disorders were very similar to characterizing depression, such that discerning between one and the other may be challenging. For instance, researchers showed a weak and positive relationship between the predicted social anxiety disorder symptom and depression when capturing social functioning of an individual based on phone call usage statistics [75].

XSL•FO
RenderX

Stress Quantification

Psychological stress is another condition that has been targeted as part of smartphone-based digital phenotyping research. Though experiencing psychological stress is not an uncommon condition for most individuals, prolonged or significant stress has been linked to a variety of physical and mental illnesses [78]. Prolonged stress may create a space for burgeoning mental illnesses such as depression and anxiety to flourish [37,78,79]. Characterizing stress typically relies on the fusion of a variety of behavioral information about an individual, such as their movement patterns, physical activity, social interaction, and sleep patterns, much like works described in the section above. Phenotyping work related to stress was in the monitoring focus category.

Concerning monitoring, in looking at monitoring stress, researchers looked at automatic stress detection in completely unconstrained environments. The study showed that it was possible to extract time and frequency domain features from accelerometer data to characterize routine and associate these features with self-reported stress levels [80]. In a similar study, the researchers explored more complex fusion models where the relationship between self-reported daily stress and sensor-derived activity changes, speech duration, changes in location, and sleep patterns was associated with changes in participants' stress levels over the course of the study [78]. In a similar work, the authors found that speech duration, movement information (measured using GPS and WiFi), and sleep duration were associated with self-reported daily stress levels, while their activity information (measured using an accelerometer) was associated with changes in their subjective loneliness [79]. Another approach to measuring stress proposed the fusion of information from call and text logs (this included observing broader patterns from calls or text, such as percentage of calls done during the night, percentage of initiated calls during the night, text response rate, text response latency, and the percentage of initiated texts) combined with proximity information (collected from Bluetooth signals from neighboring devices) along with weather information (temperatures, precipitation, humidity, visibility, etc) to accurately predict the presence of stress [81].

All these aforementioned works focused on work-related stress. Another group of individuals whose stress has been studied in some detail is college students. In 1 study, the researchers developed an app called StudentLife for continuously sensing the day-to-day impact of college workload on stress, sleep, activity, mood, sociability, mental well-being, and academic performance among a single class of 48 students across a 10-week term [82]. Analysis of the data collected shows that students at the start of the term had high positive affect and conversation levels, low stress, and healthy sleep and daily activity patterns. However, as the term progressed and the workload increased, stress appreciably increased while positive affect, sleep, conversation, and activity dropped off [82]. A follow-up study looked at a more general approach to stress quantification in college students, using a geo-fencing approach on the GPS data from the smartphones carried by students [83]. The various spaces students frequented were designated as "study spaces" and "party locations"; and in each of these locations, the student's behavior is recorded based on smartphone activity and audio collected from the phone's microphone. The results indicated that students who reported decreasing stress levels throughout the term (self-reported) obtained higher GPAs [83].

Finally, researchers used keyboard pressure as a proxy for detecting the presence of stress. Their idea was that the increase in muscle tension in response to an individual experiencing stress would lead to increased keyboard pressure when typing. In this study, the researchers asked individuals to recall a stressful and relaxing event in the recent past and write about it using an iPhone with a modified QWERTY keyboard. They then logged the timing, position, and pressure values associated with each keyboard interaction. Subsequently, the differences in typing pressure across relaxed and stressful expressive writing tasks were associated with differences in self-reported stress [84].

Delusional Disorders

A delusion is a false belief based on an inaccurate interpretation of an external reality despite evidence to the contrary [95]. In the context of delusional disorders, smartphone-based phenotyping work has exclusively focused on schizophrenia. Schizophrenia is a significant and often life-altering mental disorder characterized by episodes, which include persistent delusions and/or hallucinations, disorganized thinking and behaviors, and instances of agitation or negative symptoms. The term negative symptoms describes a lessening or absence of normal behaviors and functions related to motivation and interest, or verbal or emotional expression, which can manifest as blunted affect, reduction in speech, goal-directed activities, motivation, and experience of pleasure [96]. We summarize our findings in this space below. Table 5 provides a summary of the reviewed research in this section.

Schizophrenia

Typically, smartphone-based digital phenotyping schizophrenia research has focused on identifying predictors of negative symptoms. This is done by tracking data such as call patterns, location, physical activity, and sleep quality. The ground truth in most of these studies was collected via EMA. Phenotyping work with respect to schizophrenia space was in both monitoring and diagnosis focus categories.

Concerning monitoring, relatively few works have looked into applying phenotyping to people with schizophrenia. In 1 study, the researchers used patterns in incoming and outgoing calls and their duration, incoming and outgoing text messages, and captured the frequency and duration of nearby voices captured through a smartphone microphone to determine exactly how much social activity an individual with schizophrenia was engaging in. Overall, the researchers reported that reductions in the number and duration of outgoing calls, as well as the number of text messages, were associated with relapses [85].

In examining negative symptoms of schizophrenia, researchers explored the feasibility of clustering GPS location as a proxy for social functioning and determining where individuals with schizophrenia are spending their time. Subsequently, they added contextual labels to the location clusters to attain a general

XSL•FO
RenderX

understanding of social functioning for a specific individual [86]. In a similar study, researchers looked at GPS-derived indicators of location to determine if they could be used as a proxy for negative symptoms. They found that GPS-derived estimations of average distance traveled over time, distance from home, and the percentage of GPS samples at home were found to be highly associated with negative symptoms. Essentially, they found that less mobility was associated with greater negative symptom severity—possibly due to diminished motivation [87].

Concerning diagnosis, methods outside of GPS-based locations have also been explored to determine the level of social functioning among people with schizophrenia. In this case, routines and disturbances within those routines are seen as indicators of a relapse and may be used in the future as a marker for potential intervention [89]. Researchers have looked at smartphone sensor data (eg, call or text logs, phone screen status, accelerometer, and GPS information) as a proxy for social functioning to determine the relationship between measuring social rhythm (using phenotyping) and schizophrenia. They found that increased stability in social rhythm and routine was associated with improved symptom scores in schizophrenia. This also implies that using phenotyping data related to sociability, mobility, and physical activity, as we have described above, in detecting depression, anxiety, and bipolar disorder can all be associated with identifying negative symptoms of schizophrenia [88].

In another study involving adults with schizophrenia, the researchers looked at mobility before a relapse and found that visual hallucinations were associated with a change in sleeping patterns and the number of voice and nonvoice sounds captured, while auditory hallucinations decreased with new places visited and an increase in call duration [90]. Further, compared to individuals who did not require clinical intervention, there was a significantly higher rate of anomalies in the 2 weeks directly preceding the relapse. In a similar study, the researchers looked at movement information using purely accelerometry data from the smartphone to identify negative symptoms of schizophrenia. The findings of the study indicate that smartphone accelerometers, which are markers of movement and vigor (measured via accelerometry), hold promise in detecting negative symptoms of schizophrenia [91]. In an additional study, the researchers explored the use of accelerometer data to estimate sleep duration and sleep quality in individuals with schizophrenia. The results showed that these attributes were reliable to indicate the presence of sleep disturbances, which can be early warning signs for psychosis and relapse [92].

Disorders From the Use of Psychoactive Substances

A substance use disorder (SUD) is a mental disorder in which a person experiences an inability to control their use of psychoactive substances such as alcohol, cannabis, opioids, etc [97]. Understanding these SUDs is an emerging aspect of smartphone-based digital phenotyping. We group the effort in this space into 2 broad categories: alcohol use, and cannabis and opioid use. As with the previous efforts at phenotyping, these approaches use sensor data available from smartphones and associate behaviors such as social functioning, movement patterns, and physical activity to psychoactive substance use. Table 6 provides a summary of the research reviewed in this section. Phenotyping work in SUDs was in both monitoring and diagnosis focus categories for both alcohol use and cannabis and opioid use.

XSL•FO
RenderX

Table . Summary of smartphone-based digital phenotyping studies focused on substance use disorders (n=7). Studies examined behavioral and physiological markers related to alcohol, cannabis, and opioid use. Phenotypes primarily involved movement, sociability, and respiratory features derived from accelerometer, GPS, microphone, and app log data. Ground-truth measures included self-reported use, craving reports, and laboratory validation of BAC[a].

Citation	Condition	Phenotypes considered	Sensors used	Ground truth	Participants (n)	Duration	Focus	Study setting
[98]	SUD[b]–alcohol use	Movement, sociability	Accelerometer, battery, call or text logs, light, WiFi, Bluetooth, GPS, gyroscope, screen state	Self-reported alcohol use	30 adults	28 days	Monitoring	Naturalistic setting (location unclear)
[99]	SUD–alcohol use	Movement, sociability	Accelerometer, GPS, screen state, app logs, call or text logs, keystroke data	Mood and craving self-report	24 adults	30 days	Monitoring	Naturalistic setting (United States)
[100]	SUD–alcohol use	Physical activity	Accelerometer	Alcohol use self-report	7 adults	2 weeks	Diagnosis	Naturalistic setting (United States)
[101]	SUD–alcohol use	Physical activity	Accelerometer, gyroscope	Urine-based BAC	65 adults	1 in-laboratory session (5 h)	Diagnosis	Naturalistic setting (United States)
[102]	SUD–cannabis use	Activity, movement, sociability	GPS, accelerometer, call logs	Acute cannabis intoxication self-report	57 adults	30 days	Monitoring	Naturalistic setting (United States)
[103]	SUD–opioid use	Respiratory changes	Microphone	Simulated overdose events	209 injection events	3 - 5 min postevent	Diagnosis	Hybrid setting (United States)
[104]	SUD–opioid use	Movement patterns	GPS	Drug craving self-report	189 adults with OUD[c]	16 weeks	Diagnosis	Naturalistic setting (United States)

[a]BAC: blood alcohol concentration.

[b]SUD: substance use disorder.

[c]OUD: opioid use disorder.

Alcohol Use

One of the main areas of interest in smartphone-based phenotyping around the use of psychoactive substances is in studying alcohol use. These efforts take 2 basic forms: determining the level of intoxication and predicting cravings. Phenotyping work related to alcohol use falls under the monitoring and diagnosis categories.

Concerning monitoring, using phenotyping to determine how intoxicated an individual is has been the most common form of research in this space. One notable study in this domain looked at a wide range of mobile phone sensor data (eg, accelerometry, keyboard usage, screen state, call or text logs, battery status, etc) from young adults with a history of hazardous drinking over a period of a month. It found that time of day, day of week, screen duration, and frequency of keyboard presses had significant associations with heavy drinking, indicating behavioral changes during these episodes [98].

In a slightly different study, the researchers looked at alcohol craving in individuals with alcohol-associated liver disease via phone call status, movement patterns, phone screen state, and keystroke information. They found that the level of craving was directly proportional to loneliness, stress, and anxiety as measured primarily through a decrease in the entropy of their movement patterns [99].

Concerning diagnosis, a couple of other efforts looked at blood alcohol levels based on gait information inferred from smartphone sensors [100,101]. The researchers conducted these studies in an uncontrolled environment where it was not clear where the phone was located on the individual's body. These works used accelerometers and gyroscopes, respectively. Overall, they found some positive correlation between gait

XSL•FO
RenderX

measured through smartphone sensors and the level of intoxication of the individual.

Cannabis and Opioid Use

Not much work has been done when it comes to characterizing cannabis and opioid use using smartphone-based phenotyping. The extant work in this space can be grouped into the same 2 categories as the alcohol case: determining the level of intoxication and predicting cravings. Similar to alcohol use, phenotyping work related to cannabis and opioid use fell under the monitoring and diagnosis categories.

Concerning monitoring, for cannabis, researchers have looked at identifying slowed psychomotor function as a result of consuming cannabis. They measured psychomotor function in an individual based on their movement pattern and physical activity measured using GPS, accelerometer, and call logs. They found a smaller travel boundary and lower activity changes at times when the individual felt "high." Additionally, stronger body movements indicated by changes in the accelerometer may be indicative of impaired motor control or unsteady movements typically associated with higher levels of intoxication [102].

Concerning diagnosis, in a study related to the context of opioid use, the researchers explored the feasibility of using smartphone data to monitor respiratory changes indicative of an impending opioid overdose. They showed that they could convert a phone microphone into a short-range active sonar using frequency shift to measure chest motion and respiration. This work was done in a controlled operating room environment of the operating room, where overdose conditions were simulated; this method showed promising accuracy in identifying respiratory depression [103].

Finally, the increased interest in using smartphones to aid in the understanding and treatment of SUDs has created an opportunity to assess substance craving as reported in 1 study. By looking at the past 5 hours of solely GPS-based movement data from individuals with opioid-use disorder and correlating it with environmental exposure (measured as a function of the affluence of the locations they visited), the researchers could predict craving behaviors up to 90 minutes in the future for both heroin and cocaine. They were even able to predict stress in the individuals (indicated through EMA responses) using this information [104].

Sleep Disorders

Smartphone-based digital phenotyping research has also specifically looked at sleep disorders. However, it is somewhat limited. The main focus in phenotyping for sleep disorders has been on sleep apnea. Sleep apnea is a common condition in which a person's breathing stops and starts again throughout the night [105]. This is a potentially serious disorder that not only impacts sleep quality, but it may also lead to increased cardiovascular risk and lead to premature death if not treated correctly [105]. Smartphone-based approaches primarily monitor breathing and snoring patterns using the smartphone microphone. The ground truth for these studies was established by recruiting individuals with previously diagnosed sleep apnea or by conducting polysomnography as part of the study. Table 7 provides a summary of sleep disorder research reviewed in this section.

XSL•FO
RenderX

Table . Summary of smartphone-based digital phenotyping studies focused on sleep disorders, neurological disorders, ADHD[a], and general well-being (n=10). Studies in these domains used smartphone sensors to assess sleep quality, neurological motor symptoms, attention patterns, and general well-being. Common sensor modalities included microphones, accelerometers, and gyroscopes, with validation against polysomnogram, clinical scales, or self-reported measures shown to be strongly associated with polysomnography results for central apnea, hypopnea, and obstructive apnea [103].

Citation	Condition	Phenotypes considered	Sensors used	Ground truth	Participants (n)	Duration	Focus	Study setting
[106]	Sleep apnea	Sleep disruption, sleep patterns	Microphone	Polysomnogram taken before the study	55 samples	One 8-hour sleep period	Diagnosis	Hybrid setting (location not stated)
[107]	Sleep apnea	Sleep disruption, sleep patterns	Microphone	Polysomnogram score	55 sleep samples	One 8-hour sleep period	Diagnosis	Controlled laboratory setting (location not stated)
[108]	Parkinson disease	Tremors, FOG[b]	Accelerometer	TUG[c] test	20 adults with PD[d]	One in-laboratory assessment	Diagnosis	Controlled laboratory setting (location not stated)
[109]	Parkinson disease	FOG	Accelerometer, gyroscope	Heel-mounted foot switch, pressure sensor mat	12 adults with PD, 12 healthy controls	One in-laboratory assessment	Diagnosis	Controlled laboratory setting (Singapore)
[110]	Parkinson disease	Tremors	Accelerometer, gyroscope	UPDRS[e]	23 adults with PD	One in-laboratory assessment	Diagnosis	Controlled laboratory setting (Greece)
[111]	Parkinson disease	Tremors	Accelerometer	Diagnosis of PD or ET[f], reevaluation after 1 year	17 adults with PD, 16 adults with ET, 12 healthy controls	One in-laboratory assessment	Diagnosis	Controlled laboratory setting (location not stated)
[112]	ADHD	Loss of concentration	Text logs	Self-reported symptoms	2 adults with ADHD, 2 healthy controls	10 weeks	Monitoring	Naturalistic setting (United States)
[113]	General well-being	Movement patterns, physical activity level, sociability	Microphone, accelerometer, GPS, text logs, call logs	Self-reported mood	18,000 adults	3 years	Monitoring	Naturalistic setting (app store recruitment)
[114]	General well-being	Physical activity level, sleep disruption, sociability	Accelerometer, microphone, screen state, battery level	Self-reported physical and mental well-being	27 adults	19 days	Intervention	Naturalistic setting (app store recruitment)

[a]ADHD: attention-deficit/hyperactivity disorder.

[b]FOG: freezing of gait.

[c]TUG: Timed Up and Go.

[d]PD: Parkinson disease.

[e]UPDRS: Unified Parkinson Disease Rating Scale.

[f]ET: essential tremor.

XSL•FO
RenderX

Sleep Apnea

Traditionally, measuring the presence of sleep apnea requires an overnight stay at a sleep laboratory in order for a formal diagnosis to take place [115]. This stringent requirement may discourage individuals from actively pursuing treatment if they are unsure of the outcome of a sleep study or may perceive it to be too much of a hassle. As a result, smartphone-based digital phenotyping aims to remain as unobtrusive as possible. Phenotyping work for sleep apnea was primarily in the diagnosis focus category.

Concerning diagnosis, in 1 study, the researchers proposed a method that used accelerometer data and compared movements to a threshold to determine the extent of sleep movement. This was combined with snoring-related sound data from the smartphone microphone to determine the quality of sleep, which was divided into 3 levels of apnea: no apnea, mild apnea, and severe apnea [106]. In a similar study, researchers monitored the breathing patterns of a person using a smartphone app placed at the bedside. Their approach includes collecting information on breathing, the sound of snoring, and breathing pauses to determine the presence of sleep apnea and hypopnea, by parsing out white noise and various sounds that occur during the night [107].

The team that developed the means to determine the feasibility of using a smartphone microphone to monitor respiratory changes indicative of an impending opioid overdose also developed a similar approach that turned the microphone into a short-range active sonar to detect chest and abdomen movements. The changes in chest and abdomen movement detected from the reflected sound waves then allowed for apnea detection, which was then shown to be strongly associated with polysomnography results for central apnea, hypopnea, and obstructive apnea [103].

Neurological Disorders

A neurological disorder is any disorder of the nervous system. Examples of symptoms include paralysis, muscle weakness, poor coordination, loss of sensation, seizures, confusion, pain, and altered levels of consciousness [116]. From a smartphone-based digital phenotyping standpoint, work on neurological disorders has primarily focused on characterizing Parkinson disease. Table 7 provides a summary of the research reviewed in this section.

Parkinson disease

Parkinson disease is a movement disorder of the nervous system that progressively worsens over time. As neurons in the brain weaken and die, individuals with the disorder begin to endure tremors, impaired balance, stiffness in the limbs and trunk of the body, and difficulty with movement [117]. Like sleep apnea, phenotyping work for Parkinson disease was also primarily in the diagnosis focus category.

Concerning diagnosis, an aspect of Parkinson disease that has been explored in the smartphone-based phenotyping context is the detection of real-time freezing of gait (stopping suddenly while walking) [108]. In a similar study, 20 individuals with Parkinson disease were asked to perform a video-recorded Timed Up and Go test with and without dual-tasks while carrying a smartphone. The video and accelerometer recordings from the smartphone were synchronized in order to assess the reliability of the freezing of gait detection system in comparison to clinician ratings, with reasonable results [117]. In a similar study, the researchers looked at the accuracy of smartphone-based gait analysis using the built-in triaxial gyroscope and accelerometer and validated it against 2 heel contact-based measurements. These findings highlighted that smartphone-based gait analysis could serve as an alternative to conventional gait analysis methods (eg, foot-switch systems or sensor-embedded walkways), particularly when those methods are inconvenient (eg, cost-prohibitive) [109].

Smartphones have been evaluated as a way to measure hand tremor in Parkinson disease patients [110]. Tremors are a common symptom of Parkinson disease. In this study, accelerometers and gyroscopes were used to detect and measure hand tremors, which were then associated with clinical scores in the hand tremor components of the Unified Parkinson Disease Rating Scale. The results suggest a relatively strong correlation between the patients' Unified Parkinson Disease Rating Scale hand tremor scores and the quantitative measurement derived from smartphones [110]. In another similar study, the goal was to develop a diagnostic test to differentiate between Parkinson disease and essential tremor based on time-frequency differences using smartphone-based accelerometer data without the use of neuroimaging techniques [111]. In this case, the smartphone was placed over the back of the outstretched hand of a person with Parkinson disease while at rest. The results demonstrated that not only was the phone-based approach effective, but it was also able to diagnose the presence of Parkinson disease and essential tremor in several participants, each of whom had not been officially diagnosed [111].

Other Efforts

A small amount of work has also been done in characterizing ADHD and assessing the general mental health and well-being of individuals [112-114]. Table 7 provides a summary of the research covered in this section. Phenotyping work for ADHD was primarily focused on monitoring, whereas general well-being had work in both monitoring and diagnosis focus categories.

Concerning monitoring, with respect to ADHD, researchers collected text data and associated it with weekly self-report ADHD symptom scores. The ground truth for the ADHD symptoms was self-reported by the participants. Overall, the results of the study indicated that specific ADHD inattention, sluggish cognitive tempo, and hyperactivity symptoms could be predicted fairly accurately using the smartphone data collected from college students in a pilot study [112].

Further, smartphone-based phenotyping also focused on general well-being. These studies aimed to associate self-reported mental well-being with routine captured through a smartphone. This was a particularly large and longitudinal study. Overall, the study found that positive or negative affect for an individual could be predicted using passive sensing data (such as GPS, phone or call logs, accelerometer, and microphone data). These

XSL•FO
RenderX

results suggest that such approaches could reduce the need for self-report [113].

Concerning intervention, in the 1 intervention study, researchers deployed an app called BeWell [114]. BeWell monitored user behavior, such as sleep, physical activity, and social interaction, based on accelerometer, screen on or off, charging, and microphone data to give individuals a well-being score. The well-being scores range between 0 and 100 and are calculated for each of the 3 dimensions (sleep patterns, physical activity, and social interaction). A score of 100 indicates the person is matching or exceeding recommended guidelines. The idea is to use these scores to eventually provide feedback to individuals to maintain their well-being over time [114].

Sensor Modalities Used Across Different Health Domains

Finally, the reviewed studies showed wide variation in sensor use across health domains. While some investigations relied on a single data stream (eg, the work [69] that exclusively used GPS to track bipolar disorder or the research [91] that assessed the activity level of adults with schizophrenia using only the smartphone accelerometer), others combined multiple sources such as GPS, accelerometry, and microphone data to capture a richer behavioral context. As shown in Figure 2, GPS and accelerometer emerged as the 2 most common data streams. Accelerometers appeared as input 58% (38/65) of studies, making them the most frequently used modality overall. GPS followed close behind, appearing in 53% (35/65) studies. Communication logs were also common, with call or text logs appearing in roughly 44% (28/65) of studies. These streams formed the core of most smartphone-based digital phenotyping work due to their ability to intuitively capture mobility, activity, and daily routines. Other smartphone sensing modalities were used far less often. Microphone-based sensing appeared in 27% (18/65) of studies and was fairly evenly spread between health conditions, indicating the potential for widespread use across conditions. Screen state data also appeared in 27% (18/65) of studies, typically as a proxy for phone engagement and sleep-wake cycles. Gyroscope data were used relatively infrequently, appearing in just 10% (7/65) of studies. Environmental and proximity sensing (ambient light, WiFi, cell tower pings, app usage, and Bluetooth) appeared sparingly. At most, these streams appeared in 15% (10/65) of studies in the case of ambient light and just 6% (4/65) of studies in the case of cell tower pings. Battery percentage was used surprisingly infrequently, with only 7% (5/65) of studies using this stream. We speculate this may be due to internal differences in how Android (Google LLC) and iOS (Apple Inc) devices may collect data, but this is not made clear by the authors, who elect to use (or not use) the battery state as a data stream. Keystroke dynamics appeared the least out of all streams being used in just 3% (2/65) of studies. This is likely due to keystroke dynamics being used for specific, tailored tasks as opposed to sensing broad behavioral changes.

Figure 2. Heatmap illustrating the frequency with which different smartphone sensor modalities were used across 6 major health domains in the reviewed literature. Brighter and warmer-colored cells indicate more frequent usage of a given sensor type within that domain.

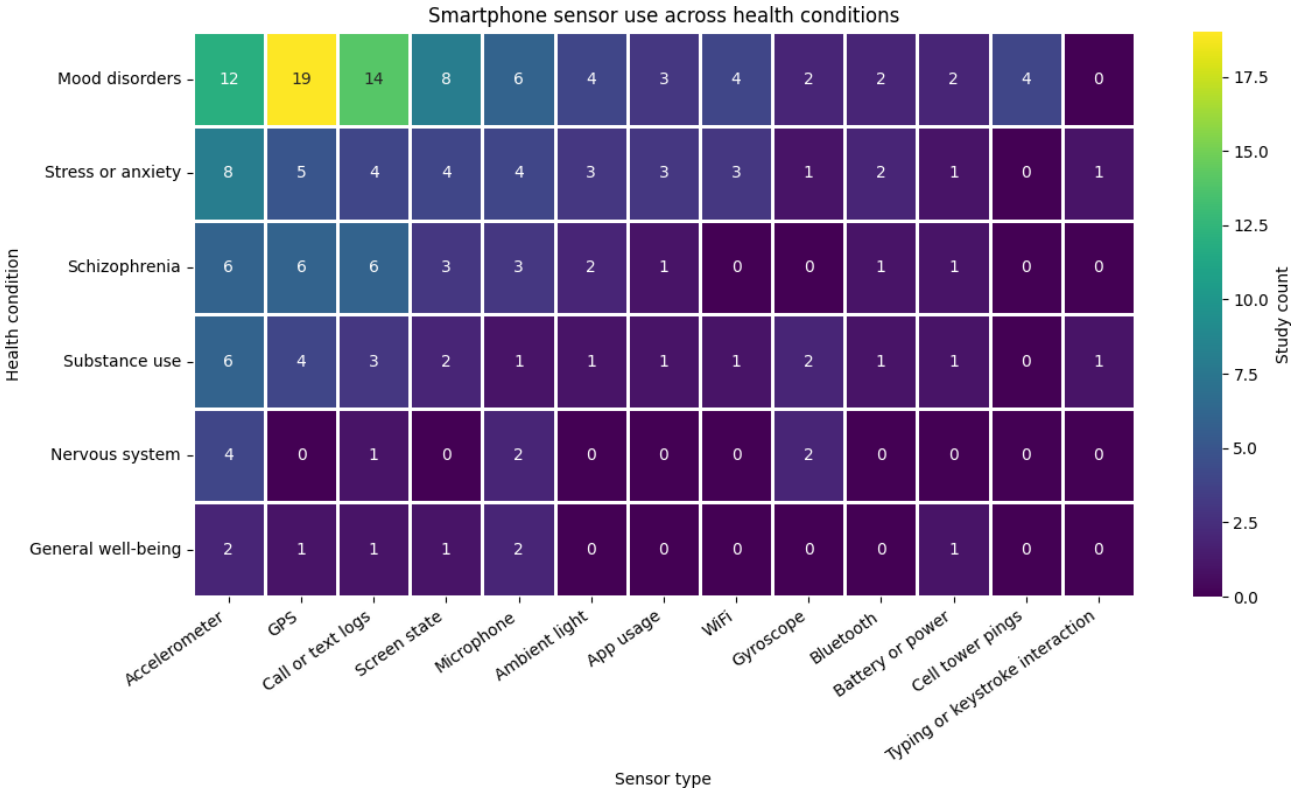

XSL•FO
RenderX

Discussion

Principal Findings

This scoping review synthesized research that used smartphone-based digital phenotyping to understand health-related conditions. Most studies in our review (n=45) used smartphone-based digital phenotyping to help with mental health conditions. Within the wider health scope, depression and bipolar disorder are of particular interest to researchers due to the ability to effectively track behavioral and mood fluctuations over time through the use of passive smartphone sensors. In contrast, relatively few studies explored other domains such as substance use (n=7), neurological disorders (n=4), or sleep disorders (n=7).

When classified by focus, about 63.5% (41/65) of the studies were conducted with the intent to monitor, which we have defined as tracking symptoms, behaviors, and environmental contexts over time. A smaller subset of these studies (n=15, 23.1%) looked at phenotypes for the purpose of diagnosis, which involves using smartphone-derived features to identify the presence or severity of health conditions. Finally, a few studies totaling 3% (2/65) implemented or evaluated intervention-based approaches in which smartphone-derived behaviors were used to inform treatment strategies or real-time interventions.

Collectively, these findings demonstrate that while smartphone-based phenotyping has some potential across monitoring, diagnostic, and intervention-focused studies for a variety of health conditions, there are several aspects of this research domain that could be improved, which we discuss below.

Methodological Inconsistencies and Gaps

We found that 15.3% (10/65) of included research studies lacked a clear description of which sensor streams were being used in their phenotyping work and why. This lack of detail about the streams takes several forms: (1) instances where the sensor streams are mentioned as being in scope but are not mentioned in the collection process described [38,39,52,68,112]; (2) instances where a variety of sensor streams are measured from participants but only a subset are used in the phenotyping work [80,92]; and (3) instances where the type of sensor stream collected to capture a phenotype is not even specified [70]. These limitations hinder continued iteration on the described work and suggest the need for guidelines for how to report data surrounding smartphone-based digital phenotyping work to enhance reproducibility, learn from the successes and failures of groups conducting this work, and ultimately advance these tools toward broader impacts. Each of these instances prevents us from fully understanding the scope of the work being done and hinders future research from properly expanding upon the findings of these works. We recommend that, at a minimum, research groups should describe the types of smartphones used in the study, their operating system, the specific sensor streams, and the strategies in which the data was obtained. This methodology could benefit others who seek to leverage smartphone-based digital phenotyping to detect and understand disease trajectories.

These reporting issues extend beyond sensor specifications. Despite all studies using smartphones as the sole sensing device, few provided sufficient technical or procedural detail to allow for replication. Only 9.2% (6/65) of the studies quantitatively reported missingness at the stream level [38,39,52,59,68,92]. Similarly, only 6.1% (4/65) of the studies explicitly reported sampling rates or duty cycles [52,82,83,85], with all other studies leaving the frequency of data capture and transmission unclear. The most common detail reported was phone operating system with 33.8% (22/65) of the studies stating the phone operating system used in data collection, indicating whether the study used Android, iOS, or mixed devices [38,39,48,51-53,59,64,68,70,71,78,80,82,83,85-87,89,90,112,113]. Finally, only 10.7% (7/65) of the studies discussed participant compliance or attrition, offering information about adherence, dropout, or data completeness at the participant level [38,39,52,68,71,82,85].

ML models are commonly used in smartphone-based digital phenotyping work to determine the presence of specific health conditions when monitoring or diagnosing them. Of the 65 studies we assessed as part of this work, about 90.7% (n=59) used ML models, while about 9.3% (n=6) did not [54,72,84,106,109,111]. For our discussion on the use of ML models, we will be noting statistics as being out of 59 so as not to unfairly include studies that do not use ML. About 69.4% (41/59) of the studies that used ML models in their smartphone-based phenotyping work delved deeper into how these models were learning and what derived features may mean in the context of a given health outcome. However, we found that a substantial number, about 30.5% (18/59) of the studies, did not engage with ML models beyond the perfunctory [38,45,53,56,66,69,77,80,81,86,100,101,103,107,108,110,112,118]. These studies reported how accurate these models are with respect to detecting the presence of a health condition in a dataset. While this kind of analysis is important to understanding if models are working, they often eschew any other kind of analysis on why the model is making its decision. We found that these emphases on how accurately an ML model performed did not provide any insight into the validity of the models themselves or an understanding of a person's behavior being used as phenotypes for characterizing a specific health condition. We believe these points point to a tendency to discover without understanding in a substantial portion of the studies that used ML approaches, which is a cause for numerous errors and bad science [119].

We believe that when ML models are discussed, an effort should be made in exploring the various features used to determine the presence of specific health conditions in addition to reporting on the global accuracy of detection. Reporting on the features allows us, at the very least, to sanity check if the model is making decisions based on known behavioral patterns that individuals with specific health conditions exhibit. However, more importantly, such an analysis has the potential to provide insights into newer ways of identifying these health conditions. This would advance the science of detecting specific health conditions and thus truly leverage the pattern-matching capabilities of ML models.

XSL•FO
RenderX

Finally, we found that despite the amount of work in smartphone-based digital phenotyping, there is a lack of publicly available datasets in this space. The primary issue with not having any data available is that researchers are required to spend a significant amount of time recruiting participants, collecting data, and cleaning data before they perform any kind of analysis. If data were to be made available, innovation and further research would be significantly easier. Only 10.7% (7/65) of studies discuss data availability. Of the studies that provide access to their data, 2 use the Dartmouth StudentLife dataset, which is openly available [82,83]. About 4.6% (3/65) of the studies indicate that the data are available upon request to the authors [49,72,87]. One study indicates that the information is individually identifiable and it must be requested through the institutional ethics board [88]. Only 1.5% (1/65) of the studies provide a working link to the GPS data collected during the study [50]. This suggests that researchers should consider strategies to provide data for other teams to advance smartphone-based digital phenotyping as an innovative strategy to address multiple health conditions. In this process, researchers should consult with their respective institutional ethics boards and ethicists to understand the implications of data sharing, especially for groups whose phenotyping data can be perceived as being more sensitive (eg, specific geolocation and text message data). In some cases, it may be necessary to restrict data access for certain streams of phenotyping data to protect the confidentiality of participants. At the same time, considerations surrounding consent during enrollment for future data sharing to advance the field should be encouraged among research teams. We encourage authors to disclose how to access the data. In the case that data is not readily accessible and must be requested through specific channels, we feel that authors should state this. In case the data is not available under any circumstances, we feel that authors should disclose this as well.

In this scoping review, we provide a comprehensive synthesis of smartphone-based digital phenotyping research across a broader range of health conditions than has been examined in prior reviews. By focusing exclusively on data collected from smartphone sensors, this work differs from existing reviews that either incorporate auxiliary sensing modalities or limit their scope to a small number of predominantly mental health conditions. From a real-world perspective, these results underscore the potential of smartphones as scalable, unobtrusive tools for monitoring, diagnosis, and intervention across mental health, physical health, and SUDs while also highlighting the need for improved reporting practices and validation to support responsible deployment.

Limitations

This scoping review has a few limitations. First, we excluded non-English and non–peer-reviewed literature, which may have resulted in missing relevant work conducted in other regions or early-stage research not yet published in academic journals. Second, we only searched Google Scholar, IEEE Xplore, ACM Digital Library, and PubMed databases for identifying the studies for this scoping review. Again, searching other databases could have revealed more studies in this space, which we did not include in this review. Third, as a scoping review, we aimed to map available literature to our 3 given focus categories rather than assessing the quality or strength of each study included. As an additional aspect to this point, studies included in this scoping review vary widely in multiple aspects, including the length of the study, the number of participants, and the study setting, which makes direct comparison between studies challenging. Finally, most studies in this review were observational and conducted in relatively small samples, often within specific demographic groups (eg, students). Due to this, the behavioral associations identified should be interpreted as correlational rather than causal, and their applicability to broader or more diverse populations remains an open question.

Despite these limitations, this scoping review provides a comprehensive overview of smartphone-based digital phenotyping across health domains and highlights important gaps that future work must address, particularly around reporting standards, validation practices, and data sharing.

Conclusion

In this scoping review, we examined the work done in the domain of smartphone-based digital phenotyping in determining the presence of various health conditions. Our scope was much more expansive than prior scoping reviews in this space and includes any and all health conditions that have been explored using smartphone-based phenotyping work. Further, and most importantly, in this work, we focus on describing the relationships between behaviors exhibited in the context of specific health conditions and not on algorithms used to extract these behaviors. Overall, based on a scoping review of 65 studies, we found that smartphone-based digital phenotyping is a powerful and informative tool that can give us insights into a variety of health conditions. Though most of the work has focused on applying smartphone-based digital phenotyping to mental health issues, such as depression, bipolar disorder, and stress, some effort has also gone into phenotyping sleep and neurological conditions.

Smartphone-based digital phenotyping offers substantial promise for understanding health-related behaviors in everyday life. Despite this promise, continued progress depends on addressing key issues surrounding data quality, reproducibility, and reporting standards. Further, as smartphone-based phenotyping scales to larger and more heterogeneous populations, privacy and consent challenges will become increasingly important. Ensuring ethical collection, sharing, and use of sensitive behavioral data will be essential to maintaining participant trust and supporting the responsible growth of this research domain. Smartphones are no longer simple communication tools, but they are emerging as primary instruments for continuous health monitoring.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank the reviewers for providing excellent comments that were very helpful in improving the quality of this scoping review. Responsibility for the final manuscript lies entirely with the authors. Generative artificial intelligence tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by AD. No artificial intelligence tool was used for this paper.

Funding

This work was supported by the National Institutes of Health grant DP2DA056107. The funders had no role in the decision to publish or in the preparation of this paper.

Authors' Contributions

Conceptualization: AD, KV, PC
Data curation: AD, JH
Formal analysis: AD, JH
Funding acquisition: PC
Investigation: AD, KV
Methodology: AD, KV, PC
Project administration: GG, PC
Resources: PC
Supervision: KV, PC
Writing – original draft: AD, KV, PC
Writing – review and editing: AD, KV

Conflicts of Interest

None declared.

Multimedia Appendix 1
Presents the full electronic search strategies used to identify relevant records.
[DOCX File, 1672 KB - jmir_v28i1e84146_app1.docx]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 74 KB - jmir_v28i1e84146_app2.docx]

References

1. Mobile technology and home broadband 2021. Pew Research Center. 2021 Jun 3. URL: https://www.pewresearch.org/internet/2021/06/03/mobile-technology-and-home-broadband-2021 [accessed 2026-02-27]
2. Ventola CL. Mobile devices and apps for health care professionals: uses and benefits. P T 2014 May;39(5):356-364. [Medline: 24883008]
3. Strawley C, Richwine C. Individuals' access and use of patient portals and smartphone health apps, 2022. In: ASTP Health IT Data Brief: Office of the Assistant Secretary for Technology Policy; 2023, Vol. 2012. 69. [Medline: 39413226]
4. Zulman DM, Verghese A. Virtual care, telemedicine visits, and real connection in the era of COVID-19: unforeseen opportunity in the face of adversity. JAMA 2021 Feb 2;325(5):437-438. [doi: 10.1001/jama.2020.27304] [Medline: 33528520]
5. Trends in individuals' access, viewing, and use of online medical records and other technology for health needs: 2022. Assistant Secretary for Technology Policy. 2019 May. URL: https://www.healthit.gov/data/data-briefs/trends-individuals-access-viewing-and-use-online-medical-records-and-other [accessed 2026-02-27]
6. Wall C, Hetherington V, Godfrey A. Beyond the clinic: the rise of wearables and smartphones in decentralising healthcare. NPJ Digit Med 2023 Nov 25;6(1):219. [doi: 10.1038/s41746-023-00971-z] [Medline: 38007554]
7. Dorsey ER, Chan YFY, McConnell MV, et al. The use of smartphones for health research. Acad Med 2017 Feb;92(2):157-160. [doi: 10.1097/ACM.0000000000001205] [Medline: 27119325]
8. Ernsting C, Dombrowski SU, Oedekoven M, et al. Using smartphones and health apps to change and manage health behaviors: a population-based survey. J Med Internet Res 2017 Apr 5;19(4):e101. [doi: 10.2196/jmir.6838] [Medline: 28381394]
9. Onnela JP. Opportunities and challenges in the collection and analysis of digital phenotyping data. Neuropsychopharmacology 2021 Jan;46(1):45-54. [doi: 10.1038/s41386-020-0771-3] [Medline: 32679583]

XSL•FO
RenderX

10. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annu Rev Clin Psychol 2017 May 8;13:23-47. [doi: 10.1146/annurev-clinpsy-032816-044949] [Medline: 28375728]
11. Harari GM, Lane ND, Wang R, Crosier BS, Campbell AT, Gosling SD. Using smartphones to collect behavioral data in psychological science: opportunities, practical considerations, and challenges. Perspect Psychol Sci 2016 Nov;11(6):838-854. [doi: 10.1177/1745691616650285] [Medline: 27899727]
12. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA 2017 Oct 3;318(13):1215-1216. [doi: 10.1001/jama.2017.11295] [Medline: 28973224]
13. Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology 2016 Jun;41(7):1691-1696. [doi: 10.1038/npp.2016.7] [Medline: 26818126]
14. Lee K, Lee TC, Yefimova M, et al. Using digital phenotyping to understand health-related outcomes: a scoping review. Int J Med Inform 2023 Jun;174:105061. [doi: 10.1016/j.ijmedinf.2023.105061] [Medline: 37030145]
15. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
16. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
17. Cornet VP, Holden RJ. Systematic review of smartphone-based passive sensing for health and wellbeing. J Biomed Inform 2018 Jan;77:120-132. [doi: 10.1016/j.jbi.2017.12.008] [Medline: 29248628]
18. Fraccaro P, Beukenhorst A, Sperrin M, et al. Digital biomarkers from geolocation data in bipolar disorder and schizophrenia: a systematic review. J Am Med Inform Assoc 2019 Nov 1;26(11):1412-1420. [doi: 10.1093/jamia/ocz043] [Medline: 31260049]
19. Jayakumar P, Lin E, Galea V, et al. Digital phenotyping and patient-generated health data for outcome measurement in surgical care: a scoping review. JPM 2020;10(4):282. [doi: 10.3390/jpm10040282]
20. Saccaro LF, Amatori G, Cappelli A, Mazziotti R, Dell'Osso L, Rutigliano G. Portable technologies for digital phenotyping of bipolar disorder: a systematic review. J Affect Disord 2021 Dec 1;295:323-338. [doi: 10.1016/j.jad.2021.08.052] [Medline: 34488086]
21. Hilty DM, Armstrong CM, Luxton DD, Gentry MT, Krupinski EA. A scoping review of sensors, wearables, and remote monitoring for behavioral health: uses, outcomes, clinical competencies, and research directions. J Technol Behav Sci 2021;6(2):278-313. [doi: 10.1007/s41347-021-00199-2]
22. De Angel V, Lewis S, White K, et al. Digital health tools for the passive monitoring of depression: a systematic review of methods. NPJ Digit Med 2022;5(1):3. [doi: 10.1038/s41746-021-00548-8]
23. Dlima SD, Shevade S, Menezes SR, Ganju A. Digital phenotyping in health using machine learning approaches: scoping review. JMIR Bioinform Biotech 2022;3(1):e39618. [doi: 10.2196/39618]
24. Mendes JPM, Moura IR, Van de Ven P, et al. Sensing apps and public data sets for digital phenotyping of mental health: systematic review. J Med Internet Res 2022 Feb 17;24(2):e28735. [doi: 10.2196/28735] [Medline: 35175202]
25. Alfalahi H, Dias SB, Khandoker AH, Chaudhuri KR, Hadjileontiadis LJ. A scoping review of neurodegenerative manifestations in explainable digital phenotyping. NPJ Parkinsons Dis 2023 Mar 30;9(1):49. [doi: 10.1038/s41531-023-00494-0] [Medline: 36997573]
26. Dobson R, Li LL, Garner K, Tane T, McCool J, Whittaker R. The use of sensors to detect anxiety for in-the-moment intervention: scoping review. JMIR Ment Health 2023 Feb 2;10(1):e42611. [doi: 10.2196/42611] [Medline: 36729590]
27. Bufano P, Laurino M, Said S, Tognetti A, Menicucci D. Digital phenotyping for monitoring mental disorders: systematic review. J Med Internet Res 2023;25:e46778. [doi: 10.2196/46778]
28. Zhang H, Ibrahim A, Parsia B, Poliakoff E, Harper S. Passive social sensing with smartphones: a systematic review. Computing 2023 Jan;105(1):29-51. [doi: 10.1007/s00607-022-01112-2]
29. Moura I, Teles A, Viana D, Marques J, Coutinho L, Silva F. Digital phenotyping of mental health using multimodal sensing of multiple situations of interest: a systematic literature review. J Biomed Inform 2023 Feb;138:104278. [doi: 10.1016/j.jbi.2022.104278] [Medline: 36586498]
30. Beames J, Han J, Shvetcov A, et al. Use of smartphone sensor data in detecting and predicting depression and anxiety in young people (12-25 years): a scoping review. SSRN J 2024;11:e47206. [doi: 10.2139/ssrn.4798261]
31. Leaning IE, Ikani N, Savage HS, et al. From smartphone data to clinically relevant predictions: a systematic review of digital phenotyping methods in depression. Neurosci Biobehav Rev 2024 Mar;158:105541. [doi: 10.1016/j.neubiorev.2024.105541] [Medline: 38215802]
32. Choi A, Ooi A, Lottridge D. Digital phenotyping for stress, anxiety, and mild depression: systematic literature review. JMIR mHealth uHealth 2024;12(1):e40689. [doi: 10.2196/40689]
33. Linardon J, Chen K, Gajjar S, et al. Smartphone digital phenotyping in mental health disorders: a review of raw sensors utilized, machine learning processing pipelines, and derived behavioral features. Psychiatry Res 2025 Jun;348:116483. [doi: 10.1016/j.psychres.2025.116483] [Medline: 40187059]
34. Kallio J, Vildjiounaite E, Tervonen J, Bordallo López M. A survey on sensor-based techniques for continuous stress monitoring in knowledge work environments. ACM Trans Comput Healthcare 2025 Jul 31;6(3):1-31. [doi: 10.1145/3720551]

XSL•FO
RenderX

35. Booth BM, Mundnich K, Feng T, et al. Multimodal human and environmental sensing for longitudinal behavioral studies in naturalistic settings: framework for sensor selection, deployment, and management. J Med Internet Res 2019 Aug 20;21(8):e12832. [doi: 10.2196/12832] [Medline: 31432781]
36. Sekhon S, Gupta V. Mood Disorder: StatPearls Publishing; 2025. URL: https://www.ncbi.nlm.nih.gov/books/NBK558911 [accessed 2026-02-27]
37. Depressive disorder (depression). World Health Organization. 2025 Aug 29. URL: https://www.who.int/news-room/fact-sheets/detail/depression [accessed 2026-02-27]
38. Choudhary S, Thomas N, Ellenberger J, Srinivasan G, Cohen R. A machine learning approach for detecting digital behavioral patterns of depression using nonintrusive smartphone data (complementary path to Patient Health Questionnaire-9 assessment): prospective observational study. JMIR Form Res 2022 May 16;6(5):e37736. [doi: 10.2196/37736] [Medline: 35420993]
39. Berrouiguet S, Ramírez D, Barrigón ML, et al. Combining continuous smartphone native sensors data capture and unsupervised data mining techniques for behavioral changes detection: a case series of the evidence-based behavior (eB2) study. JMIR mHealth uHealth 2018 Dec 10;6(12):e197. [doi: 10.2196/mhealth.9472] [Medline: 30530465]
40. Jacobson NC, Chung YJ. Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones. Sensors (Basel) 2020 Jun 24;20(12):3572. [doi: 10.3390/s20123572] [Medline: 32599801]
41. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
42. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. Appl Psychol Meas 1977;1(3):385-401. [doi: 10.1177/014662167700100306]
43. Poznanski EO, Grossman JA, Buchsbaum Y, Banegas M, Freeman L, Gibbons R. Children's Depression Rating Scale–Revised. APA PsycTests 1984. [doi: 10.1037/t55280-000]
44. Busner J, Targum SD. The clinical global impressions scale: applying a research tool in clinical practice. Psychiatry (Edgmont) 2007 Jul;4(7):28-37. [Medline: 20526405]
45. Wang Y, Ren X, Liu X, Zhu T. Examining the correlation between depression and social behavior on smartphones through usage metadata: empirical study. JMIR mHealth uHealth 2021 Jan 6;9(1):e19046. [doi: 10.2196/19046] [Medline: 33404512]
46. Saeb S, Zhang M, Karr CJ, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J Med Internet Res 2015 Jul 15;17(7):e175. [doi: 10.2196/jmir.4273] [Medline: 26180009]
47. Sarda A, Munuswamy S, Sarda S, Subramanian V. Using passive smartphone sensing for improved risk stratification of patients with depression and diabetes: cross-sectional observational study. JMIR mHealth uHealth 2019 Jan 29;7(1):e11041. [doi: 10.2196/11041] [Medline: 30694197]
48. Cao J, Truong AL, Banu S, Shah AA, Sabharwal A, Moukaddam N. Tracking and predicting depressive symptoms of adolescents using smartphone-based self-reports, parental evaluations, and passive phone sensor data: development and usability study. JMIR Ment Health 2020;7(1):e14045. [doi: 10.2196/14045]
49. Kim JS, Wang B, Kim M, et al. Prediction of diagnosis and treatment response in adolescents with depression by using a smartphone app and deep learning approaches: usability study. JMIR Form Res 2023;7(1):e45991. [doi: 10.2196/45991]
50. Saeb S, Lattie EG, Schueller SM, Kording KP, Mohr DC. The relationship between mobile phone location sensor data and depressive symptom severity. PeerJ 2016;4:e2537. [doi: 10.7717/peerj.2537] [Medline: 28344895]
51. Canzian L, Musolesi M. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Association for Computing Machinery; 2015:1293-1304. [doi: 10.1145/2750858.2805845]
52. Wahle F, Kowatsch T, Fleisch E, Rufer M, Weidt S. Mobile sensing and support for people with depression: a pilot trial in the wild. JMIR mHealth uHealth 2016 Sep 21;4(3):e111. [doi: 10.2196/mhealth.5960] [Medline: 27655245]
53. Masud MT, Mamun MA, Thapa K, Lee DH, Griffiths MD, Yang SH. Unobtrusive monitoring of behavior and movement patterns to detect clinical depression severity level via smartphone. J Biomed Inform 2020 Mar;103:103371. [doi: 10.1016/j.jbi.2019.103371] [Medline: 31935462]
54. Melcher J, Lavoie J, Hays R, et al. Digital phenotyping of student mental health during COVID-19: an observational study of 100 college students. J Am Coll Health 2023 Apr;71(3):736-748. [doi: 10.1080/07448481.2021.1905650] [Medline: 33769927]
55. Tlachac M, Flores R, Reisch M, Kayastha R, Taurich N, Melican V, et al. StudentSADD: rapid mobile depression and suicidal ideation screening of college students during the coronavirus pandemic. Proc ACM Interact Mob Wearable Ubiquitous Technol 2022;6(2):1-32. [doi: 10.1145/3534604]
56. Higuchi M, Nakamura M, Shinohara S, et al. Effectiveness of a voice-based mental health evaluation system for mobile devices: prospective study. JMIR Form Res 2020 Jul 20;4(7):e16455. [doi: 10.2196/16455] [Medline: 32554367]
57. Tlachac M, Flores R, Toto E, Rundensteiner E. Early mental health uncovering with short scripted and unscripted voice recordings. In: Deep Learning Applications, Volume 4: Springer; 2022:79-110. [doi: 10.1007/978-981-19-6153-3_4]
58. Mercier HW, Hamner JW, Torous J, Onnela JP, Taylor JA. Digital phenotyping to quantify psychosocial well-being trajectories after spinal cord injury. Am J Phys Med Rehabil 2020;99(12):1138-1144. [doi: 10.1097/PHM.0000000000001506]

XSL•FO
RenderX

59. MacLeod L, Suruliraj B, Gall D, et al. A mobile sensing app to monitor youth mental health: observational pilot study. JMIR mHealth uHealth 2021;9(10):e20638. [doi: 10.2196/20638]
60. Bipolar disorder. National Institute of Mental Health. URL: https://www.nimh.nih.gov/health/topics/bipolar-disorder [accessed 2026-02-27]
61. Hamilton M. Development of a rating scale for primary depressive illness. Br J Soc Clin Psychol 1967 Dec;6(4):278-296. [doi: 10.1111/j.2044-8260.1967.tb00530.x] [Medline: 6080235]
62. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. Br J Psychiatry 1978 Nov;133(5):429-435. [doi: 10.1192/bjp.133.5.429] [Medline: 728692]
63. Beiwinkel T, Kindermann S, Maier A, et al. Using smartphones to monitor bipolar disorder symptoms: a pilot study. JMIR Ment Health 2016;3(1):e2. [doi: 10.2196/mental.4560]
64. Grunerbl A, Muaremi A, Osmani V, et al. Smartphone-based recognition of states and state changes in bipolar disorder patients. IEEE J Biomed Health Inform 2014;19(1):140-148. [doi: 10.1109/JBHI.2014.2343154]
65. Faurholt-Jepsen M, Vinberg M, Frost M, et al. Behavioral activities collected through smartphones and the association with illness activity in bipolar disorder. Int J Methods Psychiatr Res 2016 Dec;25(4):309-323. [doi: 10.1002/mpr.1502]
66. Ryan KA, Babu P, Easter R, et al. A smartphone app to monitor mood symptoms in bipolar disorder: development and usability study. JMIR Ment Health 2020 Sep 22;7(9):e19476. [doi: 10.2196/19476] [Medline: 32960185]
67. Grünerbl A, Oleksy P, Bahle G, Haring C, Weppner J, Lukowicz P. Towards smart phone based monitoring of bipolar disorder. In: Proceedings of the Second ACM Workshop on Mobile Systems, Applications, and Services for HealthCare: Association for Computing Machinery; 2012:1-6. [doi: 10.1145/2396276.2396280]
68. Palmius N, Tsanas A, Saunders KEA, et al. Detecting bipolar depression from geographic location data. IEEE Trans Biomed Eng 2016;64(8):1761-1771. [doi: 10.1109/TBME.2016.2611862]
69. Tseng YC, Lin ECL, Wu CH, Huang HL, Chen PS. Associations among smartphone app-based measurements of mood, sleep and activity in bipolar disorder. Psychiatry Res 2022 Apr;310:114425. [doi: 10.1016/j.psychres.2022.114425] [Medline: 35152069]
70. Abdullah S, Matthews M, Frank E, Doherty G, Gay G, Choudhury T. Automatic detection of social rhythms in bipolar disorder. J Am Med Inform Assoc 2016 May;23(3):538-543. [doi: 10.1093/jamia/ocv200] [Medline: 26977102]
71. Faurholt-Jepsen M, Busk J, Frost M, et al. Voice analysis as an objective state marker in bipolar disorder. Transl Psychiatry 2016 Jul 19;6(7):e856. [doi: 10.1038/tp.2016.123] [Medline: 27434490]
72. Fellendorf FT, Hamm C, Dalkner N, et al. Monitoring sleep changes via a smartphone app in bipolar disorder: practical issues and validation of a potential diagnostic tool. Front Psychiatry 2021;12:641241. [doi: 10.3389/fpsyt.2021.641241] [Medline: 33841209]
73. Alvarez-Lozano J, Osmani V, Mayora O, Frost M, Bardram J, Faurholt-Jepsen M, et al. Tell me your apps and I will tell you your mood: correlation of apps usage with bipolar disorder state. In: Proceedings of the 7th International Conference on Pervasive Technologies Related to Assistive Environments: Association for Computing Machinery; 2014:1-7. [doi: 10.1145/2674396.2674408]
74. Cohen S, Janicki-Deverts D, Miller GE. Psychological stress and disease. JAMA 2007 Oct 10;298(14):1685-1687. [doi: 10.1001/jama.298.14.1685] [Medline: 17925521]
75. Jacobson NC, Summers B, Wilhelm S. Digital biomarkers of social anxiety severity: digital phenotyping using passive smartphone sensors. J Med Internet Res 2020;22(5):e16875. [doi: 10.2196/16875]
76. Gao Y, Li A, Zhu T, Liu X, Liu X. How smartphone usage correlates with social anxiety and loneliness. PeerJ 2016;4:e2197. [doi: 10.7717/peerj.2197]
77. Fukazawa Y, Ito T, Okimura T, Yamashita Y, Maeda T, Ota J. Predicting anxiety state using smartphone-based passive sensing. J Biomed Inform 2019 May;93:103151. [doi: 10.1016/j.jbi.2019.103151]
78. Maxhuni A, Hernandez-Leal P, Morales EF, Sucar LE, Osmani V, Mayora O. Unobtrusive stress assessment using smartphones. IEEE Trans on Mobile Comput 2020;20(6):2313-2325. [doi: 10.1109/TMC.2020.2974834]
79. Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health. Psychiatr Rehabil J 2015;38(3):218-226. [doi: 10.1037/prj0000130]
80. Garcia-Ceja E, Osmani V, Mayora O. Automatic stress detection in working environments from smartphones' accelerometer data: a first step. IEEE J Biomed Health Inform 2016 Jul;20(4):1053-1060. [doi: 10.1109/JBHI.2015.2446195] [Medline: 26087509]
81. Bogomolov A, Lepri B, Ferron M, Pianesi F, Pentland A. Daily stress recognition from mobile phone data, weather conditions and individual traits. In: MM '14: Association for Computing Machinery; 2014:477-486. [doi: 10.1145/2647868.2654933]
82. Wang R, Chen F, Chen Z, Li T, Harari G, Tignor S, et al. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Association for Computing Machinery; 2014:3-14. [doi: 10.1145/2632048.2632054]
83. Wang R, Harari G, Hao P, Zhou X, Campbell AT. SmartGPA: how smartphones can assess and predict academic performance of college students. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Association for Computing Machinery; 2015:295-306. [doi: 10.1145/2750858.2804251]

XSL•FO
RenderX

84. Exposito M, Hernandez J, Picard RW. Affective keys: towards unobtrusive stress sensing of smartphone users. In: Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct: Association for Computing Machinery; 2018:139-145. [doi: 10.1145/3236112.3236132]
85. Buck B, Scherer E, Brian R, et al. Relationships between smartphone social behavior and relapse in schizophrenia: a preliminary report. Schizophr Res 2019 Jun;208:167-172. [doi: 10.1016/j.schres.2019.03.014] [Medline: 30940400]
86. Difrancesco S, Fraccaro P, van der Veer SN, et al. Out-of-home activity recognition from GPS data in schizophrenic patients. In: 2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS): IEEE; 2016:324-328. [doi: 10.1109/CBMS.2016.54]
87. Depp CA, Bashem J, Moore RC, et al. GPS mobility as a digital biomarker of negative symptoms in schizophrenia: a case control study. NPJ Digit Med 2019;2(1):108. [doi: 10.1038/s41746-019-0182-1] [Medline: 31728415]
88. Henson P, Barnett I, Keshavan M, Torous J. Towards clinically actionable digital phenotyping targets in schizophrenia. NPJ Schizophr 2020 May 5;6(1):13. [doi: 10.1038/s41537-020-0100-1] [Medline: 32372059]
89. Barnett I, Torous J, Staples P, Sandoval L, Keshavan M, Onnela JP. Relapse prediction in schizophrenia through digital phenotyping: a pilot study. Neuropsychopharmacology 2018 Jul;43(8):1660-1666. [doi: 10.1038/s41386-018-0030-z] [Medline: 29511333]
90. Wang R, Aung MS, Abdullah S, Brian R, Campbell AT, Choudhury T, et al. CrossCheck: toward passive sensing and detection of mental health changes in people with schizophrenia. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Association for Computing Machinery; 2016:886-897. [doi: 10.1145/2971648.2971740]
91. Strauss GP, Raugh IM, Zhang L, et al. Validation of accelerometry as a digital phenotyping measure of negative symptoms in schizophrenia. Schizophrenia 2022;8(1):37. [doi: 10.1038/s41537-022-00241-z]
92. Staples P, Torous J, Barnett I, et al. A comparison of passive and active estimates of sleep in a cohort with schizophrenia. NPJ Schizophr 2017 Oct 16;3(1):37. [doi: 10.1038/s41537-017-0038-0] [Medline: 29038553]
93. Anxiety disorder. National Institute of Mental Health. URL: https://www.nimh.nih.gov/health/topics/anxiety-disorders [accessed 2026-02-27]
94. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
95. Joseph SM. Delusional Disorder: StatPearls Publishing; 2025. URL: https://www.ncbi.nlm.nih.gov/books/NBK539855 [accessed 2026-02-27]
96. Correll CU, Schooler NR. Negative symptoms in schizophrenia: a review and clinical guide for recognition, assessment, and treatment. Neuropsychiatr Dis Treat 2020 Feb 21;16:519-534. [doi: 10.2147/NDT.S225643] [Medline: 32110026]
97. Drug addiction (substance use disorder). Mayo Clinic. URL: https://www.mayoclinic.org/diseases-conditions/drug-addiction/symptoms-causes/syc-20365112 [accessed 2026-03-16]
98. Bae S, Ferreira D, Suffoletto B, et al. Detecting drinking episodes in young adults using smartphone-based sensors. Proc ACM Interact Mob Wearable Ubiquitous Technol 2017 Jun;1(2):1-36. [doi: 10.1145/3090051] [Medline: 35146236]
99. Wu T, Sherman G, Giorgi S, et al. Smartphone sensor data estimate alcohol craving in a cohort of patients with alcohol-associated liver disease and alcohol use disorder. Hepatol Commun 2023 Dec 1;7(12):e0329. [doi: 10.1097/HC9.0000000000000329] [Medline: 38055637]
100. Arnold Z, Larose D, Agu E. Smartphone inference of alcohol consumption levels from gait. In: 2015 International Conference on Healthcare Informatics (ICHI): IEEE; 2015:417-426. [doi: 10.1109/ICHI.2015.59]
101. Li R, Balakrishnan GP, Nie J, et al. Estimation of blood alcohol concentration from smartphone gait data using neural networks. IEEE Access 2021;9:61237-61255. [doi: 10.1109/ACCESS.2021.3054515]
102. Bae SW, Chung T, Islam R, et al. Mobile phone sensor-based detection of subjective cannabis intoxication in young adults: a feasibility study in real-world settings. Drug Alcohol Depend 2021 Nov;228:108972. [doi: 10.1016/j.drugalcdep.2021.108972]
103. Nandakumar R, Gollakota S, Sunshine JE. Opioid overdose detection using smartphones. Sci Transl Med 2019 Jan 9;11(474):eaau8914. [doi: 10.1126/scitranslmed.aau8914] [Medline: 30626717]
104. Epstein DH, Tyburski M, Kowalczyk WJ, et al. Prediction of stress and drug craving ninety minutes in the future with passively collected GPS data. NPJ Digit Med 2020;3(1):26. [doi: 10.1038/s41746-020-0234-6]
105. Sleep apnea. Mayo Clinic. URL: https://www.mayoclinic.org/diseases-conditions/sleep-apnea/symptoms-causes/syc-20377631 [accessed 2026-02-28]
106. Alqassim S, Ganesh M, Khoja S, Zaidi M, Aloul F, Sagahyroon A. Sleep apnea monitoring using mobile phones. In: 2012 IEEE 14th International Conference on E-Health Networking, Applications and Services (Healthcom): IEEE; 2012:443-446. [doi: 10.1109/HealthCom.2012.6379457]
107. Tiron R, Lyon G, Kilroy H, et al. Screening for obstructive sleep apnea with novel hybrid acoustic smartphone app technology. J Thorac Dis 2020 Aug;12(8):4476-4495. [doi: 10.21037/jtd-20-804] [Medline: 32944361]
108. Capecci M, Pepa L, Verdini F, Ceravolo MG. A smartphone-based architecture to detect and quantify freezing of gait in Parkinson's disease. Gait Posture 2016 Oct;50:28-33. [doi: 10.1016/j.gaitpost.2016.08.018] [Medline: 27567449]

XSL•FO
RenderX

109. Ellis RJ, Ng YS, Zhu S, et al. A validated smartphone-based assessment of gait and gait variability in Parkinson's disease. PLoS ONE 2015;10(10):e0141694. [doi: 10.1371/journal.pone.0141694]
110. Kostikis N, Hristu-Varsakelis D, Arnaoutoglou M, Kotsavasiloglou C. Smartphone-based evaluation of parkinsonian hand tremor: quantitative measurements vs clinical assessment scores. Annu Int Conf IEEE Eng Med Biol Soc 2014;2014:906-909. [doi: 10.1109/EMBC.2014.6943738] [Medline: 25570106]
111. Barrantes S, Egea AJS, Rojas HAG, et al. Differential diagnosis between Parkinson's disease and essential tremor using the smartphone's accelerometer. PLoS ONE 2017;12(8):e0183843. [doi: 10.1371/journal.pone.0183843] [Medline: 28841694]
112. Ware S, Knouse LE, Draz I, Enikeeva A. Predicting ADHD symptoms using smartphone sensing data. In: Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers: Association for Computing Machinery; 2022:503-505. [doi: 10.1145/3544793.3563430]
113. Servia-Rodríguez S, Rachuri KK, Mascolo C, Rentfrow PJ, Lathia N, Sandstrom GM. Mobile sensing at the service of mental well-being: a large-scale longitudinal study. In: Proceedings of the 26th International Conference on World Wide Web: International World Wide Web Conferences Steering Committee; 2017:103-112. [doi: 10.1145/3038912.3052618]
114. Lane ND, Lin M, Mohammod M, et al. BeWell: sensing sleep, physical activities and social interactions to promote wellbeing. Mobile Netw Appl 2014 Jun;19(3):345-359. [doi: 10.1007/s11036-013-0484-5]
115. What are sleep disorders? American Psychiatric Association. URL: https://www.psychiatry.org/patients-families/sleep-disorders/what-are-sleep-disorders [accessed 2026-02-28]
116. Thakur KT, Albanese E, Giannakopoulos P, Jette N, Linde M, Prince MJ, et al. Neurological disorders. In: Mental, Neurological, and Substance Use Disorders: Disease Control Priorities, Third Edition (Volume 4): The International Bank for Reconstruction and Development / The World Bank; 2016:87-107. [doi: 10.1596/978-1-4648-0426-7_ch5]
117. Parkinson's disease. National Institute of Neurological Disorders and Stroke. URL: https://www.ninds.nih.gov/health-information/disorders/parkinsons-disease [accessed 2026-02-28]
118. Charron E, White A, Carlston K, et al. Prospective acceptability of digital phenotyping among pregnant and parenting people with opioid use disorder: a multisite qualitative study. Front Psychiatry 2023;14:1137071. [doi: 10.3389/fpsyt.2023.1137071] [Medline: 37139320]
119. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (N Y) 2023 Sep 8;4(9):100804. [doi: 10.1016/j.patter.2023.100804] [Medline: 37720327]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
EMA: ecological momentary assessment
GAD-7: Generalized Anxiety Disorder-7
ML: machine learning
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
SUD: substance use disorder
YMRS: Young Mania Rating Scale

Edited by S Brini; submitted 15.Sep.2025; peer-reviewed by D Lottridge, V Vojdanovski; accepted 28.Jan.2026; published 24.Mar.2026.

Please cite as:
Dumas A, Hokayem J, Goodman G, Venkatasubramanian K, Chai P
Smartphone-Based Digital Phenotyping Across Health Conditions: Scoping Review
J Med Internet Res 2026;28:e84146
URL: https://www.jmir.org/2026/1/e84146
doi:10.2196/84146

© Arlen Dumas, Joanne Hokayem, Georgia Goodman, Krishna Venkatasubramanian, Peter Chai. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effects of Virtual Reality–Based Interventions for Promoting Physical Activity in Patients With Heart Failure: Systematic Review

Jeong-Ah Ahn[1], PhD; Jung Eun Lee[2], PhD; Kyoung-A Kim[3], PhD

[1]College of Nursing and Research Institute of Nursing Science, Ajou University, Suwon, Republic of Korea

[2]College of Nursing, University of Rhode Island, Kingston, RI, United States

[3]Suwon Women's University, 72, Onjeong-ro, Gwonseon-gu, Suwon, Republic of Korea

Corresponding Author:
Kyoung-A Kim, PhD
Suwon Women's University, 72, Onjeong-ro, Gwonseon-gu, Suwon, Republic of Korea

Abstract

Background: Heart failure (HF) is a progressive chronic condition associated with reduced physical and functional capacity, psychological burden, cognitive decline, and diminished quality of life (QOL). Although exercise-based cardiac rehabilitation is beneficial, participation remains low due to accessibility, physical constraints, and motivational barriers. Virtual reality (VR)–based interventions, including immersive platforms and exergaming, may enhance accessibility and engagement and promote physical activity through interactive experiences. However, evidence regarding their effectiveness in patients with HF remains fragmented.

Objective: This systematic review synthesized current evidence on the effects of VR-based interventions on physical activity, psychosocial outcomes, and self-management behaviors in patients with HF.

Methods: We systematically searched PubMed, CINAHL, Embase, and Scopus for studies published within the past 10 years. Randomized controlled trials (RCTs) and non-RCT interventional studies involving adults with HF who participated in VR-based interventions were eligible. Outcomes included physical activity or exercise capacity, psychological well-being, self-management, and QOL. The reviewers screened articles, extracted data, and assessed risk of bias using version 2 of the Cochrane risk-of-bias tool for randomized trials for RCTs and the Risk of Bias in Nonrandomized Studies of Interventions tool for non-RCTs. The review adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 and PRISMA-S (PRISMA extension for reporting literature searches) guidelines.

Results: A total of 10 studies met the inclusion criteria, comprising 7 (70%) RCTs and 3 (30%) non-RCTs. Studies were conducted across multiple countries and predominantly included older adults (mean age ≥65 years). Most interventions were home based, with exergaming as the most frequent modality, followed by immersive VR cycling and digital coaching programs, delivered over 4 to 12 weeks. Across studies, VR-based interventions were associated with improvements in exercise capacity (n=6, 60% of the studies), physical activity (n=5, 50%), and QOL (n=4, 40%). Three of the studies (30%) reported reductions in depressive symptoms, whereas effects on anxiety and self-efficacy were inconsistent. Adherence and usability were high across studies, and no intervention-related adverse events were reported. However, the risk of bias was rated as "some concerns" or "high" in several domains, and heterogeneity in intervention design and outcome measurement, along with small samples, limited pooled synthesis and overall certainty of evidence.

Conclusions: VR-based interventions show promise as accessible and engaging approaches to promote physical activity and support rehabilitation in patients with HF, particularly in home-based settings. Across the included studies, VR interventions were generally associated with improvements in exercise capacity, physical activity, QOL, and depressive symptoms, with high adherence and no reported safety concerns. However, interpretation is limited by heterogeneity in intervention design, small sample sizes, and methodological constraints. Future research should prioritize larger, rigorously designed trials to support sustained clinical impact.

(J Med Internet Res 2026;28:e86567) doi:10.2196/86567

KEYWORDS

virtual reality; heart failure; physical activity; systematic review; rehabilitation

Introduction

Heart failure (HF) is a chronic, progressive condition that affects millions of people worldwide, imposing a substantial burden on individuals, families, and health care systems. As the final pathway of various cardiovascular diseases, the prevalence of HF continues to rise due to advances in cardiac care that have improved survival rates, along with the global aging population [1,2]. HF is associated not only with physiological decline but

XSL•FO
RenderX

also with multidimensional challenges that impact patients' psychological, cognitive, and social well-being [1]. Common symptoms such as fatigue, dyspnea, and exercise intolerance limit physical activity and may lead to deconditioning, reduced mobility, and loss of independence [3]. In addition, depression, anxiety, cognitive impairment, and diminished quality of life (QOL) are highly prevalent and further complicate disease management and self-care [4,5].

Maintaining adequate physical activity is an essential component of HF management and rehabilitation. Evidence consistently indicates that regular exercise improves exercise tolerance, muscle strength, endothelial function, and overall cardiovascular health [6]. Moreover, physical activity has positive effects on psychological outcomes such as mood, motivation, and perceived QOL. However, many patients with HF remain inactive due to fear of symptom exacerbation, lack of motivation, comorbidities, and limited access to supervised rehabilitation programs [7]. Although traditional cardiac rehabilitation (CR) is effective, participation rates remain low because of transportation barriers, cost and time burdens, and reduced self-efficacy [8,9]. Therefore, innovative and patient-centered approaches are needed to enhance participation and promote sustainable physical activity among individuals with HF.

In this context, virtual reality (VR) has emerged as a promising technological tool for health promotion and disease management. VR refers to computer-generated, interactive, 3D environments that simulate real or imagined scenarios, allowing users to engage in immersive experiences [10]. These environments can elicit strong sensory and emotional engagement, making activities more enjoyable, motivating, and meaningful. The immersive nature of VR creates a sense of presence—an illusion of "being there"—which enhances focus, reduces distraction, and increases adherence to therapeutic exercises [11]. In health care, VR has been applied in various fields, including pain management, mental health therapy, motor rehabilitation, and chronic disease management [12-15]. Advances in hardware and software have made VR systems increasingly accessible, portable, and cost-effective, expanding their potential for clinical and home-based rehabilitation.

VR-based interventions have demonstrated positive outcomes across diverse patient populations. In neurological and musculoskeletal rehabilitation, VR programs have been shown to improve balance, motor coordination, and motivation [16-18]. In patients with chronic conditions such as stroke, Parkinson disease, and chronic obstructive pulmonary disease, VR-based exercise interventions have contributed to enhanced physical performance and QOL [19,20]. Moreover, VR provides a safe and controlled environment for graded physical activity tailored to patients' physical capacities and psychological needs. The real-time feedback and gamified nature of VR exercises can foster self-efficacy and promote behavior change [21]. These features are particularly valuable for patients with HF, who often face both physical limitations and psychological barriers to exercise participation. Therefore, VR-based rehabilitation may represent a paradigm shift from conventional, clinic-based rehabilitation programs toward more interactive and personalized care.

Despite growing evidence supporting the benefits of VR-based interventions, their application in HF populations remains relatively limited. Therefore, a comprehensive review of current evidence is required to understand both the potential and the limitations of VR in this population. Given the multifaceted nature of HF and the capacity of VR-based programs to improve physical activity, a systematic review on this topic has significant clinical implications. Evaluating the effects of VR interventions on physical activity and related outcomes among patients with HF can help clarify whether VR serves as an effective adjunct or alternative to traditional rehabilitation. Furthermore, identifying the types of VR interventions, their key components, and outcome measures can inform best practices for future program design.

Therefore, this systematic review aimed to synthesize current evidence on the effects of VR-based interventions for promoting physical activity, psychosocial outcomes, and self-management behaviors in patients with HF. By synthesizing and critically appraising existing evidence, this study aimed to establish a scientific foundation for incorporating VR technologies into chronic disease self-management strategies for patients with HF.

Methods

Search Strategy

This systematic review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [22] and PRISMA-S (PRISMA extension for reporting literature searches) guidelines [23]. A comprehensive search was conducted in 4 electronic databases: PubMed, CINAHL, Embase, and Scopus. Each database was searched individually on its own platform (PubMed via the National Center for Biotechnology Information, CINAHL via EBSCOhost, Embase via Elsevier, and Scopus via Elsevier); no simultaneous multidatabase searching on a single platform was conducted. The search covered the period from January 1, 2016, to December 31, 2025, and the final search was completed on January 2, 2026. All retrieved articles were compiled, and duplicates were removed. The bibliographic management software EndNote (version 21; Clarivate Analytics) was used to manage and organize references.

The search terms were developed based on the population, intervention, comparison, and outcomes framework: "heart failure" (population); "virtual reality," "immersive virtual reality," "augmented reality," "mixed reality," "exergaming," or "exergame" (intervention); "standard care" or "traditional exercise programs" (comparison); and "physical activity," "exercise capacity," "psychological and cognitive outcomes," "self-management behaviors," and "quality of life" (outcomes), including relevant synonyms and related terms. Search terms were combined using Boolean operators ("AND" and "OR") and adapted to each database's indexing system. Both MeSH (Medical Subject Headings) and free-text keywords were used to ensure a comprehensive search. The search strategy was further expanded to include a wider range of controlled vocabulary terms and text words related to HF (eg, "cardiac failure" and "cardiac insufficiency"), virtual and extended reality

XSL•FO
RenderX

technologies (eg, "virtual reality exposure therapy," "mixed reality," "smart glasses," and "head-mounted display"), exergaming and active video gaming, and physical activity–related outcomes (eg, "exercise" and "fitness"). The search strategy applied limits to publication year (past 10 years) and language (English only). The publication period was limited to the last 10 years to capture current evidence. The full reproducible search strategies for each database are provided in Multimedia Appendix 1 in accordance with PRISMA-S recommendations. The search strategy was peer reviewed by an independent librarian to ensure comprehensiveness and accuracy. Two reviewers independently conducted the database searches and verified the search outputs.

The inclusion criteria were (1) studies involving patients with HF, (2) VR-based interventional studies, (3) randomized controlled trials (RCTs) or non-RCTs, and (4) articles published in peer-reviewed journals. Eligible studies were required to include at least one outcome related to physical activity (eg, exercise capacity, activity level, or performance), whereas additional outcomes such as psychological status, self-management behaviors, or QOL were also considered. Studies that did not aim to promote or measure physical activity in patients with HF were excluded. Exclusion criteria included reviews, protocols, conference abstracts, noninterventional studies, and studies that did not report physical activity or related outcomes, as well as those lacking VR components.

Two independent reviewers conducted the screening and selection process. Any discrepancies regarding study inclusion or exclusion were resolved through discussion and consensus. When consensus could not be reached, a third reviewer was consulted. Data were extracted for all outcomes related to physical activity, exercise capacity, psychosocial status, self-management behaviors, and QOL across all reported measures and time points in each study. Physical activity–related outcomes were considered primary outcomes for synthesis and interpretation, while psychosocial and self-management outcomes were treated as secondary outcomes.

In addition to outcome data, information on study characteristics (country and study design), participant characteristics (sample size, age, and sex), and intervention characteristics (type, duration, frequency, and setting) was extracted using a predefined data extraction form. When information was missing or unclear, data were reported as not available, and no assumptions or imputations were made; study authors were not contacted for clarification.

Risk-of-Bias Assessment

The methodological quality and risk of bias of the included studies were evaluated using standardized tools appropriate to each study design. For RCTs, version 2 of the Cochrane risk-of-bias tool for randomized trials was used [24]. This tool assesses five domains of potential bias: (1) the randomization process, (2) deviations from intended interventions, (3) missing outcome data, (4) measurement of outcomes, and (5) selection of the reported results. For nonrandomized studies, the Risk of Bias in Nonrandomized Studies of Interventions tool was used [25]. The Risk of Bias in Nonrandomized Studies of Interventions evaluates seven domains: (1) confounding, (2) selection of participants, (3) classification of interventions, (4) deviations from intended interventions, (5) missing data, (6) measurement of outcomes, and (7) selection of the reported results.

Two reviewers independently assessed each study using these tools. Any disagreements were resolved through discussion or, when necessary, consultation with a third reviewer. The overall risk of bias was categorized into 3 levels: low risk, some concerns or moderate, or high or serious risk.

Synthesis Methods

Studies included in the narrative synthesis met the inclusion criteria and reported outcomes related to physical activity or exercise capacity in patients with HF. Studies were grouped by intervention type, study setting, and outcome domain and synthesized descriptively using reported data without conversion or imputation; meta-analysis was not conducted due to heterogeneity across studies, and potential sources of heterogeneity were explored descriptively.

Results

General Characteristics and Research Methodology of the Included Studies

A PRISMA flow diagram (Figure 1) was used to illustrate the study selection process. A total of 832 records were identified through PubMed, CINAHL, Embase, and Scopus. Of these 832 records, after removing 274 (32.9%) duplicates, 558 (67.1%) studies remained for title and abstract screening. Of these 558 articles, 540 (96.8%) articles were excluded for not meeting the inclusion criteria. The full texts of 18 studies were then reviewed, and 8 (44.4%) were excluded due to ineligible populations, interventions, or outcome reporting—specifically, studies aiming to improve knowledge and self-care, to manage pain, to compare hemodynamic parameters, to confirm system usability and satisfaction, or to assess symptom experience—resulting in a final selection of 10 (55.6%) articles. These comprised 7 RCTs and 3 nonrandomized single-arm studies. Table 1 and Multimedia Appendix 2 [26-35] summarize the general characteristics and research methodologies of the included studies. The studies were conducted in various countries, including Sweden, the United States, Germany, Israel, the Netherlands, Brazil, Italy, Romania, and Spain. Sample sizes ranged from 10 to 605 participants, and most studies (6/10, 60%) included older adults, with a mean age of ≥65 years. In total, 80% (n=8) of the interventions were implemented in home-based settings, whereas 20% (n=2) were conducted in hospital environments.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the study selection process. VR: virtual reality.

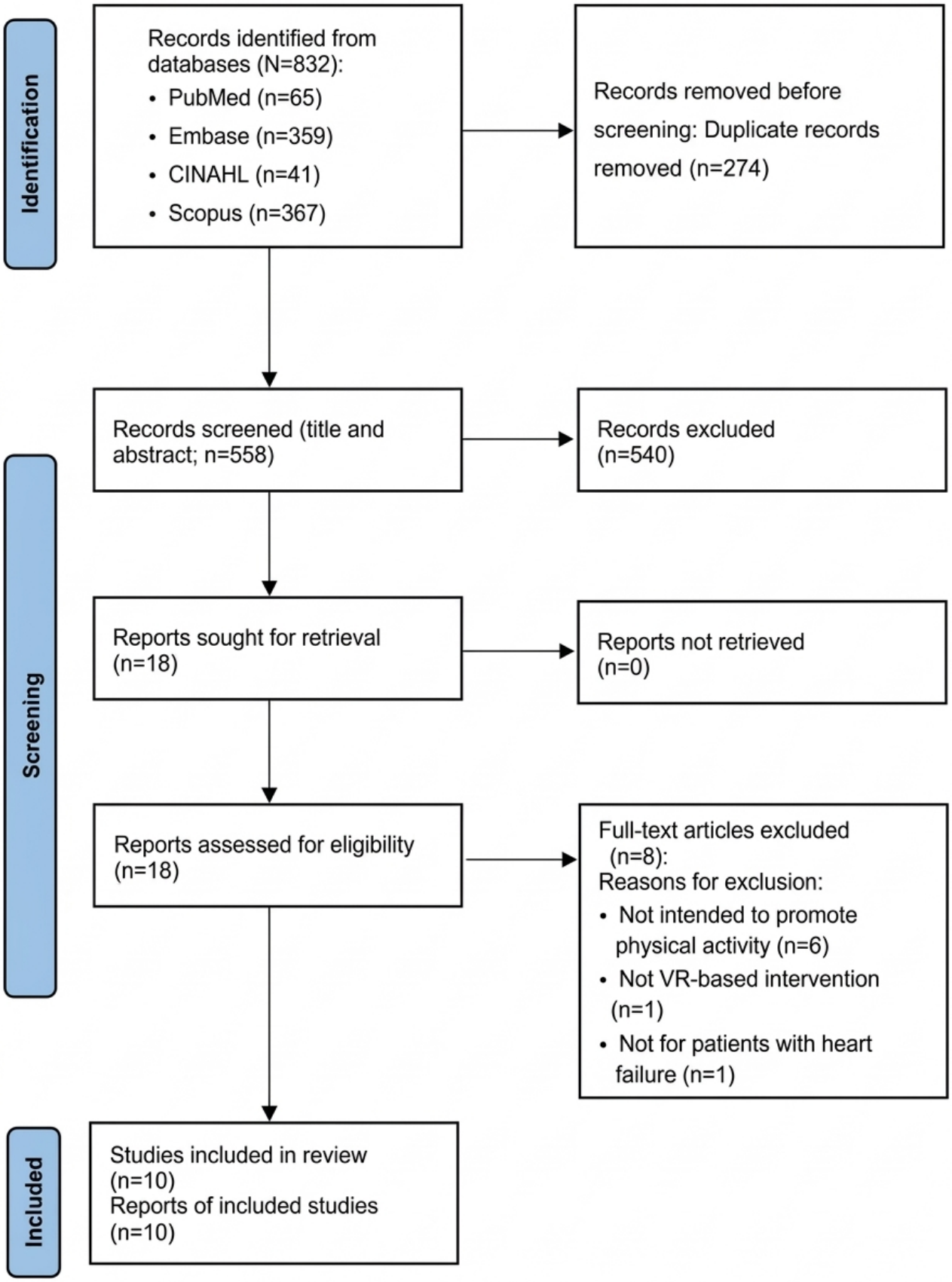

XSL•FO
RenderX

Table . Characteristics of the included studies (N=10).

Variable and category	Studies, n (%)
Disease of participants	
Heart failure	10 (100)
Country[a]	
Sweden	5 (50)
United States	4 (40)
Germany	3 (30)
Israel	3 (30)
The Netherlands	3 (30)
Brazil	2 (20)
Italy	2 (20)
Others (Romania and Spain)	2 (20)
Mean age of participants (y)	
<65	3 (30)
≥65	6 (60)
Not mentioned	1 (10)
Sample size	
<50	4 (40)
50-100	3 (30)
>100	3 (30)
Use of theoretical framework	
Yes	1 (10)
No	9 (90)
Study design	
Randomized controlled trial	7 (70)
1-group (single-arm) study	3 (30)
Setting	
Home	8 (80)
Hospital	2 (20)
Type of intervention	
Exergame	6 (60)
Immersive VR[b]-assisted cycle ergometer	2 (20)
Sensor-controlled digital game	1 (10)
Virtual application for cardiac rehabilitation	1 (10)
Intervention period	
During hospitalization	2 (20)
4 wk	2 (20)
12 wk	6 (60)
Follow-up frequency	
Once	7 (70)
3 times	3 (30)

[a]Some studies were conducted across multiple countries simultaneously. Therefore, the total number of countries listed exceeds the total number of included studies (N=10) as each multinational study was counted once per country.

XSL•FO
RenderX

[b]VR: virtual reality.

Only 10% (1/10) of the studies [29] applied a theoretical framework—the Fogg behavior model—to guide intervention design, whereas the remaining studies were empirically developed. Intervention durations varied, with most (6/10, 60%) lasting 12 weeks, 20% (2/10) lasting 4 weeks, and 20% (2/10) being conducted during hospitalization. The frequency of follow-up assessments ranged from a single posttest to up to 3 follow-up evaluations.

Regarding the types of VR-based interventions, exergaming was the most common approach, including Nintendo Wii–based programs (5/10, 50%) and a mobile exergame (Heart Farming; 1/10, 10%), followed by immersive VR-assisted cycle ergometer training (2/10, 20%), sensor-controlled digital games (1/10, 10%), and a virtual assistant–based CR program (1/10, 10%). A total of 40% (4/10) of the included studies [30-33] originated from the same RCT, the HF-Wii study. These publications represented secondary or subgroup analyses conducted in multinational cohorts (Sweden, Italy, the Netherlands, Israel, Germany, and the United States) that explored distinct outcome domains such as exercise capacity, psychological well-being, and cognitive predictors. To prevent double counting of data, these HF-Wii–derived studies were narratively synthesized and summarized separately rather than aggregated in quantitative comparisons.

Risk of Bias

Of the 7 RCTs, 6 studies were rated as having some concerns, whereas 1 study showed a high risk of bias. Of the 3 non-RCTs, 2 studies showed a moderate risk of bias, whereas 1 feasibility study was judged to have a serious risk of bias. The summarized assessments are presented in Figure 2 [29-35] and in Figure 3 [26-28].

Figure 2. Summary of the risk-of-bias assessment using version 2 of the Cochrane risk-of-bias tool for randomized controlled trials.

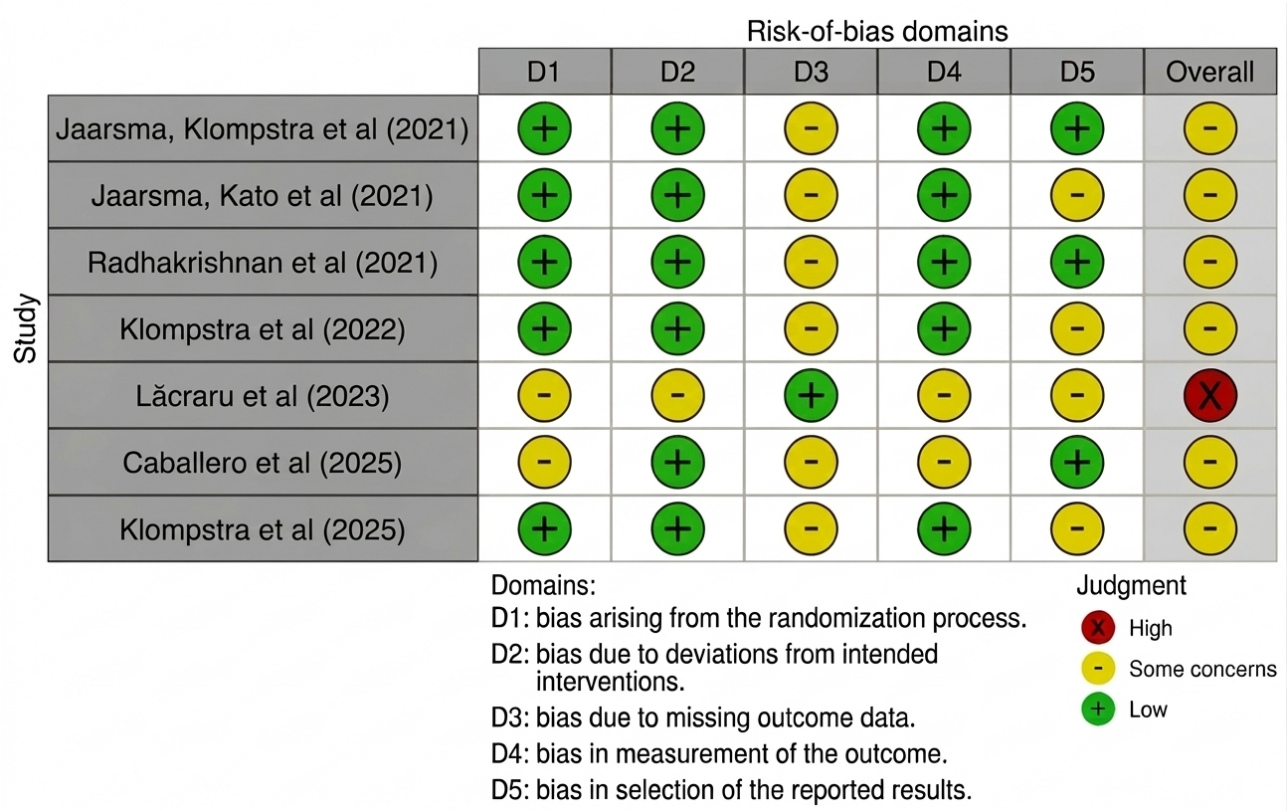

XSL•FO
RenderX

Figure 3. Summary of the risk-of-bias assessment using the Risk of Bias in Nonrandomized Studies of Interventions tool.

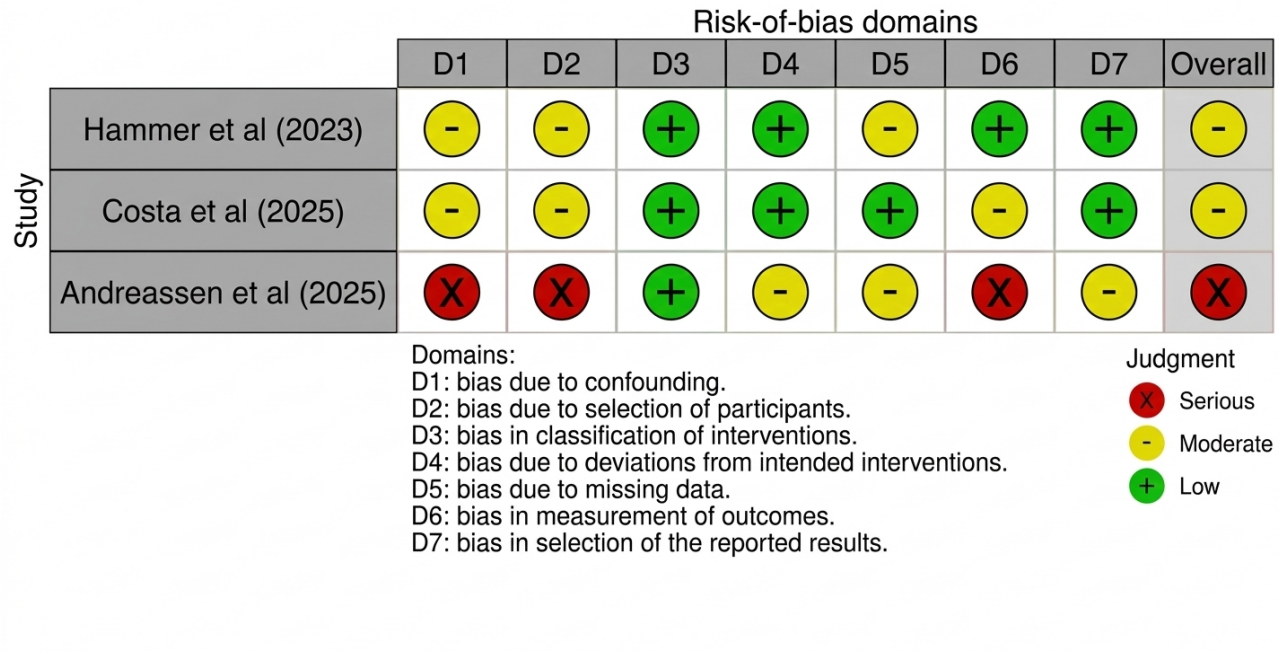

Effects of VR-Based Interventions in Patients With HF

Table 2 and Multimedia Appendix 2 [26-35] show the detailed measurements, tools, and main results related to the effects of VR-based interventions in patients with HF.

XSL•FO
RenderX

Table . Outcome measurements of the included studies (N=10).

Variable[a] and category	Studies, n (%)
Primary outcomes (physical activity related)	
Exercise capacity (6MWT[b] and VO_2 max[c])	6 (60)
Exercise motivation	3 (30)
Exercise self-efficacy	3 (30)
Self-reported physical activity	2 (20)
Physical activity behaviors	2 (20)
Muscle strength	1 (10)
Daily walking distance	1 (10)
Mobilization experience	1 (10)
Sedentary time	1 (10)
Secondary outcomes	
Anxiety and depression	5 (50)
Quality of life	4 (40)
Well-being	2 (20)
HF[d] self-management behaviors	2 (20)
HF self-management knowledge	1 (10)
HF self-efficacy	1 (10)
Cognitive function	1 (10)
HF hospitalization	1 (10)
Feasibility (acceptability, adherence, and retention)	3 (30)
Usability (user experience, system usability, and technology acceptance)	3 (30)
Safety	1 (10)

[a]In some cases, multiple reports or publications were derived from the same study. These entries share the same study population or dataset.

[b]6MWT: 6-minute walk test.

[c]VO_2 max: maximal oxygen consumption.

[d]HF: heart failure.

Primary and Secondary Outcome Measurements

The most frequently assessed primary outcomes were exercise capacity—defined as physiological performance indicators such as the 6-minute walk test (6MWT) or maximal oxygen consumption (VO_2 max)—reported in 60% (6/10) of the studies. In contrast, physical activity, treated as a distinct behavioral domain, was measured through self-reported activity levels, behavioral activity logs, muscle strength tests, mobilization experience, or sedentary time. Other primary outcomes included exercise motivation and self-efficacy (3/10, 30% of the studies each).

Secondary outcomes included anxiety and depression (5/10, 50%), QOL (4/10, 40%), well-being (2/10, 20%), and HF self-management behaviors (2/10, 20%), as well as intervention-related variables such as feasibility, usability, and safety.

Effects on Physical Outcomes

Overall, VR-based interventions demonstrated positive trends in improving physical activity and exercise performance among patients with HF. In total, 50% (5/10) of the studies, which used home-based exergaming Nintendo Wii programs, reported improvements in 6MWT distances, with 2 of these studies showing statistically significant gains over 3 to 6 months ($P<.05$). For example, Klompstra et al [32] reported significant increases in exercise capacity and reductions in fatigue and dyspnea after 3 to 6 months, although these effects diminished by 12 months. Similarly, Hammer et al [26] found that VR exergaming improved 6MWT performance and QOL in patients with HF supported by left ventricular assist devices ($P=.02$). In addition, the mobile exergame (Heart Farming) demonstrated that daily walking targets were manageable and achievable, as confirmed by in-app tracking of walking distance, supporting the feasibility of integrating regular walking activity into daily routines [27].

The vCare virtual assistant app trial by Lăcraru et al [34] demonstrated significant improvements in VO_2 max ($P=.002$)

XSL•FO
RenderX

and reductions in low-density lipoprotein cholesterol ($P<.05$) compared with standard care.

Among hospitalized patients, Caballero et al [35] and Costa et al [28] found that immersive VR-assisted mobilization and exercise were safe, feasible, and associated with high levels of exercise enjoyment and positive usability ratings.

Effects on Psychological and Cognitive Outcomes

Across the included studies, VR-based programs showed mixed but generally positive psychological effects. In total, 50% (5/10) of the studies assessed anxiety and depression, with 3 of these studies reporting significant postintervention reductions ($P<.05$). Lăcraru et al [34] found significant improvements in depression scores ($P=.03$), whereas anxiety levels remained unchanged. In contrast, Klompstra et al [32] observed no significant changes in anxiety or depression despite improvements in physical outcomes.

A total of 20% (2/10) of the studies evaluated overall well-being, reporting small to moderate improvements immediately following the intervention. Cognitive function was assessed in one secondary analysis [31], which identified lower baseline cognitive scores as a predictor of nonimprovement in exercise performance (odds ratio 0.87, 95% CI 0.80-0.94), suggesting that cognitive status may influence the effectiveness of VR-based interventions.

Effects on Self-Management and QOL Outcomes

In total, 20% (2/10) of the studies focused on self-management behaviors. The sensor-controlled digital game intervention (Heart Health Mountain) by Radhakrishnan et al [29] significantly improved HF self-management knowledge ($P<.05$), QOL ($P<.01$), and adherence behaviors such as daily weight monitoring and physical activity ($r=0.9$; $P<.001$). Self-efficacy improved over 24 weeks, although between-group differences were not statistically significant. Additionally, reduced hospitalization rates were observed in both intervention and control groups over the 6-month follow-up period. The vCare digital coaching program [34] did not directly measure self-management behaviors but functioned as a self-management support tool. The intervention improved VO_2 max ($P=.002$), QOL ($P=.007$), and depressive symptoms, whereas participants reported high usability and motivation for continued engagement in home-based rehabilitation.

QOL was assessed in 40% (4/10) of the studies [26,29,32,34], all of which reported improvements following VR-based interventions.

The Heart Health Mountain program [29] demonstrated significant increases in QOL at 6, 12, and 24 weeks ($P<.01$), along with enhanced self-efficacy and adherence to daily self-care activities.

Similarly, the vCare virtual assistant app [34] improved QOL ($P=.007$) and reduced depressive symptoms ($P=.03$), suggesting that personalized, feedback-driven home rehabilitation can yield psychosocial benefits in addition to physiological gains. In a study of left ventricular assist device–supported patients with HF, Hammer et al [26] found significant improvements in both exercise capacity and QOL after a 4-week home-based exergaming program.

Finally, Klompstra et al [32] observed notable improvements in QOL and reductions in fatigue and dyspnea during the 3- to 6-month follow-up period, although these effects diminished by 12 months.

Feasibility, Usability, and Safety of VR-Based Interventions

Feasibility was reported in 30% (3/10) of the studies [26,27,29], showing high adherence, acceptability, and retention rates for home-based exergaming programs. Usability outcomes, assessed in the vCare [34] and VR-assisted cycling studies [28], indicated good to excellent user acceptance, with mean System Usability Scale scores exceeding 68 points. Similarly, the mobile exergame (Heart Farming) was perceived as easy to use, adaptable to individual needs, and engaging, with no major acceptability concerns reported [27]. Among hospitalized patients, immersive VR-assisted mobilization and exercise were also reported to be safe, feasible, and well tolerated [28,35].

No adverse events, cybersickness, or device-related safety concerns were reported in any of the included studies, indicating that VR-based interventions were well tolerated and safe for patients with HF.

Discussion

Principal Findings

This systematic review synthesized current evidence on the effects of VR-based interventions on physical activity, psychological outcomes, and self-management behaviors in patients with HF. Overall, the findings suggest that VR-based interventions can effectively enhance physical activity, exercise performance, and selected psychosocial outcomes in this population. Improvements in exercise capacity and QOL were the most consistently reported outcomes, while effects on psychological and behavioral domains were more variable across studies. The results also confirm the feasibility, usability, and safety of VR-based interventions in both hospital and home settings, demonstrating their potential as complementary or alternative strategies to traditional CR programs for patients with HF.

Consistent with prior meta-analyses on VR-based rehabilitation in chronic disease populations [20,36,37], the included studies demonstrated that VR can improve exercise capacity and physical activity levels in patients with HF. Exercise capacity, measured primarily through the 6MWT and VO_2 max, showed significant improvements in several studies (4/10, 40%), indicating the physiological benefits of engaging in VR-mediated physical activity. Home-based exergaming programs, such as the HF-Wii intervention, encouraged patients to exercise safely and independently while maintaining adherence levels comparable to those of center-based CR [32]. These findings align with those of previous research reporting that gamified exercise enhances motivation, perceived enjoyment, and long-term engagement with physical activity among individuals with chronic cardiovascular conditions [6,21].

XSL•FO
RenderX

The ability of VR to provide immersive and interactive exercise experiences may help overcome common barriers to physical activity in HF populations, including fear of symptom exacerbation, low motivation, and limited access to supervised programs [7,8]. By integrating features such as real-time feedback, adjustable difficulty levels, and visual immersion, VR facilitates safe, individualized exercise experiences that enhance self-efficacy and engagement. Several trials (3/10, 30%) in this review reported high adherence rates (>85%) and positive user feedback regarding enjoyment, indicating that VR-based training is both acceptable and feasible for older adults. These results suggest that VR can serve as a practical extension of center-based CR carried out in the home, especially for individuals who face logistical or psychosocial barriers that hinder participation in traditional center-based programs.

Notably, the interventions in this review ranged from nonimmersive exergaming (eg, Nintendo Wii) to more immersive VR-assisted cycling systems. Although both formats generally demonstrated beneficial effects, the current evidence does not clearly indicate that higher immersion consistently produces superior outcomes, partly because intervention content, exercise intensity, and patient characteristics varied considerably across studies. Therefore, the level of immersion should be viewed as one of several interacting design features rather than the sole driver of effectiveness. Therefore, future research should explicitly compare different levels of immersion to identify which VR design characteristics—such as sensory stimulation, interaction fidelity, or gamification—are most strongly associated with adherence and clinical outcomes.

Nevertheless, the magnitude of improvement in exercise capacity varied across studies, and the sustainability of these effects beyond short-term follow-up remains uncertain. Some trials demonstrated attenuation of benefits after 6 to 12 months [32], implying that continued engagement or periodic reinforcement may be necessary to maintain gains. Future studies should explore strategies such as adaptive difficulty adjustment, remote coaching, and integration with wearable monitoring technologies to support long-term adherence and sustained physical activity.

In addition to physical outcomes, several studies (4/10, 40%) revealed potential psychological and cognitive benefits of VR-based interventions. Improvements in depressive symptoms and emotional well-being were observed in 30% (3/10) of the studies [26,29,34], consistent with previous findings showing that VR-based exercise can enhance mood, reduce stress, and foster social connectedness [11,17,38]. These effects may be mediated by mechanisms such as endorphin release from physical activity, distraction from illness-related distress, and enhanced self-efficacy through interactive goal attainment [39].

Moreover, VR environments provide an engaging, multisensory experience that may alleviate anxiety and fear associated with exercise, particularly among patients with HF who often perceive exertion as a potential trigger for worsening symptoms [4,5]. By enabling safe, controlled participation in virtual exercise, VR interventions may help reduce psychological resistance to activity. Therefore, the findings of this review support the potential of VR in promoting not only physical rehabilitation but also emotional adjustment and self-confidence in managing chronic illness.

Cognitive function, though less frequently assessed, appears to influence the efficacy of VR interventions. Jaarsma et al [30] reported that patients with lower baseline cognitive performance were less likely to achieve improvements in physical outcomes, suggesting that cognitive impairment may hinder engagement with VR technologies. This is likely because VR requires users to attend to visual-spatial cues, remember instructions, and coordinate motor actions; therefore, attention, working memory, and executive functioning play an important role in task performance and learning using VR [40]. When these abilities are reduced, patients may struggle to follow VR tasks or maintain engagement, which may attenuate treatment effects. Prior evidence indicates that cognitive decline, common in HF due to cerebral hypoperfusion, negatively affects self-care and rehabilitation adherence [41]. Accordingly, future research should consider cognitive screening to guide intervention tailoring and the potential value of incorporating cognitive training features into VR programs.

Although relatively few studies (2/10, 20%) examined self-management outcomes, there is available evidence suggesting that VR-based interventions can positively influence health-related behaviors and disease knowledge. The Heart Health Mountain program [29] demonstrated significant improvements in HF knowledge, QOL, and adherence to daily self-care tasks such as weight monitoring and physical activity logging. These results highlight VR's educational potential—its interactive and gamified format can promote active learning, repetition, and reinforcement of self-management principles [15]. Integrating personalized behavioral feedback into VR platforms may further enhance self-regulation and motivation, supporting long-term lifestyle modification.

Nevertheless, behavior changes were inconsistent across studies. Some trials 2/10 (20%) reported only transient improvements that diminished over time, possibly due to limited intervention duration or insufficient strategies to maintain engagement [32]. Future research should investigate hybrid models that combine VR-based education with telemonitoring, remote coaching, or community support to promote sustained behavior change and long-term adherence. In addition, only 10% (1/10) of the studies used a clear theoretical framework to guide intervention design. Given that many VR programs aimed not only to increase physical activity but also to influence psychosocial or self-management behaviors, theory-driven approaches will be essential in future research to clarify mechanisms of change and enhance the sustainability of intervention effects.

One of the most encouraging findings of this review is the strong evidence for the feasibility, usability, and safety of VR-based interventions for patients with HF. No adverse cardiovascular events, cybersickness, or equipment-related injuries were reported across the 10 included studies. Adherence and retention rates were high, particularly in home-based interventions, indicating good acceptability among older adults. Usability assessments consistently yielded satisfactory results, suggesting that the VR systems used were intuitive and well tolerated even

XSL•FO
RenderX

among participants with limited experience using digital technologies [18].

These findings are consistent with those of previous systematic reviews of VR-based rehabilitation in other chronic disease populations, where safety, enjoyment, and user engagement were identified as key facilitators of program success [16,20]. The use of commercially available, low-cost gaming systems (eg, Wii) in several studies (5/10, 50%) further supports the scalability and cost-effectiveness of VR interventions for HF management. As portable VR systems become more widely available, their integration into community and home settings will be increasingly feasible, offering opportunities to expand access to rehabilitation for underserved and mobility-limited populations.

Although 40% (4/10) of the included publications were derived from the same HF-Wii RCT, the findings across these reports were not identical. This variation can largely be explained by differences in follow-up duration and outcome selection. The main HF-Wii trial evaluated overall changes in exercise capacity at 3 months and reported no significant improvement, whereas the subsequent substudies targeted specific subgroups; additional follow-up periods (eg, 6 and 12 months); or secondary outcomes such as QOL, fatigue, and physical activity patterns. Consequently, some beneficial effects—particularly those observed at 3 to 6 months—appeared only in certain subanalyses and did not consistently persist across all publications. These differences highlight that the HF-Wii findings should not be interpreted as repeated independent evidence but rather as multiple analytic perspectives derived from a single parent dataset.

The findings of this review have several important implications. Clinically, VR-based interventions may serve as valuable adjuncts to traditional CR programs by offering flexible, engaging means to enhance physical activity and self-management, particularly for patients unable to attend in-person sessions. Health professionals can leverage VR platforms for remote monitoring, individualized goal setting, and motivational reinforcement. Integration of VR systems with wearable sensors and telehealth platforms could enable real-time feedback and progress tracking, fostering a more personalized and continuous rehabilitation experience [6].

From a research perspective, future RCTs should use larger sample sizes, extended follow-up periods, and standardized outcome metrics to allow for cross-study comparison and meta-analyses. Multidomain interventions combining physical, cognitive, and psychosocial training within VR environments may yield synergistic effects on overall health and well-being. Additionally, cost-effectiveness analyses and qualitative studies on patient experiences will be essential to inform clinical implementation and policy decisions. As VR technology continues to evolve, ensuring accessibility, ease of use, and cultural adaptability will be key to maximizing its impact on HF care.

Limitations

Despite this systematic review presenting promising results, several methodological limitations should be acknowledged. First, the number of high-quality RCTs was limited, and most studies (4/10, 40%) involved small sample sizes, reducing statistical power and generalizability. Second, there was substantial heterogeneity in intervention design, duration, intensity, and outcome measures, which precluded meta-analysis and limited the ability to draw direct comparisons across studies. The diversity of VR modalities—ranging from nonimmersive exergames to fully immersive environments—also complicates interpretation of specific active components responsible for the observed benefits. Third, follow-up periods were typically short (4 - 12 weeks), making it difficult to determine whether the observed improvements translate into sustained lifestyle changes or long-term clinical outcomes such as reduced hospitalization or mortality. Fourth, only 10% (1/10) of the studies used a theoretical framework to guide intervention development and behavioral engagement. Future studies should integrate behavior change theories and standardized outcome measures to strengthen conceptual and methodological rigor. In addition, 40% (4/10) of the included studies originated from the same parent trial (HF-Wii), resulting in overlapping participant samples and repeated use of the same trial dataset. To minimize potential bias from this overlap, these studies were treated as secondary analyses of a single RCT and were narratively summarized rather than combined quantitatively. Nonetheless, because these publications draw on the same underlying sample, their collective contribution should be interpreted with caution as they do not represent independent evidence. Finally, the search period was restricted to the last 10 years to reflect contemporary VR technology and current CR practice. While this approach increases clinical relevance to modern practice, it may also have excluded earlier pioneering VR studies.

Conclusions

This systematic review advances the research field by providing the first focused synthesis of recent VR-based interventions specifically designed to promote physical activity in patients with HF while concurrently evaluating psychosocial and self-management outcomes. This study critically assessed research methodologies, intervention characteristics, and clinical effectiveness within the HF-specific context. By clarifying both the demonstrated short-term benefits and the methodological limitations of existing studies, this review contributes a conceptual and empirical road map for the future development of standardized VR-supported rehabilitation in HF. Importantly, the findings highlight the real-world potential of VR as a scalable, safe, and engaging home-based strategy capable of addressing persistent barriers to traditional CR. As digital health infrastructure continues to expand, evidence-based VR interventions could become an integral component of comprehensive HF self-management and rehabilitation programs. Nevertheless, because only 10% (1/10) of the included studies explicitly incorporated a formal behavior change theory, future large-scale, rigorously designed RCTs should integrate theoretical frameworks to guide intervention development and evaluation. Theory-driven designs will be essential for enhancing methodological rigor and ensuring the long-term sustainability of physical, psychosocial, and self-management outcomes in VR-based HF rehabilitation.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank Yujin Kwon, a medical research librarian at Ajou University, for her guidance on systematic literature review methods in this study. They are also grateful to the editor and reviewers who helped improve this manuscript.

Funding

This work was supported by a National Research Foundation of Korea grant funded by the Korean government (Ministry of Science and ICT; RS-2025-24532981). This funding source had no role in the study design, data collection, data analysis, data interpretation, or writing of the manuscript.

Data Availability

All data generated or analyzed during this systematic review are included in this published article and its supplementary files.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies.
[PDF File, 118 KB - jmir_v28i1e86567_app1.pdf]

Multimedia Appendix 2
Detailed data extraction of each study.
[PDF File, 199 KB - jmir_v28i1e86567_app2.pdf]

Checklist 1
PRISMA checklist.
[PDF File, 150 KB - jmir_v28i1e86567_app3.pdf]

Checklist 2
PRISMA-S checklist.
[PDF File, 225 KB - jmir_v28i1e86567_app4.pdf]

Checklist 3
SWiM reporting items.
[PDF File, 63 KB - jmir_v28i1e86567_app5.pdf]

References

1. Authors/Task Force Members, McDonagh TA, Metra M, et al. 2023 focused update of the 2021 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure: developed by the task force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur J Heart Fail 2024 Jan;26(1):5-17. [doi: 10.1002/ejhf.3024] [Medline: 38169072]
2. Savarese G, Lund LH. Global public health burden of heart failure. Card Fail Rev 2017 Apr;3(1):7-11. [doi: 10.15420/cfr.2016:25:2] [Medline: 28785469]
3. Zheng X, Zheng Y, Ma J, et al. Effect of exercise-based cardiac rehabilitation on anxiety and depression in patients with myocardial infarction: a systematic review and meta-analysis. Heart Lung 2019 Jan;48(1):1-7. [doi: 10.1016/j.hrtlng.2018.09.011] [Medline: 30366575]
4. Al-Sutari MM, Abdalrahim MS. Symptom burden and quality of life among patients with heart failure. SAGE Open Nurs 2024 Mar 28;10:23779608241242023. [doi: 10.1177/23779608241242023] [Medline: 38559894]
5. Freedland KE, Carney RM, Rich MW, Steinmeyer BC, Skala JA, Dávila-Román VG. Depression and multiple rehospitalizations in patients with heart failure. Clin Cardiol 2016 May;39(5):257-262. [doi: 10.1002/clc.22520] [Medline: 26840627]
6. Molloy C, Long L, Mordi IR, et al. Exercise-based cardiac rehabilitation for adults with heart failure. Cochrane Database Syst Rev 2024 Mar 7;3(3):CD003331. [doi: 10.1002/14651858.CD003331.pub6] [Medline: 38451843]
7. Dalal HM, Doherty P, Taylor RS. Cardiac rehabilitation. BMJ 2015 Sep 29;351:h5000. [doi: 10.1136/bmj.h5000] [Medline: 26419744]

XSL•FO
RenderX

8. Ades PA, Keteyian SJ, Wright JS, et al. Increasing cardiac rehabilitation participation from 20% to 70%: a road map from the Million Hearts cardiac rehabilitation collaborative. Mayo Clin Proc 2017 Feb;92(2):234-242. [doi: 10.1016/j.mayocp.2016.10.014] [Medline: 27855953]
9. Anderson L, Oldridge N, Thompson DR, et al. Exercise-based cardiac rehabilitation for coronary heart disease: Cochrane systematic review and meta-analysis. J Am Coll Cardiol 2016 Jan 5;67(1):1-12. [doi: 10.1016/j.jacc.2015.10.044] [Medline: 26764059]
10. Laver KE, Lange B, George S, et al. Virtual reality for stroke rehabilitation. Cochrane Database Syst Rev 2025 Jun 20;6(6):CD008349. [doi: 10.1002/14651858.CD008349.pub5] [Medline: 40537150]
11. Maples-Keller JL, Bunnell BE, Kim SJ, Rothbaum BO. The use of virtual reality technology in the treatment of anxiety and other psychiatric disorders. Harv Rev Psychiatry 2017;25(3):103-113. [doi: 10.1097/HRP.0000000000000138] [Medline: 28475502]
12. Goudman L, Jansen J, Billot M, et al. Virtual reality applications in chronic pain management: systematic review and meta-analysis. JMIR Serious Games 2022 May 10;10(2):e34402. [doi: 10.2196/34402] [Medline: 35536641]
13. Hajesmaeel Gohari S, Gozali E, Niakan Kalhori SR. Virtual reality applications for chronic conditions management: a review. Med J Islam Repub Iran 2019;33:67. [doi: 10.34171/mjiri.33.67] [Medline: 31456991]
14. Tabassum A, Ghaznavi I, Abd-Alrazaq A, Qadir J. Exploring the application of AI and extended reality technologies in metaverse-driven mental health solutions: scoping review. J Med Internet Res 2025 Aug 19;27:e72400. [doi: 10.2196/72400] [Medline: 40829151]
15. Triberti S, Repetto C, Riva G. Psychological factors influencing the effectiveness of virtual reality-based analgesia: a systematic review. Cyberpsychol Behav Soc Netw 2014 Jun;17(6):335-345. [doi: 10.1089/cyber.2014.0054] [Medline: 24892195]
16. Cano Porras D, Siemonsma P, Inzelberg R, Zeilig G, Plotnik M. Advantages of virtual reality in the rehabilitation of balance and gait: systematic review. Neurology 2018 May 29;90(22):1017-1025. [doi: 10.1212/WNL.0000000000005603] [Medline: 29720544]
17. Kim KA, Ahn JA. Effectiveness of immersive virtual reality simulation programs using head-mounted displays in promoting physical activity in older adults: a systematic review. Clin Simul Nurs 2024 Sep;94:101593. [doi: 10.1016/j.ecns.2024.101593]
18. Rodríguez-Mansilla J, Bedmar-Vargas C, Garrido-Ardila EM, et al. Effects of virtual reality in the rehabilitation of Parkinson's disease: a systematic review. J Clin Med 2023 Jul 26;12(15):4896. [doi: 10.3390/jcm12154896] [Medline: 37568298]
19. Matamala-Gomez M, Donegan T, Bottiroli S, Sandrini G, Sanchez-Vives MV, Tassorelli C. Immersive virtual reality and virtual embodiment for pain relief. Front Hum Neurosci 2019 Aug 21;13:279. [doi: 10.3389/fnhum.2019.00279] [Medline: 31551731]
20. Rutkowski S, Rutkowska A, Kiper P, et al. Virtual reality rehabilitation in patients with chronic obstructive pulmonary disease: a randomized controlled trial. Int J Chron Obstruct Pulmon Dis 2020 Jan 13;15:117-124. [doi: 10.2147/COPD.S223592] [Medline: 32021150]
21. García-Bravo S, Cuesta-Gómez A, Campuzano-Ruiz R, et al. Virtual reality and video games in cardiac rehabilitation programs. A systematic review. Disabil Rehabil 2021 Feb;43(4):448-457. [doi: 10.1080/09638288.2019.1631892] [Medline: 31258015]
22. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. J Clin Epidemiol 2021 Jun;134:178-189. [doi: 10.1016/j.jclinepi.2021.03.001] [Medline: 33789819]
23. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
24. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
25. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016 Oct 12;355:i4919. [doi: 10.1136/bmj.i4919] [Medline: 27733354]
26. Hammer Y, Shaul AA, Ben-Avraham B, et al. Exergaming in patients with a left ventricular assist device: a feasibility study. ESC Heart Fail 2023 Feb;10(1):738-741. [doi: 10.1002/ehf2.14203] [Medline: 36251532]
27. Andreassen M, Santaularia N, Berglund A, et al. Feasibility of a mobile exergame for patients with heart failure. Games Health J 2025 Oct;14(5):380-388. [doi: 10.1089/g4h.2024.0093] [Medline: 40201979]
28. Costa AS, Barbosa CB, Guizilini S, et al. Virtual reality and physical activity in patients with heart failure: technology validation and user satisfaction – pilot study. Int J Cardiovasc Sci 2025 Feb 19;38:e20240067. [doi: 10.36660/ijcs.20240067]
29. Radhakrishnan K, Julien C, Baranowski T, et al. Feasibility of a sensor-controlled digital game for heart failure self-management: randomized controlled trial. JMIR Serious Games 2021 Nov 8;9(4):e29044. [doi: 10.2196/29044] [Medline: 34747701]
30. Jaarsma T, Klompstra L, Ben Gal T, et al. Effects of exergaming on exercise capacity in patients with heart failure: results of an international multicentre randomized controlled trial. Eur J Heart Fail 2021 Jan;23(1):114-124. [doi: 10.1002/ejhf.1754] [Medline: 32167657]

XSL•FO
RenderX

31. Jaarsma T, Perkiö Kato N, Ben Gal T, et al. Factors associated with lack of improvement in submaximal exercise capacity of patients with heart failure. ESC Heart Fail 2021 Dec;8(6):4539-4548. [doi: 10.1002/ehf2.13584] [Medline: 34480412]
32. Klompstra L, Hägglund E, Jaarsma T, Kato NP, Strömberg A. Effects of exergaming and yoga on exercise capacity and physical and mental health in heart failure patients: a randomized sub-study. Eur J Cardiovasc Nurs 2025 Apr 11;24(3):389-398. [doi: 10.1093/eurjcn/zvae155] [Medline: 39743240]
33. Klompstra L, Jaarsma T, Piepoli MF, et al. Objectively measured physical activity in patients with heart failure: a sub-analysis from the HF-Wii study. Eur J Cardiovasc Nurs 2022 Jun 30;21(5):499-508. [doi: 10.1093/eurjcn/zvab133] [Medline: 34993536]
34. Lăcraru AE, Busnatu S, Pană MA, et al. Assessing the efficacy of a virtual assistant in the remote cardiac rehabilitation of heart failure and ischemic heart disease patients: case-control study of Romanian adult patients. Int J Environ Res Public Health 2023 Feb 22;20(5):3937. [doi: 10.3390/ijerph20053937] [Medline: 36900948]
35. Caballero LG, Fraga IB, Tremea CEM, et al. Experience of heart failure patients in mobilization with virtual reality: mixed methods study. Rev Esc Enferm USP 2025 Sep 15;59:e20250114. [doi: 10.1590/1980-220X-REEUSP-2025-0114en] [Medline: 40955916]
36. Cugusi L, Prosperini L, Mura G. Exergaming for quality of life in persons living with chronic diseases: a systematic review and meta-analysis. PM R 2021 Jul;13(7):756-780. [doi: 10.1002/pmrj.12444] [Medline: 32592238]
37. Kwon SH, Park JK, Koh YH. A systematic review and meta-analysis on the effect of virtual reality-based rehabilitation for people with Parkinson's disease. J Neuroeng Rehabil 2023 Jul 20;20(1):94. [doi: 10.1186/s12984-023-01219-3] [Medline: 37475014]
38. García-Betances RI, Jiménez-Mixco V, Arredondo MT, Cabrera-Umpiérrez MF. Using virtual reality for cognitive training of the elderly. Am J Alzheimers Dis Other Demen 2015 Feb;30(1):49-54. [doi: 10.1177/1533317514545866] [Medline: 25107931]
39. Martín-Rodríguez A, Gostian-Ropotin LA, Beltrán-Velasco AI, et al. Sporting mind: the interplay of physical activity and psychological health. Sports (Basel) 2024 Jan 22;12(1):37. [doi: 10.3390/sports12010037] [Medline: 38275986]
40. Lachowicz M, Serweta-Pawlik A, Żurek A, et al. Application of short and long-term virtual reality training for visuospatial memory development in amateur e-athletes. Virtual Real 2025;29:136. [doi: 10.1007/s10055-025-01209-x]
41. Goh FQ, Kong WKF, Wong RCC, et al. Cognitive impairment in heart failure-a review. Biology (Basel) 2022 Jan 23;11(2):179. [doi: 10.3390/biology11020179] [Medline: 35205045]

Abbreviations

6MWT: 6-minute walk test
CR: cardiac rehabilitation
HF: heart failure
MeSH: Medical Subject Headings
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for reporting literature searches
QOL: quality of life
RCT: randomized controlled trial
SWiM: Synthesis Without Meta-Analysis
VO_2 max: maximal oxygen consumption
VR: virtual reality

Edited by S Brini; submitted 27.Oct.2025; peer-reviewed by AJ Umeano, C Aladeokin, Y Zhang; revised version received 22.Feb.2026; accepted 24.Feb.2026; published 24.Mar.2026.

Please cite as:
Ahn JA, Lee JE, Kim KA
Effects of Virtual Reality–Based Interventions for Promoting Physical Activity in Patients With Heart Failure: Systematic Review
J Med Internet Res 2026;28:e86567
URL: https://www.jmir.org/2026/1/e86567
doi:10.2196/86567

© Jeong-Ah Ahn, Jung Eun Lee, Kyoung-A Kim. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any

XSL•FO
RenderX

medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Current Landscape of Remote Digital Symptom Monitoring for Patients With Lung Cancer: Scoping Review

Hao Huang[1,2], MSN; Naomi Takemura[1,2], PhD; Hammoda Abu-Odah[1,2], PhD; Shuhan Li[1,2], MSN; Ting Mao[1,2], MSN; Zengjie Ye[3], PhD; Janelle Yorke[1,2,4], PhD

[1]School of Nursing, The Hong Kong Polytechnic University, 11 Yuk Choi Road, Hung Hom, Kowloon, Hong Kong, China (Hong Kong)

[2]JC STEM Lab of Digital Oncology Care Enhancement (DOCE), The Hong Kong Jockey Club Charities Trust, Hong Kong, China (Hong Kong)

[3]School of Nursing, Guangzhou Medical University, Guangzhou, China

[4]Division of Nursing, Midwifery and Social Work, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, United Kingdom

Corresponding Author:
Janelle Yorke, PhD
School of Nursing, The Hong Kong Polytechnic University, 11 Yuk Choi Road, Hung Hom, Kowloon, Hong Kong, China (Hong Kong)

Abstract

Background: Remote digital symptom monitoring systems (rSMS) have been increasingly used in recent years to monitor symptoms, health-related quality of life, and other patient-reported outcomes in lung cancer. Previous studies have demonstrated variability in study design, types of rSMS, and outcomes used to assess benefits for patients and health care systems. However, there remains a lack of synthesized evidence pertaining to the similarities and differences among rSMS, including their theoretical underpinnings, key functional components, and reported benefits and limitations.

Objective: This review aims to identify and synthesize existing research to map the current landscape of rSMS in lung cancer, including the theoretical foundations for its development and implementation, as well as its types, applications, and outcomes.

Methods: This scoping review followed the Joanna Briggs Institute scoping review framework and adhered to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines. A comprehensive literature search was conducted from database inception to October 16, 2025, across 7 English-language databases and 3 Chinese-language databases (CNKI, WanFang, and SinoMed). Eligible studies were peer-reviewed original research articles examining rSMS among adults with lung cancer. Data were independently screened and extracted by 2 reviewers, with discrepancies resolved by a third reviewer. Quantitative data were extracted using a standardized form and synthesized descriptively. Content analysis was performed to analyze the qualitative data.

Results: A total of 41 studies involving 11,765 patients and 85 health care providers were included. Twelve studies focused exclusively on advanced-stage lung cancer. Participants were generally middle-aged to older adults (mean ages 51 - 74 y), with male participants typically comprising 30% to 50% across studies. Most studies were conducted in the United States (n=19). We identified 32 patient-reported outcome measures that were used either as core rSMS components or as study outcomes. Four common functional modules were observed across rSMS: data collection, data analysis, response systems, and patient education. Qualitative evidence was limited; the most frequently reported benefit was the promotion of patient-centered care. Health care providers raised concerns about uncertain effectiveness and increased workload.

Conclusions: This scoping review highlights the promising role of rSMS in lung cancer care and provides a structured map of current evidence. It adds to prior literature in 3 ways. First, it summarizes how and how often theoretical frameworks are reported and applied in rSMS development and implementation. Second, it synthesizes and categorizes four common functional modules across systems. Third, it differentiates measures embedded as rSMS components from those used as evaluation outcomes. These contributions clarify current practices and methodological gaps and underscore the importance of theory-informed design, functional clarity, and stakeholder engagement in the development of patient-centered, clinically meaningful, and sustainable rSMS platforms.

Trial Registration: OSF Registries 9637t; https://osf.io/9637t/overview

(J Med Internet Res 2026;28:e83666) doi:10.2196/83666

KEYWORDS

digital health; electronic patient-reported outcome measure; ePROM; lung cancer; symptom experience; scoping review; symptom monitoring

XSL•FO
RenderX

Introduction

Lung cancer remains the leading cause of morbidity and mortality in men and women worldwide [1,2]. Patients with lung cancer experience a range of debilitating symptoms, including dyspnea, hoarseness, persistent cough, and fatigue, all of which substantially diminish their health-related quality of life (HRQL) throughout the disease trajectory [3,4]. Consequently, identifying and monitoring these symptoms is critical to establishing a foundation for effective interventions and optimizing patient outcomes.

Patient-reported outcome measures (PROMs) have been adopted extensively to monitor and evaluate symptoms [3,5]. These are standardized questionnaires completed by patients to capture self-assessments of symptoms, HRQL, mental well-being, and functional status [6]. In recent years, PROMs have been largely recognized as highly valid and reliable instruments for assessing health care outcomes from the patient's perspective [7-9]. Moreover, several guidelines have been established to facilitate the standardized use and enhance the effectiveness of electronic PROMs (ePROMs) in clinical practice [10-13].

Advancements in digital health technologies have facilitated the evolution of ePROMs, defined as tools for the remote, digital collection of patient-reported outcomes outside clinical settings [14,15]. The application of ePROMs overcomes traditional scalability limitations, including reliance on in-person data collection, manual workflows that delay reporting, and fragmented communication between patients and clinicians.

A variety of terms have been used to describe electronic symptom monitoring interventions, including, but not limited to, electronic symptom self-reporting systems [16], remote symptom monitoring systems [17], electronic patient-reported outcome–based symptom management [18], and patient symptom self-reporting systems [19]. Despite differences in terminology, these interventions share the common feature of enabling remote monitoring of patient symptoms through electronic platforms. To promote uniformity and standardization, this review adopts the term "remote digital symptom monitoring system (rSMS)" to refer to systems designed for digitally monitoring patient-reported outcomes.

The implementation of rSMS into routine cancer care for patients with lung cancer has gained attention in recent years, driven by its potential to enhance person-centered care [18,20]. However, evidence regarding its clinical impact remains inconsistent, reflecting heterogeneity in the evidence base, where benefits have been observed in certain domains but not consistently across all outcomes. While the efficacy of rSMS in alleviating symptom-related outcomes is widely reported, 1 randomized controlled trial (RCT) found no significant reductions in symptom burden among patients with lung cancer [21]. Furthermore, limitations in improving outcomes such as HRQL [22] and overall survival have also been documented among patients with lung cancer [17].

The integration of rSMS into lung cancer care has surged in recent years, yet the field remains marked by heterogeneity in system design, PROMs used, and clinical application. Studies vary widely in scope and sophistication: some prioritize basic functionalities, such as symptom data aggregation via mobile platforms to streamline communication between patients and health care providers [22,23], while other studies use advanced technologies to deliver real-time feedback to health care providers or patients, facilitating effective symptom management [17,24,25]. The most advanced systems integrate automated clinical alerts triggered by predefined symptom thresholds, potentially improving HRQL through timely interventions [26-28]. However, this diversity has resulted in fragmented implementation, with no unifying framework to categorize systems or evaluate their alignment with lung cancer–specific needs. It is important to note that theoretical frameworks play a critical role in enhancing the methodological rigor and reproducibility of rSMS intervention designs and implementations. Theory-driven interventions aimed at promoting behavior change are known to enhance intervention benefits [29-31].

To address these gaps, this scoping review focuses on the development and implementation of theory-driven rSMS interventions in lung cancer care. It aims to identify relevant studies and map how rSMS interventions have been designed, applied, and evaluated in this population, guided by the following questions: (1) "What theoretical frameworks have been applied to studies of rSMS in lung cancer?"; (2) "What are the most commonly used PROMs for rSMS in lung cancer?"; (3) "What are the key technology platforms, functionalities, and characteristics of rSMS?"; (4) "What are the experiences and perceptions of patients and health care providers regarding the implementation of rSMS?"; and (5) "What are the benefits and limitations of various rSMS interventions on clinical outcomes in patients with lung cancer?"

Methods

Review Design

This scoping review followed the Joanna Briggs Institute scoping review framework [32] and adhered to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines (Checklist 1) [33]. This protocol was registered on the Open Science Framework [34].

Study Sources and Strategies

The literature search reporting adhered to the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) guidelines (Checklist 2) [35]. Comprehensive literature searches were conducted from inception to April 3, 2025, in the following English and Chinese electronic databases: (1) PubMed, (2) CINAHL (via EBSCOhost), (3) Cochrane Library, (4) Embase, (5) Scopus, (6) Web of Science, (7) PsycINFO, (8) CNKI, (9) WanFang, and (10) SinoMed. The search strategy was developed collaboratively by the research team and peer-reviewed by an experienced librarian at the institution. Given the broad scope and objectives of this scoping review, the key search terms selected were "lung cancer," "symptom monitoring," "electronic patient-reported outcome," and "remote monitoring." Of note, no limits were applied for date, time period, search filters, or

XSL•FO
RenderX

study design to maximize comprehensiveness. Both MeSH terms and free-text searches were used to ensure a thorough search. Additionally, the reference lists of included studies were reviewed to ensure no relevant evidence was overlooked [33]. If any included articles were found to contain potentially relevant information not provided in the original publication, the research team planned to obtain it by contacting the corresponding authors. The detailed search strategy used in each database is presented in Multimedia Appendix 1. The literature search was rerun on October 16, 2025, to update results.

Eligibility Criteria

This study used the PCC (population, concept, and context) strategy to establish eligibility criteria in line with the Joanna Briggs Institute recommendations [32].

Population

Studies involving adults aged 18 years or older, diagnosed with any type of lung cancer (primary or secondary) across all disease stages (stages I-IV), and receiving any treatment modality, regardless of therapeutic intent (curative or palliative), were included. Studies involving mixed cancer groups were eligible if lung cancer–specific data—particularly regarding symptom severity—were extractable.

Concept

Studies focusing on the use of rSMS in lung cancer, where health-related outcomes are reported directly by patients through digital platforms.

Context

Studies examining the use of rSMS in patients with lung cancer, remotely from home or community-based settings.

Types of Studies

This review included peer-reviewed primary studies (quantitative or qualitative) published in English or Chinese. Conference abstracts, reviews, comments, letters, and case reports were excluded, as these types of studies typically provide insufficient methodological detail or lack the ability to extract outcomes relevant to this review. Studies involving populations with mixed cancers, where lung cancer–specific data could not be isolated, were also excluded.

Study Screening and Selection

All search results were exported to EndNote 2025 (Clarivate Analytics), where duplicates were removed by 2 reviewers independently (HH and TM). Titles and abstracts were thoroughly screened, followed by a full-text review to determine final inclusion or exclusion by 2 reviewers independently (HH and TM). Disagreements were discussed and resolved through consultation with 2 independent members of the team (NT and HA-O). The selection process is illustrated in a PRISMA-ScR flowchart [33].

Data Extraction

Data relevant to the research questions were systematically extracted using a predesigned table developed by the research team. Specifically, the following details were extracted from each included study: (1) study characteristics (author, publication year, country, setting, sample size, and study design); (2) theoretical frameworks or models applied; (3) technology platforms for rSMS delivery and their key features (eg, accessibility, integration, and user interface); (4) clinical workflows for rSMS implementation (eg, with real-time clinician alerts and intervention, and automated responses to patients); (5) type of PROMs used and their symptom or health outcome domains; and (6) reported benefits and limitations of rSMS from patients, health care providers, and system perspectives.

For qualitative studies, thematic data on patients' and health care providers' experiences (eg, barriers, facilitators, and satisfaction) were also extracted.

Data Synthesis

Quantitative data, including study characteristics, the use of PROMs, and reported benefits and barriers, were summarized in tabular and narrative formats, with a narrative interpretation provided. Qualitative data, such as user experiences, were analyzed using content analysis and synthesized narratively [36].

Results

Study Selection

A total of 495 articles were retrieved through the initial search, comprising 491 from English-language databases and 4 from Chinese-language databases. After screening titles and abstracts, followed by a thorough full-text review, 80 English-language articles were deemed eligible, while all Chinese-language articles were excluded due to irrelevant content or study designs that did not meet the inclusion criteria. Additionally, 3 studies were identified through citation searching. Ultimately, 41 studies were included in the final analysis. The search process is depicted in a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart (Figure 1; Checklist 3).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. ePRO: electronic patient-reported outcome.

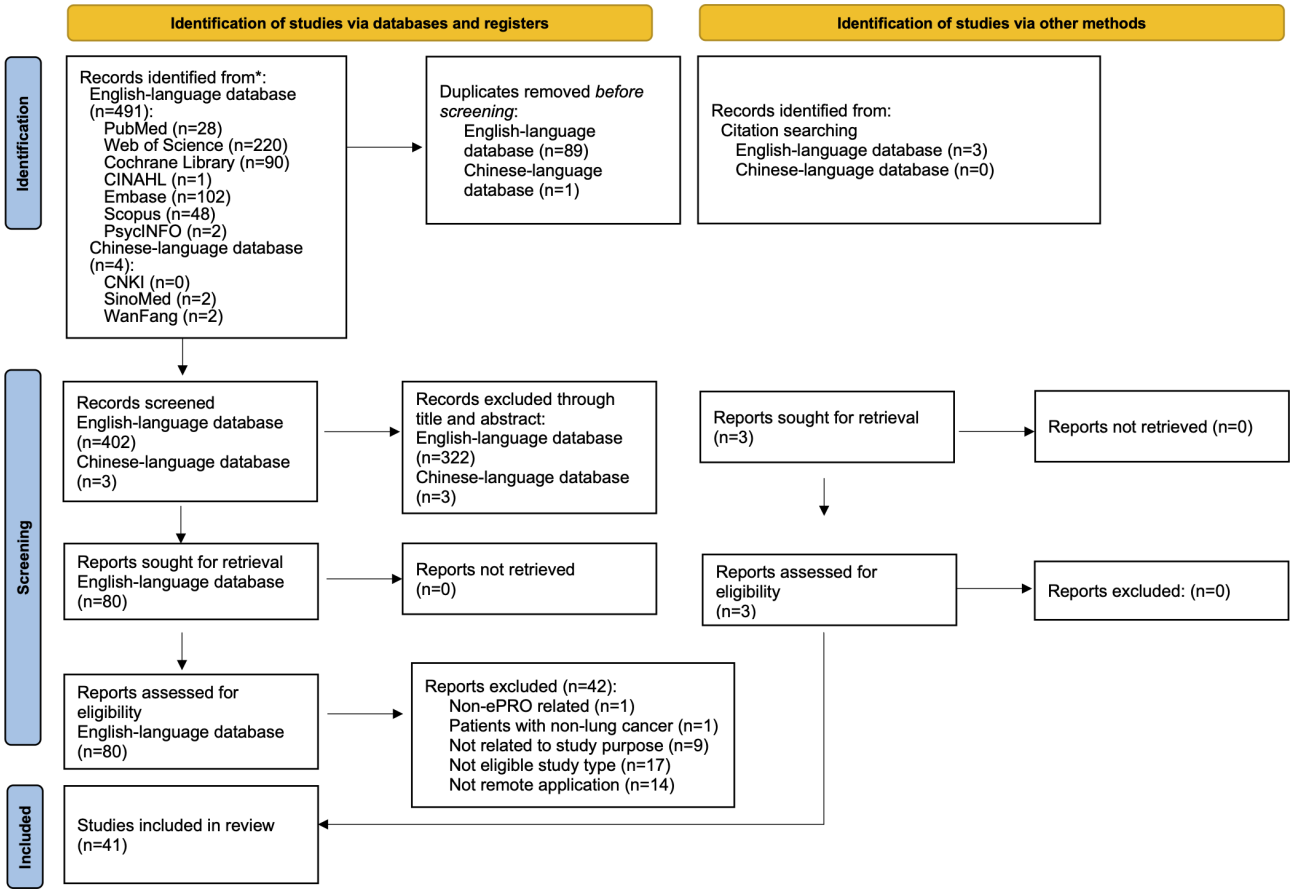

Study Characteristics

The 41 studies were published between 2007 and 2025. The majority of studies were conducted in the United States (n=19, 46%), followed by China (n=7, 17%), Denmark (n=3, 7%), the United Kingdom (n=3, 7%), and 1 (2%) study each from Australia, Germany, Spain, South Korea, Finland, France, Singapore, Italy, and Switzerland. The characteristics of the included studies are presented in Multimedia Appendix 2.

The majority of the included studies (n=19, 46%) used an RCT design, with 17 (41%) representing original RCTs [17,18,20-23,26,37-46] and 2 (5%) constituting secondary analyses derived from previously published RCTs [47,48]. The remaining publications encompassed prospective studies (n=8, 20%) [19,27,49-54], cohort studies (n=4, 10%) [55-57], mixed methods studies (n=4, 10%) [24,58-60], nonrandomized feasibility studies (n=4, 10%) [61-64], and qualitative studies (n=2, 5%) [65,66].

Among all the included studies, 5 (12%) enrolled health care providers as participants [49,50,58,60,63], whereas all others exclusively enrolled patients with lung cancer. Twelve (29%) studies focused exclusively on advanced-stage lung cancer [17,21,28,37,38,49,52,53,58,61,67], 7 (17%) included patients across all lung cancer stages [44,50,51,57,63,65,68], 3 (7%) targeted stages I to III [18,39,46], 1 (2%) concentrated on stages I to II [20], 1 (2%) focused on stage III [64], and 1 (2%) addressed stages II to IV [42]. Fifteen (37%) studies did not specify lung cancer stages [19,22-24,26,27,40,41,43,45,54-56,59,62].

Symptom monitoring across the included studies predominantly targeted respiratory complaints (cough, dyspnea, hemoptysis, voice changes, and chest pain), constitutional symptoms (fatigue, appetite loss, weight change, sleep disturbance, and cognitive difficulties), and gastrointestinal toxicities (nausea, vomiting, constipation, diarrhea, dysphagia, and sore mouth), as well as pain [17,18,20,21,39,42,46,51,55,61,62,64]. Psychological distress (anxiety, depression, and sadness) was less frequently assessed [21,42,45,46,55,64,67]. In addition, several research teams incorporated treatment-related adverse effects, such as side-effect bother, facial swelling, tumor growth sensation, and fever [17,42,45,61].

Application of Theoretical Framework

Three studies noted the application of established theoretical frameworks to guide the rSMS development or intervention implementation (Table 1).

XSL•FO
RenderX

Table . Application of theoretical frameworks or guidelines in the development and implementation of a remote digital symptom monitoring system (rSMS) for lung cancer care.

Author (Year)	Framework or guideline	Purpose	Application
Maguire et al (2015) [59]	Holistic framework to improve the uptake and impact of eHealth technologies [69]	Development of the rSMS	Prospective
Girgis et al (2022) [24]	RE-AIM[a] framework [70]	Guide the implementation	Retrospective
Mooney et al (2024) [26]	Chronic care model [71]	Development of the rSMS	Prospective

[a]RE-AIM: reach, effectiveness, adoption, implementation, and maintenance.

Development of the rSMS

The application of theoretical foundations varied across studies. Maguire et al [59] applied the holistic framework to improve the uptake and impact of eHealth technologies and to guide the development of their rSMS [69]. This framework includes 6 key foundational components, namely stakeholder engagement, iterative evaluation, intertwined development and implementation, transformation of health care organizations, persuasive design techniques, and advanced methods for impact assessment. The chronic care model [71], a patient-centered framework that emphasizes proactive, planned, and systematic interventions, was prospectively applied to design the rSMS intervention and evaluation strategies by Mooney et al [26].

Implementation of rSMS

Girgis et al [24] retrospectively applied the RE-AIM (reach, effectiveness, adoption, implementation, and maintenance) implementation science framework to guide and summarize successful intervention implementation, while concurrently identifying implementation barriers within this framework [70].

PROMs Used Within the rSMS and/or as Study Outcome Measures

A total of 32 PROMs applied across the included studies were categorized into two groups: (1) PROMs integrated into rSMS for monitoring purposes and (2) outcome measures used to evaluate the effectiveness of the rSMS intervention (Multimedia Appendix 3). Specifically, 19 PROMs were integrated within rSMS as components to monitor 3 domains—symptom assessment, HRQL, and psychological well-being. Thirteen PROMs were used as outcome measures to assess the effectiveness of the rSMS (Figure 2). Among all the 32 PROMs, three were used as both outcome measures and as components within the rSMS, including the Edmonton Symptom Assessment Scale, MD Anderson Symptom Inventory for Lung Cancer, and EuroQoL 5D-5L. Consistent with this review's focus on rSMS development and implementation within lung cancer care, detailed analysis is restricted to PROMs integrated within the rSMS, while those functioning solely as outcome measures are comprehensively cataloged in Multimedia Appendix 4.

XSL•FO
RenderX

Figure 2. Frequency of patient-reported outcome measures applied as components in remote digital symptom monitoring systems (rSMS) and as outcome measures in studies on lung cancer symptom monitoring. CTCAE: Common Terminology Criteria for Adverse Events; EORTC-QLQ-C30: European Organisation for Research and Treatment of Cancer Core Quality of Life Questionnaire; ESAS: Edmonton Symptom Assessment Scale; ESRA-C: Electronic Self Report Assessment for Cancer; FACIT-TS-S: Functional Assessment of Chronic Illness Therapy-Treatment Satisfaction-Patient Satisfaction; FACT-G: Functional Assessment of Cancer Therapy–General; FACT-L: Functional Assessment of Cancer Therapy–Lung; FLSI: Functional Assessment of Cancer Therapy–Lung Symptom Index; HADS: Hospital Anxiety and Depression Scale; MDASI-LC: MD Anderson Symptom Inventory for Lung Cancer; MSAS-SF: Memorial Symptom Assessment Scale–Short Form; PRO-CTCAE: Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events; PRO-CTCAE-k: Korean version of the Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events; PROMIS: Patient-Reported Outcomes Measurement Information System; PRSM: Patient-Reported Symptom Monitoring; QLQ-LC13: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire Lung Cancer Module; RSC: Rotterdam Symptom Checklist–Activity Subscale; SDS-15: Symptom Distress Scale–15; SIQOL: Single Item Quality of Life scale; STAI-Y: State-Trait Anxiety Inventory–Form Y; SUPPH-29: Strategies Used by Patients to Promote Health–29.

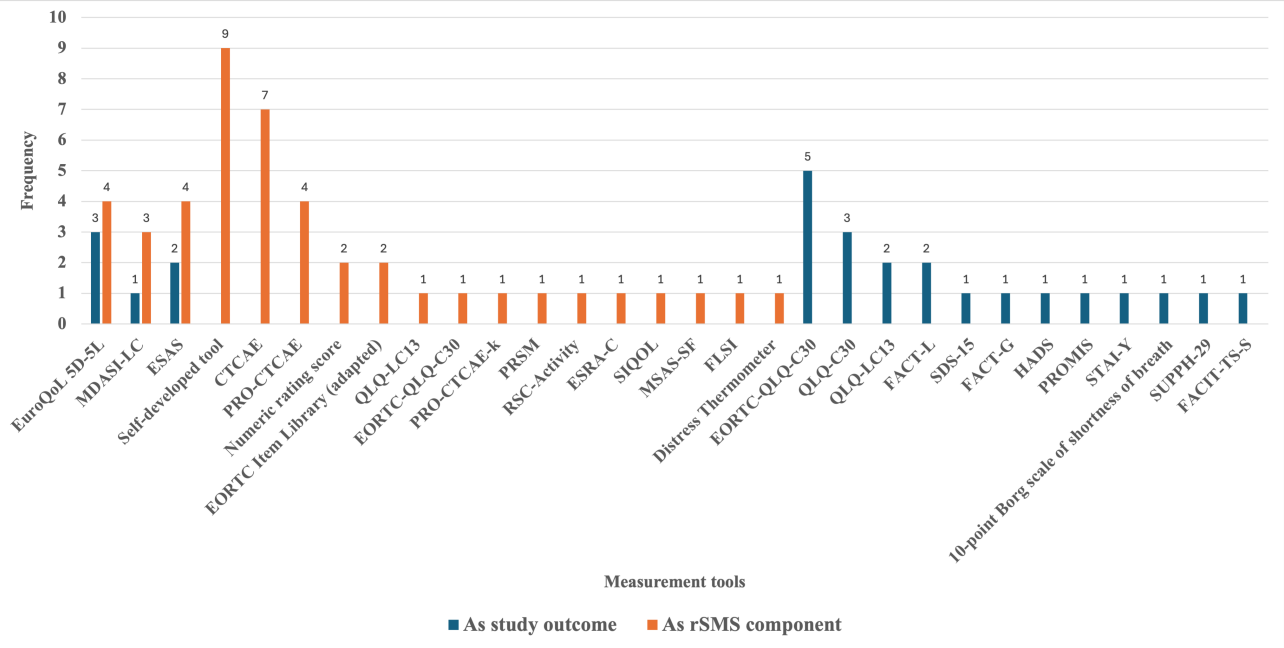

Applied Technology Platforms and the Functional Modules: Technology Platforms

The included studies used a variety of platforms for rSMS, including web-based systems, mobile phones, and wearable devices. Most studies (n=21, 51%) used web-based systems for rSMS implementation [9,17,19,22,24,27,28,43,44,51,52,54,56-58,61,62,65-67]. Mobile phones were the second most used digital platform (n=14, 34%), facilitating real-time responses and notifications between health care providers and patients [20,21,23,26,37-41,45,46,55,59,60]. Six (15%) additional studies implemented rSMS through personal electronic devices—encompassing personal computers, smartphones, or other specialized devices [18,49,50,53,63,64].

Functional Modules of the rSMS

The functional modules of rSMS were categorized into 4 key stages: data collection, data analysis, care response, and patient support (Figure 3).

XSL•FO
RenderX

Figure 3. Functional modules of the remote digital symptom monitoring system. PRO: patient-reported outcome.

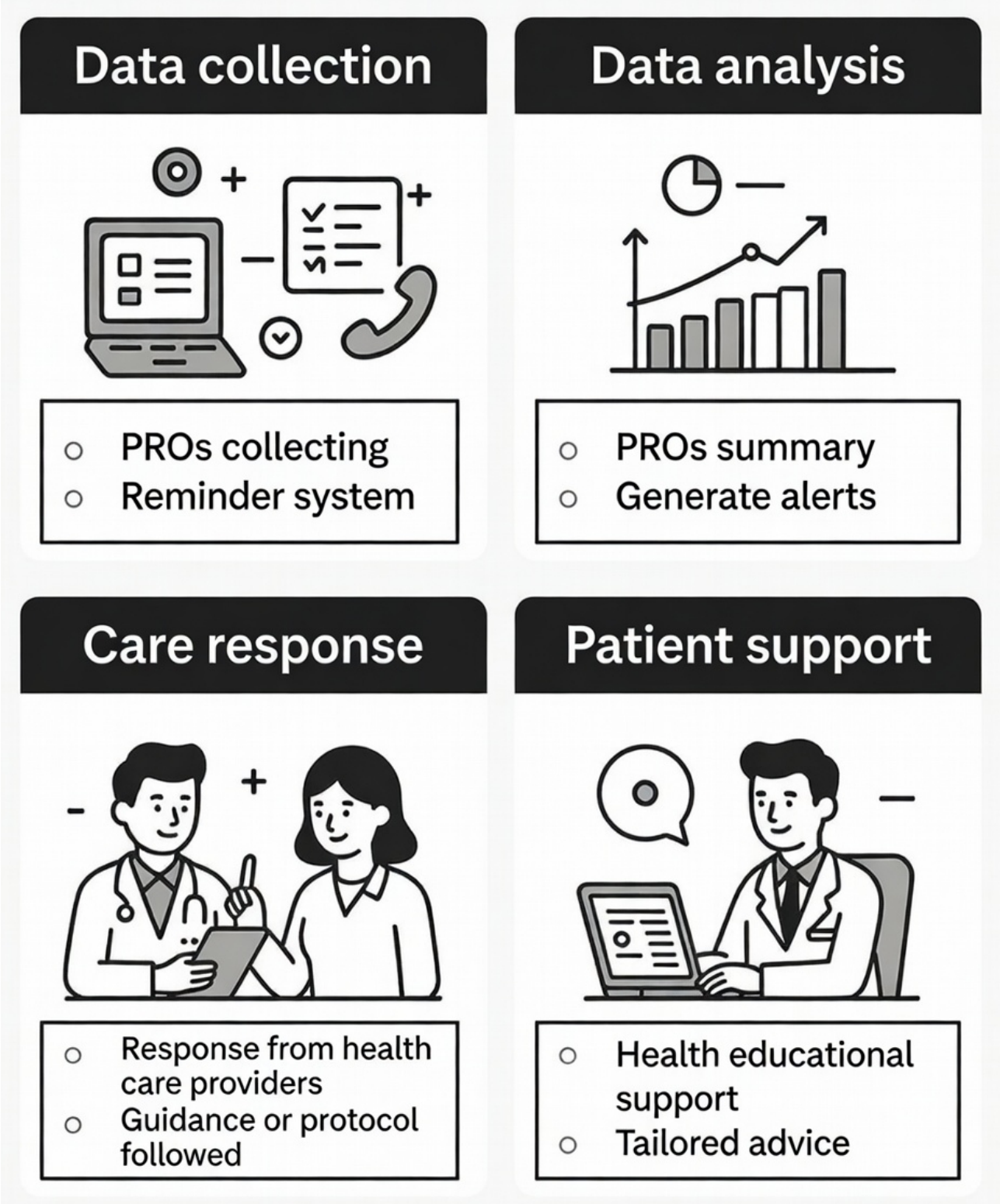

Data Collection

This stage involved symptom assessment tools to collect data, including symptom burden, symptom severity, HRQL, and other patient-reported outcomes. Additionally, 9 (22%) studies incorporated reminder systems to prompt timely patient assessments via emails, phone calls, or mobile app notifications [24,43,51,53,55,61,63,65,66].

Data Analysis

Following data collection, the data analysis stage was reported in all 41 studies. Of these, 10 (24%) studies reported systems generating summaries [22,23,38,40,42-44,60,62,72], which were provided to patients and/or health care providers, while others presented raw data or synthesized data manually. Additionally, 29 (71%) studies included notification or alert systems to inform the health care team when scores reached

XSL•FO
RenderX

predefined thresholds [17-20,24,27,28,39-42,45,46,49,50,52-55,57-61,63-67]. Typically, this module automatically conducted data analysis and generated alerts when predefined clinical criteria were met, which were then transmitted to the clinical teams—a mechanism that has been effectively implemented in current practice [17,42,49].

Care Response

The clinical response stage required mandatory responses from health care providers within a predefined time window, as reported in 11 (27%) studies [18,26,28,37,39,44-46,55,57,63]. A representative implementation by Dai et al's [18] study stipulated that when the rSMS identified a patient-reported symptom score that exceeded 4 points (indicating moderate-to-severe severity), surgeons who were involved in the patient's care were obligated to respond within 24 hours in accordance with the study protocol.

Patient Support

There were 16 (39%) studies that provided additional health education support based on the varying levels of symptom burden reported by patients [20,23,24,26,27,37,40,41,44,46,49,54,58,59,63,64]. For instance, the MyChristie-MyHealth program categorizes patients using a color-coded system based on symptom severity and provides tailored support advice [50]. Asymptomatic ("green") patients received a text message indicating that no action was required, mild symptoms ("blue") triggered an automated link to self-care advice, moderate symptoms ("orange") prompted patients to seek support from their oncology team within 1 week, and patients with severe symptoms ("red") received a message recommending urgent medical advice within 24 hours.

Experiences and Perceptions of Patients and Health Care Professionals

User experiences of both patients and health care providers were evaluated through semistructured interviews across 5 (12%) studies [45,59,60,63,66]. Questionnaire-based surveys were also used to measure satisfaction levels among patients and health care providers [62,63]. Additionally, qualitative interviews were conducted to examine the feasibility of interventions and to identify facilitating and hindering factors from the patients' perspective [65]. The findings from the content analysis are detailed in Multimedia Appendix 5 and visually synthesized in Figure 4, which highlights the key benefits and concerns expressed by patients and health care professionals regarding rSMS use.

Figure 4. Perceived benefits and concerns of the remote symptom monitoring system (rSMS) application among patients and health care professionals.

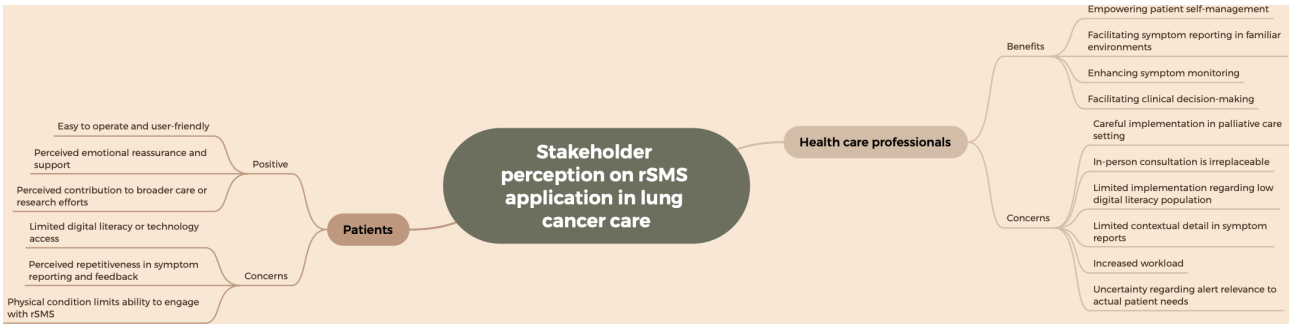

Perceptions of rSMS Applications Among Health Care Professionals

Health care providers expressed diverse benefits and barriers of rSMS. First, some health care providers in palliative care settings perceived that proactive monitoring could empower certain patients by granting them greater control over their conditions and enhancing their disease management abilities, as the system provided positive feedback that motivated patients and guided their next steps [60]. However, the same study included no patient outcome data, and most health care providers expressed doubts about implementing such interventions in palliative care settings. This aligns with the concerns voiced by clinicians in other research [63]. Additionally, the remote reporting feature allowed patients to report symptoms in familiar surroundings, potentially fostering a more relaxed reporting experience compared to hospital settings [66]. Second, for health care providers, patients' previsit self-reports enabled the care teams to prepare more effectively and facilitated early identification of symptom changes [59]. Regular monitoring of symptom fluctuations also supported disease progression tracking, aiding timely adjustments to care strategies and clinical decision-making [63,66].

Concerns about rSMS implementation were also reported. Some physicians emphasized that rSMS should supplement rather than replace clinical judgment [60,66]. Similarly, other clinicians highlighted that traditional in-person consultations and hands-on experience should not be supplanted by remote reporting, as reliance solely on patient self-reports without clinical observation poses significant challenges [60,66]. Additional concerns included the selective applicability of rSMS, particularly regarding older patients' unfamiliarity with or resistance to electronic systems, as well as its limited benefits for palliative care groups [60]. Regarding rSMS itself, some health care providers identified limitations in its ability to support rigorous clinical decisions, noting that symptom reports often lack contextual details such as frequency or type, necessitating further patient interaction [66]. Additionally, some health care providers reported that rSMS added to their existing workload, making timely responses challenging [59,66]. Furthermore, health care providers expressed concerns about alerts, questioning whether all alerts are necessary and effectively reflect patients' actual symptom burdens, while also worrying that excessive alerts might increase workloads [63].

XSL•FO
RenderX

Perceptions of rSMS Applications Among Patients

Patients reported that rSMS was generally user-friendly and easy to operate, enabling them to conveniently report their symptoms and functional status [59,65]. Timely response systems and tailored health advice were considered highly helpful [45,59,65]. Some patients felt that rSMS provided a sense of safety, encouragement, and support, making them feel attended to by health care providers outside clinical settings [45,59,65]. Additionally, patients reported feeling that they were not only receiving help during the intervention, but also contributing to helping others [45].

However, some patients expressed concerns about rSMS. System-related barriers included hardware issues (eg, unstable access in rural areas), a preference for paper-based surveys, and unfamiliarity with electronic systems [9,65]. Some perceived symptom assessments and advice as redundant with existing knowledge, reducing motivation to complete surveys [59,65]. Others questioned rSMS efficacy, describing it as "self-limited" [59,65]. High symptom burden was also cited as an engagement barrier [65].

Clinical Insights

The implementation of rSMS among patients with lung cancer has demonstrated substantial benefits, including symptom alleviation, improvements in HRQL, enhanced communication between patients and health care professionals, strengthened patient-centered care experiences, and broader health system advantages. Nevertheless, some studies have reported nonsignificant effects on clinical outcomes, underscoring the variability of evidence (Figure 5).

Figure 5. Key clinical domains impacted by remote symptom monitoring systems (rSMS).

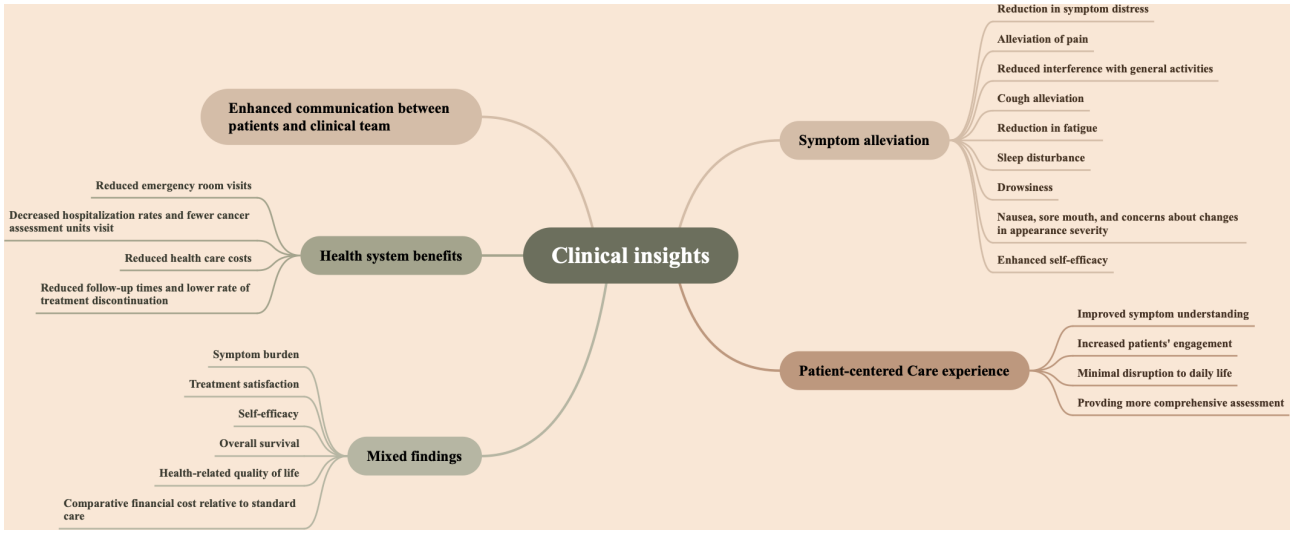

Symptom-Related Outcome

The rSMS intervention demonstrated significant reductions in overall symptoms among patients with lung cancer [18,20,22,26,39,41,46,52]. Symptom distress emerged as the most frequently documented alleviation and was evident across 4 (10%) studies [22,41,46,59]. This was followed by a significant decrease in patient-reported pain, noted in 2 RCTs [20,41]. Additionally, rSMS was reported to notably reduce interference with general activities, work, and walking, as observed in 2 RCTs [18,20]. Significant clinical improvements were reported in cough [20], fatigue [20,41], sleep disturbance [20,41], and drowsiness [59]. One (2%) study extended these findings to additional symptom categories—including nausea, sore mouth, and concerns about changes in appearance severity [41]. Beyond physical symptom management, 1 (2%) study observed enhanced patient self-efficacy, attributing these gains to rSMS-facilitated psychological distress reduction and improved adaptive functioning [59]. Conversely, another research team reported no significant improvements in symptom burden, treatment satisfaction, or self-efficacy [21]. Similarly, no significant rSMS-related benefits in overall survival were found in a separate study [17].

Health-Related Quality of Life

Five (12%) studies observed improvements in HRQL among patients using a symptom monitoring system [17,40,42,43,46]. However, 4 (10%) studies revealed no statistically significant improvements in HRQL among rSMS intervention groups [21,22,37,63].

Communication With Health Care Providers

Five (12%) studies reported that the rSMS application enhanced communication or improved discussions between patients and health care providers [22,49,50,52,57]. All these studies used questionnaire-based investigations to assess patients' perceptions. Additionally, a pooled analysis suggested that improved communication may constitute one of the potential reasons for the success of the intervention program [67]. For health care providers, rSMS primarily supported clinical workflows and optimized patient communication. This includes reducing time spent on follow-up tracking [39], enhancing work efficiency [39,40,58], and increasing confidence [24] in symptom monitoring and management.

Patient-Centered Care Experience

Several studies have demonstrated that rSMS contributes to modern patient-centered care through various approaches. Three

XSL•FO
RenderX

(7%) studies reported improved symptom understanding using a patient-reported experience measure in rSMS for lung cancer care [44,49,50]. Additionally, this intervention enhanced patient-care team interactions, increasing patients' engagement in their own care [50]. Furthermore, 2 RCTs reported that rSMS functioned effectively without disrupting daily lives, yielding high patient satisfaction in lung cancer care [19,39]. Similarly, rSMS enhanced patients' perceived benefit, making them feel their cancer care significantly improved [19,53]. Implementation also provided significant clinical benefits for specific lung cancer groups, such as patients receiving immune checkpoint inhibitors, who were reported to have a more comprehensive assessment, including toxicity grading and long-term surveillance of symptoms [40].

Health System Benefits

Reduced emergency room visits were reported in 2 (5%) studies [40,43]. Similarly, decreased hospitalization rates and fewer visits to cancer assessment units were reported in 2 (5%) other studies [56,73]. Patt et al further identified that rSMS was associated with a lower total cost of care [56]. One RCT study reported that enhanced treatment experiences and automated platforms contributed to reduced follow-up times and a lower rate of treatment discontinuation [40]. Additionally, financial costs were compared between rSMS implementation and the standard of care, with no statistically significant differences observed [45]. In the PROStep feasibility study [38], which integrated both self-reported surveys and passive activity monitoring for patients with lung and gastrointestinal cancers—whereas most other platforms in our review relied on active reporting only—the intervention group reported a high perceived understanding of symptoms and functional status by their oncology team; this did not differ from the control group [38].

Discussion

Principal Findings

This scoping review provides an overview of how rSMS has been applied in lung cancer care, with particular attention to the integration of PROMs and functional modules. The synthesis of existing studies indicates that rSMS interventions show encouraging potential to improve symptom surveillance and support timely clinical responses; however, the current evidence remains fragmented and methodologically limited. Notably, few studies have adopted theory-driven approaches in the design, implementation, or evaluation of rSMS, leaving important conceptual and practical gaps. In addition, variability in patient and health care provider engagement highlights ongoing challenges related to usability and comprehension. These findings set the stage for a more detailed discussion of the opportunities and challenges in advancing rSMS, as well as directions for future research.

Only 3 (7%) studies incorporated theory-driven approaches in designing the intervention program or evaluating the outcomes [24,26,59]. Existing evidence from implementation science underscores that theory-driven approaches enhance methodological transparency, clarify intervention mechanisms, and promote sustainability [74,75]. The current limited application of theoretical frameworks may contribute to inconsistencies in intervention implementation and hinder the comparability and validity of outcome measurements across studies. Future studies should integrate theoretical guidance at both the design and evaluation stages to maximize the effectiveness and sustainability of interventions.

To ensure the reliability of patient-reported data and the validity of study outcomes, rSMS should incorporate PROMs that have undergone rigorous psychometric validation. However, as illustrated in Figure 2, the majority of PROMs used in rSMS studies were self-developed by research teams. Of the 9 self-developed tools identified, only 4 (10%) studies explicitly reported having conducted formal validation procedures [26,42,55]. The limited reporting of validation efforts may affect the interpretability and generalizability of findings derived from these tools, underscoring the need for greater emphasis on validated PROMs in future rSMS designs.

This review further delineates four key functional modules characterizing current rSMS applications in lung cancer care, including data collection, data analysis, the response system, and patient support. While most platforms offer real-time alert capabilities, more than half report limited clinical action in response to these alerts. This lack of action is primarily attributed to barriers such as competing clinical demands and unclear delineation of responsibilities [66]. This highlights a critical need for integration strategies that link real-time symptom alerts to actionable care pathways and effective feedback mechanisms, ensuring digital monitoring enables timely intervention rather than functioning as a passive data repository.

The qualitative findings identified in this review remain limited, with reported outcomes demonstrating variability [45,59,60,63,65,66]. While health care providers value rSMS for symptom monitoring, they noted limitations, including reduced effectiveness among patients with limited digital literacy—particularly older patients with lung cancer—and perceptions of data subjectivity, possibly due to misunderstandings about the system [60,66]. Patients acknowledge rSMS benefits but express notable concerns regarding the rSMS system's trustworthiness [65]. A deeper understanding of these perceptions and experiences is needed to inform strategies for patient engagement and health care provider integration. Importantly, current evidence predominantly originates from Western high-income settings with advanced medical infrastructure, highlighting the need for research in diverse resource environments to comprehensively understand rSMS implementation.

Clinical outcomes associated with rSMS interventions were found to be heterogeneous. Several trials reported meaningful benefits, including reductions in symptom distress and improvements in pain, fatigue, or sleep disturbance [18,20,22,26,39,41,46,52]. In contrast, other studies observed limited or no impact on broader end points such as global quality of life, overall survival, or treatment satisfaction [21,22,37,63]. These discrepancies may be explained by heterogeneity in rSMS design and implementation, including differences in clinical context, monitoring frequency, patient reporting requirements,

XSL•FO
RenderX

and the extent of clinician engagement. Context-related factors may also contribute to this variability, with effectiveness potentially differing across specific care settings. For example, prior work has highlighted that the use of rSMS in palliative care requires particular caution to avoid imposing additional burdens on patients receiving end-of-life care [60].

Future Directions

In the era of precision medicine, rSMS integration can be optimized through person-centered interventions tailored to the diverse disease trajectories of patients with lung cancer. Such a tailored approach could enhance symptom assessment and management efficacy. For example, identifying variations in symptom burden can enable the integration of personalized education and technical support into the intervention program. The color-coded system identified in this review exemplifies a tailored tool for assessing these variations, thereby optimizing the application of rSMS [50].

Improving patients' understanding of rSMS implementation could facilitate its successful adoption. While the current exploratory evidence is limited, studies indicate that patients frequently hesitate to seek medical assistance despite receiving alerts [49,50]. The observed reluctance to seek medical help may reflect a limited comprehension of the intervention's objectives and functionality. As patients' understanding of rSMS can influence their adherence and engagement, future studies should prioritize preintervention education on rSMS features and clinical aims to reinforce engagement.

Limitations

This review has several limitations. First, the included studies were predominantly published in English. Although Chinese literature underwent initial screening, no studies met the inclusion criteria. Second, consistent with the guidelines of scoping review methodology, no formal quality appraisal of the included studies was performed, and therefore, the methodological quality of the evidence could not be ascertained. Third, due to the considerable heterogeneity in study designs, interventions, and outcome measures, it was challenging to draw definitive conclusions regarding clinical effectiveness.

Conclusions

This review advances the field by offering a comprehensive synthesis of the landscape of rSMS in lung cancer care. Unlike previous reviews that focused primarily on clinical outcomes or technical feasibility, this review adds to prior work by summarizing the theoretical foundations guiding rSMS, integrating functional modules, and distinguishing PROMs used as embedded system components from those reported as outcome measures. Together, these contributions clarify the current landscape and highlight methodological gaps. This is innovative because it provides a theoretical lens for comparing rSMS, not only their outcomes. To ensure long-term sustainability and clinical relevance, future research should adopt theory-driven approaches, refine functional integration, and incorporate multistakeholder perspectives to develop rSMS platforms that are patient-centered, sustainable, and adaptable across diverse clinical contexts while minimizing clinician burden.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT (Generative AI Delegation Taxonomy; 2025), the following tasks were delegated to GAI tools under full human supervision: visualization. The GAI tool used was Grok. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

The research work described in this paper was conducted in the Jockey Club Science, Technology, Engineering, and Mathematics Lab of Digital Oncology Care Enhancement funded by the Hong Kong Jockey Club Charities Trust. The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

All data generated or analyzed during this study are included in this published article (and the multimedia appendices).

Authors' Contributions

HH drafted the manuscript. NT and HA-O provided suggestions and revisions regarding the logic, presentation, and data extraction. SL and TM assisted with literature screening and data extraction. JY contributed to the conceptualization, revision, and proofreading of the draft.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of search strategy used across all databases.
[DOCX File, 26 KB - jmir_v28i1e83666_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Characteristics of included studies.
[DOCX File, 136 KB - jmir_v28i1e83666_app2.docx]

Multimedia Appendix 3
Patient-reported outcome measures serve as both components and outcome measures within remote digital symptom monitoring system.
[DOCX File, 93 KB - jmir_v28i1e83666_app3.docx]

Multimedia Appendix 4
Patient-reported outcome measures serve as outcome indicators.
[PNG File, 298 KB - jmir_v28i1e83666_app4.png]

Multimedia Appendix 5
Content analysis of qualitative studies.
[DOCX File, 43 KB - jmir_v28i1e83666_app5.docx]

Checklist 1
PRISMA-ScR checklist
[PDF File, 86 KB - jmir_v28i1e83666_app6.pdf]

Checklist 2
PRISMA-S checklist
[PDF File, 85 KB - jmir_v28i1e83666_app7.pdf]

Checklist 3
PRISMA checklist
[PDF File, 312 KB - jmir_v28i1e83666_app8.pdf]

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA A Cancer J Clinicians 2024 May;74(3):229-263. [doi: 10.3322/caac.21834]
2. Au PCM, Lee AWM, Lee VHF, Wong ICK, Hui RYM, Cheung CL. The trends in lung cancer prevalence, incidence, and survival in Hong Kong over the past two decades (2002–2021): a population-based study. Lancet Reg Health West Pac 2024 Apr;45:101030. [doi: 10.1016/j.lanwpc.2024.101030]
3. Mao S, Liu L, Miao C, et al. Electronic symptom monitoring for home-based palliative care: a systematic review. Palliat Med 2024 Sep;38(8):790-805. [doi: 10.1177/02692163241257578] [Medline: 38835179]
4. Lizán L, Pérez-Carbonell L, Comellas M. Additional value of patient-reported symptom monitoring in cancer care: a systematic review of the literature. Cancers (Basel) 2021 Sep 15;13(18):4615. [doi: 10.3390/cancers13184615] [Medline: 34572842]
5. van den Hurk CJG, Mols F, Eicher M, et al. A narrative review on the collection and use of electronic patient-reported outcomes in cancer survivorship care with emphasis on symptom monitoring. Curr Oncol 2022 Jun 17;29(6):4370-4385. [doi: 10.3390/curroncol29060349] [Medline: 35735458]
6. U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Biologics Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Devices and Radiological Health. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. Health Qual Life Outcomes 2006 Dec;4(1):79. [doi: 10.1186/1477-7525-4-79]
7. Antunes B, Harding R, Higginson IJ, EUROIMPACT. Implementing patient-reported outcome measures in palliative care clinical practice: a systematic review of facilitators and barriers. Palliat Med 2014 Feb;28(2):158-175. [doi: 10.1177/0269216313491619] [Medline: 23801463]
8. Churruca K, Pomare C, Ellis LA, et al. Patient-reported outcome measures (PROMs): a review of generic and condition-specific measures and a discussion of trends and issues. Health Expect 2021 Aug;24(4):1015-1024. [doi: 10.1111/hex.13254] [Medline: 33949755]
9. Brunelli C, Alfieri S, Zito E, et al. Patient voices: multimethod study on the feasibility of implementing electronic patient-reported outcome measures in a comprehensive cancer center. JMIR Cancer 2025 Jan 22;11:e56625. [doi: 10.2196/56625] [Medline: 39842002]

XSL•FO
RenderX

10. Ivatury SJ, Hazard-Jenkins HW, Brooks GA, McCleary NJ, Wong SL, Schrag D. Translation of patient-reported outcomes in oncology clinical trials to everyday practice. Ann Surg Oncol 2020 Jan;27(1):65-72. [doi: 10.1245/s10434-019-07749-2] [Medline: 31452053]
11. Lavallee DC, LeRouge C, Austin EJ, et al. Guidelines for integrating electronic patient-reported outcomes into health systems—final report. : Agency for Healthcare Research and Quality; 2021 URL: https://digital.ahrq.gov/sites/default/files/docs/citation/r01hs023785-lavallee-final-report-2021.pdf [accessed 2026-03-05]
12. Principles for selecting, developing, modifying, and adapting patient-reported outcome instruments for use in medical device evaluation: guidance for industry and food and drug administration staff, and other stakeholders. : US Food and Drug Administration; 2022 URL: https://www.fda.gov/media/141565/download [accessed 2026-03-05]
13. Kaneyasu T, Hoshino E, Naito M, et al. How to select and understand guidelines for patient-reported outcomes: a scoping review of existing guidance. BMC Health Serv Res 2024 Mar 6;24:334. [doi: 10.1186/s12913-024-10707-8]
14. García Abejas A, Serra Trullás A, Sobral MA, Canelas D, Leite Costa F, Salvador Verges À. Improving the understanding and managing of the quality of life of patients with lung cancer with electronic patient-reported outcome measures: scoping review. J Med Internet Res 2023 May 3;25(1):e46259. [doi: 10.2196/46259] [Medline: 37021695]
15. Salmani H, Nasiri S, Alemrajabi M, Ahmadi M. Advancing patient-centered cancer care: a systematic review of electronic patient-reported outcome measures. Front Rehabil Sci 2024;5:1427712. [doi: 10.3389/fresc.2024.1427712] [Medline: 39386354]
16. Cho Y, Zhang H, Harris MR, Gong Y, Smith EL, Jiang Y. Acceptance and use of home-based electronic symptom self-reporting systems in patients with cancer: systematic review. J Med Internet Res 2021 Mar 12;23(3):e24638. [doi: 10.2196/24638] [Medline: 33709929]
17. Friis RB, Pappot H, Hjollund NH, et al. Remote symptom monitoring of patients with advanced lung cancer (the ProWide study): a randomized controlled trial. JCO Oncol Pract 2025 Jun;21(6):813-823. [doi: 10.1200/OP-24-00562] [Medline: 39657078]
18. Dai W, Wang Y, Liao J, et al. Electronic patient-reported outcome-based symptom management versus usual care after lung cancer surgery: long-term results of a multicenter, randomized, controlled trial. J Clin Oncol 2024;42:2126-2131. [doi: 10.1200/JCO.23.01854]
19. Ma TM, Yang T, Philipson R, Kishan AU, Lee P, Raldow AC. Web-based symptom monitoring with patient-reported outcomes during definitive radiation therapy with chemotherapy (SYMPATHY): a prospective single-center phase 1 study. Adv Radiat Oncol 2023;8(3):101073. [doi: 10.1016/j.adro.2022.101073] [Medline: 37273926]
20. Lv C, Lu F, Zhou X, et al. Efficacy of a smartphone application assisting home-based rehabilitation and symptom management for patients with lung cancer undergoing video-assisted thoracoscopic lobectomy: a prospective, single-blinded, randomised control trial (POPPER study). Int J Surg 2025 Jan 1;111(1):597-608. [doi: 10.1097/JS9.0000000000001845] [Medline: 38905505]
21. Yount SE, Rothrock N, Bass M, et al. A randomized trial of weekly symptom telemonitoring in advanced lung cancer. J Pain Symptom Manage 2014 Jun;47(6):973-989. [doi: 10.1016/j.jpainsymman.2013.07.013] [Medline: 24210705]
22. Strasser F, Blum D, von Moos R, et al. The effect of real-time electronic monitoring of patient-reported symptoms and clinical syndromes in outpatient workflow of medical oncologists: E-MOSAIC, a multicenter cluster-randomized phase III study (SAKK 95/06). Ann Oncol 2016 Feb;27(2):324-332. [doi: 10.1093/annonc/mdv576] [Medline: 26646758]
23. Lee M, Kang D, Kang E, et al. Efficacy of the PRO-CTCAE mobile application for improving patient participation in symptom management during cancer treatment: a randomized controlled trial. Support Care Cancer 2023 May 6;31(6):321. [doi: 10.1007/s00520-023-07779-3] [Medline: 37148373]
24. Girgis A, Bamgboje-Ayodele A, Rincones O, et al. Stepping into the real world: a mixed-methods evaluation of the implementation of electronic patient reported outcomes in routine lung cancer care. J Patient Rep Outcomes 2022 Jun 20;6(1):70. [doi: 10.1186/s41687-022-00475-6] [Medline: 35723827]
25. Pietanza MC, Basch EM, Lash A, et al. Harnessing technology to improve clinical trials: study of real-time informatics to collect data, toxicities, image response assessments, and patient-reported outcomes in a phase II clinical trial. J Clin Oncol 2013 Jun 1;31(16):2004-2009. [doi: 10.1200/JCO.2012.45.8117] [Medline: 23630218]
26. Mooney K, Gullatte M, Iacob E, et al. Essential components of an electronic patient-reported symptom monitoring and management system: a randomized clinical trial. JAMA Netw Open 2024 Sep 3;7(9):e2433153. [doi: 10.1001/jamanetworkopen.2024.33153] [Medline: 39269704]
27. Wujcik D, Dudley WN, Dudley M, Gupta V, Brant J. Electronic patient symptom management program to support patients receiving cancer treatment at home during the COVID-19 pandemic. Value Health 2022 Jun;25(6):931-936. [doi: 10.1016/j.jval.2022.01.023] [Medline: 35339378]
28. Chih MY, DuBenske LL, Hawkins RP, et al. Communicating advanced cancer patients' symptoms via the Internet: a pooled analysis of two randomized trials examining caregiver preparedness, physical burden, and negative mood. Palliat Med 2013 Jun;27(6):533-543. [doi: 10.1177/0269216312457213] [Medline: 22988042]
29. Rothman AJ. Capitalizing on opportunities to refine health behavior theories. Health Educ Behav 2009 Oct;36(5 Suppl):150S-155S. [doi: 10.1177/1090198109340514] [Medline: 19830884]

XSL•FO
RenderX

30. Michie S, Johnston M, Francis J, Hardeman W, Eccles M. From theory to intervention: mapping theoretically derived behavioural determinants to behaviour change techniques. Appl Psychol 2008 Oct;57(4):660-680. [doi: 10.1111/j.1464-0597.2008.00341.x]
31. Davis R, Campbell R, Hildon Z, Hobbs L, Michie S. Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review. Health Psychol Rev 2015;9(3):323-344. [doi: 10.1080/17437199.2014.941722] [Medline: 25104107]
32. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020 Oct;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
33. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
34. Current digital symptom monitoring implementation among lung cancer patients: a scoping review. Open Science Framework. URL: https://osf.io/9637t/overview [accessed 2026-03-13]
35. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
36. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
37. Yang GM, Ke Y, Ng XH, Neo PSH, Cheung YB. Proactive symptom monitoring to initiate timely palliative care for patients with advanced cancer: a randomized controlled trial. Support Care Cancer 2025 Mar 5;33(3):249. [doi: 10.1007/s00520-025-09311-1] [Medline: 40038129]
38. Parikh RB, Schriver E, Ferrell WJ, et al. Remote patient-reported outcomes and activity monitoring to improve patient-clinician communication regarding symptoms and functional status: a randomized controlled trial. JCO Oncol Pract 2023 Dec;19(12):1143-1151. [doi: 10.1200/OP.23.00048] [Medline: 37816198]
39. Dai W, Feng W, Zhang Y, et al. Patient-reported outcome-based symptom management versus usual care after lung cancer surgery: a multicenter randomized controlled trial. J Clin Oncol 2022 Mar 20;40(9):988-996. [doi: 10.1200/JCO.21.01344]
40. Zhang L, Zhang X, Shen L, Zhu D, Ma S, Cong L. Efficiency of electronic health record assessment of patient-reported outcomes after cancer immunotherapy: a randomized clinical trial. JAMA Netw Open 2022 Mar 1;5(3):e224427. [doi: 10.1001/jamanetworkopen.2022.4427] [Medline: 35357459]
41. Mooney KH, Beck SL, Wong B, et al. Automated home monitoring and management of patient-reported symptoms during chemotherapy: results of the symptom care at home RCT. Cancer Med 2017 Mar;6(3):537-546. [doi: 10.1002/cam4.1002] [Medline: 28135050]
42. Denis F, Lethrosne C, Pourel N, et al. Randomized trial comparing a web-mediated follow-up with routine surveillance in lung cancer patients. J Natl Cancer Inst 2017 Sep 1;109(9). [doi: 10.1093/jnci/djx029] [Medline: 28423407]
43. Basch E, Deal AM, Kris MG, et al. Symptom monitoring with patient-reported outcomes during routine cancer treatment: a randomized controlled trial. J Clin Oncol 2016 Feb 20;34(6):557-565. [doi: 10.1200/JCO.2015.63.0830]
44. Berry DL, Hong F, Halpenny B, et al. The electronic self report assessment and intervention for cancer: promoting patient verbal reporting of symptom and quality of life issues in a randomized controlled trial. BMC Cancer 2014 Jul 12;14(1):513. [doi: 10.1186/1471-2407-14-513] [Medline: 25014995]
45. Pongiglione B, Cucciniello M, Petracca F, et al. A mobile supportive care app for patients with metastatic lung cancer: the Lung Cancer App (LuCApp) randomized controlled trial. Support Care Cancer 2025 Jun 30;33(7):641. [doi: 10.1007/s00520-025-09682-5] [Medline: 40586961]
46. Jing P, Liang Y, Tan Z, et al. Physiological and psychological symptom management based on electronic patient-reported outcomes: the TD-WELLBEING randomized clinical trial. Br J Cancer 2025 Oct;133(7):937-944. [doi: 10.1038/s41416-025-03110-5] [Medline: 40775446]
47. Dubenske LL, Chih MY, Dinauer S, Gustafson DH, Cleary JF. Development and implementation of a clinician reporting system for advanced stage cancer: initial lessons learned. J Am Med Inform Assoc 2008;15(5):679-686. [doi: 10.1197/jamia.M2532] [Medline: 18579837]
48. DuBenske LL, Gustafson DH, Namkoong K, et al. CHESS improves cancer caregivers' burden and mood: results of an eHealth RCT. Health Psychol 2014 Oct;33(10):1261-1272. [doi: 10.1037/a0034216] [Medline: 24245838]
49. Arriola E, Jaal J, Edvardsen A, et al. Feasibility and user experience of digital patient monitoring for real-world patients with lung or breast cancer. Oncologist 2024 Apr 4;29(4):e561-e569. [doi: 10.1093/oncolo/oyad289] [Medline: 38007400]
50. Payne A, Horne A, Bayman N, et al. Patient and clinician-reported experiences of using electronic patient reported outcome measures (ePROMs) as part of routine cancer care. J Patient Rep Outcomes 2023 May 4;7(1):42. [doi: 10.1186/s41687-023-00544-4] [Medline: 37140730]
51. Cheng X, Yang Y, Shentu Y, et al. Remote monitoring of patient recovery following lung cancer surgery: a messenger application approach. J Thorac Dis 2021 Feb;13(2):1162-1171. [doi: 10.21037/jtd-21-27] [Medline: 33717589]
52. Zylla DM, Gilmore GE, Steele GL, et al. Collection of electronic patient-reported symptoms in patients with advanced cancer using Epic MyChart surveys. Support Care Cancer 2020 Jul;28(7):3153-3163. [doi: 10.1007/s00520-019-05109-0] [Medline: 31701269]

XSL•FO
RenderX

53. Iivanainen S, Alanko T, Vihinen P, et al. Follow-up of cancer patients receiving anti-PD-(L) 1 therapy using an electronic patient-reported outcomes tool (KISS): prospective feasibility cohort study. JMIR Form Res 2020 Oct 28;4(10):e17898. [doi: 10.2196/17898] [Medline: 33112242]
54. Sewell M, Boerner T, Harrington C, et al. Remote symptom monitoring in thoracic surgery patients after discharge. Ann Surg 2025 Jun 1;281(6):1063-1069. [doi: 10.1097/SLA.0000000000006619] [Medline: 39774469]
55. Yu H, Lei C, Wei X, et al. Electronic symptom monitoring after lung cancer surgery: establishing a core set of patient-reported outcomes for surgical oncology care in a longitudinal cohort study. Int J Surg 2024 Oct 1;110(10):6591-6600. [doi: 10.1097/JS9.0000000000001855] [Medline: 38896873]
56. Patt DA, Patel AM, Bhardwaj A, et al. Impact of remote symptom monitoring with electronic patient-reported outcomes on hospitalization, survival, and cost in community oncology practice: the Texas two-step study. JCO Clin Cancer Inform 2023 Sep;7(7):e2300182. [doi: 10.1200/CCI.23.00182] [Medline: 37897263]
57. Mody GN, Stover AM, Wang M, et al. Electronic patient-reported outcomes monitoring during lung cancer chemotherapy: a nested cohort within the PRO-TECT pragmatic trial (AFT-39). Lung Cancer 2021 Dec;162:1-8. [doi: 10.1016/j.lungcan.2021.09.020] [Medline: 34634754]
58. Schmalz O, Jacob C, Ammann J, et al. Digital monitoring and management of patients with advanced or metastatic non-small cell lung cancer treated with cancer immunotherapy and its impact on quality of clinical care: interview and survey study among health care professionals and patients. J Med Internet Res 2020 Dec 21;22(12):e18655. [doi: 10.2196/18655] [Medline: 33346738]
59. Maguire R, Ream E, Richardson A, et al. Development of a novel remote patient monitoring system: the advanced symptom management system for radiotherapy to improve the symptom experience of patients with lung cancer receiving radiotherapy. Cancer Nurs 2015;38(2):E37-E47. [doi: 10.1097/NCC.0000000000000150] [Medline: 24836956]
60. Cox A, Illsley M, Knibb W, et al. The acceptability of e-technology to monitor and assess patient symptoms following palliative radiotherapy for lung cancer. Palliat Med 2011 Oct;25(7):675-681. [doi: 10.1177/0269216311399489] [Medline: 21474620]
61. Friis RB, Hjollund NH, Mejdahl CT, Pappot H, Skuladottir H. Electronic symptom monitoring in patients with metastatic lung cancer: a feasibility study. BMJ Open 2020 Jun 17;10(6):e035673. [doi: 10.1136/bmjopen-2019-035673] [Medline: 32554725]
62. Basch E, Iasonos A, Barz A, et al. Long-term toxicity monitoring via electronic patient-reported outcomes in patients receiving chemotherapy. J Clin Oncol 2007 Dec 1;25(34):5374-5380. [doi: 10.1200/JCO.2007.11.2243]
63. Stover AM, Deal AM, Medley CJ, et al. Feasibility, acceptability, and utility of remote patient-reported outcomes monitoring in patients with lung cancer: a Moovcare© study. Clin Lung Cancer 2025 Dec;26(8):e660-e669.e1. [doi: 10.1016/j.cllc.2025.07.016] [Medline: 40850860]
64. Nuamek T, Kwateng PAN, Payne A, et al. Integrating electronic patient-reported outcome measures (ePROMs) into personalised follow-up for patients after radiotherapy: a feasibility study. Tech Innov Patient Support Radiat Oncol 2025 Sep;35:100333. [doi: 10.1016/j.tipsro.2025.100333]
65. Boisson-Walsh A, Cox C, O'Leary M, et al. A qualitative study of electronic patient-reported outcome symptom monitoring after thoracic surgery. J Surg Res 2024 Nov;303:744-755. [doi: 10.1016/j.jss.2024.09.051] [Medline: 39461326]
66. Schougaard LMV, Friis RB, Grytnes R, et al. Exploring the nurses' perspective on using remote electronic symptom monitoring in clinical decision-making among patients with metastatic lung cancer. Semin Oncol Nurs 2023 Oct 26;39(6):151517. [doi: 10.1016/j.soncn.2023.151517] [Medline: 39492326]
67. Gustafson DH, DuBenske LL, Atwood AK, et al. Reducing symptom distress in patients with advanced cancer using an e-alert system for caregivers: pooled analysis of two randomized clinical trials. J Med Internet Res 2017 Nov 14;19(11):e354. [doi: 10.2196/jmir.7466] [Medline: 29138131]
68. Roberts NA, Pelecanos A, Alexander K, Wyld D, Janda M. Implementation of patient-reported outcomes in a medical oncology setting (the iPROMOS Study): type II hybrid implementation study. J Med Internet Res 2024 Aug 27;26:e55841. [doi: 10.2196/55841] [Medline: 39190468]
69. van Gemert-Pijnen JEWC, Nijland N, van Limburg M, et al. A holistic framework to improve the uptake and impact of eHealth technologies. J Med Internet Res 2011 Dec 5;13(4):e111. [doi: 10.2196/jmir.1672] [Medline: 22155738]
70. Glasgow RE, Harden SM, Gaglio B, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. Front Public Health 2019;7:64. [doi: 10.3389/fpubh.2019.00064] [Medline: 30984733]
71. Yeoh EK, Wong MCS, Wong ELY, et al. Benefits and limitations of implementing chronic care model (CCM) in primary care programs: a systematic review. Int J Cardiol 2018 May 1;258:279-288. [doi: 10.1016/j.ijcard.2017.11.057] [Medline: 29544944]
72. Nipp RD, Horick NK, Qian CL, et al. Effect of a symptom monitoring intervention for patients hospitalized with advanced cancer: a randomized clinical trial. JAMA Oncol 2022 Apr 1;8(4):571-578. [doi: 10.1001/jamaoncol.2021.7643] [Medline: 35142814]
73. Girgis A, Durcinoska I, Arnold A, et al. Web-based patient-reported outcome measures for personalized treatment and care (PROMPT-Care): multicenter pragmatic nonrandomized trial. J Med Internet Res 2020 Oct 29;22(10):e19685. [doi: 10.2196/19685] [Medline: 33118954]

XSL•FO
RenderX

74. Ridde V, Pérez D, Robert E. Using implementation science theories and frameworks in global health. BMJ Glob Health 2020;5(4):e002269. [doi: 10.1136/bmjgh-2019-002269] [Medline: 32377405]
75. Proctor E, Luke D, Calhoun A, et al. Sustainability of evidence-based healthcare: research agenda, methodological advances, and infrastructure support. Implementation Sci 2015 Dec;10(1):88. [doi: 10.1186/s13012-015-0274-5]

Abbreviations

ePROM: electronic patient-reported outcome measure
HRQL: health-related quality of life
PCC: population, concept, and context
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Literature Search
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analysis Extension for Scoping Reviews
PROM: patient-reported outcome measure
RCT: randomized controlled trial
RE-AIM: reach, effectiveness, adoption, implementation, and maintenance
rSMS: remote digital symptom monitoring system

Edited by S Brini; submitted 06.Sep.2025; peer-reviewed by IM Rasmussen, SS Kollala; accepted 05.Jan.2026; published 24.Mar.2026.

Please cite as:
Huang H, Takemura N, Abu-Odah H, Li S, Mao T, Ye Z, Yorke J
The Current Landscape of Remote Digital Symptom Monitoring for Patients With Lung Cancer: Scoping Review
J Med Internet Res 2026;28:e83666
URL: https://www.jmir.org/2026/1/e83666
doi:10.2196/83666

© Hao Huang, Naomi Takemura, Hammoad Abu-Odah, Shuhan Li, Ting Mao, Zengjie Ye, Janelle Yorke. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Association Between Short-Form Video Use and Mental Health: Systematic Review and Meta-Analysis

Di Tang[1,2], PhD; Xin Zhang[3], PhD; Pengpeng Gou[4,5], MSc; Jie Feng[6], Prof Dr; Rui Hu[7], PhD; Kim-wai Raymond Sum[1], Prof Dr

[1]Department of Sports Science and Physical Education, Faculty of Education, Chinese University of Hong Kong, G09, Kwok Sports Building, Hong Kong, China (Hong Kong)

[2]The Nethersole School of Nursing, Faculty of Medicine, Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

[3]Faculty of Education, Central China Normal University, Wuhan, Hubei, China

[4]School of Sports Training, Wuhan Sports University, Wuhan, Hubei, China

[5]Center of Strength and Conditioning, Wuhan Sports University, Wuhan, Hubei, China

[6]Faculty of Sports Science, Ningbo University, Ningbo, Zhejiang, China

[7]School of Sports Science and Physical Education, Southwestern University of Finance and Economics, Chengdu, Sichuan, China

Corresponding Author:
Kim-wai Raymond Sum, Prof Dr
Department of Sports Science and Physical Education, Faculty of Education, Chinese University of Hong Kong, G09, Kwok Sports Building, Hong Kong, China (Hong Kong)

Abstract

Background: Short-form videos (SFVs) have emerged as a dominant trend in digital content sharing over the past decade, gaining rapid global popularity. An increasing number of studies have explored the association between SFV use and mental health, yet current empirical evidence remains inconsistent.

Objective: This study aimed to provide a comprehensive synthesis examining the relationship between SFV usage and mental health outcomes, distinguishing between problematic and routine usage behaviors.

Methods: Web of Science, PubMed, Scopus, Embase, PsycInfo, SportDiscus, and ProQuest were searched for relevant literature up to January 3, 2026. Statistical metrics reflecting the relationship between SFV use and mental health outcome indicators were extracted for meta-analysis, such as Pearson correlation coefficients, Spearman's rank correlation coefficients, and β coefficients. For intervention and longitudinal studies, we extracted the baseline correlation coefficients. The Joanna Briggs Institute Critical Appraisal Checklist was used to assess the risk of bias.

Results: A total of 58 studies, with a cumulative sample of 96,676 participants, were included in the analysis. Results showed no significant correlations between SFV use and positive psychological states, but significant positive correlations with negative psychological states, including depression (r=0.24, 95% CI 0.15-0.33), anxiety (r=0.26, 95% CI 0.17-0.35), stress (r=0.41, 95% CI 0.38-0.56), negative affect (r=0.22, 95% CI 0.10.33), loneliness (r=0.33, 95% CI 0.25-0.41), and boredom (r=0.42, 95% CI 0.29-0.53). Subgroup analyses revealed substantial differences between usage patterns: problematic use demonstrated significant negative associations with subjective well-being (r=−0.25, 95% CI −0.35 to −0.14) and positive correlations with adverse mental health outcomes, while routine usage showed no significant associations with negative affect and exhibited a negative but nonsignificant correlation with anxiety (r=−0.02, 95% CI −0.18 to 0.14). Additionally, while some studies reported significant associations between routine SFV use (time spent or frequency) and mental health outcomes, these findings were less consistent and showed smaller effect sizes compared to the more robust correlations found with problematic use measures. This suggests that the relationship between time spent on SFV and mental health may be less stable than that of problematic use patterns, highlighting the importance of considering qualitative aspects of usage rather than merely quantitative metrics.

Conclusions: This meta-analysis suggests that SFV use is associated with adverse mental health outcomes. Future research should use objective measurement instruments to capture contemporary SFV usage patterns and differentiate between distinct types of engagement. Additionally, studies with greater geographical diversity and longitudinal or experimental designs are needed to establish causality and examine temporal changes in these associations.

(J Med Internet Res 2026;28:e82503) doi:10.2196/82503

KEYWORDS

short-form video; social media; mobile phone; addiction; mental health; meta-analysis

XSL•FO
RenderX

Introduction

In the past decade, short-form videos (SFVs) have emerged as a dominant trend in digital content sharing, rapidly gaining global popularity. Platforms such as TikTok (known as Douyin in China) have transformed the way users create and consume content by enabling the production and sharing of videos lasting only a few seconds to a few minutes [1]. By 2022, TikTok had surpassed 1.2 billion monthly active users, with the average user spending 168 minutes per day on the platform [2]. In China, approximately 95% of internet users engage in SFV consumption [3]. Historically, video-sharing platforms like YouTube primarily hosted long-form content [4]. However, in response to the growing popularity of SFVs, major platforms such as YouTube, Instagram, and Snapchat have integrated similar features to meet user demand [5]. SFVs have gained attention due to their ease of use and convenience, alongside their engaging nature, becoming a defining feature of modern social media. These platforms not only facilitate passive content consumption but also actively encourage user participation and creativity, reshaping self-expression and cultural exchange in the digital age [6,7].

Compared to other forms of social media, SFV apps have quickly captured significant market share by addressing users' psychological needs, such as fostering social connections, providing emotional gratification, and fulfilling the desire for recognition [6]. One key factor driving this success is the availability of intuitive video editing tools, which empower users to easily become content creators, thereby enhancing engagement and participation [7]. For instance, Instagram reported that by 2020, more than 500 million accounts used its "Stories" feature daily, underscoring the growing appeal of SFV consumption [8]. The preference for SFVs can also be attributed to the fast-paced nature of modern life, where users prioritize convenience and brevity. Platforms offering SFV content are particularly appealing because they are user-friendly and rely on algorithm-driven personalization to deliver tailored content based on individual preferences [5]. This personalized approach not only improves user satisfaction but also boosts traffic, accelerating the widespread adoption of these platforms.

The diverse range of content available in SFVs further contributes to its widespread popularity. These videos span a variety of topics, including beauty, cooking, entertainment, education, health, and technology [5]. While traditional social media platforms often focus on lifestyle-oriented content, platforms like TikTok have differentiated themselves by emphasizing creative, unconventional, and often humorous videos. The majority of these videos are not professionally produced, which enhances their authenticity and relatability. TikTok, for example, primarily targets younger audiences with content such as music, dance, and viral comedic trends. In contrast, Douyin, its Chinese counterpart, caters to a broader demographic by featuring content that resonates with older users and focuses more heavily on everyday life [8]. This dual strategy has enabled TikTok to expand its user base and appeal to diverse cultural and demographic groups, cementing its role as a global leader in SFV content.

Despite the significant economic value brought by SFVs and their role in enriching modern entertainment, their rapid growth and potential for overuse have raised concerns among researchers about underlying health risks [6]. The convenience and entertainment value of SFVs make it easy for the public, especially younger users, to overuse them and struggle to control their viewing habits, resulting in increased time spent on these platforms. This trend exacerbates worries regarding negative health outcomes [6,9]. Current research indicates that excessive use of SFVs exhibits similarities to internet addiction or smartphone addiction, potentially leading to addictive behaviors and health issues, such as impairing social activities, interpersonal relationships, emotional stability, and psychological well-being [1,2,10,11]. Studies have shown that the algorithmic recommendation systems of SFV apps contribute to an information cocoon that increases users' feelings of loneliness, which may subsequently lead to higher levels of depression [12]. Additionally, problematic use of SFVs has been associated with suicidal ideation and self-injurious behaviors [13]. Moreover, numerous studies highlight a positive correlation between SFV usage and negative emotions such as anxiety and stress, while revealing a negative correlation with well-being, self-esteem, and sleep quality [14-17]. However, the literature is far from conclusive. Some studies have reported no significant correlation between SFV usage and specific mental health indicators, such as depression and anxiety [17-19]. For example, Banjanin et al [20] found no significant correlation between depression and time spent on social media. Chung et al [21] and Zhai et al [22] found a positive relationship between SFVs and life satisfaction, and Chao et al [23] discovered that different usage patterns may have opposed effects on mental health. This body of research underscores the complexities surrounding the relationships in this domain [17].

In the current investigations examining the relationship between SFV usage and psychological health indicators, the selected metrics for assessing SFV usage can be broadly categorized into 2 types. The first type pertains to an individual's general engagement with SFVs, encompassing several dimensions such as frequency, duration, and other indices reflecting the extent of usage [9,24-26]. While usage frequency quantifies the time an individual allocates to SFVs, it does not capture the psychological quality of that engagement. The second focuses on problematic usage, centering on addiction severity, dependency, and compulsive behavior [27-29]. Unlike simple usage intensity, problematic usage is characterized by psychological dependence, emotional distress, and functional impairment. Consequently, high-frequency use may simply reflect strong interest or leisure engagement without necessarily implying maladaptive outcomes. Therefore, treating these 2 types of metrics as interchangeable is methodologically problematic. Specifically, the association between usage behavior and psychological health may vary significantly depending on whether usage is operationalized as addiction or problematic dependency versus simply as daily time spent or frequency. For instance, research has reported that active usage of TikTok is significantly and negatively correlated with anxiety [30], whereas problematic usage of SFVs was reported to have a significant positive correlation with anxiety in another study [31]. This distinction is likely a primary driver of the

XSL•FO
RenderX

inconsistencies observed in the existing literature. Furthermore, variations in research quality and design (such as sample size, outcome selection, and measurement methodologies) may further obscure the true relationship between SFV usage and mental health.

Despite the growing body of literature on the relationship between SFV usage and health, there remains a lack of systematic synthesis of existing findings. Given the inconsistencies in prior findings, this study aims to clarify the relationship between SFV use and mental health through a systematic review and meta-analysis. Additionally, by distinguishing between different usage patterns, this research seeks to investigate potential heterogeneity in the impact of SFV usage on mental health outcomes.

Methods

The systematic review adhered to the 2020 guidelines established by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [32] (Checklist 1). In addition, the review protocol was registered in advance in the International Prospective Register of Systematic Reviews database (Registration ID: CRD420251015620).

Search Strategy

Web of Science, PubMed, Scopus, Embase, PsycInfo, SportDiscus, and ProQuest were searched to identify relevant literature from the inception of each database to January 3, 2026. The search terms combined variations of and specific health-related outcomes, such as depression, well-being, and anxiety. A detailed description of the search strategy can be found in the Multimedia Appendix 1. All references retrieved from the searches were organized using EndNote 20 (Clarivate) software.

Eligibility Criteria

The selection criteria imposed no restrictions on participants' sex or age. Regarding study design, we primarily included cross-sectional studies. However, longitudinal studies or interventional trials were also deemed eligible if they performed a cross-sectional analysis of baseline data—specifically, analyzing the relationship between SFV use and mental health outcomes before the commencement of any intervention or follow-up period—and reported the relevant correlation coefficients.

We operationally defined SFVs as video content lasting from several seconds to a few minutes [33,34], without imposing restrictions on specific platforms or content categories. However, for multifunctional video platforms not exclusively dedicated to SFVs (eg, Instagram), inclusion was restricted to studies that specifically isolated usage behaviors related to the SFV feature (eg, intensity of use or addiction to Instagram Reels), rather than general platform usage. Metrics for SFV use were categorized into regular usage indicators, including frequency, daily and weekly duration, as well as problematic use, which encompassed measures of video addiction, dependency levels, and indicators of excessive use. Regarding mental health, studies were required to assess mental health outcomes quantitatively. These constructs could include subjective measures (eg, self-reported or perceived health) or objective measures (eg, standardized tests). To ensure systematic categorization, identified health constructs were classified into mental health domains based on the World Health Organization's International Classification of Diseases [35].

Crucially, this study focuses exclusively on users' autonomous SFV usage behaviors. Therefore, specific interventional experiments involving forced exposure to predetermined content or artificial viewing tasks were excluded. Instead, included studies operationalized autonomous usage primarily through self-reported measures of usage frequency, duration, or validated intensity scales. Furthermore, conference abstracts lacking full texts or complete data, review studies, qualitative research, editorials, and letters were excluded from consideration. Non-English or non-Chinese literature was also excluded.

Study Selection

After removing duplicates, 2 reviewers (DT and XZ) independently screened the titles and abstracts. Records that did not align with the research objectives, such as those failing to report mental health outcomes, not focusing on SFV users, or having ineligible publication types or study designs, were excluded. Subsequently, the full texts of the remaining articles were retrieved and assessed independently by the 2 reviewers to determine final inclusion based on the criteria. During the literature screening process, any disagreements between the 2 reviewers were resolved through discussion to reach a consensus. In cases where consensus could not be achieved, a third reviewer (RH) was available to make a final determination. To ensure the consistency of the screening process, we calculated the interrater reliability using Cohen kappa coefficient, which indicated strong agreement between the reviewers (κ=0.83).

Risk of Bias

Two reviewers (DT and XZ) independently evaluated the quality of the included studies using the Joanna Briggs Institute (JBI) critical appraisal checklist for cross-sectional studies. Disagreements were resolved by discussion or, if necessary, consultation with a third reviewer (RH). The assessment covered three key domains: (1) sample and setting (inclusion criteria and subject description); (2) measurement validity (exposure, condition, and outcome measurement); and (3) confounding and analysis (identification or management of confounders and statistical appropriateness). Although some included studies used a longitudinal or interventional design, our analysis specifically concentrated on baseline data collected before any intervention. Therefore, we restricted our quality assessment to the cross-sectional component of these studies to evaluate the validity and reliability of the baseline data sources, confirming the appropriateness of the JBI cross-sectional checklist for this evaluation.

Data Extraction and Analysis

The extracted data included study design, sample size, participants' age, country, SFV platforms, and measured mental health outcomes. Additionally, statistical metrics reflecting the relationship between SFV use and mental health outcome indicators were extracted, such as Pearson correlation

XSL•FO
RenderX

coefficients, Spearman's rank correlation coefficients, and β coefficients. For intervention and longitudinal studies, we extracted the correlation coefficients at baseline.

Data analysis was carried out using Review Manager version 5.3 (The Cochrane Collaboration) and R version 4.4.2 (R Core Team). To achieve an approximate normal distribution for data synthesis, correlation coefficients (r) were converted to Fisher Z-scores. In instances where correlation coefficients were not reported directly, we used formulas recommended by the Cochrane Handbook for Systematic Reviews to transform other measures of effect size into correlation coefficients [36]. Quantitative synthesis was performed only when the relationship between a specific mental health outcome and SFVs was reported by more than 3 studies. Otherwise, a narrative discussion was conducted for the outcomes.

Statistical heterogeneity was quantified using the I^2 statistic and the Q test. Random-effects models were used when the heterogeneity of the studies was high (I^2>50%), while fixed-effects models were applied when the heterogeneity was low (I^2≤50%).

Based on the existing research, we conducted further subgroup analyses of SFV use by categorizing usage indicators into two types: problematic use, which included measurements of SFV addiction, dependence, and overuse; and routine use, which included daily use duration or frequency. We further explored the relationship between these 2 types of usage indicators and mental health to determine whether there were differences in how these indicators were associated with mental health outcomes.

Potential publication bias was assessed through visual inspection of funnel plots, where asymmetry was evaluated. In accordance with the recommendations of the Cochrane Handbook, for outcomes including 10 or more studies, we further conducted quantitative assessments using Egger's linear regression test and Begg's rank correlation test. A P value of less than .05 was considered to indicate statistically significant publication bias.

Sensitivity analyses were performed to evaluate the robustness of the pooled results. First, we conducted a "leave-one-out" analysis to determine if any single study disproportionately influenced the overall effect size. Second, we assessed the impact of study quality on the outcomes by excluding studies classified as low quality based on the critical appraisal checklist.

Results

Search Results

A total of 4293 articles were retrieved from the databases. Figure 1 illustrates the identification process. Ultimately, 58 articles were included in the review, of which 53 articles were incorporated into the meta-analysis. Figure 1 presents the identification process using the PRISMA flow diagram.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow chart of the identification process. SFV: short-form video.

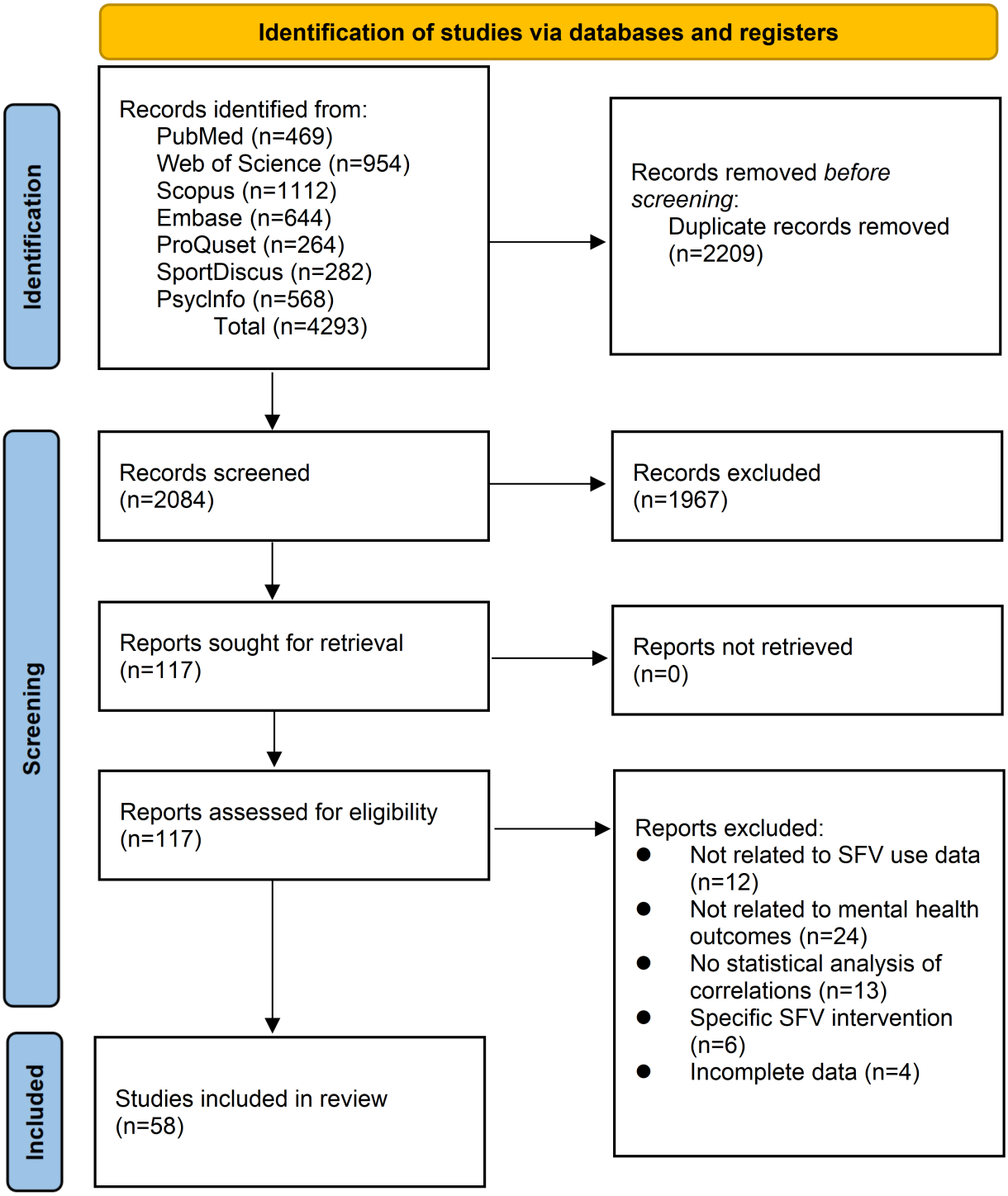

Characteristics of Studies

Among the 58 included studies, the majority were conducted in China (n=55), while a smaller number originated from other countries, including the United Arab Emirates (n=1), Pakistan (n=1), and the United States (n=1). The total accumulated sample size across all studies was 96,676, with individual study sample sizes ranging from 54 to 16,765. The mean age of participants ranged from 11.69 (SD 1.11) to 70.65 (SD 6.40) years, with some studies targeting specific populations, such as middle school students [37] and college students [29,38], while others included broader age ranges, such as 18 - 95 years [26]. Only 16 articles reported the specific SFV platforms used. The primary platforms identified were TikTok, YouTube (YouTube Shorts), and Instagram (Instagram Reels). All studies were published between 2021 and 2025.

In terms of study design, 52 studies used a cross-sectional approach, 4 used a prospective design, 1 was an intervention study, and 1 adopted a mixed methods framework. SFV use

XSL•FO
RenderX

was analyzed in various forms across the studies. SFV addiction was the most frequently examined behavior, reported in 52 studies, followed by daily use in 15 studies, problematic use in 5 studies, and overuse or dependence in a smaller subset [38-40]. Primary mental health outcomes assessed in these studies included depression (n=19), anxiety (n=13), loneliness (n=7), and well-being (n=12). Additional reported outcomes included stress, boredom, affect, suicidal ideation, and emotional dysregulation. Detailed study characteristics are shown in Multimedia Appendix 2.

Risk of Bias

Quality assessment of the 58 included studies yielded an overall mean quality score of 79.1% (SD 14.3%). Among the included studies, 27 (46.6%) achieved high quality scores of greater than or equal to 87.5%. The quality assessment revealed significant variability across methodological domains. The highest compliance rates were observed for item 2 (detailed description of study subjects and setting; 96.6%), item 7 (valid and reliable outcome measurement; 96.6%), and item 8 (appropriate statistical analysis; 94.8%). Conversely, substantial methodological limitations were identified in item 1 (clear definition of inclusion criteria; 60.3%). Detailed quality assessment results were provided in Multimedia Appendix 3.

Meta-Analysis

A total of 53 studies were included in the meta-analysis, which examined psychological outcomes categorized into negative psychological states (including depression, stress, anxiety, loneliness, boredom, and negative affect) and positive psychological states (including subjective well-being and positive affect). Studies that were not included in the meta-analysis were synthesized narratively.

SFV Use and Positive Psychological States

As shown in Figure 2, no significant correlations were observed between SFV use and either subjective well-being or positive affect. However, the 2 subgroups, problematic and routine use, exhibited marginally significant heterogeneity in their associations with subjective well-being (P=.08; I^2=68.3%). After excluding the study by Wen [40], problematic SFV use demonstrated a significant negative correlation with subjective well-being (r=−0.25, 95% CI −0.35 to −0.14).

XSL•FO
RenderX

Figure 2. Association between short-form video use and (A) well-being and (B) positive affect [21-23,26,27,34,40-48].

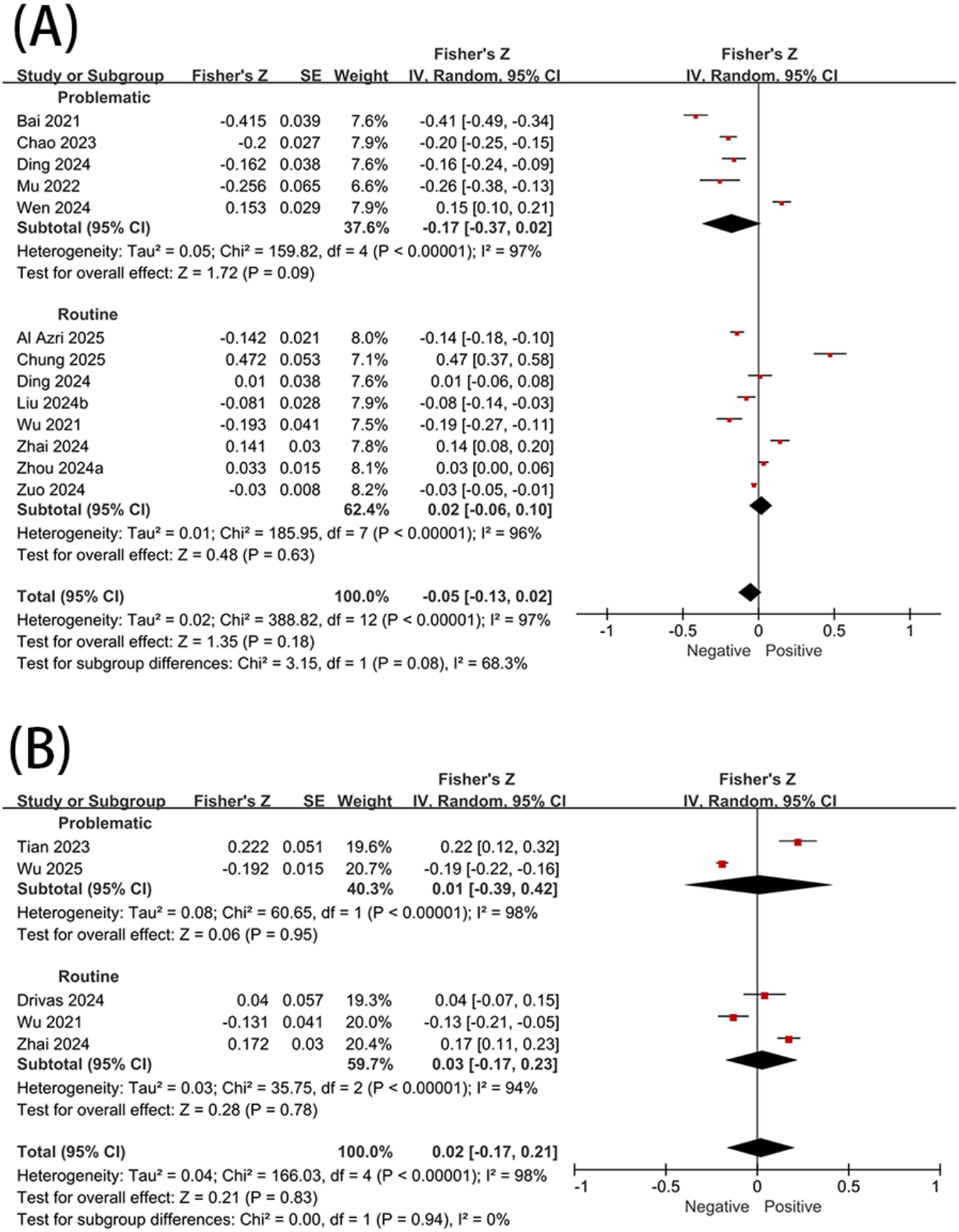

SFV Use and Negative Psychological States

As shown in Figure 3A, SFV use demonstrated a small but significant positive correlation with depression (r=0.24, 95% CI 0.15-0.33). Subgroup analysis revealed significant heterogeneity between the problematic and routine group (P<.001; I^2=96.2%). Problematic use demonstrated a significant moderate positive correlation with depression (r=0.34, 95% CI 0.28-0.39), while routine use showed a weak but significant positive association with depression (r=0.09, 95% CI 0.01-0.17). However, in the leave-one-out sensitivity analysis, the association between general use and depression became

XSL•FO
RenderX

statistically nonsignificant when [25,25], Zhang 2024 [49], Xia 2024 [50], or [44] were excluded individually. This suggests that the current evidence regarding the positive correlation between daily use and depression is not robust, and these findings should be interpreted with caution.

SFV use was significantly positively correlated with anxiety (r=0.26, 95% CI 0.17-0.35). Significant heterogeneity was found between the subgroups of problematic and routine use (P<.001; I^2=95.6%). Problematic use showed a moderate to large positive correlation with anxiety (r=0.38, 95% CI 0.33-0.43), while routine use demonstrated a nonsignificant negative association with anxiety (Figure 3B).

XSL•FO
RenderX

Figure 3. Association between short-form video use and depression (A) and anxiety (B) [13,22,23,25,28,29,31,38,39,43-45,47,49-62].

XSL•FO
RenderX

SFV use was found to be significantly associated with stress (r=0.41, 95% CI 0.38-0.56). Significant heterogeneity was also found between the subgroups of problematic and routine use (P=.006; I^2=86.6%). Specifically, problematic use demonstrated a moderate to large positive correlation with stress (r=0.44, 95% CI 0.36-0.51), whereas routine use showed a small to moderate positive association with stress (r=0.23, 95% CI 0.1-0.35). However, this result should be interpreted with caution as the routine use subgroup comprised only 1 study (Figure 4A). Regarding negative affect, SFV use was also found to be significantly positively associated (r=0.22, 95% CI 0.1-0.33). Problematic use exhibited a moderate and significant positive correlation with negative affect (r=0.3, 95% CI 0.150.43), while no significant association was observed between routine and negative affect (Figure 4B).

XSL•FO
RenderX

Figure 4. Association between short-form video use and (A) stress and (B) negative affect [3,14,15,22,23,47,48,57,58,63-66].

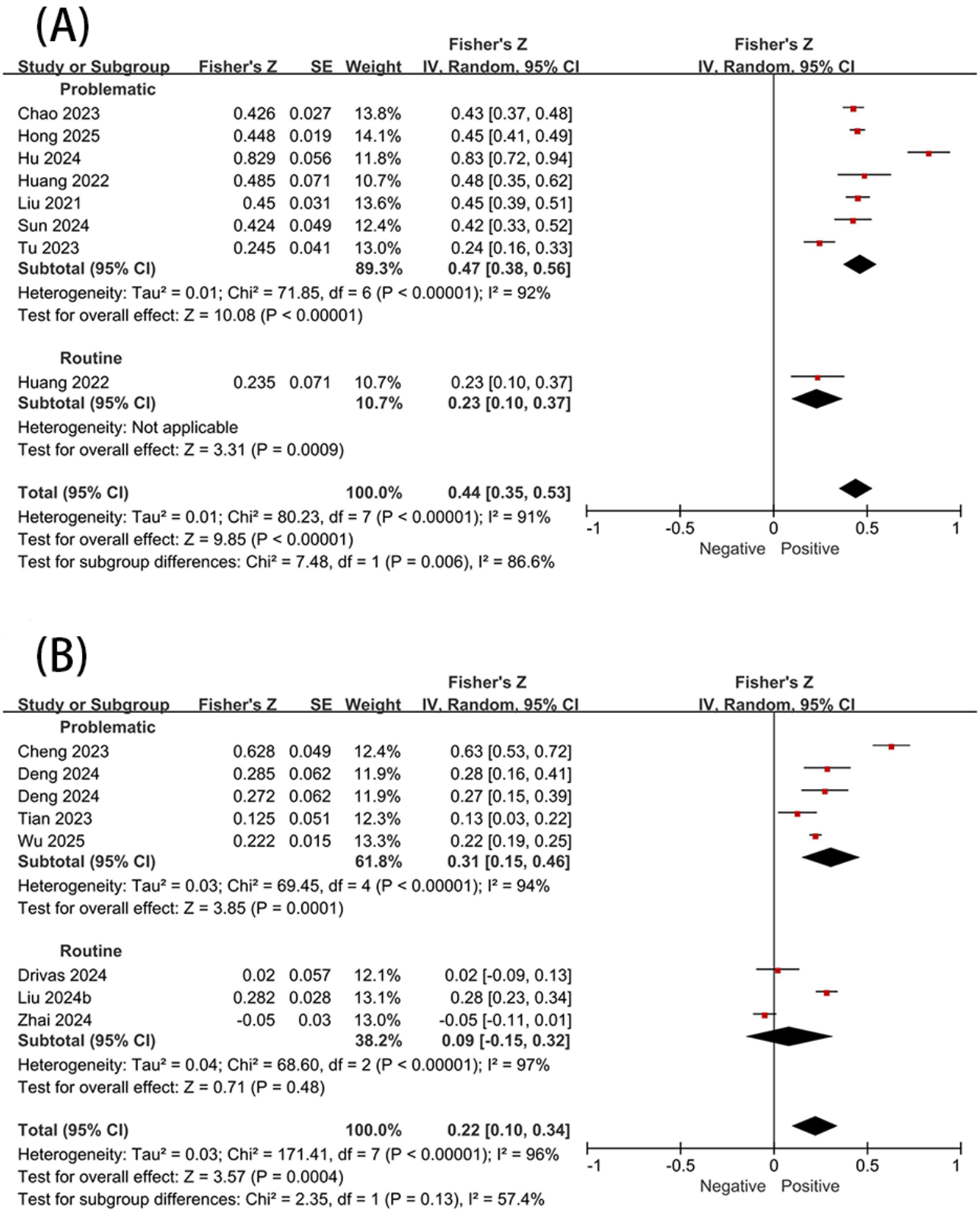

For SFV use and its associations with loneliness and boredom, only studies examining the relationships between problematic use and these 2 variables were sufficient for analysis. Therefore, no further subgroup analyses were conducted. Problematic SFV use demonstrated significant positive correlations with both loneliness (r=0.33, 95% CI 0.25-0.41) and boredom (r=0.42, 95% CI 0.29-0.53) (Figure 5).

Additionally, in studies not included in the meta-analysis, we found that problematic SFV use was positively associated with suicidal attempts (r=0.07; P<.05) [13]. SFV addiction showed positive associations with emotional suppression (r=0.29; P<.001) [37], alexithymia (r=0.339; P<.001) [67], emptiness (r=0.66, P<.001) [68], and neuroticism (r=0.3; P<.01) [69], (r=0.27; P<.01) [70], and (r=0.39; P<.001) [28]. And SFV

XSL•FO
RenderX

addiction was also found to be significantly associated with adolescent emotional dysregulation [71].

Figure 5. Association between short-form video use and (A) loneliness and (B) boredom [23,27,40,51,59,69,72-78].

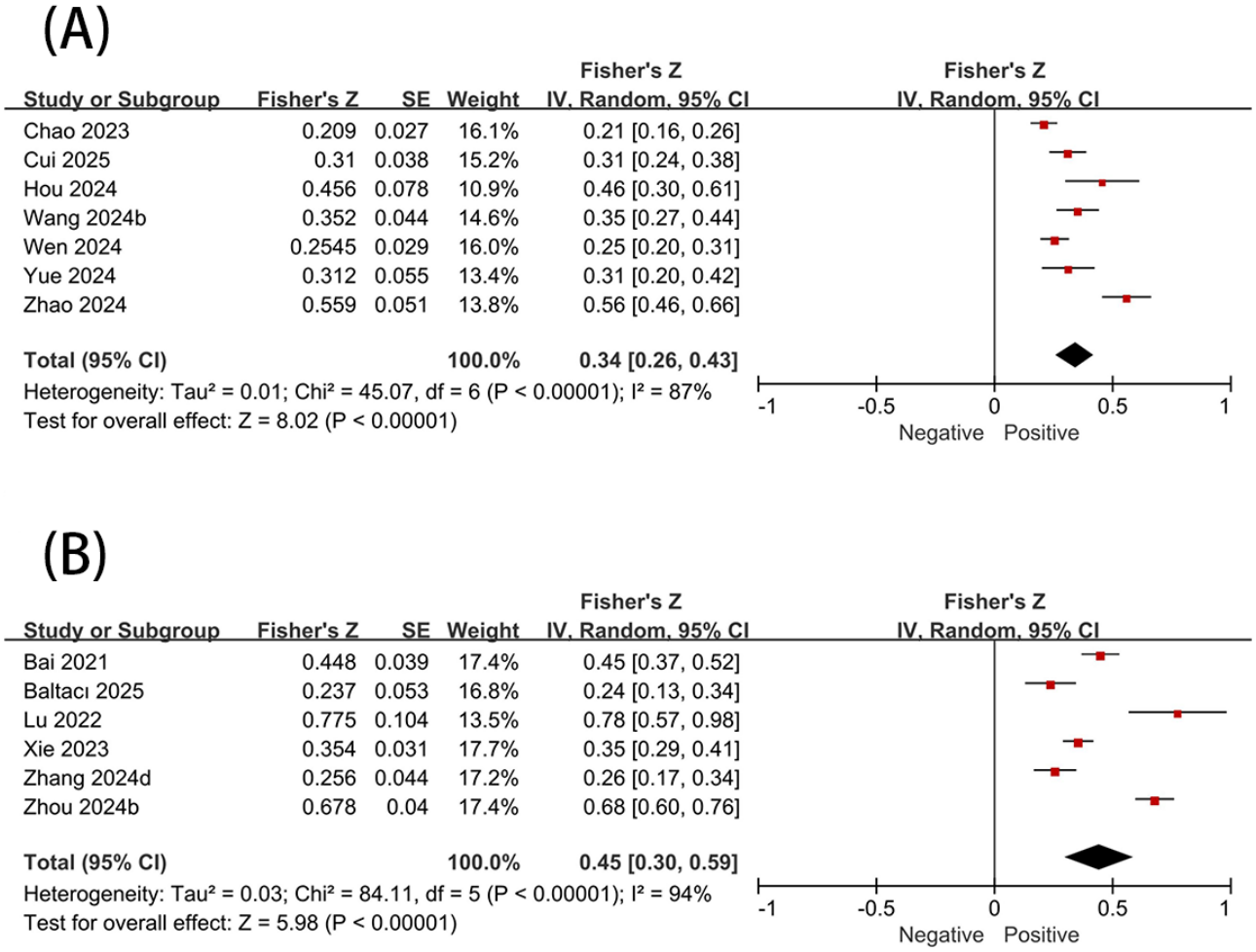

Publication Bias and Sensitivity Analysis

Funnel plots for all analyses showed no obvious asymmetry (Multimedia Appendix 4). Furthermore, for outcomes including 10 or more studies, Egger and Begg tests consistently yielded *P* values greater than .05, indicating no significant evidence of publication bias. Additionally, only 1 study was classified as low quality (JBI score <50%) [79]. Although this study was identified as low quality, it was primarily excluded from the final meta-analysis because insufficient studies reported the specific mental health indicators it analyzed to allow for pooling. Instead, it was synthesized narratively. Consequently, a sensitivity analysis comparing results before and after excluding low-quality studies was not feasible. Leave-one-out sensitivity analyses showed that the statistical significance of our main results remained consistent, confirming their overall robustness. However, in the subgroups examining the relationship between depression and routine use, and well-being and problematic use, excluding specific studies altered the significance of the results, as detailed in the Results section. Consequently, the findings concerning these specific associations appear unstable and warrant cautious interpretation.

Discussion

Principal Findings

This systematic review with meta-analysis provides a comprehensive synthesis of the relationship between SFV use and mental health. It is the first to distinguish between problematic and routine usage behaviors, revealing their divergent associations with different mental health outcomes. Analyses of current evidence indicated that SFV use demonstrated no significant associations with indicators of positive mental health, including subjective well-being and positive affect. Conversely, significant associations emerged between SFV use and various negative mental health states, including depression, anxiety, stress, negative affect, loneliness, and boredom.

Notably, our subgroup analyses revealed substantial differences between problematic and routine usage. Specifically, problematic SFV use demonstrated a significant negative association with subjective well-being and positive correlations with adverse mental health outcomes. In contrast, routine usage behaviors showed no significant associations with negative affect, and surprisingly, exhibited a significant negative association with anxiety, suggesting that nonproblematic routine engagement with SFV platforms may even serve an

XSL•FO
RenderX

anxiety-relieving function for some users [22,30]. In addition, current evidence on the relationship between routine use and depression or stress remains inconclusive. The pooled correlations between routine SFV use and depression were sensitive to the removal of individual studies, with statistical significance changing in several leave-one-out analyses, suggesting that these findings lack robustness. These findings also align with established literature indicating that heightened usage duration and elevated dependency levels are associated with increased vulnerability to adverse mental health outcomes and diminished positive psychological functioning compared to moderate or low-frequency usage patterns [23,44]. The marked distinction between problematic and routine usage in relation to mental health outcomes underscores the importance of considering qualitative differences in how individuals engage with SFV platforms, rather than treating SFV use as a monolithic construct. Overall, these findings support a more nuanced conceptualization of SFV use and its relationship with mental health, emphasizing that both the nature and extent of engagement are critical determinants of psychological outcomes.

Overall, in analyzing the relationship between problematic SFV use and mental health, the results of this meta-analysis broadly align with previous findings on problematic internet, smartphone, or social media use, both demonstrating correlations with decreased mental health and various mental health issues [80-83]. Prior research has also identified significant positive correlations between screen time and internalizing and externalizing behavior problems [84]. This suggests that despite differences in content and usage patterns between SFVs and traditional social media or video platforms, problematic SFV use may share similar mechanisms of impact on mental health with problematic internet, smartphone, and social media use.

However, it is noteworthy that our study found the strength of the association between problematic SFV use and negative mental health outcomes to be higher than previously established correlations for problematic phone use, internet use, social media use, and screen time [80-82,85,86]. In this meta-analysis, problematic SFV use shows moderate correlations with depression, anxiety, stress, loneliness, and boredom. This numerical difference might support that SFVs constitute a distinct media format, emphasizing the necessity for specialized research focused on SFVs rather than simply categorizing them under general “screen time” or “social media use” when considering their relationship with mental health.

First, the algorithmic recommendation mechanisms of SFV platforms may more intensely reinforce user immersion and platform stickiness [87-89]. Unlike traditional media, SFV platforms continuously deliver highly personalized content aligned with users’ interests through sophisticated algorithms, creating reinforcement loops that potentially lead to information cocoons and cognitive biases [90], which may further exacerbate negative mental health outcomes [12,91].

Furthermore, the short duration and rapid switching nature of SFVs may significantly impact users’ attention regulation and cognitive processing abilities. Long-term exposure potentially leads to attention fragmentation and increased cognitive burden, affecting daily functioning [78,92]. The highly condensed nature of SFV content intensifies information input, worsening cognitive load and creating information overload that may harm mental health [40].

Additionally, the rise of SFV platforms has transformed user roles, with increasing numbers of users shifting from passive information recipients to content creators compared to traditional social media or mobile phone usage [7,93]. This transition requires users to spend more time creating content on the platform, fundamentally changing usage patterns. Particularly on platforms like TikTok, users can not only create content but also monetize their videos, establishing content creation as a career option [94,95]. This behavioral pattern transformation potentially introduces psychological impacts distinctly different from those associated with traditional video consumption behaviors. For long-term content creators on SFV platforms, metrics such as view counts and video revenue may produce additional psychological effects that differ significantly from the psychological experiences of regular users who primarily engage in viewing activities [95-98]. However, current research has not effectively differentiated between these behavioral patterns, instead using simple measures of usage frequency and dependence to quantify SFV engagement without a nuanced distinction of qualitatively different usage patterns. This limitation may partly explain the substantial heterogeneity observed in our meta-analysis. Previous research on social media and internet usage has established that different usage patterns may exhibit entirely distinct relationships with mental health outcomes, such as differences between active versus passive engagement styles [22,30,86]. Future research needs to explore not only variations in usage intensity but also examine different usage patterns and their differential impacts on mental health.

Furthermore, the discussion regarding causality warrants further elaboration in the future. Although existing evidence reveals significant associations between SFV use and mental health, the cross-sectional nature of the included studies limits our ability to determine the direction of causality. This relationship is likely complex and multidimensional. First, psychological states may serve as antecedents to SFV use. For instance, individuals may engage with SFVs as a coping mechanism or for emotional regulation in response to boredom, loneliness, or stress [99,100]. Second, SFV use may precede psychological states. For example, problematic use may lead to sleep displacement or increased social comparison, thereby precipitating stress or anxiety [44,101]. Finally, a reciprocal relationship is plausible, analogous to the “internet paradox” [102] or the “reinforcing spirals model” [103]. In this scenario, adverse psychological states may precipitate problematic use, which subsequently aggravates the initial mental health concerns. Future longitudinal inquiries are indispensable for elucidating these temporal dynamics.

The effect sizes for correlations between both routine usage and problematic usage with mental health outcomes exhibited substantial heterogeneity in our findings. This variability likely stems from the wide age ranges, demographic characteristics, usage patterns, and intensity levels represented across existing studies. This heterogeneity suggests that the impact of SFV use on mental health may be influenced by multiple moderating factors, consistent with previous meta-analytic findings for other

XSL•FO
RenderX

digital media. Future research should prioritize more specific investigations to analyze particular effect sizes for specific usage patterns and demographic groups. The differences in measurement instruments highlight the importance of developing and implementing standardized, validated assessment tools for SFV usage patterns and addiction. Specifically, researchers should consider establishing measures that distinguish between content creation and content consumption to more accurately reflect the usage dynamics of modern SFV platforms.

Limitations and Future Directions

Several limitations of the current meta-analysis warrant consideration. First, the included studies were predominantly conducted in China, with limited data from other countries and regions. Given that research has demonstrated the importance of considering cultural and educational backgrounds when investigating problematic smartphone use, caution should be exercised when generalizing these findings to populations in different countries. Previous research on social media has revealed significant variations in associations across different countries and regions [81,85], highlighting the need for future cross-cultural investigations of SFV use.

Second, the limited number of available studies precluded detailed subgroup analyses, assessments of publication bias, and examination of potential moderators such as age, gender, and educational level. Understanding these factors would provide valuable insights into how SFV use and mental health relationships vary across demographic profiles. Future studies should therefore prioritize comprehensive reporting of both participant characteristics (age, gender, and education) and SFV usage patterns (platforms and content types). Particularly important is the distinction between content creators and passive viewers, alongside demographic information. Such detailed data are essential for developing a more nuanced understanding of these relationships.

Third, most of the included studies relied on self-report measures with considerable heterogeneity in assessment instruments, and many studies lacked validated methods or objective standard measurements for SFV use, such as actual usage time and frequency data obtained directly from platform analytics. To enhance cross-study comparability, there is an urgent need to develop and validate standardized measurement tools specifically designed for SFV use or addiction. Future research should differentiate between various SFV usage patterns and explore the potential differential associations between qualitatively distinct forms of engagement and mental health outcomes. Additionally, researchers should consider using more objective approaches to assess users' mental health status, including a diversified indicator system comprising physiological measures, biofeedback techniques, and behavioral data.

Finally, most of the included studies employed cross-sectional designs, which preclude causal inferences and directional conclusions regarding the relationships observed. Future research should prioritize longitudinal investigations to further elucidate causal relationships and examine whether these associations change over time.

Conclusion

This meta-analysis reveals that problematic SFV use is significantly associated with adverse mental health outcomes and reduced psychological well-being. Our findings highlight the distinct relationships between problematic versus routine usage patterns and mental health. Future research should develop standardized assessment tools that better capture contemporary SFV usage patterns and differentiate between varied usage behaviors. Longitudinal and experimental studies are needed to clarify causal relationships and examine whether these associations change over time. Additionally, as current evidence is predominantly based on Chinese samples, there is a critical need for data from diverse geographic regions to understand how these relationships may vary across different cultural contexts and user populations worldwide.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for this study.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

DT and XZ conceptualized and designed the study. DT, XZ, RH, and PG performed the literature search, study selection, data extraction, and quality assessment. DT and JF conducted the statistical analysis and interpreted the results. DT drafted the initial manuscript with XZ's substantial input. XZ, JL, RH, PG, JF, and KRS critically revised the manuscript for important intellectual content and performed content revision. All authors reviewed and approved the final version of the manuscript before submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.

XSL•FO
RenderX

[DOCX File, 11 KB - jmir_v28i1e82503_app1.docx]

Multimedia Appendix 2
Characteristics of the included studies.
[DOCX File, 99 KB - jmir_v28i1e82503_app2.docx]

Multimedia Appendix 3
Quality assessment of included studies.
[DOCX File, 29 KB - jmir_v28i1e82503_app3.docx]

Multimedia Appendix 4
Publication bias.
[DOCX File, 494 KB - jmir_v28i1e82503_app4.docx]

Checklist 1
PRISMA checklist.
[DOCX File, 270 KB - jmir_v28i1e82503_app5.docx]

References

1. Zhang X, Wu Y, Liu S. Exploring short-form video application addiction: socio-technical and attachment perspectives. Telemat Inform 2019 Sep;42:101243. [doi: 10.1016/j.tele.2019.101243]
2. Jiang A, Li S, Wang H, et al. Assessing short-video dependence for e-mental health: development and validation study of the short-video dependence scale. J Med Internet Res 2025 Mar 4;27:e66341. [doi: 10.2196/66341] [Medline: 40053762]
3. Sun R, Zhang MX, Yeh C, Ung COL, Wu AMS. The metacognitive-motivational links between stress and short-form video addiction. Technol Soc 2024 Jun;77:102548. [doi: 10.1016/j.techsoc.2024.102548]
4. Davidson J, Liebald B, Liu J, et al. The Youtube video recommendation system. Presented at: Proceedings of the 4th ACM Conference on Recommender Systems (RecSys 2010); Sep 26-30, 2010; Barcelona, Spain p. 293-296. [doi: 10.1145/1864708.1864770]
5. Wright C. Are beauty bloggers more influential than traditional industry experts? J Promot Commun 2017 Dec 31;5(3) [FREE Full text]
6. Zhang N, Hazarika B, Chen K, Shi Y. A cross-national study on the excessive use of short-video applications among college students. Comput Human Behav 2023 Aug;145:107752. [doi: 10.1016/j.chb.2023.107752]
7. Kuss DJ, Griffiths MD. Social networking sites and addiction: ten lessons learned. Int J Environ Res Public Health 2017 Mar 17;14(3):311. [doi: 10.3390/ijerph14030311] [Medline: 28304359]
8. Wang Y. Humor and camera view on mobile short-form video apps influence user experience and technology-adoption intent, an example of TikTok (DouYin). Comput Human Behav 2020 Sep;110:106373. [doi: 10.1016/j.chb.2020.106373]
9. Shimoga SV, Erlyana E, Rebello V. Associations of social media use with physical activity and sleep adequacy among adolescents: cross-sectional survey. J Med Internet Res 2019 Jun 18;21(6):e14290. [doi: 10.2196/14290] [Medline: 31215512]
10. Andreassen CS, Pallesen S. Social network site addiction - an overview. Curr Pharm Des 2014;20(25):4053-4061. [doi: 10.2174/13816128113199990616] [Medline: 24001298]
11. Qu D, Liu B, Jia L, et al. The longitudinal relationships between short video addiction and depressive symptoms: a cross-lagged panel network analysis. Comput Human Behav 2024 Mar;152:108059. [doi: 10.1016/j.chb.2023.108059]
12. He Y, Liu D, Guo R, Guo S. Information cocoons on short video platforms and its influence on depression among the elderly: a moderated mediation model. Psychol Res Behav Manag 2023;16:2469-2480. [doi: 10.2147/PRBM.S415832] [Medline: 37426388]
13. Yu Z, Zhu X, Li Y. The association between problematic short video use and suicidal ideation and self-injurious behaviors: the mediating roles of sleep disturbance and depression. BMC Public Health 2024 Jun 25;24(1):1689. [doi: 10.1186/s12889-024-19191-5] [Medline: 38915039]
14. Huang Q, Hu M, Chen H. Exploring stress and problematic use of short-form video applications among middle-aged Chinese adults: the mediating roles of duration of use and flow experience. Int J Environ Res Public Health 2021 Dec 23;19(1):132. [doi: 10.3390/ijerph19010132] [Medline: 35010389]
15. Liu Y, Ni X, Niu G. Perceived stress and short-form video application addiction: a moderated mediation model. Front Psychol 2021;12:747656. [doi: 10.3389/fpsyg.2021.747656] [Medline: 35002843]
16. Ye JH, Wu YT, Wu YF, Chen MY, Ye JN. Effects of short video addiction on the motivation and well-being of Chinese vocational college students. Front Public Health 2022;10:847672. [doi: 10.3389/fpubh.2022.847672] [Medline: 35619803]
17. Kelly Y, Zilanawala A, Booker C, Sacker A. Social media use and adolescent mental health: findings from the UK Millennium Cohort Study. EClinicalMedicine 2018 Dec;6:59-68. [doi: 10.1016/j.eclinm.2018.12.005] [Medline: 31193561]

XSL•FO
RenderX

18. Liu M, Kamper-DeMarco KE, Zhang J, Xiao J, Dong D, Xue P. Time spent on social media and risk of depression in adolescents: a dose-response meta-analysis. Int J Environ Res Public Health 2022 Apr 24;19(9):5164. [doi: 10.3390/ijerph19095164] [Medline: 35564559]
19. Zhang Y, Li S, Yu G. The relationship between social media use and fear of missing out: a meta-analysis [Article in Chinese]. Acta Psychol Sinica 2021;53(3):273-290. [doi: 10.3724/SP.J.1041.2021.01228]
20. Banjanin N, Banjanin N, Dimitrijevic I, Pantic I. Relationship between internet use and depression: focus on physiological mood oscillations, social networking and online addictive behavior. Comput Human Behav 2015 Feb;43:308-312. [doi: 10.1016/j.chb.2014.11.013]
21. Chung D, Meng Y, Wang J. The role of short-form video apps in mitigating occupational burnout and enhancing life satisfaction among healthcare workers: a serial multiple mediation model. Healthcare (Basel) 2025 Feb 7;13(4):355. [doi: 10.3390/healthcare13040355] [Medline: 39997230]
22. Zhai G, Su J, Chen Z, et al. The relationships between short video usage and subjective well-being: mediation models and network analysis. Behav Sci (Basel) 2024 Nov 12;14(11):1082. [doi: 10.3390/bs14111082] [Medline: 39594382]
23. Chao M, Lei J, He R, Jiang Y, Yang H. TikTok use and psychosocial factors among adolescents: comparisons of non-users, moderate users, and addictive users. Psychiatry Res 2023 Jul;325:115247. [doi: 10.1016/j.psychres.2023.115247] [Medline: 37167877]
24. Sharifian N, Zahodne LB. Daily associations between social media use and memory failures: the mediating role of negative affect. J Gen Psychol 2021;148(1):67-83. [doi: 10.1080/00221309.2020.1743228] [Medline: 32281502]
25. Zhu C, Jiang Y, Lei H, Wang H, Zhang C. The relationship between short-form video use and depression among Chinese adolescents: examining the mediating roles of need gratification and short-form video addiction. Heliyon 2024;10(9):e30346. [doi: 10.1016/j.heliyon.2024.e30346] [Medline: 38707427]
26. Zuo X, Wang R, Hong Z. Does using short video apps impacts life satisfaction: a perspective from psycho-social mechanism. Curr Psychol 2024 Jun;43(24):21242-21256. [doi: 10.1007/s12144-024-05896-8]
27. Bai J, Mo K, Peng Y, et al. The relationship between the use of mobile social media and subjective well-being: the mediating effect of boredom proneness. Front Psychol 2021;11:568492. [doi: 10.3389/fpsyg.2020.568492] [Medline: 33584406]
28. Yang C, Du J, Li X, et al. Association between parental phubbing and short-form video addiction: a moderated mediation analysis among Chinese adolescents. J Affect Disord 2025 Jan 15;369:523-530. [doi: 10.1016/j.jad.2024.10.023] [Medline: 39389113]
29. Zhang D, Yang Y, Guan M. A cross-lagged analysis of the relationship between short video overuse behavior and depression among college students. Front Psychol 2024;15:1345076. [doi: 10.3389/fpsyg.2024.1345076] [Medline: 39086426]
30. Liu Q, Wen J, Wang N, Wang M. Longitudinal associations between TikTok use, self-concept clarity, and anxiety among Chinese emerging adults: exploring differential impacts of active and passive TikTok use. Emerg Adulthood 2025 Apr;13(2):265-277. [doi: 10.1177/21676968241305653]
31. Jiang Y, Feng Y, Qi J, He R, Chao M. The relationship between bullying victimization and academic performance among adolescents: the chain mediating roles of social anxiety and short video addiction. Psychol Sch 2024;61(10):3859-3872. [doi: 10.1002/pits.23267]
32. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
33. Yang Y, Liu RD, Ding Y, Lin J, Ding Z, Yang X. Time distortion for short-form video users. Comput Human Behav 2024 Feb;151:108009. [doi: 10.1016/j.chb.2023.108009]
34. Wu Y, Wang X, Hong S, Hong M, Pei M, Su Y. The relationship between social short-form videos and youth's well-being: it depends on usage types and content categories. Psychol Pop Media 2021;10(4):467-477. [doi: 10.1037/ppm0000292]
35. International Statistical Classification of Diseases and Related Health Problems: Alphabetical Index: World Health Organization; 2004. URL: https://books.google.co.in/books?id=yUH0zAEACAAJ [accessed 2026-02-19]
36. Higgins JPT, Thomas J, editors. Cochrane Handbook for Systematic Reviews of Interventions: Wiley-Blackwell; 2019. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook [accessed 2026-02-19]
37. Yang J, Ti Y, Ye Y. Offline and online social support and short-form video addiction among Chinese adolescents: the mediating role of emotion suppression and relatedness needs. Cyberpsychol Behav Soc Netw 2022 May;25(5):316-322. [doi: 10.1089/cyber.2021.0323] [Medline: 35486859]
38. Zhang Q, Wang A, Zhang J. Relationship between social short-form video usage and mental health-related behaviors among college students [Article in Chinese]. Chin J Sch Health 2023;44(4):586-589. [doi: 10.16835/j.cnki.1000-9817.2023.04.025]
39. Zhang D, Wu J, Li J, Yang Y. Influence of short video overuse behavior on depression of college students: the chain mediating effect of self-concept clarity and meaning in life [Article in Chinese]. Chin J Behav Med Brain Sci 2024;33(4):342-346. [doi: 10.3760/cma.j.cn371468-20230921-00125]
40. Wen X, Zhou Y, Li Y, Li X, Qu P. Perceived overload on short video platforms and its influence on mental health among the elderly: a moderated mediation model. Psychol Res Behav Manag 2024;17:2347-2362. [doi: 10.2147/PRBM.S459426] [Medline: 38882234]

XSL•FO
RenderX

41. Ding J, Hu Z, Zuo Y, Xv Y. The relationships between short video addiction, subjective well-being, social support, personality, and core self-evaluation: a latent profile analysis. BMC Public Health 2024 Dec 18;24(1):3459. [doi: 10.1186/s12889-024-20994-9] [Medline: 39695521]
42. Mu H, Jiang Q, Xu J, Chen S. Drivers and consequences of short-form video (SFV) addiction amongst adolescents in China: stress-coping theory perspective. Int J Environ Res Public Health 2022 Oct 29;19(21):14173. [doi: 10.3390/ijerph192114173] [Medline: 36361050]
43. Al Azri M, Al Harrasi A, Al Ghailani A, et al. The moderating roles of social comparison in the relationship between Instagram reels use and mental health outcomes: a cross-sectional study in Oman. Sci Rep 2025 Nov 11;15(1):39457. [doi: 10.1038/s41598-025-22938-8] [Medline: 41219267]
44. Liu M, Zhuang A, Norvilitis JM, Xiao T. Usage patterns of short videos and social media among adolescents and psychological health: a latent profile analysis. Comput Human Behav 2024 Feb;151:108007. [doi: 10.1016/j.chb.2023.108007]
45. Zhou D, Li W. Don't say it's too late in life: how short video viewing affects the mental health of the elderly [Article in Chinese]. Doc In Knowl 2024;41(36–45):64. [doi: 10.13366/j.dik.2024.05.036]
46. Tian X, Bi X, Chen H. How short-form video features influence addiction behavior? Empirical research from the opponent process theory perspective. Inf Technol People 2023 Jan 13;36(1):387-408. [doi: 10.1108/ITP-04-2020-0186]
47. Wu Y, Bai Y, Liu X, Xu W, Liu Y. Gender differences in the relationship between short-form video addiction and adolescent depression: the mediating role of attentional bias. Cyberpsychol Behav Soc Netw 2025 Mar;28(3):169-177. [doi: 10.1089/cyber.2024.0442] [Medline: 39772883]
48. Drivas M, Reed OS, Berndt-Goke M. #WhatIEatInADay: the effects of viewing food diary TikTok videos on young adults' body image and intent to diet. Body Image 2024 Jun;49:101712. [doi: 10.1016/j.bodyim.2024.101712] [Medline: 38636388]
49. Zhang R, Su Y, Lin Z, Hu X. The impact of short video usage on the mental health of elderly people. BMC Psychol 2024 Oct 31;12(1):612. [doi: 10.1186/s40359-024-02125-6] [Medline: 39482710]
50. Xia LL, Li JD, Duan F, et al. Effects of online game and short video behavior on academic delay of gratification - mediating effects of anxiety, depression and retrospective memory. Psychol Res Behav Manag 2023;16:4353-4365. [doi: 10.2147/PRBM.S432196] [Medline: 37908679]
51. Baltacı Ö, Açar I. Does leisure boredom predict short video addiction in adolescents? Psychiatr Q 2025 Jun 12. [doi: 10.1007/s11126-025-10172-4] [Medline: 40504300]
52. Dong P, Zhang X, Yin W, et al. Relationship Between Short Video Addiction Tendency and Depression Among Rural Older Adults: Cross-Sectional Study. J Med Internet Res 2025 Jun 30;27:e75938. [doi: 10.2196/75938] [Medline: 40587851]
53. Li L, Li X, Li Y, Liu XP, Huang L. Types of short video addiction among college freshmen: effects on career adaptability, insomnia, and depressive symptoms. Acta Psychol (Amst) 2024 Aug;248:104380. [doi: 10.1016/j.actpsy.2024.104380] [Medline: 38955033]
54. Peng B, Zhu X, Han T, Zhao S, Zheng J, Jia F. Stepping away from the scroll: a chain mediation model from physical exercise and adolescent short video addiction. BMC Psychol 2025 Dec 22;13(1):1366. [doi: 10.1186/s40359-025-03710-z] [Medline: 41430732]
55. Xu X, Li H, Bai R, Liu Q. Do boys and girls display different levels of depression in response to mobile phone addiction? examining the longitudinal effects of four types of mobile phone addiction. Psychol Res Behav Manag 2024;17:4315-4329. [doi: 10.2147/PRBM.S487298] [Medline: 39711982]
56. Li C, Wang Y. Short-form video applications usage and functionally dependent adults' depressive symptoms: a cross-sectional study based on a national survey. Risk Manag Healthc Policy 2024;17:3099-3111. [doi: 10.2147/RMHP.S491498] [Medline: 39676824]
57. Liu L, Wang Y. The influence of watching short videos on the mental health of the middle-aged and elderly in China [Article in Chinese]. Mod Prev Med 2024;51:111-117. [doi: 10.20043/j.cnki.MPM.202305090]
58. Cheng X, Su X, Yang B, Zarifis A, Mou J. Understanding users' negative emotions and continuous usage intention in short video platforms. Electron Commer Res Appl 2023 Mar;58:101244. [doi: 10.1016/j.elerap.2023.101244]
59. Hou A, Hou X. Enjoyment or loneliness: the influence of cognitive absorption on addiction to short-form video apps [Article in Chinese]. Bull Educ Psychol 2024;55:537-556. [doi: 10.6251/BEP.202403_55(3).0005]
60. Li G, Geng Y, Wu T. Effects of short-form video app addiction on academic anxiety and academic engagement: the mediating role of mindfulness. Front Psychol 2024;15:1428813. [doi: 10.3389/fpsyg.2024.1428813] [Medline: 39444830]
61. Mao M, Liao F. Undergraduates short form video addiction and learning burnout association involving anxiety symptoms and coping styles moderation. Sci Rep 2025 Jul 7;15(1):24191. [doi: 10.1038/s41598-025-09656-x] [Medline: 40624123]
62. Wang P, Ouyang M, Yin L, et al. Gender differences in the relationship between parent-adolescent attachment anxiety and adolescent short-form video addiction: the mediating role of fear of missing out. Curr Psychol 2024 Aug;43(29):24216-24228. [doi: 10.1007/s12144-024-06060-y]
63. Hong X, Du W, Quan X, Yu P, Chen L. Network analysis of the interactions of short video features and personal characteristics for short video addiction in Chinese young adults. Acta Psychol (Amst) 2025 Oct;260:105482. [doi: 10.1016/j.actpsy.2025.105482] [Medline: 40885173]
64. Hu H, Huang M. How stress influences short video addiction in China: an extended compensatory internet use model. Front Psychol 2024;15:1470111. [doi: 10.3389/fpsyg.2024.1470111] [Medline: 39583000]

XSL•FO
RenderX

65. Tu W, Nie Y, Liu Q. Does the effect of stress on smartphone addiction vary depending on the gender and type of addiction? Behav Sci (Basel) 2023 Sep 30;13(10):810. [doi: 10.3390/bs13100810] [Medline: 37887460]
66. Deng J, Wang M, Mu W, et al. The relationship between addictive use of short-video platforms and marital satisfaction in older Chinese couples: an asymmetrical dyadic process. Behav Sci (Basel) 2024 Apr 25;14(5):364. [doi: 10.3390/bs14050364] [Medline: 38785855]
67. Li F, Wang G, Zhu D, et al. Parental neglect and short-form video application addiction in Chinese adolescents: the mediating role of alexithymia and the moderating role of refusal self-efficacy. Child Abuse Negl 2023 Sep;143:106345. [doi: 10.1016/j.chiabu.2023.106345] [Medline: 37451180]
68. Mu W, Kong L, He A. Understanding emptiness in a Chinese Sample: cross-cultural validation, latent profile analysis, and association with short-form video addiction. Int J Ment Health Addiction 2025 Oct;23(5):3895-3916. [doi: 10.1007/s11469-024-01327-8]
69. Cui S, Jiang J, Mu L. Impact mechanism of adult offspring's phubbing behavior on elderly short video addiction: a moderated mediation model. Applied Research Quality Life 2024 Apr;20(2):439-458. [doi: 10.1007/s11482-024-10405-x]
70. Asad K, Ali F, Awais M. Personality traits, narcissism and TikTok addiction: a parallel mediation approach. nt J Media Inf Lit 2022;7(2):293-304. [doi: 10.13187/ijmil.2022.2.293]
71. Wang J, Wang M, Lei L. Longitudinal links among paternal and maternal harsh parenting, adolescent emotional dysregulation and short-form video addiction. Child Abuse Negl 2023 Jul;141:106236. [doi: 10.1016/j.chiabu.2023.106236] [Medline: 37192589]
72. Wang P, Yue Y, Ouyang M, Yin L, Yin Y, Li B. Do narcissistic people exhibit more authentic self-disclosure to Generative AI? The roles of short-form video addiction, loneliness, and usage intention. Soc Sci Comput Rev 2025 Dec;43(6):1159-1180. [doi: 10.1177/08944393241308511]
73. Yue H, Yang G, Bao H, Bao X, Zhang X. Linking negative cognitive bias to short‐form video addiction: the mediating roles of social support and loneliness. Psychol Sch 2024 Oct;61(10):4026-4040. [doi: 10.1002/pits.23260]
74. Zhao Z, Kou Y. Effects of loneliness on short video addiction among college students: the chain mediating role of social support and physical activity. Front Public Health 2024;12:1484117. [doi: 10.3389/fpubh.2024.1484117] [Medline: 39600403]
75. Lu L, Liu M, Ge B, Bai Z, Liu Z. Adolescent addiction to short video applications in the mobile internet era. Front Psychol 2022;13:893599. [doi: 10.3389/fpsyg.2022.893599] [Medline: 35619797]
76. Xie J, Xu X, Zhang Y, et al. The effect of short-form video addiction on undergraduates' academic procrastination: a moderated mediation model. Front Psychol 2023;14:1298361. [doi: 10.3389/fpsyg.2023.1298361] [Medline: 38162977]
77. Zhang Y, Bu R, Li X. Social exclusion and short video addiction: the mediating role of boredom and self-control. Psychol Res Behav Manag 2024;17:2195-2203. [doi: 10.2147/PRBM.S463240] [Medline: 38832345]
78. Zhou L, Lv X, Zhou Y, Li J, Yu Z, Gao X. A network analysis perspective on the relationship between boredom, attention control, and problematic short video use among a sample of Chinese young adults. Int J Ment Health Addiction 2025 Dec;23(6):5022-5042. [doi: 10.1007/s11469-024-01392-z]
79. Abulibdeh ES, Alneyadi S, Skaik H, El-Saleh MS, Libdeh FIA, Naser K. Short videoaddiction and its relationship with students' academic achievement and well-being: a pilot study. Presented at: 2024 Global Digital Health Knowledge Exchange & Empowerment Conference (gDigiHealth.KEE); Sep 24-26, 2024. [doi: 10.1109/gDigiHealth.KEE62309.2024.10761759]
80. Cai Z, Mao P, Wang Z, Wang D, He J, Fan X. Associations between problematic internet use and mental health outcomes of students: a meta-analytic review. Adolesc Res Rev 2023;8(1):45-62. [doi: 10.1007/s40894-022-00201-9] [Medline: 36744121]
81. Huang C. A meta-analysis of the problematic social media use and mental health. Int J Soc Psychiatry 2022 Feb;68(1):12-33. [doi: 10.1177/0020764020978434] [Medline: 33295241]
82. Shannon H, Bush K, Villeneuve PJ, Hellemans KG, Guimond S. Problematic social media use in adolescents and young adults: systematic review and meta-analysis. JMIR Ment Health 2022 Apr 14;9(4):e33450. [doi: 10.2196/33450] [Medline: 35436240]
83. Yigiter MS, Demir S, Dogan N. The relationship between problematic social media use and depression: a meta-analysis study. Curr Psychol 2024 Mar;43(9):7936-7951. [doi: 10.1007/s12144-023-04972-9]
84. Eirich R, McArthur BA, Anhorn C, McGuinness C, Christakis DA, Madigan S. Association of screen time with internalizing and externalizing behavior problems in children 12 years or younger: a systematic review and meta-analysis. JAMA Psychiatry 2022 May 1;79(5):393-405. [doi: 10.1001/jamapsychiatry.2022.0155] [Medline: 35293954]
85. Ahmed O, Walsh EI, Dawel A, Alateeq K, Espinoza Oyarce DA, Cherbuin N. Social media use, mental health and sleep: a systematic review with meta-analyses. J Affect Disord 2024 Dec 15;367:701-712. [doi: 10.1016/j.jad.2024.08.193] [Medline: 39242043]
86. Godard R, Holtzman S. Are active and passive social media use related to mental health, wellbeing, and social support outcomes? A meta-analysis of 141 studies. J Comput Mediat Commun 2023 Nov 8;29(1):zmad055. [doi: 10.1093/jcmc/zmad055]
87. Liao M. Analysis of the causes, psychological mechanisms, and coping strategies of short video addiction in China. Front Psychol 2024;15:1391204. [doi: 10.3389/fpsyg.2024.1391204] [Medline: 39165759]

XSL•FO
RenderX

88. Qin Y, Omar B, Musetti A. The addiction behavior of short-form video app TikTok: the information quality and system quality perspective. Front Psychol 2022;13:932805. [doi: 10.3389/fpsyg.2022.932805] [Medline: 36148123]
89. Jia Y, Zheng K, Liu X, Guo X. How short video platforms retain customers: focusing on the roles of user stickiness and flow experience. Asia Pac J Mark Logist 2025 Oct 16;37(8):2308-2326. [doi: 10.1108/APJML-05-2024-0645]
90. Geschke D, Lorenz J, Holtz P. The triple-filter bubble: using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. Br J Soc Psychol 2019 Jan;58(1):129-149. [doi: 10.1111/bjso.12286] [Medline: 30311947]
91. Sunstein CR. Infotopia How Many Minds Produce Knowledge: Oxford University Press; 2006. URL: https://academic.oup.com/book/50393 [accessed 2026-02-19] [doi: 10.1093/oso/9780195189285.001.0001]
92. Firth J, Torous J, Stubbs B, et al. The “online brain”: how the Internet may be changing our cognition. World Psychiatry 2019 Jun;18(2):119-129. [doi: 10.1002/wps.20617] [Medline: 31059635]
93. Omar B, Dequan W. Watch, share or create: the influence of personality traits and user motivation on TikTok mobile video usage. Int J Interact Mob Technol 2020;14(4):121. [doi: 10.3991/ijim.v14i04.12429]
94. Abidin C. Mapping internet celebrity on TikTok: exploring attention economies and visibility labours. Cult Sci J 2020 Jan;12(1):77-103. [doi: 10.5334/csci.140]
95. Kaye DBV, Chen X, Zeng J. The co-evolution of two Chinese mobile short video apps: parallel platformization of Douyin and TikTok. Mob Media Commun 2021 May;9(2):229-253. [doi: 10.1177/2050157920952120]
96. Arriagada A, Ibáñez F. “You need at least one picture daily, if not, you’re dead:" content creators and platform evolution in the social media ecology. Social Media + Society 2020 Jul;6(3). [doi: 10.1177/2056305120944624]
97. Duffy BE, Meisner C. Platform governance at the margins: social media creators’ experiences with algorithmic (in)visibility. Media Cult Soc 2023 Mar;45(2):285-304. [doi: 10.1177/01634437221111923]
98. Duffy BE, Wissinger E. Mythologies of creative work in the social media age: fun, free, and “just being me”. Int J Commun 2017 Oct 27;11:20-4671 [FREE Full text]
99. Menon D. Factors influencing Instagram Reels usage behaviours: an examination of motives, contextual age and narcissism. Telemat Inform Rep 2022 Mar;5:100007. [doi: 10.1016/j.teler.2022.100007]
100. Deng X, Yu Z. An extended hedonic motivation adoption model of TikTok in higher education. Educ Inf Technol 2023 Oct;28(10):13595-13617. [doi: 10.1007/s10639-023-11749-x]
101. Tingrong Y, Gen Z. Magic mirror, who is the fairest one of all? testing the mediating effect between short-video social media exposure and appearance anxiety. BMC Psychol 2025 Mar 21;13(1):283. [doi: 10.1186/s40359-025-02601-7] [Medline: 40119477]
102. Kraut R, Patterson M, Lundmark V, Kiesler S, Mukophadhyay T, Scherlis W. Internet paradox: A social technology that reduces social involvement and psychological well-being? Am Psychol 1998;53(9):1017-1031. [doi: 10.1037/0003-066X.53.9.1017] [Medline: 9841579]
103. Slater MD. Reinforcing spirals: the mutual influence of media selectivity and media effects and their impact on individual behavior and social identity. Commun Theory 2007 Aug;17(3):281-303. [doi: 10.1111/j.1468-2885.2007.00296.x]

Abbreviations

JBI: Joanna Briggs Institute
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SFV: short-form video

Edited by A Schwartz, M Balcarras; submitted 16.Aug.2025; peer-reviewed by JY Fam, R Zhang, S Huiqi; accepted 23.Jan.2026; published 18.Mar.2026.

Please cite as:
Tang D, Zhang X, Gou P, Feng J, Hu R, Sum KWR
Association Between Short-Form Video Use and Mental Health: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e82503
URL: https://www.jmir.org/2026/1/e82503
doi:10.2196/82503

© Di Tang, Xin Zhang, Pengpeng Gou, Jie Feng, Rui Hu, Kim-wai Raymond Sum. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN

XSL•FO
RenderX

1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Telehealth Approaches for Pediatric Otitis Media and Clinical Outcomes: Scoping Review

Masao Noda[1*], MD, PhD, MBA; Ryohei Akiyoshi[2*], MD, PhD; Makoto Hosoya[3], MD, PhD; Chikako Shinkawa[4], MD, PhD; Ryota Koshu[1], MD; Hidekane Yoshimura[5], MD, PhD; Yukifumi Monden[6], MD, PhD; Hiroaki Fushiki[7], MD, PhD; Yasuhiro Tanaka[2], MD, PhD; Makoto Ito[1], MD, PhD

[1]Department of Otolaryngology and Head and Neck Surgery, Jichi Medical University, 3311-1, Yakushiji, Shimotsuke, Japan

[2]Department of Otolaryngology and Head and Neck Surgery, Dokkyo Medical University Saitama Medical Center, Saitama, Japan

[3]Department of Otolaryngology and Head and Neck Surgery, Keio University, Tokyo, Japan

[4]Department of Otolaryngology and Head and Neck Surgery, Yamagata University, Yamagata, Japan

[5]Department of Otolaryngology – Head and Neck Surgery, Shinshu University School of Medicine, Matsumoto, Japan

[6]Department of Pediatrics, Jichi Medical University Hospital, Shimotsuke, Japan

[7]Department of Otolaryngology, Mejiro University Ear Institute Clinic, Saitama, Japan

[*]these authors contributed equally

Corresponding Author:
Masao Noda, MD, PhD, MBA
Department of Otolaryngology and Head and Neck Surgery, Jichi Medical University, 3311-1, Yakushiji, Shimotsuke, Japan

Abstract

Background: Otitis media (OM) is a common pediatric infection worldwide. Conventionally, accurate diagnosis depends on in-person pneumatic otoscopy, which is not always accessible, contributing to delayed care and inappropriate prescribing, especially in underserved settings. Rapid advances in telemedicine and digital tools have accelerated the development of remote approaches for assessing pediatric ear diseases, while diagnostic quality maintenance, care models, and real-world outcomes have not been comprehensively mapped in children.

Objective: This study aimed to map existing telehealth technologies and operational models used for pediatric OM and report their diagnostic and implementation outcomes to guide practice and further research.

Methods: PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guided this scoping review (protocol not registered). MEDLINE (via PubMed), Scopus, Web of Science, and the Cochrane Central Register of Controlled Trials were searched for original English-language studies (published between January 1, 2010, and February 9, 2026) reporting the clinical implementation of telemedicine or digital health in children (<18 years) with suspected or confirmed OM. We excluded reviews, editorials, and protocols; conference abstracts; adult-only studies; and purely technical evaluations. Two reviewers independently extracted and charted the study characteristics: telemedicine model, technology, users, comparators, outcomes, and limitations. Findings were synthesized using practice-oriented mapping that aligned technologies with care models and implementation conditions.

Results: Fifty-two studies across 18 countries and various settings met the inclusion criteria. Sample sizes ranged from 6 to 3950, with heterogeneous reporting units (children, ears, episodes, or screening assessments). Asynchronous store-and-forward tele-otoscopy was the most common approach. The rest used synchronous or hybrid models. Image capture by trained personnel and review by experienced clinicians yielded substantial diagnostic consistency with in-person microscopy (κ 0.68 - 0.89, sensitivity 72% - 94%, specificity 93% - 98%, where available). However, the diagnostic yield was highly dependent on the training level: structured instruction improved video capture by parents and nonspecialists, whereas brief or written-only guidance resulted in low rates of diagnostically useful videos. Telemedicine approaches improved access, supported perioperative follow-up, and, in some contexts, reduced reexaminations and promoted more judicious antibiotic use. Televisits without otoscopy were associated with lower confirmation rates of middle ear effusion during tympanostomy tube placement. Evidence was heterogeneous, with predominantly small single-site studies, variable reference standards and operator training, and rapidly evolving device ecosystems.

Conclusions: This review provides a practice-oriented map of telehealth approaches for pediatric OM. Tele-otoscopy and adjunct digital tools are feasible, achieving diagnostic accuracy comparable to in-person assessments while enhancing access and service efficiency. However, important evidence gaps remain, including the need for large multisite trials, evaluation of long-term child outcomes, economic evaluations, and robust external validation of artificial intelligence–based diagnostic tools. Standardization of image capture protocols and integration into hybrid care models should be prioritized for scaling up.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e85416) doi:10.2196/85416

KEYWORDS

pediatrics; acute otitis media; otitis media with effusion; telemedicine; telehealth; smartphone otoscopy; digital otoscopy; artificial intelligence; tympanostomy

Introduction

Telemedicine has expanded rapidly across pediatrics over the past decade, further accelerated by the COVID-19 pandemic. Telehealth performance is disease specific, determined by whether robust remote proxies for the examination exist and are embedded in workflows [1,2]. In practice, the feasibility and safety of remote pediatric care depend on whether key physical findings can be reliably captured and shared. Equity concerns related to devices, connectivity, and digital literacy also underscore the need to track and address the digital divide [3]. For ear disease, this principally entails dependable tympanic membrane visualization and measures of middle ear status [4,5].

Within this context, otitis media (OM) is among the most common pediatric infections worldwide and a leading cause of health care visits, antibiotic prescriptions, and tympanostomy tube placement [6,7]. In 2021, the estimated global incident cases of OM in children were 297 million, highlighting the substantial population burden [6]. OM also contributes substantially to pediatric antibiotic exposure; a previous study reports that in a large outpatient cohort, antibiotics were prescribed in 44.8% of OM encounters [8]. Recent evidence from large outpatient datasets suggests that antibiotic prescribing patterns can differ between telemedicine and in-person visits for acute respiratory infections, raising concerns that diagnostic uncertainty—particularly when objective otoscopic findings are unavailable—may influence management [9]. Acute OM (AOM) and OM with effusion (OME) impose substantial morbidity, reduced quality of life, and increased health care costs across diverse settings [7,10], with a global review reporting annual costs as high as US $5 billion in the United States [10]. Accurate diagnosis hinges on direct visualization of the tympanic membrane and assessment of middle ear status—ideally by pneumatic otoscopy—which requires specific training and equipment [7,11]. Limited access to clinicians trained in pediatric otoscopy, particularly in rural and underserved regions, can contribute to delayed assessment and potentially inappropriate treatment [7].

Concurrently, digital health technologies relevant to ear disease have proliferated [12-14], including smartphone-connected and USB video-otoscopes, non-smartphone digital otoscopes, remote monitoring platforms, and early artificial intelligence–assisted decision support [15,16]. Beyond clinician-facilitated workflows, newer models increasingly rely on caregiver- or patient-captured otoscopic media for asynchronous review; however, interpretability can vary markedly depending on who captures the recordings, indicating that real-world deployment requires systems that standardize capture quality through clear thresholds, training, and feedback loops [17]. When paired with appropriate workflows, these tools enable asynchronous review by experts or real-time supervision and can be complemented by adjunct tests such as tympanometry or acoustic methods [18-23]. Emerging studies suggest that telemedicine can approximate in-person diagnostic performance for pediatric OM when image capture is standardized and operators are trained while also reducing barriers to care and supporting antibiotic stewardship [20,21,24]. However, important uncertainties remain regarding feasibility, usability, and sustainable integration into clinical practice.

Given the rapidly expanding yet fragmented evidence base, we conducted a scoping review to (1) map the technologies and care models used in telehealth for pediatric OM, (2) summarize diagnostic and implementation outcomes, and (3) identify evidence gaps to guide future research and clinical implementation. This review further aims to identify the key barriers and enablers of feasibility and to provide actionable guidance for implementing pediatric OM telehealth—across technologies, care models, and settings.

Methods

Overall Study Design

A scoping review was performed according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist [25]. The review process progressed through the stages outlined in the framework of Arksey and O'Malley [26], which included formulating the research question; identifying relevant studies; selecting studies; extracting and organizing the data; and synthesizing, summarizing, and presenting the findings. The protocol for this scoping review was neither registered nor published. Core decisions were defined a priori (research questions, eligibility criteria, databases, and data items). Minor refinements to the screening guidance, including operational definitions of clinical implementation and tele-otoscopy, were made iteratively during pilot screening to improve consistency and were applied uniformly thereafter. In addition, reporting of the literature search was guided by the PRISMA-S statement (an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews) to enhance transparency and reproducibility of the search methods [27].

Information Sources and Search Strategy

MEDLINE (via PubMed), Scopus, Web of Science Core Collection, and the Cochrane Central Register of Controlled Trials (CENTRAL) were searched for studies published between January 1, 2010, to February 9, 2026 (publication cutoff date), using a predefined strategy developed by 3 authors (MN, RA, and MH). Searches were last run on February 9, 2026, as the final rerun during the revision process; no publication date restrictions were applied beyond the search date, and no automated alerts were used. Each database was searched separately on its native platform rather than via a single multidatabase interface.

XSL•FO
RenderX

The strategy combined controlled vocabulary and free-text terms for three concepts: (1) otitis media, (2) pediatrics, and (3) telehealth or digital health. The start date (January 2010) was selected because smartphone-enabled imaging and telehealth implementations relevant to pediatric otoscopy began to emerge and diffuse in clinical practice around that period. Full strategies for each database are provided exactly as run in Multimedia Appendix 1. The strategy was developed de novo and was not adapted from prior reviews, and no published search filters were used.

In addition, forward citation searching was performed in Web of Science (using "Times Cited" and "Cited Reference Search") for key included articles, and reference lists of included studies were manually screened. Study registries were not searched, and study authors, experts, or manufacturers were not contacted for additional or missing information. The search strategy was not formally peer-reviewed using a dedicated tool but was iteratively reviewed within the author team. Duplicates were removed by automated matching of title, author, and publication year, followed by manual verification.

Eligibility Criteria

We defined the eligibility criteria using the population, concept, and context (PCC) framework developed by the Joanna Briggs Institute (JBI):

- Population: children and adolescents aged <18 years with suspected or confirmed OM (AOM/OME/recurrent acute OM [RAOM]), including mixed-age studies if pediatric data were separable
- Concept: telehealth or digital health approaches for middle ear evaluation and/or OM management, including tele-otoscopy (image/video capture with remote review), tele-otoscopy plus adjunct testing (tympanometry, acoustic methods), and other clinically implemented digital tools relevant to OM assessment or follow-up
- Context: any clinical or community setting (primary care, ED, specialty clinics, schools, outreach programs, home)
- Types of sources: original research reporting clinical implementation outcomes (diagnostic, clinical, implementation, or economic outcomes)

We excluded conference abstracts, protocols, reviews or editorials, adult-only studies, and purely technical evaluations without clinical implementation.

Data Items and Operational Definitions

We used "tele-otoscopy" to describe capture of otoscopic media (images/videos) for remote interpretation (store and forward or real time). We distinguished this from "video visits without otoscopy," in which no ear imaging was obtained. "Adjunct digital diagnostics" referred to nonimaging tools to assess middle ear status (eg, tympanometry and smartphone acoustics).

Study Inclusion Criteria and Exclusion Criteria

Studies were considered eligible if they were published in English between January 1, 2010, and February 9, 2026, and focused on telemedicine, telehealth, or other digital health interventions for pediatric OM. Eligible designs included qualitative, quantitative, or mixed-method approaches, provided that they reported outcomes of implementation in clinical practice. Studies that enrolled children and adolescents aged <18 years with suspected or confirmed OM—in which patients were assessed, monitored, or treated using telemedicine-supported interventions—were included.

Conference abstracts, reviews, editorials, and study protocols were excluded. Studies were also excluded if they focused exclusively on adult populations or conditions other than OM or those that did not involve the clinical application of telemedicine. Purely technical studies—including device prototype development and retrospective artificial intelligence model performance evaluations without prospective or clinical validation—and conference-only publications were excluded. Studies centered solely on audiology or hearing assessment without otoscopic or middle ear evaluation were also excluded.

Selection of Sources of Evidence

Titles and abstracts were independently screened by 2 reviewers (MN and RA). Full-text articles were retrieved for studies deemed potentially eligible and assessed against the predefined inclusion and exclusion criteria. Studies were excluded at the full-text stage for reasons including lack of clinical implementation, exclusive focus on retrospective algorithm performance, conference-only publications, absence of pediatric populations, or absence of OM-specific outcomes. Any disagreements were resolved through discussion and consensus. As part of the revision process, the rerun search results were screened using the same procedures and criteria as the original search.

Data Charting Process and Items

Data extraction was independently performed by 2 reviewers (MN and RA) using a standardized charting form. Extracted items included bibliographic details (author and year) and study design; country and setting (clinic, community, school, or home; level of care); population characteristics (age range, sample size, and OM phenotype when specified); care model (asynchronous, synchronous, or hybrid) and workflow description; operator roles (who captured and who interpreted) and training protocols (duration and content and any quality thresholds); technology and device or system characteristics (type, smartphone requirement, platform, and manufacturer when available); comparator and reference standard (otomicroscopy, pneumatic otoscopy, or surgical findings); outcomes (diagnostic performance, clinical outcomes, implementation outcomes, economic outcomes, and equity-relevant outcomes when reported); and study limitations, potential biases, and funding or conflict of interest statements when reported. Any discrepancies between reviewers were resolved through discussion and consensus. Some included studies reported diagnostic accuracy outcomes; thus, we conducted a targeted assessment for the subset of diagnostic accuracy studies using the revised Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool. Two reviewers independently assessed the risk of bias and applicability, with disagreements being resolved by consensus; detailed judgments are provided in Multimedia Appendix 2.

XSL•FO
RenderX

Synthesis of Results

We summarized the extracted data descriptively (counts/percentages) and synthesized the findings narratively. We created evidence maps (technology category × care model; operator capture vs interpretation) and an outcome-domain frequency figure to visualize evidence distribution and gaps. To apply an implementation science lens, we mapped the reported implementation outcomes and determinants across included studies to the reach, effectiveness, adoption, implementation, and maintenance (RE-AIM) framework. A summary of the RE-AIM framework analysis of each study included in this review is provided in Multimedia Appendix 3.

Results

Selection of Sources of Evidence

The database search retrieved 8356 articles from MEDLINE via PubMed (n=1687), Scopus (n=4810), Web of Science Core Collection (n=1775), and CENTRAL (n=84). In addition, 1911 records were identified through forward citation searching in Web of Science (and backward reference checking). Following the removal of duplicates, 5537 articles remained for the screening of titles and abstracts; of these, 5402 irrelevant articles were excluded, and 135 articles remained. The full texts of 135 articles were subsequently evaluated for eligibility; of these, 83 articles were excluded for the following reasons: nonpediatric population (n=9), absence of telemedicine interventions (n=17), review or commentary articles (n=13), unrelated to OM (n=13), not implemented in clinical practice (n=23), not in English (n=1), and unavailability of full text (n=7). Ultimately, 52 studies were included in this scoping review (Figure 1, PRISMA flow diagram).

Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram.

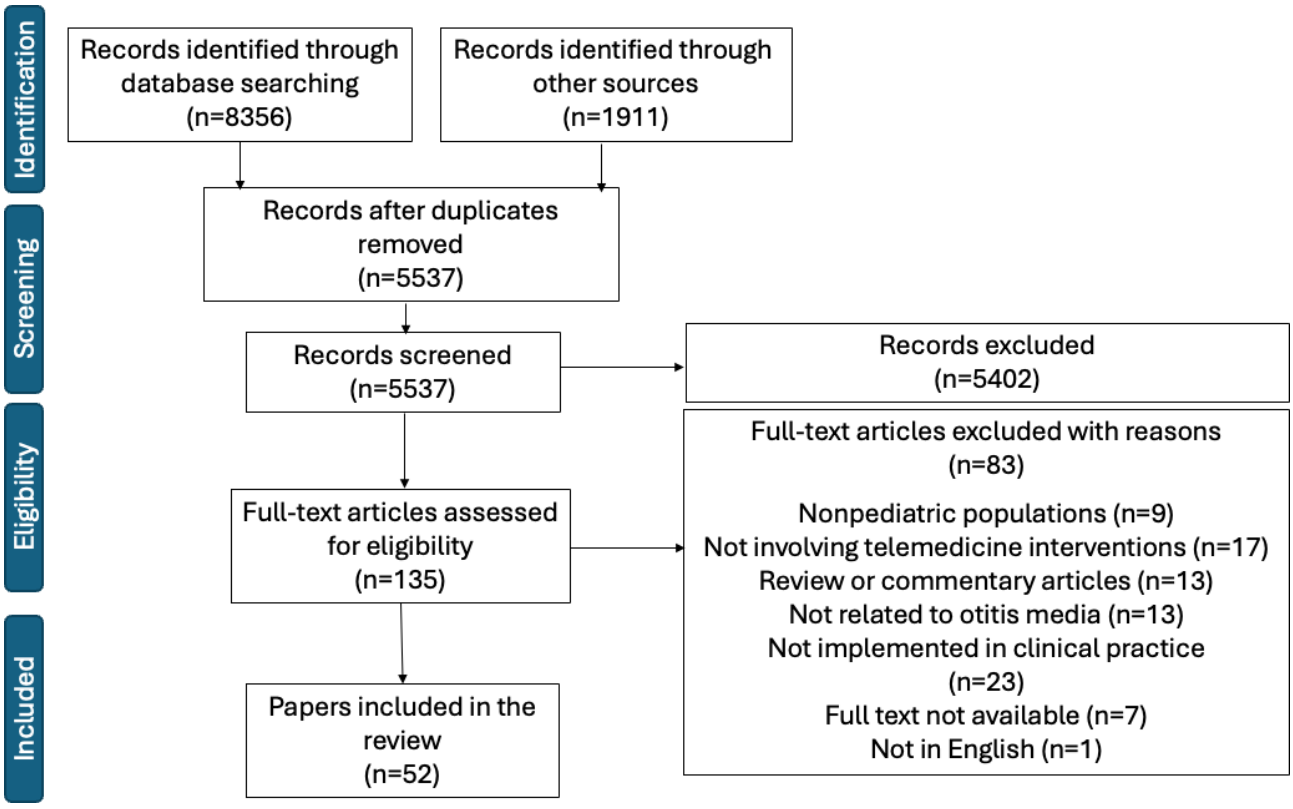

Characteristics of Sources of Evidence

The 52 included studies were published between 2012 and 2026, representing a geographically diverse evidence base encompassing high-, low-, and middle-income settings (summarized in Multimedia Appendix 2) [12-16,22-24,28-71]. QUADAS-2 judgements for diagnostic accuracy studies are summarized in Multimedia Appendix 4. The largest number of included studies (n=15) were conducted in the United States [39,41,42,44,49,52-55,58-60,66,67,71], followed by Australia (n=10) [14,22,23,28,31,32,61,64,65,70], the United Kingdom (n=5) [45,50,51,62,68], India (n=4) [35,40,46,57], and China (n=2) [16,56], with additional studies from other countries, including developing countries (Table 1).

XSL•FO
RenderX

Table . Geographic distribution of included pediatric otitis media telemedicine studies by country (N=52).

Country	Included studies, n (%)[a]	Author (year)
United States	15 (28.8%)	Rappaport et al (2016) [66], Richards et al (2015) [71], Shah et al (2018) [39], Chan et al (2019) [41], Chan et al (2019) [42], Hakimi et al (2019) [44], Jayawardena et al (2020) [49], Kolb et al (2021) [52], Kleinman et al (2021) [53], Don et al (2021) [54], Ni et al (2021) [67], Manayan et al (2022) [55], Schafer et al (2022) [58], Smola et al (2022) [59], Chan et al (2022) [60]
Australia	10 (19.2%)	Smith et al (2012) [28], Phillips et al (2014) [70], Smith et al (2015) [31], Jacups et al (2017) [32], Habib et al (2023) [61], Alenezi et al (2024) [14], Quick et al (2024) [22], Habib et al (2024) [23], Bowers et al (2025) [64], Altamimi et al (2025) [65]
United Kingdom	5 (9.6%)	Fordington and Holland Brown (2020) [45], Cottrell et al (2020) [50], Schuster-Bruce et al (2021) [51], Patel et al (2024) [62], Stancel-Lewis et al (2025) [68]
India	4 (7.7%)	Ramkumar et al (2018) [35], Bhavana et al (2018) [40], Gupta et al (2020) [46], Deshmukh et al (2022) [57]
China	2 (3.8%)	Meng et al (2022) [56], Jin and Fan (2024) [16]
France	2 (3.8%)	Venail et al (2018) [38], Dubois et al (2024) [15]
Nepal	2 (3.8%)	Mandavia et al (2018) [36], Gyawali et al (2024) [63]
South Africa	2 (3.8%)	Biagio et al (2014) [29], Lundberg et al (2017) [34]
Netherlands	1 (1.9%)	Prins-van Ginkel et al (2017) [33]
Brazil	1 (1.9%)	Wagner et al (2023) [13]
Canada	1 (1.9%)	Mousseau et al (2018) [37]
Finland	1 (1.9%)	Erkkola-Anttinen et al (2019) [12]
Greenland	1 (1.9%)	Demant et al (2019) [43]
Hungary	1 (1.9%)	Pannonhalmi et al (2025) [24]
Taiwan	1 (1.9%)	Wu et al (2014) [30]
Turkey	1 (1.9%)	Durgut et al (2020) [48]
Republic of Korea	1 (1.9%)	Cha et al (2020) [47]
Kenya	1 (1.9%)	Yancey et al (2019) [69]

[a]Percentages may not sum to 100% due to rounding.

Designs encompassed diagnostic accuracy comparisons, pragmatic service evaluations, prospective and retrospective cohorts, case series, and education and simulation trials; most studies either enrolled children exclusively or included defined pediatric subsets. Across studies, sample sizes ranged from small pilot cohorts to large-scale community and school-based screening programs.

Results of Individual Sources of Evidence

Telemedicine Models, Users, and Settings

Across the included studies, asynchronous store-and-forward approach was the most commonly used approach, accounting for 35 studies (67.3%; Table 2).

XSL•FO
RenderX

Table . Telemedicine care model classification across included studies on pediatric otitis media (N=52).

Telemedicine model	Included studies, n (%)[a]	Author (year)
Asynchronous[b]	35 (67.3%)	Smith et al (2012) [28], Biagio et al (2014) [29], Phillips et al (2014) [70], Smith et al (2015) [31], Rappaport et al (2016) [66], Prins-van Ginkel et al (2017) [33], Lundberg et al (2017) [34], Mandavia et al (2018) [36], Shah et al (2018) [39], Bhavana et al (2018) [40], Chan et al (2019) [42], Demant et al (2019) [43], Erkkola-Anttinen et al (2019) [12], Yancey et al (2019) [69], Fordington and Holland Brown (2020) [45], Gupta et al (2020) [46], Cha et al (2020) [47], Durgut et al (2020) [48], Cottrell et al (2020) [50], Don et al (2021) [54], Ni et al (2021) [67], Manayan et al (2022) [55], Meng et al (2022) [56], Deshmukh et al (2022) [57], Smola et al (2022) [59], Chan et al (2022) [60], Habib et al (2023) [61], Alenezi et al (2024) [14], Patel et al (2024) [62], Dubois et al (2024) [15], Jin and Fan (2024) [16], Habib et al (2024) [23], Gyawali et al (2024) [63], Altamimi et al (2025) [65], Pannonhalmi et al (2025) [24]
Hybrid[c]	6 (11.5%)	Ramkumar et al (2018) [35], Jayawardena et al (2020) [49], Wagner et al (2023) [13],Quick et al (2024) [22], Bowers et al (2025) [64], Stancel-Lewis et al (2025) [68]
Synchronous[d]	11 (21.2%)	Wu et al (2014) [30], Richards et al (2015) [71], Jacups et al (2017) [32], Mousseau, et al (2018) [37], Venail et al (2018) [38], Chan et al (2019) [41], Hakimi et al (2019) [44], Schuster-Bruce et al (2021) [51], Kolb et al (2021) [52], Kleinman et al (2021) [53], Schafer et al (2022) [58]

[a]Percentages may not total 100% because of rounding.

[b]Asynchronous: capture of images or videos or data with later remote interpretation.

[c]Hybrid: combination of stored media and real-time guidance or interconsultation within the care process.

[d]Synchronous: real-time assessment or interaction (typically video visit or live otoscopy visualization).

In this approach, images or videos were captured by nonphysicians or caregivers and interpreted remotely by otolaryngologists. This includes laptop-tethered video-otoscopy in primary care [29], smartphone otoscopy with consultant review in rural clinics [36], parent-submitted videos in specialty care [39], nurse or technician capture with specialist interpretation in community programs [61], and a multidisciplinary pathway using scheduled remote reviews [14].

Synchronous models accounted for 11 of 52 studies (21.2%) and included real-time postoperative reviews and acute assessments [32,37,52,53,58], whereas blended stored media with live guidance or interconsultation accounted for 6 of 52 studies (11.5%) [13,22,35,49,64,68]. Telemedicine-supported otologic care was implemented across a variety of settings, including primary care clinics, emergency departments, specialty outpatient services, schools, community screening programs, and home environments.

Technologies and Devices

A diverse range of hardware and software solutions were reported across the included studies (Table 3).

XSL•FO
RenderX

Table . Summary of technologies and key outcomes across included pediatric otitis media telemedicine studies (N=52).

References	Implementation/clinical outcome	Diagnostic performance	Typical setting	Representative device/system	Technology category
[12,36,37,41,44,49-51,53,54,71]	Repeat confirmatory examinations decreased from 97.9% to 27.2% when digital otoscopy was used in training clinics [53]. Image acquisition required a median of ~6 minutes per case in rural outpatient workflows [36]. For home tympanostomy tube surveillance, clinicians completed remote reviews in 1 - 3 minutes, families reported mean savings of ~US $127 per episode, and satisfaction was high [54].	Diagnostic concordance with in-person otoscopy was 95%, with κ=0.89, sensitivity 0.94, and specificity 0.96 [36]. In a randomized comparison, trainee diagnostic accuracy improved by 11.2% using digital otoscopy, and interrater agreement rose to Fleiss κ[a]=0.69 versus Fleiss κ=0.40 with a conventional otoscope [53]. Parent-captured videos yielded 40% diagnostic-quality submissions, allowed 87% detection or exclusion of AOM[b] during acute episodes, and achieved κ=0.69 between raters [12].	Emergency, outpatient, school, home; clinicians, trainees, parents	CellScope Oto, Cupris TYM, TYMPA system, endoscope-i	Smartphone-connected digital otoscopes or otoendoscopes
[29,32,34]	Short, focused training improved diagnostic agreement and enabled reliable store-and-forward workflows in primary care [29,34].	Agreement with otomicroscopy was Cohen κ=0.68 - 0.75, with 72.4% - 79.3% sensitivity and 93.2% - 98% specificity [29]. In facilitator-collected video-otoscopy reviewed by a general practitioner, repeat video review improved weighted κ from 0.69 to 0.82 (OMgrade), and specificity reached 0.98 [34].	Primary care; CHWs[c] or GPs[d]	Dino-Lite Pro Earscope	USB video-otoscopes
[42]	Users constructed the funnel in a mean (SD) time of 2.8 (0.9) minutes and rated usability as 8.9/10; performance was consistent across phone models following calibration.	Discrimination of middle ear effusion achieved an AUC[e] 0.898, with 84.6% sensitivity and 81.9% specificity, outperforming commercial acoustic reflectometry (AUC 0.776). Caregivers' classifications closely matched that of clinicians in 24/25 ears [42].	Home; caregivers	Paper funnel + smartphone microphone	Smartphone-based acoustic system

XSL•FO
RenderX

References	Implementation/clinical outcome	Diagnostic performance	Typical setting	Representative device/system	Technology category
[60]	Materials cost was <US $30 while maintaining functional parity with a commercial device; the clinical cohort excluded infants younger than 9 months, defining the initial applicability window.	Classification by 5 audiologists, 86% (2%) agreement; all Type B correctly identified. Bland-Altman in 45 intact ears: peak admittance bias=–0.02 (0.14) mL, peak pressure bias=–1 (17) daPa, and ear canal volume bias=–0.09 (0.25) mL [60].	Outpatient; audiologists	Open-source prototype vs GSI TympStar	Low-cost smartphone tympanometer
[33,45,70]	Adoption was high, with 87% of families downloading the app and 74% using it between visits, and clinicians reported improved information exchange during care [45].	A clinic-based app hearing test correlated with pure-tone audiometry (r=–0.656) [45]. In a cohort comparison, the app arm recorded a higher proportion of symptom days (44%) than the paper diary arm (32.5%) and achieved near-complete questionnaire return [33].	Home; parents/patients	Diary tracking, Hear Glue Ear, Mimi	mHealth applications
[15,16,47]	Systems supported self-examination and automated triage, yet generalizability across sites, devices, and pediatric populations remains limited pending broader external validation.	Internal validation of a CNN[f] model reported sensitivity and specificity >98% with AUROC[g] 1.00 across classes; external testing showed reduced sensitivity for AOM (58.3%) and OME (58.3%) with high specificity (100% and 98.1%) [15]. A low-cost USB device with DenseNet[h] achieved 95.7% overall accuracy; for OME[i], sensitivity 0.77 and specificity 0.99; for cholesteatoma, sensitivity 0.79 [16]. A rule-based model reached Top-1 accuracy of 69.5% versus 83% by human telemedicine reviewers [47].	Mixed otology; clinicians and self-examination users	Karl Storz Smart Scope + i-Nside (Inception-v2), DenseNet USB endoscope, rule-based Android	Artificial intelligence–assisted digital otoscopy

XSL•FO
RenderX

References	Implementation/clinical outcome	Diagnostic performance	Typical setting	Representative device/system	Technology category
[13,22,64]	Image capture success reached 96.3% during remote sessions [22]. School programs achieved >85% coverage and reduced costs relative to standard pathways and were rated highly by families, supporting feasibility for integrated tele-audiology [13,64].	In school-entry children, remote testing matched face-to-face within ±10 dB for air conduction at 1 kHz (98%) and 4 kHz (97.8%), bone conduction matched in 100%/95.7%, otoscopy agreement was κ=0.60, and tympanometry agreement was κ=0.90 [64]. Live video-otoscopy showed substantial agreement (κ=0.68 for videos; κ=0.83 for still images) and identified more landmarks (2.79 vs 1.68) [22]. Using a multisensor kit in pediatrics, ear-specific sensitivity was 64.4% (tympanic membrane) and 66.1% (ear canal), with specificity consistently above 70% [13].	School and pediatric tele-audiology	TytoPro; Otometrics A450 + Aurical Otocam 300 + Interacoustics MT10	Multisensor or hybrid diagnostic kits
[14,65]	Median time to diagnosis and care plan fell to 28 days compared with 450 days in standard care (*P*<.001) [14]. A hospital cost analysis estimated ~US $67 per initial visit (vs US $155) with a break-even point at 223 appointments and about 3-fold greater specialist throughput [65].	Performance was driven by the integrated, multitest pathway rather than a single diagnostic metric [14,65].	Multidisciplinary teleclinic	"Ear Portal": HearScope + HearScreen + Titan (WBT[j]) + DPOAE[k]	Integrated multidisciplinary pathway
[28,31,43,46,57,69]	In a 45-camp program, 3000 individuals were screened; 54% were referred, 13% presented to hospital, and roughly half of presenters required surgery; procedures were delivered at about 40% of usual cost [46]. Longitudinal regional programs achieved ~85% coverage and reduced referrals and waits [28,31].	In CHW-led screening, overall sensitivity was 96.9% with condition-specific concordance of CSOM[l] 96.2%, wax 98%, ASOM[m] 93.9%, foreign body 100%, and OME 100% [57]. When operators received only written instructions, diagnostically useful video yield was 18.1% with substantial interrater agreement (Fleiss κ=0.67) and higher usefulness in older children [43].	Community/schools; CHWs, teachers	ENTraview + ClickMedix; Mebird T5 video-otoscopy + EMRs[n] support follow-up	Community and school screening programs

XSL•FO
RenderX

References	Implementation/clinical outcome	Diagnostic performance	Typical setting	Representative device/system	Technology category
[48,55]	High false-positive rates led to a 66% referral rate in a cohort with 27% hearing loss prevalence; performance was limited by ambient noise and lack of standardized calibration [55].	The Hearing Test app showed κ=0.059, indicating poor agreement with conventional audiometry [48]. The hearScreen app achieved 85% sensitivity and 41% specificity, with ambient noise at 500 Hz significantly increasing false positives (*P*=.01) [55].	Clinics/non-soundproof rooms	Hearing Test, hearScreen	App-based hearing screening

[a]κ: Cohen kappa.
[b]AOM: acute otitis media.
[c]CHW: community health worker.
[d]GP: general practitioner.
[e]AUC: area under the receiver operating characteristic curve.
[f]CNN: convolutional neural network.
[g]AUROC: area under the receiver operating characteristic curve.
[h]DenseNet: Dense Convolutional Network.
[i]OME: otitis media with effusion.
[j]WBT: wideband tympanometry.
[k]DPOAE: distortion product otoacoustic emissions.
[l]CSOM: chronic suppurative otitis media.
[m]ASOM: acute suppurative otitis media.
[n]EMR: electronic medical record.

The most frequently used solution was smartphone-connected digital otoscopes and otoendoscopes, such as CellScope Oto, Cupris TYM, TYMPA system (TympaHealth), and endoscope-i, which were applied in emergency, outpatient, school, and home settings [11,36,37,41,44,49-51,53,54,71]. Earlier clinical evaluation of smartphone-enabled otoscopy for OM diagnosis and management were reported, supporting feasibility in routine care [66]. USB video-otoscopes, including the Dino-Lite Pro Earscope, were used to support store-and-forward apps in clinics [29,34]. Non-smartphone digital otoscopes, such as Welch Allyn Digital MacroView Otoscope, TeleHealth Flexican Otoscope, and Welch Allyn USB Otoscope, were used for real-time follow-up and teleconsultations [22,32].

Apart from imaging, several digital tools were used. A smartphone-only acoustic system, using a paper funnel and the phone's microphone, enabled the detection of middle ear effusion (MEE) [42]. Mobile health (mHealth) apps facilitated diary-based episode detection and home management [33,45]. By contrast, although acceptability was high, a culturally tailored multimedia or text messaging intervention for Aboriginal children with tympanic membrane perforation did not improve clinic attendance or short-term ear health outcomes [70]. A low-cost, open-source smartphone tympanometer was evaluated against the GSI TympStar systems (Grason-Stadler) [60]. Decision support innovations included the Karl Storz Smart Scope with the i-Nside app (Inception-v2) and an inexpensive (~US $10) USB endoscope integrated with a Dense Convolutional Network (DenseNet) classifier for self-examination [15,16]. A consumer Android-tethered otoendoscope, coupled with a rule-based model, demonstrated performance comparable to that of clinicians in diverse otology settings [47]. Community programs frequently used Medtronic's ENTraview integrated with ClickMedix, and 1 program used the Mebird T5 SO (Black Bee Intelligent Manufacturing [Shenzhen] Technology Co., Ltd.) for asynchronous case discussion [35,46,56,57]. A multisensor kit (TytoPro) enabled remote pediatric interconsultation [13]. The school-based hybrid audiology program integrated Madsen A450 audiometry (Otometrics Natus Medical), Natus Aurical Otocam 300 video-otoscopy, and Interacoustics MT10 tympanometry with videoconferencing [64]. For training, Otosim2 (OtoSim Inc.) via AudioProConnect and a low-cost KZYEE USB otoscope improved teaching and landmark recognition [38,44]. Teslong pen endoscopes and the Mimi Hearing Test app (Mimi Hearing Technologies) were piloted for supervised home capture and hearing checks [49,63]. A structured "Ear Portal" pathway combined hearX HearScope video-otoscopy, HearScreen audiometry, Interacoustics Titan wideband tympanometry, and distortion product otoacoustic emissions into a scheduled multidisciplinary review [14,65].

Technology Category × Care Model

Figure 2 maps technology categories against care models. Most studies used asynchronous store-and-forward workflows, primarily tele-otoscopy imaging alone or tele-otoscopy combined with adjunct testing (Figure 2). Hybrid models were less common and tended to involve multisensor kits or tele-otoscopy plus adjunct testing. Notable gaps included limited evidence for integrated pathway models beyond a small number

XSL•FO
RenderX

of programs and sparse evaluation of outcomes beyond diagnostic performance, including long-term child outcomes and cost-effectiveness.

Figure 2. Evidence gap map of technology category by care model in pediatric otitis media telemedicine. mHealth: mobile health.

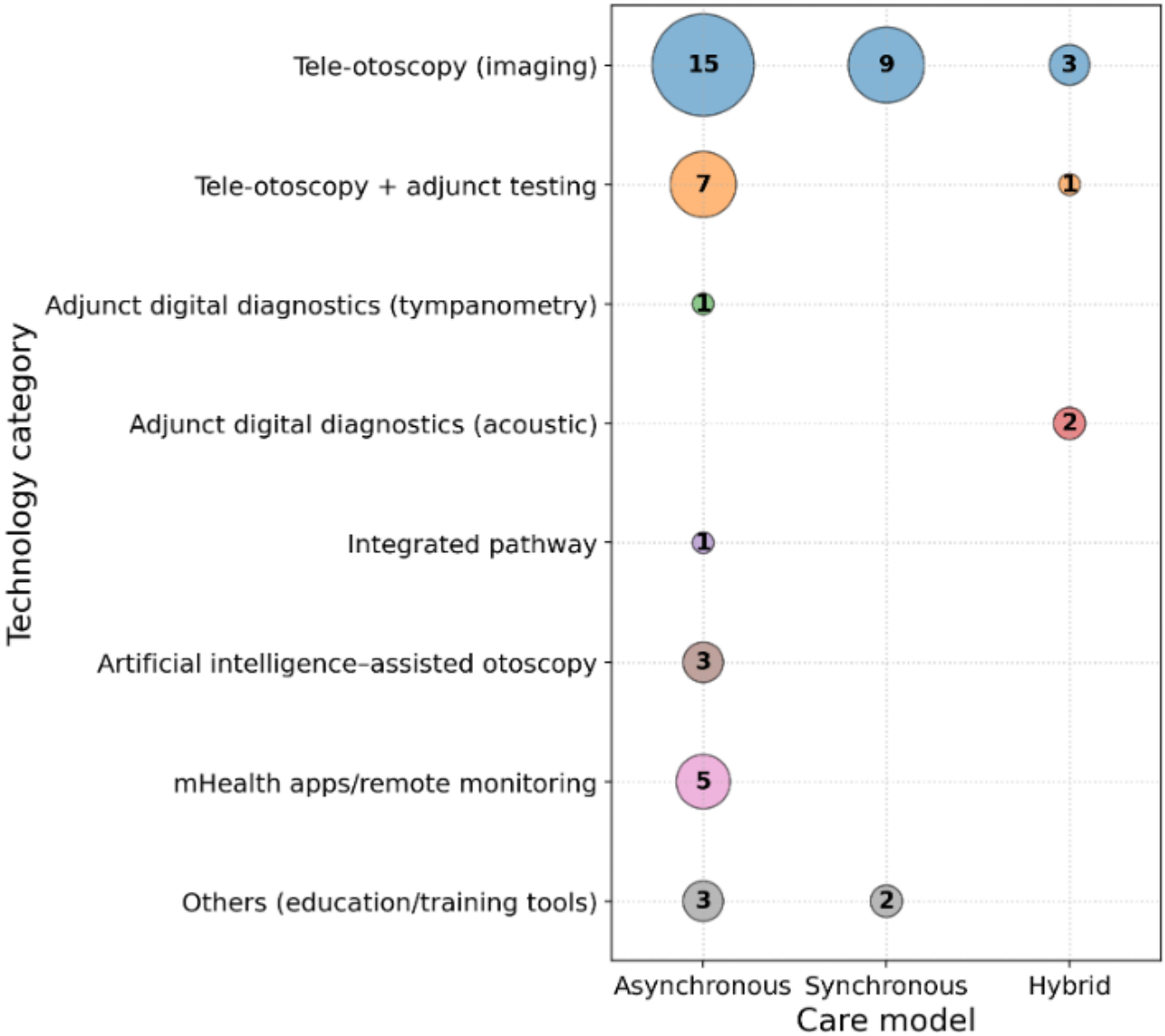

Operators and Workflow Roles

Figure 3 summarizes who captured otoscopic media (clinicians, nurses/community facilitators, caregivers/parents) and who interpreted it. Studies consistently indicated that diagnostic yield depended strongly on operator training and workflow support, which includes checklists, feedback loops, and cerumen management. Evidence gaps included limited direct comparisons of training intensity and limited reporting of minimum image quality thresholds.

XSL•FO
RenderX

Figure 3. Operator workflow map showing who captures otoscopic media and who interprets it in pediatric otitis media tele-otoscopy. CHW: community health worker, ENT: ear, nose, and throat.

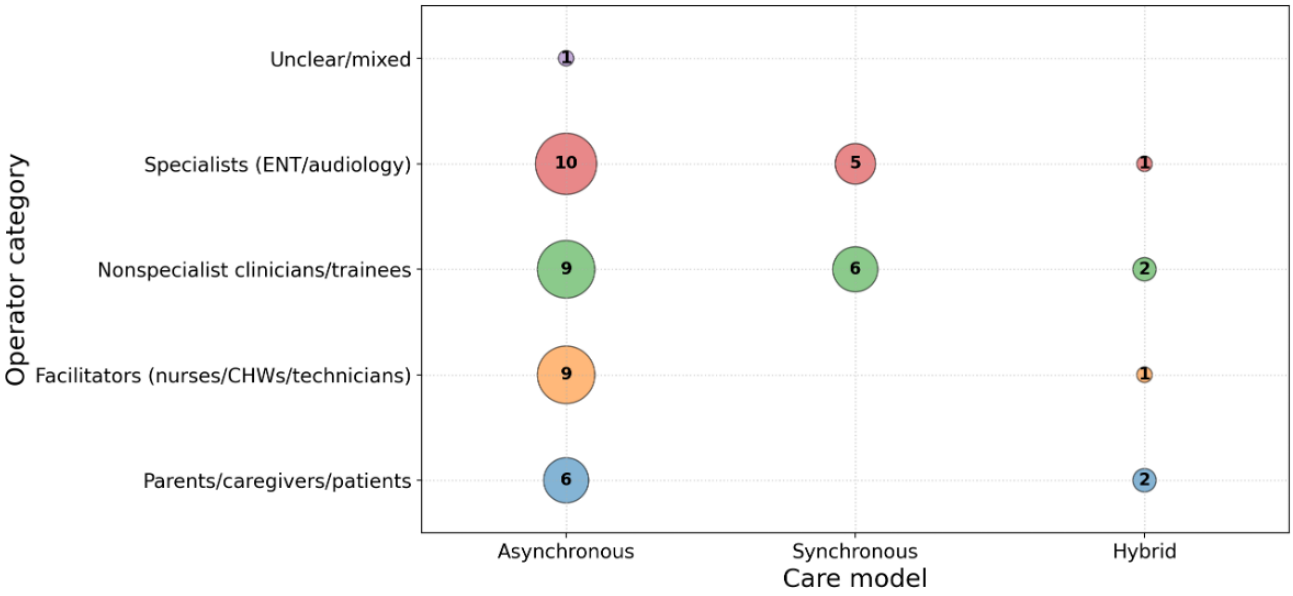

Diagnostic Performance

Across diagnostic accuracy studies, the reported κ, sensitivity, and specificity varied by key sources of clinical and methodological heterogeneity, including pediatric age strata, OM phenotype definitions (AOM vs OME vs mixed), reference standards, operator training and experience, and device and workflow-specific image quality constraints (Table 4). To improve interpretability, we summarized diagnostic performance by capture-interpretation model and reference standard.

XSL•FO
RenderX

Table . Key sources of clinical and methodological heterogeneity across diagnostic accuracy studies in pediatric otitis media tele-otoscopy.

Study and category	Key heterogeneity						Performance (κ[a]/Se[b]/Sp[c]/AUC[d])
	Age	Spectrum	Operator/training	Reference standard	Device	Quality	
(A) Trained operator/expert interpretation							
Biagio et al (2014) [29]	2-16 y	Normal/abnormal OM	Trained capture → clinician read (2 sessions)	Otomicroscopy	Asynchronous video-otoscopy	Acceptable/excellent ratings (NR criteria)	κ 0.68 - 0.75; Se 72.4% - 79.3%; Sp 93.2% - 98%
Lundberg et al (2017) [34]	Pediatrics (NR)[e]	Normal[f]/AOM[g]/CSOM[h] (abnormal pooled)	Facilitator capture → GP[i] read (repeat)	Otologist otomicroscopy	Video-otoscopy vs otoscopy	NPD[j] due to wax/degree of cooperation/low quality	wκ[k] 0.76→0.82 (video) vs 0.69 (otoscopy); Se 0.77 - 0.81; Sp 0.94 - 0.98
Mandavia et al (2018) [36]	Pediatrics (NR)	Screening (NR)	Trainees capture → remote consultant read	On-site specialist otoscopy/referral	Smartphone otoscopy (Cupris TYM)	NR	Concordance 95% (κ=0.89); Se 0.94; Sp 0.96; referral concordance 100%
(B) Parent/non-specialist capture (training dependent)							
Erkkola-Anttinen et al (2019) [12]	6-35 mo	AOM-focused	Parent capture (hands-on teaching) → remote raters	Clinician assessment (NR)	Device: smartphone video-otoscopy	Sufficient quality 62% - 64% during intervention; diagnostic quality 40%	Detect/exclude AOM κ=0.69 (87% possible)
Shah et al (2018) [39]	3 mo to 17 y	Mixed pathology	Parent or physician capture → physician read (brief tutorial)	Pneumatic otoscopy (± tympanometry subset)	iPhone/CellScope	Wax common; wax-alert κ≈0.01	Parent κ=0.42; physician κ=0.74 (vs pneumatic otoscopy)
Demant et al (2019) [43]	6-72 mo	Usefulness outcome	Nonspecialist HW capture (written instructions only) → ENT[l] raters	NR	Smartphone otoscopy (Cupris TYM)	Useful videos 6.6% (<24 mo), 22.7% (24-48 mo), 30.7% (>48 mo)	Modified Fleiss κ 0.73 (<24 mo), 0.48 (24-48 mo), 0.60 (>48 mo)
(C) Different outcomes/surrogate standards/different modalities							
Kleinman et al (2021) [53]	Pediatrics (NR)	Education/clinical trial	Trainees	Education/supervisor (NR gold standard)	Digital vs traditional otoscope	NR	Accuracy +11.2%; Fleiss κ 0.40→0.69; repeat confirmatory examinations 97.9%→27.2%
Rappaport et al (2016) [66]	6 mo to 18 y	AOM via ED[m] diagnosis (ICD-9[n] surrogate)	ED capture → blinded physician reviewers	ED diagnosis	Smartphone vs digital otoscope	NR	Overall κ=0.375; smartphone κ=0.368; digital κ=0.381
Chan et al (2019) [42]	18 mo to 17 y (median 5 y)	Chronic OME/recurrent AOM vs controls	Clinician testing;parental use tested after brief demo	Surgical fluid (myringotomy) and/or ENT pneumatic otoscopy	Smartphone acoustic method (paper funnel; ML)[o]	Phone model calibration	AUC 0.898; Se 84.6%; Sp 81.9%; parents matched clinician 24/25 ears

XSL•FO
RenderX

Study and category	Key heterogeneity						Performance (κ[a]/Se[b]/Sp[c]/AUC[d])
	Age	Spectrum	Operator/training	Reference standard	Device	Quality	
Durgut et al (2020) [48]	5-15 y	OME (50 children; 88 ears with OME + 12 controls)	In-clinic testing	Booth pure-tone audiometry	Android Hearing Test app	Environmental noise limitation	Very poor agreement: κ=0.059 (100 ears); no significant correlation 500 - 4000 Hz

[a]κ: Cohen kappa.
[b]Se: sensitivity.
[c]Sp: specificity.
[d]AUC: area under the receiver operating characteristic curve.
[e]NR: not reported.
[f]OME: otitis media with effusion.
[g]AOM: acute otitis media.
[h]CSOM: chronic suppurative otitis media.
[i]GP: general practitioner.
[j]NPD: not possible to diagnose.
[k]wκ: weighted kappa.
[l]ENT: ear, nose, and throat.
[m]ED: emergency department.
[n]ICD-9: International Classification of Diseases, 9th Revision.
[o]ML: machine learning.

Trained Operator Capture with Expert Interpretation Against Clinical Reference Standards

When images were captured by trained operators and interpreted by experienced clinicians, store-and-forward tele-otoscopy showed substantial agreement with in-person otomicroscopy. In a primary care cohort, asynchronous video-otoscopy yielded Cohen κ values of 0.68 - 0.75 across raters and 2 review sessions, 72.4% - 79.3% and specificities of 93.2% - 98% [29]. In another clinic, repeated readings improved weighted κ from 0.69 to 0.82, with specificity up to 0.98 [34]. Smartphone otoscopy used by trainees, with remote consultant review, achieved 95% ear-level concordance with on-site specialist otoscopy (κ=0.89; sensitivity 0.94; specificity 0.96) and 100% concordance with referral decisions [36]. For training, pediatric emergency department residents demonstrated similar diagnostic accuracy when using a smartphone device compared with a traditional otoscope (0.74 vs 0.69; nonsignificant). However, novice users tended to favor the conventional tools, as they felt confident with them [37]. A separate academic comparison showed that digital otoscopy improved trainee accuracy by 11.2%, increased interrater agreement (from Fleiss κof 0.40 to 0.69), and reduced repeat confirmatory examinations from 97.9% to 27.2% [53].

Parent and Nonspecialist Capture (Training- and Cooperation-Dependent Performance)

Parent and nonspecialist performance was highly dependent on training and cooperation. After structured teaching, parents produced videos of sufficient quality in 62% - 64% of submissions during the intervention periods, and the evaluators showed substantial agreement for detecting or excluding AOM (κ=0.69) [12]. In contrast, a brief tutorial yielded lower agreement with pneumatic otoscopy (parent κ=0.42 vs physician κ=0.74), while automated cerumen alerts showed negligible agreement (κ=0.01), highlighting the advantage of structured teaching [39]. Without hands-on training, local health care workers produced diagnostically useful videos in only 18% of cases, with insertion errors and cerumen as major barriers; usefulness increased proportionally with child age [43]. Educational approaches also included a smartphone-adaptable video-otoscopy quiz developed to support otoscopic image interpretation skills, reflecting ongoing efforts to standardize competency development for tele-otoscopy workflows [67].

Alternative Modalities and/or Surrogate Outcomes (Interpret with Caution)

Beyond optical otoscopy workflows, a smartphone-based acoustic method to detect MEE achieved an area under the receiver operating characteristic curve (AUC) of 0.898, with approximately 85% sensitivity and 82% specificity under surgical and/or specialist reference standards, and caregiver performance approximated clinician performance in a subset [42]. A smartphone tympanometer also demonstrated approximately 86% classification agreement with commercial devices, with only minor bias in peak admittance and pressure, supporting the feasibility of adjunct middle ear assessments beyond optical imaging [60]. In contrast, a pediatric hearing test app demonstrated very poor agreement with booth audiometry (κ≈0.06) [48,55]. These findings underscore that performance estimates are not directly comparable across studies when index test modality, outcome definition, and reference standards differ. Artificial intelligence–supported classifiers have shown high performance on curated datasets but lower

XSL•FO
RenderX

sensitivity for OM and limited external validation across devices and sites [15,16,47].

Clinical and Implementation Outcomes

Comparison With Traditional Approaches

Tele-enabled programs have broadened access and accelerated clinical decision-making. A school or community program completed 2111 assessments in 1053 children, demonstrating 85% coverage. In another evaluation, it reduced waiting times and tertiary referrals as the program expanded [28,31]. A general practitioner-to-ear, nose, and throat (ENT) specialist smartphone endoscopy pathway was able to manage 83% of the 53 referrals remotely, with a median specialist response time of 48 (IQR 22-148) minutes [50]. In a pediatric emergency department, a smartphone otoscope showed substantial within-physician agreement versus a traditional otoscope and changed the reported tympanic membrane view in 12%‑16% of examinations, including clinically relevant diagnostic changes to or from AOM in 6%‑7% of examinations [71]. Community screening of 3000 individuals generated 54% of referrals, but only 13% were ultimately presented for hospital evaluation; nearly half of those evaluated required surgery, underscoring the barriers of pathway follow-through [46]. A telehealth-supported surgical model achieved 80% clinical resolution, 88% hearing improvement, and reduced travel costs for postoperative review [32]. Home tympanostomy tube surveillance demonstrated high interrater and intrarater agreement, rapid clinician review (1‑3 min), and high family satisfaction with reported time and cost savings [54]. The Ear Portal pathway substantially shortened time to diagnosis and care plan (median 28 [IQR 19.8] days), and an economic analysis showed that it lowered per-visit costs, reached a break-even point after ~223 appointments, and increased specialist throughput [14,65]. A community service integrating smartphone otoscopy with tablet audiometry discharged most patients at their first visit at a lower cost than other services using conventional methods, with specialist review rarely leading to changed decisions [62]. However, televisits, which were performed predominantly without ear imaging, produced similar rates of tube recommendations but substantially lower documentation of effusion at surgery compared with encounters that incorporated ear imaging, indicating a potential risk of suboptimal decision-making when otoscopy is omitted [52]. During the COVID-19 pandemic, telehealth encounters were overall less likely to end with tube recommendations, underscoring the limitations imposed by restricted physical examination [58]. A multisensor interconsultation device demonstrated high agreement for several systems but lower sensitivity for ear canal and tympanic membrane findings than conventional methods, reinforcing the necessity of dedicated otoscopy using generalist kits [13]. In pediatric telehealth, live video-otoscopy provided superior image quality and landmark visualization compared to still images, supporting synchronous visualization when feasible [22]. A school-entry hybrid model demonstrated close agreement with face-to-face audiometry and tympanometry, with acceptable agreement for otoscopy [64].

Outcome Domains Reported

Table 5 summarizes the distribution of outcome domains reported across included studies (N=52). Diagnostic performance was the most frequently reported domain (38/52, 73.1%), followed by implementation and usability outcomes (29/52, 55.8%). Clinical and management impact (22/52, 42.3%) and audiology and tympanometry outcomes (19/52, 36.5%) were reported less often, while image or capture quality was reported in 15 of 52 studies (28.8%). Training and education and economic outcomes were each reported in 8 of 52 studies (15.4%), and artificial intelligence or automation was reported in 4 studies (7.7%; Table 5). Reports of longer-term patient outcomes and adverse events were uncommon, indicating priorities for future research.

Table . Outcome domains reported across included studies (N=52).

Outcome domain	Studies reporting domain, n (%)[a]
Diagnostic performance	38 (73.1%)
Implementation/usability	29 (55.8%)
Clinical/management impact	22 (42.3%)
Audiology/tympanometry outcomes	19 (36.5%)
Image/capture quality	15 (28.8%)
Training/education	8 (15.4%)
Economic outcomes	8 (15.4%)
Artificial intelligence/automation	4 (7.7%)

[a]Studies may contribute to multiple domains; therefore, percentages do not sum to 100%.

Applications and Implementation Feasibility

Tele-enabled tools were applied in 4 management scenarios for pediatric ear diseases. For AOM diagnosis, asynchronous video-otoscopy demonstrated substantial agreement with otomicroscopy when images were captured by trained operators, as reported in South Africa using a laptop-tethered Dino-Lite Pro Earscope and specialist review [29] and in a primary care study in which repeated readings of stored video sequences closely tracked the microscopic reference [34]. At the bedside, smartphone-based digital otoscopy (CellScope Oto) demonstrated accuracy comparable to, or modestly better than,

XSL•FO
RenderX

that of trainees using a conventional otoscope in a pediatric ED [37]. With structured instructions, parents successfully captured diagnostically useful otoscopy videos at home, supporting remote triage for suspected AOM [12]. For posttympanostomy tube surveillance and postoperative follow-up, home digital otoscopy with CellScope Oto enabled asynchronous specialist review and was rated highly acceptable by families. In remote Australian communities, a telehealth-supported surgical pathway cut travel and costs for postoperative care [32,54]. For screening and triage in underserved settings, school- and community-based programs—ranging from mobile ear-screening services in Queensland to large-scale campaigns using the ENTraview smartphone otoscope integrated with the ClickMedix platform—efficiently detected pathology, shortened time to specialist input, and improved surgical case prioritization, although barriers such as transport and out-of-pocket costs limited follow-through [28,46]. Additionally, a recent parent-led home ear health check pilot targeting children with complex needs highlighted feasibility and potential access benefits for populations facing barriers to clinic-based otoscopy [68].

Although feasibility was consistently high across settings and devices, the diagnostic yield depended on operator training, child cooperation, and cerumen management. Parents performed better after structured teaching, but accuracy declined with only brief instructions or when wax removal was not addressed [39]. Hands-on practice with a live tele-otoscopy simulation using Otosim2 with the AudioProConnect platform preserved accuracy and shortened task time, whereas written-only instructions under field conditions produced a low proportion of diagnostically useful videos [43]. Clinicians generally rated smartphone otoscopy imagery as adequate and highlighted supervisory advantages of shared visualization, including fewer repeat confirmatory examinations during training encounters [41,53]. Conversely, environmental noise and calibration issues materially influenced app-based hearing tools, which demonstrated poor agreement with booth-based audiometry in children with OME, and a smartphone screening app demonstrated high sensitivity but low specificity when ambient noise was not well controlled [48,55].

Risk of Bias and Study Limitations

Limitations include small, single-site samples; convenience sampling; mixed adult and pediatric cohorts in some studies; lack of blinded reference standards; heterogeneous operator training; device or discontinuation issues; and industrial involvement in some implementations. Parent and nonspecialist performance was sensitive to training and cerumen management, while omitting otoscopy in televisits risked misclassification of MEE [52].

Key Practical Takeaways

Across the included studies, the following points were notably identified as key practical takeaways:

1. Diagnostic performance depended strongly on who captured the otoscopic media and on the training and quality protocol rather than on the device alone.

2. Predefined image quality thresholds, multiview capture, and repeat capture when needed materially affected interpretability and downstream clinical decisions.

3. Telehealth encounters conducted without otoscopic imaging risk underdetecting middle ear pathology, including effusion, and may therefore alter management decisions when objective findings are required.

4. Hub-based or hybrid workflows—such as facilitated capture with remote review—appear to support access and feasibility while maintaining diagnostic reliability in many settings.

5. Evidence remains limited regarding scale-up outcomes (including sustainability/maintenance), formal economic evaluations (particularly in lower-middle-income countries), and external validation of artificial intelligence–enabled tools across devices and sites.

Discussion

Principal Findings

This scoping review mapped a rapidly evolving evidence base for telehealth in pediatric OM. Three consistent signals were observed. First, when images were captured using standardized techniques by trained operators and interpreted by experienced clinicians, store-and-forward tele-otoscopy demonstrated diagnostic performance similar to in-person examination. In primary care and community clinics, agreement with otomicroscopy was frequently in the substantial range, with high specificity for excluding the disease [29,34,36]. Second, operator training proved pivotal: performance declined markedly with brief, unsupervised capture by parents or lay workers; agreement with pneumatic otoscopy was low; and fewer than 20% of videos were diagnostically useful [43]. However, structured instruction or simulation improved the performance of workers without hands-on training [38,39,43]. Example of practical structured training can remain minimal, such as a 5- to 10-minute tutorial, a 1-page capture checklist, and brief hands-on coaching including cerumen management. Embedding a simple feedback loop on image quality further improves yield without adding substantial workload. Third, adjunct digital tools are promising: a smartphone-only acoustic method that infers tympanic membrane mobility detected MEE with the AUC approaching 0.90, whereas digital otoscopy platforms improved trainee-supervisor agreement and reduced duplicate examinations [42,53]. The contribution of this review is an implementation-facing synthesis that links “which tool,” “which care model,” and “under what conditions” to observed diagnostic performance, thereby informing scalable hybrid workflows.

Interpretation and Implications

Diagnostic Performance and Care Models

When image acquisition was standardized and performed by trained facilitators, remote specialists and generalists demonstrated diagnostic agreement with in-person reference standards that are clinically feasible for triage and management in many settings [29,34,36,50]. Programs embedding store-and-forward video-otoscopy, tympanometry, and audiometry in school- and community-based screening have

XSL•FO
RenderX

demonstrated high coverage, efficient specialist sorting, reduced wait times, and identified surgical candidates [28,31,64]. In primary care, asynchronous tele-referral pathways allowed most cases to be managed without face-to-face review while offering timely specialist input and educational feedback to referrers [50].

Simultaneously, telehealth visits performed without otoscopic visualization carry a risk of misclassification. In a pandemic-era cohort, children evaluated for RAOM via video-only visits were less likely to have effusion confirmed at the time of tympanostomy tube placement than office-based comparators (39% vs 79%), highlighting that history alone is insufficient for surgical decision-making [52]. These findings support hybrid models—telehistory plus rapid access to digital otoscopy (clinic kiosk, drive-through nurse station, and community facilitators) or validated home solutions—especially when management depends on objective middle ear findings [52,54,68].

Operators, Training, and Human Factors

Evidence has consistently demonstrated that the person capturing the image is as critical as the device used. Across included studies, capture quality and diagnostic yield varied substantially by operator type, training intensity, child cooperation, and cerumen burden [12,39,43]. Hands-on coaching and structured capture approaches were associated with higher proportions of interpretable recordings in parents and nonspecialists, whereas written-only or brief instruction yielded lower rates of diagnostically useful media in field settings [12,39,43]. In simulation settings, trained facilitators using Otosim2 with the AudioProConnect platform maintained expert accuracy without time penalties [38]. In clinical supervision settings, digital otoscopy improved trainee accuracy by approximately 11 percentage points and increased trainee-supervisor agreement from kappa 0.40 to 0.69, while reducing confirmatory reexaminations from ~98% to ~27% [53]. A randomized comparison in pediatric emergency department demonstrated similar accuracy for smartphones versus conventional otoscopy (0.74 vs 0.69) but revealed lower confidence among inexperienced users, emphasizing the value of familiarization [37]. Smartphone otoscopy may meaningfully affect clinical impressions. In a pediatric emergency department study, CellScope Oto altered the reported view in 12% to 16% of exams, with clinically relevant diagnostic changes in approximately 6%; notably, it was rated by clinicians as easy to use and valuable for teaching [71].

Given the established effectiveness of video-otoscopy, careful consideration of the methods and timing of use may enable promising integration [72]. Standardized training packages and minimum quality thresholds should be considered prerequisites for scaling, particularly for parents, community health workers, and very young children, in whom cooperation and ear canal size may complicate capture [12,39,43].

Adjunct Technologies and Artificial Intelligence–Based Decision Support

Smartphone-only acoustic reflectometry with embedded machine learning demonstrated high accuracy for effusion detection and performed comparably across platforms and even in parents' hands [42]. This suggests a low-barrier screening pathway that does not depend on optical access, which is particularly valuable in toddlers with narrow ear canals or when video quality is suboptimal [42,43]. Early studies of automated image classification demonstrated throughput advantages, but clinical readiness remains limited due to scarce external validation across devices, sites, and pediatric populations and incomplete reporting of training datasets and labeling methods, which constrains generalizability and risk-of-bias assessment [15,16,47]. Although numerous models were developed using artificial intelligence to classify middle ear diseases on the basis of tympanic membrane findings or short-wave infrared imaging, there were few studies addressing its actual clinical implementation [15,16,73-76]. It is therefore necessary to consider not only the accuracy of these models but also the optimal point and method of integration within the clinical workflow [15,47]. Most studies also provide limited information on deployment pathways, including regulatory considerations, governance for performance monitoring, and safeguards when image quality is insufficient, and direct comparisons with human experts under identical capture conditions remain uncommon [15,47]. Accordingly, priority next steps include multisite, device-agnostic external validation using real-world pediatric data and prospective evaluations of workflow-level impact and safety rather than reliance on curated test-set metrics alone. In parallel, self-management apps for OME demonstrated high acceptability and reasonable correlation with clinic audiometry, supporting family engagement and between-visit monitoring rather than diagnosis [45].

Safety, Stewardship, and Workflow

Across randomized and comparative studies, smartphone otoscopy did not increase AOM diagnosis or antibiotic prescription in emergency settings compared to conventional otoscopy, and in some trainee contexts, it was associated with fewer prescriptions, likely due to better shared visualization and supervision [41,53]. Programs that integrate remote reviews into perioperative pathways have demonstrated high postoperative success rates and measurable cost savings for families in remote regions [32,54]. Collectively, these findings support protocolized tele-otoscopy within antibiotic stewardship approaches and access pathways that do not rely exclusively on family-owned smartphones, such as facilitated capture in schools, community programs, and clinics [28,29,34,61,64]. Additionally, remote hearing screening initiatives, when paired with follow-up during rehabilitation, may integrate synergistically with OM care, ultimately contributing to comprehensive pediatric ear health services [64,77-79]. In this review, reporting on governance and medicolegal arrangements was uncommon, representing a gap in implementation evidence. Implementation should include explicit consent, minimum image quality thresholds with default to in-person examination when unmet, and red-flag escalation [52,58,72]. Future studies should report on adverse events and medicolegal frameworks.

Equity and Access

Tele-otoscopy is particularly valuable in indigenous and rural settings with few specialists, where school-based or outreach models led by local health workers can achieve over 80%

XSL•FO
RenderX

screening coverage and facilitate timely surgical intervention [28]. Jacups et al [80] reported that the telehealth model was the most cost-effective method for providing ENT surgery to children living in remote regions, primarily by reducing patient travel. However, equity implications depend strongly on implementation design, including whether capture is facilitated and whether connectivity and device requirements are minimized [28,43,50]. Not all low-burden mHealth interventions translate into short-term clinical improvements; for example, multimedia or text messaging alone did not increase clinic attendance or improve ear outcomes over 6 weeks in a high-risk remote population, despite good acceptability [70]. Digital divide factors—device availability, connectivity and bandwidth requirements, caregiver digital and health literacy, and accessibility beyond geography (including language support and cultural adaptation)—were infrequently reported or measured systematically across included studies [43,50]. Age-related constraints are also salient in pediatrics, with lower capture quality reported in children under 24 months, which can shift burden onto caregivers when programs rely primarily on home capture [43]. Care model choice has direct equity consequences. In lower-middle-income countries, implementation models pairing screening with video-otoscopy and electronic medical record capture can broaden the scope beyond hearing thresholds alone by identifying otologic pathology, even among children with normal hearing, while enabling structured follow-up workflows [69]. Hub-based or facilitated capture workflows (community-hosted capture hubs or school programs) can reduce reliance on family-owned devices and high-bandwidth connections while enabling standardized training and explicit image quality thresholds [28,29,34]. In contrast, home-based models can improve convenience but may disproportionately disadvantage families facing barriers in time, literacy, language, connectivity, or device access unless hardware and support are provisioned [43,68].

Therefore, equity-focused deployment should prioritize community-hosted capture hubs, offline-capable apps, and hardware provision [28,43,50]. In pediatrics, careful consideration of caregivers' role is essential, as their involvement can influence capture quality and follow-through [12,65]. Importantly, few included studies conducted formal health equity analyses or reported outcomes stratified by socioeconomic status, geography, or ethnicity, limiting conclusions on differential benefits across populations and underscoring the need for equity endpoints and subgroup reporting in future evaluations [43,50].

Limitations

This scoping review emphasized breadth rather than quantitative synthesis and therefore did not include a meta-analysis; however, we conducted a targeted QUADAS-2 appraisal for diagnostic accuracy studies. Limitations included the enrollment of mixed-age cohorts, lack of blinded reference standards, and the rapid evolution of device ecosystems, which may reduce the generalizability of device-specific findings. Additionally, language and publication restrictions may have led to the exclusion of relevant gray literature. Although we searched multiple bibliographic databases and used both controlled vocabulary and free-text terms, relevant studies may have been missed due to terminology variation across disciplines and incomplete indexing of emerging digital health tools. Future studies should prioritize large, multicenter evaluations that incorporate standardized blinded reference standards to strengthen diagnostic accuracy across diverse pediatric populations and care settings. External validation efforts must address varied device ecosystems and real-world image quality to ensure generalizability. Comparative research is warranted to establish the relative performance and cost-effectiveness of emerging digital otoscopy, acoustic, and decision support tools. This scoping review protocol was not prospectively registered. Core methodological decisions were defined a priori; minor refinements to screening guidance were made during pilot screening to improve consistency. Although scoping reviews do not always require registration, the absence of prospective registration may reduce transparency and should be considered when interpreting the findings. In addition, interpretation of diagnostic accuracy is constrained by substantial clinical and methodological heterogeneity across studies, which limits cross-setting comparability of κ, sensitivity, and specificity. Equity-relevant reporting was uncommon, restricting conclusions on differential benefits across populations. Finally, while implementation outcomes were frequently described, most studies did not evaluate sustainability or scale-up using an explicit implementation framework, and evidence for long-term maintenance remains limited.

Future Research Agenda

Based on mapped gaps across technologies, care models, and operators, priority research directions include (1) multisite pragmatic evaluations with standardized reporting of pediatric age strata, OM phenotype, reference standards, operator training content and competency, and explicit image quality thresholds; (2) comparative evaluation of hub-based facilitated capture versus home-based capture models, including equity-relevant endpoints and subgroup analyses; (3) workflow-focused outcomes beyond accuracy; (4) sustainability and scale-up studies using an implementation framework, including maintenance, governance, and medicolegal arrangements; and (5) for artificial intelligence–enabled tools, external validation across devices or sites with real-world image quality, transparent dataset description, and prospective evaluation of clinical impact within the care pathway rather than standalone discrimination metrics.

Conclusions

Telehealth for pediatric OM now spans mature technologies and care models. Store-and-forward tele-otoscopy is feasible and, with standardized image capture and brief structured training, can approach in-person diagnostic performance. Implementation quality requires predefined image quality thresholds and a default to in-person care when those thresholds are not met. It also requires integration into hybrid workflows and basic governance covering consent, documentation, and data protection. Programs should ensure equitable access, especially in resource-limited settings. Video-only encounters without otoscopic imaging are inadequate when decisions require objective middle ear findings. Key evidence gaps include multicenter child-level outcomes, economic

XSL•FO
RenderX

evaluations—particularly in low- and middle-income countries—and external validation of artificial intelligence–based diagnostic tools across devices and sites. These priorities should guide the next phase of research and scale-up.

Acknowledgments

This work was undertaken as part of the activities of the Domestic Scientific Committee of the Japanese Society of Pediatric Otorhinolaryngology. We would like to thank Editage (www.editage.com) for English language editing. The authors are involved in a research project supported by the Japan Science and Technology Agency (JST) under the COI-NEXT (Collaborative Creation of New Social Systems) Program (Grant Number JPMJPF2403). However, this specific study did not receive direct financial support from that grant.

Disclosure of delegation to generative artificial intelligence (GenAI): The authors declare the use of GenAI in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing and translation. The GenAI tool used was ChatGPT-5.2. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by Corresponding Author. In accordance with the GAIDeT taxonomy, GenAI tools were used solely for reformatting and limited language editing tasks under full human supervision.

Funding

This study did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

All data underlying this scoping review were derived from published studies cited in the Reference list. The study-level data extraction table and the QUADAS-2 assessments for diagnostic accuracy studies are provided in the supplementary materials (Multimedia Appendix 2, Multimedia Appendix 3, and Multimedia Appendix 4. Additional materials (the data charting form/codebook and analytic files used to generate summary tables/figures) are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: MN, RA
Data curation: MN, RA
Formal analysis: MN, RA
Investigation: RA, MN, MH
Methodology: MN, RA, MH
Project administration: MN, RA
Supervision: HY, HF, YT, MI
Validation: RA, MN, MH
Visualization: MN, RA
Writing – original draft: MN, RA
Writing – review & editing: MN, RA, MH, CS, RK, HY, YM, HF, YT, MI

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full search strategies for this scoping review.
[DOCX File, 18 KB - jmir_v28i1e85416_app1.docx]

Multimedia Appendix 2
Study-level characteristics and data charting for included studies (N=52). This appendix presents the data extraction (data charting) table for all included studies in this scoping review. Variables include study design and setting, country, participant characteristics (pediatric population), otitis media phenotype, telehealth care model (asynchronous/synchronous/hybrid), device or technology category, operator or capture and interpretation workflow, training and image quality considerations, outcome domains reported, and funding or conflict of interest statements where available.
[XLSX File, 50 KB - jmir_v28i1e85416_app2.xlsx]

XSL•FO
RenderX

Multimedia Appendix 3
RE-AIM mapping of implementation-related outcomes and reporting gaps. This appendix summarizes implementation-relevant constructs using the Reach, Effectiveness, Adoption, Implementation, and Maintenance (RE-AIM) framework across included studies. The table highlights commonly reported elements and underreported domains to inform future research and reporting in telehealth-supported pediatric otitis media care.
[XLSX File, 10 KB - jmir_v28i1e85416_app3.xlsx]

Multimedia Appendix 4
Revised Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2). This appendix reports the QUADAS-2 risk of bias and applicability assessments for the subset of included studies that evaluated diagnostic accuracy. Judgments are provided across QUADAS-2 domains (patient selection, index test, reference standard, flow and timing) and applicability domains (patient selection, index test, reference standard), with signaling questions and domain-level ratings.
[XLSX File, 12 KB - jmir_v28i1e85416_app4.xlsx]

Checklist 1
PRISMA-ScR checklist.
[PDF File, 134 KB - jmir_v28i1e85416_app5.pdf]

References

1. Bajwa NM, Perron NJ, Braillard O, et al. Has telemedicine come to fruition? Parents' and pediatricians' perceptions and preferences regarding telemedicine. Pediatr Res 2024 Oct;96(5):1332-1339. [doi: 10.1038/s41390-024-03172-w] [Medline: 38555380]
2. Howland K, Edvardsson K, Lees H, Hooker L. Telehealth use in the well-child health setting. A systematic review of acceptability and effectiveness for families and practitioners. Int J Nurs Stud Adv 2025 Jun;8:100277. [doi: 10.1016/j.ijnsa.2024.100277] [Medline: 39734584]
3. Luke MJ, Craig S, Pak-Gorstein S, et al. Narrowing the digital divide: framework for creating telehealth equity dashboards. Interact J Med Res 2024 Sep 4;13:e57435. [doi: 10.2196/57435] [Medline: 39231423]
4. Marom T, Pitaro J, Shah UK, et al. Otitis media practice during the COVID-19 pandemic. Front Cell Infect Microbiol 2021;11:749911. [doi: 10.3389/fcimb.2021.749911] [Medline: 35071032]
5. Metcalfe C, Muzaffar J, Orr L, Coulson C. A systematic review of remote otological assessment using video-otoscopy over the past 10 years: reliability and applications. Eur Arch Otorhinolaryngol 2021 Dec;278(12):4733-4741. [doi: 10.1007/s00405-020-06596-2] [Medline: 33486567]
6. Wang H, Zeng X, Miao X, et al. Global, regional, and national epidemiology of otitis media in children from 1990 to 2021. Front Pediatr 2025 Jul 2;13:1513629. [doi: 10.3389/fped.2025.1513629]
7. Schilder AGM, Chonmaitree T, Cripps AW, et al. Otitis media. Nat Rev Dis Primers 2016 Sep 8;2(1):16063. [doi: 10.1038/nrdp.2016.63] [Medline: 27604644]
8. Csonka P, Palmu S, Heikkilä P, Huhtala H, Korppi M. Outpatient antibiotic prescribing for 357,390 children with otitis media. Pediatr Infect Dis J 2022 Dec 1;41(12):947-952. [doi: 10.1097/INF.0000000000003693] [Medline: 36102742]
9. Wittman SR, Hoberman A, Mehrotra A, Sabik LM, Yabes JG, Ray KN. Antibiotic receipt for pediatric telemedicine visits with primary care vs direct-to-consumer vendors. JAMA Netw Open 2024 Mar 4;7(3):e242359. [doi: 10.1001/jamanetworkopen.2024.2359] [Medline: 38483387]
10. Bhatia R, Chauhan A, Rana M, Kaur K, Pradhan P, Singh M. Economic burden of otitis media globally and an overview of the current scenario to alleviate the disease burden: a systematic review. Int Arch Otorhinolaryngol 2024 Jul;28(3):e552-e558. [doi: 10.1055/s-0043-1767802] [Medline: 38974618]
11. El Feghaly RE, Nedved A, Katz SE, Frost HM. New insights into the treatment of acute otitis media. Expert Rev Anti Infect Ther 2023 May;21(5):523-534. [doi: 10.1080/14787210.2023.2206565] [Medline: 37097281]
12. Erkkola-Anttinen N, Irjala H, Laine MK, Tähtinen PA, Löyttyniemi E, Ruohola A. Smartphone otoscopy performed by parents. Telemed J E Health 2019 Jun;25(6):477-484. [doi: 10.1089/tmj.2018.0062] [Medline: 30040525]
13. Wagner R, Lima TC, Silva MD, et al. Assessment of pediatric telemedicine using remote physical examinations with a mobile medical device: a nonrandomized controlled trial. JAMA Netw Open 2023 Feb 1;6(2):e2252570. [doi: 10.1001/jamanetworkopen.2022.52570] [Medline: 36729459]
14. Alenezi EM, Veselinović T, Tao KF, et al. Ear Portal: an urban-based ear, nose, and throat, and audiology referral telehealth portal to improve access to specialist ear health services for children. J Telemed Telecare 2024 Dec;30(10):1581-1589. [doi: 10.1177/1357633X231158839] [Medline: 36916156]
15. Dubois C, Eigen D, Simon F, et al. Development and validation of a smartphone-based deep-learning-enabled system to detect middle-ear conditions in otoscopic images. NPJ Digit Med 2024 Jun 20;7(1):162. [doi: 10.1038/s41746-024-01159-9] [Medline: 38902477]

XSL•FO
RenderX

16. Jin X, Fan M. EarMonitor: non-clinical assessment of ear health conditions using a low-cost endoscope camera on smartphones. Proc ACM Hum-Comput Interact 2024 Sep 24;8(MHCI):1-20. [doi: 10.1145/3676499]
17. Appelberg E, Viitasalo S, Sinkkonen ST. Patient self-recorded digital otoscopy in otology diagnostics. Laryngoscope Investig Otolaryngol 2025 Dec;10(6):e70318. [doi: 10.1002/lio2.70318] [Medline: 41347050]
18. Haynes SC, Marcin JP. Pediatric telemedicine: lessons learned during the coronavirus disease 2019 pandemic and opportunities for growth. Adv Pediatr 2022 Aug;69(1):1-11. [doi: 10.1016/j.yapd.2022.04.002] [Medline: 35985702]
19. Curfman A, McSwain SD, Chuo J, et al. Pediatric telehealth in the COVID-19 pandemic era and beyond. Pediatrics 2021 Sep;148(3):e2020047795. [doi: 10.1542/peds.2020-047795] [Medline: 34215677]
20. Moberly AC, Zhang M, Yu L, et al. Digital otoscopy versus microscopy: how correct and confident are ear experts in their diagnoses? J Telemed Telecare 2018 Aug;24(7):453-459. [doi: 10.1177/1357633X17708531] [Medline: 28480781]
21. Damery L, Lescanne E, Reffet K, Aussedat C, Bakhos D. Interest of video-otoscopy for the general practitioner. Eur Ann Otorhinolaryngol Head Neck Dis 2019 Feb;136(1):13-17. [doi: 10.1016/j.anorl.2018.10.016] [Medline: 30392875]
22. Quick ME, Brennan-Jones CG, Kuthubutheen J. Remote paediatric ear examination comparing video-otoscopy and still otoscopy clinician rated outcomes. Int J Pediatr Otorhinolaryngol 2024 Feb;177:111871. [doi: 10.1016/j.ijporl.2024.111871] [Medline: 38266378]
23. Habib AR, Crossland G, Sacks R, Singh N, Patel H. Tele-otology for Aboriginal and Torres Strait Islander people living in rural and remote areas. Laryngoscope 2024 Dec;134(12):5096-5102. [doi: 10.1002/lary.31624] [Medline: 38982868]
24. Pannonhalmi Á, Posta B, Perényi Á, et al. Clinical validation of a video-otoscopy-based medical device for the remote diagnosis of ear complaints. Sensors (Basel) 2025 Jan 27;25(3):758. [doi: 10.3390/s25030758] [Medline: 39943397]
25. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
26. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
27. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
28. Smith AC, Armfield NR, Wu WI, Brown CA, Perry C. A mobile telemedicine-enabled ear screening service for Indigenous children in Queensland: activity and outcomes in the first three years. J Telemed Telecare 2012 Dec;18(8):485-489. [doi: 10.1258/jtt.2012.gth114] [Medline: 23209267]
29. Biagio L, Swanepoel DW, Laurent C, Lundberg T. Video-otoscopy recordings for diagnosis of childhood ear disease using telehealth at primary health care level. J Telemed Telecare 2014 Sep;20(6):300-306. [doi: 10.1177/1357633X14541038] [Medline: 24958356]
30. Wu CJ, Wu SY, Chen PC, Lin YS. An innovative smartphone-based otorhinoendoscope and its application in mobile health and teleotolaryngology. J Med Internet Res 2014 Mar 3;16(3):e71. [doi: 10.2196/jmir.2959] [Medline: 24590187]
31. Smith AC, Brown C, Bradford N, Caffery LJ, Perry C, Armfield NR. Monitoring ear health through a telemedicine-supported health screening service in Queensland. J Telemed Telecare 2015 Dec;21(8):427-430. [doi: 10.1177/1357633X15605407] [Medline: 26377121]
32. Jacups SP, Newman D, Dean D, Richards A, McConnon KM. An innovative approach to improve ear, nose and throat surgical access for remote living Cape York Indigenous children. Int J Pediatr Otorhinolaryngol 2017 Sep;100:225-231. [doi: 10.1016/j.ijporl.2017.07.011] [Medline: 28802377]
33. Prins-van Ginkel AC, de Hoog ML, Uiterwaal C, Smit HA, Bruijning-Verhagen PC. Detecting acute otitis media symptom episodes using a mobile app: cohort study. JMIR Mhealth Uhealth 2017 Nov 28;5(11):e181. [doi: 10.2196/mhealth.7505] [Medline: 29183869]
34. Lundberg T, Biagio de Jager L, Swanepoel DW, Laurent C. Diagnostic accuracy of a general practitioner with video-otoscopy collected by a health care facilitator compared to traditional otoscopy. Int J Pediatr Otorhinolaryngol 2017 Aug;99:49-53. [doi: 10.1016/j.ijporl.2017.04.045] [Medline: 28688565]
35. Ramkumar V, Rajendran A, Nagarajan R, Balasubramaniyan S, Suresh DK. Identification and management of middle ear disorders in a rural cleft care program: a telemedicine approach. Am J Audiol 2018 Nov 19;27(3S):455-461. [doi: 10.1044/2018_AJA-IMIA3-18-0015] [Medline: 30452749]
36. Mandavia R, Lapa T, Smith M, Bhutta MF. A cross-sectional evaluation of the validity of a smartphone otoscopy device in screening for ear disease in Nepal. Clin Otolaryngol 2018 Feb;43(1):31-38. [doi: 10.1111/coa.12898] [Medline: 28485038]
37. Mousseau S, Lapointe A, Gravel J. Diagnosing acute otitis media using a smartphone otoscope; a randomized controlled trial. Am J Emerg Med 2018 Oct;36(10):1796-1801. [doi: 10.1016/j.ajem.2018.01.093] [Medline: 29544905]
38. Venail F, Akkari M, Merklen F, et al. Evaluation of otoscopy simulation as a training tool for real-time remote otoscopy. Int J Audiol 2018 Mar;57(3):194-200. [doi: 10.1080/14992027.2017.1416190] [Medline: 29256826]
39. Shah MU, Sohal M, Valdez TA, Grindle CR. iPhone otoscopes: currently available, but reliable for tele-otoscopy in the hands of parents? Int J Pediatr Otorhinolaryngol 2018 Mar;106:59-63. [doi: 10.1016/j.ijporl.2018.01.003] [Medline: 29447893]

XSL•FO
RenderX

40. Bhavana K, Ahmad M, Sharma P. Smartphone otoscopy sans attachment: a paradigm shift in diagnosing ear pathologies. OTO Open 2018;2(3):2473974X18786496. [doi: 10.1177/2473974X18786496] [Medline: 31565683]
41. Chan KN, Silverstein A, Bryan LN, McCracken CE, Little WK, Shane AL. Comparison of a smartphone otoscope and conventional otoscope in the diagnosis and management of acute otitis media. Clin Pediatr (Phila) 2019 Mar;58(3):302-306. [doi: 10.1177/0009922818812480] [Medline: 30463434]
42. Chan J, Raju S, Nandakumar R, Bly R, Gollakota S. Detecting middle ear fluid using smartphones. Sci Transl Med 2019 May 15;11(492):eaav1102. [doi: 10.1126/scitranslmed.aav1102] [Medline: 31092691]
43. Demant MN, Jensen RG, Bhutta MF, Laier GH, Lous J, Homøe P. Smartphone otoscopy by non-specialist health workers in rural Greenland: a cross-sectional study. Int J Pediatr Otorhinolaryngol 2019 Nov;126:109628. [doi: 10.1016/j.ijporl.2019.109628] [Medline: 31445480]
44. Hakimi AA, Lalehzarian AS, Lalehzarian SP, Azhdam AM, Nedjat-Haiem S, Boodaie BD. Utility of a smartphone-enabled otoscope in the instruction of otoscopy and middle ear anatomy. Eur Arch Otorhinolaryngol 2019 Oct;276(10):2953-2956. [doi: 10.1007/s00405-019-05559-6] [Medline: 31317322]
45. Fordington S, Brown TH. An evaluation of the Hear Glue Ear mobile application for children aged 2-8 years old with otitis media with effusion. Digit Health 2020;6:2055207620966163. [doi: 10.1177/2055207620966163] [Medline: 33194218]
46. Gupta N, Chawla N, Gupta D, Dhawan N, Janaki VR. Community triage of otology patients using a store-and-forward telemedicine device: a feasibility study. Ear Nose Throat J 2020 Mar 22:145561320908953. [doi: 10.1177/0145561320908953] [Medline: 32200652]
47. Cha D, Shin SH, Kim J, et al. Feasibility of asynchronous and automated telemedicine in otolaryngology: prospective cross-sectional study. JMIR Med Inform 2020 Oct 19;8(10):e23680. [doi: 10.2196/23680] [Medline: 33027033]
48. Durgut O, Ekim B, Dikici O, Solmaz F, Ağırgöl B, Özbakan A. Evaluation of hearing thresholds by using a mobile application in children with otitis media with effusion. Audiol Neurootol 2020;25(3):120-124. [doi: 10.1159/000505309] [Medline: 31962328]
49. Jayawardena ADL, Mankarious LA, Keamy DG Jr, Cohen MS. Pediatric, family-centered, “at-home” otologic physical examination in the COVID-19 era. Otolaryngol Head Neck Surg 2020 Nov;163(5):1061-1063. [doi: 10.1177/0194599820934776] [Medline: 32513060]
50. Cottrell E, George A, Coulson C, Chambers R. Telescopic otology referrals: evaluation of feasibility and acceptability. Laryngoscope Investig Otolaryngol 2020 Apr;5(2):221-227. [doi: 10.1002/lio2.367] [Medline: 32337353]
51. Schuster-Bruce JR, Ali A, Van M, Rogel-Salazar J, Ofo E, Shamil E. A randomised trial to assess the educational benefit of a smartphone otoscope in undergraduate medical training. Eur Arch Otorhinolaryngol 2021 Jun;278(6):1799-1804. [doi: 10.1007/s00405-020-06373-1] [Medline: 32960350]
52. Kolb CM, Born K, Banker K, Barth P, Aaronson NL. Comparing telehealth with office-based visits for common pediatric otolaryngology complaints. Int J Pediatr Otorhinolaryngol 2021 Jun;145:110712. [doi: 10.1016/j.ijporl.2021.110712] [Medline: 33887549]
53. Kleinman K, Psoter KJ, Nyhan A, Solomon BS, Kim JM, Canares T. Evaluation of digital otoscopy in pediatric patients: a prospective randomized controlled clinical trial. Am J Emerg Med 2021 Aug;46:150-155. [doi: 10.1016/j.ajem.2021.04.030] [Medline: 33945977]
54. Don DM, Koempel JA, Fisher LM, Wee CP, Osterbauer B. Prospective evaluation of a smartphone otoscope for home tympanostomy tube surveillance: a pilot study. Ann Otol Rhinol Laryngol 2021 Feb;130(2):125-132. [doi: 10.1177/0003489420938820] [Medline: 32627572]
55. Manayan RC, Ladd-Luthringshauser OH, Packer A, et al. Ambient noise limits efficacy of smartphone-based screening for hearing loss in children at risk. Am J Otolaryngol 2022;43(1):103223. [doi: 10.1016/j.amjoto.2021.103223] [Medline: 34560596]
56. Meng X, Dai Z, Wang Y, et al. Application of smartphone otoscope in telemedicine in rural medical consortium in eastern China in the COVID-19 era. Front Public Health 2022;10:879410. [doi: 10.3389/fpubh.2022.879410] [Medline: 35664106]
57. Deshmukh BG, Bhisegaonkar DA, Phatak P, et al. Accuracy of the novel ENTraview mobile device to increase access to screening and provisional diagnosis of ear disease in underserved communities across India. Epidem Int 2022 Mar 27;07(1):21-24. [doi: 10.24321/2455.7048.202207]
58. Schafer A, Mousset M, Kelly N, Althubaiti A, Bourgeois T, Elmaraghy CA. Assessing the use of telehealth for the surgical management of recurrent otitis media. Int J Pediatr Otorhinolaryngol 2022 Feb;153:111036. [doi: 10.1016/j.ijporl.2021.111036] [Medline: 34998205]
59. Smola C, Shah N, Fowler A, Gubbels A, Monroe K. Assessment of pediatric residents’ comfort level with CellScope Oto to examine pediatric ear exams. Clin Pediatr (Phila) 2022 Jul;61(7):485-489. [doi: 10.1177/00099228221086374] [Medline: 35446181]
60. Chan J, Najafi A, Baker M, et al. Performing tympanometry using smartphones. Commun Med (Lond) 2022;2(1):57. [doi: 10.1038/s43856-022-00120-9] [Medline: 35721828]
61. Habib AR, Perry C, Crossland G, et al. Inter-rater agreement between 13 otolaryngologists to diagnose otitis media in Aboriginal and Torres Strait Islander children using a telehealth approach. Int J Pediatr Otorhinolaryngol 2023 May;168:111494. [doi: 10.1016/j.ijporl.2023.111494] [Medline: 37003013]

XSL•FO
RenderX

62. Patel A, Forde C, Doal S, et al. A new community-to-hospital specialist otology pathway using telemedicine: a pilot study of 50 patients. Clin Otolaryngol 2024 Sep;49(5):652-659. [doi: 10.1111/coa.14192] [Medline: 38899484]
63. Gyawali BR, Obaidullah, Ghimire S, et al. Reliability of a tailored tele-medicine model for otorhinolaryngology-head and neck surgery (ORL-HNS) clinical practice in remote areas of Nepal: a pilot study from tertiary referral center. Ear Nose Throat J 2024 Oct 15:1455613241291168. [doi: 10.1177/01455613241291168] [Medline: 39405038]
64. Bowers P, Graydon K, Tomlin D, Rance G. The reliability of remote diagnostic hearing assessment in school-entry-aged children. Int J Audiol 2025 Mar;64(3):261-270. [doi: 10.1080/14992027.2024.2333392] [Medline: 38551077]
65. Altamimi AA, Brennan-Jones CG, Robinson M, et al. A hospital-based asynchronous ENT telehealth service for children with otitis media: cost-minimisation and improved access. J Telemed Telecare 2025 Jul;31(6):818-825. [doi: 10.1177/1357633X231223994] [Medline: 38295365]
66. Rappaport KM, McCracken CC, Beniflah J, et al. Assessment of a smartphone otoscope device for the diagnosis and management of otitis media. Clin Pediatr (Phila) 2016 Aug;55(9):800-810. [doi: 10.1177/0009922815593909] [Medline: 26156976]
67. Ni G, Curtis S, Kaplon A, et al. Development of video otoscopy quiz using a smartphone adaptable otoscope. J Otol 2021 Apr;16(2):80-84. [doi: 10.1016/j.joto.2020.08.001] [Medline: 33777119]
68. Stancel-Lewis J, Joseph K, Laister S, et al. Parent-led ear health checks at home for children with complex needs pilot-project: a new approach to improving access to ear health. Int J Audiol 2025 Aug;64(8):803-812. [doi: 10.1080/14992027.2025.2456021] [Medline: 39962803]
69. Yancey KL, Cheromei LJ, Muhando J, Reppart J, Netterville JL, Jayawardena ADL. Pediatric hearing screening in low-resource settings: incorporation of video-otoscopy and an electronic medical record. Int J Pediatr Otorhinolaryngol 2019 Nov;126:109633. [doi: 10.1016/j.ijporl.2019.109633] [Medline: 31421355]
70. Phillips JH, Wigger C, Beissbarth J, McCallum GB, Leach A, Morris PS. Can mobile phone multimedia messages and text messages improve clinic attendance for Aboriginal children with chronic otitis media? A randomised controlled trial. J Paediatr Child Health 2014 May;50(5):362-367. [doi: 10.1111/jpc.12496] [Medline: 24612007]
71. Richards JR, Gaylor KA, Pilgrim AJ. Comparison of traditional otoscope to iPhone otoscope in the pediatric ED. Am J Emerg Med 2015 Aug;33(8):1089-1092. [doi: 10.1016/j.ajem.2015.04.063] [Medline: 25979304]
72. Cortés Fuentes IA, Zelada B Ú, Zitko M P. Smartphone-based video-otoscopy performed by non-ENT physicians as a complementary tool to conventional otoscopy for tympanic membrane diagnosis. Speech Lang Hear 2024 Jan 2;27(1):14-19. [doi: 10.1080/2050571X.2023.2209766]
73. Kashani RG, Młyńczak MC, Zarabanda D, et al. Shortwave infrared otoscopy for diagnosis of middle ear effusions: a machine-learning-based approach. Sci Rep 2021 Jun 15;11(1):12509. [doi: 10.1038/s41598-021-91736-9] [Medline: 34131163]
74. Tran TT, Fang TY, Pham VT, Lin C, Wang PC, Lo MT. Development of an automatic diagnostic algorithm for pediatric otitis media. Otol Neurotol 2018 Sep;39(8):1060-1065. [doi: 10.1097/MAO.0000000000001897] [Medline: 30015751]
75. Khan MA, Kwon S, Choo J, et al. Automatic detection of tympanic membrane and middle ear infection from oto-endoscopic images via convolutional neural networks. Neural Netw 2020 Jun;126:384-394. [doi: 10.1016/j.neunet.2020.03.023] [Medline: 32311656]
76. Noda M, Yoshimura H, Okubo T, et al. Feasibility of multimodal artificial intelligence using GPT-4 vision for the classification of middle ear disease: qualitative study and validation. JMIR AI 2024 May 31;3:e58342. [doi: 10.2196/58342] [Medline: 38875669]
77. Dharmar M, Simon A, Sadorra C, et al. Reducing loss to follow-up with tele-audiology diagnostic evaluations. Telemed J E Health 2016 Feb;22(2):159-164. [doi: 10.1089/tmj.2015.0001] [Medline: 26544032]
78. Landolfi E, Continisio GI, Del Vecchio V, et al. NeonaTal Assisted TelerehAbilitation (T.A.T.A. web app) for hearing-impaired children: a family-centered care model for early intervention in congenital hearing loss. Audiol Res 2022 Mar 28;12(2):182-190. [doi: 10.3390/audiolres12020021] [Medline: 35447741]
79. Ramkumar V, Joshi B D, Prabhakar A, Hall JW, Vaidyanath R. Development and beta validation of an mHealth-based hearing screener (SRESHT) for young children in resource-limited countries: pilot validation study. JMIR Form Res 2025 Jan 13;9:e53460. [doi: 10.2196/53460] [Medline: 39805111]
80. Jacups SP, Kinchin I, McConnon KM. Ear, nose, and throat surgical access for remote living Indigenous children: what is the least costly model? J Eval Clin Pract 2018 Dec;24(6):1330-1338. [doi: 10.1111/jep.13044] [Medline: 30311714]

Abbreviations

AOM: acute otitis media
AUC: area under the receiver operating characteristic curve
CSOM: chronic suppurative otitis media
DenseNet: Dense Convolutional Network
ENT: ear, nose, and throat
JBI: Joanna Briggs Institute

XSL•FO
RenderX

MEE: middle ear effusion
mHealth: mobile health
OM: otitis media
OME: otitis media with effusion
PCC: population, concept, and context
PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
QUADAS-2: revised Quality Assessment of Diagnostic Accuracy Studies tool
RAOM: recurrent acute otitis media
RE-AIM: reach, effectiveness, adoption, implementation, and maintenance
ROC: receiver operating characteristic

Edited by S Brini; submitted 07.Oct.2025; peer-reviewed by D Singh, M Mansoor; revised version received 14.Feb.2026; accepted 23.Feb.2026; published 18.Mar.2026.

Please cite as:
Noda M, Akiyoshi R, Hosoya M, Shinkawa C, Koshu R, Yoshimura H, Monden Y, Fushiki H, Tanaka Y, Ito M
Telehealth Approaches for Pediatric Otitis Media and Clinical Outcomes: Scoping Review
J Med Internet Res 2026;28:e85416
URL: https://www.jmir.org/2026/1/e85416
doi:10.2196/85416

© Masao Noda, Ryohei Akiyoshi, Makoto Hosoya, Chikako Shinkawa, Ryota Koshu, Hidekane Yoshimura, Yukifumi Monden, Hiroaki Fushiki, Yasuhiro Tanaka, Makoto Ito. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Performance of Artificial Intelligence in Classifying Molecular Markers in Adult-Type Gliomas Using Histopathological Images: Systematic Review

Obada Almaabreh[1], MD; Rukaya Al-Dafi[1], MD; Aliya Tabassum[2], PhD; Ahmad Othman[1], MD; Alaa Abd-alrazaq[3,4], PhD

[1]Faculty of Medicine, Yarmouk University, Irbid, Jordan

[2]Department of Computer Science and Engineering, College of Engineering, Qatar University, Doha, Qatar

[3]AI Center for Precision Health, Weill Cornell Medical College in Qatar, 103 Ezdan Oasis, Alwakrah, Doha, Qatar

[4]Department of Biomedical Sciences, College of Health Sciences, Qatar University, Doha, Qatar

Corresponding Author:
Alaa Abd-alrazaq, PhD
AI Center for Precision Health, Weill Cornell Medical College in Qatar, 103 Ezdan Oasis, Alwakrah, Doha, Qatar

Abstract

Background: Adult-type gliomas are among the most prevalent and lethal primary central nervous system tumors, where prompt and accurate diagnosis is essential for maximizing survival prospects. Molecular classification, particularly the detection of isocitrate dehydrogenase (IDH) mutations and 1p/19q codeletions, has become crucial for accurate diagnosis and prognosis. Artificial intelligence (AI) has emerged as a promising adjunct in enhancing diagnostic accuracy using histopathological images. Existing reviews mostly focused on radiology rather than histopathology, and no comprehensive systematic review has specifically evaluated AI performance exclusively from histopathological images for detecting these two molecular markers.

Objective: This study aims to systematically evaluate the performance of AI models in detecting and classifying IDH mutation status and 1p/19q gene codeletion in adult-type gliomas using histopathological images.

Methods: A systematic review was conducted in accordance with PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Extension for Diagnostic Test Accuracy) guidelines. Seven databases (MEDLINE, PsycINFO, Embase, IEEE Xplore, ACM Digital Library, Scopus, and Google Scholar) were searched for studies published between 2015 and 2025. Eligible studies used AI models on histopathological images for molecular classification of adult-type gliomas and reported performance metrics. Study selection, data extraction, and risk of bias assessment using a modified QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2) tool were conducted independently by two reviewers. Extracted data were synthesized narratively.

Results: A total of 2453 reports were identified, with 22 studies meeting the inclusion criteria. The pooled average accuracy, sensitivity, specificity, and area under the curve (AUC) across studies were 85.46%, 84.55%, 86.03%, and 86.53%, respectively. Hybrid models demonstrated the highest diagnostic performance (accuracy 92.80% and sensitivity 89.62%). In general, AI models that used multimodal data outperformed those that used unimodal data in terms of sensitivity (90.15% vs 84.31%) and AUC (88.93% vs 86.29%). Furthermore, models had a better overall performance in identifying IDH mutations than 1p/19q codeletions, with higher accuracy (86.13% vs 81.63%), specificity (86.61% vs 78.11%), and AUC (86.74% vs 85.15%). Unexpectedly, AI models designed for binary classification exhibited lower performance than those for multiclass classification in terms of both accuracy (91.98% vs 84.02%) and sensitivity (93.41% vs 80.18%). However, these differences should be interpreted as descriptive trends rather than statistically validated superiority, as formal between-group comparisons were not feasible.

Conclusions: AI models show strong potential as complementary tools for the molecular classification of adult-type gliomas using histopathology images, particularly for IDH mutation detection. However, these findings are constrained by the limited number of studies, the focus on adult-type gliomas, lack of meta-analysis, and restriction to English-language publications. While AI offers valuable diagnostic support, it must be integrated with expert clinical judgment. Future research should prioritize larger, more diverse datasets and multimodal AI frameworks and extend to other brain tumor types for broader applicability.

Trial Registration: PROSPERO CRD420250653668; https://www.crd.york.ac.uk/PROSPERO/view/CRD420250653668

(J Med Internet Res 2026;28:e78377) doi:10.2196/78377

XSL•FO
RenderX

KEYWORDS

glioma; brain tumors; molecular markers; artificial intelligence; histopathology; isocitrate dehydrogenase; IDH mutation; 1p/19q codeletion; systematic review

Introduction

Background

Gliomas are the most frequent primary tumors affecting the central nervous system (CNS) in adults, with a variable prognosis depending on the specific subtype and histological grade [1]. The global prevalence of CNS neoplasms in 5 years is approximately 771,110 cases. Primary brain tumors account for 1.7% of all cancers, with a global incidence of 3.9 cases per 100,000 people each year. Among these, gliomas are the most common malignant type, accounting for 75% of all malignant CNS tumors, with an incidence of 6 cases per 100,000 people annually [2]. Gliomas are among the tumors that are difficult to treat, with a 5-year overall survival of less than 35% [3]. Patients with brain tumors present with variable clinical symptoms according to the part of the brain affected. However, they usually share general symptoms that are nonspecific to anatomic location (eg, seizures, headaches).

Gliomas are neuroectodermal tumors arising from glial cells or their precursors and encompass astrocytomas, oligodendrogliomas, and ependymomas [3]. Mutations in isocitrate dehydrogenase 1 and 2 (IDH1 and IDH2) are considered early events in gliomagenesis and occur more frequently in lower-grade gliomas [3]. Diffuse gliomas harboring IDH1/2 mutations are associated with significantly improved prognosis compared with IDH–wild-type diffuse gliomas [3].

Codeletion of chromosomal arms 1p and 19q results from an unbalanced centromeric translocation, t(1;19)(q10;p10) [3]. In combination with IDH mutation, 1p/19q codeletion is a defining molecular feature required for the diagnosis of oligodendroglioma, IDH mutant, and 1p/19q codeleted. This molecular alteration is associated with a favorable prognosis among diffuse gliomas and predicts enhanced responsiveness to alkylating chemotherapy [3].

Conventional classification of gliomas relied on histologic features as a gold standard. Initially, pathologists utilized microscopic analysis of histochemically stained (eg, hematoxylin and eosin) sections for the diagnosis of gliomas [4]. However, ancillary immunohistochemistry has been increasingly used over the past few decades to enhance diagnostic accuracy [4]. With either method, pathologists followed a diagnostic algorithm to rule out nonneoplastic lesions (eg, infarcts) or other types of malignancies (eg, metastatic) based on histologic features. Clinical information (eg, age) and radiologic findings (eg, location) are also valuable clues in the diagnosis [4]. Nonetheless, the histologic-based classification system has a few limitations, particularly interobserver variability, in addition to insufficient or nonrepresentative tissue sampling [5]. Although histologic-based classification can be accurate for prototypic tumors, classification of tumors encountered in clinical practice is often more challenging due to a degree of mixed features in the same sample [5].

In order to address these limitations, molecular-based diagnosis has emerged as a possible solution. Recent advancements in genetics have helped to identify specific gene alterations involved in the development of different types of gliomas. Detection of these gene alterations is the base of molecular-based diagnosis. This method offers improved accuracy and made it possible to distinguish between primary and secondary gliomas, which was not possible with the previous histologic-based method [6]. Therefore, the World Health Organization (WHO) integrated molecular features into its most recent classification of CNS tumors, published in 2021, based on the presence or absence of IDH1 and IDH2 mutations, in addition to chromosome 1p/19q codeletion [7]. It categorized adult-type gliomas into three categories: astrocytoma with IDH mutation, oligodendroglioma with IDH mutation and 1p/19q codeletion, and glioblastoma with IDH wild-type [7]. Furthermore, a combined approach using molecular and histologic features is used for the grading of IDH-mutant astrocytomas as CNS WHO grade 2, 3, or 4. The new WHO classification also led to a more accurate prognosis of gliomas, as the absence of IDH mutation is often a sign of more aggressive tumors [7].

Neuroimaging provides noninvasive insight into the molecular landscape of diffuse gliomas. Advanced magnetic resonance imaging (MRI) features can suggest IDH mutation status, as IDH-mutant tumors typically exhibit lower perfusion and less aggressive radiological behavior than IDH–wild-type gliomas [5-7]. Magnetic resonance spectroscopy allows in vivo detection of the IDH-specific oncometabolite D-2-hydroxyglutarate, enabling preoperative identification of IDH-mutant tumors [7]. In addition, oligodendrogliomas defined by combined IDH mutation and 1p/19q codeletion display characteristic imaging features, including cortical localization, calcifications, and relatively increased perfusion, reflecting distinct tumor biology [4,7]. These radiogenomic associations support the integration of imaging into molecular stratification of adult-type diffuse gliomas [4-6].

As gliomas pose diagnostic challenges due to their complex histological and molecular features, artificial intelligence (AI) has emerged as a promising tool to enhance diagnostic accuracy and efficiency. AI models were able to offer rapid and accurate results with accuracy exceeding 90%, which can further improve the access and speed of molecular diagnosis [8]. In general, AI implementation in health care can lead to improved data synthesis (eg, patient data, medical literature) [9]. Additionally, AI can process multimodal data (two or more different data sources such as histopathological images, radiological data, clinical variables, genomic markers, and other relevant data types) to provide more accurate diagnosis and predictions. Moreover, it offers means for augmenting human performance, as human limitations can prevent processing large quantities of information and making the optimal judgment at all times. Thus, AI can improve care consistency, increase precision, accelerate discoveries, and minimize disparities [9].

XSL•FO
RenderX

Research Gaps and Aim

The use of AI in the diagnosis of brain tumors and gliomas has been widely studied, with many reviews summarizing the results of numerous studies. However, existing reviews exhibit several limitations. For instance, many reviews focused on AI implementation in other diagnostic methods of gliomas, such as radiology-based [10,11], surgery-based [12], and DNA-based diagnoses [13]. As for reviews focusing on AI implementation in the molecular diagnosis of gliomas, they fall into one of the following categories: not systematic [14], outdated [15], restricted by databases or search terms [11,16], lacking subtyping of gliomas [17], not focused on genetic markers (IDH mutations and 1p/19q codeletion) [11], or limited to a specific subfield of AI (eg, deep learning or machine learning) [16]. To address the limitations of the previous reviews, this study aims to systematically assess the performance of AI in detecting and classifying IDH mutation status and 1p/19q gene codeletion in adult-type gliomas using histopathological images.

Methods

This review was conducted in accordance with the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Extension for Diagnostic Test Accuracy) [18] (see Checklist 1) and registered with the international Prospective Register of Systematic Reviews (PROSPERO) under the number CRD420250653668.

Search Strategy

The literature search was performed on November 15, 2024, using the following digital databases: MEDLINE, PsycINFO, Embase, IEEE Xplore, ACM Digital Library, Scopus, and Google Scholar. A biweekly automated search was set for 4 months, concluding on March 15, 2025. Given the large number of results from Google Scholar, which are ranked by relevance, we reviewed only the first 100 entries (10 pages). Our query utilized a combination of proper keywords and Boolean operators, including but not limited to ((“Artificial Intelligence” OR “Machine Learning” OR “Deep Learning”) AND (“Histopatholog*” OR “Whole-Slide Imag*”) AND (“Glioma*” OR “Astrocytoma*” OR “Oligodendroglioma*” OR “Glioblastoma*”)). The search was restricted to studies published between 2015 and 2025. The search query for each database is shown in Multimedia Appendix 1. To identify additional studies, we screened the reference lists of included studies (ie, backward reference list checking) and checked studies that cited the included studies (ie, forward reference list checking).

Study Selection

The selection process of relevant studies consisted of three steps. The first step involved removing duplicates using EndNote. In the second step, the studies were assigned to two authors, who screened the papers independently by title and abstract. The final step involved each author reading the consequent papers in full text independently. Discussions were held to resolve any differences or ambiguities in steps two and three.

Study Eligibility Criteria

Studies fulfilling the following criteria were included: (1) original articles, theses, dissertations, and conference papers written in English, (2) studies published in 2015 onward, (3) studies focused on patients with adult-type glioma with no restrictions on age, (4) studies that used histopathological images gathered from tissue biopsies, tumor excision, or digital histopathological images, (5) studies that assessed the performance of AI algorithms in the detection and classification of the molecular markers, specifically the IDH mutation status and the 1p/19q gene codeletion status using histopathology data, and (6) studies that reported results related to the performance of the developed AI models. No restriction was applied regarding study settings, study design, and country of publication. Studies were excluded if they (1) used data modalities other than histopathology images (eg, radiological images, stimulated Raman histology), (2) did not employ AI algorithms for molecular marker detection or classification, (3) were not human studies, or (4) were categorized as editorials, letters to editors, posters, conference abstracts, reviews, commentaries, short communications, and brief reports.

Data Extraction

Two authors independently extracted data from the eligible studies using predefined tables in Microsoft Excel sheet. Any discrepancies in the extracted data were resolved through discussion. The extracted data included four main categories: (1) study characteristics, including surname of first author, publication type, publication year, and country of publication; (2) participants and dataset characteristics, including number of participants, mean age, female percentage, dataset size, data source, magnification, slide image resolution, level of analysis, number of classes, and ground truth assessor; (3) AI model characteristics, covering problem-solving approaches, AI classifiers, type of classification, aim of classifier, data input to AI algorithms, features extraction methods, type of validation, and performance metrics; (4) results, including accuracy, sensitivity, specificity, and area under the curve (AUC). The data extraction form is shown in Multimedia Appendix 2.

Risk of Bias Assessment

To evaluate the quality and reliability of the included studies, two reviewers independently conducted a risk of bias assessment using an adapted version of the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2) tool [19]. The original QUADAS-2 tool was modified to better suit the context of our review, which focuses on AI applications. Specifically, we incorporated relevant elements from the PROBAST (Prediction Model Risk of Bias Assessment Tool) to ensure that the assessment adequately addressed the methodological nuances of AI-based diagnostic models [20]. Each domain included four tailored signaling questions designed to determine the potential for bias and to evaluate applicability to the review’s research question. The risk of bias was rated separately for each domain, while concerns regarding applicability were assessed for the first three domains. For example, the “Participants” domain examined issues such as sample representativeness, exclusions, and subgroup balance. The “Index Test” domain evaluated the transparency and consistency of the AI model’s design and

XSL•FO
RenderX

feature use. The "Reference Standard" focused on the validity and consistency of outcome definitions and assessor qualifications, while the "Analysis" domain addressed data inclusion, preprocessing, set division, and model performance evaluation. In addition to risk of bias, we also assessed concerns regarding applicability in the first three domains, evaluating whether the participants, AI model, and outcome definitions were aligned with the review's objectives. To refine and validate our modified tool, we conducted a pilot assessment using five studies, which allowed us to fine-tune the criteria before full application. Subsequently, two reviewers independently assessed all included studies using the finalized tool (Multimedia Appendix 3). Any discrepancies in assessments were resolved through discussion and consensus.

Data Synthesis

A narrative approach was used to synthesize the data extracted from the included studies. Specifically, we used texts and tables to summarize and describe the characteristics of the included studies and results. We could not perform a meta-analysis because fewer than two studies reported the confusion matrices or other necessary details (eg, number of cases and controls in the test set) required to calculate the numerators and denominators for the same performance measures (accuracy, sensitivity, specificity, and AUC), AI algorithms, and measured outcomes (detection of IDH and 1p/19q). Instead of meta-analysis, we calculated the traditional mean for each metric (ie, accuracy, sensitivity, specificity, and AUC).

Results

Search Results

As shown in Figure 1, the literature searches identified a total of 2453 reports from databases. After removing duplicate records (n=583), we screened the titles and abstracts of 1870 reports. This led to the exclusion of 1753 reports. Of the remaining 117 reports, we could not retrieve the full text for 5 reports. Subsequently, we checked the eligibility of the remaining 112 reports. This led to the exclusion of 91 reports for the following reasons: not related to glioma (n=7), not related to molecular classification (n=53), not pertaining to histopathology (n=11), inappropriate publication type (n=19), and not in the English language (n=1). We included an additional study by checking the reference list of the included studies. Ultimately, only 22 reports met the inclusion criteria and were included in the review [21-42].

XSL•FO
RenderX

Figure 1. Flowchart of the study selection process.

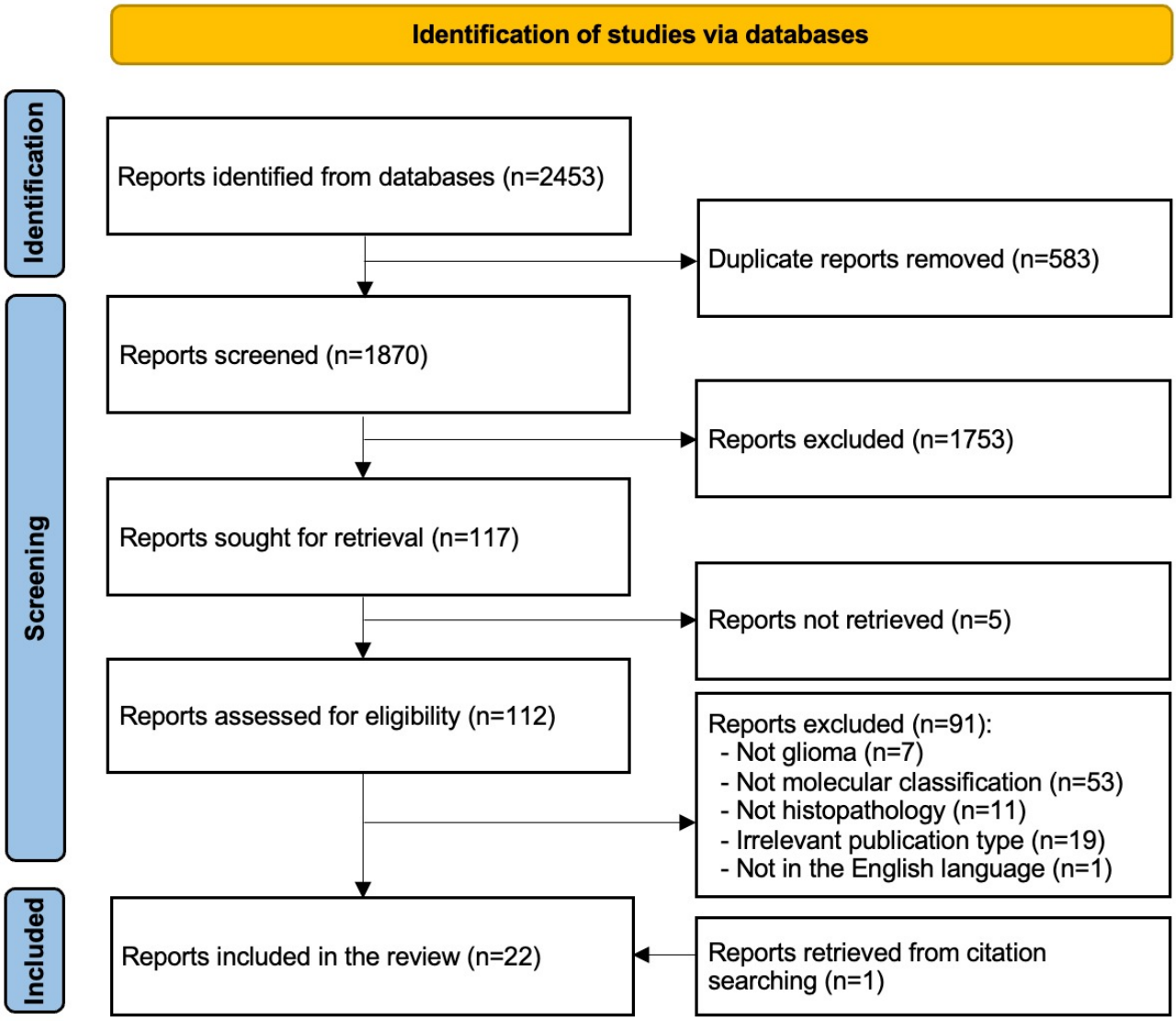

Characteristics of the Included Studies

The included studies were published between 2015 and 2024 (Table 1). The majority of studies were published in 2023 (9/22, 40.91%), followed by 2024 (5/22, 22.73%). Only 3 of the 22 studies (13.64%) are conference papers, whereas 16 (72.73%) are journal articles and 3 (13.64%) are preprints. Studies were carried out in different countries, with the United States and China being the most frequent contributors (6/22, 27.27%) and (5/22, 22.73%), respectively. The number of participants in the included studies ranges from 29 to 2845, with an average of 765.1 (SD 815.6). The mean of female participation rate is 42.45% with a range between 34.5% and 49.3%. The dataset size (ie, number of whole-slide image or patches) varied widely (47 - 27,000), with an average of 2464 (SD 5851). It is worth mentioning that the unit of analysis used to compute performance metrics was predominantly whole-slide image level (21/22, 95.5%), while it was at the patch level in 1 (4.5%) study. The participants' age ranged from 42.2 to 54.5 with an average of 48.03 (SD 4.2; Table 1). The characteristics of each included study are described in Multimedia Appendix 4.

XSL•FO
RenderX

Table . Characteristics of included studies.

Features	Number of studies	References
Year of publication, n (%)		
2024	5 (22.73)	[24,28,35,37,42]
2023	9 (40.91)	[21,22,25,27,30-33,40]
2022	2 (9.09)	[26,34]
2021	3 (13.64)	[29,38,41]
2020	2 (9.09)	[23,36]
2019	1 (4.55)	[39]
Publication type, n (%)		
Journal article	16 (72.73)	[23,24,26,27,29-31,33-38,40-42]
Conference paper	3 (13.64)	[21,28,32]
Preprint	3 (13.64)	[22,25,39]
Country of publication, n (%)		
United States	6 (27.27)	[28,29,32,34,38,39]
China	5 (22.73)	[23,25,35,40,42]
Germany	2 (9.09)	[21,27]
Australia	2 (9.09)	[36,41]
India	1 (4.55)	[22]
Luxembourg	1 (4.55)	[24]
Canada	1 (4.55)	[26]
Switzerland	1 (4.55)	[30]
South Korea	1 (4.55)	[31]
Saudi Arabia	1 (4.55)	[33]
Japan	1 (4.55)	[37]
Number of participants[a], mean (SD); range (%)	765.1 (815.6); 29 - 2845 (95.45)	[21,22,24-42]
Female percentage[b] (%), mean (SD); range (%)	42.45 (0.03); 34.5 - 49.3 (59.09)	[24,25,29,31,33-37,39-42]
Dataset size[c], mean (SD); range (%)	2464 (5851); 47 - 27,000 (95.45)	[22-42]
Mean age[d] (year), mean (SD); range (%)	48.03 (4.2); 42.2 - 54.5 (59.09)	[24,25,29,31,33-37,39-42]

[a]One study did not report the mean age [23].

[b]Studies did not report the female percentage [21-23,26-28,30,32,38].

[c]One study did not report the dataset size [21].

[d]Studies did not report the mean age [21-23,26-28,30,32,38].

Features of Histopathological Images

As presented in Table 2, the histopathological images were captured at various magnification levels, including 2.5×, 5×, 10×, 20×, 40×, and 100×. Among these, 20× magnification was the most commonly used (12/22, 54.55%). While 14 (63.64%) studies focused on architectural-level analysis, 13 (59.09%) studies focused on cellular-level analysis. Slide resolutions varied across studies, with 256×256 pixels being the most common (8/22, 36.36%). Datasets were sourced from either open-source databases (16/22, 72.73%) or closed-source datasets (13/22, 59.09%) (ie, data were collected by the study authors or obtained from previous studies). The open datasets used include The Cancer Genome Atlas (15/22, 68.18%) and the Digital Brain Tumor Atlas (1/22, 4.55%), with some studies combining open and closed sources. The histopathological images were labeled into 2 classes in about two-thirds of the studies (14/22, 63.64%). The features of histopathological images in each study are described in Multimedia Appendix 5.

XSL•FO
RenderX

Table . Features of histopathological images[a].

Features	Number of studies, n (%)	References
Magnification		
2.5×	1 (4.55)	[34]
5×	2 (9.09)	[32,34]
10×	4 (18.18)	[24,29,34,36]
20×[b]	*12* (54.55)	[22,24-26,28,31,32,34,35,40-42]
40×	5 (22.73)	[24,30,31,35,41]
100×	1 (4.55)	[24]
Not reported	7 (31.82)	[21,23,27,33,37-39]
Level of analysis		
Architecture level	*14* (63.64)	[22,24-26,28,29,31,32,34-36,40-42]
Cell level	13 (59.09)	[22,24-26,28,30-32,34,35,40-42]
Not reported	7 (31.82)	[21,23,27,33,37-39]
Slide image resolution		
224×224	4 (18.18)	[29,31,37,42]
256×256	*8* (36.36)	[22,23,25,28,32,34,36,37]
512×512	3 (13.64)	[21,23,38]
1024×1024	5 (22.73)	[23,33,35,36,40]
Other	7 (31.82)	[21,23,24,26,30,41,42]
Not reported	2 (9.09)	[27,39]
Data sources		
Open source	*16* (72.73)	[21-23,25,27-29,31-34,36-39,41]
Closed source	13 (59.09)	[24-27,30,31,33-36,40-42]
Open dataset name		
TCGA[c]	*15* (68.18)	[21,23,25,27-29,31-34,36-39,41]
DBTA[d]	1 (4.55)	[22]
Not applicable	6 (27.27)	[24,26,30,35,40,42]
Number of classes		
Two	*14* (63.64)	[22,23,28-37,41,42]
Three	2 (9.09)	[21,26]
Four	4 (18.18)	[25,27,38,39]
Five	2 (9.09)	[24,33]
Six	1 (4.55)	[40]

[a]Studies used both data sources, open and closed [25,31,33,34,36,41].

[b]Values in italics indicate highest score.

[c]TCGA: The Cancer Genome Atlas.

[d]DBTA: Digital Brain Tumor Atlas.

Features of AI

AI was primarily used for binary classification in 86.36% (19/22) of studies, where models differentiate between two outcomes (eg, IDH-mutant vs wild-type), whereas only 22.73% (5/22) employed multiclass classification, which predicts among 3 or more glioma subtypes (Table 3). The majority of studies (20/22, 90.91%) used AI for IDH subtyping (ie, distinguishing between IDH mutant and IDH wild-type), whereas only 6 (27.27%) studies applied AI for subtyping IDH mutant tumors in relation to 1p/19q gene codeletion. Convolutional neural networks (CNNs) were the most commonly used classifier and feature extraction method, appearing in 54.55% (12/22) and 90.91% (20/22) of studies. For specific classifiers, CNNs are

XSL•FO
RenderX

the most utilized, appearing in 54.55% (12/22) of studies, followed by multiple instance learning (MIL) (which handles weakly labeled slide-level data by learning from sets of image patches without requiring pixel-level annotations) at 50% (11/22), transformers at 18.18% (4/22), and hybrid models (ie, single end-to-end architectures that integrate convolutional components with attention-based mechanisms [eg, CNN-transformer models] within a unified computational graph and are trained jointly) also at 18.18% (4/22). For specific feature extraction models, ResNet50 leads with 46.67% (10/22) usage, followed by densely connected convolutional network 121 (DenseNet121) and InceptionV3, both at 19.05% (4/22). Among the included studies, most models were developed using histopathological images alone (17/22, 77.3%), while a smaller number (5/22, 22.7%) incorporated additional modalities such as MRI scans (2/22, 9.1%), demographic information (3/22, 13.6%), and clinical variables (2/22, 9.1%). The most widely used method for model validation was k-fold cross-validation, used in 50% (11/22) of studies, followed by the train-test split approach in 45.45% (10/22) of studies. AI model performance was assessed using multiple metrics, with accuracy being the most frequently reported (19/22, 86.36%), followed by AUC and sensitivity, each used in 68.18% (15/22) of studies. A detailed summary of the AI models used in each study is provided in Multimedia Appendix 6.

XSL•FO
RenderX

Table . Features of artificial intelligence (AI).

Feature	Number of studies, n (%)	References
Type of classification		
Binary[a]	*19* (86.36)	[22,23,25,27-42]
Multiclass	5 (22.73)	[21,24,26,33,40]
Aim of AI algorithms[b]		
Subtyping IDH[c]	*20* (90.91)	[21-29,32-42]
Subtyping 1p/19q codeletion	6 (27.27)	[25,27,30,31,38,39]
Classifier		
Convolutional neural networks	*12* (54.55)	[24,26,29,30,33-36,38-40,42]
Multiple instance learning	11 (50)	[21-23,25,27,28,31-33,37,42]
Transformers	4 (18.18)	[30,33,35,42]
Hybrid models	2 (9.09)	[35,42]
Ensemble model	2 (9.09)	[34,37]
Logistic regression	2 (9.09)	[31,36]
Random forest	1 (4.55)	[41]
Multilayer perceptron	1 (4.55)	[21]
Specific classifier		
AttMIL[d]	*7* (31.82)	[22,23,25,27,28,33,37]
ResNet50	5 (22.73)	[24,35,36,40,42]
Traditional MIL[e]	4 (18.18)	[21,31,33,42]
InceptionV3	4 (18.18)	[24,33,35,36]
ViT[f]	4 (18.18)	[24,33,35,42]
DenseNet121[g]	3 (13.64)	[24,34,35]
VGG19[h]	3 (13.64)	[24,26,36]
Features extraction methods		
Convolutional neural networks	*20* (90.91)	[21-26,28-40,42]
Transformers	4 (18.18)	[33,35,37,42]
Hybrid models	4 (18.18)	[27,35,37,42]
MIL	1 (4.55)	[42]
Random forest	1 (4.55)	[41]
Specific features extraction model		
ResNet50	*10* (45.45)	[22,24,25,28,31,35-37,40,42]
DenseNet121	5 (22.73)	[21,22,24,34,35]
InceptionV3	4 (18.18)	[24,33,36,37]
ResNet18	3 (13.64)	[24,29,32]
VGG19	2 (9.09)	[24,26,36]
ViT	2 (9.09)	[33,42]
Not reported	2 (9.09)	[23,30]
Data modality		
Unimodel data	17 (77.3)	[22-35,39,40,42]
Multimodel data	5 (22.7)	[21,36-38,41]

XSL•FO
RenderX

Feature	Number of studies, n (%)	References
Types of validation		
k-fold cross-validation	*11* (50)	[21,23,25,27,28,31,33,37-39,41]
Training-test split	10 (45.45)	[22,24,26,29,30,32,34-36,40]
External validation	7 (31.82)	[25,27,29,31,34,40,42]
Performance measures		
Accuracy	*19* (86.36)	[21-25,27,28,31-42]
AUC[i]	15 (68.18)	[21-23,26,28,29,31,33-36,39-42]
Sensitivity	15 (68.18)	[23-25,27,28,30-32,34-37,39,40,42]
F_1-score	9 (40.91)	[21,24,25,31,32,35,37,40,42]
Precision	8 (36.36)	[23,24,27,30-32,35,37]
Specificity	8 (36.36)	[25,27,28,34,36,39,40,42]
Others	8 (36.36)	[21,22,25,27-29,37,39]

[a]Values in italic indicate highest score.

[b]Studies reported results for both IDH mutation status and 1p/19q codeletion status [25,27,38,39].

[c]IDH: isocitrate dehydrogenase.

[d]AttMIL: attention-based multiple instance learning.

[e]MIL: multiple instance learning.

[f]ViT: vision transformer.

[g]DenseNet121: densely connected convolutional network 121.

[h]VGG: visual geometry group.

[i]AUC: area under the curve.

Results of the Risk of Bias and Applicability

In the selection of participants domain, nearly a quarter of the included studies (5/22, 23%) provided sufficient information to confirm the use of an appropriate consecutive or random sampling of eligible participants. Nearly all studies (21/22, 95%) reported a sufficient sample size. Additionally, 8 (36%) studies avoided inappropriate exclusions, while over half (12/22, 55%) demonstrated balanced subgroup representation. Consequently, a small proportion of the studies (2/22, 9%) were assessed as having a low risk of bias, while a larger portion (8/22, 36%) were rated as having an unclear risk of bias in the "selection of participants" domain, as shown in Figure 2.

In terms of matching participants to the predefined requirements of the review question, a low level of concern was identified in 95% (21/22) of the included studies, as shown in Figure 3.

Figure 2. Results of the assessment of risk of bias in the included studies.

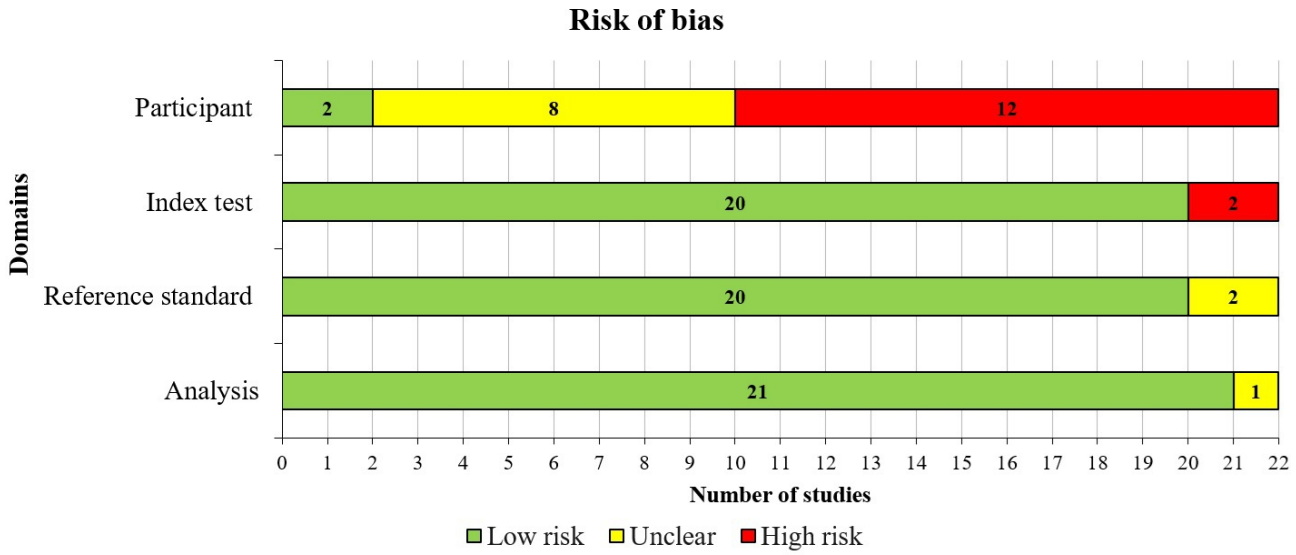

XSL•FO
RenderX

Figure 3. Results of the assessment of applicability concerns in the included studies.

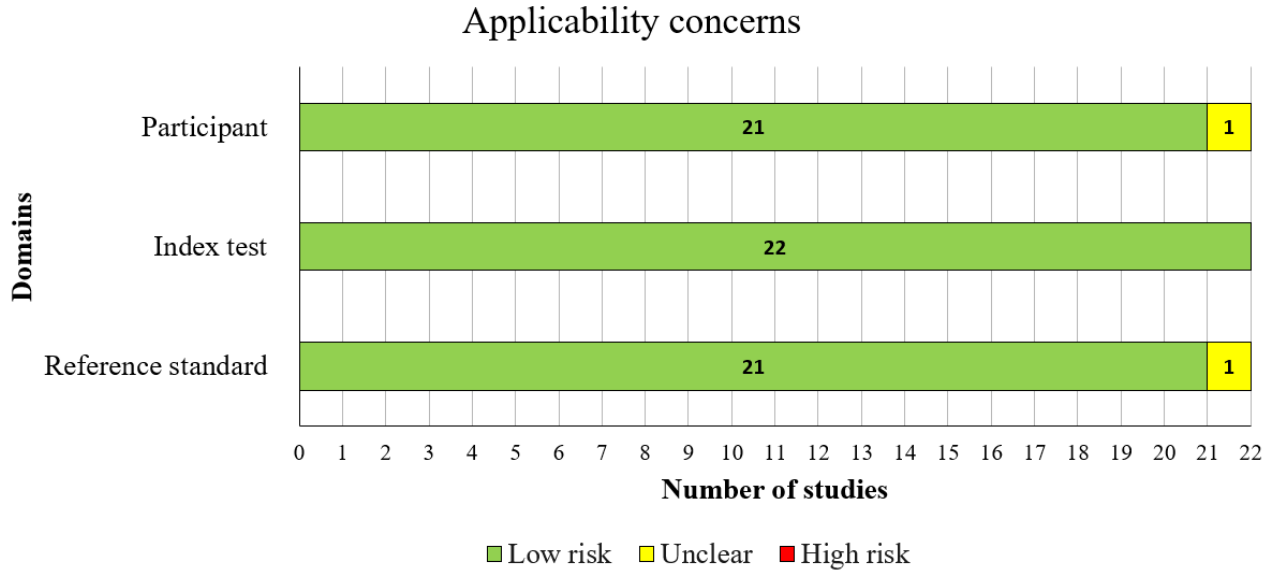

A substantial majority of the included studies comprehensively detailed their AI models, with 21 of the 22 (95%) studies providing thorough descriptions. An overwhelming majority, all 22 studies (100%) clearly reported the features (predictors) used. Moreover, all 22 studies (100%) ensured that these features were sourced without prior knowledge of the outcome data. Consistency in feature assessment across participants was observed in 21 of the 22 (95%) studies. Consequently, the potential for bias in the "index test" domain was assessed as low in 20 of the 22 (91%) studies**,** as shown in Figure 2. In addition, all 22 (100%) studies were found to have minimal concerns regarding the alignment between the model's predictors and the review question's criteria, as illustrated in Figure 3.

In most of the included studies (20/22, 91%), the outcome of interest, histopathology image analysis, was conducted by an expert (eg, pathologist or neuropathologist), ensuring the necessary expertise to accurately classify the outcomes. Nearly all studies (22/22, 100%) applied consistent and uniform outcome definitions across all participants, with outcome classification (ie, image labeling) blinded to the AI model predictions. An overwhelming majority (22/22, 100%) determined the outcome without prior knowledge of the predictor information. Furthermore, in all 22 (100%) studies, the diagnostic process was carried out over an appropriate duration, using the same diagnostic criteria for all participants. As a result, the risk of bias in the "reference standard" domain was considered low in the majority of studies (20/22, 91%), as shown in Figure 2. Additionally, a similar proportion of studies (21/22, 95%) exhibited minimal concerns regarding discrepancies in outcome definition, timing, and determination methods, as shown in Figure 3.

Finally, a significant majority of the studies (22/22, 100%) ensured the inclusion of all enrolled participants in the data analysis. A substantial number of these studies (21/22, 95%) reported no missing values, or if missing values were present, they were handled appropriately (eg, through multiple imputation).

Similarly, a high proportion of studies (22/22, 100%) adopted suitable measures to evaluate the performance of their models. The confusion matrix was presented, or more than one evaluation measure was used, and the selected measures were deemed appropriate.

Nearly half of the studies (21/22, 95%) demonstrated an appropriate split among training, validation, and test sets. The chosen distribution aligned with best practices in the field (eg, 70% - 80% for training, 10% - 15% for validation, and 10% - 20% for testing).

Consequently, the conduct and interpretation of the analysis did not introduce bias in most studies, with 21/22 (95%) having a low risk of bias in the "analysis" domain, as shown in Figure 2. A detailed breakdown of the "risk of bias" and "applicability concerns" for each domain in every study is available in Multimedia Appendix 7.

Results of the Included Studies

As shown in Table 4, the accuracy of AI models in detecting or subtyping molecular markers in adult-type gliomas ranged from 58% to 100%, with a mean accuracy of 85.46%. Sensitivity ranged from 62% to 100%, averaging 84.55%, while specificity ranged from 46% to 100%, with a mean of 86.03%. The AUC values ranged from 46% to 99%, with an average of 86.53%. The hybrid model demonstrated the highest accuracy (92.80%) and sensitivity (89.62%), while the CNN model achieved the highest specificity (89.30%). Logistic regression outperformed other models in terms of AUC (93.05%). Among specific CNN architectures, DenseNet121 achieved the highest accuracy (90.82%), sensitivity (89.28%), and AUC (91.05%). Moreover, AI models that used multimodal data outperformed those that used unimodal data in terms of sensitivity (90.15% vs 84.31%) and AUC (88.93% vs 86.29%). In outcome prediction, AI models showed better performance in detecting IDH mutations compared to 1p/19q codeletions, with higher accuracy (86.13% vs 81.63%), specificity (86.61% vs 78.11%), and AUC (86.74% vs 85.15%). AI models also performed better in IDH subtyping for multiclass classification compared to binary classification

XSL•FO
RenderX

in terms of accuracy (91.98% vs 84.02%) and sensitivity (93.41% vs 80.18%). Conversely, AI models for binary classification yielded a lower performance than multiclass classification in terms of accuracy (91.98% vs 84.02%) and sensitivity (93.41% vs 80.18%). However, these differences should be interpreted as descriptive trends rather than statistically validated superiority, as formal between-group comparisons were not feasible.

XSL•FO
RenderX

Table . Summary of overall performance of all models based on number of classes, type of data, and artificial intelligence (AI) algorithms in the target variable[a].

Groups	Accuracy			Sensitivity			Specificity			AUC[b]		
	Number of studies	Mean % (SD)	Range	Number of studies	Mean % (SD)	Range	Number of studies	Mean % (SD)	Range	Number of studies	Mean % (SD)	Range
Algorithms												
CNNs[c]	8	85.60 (0.09)	0.72 - 1.00	8	84.09 (0.11)	0.64 - 1.00	7	*89.30*[d] (0.09)	0.75 - 1.00	8	86.54 (0.06)	0.76 - 0.99
MIL[e]	10	80.83 (0.09)	0.58 - 0.91	8	82.54 (0.13)	0.62 - 0.99	3	68.69 (0.23)	0.46 - 0.94	10	81.73 (0.16)	0.46 - 0.97
Transformer	4	88.27 (0.03)	0.83 - 0.92	3	81.62 (0.10)	0.69 - 0.92	—	—	—	3	92.37 (0.04)	0.89 - 0.96
Hybrid model	3	*92.80* (0.05)	0.85 - 0.97	3	*89.62* (0.08)	0.81 - 0.97	—	—	—	2	92.20 (0.04)	0.89 - 0.95
Ensemble model	2	86.35 (0.12)	0.78 - 0.95	2	89.60 (0.05)	0.86 - 0.93	—	—	—	2	92.00 (0.09)	0.85 - 0.99
Logistic regression	2	87.15 (0.01)	0.86 - 0.88	2	86.00 (0.12)	0.78 - 0.94	—	—	—	2	*93.05* (0.00)	0.93 - 0.931
CNN models												
ResNet (ResNet18,50)	7	87.04 (0.09)	0.72 - 1.00	6	87.51 (0.11)	0.69 - 1.00	5	*91.48* (0.07)	0.84 - 1.00	6	90.19 (0.06)	0.81 - 0.99
VGG[f] (VGG16,19)	4	89.67 (0.09)	0.79 - 1.00	4	87.03 (0.13)	0.69 - 1.00	2	90.65 (0.13)	0.81 - 1.00	2	81.55 (0.01)	0.81 - 0.82
DenseNet[g]	3	*90.82* (0.10)	0.79 - 0.97	3	*89.28* (0.10)	0.78 - 0.97	—	—	—	2	*91.05* (0.08)	0.86 - 0.97
InceptionV3	3	80.15 (0.05)	0.77 - 0.86	3	75.82 (0.11)	0.64 - 0.86	—	—	—	3	83.58 (0.03)	0.80 - 0.87
Type of data												
Unimodel data	20	*85.67* (0.09)	0.58 - 1.00	16	84.31 (0.11)	0.62 - 1.00	9	*86.27* (0.15)	0.46 - 1.00	16	86.29 (0.10)	0.46 - 0.99
Multimodel data	5	83.64 (0.05)	0.78 - 0.90	2	*90.15* (0.06)	0.86 - 0.94	—	—	—	4	*88.93* (0.03)	0.85 - 0.93
Outcome prediction												
1p/19q	6	81.63 (0.12)	0.58 - 1.00	6	*87.64* (0.09)	0.75 - 1.00	3	78.11 (0.28)	0.46 - 1.00	5	85.15 (0.19)	0.46 - 0.99
IDH[h]	19	*86.13* (0.08)	0.72 - 1.00	14	83.96 (0.11)	0.62 - 1.00	8	*86.61* (0.12)	0.52 - 1.00	16	*86.74* (0.08)	0.52 - 0.99
Class prediction												
Binary IDH	16	84.02 (0.07)	0.72 - 1.00	12	80.18 (0.11)	0.62 - 1.00	8	*86.61* (0.12)	0.52 - 1.00	15	*86.87* (0.08)	0.52 - 0.99
Multiclass IDH	3	*91.98* (0.07)	0.75 - 1.00	2	*93.41* (0.05)	0.85 - 1.00	—	—	—	—	—	—
Overall	*20*	*85.46* (0.08)	*0.58 - 1.00*	*16*	*84.55* (0.01)	*0.62 - 1.00*	*9*	*86.03* (0.15)	*0.46 - 1.00*	*17*	*86.53* (0.10)	*0.46 - 0.99*

[a]Specificity not reported in all studies.

[b]AUC: area under the curve.

[c]CNNs: convolutional neural networks.

XSL•FO
RenderX

[d]Values in italics indicate highest score.

[e]MIL: multiple instance learning.

[f]VGG: visual geometry group.

[g]DenseNet121: densely connected convolutional network 121.

[h]IDH: isocitrate dehydrogenase.

Discussion

Principal Findings

This review aimed to systematically evaluate the performance of AI models in detecting and classifying IDH mutation status and 1p/19q codeletion in adult-type gliomas using histopathological images. Our analysis indicates that these models demonstrate promising diagnostic capabilities, achieving pooled mean values of 85.46% for accuracy, 84.55% for sensitivity, 86.03% for specificity, and 86.53% for the area under the receiver operating characteristic curve. As noticed, the performance across all metrics was comparable, suggesting that AI models offer balanced diagnostic capabilities. Despite these promising results, they remain below the commonly cited thresholds for clinical adoption, which often exceed 90% for high-stakes diagnostic tasks [42]. As such, the current evidence is not sufficient to support stand-alone clinical decision-making. Therefore, the results of this meta-analysis should be interpreted as early indicators of feasibility rather than clinically actionable performance, underscoring the substantial methodological and validation work still required before these models can be safely integrated into clinical workflows.

These results are generally consistent with prior reviews. Specifically, a systematic review conducted by Farahani et al [43] reported a similar performance in IDH prediction using MRI, with pooled sensitivity of 84% and specificity of 87% and an AUC of 0.89. Further, Chen et al [44] found pooled sensitivity of 83%, specificity of 84%, and AUC of 0.90 for predicting IDH mutations using machine learning–based radiomics. Taken together, these comparisons suggest that our findings are in line with the broader body of research, reinforcing the reliability and clinical promise of AI-based histopathological approaches for glioma molecular classification.

Individual studies showed considerable variability across performance metrics; for instance, specificity values ranged from 46% to 100%. This variability likely stems from population differences, sample sizes, model architectures, and the inherent challenge of predicting rare molecular alterations (eg, telomerase reverse transcriptase mutations) from histopathology images, compounded by class imbalance, domain shift (ie, performance degradation when models are applied to data from different sources), and the absence of distinct histological correlates. While accuracy and sensitivity remain higher for well-characterized targets (eg, IDH, 1p/19q), specificity suffers when the model encounters rare or ambiguous cases in external validation.

Hybrid models achieved the highest mean accuracy (92.80%) and sensitivity (89.62%), while logistic regression models achieved the top mean AUC just slightly higher than hybrid models (93.05% vs 92.20%). These results suggest that hybrid models deliver the strongest overall performance. Hybrid models are defined as AI frameworks that integrate multiple deep learning architectures to leverage the complementary strengths of each component. In the context of glioma diagnosis, most recent reviews have highlighted the effectiveness of hybrid models combining CNNs with transformer-based architectures [45-47]. CNNs excel at extracting local spatial features from histopathological images, such as cellular morphology and texture patterns, while transformers are adept at capturing long-range dependencies and global contextual information through their self-attention mechanisms. By combining these capabilities, hybrid CNN-transformer models can achieve a more comprehensive understanding of complex tissue structures and molecular characteristics within gliomas, leading to improved predictive performance. Our findings align with a recent review by Redlich et al [45], which analyzed 83 studies on AI-based methods for whole-slide histopathology image analysis of gliomas. The review highlighted that hybrid models combining CNNs with other architectures, including transformers, consistently outperform single-architecture models in tasks such as subtyping, grading, and molecular marker prediction [45]

Among pure CNN backbones, DenseNet121 demonstrated the best performance, achieving an accuracy of 90.82%, sensitivity of 89.28%, and an AUC of 91.05%. DenseNet121's superior performance may be due to its unique architectural features, notably dense connectivity, which enhance training dynamics, feature representation, and parameter efficiency.

In terms of data modalities employed by AI models, our analysis demonstrated that multimodal approaches exhibited superior performance in sensitivity (90.15% vs 84.31%) and AUC (88.93% vs 86.29%) compared to unimodal models. This suggests that multimodal models benefit from richer and more diverse feature representations, thereby enhancing their diagnostic capability. However, unimodal models achieved slightly higher accuracy (85.67% vs 83.64%), which may be attributed to the fact that none of the five multimodal studies included in the accuracy analysis employed hybrid modeling. Instead, they utilized simple model ensembles or comparisons involving architectures such as multilayer perceptron, logistic regression, random forests, and CNNs. Our findings are consistent with existing literature. For instance, a systematic review by Alleman et al [48] concluded that multimodal deep learning approaches enhance prognostic accuracy in glioma patients compared to unimodal models, particularly in predicting overall survival. Similarly, d'Este et al [49] demonstrated that combining multimodal imaging with AI techniques improves visualization of glioma infiltration, surpassing the capabilities of single-modality approaches. Furthermore, the GlioMT model, introduced by Byeon et al [50], which integrates imaging and clinical data, exhibited superior performance in predicting adult-type diffuse gliomas compared to conventional CNNs and visual transformers.

XSL•FO
RenderX

Regarding outcome prediction and AI model performance in detecting molecular markers, our analysis revealed that AI models demonstrated stronger performance in identifying IDH mutations compared to 1p/19q codeletions. Specifically, IDH detection yielded higher accuracy (86.13% vs 81.63%), specificity (86.61% vs 78.11%), and AUC (86.74% vs 85.15%). In contrast, AI models achieved slightly better sensitivity for 1p/19q codeletion detection (87.64% vs 83.96%). The discrepancy in performance between IDH mutation and 1p/19q codeletion detection may be attributed to several factors. IDH mutations often result in more distinct histopathological features, making them more amenable to detection by AI models. In contrast, 1p/19q codeletions may not produce as pronounced morphological changes, posing a greater challenge for AI-based detection. The findings of published studies align with our results. For example, the review by Farahani et al [43] showed higher performance in identifying IDH mutations compared to 1p/19q codeletions using MRI in terms of sensitivity (84% vs 76%) and specificity (87% vs 85%).

In our systematic review, CNNs and MIL algorithms were the most predominantly used in AI models, likely due to their suitability for histopathological image analysis. CNNs are designed to capture spatial hierarchies in image data through convolutional layers, making them effective at identifying morphological patterns such as cellular structures and tissue architecture that are critical in tumor classification. Histopathological slides of gliomas are typically of high resolution and contain complex visual features, which CNNs can learn to recognize with high accuracy. Meanwhile, MIL frameworks address the challenge of weakly labeled data, which is common in medical imaging, where only slide-level labels (eg, tumor type or mutation status) are available, rather than precise annotations at the pixel or region level. MIL allows models to learn from sets of image patches (instances) while associating predictions with the whole slide (bag), making it particularly useful in computational pathology where exhaustive labeling is impractical.

Practical and Research Implications

While our results are encouraging, they remain not optimal and arguably lower than that acceptable in clinical practice [42]. Thus, our findings should be interpreted with caution due to several limitations. First, 15 of the 22 studies included in the review utilized the same dataset, raising concerns regarding the generalizability of the results. Second, many of the studies were based on small sample sizes, limiting the robustness and reliability of the conclusions. Third, several key findings were derived from fewer than 5 studies, reducing the strength of the evidence. Lastly, the aggregation of results using simple arithmetic means rather than formal meta-analytic techniques may have limited the ability to draw definitive conclusions about the overall efficacy of AI models.

Given these limitations, AI should not be used as a stand-alone diagnostic tool for the detection or subtyping of adult-type gliomas but rather as a complementary approach alongside conventional methods. Future research should prioritize the development of AI models using larger and more diverse datasets to enhance performance and generalizability. Ensuring adequate sample sizes and improving the transparency of reporting, such as including confusion matrices, will also facilitate more rigorous meta-analyses. Notably, the majority of studies reviewed developed AI models exclusively from histopathological images. Developing multimodal AI models that incorporate additional data types (eg, radiological and clinical information) may further enhance diagnostic accuracy. Thus, continued efforts are needed to advance the integration of multimodal data in AI model development for the detection and subtyping of adult-type gliomas.

In this review, the majority of studies used AI algorithms that were either MIL or CNNs, indicating a research gap in exploring the performance and effectiveness of other types of AI algorithms (eg, transformers, hybrid models, ensemble models [ie, approaches that combine predictions from multiple independently trained models using averaging, voting, or weighted fusion strategies], or logistic regression). Additionally, we observed a lack of research assessing the performance of AI models that use multimodal data, which may have affected the accuracy of the results. Moreover, the number of studies evaluating AI models' performance in detecting IDH mutations was noticeably greater than those that used 1p/19q as a predicted outcome. These studies predominantly employed binary IDH classification (ie, IDH mutation vs IDH wild-type), whereas studies using IDH multiclass classification (eg, astrocytoma, oligodendroglioma) were markedly fewer. Future research should address these gaps to produce more accurate and representative findings.

Our findings that AI hybrid models performed better in detecting adult-type gliomas than other models underscore the need for more advanced approaches to explore their effectiveness and their potential to achieve diagnostic outcomes comparable to those of histopathologists. However, their real-world applicability requires consideration of computational and operational demands. These models often require greater processing power (eg, GPU-enabled servers or high-performance cloud computing) and longer inference times compared to conventional CNNs. Such requirements may pose challenges for pathology laboratories lacking advanced digital infrastructure. Nonetheless, many modern clinical centers are increasingly equipped with GPU-accelerated digital pathology ecosystems, and cloud-based deployment can significantly reduce the need for on-site hardware. From a cost-benefit perspective, the additional computational resources required by hybrid models may be justified given their superior accuracy and robustness, particularly for molecular prediction tasks where diagnostic errors carry serious clinical implications. However, successful implementation in routine practice will depend on streamlined model optimization, efficient inference pipelines, and prospective validation of runtime performance to ensure that diagnostic gains meaningfully outweigh operational costs.

As our findings indicated superior performance of multimodal AI models in sensitivity and AUC, we suggest a complementary approach: multimodal AI models may serve as effective initial diagnostic tools due to their comprehensive data integration, while unimodal models can be employed subsequently for more accurate and specific outcome determination. Moreover, our results reflect a stronger performance of AI models in the

XSL•FO
RenderX

detection of IDH mutations compared to 1p/19q codeletion in all metrics except sensitivity. Therefore, we suggest that while IDH mutation detection should remain a primary focus for model refinement and deployment due to its balanced high performance, models targeting 1p/19q codeletions may serve as effective initial diagnostic tools in adult-type gliomas due to their higher sensitivity, which is critical in early disease detection.

Given that the current review focused on IDH mutations and 1p/19q codeletions, we encourage researchers to undertake further reviews covering other glioma molecular markers, including MGMT methylation, ATRX, and telomerase reverse transcriptase mutations. Moreover, further research should extend to other types of brain tumors, such as meningiomas, pituitary adenomas, and schwannomas. These tumors differ significantly in terms of biological behavior, treatment options, and clinical outcomes, warranting a comprehensive evaluation of AI's potential in diagnosing and subtyping each tumor type.

Limitations

This systematic review is subject to several limitations. First, the generalizability of the findings is limited, as approximately 68.18% of the included studies relied on the same dataset, and the number of included studies is relatively small (n=22). This raises concerns about the external validity of the results and their applicability to the broader population of adult-type glioma patients. To address the potential duplication of patients or slides across studies (ie, those relying on The Cancer Genome Atlas), we reviewed the data sources reported by each study to identify where overlap was likely. Because individual-level data were not available, we were unable to verify or remove duplicate cases. Instead, we treated studies drawing from the same dataset as nonindependent samples and interpreted pooled estimates with caution. As a result, our summary metrics should be viewed as descriptive indicators rather than independent, unbiased effect estimates. Second, the review focused exclusively on adult-type gliomas, thereby restricting the ability to draw conclusions about the performance of AI in detecting other adult brain tumors or brain tumors in the general population. Third, the analysis was confined to studies utilizing histopathological images, limiting the assessment of AI performance with alternative imaging modalities, such as MRI or computed tomography scans. Fourth, meta-analysis was not feasible due to incomplete reporting of key performance metrics across studies. Instead, mean values were calculated to provide a descriptive summary, though this approach lacks the statistical rigor and precision of formal meta-analysis. Finally, the inclusion of only English-language studies may have introduced language bias and led to the exclusion of relevant research published in other languages, potentially affecting the comprehensiveness of the review.

Conclusion

This systematic review demonstrates that AI models applied to histopathological images show strong potential for the molecular classification of adult-type gliomas, particularly in detecting IDH mutations and 1p/19q codeletions. Our findings indicate that while overall diagnostic performance is promising—especially for IDH mutations—variability in model performance and methodological limitations across studies temper the generalizability of results. Hybrid models and multimodal approaches emerged as particularly effective strategies, offering enhanced sensitivity and discriminative capability. However, the limited use of diverse datasets, the predominance of unimodal designs, and the lack of standardized reporting remain critical barriers to clinical translation. Given these constraints, AI models should currently serve as complementary tools rather than stand-alone diagnostic systems, functioning primarily to assist pathologists in pattern recognition and decision support while final diagnoses remain clinician-led. Future research should prioritize methodological standardization, broader data integration, and validation across larger, heterogeneous populations to support the clinical deployment of AI in glioma diagnostics.

Acknowledgments

The authors declare the use of generative AI in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to generative artificial intelligence (GAI) tools under full human supervision: Proofreading and editing. The GAI tool used was ChatGPT-4. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. All intellectual content, study design, data collection, analysis, and final interpretations are the sole responsibility of the authors.

Funding

The publication of this article was funded by the Weill Cornell Medicine–Qatar Health Sciences Library.

Data Availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Authors' Contributions

Concept for the paper and experimental design: AA, RAD, and OA
Data collection and extraction: AA, RAD, and OA
Data analysis: AA and OA

XSL•FO
RenderX

Risk of bias assessment: AT
Writing original draft preparation: AO, RAD, and OA
Writing-review and editing: AA, AO, RAD, and OA
Supervision: AA
All authors agreed to submission of the manuscript.

Conflicts of Interest

AA is an associate editor of *JMIR Nursing* at the time of publication. All other authors declare no conflict of interest.

Multimedia Appendix 1
Search strategy.
[DOCX File, 30 KB - jmir_v28i1e78377_app1.docx]

Multimedia Appendix 2
Data extraction form.
[DOCX File, 23 KB - jmir_v28i1e78377_app2.docx]

Multimedia Appendix 3
Modified version of QUADAS-2.
[DOCX File, 28 KB - jmir_v28i1e78377_app3.docx]

Multimedia Appendix 4
Characteristics of each included study.
[DOCX File, 43 KB - jmir_v28i1e78377_app4.docx]

Multimedia Appendix 5
Features of histopathological images in each study.
[DOCX File, 39 KB - jmir_v28i1e78377_app5.docx]

Multimedia Appendix 6
Features of artificial intelligence in each study.
[DOCX File, 44 KB - jmir_v28i1e78377_app6.docx]

Multimedia Appendix 7
Reviewers' judgments about each domain in "risk of bias" and "applicability concerns" for each included study.
[DOCX File, 218 KB - jmir_v28i1e78377_app7.docx]

Checklist 1
PRISMA-DTA checklist.
[DOC File, 67 KB - jmir_v28i1e78377_app8.doc]

References

1. Ostrom QT, Cioffi G, Waite K, Kruchko C, Barnholtz-Sloan JS. CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2014–2018. Neuro Oncol 2021 Oct 5;23(12 Suppl 2):iii1-iii105. [doi: 10.1093/neuonc/noab200] [Medline: 34608945]
2. Reynoso-Noverón N, Mohar-Betancourt A, Ortiz-Rafael J. Epidemiology of brain tumors. In: Principles of Neuro-Oncology: Springer; 2020:15-25. [doi: 10.1007/978-3-030-54879-7_2]
3. Lapointe S, Perry A, Butowski NA. Primary brain tumours in adults. The Lancet 2018 Aug;392(10145):432-446. [doi: 10.1016/S0140-6736(18)30990-5]
4. Perry A, Wesseling P. Histologic classification of gliomas. Handb Clin Neurol 2016;134:71-95. [doi: 10.1016/B978-0-12-802997-8.00005-0] [Medline: 26948349]
5. Chen R, Smith-Cohn M, Cohen AL, Colman H. Glioma subclassifications and their clinical significance. Neurotherapeutics 2017 Apr;14(2):284-297. [doi: 10.1007/s13311-017-0519-x] [Medline: 28281173]
6. Kalidindi N, Or R, Babak S, Mason W. Molecular classification of diffuse gliomas. Can J Neurol Sci 2020 Jul;47(4):464-473. [doi: 10.1017/cjn.2020.10] [Medline: 31918786]

XSL•FO
RenderX

7. Antonelli M, Poliani PL. Adult type diffuse gliomas in the new 2021 WHO Classification. Pathologica 2022 Dec;114(6):397-409. [doi: 10.32074/1591-951X-823] [Medline: 36534419]
8. Hollon T, Jiang C, Chowdury A, et al. Artificial-intelligence-based molecular classification of diffuse gliomas using rapid, label-free optical imaging. Nat Med 2023 Apr;29(4):828-832. [doi: 10.1038/s41591-023-02252-4] [Medline: 36959422]
9. Koski E, Murphy J. AI in healthcare. Stud Health Technol Inform 2021 Dec 15;284:295-299. [doi: 10.3233/SHTI210726] [Medline: 34920529]
10. van Kempen EJ, Post M, Mannil M, et al. Accuracy of machine learning algorithms for the classification of molecular features of gliomas on MRI: a systematic literature review and meta-analysis. Cancers (Basel) 2021 May 26;13(11):2606. [doi: 10.3390/cancers13112606] [Medline: 34073309]
11. Farahani S, Hejazi M, Moradizeyveh S, Di Ieva A, Fatemizadeh E, Liu S. Diagnostic accuracy of deep learning models in predicting glioma molecular markers: a systematic review and meta-analysis. Diagnostics (Basel) 2025 Mar 21;15(7):797. [doi: 10.3390/diagnostics15070797] [Medline: 40218147]
12. Puustinen S, Vrzáková H, Hyttinen J, et al. Hyperspectral imaging in brain tumor surgery-evidence of machine learning-based performance. World Neurosurg 2023 Jul;175:e614-e635. [doi: 10.1016/j.wneu.2023.03.149] [Medline: 37030483]
13. Singh J, Sahu S, Mohan T, et al. Current status of DNA methylation profiling in neuro-oncology as a diagnostic support tool: a review. Neuro oncol Pract 2023 Dec;10(6):518-526. [doi: 10.1093/nop/npad040] [Medline: 38009119]
14. Liu Y, Wu M. Deep learning in precision medicine and focus on glioma. Bioeng Transl Med 2023 Sep;8(5):e10553. [doi: 10.1002/btm2.10553]
15. Zhao J, Huang Y, Song Y, et al. Diagnostic accuracy and potential covariates for machine learning to identify IDH mutations in glioma patients: evidence from a meta-analysis. Eur Radiol 2020 Aug;30(8):4664-4674. [doi: 10.1007/s00330-020-06717-9] [Medline: 32193643]
16. Lv Q, Liu Y, Sun Y, Wu M. Insight into deep learning for glioma IDH medical image analysis: a systematic review. Medicine (Baltimore) 2024 Feb 16;103(7):e37150. [doi: 10.1097/MD.0000000000037150] [Medline: 38363910]
17. Guha A, Halder S, Shinde SH, et al. How does deep learning/machine learning perform in comparison to radiologists in distinguishing glioblastomas (or grade IV astrocytomas) from primary CNS lymphomas? A meta-analysis and systematic review. Clin Radiol 2024 Jun;79(6):460-472. [doi: 10.1016/j.crad.2024.03.007] [Medline: 38614870]
18. McInnes MDF, Moher D, Thombs BD, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA 2018 Jan 23;319(4):388-396. [doi: 10.1001/jama.2017.19163] [Medline: 29362800]
19. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
20. Wolff RF, Moons KGM, Riley RD, et al. Probast: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019 Jan 1;170(1):51-58. [doi: 10.7326/M18-1376] [Medline: 30596875]
21. Albuquerque T, Fang ML, Wiestler B, et al. Multimodal context-aware detection of glioma biomarkers using MRI and WSI. 2023 Presented at: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 Workshops; Oct 8-12, 2023; Vancouver, BC, Canada p. 157-167. [doi: 10.1007/978-3-031-47425-5_15]
22. Chitnis SR, Liu S, Dash T, et al. Domain-specific pre-training improves confidence in whole slide image classification. 2023 Presented at: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Jul 24-27, 2023. [doi: 10.1109/EMBC40787.2023.10340659]
23. Cui D, Liu Y, Liu G, Liu L. A multiple-instance learning-based convolutional neural network model to detect the IDH1 mutation in the histopathology images of glioma tissues. J Comput Biol 2020 Aug;27(8):1264-1272. [doi: 10.1089/cmb.2019.0410] [Medline: 31905004]
24. Despotovic V, Kim SY, Hau AC, et al. Glioma subtype classification from histopathological images using in-domain and out-of-domain transfer learning: an experimental study. Heliyon 2024 Mar 15;10(5):e27515. [doi: 10.1016/j.heliyon.2024.e27515] [Medline: 38562501]
25. Fang Z, Liu Y, Wang Y, et al. Deep learning predicts biomarker status and discovers related histomorphology characteristics for low-grade glioma. arXiv. Preprint posted online on Oct 11, 2023. [doi: 10.48550/arXiv.2310.07464]
26. Faust K, Lee MK, Dent A, et al. Integrating morphologic and molecular histopathological features through whole slide image registration and deep learning. Neurooncol Adv 2022;4(1):vdac001. [doi: 10.1093/noajnl/vdac001] [Medline: 35156037]
27. Hewitt KJ, Löffler CML, Muti HS, et al. Direct image to subtype prediction for brain tumors using deep learning. Neurooncol Adv 2023;5(1):vdad139. [doi: 10.1093/noajnl/vdad139] [Medline: 38106649]
28. Innani S, Baheti B, Nasrallah MP, Bakas S. Weakly supervised IDH-status glioma classification from H&E-stained whole slide images. 2024 Presented at: 2024 IEEE International Symposium on Biomedical Imaging (ISBI); May 27-30, 2024. [doi: 10.1109/ISBI56570.2024.10635869]
29. Jiang S, Zanazzi GJ, Hassanpour S. Predicting prognosis and IDH mutation status for patients with lower-grade gliomas using whole slide images. Sci Rep 2021 Aug 19;11(1):16849. [doi: 10.1038/s41598-021-95948-x] [Medline: 34413349]

XSL•FO
RenderX

30. Jungo P, Hewer E. Code-free machine learning for classification of central nervous system histopathology images. J Neuropathol Exp Neurol 2023 Feb 21;82(3):221-230. [doi: 10.1093/jnen/nlac131] [Medline: 36734664]
31. Kim GJ, Lee T, Ahn S, Uh Y, Kim SH. Efficient diagnosis of IDH-mutant gliomas: 1p/19qNET assesses 1p/19q codeletion status using weakly-supervised learning. NPJ Precis Onc 2023;7(1):94. [doi: 10.1038/s41698-023-00450-4]
32. Krebs O, Agarwal S, Tiwari P. Self-supervised deep learning to predict molecular markers from routine histopathology slides for high-grade glioma tumors. Presented at: Proceedings Volume 12471, Medical Imaging 2023: Digital and Computational Pathology; Feb 19-23, 2023. [doi: 10.1117/12.2653929]
33. Li Z, Cong Y, Chen X, et al. Vision transformer-based weakly supervised histopathological image analysis of primary brain tumors. iScience 2023 Jan;26(1):105872. [doi: 10.1016/j.isci.2022.105872]
34. Liechty B, Xu Z, Zhang Z, et al. Machine learning can aid in prediction of IDH mutation from H&E-stained histology slides in infiltrating gliomas. Sci Rep 2022 Dec 31;12(1):22623. [doi: 10.1038/s41598-022-26170-6] [Medline: 36587030]
35. Liu X, Hu W, Diao S, Abera DE, Racoceanu D, Qin W. Multi-scale feature fusion for prediction of IDH1 mutations in glioma histopathological images. Comput Methods Programs Biomed 2024 May;248:108116. [doi: 10.1016/j.cmpb.2024.108116] [Medline: 38518408]
36. Liu S, Shah Z, Sav A, et al. Isocitrate dehydrogenase (IDH) status prediction in histopathology images of gliomas using deep learning. Sci Rep 2020;10(1):7733. [doi: 10.1038/s41598-020-64588-y]
37. Nakagaki R, Debsarkar SS, Kawanaka H, Aronow BJ, Prasath VBS. Deep learning-based IDH1 gene mutation prediction using histopathological imaging and clinical data. Comput Biol Med 2024 Sep;179:108902. [doi: 10.1016/j.compbiomed.2024.108902] [Medline: 39038392]
38. Pei L, Jones KA, Shboul ZA, Chen JY, Iftekharuddin KM. Deep neural network analysis of pathology images with integrated molecular data for enhanced glioma classification and grading. Front Oncol 2021;11:668694. [doi: 10.3389/fonc.2021.668694] [Medline: 34277415]
39. Rathore S, Iftikhar M, Nasrallah M, Gurcan M, Rajpoot N, Mourelatos Z. Tmod-35. Prediction of overall survival, and molecular markers in gliomas via analysis of digital pathology images using deep learning. Neuro Oncol 2019 Nov 11;21(Supplement_6):vi270. [doi: 10.1093/neuonc/noz175.1134]
40. Wang W, Zhao Y, Teng L, et al. Neuropathologist-level integrated classification of adult-type diffuse gliomas using deep learning from whole-slide pathological images. Nat Commun 2023;14(1):6359. [doi: 10.1038/s41467-023-41195-9]
41. Wang D, Liu C, Wang X, et al. Automated machine-learning framework integrating histopathological and radiological information for predicting IDH1 mutation status in glioma. Front Bioinform 2021;1:718697. [doi: 10.3389/fbinf.2021.718697] [Medline: 36303770]
42. Zhao Y, Wang W, Ji Y, et al. Computational pathology for prediction of isocitrate dehydrogenase gene mutation from whole slide images in adult patients with diffuse glioma. Am J Pathol 2024 May;194(5):747-758. [doi: 10.1016/j.ajpath.2024.01.009] [Medline: 38325551]
43. Farahani S, Hejazi M, Tabassum M, Di Ieva A, Mahdavifar N, Liu S. Diagnostic performance of deep learning for predicting glioma isocitrate dehydrogenase and 1p/19q co-deletion in MRI: a systematic review and meta-analysis. Eur Radiol 2025 Aug 16. [doi: 10.1007/s00330-025-11898-2] [Medline: 40817944]
44. Chen X, Lei J, Wang S, Zhang J, Gou L. Diagnostic accuracy of a machine learning-based radiomics approach of MR in predicting IDH mutations in glioma patients: a systematic review and meta-analysis. Front Oncol 2024;14:1409760. [doi: 10.3389/fonc.2024.1409760]
45. Redlich JP, Feuerhake F, Weis J, et al. Applications of artificial intelligence in the analysis of histopathology images of gliomas: a review. Npj Imaging 2024 Jul 1;2(1):16. [doi: 10.1038/s44303-024-00020-8] [Medline: 40603567]
46. Nobel SMN, Swapno S, Islam MB, et al. A novel mixed convolution transformer model for the fast and accurate diagnosis of glioma subtypes. Adv Intell Syst 2025 May;7(5):2400566. [doi: 10.1002/aisy.202400566]
47. Henry EU, Emebo O, Omonhinmin CA. Vision transformers in medical imaging: a review. arXiv. Preprint posted online on Nov 18, 2022. [doi: 10.48550/arXiv.2211.10043]
48. Alleman K, Knecht E, Huang J, Zhang L, Lam S, DeCuypere M. Multimodal deep learning-based prognostication in glioma patients: a systematic review. Cancers (Basel) 2023 Jan 16;15(2):545. [doi: 10.3390/cancers15020545] [Medline: 36672494]
49. d'Este SH, Nielsen MB, Hansen AE. Visualizing glioma infiltration by the combination of multimodality imaging and artificial intelligence, a systematic review of the literature. Diagnostics (Basel) 2021 Mar 25;11(4):592. [doi: 10.3390/diagnostics11040592] [Medline: 33806195]
50. Byeon Y, Park YW, Lee S, et al. Interpretable multimodal transformer for prediction of molecular subtypes and grades in adult-type diffuse gliomas. NPJ Digit Med 2025 Mar 5;8(1):140. [doi: 10.1038/s41746-025-01530-4] [Medline: 40044878]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
CNN: convolutional neural network
CNS: central nervous system

XSL•FO
RenderX

DBTA: Digital Brain Tumor Atlas
DenseNet121: densely connected convolutional network 121
IDH: isocitrate dehydrogenase
MIL: multiple instance learning
MRI: magnetic resonance imaging
PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Extension for Diagnostic Test Accuracy
PROBAST: Prediction Model Risk of Bias Assessment Tool
PROSPERO: International Prospective Register of Systematic Reviews
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2
WHO: World Health Organization

Edited by A Schwartz; submitted 01.Jun.2025; peer-reviewed by BS Ibrahim, K Dhou, Y Lu; accepted 06.Jan.2026; published 13.Mar.2026.

Please cite as:
Almaabreh O, Al-Dafi R, Tabassum A, Othman A, Abd-alrazaq A
The Performance of Artificial Intelligence in Classifying Molecular Markers in Adult-Type Gliomas Using Histopathological Images: Systematic Review
J Med Internet Res 2026;28:e78377
URL: https://www.jmir.org/2026/1/e78377
doi:10.2196/78377

© Obada Almaabreh, Rukaya Al-Dafi, Aliya Tabassum, Ahmad Othman, Alaa Abd-alrazaq. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

eHealth Literacy and Type 2 Diabetes Prevention Among At-Risk Populations: Mechanistic Systematic Review Using Theory-Driven Thematic Analysis

Jingyi Li[1,2], MA; Arina Anis Azlan[1,3], PhD; Nurzihan Hassim[1], PhD; Yuan Wang[1,4], MSocSci; Ruina Guo[1], MJC

[1]Centre for Research in Media and Communication, Faculty of Social Sciences and Humanities, National University of Malaysia, UKM, Selangor, Malaysia

[2]Shanxi Technology and Business University, Taiyuan, Shanxi, China

[3]Komunikasi Kesihatan (Healthcomm) - UKM Research Group, National University of Malaysia, Selangor, Malaysia

[4]Yuncheng University, Yuncheng, Shanxi, China

Corresponding Author:
Arina Anis Azlan, PhD
Centre for Research in Media and Communication, Faculty of Social Sciences and Humanities, National University of Malaysia, UKM, Selangor, Malaysia

Abstract

Background: Type 2 diabetes (T2D) is emerging as a growing global public health crisis. Early and effective interventions can reduce T2D incidence among at-risk populations. Compared with traditional approaches, digital health technologies offer promising opportunities for prevention, with eHealth literacy (eHL) emerging as a critical determinant of digital prevention outcomes.

Objective: This systematic review aims to synthesize and explain the pathways and mechanisms through which eHL supports T2D prevention among at-risk populations.

Methods: We searched Scopus, Web of Science, and PubMed databases for English-language original research published between January 1, 2000, and August 14, 2025. Studies included were prevention research involving eHL engagement among populations at risk for T2D. Nonoriginal literature, such as editorials and abstracts, as well as research protocols, was excluded. The findings were synthesized using a thematic analysis approach, integrating the Theoretical Domains Framework with the eHL model. Two reviewers independently screened literature and extracted data, and discrepancies were resolved by a third reviewer. The Mixed Methods Appraisal Tool was used to assess risk of bias.

Results: This review included 28 studies (n=13,100), mostly quantitative and published within the past decade, targeting people with prediabetes, prior gestational diabetes, and overweight/metabolic risk. Study quality was moderate to high (Mixed Methods Appraisal Tool 60% - 100%) with no high risk of bias. eHL supported prevention mainly through knowledge (28/28), behavioral regulation (16/28), social influences (15/28), environmental resources (12/28), and goals (11/28), while emotions, memory, attention, decision process, and beliefs about competence were rarely addressed. Health literacy (27/28), information literacy (20/28), and communicative eHL (20/28) were most common; critical eHL and media literacy were not addressed. Studies reported positive outcomes: high engagement, weight loss (≥5%), improved glycemic markers, and enhanced lifestyle behaviors.

Conclusions: This is the first systematic exploration of eHL mechanism pathways in T2D prevention via theoretical mapping. We found interventions yield positive effects despite highly uneven mechanism application: extant research relies excessively on knowledge and behavioral pathways while underemphasizing emotional support, autonomy, and critical evaluation—factors linked to long-term adherence. We provide a mechanism-based framework and identify critical gaps, including the absence of focus on critical eHL and media literacy. This review is limited by substantial variation across studies that did not allow for meta-analysis and by the limited evidence base on eHL. Future interventions should explore and test emotional and autonomy support, information discernment training, and accessibility optimization in T2D prevention. These comprehensive, equity-focused intervention approaches will help ensure that eHL becomes a truly effective public health tool that benefits everyone, especially at-risk and vulnerable populations.

Trial Registration: PROSPERO CRD42025630395; https://www.crd.york.ac.uk/PROSPERO/view/CRD42025630395

(J Med Internet Res 2026;28:e77788) doi:10.2196/77788

KEYWORDS

eHealth literacy; digital health literacy; type 2 diabetes prevention; type 2 diabetes mellitus; self-management; chronic disease self-care

XSL•FO
RenderX

Introduction

Background

Type 2 diabetes (T2D) is becoming a growing public health crisis worldwide [1], accounting for 90% to 95% of all diabetes cases [2]. The global number of adults with diabetes is expected to surge from 589 million in 2024 to 853 million in 2050 [2], placing a huge burden on health care systems worldwide.

The development of T2D is driven by a variety of factors, among which unhealthy dietary habits and a sedentary lifestyle are the main triggers [3,4]. In addition, at-risk factors include overweight, obesity, insulin resistance, metabolic syndrome, dyslipidemia, and hypertension [5-7], genetics [8,9], nonalcoholic fatty liver disease, history of gestational diabetes [10], chronic stress [11], depression [12], sleep deprivation [13], etc.

Long-term exposure to or persistent presence of the aforementioned risk factors may increase the risk of developing dysglycemia and progressing to prediabetes. Prediabetes is an independent predictor of progression to T2D, and without timely intervention, individuals with prediabetes face a significantly increased risk of developing diabetes. Therefore, populations exhibiting one or more of these characteristics constitute the “at-risk population” for T2D [14].

Focusing prevention efforts on this at-risk population holds significant public health implications. Approximately 5% - 10% of individuals with prediabetes progress to diabetes annually, with a lifetime risk reaching up to 70% [15]. Notably, prediabetes and related risk states are not irreversible. Research indicates that individuals can regain normal glucose regulation through intensive lifestyle interventions [16], offering a critical window for preventive action. Furthermore, targeting prevention efforts toward at-risk populations demonstrates favorable cost-effectiveness in economic evaluations [17]. Despite the significant benefits of preventive interventions, the landmark Diabetes Prevention Program (DPP) study proved that intensive lifestyle intervention can reduce the risk of T2D by 58% in at-risk populations [18]. However, traditional face-to-face DPP programs face severe challenges, such as insufficient geographical and resource distribution limiting accessibility [19], high promotion cost [20], and low completion rate [21], which make it difficult to achieve comprehensive coverage. These limitations highlight the urgent need for innovative prevention models.

Digital health technologies offer promising solutions to overcome traditional barriers to prevention [22]. Interventions based on digital platforms have shown potential to improve access and reduce implementation costs [23]. However, with the rapid growth of web-based health information, there are significant differences in information quality and availability [24]. Some information may even mislead users and cause anxiety, which in turn has a negative impact on health decision-making [25].

In this digital context, eHealth literacy (eHL) has become a key factor in determining the effectiveness of digital prevention. Norman and Skinner [26] defined eHL as "the ability to seek, discover, understand, and evaluate health information from electronic resources, and to apply the knowledge gained to address or solve health problems” and divided it into 6 dimensions. In 2018, Paige et al [27] proposed a conceptual model in which eHL is described as an interpersonal skill and divided into 4 dimensions. Recent conceptual updates emphasize “the ability to use digital technologies to achieve health goals in effective, safe and beneficial ways” [28], reflecting the evolving nature of eHL in digital health settings.

Studies have confirmed that eHL significantly affects diabetes prevention behaviors [29]. High eHL can promote a healthy diet, regular exercise, and good sleep [30], and improve the use of the medical system [31]. Lack of eHL may hinder disease prevention and health management [32].

Despite the increasing importance of eHL, there is still a lack of mechanistic understanding and theory-driven synthesis regarding how eHL influences the prevention of T2D in at-risk populations. On the one hand, existing systematic reviews have generally studied the relationship between eHL and health behavior [33,34]; on the other hand, they have only examined the impact of digital health intervention or telemedicine on prevention, lacking attention to literacy. Or focus on disease management in patients with established disease [35-40].

This study aims to address this research gap by conducting a mechanistic systematic review using theory-driven thematic analysis. The main objective was to study the effect path of eHL on the prevention of T2D in people at risk of T2D and the participation ways of different dimensions. The findings are expected to provide evidence-based guidance to health educators, intervention designers, and policy makers to develop T2D prevention strategies tailored to at-risk populations.

Research Questions

This mechanistic systematic review aimed to explore the pathways of eHL in the prevention of T2D among at-risk populations to fill the existing literature gap. Specifically, this study sought to answer the following research question (RQ):

- RQ1. What are the mechanistic actions of eHealth literacy in type 2 diabetes prevention among at-risk populations?
- RQ2. How do the dimensions of eHealth literacy facilitate type 2 diabetes prevention among at-risk populations?

Methods

Overview

The study protocol for this systematic review was registered with PROSPERO (registration number CRD42025630395 on January 16, 2025). This study was written and reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guideline [41] (Checklist 1) and the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search) [41] extension for reporting literature searches (Checklist 2).

Data Resources and Search Strategy

Scopus, Web of Science (WOS), and PubMed were selected as the databases for this review. The literature search was

XSL•FO
RenderX

conducted from January 1, 2000, to August 14, 2025. The year 2000 was chosen as the starting year because the concept of eHealth was formally proposed in 2001 [42]. Although the term eHL was proposed by Norman and Skinner in 2006 [26], earlier studies may have used related terms such as digital health literacy, health information literacy, etc, so a broad time frame was used to ensure the comprehensibility of the search.

In addition to database searching, we did not search trial or study registries; conduct any targeted searching or browsing of online or print resources; perform reference list screening or citation tracking; obtain additional studies or data by contacting authors, experts, or manufacturers; or use any other information sources or search methods. The search strategy was developed de novo, without adapting or reusing strategies from existing reviews, and no search update was performed (Multimedia Appendix 1). The strategy was peer-reviewed by an information scientist with extensive expertise in systematic review methods and database search strategy design.

Search terms were generated around the research topic. Two researchers (WY and RG) independently screened and identified relevant studies using retrieval strings based on the Population, Interest/Intervention, and Context model combined with Boolean logic. Disagreements were resolved by discussion and, if necessary, consensus was reached by consulting other investigators (AAA and ZZH). All keywords were searched in the database. A comprehensive list of search queries is presented in Table 1. Individual search strings for each database are shown in Checklist 2.

Table . Keywords search strings.

Concept	Keywords/search terms
eHealth	“eHealth” OR “e-health” OR “mHealth” OR “m-health” OR “mobile health” OR “digital health” OR “internet health” OR “telehealth” OR “telemedic*” OR “computer health” OR “computer-based health” OR “web health” OR “web-based health” OR “web based health” OR “online health” OR “online-based health” OR “online based health” OR “health” OR “media” OR “computer” OR “virtual” OR “intelligent” OR “technolog*” OR “Web 2.0” OR “web based” OR “online based” OR “phone*” OR “app*”
Literacy	“literac*” OR “comprehension” OR “skill*” OR “ability” OR “knowledge” OR “seek*“OR “find*” OR “search*” OR “understand*” OR “apprais*” OR “evaluat*” OR “assess*” OR “access*” OR “apply” OR “use*” OR “communication” OR “interaction” OR “efficacy” OR “engagement”
Type 2 diabetes prevention	“diabetes prevent*” OR “diabetes prevention” OR “type 2 diabetes prevent*” OR “type II diabetes prevent*” OR “T2D prevent*” OR “T2DM prevent*” OR “prediabetes prevent*” OR “non-insulin-dependent diabetes mellitus prevent*” OR “adult-onset diabetes prevent*” OR “NIDDM prevent*”
At-risk population	“at-risk” OR “high-risk” OR “risk population” OR “prediabetes” OR “overweight” OR “fat” OR “obesity” OR “sedentary behavi*” OR “sedentary lifestyle*” OR “physical inactiv*” OR “lack of exercise” OR “genetics” OR “family history” OR “gestational diabetes” OR “hypertens*” OR “dyslipid*” OR “insulin resistance” OR “acanthosis nigricans” OR “non-alcoholic fatty liver disease” OR “NAFLD” OR “metabolic syndrome” OR “chronic stress” OR “depression” OR “poor sleep” OR “western diet” OR “unhealthy diet” OR “suboptimal diet”

Selection Criteria

Papers included in the review had to meet prespecified inclusion criteria. First, selected articles should be original research articles written in English and published between January 1, 2000, and August 14, 2025. Other literature types, such as reviews, editorials, conference papers, letters, briefings, reviews, book chapters, monographs, or research protocols, were excluded from this literature review. Second, people at risk for T2D. Patients included those with prediabetes, family history of diabetes, overweight or obesity, history of gestational diabetes, sedentary or lack of exercise, hypertension, dyslipidemia, insulin resistance, acanthosis nigricans, nonalcoholic fatty liver disease, metabolic syndrome, chronic stress, depression, inadequate sleep, or unhealthy diet. In addition, studies were excluded if they were (1) not specifically targeting T2D, but referring to diabetes or type 1 diabetes; (2) conducted via nonelectronic means such as text messages and phone calls; (3) only describing barriers or benefits of online prevention without explicitly mentioning prevention measures related to eHL; and (4) irrelevant to T2D prevention. Detailed inclusion and exclusion criteria are provided in Table 2.

XSL•FO
RenderX

Table . Inclusion and exclusion criteria.

Components	Inclusion	Exclusion
Population	Individuals at risk for type 2 diabetes, including those with prediabetes: a family history of diabetes; overweight or obesity; a history of gestational diabetes; a sedentary lifestyle or physical inactivity; hypertension; dyslipidemia; insulin resistance; acanthosis nigricans; nonalcoholic fatty liver disease; metabolic syndrome; chronic stress; depression; insufficient sleep; or an unhealthy diet.	• Refers to diabetes without specifying people with type 2 diabetes, people with type 1 diabetes, and people without risk of type 2 diabetes.
Interest	Studies explicitly involving eHealth literacy in type 2 diabetes prevention among at-risk populations, either as an overall construct or through specific dimensions, using digital or online approaches.	• Not specifically targeting type 2 diabetes, but referring to diabetes or type 1 diabetes. • Nonelectronic means such as text messages and phone calls.
Context	Studies must report the impact of eHealth literacy on the prevention of type 2 diabetes among at-risk populations.	• Only describe barriers or benefits of online prevention, without explicitly mentioning prevention related to eHealth literacy. • Not associated with type 2 diabetes prevention.
Timeline	January 1, 2000, to August 14, 2025	• Before 2000.
Article type	Primary research journal articles on quantitative research, qualitative research, and mixed methods research	• Reviews, editorials, conference papers, letters, notes, commentaries, book chapters, books, or study protocols.
Language	English	• Non-English.

Systematic Review Process

All papers retrieved from WOS, Scopus, and PubMed databases were imported into Rayyan Citation Manager [43], and duplicates were removed. Two authors (WY and RG) independently screened studies according to prespecified inclusion and exclusion criteria. Studies that may meet the criteria will be screened for full text for a final decision on inclusion in our review. Any inconsistencies were resolved through discussion and consensus among the reviewers. If consensus was not reached, a third reviewer (AAA) made the final decision.

Risk of Bias Assessment and Reporting Bias Assessment

The risk of bias of all included studies was independently assessed by 2 reviewers (WY and RG) and cross-validated by each other. Any disagreements were resolved by discussion with a third evaluator (AAA). The Mixed Methods Appraisal Tool (MMAT, version 2018) was used to assess the quality of the eligible studies. This tool is primarily used to assess the quality of mixed methods studies and also includes the assessment of qualitative and quantitative methods. MMAT included 2 screening questions for each study and then 5 questions for each design, depending on the category of study design. Each question was rated as “no,” “yes,” and “unable to judge.” The risk of bias was determined for each study based on the percentage of questions answered “yes,” with scores ranging from 0% to 100% [44]. Prior to conducting the assessment, reviewers completed the official MMAT online training tutorial developed by Pace et al [45] to ensure consistent and informed use of the tool.

Theoretical Underpinnings

This systematic review uses the Theoretical Domains Framework (TDF) and the eHL model as its theoretical foundation for constructing a coding framework, organizing and interpreting evidence, and identifying behavioral mechanisms and literacy dimensions that are adequately covered versus relatively underrepresented in current intervention designs during the thematic synthesis process.

The TDF encompasses 14 domains, including knowledge, skills, social or professional roles and identities, beliefs about competence, optimism, reinforcement, intentions, goals, memory, attention and decision processes, environmental background and resources, social influences, and emotions and behavioral regulation [46]. This framework provides a systematic and actionable theoretical classification to elucidate the potential mechanisms through which eHL may influence diabetes prevention-related behaviors, further pointing to intervention targets for optimization.

To comprehensively describe the dimensional coverage and mechanistic actions of eHL in the included studies, this research uses 2 complementary eHL models to balance breadth and depth. At the breadth level, the Norman and Skinner framework conceptualizes eHL as 6 foundational literacies (traditional literacy, health literacy, information literacy, scientific literacy, computer literacy, and media literacy) [26], facilitating cross-comparisons of eHL dimensional coverage across studies.

XSL•FO
RenderX

At the depth level, the interactive framework of Paige et al [27] refines eHL into functional eHL, communicative eHL, critical eHL, and translational eHL, emphasizing interaction, comprehension, and application transfer within specific digital health contexts. The integration of these models enables this review to provide a more systematic interpretation and comparative analysis of eHL's multidimensional mechanisms within the context of T2D prevention.

Data Extraction and Analysis

To gain a comprehensive understanding of the essential characteristics and trends of the included studies, we developed a set of standardized data extraction tables for summarizing key study characteristics. The table collected detailed information from the included literature, including author, year of publication, country/region, age of participants, study methods, and main intervention outcomes.

In addition, we use thematic analysis as the main method to identify patterns or themes in qualitative data [47]. It is widely regarded as one of the most appropriate methods for integrating mixed study designs [48]. Thematic analysis is a descriptive strategy that enables flexible reduction of data and can be used in combination with other data analysis techniques [49].

Therefore, we conducted a thematic analysis to identify common themes that emerged in the data. These themes were derived from the objectives of this review.

First, addressing RQ 1, two reviewers (WY and RG) classified relevant findings by mapping intervention components and reported implementation strategies to the 14 domains of the TDF [46], to summarize which behavior change mechanisms were reflected across studies.

Second, addressing RQ 2, we examined how different dimensions of eHL contribute to the prevention process by mapping eHL-related content from included studies onto 2 complementary eHL models [26,27].

In addition, we used the same thematic analysis method to generalize the intervention outcomes reported by the studies into weight change, diabetes-related measures, behavior change, self-efficacy, and engagement and acceptability.

For studies independently reviewed by 2 researchers (WY and RG), if consensus cannot be reached, the final decision will be made by a third reviewer (AAA).

Results

Study Screen

Overview

In accordance with PRISMA 2020 reporting guidelines, we conducted a comprehensive search on WOS, Scopus, and PubMed to find 5409 articles published in English. During the initial title and abstract screening process, we identified and removed 2593 duplicate records, resulting in 2816 articles progressing to the next stage.

After further review of titles and abstracts, 2702 articles were excluded due to lack of relevance to eHL or T2D prevention in at-risk populations, leaving 114 articles for full-text screening. Of these, 86 articles were subsequently excluded. Ultimately, 28 studies were included. The PRISMA screening flow chart is shown in Figure 1 for details.

XSL•FO
RenderX

Figure 1. Flowchart of systematic review search results. eHL: eHealth literacy.

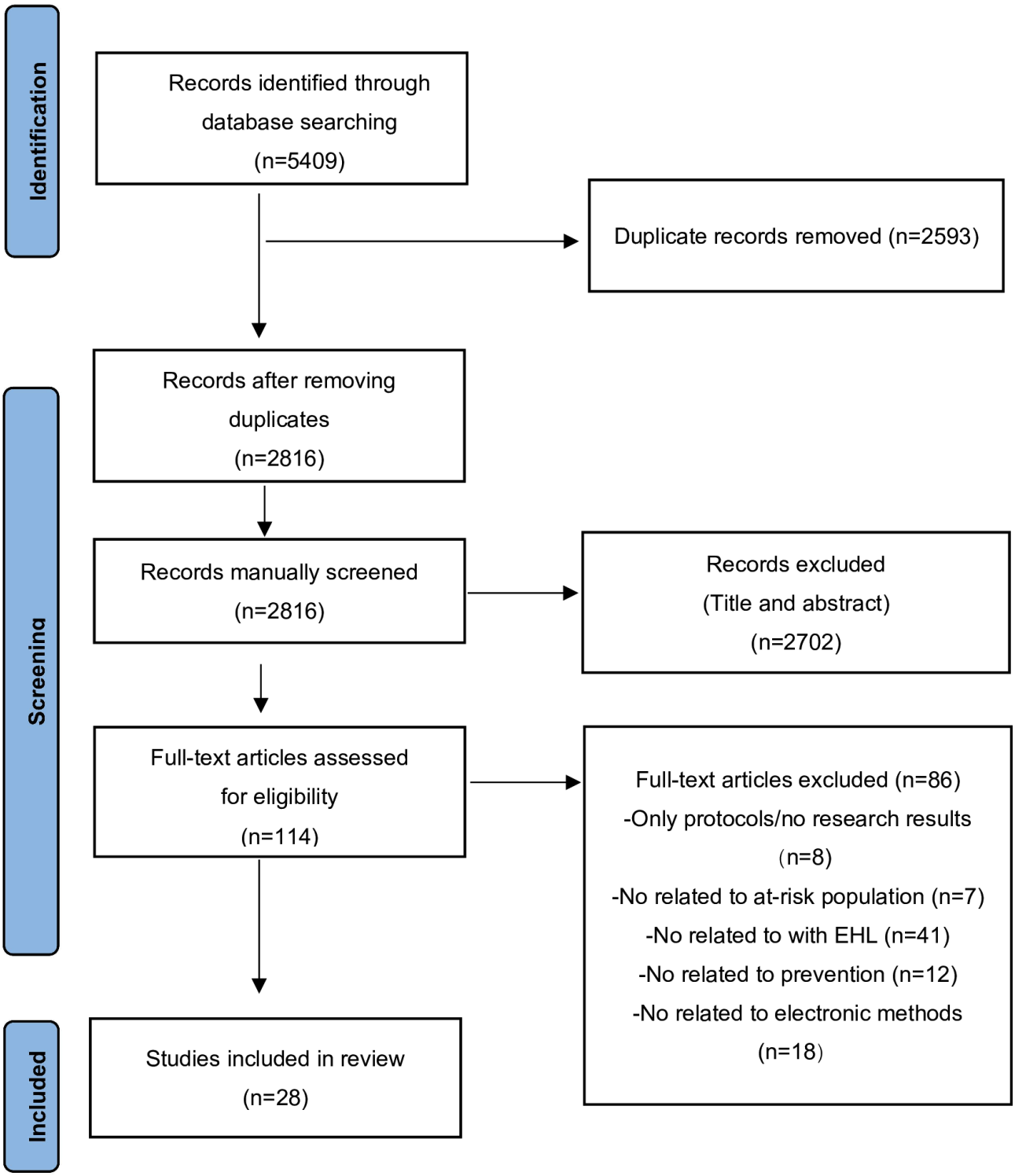

Description of Studies

This review includes 28 studies that met the inclusion criteria, with their characteristics detailed in Table 3.

The majority of studies were conducted within the last decade (27/28), with a particular focus on 2019, 2022, and 2023 [50-62]. One such study was conducted in 2010 [63].

A total of 86% (24/28) of the studies were from the United States. In addition, there was one study from Canada (n=1), one study from Malaysia (n=1), one study from Taiwan, China (n=1), and one study from Singapore (n=1).

In terms of population at-risk, 23 studies focused on people with prediabetes [50-72], 3 studies focused on women with a history of gestational diabetes [59,63,73], and 10 studies examined individuals who were overweight, obese, had excessive BMI, or had metabolic syndrome [53,56,58,59,67-69,71,74,75]. In addition, one study included individuals who had previously given birth to a baby weighing

XSL•FO
RenderX

more than 9 lb (about 3.7 kg) and had hypertension or dyslipidemia [63].

Seven studies [51,53,54,58,60,64,76] specifically addressed special populations, including female veterans [64], low-income individuals enrolled in Medicaid [51], self-identified as Asian [53], adolescents [54], older at-risk individuals [58,76], and Chinese Americans [60].

A total of 18 studies used web-based DPP interventions [51,53,55,57-60,62-64,66,67,69,70,72,74,76,77], and one study specifically recruited a population not previously enrolled in a DPP program [52].

A total of 28 intervention studies were ultimately included, encompassing 13,100 participants. Intervention durations ranged from 12 weeks to 3 years, with participant ages spanning adolescents to older adults (approximately 14.6 to 69.2 y).

XSL•FO
RenderX

Table . Characteristics and key findings of included studies on type 2 diabetes prevention among at-risk populations.

Study	Country	Study design	Sample size	Characteristics of at-risk populations	Average age (years)	Gender	Intervention delivery modality	Time frame	Key findings
Vadheim et al (2010) [63]	United States	Quantitative nonrandomized	29	A history of prediabetes, gestational diabetes mellitus (GDM), or delivery of a baby weighing more than 9 pounds, hypertension, or dyslipidemia	51.4	81.4% (22/27) women	Remote videoconference	16 weeks	A total of 88% (14/16) of participants completed the 16-week program; More than 40% (12/29) of participants achieved the 7% weight loss goal.
Moin et al (2015) [64]	United States	Qualitative study	15	Female veterans with prediabetes	56.8	100% (15/15) women	Online virtual platform, pedometer, tracker, wireless scale	16 weeks	Perception: the plan was well aligned with perceived health needs; The program is convenient; Integration of projects into daily life; A sense of responsibility; Hate records; Prefer traditional offline; Difficult to understand; Mean weight loss was 5.24%.
Sepah et al (2015) [74]	United States	Quantitative nonrandomized	220	BMI>24 kg/m^2 over 18 years of age	43.6	17.3% (38/220) men	Online web platform, digital tracking tools (wireless scale and pedometer tracking)	2 years	Users had significant reductions in body weight and HbA_{1c}[a] that were sustained 2 years later.

XSL•FO
RenderX

Study	Country	Study design	Sample size	Characteristics of at-risk populations	Average age (years)	Gender	Intervention delivery modality	Time frame	Key findings
Block et al (2015) [65]	United States	RCT[b]	339	Prediabetes	55	68.7% (233/339) men	Alive-PD app, personalized website, and interactive email delivery	6 months	Reductions in HbA_{1c}, fasting glucose, and body weight were significantly better in the intervention group than in the control group. Health behaviors and lifestyle improved significantly, and program participation was high.
Sepah et al (2017) [66]	United States	Quantitative nonrandomized	220	Prediabetes	__[c]	82.7% (182/220) women	An Internet-connected desktop computer or mobile device	3 years	There was a significant decrease in body weight and glycated hemoglobin at week 16, year 1, and year 2. At the third year, there was a significant reduction in body weight and HbA_{1c}.
Stein and Brooks (2017) [75]	United States	Quantitative nonrandomized	70	Overweight and obese adults	47	74.5% (35/47) women	Artificial Intelligence Health Coaching app (HCAI app)	15 weeks	A total of 75.7% (53/70) of the users lost weight in this plan, and the percentage of healthy diets increased by 31%. The response rate of the user trust survey is 100%.

XSL•FO
RenderX

Study	Country	Study design	Sample size	Characteristics of at-risk populations	Average age (years)	Gender	Intervention delivery modality	Time frame	Key findings
Michaelides et al (2018) [67]	United States	Quantitative nonrandomized	59	Overweight or obese adults with prediabetes	51	81% (48/59) women	The Noom application	65 weeks	Weight and BMI were significantly reduced, with 53% (25/47) of the completion group and 66% (21/32) of the maintenance group achieving ≥5% weight loss.
Everett et al (2018) [68]	United States	Quantitative nonrandomized	55	Adults with prediabetes and BMI between 24 and 40 kg/m^2	55	60% (33/55) women	Noom application, Sweetch application, and Digital Weight Scales (DBWS)	3 months	The retention rate and acceptance of the study were relatively high. Participants lost 1.6 kilograms in weight, their BMI decreased by 0.6 $kilograms/m^2$, their waist circumference reduced by 1.4 centimeters, and their physical activity significantly increased. Fasting blood glucose showed no significant change compared with the baseline.

XSL•FO
RenderX

Study	Country	Study design	Sample size	Characteristics of at-risk populations	Average age (years)	Gender	Intervention delivery modality	Time frame	Key findings
Castro Sweet et al (2018) [69]	United States	Quantitative nonrandomized	501	Have prediabetes or metabolic syndrome	68.8	64% women[d]	Connected devices, wireless scales	12 months	The participants lost an average of 6.5% in weight over 16 weeks. At the 6th month, the weight decreased by 8.0%, and at the 12th month, it decreased by 7.5%. The participants' blood sugar control was improved and their total cholesterol was reduced. Participants' self-reported improvements in well-being, depression, and self-care.
Srivastava et al (2019) [50]	United States	Quantitative nonrandomized	10	Prediabetes	56.1	70% (7/10) women	Cloud-based type 2 diabetes prevention module	6 months	The success rate of the module is approximately 60%. The average weight loss of users is approximately 9.0%, and their weekly physical activity time is 192 minutes.
Kim et al (2019) [51]	United States	Quantitative nonrandomized	227	Low-income people with prediabetes enrolled in Medicaid	48.2	81.3% women[d]	Online digital platforms	12 months	The intervention group had an average weight loss of 4.4% and an average BMI change of –1.6 over 12 months.

XSL•FO
RenderX

Study	Country	Study design	Sample size	Characteristics of at-risk populations	Average age (years)	Gender	Intervention delivery modality	Time frame	Key findings
Griauzde et al (2019) [52]	United States	Mixed methods	69	People with prediabetes who did not participate in DPP[e]	51.7	64% (44/69) women	Apps, Fitbit devices	12 weeks	8 out of 13 respondents liked some features of the app, and 13 out of 13 respondents reported changes in health behaviors.
Alwashmi et al (2019) [53]	United States	Quantitative nonrandomized	273	Prediabetes, BMI>25 kg/m^2, or self-identified as Asian	54	70.3% (192/273) women	Interactive mobile computing (eg, smartphone applications), wearable tracking devices (eg, activity trackers), remote health monitoring hardware (eg, electronic scales)	4 months	The average weight loss was 6.5%, the average BMI decreased by 1.9 kg/m^2, the frequency of exercise increased by 1.7 days per week, and the monthly absenteeism rate decreased by nearly half a day.
Toro-Ramos et al (2020) [70]	United States	RCT	202	Prediabetes	56.6	71.3% (144/202) women	The Noom application	12 months	The changes in body weight and BMI in the intervention group were significantly reduced, but the effect on HbA_{1c} was not significant in the intention-to-treat analysis.

XSL•FO
RenderX

Study	Country	Study design	Sample size	Characteristics of at-risk populations	Average age (years)	Gender	Intervention delivery modality	Time frame	Key findings
Ferrara et al (2020) [77]	United States	RCT	158	People with a risk of more than 9 points were screened by the tool	52.2	72.8% (115/158) women	Canary health online platform	12 months	The online group achieved significant weight loss within 6 months, with an average weight change of –1.97 kilograms. 12 months –1.27 kilograms; At 6 months and 12 months, the average total calorie intake, dietary fat intake, and saturated fat intake of the online group all decreased significantly. However, it did not lead to significant changes in physical activity.
Ekezie et al (2021) [73]	United States	Qualitative study	23	Women with previous Gestational Diabetes Melilitus	—	100% (23/23) women	Baby steps app	11 months	Changes in lifestyle; Increasing physical activity and steps; Promoting competition and communication; Increase motivation; Bring in peer support.
Islam et al (2022) [54]	United States	Qualitative study	14	Adolescents with prediabetes	14.6	36% (5/14) women	Zoom	12 weeks	Participants indicated that the online adolescent health education program was beneficial and applicable to their lives, and gave them a deeper understanding of healthy eating and active lifestyle.

XSL•FO
RenderX

Study	Country	Study design	Sample size	Characteristics of at-risk populations	Average age (years)	Gender	Intervention delivery modality	Time frame	Key findings
Katula et al (2022) [55]	United States	RCT	599	Prediabetes	55.4	61.4% (368/599) women	Internet-connected device (laptop, tablet, or smartphone)	12 months	HbA_{1c}, body weight, and cholesterol/high-HDL[f] ratio decreased significantly. Reductions in HbA_{1c} and body weight were greater in the older participants aged ≥65 years.
Lim et al (2022) [56]	Singapore	RCT	148	Adults with prediabetes and BMI ≥23 kg/m^2	53.1	39.9% (59/148) women	nBuddy Diabetes app	6 months	The intervention group was 2.1 times more likely to achieve euglycemia (HbA_{1c}<5.7%) than the control group. Within-group improvements in systolic and diastolic blood pressure, total cholesterol, and low-density lipoprotein (LDL) cholesterol were significant in the intervention group. The intervention group had significantly greater reductions in daily intake of total calories, carbohydrates, total fat, saturated fat, and sugar than the control group.

XSL•FO
RenderX

Study	Country	Study design	Sample size	Characteristics of at-risk populations	Average age (years)	Gender	Intervention delivery modality	Time frame	Key findings
Batten et al (2022) [57]	Canada	Quantitative nonrandomized	1095	Prediabetes	53.6	67.7% (741/1095) women	Smartphone applications, wireless scales, wearable tracking devices	12 months	After at least 9 months of participation, the average weight loss was 11.4 pounds, with a rate of 5.5%. Mean weekly physical activity increased to 132.9 minutes.
Fitzpatrick et al (2022) [58]	United States	Quantitative nonrandomized	3904	65-75 years with prediabetes and obesity	69.2	55.2% (2155/3904) women	Electronic patient portal, wireless scale, and pedometer	24 months	The average weight loss was 8.6 pounds, and participation was high.
Collins et al (2023) [59]	United States	Quantitative nonrandomized	2390	Adults with BMI >25 and a history of gestational diabetes or prediabetes	54.9	89.2% (2132/2389) women	Video conferencing, secure online portals	12 months	HbA_{1c} monitoring decreased significantly, and those who completed the program lost an average weight of 5.66%. Participants showed statistically significant changes in their self-reported confidence in their ability to perform all 18 health promotion behaviors.

XSL•FO
RenderX

Study	Country	Study design	Sample size	Characteristics of at-risk populations	Average age (years)	Gender	Intervention delivery modality	Time frame	Key findings
Yeh et al (2023) [60]	United States	Mixed methods	13	Chinese Americans with prediabetes	66	85% (11/13) women	Facebook groups, Qualtrics links	12 months	The completion rate was high. The completion rate of the maintenance phase decreased slightly; the weekly modules were generally considered to be clear and easy to understand. All of the participants (100%) liked the online DPP, and the overall satisfaction of the participants was high. Body weight decreased significantly at 2, 3, 4, 8, and 10 months.

XSL•FO
RenderX

Study	Country	Study design	Sample size	Characteristics of at-risk populations	Average age (years)	Gender	Intervention delivery modality	Time frame	Key findings
Zahedani et al (2023) [61]	United States	Quantitative nonrandomized	2217	Prediabetes	49	51% women[d]	January AI app, wearing CGM (Freestyle Libre, Abbott), heart rate monitor (Apple Watch or Fitbit)	About 3 months	The GMI[g] was significantly reduced, and the number of prediabetic hyperglycemia events was the largest reduction. Prediabetic individuals lost an average of 2.5 pounds, roughly doubled their physical activity, and increased their heart rate by 6.3 minutes/day; Calorie, carbohydrate, sugar, and saturated fat intakes all decreased, whereas protein, total fat, and dietary fiber intakes increased.
Chung et al (2023) [62]	Taiwan, China	RCT	121	Prediabetes	58.1	52.9% (64/121) women	TCM mHealth app	About 4 months	The TCM mobile health app intervention was superior to the control group in blood glucose control (HbA_{1c} reduction), body constitution (Yang deficiency, phlegm stasis improvement), body energy improvement, health-related quality of life (PCS, MCS), and BMI reduction.

XSL•FO
RenderX

Study	Country	Study design	Sample size	Characteristics of at-risk populations	Average age (years)	Gender	Intervention delivery modality	Time frame	Key findings
Gerber et al (2024) [76]	United States	RCT	20	Older people with excessive BMI	65.4	45% (9/20) women	Zoom, Fitbit Inspire 2 /Fitbit Aria Air, Fitbit App	6 months	Meeting participation was high, with 74% calorie intake compliance; Body weight and BMI changed significantly. The intervention group achieved an 8.4% reduction in body weight at 6 months while BMI decreased by 2.3 units at 3 months.
Ng et al (2025) [71]	Malaysia	RCT	82	Overweight/obese prediabetic patients	49.8	53.7% (44/82) women	Blood glucose meters and wearable fitness trackers, PRIME app, WhatsApp	6 months	The dietary quality score, vegetable score, and sodium score were significantly improved in the intervention group. The intervention group showed beneficial changes in refined grains, vegetable consumption, and sodium intake.
Audet et al (2025) [72]	United States	Qualitative study	27	Prediabetes	56	89% (24/27) women	HP-DPP program (Zoom+REDCap app)	12 months	The participants' knowledge, motivation, and skills were improved.

[a]HbA_{1c}: glycosylated hemoglobin A_{1c}.

[b]RCT: randomized controlled trial.

[c]Not available.

[d]N values are not available.

[e]DPP: Diabetes Prevention Program.

[f]HDL: density lipoprotein cholesterol.

[g]GMI: glucose management index.

Most studies were quantitative (n=22), including randomized controlled trials (n=8) [55,56,62,65,70,71,76,77] and nonrandomized quantitative studies (n=14). These include retrospective studies [50,53,57], interventional studies [51,61], observational studies [58,59,67-69,75], experimental studies [66,74], and pilot studies [63].

In addition to this, 2 studies used a mixed methods design [52,60] and another 4 studies were qualitative [54,64,72,73].

XSL•FO
RenderX

Among the 28 included studies, the delivery modalities used to deliver T2D prevention interventions were categorized into 4 groups. The first 10 involved apps for diabetes prevention [53,56,61,62,65,67,70,72,73,75] and the second 13 involved the use of web-based platforms or websites [50,51,53,55,58-60,64-66,69,74,77]. In addition, 11 trials used wearable devices or digital tracking tools such as a wireless scale [52,53,57,58,61,64,68,69,72,74,76]. Six studies used video conferencing or social software such as Zoom and Facebook [54,59,60,63,72,76].

Bias Reporting of the Eligible Studies

The MMAT was used to assess the risk of bias across the 28 included studies, with none identified as having a high risk of bias. The overall methodological quality was rated as moderate to high, with MMAT scores ranging from 60% to 100% across the studies (Table 4).

XSL•FO
RenderX

Table . Quality assessment.

Study	Research design	Q1	Q2	Q3	Q4	Q5	Quality	Inclusion in the review
Vadheim et al (2010) [63]	QN (NR)[a]	1	1	1	0	1	80%	✓
Moin et al (2015) [64]	QA[b]	1	1	1	1	1	100%	✓
Sepah et al (2015) [74]	QN (NR)	1	1	0	0	1	60%	✓
Block et al (2015) [65]	QN (RC)[c]	1	1	1	1	1	100%	✓
Sepah et al (2017) [66]	QN (NR)	1	1	1	0	1	80%	✓
Stein and Brooks (2017) [75]	QN (NR)	1	1	1	0	1	80%	✓
Michaelides et al (2018) [67]	QN (NR)	1	1	1	0	1	80%	✓
Everett et al (2018) [68]	QN (NR)	1	1	1	0	1	80%	✓
Castro Sweet et al (2018) [69]	QN (NR)	1	1	1	0	1	80%	✓
Srivastava et al (2019) [50]	QN (NR)	1	1	0	0	1	60%	✓
Kim et al (2019) [51]	QN (NR)	1	1	1	1	1	100%	✓
Griauzde et al (2019) [52]	MX	1	1	1	0	1	80%	✓
Alwashmi et al (2019) [53]	QN (NR)	1	1	1	0	1	80%	✓
Toro-Ramos et al (2020) [70]	QN (RC)	1	1	1	1	1	100%	✓
Ferrara et al (2020) [77]	QN (RC)	1	1	0	1	1	80%	✓
Ekezie et al (2021) [73]	QA	1	1	1	1	1	100%	✓
Islam et al (2022) [54]	QA	1	1	1	1	1	100%	✓
Katula et al (2022) [55]	QN (RC)	0	1	1	1	1	80%	✓
Lim et al (2022) [56]	QN (RC)	1	1	1	1	1	100%	✓
Batten et al (2022) [57]	QN (NR)	1	1	1	0	1	80%	✓
Fitzpatrick et al (2022) [58]	QN (NR)	1	1	1	1	1	100%	✓
Collins et al (2023) [59]	QN (NR)	1	1	0	1	1	80%	✓
Yeh et al (2023) [60]	MX[d]	1	1	1	0	1	80%	✓
Zahedani et al (2023) [61]	QN (NR)	1	1	1	0	1	80%	✓

XSL•FO
RenderX

Study	Research design	Q1	Q2	Q3	Q4	Q5	Quality	Inclusion in the review
Chung et al (2023) [62]	QN (RC)	1	1	1	1	1	100%	✓
Gerber et al (2024) [76]	QN (RC)	1	1	1	1	1	100%	✓
Ng et al (2025) [71]	QN (RC)	0	1	1	0	1	60%	✓
Audet et al (2025) [72]	QA	1	1	1	1	1	100%	✓

[a]QN (NR): Quantitative nonrandomized.

[b]QA: Qualitative assessment.

[c]QN (RC): Quantitative randomized controlled trials.

[d]MX: Mixed Method.

The Mechanism of Action of eHL

Knowledge

Figure 2 illustrates the pathways and mechanisms through which eHL supports T2D prevention among at-risk populations.

Figure 2. Mechanistic pathways of eHL in type 2 diabetes prevention among at-risk populations [26,27]. eHL: eHealth literacy.

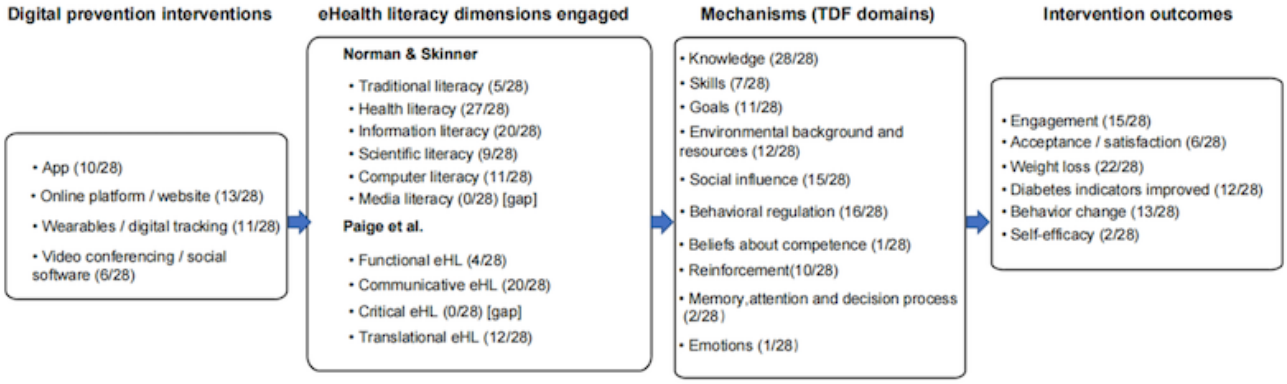

Overview

All 28 included studies (100%) addressed the mechanistic actions of health literacy in the prevention of T2D. The content covered disease cognition, nutrition and diet, physical activity, and standardized DPP courses. These interventions impact diabetes prevention through 3 pathways as follows.

Information Acquisition and Understanding—Disease Knowledge (n=3)

Such as prediabetes and strategies to prevent progression of T2D [52], understanding prediabetes [62], and the pathophysiology and complications of prediabetes [71], the acquisition of disease knowledge builds the knowledge base for prevention.

Behavior Guidance—Diet and Exercise Knowledge (n=8)

Nutrition and diet knowledge was the most common content, Content covering topics such as dietary sources of saturated fats and sugars, portion control [65], weight management, diabetes prevention, healthy meal planning, the relationship between carbohydrates and glycemic response [56], balanced meals following the US Department of Agriculture's My Plate program [57,58], diet [60], reducing foods that raise blood glucose, restricting calories, and increasing fiber intake [61], the Dietary Approaches to Stop Hypertension diet [62], practical nutrition advice for people with prediabetes, including energy balance, portion control, healthy eating based on local foods, glycemic index, fiber intake, variety of fat and protein foods, alcohol intake, etc [71].

Physical activity knowledge (n=6) included teaching classes related to physical activity, such as aerobic exercise, strength training, yoga, dance, pilates, water aerobics, etc [63], physical exercise guidelines [65], educational videos on physical activity [56], benefits and recommendations for increasing activity [61,62] of physical activity [71]. The guidance of diet and exercise knowledge promotes prevention from information to practice.

Digital Intervention Program—Systematizing Knowledge (n=18)

A total of 18 studies [51,53,55,57-60,62-64,66,67,69,70,72,74,76,77] were based on the DPP or its digital adaptations, including 2 from the Prevent project [64,74] and 5 from the Omada project [55,57,58,66,69]. And 11 derived from DPP and its derivative programs

XSL•FO
RenderX

[51,53,59,60,62,63,67,70,72,76,77], whose curricula deliver knowledge through integrated multimodal digital channels.

Skills

Overview

Seven studies [52,54,56,62,67,70,77] reported that eHL promotes prevention through skill building.

Digital Tool Operation Skills (n=6)

eHL is reflected in the ability of participants to learn and master digital intervention tools. Six studies provided training in tool use [52,56,62,67,70,77]. These included learning how to operate the Noom App and interacting with a web-based coach [67,70], receiving instructions for using the App and Fitbit devices [52], learning how to use the online platform through face-to-face group coaching, goal setting, and behavioral techniques for self-monitoring diet, physical activity, and weight using the online platform [77], and downloading and familiarizing yourself with mobile health applications [56,62].

Health Behavior–Related Skills (n=1)

One study [54] reported participants learning practical health management skills through a digital platform, including reading nutrition labels, making plate plans, and controlling portion sizes.

Goals

Overview

A total of 11 studies [54,56-58,62,65,68-71,73,75] provide the function of goal setting in diabetes prevention. There are 3 main patterns of goal setting in research.

Algorithm-Driven Personalized Target Recommendation (n=3)

Three studies used algorithms to generate personalized goals for participants [62,65,68]. In a high-quality study, the Alive-PD system recommended weekly that participants choose one eating habit goal and one physical activity goal from among the suggested goals. In addition to long-term goal setting, individually relevant small-step goals were recommended weekly according to the participant. For example, participants who consumed more than the recommended level of added sugars were provided with weekly targets based on their reported sources of added sugars. The target-recommendation algorithm adapts to the participant's previous performance. They were also prompted on a weekly basis to report on the completion of goals in the previous week. The system tracks the number and type of goals that respondents successfully achieve [65]. In another study, Sweetch used advanced algorithms to provide personalized, contextualized, immediate, and timely advice to guide users toward recommended activity, weight loss, and dietary goals in a way that aligns with the user's actual lifestyle. In another high-quality study, the Sweetch technique was able to learn which types of messages led to better compliance for specific users in specific contexts. Breaking down the weekly 150-minute physical training goal into small pieces throughout the day makes the goal easier to achieve. These snippets will continue to be personalized for each user. Personalized goals are constantly adjusted according to the actual behavior of the user to best fit the actual life capabilities of the user [68]. Another study set up a "milestone" function so that participants could review their goals and thus make adjustments to meet their monthly goals. For example, if a participant sets a fasting glucose target of 60 to 99 mg/dL, a pop-up window displays achievement when the target is reached. Participants can also view weekly or month-long bar and line charts to compare actual versus expected values at different time points and actual versus ideal behavior in order to adjust desired goals as needed [62].

Goal Setting Assisted by a Web-Based Coach (n=4)

Four studies [57,58,70,71] involved health coaches assisting participants to help users set specific, measurable, achievable, and realistic goals. Alternatively, in one study, the intervention group was instructed to set specific, measurable, realistic, achievable, and timely goals by selecting small but meaningful changes to fit into their daily lives, which were constantly reviewed and revised during subsequent follow-up [71].

Autonomous Goal Tracking and Self-Regulation (n=5)

In addition, goal setting can be done through personal records and individual settings [54,56,69,73,75]. In one study, users were guided through goal-setting modules on weight loss and food choice [75]. In another study, individuals were able to set step goals within an app [56].

Environmental Background and Resources

Overview

Twelve studies created empowering prevention resources for participants by providing access to digital tracking tools and wearable devices [52,53,55,56,61,65,66,69,71,73,74,76]. These devices fall into 3 main categories as follows.

Weight Monitoring Equipment (n=7)

Seven studies equipped participants with wireless or Internet-connected scales to automatically track weight changes and transfer data to a database or personal dashboard [52,55,66,69,71,74,76].

Motion Monitoring Equipment (n=11)

Exercise monitoring equipment was the most widely used form of resources, with a total of 11 items. Including a pedometer, wrist wear activity tracker, and other intelligent wearable devices [52,53,55,56,65,69,71,73,74,76]. One study [55] allowed participants to use their own wearable devices (eg, Fitbit and Apple Watch).

Other Health Monitoring Tools (n=2)

In addition to weight and exercise tracking, 2 studies combined more integrated digital health tools such as continuous glucose monitors, which track blood glucose fluctuations in real time, and heart rate monitors [61]. Alternatively, in the study by Ng et al [71], a glucose meter was provided for self-blood glucose monitoring, and data on the glucose meter and tracker were gradually tracked and synchronized into the application.

XSL•FO
RenderX

Social Influence

A total of 15 studies identified social influence as a key mechanistic pathway through which eHL facilitated diabetes prevention behaviors [51,53,55,57-59,62,64,66,67,69,71,72,74,76]. The forms can be divided into 2 types. One group used closed digital communities, with 12 studies interacting through web-based groups, private social forums, private chat groups, third-party messaging tools, where participants posted updates, responded to comments, shared lifestyle change experiences, and received feedback from wellness coaches or other members to get online peer support [51,53,55,57,58,62,64,66,67,69,71,74]. For example, in one study, the intervention group was invited to join a peer support chat group hosted by WhatsApp. They could ask questions related to prediabetes and exchange experiences on lifestyle modification, including best practices, challenges faced, and ways to overcome, throughout the study period [71]. In addition to digital communities, 3 studies used structured online discussion activities to enhance interactivity and interaction during continuous live lectures [59,72,76]. For example, in one study, group members explored interactively around each member's individual case and provided integration opportunities to advise each other based on feedback or their own experience [76].

Behavioral Regulation

Of the included studies, 16 studies included behavioral regulatory mechanisms as an important component of promoting lifestyle change [50,52,53,56-60,62,63,66,67,70-72,75].

Thirteen studies asked participants to regularly self-record health-related behaviors such as body weight, dietary intake, physical activity, and sleep [50,52,56,59,60,62,63,66,67,70-72,75]. Among them, one high-quality study used the Artificial Intelligence Health Coaching (HCAI) application to prompt users to enter their weight weekly and record the intake of meals and snacks. The system automatically classifies foods as "healthy," "unhealthy," or "neutral" to help users form a clearer perception of diet [75]. Another high-quality study introduced the "SPACE" model, which encouraged participants to record five types of daily health-related habits—Sleep, Presence, Activity, Creativity, and Eating. To enhance overall self-awareness [52].

Three other studies used photo-capable "food diaries" in addition to self-recording measures, in which participants were asked to upload photos of each meal, snack, or drink to an App. Health coaches reviewed the records weekly and provided personalized feedback, thereby reinforcing behaviors [53,57,58].

Beliefs About Competence

Only one study addressed the mechanistic pathway of beliefs about capabilities in diabetes prevention [53]. Based on the concept of social cognitive theory, the Transform program aims to help participants improve self-efficacy, enhance participation, and influence behavior.

Reinforcement

Overview

Among the included studies, 10 reported using reinforcement strategies to enhance participants' self-efficacy and support the maintenance of health behaviors. These strategies included structured encouragement, coach-delivered encouragement, and incentive-based approaches such as points-based rewards and gamification [50,59,62,64-68,76,77].

Systematic Encouragement (n=4)

Four studies used programmed systematic incentive strategies to continuously encourage and strengthen participants' participation through algorithmically driven feedback [50,64,65,76]. For example, users can see progress and send encouragement notifications. Users are provided with daily and weekly summaries, with weekly summaries graphically providing the amount consumed, total calories consumed, calories burned, and projected weight loss goals for the week for each food group [50].

Coach Encouragement (n=4)

Four studies emphasized reinforcement strategies at the interpersonal level, which continuously strengthened participants' competency beliefs through encouragement of structured programs by health coaches [66], daily information encouragement [67], progress encouragement [77], and social support [59]. In a high-quality study, participants received daily messages from their coaches, such as "No pains, no gains. Focus on the application as much as possible. Record as much as you can, as this gives your coach a sense of your health habits and provides background information for all discussions. The longer you use the app, the more likely you are to stick to your health goals" [67].

Points and Gamification Encouragement (n=2)

In one study, participants earned points for setting and achieving goals, which were redeemed for small cash rewards to encourage them to continue reporting health behaviors and achieving goals [68]. Another study added a digital gold reward system to the group chat system, where participants could collect digital currency by completing questionnaires, quizzes, or meeting goals and redeem it for actual prizes. Researchers have further embedded gamification elements to improve participants' participation in prevention programs [62].

Memory, Attention, and Decision Processes

Two studies referred to studies related to memory, attention, and decision processes [52,65]. One study helps participants to maintain focus and self-regulation by providing behavioral strategies of mindfulness and decompression [65]. In addition, it predicts the energy and willpower of the individual for a future day. Help individuals understand and control the factors that affect their health behaviors and strengthen their decision-making awareness [52].

Emotions

Only one study incorporated emotion-related mechanisms in prevention by using artificial intelligence (AI)–enabled guidance to provide compassionate, feedback-informed support. Providing

XSL•FO
RenderX

strategic help and emotional support, AI incorporates interactive elements such as reflection, legitimation, respect, support, and partnership to make users feel valued [50].

The Dimensions of eHL and Intervention Outcomes

The classical framework [26] divides eHL into 6 dimensions. Later, Paige et al [27] proposed a new classification after 2018, which emphasized the interactive use, communication, and behavioral transformation of digital health information in real situations.

Traditional Literacy

Five studies addressed traditional literacy [51,54,63,65,69]. Four of these directly intervened in participants' reading comprehension through a forced reading task. For example, in the study by Block et al [65], participants were required to read 2-weekly health notes written by certified diabetes educators and behavioral specialists. In addition, one study conducted an adaptive intervention at the language and cultural level according to the traditional literacy level of participants [51]. The study adapted the DPP curriculum for low-income people, including rewriting content at the fourth and fifth grade reading level, cultural adjustments, and Spanish translation, and adding bilingual and biculture-based health coaches in English and Spanish to help them better understand prevention programs.

Health Literacy

A total of 27 studies addressed health literacy [50-60,62-77], and interventions were broadly divided into 2 aspects.

First, most studies used health education courses as the intervention core (n=16) [50,51,53,55,57-59,63,64,66,67,69,74-77], covering topics such as healthy diet [54,62,63,65,71], physical activity [54,62,63,65], knowledge of prediabetes [71], etc. Three studies integrated the official Centers for Disease Control and Prevention (CDC)-DPP curriculum.

Second, 7 studies incorporated personalized content into health literacy interventions [6,8,12,19,20,26,27]. In particular, 2 of these studies introduced AI and algorithm-driven personalized health content. The research of Stein and Brooks [75] uses AI to simulate health coaches to provide counseling support, provide strategic help and emotional support, and AI integrates interactive elements such as reflection, legitimization, respect, support, and partnership to make users feel valued. And provide personalized content. The research of Everett et al [68] uses the advanced Sweetch algorithm to provide users with personalized, contextualized, immediate, and timely recommendations to guide users to achieve recommended activity, weight loss, and diet goals in a way that conforms to users' actual living habits. Personalized goals are constantly adjusted according to the actual behavior of the user to best fit the actual life capabilities of the user. Notifications constitute "teachable moments" that can bring about sustained behavioral change.

Information Literacy

A total of 20 studies addressed information literacy [50,52-54,56-58,60,62-67,69-72,75,76]. Among them, information recording is the most common type, and participants in 11 studies actively record personal health data such as weight, diet, step count, physical activity, sleep, etc, through apps, dashboards, or online platforms [52,56,60,63,66,67,69,70,72,75,76]. Of these, 4 studies specifically used "health diaries" to enhance information literacy through daily journaling [50,53,62,71]. For example, the study by Srivastava et al [50] entered participants' food eaten, exercise completed, educational sessions completed, and body weight into a diary, which resulted in automated daily and weekly performance summaries and performance-based tailored daily and weekly recommendations.

Second, information was understood and evaluated through quizzes. In 11 studies, participants were required to complete a course quiz or answer questions [50,53,54,57,58,60,62,64,65,69,71]. For example, in the study by Fitzpatrick et al [58], participants were required to complete questionnaires, quizzes, and open-ended questions. Course completion was defined as completion of the quiz attached to the end of the course content.

Science Literacy

Among the included studies, 9 studies involved the cultivation of scientific literacy [50,52,56,61-63,71,75,76]. Of these, 7 studies focused on dietary knowledge based on scientific principles [50,56,61,63,71,75,76]. For example, in the study of Stein and Brooks [75], the AI-based text interactive HCAI system will classify the food entered by the user into 3 categories of "healthy," "unhealthy," and "neutral" to facilitate the user to establish a food recognition mechanism. In the study by Ng et al [71], the application module covered multiple aspects of healthy eating and practical nutrition advice for people with prediabetes, including content on energy balance, portion control, healthy eating based on Malaysian indigenous foods, glycemic index, dietary fiber intake, fat and protein food types, and alcohol intake.

In addition, 2 studies addressed scientific mechanisms of disease [62,71]. For example, in the study of Chung et al [62], the TCM mHealth app included knowledge about constitution classification, meridian energy, and descriptive illustrations and video teaching of Qigong types such as Baduanjin and abdominal breathing. The Ng et al [71] study provided knowledge on the pathophysiology and complications of prediabetes, health benefits, and specific recommendations of physical activity, health risks of smoking, stress management strategies, and motivation training for lifestyle changes in the app. These science literacy training strategies help participants to understand the causal relationship between health behaviors and disease prevention from a scientific perspective, thereby improving health decision-making ability.

Media Literacy

Of the 28 included studies, none explicitly addressed the cultivation of media literacy, which is the ability to critically evaluate digital health information sources, identify misleading content, or understand algorithmic mechanisms, among other things.

XSL•FO
RenderX

Computer Literacy

According to the analysis of the included studies, 11 studies involved the cultivation of computer literacy [52,53,56-58,62,63,67,70,71,77]. These interventions mainly help participants master the skills of using digital health tools through platform operation training and multimedia content creation and uploading.

First, platform operation training is the basis of computer literacy training [52,56,62,63,67,70,71,77]. For example, in the study by Lim et al [56], participants in the intervention group were introduced to the nBuddy Diabetes mobile app. They were asked to download the nBuddy Diabetes app and receive training, and in the study by Michaelides et al [67] and Toro-Ramos et al [70], participants were asked to learn how to use the Noom app and interact online with a health coach.

Second, 3 studies involved multimedia content creation and uploading. For example, participants were asked to track eating behavior by taking and uploading eating photos [53,57,58].

Functional eHealth Literacy

Of the included studies, 4 studies addressed functional eHL [51,54,63,65].

First, the ability to read health information is a core competency of functional literacy [54,63,65]. For example, in the study by Vadheim et al [63], each participant was asked to read books that estimated the fat and calorie content of specific foods and to correctly record fat, body weight, calories, and minutes of activity in an electronic diary. The study by Islam et al [54] designed practice activities in workshops to help participants master important nutrition-related skills, such as reading food nutrition labels.

Second, culturally adaptive content adjustment reflects the attention paid to people with different levels of functional literacy [51]. In the study by Kim et al [51], the DPP curriculum was adapted for low-income populations, including rewriting content to fourth- and fifth-grade reading levels, cultural adjustment and Spanish translation, and adding bilingual and bicultural health coaches in English and Spanish. This design lowers the reading threshold and ensures that participants of different educational backgrounds and language abilities can effectively understand and use digital health information.

Communicative eHealth Literacy

According to the analysis of the included studies, 20 studies related to communicative eHL [51,53-59,64-67,69-76].

A total of 11 studies demonstrated the cultivation of participants' collaboration ability [51,54,57-59,66,69,71,73,74,76]. These studies promote participant prevention through discussion, providing social support, sharing experiences, etc. For example, in the study of Islam et al [54], the study set up a "brainstorming" session, where peer leaders shared questions/prompts related to healthy eating, physical activity, and healthy lifestyle barriers (stress, difficult emotions, school/work schedule, family stress, etc), and participants were invited to share ideas. In the study by Collins et al [59], program participants were allowed to interact during live lectures to share successes/challenges and provide support to each other.

Translational eHealth Literacy

Of the included studies, 12 involved translational eHL [50,52-54,56,61,62,65,68,70,71,76]. The common feature of these studies is the immediate translation of digital health information and self-monitoring data into concrete, executable behavioral goals through regular individualized feedback and the dynamic revision of these goals in subsequent stages. For example, in the study of Block et al [65], the system generates personalized dietary and exercise goals for participants every week and automatically adjusts small step goals for the next week based on their actual achievement in the previous week. Participants are required to report their goal completion every week, and the system continuously updates the follow-up plan accordingly. In contrast, in the study by Alwashmi et al [53], this regulatory process was mainly accomplished by human coaches. The coaches reviewed participants' follow-up weekly, provided specific recommendations for diet and meal plan changes, and helped participants set more feasible goals for the coming week, thereby reinforcing behavior change on the basis of progress.

Critical eHealth Literacy

Among the 28 studies included, none addressed critical eHL, which is the ability to evaluate the credibility, relevance, and risks of health information shared and received online.

Intervention Outcomes

Engagement

Overall, studies reported high levels of participant engagement [50,52,55,58-60,63-65,68,69,74-77]. Of note, 2 studies achieved 100% participation [52,76]. In one of these studies, app+ completion rates reached 100% in the intervention group. Participants used app+ for an average of 37 days over 12 weeks [52]. In the other study, meeting attendance was 100%, participants had a mean age of 65.4 years, wore a fitness tracker daily, were weighed more than 3 times per week, and recorded food intake more than 3 times per week. During the 6-month reporting period, participants adhered to the target calorie range 74% of the time. The initial daily step goal was 5000, and 92% of participants maintained this goal through 6 months [76].

In terms of course completion, participation was better [50,58,59,63,69,77]. Among them, in one study, 88% of participants completed 16 weeks of courses [63]. In another study, 95% completed at least 4 weekly education courses, and 92% completed 9 or more courses [69]. Across different educational stages, nearly 65% of high school or equivalent people completed the first stage of courses. For participants who graduated from college, the completion rate of phase 1 reached 75.5% [59].

In terms of online platform use, participants similarly demonstrated strong engagement. The study showed that all participants completed at least 4 core modules, and participants logged in to the online intervention 76 times on average, weighed 89 times, sent 46.5 group messages, and sent 20.5 messages to the health coach [64]. Another study reported that

XSL•FO
RenderX

70.8% of intervention group participants interacted with the program at least once a week. After 6 months, 70.6% were still actively using the Alive-PD system [65]. In another study, participants logged on an average of 22.5 times weekly, weighed themselves 7.3 times weekly, performed 11 diet/activity tracking sessions, and had 1 private message with a health coach. In addition, 96% (n=455) of participants were enrolled in the digital DPP program for the full 12 months [58].

Acceptance or Satisfaction

Six of the included studies [52,54,60,64,68,75] explored acceptance or satisfaction. Six studies received positive reviews from participants, and in 2 of these studies, participants clearly expressed a preference for prevention programs [52,60]. In the study by Moin et al [64], women veterans with prediabetes after the prevention program rated the program as a good fit with perceived health needs, "it's perfect," "it can keep its own pace," "flexible, it can be free from time constraints," the program was integrated into daily life and made people feel responsible. In the study of Everett et al [68], Sweetch mobile platform had a high acceptance among participants. Most participants agreed or strongly agreed that they would like to use the Sweetch app often (74%), thought the app was easy to use (83%), thought the features were well integrated (72%), thought most people could learn to use the app quickly (89%), and felt confident in using the app (77%). The value and practicability of the project were highly recognized by the participants. In the study by Islam et al [54], participants indicated that the online adolescent health education program was beneficial and applicable to their lives, giving them a deeper understanding of healthy eating and an active lifestyle. When asked to describe items with adjectives, the most commonly used words were "interesting" and "useful." Participants appreciated the Let's Move activities and discussed how the activities had helped them understand healthy eating and share what they had learned with friends and family. Most participants found the game both fun and rewarding, enjoyed teamwork, enjoyed competition, felt that the game helped participants become more comfortable with each other, and appreciated that the game challenged them to remember important concepts they had learned.

In addition, 3 studies [60,64,72] also pointed out the challenges of eHL participation in prevention. In the study by Moin et al [64], participants expressed an aversion to records, a lack of openness online, and a lack of computer literacy, leading to comprehension difficulties.

In the study of Yeh et al [60], computer literacy also encountered challenges. Some participants experienced difficulties in the first 2 weeks, including difficulty recording food intake and low Facebook group participation, with only 2 people interacting. Participants had barriers to software use. In the study by Audet et al [72], older participants still complained about the use of video technology in group meetings and expressed difficulties in developing computer literacy.

In addition, the study of Yeh et al [60] also pointed out the challenges brought by cultural differences. Chinese-American participants expressed challenges in recording weight in prevention programs. Because in Chinese culture, meals are usually shared in a family format, and there are no individual portion sizes.

Body Weight

Of the 28 included studies, 22 reported on weight management results, showing a significant effect of eHL in promoting weight management in people at risk of diabetes [50,51,53,55-61,63-70,74-77].

In short-term intervention studies (≤6 mo) [50,53,56,61,63-65,68,75-77], weight loss has been substantial. Of these, 6 studies explicitly reported that participants lost 5% or more of their body weight [50,53,63-65,76]. For example, in the study by Srivastava et al [50], all 6 users who completed the study achieved the CDC target of 5% weight loss at 6 months, with an average weight loss of approximately 9.0%. In another high-quality study [76], weight was reduced by 5.3% at 3 months and 8.4% at 6 months in the intervention group, with 90% of participants achieving and maintaining the target weight loss of 5% or more and a reduction in BMI from baseline 31.3 to 29.06.

Weight loss was maintained in medium- and long-term intervention (≥9 mo) studies. Five studies [55,57,59,67,70] reported weight changes in participants from 9 months to 65 weeks later. For example, Katula et al [55] showed that patients in the d-DPP group lost an average of 5.49% of their initial weight at 12 months and were 61% more likely to have clinically significant weight loss (≥5%) than those in the control group, and that greater participation was associated with greater weight loss, especially among participants 65 years of age or older.

Six studies reported trends in weight change at different time points [51,58,60,66,69,74]. For example, Castro Sweet et al [69] showed that participants lost an average of 6.5% of body weight at 16 weeks, increased to 8.0% at 6 months, and remained at 7.5% at 12 months, with an average weight loss of 13 - 14 pounds. Long-term follow-up by Sepah et al [74] showed a mean weight loss of 4.9% after 1 year and 4.3% after 2 years among patients who completed the program, demonstrating sustained weight loss maintenance. Of note, in the study by Fitzpatrick et al [58], the average weight loss was 8.6 pounds (4.0%) within 12 months, which was clinically significant, but at 24 months, the percentage of weight loss was 5.7 pounds, which was no longer statistically significant, and long-term weight maintenance was still a challenge.

Diabetes Indicators

Of the 28 included studies, 12 reported improvements in measures of glycemic control [51,55,56,59,61,62,65,66,68-70,74]. Ten of these studies showed positive results [51,55,56,59,61,62,65,66,69,74], and 2 studies [68,70] reported no significant change from baseline in fasting plasma glucose and glycosylated hemoglobin (HbA_{1c}).

Seven studies reported reductions in HbA_{1c} [51,55,59,62,66,69,74]. Among them, 3 studies [62,66,69] found sustained and sustained reductions in HbA_{1c} over time after eHL engagement prevention studies. For example, in the study by Sepah et al [74], the mean reduction in glycated hemoglobin levels was 0.38% after one year and 0.43% after 2 years. In a

XSL•FO
RenderX

long-term study by Sepah et al [66], significant reductions in glycated hemoglobin levels were observed at week 16 and at years 1, 2, and 3.

In addition, 5 studies reported improvements in fasting plasma glucose [56,61,62,65,69]. In a high-quality study, participants' fasting blood glucose decreased by 0.41 at 6 months, and their Framingham 8-year risk for diabetes decreased from 16% to 11%. Most encouragingly, 40.5% of the intervention group achieved normal fasting glucose levels at 6 months.

In terms of other diabetes-related measures, one study showed significant within-group improvements in systolic and diastolic blood pressure, total cholesterol, and low-density lipoprotein cholesterol after 6 months in the intervention group [56], and another study combined a TCM mobile health app intervention. It was shown that participants performed better than the control group in terms of body constitution (improvement of Yang deficiency and sputum stasis), body energy improvement, and health-related quality of life (PCS and MCS) after 4 months. Prevention that demonstrates eHL engagement appears to be more likely to improve physical energy and health-related quality of life [62].

Behavioral Change

Results of a total of 13 studies showed that eHL participation in prevention research promoted the improvement of health behaviors of participants [50,52-54,56,61,65,68,72,73,75,77]. Of these, 8 studies involved improvements in dietary behavior among participants [52,54,56,61,65,72,75,77]. For example, in the study by Lim et al [56], the daily intake of total energy, carbohydrates, total fat, saturated fat, and sugar decreased significantly after 6 months in the intervention group. Another pilot study [54] found that participants learned about healthy eating during the program and were able to share what they learned with family and friends. As one participant commented, "The program helped me to adjust my eating habits, to know what to eat, how much to eat, and how much to plan."

Second, 9 studies showed changes in physical activity among participants [50,52,53,57,61,65,68,72,73]. For example, in the study by Ekezie et al [73], participants significantly increased their physical activity levels and average daily steps. A wrist-worn activity monitor linked to a digital app tracked daily steps in real time, and the "physical activity challenge" feature of the program was considered a key enabler of goal attainment, helping participants stay focused and motivated to continue achieving their goals over a longer period of time, they said. Another study further validated the improvement in motor behavior. Zahedani et al [61] found that participants with prediabetes had a significant increase in their daily physical activity after 3 months of intervention, with the average duration of physical activity approximately doubling. As an objective measure, the mean daily heart rate rose by 6.3 minutes/day.

Self-Efficacy

In addition, 2 studies also reported significant improvements in self-efficacy among participants [59,65]. In the study by Collins et al [59], participants showed statistically significant improvements in the ability to perform 18 self-reported health promotion behaviors.

Discussion

Principal Findings

To the best of our knowledge, this is the first systematic literature review of T2D prevention in at-risk populations, which aims to deeply explore the mechanism of eHL in T2D prevention in at-risk populations and the participation and effects of multidimensional prevention. Our core finding is that eHL, although extremely uneven in its application, still achieves significant preventive effects. This reveals systemic problems in current prevention research and future directions for improvement.

In this review, we mapped intervention components and reported implementation strategies from the included studies to the TDF domains and quantified how frequently each domain was reflected across the 28 studies. The use of these pathways is extremely uneven. Almost all studies relied on the "knowledge" approach (28/28, 100%), while behavioral regulation (16/28, 57.1%) and social influence (15/28, 53.6%) were also widely used. However, few studies involved emotions (1/28, 3.6%), beliefs about competence (1/28, 3.6%), memory, attention, and decision processes (2/28, 7.1%). This implies that current research constructs a "surface" rather than a "deep" prevention model, with much focus on knowledge that is easily transmitted and behaviors that are easily monitored, but the potential to stimulate intrinsic motivation for change has been largely neglected.

In all the studies included in this review, health literacy was central in terms of understanding. The original intention of such prevention is consistent with the CDC in the United States, which is to translate scientific knowledge into action to improve public health [78]. In addition, 18 studies showed a trend toward more standardized prevention content based on DPP content. Similarly, behavioral regulation and social incentives are relatively easy to achieve through regular recording with various types of incentives. Perhaps the reason for the widespread adoption of easy-to-implement mechanisms lies in their operability. The convenience of such preventive measures may obscure a fundamental question: Is the easy-to-implement path the most long-lasting mechanism?

Our findings suggest that, in terms of implementation, the included studies focused primarily on the goals of algorithmic systems and digital coaches assisting design, and that this design difference has theoretical and practical implications for self-determination theory. Self-determination theory emphasizes autonomy as a key intrinsic motivation for promoting behavior change [79]. Overreliance on algorithms or digital coaches may undermine participant autonomy. Second, the emotional and psychological pathways are also neglected. This is consistent with previous findings that psychological factors have been underreported in digital health care interventions [80]. For example, negative emotions such as the frustration caused by weight loss and blood glucose rebound were obviously ignored in the research, and negative emotions are often one of the reasons for interrupting self-control behavior [81]. Such neglect tends to lead to long-term prevention failure.

XSL•FO
RenderX

Second, eHL is by no means a trivial set of skills in a highly digital environment. Rather, it combines knowledge and skills from a variety of fields [82]. Our review found that the framework of Norman et al [26] and Paige et al [27] is highly fragmented and selectively applied in practice.

Specifically, health literacy and information literacy were the 2 most frequently involved dimensions. On the contrary, although the attention to scientific literacy increased year by year, the attention to traditional literacy was completely absent after 2022. Media literacy is more worrisome, and none of the prevention studies addressed the cultivation of media literacy.

There are many explanations for the uneven distribution of dimension involvement in prevention. First, the concept of eHL has been discussed and applied in the field of public health and health communication since it was proposed by Norman et al [26]. In contrast, media literacy stems more from the field of media studies and has not yet been fully integrated. This has led to a critical gap: the Internet has become an important channel for the dissemination of health information for the prevention of T2D, but the requirements of media literacy, such as how to identify the correct medical information, have not been systematically incorporated into preventive medical interventions.

However, the interest in scientific literacy has been increasing in recent years, especially after the COVID-19 pandemic, which is a positive sign to some extent. However, it may simply be because of the external pressure of the global nature of the pandemic.

In addition, the lack of traditional literacy, perhaps due to the popularity of digitalization, researchers have neglected that online functions still require basic reading and writing skills. It has been found that a disadvantage in written and oral language skills can be a barrier to accessing web-based health information [83]. This means that current prevention studies may be inadvertently excluding participants who have low literacy and difficulty reading and typing. People with lower education levels are generally at higher risk of chronic diseases [84]. Thus, the lack of traditional literacy is not only a study design issue, but also a health equity issue. To ignore traditional literacy is to exclude the most vulnerable from digital health help, exacerbating divisions.

In 2018, Paige et al [27] classified the new model as a further improvement of the old model, from “instrumental skills” to “interactive application ability.” However, this study found further “selective neglect” in the study: communicative eHL and translational eHL were more involved in prevention, but functional eHL was less involved, and critical eHL was completely missing. Once again, this demonstrates the “selective” application mechanism of eHL in practical prevention research.

Encouragingly, research has begun to directly acknowledge and respond to inequalities in functional literacy. The diabetes prevention curriculum was adapted for low-income people, rewritten to fourth to fifth grade reading difficulty, culturally adjusted, and translated into Spanish, and added bilingual and bicultural health coaches. The breakthrough of this approach is that it takes the initiative to consider and lower the threshold of health information, so that participants with different cultural backgrounds and language abilities have the opportunity to participate in digital health, which is particularly critical in the field of diabetes prevention. Because the lack of functional eHL can further lead to the emergence of a digital divide, there are significant differences in the need for technical training among different populations. Among the included studies, studies involving, for example, older adults [58,76], their training in the use of digital tools need more targeted attention.

Second, communicative eHL had the highest frequency, also responding to the absence of the “emotional information factor” in the path effect. For those at risk, the experience of being “understood and supported” is itself a source of motivation to continue to participate. This motivation is further strengthened by the translational eHealth hormone. Participants were trained and guided repeatedly, leading to more long-term, routine behavior change. An emerging trend observed in this study is that AI is gradually becoming an important tool for eHL interventions. In one of the included studies, an AI health guidance application (HCAI application) provided a human-like guidance experience and personalized content [75]. AI is playing an increasingly important role in enhancing diabetes prevention and management [85]. The introduction of AI, to some extent, strengthens “translational eHealth literacy,” and AI shows great potential in long-term prevention.

The absence of critical eHL, however, is consistent with media literacy. This is also one of the biggest concerns of this review. The current digital environment is extremely complex, and it is very important for individuals to be able to include and judge the credibility and reliability of information sources. The included studies were mainly conducted within the framework of guidance from medical professionals, and participants were not required to assess whether such programs were disinformative. However, once this support is withdrawn, and the participants’ lack of autonomy is fostered, it is extremely easy to lead to difficulties in self-management. Therefore, we suggest that future prevention research could embed training modules on “how to avoid false judgments” while maintaining “learning to make good decisions,” for example, scoring the credibility of health information, exercises on checking health information, etc, mitigating the double risks of information overload and misinformation.

More urgently, the review 5 years ago already noted that while health literacy was the most assessed domain, none of the studies evaluated all 6 domains of the eHL model [86]. The results of this review still show that none of the 28 studies designed and evaluated eHL as a comprehensive, unified concept. Although the path application of eHL was not balanced with the dimensions of participation, our review suggests that eHL is capable of achieving positive outcomes in terms of T2D prevention outcomes for at-risk populations. The cultivation of eHL can improve the participation, satisfaction, weight management, diabetes-related physiological indicators, health behavior change, and self-efficacy of the prevention of T2D in the risk population. This finding is consistent with existing results on the impact of eHL on health behaviors [87-92].

XSL•FO
RenderX

Therefore, the importance and effectiveness of eHL in the prevention of T2D in the at-risk population are clear. But our findings need to highlight how eHL can be a more sustainable long-term driver that needs to be addressed by health communication. In the future of T2D prevention, especially with the rapid development of AI, there is an urgent need to incorporate "autonomous" literacy training without professional guidance into prevention design, rather than continue to assume that participants will automatically have these abilities. Only in this way can eHL help public health mitigate the imbalances brought about by the information age, especially in at-risk and vulnerable populations, and reduce the incidence of T2D in the long term.

This review is novel in providing a theory-driven, mechanistic synthesis of how eHL supports T2D prevention in at-risk populations. We mapped and quantified intervention components and implementation strategies against the TDF and multidimensional eHL frameworks, revealing a marked imbalance in mechanisms: studies largely rely on knowledge and easily operationalized behavioral pathways, while underrepresenting mechanisms critical for long-term maintenance (eg, emotions, support toward autonomy/self-efficacy, and cognitive decision processes). Our contribution is a practical mechanism-by-dimension framework and an evidence-gap agenda, highlighting common omissions such as critical eHL and media literacy, and limited attention to functional literacy and access barriers as determinants of equity. In real-world implementation, future digital/AI-enabled prevention should move beyond information delivery to embed affective and autonomy-supportive design, misinformation-resilience (critical appraisal) training, and accessibility adaptations for vulnerable groups; AI should be integrated cautiously to avoid undermining autonomy and to enhance long-term sustainability and health equity.

Limitations

This review has several limitations. First, although the study aimed to explore the impact of eHL on the prevention of T2D, we did not conduct a meta-analysis due to the substantial heterogeneity observed among the included studies in terms of research design, intervention approaches, and outcome measures.

Second, as "eHealth literacy" remains an emerging concept in the research landscape, the volume of available literature is still limited. Consequently, the relatively small sample of studies included in this review may restrict the comprehensiveness of our understanding of the field.

Third, the thematic analysis used in this review was qualitative in nature. Most of the included studies did not directly investigate the impact of eHL on T2D prevention but instead provided indirect evidence or descriptions of relevant interventions. As a result, the mechanisms through which different dimensions of eHL contribute to the prevention process were derived primarily through interpretive synthesis rather than quantitative validation. This type of analysis may be subject to researcher subjectivity and carries a risk of reporting bias. To mitigate this risk and enhance the transparency of the findings, all authors were involved in reviewing each stage of the data analysis process.

Accordingly, several strategies are proposed for future research. First, to address the high degree of heterogeneity across studies, future research should consider conducting more randomized controlled trials with clearly defined control groups and standardized intervention protocols. Such studies would help generate high-quality empirical data suitable for quantitative synthesis and lay the groundwork for future meta-analyses.

Second, future studies should conceptualize eHL as an integrated and unified construct rather than examining its dimensions in isolation. Researchers are encouraged to adopt standardized measurement tools, such as the eHL Scale, for systematic assessment. Moreover, rather than focusing solely on the indirect role of eHL in digital interventions, future studies should investigate its direct effects as a core variable. Doing so will help clarify its independent contributions and support the standardization and theoretical development of eHL in the context of disease prevention.

Third, to reduce potential bias and enhance the credibility and validity of findings, future research may consider incorporating external peer review or expert consensus methods in the analytical process. For example, using the Delphi method to collect structured input from multidisciplinary experts can enhance the robustness and transparency of the results. These strategies help ensure that the conclusions are not solely reliant on the subjective judgments of the primary research team.

Conclusions

This mechanistic systematic review integrated the TDF and eHL models through theory-driven thematic analysis. It focuses on the pathways through which eHL influences prevention among individuals at risk for T2D. Findings indicate that eHL supports sustained engagement and self-management through mechanisms such as knowledge and skill acquisition, goal setting, behavioral regulation and self-monitoring, resource usage, and social influence, and is associated with outcomes including weight control, improved glycemic indicators, self-efficacy, and lifestyle change. Most studies reported high engagement rates and satisfaction levels, suggesting good acceptability and feasibility of intervention programs.

However, the exploration of mechanisms for T2D prevention across current studies remains uneven. Emotional support, beliefs about competence, and autonomy were underaddressed, while critical eHL and media literacy dimensions were largely absent. The default assumption of traditional and functional eHL's importance may inadvertently exclude resource-limited, at-risk populations, exacerbating health inequities and the digital divide.

Future digital DPPs must better serve diverse populations. Our findings reveal that most programs focus heavily on information delivery but neglect critical factors that help people sustain behavior change: managing emotional setbacks, supporting personal autonomy, and distinguishing reliable health information from misinformation. We recommend that programs incorporate emotional support to address frustration and disappointment, empower users to make independent choices rather than relying solely on algorithm recommendations, and teach simple skills for evaluating web-based health information

XSL•FO
RenderX

credibility. Additionally, programs should include media literacy training, use engagement mechanisms such as incentives and encouragement, and adapt designs to accommodate different ages, education levels, and digital access capabilities. Only through such comprehensive, equity-focused approaches can eHL serve as an effective public health tool for broader populations, particularly vulnerable and at-risk groups.

Funding

This study did not receive external funding.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

JL, AAA, and ZZH contributed to the conceptualization and methodology of the study. JL was responsible for data curation, investigation, formal analysis, provision of resources, and project administration. WY and RG conducted the screening and selection of studies, performed data extraction, and carried out the critical appraisal. Supervision was provided by AAA and ZZH. JL prepared the original draft of the manuscript, and JL, AAA, and ZZH contributed to the review and editing of the manuscript. All authors have read and approved the final version of the manuscript. No generative artificial intelligence tools were used in the writing of this manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Searching strategies.
[DOCX File, 22 KB - jmir_v28i1e77788_app1.docx]

Checklist 1
PRISMA 2020 checklist.
[PDF File, 169 KB - jmir_v28i1e77788_app2.pdf]

Checklist 2
PRISMA-S checklist.
[DOCX File, 15 KB - jmir_v28i1e77788_app3.docx]

References

1. Lu X, Xie Q, Pan X, et al. Type 2 diabetes mellitus in adults: pathogenesis, prevention and therapy. Sig Transduct Target Ther 2024 Oct 2;9(1):262. [doi: 10.1038/s41392-024-01951-9]
2. IDF global clinical practice recommendations for managing type 2 diabetes 2025. Diabetes Res Clin Pract 2025 Jun;224:112238. [doi: 10.1016/j.diabres.2025.112238]
3. Udler MS, McCarthy MI, Florez JC, Mahajan A. Genetic risk scores for diabetes diagnosis and precision medicine. Endocr Rev 2019 Dec 1;40(6):1500-1520. [doi: 10.1210/er.2019-00088] [Medline: 31322649]
4. Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. Nat Rev Endocrinol 2018 Feb;14(2):88-98. [doi: 10.1038/nrendo.2017.151] [Medline: 29219149]
5. DeFronzo RA, Ferrannini E. Insulin resistance: a multifaceted syndrome responsible for NIDDM, obesity, hypertension, dyslipidemia, and atherosclerotic cardiovascular disease. Diabetes Care 1991 Mar 1;14(3):173-194. [doi: 10.2337/diacare.14.3.173]
6. Reaven GM. Banting lecture 1988. role of insulin resistance in human disease. Diabetes 1988 Dec;37(12):1595-1607. [doi: 10.2337/diab.37.12.1595] [Medline: 3056758]
7. Rossetti G, Ermer JA, Stentenbach M, et al. A common genetic variant of a mitochondrial RNA processing enzyme predisposes to insulin resistance. Sci Adv 2021 Sep 24;7(39):eabi7514. [doi: 10.1126/sciadv.abi7514] [Medline: 34559558]
8. Cowie CC, Harris MI, Silverman RE, Johnson EW, Rust KF. Effect of multiple risk factors on differences between blacks and whites in the prevalence of non-insulin-dependent diabetes mellitus in the United States. Am J Epidemiol 1993 Apr 1;137(7):719-732. [doi: 10.1093/oxfordjournals.aje.a116732] [Medline: 8484363]
9. Ferrannini E, Gastaldelli A, Matsuda M, et al. Influence of ethnicity and familial diabetes on glucose tolerance and insulin action: a physiological analysis. J Clin Endocrinol Metab 2003 Jul;88(7):3251-3257. [doi: 10.1210/jc.2002-021864] [Medline: 12843172]

XSL•FO
RenderX

10. Tam WH, Ma RW, Yang X, et al. Cardiometabolic risk in Chinese women with prior gestational diabetes: a 15-year follow-up study. Gynecol Obstet Invest 2012;73(2):168-176. [doi: 10.1159/000329339] [Medline: 22179684]
11. Lisco G, Giagulli VA, De Pergola G, et al. Chronic stress as a risk factor for type 2 diabetes: endocrine, metabolic, and immune implications. Endocr Metab Immune Disord Drug Targets 2024 Mar;24(3):321-332. [doi: 10.2174/1871530323666230803095118]
12. Wicke FS, Otten D, Schulz A, et al. Current and past depression as risk factors for incident type 2 diabetes mellitus and pre-diabetes in men and women: evidence from a longitudinal community cohort. Diabetol Metab Syndr 2024;16(1). [doi: 10.1186/s13098-024-01273-4]
13. Darraj A. The link between sleeping and type 2 diabetes: a systematic review. Cureus 2023 Nov;15(11):e48228. [doi: 10.7759/cureus.48228] [Medline: 38050514]
14. Hu S, Ji W, Zhang Y, Zhu W, Sun H, Sun Y. Risk factors for progression to type 2 diabetes in prediabetes: a systematic review and meta-analysis. BMC Public Health 2025 Mar 31;25(1):1220. [doi: 10.1186/s12889-025-21404-4]
15. Tabák AG, Herder C, Rathmann W, Brunner EJ, Kivimäki M. Prediabetes: a high-risk state for diabetes development. The Lancet 2012 Jun;379(9833):2279-2290. [doi: 10.1016/S0140-6736(12)60283-9]
16. Nah EH, Chu J, Kim S, Cho S, Kwon E. Efficacy of lifestyle interventions in the reversion to normoglycemia in Korean prediabetics: one-year results from a randomised controlled trial. Prim Care Diabetes 2019 Jun;13(3):212-220. [doi: 10.1016/j.pcd.2018.11.017] [Medline: 30583933]
17. Herman WH, Hoerger TJ, Brandle M, et al. The cost-effectiveness of lifestyle modification or metformin in preventing type 2 diabetes in adults with impaired glucose tolerance. Ann Intern Med 2005 Mar 1;142(5):323-332. [doi: 10.7326/0003-4819-142-5-200503010-00007] [Medline: 15738451]
18. Diabetes Prevention Program (DPP) Research Group. The diabetes prevention program (DPP): description of lifestyle intervention. Diabetes Care 2002 Dec;25(12):2165-2171. [doi: 10.2337/diacare.25.12.2165] [Medline: 12453955]
19. Auster-Gussman LA, Lockwood KG, Graham SA, Stein N, Branch OH. Reach of a fully digital diabetes prevention program in health professional shortage areas. Popul Health Manag 2022 Aug;25(4):441-448. [doi: 10.1089/pop.2021.0283] [Medline: 35200043]
20. Parsons AS, Raman V, Starr B, Zezza M, Rehm CD. Medicare underpayment for diabetes prevention program: implications for DPP suppliers. Am J Manag Care 2018 Oct;24(10):475-478. [Medline: 30325189]
21. Ely EK, Gruss SM, Luman ET, et al. A national effort to prevent type 2 diabetes: participant-level evaluation of CDC's national diabetes prevention program. Diabetes Care 2017 Oct 1;40(10):1331-1341. [doi: 10.2337/dc16-2099]
22. Lee EC, Grigorescu V, Enogieru I. Updated National Survey Trends in Telehealth Utilization and Modality (2021-2022): Issue Brief: Washington (DC): Office of the Assistant Secretary for Planning and Evaluation (ASPE); 2023. [Medline: 38913813]
23. Iyengar V, Wolf A, Brown A, Close K. Challenges in diabetes care: can digital health help address them? Clin Diabetes 2016 Jul;34(3):133-141. [doi: 10.2337/diaclin.34.3.133] [Medline: 27621530]
24. Haseldine C, O'Donoghue G, Kearney PM, et al. Characteristics of participants in the first fully online national diabetes prevention programme: a quantitative survey. HRB Open Res 2023;6:61. [doi: 10.12688/hrbopenres.13807.4] [Medline: 39931144]
25. Peng RX. How online searches fuel health anxiety: investigating the link between health-related searches, health anxiety, and future intention. Comput Human Behav 2022 Nov;136:107384. [doi: 10.1016/j.chb.2022.107384]
26. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9. [doi: 10.2196/jmir.8.2.e9]
27. Paige SR, Stellefson M, Krieger JL, Anderson-Lewis C, Cheong J, Stopka C. Proposing a transactional model of eHealth literacy: concept analysis. J Med Internet Res 2018 Oct 2;20(10):e10175. [doi: 10.2196/10175]
28. Milanti A, Norman C, Chan DNS, So WKW, Skinner H. eHealth literacy 3.0: updating the Norman and Skinner 2006 model. J Med Internet Res 2025 Mar 11;27(1):e70112. [doi: 10.2196/70112]
29. Nagori A, Keshvani N, Patel L, Dhruve R, Sumarsono A. Electronic health Literacy gaps among adults with diabetes in the United States: role of socioeconomic and demographic factors. Prev Med Rep 2024 Nov;47:102895. [doi: 10.1016/j.pmedr.2024.102895] [Medline: 39823076]
30. Hsu W, Chiang C, Yang S. The effect of individual factors on health behaviors among college students: the mediating effects of eHealth literacy. J Med Internet Res 2014 Dec 12;16(12):e287. [doi: 10.2196/jmir.3542] [Medline: 25499086]
31. Busse TS, Nitsche J, Kernebeck S, et al. Approaches to improvement of digital health literacy (eHL) in the context of person-centered care. Int J Environ Res Public Health 2022 Jul 7;19(14):8309. [doi: 10.3390/ijerph19148309]
32. Watkins I, Xie B. eHealth literacy interventions for older adults: a systematic review of the literature. J Med Internet Res 2014 Nov 10;16(11):e225. [doi: 10.2196/jmir.3318] [Medline: 25386719]
33. Kim K, Shin S, Kim S, Lee E. The relation between eHealth literacy and health-related behaviors: systematic review and meta-analysis. J Med Internet Res 2023 Jan 30;25:e40778. [doi: 10.2196/40778]
34. Jacobs RJ, Lou JQ, Ownby RL, Caballero J. A systematic review of eHealth interventions to improve health literacy. Health Informatics J 2016 Jun;22(2):81-98. [doi: 10.1177/1460458214534092] [Medline: 24916567]

XSL•FO
RenderX

35. Verweel L, Newman A, Michaelchuk W, Packham T, Goldstein R, Brooks D. The effect of digital interventions on related health literacy and skills for individuals living with chronic diseases: a systematic review and meta-analysis. Int J Med Inform 2023 Sep;177:105114. [doi: 10.1016/j.ijmedinf.2023.105114]
36. McDaniel CC, Kavookjian J, Whitley HP. Telehealth delivery of motivational interviewing for diabetes management: a systematic review of randomized controlled trials. Patient Educ Couns 2022 Apr;105(4):805-820. [doi: 10.1016/j.pec.2021.07.036] [Medline: 34366228]
37. Bingham JM, Black M, Anderson EJ, et al. Impact of telehealth interventions on medication adherence for patients with type 2 diabetes, hypertension, and/or dyslipidemia: a systematic review. Ann Pharmacother 2021 May;55(5):637-649. [doi: 10.1177/1060028020950726]
38. Bassi G, Mancinelli E, Dell'Arciprete G, Rizzi S, Gabrielli S, Salcuni S. Efficacy of eHealth interventions for adults with diabetes: a systematic review and meta-analysis. Int J Environ Res Public Health 2021 Aug 26;18(17):8982. [doi: 10.3390/ijerph18178982] [Medline: 34501572]
39. Duong T, Olsen Q, Menon A, et al. Digital health interventions to prevent type 2 diabetes mellitus: systematic review. J Med Internet Res 2025 Apr 25;27:e67507. [doi: 10.2196/67507] [Medline: 40280555]
40. Suhlrie L, Ayyagari R, Mba C, et al. The effectiveness of telemedicine in the prevention of type 2 diabetes mellitus: a systematic review and meta-analysis of interventions. Diabetes Metab Syndr Clin Res Rev 2025 May;19(5):103252. [doi: 10.1016/j.dsx.2025.103252]
41. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
42. Eysenbach G. What is e-health? J Med Internet Res 2001;3(2):E20. [doi: 10.2196/jmir.3.2.e20] [Medline: 11720962]
43. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 5;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
44. Hong QN, Fàbregues S, Bartlett G, et al. The mixed methods appraisal tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018 Dec 18;34(4):285-291. [doi: 10.3233/EFI-180221]
45. Pace R, Pluye P, Bartlett G, et al. Testing the reliability and efficiency of the pilot mixed methods appraisal tool (MMAT) for systematic mixed studies review. Int J Nurs Stud 2012 Jan;49(1):47-53. [doi: 10.1016/j.ijnurstu.2011.07.002] [Medline: 21835406]
46. Atkins L, Francis J, Islam R, et al. A guide to using the theoretical domains framework of behaviour change to investigate implementation problems. Implement Sci 2017 Jun 21;12(1):77 [FREE Full text] [doi: 10.1186/s13012-017-0605-9] [Medline: 28637486]
47. Naeem M, Ozuem W, Howell K, Ranfagni S. A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. Int J Qual Methods 2023 Oct;22. [doi: 10.1177/16094069231205789]
48. Genapathy S, Mohamad E, Damanhuri H, et al. Communication matters: a systematic literature review on malaria prevention behaviour intervention. J Komun. Malays J Commun 2024;40(4):269-286. [doi: 10.17576/JKMJC-2024-4004-15]
49. Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. Nurs Health Sci 2013 Sep;15(3):398-405. [doi: 10.1111/nhs.12048] [Medline: 23480423]
50. Srivastava P, Verma A, Geronimo C, Button TM. Behavior stages of a physician- and coach-supported cloud-based diabetes prevention program for people with prediabetes. SAGE Open Med 2019;7:2050312119841986. [doi: 10.1177/2050312119841986] [Medline: 31105938]
51. Kim SE, Castro Sweet CM, Cho E, Tsai J, Cousineau MR. Evaluation of a digital diabetes prevention program adapted for low-income patients, 2016-2018. Prev Chronic Dis 2019 Nov 27;16:E155. [doi: 10.5888/pcd16.190156] [Medline: 31775010]
52. Griauzde D, Kullgren JT, Liestenfeltz B, et al. A mobile phone-based program to promote healthy behaviors among adults with prediabetes who declined participation in free diabetes prevention programs: mixed-methods pilot randomized controlled trial. JMIR Mhealth Uhealth 2019 Jan 9;7(1):e11267. [doi: 10.2196/11267] [Medline: 30626566]
53. Alwashmi MF, Mugford G, Abu-Ashour W, Nuccio M. A digital diabetes prevention program (transform) for adults with prediabetes: secondary analysis. JMIR Diabetes 2019 Jul 26;4(3):e13904. [doi: 10.2196/13904] [Medline: 31350833]
54. Islam S, Elaiho C, Arniella G, Rivera S, Vangeepuram N. A pilot study to examine the feasibility and acceptability of a virtual adaptation of an in-person adolescent diabetes prevention program. IJERPH 2022 Sep 27;19(19):12286. [doi: 10.3390/ijerph191912286]
55. Katula JA, Dressler EV, Kittel CA, et al. Effects of a digital diabetes prevention program: an RCT. Am J Prev Med 2022 Apr;62(4):567-577. [doi: 10.1016/j.amepre.2021.10.023] [Medline: 35151522]
56. Lim SL, Ong KW, Johal J, et al. A smartphone app-based lifestyle change program for prediabetes (D'LITE study) in a multiethnic asian population: a randomized controlled trial. Front Nutr 2022;8:780567. [doi: 10.3389/fnut.2021.780567]
57. Batten R, Alwashmi MF, Mugford G, Nuccio M, Besner A, Gao Z. A 12-month follow-up of the effects of a digital diabetes prevention program (VP transform for prediabetes) on weight and physical activity among adults with prediabetes: secondary analysis. JMIR Diabetes 2022 Jan 14;7(1):e23243. [doi: 10.2196/23243]
58. Fitzpatrick SL, Mayhew M, Rawlings AM, et al. Evaluating the Implementation of a digital diabetes prevention program in an integrated health care delivery system among older adults: results of a natural experiment. Clin Diabetes 2022 Jul 1;40(3):345-353. [doi: 10.2337/cd21-0114]

XSL•FO
RenderX

59. Collins C, McCallum E, Nordby K, et al. Enrollment, engagement, and effectiveness of a large-scale diabetes prevention program using synchronous distance technology. Diabetology 2023 Dec 11;4(4):574-585. [doi: 10.3390/diabetology4040051]
60. Yeh MC, Lau W, Keady CA, et al. Evaluation of feasibility and acceptability of a web-based diabetes prevention program (DPP) for diabetes risk reduction in Chinese Americans in New York City. Front Public Health 2023;11:1199746. [doi: 10.3389/fpubh.2023.1199746]
61. Zahedani AD, McLaughlin T, Veluvali A, et al. Digital health application integrating wearable data and behavioral patterns improves metabolic health. NPJ Digit Med 2023 Nov 25;6(1):216. [doi: 10.1038/s41746-023-00956-y] [Medline: 38001287]
62. Chung HW, Tai CJ, Chang P, Su WL, Chien LY. The effectiveness of a traditional Chinese medicine–based mobile health app for individuals with prediabetes: randomized controlled trial. JMIR Mhealth Uhealth 2023 Jun 20;11:e41099. [doi: 10.2196/41099]
63. Vadheim LM, McPherson C, Kassner DR, et al. Adapted diabetes prevention program lifestyle intervention can be effectively delivered through telehealth. Diabetes Educ 2010;36(4):651-656. [doi: 10.1177/0145721710372811] [Medline: 20534873]
64. Moin T, Ertl K, Schneider J, et al. Women veterans' experience with a web-based diabetes prevention program: a qualitative study to inform future practice. J Med Internet Res 2015 May 25;17(5):e127. [doi: 10.2196/jmir.4332] [Medline: 26006697]
65. Block G, Azar KM, Romanelli RJ, et al. Diabetes prevention and weight loss with a fully automated behavioral intervention by email, web, and mobile phone: a randomized controlled trial among persons with prediabetes. J Med Internet Res 2015 Oct 23;17(10):e240. [doi: 10.2196/jmir.4897] [Medline: 26499966]
66. Sepah SC, Jiang L, Ellis RJ, McDermott K, Peters AL. Engagement and outcomes in a digital diabetes prevention program: 3-year update. BMJ Open Diabetes Res Care 2017;5(1):e000422. [doi: 10.1136/bmjdrc-2017-000422] [Medline: 28948027]
67. Michaelides A, Major J, Pienkosz Jr E, Wood M, Kim Y, Toro-Ramos T. Usefulness of a novel mobile diabetes prevention program delivery platform with human coaching: 65-week observational follow-up. JMIR Mhealth Uhealth 2018 May 3;6(5):e93. [doi: 10.2196/mhealth.9161]
68. Everett E, Kane B, Yoo A, Dobs A, Mathioudakis N. A novel approach for fully automated, personalized health coaching for adults with prediabetes: pilot clinical trial. J Med Internet Res 2018;20(2):e72. [doi: 10.2196/jmir.9723]
69. Castro Sweet CM, Chiguluri V, Gumpina R, et al. Outcomes of a digital health program with human coaching for diabetes risk reduction in a medicare population. J Aging Health 2018 Jun;30(5):692-710. [doi: 10.1177/0898264316688791] [Medline: 28553807]
70. Toro-Ramos T, Michaelides A, Anton M, et al. Mobile delivery of the diabetes prevention program in people with prediabetes: randomized controlled trial. JMIR Mhealth Uhealth 2020 Jul 8;8(7):e17842. [doi: 10.2196/17842] [Medline: 32459631]
71. Ng CM, Teoh KW, Chong CW, Ng YL, Bell JS, Lee SWH. A digital health-supported prediabetes intervention, management and evaluation (PRIME) program improves healthy eating index score: a cluster randomized controlled trial. Nutr Res 2025 Jun;138:151-161. [doi: 10.1016/j.nutres.2025.04.005]
72. Audet CM, Roller LJ, Aliyu MH, et al. Evaluation of a workplace diabetes prevention program delivered via distance learning: a qualitative study. Health Educ Behav 2025 Apr;52(2):156-165. [doi: 10.1177/10901981241285433]
73. Ekezie W, Dallosso H, Saravanan P, Khunti K, Hadjiconstantinou M. Experiences of using a digital type 2 diabetes prevention application designed to support women with previous gestational diabetes. BMC Health Serv Res 2021 Aug 5;21(1):772. [doi: 10.1186/s12913-021-06791-9] [Medline: 34348719]
74. Sepah SC, Jiang L, Peters AL. Long-term outcomes of a web-based diabetes prevention program: 2-year results of a single-arm longitudinal study. J Med Internet Res 2015 Apr 10;17(4):e92. [doi: 10.2196/jmir.4052] [Medline: 25863515]
75. Stein N, Brooks K. A fully automated conversational artificial intelligence for weight loss: longitudinal observational study among overweight and obese adults. JMIR Diabetes 2017 Nov 1;2(2):e28. [doi: 10.2196/diabetes.8590] [Medline: 30291087]
76. Gerber S, Silver RE, Das SK, et al. Development and feasibility of an eHealth diabetes prevention program adapted for older adults—results from a randomized control pilot study. Nutrients 2024 Mar 23;16(7):930. [doi: 10.3390/nu16070930]
77. Ferrara A, McDonald JC, Brown SD, et al. Comparative effectiveness of 2 diabetes prevention lifestyle programs in the workplace: the city and county of San Francisco diabetes prevention trial. Prev Chronic Dis 2020 May 28;17:E38. [doi: 10.5888/pcd17.190396] [Medline: 32463786]
78. Burd C, Gruss S, Albright A, Zina A, Schumacher P, Alley D. Translating knowledge into action to prevent type 2 diabetes: medicare expansion of the national diabetes prevention program lifestyle intervention. Milbank Q 2020 Mar;98(1):172-196. [doi: 10.1111/1468-0009.12443] [Medline: 31994260]
79. Ryan RM, Deci EL. Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness: The Guilford Press; 2017. [doi: 10.1521/978.14625/28806]
80. Patel SB, Iqbal FM, Lam K, Acharya A, Ashrafian H, Darzi A. Characterizing behaviors that influence the implementation of digital-based interventions in health care: systematic review. J Med Internet Res 2025 Jun 12;27:e56711. [doi: 10.2196/56711] [Medline: 40505130]
81. Chester DS, Lynam DR, Milich R, Powell DK, Andersen AH, DeWall CN. How do negative emotions impair self-control? a neural model of negative urgency. Neuroimage 2016 May;132:43-50. [doi: 10.1016/j.neuroimage.2016.02.024]
82. Levin-Zamir D, Bertschi I. Media health literacy, eHealth literacy, and the role of the social environment in context. Int J Environ Res Public Health 2018 Aug 3;15(8):1643. [doi: 10.3390/ijerph15081643]

XSL•FO
RenderX

83. Zanchetta MS, Poureslami IM. Health literacy within the reality of immigrants' culture and language. Can J Public Health 2006;97 Suppl 2:S26-S30. [Medline: 16805158]
84. Liao YS, Tsai WC, Chiu LT, Kung PT. Educational attainment affects the diagnostic time in type 2 diabetes mellitus and the mortality risk of those enrolled in the diabetes pay-for-performance program. Health Policy 2023 Dec;138:104917. [doi: 10.1016/j.healthpol.2023.104917] [Medline: 37776765]
85. Khalifa M, Albadawy M. Artificial intelligence for diabetes: enhancing prevention, diagnosis, and effective management. Comput Methods Programs Biomed Update 2024;5:100141. [doi: 10.1016/j.cmpbup.2024.100141]
86. El Benny M, Kabakian-Khasholian T, El-Jardali F, Bardus M. Application of the eHealth literacy model in digital health interventions: scoping review. J Med Internet Res 2021 Jun 3;23(6):e23473. [doi: 10.2196/23473] [Medline: 34081023]
87. Lee EH, Lee YW, Kang EH, Kang HJ. Relationship between electronic health literacy and self-management in people with type 2 diabetes using a structural equation modeling approach. J Nurs Res 2024 Jan 1;32(1):e315. [doi: 10.1097/jnr.0000000000000588]
88. Lee S. Association between Korean adults' electronic health literacy and active participation in health decision-making. Comput Inform Nurs 2025 Jan 1;43(1):e01175. [doi: 10.1097/CIN.0000000000001175]
89. Cheun J. How eHealth Literacy Impacts Patient-Provider Relationships: A Study on Trust, Self-Care, and Patient Satisfaction [Doctoral]: University of North Texas; 2017. [doi: 10.12794/metadc1011860]
90. Bello CB, Balogun MO, Ogundipe L, Olubiyi SK, Bamigboye TO, Esan DT. Influence of eHealth literacy and health promotion behavior on body mass index of workers in the public sector. SAGE Open Nurs 2024;10:23779608241274253. [doi: 10.1177/23779608241274253] [Medline: 39165911]
91. Sadeghi S, Borzu ZA, Shirzadi S, Jahanfar S, Abadi YG, Taziki M. Investigating e-health literacy and its relationship with self-care behaviors among Iranian middle-aged type 2 diabetes: a cross-sectional study. BMC Public Health 2025 Dec;25(1):1376. [doi: 10.1186/s12889-025-22648-w]
92. Shudayfat T, Hani SB, Shdaifat E, Al-Mugheed K, Alsenany SA, Farghaly Abdelaliem SM. Electronic health literacy and its association with lifestyle behavior among undergraduate students: a cross-sectional survey. Digit Health 2023;9:20552076231185429. [doi: 10.1177/20552076231185429] [Medline: 37434731]

Abbreviations

AI: artificial intelligence
CDC: Centers for Disease Control and Prevention
DPP: Diabetes Prevention Program
eHL: eHealth literacy
HbA_{1c}: glycosylated hemoglobin
HCAI: Artificial Intelligence Health Coaching
MMAT: Mixed Methods Appraisal Tool
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search
RQ: research question
T2D: type 2 diabetes
TDF: Theoretical Domains Framework
WOS: Web of Science

Edited by N Cahill, S Brini; submitted 20.May.2025; peer-reviewed by M Tanhapour, S Ranjbaran, V Mohan; revised version received 11.Jan.2026; accepted 13.Jan.2026; published 10.Mar.2026.

Please cite as:
Li J, Azlan AA, Hassim N, Wang Y, Guo R
eHealth Literacy and Type 2 Diabetes Prevention Among At-Risk Populations: Mechanistic Systematic Review Using Theory-Driven Thematic Analysis
J Med Internet Res 2026;28:e77788
URL: https://www.jmir.org/2026/1/e77788
doi:10.2196/77788

© Jingyi Li, Arina Anis Azlan, Nurzihan Hassim, Yuan Wang, Ruina Guo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is

XSL•FO
RenderX

properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Artificial Intelligence Applications in Medical Devices for Personalized Health Care Solutions: Systematic Review

Hemanth Ponnambalath Mohanadas[1], MS; A Manikandan[2], PhD; Ahmad Fauzi Ismail[3], PhD; Nick Tucker[4], PhD; Saravana Kumar Jaganathan[5,6,7], PhD

[1]Device Development, Abbott Diabetes Care, Alameda, CA, United States

[2]Centre for Material Chemistry, Department of Chemistry, Karpagam Academy of Higher Education, Coimbatore, India

[3]School of Chemical and Energy Engineering, Advanced Membrane Technology Research Centre (AMTEC), University of Technology Malaysia, Skudai, Malaysia

[4]School of Engineering and Physical Sciences, College of Science, University of Lincoln, Lincoln, United Kingdom

[5]Institute of Research and Development, Duy Tan University, 03 Quang Trung, Da Nang, Vietnam

[6]School of Engineering & Technology, Duy Tan University, Da Nang, Vietnam

[7]Biomedical Engineering Research Group, School of Engineering, University of Leicester, Leicester, United Kingdom

Corresponding Author:
Saravana Kumar Jaganathan, PhD
Institute of Research and Development, Duy Tan University, 03 Quang Trung, Da Nang, Vietnam

Abstract

Background: The integration of artificial intelligence (AI) in medical devices is transforming health care by enabling enhanced personalization and precision medicine. AI-driven medical devices can tailor treatments based on individual patient profiles, including genetic data, medical history, and physiological parameters. This advancement holds the potential to refine therapeutic interventions, improve patient outcomes, and streamline health care delivery. However, challenges such as data quality, algorithmic bias, patient privacy, and regulatory complexities hinder the full realization of AI-driven personalization. By 2030, the global AI in health care market is projected to exceed US $187.95 billion, growing at a compound annual growth rate of 37% from US $15.1 billion in 2022.

Objective: This review aims to explore the scope and impact of AI-driven personalization in medical devices. It seeks to analyze key technological innovations that have enabled AI integration, identify the critical challenges impeding progress, and evaluate strategies to address these challenges. Additionally, it highlights future research directions and innovation opportunities in this evolving field.

Methods: A systematic review was conducted, drawing from scholarly literature, industry analyses, and regulatory advisories. Relevant studies and case examples were analyzed to assess the current applications of AI in medical devices, the barriers to its implementation, and best practices for overcoming these barriers. Ethical, technical, and regulatory considerations were also examined. The review included studies published between 2016 and 2023, covering over 100 peer-reviewed articles and reports.

Results: The review highlights significant advancements in AI-driven medical devices, including applications in diagnostics, treatment personalization, wearable health monitoring, and smart prosthetics. AI-based diagnostic tools have achieved up to 98.88% accuracy in multiclass disease classification from X-ray images and 95% accuracy in insulin injection site recognition. It identifies key challenges such as data security risks, algorithmic biases, regulatory constraints, and integration issues with existing health care infrastructures. Currently, more than 70% of clinical decisions rely on diagnostic tests, yet AI-driven automation could reduce diagnostic delays by up to 50%. Several strategies, including improved data validation techniques, regulatory frameworks for AI approval, and ethical guidelines, were found to be effective in mitigating these challenges. Case studies demonstrate how AI has enhanced medical device functionality and patient outcomes.

Conclusions: AI-driven personalization in medical devices holds immense potential to revolutionize health care, offering more precise, adaptive, and patient-centered solutions. However, successful implementation requires addressing technical, ethical, and regulatory challenges. Emerging technologies such as quantum computing could improve AI-driven medical diagnoses by 10‑20 times in processing efficiency, while blockchain-based patient data management could reduce security breaches by more than 30%. This review serves as a valuable resource for researchers, health care professionals, policymakers, and industry leaders, fostering informed discussions and guiding future advancements in AI-enabled personalized medicine.

(J Med Internet Res 2026;28:e72410) doi:10.2196/72410

XSL•FO
RenderX

KEYWORDS

artificial intelligence; personalized medicine; medical devices; digital health; predictive analytics

Introduction

Background

The intersection of artificial intelligence (AI) with medical device technology marks a cornerstone in the evolution of health care, introducing a new epoch where personalization and precision medicine are not just aspirational goals but achievable realities [1]. AI refers to computational systems that perform tasks typically requiring human intelligence, such as learning and decision-making. Personalization in this context is defined as the adaptation of medical devices and interventions to individual patient characteristics, enabled by AI algorithms. AI-driven personalization thus describes the integration of AI into medical devices to dynamically tailor diagnostics, treatments, and monitoring to each patient. This transformative synergy is at the forefront of a health care revolution, offering personalized health care solutions that were once deemed futuristic. Despite these advancements, existing research lacks a comprehensive analysis of how AI-driven personalization is transforming medical devices, particularly in terms of regulatory challenges, clinical adoption, and long-term impact on patient outcomes. While studies have explored AI applications in diagnostics and predictive analytics, a holistic evaluation of AI's role across the entire medical device spectrum, spanning real-time monitoring, adaptive therapeutics, and regulatory compliance, remains underexplored.

Based on the Precedence Research report, the global AI in the health care market was valued at US $15.1 billion in 2022 and is projected to exceed US $187.95 billion by 2030, growing at a compound annual growth rate of 37% during the forecast period. This substantial growth highlights the increasing adoption of AI technologies in health care, driven by their potential to enhance patient care and operational efficiency. AI-driven personalization in medical devices epitomizes a leap toward a future where medical care is uniquely customized to everyone's genetic, environmental, and lifestyle specifics, revolutionizing patient care and treatment outcomes [2]. However, despite rapid AI integration, real-world deployment remains constrained by technical limitations, algorithmic biases, and evolving regulatory frameworks. This underscores the need for a systematic evaluation of AI-driven medical devices, addressing both technological breakthroughs and the challenges hindering large-scale implementation.

For instance, additive manufacturing (3D printing) is being harnessed to create bespoke implants and prosthetics tailored to the unique anatomical features of patients, enhancing the effectiveness and comfort of these medical solutions [3]. While promising, the clinical validation of these AI-driven medical solutions remains a major hurdle, with only a fraction of AI-based medical devices receiving regulatory approval due to concerns over accuracy, interpretability, and patient safety.

AI, at its core, involves the development of algorithms and systems that can perform tasks requiring human intelligence, such as learning, reasoning, problem-solving, and decision-making [4]. In the medical device sector, AI technologies are used to process vast amounts of health data to uncover patterns and insights that inform clinical decisions. This includes analyzing genomic data, electronic health records, biometric readings, and real-time physiological data [5]. Advanced AI algorithms and machine learning (ML) techniques are crucial in creating medical devices that can adapt, predict, and respond to the specific needs of individual patients, offering unprecedented levels of accuracy and efficiency [6]. Medical devices integrated with AI range from diagnostic tools capable of early disease detection to therapeutic devices that adjust treatment protocols in real time [7]. These devices not only improve the precision of diagnostics and treatment but also facilitate continuous monitoring and proactive health care management. Applications of AI in medical devices are diverse, including AI-powered imaging tools for cancer detection [8], smart prosthetics that adjust to physiological changes [9], and wearable technologies for real-time health tracking [10]. The regulation of AI in medical devices is a complex but essential aspect to ensure patient safety and efficacy of the devices [11]. Regulatory bodies such as the US Food and Drug Administration (FDA) and the European Medicines Agency are continuously evolving their frameworks to keep pace with rapid advancements in AI technology. These regulations cover various aspects, such as algorithm transparency, data security, patient privacy, and postmarket surveillance [12]. Despite these efforts, regulatory inconsistencies across regions, coupled with the lack of standardized validation protocols for AI-based medical devices, remain a critical research gap. Understanding these challenges is vital for the ethical and effective deployment of AI-driven medical devices.

The scope of AI-driven personalization in medical devices is broad, encompassing a diverse array of applications that promise to enhance disease diagnosis, treatment efficacy, and patient monitoring. For instance, AI-powered imaging tools are revolutionizing diagnostics by providing deeper insights into medical images, enabling early detection of conditions such as cancer with higher precision than ever before [13]. In the realm of treatment, we are witnessing the emergence of smart prosthetics and implants that adapt to patients' physiological changes and usage patterns, thereby improving functionality and comfort [14]. Furthermore, wearable technologies equipped with AI algorithms are transforming patient monitoring by offering continuous, real-time health tracking and analysis, facilitating early intervention and personalized health management strategies [15].

Despite the great promise of AI in personalizing medical devices, there are many challenges to overcome, including technical, ethical, and regulatory issues. Technical challenges include ensuring the accuracy, reliability, and robustness of AI algorithms, as well as securing high-quality, representative data to train these models effectively [16]. Ethical considerations revolve around issues such as algorithmic bias, which can lead to unequal treatment outcomes, and the imperative to maintain patient privacy and data security in an era of ubiquitous data

XSL•FO
RenderX

collection [17]. Furthermore, navigating the regulatory landscape presents its own set of hurdles, as policymakers and regulatory bodies strive to keep pace with the rapid advancements in AI technology while ensuring patient safety and the efficacy of AI-integrated medical devices [18].

This study aims to bridge these gaps by systematically reviewing the role of AI in medical devices, with a focus on its applications, regulatory landscape, and clinical implementation. The objectives of this review are threefold: (1) to explore and categorize AI-driven advancements in medical devices, including their technological underpinnings and real-world use cases; (2) to identify and critically assess key challenges, technical, ethical, and regulatory, affecting AI adoption in medical devices; and (3) to propose strategic recommendations for overcoming these barriers, facilitating a smoother transition of AI-powered devices from research to clinical practice. Our hypothesis is that while AI-driven personalization has the potential to revolutionize medical device functionality, its widespread adoption remains limited due to unresolved challenges in data validation, regulatory approval, and integration into existing health care infrastructures. By systematically analyzing these aspects, this review aims to provide actionable insights for researchers, policymakers, and industry leaders striving to harness AI for personalized medicine.

Historical Development of AI in Medical Devices

The integration of AI into medical devices has roots that trace back to the mid-20th century, with the advent of computers and early AI research [19]. Initially, these endeavors focused on automating simple tasks and analyzing medical data. A pivotal moment came with the development of expert systems in the 1970s and 1980s, such as MYCIN, a computer-based consultation system designed to assist physicians in diagnosing bacterial infections and recommending antibiotics [20]. These systems laid the groundwork for the incorporation of AI in health care by demonstrating the potential of machine-based decision-making support.

The subsequent decades saw incremental advancements, with AI applications expanding into diagnostic imaging, patient monitoring, and robotic surgery. However, the real explosion of AI in medical devices coincided with the advent of more sophisticated ML algorithms and the exponential growth in computational power and data availability in the 21st century [21]. This era ushered in devices capable of complex analyses and decisions, from predicting patient outcomes to personalizing treatment plans based on individual patient data.

Basic Principles of AI and ML Relevant to Health Care

AI encompasses a broad range of technologies that enable machines to sense, comprehend, act, and learn. ML, a subset of AI, involves the development of algorithms that can learn from and make predictions or decisions based on data. This is particularly relevant in health care, where ML models are trained on vast datasets of medical records, images, and genetic information to identify patterns and anomalies [6].

Among the diverse techniques under the ML umbrella, 2 approaches are notably significant for personalizing medical devices. Supervised learning, with its basis in training on labeled datasets, proves indispensable in diagnostic realms such as imaging, where algorithms learn from vast databases of prediagnosed images to identify diseases. This training allows machines to recognize complex patterns and anomalies in patient data, making precise predictions or decisions that support clinical diagnoses. On the other hand, unsupervised learning ventures into the analysis of unlabeled data, uncovering hidden patterns without predefined outcomes. This approach is particularly valuable in genomic studies and identifying patient subgroups for targeted treatment approaches, where the absence of labeled data does not hinder the algorithm's ability to generate insights. Together, these methodologies underscore a transformative shift toward more personalized, data-informed health care solutions, leveraging the intrinsic patterns within vast datasets to cater to individual patient needs [22].

Further advancement within ML, deep learning, delves into complex neural networks with multiple layers, offering profound insights into the intricate datasets typical in health care, such as detailed medical imaging [23]. For instance, deep learning algorithms are used in radiology to enhance the accuracy of detecting abnormalities in X-rays and computed tomography (CT) scans. These algorithms can identify subtle signs of diseases like lung cancer at stages earlier than human radiologists might detect [24]. Another example is the use of deep learning in pathology, where neural networks analyze tissue samples to detect cancerous cells with high precision, significantly reducing the risk of misdiagnosis [25]. In cardiology, deep learning applications include analyzing electrocardiograms to predict the likelihood of heart conditions such as atrial fibrillation or heart failure [26]. These predictive models can alert health care providers to intervene early, potentially preventing serious cardiac events. This progression underscores the pivotal role of AI and ML in not only advancing medical technology but also in shaping the future of personalized health care, illustrating a journey from data to diagnosis that is increasingly nuanced and tailored to individual patient needs.

The Evolution of Medical Devices From Generic to Personalized Solutions

The journey of medical devices from one-size-fits-all to personalized solutions reflects broader shifts toward precision medicine and patient-centered care [27]. Initially, medical devices were designed for the average patient, with limited ability to adjust to individual needs [28]. However, the recognition that health care efficacy can significantly improve by considering individual differences in genetics, environment, and lifestyle catalyzed the shift toward personalization [29].

Personalized medical devices leverage AI and ML to adapt their functionality to individual patients. For example, wearable devices now monitor health metrics in real-time, adjusting treatment regimens accordingly [30]. Similarly, AI-enabled implants can adjust their behavior based on changes in the patient's physiological data, optimizing therapeutic outcomes [31].

This evolution is not just technological but also philosophical, representing a shift in health care paradigms toward more holistic, tailored care. It underscores the recognition of the complex interplay between various factors affecting health and

XSL•FO
RenderX

the need for treatments that accommodate this complexity [32]. Personalized medical devices, powered by AI, are at the forefront of this shift, offering the promise of health care that is not only more effective but also more attuned to the unique needs and circumstances of each patient.

In summary, the integration of AI into medical devices is a culmination of decades of technological advancement and a reflection of a broader shift toward personalized health care. By leveraging the power of AI and ML, medical devices are evolving to meet the unique needs of individual patients, promising a new era of precision medicine that is more effective, efficient, and patient-centered.

Methods

Overview

To comprehensively review the current state of AI-driven personalization in medical devices, a systematic literature search was conducted across multiple databases, including PubMed, IEEE Xplore, Scopus, Web of Science, Embase, and Google Scholar, ensuring broad coverage of relevant studies. To minimize potential gaps, additional sources such as preprint repositories and conference proceedings were also considered. The search strategy incorporated a diverse range of keywords and synonyms, including “artificial intelligence in medical devices,” “personalized healthcare,” “AI-driven diagnostics,” “machine learning in medical technology,” “smart prosthetics,” “ethical considerations in AI-driven healthcare,” and “regulatory frameworks for AI-integrated medical devices.” Boolean operators (AND, OR) were systematically applied, and medical subject heading terms were used in PubMed to enhance precision, with truncation and wildcard operators used where necessary. The search was limited to peer-reviewed studies published between January 2016 and December 2023 to ensure relevance to contemporary developments. While studies were restricted to English-language publications, a cross-verification process was implemented to identify key non-English papers referenced in systematic reviews and meta-analyses, allowing for their inclusion through reliable academic translation tools. This structured search strategy was designed to maximize coverage, ensure methodological rigor, and minimize potential biases, thereby providing a comprehensive and objective review of AI-driven personalization in medical devices.

Eligibility Criteria

Studies were included if they focused on AI-driven personalization in medical devices, were published in peer-reviewed journals, and provided empirical data, whether quantitative or qualitative. Additionally, studies needed to be available in full text and in the English language. Exclusion criteria encompassed non-English articles, opinion pieces, editorials, letters, studies without accessible full text, and articles unrelated to medical devices or AI. This approach ensured the selection of high-quality and relevant studies for the review.

Study Selection and Agreement Assessment

The study selection process was conducted by the authors, who were divided into 2 independent groups. Each group systematically screened the titles and abstracts of all identified articles based on predefined inclusion and exclusion criteria. Full-text articles were retrieved for studies deemed potentially relevant. After the initial screening, both groups compared their selections, and any discrepancies were resolved through discussion. In cases where disagreements persisted, a third reviewer (SKJ) was consulted to reach a final decision, ensuring a transparent and unbiased selection process. The overall agreement rate before resolution was 90%, reflecting a high level of consistency in study selection. The detailed study selection process is depicted in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram, which illustrates the number of studies identified, screened, excluded, and included in the review. Following study selection, articles were categorized into three key themes: (1) case studies illustrating state-of-the-art AI-integrated medical devices, (2) applications of AI-driven personalization in various medical devices, and (3) challenges associated with AI-driven medical personalization. This thematic categorization enabled a structured synthesis of findings, ensuring clarity in the presentation of results.

Data Extraction

Data extraction was performed using a standardized form developed in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Scoping Reviews) to ensure consistency and methodological rigor. The extracted data included study characteristics (authors, year of publication, and country of origin), AI techniques used (ML models, algorithms, and tools), application in medical devices (type of device and clinical use case), outcomes measured (accuracy, effectiveness, and patient outcomes), and key findings (main results and conclusions). Two authors (HPM and MA) independently extracted data following this predefined protocol to enhance reliability. Any discrepancies were resolved through discussion, and if necessary, by consulting additional authors to achieve consensus. The extracted data were then synthesized according to the predefined thematic categories to align with the structure of the Results and Discussion sections. This structured approach ensured a transparent, reproducible, and comprehensive data collection process.

Reference Screening

In addition to the database searches, the reference lists of all included articles were manually screened to identify further relevant studies. This snowballing technique ensured comprehensive coverage of the literature.

Results

Overview

The results of this review paper are categorized into three major sections: (1) case studies of state-of-the-art AI-integrated medical devices, (2) applications of AI-driven personalization in medical devices, and (3) challenges in AI-driven personalization (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram illustrating the systematic review process for artificial intelligence–driven personalization in medical devices. AI: artificial intelligence.

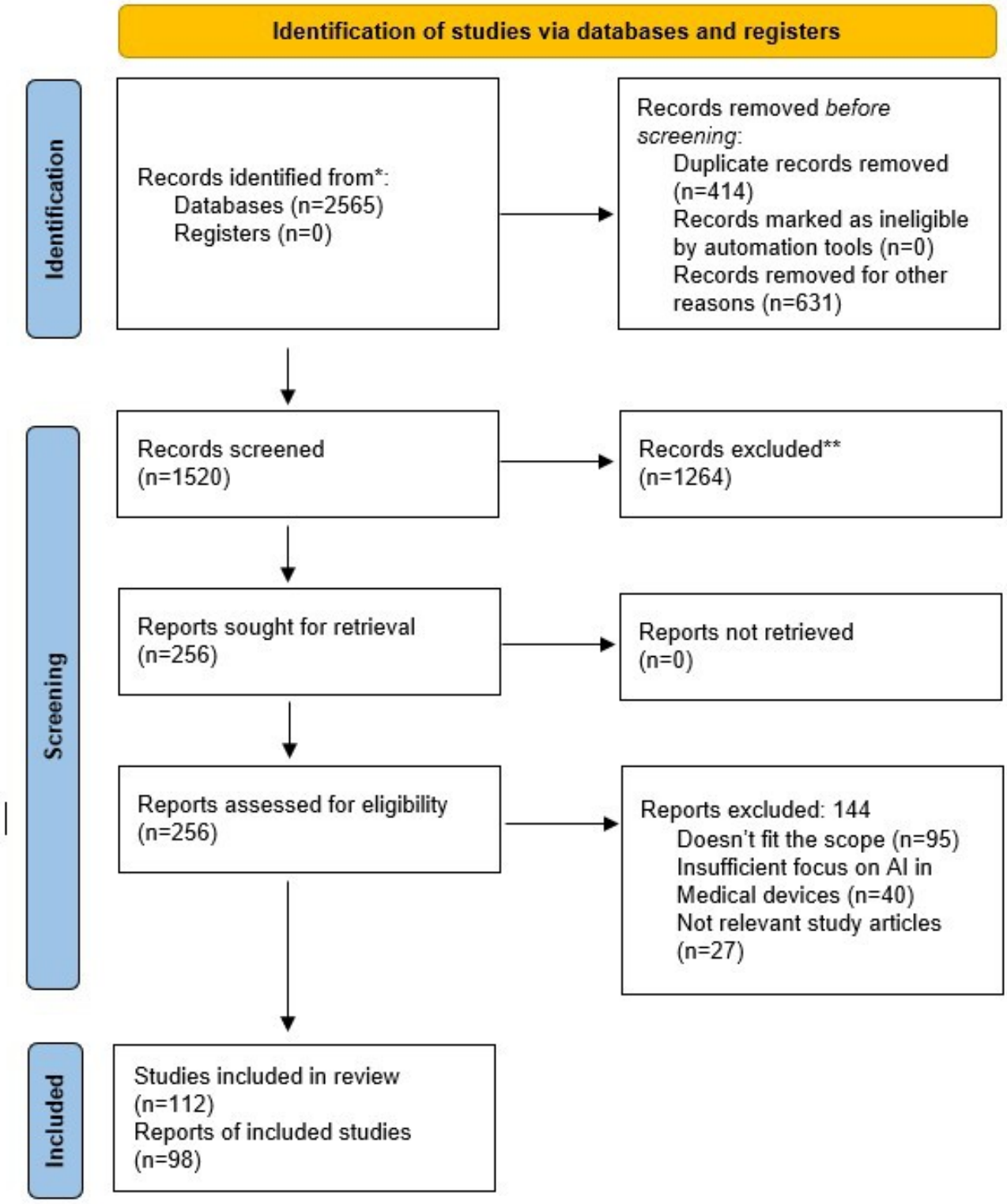

Case Studies of State-of-the-Art AI-Integrated Medical Devices

Smart Insulin Pumps

Smart insulin pumps represent a transformative advancement in diabetes management, addressing many challenges associated with traditional methods that often require frequent manual adjustments [33]. These sophisticated devices use advanced ML algorithms to process real-time data from continuous glucose monitors (CGMs). By analyzing trends and fluctuations in glucose levels, the pumps autonomously adjust insulin delivery, significantly enhancing glycemic control [34]. Adding to this innovation, recent research has successfully applied AI in smart insulin pens, achieving more than 95% accuracy in both detecting proper injection sites and identifying lipodystrophies using a sensor with dual LEDs, thereby

XSL•FO
RenderX

promoting better injection practices and reducing tissue complications [35]. Furthermore, another pivotal development includes a smart AI-based self-care device that predicts future blood glucose fluctuations to proactively manage type-1 diabetes, significantly reducing the risk of hypoglycemia and hyperglycemia by intelligently controlling insulin dosage during automatic injections [36]. The first FDA-approved smart insulin pen in 2014, which records and wirelessly transmits insulin dosing data, exemplifies these technological advances, providing clinical data to assist individualized diabetes management and overcoming barriers such as poor adherence and clinical inertia [37]. This dynamic approach not only alleviates the daily burden on patients but also improves their quality of life by minimizing the risk of glucose variability-related complications and enhancing adherence to injection therapy. Studies have shown that such automated systems can lead to better long-term health outcomes, demonstrating a critical evolution in diabetes care technology [38].

AI-Powered Prosthetics

AI-powered prosthetic limbs are revolutionizing the field of prosthetics, providing amputees with unprecedented levels of functionality and adaptability. These sophisticated devices integrate sensors and advanced AI algorithms that interpret the user's muscle signals and movements, allowing the limb to respond in a natural and intuitive manner. As these AI systems interact with the user, they continuously learn and adapt, fine-tuning their responses to better align with individual needs. This technology not only captures and interprets complex muscle signals but also adapts to the user's unique physiological changes over time, enhancing the effectiveness of rehabilitative practices, as demonstrated in recent studies on surface electromyographic-based systems for hand rehabilitation using AI-powered gesture recognition [39]. Additionally, innovations in lower limb assistive technologies, such as adaptable central pattern generators and personalized motion planning, are enhancing capabilities in gait analysis and stability, crucial for preventing falls and ensuring safe mobility in various environments [40]. Furthermore, AI applications extend beyond mobility aids to the assessment of medical conditions, where tools have significantly improved the evaluation of lower limb radiographs before and after knee arthroplasty, speeding up assessments while maintaining high accuracy [41]. This broad spectrum of AI applications illustrates a transformative impact on patient care, making daily activities and medical diagnostics more efficient and tailored to individual health needs.

Diagnostic and Imaging Tools

AI-integrated diagnostic tools have made significant strides, particularly in the field of medical imaging, where deep learning algorithms excel at detecting subtle patterns indicative of diseases like cancer, often outperforming human radiologists in accuracy. These sophisticated tools adapt their analyses based on individual patient data, greatly enhancing diagnostic precision and enabling timely interventions. For instance, deep learning's integration within imaging modalities has notably advanced pancreatic cancer diagnostics, leading to more precise detection and improved clinical decision-making [42]. Additionally, the application of AI in optical imaging techniques, such as optical coherence tomography and photoacoustic imaging, facilitates high-contrast, low-cost visualization of tumor tissues, thereby enhancing both the detection and theranostics of cancer, which is paving the way for precision oncology [43]. Moreover, the development of a novel deep learning framework for multiclass diagnosis of lung diseases from chest X-ray images shows significant improvements in accuracy and efficiency, exemplifying how AI can transform pulmonary diagnostics and thereby ensure better patient outcomes through enhanced feature extraction capabilities [44]. These advancements underscore the transformative impact of AI in medical diagnostics, revolutionizing early disease detection and treatment strategies, and thereby playing a crucial role in enhancing personalized and effective health care solutions.

AI-Powered Humanoids

AI-powered humanoids are revolutionizing various sectors by merging advanced robotics with deep learning technologies to perform complex tasks. These humanoids, equipped with AI, are capable of high-level functions such as speech recognition, comprehensive comprehension, and sophisticated interactions, making them invaluable in industries like health care, education, and space exploration. For instance, AI-based 3D-printed humanoids represent a significant leap forward, offering customizable interactions and decision-making capabilities that mimic human actions [45]. Similarly, AI-powered smart glasses enhance human-robot interaction by allowing dynamic adaptation to varied walking terrains, demonstrating the potential of embedded devices in everyday environments [46]. Furthermore, discussions about responsibly regulating AI to prevent ethical dilemmas and enhance user safety emphasize the need for knowledge about AI's operational boundaries [47]. In customer service, AI robots are researched for their emotional and functional impact on customer engagement, showcasing their potential to deliver meaningful and personalized service experiences [48]. Moreover, in health care, AI-powered avatars are being evaluated for their acceptance by patients, particularly in mental health services where they can act as assistants to physicians in resource-limited settings [49]. Emotional intelligence in robots, allowing them to recognize and respond to human emotions, is another frontier being explored to make interactions with humanoids more natural and effective [50]. Collectively, these advancements are not just enhancing the capabilities of AI humanoids but are also setting the stage for future innovations that could redefine human-machine collaboration.

Other AI-Powered Applications

AI-powered applications are expanding their reach and transforming industries with innovative and practical solutions. In dentistry, the integration of AI and augmented reality is revolutionizing both practice and education, offering new possibilities for enhanced clinical precision and interactive learning environments [51]. Similarly, the application of AI in the fabrication of 3D-printed materials is propelling the creation of patient-specific wearable devices and smart biomedical implants, streamlining manufacturing processes and advancing personalized medicine [52]. In the field of orthopedics, AI algorithms are being used to standardize the measurement of

XSL•FO
RenderX

implant alignments in radiography, improving the reliability and efficiency of postoperative assessments [53]. The introduction of AI-powered chatbots in medical consultations offers immediate, personalized medical advice, drastically enhancing patient engagement and health care delivery [54]. Moreover, AI is pivotal in field operations, where it aids in the digitalization and enhancement of inspection rounds, significantly optimizing process monitoring and anomaly detection in industrial settings [55]. Additionally, the development of AI-powered neural implants is opening new frontiers in the treatment of conditions like Alzheimer disease, enhancing communication capabilities and patient care through advanced brain-machine interfaces [56]. Each of these applications mentioned in Table 1 not only highlights the diverse capabilities of AI across various sectors but also underscores the significant impact of intelligent technology in improving operational efficiency, accuracy, and personalized solutions.

XSL•FO
RenderX

Table . Case studies: artificial intelligence applications in personalized health care devices.

Case studies	Device type	Specific AI[a] technology	Applications	Outcomes
Exploiting AI to make insulin pens smart: injection site recognition and lipodystrophy detection [35]	Smart insulin pen	Machine learning (one-class SVM[b])	Diabetes management	Achieved more than 95% accuracy in injection site recognition and lipodystrophy detection, promoting better adherence to injection therapy and improving patient quality of life
Smart AI-based self-care device for patients with diabetes [36]	Insulin pump	AI-based predictive algorithms	Managing type-1 diabetes	Intelligently controls insulin dosage during automatic injection, predicts future blood glucose levels, monitors and reports critical conditions, sends summarized reports to health care centers, maintains a timeline of data and insulin dosages
Empowering hand rehabilitation with AI-powered gesture recognition: a study of an sEMG[c]-based system [39]	Wearable rehabilitation glove	Deep learning (1D-CNN[d], InceptionTime [an advanced deep learning architecture designed for time series classification])	Hand rehabilitation	Enhanced hand gesture recognition with 90.89% accuracy, aiding in the rehabilitation of stroke patients through mirror therapy and task-oriented therapy. Improved motor function and usability in both clinical and home settings.
The role and efficiency of an AI-powered software in the evaluation of lower limb radiographs before and after total knee arthroplasty [41]	Radiograph analysis tool	AI-powered software (LAMA, U-net-based convolutional neural network)	Evaluation of lower limb radiographs to assess alignment and angles before and after total knee arthroplasty	Good to excellent agreement with expert assessments (ICC[e]=0.78 - 1.00), significant time efficiency (twice as fast as experts), but struggled with certain measurements (eg, joint line convergence angle, Mikulicz line, and patients with high BMI)
AI-driven diagnosis of pancreatic cancer [42]	Diagnostic imaging	Deep learning and ML[f]	Early detection and accurate diagnosis of pancreatic cancer	Improved accuracy in diagnosis, early detection, better clinical decision-making, and reduced mortality risk
An advanced deep learning framework for multiclass diagnosis from chest x-ray images [44]	Diagnostic tool	Deep learning (CNN[g])	Multiclass diagnosis of lung diseases (fibrosis, opacity, tuberculosis, normal, viral pneumonia, and COVID-19 pneumonia)	Achieved superior performance with an accuracy of 98.88%, precision of 0.9870, recall of 0.9904, F_1-score of 0.9887, and AUC of 0.9939
AI enables reliable and standardized measurements of implant alignment in long leg radiographs with total knee arthroplasties [53]	Diagnostic tool	AI algorithm for image analysis	Evaluation of long leg radiographs (LLR) post-total knee arthroplasties (TKA)	The AI algorithm showed excellent reliability (ICC>0.97) and reproducibility (96 % reproducible, 92.1 % reliable) in measuring hip–knee–ankle (HKA), femoral component (FCA), and tibial component (TCA) angles.
Deep neural networks improve radiologists' performance in breast cancer screening [57]	Diagnostic tool	Deep CNN	Breast cancer screening exam classification	Achieved an AUC[h] of 0.895; model as accurate as experienced radiologists; hybrid model (AI+ radiologist) more accurate than either alone

XSL•FO
RenderX

Case studies	Device type	Specific AI[a] technology	Applications	Outcomes
Deep learning for the classification of small (≤2 cm) pulmonary nodules on CT[i] imaging: a preliminary study [58]	Diagnostic tool (CT imaging)	Deep learning (CT-lungNET, a deep learning-based malignancy prediction model)	Classification of small pulmonary nodules on CT imaging	Improved AUROC[j] (0.85 vs 0.82 for AlexNET), faster processing time, and enhanced diagnostic performance for nonradiologists
AI for cardiac diseases diagnosis and prediction using ECG[k] images on embedded systems [59]	Diagnostic tool	Deep learning algorithms (MobileNetV2 and VGG16)	Analysis and classification of ECG signals to predict cardiovascular diseases	High validation accuracy of 0.95 for MobileNetV2 and VGG16; after implementation on Raspberry Pi, accuracy slightly decreased to 0.94 and 0.90, respectively. Capable of real-time monitoring using smart mobile tools, such as mobile phones, smartwatches, and connected T-shirts.
Personalized diabetes monitoring platform leveraging IoMT[l] and AI for noninvasive estimation [60]	Wearable device	Deep neural network (GlucoNet)	Noninvasive real-time blood glucose estimation	The AI-powered platform demonstrated an MAPE[m] of 17.8 % with 100 % of predictions falling in the clinically acceptable zones A and B of the Clarke-error grid. Clinically validated on over 350 patients, capable of handling large datasets, and does not require frequent calibration.
A predictive machine learning tool for asthma exacerbations: results from a 12-week, open-label study using an electronic multidose dry powder inhaler with integrated sensors [61]	Electronic Multi-Dose Dry Powder Inhaler (eMDPI)	Machine learning (gradient-boosting trees)	Predicting impending asthma exacerbations	The AI model predicted asthma exacerbations with an ROC[n] AUC of 0.83. The model was based on inhaler use data and inhalation parameters, providing a proactive approach to asthma management by identifying exacerbations up to 5 days in advance.

[a]AI: artificial intelligence.
[b]SVM: support vector machine.
[c]sEMG: surface electromyographic.
[d]1D-CNN: One-Dimensional Convolutional Neural Network.
[e]ICC: intraclass correlation coefficient.
[f]ML: machine learning.
[g]CNN: convolutional neural network.
[h]AUC: area under the curve.
[i]CT: computed tomography.
[j]AUROC: area under the receiver operating characteristic.
[k]ECG: electrocardiogram.
[l]IoMT: Internet of Medical Things.
[m]MAPE: mean absolute percentage error.
[n]ROC: receiver operating characteristic.

Applications of AI-Driven Personalization in Medical Devices

Overview

The integration of AI into medical devices has opened a new frontier in health care, enabling personalized medicine that promises more accurate diagnoses, tailored treatments, and continuous patient monitoring. This section delves into the specific applications of AI-driven personalization across various categories of medical devices, highlighting the transformative impact on patient care and health management.

XSL•FO
RenderX

Diagnostic Devices: Improvements in Early Detection and Accuracy

About 70% of clinical decisions rely on diagnostic tests, which play a key role in evidence-based patient care. However, the current method of sending samples to centralized labs often causes delays of several days in getting results. This delay can limit the usefulness of diagnostics and increase patient stress while waiting for important test results [62]. AI-driven diagnostic devices mark a transformative advance in this realm of medical diagnostics, significantly enhancing early detection and improving accuracy. These devices leverage sophisticated algorithms to sift through and analyze complex medical data, including imaging scans [63], blood tests [64], and genetic information [65]. This capability allows them to detect patterns and anomalies that might be overlooked by human evaluators.

In the field of medical imaging, AI algorithms have demonstrated exceptional proficiency. They interpret data from X-rays, magnetic resonance imaging, and CT scans more quickly and accurately than traditional methods. These algorithms excel at detecting subtle abnormalities that are often indiscernible to the human eye, thus facilitating earlier and more accurate diagnoses of conditions such as cancer, cardiovascular diseases, and neurological disorders [66]. The table below (Table 2) summarizes the distribution of FDA-approved diagnostic devices by clinical specialty, underscoring the dominant role of radiology in this field. According to the FDA article released on December 20, 2024, a total of 1016 diagnostic devices have been approved, with an overwhelming 76% (777/1016) categorized under radiology. This statistic highlights the transformative advancements in AI-powered radiology tools, which have played a pivotal role in improving early detection, especially in imaging modalities like X-rays, CT scans, and magnetic resonance imaging. The remaining categories include devices for genetic testing, blood diagnostics, and other fields, reflecting the breadth of AI integration across diverse diagnostic platforms.

Table . Distribution of US Food and Drug Administration–approved artificial intelligence or machine learning–enabled medical devices across medical specialties [67].

Categories	Number of devices (N=1016), n (%)
Radiology	777 (76)
Cardiovascular	104 (10)
Neurology	42 (4)
Other categories	27 (3)
Anesthesiology	17 (2)
Hematology	17 (2)
Clinical chemistry	9 (1)
Gastroenterology – urology	14 (1)
Ophthalmic	9 (1)

AI-powered diagnostic tools are reshaping medical imaging by surpassing human radiologists in detecting subtle disease markers, particularly in cancer diagnostics. Deep learning models, tailored to patient-specific data, have improved the early detection of diseases like pancreatic cancer and enhanced clinical decision-making [42]. AI also advances optical imaging techniques, such as optical coherence tomography and photoacoustic imaging, by offering affordable, high-contrast tumor visualization, which supports precision oncology [43]. Additionally, AI frameworks for multiclass lung disease diagnosis via chest X-rays have boosted both accuracy and efficiency, transforming pulmonary care [44].

For example, a study comparing the diagnostic performance of 14 radiologists in breast cancer screening with a deep learning model illustrates the benefits of deep convolutional neural networks (CNNs). This study demonstrates that CNNs can significantly enhance radiologists' performance, achieving an area under the curve of 0.895. Area under the curve is a statistical measure used to gauge the performance of binary classification models, representing the likelihood that the model correctly distinguishes between positive and negative cases. The study also highlights the successful integration of CNNs with radiologists' evaluations, leading to improved diagnostic outcomes [57]. Similarly, in lung cancer, AI tools applied to low-dose CT scans can identify small pulmonary nodules that may be early indicators of cancer. The precision of AI in these applications not only improves the outcomes for patients through earlier treatment but also reduces the number of false positives and unnecessary biopsies, thereby optimizing health care resources [58]. In ophthalmology, AI-driven tools are used to diagnose diseases like diabetic retinopathy and glaucoma by analyzing retinal images with high accuracy. Various AI systems, including IDx-DR (Digital Diagnostics Inc.), a tool used to diagnose diabetic retinopathy, and EyeArt (Eyenuk Inc.), a system that autonomously analyzes patients' retinal images, which have been approved for clinical use, have demonstrated high sensitivity and specificity in identifying diabetic retinopathy through color fundus photography [68]. These models are trained on large datasets of retinal images to identify specific patterns, such as microaneurysms [69], hemorrhages [70], and exudates [71], which are indicative of diabetic retinopathy. The AI algorithms can assess the severity of retinopathy by evaluating changes in the retina's vascular structure, offering a noninvasive method to monitor the progression of diabetes. Similarly, AI applications in

XSL•FO
RenderX

pulmonology interpret chest X-rays and CT scans to identify signs of lung cancer at early stages [72]. During the COVID-19 pandemic, the integration of AI-driven diagnostic tools significantly enhanced the detection and management of the virus using imaging and blood tests, as highlighted by several studies. For instance, AI algorithms applied to chest X-rays and CT scans swiftly improved the diagnostic accuracy for COVID-19, enabling better prediction of disease severity and aiding in patient management [73]. In parallel, simultaneously, innovative AI applications analyzing changes in blood parameters demonstrated the potential of routine blood tests in COVID-19 diagnosis, achieving high classification performance and offering a rapid, cost-effective alternative to standard reverse transcription polymerase chain reaction [74,75]. These advancements underline the transformative impact of AI in health care, providing crucial support in diagnosing and managing COVID-19 effectively.

AI algorithms are making significant inroads in the field of genetic testing as well. They analyze genetic data to predict the risk of genetic disorders and conditions swiftly and with high precision. A notable example is the AI-based clinical decision support tool, Fabric GEM (Fabric Genomics, Inc.), which enhances the diagnosis of rare genetic diseases by effectively prioritizing genetic variants and integrating deep phenotyping through clinical natural language processing. This tool substantially improves the diagnostic accuracy and efficiency for rare genetic disorders, as demonstrated in a retrospective cohort study [76]. Similarly, another study reviews AI and ML approaches using gene expression and variant data to advance personalized medicine, highlighting the potential of these technologies to enhance disease diagnosis, risk assessment, and treatment outcomes across diverse populations and a range of diseases, including cancer, cardiovascular, and neurodegenerative diseases [77]. Furthermore, a separate research effort using supervised ML showcases how models such as support vector machines and K-nearest neighbors can be applied to multifactorial genetic inheritance disorders to predict conditions such as cancer, dementia, and diabetes with high accuracy. Support vector machine is a classification technique that finds the hyperplane that maximizes the margin between different classes in the data, while K-nearest neighbors classifies data based on the majority class among the k-nearest samples, providing a straightforward but powerful approach for predictive modeling [78]. Despite these advancements, the deployment of AI in diagnostic devices comes with challenges such as ensuring data diversity to prevent bias, enhancing the interpretability of AI decisions for clinical use, and navigating complex regulatory and ethical considerations [79,80].

Therapeutic Devices: Dynamic Treatment Adaptations and Real-Time Monitoring

AI-enabled therapeutic devices are revolutionizing patient care by adapting treatments in real-time, based on continuous monitoring and dynamic feedback. This capability leads to more effective and personalized treatment plans, enhancing overall patient outcomes [30]. By leveraging real-time data analysis, these devices can respond immediately to changes in a patient's condition, ensuring that treatments are both timely and precisely tailored to individual needs.

One key application of AI in therapeutic devices is in the management of diabetes through AI-driven real-time insulin pumps, which use continuous monitoring and feedback to deliver personalized and effective treatment plans. This approach is underscored by recent research that introduces a closed-loop electronic system using a proportional-integral-derivative controller and WiFi-controlled voltage to automatically adjust insulin levels based on glucose sensor readings, enhancing diabetes management [81]. These types of dynamic management help maintain optimal blood sugar levels, significantly reducing the risk of complications associated with diabetes. Research developed for an AI expert system based on CGM data that not only automates insulin delivery but also facilitates forward-thinking and collaborative health management between sensors and patients [82].

Another significant advancement is in cardiac care, where AI-powered devices such as pacemakers and defibrillators adjust their functioning based on the patient's activity level and physiological signals [83]. This ensures that the heart operates as efficiently as possible, providing personalized care that adapts to the daily life and needs of the patient, thereby enhancing treatment efficacy and patient safety. Complementing these innovations, a recent study introduces a neural network capable of identifying the manufacturer and model group of cardiac rhythm devices from chest radiographs with exceptional accuracy. This tool, which achieves 99.6% accuracy for manufacturer identification and 96.4% for model groups, outperforms traditional methods used by cardiologists, offering a quicker and more precise approach to managing patients with implanted cardiac devices [84]. Furthermore, another pivotal study uses CNNs to predict the risk of pacemaker-induced cardiomyopathy using clinical data available prior to pacemaker implantation. Achieving an accuracy rate of 75.8%, this model highlights the potential for AI to proactively manage patients' health, potentially preventing the onset of cardiomyopathy post-implantation [85].

Expanding on this, AI technologies are being increasingly integrated into various cardiac monitoring tools, leveraging advanced algorithms and wireless connectivity to deliver expert-level diagnostics and continuous heart health monitoring. For example, continuous electrocardiogram recordings help redefine arrhythmia phenotypes and predict the success of antiarrhythmic therapies, while wearable activity trackers measure vital physiological indices that can detect irregularities such as atrial fibrillation. Additionally, smartphone apps contribute to reducing response times for sudden cardiac events, showcasing the integration of novel monitoring devices with ML for scalable cardiovascular management. These advancements in cardiovascular monitoring technologies not only promise to improve personalized care but also pose new challenges and opportunities for the future of cardiac health [86].

These AI advancements in cardiac care not only enhance diagnostic and predictive capabilities but also significantly improve the personalized management and safety of patients requiring cardiac rhythm management. Together, these applications of AI in therapeutic devices demonstrate a shift toward more adaptive, responsive medical treatments that

XSL•FO
RenderX

promise to improve patient outcomes by integrating sophisticated technology with real-time health data [87]. Supporting this, a study develops predictive models based on mobile clinical data for patients with heart failure, significantly reducing false alerts and improving the management of patient decompensation through telemonitoring systems [88]. These advances support the ongoing development and refinement of therapeutic devices that not only anticipate health needs but also actively engage with them, paving the way for highly personalized and proactive treatment strategies.

Wearable Technology: Continuous Health Tracking and Disease Management

AI-enabled wearable technologies are revolutionizing health management by providing continuous tracking and dynamic feedback, allowing for unprecedented personalization in medical care. Equipped with advanced AI algorithms, these devices analyze data from everyday activities to offer personalized health insights and recommendations, enhancing the effectiveness and timeliness of interventions [89]. This capability helps users manage their health more proactively and enables early detection of potential health issues. Recent advancements in wearable health technology have significantly improved the performance of wearable sensors, enabling real-time health monitoring and early disease detection. Additionally, for chronic conditions like hypertension and heart disease, wearable technologies equipped with AI can predict exacerbations, enabling preemptive medical interventions that tailor treatments to the nuances of an individual's physiological data [90]. This advanced approach not only optimizes patient care but also integrates seamlessly with the broader goals of precision medicine, underscoring the profound impact of AI in enhancing daily health management.

Wearable devices like fitness trackers and smartwatches have become essential tools in everyday health management, capturing a wide array of data such as heart rate [91], sleep patterns [92], and physical activity levels [93]. These devices leverage sophisticated AI algorithms to analyze this wealth of information, providing personalized insights and actionable recommendations. This functionality not only supports users in maintaining or improving their lifestyle but also plays a crucial role in the early detection of potential health issues [10]. Notably, changes in heart rate variability and sleep quality, as measured by these wearables, may indicate stress or cardiovascular issues. A study found that heart rate variability, tracked through wearable devices, effectively predicts the effectiveness of therapeutic interventions for anxiety and depression, underscoring its value in preemptive health measures [94]. This proactive approach to health monitoring highlights the significant role that wearable technology plays in preventive health care, empowering users to manage their well-being with precision and foresight.

Figure 2 illustrates various innovative CGM biosensors designed for CGM using different platforms such as interstitial fluid, tears, saliva, and sweat. These wearable devices represent the cutting edge of personalized health monitoring technologies, integrating advanced sensor systems and AI to provide real-time health data, crucial for managing chronic conditions like diabetes [95].

XSL•FO
RenderX

Figure 2. Representative examples of wearable continuous glucose monitoring biosensors [95]. (**A**) An electrochemical nonenzymatic continuous glucose monitoring system. Reproduced from Yoon et al [96] (**B**) A fluorescence-based continuous glucose monitoring biosensor. Reproduced from Sawayama and Takeuchi [97]. (**C**) Commercialized Senseonics continuous glucose monitoring sensor. (**D**) A contact lens biosensor for real-time glucose monitoring. Reproduced from Moreddu et al [98] (**E**) A wearable mouthguard continuous glucose monitoring biosensor. Reproduced from Arakawa et al [99] (**F**) A smart wristband for monitoring glucose in sweat. Reproduced from Gao et al [100]. Full figure reproduced from Jin et al. [95], licensed under CC BY 4.0. CGM: continuous glucose monitoring; GF-PEG: glucose-responsive fluorescent polyethylene glycol; ISF: interstitial fluid.

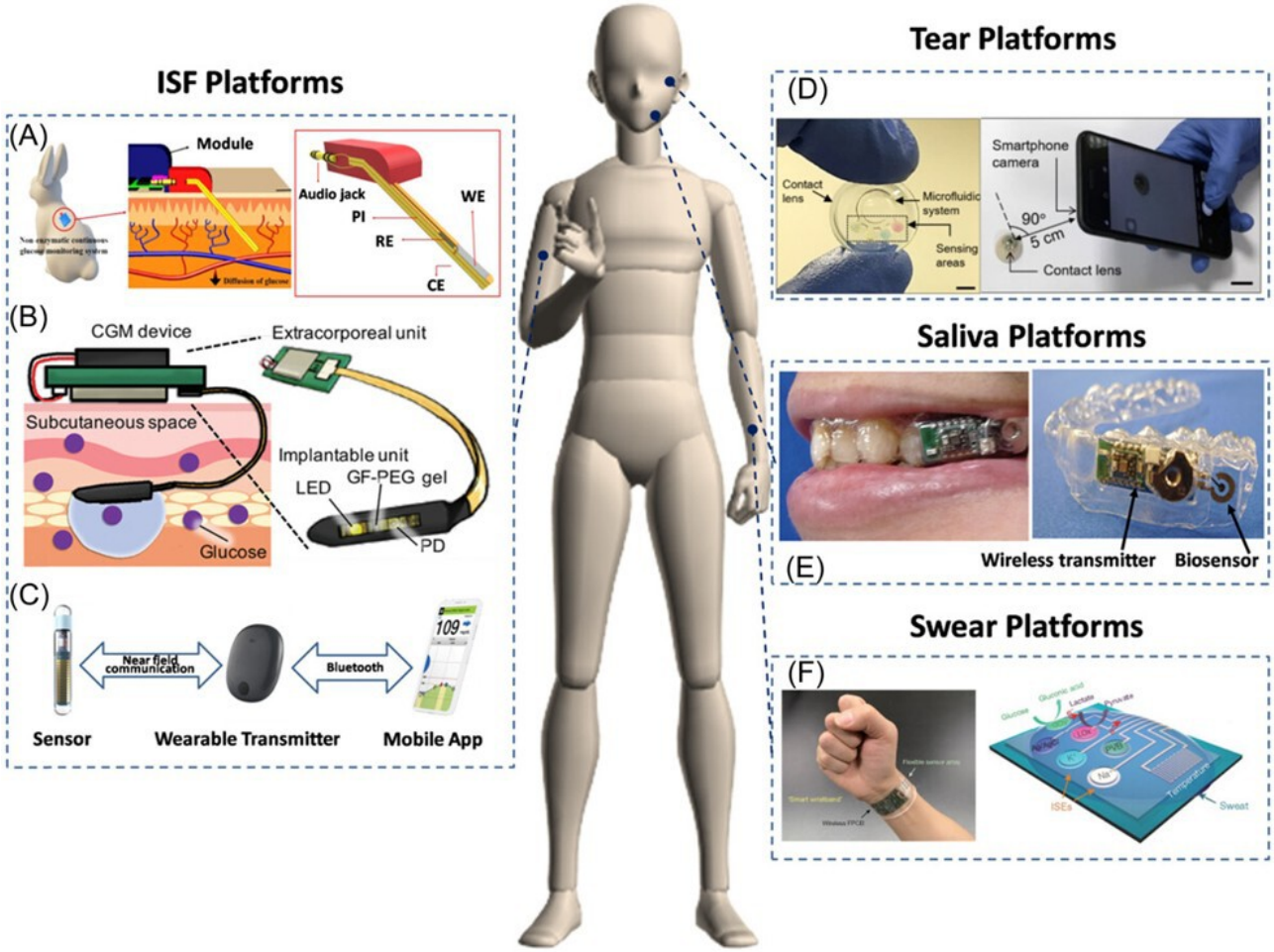

Personalized eHealth programs have demonstrated significant lifestyle improvements among patients with cardiac issues through tailored interventions based on data captured by wearable devices. For instance, the "Do Cardiac Health Advanced New Generation Ecosystem (Do CHANGE 2)" program effectively demonstrates the impact of personalized eHealth interventions, which use wearable technology to promote lifestyle changes and improve the quality of life among patients with cardiac issues [101]. This intervention, documented to significantly enhance lifestyle behaviors through continuous health behavior assessments and personalized feedback, underscores the potential of wearable devices in facilitating substantial health improvements and personalized care. Furthermore, the AI4FoodDB project has innovatively used wearable devices to gather and analyze dietary and lifestyle data, offering groundbreaking personalization in nutritional advice, which significantly impacts lifestyle improvements [102]. Similarly, personalized activity eCoaching, enhanced by AI and semantic ontology, exemplifies how technological advancements can foster more accurate health monitoring and lifestyle recommendations, effectively encouraging healthier behavior patterns [103]. Additionally, real-time models to predict stress-induced blood pressure spikes using wearable devices underline the potential of AI-driven technologies to proactively manage and mitigate health risks [104]. These examples collectively underscore the profound influence of AI-enhanced wearable technology in shaping future health care paradigms, focusing on personalized care and proactive health management strategies.

In the realm of chronic disease management, wearable technologies equipped with AI are particularly transformative. A systematic review of studies involving patients with various chronic conditions demonstrated that wearable devices could facilitate active self-management and real-time health monitoring, thereby playing a crucial role in managing chronic diseases. This data can then be leveraged through AI algorithms to provide actionable insights, which may aid in the early detection and management of potential complications associated with chronic diseases [105]. For instance, wearables have been shown to impact health positively by aiding in the management of diseases like diabetes, where they can monitor and provide feedback on blood sugar levels, or in cardiovascular diseases, where they assist in monitoring heart function [106]. Devices designed for monitoring hypertension and heart disease continuously track vital signs like blood pressure and heart

XSL•FO
RenderX

rhythm, transmitting this data in real time to health care providers [107]. AI plays a critical role in this process by analyzing trends and variations in the data to predict potential exacerbations. This proactive approach allows health care providers to offer preemptive medical interventions, potentially reducing hospital visits and improving patient outcomes [108].

This evolution in wearable technology emphasizes a shift from reactive to proactive and personalized medical care. The continuous monitoring facilitated by these devices, combined with AI-driven analysis, ensures that each patient receives care that is tailored not just to a condition but to their unique physiological makeup and lifestyle. This not only enhances patient engagement by providing them with real-time, actionable health insights but also significantly improves the efficacy of interventions, aligning perfectly with the goals of precision medicine [109]. As these technologies advance, they promise to further refine our approach to health management, making preventative care more accessible and effective than ever before.

Prosthetics and Implants: Customization and Adaptive Functionalities

The advent of AI-driven prosthetics and implants marks a significant leap forward in the field of medical devices, enhancing the integration of technology and human physiology. These advanced devices harness AI to provide functionalities that closely emulate natural human movements, offering a more intuitive and seamless experience for users [110]. This technological evolution not only improves the quality of life for individuals but also pushes the boundaries of what medical devices can achieve in terms of personalization and performance.

In the realm of smart prosthetics, AI plays a critical role in enabling these devices to learn from and adapt to the user's specific movement patterns. This capability allows for fluid and natural motions, such as a prosthetic arm that adjusts its grip based on the object it is interacting with. This is achieved through advanced sensors and actuators that continually collect and analyze data, allowing the prosthetic to make real-time adjustments. For example, shared control systems in assistive robots predict user intent and react accordingly, enhancing the functionality of the prosthetic while maintaining a balance between user control and automated responses [111]. Similarly, demonstration units for upper-limb prostheses provide valuable insights during the calibration and learning phases, further personalizing the user experience [112]. A recent study developed an automated pipeline using ML methods to design custom total knee replacement implants from CT scans. This pipeline demonstrated high accuracy and repeatability, significantly reducing the time and cost associated with traditional manual processes, highlighting the potential for broader applications in custom medical implants (Figure 3, which illustrates the workflow of the CT-3D surface model prediction process for custom total knee replacement design) [113].

Figure 3. Workflow of computed tomography-3D surface model prediction process for custom total knee replacement design: (**A**) Load digital imaging and communications in medicine, (**B**) classify slices, (**C**) identify region of interest, (**D**) crop slices to region of interest, (**E**) segment femur and tibia bones, (**F**) create contour stacks, (**G**) align contour stacks with statistical shape model base shapes, (**H**) morph statistical shape models to fit contour points and create 3D model predictions [113]. Full figure reproduced from Burge et al. [95,113], licensed under CC BY 4.0. CT: computed tomography; DICOM: digital imaging and communications in medicine; ICP: iterative closest point; ROI: region of interest; SSM: statistical shape model.

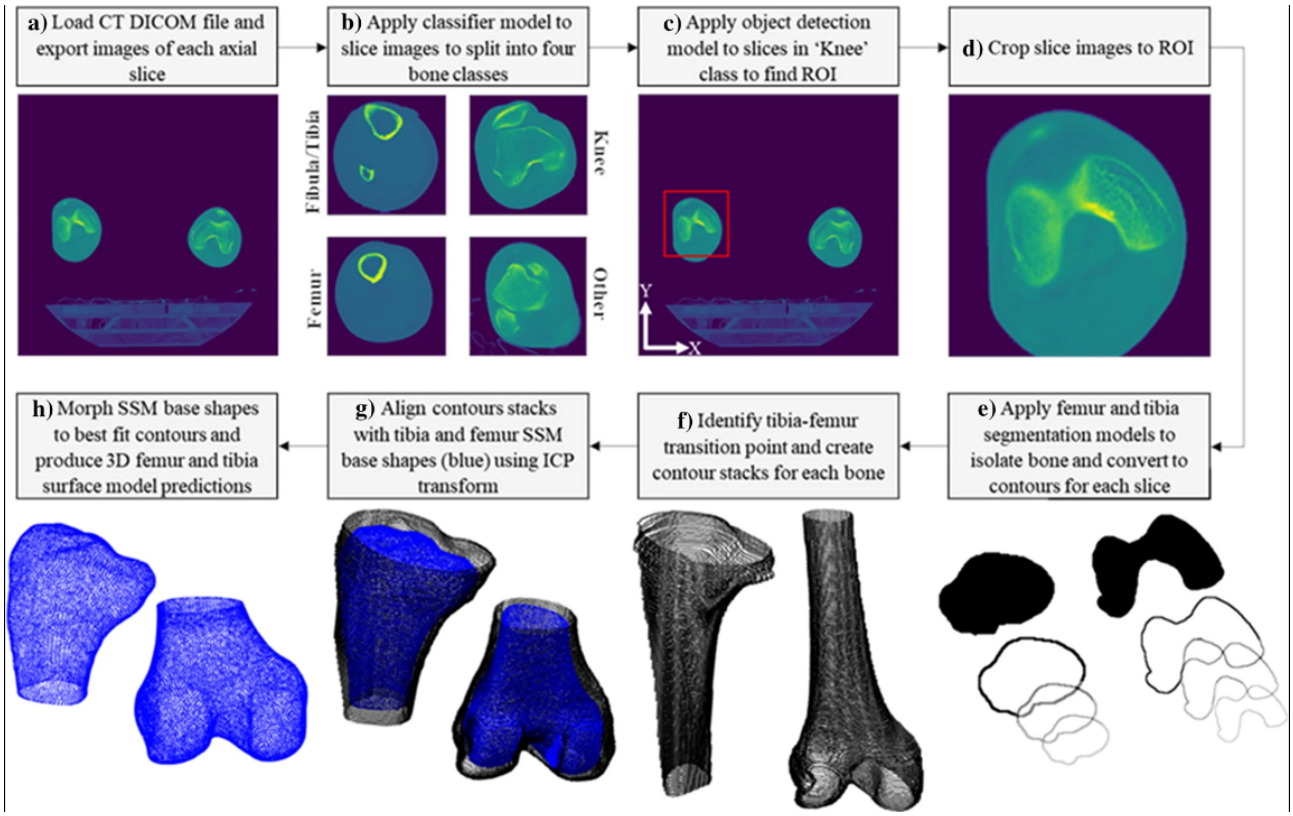

XSL•FO
RenderX

The scope of AI's impact extends beyond prosthetics to include adaptive implants, which are revolutionizing areas such as orthopedics. Adaptive implants also benefit from AI integration, particularly in orthopedics, where implants can adjust over time to changes in the body or the mechanical loads they must endure. This adaptability can significantly enhance the implant's longevity and the comfort it offers, making it a vital feature in modern medical practice [114]. For instance, recent advancements in knee arthroplasty involve using ML techniques to automate the customization of stiffness-matched knee implants, ensuring they closely mimic the mechanical properties of the patient's own bones, thereby improving post-surgical outcomes and promoting natural bone remodeling [115]. Furthermore, new designs in transfemoral prosthetic sockets incorporate a combination of rigid frames and silicone structures that not only offer better biomechanical coupling with the residual limb but also integrate smart technologies for health monitoring and feedback [116]. These AI-enhanced implants can respond to the body's changes over time, adjusting to shifts in load and movement. In spinal and joint replacement surgeries, AI assisted in the design of implants that align with the patient's unique biomechanical structure, potentially enhancing the longevity and comfort of these devices [117]. This holistic approach to prosthetics and implants, driven by AI, represents a significant step forward in personalized medical devices, offering enhanced functionality and improved patient comfort through adaptive capabilities.

The potential of AI in prosthetics and implants extends into holistic health management, where these technologies facilitate continuous monitoring and condition management, contributing to preventive health care. Innovations such as electromyography and gyroscope integration in sports training assist systems exemplify how AI-driven devices can enhance physical activities and training effectiveness [118]. Each advancement not only exemplifies the integration of AI in improving device functionality but also highlights the shift toward more personalized, responsive medical care, aligning with the broader goals of enhancing patient outcomes and optimizing therapeutic efficacy.

Challenges in AI-Driven Personalization

Overview

The integration of AI into medical devices heralds significant advancements in personalized health care. However, this integration is not without its challenges, which span technical, ethical, social, and regulatory domains. Addressing these challenges is crucial for the successful implementation and acceptance of AI-driven personalization in medical devices.

Technical Challenges

In the realm of AI-driven personalization in health care, technical challenges significantly impact the development and implementation of these innovative technologies. One of the foundational hurdles is the issue of data quality and availability. Effective AI models rely on high-quality, comprehensive datasets, yet health care often struggles with accessing such data [119]. Patient information is frequently dispersed across various institutions and platforms, creating considerable gaps in data necessary for AI training. Additionally, privacy concerns and stringent regulatory frameworks can restrict access to vital data, compounding the difficulty in ensuring data quality, specifically in terms of accuracy, completeness, and relevance, due to varied data collection methodologies and the potential for human error [120,121]. For instance, the integration of wearables in clinical settings necessitates additional validation to confirm their accuracy, reliability, and effectiveness in enhancing outcomes for chronic conditions. Issues such as data privacy, standardization of device functionalities, and integration into existing health care infrastructures also need to be addressed to leverage their potential in chronic disease management [106].

Another significant technical challenge is the development of reliable and valid AI algorithms that perform consistently well across diverse patient populations and clinical environments [122]. Algorithmic reliability is crucial for clinician acceptance and patient safety. Validation of these algorithms demands extensive testing to ensure they deliver accurate predictions and recommendations under various circumstances [123]. However, the opaque nature of some AI models complicates this validation process, leading to concerns about their interpretability and the feasibility of auditing their decision-making processes [124].

Moreover, the successful deployment of AI-driven devices hinges on their seamless integration with existing health care systems and workflows. This integration presents substantial technical challenges, as health care infrastructures are often intricate and composed of outdated software and hardware that may not readily accommodate new AI functionalities [32]. Achieving interoperability between AI systems and electronic health records is essential for real-time data exchange and analysis, yet often faces obstacles due to compatibility issues [125]. These technical challenges must be navigated carefully to realize the full potential of AI in personalizing medical treatment and enhancing patient care.

Ethical and Social Challenges

The integration of AI into medical devices, while promising, brings with it a myriad of ethical and social challenges that require diligent attention and strategic management. Privacy and data security concerns are at the forefront of these challenges [126]. The extensive collection, storage, and analysis of sensitive patient data inherent in AI applications necessitate rigorous cybersecurity measures to protect against unauthorized access and breaches [127]. Ensuring the confidentiality, integrity, and availability of this data is not only a technical requirement but also a moral imperative, demanding strict adherence to evolving privacy regulations and standards [120].

Another significant ethical concern is addressing algorithmic bias and fairness. AI systems, inherently dependent on the data with which they are trained, can perpetuate existing biases if the data reflect historical inequities in health care [128]. Such biases could result in AI-driven medical devices performing differently across diverse patient groups, potentially exacerbating health disparities rather than alleviating them [129]. To counteract this, there needs to be a proactive effort in the selection and scrutiny of training datasets and the design of AI models, ensuring they are both inclusive and equitable.

XSL•FO
RenderX

Furthermore, the deployment of AI in health care raises complex issues around patient consent and autonomy. It is essential that patients are thoroughly informed about how their data will be used, the involvement of AI in their treatment, and the implications of decisions made by AI systems [130]. This is particularly challenging given the complexity of AI technologies and the potential for misunderstandings among patients lacking technical expertise. Moreover, care must be taken to ensure that AI-enhanced medical devices support, rather than undermine, patient autonomy, allowing individuals to remain central in making informed decisions about their health care [131].

These ethical and social challenges underscore the need for a balanced approach to the deployment of AI in medical devices, where innovation is matched by an unwavering commitment to ethical standards, patient rights, and social justice. As AI technology continues to evolve, so must the strategies for addressing the ethical implications it brings to the health care landscape.

Regulatory and Policy Challenges

Navigating the complex regulatory landscape for AI-enabled medical devices presents a significant challenge as authorities worldwide strive to balance innovation with patient safety. Regulatory bodies are actively developing frameworks that ensure the safety and efficacy of these technologies [12]. This entails rigorous processes for AI development, validation, and ongoing post-market monitoring [132]. The task is particularly daunting due to the rapid technological advances in AI, which require regulators to continually adapt and update guidelines to keep pace with innovation [133].

In addition to regulatory hurdles, the standardization and certification of AI-driven medical devices are critical yet complex issues. Establishing universal protocols for development, testing, and certification is essential to ensure the reliability and interoperability of these devices [134]. However, achieving consensus on these standards involves intricate negotiations among diverse stakeholders from the health care, technology, and regulatory sectors. These standards must not only be robust enough to ensure safety and efficacy but also flexible enough to adapt to future technological innovations [135].

Moreover, the international dimension of AI-driven personalization in medical devices adds another layer of complexity. Global collaboration is necessary to tackle issues like data sharing, privacy protection, and the alignment of regulatory standards [136]. Harmonizing these standards across different jurisdictions can facilitate the widespread adoption of AI technologies in health care [137]. Yet, this requires a concerted effort to reconcile varying regulatory philosophies, health care systems, and cultural attitudes toward privacy and autonomy.

Addressing these regulatory and policy challenges necessitates a multidisciplinary approach that involves experts from health care, technology, ethics, and law. By fostering open collaboration and dialogue among these stakeholders, the sector can effectively navigate the intricacies of AI-driven personalization in medical devices. This collaborative approach is vital for unleashing the transformative potential of AI in health care, enhancing global health outcomes while ensuring compliance with rigorous safety standards.

Discussion

Principal Findings

The comprehensive evaluation of AI-integrated medical devices highlights a remarkable evolution in health care technology, spanning multiple domains including diagnostics, therapeutics, prosthetics, and robotics. The results demonstrate how AI-powered solutions are reshaping health care delivery, enabling real-time monitoring, data-driven interventions, and enhanced patient outcomes. In diabetes management, smart insulin pumps exemplify the shift toward precision medicine. These devices leverage real-time data from CGMs and ML algorithms to autonomously modulate insulin delivery, addressing long-standing challenges such as glycemic variability and poor treatment adherence. Complementing this, AI-enabled insulin pens that achieve over 95% accuracy in detecting injection sites and lipodystrophies illustrate how intelligent devices can mitigate complications and support better patient compliance [35]. Further advancements, such as predictive self-care devices, highlight how AI facilitates proactive diabetes management, reducing the risk of hypoglycemia and hyperglycemia [138].

AI-powered prosthetics demonstrate a significant leap in rehabilitation and mobility aids. These systems, particularly those incorporating surface electromyographic-based systems and deep learning frameworks, adaptively interpret complex muscle signals, enhancing patient-robot interaction and optimizing motor function restoration. Innovations in gait analysis and lower limb assistive technologies further contribute to fall prevention and stability in dynamic environments [39]. Beyond mobility, AI-powered tools also enhance clinical workflow, with AI-driven radiograph analysis improving the speed and accuracy of pre- and post-operative orthopedic evaluations.

The diagnostic sector has witnessed transformative developments, particularly with deep learning models outperforming traditional diagnostic approaches. In medical imaging, AI-enabled systems detect cancerous lesions and other pathologies with higher sensitivity and specificity compared to manual interpretation. AI-enhanced frameworks have been pivotal in diagnosing pancreatic and lung cancers, improving early detection rates and clinical decision-making. The successful application of deep learning in chest X-ray analysis, achieving accuracy rates nearing 99%, reflects the capacity of AI to revolutionize pulmonary diagnostics. Moreover, the incorporation of AI into ophthalmology and cardiology diagnostics has led to non-invasive, precise evaluations of diabetic retinopathy and cardiovascular conditions.

Comparison With Prior Work

In comparison to prior literature, our findings are consistent with earlier studies that highlight the critical role of AI in personalizing medical interventions. For example, the diagnostic accuracy of AI models surpassing traditional radiological

XSL•FO
RenderX

assessments aligns with recent meta-analyses showing AI's superior sensitivity and specificity in medical imaging [43]. Additionally, the effectiveness of AI-powered wearables and biosensors in cardiovascular and metabolic monitoring complements existing literature on personalized chronic disease management [101].

The reviewed case studies build upon previous research by providing concrete evidence of AI's integration into rehabilitation devices, surgical planning tools, and patient communication platforms, such as AI-powered chatbots. Furthermore, AI-driven technologies are making significant contributions to medical manufacturing, education, and patient communication. AI-enhanced 3D printing supports the creation of personalized implants and wearable devices, while AI-powered chatbots revolutionize patient consultation by delivering timely, accurate medical advice.

Additionally, our review underscores how AI customization in prosthetics and implants contributes to reducing production times and improving biomechanical outcomes, which reflects a growing trend observed in related engineering and biomedical studies. The reviewed studies further elaborate on AI's role in adaptive implants and intelligent prosthetic sockets equipped with biosensors, offering continuous feedback, improving comfort, and enhancing functional outcomes for users [114].

Strengths and Limitations

One of the strengths of this review is the comprehensive coverage of AI applications across a broad spectrum of medical devices, ranging from diagnostic imaging to rehabilitation tools and patient engagement platforms. The systematic methodology ensured the inclusion of peer-reviewed studies from diverse health care domains, enhancing the robustness and generalizability of our findings. Additionally, by using a thematic categorization, we provided clarity in synthesizing complex data. The use of a standardized data extraction protocol based on PRISMA-ScR guidelines ensured methodological rigor and minimized bias. Further, the incorporation of snowballing techniques allowed us to capture relevant studies that may have been missed during the initial database searches, contributing to the depth and breadth of our review.

The review has a number of limitations. The restriction to English-language publications may have excluded valuable studies in other languages, potentially introducing language bias. To address this, we conducted cross-verification using key non-English papers referenced in systematic reviews and meta-analyses. Additionally, the study did not fully explore how specific patient demographics, such as age, gender, socioeconomic status, and comorbidities, may influence responses to AI-driven medical devices, which could affect generalizability and clinical relevance. Another limitation lies in the rapidly evolving nature of AI technology, which may render some findings subject to change as new developments emerge. Additionally, while our review emphasizes AI's potential in health care, real-world implementation challenges such as regulatory compliance, data privacy concerns, and the need for interdisciplinary collaboration remain critical issues for future research.

Future Directions

Future research should continue to explore the integration of emerging technologies such as quantum computing, blockchain, and AI-driven personalization into health care. These technologies offer opportunities to enhance the precision, security, and scalability of medical devices and health care delivery [139,140].

Quantum computing holds significant promise for advancing the speed and complexity of data processing, which could improve the performance of AI algorithms used in diagnostics, disease monitoring, and predictive analytics [141]. This computational capacity may enable the development of adaptive medical devices that respond dynamically to patients' evolving health needs [142].

Blockchain technology has the potential to improve data security and integrity by enabling decentralized, tamper-proof health information systems [143]. This could support the development of more trustworthy AI models, particularly in sensitive domains such as patient data management and cross-institutional interoperability [131].

In addition, future research should investigate the role of AI-driven personalized medical devices in addressing global health disparities. Portable AI-enabled diagnostic tools and wearable technologies may help expand health care access in underserved and remote regions [144]. Such innovations could also inform targeted public health strategies, contributing to a more equitable global health landscape. As these technologies become more integrated into health care ecosystems, continued interdisciplinary collaboration will be essential to navigate ethical, regulatory, and technical challenges.

Conclusions

The integration of AI into medical devices is fundamentally reshaping modern health care by enabling the shift toward personalized and precision medicine. AI-powered innovations, from smart insulin pumps to diagnostic imaging tools and wearable biosensors, are transforming how health care providers diagnose, treat, and monitor patients. These technologies harness advanced ML algorithms and real-time data to tailor medical interventions to individual needs, enhancing treatment efficacy and improving patient outcomes. The dynamic capabilities of AI in automating tasks, recognizing patterns in complex datasets, and adapting to physiological feedback exemplify its critical role in evolving health care delivery toward a more patient-centric model.

The technological landscape continues to expand with advancements in deep learning, natural language processing, and edge computing, which have accelerated the development of AI-enhanced therapeutic and assistive devices. Wearable technologies now enable continuous health monitoring, empowering patients to actively participate in managing chronic conditions such as diabetes and cardiovascular diseases. Meanwhile, AI-driven prosthetics and adaptive implants offer improved functionality and comfort by personalizing device behavior to the user's biomechanical and physiological characteristics. This fusion of AI with medical devices not only augments clinical workflows but also enhances the quality of

XSL•FO
RenderX

life for patients through intuitive, real-time, and highly responsive health care solutions.

Despite these strides, realizing the full potential of AI in personalizing medical devices requires addressing key challenges related to data privacy, algorithmic bias, system interoperability, and regulatory compliance. Ensuring AI's ethical deployment, maintaining data integrity, and aligning with global safety standards are essential to prevent disparities in health care delivery. Future innovations, including the integration of quantum computing to optimize AI algorithms and blockchain for secure data management, hold promises for overcoming these barriers. As the health care ecosystem embraces interdisciplinary collaboration across technology, ethics, and policy, AI-driven medical devices will increasingly define the future of global health—advancing precision medicine to be more accessible, equitable, and effective for all populations.

Conflicts of Interest

None declared.

Checklist 1
PRISMA checklist.
[DOCX File, 32 KB - jmir_v28i1e72410_app1.docx]

References

1. Johnson KB, Wei WQ, Weeraratne D, et al. Precision medicine, AI, and the future of personalized health care. Clin Transl Sci 2021 Jan;14(1):86-93. [doi: 10.1111/cts.12884] [Medline: 32961010]
2. Li YH, Li YL, Wei MY, Li GY. Innovation and challenges of artificial intelligence technology in personalized healthcare. Sci Rep 2024;14(1):18994-19728. [doi: 10.1038/s41598-024-70073-7]
3. Mohanadas HP, Nair V, Doctor AA, et al. A systematic analysis of additive manufacturing techniques in the bioengineering of in vitro cardiovascular models. Ann Biomed Eng 2023 Nov;51(11):2365-2383. [doi: 10.1007/s10439-023-03322-x] [Medline: 37466879]
4. Korteling JEH, van de Boer-Visschedijk GC, Blankendaal RAM, Boonekamp RC, Eikelboom AR. Human- versus artificial intelligence. Front Artif Intell 2021;4:622364. [doi: 10.3389/frai.2021.622364] [Medline: 33981990]
5. Khatiwada P, Yang B, Lin JC, Blobel B. Patient-generated health data (PGHD): understanding, requirements, challenges, and existing techniques for data security and privacy. J Pers Med 2024 Mar 3;14(3):282. [doi: 10.3390/jpm14030282] [Medline: 38541024]
6. Ahmed Z, Mohamed K, Zeeshan S, Dong X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. Database (Oxford) 2020 Jan 1;2020:baaa010. [doi: 10.1093/database/baaa010] [Medline: 32185396]
7. Manickam P, Mariappan SA, Murugesan SM, et al. Artificial intelligence (AI) and internet of medical things (IoMT) assisted biomedical systems for intelligent healthcare. Biosensors (Basel) 2022 Jul 25;12(8):562. [doi: 10.3390/bios12080562] [Medline: 35892459]
8. Iqbal MJ, Javed Z, Sadia H, et al. Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future. Cancer Cell Int 2021 May 21;21(1):270. [doi: 10.1186/s12935-021-01981-1] [Medline: 34020642]
9. Malcangi M. AI-based methods and technologies to develop wearable devices for prosthetics and predictions of degenerative diseases. Methods Mol Biol 2021;2190:337-354. [doi: 10.1007/978-1-0716-0826-5_17] [Medline: 32804375]
10. Nahavandi D, Alizadehsani R, Khosravi A, Acharya UR. Application of artificial intelligence in wearable devices: opportunities and challenges. Comput Methods Programs Biomed 2022 Jan;213:106541. [doi: 10.1016/j.cmpb.2021.106541] [Medline: 34837860]
11. Pesapane F, Volonté C, Codari M, Sardanelli F. Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. Insights Imaging 2018 Oct;9(5):745-753. [doi: 10.1007/s13244-018-0645-y] [Medline: 30112675]
12. Gilbert S, Anderson S, Daumer M, Li P, Melvin T, Williams R. Learning from experience and finding the right balance in the governance of artificial intelligence and digital health technologies. J Med Internet Res 2023 Apr 14;25(1):e43682. [doi: 10.2196/43682] [Medline: 37058329]
13. Deepa R, Arunkumar S, Jayaraj V, Sivasamy A. Healthcare's new Frontier: AI-driven early cancer detection for improved well-being. AIP Adv 2023 Nov 1;13(11):11. [doi: 10.1063/5.0177640]
14. Kulkarni PG, Paudel N, Magar S, et al. Overcoming challenges and innovations in orthopedic prosthesis design: an interdisciplinary perspective. Biomed Mater Devices 2023 May 12;2(1):1-12. [doi: 10.1007/s44174-023-00087-8] [Medline: 37363137]
15. Junaid SB, Imam AA, Abdulkarim M, et al. Recent advances in artificial intelligence and wearable sensors in healthcare delivery. Appl Sci (Basel) 2022;12(20):10271. [doi: 10.3390/app122010271]

XSL•FO
RenderX

16. Albahri AS, Duhaim AM, Fadhel MA, et al. A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion. Information Fusion 2023 Aug;96:156-191. [doi: 10.1016/j.inffus.2023.03.008]
17. Nasir S, Khan RA, Bai S. Ethical framework for harnessing the power of AI in healthcare and beyond. IEEE Access 2024;12:31014-31035. [doi: 10.1109/ACCESS.2024.3369912]
18. Kasula BY. Framework development for artificial intelligence integration in healthcare: optimizing patient care and operational efficiency. Transactions on Latest Trends in IoT 2023;6(6):77-83 [FREE Full text]
19. Nakamura T, Sasano T. Artificial intelligence and cardiology: current status and perspective. J Cardiol 2022 Mar;79(3):326-333. [doi: 10.1016/j.jjcc.2021.11.017] [Medline: 34895982]
20. Schurink CAM, Lucas PJF, Hoepelman IM, Bonten MJM. Computer-assisted decision support for the diagnosis and treatment of infectious diseases in intensive care units. Lancet Infect Dis 2005 May;5(5):305-312. [doi: 10.1016/S1473-3099(05)70115-8] [Medline: 15854886]
21. George AH, George AS. From pulse to prescription: exploring the rise of AI in medicine and its implications. Partners Universal International Innovation Journal 2023;1(6):38-54 [FREE Full text]
22. Guni A, Normahani P, Davies A, Jaffer U. Harnessing machine learning to personalize web-based health care content. J Med Internet Res 2021 Oct 19;23(10):e25497. [doi: 10.2196/25497] [Medline: 34665146]
23. Li M, Jiang Y, Zhang Y, Zhu H. Medical image analysis using deep learning algorithms. Front Public Health 2023;11:1273253. [doi: 10.3389/fpubh.2023.1273253] [Medline: 38026291]
24. Al-qaness MAA, Zhu J, AL-Alimi D, et al. Chest X-ray images for lung disease detection using deep learning techniques: a comprehensive survey. Arch Computat Methods Eng 2024 Aug;31(6):3267-3301. [doi: 10.1007/s11831-024-10081-y]
25. Chen C, Zheng S, Guo L, et al. Identification of misdiagnosis by deep neural networks on a histopathologic review of breast cancer lymph node metastases. Sci Rep 2022;12(1):13482. [doi: 10.1038/s41598-022-17606-0]
26. Khurshid S, Friedman S, Reeder C, et al. ECG-based deep learning and clinical risk factors to predict atrial fibrillation. Circulation 2022 Jan 11;145(2):122-133. [doi: 10.1161/CIRCULATIONAHA.121.057480] [Medline: 34743566]
27. Abdulrahman Mohammed Al M, Zakiuddin A, Abdullah M. A. Paradigm shift in healthcare through technology and patient-centeredness. Int Arch Public Health Community Med 2018;2(1):1710015. [doi: 10.23937/iaphcm-2017/1710015]
28. Dhawan AP. Collaborative paradigm of preventive, personalized, and precision medicine with point-of-care technologies. IEEE J Transl Eng Health Med 2016;4:2800908. [doi: 10.1109/JTEHM.2016.2635126] [Medline: 28560119]
29. Gameiro GR, Sinkunas V, Liguori GR, Auler-Júnior JOC. Precision medicine: changing the way we think about healthcare. Clinics (Sao Paulo) 2018 Dec 3;73:e723. [doi: 10.6061/clinics/2017/e723] [Medline: 30517307]
30. Haick H, Tang N. Artificial intelligence in medical sensors for clinical decisions. ACS Nano 2021 Mar 23;15(3):3557-3567. [doi: 10.1021/acsnano.1c00085] [Medline: 33620208]
31. Engineering innovations in healthcare: a review of developments in the USA. Eng Sci Technol J 2023;4(6):381-400. [doi: 10.51594/estj.v4i6.638]
32. Esmaeilzadeh P. Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: a perspective for healthcare organizations. Artif Intell Med 2024 May;151:102861. [doi: 10.1016/j.artmed.2024.102861] [Medline: 38555850]
33. Klonoff DC, Kerr D. Smart pens will improve insulin therapy. J Diabetes Sci Technol 2018 May;12(3):551-553. [doi: 10.1177/1932296818759845] [Medline: 29411641]
34. Vettoretti M, Cappon G, Facchinetti A, Sparacino G. Advanced diabetes management using artificial intelligence and continuous glucose monitoring sensors. Sensors (Basel) 2020 Jul 10;20(14):3870. [doi: 10.3390/s20143870] [Medline: 32664432]
35. Torre E, Francini L, Cordelli E, et al. Exploiting AI to make insulin pens smart: injection site recognition and lipodystrophy detection. 2022 Presented at: 2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS); Jul 21-23, 2022; Shenzen, China p. 211-216. [doi: 10.1109/CBMS55023.2022.00044]
36. Almulla SK, Zaatar O, Ahmed HS, ul Misbah Khanum N, Jarndal A. Smart artificial-intelligence based self-care-device for diabetic patients. 2022 Presented at: 2022 Advances in science and engineering technology international conferences (ASET); Feb 21-24, 2022; Dubai, United Arab Emirates p. 1-5. [doi: 10.1109/ASET53988.2022.9734874]
37. Kesavadev J, Saboo B, Krishna MB, Krishnan G. Evolution of insulin delivery devices: from syringes, pens, and pumps to DIY artificial pancreas. Diabetes Ther 2020 Jun;11(6):1251-1269. [doi: 10.1007/s13300-020-00831-z] [Medline: 32410184]
38. Bozzetto L, Corrado A, Scidà G. Dietary treatment of type 1 diabetes: beyond carbohydrate counting to fight cardiovascular risk. Nutr Metab Cardiovasc Dis 2023 Feb;33(2):299-306. [doi: 10.1016/j.numecd.2022.12.012] [Medline: 36642609]
39. Guo K, Orban M, Lu J, Al-Quraishi MS, Yang H, Elsamanty M. Empowering hand rehabilitation with AI-powered gesture recognition: a study of an sEMG-based system. Bioengineering (Basel) 2023 May 6;10(5):557. [doi: 10.3390/bioengineering10050557] [Medline: 37237627]
40. Mehr JK, Akbari M, Faridi P, Xing H, Mushahwar VK, Tavakoli M. Artificial‐intelligence‐powered lower limb assistive devices: future of home care technologies. Adv Intell Syst 2023 Jun;5(6):2200361. [doi: 10.1002/aisy.202200361]

XSL•FO
RenderX

41. Pagano S, Müller K, Götz J, et al. The role and efficiency of an AI-powered software in the evaluation of lower limb radiographs before and after total knee arthroplasty. J Clin Med 2023 Aug 24;12(17):5498. [doi: 10.3390/jcm12175498] [Medline: 37685563]
42. Hameed BS, Krishnan UM. Artificial intelligence-driven diagnosis of pancreatic cancer. Cancers (Basel) 2022 Oct 31;14(21):5382. [doi: 10.3390/cancers14215382] [Medline: 36358800]
43. Xu M, Chen Z, Zheng J, Zhao Q, Yuan Z. Artificial intelligence-aided optical imaging for cancer theranostics. Semin Cancer Biol 2023 Sep;94:62-80. [doi: 10.1016/j.semcancer.2023.06.003] [Medline: 37302519]
44. Sanida MV, Sanida T, Sideris A, Dasygenis M. An advanced deep learning framework for multi-class diagnosis from chest X-ray images. J 2024;7(1):48-71. [doi: 10.3390/j7010003]
45. Papadopoulos GT, Antona M, Stephanidis C. Towards open and expandable cognitive AI architectures for large-scale multi-agent human-robot collaborative learning. IEEE Access 2021;9:73890-73909. [doi: 10.1109/ACCESS.2021.3080517]
46. Rossos D, Mihailidis A, Laschowski B. AI-powered smart glasses for sensing and recognition of human-robot walking environments. bioRxiv Bioengineering. Preprint posted online on Oct 29, 2023. [doi: 10.1101/2023.10.24.563804]
47. Robbins S. AI and the path to envelopment: knowledge as a first step towards the responsible regulation and use of AI-powered machines. AI & Soc 2020 Jun;35(2):391-400. [doi: 10.1007/s00146-019-00891-1]
48. Hlee S, Park J, Park H, Koo C, Chang Y. Understanding customer's meaningful engagement with AI-powered service robots. ITP 2023 Apr 11;36(3):1020-1047. [doi: 10.1108/ITP-10-2020-0740]
49. Moriuchi E. Leveraging the science to understand factors influencing the use of AI-powered avatar in healthcare services. J Technol Behav Sci 2022;7(4):588-602. [doi: 10.1007/s41347-022-00277-z]
50. Erol BA, Majumdar A, Benavidez P, Rad P, Choo KKR, Jamshidi M. Toward artificial emotional intelligence for cooperative social human–machine interaction. IEEE Trans Comput Soc Syst 2019;7(1):234-246. [doi: 10.1109/TCSS.2019.2922593]
51. Mladenovic R. AI-powered and "augmented" dentistry: applications, implications and limitations. In: Augmented Reality and Artificial Intelligence: The Fusion of Advanced Technologies: Springer Nature Switzerland; 2023:211-226 URL: https://doi.org/10.1007/978-3-031-27166-3_12 [accessed 2026-03-02]
52. Zhu Z, Ng DWH, Park HS, McAlpine MC. 3D-printed multifunctional materials enabled by artificial-intelligence-assisted fabrication technologies. Nat Rev Mater 2021;6(1):27-47. [doi: 10.1038/s41578-020-00235-2]
53. Schwarz GM, Simon S, Mitterer JA, et al. Artificial intelligence enables reliable and standardized measurements of implant alignment in long leg radiographs with total knee arthroplasties. Knee Surg Sports Traumatol Arthrosc 2022 Aug;30(8):2538-2547. [doi: 10.1007/s00167-022-07037-9] [Medline: 35819465]
54. Itelman E, Golovchiner G, Barsheshet A, Kornowski R, Erez A. Balancing innovation and professionalism: the emerging role of AI-powered chatbots in medical consultation. Heart Rhythm 2024 Oct;21(10):2037-2039. [doi: 10.1016/j.hrthm.2024.04.010] [Medline: 38588991]
55. Samanoodi R, Alrasheed A, Felemban A, et al. AI-powered mobility application for field inspection rounds. 2023 Oct 2 Presented at: ADIPEC; Oct 2-5, 2023. [doi: 10.2118/216996-MS]
56. Sudharson NA, Joseph M, Kurian N, Varghese KG, Wadhwa S, Thomas HA. AI-powered neural implants. Br Dent J 2023 Mar;234(6):359-360. [doi: 10.1038/s41415-023-5698-8] [Medline: 36964340]
57. Wu N, Phang J, Park J, et al. Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans Med Imaging 2020 Apr;39(4):1184-1194. [doi: 10.1109/TMI.2019.2945514] [Medline: 31603772]
58. Chae KJ, Jin GY, Ko SB, et al. Deep learning for the classification of small (≤2 cm) pulmonary nodules on CT imaging: a preliminary study. Acad Radiol 2020 Apr;27(4):e55-e63. [doi: 10.1016/j.acra.2019.05.018] [Medline: 31780395]
59. Mhamdi L, Dammak O, Cottin F, Dhaou IB. Artificial intelligence for cardiac diseases diagnosis and prediction using ECG images on embedded systems. Biomedicines 2022 Aug 19;10(8):8. [doi: 10.3390/biomedicines10082013] [Medline: 36009560]
60. Padmavilochanan D, Pathinarupothi RK, Menon KAU, et al. Personalized diabetes monitoring platform leveraging IoMT and AI for non-invasive estimation. Smart Health (2014) 2023 Dec;30:100428. [doi: 10.1016/j.smhl.2023.100428]
61. Lugogo NL, DePietro M, Reich M, et al. A predictive machine learning tool for asthma exacerbations: results from a 12-week, open-label study using an electronic multi-dose dry powder inhaler with integrated sensors. J Asthma Allergy 2022;15:1623-1637. [doi: 10.2147/JAA.S377631] [Medline: 36387836]
62. McRae MP, Rajsri KS, Alcorn TM, McDevitt JT. Smart diagnostics: combining artificial intelligence and in vitro diagnostics. Sensors (Basel) 2022 Aug 24;22(17):6355. [doi: 10.3390/s22176355] [Medline: 36080827]
63. Luchini C, Pea A, Scarpa A. Artificial intelligence in oncology: current applications and future perspectives. Br J Cancer 2022 Jan;126(1):4-9. [doi: 10.1038/s41416-021-01633-1] [Medline: 34837074]
64. Rahmani F. The application of machine learning in the detection of Covid-19 in blood tests. Int J New Chem 2021;8(5):103-126 [FREE Full text]
65. Krittanawong C, Johnson KW, Choi E, et al. Artificial intelligence and cardiovascular genetics. Life (Basel) 2022 Feb 14;12(2):279. [doi: 10.3390/life12020279] [Medline: 35207566]
66. Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med 2021 Apr 7;4(1):65. [doi: 10.1038/s41746-021-00438-z] [Medline: 33828217]

XSL•FO
RenderX

67. Artificial intelligence-enabled medical devices. FDA. 2024 Dec 20. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices [accessed 2026-02-05]
68. Parmar UPS, Surico PL, Singh RB, et al. Artificial intelligence (AI) for early diagnosis of retinal diseases. Medicina (Kaunas) 2024 Mar 23;60(4):527. [doi: 10.3390/medicina60040527] [Medline: 38674173]
69. Chudzik P, Majumdar S, Calivá F, Al-Diri B, Hunter A. Microaneurysm detection using fully convolutional neural networks. Comput Methods Programs Biomed 2018 May;158:185-192. [doi: 10.1016/j.cmpb.2018.02.016] [Medline: 29544784]
70. Wang H, Yuan G, Zhao X, et al. Hard exudate detection based on deep model learned information and multi-feature joint representation for diabetic retinopathy screening. Comput Methods Programs Biomed 2020 Jul;191:105398. [doi: 10.1016/j.cmpb.2020.105398] [Medline: 32092614]
71. Murugan R, Roy P. MicroNet: microaneurysm detection in retinal fundus images using convolutional neural network. Soft Comput 2022 Feb;26(3):1057-1066. [doi: 10.1007/s00500-022-06752-2]
72. Chiu HY, Chao HS, Chen YM. Application of artificial intelligence in lung cancer. Cancers (Basel) 2022 Mar 8;14(6):1370. [doi: 10.3390/cancers14061370] [Medline: 35326521]
73. Fusco R, Grassi R, Granata V, et al. Artificial intelligence and COVID-19 using chest CT scan and chest X-ray images: machine learning and deep learning approaches for diagnosis and treatment. J Pers Med 2021 Sep 30;11(10):993. [doi: 10.3390/jpm11100993] [Medline: 34683133]
74. de Freitas Barbosa VA, Gomes JC, de Santana MA, et al. Heg.IA: an intelligent system to support diagnosis of Covid-19 based on blood tests. Res Biomed Eng 2022 Mar;38(1):99-116. [doi: 10.1007/s42600-020-00112-5]
75. Kukar M, Gunčar G, Vovko T, et al. COVID-19 diagnosis by routine blood tests using machine learning. Sci Rep 2021 May 24;11(1):10738. [doi: 10.1038/s41598-021-90265-9] [Medline: 34031483]
76. De La Vega FM, Chowdhury S, Moore B, et al. Artificial intelligence enables comprehensive genome interpretation and nomination of candidate diagnoses for rare genetic diseases. Genome Med 2021 Oct 14;13(1):153. [doi: 10.1186/s13073-021-00965-0] [Medline: 34645491]
77. Vadapalli S, Abdelhalim H, Zeeshan S, Ahmed Z. Artificial intelligence and machine learning approaches using gene expression and variant data for personalized medicine. Brief Bioinform 2022 Sep 20;23(5):bbac191. [doi: 10.1093/bib/bbac191] [Medline: 35595537]
78. Ghazal TM, Al Hamadi H, Umar Nasir M, et al. Supervised machine learning empowered multifactorial genetic inheritance disorder prediction. Comput Intell Neurosci 2022;2022(1):1051388. [doi: 10.1155/2022/1051388] [Medline: 35685134]
79. Goodman KE, Morgan DJ, Hoffmann DE. Clinical algorithms, antidiscrimination laws, and medical device regulation. JAMA 2023 Jan 24;329(4):285-286. [doi: 10.1001/jama.2022.23870] [Medline: 36602795]
80. Saw SN, Ng KH. Current challenges of implementing artificial intelligence in medical imaging. Phys Med 2022 Aug;100:12-17. [doi: 10.1016/j.ejmp.2022.06.003] [Medline: 35714523]
81. Medanki S, Dommati N, Bodapati HH, et al. Artificial intelligence powered glucose monitoring and controlling system: pumping module. World J Exp Med 2024 Mar 20;14(1):87916. [doi: 10.5493/wjem.v14.i1.87916] [Medline: 38590308]
82. Marian CV. Artificial intelligence expert system based on continuous glucose monitoring (CGM) data for auto-adaptive adjustment therapy protocol – how to make sensors and patients to think forward and work together? Presented at: 2021 International Conference on e-Health and Bioengineering (EHB); Nov 18-19, 2021. [doi: 10.1109/EHB52898.2021.9657707]
83. Choudhury AD, Banerjee R, Kimbahune S, Pal A. New Frontiers of Cardiovascular Screening Using Unobtrusive Sensors, AI, and IoT: Academic Press; 2022. [doi: 10.1016/C2019-0-05510-0]
84. Howard JP, Fisher L, Shun-Shin MJ, et al. Cardiac rhythm device identification using neural networks. JACC Clin Electrophysiol 2019 May;5(5):576-586. [doi: 10.1016/j.jacep.2019.02.003] [Medline: 31122379]
85. Oida M, Mizutani T, Hasumi E, et al. Prediction of pacemaker-induced cardiomyopathy using a convolutional neural network based on clinical findings prior to pacemaker implantation. Sci Rep 2024 Mar 22;14(1):6916. [doi: 10.1038/s41598-024-57418-y] [Medline: 38519537]
86. Krittanawong C, Rogers AJ, Johnson KW, et al. Integration of novel monitoring devices with machine learning technology for scalable cardiovascular management. Nat Rev Cardiol 2021 Feb;18(2):75-91. [doi: 10.1038/s41569-020-00445-9] [Medline: 33037325]
87. Venkatesan D, Elangovan A, Winster H, et al. Diagnostic and therapeutic approach of artificial intelligence in neuro-oncological diseases. Biosensors and Bioelectronics: X 2022 Sep;11:100188. [doi: 10.1016/j.biosx.2022.100188]
88. Larburu N, Artetxe A, Escolar V, Lozano A, Kerexeta J. Artificial intelligence to prevent mobile heart failure patients decompensation in real time: monitoring-based predictive model. Mob Inf Syst 2018 Nov 5;2018(1):1-11. [doi: 10.1155/2018/1546210]
89. Shajari S, Kuruvinashetti K, Komeili A, Sundararaj U. The emergence of AI-based wearable sensors for digital health technology: a review. Sensors (Basel) 2023 Nov 29;23(23):9498. [doi: 10.3390/s23239498] [Medline: 38067871]
90. Xie Y, Lu L, Gao F, et al. Integration of artificial intelligence, blockchain, and wearable technology for chronic disease management: a new paradigm in smart healthcare. CURR MED SCI 2021 Dec;41(6):1123-1133. [doi: 10.1007/s11596-021-2485-0] [Medline: 34950987]

XSL•FO
RenderX

91. Karimi F, Amoozgar Z, Reiazi R, Hosseinzadeh M, Rawassizadeh R. Longitudinal analysis of heart rate and physical activity collected from smartwatches. CCF Trans Pervasive Comp Interact 2024 Mar;6(1):18-35. [doi: 10.1007/s42486-024-00147-y]
92. Arora A, Chakraborty P, Bhatia MPS. Intervention of wearables and smartphones in real time monitoring of sleep and behavioral health: an assessment using adaptive neuro-fuzzy technique. Arab J Sci Eng 2022 Feb;47(2):1999-2024. [doi: 10.1007/s13369-021-06078-5]
93. Esakia A, McCrickard S, Harden S, Horning M, Ramalingam N. Using smartwatches to facilitate a group dynamics-based statewide physical activity intervention. Int J Hum Comput Stud 2020 Oct;142:102501. [doi: 10.1016/j.ijhcs.2020.102501]
94. Lin Z, Zheng J, Wang Y, et al. Prediction of the efficacy of group cognitive behavioral therapy using heart rate variability based smart wearable devices: a randomized controlled study. BMC Psychiatry 2024 Mar 6;24(1):187. [doi: 10.1186/s12888-024-05638-x] [Medline: 38448895]
95. Jin X, Cai A, Xu T, Zhang X. Artificial intelligence biosensors for continuous glucose monitoring. Interdisciplinary Materials 2023 Mar;2(2):290-307 [FREE Full text] [doi: 10.1002/idm2.12069]
96. Yoon H, Xuan X, Jeong S, Park JY. Wearable, robust, non-enzymatic continuous glucose monitoring system and its in vivo investigation. Biosens Bioelectron 2018 Oct 15;117:267-275. [doi: 10.1016/j.bios.2018.06.008] [Medline: 29909198]
97. Sawayama J, Takeuchi S. Long‐term continuous glucose monitoring using a fluorescence‐based biocompatible hydrogel glucose sensor. Adv Healthcare Materials 2021 Feb;10(3):2001286. [doi: 10.1002/adhm.202001286]
98. Moreddu R, Wolffsohn JS, Vigolo D, Yetisen AK. Laser-inscribed contact lens sensors for the detection of analytes in the tear fluid. Sensors and Actuators B: Chemical 2020 Aug;317:128183. [doi: 10.1016/j.snb.2020.128183]
99. Arakawa T, Tomoto K, Nitta H, et al. A wearable cellulose acetate-coated mouthguard biosensor for in vivo salivary glucose measurement. Anal Chem 2020 Sep 15;92(18):12201-12207. [doi: 10.1021/acs.analchem.0c01201] [Medline: 32927955]
100. Gao W, Emaminejad S, Nyein HYY, et al. Fully integrated wearable sensor arrays for multiplexed in situ perspiration analysis. Nature New Biol 2016 Jan 28;529(7587):509-514. [doi: 10.1038/nature16521] [Medline: 26819044]
101. Broers ER, Widdershoven J, Denollet J, et al. Personalized eHealth program for life-style CHANGE: results from the “Do Cardiac Health Advanced New Generated Ecosystem (Do CHANGE 2)” randomized controlled trial. Psychosom Med 2020 May;82(4):409-419. [doi: 10.1097/PSY.0000000000000802] [Medline: 32176191]
102. Romero-Tapiador S, Lacruz-Pleguezuelos B, Tolosana R, et al. AI4FoodDB: a database for personalized e-Health nutrition and lifestyle through wearable devices and artificial intelligence. Database (Oxford) 2023 Jul 18;2023:baad049. [doi: 10.1093/database/baad049] [Medline: 37465917]
103. Chatterjee A, Pahari N, Prinz A, Riegler M. AI and semantic ontology for personalized activity eCoaching in healthy lifestyle recommendations: a meta-heuristic approach. BMC Med Inform Decis Mak 2023 Dec 1;23(1):278. [doi: 10.1186/s12911-023-02364-4] [Medline: 38041041]
104. Kargarandehkordi A, Slade C, Washington P. Personalized AI-driven real-time models to predict stress-induced blood pressure spikes using wearable devices: proposal for a prospective cohort study. JMIR Res Protoc 2024 Mar 25;13(1):e55615. [doi: 10.2196/55615] [Medline: 38526539]
105. Mattison G, Canfell O, Forrester D, et al. The influence of wearables on health care outcomes in chronic disease: systematic review. J Med Internet Res 2022 Jul 1;24(7):e36690. [doi: 10.2196/36690] [Medline: 35776492]
106. Ahmed A, Aziz S, Qidwai U, Abd-Alrazaq A, Sheikh J. Performance of artificial intelligence models in estimating blood glucose level among diabetic patients using non-invasive wearable device data. Comput Methods Programs Biomed Update 2023;3:100094. [doi: 10.1016/j.cmpbup.2023.100094]
107. Kario K. Management of hypertension in the digital era: small wearable monitoring devices for remote blood pressure monitoring. Hypertension 2020 Sep;76(3):640-650. [doi: 10.1161/HYPERTENSIONAHA.120.14742] [Medline: 32755418]
108. Wilmink G, Dupey K, Alkire S, et al. Artificial intelligence-powered digital health platform and wearable devices improve outcomes for older adults in assisted living communities: pilot intervention study. JMIR Aging 2020 Sep 10;3(2):e19554. [doi: 10.2196/19554] [Medline: 32723711]
109. Hughes A, Shandhi MMH, Master H, Dunn J, Brittain E. Wearable devices in cardiovascular medicine. Circ Res 2023 Mar 3;132(5):652-670. [doi: 10.1161/CIRCRESAHA.122.322389] [Medline: 36862812]
110. Wang C, He T, Zhou H, Zhang Z, Lee C. Artificial intelligence enhanced sensors - enabling technologies to next-generation healthcare and biomedical platform. Bioelectron Med 2023 Aug 2;9(1):17. [doi: 10.1186/s42234-023-00118-1] [Medline: 37528436]
111. Menon A, Olascoaga LIG, Balanaga V, et al. Shared control of assistive robots through user-intent prediction and hyperdimensional recall of reactive behavior. 2023 Presented at: 2023 IEEE International Conference on Robotics and Automation (ICRA); May 29-2, 2023; London, United Kingdom p. 12638-12644. [doi: 10.1109/ICRA48891.2023.10161509]
112. Potenza DC, Grazioso A, Gaetani F, Mantriota G, Reina G. A smart demonstration unit for upper-limb myoelectric prostheses. Smart Health (2014) 2024 Jun;32:100481. [doi: 10.1016/j.smhl.2024.100481]
113. Burge TA, Jeffers JRT, Myant CW. Applying machine learning methods to enable automatic customisation of knee replacement implants from CT data. Sci Rep 2023 Feb 27;13(1):3317. [doi: 10.1038/s41598-023-30483-5] [Medline: 36849812]

XSL•FO
RenderX

114. Pandelani T, Nemavhola FJ, Subramaniyan AK. 3D-printed smart implants in orthopedic surgery. In: Digital Design and Manufacturing of Medical Devices and Systems: Springer; 2024:187-211. [doi: 10.1007/978-981-99-7100-8_10]
115. Burge TA, Munford MJ, Kechagias S, Jeffers JRT, Myant CW. Automating the customization of stiffness-matched knee implants using machine learning techniques. Int J Adv Manuf Technol 2023 Jun;126(7-8):3725-3737. [doi: 10.1007/s00170-023-11357-6]
116. Paterno L, Filosa M, Anselmino E, et al. Soft transfemoral prosthetic socket with EMG sensing and augmenting feedback: a case study. Research Square. Preprint posted online on May 19, 2023. [doi: 10.21203/rs.3.rs-2889998/v1]
117. Ramakrishna VAS, Chamoli U, Rajan G, Mukhopadhyay SC, Prusty BG, Diwan AD. Smart orthopaedic implants: a targeted approach for continuous postoperative evaluation in the spine. J Biomech 2020 May 7;104:109690. [doi: 10.1016/j.jbiomech.2020.109690] [Medline: 32139096]
118. Lin KC, Cheng IL, Huang YC, et al. The effects of the badminton teaching-assisted system using electromyography and gyroscope on learners' badminton skills. IEEE Trans Learning Technol 2023;16(5):780-789. [doi: 10.1109/TLT.2023.3292215]
119. Aminizadeh S, Heidari A, Dehghan M, et al. Opportunities and challenges of artificial intelligence and distributed systems to improve the quality of healthcare service. Artif Intell Med 2024 Mar;149:102779. [doi: 10.1016/j.artmed.2024.102779] [Medline: 38462281]
120. Thapa C, Camtepe S. Precision health data: requirements, challenges and existing techniques for data security and privacy. Comput Biol Med 2021 Feb;129:104130. [doi: 10.1016/j.compbiomed.2020.104130] [Medline: 33271399]
121. Ross P, Spates K. Considering the safety and quality of artificial intelligence in health care. Jt Comm J Qual Patient Saf 2020 Oct;46(10):596-599. [doi: 10.1016/j.jcjq.2020.08.002] [Medline: 32878718]
122. Bernstam EV, Shireman PK, Meric-Bernstam F, et al. Artificial intelligence in clinical and translational science: successes, challenges and opportunities. Clin Transl Sci 2022 Feb;15(2):309-321. [doi: 10.1111/cts.13175] [Medline: 34706145]
123. Choudhury A. Toward an ecologically valid conceptual framework for the use of artificial intelligence in clinical settings: need for systems thinking, accountability, decision-making, trust, and patient safety considerations in safeguarding the technology and clinicians. JMIR Hum Factors 2022 Jun 21;9(2):e35421. [doi: 10.2196/35421] [Medline: 35727615]
124. Antoniadi AM, Du Y, Guendouz Y, et al. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. Appl Sci (Basel) 2021;11(11):5088. [doi: 10.3390/app11115088]
125. Manias G, Op Den Akker H, Azqueta A, et al. IHELP: personalised health monitoring and decision support based on artificial intelligence and holistic health records. 2021 Presented at: 2021 IEEE Symposium on Computers and Communications (ISCC); Sep 5-8, 2021; Athens, Greece p. 1-8. [doi: 10.1109/ISCC53001.2021.9631475]
126. Larson DB, Magnus DC, Lungren MP, Shah NH, Langlotz CP. Ethics of using and sharing clinical imaging data for artificial intelligence: a proposed framework. Radiology 2020 Jun;295(3):675-682. [doi: 10.1148/radiol.2020192536] [Medline: 32208097]
127. Almalawi A, Khan AI, Alsolami F, Abushark YB, Alfakeeh AS. Managing security of healthcare data for a modern healthcare system. Sensors (Basel) 2023 Mar 30;23(7):3612. [doi: 10.3390/s23073612] [Medline: 37050672]
128. McCradden MD, Joshi S, Mazwi M, Anderson JA. Ethical limitations of algorithmic fairness solutions in health care machine learning. Lancet Digit Health 2020 May;2(5):e221-e223. [doi: 10.1016/S2589-7500(20)30065-0] [Medline: 33328054]
129. Chen RJ, Wang JJ, Williamson DFK, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat Biomed Eng 2023 Jun;7(6):719-742. [doi: 10.1038/s41551-023-01056-8] [Medline: 37380750]
130. Lorenzini G, Shaw DM, Arbelaez Ossa L, Elger BS. Machine learning applications in healthcare and the role of informed consent: ethical and practical considerations. Clin Ethics 2023 Dec;18(4):451-456. [doi: 10.1177/14777509221094476]
131. Maccaro A, Stokes K, Statham L, et al. Clearing the fog: a scoping literature review on the ethical issues surrounding artificial intelligence-based medical devices. J Pers Med 2024 Apr 23;14(5):443. [doi: 10.3390/jpm14050443] [Medline: 38793025]
132. Higgins DC, Johner C. Validation of artificial intelligence containing products across the regulated healthcare industries. Ther Innov Regul Sci 2023 Jul;57(4):797-809. [doi: 10.1007/s43441-023-00530-4] [Medline: 37202591]
133. Fraser AG, Biasin E, Bijnens B, et al. Artificial intelligence in medical device software and high-risk medical devices - a review of definitions, expert recommendations and regulatory initiatives. Expert Rev Med Devices 2023 Jun;20(6):467-491. [doi: 10.1080/17434440.2023.2184685] [Medline: 37157833]
134. Lewis D, Hogan L, Filip D, Wall PJ. Global challenges in the standardization of ethics for trustworthy AI. JICTS 2020;8(2):123-150. [doi: 10.13052/jicts2245-800X.823]
135. Winter JS, Davidson E. Harmonizing regulatory regimes for the governance of patient-generated health data. Telecomm Policy 2022 Jun;46(5):102285. [doi: 10.1016/j.telpol.2021.102285]
136. Federico CA, Trotsyuk AA. Biomedical data science, artificial intelligence, and ethics: navigating challenges in the face of explosive growth. Annu Rev Biomed Data Sci 2024 Aug;7(1):1-14. [doi: 10.1146/annurev-biodatasci-102623-104553] [Medline: 38598860]
137. Palaniappan K, Lin EYT, Vogel S. Global regulatory frameworks for the use of artificial intelligence (AI) in the healthcare services sector. Healthcare (Basel) 2024 Feb 28;12(5):562. [doi: 10.3390/healthcare12050562] [Medline: 38470673]

XSL•FO
RenderX

138. Mackenzie SC, Sainsbury CAR, Wake DJ. Diabetes and artificial intelligence beyond the closed loop: a review of the landscape, promise and challenges. Diabetologia 2024 Feb;67(2):223-235. [doi: 10.1007/s00125-023-06038-8] [Medline: 37979006]
139. Farouk A, Alahmadi A, Ghose S, Mashatan A. Blockchain platform for industrial healthcare: vision and future opportunities. Comput Commun 2020 Mar;154:223-235. [doi: 10.1016/j.comcom.2020.02.058]
140. Mahajan HB. Emergence of healthcare 4.0 and blockchain into secure cloud-based electronic health records systems: solutions, challenges, and future roadmap. Wireless Pers Commun 2022 Oct;126(3):2425-2446. [doi: 10.1007/s11277-022-09535-y]
141. Aggarwal L, Sachdeva S, Goswami P. Quantum healthcare computing using precision based granular approach. Appl Soft Comput 2023 Sep;144:110458. [doi: 10.1016/j.asoc.2023.110458]
142. Davids J, Lidströmer N, Ashrafian H. Artificial intelligence in medicine using quantum computing in the future of healthcare. In: Artificial Intelligence in Medicine: Springer International Publishing; 2022:423-446. [doi: 10.1007/978-3-030-64573-1_338]
143. Tagde P, Tagde S, Bhattacharya T, et al. Blockchain and artificial intelligence technology in e-health. Environ Sci Pollut Res Int 2021 Oct;28(38):52810-52831. [doi: 10.1007/s11356-021-16223-0] [Medline: 34476701]
144. Mbunge E, Muchemwa B, Jiyane S, Batani J. Sensors and healthcare 5.0: transformative shift in virtual care through emerging digital health technologies. Global Health Journal 2021 Dec;5(4):169-177. [doi: 10.1016/j.glohj.2021.11.008]

Abbreviations

AI: artificial intelligence
CGM: continuous glucose monitor
CNN: convolutional neural network
CT: computed tomography
FDA: US Food and Drug Administration
ML: machine learning
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR : Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Scoping Reviews

Edited by J Sarvestan; submitted 09.Feb.2025; peer-reviewed by G Wilson, M Al-Raeei, MM Khan; revised version received 17.Apr.2025; accepted 18.Apr.2025; published 04.Mar.2026.

Please cite as:
Ponnambalath Mohanadas H, Manikandan A, Ismail AF, Tucker N, Jaganathan SK
Artificial Intelligence Applications in Medical Devices for Personalized Health Care Solutions: Systematic Review
J Med Internet Res 2026;28:e72410
URL: https://www.jmir.org/2026/1/e72410
doi:10.2196/72410

© Hemanth Ponnambalath Mohanadas, A Manikandan, Ahmad Fauzi Ismail, Nick Tucker, Saravana Kumar Jaganathan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 4.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effectiveness of Interventions for Internet, Smartphone, and Gaming Addictions: Umbrella Review and Meta–Meta-Analysis

Minggang Zhang[1*], PhD; Tong Lan[2*], MSc; Tao Song[1], PhD; Xiaochun Wang[1], PhD

[1]School of Psychology, Shanghai University of Sport, 399 ChangHai Road, Shanghai, 200438, P.R. China, Shanghai, China

[2]Psychological Counseling Center, Nanchang Institute of Technology, Nanchang, China

[*]these authors contributed equally

Corresponding Author:
Xiaochun Wang, PhD
School of Psychology, Shanghai University of Sport, 399 ChangHai Road, Shanghai, 200438, P.R. China, Shanghai, China

Abstract

Background: Digital addiction, including internet, smartphone, and gaming addiction, has emerged as a significant global health concern. Although a wide range of interventions has been evaluated, the fragmented and siloed nature of existing meta-analyses limits a clear understanding of the comparative effectiveness of different interventions across addiction subtypes.

Objective: This umbrella review and meta–meta-analysis aimed to estimate the overall effectiveness of interventions for digital addiction and examine differential effects according to addiction subtype, intervention modality, study design, and control condition.

Methods: A systematic search of 5 electronic databases (PubMed, Web of Science, Scopus, APA PsycInfo, and the Cochrane Library) was conducted from inception to June 24, 2025. Eligible studies were systematic reviews with meta-analyses evaluating interventions for internet, smartphone, or gaming addiction. Random-effects models were applied to synthesize standardized mean differences (SMDs). Methodological quality and certainty of evidence were assessed using A Measurement Tool to Assess Systematic Reviews 2 and the Grading of Recommendations Assessment, Development, and Evaluation framework.

Results: A total of 29 meta-analyses, comprising 52 effect sizes and 66,530 participants, were included (I^2=95.13%). Overall, interventions demonstrated a large and statistically significant effect in reducing digital addiction symptoms (SMD=–1.44, 95% CI –1.67 to –1.21; *P*=.003). Subgroup analyses indicated that the largest effects were observed for internet addiction (SMD=–1.70, 95% CI –1.99 to –1.42), followed by gaming addiction (SMD=–0.82, 95% CI –1.09 to –0.56) and smartphone addiction (SMD=–0.80, 95% CI –1.39 to –0.21). Exercise-based interventions, particularly those integrated with psychological approaches, showed large effect sizes (SMD=–3.14, 95% CI –4.30 to –1.97); however, this finding was based on a very limited number of effect sizes and should be interpreted cautiously. In addition, randomized controlled trials yielded larger effects than mixed study designs, and no-intervention controls were associated with larger effect sizes than mixed control conditions. The certainty of evidence was generally low.

Conclusions: Interventions for digital addiction are effective, although their magnitude of benefit varies by addiction subtype and intervention modality. These findings support the use of tailored and multimodal intervention strategies while highlighting the need for more rigorous, high-quality, and balanced evidence across different forms of digital addiction.

Trial Registration: PROSPERO CRD420251084061; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251084061

(J Med Internet Res 2026;28:e81705) doi:10.2196/81705

KEYWORDS

internet addiction; smartphone addiction; gaming disorder; umbrella review; meta–meta-analysis; intervention effectiveness; PRISMA

Introduction

In the digital era, behavioral addictions associated with the excessive use of internet-based technologies have emerged as major public health concerns. A recent global meta-analysis synthesizing 504 prevalence studies involving more than 2 million individuals across 64 countries reported pooled prevalence rates of 14.2% for internet addiction, 26.99% for smartphone addiction, and 6.04% for gaming addiction, with pronounced increases observed during the COVID-19 pandemic [1]. Consistent with these findings, pandemic-restricted analyses have shown prevalence rates rising to 30.7% for smartphone addiction and 10.6% for internet addiction [2,3], underscoring a growing public health burden that spans age groups, sexes, and cultural contexts. The clinical relevance of this burden is increasingly reflected in international diagnostic frameworks

XSL•FO
RenderX

[4]. In 2019, the World Health Organization incorporated gaming disorder into the *International Classification of Diseases, 11th Revision* (*ICD-11*), formally assigning it to the "Disorders due to addictive behaviors" section [5]. Although internet and smartphone addiction have not yet been formally recognized within the *ICD-11*, accumulating evidence links all 3 conditions to emotional dysregulation, sleep disturbances, impaired academic and occupational functioning, and reduced quality of life [6-9].

Over the past decade, a broad range of interventions has been evaluated to address digital addictions, including cognitive behavioral therapy (CBT), group counseling, exercise training, mindfulness-based approaches, and multicomponent programs. A number of meta-analyses have synthesized evidence on the effectiveness of these interventions in reducing symptoms of internet, gaming, and smartphone addiction. For example, a meta-analysis of 59 randomized trials reported large overall effects for interventions targeting internet addiction and identified combined CBT and exercise interventions as particularly effective [10]. Similarly, exercise-based interventions have demonstrated substantial reductions in smartphone dependence among children and adolescents, with optimal effects observed for moderate-intensity aerobic exercise performed approximately 3 times per week over a 12-week period [11]. In the context of gaming disorder, a synthesis of 38 studies involving 9524 participants reported a moderate but highly heterogeneous treatment effect [12].

Despite these encouraging findings, the evidence base for interventions targeting digital addictions remains fragmented. Most meta-analyses to date focus exclusively on a single type of addiction, such as internet use, smartphone dependence, or gaming disorder, while adopting diverse outcome measures, inconsistent risk-of-bias frameworks, and varying inclusion criteria [12-14]. This fragmentation limits the comparability of intervention effects across different behavioral addictions and hinders the development of unified, evidence-based clinical guidelines. Consequently, a higher-level synthesis is needed to systematically integrate and compare findings across multiple meta-analyses.

Umbrella reviews, which synthesize evidence from multiple systematic reviews and meta-analyses, provide a robust methodological framework for addressing these challenges [15]. Although recent umbrella reviews have made initial progress, important gaps remain. For instance, one umbrella review pooled 5 meta-analyses addressing digital addiction, encompassing internet, smartphone, and gaming addiction and general screen use. While cognitive behavioral and exercise-based interventions emerged as promising, the study did not differentiate effects by addiction subtype or conduct quantitative meta–meta-analysis comparisons, limiting conclusions regarding relative effectiveness [16]. Another umbrella review focused on internet and smartphone addiction and confirmed potential benefits of psychotherapeutic and exercise-based interventions; however, it similarly lacked quantitative comparisons across addiction categories or intervention modalities [17].

To address these limitations, this study conducted an umbrella review and meta–meta-analysis of interventions targeting internet, smartphone, and gaming addictions. Specifically, we aimed to (1) quantify the overall effectiveness of major intervention modalities on addiction subtypes and (2) compare effect sizes across internet, smartphone, and gaming addictions. By providing a comparative, high-level synthesis of the current intervention landscape, this study sought to inform clinicians, researchers, and policymakers working to mitigate the psychosocial and health consequences of digital-age behavioral addictions.

Methods

Protocol and Registration

This umbrella review was preregistered in PROSPERO (CRD420251084061). This review was conducted and reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement [18].

Search Strategy

Five electronic databases (PubMed, Web of Science, Scopus, APA PsycInfo, and the Cochrane Library) were systematically searched from inception to June 24, 2025. The search strategy included a combination of 4 key blocks of terms related to internet addiction, intervention, and meta-analyses. Table S1 in Multimedia Appendix 1 provides our electronic search strategy.

Eligibility Criteria

We included systematic reviews with a meta-analysis examining the effects of interventions on internet, smartphone, or gaming addiction. For addiction subtype classification, studies were categorized as targeting internet addiction, smartphone addiction, or gaming addiction based on the primary addiction construct explicitly defined by the authors, regardless of the device used (eg, gaming on a smartphone). No reclassification of primary studies was performed to avoid introducing additional subjectivity.

Specific inclusion criteria were also defined using the population, intervention, comparator, outcome, and study design framework as follows:

- Population—reviews were eligible if they included studies that enrolled individuals with addictive or excessive or problematic internet or smartphone use or gaming.
- Intervention—reviews were eligible if they included studies that investigated any type of psychological, behavioral, or physical activity intervention, among other types, that was implemented for individuals as an addiction intervention. Studies evaluating pharmacological interventions were not eligible for inclusion.
- Comparator—reviews were eligible if they included controlled studies in which the addiction intervention was compared with any type of control group.
- Outcomes—reviews were eligible if they included studies that reported the results of at least 1 meta-analysis of an intervention for internet, smartphone, or gaming addiction at a postintervention time point vs controls. The outcome measures needed to use valid and reliable rating scales suitable for internet, smartphone, or gaming addiction.

XSL•FO
RenderX

When more than one measure was described by a study, the primary outcome was used. If the primary outcome was not specified, the data from the measure reported first were extracted. Meta-analyses without any effect size were excluded.

- Study design—reviews were eligible if they assessed quantitative randomized controlled trials (RCTs) only and if they assessed a mixed evidence base (RCTs with other designs, such as non-RCTs and quasi-experimental studies). Reviews were included if they were published in English in peer-reviewed journals. Conferences, theses, and commentaries were not included.

Data Extraction

Study characteristics and outcome measures were extracted by MZ and TL. Data entry was independently checked by TS after the data were entered into a Microsoft Excel spreadsheet. The following data from two domains were extracted: (1) study characteristics (author, year of publication, country, mean age of participants, number of included studies, and sample size [percentage of participants by gender]) and (2) interventions (addiction types, intervention types, control group conditions, and study design). All the extracted data were combined with those from the original review.

Evaluations of Methodological Quality and Quality of Evidence

The methodological quality of eligible systematic reviews with meta-analyses was evaluated using a critical appraisal tool: A Measurement Tool to Assess Systematic Reviews 2 (AMSTAR 2) [19]. AMSTAR 2 contains 16 checklist items rated as "yes" (1 point), "partial yes" (0.5 points), and "no" (0 points) and is used to identify the risk of bias for systematic reviews that involve randomized and/or nonrandomized intervention studies. Among the AMSTAR 2 items, 7 are considered "critical" and 9 are considered "noncritical." Reviews were scored as follows: "critically low" (>1 critical weakness and 3 noncritical weaknesses), "low" (>1 critical weakness and <3 noncritical weaknesses), "moderate" (1 critical weakness and <3 noncritical weaknesses), or "high" (no critical weaknesses and <3 noncritical weaknesses).

The certainty of evidence was assessed using the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach [20]. Five domains were considered: risk of bias, inconsistency, indirectness, imprecision, and publication bias. Each outcome of internet-related addiction was assigned a certainty rating of "high," "moderate," "low," or "very low" based on these criteria. Two authors independently performed the AMSTAR 2 and GRADE evaluations, with the third author solving any discrepancies through consensus.

Statistical Analysis

Calculation of Effect Sizes

A meta-analysis was performed in this meta-review through Stata/MP (version 18.0; StataCorp) to calculate individual study and pooled effect sizes. To synthesize the findings on the effectiveness of interventions for internet, smartphone, and gaming addictions, meta-analyses were performed by combining the effect sizes and 95% CIs reported in each review via a random-effects model. Standardized effect sizes (standardized mean differences; SMDs) with 95% CIs were used as the effect measure for the meta-analyses. For consistency across studies, all effect sizes were converted such that negative values reflected beneficial intervention effects. All statistical tests were 2 tailed, and statistical significance was set at a P value of less than .05.

Testing Homogeneity and Potential Publication Bias

The heterogeneity of effect sizes was tested using the Cochran Q test and I^2 statistic, which indicates heterogeneity in percentages [21]. Significant heterogeneity was deemed present if the P value of the Q test was less than .10, and 0% indicated no heterogeneity, whereas 25%, 50%, and 75% indicated low, moderate, and high heterogeneity, respectively. To consider potential publication bias, a visual inspection of a funnel plot was conducted, and we also applied the nonparametric trim-and-fill method using a random-effects model. The method was implemented using a linear estimator with imputation on the right side of the funnel plot, reflecting a scenario in which studies with smaller or nonsignificant (less negative) effect sizes may be underrepresented.

Subgroup Analyses

For subgroup analyses, we extracted subgroup-specific effect sizes and corresponding data as reported in the included systematic reviews and pooled these results across reviews. As in the original review, the subgroup analyses were planned according to addiction type, intervention type, control group type, and study design. We used a random-effects analysis to estimate the effect size and compare subgroups.

Results

Search Results and Study Characteristics

A total of 1719 studies were identified through the database searches, and after the titles and abstracts were initially checked, 49 (2.9%) of the studies were included for full-text screening. We ultimately included 29 review and meta-analysis studies that met the full inclusion criteria. The PRISMA flowchart and checklist are presented in Figure 1 and Checklist 1, respectively, and the list of reasons for exclusion after full-text review is presented in Table S2 in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.

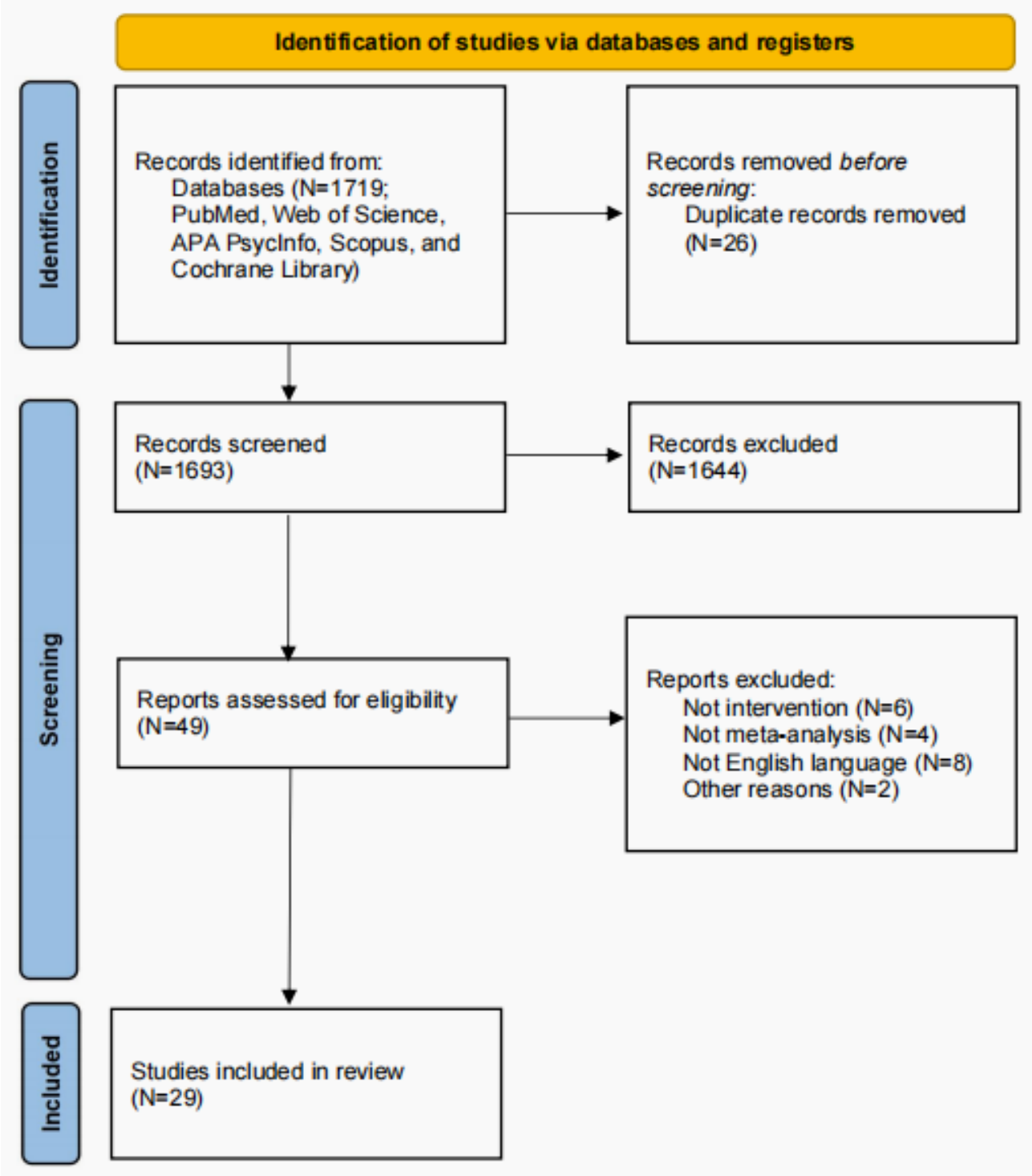

The eligible studies were published from 2017 to 2025 and were conducted in China (15/29, 51.7%), South Korea (6/29, 20.7%), Austria (2/29, 6.9%), Australia (1/29, 3.4%), Lithuania (1/29, 3.4%), Spain (1/29, 3.4%), the United States (1/29, 3.4%), Norway (1/29, 3.4%), and Turkey (1/29, 3.4%). Of these 29 studies, 18 (62.1%) assessed the intervention effect on internet addiction by calculating 38 effect sizes, 8 (27.6%) assessed the intervention effect on internet gaming addiction by calculating 10 effect sizes, and 4 (13.8%) assessed the intervention effect on smartphone addiction by calculating 4 effect sizes. A total of 31 effect sizes were calculated to assess psychological interventions for overall digital addiction, 14 effect sizes were calculated to assess exercise-based interventions, and 7 effect sizes were calculated to assess other types of interventions. In total, 41.4% (12/29) of the studies conducted a meta-analysis of RCTs, 58.6% (17/29) of the studies assessed mixed study designs, 24.1% (7/29) of the studies specifically assessed no-intervention controls, and 75.9% (22/29) of the studies assessed mixed controls. A summary of the participant demographics is provided in Table S3 in Multimedia Appendix 1.

XSL•FO
RenderX

Regarding the methodological quality and quality of evidence of all included studies, the percentages of studies with low and critically low methodological quality were 13.8% (4/29) and 86.2% (25/29), respectively, according to the AMSTAR 2 score (Table S4 in Multimedia Appendix 1). The percentages of reviews with moderate, low, and very low quality of evidence were 3.4% (1/29), 20.7% (6/29), and 75.9% (22/29), respectively, according to the GRADE score (Table S5 in Multimedia Appendix 1).

The visual and statistical results indicated significant asymmetry in the funnel plot (β=–3.00; *P*<.001), suggesting potential publication bias or small-study effects (Figure 2). The trim-and-fill procedure identified 14 potentially missing studies, resulting in a total of 66 studies (n=52, 78.8% observed; n=14, 21.2% imputed; Figure S1 in Multimedia Appendix 1). The observed pooled effect size was SMD=–1.439 (95% CI –1.67 to –1.21). After imputing the potentially missing studies, the adjusted pooled effect size was SMD=–1.09 (95% CI –1.38 to –0.81). Although the adjusted effect size was smaller in magnitude, it remained statistically significant, indicating a robust intervention effect despite potential publication bias.

Figure 2. Funnel plot of all included effect sizes. REML: random effect model; SMD: standardized mean difference.

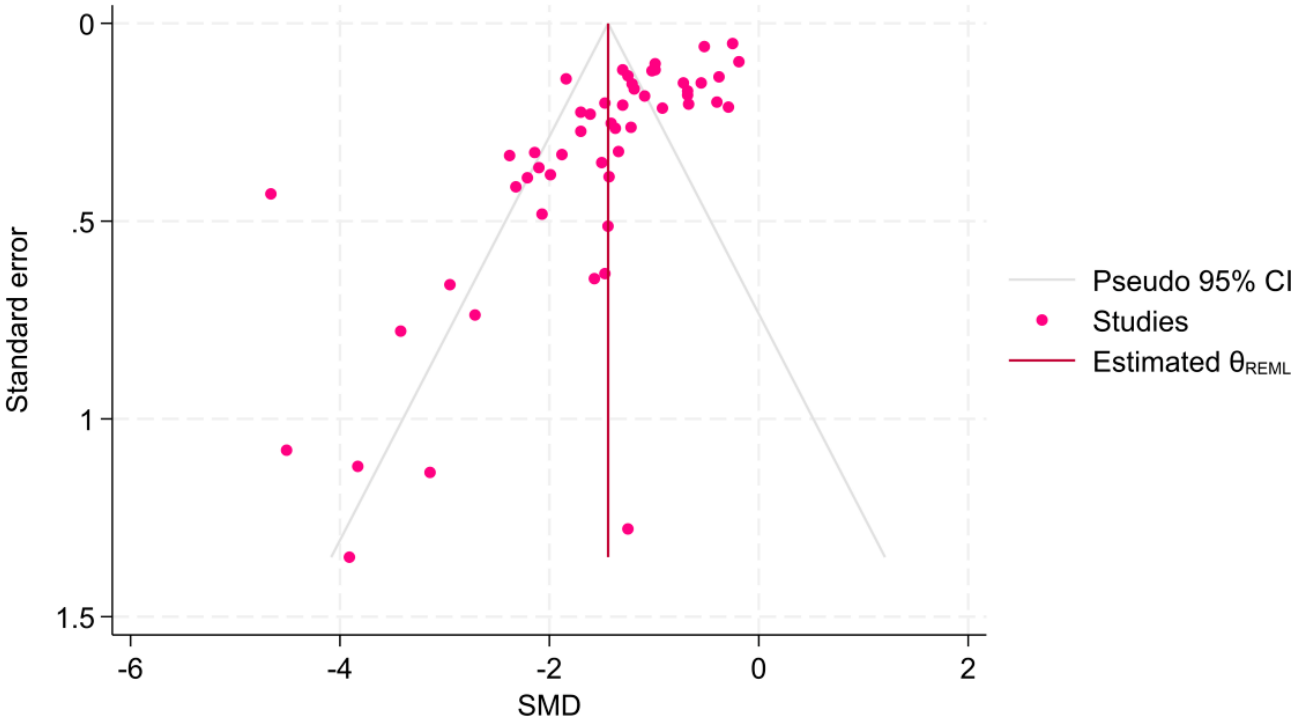

Meta–Meta-Analysis Results

The overall analysis revealed that interventions significantly reduced digital addiction symptoms, with an SMD=–1.44 (95% CI –1.67 to –1.21; *P*=.003; I^2=95.13%; Figure 3 [10,12,13,22-47]). Table 1 summarizes the results of the overall and subgroup meta-analyses. For different addiction types, the subgroup analysis revealed a significant effect (Q_b=21.43; *df*=2; *P*=.003), and the interventions had the largest reducing effect on internet addiction (SMD=–1.70, 95% CI –1.99 to –1.42), followed by internet gaming addiction (SMD=–0.82, 95% CI –1.09 to –0.56) and smartphone addiction (SMD=–0.80, 95% CI –1.39 to –0.21).

XSL•FO
RenderX

Figure 3. Forest plot for overall addiction and subaddictions [10,12,13,22-47]. SMD: standardized mean difference.

XSL•FO
RenderX

Table . Results of overall and subgroup meta-analyses.

	Number of effects	SMD[a] (95% CI)	I^2 (%)	Test of subgroup differences	
				Q_b (*df*)	*P* value
Overall	52	–1.44 (–1.67 to –1.21)	95.13	—[b]	—
Subgroup analysis of all data					
Addiction				21.43 (2)	.003
Internet addiction	38	–1.7 (–1.99 to –1.42)	93.96		
Internet gaming addiction	10	–0.82 (–1.09 to –0.56)	88.43		
Smartphone addiction	4	–0.8 (–1.39 to –0.21)	86.9		
Intervention				58.68 (5)	.002
Psychological	31	–1.16 (–1.38 to –0.95)	92.53		
Exercise	14	–2.01 (–2.52 to –1.50)	93.9		
Exercise+psychological	2	–3.14 (–4.30 to –1.97)	—		
Biofeedback	2	–2.7 (–5.91 to 0.51)	88.6		
Behavioral	2	–0.46 (–0.65 to –0.26)	—		
Electrotherapy	1	–1.25 (–3.75 to 1.25)	—		
Study design				11.89 (1)	.008
RCT[c]	29	–1.83 (–2.20 to –1.46)	91.7		
Mixed	23	–1.06 (–1.29 to –0.82)	94.05		
Control				17.97 (1)	.007
No intervention	16	–2.45 (–3.04 to –1.86)	85.65		
Mixed	36	–1.12 (–1.30 to –0.93)	92.12		
Subgroup analysis for internet addiction					
Intervention				13.33 (4)	.01
Psychological	20	–1.34 (–1.63 to –1.06)	91.6		
Exercise	13	–2.08 (–2.63 to –1.53)	92.87		
Exercise+psychological	2	–3.14 (–4.30 to –1.97)	—		
Biofeedback	2	–2.7 (–5.91 to 0.51)	88.6		
Electrotherapy	1	–1.25 (–3.75 to 1.25)	—		
Control				12.44 (1)	.006
No intervention	16	–2.45 (–3.04 to –1.86)	85.65		
Mixed	22	–1.31 (–1.54 to –1.08)	90.41		
Study design				9.22 (1)	.005
RCT	24	–2.05 (–2.45 to –1.65)	88.14		
Mixed	14	–1.24 (–1.57 to –0.92)	94.13		
Subgroup analysis for gaming addiction					
Intervention				6.43 (1)	.01
Psychological	8	–0.93 (–1.24 to –0.62)	88.37		
Behavioral	2	–0.46 (–0.65 to –0.26)	—		
Study design				0.36 (1)	.55
RCT	3	–1.02 (–1.75 to –0.28)	83.03		

XSL•FO
RenderX

	Number of effects	SMD[a] (95% CI)	I^2 (%)	Test of subgroup differences	
				Q_b (*df*)	*P* value
Mixed	7	–0.77 (–1.06 to –0.49)	88.64		
Subgroup analysis for smartphone addiction					
Intervention				23.60 (1)	.005
Psychological	3	–0.41 (–0.68 to –0.13)	—		
Behavioral	1	–1.3 (–1.53 to –1.07)	—		
Study design				0.00 (1)	.99
RCT	2	–0.81 (–1.80 to 0.18)	94.26		
Mixed	2	–0.82 (–1.93 to 0.28)	66.68		

[a]SMD: standardized mean difference.

[b]Not applicable.

[c]RCT: randomized controlled trial.

For the intervention types, the overall subgroup analysis revealed significant differences in efficacy among the included interventions ($Q_b(5)=58.68$; $P=.002$). Although only 2 effect sizes were calculated to assess the efficacy of a psychological intervention plus exercise, this intervention had the greatest effect on overall addiction (SMD=–3.14, 95% CI –4.30 to –1.97), followed by biofeedback (SMD=–2.70, 95% CI –5.91 to 0.51) and exercise (SMD=–2.01, 95% CI –2.52 to –1.50) interventions. Specifically, the interventions also had significant effects on reducing addictive symptoms for internet addiction ($Q_b(4)=13.33$; $P=.01$), internet gaming addiction ($Q_b(1)=6.43$; $P=.01$), and smartphone addiction ($Q_b(1)=23.60$; $P=.005$). For internet addiction, the intervention with the largest reducing effect was psychological plus exercise (SMD=–3.14, 95% CI –4.30 to –1.97), followed by biofeedback (SMD=–2.70, 95% CI –5.91 to 0.51) and exercise (SMD=–2.08, 95% CI –2.63 to –1.53) interventions. For internet gaming addiction, only 2 types of interventions were assessed. Compared with those of the behavioral intervention, the larger effect sizes of the psychological intervention indicated that it had greater effects on reducing addictive symptoms (SMD=–0.93, 95% CI –1.24 to –0.62). Finally, for smartphone addiction, only 3.4% (1/29) of the studies assessed a behavioral intervention and reported greater effects than those of a psychological intervention (SMD=–0.41, 95% CI –0.68 to –0.31).

With respect to study design, the subgroup analysis revealed that RCTs found a significantly greater effect in reducing overall addictive symptoms (SMD=–1.83, 95% CI –2.20 to –1.46) and internet addiction symptoms (SMD=–2.05, 95% CI –2.25 to –1.65) but not smartphone addiction ($Q_b(1)=0.00$; $P=.99$) or internet gaming addiction ($Q_b(1)=0.36$; $P=.55$) symptoms.

For the control conditions, the subgroup analysis revealed that a no-intervention control had a significant and greater effect on reducing overall and internet addiction with the same included effect sizes (SMD=–2.45, 95% CI –3.04 to –1.86) than a mixed control. No subgroup analyses were conducted for internet gaming addiction or smartphone addiction, as only 1 control was included.

The results of the subgroup analyses for different addiction types, intervention types, study designs, and control group types are presented as forest plots in Figures S2 to S11 in Multimedia Appendix 1.

Discussion

Principal Findings

In this umbrella review and meta–meta-analysis, evidence from 29 systematic reviews and meta-analyses was synthesized to evaluate the effectiveness of interventions targeting internet, smartphone, and gaming addictions. Overall, interventions demonstrated large and statistically significant effects in reducing digital addiction symptoms. Among the 3 addiction subtypes, interventions targeting internet addiction yielded the largest pooled effect, whereas those targeting gaming and smartphone addiction produced more modest but still clinically meaningful reductions. These findings support the overall efficacy of interventions for digital behavioral addictions and suggest that responsiveness to intervention may differ across addiction subtypes. Notably, to our knowledge, this study represents the first quantitative umbrella synthesis that directly compares effect sizes across different categories of digital addiction, thereby providing a more granular and comparative perspective on intervention effectiveness.

Our findings indicate clear differences in intervention effects across addiction subtypes. Interventions for internet addiction showed the largest pooled effect (SMD=–1.70), followed by interventions for gaming addiction (SMD=–0.82) and smartphone addiction (SMD=–0.80). These results are broadly consistent with those of prior meta-analyses reporting moderate effects for gaming disorder and smartphone addiction [11,12], while extending previous work by demonstrating the comparatively greater responsiveness of internet addiction to intervention. One potential explanation is that internet addiction, often conceptualized as a broader and more heterogeneous construct, encompasses diverse behaviors such as social media use, web browsing, and streaming, which may be more amenable to general behavioral regulation strategies [48,49]. In contrast, gaming and smartphone addiction may involve more entrenched

XSL•FO
RenderX

reward-based and habit-forming mechanisms frequently embedded within immersive or socially reinforced contexts, thereby posing greater challenges for behavior change [50,51]. These distinctions underscore the importance of tailoring intervention strategies to the specific behavioral and cognitive characteristics of each addiction subtype.

With respect to intervention modalities, our results highlight a notable and often underemphasized finding. Exercise-based interventions, both as stand-alone approaches and in combination with psychological therapies, were associated with particularly large effect sizes. The combined psychological and exercise intervention demonstrated the largest pooled effect (SMD=–3.14), followed by exercise-only interventions (SMD=–2.01) and biofeedback-based approaches (SMD=–2.70). This pattern diverges from much of the existing literature, which has predominantly emphasized cognitive behavioral and psychotherapeutic interventions [17]. Nevertheless, our findings are consistent with emerging evidence supporting the benefits of physical activity for reducing internet and smartphone addiction symptoms [14,22]. Potential mechanisms underlying these effects include improvements in executive functioning, stress regulation, emotional control, and sleep quality [52,53]. At the same time, the strong performance of psychological-only interventions reinforces their established role in treatment and suggests that psychological and exercise-based modalities may exert complementary effects.

Subgroup analyses further revealed substantial heterogeneity in intervention effects not only across addiction subtypes but also within each subtype, depending on intervention modality, study design, and control condition. For internet addiction, combined psychological and exercise-based interventions yielded the largest effects. In contrast, behavioral interventions appeared more effective than psychological interventions for smartphone addiction, aligning with prior evidence regarding the role of physical activity in enhancing self-regulation and reducing screen-based behaviors [14,22]. For gaming addiction, relatively larger effects were observed for psychological interventions compared with behavioral approaches. These patterns support the increasingly accepted view that different forms of digital addiction may be underpinned by distinct neurocognitive and behavioral profiles and, therefore, require intervention strategies tailored to the specific structure and reinforcement mechanisms of each disorder [51].

Beyond intervention modality, methodological characteristics, including study design and control condition, emerged as important moderators of treatment effects. Across all 3 addiction subtypes, RCTs yielded larger effect sizes than non-RCT or mixed designs, with this difference being most pronounced for internet addiction. This finding reinforces the importance of experimental rigor in evaluating intervention efficacy [15,54]. Regarding control conditions, studies using no-intervention or waitlist controls consistently reported larger effect sizes than those using active or mixed control groups, particularly for internet and smartphone addiction. While this pattern may partly reflect expectancy effects or the absence of placebo balancing, it also raises concerns about potential effect size inflation in less rigorously controlled designs [55,56]. Similar trends have been documented in psychotherapy research, where studies using no-treatment controls often report larger effects than those using active or placebo comparators [57]. These observations highlight the necessity of incorporating well-designed active or attention-matched control conditions in future trials to enhance interpretability and comparability.

Although the meta–meta-analytic findings provide strong support for the effectiveness of interventions, caution is warranted when interpreting the apparent superiority of certain intervention types. In this synthesis, interventions for internet addiction accounted for a disproportionately large share of the available effect sizes, whereas smartphone addiction and gaming addiction were represented by substantially fewer meta-analyses and primary studies. This imbalance likely reflects historical research priorities and differences in diagnostic recognition, particularly given the relatively recent inclusion of gaming disorder in the *ICD-11* [5,58]. However, such disparities limit the robustness of cross-subtype comparisons. Specifically, the smaller evidence base for smartphone and gaming addiction may reduce statistical power, increase sensitivity to individual study characteristics, and obscure true intervention effects. Consequently, the comparatively lower pooled effects observed for these subtypes should not be interpreted as definitive evidence of inferior treatment efficacy. Addressing this limitation will require greater investment in high-quality, subtype-specific intervention research.

Several contextual factors should also be considered when interpreting these findings. First, most included meta-analyses were conducted in middle- and high-income countries, which may limit the generalizability of the results to settings affected by the digital divide, where access to digital technologies and intervention resources differs substantially. Second, reverse causation cannot be fully excluded, as individuals with more severe addiction symptoms may be more likely to participate in intervention studies, potentially inflating observed effects, particularly in non-RCT designs. Moreover, intervention effectiveness may vary by gender, age, and regional context; however, most included reviews did not report stratified analyses for these variables. Future studies should address these gaps to improve the external validity of digital addiction interventions.

Conceptual overlap among digital addiction constructs should also be considered. In practice, behaviors such as gaming may occur on multiple platforms, including smartphones and computers, which may blur the boundaries among internet, smartphone, and gaming addiction. This overlap could partially attenuate observed differences in intervention effectiveness across addiction subtypes and should be taken into account when interpreting comparative results. Future research should also establish a more detailed distinction among these potential impacts. Finally, the substantial heterogeneity observed in the overall analysis (I^2=95.13%) likely reflects multiple sources of variability beyond intervention type. These include differences in diagnostic criteria and measurement instruments, variation in symptom severity and age composition of samples, and cultural and regional differences in digital use patterns. Additionally, the inclusion of both RCTs and mixed study designs may have further contributed to heterogeneity. Such

XSL•FO
RenderX

variability underscores the need for cautious interpretation of pooled estimates.

This study has several key strengths. First, it is the first umbrella review and meta–meta-analysis to systematically compare the effectiveness of interventions across 3 distinct digital addiction types, addressing a major gap in prior literature. Second, it evaluated multiple intervention modalities, offering a broader understanding of the treatment landscape beyond the dominant CBT-centric models. Third, by applying rigorous tools such as AMSTAR 2 and GRADE, it appraised not only effect sizes but also the credibility of the evidence. Notably, the use of these tools revealed that all of the included meta-analyses were rated as low or critically low in methodological quality, and most effect estimates were graded as low or very low in certainty. This outcome reflects the stringent criteria of both appraisal systems. AMSTAR 2 places strong emphasis on protocol registration, risk-of-bias assessment in primary studies, and reporting transparency—criteria often underreported in existing meta-analyses [19]. Similarly, GRADE tends to downgrade evidence derived from observational studies or indirect comparisons and penalizes for inconsistency, imprecision, and publication bias, which were common across the included reviews. While these findings do not negate the observed intervention effects, they underscore the need to improve the methodological transparency and consistency of future meta-analyses in this domain [59].

Nevertheless, several limitations must be acknowledged. All of the included meta-analyses were rated as low or critically low in methodological quality, and the overall certainty of evidence was predominantly classified as very low. Substantial heterogeneity and indications of publication bias further limit the interpretability and generalizability of the findings. In addition, studies focusing on gaming and smartphone addiction were underrepresented relative to those addressing internet addiction, and several intervention approaches were supported by only a small number of studies. These limitations highlight the need for more rigorous, targeted, and subtype-specific intervention research. Finally, the larger effect sizes observed in studies using no-intervention controls are consistent with prior psychotherapy research and likely reflect expectancy and placebo-related effects. Future intervention trials should prioritize the use of active or attention-matched control conditions to reduce bias and improve the interpretability of treatment effects.

Despite these limitations, the findings of this study have important implications for clinical and applied practice. Given the comparatively strong effects observed for integrated interventions, particularly those combining psychological therapy with exercise, clinicians and educators may consider adopting multimodal treatment strategies, especially in school-based or youth-oriented settings. Moreover, the observed benefits of stand-alone exercise and biofeedback interventions underscore the potential value of nonpharmacological, resource-efficient approaches that may be scalable in low-resource or community contexts. Tailoring interventions to specific addiction subtypes, developmental stages, and cultural backgrounds while maintaining rigorous study designs and standardized outcome measures may further enhance the effectiveness and ecological validity of future digital addiction interventions.

Conclusions

This umbrella review demonstrates that interventions targeting internet, smartphone, and gaming addictions are generally effective. Integrated psychological-exercise interventions appear particularly promising; however, the current evidence remains preliminary due to the limited number of available studies. High-quality evidence is especially scarce for smartphone and gaming addictions. Future research should prioritize methodological rigor, expand intervention trials for underrepresented addiction subtypes, and develop tailored strategies based on distinct addiction profiles. Collectively, these findings provide an evidence-based foundation for designing more precise, scalable, and context-sensitive interventions in both clinical and educational settings.

Acknowledgments

Generative artificial intelligence was not used to draft any portion of this manuscript.

Funding

This work was supported by the Ministry of Education of the People’s Republic of China (grant 21JZDW007), the Shanghai Municipal Education Commission “Primary and Secondary School Kindergarten Student Care Project Research” special project (Runxin Project), and the China Scholarship Council (grant 202408310317).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

MZ and TL were responsible for the conception, design, and hypothesis. MZ, TL, and TS conducted the systematic search, data extraction, and assessment of risk of bias. MZ and TL completed all statistical analyses. MZ, TL, and XW contributed to the drafting of the manuscript and its revisions, including the final version. XW supervised the work. All authors read and approved a draft of the manuscript and the final version.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy; Excluded studies; characteristics of the included studies; A Measurement Tool to Assess Systematic Reviews 2; and Grading of Recommendations Assessment, Development, and Evaluation tool.
[DOC File, 519 KB - jmir_v28i1e81705_app1.doc]

Checklist 1
PRISMA checklist.
[DOCX File, 23 KB - jmir_v28i1e81705_app2.docx]

References

1. Meng SQ, Cheng JL, Li YY, et al. Global prevalence of digital addiction in general population: a systematic review and meta-analysis. Clin Psychol Rev 2022 Mar;92:102128. [doi: 10.1016/j.cpr.2022.102128] [Medline: 35150965]
2. Liu X, Gui Z, Chen ZM, et al. Global prevalence of internet addiction among university students: a systematic review and meta-analysis. Curr Opin Psychiatry 2025 May 1;38(3):182-199. [doi: 10.1097/YCO.0000000000000994] [Medline: 40009750]
3. Alimoradi Z, Lotfi A, Lin CY, Griffiths MD, Pakpour AH. Estimation of behavioral addiction prevalence during COVID-19 pandemic: a systematic review and meta-analysis. Curr Addict Rep 2022;9(4):486-517. [doi: 10.1007/s40429-022-00435-6] [Medline: 36118286]
4. Olson JA, Sandra DA, Colucci É, et al. Smartphone addiction is increasing across the world: a meta-analysis of 24 countries. Comput Human Behav 2022 Apr;129:107138. [doi: 10.1016/j.chb.2021.107138]
5. Brand M, Antons S, Bőthe B, et al. Current advances in behavioral addictions: from fundamental research to clinical practice. Am J Psychiatry 2025 Feb 1;182(2):155-163. [doi: 10.1176/appi.ajp.20240092] [Medline: 39659159]
6. Amin MH, Alhadi HA, Mohamed AE, et al. Impact of smartphone addiction on health status, mental well-being, and sleep quality among medical students in Sudan. BMC Psychiatry 2024 Dec 31;24(1):965. [doi: 10.1186/s12888-024-06377-9] [Medline: 39741229]
7. Ye XL, Zhang W, Zhao FF. Depression and internet addiction among adolescents: a meta-analysis. Psychiatry Res 2023 Aug;326:115311. [doi: 10.1016/j.psychres.2023.115311] [Medline: 37348449]
8. Alimoradi Z, Lin CY, Broström A, et al. Internet addiction and sleep problems: a systematic review and meta-analysis. Sleep Med Rev 2019 Oct;47:51-61. [doi: 10.1016/j.smrv.2019.06.004] [Medline: 31336284]
9. Sun Y, Wang Z, Liu T. Association of internet addiction with psychiatric symptom levels and sleep disorders: a systematic review and meta-analysis. Front Psychol 2025 Apr 17;16:1573058. [doi: 10.3389/fpsyg.2025.1573058] [Medline: 40313890]
10. Zhang X, Zhang J, Zhang K, et al. Effects of different interventions on internet addiction: a meta-analysis of random controlled trials. J Affect Disord 2022 Sep 15;313:56-71. [doi: 10.1016/j.jad.2022.06.013] [Medline: 35760188]
11. Chen L, Ren Y, Zhu P, Yang Y, Zhu F. Meta-analysis on intervention effects of exercise on Chinese children and adolescents with mobile phone dependence. Front Psychol 2025 Apr 11;16:1487944. [doi: 10.3389/fpsyg.2025.1487944] [Medline: 40290548]
12. Danielsen PA, Mentzoni RA, Låg T. Treatment effects of therapeutic interventions for gaming disorder: a systematic review and meta-analysis. Addict Behav 2024 Feb;149:107887. [doi: 10.1016/j.addbeh.2023.107887] [Medline: 37826910]
13. Zhu Y, Chen H, Li J, Mei X, Wang W. Effects of different interventions on internet addiction: a systematic review and network meta-analysis. BMC Psychiatry 2023 Dec 8;23(1):921. [doi: 10.1186/s12888-023-05400-9] [Medline: 38066462]
14. Li Z, Xia X, Sun Q, Li Y. Exercise intervention to reduce mobile phone addiction in adolescents: a systematic review and meta-analysis of randomized controlled trials. Front Psychol 2023 Dec 13;14:1294116. [doi: 10.3389/fpsyg.2023.1294116] [Medline: 38192396]
15. Gates M, Gates A, Pieper D, et al. Reporting guideline for overviews of reviews of healthcare interventions: development of the PRIOR statement. BMJ 2022 Aug 9;378:e070849. [doi: 10.1136/bmj-2022-070849] [Medline: 35944924]
16. Lu P, Qiu J, Huang S, et al. Interventions for digital addiction: umbrella review of meta-analyses. J Med Internet Res 2025 Feb 11;27:e59656. [doi: 10.2196/59656] [Medline: 39933164]
17. Yang H, Guo H, Zhu Z, et al. Intervention of internet addiction and smartphone addiction: an umbrella review of systematic reviews and meta-analyses. Curr Addict Rep 2023;11(9):1-24. [doi: 10.1007/s40429-023-00536-w]
18. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
19. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017 Sep 21;358:j4008. [doi: 10.1136/bmj.j4008] [Medline: 28935701]

XSL•FO
RenderX

20. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011 Apr;64(4):383-394. [doi: 10.1016/j.jclinepi.2010.04.026] [Medline: 21195583]
21. Cumpston M, Li T, Page MJ, et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane Handbook for Systematic Reviews of Interventions. Cochrane Database Syst Rev 2019 Oct 3;10(10):ED000142. [doi: 10.1002/14651858.ED000142] [Medline: 31643080]
22. Zhou Z, Wan Y, Li C, et al. Effectiveness of sports intervention: a meta-analysis of the effects of different interventions on adolescent internet addiction. J Affect Disord 2024 Nov 15;365:644-658. [doi: 10.1016/j.jad.2024.08.064] [Medline: 39147163]
23. Augner C, Vlasak T, Aichhorn W, Barth A. Tackling the “digital pandemic”: The effectiveness of psychological intervention strategies in problematic Internet and smartphone use-A meta-analysis. Aust N Z J Psychiatry 2022 Mar;56(3):219-229. [doi: 10.1177/00048674211042793] [Medline: 34477009]
24. Chun JS, Shim HS, Kim S. A meta-analysis of treatment interventions for internet addiction among Korean adolescents. Cyberpsychol Behav Soc Netw 2017 Apr;20(4):225-231. [doi: 10.1089/cyber.2016.0188]
25. Gao W, Zhou H, Zou X, Li M, Lin X. The impact of exercise and exercise-based combined therapies on adolescent internet addiction: a systematic review and network meta-analysis. Curr Psychol 2025 Apr;44(8):7095-7110. [doi: 10.1007/s12144-025-07719-w]
26. Goslar M, Leibetseder M, Muench HM, Hofmann SG, Laireiter AR. Treatments for internet addiction, sex addiction and compulsive buying: a meta-analysis. J Behav Addict 2020 Apr 1;9(1):14-43. [doi: 10.1556/2006.2020.00005] [Medline: 32359229]
27. Jiang YS, Liu TH, Qin D, Wang ZP, He XY, Chen YN. Effects of non-pharmacological interventions on youth with internet addiction: a systematic review and meta-analysis of randomized controlled trials. Front Psychiatry 2023;14:1327200. [doi: 10.3389/fpsyt.2023.1327200] [Medline: 38274427]
28. Tian JJ, He XY, Guo Z. Optimal non-pharmacological interventions for reducing problematic internet use in youth: a systematic review and Bayesian network meta-analysis. Behav Sci (Basel) 2025 Jan 20;15(1):98. [doi: 10.3390/bs15010098] [Medline: 39851902]
29. Kim J, Lee S, Lee D, et al. Psychological treatments for excessive gaming: a systematic review and meta-analysis. Sci Rep 2022;12(1). [doi: 10.1038/s41598-022-24523-9]
30. Kim S, Noh D. The current status of psychological intervention research for internet addiction and internet gaming disorder. Issues Ment Health Nurs 2019 Apr;40(4):335-341. [doi: 10.1080/01612840.2018.1534910] [Medline: 30742546]
31. Kurnaz MF, Koçtürk N. Are CBT-based interventions effective for pathologic technology use? A meta-analysis of experimental studies. Curr Psychol 2024 Mar;43(12):11127-11138. [doi: 10.1007/s12144-023-05206-8]
32. Liu J, Nie J, Wang Y. Effects of group counseling programs, cognitive behavioral therapy, and sports intervention on internet addiction in East Asia: a systematic review and meta-analysis. IJERPH 2017;14(12):1470. [doi: 10.3390/ijerph14121470]
33. Liu S, Xiao T, Yang L, Loprinzi PD. Exercise as an alternative approach for treating smartphone addiction: a systematic review and meta-analysis of random controlled trials. IJERPH 2019;16(20):3912. [doi: 10.3390/ijerph16203912]
34. Lo CKM, Chan KL, Yu L, Chui WWH, Ip P. Long-term effects of psychosocial interventions on internet-related disorders: a meta-analysis. Comput Human Behav 2023 Jan;138:107465. [doi: 10.1016/j.chb.2022.107465]
35. Malinauskas R, Malinauskiene V. A meta-analysis of psychological interventions for Internet/smartphone addiction among adolescents. J Behav Addict 2019 Dec 1;8(4):613-624. [doi: 10.1556/2006.8.2019.72] [Medline: 31891316]
36. Ock CM, Lee HS, Chae J, Kim H. Effectiveness of non-pharmacological interventions on gaming disorder: a systematic review and meta-analysis. Psychiatry Investig 2025 May;22(5):490-503. [doi: 10.30773/pi.2024.0358] [Medline: 40404379]
37. Reangsing C, Wongchan W, Trakooltorwong P, Thaibandit J, Oerther S. Effects of cognitive behavioral therapy (CBT) on addictive symptoms in individuals with internet gaming disorders: a systematic review and meta-analysis. Psychiatry Res 2025 Jun;348:116425. [doi: 10.1016/j.psychres.2025.116425] [Medline: 40054044]
38. Kwon MS, Han SJ, Yeun YR. Effectiveness of smartphone addiction intervention for university students: a systematic review and meta analysis. J Korea Converg Soc 2020;11(5):399-412. [doi: 10.15207/JKCS.2020.11.5.399]
39. Stevens MWR, King DL, Dorstyn D, Delfabbro PH. Cognitive–behavioral therapy for internet gaming disorder: a systematic review and meta‐analysis. Clin Psychology and Psychoth 2019 Mar;26(2):191-203. [doi: 10.1002/cpp.2341]
40. Vicente-Escudero JL. Paradox or efficacy. Meta-analysis of internet-based psychological interventions to reduce internet addiction, anxiety and depression in internet-addicted people. Curr Psychol 2025 Jun;44(12):12807-12819. [doi: 10.1007/s12144-025-08096-0]
41. Wang H, Li X, Lok GKI, et al. Family-based therapy for internet addiction among adolescents and young adults: a meta-analysis. J Behav Addict 2024 Jun 26;13(2):295-312. [doi: 10.1556/2006.2024.00015] [Medline: 38635339]
42. Yan Y, Qin X, Liu L, Zhang W, Li B. Effects of exercise interventions on Internet addiction among college students: a systematic review and meta-analysis of randomized controlled trials. Addict Behav 2025 Jan;160:108159. [doi: 10.1016/j.addbeh.2024.108159] [Medline: 39303655]

XSL•FO
RenderX

43. Yeun YR, Han SJ. Effects of psychosocial interventions for school-aged children's internet addiction, self-control and self-esteem: meta-analysis. Healthc Inform Res 2016 Jul;22(3):217-230. [doi: 10.4258/hir.2016.22.3.217] [Medline: 27525163]
44. Zhang K, Lu X, Zhang X, et al. Effects of psychological or exercise interventions on problematic mobile phone use: a systematic review and meta-analysis. Curr Addict Rep 2023;10(2):230-253. [doi: 10.1007/s40429-023-00471-w]
45. Zhang M, Meng SQ, Hasan AJ, et al. Network meta-analysis of the effectiveness of different interventions for internet addiction in college students. J Affect Disord 2024 Oct;363:26-38. [doi: 10.1016/j.jad.2024.07.032]
46. Zhang Y, Li G, Liu C, Chen H, Guo J, Shi Z. Mixed comparison of interventions for different exercise types on students with internet addiction: a network meta-analysis. Front Psychol 2023;14. [doi: 10.3389/fpsyg.2023.1111195]
47. Wang X, Zhang Y, Lin J, et al. Treatments of internet gaming disorder and comorbid mental disorders: a systematic review and meta-analysis. Comput Human Behav 2023 Dec;149:107947. [doi: 10.1016/j.chb.2023.107947]
48. Montag C, Wegmann E, Sariyska R, Demetrovics Z, Brand M. How to overcome taxonomical problems in the study of internet use disorders and what to do with "smartphone addiction"? J Behav Addict 2021 Jan 15;9(4):908-914. [doi: 10.1556/2006.8.2019.59] [Medline: 31668089]
49. Dong G, Potenza MN. A cognitive-behavioral model of internet gaming disorder: theoretical underpinnings and clinical implications. J Psychiatr Res 2014 Nov;58:7-11. [doi: 10.1016/j.jpsychires.2014.07.005] [Medline: 25062755]
50. Stockings E, Shakeshaft A, Farrell M. Community approaches for reducing alcohol-related harms: an overview of intervention strategies, efficacy, and considerations for future research. Curr Addict Rep 2018 Jun;5:274-286. [doi: 10.1007/s40429-018-0210-2]
51. Brand M, Wegmann E, Stark R, et al. The Interaction of Person-Affect-Cognition-Execution (I-PACE) model for addictive behaviors: update, generalization to addictive behaviors beyond internet-use disorders, and specification of the process character of addictive behaviors. Neurosci Biobehav Rev 2019 Sep;104:1-10. [doi: 10.1016/j.neubiorev.2019.06.032] [Medline: 31247240]
52. Patterson MS, Spadine MN, Graves Boswell T, et al. Exercise in the treatment of addiction: a systematic literature review. Health Educ Behav 2022 Apr 29:10901981221090155. [doi: 10.1177/10901981221090155] [Medline: 35484950]
53. Chen Y, Lu J, Wang L, Gao X. Effective interventions for gaming disorder: a systematic review of randomized control trials. Front Psychiatry 2023 Feb 6;14:1098922. [doi: 10.3389/fpsyt.2023.1098922] [Medline: 36815197]
54. Ioannidis JP. Integration of evidence from multiple meta-analyses: a primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. CMAJ 2009 Oct 13;181(8):488-493. [doi: 10.1503/cmaj.081086] [Medline: 19654195]
55. Michopoulos I, Furukawa TA, Noma H, et al. Different control conditions can produce different effect estimates in psychotherapy trials for depression. J Clin Epidemiol 2021 Apr;132:59-70. [doi: 10.1016/j.jclinepi.2020.12.012] [Medline: 33338564]
56. Laws KR, Pellegrini L, Reid JE, Drummond LM, Fineberg NA. The inflating impact of waiting-list controls on effect size estimate. Front Psychiatry 2022 Jun 22;13:877089. [doi: 10.3389/fpsyt.2022.877089] [Medline: 35815038]
57. Mohr DC, Ho J, Hart TL, et al. Control condition design and implementation features in controlled trials: a meta-analysis of trials evaluating psychotherapy for depression. Transl Behav Med 2014 Dec;4(4):407-423. [doi: 10.1007/s13142-014-0262-3] [Medline: 25584090]
58. Reed GM, First MB, Billieux J, et al. Emerging experience with selected new categories in the ICD-11: complex PTSD, prolonged grief disorder, gaming disorder, and compulsive sexual behaviour disorder. World Psychiatry 2022 Jun;21(2):189-213. [doi: 10.1002/wps.20960] [Medline: 35524599]
59. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-926. [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]

Abbreviations

AMSTAR 2: A Measurement Tool to Assess Systematic Reviews 2
CBT: cognitive behavioral therapy
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
***ICD-11*:** *International Classification of Diseases, 11th Revision*
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SMD: standardized mean difference

XSL•FO
RenderX

Edited by A Schwartz, M Balcarras; submitted 01.Aug.2025; peer-reviewed by H Rajesh, H Thongdaeng, Z Zhang; revised version received 18.Jan.2026; accepted 04.Feb.2026; published 05.Mar.2026.

Please cite as:
Zhang M, Lan T, Song T, Wang X
Effectiveness of Interventions for Internet, Smartphone, and Gaming Addictions: Umbrella Review and Meta–Meta-Analysis
J Med Internet Res 2026;28:e81705
URL: https://www.jmir.org/2026/1/e81705
doi:10.2196/81705

© Minggang Zhang, Tong Lan, Tao Song, Xiaochun Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effectiveness of Virtual Reality–Based Early Rehabilitation Strategies on Pain, Sleep, Anxiety, Balance, Cognition, and Limb Motor Function in Adult Intensive Care Unit Patients: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Fei Wu[1,2], MSN; Yunting Wu[1,2], MSN; Yana Xing[1,3], MSN; Weixin Cai[1*], MSN; Ran Zhang[1,3*], MSN

[1]Nursing Department, Beijing Tiantan Hospital, Capital Medical University, No. 119, South 4th Ring West Road, Fengtai District, Beijing, China, Beijing, Beijing, China

[2]School of Nursing, Capital Medical University, Beijing, Beijing, China

[3]Department of Critical Care Medicine, Beijing Tiantan Hospital, Capital Medical University, Beijing, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Ran Zhang, MSN
Nursing Department, Beijing Tiantan Hospital, Capital Medical University, No. 119, South 4th Ring West Road, Fengtai District, Beijing, China, Beijing, Beijing, China

Abstract

Background: Early rehabilitation is vital for functional recovery in critically ill patients. Virtual reality–based early rehabilitation intervention (VR-ERI) is an emerging strategy, but evidence on its feasibility, safety, and efficacy remains inconsistent and unsynthesized.

Objective: We synthesized evidence from randomized controlled trials (RCTs) on the feasibility and safety of VR-ERI in adult critically ill patients and evaluated its effects on functional outcomes during intensive care unit (ICU) stay and at short-term follow-up (≤3 months post ICU).

Methods: Following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines, we searched 10 databases (eg, PubMed, Web of Science, Cochrane Library, and Embase) from inception to October 5, 2025, for Chinese and English publications. We included RCTs comparing VR-ERI with control measures initiated early (within 72 h of ICU admission) in adults. FW and YW independently screened studies and extracted data; YW and YX independently conducted the revised Cochrane Risk of Bias tool assessment. Data were synthesized narratively or via meta-analysis in R Studio using a random-effects model (Hartung-Knapp-Sidik-Jonkman adjustment). Effects were expressed as standardized mean differences (SMD; with Hedges g correction) with 95% CIs and 95% prediction intervals (95% PIs). Subgroup and sensitivity analyses were performed to explore heterogeneity and assess robustness. The Grading of Recommendations Assessment, Development, and Evaluation framework was used to assess the quality of evidence.

Results: Sixteen RCTs (published 2020 - 2025) involving 1356 patients were included. Bias assessment found 1 study at low risk, 5 with some concerns, and 10 at high risk. Meta-analysis suggested potential trends for VR-ERI in improving ICU anxiety (SMD −0.86, 95% CI −1.85 to 0.13, 95% PI −3.75 to 2.03; very low certainty) and subjective sleep quality (SMD 3.36, 95% CI 0.77-5.94; very low certainty), with a more pronounced effect in the Richards-Campbell Sleep Questionnaire–assessed subgroup (SMD 5.12, 95% CI 0.54-9.71). At follow-up, VR-ERI showed trends toward improved balance (Berg Balance Scale: SMD 0.97, 95% CI 0.74-1.20, 95% PI 0.37-1.58; moderate certainty), limb motor function (Fugl-Meyer: SMD 1.40, 95% CI −0.23 to 3.02; low certainty), and cognitive function (SMD 0.78, 95% CI 0.16-1.39; low certainty). No significant differences were found for objective sleep measures or ICU pain (low to very low certainty). No serious adverse events were reported; only a few studies mentioned mild reactions, such as dizziness, nausea, and fatigue.

Conclusion: This review indicates VR-ERI's potential to improve anxiety, subjective sleep, balance, cognition, and motor function in early ICU rehabilitation, while its effects on pain and objective sleep remain unclear and safety protocols need refinement. Given the high risk of bias, substantial heterogeneity, and imprecision, the overall certainty of evidence is low. Thus, VR-ERI may serve as a nonpharmacological adjunct, but its clinical translation requires consideration of cost and patient suitability, and more rigorous research is needed.

Trial Registration: PROSPERO CRD420251114439; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251114439

(J Med Internet Res 2026;28:e81865) doi:10.2196/81865

XSL•FO
RenderX

KEYWORDS

virtual reality; intensive care unit; early rehabilitation; randomized controlled trial; systematic review; meta-analysis

Introduction

Background

Patients in the intensive care unit (ICU) often experience a series of physical, cognitive, and psychological impairments after discharge due to factors such as physiological dysfunction, surgical trauma, pain, and reduced social interaction [1]. In 2010, the Society of Critical Care Medicine first introduced the term "post-intensive care syndrome (PICS)" to define this phenomenon [2,3]. The manifestations commonly include weakness, pain, anxiety, sleep disturbances, and cognitive dysfunction, which can persist for months or even years. Studies indicate that 47.9% of patients still exhibit cognitive impairment 3 - 6 months after transfer from the ICU [4]; 68% (355/523) report persistent pain after discharge, with pain intensity remaining at a moderate level even 1 year later [5]. Additionally, 55% of patients experience sleep disturbances within the first month after discharge [6]. Furthermore, 12.3% (98,530/799,645) of critically ill patients are diagnosed with depression or anxiety after discharge, and these patients demonstrate higher long-term mortality rates [7]. PICS severely compromises patients' quality of life, reduces functional independence, and imposes a significant care burden on families and society [8].

To reduce long-term impairments after critical illness, the Society of Critical Care Medicine recommends that health care professionals assess patients and coordinate care strategies for injury management beginning at ICU admission [9]. As clinical treatment advances, early functional recovery has become a key focus for preventing PICS [10]. Early rehabilitation is generally initiated within 72 hours of ICU admission, following achievement of hemodynamic and respiratory stability (typically within 24 - 48 h) [11,12]. It commonly includes early physical activity [13,14], neurocognitive stimulation [15,16], psychological intervention [17], and multimodal programs [18]. Existing meta-analyses confirm that routine early rehabilitation started within 72 hours may improve physical and cognitive function, help prevent PICS without increasing adverse events [19], and reduce the incidence of ICU-acquired weakness [20]. However, no differences have been found in cognitive-related delirium-free days or Hospital Anxiety and Depression Scale scores. Another meta-analysis on nurse-led early rehabilitation showed it significantly shortens ICU length of stay, although effects on functional outcomes, such as mobility, muscle strength, and mortality, were not significant [21]. These inconsistencies may relate to factors including the time-intensive nature of current measures, insufficient standardization in implementation, and low patient participation and adherence.

Virtual reality (VR), a term first introduced by computer scientist Jaron Lanier in 1989, refers to a simulation technology providing users with an immersive environmental experience [22]. VR technology integrates computer, electronic information, and simulation technologies to enable users to interact with the VR system and experience an immersive, tactile, visual, and multisensory environment via VR helmets, all-in-one devices, and other gadgets [23]. The system is known for its novelty, interactivity, and immersive qualities [24]. VR technology–based rehabilitation techniques have been widely adopted in clinical rehabilitation. There is evidence suggesting that these techniques can effectively improve cognitive function in patients with stroke [25], reduce pain levels in patients with chronic low back pain and neck pain [26,27], and alleviate anxiety and depression [28,29].

Virtual reality–based early rehabilitation intervention (VR-ERI) is an emerging ICU recovery strategy that is gaining attention and has shown preliminary application results. Using VR technology, it provides ICU patients with specific virtual environments, such as natural landscapes [30], meditation music [31], or virtual tasks (such as memory games [32]), and cognitive exercises like "catching coconuts" and "avoiding obstacles" [33]. This aims to enhance rehabilitation motivation, participation, and physiological and psychological function. Its interactive, immersive, and feedback-driven design makes VR-ERI a promising tool for personalized early rehabilitation in critically ill patients. However, current research findings are inconsistent. Some studies report that VR-ERI effectively reduces anxiety [32], while others show no significant difference [34,35]. Such discrepancies may affect the certainty and consistency of clinical decision-making, potentially hindering the technology's wider adoption in ICU early rehabilitation.

To our knowledge, researchers [36] have explored the interventional effects of VR therapy on patients' PICS through a meta-analysis, but its included population did not specifically focus on early rehabilitation strategies during the ICU period. This analysis pooled data from patients receiving VR therapy both within the ICU and during follow-up after discharge. Due to differences between these 2 populations in aspects such as intervention timing and environment, combining them in the analysis may introduce clinical heterogeneity, thereby weakening the credibility of the results. Another meta-analysis indicated that compared with traditional rehabilitation, VR-based early motor rehabilitation can improve balance, functional walking, and upper and lower limb motor function in critically ill patients [37]. However, this review only assessed motor function and did not comprehensively examine the effects on other early rehabilitation outcome indicators.

Objectives

This systematic review and meta-analysis aim to (1) characterize the types and content of VR-ERI activities and analyze the feasibility and safety of VR technology implementation in critically ill patients; (2) systematically evaluate the effectiveness of VR-ERI on physiological, psychological, and cognitive outcomes during ICU treatment and at short-term follow-up (within 3 months post-ICU discharge); and (3) provide recommendations for future research.

XSL•FO
RenderX

Methods

Study Protocol and Registration

The review protocol was registered in PROSPERO. This systematic review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement [38] to ensure transparency (Checklists 1-3).

Deviations From the PROSPERO Protocol

During implementation, the search language was limited to Chinese and English to ensure assessment accuracy, and the search cutoff date was extended to October 5, 2025, to incorporate the latest evidence. Regarding statistical analysis, due to heterogeneity among the included studies, all analyses adopted the more conservative random-effects model. To enhance statistical robustness, the analysis was switched to the R language (version 4.4.1; R Core Team), and the Hartung-Knapp-Sidik-Jonkman method was applied to correct random-effects outcomes. This method provides more conservative CIs when the number of studies is small, reducing the risk of false-positive results in the presence of high heterogeneity. Apart from the above adjustments, all other procedures remained consistent with the original protocol.

Data Sources and Search Strategy

Two independent researchers (FW and YW) conducted literature searches across 6 English-language and 4 Chinese-language databases. MEDLINE and Embase were searched simultaneously via the Ovid platform. The other English databases were searched on their respective platforms—PubMed (via National Center for Biotechnology Information), Web of Science (via Clarivate Analytics), the Cochrane Library, and PsycINFO (via ProQuest). The Chinese databases searched were China National Knowledge Infrastructure (CNKI), WanFang Data, VIP Database (VIP), and the SinoMed Service System (SinoMed). The search strategy combined Medical Subject Headings with free-text terms using Boolean operators. The search was performed in 2 stages. An initial search covered records from the inception of each database until May 11, 2025, and to incorporate the most recent evidence, the full search strategy was rerun on October 5, 2025, extending the cutoff date accordingly. All searches were limited to Chinese and English publications to align with the team's language capabilities and ensure feasibility.

All search strategies were independently designed based on the research question and were peer-reviewed and refined by experts in evidence-based medicine and critical care medicine in accordance with the PRISMA-S (Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension) checklist [39]. No published search filters were used, nor were search strategies from other systematic reviews adapted or reused. To cover the literature as comprehensively as possible, we also used the snowballing method to trace references and citations and attempted to obtain full-text articles by contacting librarians or the original authors. Furthermore, as this study focused on published efficacy evidence, we did not specifically search clinical trial registries or gray literature (including conference proceedings). The complete search strategies and procedures are detailed in Multimedia Appendix 1.

Study Selection

The rationale for formulating inclusion and exclusion criteria was based on the Population, Intervention, Comparison, Outcome, and Study Design (PICOS) framework, (1) population: adults aged ≥18 years receiving treatment during their ICU stay; (2) intervention: VR-ERI; (3) control: no treatment, usual care measures, or non–VR-based interventions; (4) outcomes: studies reporting at least 1 physical, psychological, or cognitive function outcome (such as cognition, pain, delirium, anxiety, depression, and sleep quality) or feasibility and safety outcomes during ICU treatment or after ICU transfer; and (5) study design: randomized controlled trials (RCTs) published in Chinese or English. The exclusion criteria were (1) studies that did not report original outcome measures and for which relevant outcome data could not be obtained after contacting the authors; (2) duplicate publications; and (3) case reports, letters, conference abstracts, and review articles.

Screening

In total, 2 independent researchers (FW and YW) conducted the literature screening and data extraction, followed by cross-checking. Any uncertainties were resolved by WC. First, duplicate records were automatically identified and removed using the built-in function of the EndNote X9 (Clarivate) reference management software, based on fields such as title, author, publication year, and journal name. Subsequently, the 2 researchers (FW and YW) independently examined the titles, authors, and other key information of the remaining records to further exclude any duplicates not detected by the software. Next, titles and abstracts were reviewed to exclude obviously irrelevant literature. Finally, full texts were retrieved and assessed against the eligibility criteria to determine the final set of included studies.

Data Extraction

A total of 2 independent researchers (FW and YW) performed data extraction using a uniformly developed standardized form without using automated tools. The data that were systematically extracted were (1) general information, including first author, country, and publication year; (2) participant characteristics, including disease type, sample size, and baseline demographic and clinical characteristics; and (3) intervention characteristics, including specific type of VR, content of the VR-based intervention, content of the control group intervention, intervention duration, intervention frequency, outcome measures, measurement tools, time points of data collection and raw data (such as mean, SD, median, and IQR), for outcomes reported at baseline, post intervention, or follow-up for both groups.

To manage multiple data points objectively and consistently, we predefined selection criteria following the Cochrane Handbook [40]. The core principle was to prioritize data that best aligned with the review's primary objective, had the highest validity, or greatest clinical representativeness. For multiple measures, we selected the tool that best matched the outcome definition or had the highest validity; for multiple intervention

XSL•FO
RenderX

groups, we included the primary intervention group, the most clinically representative, or the largest in sample size; for multiple control groups, we chose the "most representative usual care" group. Where data were missing, unclear, or incomplete, we contacted corresponding authors by email for clarification or missing data. If no response was received after 2 attempts, we used recommended formulas to estimate means and SDs from available data (eg, based on sample size, median, and range) [41].

Risk-of-Bias Assessment

The second author (YW) and the third author (YX) independently assessed the risk of bias of the included studies using the revised Cochrane Risk of Bias tool (RoB 2) for randomized trials [42]. The RoB 2 tool assesses five domains for randomized trials, which are (1) bias arising from the randomization process, (2) bias due to deviations from the intended interventions, (3) bias due to missing outcome data, (4) bias in the measurement of the outcome, and (5) bias in the selection of the reported result. Each domain is judged against a series of signaling questions. The overall risk of bias for a study is categorized as "low," "some concerns," or "high." A study is judged to have a low overall risk of bias if all domains are rated as "low." An overall rating of "some concerns" is given if at least 1 domain is rated as having "some concerns" and no domain is rated as "high." An overall "high" risk of bias is judged if any single domain is rated as "high" or if multiple domains are rated as having "some concerns."

Data Synthesis and Analysis

To ensure transparency in the data synthesis process, we followed a systematic decision-making procedure. First, all studies had to meet the predefined PICOS criteria and provide the raw data for the relevant outcome measures. Second, we established a rule that each study would contribute only a single data point for a given outcome. If a study contained multiple measurements or multiple group comparisons, a single dataset was selected for pooling according to the prespecified criteria, aiming to minimize heterogeneity as much as possible.

All analyses were performed using the R language (version 4.4.1) within the R Studio integrated development environment (version 2023.12). The metafor package was used for effect size pooling and forest plot generation [43], providing visual representations of each study's effect size with its 95% CI, along with the pooled overall effect size. All included outcome measures were continuous variables. We conducted a meta-analysis using change scores (from baseline to postintervention or follow-up) in means and SDs for both groups. For studies reporting only pre- and postintervention means and SDs, the mean and SD of change scores were calculated as $Mean_{change}=Mean_{post}-Mean_{pre}$ and $SD_{_change}=\sqrt{(SD_{_pre}^2+SD_{_post}^2-2\times r\times SD_{_pre}\times SD_{_post})}$, where r represents the correlation coefficient between pre- and postintervention values. When correlations were not reported in the original publications, they were calculated according to the Cochrane Handbook for Systematic Reviews of Interventions. To harmonize different assessment tools and ensure consistent direction of results, we reversed the scoring for subjective sleep quality assessed by the Pittsburgh Sleep Quality Index and cognitive scores assessed by the Perceived Deficits Questionnaire for Depression, following Cochrane Handbook guidelines [40]. Consequently, higher scores in the pooled results indicate better sleep quality and cognitive function in patients. Additionally, we categorized outcome measures by assessment time points, specifically divided into (1) during ICU treatment and (2) short-term follow-up (2 wk to 3 mo after ICU discharge).

Given the variability in assessment tools and the presence of clinical and methodological heterogeneity among the included studies, the standardized mean difference (SMD) was adopted as the effect size measure for meta-analysis outcomes. The adjusted Hedges g was used to estimate the effect sizes [44,45]. Absolute SMD values of 0.2, 0.5, and 0.8 were interpreted as small, medium, and large effect sizes, respectively [46]. A random-effects model was applied to derive more conservative effect estimates. To enhance the robustness of statistical inference, particularly in reducing the risk of false-positive results when the number of studies was limited, the Hartung-Knapp-Sidik-Jonkman adjustment was used during model fitting to estimate the pooled effect size and its 95% CI [47].

To quantify between-study heterogeneity and enhance the generalizability of the results in real-world settings, we used a comprehensive assessment strategy. First, absolute heterogeneity was quantified using τ^2 (between-study variance) and its square root τ (between-study SD) [48]. τ^2 provides an absolute measure of the variation between studies, while τ, being on the same scale as the effect size, facilitates clinical interpretation. Second, the Cochran Q test was used for the hypothesis testing of effect size homogeneity, supplemented by the I^2 statistic to quantify the relative extent of heterogeneity [49], with thresholds of 25%, 50%, and 75% used to define low, moderate, and high levels, respectively [50]. Finally, for outcome measures involving at least 3 studies, 95% prediction intervals (PIs) were calculated [51] to estimate the potential range of the true effect size in future similar interventions, thereby providing more clinically informative guidance for practice.

Subgroup analyses were performed to explore potential sources of heterogeneity. A leave-one-out sensitivity analysis was conducted to assess the robustness of the pooled results. For outcome measures with fewer than 2 included studies, descriptive analyses were performed. Additionally, to assess the potential for small-study effects (such as publication bias), we performed several checks for the anxiety outcome, which had the largest number of included studies (n=8). These included a visual inspection of funnel plot symmetry, the Egger test (regression test), and the Begg test (rank correlation test). Furthermore, the trim-and-fill method was used to estimate the potential number of missing studies and to compute an adjusted effect size, thereby evaluating the potential impact of small-study effects on the pooled result. The threshold for statistical significance was set at $P\leq.05$.

Evidence Quality

To evaluate the reliability of the evidence, we used the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) framework [52] to rate the certainty of evidence for

XSL•FO
RenderX

each outcome. In addition, 2 independent raters (FW and YW) assessed the certainty of evidence using the official GRADE pro GDT software (Evidence Prime Inc), with any disagreements resolved by the corresponding author (RZ). The evidence was rated as very low, low, moderate, or high. Footnotes were provided to explain the reasons for any downgrading of the evidence, which included considerations of risk of bias, inconsistency, indirectness, imprecision, and publication bias.

Ethical Considerations

This study used literature data and did not require approval from an ethics review committee or patient consent.

Results

Search Results

A total of 3737 studies were identified from 10 databases. After removing 243 duplicate articles, the titles and abstracts of the studies were initially screened, and 58 full-text manuscripts relevant to this study were identified. Following full-text reading and quality assessment, 16 RCTs [16,17,30,32-35,53-61] were ultimately included for analysis. The specific search and inclusion process is illustrated in Figure 1, and the excluded articles after rescreening, along with the reasons, are detailed in Multimedia Appendix 2.

Figure 1. Flowchart of literature screening. CNKI: China National Knowledge Infrastructure; SinoMed: SinoMed Service System; VIP: VIP Database.

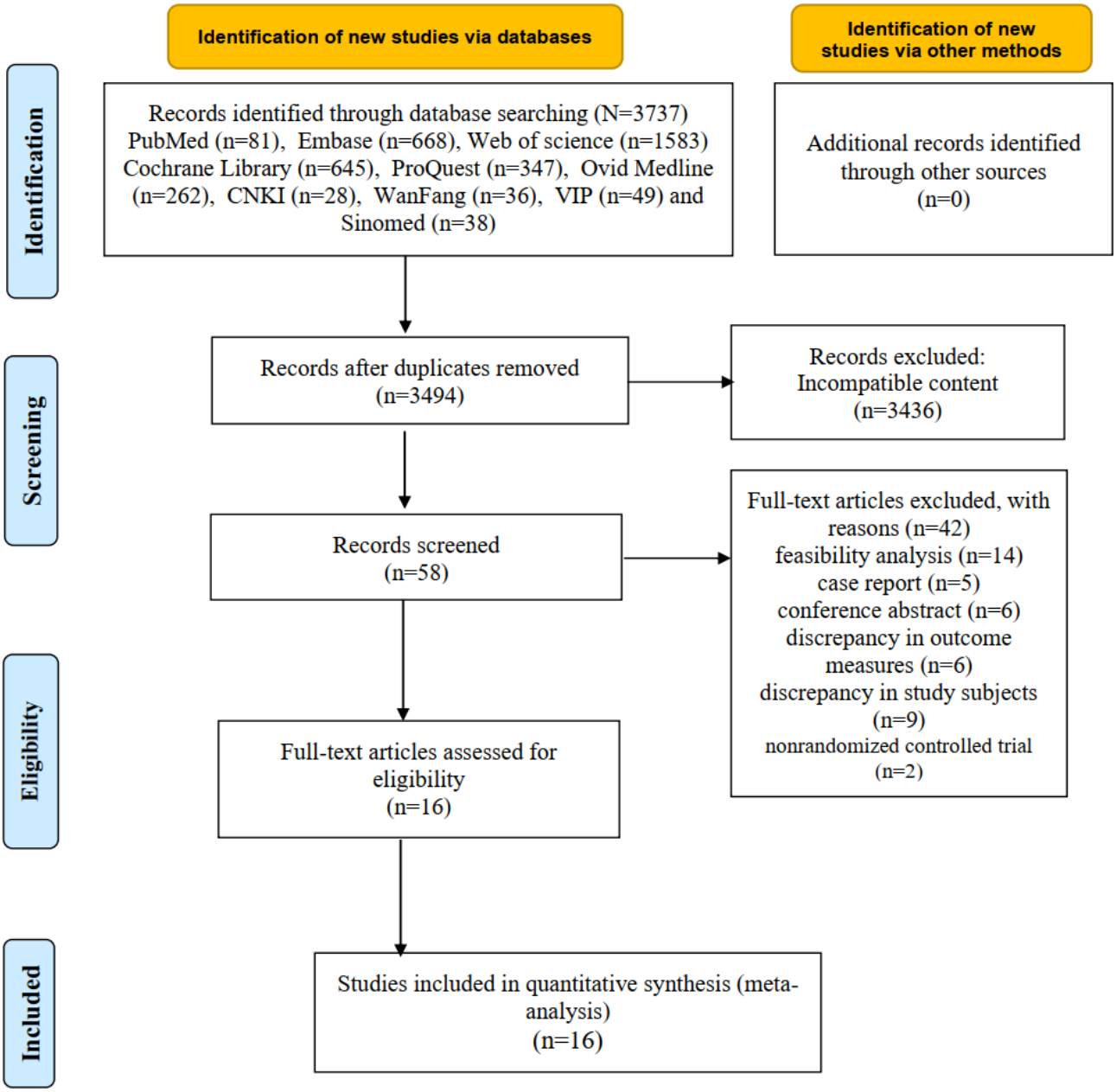

Characteristics of the Studies

Multimedia Appendix 3 summarizes the basic characteristics of the included studies. The 16 studies (published between 2020 and 2025) comprised a total of 1356 participants, with sample sizes ranging from 42 to 180. The studies originated from Brazil [53], Korea [30,54], France [17,35], Spain [16], Turkey [34], and China [32,33,55-61]. Patient conditions included cardiovascular diseases [30,35,53,55,60], postcraniotomy [33], postliver transplantation [61], postlaparoscopic sleeve

XSL•FO
RenderX

gastrectomy [34], conventional oncologic surgery [32], severe respiratory diseases [58], and postmultiple fracture surgery [56], among others. The mean age of patients was generally between 60 and 70 years. Interventions were initiated within 24 hours after ICU admission in 3 studies [35,53,59], within 6 hours in 2 studies [32,34], and on the day of ICU admission once the patient's condition was stabilized in the remaining 11 studies [16,17,30,33,54-58,60,61].

One study [16] used nonimmersive VR technology, while 15 studies [17,30,32-35,53-61] used fully immersive VR. The intervention directions included (1) early cognitive rehabilitation (training reaction and executive function through games or tasks) [16,33,55,57], (2) early motor rehabilitation (engaging in virtual activities, such as fruit cutting or cycling combined with motion capture) [53,56,58,59,61], (3) early delirium prevention (providing emotional support and cognitive training for multidimensional prevention) [32], (4) VR relaxation therapy (conducting physical and mental relaxation in natural scenes) [17,34,35], (5) VR meditation (improving sleep and mood by combining mindfulness) [30,54], and (6) VR preoperative visit (structured introduction to the ICU environment, treatment process, and rehabilitation) [60]. The intervention protocols were categorized into (1) single-session short-term intervention (15 - 30 min at key clinical time points) [17,30,34,35,60], (2) high-frequency short-term intervention during ICU stay (approximately 20 min per session, 1 - 3 times daily, lasting 3 - 7 d) [16,32,54,55,57-59], (3) medium- to long-term structured intervention (30 - 40 min per session, 5 d per wk, lasting 2 wk to 3 mo) [33,56,61], and (4) flexible individualized protocol (adjusted based on patient tolerance) [53]. The studies involved a total of 24 outcome measures, such as cognitive function, delirium, anxiety, depression, pain, sleep, and quality of life, although the assessment tools and observation windows varied across studies.

Risk of Bias

The overall risk of bias for the studies included in this review was high, with summary results presented in Figures 2 and 3, and detailed assessment results available in Multimedia Appendix 4. Furthermore, 1 study [58] was rated as "low," 5 studies [17,32,34,54,55] raised "some concerns," and 10 [16,30,33,35,53,56,57,59-61] were rated as "high." In domain 1 (randomization process), 7 studies [33-35,53,56,60,61] did not describe allocation concealment, and 1 study [32] described neither the method for generating the random sequence nor allocation concealment; all were rated as "some concerns." For domain 2 (deviations from intended interventions), 4 studies [16,32,55,57] were rated as "some concerns" due to lack of blinding of patients and implementers; 1 study [59] was rated as "high" because it lacked blinding and did not adhere to the intention-to-treat analysis principle. In domain 3 (missing outcome data), 2 studies [16,59] were rated as "high" due to a high rate of loss to follow-up and because the reasons for missing data were potentially related to health status. For domain 4 (measurement of the outcome), 11 studies [16,17,30,32,33,35,53,56,57,60,61] had issues with the lack of blinding of outcome assessors. Among these, 9 studies [16,30,33,35,53,56,57,60,61] were rated as "high" because they measured subjective patient-reported outcomes (eg, pain and anxiety), and 2 studies [17,32] were rated as "some concerns" as their outcome measures were relatively objective. In domain 5 (selection of the reported result), transparency was generally insufficient. Moreover, 4 studies [16,17,35,55] mentioned trial registration but lacked verifiable prespecified analysis descriptions for the reported results, while the remaining 11 studies [30,32-34,53,54,56,57,59-61] did not mention a preregistered protocol; consequently, all were rated as "some concerns."

Figure 2. Risk of bias graph for proportion of judgements per item in included studies.

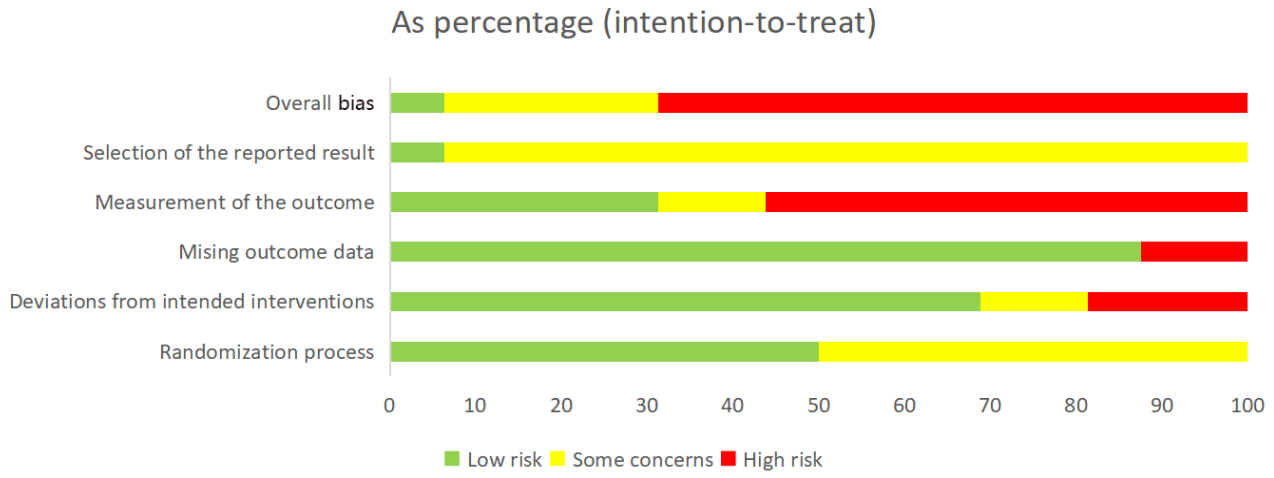

XSL•FO
RenderX

Figure 3. Risk of bias summary Items assessed in each included study [16,17,30,32-35,53-61].

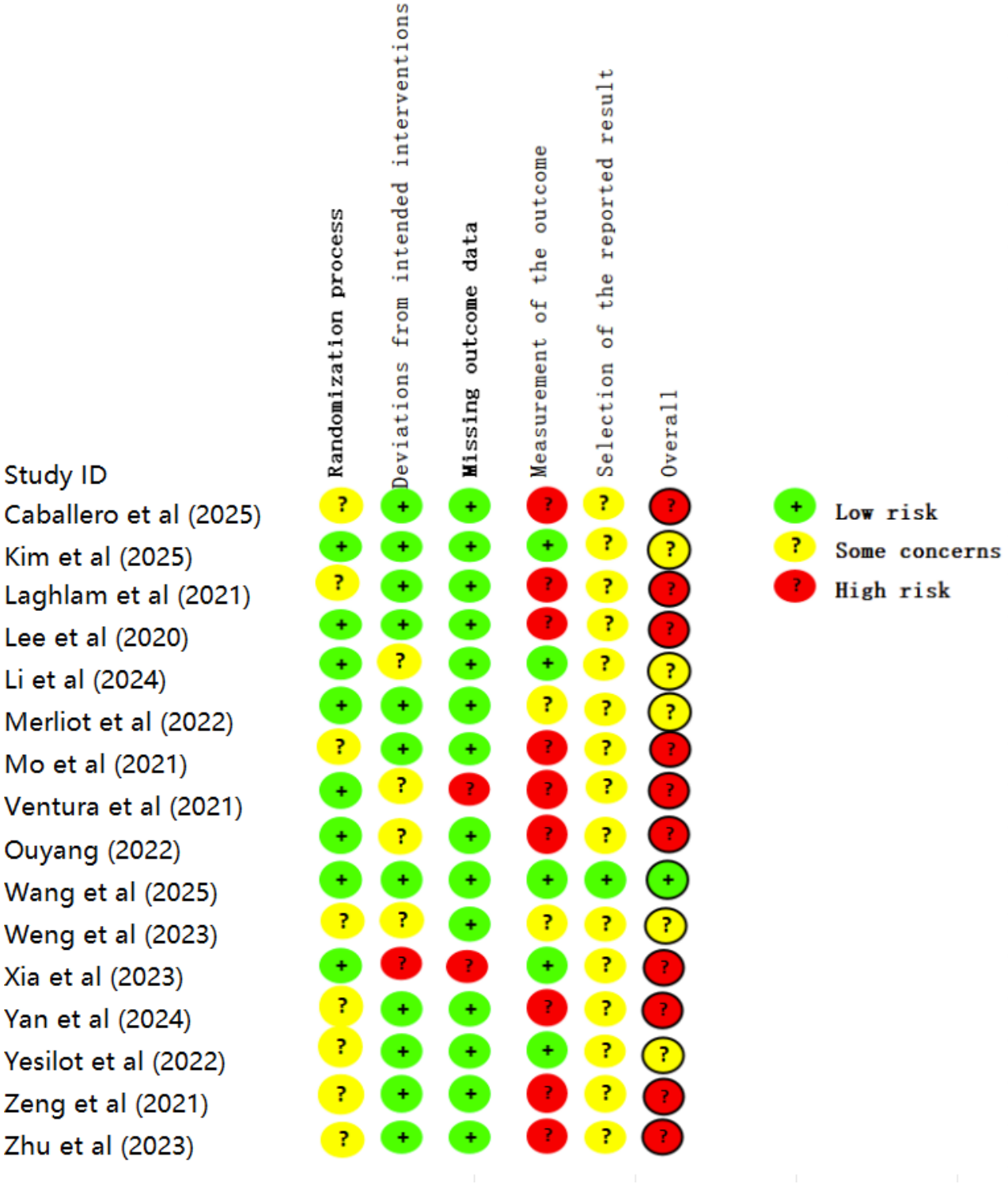

Clinical Safety and Feasibility

In total, 8 studies [16,17,35,53,55,57-59] reported on safety, 5 studies [16,55,57-59] confirmed the absence of serious adverse events but documented related reactions including dizziness [16,35,57], nausea [35,57], fatigue [16,17], anxiety [16,17], confusion [16,55], visual disturbance [17], headache or discomfort from wearing the headset [55], claustrophobia, agitated withdrawal, and dyspnea due to device displacement [17]. Furthermore, 1 study reported a significant drop in blood pressure during intervention in a patient [59], another reported a case of excessive drowsiness [16], and 1 noted that vital signs in all patients fluctuated within the normal range (±10%) during the intervention [58]. Regarding feasibility, 1 study reported a 78.3% patient recommendation rate and a 90% "good/excellent" experience rating [53]. Meanwhile, 2 studies reported that patients in the VR group demonstrated significantly better outcomes in terms of daily total exercise time, average daily out-of-bed activity time [58], and exercise adherence (with a high adherence rate of 71.74% vs 54.35% at 3 months) [56].

XSL•FO
RenderX

Results of the Meta-Analysis

Among the included trials, due to the limited number of studies, we conducted meta-analyses on outcome indicators with at least 2 studies. The specific analysis results are as follows.

Anxiety

A total of 8 studies (n=801) included anxiety as an outcome measure. According to the predefined data pooling criteria, for the study by Li et al [55], we only pooled the data assessed during ICU treatment using the State-Trait Anxiety Inventory scale. For the study by Merliot-Gailhoustet et al [17], we only pooled the data from the group receiving the DEEPSEN VR system (Saint-Didier-au-Mont-d'Or, France). All studies were analyzed using change-from-baseline values for the pooling.

The meta-analysis revealed that, in terms of the average effect, VR-ERI did not show a statistically significant effect on reducing patient anxiety (SMD −0.86, 95% CI −1.85 to 0.13). Given that higher scores on all assessment tools indicated greater anxiety, the negative point estimate suggests that VR-ERI may reduce anxiety. However, the 95% PI was wide and included the null value (−3.75 to 2.03), indicating considerable uncertainty about the intervention's effect in future settings, ranging from a substantial benefit to potential harm. Therefore, these findings require validation in further studies.

Extremely high heterogeneity was detected among the studies (I^2=95.4%, τ=1.1480, τ^2=1.3179; Cochran Q=150.82, df=7; P<.01; Figure 4). Exploratory subgroup analyses were performed based on VR intervention content, timing, patient population, and type of anxiety assessment tool, but no potential source of heterogeneity was identified. Sensitivity analysis (leave-one-out method) indicated that the pooled effect size was sensitive to the removal of individual studies, with a maximum change of 0.3478, exceeding the prespecified threshold of 0.1, suggesting that the results may not be robust. The study by Weng et al [32] had the most substantial influence on the overall effect estimate. After excluding this study, the pooled effect size remained statistically nonsignificant (SMD −0.51, 95% CI −1.17 to 0.15; 95% PI −2.31 to 1.28), and heterogeneity remained high (I^2=92.5%, τ=0.6806, τ^2=0.4632; Cochran Q=132.89, df=6; P<.01), further supporting the instability of the findings.

Figure 4. Forest plot of the meta-analysis for anxiety outcomes [16,17,32,34,35,55,57,60].

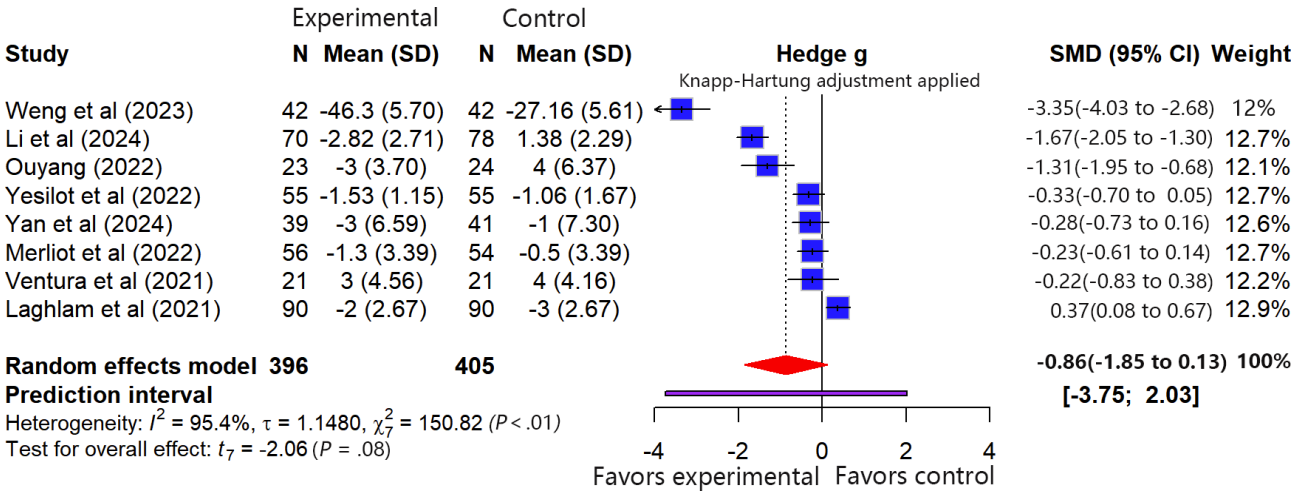

According to the GRADE framework, the certainty of evidence for anxiety was rated as "very low." The downgrading was based on (1) serious risk of bias in the included studies (4 studies [16,35,57,60] with high risk), (2) very high heterogeneity among studies, (3) imprecision of the pooled effect estimate (the 95% CI spanned both the line of null effect and the threshold for clinically meaningful benefit), and (4) potential publication bias. Although the pooled effect size suggested a potential moderate-to-large effect, an upgrade was not considered after weighing the above limitations (detailed assessment is available in Multimedia Appendix 5).

Subjective Sleep Quality (Scale Assessment)

In total, 6 studies [30,32,54,55,57,60] (n=500) assessed subjective sleep quality during ICU stay. In accordance with predefined pooling criteria, we extracted data from the second intervention day for Kim and Kang [54] and the Richards-Campbell Sleep Questionnaire (RCSQ)–measured sleep data only for Ouyang [57]. While all studies reported comparable baseline characteristics, specific data were provided by only 3 [32,55,60]. To maximize data utility, the meta-analysis was performed using postintervention outcomes.

Meta-analysis revealed considerable heterogeneity (I^2=97.5%, τ=2.4143, τ^2=5.8288; Cochran Q=198.44, df=5; P<.01), likely due to differences in populations, interventions, and assessments. Sensitivity analysis (leave-one-out method) confirmed the robustness of the results, as the maximum change in effect size after excluding any single study was 0.0347 (below the prespecified 0.1 threshold), with no specific source of heterogeneity identified. The pooled results (Figure 5) indicated a significant positive effect of VR-ERI on improving patients' subjective sleep quality compared with usual care (SMD 3.36, 95% CI 0.77-5.94). For consistent interpretation, Pittsburgh Sleep Quality Index scores were reversed so that higher scores uniformly indicated better sleep quality across all scales; thus, the positive SMD reflects an enhancing effect of VR-ERI, although more studies are needed to confirm the precision of this estimate.

XSL•FO
RenderX

Figure 5. Forest plot of the subgroup analysis for subjective sleep quality [30,32,54,55,57,60].

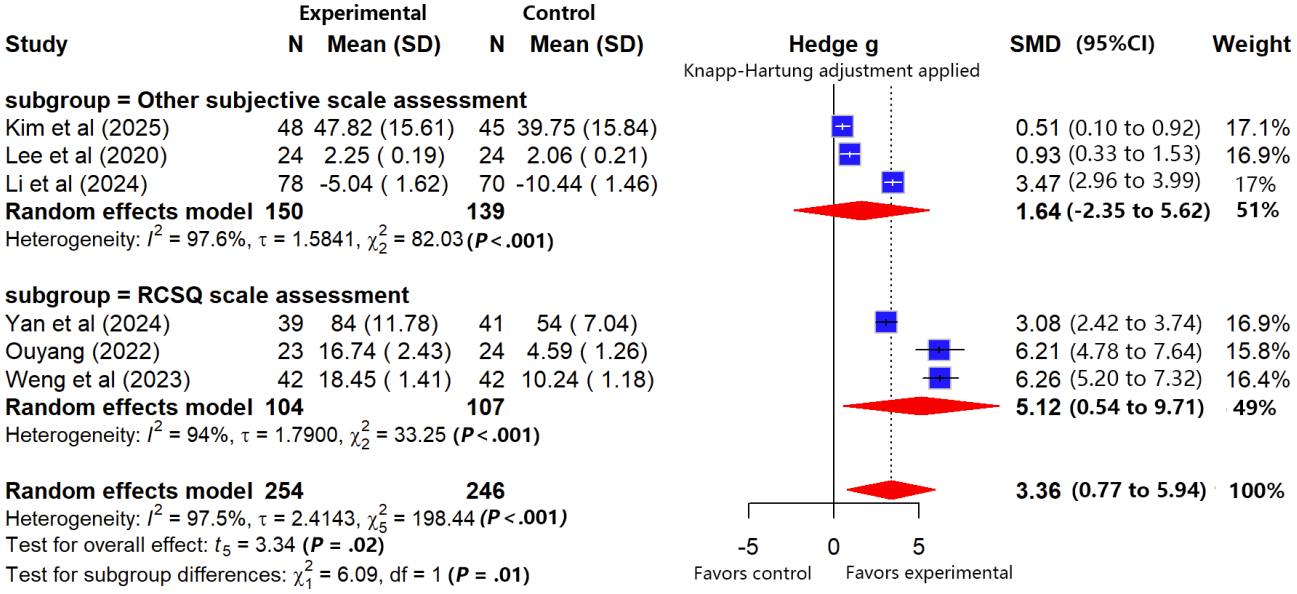

Subgroup analysis was conducted after unifying scale directions to assess how the inherent characteristics of different instruments might influence the pooled result. The analysis showed that VR-ERI did not significantly improve sleep when assessed with generic scales (SMD 1.64, 95% CI −2.35 to 5.61). In contrast, a positive effect was observed with the ICU-specific RCSQ (SMD 5.12, 95% CI 0.54 to 9.71). The between-subgroup difference was statistically significant (χ^2_1=6.1; *P*=.014), indicating that the type of assessment tool is a significant effect modifier, with the RCSQ potentially being more sensitive to VR-ERI–related sleep improvements in the ICU setting.

According to the GRADE approach, the certainty of evidence for subjective sleep quality during ICU treatment was rated as "very low." The reasons for downgrading included (1) a serious risk of bias in the included studies (3 studies [30,57,60] with high risk), (2) very high heterogeneity among studies, and (3) extreme imprecision of the pooled effect estimate (the 95% CI was extremely wide, spanning the range from moderate to very large effects). Considering all the downgrading factors above, an upgrade was not applied despite the large pooled effect size (for detailed assessment, refer to Multimedia Appendix 5).

Objective Sleep Quality (Smart Wristband Monitoring)

In total, 2 studies [30,54] (n=141) monitored objective sleep quality using smart wristbands. Key metrics included total sleep time (TST), defined as the actual duration of sleep from onset to awakening; wake after sleep onset (WASO), representing total wake time during the sleep period as an indicator of sleep fragmentation; and sleep efficiency, calculated as the percentage of TST relative to total time in bed, with higher values indicating better sleep quality. A pooled analysis was performed using change-from-baseline values, with specific results as follows.

TST

The meta-analysis indicated no heterogeneity among the studies (I^2=0%, τ=0, τ^2=0; Cochran Q=0, df=1; *P*=.95). Sensitivity analysis (leave-one-out method) indicated that the results were robust. The maximum change in the effect size after excluding any single study was 0.0375, which is below the prespecified threshold of 0.1. The meta-analysis (Figure 6) showed that, regarding the average effect, VR-ERI did not lead to a statistically significant improvement in patients' TST compared with usual care (SMD 0.08, 95% CI −0.07 to 0.22). The positive SMD suggests that the direction of the effect may favor VR-ERI, indicating an increase in TST. According to the GRADE approach, the certainty of evidence for objective TST during ICU treatment was rated as "moderate." The rating was downgraded due to a high risk of bias in 1 study [30]. No downgrading was applied for other domains, as there was no heterogeneity among studies and the pooled effect estimate was relatively precise (for detailed assessment, refer to Multimedia Appendix 5).

XSL•FO
RenderX

Figure 6. Forest plot of the pooled analysis for objective sleep quality [30,54].

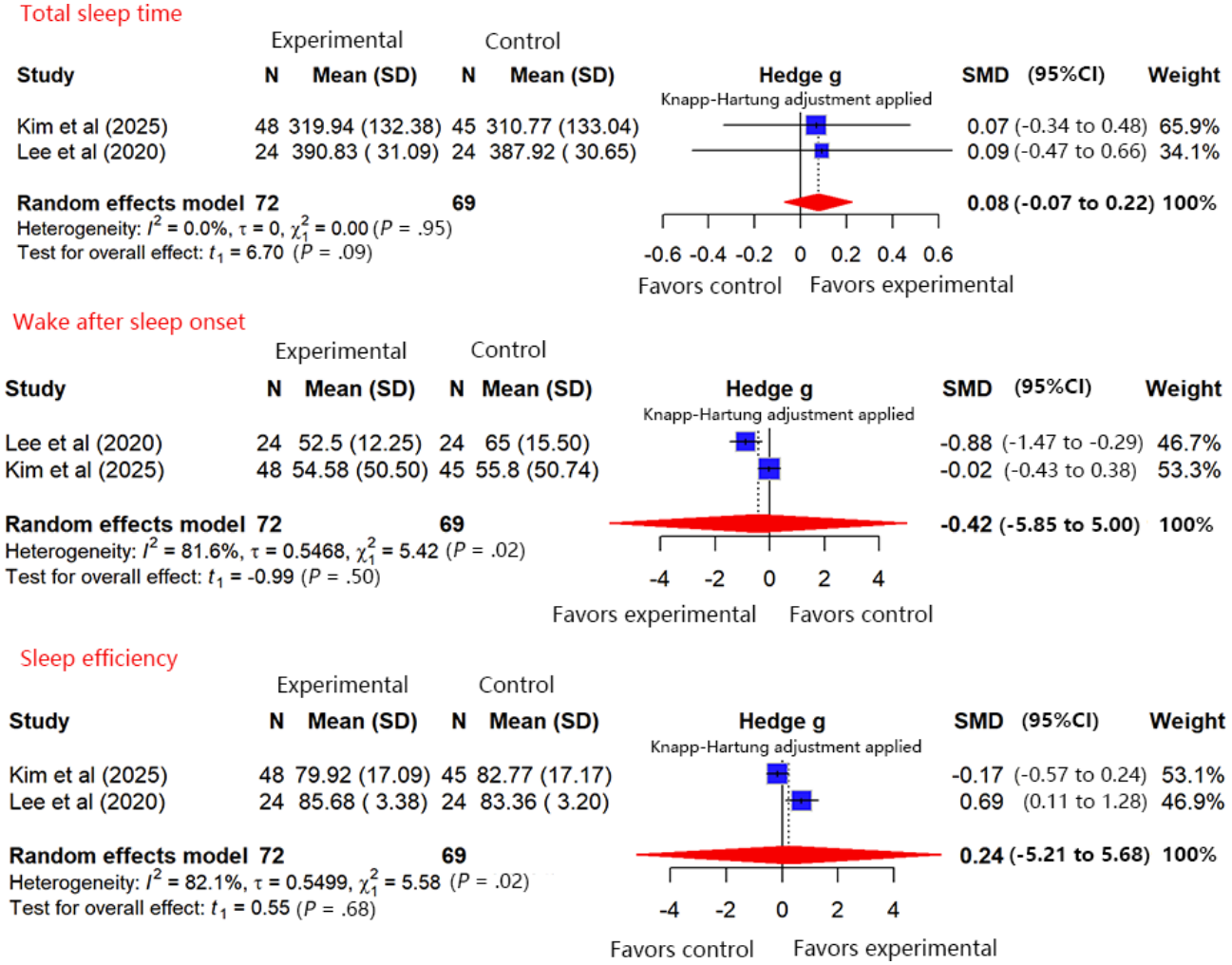

WASO

The meta-analysis indicated substantial heterogeneity among the studies (I^2=81.6%, τ=0.5468, τ²=0.2990; Cochran Q=5.42, df=1; *P*=.02). Sensitivity analysis showed that the results were robust (the maximum change in the effect size after excluding any single study was 0.0397, below the threshold of 0.1). The meta-analysis (Figure 6) showed that, regarding the average effect, VR-ERI did not lead to a statistically significant reduction in WASO compared with usual care (SMD −0.42, 95% CI −5.85 to 5.00). The negative SMD indicates that the direction of the effect points toward a reduction in WASO, signifying an improvement in sleep maintenance. According to the GRADE approach, the certainty of evidence for objective WASO during ICU treatment was rated as "very low." The rating was downgraded due to (1) a very limited number of included studies with risk of bias, (2) high heterogeneity among studies, and (3) extreme imprecision of the pooled effect estimate (the 95% CI was exceptionally wide; for detailed assessment, refer to Multimedia Appendix 5).

Sleep Efficiency

The meta-analysis indicated substantial heterogeneity among the studies (I^2=82.1%, τ=0.5499, τ²=0.3024; Cochran Q=5.58, df=1; *P*=.02). Sensitivity analysis indicated robust results (the maximum change in the effect size after excluding any single study was 0.0432, below the threshold of 0.1). Regarding the average effect, the meta-analysis (Figure 6) showed that VR-ERI did not lead to a statistically significant improvement in sleep efficiency compared with usual care (SMD 0.24, 95% CI −5.21 to 5.68). The positive SMD indicates that the direction of the effect points toward an improvement in sleep efficiency. According to the GRADE approach, the certainty of evidence for objective sleep efficiency during ICU treatment was rated as "very low." The rating was downgraded due to (1) risk of bias in the included studies, (2) substantial heterogeneity among studies, and (3) extreme imprecision of the pooled effect estimate (refer to Multimedia Appendix 5 for detailed assessment).

Pain

In total, 3 studies included pain level during ICU treatment as an outcome measure. According to the predefined data pooling criteria, for the study by Merliot-Gailhoustet et al [17], we only included the pain data from patients who received the DEEPSEN VR system. The pain outcome measure was analyzed using change-from-baseline values for the pooled analysis.

Initial meta-analysis revealed extremely high heterogeneity among the studies (I^2=95.1%, τ=0.7550, τ²=0.5700; Cochran Q=41.10, df=2; *P*<.01). Regarding the average effect, VR-ERI did not show a statistically significant difference in reducing patients' ICU-related pain compared with usual care (SMD

XSL•FO
RenderX

−0.13, 95% CI −2.06 to 1.79). Sensitivity analysis (leave-one-out method) indicated that the results may not be robust. The maximum change in the effect size after excluding a single study was 0.4062, which exceeds the prespecified threshold of 0.1. The study with the largest influence was by Laghlam et al [35]. Upon verification, this study was a noninferiority trial comparing VR therapy with another analgesic intervention. Pooling its control group with standard usual care controls could confound the pain outcome. Therefore, the study by Laghlam et al [35] was excluded, and the pooled result from the remaining 2 studies [17,34] (n=220) was used as the primary evidence for assessing the effect of VR-ERI on pain during ICU treatment. The meta-analysis of the 2 remaining studies showed low heterogeneity (I^2=23.7%, τ=0.1086, τ^2=0.0118; Cochran Q=1.31, df=1; *P*=.25; Figure 7). Regarding the average effect, VR-ERI did not demonstrate a statistically significant effect on reducing pain during ICU treatment compared with usual care (SMD −0.58, 95% CI −2.58 to 1.43). Subgroup analysis was not performed due to the limited number of studies.

Figure 7. Forest plot for pain outcomes during the intensive care unit stay [17,34].

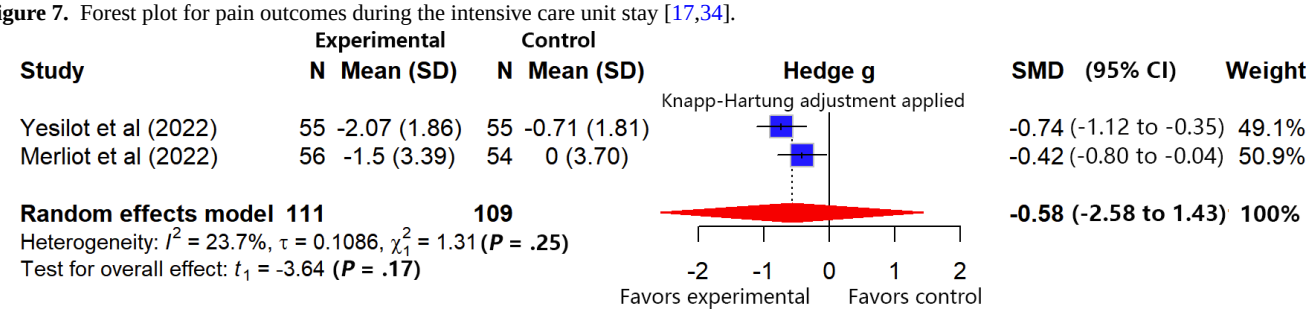

According to the GRADE criteria, the certainty of evidence for the outcome of pain during ICU treatment was rated as "low." The reasons for downgrading include (1) a high risk of bias in the included primary studies and (2) serious imprecision, as the 95% CI of the pooled effect is extremely wide and includes the line of null effect. As the heterogeneity among the remaining studies was low after the exclusion of 1 study based on clinical heterogeneity, no downgrading was applied for inconsistency (refer to Multimedia Appendix 5 for detailed assessment).

Balance

In total, 3 studies [33,56,61] (n=228) assessed patients' balance function using the Berg Balance Scale at short-term follow-up. We pooled change-from-baseline values for analysis. No heterogeneity was observed among the included studies (I^2=0%, τ=0, τ^2=0; Cochran Q=0.29, df=2; *P*=.86). The meta-analysis showed that, in terms of the average effect, VR-ERI improved patients' balance function during short-term follow-up after discharge compared with usual care (SMD 0.97, 95% CI 0.74 to 1.20). Furthermore, the 95% PI (0.37-1.58) lies entirely to the right of the line of null effect, suggesting that the positive effect of VR-ERI on balance function is likely to remain at a clinically meaningful level in most future, similar studies (Figure 8). Sensitivity analysis (leave-one-out method) indicated that the results were robust. The maximum change in the effect size after excluding any single study was 0.0461, which is below the prespecified threshold of 0.1. Subgroup analysis was not performed due to the limited number of studies.

Figure 8. Forest plot for balance outcomes [33,56,61].

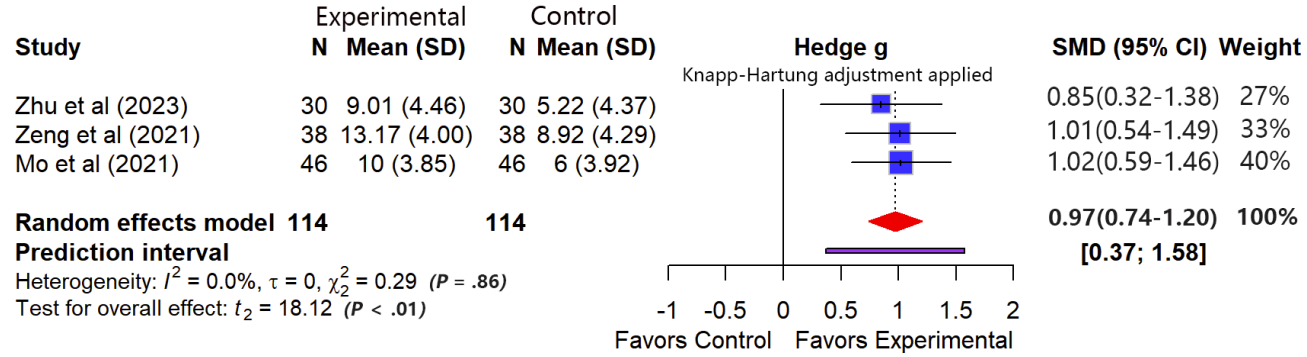

According to the GRADE approach, the certainty of evidence for balance function at short-term follow-up was rated as "moderate." The rating was downgraded because all included primary studies were assessed as having a high risk of bias. No downgrading was applied for inconsistency or imprecision, as no heterogeneity was observed among studies, and the pooled effect estimate was precise. Although the pooled effect size was large and clinically meaningful, an upgrade was not performed. This decision was made to maintain a cautious interpretation, considering that the methodological limitations commonly present in the existing studies could affect the validity of the effect estimate (refer to Multimedia Appendix 5 for the detailed evaluation process).

Cognitive Function

A total of 3 studies [33,55,57] (n=271) were included that assessed cognitive function at short-term follow-up. According to the predefined data pooling criteria, for the study by Zeng [33], we only included the results obtained using the Montreal Cognitive Assessment (MoCA) scale. All analyses were based on change-from-baseline values.

XSL•FO
RenderX

The meta-analysis (Figure 9) indicated mild heterogeneity among the studies (I^2=20.6%, τ=0.1498, τ²=0.0224; Cochran Q=2.52, df=2; *P*=.28). Sensitivity analysis (leave-one-out method) showed that the results are robust, as the maximum change in the effect size after excluding any single study was 0.0218, which is below the threshold of 0.1. The pooled results indicated that, on average, VR-ERI showed a statistically significant trend toward improving patients' cognitive function compared with usual care (SMD 0.78, 95% CI 0.16 to 1.39). However, the 95% PI (−0.16 to 1.71) crossed the line of null effect, suggesting that the beneficial effect of VR-ERI on cognitive function in most future, similar studies is unstable—it could be beneficial or show no difference and therefore requires further exploration in future research. Subgroup analysis was not performed due to the limited number of included studies.

Figure 9. Forest plot of the meta-analysis for cognitive function outcomes [33,55,57].

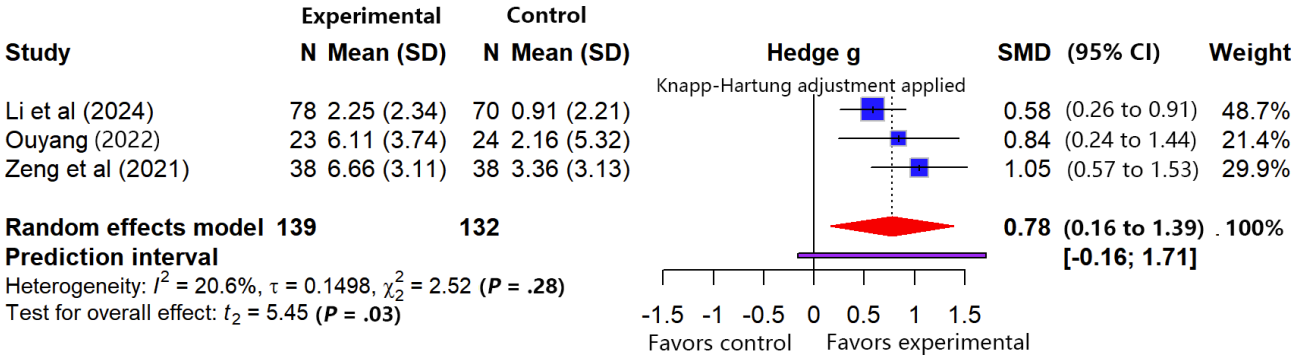

According to the GRADE framework, the quality of evidence was rated as "low." This rating was based on (1) all 3 included studies [33,55,57] were rated as having a high risk of bias, and (2) imprecision of the pooled effect estimate. Although the pooled effect size suggests a moderate effect, the methodological flaws in the studies significantly reduce our confidence in this estimate. Specific rating details are provided in Multimedia Appendix 5.

Limb Motor Function

In total, 2 studies [33,56] (n=168) assessed the improvement of patients' limb motor function during short-term follow-up using the Fugl-Meyer Assessment. As the study by Zeng [33] reported upper limb and lower limb function scores separately, we applied the weighted mean method to combine the function scores from different body parts within the same study into an overall limb motor function score. All analyses were based on change-from-baseline values.

No heterogeneity was observed between the studies (I^2=0%, τ=0, τ²=0; Cochran Q=0.55, df=1; *P*=.46). The meta-analysis (Figure 10) indicated that, based on the average effect, VR-ERI showed a trend toward improving patients' limb motor function at short-term follow-up after discharge compared to usual care (SMD 1.40, 95% CI −0.23 to 3.02), although the difference was not statistically significant and requires further verification. According to the GRADE framework, the quality of evidence was rated as "low." This rating was based on (1) both included studies had a high risk of bias, and (2) the pooled effect estimate was extremely imprecise. Although the pooled effect size suggests a potentially large effect, we did not upgrade the rating due to the very limited evidence base and the low quality of the studies (Multimedia Appendix 5).

Figure 10. Forest plot for limb motor function [33,56].

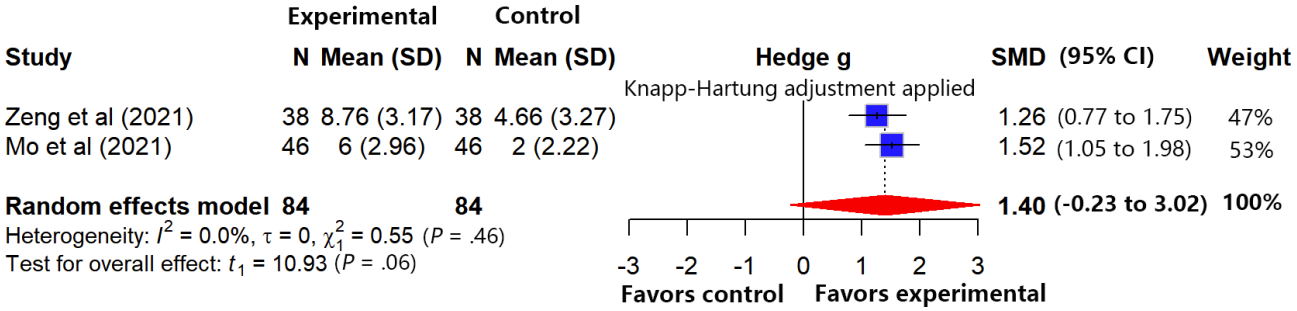

Descriptive Integration Results

Overview

We performed descriptive analysis on outcomes with fewer than 2 included studies and those unsuitable for meta-analysis due to differences in assessment methods or data collection time points. Given the limited number of studies, further validation of the intervention effects on these outcomes is currently lacking, and more high-quality RCTs are needed in the future to evaluate them.

Motor Function Outcomes (Muscle Strength, Grip Strength, Patient Engagement, and Functional Outcomes)

VR-ERI demonstrated multifaceted benefits on motor function, including improved muscle strength. In total, 2 studies supported this finding; Zhu et al [61] reported significantly higher Manual

XSL•FO
RenderX

Muscle Testing scores in the VR-ERI group versus controls (t_{58}=2.33, P=.02), while Wang et al [58] observed stronger muscle strength via the Medical Research Council scale ($F_{1,66}$=7.9, P=.006), with a maintained intervention effect over time ($F_{7,66}$=50.1, P<.001). Due to insufficient raw data, a meta-analysis was not conducted, and further studies are needed to confirm these results. In contrast, findings regarding grip strength improvement were inconsistent. Both Xia et al [59] and Wang et al [58], using a dynamometer, reported no statistically significant between-group difference in absolute grip strength post intervention. Notably, Wang et al [58] observed significant improvement over time in both groups and a significant group-time interaction, suggesting VR-ERI may influence the recovery trajectory. However, its advantage over conventional rehabilitation requires further validation.

Furthermore, VR-ERI also enhanced patient engagement. Both Xia et al [59] and Wang et al [58] reported that the intervention group had significantly longer daily exercise duration than the control group. Mo et al [56] further supported this, finding a higher proportion of patients with high exercise compliance in the VR-ERI group (33/46, 71.74%) versus the control group (25/46, 54.35%) via a self-developed questionnaire. This may be attributed to the inherent enjoyment and real-time feedback of VR-based training, although limited raw data necessitates further research. Regarding functional outcomes, VR-ERI also demonstrated positive effects. Mo et al [56] observed a higher proportion of patients attaining the highest walking ability level (Functional Ambulation Category grade 5) in the intervention group (22/46, 47.93%) versus the control group (14/46, 30.43%). Meanwhile, Xia et al [59] reported a significant reduction in ICU-acquired weakness incidence with VR-ERI (intervention: 13.5% vs control: 35.5%). Future high-quality studies are needed to confirm these effects.

Psychological Function Outcomes (Depression, Posttraumatic Stress Disorder, Somatic Symptoms, and Perceived Well-Being)

The effects of VR-ERI on psychological function were inconsistent. Navarra-Ventura et al [16] found no significant between-group differences in depression or posttraumatic stress disorder at 1 and 12 months post-ICU discharge. In contrast, Ouyang [57] reported a significant reduction in depression during ICU treatment. However, due to differing assessment timepoints between the 2 studies, meta-analysis was not performed, and these divergent conclusions require further validation. Li et al [55] observed better improvement in somatic symptoms and higher perceived well-being in the intervention group at 1 week during ICU stay, but these differences were not maintained 1 month after discharge. Due to differing assessment timepoints, meta-analysis was not performed, and the long-term psychological effects of VR-ERI require further investigation.

Cognitive Function Outcomes (Delirium)

Kim and Kang [54] used the Confusion Assessment Method (CAM) for ICU scale to evaluate the effect of VR-ERI on delirium incidence and found no significant difference between the 2 groups (12.2% vs 12.8%, χ^2_1=0.01; P=.94). In contrast, Weng et al [32] used the CAM scale to assess delirium severity and reported a statistically significant difference in CAM scores between groups (mean 2.25, SD 0.44 vs mean 6.37, SD 1.05, t_{82}=23.453; P<.05). Future high-quality studies with standardized assessment criteria are recommended to validate the intervention effects.

ICU Length of Stay and Recovery Quality

Regarding ICU length of stay, both Xia et al [59] and Wang et al [58] reported significantly shorter stays for the VR-ERI group versus controls, suggesting faster patient turnover. However, a meta-analysis was not feasible due to incomplete raw data. For quality of recovery, Zhu et al [61] (using the Postoperative Quality Recovery Scale) found significantly better recovery rates in the VR-ERI group for physiological function and activity domains. Li et al [55] noted improved mental health scores (12-Item Short Form Health Survey Mental Component Score) at 1 week post intervention, though physical health scores (12-Item Short Form Health Survey Physical Component Score) showed no difference, and no between-group differences remained significant at the 3-month follow-up.

Small-Study Effects

Although the number of included studies for all outcomes in this review did not meet the recommended threshold for conventional statistical tests of publication bias (≥10 studies), we still performed an exploratory analysis on the anxiety outcome, which had the largest number of included studies (n=8). This was done to preliminarily assess the potential impact of bias on the robustness of the conclusions.

The funnel plot (Figure 11) appeared approximately symmetrical by visual inspection. Statistical test results showed that the Egger linear regression test did not indicate significant small-study effects (t_6=−1.94; P=.10). However, the Begg rank correlation test suggested the possible presence of publication bias (z=−2.47, P=.01). To quantify the potential impact of missing studies on the pooled result, a trim-and-fill analysis was conducted. The analysis estimated that 1 potentially missing study needed to be imputed. The adjusted pooled effect size changed from the original SMD −0.862 (95% CI −1.853 to 0.129) to SMD −0.495 (95% CI −1.719 to 0.729), a change of 0.367. The decrease in the absolute value suggests that, if bias exists, its influence may have led to an overestimation of the intervention effect.

XSL•FO
RenderX

Figure 11. Funnel plot for anxiety outcomes[16,17,32,34,35,55,57,60].

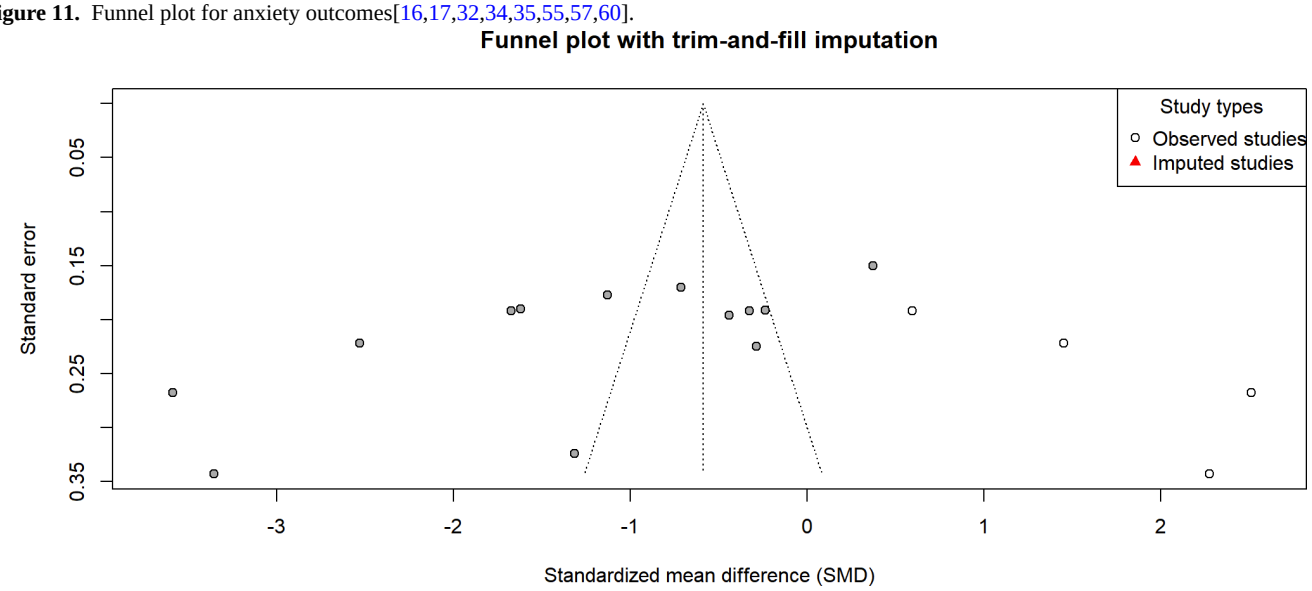

We fully acknowledge that with only 8 included studies [16,17,32,34,35,55,57,60] and substantial heterogeneity present, the statistical power of these tests is limited. Furthermore, heterogeneity itself can interfere with the interpretation of funnel plot asymmetry. Therefore, the above results should be considered exploratory. In summary, the current analysis, while suggesting that publication bias cannot be entirely ruled out, does not provide evidence from the Egger test supporting significant small-study effects. Moreover, the CI for the effect size adjusted by the trim-and-fill method remained wide and crossed the line of null effect. This indicates that, even if potential bias exists, the current conclusion regarding the anxiety outcome is stable. However, it serves as a reminder to interpret the magnitude of this effect size with caution in the future.

Discussion

Principal Results

This review systematically assessed the feasibility, safety, and rehabilitation outcomes of VR-ERI in critically ill patients across 16 RCTs [16,17,30,32-35,53-61] (n=1356). VR-ERI was well-accepted and associated with high exercise adherence. Vital signs remained within acceptable ranges during use, and no serious adverse events were reported; however, mild dizziness, nausea, and fatigue occurred in some studies, underscoring the need for standardized safety monitoring. Regarding efficacy, VR-ERI improved balance during short-term follow-up, supported by low heterogeneity, narrow PIs, and moderate-certainty evidence, indicating robust and potentially generalizable benefit. VR-ERI also shows potential to improve in-ICU anxiety and subjective sleep quality, as well as short-term cognitive and motor function, particularly sleep, when assessed with the RCSQ. However, it did not significantly improve objective sleep parameters or in-ICU pain. For most outcomes, the certainty of evidence was low or very low due to implementation bias (eg, lack of blinding), study inconsistency, and imprecision in pooled estimates. Thus, results should be interpreted cautiously, and current evidence remains insufficient for definitive conclusions. Some outcomes could not be pooled due to limited studies. Further high-quality trials are needed to confirm these findings and clarify the clinical value of VR-ERI in critical care.

VR-ERI may reduce anxiety in critically ill patients, a finding consistent with recent reviews in both critical care [36] and elective surgery [62], suggesting VR could broadly mitigate hospital-related anxiety. Stress recovery theory [63] offers a possible explanation, proposing that exposure to natural landscapes can activate the parasympathetic nervous system and promote emotional relaxation. Research confirms that biophilic elements aid stress recovery [64]. Most VR-ERI scenarios in the included studies incorporated natural views (eg, mountains and rivers), which may help block typical ICU stressors, such as constant light, noise, and isolation. However, the pooled effect in this analysis was not statistically significant, and the 95% PI crossed zero, indicating instability in the estimate. The included studies also had a high risk of bias, with substantial inconsistency and imprecision, yielding very low certainty of evidence. Thus, future rigorous RCTs with adequate sample sizes and transparent reporting are needed to clarify VR's effect on anxiety in this population.

VR-ERI shows potential in improving ICU patients' subjective sleep quality, likely due to its intervention design—2 protocols [30,54] used VR meditation and hypnosis modules to promote relaxation and distraction, enhancing perceived sleep quality [30,54]. However, objective parameters measured with smart wristbands (eg, TST and sleep efficiency) did not improve concurrently, suggesting VR may modulate sleep perception rather than alter sleep physiology. A recent network meta-analysis [65] found that simple interventions like eye masks and earplugs offer more consistent sleep benefits in critically ill patients, while the effect of VR meditation remains unestablished, underscoring VR's uncertain role in sleep intervention. Subgroup analysis indicated positive effects only when sleep was assessed with the RCSQ, not with other subjective scales. This highlights how tool selection influences outcomes, possibly because the RCSQ focuses more on immediate sleep depth and is designed specifically for critically ill patients. Future research should use standardized, population-specific sleep assessment tools to accurately capture

XSL•FO
RenderX

intervention effects. Current evidence is limited by low certainty. Future studies should develop standardized VR sleep protocols, clarify optimal dosage, and incorporate objective measures, such as polysomnography or validated wearable monitoring, to systematically evaluate VR's impact on sleep architecture, thereby building a more reliable evidence base for practice.

We also observed that VR-ERI can improve balance (measured by Berg Balance Scale) and limb motor function (measured by Fugl-Meyer Assessment) in critically ill patients during short-term follow-up. This finding is consistent with the review by He et al [37] on VR for early motor intervention in this population. By providing an immersive environment, VR compensates for the sensory and interaction deficits resulting from prolonged immobilization and isolation. It enhances sensorimotor pathways and balance-related neuroplasticity through visual-vestibular-proprioceptive integration [66]. Furthermore, VR-ERI engages both cognitive and motor functions simultaneously [67] and promotes cortical activation via feedback-driven learning and task-specific adaptation [68,69]. The engaging and gamified design of VR-ERI may also improve patient adherence to exercise, which could explain its advantage over traditional motor rehabilitation. Improvement in balance is supported by moderate-quality evidence and relatively narrow PIs, suggesting that standardized VR-ERI motor activities could be implemented early in critical care rehabilitation, with attention to personalized parameters and long-term effect tracking. However, evidence for limb motor improvement remains low in certainty and requires further rigorous validation.

However, VR-ERI did not demonstrate a statistically significant effect on reducing pain in critically ill patients during ICU treatment, a result that may be associated with several factors. The pain experience of critically ill patients is particularly complex, as frequent threats to psychological and physical integrity make their pain difficult to manage effectively [70]. Neuroimaging studies have shown that receiving VR intervention during painful stimuli can significantly reduce excitability in brain regions associated with pain processing, with a higher level of VR immersion correlating with stronger inhibitory effects [71]. Concurrently, immersive VR devices can divert attention away from nociceptive stimuli by isolating patients' audiovisual connection to the external environment [72]. Nevertheless, in this study, the reported pain scores were relatively low in both groups, which might be related to insufficient immersion of the VR experience, inadequate intervention dosage, or the widespread use of analgesic medications. This warrants further investigation in future studies.

VR-ERI shows a potential trend for improving cognitive function in short-term follow-up, although the 95% PI crosses the null effect line, indicating unstable results. This may stem from several factors. First, the pathophysiology of post-ICU cognitive impairment is complex (eg, neuroinflammation and blood-brain barrier disruption) [73]; for instance, underlying neural damage in conditions like posterior cortical atrophy may limit response to conventional interventions [74], highlighting the need for more targeted programs. Second, widely used screening tools, such as the Mini-Mental State Examination and MoCA, may lack sensitivity to detect subtle, clinically meaningful improvements in critically ill patients. The Mini-Mental State Examination, although brief and simple, has a ceiling effect and inadequately assesses executive function and complex attention [75]. The MoCA is more sensitive to mild impairment but remains a broad screening tool that may miss specific post-ICU executive deficits (eg, inhibitory control and cognitive flexibility) [76]. Therefore, future studies should develop VR modules tailored to ICU-related cognitive impairment and consider longer intervention duration. A multilevel assessment strategy is recommended, combining screening scales with computerized neuropsychological tests (eg, flanker tasks) for objective, quantitative data [77]. Integrating VR with other rehabilitation methods into a multidimensional intervention may also help achieve meaningful cognitive gains. Given the low certainty of current evidence, these findings should be interpreted with caution.

This systematic review is the first to focus on VR intervention studies conducted during the early ICU rehabilitation phase. By systematically searching ten Chinese and English databases and implementing independent dual-reviewer screening and data extraction, it strictly adhered to the PRISMA 2020 guidelines, ensuring a transparent process and reproducible results. The review comprehensively used the Cochrane RoB 2.0 tool for risk of bias assessment and the GRADE framework for evidence grading, supplemented by sensitivity analyses to verify the robustness of the findings. It summarizes and analyzes the safety and clinical feasibility of VR intervention in critically ill patients, while systematically elucidating the efficacy profile of VR on relevant outcomes during both the ICU stay and at short-term follow-up. This provides an important reference for advancing the translational application of VR technology in this patient population.

Limitations and Future Recommendations

Several limitations should be acknowledged. First, the literature search was restricted to Chinese and English publications and excluded trial registries and gray literature, potentially introducing omission and language bias. Second, the limited number of studies and small total sample size reduce statistical power, which may affect the robustness and generalizability of the pooled results. Third, substantial clinical and methodological heterogeneity existed across studies, including variations in VR immersion levels, protocol design, dosage, and outcome measures, contributing to high statistical heterogeneity. Inadequate reporting of precise dosage parameters also limited dose-response analyses. Fourth, most included studies had a high risk of bias due to issues such as unclear allocation concealment, lack of assessor blinding, and inherent challenges in blinding participants and personnel. Consequently, the certainty of evidence for most outcomes was low or very low, necessitating cautious interpretation. Finally, some outcomes were supported by only a few studies, resulting in meta-analyses with limited sample sizes and potentially unstable effect estimates. In summary, further well-designed, large-scale studies are required to validate these findings.

At the clinical level, VR-ERI can supplement traditional early rehabilitation, but widespread adoption faces systematic challenges. First, cost remains a key barrier. Although

XSL•FO
RenderX

commercial VR prices are falling, expenses may still burden some institutions or patients, with VR for early pain management costing approximately €47.48 (approximately US $51.30) per patient [78]. Policymakers should thus consider including VR in insurance reimbursement or supporting it via dedicated funding. Second, patient vulnerability limits generalizability [79]. Those who are older adults, or have vestibular dysfunction, severe cognitive impairment, or complex life-support needs may be unable to tolerate VR. Health care institutions should form multidisciplinary teams to create standardized screening protocols—focusing on vestibular and cognitive function—and implement a safety system covering pre-, intra-, and postsession phases. Implementation should be phased and individualized—starting at low intensity and adjusted gradually under therapist supervision, while training ward staff and patients in device use. Finally, before broader promotion, safety and feasibility must be validated in heterogeneous critically ill populations [14].

At the policy level, relevant departments should establish a comprehensive support framework by developing evidence-based clinical guidelines for VR rehabilitation that specify indications, contraindications, and protocols, and implementing unified device certification and quality control systems to ensure safe and standardized application. Future research should adopt a phased strategy—priorities include conducting large-scale, rigorous pragmatic trials using blinded outcome assessment, diverse control groups, objective monitoring, and long-term follow-up to minimize bias; creating personalized interventions for specific patient groups, defining a core outcome set, and using objective, standardized assessment tools; and enhancing implementation science to identify optimal rollout pathways across different resource settings—beginning with pilot projects at regional medical centers before scaling to primary care. Concurrently, interdisciplinary collaboration among engineering, rehabilitation medicine, and critical care nursing should be strengthened to jointly advance technology innovation and support the development of a more integrated early rehabilitation system for critically ill patients.

Conclusions

This review is the first to examine VR rehabilitation effectiveness specifically during early intervention in the ICU. Unlike previous reviews, we focused strictly on the in-ICU early rehabilitation period to establish an evidence base for initiating such care and to minimize the clinical heterogeneity from combining interventions across different care settings. Evidence was synthesized across physiological, psychological, and cognitive domains. Results indicate that VR-ERI shows considerable potential for improving anxiety and subjective sleep quality during ICU treatment, as well as balance, limb motor function, and cognitive function in short-term follow-up. However, effects on pain and objective sleep quality in the ICU remain unclear. Given the high risk of bias in included studies, substantial heterogeneity, and imprecise pooled effect estimates, the evidence certainty for most outcomes is low, and conclusions should be interpreted cautiously. Additionally, establishing a comprehensive safety monitoring system is essential for future clinical translation. This review offers a new perspective on early ICU rehabilitation, suggesting VR technology could serve as a nonpharmacological adjunct in this phase. However, the current evidence is insufficient to support widespread implementation. Clinical translation requires careful consideration of cost and patient suitability, and long-term benefits need validation through large-scale and rigorous research.

Acknowledgments

The authors declare that no generative artificial intelligence was used at any stage of this study.

Funding

This work was supported by the Capital Medical University Digital Intelligence Nursing Basic Clinical Collaborative Research Project (Grant SZHL23Z02) and the Beijing Hospitals Authority Youth Programme (Grant QML20230515). The former was led by principal investigator (WC), who supervised the research and reviewed the manuscript. The latter was led by the corresponding author (RZ), who was responsible for the study conception and project guidance.

Data Availability

Relevant data are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed search strategy.
[DOCX File, 48 KB - jmir_v28i1e81865_app1.docx]

Multimedia Appendix 2
Excluded literature and reasons.
[DOCX File, 69 KB - jmir_v28i1e81865_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Characteristics of included studies.
[DOCX File, 34 KB - jmir_v28i1e81865_app3.docx]

Multimedia Appendix 4
Dual Revised Cochrane Risk-of-Bias tool for randomized trials assessment process with interrater consistency.
[XLS File, 1604 KB - jmir_v28i1e81865_app4.xls]

Multimedia Appendix 5
Grading of Recommendations Assessment, Development, and Evaluation assessment results.
[DOCX File, 34 KB - jmir_v28i1e81865_app5.docx]

Checklist 1
PRISMA checklist for abstracts.
[PDF File, 119 KB - jmir_v28i1e81865_app6.pdf]

Checklist 2
PRISMA-S checklist.
[PDF File, 66 KB - jmir_v28i1e81865_app7.pdf]

Checklist 3
PRISMA checklist.
[PDF File, 321 KB - jmir_v28i1e81865_app8.pdf]

References

1. Gao S, Liang X, Pan Z, Zhang X, Zhang L. Effect size estimates of risk factors for post-intensive care syndrome: a systematic review and meta-analysis. Intensive Crit Care Nurs 2025 Apr;87:103888. [doi: 10.1016/j.iccn.2024.103888] [Medline: 39561481]
2. Fresenko LE, Rivera ZC, Parry SM, Mayer KP. Post-intensive care syndrome: physical impairments and function. Crit Care Clin 2025 Jan;41(1):1-20. [doi: 10.1016/j.ccc.2024.08.004] [Medline: 39547718]
3. Needham DM, Davidson J, Cohen H, et al. Improving long-term outcomes after discharge from intensive care unit: report from a stakeholders' conference. Crit Care Med 2012 Feb;40(2):502-509. [doi: 10.1097/CCM.0b013e318232da75] [Medline: 21946660]
4. Ho MH, Lee YW, Wang L. Estimated prevalence of post-intensive care cognitive impairment at short-term and long-term follow-ups: a proportional meta-analysis of observational studies. Ann Intensive Care 2025 Jan 10;15(1):3. [doi: 10.1186/s13613-025-01429-z] [Medline: 39792310]
5. Valsø Å, Rustøen T, Småstuen MC, et al. Occurrence and characteristics of pain after ICU discharge: a longitudinal study. Nurs Crit Care 2022 Sep;27(5):718-727. [doi: 10.1111/nicc.12701] [Medline: 34382725]
6. Fukamachi Y, Sakuramoto H, Sato T, et al. Prevalence, risk factors, and intervention of long-term sleep disturbance after intensive care unit discharge: a scoping review. Cureus 2025 Apr;17(4):e83011. [doi: 10.7759/cureus.83011] [Medline: 40421347]
7. Yoo KH, Lee J, Oh J, et al. Depression or anxiety and long-term mortality among adult survivors of intensive care unit: a population-based cohort study. Crit Care 2025 May 6;29(1):179. [doi: 10.1186/s13054-025-05381-z] [Medline: 40329374]
8. Ahn S, LaNoue M, Su H, et al. Post-intensive care syndrome and caregiver burden: a post hoc analysis of a randomized clinical trial. JAMA Netw Open 2025 Apr 1;8(4):e253443. [doi: 10.1001/jamanetworkopen.2025.3443] [Medline: 40198074]
9. Mikkelsen ME, Still M, Anderson BJ, et al. Society of Critical Care Medicine's International Consensus Conference on Prediction and Identification of Long-Term Impairments After Critical Illness. Crit Care Med 2020 Nov;48(11):1670-1679. [doi: 10.1097/CCM.0000000000004586] [Medline: 32947467]
10. Dayton K, Lindroth H, Engel HJ, et al. Creating a culture of an awake and walking intensive care unit: in-hospital strategies to mitigate post-intensive care syndrome. Crit Care Clin 2025 Jan;41(1):121-140. [doi: 10.1016/j.ccc.2024.08.002] [Medline: 39547720]
11. Tipping CJ, Harrold M, Holland A, Romero L, Nisbet T, Hodgson CL. The effects of active mobilisation and rehabilitation in ICU on mortality and function: a systematic review. Intensive Care Med 2017 Feb;43(2):171-183. [doi: 10.1007/s00134-016-4612-0] [Medline: 27864615]
12. Ding N, Zhang Z, Zhang C, et al. What is the optimum time for initiation of early mobilization in mechanically ventilated patients? A network meta-analysis. PLoS ONE 2019;14(10):e0223151. [doi: 10.1371/journal.pone.0223151] [Medline: 31589642]

XSL•FO
RenderX

13. Haghedooren E, Haghedooren R, Langer D, Gosselink R. Feasibility and safety of interactive virtual reality upper limb rehabilitation in patients with prolonged critical illness. Aust Crit Care 2024 Nov;37(6):949-956. [doi: 10.1016/j.aucc.2024.06.004] [Medline: 39054204]
14. de Vries M, Beumeler LFE, van der Meulen J, Bethlehem C, den Otter R, Boerma EC. The feasibility of virtual reality therapy for upper extremity mobilization during and after intensive care unit admission. PeerJ 2025;13:e18461. [doi: 10.7717/peerj.18461] [Medline: 39802193]
15. Godoy-González M, López-Aguilar J, Fernández-Gonzalo S, et al. Efficacy and safety of a non-immersive virtual reality-based neuropsychological intervention for cognitive stimulation and relaxation in patients with critical illness: study protocol of a randomized clinical trial (RGS-ICU). BMC Psychiatry 2024 Dec 18;24(1):917. [doi: 10.1186/s12888-024-06360-4] [Medline: 39696098]
16. Navarra-Ventura G, Gomà G, de Haro C, et al. Virtual reality-based early neurocognitive stimulation in critically ill patients: a pilot randomized clinical trial. J Pers Med 2021 Nov 29;11(12):1260. [doi: 10.3390/jpm11121260] [Medline: 34945732]
17. Merliot-Gailhoustet L, Raimbert C, Garnier O, et al. Discomfort improvement for critically ill patients using electronic relaxation devices: results of the cross-over randomized controlled trial E-CHOISIR (Electronic-CHOIce of a System for Intensive care Relaxation). Crit Care 2022 Sep 3;26(1):263. [doi: 10.1186/s13054-022-04136-4] [Medline: 36057612]
18. Alvarado Sánchez JI, Castillo Morales LM, Cardenas Bolivar YR, et al. Efficacy and safety of the early implementation of a multimodal rehabilitation program in mechanically ventilated patients: a randomized clinical trial protocol. PLoS ONE 2025;20(5):e0324335. [doi: 10.1371/journal.pone.0324335] [Medline: 40388519]
19. Matsuoka A, Yoshihiro S, Shida H, et al. Effects of mobilization within 72 h of ICU admission in critically ill patients: an updated systematic review and meta-analysis of randomized controlled trials. J Clin Med 2023 Sep 11;12(18):5888. [doi: 10.3390/jcm12185888] [Medline: 37762829]
20. Fuke R, Hifumi T, Kondo Y, et al. Early rehabilitation to prevent postintensive care syndrome in patients with critical illness: a systematic review and meta-analysis. BMJ Open 2018 May 5;8(5):e019998. [doi: 10.1136/bmjopen-2017-019998] [Medline: 29730622]
21. Xu J, Wang S, Zhang Q, Yao Y, Yu J. Effectiveness of nurse-led early mobility protocols on the outcomes of critical care patients: a systematic review and meta-analysis. Nurs Open 2025 May;12(5):e70206. [doi: 10.1002/nop2.70206] [Medline: 40371936]
22. Lanier J. Virtual Reality: CITIC Publishing Group Corporation Ltd; 2018.
23. Abrams Z. Enhancing mental health care with VR. IEEE Pulse 2022;13(5):16-20. [doi: 10.1109/MPULS.2022.3208824] [Medline: 36301813]
24. Bauer ACM, Andringa G. The potential of immersive virtual reality for cognitive training in elderly. Gerontology 2020;66(6):614-623. [doi: 10.1159/000509830] [Medline: 32906122]
25. Zhang Q, Fu Y, Lu Y, et al. Impact of virtual reality-based therapies on cognition and mental health of stroke patients: systematic review and meta-analysis. J Med Internet Res 2021 Nov 17;23(11):e31007. [doi: 10.2196/31007] [Medline: 34787571]
26. Li R, Li Y, Kong Y, et al. Virtual reality-based training in chronic low back pain: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2024 Feb 26;26:e45406. [doi: 10.2196/45406] [Medline: 38407948]
27. Guo Q, Zhang L, Gui C, et al. Virtual reality intervention for patients with neck pain: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2023 Apr 3;25:e38256. [doi: 10.2196/38256] [Medline: 37010891]
28. Schröder D, Wrona KJ, Müller F, Heinemann S, Fischer F, Dockweiler C. Impact of virtual reality applications in the treatment of anxiety disorders: a systematic review and meta-analysis of randomized-controlled trials. J Behav Ther Exp Psychiatry 2023 Dec;81:101893. [doi: 10.1016/j.jbtep.2023.101893] [Medline: 37453405]
29. Yen HY, Chiu HL. Virtual reality exergames for improving older adults' cognition and depression: a systematic review and meta-analysis of randomized control trials. J Am Med Dir Assoc 2021 May;22(5):995-1002. [doi: 10.1016/j.jamda.2021.03.009] [Medline: 33812843]
30. Lee SY, Kang J. Effect of virtual reality meditation on sleep quality of intensive care unit patients: a randomised controlled trial. Intensive Crit Care Nurs 2020 Aug;59:102849. [doi: 10.1016/j.iccn.2020.102849] [Medline: 32241625]
31. Ong TL, Ruppert MM, Akbar M, et al. Improving the intensive care patient experience with virtual reality-a feasibility study. Crit Care Explor 2020 Jun;2(6):e0122. [doi: 10.1097/CCE.0000000000000122] [Medline: 32695991]
32. Weng D, Cui M, Jiang J. Application of 5g enabled virtual reality technology in the intervention of delirium in intensive care unit patients. Chinese Journal of Integrated Traditional and Western Medicine in Intensive and Critical Care 2023;30(6):693-696. [doi: 10.3969/j.issn.1008-9691.2023.06.011]
33. Zeng Y. Application effects of cognitive behavioral intervention combined with virtual reality training in postoperative patients with severe traumatic brain injury. Henan Medical Research 2021;30(28). [doi: 10.3969/j.issn.1004-437X.2021.28.058]
34. Yesilot SB, Yeşilkuş R, Beyaz F. Use of virtual reality for reducing pain and anxiety after laparoscopic sleeve gastrectomy: a randomized controlled trial. Pain Manag Nurs 2022 Dec;23(6):826-831. [doi: 10.1016/j.pmn.2022.07.001] [Medline: 35934661]

XSL•FO
RenderX

35. Laghlam D, Naudin C, Coroyer L, et al. Virtual reality vs. Kalinox® for management of pain in intensive care unit after cardiac surgery: a randomized study. Ann Intensive Care 2021 May 13;11(1):74. [doi: 10.1186/s13613-021-00866-w] [Medline: 33983498]
36. Liu Q, Zhang Y, Chen X, Ding B, Guo X, Gai Y. Efficacy of virtual reality in alleviating post-ICU syndrome symptoms: a systematic review and meta-analysis. Nurs Crit Care 2025 Mar;30(2):e70004. [doi: 10.1111/nicc.70004] [Medline: 40070039]
37. He Y, Yang Q, Dai X, et al. Effects of virtual reality technology on early mobility in critically ill adult patients: a systematic review and meta-analysis. Front Neurol 2024;15:1469079. [doi: 10.3389/fneur.2024.1469079] [Medline: 39975851]
38. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
39. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc 2021 Apr 1;109(2):174-200. [doi: 10.5195/jmla.2021.962] [Medline: 34285662]
40. Higgins JPT, Thomas J. In: Chandler J, editor. Cochrane Handbook for Systematic Reviews of Interventions Version 65: Cochrane; 2024. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook [accessed 2025-12-04]
41. Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. BMC Med Res Methodol 2005;5(1):13. [doi: 10.1186/1471-2288-5-13]
42. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
43. Viechtbauer W. Conducting meta-analyses in r with the metafor package. J Stat Softw 2010;36(3). [doi: 10.18637/jss.v036.i03]
44. Murad MH, Wang Z, Chu H, Lin L. When continuous outcomes are measured using different scales: guide for meta-analysis and interpretation. BMJ 2019 Jan 22;364:k4817. [doi: 10.1136/bmj.k4817] [Medline: 30670455]
45. Lin L, Aloe AM. Evaluation of various estimators for standardized mean difference in meta-analysis. Stat Med 2021 Jan 30;40(2):403-426. [doi: 10.1002/sim.8781] [Medline: 33180373]
46. Braitman LE. Statistical power analysis in medical research. Ann Intern Med 1983 Aug;99(2):269-271. [doi: 10.7326/0003-4819-99-2-269] [Medline: 6349459]
47. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
48. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986 Sep;7(3):177-188. [doi: 10.1016/0197-2456(86)90046-2] [Medline: 3802833]
49. Bowden J, Tierney JF, Copas AJ, Burdett S. Quantifying, displaying and accounting for heterogeneity in the meta-analysis of RCTs using standard and generalised Q statistics. BMC Med Res Methodol 2011 Apr 7;11:41. [doi: 10.1186/1471-2288-11-41] [Medline: 21473747]
50. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 6;327(7414):557-560. [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
51. IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open 2016 Jul 12;6(7):e010247. [doi: 10.1136/bmjopen-2015-010247] [Medline: 27406637]
52. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-926. [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
53. Caballero LG, Fraga IB, Tremea CEM, et al. Experience of heart failure patients in mobilization with virtual reality: mixed methods study. Rev Esc Enferm USP 2025;59:e20250114. [doi: 10.1590/1980-220X-REEUSP-2025-0114en] [Medline: 40955916]
54. Kim S, Kang J. Effects of virtual reality meditation on sleep and delirium in ICU patients: a randomized controlled trial. Comput Inform Nurs 2025 Apr 2;43(8):e01307. [doi: 10.1097/CIN.0000000000001307] [Medline: 40194914]
55. Li YY, Peng J, Ping YY, et al. Virtual reality-based cognitive-behavioural therapy for the treatment of anxiety in patients with acute myocardial infarction: a randomised clinical trial. Gen Psychiatr 2024;37(2):e101434. [doi: 10.1136/gpsych-2023-101434] [Medline: 38645380]
56. Mo Y, Wang H, Ye P. Application of strengthening rehabilitation nursing multidisciplinary combined with virtual reality training in severe patients with multiple fractures. China Medical Herald 2021;18(14):173-176. [doi: 10.20047/j.issn1673-7210.2021.14.042]
57. Ouyang Y. Study on Influencing Factors of Cognitive Dysfunction and Intervention of Virtual Reality Nature Therapy in Critical Patients: Fujian Medical University; 2022. [doi: 10.27020/d.cnki.gfjyu.2022.001308]
58. Wang J, Shi C, Jia Y, Xiao Q. Effectiveness of virtual reality assisted active limb movement exercises for patients in the respiratory intensive care unit: a randomized pilot study. J Rehabil Med 2025 Jun 3;57:jrm28399. [doi: 10.2340/jrm.v57.28399] [Medline: 40462528]
59. Xia Y, Pan W, Zhao Y, Chen Z, Cai S, Tang Y, et al. Application of virtual reality technology in early activities of ICU patients. Chinese Journal of Nursing 2023;58(9):1050-1055.

XSL•FO
RenderX

60. Yani Y, Wenyi T, Zhenhong F, Wenyu L, Mengyao W, Ruhui C. Improvement of sleep quality and anxiety among cardiac surgery patients by preoperative visits based on virtual reality technology. Journal of Nursing Science 2024;39(22):112-115. [doi: 10.3870/j.issn.1001-4152.2024.22.112]
61. Zhu H, Jie H, Wang X, et al. Application of virtual reality technology in early rehabilitation training of patients after liver transplantation. Nurs Res 2023;37(20):3790-3792 [FREE Full text] [Medline: 37916845]
62. Li H, Chiu PL, Efendi D, Huang H, Ko KY, Wong CL. Effects of virtual reality-based interventions on preoperative anxiety in patients undergoing elective surgery with anesthesia: systematic review and meta-analysis. J Med Internet Res 2025 Apr 30;27:e55291. [doi: 10.2196/55291] [Medline: 40305092]
63. Brosschot JF, Pieper S, Thayer JF. Expanding stress theory: prolonged activation and perseverative cognition. Psychoneuroendocrinology 2005 Nov;30(10):1043-1049. [doi: 10.1016/j.psyneuen.2005.04.008] [Medline: 15939546]
64. Yin J, Yuan J, Arfaei N, Catalano PJ, Allen JG, Spengler JD. Effects of biophilic indoor environment on stress and anxiety recovery: a between-subjects experiment in virtual reality. Environ Int 2020 Mar;136:105427. [doi: 10.1016/j.envint.2019.105427] [Medline: 31881421]
65. Huang D, Li Y, Ye J, Liu C, Shen D, Lv Y. Different nursing interventions on sleep quality among critically ill patients: a systematic review and network meta-analysis. Medicine (Baltimore) 2023 Dec 29;102(52):e36298. [doi: 10.1097/MD.0000000000036298] [Medline: 38206715]
66. Wankhede NL, Koppula S, Ballal S, et al. Virtual reality modulating dynamics of neuroplasticity: innovations in neuro-motor rehabilitation. Neuroscience 2025 Feb 6;566:97-111. [doi: 10.1016/j.neuroscience.2024.12.040] [Medline: 39722287]
67. Morone G, Spitoni GF, De Bartolo D, et al. Rehabilitative devices for a top-down approach. Expert Rev Med Devices 2019 Mar;16(3):187-195. [doi: 10.1080/17434440.2019.1574567] [Medline: 30677307]
68. Cervera MA, Soekadar SR, Ushiba J, et al. Brain-computer interfaces for post-stroke motor rehabilitation: a meta-analysis. Ann Clin Transl Neurol 2018 May;5(5):651-663. [doi: 10.1002/acn3.544] [Medline: 29761128]
69. Lavi R, Pau M, Dreyer-Alster S, Drori T, Cocco E, Kalron A. Real-time cortical activity during virtual reality practice in people with multiple sclerosis: a pilot fNIRS study. J Neuroeng Rehabil 2025 Jul 4;22(1):148. [doi: 10.1186/s12984-025-01681-1] [Medline: 40615921]
70. Leininger S. The challenges of managing pain in the intensive care unit. Crit Care Nurs Q 2022;45(4):325-331. [doi: 10.1097/CNQ.0000000000000424] [Medline: 35980794]
71. Hoffman HG, Richards TL, Van Oostrom T, et al. The analgesic effects of opioids and immersive virtual reality distraction: evidence from subjective and functional brain imaging assessments. Anesth Analg 2007 Dec;105(6):1776-1783. [doi: 10.1213/01.ane.0000270205.45146.db] [Medline: 18042882]
72. Matamala-Gomez M, Donegan T, Świdrak J. VR for pain relief. Curr Top Behav Neurosci 2023;65:309-336. [doi: 10.1007/7854_2022_402] [Medline: 36592274]
73. Pandharipande PP, Girard TD, Jackson JC, et al. Long-term cognitive impairment after critical illness. N Engl J Med 2013 Oct 3;369(14):1306-1316. [doi: 10.1056/NEJMoa1301372] [Medline: 24088092]
74. Suárez-González A, Bier N, Sauvageau H, Pelak VS, Ahmed S. Cognitive rehabilitation in posterior cortical atrophy. Pract Neurol 2025 Jan 16;25(1):40-44. [doi: 10.1136/pn-2024-004259] [Medline: 39515839]
75. Pfoh ER, Chan KS, Dinglas VD, et al. Cognitive screening among acute respiratory failure survivors: a cross-sectional evaluation of the Mini-Mental State Examination. Crit Care 2015 May 5;19(1):220. [doi: 10.1186/s13054-015-0934-5] [Medline: 25939482]
76. Fu C, Jin X, Chen B, et al. Comparison of the Mini-Mental State Examination and Montreal Cognitive Assessment executive subtests in detecting post-stroke cognitive impairment. Geriatr Gerontol Int 2017 Dec;17(12):2329-2335. [doi: 10.1111/ggi.13069] [Medline: 28675607]
77. Lin LT, Lin HA, Bai CH, Lin SF. Flanker task assessment of inhibitory control in ICU survivors: preliminary insights into cognitive impairment. Behav Brain Res 2025 Oct 18;495:115791. [doi: 10.1016/j.bbr.2025.115791] [Medline: 40907667]
78. Groenveld TD, van Dorst P, Ten Broek RPG, et al. Early health economic analysis of virtual reality therapy for pain management after surgery. Int J Surg 2025 Jun 1;111(6):3798-3809. [doi: 10.1097/JS9.0000000000002385] [Medline: 40387696]
79. Li M, Race M, Huang F, Escalon MX. Role of virtual reality to promote mobilization in the critical care setting: a narrative review. Am J Phys Med Rehabil 2025 May 1;104(5):487-494. [doi: 10.1097/PHM.0000000000002648] [Medline: 39642293]

Abbreviations

CAM: Confusion Assessment Method
CNKI: China National Knowledge Internet
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
ICU: intensive care unit
MoCA: Montreal Cognitive Assessment
PI: prediction interval
PICOS: Population, Intervention, Comparison, Outcome, and Study Design

XSL•FO
RenderX

PICS: post-intensive care syndrome
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
RCSQ: Richards-Campbell Sleep Questionnaire
RCT: randomized controlled trials
RoB 2: revised Cochrane Risk of Bias tool
SinoMed: SinoMed Service System
SMD: standardized mean difference
TST: total sleep time
VIP: VIP Database
VR: virtual reality
VR-ERI: virtual reality–based early rehabilitation intervention
WASO: wake after sleep onset

Edited by S Brini; submitted 18.Aug.2025; peer-reviewed by A Shivanna, MDP Beristain-Colorado, O Oluwole, U Ogbonna; accepted 29.Dec.2025; published 06.Mar.2026.

Please cite as:
Wu F, Wu Y, Xing Y, Cai W, Zhang R
Effectiveness of Virtual Reality–Based Early Rehabilitation Strategies on Pain, Sleep, Anxiety, Balance, Cognition, and Limb Motor Function in Adult Intensive Care Unit Patients: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e81865
URL: https://www.jmir.org/2026/1/e81865
doi:10.2196/81865

© Fei Wu, Yunting Wu, Yana Xing, Weixin Cai, Ran Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Artificial Intelligence for Medicines Information: Scoping Review of Clinical Applications and Digital Health Inequalities

Shahd Al-Arkee[1], PhD; Josephine Falade[2], MSc; Vibhu Paudyal[3], PhD

[1]Department of Life Sciences, University of Bath, Cleaverton Down, Bath, United Kingdom
[2]Research Department of Practice and Policy, UCL School of Pharmacy, University College London, London, United Kingdom
[3]Department of Health Services and Population Research, King's College London, London, United Kingdom

Corresponding Author:
Shahd Al-Arkee, PhD
Department of Life Sciences, University of Bath, Cleaverton Down, Bath, United Kingdom

Abstract

Background: Artificial intelligence (AI) has the potential to support medicines information services. However, a comprehensive mapping of its use, particularly within pharmacy practice and in the context of digital health inequalities, is lacking.

Objective: This scoping review mapped existing evidence on AI-driven medicines information, focusing on the accuracy and completeness of AI-generated content, the role of health care professionals (HCPs), particularly pharmacists, and the impact of digital health inequalities on AI adoption.

Methods: This scoping review was informed by the methodological framework proposed by Levac et al, which includes modifications to the original Arksey and O'Malley scoping review framework. A systematic search was conducted across MEDLINE (Ovid), PubMed Central, Cochrane Library, CINAHL Plus (EBSCOhost), International Pharmaceutical Abstracts (IPA), Web of Science, and Google Scholar from inception to January 2025, which served as the search cutoff date. Peer-reviewed studies in English evaluating the role of AI in medicines information across any health care settings (including patient homes) were included. The results are reported in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines.

Results: A total of 1911 citations were identified, with 14 studies meeting the inclusion criteria. AI tools showed promise in supporting medicines information services but were found to have limitations in accuracy, particularly when applied to complex clinical queries. Pharmacists were the most engaged HCPs in the evaluation of AI-generated content. Only 3 studies explored digital health inequalities in the context of AI and access to medicines information. Reported barriers included misinformation risks, regulatory gaps, and digital health inequalities, particularly infrastructure limitations and disparities in digital literacy, which affected AI adoption.

Conclusions: AI-driven tools show promise in supporting medicines information services, but concerns remain. HCPs, particularly pharmacists, play a critical role in AI evaluation and validation, yet their involvement remains ill-defined. Addressing digital health inequalities is essential for effective AI integration. Future research should focus on identifying and minimizing digital health inequalities, as well as evidence-informed AI implementation in medicines information services.

Trial Registration: OSF Registries 2byvn; https://osf.io/2byvn/overview

(J Med Internet Res 2026;28:e77747) doi:10.2196/77747

KEYWORDS

artificial intelligence; medicines information; pharmacy; digital health inequalities; scoping review

Introduction

Artificial intelligence (AI), often described as "computational intelligence" or the "science and engineering of creating intelligent machines" [1], is a rapidly evolving field focused on replicating human-like behavior in computers and related technologies [2]. A key subset of AI, machine learning (ML), enables computers to analyze vast datasets and improve predictive accuracy without explicit programming [3]. AI applications in health care range from rule-based decision systems to advanced ML tools. By identifying patterns and learning from experience, ML algorithms can assist in predicting patient outcomes and supporting clinical decision-making across various health care settings [4].

AI has recently demonstrated increasing potential in pharmacy practice, particularly in medication management. A recent systematic review analyzing multiple AI algorithms reported that technology-driven approaches can enhance medication management in primary care [5]. The growing integration of AI into health care has led to the development of various AI

XSL•FO
RenderX

interventions, including algorithms designed to support both patients and health care professionals (HCPs) in disease management [6]. Several reviews have evaluated AI beyond medicines information, encompassing interventions in disease management [7], diagnosis [8], and shared decision-making [9]. Other reviews highlighted AI's role in pharmacy practice, demonstrating its effectiveness in optimizing medication management, enhancing drug safety, and improving clinical decision support systems [10]. However, the application of AI specifically to medicines information remains underexplored. Medicines information is recognized as a key component of promoting rational use of medicines, ensuring prescribers, dispensers, and consumers have access to independent and unbiased information about medication use [11].

Despite the growing interest in AI applications, there remains a need to comprehensively map how AI is used in medicines information services. Existing research lacks a clear synthesis of applications of AI, the accuracy and completeness of AI-generated responses, and the challenges influencing its adoption. Pharmacists and other HCPs play a crucial role in evaluating and integrating AI-driven tools, yet the extent of their engagement and its implications for pharmacy practice remain unclear. Furthermore, digital health inequalities refer to disparities in access to and use of digital health technologies, which are often shaped by factors such as internet connectivity and geographic location. These may influence the equitable implementation of AI across diverse health care settings.

This scoping review aims to map the current landscape of AI in medicines information, focusing on its applications, content accuracy and completeness, and adoption within pharmacy practice.

Methods

Information Sources and Search Strategy

This scoping review was conducted following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [12], available in Checklist 1. The methodological framework proposed by Levac et al [13], which expands on the original approach by Arksey and O'Malley [14], was used to guide the review. The protocol was developed in accordance with the following steps that are (1) identifying the research questions; (2) identifying relevant studies; (3) selecting studies through a team-based approach for study selection and data extraction; (4) charting the data using a structured approach; and (5) collating, summarizing, and reporting the results. This review was guided by four research questions that are (1) What impact do AI tools have on the dissemination and accessibility of medicines information in different health care settings, particularly in pharmacy practice? (2) What is the reported accuracy and completeness of AI-generated medicines information, and what factors influence its reliability? (3) What are the barriers and facilitators to AI adoption, particularly in addressing digital health inequalities? (4) How do HCPs, particularly pharmacists, perceive and engage with AI tools for medicines information? This protocol was registered and available on the Open Science Framework website [15].

A systematic literature search was conducted across MEDLINE (Ovid), PubMed Central, Cochrane Library, CINAHL Plus (EBSCOhost), International Pharmaceutical Abstracts (IPA), Web of Science, and Google Scholar from inception to January 2025. The database searches were conducted between January 10 and 20, 2025, with January 31, 2025, set as the inclusion cutoff date. A 4-domain search strategy was used, including terms related to AI, medicines information, digital health inequalities, and the role of the pharmacy workforce, particularly pharmacists, in AI adoption. The search strategy was developed and refined by 2 reviewers in consultation with an expert from the University College London Pharmacy School library. The final search strategy is presented in Multimedia Appendix 1 and was formatted in line with PRESS (Peer Review of Electronic Search Strategies) guidelines [16].

Study Selection and Data Extraction

Studies were included if they were qualitative, quantitative, or mixed methods studies, published in English in peer-reviewed journals, and explored the role of AI in enhancing medicines information. Medicines information was defined as a key component of promoting rational use of medicines, ensuring prescribers, dispensers, and consumers have access to independent and unbiased information about medication use [11] and guided the eligibility assessment. Eligible studies involved AI-enabled tools, including rule-based expert systems and ML models. Tools lacking any AI or ML functionality, such as conventional digital databases, were excluded. Reviews, editorials, commentaries, letters, media data, gray literature (eg, reports, theses, and white papers), and conference abstracts were also excluded to ensure methodological rigor and consistency.

A 2-stage selection process was used. In the first stage, one reviewer (SA) screened titles and abstracts for relevance and removed duplicate records, and the second reviewer (VP) thoroughly assessed the titles and abstracts for inclusion. In the second stage, 2 reviewers independently evaluated the full-text studies against the prespecified eligibility criteria. Disagreements were resolved through discussion and consensus. Bibliographies of selected studies were manually searched to identify additional relevant references.

Data extraction was performed using a standardized form developed specifically for this review. Extracted data included study characteristics, AI intervention details, and key findings related to AI-generated medicines information.

Data Synthesis and Quality Assessment

Data were synthesized using a descriptive approach, following a narrative synthesis method. The evidence was mapped to the predefined categories of HCP involvement, accuracy and completeness of AI-generated content, and implications for digital health inequalities. Study characteristics were also summarized, and gaps in the literature were highlighted. A formal quality assessment of the included studies was not conducted, in line with Arksey and O'Malley methodological framework for scoping reviews [14]. However, variations in study quality and methodological limitations were described narratively to support transparency and interpretation.

XSL•FO
RenderX

Results

Search Results and Study Selection

The systematic search yielded 1911 citations, of which 447 duplicates were removed, resulting in 1464 records screened by title and abstract. Of these, 20 studies were assessed for full-text review. Six studies were excluded, and no additional records were identified through manual searching. Therefore, 14 studies met the inclusion criteria and were included in this review [17-30]. A PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart summarizing the study selection process is shown in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram depicting study selection. AI: artificial intelligence.

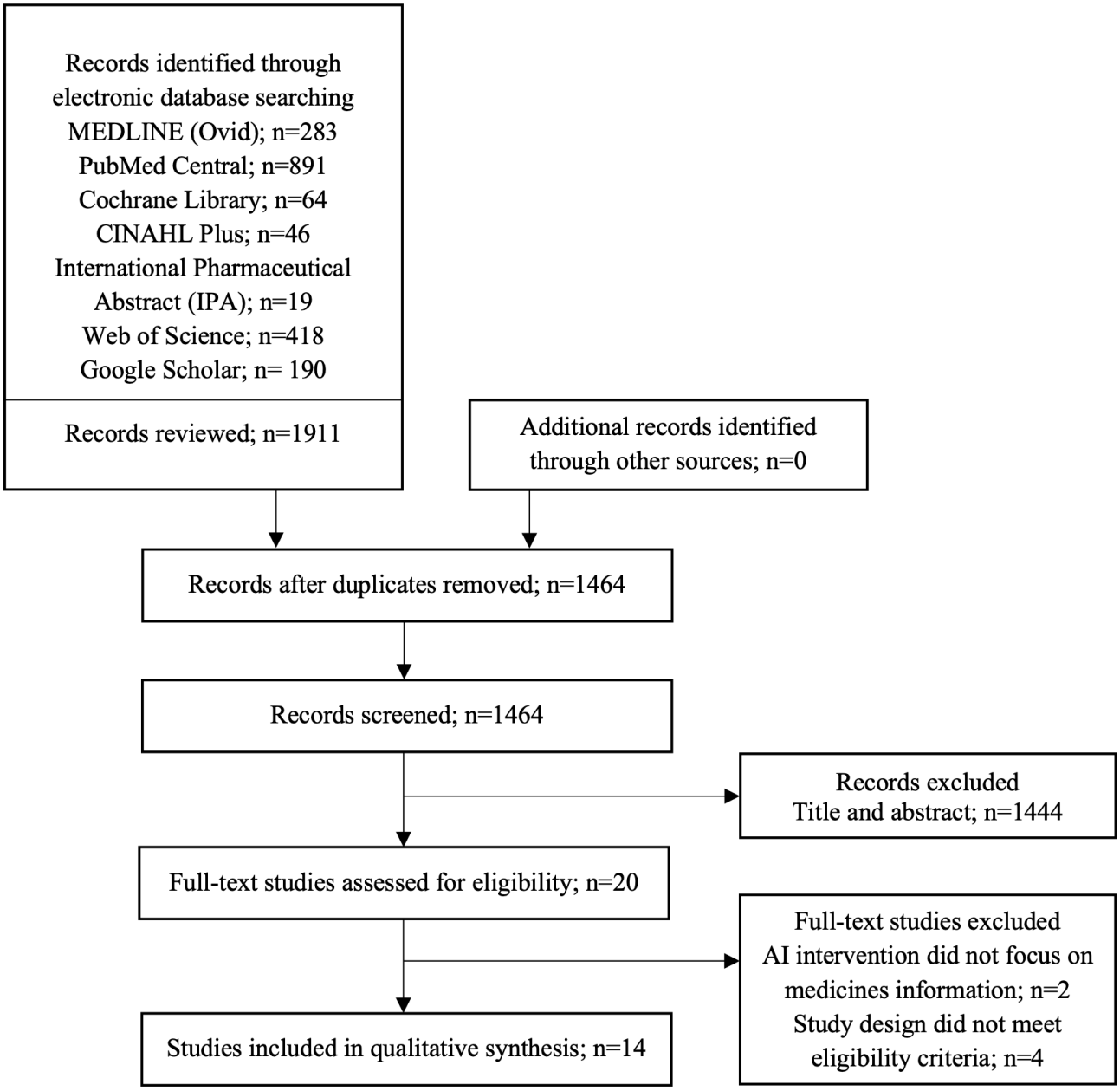

Study Characteristics and Design

The included studies were published between 2021 and 2025. Most originated from high-income economies, as classified by the World Bank, including Australia [26], Canada [21], Germany [17], the Netherlands [30], Singapore [27], and the United States [18-20,22,25,29]. In contrast, only a few studies originated from middle- and low-income countries, such as Egypt [28], Ethiopia [24], and India [23]. All studies focused on the application of AI for delivering medicines information (Table 1).

XSL•FO
RenderX

Table . Characteristics of the included studies.

Study; country	Health care setting	Data sample size	Research purpose	Reported outcomes
Andrikyan et al [17]; Germany	Online patient drug information search	500 chatbot responses (readability, completeness, and accuracy), 20 responses (safety)	Assess chatbot-generated drug advice for quality, readability, and safety concerns	• Readability • Completeness • Accuracy • Safety
Beavers et al [18]; United States	Inpatient hospital	200 real-world medication-use questions	Analyze the clinical completeness, correctness, usefulness, and safety of chatbot and medication database responses to everyday inpatient medication-use questions	• Clinical correctness • Completeness • Safety
Cornelison et al [19]; United States	AI[a]-driven medication counseling	240 chatbot responses	Evaluate chatbot accuracy and completeness in answering medication-related patient questions	• Accuracy • Completeness
Grossman et al [20]; United States	Academic drug information service	39 medication-related questions	Evaluate ChatGPT's accuracy, completeness, and relevance in drug inquiries	• Accuracy • Completeness • Satisfactory • Performance by question complexity
Laymouna et al [21]; Canada	Pharmacists in HIV care (community and hospital)	41 pharmacists from 15 municipalities	Assess pharmacists' knowledge, attitudes, and practices in HIV care and evaluate the usability of an AI-based chatbot for medicines information support	• Pharmacists' knowledge, attitudes, practices, and perceived barriers • Usability of the chatbot
Munir et al [22]; United States	Clinical pharmacy practice	32 pharmacy-based clinical questions	Assess ChatGPT's accuracy in answering pharmacy-related clinical queries	• Accuracy • Case-based questions
Ramasubramanian et al [23]; India	Not stated	462 medication dosage-related queries	Evaluate the accuracy of ChatGPT 3.5, ChatGPT 4, and Google Bard in providing medical drug dosages based on Harrison's Principles of Internal Medicine	• Accuracy
Sendekie et al [24]; Ethiopia	Community pharmacies	225 pharmacists	Investigate pharmacists' perceptions and willingness to use AI in practices, and their perceived barriers to its implementation	• Perception of AI and willingness to use AI • Barriers to AI adoption
Sheikh et al [25]; United States	Nephrology and pharmacy practice	124 nonprescription drugs and supplements assessed	Evaluate ChatGPT in drug safety for patients with kidney disease	• Accuracy • Safety
Stanceski et al [26]; Australia	Hospital discharge summaries (MIMIC-IV database)	100 chatbot responses (summaries)	Evaluate responses for safety, accuracy, and language simplification for AI-generated patient-centered discharge instructions	• Accuracy • Language simplification
Sumner et al [27]; Singapore	Outpatient clinics (hospital)	20 (10 health care providers, 10 patients)	Develop and evaluate an AI-driven nudge intervention for medication adherence	• Acceptance • Preference and concerns
Taha et al [28]; Egypt	Pharmacy practice (community and hospital pharmacies)	428 pharmacists	Explore pharmacists' perceptions, practices, and concerns regarding ChatGPT in pharmacy practice	• Perceptions on accuracy • Privacy concerns

XSL•FO
RenderX

Study; country	Health care setting	Data sample size	Research purpose	Reported outcomes
Triplett et al [29]; United States	Academic-based drug information center	84 ChatGPT responses analyzed	Assess the accuracy, completeness, and consistency of ChatGPT responses to drug information inquiries compared to drug information center responses	• Accuracy • Readability
van Nuland et al [30]; The Netherlands	Hospital-based clinical pharmacy practice	30 clinical pharmacy questions	Evaluate ChatGPT's ability to provide appropriate responses to clinical pharmacy questions and assess accuracy and consistency	• Accuracy • Completeness

[a]AI: artificial intelligence.

The included studies used a range of study designs. Eight studies used comparative and evaluation methodologies to assess AI-generated medicines information accuracy and usability [18-20,22,23,25,26,29]. Two studies used cross-sectional survey designs to explore user perceptions, knowledge, and attitudes toward AI-based tools [17,28]. Two studies adopted a human-centered design approach to assess AI-driven interventions: one evaluating AI-generated patient discharge instructions [26], and another developing an AI-driven nudge tool for medication adherence [27]. While most studies used quantitative research approaches**,** 2 studies used survey-based needs assessment designs to explore contextual factors and AI implementation considerations [21,24]. The studies were conducted across diverse health care settings, including community-based and primary care [21,24,28], secondary care (outpatients) [22,26,27], tertiary care settings [18,25,30], and web-based patient communities (online platforms without direct clinical engagement) [17,19,20,29]. One study did not specify a health care setting [23].

AI Intervention Characteristics

The AI tools evaluated across the included studies showed diverse functionalities in delivering medicines information (Table 2). Most AI systems used natural language processing (NLP) to generate medication-related responses [17-20,22,23]. Microsoft Bing Copilot analyzed user queries and provided structured drug information, while Micromedex with Watson generated text-based responses but struggled with complex clinical scenarios [17,18]. ChatGPT was widely assessed using various versions, with studies evaluating its accuracy in answering patient and academic drug inquiries, pharmacy-based clinical questions, and drug dosage recommendations [19,20,22,23]; among these, ChatGPT-4 (OpenAI) outperformed other models in a comparative evaluation [23].

Other included studies focused more on AI applications, adoption barriers, or decision-making, rather than primarily evaluating NLP-generated responses [21,24-30]. The MARVIN-Pharma chatbot supported pharmacists in HIV care, enhancing access to adherence counseling and clinical guidelines [21]. An evaluation of AI adoption in community pharmacies revealed that while AI could streamline medication management, infrastructure, and training barriers limited its potential [24]. Another study explored the utility of AI in nephrology and pharmacy practice, finding that ChatGPT sometimes misclassified supplement safety, thereby indicating the need for human oversight [25]. In hospital settings, AI-generated discharge summaries improved readability but also raised safety concerns [26]. AI-driven nudges enhanced medication adherence through personalized reminders for outpatients [27]. While AI tools improved efficiency and medication counseling, concerns persisted regarding misinformation, privacy, and patient-specific recommendations [28]. AI-generated drug information improved clarity but required pharmacist oversight, and hospital pharmacy AI responses often lacked reliability, raising concerns about clinical decision-making [29,30].

XSL•FO
RenderX

Table . Characteristics of the AI[a] interventions.

Study; country	AI name and functionality	AI potential risk	Research conclusion
Andrikyan et al [17]; Germany	Microsoft Bing Copilot: for patient drug-related queries	Incomplete or inaccurate drug information, and readability issues	• Capable of providing complete and accurate information • Not reliable for medical advice without validation
Beavers et al [18]; United States	Micromedex with Watson (IBM Watson Health): AI chatbot integrated with an evidence-based pharmacological knowledge base (MDX), using natural language processing to provide accurate, complete, and safe responses to medication-use questions	Limited scope, incomplete answers, and potential safety concerns	• Traditional drug databases outperformed AI chatbots in completeness, safety, and usefulness • Chatbot responses were clinically correct but not always complete or safe
Cornelison et al [19]; United States	ChatGPT-3.5: general AI chatbot for patient medication-related questions	Incomplete or inaccurate drug information	• Mostly accurate for answering common medication-related questions • May provide incomplete or inaccurate information
Grossman et al [20]; United States	ChatGPT 3.5: general AI chatbot for pharmaceutical medication-related questions	Incomplete or inaccurate drug information and fabricated references	• Unreliable for medication-related inquiries due to frequent inaccuracies, incomplete responses, and misleading references
Laymouna et al [21];Canada	MARVIN-Pharma: AI chatbot designed to support pharmacists in delivering evidence-based medicines information for HIV care	Usability concerns, with pharmacist engagement required	• AI could support HIV care • Needs pharmacist engagement in the chatbot's development
Munir et al [22]; United States	ChatGPT 3.5: generative AI language model designed to answer clinical questions in pharmacy practice	Misleading information, poor performance in complex cases, and safety concerns	• Limited success • Excels in calculations • Struggles with drug information and patient-specific cases
Ramasubramanian et al [23]; India	ChatGPT 3.5, ChatGPT 4, Google Bard: AI models assessed for drug dosage accuracy, based on Harrison's Principles of Internal Medicine	Accuracy varied across diseases and organ systems	• ChatGPT 4 exhibited the highest accuracy • Requires further validation before clinical application
Sendekie et al [24]; Ethiopia	Not specified: generic AI applications in pharmacy: for personalized patient care and decision support	Needs for resources, policies, training, and infrastructure	• AI is recognized as beneficial • Needs training and policy development for successful integration
Sheikh et al [25]; United States	ChatGPT 3.5 and ChatGPT 4: evaluated for assessing the safety of nonprescription medications and supplements in kidney disease, compared to Micromedex	Inaccuracy, lack of reliability, and patient safety concerns	• ChatGPT 4 has potential, but remains unreliable compared to established databases • Further AI development is needed before clinical adoption
Stanceski et al [26]; Australia	ChatGPT 3.5 (via Microsoft Azure OpenAI): AI-generated patient-friendly discharge instructions from hospital summaries	Potential medication errors	• AI tools can generate patient-centered discharge instructions • Require clinician oversight before implementation
Sumner et al [27]; Singapore	AI-driven nudge system for medication adherence, integrating reminders, social references, and automated interventions	Digital literacy barriers for older adults, and data privacy concerns	• Promising for improving medication adherence but must be tailored to user needs • Needs flexible delivery, simplified data entry, and caregiver involvement

XSL•FO
RenderX

Study; country	AI name and functionality	AI potential risk	Research conclusion
Taha et al [28]; Egypt	ChatGPT: AI-driven drug information tool assisting pharmacists with counseling, drug interactions, and clinical queries	Concerns on accuracy, data privacy, and bias	• Potential applications in pharmacy but requires accuracy improvements and regulatory guidelines
Triplett et al [29]; United States	ChatGPT 3.5: AI-generated responses for pharmaceutical inquiries, based on NLP[b] and public databases	Accuracy concerns, lack of references, requiring pharmacist validation	• ChatGPT provided clear and readable responses • Lacked accuracy, requiring expert review for reliability
van Nuland et al [30]; The Netherlands	ChatGPT 3.5: AI-generated responses for hospital pharmacy inquiries, including dosing, drug interactions, and therapeutic drug monitoring	Lack of response accuracy, inconsistent AI-generated information, risk of medication errors	• ChatGPT offered clinical advice • Inconsistent accuracy and reproducibility, making it unsuitable for independent clinical use

[a]AI: artificial intelligence.

[b]NLP: natural language processing.

Accuracy and Completeness

Accuracy was primarily evaluated based on the correctness of AI-generated responses in providing medication-related information, but only 5 studies [22,23,25,26,29] conducted a direct assessment. A study evaluated ChatGPT-3.5, ChatGPT-4, and Google Bard for drug dosage recommendations. ChatGPT-4 demonstrated the highest performance (83.77%), but accuracy varied across diseases and organ systems [23]. Another study assessing AI's classification of drug safety in patients with kidney disease found that ChatGPT-4 had higher agreement (81.4%) with Micromedex than ChatGPT-3.5 (64.5%), yet it remained unreliable for independent clinical use [25]. Similarly, a study analyzing AI-generated discharge summaries found that 18% of responses contained potentially harmful safety issues, including 6% with "hallucinated information," meaning responses that appeared plausible but were actually inaccurate and 3% introducing unprescribed medications [26]. Another study compared ChatGPT's accuracy in drug information inquiries against a drug information center, reporting an accuracy rate of 50%. ChatGPT lacked proper references and reliability for clinical use [29]. Another study identified significant limitations in ChatGPT's accuracy, particularly in patient-specific scenarios, emphasizing the need for verification before integration into pharmacy practice [22].

Five additional studies evaluated both accuracy and completeness, evaluating how well AI responses covered all essential information. Microsoft Bing Copilot achieved 100% median accuracy and completeness across 500 responses, yet experts reviewing a subset of 20 chatbot responses found that 66% were potentially harmful, with 22% classified as life-threatening [17]. Micromedex with Watson produced 85% clinically correct responses, but only 65% were complete and 71% acceptable for safety [18]. AI-generated responses to hospital pharmacy inquiries were only 26% correct and complete, while 22% were correct but incomplete, 30% partially correct, and 22% completely incorrect [30]. One study found ChatGPT-generated responses were 92.5% accurate and 80.8% complete, though 4.2% contained minor errors [19]. Another study reported that only 26% of ChatGPT's responses were satisfactory, while 38% lacked a direct response, 38% contained inaccuracies, and 41% lacked completeness [20]. A detailed grouped summary of all included studies by AI task and evaluation metric is presented in Table 3.

XSL•FO
RenderX

Table . Grouped summary by AI[a] task and metric.

AI task	Study; country	Evaluation metric	Outcome
Drug information	Andrikyan et al [17]; Germany	• Accuracy • Safety	• 100% median accuracy • 66% potentially harmful (subset safety analysis)
	Beavers et al [18]; United States	• Accuracy • Completeness • Safety	• 85% accuracy • 65% completeness • 71% safety
	Cornelison et al [19]; United States	• Accuracy • Completeness	• 92.5% accuracy • 80.8% completeness
	Grossman et al [20]; United States	• Accuracy • Completeness	• 26% satisfactory • Among unsatisfactory responses, 38% lacked accuracy and 41% lacked completeness
	Taha et al [28]; Egypt	• Perception • Concerns	• Perceived benefits (73.6%) • Concerns about privacy, bias, and accuracy
	Triplett et al [29]; United States	• Accuracy • Reference • Quality	• 50% accuracy • Poor reference quality
	van Nuland et al [30]; The Netherlands	• Accuracy • Completeness	• 26% correct and complete • 22% correct but incomplete
Dosage recommendations	Ramasubramanian et al [23]; India	• Accuracy	• 83.77% accuracy • Performance varied by disease
HIV care	Laymouna et al [21]; Canada	• Perception • Usability	• Perceived potential to support HIV care, pharmacist engagement essential
Clinical queries	Munir et al [22]; United States	• Expert judgment	• Good in calculation (100%) • Poor in complex cases (20%) • Safety concerns
System-level use	Sendekie et al [24]; Ethiopia	• Infrastructure readiness • Perception	• Positive perception • Policy and training needed for integration
Safety assessment	Sheikh et al [25]; United States	• Agreement • Safety	• 81.4% agreement (ChatGPT-4) with Micromedex • Not yet reliable for clinical use
Discharge instructions	Stanceski et al [26]; Australia	• Safety	• 18% potentially harmful • Incorrect medication content
Adherence support	Sumner et al [27]; Singapore	• Perception	• Acceptable and promising • User-centered design emphasized

[a]AI: artificial intelligence.

Digital Health Inequalities

Aspects of digital health inequalities investigated within the included studies related to AI-driven interventions to enhance health care access, support medication adherence, and provide drug information. Key barriers identified were limited infrastructure and digital literacy (Table 4).

XSL•FO
RenderX

Table . Summary of AI[a] adoption barriers by country context.

Study; country	Research setting (geography)	Barriers category	Key barriers identified	Patient-level inequities
Sendekie et al [24]; Ethiopia	LMIC[b]	Structural or institutional	• Lack of internet • Lack of AI tools • Limited training	• Infrastructure • Digital literacy
Sumner et al [27]; Singapore	HIC[c]	Personal	• Digital literacy among older adults • Privacy concerns	• Age • Data privacy concerns
Taha et al [28]; Egypt	LMIC	Personal or demographic	• Regional gaps in awareness and training • Limited accuracy of AI • Data privacy	• Geographic disparities • Low income and education levels

[a]AI: artificial intelligence.

[b]LMIC: low- and middle-income country.

[c]HIC: high-income country.

AI-driven medication adherence tools showed promise, with one study demonstrating improved adherence through personalized reminders and automated interventions, but digital literacy barriers among older adults and privacy concerns limited their effectiveness [27]. Infrastructure and accessibility challenges were key barriers to AI adoption. A study in Ethiopia found that 89.3% of surveyed pharmacists cited lack of internet availability, 88.2% lacked AI-related software and hardware, and 80.9% reported insufficient training. More than 90% emphasized the need for structured policies, better internet access, and AI-focused training to support AI integration in pharmacy services [24].

Regional disparities in AI awareness and adoption were evident. A study evaluating Egyptian pharmacists' perceptions of ChatGPT found that those in Greater Cairo showed the highest level of AI awareness, whereas those in South Upper Egypt had the lowest, reflecting geographic variations in AI exposure. Limited AI training further hindered equitable adoption and impacted AI-supported decision-making. Concerns about AI accuracy, data privacy, and bias reinforced skepticism toward AI as a standalone clinical tool, emphasizing the need for expert validation in pharmacy practice [28].

Involvement of HCPs

The involvement of HCPs in AI interventions varied across the included studies, with pharmacists being the primary group engaged. This aligns with the focus of our scoping review on their role in AI-driven medicines information. Five studies specifically involved pharmacists, either as primary users of AI [20,29] or as evaluators of AI-generated responses [22,25,30]. Three studies explored pharmacists' perceptions of AI, assessing their willingness to integrate AI into practice, perceived barriers to adoption, and concerns regarding accuracy, data privacy, and trust in AI-based recommendations [21,24,28]. These studies provided valuable insights into pharmacists' readiness to adopt AI-driven health care solutions while highlighting challenges related to infrastructure, regulatory frameworks, and professional training.

Four studies included multidisciplinary HCPs, such as physicians and nurses, in evaluating and validating AI interventions [17,18,26,27]. In contrast, 2 studies reported no direct involvement of HCPs in AI use or evaluation, as AI-generated responses were independently assessed by reviewers whose professional backgrounds were not specified [19,23].

Narrative Appraisal of Included Studies

Several studies relied on small or narrowly defined datasets [22,25], hypothetical scenarios without clinical validation [17,26], or educational and nonclinical settings [19,20,24,29], which limited their real-world applicability. Evaluations were frequently conducted internally, without independent or blinded assessors [21,30], and some lacked comparisons to human experts or assessments of clinical consequences [22,23]. In addition, poor reproducibility [30], limited topic scope [25], single-center study design [18], and nonsystematic prompt design were apparent [27,28].

Discussion

Principal Findings

This scoping review provided an overview of AI-driven tools in medicines information, focusing on HCP involvement, AI accuracy and completeness, and digital health inequalities. Some models accurately address general medication inquiries but struggle with complex clinical questions, raising concerns regarding reliability, data privacy, and usability. With respect to HCP involvement, studies highlighted the multidisciplinary engagement, particularly in medication adherence and discharge planning. Pharmacists played a key role in AI evaluation and integration. Regarding equity, the review highlighted disparities in infrastructure, digital literacy, and access to technology, which may hinder equitable implementation. One study examined digital health inequalities from the patient perspective.

Although a formal quality appraisal was not conducted, several methodological limitations were observed across the included

XSL•FO
RenderX

studies, which should be considered when interpreting the overall strength and applicability of the evidence.

Comparison With Prior Work

The included studies presented AI interventions used for medicines information, particularly in their potential applications in medication inquiries, adherence monitoring, and discharge planning. These findings align with previous reviews, which have explored the capacity of AI to support medication adherence through predictive models and real-time monitoring. AI-driven tools, including ML-driven monitoring systems, have been investigated for their ability to assess adherence patterns and identify at-risk patients; however, their effectiveness varies based on implementation and patient engagement [31]. Similarly, ML and statistical models have been explored for their potential in optimizing discharge planning by predicting patient outcomes, which may help reduce readmission rates and improve hospital efficiency [32]. Despite these promising applications, challenges remain regarding the accuracy, data privacy, and usability of AI-generated outputs. A recent systematic review highlighted that while AI has been explored in various aspects of medication use, its application in supporting clinical decision-making and complex, patient-specific inquiries remains limited [33]. Although AI models often demonstrate high accuracy when handling standard medication-related questions, their performance tends to decline in more complex clinical contexts, reinforcing the importance of human oversight. This review builds on these findings by specifically examining how AI is being applied to support medicines information services and evaluating the quality and focus of these emerging interventions.

The successful integration of AI into medication management and the provision of medicines information requires the active involvement of HCPs, particularly pharmacists, physicians, and nurses. The included studies indicate that pharmacists are key stakeholders in the adoption, evaluation, validation, and clinical implementation of AI tools. Similarly, a systematic review highlighted the potential of AI to enhance decision-making, emphasizing the need for expert validation to ensure clinical relevance and safety [34]. Furthermore, multidisciplinary engagement is essential for successful AI adoption; involving HCPs in AI development and validation improves usability and fosters trust. A narrative review identified key factors influencing trust in medical AI—such as explainability, transparency, and usability—reinforcing the importance of clinician involvement in ensuring effective decision-making [35].

Challenges remain in AI implementation, particularly in relation to digital health inequalities. Some included studies highlighted regional disparities in AI adoption, with pharmacists working in urban areas showing greater familiarity with AI tools compared to their counterparts working in rural settings [24,28]. While these studies provide valuable insights into pharmacists' access to and perceptions of digital tools, only one study specifically addressed digital health barriers from the patient perspective [27]. This highlights a key research gap in understanding how AI-driven pharmacy interventions affect patients, particularly vulnerable populations. Furthermore, limited AI-related training and infrastructure barriers, particularly in low-resource settings, hinder AI adoption [24], whereas the study from a high-income country emphasized individual-level concerns [27]. These findings align with broader research highlighting technological gaps and digital literacy disparities as significant barriers to AI adoption in medicines information services [36].

A key challenge in AI-generated medicines information involves potential risks related to data privacy, regulatory gaps, and ethical concerns. Two included studies reported inaccuracies or inconsistencies in AI-generated responses, raising concerns about their reliability and implications for patient safety [23,25]. This aligns with findings from a previous narrative review on AI governance, which highlighted the need for clear regulatory frameworks to support the adoption of AI in medicines information, particularly in addressing issues of bias, transparency, and accountability [37].

Strengths and Limitations

To the best of the authors' knowledge, this is the first scoping review to comprehensively map the use of AI for medicines information and evaluate its accuracy. A key strength is its systematic approach, adherence to established scoping review methodology, incorporating both qualitative and quantitative study designs.

Certain limitations should be acknowledged. The exclusion of gray literature, while consistent with our protocol and justified by the need to prioritize methodological rigor, may have introduced publication bias. Given the fast-evolving nature of AI research, new AI applications may be reported initially in non–peer-reviewed sources. Future updates of this review may benefit from a broader inclusion strategy or a targeted sensitivity search of these sources. Other limitations relate to the heterogeneity of included studies. The limited number of evaluated AI tools, small sample size, and small number of countries and within specific health care settings restrict the ability to draw generalizable conclusions. Inconsistencies were observed in how included studies defined and measured key outcomes such as "accuracy" and "completeness." Some studies assessed accuracy based on comparison to expert responses, while others used subjective scoring or evaluated only surface-level correctness. Similarly, the concept of "completeness" varied, ranging from the presence of keywords to the inclusion of all critical information. This variability complicates direct comparisons across studies and limits the ability to synthesize consistent conclusions regarding AI performance. Moreover, some application areas remain underexplored. Importantly, no formal quality appraisal of the included studies was performed, which may affect the ability to assess the robustness of individual findings—although this is consistent with established scoping review methodology.

Implications for Pharmacy Practice

This review highlights pharmacists as the most actively engaged HCPs in the development, evaluation, and implementation of AI tools for medicines information. Their clinical expertise, accessibility, and experience in medication counseling position them as key stakeholders in ensuring the safe and effective use

XSL•FO
RenderX

of AI-driven systems. Pharmacists require targeted training to critically evaluate AI-generated information, understand algorithmic limitations, and interpret outputs in clinical context. Embedding digital health and AI literacy into pharmacy education and continuing professional development will be critical to building confidence and capability in using these tools. Particular attention should be paid to pharmacists working in low-resource settings who lack access to digital health technologies.

Workflow integration also requires attention. AI systems should be incorporated seamlessly into pharmacy practice to enhance efficiency without disrupting established routines. This may involve embedding AI within electronic health records, clinical decision support systems, and patient counseling processes. Co-designing tools with pharmacists can ensure alignment with routine tasks and improve adoption.

Policy and governance considerations are equally important. Pharmacists should contribute actively to shaping standards for validation, data privacy, and accountability, especially in patient-facing contexts. Interdisciplinary collaborations will be essential as medicines information becomes increasingly digitized. Pharmacists are well-positioned to lead interdisciplinary teams in the development, use, and appraisal of AI applications and the information they generate.

Implication for Research

This review provides an overview of AI's role in medicines information; however, several gaps remain that warrant further investigation. Future research should focus on improving the accuracy and reliability of AI in handling complex clinical inquiries, as current models demonstrate inconsistencies in delivering clinically relevant recommendations. Equally important is the evaluation of the real-world effectiveness of AI in supporting medication adherence and discharge planning through personalized interventions.

Research should also assess the impact of digital health inequalities on AI adoption. In this review, only a small subset of studies explicitly examined aspects such as disparities in access to AI tools, language inclusivity, or representation of vulnerable populations. This limited evidence base restricts our ability to draw conclusions about the equity implications of AI-generated medicines information. Given that AI systems can perpetuate or even exacerbate existing health care disparities, future research should assess how factors such as access to digital technologies, digital literacy, and systemic bias influence access to and trust in AI tools, particularly in underresourced settings. Particular attention needs to be paid to inequalities that intersect with digital literacy skills, including older age, socioeconomic disadvantage, and people living in remote and rural areas. Frameworks such as the Digital Health Equity Framework [38] or the PROGRESS-Plus (Place of residence, Race/ethnicity/culture/language, Occupation, Gender/sex, Religion, Education, Socioeconomic status, and Social capital) lens [39] can provide conceptual guidance for equity-centered investigations. Future research should also explore ways to facilitate accessibility, affordability, and digital literacy skills from the perspectives of HCPs who practice in low-resource settings.

While this review identified pharmacists as key stakeholders in AI evaluation, further studies are needed to assess how AI influences clinical decision-making, workflow integration, and patient communication across various health care settings. Equally important is the integration of patient perspectives in the design and evaluation of AI tools. Despite their critical role as end users, patients were largely underrepresented in the included studies, indicating a need for more patient-centered research to ensure AI solutions are usable, trustworthy, and aligned with patient needs.

Conclusions

This scoping review highlights the early promise of AI in supporting medicines information, but its reliability, especially in complex clinical scenarios, remains uncertain. At present, AI should be viewed as a complementary tool rather than a replacement for human expertise, with HCPs playing a critical role in its evaluation and integration. Key barriers such as digital health inequalities, ethical concerns, and regulatory gaps must be addressed to enable the safe and equitable implementation. Future research should focus on identifying and mitigating digital health inequalities in this context, including intersectional disadvantages as well as improving evidence-based use of AI, expanding training for HCPs, and developing robust governance frameworks to support standardized AI integration into clinical practice.

Acknowledgments

The authors gratefully acknowledge the library and database resources provided by University College London, which supported the literature search and development of this review.

Funding

No financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Authors' Contributions

Conceptualization: SA (lead), VP (equal)
Methodology: SA (lead), VP (equal), JF (supporting)
Formal analysis: SA (lead), VP (equal)

XSL•FO
RenderX

Writing – original draft: SA (lead), VP (equal)
Writing – review & editing: SA (lead), VP (equal), JF (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy (PRESS) Format.
[PDF File, 86 KB - jmir_v28i1e77747_app1.pdf]

Checklist 1
PRISMA-ScR guidelines.
[PDF File, 61 KB - jmir_v28i1e77747_app2.pdf]

References

1. McCarthy J. What Is Artificial Intelligence: MIT Press; 2007. URL: https://www-formal.stanford.edu/jmc/whatisai.pdf
2. Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. J Family Med Prim Care 2019 Jul;8(7):2328-2331. [doi: 10.4103/jfmpc.jfmpc_440_19] [Medline: 31463251]
3. What is machine learning? IBM. URL: https://www.ibm.com/think/topics/machine-learning [accessed 2025-04-04]
4. Abbasgholizadeh Rahimi S, Légaré F, Sharma G, et al. Application of artificial intelligence in community-based primary health care: systematic scoping review and critical appraisal. J Med Internet Res 2021 Sep 3;23(9):e29839. [doi: 10.2196/29839] [Medline: 34477556]
5. Damiani G, Altamura G, Zedda M, et al. Potentiality of algorithms and artificial intelligence adoption to improve medication management in primary care: a systematic review. BMJ Open 2023 Mar 23;13(3):e065301. [doi: 10.1136/bmjopen-2022-065301] [Medline: 36958780]
6. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
7. Singareddy S, Sn VP, Jaramillo AP, et al. Artificial intelligence and its role in the management of chronic medical conditions: a systematic review. Cureus 2023 Sep;15(9):e46066. [doi: 10.7759/cureus.46066] [Medline: 37900468]
8. Kumar Y, Koul A, Singla R, Ijaz MF. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. J Ambient Intell Humaniz Comput 2023;14(7):8459-8486. [doi: 10.1007/s12652-021-03612-z] [Medline: 35039756]
9. Abbasgholizadeh Rahimi S, Cwintal M, Huang Y, et al. Application of artificial intelligence in shared decision making: scoping review. JMIR Med Inform 2022 Aug 9;10(8):e36199. [doi: 10.2196/36199] [Medline: 35943793]
10. Choudhury A, Asan O. Role of artificial intelligence in patient safety outcomes: systematic literature review. JMIR Med Inform 2020 Jul 24;8(7):e18599. [doi: 10.2196/18599] [Medline: 32706688]
11. The pursuit of responsible use of medicines: sharing and learning from country experiences. World Health Organization. 2012. URL: https://www.who.int/publications/i/item/WHO-EMP-MAR-2012.3 [accessed 2025-04-17]
12. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
13. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5(1):69. [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
14. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
15. Al-Arkee S, Falade J, Paudyal V. Artificial intelligence for medicines information: scoping review protocol. URL: https://osf.io/p3tgs [accessed 2025-04-06]
16. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS peer review of electronic search strategies: 2015 Guideline Statement. J Clin Epidemiol 2016 Jul;75:40-46. [doi: 10.1016/j.jclinepi.2016.01.021] [Medline: 27005575]
17. Andrikyan W, Sametinger SM, Kosfeld F, et al. Artificial intelligence-powered chatbots in search engines: a cross-sectional study on the quality and risks of drug information for patients. BMJ Qual Saf 2025 Jan 28;34(2):100-109. [doi: 10.1136/bmjqs-2024-017476] [Medline: 39353736]
18. Beavers J, Schell RF, VanCleave H, et al. Evaluation of inpatient medication guidance from an artificial intelligence chatbot. Am J Health Syst Pharm 2023 Dec 5;80(24):1822-1829. [doi: 10.1093/ajhp/zxad193] [Medline: 37611187]
19. Cornelison BR, Erstad BL, Edwards C. Accuracy of a chatbot in answering questions that patients should ask before taking a new medication. J Am Pharm Assoc (2003) 2024;64(4):102110. [doi: 10.1016/j.japh.2024.102110] [Medline: 38670493]
20. Grossman S, Zerilli T, Nathan JP. Appropriateness of ChatGPT as a resource for medication-related questions. Br J Clin Pharmacol 2024 Oct;90(10):2691-2695. [doi: 10.1111/bcp.16212] [Medline: 39096130]

XSL•FO
RenderX

21. Laymouna M, Ma Y, Lessard D, et al. Needs-assessment for an artificial intelligence-based chatbot for pharmacists in HIV care: results from a knowledge-attitudes-practices survey. Healthcare (Basel) 2024 Aug 20;12(16):1661. [doi: 10.3390/healthcare12161661] [Medline: 39201222]
22. Munir F, Gehres A, Wai D, Song L. Evaluation of ChatGPT as a tool for answering clinical questions in pharmacy practice. J Pharm Pract 2024 Dec;37(6):1303-1310. [doi: 10.1177/08971900241256731] [Medline: 38775367]
23. Ramasubramanian S, Balaji S, Kannan T, et al. Comparative evaluation of artificial intelligence systems' accuracy in providing medical drug dosages: a methodological study. World J Methodol 2024 Dec 20;14(4):92802. [doi: 10.5662/wjm.v14.i4.92802] [Medline: 39712564]
24. Sendekie AK, Limenh LW, Abate BB, et al. Artificial intelligence in community pharmacy practice: pharmacists' perceptions, willingness to utilize, and barriers to implementation. Explor Res Clin Soc Pharm 2024 Dec;16:100542. [doi: 10.1016/j.rcsop.2024.100542] [Medline: 39687445]
25. Sheikh MS, Barreto EF, Miao J, et al. Evaluating ChatGPT's efficacy in assessing the safety of non-prescription medications and supplements in patients with kidney disease. Digit Health 2024;10:20552076241248082. [doi: 10.1177/20552076241248082] [Medline: 38638404]
26. Stanceski K, Zhong S, Zhang X, et al. The quality and safety of using generative AI to produce patient-centred discharge instructions. NPJ Digit Med 2024 Nov 20;7(1):329. [doi: 10.1038/s41746-024-01336-w] [Medline: 39567722]
27. Sumner J, Bundele A, Lim HW, Phan P, Motani M, Mukhopadhyay A. Developing an artificial intelligence-driven nudge intervention to improve medication adherence: a human-centred design approach. J Med Syst 2023 Dec 8;48(1):3. [doi: 10.1007/s10916-023-02024-0] [Medline: 38063940]
28. Taha T, Abdel-Qader DH, Alamiry KR, Fadl ZA, Alrawi A, Abdelsattar NK. Perception, concerns, and practice of ChatGPT among Egyptian pharmacists: a cross-sectional study in Egypt. BMC Health Serv Res 2024 Nov 28;24(1):1500. [doi: 10.1186/s12913-024-11815-1] [Medline: 39609697]
29. Triplett S, Ness-Engle GL, Behnen EM. A comparison of drug information question responses by a drug information center and by ChatGPT. Am J Health Syst Pharm 2025 Apr 17;82(8):448-460. [doi: 10.1093/ajhp/zxae316] [Medline: 39450858]
30. van Nuland M, Lobbezoo AFH, van de Garde EMW, et al. Assessing accuracy of ChatGPT in response to questions from day to day pharmaceutical care in hospitals. Explor Res Clin Soc Pharm 2024 Sep;15:100464. [doi: 10.1016/j.rcsop.2024.100464] [Medline: 39050145]
31. Bohlmann A, Mostafa J, Kumar M. Machine learning and medication adherence: scoping review. JMIRx Med 2021 Nov 24;2(4):e26993. [doi: 10.2196/26993] [Medline: 37725549]
32. Pahlevani M, Taghavi M, Vanberkel P. A systematic literature review of predicting patient discharges using statistical methods and machine learning. Health Care Manag Sci 2024 Sep;27(3):458-478. [doi: 10.1007/s10729-024-09682-7] [Medline: 39037567]
33. Khosravi M, Zare Z, Mojtabaeian SM, Izadi R. Artificial intelligence and decision-making in healthcare: a thematic analysis of a systematic review of reviews. Health Serv Res Manag Epidemiol 2024;11:23333928241234863. [doi: 10.1177/23333928241234863] [Medline: 38449840]
34. Ahmed MI, Spooner B, Isherwood J, Lane M, Orrock E, Dennison A. A systematic review of the barriers to the implementation of artificial intelligence in healthcare. Cureus 2023 Oct;15(10):e46454. [doi: 10.7759/cureus.46454] [Medline: 37927664]
35. Tucci V, Saary J, Doyle TE. Factors influencing trust in medical artificial intelligence for healthcare professionals: a narrative review. J Med Artif Intell 2021;5:4-4. [doi: 10.21037/jmai-21-25]
36. Zuhair V, Babar A, Ali R, et al. Exploring the impact of artificial intelligence on global health and enhancing healthcare in developing nations. J Prim Care Community Health 2024;15:21501319241245847. [doi: 10.1177/21501319241245847] [Medline: 38605668]
37. Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: a narrative review. Heliyon 2024 Feb 29;10(4):e26297. [doi: 10.1016/j.heliyon.2024.e26297] [Medline: 38384518]
38. Rowland S, Brewer LC, Rosas LG. Digital health equity - a call to action for clinical and translational scientists. J Clin Transl Sci 2024;8(1):e145. [doi: 10.1017/cts.2024.564] [Medline: 39478785]
39. O'Neill J, Tabish H, Welch V, et al. Applying an equity lens to interventions: using PROGRESS ensures consideration of socially stratifying factors to illuminate inequities in health. J Clin Epidemiol 2014 Jan;67(1):56-64. [doi: 10.1016/j.jclinepi.2013.08.005] [Medline: 24189091]

Abbreviations

AI: artificial intelligence
HCP: health care professional
IPA: International Pharmaceutical Abstracts
ML: machine learning
NLP: natural language processing
PRESS: Peer Review of Electronic Search Strategies

XSL•FO
RenderX

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
PROGRESS-Plus: Place of residence, Race/ethnicity/culture/language, Occupation, Gender/sex, Religion, Education, Socioeconomic status, and Social capital

Edited by A Coristine, TDA Cardoso; submitted 19.May.2025; peer-reviewed by N Acharya, P Sharma, Y Olanlokun; revised version received 27.Dec.2025; accepted 27.Dec.2025; published 06.Mar.2026.

Please cite as:
Al-Arkee S, Falade J, Paudyal V
Artificial Intelligence for Medicines Information: Scoping Review of Clinical Applications and Digital Health Inequalities
J Med Internet Res 2026;28:e77747
URL: https://www.jmir.org/2026/1/e77747
doi:10.2196/77747

© Shahd Al-Arkee, Josephine Falade, Vibhu Paudyal. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Machine Learning in Left Ventricular Hypertrophy Detection: Systematic Review and Meta-Analysis

Yilin Li[1], MD; Ke Zhao[2], MM; Jing Wu[1], BS

[1]Department of Geriatrics, The Third People's Hospital of Chengdu, 82 Qinglong Street, Qingyang District, Chengdu, Sichuan Province, China, 610031, Chengdu, Sichuan, China

[2]Department of Critical Care Medicine, Affiliated Hospital of North Sichuan Medical College, Nanchong, Sichuan, China

Corresponding Author:
Jing Wu, BS
Department of Geriatrics, The Third People's Hospital of Chengdu, 82 Qinglong Street, Qingyang District, Chengdu, Sichuan Province, China, 610031, Chengdu, Sichuan, China

Abstract

Background: In recent years, researchers have investigated machine learning (ML)–based approaches for the detection of left ventricular hypertrophy (LVH). However, the accuracy of ML in detecting LVH varies across different modeling variables and models. Systematic evidence is lacking in understanding how different ML approaches affect LVH detection accuracy.

Objective: The aim of this study is to systematically assess the diagnostic accuracy of these ML approaches to inform the development of artificial intelligence tools.

Methods: PubMed, Embase, Cochrane Library, and Web of Science were comprehensively searched up to November 12, 2025. The Prediction Model Risk of Bias Assessment Tool was used to evaluate the risk of bias. Subgroup analyses were performed based on ML model types and modeling variables (electrocardiogram [ECG], clinical features, and echocardiography). Only diagnostic 2×2 tables from validation sets were pooled for meta-analysis, with all statistical analyses performed using Stata.

Results: A total of 25 studies were included in the analysis. The performance of ML models varied with input data types and algorithms. A meta-analysis showed that ECG-based models, in comparison, exhibited a sensitivity of 0.76 (95% CI 0.66 - 0.84) and a specificity of 0.84 (95% CI 0.78 - 0.89). Echocardiography-based models had a sensitivity ranging from 0.71 to 0.94 and a specificity ranging from 0.67 to 0.96. The models based on clinical features had a sensitivity of 0.78 (95% CI 0.69 - 0.85) and a specificity of 0.71 (95% CI 0.65 - 0.76). A subgroup analysis of the ECG-based models revealed that the deep learning model produced a sensitivity of 0.71 (95% CI 0.60 - 0.80) and a specificity of 0.79 (95% CI 0.65 - 0.88).

Conclusions: ML demonstrates reasonably high accuracy in detecting LVH. However, these conclusions are derived from limited evidence. Meanwhile, the extreme heterogeneity reported in the meta-analysis requires more critical interpretation. Current conclusions regarding model accuracy should be interpreted with caution. Therefore, future research should focus on constructing high-performance ML models based on imaging data for LVH diagnosis.

(J Med Internet Res 2026;28:e76637) doi:10.2196/76637

KEYWORDS

artificial intelligence; AI; machine learning; left ventricular hypertrophy; meta-analysis; deep learning; electrocardiogram; ECG; echocardiography; cardiovascular risk

Introduction

Left ventricular hypertrophy (LVH) represents an increase in left ventricular mass driven by various cardiovascular risk factors [1], and it confers a 5 - 10 times greater risk of developing cardiovascular disease [2,3]. Accumulating evidence demonstrates that LVH is an independent predictor of cardiovascular risk [4-6]. However, the pathological and physiological alterations induced by LVH occur without overt symptoms or clinical signs [1]. Therefore, the early detection of LVH through sensitive screening methods is imperative, as it holds significant positive implications for patient health outcomes.

Currently, echocardiography (ECHO) and electrocardiogram (ECG) are the conventional diagnostic methods for detecting LVH [7,8]. While ECHO is a relatively accurate approach [9], it is costly and requires specialized equipment and trained operators, which poses challenges for early diagnosis [9,10]. Studies also show high rates of disagreement in ECHO readings among cardiologists, sometimes as much as 42.1% [11]. Cardiologists may mistakenly read ECHOs, particularly when tired, possibly resulting in misdiagnoses and misinterpretations [12,13]. Machine learning (ML) has garnered increasing

XSL•FO
RenderX

attention in the medical field over the past decade, particularly in cardiovascular medicine [14-16]. ML is important since it can discern critical features from complex datasets [17].

Given its capacity to process intricate, high-volume health data, ML has ascended as an indispensable tool for developing predictive models in medicine [17,18]. In this context, researchers have begun constructing ML-based approaches for diagnosing LVH.

Nonetheless, there are various models and modeling variables for constructing ML models [17,19,20]. Multiple ML models have been proposed for detecting LVH [1,10,21,22], yet their comparative diagnostic accuracy remains debated. Robust evidence assessing both the accuracy and real-world implementation of ML models for LVH identification remains insufficient. Therefore, a systematic review and meta-analysis was conducted in this study to evaluate the performance of ML in detecting LVH and furnish empirical support for artificial intelligence (AI) development in this field.

Methods

Study Registration

This study followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines and was prospectively registered in PROSPERO (registration ID: CRD42024617183) [23].

Eligibility Criteria

Detailed inclusion and exclusion criteria were formulated for this review. To improve the visualization effect, these standards were presented in tabular form (Textbox 1).

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Population: individuals suspected of having left ventricular hypertrophy
- Intervention: machine learning (ML) models for left ventricular hypertrophy identification
- Comparison: no diagnostic tool was used as a comparator
- Outcome: quantitative diagnostic performance metrics, including C-statistic, sensitivity, specificity, accuracy, recall, precision, confusion matrices, diagnostic 2×2 tables, F_1-score, or calibration curves
- Study design: case-control, cohort, or cross-sectional studies published in English

Exclusion criteria

- Intervention: studies that only examined differential factors without developing complete ML models
- Outcomes: studies lacking the following metrics for ML performance: C-statistic, sensitivity, specificity, accuracy, recall, precision, confusion matrices, diagnostic 2×2 tables, F_1-score, or calibration curves
- Study design: meta-analyses, reviews, guidelines, expert opinions, and unpublished conference abstracts

Data Sources and Search Strategy

PubMed, Embase, Cochrane Library, and Web of Science were systematically searched up to November 12, 2025, using both Medical Subject Headings (MeSH) terms and free-text words. Search terms included combinations of “artificial intelligence,” “deep learning,” “machine learning,” “random forest,” “support vector machine,” and “left ventricular hypertrophy.” No filters were applied to constrain studies by publication year or geographic region. The details on the search strategy are available in Multimedia Appendix 1.

Study Selection and Data Extraction

All identified records were uploaded to EndNote, and duplicates were filtered out. The titles and abstracts of the remaining studies were screened for potentially relevant studies. Full texts of the selected studies were obtained and evaluated for eligibility. An electronic extraction sheet was prepared in advance to record author information, publication year, country of the author, study design, patient sources, generation method of validation set, model types, and modeling variables. The entire screening process and data extraction were performed independently by 2 investigators, followed by cross-verification. Discrepancies were addressed through discussion with a third researcher.

Risk of Bias in Studies

The Prediction Model Risk of Bias Assessment Tool (PROBAST) was leveraged to appraise the risk of bias (RoB) in the included original studies [24]. The evaluation covers 4 domains: participants, predictor variables, outcomes, and statistical analysis, which collectively determine the overall RoB and applicability. Each domain contains specific signaling questions (2, 3, 6, and 9 questions, respectively), with 3 possible responses: “yes/probably yes” (low RoB), “no or probably no” (high RoB), and “no information” (unclear). Two researchers independently conducted the PROBAST assessments, followed by cross-checking. Any dissents were addressed through discussion with a third researcher.

Synthesis Methods

Some studies had multiple validation sets. When both an internal and an external validation set were available, only the results from the external set were retained. If an external validation set is unavailable, the results from the internal validation set with optimal performance were selected.

XSL•FO
RenderX

A meta-analysis was conducted on the validation set, leveraging diagnostic 2×2 tables. For studies without 2×2 tables, sensitivity, specificity, positive predictive value, and accuracy, combined with case numbers, were used for estimation (formulas 1 - 5). The bivariate mixed-effects model was used to pool diagnostic metrics (sensitivity, specificity, positive likelihood ratio [PLR], negative likelihood ratio [NLR], diagnostic odds ratio [DOR], and summary receiver operating characteristic [SROC]). Deek funnel plot was used for assessing publication bias [25], and Fagan nomogram was adopted for evaluating clinical utility [26]. Subgroup analyses were carried out based on modeling variables and model types. A significance threshold of $P<.05$ was established.

TP=Sensitivity×Events (1)

TN=Specificity×(sample size − Events) (2)

FN=Events-TP (3)

FP=samplesize-Events-TN (4)

OR

FP=TP/precision-TP (5)

Results

Study Selection

A total of 1382 records were searched. After removing duplicates, 1086 records remained. Then, 1019 records were excluded through title and abstract screening, leaving 67 records for full-text review. Of those, 42 reports were removed due to incomplete data, incompatible outcomes, and incompatible interventions. Ultimately, 25 studies were selected for analysis (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart for literature screening.

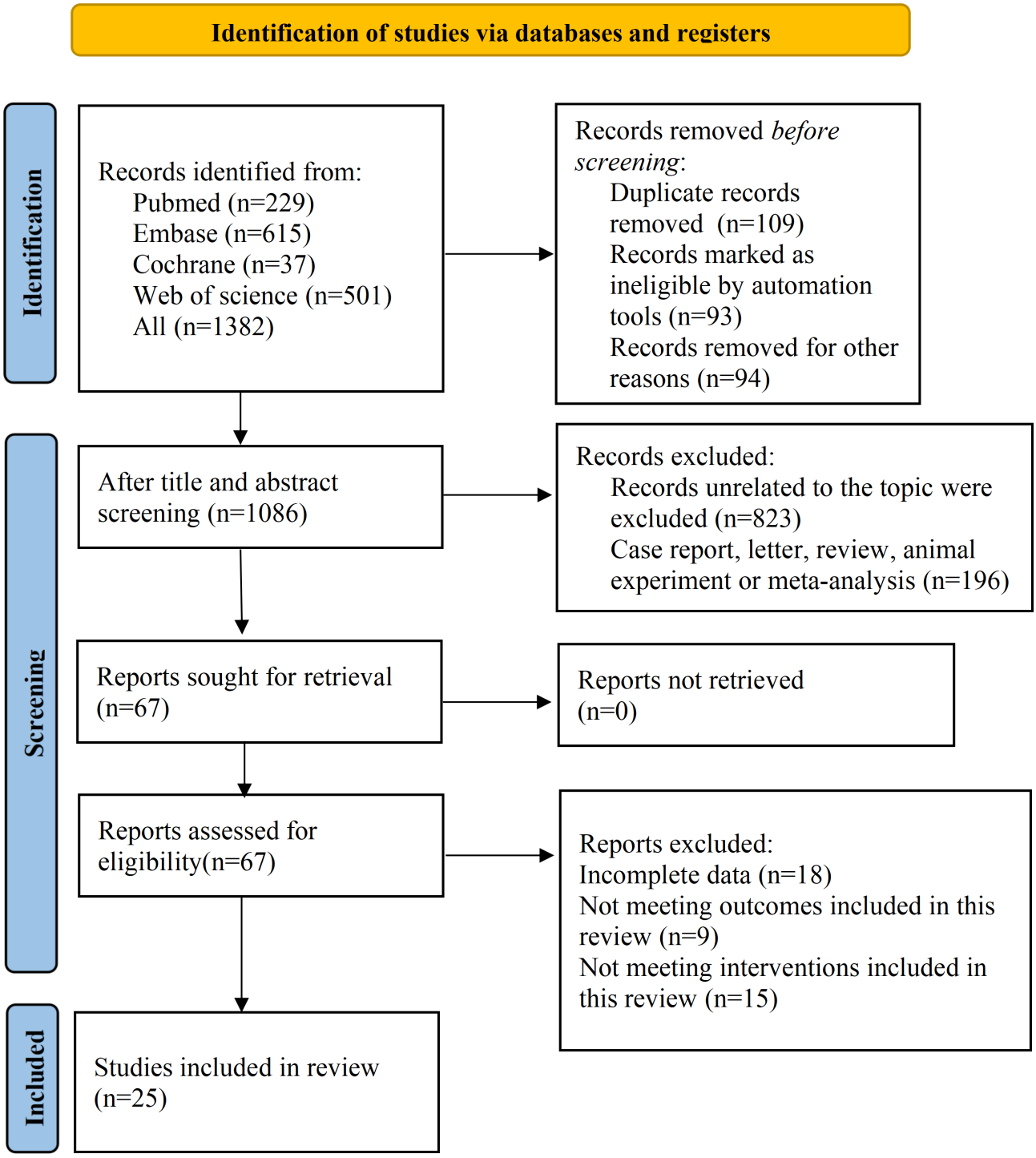

Study Characteristics

Of the 25 case-control studies included (2018 - 2025) [1,10,21,22,27-47], 19 used single-center designs, and 6 adopted multicenter approaches (including 4 studies based on databases that leveraged the UK Biobank resource). These studies were conducted across diverse regions: 10 studies in China, 6 in the United States, 2 in Japan, 2 in Korea, 1 in Mexico, 1 in the United Kingdom, 1 in Greece, 1 in Italy, and 1 in the United Arab Emirates. ECG-based models were included in 16 studies, clinical feature-based models were included in 5 studies, and ECHO-based models were included in 4 studies. Six studies performed external validation, including 5 ECG-based models and 1 clinical feature-based model. The main features of these studies are summarized in Table 1.

XSL•FO
RenderX

Table . Basic features of the included studies.

Author and publication year	Country of author	Study design	Patient source	Validation set generation method	Model type	Modeling variable	Reference standard
Ryu et al [22] (2023)	Korea	Case-control study	Single center	RS[a]	DL[b]	12-lead ECG[c]	ECHO[d]
Farhad et al [28] (2023)	United Arab Emirates	Case-control study	LVH[e] database	RS 10 folds	DL	ECHO	Clinical consensus
Zhao et al [32] (2022)	China	Case-control study	Single center	Folds RS	DL	12-lead ECG	Clinical consensus
Liu et al [33] (2022)	China	Case-control study	Multicenter	Folds EV[f]	DL	12-lead ECG	Clinical consensus
De la Garza Salazar et al [45] (2021)	Mexico	Case-control study	Single center	RS EV	DT[g]	12-lead ECG	Clinical consensus
Kwon et al [40] (2020)	Korea	Case-control study	Multicenter	RS EV	DL	12-lead ECG	Clinical consensus
Ghorbani et al [39] (2020)	United States	Case-control study	Single center	RS	DL	ECHO	Clinical consensus
Sparapani et al [43] (2019)	United States	Case-control study	The MESA[h] database	RS	DT	12-lead ECG	CMR[i]
Madani et al [44] (2018)	United States	Case-control study	Single center	RS	DL	ECHO	Clinical consensus
Yuan et al [29] (2023)	China	Case-control study	Single center	RS	LR[j]	Clinical features	Clinical consensus
Dwivedi et al [30] (2023)	United States	Case-control study	Single center	RS	RF[k]	12-lead ECG	ECHO
Yu et al [35] (2022)	China	Case-control study	Single center	RS	DL	ECHO	Clinical consensus
Wu et al [34] (2022)	China	Case-control study	Single center	EV	LR	Clinical features	Clinical consensus
Kokubo et al [36] (2022)	Japan	Case-control study	Single center	RS	DL	12-lead ECG	Clinical consensus
Ye et al [37] (2021)	China	Case-control study	Single center	10 folds[l]	LASSO[m] regression	Clinical features	Clinical consensus
Angelaki et al [38] (2021)	Greece	Case-control study	Single center	RS	RF	12-lead ECG	Clinical consensus
Lin and Liu [41] (2020)	United States	Case-control study	Single center	RS 4 folds	SVM[o]	12-lead ECG	ECHO
Tison et al [42] (2019)	United States	Case-control study	Single center	RS	GBM[n]	12-lead ECG	ECHO
Liu et al [31] (2023)	China	Case-control study	Single center	RS	ANN[p]	12-lead ECG	Clinical consensus
Zhang et al [1] (2024)	China	Case-control study	Single center	RS	LR	Clinical features	Clinical consensus
Wan et al [27] (2024)	China	Case-control study	Single center	RS	LR	Clinical features	Clinical consensus
Naderi et al [21] (2024)	United Kingdom	Case-control study	UK Biobank Database+ Single center	RS EV	SVM	12-lead ECG	CMR
Cai et al [10] (2024)	Japan	Case-control study	Single center	RS	DL	12-lead ECG	Clinical consensus
Huang et al [46] (2025)	China	Case-control study	Single center	RS	CatBoost	12-lead ECG	Clinical consensus

XSL•FO
RenderX

Author and publication year	Country of author	Study design	Patient source	Validation set generation method	Model type	Modeling variable	Reference standard
Taconné et al [47] (2025)	Italy	Case-control study	PTB-XL ECG database+ Georgia 12-lead ECG Challenge Database	10 folds EV	SVM	12-lead ECG	Clinical consensus

[a]RS: random sampling.
[b]DL: deep learning.
[c]ECG: electrocardiography.
[d]ECHO: echocardiography.
[e]LVH: left ventricular hypertrophy.
[f]EV: external validation.
[g]DT: decision tree.
[h]MESA: Multi-Ethnic Study of Atherosclerosis.
[i]CMR: cardiac magnetic resonance.
[j]LR: logistic regression.
[k]RF: random forest.
[l]10 folds: 10-fold cross-validation.
[m]LASSO: least absolute shrinkage and selection operator.
[n]GBM: gradient boosted machine.
[o]SVM: support vector machine.
[p]ANN: artificial neural network.

RoB in Studies

A high RoB was identified in 21 case-control studies (single- and multi-center) due to potential selection bias. An unclear RoB was assigned to 4 studies based on public databases. The 21 case-control studies were rated as having a high RoB due to selection bias. Due to inherent selection biases, such as the "healthy volunteer" effect commonly found in biobanks, the 4 studies based on public databases were rated as having an unclear RoB. Five case-control studies utilized clinical features for modeling and were rated as having a high RoB due to case-control influence during data collection. The remaining studies, which used image-based ML models, were assessed as having a low RoB in terms of predictors. All studies used valid outcome definitions, and the avoidance of predictor variables in model construction ensured a low RoB. For the imaging-based validation set, the accurate calculation of events per variable (EPV) proved challenging, leading to an uncertain RoB. Developers typically rely on EPV, especially EPV of 10 higher, to determine the minimum sample size required and the maximum number of candidate predictions that can be tested [48]. However, when building models based on images, the basic unit of validation data is the image, not the traditional "patient case." Therefore, we cannot calculate the sample size that meets the EPV requirements based on the number of patient cases, making it difficult to directly apply this standard in such studies. In contrast, studies using clinical features for modeling consistently met the EPV>10 criterion and thus were classified as low RoB. In 2 studies, data were segmented according to images, and patients' images were shared between the training set and the validation set, which may lead to data leakage, and thus were classified as high RoB (Figure 2).

XSL•FO
RenderX

Figure 2. Risk of bias of the included machine learning models.

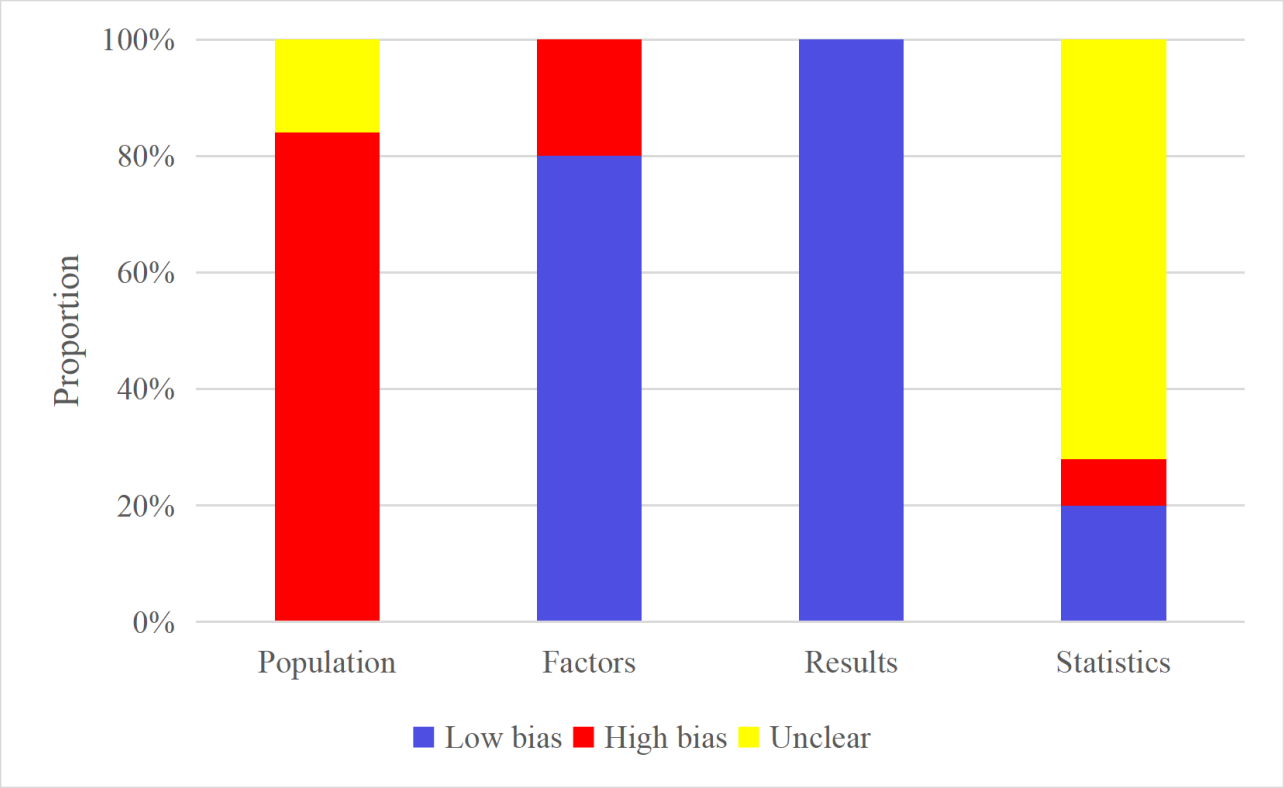

Meta-Analysis

ECG-Based ML Model for LVH Detection

In the validation set, the performance of ECG-based ML models was evaluated in 16 diagnostic 2×2 tables. The meta-analysis results demonstrated the following estimates: sensitivity 0.76 (95% CI 0.66 - 0.84), specificity 0.84 (95% CI 0.78 - 0.89), PLR 4.8 (95% CI 3.3 - 7.0), NLR 0.28 (95% CI 0.19 - 0.41), DOR 17 (95% CI 9 - 33), and SROC 0.88 (95% CI 0.75 - 0.95; Figure 3 and Figure S1 in Multimedia Appendix 2). The Deek funnel plot showed no significant publication bias among the included studies (*P*=.43; Figure S2 in Multimedia Appendix 2). Given a pretest high-risk probability of 10%, if the model result indicated LVH, the posterior probability of LVH was 35%; if the model result indicated non-LVH, the posterior probability of LVH was 3% (Figure S3 in Multimedia Appendix 2).

XSL•FO
RenderX

Figure 3. Forest plot of sensitivity and specificity for electrocardiogram-based models [10,21,22,30-33,36,38,40-43,45-47]. ANN: artificial neural network; DT: decision tree; GBM: gradient boosting machine; RF: random forest; SVM: support vector machine.

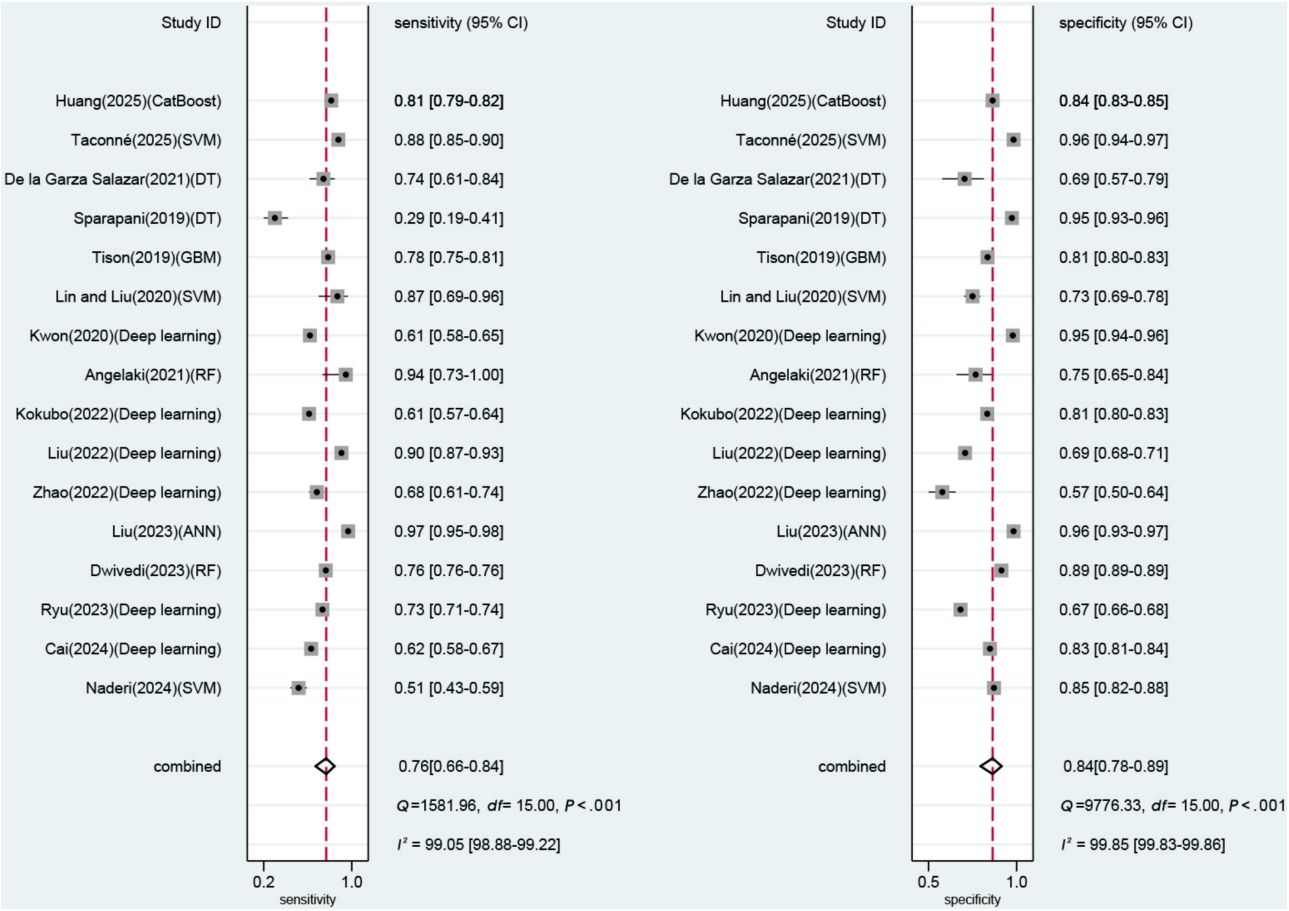

Subgroup analyses were performed across various ML approaches. The deep learning (DL) model produced a sensitivity of 0.71 (95% CI 0.60 - 0.80) and a specificity of 0.79 (95% CI 0.65 - 0.88).

Subgroup analyses were performed across various validation approaches. The results showed that internal validation produced a sensitivity of 0.77 (95% CI 0.64 - 0.86) and a specificity of 0.83 (95% CI 0.75 - 0.89), while external validation produced a sensitivity of 0.76 (95% CI 0.60 - 0.87) and a specificity of 0.87 (95% CI 0.74 - 0.94; Table 2).

Table . Subgroup analyses of electrocardiography (ECG)–based machine learning (ML) models across various validation approaches.

Methods	Machine learning approaches, n	Sen[a] (95% CI)	Spe[b] (95% CI)	PLR[c] (95% CI)	NLR[d] (95% CI)	DOR[e] (95% CI)	SROC[f] (95% CI)
IV[g]	11	0.77 (0.64 - 0.86)	0.83 (0.75 - 0.89)	4.5 (2.9 - 6.8)	0.28 (0.17 - 0.46)	16 (7-35)	0.87 (0.74 - 0.94)
EV[h]	5	0.76 (0.60 - 0.87)	0.87 (0.74 - 0.94)	5.8 (2.8 - 12.2)	0.28 (0.16 - 0.48)	21 (7-60)	0.89 (0.52 - 0.98)
Overall	16	0.76 (0.66 - 0.84)	0.84 (0.78 - 0.89)	4.8 (3.3 - 7.0)	0.28 (0.19 - 0.41)	17 (9-33)	0.88 (0.75 - 0.95)

[a]Sen: sensitivity.

[b]Spe: specificity.

[c]PLR: positive likelihood ratio.

[d]NLR: negative likelihood ratio.

[e]DOR: diagnostic odds ratio.

[f]SROC: summary receiver operating characteristic.

[g]IV: internal validation.

[h]EV: external validation.

XSL•FO
RenderX

As for internal validation, the meta-analysis results demonstrated an SROC of 0.87 (95% CI 0.74 - 0.94; Figure S4 in Multimedia Appendix 2). The Deek funnel plot showed no significant publication bias among the included studies (*P*=.30; Figure S5 in Multimedia Appendix 2). Given a pretest high-risk probability of 10%, if the model result indicated LVH, the posterior probability of LVH was 33%; if the model result indicated non-LVH, the posterior probability of LVH was 3% (Figure S6 in Multimedia Appendix 2).

As for external validation, the meta-analysis results showed an SROC of 0.89 (95% CI 0.52 - 0.98; Figure S7 in Multimedia Appendix 2). The Deek funnel plot showed no significant publication bias across the included studies (*P*=.37; Figure S8 in Multimedia Appendix 2). Given a pretest high-risk probability of 10%, if the model result indicated LVH, the posterior probability of LVH was 39%; if the model result indicated non-LVH, the posterior probability of LVH was 3% (Figure S9 in Multimedia Appendix 2).

Dwivedi et al [30] used image-level splitting, which involved sharing patient images between training and validation. This method posed a high RoB due to data leakage. Accordingly, a sensitivity analysis was conducted on the ECG-based ML model group to exclude this study. The meta-analysis results demonstrated the following estimates: sensitivity 0.76 (95% CI 0.66 - 0.85), specificity 0.84 (95% CI 0.77 - 0.89), PLR 4.7 (95% CI 3.2 - 7.0), NLR 0.28 (95% CI 0.19 - 0.42), DOR 17 (95% CI 8 - 33), and SROC 0.88 (95% CI 0.74 - 0.94).

Clinical Feature–Based ML Model for LVH Detection

In the validation set, there were 5 diagnostic 2×2 tables for evaluating the performance of ML models based on clinical features. The meta-analysis results demonstrated the following estimates: sensitivity 0.78 (95% CI 0.69 - 0.85), specificity 0.71 (95% CI 0.65 - 0.76), PLR 2.7 (95% CI 2.1 - 3.4), NLR 0.31 (95% CI 0.21 - 0.46), DOR 9 (95% CI 5 - 16), and SROC 0.79 (95% CI 0.62 - 0.90; Figure 4 and Figure S10 in Multimedia Appendix 2). The Deek funnel plot showed no significant publication bias among the included studies (*P*=.09; Figure S11 in Multimedia Appendix 2). Given a pretest high-risk probability of 10%, if the model result indicated LVH, the posterior probability of LVH was 23%; if the model result indicated non-LVH, the posterior probability of LVH was 3% (Figure S12 in Multimedia Appendix 2).

Figure 4. Forest plot of sensitivity and specificity for clinical feature-based models [1,27,29,34,37]. LASSO: least absolute shrinkage and selection operator; LR: logistic regression.

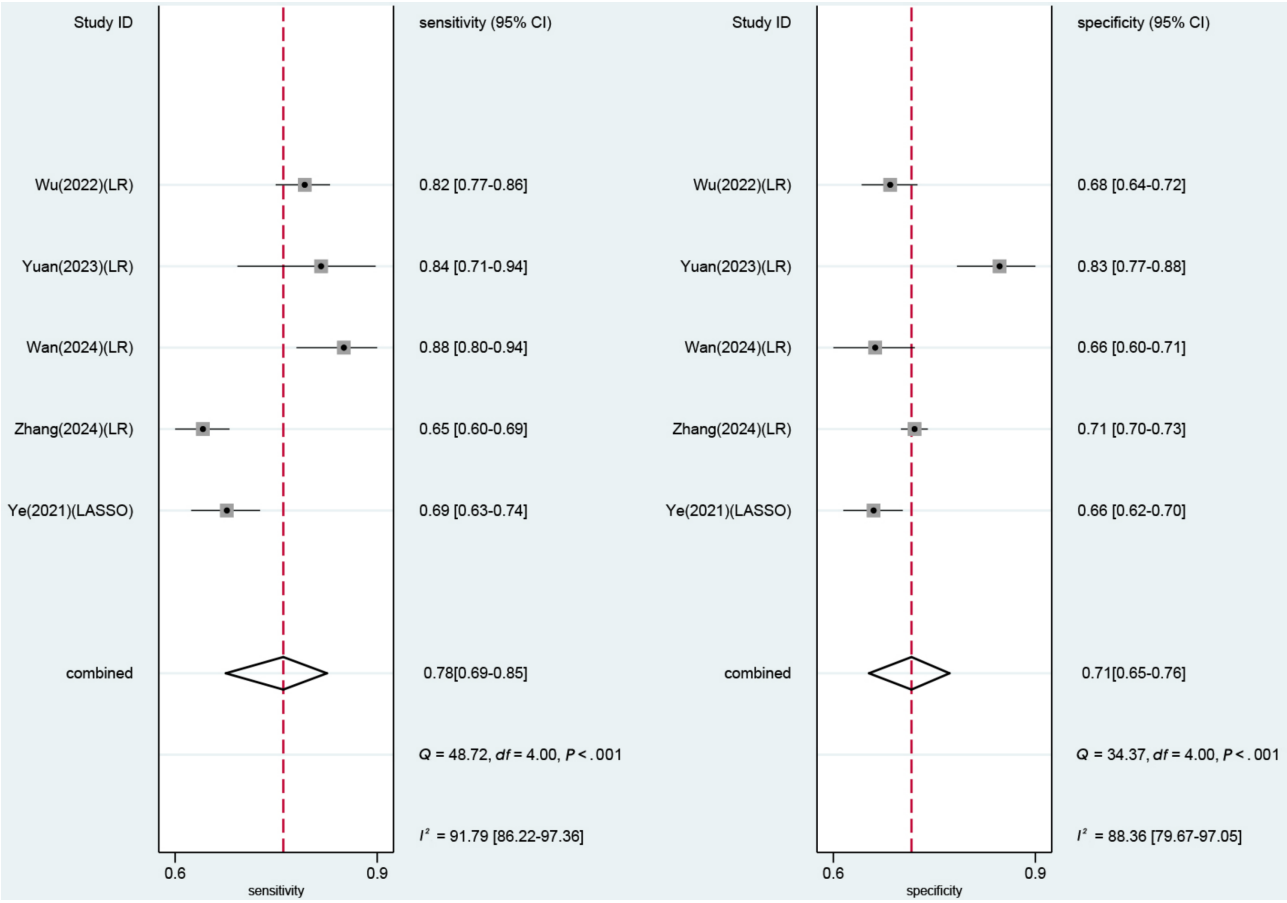

ECHO-Based ML Model for LVH Detection

Farhad et al [28] used image-level splitting, which involved sharing patient images between training and validation. This method posed a high RoB due to data leakage. Accordingly, a sensitivity analysis was conducted in the ECHO-based ML model group to exclude this study. Ghorbani et al [39] developed a DL model based on ECHO. In their validation set of 373 cases (including 142 LVH cases), the model had a pooled sensitivity of 0.71 and a specificity of 0.67. Madani et al [44] also developed a DL model using echocardiographic data. Their validation set comprised 200 cases (100 LVH cases), with the

XSL•FO
RenderX

model demonstrating a pooled sensitivity of 0.78 and a specificity of 0.96. Similarly, Yu et al [35] implemented a DL approach for ECHO. Their validation set included 150 cases (97 LVH cases), and the model yielded a pooled sensitivity of 0.94 and a specificity of 0.92.

Discussion

Principal Findings

This study examined the efficacy of 3 types of ML models for LVH detection. In the validation cohort, ECG-based models had a sensitivity of 0.76 (95% CI 0.66 - 0.84) and a specificity of 0.84 (95% CI 0.78 - 0.89). The clinical feature-based models had a sensitivity of 0.78 (95% CI 0.69 - 0.85) and a specificity of 0.71 (95% CI 0.65 - 0.76). The ECHO-based models had a sensitivity ranging from 0.71 to 0.94 and a specificity ranging from 0.67 to 0.96. These findings confirm the significant efficacy of ML in LVH detection.

Comparison to Prior Work

ML is increasingly accurate in LVH detection, which enables the identification and treatment of earlier subclinical cardiovascular diseases [49]. In clinical practice, ECG is now widely used as a screening method for LVH due to its affordability and high accessibility, and is recommended as part of standard clinical evaluation for patients with hypertension in the Chinese Hypertension Clinical Practice Guidelines [50]. ECG-based LVH detection has been a major research priority in hypertensive cardiac target organ damage studies [1]. The clinical diagnosis of LVH via 12-lead ECG relies principally on the quantitative assessment of QRS complex (Q-wave, R-wave, S-wave) voltage parameters [51,52]. Despite recent innovations in diagnostic criteria, the overall sensitivity remains suboptimal, particularly in Chinese and other East Asian populations [1,53,54]. The low sensitivity of ECG in LVH detection limits its clinical application [32]. Consistent with existing research, our results showed that ECG-based models produced a sensitivity of 0.76 (95% CI 0.66 - 0.84) and a specificity of 0.84 (95% CI 0.78 - 0.89). Although limited by its sensitivity, studies have confirmed that ECG-detected LVH is associated with adverse cardiovascular outcomes [32,55]. Given its widespread availability and low cost, ECG remains an ideal initial assessment tool for LVH screening in patients with hypertension [56]. Therefore, some researchers recommend using ECG primarily for the risk stratification and prognosis prediction of LVH [43,57].

It has also been noted that different ML methods show significant differences in their ability to identify positive or outcome events. With its powerful automatic feature extraction capabilities, DL models can directly identify complex and subtle patterns from imaging data, integrate multiple imaging variables, and improve diagnostic accuracy [58,59]. Its automated segmentation reduces the subjective bias of traditional manual segmentation in radiomics and can integrate segmentation with diagnosis, offering significant advantages [60,61]. Therefore, it is imperative to prioritize the development of automated methods based on DL. Decision tree models are highly interpretable models that generate rules by recursively splitting variables in a dataset [62]. Decision tree models add a penalty term proportional to the absolute value of the feature coefficient for feature selection and model building [63]. Multiple studies have confirmed that RF models typically exhibit higher predictive performance than other ML methods because they can model nonlinear relationships and are insensitive to overfitting [64]. RF models demonstrate significant advantages in handling imbalanced datasets [65]. RF models excel at handling nonlinear and missing data, are conducive to high-dimensional data analysis, and demonstrate strong predictive performance [66]. Support vector machine (SVM) models have high predictive accuracy [67]. For structured and well-defined data, SVM models can skillfully capture the association between biomarkers and diseases through mathematical and statistical principles, thereby achieving high diagnostic accuracy [68,69]. SVM models are known for their ability to find the optimal decision boundary in high-dimensional feature spaces, resulting in effective classification based on extracted features [70,71].

ML models that incorporated clinical features demonstrated a sensitivity of 0.78 (95% CI 0.69 - 0.85) and a specificity of 0.71 (95% CI 0.65 - 0.76) in the validation set. Unlike traditional ECG criteria, the clinical features used to construct these models encompassed demographic characteristics (eg, sex, age) and clinical factors (eg, systolic blood pressure, duration of hypertension, and abdominal obesity). The possible reasons for using these factors are as follows.

Sex constitutes a significant determinant of LVH [1,37]. Substantial evidence has established the female sex as an independent predictor. In previous predictive models, women scored 70, while men scored 0, possibly due to the older study population, where most women had reached menopause [1]. Women exhibit higher salt sensitivity than men due to female hormonal influences, which amplifies adverse renin-angiotensin-aldosterone system effects [72,73]. Regardless of race or menopausal status, women with salt-sensitive hypertension face a higher LVH risk because this sex-specific mechanism promotes aldosterone-driven endothelial dysfunction [72,73]. This association is further exacerbated by diminished cardiovascular protection due to decreased estrogen levels in postmenopausal women [37].

Age plays a pivotal role in forecasting LVH in young patients with salt-sensitive hypertension. The evidence reveals a dual pattern: (1) patients 40 years or older demonstrate a significantly higher risk than younger populations, consistent with age-related left ventricular mass increase [1,37,74], while (2) patients with early-onset hypertension exhibit more severe cardiac damage and elevated cardiovascular mortality [74,75]. Such observations suggest that the interaction between age and the timing of hypertension requires special attention.

The development of LVH is closely associated with systolic blood pressure level and duration of hypertension. Prospective studies demonstrate that each 19 mm Hg increase in systolic blood pressure elevates the LVH risk by 49% [6], while effective systolic blood pressure control markedly reduces the disease risk in old patients [76]. These findings provide crucial evidence for clinical treatment decisions. Hypertension duration, as a cumulative exposure metric, has been consistently validated in

XSL•FO
RenderX

multiple studies to be significantly associated with LVH. Shortening the course of hypertension can effectively decrease the risk of myocardial remodeling, which shows the importance of early intervention in delaying LVH progression [77].

Obesity constitutes an independent risk factor for LVH [78,79]. Waist circumference, the principal diagnostic metric for abdominal obesity, is linked with both cardiovascular and metabolic disorders [80]. The pathophysiological linkage between abdominal obesity and LVH can be attributed to the robust correlation between waist circumference and excessive visceral adipose tissue accumulation [80,81]. As a metabolically active depot, visceral adipose tissue exhibits substantially higher energy demands compared to subcutaneous fat [81]. This augmented metabolic requirement exacerbates cardiac workload in patients with hypertension, ultimately resulting in LVH [78,81]. Studies have identified waist circumference as a superior predictor of cardiovascular risk and mortality compared to BMI [80,82]. Notably, elevated systolic blood pressure and prolonged hypertension duration may synergistically accelerate concentric cardiac remodeling, while abdominal obesity plus aging compounds metabolic risk.

The complex interactions among clinical features were quantified and integrated into ML models. While previous studies have confirmed the underlying mechanism of individual factors, models combining features from diverse dimensions provide more complete predictive insights. Specifically, sex-specific risk mechanisms dynamically interact with age-related cardiovascular changes, while blood pressure's temporal cumulative effects and the spatial distribution of metabolic factors form a 3D LVH prediction network. This finding provides a theoretical foundation for developing clinical feature–based ML tools for precise prediction.

ECHO serves as the primary imaging tool for LVH identification. However, its clinical application faces multiple technical challenges, including a relatively low signal-to-noise ratio, inconsistent image quality, and suboptimal reproducibility [28,35]. These limitations are particularly pronounced in patients with obesity due to enhanced chest wall and subcutaneous tissue thickness, which significantly impede image acquisition [28,35]. ML technology offers a novel approach to overcoming these technical barriers. ML models incorporating ECHO demonstrated a sensitivity range of 0.71 to 0.94 and a specificity range of 0.67 to 0.96.

Existing hospital ECHO systems are integrated with this technology to enable automated image analysis and real-time LVH diagnosis. Without replacing existing manual verification procedures, it markedly increases diagnostic efficiency and proves particularly beneficial as a clinical decision aid for less experienced physicians [28]. In contemporary practice, LVH detection chiefly relies on ECHO, where ML proves to be powerful in image processing stages [35,83]. Conventional approaches that depend on manual segmentation and feature extraction are prone to information loss, whereas DL models incorporate image processing, texture extraction, and feature selection throughout the modeling process, which better preserves original image information [35,84,85]. All ML models based on ECHO in this study were DL models. Beyond improved diagnostic efficiency, ML models can automatically identify high-risk cases and enable cardiologists to prioritize urgent interventions, thus shortening care pathways and improving outcomes [49]. Additionally, this technique can enhance diagnostic capabilities at primary care levels, optimize the allocation of cardiovascular imaging resources, reduce unnecessary tests, and be significantly cost-effective in population screening and subclinical disease management [28].

Despite the superior image-processing capabilities of DL models, their clinical integration with ultrasound systems must overcome real-world challenges, including real-time performance, device compatibility, and patient data privacy. Furthermore, the stability of ML models across demographics needs further validation in multicenter research. Future research should prioritize enlarging and diversifying ECHO image databases and validating the actual performance of ML in the detection of LVH in prospective trials.

Meta-analysis of ECG-based ML models for detecting LVH revealed substantial heterogeneity. This restricts the statistical reliability of pooled performance metrics, preventing their interpretation as definitive generalization conclusions. Potential sources of such heterogeneity are manifold. Methodologically, substantial variation in ECG signal preprocessing, feature extraction, and used algorithms introduces fundamental differences. Clinically, diverse underlying disease profiles and severity among studied populations affect model consistency. Regional demographic variations further increase result dispersion. Consequently, pooled estimates from our research require cautious interpretation. Their value lies more in demonstrating the broad performance range and limited confirmatory nature of current evidence than in providing a stable performance expectation. Future efforts should prioritize methodological standardization and prospective multicenter validation to enhance evidence quality in this domain.

When the PROBAST was used to assess model quality, we found that the results raised some concerns, particularly in terms of its strict evaluation of the study population. We believe that some of its criteria may be overly stringent. Although the PROBAST is suitable for evaluating multivariate models for diagnosis and prediction, it is primarily based on retrospective case-control studies, with few prospective studies included [86]. This design preference often leads to a high risk of bias in assessment results, which poses a significant challenge to diagnostic model research. In the studies we included, most involved diagnostic models, and the case selection process itself might introduce bias. The PROBAST uses the EPV principle for sample size estimation, but image-based ML models face challenges in applying EPV rules for bias risk assessment. In future studies, we may explore prior sample size estimation during the research design phase, leveraging ML-based expected performance metrics, such as area under the curve and its 95% CI. Moreover, learning curve analysis serves as a powerful empirical tool. By plotting how a model's performance on both the training and validation sets changes as the sample size increases, we can intuitively determine whether the model's performance has reached a plateau, thereby establishing the optimal number of cases. In the field of DL, there exist several widely accepted rules. For instance, in complex image

XSL•FO
RenderX

classification tasks, it is generally recommended that each category requires at least several thousand or even tens of thousands of training images to achieve good generalization performance. Although these guidelines are somewhat rough, they provide a practical benchmark for assessing sample sizes in large-scale imaging studies. In addition, the PROBAST requires the verification of the consistency between model statistical weights and actual reported results, which is particularly difficult for models with low interpretability (such as SVM), as they typically do not provide the weight information required in the original validation set [24]. These issues indicate that while the PROBAST is widely used for bias risk assessment in ML research, the observed limitations arise more fundamentally from the inherent biases of the retrospective case-control design—which is applied by most of the current ML diagnostic studies—than from an inherent excessive stringency in the PROBAST criteria. Advancing this field requires constructing and validating models through prospective, multicenter, double-blinded cohort studies, thereby improving evidence quality at its source. Specifically, future work should emphasize prospective sample size estimation and learning curve analysis during development, enhance model interpretability, conduct independent external validation, and ultimately assess the real-world impact of a model on clinical workflows and patient outcomes via pragmatic effectiveness studies. Additionally, we recommend using QUADAS-AI (Quality Assessment of Diagnostic Accuracy Studies Using AI) to assess the risk of bias in included original studies for diagnostic ML tasks [87]. QUADAS-AI focuses on data sources, sample size, eligibility criteria, the rationale for splitting training, validation, test sets, imaging protocols, and preprocessing methods—rather than simply labeling all case-control studies as "high risk." In the studies we included, a substantial number of studies clearly described the source, size, and eligibility criteria, the principles of splitting into training, validation, and test sets, as well as the imaging protocols and preprocessing methods, thereby mitigating the high risk of bias associated with the study design.

This study has high statistical heterogeneity. Even when we conducted subgroup analyses using various ML approaches and validation methods to further elucidate the sources of heterogeneity, our explanatory power remains limited. This may be due to underlying heterogeneity present during model training. We consider that the following ML processes may introduce potential heterogeneity. First, regarding image preprocessing, the included studies exhibited certain variations in their approaches. These differences may introduce significant heterogeneity. For different images, there is considerable variation in image parameters, yet there is a lack of effective discussion on how this impacts modeling. The images in the original studies exhibit significant variations across imaging protocols. Although we performed subgroup analyses based on different models, even within the same model, differences in model iterations and parameter adjustment rules persist, potentially contributing to heterogeneity.

Strengths and Limitations of the Study

Our systematic review is the first to quantitatively examine ML performance in LVH detection. While informative, several limitations should be considered when interpreting these results. First, the systematic search, while rigorous, identified solely small case-control studies, thereby narrowing the spectrum of studies for subgroup comparisons. Second, the limited number of included studies precluded meaningful comparisons between different ML methodologies. Third, while most studies completed external validation, some relied solely on random sampling for internal validation. Owing to the limited number of publications, a systematic comparison between internal and external validation approaches is infeasible. Fourth, the RoB in the included studies is high, which is inevitable given the stringent criteria for evaluating ML-based research. This observation corresponds with methodological appraisals indicating that 87% of ML models for medical prediction carry a high RoB [88]. Future studies should optimize design, implementation, reporting, and validation approaches to facilitate clinical translation of ML-based prediction models. Fifth, some studies divide data according to images, and patients' images are shared between the training set and validation sets, which may result in data leakage. Sixth, there are different reference standards for different models. This discrepancy in reference standards indicates that any comparison of diagnostic accuracy between different model types is inherently confounded. The presented pooled estimates for each subgroup should be interpreted independently and not compared directly to each other. Seventh, the findings of this study may have been influenced to some extent by the fact that 3 papers used the 2×2 formula for estimation. The results should be interpreted with caution. Eighth, several included studies used "clinical consensus" as the reference standard for diagnosing LVH. While this consensus typically adhered to established guidelines (defining LVH as left ventricular mass index >115 g/m² in men and >95 g/m² in women [89]), original studies generally did not explicitly describe whether the consensus process was entirely independent of the outputs of evaluated ML models. Theoretically, "incorporation bias" could be introduced if the predictions of a model were known or incorporated during the adjudication of the reference standard, potentially leading to an overestimation of its diagnostic accuracy. This constitutes a key limitation for evidence interpretation in this systematic review and underscores the necessity for future original research to explicitly report reference standard independence in methodological descriptions.

Future Directions

ML is widely used in LVH research, but its clinical application still faces multiple challenges. Many models are developed based on limited samples and lack standardized processes for key steps, such as image selection, segmentation, and feature extraction, resulting in high reliance of models on the experience of developers [62,90]. This raises concerns about the robustness of the models. Significant differences in image quality across different institutions further weaken the universality of the models. Existing models often focus on accuracy and neglect population risk stratification and regional differences in medical resources, limiting their clinical utility. Considering interpretability and a lack of trust among physicians, it remains difficult to integrate ML into clinical practice [91,92]. Although some models perform well in retrospective studies, their

XSL•FO
RenderX

effectiveness and applicability in real-world clinical settings still need to be verified through rigorous prospective trials. Before clinical application, it is necessary to fully address the issues of patient data privacy protection, lack of trust among doctors due to insufficient model interpretability, and ethical considerations associated with new systems [93-95]. Additionally, we must be wary of overreliance on technological means at the expense of humanistic care. The need for systematic management and individualized treatment should not be overlooked. Technology should serve, rather than replace, patient-centered medical practices.

To promote the safe, effective, and responsible application of ML in LVH clinical practice, it is necessary to standardize the model development process, strengthen research on model generalizability and interpretability, conduct rigorous prospective validation, and focus on risk stratification and resource differences. Furthermore, it is necessary to establish a comprehensive ethical and regulatory framework covering data privacy, patient rights, algorithm transparency, and clinical workflow integration. Clear implementation guidelines are needed to ensure that technological development truly benefits patients and improves the quality of medical care.

Conclusions

This study suggests that ML demonstrates diagnostic potential in detecting LVH. However, the overall certainty of the evidence is low, primarily constrained by the limited number of included studies and substantial heterogeneity. Consequently, current conclusions about model performance should be cautiously interpreted. Future investigations should adopt prospective designs and implement standardized data acquisition and model validation protocols to develop and evaluate ML models with enhanced robustness and clinical interpretability. This will advance this technology and make it a reliable diagnostic tool.

Funding

This research was not supported by any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

All data generated or analyzed during this study are included in this published article and its multimedia appendices.

Authors' Contributions

Conceptualization: JW, KZ, YL
Formal analysis and investigation: JW, KZ, YL
Methodology: JW, KZ, YL
Supervision: JW, KZ, YL
Writing – original draft preparation: JW, KZ, YL
Writing – review and editing: JW, KZ, YL
All authors commented on previous versions of the manuscript, read and approved the final manuscript, and contributed to the study conception and design.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Literature search strategy.
[DOCX File, 20 KB - jmir_v28i1e76637_app1.docx]

Multimedia Appendix 2
Summary receiver operating characteristic curves, funnel plots, and nomograms for electrocardiogram (overall and validation analyses) and clinical features.
[DOCX File, 1410 KB - jmir_v28i1e76637_app2.docx]

Checklist 1
PRISMA flowchart.
[PDF File, 99 KB - jmir_v28i1e76637_app3.pdf]

References

1. Zhang X, He C, Lu S, et al. Construction and validation of a nomogram to predict left ventricular hypertrophy in low‐risk patients with hypertension. J Clin Hypertens 2024 Mar;26(3):274-285. [doi: 10.1111/jch.14780]
2. Đorđević DB, Koračević GP, Đorđević AD, Lović DB. Hypertension and left ventricular hypertrophy. J Hypertens 2024 Sep 1;42(9):1505-1515. [doi: 10.1097/HJH.0000000000003774] [Medline: 38747417]

XSL•FO
RenderX

3. Han Y, Li Y, Wu Z, et al. Progress in diagnosis and treatment of hypertension combined with left ventricular hypertrophy. Ann Med 2024 Dec;56(1):2405080. [doi: 10.1080/07853890.2024.2405080] [Medline: 39301864]
4. Miller RJH, Mikami Y, Heydari B, et al. Sex-specific relationships between patterns of ventricular remodelling and clinical outcomes. Eur Heart J Cardiovasc Imaging 2020 Sep 1;21(9):983-990. [doi: 10.1093/ehjci/jeaa164] [Medline: 32594163]
5. Lewis AA, Ayers CR, Selvin E, et al. Racial differences in malignant left ventricular hypertrophy and incidence of heart failure: a multicohort study. Circulation 2020 Mar 24;141(12):957-967. [doi: 10.1161/CIRCULATIONAHA.119.043628] [Medline: 31931608]
6. Cao X, Broughton ST, Waits GS, Nguyen T, Li Y, Soliman EZ. Interrelations between hypertension and electrocardiographic left ventricular hypertrophy and their associations with cardiovascular mortality. Am J Cardiol 2019 Jan 15;123(2):274-283. [doi: 10.1016/j.amjcard.2018.10.006] [Medline: 30390988]
7. Siranart N, Deepan N, Techasatian W, et al. Diagnostic accuracy of artificial intelligence in detecting left ventricular hypertrophy by electrocardiograph: a systematic review and meta-analysis. Sci Rep 2024 Jul 10;14(1):15882. [doi: 10.1038/s41598-024-66247-y] [Medline: 38987652]
8. Pedersen LR, Kristensen AMD, Petersen SS, et al. Prognostic implications of left ventricular hypertrophy diagnosed on electrocardiogram vs echocardiography. J Clin Hypertens (Greenwich) 2020 Sep;22(9):1647-1658. [doi: 10.1111/jch.13991] [Medline: 32813895]
9. Woythaler JN, Singer SL, Kwan OL, et al. Accuracy of echocardiography versus electrocardiography in detecting left ventricular hypertrophy: comparison with postmortem mass measurements. J Am Coll Cardiol 1983 Aug;2(2):305-311. [doi: 10.1016/s0735-1097(83)80167-3] [Medline: 6223063]
10. Cai C, Imai T, Hasumi E, Fujiu K. One-shot screening: utilization of a two-dimensional convolutional neural network for automatic detection of left ventricular hypertrophy using electrocardiograms. Comput Methods Programs Biomed 2024 Apr;247:108097. [doi: 10.1016/j.cmpb.2024.108097] [Medline: 38428250]
11. Spahillari A, McCormick I, Yang JX, Quinn GR, Manning WJ. On-call transthoracic echocardiographic interpretation by first year cardiology fellows: comparison with attending cardiologists. BMC Med Educ 2019 Jun 14;19(1):213. [doi: 10.1186/s12909-019-1634-7] [Medline: 31200721]
12. Rao S, Ferris TG, Hidrue MK, et al. Physician burnout, engagement and career satisfaction in a large academic medical practice. Clin Med Res 2020 Mar;18(1):3-10. [doi: 10.3121/cmr.2019.1516] [Medline: 31959669]
13. Quinn GR, Ranum D, Song E, et al. Missed diagnosis of cardiovascular disease in outpatient general medicine: insights from malpractice claims data. Jt Comm J Qual Patient Saf 2017 Oct;43(10):508-516. [doi: 10.1016/j.jcjq.2017.05.001] [Medline: 28942775]
14. Wang Z, Gu Y, Huang L, et al. Construction of machine learning diagnostic models for cardiovascular pan-disease based on blood routine and biochemical detection data. Cardiovasc Diabetol 2024 Sep 28;23(1):351. [doi: 10.1186/s12933-024-02439-0] [Medline: 39342281]
15. Qi X, Wang S, Fang C, Jia J, Lin L, Yuan T. Machine learning and SHAP value interpretation for predicting comorbidity of cardiovascular disease and cancer with dietary antioxidants. Redox Biol 2025 Feb;79:103470. [doi: 10.1016/j.redox.2024.103470] [Medline: 39700695]
16. Layton AT. AI, machine learning, and ChatGPT in hypertension. Hypertension 2024 Apr;81(4):709-716. [doi: 10.1161/HYPERTENSIONAHA.124.19468] [Medline: 38380541]
17. Ying Y, Ju R, Wang J, et al. Accuracy of machine learning in diagnosing microsatellite instability in gastric cancer: a systematic review and meta-analysis. Int J Med Inform 2025 Jan;193:105685. [doi: 10.1016/j.ijmedinf.2024.105685] [Medline: 39515046]
18. Beam AL, Kohane IS. Big data and machine learning in health care. J Am Med Assoc 2018 Apr 3;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
19. Zhao K, Zhu Y, Chen X, et al. Machine learning in hypertrophic cardiomyopathy: nonlinear model from clinical and CMR features predicting cardiovascular events. JACC Cardiovasc Imaging 2024 Aug;17(8):880-893. [doi: 10.1016/j.jcmg.2024.04.013] [Medline: 39001729]
20. You J, Guo Y, Kang JJ, et al. Development of machine learning-based models to predict 10-year risk of cardiovascular disease: a prospective cohort study. Stroke Vasc Neurol 2023 Dec 29;8(6):475-485. [doi: 10.1136/svn-2023-002332] [Medline: 37105576]
21. Naderi H, Ramírez J, van Duijvenboden S, et al. Diagnostic and prognostic value of ECG-predicted hypertension-mediated left ventricular hypertrophy using machine learning. medRxiv. Preprint posted online on Apr 23, 2024. [doi: 10.1101/2024.04.22.24306204]
22. Ryu JS, Lee S, Chu Y, Ahn MS, Park YJ, Yang S. CoAt-mixer: self-attention deep learning framework for left ventricular hypertrophy using electrocardiography. PLoS ONE 2023;18(6):e0286916. [doi: 10.1371/journal.pone.0286916] [Medline: 37289800]
23. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]

XSL•FO
RenderX

24. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019 Jan 1;170(1):W1-W33. [doi: 10.7326/M18-1377] [Medline: 30596876]
25. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol 2005 Sep;58(9):882-893. [doi: 10.1016/j.jclinepi.2005.01.016] [Medline: 16085191]
26. Fagan TJ. Letter: nomogram for Bayes's theorem. N Engl J Med 1975 Jul 31;293(5):257. [doi: 10.1056/NEJM197507312930513] [Medline: 1143310]
27. Wan J, Wang P, Liu S, Wang X, Zhou P, Yang J. Risk factors and a predictive model for left ventricular hypertrophy in young adults with salt-sensitive hypertension. J Clin Hypertens (Greenwich) 2024 Aug;26(8):933-944. [doi: 10.1111/jch.14863] [Medline: 38940286]
28. Farhad M, Masud MM, Beg A, Ahmad A, Ahmed LA, Memon S. A data-efficient zero-shot and few-shot Siamese approach for automated diagnosis of left ventricular hypertrophy. Comput Biol Med 2023 Sep;163:107129. [doi: 10.1016/j.compbiomed.2023.107129] [Medline: 37343469]
29. Yuan R, Chen J, Zhang S, Zhang X, Yu J. Establishment of an individual-specific nomogram for predicting the risk of left ventricular hypertrophy in Chinese postmenopausal hypertensive women. Medicina (Kaunas) 2023 Mar 17;59(3):598. [doi: 10.3390/medicina59030598] [Medline: 36984599]
30. Dwivedi T, Xue J, Treiman D, Dubey A, Albert D. Machine learning models of 6-lead ECGs for the interpretation of left ventricular hypertrophy (LVH). J Electrocardiol 2023;77:62-67. [doi: 10.1016/j.jelectrocard.2022.12.001] [Medline: 36641988]
31. Liu CW, Wu FH, Hu YL, et al. Left ventricular hypertrophy detection using electrocardiographic signal. Sci Rep 2023 Feb 13;13(1):2556. [doi: 10.1038/s41598-023-28325-5] [Medline: 36781924]
32. Zhao X, Huang G, Wu L, et al. Deep learning assessment of left ventricular hypertrophy based on electrocardiogram. Front Cardiovasc Med 2022;9:952089. [doi: 10.3389/fcvm.2022.952089] [Medline: 36035939]
33. Liu CM, Hsieh ME, Hu YF, et al. Artificial intelligence-enabled model for early detection of left ventricular hypertrophy and mortality prediction in young to middle-aged adults. Circ Cardiovasc Qual Outcomes 2022 Aug;15(8):e008360. [doi: 10.1161/CIRCOUTCOMES.121.008360] [Medline: 35959675]
34. Wu Z, Shi M, Wang L, Yao Y. Identification of major risk factors and non-linear effects to the development of left ventricular hypertrophy in chronic kidney disease by constructing and validation of nomograms. Front Med (Lausanne) 2022;9:914800. [doi: 10.3389/fmed.2022.914800] [Medline: 35911424]
35. Yu X, Yao X, Wu B, et al. Using deep learning method to identify left ventricular hypertrophy on echocardiography. Int J Cardiovasc Imaging 2022 Apr;38(4):759-769. [doi: 10.1007/s10554-021-02461-3] [Medline: 34757566]
36. Kokubo T, Kodera S, Sawano S, et al. Automatic detection of left ventricular dilatation and hypertrophy from electrocardiograms using deep learning. Int Heart J 2022 Sep 30;63(5):939-947. [doi: 10.1536/ihj.22-132] [Medline: 36104234]
37. Ye C, Wang T, Gong J, et al. Development of a nomogram for screening the risk of left ventricular hypertrophy in Chinese hypertensive patients. J Clin Hyperten 2021 Jun;23(6):1176-1185. [doi: 10.1111/jch.14240]
38. Angelaki E, Marketou ME, Barmparis GD, et al. Detection of abnormal left ventricular geometry in patients without cardiovascular disease through machine learning: an ECG-based approach. J Clin Hypertens (Greenwich) 2021 May;23(5):935-945. [doi: 10.1111/jch.14200] [Medline: 33507615]
39. Ghorbani A, Ouyang D, Abid A, et al. Deep learning interpretation of echocardiograms. NPJ Digit Med 2020;3(10):10. [doi: 10.1038/s41746-019-0216-8] [Medline: 31993508]
40. Kwon JM, Jeon KH, Kim HM, et al. Comparing the performance of artificial intelligence and conventional diagnosis criteria for detecting left ventricular hypertrophy using electrocardiography. Europace 2020 Mar 1;22(3):412-419. [doi: 10.1093/europace/euz324] [Medline: 31800031]
41. Lin GM, Liu K. An electrocardiographic system with anthropometrics via machine learning to screen left ventricular hypertrophy among young adults. IEEE J Transl Eng Health Med 2020;8:1800111. [doi: 10.1109/JTEHM.2020.2990073] [Medline: 32419990]
42. Tison GH, Zhang J, Delling FN, Deo RC. Automated and interpretable patient ECG profiles for disease detection, tracking, and discovery. Circ Cardiovasc Qual Outcomes 2019 Sep;12(9):e005289. [doi: 10.1161/CIRCOUTCOMES.118.005289] [Medline: 31525078]
43. Sparapani R, Dabbouseh NM, Gutterman D, et al. Detection of left ventricular hypertrophy using Bayesian additive regression trees: the MESA. J Am Heart Assoc 2019 Mar 5;8(5):e009959. [doi: 10.1161/JAHA.118.009959] [Medline: 30827132]
44. Madani A, Ong JR, Tibrewal A, Mofrad MRK. Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. NPJ Digit Med 2018;1(1):59. [doi: 10.1038/s41746-018-0065-x] [Medline: 31304338]
45. De la Garza Salazar F, Romero Ibarguengoitia ME, Azpiri López JR, González Cantú A. Optimizing ECG to detect echocardiographic left ventricular hypertrophy with computer-based ECG data and machine learning. PLoS ONE 2021;16(11):e0260661. [doi: 10.1371/journal.pone.0260661] [Medline: 34847202]

XSL•FO
RenderX

46. Huang JT, Tseng CH, Huang WM, et al. Comparison of machine learning and conventional criteria in detecting left ventricular hypertrophy and prognosis with electrocardiography. Eur Heart J Digit Health 2025 Mar;6(2):252-260. [doi: 10.1093/ehjdh/ztaf003] [Medline: 40110220]
47. Taconne M, Corino VDA, Mainardi L. An ECG-based model for left ventricular hypertrophy detection: a machine learning approach. IEEE Open J Eng Med Biol 2025;6:219-226. [doi: 10.1109/OJEMB.2024.3509379] [Medline: 39698126]
48. van Smeden M, Moons KG, de Groot JA, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. Stat Methods Med Res 2019 Aug;28(8):2455-2474. [doi: 10.1177/0962280218784726] [Medline: 29966490]
49. Duffy G, Cheng PP, Yuan N, et al. High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning. JAMA Cardiol 2022 Apr 1;7(4):386-395. [doi: 10.1001/jamacardio.2021.6059] [Medline: 35195663]
50. Liu LS, Writing Group of 2010 Chinese Guidelines for the Management of Hypertension. 2010 Chinese guidelines for the management of hypertension. Zhonghua Xin Xue Guan Bing Za Zhi 2011 Jul;39(7):579-615. [Medline: 22088239]
51. Casale PN, Devereux RB, Kligfield P, et al. Electrocardiographic detection of left ventricular hypertrophy: development and prospective validation of improved criteria. J Am Coll Cardiol 1985 Sep;6(3):572-580. [doi: 10.1016/s0735-1097(85)80115-7] [Medline: 3161926]
52. Sokolow M, Lyon TP. The ventricular complex in left ventricular hypertrophy as obtained by unipolar precordial and limb leads. Am Heart J 1949 Feb;37(2):161-186. [doi: 10.1016/0002-8703(49)90562-1] [Medline: 18107386]
53. Xia Y, Li X, Zhang H, et al. Diagnostic capability and influence factors for a new electrocardiogram criterion in the diagnosis of left ventricular hypertrophy in a Chinese population. Cardiology 2020;145(5):294-302. [doi: 10.1159/000505421] [Medline: 32289773]
54. Wang D, Xu JZ, Zhang W, et al. Performance of electrocardiographic criteria for echocardiographically diagnosed left ventricular hypertrophy in Chinese hypertensive patients. Am J Hypertens 2020 Sep 10;33(9):831-836. [doi: 10.1093/ajh/hpaa083] [Medline: 32484222]
55. Bang CN, Devereux RB, Okin PM. Regression of electrocardiographic left ventricular hypertrophy or strain is associated with lower incidence of cardiovascular morbidity and mortality in hypertensive patients independent of blood pressure reduction—a LIFE review. J Electrocardiol 2014;47(5):630-635. [doi: 10.1016/j.jelectrocard.2014.07.003] [Medline: 25052475]
56. Hancock EW, Deal BJ, Mirvis DM, et al. AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: part V: electrocardiogram changes associated with cardiac chamber hypertrophy: a scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society: endorsed by the International Society for Computerized Electrocardiology. Circulation 2009 Mar 17;119(10):e251-e261. [doi: 10.1161/CIRCULATIONAHA.108.191097] [Medline: 19228820]
57. Rautaharju PM, Soliman EZ. Electrocardiographic left ventricular hypertrophy and the risk of adverse cardiovascular events: a critical appraisal. J Electrocardiol 2014;47(5):649-654. [doi: 10.1016/j.jelectrocard.2014.06.002] [Medline: 25012077]
58. Jiang Y, Yang M, Wang S, Li X, Sun Y. Emerging role of deep learning-based artificial intelligence in tumor pathology. Cancer Commun (Lond) 2020 Apr;40(4):154-166. [doi: 10.1002/cac2.12012] [Medline: 32277744]
59. Jiang X, Li J, Kan Y, et al. MRI based radiomics approach with deep learning for prediction of vessel invasion in early-stage cervical cancer. IEEE/ACM Trans Comput Biol Bioinform 2021;18(3):995-1002. [doi: 10.1109/TCBB.2019.2963867] [Medline: 31905143]
60. She L, Li Y, Wang H, et al. Imaging-based AI for predicting lymphovascular space invasion in cervical cancer: systematic review and meta-analysis. J Med Internet Res 2025 Jun 16;27:e71091. [doi: 10.2196/71091] [Medline: 40523236]
61. Zuo H, Huang B, He J, Fang L, Huang M. Machine learning approaches in high myopia: systematic review and meta-analysis. J Med Internet Res 2025 Jan 3;27:e57644. [doi: 10.2196/57644] [Medline: 39753217]
62. Zhang H, Zou P, Luo P, Jiang X. Machine learning for the early prediction of delayed cerebral ischemia in patients with subarachnoid hemorrhage: systematic review and meta-analysis. J Med Internet Res 2025 Jan 20;27:e54121. [doi: 10.2196/54121]
63. Zhu J, Yang F, Wang Y, et al. Accuracy of machine learning in discriminating Kawasaki disease and other febrile illnesses: systematic review and meta-analysis. J Med Internet Res 2024 Nov 18;26:e57641. [doi: 10.2196/57641] [Medline: 39556821]
64. Becker T, Rousseau AJ, Geubbelmans M, Burzykowski T, Valkenborg D. Decision trees and random forests. Am J Orthod Dentofacial Orthop 2023 Dec;164(6):894-897. [doi: 10.1016/j.ajodo.2023.09.011] [Medline: 38008491]
65. Liu L, Li Z, Hu Y, et al. Predictive performance of machine learning for suicide in adolescents: systematic review and meta-analysis. J Med Internet Res 2025 Jun 16;27:e73052. [doi: 10.2196/73052] [Medline: 40522723]
66. Yang J, Zeng S, Cui S, Zheng J, Wang H. Predictive modeling of acute respiratory distress syndrome using machine learning: systematic review and meta-analysis. J Med Internet Res 2025 May 13;27:e66615. [doi: 10.2196/66615] [Medline: 40359510]
67. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov 2019;9(4):e1312. [doi: 10.1002/widm.1312] [Medline: 32089788]
68. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol 2019 May;20(5):e262-e273. [doi: 10.1016/S1470-2045(19)30149-4] [Medline: 31044724]

XSL•FO
RenderX

69. Parvatikar PP, Patil S, Khaparkhuntikar K, et al. Artificial intelligence: machine learning approach for screening large database and drug discovery. Antiviral Res 2023 Dec;220:105740. [doi: 10.1016/j.antiviral.2023.105740] [Medline: 37935248]
70. Tsai CA, Chang YJ. Efficient selection of Gaussian kernel SVM parameters for imbalanced data. Genes (Basel) 2023 Feb 25;14(3):583. [doi: 10.3390/genes14030583] [Medline: 36980852]
71. Park JS, Park SY, Moon JW, Kim K, Suh DI. Artificial intelligence models for pediatric lung sound analysis: systematic review and meta-analysis. J Med Internet Res 2025 Apr 18;27:e66491. [doi: 10.2196/66491] [Medline: 40249944]
72. Liao YY, Gao K, Fu BW, et al. Risk factors for electrocardiographic left ventricular hypertrophy in a young Chinese general population: the Hanzhong adolescent cohort study. BMC Cardiovasc Disord 2021 Mar 31;21(1):159. [doi: 10.1186/s12872-021-01966-y] [Medline: 33789587]
73. Tiyerili V, Müller CFH, Fung S, Panek D, Nickenig G, Becher UM. Estrogen improves vascular function via peroxisome-proliferator-activated-receptor-γ. J Mol Cell Cardiol 2012 Aug;53(2):268-276. [doi: 10.1016/j.yjmcc.2012.05.008] [Medline: 22634137]
74. Vasan RS, Song RJ, Xanthakis V, et al. Hypertension-mediated organ damage: prevalence, correlates, and prognosis in the community. Hypertension 2022 Mar;79(3):505-515. [doi: 10.1161/HYPERTENSIONAHA.121.18502] [Medline: 35138872]
75. Wang C, Yuan Y, Zheng M, et al. Association of age of onset of hypertension with cardiovascular diseases and mortality. J Am Coll Cardiol 2020 Jun 16;75(23):2921-2930. [doi: 10.1016/j.jacc.2020.04.038] [Medline: 32527401]
76. Zhou B, Li C, Shou J, Zhang Y, Wen C, Zeng C. The cumulative blood pressure load and target organ damage in patients with essential hypertension. J Clin Hypertens (Greenwich) 2020 Jun;22(6):981-990. [doi: 10.1111/jch.13875] [Medline: 32427414]
77. Carlsson AC, Ruge T, Sundström J, et al. Association between circulating endostatin, hypertension duration, and hypertensive target-organ damage. Hypertension 2013 Dec;62(6):1146-1151. [doi: 10.1161/HYPERTENSIONAHA.113.02250] [Medline: 24082055]
78. Zhang X, Li G, Zhang D, Sun Y. Influence of hypertension and global or abdominal obesity on left ventricular hypertrophy: a cross-sectional study. J Clin Hypertens (Greenwich) 2024 Dec;26(12):1342-1350. [doi: 10.1111/jch.14716] [Medline: 37794691]
79. Tsujimoto T, Kajio H. Abdominal obesity is associated with an increased risk of all-cause mortality in patients with HFpEF. J Am Coll Cardiol 2017 Dec 5;70(22):2739-2749. [doi: 10.1016/j.jacc.2017.09.1111] [Medline: 29191321]
80. Choi D, Choi S, Son JS, Oh SW, Park SM. Impact of discrepancies in general and abdominal obesity on major adverse cardiac events. J Am Heart Assoc 2019 Sep 17;8(18):e013471. [doi: 10.1161/JAHA.119.013471] [Medline: 31480883]
81. Ross R, Neeland IJ, Yamashita S, et al. Waist circumference as a vital sign in clinical practice: a Consensus Statement from the IAS and ICCR Working Group on Visceral Obesity. Nat Rev Endocrinol 2020 Mar;16(3):177-189. [doi: 10.1038/s41574-019-0310-7] [Medline: 32020062]
82. Wang Y, Howard AG, Adair LS, Wang H, Avery CL, Gordon-Larsen P. Waist circumference change is associated with blood pressure change independent of BMI change. Obesity (Silver Spring) 2020 Jan;28(1):146-153. [doi: 10.1002/oby.22638] [Medline: 31755247]
83. Wu D, Ono R, Wang S, Kobayashi Y, Sughimoto K, Liu H. Pulse wave signal-driven machine learning for identifying left ventricular enlargement in heart failure patients. Biomed Eng Online 2024 Jun 22;23(1):60. [doi: 10.1186/s12938-024-01257-5] [Medline: 38909231]
84. Lau ES, Di Achille P, Kopparapu K, et al. Deep learning-enabled assessment of left heart structure and function predicts cardiovascular outcomes. J Am Coll Cardiol 2023 Nov 14;82(20):1936-1948. [doi: 10.1016/j.jacc.2023.09.800] [Medline: 37940231]
85. Wehbe RM, Katsaggelos AK, Hammond KJ, et al. Deep learning for cardiovascular imaging: a review. JAMA Cardiol 2023 Nov 1;8(11):1089-1098. [doi: 10.1001/jamacardio.2023.3142] [Medline: 37728933]
86. Maleki F, Ovens K, Gupta R, Reinhold C, Spatz A, Forghani R. Generalizability of machine learning models: quantitative evaluation of three methodological pitfalls. Radiol Artif Intell 2023 Jan;5(1):e220028. [doi: 10.1148/ryai.220028] [Medline: 36721408]
87. Sounderajah V, Ashrafian H, Rose S, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. Nat Med 2021 Oct;27(10):1663-1665. [doi: 10.1038/s41591-021-01517-0] [Medline: 34635854]
88. Andaur Navarro CL, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ 2021 Oct 20;375:n2281. [doi: 10.1136/bmj.n2281] [Medline: 34670780]
89. Lang RM, Badano LP, Mor-Avi V, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. J Am Soc Echocardiogr 2015 Jan;28(1):1-39. [doi: 10.1016/j.echo.2014.10.003] [Medline: 25559473]
90. Wang B, Jiang B, Liu D, Zhu R. Early predictive accuracy of machine learning for hemorrhagic transformation in acute ischemic stroke: systematic review and meta-analysis. J Med Internet Res 2025 May 23;27:e71654. [doi: 10.2196/71654] [Medline: 40408765]

XSL•FO
RenderX

91. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019 Oct 29;17(1):195. [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
92. Alkhanbouli R, Matar Abdulla Almadhaani H, Alhosani F, Simsekler MCE. The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions. BMC Med Inform Decis Mak 2025 Mar 4;25(1):110. [doi: 10.1186/s12911-025-02944-6] [Medline: 40038704]
93. Aggarwal R, Farag S, Martin G, Ashrafian H, Darzi A. Patient perceptions on data sharing and applying artificial intelligence to health care data: cross-sectional survey. J Med Internet Res 2021 Aug 26;23(8):e26162. [doi: 10.2196/26162] [Medline: 34236994]
94. Murdoch B. Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Med Ethics 2021 Sep 15;22(1):122. [doi: 10.1186/s12910-021-00687-3] [Medline: 34525993]
95. Larson DB, Magnus DC, Lungren MP, Shah NH, Langlotz CP. Ethics of using and sharing clinical imaging data for artificial intelligence: a proposed framework. Radiology 2020 Jun;295(3):675-682. [doi: 10.1148/radiol.2020192536] [Medline: 32208097]

Abbreviations

AI: artificial intelligence
DL: deep learning
DOR: diagnostic odds ratio
ECG: electrocardiogram
ECHO: echocardiography
EPV: events per variable
LVH: left ventricular hypertrophy
MeSH: Medical Subject Headings
ML: machine learning
NLR: negative likelihood ratio
PLR: positive likelihood ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST: Prediction Model Risk of Bias Assessment Tool
QUADAS-AI: Quality Assessment of Diagnostic Accuracy Studies Using AI
RoB: risk of bias
SROC: summary receiver operating characteristic
SVM: support vector machine

Edited by A Coristine; submitted 28.Apr.2025; peer-reviewed by N Acharya, N Shah; revised version received 28.Jan.2026; accepted 30.Jan.2026; published 27.Feb.2026.

Please cite as:
Li Y, Zhao K, Wu J
Machine Learning in Left Ventricular Hypertrophy Detection: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e76637
URL: https://www.jmir.org/2026/1/e76637
doi:10.2196/76637

© Yilin Li, Ke Zhao, Jing Wu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Accuracy of Medical Image–Based Deep Learning for Detecting Microvascular Invasion in Hepatocellular Carcinoma: Systematic Review and Meta-Analysis

Wei Feng[1*], MM; Bo Qu[2*], MM; Shuo Han[3], PhD

[1]Department of Ultrasound, the Fourth Affiliated Hospital, China Medical University, Shenyang, China

[2]Department of Urology, Jinqiu Hospital of Liaoning Province, Shenyang, China

[3]Department of Cardiology, the Fourth Affiliated Hospital, China Medical University, No. 4, Chongshan East Road, Huanggu District, Shenyang, China

[*]these authors contributed equally

Corresponding Author:
Shuo Han, PhD
Department of Cardiology, the Fourth Affiliated Hospital, China Medical University, No. 4, Chongshan East Road, Huanggu District, Shenyang, China

Abstract

Background: Hepatocellular carcinoma (HCC) is a leading cause of cancer-related mortality worldwide. Microvascular invasion (MVI) is a critical pathological indicator of postoperative recurrence and poor prognosis in patients with HCC. Some researchers have explored the diagnostic accuracy of deep learning (DL) based on various imaging modalities for MVI.

Objective: This meta-analysis aimed to systematically evaluate the preoperative diagnostic performance of DL models using medical images to predict MVI in HCC, and to investigate the impact of different imaging modalities and validation strategies on model performance and generalizability.

Methods: PubMed, Cochrane Library, Embase, and Web of Science were searched up to October 16, 2025. Studies investigating the detection of MVI in HCC using imaging-based DL techniques were eligible. Studies focusing solely on image segmentation were excluded. The Quality Assessment of Diagnostic Accuracy Studies-2 tool was used to assess risk of bias. A bivariate mixed-effects meta-analysis was performed to calculate the pooled sensitivity, specificity, and area under the summary receiver operating characteristic curve (SROC). Subgroup analyses were conducted by imaging modality and validation set generation method.

Results: This meta-analysis included 52 studies with 19,531 patients with HCC. The pooled analysis revealed that imaging-based DL models had an overall sensitivity of 0.80 (95% CI 0.78‐0.83), a specificity of 0.82 (95% CI 0.80‐0.85), and an SROC of 0.88 for MVI prediction. Subgroup analysis showed that models based on preoperative contrast-enhanced computed tomography performed excellently, with a sensitivity of 0.84 (95% CI 0.79‐0.88), a specificity of 0.83 (95% CI 0.77‐0.88), and an SROC of 0.90. These results suggest that contrast-enhanced computed tomography is the most promising noninvasive method for current clinical applications. Meanwhile, DL models using pathological sections achieved the highest diagnostic performance: a sensitivity of 0.91 (95% CI 0.87‐0.94), a specificity of 0.90 (95% CI 0.68‐0.97), and an SROC of 0.92. This establishes the ultimate benchmark for performance optimization for all noninvasive models. A key finding was that model performance was less consistent in independent external validation (SROC: 0.85) than in internal validation (SROC: 0.90). This discrepancy indicates that overreliance on internal validation may overestimate model efficacy and underscores the decisive role of rigorous external validation in assessing real-world generalizability.

Conclusions: This study is the first to systematically assess the use of imaging-based DL for diagnosing MVI in HCC. The results demonstrate a significant potential for these models in predicting MVI. However, their clinical applicability requires rigorous evaluation, given the scarcity of independent external validation cohorts, notable heterogeneity among them, and the observed decline in model performance. Therefore, prospective, multicenter studies following standardized reporting guidelines are a critical future direction. These studies should also focus on developing integrated algorithms that translate histopathological insights into preoperative imaging data to establish robust clinical tools.

Trial Registration: PROSPERO CRD42024613733; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024613733

(J Med Internet Res 2026;28:e82000) doi:10.2196/82000

KEYWORDS

deep learning; hepatocellular carcinoma; medical imaging; microvascular invasion; artificial intelligence

XSL•FO
RenderX

Introduction

Hepatocellular carcinoma (HCC) is the most common pathological subtype of primary liver cancer, accounting for about 90% of cases [1]. Globally, HCC is the fourth leading cause of cancer-related deaths [2]. According to recent epidemiological data, the age-standardized incidence and mortality rates of HCC are highest in Africa and the Western Pacific region. Over 70% of global HCC cases occur in Asia [1]. Despite advancements in treatment modalities for HCC, including liver transplantation, surgical resection, transarterial chemoembolization, local ablation, targeted therapy, and immunotherapy, the 5-year relative survival rate remains below 20% [3]. Even after complete surgical tumor removal, around 50% - 70% of patients with HCC experience tumor recurrence within 5 years postsurgery [4]. Consequently, HCC has become a significant oncological burden that threatens human life.

Microvascular invasion (MVI) is the pathological process by which tumor cells invade the microvascular structures of the liver tissue surrounding an HCC lesion. MVI occurs in approximately 30% - 50% of cases. It is a significant factor in HCC recurrence after surgery and is associated with poor prognoses in patients with HCC [4,5]. Studies have shown that individuals with HCC and MVI-positive status have significantly lower 5-year disease-free survival and overall survival rates than those with MVI-negative status [6, 7]. Notably, MVI status directly influences treatment strategy selection. Wide-range hepatectomy (resection of ≥ 3 liver segments) is recommended for individuals at high preoperative risk for MVI. This method has a significantly lower 5-year cumulative recurrence rate than limited resection (26.6% vs 58.3%, P=.040) [8]. However, other studies have found that adjuvant hepatic arterial infusion chemotherapy does not significantly improve the survival of high-risk MVI-positive individuals compared with the untreated group (P=.61). Nevertheless, hepatic arterial infusion chemotherapy significantly improves the prognosis of low-risk patients with MVI (P<.001) [9]. Furthermore, individuals with MVI undergoing radiofrequency ablation have a significantly higher recurrence risk than those undergoing radical surgery (P<.05) [10]. Therefore, accurately identifying MVI preoperatively is significant for formulating individualized, comprehensive treatment regimens and improving patient prognosis [11].

Currently, a definitive diagnosis of MVI relies on a postoperative pathological examination. However, this examination is subject to biases resulting from the quality of slide preparation and interobserver heterogeneity. These factors may lead to diagnostic inaccuracies. Furthermore, the absence of preoperative MVI information restricts its use in personalized treatment decisions. Therefore, developing efficient MVI auxiliary detection tools is crucial for optimizing clinical management of HCC.

While imaging examinations are crucial for evaluating MVI, predictions based on traditional imaging features rely heavily on radiologists' subjective interpretations. A systematic review and meta-analysis of 19 studies involving 1920 patients revealed that traditional contrast-enhanced features on magnetic resonance imaging (MRI) showed poor overall diagnostic performance in predicting MVI. Only peritumoral enhancement in the arterial phase exhibited moderate diagnostic accuracy. The combined efficacy of other features, such as peritumoral hypointensity in the hepatobiliary phase and irregular margins, was insufficient to meet the requirements for precise preoperative clinical assessment [12]. Recent progress in data mining techniques has accelerated the growth of radiomics. This technique assists in the analysis of imaging features (eg, shape, intensity, and texture) that are difficult for the human eye to perceive. It can overcome some of the limitations of subjectivity. However, radiomic features are mostly low- or mid-level and susceptible to noise interference. They may also not fully reflect tumor heterogeneity [13,14]. Zhang et al [15] noted that radiomics based on single-modality medical imaging is inherently limited. Due to constraints in imaging principles, such approaches can only reflect partial tumor information. Furthermore, when features are extracted using the entire tumor as the region of interest, information about intratumoral heterogeneity is inevitably lost. Additionally, these single-modality radiomic features are susceptible to image noise and variations in scanning parameters, which further compromise the model's ability to capture tumor heterogeneity comprehensively. In contrast, deep learning (DL) uses multi-layer neural networks and an end-to-end learning mode to directly extract multi-level abstract high-order features from original images. This improves the predictive performance, interpretability, and generalizability of models. DL is expected to provide a new paradigm for the preoperative, noninvasive assessment of MVI [16,17]. However, existing studies often focus on a single imaging modality or have small sample sizes. These studies lack a systematic comparison of DL model performance across different imaging modalities, which limits the interpretation of the advantages of DL in detecting MVI and poses challenges to the development or update of intelligent auxiliary diagnostic tools. Consequently, this meta-analysis was conducted to systematically evaluate the diagnostic efficacy of DL models based on medical images for MVI, as well as to explore the impact of different imaging modalities, validation strategies, and sources of heterogeneity on model performance and generalizability, to provide evidence-based support for the development or update of future intelligent auxiliary diagnostic tools.

Methods

Study Registration

This meta-analysis was prospectively registered with the PROSPERO (CRD42024613733). This systematic review and meta-analysis of diagnostic test accuracy was reported in accordance with the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Diagnostic Test Accuracy studies) [18] guidelines in Checklist 1. Due to the absence of subject information collection and its lack of impact on clinical diagnosis and treatment, ethical approval and informed consent were waived.

XSL•FO
RenderX

Eligibility Criteria

Inclusion Criteria

The inclusion criteria are as follows:

- Original research with full text published in English (including cohort, case-control, and cross-sectional studies).
- MVI status in HCC individuals had to be confirmed by histopathology or biopsy.
- Studies had to develop complete DL models based on medical images to detect MVI status in patients with HCC.
- English-language studies.

Exclusion Criteria

The exclusion criteria are as follows:

- Meta-analyses, reviews, guidelines, or expert opinions.
- Only differential factor analysis was implemented without a comprehensive DL model.
- Studies lacking outcome measures of predictive accuracy for the machine learning model, including sensitivity, C-index, accuracy, specificity, precision, F_1-score, and confusion matrix.
- Only image segmentation was performed.

Search Strategies and Data Sources

We conducted a systematic literature search in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for literature searches; completed checklist is available in Checklist 2) [19]. Relevant English-language publications were retrieved from PubMed, Web of Science, the Cochrane Library, and Embase, with the search covering all records up to October 16, 2025. The search strategy used both MeSH (Medical Subject Headings) and free-text keywords, including MeSH terms such as “Carcinoma, Hepatocellular,” “liver cell carcinoma,” and “deep learning.” Boolean operators were used to integrate MeSH terms and free-text terms, constructing tailored search queries for each database. Furthermore, the reference lists of identified review articles were manually screened to locate any additional eligible studies. No prospective study registries were searched, and no attempts were made to obtain unpublished data or to contact study authors for further information. The complete search strings for each database are provided in Table S1 in Multimedia Appendix 1.

Study Selection

Retrieved articles were imported into EndNote. After removing duplicates, the remaining articles were reviewed by title and abstract to identify initially eligible articles. Then, the full texts were downloaded and screened to determine the final eligible articles. Two researchers (WF and BQ, with 6 and 4 years of meta-analysis experience, respectively) performed the review independently. Interresearcher agreement during the literature screening was assessed using the κ coefficient (κ=0.93). Any disagreements were resolved in a consensus meeting with a third researcher (SH, with 10 years of experience in meta-analysis).

Data Extraction

Before data extraction, a standardized spreadsheet was generated. The content to be extracted included the following: publication year, patient source, author, image source, manual segmentation, number of patients with MVI, total number of patients with HCC, number of patients with MVI in the training set, number of patients with HCC in the training set, validation set generation method, number of patients with MVI in the validation set, number of patients with HCC in the validation set, confusion matrix, sensitivity, specificity, precision, and accuracy.

Two researchers (WF and BQ, with 14 and 16 years of medical experience, respectively) carried out the data extraction independently. They then cross-checked their results. Any discrepancies were resolved through consultation with a third researcher (SH, with 13 years of medical experience).

Risk of Bias in Studies

The Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool was used to evaluate the overall risk of bias (ROB) and applicability of the eligible studies. The QUADAS-2 instrument encompasses 4 domains: index test, patient selection, flow and timing, and reference standard. Each domain includes specific questions. Answers to these questions are categorized as “yes,” “no,” or “unclear.” These answers correspond to ROB ratings of “unclear,” “high,” or “low.” If all key questions within a domain received a “yes” answer, the domain was rated as having a low ROB. If any key question received a “no” answer, a potential ROB was indicated, and the researcher judged the ROB according to established guidelines. An “unclear” rating was assigned when the literature did not provide sufficient detail for the researcher to make a judgment.

Two researchers completed the QUADAS-2 assessment independently. Any discrepancies were resolved through discussion with a third researcher.

Synthesis Methods

A meta-analysis was performed using a bivariate mixed-effects model based on diagnostic 2×2 tables [20]. For studies that did not report these tables directly, we derived them from the available specificity, sensitivity, positive predictive value, accuracy, F_1-score, and case numbers. Throughout the analysis, the explicitly defined independent validation cohort from each study served as the unit of analysis. Each data point corresponded to distinct and nonoverlapping patient samples, which ensured the independence of the pooled results and mitigated potential bias from data reuse. Using the bivariate mixed-effects model, we computed the pooled estimates for specificity, sensitivity, negative likelihood ratio (LR–), positive likelihood ratio (LR+), diagnostic odds ratio (DOR), and area under the summary receiver operating characteristic (SROC) curve, along with their 95% CIs [20]. The Spearman correlation coefficient was used to evaluate the threshold effect and its contribution to between-study heterogeneity. Small-study effects were assessed using Deeks funnel plot asymmetry test. For subgroups with fewer than 10 studies, a Doi plot was used to informally assess publication bias. The degree of bias was determined based on the absolute value of the Luis

XSL•FO
RenderX

Furuya-Kanamori (LFK) index. A value less than 1 suggests minor publication bias, a value between 1 and 2 indicates moderate publication bias, and a value exceeding 2 suggests substantial publication bias. During the meta-analysis, the validation set was used, and subgroup analyses were performed according to the validation set generation method and image type. A *P* value less than .05 was considered statistically significant.

Results

Study Selection

Database retrieval yielded 11,266 articles, of which 7539 remained after duplicate removal. Following title and abstract screening, 62 articles were selected for a full-text review. The full-text review subsequently excluded 10 records: 2 non-DL studies, 3 studies with insufficient data to construct diagnostic 2×2 tables, and 5 studies that used DL solely for medical image segmentation without establishing HCC MVI prediction models. Finally, 52 articles [14,16,21-70] met the eligibility criteria. The detailed process is illustrated in Figure 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the study screening and selection process for this meta-analysis.

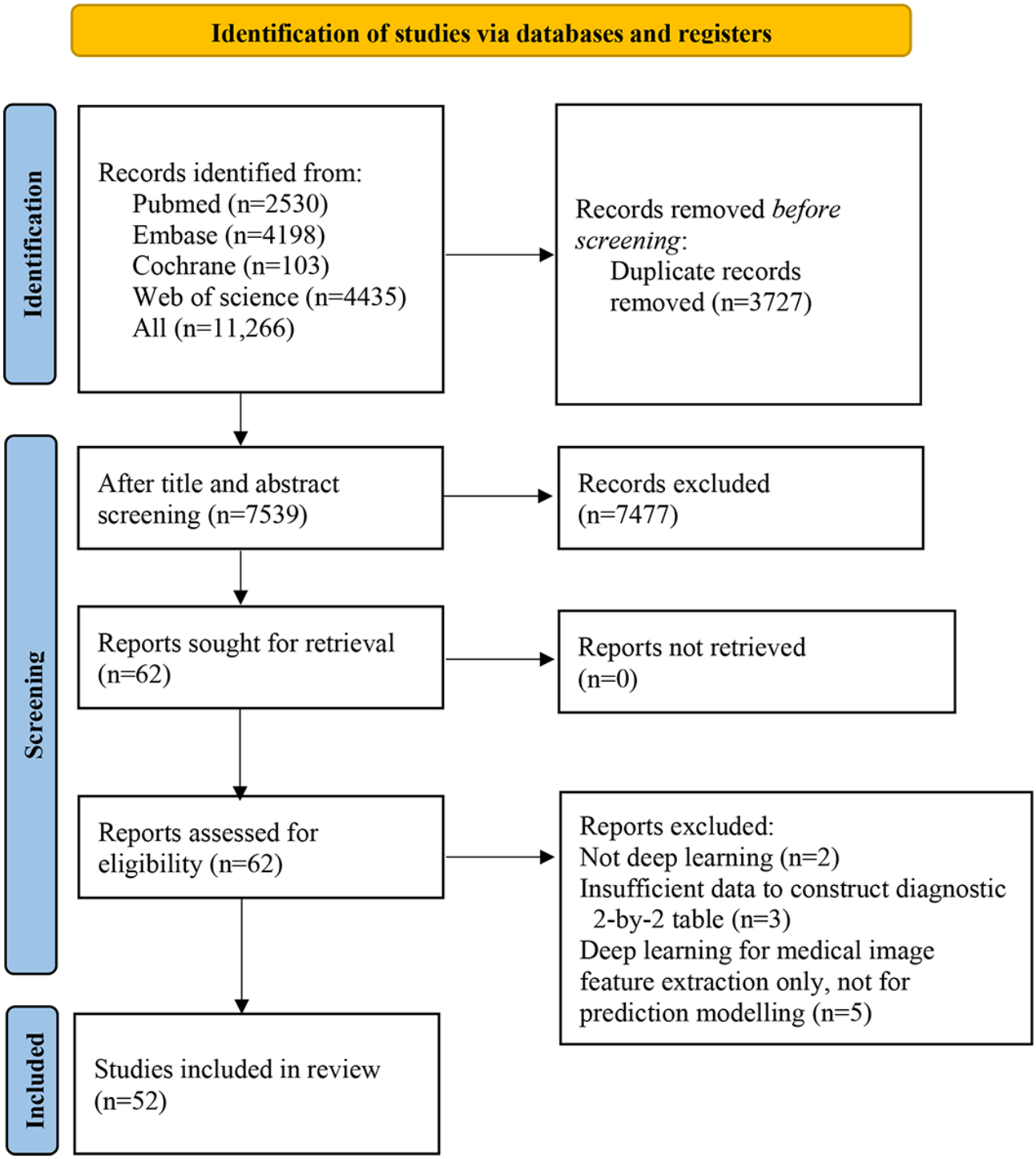

Study Characteristics

This meta-analysis included 52 studies published between 2019 and 2025. All of the studies used histopathological diagnosis as the gold standard for MVI. All 52 studies were case-control studies, encompassing 19,531 individuals with HCC, of whom 8161 were MVI cases. Regarding population source, 32 (61.5%) of the studies were single-center, 19 (36.5%) studies were multicenter, and 1 (1.9%) study was based on a registry database.

These studies primarily used single-modality imaging techniques for image modeling methods: contrast-enhanced computed tomography (CECT, n=16), contrast-enhanced magnetic resonance imaging (CEMRI, n=19), MRI (n=5), contrast-enhanced ultrasound (CEUS, n=5), and pathological sections (n=2). Five additional studies integrated multimodal imaging (computed tomography + positron emission tomography: 1, CECT +CEMRI: 4) for modeling. Regarding image segmentation methods, 38 of the 52 (73.1%) studies used manual segmentation, while the remaining studies used fully automatic segmentation (n=5) or semiautomatic segmentation (n=9).

XSL•FO
RenderX

Concerning model validation strategies, 23 studies used randomly sampled internal validation, 15 used performance through cross-validation, and 14 used external validation with independent cohorts (Table 1).

XSL•FO
RenderX

Table . Basic characteristics of the 52 eligible studies evaluating medical image-based deep learning models for preoperatively detecting microvascular invasion (MVI) in patients with hepatocellular carcinoma.

Author (year of publication)	Basic characteristics						Training versus validation cohort characteristics				
	Source of the patients[a]	Imaging modality	Segmentation	Diagnostic criteria for MVI	Patients with MVI	Sample size	Patients with MVI in the training set	Sample size in the training set	Type of validation	Patients with MVI in the validation set	Sample size in the validation set
Zhang et al (2024)[a] [21]	Single center +registration database (TCGA)[b]	Pathological section	Semiautomatic segmentation	Pathological diagnosis	576	1111	328	530	Random sampling (7:3), external validation (registration database TCGA)	Validation set 1: 143; Validation set 2: 105	Validation set 1: 223; Validation set 2: 358
Lei et al (2024) [22]	Multicenter	CECT[c]+CEMRI[d]	Semiautomatic segmentation	Pathological diagnosis	104	345	89	301	Random sampling (6:4)	15	44
Liu et al (2024) [23]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	123	265	97	211	Random sampling (8:2)	26	54
Zhou et al (2024) [24]	Multicenter	CECT	Semiautomatic segmentation	Pathological diagnosis	54	140	38	98	Random sampling (7:3)	16	42
He et al (2024) [25]	Multicenter	CECT	Semiautomatic segmentation	Pathological diagnosis	299	640	172	368	Internal validation	Validation set 1: 63; Validation set 2: 64	Validation set 1: 134; Validation set 2: 138
Zhong et al (2024) [26]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	82	173	57	120	Random sampling (7:3)	25	53
Wang et al (2024)[a] [27]	Multicenter	CEMRI	Manual segmentation	Pathological diagnosis	274	725	109	234	Random sampling (4:1); External validation (multicenter)	Validation set 1: 20; Validation set 2.1: 82; Validation set 2.2: 37; Validation set 2.3: 26	Validation set 1: 58; Validation set 2.1: 212; Validation set 2.2: 111; Validation set 2.3: 110
Yu et al (2024)[a] [28]	Multicenter	CECT	Semiautomatic segmentation	Pathological diagnosis	78	205	39	119	Random sampling (7:3); External validation (multicenter)	Validation set 1: 15; Validation set 2: 24	Validation set 1: 44; Validation set 2: 42
Ma et al (2025) [29]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	52	117	42	94	Random sampling (8:2); 10-fold cross-validation	10	23

XSL•FO
RenderX

Author (year of publication)	Basic characteristics						Training versus validation cohort characteristics				
	Source of the patients[a]	Imaging modality	Segmentation	Diagnostic criteria for MVI	Patients with MVI	Sample size	Patients with MVI in the training set	Sample size in the training set	Type of validation	Patients with MVI in the validation set	Sample size in the validation set
Zhang et al (2024)[a] [30]	Multicenter	CEUS[e]	Manual segmentation	Pathological diagnosis	219	576	175	461	Single center random sampling (8:2); External validation (multicenter)	—[f]	—
Wang et al (2023)[a] [31]	Multicenter	CECT+CEMRI	Manual segmentation	Pathological diagnosis	150	397	119	297	External validation (multicenter)	31	100
You et al (2023) [32]	Single center	CEMRI	Semiautomatic segmentation	Pathological diagnosis	70	210	56	168	Random sampling (4:1); 5-fold cross-validation	14	42
Qin et al (2023) [33]	Multicenter	CEUS	Manual segmentation	Pathological diagnosis	92	252	71	198	Random sampling (8:2)	21	54
Li et al (2023) [34]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	146	283	117	226	Random sampling (4:1); 5-fold cross-validation	29	57
Cao et al (2023) [35]	Single center	CECT	Manual segmentation	Pathological diagnosis	149	559	120	448	Random sampling (4:1)	29	111
Wang et al (2023) [36]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	109	233	76	163	Random sampling (7:3)	33	70
Ye et al (2023) [37]	Multicenter	CECT+PET[g]	Manual segmentation	Pathological diagnosis	41	100	29	70	Random sampling (7:3); 5-fold cross-validation	12	30
Xu et al (2023) [38]	Single center	CECT	Manual segmentation	Pathological diagnosis	99	305	79	244	Random sampling (8:2); 5-fold cross-validation	20	61
Deng et al (2022) [39]	Single center	CECT+CEMRI	Manual segmentation	Pathological diagnosis	44	103	35	82	Random sampling (4:1); 5-fold cross-validation	9	21
Li et al (2022) [40]	Multicenter	CECT	Manual segmentation	Pathological diagnosis	433	1116	346	892	Random sampling (4:1)	87	224

XSL•FO
RenderX

Author (year of publication)	Basic characteristics						Training versus validation cohort characteristics				
	Source of the patients[a]	Imaging modality	Segmentation	Diagnostic criteria for MVI	Patients with MVI	Sample size	Patients with MVI in the training set	Sample size in the training set	Type of validation	Patients with MVI in the validation set	Sample size in the validation set
Chen et al (2022)[a] [41]	Multicenter	Pathological section	Automatic segmentation	Pathological diagnosis	224	470	137	270	Random sampling external validation (multicenter)	Validation set 1: 43; Validation set 2: 44	Validation set 1: 80; Validation set 2: 120
Zhang et al (2022) [42]	Single center	CEUS	Manual segmentation	Pathological diagnosis	150	436	103	301	Random sampling (3:1); Internal validation	Validation set 1: 35; Validation set 2: 12	Validation set 1: 02; Validation set 2: 33
Liu et al (2022) [43]	Single center	MRI[h]	Manual segmentation	Pathological diagnosis	43	114	28	74	Random sampling	15	40
Sun et al (2022) [44]	Single center	CECT	Semiautomatic segmentation	Pathological diagnosis	134	358	77	193	Random sampling	Validation set 1: 23; Validation set 2: 34	Validation set 1: 61; Validation set 2: 104
Wang et al (2022) [45]	Single center	CECT	Automatic segmentation	Pathological diagnosis	68	138	54	110	Random sampling (8:1:1); 5-fold cross-validation	Validation set 1: 7; Validation set 2: 7	Validation set 1: 14; Validation set 2: 14
Xiao et al (2022)[a] [46]	Multicenter	CECT	Automatic segmentation	Pathological diagnosis	1103	2096	458	876	Random sampling (3:1); External validation (multicenter)	Validation set 1: 152; Validation set 2: 327; Validation set 3: 166	Validation set 1: 292; Validation set 2: 578; Validation set 3: 350
Yang et al (2022) [47]	Single center	CECT	Manual segmentation	Pathological diagnosis	36	283	25	198	Random sampling (198:85)	11	85
Sun et al (2022) [48]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	185	321	86	149	Internal validation	99	172
Dai et al (2022) [49]	Single center	CECT	Manual segmentation	Pathological diagnosis	215	400	172	320	Random sampling (80:10:10)	Validation set 1: 21; Validation set 2: 22	Validation set 1: 40; Validation set 2: 40
Zhang et al (2021) [50]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	92	237	61	158	Random sampling	31	79
Liu et al (2021)[a] [51]	Multicenter	CECT	Manual segmentation	Pathological diagnosis	135	473	68	216	Random sampling (70:30); External validation (multicenter)	Validation set 1: 28; validation set 2: 39	Validation set 1: 93; Validation set 2: 164

XSL•FO
RenderX

Author (year of publication)	Basic characteristics						Training versus validation cohort characteristics				
	Source of the patients[a]	Imaging modality	Segmentation	Diagnostic criteria for MVI	Patients with MVI	Sample size	Patients with MVI in the training set	Sample size in the training set	Type of validation	Patients with MVI in the validation set	Sample size in the validation set
Wei et al (2021)[a] [52]	Multicenter	CECT + CEMRI	Manual segmentation	Pathological diagnosis	270	750	216	635	External validation (prospective, multicenter)	54	115
Zhou et al (2021) [53]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	—	114	—	—	Random sampling	—	—
Wang et al (2021) [54]	Single center	MRI	Manual segmentation	Pathological diagnosis	43	100	24	60	Random sampling	19	40
Zeng et al (2021) [55]	Single center	MRI	Manual segmentation	Pathological diagnosis	38	98	25	64	Random sampling; 4-fold cross-validation	13	34
Gao et al (2021) [56]	Single center	MRI	Manual segmentation	Pathological diagnosis	74	225	56	168	Random sampling	18	57
Jiang et al (2020) [57]	Single center	CECT	Manual segmentation	Pathological diagnosis	220	405	176	324	Random sampling (8:2)	44	81
Song et al (2021) [58]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	225	601	174	461	Random sampling	51	140
Men et al (2019) [59]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	28	63	21	47	4-fold cross-validation	7	16
Zhou et al (2022) [60]	Single center	CECT	Manual segmentation	Pathological diagnosis	145	466	97	311	3-fold cross-validation	48	155
Chu et al (2022) [61]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	51	133	35	93	Random sampling (7:3)	16	40
Huang et al (2022) [62]	Single center	MRI	Manual segmentation	Pathological diagnosis	43	114	32	86	4-fold cross-validation	11	28
Wang et al (2025) [63]	Single center	CEUS	Semiautomatic segmentation	Pathological diagnosis	142	318	99	222	Random sampling; 5-fold cross-validation	43	96
Cen et al (2025) [64]	Single center	CECT	Semiautomatic segmentation	Pathological diagnosis	68	192	47	134	Random sampling (7:3)	21	58
Huang et al (2025) [65]	Multicenter	CEMRI	Manual segmentation	Pathological diagnosis	124	300	87	210	Random sampling; 5-fold cross-validation	37	90

XSL•FO
RenderX

Author (year of publication)	Basic characteristics						Training versus validation cohort characteristics				
	Source of the patients[a]	Imaging modality	Segmentation	Diagnostic criteria for MVI	Patients with MVI	Sample size	Patients with MVI in the training set	Sample size in the training set	Type of validation	Patients with MVI in the validation set	Sample size in the validation set
Miao et al (2025)[a] [66]	Multicenter	CECT	Manual segmentation	Pathological diagnosis	206	483	136	311	Random sampling (8:2); External validation (multicenter)	Validation set 1: 32; Validation set 2: 38	Validation set 1: 77; Validation set 2: 95
Zhu et al (2025)[a] [14]	Multicenter	CEMRI	Manual segmentation	Pathological diagnosis	120	304	90	216	External validation (multicenter, retrospective)	30	88
Dong et al (2025)[a] [67]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	188	519	100	263	Random sampling (4:1); External validation (multicenter)	Validation set 1: 26; Validation set 2: 27; Validation set 3: 35	Validation set 1: 66; Validation set 2: 93; Validation set 3: 97
Zhang et al (2025) [68]	Single center	CEMRI	Automatic segmentation	Pathological diagnosis	142	270	114	216	5-fold cross-validation	28	54
Zheng et al (2025)[a] [16]	Multicenter	CEMRI	Automatic segmentation	Pathological diagnosis	292	589	154	317	Random sampling (7:3); External validation (multicenter)	Validation set 1: 52; Validation set 2: 86	Validation set 1: 106; Validation set 2: 166
Zhao et al (2025)[a] [69]	Multicenter	CEMRI	Manual segmentation	Pathological diagnosis	51	145	25	66	10-fold cross-validation; External validation (multicenter)	26	79
Qin et al (2025) [70]	Single center	CEUS	Manual segmentation	Pathological diagnosis	65	164	44	114	Random sampling (7:3); 10-fold cross-validation	21	50

[a]Studies used external validation with an independent cohort.

[b]TCGA: The Cancer Genome Atlas.

[c]CECT: contrast-enhanced computed tomography.

[d]CEMRI: contrast-enhanced magnetic resonance imaging.

[e]CEUS: contrast-enhanced ultrasound.

[f]Not available.

[g]PET: positron emission tomography.

[h]MRI: magnetic resonance imaging.

XSL•FO
RenderX

ROB in Studies

Regarding patient selection, all eligible studies included consecutive or random cases. According to the QUADAS-2 assessment criteria, this study design carries an inherent high ROB in the "Patient Selection" domain. Consequently, all included studies received a "high" rating for ROB in this domain.

Due to the adoption of supervised DL, model training was based on clear pathological outcomes. However, since DL models predict by extracting inherent image features, rather than directly relying on clinical covariates, their training process's dependence on known outcomes did not result in diagnostic information leakage. This led to a low ROB. Manual segmentation was used in 38 studies, which could have introduced operator subjectivity and led to a high ROB.

Regarding the implementation of the gold standard, all studies used histopathological diagnosis as the gold standard for MVI, ensuring the objectivity and consistency of disease classification. This indicated a low ROB in the implementation of the gold standard.

Regarding the item "the match between the conduct and interpretation of the index test and the review question" in the QUADAS-2 scale, 5 (9.6%) out of the 52 studies did not directly report specificity values. Since specificity is a key indicator for verifying the match between a DL model and a clinical question, the absence of such data may lead to an incomplete assessment of a model's diagnostic efficacy and weaken the reliability of a study's conclusions. Therefore, these 5 studies were determined to have a high ROB. The remaining 47 (90.4%) studies fully reported diagnostic performance indicators, ensuring the transparency in test conduct and interpretation, and indicating a low ROB.

There was a reasonable and appropriate time interval between imaging examinations and pathological diagnoses in all eligible studies. Therefore, there did not appear to be a significant impact on the cases' process (Figure 2).

Figure 2. Methodological quality assessment of the included studies based on the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) scale. (**A**) Summary, (**B**) Individual studies [14,16,21-70].

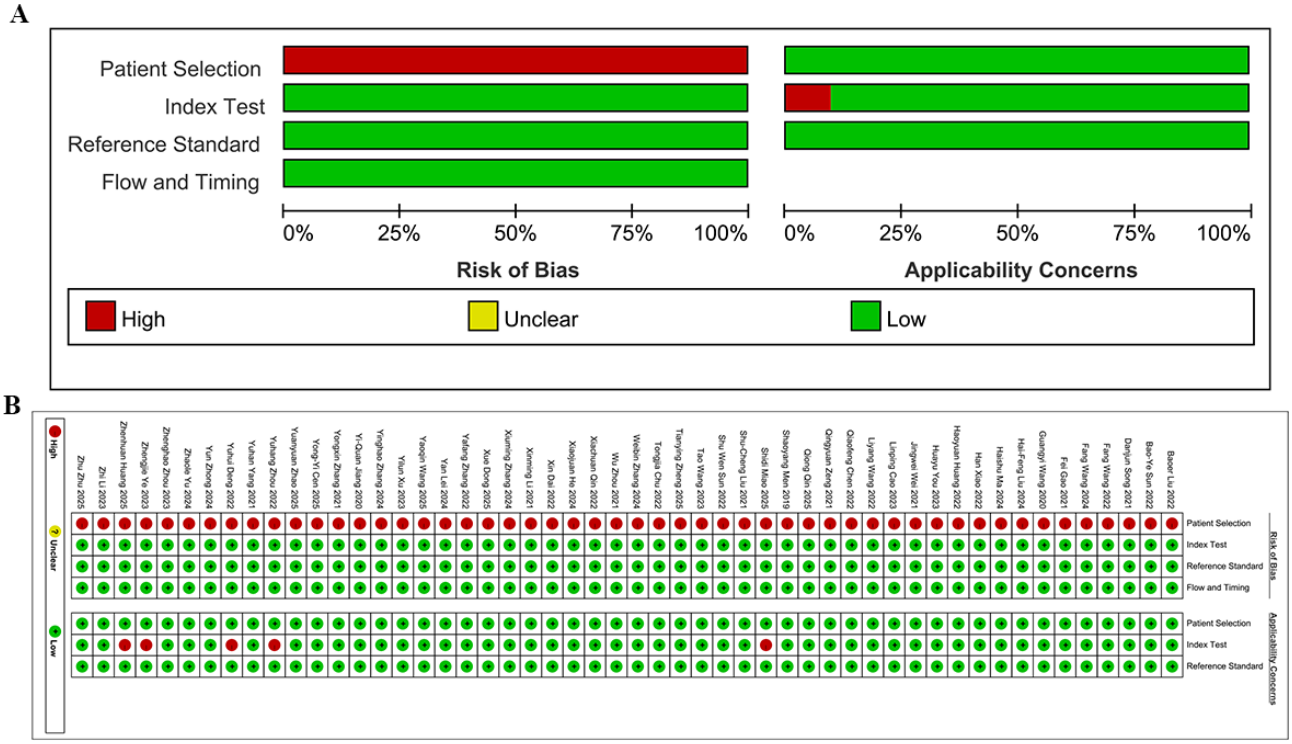

Meta-Analysis

Overall

The model's accuracy was validated using 68 diagnostic fourfold tables. The Spearman correlation coefficient was 0.02, indicating a minimal threshold effect. This effect accounted for none of the observed between-study heterogeneity. The pooled analysis revealed the following results: sensitivity 0.80 (95% CI 0.78 - 0.83, I^2=65.52%), specificity 0.82 (95% CI 0.80 - 0.85, I^2=79.13%), LR+4.6 (95% CI 3.9 - 5.3), LR- 0.24 (95% CI 0.21 - 0.27), DOR 19 (95% CI 15 - 25), and SROC 0.88 (95% CI 0.56 - 0.98; Table 2 and Figures 3 and 4).

XSL•FO
RenderX

Table . Meta-analysis results of deep learning for microvascular invasion diagnosis under different image sources and validation set generation methods.

Subgroup	N	SENS[a] (95% CI)	SPEC[b] (95% CI)	PLR[c] (95% CI)	NLR[d] (95% CI)	DOR[e] (95% CI)	SROC[f] (95% CI)	Deeks	Doi
Overall	68	0.80 (0.78-0.83)	0.82 (0.80-0.85)	4.6 (3.9-5.3)	0.24 (0.21-0.27)	19 (15-25)	0.88 (0.56-0.98)	0.77	—[g]
Validation set generation methods									
Internal validation	49	0.82 (0.79-0.85)	0.83 (0.80-0.86)	4.9 (4.1-6.0)	0.22 (0.18-0.26)	23 (17-32)	0.90 (1.00 - 0.00)	0.85	—
External validation	19	0.77 (0.72-0.82)	0.80 (0.74-0.85)	3.9 (3.0-5.0)	0.29 (0.23-0.36)	13 (9-20)	0.85 (1.00 - 0.00)	0.22	—
The number of cases in the validation set									
<100	38	0.79 (0.75-0.83)	0.83 (0.79-0.87)	4.7 (3.8-5.8)	0.25 (0.20-0.30)	19 (14-26)	0.88 (0.63 - 0.97)	0.15	—
≥100	30	0.82 (0.78-0.85)	0.81 (0.77-0.85)	4.3 (3.4-5.5)	0.23 (0.18-0.28)	19 (13-29)	0.88 (0.66 - 0.97)	0.5	—
Image sources									
CECT[h]	20	0.84 (0.79-0.88)	0.83 (0.77-0.88)	5.0 (3.6-6.9)	0.19 (0.15-0.25)	26 (16-42)	0.90 (1.00 - 0.00)	0.09	
CECT internal validation	16	0.86 (0.80-0.90)	0.85 (0.78-0.90)	5.7 (3.8-8.4)	0.17 (0.12-0.24)	33 (19-59)	0.92 (1.00 - 0.00)	0.19	—
CECT external validation	4	0.82 (0.73-0.88)	0.76 (0.64-0.85)	3.4 (2.3-5.0)	0.24 (0.17-0.34)	14 (9-23)	0.86 (1.00 - 0.00)	—	−1.14
CEUS[i]	5	0.70 (0.58-0.80)	0.88 (0.82-0.92)	5.6 (3.5-9.0)	0.34 (0.23-0.51)	17 (7-37)	0.89 (1.00 - 0.00)	—	0.46
CEMRI[j]	24	0.78 (0.73-0.83)	0.81 (0.76-0.85)	4.0 (3.2-5.0)	0.27 (0.22-0.33)	15 (10-22)	0.86 (1.00 - 0.00)	0.15	
CEMRI internal validation	16	0.81 (0.75-0.86)	0.81 (0.75-0.85)	4.3 (3.2-5.6)	0.24 (0.17-0.32)	18 (11-30)	0.88 (1.00 - 0.00)	0.17	—
CEMRI external validation	8	0.74 (0.65-0.82)	0.80 (0.71-0.87)	3.7 (2.7-5.2)	0.32 (0.24-0.43)	12 (7-18)	0.84 (1.00 - 0.00)	—	1.74
MRI[k]	5	0.76 (0.68-0.82)	0.80 (0.70-0.87)	3.7 (2.5-5.5)	0.30 (0.23-0.40)	12 (7-21)	0.84 (1.00 - 0.00)	—	3.11
Multimodal	10	0.74 (0.68-0.80)	0.79 (0.75-0.83)	3.6 (2.9-4.4)	0.33 (0.26-0.41)	11 (7-16)	0.82 (1.00 - 0.00)	0.44	
Multimodal internal validation	6	0.79 (0.69-0.87)	0.79 (0.73-0.85)	3.8 (2.8-5.3)	0.26 (0.17-0.41)	15 (7-29)	0.86 (1.00 - 0.00)	—	1.6
Multimodal external validation	4	0.71 (0.64-0.78)	0.78 (0.73-0.83)	3.3 (2.6-4.2)	0.36 (0.28-0.47)	9 (6-14)	0.82 (1.00 - 0.00)	—	1.93
Pathological Sections	4	0.91 (0.87-0.94)	0.90 (0.68-0.97)	9.2 (2.5-33.6)	0.09 (0.06-0.15)	97 (20-465)	0.92 (1.00 - 0.00)	—	3.08

[a]SENS: sensitivity.

[b]SPEC: specificity.

[c]PLR: positive likelihood ratio.

[d]NLR: negative likelihood ratio.

[e]DOR: diagnostic odds ratio.

[f]SROC: summary receiver operating characteristic.

XSL•FO
RenderX

[g]Not applicable.

[h]CECT: contrast-enhanced computed tomography.

[i]CEUS: contrast-enhanced ultrasound.

[j]CEMRI: contrast-enhanced magnetic resonance imaging.

[k]MRI: magnetic resonance imaging.

Figure 3. Meta-analysis forest plots: specificity and sensitivity of image-based deep learning in microvascular invasion diagnosis [14,16,21-70].

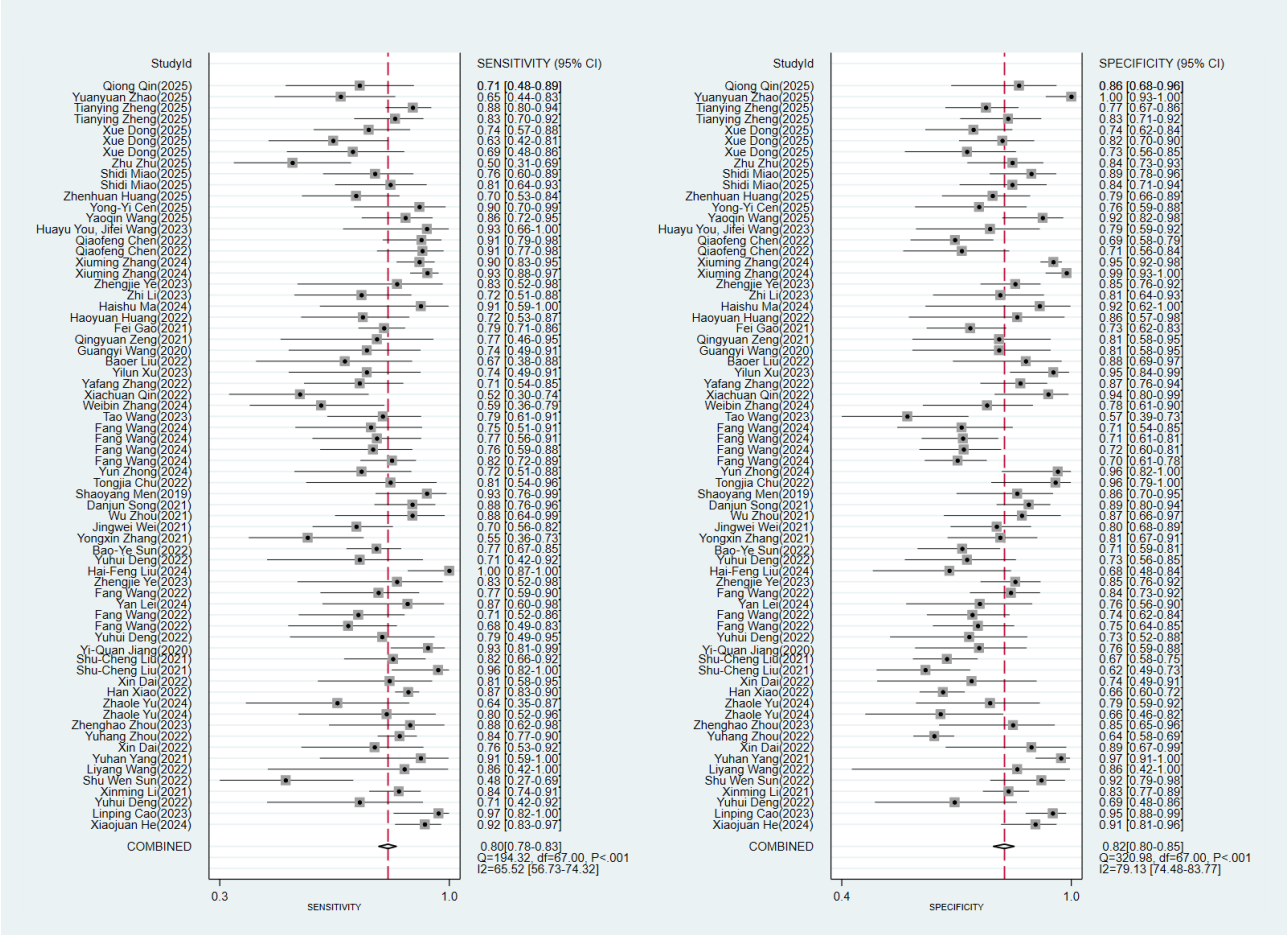

XSL•FO
RenderX

Figure 4. Meta-analysis SROC: specificity and sensitivity of image-based deep learning in microvascular invasion diagnosis. SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

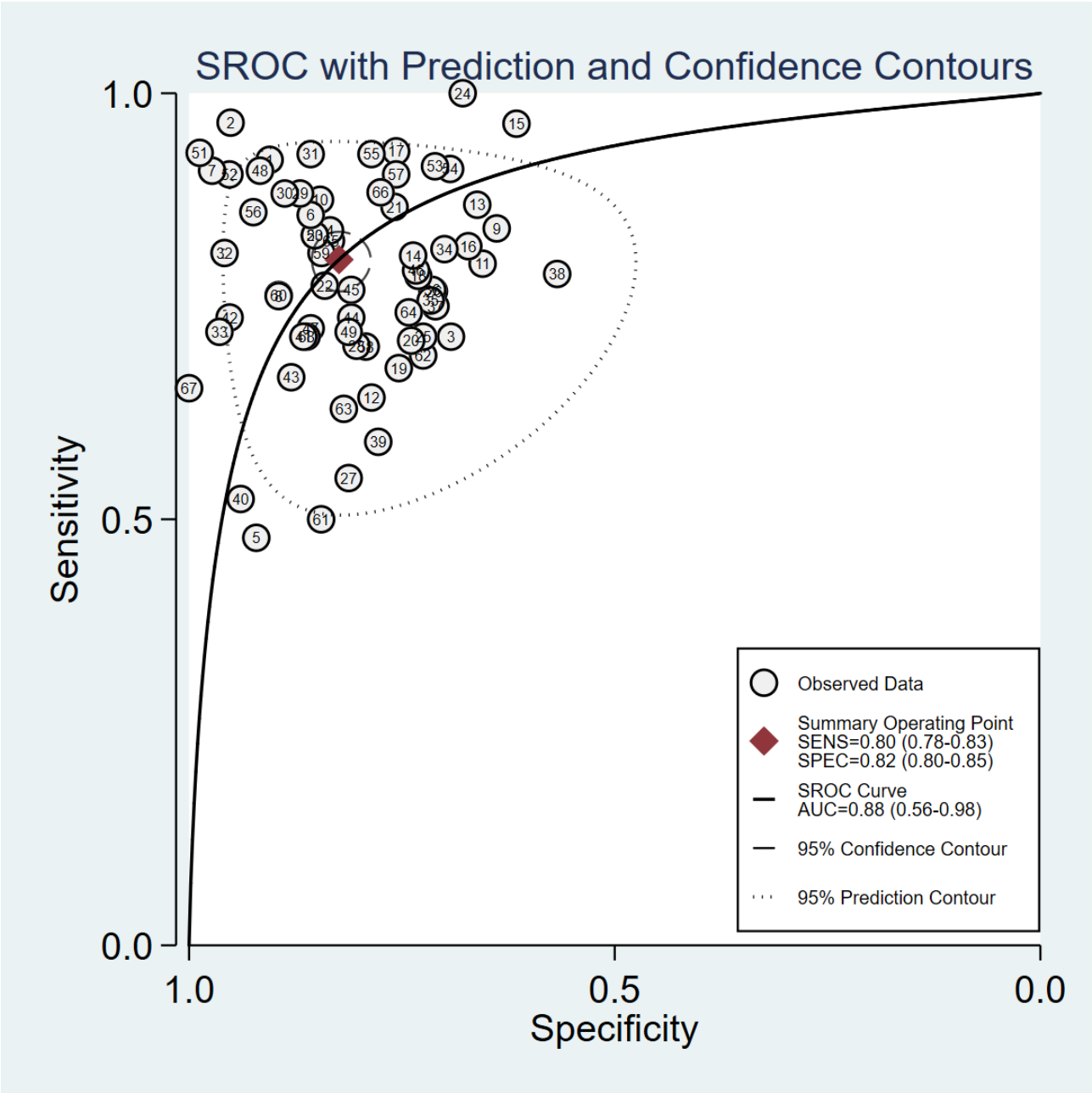

Deeks' funnel plot indicated no significant small-study effects (*P*=.77, Table 2 and Figure 5). When the prior probability of MVI was set to 40%, the posterior probabilities corresponding to positive and negative DL model detection results were 75% and 14%, respectively. Fagan's nomogram analysis showed that a positive detection result increased the posterior probability by 35% compared with the prior probability, whereas a negative detection result decreased it by 26% (Figure 6).

XSL•FO
RenderX

Figure 5. Deeks' funnel plot from meta-analysis of the specificity and sensitivity of image-based deep learning in microvascular invasion diagnosis.

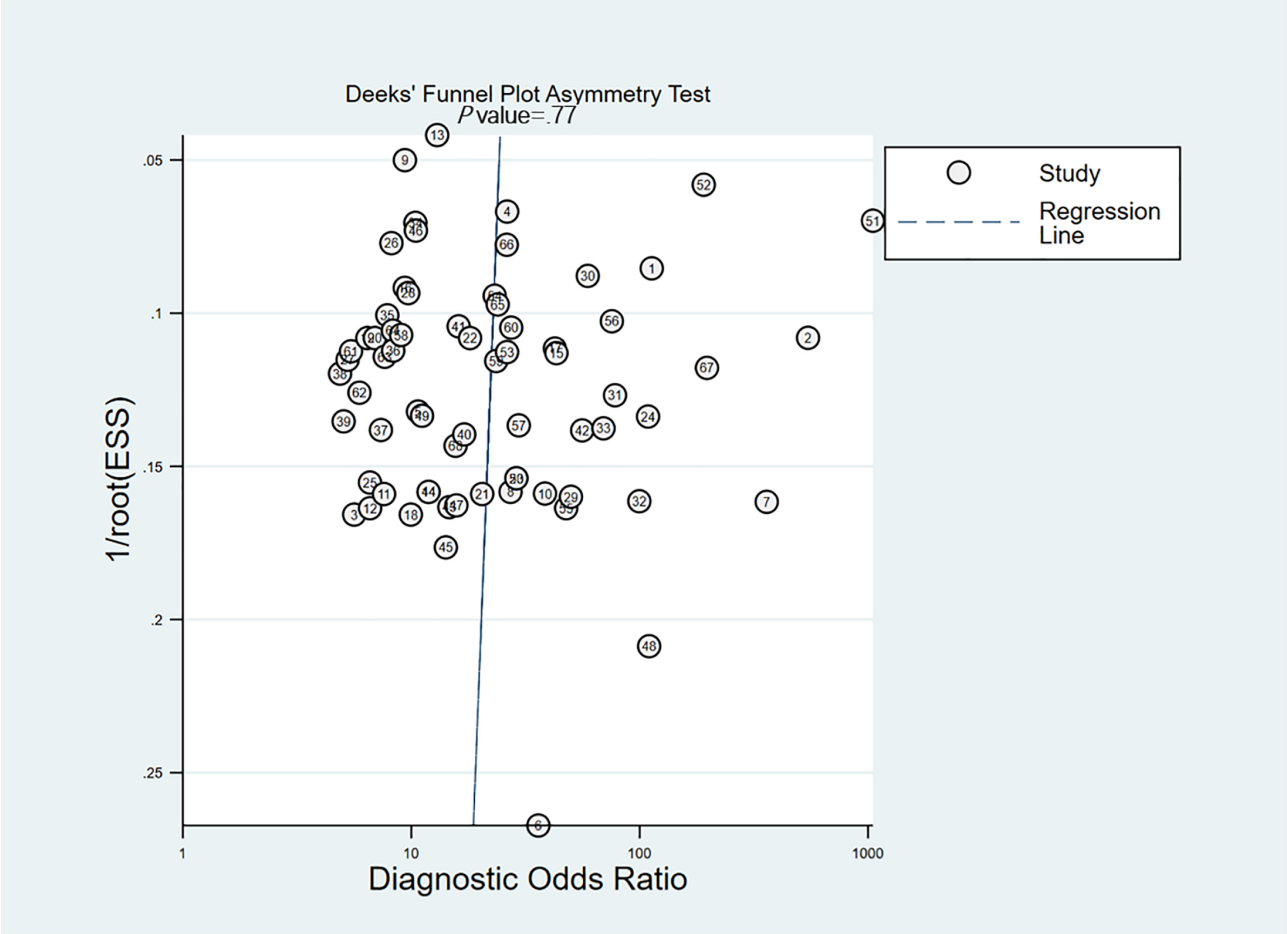

XSL•FO
RenderX

Figure 6. Fagan's nomogram from meta-analysis of the specificity and sensitivity of image-based deep learning in microvascular invasion diagnosis.

XSL•FO
RenderX

Subgroup Analysis by Validation Set Generation Method

Internal Validation

The model's accuracy was validated using 49 diagnostic fourfold tables in internal validation. The Spearman correlation coefficient was 0.04, suggesting a minimal threshold effect. This effect accounted for none of the observed between-study heterogeneity. The pooled analysis yielded the following results: sensitivity 0.82 (95% CI 0.79 - 0.85, I^2=61.27%), specificity 0.83 (95% CI 0.80 - 0.86, I^2=72.22%), LR+4.9 (95% CI 4.1 - 6.0), LR–0.22 (95% CI 0.18 - 0.26), DOR 23 (95% CI 17 - 32), and SROC 0.90 (95% CI 1.00 - 0.00; Table 2 and Figures S1A,B in Multimedia Appendix 2).

Deeks' funnel plot illustrated no significant small-study effects (*P*=.85, Table 2 and Figure S1C in Multimedia Appendix 2). When the prior probability of MVI was set to 40%, the corresponding posterior probabilities for positive and negative DL model detection results were 77% and 13%, respectively. Fagan's nomogram analysis showed that a positive detection result increased the posterior probability by 37%, compared with the prior probability; a negative detection result decreased the posterior probability by 27% (Figure S1D in Multimedia Appendix 2).

External Validation

The model's accuracy was validated using 19 diagnostic fourfold tables in external validation. The Spearman correlation coefficient was −0.15, indicating a minimal threshold effect. This effect accounted for 2% of the observed between-study heterogeneity. The pooled analysis yielded the following results: sensitivity 0.77 (95% CI 0.72 - 0.82, I^2=73.40%), specificity 0.80 (95% CI 0.74 - 0.85, I^2=83.19%), LR+3.9 (95% CI 3.0 - 5.0), LR–0.29 (95% CI 0.23 - 0.36), DOR 13 (95% CI 9 - 20), and SROC 0.85 (95% CI 1.00 - 0.00; Table 2 and Figures S2A,B in Multimedia Appendix 3).

Deeks' funnel plot revealed no significant small-study effects (*P*=.22, Table 2 and Figure S2C in Multimedia Appendix 3). When the prior probability of MVI was set to 39%, the corresponding posterior probabilities for positive and negative DL model detection results were 71% and 15%, respectively. Fagan's nomogram analysis showed that a positive detection result increased the posterior probability by 32% compared with the prior probability, whereas a negative detection result decreased it by 24% (Figure S2D inMultimedia Appendix 3).

Validation Set Size < 100 Cases

Among studies with a validation set size <100 cases, the model's accuracy was validated using 38 diagnostic fourfold tables. The Spearman correlation coefficient was −0.44, suggesting a minimal threshold effect. This effect accounted for 19% of the observed between-study heterogeneity. The pooled analysis yielded the following results: sensitivity 0.79 (95% CI 0.75 - 0.83, I^2=56.49%), specificity 0.83 (95% CI 0.79 - 0.87, I^2=67.03%), LR+4.7 (95% CI 3.8 - 5.8), LR−0.25 (95% CI 0.20 - 0.30), DOR 19 (95% CI 14 - 26), and SROC 0.88 (95% CI 0.63 - 0.97; Table 2 and Figure S3A,B in Multimedia Appendix 4).

Deeks' funnel plot revealed no significant small-study effects (*P*=.15, Table 2 and Figure S3C in Multimedia Appendix 4). When the prior probability of MVI was set to 41%, the corresponding posterior probabilities for positive and negative DL model detection results were 77% and 15%, respectively. Fagan's nomogram analysis displayed that a positive detection result increased the posterior probability by 36% compared with the prior probability, whereas a negative detection result decreased it by 26% (Figure S3D in Multimedia Appendix 4).

Validation Set Size ≥ 100 Cases

Among studies with a validation set size ≥100 cases, the model's accuracy was validated using 30 diagnostic fourfold tables. The Spearman correlation coefficient was 0.47, indicating a minimal threshold effect. This effect accounted for 22% of the observed between-study heterogeneity. The pooled analysis yielded the following results: sensitivity 0.82 (95% CI 0.78 - 0.85, I^2=69.35%), specificity 0.81 (95% CI 0.77 - 0.85, I^2=85.58%), LR+4.3 (95% CI 3.4 - 5.5), LR−0.23 (95% CI 0.18 - 0.28), DOR 19 (95% CI 13 - 29), and SROC 0.88 (95% CI 0.66 - 0.97; Table 2 and Figure S4A,B in Multimedia Appendix 5).

Deeks' funnel plot revealed no significant small-study effects (*P*=.50; Table 2 and Figure S4C in Multimedia Appendix 5). When the prior probability of MVI was set to 39%, the corresponding posterior probabilities for positive and negative DL model detection results were 73% and 13%, respectively. Fagan's nomogram analysis showed that a positive detection result increased the posterior probability by 34% compared with the prior probability, whereas a negative detection result decreased it by 26% (Figure S4D in Multimedia Appendix 5).

Subgroup Analysis by Image Source

DL Based on CECT

In CECT, the model's accuracy was validated using 20 diagnostic fourfold tables. The Spearman correlation coefficient was −0.28, suggesting a minimal threshold effect. This effect accounted for 8% of the observed between-study heterogeneity. The pooled analysis revealed the following results: sensitivity 0.84 (95% CI 0.79 - 0.88, I^2=71.67%), specificity 0.83 (95% CI 0.77 - 0.88, I^2=87.84%), LR+5.0 (95% CI 3.6 - 6.9), LR−0.19 (95% CI 0.15 - 0.25), DOR 26 (95% CI 16 - 42), and SROC 0.90 (95% CI 1.00 - 0.00; Table 2 and Figure S5 in Multimedia Appendix 6). Deeks funnel plot revealed no significant small-study effects (*P*=.09, Table 2).

Among CECT-based models with internal validation, the accuracy was validated using 16 diagnostic fourfold tables. The Spearman correlation coefficient was −0.26, indicating a minimal threshold effect. This effect accounted for 7% of the observed between-study heterogeneity. The pooled analysis yielded the following results: sensitivity 0.86 (95% CI 0.80 - 0.90, I^2=71.85%), specificity 0.85 (95% CI 0.78 - 0.90, I^2=88.74%), LR+5.7 (95% CI 3.8 - 8.4), LR−0.17 (95% CI 0.12 - 0.24), DOR 33 (95% CI 19 - 59), and SROC 0.92 (95% CI 1.00 - 0.00; Table 2 and Multimedia Appendix 7). Deeks' funnel plot revealed no significant small-study effects (*P*=.19, Table 2).

XSL•FO
RenderX

Among CECT-based models with external validation, the accuracy was validated using four diagnostic fourfold tables. The pooled analysis yielded the following results: sensitivity 0.82 (95% CI 0.73 - 0.88, I^2=62.47%), specificity 0.76 (95% CI 0.64 - 0.85, I^2=77.88%), LR+3.4 (95% CI 2.3 - 5.0), LR−0.24 (95% CI 0.17 - 0.34), DOR 14 (95% CI 9 - 23), and SROC 0.86 (95% CI 1.00 - 0.00; Table 2 and Multimedia Appendix 7). Subsequent Doi plot analysis showed moderate publication bias among the included studies (LFK index=−1.14, Table 2).

DL Based on CEUS

In CEUS, the model's accuracy was validated using 5 diagnostic fourfold tables. The pooled analysis yielded the following results: sensitivity 0.70 (95% CI 0.58 - 0.80, I^2=58.75%), specificity 0.88 (95% CI 0.82 - 0.92, I^2=30.18%), LR+5.6 (95% CI 3.5 - 9.0), LR−0.34 (95% CI 0.23 - 0.51), DOR 17 (95% CI 7 - 37), and SROC 0.89 (95% CI 1.00 - 0.00; Table 2 and Multimedia Appendix 8). Analysis using the Doi plot revealed minimal publication bias among the included studies (LFK index=0.46, Table 2).

DL Based on CEMRI

In CEMRI, the model's accuracy was validated using 24 diagnostic fourfold tables. The Spearman correlation coefficient was −0.22, suggesting a minimal threshold effect. This effect accounted for 5% of the observed between-study heterogeneity. The pooled analysis yielded the following results: sensitivity 0.78 (95% CI 0.73 - 0.83, I^2=61.70%), specificity 0.81 (95% CI 0.76 - 0.85, I^2=65.14), LR+4.0 (95% CI 3.2 - 5.0), LR− 0.27 (95% CI 0.22 - 0.33), DOR 15 (95% CI 10 - 22), and SROC 0.86 (95% CI 1.00 - 0.00; Table 2 and Multimedia Appendix 9). Deeks funnel plot revealed no significant small-study effects (P=0.15, Table 2).

Among CEMRI models with internal validation, the accuracy was validated using 16 diagnostic fourfold tables. The Spearman correlation coefficient was 0.17, indicating a minimal threshold effect. This effect accounted for 3% of the observed between-study heterogeneity. The pooled analysis yielded the following results: sensitivity 0.81 (95% CI 0.75 - 0.86, I^2=57.28%), specificity 0.81 (95% CI 0.75 - 0.85, I^2=61.32), LR+4.3 (95% CI 3.2 - 5.6), LR− 0.24 (95% CI 0.17 - 0.32), DOR 18 (95% CI 11 - 30), and SROC 0.88 (95% CI 1.00 - 0.00; Table 2 and Figure S9 in Multimedia Appendix 10). Deeks funnel plot revealed no significant small-study effects (P=.17, Table 2).

Among CEMRI models with external validation, the accuracy was validated using 8 diagnostic fourfold tables. The Spearman correlation coefficient was −0.72, suggesting a significant threshold effect. This effect accounted for 52% of the observed between-study heterogeneity. The pooled analysis yielded the following results: sensitivity 0.74 (95% CI 0.65 - 0.82, I^2=70.90%), specificity 0.80 (95% CI 0.71 - 0.87, I^2=72.55%), LR+3.7 (95% CI 2.7 - 5.2), LR−0.32 (95% CI 0.24 - 0.43), DOR 12 (95% CI 7 - 18), and SROC 0.84 (95% CI 1.00 - 0.00; Table 2 and Figure S9 in Multimedia Appendix 10). Further Doi plot analysis revealed moderate publication bias among the included studies (LFK index=1.74, Table 2).

DL Based on MRI

In MRI, the model's accuracy was validated using 5 diagnostic fourfold tables. The pooled analysis revealed the following results: sensitivity 0.76 (95% CI 0.68 - 0.82, I^2=0.00%), specificity 0.80 (95% CI 0.70 - 0.87, I^2=0.00%), LR+3.7 (95% CI 2.5 - 5.5), LR−0.30 (95% CI 0.23 - 0.40), DOR 12 (95% CI 7 - 21), and SROC, 0.84 (95% CI 1.00-0.00; Table 2 and Multimedia Appendix 11). Analysis using the Doi plot revealed substantial publication bias among the included studies (LFK index=3.11, Table 2).

DL Based on Multimodal Imaging

In multimodal medical images, the model's accuracy was validated using 10 diagnostic fourfold tables. The pooled analysis yielded the following results: sensitivity 0.74 (95% CI 0.68 - 0.80, I^2=0.00%), specificity 0.79 (95% CI 0.75 - 0.83, I^2=11.32%), LR+3.6 (95% CI 2.9 - 4.4), LR− 0.33 (95% CI 0.26 - 0.41), DOR 11 (95% CI 7 - 16), and SROC 0.82 (95% CI 1.00 - 0.00; Table 2 and Multimedia Appendix 12). Deeks' funnel plot revealed no significant small-study effects (P=.44, Table 2).

In multimodal internal validation, the model's accuracy was validated using 6 diagnostic fourfold tables. The pooled analysis revealed the following results: sensitivity 0.79 (95% CI 0.69 - 0.87, I^2=0.00%), specificity 0.79 (95% CI 0.73 - 0.85, I^2=31.69%), LR+3.8 (95% CI 2.8 - 5.3), LR− 0.26 (95% CI 0.17 - 0.41), DOR 15 (95% CI 7 - 29), and SROC 0.86 (95% CI 1.00 - 0.00; Table 2 and Multimedia Appendix 13). Analysis using the Doi plot revealed moderate publication bias among the included studies (LFK index=1.6; Table 2).

In multimodal external validation, the model's accuracy was validated using 4 diagnostic fourfold tables. The pooled analysis showed the following results: sensitivity 0.71 (95% CI 0.64 - 0.78, I^2=0.00%), specificity 0.78 (95% CI 0.73 - 0.83, I^2=0.00%), LR+3.3 (95% CI 2.6 - 4.2), LR− 0.36 (95% CI 0.28 - 0.47), DOR 9 (95% CI 6 - 14), and SROC 0.82 (95% CI 1.00 - 0.00; Table 2 and Multimedia Appendix 13). Subsequent Doi plot analysis indicated moderate publication bias (LFK index=1.93; Table 2).

DL Based on Pathological Sections

In pathological sections, the model's accuracy was validated using 4 diagnostic fourfold tables. The pooled analysis revealed the following results: sensitivity 0.91 (95% CI 0.87 - 0.94, I^2=0.00%), specificity 0.90 (95% CI 0.68 - 0.97, I^2=94.83%), LR+9.2 (95% CI 2.5 - 33.6), LR− 0.09 (95% CI 0.06 - 0.15), DOR 97 (95% CI 20 - 465), and SROC 0.92 (95% CI 1.00 - 0.00; Table 2 and Multimedia Appendix 14). Further Doi plot analysis revealed substantial publication bias (LFK index=3.08; Table 2).

Discussion

Summary of the Main Findings

The current meta-analysis revealed that the modeling methods used for DL-based prediction of HCC MVI primarily used

XSL•FO
RenderX

CECT, CEUS, CEMRI, MRI, multimodal imaging, and pathological image techniques. DL models based on medical imaging showed favorable overall diagnostic performance in predicting HCC MVI, with a pooled sensitivity of 0.80 (95% CI 0.78 - 0.83) and specificity of 0.82 (95% CI 0.80 - 0.85). Further analysis across imaging modalities revealed that CECT-based models achieved the highest diagnostic efficacy, showing a sensitivity of 0.84 (95% CI 0.79 - 0.88) and specificity of 0.83 (95% CI 0.77 - 0.88). Models based on CEUS exhibited particularly high specificity (0.88, 95% CI 0.82 - 0.92). Furthermore, models using pathological slides, considered the diagnostic reference standard, attained the highest overall performance, with a sensitivity of 0.91 (95% CI 0.87 - 0.94) and specificity of 0.90 (95% CI 0.68 - 0.97). Therefore, these models appear promising as a diagnostic approach for MVI in HCC.

Comparison With Other Previous Reviews

This study also noted that some researchers have discussed the use of machine learning for MVI in HCC. Xiao et al [71] and Liang et al [72], who focused on MRI and ultrasound radiomics, respectively, validated the predictive potential of single modalities, with pooled area under the curves (AUCs) of 0.87 and 0.81, respectively. However, their analyses included a limited number of studies and focused only on a single imaging modality, limiting the generalizability of their conclusions. Li et al [73] integrated multiple imaging modalities across 22 studies (involving 4129 participants), reporting a pooled AUC of 0.90 for radiomic models. However, their analysis primarily incorporated traditional machine learning models and did not stratify performance based on the validation set generation method. This limitation may lead to an overly optimistic assessment of model generalizability.

Unlike previous systematic reviews that focused on traditional machine learning models, our meta-analysis focuses on the value of DL algorithms for diagnosing HCC MVI. Methodologically, this study provides a deeper exploration of the sources of performance heterogeneity through subgroup analyses based on the validation set generation method and image source. A particular methodological strength is the inclusion of pathological sections, the diagnostic gold standard, as a benchmark for performance optimization. This meta-analysis synthesized data from 52 studies involving 19,531 patients and provided more robust and reliable conclusions than those from analyses with smaller sample sizes.

Influence of Imaging Modalities on DL

The imaging modalities used to construct DL models can be categorized into 2 main types: noninvasive and invasive. This study showed that the properties of different modalities and their ability to extract biological features directly affected the diagnostic efficacy of corresponding models.

Among noninvasive imaging modalities, CECT can effectively capture tumor heterogeneity, enhancement patterns, and peritumoral microenvironment changes due to its high spatial resolution and multi-phase dynamic imaging capabilities. These features support the superior diagnostic performance of DL models (AUC=0.90) [74,75]. Models developed from CEUS achieved commendable performance (AUC=0.89) and exhibited a high pooled specificity (0.88). This finding suggests that CEUS's real-time hemodynamic properties may be valuable in ruling out MVI-negative cases. However, its comparatively lower sensitivity (0.70) concurrently suggests ongoing challenges in consistently identifying MVI-positive features. Integrating Sonazoid-based functional imaging with conventional ultrasound characteristics and serum markers in the future may be a promising way to improve performance [76]. CEMRI has unique advantages in depicting tumor boundaries and detecting subtle changes in the peritumoral liver parenchyma due to its excellent soft tissue contrast [77]. The model's performance (sensitivity: 0.78, specificity: 0.81, AUC=0.86) surpasses the assessment that relies on traditional visual features, as reported by Wu et al [78] (sensitivity: 0.55, specificity: 0.87, AUC=0.80). However, its widespread adoption is limited by long examination times, high costs, and dependence on equipment. Nonenhanced MRI models also demonstrate predictive potential (sensitivity: 0.76, specificity: 0.80, AUC=0.84). Notably, multimodal fusion models, which are designed to integrate complementary information, have not demonstrated significant advantages in the studies included (AUC=0.82). This may be due to the simplified fusion strategies or data heterogeneity.

Among invasive imaging modalities, DL models based on pathological sections demonstrated the highest diagnostic efficacy in this study (AUC=0.92). This highlights the gold-standard status of pathology in evaluating MVI. However, the inherent invasiveness of pathological examination precludes its use in preoperative decision-making. Therefore, a central challenge for future research is to effectively transfer and integrate the gold-standard-level diagnostic insights from pathological sections into preoperative, noninvasive imaging. This objective requires leveraging advanced methodologies, such as transfer learning, cross-modal fusion, and generative models. Augmenting existing advantageous modalities, such as CECT, with these advancements could ultimately pave the way for a clinically viable preoperative "virtual biopsy."

Image Segmentation

Accurate image segmentation is essential for building reliable DL models. However, the literature incorporated in this meta-analysis primarily relies on manual segmentation. This method can introduce subjective bias (38/52, 73.1%). While advanced network architectures have achieved expert-level precision in segmenting normal liver tissue, as demonstrated by Dice coefficients ranging from 0.968 to 0.982 [79,80], comparative data from the Liver Tumor Segmentation challenge reveal significant discrepancies. Specifically, the accuracy of liver tumor segmentation (Dice: 0.739) is considerably lower than that of liver parenchyma segmentation (Dice: 0.963) [81]. This discrepancy primarily stems from the heterogeneity of HCC lesions, suboptimal image contrast, and a lack of high-quality annotated data. Concurrently, selecting the appropriate segmentation strategy is critical. While 3D segmentation comprehensively captures spatial heterogeneity, the more clinically feasible 2D approach sacrifices substantial volumetric information [82]. Given the demand for submillimeter-level precision in MVI prediction, there are 2

XSL•FO
RenderX

key technological pathways for enhancing model stability. First, exploring segmentation paradigms that require fewer annotations, such as weakly or self-supervised learning, can reduce dependency on annotations. Second, developing novel network architectures designed specifically to address HCC heterogeneity and boundary ambiguity is equally crucial.

Validation Set Generation Method

The rigor of validation strategies is paramount for evaluating the real-world generalizability of DL models in predicting MVI of HCC [83,84]. The present analysis reveals that, despite exemplary performance during internal validation, consistent performance declines emerge in independent external validation cohorts. This finding clearly shows that relying too much on internal validation can lead to overestimating a model's true efficacy [85]. Concern regarding validation strategies is not unique to DL research. A recent meta-analysis [86] that focused on MRI-based radiomics for predicting HCC recurrence and MVI similarly concluded that the current predominance of internal validation results in an overestimation of model generalizability as well. Consequently, any proclaimed superior performance may substantially diminish when confronting real-world heterogeneity, if the evaluation framework remains confined to internal validation, irrespective of the underlying algorithm—be it DL or radiomics. Therefore, promoting rigorous external validation and establishing standardized, cross-institutional imaging protocols are essential steps toward reliable clinical translation in this field.

Heterogeneity Analysis

There is substantial heterogeneity among the included studies. Notably, high levels of heterogeneity persist within subgroups, even after stratifying analyses by imaging modality and validation strategy. This observation objectively reflects the inherent complexity of artificial intelligence (AI)–based medical imaging research and is a common challenge for meta-analyses in this field. The heterogeneity primarily stems from 3 levels. Technically, variations in critical parameters, including imaging equipment, magnetic field strength, and slice thickness, directly influence image texture and quality. These variations are a significant technical source of variability in model performance. Methodologically, diversity in study design, validation strategies, segmentation techniques, and DL network architectures introduces additional variation in model construction and performance interpretation. Clinically, differences in patient populations regarding geographic distribution, underlying liver disease etiology, and disease stage may also affect model performance and generalizability. While this heterogeneity limits the direct interpretability of the pooled results to some extent, it accurately reflects the diversity in methodology and clinical practice within the field. Future investigations should adhere to the Findable, Accessible, Interoperable, and Reusable principles, providing detailed reporting of imaging acquisition parameters, model architectures, and training specifics. Such comprehensive reporting will facilitate the in-depth exploration of heterogeneity sources via methods such as meta-regression. This will promote the identification of key influencing factors and the standardization of methodologies.

Methods for ROB Assessment

This systematic review used a composite strategy to assess the ROB. First, all included studies were rigorously evaluated according to the QUADAS-2 guidelines. The results showed that all primary studies were rated as having a high ROB in the "Patient Selection" domain of QUADAS-2 due to the widespread use of the retrospective case-control design. While this outcome aligns with the QUADAS-2 assessment principles, it also reveals a limitation of the tool when evaluating machine learning-based diagnostic studies that use retrospective data. The tool struggles to differentiate nuances in data construction quality among studies.

To conduct a more granular assessment of data-level bias risk, a supplemental analysis was performed using items from the Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence (QUADAS-AI) tool targeting "Study Participant Selection." The QUADAS-AI tool provides more detailed criteria for this dimension, including an explicit description of data source, size, and quality characteristics, use of open-source datasets, a clear rationale for splitting data into training, validation, and test sets, performance of image preprocessing, and provision of scanner model information. The analysis using QUADAS-AI items showed that all 52 studies appropriately described the source, size, and quality of the input data and clearly defined the patient inclusion criteria. Among these studies, only one used an open-source dataset. All studies provided a rationale for the data split. Image preprocessing was performed in all studies. However, 11 studies did not report the scanner model used for image acquisition (Table S2 in Multimedia Appendix 15 [14,16,21-70]).

Advantages and Limitations

A primary strength of this research is that it is the first large-scale, systematic meta-analysis to evaluate medical imaging-based DL models for predicting MVI in HCC. The analysis included 52 studies with 19,531 patients, providing the field with comprehensive evidence. The analytical process strictly adhered to the PRISMA guidelines, and bias risk was evaluated using QUADAS-2. These measures ensured methodological rigor and transparency. However, several limitations warrant consideration. First**,** the training cohorts in most primary studies were small. Only 3 studies had a sample size greater than 1000 patients. The robustness of DL models depends heavily on large volumes of high-quality data. Therefore, restricted training sample sizes are a potential methodological limitation. This likely contributes to overfitting in some models, which is probably an internal reason for the observed performance degradation during external validation. This finding underscores the fundamental importance of acquiring large-scale, high-quality datasets to enhance model generalizability in the development of current DL models. Second, the 95% CIs for the pooled AUCs and for most subgroup analyses were exceptionally wide. This constrained the interpretation of the result precision to some extent. This primarily stems from the significant heterogeneity among the included studies. The limited number of studies in subgroup analyses exacerbated data sparsity. Third, Doi plot analyses for some subgroups indicated moderate to substantial publication

XSL•FO
RenderX

bias. These analyses showed that the pooled results for these subgroups may be influenced by unpublished negative studies, which poses a risk of overestimating diagnostic performance. Fourth, the rigor of the validation strategies needs to be improved. Most studies relied on internal validation. Only a few conducted stringent external validation. This reliance may introduce an optimistic bias into the overall assessment of the models' real-world generalizability. While the overall performance on external validation sets was discussed, the limited quantity of external validation data prevented a more in-depth subgroup analysis of validation strategies by different imaging modalities. Fifth, the vast majority of primary studies inadequately reported model calibration metrics or details about network complexity. This omission hinders a quantitative, systematic evaluation of predictive reliability and overfitting risk at the review level and impacts the comprehensive assessment of clinical applicability. This reflects a general deficiency in the transparency of methodological reporting within the current research landscape.

Clinical Implications and Future Perspectives

This study indicates that medical imaging-based DL models, particularly those using preoperative CECT, demonstrate promising diagnostic performance in predicting MVI of HCC. These models have the potential to assist in personalized surgical planning. However, translating them into clinical practice faces multiple challenges. One primary issue is establishing clinical decision thresholds. While this analysis quantified predictive probabilities via Fagan nomograms, there is currently a lack of evidence-based guidelines defining "at what predicted probability of MVI the surgical margin should be adjusted." Future work must integrate clinical outcome data and empirically explore the net benefit of different thresholds using methods such as decision curve analysis. Second, these models' generalizability and deployment feasibility need urgent enhancement. Performance degradation during external validation suggests susceptibility to variations in imaging protocols, equipment, and patient populations. Furthermore, model interpretability is crucial for gaining clinical trust. This necessitates developing transparent methods for presenting decision rationale.

To bridge the gap between "high performance" and "high utility," future efforts must focus on 3 interconnected levels. At the research level, the focus must shift from model construction to rigorous, prospective, multicenter validation to unequivocally assess generalizability. At the algorithmic level, it is essential to explore cross-modal information fusion, especially using transfer learning, to bring "gold standard"–level diagnostic insights from histopathological sections to preoperative, noninvasive imaging. At the clinical level, establishing standardized, cross-institutional imaging protocols and developing decision support systems that integrate seamlessly into clinical workflows is imperative. This integrated approach is vital for reliably translating technology from innovation to tangible patient benefit.

Conclusions

This systematic review and meta-analysis demonstrate that medical imaging-based DL models, especially those leveraging preoperative CECT, hold significant promise for the noninvasive preoperative prediction of MVI in HCC. Unlike previous reviews that focused on radiomics or single imaging modalities, this study conducted a comprehensive comparison across multiple modalities. The study also emphasizes the critical role of external validation in the real-world generalizability of a model. However, substantial heterogeneity across studies and the performance degradation observed during independent external validation suggest that their generalizability to the real world must be confirmed through more rigorous study designs. Consequently, future research should prioritize establishing model robustness via prospective, multicenter external validation, coupled with efforts to standardize methodologies and improve reporting transparency. A critical step toward reliable clinical translation and achieving the ultimate goal of a "virtual biopsy" is developing algorithms that can translate pathology-grade diagnostic insights into preoperative, noninvasive imaging.

Funding

This study was funded by the General Project of the Liaoning Provincial Department of Education (grant number JYTMS20230109) and the Liaoning Provincial Science and Technology Program Joint Plan (Natural Science Foundation General Program) (grant number 2025-MSLH-742). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

All authors contributed to the study conception and design. WF and BQ prepared the original draft of the manuscript and contributed to the methodology, formal analysis, and investigation. SH was responsible for conceptualization and funding acquisition and provided supervision. SH also reviewed and edited the manuscript. All authors commented on previous versions of the manuscript and also read and approved the final manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Flow diagram illustrating the search strategy.
[DOCX File, 22 KB - jmir_v28i1e82000_app1.docx]

Multimedia Appendix 2
(A) Meta-analysis forest plots: specificity and sensitivity of image-based deep learning (DL) in microvascular invasion (MVI) diagnosis in internal validation; (B) Meta-analysis summary receiver operating characteristic: specificity and sensitivity of image-based DL in MVI diagnosis in internal validation; (C) Deeks funnel plot from meta-analysis of specificity and sensitivity of image-based DL in MVI diagnosis in internal validation; (D) Fagan nomogram from meta-analysis of specificity and sensitivity of image-based DL in MVI diagnosis in internal validation.
[PNG File, 606 KB - jmir_v28i1e82000_app2.png]

Multimedia Appendix 3
(A) Meta-analysis forest plots: specificity and sensitivity of image-based deep learning (DL) in microvascular invasion (MVI) diagnosis in external validation; (B) Meta-analysis summary receiver operating characteristic: specificity and sensitivity of image-based DL in MVI diagnosis in external validation; (C) Deeks funnel plot from meta-analysis of specificity and sensitivity of image-based DL in MVI diagnosis in external validation; (D) Fagan nomogram from meta-analysis of specificity and sensitivity of image-based DL in MVI diagnosis in external validation.
[PNG File, 402 KB - jmir_v28i1e82000_app3.png]

Multimedia Appendix 4
(A) Meta-analysis forest plots: specificity and sensitivity of image-based deep learning (DL) in microvascular invasion (MVI) diagnosis in Validation set size of <100 cases; (B) Meta-analysis summary receiver operating characteristic: specificity and sensitivity of image-based DL in MVI diagnosis in Validation set size < 100 cases; (C) Deeks funnel plot from meta-analysis of specificity and sensitivity of image-based DL in MVI diagnosis in the validation set size of <100 cases; (D) Fagan nomogram from meta-analysis of specificity and sensitivity of image-based DL in MVI diagnosis in the validation set size of <100 cases.
[PNG File, 519 KB - jmir_v28i1e82000_app4.png]

Multimedia Appendix 5
(A) Meta-analysis forest plots: specificity and sensitivity of image-based deep learning (DL) in microvascular invasion (MVI) diagnosis in Validation set size of ≥100 cases; (B) Meta-analysis summary receiver operating characteristic: specificity and sensitivity of image-based DL in MVI diagnosis in Validation set size of ≥100 cases; (C) Deeks funnel plot from meta-analysis of specificity and sensitivity of image-based DL in MVI diagnosis in the validation set size ≥100 cases; (D) Fagan nomogram from meta-analysis of specificity and sensitivity of image-based DL in MVI diagnosis in Validation set size of ≥100 cases.
[PNG File, 528 KB - jmir_v28i1e82000_app5.png]

Multimedia Appendix 6
(A) Meta-analysis forest plots: specificity and sensitivity of contrast-enhanced computed tomography (CECT)–based deep learning (DL) in microvascular invasion (MVI) diagnosis; (B) Meta-analysis SROC: specificity and sensitivity of CECT-based DL in MVI diagnosis.
[PNG File, 320 KB - jmir_v28i1e82000_app6.png]

Multimedia Appendix 7
(A) Meta-analysis forest plots: specificity and sensitivity of contrast-enhanced computed tomography (CECT)–based deep learning (DL) in microvascular invasion (MVI) diagnosis in internal validation; (B) Meta-analysis SROC: specificity and sensitivity of CECT-based DL in MVI diagnosis in internal validation; (C) Meta-analysis forest plots: specificity and sensitivity of CECT-based DL in MVI diagnosis in external validation; (D) Meta-analysis summary receiver operating characteristic: specificity and sensitivity of CECT-based DL in MVI diagnosis in external validation.
[PNG File, 394 KB - jmir_v28i1e82000_app7.png]

Multimedia Appendix 8
(A) Meta-analysis forest plots: specificity and sensitivity of contrast-enhanced ultrasound (CEUS)–based deep learning (DL) in microvascular invasion (MVI) diagnosis; (B) Meta-analysis summary receiver operating characteristic: specificity and sensitivity of CEUS-based DL in MVI diagnosis.
[PNG File, 172 KB - jmir_v28i1e82000_app8.png]

Multimedia Appendix 9

XSL•FO
RenderX

(A) Meta-analysis forest plots: specificity and sensitivity of contrast-enhanced magnetic resonance imaging (CEMRI)–based deep learning (DL) in microvascular invasion (MVI) diagnosis; (B) Meta-analysis summary receiver operating characteristic: specificity and sensitivity of CEMRI-based DL in MVI diagnosis.
[PNG File, 374 KB - jmir_v28i1e82000_app9.png]

Multimedia Appendix 10
(A) Meta-analysis forest plots: specificity and sensitivity of contrast-enhanced magnetic resonance imaging (CEMRI)–based deep learning (DL) in microvascular invasion (MVI) diagnosis in internal validation; (B) Meta-analysis summary receiver operating characteristic: specificity and sensitivity of CEMRI-based DL in MVI diagnosis in internal validation; (C) Meta-analysis forest plots: specificity and sensitivity of CEMRI-based DL in MVI diagnosis in external validation; (D) Meta-analysis SROC: specificity and sensitivity of CEMRI-based DL in MVI diagnosis in external validation.
[PNG File, 424 KB - jmir_v28i1e82000_app10.png]

Multimedia Appendix 11
(A) Meta-analysis forest plots: specificity and sensitivity of magnetic resonance imaging (MRI)–based deep learning (DL) in microvascular imaging (MVI) diagnosis; (B) Meta-analysis summary receiver operating characteristic: specificity and sensitivity of MRI-based DL in MVI diagnosis.
[PNG File, 171 KB - jmir_v28i1e82000_app11.png]

Multimedia Appendix 12
(A) Meta-analysis forest plots: specificity and sensitivity of multimodal imaging-based deep learning (DL) in microvascular invasion (MVI) diagnosis; (B) Meta-analysis summary receiver operating characteristic: specificity and sensitivity of multimodal imaging-based DL in MVI diagnosis.
[PNG File, 222 KB - jmir_v28i1e82000_app12.png]

Multimedia Appendix 13
(A) Meta-analysis forest plots: specificity and sensitivity of multimodal imaging-based deep learning (DL) in microvascular invasion (MVI) diagnosis in internal validation; (B) Meta-analysis summary receiver operating characteristic (SROC): specificity and sensitivity of multimodal imaging-based DL in MVI diagnosis in internal validation; (C) Meta-analysis forest plots: specificity and sensitivity of multimodal imaging-based DL in MVI diagnosis in external validation; (D) Meta-analysis SROC: specificity and sensitivity of multimodal imaging-based DL in MVI diagnosis in external validation.
[PNG File, 313 KB - jmir_v28i1e82000_app13.png]

Multimedia Appendix 14
(A) Meta-analysis forest plots: specificity and sensitivity of pathological sections-based deep learning (DL) in microvascular invasion (MVI) diagnosis; (B) Meta-analysis summary receiver operating characteristic: specificity and sensitivity of pathological sections-based DL in MVI diagnosis.
[PNG File, 163 KB - jmir_v28i1e82000_app14.png]

Multimedia Appendix 15
Summary of bias risk for each study included in the paper according to the QUADAS-AI (artificial intelligence–specific quality assessment of diagnostic accuracy studies) domains.
[PDF File, 203 KB - jmir_v28i1e82000_app15.pdf]

Checklist 1
PRISMA-DTA checklist.
[DOCX File, 32 KB - jmir_v28i1e82000_app16.docx]

Checklist 2
PRISMA-S checklist.
[DOCX File, 19 KB - jmir_v28i1e82000_app17.docx]

References

1. Hwang SY, Danpanichkul P, Agopian V, et al. Hepatocellular carcinoma: updates on epidemiology, surveillance, diagnosis and treatment. Clin Mol Hepatol 2025 Feb;31(Suppl):S228-S254. [doi: 10.3350/cmh.2024.0824] [Medline: 39722614]
2. Brown ZJ, Tsilimigras DI, Ruff SM, et al. Management of hepatocellular carcinoma: a review. JAMA Surg 2023 Apr 1;158(4):410-420. [doi: 10.1001/jamasurg.2022.7989] [Medline: 36790767]

XSL•FO
RenderX

3. Vogel A, Meyer T, Sapisochin G, Salem R, Saborowski A. Hepatocellular carcinoma. Lancet 2022 Oct 15;400(10360):1345-1362. [doi: 10.1016/S0140-6736(22)01200-4] [Medline: 36084663]
4. Zhang ZH, Jiang C, Qiang ZY, et al. Role of microvascular invasion in early recurrence of hepatocellular carcinoma after liver resection: a literature review. Asian J Surg 2024 May;47(5):2138-2143. [doi: 10.1016/j.asjsur.2024.02.115] [Medline: 38443255]
5. Fuster-Anglada C, Mauro E, Ferrer-Fàbrega J, et al. Histological predictors of aggressive recurrence of hepatocellular carcinoma after liver resection. J Hepatol 2024 Dec;81(6):995-1004. [doi: 10.1016/j.jhep.2024.06.018] [Medline: 38925272]
6. Li SH, Mei J, Cheng Y, et al. Postoperative adjuvant hepatic arterial infusion chemotherapy with FOLFOX in hepatocellular carcinoma with microvascular invasion: a multicenter, phase III, randomized study. JCO 2023 Apr 1;41(10):1898-1908. [doi: 10.1200/JCO.22.01142]
7. Kang I, Jang M, Lee JG, et al. Subclassification of microscopic vascular invasion in hepatocellular carcinoma. Ann Surg 2021;274(6):e1170-e1178. [doi: 10.1097/SLA.0000000000003781]
8. Kim NR, Bae H, Hwang HS, et al. Preoperative prediction of microvascular invasion with gadoxetic acid-enhanced magnetic resonance imaging in patients with single hepatocellular carcinoma: the implication of surgical decision on the extent of liver resection. Liver Cancer 2024 Apr 24;13(2):181-192. [doi: 10.1159/000531786]
9. Ma L, Zhang C, Wen Y, et al. Imaging-based surrogate classification for risk stratification of hepatocellular carcinoma with microvascular invasion to adjuvant hepatic arterial infusion chemotherapy: a multicenter retrospective study. Int J Surg 2025 Jan 1;111(1):872-883. [doi: 10.1097/JS9.0000000000001903]
10. Lee S, Kang TW, Song KD, et al. Effect of microvascular invasion risk on early recurrence of hepatocellular carcinoma after surgery and radiofrequency ablation. Ann Surg 2021;273(3):564-571. [doi: 10.1097/SLA.0000000000003268]
11. Xu XF, Diao YK, Zeng YY, et al. Association of severity in the grading of microvascular invasion with long-term oncological prognosis after liver resection for early-stage hepatocellular carcinoma: a multicenter retrospective cohort study from a hepatitis B virus-endemic area. Int J Surg 2023 Apr 1;109(4):841-849. [doi: 10.1097/JS9.0000000000000325] [Medline: 36974673]
12. Wang L, Jia M, Wen X, Shen J, Yang H. Diagnostic value of magnetic resonance imaging features of microvascular invasion in hepatocellular carcinoma: a meta-analysis. Diagn Interv Radiol 2022 Sep;28(5):428-440. [doi: 10.5152/dir.2022.21108] [Medline: 36218148]
13. Xie XY, Chen R. Research progress of MRI-based radiomics in hepatocellular carcinoma. Front Oncol 2025;15:1420599. [doi: 10.3389/fonc.2025.1420599] [Medline: 39980543]
14. Zhu Z, Wu K, Lu J, et al. Gd-EOB-DTPA-enhanced MRI radiomics and deep learning models to predict microvascular invasion in hepatocellular carcinoma: a multicenter study. BMC Med Imaging 2025 Mar 31;25(1):40165094. [doi: 10.1186/s12880-025-01646-9]
15. Zhang L, Wang Y, Peng Z, et al. The progress of multimodal imaging combination and subregion based radiomics research of cancers. Int J Biol Sci 2022;18(8):3458-3469. [doi: 10.7150/ijbs.71046]
16. Zheng T, Zhu Y, Jiang H, et al. MRI-based topology deep learning model for noninvasive prediction of microvascular invasion and assisting prognostic stratification in HCC. Liver Int 2025 Mar;45(3):e16205. [doi: 10.1111/liv.16205] [Medline: 39992060]
17. Zhu Y, Wang P, Wang B, et al. Accelerated multi-b-value DWI using deep learning reconstruction: image quality improvement and microvascular invasion prediction in BCLC stage A hepatocellular carcinoma. Acad Radiol 2025 Jul;32(7):3924-3937. [doi: 10.1016/j.acra.2025.01.043]
18. McInnes MDF, Moher D, Thombs BD, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies. JAMA 2018 Jan 23;319(4):388. [doi: 10.1001/jama.2017.19163]
19. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc 2021 Apr 1;109(2):174-200. [doi: 10.5195/jmla.2021.962] [Medline: 34285662]
20. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005 Oct;58(10):982-990. [doi: 10.1016/j.jclinepi.2005.02.022]
21. Zhang X, Yu X, Liang W, et al. Deep learning-based accurate diagnosis and quantitative evaluation of microvascular invasion in hepatocellular carcinoma on whole-slide histopathology images. Cancer Med 2024 Mar;13(5):e7104. [doi: 10.1002/cam4.7104] [Medline: 38488408]
22. Lei Y, Feng B, Wan M, et al. Predicting microvascular invasion in hepatocellular carcinoma with a CT- and MRI-based multimodal deep learning model. Abdom Radiol 2024 May;49(5):1397-1410. [doi: 10.1007/s00261-024-04202-1]
23. Liu HF, Wang M, Lu YJ, et al. CEMRI-based quantification of intratumoral heterogeneity for predicting aggressive characteristics of hepatocellular carcinoma using habitat analysis: comparison and combination of deep learning. Acad Radiol 2024 Jun;31(6):2346-2355. [doi: 10.1016/j.acra.2023.11.024] [Medline: 38057182]
24. Zhou Z, Xia T, Zhang T, et al. Prediction of preoperative microvascular invasion by dynamic radiomic analysis based on contrast-enhanced computed tomography. Abdom Radiol 2024 Feb;49(2):611-624. [doi: 10.1007/s00261-023-04102-w]

XSL•FO
RenderX

25. He X, Xu Y, Zhou C, et al. Prediction of microvascular invasion and pathological differentiation of hepatocellular carcinoma based on a deep learning model. Eur J Radiol 2024 Mar;172:111348. [doi: 10.1016/j.ejrad.2024.111348] [Medline: 111348]
26. Zhong Y, Chen L, Ding F, Ou W, Zhang X, Weng S. Assessing microvascular invasion in HBV-related hepatocellular carcinoma: an online interactive nomogram integrating inflammatory markers, radiomics, and convolutional neural networks. Front Oncol 2024;14:1401095. [doi: 10.3389/fonc.2024.1401095] [Medline: 39351352]
27. Wang F, Zhan G, Chen QQ, et al. Multitask deep learning for prediction of microvascular invasion and recurrence-free survival in hepatocellular carcinoma based on MRI images. Liver Int 2024 Jun;44(6):1351-1362. [doi: 10.1111/liv.15870] [Medline: 38436551]
28. Yu Z, Liu Y, Dai X, Cui E, Cui J, Ma C. Enhancing preoperative diagnosis of microvascular invasion in hepatocellular carcinoma: domain-adaptation fusion of multi-phase CT images. Front Oncol 2024;14:1332188. [doi: 10.3389/fonc.2024.1332188] [Medline: 38333689]
29. Ma H, Wang L, Sun L, et al. Preoperative prediction of microvascular invasion in hepatocellular carcinoma from multi-sequence magnetic resonance imaging based on deep fusion representation learning. IEEE J Biomed Health Inform 2025 May;29(5):3259-3271. [doi: 10.1109/JBHI.2024.3451331] [Medline: 39196745]
30. Zhang W, Guo Q, Zhu Y, et al. Cross-institutional evaluation of deep learning and radiomics models in predicting microvascular invasion in hepatocellular carcinoma: validity, robustness, and ultrasound modality efficacy comparison. Cancer Imaging 2024 Oct 22;24(1):142. [doi: 10.1186/s40644-024-00790-9] [Medline: 39438929]
31. Wang F, Chen Q, Chen Y, et al. A novel multimodal deep learning model for preoperative prediction of microvascular invasion and outcome in hepatocellular carcinoma. Eur J Surg Oncol 2023 Jan;49(1):156-164. [doi: 10.1016/j.ejso.2022.08.036] [Medline: 36333180]
32. You H, Wang J, Ma R, et al. Clinical interpretability of deep learning for predicting microvascular invasion in hepatocellular carcinoma by using attention mechanism. Bioengineering (Basel) 2023 Aug 9;10(8):948. [doi: 10.3390/bioengineering10080948] [Medline: 37627833]
33. Qin X, Zhu J, Tu Z, Ma Q, Tang J, Zhang C. Contrast-enhanced ultrasound with deep learning with attention mechanisms for predicting microvascular invasion in single hepatocellular carcinoma. Acad Radiol 2023 Sep;30 Suppl 1:S73-S80. [doi: 10.1016/j.acra.2022.12.005] [Medline: 36567144]
34. Li Z, Wang Y, Zhu Y, et al. Modality-based attention and dual-stream multiple instance convolutional neural network for predicting microvascular invasion of hepatocellular carcinoma. Front Oncol 2023;13:1195110. [doi: 10.3389/fonc.2023.1195110] [Medline: 37434971]
35. Cao L, Wang Q, Hong J, et al. MVI-TR: a transformer-based deep learning model with contrast-enhanced CT for preoperative prediction of microvascular invasion in hepatocellular carcinoma. Cancers (Basel) 2023 Feb 28;15(5):1538. [doi: 10.3390/cancers15051538] [Medline: 36900327]
36. Wang T, Li Z, Yu H, et al. Prediction of microvascular invasion in hepatocellular carcinoma based on preoperative Gd-EOB-DTPA-enhanced MRI: comparison of predictive performance among 2D, 2D-expansion and 3D deep learning models. Front Oncol 2023;13:987781. [doi: 10.3389/fonc.2023.987781] [Medline: 36816963]
37. Ye Z, Zhang J, Wu N, et al. PET-guided attention for prediction of microvascular invasion in preoperative hepatocellular carcinoma on PET/CT. Ann Nucl Med 2023 Apr;37(4):238-245. [doi: 10.1007/s12149-023-01822-3] [Medline: 36723705]
38. Xu Y, Chen Y, Wu J, Pan J, Liao C, Su H. The utility of vision transformer in preoperatively predicting microvascular invasion status of hepatocellular carcinoma. HPB (Oxford) 2023 May;25(5):533-542. [doi: 10.1016/j.hpb.2023.01.015]
39. Deng Y, Jia X, Yu G, et al. Can a proposed double branch multimodality-contribution-aware TripNet improve the prediction performance of the microvascular invasion of hepatocellular carcinoma based on small samples? Front Oncol 2022;12:1035775. [doi: 10.3389/fonc.2022.1035775] [Medline: 36387069]
40. Li X, Qi Z, Du H, et al. Deep convolutional neural network for preoperative prediction of microvascular invasion and clinical outcomes in patients with HCCs. Eur Radiol 2022 Feb;32(2):771-782. [doi: 10.1007/s00330-021-08198-w]
41. Chen Q, Xiao H, Gu Y, et al. Deep learning for evaluation of microvascular invasion in hepatocellular carcinoma from tumor areas of histology images. Hepatol Int 2022 Jun;16(3):590-602. [doi: 10.1007/s12072-022-10323-w]
42. Zhang Y, Wei Q, Huang Y, et al. Deep learning of liver contrast-enhanced ultrasound to predict microvascular invasion and prognosis in hepatocellular carcinoma. Front Oncol 2022;12:878061. [doi: 10.3389/fonc.2022.878061] [Medline: 35875110]
43. Liu B, Zeng Q, Huang J, et al. IVIM using convolutional neural networks predicts microvascular invasion in HCC. Eur Radiol 2022 Oct;32(10):7185-7195. [doi: 10.1007/s00330-022-08927-9]
44. Sun SW, Xu X, Liu QP, et al. LiSNet: an artificial intelligence - based tool for liver imaging staging of hepatocellular carcinoma aggressiveness. Med Phys Mex Symp Med Phys 2022 Nov;49(11):6903-6913. [doi: 10.1002/mp.15972]
45. Wang L, Wu M, Li R, Xu X, Zhu C, Feng X. MVI-Mind: a novel deep-learning strategy using computed tomography (CT)-based radiomics for end-to-end high efficiency prediction of microvascular invasion in hepatocellular carcinoma. Cancers (Basel) 2022;14(12):2956. [doi: 10.3390/cancers14122956]
46. Xiao H, Guo Y, Zhou Q, et al. Prediction of microvascular invasion in hepatocellular carcinoma with expert-inspiration and skeleton sharing deep learning. Liver Int 2022 Jun;42(6):1423-1431. [doi: 10.1111/liv.15254] [Medline: 35319151]

XSL•FO
RenderX

47. Yang Y, Zhou Y, Zhou C, Ma X. Deep learning radiomics based on contrast enhanced computed tomography predicts microvascular invasion and survival outcome in early stage hepatocellular carcinoma. Eur J Surg Oncol 2022 May;48(5):1068-1077. [doi: 10.1016/j.ejso.2021.11.120]
48. Sun BY, Gu PY, Guan RY, et al. Deep-learning-based analysis of preoperative MRI predicts microvascular invasion and outcome in hepatocellular carcinoma. World J Surg Oncol 2022 Jun 8;20(1):189. [doi: 10.1186/s12957-022-02645-8] [Medline: 35676669]
49. Dai X, Cao S, Guo Y, et al. Preoperative identification of microvascular invasion in hepatocellular carcinoma based on multi-modal and multi response convolutional neural network. Presented at: Fourteenth International Conference on Graphics and Image Processing (ICGIP 2022); Oct 21-23, 2022. [doi: 10.1117/12.2680693]
50. Zhang Y, Lv X, Qiu J, et al. Deep learning with 3D convolutional neural network for noninvasive prediction of microvascular invasion in hepatocellular carcinoma. Magn Reson Imaging 2021 Jul;54(1):134-143. [doi: 10.1002/jmri.27538]
51. Liu SC, Lai J, Huang JY, et al. Predicting microvascular invasion in hepatocellular carcinoma: a deep learning model validated across hospitals. Cancer Imaging 2021 Oct 9;21(1):56. [doi: 10.1186/s40644-021-00425-3] [Medline: 34627393]
52. Wei J, Jiang H, Zeng M, et al. Prediction of microvascular invasion in hepatocellular carcinoma via deep learning: a multi-center and prospective validation study. Cancers (Basel) 2021 May 14;13(10):2368. [doi: 10.3390/cancers13102368]
53. Zhou W, Jian W, Cen X, et al. Prediction of microvascular invasion of hepatocellular carcinoma based on contrast-enhanced MR and 3D convolutional neural networks. Front Oncol 2021;11:588010. [doi: 10.3389/fonc.2021.588010] [Medline: 33854959]
54. Wang G, Jian W, Cen X, et al. Prediction of microvascular invasion of hepatocellular carcinoma based on preoperative diffusion-weighted MR using deep learning. Acad Radiol 2021 Nov;28:S118-S127. [doi: 10.1016/j.acra.2020.11.014]
55. Zeng Q, Liu B, Xu Y, Zhou W. An attention-based deep learning model for predicting microvascular invasion of hepatocellular carcinoma using an intra-voxel incoherent motion model of diffusion-weighted magnetic resonance imaging. Phys Med Biol 2021 Sep 21;66(18):185019. [doi: 10.1088/1361-6560/ac22db]
56. Gao F, Qiao K, Yan B, et al. Hybrid network with difference degree and attention mechanism combined with radiomics (H-DARnet) for MVI prediction in HCC. Magn Reson Imaging 2021 Nov;83:27-40. [doi: 10.1016/j.mri.2021.06.018] [Medline: 34147593]
57. Jiang YQ, Cao SE, Cao S, et al. Preoperative identification of microvascular invasion in hepatocellular carcinoma by XGBoost and deep learning. J Cancer Res Clin Oncol 2021 Mar;147(3):821-833. [doi: 10.1007/s00432-020-03366-9] [Medline: 32852634]
58. Song D, Wang Y, Wang W, et al. Using deep learning to predict microvascular invasion in hepatocellular carcinoma based on dynamic contrast-enhanced MRI combined with clinical parameters. J Cancer Res Clin Oncol 2021 Dec;147(12):3757-3767. [doi: 10.1007/s00432-021-03617-3] [Medline: 33839938]
59. Men S, Ju H, Zhang L, Zhou W. Prediction of microvascular invasion of hepatocellar carcinoma with contrast-enhanced MR using 3D CNN and LSTM. Presented at: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI); Oct 8-11, 2019. [doi: 10.1109/ISBI.2019.8759582]
60. Zhou Y, Sun SW, Liu QP, Xu X, Zhang Y, Zhang YD. TED: two-stage expert-guided interpretable diagnosis framework for microvascular invasion in hepatocellular carcinoma. Med Image Anal 2022 Nov;82:102575. [doi: 10.1016/j.media.2022.102575] [Medline: 36063747]
61. Chu T, Zhao C, Zhang J, et al. Application of a convolutional neural network for multitask learning to simultaneously predict microvascular invasion and vessels that encapsulate tumor clusters in hepatocellular carcinoma. Ann Surg Oncol 2022 Oct;29(11):6774-6783. [doi: 10.1245/s10434-022-12000-6] [Medline: 35754067]
62. Huang H, Liu B, Zhang L, Xu Y, Zhou W, editors. Transformer Based Multi-Task Deep Learning with Intravoxel Incoherent Motion Model Fitting for Microvascular Invasion Prediction of Hepatocellular Carcinoma: Springer Nature Switzerland; 2022. [doi: 10.1007/978-3-031-16449-1_26]
63. Wang Y, Xie W, Li C, et al. Automated microvascular invasion prediction of hepatocellular carcinoma via deep relation reasoning from dynamic contrast-enhanced ultrasound. Comput Med Imaging Graph 2025 Sep;124:102606. [doi: 10.1016/j.compmedimag.2025.102606] [Medline: 40680384]
64. Cen YY, Nong HY, Huang XX, et al. Computed tomography-based deep learning and multi-instance learning for predicting microvascular invasion and prognosis in hepatocellular carcinoma. World J Gastroenterol 2025 Aug 14;31(30):109186. [doi: 10.3748/wjg.v31.i30.109186] [Medline: 40933208]
65. Huang Z, Huang W, Jiang L, et al. Decision fusion model for predicting microvascular invasion in hepatocellular carcinoma based on multi-MR habitat imaging and machine-learning classifiers. Acad Radiol 2025 Apr;32(4):1971-1980. [doi: 10.1016/j.acra.2024.10.007] [Medline: 39472207]
66. Miao S, Sun M, Li X, et al. Deep learning-based prediction of microvascular invasion and survival outcomes in hepatocellular carcinoma using dual-phase CT imaging of tumors and lesser omental adipose: a multicenter study. Acad Radiol 2025 Oct;32(10):5789-5801. [doi: 10.1016/j.acra.2025.07.015]
67. Dong X, Jia X, Zhang W, et al. Interpretable and generalizable deep learning model for preoperative assessment of microvascular invasion and outcome in hepatocellular carcinoma based on MRI: a multicenter study. Insights Imaging 2025 Jul 3;16(1):151. [doi: 10.1186/s13244-025-02035-0] [Medline: 40610844]

XSL•FO
RenderX

68. Zhang Y, Liu H, Zhu L, et al. Modality-aware distillation network for microvascular invasion prediction of hepatocellar carcinoma from MRI images. IEEE Trans Biomed Eng 2025 Jun;72(6):1825-1836. [doi: 10.1109/TBME.2024.3523921]
69. Zhao Y, Huang X, Sun M, et al. Predicting microvascular invasion plus cytokeratin 19 expression positivity in hepatocellular carcinoma based on EOB-MRI using multitask deep learning. Hepatoma Res 2025;10:12. [doi: 10.20517/2394-5079.2024.143]
70. Qin Q, Pang J, Li J, et al. Transformer model based on Sonazoid contrast-enhanced ultrasound for microvascular invasion prediction in hepatocellular carcinoma. Med Phys 2025 Jul;52(7):e17895. [doi: 10.1002/mp.17895] [Medline: 40384312]
71. Xiao Q, Zhu W, Tang H, Zhou L. Ultrasound radiomics in the prediction of microvascular invasion in hepatocellular carcinoma: a systematic review and meta-analysis. Heliyon 2023 Jun;9(6):e16997. [doi: 10.1016/j.heliyon.2023.e16997]
72. Liang G, Yu W, Liu S, et al. The diagnostic performance of radiomics-based MRI in predicting microvascular invasion in hepatocellular carcinoma: a meta-analysis. Front Oncol 2022;12:960944. [doi: 10.3389/fonc.2022.960944] [Medline: 36798691]
73. Li L, Wu C, Huang Y, Chen J, Ye D, Su Z. Radiomics for the preoperative evaluation of microvascular invasion in hepatocellular carcinoma: a meta-analysis. Front Oncol 2022;12:831996. [doi: 10.3389/fonc.2022.831996] [Medline: 35463303]
74. He K, Liu X, Shahzad R, et al. Advanced deep learning approach to automatically segment malignant tumors and ablation zone in the liver with contrast-enhanced CT. Front Oncol 2021;11:669437. [doi: 10.3389/fonc.2021.669437] [Medline: 34336661]
75. Kinoshita M, Ueda D, Matsumoto T, et al. Deep learning model based on contrast-enhanced computed tomography imaging to predict postoperative early recurrence after the curative resection of a solitary hepatocellular carcinoma. Cancers (Basel) 2023 Apr 4;15(7):2140. [doi: 10.3390/cancers15072140] [Medline: 37046801]
76. Lu D, Wang LF, Han H, et al. Prediction of microvascular invasion in hepatocellular carcinoma with conventional ultrasound, Sonazoid-enhanced ultrasound, and biochemical indicator: a multicenter study. Insights Imaging 2024 Oct 28;15(1):261. [doi: 10.1186/s13244-024-01743-3] [Medline: 39466459]
77. Yan M, Zhang X, Zhang B, et al. Deep learning nomogram based on Gd-EOB-DTPA MRI for predicting early recurrence in hepatocellular carcinoma after hepatectomy. Eur Radiol 2023 Jul;33(7):4949-4961. [doi: 10.1007/s00330-023-09419-0] [Medline: 36786905]
78. Wu Y, Zhu M, Liu Y, Cao X, Zhang G, Yin L. Peritumoral imaging manifestations on Gd-EOB-DTPA-enhanced MRI for preoperative prediction of microvascular invasion in hepatocellular carcinoma: a systematic review and meta-analysis. Front Oncol 2022;12:907076. [doi: 10.3389/fonc.2022.907076] [Medline: 35814461]
79. Li C, Tan Y, Chen W, et al. Attention unet++: a nested attention-aware u-net for liver CT image segmentation. 2020 Presented at: 2020 IEEE International Conference on Image Processing (ICIP); Oct 25-28, 2020; Abu Dhabi, United Arab Emirates p. 345-349. [doi: 10.1109/ICIP40778.2020.9190761]
80. Huang H, Lin L, Tong R, et al. UNet 3+: a full-scale connected unet for medical image segmentation. Presented at: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 4-8, 2020; Barcelona, Spain p. 1055-1059. [doi: 10.1109/ICASSP40776.2020.9053405]
81. Bilic P, Christ P, Li HB, et al. The liver tumor segmentation benchmark (LiTS). Med Image Anal 2023 Feb;84:102680. [doi: 10.1016/j.media.2022.102680] [Medline: 36481607]
82. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology 2016 Feb;278(2):563-577. [doi: 10.1148/radiol.2015151169]
83. Riley RD, Archer L, Snell KIE, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ 2024 Jan 15;384:e074820. [doi: 10.1136/bmj-2023-074820]
84. Collins GS, Dhiman P, Ma J, et al. Evaluation of clinical prediction models (part 1): from development to external validation. BMJ 2024 Jan 8;384:e074819. [doi: 10.1136/bmj-2023-074819]
85. Strandberg R, Jepsen P, Hagström H. Developing and validating clinical prediction models in hepatology - an overview for clinicians. J Hepatol 2024 Jul;81(1):149-162. [doi: 10.1016/j.jhep.2024.03.030] [Medline: 38531493]
86. Maung ST, Tangkijvanich P, Chaiteerakij R. Magnetic resonance imaging radiomics for predicting hepatocellular carcinoma recurrence following resection or ablation: a systematic review and meta-analysis. Abdom Radiol (NY) 2025 Sep 22. [doi: 10.1007/s00261-025-05184-4] [Medline: 40981987]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
CECT: contrast-enhanced computed tomography
CEMRI: contrast-enhanced magnetic resonance imaging
CEUS: contrast-enhanced ultrasound
DL: deep learning
DOR: diagnostic odds ratio
HCC: hepatocellular carcinoma

XSL•FO
RenderX

LFK: Luis Furuya-Kanamori
LR: likelihood ratio
MeSH: Medical Subject Headings
MRI: magnetic resonance imaging
MVI: microvascular invasion
PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Diagnostic Test Accuracy
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for literature searches
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2
QUADAS-AI: Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence
ROB: risk of bias
SROC: summary receiver operating characteristic

Edited by S Brini; submitted 07.Aug.2025; peer-reviewed by L Zhou, ST Maung; accepted 08.Jan.2026; published 02.Mar.2026.

Please cite as:
Feng W, Qu B, Han S
Accuracy of Medical Image–Based Deep Learning for Detecting Microvascular Invasion in Hepatocellular Carcinoma: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e82000
URL: https://www.jmir.org/2026/1/e82000
doi:10.2196/82000

© Wei Feng, Bo Qu, Shuo Han. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Predictive Value of Machine Learning for Postoperative Delirium in Cardiac Surgery: Systematic Review and Meta-Analysis

Yi Guo[1], MS; Hong Xu[2], BS; Ankui Wang[1], BS; Mingming Zhang[1], MS; Shuai Zhang[1], BS; Peng Xie[3], MS

[1]Department of Anesthesiology, Norinco General Hospital, Xi'an, China

[2]Department of Operating Room, Norinco General Hospital, Xi'an, China

[3]Department of Hepatobiliary and Vascular Surgery, Norinco General Hospital, No. 12, Zhangba East Road, Yanta District, Xi'an, China

Corresponding Author:
Peng Xie, MS
Department of Hepatobiliary and Vascular Surgery, Norinco General Hospital, No. 12, Zhangba East Road, Yanta District, Xi'an, China

Abstract

Background: Postoperative delirium (POD) following cardiac surgery is a severe complication, and early identification of delirium risk remains a challenge in clinical practice. While machine learning (ML) has garnered increasing attention in health care applications, effective early prediction tools remain limited in current clinical practice. Recent investigations have explored the effectiveness of ML-based methods for identifying the risk of POD in patients undergoing cardiac surgery. However, more evidence is required to validate the feasibility of these methods.

Objectives: This study aims to ascertain the performance of ML in identifying the risk of POD following cardiac surgery, providing evidence for the development or updating of future ML-based assessment tools.

Methods: A comprehensive literature search was conducted across 4 databases—PubMed, the Cochrane Library, Embase, and Web of Science—through August 30, 2024, to identify studies investigating individual POD risk prediction using ML approaches and nomograms. The risk of bias of the models in the included studies was assessed leveraging the Prediction Model Bias Risk Assessment Tool. Subgroup analyses were performed based on datasets, validation methods, study types, risk of bias, and model types.

Results: The analysis incorporated 28 original studies comprising 80,143 patients undergoing cardiac surgery, of whom 6326 developed POD. Meta-analysis revealed that, in validation datasets, the c-index, sensitivity, and specificity for delirium prediction reached 0.805 (95% CI 0.759 - 0.852), 0.72 (95% CI 0.65 - 0.79), and 0.78 (95% CI 0.71 - 0.83), respectively. Logistic regression was the primary modeling method. In validation datasets, the c-index, sensitivity, and specificity reached 0.773 (95% CI 0.724 - 0.823), 0.73 (95% CI 0.64 - 0.80), and 0.70 (95% CI 0.65 - 0.74), respectively.

Conclusions: ML-based prediction tools for POD following cardiac surgery demonstrate promising performance. However, the limited number of studies and validation approaches necessitate cautious interpretation of these findings. Future multicenter studies are warranted to develop more robust ML-based prediction tools, enabling precise risk stratification and targeted preventive interventions for POD.

Trial Registration: PROSPERO CRD42024588522; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024588522

(J Med Internet Res 2026;28:e72304) doi:10.2196/72304

KEYWORDS

cardiac surgery; delirium; machine learning; predictive model; systematic review

Introduction

The number of deaths from cardiovascular disease far exceeds those from cancer [1], and cardiac surgery is often required for its treatment. Common types of cardiac surgery include congenital heart disease surgery, coronary artery bypass grafting, cardiac valve surgery, and ascending aorta and arch surgery. Although cardiac surgery has benefitted a large number of patients, postoperative complications remain a significant clinical challenge. Common complications include excessive bleeding, refractory shock, postoperative cardiac arrest, neurologic injury, respiratory failure, and acute respiratory distress syndrome [2], as well as gastrointestinal complications [3]. Delirium, characterized by acute changes in mental status with confusion and impaired attention, is one of the most common and serious neurological complications following cardiac surgery, with reported incidence rates exceeding 50% [4,5]. A study by Salluh et al [6] found that delirium correlates with higher in-hospital mortality rates, extended hospital stays, and postdischarge cognitive impairment. Therefore, early

XSL•FO
RenderX

prediction of delirium risk and the development of effective, targeted preventive strategies are of considerable clinical importance.

Currently, there is no universally recognized or effective tool for predicting postoperative delirium (POD) in cardiac surgery. Machine learning (ML) can integrate high-dimensional data to develop and validate clinical prediction tools for disease prediction, diagnosis, treatment response, prognosis, and adverse events. Several researchers have constructed predictive models for POD using ML techniques [7-9]; however, the accuracy of these models requires further validation. In the Shining Cai study [7], prediction models for postcardiac surgery delirium were summarized, but the number of included studies was limited, possibly due to the search strategy. Both Lee et al [8] and Chen et al [9] focused on POD prediction models and demonstrated promising predictive value. However, the study by Lee et al [8] specifically focused on delirium occurring in the intensive care unit after cardiac surgery, whereas Chen et al [9] examined POD without specifying cardiac surgery. Therefore, this study aims to evaluate the effectiveness of ML-based predictive models for delirium after cardiac surgery and to inform the future development or refinement of simplified clinical scoring tools.

Methods

Study Registration

This study adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines and was registered with PROSPERO (registration number CRD42024588522).

Inclusion and Exclusion Criteria

Eligibility criteria were defined before literature screening (Table 1).

Table . Eligibility criteria for original studies included in this systematic review.

Items	Inclusion criteria	Exclusion criteria
P (Population)	Patients who underwent cardiac surgery, including CABG[a], OPCABG[b], valve surgery, ascending aorta and arch surgery, and minimally invasive procedures such as TAVR[c]	Studies that did not clearly distinguish cardiac surgery from other surgical procedures
I (Intervention)	Studies that developed a complete predictive model for POD[d] after cardiac surgery	Studies limited to risk factor identification without a complete ML[e] model, or studies that assessed only univariate predictive accuracy
C (Control)	None	None
O (Outcomes)	Evaluation of ML model performance using metrics such as c-index, sensitivity (Sen), specificity (Spe), accuracy, precision, diagnostic contingency tables, or F_1-score	Studies lacking outcome measures for model accuracy
S (Study design)	Case-control, cohort, or cross-sectional studies published in English	Studies evaluating only established scales (for studies using overlapping datasets, only the study with the largest sample size was retained)

[a]CABG: coronary artery bypass graft.

[b]OPCABG: off-pump coronary artery bypass grafting.

[c]TAVR: transcatheter aortic valve replacement.

[d]POD: postoperative delirium.

[e]ML: machine learning.

Data Sources and Search Strategy

A comprehensive literature search was executed in 4 databases: PubMed, the Cochrane Library, Embase, and Web of Science, through August 30, 2024. Both Medical Subject Headings and free-text terms were used for the search. No restrictions were imposed on regions and year of publication (Table S1 in Multimedia Appendix 1).

Study Selection and Data Extraction

The retrieved records were imported into EndNote, and after removing duplicates, titles and abstracts were reviewed. Full texts of potentially relevant articles were obtained and checked. To reduce population heterogeneity, this review focused exclusively on cardiac surgery. Because some relevant studies did not explicitly mention cardiac surgery in the title or abstract, the term “cardiac surgery” was not included as a search keyword; instead, noncardiac surgical studies were excluded manually during the screening process.

A standardized digital form was leveraged for data collection. Extracted information included title, DOI, authors, year of publication, study design, patient source, type of cardiac surgery, diagnostic criteria for delirium, number of patients with delirium, total number of cardiac surgery patients, number of delirium patients in both training and validation sets, validation set formation method, total number of patients in the validation set, missing data and methods for handling missing data, variable screening methods, modeling methods, and predictive factors.

XSL•FO
RenderX

Two researchers independently reviewed the literature and cross-checked the extracted data, with a third investigator resolving any disagreements.

Risk of Bias in Included Studies

Risk of bias in the included studies was appraised using the Prediction Model Risk of Bias Assessment Tool (PROBAST) [10], which included questions across domains including participants, predictors, outcomes, and statistical analysis. These domains reflected the overall risk of bias and applicability of the studies. Specific questions within domains had 3 possible responses: yes/likely, no/likely not, and no information. If the risk of bias in all domains was low for a study, the study was marked as low risk. If the risk of bias in at least a domain was high, the study was graded to have a high risk of bias. Two researchers independently performed and cross-checked the risk of bias assessments, with a third researcher resolving any disagreements.

Statistical Analysis

A meta-analysis was performed to evaluate the overall performance of ML models using the c-index. For studies lacking CIs or SEs of c-index, SEs were estimated according to the study by Debray et al [11]. The heterogeneity between studies was quantified by leveraging the I^2 statistic. If I^2 was greater than 50%, a random-effects model was utilized. If I^2 was less than 50%, a fixed-effects model was applied. Funnel plots were employed to detect publication bias of the c-index. The Egger test was applied for statistical assessment of publication bias.

A bivariate mixed-effects model was used to synthesize sensitivity and specificity data. For studies without diagnostic contingency tables, these were reconstructed using sensitivity, specificity, precision, and case numbers for calculations. Subgroup analyses were performed based on validation methods, study types, risk of bias, and model types. The statistical computation was conducted using Stata version 15.0.

Results

Study Selection

The initial search yielded 5868 articles. After the removal of 1981 duplicates, 3760 studies were excluded based on titles and abstracts because they were unrelated to the research theme or inconsistent with the original study design. A total of 127 studies underwent full-text review. Of these, 10 unpublished conference abstracts, 25 studies limited to risk factor analysis without construction of a complete prediction model, 62 studies that failed to clearly distinguish cardiac surgery from noncardiac surgery, and 2 studies with overlapping datasets were excluded. Eventually, 28 studies [12-39] were included in the final analysis (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the study selection process.

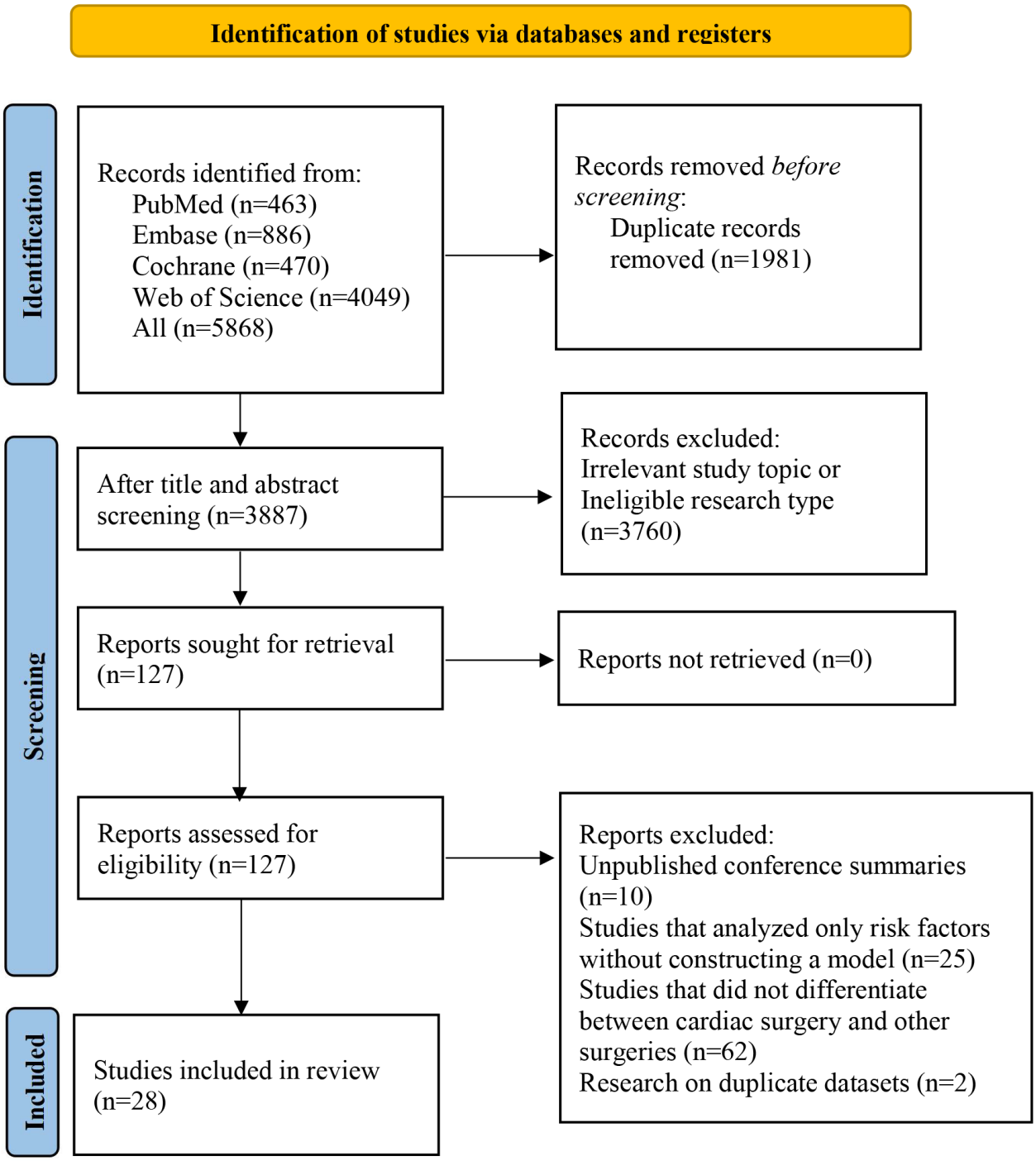

Study Characteristics

The analysis included 28 publications from 2012 to 2024. The studies collectively involved 80,143 patients undergoing cardiac surgery, of whom 6326 developed delirium. These studies included 12 retrospective cohort studies, 14 prospective cohort studies, and 2 retrospective (for model development) and prospective (for model validation) cohort studies. These studies spanned 12 countries, with 13 studies originating from China. Five studies were multicenter, one was based on a registry database, and 23 studies were single-center.

The studies included coronary artery bypass grafting, valve surgery, congenital heart surgery (patients aged ≤8 y), cardiac tumor resection, aortic arch repairs, and type A aortic dissection treated with open-heart surgery. Most studies described multiple types of cardiac surgery, whereas only a few studies focused on a single surgery type. All 28 studies clearly outlined the diagnostic methods or steps for identifying delirium. To minimize population heterogeneity, the 2 pediatric studies and the remaining 26 adult studies were analyzed separately. Among the 26 adult studies, 21 included a validation set and described its generation method, comprising 1 external validation, 4 temporal validations, 11 random split validations, 3 k-fold

XSL•FO
RenderX

cross-validations, and 2 bootstrap resampling validations. One study performed temporal validation but did not report validation set results due to inadequate sample size. To avoid sample size inflation, only the best-performing model from each study was selected, resulting in 9 model types, with logistic regression models being the most prevalent (Table S2 in Multimedia Appendix 2) [12-39].

In addition, the variables used for modeling were primarily clinical features. The modeling variables are presented in Table S3 in Multimedia Appendix 3.

Risk of Bias Assessment

Risk of bias was assessed for the 28 models from the 28 studies. In the participant domain, some studies had only validated a single model, which contributed to potential bias. Twelve models were from retrospective cohort studies, resulting in a high risk of bias in participant selection. Given that all studies constructed models based on clinical features, the retrospective cohort models may introduce bias in assessing clinical characteristics, thereby increasing the risk of bias in the prediction factors as well. All included studies used established diagnostic criteria for delirium, and no prediction factors that required delirium assessment were incorporated into the modeling process, indicating low bias risk in outcome assessment.

In the predictors domain, only one study was rated as high risk of bias due to the incorporation of outcome information during predictor assessment, whereas all other studies demonstrated low risk of bias.

In the outcomes domain, none of the studies explicitly reported blinding; consequently, all models were rated as having unclear risk of bias for this domain.

Fifteen models had a high risk of bias from insufficient events per variable (EPV <20) or lack of independent validation. Twenty-three models showed a high risk of bias in missing data handling, including direct deletion of missing values. Two studies showed unclear risk of bias due to inadequate description of predictor selection methods. Seventeen models demonstrated a high risk of bias for inadequate assessment of model fit or overfitting [40]. Overall, 26 models were identified as high-risk and 2 low-risk prediction models in statistical analysis (Figure 2).

Figure 2. Risk of bias of the included predictive models assessed with the Prediction Model Risk of Bias Assessment Tool.

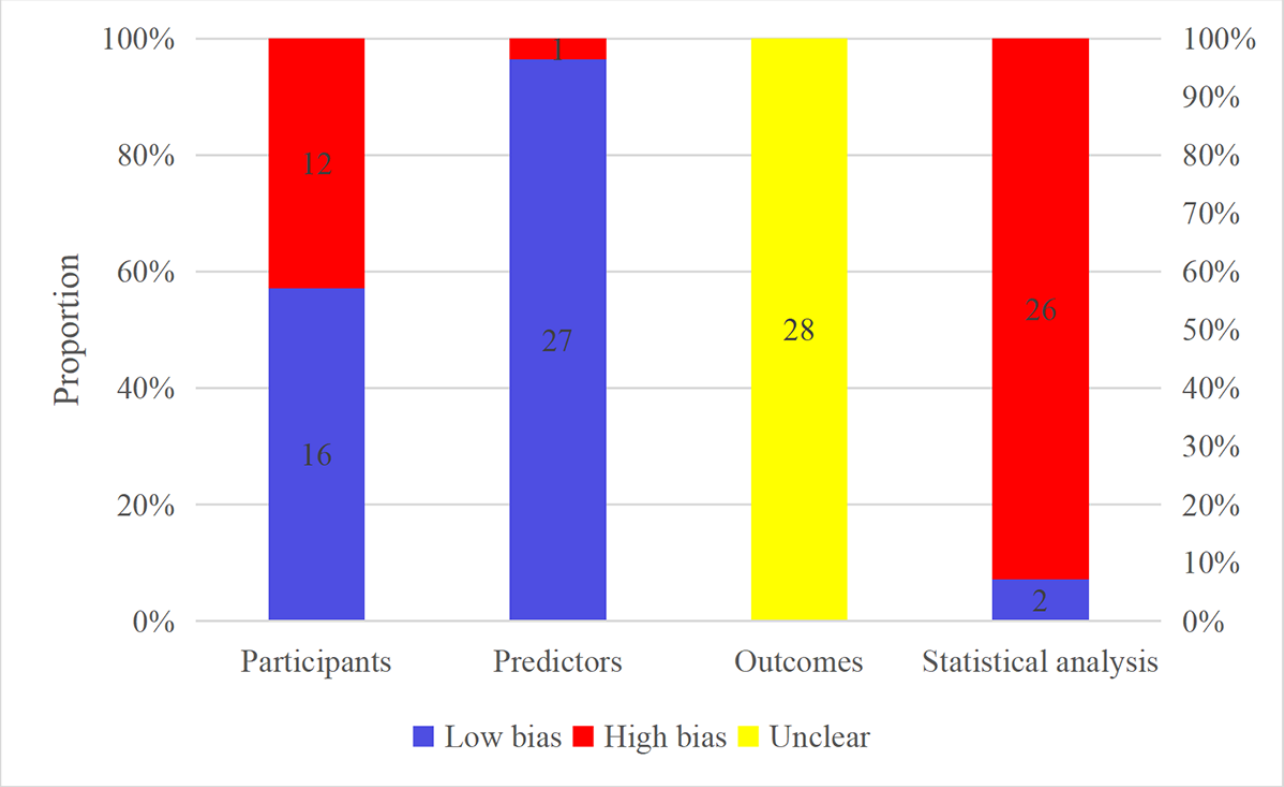

Meta-Analysis

C-Index

Because training set data were unavailable from 9 studies, only 17 prediction models for POD following cardiac surgery were reported in the training set. A random-effects model was leveraged, and the combined c-index was 0.815 (95% CI 0.754 - 0.876). For validation, 20 models were reported for validating the predictive accuracy of POD following cardiac surgery. The pooled c-index, also derived from a random-effects model, was 0.805 (95% CI 0.759 - 0.852) (Table 2).

XSL•FO
RenderX

Table . Meta-analysis results of c-index in the included studies for machine learning prediction of postoperative delirium following cardiac surgery in adults.

Subgroups	Training set			Validation set		
	Studies, n	c-index (95% CI)	I^2 (%)	Studies, n	c-index (95% CI)	I^2 (%)
Validation methods						
Random split validation	7	0.769 (0.692 - 0.847)	97.6	11	0.791 (0.722 - 0.860)	98.4
K-fold cross-validation	__[a]	—	—	3	0.845 (0.769 - 0.921)	73.9
Bootstrap resampling validation	—	—	—	2	0.805 (0.745 - 0.866)	47.1
External validation	1	0.740 (0.665 - 0.815)	—	1	0.750 (0.667 - 0.833)	—
Temporal validation	4	0.932 (0.895 - 0.969)	94.2	3	0.848 (0.807 - 0.889)	16.8
No validation	5	0.801 (0.744 - 0.857)	74.8	—	—	—
Study types						
Retrospective cohort study	6	0.819 (0.696 - 0.943)	99.5	10	0.803 (0.734 - 0.871)	98.4
Prospective cohort study	9	0.809 (0.758 - 0.860)	88.7	8	0.817 (0.762 - 0.872)	82.6
Retrospective and prospective cohort study	2	0.832 (0.599 - 1.065)	98.7	2	0.789 (0.621 - 0.958)	93.8
Risk of bias (statistic analysis)						
Low	1	0.740 (0.665 - 0.815)	—	1	0.750 (0.667 - 0.833)	—
High	16	0.819 (0.756 - 0.882)	98.9	19	0.808 (0.761 - 0.856)	97.4
Model types						
ANN[b]	2	0.867 (0.744 - 0.989)	98.5	2	0.781 (0.742 - 0.820)	0
DT[c]	—	—	—	1	0.930 (0.885 - 0.975)	—
KNN[d]	—	—	—	1	0.901 (0.866 - 0.935)	—
LASSO[e]	1	0.741 (0.678 - 0.804)	—	1	0.640 (0.531 - 0.749)	—
LR[f]	13	0.801 (0.715 - 0.887)	98.9	11	0.773 (0.724 - 0.823)	91.1
RF[g]	—	—	—	1	0.920 (0.915 - 0.925)	—
SVM[h]	—	—	—	1	0.941 (0.919 - 0.963)	—
XGBoost[i]	—	—	—	1	0.760 (0.651 - 0.869)	—
LGBM[j]	1	0.950 (0.935 - 0.965)	—	1	0.877 (0.808 - 0.946)	—
Overall	17	0.815 (0.754 - 0.876)	98.8	20	0.805 (0.759 - 0.852)	97.3

XSL•FO
RenderX

[a]Not applicable.
[b]ANN: artificial neural network.
[c]DT: decision tree.
[d]KNN: k-nearest neighbor.
[e]LASSO: least absolute shrinkage and selection operator.
[f]LR: logistic regression.
[g]RF: random forest.
[h]SVM: support vector machine.
[i]XGBoost: extreme gradient boosting.
[j]LightGBM: light gradient boosting machine.

Subgroup analysis results are also presented in Table 2. Of note, Subgroup analysis by model type indicated that most models were based on logistic regression. This approach achieved a c-index of 0.801 (95% CI 0.715 - 0.887) in the training set and 0.773 (95% CI 0.724 - 0.823) in the validation set. These values were lower than the overall c-index, suggesting that the performance of logistic regression models may be less effective than non-logistic regression models.

Notably, subgroup analyses revealed that heterogeneity remained extremely high across most subgroups and overall (I^2>90%), indicating substantial unexplained variability among performance estimates. This extreme heterogeneity suggests that while pooled c-indices were calculated, their clinical interpretability is limited; these values should be considered a broad approximation of possible outcomes rather than precise summary effect estimates.

Sensitivity and Specificity

In the subgroup analysis by model types, diagnostic 2 × 2 tables from 13 models were available in the training set, and 18 models were available in the validation set. A bivariate random-effects model was leveraged. The majority of models were based on logistic regression, with pooled sensitivity and specificity of 0.75 (95% CI 0.62 - 0.85) and 0.79 (95% CI 0.72 - 0.85) in the training set, and 0.73 (95% CI 0.64 - 0.80) and 0.70 (95% CI 0.65 - 0.74) in the validation set. Sensitivity and specificity results for other subgroups are presented in Table 3.

XSL•FO
RenderX

Table . Meta-analysis results of sensitivity and specificity in the included studies for machine learning prediction of postoperative delirium following cardiac surgery in adults.[a]

Subgroups	Training set			Validation set		
	Studies, n	Sensitivity (95% CI) or range	Specificity (95% CI) or range	Studies, n	Sensitivity (95% CI) or range	Specificity (95% CI) or range
Verification methods						
Random split validation	7	0.68 (0.56 - 0.78)	0.77 (0.69 - 0.83)	10	0.72 (0.62 - 0.80)	0.76 (0.67 - 0.82)
K-fold cross-validation	—[b]	—	—	3	0.67 - 0.86	0.72 - 0.91
Bootstrap resampling validation	—	—	—	2	0.71 - 0.91	0.67 - 0.68
External validation						
Temporal validation	3	0.72 - 0.97	0.80 - 0.93	3	0.35 - 0.79	0.68 - 0.96
No validation	3	0.66 - 0.81	0.66 - 0.82	—	—	—
Study types						
Retrospective cohort study	5	0.80 (0.52 - 0.93)	0.82 (0.72 - 0.89)	10	0.74 (0.66 - 0.80)	0.77 (0.69 - 0.84)
Prospective cohort study	7	0.76 (0.70 - 0.82)	0.78 (0.72 - 0.84)	6	0.77 (0.67 - 0.85)	0.70 (0.65 - 0.75)
Retrospective and prospective cohort study	1	0.64	0.64	2	0.35 - 0.53	0.77 - 0.96
Risk of bias (statistic analysis)						
Low	—	—	—	—	—	—
High	13	0.76 (0.66 - 0.84)	0.79 (0.73 - 0.84)	18	0.72 (0.65 - 0.79)	0.78 (0.71 - 0.83)
Model types						
ANN[c]	2	0.72 - 0.91	0.72 - 0.86	2	0.68 - 0.73	0.73 - 0.78
KNN[d]	—	—	—	1	0.67	0.91
LASSO[e]	1	0.72	0.73	1	0.60	0.62
LR[f]	10	0.75 (0.62 - 0.85)	0.79 (0.72 - 0.85)	10	0.73 (0.64 - 0.80)	0.70 (0.65 - 0.74)
RF[g]	—	—	—	1	0.86	0.92
SVM[h]	—	—	—	1	0.91	0.87
XGBoost[i]	—	—	—	1	0.67	0.79
LGBM[j]	—	—	—	1	0.35	0.96
Overall	13	0.76 (0.66 - 0.84)	0.79 (0.73 - 0.84)	18	0.72 (0.65 - 0.79)	0.78 (0.71 - 0.83)

[a]For subgroups with fewer than 4 studies (n<4), quantitative pooling was not performed; values presented represent the range from original studies. For subgroups with 1 study (n=1), the presented values are single numbers.

[b]Not applicable.

[c]ANN: artificial neural network.

[d]KNN: k-nearest neighbor.

[e]LASSO: least absolute shrinkage and selection operator.

[f]LR: logistic regression.

[g]RF: random forest.

XSL•FO
RenderX

[h]SVM: support vector machine.

[i]XGBoost: extreme gradient boosting.

[j]LightGBM: light gradient boosting machine.

The pooled sensitivity and specificity were 0.76 (95% CI 0.66 - 0.84) and 0.79 (95% CI 0.73 - 0.84) in the training set and 0.72 (95% CI 0.65 - 0.79) and 0.78 (95% CI 0.71 - 0.83) in the validation set, respectively (Table 3).

Publication Bias Analysis

Funnel plots were generated separately for the training and validation sets to assess publication bias. The results indicated no significant publication bias among the included studies in either the training (*P* value for Egger test was .65; Figure 3) or validation set (*P* value for Egger test was .59; Figure 4).

Figure 3. Publication bias analysis in the training set.

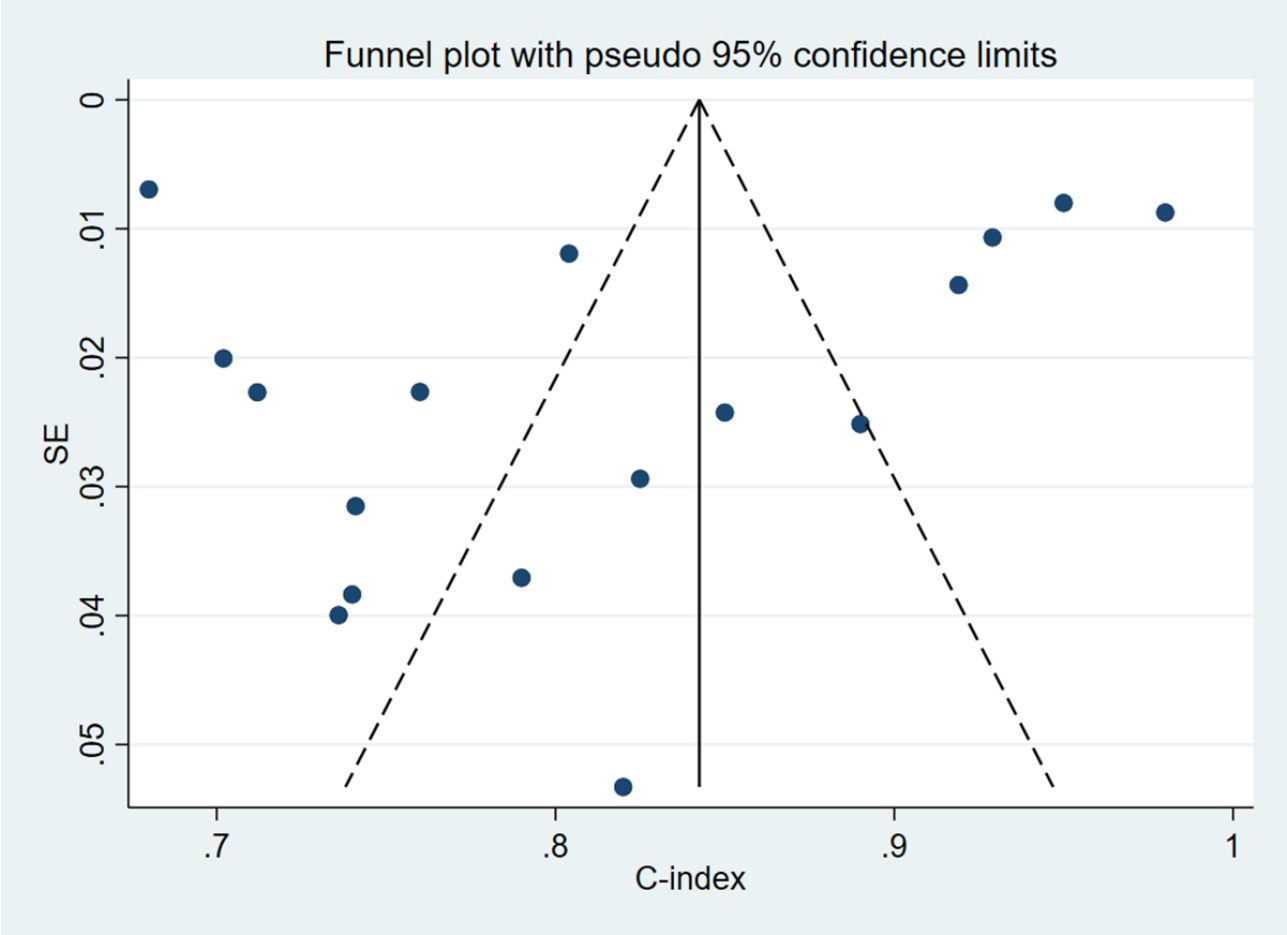

XSL•FO
RenderX

Figure 4. Publication bias analysis in the validation set.

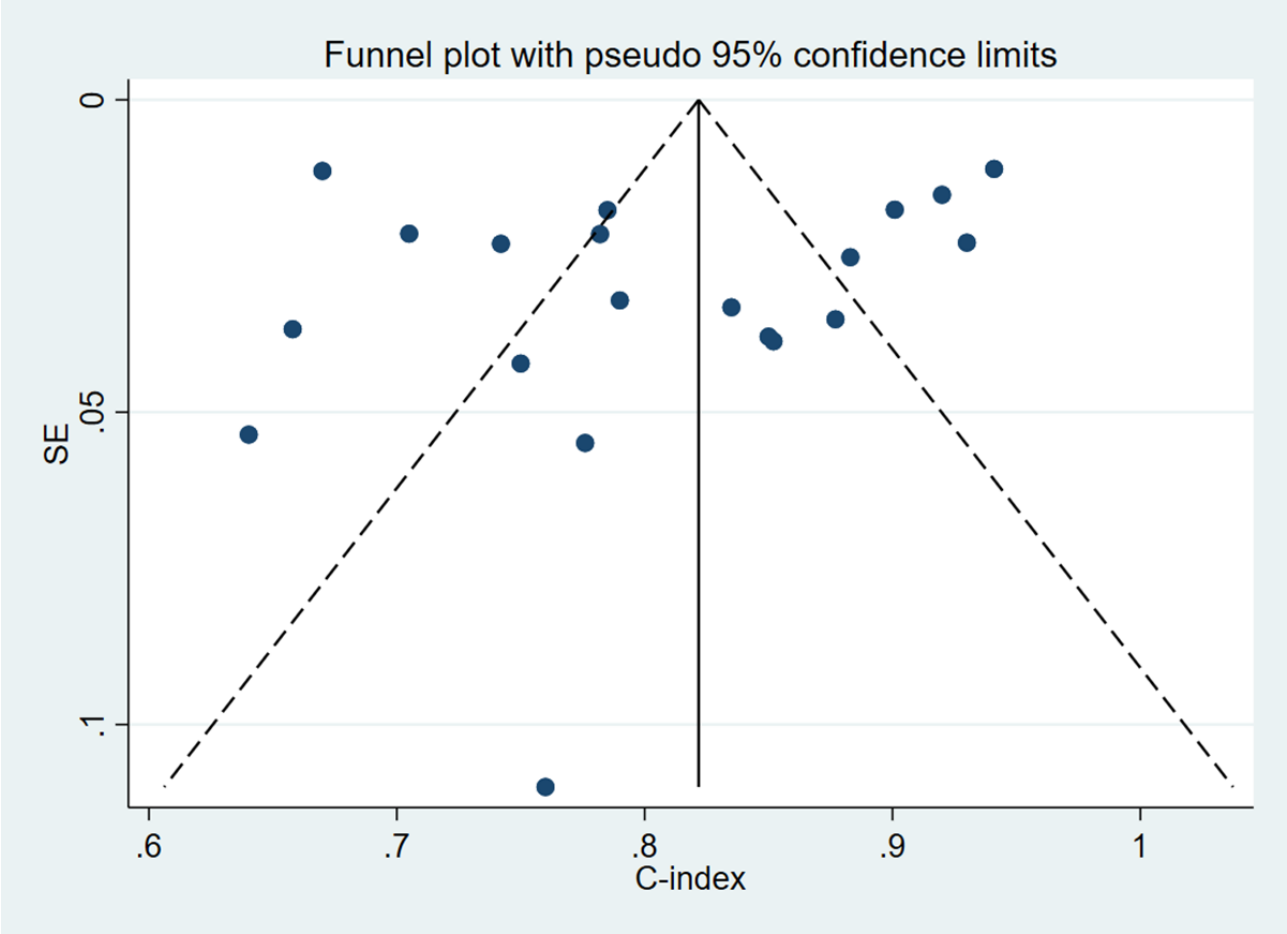

Studies for Prediction of POD Following Cardiac Surgery in Children

Two studies focusing on pediatric patients for predicting POD following cardiac surgery were ultimately included. Both were single-center studies that employed least absolute shrinkage and selection operator regression combined with multivariable logistic regression for variable selection and constructed nomogram models. The validation sets utilized random split validation and temporal validation, respectively. A meta-analysis was performed to synthesize model performance. In the validation set, the c-index, sensitivity, and specificity were 0.845 (95% CI 0.789 - 0.900), 0.84, and 0.80, respectively.

Discussion

Main Findings

This study revealed growing research interest in early prediction of POD following cardiac surgery, with ML models demonstrating promising effectiveness. However, the majority of existing models are characterized by substantial heterogeneity and high risk of bias. Logistic regression remained the predominant modeling method, showing favorable performance.

Comparison With Previous Reviews

Previous studies have reviewed early prediction methods for POD following cardiac surgery. For instance, a study by Cai et al [7] reviewed the use of ML in predicting POD in cardiac surgery. However, the included studies were limited, as the search strategy did not specifically account for the types of cardiac surgery. Furthermore, their study only provided the range of c-index values for the training set (0.74 - 0.91) and external validation set (0.54 - 0.90) without a quantitative systematic review. This limitation hindered a thorough interpretation of the value of ML in predicting POD in cardiac surgery.

In contrast, a study by Lee et al [8] summarized 3 well-established scoring tools for predicting POD in cardiac surgery. Their meta-analysis showed a pooled c-index of 0.62 (95% CI 0.58 - 0.66), while the international recalibrated PREDELIRIC model demonstrated a c-index of 0.75 (95% CI 0.72 - 0.79) in external validation. While these tools are well-established, they do not address the performance of ML. Our study has found that ML models appear to outperform these traditional scoring tools. Moreover, a review by Chen et al [9] examined POD in adults but did not specifically focus on cardiac surgery. The ML models included in their review demonstrated a c-index of 0.792 (95% CI 0.769 - 0.816) for predicting POD following cardiac surgery. Our study, based on a more comprehensive body of evidence, systematically summarizes the predictive value of ML for POD following cardiac surgery, with results from validation sets demonstrating strong predictive performance.

Previous studies have reviewed the main factors predicting POD in cardiac surgery. Koster et al [41] identified predisposing risk

XSL•FO
RenderX

factors such as atrial fibrillation, cognitive impairment, depression, stroke history, advanced age, and peripheral vascular disease. The frequency statistics in Table S3 in Multimedia Appendix 3 of the present study show that age, creatinine level, cardiopulmonary bypass duration, Mini-Mental State Examination score, and left ventricular ejection fraction are the most frequently used modeling variables in existing studies. The differences between the risk factors highlighted in this review and those reported in prior models may reflect variations in study design or data sources. Due to significant differences in modeling methods and population characteristics among the included studies, this review did not conduct formal predictive efficacy analysis to determine key predictive variables. The actual predictive value of these high-frequency variables requires further verification through more standardized designed studies.

Challenges in Clinical Translation

The studies incorporated in this review involve diverse ML models, with some observed differences in the predictive performance of these models, prompting us to summarize their predictive value for POD following cardiac surgery. The study has found that logistic regression is currently the predominant predictive model, as it allows for the construction of widely applicable or simpler predictive nomograms. These nomograms provide excellent interpretability of the relationships between clinical indicators and delirium. Although the interpretability of predictive models is a crucial measure in clinical research, we found that the accuracy of logistic regression is lower than that of several other ML models. Other models are less interpretable, which hinders their further application. Additionally, literature on these other models remains relatively scarce. Therefore, future studies should develop ML models with potentially higher accuracy, enabling earlier identification of delirium risk in cardiac surgery patients and the implementation of preventive measures.

Among included studies, the PROBAST assessment revealed high risk of bias. In the 28 included studies, high risk of bias primarily stemmed from participant selection bias (eg, some studies did not explicitly exclude patients with severe neurological comorbidities), inadequate handling of missing data, and lack of external validation or reliance solely on internal validation with small samples. These methodological limitations impose certain constraints on the interpretation of our findings: although the pooled results reflect the overall trends observed across existing studies, the actual clinical reliability of ML models warrants cautious evaluation due to methodological deficiencies in the original studies. These findings also offer important implications for future research: subsequent studies should rigorously adhere to the PROBAST assessment criteria and optimize study designs—such as refining participant selection processes, strengthening data quality control, and conducting multicenter external validation—to enhance the reliability and generalizability of ML models.

Although the predictive value of ML for POD following cardiac surgery was summarized and demonstrated relatively promising results, substantial heterogeneity cannot be ignored. This challenge is commonly faced by current meta-analyses on ML. To better explain heterogeneity, subgroup analyses were conducted based on data sets (training sets and validation sets), validation methods, study types, bias risk, and model types. Substantial heterogeneity remained among different subgroups. This heterogeneity may result from the diverse modeling variables in the model training sets. Even within the same type of model, differences may exist in model parameters and selected modeling variables. Medical conditions in different countries also seem to have potential impacts. Given these potential influences, a random-effects model was adopted for pooling. Overall, due to the impact of heterogeneity, these results require cautious interpretation. In clinical applications, comprehensive judgment based on specific population characteristics and diagnosis and treatment scenarios is necessary to avoid direct application.

Limitations

This study has several limitations. First, despite comprehensive systematic searching, the limited number of included studies may restrict result interpretation, particularly for models based on few studies. Second, original studies did not differentiate between cardiac surgery types and delirium occurrence locations (ICU vs general wards), precluding more detailed analysis. Third, most studies utilized internal validation without independent external validation, potentially limiting generalizability. Fourth, due to language search barriers, only English-language studies were included, potentially introducing bias and limiting result interpretation. Fifth, included populations showed limitations regarding source, sample size, and modeling method standardization, creating clinical application challenges without effective model stratification analysis. Sixth, this study can only reflect the variable usage trends in existing studies and cannot directly infer the predictive importance of variables. This may also result in certain established risk factors (eg, gender) appearing less frequently than their recognized importance in general clinical research would suggest. Such patterns reflect the variable selection outcomes of the ML models employed rather than the biological significance of these factors and therefore warrant cautious interpretation. Seventh, inherent limitations in the included studies—predominantly single-center, retrospective designs lacking rigorous external validation—contributed to the high risk of bias in the models, suggesting that the pooled predictive performance may be overestimated. The generalizability of these models requires further validation through prospective, multicenter studies. Therefore, future research should standardize model construction methods, enhance model transparency, and reduce bias risk to broaden clinical application. Additionally, our search strategy primarily focused on “nomogram” and specific ML terminology, potentially missing studies that reported traditional prediction models using only terms such as “logistic regression” or “multivariate analysis,” which may have introduced search bias. Although “nomogram” is a common presentation format for logistic regression models, future systematic reviews should consider more comprehensive search terms.

Conclusions

This systematic review suggests that ML-based prediction tools for POD following cardiac surgery appear to demonstrate promising performance. However, the majority of existing

XSL•FO
RenderX

models carry high risk of bias and lack rigorous external validation, potentially resulting in overly optimistic performance estimates. Current evidence is substantially constrained by extreme heterogeneity and low methodological quality. Future research should therefore standardize ML model development workflows, prioritize prospective designs, adhere to prediction model reporting standards, conduct extensive external validation, and perform comparative model studies to establish the actual clinical utility and feasibility of these tools in real-world settings.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Authors' Contributions

Writing - original draft preparation: YG
Writing - review and editing: YG
Conceptualization: PX
Methodology: PX
Formal analysis and investigation: SZ
Funding acquisition: AW
Resources: HX
Supervision: MZ
All authors contributed to the study conception and design and authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Literature search strategy.
[DOCX File, 18 KB - jmir_v28i1e72304_app1.docx]

Multimedia Appendix 2
Basic information of included studies.
[DOCX File, 41 KB - jmir_v28i1e72304_app2.docx]

Multimedia Appendix 3
Modeling variables in included studies.
[DOCX File, 26 KB - jmir_v28i1e72304_app3.docx]

Checklist 1
PRISMA 2020 checklist.
[PDF File, 101 KB - jmir_v28i1e72304_app4.pdf]

References

1. Timmis A, Vardas P, Townsend N, et al. European Society of Cardiology: cardiovascular disease statistics 2021. Eur Heart J 2022 Feb 22;43(8):716-799. [doi: 10.1093/eurheartj/ehab892] [Medline: 35016208]
2. Stephens RS, Whitman GJR. Postoperative critical care of the adult cardiac surgical patient: part II: procedure-specific considerations, management of complications, and quality improvement. Crit Care Med 2015 Sep;43(9):1995-2014. [doi: 10.1097/CCM.0000000000001171] [Medline: 26136101]
3. Andersson B, Nilsson J, Brandt J, Höglund P, Andersson R. Gastrointestinal complications after cardiac surgery. Br J Surg 2005 Mar;92(3):326-333. [doi: 10.1002/bjs.4823] [Medline: 15672438]
4. Newman MF, Kirchner JL, Phillips-Bute B, et al. Longitudinal assessment of neurocognitive function after coronary-artery bypass surgery. N Engl J Med 2001 Feb 8;344(6):395-402. [doi: 10.1056/NEJM200102083440601] [Medline: 11172175]
5. Saczynski JS, Marcantonio ER, Quach L, et al. Cognitive trajectories after postoperative delirium. N Engl J Med 2012 Jul 5;367(1):30-39. [doi: 10.1056/NEJMoa1112923] [Medline: 22762316]

XSL•FO
RenderX

6. Salluh JIF, Wang H, Schneider EB, et al. Outcome of delirium in critically ill patients: systematic review and meta-analysis. BMJ 2015 Jun 3;350:h2538. [doi: 10.1136/bmj.h2538] [Medline: 26041151]
7. Cai S, Li J, Gao J, Pan W, Zhang Y. Prediction models for postoperative delirium after cardiac surgery: systematic review and critical appraisal. Int J Nurs Stud 2022 Dec;136:104340. [doi: 10.1016/j.ijnurstu.2022.104340] [Medline: 36208541]
8. Lee A, Mu JL, Joynt GM, et al. Risk prediction models for delirium in the intensive care unit after cardiac surgery: a systematic review and independent external validation. Br J Anaesth 2017 Mar 1;118(3):391-399. [doi: 10.1093/bja/aew476] [Medline: 28186224]
9. Chen H, Yu D, Zhang J, Li J. Machine learning for prediction of postoperative delirium in adult patients: a systematic review and meta-analysis. Clin Ther 2024 Dec;46(12):1069-1081. [doi: 10.1016/j.clinthera.2024.09.013] [Medline: 39395856]
10. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019 Jan 1;170(1):51-58. [doi: 10.7326/M18-1376] [Medline: 30596875]
11. Debray TP, Damen JA, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. Stat Methods Med Res 2019 Sep;28(9):2768-2786. [doi: 10.1177/0962280218785504] [Medline: 30032705]
12. Hatano Y, Narumoto J, Shibata K, et al. White-matter hyperintensities predict delirium after cardiac surgery. Am J Geriatr Psychiatry 2013 Oct;21(10):938-945. [doi: 10.1016/j.jagp.2013.01.061] [Medline: 24029014]
13. Mao D, Fu L, Zhang W. Risk factors and nomogram model of postoperative delirium in children with congenital heart disease: a single-center prospective study. Pediatr Cardiol 2024 Jan;45(1):68-80. [doi: 10.1007/s00246-023-03297-5] [Medline: 37741935]
14. Han C, Kim HI, Soh S, Choi JW, Song JW, Yoon D. Machine learning with clinical and intraoperative biosignal data for predicting postoperative delirium after cardiac surgery. iScience 2024 Jun 21;27(6):109932. [doi: 10.1016/j.isci.2024.109932] [Medline: 38799563]
15. Zhao X, Li J, Xie X, et al. Online interpretable dynamic prediction models for postoperative delirium after cardiac surgery under cardiopulmonary bypass developed based on machine learning algorithms: a retrospective cohort study. J Psychosom Res 2024 Jan;176:111553. [doi: 10.1016/j.jpsychores.2023.111553] [Medline: 37995429]
16. Lin N, Lv M, Li S, Xiang Y, Li J, Xu H. A nomogram for predicting postoperative delirium in pediatric patients following cardiopulmonary bypass: a prospective observational study. Intensive Crit Care Nurs 2024 Aug;83:103717. [doi: 10.1016/j.iccn.2024.103717] [Medline: 38692080]
17. de la Varga-Martínez O, Gómez-Pesquera E, Muñoz-Moreno MF, et al. Development and validation of a delirium risk prediction preoperative model for cardiac surgery patients (DELIPRECAS): an observational multicentre study. J Clin Anesth 2021 May;69:110158. [doi: 10.1016/j.jclinane.2020.110158] [Medline: 33296785]
18. Zhang Y, Ren M, Zhai W, Han J, Guo Z. Construction and validation of a risk prediction model for postoperative delirium in patients with off - pump coronary artery bypass grafting. J Thorac Dis 2024 Jun;16(6):3944-3955. [doi: 10.21037/jtd-24-578] [Medline: 38983165]
19. Nagata C, Hata M, Miyazaki Y, et al. Development of postoperative delirium prediction models in patients undergoing cardiovascular surgery using machine learning algorithms. Sci Rep 2023 Nov 30;13(1):21090. [doi: 10.1038/s41598-023-48418-5] [Medline: 38036664]
20. Lin JL, Zheng GZ, Chen LW, Luo ZR. A nomogram model for assessing predictors and prognosis of postoperative delirium in patients receiving acute type A aortic dissection surgery. BMC Cardiovasc Disord 2023 Feb 7;23(1):72. [doi: 10.1186/s12872-023-03111-3] [Medline: 36750929]
21. Hata M, Miyazaki Y, Nagata C, et al. Predicting postoperative delirium after cardiovascular surgeries from preoperative portable electroencephalography oscillations. Front Psychiatry 2023;14:1287607. [doi: 10.3389/fpsyt.2023.1287607] [Medline: 38034919]
22. Xu Y, Meng Y, Qian X, et al. Prediction model for delirium in patients with cardiovascular surgery: development and validation. J Cardiothorac Surg 2022 Oct 1;17(1):247. [doi: 10.1186/s13019-022-02005-3] [Medline: 36183105]
23. Segernäs A, Skoog J, Ahlgren Andersson E, Almerud Österberg S, Thulesius H, Zachrisson H. Prediction of postoperative delirium after cardiac surgery with a quick test of cognitive speed, mini-mental state examination and hospital anxiety and depression scale. Clin Interv Aging 2022;17:359-368. [doi: 10.2147/CIA.S350195] [Medline: 35400995]
24. Cai S, Cui H, Pan W, Li J, Lin X, Zhang Y. Two-stage prediction model for postoperative delirium in patients in the intensive care unit after cardiac surgery. Eur J Cardiothorac Surg 2022 Dec 2;63(1):ezac573. [doi: 10.1093/ejcts/ezac573] [Medline: 36579859]
25. He J, Ling Q, Chen Y. Construction and application of a model for predicting the risk of delirium in postoperative patients with type a aortic dissection. Front Surg 2021;8:772675. [doi: 10.3389/fsurg.2021.772675] [Medline: 34869569]
26. Kotfis K, Ślozowska J, Safranow K, Szylińska A, Listewnik M. The practical use of white cell inflammatory biomarkers in prediction of postoperative delirium after cardiac surgery. Brain Sci 2019 Nov 2;9(11):308. [doi: 10.3390/brainsci9110308] [Medline: 31684066]
27. Ten Broeke M, Koster S, Konings T, Hensens AG, van der Palen J. Can we predict a delirium after cardiac surgery? A validation study of a delirium risk checklist. Eur J Cardiovasc Nurs 2018 Mar;17(3):255-261. [doi: 10.1177/1474515117733365] [Medline: 28980478]

XSL•FO
RenderX

28. Price CC, Garvan C, Hizel LP, Lopez MG, Billings FT 4th. Delayed recall and working memory MMSE domains predict delirium following cardiac surgery. J Alzheimers Dis 2017;59(3):1027-1035. [doi: 10.3233/JAD-170380] [Medline: 28697572]
29. Koster S, Hensens AG, Schuurmans MJ, van der Palen J. Prediction of delirium after cardiac surgery and the use of a risk checklist. Eur J Cardiovasc Nurs 2013 Jun;12(3):284-292. [doi: 10.1177/1474515112450244] [Medline: 22694810]
30. Li Q, Li J, Chen J, et al. A machine learning-based prediction model for postoperative delirium in cardiac valve surgery using electronic health records. BMC Cardiovasc Disord 2024 Jan 18;24(1):56. [doi: 10.1186/s12872-024-03723-3] [Medline: 38238677]
31. Tian Y, Ji B, Diao X, et al. Dynamic predictive scores for cardiac surgery-associated agitated delirium: a single-center retrospective observational study. J Cardiothorac Surg 2023 Jul 6;18(1):219. [doi: 10.1186/s13019-023-02339-6] [Medline: 37415226]
32. Li X, Cheng W, Zhang J, Li D, Wang F, Cui N. Early alteration of peripheral blood lymphocyte subsets as a risk factor for delirium in critically ill patients after cardiac surgery: a prospective observational study. Front Aging Neurosci 2022;14:950188. [doi: 10.3389/fnagi.2022.950188]
33. Andrási TB, Talipov I, Dinges G, Arndt C, Rastan AJ. Risk factors for postoperative delirium after cardiac surgical procedures with cardioplegic arrest. Eur J Cardiothorac Surg 2022 Jun 15;62(1):ezab570. [doi: 10.1093/ejcts/ezab570] [Medline: 35037042]
34. Rudolph JL, Jones RN, Levkoff SE, et al. Derivation and validation of a preoperative prediction rule for delirium after cardiac surgery. Circulation 2009 Jan 20;119(2):229-236. [doi: 10.1161/CIRCULATIONAHA.108.795260] [Medline: 19118253]
35. Wang Y, Shen B, Zhu C, et al. Unveiling the nexus of postoperative fever and delirium in cardiac surgery: identifying predictors for enhanced patient care. Front Cardiovasc Med 2023;10:1237055. [doi: 10.3389/fcvm.2023.1237055]
36. Yang T, Yang H, Liu Y, et al. Postoperative delirium prediction after cardiac surgery using machine learning models. Comput Biol Med 2024 Feb;169:107818. [doi: 10.1016/j.compbiomed.2023.107818] [Medline: 38134752]
37. Sadlonova M, Hansen N, Esselmann H, et al. Preoperative delirium risk screening in patients undergoing a cardiac surgery: results from the prospective observational FINDERI study. Am J Geriatr Psychiatry 2024 Jul;32(7):835-851. [doi: 10.1016/j.jagp.2023.12.017] [Medline: 38228452]
38. Mufti HN, Hirsch GM, Abidi SR, Abidi SSR. Exploiting machine learning algorithms and methods for the prediction of agitated delirium after cardiac surgery: models development and validation study. JMIR Med Inform 2019 Oct 23;7(4):e14993. [doi: 10.2196/14993] [Medline: 31558433]
39. Lapp L, Roper M, Kavanagh K, Schraag S. Predicting the onset of delirium on hourly basis in an intensive care unit following cardiac surgery. Presented at: 2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS); Jul 21-23, 2022; Shenzen, China p. 234-239. [doi: 10.1109/CBMS55023.2022.00048]
40. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019 Jan 1;170(1):W1-W33. [doi: 10.7326/M18-1377] [Medline: 30596876]
41. Koster S, Hensens AG, Schuurmans MJ, van der Palen J. Risk factors of delirium after cardiac surgery: a systematic review. Eur J Cardiovasc Nurs 2011 Dec;10(4):197-204. [doi: 10.1016/j.ejcnurse.2010.09.001] [Medline: 20870463]

Abbreviations

ML: machine learning
POD: postoperative delirium
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST: Prediction Model Risk of Bias Assessment Tool

Edited by A Coristine; submitted 07.Feb.2025; peer-reviewed by A Jafarizadeh, J Zeng; accepted 12.Jan.2026; published 23.Feb.2026.

Please cite as:
Guo Y, Xu H, Wang A, Zhang M, Zhang S, Xie P
The Predictive Value of Machine Learning for Postoperative Delirium in Cardiac Surgery: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e72304
URL: https://www.jmir.org/2026/1/e72304
doi:10.2196/72304

© Yi Guo, Hong Xu, Ankui Wang, Mingming Zhang, Shuai Zhang, Peng Xie. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Feb.2026. This is an open-access article distributed under the terms of the Creative

XSL•FO
RenderX

Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Securing Federated Learning With Blockchain in the Medical Field: Systematic Literature Review

Xudong Wang[1*], MD; Yi Xie[2*], MD; Xiaoliang Chen[1,3*], MD; Jiaming Yang[1], MD; Ruiyuan Li[1], MD; Weihang Gao[1], MD; Zineng Yan[1], MD; Hong Zhou[1], MD; Zhewei Ye[1], Prof Dr Med

[1]Department of Orthopedics, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, 1277 Jiefang Avenue, Wuhan, Hubei Province, China

[2]Renmin Hospital of Wuhan University, Wuhan, Hubei Province, China

[3]Department of Orthopedics, People's Hospital of Ningxia Hui Autonomous Region, Ningxia Medical University, Yinchuan, China

[*]these authors contributed equally

Corresponding Author:
Zhewei Ye, Prof Dr Med
Department of Orthopedics, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, 1277 Jiefang Avenue, Wuhan, Hubei Province, China

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2026/1/e95788

Abstract

Background: The exponential growth of medical data and advancements in artificial intelligence (AI) have accelerated the development of data-driven health care. However, the secure and efficient sharing of sensitive medical data across institutions remains a major challenge due to privacy concerns, data silos, and regulatory restrictions. Traditional centralized systems are prone to data breaches and single points of failure, while existing privacy-preserving techniques face high computational and communication costs.

Objective: This study aims to provide a comprehensive review of the recent advances in blockchain-based federated learning (BCFL) within the medical field. By exploring the synergistic integration of federated learning and blockchain, this review evaluates how BCFL enhances data security, supports privacy-preserving cross-institutional collaboration, and facilitates practical applications in health care, including medical data sharing, Internet of Medical Things, public health surveillance, and telemedicine.

Methods: We conducted a systematic literature review using databases such as PubMed, IEEE Xplore, Web of Science, and Google Scholar. Boolean logic and domain-specific keywords were used to retrieve studies from 2018 to 2025. After automated deduplication and multistage manual screening, over 100 high-quality papers were included. These works cover BCFL's theoretical foundations, system architectures, application domains, limitations, and future directions.

Results: BCFL frameworks combine the decentralized trust and auditability of blockchain with the privacy-preserving collaborative learning capabilities of federated learning. This integration mitigates risks such as model tampering, data leakage, and a lack of incentives in federated systems. Applications span across cross-institutional medical data sharing, Internet of Medical Things, epidemic forecasting, and telemedicine. Architectures including fully coupled, flexibly coupled, and loosely coupled models offer varying trade-offs between efficiency, scalability, and security.

Conclusions: BCFL represents a transformative paradigm for secure, collaborative, and privacy-preserving medical AI. By combining decentralized trust, incentive-driven participation, and privacy-enhancing machine learning, BCFL paves the way for next-generation smart health care systems. Despite current technical and practical challenges, BCFL demonstrates strong potential to support precision medicine, global health data collaboration, and large-scale AI deployment in health care.

(J Med Internet Res 2026;28:e79052) doi:10.2196/79052

KEYWORDS

blockchain; federated learning; health care; health data; review; Internet of Medical Things; COVID-19; IoMT

XSL•FO
RenderX

Introduction

Background

With the continuous advancement of medical informatization, the volume of medical big data is growing exponentially, offering significant opportunities for the application of artificial intelligence (AI) in health care, particularly in areas such as assisted diagnosis, personalized treatment, and disease prediction. Especially in recent years, advances in computing power and algorithmic innovation have made machine learning (ML) models a cornerstone of medical intelligence, with their efficient training and optimization relying heavily on large-scale, high-quality datasets from multiple sources [1-3]. However, the acquisition and sharing of medical data face significant obstacles, including privacy concerns, data security risks, and regulatory constraints [4].

In the digital medicine era, the secure exchange and control of sensitive health information have become central concerns in modern health care systems [5,6]. Meanwhile, strict legal and regulatory requirements—such as HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation)—must be met when handling these personal data [7]. Additionally, the competition between different medical centers and hospitals has led to data often remaining siloed [8]. These challenges not only impede the effective integration of multisource health care data but also hinder the translation of ML models from theoretical research to clinical practice.

While conventional centralized data storage architectures can partially facilitate model development, their dependence on centralized infrastructures has become increasingly problematic, rendering them vulnerable to single-point failures and malicious cyberattacks [9,10]. Consequently, establishing secure frameworks for cross-institutional data sharing and intelligent processing—while rigorously protecting data security and patient confidentiality—has emerged as a pivotal challenge impeding progress in medical AI. The fundamental tension between aggregating health care data for scientific progress and preserving individual privacy and data security has spurred the development of novel computational approaches. Federated learning (FL), as an emerging ML paradigm, addresses part of this problem by allowing institutions to collaboratively train models without exchanging raw data [11,12]. However, in practice, the deployment of FL in health care faces practical obstacles: the integrity and authenticity of model updates may be subject to security threats, including malicious attacks, client-side data tampering, and model forgery—all of which can reduce the accuracy of the global model. Furthermore, at present, there is a lack of reliable incentive structures for continuous participation, with limited auditability. The heterogeneity between institutions and edge devices undermines the integration and generalization of the model. Blockchain technology, with its decentralized architecture, immutable ledger system, and transparent traceability, has become a potential solution to address these limitations of logical reasoning [2,13,14]. When the two are integrated, blockchain can provide a verifiable source for model contributions, automated and transparent incentive mechanisms [15,16], tamper-proof logs for auditing, and a governance layer that supports cross-institutional workflows [17]. Therefore, the combined paradigm of blockchain-based federated learning (BCFL) is expected to become a practical approach to coordinating data privacy, trust, and collaborative intelligence in medicine.

Although research on FL and blockchain has accelerated in recent years, existing reviews exhibit clear gaps in focus and depth, limiting their value for clinical researchers and multidisciplinary audiences. The main shortcomings can be summarized as follows. First, most reviews examine either FL or blockchain technology in isolation, without providing a systematic analysis of how these two approaches can be integrated to address concrete challenges in health care. Second, prior reviews tend to emphasize algorithmic and technical details, while offering limited discussion of how BCFL can be adapted to real-world medical scenarios—such as cross-hospital electronic health record (EHR) integration, collaborative training of medical imaging models, Internet of Medical Things (IoMT) device coordination, and epidemic surveillance. Finally, few reviews adequately address the challenges, regulatory considerations, and future development trends of BCFL within the constraints of modern health care governance frameworks.

To address these gaps, this review makes three key contributions. First, it provides a medical demand–oriented, systematic classification of BCFL frameworks and outlines their typical workflows. Second, it maps different BCFL architectures to representative health care application scenarios, clarifying their practical relevance. Third, it analyzes the technical, regulatory, and implementation challenges that currently hinder BCFL adoption and identifies promising directions for future research, providing evidence-based insights for clinical translation and decision-making in the health care and biomedical research communities.

Objective

Although previous studies have analyzed BCFL frameworks from the perspectives of technology and commercial applications, the review of their practical application scenarios in health care remains limited. Especially in the era of AI, with the explosive growth of medical data, the application of large models in medical practice has encountered some development obstacles, bringing some new ideas for the application prospects of BCFL in health care. This study investigates the synergistic integration of FL and blockchain technology, elucidating their combined architectural framework and operational mechanisms. We demonstrate how this technological convergence enables secure, privacy-preserving health care data sharing and collaborative model development across distributed health care institutions. Although blockchain-enhanced FL has recently emerged as a promising approach in medical research, the field remains nascent, constrained by technical limitations, unresolved privacy issues, and implementation barriers. Furthermore, we present a systematic review of current advancements in blockchain-assisted FL for medical applications and propose actionable research directions to overcome existing challenges.

Key contributions of this work include:

XSL•FO
RenderX

1. This study systematically reviews the theoretical foundations of FL and blockchain technology and elaborates on their potential and advantages in the medical field.
2. A comprehensive taxonomy of existing integration frameworks, accompanied by a mechanistic analysis of bidirectional benefits: how blockchain enhances FL security and trustworthiness, and how FL expands blockchain's utility in distributed computing scenarios.
3. To summarize recent advancements and practical applications of BCFL across key health care domains, including cross-institutional medical data sharing, IoMT, public health surveillance, and telemedicine.
4. The current technological limitations and application challenges are examined, and key future research directions are proposed to address these gaps and advance real-world implementation.

By synthesizing current research and offering a structured analytical framework, we aim to provide a comprehensive reference for medical personnel, researchers, and health care policymakers. Ultimately, it fosters the development of trustworthy, privacy-preserving, and collaborative AI systems that can support precision medicine and smart health care in a decentralized digital era.

Methods

Overview

This review systematically summarizes recent advances in BCFL within the medical field, following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist (Checklist 1). To ensure methodological rigor and reproducibility, we adopted a transparent multistep process including comprehensive literature search, independent screening, quality assessment, and evidence synthesis.

Search Strategy

The literature data were primarily retrieved from several prominent academic databases, including PubMed, IEEE Xplore, Web of Science, and Google Scholar. To ensure the timeliness of the research, the review focuses on literature published between January 2018 and February 2025, while also incorporating some early seminal studies to trace the theoretical development and technological evolution of BCFL in the medical field.

After that, the search strategy uses Boolean logic operators to formulate a comprehensive formula, for example, ("blockchain") AND ("federated learning") AND ("medical" OR "healthcare"). To enhance retrieval efficiency and encompass interdisciplinary intersections, the search terms are further expanded, such as "(blockchain-enabled federated learning)," "(distributed machine learning)," "(decentralization)," "(Internet of Medical Things (IoMT)," "(telemedicine)," "(EMR)," and "(epidemics)" are incorporated to ensure literature comprehensiveness. The exact search string is as follows: ("blockchain" OR "distributed ledger technology") AND ("federated learning" OR "collaborative learning" OR "distributed machine learning") AND ("healthcare" OR "medical" OR "clinical" OR "EMR" OR "(epidemics)" OR "IoMT" OR "telemedicine"). Gray literature (conference proceedings and preprints from arXiv) was screened manually and included only if it contained original data or detailed technical methodology relevant to BCFL.

Inclusion and Exclusion Criteria

To ensure relevance and academic rigor, this review establishes strict inclusion and exclusion criteria. The inclusion criteria are as follows: (1) the studies must involve blockchain and FL technologies and explore their medical applications; (2) the literature published in peer-reviewed journals and reviews indexed in the Science Citation Index or Social Sciences Citation Index or in top-tier international conferences (eg, IEEE and Association for Computing Machinery) and seminal papers with >30 citations were included regardless of publication venue; and (3) reported theoretical frameworks, system architectures, empirical evaluations, or case studies. The exclusion criteria then include (1) studies focusing solely on blockchain or FL without medical applications, (2) editorial and opinion articles that lack technical details or empirical validation, and (3) duplicate reports or low-quality publications from non–peer-reviewed sources.

Literature Screening and Processing

The above search strategy initially retrieved 2547 documents. After automatic deduplication by EndNote, 1327 were retained. Subsequently, two independent reviewers (XW and XC) manually screened all the literature in two stages based on the inclusion and exclusion criteria. First, based on the title and abstract, high-quality reviews, papers from top journals and conferences, and literature that clearly explained the application of BCFL in the health care field were selected, totaling 319 articles. In the second stage, further review was conducted. Through full-text reading, studies lacking empirical verification, technical details, or experimental data were eliminated, and ultimately 111 high-quality documents were retained. In addition, to ensure the comprehensiveness of the literature, this review also referred to the latest review papers and included the core research results cited therein to avoid missing key progress. Detailed literature retrieval strategies can be found in Multimedia Appendix 1. After the screening process, the reviewers evaluated the quality of the literature. Since the corpus of this review mainly comes from the cross-disciplinary research of medicine, engineering, and computer science, the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) framework commonly used in clinical evidence-based studies was not applicable. We used the Systematic Literature Review quality checklist based on the Kitchenham principle [18] to score each of the 12 indicators of the included studies, including reproducibility, method transparency, evaluation design, data description, experimental validity, attack/privacy discussion, and reproducibility, item by item (1/0.5/0). Two independent reviewers (XW and XC) assessed each document. When there were significant differences in the scores (the total score difference of a single document was greater than 2 or there were differences in key items), a third senior reviewer (YX) arbitrated. Ultimately, the studies were classified into three quality grades based on the total score: high (≥9), medium (5-8), and low (<5). When writing summaries and conclusions, priority was given to citing

XSL•FO
RenderX

high-quality research. Multimedia Appendix 2 [1-130] provides complete assessment criteria, scoring rules, and statistics on peer-to-peer agreements.

Results

Study Findings

Finally, this review is founded on over 100 strictly selected papers, encompassing theoretical research, technical architecture, application scenarios, challenges, and future trends. These studies provide a robust academic foundation for an in-depth exploration of BCFL applications in medicine. A PRISMA flow diagram illustrates the systematic selection process (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the systematic review phases. FL: federated learning.

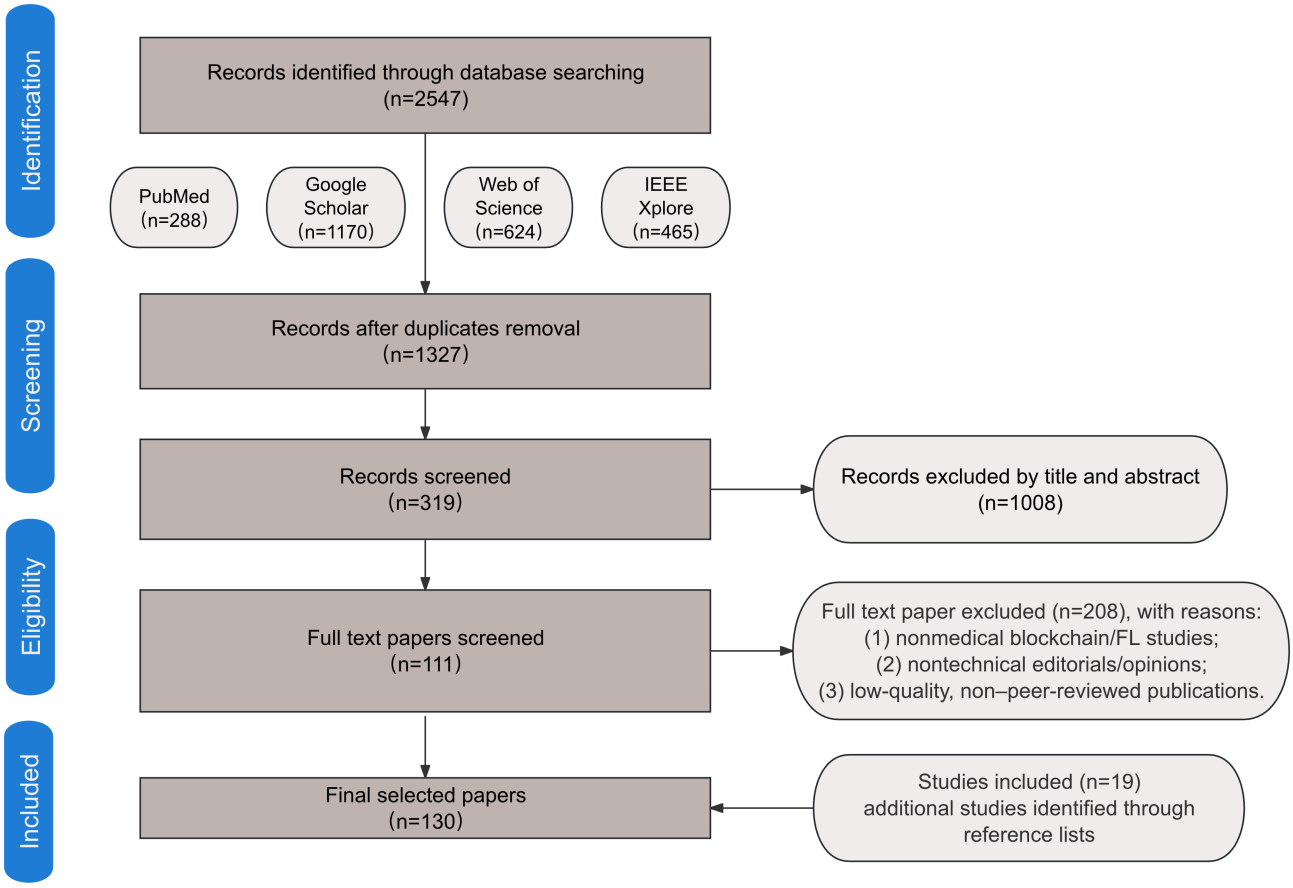

Current Status of Artificial Intelligence in Medicine

In the context of the rapid advancements in AI, the performance of ML models heavily depends on access to large volumes of high-quality data. However, the health care sector has long struggled with data sharing due to concerns over privacy, security, and regulatory compliance. The prevalence of "data silos" in health care impedes the advancement of AI and hinders the translation of research findings into practical clinical applications.

Traditional centralized ML approaches require aggregating raw data from multiple sources to a central server for training. While this approach allows for extensive data usage, it still presents significant challenges in practical implementation. First, data stored and transmitted in a centralized system is susceptible to network attacks, which can lead to sensitive information theft or tampering, posing a serious privacy risk. Second, centralized storage faces significant compliance challenges, especially under stringent privacy regulations such as GDPR [7] and HIPAA, which further restrict the cross-organizational exchange of health care data. Moreover, health care organizations, research institutions, and enterprises are often reluctant to share critical data due to competitive concerns and resource protection, thus further exacerbating data silos and hindering cross-organizational collaboration in building high-quality ML models.

Federated Learning in Health Care

In this context, the limitations of traditional algorithms are becoming increasingly apparent, driving researchers to explore innovative solutions. In 2016, Google introduced the concept of federated learning [19], a distributed and collaborative ML paradigm. At its core, FL enables multiple data holders to collaboratively develop a global model by training locally and exchanging model parameters without sharing raw data. This paradigm is well-suited for scenarios with strict privacy requirements and decentralized data that cannot be centrally stored, such as health care, finance, and smart cities [20-23].

XSL•FO
RenderX

This approach not only effectively enhances data privacy protection but also overcomes the limitation of data silos, ushering in a new era of privacy-preserving collaborative learning.

FL operates on the principles of distributed model training and global parameter aggregation. One of its core algorithms is the Federated Averaging (FedAvg) algorithm. As illustrated in Figure 2, the fundamental process can be summarized as follows [24,25]. The first step is that a central coordination server initializes the global model and distributes it to all participating clients (eg, hospitals or mobile devices). Then each client trains the model locally using its private dataset and transmits the updated model parameters (or gradients) back to the central server in an encrypted form. Following this, the central server collects and aggregates the model parameters uploaded by all clients and updates the global model based on a predefined aggregation algorithm (eg, FedAvg). The updated model is then redistributed to the clients for the next training round. This process is iteratively performed over multiple communication rounds until the global model converges or predefined performance metrics are met.

Figure 2. Federated learning architecture and workflow in the medical field.

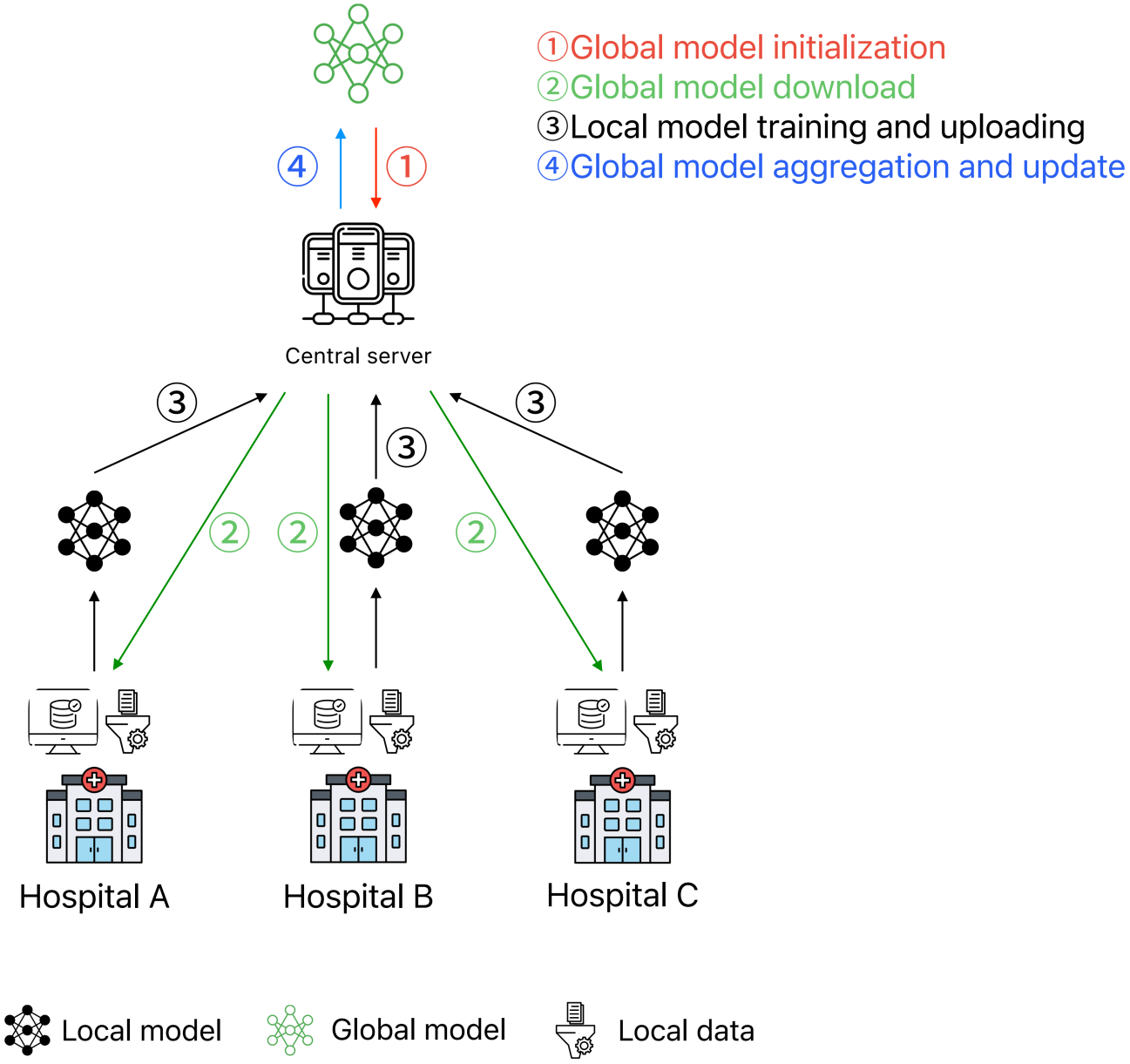

Based on the above operating principle of FL, its unique technical characteristics give it different advantages in the health care field: if it can use large-scale, diverse, and geographically distributed datasets without compromising patient privacy, this decentralized approach not only mitigates the impact of data silos, which are common in hospitals and research institutions, but also captures a broader range of features by leveraging heterogeneous real datasets, ultimately achieving an AI model with higher efficiency, robustness, and accuracy. Therefore, since its proposal, FL has shown great application potential in the field of health care [26]. For instance, FL has been applied in drug research, allowing pharmaceutical companies to leverage shared algorithmic models to accelerate drug discovery while avoiding direct data sharing [27]. Additionally, FL-based collaborative protocols, involving multiple hospitals and cloud servers, have been developed for EHR analytics [28]. Moreover,

XSL•FO
RenderX

FL has been investigated for predicting hospitalizations in cardiac patients [29]. FL has also been widely studied in medical imaging applications, such as prostate cancer detection, brain tumor segmentation [30], and MRI analysis for Alzheimer and Parkinson diseases. Moreover, FL-based approaches have been proposed for detecting coronavirus infections [31,32].

Despite the promise of FL in medical applications, its practical application still faces the following challenges: a central concern is its dependence on a centralized server for model coordination and parameter aggregation. However, the central server can be a source of a single point of failure and remains susceptible to man-in-the-middle attacks [33,34]. Moreover, as an increasing number of local devices simultaneously transmit model parameters to the central server, it places significant pressure on server bandwidth and scalability, thereby elevating the risk of network congestion [8,35,36]. Besides, although FL avoids direct sharing of raw data, its training process may still inadvertently expose sensitive information through model parameters. For instance, via gradient inversion attacks [37], adversaries can reconstruct sensitive training data from shared model parameters. The risk is further heightened by the presence of malicious participants, who may compromise the integrity of FL by uploading falsified or adversarial local training updates, such as poisoning attacks [38]. These attacks degrade the reliability of the global model, reducing its accuracy and overall utility. Additionally, the lack of a robust incentive mechanism poses a practical barrier to adoption. FL implicitly assumes that all participants willingly contribute data and computational resources without direct compensation. However, this assumption is difficult to uphold in real-world scenarios. The absence of a fair and transparent incentive mechanism may diminish participants' motivation to contribute high-quality updates, ultimately degrading system performance [39-41]. Moreover, some participants may receive rewards without actively contributing data, leading to unfair financial compensation.

When FL is applied to the medical domain, heterogeneity across systems, data, and distributions presents a fundamental challenge to model reliability and practical deployment. System heterogeneity arises from disparities in computing power, memory, network stability, and energy availability among participating devices [42]. This is particularly critical in resource-constrained environments, where limited hardware capabilities and intermittent connectivity can hinder local training and delay model updates. Data heterogeneity further complicates collaboration, as medical institutions often store data in different formats and standards, leading to integration difficulties [43,44]. Variations in data quality—such as incomplete records or inconsistent annotations—and significant differences in dataset sizes across institutions can distort the learning process and reduce model performance. Most critically, distribution heterogeneity, or the non-independent and identically distributed (non-IID) data problem, undermines generalizability. Institutions serve distinct patient populations, resulting in divergent feature and label distributions; for example, a model trained predominantly on "healthy" samples may struggle with accurate predictions when exposed to datasets with a higher prevalence of "diseased" samples, reducing its generalizability.

These advantages and limitations together underscore the urgent need for complementary technologies to enhance trust, security, and coordination in FL workflows. Blockchain, with its decentralized, tamper-resistant, and auditable infrastructure, offers a promising solution to many of these pain points. In the following section, we explore how blockchain can be integrated with FL to build more robust, transparent, and privacy-preserving frameworks for medical applications.

Blockchain in Health Care

Introduced in 2008 as the foundational technology behind the Bitcoin system, it was originally designed to solve trust-related issues in digital currency transactions [45]. Essentially, it functions as a distributed digital ledger, where transaction records are maintained and shared across all participants through a peer-to-peer network. Unlike traditional centralized systems, blockchain eliminates reliance on a trusted third party, ensuring decentralized data storage [46].

The implementation of blockchain technology relies on multiple technical layers and core components. To fully understand the principles of blockchain, it is essential to analyze its specific components. The fundamental unit of blockchain is the "block," comprising two primary components: the block header and the block body, which is shown in Figure 3. The block header contains metadata, including the hash value of the preceding block (which links the current block to its predecessor), timestamps, random numbers, and block version numbers. This metadata establishes the linkage between blocks. The block body holds smart contracts and actual data, including transaction records. These blocks are cryptographically linked in chronological order, forming an immutable chain structure. Each block is closely linked to its predecessor via a hash value, ensuring that any tampering with a single block will alter the hash values of all subsequent blocks. This change is detectable and will be rejected by the entire network, thereby preserving data authenticity and integrity.

XSL•FO
RenderX

Figure 3. A schematic diagram of the blockchain structure.

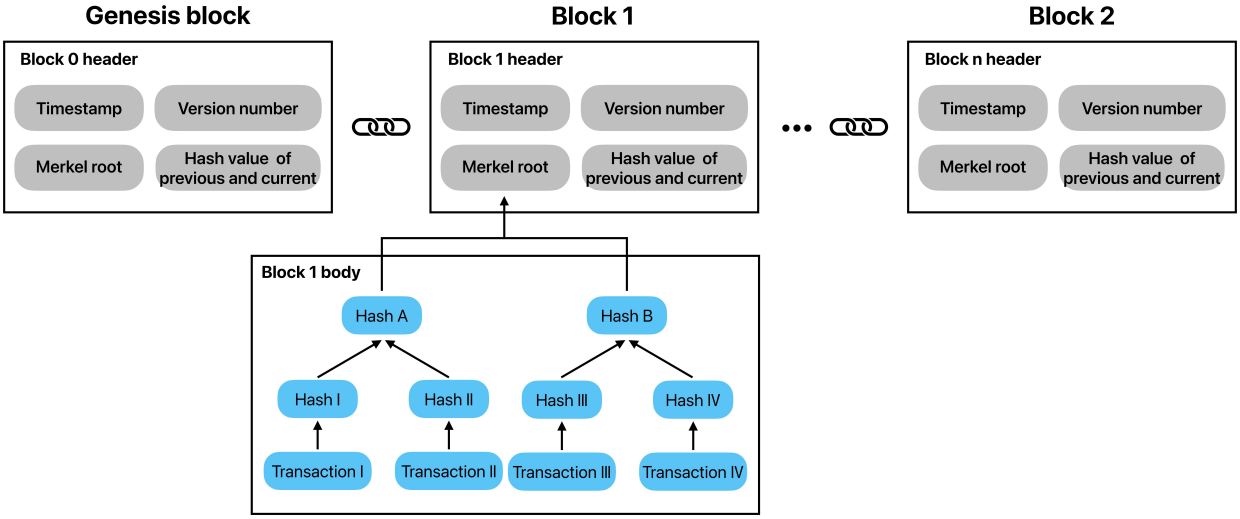

Furthermore, to fully understand the potential of blockchain in such applications, it is essential to examine its underlying architecture and systematically analyze its functional layers (shown in Figure 4).

Figure 4. The system architecture of blockchain. P2P: Peer to Peer; PBFT: Practical Byzantine Fault Tolerance; PoQ: Proof of Quality; PoS: Proof of Stake; PoW: Proof of Work; SCM: Supply Chain Management.

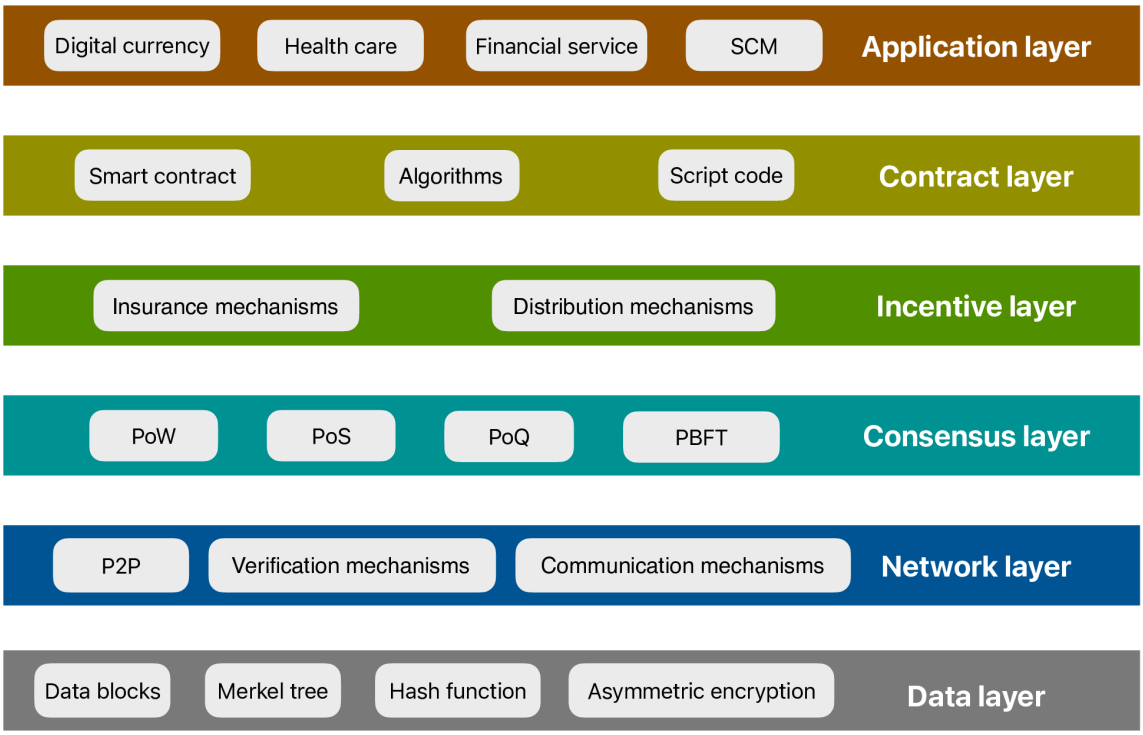

Fundamentally, blockchain systems are structured into 6 tightly interwoven layers, each addressing specific functional needs. At the core lies the data layer, which serves as the foundation for storing and organizing transaction records. The data layer organizes and secures transaction records through cryptographic techniques. This layer not only guarantees data integrity and immutability but also significantly reduces storage overhead, which is critical for handling the growing volume of medical data [47]. Above this, the network layer enables efficient and reliable information exchange across decentralized nodes via peer-to-peer communication protocols [47]. In the context of high-stakes environments like health care, securing consensus across distributed and often mutually untrusting participants is the responsibility of the consensus layer. By establishing agreement on the validity and chronological order of transactions, consensus mechanisms prevent malicious interference and ensure the reliability of shared records [48,49]. The widely adopted algorithms include Proof of Work (PoW),

XSL•FO
RenderX

Proof of Stake (PoS), and Practical Byzantine Fault Tolerance (PBFT) [16,50,51]. To sustain long-term network participation, the incentive layer introduces economic mechanisms that reward nodes for validating transactions and generating new blocks. This is especially important in medical applications, where incentivized behavior facilitates the uploading of more high-quality medical data, while the authenticity and availability of data directly affects patient outcomes. Building upon these foundations, the contract layer operationalizes complex protocols through smart contracts, which autonomously execute predefined rules and transactions without the need for third-party intermediaries [52]. In the health care sector, this translates into streamlined workflows for processes such as insurance claims, consent management, and secure data access, offering enhanced transparency, reduced administrative overhead, and minimized human error. Finally, the application layer serves as the interface between users and the blockchain, bridging technical functionality with real-world applications. Across industries, this layer has fueled innovations in financial services, supply chain management, and beyond [53,54]. In health care, it enables secure sharing of EHRs, protecting patient privacy, and supporting collaborative research efforts across institutions [55]. By facilitating trusted interactions in environments where data sensitivity and security are paramount, blockchain's multilayered architecture offers an indispensable foundation for integrating advanced paradigms such as FL, thereby unlocking new possibilities for privacy-preserving, distributed medical intelligence.

Through the analysis of the structure and underlying architecture of blockchain, we can better understand that blockchain has unique characteristics that distinguish it from traditional systems, including decentralization, immutability, transparency, traceability, security, anonymity, and high availability [8,13].

These characteristics are particularly valuable in the field of health care. For instance, decentralization eliminates the inherent single point of failure in centralized medical record systems, thereby enhancing the system's resilience and reducing its vulnerability to cyber attacks. This feature is particularly suitable for scenarios with high trust requirements, such as medical record management and supply chain supervision. Immutability and traceability not only ensure that clinical data, diagnostic results, and patient consent records cannot be altered or deleted, guaranteeing the auditing of medical data, but also are crucial for tracking the source of drugs and ensuring the integrity of the supply chain, thereby reducing the risk of counterfeit drugs. Furthermore, blockchain offers a transparent and privacy-protecting environment that enables authorized clinicians, researchers, and insurance companies to verify data sources without exposing sensitive patient information, thereby enhancing accountability in medical research and drug development. The security of blockchain stems from smart contracts, consensus mechanisms, etc, maintaining transaction integrity and preventing malicious nodes from arbitrarily altering the block addition process [56,57].

As illustrated in Multimedia Appendix 3, a comparison of the technical characteristics of blockchain and FL is presented. Simultaneously, given these characteristics, blockchain facilitates trust establishment among untrusted participants in wireless networks [58,59]. Consequently, blockchain demonstrates great potential across various domains, including cryptocurrencies, health care, and the Internet of Things (IoT) [45,60,61]. In health care, blockchain technology enables the storage and verification of IoT data within patients' electronic medical records (EMRs), clinical trials, and sensors, granting patients control over their own medical data [62-64]. During AI-driven sample learning, medical data from various institutions—including x-rays, CT scans, MRI reports, and pathological examinations—are securely stored on the blockchain. Predefined entity and event annotation platforms facilitate data labeling within each institution, followed by model training on in-hospital servers [65,66]. Kordestain et al [67] proposed HapiChain, a telemedicine platform built on a patient-centered blockchain infrastructure, ensuring the security of remote consultations between patients and doctors. In addition, decentralized blockchain solutions, such as Drug-ledger and Med-ledger, have been proposed to enhance traceability and security in the pharmaceutical supply chain [68,69].

Given blockchain's significant advantages in decentralization, privacy protection, data immutability, incentives, and automation, its attributes fit well with FL's requirements for secure data sharing and distributed modeling. Blockchain can serve as a secure and reliable collaborative infrastructure for FL, addressing challenges such as trust deficiency, data integrity, and transparency [4,22,70]. Conversely, the decentralized data processing mechanism of FL can compensate for blockchain's limitations in scalability and computational efficiency. Therefore, integrating blockchain with FL has the potential not only to address their respective challenges but also to unlock new application scenarios and possibilities. The following section examines the necessity and feasibility of integrating blockchain with FL.

Integration of Blockchain and Federated Learning in Health Care

Overview

BCFL has emerged as a promising solution in health care, fostering a balance between privacy protection and data collaboration while unlocking new opportunities for data-driven health care innovation [71,72]. The following section analyzes the complementary strengths and performance of integrating blockchain with FL, illustrating how blockchain mitigates the limitations of FL (as illustrated in Multimedia Appendix 3) and how FL benefits blockchain.

Blockchain Empowers Federated Learning in Health Care

Decentralization Mitigates Single Points of Failure and Scalability Bottlenecks

The traditional architecture of FL primarily depends on a central server to manage and coordinate participants. This architecture is susceptible to single points of failure and can also result in bandwidth and computational resource bottlenecks as the number of clients increases [70,73,74]. In contrast, blockchain's decentralized architecture eliminates reliance on a central server by leveraging a distributed network, allowing automated data

XSL•FO
RenderX

collaboration across multiple nodes and effectively mitigating the risk of single points of failure [13]. In this context, temporary aggregators are selected based on blockchain's consensus mechanisms (eg, PoW or PoS) [8,36]. Moreover, blockchain's Byzantine fault tolerance, which enables the dynamic management of unreliable nodes through consensus mechanisms, further enhances the stability of FL in large-scale distributed environments [75].

Incentives to Enhance Participant Motivation

Traditional FL systems often lack effective incentives, especially in environments where resources are unevenly distributed. High-performing participants may struggle to sustain long-term contributions due to the absence of direct benefits [76,77]. Blockchain's built-in incentive mechanisms address this challenge by rewarding data contributors, validators, and maintainers through tokens or other financial instruments, introducing an economic driver for FL [78]. This incentive model encourages the contribution of high-quality data while discouraging unreliable participation, ultimately improving the overall performance and stability of FL models. For instance, Weng et al [79] proposed an incentive mechanism designed to promote collaboration in training deep learning models. The mechanism introduces two key concepts: compatibility and activity. Compatibility ensures that each participant receives optimized rewards based on their contribution, while activity incentivizes participants to update the local model and aggregate the global model actively. Upon each global model update, rewards are distributed to local devices and miners based on individual contributions. Similarly, Kang et al [77] introduced a reputation-based incentive model to measure client trustworthiness. By leveraging blockchain's immutability, the system ensures distributed reputation management and evaluates participants based on model quality and computational contributions.

Privacy Protection and Attack Resistance Enhancement

Although FL protects raw data privacy, it remains vulnerable to adversarial threats such as poisoning attacks and Byzantine attacks, which can mislead model training and hinder convergence [36,80]. Blockchain strengthens FL security through its immutability and tamper-proof nature. Its authentication mechanisms detect and exclude malicious nodes, ensuring that only authorized participants can access FL data, thereby enhancing privacy protection. Moreover, blockchain's cryptographic techniques and anonymity mechanisms reduce the risks of background knowledge attacks and collusion attacks [78]. Additionally, its consensus mechanism ensures data consistency across all nodes while using sophisticated algorithms to prevent malicious nodes from compromising the network [81,82]. For example, in medical data collaboration, blockchain records the training processes and contributions of each participating hospital, ensuring both data integrity and privacy while mitigating risks of data contamination and adversarial attacks. Shayan et al [83] proposed a multi-Krum consensus mechanism to counter poisoning attacks by electing a validation peer committee that filters out malicious model updates. Similarly, Chen et al [84] used a blockchain-based validation voting mechanism, where nodes vote on model update validity and remove malicious devices based on consensus results.

Transparency and Auditability

Blockchain's transparency and auditability effectively address the challenges of trust deficits and compliance difficulties in FL. By leveraging blockchain's transparent data-sharing mechanisms, FL participants can verify the source and integrity of health care data and model updates in real time, ensuring fairness and reliability in contributions [72]. Additionally, blockchain's immutable records establish an accountability framework for FL [33,85]. These records facilitate anomaly detection and responsibility attribution throughout the model training process, strengthening compliance and governance mechanisms. This transparent and auditable nature not only fosters trust in collaborative learning but also provides a technological foundation for regulatory compliance.

Automated Management With Smart Contracts

Smart contracts enable automated execution of key processes in FL, including model update sharing, model update validation, and global model aggregation. By enforcing predefined rules, smart contracts eliminate human intervention, ensuring an unbiased and tamper-proof process. Moreover, they can dynamically allocate resources and rewards through conditional triggering mechanisms—such as when a model reaches an expected accuracy or when a node successfully completes a specific task [86,87]. This automation enhances the autonomy and reliability of FL while ensuring transparency and fairness through open code logic. By reducing administrative overhead and mitigating trust concerns, smart contracts introduce a novel and efficient approach for managing decentralized collaborative learning [33].

How Federated Learning Can Benefit Blockchain

Enhancing Blockchain Consensus Efficiency

The blockchain consensus mechanism, while ensuring network security and data consistency, is often associated with substantial computational costs and energy consumption, a challenge that is particularly serious in the PoW mechanism [8,88]. PoW relies on miners solving complex hashing operations to compete for block generation, necessitating the continuous operation of high-performance computing hardware, which in turn results in substantial global energy consumption. Studies indicate that the annual energy consumption of the Bitcoin network is comparable to that of a small- to medium-sized country. This highly inefficient competition leads to an enormous waste of computational resources—only the first miner to discover a valid hash can package the transaction and claim the reward, rendering all other computational efforts futile. Furthermore, the excessive energy consumption of PoW constrains blockchain's sustainable adoption in critical domains such as health care data management and edge computing, necessitating the development of more energy-efficient consensus optimization strategies. In this context, FL offers a novel approach to optimizing blockchain consensus mechanisms. Integrating the blockchain consensus process with FL allows miners to contribute to model training while competing for block validation, effectively repurposing computational resources that would otherwise be wasted. This approach not only reduces energy consumption but also enhances

XSL•FO
RenderX

computational resource efficiency, rendering the consensus process more practically valuable [89].

Facilitating Cross-Chain Data Collaboration

As blockchain applications continue to expand, the demand for data exchange across different blockchains has grown, making cross-chain technology a crucial solution to addressing "data silos" across various domains. FL and its variants, such as federated transfer learning, can establish a unified model collaboration framework across different blockchain networks, enabling privacy-preserving data sharing and joint modeling. By maintaining a shared ML model, disparate blockchains can collaborate while preserving autonomy and privacy, thereby facilitating cross-chain applications in finance, health care, and other sectors [90].

Enhancing Blockchain Scalability

Blockchain faces storage and computational bottlenecks when handling large-scale data. FL, by adopting a local training model that eliminates the need to upload raw data to the blockchain, significantly reduces on-chain storage demands. Additionally, FL alleviates blockchain's computational burden by distributing processing tasks among participating nodes, thereby providing a scalable foundation for large-scale collaboration.

Architectural Frameworks for Integrating Blockchain and Federated Learning

A BCFL typically adopts one of three architectural paradigms: fully coupled, flexibly coupled, and loosely coupled architectures [41,85,91]. These architectures differ in terms of the degree of coupling between blockchain nodes and FL clients, each offering unique characteristics in function allocation, resource usage, and system structure.

Fully Coupled BCFL

The fully coupled architecture represents a highly integrated design, wherein FL clients simultaneously function as blockchain nodes, assuming dual roles. Consequently, each node is responsible for local model training, update validation, global model aggregation, and new block generation. These tasks are executed on a single node, fostering a fully decentralized collaborative model [36,74,79,92].

Global model aggregation can be carried out either by selected nodes or collaboratively by all nodes, depending on the network's design strategy. Moreover, the blockchain's distributed ledger not only records local model updates but also stores global models and other relevant information generated during training, ensuring data integrity and traceability.

Flexibly Coupled BCFL

The flexibly coupled architecture achieves higher design flexibility by separating FL clients from blockchain nodes. In this architecture, FL clients primarily handle local data collection and model training, whereas blockchain nodes are responsible for validating model updates, storing the global model, and maintaining the ledger [2,40,93]. The blockchain can aggregate global models via selected nodes, which typically possess superior computing resources and reliability, thereby reducing resource consumption and enhancing system efficiency. Alternatively, aggregation can be performed collaboratively by all nodes, achieving full decentralization and mitigating the risk of a single point of failure.

This architecture significantly lowers the resource requirements for FL clients, allowing them to function in different network environments while preserving blockchain's inherent advantages in data security and transparency. Due to its high adaptability, this architecture has become a preferred choice for large-scale distributed systems, such as health care data sharing and cross-organizational collaboration.

Loosely Coupled BCFL

The loosely coupled architecture further weakens the coupling between blockchain nodes and FL clients by optimizing functional allocation. FL clients primarily perform local model training and upload updates to the blockchain for validation, whereas the blockchain handles authentication, model update validation, and participant reputation management.

In this architecture, the blockchain does not store model updates but instead records only reputation-related data. A reputation mechanism is implemented as a key criterion for assessing participant reliability, thereby incentivizing them to contribute high-quality data and updates [77,94,95]. This design enhances system scalability by alleviating storage pressure on the ledger while ensuring the trustworthiness of participant behavior.

Workflow in BCFL

Overview

In BCFL systems, the flexibly coupled architecture has emerged as the predominant choice for real-world applications due to its optimal balance of efficiency and adaptability. By separating FL clients from blockchain nodes, this architecture allows them to operate on different networks and devices, thereby reducing system communication overhead and latency. Additionally, it alleviates the computational and storage burden on client devices while preserving key advantages such as data privacy protection and blockchain-based verification, ultimately achieving an optimal balance between efficiency and privacy. Leveraging these advantages, the flexibly coupled architecture has demonstrated significant potential in practical applications, including medical data sharing and cross-organizational collaboration.

As illustrated in Figure 5, the following section focuses on the specific workflows of mainstream BCFL frameworks, analyzing their distinct advantages in practical applications.

XSL•FO
RenderX

Figure 5. Flexibly coupled blockchain-based federated learning architecture and workflow. gRPC API: Google Remote Procedure Calls – Application Programming Interface; REST-API: Representational State Transfer – Application Programming Interface.

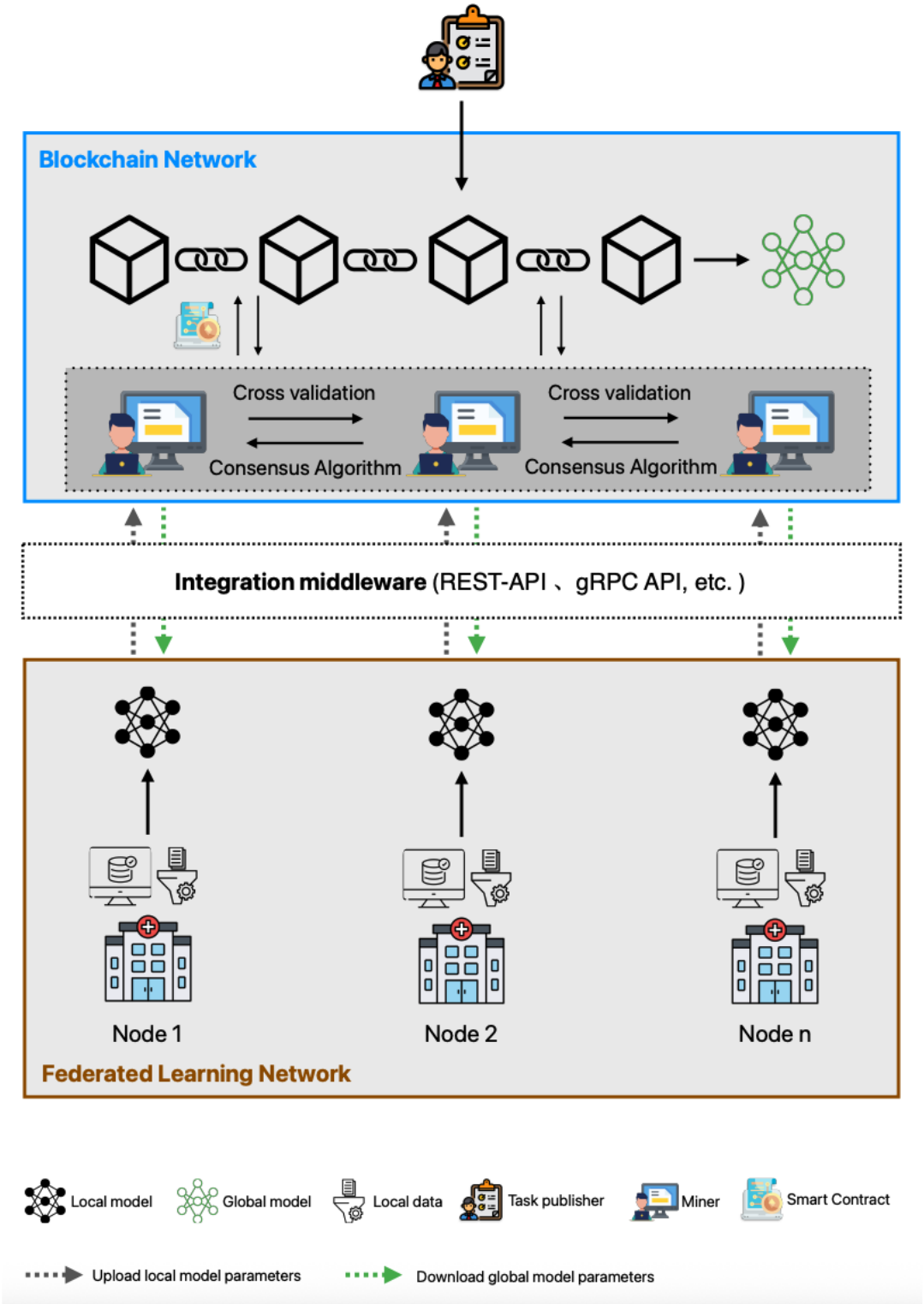

Task Release

Task initiators release FL tasks and requirements on the blockchain, specifying details such as data volume and type, hardware specifications, and the number of training rounds. Leveraging blockchain's transparency and decentralization, this process ensures fair and open task distribution while fostering participant trust.

Local Model Training and Update Transmission

Each FL client downloads the initial global model from the blockchain, after which it preprocesses local data, extracts features, and uses this data for model training, subsequently generating local model updates. These updates are then transmitted to the blockchain network in encrypted form.

XSL•FO
RenderX

It is important to note that, in the flexibly coupled architecture, FL clients and blockchain nodes operate within different networks and systems, each with clearly defined responsibilities. Therefore, this architecture heavily relies on integrated middleware, which serves as a communication bridge and coordinator between the two components. In a research study, Lamken et al [96] used REST-API (Representational State Transfer – Application Programming Interface) for communication with the Hyperledger Fabric blockchain, facilitating the recording and incentivization of gradient uploads. Additionally, the Remote Procedure Calls (RPC) protocol developed by Google, known as the gRPC API, facilitates data exchange between FL clients and the Ethereum blockchain network [33,96].

Blockchain Node Verification Update

Blockchain nodes (ie, miners) verify the uploaded model updates using a predefined validation mechanism. Concurrently, miners exchange their validated local model updates with each other. A consensus algorithm guarantees that only validated updates contribute to the global model aggregation.

Global Model Aggregation

Subsequently, the blockchain selects an interim leader among its nodes through a consensus mechanism. The selected node(s) then collect verified model updates and aggregate them to construct the global model [36]. The flexibly coupled architecture enables this process to be executed by selected nodes or collectively by all nodes, thereby opening up the possibility of optimizing efficiency across various scenarios.

New Block Generation and Model Storage

Validated model updates and global models are packaged by selected blockchain nodes to generate new blocks. Upon adding the block header information, the legitimacy of the block is verified through a consensus mechanism among the nodes.

Distributed Ledger Update

The newly generated blocks are broadcast across the entire network, and all blockchain nodes update their local ledgers after verification. This process ensures the transparency and traceability of the global model and its associated information throughout the network.

Reward Distribution and Incentives

The system allocates rewards, such as cryptocurrency or reputation scores, based on client performance. This incentive mechanism not only motivates participants to contribute high-quality updates but also deters malicious behavior, thereby enhancing the accuracy and reliability of the model.

Global Model Download

After the training is completed, all participating clients can download a newly generated block containing the updated global model parameters from the blockchain. Clients can then independently decide whether to participate in the next training round based on their specific needs. This mechanism enhances both system flexibility and participant autonomy.

As illustrated in Table 1, a comparison of BCFL integration architectures is presented. In future practical applications, the selection of a specific architecture must be carefully evaluated based on scenario requirements, resource constraints, and design objectives to achieve optimal collaboration and technical performance.

XSL•FO
RenderX

Table . Comparison of blockchain-based federated learning integration architectures.

Architecture type	Characteristics	Advantages	Disadvantages	Applicable scenarios
Fully coupled BCFL[a]	• High integration: FL[b] clients and blockchain nodes are fully merged • Fully decentralized: All nodes work together through a consensus mechanism	• High transparency: All transactions and model updates are recorded on the blockchain • Strong security: Resistant to single-point failures and man-in-the-middle attacks • Strict control: Highly controlled over data and models	• High resource demand: Requires significant computational and storage resources • High network complexity: All nodes participate in the consensus mechanism, and the network complexity is high • Intensive coordination: Frequent internode communication is required	• Large-scale distributed environments: Suitable for large medical institutions and research centers • Strict control and security requirements: Scenarios involving the sharing and analysis of sensitive medical data [97,98]
Flexibly coupled BCFL	• Functional separation: FL clients operate independently from blockchain nodes • Computational offloading: Model aggregation occurs at selected nodes	• Enhanced efficiency: Optimized allocation of computing and storage resources • Greater flexibility: Can be adapted to different application scenarios • Improved scalability: Supports large-scale data sharing and collaborative learning	• Complex coordination: The responsibilities of clients and nodes are separated, and complex coordination and management mechanisms are required • Centralization risks: Use of a centralized aggregator may introduce a single point of failure • Node selection challenges: Issues such as node selection criteria and fairness are involved	• Dynamic collaboration settings: Suitable for cross-institutional medical data sharing, IoMT[c] device management [40,99]
Loosely coupled BCFL	• Minimal integration: FL clients and blockchain nodes operate independently • Lightweight blockchain: Primarily used for identity authentication and reputation management	• Reduced overhead: Reduce the operating cost of blockchain and improve system performance • Enhanced privacy: Reduce on-chain storage pressure and improve scalability • Optimized incentives: Reputation-based mechanisms encourage high-quality contributions	• Lower decentralization: May still rely on trusted central nodes for model aggregation • Data integrity risks: Blockchain does not store model updates	• Resource-constrained environments: Suitable for wearable medical devices and real-time health monitoring • Small-scale institutions: Ideal for personal mobile health applications and smaller clinics [100]

[a]BCFL: blockchain-based federated learning.

[b]FL: federated learning.

[c]IoMT: Internet of Medical Things.

BCFL in Medicine

Overview

As the demand for data-driven technologies in health care continues to grow, the BCFL framework presents significant potential due to its advantages in privacy preservation, data security, and collaborative efficiency. BCFL facilitates cross-organizational data sharing and collaborative analytics, optimizing personalized health care solutions while driving advancements in areas such as telemedicine, IoMT, and public health monitoring, which are shown in Figure 6. In the following section, we will discuss the various applications of BCFL in the medical field and analyze its key role and potential value in addressing real-world challenges.

XSL•FO
RenderX

Figure 6. Blockchain-based federated learning framework for different domains in health care. EMR: electronic medical record.

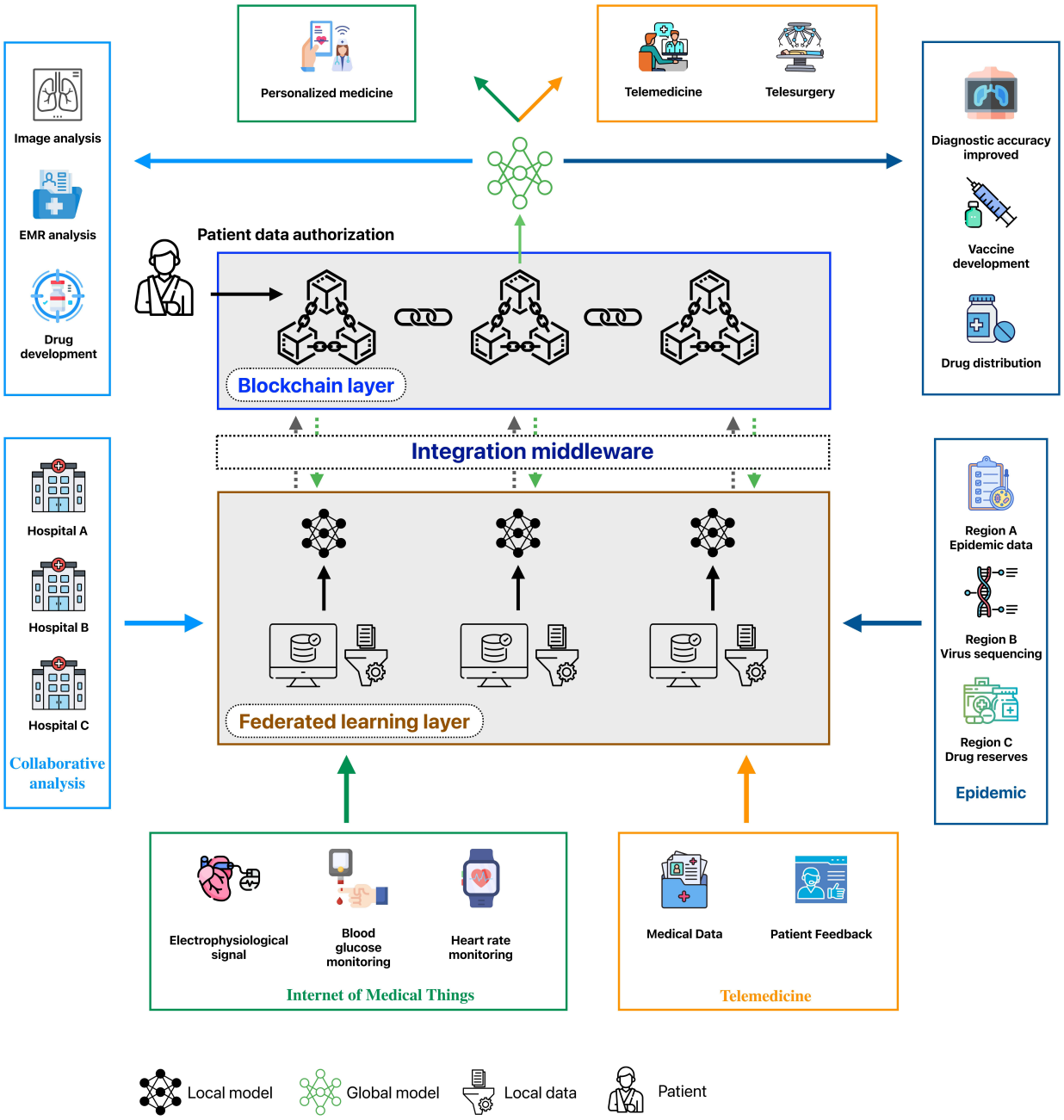

Cross-Institutional Medical Data Sharing and Collaborative Analysis

In modern health care, data serves as a crucial resource for driving innovation and enhancing treatment efficacy. However, data sharing among health care institutions is hindered by concerns over privacy, data security, and regulatory compliance. The integration of blockchain and FL offers an innovative solution for cross-institutional health care data sharing and collaborative analysis. While numerous studies have demonstrated the feasibility of BCFL in various medical domains, the strength of evidence supporting these applications varies considerably, and critical challenges remain.

Several studies have investigated BCFL in the context of chronic disease management, particularly diabetes prediction. Hasan et al [101] developed a blockchain-FL framework that reported a 15% improvement in predictive performance across multiple metrics. Although these results are encouraging, the framework relied primarily on public diabetes datasets with limited diversity, raising questions about its generalizability to heterogeneous real-world populations. Similarly, Moulahi et al [102] evaluated a BCFL model on the Pima Indians Diabetes dataset, achieving a multilayer perceptron accuracy rate of 97.11% and an average FL accuracy rate of 93.95% while protecting privacy. Yet, the reliance on small, well-characterized datasets constrains the robustness of the findings. Taken together, these studies suggest that BCFL holds promise for chronic disease prediction, but the supporting evidence remains preliminary, and large-scale multi-institutional validation is still lacking.

XSL•FO
RenderX

In the realm of the IoMT, Ramani et al [103] introduced the ODMSM-FL (Optimized Data Management and Secured Federated Learning) approach to address secure data storage and exchange using EHR datasets from HealthData.gov. This research report presents a set of numerical results on key performance indicators: transaction throughput (102.75 Kbps), data retrieval delay (64.02 ms), security (88.97%), and accuracy (86.32%). The research results highlight the great potential of ODMSM-FL in effectively addressing the urgent data management and security issues in IoMT. However, the system was evaluated under controlled experimental conditions rather than real-world clinical settings, limiting its immediate applicability. By contrast, research in medical imaging tasks, such as brain tumor segmentation, has placed greater emphasis on model accuracy and privacy preservation. For example, Kumar et al [104] proposed a permissioned blockchain-based federated framework with quality-aware model aggregation, achieving improved segmentation metrics on the BraTS 2020 dataset. Specifically, compared with the baseline method, our approach increased the Dice similarity coefficient of enhanced tumors by 1.99% and reduced the Hausdorff distance of the overall tumor by 19.08%. Although the study demonstrated methodological innovation, it still relied on benchmark imaging datasets rather than prospective clinical data, which restricts the strength of evidence regarding its clinical translatability.

Other investigations have targeted specific diagnostic applications. Heidari et al [105] designed the FBCLC-Rad (Federated Learning–Enabled Blockchain CapsNets Lung Cancer Radiologist) framework for lung cancer detection, achieving near-perfect accuracy on nodule identification tasks. This technology achieved an accuracy rate of 99.69% with the lowest classification error. While technically impressive, results derived from controlled datasets may not fully reflect the complexity of real-world diagnostic workflows. Liang et al [106] extended BCFL applications to clinical trials, where blockchain ensured data authenticity and traceability, and FL supported participant screening across organizations. This study represents an important step toward integrating BCFL into the clinical research pipeline, but it remains largely conceptual, with limited empirical validation in actual trial environments.

A growing body of work has also highlighted the integration of BCFL with EMRs to facilitate precision medicine [99,107-110]. This approach has demonstrated significant effectiveness in enhancing diagnostic accuracy, optimizing treatment planning, identifying patient subgroups for clinical trials, and accelerating the development of novel therapeutics. Within the paradigm of precision medicine, such a framework facilitates a transition from the traditional “one-size-fits-all” treatment model to a more personalized and adaptive intervention strategy. However, despite their conceptual appeal, most studies are limited to prototype frameworks or simulations and have not yet undergone prospective evaluation in clinical practice. As such, the current evidence supporting BCFL in EMR-based precision medicine remains promising but immature.

Overall, existing literature demonstrates the conceptual feasibility and technical potential of BCFL for cross-institutional health care data sharing and analysis. Nevertheless, the evidence base is uneven: studies using small public datasets provide only preliminary support, while those addressing more complex tasks such as imaging or clinical trials often lack real-world validation.

Internet of Medical Things

The IoMT is an advanced technological ecosystem that integrates internet technology with medical devices, enabling real-time data collection, exchange, and analysis to enhance clinical decision-making, disease prevention, and patient care. With the rapid advancement of the IoMT, the traditional hospital-centric model has evolved into a patient-centered health care system driven by comprehensive clinical analysis. The widespread adoption of IoMT enables individuals to conveniently monitor their health at home, thereby streamlining diagnosis and treatment while allowing patients to enjoy more efficient, personalized health care. However, despite its growing adoption, IoMT in health care is also facing significant challenges. One of the primary concerns is data privacy and security. IoMT devices collect and transmit vast amounts of sensitive data, including patient identities, insurance details, and payment information. Once these data are accessed by malicious individuals, it could lead to serious consequences. Moreover, the absence of standardized security protocols among IoMT devices exacerbates the risks of data leakage and device manipulation. Consequently, device manufacturers and health care institutions face immense pressure to ensure data privacy and regulatory compliance. Additionally, IoMT devices frequently encounter challenges such as high computational complexity, elevated costs, and communication delays due to resource limitations.

To address these limitations, several studies have explored the integration of blockchain and FL in IoMT. For instance, Rahman et al [111] proposed a lightweight hybrid FL framework that leverages blockchain to secure health data provenance and uses smart contracts to coordinate model training and trust management. While this framework demonstrates theoretical scalability and robust traceability, its evaluation was primarily conducted in simulated settings, raising concerns about its applicability in heterogeneous and large-scale real-world health care environments.

Muazu et al [112] combined BCFL with edge computing to improve resource allocation, reduce computational costs, and enhance IoMT data security. By offloading intensive computations to edge nodes, the study reported reductions in latency and energy consumption. Meanwhile, the performance of the proposed model offers a higher precision of 83% and an accuracy rate of 78%. However, the framework largely relies on linear regression as the global learning model, which—although interpretable and useful for basic clinical predictions—may not adequately capture the complexity of real-world multimodal medical data. Compared with Rahman et al [111], this work provides stronger performance evidence in terms of latency and efficiency but weaker generalizability for complex clinical prediction tasks.

Dhasaratha et al [113] extended BCFL by incorporating reinforcement learning and distributed computing to improve risk factor monitoring and COVID-19 patient prediction. The dynamic optimization enabled by reinforcement learning is a

XSL•FO
RenderX

notable strength, allowing adaptive performance improvements in evolving environments. Compared with the previous two documents, this approach offers methodological novelty but lacks equally rigorous performance benchmarking across standard IoMT metrics.

To tackle fraud detection and scheduling issues, Lakhan et al [114] introduced the FL-BETS (Federated Learning–Based Blockchain-Enabled Task Scheduling) framework, which integrates BCFL with dynamic heuristics for task scheduling across fog and cloud nodes. The experimental results show that the framework, from the initial 60:80 fraud delay ratio to a 10:10 ratio, demonstrates better performance in energy-delay trade-offs and antifraud behavior. Yet, the framework emphasizes technical efficiency rather than clinical utility, and its reliance on hard and soft scheduling constraints may limit adaptability in unpredictable medical environments. Compared with Dhasaratha et al [113], which focuses on patient-specific outcomes, this work provides stronger technical validation but weaker clinical alignment.

In the context of wearable IoMT devices, Baucas et al [115] designed a system that integrates FL with a private blockchain in a fog computing architecture to enhance privacy and adaptability. Their framework demonstrated efficiency in resource-constrained environments and produced accurate predictive models while safeguarding patient privacy. Unlike the literature above, this study directly validated its framework on wearable health care devices, thereby providing more immediate clinical relevance. However, the scalability of the approach for larger IoMT networks remains uncertain.

Overall, these studies collectively highlight the potential of BCFL in overcoming IoMT's inherent privacy, security, and performance limitations. Yet, their evidence strength varies significantly: some emphasize theoretical frameworks validated in simulations [111], while others demonstrate more robust experimental performance [112,114] or closer alignment with clinical practice [115].

Public Health Surveillance and Epidemiological Forecasting

The global outbreak of COVID-19 highlighted the limitations of existing surveillance infrastructures, particularly the inability to provide accurate, real-time epidemic monitoring. Traditional methods often struggle with the rapid spread and variability of epidemics, while stringent privacy requirements hinder effective collaboration across institutions. For instance, during global outbreaks such as COVID-19, the inability of national and regional health care organizations to efficiently integrate data has impeded comprehensive analyses of epidemic progression [116]. This phenomenon of data silos delays the formulation of precise response strategies and undermines the efficiency of vaccine distribution and health care resource allocation. Therefore, achieving efficient and secure data integration while preserving privacy has emerged as a critical challenge in public health. The BCFL framework not only integrates anonymized health data from diverse regions but also facilitates the efficient construction of predictive models for epidemic spread. This framework enables multiple health care organizations and research institutions to collaborate securely without compromising patient privacy, thereby providing robust data support for early epidemic detection, transmission trend analysis, and the formulation of intervention strategies. However, the strength of evidence supporting BCFL frameworks varies considerably across different studies, depending on data scale, validation methods, and implementation feasibility.

FedMedChain [117] represents an early attempt to address these challenges. By combining blockchain with FL and leveraging the IoMT, it enhances the trustworthiness of public health communication and mitigates risks associated with centralized data transmission. Its contribution lies in demonstrating that blockchain can ensure data transparency and tamper resistance while maintaining privacy. Nevertheless, FedMedChain was mainly verified through small-scale simulation experiments rather than real-world deployments, which limited the strength of the evidence and its direct clinical applicability.

In contrast, Kumar et al [118] explored a BCFL framework for processing heterogeneous CT images using capsule networks. This method achieves a high detection accuracy on the CC-19 dataset, and its research results include 98.68% specificity and 98% sensitivity. While the study demonstrates the feasibility of applying BCFL to medical imaging and highlights the benefits of privacy-preserving collaboration, the restricted dataset size and limited institutional diversity weaken its external validity. Compared with FedMedChain, this workplaces greater emphasis on model performance but provides weaker evidence regarding scalability and generalizability.

Durga and Poovammal [119] extended this direction by proposing the FLED-Block (Federated Learning–Ensembled Deep Learning Blockchain Model) framework, which integrates blockchain with FL for COVID-19 prediction using multisource heterogeneous CT datasets. This framework improves the classification accuracy by using capsule networks for feature extraction and extreme learning machines for efficient classification. The research results include an accuracy of 98.2%, a precision of 97.3%, and a recall rate of 96.5%. Importantly, it integrates blockchain to share model weights without the need to exchange raw data, thereby resolving privacy issues. Compared with Kumar's study [118], FLED-Block was supported by evidence from more hospitals, providing stronger validation. However, the latency of blockchain is regarded as a limitation, raising questions about its applicability in real-time clinical diagnosis. Therefore, although this framework demonstrates outstanding technical performance, its transformation in emergency health care settings remains uncertain.

Abdel-Basset et al [120] proposed the blockchain-based federated learning for pandemic diagnosis (BFLPD) framework, which stands out for its focus on system security and robustness in the context of smart cities. Unlike FedMedChain and FLED-Block, BFLPD combines more encryption technologies, including secure aggregation, homomorphic encryption (Cheon-Kim-Kim-Song scheme), and consensus mechanisms (PBFT), to mitigate malicious attacks and improve reliability. The classification accuracy of BFLPD reaches 95.14%, exceeding the benchmark set by the most advanced distributed models. In addition, this framework also demonstrated

XSL•FO
RenderX

significant precision and recall rates (95.26% and 95.77%, respectively) and a relatively high F_1-score (95.52%). Meanwhile, the authors incorporated heat map visualization, further enhancing its clinical application value. This framework provides stronger evidence than earlier works, as it addresses adversarial threats that are often overlooked in BCFL studies. Nevertheless, its reliance on complex cryptographic and consensus algorithms introduces implementation challenges, such as high computational overhead, which could hinder real-world deployment.

Telemedicine and Telesurgery

In recent years, telemedicine has experienced rapid advancements, especially in response to the global COVID-19 pandemic, which has significantly increased its role in modern health care systems. Telemedicine leverages modern information and communication technologies to facilitate medical information exchange across geographic boundaries, encompassing various applications such as remote diagnosis, remote consultation, remote treatment, and continuous health monitoring [5]. By providing on-demand, personalized health care services, telemedicine optimizes medical resource allocation, effectively addressing the challenge of unequal distribution of traditional health care resources and ensuring medical support for patients in remote or underserved areas. Despite its potential, telemedicine faces several critical challenges in practical implementation:

- Data Security and Privacy Risks: Most telemedicine systems rely on centralized cloud servers to store patient health data, making them vulnerable to single points of failure.
- Lack of Data Access Control Mechanisms: Many existing telemedicine platforms do not offer a robust data access control framework, meaning that once patient data are uploaded to the cloud, patients often lose ownership and control over their own health records.
- High Infrastructure and Computational Costs: Telemedicine demands substantial computational resources, high-speed communication networks, and specialized medical equipment, particularly for real-time diagnosis and treatment.

Building upon these advantages, recent studies have proposed generalized frameworks that integrate blockchain and FL to support secure and scalable telemedicine systems. For example, Hiwale et al [5] highlighted the importance of incorporating privacy-preserving technologies into BCFL, laying the groundwork for reliable, privacy-compliant telemedicine applications. Within such frameworks, blockchain's distributed ledger technology enables decentralized data storage, reducing the risks of single points of failure and data breaches, while ensuring transparency and traceability in data access. Simultaneously, FL enhances data privacy by enabling local model training, thereby minimizing the exposure of sensitive health information. Although valuable as a theoretical framework, this study provides limited experimental validation and thus represents weak evidence for clinical applicability. Gupta et al [121] enhanced trust between patients and providers by designing a smart contract system based on public blockchains. The framework allows patients to retain ownership and fine-grained control of health data, addressing a central shortcoming of conventional telemedicine platforms. Compared to Hiwale et al [5] conceptual work, Gupta et al [121] system offers a more concrete mechanism for authorization and data sharing. Nonetheless, its validation remains restricted to simulation environments, with no real-world deployment or clinical evaluation. As such, while it provides moderate evidence of feasibility, its generalizability remains uncertain.

With the rapid advancement of technology and the increasing improvement of medical demands, traditional telemedicine models are evolving beyond routine diagnosis and treatment. Among these advancements, telesurgery—a critical extension of telemedicine—is emerging as a transformative innovation. However, this technology imposes stringent requirements on real-time data synchronization, precise coordination of surgical equipment, and robust data security, introducing new challenges to the reliability of underlying technological infrastructures. For instance, Chaudjary et al [122] proposed a secure telesurgery system that integrates blockchain and FL with 6G communication networks and the Interplanetary File System protocol. This study demonstrated notable improvements in latency reduction, storage efficiency, and transmission reliability compared with traditional telesurgery systems. Unlike earlier works by Hiwale et al [5] and Gupta et al [121], Chaudhary et al [122] provided more systematic experimental results, suggesting stronger evidence of technical feasibility. However, these results were still derived from controlled simulations rather than real-world surgical environments, and issues such as blockchain latency and computational overhead remain unresolved. Therefore, while the study represents the strongest evidence among current works, its translation into clinical practice requires further validation.

In the past few years, BCFL's research in the health care field has shown significant growth. Multiple studies have confirmed the improvement of model performance in public datasets or experimental environments, such as enhancing the accuracy of disease prediction, strengthening image diagnostic capabilities, or improving edge device management.

However, when examined from the perspective of evidence, most of these achievements are still at the stage of simulation experiments, prototype systems, or preclinical validation. Research is usually based on controlled datasets or static data scenarios, and there is a significant gap between the model performance and the actual clinical diagnosis and treatment process. Importantly, current literature pays more attention to technical indicators (such as accuracy, Dice, and delay) rather than specific medical endpoints, such as changes in misdiagnosis rates, shortened treatment duration, or improved patient prognosis. Therefore, a direct chain of evidence has not yet been formed between the technical performance of BCFL and its actual medical value.

Furthermore, the BCFL architecture is not inherently compatible with the real deployment environment. The medical system features a complex governance structure, compliance requirements, and heterogeneous infrastructure. However, existing research often assumes node autonomy, network stability, or institutional equivalence, while neglecting key issues

XSL•FO
RenderX

such as data authorization, responsibility division, and system compatibility. Although blockchain consensus, smart contracts, and high-intensity encryption enhance security, they also bring about latency, energy consumption, and maintenance costs, which conflict with the real-time and reliability requirements of clinical practice. These mismatches have led to many frameworks performing well in experimental settings but being difficult to migrate to real medical scenarios.

Given the above limitations, relying solely on technical performance indicators cannot accurately reflect the maturity of BCFL in medical scenarios. To more systematically assess the application level of existing research, we adopted an evidence stratification strategy to categorize the existing literature from multiple dimensions such as architectural innovation, deployment environment, verification depth, clinical relevance, and potential risks. This stratification aims to reveal the gap between "technical performance" and "medical practice value," identify the most critical bottlenecks in the process from conceptual framework to clinical validation, and provide directional references for future research design. Table 2 summarizes the evidence stratification of typical BCFL studies in the health care field.

XSL•FO
RenderX

Table . Evidence stratification of blockchain-based federated learning studies in health care.

Reference	Application domain	BCFL[a] architecture/contribution	Deployment environment	Validation depth	Clinical relevance/risk	Evidence level/maturity
[101]	Cross-Institutional Medical Data Sharing (Chronic Disease)	Proposed a decentralized and privacy-preserving collaboration framework that integrates blockchain and FL[b], enhancing the predictive performance of diabetes models while ensuring data security and reducing communication overhead	Evaluation on public dataset (unspecified diabetes data)	Retrospective data validation	Population heterogeneity is not covered; it is difficult to extrapolate to real clinical patients	Level 3: Preclinical
[102]	Cross-Institutional Medical Data Sharing (Chronic Diseases)	Developed a blockchain-integrated FL mechanism to enhance IoMT[c] data privacy and improve diabetes prediction accuracy, achieving 97.11% accuracy with a multilayer perceptron model	Evaluation on public dataset (Pima Indians Diabetes)	Retrospective data validation	Relying on a small and well-defined dataset limits the robustness of the findings	Level 3: Preclinical
[103]	Cross-Institutional Medical Data Sharing (Data Management)	Introduced the ODMSM-FL[d] framework, which optimizes storage, management, and privacy protection for IoMT data, enhancing data security and system efficiency	Evaluation on public EHR[e] dataset (HealthData.gov)	Controlled experimental conditions	Data latency, human-machine device heterogeneity; The real IoMT network is uncontrollable	Level 2: Prototype validation
[104]	Cross-Institutional Medical Data Sharing (Medical Imaging)	Designed a blockchain-powered FL framework for brain tumor segmentation using 3D U-Net, achieving significant improvements in Dice similarity coefficient and Hausdorff distance	Evaluation on public benchmark dataset (BraTS 2020)	Retrospective data validation	Relying on benchmark datasets rather than prospective clinical data limits the clinical translational application	Level 2: Prototype validation
[105]	Cross-Institutional Medical Data Sharing (Medical Imaging)	Proposed the FB-CLC-Rad[f] framework, integrating CapsNets, blockchain, and FL to enhance lung cancer nodule detection accuracy in CT scans, reaching 99.69% accuracy	Evaluation on public and local dataset (Cancer Imaging Archive [CIA], Kaggle Data Science Bowl [KDSB], LUNA 16, and local datasets)	Retrospective data validation	The process of not covering the real image; lack of doctor decision-making and workflow verification	Level 2: Prototype validation

XSL•FO
RenderX

Reference	Application domain	BCFL[a] architecture/contribution	Deployment environment	Validation depth	Clinical relevance/risk	Evidence level/maturity
[106]	Cross-Institutional Medical Data Sharing (Drug Discovery)	Designed RahasakML, a decentralized blockchain-FL platform enabling multi-institutional collaboration with enhanced transparency and security in drug discovery	Theoretical analysis; limited empirical validation	Conceptual framework	Empirical verification in actual test environments is limited	Level 1: Conceptual
[107]	Cross-Institutional Medical Data Sharing (EMR[g])	Integrated FL and blockchain for cloud-based medical record recommendation systems, leveraging Hyperledger Fabric, IPFS[h], LightGBM, and N-Gram models for collaborative learning	Evaluation on public EHR dataset (not specified)	Simulated	Limited to prototype frameworks or simulations, not prospectively evaluated in clinical practice	Level 2: Prototype validation
[108]	Cross-Institutional Medical Data Sharing (EMR)	Proposed a blockchain-FL framework for EHR privacy protection, achieving 92.5% global model accuracy and 88.33% local model accuracy using a deep neural network	Evaluation on public EHR dataset (Chronic Kidney Disease [CKD] dataset [UCI Machine Learning Repository])	Retrospective data validation	Limited to prototype frameworks or simulations, not prospectively evaluated in clinical practice	Level 3: Preclinical
[109]	Cross-Institutional Medical Data Sharing (EMR)	Used lightweight encryption and FL to secure EHR data in an Ethereum test environment, reducing reliance on trusted third parties	Evaluation on public EHR dataset (Simulation in Ethereum test environment)	Simulated	Limited to prototype frameworks or simulations, not prospectively evaluated in clinical practice	Level 2: Prototype validation
[110]	Cross-Institutional Medical Data Sharing (EMR)	Combined CNN[i] and blockchain-FL to enhance EHR data security and detect abnormal user behaviors automatically	Evaluation on public EHR dataset (Python-based simulation)	Simulated	Limited to prototype frameworks or simulations, not prospectively evaluated in clinical practice	Level 2: Prototype validation
[99]	Cross-Institutional Medical Data Sharing (EMR)	Explored blockchain-FL applications in precision medicine, emphasizing diagnostic accuracy, treatment optimization, clinical trial subpopulation identification, and drug development acceleration	Evaluation on public EHR dataset (not specified)	Simulated	Limited to prototype frameworks or simulations, not prospectively evaluated in clinical practice	Level 2: Prototype validation

XSL•FO
RenderX

Reference	Application domain	BCFL[a] architecture/contribution	Deployment environment	Validation depth	Clinical relevance/risk	Evidence level/maturity
[111]	IoMT (Data Security)	Proposed a lightweight hybrid FL framework with blockchain smart contracts for edge training plan management, trust evaluation, and authentication in IoMT networks	Evaluation on public COVID-19 dataset (not specified)	Simulated	The device has heavy computational burden, high system complexity, and difficult clinical translation	Level 1: Proof-of-concept
[112]	IoMT (Data Management)	Developed a blockchain-FL system leveraging edge computing and Paillier encryption to securely manage medical resource transactions in IoMT environments	Evaluation on public dataset (unspecified diabetes data)	Retrospective data validation	The device has heavy computational burden, high system complexity, and difficult clinical translation	Level 2: Prototype validation
[113]	IoMT (Data Security)	Introduced a distributed reinforcement learning method integrating blockchain and FL for improved data privacy and security in IoMT applications	Evaluation on public COVID-19 dataset (not specified)	Simulated	The device has heavy computational burden, high system complexity, and difficult clinical translation	Level 2: Prototype validation
[114]	IoMT (Data Security)	Proposed the FL-BETS[j] framework, leveraging fog computing and blockchain to minimize energy consumption and latency while enhancing fraud detection in health care	Evaluation on Private dataset focusing on medical insurance fraud (Kaggle)	Simulated	The device has heavy computational burden, high system complexity, and difficult clinical translation	Level 1: Proof-of-concept
[115]	IoMT (Data Security)	Developed a fog computing IoT platform that integrates FL and private blockchain technology to enhance privacy protection in wearable IoMT devices	Evaluation on human activity recognition dataset (UCI[k] Machine Learning Library)	Simulated	The device has heavy computational burden, high system complexity, and difficult clinical translation	Level 2: Prototype validation
[117]	Public Health (COVID-19 Imaging)	Proposed a blockchain-FL-based IoMT architecture for COVID-19 detection and epidemic management; the architecture enhances data privacy through FL and ensures data transparency and immutability via blockchain	Evaluation on public COVID-19 dataset (Centers for Disease Control [CDC] data)	Simulated	High real-time requirements in epidemic environment; blockchain delay is not resolved	Level 1: Proof-of-concept

XSL•FO
RenderX

Reference	Application domain	BCFL[a] architecture/contribution	Deployment environment	Validation depth	Clinical relevance/risk	Evidence level/maturity
[118]	Public Health (COVID-19 Imaging)	Developed a blockchain-based FL framework for COVID-19 detection, using Capsule Networks for image segmentation and classification to enhance data privacy and model accuracy	Evaluation on Private COVID-19 dataset (CC-19)	Retrospective data validation	The limited scale and types of data restrict the generalization ability of the model	Level 3: Preclinical
[119]	Public Health (COVID-19 Imaging)	Introduced FLED-Block[l], a blockchain-based FL model integrating Capsule Networks for image feature extraction and extreme learning machines (ELM) for classification, achieving high accuracy with strong privacy protection	Evaluation on public COVID-19 dataset (CT data from multiple hospitals)	Retrospective data validation (multi-source datasets)	The computational complexity and the feasibility of actual deployment require further research	Level 3: Preclinical
[120]	Public Health (Pandemic Diagnosis)	Designed BFLPD[m], a blockchain-FL framework for epidemic diagnosis in smart cities, particularly for COVID-19; the framework ensures secure model aggregation and enhances global model integrity and efficiency	Evaluation on public ultrasound COVID-19 dataset (POCUS, ICLUS-DB, and COVIDx-US)	Retrospective data validation (multi-source datasets)	The huge computational overhead and the high resistance to actual deployment	Level 3: Preclinical
[5]	Telemedicine (Telemedicine System)	Proposed a blockchain-FL application framework for telemedicine, analyzing how these technologies improve data accessibility, security, and privacy in remote health care	Based on theoretical simulations or examples	Conceptual framework	Mainly focuses on the theoretical framework, the lack of a real-world application cases	Level 1: Proof-of-concept

XSL•FO
RenderX

Reference	Application domain	BCFL[a] architecture/contribution	Deployment environment	Validation depth	Clinical relevance/risk	Evidence level/maturity
[121]	Telemedicine (Remote Surgery System)	An intelligent remote surgery framework named BITS[n], which is based on blockchain and artificial intelligence, is proposed; this architecture integrates blockchain technologies (such as Ethereum and IPFS protocols), 6G communication networks, and federated learning (or AI[o] algorithms), aiming to enhance the security, privacy, and real-time performance of remote surgery systems	Based on theoretical simulations or examples	Simulated	Limited to a simulated environment, no actual deployment or clinical evaluation has been carried out yet	Level 1: Proof-of-concept
[122]	Telemedicine (Remote Surgery System)	Developed a remote surgery system framework leveraging blockchain and FL to enhance data security, reliability, and real-time processing; the framework integrates 6G networks and IPFS for low-latency and high-reliability data transmission	Based on theoretical simulations or examples	Conceptual framework	There is insufficient discussion on the specific implementation details of BCFL in remote surgery and a lack of clinical deployment	Level 1: Proof-of-concept

[a]BCFL: blockchain-based federated learning.
[b]FL: federated learning.
[c]IoMT: Internet of Medical Things.
[d]ODMSM-FL: Optimized Data Management and Secured Federated Learning.
[e]EHR: electronic health record.
[f]FBCLC-Rad: Federated Learning–Enabled Blockchain CapsNets Lung Cancer Radiologist.
[g]EMR: electronic medical record.
[h]IPFS: Interplanetary File System.
[i]CNN: convolutional neural network.
[j]FL-BETS: Federated Learning–Based Blockchain-Enabled Task Scheduling.
[k]UCI: University of California, Irvine.
[l]FLED-Block: Federated Learning–Ensembled Deep Learning Blockchain Model.
[m]BFLPD: blockchain-based federated learning for pandemic diagnosis.
[n]BITS: Blockchain-Driven Intelligent Scheme for Telesurgery System.
[o]AI: artificial intelligence.

As shown in the table, most BCFL studies are still focused on the concept or prototype stage, lacking multicenter real data validation and evaluation corresponding to clinical endpoint indicators. To promote the clinical application of BCFL, improvements need to be made in three aspects: (1) conduct cross-institutional and prospective validations to evaluate the model's performance in real patient populations and medical processes; (2) strike a balance among security, latency, and maintainability to avoid the unavailability caused by simply pursuing complex encryption or on-chain computing; and (3) achieve integration with medical information systems, data governance and regulatory frameworks, and deploy the system

XSL•FO
RenderX

under the premise of clearly defining data responsibilities and authorities. Only when verified under the joint constraints of clinical workflow, patient heterogeneity, and compliance requirements can BCFL gradually evolve from a conceptual technology to a usable medical solution.

Discussion

Challenge

Although BCFL shows great potential in transforming health care data sharing, its deployment in real medical environments is still highly limited. The vision of change for BCFL must be balanced with technical limitations and the specific complexity of applications (particularly the interoperability gap, high implementation costs, and unresolved scalability bottlenecks), which prevent it from being transformed from a conceptual framework into regular clinical practice.

There are many application challenges in the field of medical data sharing, including the lack of system interoperability and the absence of standardized benchmark datasets specifically designed for medical applications. Currently, various health care information systems (eg, EMR systems) use diverse system architectures, data formats, and operational standards, lacking a unified interoperability framework. Moreover, unlike conventional FL research that often leverages open-access datasets such as CIFAR or MNIST, health care data are inherently sensitive, fragmented, and institution-specific, which makes reproducibility and cross-study comparability particularly difficult.

Scalability and communication efficiency also present critical obstacles. As FL tasks expand, the number of health care data sources and the complexity of training increase significantly. However, in practical deployments, the scalability and throughput limitations of blockchain are particularly pronounced. Even in permissioned frameworks such as Hyperledger Fabric, which offer improved throughput, empirical benchmarks still report end-to-end latencies of several seconds per block under moderate workloads. Insufficient mining resources slow down block generation and verification, hindering the efficient execution of large-scale tasks [123]. Moreover, the influx of numerous participants in distributed health care environments amplifies the load on the blockchain network, while the efficiency of existing consensus mechanisms is difficult to meet the demands of health care applications [124]. When applied to BCFL, these constraints imply that each iteration of local training, parameter aggregation, and block creation may introduce cumulative delays that significantly slow model convergence. Furthermore, as the number of blockchain nodes rises, communication costs increase exponentially. The network delays and communication efficiency degradation caused by high communication costs directly impact the training speed and overall model performance.

High implementation and maintenance costs represent another practical barrier. Establishing a BCFL infrastructure requires significant upfront investment in blockchain nodes, secure servers, storage, and high-speed networking. Additionally, energy consumption associated with blockchain consensus protocols, as well as the operational costs of managing frequent model updates across institutions, may exceed the financial capacity of many health care providers, especially in resource-limited settings. Without clear evidence of cost-benefit balance, hospitals and regulators may be reluctant to adopt BCFL at scale.

Additional difficulties arise from the integration of BCFL within IoMT environments. The heterogeneity of IoMT devices results in substantial disparities in storage capacity, computational power, energy consumption, and communication capabilities. For instance, advanced hospital equipment often features powerful processors, stable power supplies, and ample storage space, whereas wearable medical devices typically operate on low-power batteries, constrained network bandwidth, and limited computational resources. This disparity in device capabilities poses a significant challenge to the deployment of FL models. Moreover, energy constraints and unstable network connections make edge medical devices prone to data transmission failures or system disconnections, ultimately resulting in end-device desynchronization. This issue not only hampers the timeliness of data uploads and model updates but may also prevent the global model from converging efficiently.

Compounding these challenges is the heterogeneity of health care data within IoMT systems. Data generated by various devices exhibit significant diversity, often displaying uneven distributions and violating the non-IID assumption. For instance, hospital A may collect dynamic ECG signals, whereas hospital B primarily acquires static medical images. Such disparities in data distribution exacerbate the complexity of model training. Moreover, variations in medical coding standards across countries and regions (eg, ICD-10 [*International Statistical Classification of Diseases, Tenth Revision*] in the United Kingdom vs ICD-10-CM [*International Classification of Diseases, Tenth Revision, Clinical Modification*] in the United States) contribute to data standard inconsistencies. Consequently, such heterogeneities complicate global model training, analysis, and evaluation, ultimately impairing the model's generalization across diverse clients [125].

Beyond the application-level challenges, BCFL also faces significant technical limitations that must be addressed to realize its full potential in medical settings. Although BCFL integrates the decentralized nature of blockchain with the "data availability without visibility" principle of FL to provide an initial level of privacy protection and enhance overall system security, it does not fully resolve privacy concerns. The sensitivity of patient data and the stringent privacy requirements in the medical field necessitate addressing a series of complex security challenges. The system remains vulnerable to various types of malicious attacks that pose significant risks to patient confidentiality and institutional trust. For instance, background knowledge attacks involve adversaries inferring sensitive information using previously known data and analyzing shared model parameters [3]. In conspiracy attacks, multiple nodes conspire to steal data features from other participants by exchanging local training information [126]. Inference attacks similarly analyze model parameter updates, and attackers can infer private details about patient data [1,43]. These threats highlight the critical need for

XSL•FO
RenderX

advanced privacy-preserving mechanisms within BCFL to ensure the safety and integrity of medical data.

While existing privacy-preserving technologies provide preliminary protections, they are often inadequate when faced with the dual demands of strong privacy and high model utility. Homomorphic encryption allows computation directly on encrypted data, preventing plaintext exposure. However, its computational inefficiency makes it unsuitable for large-scale, complex operations. Differential privacy introduces noise to model parameters to obscure individual data contributions; however, this can significantly compromise model accuracy and system performance if not carefully balanced. Secure Multiparty Computation (SMPC) offers robust data confidentiality through distributed computation, yet its reliance on frequent interaction between parties contradicts the low-interaction protocols typically favored in BCFL for efficient aggregation. These limitations underscore the urgent need for lightweight, efficient, and scalable privacy-preserving solutions specifically tailored to the BCFL context.

Another critical technical issue is the optimization of incentive mechanisms. In traditional blockchain systems, fixed token-based reward structures fail to reflect the true value of each participant's contribution [8]. This misalignment can result in low-quality nodes receiving undeserved rewards, while high-contribution nodes may become demotivated due to insufficient compensation. In BCFL, this issue is further complicated by the heterogeneity of participants, who differ in computational capabilities, data quality, and participation frequency. Resource-constrained nodes, in particular, may lack sufficient incentives to participate, ultimately affecting the quality and diversity of the global model. To address these disparities, incentive mechanisms in BCFL must go beyond simple token rewards and instead adopt dynamic, contribution-aware frameworks that account for the multidimensional nature of participant involvement. A well-designed incentive system can encourage broader and more sustained engagement, improve fairness, and enhance the overall efficiency and robustness of BCFL. Therefore, developing more sophisticated and adaptable incentive mechanisms is a pressing direction for future research.

While the immutability of blockchain and the tamper-resistant nature of smart contracts ensure data integrity and trustworthiness, these characteristics also introduce rigidity, posing challenges in dynamic health care environments [33]. Several scenarios highlight the importance for greater flexibility: when patient data are entered incorrectly or require modification, the immutable blockchain structure cannot accommodate efficiently; then patients may request the deletion or modification of their data to preserve privacy, particularly to comply with data protection regulations; moreover, in public health emergencies, ensuring timely access to accurate health care data is crucial for effective crisis management, necessitating mechanisms for controlled updates within blockchain systems. These scenarios underscore the need for greater flexibility within BCFL systems, where mechanisms must be designed to support controlled edits under predefined conditions—balancing the need for data integrity with the operational demands of evolving health care environments.

What's more, another important consideration in the medical application of BCFL is model interpretability, which directly impacts the reliability of clinical decisions, patient trust, and regulatory compliance. In medical decision-making, the ability to interpret model outcomes is essential to ensuring safety, transparency, and credibility. However, the "black-box" nature of many complex deep learning models limits their interpretability, posing significant challenges for medical applications [106]. The necessity of interpretability can be emphasized from multiple perspectives: For health care professionals, AI model predictions must be interpretable to allow physicians to understand the underlying reasoning and effectively integrate them into diagnosis, treatment planning, and patient monitoring. Moreover, interpretability enables researchers and clinicians to identify and trace the sources of bias or errors in model predictions, facilitating performance optimization and improving diagnostic accuracy and reliability. Simultaneously, in terms of patients, the widespread adoption of AI in medicine inevitably raises concerns regarding privacy and ethics. Enhancing model interpretability can build patient trust in AI-assisted diagnosis and treatment by clarifying the model's reliability and limitations, thereby mitigating concerns over "black-box" decision-making. Moreover, at the regulatory level, numerous countries and regions have mandated transparency and auditability in medical AI systems to ensure that decision-making processes align with ethical and legal standards. Furthermore, medical AI operates within a highly interdisciplinary environment, encompassing physicians, technology developers, data scientists, and other professionals. Interpretability serves as a crucial bridge for communication among experts from diverse disciplines, facilitating the effective implementation of medical AI technologies and ultimately enhancing the quality and accessibility of health care services.

Finally, summarize the limitations of the literature included in the review: first of all, the risk of prejudice is very common. Many available studies are conceptual frameworks, simulations, or small-scale case studies, rather than large-scale clinical implementations, without independent external validation. Common methodological flaws include the selective presentation of favorable performance metrics, limited or lack of adversarial and privacy attack tests, and the absence of long-term or actionable measurements (such as maintenance burden, interoperability failures, or ongoing participation rates). Unavailable code, undisclosed model/configuration details, and dependencies on nonshareable datasets often compromise reproducibility. In conclusion, these factors have created systemic uncertainties, posing the risk of overestimating feasibility and underestimating the actual deployment challenges. Second, the included studies demonstrated substantial inconsistencies in the key dimensions of system design and evaluation. Research in BCFL architecture (fully coupled, flexibly coupled, and loosely coupled), blockchain configuration (private and public), privacy countermeasures (secure aggregation, differential privacy, homomorphic encryption, and SMPC), and data sources (public benchmarks, single-center clinical records, and various IoMT streams) varies greatly. The result measurement criteria have not been standardized: some papers prioritize predictive performance, others emphasize communication or computational overhead,

XSL•FO
RenderX

and still others focus on source or incentive metrics. This heterogeneity has led to different discoveries. Furthermore, there are potential biases because many studies are conducted in a controlled environment with carefully curated datasets, which may not reflect the heterogeneity and noise of real-world medical data. Finally, due to the limited performance index reports and insufficient longitudinal validation, statistical uncertainties (confidence intervals and variability between runs) are rarely reported; large-scale, multi-institutional deployments remain uncommon; and few studies have evaluated long-term stability, scalability under large volumes of clinical data, or regulatory compliance under real-world conditions. These gaps prominently indicate the need for more robust and well-designed research, including prospective clinical trials, to verify the effectiveness, safety, and interoperability of the BCFL framework in actual health care settings.

Future Prospects

The main obstacles to data sharing and collaboration in the health care system lie in the lack of interoperability and standardization, as well as the absence of standardized benchmark datasets in medical applications in the context of FL and blockchain applications. Therefore, future research must integrate the joint efforts of the government, regulatory authorities, and industry leaders to establish unified technical standards and policy frameworks and give priority to the development of standardized BCFL benchmark datasets. These standards should encompass data formats, transmission protocols, privacy safeguards, and technology implementation guidelines to facilitate seamless data integration and collaboration across diverse health care entities. Standardized datasets should reflect the realistic heterogeneity of medical imaging, EHRs, and multimodal data streams. By providing a common reference point, these benchmarks will enable fair and transparent algorithm comparisons, promote the reproducibility of results, and accelerate the transformation of BCFL innovations into clinical validation tools. In addition, benchmark datasets can be stratified based on disease types, patterns, and clinical tasks (eg, diagnosis, prognosis, and treatment response prediction), thereby allowing for more fine-grained evaluations of system performance in different health care settings. Moreover, regulatory frameworks should prioritize patient privacy and data security, delineate the rights and responsibilities of stakeholders, and foster the compliant adoption of BCFL technology. Comprehensive policy support and technical guidance will be instrumental in mitigating data silos and enhancing the efficiency of collaborative model training.

Scalability and communication efficiency remain two of the most critical technical challenges facing BCFL, particularly in large-scale and high-frequency health care environments. To address the scalability limitations, future research should focus on several key technical optimizations. The first is off-chain computing and side-chain technology [43]. Off-chain computing enables complex training tasks to be executed off-chain, reducing the computational burden on the main blockchain. Meanwhile, side chains can independently handle task-specific transactions, alleviating congestion on the main chain. Layer 2 protocols, such as stateful channels and plasma technology, offer a viable solution by enabling faster transaction processing while preserving blockchain security [43]. These technologies enhance scalability by minimizing main-chain data storage requirements. Another crucial research direction is cross-chain technologies [42]. This approach not only decentralizes workloads and mitigates single-chain bottlenecks but also enhances system-wide parallel processing capabilities. Moreover, cross-chain FL is particularly well-suited for cross-regional and cross-organizational health care collaborations, further strengthening resource management and system security.

Regarding communication cost and efficiency, the following directions are worth exploring in depth. One promising direction is gradient compression. Using gradient compression techniques helps reduce communication overhead. For instance, Konecny et al [127] proposed that structured updates and sketch updates can significantly lower communication costs. However, this approach also introduces potential challenges, such as the loss of relevant information during compression, which may have an impact on the performance of the global model. Therefore, future research should focus on achieving an optimal balance between gradient compression and global model accuracy [128]. Lightweight consensus protocols are also another crucial research direction. Another key technological direction is the Digital Twin [14], which minimizes the need for long-distance data transmission by enabling the generation of virtual models directly on miner nodes. This approach significantly decreases communication latency and costs, making it particularly well-suited for resource-constrained health care environments.

In the field of IoMT, to address the challenges posed by heterogeneous storage, computing, and communication capabilities of medical devices and sensors, future research should prioritize optimizing resource usage. First, lightweight ML models and algorithms, such as model compression and pruning techniques [128], can be developed to alleviate the computational burden on local devices. This not only enhances device operational efficiency but also significantly reduces communication overhead. At the task allocation level, device grouping and hierarchical architectures can be leveraged to allocate computation and data aggregation tasks to devices or intermediate nodes with higher computational capacity, thereby forming a resource-optimized collaborative network.

Another critical aspect is enhancing system robustness in the face of equipment instability. To mitigate this issue, future research should focus on optimizing both failure recovery mechanisms and participant selection strategies. On the one hand, an adaptive connection protocol can be designed to enable devices to automatically rejoin the training process after a connection interruption, ensuring that the global model's convergence remains unaffected. On the other hand, an optimization strategy based on device availability and performance can be implemented to prioritize the selection of more stable devices for training. Moreover, incorporating a flexible time window would allow devices to complete tasks within a predefined period, thereby enhancing the overall system's fault tolerance and training efficiency.

Given the highly diverse and statistically heterogeneous nature of medical data in IoMT environments, achieving robust model

XSL•FO
RenderX

generalization is another priority. The prevalence of non-IID data across institutions often results in significant discrepancies between global and local models, undermining convergence and performance. Consequently, Wu and Wang [125] have proposed an optimal aggregation algorithm that dynamically adjusts the selection probability of each trainer based on the algorithm's output. However, trainer selection based on model preferences can severely compromise the generalization capability of the global model. Future research should focus on developing a highly accurate BCFL system with enhanced model generalization. One promising direction is the development of automated data normalization tools, which are capable of recognizing various dataset formats and characteristics, performing automatic transformations to ensure that the data can be directly used by model training [129].

The technical limitations faced by BCFL systems need to be well addressed in the future, and privacy protection is one of the fundamental challenges. Future research should further weigh the relationship between privacy and utility, selecting appropriate privacy-preserving techniques based on system priorities. For instance, in certain scenarios, prioritizing privacy may necessitate the adoption of computationally intensive yet more secure techniques, whereas high-performance requirements may favor more lightweight solutions. Additionally, more advanced privacy-preserving techniques can be explored. For example, zero-knowledge proof (ZKP) is a promising cryptographic method that enables a prover to demonstrate the validity of a statement to a verifier without disclosing any additional information. This ensures that the data owner can validate the accuracy of an update while keeping the original data confidential. This approach not only minimizes the trust overhead among participants but also alleviates the verification burden on clients. Future research should investigate how ZKP can be seamlessly integrated into the BCFL architecture to optimize the balance between performance and privacy, paving the way for a more efficient and secure system with enhanced privacy measures.

Another pressing issue in the development of BCFL systems is the design of fair and dynamic incentive mechanisms to ensure equitable resource allocation among contributors. On the one hand, the Shapley value-based contribution quantification method can be used to assess each participant's impact on global model performance improvement, thereby enabling a fairer incentive distribution. On the other hand, since variations in data quality directly impact model training effectiveness, data quality-driven incentives can be introduced so that participants contributing high-quality or more representative data receive greater rewards. Moreover, integrating penalty mechanisms is also another crucial research direction [85]. For instance, Cui et al [130] proposed withdrawing tokens when a trainer's behavior is identified as malicious. Similarly, Weng et al [79] suggested requiring trainers to predeposit tokens, which are forfeited upon detection of malicious activity, but the fairness of this deposit mechanism remains uncertain. Therefore, how to reasonably set the punishment rules under the premise of ensuring fairness still needs further research. These optimization strategies will contribute to building a more effective and equitable BCFL ecosystem, fostering its sustainable development in medical data sharing.

It is undeniable that token-based incentive mechanisms have been widely discussed as a promising approach to encouraging participation in the BCFL network. Although such mechanisms can enhance participation and promote fair resource allocation, they also introduce complex ethical issues. Therefore, in future research, ethical supervision should be embedded in token-based system design, and new solutions should be continuously explored.

Editable mechanisms are expected to be implemented in the near future. Future research could explore how to introduce a moderate degree of editability while ensuring data integrity and security. For instance, the modified Chameleon hash function (also known as the trapdoor hash function) enables controlled modification of blockchain data. When the trapdoor information is available, hash collisions can be efficiently identified, allowing input modifications without altering the hash output. This mechanism facilitates the correction of inaccurate or incomplete data while preserving the structural integrity of the blockchain [8].

In parallel, the interpretability of AI models remains a central concern for the deployment of BCFL in clinical practice. Trust, transparency, and accountability are all closely tied to how well clinicians and patients can understand the rationale behind AI-generated predictions. Existing mainstream interpretability tools (eg, SHAP and LIME) have limitations in the medical field, including difficulties in handling complex distributed data environments and the inability to provide clinically relevant interpretations. Future research should focus on enhancing existing tools and developing interpretability methods adapted to BCFL, such as global-local model comparison techniques, to provide more intuitive and trustworthy interpretations.

In addition to technological and application-level innovations, the future development of BCFL also needs to confront the issue that research mainly remains at the simulation, prototype, and preclinical stages, and these studies have not yet formed a direct correspondence with real medical endpoint indicators. Future work should start from real scenarios: (1) conduct prospective validations in multi-institutional, heterogeneous data and complete clinical workflows to evaluate their actual impact on diagnostic efficiency, therapeutic effect improvement, and resource usage; (2) maintain a balance between security, latency, and maintainability in system design to avoid unacceptable computing and communication costs caused by encryption and on-chain operations; and (3) at the governance level, strengthen the connection with the medical system, regulations and ethical frameworks, and clarify data ownership, responsibility attribution, and auditing mechanisms. Only when technical performance is validated in real clinical settings, system design remains practically deployable, and governance mechanisms are clearly defined, BCFL can progress from laboratory prototypes to clinically reliable infrastructure.

While the above guidance outlines strategic directions for policymakers, clinicians, and implementers, these recommendations remain high-level. To move from broad vision to actionable progress, there is a pressing need for specific

XSL•FO
RenderX

evaluative structures that can translate theories into measurable outcomes. Therefore, establishing a good evaluation framework is indispensable, which can provide a foundation for future research, allowing for comparison, verification, and ultimately integration into clinical workflows.

A unified, multidimensional evaluation framework is central to the advancement of BCFL in medicine. From a technical perspective, standardized metrics such as model accuracy, convergence speed, robustness against adversarial attacks, communication latency, and scalability across heterogeneous institutional datasets are indispensable. These indicators establish the baseline scientific validity of BCFL. Equally important afterwards are the clinical assessment criteria, which should reflect the sensitivity and specificity of the diagnosis, its universality in different patient populations and multicenter environments, reduce algorithmic bias, and bring about a tangible improvement in patient prognosis. By embedding clinical endpoints into the assessment, this framework can ensure that technological progress is in line with the real-world health care needs. Finally, there are operational indicators, which should cover interoperability with existing medical data systems, cost-effectiveness, and the long-term sustainability of deployment. By integrating these three dimensions into a unified evaluation system, researchers and clinical workers can establish benchmarks that can both horizontally compare different studies and vertically track research progress. This evaluation framework not only enhances the scientific rigor of BCFL research but also provides an evidence-based basis for decision-making sharing, thereby accelerating the transformation from experimental prototypes to clinical applications.

Conclusions

This review systematically compiles the research progress of blockchain and FL in the medical field. First, we introduce the theoretical foundations and core technical features of both technologies, analyzing how blockchain enhances the security, privacy protection, and decentralization characteristics of FL, while FL improves the computational efficiency and scalability of blockchain. In addition, we describe the three frameworks and workflows of BCFL. Next, we summarize the research progress in BCFL applications across cross-institutional health care data sharing, the IoMT, public health monitoring, and telemedicine, highlighting its practical value in privacy protection and data collaboration. Moreover, we discuss key challenges in BCFL, including computational efficiency, scalability, data privacy, and incentive mechanism design, while proposing potential solutions and future research directions. The significance of this review lies in providing a comprehensive overview of how BCFL could reshape medical data collaboration and security paradigms, thereby offering valuable insights for researchers and practitioners exploring this interdisciplinary field. However, it is important to acknowledge certain limitations of this review. This review primarily focuses on the theoretical principles and current applications of BCFL, with relatively limited exploration of specific implementation details and performance evaluations. Additionally, most existing BCFL studies rely on simulation experiments or public datasets, lacking validation with large-scale real-world medical data, which affects assessments of its practical feasibility. As such, although BCFL shows promise in supporting future intelligent diagnostics, precision medicine, and collaborative health care systems, claims about clinical readiness remain premature. Future work should focus on prospective, large-scale validation studies, interdisciplinary collaboration with health care providers, and the development of standardized evaluation protocols to ensure that BCFL solutions are clinically safe, ethically sound, and operationally feasible.

Acknowledgments

We sincerely thank the journal editor Javad Sarvestan, PhD, and the peer reviewers for their valuable feedback during this study, which has greatly enhanced our research. We used ChatGPT (OpenAI) to check and revise the vocabulary and grammar of this manuscript and polished the language in the introduction.

Funding

This research was funded by the National Natural Science Foundation of China (project no. 81974355 and 82172524), the Key Research and Development Program of Hubei Province (project no. 2021BEA161), the National Innovation Platform Cultivation Program (project no. 2020021105012440), and the Education Reform Project of Ningxia Medical University (funded by NYJY2025057).

Data Availability

All data generated or analyzed during this study are included in Multimedia Appendices 1 and 2.

Authors' Contributions

Conceptualization: XW (lead), YX (equal), XC (equal)
Data curation: XW (lead), YX (supporting), XC (supporting), JY (supporting)
Formal analysis: XW
Funding acquisition: ZY
Investigation: XW (lead), YX (supporting), XC (supporting), JY (supporting), RL (supporting)
Methodology: XW (lead), YX (supporting), XC (supporting), JY (supporting)

XSL•FO
RenderX

Project administration: XW (lead), YX (supporting), XC (supporting), JY (supporting)
Resources: XW (lead), YX (equal), XC (supporting)
Supervision: ZY (lead), YX (supporting)
Validation: YX (lead), ZY (supporting)
Visualization: XW (lead), YX (equal), XC (supporting)
Writing – original draft: XW (lead), YX (equal), XC (supporting), JY (supporting)
Writing – review & editing: XW (lead), YX (supporting), XC (supporting), ZY (supporting), JY (supporting), RL (supporting), WG (supporting), ZY (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Databases search strategy.
[DOCX File, 22 KB - jmir_v28i1e79052_app1.docx]

Multimedia Appendix 2
Scoring statistics and agreement data.
[XLSX File, 13 KB - jmir_v28i1e79052_app2.xlsx]

Multimedia Appendix 3
Additional tables.
[DOCX File, 14 KB - jmir_v28i1e79052_app3.docx]

Checklist 1
PRISMA checklist.
[DOCX File, 31 KB - jmir_v28i1e79052_app4.docx]

Checklist 2
Kitchenham SLR Quality Checklist.
[DOCX File, 17 KB - jmir_v28i1e79052_app5.docx]

References

1. Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning. In: 2019 IEEE Symposium on Security and Privacy: IEEE; 2019:739-753. [doi: 10.1109/SP.2019.00065]
2. Qu Y, Gao L, Luan TH, et al. Decentralized privacy using blockchain-enabled federated learning in fog computing. IEEE Internet Things J 2020 Jun;7(6):5171-5183. [doi: 10.1109/JIOT.2020.2977383]
3. Wang Z, Song M, Zhang Z, Song Y, Wang Q, Qi H. Beyond inferring class representatives: user-level privacy leakage from federated learning. In: IEEE INFOCOM 2019 – IEEE Conference on Computer Communications: IEEE; 2019:2512-2520. [doi: 10.1109/INFOCOM.2019.8737416]
4. Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning. ACM Trans Intell Syst Technol 2019 Mar 31;10(2):1-19. [doi: 10.1145/3298981]
5. Hiwale M, Walambe R, Potdar V, Kotecha K. A systematic review of privacy-preserving methods deployed with blockchain and federated learning for the telemedicine. Healthc Anal 2023 Nov;3:100192. [doi: 10.1016/j.health.2023.100192]
6. Boland MR, Parhi P, Li L, et al. Uncovering exposures responsible for birth season – disease effects: a global study. J Am Med Inform Assoc 2018 Mar 1;25(3):275-288. [doi: 10.1093/jamia/ocx105]
7. Freire-Garabal y Núñez M. General vision of the Regulation (EU) 2016/679 of the European Parliament and of the Council of April 27, 2016. Al-Khalifa Business School 2020 Jun 21. [doi: 10.21428/18D9181C.D22AA10A]
8. Qu Y, Uddin MP, Gan C, Xiang Y, Gao L, Yearwood J. Blockchain-enabled federated learning: a survey. ACM Comput Surv 2023 Apr 30;55(4):1-35. [doi: 10.1145/3524104]
9. Uddin MA, Stranieri A, Gondal I, Balasubramanian V. A survey on the adoption of blockchain in IoT: challenges and solutions. Blockchain Res Appl 2021;2(2):100006. [doi: 10.1016/j.bcra.2021.100006]
10. Alsamhi SH, Myrzashova R, Hawbani A, et al. Federated learning meets blockchain in decentralized data sharing: healthcare use case. IEEE Internet Things J 2024 Jun 1;11(11):19602-19615. [doi: 10.1109/JIOT.2024.3367249]
11. Topaloglu MY, Morrell EM, Rajendran S, Topaloglu U. In the pursuit of privacy: the promises and predicaments of federated learning in healthcare. Front Artif Intell 2021 Oct 6;4:746497. [doi: 10.3389/frai.2021.746497]

XSL•FO
RenderX

12. Li J, Meng Y, Ma L, et al. A federated learning based privacy-preserving smart healthcare system. IEEE Trans Ind Inf 2022 Mar;18(3):2021-2031. [doi: 10.1109/TII.2021.3098010]
13. Zyskind G, Nathan O, Pentland A. Decentralizing privacy: using blockchain to protect personal data. In: 2015 IEEE Security and Privacy Workshops: IEEE; 2015:180-184. [doi: 10.1109/SPW.2015.27]
14. Liu K, Yan Z, Liang X, Kantola R, Hu C. A survey on blockchain-enabled federated learning and its prospects with digital twin. Digit Commun Netw 2024;10(2):248-264. [doi: 10.1016/j.dcan.2022.08.001]
15. Dinh TTA, Liu R, Zhang M, Chen G, Ooi BC, Wang J. Untangling blockchain: a data processing view of blockchain systems. IEEE Trans Knowl Data Eng 2018 Jul 1;30(7):1366-1385. [doi: 10.1109/TKDE.2017.2781227]
16. Gervais A, Karame GO, Wüst K, Glykantzis V, Ritzdorf H, Capkun S. On the security and performance of proof of work blockchains. In: CCS '16: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security: IEEE; 2016:3-16. [doi: 10.1145/2976749.2978341]
17. Kim H, Park J, Bennis M, Kim SL. Blockchained on-device federated learning. IEEE Commun Lett 2020 Jun;24(6):1279-1283. [doi: 10.1109/LCOMM.2019.2921755]
18. Kitchenham B, Charters S. Guidelines for performing systematic literature reviews in software engineering. : Keele University and Durham University Joint Report; 2007 Jul EBSE 2007-001 URL: https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf [accessed 2026-01-29]
19. McMahan HB, Moore E, Ramage D, Hampson S, Agüera y Arcas B. Communication-efficient learning of deep networks from decentralized data. arXiv. Preprint posted online on Jan 26, 2023. [doi: 10.48550/arXiv.1602.05629]
20. Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. J Healthc Inform Res 2021;5(1):1-19. [doi: 10.1007/s41666-020-00082-4] [Medline: 33204939]
21. Chen Y, Qin X, Wang J, Yu C, Gao W. FedHealth: a federated transfer learning framework for wearable healthcare. IEEE Intell Syst 2020 Jul 1;35(4):83-93. [doi: 10.1109/MIS.2020.2988604]
22. Long G, Tan Y, Jiang J, Zhang C. Federated learning for open banking. In: Yang Q, Fan L, Yu H, editors. Federated Learning: Privacy and Incentive: Springer; 2020:240-254. [doi: 10.1007/978-3-030-63076-8_17]
23. Zheng Z, Zhou Y, Sun Y, Wang Z, Liu B, Li K. Applications of federated learning in smart cities: recent advances, taxonomy, and open challenges. Conn Sci 2022 Dec 31;34(1):1-28. [doi: 10.1080/09540091.2021.1936455]
24. Zhang C, Xie Y, Bai H, Yu B, Li W, Gao Y. A survey on federated learning. Knowl Based Syst 2021 Mar;216:106775. [doi: 10.1016/j.knosys.2021.106775]
25. Nishio T, Yonetani R. Client selection for federated learning with heterogeneous resources in mobile edge. In: 2019 IEEE International Conference on Communications: IEEE; 2019:1-7. [doi: 10.1109/ICC.2019.8761315]
26. Nguyen DC, Pham QV, Pathirana PN, et al. Federated learning for smart healthcare: a survey. ACM Comput Surv 2023 Mar 31;55(3):1-37. [doi: 10.1145/3501296]
27. Heyndrickx W, Mervin L, Morawietz T, et al. MELLODDY: cross-pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information. J Chem Inf Model 2024 Apr 8;64(7):2331-2344. [doi: 10.1021/acs.jcim.3c00799] [Medline: 37642660]
28. Hao M, Li H, Xu G, Liu Z, Chen Z. Privacy-aware and resource-saving collaborative learning for healthcare in cloud computing. In: 2020 IEEE International Conference on Communications: IEEE; 2020:1-6. [doi: 10.1109/ICC40277.2020.9148979]
29. Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis I, Shi W. Federated learning of predictive models from federated electronic health records. Int J Med Inform 2018 Apr;112:59-67. [doi: 10.1016/j.ijmedinf.2018.01.007]
30. Li W, Milletarì F, Xu D, et al. Privacy-preserving federated brain tumour segmentation. In: Suk HI, Liu M, Yan P, Lian C, editors. Machine Learning in Medical Imaging: Springer; 2019:133-141. [doi: 10.1007/978-3-030-32692-0_16]
31. Zhang W, Zhou T, Lu Q, et al. Dynamic-fusion-based federated learning for COVID-19 detection. IEEE Internet Things J 2021 Nov 1;8(21):15884-15891. [doi: 10.1109/JIOT.2021.3056185] [Medline: 35663640]
32. Ulhaq A, Burmeister O. COVID-19 imaging data privacy by federated learning design: a theoretical framework. arXiv. Preprint posted online on Oct 13, 2020. [doi: 10.48550/arXiv.2010.06177]
33. Qammar A, Karim A, Ning H, Ding J. Securing federated learning with blockchain: a systematic literature review. Artif Intell Rev 2023;56(5):3951-3985. [doi: 10.1007/s10462-022-10271-9] [Medline: 36160367]
34. Feng L, Zhao Y, Guo S, Qiu X, Li W, Yu P. BAFL: a blockchain-based asynchronous federated learning framework. IEEE Trans Comput 2022 May 1;71(5):1092-1103. [doi: 10.1109/TC.2021.3072033]
35. Kim YJ, Hong CS. Blockchain-based node-aware dynamic weighting methods for improving federated learning performance. In: 2019 20th Asia-Pacific Network Operations and Management Symposium: IEEE; 2019:1-4. [doi: 10.23919/APNOMS.2019.8893114]
36. Li Y, Chen C, Liu N, Huang H, Zheng Z, Yan Q. A blockchain-based decentralized federated learning framework with committee consensus. IEEE Netw 2021 Jan;35(1):234-241. [doi: 10.1109/MNET.011.2000263]
37. Bhowmick A, Duchi J, Freudiger J, Kapoor G, Rogers R. Protection against reconstruction and its applications in private federated learning. arXiv. Preprint posted online on Jun 3, 2019. [doi: 10.48550/arXiv.1812.00984]
38. Zhang J, Chen J, Wu D, Chen B, Yu S. Poisoning attack in federated learning using generative adversarial nets. In: 2019 18th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/13th IEEE

XSL•FO
RenderX

International Conference on Big Data Science and Engineering (TrustCom/BigDataSE): IEEE; 2019:374-380. [doi: 10.1109/TrustCom/BigDataSE.2019.00057]
39. Kang J, Xiong Z, Niyato D, Yu H, Liang YC, Kim DI. Incentive design for efficient federated learning in mobile networks: a contract theory approach. In: 2019 IEEE VTS Asia Pacific Wireless Communications Symposium (APWCS): IEEE; 2019:1-5. [doi: 10.1109/VTS-APWCS.2019.8851649]
40. Passerat-Palmbach J, Farnan T, Miller R, Gross MS, Flannery HL, Gleim B. A blockchain-orchestrated federated learning architecture for healthcare consortia. arXiv. Preprint posted online on Oct 12, 2019. [doi: 10.48550/arXiv.1910.12603]
41. Majeed U, Hong CS. FLchain: federated learning via MEC-enabled blockchain network. In: 2019 20th Asia-Pacific Network Operations and Management Symposium (APNOMS): IEEE; 2019. [doi: 10.23919/APNOMS.2019.8892848]
42. Sameera KM, Nicolazzo S, Arazzi M, et al. Privacy-preserving in blockchain-based federated learning systems. Comput Commun 2024 Jun;222:38-67. [doi: 10.1016/j.comcom.2024.04.024]
43. Liu J, Chen C, Li Y, et al. Enhancing trust and privacy in distributed networks. arXiv. Preprint posted online on Mar 28, 2024. [doi: 10.48550/arXiv.2403.19178]
44. Shahsavari Y, Dambri OA, Baseri Y, Hafid AS, Makrakis D. Integration of federated learning and blockchain in healthcare: a tutorial. arXiv. Preprint posted online on Apr 15, 2024. [doi: 10.48550/arXiv.2404.10092]
45. Nakamoto S. Bitcoin: a peer-to-peer electronic cash system. SSRN. 2008 Aug 21. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3440802 [accessed 2026-02-12]
46. Zheng Z, Xie S, Dai H, Chen X, Wang H. An overview of blockchain technology: architecture, consensus, and future trends. In: 2017 IEEE International Congress on Big Data (BigData Congress): IEEE; 2017:557-564. [doi: 10.1109/BigDataCongress.2017.85]
47. Nofer M, Gomber P, Hinz O, Schiereck D. Blockchain. Bus Inf Syst Eng 2017 Jun;59(3):183-187. [doi: 10.1007/s12599-017-0467-3]
48. Garay J, Kiayias A. SoK: a consensus taxonomy in the blockchain era. In: Jarecki S, editor. Top Cryptol – CT-RSA 2020: Springer; 2020:284-318. [doi: 10.1007/978-3-030-40186-3_13]
49. Wang W, Hoang DT, Hu P, et al. A survey on consensus mechanisms and mining strategy management in blockchain networks. IEEE Access 2019;7:22328-22370. [doi: 10.1109/ACCESS.2019.2896108]
50. Sukhwani H, Martinez JM, Chang X, Trivedi KS, Rindos A. Performance modeling of PBFT consensus process for permissioned blockchain network (Hyperledger Fabric). In: 2017 IEEE 36th Symposium on Reliable Distributed Systems (SRDS): IEEE; 2017. [doi: 10.1109/SRDS.2017.36]
51. Nguyen CT, Hoang DT, Nguyen DN, Niyato D, Nguyen HT, Dutkiewicz E. Proof-of-stake consensus mechanisms for future blockchain networks: fundamentals, applications and opportunities. IEEE Access 2019;7:85727-85745. [doi: 10.1109/ACCESS.2019.2925010] [Medline: 33777591]
52. Huang H, Li K, Chen X. Blockchain‐based fair three‐party contract signing protocol for fog computing. Concurr Comput 2019 Nov 25;31(22):e4469. [doi: 10.1002/cpe.4469]
53. Wu H, Jiang S, Cao J. High-efficiency blockchain-based supply chain traceability. IEEE Trans Intell Transport Syst 2023;24(4):3748-3758. [doi: 10.1109/TITS.2022.3205445]
54. Dos Santos S, Singh J, Thulasiram RK, Kamali S, Sirico L, Loud L. A new era of blockchain-powered decentralized finance (DeFi) – a review. In: 2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC): IEEE; 2022:1286-1292. [doi: 10.1109/COMPSAC54236.2022.00203]
55. Kuo TT, Kim HE, Ohno-Machado L. Blockchain distributed ledger technologies for biomedical and health care applications. J Am Med Inform Assoc 2017 Nov 1;24(6):1211-1220. [doi: 10.1093/jamia/ocx068]
56. Mohanta BK, Jena D, Panda SS, Sobhanayak S. Blockchain technology: a survey on applications and security privacy challenges. Internet Things 2019 Dec;8:100107. [doi: 10.1016/j.iot.2019.100107]
57. Huynh TT, Nguyen TD, Tan H. A survey on security and privacy issues of blockchain technology. In: 2019 International Conference on System Science and Engineering (ICSSE): IEEE; 2019:362-367. [doi: 10.1109/ICSSE.2019.8823094]
58. Dinh TN, Thai MT. AI and blockchain: a disruptive integration. Computer (Long Beach Calif) 2018 Sep;51(9):48-53. [doi: 10.1109/MC.2018.3620971]
59. Salah K, Rehman MHU, Nizamuddin N, Al-Fuqaha A. Blockchain for AI: review and open research challenges. IEEE Access 2019;7:10127-10149. [doi: 10.1109/ACCESS.2018.2890507]
60. Wang Q, Zhu X, Ni Y, Gu L, Zhu H. Blockchain for the IoT and industrial IoT: a review. Internet Things 2020 Jun;10:100081. [doi: 10.1016/j.iot.2019.100081]
61. Agbo CC, Mahmoud QH, Eklund JM. Blockchain technology in healthcare: a systematic review. Healthcare (Basel) 2019 Apr 4;7(2):56. [doi: 10.3390/healthcare7020056] [Medline: 30987333]
62. Kang M, Park E, Cho BH, Lee KS. Recent patient health monitoring platforms incorporating Internet of Things-enabled smart devices. Int Neurourol J 2018 Jul;22(Suppl 2):S76-S82. [doi: 10.5213/inj.1836144.072] [Medline: 30068069]
63. Yang G, Li C. A design of blockchain-based architecture for the security of electronic health record (EHR) systems. In: 2018 IEEE International Conference on Cloud Computing Technology and Science (CloudCom): IEEE; 2018:261-265. [doi: 10.1109/CloudCom2018.2018.00058]

XSL•FO
RenderX

64. Yuan WX, Yan B, Li W, Hao LY, Yang HM. Blockchain-based medical health record access control scheme with efficient protection mechanism and patient control. Multimed Tools Appl 2023 May;82(11):16279-16300. [doi: 10.1007/s11042-022-14023-3]
65. Zhan W, Chen CL, Weng W, Tsaur WJ, Lim ZY, Deng YY. Incentive EMR sharing system based on consortium blockchain and IPFS. Healthcare (Basel) 2022 Sep 22;10(10):1840. [doi: 10.3390/healthcare10101840]
66. Amanat A, Rizwan M, Maple C, Zikria YB, Almadhor AS, Kim SW. Blockchain and cloud computing-based secure electronic healthcare records storage and sharing. Front Public Health 2022;10:938707. [doi: 10.3389/fpubh.2022.938707] [Medline: 35928494]
67. Kordestani H, Barkaoui K, Zahran W. HapiChain: a blockchain-based framework for patient-centric telemedicine. In: 2020 IEEE 8th International Conference on Serious Games and Applications for Health (SeGAH): IEEE; 2020:1-6. [doi: 10.1109/SeGAH49190.2020.9201726]
68. Uddin M. Blockchain Medledger: hyperledger fabric enabled drug traceability system for counterfeit drugs in pharmaceutical industry. Int J Pharm 2021 Mar;597:120235. [doi: 10.1016/j.ijpharm.2021.120235]
69. Huang Y, Wu J, Long C. Drugledger: a practical blockchain system for drug traceability and regulation. In: 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData): IEEE; 2018:1137-1144. [doi: 10.1109/Cybermatics_2018.2018.00206]
70. Qu Y, Pokhrel SR, Garg S, Gao L, Xiang Y. A blockchained federated learning framework for cognitive computing in Industry 4.0 networks. IEEE Trans Ind Inf 2021 Apr;17(4):2964-2973. [doi: 10.1109/TII.2020.3007817]
71. El Rifai O, Biotteau M, Megdiche I, Ravat F, Teste O. Blockchain-based federated learning in medicine. In: Michalowski M, Moskovitch R, editors. Artificial Intelligence in Medicine: Springer; 2020:214-224. [doi: 10.1007/978-3-030-59137-3_20]
72. Aich S, Sinai NK, Kumar S, et al. Protecting personal healthcare record using blockchain & federated learning technologies. In: 2022 24th International Conference on Advanced Communication Technology (ICACT): IEEE; 2022:109-112. [doi: 10.23919/ICACT53585.2022.9728772]
73. Xu G, Zhou Z, Dong J, Zhang L, Song X. A blockchain-based federated learning scheme for data sharing in Industrial Internet of Things. IEEE Internet Things J 2023 Dec 15;10(24):21467-21478. [doi: 10.1109/JIOT.2023.3298196]
74. Lu Y, Huang X, Dai Y, Maharjan S, Zhang Y. Blockchain and federated learning for privacy-preserved data sharing in Industrial IoT. IEEE Trans Ind Inf 2020 Jun;16(6):4177-4186. [doi: 10.1109/TII.2019.2942190]
75. Lu Y, Huang X, Zhang K, Maharjan S, Zhang Y. Blockchain and federated learning for 5G Beyond. IEEE Netw 2021 Jan;35(1):219-225. [doi: 10.1109/MNET.011.1900598]
76. Fan S, Zhang H, Zeng Y, Cai W. Hybrid blockchain-based resource trading system for federated learning in edge computing. IEEE Internet Things J 2021;8(4):2252-2264. [doi: 10.1109/JIOT.2020.3028101]
77. Kang J, Xiong Z, Niyato D, Xie S, Zhang J. Incentive mechanism for reliable federated learning: a joint optimization approach to combining reputation and contract theory. IEEE Internet Things J 2019 Dec;6(6):10700-10714. [doi: 10.1109/JIOT.2019.2940820]
78. Moudoud H, Cherkaoui S, Khoukhi L. Towards a secure and reliable federated learning using blockchain. In: 2021 IEEE Global Communications Conference (GLOBECOM): IEEE; 2021:1-6. [doi: 10.1109/GLOBECOM46510.2021.9685388]
79. Weng J, Weng J, Zhang J, Li M, Zhang Y, Luo W. DeepChain: auditable and privacy-preserving deep learning with blockchain-based incentive. IEEE Trans Dependable Secure Comput 2021;18(5):2438-2455. [doi: 10.1109/TDSC.2019.2952332]
80. Zhao Y, Zhao J, Jiang L, et al. Privacy-preserving blockchain-based federated learning for IoT Devices. IEEE Internet Things J 2021 Feb;8(3):1817-1829. [doi: 10.1109/JIOT.2020.3017377] [Medline: 35783535]
81. Prokop K, Połap D, Srivastava G, Lin JCW. Blockchain-based federated learning with checksums to increase security in Internet of Things solutions. J Ambient Intell Human Comput 2023 May;14(5):4685-4694. [doi: 10.1007/s12652-022-04372-0]
82. Kalapaaking AP, Khalil I, Yi X. Blockchain-based federated learning with SMPC model verification against poisoning attack for healthcare systems. IEEE Trans Emerg Topics Comput 2024 Jan;12(1):269-280. [doi: 10.1109/TETC.2023.3268186]
83. Shayan M, Fung C, Yoon CJM, Beschastnikh I. Biscotti: a blockchain system for private and secure federated learning. IEEE Trans Parallel Distrib Syst 2021 Jul 1;32(7):1513-1525. [doi: 10.1109/TPDS.2020.3044223]
84. Chen H, Asif SA, Park J, Shen CC, Bennis M. Robust blockchained federated learning with model validation and proof-of-stake inspired consensus. arXiv. Preprint posted online on Jan 9, 2021. [doi: 10.48550/arXiv.2101.03300]
85. Wang Z, Hu Q, Xu M, Zhuang Y, Wang Y, Cheng X. A systematic survey of blockchained federated learning. arXiv. Preprint posted online on Jun 1, 2024. [doi: 10.48550/arXiv.2110.02182]
86. Behera MR, Upadhyay S, Shetty S. Federated learning using smart contracts on blockchains. arXiv. Preprint posted online on Mar 26, 2022. [doi: 10.48550/arXiv.2107.10243]
87. Awan S, Li F, Luo B, Liu M. Poster: a reliable and accountable privacy-preserving federated learning framework using the blockchain. In: CCS '19: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security: Association for Computing Machinery; 2019:2561-2563. [doi: 10.1145/3319535.3363256]

XSL•FO
RenderX

88. Shoker A. Brief announcement: sustainable blockchains through proof of exercise. In: PODC '18: Proceedings of the 2018 ACM Symposium on Principles of Distributed Computing: Association for Computing Machinery; 2018:269-271. [doi: 10.1145/3212734.3212781]
89. Qu X, Wang S, Hu Q, Cheng X. Proof of federated learning: a novel energy-recycling consensus algorithm. IEEE Trans Parallel Distrib Syst 2021 Aug 1;32(8):2074-2085. [doi: 10.1109/TPDS.2021.3056773]
90. Wang X, Garg S, Lin H, Piran M, Hu J, Hossain MS. Enabling secure authentication in Industrial IoT with transfer learning empowered blockchain. IEEE Trans Ind Inf 2021 Nov;17(11):7725-7733. [doi: 10.1109/TII.2021.3049405]
91. Zhu J, Cao J, Saxena D, Jiang S, Ferradi H. Blockchain-empowered federated learning: challenges, solutions, and future directions. ACM Comput Surv 2023 Nov 30;55(11):1-31. [doi: 10.1145/3570953]
92. Bao X, Su C, Xiong Y, Huang W, Hu Y. FLChain: a blockchain for auditable federated learning with trust and incentive. In: 2019 5th International Conference on Big Data Computing and Communications (BIGCOM): IEEE; 2019:151-159. [doi: 10.1109/BIGCOM.2019.00030]
93. Lu Y, Huang X, Zhang K, Maharjan S, Zhang Y. Blockchain empowered asynchronous federated learning for secure data sharing in Internet of Vehicles. IEEE Trans Veh Technol 2020 Apr;69(4):4298-4311. [doi: 10.1109/TVT.2020.2973651]
94. Habib ur Rehman M, Salah K, Damiani E, Svetinovic D. Towards blockchain-based reputation-aware federated learning. In: IEEE INFOCOM 2020 – IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS): IEEE; 2020:183-188. [doi: 10.1109/INFOCOMWKSHPS50562.2020.9163027]
95. Qi J, Lin F, Chen Z, Tang C, Jia R, Li M. High-quality model aggregation for blockchain-based federated learning via reputation-motivated task participation. IEEE Internet Things J 2022 Oct 1;9(19):18378-18391. [doi: 10.1109/JIOT.2022.3160425]
96. Lamken D, Wagner T, Hoiss T, et al. Design patterns and framework for blockchain integration in supply chains. In: 2021 IEEE International Conference on Blockchain and Cryptocurrency (ICBC): IEEE; 2021:1-3. [doi: 10.1109/ICBC51069.2021.9461062]
97. Farooq K, Syed HJ, Alqahtani SO, Nagmeldin W, Ibrahim AO, Gani A. Blockchain federated learning for in-home health monitoring. Electronics (Basel) 2022 Dec 28;12(1):136. [doi: 10.3390/electronics12010136]
98. Chang Y, Fang C, Sun W. A blockchain-based federated learning method for smart healthcare. Comput Intell Neurosci 2021;2021(1):4376418. [doi: 10.1155/2021/4376418] [Medline: 34868289]
99. Ali A, Al-rimy BAS, Tin TT, Altamimi SN, Qasem SN, Saeed F. Empowering precision medicine: unlocking revolutionary insights through blockchain-enabled federated learning and electronic medical records. Sensors (Basel) 2023 Aug 28;23(17):7476. [doi: 10.3390/s23177476]
100. Kumar O, Sudhakaran G, Balaji V, Nhaveen A, Sai Balakrishnan S. Securing health care data through blockchain enabled collaborative machine learning. Research Square. Preprint posted online on May 23, 2023. [doi: 10.21203/rs.3.rs-2205379/v1]
101. Hasan MR, Li Q, Saha U, Li J. Decentralized and secure collaborative framework for personalized diabetes prediction. Biomedicines 2024 Aug 21;12(8):1916. [doi: 10.3390/biomedicines12081916]
102. Moulahi W, Jdey I, Moulahi T, Alawida M, Alabdulatif A. A blockchain-based federated learning mechanism for privacy preservation of healthcare IoT data. Comput Biol Med 2023 Dec;167:107630. [doi: 10.1016/j.compbiomed.2023.107630] [Medline: 37952305]
103. Ramani R, Rosline Mary A, Edwin Raja S, Arun Shunmugam D. Optimized data management and secured federated learning in the Internet of Medical Things (IoMT) with blockchain technology. Biomed Signal Process Control 2024 Jul;93:106213. [doi: 10.1016/j.bspc.2024.106213]
104. Kumar R, Bernard CM, Ullah A, et al. Privacy-preserving blockchain-based federated learning for brain tumor segmentation. Comput Biol Med 2024 Jul;177:108646. [doi: 10.1016/j.compbiomed.2024.108646]
105. Heidari A, Javaheri D, Toumaj S, Navimipour NJ, Rezaei M, Unal M. A new lung cancer detection method based on the chest CT images using federated learning and blockchain systems. Artif Intell Med 2023 Jul;141:102572. [doi: 10.1016/j.artmed.2023.102572]
106. Liang X, Bandara E, Zhao J, Shetty S. A blockchain-empowered federated learning system and the promising use in drug discovery. In: Charles W, editor. Blockchain in Life Sciences: Springer; 2022:113-139. [doi: 10.1007/978-981-19-2976-2_6]
107. Hai T, Zhou J, Srividhya SR, Jain SK, Young P, Agrawal S. BVFLEMR: an integrated federated learning and blockchain technology for cloud-based medical records recommendation system. J Cloud Comp 2022 Dec;11(1):22. [doi: 10.1186/s13677-022-00294-6]
108. Gayathri Hegde M, Shrishti Bekal M, Deepa Shenoy P, Venugopal KR. Preserving privacy and security of electronic health records using blockchain-based federated learning (BFL) framework. In: 2023 IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC): IEEE; 2023:853-859. [doi: 10.1109/R10-HTC57504.2023.10461823]
109. Guduri M, Chakraborty C, Maheswari U, Margala M. Blockchain-based federated learning technique for privacy preservation and security of smart electronic health records. IEEE Trans Consumer Electron 2024 Feb;70(1):2608-2617. [doi: 10.1109/TCE.2023.3315415]
110. Alzubi JA, Alzubi OA, Singh A, Ramachandran M. Cloud-IIoT-based electronic health record privacy-preserving by CNN and blockchain-enabled federated learning. IEEE Trans Ind Inf 2023 Jan;19(1):1080-1087. [doi: 10.1109/TII.2022.3189170]

XSL•FO
RenderX

111. Rahman MA, Hossain MS, Islam MS, Alrajeh NA, Muhammad G. Secure and provenance enhanced Internet of Health Things framework: a blockchain managed federated learning approach. IEEE Access 2020;8:205071-205087. [doi: 10.1109/ACCESS.2020.3037474]
112. Muazu T, Yingchi M, Muhammad AU, Ibrahim M, Samuel O, Tiwari P. IoMT: a medical resource management system using edge empowered blockchain federated learning. IEEE Trans Netw Serv Manage 2024 Feb;21(1):517-534. [doi: 10.1109/TNSM.2023.3308331]
113. Dhasaratha C, Hasan MK, Islam S, et al. Data privacy model using blockchain reinforcement federated learning approach for scalable Internet of Medical Things. CAAI Trans Intel Tech 2024. [doi: 10.1049/cit2.12287]
114. Lakhan A, Mohammed MA, Nedoma J, et al. Federated-learning based privacy preservation and fraud-enabled blockchain IoMT system for healthcare. IEEE J Biomed Health Inform 2023 Feb;27(2):664-672. [doi: 10.1109/JBHI.2022.3165945]
115. Baucas MJ, Spachos P, Plataniotis KN. Federated learning and blockchain-enabled fog-IoT platform for wearables in predictive healthcare. IEEE Trans Comput Soc Syst 2023;10(4):1732-1741. [doi: 10.1109/TCSS.2023.3235950]
116. Raisaro JL, Marino F, Troncoso-Pastoriza J, et al. SCOR: a secure international informatics infrastructure to investigate COVID-19. J Am Med Inform Assoc 2020 Nov 1;27(11):1721-1726. [doi: 10.1093/jamia/ocaa172] [Medline: 32918447]
117. Samuel O, Omojo AB, Onuja AM, et al. IoMT: a COVID-19 healthcare system driven by federated learning and blockchain. IEEE J Biomed Health Inform 2023 Feb;27(2):823-834. [doi: 10.1109/JBHI.2022.3143576]
118. Kumar R, Khan AA, Kumar J, et al. Blockchain-federated-learning and deep learning models for COVID-19 detection using CT imaging. IEEE Sensors J 2021 Jul 15;21(14):16301-16314. [doi: 10.1109/JSEN.2021.3076767]
119. Durga R, Poovammal E. FLED-Block: federated learning ensembled deep learning blockchain model for COVID-19 prediction. Front Public Health 2022 Jun 17;10:892499. [doi: 10.3389/fpubh.2022.892499]
120. Abdel-Basset M, Alrashdi I, Hawash H, Sallam K, Hameed IA. Towards efficient and trustworthy pandemic diagnosis in smart cities: a blockchain-based federated learning approach. Mathematics 2023 Jul 13;11(14):3093. [doi: 10.3390/math11143093]
121. Gupta R, Thakker U, Tanwar S, Obaidat MS, Hsiao KF. BITS: a blockchain-driven intelligent scheme for telesurgery system. In: 2020 International Conference on Computer, Information and Telecommunication Systems (CITS): IEEE; 2020:1-5. [doi: 10.1109/CITS49457.2020.9232662]
122. Chaudjary S, Kakkar R, Gupta R, Tanwar S, Agrawal S, Sharma R. Blockchain and federated learning-based security solutions for telesurgery system: a comprehensive review. Turk J Electr Eng Comput Sci 2022;30(7):2446-2488. [doi: 10.55730/1300-0632.3950]
123. Martínez Beltrán ET, Pérez MQ, Sánchez PMS, et al. Decentralized federated learning: fundamentals, state of the art, frameworks, trends, and challenges. IEEE Commun Surv Tutorials 2023;25(4):2983-3013. [doi: 10.1109/COMST.2023.3315746]
124. Wu L, Ruan W, Hu J, He Y. A survey on blockchain-based federated learning. Future Internet 2023 Dec 12;15(12):400. [doi: 10.3390/fi15120400]
125. Wu H, Wang P. Node selection toward faster convergence for federated learning on non-IID data. IEEE Trans Netw Sci Eng 2022 Sep 1;9(5):3099-3111. [doi: 10.1109/TNSE.2022.3146399]
126. Melis L, Song C, De Cristofaro E, Shmatikov V. Exploiting unintended feature leakage in collaborative learning. In: 2019 IEEE Symposium on Security and Privacy (SP): IEEE; 2019:691-706. [doi: 10.1109/SP.2019.00029]
127. Konečný J, McMahan HB, Yu FX, Richtárik P, Suresh AT, Bacon D. Federated learning: strategies for improving communication efficiency. arXiv. Preprint posted online on Oct 30, 2017. [doi: 10.48550/arXiv.1610.05492]
128. Zhang H, Liu J, Jia J, Zhou Y, Dai H, Dou D. FedDUAP: federated learning with dynamic update and adaptive pruning using shared data on the server. arXiv. Preprint posted online on Apr 25, 2022. [doi: 10.48550/ARXIV.2204.11536]
129. Antunes RS, André da Costa C, Küderle A, Yari IA, Eskofier B. Federated learning for healthcare: systematic review and architecture proposal. ACM Trans Intell Syst Technol 2022 Aug 31;13(4):1-23. [doi: 10.1145/3501813]
130. Cui L, Su X, Ming Z, et al. CREAT: blockchain-assisted compression algorithm of federated learning for content caching in edge computing. IEEE Internet Things J 2022 Aug 15;9(16):14151-14161. [doi: 10.1109/JIOT.2020.3014370]

Abbreviations

AI: artificial intelligence
BCFL: blockchain-based federated learning
BFLPD: blockchain-based federated learning for pandemic diagnosis
EHR: electronic health record
EMR: electronic medical record
FBCLC-Rad: Federated Learning–Enabled Blockchain CapsNets Lung Cancer Radiologist
FedAvg: Federated Averaging algorithm
FL: federated learning
FL-BETS: Federated Learning–Based Blockchain-Enabled Task Scheduling
FLED-Block: Federated Learning–Ensembled Deep Learning Blockchain Model

XSL•FO
RenderX

GDPR: General Data Protection Regulation
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
gRPC API: Google Remote Procedure Calls – Application Programming Interface
HIPAA: Health Insurance Portability and Accountability Act
ICD-10: *International Statistical Classification of Diseases, Tenth Revision*
ICD-10-CM: *International Classification of Diseases, Tenth Revision, Clinical Modification*
IoMT: Internet of Medical Things
IoT: Internet of Things
ML: machine learning
non-IID: non-independent and identically distributed
ODMSM: Optimized Data Management and Secured Federated Learning
PBFT: Practical Byzantine Fault Tolerance
PoS: Proof of Stake
PoW: Proof of Work
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
REST-API: Representational State Transfer – Application Programming Interface
RPC: Remote Procedure Calls
SMPC: Secure Multiparty Computation
ZKP: zero-knowledge proof

Edited by J Sarvestan; submitted 14.Jun.2025; peer-reviewed by S Fitzek, S Poruran; accepted 27.Nov.2025; published 19.Feb.2026.

Please cite as:
Wang X, Xie Y, Chen X, Yang J, Li R, Gao W, Yan Z, Zhou H, Ye Z
Securing Federated Learning With Blockchain in the Medical Field: Systematic Literature Review
J Med Internet Res 2026;28:e79052
URL: https://www.jmir.org/2026/1/e79052
doi:10.2196/79052

© Xudong Wang, Yi Xie, Xiaoliang Chen, Jiaming Yang, Ruiyuan Li, Weihang Gao, Zineng Yan, Hong Zhou, Zhewei Ye. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Inequalities in the Use of eHealth Services in European Public Health Care Systems: Systematic Review of Observational Studies

Garazi Monasterio[1,2], MSc; Marcos José Fernández-López[2], MSc; Erika Valero[1,2], PhD; Unai Martin[2,3,4], PhD; Amaia Ayala-García[2,3,4], PhD

[1]Department of Nursing I, Faculty of Medicine and Nursing, University of the Basque Country (UPV/EHU), Barrio Sarriena s/n, Leioa, Spain

[2]Research Group on Social Determinants of Health and Demographic Change-OPIK, University of the Basque Country (UPV/EHU), Leioa, Spain

[3]Department of Sociology and Social Work, Faculty of Social Sciences and Communication, University of the Basque Country (UPV/EHU), Leioa, Spain

[4]Biobizkaia Health Research Institute, Barakaldo, Spain

Corresponding Author:
Garazi Monasterio, MSc
Department of Nursing I, Faculty of Medicine and Nursing, University of the Basque Country (UPV/EHU), Barrio Sarriena s/n, Leioa, Spain

Abstract

Background: European public health care systems are expanding eHealth tools such as teleconsultations, online appointment bookings, and electronic health records to improve efficiency and access to health care. However, their use depends on factors such as digital skills and internet access, which are unequally distributed across socioeconomic and demographic determinants. Most existing evidence on these inequalities is qualitative or outside universal health care systems.

Objective: This systematic review aims to synthesize quantitative evidence on social inequalities in access to and use of eHealth services within European public health care systems. Specifically, we sought to identify which social determinants were most consistently associated with unequal use of online appointment booking, teleconsultations, electronic health records, and eHealth portals, across major social determinants of health.

Methods: A systematic search was conducted across PubMed, Scopus, Web of Science, and PsycINFO for studies published in English or Spanish between 2015 and October 2025. Eligible quantitative studies included adults (≥18 years) using public health care systems in European countries. The primary outcome was differential access to or use of eHealth tools by social determinants in any level of care. Screening and data extraction were independently performed by 3 reviewers using Rayyan, resolving disagreements through consensus. Data extracted covered study design, population, eHealth tools, social determinants, and outcomes. Risk of bias was evaluated using Joanna Briggs Institute tools. Due to study heterogeneity in digital tools and inequality dimensions, results were synthesized narratively by tool type and social inequality factors. Point estimates and 95% CIs were extracted when available.

Results: Of the 2366 records retrieved, 18 observational studies met the inclusion criteria: 13 cross-sectional, 3 prevalence, 1 retrospective cohort, and 1 ecological cohort. Publication output increased from 2020 onward, mostly driven by cross-sectional studies from northern and western Europe. Findings revealed consistent social gradients in eHealth use: older adults, individuals with lower educational or socioeconomic level, ethnic minorities, and those with limited digital skills or poorer health were less likely to use eHealth tools. Most studies were rated as high quality (78%), and the remainder as moderate, heterogeneity in designs, outcomes, and populations may limit generalizability.

Conclusions: Digital transformation in European public health systems has not benefited all groups equally. This review highlights persistent social inequalities in the use of key digital health tools. While many included studies were of high quality, heterogeneity in study designs, populations, and outcomes, as well as risk of bias, limits causal inference and the direct translation of findings into policy and practice. The findings nonetheless reveal systematic patterns of exclusion that are highly relevant for policy. Emphasizing an intersectional approach and standardizing measures of digital access will be essential to develop effective, equity-focused policies that ensure inclusive digital health services for all.

Trial Registration: PROSPERO CRD420251015756; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251015756

(J Med Internet Res 2026;28:e81841) doi:10.2196/81841

XSL•FO
RenderX

KEYWORDS

eHealth; digital health; health services accessibility; delivery of health care; health inequities; health care disparities; social determinants of health; public health infrastructure

Introduction

Health systems worldwide are undergoing a strong shift toward eHealth [1,2], also referred to as digital health, telehealth, or telemedicine. This transition, long encouraged by institutions such as the World Health Organization (WHO) [3] to achieve the 2030 Sustainable Development Goals, was accelerated by the COVID-19 pandemic [4,5].

One of the earliest and most widespread applications of eHealth is the online appointment booking, teleconsultations, access to electronic health records (EHRs), and integrated digital health portals that bundle multiple services in a single interface [6,7]. These 4 modalities constitute the most common digital interaction points with health care services across European public health systems and form the conceptual foundation for evaluating digital health access and usage in this review [8]. Beyond offering advantages for health care systems and for patients [9,10], these tools have also been identified as potential levers for improving access in underserved settings, such as rural or remote areas [11].

However, there is comparatively less research on how these same tools may reinforce or even exacerbate existing social inequalities in access and usage. As Western et al [12] have conceptualized, these inequalities manifest across 3 interrelated levels of the digital divide: (1) access to internet connectivity and digital devices; (2) digital skills and literacy, trust in technology, and willingness to adopt new platforms; and (3) disparities in actual outcomes or health improvements derived from digital tool use. Only those who overcome these first 2 levels can potentially benefit, leading to the third digital health divide. Breakdowns at any of these levels can cascade to reinforce exclusion at the next, perpetuating digital and health inequities. Quantitative studies, many outside universal health care settings, point out that older adults, people with lower socioeconomic status or education [13], migrated population, ethnic minorities, rural population, or in low-skilled jobs [14], individuals with impairments [15], and other socially disadvantaged groups often lack these prerequisites [16], revealing measurable gaps in digital health use [17,18]. Consequently, digital health strategies may risk amplifying pre-existing health disparities [19]. In addition, digital divide measurement methods and indicators vary widely, from composite indexes (aggregated scores combining multiple dimensions of access and use) to simple usage metrics (such as number of logins or appointments booked) [18,20]. Yet, the absence of harmonized indicators makes it difficult to compare findings across countries, especially in health care systems with very different structures.

Within Europe's public health care systems—across all levels of care—there remains scant data quantifying inequalities in basic digital functions such as booking appointments, accessing health portals, or having teleconsultations.

Despite these concerns, public health systems across Europe (and beyond) are investing heavily in digital care models [21]. While this expansion signals progress, it also highlights the lack of robust monitoring of equity impacts. Most existing studies emphasize qualitative evidence, which provides valuable perspectives but limited measurement of scale [22]. Quantitative research is less common, and when available, it rarely disaggregates results by key social determinants, leaving important gaps in understanding how digital health inequalities unfold across vulnerable groups.

In recent years, this issue has attracted renewed policy and research attention, reflecting a growing recognition that digital transformation alone does not guarantee equitable access [23]. Several recent analyses and policy reports have documented substantial advances in national eHealth strategies and portal deployment across Europe, while noting that evaluation frameworks for equity and inclusion remain limited [24,25]. At the same time, comparative assessments by international organizations emphasize the need for harmonized indicators and monitoring systems to assess how digitalization affects different population groups within public health care [26]. This evolving evidence base reinforces the importance of systematically consolidating quantitative findings to clarify where inequalities persist and how they are measured within European public health systems.

In this context and considering that the WHO's Global Strategy on Digital Health 2020 - 2025 (extended to 2027) emphasizes that eHealth should enhance universal access, quality, efficiency, and equity [8], there is an urgent need to identify who is falling behind. To date, no systematic review has synthesized quantitative evidence on these digital inequalities across European public health care systems.

The objective of this systematic review is, therefore, to synthesize the available quantitative evidence on inequalities in access to and use of core eHealth services—including online appointment booking, teleconsultations, and EHRs—within European public health care systems between 2015 and 2025, examining how key social determinants, such as age, gender, socioeconomic status, education, migration background, and digital literacy, are associated with the use of these services.

Methods

Overview

A systematic review was conducted to identify studies examining inequalities in the use of eHealth technologies to access health care, provided by public health care systems, in relation to social determinants of health in Europe. This review was carried out in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [27] guidelines to ensure a rigorous and transparent evaluation of the review process. For details of items included in the PRISMA checklist, please see Checklist 1. The protocol was

XSL•FO
RenderX

registered in the PROSPERO database (registration number: CRD420251015756).

Information Sources and Search Strategy

A systematic search was performed in PubMed, Scopus, PsycINFO, and Web of Science up to April 2025. The search strategy combined controlled vocabulary and free-text terms related to digital health (eg, "telemedicine," "eHealth," "digital health," and "patient portal"), social inequalities (eg, "health disparities," "digital divide," "inequality," and "equity"), social determinants and user characteristics (eg, "age," "gender," "ethnic minorities," "disability," and "socioeconomic"), health care access and use (eg, "access," "use," "user," and "nonuser") and geographical terms referring to Europe and individual European countries (eg, "Europe," "United Kingdom," "Germany," and "Spain"). Truncation and Boolean operators (AND, OR, NOT) were used to combine these concepts and optimize sensitivity and specificity. The search strategy was developed collaboratively by the research team, without consultation with an information scientist. Complete search strategies and term lists are available in the Multimedia Appendix 1.

The search process and its reporting followed the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) guidelines to ensure transparency, completeness, and reproducibility of each search component [28]. In addition, a backward citation search (snowballing) was conducted from the reference lists of the included studies. The literature search was rerun on October 20, 2025, to capture any recent publications since the initial search.

Inclusion and Exclusion Criteria

Studies were eligible for inclusion if they met the following criteria: (1) study population focused on adults aged 18 years or older. Studies that included younger populations but allowed for interpretation of results in adults more than 18 years were also included; (2) assessed any type of eHealth technology designed to connect individuals to public health systems; (3) reported quantitative data related to use of digital health services provided by public health systems in association with social determinants of health; (4) were conducted in European countries, defined as member states of the European Union, the European Economic Area (Iceland, Liechtenstein, and Norway), the United Kingdom, and Switzerland. This definition was selected to ensure conceptual and structural comparability of health care systems with public health care systems coverage and established digital health infrastructure; and (5) were peer-reviewed studies published in English or Spanish between January 2015 and April 2025.

The following exclusion criteria were applied: (1) studies focusing exclusively on populations using private health care services; (2) focusing on the effects of the COVID-19 pandemic on eHealth usage; (3) focusing on health apps designed for preventive purposes, such as interventions, general health information provision, self-care, or supporting the management of specific conditions; and (4) qualitative studies, study protocols, conference abstracts, theses, editorials, opinion pieces, and systematic reviews.

Data Management and Study Selection

Study screening was conducted using the Rayyan web-based tool [29]. No automation tools were used in this process. Two researchers independently screened titles and abstracts (GM and AA), followed by full-text review of potentially eligible studies. Disagreements were resolved through a third reviewer (MF). Study authors were contacted when additional information or clarification was required.

Data Extraction

Data extraction was carried out collaboratively by 3 researchers (GM, AA, and MF) and reviewed by the other team members to ensure accuracy and consistency. The following information was extracted: authorship, year of publication, country, study design, population characteristics, type of eHealth technology, outcome, social determinants of health included, and key quantitative findings and their narrative description. In addition, eHealth tools identified in the studies were regrouped into 4 categories according to their functionality in (1) eHealth portals—which typically integrate multiple functionalities into a single tool. These include services such as prescription renewal, appointment reminders, asynchronous messaging with health care professionals, access to test results, and in some cases, teleconsultation features or visualization of EHRs; (2) EHRs—this category includes those tools that allow users to consult their medical results, diagnoses, clinical notes, and other information recorded in their health record. These functionalities may be, and often are, integrated into digital portals managed by public health services or specific hospital apps. Notable examples include platforms such as MyCare, myHealth@QEHB, or the Care Information Exchange in the United Kingdom; (3) telemedicine and remote primary care—understood as the provision of direct health care through digital channels. It includes video medical consultations, direct-to-consumer digital care models, as well as the remote provision of primary care; (4) online appointment booking. This classification enabled a more nuanced analysis of inequalities based on the specific functionality assessed in each study. To ensure clarity in the tables, this categorization is represented numerically (1-4).

Given the variability in terminology and definitions across included studies, we compiled a comparative table showing how each study defined key determinants and eHealth usage, which is provided in Multimedia Appendix 2.

Risk of Bias Assessment

Risk of bias was assessed using the appropriate critical appraisal tools developed by the Joanna Briggs Institute (JBI), selected based on each study's methodological design (eg, cross-sectional and cohort) [30]. The assessment was conducted at the study level rather than at the outcome level. This approach was chosen because the included studies were primarily observational and reported heterogeneous outcomes, making study-level assessment the most consistent and feasible strategy. The appraisal focused on key domains affecting internal validity, including participant selection, measurement of exposures and outcomes, and control of confounding factors. Studies were

XSL•FO
RenderX

classified into 3 quality categories based on the proportion of affirmative responses to the JBI appraisal items: high quality (≥70%), moderate quality (40% - 69%), and low quality (≤39%) [31]. A detailed summary of the risk of bias assessment is provided in Multimedia Appendix 3 [32-49], using the corresponding JBI Critical Appraisal Tool.

At least 2 reviewers assessed the risk of bias independently. Discrepancies were resolved through discussion and consensus, involving 3 researchers (GM, AA, and MF) during this phase.

Data Synthesis and Effect Measures

The effect measures reported in the studies varied. Some studies presented descriptive comparisons between population subgroups (such as percentages or proportions), while others reported adjusted measures of association derived from regression models, such as odds ratios (OR) with corresponding CIs. When available, information was classified according to the type of eHealth service analyzed.

In addition, some studies applied more advanced equity metrics, such as the concentration index and the horizontal inequity index. The concentration index summarizes whether use of services is disproportionately concentrated among socioeconomically advantaged or disadvantaged groups. The horizontal inequity index adjusts for differences in health care needs (eg, morbidity), isolating inequities that persist beyond expected variations in medical necessity. A few studies also used decomposition analysis to examine the contribution of individual determinants (eg, education, income, and employment) to overall inequality, and indirect standardization to account for differences in health needs when comparing groups.

Due to the heterogeneity in study designs, eHealth technologies evaluated, and inequality dimensions analyzed, a narrative synthesis was therefore conducted to summarize the findings. Results were presented in evidence tables and described narratively in the text. Studies with comparable outcomes were grouped, as explained in the “Data Extraction” section.

Results

Study Selection

The initial search identified 2366 records from 4 databases. Of these, ultimately, 18 studies [32-49] were included in the systematic review. A PRISMA 2020 flow diagram detailing the selection process is presented in Figure 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram of study selection process.

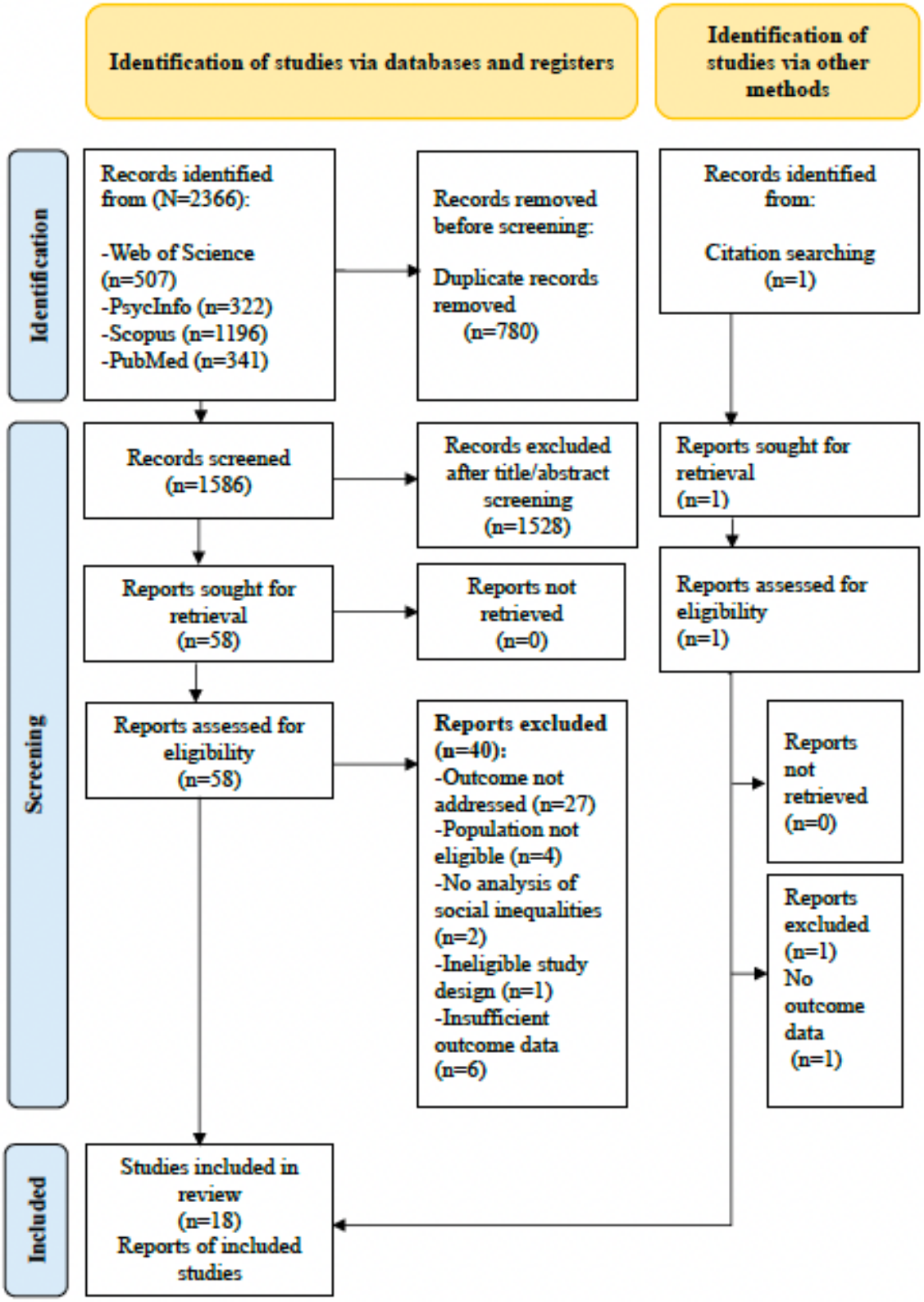

Characteristics of the Studies Included

The included studies were published between 2017 and 2025. A general upward trend in the number of publications was observed over time, with a noticeable increase in high-quality studies in the most recent years. The distribution of included studies per year according to their methodological quality category is illustrated in Figure 2, where the trend line represents the total number of studies published per year, and the green and yellow "x"'s represent the quality of the individual studies per year.

XSL•FO
RenderX

Figure 2. Publications per year by quality category.

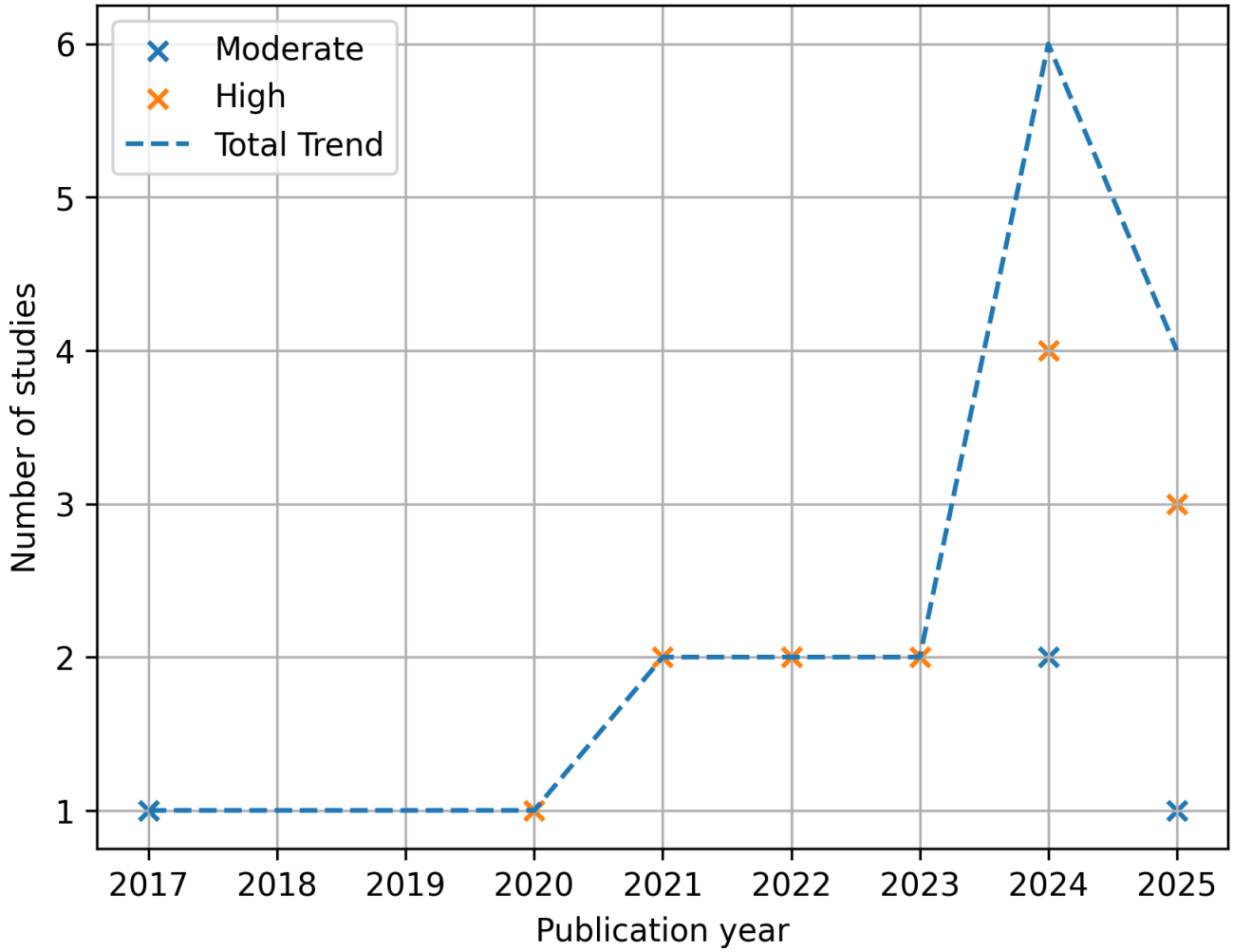

The most frequent study design was cross-sectional observational, accounting for 72% (13/18) of the studies [32-37,39-42,45,46,49], followed by prevalence studies at 17% (3/18) [38,43,47], with one retrospective cohort [44] and one ecological cohort study each representing 6% (1/18) [48]. The 18 studies [32-49] were conducted across 10 European countries, with the majority from Northern and Western Europe. Sweden had the highest number of studies, representing 28% (5/18) [33,36,37,44,45], followed by England at 22% (4/18) [35,46,48,49] and Finland at 11% (2/18) [32,41]. Other countries contributed one study each: Spain [42], Norway [47], Germany [40], Iceland [38], Denmark [43], the United Kingdom (distinct from England) [34], and one pan-European study [39]. Sample sizes ranged from under 438 to 1,991,995 participants. They varied from small studies such as Hörhammer et al [32] in Finland (n=438) to very large population-based datasets such as Dahlgren et al [33] in Sweden (n=1,991,995). Specifically, 17% (3/18) of studies had sample sizes below 1000 [32,34,35], 44% (8/18) ranged between 1000 and 10,000 [36-43], 28% (5/18) exceeded 10,000 participants [33,44-47], and 11% (2/18) did not report sample size [48,49]. Most studies focused on the general adult population aged 18 years and older.

Risk of Bias Assessment

According to the risk of bias assessment, 22% (4/18) of the studies [40,42,43,48] were classified as of moderate quality, while the remaining 78% (14/18) were rated as high quality [32-39,41,44-47,49]. No studies were considered to have low methodological quality, suggesting that the overall quality of the evidence included was generally high.

The eHealth tools analyzed were regrouped into 4 functional categories as described in the “Methods” section: (1) eHealth portals, (2) EHRs, (3) telemedicine and remote primary care, and (4) online appointment booking. Of the 18 included studies [32-49], 56% (10/18) analyzed eHealth portals [32,34,35,38-41,43,48,49], 33% (6/18) examined EHRs [37,41,42,46-48], 22% (4/18) addressed telemedicine and remote primary care [33,36,44,45], and only 11.1% (2/18) focused on online appointment booking [41,48]. Figure 3 presents the distribution of eHealth tool categories by publication year, highlighting a growing diversification in digital health functionalities over time. Studies appear more than once when assessing more than one tool usage.

XSL•FO
RenderX

Figure 3. Assessment of eHealth tools per year. EHR: electronic health record.

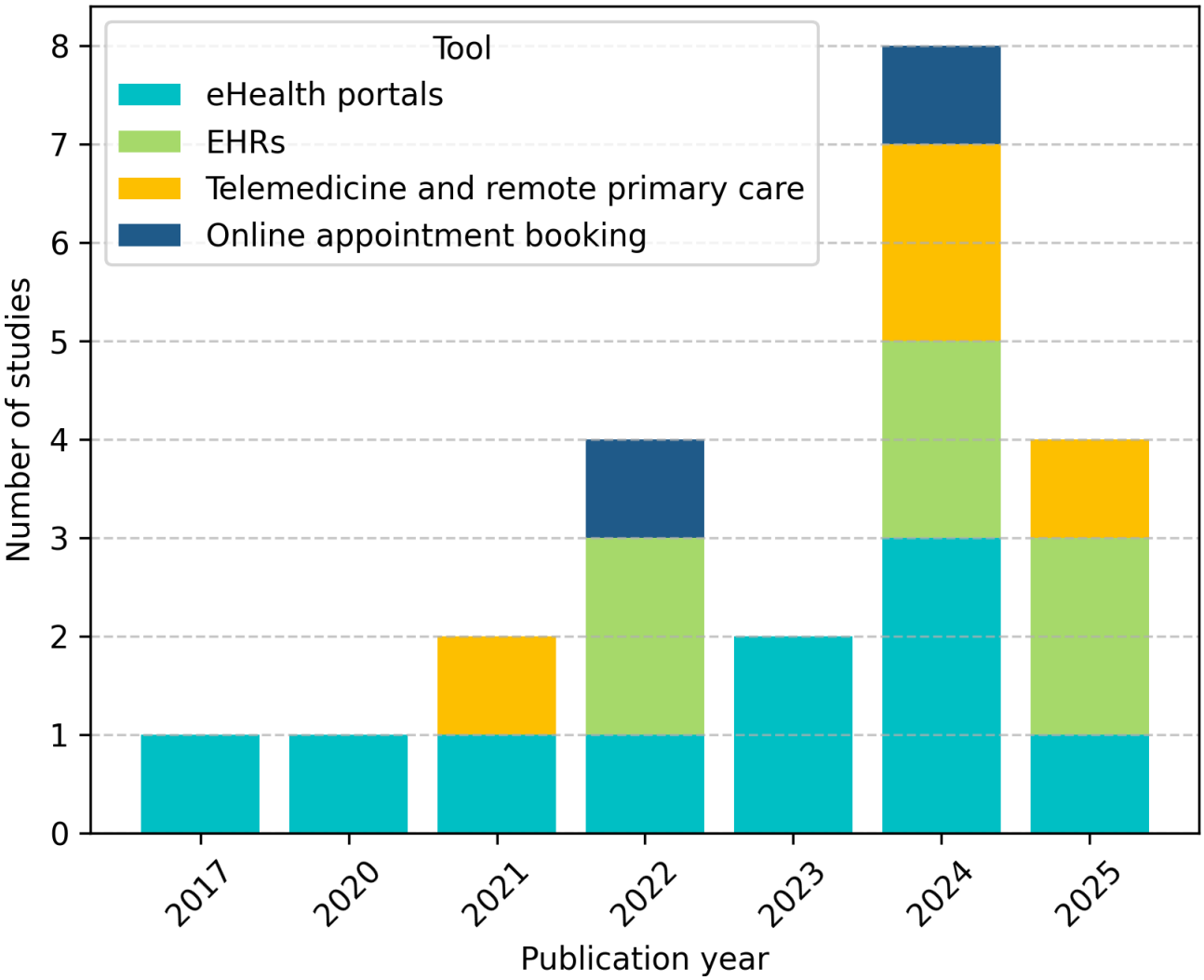

All 18 studies reviewed examined social inequalities in the use of eHealth tools, with age and gender being the most analyzed determinants [32-49]. A second major block of variables related to socioeconomic position, especially social class [39,43,44], education [33,34,36,37,42,45], employment [32,36,40,45], income [33,45], and deprivation [35,46,48,49], often using both individual-level and area-based indicators. Fewer studies addressed cultural-related factors such as ethnicity [34,46,48,49], country of birth [33,45], nationality [42], and language [35,40,46], though these were sometimes used as proxies for integration and accessibility. Determinants related to residential context (eg, urban vs rural [40,45,49] and distance to care [33]) and family structure (eg, marital status [37,39,40], household size [39], and parity [35]) were included less frequently. Health status and clinical complexity were operationalized in diverse ways, including self-rated health [34,36,37], chronic conditions [32,33,48,49], diagnoses or number of specialties involved [45,46] or resource use band [44]. Finally, several studies assessed digital competence [41,42], literacy [32,34], and internet habits [36], highlighting the role of skills and confidence in shaping access to digital health care. Figure 4 presents a heatmap summarizing the extent to which each social determinant was not only examined but also found to be associated with inequalities in the use of specific eHealth tools. Each cell displays the proportion of studies that reported significant disparities out of those that analyzed a given determinant for each tool. For example, a value of “5/5” under “gender” and “EHRs” indicates that all 5 studies assessing gender differences in the use of EHRs reported significant inequalities by this determinant. This visual summary highlights both the breadth of research on each determinant and the consistency of evidence across different tools. eHealth portals were the most frequently studied, covering a wide range of social factors; however, the findings were somewhat mixed, with inequalities observed in some areas (eg, age and education) but not others (eg, gender and ethnicity). In contrast, studies on EHRs and telemedicine and remote primary care, though less numerous, tended to yield more consistent evidence of social inequalities, particularly along the lines of age, gender, and socioeconomic status. Finally, online appointment booking was the least explored modality, with very limited evidence available.

XSL•FO
RenderX

Figure 4. Evidence of social inequalities in eHealth tool use. EHR: electronic health record.

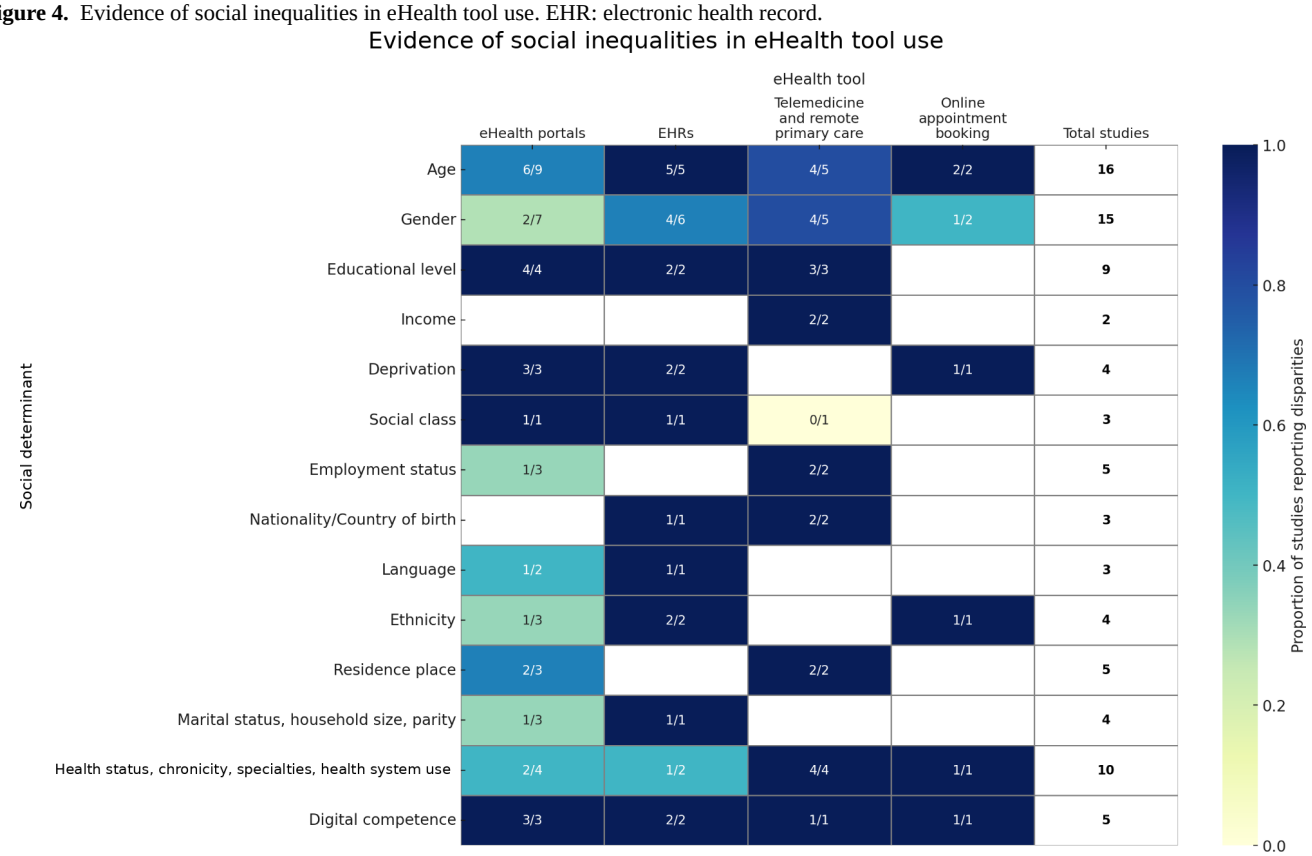

A summary of the main characteristics and key findings of the included studies is presented in Table 1. A more detailed version is available in Multimedia Appendix 4 [32-49].

XSL•FO
RenderX

Table . Evidence table summarizing the main information of each study included in the systematic review, ordered by study design.

Author (year); Country	Design, Sample, Population, and Statistical analysis	eHealth Cat[a] and study Outcome	SDH[b]	Description of main finding of inequalities in the studies
González-Cacheda et al [42] (2025); Spain	• Cross-sectional • Population present in the Health Barometer 2022 (n=7454) • Univariate logistic regressions	(2) Use of medical records via internet	• Age • Gender • Education • Nationality • Social class • Device use	• Older adults: higher awareness (OR[c] 0.86) but lower effective use than younger and middle-aged. • No gender differences. • Higher education (OR 1.41) and digital experience (OR 1.51) associated with greater use. • Spanish nationals are more likely to use digital medical records (OR 1.45). 95% CI not reported in the primary study.
Hörhammer et al [32] (2025); Finland	• Cross-sectional • Patients from a Finnish mental health and substance abuse unit (n=438) • Univariate and multivariate logistic regressions	(1) Use of digital services	• Age • Gender • Health confidence • Employment • Duration of care	• Older adults (≥65 years; OR 0.40, 95% CI 0.21 - 0.75) less likely to use digital services. • Low health confidence is associated with reduced use (OR 0.62, 95% CI 0.45 - 0.86).
Knöchelmann et al [40] (2024); Germany	• Cross-sectional • Adults (≥ 18 years) from the HeReCa panel (n=1821) • Descriptive analysis, latent class analysis, and multinomial logistic regression	(1) Use of digitalized health care services	• Age • Gender • Education • Language • Employment • Place of residence • Marital status	• No significant associations for women (OR 0.78, 95% CI 0.56 - 1.10), age (OR 0.88, 95% CI 0.71 - 1.08), or marital status (married vs others; OR 0.77, 95% CI 0.53 - 1.13). • Higher education is linked to greater odds of rejecting participation (OR 1.23) but lower odds of active use (OR 0.68). 95% CI not reported in the primary study. • Lower educational attainment (OR 0.67, 95% CI 0.47 - 0.95) and unemployment (OR 0.57, 95% CI 0.38 - 0.84) associated with reduced active use. • Number of previous illnesses not associated with use.

XSL•FO
RenderX

Author (year); Country	Design, Sample, Population, and Statistical analysis	eHealth Cat[a] and study Outcome	SDH[b]	Description of main finding of inequalities in the studies
Söderberg et al [36]; Sweden	• Cross-sectional • Adults (≥ 18 years) residing in Sweden (n=2716) • Multivariate logistic regression	(3) Use of digital primary care	• Age • Gender • Occupation • Education • Internet habits • Self-rated health	• Younger age associated with higher likelihood of seeking digital primary care (95% CI not reported in the primary study). • University education (OR 1.41, 95% CI 1.19 - 1.67) and daily internet use (OR 3.21, 95% CI 1.30 - 7.90) linked to greater use. • Retired individuals had lower odds than those working (OR 0.68, 95% CI 0.46 - 1.02; borderline nonsignificant).
Wilkens et al (2024) [45]; Sweden	• Cross-sectional • Patients ≥18 years (n=726,087) • Concentration index and curves; decomposition analysis; indirect standardization; horizontal inequity index	(3) Use of digital primary care	• Age • Gender • Education • Morbidity • Country of birth • Geographic region • Employment	• Digital care users were younger and more often women. • Low-income patients used more office-based visits, whereas high-income patients had more digital contacts; income and employment explained most inequalities in office visits. • Pro-rich inequality in digital contacts was partly explained by higher education and being born in Sweden; digital users were generally healthier.
Muli et al (2024) [37]; Sweden	• Cross-sectional • Patients ≥18 years and guardians of minors living in Region Stockholm who had a consultation with a physician (n=3421) • Multivariate logistic regressions	(2) Access to patient-accessible electronic health records	• Age • Gender • Education • Marital status • Health status	• Younger age associated with higher likelihood of reading records (OR 0.97, 95% CI 0.95 - 0.98). • Women are more likely to have read their electronic health records (77% readers overall). • Partnered individuals (vs single; OR 0.60, 95% CI 0.44 - 0.99) and those with higher education were more likely to be readers. • No differences were observed by health status.
Zhang et al [49] (2023); England		(1) Use of National Health Service (NHS) app (NHS Digital) and a primary care portal	• Age • Deprivation • Ethnicity • Residence place • Long-term condition	

XSL•FO
RenderX

Author (year); Country	Design, Sample, Population, and Statistical analysis	eHealth Cat[a] and study Outcome	SDH[b]	Description of main finding of inequalities in the studies
	• Cross-sectional • Populations of primary care practices in National Health Service (no n reported) • Multivariate linear regression			• Practices with more male or chronically ill patients had lower usage (*P*<.01). • Strong socioeconomic gradient: all more deprived quintiles showed lower digital health use than the least deprived (Q1; difference = –2.047 units; *P*<.001). • Higher use in practices with more White and younger patients (15 - 34 years; +10.39% registration, *P*<.05) and in larger practices (*P*<.05).
Pierce et al [35] (2023); England	• Cross-sectional • Pregnant women booked into UCLH[d] for their initial antenatal appointment in February 2022, identified through the EPIC platform (n=636) • Descriptive	(1) Use of MyCare, an electronic patient portal	• Age • Parity • Language • Deprivation	• Non-users of MyCare were younger (mean age 30) and showed lower engagement with higher parity (mean parity =0.94, 1.47, 1.78 for high-, low-, and nonusers). • Women whose first language was not English were more common among non-users (48.7%), who also had lower socioeconomic status (mean SDI[e]=3.72).
Heponiemi et al [41] (2022); Finland	• Cross-sectional • Sample from the Population Register of Finland • >20 years (n=4495) • Multivariate logistic regression	(1,2, and 4) Use of online health services	• Age • Gender • Digital competence	• Use of all digital services declined sharply after age 60, except for appointment booking. • Women used online services 8% more than men. • Good digital competence strongly increased use: test results (OR 12.61, 95% CI 8.52 - 18.64), prescription renewal (OR 8.82, 95% CI 6.15 - 12.64), appointment scheduling (OR 10.91, 95% CI 7.24 - 16.44), and online appointments (OR 6.48, 95% CI 0.93 - 45.12; not significant).
Chapman et al [46] (2022); England		(2) Sign up and activation of myHealth @QEHB, a hospital-based Personal Health Record		

XSL•FO
RenderX

Author (year); Country	Design, Sample, Population, and Statistical analysis	eHealth Cat[a] and study Outcome	SDH[b]	Description of main finding of inequalities in the studies
	• Cross-sectional • Adults ≥18 years receiving hospital outpatient care (n=28,637) • Descriptive, univariate, and multivariate logistic regression		• Age • Gender • Deprivation • Ethnicity • Interpreted need • Number of hospital specialties	• Activation highest among adults aged 35 - 54 years; lower for ages 16 - 34 years (OR 0.80, 95% CI 0.70 - 0.91) and lowest for ≥75 years (OR 0.39, 95% CI 0.32 - 0.47). • Males are less likely to activate their Personal Health Record (OR 0.85, 95% CI 0.78 - 0.94) but more likely to sign up (OR 1.10, 95% CI 1.04 - 1.16); activation is 3 times higher in the least versus the most deprived areas (OR 2.99, 95% CI 2.40 - 3.71). • Asian (aOR[f] 0.61, 95% CI 0.53 - 0.71), Black (aOR 0.45, 95% CI 0.36 - 0.56), and mixed ethnic groups (aOR 0.77, 95% CI 0.60 - 0.97) less likely to activate than White patients; those not needing an interpreter had higher activation (OR 3.16, 95% CI 1.96 - 5.09). • Number of clinical specialties associated with sign-up (aOR 2.54, 95% CI 2.30 - 2.82) but not activation (aOR 0.88, 95% CI 0.79 - 0.97).
Neves et al [34] (2021); United Kingdom	• Cross-sectional • Patients >18 years of hospitals and primary care in London, registered in the Care Information Exchange (n=650) • Descriptive, univariate, and multivariate logistic regression	(1) Use of Care Information Exchange, a patient portal containing patient information	• Age • Gender • Education • Ethnicity • Health status eHEALS[g] score	• Higher education associated with increased odds of use: undergraduate and professional (aOR 1.58, 95% CI 1.04 - 2.39) and postgraduate and higher (aOR 2.38, 95% CI 1.42 - 4.02). • Greater digital literacy (>30) strongly predicted use (aOR 2.96, 95% CI 2.02 - 4.35). • Good health status linked to lower odds of use (aOR 0.58, 95% CI 0.37 - 0.91).
Dahlgren et al [33] (2021); Sweden		(3) Use of telemedicine to provide traditional primary care		

XSL•FO
RenderX

Author (year); Country	Design, Sample, Population, and Statistical analysis	eHealth Cat[a] and study Outcome	SDH[b]	Description of main finding of inequalities in the studies
	• Cross-sectional • Adults (≥18 years) residing in Stockholm County and registered with a publicly funded Primary Health Care provider (n=1,991,995) • Descriptive and multivariate logistic regression		• Age • Gender • Education • Income • Country of birth • Diagnosis of chronic conditions • Primary health care accessibility • Distance to primary health care	• Greater telemedicine use among women (OR 1.60, 95% CI 1.58 - 1.62), younger adults (19–25), those born in Sweden, and individuals with higher education and income. • Lower use among older adults (≥65 years), those born outside the EU28, with lower education or income, or with heart failure and diabetes; depression and COPD[h] or asthma predicted higher telemedicine use. • Digital care use is more concentrated in urban areas (67.5% of users).
Merkel et al [39] (2020); Europe	• Cross-sectional • Adults aged >65 years who use the internet (n=6900) • Multilevel logistic regression	(1) Use of internet-based health care services (any service, not specified)	• Age • Gender • Education • Social class • Employment • Marital status • Household size • Population density	• eHealth users were younger, better educated, and of higher social class. • Older age was associated with lower use (OR 0.97, 95% CI 0.96 - 0.98; *P*<.001) • Higher education (16 - 19 years: OR 1.43, 95% CI 1.15 - 2.79; ≥20 years: OR 1.95, 95% CI 1.54 - 2.46; *P*<.001) and higher social class (medium: OR 1.45, 95% CI 1.23 - 1.71; high: OR 2.00, 95% CI 1.53 - 2.61; *P*<.001) predicted greater eHealth use. • Living with a partner increased, and living alone reduced, eHealth use. • Urban residence (OR 1.23, 95% CI 1.02 - 1.48; *P*=.03) and higher national education participation among older adults (OR 1.06, 95% CI 1.01 - 1.13; *P*=.02) were positively associated with use.
Kharko et al [47] (2025); Norway Sweden Finland	• Prevalence • NORDeHEALTH 2022 Survey (n=27,038) • Descriptive	(2) Access to EHR[i]	• Gender	• Women less likely to have never accessed the EHR or to be first-time users. • Women are more likely to have visited the EHR >20 times (χ^2_4=57; *P*<.001).

XSL•FO
RenderX

Author (year); Country	Design, Sample, Population, and Statistical analysis	eHealth Cat[a] and study Outcome	SDH[b]	Description of main finding of inequalities in the studies
Pálsdóttir et al [38]; Iceland	• Prevalence • Adults (≥ 18 years) • Residing in Iceland (n=3000) • Descriptive	(1) Use of national digital health care system for communication with health professionals or access personal health information	• Age • Gender	• Women aged 18 - 35 years (*P*<.001) and 36 - 55 years (*P*<.10) used the digital system more frequently. • Women >56 years showed lower use compared to younger women.
Petersen et al [43] (2017); Denmark	• Prevalence • Danish adults (n=1059) • Descriptive	(1) Access and use of Sundhed.dk, a national health portal	• Education level	• Individuals with only primary education were less likely to use the Danish National Health Portal than those with higher education. • Portal use was 21% among those with primary education versus 60% among higher-educated individuals.
Eriksson et al [44] (2025); Sweden	• Retrospective cohort (registry-based) • Individuals of any age registered with Primary Health Care Centers, and having made at least 1 outpatient consultation during the study period (n=73,486) • Multivariable Logistic Regression	• (3) Use of telemedicine consultation in primary health care and EHR	• Age	• Younger adults (20-39 years) more likely to use telemedicine; older adults (60 - 79 and ≥80 years) less likely (IRR[j] 0.19, 95% CI 0.16 - 0.25). • Women showed higher telemedicine use (IRR 1.39, 95% CI 1.28 - 1.51). • Patients with high resource use (RUB 5; IRR[e] 2.67, 95% CI 0.20 - 3.54) and more EHR entries (IRR 1.50, 95% CI 1.32 - 1.72) showed higher use; no significant differences by Care Need Index.
Kc et al [48] (2024); England	• Ecological cohort • GP-registered[k] patients in practices with >200 patients between March 23, 2020, and June 27, 2022, aged ≥15, (no n reported) • Univariate negative binomial regressions	(1, 2, and 4) Use of the NHS app for registrations, log-ins, appointments booked, prescriptions, medical record views	• Age • Gender • Ethnicity • Deprivation • Health status	• Higher use in practices with more patients aged 15 - 34 years (+10.39% registration). • Lower use across more deprived areas (up to 38.84% lower registration in most vs least deprived quintile). • Use higher in larger practices and those with more White patients; lower where more male or chronically ill patients were registered.

[a]Author-defined grouping of eHealth tools categories.

[b]Social determinants of health assessed.

[c]OR: odds ratio.

XSL•FO
RenderX

[d]UCLH: University College London Hospitals.
[e]SDI: Social Deprivation Index.
[f]aOR: adjusted odds ratio.
[g]eHEALS: eHealth Literacy Scale.
[h]COPD: chronic obstructive pulmonary disease.
[i]EHR: electronic health record.
[j]IRR: incidence rate ratio.
[k]GP: general practitioner.

Overall Summary

The included studies revealed consistent social inequalities in access to telemedicine, structured along key social determinants of health and disproportionately affecting vulnerable groups.

Age

Age was the most consistently associated determinant, with lower use among adults aged 60 years and older, and higher engagement among younger and middle-aged [32,33,35-39,41,44-46,48,49].

Gender

Findings regarding gender were more heterogeneous. Several studies showed lower engagement among men across different modalities [32,33,37-39,44-48], while others found no significant differences [32,34,36,39,40].

Socioeconomic

Lower education was strongly associated with reduced use across most studies [33,34,36,37,39,42,43,45]. Similar patterns were seen for income [33,45], social class [39,42], and area deprivation [35,46,48,49]. Unemployment [36,45] was generally linked to reduced use, though findings were mixed in some contexts [32,39,40].

Migration, Ethnicity, and Language

Individuals born outside the country of study [33,42,45], belonging to ethnic minorities [46,48,49], or requiring interpreter support [35,46] were less likely to access digital health care services.

Place of Residence

In general, people living in rural or remote areas showed lower usage levels [33,45,49] compared to those in urban or more densely populated settings. However, some studies showed mixed or context-dependent patterns [39,40].

Household Composition

Results were less consistent. Some studies found lower engagement among individuals without a partner [37] or with more children [35], while others reported no significant associations [39,40].

Health Status

The findings were mixed. Chronic illness or poor self-rated health sometimes reduced engagement [33,36,45,48,49], while others observed higher usage among those with more frequent contact with the health care system [44].

Digital Skills

Across the studies that assessed this dimension, lower levels of digital competence [41], limited prior digital experience [42], lack of access to a device [36], low health literacy [34], and low confidence in using technology [32] were all associated with significantly lower use of digital health services.

In the next section, we break down these findings by type of eHealth service used.

eHealth Portal

Among the 10 studies that analyzed eHealth portals [32,34,35,38-41,43,48,49], 9 included age as a determinant [32,34,35,38-41,48,49]. Of these, 78% (6/9) found age to be a consistent factor associated with portal use [32,35,38,39,41,49]. Younger adults (15-34 years) showed higher use, and use declined after the age of 60 years [45]. Two other studies found no differences [34,46].

Regarding gender, 70% (7/10) of the studies analyzing eHealth portals included this variable [32,34,38-41,48]. Findings were mixed. Some studies found no significant differences [32,34,39,40], while others reported higher usage among women [48], and greater overall use by women [38].

Regarding socioeconomic factors, 70% (7/10) of the studies [34,35,39,40,43,48,49] examined associations between portal use and indicators such as education, income, social class, employment, and area-based deprivation. Higher education [39,40,42,45] and less deprivation [38,41,49] were consistently associated with greater use (eg, postgraduate adjusted odds ratio [aOR] 2.38, 95% CI 1.42 - 4.02) [40].

Employment status was analyzed in 30% (3/10) of studies, showing inconsistent results [32,39,40]: only one study found that being used was related to lower use (OR 0.57, 95% CI 0.38 - 0.84) [45]. The other two reported no associations [32,39].

Household composition, including size of household and marital status, was found as not associated in 2 of the 3 studies that analyzed it (2/10) [39,40]. However, Pierce et al [35] tied higher parity to lower use.

Cultural background—analyzed in 5 [34,35,40,48,49] of the 10 portal-focused studies (50%) [32,34,35,38-41,43,48,49]—showed no strong effects, with no significant differences in 2 studies [34,40]. Pierce et al [35] reported that 48.7% of nonusers were non-native compared to 24.4 % of those with high engagement; Kc et al [48] observed higher use in White patients [48]; Zhang [49] noted higher Asian patient use postadjustment.

XSL•FO
RenderX

In terms of area of residence, 2 [39,49] of the 3 (3/10) studies [39,40,49] that analyzed it found higher use among urban areas (OR 1.23, 95% CI 1.02 - 1.48) [39] and activation [49]. Knöchelmann et al [40] found no differences.

Health status was examined in 40% (4/10) of the studies [32,34,48,49], with mixed associations: some studies reported lower use among individuals with long-term conditions [48,49], while others found higher use among those with poor health [34] or no significant differences [32].

Finally, digital competence was analyzed in 30% (3/10) of the studies, where it was described as a strong predictor [32,34,41]: (OR 8.82, 95% CI 6.15 - 12.64) [41] and (aOR 2.96, 95% CI 2.02 - 4.35) for high literacy [34]. Lower health literacy was associated with lower use [32].

Electronic Health Records (EHRs)

Among the 33% (6/18) of studies [37,41,42,46-48] that analyzed EHR use, age was examined in 5, and all of them found that younger adults were more likely to use these tools [37,41,42,46,48]. Engagement declined in older adults, especially ≥75 years, with reduced registration (aOR 0.40, 95% CI 0.36 - 0.44) and activation (aOR 0.39, 95% CI 0.32 - 0.47) [46].

Gender was assessed in all 6 studies [37,41,42,46-48]. Four (67%) found higher EHR use among women [37,41,46-48]. In contrast, England reported higher male registration into the tool (aOR 1.10, 95% CI 1.04 - 1.16) but lower activation of it (aOR 0.85, 95% CI 0.78 - 0.94) [46], which was necessary for usage. González-Cacheda et al [42] found no differences.

Socioeconomic status was examined in 67% (4/6) of the studies [37,42,46,48]. All confirmed that higher levels of education increased use [37,42], and deprivation showed a clear gradient: individuals in more affluent areas were more likely to register and activate the tool [46,48].

Cultural background, assessed in 50% (3/6) of the studies [42,46,48], predicted lower activation in Asian (aOR 0.61, 95% CI 0.53 - 0.71), Black (aOR 0.45, 95% CI 0.36 - 0.56), and mixed (aOR 0.77, 95% CI 0.60 - 0.97) groups compared to White [46]. Evidence comes primarily from a large cross-sectional study conducted in England (n=28,637) [46], complemented by national datasets from England [48] and Spain [42]. Language barriers reduced use as no interpreter need increased registration (aOR 1.63, 95% CI 1.33 - 1.99) and activation (aOR 3.16, 95% CI 1.96 - 5.09) [46]. National [42] and White patients [48] had higher use.

Only one study (17%, 1/6) [37], a Swedish survey of over 3000 adults, analyzed marital status, reporting that being in a partnership was linked to higher engagement [37].

Health status findings showed mixed results among the 50% of studies (3/6): one study found no differences [37], another observed lower use in long-term conditions [48], while Chapman reported higher registration among patients with complex conditions (aOR 2.54, 95% CI 2.30 - 2.82) [46].

Finally, both studies assessing digital competence were conducted in Finland (n=4495) [41] and Spain (n=7454) [42], using large national datasets, and both found that lower digital skills were associated with reduced use of EHRs [41,42].

Telemedicine and Remote Primary Care

Among the 22% (4/18) of studies [33,36,44,45] that analyzed telemedicine and remote primary care, 100% (4/4) examined age and consistently identified younger adults as the main users, while older adults showed lower use [33,36,44,45]. In Sweden, those aged 60 - 79 years had substantially lower use (IRR [incidence rate ratio [0.45], 95% CI 0.40 - 0.50) and those ≥80 years even lower (IRR 0.19, 95% CI 0.16 - 0.25) [44].

Gender was assessed in 100% (4/4) of the studies analyzing telemedicine and remote primary care [33,36,44,45], with 75% (3/4) reporting higher usage among women [33,44,45]. In Sweden, women accounted for 65.4% of users (IRR 1.39, 95% CI 1.28 - 1.51) [44], while one study found no gender difference [36].

Socioeconomic factors were assessed in 100% (4/4) of the studies [33,36,44,45]. Higher education was associated with greater use (OR 1.41, 95% CI 1.19 - 1.67) [36], with postsecondary users showing the highest engagement (7.2%; OR 1.51, 95% CI 1.74 - 1.56) [33]. More users were observed among individuals with higher incomes (7.7% vs 4.0%; OR 2.46, 95% CI 2.38 - 2.54) [33]. Employment was also linked to usage: 85.7% of users were employed versus 14.3% unemployed [45], while one study reported lower odds among retirees (OR 0.68, 95% CI 0.46 - 1.02) [36]. One study found no differences across social class [44].

Cultural background and origin were assessed in 50% (2/4) of the studies, both of which were based on large population-based datasets [33,45]. Both studies reported lower uptake among foreign-born users compared to natives (3.2% vs 7.2%; OR 0.53, 95% CI 0.52 - 0.55) [33].

Urban residency was analyzed in 2 [33,45] of 4 studies [33,36,44,45]. One study associated it with higher use [45], while residing ≥10 km from health care services was significantly associated with higher telemedicine use (7.1%; OR≤1.11) [33].

Health status was analyzed in 100% (4/4) of the studies [33,36,44,45], showing mixed results. Better health predicted higher use (OR 1.31, 95% CI 1.01 - 1.69) [36]. In contrast, heart failure was associated with lower use (OR 0.49, 95% CI 0.40 - 0.59) [33]. Some conditions, such as depression, asthma, and chronic obstructive pulmonary disease, were linked to higher use, with greater morbidity also associated with increased engagement (IRR 2.67, 95% CI 2.02 - 3.54) [44,45].

Online Appointment Booking

Inequalities in online appointment booking were examined in 11% (2/18) of the studies [41,48], both showing that younger adults were more likely to use these systems. In Finland, use declined sharply with age [41], while in England, practices with more patients aged 15 - 34 years had slightly higher booking rates (+1.35%) [48]. Gender differences were assessed in 100% (2/2) of the studies. Both studies reported lower online booking among men [41,48].

XSL•FO
RenderX

Socioeconomic and cultural factors were analyzed in 50% (1/2) of the studies, revealing lower booking rates in practices located in more deprived areas, as well as reduced uptake among non-White populations [48].

Health status was assessed in 50% (1/2) of the studies, showing that patients with long-term conditions booked less frequently (–0.77%) [48].

Finally, digital competence was evaluated in 50% (1/2) of the studies and was strongly associated with use; patients with good digital skills were over 10 times more likely to book online (OR 10.91, 95% CI 7.24 - 16.44) [41].

Discussion

Principal Findings

Although European public health care systems are characterized by universal health care coverage, this systematic review aimed to examine inequalities in access to and use of eHealth tools across social determinants of health. The findings reveal persistent and significant social inequalities, with lower usage consistently reported among older adults, individuals with lower socioeconomic or educational status, rural populations, ethnic minorities, people with poor health, and those with limited digital competence. Findings regarding gender were more mixed, with some studies reporting slightly lower usage among men. The majority of included studies were methodologically robust, with most rated high quality and a few moderate quality according to JBI appraisal and conducted with widely varying sample sizes, which should be considered when interpreting these patterns and their potential translation into real-world health care practice. eHealth portals were the most investigated tool, but EHRs and telemedicine showed more consistent evidence of social gradients. Online appointment booking was the least studied category, limiting the strength of conclusions in this area.

Interpretation of Key Findings

Although digitalization in health care has been presented as improving access and system efficiency [1,2,50], our findings reveal a generational gap across all tool types [32,33,35-39,41,42,44-46,48,49]. Lower usage among older adults reflects not only differences in digital competence—such as less experience and lower technological confidence [51,52]—but also a lack of service adaptation to their needs and potential age-related cognitive barriers that limit effectiveness [53]. These observations, derived mostly from high-quality studies, strengthen the reliability of the age-related patterns observed. They were consistently found both in medium-sized samples [37] and in large administrative datasets [33,44], reinforcing the robustness of these patterns across study designs. At the same time, heterogeneity in measurement and study scale reduces the precision of effect estimates, particularly in smaller studies. In addition, unintuitive interfaces, low perceived usefulness, and a preference for in-person care further contribute to lower uptake [54]. While family support may partially mitigate this gap [54,55], findings are inconsistent [40,56]. This raises concerns about reduced autonomy, informal caregiver burden—often falling on women—and the exclusion of those without social support.

Gender-related patterns were less consistent. Some studies showed higher usage among women for EHRs and telemedicine, while differences in other services were weaker, albeit with a general trend toward greater women's engagement. These mixed results may reflect contextual factors such as national health care organization, cultural norms, and sample composition. Some null findings came from smaller or moderate samples [32,40], whereas large studies such as Dahlgren et al [33] consistently showed gender differences. Greater women's involvement may be interpreted in light of traditional gender roles, which assign women the responsibility for caregiving, reflected in a higher willingness to seek care, familiarity with health services, and confidence in sustained, active communication [57] as well as in more continuous use, as suggested by one of the included studies [46]. However, previous literature shows that in certain contexts, the gender gap is reversed, possibly due to broader disparities in digital engagement. Women tend to use a more limited range of digital services and engage with them less frequently and intensively than men, reflecting differences in digital familiarity and confidence [58,59].

It is well established that the socioeconomic gradient not only determines access to material resources but also shapes individuals' ability to navigate digital environments. In our review, educational attainment, income level, and residence in disadvantaged areas were systematically associated with lower use of digital health tools, in line with studies from other digitalizing systems [60]. Evidence regarding employment status was less consistent, with some studies suggesting lower use among unemployed individuals. Beyond material barriers (access to internet or digital devices), differences in digital literacy, confidence, and prior experience—the so-called third-level digital divide—further perpetuated exclusion [61,62]. Although education strongly predicted competences, structural conditions such as local digital infrastructure and social position also shaped digital engagement. In disadvantaged regions, limited digital services may affect both patients and health care professionals, potentially reinforcing existing health care disparities [63]. Given the overall rigor of the included studies, these findings appear reliable; however, their transferability to different health care contexts should be approached with caution. The WHO stresses that closing this gap requires not only investment in infrastructure, but also training and support strategies specifically targeted at the most vulnerable groups [64]. The consistency of socioeconomic gradients across both small samples [34] and very large datasets [33] strengthens confidence in these associations and suggests that they are not driven by sample-specific variability.

Beyond socioeconomic status, ethnic minorities have shown lower engagement with all the tools analyzed, even after adjusting for socioeconomic variables. Language barriers, lack of familiarity with the health care system, distrust of institutional platforms, and poor cultural adaptation of services emerge as recurring obstacles [46,65,66].

XSL•FO
RenderX

Although urban environments are generally associated with higher use of eHealth portals [39,49], one study showed greater use of video consultations in rural areas [33], consistent with previous literature [67], which suggests that digitalization compensates for limited in-person services. This pattern, particularly in studies with careful design and with large samples such as Dahlgren et al [33], suggests that geographical effects are not artifacts of small-sample instability. However, rural areas often have lower educational attainment, which can reduce digital competence [42,45]. These findings indicate that geography can both mitigate and amplify inequalities, depending on the system's capacity to adapt.

Regarding health status, some studies showed lower use among people with multimorbidity, in line with previous literature [68,69], reflecting the Inverse Care Law, where populations with greater health care needs often face reduced access [70,71]. However, other studies found higher use among people with poorer health, likely due to more frequent contact with the health care system [40,48]. These mixed results are partly attributable to methodological differences and varying study quality, with patterns more clearly observed in studies with comprehensive data and stronger design [33,37]. Variation in sample sizes also contributed: larger datasets such as Eriksson et al [44] and Dahlgren et al (N≈2 million) [33] produced more stable health-related estimates than smaller clinical samples, helping explain discrepancies in the direction and magnitude of associations. Overall, individuals with greater clinical needs may face additional barriers when poor health intersects with low education or socioeconomic status.

Although the overall methodological quality of the included studies was acceptable—with most rated high quality and a few moderate quality according to the JBI appraisal—the predominance of cross-sectional designs limits causal inference. Substantial heterogeneity in study populations, digital tools, and outcome indicators also affects comparability across studies. Furthermore, the wide variability in sample sizes introduces important differences in estimate precision, which should be taken into account when interpreting inconsistent findings across determinants. These methodological and contextual differences influence how findings can be translated into real-world practice: variations in health care organization, digital infrastructure, and population characteristics across countries may alter the magnitude and direction of observed inequalities. Therefore, interpretation and policy translation should be undertaken with caution, recognizing that local context strongly shapes the impact of the digital divide.

Limitations and Strengths

This systematic review has several limitations. One limitation is that the search strategy was not developed in collaboration with an information scientist, which might have influenced the sensitivity and comprehensiveness of the retrieved records. The conclusions may also be affected by differences in study design, sample representativeness, and reporting. Although most studies were rated as high quality according to the JBI appraisal, some moderate-quality designs introduce a potential risk of bias that should be considered when interpreting the results. The exclusion of qualitative studies—although methodologically justified to ensure comparability—may have constrained the review's ability to capture rich contextual insights, such as user experiences, perceived barriers, and cultural or behavioral factors influencing digital health use, which are crucial for understanding the mechanisms behind observed inequalities. Heterogeneity in populations, definitions of eHealth use, and measurement of social determinants limited comparability and prevented meta-analysis, requiring a narrative synthesis. Restricting the search to peer-reviewed publications in English or Spanish may have introduced language and publication bias, excluding relevant evidence published in other languages or in gray literature. Nevertheless, this criterion enhances replicability and scientific robustness. The focus on public health care systems in Europe restricts generalizability to other regions. Additionally, the studies' highly variable sample sizes, while not undermining the findings, do influence precision and should be considered when weighing the strength of evidence across determinants.

Despite these limitations, the review has several strengths. It followed PRISMA guidelines, used a comprehensive search across 4 major databases, supplemented by backward citation searching, and included independent screening and data extraction by multiple reviewers to improve validity. By synthesizing evidence across diverse European settings, this review offers an up-to-date and policy-relevant overview of eHealth-related inequalities, and its classification by type of digital service enables a more granular understanding of patterns of exclusion. No amendments were made to the registered protocol after PROSPERO registration.

These findings carry important implications for real-world health care policy and practice. By highlighting persistent inequalities in access and use of digital health tools across multiple social determinants, the review underscores the need for targeted interventions, including digital literacy programs, inclusive eHealth platform design, and support strategies for older adults, ethnic minorities, and socioeconomically disadvantaged populations. The innovative contribution of this review lies in its focus on quantitative evidence from European public health care systems, integrating results across diverse social determinants and types of eHealth tools. This provides a policy-relevant, evidence-based perspective to guide equitable digital transformation and foster inclusive health systems.

Future Research Directions

Future research should explore the long-term effects of digital exclusion on health outcomes, assess the equity impact of digital health policies, and promote participatory research with underserved communities to co-develop inclusive digital solutions. There is a clear need for intersectional frameworks to understand how intersecting inequalities shape digital health access, especially in highly vulnerable populations. Additionally, research is scarce in low-income countries, where sociodemographic and economic contexts may lead to different and sometimes contrasting patterns, for example regarding gender. Further studies are also needed to evaluate the actual impact of digital health tools on access to health care services more broadly—not only their uptake or use but also how they affect entry into the health system and care continuity.

XSL•FO
RenderX

Moreover, more attention should be paid to the underlying mechanisms driving these inequalities, to move beyond description and identify structural, institutional, or behavioral factors that generate unequal access. Finally, while the number of studies exploring social disparities in digital health access has increased, there remains a gap in longitudinal and in-depth research that can capture the complexity and dynamics of these inequities over time.

Conclusions

This systematic review shows that digital transformation in European public health systems is not equally benefiting all population groups. Across the studies analyzed, persistent social inequalities in the use of digital health tools—particularly EHRs, eHealth portals, and telemedicine and remote primary care—were observed. The most excluded from digital care services are older adults, individuals with lower socioeconomic or educational status, ethnic minorities, rural populations, people in poor health, and those with limited digital competence.

Rather than closing gaps, digitalization may replicate or deepen existing health inequalities if equity is not explicitly addressed. Ensuring inclusive design, supporting digital literacy, and engaging underserved communities are essential steps to make digital health a tool for reducing—not reinforcing—structural exclusion. Although a few studies explicitly explored interactions between determinants, the detected patterns suggest that the combination of factors such as advanced age, low socioeconomic status, migrant background, and limited digital competence tends to produce cumulative effects that significantly limit the adoption of eHealth tools. In interpreting these findings, both the heterogeneity of studies and their risk of bias should be taken into account. While these findings offer a policy-relevant perspective, the heterogeneity of study designs, populations, and outcomes, as well as the risk of bias, limits direct translation into policy and practice. By synthesizing quantitative evidence across diverse social determinants and types of digital health tools, this review not only highlights persistent inequalities but also offers a policy-relevant perspective. These findings can inform equity-oriented interventions, guide inclusive eHealth design, and support digital literacy programs, helping policymakers and health systems ensure that digital transformation benefits all population groups.

Acknowledgments

The authors gratefully acknowledge the support of the “Instituto de Salud Carlos III” and the European Union, which co-funded this study under the 2023 call for Health R&D Projects.

Generative artificial intelligence tools were used solely to assist with the review and refinement of the English language in the manuscript. No artificial intelligence tools were used for content generation, data analysis, or interpretation.

Funding

This study was funded by the “Instituto de Salud Carlos III” and co-funded by the European Union (Grant No. PI23/01063), within the framework of the 2023 call for Health R&D Projects. The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during this study. All data supporting the findings of this study are derived from previously published articles, which are fully referenced in the manuscript. The complete search strategies are provided in Multimedia Appendix 1, and summary data extraction tables are available in Multimedia Appendix 4. Additional extracted datasets (eg, the complete data extraction matrix) can be made available from the corresponding author upon reasonable request.

Authors' Contributions

GMG, AAG, and UM conceptualized the study. GMG, AAG, and MFL designed the data search strategy and conducted data extraction. GMG drafted the initial version of the manuscript in close collaboration with AAG. MFL revised multiple versions of the manuscript, contributing substantially to the content and wording. UM and EV reviewed the final versions of the manuscript and provided critical input to the drafting process. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete search strategies.
[DOCX File, 16 KB - jmir_v28i1e81841_app1.docx]

Multimedia Appendix 2
Study-specific definitions of determinants and eHealth use.

XSL•FO
RenderX

[XLSX File, 11 KB - jmir_v28i1e81841_app2.xlsx]

Multimedia Appendix 3
Risk of bias assessment of each study.
[DOCX File, 24 KB - jmir_v28i1e81841_app3.docx]

Multimedia Appendix 4
Extended evidence table summarizing the main information of each study included in the systematic review.
[DOCX File, 113 KB - jmir_v28i1e81841_app4.docx]

Checklist 1
PRISMA checklist.
[DOCX File, 282 KB - jmir_v28i1e81841_app5.docx]

References

1. Oh H, Rizo C, Enkin M, Jadad A, Powell J, Pagliari C. What is eHealth (3): a systematic review of published definitions. J Med Internet Res 2005 Feb 24;7(1):e1. [doi: 10.2196/jmir.7.1.e1] [Medline: 15829471]
2. Global digital health monitor. World Health Organization. 2023. URL: https://data.who.int/dashboards/gdhm/overview [accessed 2025-08-03]
3. Resolution WHA71.7: digital health. World Health Organization. 2018. URL: https://www.who.int/publications/i/item/10665-279505 [accessed 2025-12-11]
4. Special report 25/2024: digitalisation of healthcare:EU support for member states effective overall, but difficulties in using EU funds. : European Court of Auditors; 2024 URL: https://www.eca.europa.eu/ECAPublications/SR-2024-25/SR-2024-25_EN.pdf [accessed 2026-01-14]
5. Jonsson M, Johansson S, Hussain D, Gulliksen J, Gustavsson C. Development and evaluation of eHealth services regarding accessibility: scoping literature review. J Med Internet Res 2023 Aug 17;25:e45118. [doi: 10.2196/45118] [Medline: 37590050]
6. Atherton H, Eccles A, Poltawski L, Dale J, Campbell J, Abel G. Investigating patient use and experience of online appointment booking in primary care: mixed methods study. J Med Internet Res 2024 Jul 8;26:e51931. [doi: 10.2196/51931] [Medline: 38976870]
7. Giebel GD, Speckemeier C, Abels C, et al. Problems and barriers related to the use of digital health applications: scoping review. J Med Internet Res 2023 May 12;25:e43808. [doi: 10.2196/43808] [Medline: 37171838]
8. EGovernment benchmark 2022. synchronising digital governments.INSIGHT REPORT. : Publications Office of the European Union; 2022 URL: https://op.europa.eu/en/publication-detail/-/publication/cc953c44-1468-11ed-8fa0-01aa75ed71a1/language-en [accessed 2026-01-14]
9. Abernethy A, Adams L, Barrett M, et al. The promise of digital health: then, now, and the future. NAM Perspect 2022;2022(22). [doi: 10.31478/202206e] [Medline: 36177208]
10. Global Strategy on Digital Health 2020-2025: World Health Organization; 2021.
11. Gobburi RK, Olawade DB, Olatunji GD, Kokori E, Aderinto N, David-Olawade AC. Telemedicine use in rural areas of the United Kingdom to improve access to healthcare facilities: a review of current evidence. Informatics Health 2025 Mar;2(1):41-48. [doi: 10.1016/j.infoh.2025.01.003]
12. Western MJ, Smit ES, Gültzow T, et al. Bridging the digital health divide: a narrative review of the causes, implications, and solutions for digital health inequalities. Health Psychol Behav Med 2025;13(1):2493139. [doi: 10.1080/21642850.2025.2493139] [Medline: 40276490]
13. Western MJ, Armstrong MEG, Islam I, Morgan K, Jones UF, Kelson MJ. The effectiveness of digital interventions for increasing physical activity in individuals of low socioeconomic status: a systematic review and meta-analysis. Int J Behav Nutr Phys Act 2021 Nov 9;18(1):148. [doi: 10.1186/s12966-021-01218-4] [Medline: 34753490]
14. Szinay D, Forbes CC, Busse H, DeSmet A, Smit ES, König LM. Is the uptake, engagement, and effectiveness of exclusively mobile interventions for the promotion of weight-related behaviors equal for all? A systematic review. Obes Rev 2023 Mar;24(3):e13542. [doi: 10.1111/obr.13542] [Medline: 36625062]
15. Pettersson L, Johansson S, Demmelmaier I, Gustavsson C. Disability digital divide: survey of accessibility of eHealth services as perceived by people with and without impairment. BMC Public Health 2023 Jan 27;23(1):181. [doi: 10.1186/s12889-023-15094-z] [Medline: 36707791]
16. Chae J. A comprehensive profile of those who have health-related apps. Health Educ Behav 2018 Aug;45(4):591-598. [doi: 10.1177/1090198117752784] [Medline: 29353544]
17. Ratcliff CL, Krakow M, Greenberg-Worisek A, Hesse BW. Digital health engagement in the US population: insights from the 2018 Health Information National Trends Survey. Am J Public Health 2021 Jul;111(7):1348-1351. [doi: 10.2105/AJPH.2021.306282] [Medline: 34014759]

XSL•FO
RenderX

18. Hegeman PC, Vader DT, Kamke K, El-Toukhy S. Patterns of digital health access and use among US adults: a latent class analysis. BMC Digit Health 2024;2(1):42. [doi: 10.1186/s44247-024-00100-0] [Medline: 39544227]
19. Jahnel T, Dassow HH, Gerhardus A, Schüz B. The digital rainbow: digital determinants of health inequities. Digit Health 2022;8:20552076221129093. [doi: 10.1177/20552076221129093] [Medline: 36204706]
20. Torrent-Sellens J, Díaz-Chao Á, Soler-Ramos I, Saigí-Rubió F. Modelling and predicting eHealth usage in Europe: a multidimensional approach from an online survey of 13,000 European Union internet users. J Med Internet Res 2016 Jul 22;18(7):e188. [doi: 10.2196/jmir.5605] [Medline: 27450189]
21. Wong BLH, Maaß L, Vodden A, et al. The dawn of digital public health in Europe: implications for public health policy and practice. Lancet Reg Health Eur 2022 Mar;14:100316. [doi: 10.1016/j.lanepe.2022.100316] [Medline: 35132399]
22. Wilson S, Tolley C, McArdle R, Hassan N, Slight R, Slight S. Strategies to advance digital health equity: a qualitative study with underserved groups. Lancet 2024 Nov;404:S37. [doi: 10.1016/S0140-6736(24)02004-X]
23. Woolley KE, Bright D, Ayres T, Morgan F, Little K, Davies AR. Mapping inequities in digital health technology within the World Health Organization's European Region using PROGRESS PLUS: scoping review. J Med Internet Res 2023 Apr 28;25:e44181. [doi: 10.2196/44181] [Medline: 37115613]
24. Digital health not accessible by everyone equally, new study finds. World Health Organization. 2022 Dec 21. URL: https://www.who.int/europe/news/item/21-12-2022-digital-health-not-accessible-by-everyone-equally-new-study-finds [accessed 2025-12-11]
25. Organisation for Economic Co-Operation and Development. Health at a Glance 2023: OECD Indicators: OECD Publishing; 2023. [doi: 10.1787/4dd50c09-en]
26. European Commission. Case study on the digitalisation of health (ehealth). : Luxembourg: Publications Office of the European Union; 2024 URL: https://commission.europa.eu/document/download/652a3175-d410-4608-8f0e-642049433c35_en [accessed 2025-06-26]
27. Page MJ, McKenzie JE, Bossuyt PM, et al. PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
28. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
29. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 5;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
30. Ma LL, Wang YY, Yang ZH, Huang D, Weng H, Zeng XT. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? Military Med Res 2020 Dec;7(1):7. [doi: 10.1186/s40779-020-00238-8]
31. Xu Y, Wang M, Chen D, Jiang X, Xiong Z. Inflammatory biomarkers in older adults with frailty: a systematic review and meta-analysis of cross-sectional studies. Aging Clin Exp Res 2022 May;34(5):971-987. [doi: 10.1007/s40520-021-02022-7] [Medline: 34981430]
32. Hörhammer I, Suvanto J, Kinnunen M, Kujala S. Usefulness of self-guided digital services among mental health patients: the role of health confidence and sociodemographic characteristics. Int J Med Inform 2025 Feb;194:105693. [doi: 10.1016/j.ijmedinf.2024.105693] [Medline: 39556968]
33. Dahlgren C, Dackehag M, Wändell P, Rehnberg C. Determinants for use of direct-to-consumer telemedicine consultations in primary healthcare-a registry based total population study from Stockholm, Sweden. BMC Fam Pract 2021 Jun 26;22(1):133. [doi: 10.1186/s12875-021-01481-1] [Medline: 34172009]
34. Neves AL, Smalley KR, Freise L, Harrison P, Darzi A, Mayer EK. Determinants of use of the care information exchange portal: cross-sectional study. J Med Internet Res 2021 Nov 11;23(11):e23481. [doi: 10.2196/23481] [Medline: 34762063]
35. Pierce P, Whitten M, Hillman S. The impact of digital healthcare on vulnerable pregnant women: a review of the use of the MyCare app in the maternity department at a central London tertiary unit. Front Digit Health 2023;5:1155708. [doi: 10.3389/fdgth.2023.1155708] [Medline: 37153515]
36. Söderberg D, Bonn SE, Sjöblom L, et al. Individual patient factors associated with the use of physical or digital primary care in Sweden. Telemed J E Health 2024 May;30(5):1289-1296. [doi: 10.1089/tmj.2023.0351] [Medline: 38394275]
37. Muli I, Cajander Å, Hvitfeldt H, et al. To read or not to read - a cross-sectional study of Swedish primary care patients' adoption of patient accessible electronic health records. Digit Health 2024;10:20552076241287636. [doi: 10.1177/20552076241287636] [Medline: 39381824]
38. Pálsdóttir Á. Development in the adoption of a national digital healthcare system and experience at taking new technology in use – changes from 2019 to 2022. Inf Res 2024;29(2):464-482. [doi: 10.47989/ir292841]
39. Merkel S, Hess M. The use of internet-based health and care services by elderly people in Europe and the importance of the country context: multilevel study. JMIR Aging 2020 Jun 3;3(1):e15491. [doi: 10.2196/15491] [Medline: 32490837]
40. Knöchelmann A, Healy K, Frese T, et al. User profiles in digitalized healthcare: active, potential, and rejecting — a cross-sectional study using latent class analysis. BMC Health Serv Res 2024;24(1):1083. [doi: 10.1186/s12913-024-11523-w]
41. Heponiemi T, Kaihlanen AM, Kouvonen A, Leemann L, Taipale S, Gluschkoff K. The role of age and digital competence on the use of online health and social care services: a cross-sectional population-based survey. Digit Health 2022;8:20552076221074485. [doi: 10.1177/20552076221074485] [Medline: 35111333]

XSL•FO
RenderX

42. González-Cacheda B, Outeda CC. Understanding attitudes, knowledge, and use of e-health services in the health system in Spain. J Public Health Policy 2025 Jun;46(2):295-302. [doi: 10.1057/s41271-025-00549-x] [Medline: 39827313]
43. Petersen LS, Bertelsen P. Equality challenges in the use of eHealth: selected results from a Danish citizens survey. Stud Health Technol Inform 2017;245:793-797. [Medline: 29295207]
44. Eriksson P, Randjelovic M, Thulesius H, Hammar T, Lagrosen S, Nilsson E. Differences in use of telemedicine integrated into traditional primary health care - a comparative observational study. Scand J Prim Health Care 2025 Jun;43(2):476-487. [doi: 10.1080/02813432.2025.2457542] [Medline: 39915941]
45. Wilkens J, Thulesius H, Ekman B. From office to digital primary care services: analysing income-related inequalities in utilization. Int J Equity Health 2024 Apr 30;23(1):86. [doi: 10.1186/s12939-024-02184-6] [Medline: 38689241]
46. Chapman R, Haroon S, Simms-Williams N, et al. Socioeconomic deprivation, age and language are barriers to accessing personal health records: a cross-sectional study of a large hospital-based personal health record system. BMJ Open 2022 Jan 20;12(1):e054655. [doi: 10.1136/bmjopen-2021-054655] [Medline: 35058264]
47. Kharko A, Hagström J, Simola S, Cajander Å, Blease C, Hägglund M. The gender gap in EHR experiences and preferences: results from an international cross-sectional survey of patients. Stud Health Technol Inform 2025 May 15;327:939-943. [doi: 10.3233/SHTI250510] [Medline: 40380617]
48. Kc S, Papoutsi C, Reidy C, et al. Differences in use of a patient portal across sociodemographic groups: observational study of the NHS app in England. J Med Internet Res 2024 Nov 13;26:e56320. [doi: 10.2196/56320] [Medline: 39536310]
49. Zhang J, Gallifant J, Pierce RL, et al. Quantifying digital health inequality across a national healthcare system. BMJ Health Care Inform 2023 Nov 24;30(1):e100809. [doi: 10.1136/bmjhci-2023-100809] [Medline: 38007224]
50. Al-Shorbaji N. Improving healthcare access through digital health: the use of information and communication technologies. In: Healthcare Access: IntechOpen; 2022. [doi: 10.5772/intechopen.99607]
51. Yusif S, Soar J, Hafeez-Baig A. Older people, assistive technologies, and the barriers to adoption: a systematic review. Int J Med Inform 2016 Oct;94:112-116. [doi: 10.1016/j.ijmedinf.2016.07.004] [Medline: 27573318]
52. Klaver NS, van de Klundert J, van den Broek R, Askari M. Relationship between perceived risks of using mHealth applications and the intention to use them among older adults in the Netherlands: cross-sectional study. JMIR Mhealth Uhealth 2021 Aug 30;9(8):e26845. [doi: 10.2196/26845] [Medline: 34459745]
53. Çöme O, Gökdemir Ö, Bilik Sezer B, Kasapoğlu SS, Kjær NK, Güldal D. The interplay between cognitive function and digital health literacy among older adults: implications for e-health equity and accessibility. Int J Med Inform 2025 Sep;201:105934. [doi: 10.1016/j.ijmedinf.2025.105934] [Medline: 40273597]
54. Wildenbos GA, Jaspers MWM, Schijven MP, Dusseljee-Peute LW. Mobile health for older adult patients: using an aging barriers framework to classify usability problems. Int J Med Inform 2019 Apr;124:68-77. [doi: 10.1016/j.ijmedinf.2019.01.006] [Medline: 30784429]
55. Schroeder T, Dodds L, Georgiou A, Gewald H, Siette J. Older adults and new technology: mapping review of the factors associated with older adults' intention to adopt digital technologies. JMIR Aging 2023 May 16;6:e44564. [doi: 10.2196/44564] [Medline: 37191976]
56. Garcia Reyes EP, Kelly R, Buchanan G, Waycott J. Understanding older adults' experiences with technologies for health self-management: interview study. JMIR Aging 2023 Mar 21;6:e43197. [doi: 10.2196/43197] [Medline: 36943333]
57. Jiménez Ruiz I, Moya Nicolás M. The family caregiver: the naturalized sense of obligation in women to be caregivers {Article in Spanish]. Enfermería Global . 2017;17(1):420. [doi: 10.6018/eglobal.17.1.292331]
58. Tyers-Chowdhury A, Binder G. What we know about the gender digital divide for girls: a literature review. : United Nations Children's Fund (UNICEF); 2021 URL: https://www.unicef.org/eap/media/8311/file/What%20we%20know%20about%20the%20gender%20digital%20divide%20for%20girls:%20a%20literature%20review.pdf [accessed 2025-01-12]
59. Ek S. Gender differences in health information behaviour: a Finnish population-based survey. Health Promot Int 2015 Sep;30(3):736-745. [doi: 10.1093/heapro/dat063] [Medline: 23985248]
60. Standaar L, van Tuyl L, Suijkerbuijk A, Brabers A, Friele R. Differences in eHealth access, use, and perceived benefit between different socioeconomic groups in the Dutch context: secondary cross-sectional study. JMIR Form Res 2025 Jan 7;9:e49585. [doi: 10.2196/49585] [Medline: 39773883]
61. van Deursen A, Helsper EJ. The third-level digital divide: who benefits most from being online. In: Communication and Information Technologies Annual: Emerald Group Publishing; 2015:29-52. [doi: 10.1108/S2050-206020150000010002]
62. Milanti A, Chan DNS, Parut AA, So WKW. Determinants and outcomes of eHealth literacy in healthy adults: a systematic review. PLoS One 2023;18(10):e0291229. [doi: 10.1371/journal.pone.0291229] [Medline: 37792773]
63. Alam K, Mahumud RA, Alam F, Keramat SA, Erdiaw-Kwasie MO, Sarker AR. Determinants of access to eHealth services in regional Australia. Int J Med Inform 2019 Nov;131:103960. [doi: 10.1016/j.ijmedinf.2019.103960] [Medline: 31518858]
64. Equity within digital health technology within the WHO european region: a scoping review. World Health Organization. 2022. URL: https://www.who.int/europe/publications/i/item/WHO-EURO-2022-6810-46576-67595 [accessed 2025-08-03]
65. Bergman L, Nilsson U, Dahlberg K, Jaensson M, Wångdahl J. Health literacy and e-health literacy among Arabic-speaking migrants in Sweden: a cross-sectional study. BMC Public Health 2021 Nov 25;21(1):2165. [doi: 10.1186/s12889-021-12187-5] [Medline: 34823499]

XSL•FO
RenderX

66. Whitehead L, Talevski J, Fatehi F, Beauchamp A. Barriers to and facilitators of digital health among culturally and linguistically diverse populations: qualitative systematic review. J Med Internet Res 2023 Feb 28;25:e42719. [doi: 10.2196/42719] [Medline: 36853742]
67. Ratna S, Behera KL, Renuka JR. Exploring healthcare professionals' perceptions and practices on telemedicine adoption in rural regions. Semin Med Writ Educ 2023;2:139. [doi: 10.56294/mw2023139]
68. Manning SE, Wang H, Dwibedi N, et al. Association of multimorbidity with the use of health information technology. Digit Health 2023;9:20552076231163797. [doi: 10.1177/20552076231163797] [Medline: 37124332]
69. Runz-Jørgensen SM, Schiøtz ML, Christensen U. Perceived value of eHealth among people living with multimorbidity: a qualitative study. J Comorb 2017;7(1):96-111. [doi: 10.15256/joc.2017.7.98] [Medline: 29359124]
70. Pathirana TI, Jackson CA. Socioeconomic status and multimorbidity: a systematic review and meta-analysis. Aust N Z J Public Health 2018 Apr;42(2):186-194. [doi: 10.1111/1753-6405.12762] [Medline: 29442409]
71. Tudor Hart J. The inverse care law. Lancet 1971 Feb;297(7696):405-412. [doi: 10.1016/S0140-6736(71)92410-X]

Abbreviations

EHR: electronic health record
JBI: Joanna Briggs Institute
OR: odds ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
WHO: World Health Organization

Edited by A Mavragani, S Brini; submitted 05.Aug.2025; peer-reviewed by H Maheshwari, K Triep; revised version received 24.Nov.2025; accepted 24.Nov.2025; published 09.Feb.2026.

Please cite as:
Monasterio G, Fernández-López MJ, Valero E, Martin U, Ayala-García A
Digital Inequalities in the Use of eHealth Services in European Public Health Care Systems: Systematic Review of Observational Studies
J Med Internet Res 2026;28:e81841
URL: https://www.jmir.org/2026/1/e81841
doi:10.2196/81841

© Garazi Monasterio, Marcos José Fernández-López, Erika Valero, Unai Martin, Amaia Ayala-García. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Data-Driven Guideline Adherence in Data Representation and Compliance Measurement: Scoping Review

Minh Trang Hoang[1], BMedSci (Hon); Candice Donnelly[1], PhD; Christina Igasto[1,2], MSc, PhD; Amith Shetty[1,3,4], MBBS, FACEM, PhD; Malcolm Pradhan[1], MBBS, FACHI, PhD; Tim Shaw[1], PhD

[1]Biomedical Informatics and Digital Health, School of Medical Sciences, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia
[2]Service NSW, Sydney, Australia
[3]NSW Ministry of Health, Sydney, Australia
[4]Western Sydney Local Health District, Westmead, Australia

Corresponding Author:
Minh Trang Hoang, BMedSci (Hon)
Biomedical Informatics and Digital Health, School of Medical Sciences, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia

Abstract

Background: Best practice standards aim to standardize care and improve outcomes. However, variation in clinical practice exists, and not all deviations are inappropriate. Measuring adherence to best practice standards remains challenging due to limitations in representation methods and data fidelity.

Objective: This scoping review aims to survey and synthesize the existing literature on the computable representation of guideline recommendations and to explore methods for detecting and quantifying deviations from best practice standards.

Methods: We followed the Arksey and O'Malley framework and PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines. Five databases (Ovid Medline, EMBASE, IEEE Xplore, Web of Science, and Scopus) were searched in November 2025. Studies were included if they either (1) described a computer representation of best practice standards or (2) assessed adherence to such standards using patient data, including patient data derived from electronic medical records or event logs. Screening was done using Covidence (Veritas Health Innovation). Data were extracted on representation, clinical context, data sources, adherence metrics, and modeling techniques. A narrative synthesis was conducted to identify themes.

Results: Twenty-four studies were included. Most studies were published as conference proceedings (13/24, 54%). Fourteen studies (14/24, 58%) included measurement of adherence to best practice standards. Cardiovascular conditions were the most common focus (13/24, 54%). Data sources included Health Level Seven (HL7) messages, structured electronic medical record data, event logs, and Fast Healthcare Interoperability Resources (FHIR)-transformed data. Best practice standards were formalized using Business Process Model and Notation (BPMN; 6/24, 25%), ontologies (7/24, 29%), FHIR (4/24, 17%), or hybrid approaches (4/24, 17%). The most common method for adherence measurement was rule-based alignment. Several studies incorporated weighted scoring to differentiate the severity of deviations. Process mining was used in a subset to detect sequence and timing variations. However, most models lacked contextual sensitivity and rarely incorporated patient-specific factors, such as comorbidities, patient acuity, or clinician rationale. Consequently, although deviations can be automatically identified, determining whether they were clinically warranted remained largely unresolved.

Conclusions: Despite promising advances, challenges persist in computer-interpretable representation and measuring adherence in a clinically meaningful way. Current approaches predominantly assess technical alignment rather than clinical relevance and are limited by data quality and standardization, thereby limiting real-world utility. This scoping review offers an innovative contribution by synthesizing evidence from 2 separate domains—the computable representation of best practice standards and the measurement of adherence. The findings emphasize the need for context-aware, standardized modeling and integration with clinical workflows to distinguish warranted from unwarranted deviations. Such advances are essential for scalable, transparent, and real-time adherence monitoring—ultimately driving safer, patient-centered care delivery.

(J Med Internet Res 2026;28:e79937) doi:10.2196/79937

KEYWORDS

clinical practice guidelines; best practice; guideline compliance; computer-interpretable guidelines; electronic medical records; clinical decision support

XSL•FO
RenderX

Introduction

Delivering patient care that aligns with evidence-based best practice has been associated with many benefits, including increased hospital efficiency, decreased operational costs, and improved patient outcomes [1-4], especially reduced length of stay and decreased mortality rate [4]. This is achievable through best practice standards, which are typically reflected in clinical practice guidelines (CPGs). The most cited definition refers to CPGs as "systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances" [5]. In the literature, terms such as "protocols," "practice policies," "clinical algorithms," and "clinical pathways" are sometimes used to refer to CPGs, although they differ in scope, specificity, and intended use [6]. While CPGs are narrative documents that provide overall evidence-based recommendations, clinical pathways (CPs) translate these guidelines into structured, time-sequenced clinical processes of care for a specific problem within a particular population [7]. CPs are typically developed at the institutional or regional level and are informed by international or national guidelines but tailored to local contexts [7]. Their structured, time-sequential nature may make them well-suited as benchmarks for monitoring and evaluating clinical practice.

Despite the recognized benefits and availability of CPGs and CPs, a persistent gap exists between recommended care and clinical practice delivery [8,9]. This discrepancy, often referred to as practice variations or pathway deviations, can reflect both unwarranted and warranted deviations from practice. Practice variations may indicate erroneous care resulting in adverse patient outcomes and increased health care expenses and are often referred to as unwarranted deviations. While some practice variations may be deemed warranted deviation when individualized care is provided to accommodate patient-specific needs. Distinguishing between unwarranted and warranted deviations is crucial to understanding and improving clinical practice. However, existing approaches to measuring adherence often rely on aggregated metrics or single process indicators and fail to capture the nuance and temporal complexity of clinical care.

Advancements in digital technology play a pivotal role in delivering health care and enabling systematic and data-driven monitoring of adherence to best practice standards [10]. Electronic medical records (eMRs) offer a large volume of time-stamped, structured clinical data that can be used to evaluate adherence to best practice. In addition, Clinical Decision Support Systems (CDSS) embedded within eMRs can promote adherence to best practice standards by delivering evidence-based recommendations in a variety of care aspects [11,12]. As a result, it has also led to the emergence of computer-interpretable guidelines (CIGs), which transform narrative CPGs or CPs into machine-readable, logical representations to support automated assessment of adherence.

Multiple methodologies have been developed to support the representation of best practice standards in computable formats. Guideline modeling, such as GLIF3, Asbru, SAGE, PROforma, as well as workflow and interoperability frameworks such as Business Process Model and Notation (BPMN) and Fast Healthcare Interoperability Resources (FHIR), offer different strengths in terms of expressiveness, configurability, and system integration [13-17]. In addition, Clinical Quality Language (CQL) [18], a high-level, domain-specific language used to support the creation and execution of clinical decision support rules and algorithms. However, the heterogeneity of these approaches introduces challenges in clinical implementation and usability, and cross-system interoperability [17]. Additionally, the availability of data and different interpretations of adherence measures continue to pose a challenge to the measurement of adherence to CPGs [19,20]. Nevertheless, a lack of standardized metrics and limited consideration of clinical context continue to constrain the practical utility and adoption of these methods in clinical practice.

A recent scoping review provides summaries on the role of informatics in addressing practice variations and the clinical tasks in which practice variations occur [21]. There remains a need to understand the methodologies behind the representation of best practice standards and how these representations enable the measurement of adherence to best practice. This scoping review aims to survey and synthesize the existing literature on the computable representation of guideline recommendations and to explore the extent to which these representations have been used to detect and quantify deviations from best practice. By doing so, this review highlights methodological and operational gaps in current approaches and identifies opportunities to advance adherence measurement in digital health systems.

Methods

Study Design and Framework

This scoping review was conducted following the methodology and guidelines by Arksey and O'Malley [22]. In line with scoping review conventions, no quality appraisal of the included studies was conducted, as the objective was to map the breadth of a diverse range of work.

Study Protocol

The PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) [23] was used to develop our review protocol. As a recommended practice, we used a checklist when reporting results to enhance transparency and reproducibility (eg, by sharing the exact search strategy and source databases). A complete checklist is available in Checklist 1.

Information Sources and Search Strategy

We searched 5 databases: Ovid Medline, EMBASE, IEEE Xplore, Web of Science, and Scopus. No study registries were searched. We also did not purposefully look for any online or print source. The last search was conducted in November 2025. The search strategy was developed in consultation with an experienced university librarian, without adapting previous search strategies from other literature reviews. It included combinations of controlled vocabulary and keywords related to clinical guidelines, clinician adherence, variation, and eMRs. We did not impose any time limit on obtaining a comprehensive

XSL•FO
RenderX

overview of published articles. We reported the search according to PRISMA-S (an Extension of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Statement) guidelines [24], which are available in Checklist 2. Additional articles were identified through backward searching of the reference lists in the included studies and the authors' collection. No additional studies or data were sought by contacting authors, experts, and manufacturers.

The references were imported to Covidence (Veritas Health Innovation) [25] for deduplication. During title and abstract screening, additional duplicates and irrelevant entries were manually removed. Covidence was also used for full-text screening, data extraction, and charting.

Eligibility Criteria

In this review, CPGs and CPs were considered as best practice standards as they provide evidence-based actionable recommendations. Our focus was on studies conducted in tertiary care settings, where guideline implementation and adherence monitoring are most frequently studied.

To maintain a focus on comprehensive care delivery and computable representations of best practice, we included studies that either (1) described or modeled a computer representation of best practice standards, or (2) assessed adherence to such standards using patient data, such as eMR-derived patient data or patient event logs. Studies were eligible if they captured multiple (>1) guidelines or components of care in the patient pathway. Studies focused solely on a single component, such as medication adherence—while important—were excluded, as they addressed only a single aspect of care. We also included studies that aimed to detect or quantify deviations from best practice across multiple components outlined by CPGs and CPs. Qualitative analyses (eg, interviews and ethnographic observations) that did not offer computer representations or measurable assessments of adherence were excluded. Additionally, studies that used a single process indicator as a proxy for adherence were excluded if they lacked a broader evaluation of best practice adherence.

Articles were excluded if they were not in English, had no full text available, or were not aligned with the research questions. Given the emerging nature of this topic and the evolving technological landscape, conference proceedings were also considered to capture developments in the field.

Study Selection

The retrieved articles were initially screened by applying the eligibility criteria by one reviewer (MTH). Concurrently, 5% of the total articles were randomly selected and separately screened by a second reviewer (MS). Discrepancies were discussed and resolved through consensus. The study selection process is illustrated in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

Data Charting and Data Items

A structured data charting form was developed to extract the following variables from each included study:

1. Bibliographic information (title, authors, year of publication, and publisher)
2. Study aims
3. Representation of clinical data
4. Representation of best practice standards
5. Whether deviations from clinical pathways were detected and measured
6. The mechanism used to measure adherence
7. Reported limitations

Full extraction details are available in Multimedia Appendix 1 [20,26-48].

Synthesis of the Result

Extracted data were analyzed using a narrative synthesis approach. Studies were grouped thematically based on their technical approaches to best practice representation and adherence measurement. Key features, methodological innovations, and limitations were summarized across studies. All authors contributed to iterative discussions, the refinement of the thematic structure, and the interpretation and presentation of the results.

Results

Included Articles and Metadata

For this scoping review, 24 articles [20,26-46,47,48] were included for data extraction (Figure 1). Most studies were conducted in China (6/24, 25%), the United States (5/24, 21%), and Germany (3/24, 13%). Of these, more than half (13/24, 54%) were published as conference proceedings rather than full-length journal articles, likely reflecting the technically innovative but emerging nature of this research area. Of 24 studies [20,26-46,47,48], 14 (58%) [20,26-36,46,48] included a measure of adherence to best practice standards (Figure 2). The number indicates how frequently each combination was used in the included studies. Formal knowledge representations—such as BPMN and Semantic Web Ontology Language (OWL)—combined with rule-based alignment are the most commonly used approach for representing best practice standards and measuring adherence.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram demonstrates the article search and selection process.

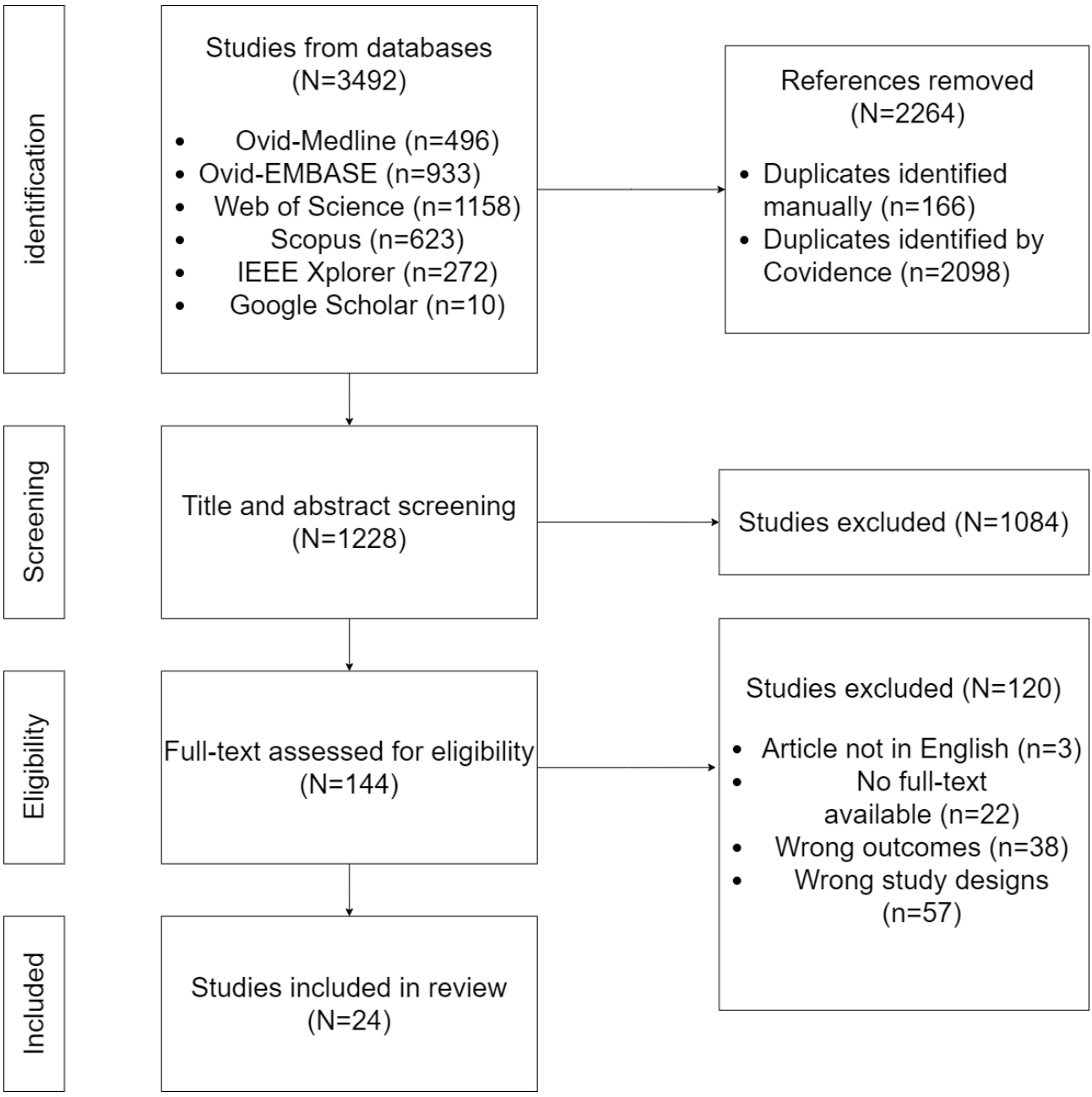

XSL•FO
RenderX

Figure 2. Map of best practice standards representation by approach to measure adherence in the included literature (N=14). BPMN: Business Process Model and Notation; FHIR: Fast Healthcare Interoperability Resources; OWL: Semantic Web Ontology Language.

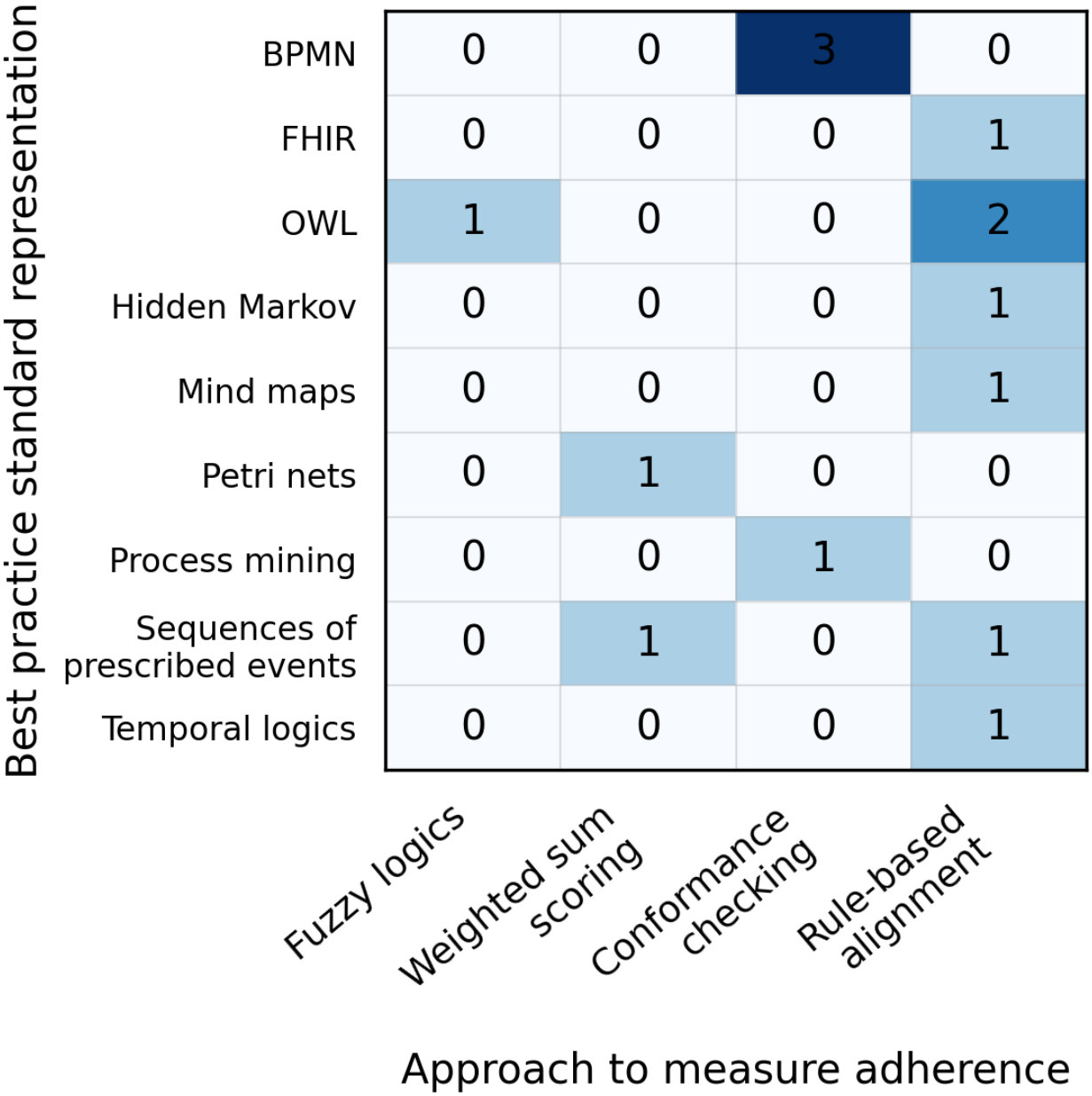

Clinical Conditions

Among the 24 included studies [20,26-46,47,48], only 1 study did not provide any clinical setting [26]. Cardiovascular disease was the most common clinical area, accounting for 54% (13/24) of the sample. Two studies applied their models across multiple clinical conditions (Table 1).

XSL•FO
RenderX

Table . Clinical conditions whose clinical practice guidelines or clinical pathways were translated to computer-interpretable guidelines in the included studies (N=24).

Clinical condition	Count
Weaning protocol in Intensive Care Unit (ICU)	1
Diabetes	1
Major joint replacement	1
Nutrition	1
Thyroid disease	1
No clinical case	1
COVID-19	2
Family planning and sexually transmitted infections	1
Cardiovascular disease	13
Mixed clinical conditions	2

Data Sources Used for Monitoring Adherence to Best Practice Standards

Various sources of clinical data were used in the included studies, including Health Level Seven (HL7) messages, structured clinical data extracted from eMR [27,28], patient events trace derived by sorting timestamped entries from eMR [20,29-31], and system-generated event logs [32-35]. Some studies transformed eMR data into standardized formats using the FHIR specification [36-38]. Conversely, several studies focused solely on the conceptual representation of best practice standards without implementation [26,39-46].

Representation of Best Practice Standards

Of the 24 studies [20,26-46,47,48], 13 [27,28,32,36,38-42,44-47] used CPGs as their reference for best practice, while the remaining studies used CPs. Various methods were used to transform CPGs and CPs to CIG format. The most used approaches were BPMN (6/24, 25%) and FHIR (4/24, 16%). For example, Yan et al [35] used BPMN to model a task-time matrix CP for unstable angina pectoris. The authors mapped terminology from eMR to pathway activities, identified key activities as milestones, and then embedded the remaining activities between those milestones to construct the process [35]. Parallel and inclusive gateways were used to represent concurrent activities, while exclusive gateways captured alternate clinical routes [35]. This BPMN representation yielded a computer-interpretable pathway model that served as a reference standard for detecting deviations in real-world care delivery [35]. Meanwhile, Lichtner et al [45] represented CPGs in FHIR by identifying the required elements and metadata from the CPGs to be captured in FHIR representation. An information model was built by organizing these elements into 3 main categories—medical relationships, metadata, and justification for the recommendations—and sorting them into 3 layers of certainty of evidence (primary clinical studies, systematic review and outcome-level assessment, and final recommendation derivation with evidence grading) [45]. Finally, the model was mapped into EBM-on-FHIR and CPG-on-FHIR profiles using resources such as PlanDefinition, EvidenceVariable, and ArtifactAssessment, thus enabling a computer-interpretable representation of CPG recommendations with underlying evidence certainty [45]. Details of the common methods used to represent CIGs are presented in Table 2.

XSL•FO
RenderX

Table . Description of the most used approaches to represent computer-interpretable guidelines.

Approach name	Included studies	Mechanism
Business Process Model and Notation (BPMN)	[32,33,35,37,44]	• Map variables extracted from eMR[a] to those components of CPGs[b] and CPs[c] • Identify key activities in CPGs and CPs to act as milestones that connect different phases of the care process • Add other care activities in between key activities • Activities are sequenced based on the clinical logic: whether they occur simultaneously or exclusively, and the order in which activities occur first
Ontology (OWL)	[26,27,30,43,44]	• Clinical concepts are defined as classes • A clinical concept can have object properties which describe the relationship between clinical concepts and data properties which record its quantitative aspect • eMR variables are then mapped to the newly created ontology to create instances of classes • Decision rules are created based on the if-then logic to encompass the conditions as outlined in CPGs and CPs
FHIR	[36,37,40,45]	• Identify the contents and metadata of CPGs and CPs to be captured in the FHIR-based[d] representation • Information model is built to understand the clinical concepts and relationships between them • Components of CPGs and CPs are mapped to the corresponding FHIR resources • Specifically, the PlanDefinition resource is used to model the care processes as outlined in CPGs and CPs. Meanwhile, the ActivityDefinition resource represents each care activity within the whole care process

[a]eMR: electronic medical record.

[b]CPG: clinical practice guideline.

[c]CP: clinical pathway.

[d]FHIR: Fast Healthcare Interoperability Resources.

Interestingly, some studies used hybrid approaches to represent best practice standards. For example, Dos Santos Leandro et al [37] used BPMN and FHIR to model stroke care CPs, first structuring the care process with BPMN, then mapping the components to corresponding FHIR resources to enable semantic interoperability. Whereas 2 studies [27,47] defined their best practice standards as linear sequences of prescribed actions without using formal modeling tools.

In addition to these core approaches, other approaches used to transform CPGs and CPs to CIGs included Petri Net [29], a combination of Natural Rule Language (NRL) and Object Constraint Language (OCL) [47], fuzzy logics [38], process mining [34], decision tree [28,39], hidden Markov model [31], and the formal representation language PROforma [41,42]. These studies thus demonstrate the use of a diverse range of formalisms to represent best practice standards; however, some of them lacked clinical integration, as observed in studies such as [39,41,42].

Approaches to Measuring Adherence

The included studies used a range of techniques to detect and quantify deviations from best practices standards. The most commonly used approach involved rule-based alignment, wherein individual patient records were assessed against predefined care steps derived from CPGs and CPs [20,27-29,36,46,48]. In this approach, CPG and CP components are translated into computable “Event-Condition-Action” (ECA) constraints, each constraint representing a specific care step. Individual patient records were then systematically aligned with these predefined constraints using structured data extracted from eMRs. For example, Kaur et al [27] translated the nutrition recommendations into “rules” that dictate actions based on specific conditions or criteria using logical expressions or attributes (day of life, birth weight, and current health

XSL•FO
RenderX

conditions). Similarly, Lichtner et al [36] used guideline recommendations in a FHIR-based declarative format and built executable ECA rules based on the criteria specified in the recommendations and then executed against standardized patient data to determine whether the patients' data aligned with the recommendations. This rule-based alignment enables automated, transparent, and scalable assessment of adherence to best practice standards across large patient cohorts.

Beyond rule-based alignment, studies [20,29,48] incorporated weighting mechanisms to reflect the clinical significance of different types of deviations. Instead of treating all deviations from best practice standards as equally significant, these weighted scoring systems distinguished between addition, omission, and substitution of care activities and assigned differential weights. Implementation typically involves differentiating between types of activities that mismatch the prescribed activities in CPGs and CPs and those observed in patient data. In this context, omissions refer to activities specified in CPGs and CPs but not observed in the patient event trace, while additions refer to activities observed in the patient event trace that are not prescribed in CPGs and CPs.

By using weighting mechanisms, the adherence score reflects not only alignment with the referenced CPGs and CPs but also the severity of deviations. For example, van de Klundert et al [20] developed and tested dynamic programming formulations to match patient trajectories to the prescribed CPs, allowing partial adherence and assigning varying penalties to deviations based on their severity. Konrad et al [48] not only assigned equal weights to the addition and omission of activities compared to what was prescribed in the CP, but also to whether patient outcomes remained within target ranges during specified time intervals. Meanwhile, Zema et al [29] did not consider weighting patient outcomes, but they emphasized documentation quality in their Clinical Pathway Deviation Index (CPDI). CPDI was the weighted sum of 5 indicators: addition, omission, variation in activities sequence and duration, and missing in documentation. While the least weight was assigned to missing in documentation, equivalent weights were assigned to the other indicators.

Process mining techniques were used in a subset of studies to reconstruct care processes from event logs or event traces. Authors in several studies [29,30,32] used conformance checking and an optimal alignment algorithm to detect deviations. Deviations were identified by misalignment between observed activities and expected pathways, with some studies applying A* based algorithm—a graph traversal and path-finding algorithm—to determine the shortest or most efficient match between the observed and expected activities.

Unlike traditional health service research, which relies on single-point indicators, these studies adopted a process-level alignment approach. Rather than evaluating isolated components of CPGs and CPs, they aimed to align entire sequences of care activities with components extracted from best practice standards. This approach captures both what and when actions were performed, thus reflecting the temporal flow of care delivery. For example, Yan et al [26] described the modeling of the relative temporal relations between 2 clinical activities. They used temporal rules expressed in the Semantic Web Rule Language (SWRL) to describe the 30-minute interval between administering intravenous antibiotics and the cesarean delivery treatment intervention, as shown in Textbox 1 [26]. By analyzing these sequences, researchers can capture deviations in timing and sequencing, providing a more holistic and dynamic assessment of adherence to best practice standards.

Textbox 1. Example of modeling the relative temporal relations between clinical activities using a Semantic Web Ontology Language (OWL).

begins(IVPuchAntibiotics, ?t1) ^

begins(CesareanDelivery, ?t2) ^ swrlb:

subtractDateTimesYieldingDayTimeDuration(?t, ?t2, ?t1)

⇒swrlb:greaterThan (?t,T30M).

Despite these methodological innovations, most studies focused on structural or process-level compliance and did not consider contextual factors such as patient conditions, comorbidities, or clinician intent. As a result, the distinction between warranted deviations and unwarranted deviations remained largely unresolved across the included studies.

Discussion

Principal Findings

This scoping review explored existing literature on the computable representations of best practice standards and methods for detecting and quantifying deviations from best practice using data extracted from eMRs. We identified 24 studies [20,26-46,47,48], a majority of which were published as conference proceedings. Cardiovascular disease was the most common clinical domain, most likely due to the extensive evidence available of CPGs in this clinical specialty [49].

A wide variety of methods were used to represent patient data, including system-generated event logs, patient event traces, and FHIR. Similarly, CIG models were constructed using BPMN, ontology frameworks, FHIR, or hybrid combinations of these. These models aim to enable structured and machine-readable, potentially interoperable standard-based guideline representation exchange format [36].

While not all studies detailed explicit methods for measuring adherence, the most common approach was rule-based alignment, wherein individual patient records were assessed against predefined care steps derived from CPGs and CPs. Some studies incorporated weighting mechanisms to differentiate the clinical significance of deviations, reflecting the recognition that not all deviations from recommendations are negative.

Three principal insights were identified. First, a diverse range of formalism types was used to represent best practice standards, though few were implemented in a clinically integrated fashion.

XSL•FO
RenderX

Second, approaches to measuring adherence varied significantly, ranging from rule-based alignment to more nuanced methods involving weighted scoring and process mining. Finally, the effectiveness of adherence measurement was closely linked to the quality and structure of the underlying data sources, which varied across studies.

Together, these findings demonstrate both the progress made and the challenges that remain in operationalizing and measuring adherence to best practice using eMR-derived data. The results of this study highlight the need for standardized, context-sensitive, and interoperable systems that can support meaningful adherence monitoring to best practice standards as well as more quality and peer-reviewed research in the area.

Representing Best Practice: Challenges in Formalization

Translating narrative CPGs and CPs into structured, computer-interpretable formats is inherently complex. The narrative language used in CPGs is often characterized by ambiguity, conditional logic, and implicit assumptions, making it difficult to convert into electronic formats without loss of fidelity. Many studies in this review used hierarchical decomposition of narrative text into modular components combined with temporal logic and decision rules to form the basis of CIGs. However, even when the logic is accurately encoded, clinicians may find it challenging to validate outputs from these systems as CIG languages are not always accessible for clinicians [50].

Moreover, the global implementation of these systems is often constrained by their design. Many of them were developed as standalone tools that do not communicate with each other or with existing health information systems [51]. Additionally, these systems are also highly tailored to specific use cases and cannot be easily modified or maintained, except by their original authors, making them unsustainable for broader adoption [17,52,53].

BPMN has the advantage of being a graphical representation that can visualize workflow processes, as demonstrated in the included studies (Table 2). In addition, a systematic review has also reinforced BPMN's potential to standardize and redesign clinical processes [54]. However, there are certain limitations to using BPMN to represent best practice standards. Despite being excellent at modeling workflow, BPMN is primarily designed to model generic business processes and may not be able to capture the complexity and dynamics of health care processes [55,56]. Implementing BPMN in modeling clinical pathways requires a considerable investment in finance, time, and human resources [57]. At the same time, its technical complexity also poses a barrier to adoption among clinicians, especially those without formal training in informatics.

These limitations emphasize the need for models that balance clinical expressiveness with technical interoperability and usability, ensuring sustainability, scalability, and the ability to provide meaningful insights into clinical practice.

Measuring Adherence: Methodological Tensions

Approaches to measuring adherence to best practice standards varied significantly, reflecting a wide methodological spectrum. Traditional approaches to measuring adherence to best practice standards often rely on external, aggregated metrics, such as length of stay, cost, infection rate, mortality, etc. Although these indicators can provide valuable insights at a system level, they offer little information about the underlying process of care, as well as patient-level information [58]. Such measures are insufficient to determine whether care was delivered in accordance with the sequence, timing, and logic prescribed by best practice standards.

Several studies in this review sought to overcome this barrier by aligning process-level components with guideline-specified components and patient data. These approaches thus provide a more granular view of whether the intended sequence and logic of care delivery were followed. Despite their methodological sophistication, most of these models did not integrate contextual data such as patient comorbidities, preferences, or clinical urgency.

This gap presents persistent tension. While current models are effective at detecting technical deviation, they remain limited in their ability to distinguish between clinically necessary deviations from those that are undesirable. The future approach to adherence measurement must evolve to be more patient-centered and context-aware.

Limitations of Process Mining and Data Source Fidelity

Clinical decision-making is primarily guided by a patient's health status and context, which are often outside the stated space that the traditional process mining techniques rely on [59]. Many process mining techniques can tell us which treatment behaviors are performed and in which order in accordance with established best practice standards [60]. These methods are attractive due to their scalability and ability to visualize workflow variation. However, their utility in clinical settings is often limited by the eMR data structure and granularity [61].

The effectiveness of process mining techniques is often hindered by the inconsistent, incomplete, or unstructured nature of eMR data. A recent scoping systematic review shows that process mining cannot distinguish the differences in deviations due to the complex nature of clinical data [62]. In addition, the authors mentioned that process mining becomes more beneficial as the input data become more structured, which is usually not the case, especially for data extracted from eMR [62].

Consequently, the accuracy of deviation detection in process mining is heavily dependent on data fidelity. Even when the result produced by process mining is technically significant, it may lack clinical relevance to clinicians.

Since clinical decision-making is primarily driven by patient-specific factors, many of which are not captured in structured eMR fields, process mining risks oversimplifying the complex care delivery process. Without integration of outcome data, clinician notes, or patient-specific modifiers,

XSL•FO
RenderX

these methods offer only a partial view of adherence and may not resonate with clinical end users.

Recent Approach Using FHIR for Interoperability and Real-Time Monitoring of Adherence to Best Practice Standards

Among the most promising recent developments in this domain is the adoption of FHIR. As an emerging interoperability standard, FHIR aims to advance the messaging standard to enhance the exchange and interpretation of information across disparate and heterogeneous systems [63]. In the included studies, FHIR was used to represent patient data in a format that supports automated adherence analysis. Therefore, the use of FHIR not only provides concise structural data input but also promotes the reusability of clinical data for research and accelerates data availability.

A recent scoping review, conducted by Tabari et al [64], provided a summary of the tools and techniques used to represent the FHIR-based data model, demonstrating the promising use of FHIR to facilitate the integration, transformation, and analysis of clinical data. Indeed, in this scoping review, as shown in Table 2, FHIR was used to represent best practice standards rather than solely represent patient data. FHIR was demonstrated to successfully translate guideline recommendations into a standardized, computer-interpretable format, coupled with a rating system that can assess the level of certainty of evidence [45]. The authors also performed a study on automating the integration of guideline representation with real-time clinical data. They found a promising possibility of real-time monitoring adherence to guideline recommendations [36]. In addition, a recent publication demonstrates that it is possible to set up a trustworthy process to develop computer-interpretable representations of quality indicators for future automated quality management systems [65]. As a result, advancements in adopting FHIR to monitor adherence to best practice standards demonstrate the possibility of a real-time adherence monitoring system that not only detects deviations but also weights them according to the certainty of the supporting evidence.

Future Approach to Move Toward a More Meaningful Evaluation of Appropriate Care Delivery

Given the limitations of eMR data in measuring adherence to best practice standards, further research should aim to improve the applicability of eMR data for this purpose. This could be addressed by assessing the quality of eMR data characteristics—accuracy, completeness, reliability, relevance, and timelines—before evaluating adherence to CPGs and CPs. Furthermore, it could be beneficial to investigate the use of large language model technologies for the representation of CPGs and CPs and the measurement of adherence to best practice standards.

Despite these methodological innovations, most studies primarily focus on structural adherence, that is, whether the documented care sequence matched the prescribed components of CPGs and CPs, rather than on whether it was clinically appropriate in the specific care context. Consequently, while deviations from best practice standards can be detected with high accuracy, the crucial distinction between warranted deviations, which reflect high-quality patient-centered care, and unwarranted deviations, which can result in patient harm and unnecessary use of health care resources, remained largely unresolved.

Future studies can address this limitation by accounting for patient characteristics, outcomes, and clinician intent. For instance, detected deviations could be reclassified based on patient-specific conditions, allowing for outcome-informed validation of adherence measurement. By combining rule-based alignment and contextual reasoning, future approaches can advance beyond static measurement of adherence to best practice standards toward a more meaningful evaluation of appropriate care delivery. Additionally, future research should consider other potential factors influencing adherence to best practice standards, such as the structure of medical billing items, which could lead to deviations from the CPGs and CPs.

This scoping review makes several novel contributions to the existing literature by providing a comprehensive mapping of the computable representation of best practice standards and how adherence is measured—2 domains that have often been examined separately. Unlike previous reviews that focus only on either the technical approach or clinical evaluation, this scoping review synthesizes evidence from both perspectives and demonstrates how computer-interpretable format representation enables automated adherence assessment. It thus establishes a conceptual link between the modeling framework used for computable representation and the analytical methods used for evaluating practice variation. Notably, the findings promote the shift toward dynamic, computerized real-time monitoring of best practice standards, which can serve as a foundation for health care systems aiming to leverage eMR data for quality improvement. This also highlights the potential for a scalable, transparent, and clinically meaningful system for monitoring best practices to reduce unwarranted variation, strengthen guideline implementation, and ultimately improve patient care.

Limitations

This scoping review has some limitations, including the restriction to English-language publications and databases indexed up to November 2025, which may have excluded relevant literature, especially from non–English-speaking countries. We also did not register this study protocol on PROSPERO. In addition, due to the heterogeneity of study designs, a formal quality appraisal was not performed. We also focused on the quantitative measurement of adherence to best practice standards. Other works had been done on investigating the deviations from best practice using qualitative approaches, which were not included and analyzed in this scoping review. However, the qualitative type of studies often focus on providing insights into the nature of deviations, as demonstrated in the studies done by Kahol et al [66], which may not fit well with our aims, as our study focuses on quantitative measurements. We also explored studies that proposed conceptual models rather than real-world implementation, limiting the generalizability of findings. Finally, we sourced studies with data extracted from

XSL•FO
RenderX

eMR without considering additional data sources which again could provide the context of guideline deviations.

Conclusions

This scoping review offers an innovative contribution by surveying and synthesizing the existing literature that highlights the significant progress made in the formalization of best practice standards and the development of quantitative measures for adherence monitoring. However, significant gaps remain. The representation of best practice standards is heterogeneous and often disconnected from clinical workflows. Adherence measurement tends to prioritize detectability over clinical relevance. Although various data sources have been leveraged for these approaches, many remain incomplete or underused.

The increasing adoption of FHIR highlights the potential to improve eMR-based adherence measurement; however, technical and operational barriers must be addressed to achieve these goals. To realize the full potential of eMR-based adherence monitoring, future work must focus on clinically meaningful modeling, standardized yet flexible adherence metrics, and real-time, interoperable systems that integrate seamlessly with care delivery. Such advances will be critical to supporting evidence-based, adaptive, and patient-centered health care systems.

Acknowledgments

This research is supported by the Digital Health Cooperative Research Centre Limited (DHCRC). DHCRC is funded under the Australian Commonwealth's Cooperative Research Centres (CRC) Program. Author MTH holds a higher degree of research scholarship funded through a DHCRC project grant. The authors would like to thank Malhar Soni for screening the articles for this scoping review.

Funding

The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

All extraction data from the included studies are presented in Multimedia Appendix 1. All other data for title and abstract screening and full-text screening are available upon request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed properties of included studies.
[DOCX File, 32 KB - jmir_v28i1e79937_app1.docx]

Checklist 1
PRISMA checklist for scoping review.
[DOCX File, 111 KB - jmir_v28i1e79937_app2.docx]

Checklist 2
PRISMA-S checklist search strategies and search strings.
[DOCX File, 26 KB - jmir_v28i1e79937_app3.docx]

References

1. Barbieri A, Vanhaecht K, Van Herck P, et al. Effects of clinical pathways in the joint replacement: a meta-analysis. BMC Med 2009 Jul 1;7(1-11):32. [doi: 10.1186/1741-7015-7-32] [Medline: 19570193]
2. Brusco NK, Sykes K, Cheng AC, et al. A state-wide implementation of a whole of hospital sepsis pathway with a mortality based cost-effectiveness analysis from a healthcare sector perspective. PLOS Glob Public Health 2023;3(5):e0000687. [doi: 10.1371/journal.pgph.0000687] [Medline: 37205639]
3. Thursky K, Lingaratnam S, Jayarajan J, et al. Implementation of a whole of hospital sepsis clinical pathway in a cancer hospital: impact on sepsis management, outcomes and costs. BMJ Open Qual 2018;7(3):e000355. [doi: 10.1136/bmjoq-2018-000355] [Medline: 30019016]
4. Connor L, Dean J, McNett M, et al. Evidence-based practice improves patient outcomes and healthcare system return on investment: findings from a scoping review. Worldviews Evid Based Nurs 2023 Feb;20(1):6-15. [doi: 10.1111/wvn.12621] [Medline: 36751881]
5. Field MJ, Lohr KN, editors. Clinical practice guidelines. In: Directions for a New Program: National Academies Press; 1990. [doi: 10.17226/1626] [Medline: 25144032]

XSL•FO
RenderX

6. A guide to the development, implementation and evaluation of clinical practice guidelines. National Health and Medical Research Council. 1999. URL: https://www.nhmrc.gov.au/sites/default/files/images/a-guide-to-the-development-and-evaluation-of-clinical-practice-guidelines.pdf [accessed 2026-01-15]
7. Rotter T, Jong RB, Lacko SE, Ronellenfitsch U, Kinsman L, editors. Clinical pathways as a quality strategy. In: Improving Healthcare Quality in Europe 2019. URL: https://www.ncbi.nlm.nih.gov/books/NBK549262 [accessed 2026-01-15]
8. Hay MC, Weisner TS, Subramanian S, Duan N, Niedzinski EJ, Kravitz RL. Harnessing experience: exploring the gap between evidence-based medicine and clinical practice. J Eval Clin Pract 2008 Oct;14(5):707-713. [doi: 10.1111/j.1365-2753.2008.01009.x] [Medline: 19018899]
9. Kennedy PJ, Leathley CM, Hughes CF. Clinical practice variation. Med J Aust 2010 Oct 18;193(S8):S97-S99. [doi: 10.5694/j.1326-5377.2010.tb04021.x] [Medline: 20955142]
10. Campanella P, Lovato E, Marone C, et al. The impact of electronic health records on healthcare quality: a systematic review and meta-analysis. Eur J Public Health 2016 Feb;26(1):60-64. [doi: 10.1093/eurpub/ckv122] [Medline: 26136462]
11. Jaspers MWM, Smeulers M, Vermeulen H, Peute LW. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. J Am Med Inform Assoc 2011 May 1;18(3):327-334. [doi: 10.1136/amiajnl-2011-000094] [Medline: 21422100]
12. Garg AX, Adhikari NKJ, McDonald H, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005 Mar 9;293(10):1223-1238. [doi: 10.1001/jama.293.10.1223] [Medline: 15755945]
13. Boxwala AA, Peleg M, Tu S, et al. GLIF3: a representation format for sharable computer-interpretable clinical practice guidelines. J Biomed Inform 2004 Jun;37(3):147-161. [doi: 10.1016/j.jbi.2004.04.002] [Medline: 15196480]
14. Seyfang A, Miksch S, Marcos M. Combining diagnosis and treatment using ASBRU. Int J Med Inform 2002 Dec 18;68(1-3):49-57. [doi: 10.1016/s1386-5056(02)00064-3] [Medline: 12467790]
15. Tu SW, Campbell JR, Glasgow J, et al. The SAGE Guideline Model: achievements and overview. J Am Med Inform Assoc 2007;14(5):589-598. [doi: 10.1197/jamia.M2399] [Medline: 17600098]
16. Riaño D, Peleg M, Ten Teije A. Ten years of knowledge representation for health care (2009-2018): topics, trends, and challenges. Artif Intell Med 2019 Sep;100:101713. [doi: 10.1016/j.artmed.2019.101713] [Medline: 31607346]
17. Peleg M. Computer-interpretable clinical guidelines: a methodological review. J Biomed Inform 2013 Aug;46(4):744-763. [doi: 10.1016/j.jbi.2013.06.009] [Medline: 23806274]
18. Clinical Quality Language specification. CQL. URL: https://cql.hl7.org [accessed 2026-01-15]
19. Kardas P, Aguilar-Palacio I, Almada M, et al. The need to develop standard measures of patient adherence for big data: viewpoint. J Med Internet Res 2020 Aug 27;22(8):e18150. [doi: 10.2196/18150] [Medline: 32663138]
20. van de Klundert J, Gorissen P, Zeemering S. Measuring clinical pathway adherence. J Biomed Inform 2010 Dec;43(6):861-872. [doi: 10.1016/j.jbi.2010.08.002] [Medline: 20696277]
21. Sohn S, Moon S, Prokop LJ, Montori VM, Fan JW. A scoping review of medical practice variation research within the informatics literature. Int J Med Inform 2022 Sep;165:104833. [doi: 10.1016/j.ijmedinf.2022.104833] [Medline: 35868231]
22. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
23. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
24. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
25. Covidence systematic review software, Veritas Health Innovation. Covidence. 2025. URL: www.covidence.org [accessed 2026-01-15]
26. Yan Y, Zhibin J, Dong Y, Gang D. A semantics-based clinical pathway workflow and variance management framework. Presented at: 2008 IEEE International Conference on Service Operations and Logistics, and Informatics; Oct 12-15, 2008. [doi: 10.1109/SOLI.2008.4686499]
27. Kaur R, Jain M, McAdams RM, et al. An ontology and rule-based clinical decision support system for personalized nutrition recommendations in the neonatal intensive care unit. IEEE Access 2023;11:142433-142446. [doi: 10.1109/ACCESS.2023.3341403]
28. Yu HW, Hussain M, Afzal M, et al. Use of mind maps and iterative decision trees to develop a guideline-based clinical decision support system for routine surgical practice: case study in thyroid nodules. J Am Med Inform Assoc 2019 Jun 1;26(6):524-536. [doi: 10.1093/jamia/ocz001] [Medline: 31087071]
29. Zema M, Rosati S, Duran Carvajal JE, Balestra G. CPDI: an index for measuring deviations in clinical pathways. Annu Int Conf IEEE Eng Med Biol Soc 2015 Aug;2015:1385-1388. [doi: 10.1109/EMBC.2015.7318627] [Medline: 26736527]
30. Huang ZX, Bao YR, Dong W, Lu XD, Duan HL. Online treatment compliance checking for clinical pathways. J Med Syst 2014 Oct;38(10):123. [doi: 10.1007/s10916-014-0123-0] [Medline: 25149871]
31. Li X, Liu H, Zhang S, et al. Automatic variance analysis of multistage care pathways. Stud Health Technol Inform 2014;205:715-719. [doi: 10.3233/978-1-61499-432-9-715] [Medline: 25160280]

XSL•FO
RenderX

32. Xu H, Yan H, Pang J, Nan S, Yang X, Zhao D. Evaluating the relative value of care interventions based on clinical pathway variation detection and propensity score. Presented at: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 16-19, 2020; Seoul, Korea (South p. 1184-1187 URL: https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=9312958 [doi: 10.1109/BIBM49941.2020.9313546]
33. Yan H, Lu X, Van Gorp P, et al. On accurate, automated and insightful deviation analysis of clinical protocols. Presented at: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 3-6, 2018; Madrid, Spain p. 1130-1134. [doi: 10.1109/BIBM.2018.8621133]
34. Noshad M, Rose CC, Chen JH. Signal from the noise: a mixed graphical and quantitative process mining approach to evaluate care pathways applied to emergency stroke care. J Biomed Inform 2022 Mar;127(no):104004. [doi: 10.1016/j.jbi.2022.104004] [Medline: 35085813]
35. Yan H, Van Gorp P, Kaymak U, et al. Variance analysis in Task-Time matrix clinical pathways. Presented at: 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI); Orland, FL, USA p. 253-256. [doi: 10.1109/BHI.2017.7897253]
36. Lichtner G, Spies C, Jurth C, et al. Automated monitoring of adherence to evidenced-based clinical guideline recommendations: design and implementation study. J Med Internet Res 2023 May 4;25:e41177. [doi: 10.2196/41177] [Medline: 36996044]
37. Dos Santos Leandro G, Moro CMC, Cruz-Correia RJ, Portela Santos EA. FHIR implementation guide for stroke: a dual focus on the patient's clinical pathway and value-based healthcare. Int J Med Inform 2024 Oct;190:105525. [doi: 10.1016/j.ijmedinf.2024.105525] [Medline: 39033722]
38. Minutolo A, Esposito M, Pietro G, Soonchunhyang U, Icar Cnr I, Networking V, et al. Fuzzy on FHIR: a decision support service for healthcare applications. Presented at: International Conference on P2P, Parallel, Grid, Cloud and Internet Computing; Republic of Korea p. 163-172. [doi: 10.1007/978-3-319-49109-7_16]
39. Ugon A, Bouzid AIH, Jaulent MC, Favre M, Duclos C, Jobez E, et al. Building a knowledge-based tool for auto-assessing the cardiovascular risk. Stud Health Technol Inform 2018;247:735-739. [Medline: 29678058]
40. Shivers J, Amlung J, Ratanaprayul N, Rhodes B, Biondich P. Enhancing narrative clinical guidance with computer-readable artifacts: authoring FHIR implementation guides based on WHO recommendations. J Biomed Inform 2021 Oct;122:103891. [doi: 10.1016/j.jbi.2021.103891] [Medline: 34450285]
41. Oliveira T, Neves J, Costa A, Novais P, Neves J, et al. An interpretable guideline model to handle incomplete information. In: Omatu S, De Paz Santana J, González S, Molina J, Bernardos A, Rodríguez J, editors. Distributed Computing and Artificial Intelligence Advances in Intelligent and Soft Computing: Springer, Berlin, Heidelberg; 2012, Vol. 151. URL: https://link.springer.com/chapter/10.1007/978-3-642-28765-7_52#citeas [accessed 2026-01-27] [doi: 10.1007/978-3-642-28765-7_52]
42. Martínez-Salvador B, Marcos M, Palau P, Domínguez Mafé E. A model-driven transformation approach for the modelling of processes in clinical practice guidelines. Artif Intell Med 2023 Mar;137:102495. [doi: 10.1016/j.artmed.2023.102495] [Medline: 36868689]
43. Hu Z, Li JS, Yu HY, Zhang XG, Suzuki M, Araki K. Modeling of clinical pathways based on ontology. Presented at: 2009 IEEE International Symposium on IT in Medicine & Education; Aug 14-16, 2009. [doi: 10.1109/ITIME.2009.5236262]
44. Madhusanka S, Walisadeera A, Dantanarayana G, Goonetillake J, Ginige A. An ontological clinical decision support system based on clinical guidelines for diabetes patients in Sri Lanka. Healthcare (Basel) 2020 Dec 18;8(4):573. [doi: 10.3390/healthcare8040573] [Medline: 33352875]
45. Lichtner G, Alper BS, Jurth C, et al. Representation of evidence-based clinical practice guideline recommendations on FHIR. J Biomed Inform 2023 Mar;139(104305):104305. [doi: 10.1016/j.jbi.2023.104305] [Medline: 36738871]
46. Autexier S, Bawadekji M, Hutter D, Wolters R. Supporting clinical guidelines using DL-temporal reasoning. Presented at: 2013 10th International Conference & Expo on Emerging Technologies for a Smarter World (CEWIT); Oct 21-22, 2013; Melville, NY p. 1-6. [doi: 10.1109/CEWIT.2013.6713766]
47. Farkash A, Timm JTE, Waks Z. A model-driven approach to clinical practice guidelines representation and evaluation using standards. Stud Health Technol Inform 2013;192:200-204. [Medline: 23920544]
48. Konrad R, Tulu B, Lawley M. Monitoring adherence to evidence-based practices: a method to utilize HL7 messages from hospital information systems. Appl Clin Inform 2013;4(1):126-143. [doi: 10.4338/ACI-2012-06-RA-0026] [Medline: 23650493]
49. Writing Committee Members, Virani SS, Newby LK, et al. 2023 AHA/ACC/ACCP/ASPC/NLA/PCNA Guideline for the management of patients with chronic coronary disease: a report of the American Heart Association/American College of Cardiology Joint Committee on Clinical Practice Guidelines. J Am Coll Cardiol 2023 Aug 29;82(9):833-955. [doi: 10.1016/j.jacc.2023.04.003] [Medline: 37480922]
50. Martínez-Salvador B, Marcos M. Supporting the refinement of clinical process models to computer-interpretable guideline models. Bus Inf Syst Eng 2016 Oct;58(5):355-366. [doi: 10.1007/s12599-016-0443-3]
51. Wong D, Peek N. Does not compute: challenges and solutions in managing computable biomedical knowledge. BMJ Health Care Inform 2020 Jul;27(2):e100123. [doi: 10.1136/bmjhci-2019-100123] [Medline: 32723852]

XSL•FO
RenderX

52. De Clercq P, Kaiser K, Hasman A. Computer-interpretable guideline formalisms. Stud Health Technol Inform 2008;139:22-43. [Medline: 18806319]
53. Iglesias N, Juarez JM, Campos M. Business Process Model and Notation and openEHR task planning for clinical pathway standards in infections: critical analysis. J Med Internet Res 2022 Sep 15;24(9):e29927. [doi: 10.2196/29927] [Medline: 36107480]
54. De Ramón Fernández A, Ruiz Fernández D, Sabuco García Y. Business Process Management for optimizing clinical processes: a systematic literature review. Health Informatics J 2020 Jun;26(2):1305-1320. [doi: 10.1177/1460458219877092] [Medline: 31581880]
55. Rebuge Á, Ferreira DR. Business process analysis in healthcare environments: a methodology based on process mining. Inf Syst 2012 Apr;37(2):99-116. [doi: 10.1016/j.is.2011.01.003]
56. Pufahl L, Zerbato F, Weber B, Weber I. BPMN in healthcare: challenges and best practices. Inf Syst 2022 Jul;107:102013. [doi: 10.1016/j.is.2022.102013]
57. Scheuerlein H, Rauchfuss F, Dittmar Y, et al. New methods for clinical pathways-Business Process Modeling Notation (BPMN) and Tangible Business Process Modeling (t.BPM). Langenbecks Arch Surg 2012 Jun;397(5):755-761. [doi: 10.1007/s00423-012-0914-z] [Medline: 22362053]
58. van der Lubbe SCC, Goh LH, Kontopantelis E, Tam WW, Valderas JM. Measuring guideline concordance via electronic health records: a new model for estimating concordance scores. BMJ Qual Saf 2025;PMID:bmjqs-b2025. [doi: 10.1136/bmjqs-2025-018494]
59. Grüger J, Geyer T, Kuhn M, Braun SA. Eighted violations in alignment-based conformance checking. In: Bergmann R, editor. Process Mining Workshop: Springer. [doi: 10.1007/978-3-031-27815-0_21]
60. Munoz-Gama J, Martin N, Fernandez-Llatas C, Johnson OA, Sepúlveda M. Innovative informatics methods for process mining in health care. J Biomed Inform 2022 Oct;134:104203. [doi: 10.1016/j.jbi.2022.104203] [Medline: 36113758]
61. Augusto A, Deitz T, Faux N, Manski-Nankervis JA, Capurro D. Process mining-driven analysis of COVID-19's impact on vaccination patterns. J Biomed Inform 2022 Jun;130:104081. [doi: 10.1016/j.jbi.2022.104081] [Medline: 35525400]
62. Oliart E, Rojas E, Capurro D. Are we ready for conformance checking in healthcare? Measuring adherence to clinical guidelines: a scoping systematic literature review. J Biomed Inform 2022 Jun;130:104076. [doi: 10.1016/j.jbi.2022.104076] [Medline: 35525401]
63. Stan O, Miclea L. Local EHR management based on FHIR. Presented at: 2018 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR); May 24-26, 2018. [doi: 10.1109/AQTR.2018.8402719]
64. Tabari P, Costagliola G, De Rosa M, Boeker M. State-of-the-art Fast Healthcare Interoperability Resources (FHIR)-based data model and structure implementations: systematic scoping review. JMIR Med Inform 2024 Sep 24;12(1):e58445. [doi: 10.2196/58445] [Medline: 39316433]
65. von Dincklage F, Bublitz VK, Kumpf O, et al. Computer-interpretable quality indicators for intensive care medicine: development and validation study. J Med Internet Res 2025 Sep 26;27:e77077. [doi: 10.2196/77077] [Medline: 41004802]
66. Kahol K, Vankipuram M, Patel VL, Smith ML. Deviations from protocol in a complex trauma environment: errors or innovations? J Biomed Inform 2011 Jun;44(3):425-431. [doi: 10.1016/j.jbi.2011.04.003] [Medline: 21496496]

Abbreviations

BPMN: Business Process Model and Notation
CDSS: Clinical Decision Support Systems
CIG: computer-interpretable guideline
CP: clinical pathway
CPDI: Clinical Pathway Deviation Index
CPG: clinical practice guideline
CQL: Clinical Quality Language
ECA: Event-Condition-Action
eMR: electronic medical record
FHIR: Fast Healthcare Interoperability Resources
HL7: Health Level Seven
NRL: Natural Rule Language
OCL: Object Constraint Language
OWL: Semantic Web Ontology Language
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: An Extension of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Statement
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
SWRL: Semantic Web Rule Language

XSL•FO
RenderX

Edited by S Brini; submitted 01.Jul.2025; peer-reviewed by FV Dincklage, G Lichtner, X Huang; accepted 27.Nov.2025; published 09.Feb.2026.

Please cite as:
Hoang MT, Donnelly C, Igasto C, Shetty A, Pradhan M, Shaw T
Data-Driven Guideline Adherence in Data Representation and Compliance Measurement: Scoping Review
J Med Internet Res 2026;28:e79937
URL: https://www.jmir.org/2026/1/e79937
doi:10.2196/79937

© Minh Trang Hoang, Candice Donnelly, Christina Igasto, Amith Shetty, Malcolm Pradhan, Tim Shaw. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effects of Digital Health Interventions to Promote Safer Sex Behaviors Among Youth: Systematic Review and Bayesian Network Meta-Analysis

Yiran Zhu[1*], MSc; Wenwen Peng[1*], MSc; Die Hu[1], BN; Edmond Pui Hang Choi[2], PhD; Maritta Anneli Välimäki[3,4], PhD; Ci Zhang[1], PhD; Xianhong Li[1,5], PhD

[1]Xiangya School of Nursing, Central South University, 172 Tongzipo Road, Changsha, China
[2]School of Nursing, University of Hong Kong, Hong Kong SAR, China (Hong Kong)
[3]School of Public Health, Faculty of Medicine, University of Helsinki, Helsinki, Finland
[4]Helsinki University Hospital, Nursing Research Center (NRC), Helsinki, Finland
[5]JBI Xiangya Research Centre for Evidence-based Healthcare Innovation, Central South University, Changsha, China
[*]these authors contributed equally

Corresponding Author:
Xianhong Li, PhD
Xiangya School of Nursing, Central South University, 172 Tongzipo Road, Changsha, China

Abstract

Background: Youth aged 15 - 24 years carry a disproportionate HIV/sexually transmitted infections (STIs) burden. In recent years, different modalities of digital health interventions (DHIs) have been explored to promote safer sex behaviors among youth, but their comparative effectiveness across modalities and relative to nondigital interventions (NDIs) remains unclear.

Objective: This study aimed to compare DHI modalities on safer sex behaviors and HIV/STI incidence, rank modalities using Bayesian network meta-analysis (NMA), and position their effectiveness relative to NDIs.

Methods: A systematic review and Bayesian NMA of randomized controlled trials were conducted by comprehensively searching PubMed, EMBASE, Web of Science, and Cochrane Library (inception to November 2025). Eligible studies were those that enrolled youth aged 15 - 24 years and evaluated mobile app-based intervention, telecommunication-based intervention (TCI), static web-based intervention (SWI), or interactive online-based intervention (IOI)—with an NDI or another DHI. Primary outcomes were condom use at last sexual contact, consistent condom use, and proportion of condom use. Secondary outcomes included condom use self-efficacy, number of sexual partners, and STI incidence (including HIV). Risk of bias was assessed with the Cochrane Risk of Bias 2 tool, and certainty of evidence with GRADE/CINeMA (Confidence in NMA). Bayesian random-effects NMAs estimated odds ratios (ORs) with 95% credible intervals (CrIs), and complementary frequentist NMAs provided 95% CIs and 95% prediction intervals.

Results: Twenty-four randomized controlled trials (20,134 participants) were included, forming treatment networks across 5 intervention types. TCI was the only intervention that significantly improved condom use at last sex compared with NDI (OR 1.13, 95% CrI 1.02 - 1.26). For consistent condom use, SWI and IOI outperformed TCI (SWI vs TCI: OR 1.77, 95% CrI 1.03 - 3.06; IOI vs TCI: OR 1.68, 95% CrI 1.02 - 2.76). For the proportion of condom use, IOI outperformed SWI (OR 1.34, 95% CrI 1.01 - 1.80), and mobile app-based intervention ranked highest in probability rankings, though estimates lacked precision. For STI incidence, NDI was associated with fewer STIs than SWI (OR 0.61, 95% CrI 0.46 - 0.82).

Conclusions: This is the first NMA to compare the effectiveness of DHIs on condom use and HIV/STI outcomes among youth populations. It demonstrates that the impact of DHIs on HIV prevention varies substantially by intervention modality and outcome type. While TCI demonstrates the most consistent improvement in condom use at last sex, SWI and IOI may be more effective for promoting consistent condom use, though estimates remain imprecise. However, wide prediction intervals and low-certainty evidence suggest that self-reported behavioral changes may not translate into reductions in HIV/STI incidents without integration with offline services and broader structural support. Future trials might consider including standardized outcome indicators and longer follow-up to generate more precise estimates of the effectiveness of DHIs and guide generalization of youth-centered digital HIV/STIs prevention.

Trial Registration: PROSPERO CRD42024527317; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024527317

(J Med Internet Res 2026;28:e87071) doi:10.2196/87071

XSL•FO
RenderX

KEYWORDS

digital health intervention; HIV prevention; safer sex behavior; youth; mobile health; telecommunication-based intervention; web-based intervention; network meta-analysis; mHealth

Introduction

Adolescents and young adults aged 15 - 24 years, defined as "youth" by the United Nations [1-3], are disproportionately affected by HIV around the globe. This age range is widely used in international health research and reporting, which allows comparability across studies and alignment with global HIV surveillance data. Alarmingly, in 2023, youth in this age group comprised nearly one-third of the 3600 daily new HIV infections recorded worldwide. Youth are especially vulnerable to HIV due to high rates of unprotected sex, inconsistent condom use, and co-occurring risk behaviors such as alcohol and drug use [4,5]. Still, a significant proportion of youth around the world lack access to accurate and age-appropriate information on sexual and reproductive health, rendering them susceptible to misinformation, psychological distress, and engagement in high-risk sexual behaviors [6]. To address this public crisis, scalable evidence-based health interventions targeting safer sex practices should be prioritized in this vulnerable population [7].

Digital health interventions (DHIs) have emerged as a promising strategy for health promotion in recent years [8]. Digital health, conceptualized as an umbrella term by the World Health Organization [9], refers to the use of digital and wireless platforms to facilitate health care delivery or health interventions, including but not limited to electronic health, mobile health, telehealth, and artificial intelligence-based applications. On the other hand, with the growing accessibility of smartphones and internet services among youth, digital technologies have become a dominant force to shape their sexual behaviors [10,11]. It is more convenient for young people to meet sexual partners, including casual, one-night, and anonymous partners, through web-based platforms, dating apps, and social networking sites [12,13], which further increases the likelihood of having unprotected sex frequency [14]. While digital technologies have facilitated riskier sexual behaviors among youth, they also create opportunities for DHIs that leverage young people's existing online engagement patterns and preferences [15-19].

Accumulating evidence suggests that DHIs can improve HIV-related knowledge, risk perception, prevention intentions, and behavioral outcomes among youth [20], with the types of DHIs including mobile apps, text messaging, online videos, social media platforms, and interactive websites [21-23]. A stage-based computer-delivered intervention, for example, targeting heterosexual young men demonstrated significant improvements in condom use intention and subsequent condom use behavior [24]. Similarly, a study evaluating a social media-based intervention via Facebook reported a 23% increase in condom use and a 54% reduction in chlamydia incidence among adolescents [25]. In contrast, a large (randomized controlled trial (RCT) delivering sexual health promotion via SMS and email enhanced sexually transmitted infection (STI) knowledge and testing uptake, particularly among women, but showed no significant impact on condom use [26]. Another study reported that intervention based on a peer-led safer-sex Facebook group for Chinese college students found no significant change in contraceptive use intention or frequency [27]. Similarly, a social media-based crowdsourced HIV testing intervention among youth did not increase facility-based HIV testing, condom use, or syphilis testing [28]. Therefore, different types of DHIs may differentially affect sexual health outcomes, yet existing trials rarely distinguish the relative effectiveness of each DHI modality. Clarifying which intervention types are most effective for specific behavioral and biological outcomes is essential for optimizing digital HIV prevention strategies among youth [7,29].

In addition, several systematic reviews (SRs) have synthesized the evidence on DHIs for HIV prevention, but important limitations remain. Some SRs are purely descriptive, lacking quantitative synthesis [22,30-32]. Other reviews have focused narrowly on specific DHI types (eg, social media or telehealth) [33,34], or have failed to examine key behavioral outcomes like condom use [35,36]. In addition, traditional meta-analyses are constrained to pairwise comparisons [37], leaving uncertainty about which types of DHIs are most effective in head-to-head comparisons [30]. To address these knowledge gaps, we conducted a SR and network meta-analysis (NMA) to evaluate and compare the effectiveness of different DHIs in promoting safer sex behaviors among youth. The study aimed to: (1) identify the most effective types of DHIs in promoting safe sex among youth; (2) construct a network-based ranking of intervention effectiveness; and (3) inform the design of scalable, evidence-based digital health programs for HIV prevention among youth.

Methods

Overview

This SR and NMA follow the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA-NMA) guidelines [38]. The completed PRISMA-NMA checklist is provided in Checklist 1. The protocol for this study has been registered with PROSPERO (CRD42024527317).

Eligibility Criteria

Types of Population

Studies were eligible for inclusion if they involved participants aged 15 - 24 years or if at least half of the participants were within this age range. Those that did not report, or for which data could not be extracted, for this specific age group were excluded.

Types of Interventions and Comparison

The interventions included in this review were DHIs, which were defined in accordance with the World Health Organization's broad definition of digital health technologies [9]. For the purpose of this review, the included DHIs were further classified into four mutually exclusive categories based

XSL•FO
RenderX

on their delivery modes and characteristics: (1) mobile app-based interventions (MAIs), (2) telecommunication-based interventions (TCI), (3) static web-based interventions (SWIs), and (4) interactive online-based interventions. This operational classification was developed to reflect the interventions identified in the included studies and to avoid overlap across categories. A detailed description of common DHI subcategories was provided in Table 1; the subcategories of the DHIs were based on a previous SR [30]. The control group received nondigital interventions (NDI), referring to traditional approaches without digital technology, such as face-to-face counseling, printed materials, or group sessions.

Table . Subcategories of digital health interventions (DHIs) and abbreviations used to classify interventions evaluated.

Subcategories of DHIs	Abbreviation	Description
Mobile app–based interventions	MAI	Programs delivered primarily via dedicated software apps installed on smartphones or tablets, leveraging device features (eg, notifications, sensors, data storage) to provide interactive content, personalized feedback, tracking, and behavior change support.
Telecommunication-based intervention	TCI	Interventions using traditional telecommunication methods such as SMS text messages or telephone calls. These interventions typically involve sending reminders, educational messages, or conducting counseling via phone communication without the need for internet-based platforms.
Static web-based interventions	SWI	Interventions provided through websites that offer static, noninteractive content. This may include informational pages, downloadable resources, or educational materials without features for user engagement or real-time feedback.
Interactive online-based interventions	IOI	Interventions delivered via web-based platforms or websites that enable user interaction, such as quizzes, tailored feedback, chatbots, or real-time communication with health professionals. These platforms actively engage users to enhance learning and behavior change.

Any appropriate comparator group was included, such as usual care, placebo, no intervention, waitlist, attention control, or different DHIs. To reduce inconsistency among trials, we excluded studies that combined non-DHIs with DHIs, unless the distinction between the intervention and control groups lay solely in the DHIs. In multi-arm trials, intervention arms representing the same modality without meaningful differences in content, intensity, or delivery were considered a single treatment node for eligibility purposes and later combined analytically to avoid double-counting.

Outcomes

The primary outcomes were specific condom use behaviors, defined as follows:

1. Condom Use Rate in the Last Sexual Contact: The percentage of individuals reporting condom use during their most recent penetrative sexual act [39].
2. Consistent Condom Use Rate: The percentage of individuals reporting consistent condom use during all their penetrative sexual acts over the recall period specified in each study [40].
3. Proportion of Condom Use: The overall proportion of sexual acts in which a condom was used, calculated as the total number of times a condom was used divided by the total number of sexual acts. Unlike the consistent condom use rate, which measures whether individuals always use a condom, this indicator captures the frequency of condom use across all reported sexual encounters, allowing for partial or occasional use [41].

The secondary outcomes included (1) self-efficacy for condom use, measured by the overall mean score on a validated condom use self-efficacy scale, such as Brafford and Beck's [42] condom use self-efficacy scale, Lawrance et al's [43] self-efficacy for HIV prevention scale and others, (2) number of sexual partners, and (3) the incidence rate of STIs (including HIV). Because follow-up length varied substantially across trials, we included studies reporting at least one postintervention follow-up outcome and extracted the longest follow-up time point for synthesis to enhance comparability [44,45].

Types of Studies

Only RCTs were included, including crossover trials and cluster-randomized trials. Studies using nonrandomized, quasi-experimental, observational, or qualitative designs were excluded. Only peer-reviewed articles published in English were eligible, as non-English or non-peer-reviewed sources lack sufficient methodological detail for reliable data extraction and risk-of-bias assessment.

Search Strategy

The search strategy was developed and reported in accordance with the PRISMA-S guideline [46]. Searches for RCTs were conducted in PubMed (including MEDLINE), EMBASE, Web of Science, and the Cochrane Library. Searches were performed

XSL•FO
RenderX

through the native interfaces of each database (PubMed via NCBI, EMBASE via Elsevier, Web of Science via Clarivate, and Cochrane Library via Wiley). In addition, the reference lists of relevant SRs were checked to ensure that no eligible trials were missed.

The search terms were formulated according to the PICOs framework, including participants or populations, interventions, outcomes, and types of research design. Both Medical Subject Headings and free-text terms were included as appropriate. Boolean operators ("AND," "OR") were used to combine search terms, and database-specific search techniques such as truncation, phrase marks, and wildcards were applied. The complete search terms and algorithm were provided in Multimedia Appendix 1, and search strategies for the other databases were adapted accordingly. The search was designed and executed by 2 reviewers (YZ and WP), who were trained in SR methodology, and the strategy was cross-checked for completeness and accuracy.

The literature search was initially conducted on June 13, 2024 and was last updated on November 15, 2025. All retrieved records were imported into EndNote X9 for citation management, and duplicates were removed using both automated and manual deduplication. Additional search methods included checking the reference lists of the included studies or relevant SRs. Gray literature was also searched via Google Scholar, OpenGrey, and ProQuest Dissertations. The search was limited to studies published in English due to resource constraints for translation.

Selection Procedure and Data Extraction

Two reviewers (YZ and DH) independently screened the titles and abstracts against predefined protocol criteria. Full texts were retrieved for all potentially eligible studies. When multiple articles were identified from the same randomized controlled trial, the most recent or most comprehensive publication was retained for data extraction. Earlier reports were used to supplement missing information on study design, intervention details, or outcomes when necessary. Any discrepancies between the 2 reviewers were resolved by discussion. If disagreements persisted, a third reviewer (WP) was invited for adjudication. At the title and abstract screening stage, we excluded 14,788 records that clearly did not meet the eligibility criteria, most commonly because of wrong study design (eg, cross-sectional surveys, qualitative studies, reviews, protocols), wrong population (non-youth samples), ineligible intervention or comparator (ie, the difference between study arms did not lie in the use of a DHI), or an unrelated topic.

Data were extracted using a standardized and piloted form. Extracted variables included: first author, publication year, recruitment region, participant characteristics (mean age, SD, gender distribution), type of intervention and comparator, sample size per arm, intervention duration, and reported endpoints. Detailed characteristics of the DHIs were also extracted to facilitate subcategorization (Table 1). Further, outcomes and corresponding measurement methods were recorded, such as self-reported condom use and validated self-efficacy scales.

When outcome data were incomplete or unclear, study authors were contacted by email for clarification; trials with essential missing data were excluded from the quantitative synthesis and documented in Multimedia Appendix 2. All eligible studies were included in the SR, and only studies with usable and connected outcome data were included in the NMA.

Risk of Bias Assessment

The methodological quality of the included studies was independently assessed by 2 reviewers (YZ and WP), with disagreements resolved by a third reviewer (CZ). Risk of bias was evaluated using version 2 of the Cochrane risk of bias 2 tool (RoB 2) for randomized trials [47]. For each domain, studies were rated as having "low risk," "some concerns," or "high risk" of bias according to the Cochrane Handbook (version 6.5) [48]. Domain-level risk of bias judgments for each trial are summarized in Figure 1, with extended graphs and contribution matrices provided in Multimedia Appendix 3. The reference list of included studies is provided in Multimedia Appendix 4 [26,49-71].

XSL•FO
RenderX

Figure 1. Risk of bias assessment of included randomized controlled trials using Cochrane Risk of Bias 2 tool [26,49-71].

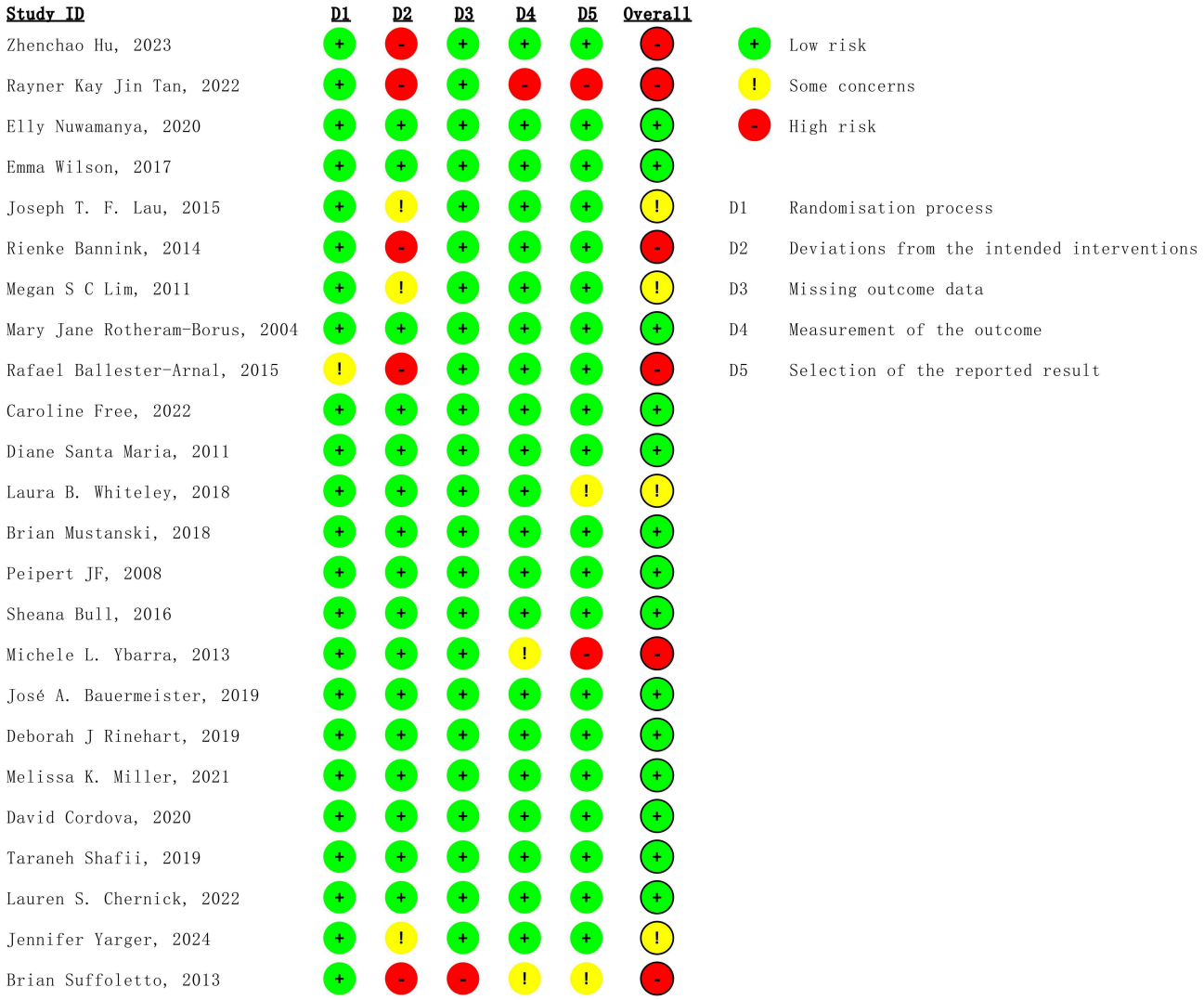

Because blinding was generally not feasible for these nonpharmacological interventions, many trials were judged at high risk of bias in the domain of deviations from intended interventions [72,73]. As this limitation was expected and unlikely to influence objectively measured outcomes, we did not consider this domain when grading the certainty of evidence. The overall certainty of evidence was determined using the CINeMA (Confidence in NMA) web application, which is based on the GRADE framework [74,75]. In addition, we constructed GRADE "Summary of Findings" tables using the official template provided by the GRADE Working Group to summarize the key relative and absolute effects and certainty ratings for the primary outcomes (Multimedia Appendix 3).

RoB 2 assessments were incorporated into the interpretation of NMA findings and into the GRADE/CINeMA evaluation of certainty, but they were not used to weight studies in the statistical synthesis.

Statistical Analysis

Geometry of the Evidence Network

We examined the geometry of the treatment network by mapping each trial arm to one of the predefined intervention nodes and summarizing the pattern of direct comparisons. A network plot was generated to visually depict the evidence base, with node size proportional to the number of randomized participants and edge thickness reflecting the number of trials informing each comparison. We further assessed potential network-related biases by identifying sparse nodes, single-study comparisons, and imbalance in the distribution of direct evidence.

Model Specification and Synthesis Methods

Model convergence was assessed through Markov Chain Monte Carlo diagnostics, including the Gelman-Rubin potential scale reduction factor and inspection of leverage plots. The number of adaptation iterations, burn-in period, and total iterations were set to ensure adequate mixing and convergence. Effect estimates were expressed as pooled odds ratios (ORs) and 95% credible intervals (CrIs), which served as the primary summary measure for all dichotomous outcomes.

Bayesian NMAs were conducted using the R package *BUGSnet* to compare the effectiveness of 4 subcategories of DHIs and control groups. Binomial likelihood models with a logit link function were specified, and both fixed-effect and random-effects consistency models were fitted. Given anticipated clinical heterogeneity, random-effects models were treated as primary, with fixed-effect models used in sensitivity analyses. Noninformative priors were assigned to treatment

XSL•FO
RenderX

effects and heterogeneity parameters to minimize prior influence. Model fit and parsimony were evaluated using the deviance information criterion (DIC), with lower values indicating better fit.

To evaluate the transitivity assumption, we compared mean age, sex distribution, intervention intensity, and follow-up duration across treatment comparisons. We further restricted inclusion to trials in which ≥50% of participants were aged 15 - 24 years, excluded trials in which nondigital components were offered only to one arm, and extracted outcomes at the longest reported follow-up to harmonize follow-up time. These design and population characteristics showed broadly overlapping ranges across interventions and no systematic differences between comparisons, so transitivity was judged plausible. Forest plots of posterior ORs with 95% CrIs from the Bayesian consistency model were generated to summarize the magnitude and uncertainty of estimated treatment effects.

To complement these Bayesian estimates and quantify uncertainty in effects that might be observed in new settings, we also performed frequentist random-effects NMAs using the netmeta package in R, specifying NDI as the reference group [76]. For each outcome, we estimated ORs and 95% CIs for each intervention versus NDI and derived 95% prediction intervals (PIs) by combining the average treatment effect with between-study heterogeneity, in line with recent recommendations that NMAs should routinely report PIs when heterogeneity is present [77].

Assessment of Inconsistency and Heterogeneity

Consistency between direct and indirect evidence was assessed by comparing the DIC between consistency and inconsistency models. A substantially lower DIC in the consistency model indicated acceptable agreement between sources of evidence. Due to the limited number of included studies, we did not formally investigate small-study effects (eg, using funnel plots or Egger's regression test), which are typically used to explore potential publication bias as one of several possible explanations for such effects. Selective outcome reporting could not be formally assessed due to insufficient reporting in the included trials; however, the potential for selective reporting was considered when interpreting the cumulative evidence. Because the number of studies informing most comparisons was limited, local inconsistency (eg, node-splitting) could not be reliably assessed; in the presence of any potential inconsistency, we planned to explore differences in study characteristics and reassess the plausibility of the transitivity assumption.

Handling of Multi-Arm Trials and Node Merging

For multi-arm trials, if 2 or more arms delivered essentially the same intervention category (eg, different versions of the same SWI content without meaningful variation in delivery or timing), we merged these arms by summing the number of events and participants. This ensured that each intervention was represented by a single node in the network and avoided duplicate contributions from the same trial [78].

Ranking of Interventions

Ranking probabilities and surface under the cumulative ranking curve (SUCRA) were computed to summarize the relative effectiveness of each intervention across the posterior distribution. Rankograms and cumulative ranking plots were used to visualize intervention hierarchies, and league tables and heatmaps were generated to present pairwise comparisons and their relative effect estimates. No additional analyses, such as sensitivity analyses, subgroup analyses, or meta-regression, were conducted because the limited number of studies and the sparse network geometry did not permit reliable implementation of these methods. All statistical analyses were conducted using R (version 4.3.2; R Foundation for Statistical Computing) with the gemtc, BUGSnet, and netmeta packages.

Results

Description of Included Studies

From a total of 25,659 records initially retrieved, 24 RCTs published between 2004 and 2024 were included in the final analysis (Figure 2). These studies were conducted across 8 diverse countries, predominantly in the United States (n=14), with the remainder from China (n=2), the United Kingdom (n=2), Uganda (n=2), Singapore (n=1), the Netherlands (n=1), Australia (n=1), and Spain (n=1). The trials collectively enrolled 20,134 participants (range 50 - 6248; mean 838.9, SD 1358.2), with a mean age of 19.5 years. Overall, 10,228 participants (53.4%) were male, although sex composition varied substantially—some studies enrolled only males [49-51], only females [52,53], or mixed populations. Of the 24 included studies, 21 studies were two-arm, and 3 studies were multi-arm. Across the included studies, 6 trials evaluated TCI, 8 assessed interactive online-based intervention (IOI), 6 examined MAI, and 8 investigated SWI. The total number of intervention approaches (n=28) exceeded the number of included studies (n=24) because several trials directly compared 2 or more active interventions (eg, IOI vs SWI) without including a conventional control group. Intervention durations ranged from brief sessions lasting 10 - 20 minutes up to 12 months. Follow-up periods were heterogeneous, spanning from immediate postintervention assessments to 24 months. Most studies reported outcomes at 3 - 6 months, while only a few provided longer-term follow-up beyond 12 months. Five studies performed analyses for more than one time point. To enhance consistency and reduce potential bias associated with short-term variability, we extracted outcomes at the longest follow-up time point, thereby facilitating a more comprehensive evaluation of the intervention's long-term effectiveness [44,45].

XSL•FO
RenderX

Figure 2. The flow diagram of the literatureease clarify th selection process for randomized controlled trials of digital health interventions included in this review.

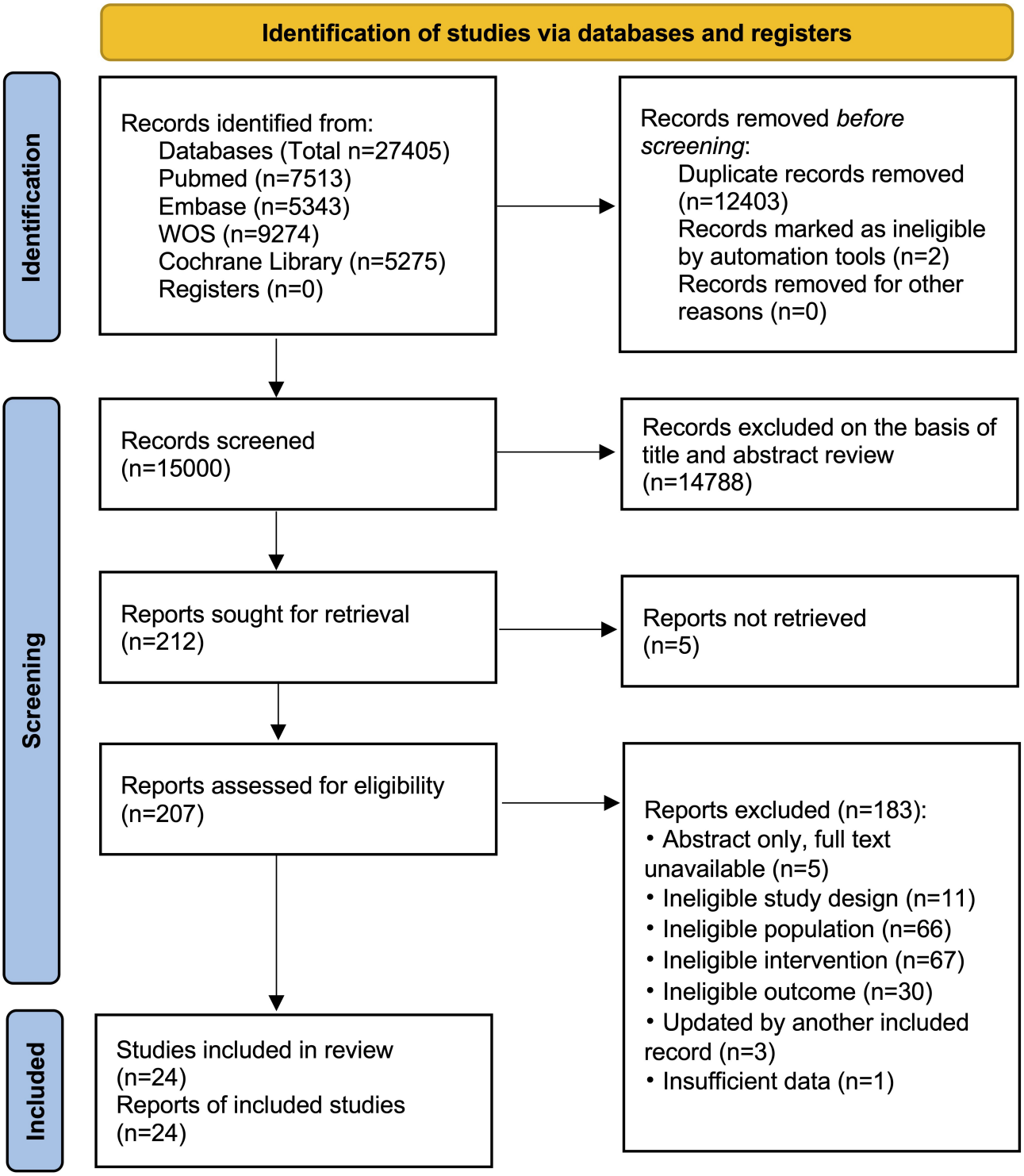

For condom use at last sexual contact, 4 out of 7 studies reported ORs greater than 1, suggesting a possible beneficial effect of the interventions; however, only one trial [54] showed a clear statistical significance (OR 1.13, 95% CI 1.01 - 1.25). Regarding consistent condom use, 6 of 11 studies showed ORs above 1, with substantial heterogeneity; one study reported a very large effect (OR 4.55, 95% CI 1.15 - 17.95) [55]. For the proportion of condom use, 5 out of 6 trials reported ORs above 1, suggesting a tendency towards higher condom use in the intervention groups; however, CIs were wide and often included the null (overall OR range 0.48 - 2.43), indicating that the evidence for this outcome is imprecise. Finally, for STIs incidence, including HIV, effects varied substantially across 7 trials, with ORs ranging from 0.53 to 2.10. Four of the 7 trials had point estimates below 1 [49,50,53,79], and 3 trials had 95% CI that excluded 1 [49,56,57], indicating heterogeneous and partly conflicting evidence for this outcome. These results summarized the observed effects across trials, highlighting that effect estimates varied considerably across outcomes and

XSL•FO
RenderX

studies. The characteristics and effect estimates of the included studies were summarized in Table 2.

XSL•FO
RenderX

Table . Study characteristics and effect estimates of the included trials.

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	Endpoints	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
Hu et al [57]	China	RCT[a]	16.09 (0.84)	—[b]	1691 (53.67)	"You and Me	—	Online education sessions; Internet-based educational platform; cartoon videos; PowerPoint slides	IOI[c] / NDI[d]	1760/1391	45 min sessions/8 wk	End of intervention / 12 wk	①, ④	• 0.72 (0.[illegible]-1.19) ① • 1.42 (1.[illegible]-1.74) ④
Tan et al [50]	Singapore	RCT	23.90 (2.98)	—	300 (100.00)	People Like Us (PLU) web drama video series	Predetermined theory of behavior change	Web-based	SWI / NDI	150/150	6 videos, each about 10 min in length/1 wk	6 mo	②, ④	• 1.22 (0.[illegible]-2.50) ② • 0.82 (0.[illegible]-1.70) ④
Nuwamanya et al [58]	Uganda	RCT	21.00 (2.00)	—	407 (36.60)	MPA-SRH (mobile phone apps-sexual reproductive health)	—	Mobile app	MAI / NDI	556/556	Within 6 mo, open for use to the intervention group.	End of intervention	①	• 1.23 (0.[illegible]-1.18)

XSL•FO
RenderX

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	End-points	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
Wilson et al [56]	The United Kingdom	RCT	23.00 (3.55)	—	846 (41.01)	SH:24 website	—	Website, online service	SWI[e] / NDI	1031/1032	All participants were free to use any other sexual health services or interventions during the trial period/Median =28.8 d (participant-dependent duration)	6 wk	④	• 2.01 (1.40)
Lau et al [51]	China	RCT	—	—	396 (100.00)		—	Online videos (Sent Intervention package email）	SWI / NDI	261/135	10 - 20 min	3 mo	③	• 1.10 (0.97)

XSL•FO
RenderX

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	End-points	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
						Online intervention based on STD-related cognitions involving videos (SC)/online intervention based on both STD-related cognitions and emotions (eg, fear) involving fear-arousing imagery and videos (SCFI)								
Bannink et al [59]	Netherlands	RCT	15.81 (0.68)	—	446 (54.00)	The E-health4Uth Intervention	—	Web-based tailored messages	IOI / NDI	392/434	1 mo	4 mo	②	• 1.59 (0.94-2.70)
Lim et al [26]	Australia	RCT	—	19	417 (41.99)	Email and SMS to a group of young people (intervention gro	—	SMS, email	TCI[f] / NDI	507/486		end of intervention (6 mo) / end of intervention (12 mo)	②	• 0.75 (0.46-1.24)

XSL•FO
RenderX

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	End-points	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
											SMS messages were sent every 3 - 4 wk (a total of 14) emails were sent less than monthly (a total of 8)/12 mo			
Rotheram-Borus et al [60]	The United States	RCT	—	23	136 (77.71)	Telephone intervention	—	Telephone sessions	TCI / NDI	59/116	2 h/6 wk	15 mo	②	• 0.76 (0.40-1.42)
Ballester-Arnal et al [61]	Spain	RCT	20.90 (1.90)	—	60 (25.00)	Fear induction group/website group	Information-Motivation-Behavioral Skills (IMB) Model	Video, music/website; computer based	SWI / NDI	Estimated 34 - 35 participants per group (total n=239; balanced allocation across 7 groups)	1 h	1 months / 4 mo	③	• 1.54 (0.48-5.44)
Free et al [54]	The United Kingdom	Parallel group RCT	20.35 (2.1)	—	2162 (34.60)	text messaging intervention (safetxt)	COM-B (capability, opportunity, motivation, and behavior) model	Delivered by text messages to improve safer sex behaviors	TCI / NDI	3123/3125		4 weeks / 12 mo	①, ④	• 1.13 (0.1-12.5) ① • 1.11 (0.1-12.5) ④

XSL•FO
RenderX

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	End-points	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
											Day 1 - 3: 4 items per day Day 4 - 28: 1 - 2 items per day Second month: 2 - 3 items per week Month 3 - 12: 2 - 5 items per month/12 mo			
Santa Maria et al [62]	The United States	Pilot RCT	21.20 (2.10)	—	56 (52.34)	MYRID (Motivating Youth to Reduce Infection and Disconnection)	Information-Motivation-Behavioral Skills (IMB) Model	Smartphone-based Just-in-Time Adaptive Intervention (JITAI) Android smartphones with unlimited data were provided to participants	MAI[g] / NDI	48/49		end of intervention	②	• 20 (6-53)

XSL•FO
RenderX

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	End-points	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
											Participants will receive a customized message for each EMA assessment completed.• Weeks 1 - 2: 3 times per day• Weeks 3 - 4: 2 times per day• Weeks 5 - 6: 1 time per day• Each EMA took 1 - 5 min to complete/6 wk			
Whiteley et al [55]	The United States	Pilot RCT	18.60 (2.30)	—	37 (61.67)	Online HIV/STI[h] prevention intervention	Information-Motivation-Behavioral Skills (IMB) Model		SWI / NDI	31/29		3 mo	②	• 45 (5-129)

XSL•FO
RenderX

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	End-points	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
								Publicly available websites and YouTube videos related to HIV/STI prevention Content accessible via computer, smartphone, or tablet			Twice per week for 4 wk, each contained 2 - 3 website or video links Total exposure: 8 emails with up to 19 different resources/4 wk			
Mustanski et al [49]	The United States	RCT	23.82[i]	—	901 (100.00)	"Keep It Up!"	Information-Motivation-Behavioral Skills (IMB) model	Fully online (eHealth intervention), delivered via computers and tablets (not available on mobile phones).	IOI / SWI	445/456	Each session lasted about 1 h./≥3 d for core intervention (3 sessions ≥24 h apart); booster sessions at 3 and 6 mo	12 mo	③, ④	• 1.34 (0.90-1.98) ③ • 0.86 (0.56-1.32) ④
Peipert et al [53]	The United States	RCT	—	22	0 (0)	Project PROTECT	Transtheoretical (TTM) model of behavior change	Computer-based multimedia program	IOI / SWI	272/270	Three computer-based sessions /80 d	24 mo	②, ④	• 0.99 (0.70-1.39) ② • 0.96 (0.61-1.53) ④
Bull et al [63]	The United States	RCT	14.94 (1.08)	—	415 (48.71)		Integrated Theory of mHealth	Text messages based on social media	TCI / NDI	436/416	5 and 7 messages weekly/25 wk	end of intervention	③	• 1.13 (0.76-1.53)

XSL•FO
RenderX

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	End-points	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
						Teen Outreach Program (TOP)+ text message program: Youth All Engaged (YAE!)								
Ybarra et al [64]	Uganda	RCT	16.10 (1.40)	—	307 (83.88)	CyberSenga	Information-Motivation-Behavioral Skills (IMB) model	Website based; computer-based	IOI / NDI	183/183	One module per week, a total of 5 modules need to be completed./5 wk	3 mo	②	• 0.91 (0.60-1.16)
Bauermeister et al [65]	The United States	Pilot RCT	21.67 (1.81)	—	123 (100.00)	myDEx(My Desires & Expectations)	The dual processing cognitive-emotional decision-making framework	Online-delivered HIV prevention interventions; App	MAI / SWI	95/28		end of intervention	③	• 2.43 (0.30-19.50)

XSL•FO
RenderX

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	End-points	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
											myDEx intervention includes 6 personalized online courses, completed within 3 mo, with participants logging in an average of about 5 times, and a total conversation volume of about 7 times/3 mo			
Rinehart et al [66]	The United States	Pilot RCT	15.90 (1.60)	—	0 (0)	Texts for Sexual Health Education and Empowerment (t4she)	Health belief model	Text message	TCI / NDI	122/122	58 automated messages sent over 12 wk/12 wk	3 months / 6 mo	①	• 1.40 (0.66-2.96)
Miller et al [67]	The United States	Pilot RCT	16.90 (1.00)	—	26 (28.57)	Sex-Health	Theory of Planned Behavior to inform intervention content and the Social Ecological Model		IOI / NDI	44/47	25 min	6 mo	①	• 0.50 (0.10-2.40)

XSL•FO
RenderX

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	End-points	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
								A tablet-based, interactive intervention: The educator used a tablet to deliver the intervention, intermittently sharing the screen with the participant.						
Cordova et al [68]	The United States	Pilot RCT	18.82 (2.1)	—	4 (8.00)	Storytelling 4 Empowerment (S4E)	An ecodevelopment and empowerment framework	Multi-level Mobile Health App	MAI / NDI	25/25	Complete 3 interactive modules at once /30 - 45 min	1 mo	③	• 0.48 (0.14-1.66)
Shafii et al [69]	The United States	Pilot RCT	21[i]	—	176 (64.71)	e-KISS	Information-Motivation-Behavioral Skills (IMB) model	An interactive computer-based intervention; video	IOI / NDI	130/142	15 - 20 min, a single, one-time interactive /15 - 20 min	2 mo	④	• 0.53 (0.26-1.09)
Chernick et al [52]	The United States	Pilot RCT	17.74 (1.27)	—	0 (0)			Multimedia text messaging; video; mobile-based	IOI / NDI	72/74		3 mo	①, ②	• 0.95 (0.40-2.21) ① • 1.53 (0.56-4.25) ②

XSL•FO
RenderX

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	End-points	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
						Dr. Erica (Emergency Room Interventions to improve the Care of Adolescents)	Intervention mapping, a program planning framework; the Social Cognitive Theory and Motivational Interviewing				A minimum of 56 and maximum of 121 texts, with additional texts sent based on keywords./3 mo			
Yarger et al [70]	The United States	A Cluster Randomized Trial	15.7[i]	—	340 (40.72)	In the Know—an in-person, group-based sexual health education program integrating digital technologies,	—	Technology-based; mobile-based; app	MAI / NDI	348/487	1.5 h/4 wk	3 mo	②	• 0.86 (0.[illegible]-1.27)
Suffoletto et al [71]	The United States	Pilot RCT	21.44 (2.04)	—	0 (0)	SMS program	The Health Belief Model; the Information Motivation Behavior model	Text message	TCI / NDI	23/29	Once a week, send a text message at noon every Sunday/12 wk	3 mo	①, ②	• 1.86 ([illegible]-7.12) ① • 1.60 ([illegible]-6.66) ②

[a]RCT: randomized controlled trial.

[b]Not available.

[c]IOI: interactive online-based intervention.

[d]NDI: nondigital intervention.

[e]SWI: static web-based intervention.

[f]TCI: telecommunication-based intervention.

XSL•FO
RenderX

[g]MAI: mobile app-based intervention.

[h]STI: sexually transmitted infection.

[i]The included studies did not report SD values for these mean estimates, and the SDs cannot be derived from the available information.

Risk-of-bias assessments using the RoB 2 tool are summarized in Figure 1. Overall, most trials were judged to be at low risk of bias for the randomization process, outcome measurement, and selection of the reported result. However, a substantial minority of studies had some concerns or high risk of bias in at least one domain, most frequently for deviations from the intended interventions and missing outcome data. Consequently, several trials were rated as having some concerns or a high overall risk of bias.

Although condom use self-efficacy and number of sexual partners were prespecified as secondary outcomes, too few studies reported these measures to allow meta-analysis. Only one trial evaluated condom use self-efficacy. In Rinehart et al [66], this construct was assessed using 3 items developed within the Health Belief Model (range 0 - 12; Cronbach α=0.72). At the 3rd month, the intervention group reported significantly higher self-efficacy scores than the control group (7.38 vs 6.68; *P*=.04), but this difference was no longer significant at the 6th month (7.39 vs 6.99; *P*=.20). Two trials reported on the number of sexual partners. In Shafii et al [79], participants in the intervention arm reported a 29% reduction in the number of sexual partners at follow-up, although the effect did not reach statistical significance (IRR 0.71, 95% CI 0.50 - 1.03, *P*=.07). Changes in the control group were not reported. By contrast, Free et al [54] examined the proportion of participants reporting 2 or more sexual partners over 12 months. At one year, this outcome was reported by 56.9% of intervention participants compared with 54.8% of controls (OR 1.11, 95% CI 1.00 - 1.24, *P*=.06). Overall, the evidence on the impact of digital interventions on the number of sexual partners remains limited and inconsistent.

Results of Network Meta-Analysis

Overview

A total of 24 RCTs were included to evaluate the comparative effectiveness among 5 intervention types—4 DHIs (TCI, IOI, MAI, and SWI) and NDI—across the four analyzable outcomes: (1) condom use at last sexual contact, (2) consistent condom use, (3) overall proportion of condom use, and (4) incidence of STIs (including HIV). The remaining 2 outcomes of self-efficacy were excluded due to insufficient network connectivity. The network structures for each outcome were shown in Figure 3, where the thickness of the lines was proportional to the number of comparisons, and the size of the nodes reflected the number of studies involving each intervention. Across outcomes, the treatment network was dominated by comparisons of each DHI category versus NDI, whereas head-to-head trials comparing different DHIs were rare. Several DHI-DHI contrasts and some STI outcomes were informed by only one or two small trials, and self-efficacy outcomes formed disconnected subnetworks. Thus, the network geometry was relatively sparse and heavily anchored on NDI, implying that several treatment rankings rely mainly on indirect evidence. The indirect comparative effectiveness of DHIs was summarized in Multimedia Appendix 5. Forest plots of posterior ORs with 95% CrIs for each intervention versus NDI across all 4 outcomes are provided in Multimedia Appendix 6.

XSL•FO
RenderX

Figure 3. Network structure diagrams for randomized controlled trials of digital health interventions among youth, by outcome: (**A**) Condom use rate in the last sexual contact; (**B**) Consistent condom use rate; (**C**) Proportion of condom use; (**D**) The incidence rate of sexually transmitted infections (including HIV). The thicknesses of the lines were proportional to the number of comparisons; the diameters of the circles were proportional to the number of treatments. IOI: interactive online-based intervention; MAI: mobile app-based intervention; NDI: nondigital intervention; SWI: static web-based intervention; TCI: telecommunication-based intervention.

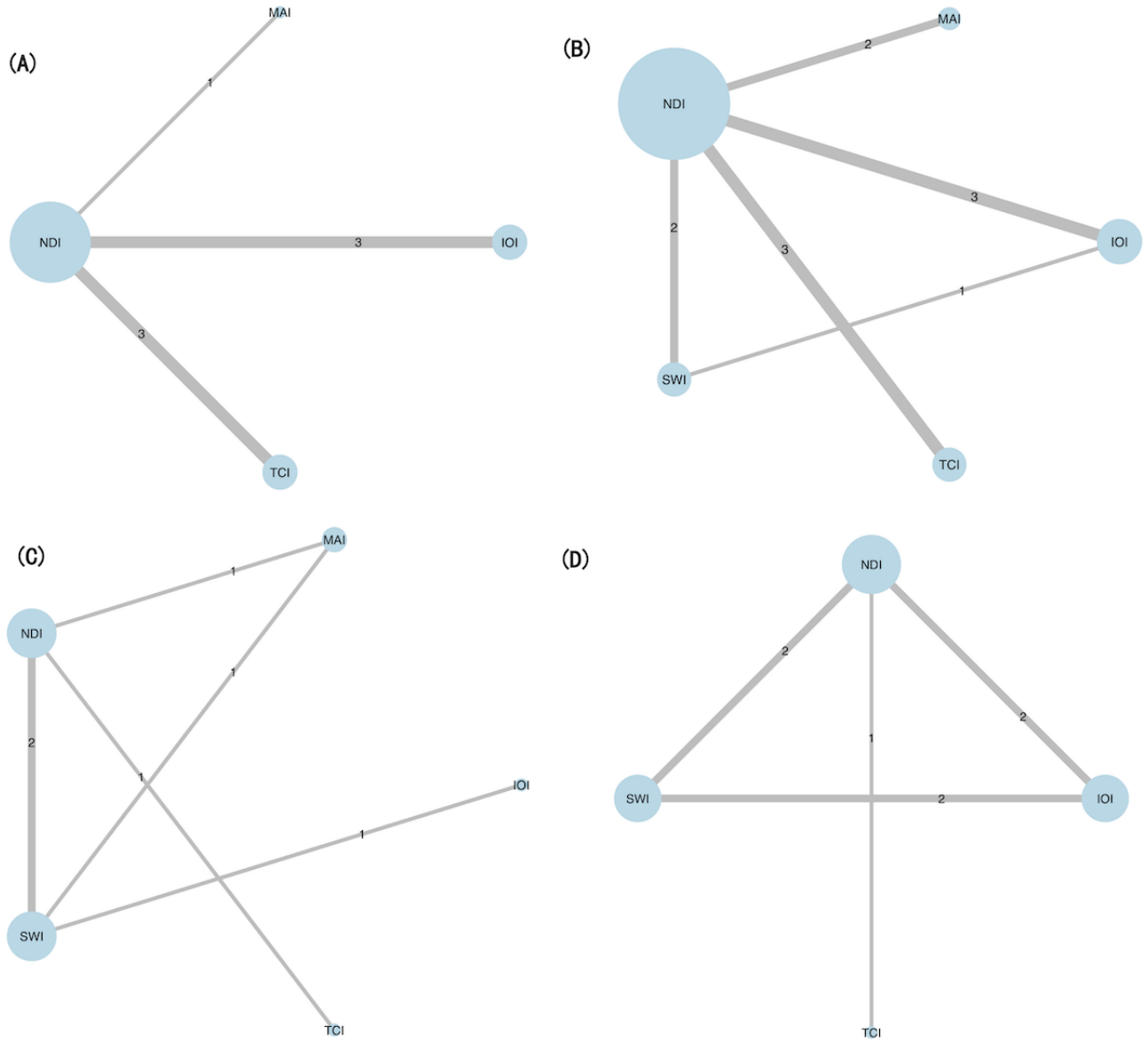

In the complementary frequentist random-effects NMAs, point estimates and 95% CIs for each intervention versus NDI were broadly consistent with the Bayesian results (Multimedia Appendix 7). Across all 4 outcomes, 95% PIs were noticeably wider than the corresponding CIs and frequently included the null value, even when the average effects suggested benefit. For example, for condom use at last sexual contact and for consistent condom use, TCI, IOI, and MAI tended to favor improved condom use versus NDI, but their PIs indicated that future trials conducted in different settings could plausibly observe smaller benefits or no clear difference from NDI. Similar patterns were observed for the proportion of condom-protected acts and for STI incidence, highlighting that between-study heterogeneity and contextual differences may lead to substantial variability in the effects realized in new populations.

Condom Use Rate in the Last Sexual Contact

Seven studies involving 4 DHIs with a total of 10,285 participants were included in the analysis of condom use at last sexual contact. The random-effects consistency model was selected based on model fit, as it showed comparable DIC and residual deviance values to the inconsistency model, indicating no substantial inconsistency. Among the interventions, only TCI showed a statistically significant improvement compared with NDI (OR 1.13, 95% CrI 1.02 - 1.26). Although MAI had the highest SUCRA value (83.44%) and was most likely to rank first (65.61%), its effect was not statistically significant. The rank probabilities for all interventions were summarized in Table 3A and illustrated in Figures 4 and 5, showing the descending order of MAI, TCI, NDI, and IOI. As shown in Multimedia Appendix 6, TCI was the only intervention with its 95% CrI entirely to the right of the line of no effect, suggesting a modest but relatively certain increase in condom use at last

XSL•FO
RenderX

sex compared with NDI. IOI and MAI showed point estimates on either side of 1 with wide CrIs, indicating no clear difference from NDI.

Table . Rank probabilities and surface under the cumulative ranking curve (SUCRA) values for digital health intervention categories in the network meta-analysis of randomized controlled trials assessing sexual health outcomes among youth.

Rank	IOI	MAI	NDI	SWI	TCI
(A) rank probability of condom use rate in the last sexual contact					
1	5.76	65.61	0.05	—[a]	28.58
2	12.02	87.34	8.58	—	92.05
3	20.34	97.38	82.46	—	99.82
4	100	100	100	—	100
SUCRA	12.71	83.44	30.36	—	73.48
(B) Rank probability of consistent condom use rate					
1	33.07	4.59	0.54	61.09	0.7
2	90.77	10.95	3.35	92.76	2.16
3	97.78	41.8	52.43	97.71	10.27
4	99.59	77.38	94.65	99.5	28.87
5	100	99.99	99.99	100	100
SUCRA	80.3	33.68	37.74	87.77	10.5
(C) Rank probability of proportion of condom use					
1	12.6	69.18	0.8	0.04	17.39
2	65.17	88.68	6.82	2.82	36.53
3	92.3	94.87	24.8	37.75	50.31
4	99.3	97.89	65.6	74.13	63.11
5	100	100	100	100	100
SUCRA	67.34	87.66	24.5	28.68	41.84
(D) Rank probability of the incidence rate of STIs (including HIV)					
1	0.38	—	91.78	0.04	7.81
2	7	—	99.95	0.43	92.63
3	95.79	—	100	4.5	99.72
4	100	—	100	100	100
SUCRA	34.39	—	97.25	1.66	66.72

[a]Not available.

XSL•FO
RenderX

Figure 4. Rank of probabilities of digital health intervention categories in the network meta-analysis of randomized controlled trials assessing sexual health outcomes among youth: (**A**) Condom use rate in the last sexual contact; (**B**) Consistent condom use rate; (**C**) Proportion of condom use; (**D**) The incidence rate of sexually transmitted infections (including HIV). Stacked bars show the probability that each digital health intervention category occupies each possible rank (from best to worst).

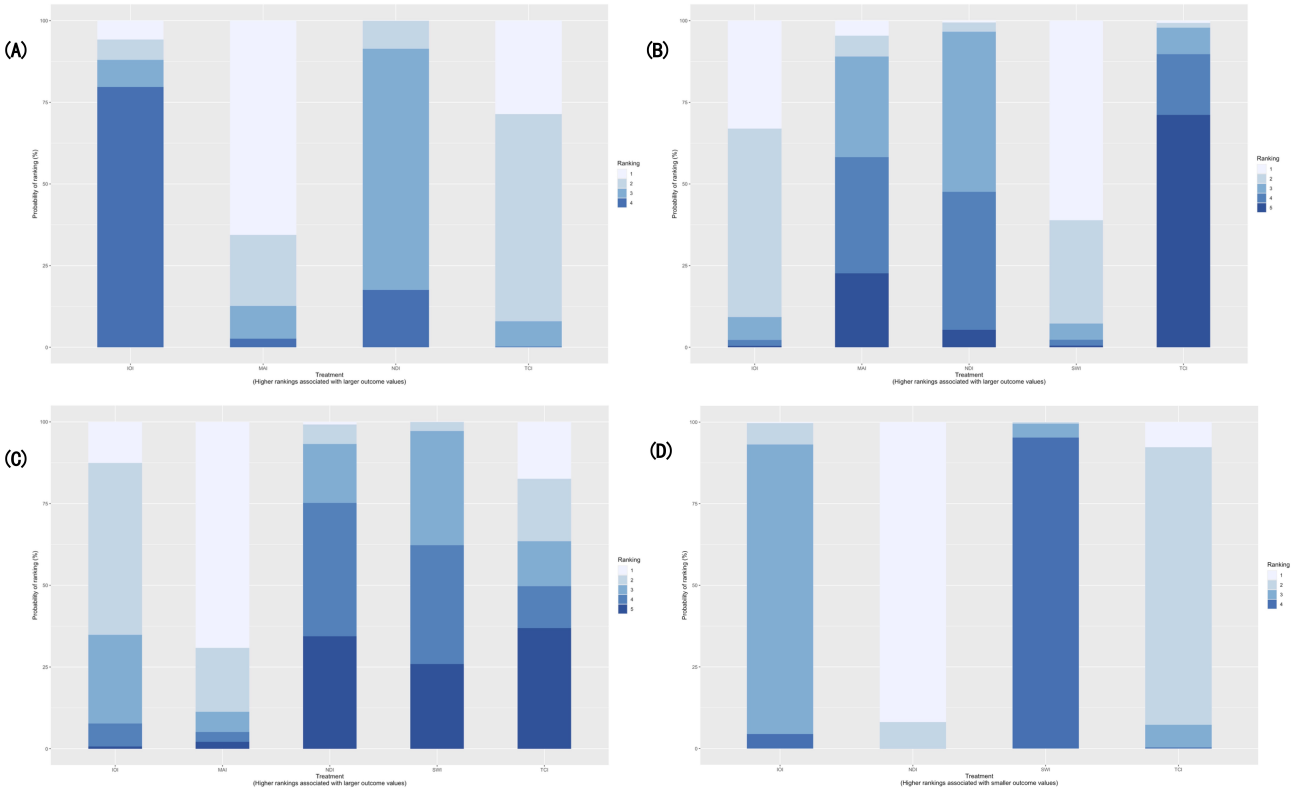

Figure 5. Cumulative rank plot for digital health intervention categories in the network meta-analysis of randomized controlled trials assessing sexual health outcomes among youth: (**A**) Condom use rate in the last sexual contact; (**B**) Consistent condom use rate; (**C**) Proportion of condom use; (**D**) The incidence rate of sexually transmitted infections (including HIV). Lines represent the cumulative probability of each digital health intervention category being ranked at or above each position.

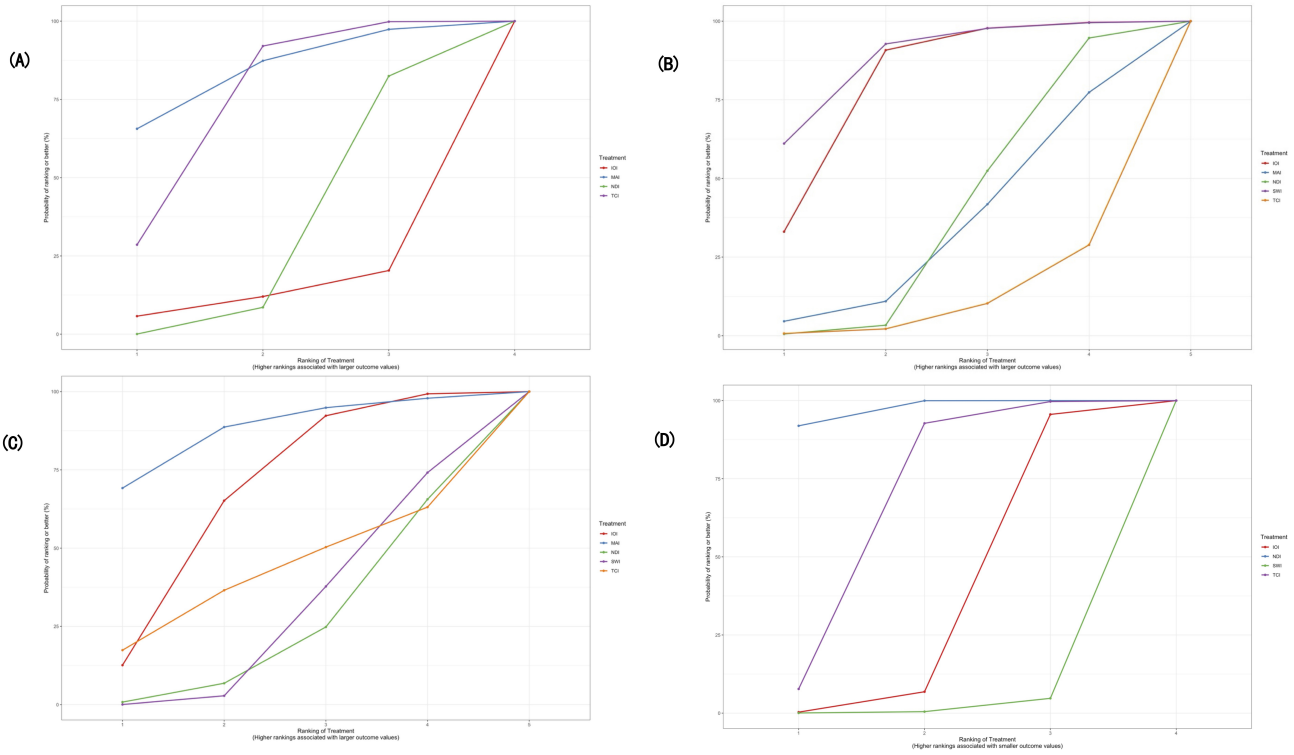

XSL•FO
RenderX

Consistent Condom Use Rate

Eleven studies involving all 5 DHIs with a total of 4881 participants were included in the analysis of consistent condom use. A random-effects consistency model was selected based on a slightly better model fit (DIC=39.49 vs 41.17) and no substantial evidence of inconsistency. Both SWI and IOI were significantly more effective than TCI (SWI vs TCI: OR 1.77, 95% CrI 1.03 - 3.06; IOI vs TCI: OR 1.68, 95% CrI 1.02 - 2.76). SWI had the highest probability of being the most effective intervention (SUCRA=87.77%), followed by IOI (80.3%). TCI ranked the lowest (10.5%), while MAI and NDI showed moderate effectiveness. The ranking probabilities were summarized in Table 3B and visualized in Figures 3B and 4B. Multimedia Appendix 6 shows that IOI and SWI had posterior ORs above 1, implying a tendency toward improved consistent condom use, whereas MAI and TCI showed ORs close to or below 1. However, all CrIs crossed 1, suggesting that these differences were uncertain.

Proportion of Condom Use

Six studies involving 5 DHIs with a total of 2048 participants were included in the analysis of the proportion of condom use. Given the slightly lower DIC (22.45 vs 22.83) and similar model complexity and fit, the consistency model was deemed preferable. Only IOI showed a statistically significant improvement compared with SWI (OR 1.34, 95% CrI 1.01 - 1.80). Regarding the ranking probabilities (Table 3C), MAI had the highest probability of being the most effective intervention (SUCRA=87.66%), followed by IOI (SUCRA=67.34%). In contrast, NDI (SUCRA=24.5%) and SWI (SUCRA=28.68%) ranked relatively low. The distribution of rank probabilities was presented in Table 3C and visualized in Figures 3C and 4C. In Appendix 4C, all interventions showed ORs >1 relative to NDI, with MAI and IOI having the largest point estimates. Nevertheless, the CrIs were wide and crossed 1, indicating that although the direction of effect generally favored digital interventions, the precision of the estimates was limited.

The Incidence Rate of STIs (Including HIV)

Seven studies involving 4 DHIs and a total of 14,966 participants were included in the analysis of STIs incidence. NDI was significantly more effective than IOI (OR 0.78, 95% CrI 0.65 - 0.93) and SWI (OR 0.61, 95% CrI 0.46 - 0.82), while TCI also showed a significant advantage over SWI (OR 0.67, 95% CrI 0.49 - 0.92). Based on rank probabilities and SUCRA values, NDI had the highest likelihood of being the most effective intervention (91.78% probability of ranking first; SUCRA=97.25%), followed by TCI (7.81%; SUCRA=66.72%). IOI and SWI had considerably lower rankings, with SUCRA values of 34.39% and 1.66%, respectively. Table 3D summarizes the intervention rankings, and Figures 3D and 4D illustrate the rank probability and cumulative ranking plots. Multimedia Appendix 6 displays ORs >1 for IOI, SWI, and TCI compared with NDI, and all CrIs lie entirely to the right of 1. This pattern suggests that these digital interventions were associated with equal or higher STI incidence, with SWI showing the highest point estimate.

Consistency and Visualization

For all 4 outcomes, consistency between direct and indirect comparisons was assessed by comparing the DIC values between the consistency and inconsistency models. In all cases, the DIC differences were less than 5, indicating no evidence of global inconsistency (Multimedia Appendix 8). In complementary frequentist random-effects NMAs conducted with the netmeta package in R, we estimated ORs with 95% CIs and 95% PIs for each intervention versus NDI for all 4 outcomes (Multimedia Appendix 7). Across outcomes, PIs were wider than the corresponding CIs and frequently crossed the null, indicating substantial uncertainty in the effects that might be observed in future implementation settings despite the direction of the average effects.

Strength of Evidence

All of these enrolled studies were RCTs, and the quality of evidence was evaluated by the Cochrane Handbook and graded each potential source of bias as low, high, or some concerns; the details were displayed in Figure 1. We assessed confidence in the results of the NMA using the CINeMA framework. Of the 4 outcomes analyzed, only “consistent condom use” met the criteria for CINeMA assessment. The remaining outcomes were excluded due to insufficient numbers of studies or disconnected network structures. For consistent condom use, certainty of evidence was mainly downgraded for within-study bias and imprecision, resulting in overall ratings of “low” to “very low” confidence. Among the 10 comparisons, 1 (10%) was rated as “very low” and 9 (90%) as “low” certainty. For condom use at last sex, the proportion of condom-protected acts, and STI incidence, the certainty of evidence is less well characterized, but given the sparse data, risk of bias, and wide intervals, these estimates should similarly be interpreted as low certainty. A detailed summary of risk-of-bias judgements, CINeMA assessments, and the corresponding GRADE “Summary of Findings” information for each primary outcome is provided in Figure 1 and Table 4, with full GRADE “Summary of Findings” tables formatted according to the GRADE Working Group template available in Multimedia Appendix 3 to aid interpretation of the magnitude and certainty of the main comparisons.

XSL•FO
RenderX

Table . Summary of confidence in the evidence for consistent condom use, assessed using Confidence in Network Meta-Analysis and GRADE.

Comparison	Number of studies	Within-study bias	Reporting bias	Indirectness	Imprecision	Heterogeneity	Incoherence	Overall confidence rating (GRADE)	Reasons for downgrading
IOI:NDI[a, b]	3	Major concerns	Low risk	No concerns	Major concerns	No concerns	No concerns	Very low	Within - study bias and imprecision
IOI:SWI[c]	1	No concerns	Low risk	No concerns	Major concerns	No concerns	No concerns	Low	Imprecision only
MAI:NDI[d]	2	Some concerns	Low risk	No concerns	Major concerns	No concerns	No concerns	Low	Within - study bias and imprecision
NDI:SWI	2	Some concerns	Low risk	No concerns	Major concerns	No concerns	No concerns	Low	Within - study bias and imprecision
NDI:TCI[e]	3	Some concerns	Low risk	No concerns	Major concerns	No concerns	No concerns	Low	Within - study bias and imprecision

[a]IOI: interactive online-based intervention.

[b]NDI: nondigital intervention.

[c]SWI: static web-based intervention.

[d]MAI: mobile app-based intervention.

[e]TCI: telecommunication-based intervention.

Discussion

Principal Findings

To the best of our knowledge, this is the first NMA to systematically evaluate the comparative effectiveness of different modalities of DHIs on promoting safer sexual behaviors among youth. By simultaneously examining 4 distinct digital modalities and 3 behavioral outcomes, and by incorporating STI incidence (including HIV) as a biological endpoint, this review expands the current evidence base and clarifies which intervention types are better suited for immediate versus sustained behavior change, and highlights the gap between improvements in self-reported safer behaviors and reductions in biological HIV/STI infection. Drawing upon data from 24 RCTs across diverse contexts, our findings offer comprehensive insights to inform future development, optimization, and implementation of DHIs in HIV/STIs prevention, underscore the need for designing multimodal, context-aware digital interventions that integrate behavioral support with access to testing and care services, and provide practical considerations for policymakers, program designers, and digital platform developers who seek to tailor DHIs to youth populations.

Across outcomes, between-study heterogeneity and statistical inconsistency were generally low to moderate, and the network satisfied the assumptions of transitivity and global consistency. However, most comparisons were informed by a small number of trials, many of which had some concerns or a high risk of bias in at least one RoB 2 domain. The complementary frequentist NMAs showed that 95% PIs were typically wide and often crossed the null, even when average effects appeared beneficial. Consistent condom use was the only outcome that met CINeMA requirements, and all network comparisons for this outcome were rated as having low or very low certainty. Together, these features suggest that our estimates reflect uncertain average effects rather than precise predictions for specific programs or settings. These patterns of risk of bias, particularly deviations from intended interventions and missing outcome data, may have led to overestimation of some intervention effects or increased uncertainty in the network estimates and contributed to downgrading the certainty of evidence in our GRADE/CINeMA assessment.

In assessing condom use at the last sexual encounter, TCI emerged as the only approach showing statistically significant improvement compared with NDI. This finding aligns with prior studies reporting absolute increases in condom use among participants receiving SMS or phone-based reminders [80]. The relatively stronger performance of TCI may be attributable to its simplicity and immediacy, directly prompting protective behaviors without requiring advanced digital literacy or prolonged engagement [81]. Previous evidence has also highlighted TCI as one of the more acceptable and widely used forms of digital intervention among young people [30]. Thus, TCIs may offer an immediate behavioral benefit, especially for outcomes tied to the most recent sexual event. Interestingly, while MAI ranked highest in SUCRA probability, its effect did not reach significance, suggesting inconsistency between ranking and statistical evidence. This discrepancy could stem from limited trial numbers, variability in app engagement, or short

XSL•FO
RenderX

intervention duration, which may have constrained the power to detect statistical changes. However, this apparent benefit of TCI is based on a few trials with some concerns or high risk of bias, and the wide PIs suggest that similar effects may not be consistently achievable in all implementation settings.

In this study, consistent condom use is improved more effectively by SWI and IOI than by TCI, with SWI ranking the highest. This finding aligns with prior evidence indicating that web-based and online interventions contributed to a substantial proportion of effective digital interventions [21,30]. First, that SWI outperformed IOI may seem counterintuitive, given the absence of interactive features. However, SWI could deliver standardized, theory-based content in a less distracting, user-driven format, allowing youth to process key prevention messages at their own pace. Moreover, while the IOIs included in this review are mostly delivered via computers or tablets, SWIs—though static—are often accessible on smartphones or distributed through popular platforms such as Facebook or WeChat with text, images, or videos [82]. This accessibility and portability may explain why SWI demonstrates stronger effects on consistent condom use. Second, TCI was less effective than both SWI and IOI for improving consistent condom use. Previous studies have suggested that TCI, especially SMS-based reminders, remain controversial in terms of their long-term effectiveness for HIV prevention behaviors [83]. Therefore, although TCI can effectively prompt immediate behaviors, its brief and repetitive messages may lack the depth and reinforcement required to sustain consistent condom use over time. Nevertheless, most trials contributing to this outcome had some concerns or high risk of bias, and CINeMA rated the certainty of these network estimates as low to very low, so the apparent superiority of SWI and IOI should be considered tentative.

When examining the proportion of condom use, IOI demonstrated a significant advantage over SWI, indicating the added value of interactive engagement. This aligns with prior evidence showing that increases in condom use were significantly associated with the use of tailored strategies [21], feedback provision, and guided navigation in digital interventions. Unlike static websites, IOIs integrated tailored feedback, quizzes, or real-time support, which represent core behavior change techniques such as personalized feedback, problem-solving, and self-regulation strategies [84]. These core behavioral change techniques are particularly effective in influencing situational decisions and negotiations during sexual encounters, thereby enhancing the overall proportion of condom use [85,86]. In other words, consistent condom use reflects the internalization of long-term protective norms, and it could be reinforced by standardized and less distracting formats like SWI, while the proportion of condom use is more sensitive to moment-to-moment decision-making. This divergence in findings—SWI being more effective for consistent use, while IOI excels in overall proportion—highlights that different behavioral outcomes may respond to distinct mechanisms of action. Besides, the evolution of technology-based intervention modes has gradually expanded from web-based formats to SMS and social media [87-89]. This trend further supports the notion that while static formats may effectively reinforce long-term norms, interactive platforms provide additional advantages for immediate behavioral decisions during sexual encounters. Yet the CrIs and PIs for this outcome were wide and frequently included the null, indicating considerable heterogeneity and imprecision and implying that any average benefit in the proportion of condom-protected acts may not translate into clear improvements in every context.

Unlike the behavioral outcomes, the effectiveness of DHIs on reducing STI infection reveals a different pattern: NDI and TCI performed more favorably, while IOI and SWI ranked lowest. This finding contrasts with prior studies showing that digital interventions can improve HIV/STI care engagement, such as testing uptake or service use [90-92], highlighting that improvements in care engagement do not necessarily translate into reductions in biological outcomes like STI incidence. Several factors may explain this inconsistency. First, STI incidence represents a distal biological endpoint that may require longer follow-up to capture meaningful reductions, and improvements in self-reported behaviors, such as condom use, may not directly translate into biological protection due to reporting bias or inconsistent application in high-risk contexts. Second, reductions in STI incidence depend not only on safer behaviors but also on timely testing, treatment, and linkage to care. However, evidence shows that stigma related to gender identity, socioeconomic status, race, and ethnicity can delay care-seeking and discourage individuals from accessing necessary services [93]. Nondigital approaches, such as community outreach or peer education programs, often combine behavioral education with direct access to services such as STI testing, treatment linkage, and ongoing support from trained staff or peers, which can directly impact biological outcomes. In contrast, many DHIs focus primarily on education and motivation, without providing structured access to testing or clinical care. Additionally, NDIs may facilitate stronger trust and engagement through in-person interactions, which can overcome barriers related to stigma, confidentiality concerns, or digital literacy limitations. Therefore, while DHIs can effectively change self-reported behaviors, the integrated, multi-component structure of NDIs may explain their relative advantage in reducing actual STI incidence (including HIV). Besides, in this study, TCI is the only digital intervention showing a relatively favorable effect on reducing STI incidence, second only to NDI. Prior studies have demonstrated that TCI can facilitate access to HIV prevention services for youth and achieve high patient and provider satisfaction [93,94]. By providing a comfortable, judgment-free platform, telemedicine may be particularly preferred by marginalized populations, especially transgender youth [95,96]. Given the small and heterogeneous evidence base, important risks of bias in several trials, and wide PIs, these findings on STI incidence should be regarded as exploratory and hypothesis-generating rather than definitive.

Our use of PIs further illustrates the extent to which the observed benefits of DHIs may vary across settings. For most comparisons, the 95% PIs were wide and often crossed the null value, even where the corresponding credible or CIs suggested modest advantages over NDI. This pattern indicates that, although certain DHI modalities tend to improve condom use

XSL•FO
RenderX

on average, implementation in new populations or health systems may yield smaller effects or no clear benefit, underscoring the need for careful adaptation, monitoring, and evaluation when scaling up digital prevention programs.

Implications

These findings have several implications. First, the differential effectiveness of DHIs suggests tailoring intervention types to targeted outcomes. TCI showed immediate benefits for condom use at last sexual contact and a relative advantage for STI incidence, indicating that simple, low-burden interventions may prompt rapid behavior change and reach marginalized youth. In contrast, SWI and IOI were more effective for consistent and habitual condom use, highlighting the value of self-paced content and behavior change techniques that enhance motivation and self-regulation. Multi-modal approaches combining these strengths may maximize overall effectiveness. Second, DHIs alone may be insufficient to reduce STI incidence. Biological outcomes require longer follow-up, and self-reported behavior change does not always translate to infection reduction. Integrating DHIs with offline services, such as condom distribution, PrEP promotion, routine testing, and clinical linkage, is likely necessary to achieve meaningful improvements. Third, these results have implications for intervention design and digital health policy. When targeting youth populations, accessibility, acceptability, and engagement should be prioritized. For example, interventions delivered via mobile platforms or telecommunication may overcome barriers related to stigma or limited digital literacy, while interactive online content can leverage BCTs to support skill acquisition and habitual behavior change. Intervention planners should also consider the balance between immediacy and sustainability of effects: brief, repeated prompts may drive immediate behavior, whereas structured, self-paced content may reinforce long-term habits.

Limitations

Several limitations of this study should be noted. First, substantial clinical and methodological heterogeneity across trials—including differences in intervention intensity and duration, digital platforms, follow-up periods, and outcome definitions—may have contributed to between-study heterogeneity and could challenge the transitivity assumption underpinning some indirect comparisons in the NMA. Second, we restricted inclusion to randomized, prevention-focused DHIs among HIV-negative or status-unknown youth and excluded nonrandomized studies and trials targeting HIV-positive adolescents; the findings may therefore not generalize to these subgroups or to broader digital programs. Third, all behavioral outcomes relied on self-report, and secondary outcomes such as self-efficacy and number of sexual partners were too sparsely and inconsistently reported to be synthesized, limiting our ability to evaluate the broader psychosocial impact of DHIs beyond condom use. Fourth, the number of trials assessing certain interventions, particularly MAI, was limited, which may reduce statistical power and the precision of effect estimates. Fifth, we were unable to formally assess small-study or publication bias, and selective nonpublication or outcome reporting cannot be ruled out. Finally, most network comparisons were rated as low or very low certainty because of within-study bias and imprecision, and PIs, estimated using standard random-effects methods in netmeta, were wide; the true comparative effects may therefore differ meaningfully from our estimates, underscoring the need for rigorous, adequately powered RCTs with standardized outcomes and longer follow-up.

Conclusions

In conclusion, this NMA provides the most comprehensive synthesis to date on the comparative effectiveness of DHIs in promoting safer sexual behaviors among youth. This NMA highlights that the effectiveness of DHIs for HIV prevention among youth depends on both intervention modality and targeted outcomes. While DHIs can enhance knowledge and protective behaviors, their impact on biological endpoints remains limited without integration with offline services and broader structural support. Tailoring interventions to behavioral targets, engagement strategies, and contextual factors is essential to maximize their potential in promoting youth sexual health. We hope that these results will inform the design of youth-centered digital prevention programs, guide clinicians and educators in selecting appropriate modalities, and support policymakers and guideline developers in integrating digital strategies into national HIV prevention frameworks. Future studies should focus on the specific characteristics of patients to provide personalized estimates of comparative effectiveness and individualized predictions regarding the probability of response to treatment and of side effects.

Acknowledgments

YZ and WP contributed equally to this work and are co-first authors. The authors thank all the reviewers for their assistance and support. A generative artificial intelligence tool (ChatGPT, OpenAI) was used solely for language refinement. All artificial intelligence–assisted edits were reviewed, verified, and approved by the authors, who take full responsibility for the final manuscript.

Funding

This research is funded by the China Medical Board (grant number 22- - 465), Hainan Provincial Department of Science and Technology (grant number ZDYF2024SHFZ042), and the National Nature Science Foundation of China (grant number 72574244). The funding sources had no involvement in the study design, data collection, analysis, interpretation of results, or manuscript writing.

XSL•FO
RenderX

Data Availability

The datasets analyzed during this study were derived from previously published studies and are available in the respective articles. All data extracted and analyzed for this systematic review and Bayesian network meta-analysis are included in this published article and its multimedia appendices.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search String.
[DOCX File, 27 KB - jmir_v28i1e87071_app1.docx]

Multimedia Appendix 2
Reference list of excluded studies.
[DOCX File, 50 KB - jmir_v28i1e87071_app2.docx]

Multimedia Appendix 3
Risk of bias, CINeMA assessments, and GRADE Summary of Findings tables.
[DOCX File, 210 KB - jmir_v28i1e87071_app3.docx]

Multimedia Appendix 4
Reference list of included studies.
[DOCX File, 22 KB - jmir_v28i1e87071_app4.docx]

Multimedia Appendix 5
Indirect comparative effectiveness DHIs.
[DOCX File, 19 KB - jmir_v28i1e87071_app5.docx]

Multimedia Appendix 6
Forest plots of posterior ORs with 95.
[DOCX File, 109 KB - jmir_v28i1e87071_app6.docx]

Multimedia Appendix 7
Random‐effects network meta‐analysis of digital health interventions versus non‐digital interventions.
[DOCX File, 19 KB - jmir_v28i1e87071_app7.docx]

Multimedia Appendix 8
Model fit diagnostics for consistency and inconsistency models in the network meta.
[DOCX File, 1102 KB - jmir_v28i1e87071_app8.docx]

Checklist 1
PRISMA-NMA checklist.
[DOCX File, 166 KB - jmir_v28i1e87071_app9.docx]

References

1. Youth. United Nations. URL: https://www.un.org/en/global-issues/youth [accessed 2025-11-01]
2. Rotheram-Borus MJ, Davis E, Rezai R. Stopping the rise of HIV among adolescents globally. Curr Opin Pediatr 2018 Feb;30(1):131-136. [doi: 10.1097/MOP.0000000000000580] [Medline: 29315110]
3. Zhu Q, Wang Y, Liu J, et al. Identifying major drivers of incident HIV infection using recent infection testing algorithms (RITAs) to precisely inform targeted prevention. Int J Infect Dis 2020 Dec;101:131-137. [doi: 10.1016/j.ijid.2020.09.1421] [Medline: 32987184]
4. Barrett M, Laris BA, Anderson P, et al. Condom use and error experience among young adolescents: implications for classroom instruction. Health Promot Pract 2021 May;22(3):313-317. [doi: 10.1177/1524839920935431] [Medline: 32536213]

XSL•FO
RenderX

5. Zhang X, Tang W, Li Y, et al. The HIV/AIDS epidemic among young people in China between 2005 and 2012: results of a spatial temporal analysis. HIV Med 2017 Mar;18(3):141-150. [doi: 10.1111/hiv.12408] [Medline: 27552922]
6. Digitizing sexual health information in Kenya and Peru. World Health Organization. URL: https://www.who.int/news-room/feature-stories/detail/digitizing-sexual-health-information-in-kenya-and-peru [accessed 2025-11-01]
7. Brasileiro J, Queiroz A, Hightow-Weidman LB, Muessig KE. Implementation strategies for Digital HIV prevention and care interventions for youth: a scoping review. Curr HIV/AIDS Rep 2025 Mar 13;22(1):23. [doi: 10.1007/s11904-025-00732-5] [Medline: 40080278]
8. Guiding optimal development and use of digital health towards improved health outcomes. World Health Organization. URL: https://www.who.int/westernpacific/activities/guiding-optimal-development-and-use-of-digital-health-towards-improved-health-outcomes [accessed 2025-11-01]
9. Global Strategy on Digital Health 2020-2025: World Health Organization; 2021. URL: https://www.who.int/publications/i/item/9789240020924 [accessed 2025-03-12]
10. Jennings Mayo-Wilson L, Glass NE, Labrique A, et al. Feasibility of assessing economic and sexual risk behaviors using text message surveys in African-American young adults experiencing homelessness and unemployment: single-group study. JMIR Form Res 2020 Jul 17;4(7):e14833. [doi: 10.2196/14833] [Medline: 32706656]
11. Ippoliti NB, L'Engle K. Meet us on the phone: mobile phone programs for adolescent sexual and reproductive health in low-to-middle income countries. Reprod Health 2017 Jan 17;14(1):11. [doi: 10.1186/s12978-016-0276-z] [Medline: 28095855]
12. Castro Á, Barrada JR, Ramos-Villagrasa PJ, Fernández-Del-Río E. Profiling dating apps users: sociodemographic and personality characteristics. Int J Environ Res Public Health 2020 May 22;17(10):3653. [doi: 10.3390/ijerph17103653] [Medline: 32455986]
13. Watchirs Smith L, Guy R, Degenhardt L, et al. Meeting sexual partners through internet sites and smartphone apps in Australia: national representative study. J Med Internet Res 2018 Dec 18;20(12):e10683. [doi: 10.2196/10683] [Medline: 30563809]
14. Smith LW, Liu B, Degenhardt L, et al. Is sexual content in new media linked to sexual risk behaviour in young people? A systematic review and meta-analysis. Sex Health 2016 Nov 28;13(6):501-515. [doi: 10.1071/SH16037]
15. Cornelius JB, Whitaker-Brown C, Neely T, Kennedy A, Okoro F. Mobile phone, social media usage, and perceptions of delivering a social media safer sex intervention for adolescents: results from two countries. Adolesc Health Med Ther 2019;10:29-37. [doi: 10.2147/AHMT.S185041] [Medline: 31118855]
16. Burns JC, Chakraborty S, Saint Arnault D. Social media preference and condom use behaviors: an analysis of digital spaces with young African American males. Health Educ Behav 2021 Apr;48(2):190-198. [doi: 10.1177/1090198121993043] [Medline: 33703958]
17. Bailey J, Mann S, Wayal S, Abraham C, Murray E. Digital media interventions for sexual health promotion-opportunities and challenges: a great way to reach people, particularly those at increased risk of sexual ill health. BMJ 2015 Mar 3;350:h1099. [doi: 10.1136/bmj.h1099] [Medline: 25736806]
18. Berendes S, Gubijev A, French R, Hickson FCI, Free C. A qualitative study exploring experiences of the safetxt digital health intervention to reduce sexually transmitted infections in young people in the UK. BMJ Open 2023 Oct 24;13(10):e072701. [doi: 10.1136/bmjopen-2023-072701] [Medline: 37879678]
19. Eleuteri S, Toso M. How the smartphone apps can improve your sexual wellbeing. Int J Impot Res 2024 Nov;36(7):781-785. [doi: 10.1038/s41443-023-00730-4] [Medline: 37414872]
20. Bailey JV, Wayal S, Aicken CRH, et al. Interactive digital interventions for prevention of sexually transmitted HIV. AIDS 2021 Mar 15;35(4):643-653. [doi: 10.1097/QAD.0000000000002780] [Medline: 33259345]
21. Xin M, Viswanath K, Li AYC, et al. The effectiveness of electronic health interventions for promoting HIV-preventive behaviors among men who have sex with men: meta-analysis based on an integrative framework of design and implementation features. J Med Internet Res 2020 May 25;22(5):e15977. [doi: 10.2196/15977] [Medline: 32449685]
22. Wadham E, Green C, Debattista J, Somerset S, Sav A. New digital media interventions for sexual health promotion among young people: a systematic review. Sex Health 2019 Apr;16(2):101-123. [doi: 10.1071/SH18127] [Medline: 30819326]
23. Javidi H, Widman L, Lipsey N, Brasileiro J, Javidi F, Jhala A. Redeveloping a digital sexual health intervention for adolescents to allow for broader dissemination: implications for HIV and STD prevention. AIDS Educ Prev 2021 Apr;33(2):89-102. [doi: 10.1521/aeap.2021.33.2.89] [Medline: 33821678]
24. Carvalho T, Alvarez MJ, Pereira C, Schwarzer R. Stage-based computer-delivered interventions to increase condom use in young men. Int J Sex Health 2016 Apr 2;28(2):176-186. [doi: 10.1080/19317611.2016.1158764]
25. Jones K, Baldwin KA, Lewis PR. The potential influence of a social media intervention on risky sexual behavior and chlamydia incidence. J Community Health Nurs 2012;29(2):106-120. [doi: 10.1080/07370016.2012.670579] [Medline: 22536914]
26. Lim MSC, Hocking JS, Aitken CK, et al. Impact of text and email messaging on the sexual health of young people: a randomised controlled trial. J Epidemiol Community Health 2012 Jan;66(1):69-74. [doi: 10.1136/jech.2009.100396] [Medline: 21415232]

XSL•FO
RenderX

27. Sun WH, Wong CKH, Wong WCW, Peer-Led A. A peer-led, social media-delivered, safer sex intervention for Chinese college students: randomized controlled trial. J Med Internet Res 2017 Aug 9;19(8):e284. [doi: 10.2196/jmir.7403] [Medline: 28793980]
28. Tang W, Wei C, Cao B, et al. Crowdsourcing to expand HIV testing among men who have sex with men in China: a closed cohort stepped wedge cluster randomized controlled trial. PLoS Med 2018 Aug;15(8):e1002645. [doi: 10.1371/journal.pmed.1002645] [Medline: 30153265]
29. Evans WD, Ulasevich A, Hatheway M, Deperthes B. Systematic review of peer-reviewed literature on global condom promotion programs. Int J Environ Res Public Health 2020 Mar 27;17(7):2262. [doi: 10.3390/ijerph17072262] [Medline: 32230929]
30. Sewak A, Yousef M, Deshpande S, Seydel T, Hashemi N. The effectiveness of digital sexual health interventions for young adults: a systematic literature review (2010-2020). HEALTH Promot Int 2023 Feb 1;38(1):daac104. [doi: 10.1093/heapro/daac104] [Medline: 36757346]
31. Guse K, Levine D, Martins S, et al. Interventions using new digital media to improve adolescent sexual health: a systematic review. J Adolesc Health 2012 Dec;51(6):535-543. [doi: 10.1016/j.jadohealth.2012.03.014] [Medline: 23174462]
32. Nwaozuru U, Obiezu-Umeh C, Shato T, et al. Mobile health interventions for HIV/STI prevention among youth in low- and middle-income countries (LMICs): a systematic review of studies reporting implementation outcomes. Implement Sci Commun 2021 Nov 6;2(1):126. [doi: 10.1186/s43058-021-00230-w] [Medline: 34742357]
33. Saragih ID, Imanuel Tonapa S, Porta CM, Lee BO. Effects of telehealth interventions for adolescent sexual health: a systematic review and meta-analysis of randomized controlled studies. J Telemed Telecare 2024 Feb;30(2):201-214. [doi: 10.1177/1357633X211047762]
34. Teadt S, Burns JC, Montgomery TM, Darbes L. African American adolescents and young adults, new media, and sexual health: scoping review. JMIR Mhealth Uhealth 2020 Oct 5;8(10):e19459. [doi: 10.2196/19459] [Medline: 33016890]
35. Goldstein M, Archary M, Adong J, et al. Systematic review of mHealth interventions for adolescent and young adult HIV prevention and the adolescent HIV continuum of care in low to middle income countries. AIDS Behav 2023 May;27(Suppl 1):94-115. [doi: 10.1007/s10461-022-03840-0] [Medline: 36322217]
36. Widman L, Nesi J, Kamke K, Choukas-Bradley S, Stewart JL. Technology-based interventions to reduce sexually transmitted infections and unintended pregnancy among youth. J Adolesc Health 2018 Jun;62(6):651-660. [doi: 10.1016/j.jadohealth.2018.02.007] [Medline: 29784112]
37. Buti J, Glenny AM, Worthington HV, Nieri M, Baccini M. Network meta-analysis of randomised controlled trials: direct and indirect treatment comparisons. Eur J Oral Implantol 2011;4(1):55-62. [Medline: 21594220]
38. Hutton B, Salanti G, Caldwell DM, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. Ann Intern Med 2015 Jun 2;162(11):777-784. [doi: 10.7326/M14-2385] [Medline: 26030634]
39. Younge SN, Salazar LF, Crosby RF, DiClemente RJ, Wingood GM, Rose E. Condom use at last sex as a proxy for other measures of condom use: is it good enough? Adolescence 2008;43(172):927-931. [Medline: 19149154]
40. Pallonen UE, Williams ML, Timpson SC, Bowen A, Ross MW. Personal and partner measures in stages of consistent condom use among African-American heterosexual crack cocaine smokers. AIDS Care 2008 Feb;20(2):205-213. [doi: 10.1080/09540120701513669] [Medline: 18293131]
41. Slaymaker E, Zaba B. Measurement of condom use as a risk factor for HIV infection. Reprod Health Matters 2003 Nov;11(22):174-184. [doi: 10.1016/s0968-8080(03)02299-7] [Medline: 14708408]
42. Brafford LJ, Beck KH. Development and validation of a condom self-efficacy scale for college students. J Am Coll Health 1991 Mar;39(5):219-225. [doi: 10.1080/07448481.1991.9936238] [Medline: 1783705]
43. Lawrance L, Levy SR, Rubinson L. Self-efficacy and AIDS prevention for pregnant teens. J Sch Health 1990 Jan;60(1):19-24. [doi: 10.1111/j.1746-1561.1990.tb04771.x] [Medline: 2299814]
44. Goldkuhle M, Bender R, Akl EA, et al. GRADE guidelines: 29. rating the certainty in time-to-event outcomes-study limitations due to censoring of participants with missing data in intervention studies. J Clin Epidemiol 2021 Jan;129:126-137. [doi: 10.1016/j.jclinepi.2020.09.017] [Medline: 33007458]
45. Renson A, Hudgens MG, Keil AP, Zivich PN, Aiello AE. Identifying and estimating effects of sustained interventions under parallel trends assumptions. Biometrics 2023 Dec;79(4):2998-3009. [doi: 10.1111/biom.13862] [Medline: 36989497]
46. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
47. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898. [doi: 10.1136/bmj.l4898]
48. Cochrane Handbook for Systematic Reviews of Interventions: Cochrane; 2024. URL: https://www.cochrane.org/handbook [accessed 2025-02-05]
49. Mustanski B, Parsons JT, Sullivan PS, Madkins K, Rosenberg E, Swann G. Biomedical and behavioral outcomes of keep it up!: an eHealth HIV prevention program RCT. Am J Prev Med 2018 Aug;55(2):151-158. [doi: 10.1016/j.amepre.2018.04.026] [Medline: 29937115]

XSL•FO
RenderX

50. Tan RKJ, Koh WL, Le D, et al. Effect of a popular web drama video series on HIV and other sexually transmitted infection testing among gay, bisexual, and other men who have sex with men in Singapore: community-based, pragmatic, randomized controlled trial. J Med Internet Res 2022 May 6;24(5):e31401. [doi: 10.2196/31401] [Medline: 35522470]
51. Lau JTF, Lee AL, Tse WS, et al. A randomized control trial for evaluating efficacies of two online cognitive interventions with and without fear-appeal imagery approaches in preventing unprotected anal sex among Chinese men who have sex with men. AIDS Behav 2016 Sep;20(9):1851-1862. [doi: 10.1007/s10461-015-1263-z]
52. Chernick LS, Santelli J, Stockwell MS, et al. A multi-media digital intervention to improve the sexual and reproductive health of female adolescent emergency department patients. Acad Emerg Med 2022 Mar;29(3):308-316. [doi: 10.1111/acem.14411] [Medline: 34738284]
53. Peipert JF, Redding CA, Blume JD, et al. Tailored intervention to increase dual-contraceptive method use: a randomized trial to reduce unintended pregnancies and sexually transmitted infections. Am J Obstet Gynecol 2008 Jun;198(6):630. [doi: 10.1016/j.ajog.2008.01.038] [Medline: 18395692]
54. Free C, Palmer MJ, McCarthy OL, et al. Effectiveness of a behavioural intervention delivered by text messages (safetxt) on sexually transmitted reinfections in people aged 16-24 years: randomised controlled trial. BMJ 2022 Sep 28;378(378):e070351. [doi: 10.1136/bmj-2022-070351] [Medline: 36170988]
55. Whiteley LB, Brown LK, Curtis V, Ryoo HJ, Beausoleil N. Publicly available internet content as a HIV/STI prevention intervention for urban youth. J Prim Prev 2018 Aug;39(4):361-370. [doi: 10.1007/s10935-018-0514-y] [Medline: 30003459]
56. Wilson E, Free C, Morris TP, et al. Internet-accessed sexually transmitted infection (e-STI) testing and results service: a randomised, single-blind, controlled trial. PLoS Med 2017 Dec;14(12):e1002479. [doi: 10.1371/journal.pmed.1002479] [Medline: 29281628]
57. Hu Z, Fu Y, Wang X, et al. Effects of sexuality education on sexual knowledge, sexual attitudes, and sexual behaviors of youths in China: a cluster-randomized controlled trial. J Adolesc Health 2023 Apr;72(4):607-615. [doi: 10.1016/j.jadohealth.2022.11.006] [Medline: 36604206]
58. Nuwamanya E, Nalwanga R, Nuwasiima A, et al. Effectiveness of a mobile phone application to increase access to sexual and reproductive health information, goods, and services among university students in Uganda: a randomized controlled trial. Contracept Reprod Med 2020 Oct 31;5(1):31. [doi: 10.1186/s40834-020-00134-5] [Medline: 33292724]
59. Bannink R, Broeren S, Joosten-van Zwanenburg E, van As E, van de Looij-Jansen P, Raat H. Effectiveness of a Web-based tailored intervention (E-health4Uth) and consultation to promote adolescents' health: randomized controlled trial. J Med Internet Res ;16(5). [doi: 10.2196/jmir.3163]
60. Rotheram-Borus MJ, Swendeman D, Comulada WS, Weiss RE, Lee M, Lightfoot M. Prevention for substance-using HIV-positive young people. J Acquir Immune Defic Syndr 2004;37(Supplement 2):S68-S77. [doi: 10.1097/01.qai.0000140604.57478.67]
61. Ballester-Arnal R, Gil-Llario MD, Giménez-García C, Kalichman SC. What works well in HIV prevention among Spanish young people? An analysis of differential effectiveness among six intervention techniques. AIDS Behav 2015 Jul;19(7):1157-1169. [doi: 10.1007/s10461-014-0863-3] [Medline: 25085080]
62. Santa Maria D, Padhye N, Businelle M, et al. Efficacy of a just-in-time adaptive intervention to promote HIV risk reduction behaviors among young adults experiencing homelessness: pilot randomized controlled trial. J Med Internet Res 2021 Jul 6;23(7):e26704. [doi: 10.2196/26704] [Medline: 34255679]
63. Bull S, Devine S, Schmiege SJ, Pickard L, Campbell J, Shlay JC. Text messaging, teen outreach program, and sexual health behavior: a cluster randomized trial. Am J Public Health 2016 Sep;106(S1):S117-S124. [doi: 10.2105/AJPH.2016.303363]
64. Ybarra ML, Bull SS, Prescott TL, Korchmaros JD, Bangsberg DR, Kiwanuka JP. Adolescent abstinence and unprotected sex in CyberSenga, an internet-based HIV prevention program: randomized clinical trial of efficacy. PLoS ONE 2013;8(8):e70083. [doi: 10.1371/journal.pone.0070083] [Medline: 23967069]
65. Bauermeister JA, Tingler RC, Demers M, et al. Acceptability and preliminary efficacy of an online HIV prevention intervention for single young men who have sex with men seeking partners online: the myDEx project. AIDS Behav 2019 Nov;23(11):3064-3077. [doi: 10.1007/s10461-019-02426-7] [Medline: 30762190]
66. Rinehart DJ, Leslie S, Durfee MJ, et al. Acceptability and efficacy of a sexual health texting intervention designed to support adolescent females. Acad Pediatr 2020;20(4):475-484. [doi: 10.1016/j.acap.2019.09.004] [Medline: 31560971]
67. Miller MK, Catley D, Adams A, et al. Brief motivational intervention to improve adolescent sexual health service uptake: a pilot randomized controlled trial in the emergency department. J Pediatr 2021 Oct;237:250-257. [doi: 10.1016/j.jpeds.2021.06.007] [Medline: 34144031]
68. Cordova D, Munoz-Velazquez J, Mendoza Lua F, et al. Pilot study of a multilevel mobile health app for substance use, sexual risk behaviors, and testing for sexually transmitted infections and HIV among youth: randomized controlled trial. JMIR Mhealth Uhealth 2020 Mar 17;8(3):e16251. [doi: 10.2196/16251] [Medline: 32181747]
69. Shafii T, Benson SK, Morrison DM, Hughes JP, Golden MR, Holmes KK. Results from e-KISS: electronic-KIOSK intervention for safer sex: a pilot randomized controlled trial of an interactive computer-based intervention for sexual health in adolescents and young adults. PLOS ONE 2019;14(1):e0209064. [doi: 10.1371/journal.pone.0209064] [Medline: 30673710]

XSL•FO
RenderX

70. Yarger J, Gutmann-Gonzalez A, Borgen N, Romero J, Decker MJ. In the know: a cluster randomized trial of an in-person sexual health education program integrating digital technologies for adolescents. J Adolesc Health 2024 May;74(5):1019-1025. [doi: 10.1016/j.jadohealth.2023.12.012] [Medline: 38323966]
71. Suffoletto B, Akers A, McGinnis KA, Calabria J, Wiesenfeld HC, Clark DB. A sex risk reduction text-message program for young adult females discharged from the emergency department. J Adolesc Health 2013 Sep;53(3):387-393. [doi: 10.1016/j.jadohealth.2013.04.006] [Medline: 23707402]
72. Stoffers JM, Völlm BA, Rücker G, Timmer A, Huband N, Lieb K. Psychological therapies for people with borderline personality disorder. Cochrane Database Syst Rev 2012 Aug 15;2012(8):CD005652. [doi: 10.1002/14651858.CD005652.pub2] [Medline: 22895952]
73. Zhao T, Tang C, Yan H, Lu Q, Guo M, Wang H. Comparative efficacy and acceptability of non-pharmacological interventions for depression in people living with HIV: a systematic review and network meta-analysis. Int J Nurs Stud 2023 Apr;140:104452. [doi: 10.1016/j.ijnurstu.2023.104452] [Medline: 36821952]
74. Nikolakopoulou A, Higgins JPT, Papakonstantinou T, et al. CINeMA: an approach for assessing confidence in the results of a network meta-analysis. PLoS Med 2020 Apr;17(4):e1003082. [doi: 10.1371/journal.pmed.1003082] [Medline: 32243458]
75. Puhan MA, Schünemann HJ, Murad MH, et al. A GRADE working group approach for rating the quality of treatment effect estimates from network meta-analysis. BMJ 2014 Sep 24;349:g5630. [doi: 10.1136/bmj.g5630] [Medline: 25252733]
76. Balduzzi S, Ruecker G, Nikolakopoulou OS. Guido: netmeta: an R package for network meta-analysis using frequentist methods. J Stat Softw 2023;106(2):1-40. [doi: 10.18637/jss.v106.i02]
77. Noma H, Hamura Y, Sugasawa S, Furukawa TA. Improved methods to construct prediction intervals for network meta-analysis. Res Synth Methods 2023 Nov;14(6):794-806. [doi: 10.1002/jrsm.1651] [Medline: 37399809]
78. Chaimani A, Higgins JPT, Mavridis D, Spyridonos P, Salanti G. Graphical tools for network meta-analysis in STATA. PLoS ONE 2013;8(10):e76654. [doi: 10.1371/journal.pone.0076654] [Medline: 24098547]
79. Shafii T, Benson SK, Morrison DM, Hughes JP, Golden MR, Holmes KK. Results from e-KISS: electronic-KIOSK ts and young adults. PLoS ONE 2019;14(1):e0209064. [doi: 10.1371/journal.pone.0209064] [Medline: 30673710]
80. Huang W, Stegmueller D, Ong JJ, et al. Technology-based HIV prevention interventions for men who have sex with men: systematic review and meta-analysis. J Med Internet Res 2025 Apr 28;27:e63111. [doi: 10.2196/63111] [Medline: 40293786]
81. Alkhaldi G, Modrow K, Hamilton F, Pal K, Ross J, Murray E. Promoting engagement with a digital health intervention (HeLP-diabetes) using email and text message prompts: mixed-methods study. Interact J Med Res 2017 Aug 22;6(2):e14. [doi: 10.2196/ijmr.6952] [Medline: 28829328]
82. Plantin JC, de Seta G. WeChat as infrastructure: the techno-nationalist shaping of Chinese digital platforms. Chinese Journal of Communication 2019 Jul 3;12(3):257-273. [doi: 10.1080/17544750.2019.1572633]
83. Fischer AE, Hanif H, Stocks JB, et al. Mobile health intervention tools promoting HIV pre-exposure prophylaxis among adolescent girls and young women in Sub-Saharan Africa: scoping review. JMIR Mhealth Uhealth 2025 Jun 20;13:e60819. [doi: 10.2196/60819] [Medline: 40540732]
84. Mo PH, Xie L, Lee TC, Li AYC. Use of behavior change techniques in digital HIV prevention programs for adolescents and young people: systematic review. JMIR Public Health Surveill 2025 Apr 28;11:e59519. [doi: 10.2196/59519] [Medline: 40293783]
85. Kwasnicka D, Dombrowski SU, White M, Sniehotta F. Theoretical explanations for maintenance of behaviour change: a systematic review of behaviour theories. Health Psychol Rev 2016 Sep;10(3):277-296. [doi: 10.1080/17437199.2016.1151372] [Medline: 26854092]
86. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. J Med Internet Res 2017 Jun 29;19(6):e232. [doi: 10.2196/jmir.7126] [Medline: 28663162]
87. Cao B, Gupta S, Wang J, et al. Social media interventions to promote HIV testing, linkage, adherence, and retention: systematic review and meta-analysis. J Med Internet Res 2017 Nov 24;19(11):e394. [doi: 10.2196/jmir.7997] [Medline: 29175811]
88. Nguyen LH, Tran BX, Rocha LEC, et al. A systematic review of ehealth interventions addressing HIV/STI prevention among men who have sex with men. AIDS Behav 2019 Sep;23(9):2253-2272. [doi: 10.1007/s10461-019-02626-1] [Medline: 31401741]
89. Muessig KE, Nekkanti M, Bauermeister J, Bull S, Hightow-Weidman LB. A systematic review of recent smartphone, internet and web 2.0 interventions to address the HIV continuum of care. Curr HIV/AIDS Rep 2015 Mar;12(1):173-190. [doi: 10.1007/s11904-014-0239-3] [Medline: 25626718]
90. McNeill IM, Borland R, Abraham C. Digital health interventions providing behavioral assessment and goal prioritization support: scoping review. J Med Internet Res 2025 Aug 28;27:e68112. [doi: 10.2196/68112] [Medline: 40875981]
91. Wantland DJ, Portillo CJ, Holzemer WL, Slaughter R, McGhee EM. The effectiveness of web-based vs. non-web-based interventions: a meta-analysis of behavioral change outcomes. J Med Internet Res 2004 Nov 10;6(4):e40. [doi: 10.2196/jmir.6.4.e40] [Medline: 15631964]
92. Noar SM, Black HG, Pierce LB. Efficacy of computer technology-based HIV prevention interventions: a meta-analysis. AIDS 2009 Jan 2;23(1):107-115. [doi: 10.1097/QAD.0b013e32831c5500] [Medline: 19050392]

XSL•FO
RenderX

93. Radix AE, Bond K, Carneiro PB, Restar A. Transgender individuals and digital health. Curr HIV/AIDS Rep 2022 Dec;19(6):592-599. [doi: 10.1007/s11904-022-00629-7] [Medline: 36136217]
94. Waad A. Caring for our community: telehealth interventions as a promising practice for addressing population health disparities of LGBTQ+ communities in health care settings. Delaware Journal of Public Health 2019;5(3):12-15. [doi: 10.32481/djph.2019.06.005]
95. Homkham N, Manojai N, Patpeerapong P, et al. A comparative study of transgender women accessing HIV testing via face-to-face and telemedicine services in Chiang Mai, Thailand during the COVID-19 pandemic and their risk of being HIV-positive. BMC Public Health 2023 Nov 4;23(1):2161. [doi: 10.1186/s12889-023-17124-2] [Medline: 37925430]
96. Atkinson E, Galinkala P, Campos-Castillo C. Telehealth use in 2022 among US adults by sexual orientation. Am J Manag Care 2024 Jan 1;30(1):e19-e25 [FREE Full text] [doi: 10.37765/ajmc.2024.89490] [Medline: 38271570]

Abbreviations

CINeMA: Confidence in Network Meta-Analysis
CrI: credible interval
DHI: digital health intervention
DIC: deviance information criterion
IOI: interactive online-based intervention
MAI: mobile app-based intervention
NDI: nondigital intervention
NMA: network meta-analysis
OR: odds ratio
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyzes
PRISMA-NMA: PRISMA extension for network meta-analyzes
RCT: randomized controlled trial
RoB 2: Cochrane Risk of Bias 2 tool
SR: systematic review
STI: sexually transmitted infection
SUCRA: surface under the cumulative ranking curve
SWI: static web-based intervention
TCI: telecommunication-based intervention

Edited by S Brini; submitted 04.Nov.2025; peer-reviewed by C Aladeokin, JR de-Moya-Romero; accepted 09.Dec.2025; published 04.Feb.2026.

Please cite as:
Zhu Y, Peng W, Hu D, Choi EPH, Välimäki MA, Zhang C, Li X
Effects of Digital Health Interventions to Promote Safer Sex Behaviors Among Youth: Systematic Review and Bayesian Network Meta-Analysis
J Med Internet Res 2026;28:e87071
URL: https://www.jmir.org/2026/1/e87071
doi:10.2196/87071

© Yiran Zhu, Wenwen Peng, Die Hu, Edmond Pui Hang Choi, Maritta Anneli Välimäki, Ci Zhang, Xianhong Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 4.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Characterization of Models for Identifying Physical and Cognitive Frailty in Older Adults With Diabetes: Systematic Review and Meta-Analysis

Xia Wang[1*], MS; Shujie Meng[1*], MS; Xiang Xiao[2], BS; Liu Lu[3]; Hongyan Chen[1], MS; Yong Li[4], MEd; Rong Zhang[4]; Qiwu Jiang[5], BM; Shan Liu[6], MS; Ru Gao[6], PhD

[1]School of Basic Medical Sciences and School of Nursing, Chengdu University, No. 2025, Chengluo Avenue, Chengdu, China

[2]School of Nursing, Yibin Vocational College of Medicine and Health, Yibin, China

[3]Nursing Department, The Fourth People's Hospital of Yibin, Yibin, China

[4]Rehabilitation College, Sichuan Health Rehabilitation Vocational College, Zigong, China

[5]School of Rehabilitation and Wellness, Yibin Vocational College of Medicine and Health, Yibin, China

[6]Nursing Department, Wenjiang District People's Hospital, No.86 Kangtai Road Wenjiang District, Chengdu, Sichuan Province, China

[*]these authors contributed equally

Corresponding Author:
Ru Gao, PhD
Nursing Department, Wenjiang District People's Hospital, No.86 Kangtai Road Wenjiang District, Chengdu, Sichuan Province, China

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2026/1/e92717

Abstract

Background: Physical frailty and cognitive frailty are increasingly recognized as critical geriatric syndromes among older adults with diabetes, contributing to adverse outcomes such as disability, hospitalization, and mortality. Early identification of individuals at high risk is therefore essential for timely prevention and intervention. Although a growing number of prediction models have been developed for this population, evidence regarding their methodological rigor, predictive performance, and generalizability remains fragmented.

Objective: This study aims to evaluate and characterize existing models for detecting or predicting physical frailty and cognitive frailty in older adults with diabetes.

Methods: PubMed, Embase, Web of Science, China National Knowledge Infrastructure (CNKI), Wanfang, and VIP databases were searched from their inception to December 2025. Retrospective, cross-sectional, and prospective studies that developed or validated models predicting frailty or cognitive frailty in older adults with diabetes were included. The Prediction Model Study Risk Of Bias Assessment Tool (PROBAST) was used to assess risk of bias and applicability. Random effects meta-analyses using the Hartung-Knapp-Sidik-Jonkman method were conducted to synthesize model performance, including the pooled area under the receiver operating characteristic curve (AUC). Heterogeneity was explored through subgroup and sensitivity analyses. Small study effects were evaluated using funnel plots, the Egger test, and the Deeks funnel plot asymmetry test.

Results: A total of 24 studies comprising 32 diagnostic models were included. The overall pooled analysis demonstrated an AUC of 0.851 (95% CI 0.820 - 0.882) with a 95% prediction interval of 0.710 - 0.992, sensitivity of 0.810 (95% CI 0.740 - 0.850), and specificity of 0.850 (95% CI 0.810 - 0.890). Statistical comparisons in the modeling approach revealed that logistic regression models achieved a significantly higher pooled AUC (0.850) compared with machine learning models (0.785; P=.003). Similarly, retrospective studies demonstrated superior performance, with an AUC of 0.900 compared with 0.843 for cross-sectional studies (P=.03). Conversely, no significant differences were observed across subgroups stratified by data source (P=.42), patient characteristics (P=.77), validation methods (P=.16), or specific outcomes (P=.94). The most common predictors identified were depression, age, and regular exercise; however, all included studies were assessed as having a high risk of bias.

Conclusions: To our knowledge, this review provides the first comprehensive synthesis of models for risk stratification of physical frailty and cognitive frailty in older adults with diabetes. The findings indicate that existing models demonstrate satisfactory discrimination; specifically, CIs confirmed a robust average effect, while prediction intervals suggested that performance in future settings, though variable, is likely to remain acceptable. However, clinical utility is currently constrained by high risk of bias and

XSL•FO
RenderX

limited external validation. Future research must prioritize rigorous, prospective, multicenter studies adhering to standard reporting guidelines (eg, TRIPOD [Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis]) to establish valid, generalizable, and clinically actionable prognostic instruments.

Trial Registration: PROSPERO CRD420251019308; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251019308

(J Med Internet Res 2026;28:e84617) doi:10.2196/84617

KEYWORDS

prediction model; diabetes; frailty; systematic review; meta-analysis

Introduction

Background

Diabetes mellitus has evolved into one of the most critical global public health challenges of the 21st century. According to the International Diabetes Federation (IDF), approximately 589 million adults were living with diabetes globally in 2024, with projections indicating this number could rise to 853 million by 2050 [1]. In the absence of optimal management, patients with diabetes are predisposed to micro- and macrovascular complications that significantly shorten life expectancy [2,3]. Recent data from the Global Burden of Disease study indicate that the prevalence of diabetes increases with age, reaching 24.4% among individuals aged ≥75 years [4]. Older adults are especially vulnerable to diabetes-related complications due to greater medical complexity and a higher likelihood of frailty compared with younger populations [5].

Frailty is regarded as a consequence of the decline in function and reserve of multiple organs with age, particularly involving the neuromuscular, endocrine, and immune systems [6]. Notably, frailty is particularly prevalent among patients with diabetes, with reported prevalence rates ranging from 10.4% to 20.8% across different studies [7-10]. Furthermore, previous studies indicated that individuals with diabetes have an approximately 1.6-fold higher risk of developing frailty than those without diabetes [11]. However, frailty frequently co-occurs with cognitive impairment [12]; their simultaneous presence is termed cognitive frailty, a distinct clinical entity that represents a crucial subtype of frailty requiring specific attention.

Once established, frailty typically follows a progressive trajectory, increasing the likelihood of adverse clinical outcomes such as falls, incontinence, rapid functional decline, pressure ulcers, and delirium [13-17]. In addition to these risks, frailty is linked to higher rates of hospitalization, emergency department visits, prolonged inpatient stays, and mortality [18,19]. Of particular concern is that the coexistence of physical and cognitive impairment further amplifies these risks, leading to greater adverse outcomes [20,21].

Evidence suggests a bidirectional relationship between diabetes and frailty, often creating a cycle where each condition exacerbates the other [22]. The presence of physical or cognitive frailty introduces significant complexity to diabetes management [23]. In frail patients, physiological deterioration and multi-organ dysfunction fundamentally alter the pharmacokinetics of antihyperglycemic agents [24]. Specifically, sarcopenia, increased adiposity, and compromised renal or hepatic clearance heighten the susceptibility to adverse drug events, such as hypoglycemia and unintended weight loss. Additionally, the decreased caloric intake typical of this population further aggravates the risk of hypoglycemia and hinders recovery from hypoglycemic events [25,26].

In recent years, physical frailty and cognitive frailty have been increasingly conceptualized as dynamic and potentially preventable or reversible conditions, especially when identified at an early stage [27,28]. In patients with diabetes, nonpharmacological interventions—including structured physical activity, nutritional optimization, and multimodal strategies—have demonstrated potential benefits for mitigating frailty progression. Consequently, early identification of individuals at high risk has become a cornerstone of effective prevention and management strategies. To this end, diagnostic and prognostic models designed to detect physical or cognitive frailty integrate multiple demographic, clinical, and psychosocial factors to estimate an individual's risk profile. These models serve to support health care professionals with stratifying risk, facilitating timely and targeted interventions, and optimizing the allocation of health care resources.

However, the clinical application of these models may be hindered due to insufficient evidence regarding their performance, risk of bias, and applicability in routine practice. Although individual studies exist, no systematic review has yet comprehensively evaluated these models for both physical frailty and cognitive frailty in older adults with diabetes. Therefore, it is essential to conduct a systematic review that thoroughly assesses the methodological quality and clinical applicability of existing models.

Objectives

The aim of this systematic review and meta-analysis was to evaluate the methodological quality and clinical utility of existing models designed for the identification or prediction of physical frailty and cognitive frailty in older adults with diabetes. The specific aims included the following: (1) to determine the characteristics and most frequent predictors of risk prediction models developed for physical frailty and cognitive frailty in this population; (2) to analyze the methodological limitations and risk of bias of these models using the Prediction Model Study Risk Of Bias Assessment Tool (PROBAST); and (3) to investigate the pooled predictive performance of these tools to assess their potential for real-world clinical implementation.

XSL•FO
RenderX

Methods

Search Strategy and Selection Criteria

This systematic review and meta-analysis was registered on PROSPERO (CRD420251019308). The study followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) expanded checklist [29] and the PRISMA extension for diagnostic test accuracy (PRISMA-DTA) [30], while the literature search was conducted and reported in accordance with PRISMA-S (PRISMA Search) [31]. A comprehensive literature search was conducted across the PubMed, EMBASE, Web of Science, China National Knowledge Infrastructure (CNKI), Wanfang, and VIP databases, covering records from database inception to December 2025. We developed the search strategy based on the PITROS (Participants, Index Test, Target Conditions, Reference Standard, Outcomes, Settings) framework (Table S1 in Multimedia Appendix 1). The strategy combined Medical Subject Headings (MeSH) with free-text terms (Table S2 in Multimedia Appendix 1). The search strategy was independently evaluated by another librarian in accordance with the PRESS (Peer Review of Electronic Search Strategies) guidelines. In addition, references of relevant studies, guidelines, and reviews were manually searched, and citation tracking was performed using the Web of Science database to identify other relevant studies. No study registries were searched.

In clinical settings, prediction encompasses both diagnostic models (estimating the probability of a particular condition being present) and prognostic models (forecasting the likelihood of future outcomes) [32,33]. This review included all primary studies describing the development and/or validation of prediction models, tools, or scores for estimating the risk of physical frailty or cognitive frailty in older adults with diabetes. The inclusion criteria were (1) participant age ≥60 years and presence of diabetes, including diabetes only and diabetes with other comorbidities or complications; (2) research content involving the construction of a predictive model for identifying physical frailty or cognitive frailty in individuals with diabetes; (3) retrospective studies, cross-sectional studies, and prospective studies; and (4) published in English or Chinese. The exclusion criteria were (1) duplicate publications; (2) reviews, case reports, or conference abstracts; (3) literature that could not be obtained from the original text; (4) literature that could not provide valid data; and (5) studies in which the model contained only a single predictor.

Data Extraction

This study used the reference management software EndNote X9 to identify and remove duplicate records. We then eliminated literature unrelated to the research topic by screening titles and abstracts. Finally, the full texts were reviewed to identify studies satisfying both the inclusion and exclusion criteria. Upon completion of the literature screening process, a data extraction form was devised in accordance with the Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modeling Studies (CHARMS) [34]. The contents included (1) basic information of the study including author, year of publication, study location, study design, and source of data used; (2) patient characteristics including sample size and patient diagnosis; (3) number of predictors, predictor type, most important predictors, and predictor screening methods; (4) model characteristics including outcome indicators, modeling methods, model verification methods, and missing data processing methods; (5) model presentation; and (6) model performance as measured using the area under the curve (AUC), sensitivity, and specificity. In studies where multiple models were developed and the best-performing model was explicitly reported, we included the best model in the analysis. For studies that reported multiple models without specifying a preferred one, we selected the model with the highest AUC to represent the study. We prioritized internally validated estimates, resorting to development performance or external validation data only when internal estimates were unavailable. For studies with unclear or incomplete data, attempts were made to contact the corresponding authors. To ensure the consistency and accuracy of the final data, two researchers (XW and SM) independently extracted the data, and the extracted results were compared and checked. Inconsistencies were resolved through discussion and consultation, and a third researcher (RG) was asked to assist in judgment when necessary.

Quality Assessment

Two reviewers independently used PROBAST [35] to appraise the risk of bias and the applicability of each included study. Disagreements were resolved by discussion or by consulting a third reviewer. PROBAST evaluates 4 domains: participants, predictors, outcome, and analysis. Each domain is rated as high, unclear, or low risk of bias. The applicability evaluation focuses on the 3 areas of research (subjects, predictors, and results), and its evaluation process is similar to the risk of bias assessment.

Statistical Analysis

Meta-analyses were conducted using Stata 18 (specifically the midas and metan commands) and R version 4.3.2 (the metafor package). Using AUC values derived from models, we calculated the pooled AUC and produced an AUC forest plot. An AUC below 0.7 signified inadequate discrimination, an AUC ranging from 0.7 to 0.8 denoted moderate discrimination, and an AUC exceeding 0.8 suggested excellent discrimination [36]. Additionally, models that reported sample sizes and sensitivity and specificity were extracted. The true positive (TP), false positive (FP), false negative (FN), and true negative (TN) for each model were calculated using the formulas sensitivity=TP/(TP+ FN) and specificity=TN/(FP+ TN). Pooled sensitivity and pooled specificity were computed based on TP, FP, FN, and TN, and corresponding forest plots of sensitivity and specificity were constructed. Subsequently, a summary receiver operating characteristic curve was generated. The degree of heterogeneity across the models under consideration was evaluated using the Q test and measured using the I^2 statistic (where an I^2 value <25% signifies low heterogeneity, between 25% and 50% indicates moderate heterogeneity, and >50% denotes high heterogeneity) [37]. To account for between-study heterogeneity and provide more robust variance estimation, we used the Hartung-Knapp-Sidik-Jonkman method for a random effects meta-analysis on the logit scale [38]. For studies with significant heterogeneity, subgroup analyses, sensitivity

XSL•FO
RenderX

analyses, or only descriptive analyses were performed. The presence of small study effects was evaluated using the Egger test, funnel plots, and the Deeks funnel plot [39]. A *P* value <.05 was deemed to indicate statistical significance.

Results

Study Selection

The initial search identified a total of 4873 records. After removing duplicates, 3124 records remained. Following title and abstract screening, 88 studies were selected for full-text review. We could not retrieve 4 studies. During the full-text assessment, 18 studies were excluded because the study population did not have diabetes mellitus. Additionally, 16 studies were excluded because the outcome variable was not physical frailty or cognitive frailty; 16 were excluded because they were nonoriginal studies, such as reviews and meta-analyses; and 10 were excluded due to an inappropriate study design. Ultimately, 24 studies [40-63] were included in this meta-analysis. The literature screening process and results are illustrated in Figure 1.

Figure 1. Literature screening flow chart. CNKI: China National Knowledge Infrastructure.

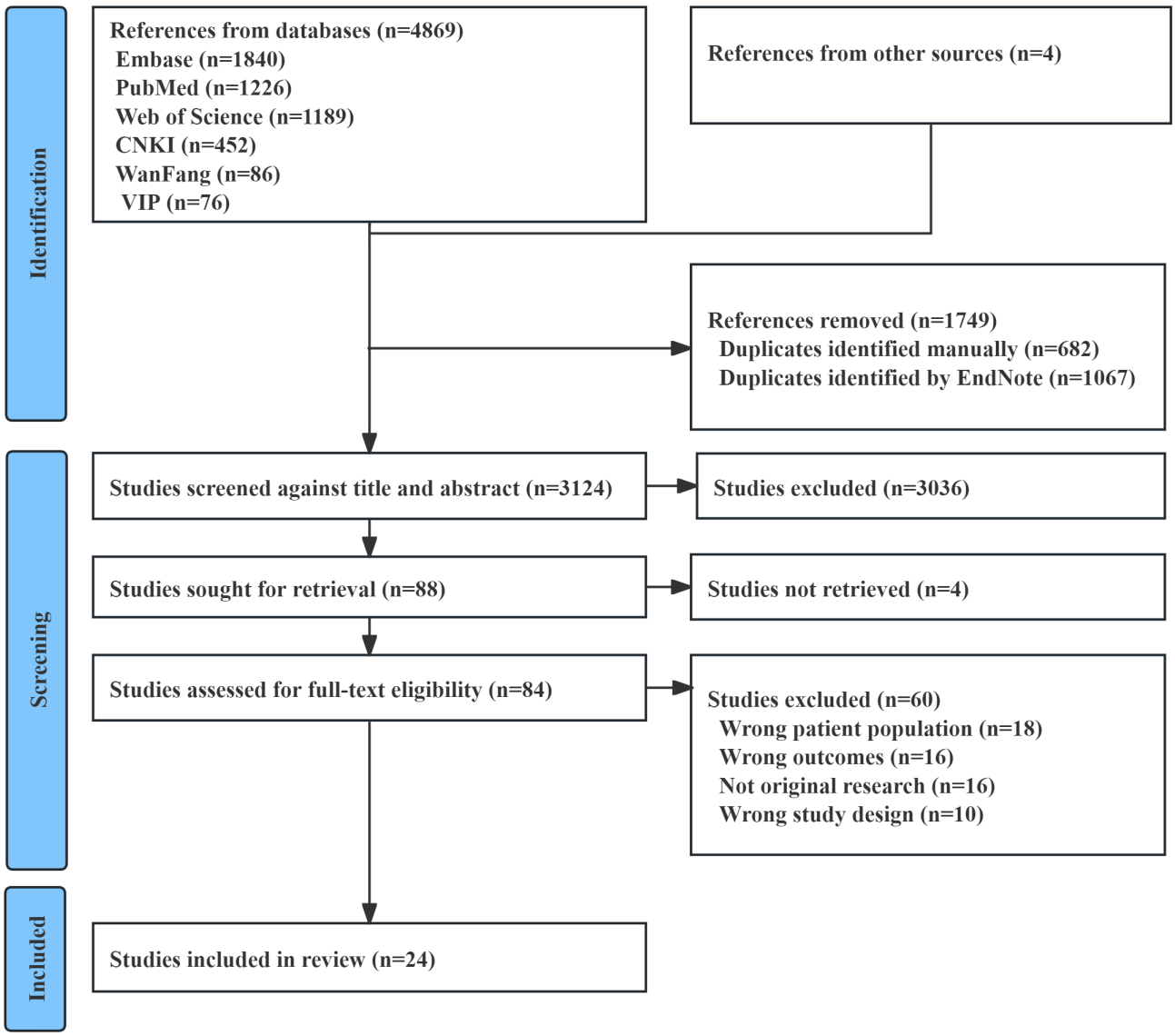

Characteristics of Included Studies

The specific characteristics of the included studies are detailed in Tables 1 and 2. A total of 24 studies reported 32 diagnostic models for physical frailty and cognitive frailty in older adults with diabetes. The publication years of these papers spanned from 2023 to 2025. All the studies were conducted in China. The number of participants in the included studies ranged from 152 to 1436. The prevalence of frailty varied from 10.1% to 51.2%, while that of cognitive frailty ranged from 20.3% to 62.1%. Regarding study design, 21 studies [41-46,48-60,62,63] used a cross-sectional design, while 3 were retrospective studies [40,47,61]. In addition, 18 studies [41-48,50,51,53-59,63] were conducted at single centers, whereas 6 studies [40,49,52,60-62] were multicenter studies. In terms of the target population, 7 studies [40-42,57,58,60,61] included all types of diabetes, 14 studies [43,45,47-51,53-56,59,62,63] focused specifically on patients with type 2 diabetes, and 3 studies [44,46,52] included patients with diabetes and other comorbidities or complications. Physical frailty was the primary outcome in 13 studies [40-42,44,45,51,52,55,57,58,60-62], while the remaining 11 studies [43,46-50,53,54,56,59,63] investigated cognitive frailty.

XSL•FO
RenderX

For frailty screening, 22 studies [40-51,53-57,59-63] used the Fatigue, Resistance, Ambulation, Illnesses, and Loss of Weight (FRAIL) scale, while 2 studies [52,58] used the Tilburg Frailty Indicator. To evaluate cognitive function in the 11 cognitive frailty studies, the Montreal Cognitive Assessment was the predominant tool used in 10 studies [43,46-50,53,54,59,63], whereas the Mini-Mental State Examination was used in only 1 study [56]. Regarding data handling, missing data were not reported in 9 studies [42,43,45-47,49,51,56,59], 3 studies [40,57,62] used imputation methods, and 12 studies [41,44,48,50,52-55,58,60,61,63] excluded participants with missing data. Continuous variables were maintained as continuous in 14 studies [40,42-44,47,51,52,55-59,61,62] and transformed into categorical variables in 10 studies [41,45,46,48-50,53,54,60,63]. The majority of models were presented as nomograms: Specifically, 13 studies [40,43,45,48,50,51,54-57,59,61,63] presented results solely as nomograms, 4 studies [41,42,44,46] provided both nomograms and full equations, 1 study [52] presented a nomogram and risk chart, and 1 study [58] presented a nomogram and decision tree. Additionally, 3 studies [53,60,62] developed risk sum scores, and 2 studies [47,49] provided logistic regression (LR) equations.

XSL•FO
RenderX

Table . Overview of the basic characteristics of included studies (n=24) identifying physical and cognitive frailty in older adults with diabetes.

Author	Year	Outcome	Definition of outcome	Sample size, n	Event rate, n (%)	Modeling algorithms used	Population	Internal validation	Predictors, n	Model presentation
Wu et al [62]	2025	Frailty	FRAIL[a]	509	148 (29.1)	LR[b], SVM[c], GBM[d], RF[e], CatBoost	T2DM[f]	Cross-validation	7	Sum score
Xiao et al [61]	2025	Frailty	FRAIL	1107	113 (10.2)	LR	Diabetes	Random split	7	Nomogram
Wang et al [55]	2025	Frailty	FRAIL	152	47 (31. 1)	LR	T2DM	Random split	4	Nomogram
Du et al [45]	2024	Frailty	FRAIL	458	83 (18.1)	LR	T2DM	Bootstrap	8	Nomogram
Tang et al [51]	2024	Frailty	FRAIL	566	213 (37.6)	LR	T2DM	Random split	6	Nomogram
Wang et al [52]	2024	Frailty	TFI[g]	491	216 (44.0)	LR, NN[h]	Diabetes with diabetic foot	Random split	8	Nomogram and risk chart
Dang [42]	2024	Frailty	FRAIL	360	115 (32.0)	LR	Diabetes	Random split (+ external)	5	Nomogram and full equation
Xi [57]	2024	Frailty	FRAIL	338	130 (38.5)	LR	Diabetes	Random split	6	Nomogram
Cheng [41]	2024	Frailty	FRAIL	317	118 (37.2)	LR	Diabetes	Bootstrap	5	Nomogram and full equation
Yin [58]	2024	Frailty	TFI	379	194 (51.2)	LR, DT[i]	Diabetes	Bootstrap	8	Nomogram and decision tree
Zheng [60]	2024	Frailty	FRAIL	380	112 (29.5)	RF, SVM, KNN[j]	Diabetes	Cross-validation	7	Sum score
Bu et al [40]	2023	Frailty	FRAIL	1436	145 (10.1)	LR	Diabetes	Random split	7	Nomogram
Dong et al [44]	2023	Frailty	FRAIL	485	211 (43.5)	LR	Diabetes with diabetic retinopathy	Bootstrap	7	Nomogram and full equation
Ma et al [49]	2025	Cognitive frailty	FRAIL, MoCA[k]	253	76 (30.0)	LR	T2DM	None	5	Full equation
Wang et al [53]	2025	Cognitive frailty	FRAIL, MoCA	202	80 (39.6)	DT	T2DM	Random split	11	Sum score
Liang et al [46]	2024	Cognitive frailty	FRAIL, MoCA	265	93 (35.1)	LR	Diabetes with COPD[l]	Random split (+ external)	7	Nomogram and full equation
Yu and Yu [63]	2024	Cognitive frailty	FRAIL, MoCA	430	132 (30.7)	LR	T2DM	Bootstrap (+ external)	7	Nomogram
Zhang et al [59]	2024	Cognitive frailty	FRAIL, MoCA	215	66 (30.7)	LR	T2DM	Bootstrap	5	Nomogram
Liu [47]	2024	Cognitive frailty	FRAIL, MoCA	220	137 (62.1)	LR	T2DM	Random split	4	Full equation
Deng et al [43]	2023	Cognitive frailty	FRAIL, MoCA	315	87 (27.6)	LR	T2DM	Random split	6	Nomogram

XSL•FO
RenderX

Author	Year	Outcome	Definition of outcome	Sample size, n	Event rate, n (%)	Modeling algorithms used	Population	Internal validation	Predictors, n	Model presentation
Wang and Xu [54]	2023	Cognitive frailty	FRAIL, MoCA	262	85 (32.4)	LR	T2DM	Bootstrap	8	Nomogram
Wang et al [56]	2023	Cognitive frailty	FRAIL, MMSE[m]	321	85 (26.5)	LR	T2DM	Bootstrap	5	Nomogram
Liu [48]	2023	Cognitive frailty	FRAIL, MoCA	483	98 (20.3)	LR	T2DM	Random split	6	Nomogram
Meng [50]	2022	Cognitive frailty	FRAIL, MoCA	508	117 (23.0)	LR	T2DM	Bootstrap (+ external)	6	Nomogram

[a]FRAIL: Fatigue, Resistance, Ambulation, Illness, and Loss of Weight scale.

[b]LR: logistic regression.

[c]SVM: support vector machine.

[d]GBM: gradient boosting machine.

[e]RF: random forest.

[f]T2DM: type 2 diabetes mellitus.

[g]TFI: Tilburg Frailty Indicator

[h]NN: neural network.

[i]DT: decision tree.

[j]KNN: k-nearest neighbors.

[k]MoCA: Montreal Cognitive Assessment.

[l]COPD: chronic obstructive pulmonary disease.

[m]MMSE: Mini-Mental State Examination.

XSL•FO
RenderX

Table . Methodological and clinical characteristics of included studies (n=24) identifying physical frailty and cognitive frailty in older adults with diabetes.

Characteristic	Studies, n (%)
Study design	
Retrospective studies	3 (13)
Cross-sectional study	21 (88)
Source of data used	
Single center	18 (75)
Multicenter	6 (25)
Missing data handling	
Not reported	9 (38)
Exclusion	12 (50)
Imputation	3 (13)
Handling of continuous data	
Continuous	14 (58)
Categorical or dichotomous	10 (42)
Feature selection	
Univariate analysis	6 (25)
Multivariate analysis	7 (29)
Univariate analysis and multivariate analysis	11 (46)
Calibration method	
Hosmer-Lemeshow test	2 (8)
Calibration plot	4 (17)
Hosmer-Lemeshow test and calibration plot	14 (58)
None	4 (17)
Validation method	
Internal validation	19 (79)
External validation and internal validation	4 (17)
None	1 (4)

Characteristics of Included Prediction Models

Regarding modeling methods, among the 32 models included, LR was the most commonly used algorithm. LR analyses were used in 22 models, while 10 models used machine learning (ML) techniques, including random forest (n=2), support vector machines (n=2), decision trees (n=2), k-nearest neighbors (n=1), CatBoost (n=1), gradient boosting machine (n=1), and neural networks (n=1). Model discrimination was reported for all models, with AUC values ranging from 0.703 to 0.983 (Table S3 in Multimedia Appendix 1). Specificity and sensitivity were reported in 17 studies [40-42,44,46-50,52,53,56-58,60-62] involving 25 models. Specifically, sensitivity ranged from 0.102 to 0.955, and specificity varied from 0.505 to 0.990. However, model calibration was not reported in 4 studies [52,53,60,62]. *P* values from both the Hosmer-Lemeshow test and calibration plots were used in 14 studies [40-44,48,50,51,54-59], 2 studies [47,49] used *P* values only from the Hosmer-Lemeshow test, and 4 studies [45,46,61,63] used only calibration plots. Regarding model validation, 1 study [49] developed models without validation, and 19 studies [40,41,43-45,47,48,51-62] conducted only internal validation without external validation.

Features

All features covered a wide range of factors, including sociodemographic characteristics, lifestyle factors, health-related factors, mental health status, laboratory test indicators, and anthropometric measurements. A total of 33 features were involved in the studies. The number of features incorporated into each study varied from 4 to 11. Among the features, the 5 most frequently occurring were depression, age, regular exercise, social activity, and duration of diabetes. The frequency distribution of all features is illustrated in Figure 2.

XSL•FO
RenderX

Figure 2. Frequency of predictors in the included studies. GFR: glomerular filtration rate; HbA_{1c}: glycated hemoglobin; L/A: ratio of serum leptin to adiponectin.

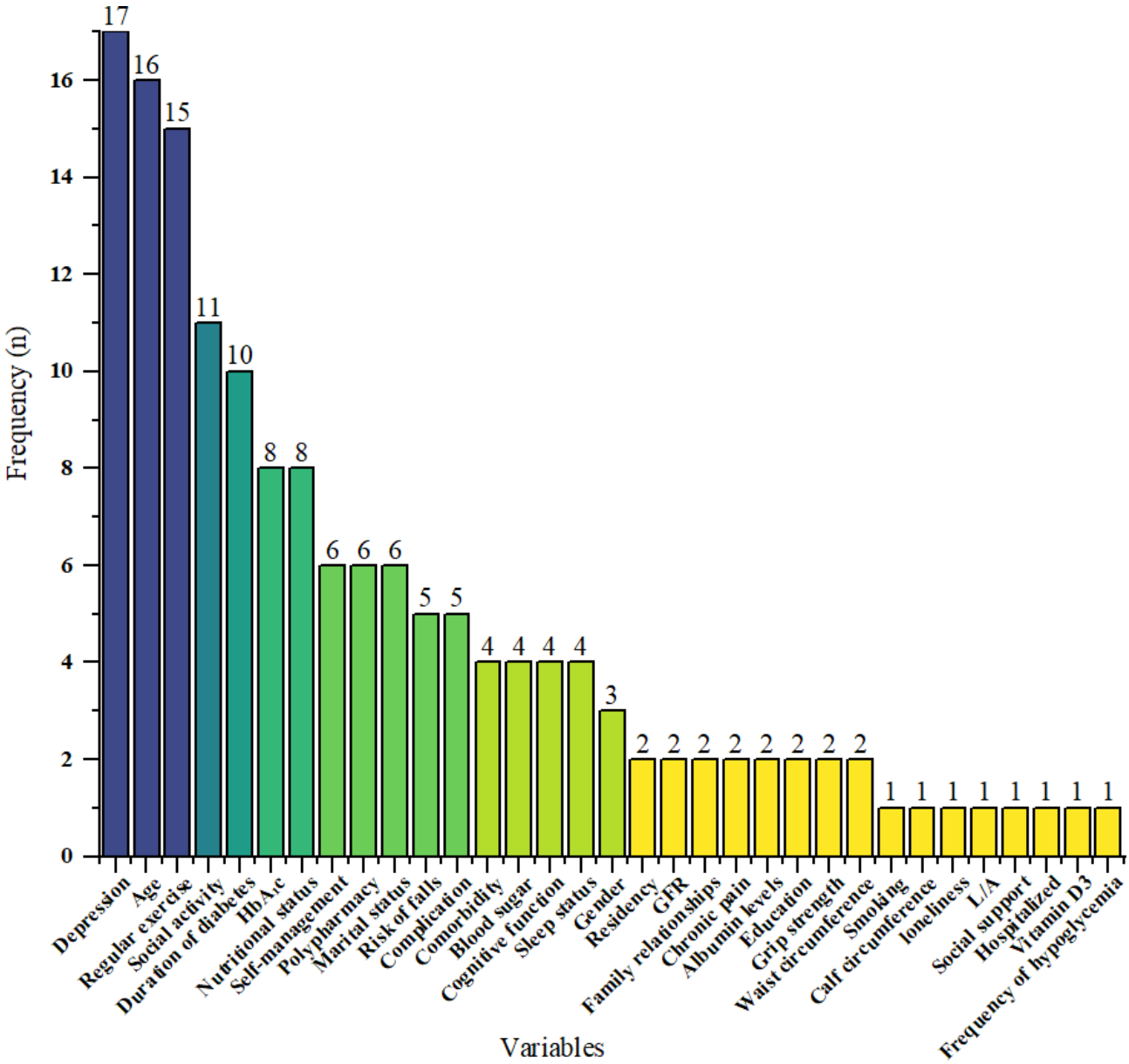

Quality Assessment

Overview

The PROBAST tool was used to assess the risk of bias and the applicability of the included prediction model studies (Figure 3 and Table S4 in Multimedia Appendix 1). According to the established criteria, all 24 studies, which encompassed 32 models, were identified as having a high risk of bias. In terms of applicability, 13 studies [43,44,46-50,52-54,56,59,63], including 14 models, were deemed to have high concerns regarding applicability. Conversely, the remaining 11 studies [40-42,45,51,55,57,58,60-62], which included 18 models, were considered to have low concerns regarding applicability. Notably, 4 studies [52,58,60,62] included multiple models each; however, there was no difference in the quality assessment results between the models within these studies.

XSL•FO
RenderX

Figure 3. Prediction Model Study Risk Of Bias Assessment Tool (PROBAST) risk of bias (ROB) and applicability assessment for all included studies.

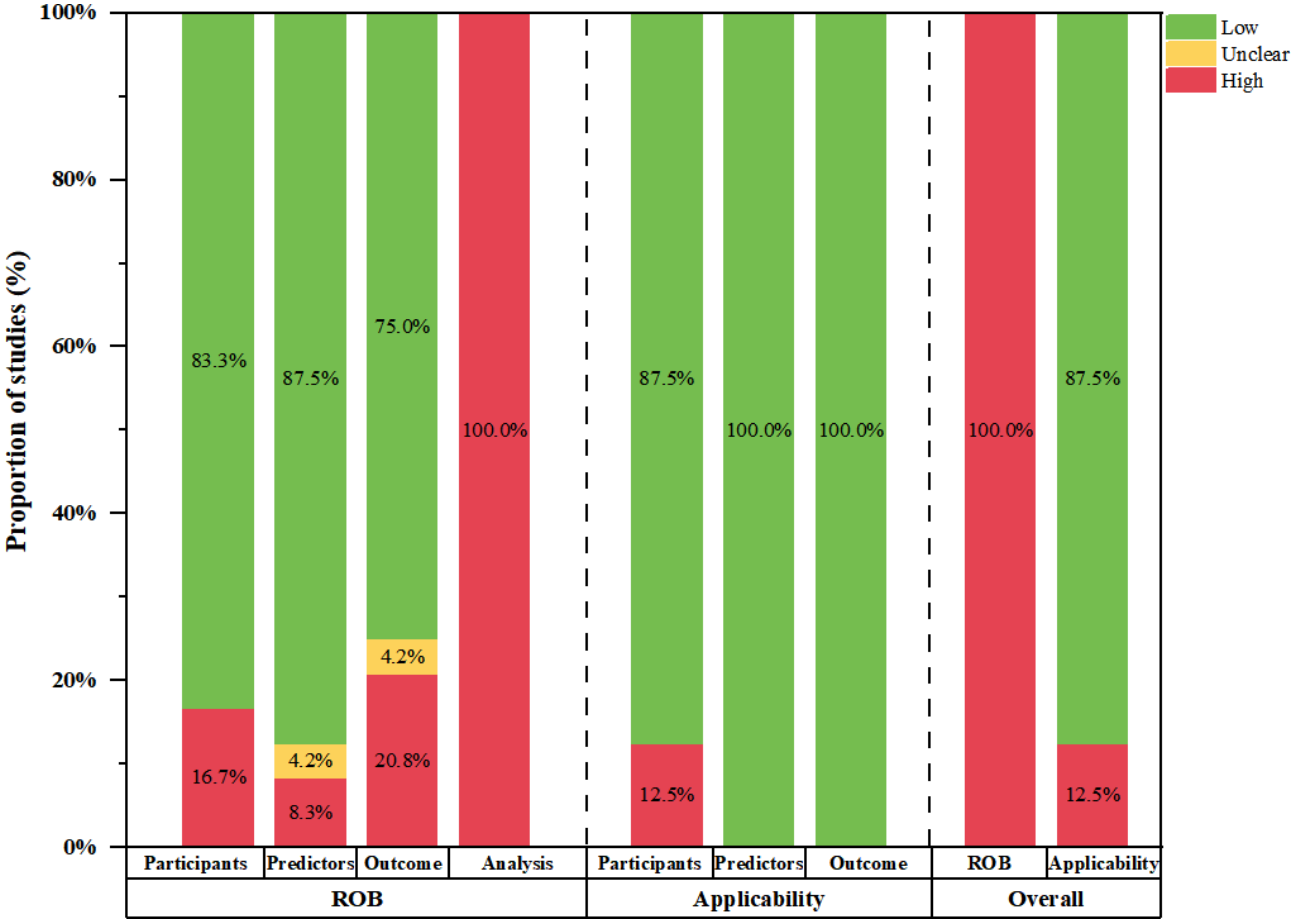

Risk of Bias Assessment

Within the participant domain, 4 studies [40,43,47,61] were recognized as exhibiting a high risk of bias. Of these, 3 studies [40,47,61] were deemed as having a high risk owing to their study designs, while the remaining study [43] was classified as such due to the exclusion of specific subgroups that could potentially alter the performance of the prediction model. In the predictor domain, 2 studies [40,63] were assessed as having a significant risk of bias due to the use of outcome information in the evaluation of predictors, 1 study [55] was rated as having an unclear risk of bias because the researchers did not report whether they used the same assessment measures when evaluating the predictors. In the outcome domain, 3 studies [40,62,63] had a significant risk of bias because the definition of outcomes included ≥1 predictor, 2 studies [52,53] were rated as having a high risk of bias due to the potentially inappropriate time interval between predictor assessment and outcome determination, and 1 study [49] was deemed to be at unclear risk of bias as they did not report information on the method of outcome classification. In the analysis domain, all studies were judged to have a high risk of bias. Current guidance recommends that studies developing predictive models achieve at least 20 events per variable (EPV). However, 13 studies [43,45,46,48,49,53-56,59-61,63] did not meet this requirement. Moreover, 10 studies [41,45,46,48-50,53,54,60,63] transformed continuous variables into categorical variables, either in part or entirely, and the authors did not report whether standard definitions were used for the categorization; 1 study [40] partially excluded participants for unreasonable reasons. Regarding the handling of missing data, 12 studies [41,42,44,48,50,52-54,58,60,61,63] directly excluded cases with missing data, while 9 studies [43,45-47,49,51,55,56,59] did not explicitly report whether data were missing. In addition, 6 studies [44,49,50,52,54,59] did not avoid selecting variables based solely on univariate analysis; 3 studies [52,53,60] did not comprehensively assess the predictive performance of their models, using only discrimination measures without calibration; 6 studies [41,44,48,50,55,57] neglected to evaluate the risk of overfitting, underfitting, or optimism that could bias the apparent performance of their predictive models; 1 study [49] developed models without validation; and 19 studies [40,41,43-45,47,48,51-62] conducted only internal validation without external validation.

Applicability Risk Assessment

In the participant domain, 3 studies [44,46,52] had a high risk of applicability concerns due to the inclusion of individuals with other comorbidities or complications.

Meta-Analysis

A random effects meta-analysis using the Hartung-Knapp-Sidik-Jonkman method was performed to evaluate the predictive performance at both the study and model levels. Regarding the analysis of the 24 included studies, the overall pooled AUC was 0.851 (95% CI 0.820 - 0.882), with a 95% prediction interval (PI) of 0.710 to 0.992 (P<.001; I^2=92.0%; Figure 4). When analyzing the 32 models, the overall

XSL•FO
RenderX

pooled AUC was 0.829 (95% CI 0.802 - 0.856), with a 95% PI of 0.686 to 0.972 (P<.001; I^2=92.5%; Figure S1 in Multimedia Appendix 1).

Figure 4. Forest plot of the random effects meta-analysis of pooled area under the curve (AUC) estimates for 29 validation models [40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62]. HKSJ: Hartung-Knapp-Sidik-Jonkman.

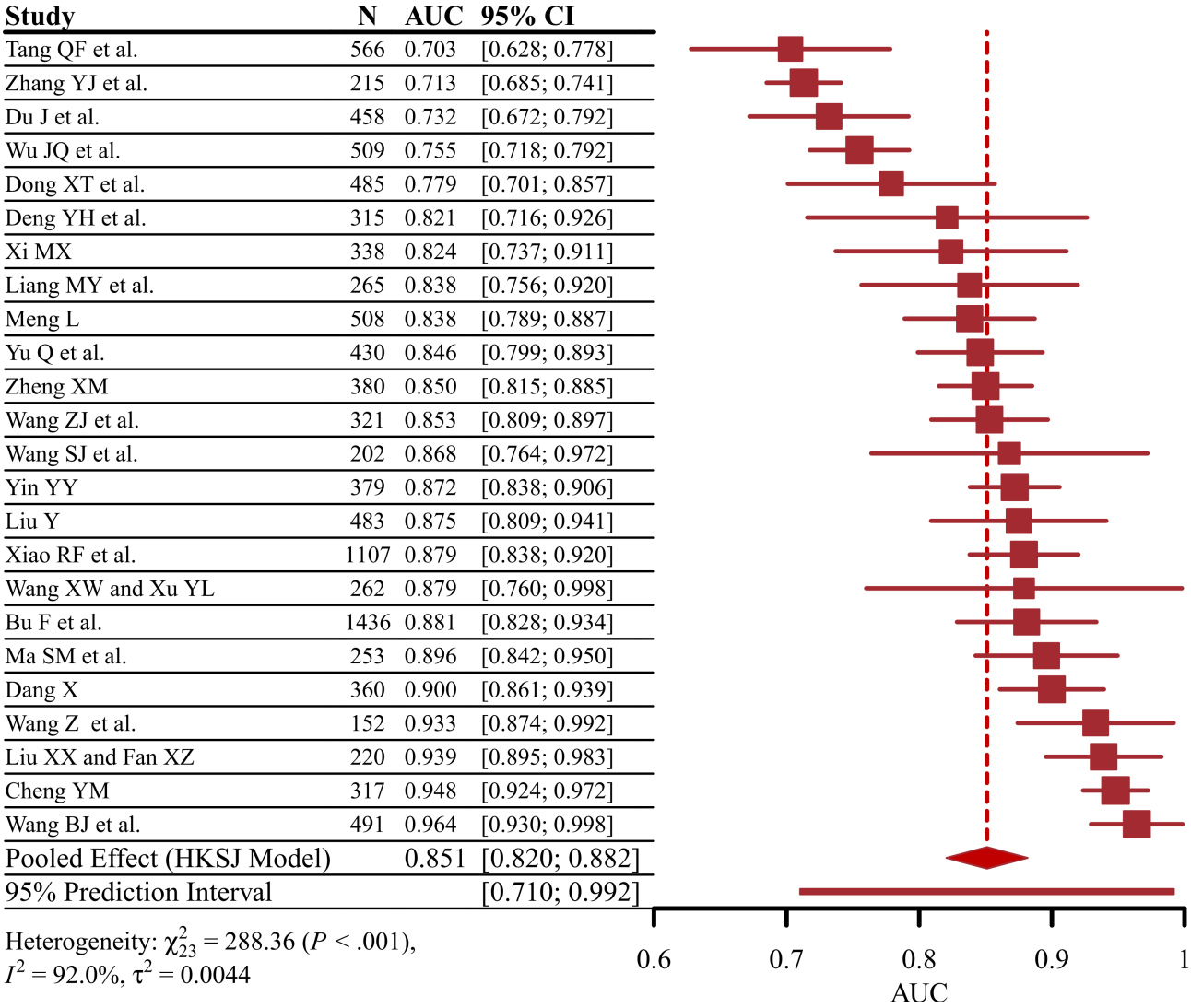

Additional data on sample size, sensitivity, and specificity were extracted from 17 studies to calculate TP, FP, FN, and TN (Table S5 in Multimedia Appendix 1). Based on these values, the pooled sensitivity was 0.810 (95% CI 0.740 - 0.850; I^2=92.26%), as illustrated in the forest plot. The pooled specificity was 0.850 (95% CI 0.810 - 0.890; I^2=92.44%), with the corresponding forest plot also presented (Figure 5B). Furthermore, a summary receiver operating characteristic curve was generated, as depicted in Figure 5A. These results indicate significant heterogeneity across the models regarding AUC, sensitivity, and specificity.

XSL•FO
RenderX

Figure 5. (A) Summary receiver operating characteristic curve and (B) forest plots of the random effects meta-analysis of the sensitivity and specificity for 22 validation models [40-42,44,46-50,52,53,56-58,60-62]. AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

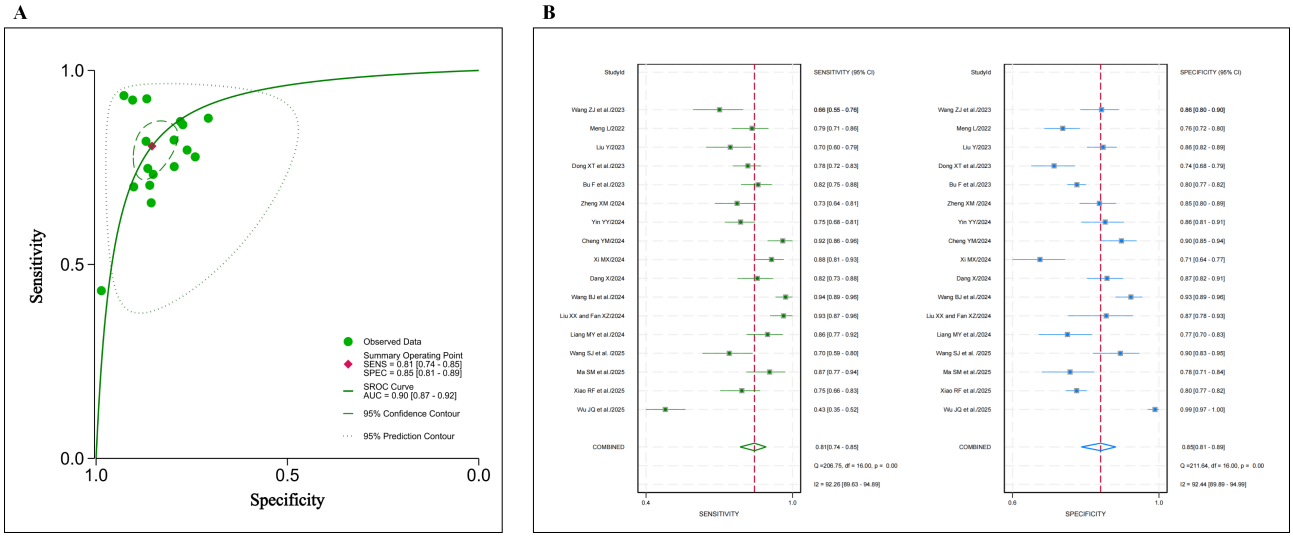

Sensitivity Analysis

Sensitivity analyses were undertaken by sequentially removing one study at a time. The point estimates obtained after excluding any single study all fell within the 95% CI of the overall effect size (Figure 6A). This indicates that the removal of any individual study did not significantly influence the pooled AUC. Therefore, the combined results were relatively stable.

Figure 6. Assessment of robustness, heterogeneity, and small study effects in the meta-analysis of 24 studies evaluating physical frailty and cognitive frailty in older adults with diabetes: (A) sensitivity analysis, (B) subgroup forest plot stratified by modeling approach (32 models), and (C) funnel plot [40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63]. AUC: area under the curve; HKSJ: Hartung-Knapp-Sidik-Jonkman; LR: logistic regression; ML: machine learning.

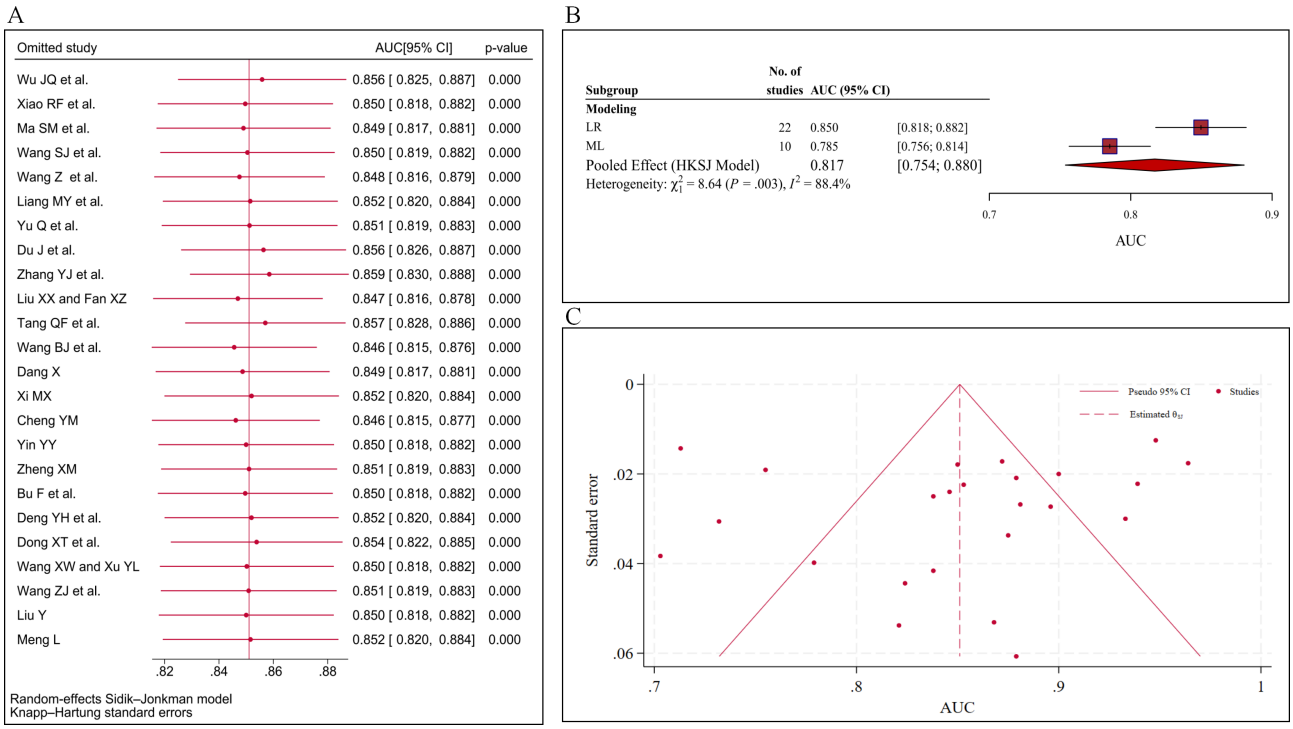

Subgroup Analysis

Subgroup analyses were performed based on modeling approach, study design, data source, population, outcome, and validation method (Table 3 and Figure 6B). Significant differences were observed in modeling approaches (*P*=.003), where LR models yielded a higher pooled AUC (0.850) compared with ML models (0.785). Similarly, study design showed significant heterogeneity (*P*=.03), with retrospective studies demonstrating superior diagnostic performance (AUC=0.900) than cross-sectional studies (AUC=0.843). However, no statistically significant differences were found across subgroups stratified by data source (*P*=.42), patient characteristics (*P*=.77), validation methods (*P*=.16), or specific outcomes (*P*=.94).

XSL•FO
RenderX

Table . Subgroup analysis of the pooled area under the curve (AUC) of studies of physical frailty and cognitive frailty in older adults with diabetes to explore potential sources of heterogeneity (n=24).

Subgroup	Studies, n	AUC (95% CI)	I^2, %	*P* value
Population			0	.77
Diabetes	21	0.849 (0.818-0.879)		
Diabetes with comorbidities or complications	3	0.866 (0.757-0.975)		
Study design			79.6	.03
Cross-sectional study	21	0.843 (0.812-0.875)		
Retrospective study	3	0.900 (0.861-0.939)		
Data source			0	.42
Multicenter	6	0.871 (0.815-0.926)		
Single center	18	0.844 (0.809-0.878)		
Outcome			0	.94
Physical frailty	13	0.851 (0.806-0.896)		
Cognitive frailty	11	0.849 (0.812-0.886)		
Validation			42.5	.16
Random split	12	0.875 (0.836-0.914)		
Bootstrap	9	0.829 (0.779-0.879)		
Cross-validation	2	0.803 (0.710-0.896)		
None	1	0.896 (0.842-0.950)		

Small Study Effects Assessment

We examined small study effects using the Egger test, funnel plots, and the Deeks funnel plot. The Egger test yielded a coefficient of −1.07 (*P*=.40), indicating no statistically significant small study effects. This result aligns with the visual inspection of the contour-enhanced funnel plot (Figure 6C), which displayed a relatively symmetrical distribution of the studies, suggesting no obvious small study effects among the included studies. Furthermore, the Deeks funnel plot asymmetry test yielded a nonsignificant *P* value of .94, providing no evidence of significant small study effects (Figure S2 in Multimedia Appendix 1).

However, these results must be interpreted with caution. Funnel plot symmetry or nonsignificant statistical tests do not definitively rule out publication bias. Conversely, asymmetry may arise from heterogeneity or methodological limitations not merely from publication bias. Therefore, the potential for publication bias cannot be entirely excluded, particularly due to limitations in study selection. Although our search strategy was designed to be global, the final pool of eligible studies consisted exclusively of research conducted in China. Additionally, the restriction of inclusion criteria to English and Chinese languages may have introduced language bias by excluding relevant data published in other languages.

Discussion

Principal Findings

To our knowledge, this is the first systematic review and meta-analysis to comprehensively evaluate the performance of prediction models for physical frailty and cognitive frailty specifically in older adults with diabetes. Our analysis included 24 studies encompassing 32 models. At the study level, the pooled AUC was 0.851 (95% CI 0.820 - 0.882), while the model-level analysis yielded a similarly high pooled AUC of 0.829 (95% CI 0.802 - 0.856). In addition, the pooled sensitivity was 0.810, and the pooled specificity was 0.850, indicating that these models demonstrated reasonable discriminative performance for identifying physical frailty and cognitive frailty in this high-risk population.

A notable finding from our subgroup analysis was the difference in performance based on the modeling approach (*P*=.003). LR models yielded a higher pooled AUC (0.850) than ML models (0.785). However, this finding should be interpreted with caution. Consistent with the systematic review by Christodoulou et al [64], we observed no consistent performance advantage of ML over LR analysis in this dataset. The apparent disparity may be attributed to the smaller number of ML studies included, heterogeneity in populations and predictors, and potential risk of bias, rather than an inherent superiority of LR. For instance, Wang et al [52] and Yin [58] developed models using both approaches and found LR performed slightly better. Ultimately, there is no “one-size-fits-all” modeling method; performance often depends on the specific data structure and clinical context. Therefore, future research should prioritize rigorous comparisons

XSL•FO
RenderX

of multiple modeling approaches—including proper hyperparameter tuning for ML—to identify the optimal strategy for specific physical frailty and cognitive frailty prediction scenarios. Additionally, we observed that retrospective studies yielded significantly higher AUC values (0.900) than cross-sectional studies (0.843; P=.03). This phenomenon likely stems from the inherent selection bias and better data quality control often present in retrospective cohorts, potentially leading to overoptimistic performance estimates.

The diagnostic models included in this review possess meaningful clinical implications, facilitating a shift toward the precise identification and risk stratification of physical frailty and cognitive frailty in clinical settings. Our analysis revealed that the most frequently used features—depression, age, regular exercise, social activity, and diabetes duration—provide concrete metrics for identifying these concurrent conditions. Notably, depression emerged as the most consistent and prominent feature across multiple studies, highlighting its strong correlation with both physical frailty and cognitive frailty. Existing evidence indicates a bidirectional relationship between depression and these conditions, potentially mediated through inflammatory pathways, endocrine dysregulation, and overlapping symptoms such as fatigue and psychomotor slowing [65,66]. Consequently, assessing mental health not only is vital for psychological well-being but also serves as a critical entry point for identifying patients who may already be experiencing physical frailty or cognitive frailty. Advanced age was also a predominant feature, consistent with reports by Kong et al [9] and Wang et al [67]. This association likely reflects immunosenescence, chronic inflammation, and metabolic dysregulation, which collectively contribute to sarcopenia and functional decline [68-70]. These findings suggest that age remains a fundamental stratification factor, warranting heightened clinical vigilance for physical and cognitive frailty in older cohorts. Regular exercise was identified as a powerful discriminatory feature. Physiologically, physical activity is known to reduce inflammation, preserve muscle function, and support cognitive health [71]. In the context of these diagnostic models, the absence of regular exercise serves as a robust, easily accessible clinical marker for detecting potential physical frailty and cognitive frailty. This allows health care providers to efficiently target vulnerable populations in community settings where elaborate geriatric assessments may be impractical. Similarly, lower levels of social activity emerged as a significant indicator. This suggests that social isolation often co-occurs with physical frailty and cognitive frailty, making social history a valuable component of the risk stratification process for older adults with diabetes. Finally, the duration of diabetes was a frequent feature in the included models. The strong association between longer disease duration and physical frailty and cognitive frailty likely reflects the cumulative burden of chronic hyperglycemia, insulin resistance, and related complications over time [8,26]. As diabetes duration increases, physiological and cognitive reserves decline, thereby increasing the probability of concurrent physical frailty and cognitive frailty. These insights emphasize that patients with a long history of diabetes represent a high-risk group requiring prioritized screening and comprehensive management.

Comparison With Prior Work

Previous systematic reviews and narrative summaries on physical frailty or cognitive frailty in older adults with diabetes have primarily concentrated on estimating prevalence, identifying associated risk factors, or describing frailty phenotypes, rather than systematically evaluating multivariable prediction or identification models [8,9,72,73]. Moreover, most prior reviews did not attempt a quantitative synthesis of model performance, likely due to the substantial methodological and clinical heterogeneity across studies, which is commonly encountered in the evaluation of prediction models. Consistent with this literature, we observed considerable variation among included studies in terms of study populations, definitions of physical frailty and cognitive frailty, predictor selection, model development strategies, and validation approaches. These differences reflect the evolving and fragmented nature of model development in this field and pose challenges for direct comparison across studies. Nevertheless, by conducting a meta-analysis of discrimination performance, particularly through the synthesis of the AUC, our study provides a quantitative overview of the overall performance of existing models for identifying physical frailty and cognitive frailty in older adults with diabetes.

Heterogeneity

Substantial heterogeneity was observed across the included studies (I^2>90% for sensitivity, specificity, and AUC). This is a common challenge in diagnostic meta-analyses and may be attributed to variations in study design, population characteristics, and modeling methodologies. Our subgroup analysis identified that the modeling approach (ML vs LR) and study design (retrospective vs cross-sectional) were significant sources of heterogeneity (P<.05). However, other factors such as data source (single vs multicenter) and outcome definitions (physical frailty vs cognitive frailty) did not significantly contribute to the observed variance. It is also important to note the variability in feature selection across studies. With 33 different features identified—ranging from depression and age to regular exercise—the lack of a standardized set of predictors likely contributes to the heterogeneity in model performance. This diversity reflects the multifaceted pathophysiology of frailty in diabetes but complicates the direct comparison of models.

Importantly, beyond the conventional I^2 statistic, we further quantified between-study heterogeneity using 95% PIs, which provide a clinically meaningful estimate of the expected range of model performance in future settings. Although I^2 values exceeding 90% indicate substantial relative heterogeneity, they do not convey the absolute extent to which predictive performance may vary across populations and clinical contexts. In contrast, PIs directly address this limitation by reflecting the dispersion of true effects on the original AUC scale. At the study level, the pooled AUC of 0.851 was accompanied by a 95% PI ranging from 0.710 to 0.992, indicating that, although predictive performance may vary considerably across different real-world settings, most future applications are still likely to achieve at least acceptable discrimination.

XSL•FO
RenderX

Methodological Quality and Risk of Bias

Evaluation using the PROBAST checklist indicated that all included studies exhibited a high risk of bias, predominantly in the analysis domain. Consequently, the pooled performance estimates reported in this review should be interpreted with caution, likely representing "best-case scenarios" or optimistic estimates rather than robust predictions of real-world performance.

In the participant domain, specifically regarding data sources, although retrospective designs were identified as a source of bias for a few studies, the overarching issue remains the analytical approach. We recommend using prospective data or registry data for model development in future optimization efforts to reduce the risk of bias arising from data sources. Additionally, the evaluation of model applicability indicated that certain studies included not only patients with diabetes but also those with other comorbidities or complications. These factors limited the applicability of the respective models to the general diabetes population. In the outcome domain, some studies inappropriately incorporated outcome-related information into the predictor assessment, leading to information leakage and inflated model performance. In addition, inadequate reporting regarding the consistency of predictor measurement raised concerns about reproducibility in at least one study. In the outcome domain, several studies were judged to have a high risk of bias due to problematic outcome definitions.

The analysis domain had the highest frequency of a high risk of bias, with all studies rated as high risk in this domain. According to the PROBAST assessment tool, an EPV≥20 is commonly used as a heuristic to indicate an adequate sample size for developing prediction models. In this review, 13 studies had an EPV <20, which may suggest an increased risk of bias related to model overfitting. However, relying solely on fixed rules of thumb may be insufficient; therefore, future studies should prioritize formal, model-tailored sample size calculations (eg, approaches proposed by Riley et al [74]) to ensure precise estimation and adequate statistical power. During the predictor selection process, several studies relied solely on univariate screening, which often fails to identify confounding factors and can lead to model overfitting. Therefore, predictor selection should not solely depend on univariate screening but should also be combined with clinical practice. Moreover, an increasing number of studies are using least absolute shrinkage and selection operator (LASSO) regression to handle high-dimensional data and select potential variables. By introducing a penalty term, LASSO regression reduces the estimates of extreme variables, thereby effectively enhancing the accuracy of model estimation and decreasing the likelihood of overfitting [75]. Moreover, the handling of missing data was suboptimal. Cases with missing data were directly excluded in 12 studies, which can introduce bias and reduce statistical power, while 9 studies failed to report how missing data were handled. A minority of studies used appropriate methods such as multiple imputation. Accurate data reporting and careful handling of missing observations help reduce model overfitting. It is recommended that future studies strengthen the management of missing data to ensure the integrity of the study. When dealing with continuous variables, transforming continuous data into categorical variables for modeling may lead to a significant loss of model efficacy [76]; however, 10 studies in our review performed such transformations without reporting standard definitions. Although data transformation can be considered to enhance the convenience of application for researchers during the clinical dissemination phase, it should be done with caution during development. Crucially, the majority of models lacked external validation. Although they demonstrated good discrimination in derivation cohorts, their performance remains inherently tied to their specific development settings.

Therefore, the primary implication of this review is methodological: Rather than endorsing specific existing tools for immediate clinical use, we emphasize the urgent need for better-designed research. Future studies must strictly adhere to TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) guidelines and PROBAST standards—specifically ensuring adequate sample sizes, appropriate handling of continuous variables and missing data, and rigorous external validation in independent, geographically distinct populations—to develop robust and transportable prediction models.

Limitations

This study has several limitations that warrant consideration. First, the predominance of cross-sectional and retrospective designs among the included studies restricts the scope of these models to the diagnostic identification of prevalent frailty. Consequently, they function as concurrent screening tools rather than prognostic instruments for predicting future incidence, precluding the ability to infer causal relationships between predictors and outcomes. Second, according to the PROBAST assessment, all included studies exhibited a high risk of bias, particularly within the analysis domain. This methodological weakness likely results in overoptimistic performance estimates and limits the transportability of these models to diverse clinical populations. Third, substantial statistical heterogeneity (I^2) was observed, stemming from variations in study design and modeling methodologies. We calculated 95% PIs to provide a more clinically meaningful estimate of the expected range of model performance in future settings. Finally, all included studies were conducted in China, and the review was restricted to English and Chinese literature. Although this reflects the rapid emergence of this research focus within China, the lack of geographic and ethnic diversity limits the generalizability of our findings to other global populations and health care systems.

Conclusion

This review provides the first comprehensive synthesis of models for risk stratification of physical and cognitive frailty in older adults with diabetes. The findings indicate that existing models demonstrate satisfactory pooled discriminative performance. Specifically, although the CIs confirm a robust average effect, the 95% PIs indicate that the distribution of predictive performance in future real-world settings is expected to vary across different clinical contexts, yet likely remaining within an acceptable range. Nevertheless, their clinical utility is currently constrained by significant methodological limitations. Specifically, the identified models rely heavily on readily available clinical and psychosocial predictors, such as

XSL•FO
RenderX

depression, age, regular exercise, and social activity, suggesting that early risk stratification is feasible in routine practice. However, the evidence is underpinned by a pervasive high risk of bias, primarily due to analytical shortcomings, small sample sizes, and a lack of rigorous external validation. Furthermore, the predominance of cross-sectional designs and the geographic restriction of studies to China limit the generalizability of these tools to broader global populations and their ability to function as true prognostic instruments for future risk. Consequently, although current models show promise for screening and identifying prevalent physical and cognitive frailty, they are not yet sufficiently robust for widespread deployment in diverse clinical settings. Future research must pivot from developing new, redundant models to conducting robust, prospective, multicenter studies that adhere strictly to TRIPOD guidelines. Emphasis should be placed on external validation and the development of longitudinal prognostic tools to ensure reliable, transportable, and clinically actionable risk stratification for this vulnerable population.

Acknowledgments

The authors declare that generative artificial intelligence (AI) was not used in the creation of this manuscript.

Funding

This study was funded by grants from the nursing research project of Sichuan Province (number H23003) and the research project of the Science and Technology Department of Sichuan Province (number 2024ZYD0338). The funders played no part in the study design, data collection, analysis, interpretation of the results, or the writing of the manuscript.

Disclaimer

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: XW
Data curation: LL, RZ, QJ
Formal analysis: XW, XX, HC
Funding acquisition: YL, RZ, RG
Methodology: XW, SM, XX, RG
Project administration: RG
Resources: LL
Software: QJ
Writing - original draft: XW, XX
Writing - review & editing: XW, XM, XX, LL, HC, YL, RZ, QJ, SL, RG, SM

Conflicts of Interest

None declared.

Multimedia Appendix 1
Inclusion criteria, search terms, performance measures, and assessments of bias.
[DOCX File, 495 KB - jmir_v28i1e84617_app1.docx]

Checklist 1
PRISMA checklist.
[PDF File, 106 KB - jmir_v28i1e84617_app2.pdf]

References

1. Genitsaridi I, Salpea P, Salim A, et al. 11th edition of the IDF Diabetes Atlas: global, regional, and national diabetes prevalence estimates for 2024 and projections for 2050. Lancet Diabetes Endocrinol 2026 Feb;14(2):149-156. [doi: 10.1016/S2213-8587(25)00299-2] [Medline: 41412135]

XSL•FO
RenderX

2. Kobayashi E, Linden-Santangeli NJ, Chan N, et al. Longitudinal metabolic trajectories in diabetes prevention program participants reveal subgroups with varying micro- and macrovascular complication risks. Diabetes Care 2025 Oct 1;48(10):1704-1712. [doi: 10.2337/dc25-0866] [Medline: 40857122]
3. Xu X, Wang F, Liao S, Liu J, Xiao L. Tele-cognitive behavioral therapy for the treatment of diabetes-related distress in individuals With diabetes mellitus: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2025 Dec 24;27:e80476. [doi: 10.2196/80476] [Medline: 41442672]
4. Ong KL, Stafford LK, McLaughlin SA. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet 2023 Jul 15;402(10397):203-234. [doi: 10.1016/S0140-6736(23)01301-6] [Medline: 37356446]
5. O'Neil H, Todd A, Pearce M, Husband A. What are the consequences of over and undertreatment of type 2 diabetes mellitus in a frail population? A systematic review. Endocrinol Diabetes Metab 2024 Mar;7(2):e00470. [doi: 10.1002/edm2.470] [Medline: 38411378]
6. Clegg A, Young J, Iliffe S, Rikkert MO, Rockwood K. Frailty in elderly people. The Lancet 2013 Mar;381(9868):752-762. [doi: 10.1016/S0140-6736(12)62167-9]
7. Wu Y, Xiong T, Tan X, Chen L. Frailty and risk of microvascular complications in patients with type 2 diabetes: a population-based cohort study. BMC Med 2022 Dec 8;20(1):473. [doi: 10.1186/s12916-022-02675-9] [Medline: 36482467]
8. Hanlon P, Fauré I, Corcoran N, et al. Frailty measurement, prevalence, incidence, and clinical implications in people with diabetes: a systematic review and study-level meta-analysis. Lancet Healthy Longev 2020 Dec;1(3):e106-e116. [doi: 10.1016/S2666-7568(20)30014-3] [Medline: 33313578]
9. Kong LN, Lyu Q, Yao HY, Yang L, Chen SZ. The prevalence of frailty among community-dwelling older adults with diabetes: a meta-analysis. Int J Nurs Stud 2021 Jul;119:103952. [doi: 10.1016/j.ijnurstu.2021.103952] [Medline: 34022743]
10. Huang ST, Chen LK, Hsiao FY. Clinical impacts of frailty on 123,172 people with diabetes mellitus considering the age of onset and drugs of choice: a nationwide population-based 10-year trajectory analysis. Age Ageing 2023 Jul 1;52(7):afad128. [doi: 10.1093/ageing/afad128] [Medline: 37505989]
11. Walston J, McBurnie MA, Newman A, et al. Frailty and activation of the inflammation and coagulation systems with and without clinical comorbidities: results from the Cardiovascular Health Study. Arch Intern Med 2002 Nov 11;162(20):2333-2341. [doi: 10.1001/archinte.162.20.2333] [Medline: 12418947]
12. Zhang X, Bo Y, Li Z, et al. Association between frailty and cognitive function: a pooled analysis of three ageing cohorts. Transl Psychiatry 2025 Nov 18;15(1):486. [doi: 10.1038/s41398-025-03674-z]
13. Li CL, Stanaway FF, Lin JD, Chang HY. Frailty and health care use among community-dwelling older adults with diabetes: a population-based study. Clin Interv Aging 2018;13:2295-2300. [doi: 10.2147/CIA.S183681] [Medline: 30519011]
14. Ferri-Guerra J, Aparicio-Ugarriza R, Salguero D, et al. The association of frailty with hospitalizations and mortality among community dwelling older adults with diabetes. J Frailty Aging 2020;9(2):94-100. [doi: 10.14283/jfa.2019.31] [Medline: 32259183]
15. Wang X, Chen Z, Li Z, et al. Association between frailty and risk of fall among diabetic patients. Endocr Connect 2020 Oct;9(10):1057-1064. [doi: 10.1530/EC-20-0405] [Medline: 33112808]
16. Santulli G, Sabatelli G, Wang B, et al. Interplay between frailty and cardiometabolic disorders: from pathophysiology to clinical implications. Cardiovasc Diabetol 2025 Dec 8;25(1):1. [doi: 10.1186/s12933-025-03022-x] [Medline: 41361441]
17. Zhang RH, Wang J, Wang Y, et al. Frailty and transitions across cardiometabolic disease states: evidence from multistate models in a 16-year Chinese cohort. NPJ Aging 2025 Nov 29;12(1):4. [doi: 10.1038/s41514-025-00301-5] [Medline: 41318578]
18. Fan J, Yu C, Guo Y, et al. Frailty index and all-cause and cause-specific mortality in Chinese adults: a prospective cohort study. Lancet Public Health 2020 Dec;5(12):e650-e660. [doi: 10.1016/S2468-2667(20)30113-4] [Medline: 33271078]
19. Haapanen MJ, Jansson Sigfrids F, Ylinen A, et al. Frailty outperforms conventional risk factors for predicting complications and death in type 1 diabetes. Diabetes Res Clin Pract 2025 Dec;230:112984. [doi: 10.1016/j.diabres.2025.112984] [Medline: 41240994]
20. Guo X, Pei J, Ma Y, et al. Cognitive frailty as a predictor of future falls in older adults: a systematic review and meta-analysis. J Am Med Dir Assoc 2023 Jan;24(1):38-47. [doi: 10.1016/j.jamda.2022.10.011] [Medline: 36423679]
21. Ren M, Guo H, Guo Y, Guo W, Zhu L. The risk prediction models for cognitive frailty in the older people in China: a systematic review and meta-analysis. BMC Geriatr 2025 May 22;25(1):365. [doi: 10.1186/s12877-025-05961-2] [Medline: 40405068]
22. Si H, Zhang Y, Zhao P, et al. Bidirectional relationship between diabetes and frailty in middle-aged and older adults: a systematic review and meta-analysis. Arch Gerontol Geriatr 2025 Aug;135:105880. [doi: 10.1016/j.archger.2025.105880] [Medline: 40319625]
23. Cheng M, He M, Ning L, et al. The impact of frailty on clinical outcomes among older adults with diabetes: a systematic review and meta-analysis. Medicine (Abingdon) 2024;103(26):e38621. [doi: 10.1097/MD.0000000000038621]
24. Sinclair AJ, Pennells D, Abdelhafiz AH. Hypoglycaemic therapy in frail older people with type 2 diabetes mellitus-a choice determined by metabolic phenotype. Aging Clin Exp Res 2022 Sep;34(9):1949-1967. [doi: 10.1007/s40520-022-02142-8] [Medline: 35723859]

XSL•FO
RenderX

25. Ji CH, Huang XQ, Li Y, Muheremu A, Luo ZH, Dong ZH. The relationship between physical activity, nutritional status, and sarcopenia in community- dwelling older adults with type 2 diabetes: a cross-sectional study. BMC Geriatr 2024 Jun 7;24(1):506. [doi: 10.1186/s12877-024-05038-6] [Medline: 38849763]
26. Sinclair AJ, Abdelhafiz AH. Metabolic impact of frailty changes diabetes trajectory. Metabolites 2023 Feb 16;13(2):295. [doi: 10.3390/metabo13020295] [Medline: 36837914]
27. Zhong W, Huang W, Deng H, Qiu S, Yang Q, Jia H. A randomized controlled trial to assess the efficacy of standardized tai chi in prefrail older adults with immunosenescence: design and protocol. BMC Complement Med Ther 2025 Jan 3;25(1):1. [doi: 10.1186/s12906-024-04732-7] [Medline: 39754159]
28. Sobrinho ACDS, de Paula Venancio RC, da Silva Rodrigues G, et al. Systematic review of interventions for pre-frail and frail older adults: evidence from clinical trials on frailty levels. Arch Gerontol Geriatr 2025 Jul;134:105851. [doi: 10.1016/j.archger.2025.105851] [Medline: 40262339]
29. Page MJ, Moher D, Bossuyt PM, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021 Mar 29;372:n160. [doi: 10.1136/bmj.n160] [Medline: 33781993]
30. McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: the PRISMA-DTA statement. JAMA 2018 Jan 23;319(4):388-396. [doi: 10.1001/jama.2017.19163] [Medline: 29362800]
31. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
32. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019 Jan 1;170(1):51-58. [doi: 10.7326/M18-1376] [Medline: 30596875]
33. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015 Jan 6;162(1):W1-W73. [doi: 10.7326/M14-0698] [Medline: 25560730]
34. Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 2014 Oct;11(10):e1001744. [doi: 10.1371/journal.pmed.1001744] [Medline: 25314315]
35. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019 Jan 1;170(1):W1-W33. [doi: 10.7326/M18-1377] [Medline: 30596876]
36. Šimundić AM. Measures of diagnostic accuracy: basic definitions. EJIFCC 2009 Jan;19(4):203-211. [Medline: 27683318]
37. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 6;327(7414):557-560. [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
38. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
39. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634. [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
40. Bu F, Deng XH, Zhan NN, et al. Development and validation of a risk prediction model for frailty in patients with diabetes. BMC Geriatr 2023 Mar 27;23(1):172. [doi: 10.1186/s12877-023-03823-3] [Medline: 36973658]
41. Cheng YM. Construction of a predictive model for the risk of frailty in hospitalized elderly patients with diabetes mellitus [Master's thesis]. : Gannan Medical University; 2024 URL: https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202501&filename=1024557771.nh&uniplatform=OVERSEA&v=aCGchnfEAAyXVFPDXN9Dm_iJvJYtdsledsnnFwYOOKq65nJCZmCdm-7Y3NSbCg_U [accessed 2026-01-22] [doi: 10.27959/d.cnki.ggnyx.2024.000179]
42. Dang X. Study on the construction of a model for predicting the risk of debility in hospitalized elderly patients with diabetes mellitus [Master's thesis]. : Changchun University of Chinese Medicine; 2024 URL: https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202501&filename=1024551416.nh&uniplatform=OVERSEA&v=saoWsbwFh_hMUrvE2hYKMbmPvq6dTUrm6EUZP7P8Emfni9qewTW8gjWKg0liRN7b [accessed 2026-01-22] [doi: 10.26980/d.cnki.gcczc.2024.000284]
43. Deng YH, Li N, Wang Y, Xiong C, Zou X. Risk factors and prediction nomogram of cognitive frailty with diabetes in the elderly. Diabetes Metab Syndr Obes 2023;16:3175-3185. [doi: 10.2147/DMSO.S426315] [Medline: 37867632]
44. Dong XT, Wang XH, Sheng YH, Wang GP, Gu T. Construction of frailty risk prediction model in elderly patients with diabetic retinopathy. Journal of Gannan Medical University 2023;43(12):1275-1281. [doi: 10.3969/j.issn.1001-5779.2023.12.013]
45. Du J, Zhang D, Chen Y, Zhang W. Development of a prediction model for frailty among older Chinese individuals with type 2 diabetes residing in the community. Public Health Nurs 2024;41(6):1271-1280. [doi: 10.1111/phn.13377] [Medline: 39101656]

XSL•FO
RenderX

46. Liang MY, Li R, Feng L, Qian WP. Construction and verification of a risk prediction model for cognitive frailty in older patients with chronic obstructive pulmonary disease and diabetes mellitus. J Int Med Res 2024 Sep;52(9):3000605241274211. [doi: 10.1177/03000605241274211] [Medline: 39224937]
47. Liu XX, Fan XZ. Establishment and validation of a prediction model for cognitive frailty in elderly patients with type 2 diabetes mellitus. Anhui Medical Journal 2024;45(12):1543-1548. [doi: 10.3969/j.issn.1000-0399.2024.12.014]
48. Liu Y. Construction and verification of cognitive frailty risk prediction model in elderly patients with type 2 diabetes [Master's thesis]. : Tianjin University of Traditional Chinese Medicine; 2023 URL: https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202501&filename=1023871608.nh&uniplatform=OVERSEA&v=bMx3cCr3OPvyjITKNyqnW0JcHM8BIfzLFhPkowxsujZYch6jIjA1m7lsencsQrYF [accessed 2026-01-22] [doi: 10.27368/d.cnki.gtzyy.2023.000432]
49. Ma SM, Ni LL, Guo L, Gu LP. Construction of a risk prediction model for cognitive decline in elderly patients with diabetic foot ulcers. Psychological Monthly 2025;20(2):52-54 [FREE Full text] [doi: 10.19738/j.cnki.psy.2025.02.014]
50. Meng L. Research on the influencing factors of cognitive frailty and the construction of risk prediction model in elderly diabetic patients [Master's thesis]. : Zunyi Medical University; 2023 URL: https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202501&filename=1024333258.nh&uniplatform=OVERSEA&v=OQcVHZS3INx0BTBGPkZnRQy4Deev1Fhcs49_DXQJd8j1MnacoqsS81_n7RhEcusZ [accessed 2026-01-26] [doi: 10.27680/d.cnki.gzyyc.2023.000437]
51. Tang QF, Wang JJ, Zhao HY, Wang HY. Establishment and validation of a frailty risk prediction model for elderly patients with diabetes mellitus. Chinese Nursing Research 2024;38(17):3065-3071. [doi: 10.12102/j.issn.1009-6493.2024.17.009]
52. Wang BJ, Liang Q, Liu Y, Cheng Y, Zhang CM. Construction and validation of frailty risk prediction model in elderly patients with diabetic foot. Mil Nurs 2024;41(5):6-10. [doi: 10.3696/j.issn.2097-1826.2024.05.002]
53. Wang SJ, Tan TT, Wang QQ, et al. Construction of a risk prediction model for cognitive decline in elderly hospitalized patients with type 2 diabetes based on decision tree. Sichuan Medical Journal 2025;46(2):204-210. [doi: 10.16252/j.cnki.issn1004-0501-2025.02.017]
54. Wang XW, Xu YL. Prediction of cognitive decline among elderly patients with type 2 diabetes mellitus. China Preventive Medicine Journal 2023;35(12):1037-1042. [doi: 10.19485/j.cnki.issn2096-5087.2023.12.006]
55. Wang Z, Zheng HF, Liang LL. Analysis of risk factors of elderly patients with type 2 diabetes complicated with frailty and establishment of prediction model. Int J Geriatr 2025;46(2):162-168. [doi: 10.3969/j.issn.1674-7593.2025.02.007]
56. Wang ZJ, Chen JY, Li MX. Establishment of a nomogram predictive model based on serum leptin-to-adiponectin ratio for cognitive frailty in elderly patients with type 2 diabetes mellitus. Shandong Medical Journal 2023;63(34):31-37. [doi: 10.3969/j.issn.1002-266X.2023.34.007]
57. Xi MX. Construction of prediction model for frailty in hospitalized elderly patients with diabetes mellitus [Master's thesis]. : Yan'an University; 2024 URL: https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202501&filename=1024455798.nh&uniplatform=OVERSEA&v=S5OEfMlXAp2MuMsWXc_GQtzuxBN-77f0i0NyJbd51w8xQlUaiOwjOVtxUElm5mcS [accessed 2026-01-22] [doi: 10.27438/d.cnki.gyadu.2024.000104]
58. Yin YY. Construction and verification of frailty risk prediction model in elderly diabetic patients [Master's thesis]. : Hebei University; 2024 URL: https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202501&filename=1024530021.nh&uniplatform=OVERSEA&v=SApAmgAzcplPDX5BRNmZPhLMUnWViopteNW4E_KeL-FHGr3Rn323ES_7vpnrm8Sb [accessed 2026-01-22] [doi: 10.27103/d.cnki.ghebu.2024.000915]
59. Zhang YJ, Tian XF, Zhang H, Tang ZY, Zhang ZY. Influencing factors of cognitive decline in elderly patients with type 2 diabetes mellitus and hypertension and construction of nomogram model for predicting its risk. Practical Journal of Cardiac Cerebral Pneumal and Vascular Disease 2024;32(12):49-54 [FREE Full text] [doi: 10.12114/j.issn.1008-5971.2024.00.257]
60. Zheng XM. Research on the construction of frailty prediction model for elderly hospitalized patients with diabetes based on machine learning algorithm. : Yangtze University; 2024 URL: https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202501&filename=1024659916.nh&uniplatform=OVERSEA&v=iU31hKCfxDkcQQIIuCxhedEIj2mwaaiQ5ki4TyGPmRJytm6uY7RU4TpTMk0kFlYT [accessed 2026-01-22] [doi: 10.26981/d.cnki.gjhsc.2024.000756]
61. Xiao RF, Wang R, Xu L, Xu MJ. Risk factors for the development of frailty in Chinese elderly diabetic patients and the predictive value of the nomogram model. Journal of Lanzhou University (Medical Sciences) 2025;51(10):20-25. [doi: 10.13885/j.issn.2097-681X.2025.10.004]
62. Wu JQ, Fang SZ, Li K, et al. Construction and validation of frailty risk prediction model for hospitalized elderly patients with type 2 diabetes based on machine learning and SHAP explainabilility. Journal of Nursing (China) 2025;32(11):7-12. [doi: 10.16460/j.issn2097-6569.2025.11.007]
63. Yu Q, Yu H. Development and validation of a risk prediction model for cognitive frailty in elderly patients with type 2 diabetes mellitus. J Clin Nurs 2025 Aug;34(8):3261-3275. [doi: 10.1111/jocn.17508] [Medline: 39809596]

XSL•FO
RenderX

64. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019 Jun;110:12-22. [doi: 10.1016/j.jclinepi.2019.02.004] [Medline: 30763612]
65. Chu W, Chang SF, Ho HY, Lin HC. The relationship between depression and frailty in community-dwelling older people: a systematic review and meta-analysis of 84,351 older adults. J Nurs Scholarsh 2019 Sep;51(5):547-559. [doi: 10.1111/jnu.12501] [Medline: 31328878]
66. Borges MK, Aprahamian I, Romanini CV, et al. Depression as a determinant of frailty in late life. Aging Ment Health 2021 Dec;25(12):2279-2285. [doi: 10.1080/13607863.2020.1857689] [Medline: 33307781]
67. Wang Q, Wang J, Dai G. Prevalence, characteristics, and impact on health outcomes of frailty in elderly outpatients with diabetes: a cross-sectional study. Medicine (Abingdon) 2023;102(47):e36187. [doi: 10.1097/MD.0000000000036187]
68. Ferrucci L, Fabbri E. Inflammageing: chronic inflammation in ageing, cardiovascular disease, and frailty. Nat Rev Cardiol 2018 Sep;15(9):505-522. [doi: 10.1038/s41569-018-0064-2] [Medline: 30065258]
69. Zeng M, Li Y, Zhu Y, Sun Y. Inflammatory markers and clinical factors as key independent risk factors for frailty: a retrospective study. BMC Geriatr 2025 Jun 4;25(1):404. [doi: 10.1186/s12877-025-06033-1] [Medline: 40468183]
70. Straub RH. Interaction of the endocrine system with inflammation: a function of energy and volume regulation. Arthritis Res Ther 2014 Feb 13;16(1):203. [doi: 10.1186/ar4484] [Medline: 24524669]
71. Sutkowy P, Woźniak A, Mila-Kierzenkowska C, et al. Physical activity vs. redox balance in the brain: brain health, aging and diseases. Antioxidants (Basel) 2021 Dec 30;11(1):95. [doi: 10.3390/antiox11010095] [Medline: 35052600]
72. Wen L, Lu Y, Li X, An Y, Tan X, Chen L. Association of frailty and pre-frailty with all-cause and cardiovascular mortality in diabetes: three prospective cohorts and a meta-analysis. Ageing Res Rev 2025 Apr;106:102696. [doi: 10.1016/j.arr.2025.102696] [Medline: 39971101]
73. Lyu Q, Guan CX, Kong LN, Zhu JL. Prevalence and risk factors of cognitive frailty in community-dwelling older adults with diabetes: a systematic review and meta-analysis. Diabet Med 2023 Jan;40(1):e14935. [doi: 10.1111/dme.14935] [Medline: 35962598]
74. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ 2020 Mar 18;368:m441. [doi: 10.1136/bmj.m441] [Medline: 32188600]
75. Zarringhalam K, Degras D, Brockel C, Ziemek D. Robust phenotype prediction from gene expression data using differential shrinkage of co-regulated genes. Sci Rep 2018 Jan 19;8(1):1237. [doi: 10.1038/s41598-018-19635-0] [Medline: 29352257]
76. Rhemtulla M, Brosseau-Liard P, Savalei V. When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. Psychol Methods 2012 Sep;17(3):354-373. [doi: 10.1037/a0029315] [Medline: 22799625]

Abbreviations

AUC: area under the curve
CHARMS: Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modeling Studies
CNKI: China National Knowledge Infrastructure
EPV: events per variable
FN: false negative
FP: false positive
FRAIL: Fatigue, Resistance, Ambulation, Illnesses, and Loss of Weight
IDF: International Diabetes Federation
LASSO: least absolute shrinkage and selection operator
LR: logistic regression
MeSH: Medical Subject Headings
ML: machine learning
PI: prediction interval
PITROS: Participants, Index Test, Target Conditions, Reference Standard, Outcomes, Settings
PRESS: Peer Review of Electronic Search Strategies
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST: Prediction Model Study Risk Of Bias Assessment Tool
TN: true negative
TP: true positive
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

XSL•FO
RenderX

Edited by S Brini; submitted 23.Oct.2025; peer-reviewed by A Shivanna, CY Su; revised version received 04.Jan.2026; accepted 05.Jan.2026; published 29.Jan.2026.

Please cite as:
Wang X, Meng S, Xiao X, Lu L, Chen H, Li Y, Zhang R, Jiang Q, Liu S, Gao R
Characterization of Models for Identifying Physical and Cognitive Frailty in Older Adults With Diabetes: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e84617
URL: https://www.jmir.org/2026/1/e84617
doi:10.2196/84617

© Xia Wang, Shujie Meng, Xiang Xiao, Liu Lu, Hongyan Chen, Yong Li, Rong Zhang, Qiwu Jiang, Shan Liu, Ru Gao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Predictive Performance of Artificial Intelligence Algorithms for Gestational Diabetes Mellitus in Pregnant Women: Systematic Review and Meta-Analysis

Yingni Liang[1*], MM; Anran Dai[1*], MM; Meiyan Luo[2*], MM; Zhuolian Zheng[1], MM; Jiayu Shen[1], MM; Yinhua Su[1*], PhD; Zhongyu Li[1*], PhD

[1]School of Nursing, University of South China, No. 28, Changsheng West Road, Hengyang, Hunan, China

[2]Department of Obstetrics, Second Affiliated Hospital of the University of South China, Hunan, China

[*]these authors contributed equally

Corresponding Author:
Zhongyu Li, PhD
School of Nursing, University of South China, No. 28, Changsheng West Road, Hengyang, Hunan, China

Abstract

Background: Gestational diabetes mellitus (GDM) is a common complication during pregnancy, with its incidence increasing year by year. It poses numerous adverse health effects on both mothers and newborns. Accurate prediction of GDM can significantly improve patient prognosis. In recent years, artificial intelligence (AI) algorithms have been increasingly used in the construction of GDM prediction models. However, there is still no consensus on the most effective algorithm or model.

Objective: This study aimed to evaluate and compare the performance of existing GDM prediction models constructed using AI algorithms and propose strategies for enhancing model generalizability and predictive accuracy, thereby providing evidence-based insights for the development of more accurate and effective GDM prediction models.

Methods: A comprehensive search was conducted across PubMed, Web of Science, Cochrane Library, EMBASE, Scopus, and OVID, covering publications from the inception of databases to June 1, 2025, to include studies that developed or validated GDM prediction models based on AI algorithms. Study selection, data extraction, and risk of bias assessment using the Prediction Model Risk of Bias Assessment Tool were performed independently by 2 reviewers. A bivariate mixed-effects model was used to summarize sensitivity and specificity and to generate a summary receiver operating characteristic (SROC) curve, calculating area under the curve (AUC). The Hartung-Knapp-Sidik-Jonkman method was further used to adjust for the pooled sensitivity and specificity. Between-study standard deviation (τ) and variance (τ^2) were extracted from the bivariate model to quantify absolute heterogeneity. The Deek test was used to evaluate small-study effects among included studies. Additionally, subgroup analysis and meta-regression were conducted to compare the performance differences among algorithms and to explore sources of heterogeneity.

Results: Fourteen studies reported on the predictive value for AI algorithms for GDM. After adjustment with the Hartung-Knapp-Sidik-Jonkman method, the pooled sensitivity and specificity were 0.78 (95% CI 0.69 - 0.86; τ=0.15, τ^2=0.02; PI 0.47 - 1.09) and 0.85 (95% CI 0.78 - 0.92; τ=0.11, τ^2=0.01; PI 0.59 - 1.11), respectively. The SROC curve showed that the AUC for predicting GDM using AI algorithms was 0.94 (95% CI 0.92 - 0.96), indicating a strong predictive capability. Deek test (*P*=.03) and the funnel plot both showed clear asymmetry, suggesting the presence of small-study effects. Subgroup analysis showed that the random forest algorithm exhibited the highest sensitivity (0.83, 95% CI 0.74 - 0.93), while the extreme gradient boosting algorithm exhibited the highest specificity (0.82, 95% CI 0.77 - 0.87). Meta-regression further revealed an evaluation in predictive accuracy in prospective study designs (regression coefficient=2.289, *P*=.001).

Conclusions: Unlike previous narrative reviews, this systematic review innovatively provided a comparative and quantitative synthesis of AI algorithms for GDM prediction. This established an evidence-based framework to guide model selection and identified a critical evidence gap. The key implication for real-world application was the demonstrated necessity of local validation before clinical adoption. Therefore, future work should focus on large-scale, prospective validation studies to develop clinically applicable tools.

Trial Registration: PROSPERO CRD42025645913; https://www.crd.york.ac.uk/PROSPERO/view/CRD42025645913

(J Med Internet Res 2026;28:e79729) doi:10.2196/79729

XSL•FO
RenderX

KEYWORDS

gestational diabetes mellitus; artificial intelligence; prediction; meta-analysis; PRISMA; Preferred Reporting Items for Systematic Reviews and Meta-Analysis

Introduction

Gestational diabetes mellitus (GDM) is one of the most common metabolic disorders during pregnancy, characterized by glucose metabolism abnormalities that first appear during gestation [1]. The incidence of GDM has risen to 15.8% due to factors like increased childbearing age, dietary changes, and pre-pregnancy obesity [2-4]. GDM not only significantly increased the risk of adverse pregnancy outcomes for pregnant women, such as macrosomia, preterm birth, and preeclampsia, but also had a profound impact on the long-term health of their offspring, including an increased risk of developing obesity, type 2 diabetes, and other metabolic disorders in the future [5-7]. Therefore, early prediction and management of GDM could effectively reduce the incidence of GDM and its associated maternal and neonatal complications, thereby optimizing perinatal care and improving long-term health outcomes.

The emergence of artificial intelligence (AI) algorithms in medicine has opened new frontiers for predictive analytics, offering the potential to model complex, non-linear interactions within multidimensional health data [8]. In fields such as oncology, cardiology, and endocrinology, AI-driven prediction models have demonstrated superior discriminative accuracy compared to conventional statistical approaches, largely by capturing subtle patterns and interactions among risk factors that traditional methods might overlook [9-12]. This capability was particularly salient for GDM, a condition influenced by a dynamic interplay of genetic, metabolic, hormonal, and lifestyle factors [13].

Building on this general capability, the application of AI algorithms for the specific task of GDM prediction has gained considerable momentum, with primary attention to 2 domains: machine learning (ML) and deep learning (DL) [14-16]. Commonly used ML algorithms, such as random forest (RF), support vector machine, and extreme gradient boosting (XGBoost), have been applied to structured clinical and biomarker data, while DL algorithms typically use neural networks to exploit high-dimensional inputs, including eHealth records and even image-based data [17]. Despite promising reported accuracies, a critical and persistent challenge is the marked heterogeneity in model performance across different populations and settings [18-20]. The ML model developed by Gallardo et al [21], based on routine early-pregnancy examination data, showed high predictive accuracy in a particular population but performed poorly in other GDM populations due to differences in data characteristics. This discrepancy revealed a severe methodological inconsistency in these studies, such as the lack of standardized data preprocessing, non-uniform validation strategies, and incomplete reporting of performance metrics. This heterogeneity made it difficult to directly compare and integrate the results of different studies.

Consequently, although a growing body of primary studies investigating AI models for GDM prediction, the evidence in this field remained fragmented and methodologically heterogeneous. Currently, for the prediction of GDM, there was still a lack of systematic reviews and meta-analyses that could directly compare multiple AI algorithms head-to-head, quantitatively assess their cross-population applicability, and systematically examine methodological rigor. The majority of existing original studies have developed single-algorithm models and validated them only within mono-ethnic or single-center cohorts [16,17,21,22]. Consequently, clinicians lack the high-level evidence required to determine which algorithm is superior and whether reported accuracies generalize to other settings, which markedly impedes the credible clinical adoption and broader dissemination of AI-based prediction models.

To address these evidence gaps, this systematic review and meta-analysis aimed to quantitatively synthesize the predictive performance of prediction models constructed using AI algorithms across different scenarios for GDM, compare the effectiveness of different AI algorithms, and identify the key factors influencing performance. By providing a rigorous, evidence-based framework for evaluating and comparing AI prediction models in GDM, this systematic review sought to inform the future development of more robust, generalizable, and clinically actionable tools, thereby supporting efforts toward early identification, risk stratification, and personalized management of GDM.

Methods

Registration and Protocol

This systematic review adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) 2020 extended checklist, with extensions for Diagnostic Test Accuracy (PRISMA-DTA) and literature search reporting (PRISMA-S) [23-25]. The protocol was prospectively registered with PROSPERO (International Prospective Register of Systematic Reviews; ID CRD42025645913). And the registration was completed on February 13, 2025, prior to the commencement of data extraction and analysis (Checklist 1).

Information Sources and Search Strategy

A comprehensive search was conducted across 6 databases, including PubMed, Web of Science, Cochrane Library, Scopus, EMBASE, and OVID, from the inception of each database to June 1, 2025. To enhance the accuracy of the search results and avoid the omission of relevant studies, the research term developed a rigorous search strategy by combining Medical Subject Headings terms, keywords, and synonyms. No previously published search filters were applied so as to maintain a highly sensitive search strategy. Table 1 summarizes the core search concepts and representative terms. And the detailed search strategy is presented in Multimedia Appendix 1. In addition, we also reviewed the reference lists of relevant

XSL•FO
RenderX

literature, particularly systematic reviews related to the topic of this study, and conducted additional searches in the electronic databases to minimize the omission of the key literature as much as possible. All searches were conducted under the supervision of an academic librarian.

Table . Search strategy using the population, Intervention framework for artificial intelligence–based gestational diabetes mellitus prediction studies.

Concept	Key terms (PubMed example)
Population	"Gestational Diabetes Mellitus" OR "Pregnancy-induced Diabetes" OR "GDM" OR "Diabetes in Pregnancy" OR "Maternal Diabetes"
Intervention	"Artificial Intelligence" OR "Machine Learning Algorithms" OR "Deep Learning Algorithms" OR "Ensemble Learning Algorithms"

Eligibility Criteria

To screen out the original studies relevant to this systematic review from the retrieved literature, detailed inclusion and exclusion criteria were defined (Textbox 1).

Textbox 1. Inclusion and exclusion criteria

Inclusion criteria

- Studies that conducted among pregnant women with gestational diabetes mellitus (GDM) or those at risk of developing GDM.
- Studies that completely constructed one or more predictive models for predicting GDM.
- Studies that used AI algorithms for the construction of a predictive model.
- Studies published in English.

Exclusion criteria

- Reviews, meta-analysis, protocols, letters, conference abstracts, case reports, and animal studies.
- Studies on the predictive accuracy of single-factor predictors.
- Studies only conducted a risk factor analysis without constructing a predictive model.
- Studies did not include any outcome measures for assessing the predictive accuracy of the predictive model.

Selection and Data Collection Process

Following the completion of the systematic research, all records were imported into the reference management software Endnote 21. After removing duplicate records, 2 reviewers independently examined the titles and abstracts of each study. Studies not reporting AI-based predictive models were discarded. Subsequently, a thorough full-text assessment was conducted for all studies that initially met the criteria, and the reasons for excluding each study were recorded in detail. In the predesigned Excel spreadsheet, data was extracted from studies that qualified based on the inclusion criteria. The extracted information included: characteristics of the study (authors, country, publication year, study design, and sample size), characteristics of the participants (diagnostic criteria for GDM and number of GDM cases), intervention features (model development process, types of AI algorithms used, methods for handling missing data, predictors, and model validation), and study outcomes (assessment of model accuracy). In cases where the information presented in the literature was ambiguous, the researchers would proactively contact the corresponding author to acquire the relevant information. The aforementioned process was independently conducted by 2 authors. Any discrepancies were discussed and resolved with a third author.

Study Risk-of-Bias Assessment

The Prediction Model Risk of Bias Assessment Tool (PROBAST) was used to assess the risk of bias (ROB) for each study. PROBAST consisted of four domains: participants, predictors, outcomes, and analysis [26]. Based on the responses to the items provided in the PROBAST checklist, a ROB rating (high, low, or unclear) was assigned to each domain. The criteria for assessment were detailed below: (1) the overall ROB was deemed "low" when all domains were classified as "low risk"; (2) the overall ROB was considered "high" if any domain was identified as "high risk"; (3) the overall ROB was determined to be "unclear" when there was at least one domain with an "unclear" rating, while the other domains were classified as "low risk" [26]. The quality assessment was conducted by the same 2 authors who performed the study selection and data extraction. Any disagreements between the 2 authors were resolved through consultation with a third author.

Effect Measures and Synthesis Methods

Statistical analyses were performed using Stata (version 17.0; StataCorp LLC), R (version 4.2.0; R Development Core Team), and Meta DiSc (version 1.4; Clinical Biostatistics Unit) software. A bivariate mixed-effects model was used to pool sensitivity and specificity, generate a summary receiver operating characteristic (SROC) curve, and calculate area under

XSL•FO
RenderX

the curve (AUC). The Hartung-Knapp-Sidik-Jonkman method was further used to adjust the pooled estimates. All results were reported with 95% CI values. Between-study standard deviation (τ) and variance (τ^2) were extracted from the bivariate model to quantify absolute heterogeneity. And prediction intervals (PIs) were subsequently computed to estimate the range within which the true sensitivity or specificity of a future study was expected to lie, providing a clinically interpretable measure of real-world dispersion. Moreover, the Fagan nomogram was used to explore the relationship between pretest probability, likelihood ratios (LR), and post-test probability. The LR dot plot, divided into 4 quadrants based on the strength of evidence threshold, was used to determine the exclusion and confirmation of the AI model. Additionally, a bivariate boxplot was drawn to detect heterogeneity caused by threshold effects. And subgroup analysis was used to compare the predictive capabilities of different AI algorithms in GDM prediction. In line with current recommendations for interpreting heterogeneity, we quantified real-world dispersion primarily using the τ, τ^2, and calculated PIs as the key measure of practical uncertainty [27]. The I^2 statistic was considered but not emphasized, given its limited use informing the generalizability of findings compared to PIs [27]. Based on the clinical and methodological characteristics anticipated to cause heterogeneity across studies, a meta-regression analysis was used to explore and explain such heterogeneity. It aimed to uncover potential influencing factors and analyze which variables might account for variations in the effect sizes. And the Deek test was used to evaluate small-study effects among the included studies, with $P<.05$ indicating funnel-plot asymmetry.

Ethical Considerations

This systematic review and meta-analysis was conducted exclusively with published aggregate data. No individual-level or identifiable participant information was involved. Therefore, informed consent, institutional review board approval, privacy protection, and participant compensation were not applicable.

Results

Study Selection and Characteristics of Included Studies

A total of 2790 studies were retrieved from the database. After removing duplicates, the titles and abstracts of 1455 studies were reviewed, and the full texts of 116 studies were screened. Finally, 22 studies were included in this study, with 8 studies [14,15,28-35] being included in the systematic review and 14 studies being incorporated into the meta-analysis [15,16,21,22,28,36-44]. The detailed process of the literature screening is illustrated in Figure 1. The fundamental characteristics of the included studies are presented in Table 2. The included studies were conducted in 11 countries, with 12 being single-center studies [14,16,21,28,30-32,36,37,40,41,43], 10 being multicenter studies [15,22,29,33-35,38,39,42,44], 14 being retrospective studies [14,16,21,22,28-30,32,36,37,39-41,43], and 8 being prospective studies [15,31,33-35,38,42,44]. All 22 studies used ML algorithms, and 2 of them further used DL algorithms [16,42]. To evaluate the predictive performance of the models, 12 studies conducted internal validation [14,15,21,22,31,32,34,37,38,40-42], and 8 studies performed external validation [16,28-30,32,37,39,42]. Multimedia Appendix 2 provides a detailed record of the model performance parameters for each study.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) flow diagram for study selection. This figure illustrates the process of identifying, screening, and selecting studies for inclusion in the systematic review, showing the number of records at each stage and reasons for exclusions. AI: artificial intelligence; GDM: gestational diabetes mellitus.

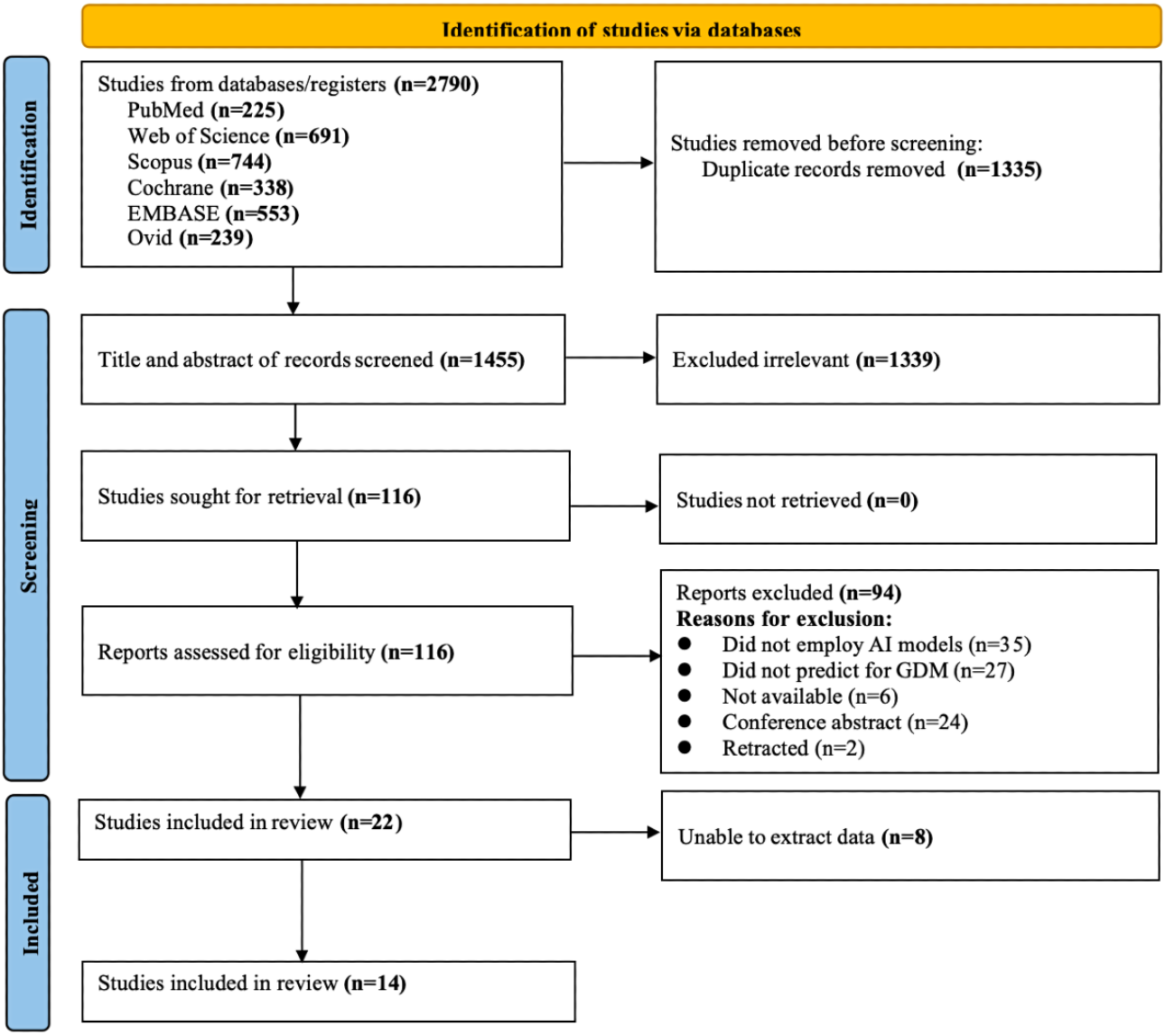

XSL•FO
RenderX

Table . General characteristics of the 22 studies included in the systematic review of artificial intelligence models for gestational diabetes mellitus prediction.

Study	Country	Study type	Single-center or multicenter	Diagnostic criteria	Sample size	Type of model
Belsti et al (2023) [22]	Australia	Retrospective	Multicenter	IADPSG[a]	48,502	ML[b]
Ali et al (2022) [33]	United Arab Emirates	Prospective	Multicenter	IADPSG	3858	ML
Wu et al (2021) [16]	China	Retrospective	Single-center	IADPSG	32,190	ML and DL[c]
Lin and Fang (2023) [36]	China	Retrospective	Single-center	IADPSG	406	ML
Ye et al (2020) [37]	China	Retrospective	Single-center	IADPSG	22,242	ML
Wang et al (2022) [30]	China	Retrospective	Single-center	IADPSG	1075	ML
Wu et al (2021) [28]	China	Retrospective	Single-center	IADPSG	17,005	ML
Wang et al (2021) [38]	China	Prospective	Multicenter	IADPSG	1139	ML
Syngelaki et al (2025) [31]	England	Prospective	Single-center	NICE[d]	41,587	ML
Donovan et al (2019) [39]	America	Retrospective	Multicenter	NIH[e]	11,56,708	ML
Kaya et al (2024) [40]	Turkey	Retrospective	Single-center	IADPSG	97	ML
Hu et al (2023) [41]	China	Retrospective	Single-center	IADPSG	735	ML
Liu et al (2022) [34]	China	Prospective	Multicenter	IADPSG	6848	ML
Lee et al (2021) [42]	Korea	Prospective	Multicenter	NIH	1443	ML and DL
Kumar et al (2022) [35]	Singapore	Prospective	Multicenter	IADPSG	222	ML
Bigdeli et al (2025) [14]	Iran	Retrospective	Single-center	NIH	743	ML
Kurt et al (2023) [15]	Turkey	Prospective	Multicenter	IADPSG	489	DL
Cubillos et al (2023) [21]	Chile	Retrospective	Single-center	IADPSG	1611	ML
Ding et al (2024) [43]	China	Retrospective	Single-center	IADPSG	554	ML
Kang et al (2023) [29]	Korea	Retrospective	Multicenter	NIH	34,387	ML
Zhao et al (2025) [32]	China	Retrospective	Single-center	IADPSG	1,03,172	ML
Liu et al (2020) [44]	China	Prospective	Multicenter	IADPSG	19,331	ML

XSL•FO
RenderX

[a]IADPSG: International Association of Diabetes and Pregnancy Study Groups.

[b]ML: machine learning.

[c]DL: deep learning.

[d]NICE: National Institute for Health and Care Excellence.

[e]NIH: National Institutes of Health.

ROB in Studies

Based on the PROBAST checklist, each study was assessed in terms of participants, predictors, outcomes, and analysis (Figure 2). The majority of studies consistently demonstrated low overall ROB and high applicability, indicating reliable methodology. However, in terms of overall ROB, 5 studies were rated as "unclear" [15,33,35,37,40]. One study was identified as having "high risk" in overall applicability due to insufficiently detailed descriptions of predictors used in model development [15]. Additionally, within the analysis domain, 2 studies were rated as "unclear" due to relatively small sample size, and this might also be one of the potential sources of bias [35,40]. In summary, most studies exhibited strong methodological quality and applicability. The detailed quality assessment of the included studies is detailed in Multimedia Appendix 3.

Figure 2. Risk assessment of the included models. This graph summarizes the methodological quality of the included prediction models, categorizing ROB across key domains to help readers assess the reliability of the evidence. ROB: risk of bias

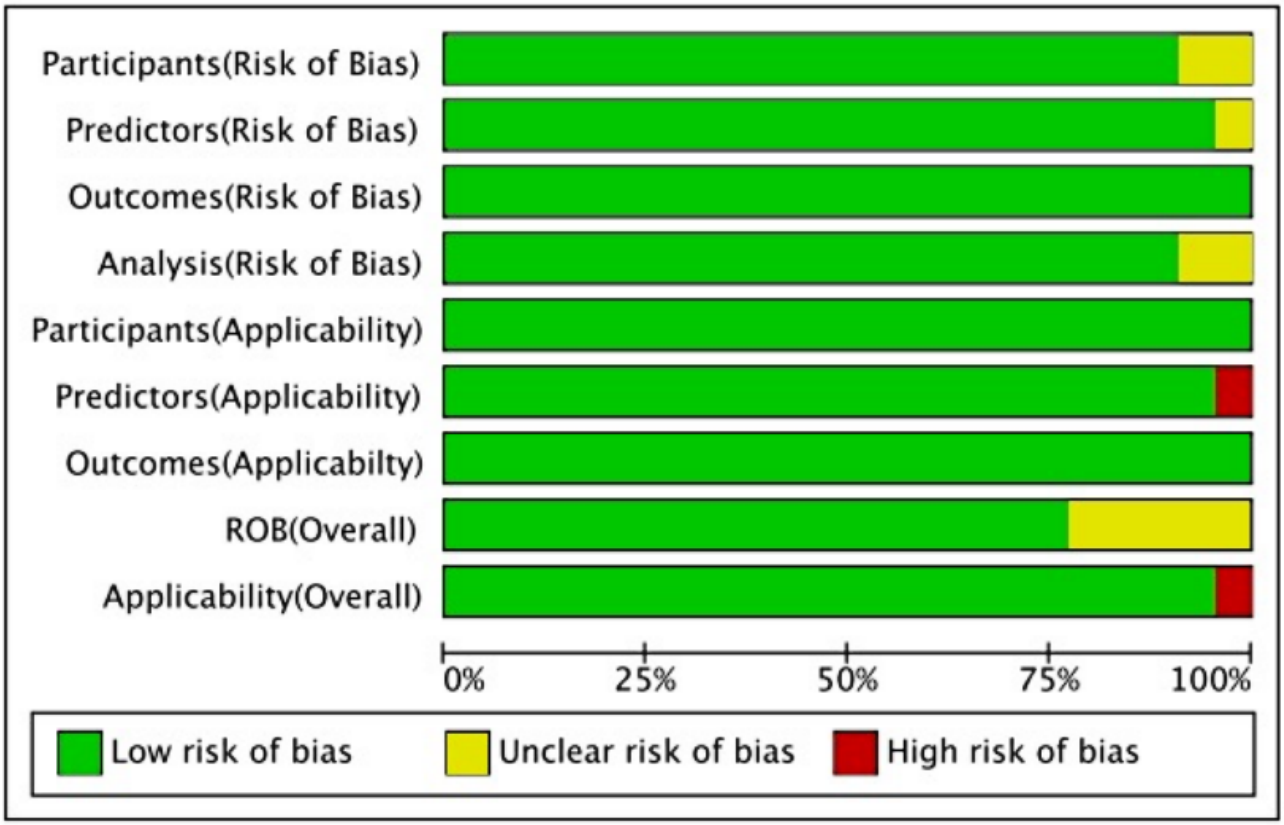

Performance of AI Algorithms for GDM

A total of 14 studies conducted on independent patient populations were included with the aim of evaluating the predictive value of AI algorithms for GDM [15,16,21,22,28,29,36-43]. Since some studies used multiple AI algorithms to construct several prediction models, this systematic review selected the model with the best performance reported in each study for meta-analysis. The pooled sensitivity was 0.78 (95% CI 0.69 - 0.86; τ=0.15, τ^2=0.02; PI 0.47 - 1.09), and specificity was 0.85 (95% CI 0.78 - 0.92; τ=0.11, τ^2=0.01; PI 0.59 - 1.11) after adjustment for the Hartung-Knapp-Sidik-Jonkman method (Figure 3). The wide PIs indicated substantial heterogeneity in real-world performance across populations, supporting the recommendation for local validation in the target population before clinical deployment. Note that the upper bounds of the PIs exceeded 1.0, specifically reaching 1.09 and 1.11. This occurred as a result of back-transformation from the logit scale and was a recognized statistical artifact, which did not indicate actual predictive performance greater than 100%.

As depicted in Figure 4, the SROC curve revealed the AUC of 0.94 (95% CI 0.92 - 0.96) for AI algorithms predicting GDM, suggesting a strong predictive capability. Furthermore, we set the pretest probability at 20% based on the pretest probability of the disease. At this level, when patients were predicted to have GDM by the AI algorithms, the true positive rate was 79%, and when the prediction was not GDM, the false negative rate was 4% (Figure 5). Moreover, the model demonstrated a positive LR of 15 and a negative LR of 0.17 (Figure 5). However, the

XSL•FO
RenderX

summary LR plot for the AI algorithms was located in the upper right quadrant (positive LR>10 and negative LR>0.1: confirmation only), and the individual studies were widely dispersed (Figure 6). The results indicated that while the prediction models built on AI algorithms generally demonstrated acceptable performance, they were not yet adequate for definitive diagnosis or exclusion of GDM. Additionally, there were notable variations in performance among the existing models.

XSL•FO
RenderX

Figure 3. Forest plots of sensitivity and specificity in 14 included studies on using artificial intelligence algorithms for predicting gestational diabetes mellitus. Each horizontal line represents the performance estimate of an individual study, with the diamond indicating the pooled result. The wide variability across studies highlights substantial heterogeneity in model performance [15,16,21,22,28,36-44]. DNN: deep neural network; GBDT: gradient-boosting decision tree; LR: logistic regression; RF: random forest; RNN-LSTM: recurrent neural network-long short-term memory; SVM: support vector machine; XGBoost: extreme-gradient boosting.

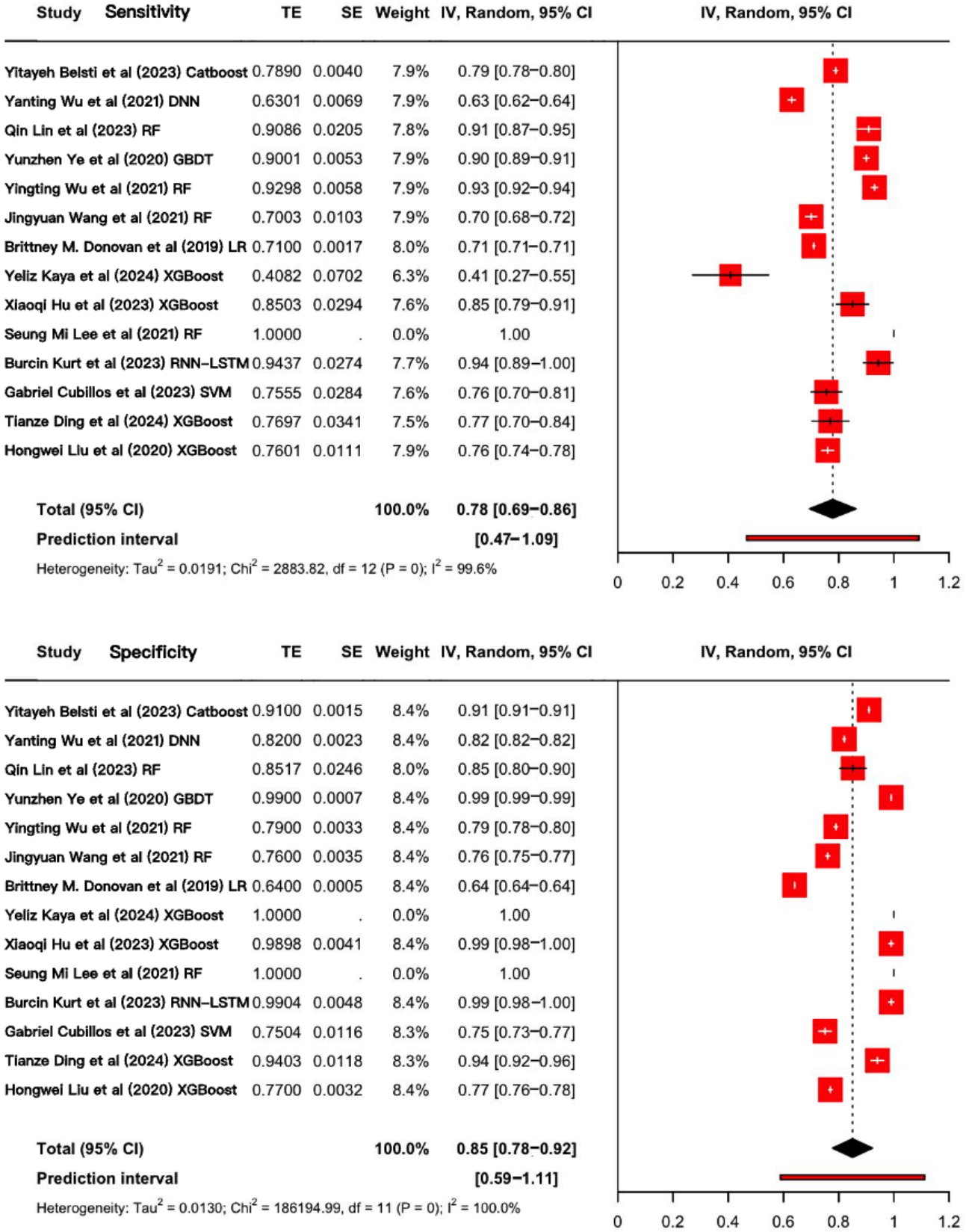

XSL•FO
RenderX

Figure 4. SROCs of included studies. This plot shows the overall diagnostic accuracy of artificial intelligence algorithms, with the curve position indicating the trade-off between sensitivity and specificity across different thresholds. The high AUC (0.87) reflects strong average discriminatory power. SROC: summary receiver operating characteristic curve; AUC: area under the curve.

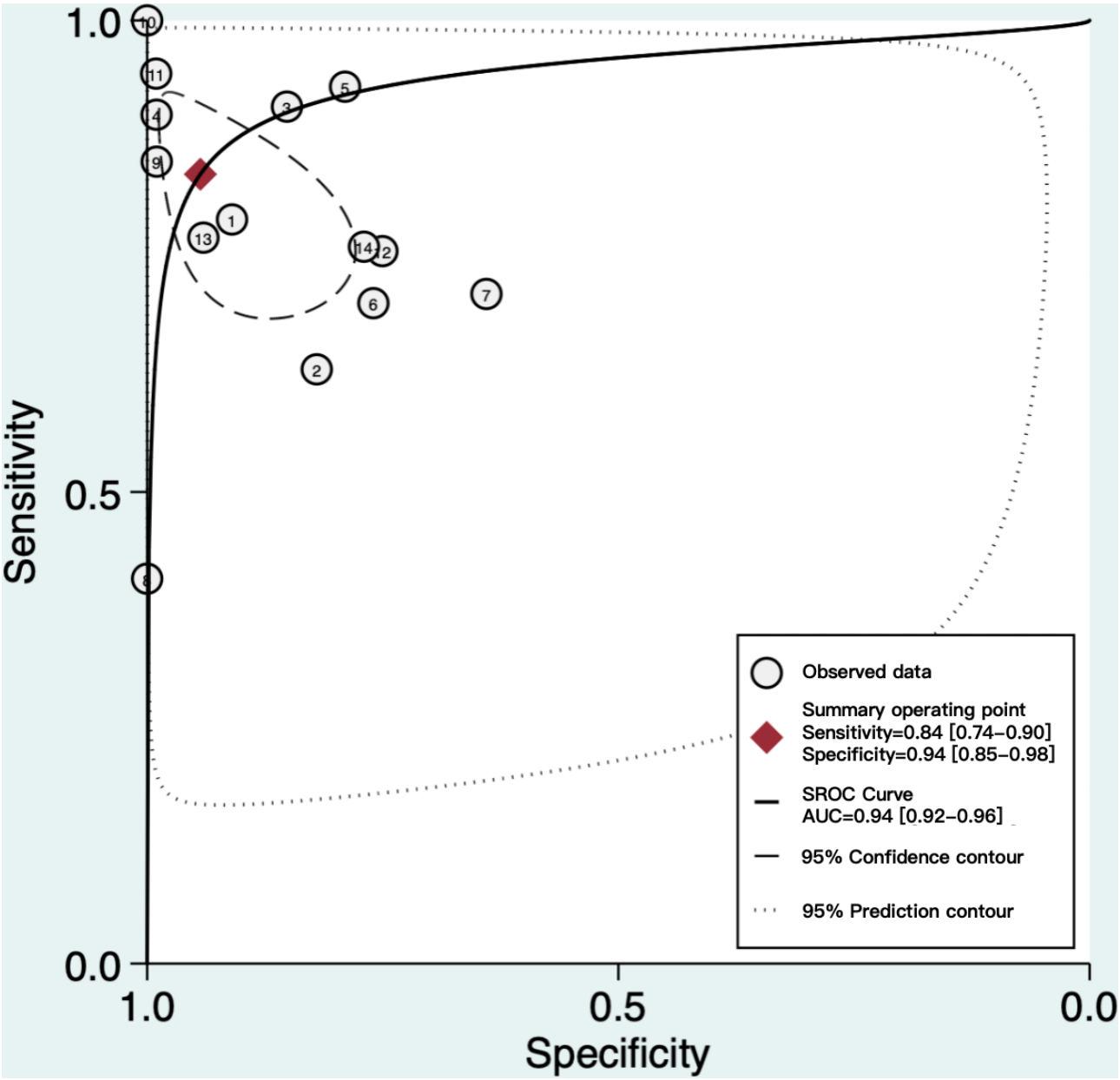

XSL•FO
RenderX

Figure 5. Fagan nomogram of artificial intelligence (AI) algorithms for predicting gestational diabetes mellitus. The first column of this nomogram represents the pretest probability, the second column represents the likelihood ratio, and the third shows the posttest probability. Interpretation: This tool helps clinicians estimate how a positive or negative test result changes the probability of gestational diabetes mellitus. The limited shift from pre to posttest probability indicates that current AI models provide only modest diagnostic value in clinical practice.

XSL•FO
RenderX

Figure 6. Likelihood ratio dot plot of artificial intelligence algorithms for predicting gestational diabetes mellitus. The position of the summary point in the upper right quadrant indicates that current artificial intelligence algorithms have confirmation but limited exclusion ability (positive likelihood ratio>10 and negative likelihood ratio>0.1), supporting their role as screening adjuncts rather than definitive diagnostic tools. LRN: likelihood ratio for a negative test; LRP: likelihood ratio for a positive test.

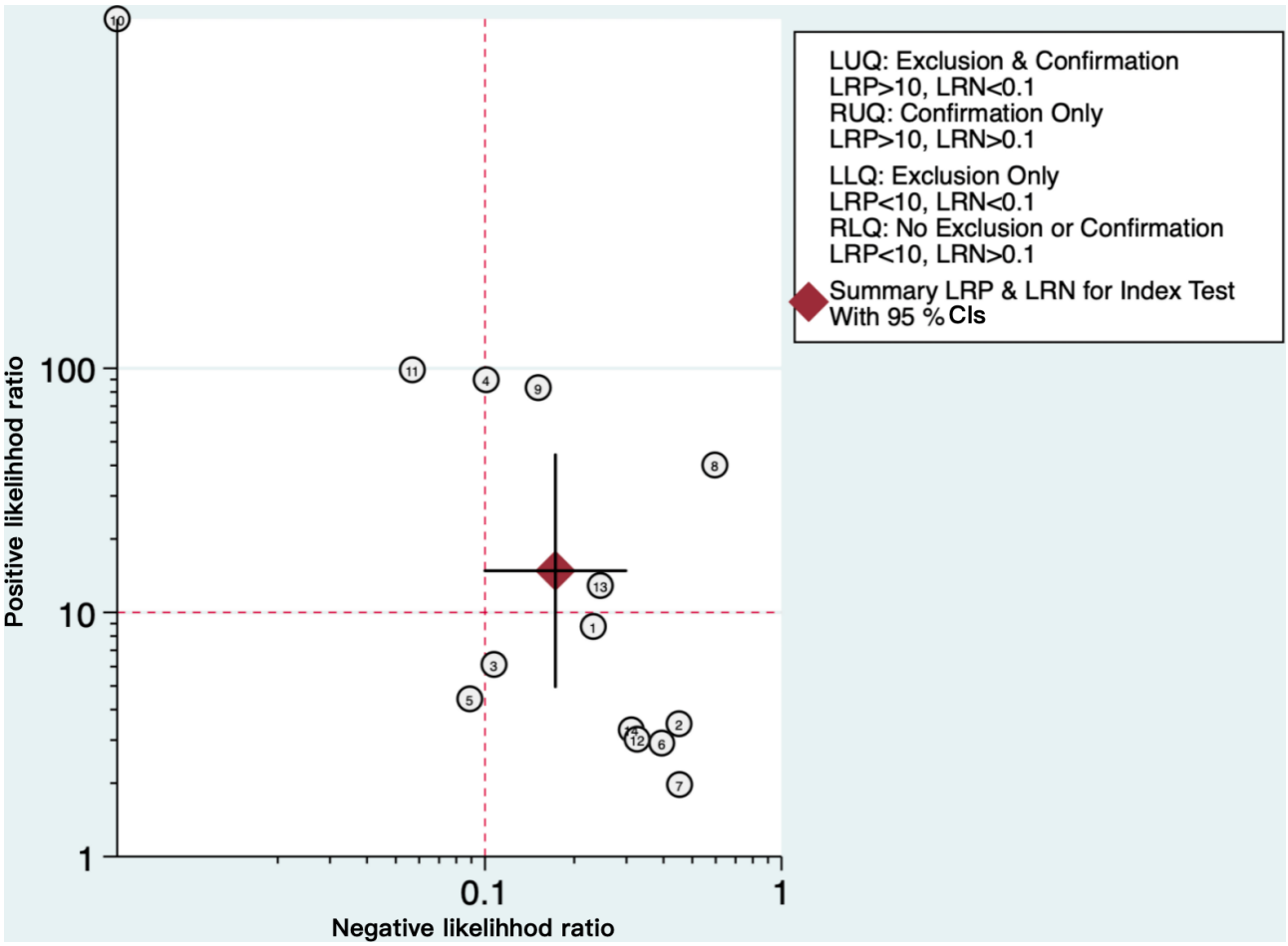

Predictors

From the models included in this systematic review, all reported predictors were systematically extracted and cataloged. The selection of key predictors for presentation and further analysis was based on three principal criteria: (1) clinical and pathophysiological relevance to GDM development, as established in prior literature and clinical guidelines; (2) frequency of reporting across the included studies, ensuring the findings were representative of common modeling practices; and (3) feasibility of meta-analytic synthesis, prioritizing variables with consistent definitions and measurements.

The consistently reported and clinically salient predictors were age, pre-pregnancy body mass index, first-trimester fast blood glucose, family history of diabetes, parity, gravidity, and history of GDM. These factors were well-recognized risk determinants in existing GDM etiological research and screening protocols. Detailed information is provided in Multimedia Appendix 4.

Assessment of Small-Study Effects

Deek test (P=.03) and the funnel plot (Figure 7) both showed clear asymmetry, suggesting the presence of small-study effects. This asymmetry might stem from publication bias, selective reporting, and methodological differences among smaller studies.

XSL•FO
RenderX

Figure 7. Deek funnel plot asymmetry test of small-study effects. The asymmetric distribution of studies suggests potential publication bias, where smaller studies reporting higher accuracy may be overrepresented in the literature.

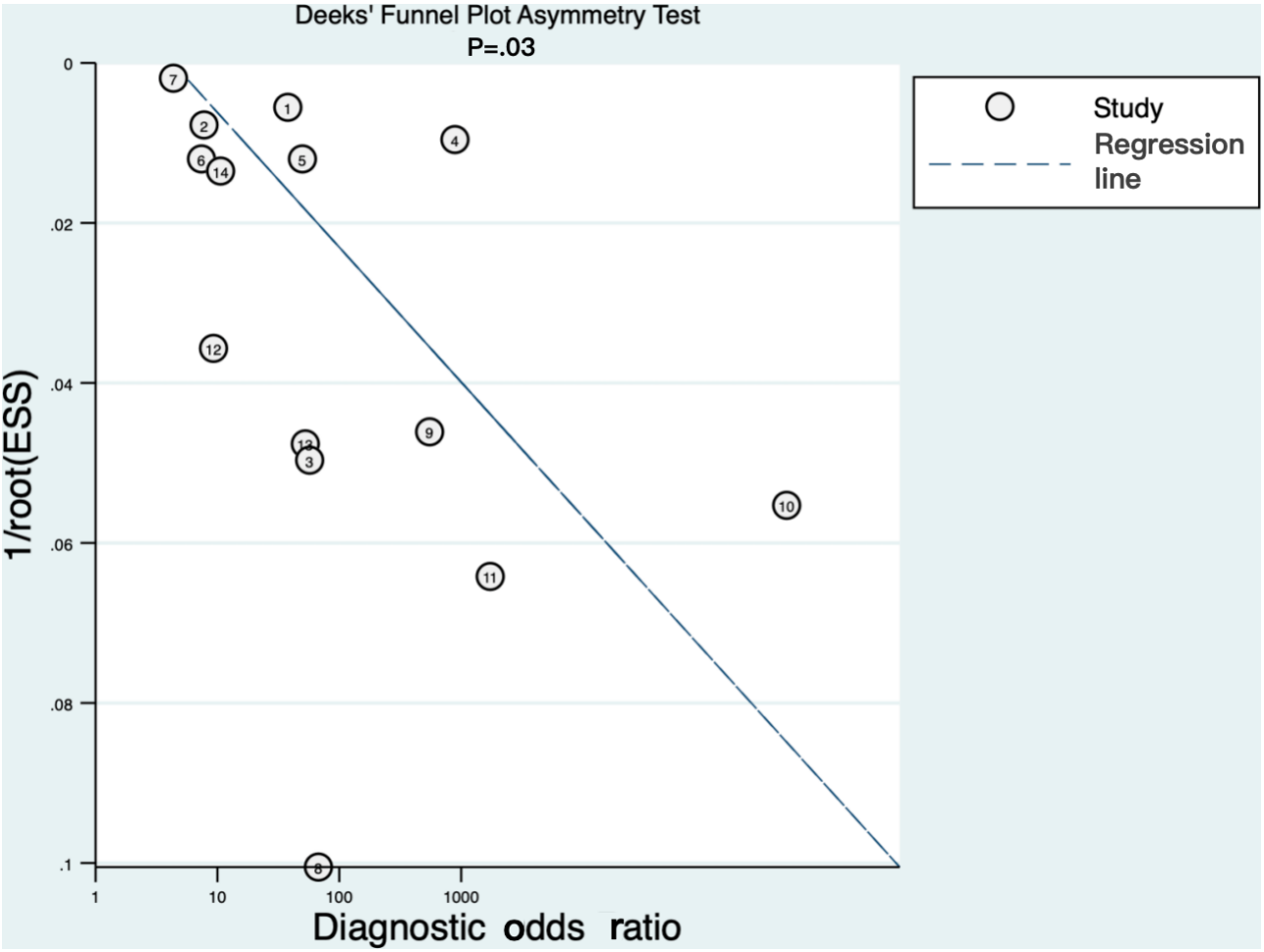

Heterogeneity Analysis

Threshold Effect Analysis

Bivariate boxplot (Figure 8) showed a positive correlation between the sensitivity and specificity of the included studies, indicating the absence of a threshold effect among the studies included in this systematic review. Moreover, some individual studies fell outside the shaded area, indicating the potential presence of heterogeneity.

XSL•FO
RenderX

Figure 8. Bivariate boxplot of threshold effect analysis. This plot evaluates whether heterogeneity in results can be explained by differences in diagnostic thresholds used across studies, with dispersed points indicating substantial variability beyond what threshold effects alone can explain.

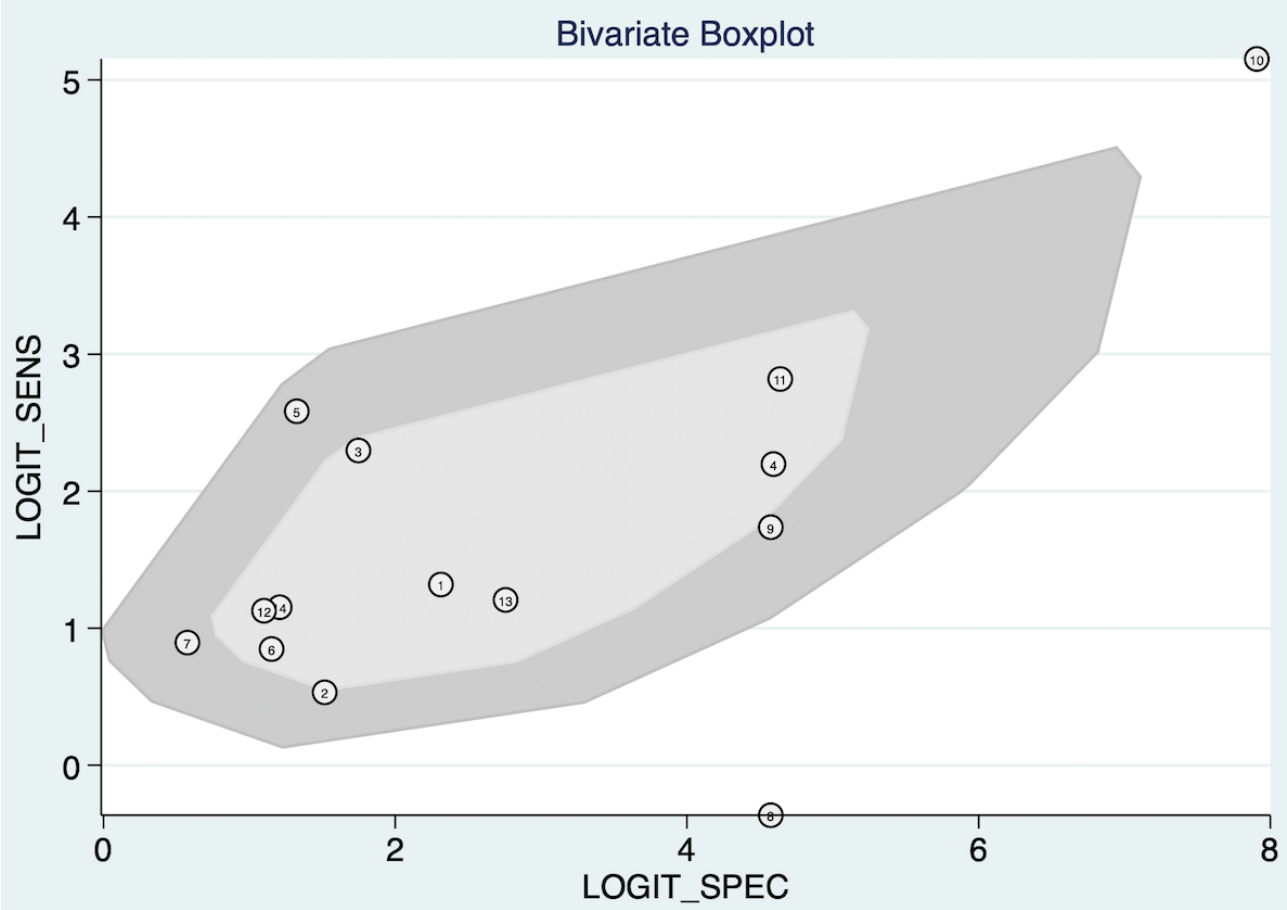

Subgroup Analysis

To evaluate the performance of prediction models constructed using various algorithms, subgroup analyses were performed on models that had been used in at least 3 studies, after first excluding 2 studies with extreme values caused by sparse data [40,42]. The performance of each algorithm was assessed using AUC, sensitivity, specificity, positive LR, negative LR, and diagnostic odds ratio (DOR). Details are presented in Table 3 and forest plots for sensitivity and specificity are shown in Figure 9. Among the subgroup models with sparse-data studies removed, the models using the RF algorithm exhibited the highest AUC, followed by those using the XGBoost algorithm, while the models using the logistic regression algorithm demonstrated the lowest AUC performance. Additionally, these models demonstrated varying performance across different metrics. The RF algorithm exhibited the highest sensitivity (0.83, 95% CI 0.74 - 0.93), while the XGBoost algorithm demonstrated the highest specificity (0.82, 95% CI 0.77 - 0.87) and DOR (49, 95% CI 11 - 211).

XSL•FO
RenderX

Table . Subgroup analysis of predictive performance across different artificial intelligence algorithms.

Models	Logistic regression	Random forest	XGBoost[a]	SVM[b]	*P* value
Number	8	4	4	4	__[c]
AUC[d]	0.75	0.87	0.86	0.78	<.001
Sensitivity (95% CI)	0.67 (0.62 - 0.72)	0.83 (0.74 - 0.93)	0.82 (0.79 - 0.85)	0.61 (0.36 - 0.86)	<.001
Specificity (95% CI)	0.72 (0.66 - 0.79)	0.80 (0.75 - 0.85)	0.82 (0.77 - 0.87)	0.80 (0.61 - 0.99)	.03
Positive LR[e] (95% CI)	2.8 (1.7 - 4.7)	4.5 (3.5 - 5.7)	10.1 (2.9 - 35.3)	4.2 (1.9 - 9.2)	<.001
Negative LR (95% CI)	0.42 (0.31 - 0.55)	0.17 (0.09 - 0.31)	0.21 (0.16 - 0.27)	0.45 (0.34 - 0.60)	<.001
DOR[f] (95% CI)	7 (3-15)	26 (12-58)	49 (11-211)	9 (5-17)	<.001

[a]XGBoost: extreme gradient boosting.
[b]SVM: support vector machine.
[c]Not applicable.
[d]AUC: area under the curve.
[e]LR: likelihood ratio.
[f]DOR: diagnostic odds ratio.

XSL•FO
RenderX

Figure 9. Forest plots of sensitivity and specificity in subgroup analysis. This forest plot presents pooled sensitivity and specificity estimates stratified by algorithm type (logistic regression, random forest, XGBoost, and SVM), allowing visual assessment of performance variability across model subgroups. The width of CI values reflects the precision of each estimate, while consistent point estimates across studies within a subgroup indicate algorithm-specific stability in diagnostic performance [16,21,22,28,36,38,39,41-44]. DL: deep learning; SVM: support vector machine; XGBoost: extreme gradient boosting.

XSL•FO
RenderX

Meta-Regression Analysis

To further explore the potential sources of heterogeneity in the performance of prediction models, a meta-regression analysis was conducted by including the study design (whether the study was conducted in Asia), study type (whether it was prospective), study design (whether it was multicenter), sample size (whether it exceeded 1000), GDM diagnostic criteria (whether it was based on IADPSG), and the timing of model use (whether it was in first trimester). Through meta-regression, we identified sources of heterogeneity among studies and evaluated their impact on diagnostic outcomes. The results indicated that study type significantly influenced heterogeneity among studies, with a trend toward increased predictive accuracy in prospective study designs (regression coefficient=2.289; *P*=.001). And the sample size had a substantial impact on the heterogeneity across studies, with predictive accuracy declining as the sample size increased (regression coefficient=−2.535; *P*=.001; Figure 10). This might reflect overfitting in small single-center datasets and greater clinical heterogeneity in large multicenter cohorts. Moreover, given the disparities among regions, the study area also served as one of the potential sources of heterogeneity (regression coefficient=−2.139, *P*=.002). The detailed procedures of the meta-regression are provided in Multimedia Appendix 5.

Figure 10. Bubble plot of meta-regression examining the association between sample size and predictive accuracy. This bubble plot visualizes the relationship between study sample size (log-transformed) and predictive accuracy (log-transformed diagnostic odds ratio) across all included studies. Each circle represents an individual study. The fitted regression line demonstrates a significant negative association, indicating that larger sample sizes tend to be associated with lower diagnostic accuracy. The plot provides an intuitive graphical confirmation of the quantitative meta-regression results, highlighting sample size as an important source of heterogeneity in model performance. log(DOR) values were truncated at ±3 for extreme cases (fp=0 or fn=0).

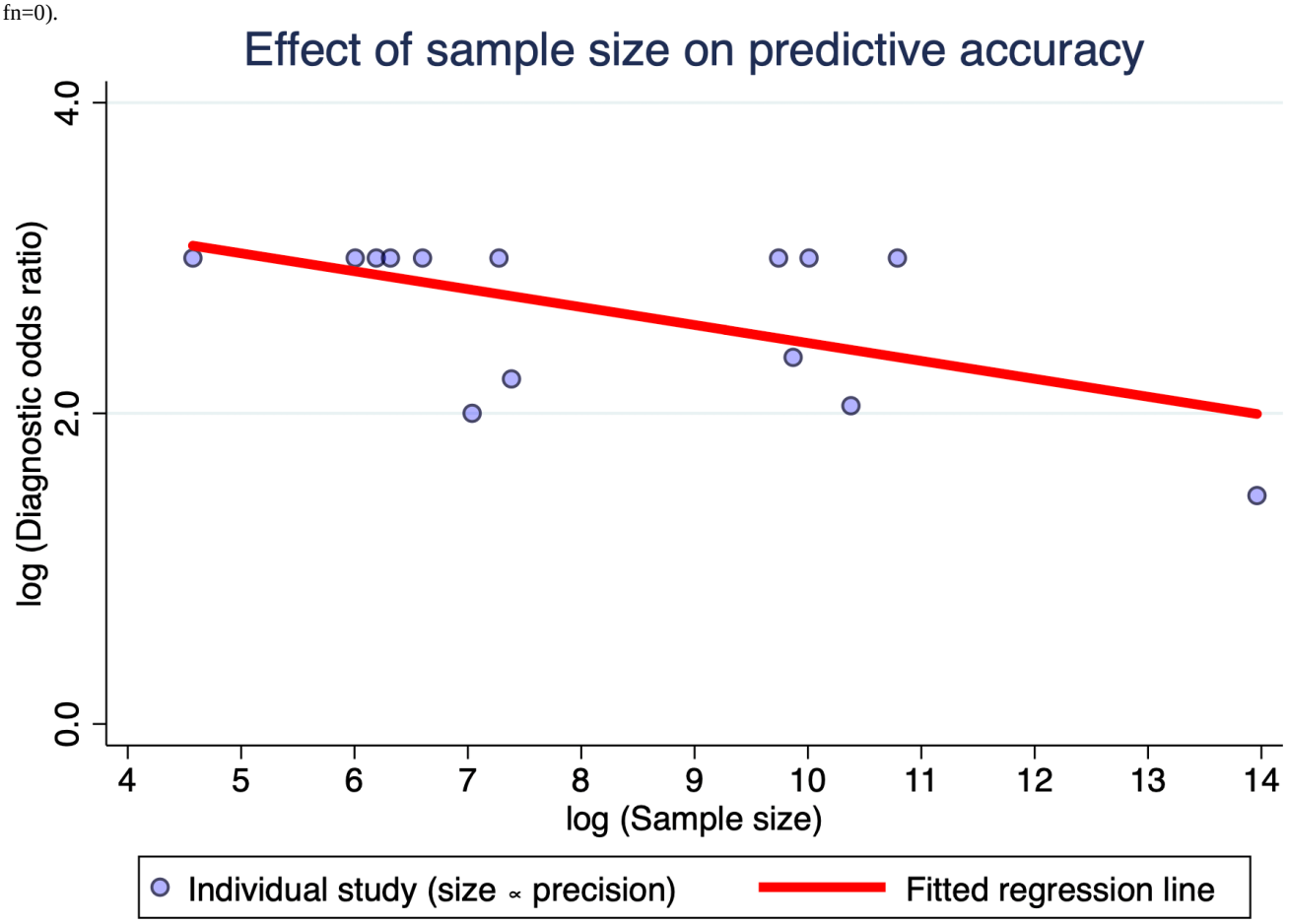

Discussion

Overview

This systematic review and meta-analysis aimed to evaluate the predictive performance of AI algorithms for GDM, compare the efficacy of different algorithms, and determine the key performance determinants. The pooled analysis revealed that AI-based models exhibited robust predictive capability for GDM prediction. However, the wide PIs revealed substantial performance heterogeneity in real-world applications, urging cautious interpretation of the currently summarized high-performance metrics. In addition, the summary LR plot and Fagan nomogram analyses indicated that existing models were insufficient to independently confirm or exclude GDM, so their present role should be positioned as an adjunct screening tool.

Consistent with the mainstream research trend, this systematic review further confirmed the dominant role of the RF algorithm in predicting GDM, which corroborated the findings of prior systematic reviews that highlighted ensemble methods for their robustness [14,45]. However, our analysis moved beyond merely confirming superiority by quantifying its extent and contrasting

XSL•FO
RenderX

it with other algorithms. Specifically, the RF algorithm performed the best in key metrics such as AUC and sensitivity, mainly because it handles the complex, non-linear relationships inherent in GDM prediction more effectively than linear models [46,47]. This was particularly relevant in clinical settings where data could be incomplete; the inherent ability of RF to handle missing values gracefully contributed to its stronger robustness when dealing with the imperfect data often presented in routine care, which was a critical practical advantage for implementation in real-world settings [47]. In contrast, the XGBoost algorithm demonstrated higher specificity, probably benefiting from its built-in regularization and feature-importance ranking, which made it more proficient at identifying true-negative cases [48,49]. It was worth noting that the 95% CI for the DOR of XGBoost was wide, reflecting marked between-study differences in sample size, event rate, or clinical heterogeneity and indicating that its actual diagnostic consistency was highly dependent on specific population characteristics and implementation settings. Notably, this systematic review identified and emphasized methodological heterogeneity as a key driver of performance disparities. Inconsistencies across studies in data preprocessing (eg, handling of missing values and feature scaling), validation strategies (eg, data split ratios and internal validation methods), and performance reporting standards significantly hindered the comparability and integrability of research outcomes. Therefore, while pursuing superior algorithms, future studies should prioritize the establishment and adherence to methodological reporting standards for the development and validation of AI-based prediction models.

To further elucidate the real-world implications of our findings, our meta-regression analysis identified several influential factors related to variations in model accuracy, providing a more nuanced understanding than simple performance pooling. Specifically, we found that prospective study design was associated with significantly higher predictive accuracy. This might be attributed to more standardized data collection procedures and better control of confounders in prospective settings, whereas retrospective studies often relied on preexisting eHealth record data, which could be heterogeneous and incomplete [50-52]. These findings aligned with the results reported by Liu et al [53], who reported that AI-based models in prospective cohort studies achieved AUC values 4% - 7% higher than those from retrospective studies. This consistency across different analyses strengthened the argument for prioritizing prospective validation designs. Additionally, we observed that studies with larger sample sizes tended to report lower accuracy estimates. This counterintuitive finding was crucial, as it likely reflected greater demographic and clinical diversity in larger cohorts, thereby reducing overfitting and offering a more realistic, generalizable performance assessment than optimistic estimates from small, homogeneous samples. This underscored that larger, more diverse studies provided a more trustworthy evidence base for clinical deployment. Similarly, studies conducted in certain geographic regions also showed systematically lower accuracy, possibly due to regional differences in diagnostic criteria, risk factor prevalence, or health care infrastructure. These findings indicated that the performance of a model depended not only on the algorithm itself but was also profoundly shaped by the environment in which it was developed and validated. This had direct implications for implementation: a model successful in one region might not translate directly to another without adaptation and local validation.

Despite the strong performance of some algorithms, AI models still faced critical barriers to clinical deployment that should be addressed to realize their potential [54]. These included the “black-box” nature leading to limited interpretability, a persistent lack of large-scale external validation in diverse populations, and the absence of standardized interfaces for integration with existing clinical workflows—especially eHealth record systems [55,56]. To overcome these barriers, future efforts should adopt a multifaceted implementation-science approach. This entails: (1) prioritizing prospective, multicenter validation studies to generate high-grade, generalizable evidence; and (2) incorporating explainable AI techniques to enhance model interpretability and foster clinician confidence. Ultimately, realizing the full potential of AI in GDM prediction requires a concerted shift from merely developing accurate algorithms to engineering clinically viable, trustworthy, and deployable solutions.

Limitations

However, several limitations exist in this systematic review and meta-analysis. First, most included studies and citations focused on East Asian populations, which might limit the generalizability of our findings to multi-ethnic or low-resource settings. External validation in diverse cohorts from Europe, North America, and Africa should therefore be needed to assess global applicability and to examine performance after feature-set simplification. Second, owing to limited application frequency, several emerging algorithms such as artificial neural networks and DL were not included in the subgroup analysis. Future studies should pay attention to the development of these emerging algorithms, verify their performance through more empirical studies, and explore their unique value in GDM prediction. Third, the Deek funnel plot asymmetry test indicated potential publication bias, suggesting that studies reporting higher performance metrics might be overrepresented. This could inflate the pooled estimates and limit generalizability. Future studies should consider preregistering protocols and sharing analysis code and datasets to improve reproducibility and reduce selective reporting.

Conclusions

This systematic review and meta-analysis confirmed the strong discriminative performance of AI models for GDM prediction. However, substantial heterogeneity, publication bias, and small-study effects currently limited their readiness for direct clinical deployment. Unlike previous narrative reviews, this study innovatively provided the first direct comparative and quantitative synthesis of multiple AI algorithms in this field. This approach filled a critical gap in existing literature by offering an evidence-based framework to guide algorithm selection, rather than merely summarizing performance metrics. The key implication for real-world application was the demonstrated need for local validation in target populations before implementation. To translate this potential into practice,

XSL•FO
RenderX

future studies must prioritize prospective, multicenter, large-scale external validations. The ultimate goal was to develop AI tools that were not only accurate but also interpretable and seamlessly integrable into clinical workflows, thereby enabling reliable AI-driven early prediction and management of GDM.

Acknowledgments

This manuscript did not use generative artificial intelligence for content creation, data analysis, or study design.

Funding

This study was supported by the National Natural Science Foundation of China (32070189), Postgraduate Scientific Research Innovation Project of Hunan Province (CX20251466), and Hunan Province Health Commission Scientific Research Project (D202314038701). The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the article.

Data Availability

All data analyzed in this study are included in this published article and its supplementary files.

Authors' Contributions

Conceptualization: YL (lead), YS (equal), and ZL (equal)
Data curation: AD (lead) and ML (supporting)
Formal analysis: YL (lead), AD (supporting), and ML (supporting)
Funding acquisition: YL, ZL, and ML
Investigation: ML (lead) and ZZ (supporting)
Methodology: YL (lead), AD (supporting), and ML (supporting)
Project administration: YL (lead) and JS (supporting)
Software: YL
Supervision: ZL (lead) and YS (supporting)
Validation: SJ (equal) and ZZ (equal)
Visualization: YL
Writing – original draft: YL (lead), AD (equal), and ML (equal)
Writing – review and editing: ZL (lead) and YS (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details of search strategy and result.
[DOCX File, 22 KB - jmir_v28i1e79729_app1.docx]

Multimedia Appendix 2
Details of model performance parameters for each study.
[DOCX File, 26 KB - jmir_v28i1e79729_app2.docx]

Multimedia Appendix 3
Result of Prediction Model Risk of Bias Assessment Tool assessment.
[DOCX File, 417 KB - jmir_v28i1e79729_app3.docx]

Multimedia Appendix 4
Predictive factors.
[DOCX File, 16 KB - jmir_v28i1e79729_app4.docx]

Multimedia Appendix 5
Result of meta-regression.
[DOCX File, 16 KB - jmir_v28i1e79729_app5.docx]

XSL•FO
RenderX

Checklist 1
PRISMA checklists.
[DOCX File, 61 KB - jmir_v28i1e79729_app6.docx]

References

1. Sweeting A, Wong J, Murphy HR, Ross GP. A clinical update on gestational diabetes mellitus. Endocr Rev 2022 Sep 26;43(5):763-793. [doi: 10.1210/endrev/bnac003] [Medline: 35041752]
2. Yang W, Liu J, Li J, et al. Interactive effects of prepregnancy overweight and gestational diabetes on macrosomia and large for gestational age: a population-based prospective cohort in Tianjin, China. Diabetes Res Clin Pract 2019 Aug;154:82-89. [doi: 10.1016/j.diabres.2019.06.014] [Medline: 31271809]
3. Gao C, Sun X, Lu L, Liu F, Yuan J. Prevalence of gestational diabetes mellitus in mainland China: a systematic review and meta-analysis. J Diabetes Investig 2019 Jan;10(1):154-162. [doi: 10.1111/jdi.12854] [Medline: 29683557]
4. Duong TL, Shahunja KM, Le M, McIntyre DH, Ward J, Mamun AA. Gestational diabetes mellitus and its impact on maternal and neonatal outcomes in Indigenous populations: a systematic review and meta-analysis. Diabetes Res Clin Pract 2025 Nov;229:112462. [doi: 10.1016/j.diabres.2025.112462] [Medline: 40947022]
5. Yang F, Liu H, Ding C. Gestational diabetes mellitus and risk of neonatal respiratory distress syndrome: a systematic review and meta-analysis. Diabetol Metab Syndr 2024 Dec 5;16(1):294. [doi: 10.1186/s13098-024-01539-x] [Medline: 39639383]
6. Zhang Y, Chen L, Ouyang Y, et al. A new classification method for gestational diabetes mellitus: a study on the relationship between abnormal blood glucose values at different time points in oral glucose tolerance test and adverse maternal and neonatal outcomes in pregnant women with gestational diabetes mellitus. AJOG Glob Rep 2024 Nov;4(4):100390. [doi: 10.1016/j.xagr.2024.100390] [Medline: 39309607]
7. Rizzo HE, Escaname EN, Alana NB, et al. Maternal diabetes and obesity influence the fetal epigenome in a largely Hispanic population. Clin Epigenetics 2020 Feb 19;12(1):34. [doi: 10.1186/s13148-020-0824-9] [Medline: 32075680]
8. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J 2021 Jul;8(2):e188-e194. [doi: 10.7861/fhj.2021-0095] [Medline: 34286183]
9. Helm JM, Swiergosz AM, Haeberle HS, et al. Machine learning and artificial intelligence: definitions, applications, and future directions. Curr Rev Musculoskelet Med 2020 Feb;13(1):69-76. [doi: 10.1007/s12178-020-09600-8] [Medline: 31983042]
10. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol 2020 Feb 27;9(2):14. [doi: 10.1167/tvst.9.2.14] [Medline: 32704420]
11. Segar MW, Jaeger BC, Patel KV, et al. Development and validation of machine learning-based aace-specific Mmodels to predict 10-year risk of heart failure: a multicohort analysis. Circulation 2021 Jun 15;143(24):2370-2383. [doi: 10.1161/CIRCULATIONAHA.120.053134] [Medline: 33845593]
12. Luo D, Zhang J, Xie L, Liu C, Wang R. Early detection of positive urine culture in patients with urolithiasis: a machine learning model with dynamic online nomogram. Ann Med 2025 Dec;57(1):2550582. [doi: 10.1080/07853890.2025.2550582] [Medline: 40853723]
13. Ma L, Yang L, Wang Y, et al. Enhancing early gestational diabetes mellitus prediction with imputation-based machine learning framework: a comparative study on real-world clinical records. Digit Health 2025;11:20552076251352436. [doi: 10.1177/20552076251352436] [Medline: 40755962]
14. Bigdeli SK, Ghazisaedi M, Ayyoubzadeh SM, Hantoushzadeh S, Ahmadi M. Predicting gestational diabetes mellitus in the first trimester using machine learning algorithms: a cross-sectional study at a hospital fertility health center in Iran. BMC Med Inform Decis Mak 2025 Jan 3;25(1):3. [doi: 10.1186/s12911-024-02799-3] [Medline: 39754258]
15. Kurt B, Gürlek B, Keskin S, et al. Prediction of gestational diabetes using deep learning and Bayesian optimization and traditional machine learning techniques. Med Biol Eng Comput 2023 Jul;61(7):1649-1660. [doi: 10.1007/s11517-023-02800-7] [Medline: 36848010]
16. Wu YT, Zhang CJ, Mol BW, et al. Early prediction of gestational diabetes mellitus in the Chinese population via advanced machine learning. J Clin Endocrinol Metab 2021 Mar 8;106(3):e1191-e1205. [doi: 10.1210/clinem/dgaa899] [Medline: 33351102]
17. AlSaad R, Elhenidy A, Tabassum A, et al. Artificial intelligence in gestational diabetes care: a systematic review. J Diabetes Sci Technol 2025 Aug 25;25:19322968251355967. [doi: 10.1177/19322968251355967] [Medline: 40855734]
18. Kokori E, Olatunji G, Aderinto N, et al. The role of machine learning algorithms in detection of gestational diabetes; a narrative review of current evidence. Clin Diabetes Endocrinol 2024 Jun 25;10(1):18. [doi: 10.1186/s40842-024-00176-7] [Medline: 38915129]
19. Snell KI, Ensor J, Debray TP, Moons KG, Riley RD. Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures? Stat Methods Med Res 2018 Nov;27(11):3505-3522. [doi: 10.1177/0962280217705678] [Medline: 28480827]
20. Hah H, Goldin DS. How clinicians perceive artificial intelligence-assisted technologies in diagnostic decision making: mixed methods approach. J Med Internet Res 2021 Dec 16;23(12):e33540. [doi: 10.2196/33540] [Medline: 34924356]

XSL•FO
RenderX

21. Cubillos G, Monckeberg M, Plaza A, et al. Development of machine learning models to predict gestational diabetes risk in the first half of pregnancy. BMC Pregnancy Childbirth 2023 Jun 23;23(1):469. [doi: 10.1186/s12884-023-05766-4] [Medline: 37353749]
22. Belsti Y, Moran L, Du L, et al. Comparison of machine learning and conventional logistic regression-based prediction models for gestational diabetes in an ethnically diverse population; the Monash GDM Machine learning model. Int J Med Inform 2023 Nov;179:105228. [doi: 10.1016/j.ijmedinf.2023.105228] [Medline: 37774429]
23. Salameh JP, Bossuyt PM, McGrath TA, et al. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. BMJ 2020 Aug 14;370:m2632. [doi: 10.1136/bmj.m2632] [Medline: 32816740]
24. Rethlefsen ML, Page MJ. PRISMA 2020 and PRISMA-S: common questions on tracking records and the flow diagram. J Med Libr Assoc 2022 Apr 1;110(2):253-257. [doi: 10.5195/jmla.2022.1449] [Medline: 35440907]
25. Ivaldi D, Burgos M, Oltra G, Liquitay CE, Garegnani L. Adherence to PRISMA 2020 statement assessed through the expanded checklist in systematic reviews of interventions: a meta-epidemiological study. Cochrane Evid Synth Methods 2024 May;2(5):e12074. [doi: 10.1002/cesm.12074] [Medline: 40476264]
26. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019 Jan 1;170(1):W1-W33. [doi: 10.7326/M18-1377] [Medline: 30596876]
27. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023 Dec;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
28. Wu Y, Ma S, Wang Y, et al. A risk prediction model of gestational diabetes mellitus before 16 gestational weeks in Chinese pregnant women. Diabetes Res Clin Pract 2021 Sep;179:109001. [doi: 10.1016/j.diabres.2021.109001] [Medline: 34390760]
29. Kang BS, Lee SU, Hong S, et al. Prediction of gestational diabetes mellitus in Asian women using machine learning algorithms. Sci Rep 2023 Aug 16;13(1):13356. [doi: 10.1038/s41598-023-39680-8] [Medline: 37587201]
30. Wang N, Guo H, Jing Y, et al. Development and validation of risk prediction models for gestational diabetes mellitus using four different methods. Metabolites 2022 Oct 29;12(11):1040. [doi: 10.3390/metabo12111040] [Medline: 36355123]
31. Syngelaki A, Wright A, Gomez Fernandez C, Mitsigiorgi R, Nicolaides KH. First-trimester prediction of gestational diabetes mellitus based on maternal risk factors. BJOG 2025 Jun;132(7):972-982. [doi: 10.1111/1471-0528.18110] [Medline: 40000426]
32. Zhao M, Su X, Huang L. Early gestational diabetes mellitus risk predictor using neural network with NearMiss. Gynecol Endocrinol 2025 Dec;41(1):2470317. [doi: 10.1080/09513590.2025.2470317] [Medline: 39992231]
33. Ali N, Khan W, Ahmad A, Masud MM, Adam H, Ahmed LA. Predictive modeling for the diagnosis of gestational diabetes mellitus using epidemiological data in the United Arab Emirates. Information 2022;13(10):485. [doi: 10.3390/info13100485]
34. Liu R, Zhan Y, Liu X, et al. Stacking ensemble method for gestational diabetes mellitus prediction in Chinese pregnant women: a prospective cohort study. J Healthc Eng 2022;2022:8948082. [doi: 10.1155/2022/8948082] [Medline: 36147870]
35. Kumar M, Ang LT, Png H, et al. Automated machine learning (AutoML)-derived preconception predictive risk model to guide early intervention for gestational diabetes mellitus. Int J Environ Res Public Health 2022 Jun 1;19(11):6792. [doi: 10.3390/ijerph19116792] [Medline: 35682375]
36. Lin Q, Fang ZJ. Establishment and evaluation of a risk prediction model for gestational diabetes mellitus. World J Diabetes 2023 Oct 15;14(10):1541-1550. [doi: 10.4239/wjd.v14.i10.1541] [Medline: 37970129]
37. Ye Y, Xiong Y, Zhou Q, Wu J, Li X, Xiao X. Comparison of machine learning methods and conventional logistic regressions for predicting gestational diabetes using routine clinical data: a retrospective cohort study. J Diabetes Res 2020;2020:4168340. [doi: 10.1155/2020/4168340] [Medline: 32626780]
38. Wang J, Lv B, Chen X, et al. An early model to predict the risk of gestational diabetes mellitus in the absence of blood examination indexes: application in primary health care centres. BMC Pregnancy Childbirth 2021 Dec 8;21(1):814. [doi: 10.1186/s12884-021-04295-2] [Medline: 34879850]
39. Donovan BM, Breheny PJ, Robinson JG, et al. Development and validation of a clinical model for preconception and early pregnancy risk prediction of gestational diabetes mellitus in nulliparous women. PLoS ONE 2019;14(4):e0215173. [doi: 10.1371/journal.pone.0215173] [Medline: 30978258]
40. Kaya Y, Bütün Z, Çelik Ö, Salik EA, Tahta T, Yavuz AA. The early prediction of gestational diabetes mellitus by machine learning models. BMC Pregnancy Childbirth 2024 Aug 31;24(1):574. [doi: 10.1186/s12884-024-06783-7] [Medline: 39217284]
41. Hu X, Hu X, Yu Y, Wang J. Prediction model for gestational diabetes mellitus using the XG Boost machine learning algorithm. Front Endocrinol (Lausanne) 2023;14:1105062. [doi: 10.3389/fendo.2023.1105062] [Medline: 36967760]
42. Lee SM, Hwangbo S, Norwitz ER, et al. Nonalcoholic fatty liver disease and early prediction of gestational diabetes mellitus using machine learning methods. Clin Mol Hepatol 2022 Jan;28(1):105-116. [doi: 10.3350/cmh.2021.0174] [Medline: 34649307]
43. Ding T, Liu P, Jia J, Wu H, Zhu J, Yang K. Application of machine learning algorithm incorporating dietary intake in prediction of gestational diabetes mellitus. Endocr Connect 2024 Dec 1;13(12):e240169. [doi: 10.1530/EC-24-0169] [Medline: 39393404]

XSL•FO
RenderX

44. Liu H, Li J, Leng J, et al. Machine learning risk score for prediction of gestational diabetes in early pregnancy in Tianjin, China. Diabetes Metab Res Rev 2021 Jul;37(5):e3397. [doi: 10.1002/dmrr.3397] [Medline: 32845061]
45. Kolozali S, White SL, Norris S, Fasli M, van Heerden A. Explainable early prediction of gestational diabetes biomarkers by combining medical background and wearable devices: a pilot study with a cohort group in South Africa. IEEE J Biomed Health Inform 2024 Apr;28(4):1860-1871. [doi: 10.1109/JBHI.2024.3361505] [Medline: 38345955]
46. Speiser JL. A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data. J Biomed Inform 2021 May;117:103763. [doi: 10.1016/j.jbi.2021.103763] [Medline: 33781921]
47. Uddin S, Lu H. Confirming the statistically significant superiority of tree-based machine learning algorithms over their counterparts for tabular data. PLoS ONE 2024;19(4):e0301541. [doi: 10.1371/journal.pone.0301541] [Medline: 38635591]
48. Wang X, Ren H, Ren J, et al. Machine learning-enabled risk prediction of chronic obstructive pulmonary disease with unbalanced data. Comput Methods Programs Biomed 2023 Mar;230:107340. [doi: 10.1016/j.cmpb.2023.107340] [Medline: 36640604]
49. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot 2013;7(21):21. [doi: 10.3389/fnbot.2013.00021] [Medline: 24409142]
50. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. Annu Rev Public Health 2016;37:61-81. [doi: 10.1146/annurev-publhealth-032315-021353] [Medline: 26667605]
51. Behman R, Bubis L, Karanicolas P. Prospective and retrospective cohort studies. In: Evidence-Based Surgery: A Guide to Understanding and Interpreting the Surgical Literature: Springer International Publishing; 2019:159-170. [doi: 10.1007/978-3-030-05120-4_16]
52. Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet 2002 Jan 19;359(9302):248-252. [doi: 10.1016/S0140-6736(02)07451-2] [Medline: 11812579]
53. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019 Oct;1(6):e271-e297. [doi: 10.1016/S2589-7500(19)30123-2] [Medline: 33323251]
54. Ioannidis JPA, Greenland S, Hlatky MA, et al. Increasing value and reducing waste in research design, conduct, and analysis. Lancet 2014 Jan 11;383(9912):166-175. [doi: 10.1016/S0140-6736(13)62227-8] [Medline: 24411645]
55. Shim JM, Shin E, Johnson TP. Self-rated health assessed by web versus mail modes in a mixed mode survey: the digital divide effect and the genuine survey mode effect. Med Care 2013 Sep;51(9):774-781. [doi: 10.1097/MLR.0b013e31829a4f92] [Medline: 23774510]
56. Lee KJ, Tilling KM, Cornish RP, et al. Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. J Clin Epidemiol 2021 Jun;134:79-88. [doi: 10.1016/j.jclinepi.2021.01.008] [Medline: 33539930]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
DL: deep learning
DOR: diagnostic odds ratio
GDM: gestational diabetes mellitus
LR: likelihood ratio
ML: machine learning
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis
PROBAST: Prediction Model Risk of Bias Assessment Tool
RF: random forest
ROB: risk of bias
SROC: summary receiver operating characteristic curve
XGBoost: extreme gradient boosting
τ: standard deviation
τ^2: standard variance

XSL•FO
RenderX

Edited by S Brini; submitted 26.Jun.2025; peer-reviewed by F Khosravi, X Guo; accepted 05.Jan.2026; published 30.Jan.2026.

Please cite as:
Liang Y, Dai A, Luo M, Zheng Z, Shen J, Su Y, Li Z
Predictive Performance of Artificial Intelligence Algorithms for Gestational Diabetes Mellitus in Pregnant Women: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e79729
URL: https://www.jmir.org/2026/1/e79729
doi:10.2196/79729

© Yingni Liang, Anran Dai, Meiyan Luo, Zhuolian Zheng, Jiayu Shen, Yinhua Su, Zhongyu Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Barriers to Digital Health Adoption in Older Adults: Scoping Review Informed by Innovation Resistance Theory

Yosefa Birati[1,2], RN, PhD; Roy Tzemah-Shahar[1], PhD

[1]The Cheryl Spencer Department of Nursing, Faculty of Social Welfare and Health Sciences, University of Haifa, Haifa, Israel

[2]The Center of Research and Study of Aging, University of Haifa, Haifa, Israel

Corresponding Author:
Yosefa Birati, RN, PhD
The Cheryl Spencer Department of Nursing, Faculty of Social Welfare and Health Sciences, University of Haifa, Haifa, Israel

Abstract

Background: The transformation of digital health technologies has reshaped health care delivery in primary care. Despite these benefits, older adults remain among the most resistant users. Traditional technology adoption models may not fully capture this reluctance, which is shaped not only by usability challenges but also by emotional, psychological, and identity-related concerns. Innovation resistance theory (IRT) offers a complementary framework focused on barriers to adoption rather than solely on facilitators.

Objective: This study aims to map and synthesize evidence on older adults' resistance to digital health in primary care through the lens of IRT, and to examine how resistance factors align with, extend, or refine IRT's functional and psychological barriers.

Methods: A scoping review with concept-driven thematic synthesis was conducted. A search for studies published between 2014 and 2025 was conducted across 5 databases: PubMed, CINAHL, Ovid Medline, Web of Science, and Scopus; the final search was completed in November 2025. Eligible studies were those that examined barriers or resistance to digital health use among adults aged 60 years and older in primary care settings. Search terms included "older adults," "digital health/eHealth," and "technology resistance." We excluded studies outside primary care and in which caregivers or health care professionals were the primary users. Data were extracted into a structured matrix and coded to the IRT domains: usage, value, risk, tradition, and image barriers. Relational integration was used to examine co-occurrence and linkages among barriers to inform the conceptual model.

Results: Seventeen studies were included, comprising 6822 participants (sample sizes ranged from 11 to 4525). Most studies were conducted in high-income Western countries, predominantly with qualitative designs, alongside mixed-methods and cross-sectional surveys. Functional barriers included usability challenges, interface complexity, and age-related impairments. Psychological resistance was linked to emotional discomfort, symbolic misalignment, and concerns about the loss of relational care. Value and risk concerns included distrust in diagnostic accuracy, privacy and data security, and skepticism about care quality. Traditional preferences for face-to-face interactions and generational digital divides reinforced image-based resistance. Interactions between barriers were identified, with low self-efficacy and technology anxiety creating feedback loops that reinforce avoidance behaviors.

Conclusions: Older adults' resistance to digital health is not simply a lack of adoption but a complex, emotionally grounded process involving functional, psychological, and identity-based barriers. This review applies IRT to primary care digital health, shifting the focus from adoption facilitators to resistance mechanisms and integrating co-occurrence patterns into a conceptual model. The synthesis reveals interacting factors of usability, self-efficacy, anxiety, trust, and legitimacy concerns that reinforce avoidance, suggesting that implementation strategies should extend beyond technical usability to rebuild trust, preserve relational care, and align digital solutions with older adults' values. Review limitations include the predominance of Western-based studies and limited longitudinal data on how resistance evolves.

(J Med Internet Res 2026;28:e75591) doi:10.2196/75591

KEYWORDS

older adults; digital health; telemedicine; technology resistance; technology adoption; primary health care; innovation resistance theory

XSL•FO
RenderX

Introduction

Background

The digital transformation of health care has been driven by the integration of telemedicine, mobile health (mHealth) applications, electronic health records, and wearable devices, which have significantly reshaped the delivery of medical services. These innovations address the limitations of traditional care models, which often struggle to meet the evolving demands of health care, particularly for aging populations in rural or underserved areas [1]. By improving access to care, supporting chronic disease management, and promoting preventive health care initiatives, digital health technologies offer promising solutions.

Notably, older adults, who often face mobility limitations, chronic illnesses, and restricted access to traditional health care services, are likely to gain substantial benefits from these digital health innovations [2,3]. However, despite the potential benefits, older adults remain among the most resistant groups to adopting these technologies [4]. This reluctance is widely documented in prior research and often attributed to multiple factors, including limited digital literacy, usability concerns, lower self-efficacy, privacy concerns, and a strong preference for in-person health care interactions. These barriers contribute to older adults' limited willingness to engage with digital health care solutions [5-7].

The persistence of this reluctance suggests that adoption-centric models may offer an incomplete explanation, highlighting the need for complementary resistance-focused frameworks. To better understand these patterns, we first reviewed established technology adoption models used in health care, clarifying their scope and limitations, and then introduced innovation resistance theory (IRT) as a complementary resistance-focused framework.

Existing Technology Adoption Models

Established theoretical models, such as the technology acceptance model (TAM) and the Unified Theory of Acceptance and Use of Technology (UTAUT), have been widely used to explain individuals' adoption and use of new technologies [8]. These models highlight factors such as perceived usefulness, ease of use, performance expectancy, and social influence as key determinants of technology adoption [9,10]. Complementary to these, Rogers' Diffusion of Innovations Theory describes how new technologies spread through populations by considering factors such as adopters' characteristics, communication channels, and social systems [11]. These frameworks have been extensively validated and remain central tools for understanding and measuring technology acceptability and usage intentions in health care.

In the context of older adults' digital health use, adoption-focused models provide valuable insights into factors associated with acceptance and initial uptake; however, prior literature suggests that older adults' persistent nonuse and resistance are also shaped by affective, psychological, and contextual factors that are not always represented as central constructs in these models [12]. For example, a scoping review by Wilson et al [13] applied UTAUT2 as an analytic framework to map barriers and facilitators to eHealth use among older adults. They identified gaps in the evidence base for certain UTAUT2 constructs (eg, habit and hedonic motivation) alongside recurring concerns related to privacy, trust, and support needs [13]. Another empirical study showed that older adults' intention to use mHealth was not explained solely by perceived ease of use and perceived usefulness, with person-related, technology-related, and contextual barriers influencing adoption [14]. Fox and Connolly further argue that research on older adults' resistance to mHealth remains limited and therefore examine how privacy concerns, trust, and risk beliefs influence willingness to adopt beyond standard adoption-model constructs [15]. Taken together, these findings suggest that complementing adoption-focused models with resistance-oriented frameworks may better capture why some older adults actively avoid digital health technologies, including perceived risks, emotional discomfort, and contextual constraints [12,16-18]. Accordingly, adoption-focused models may emphasize intention and perceived benefits, whereas nonadoption can also reflect an active decision-making process shaped by perceived risks and psychological discomfort. Therefore, we propose complementing adoption-focused theories with a resistance-oriented framework, such as IRT.

Innovation Resistance Theory (IRT) as a Conceptual Framework

IRT, introduced by Ram and Sheth [19], was developed to understand consumer resistance to marketing innovations and their behavior. Unlike models that emphasize adoption facilitators, IRT focuses on understanding why individuals hesitate or actively refuse to adopt new products, services, and ideas, even when they offer potential benefits [19]. The strength of IRT lies in its focus on perceived barriers rather than enablers, making it well-suited for populations such as older adults, where complex emotional, cognitive, and contextual factors influence nonuse. By focusing on the barriers, IRT offers a different perspective that shifts attention from the characteristics of innovations themselves to the reasons behind consumer reluctance to adopt them, especially when such adoption threatens established habits and routines or involves perceived risks [20-22]. In this view, resistance is not merely a lack of adoption but an active process that focuses on barriers to acceptance, including functional, psychological, and social resistance factors [19].

Resistance is defined as a multidimensional construct encompassing 3 dimensions: cognitive resistance, which involves individuals' appraisal of innovations and their perceived risks; affective resistance, which stems from emotional responses such as fear, frustration, or anxiety; and behavioral resistance, which manifests in actions ranging from passive disengagement to active opposition [23,24]. Within the IRT, these dimensions are further classified into functional and psychological barriers. Functional barriers include the usage barrier, which reflects the extent to which an innovation is perceived as requiring changes to established routines or habits; the value barrier, which arises when the individual perceives that the benefits of an innovation do not outweigh its costs; and the risk barrier, which represents concerns about the financial, functional, and social consequences of adopting an innovation.

XSL•FO
RenderX

Psychological barriers encompass traditional barriers, which refer to the degree to which an innovation forces an individual to accept changes that challenge cultural norms or long-standing behaviors, and image barriers, which relate to the degree to which an innovation is perceived as having an unfavorable image or negative associations [19,25]. These psychological categories often reflect deeper symbolic concerns, such as identity, generational belonging, or perceived legitimacy of digital care. This classification allows IRT to capture the multifaceted nature of resistance in older populations, particularly their emotional unease, normative preferences, and experiential distrust of digital systems. By categorizing resistance into functional and psychological barriers, IRT may provide a comprehensive framework for understanding why older adults struggle to adopt digital health solutions.

Over time, IRT has gained strong empirical support across different service and technology contexts. For example, in mobile banking research across Thailand and Taiwan, IRT barriers explained 60% - 66% of the variance in resistance intentions, with usage, value, risk, and image barriers showing statistically significant effects [21]. In a large Italian survey, Spinelli et al [26] showed that usage barriers and value-related concerns significantly reduced both actual mobile payment use and intention to adopt, whereas risk and image barriers had weaker or nonsignificant effects, and their impact varied across technology-readiness clusters [26]. Similarly, a study of Internet and mobile banking in Finland found that the value barrier was the dominant inhibitor of adoption and intention to adopt, while image and tradition barriers differentiated postponers from rejecters across seemingly similar service innovations [20].

Together, these findings demonstrate that IRT-based barriers have substantial explanatory power for resistance to digital innovations. Therefore, in this review, we apply IRT to structure the evidence on older adults' resistance toward digital health technologies and to examine whether the identified resistance factors map onto, extend, or refine the original IRT barrier categories. The aim of this scoping review was to synthesize and conceptualize evidence on older adults' resistance to digital health technologies in primary care using IRT. Specifically, we aimed to identify and categorize resistance factors into IRT functional and psychological barriers and to examine how these barriers co-occur and interact to inform a conceptual model of resistance. The review was guided by the following research questions: (1) What is known from the existing literature about older adults' resistance to using digital health technologies for monitoring and management in primary health care? (2) What are the functional (usage, value, risk) and psychological (tradition, image) IRT barriers reported across studies? (3) How do IRT barriers co-occur and link within and across studies?

Methods

Study Design

The methodology for this scoping review follows the framework proposed by Arksey and O'Malley [27], incorporating refinements by Levac et al [28], and the Joanna Briggs Institute (JBI) Reviewers' Manual [29]. We selected the scoping review approach to explore the current body of knowledge regarding older adults' resistance to digital health technologies through the lens of IRT, as it is well-suited to mapping the existing literature, identifying and interpreting patterns of functional and psychological resistance across heterogeneous study types. Within this design, our goal was to provide a theory-informed synthesis that evaluates how well IRT accounts for older adults' resistance to digital health in primary care and to identify conceptual and empirical gaps that warrant further investigation and measurement development. The reporting of this scoping review was guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) guidelines [30]. Reporting of the search methods followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension checklist (PRISMA-S) [31] to ensure transparent and complete reporting. The completed PRISMA-ScR checklist is provided in Checklist 2, and the PRISMA-S checklist is provided in Checklist 1.

Stage 1: Identifying the Research Question

The review was guided by predefined research questions (presented at the end of the Introduction section) informed by IRT and scoping review guidance.

Stage 2: Searching and Identifying Relevant Studies

A literature search was conducted across 5 major databases: PubMed, CINAHL, Ovid Medline, Web of Science, and Scopus, to identify peer-reviewed publications relevant to the research question. These databases were selected for their broad coverage of health, behavioral, and interdisciplinary studies on older adults and digital health. Each database was searched separately through its web interface, and all retrieved records were exported to Mendeley (version 1.63.0; Mendeley Reference Manager) for deduplication. Review studies were not included in this scoping review; however, their reference lists were screened to identify potentially eligible primary studies. No study registries were searched. Apart from reference-list screening, no additional sources were searched, and no citation searching was undertaken. We did not contact authors to identify additional studies, and no other search methods were used beyond those described. We did not use any previously validated search filters. Search strategies were developed specifically for this scoping review by the authors and were not peer reviewed by an independent expert before execution. We did not adapt or reuse search strategies from previous literature reviews for any substantive part of our search.

The search was carried out on December 20, 2024, and was rerun on November 18, 2025, to identify newly published studies since the initial search. The search followed the JBI PCC structure (Participants, Concept, Context) and combination of the following keywords and MeSH terms: "older adults," "elderly," phenomena of "digital health," "eHealth," "Telemedicine," and context of "primary health care," and "barriers to health technology." Boolean operators were used to combine search strings (eg, AND, OR). Title and abstract screening and full-text review were conducted by 2 independent reviewers (YB and RTS). The search strategy and keyword combination can be found in Multimedia Appendix 1. Additionally, reference lists of included studies were manually

XSL•FO
RenderX

screened to identify relevant studies not captured in the initial searches.

Stage 3: Selecting the Relevant Studies

Inclusion and Exclusion Criteria

The review included papers that met predefined inclusion and exclusion criteria aligned with the JBI PCC framework for scoping reviews (Table 1).

Table . Study eligibility criteria (Population-Concept-Context) for the scoping review.

Criteria	Inclusion	Exclusion
Participants/population	• Older adults aged 60 years and older	• Children, adolescents, and younger adults aged <60 years • Caregivers • Health care professionals
Concept (intervention)	• Use of mHealth[a] for monitoring and management • mHealth: telemedicine, mobile phone apps, smartphone apps, web-based systems • Resistance or barriers to the use of digital health technologies	• Use of mHealth telemonitoring for patients who are not adults and younger adults aged <60 years • Use of mHealth telemonitoring by caregivers or health care professionals
Context (cultural factors, geographic location, setting)	• Use of digital health technologies in primary health care	• Secondary/tertiary care: hospital inpatient wards, surgical centers • Emergency/urgent care
Type of studies	• Qualitative, quantitative, or mixed methods studies • Observational and experimental, cross-sectional, or longitudinal, randomized controlled trial or nonrandomized or noncontrolled trial, case series or case reports	• Conference abstracts, editorials, commentaries, letters to editor, essays, book chapters, and books
Language	• English	• Language other than English
Publication date	• From 2014	__[b]

[a]mHealth: mobile health.

[b]Not applicable.

The context of the review centered on the resistance to digital health within the framework of IRT. Publications addressing the use of digital health within the resistance domains of usage, value, risk, traditional, and image barriers were considered, while those focusing solely on the description of digital health adoption and facilitators were excluded. Also, no minimum sample size threshold was applied. Consistent with the objectives of a scoping review, studies were eligible regardless of their sample size to maximize coverage of designs (qualitative, quantitative, mixed methods) and contexts.

Study Selection Process

The studies were screened against the inclusion and exclusion criteria developed by the authors. The selection process followed three steps: (1) Title and abstract screening to remove irrelevant or duplicate records; (2) Full-text review based on predefined inclusion and exclusion criteria; and (3) Final inclusion based on relevance for examination of resistance to digital health technologies among older adults.

A total of 4976 records were identified through database searches, and 2387 duplicates were removed. After screening 2589 titles and abstracts, 227 full-text articles were reviewed. Two independent reviewers (YB and RTS) evaluated relevant publications for eligibility and selected qualifying publications based on the inclusion/exclusion criteria. We used a consensus-based approach, prioritizing unanimous agreement through re-evaluation of the eligibility criteria; if consensus could not be reached, a third reviewer would adjudicate. An initial pilot screening was conducted independently by both reviewers to ensure consistent interpretation of the eligibility criteria. Discrepancies identified at this stage were resolved through discussion and used to refine the criteria, resulting in full agreement during subsequent screening. A total of 17 studies met the inclusion criteria and were included in the final synthesis. A PRISMA flow diagram illustrates the selection process.

Stage 4: Charting the Data - Data Extraction and Synthesis

Two authors independently extracted data from all included studies. Data were charted using a standardized extraction form developed for this review, capturing study design, aims, population, type of digital health intervention, and resistance-related findings. Using a concept-driven thematic synthesis, findings were organized into 5 resistance categories from the IRT: usage, value, risk, tradition, and image barriers.

XSL•FO
RenderX

A structured matrix was used to map resistance dimensions across the studies. Barrier statements were first open-coded descriptively and then mapped to one IRT family using prespecified rules. Data charting was conducted by the 2 authors, and disagreements were resolved by consensus.

Stage 5: Collating, Summarizing, and Reporting the Results

Findings were organized in three layers: (1) mapping of the evidence base (study characteristics, settings, modalities), (2) concept-driven qualitative synthesis using IRT classification (usage, value, risk, tradition, and image), and (3) relational integration examined interconnections across IRT barriers. We extracted and coded barrier co-occurrences and linkages reported in the studies' results sections and participant quotations when two or more barriers were described as co-occurring or interacting. Links were considered explicit when directly stated, inferential when implied within a study's narrative context, and integrative when consistent patterns recurred across multiple studies (worked examples are provided in the Results).

Results

General Characteristics of the Studies

The database search initially identified 4976 records. After removing duplicates, screening titles and abstracts, and full papers, 17 studies were included in the final synthesis (Figure 1). The included papers represent a predominantly high-income Western countries from the United States (n=4), Sweden (n=3), the Netherlands (n=3), Canada (n=2), Finland (n=1), Norway (n=1), and the United Kingdom (n=1), with only 2 studies from non-Western settings Israel (n=1) and Indonesia (n=1). Eleven studies were qualitative, 3 studies were cross-sectional, and 4 studies were mixed methods designs. Sample sizes ranged from 11 to over 4500 participants, though qualitative samples were generally smaller and in-depth. In terms of digital health modalities, most studies focused on telemedicine or digital consultations (12/17) and patient portals or eHealth services (8/17), with comparatively few studies examining mobile apps or tablets (3/17) and wearables or remote monitoring (2/17) (Table 2).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the screening and selection process for the Scoping Review on resistance to digital health among older adults.

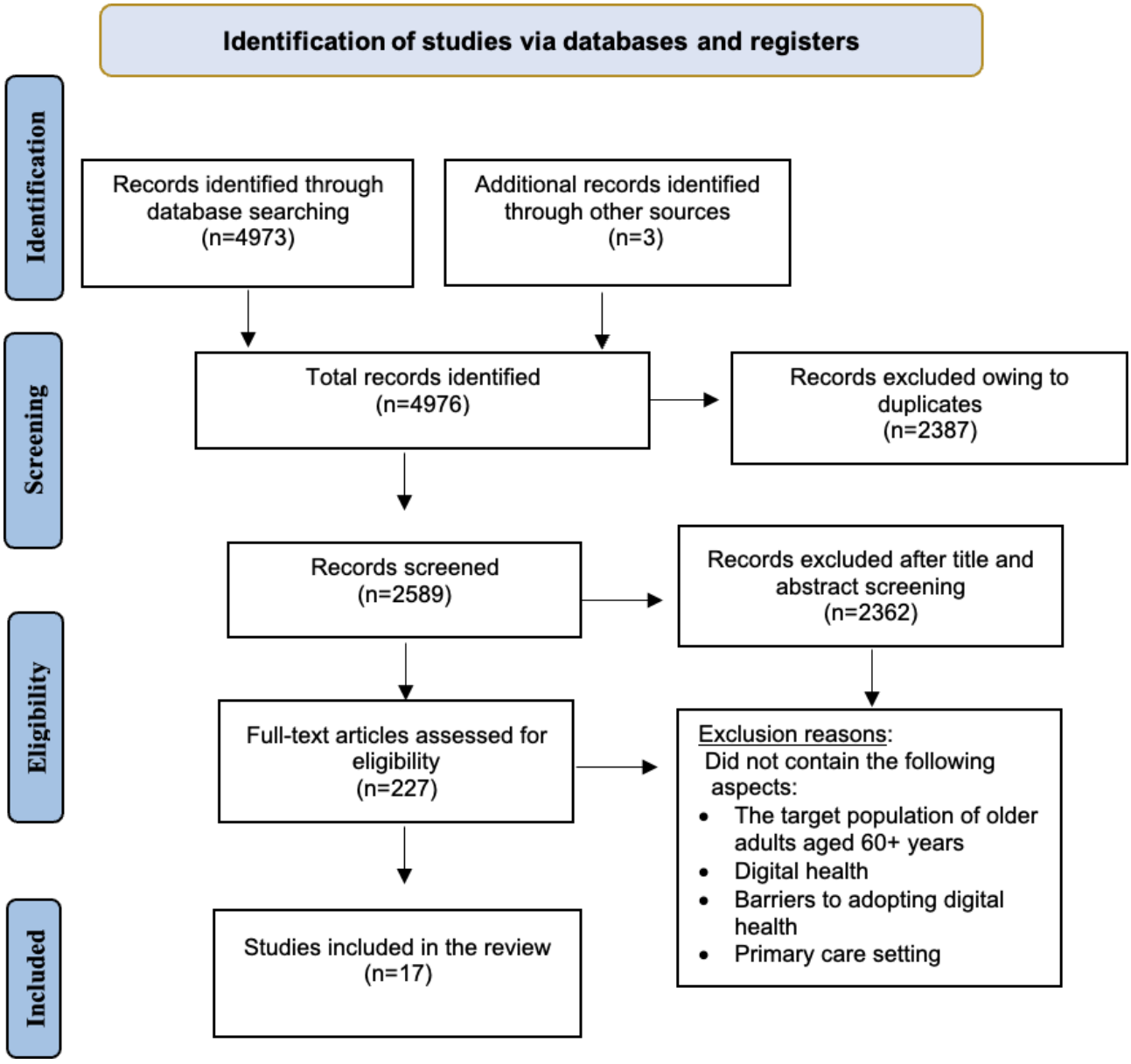

XSL•FO
RenderX

Table . Included studies on older adults' resistance to digital health technologies (n=17). This table presents the country, study design, population sample size, age range, and digital health modality.

	Study design	Aims	Study population	Digital health
Khanassov et al [32] (Canada)	Qualitative study: semistructured interviews and 3 focus groups to explore the experiences of both older adults and health care professionals	• How do older adults and health care professionals experience the use of telemedicine? • What are the facilitators and barriers to telemedicine use in the care of older adults? • What recommendations can enhance telemedicine engagement for older adults and health care professionals	• 29 older adults and health care professionals (family physicians, nurses, social workers, physiotherapists) • Age range 65 - 90 years	• Telemedicine in primary care
Vergouw et al [33] (Netherlands)	Qualitative study: semistructured and think-aloud interviews	• Identify the needs, barriers, and facilitators among community-dwelling older adults (60 years and older) with chronic health conditions in using web-based eHealth applications to support general practice services	• 19 community-dwelling older adults with at least one chronic condition • Mean age 73 (SD 5.3) years	eHealth applications for: • e-Consultation • Schedule e-Appointment • e-Prescription ordering • e-Lab results viewing • Access to e-File
Knotnerus et al [34](Netherlands)	Qualitative study: semistructured interviews thematic analysis	• Investigate the experiences of older patients (65 years and older) who use a digital health platform in general practice • Identify barriers and facilitators for using digital health • Examine whether a practice's focus on digital health influences older patients' choice to become a patient at the practice	• 18 older patients enrolled in 2 general practices • Age range 68 - 89 years	Digital health platform for: • Communicate with general practitioners • Appointment scheduling • Order repeat medications
Bhatia et al [35] (United States)	Cross-sectional multimethod study: mixed methods (Quantitative and Qualitative: close and open-ended questions)	• Understand older adults' experience with primary care telemedicine since the COVID-19 pandemic • Identify satisfaction levels and technical challenges in telemedicine use • Provide policy recommendations for the future of telemedicine services	• 208 older adults (≥65 y) who had a telemedicine visit within primary care visit • Mean age 74.4 (SD 4.4)	• Telemedicine (telephone and video visits)

XSL•FO
RenderX

	Study design	Aims	Study population	Digital health
Lam et al [36] (United States)	Cross-sectional study: data from the 2018 National Health and Aging Trends Study (NHATS)	• Assess the prevalence of telemedicine unreadiness and how older adults may be left behind in the United States when the migration to telemedicine occurred • Identify key barriers preventing the use of video-based telemedicine • Examine disparities in telemedicine access based on demographic, socioeconomic, and health-related factors	• 4525 community-dwelling older adults (≥65 y) in the United States • Mean age 79.6 (SD 6.9)	• Telemedicine (video and telephone visits)
Nymberg et al [37] (Sweden)	Qualitative research using focus group interviews thematic content analysis	• Explore older adults' beliefs, attitudes, experiences, and expectations regarding eHealth services in primary health care • Understand factors influencing adherence to eHealth tools in primary care among elderly patients • Identify barriers and facilitators affecting older adults' engagement with eHealth services	• 15 elderly patients from 3 primary health care centers in Southern Sweden, selected based on chronic disease status and medication use • Age range 65 - 80 years	eHealth services and use of the mobile phone for: • Contacting the health care system via web • Self-monitoring of chronic illnesses • Seeking medical information
van Houwelingen et al [38] (Netherlands)	A mixed method triangulation design, including a cross-sectional survey study (quantitative phase) and qualitative observations of older adults performing digital tasks in their daily lives	• Understand older adults' readiness for telehealth, particularly videoconferencing • Identify factors influencing their intention to use videoconferencing • Examine their capacities and barriers in using digital technology in daily life	• 256 participants in the survey and 15 older adults aged 65 years or older in the qualitative observations • Median (IQR) age=71 (67 - 76) years	• Telehealth, focused particularly on the use of videoconferencing for health care consultations
Laukka et al [39] (Finland)	Survey study with qualitative inductive content analysis of open-ended questions	• Investigate the preferences and needs of older adults regarding the use and development of digital health and social services • Understand how digital health and social services can be designed to more effectively meet the needs of older adults	• 1100 Finnish individuals aged 75 and older • Age range 75 - 99 years	• Telemedicine consultations • eHealth services
Rochmawati et al [40] (Indonesia)	Exploratory qualitative study using semistructured interviews, thematic analysis			• Digital health technologies (mobile apps, smartwatches) for health monitoring.

XSL•FO
RenderX

	Study design	Aims	Study population	Digital health
		• Explore the acceptance of eHealth technology among older adults in primary care • Examine perceptions, attitudes, experiences, and expectations of older people patients regarding eHealth services used in primary care	• 11 Older adults with chronic conditions (diabetes, hypertension) from a suburban primary health clinic in Indonesia • Mean age 66.9 years	
Fjellså et al [41] (Norway)	Explorative qualitative study using semistructured interviews	• To explore multimorbid older adults' experiences with participation and eHealth in care coordination with the support of general practitioners and district nurses	• 20 older adults with multimorbidity (COPD, heart failure, diabetes, and physical disabilities) receiving primary care services • Mean age 82 (range 71 - 98) years	• Patient portals to share and access information • Electronic messaging with general practitioners • Schedule appointments • Order prescriptions
Mao et al [42] (United States)	Mixed methods needs assessment (cross-sectional survey and qualitative interviews)	• Identify barriers to telemedicine video visits among older adults with differing socioeconomic backgrounds and primary spoken languages • Understand technological, cognitive, and language-related obstacles to telemedicine use • Provide recommendations to improve access and engagement with telemedicine	• 249 older adults from 2 independent living facilities • Mean age 84.6 (SD 6.6) years	• Telemedicine visits.
Frishammar et al [43] (Sweden)	Qualitative interviews and process data from a Swedish DHP provider	• To investigate adoption and usage barriers of digital health platforms among older adults • To understand how to facilitate increased adoption and usage of digital health platforms among the elderly	• 22 older adults aged ≥65 years, including both users and nonusers of digital health platforms, as well as individuals with experience in digital health development • Age range 65 - 80 years	• Video calls • Chats • Asynchronous messaging
Haimi et al [44] (Israel)	Qualitative study using semistructured interviews.	• To identify the challenges and barriers faced by the senior population when utilizing telemedicine services	• 14 elderly individuals from a primary health care clinic in Israel • Mean age=73 (range 66 - 85) years	• Telemedicine (phone and video visits) • electronic medical records prescription refills • Digital referrals • Electronic messages with the medical provider
Landgren and Cajander [45] (Sweden)	Qualitative, semistructured interviews.		• 13 participants aged >65 years • Mean age 74 years	• Digital health consultations delivered by video or chat/phone applications in primary care settings

XSL•FO
RenderX

	Study design	Aims	Study population	Digital health
		• To identify reasons for nonuse of digital health consultations among elderly in rural areas • To describe their attitudes toward technology, and possible challenges and opportunities.		
Ahmed et al [46] (United Kingdom)	Qualitative, focus group study.	• To explore the experiences, perceptions, and expectations of older adults from 3 minoritized ethnic group backgrounds regarding digitalized primary care services since the beginning of COVID-19.	• 27 participants age >65 years • Median (IQR) age=69 (66.5 - 72.5) years	• Telemedicine (phone and video visits) • Web-based services: View medical records Schedule appointments Order prescriptions
Ufholz et al [47] (United States)	Cross-sectional survey.	• To assess telemedicine preparedness of older primary care patients: internet use, device ownership, prior telemedicine experience, concerns, and perceived barriers	• 30 community-dwelling adults aged ≥65 • Age range 65 - 89 years	• Telemedicine for primary care (video/online visits)
Sproul et al [48] (Canada)	Cross-sectional survey	• To determine what technologies and apps are in current use by older adults, to explore the types of technologies and apps that may be of interest to people in this age group, to explore concerns about technologies, and to examine any age-related differences	• 266 participants aged ≥60 years • 60.2% participants were 60 - 74 years and 39.8% participants were 75 years or older	• Mobile phones • Tablets • Health-related apps

The IRT framework was used to guide the coding of extracted findings into the 5 barrier domains (usage, value, risk, tradition, and image).

The findings synthesis is presented in the following sections and summarized in Tables 3 and 4.

XSL•FO
RenderX

Table . Matrix mapping of innovation resistance theory (IRT) functional and psychological barrier domains (usage, value, risk, tradition, and image) across included studies of older adults' resistance to digital health in primary care (n=17).

	Functional barriers			Psychological barriers	
	Usage barriers	Value barriers	Risk barriers	Tradition barriers	Image barriers
Khanassov et al [32]	• Technical challenges • Symptom articulation • Technology usability	• Informality bias • Limited use perception	• Diagnostic uncertainty • Missed diagnosis concern • Technology misuse anxiety	• In-person preference	• Legitimacy gap • Unsuitable for complex care
Vergouw et al [33]	• Digital learning curve • Technology usability • Interface complexity	• Limited use perception	• Privacy and security concerns • Technology misuse anxiety	• Symptom articulation • In-person preference	• Legitimacy gap
Knotnerus et al [34]	• Technology usability • Interface complexity • Symptom articulation • Digital learning curve	• Limited use perception • Disrupted continuity of care	• Privacy and security concerns • Technology misuse anxiety	• In-person preference	• Legitimacy gap
Bhatia et al [35]	• Symptom articulation • Technology usability • Cognitive and sensory limitations • Digital learning curve	N/A[a]	• Missed diagnosis concern • Technology misuse anxiety • Missed diagnosis concern	• In-person preference	• Legitimacy gap • Unsuitable for complex care
Lam et al [36]	• Digital learning curve • Symptom articulation	N/A	N/A	• In-person preference	N/A
Nymberg et al [37]	• Digital learning curve • Tech usability. • Technology anxiety • Physical and sensory impairments	• Limited use perception	• Privacy and security concerns • Technology misuse anxiety	• In-person preference • Preference for physical documentation	• Legitimacy gap • Generational digital divide
Houwelingen et al [38]	• Digital learning curve • Technology anxiety • Self-efficacy deficit	N/A	• Technology misuse anxiety • Privacy and security concerns	N/A	N/A

XSL•FO
RenderX

	Functional barriers			Psychological barriers	
Laukka et al [39]	• Interface complexity • Self-efficacy deficit • Language and terminology complexity • Physical and sensory impairments • Technology usability	• Limited use perception	• Fraud and scam concerns • Privacy and security concerns	• In-person preference • Need for familiarity in care	• Generational digital divide • Unsuitable for complex care
Rochmawati et al [40]	• Self-efficacy deficit • Digital learning curve	• Limited use perception • Informality bias	N/A	• In-person preference • Need for familiarity in care	N/A
Fjellså et al [41]	• Technology usability • Interface complexity	• Limited use perception • Informality bias	• Diagnostic uncertainty • Missed diagnosis concern • Technology misuse anxiety	• In-person preference • Need for familiarity in care	• Generational digital divide
Mao et al [42]	• Physical and sensory impairments • Digital learning curve technical challenges language barriers • Cognitive and sensory impairments • Symptom articulation • Physical and sensory impairments • Technical challenges • Technology anxiety • Interface complexity	• Limited use perception • Limited use perception	• Diagnostic uncertainty	• In-person preference	• Unsuitable for complex care
Frishammar et al [43]	• Digital learning curve • Self-efficacy deficit • Technology anxiety	• Limited use perception • Informality bias	• Diagnostic uncertainty • Missed diagnosis concern • Privacy and security concerns	• In-person preference	• Unsuitable for complex care • Legitimacy gap
Haimi et al [44]	• Symptom articulation • Technology anxiety • Language and terminology complexity • Technical challenges • Physical and sensory impairments	N/A	• Missed diagnosis concern	• In-person preference	N/A
Landgren and Cajander [45]		• Limited use perception • Informality bias		• In-person preference	• Generational digital divide

XSL•FO
RenderX

	Functional barriers			Psychological barriers	
	• Interface complexity • Digital learning curve • Self-efficacy deficit • Technology anxiety		• Diagnostic uncertainty • Missed diagnosis concern		
Ahmed et al [46]	• Technology usability • Language and terminology complexity • Interface complexity	• Limited use perception	• Diagnostic uncertainty • Technology misuse anxiety	• In-person preference • Need for familiarity in care	N/A
Ufholz et al [47]	N/A	• Limited use perception	• Privacy and security concerns • Diagnostic uncertainty	• In-person preference	N/A
Sproul et al [48]	• Technology usability	• Limited use perception	• Privacy and security concerns	N/A	• Legitimacy gap

[a]Not applicable.

XSL•FO
RenderX

Table . Thematic categorization and definitions of digital health resistance barriers subcategories among older adults.

Category and subcategory	Definition
Usage barriers	
Symptom articulation [32,34-36,42,44,45]	Difficulty in effectively describing symptoms or raising multiple health concerns during telemedicine or digital health interactions, often due to sensory limitations, cognitive strain, or unfamiliarity with web-based communication formats
Technology usability [32-35,37,39,41,44-46,48]	Difficulties interacting with digital health tools due to poor interface design, complex navigation, multi-step login processes, or lack of age-appropriate accessibility features
Digital learning curve [33-38,40,42,43,45]	Challenges individuals face in acquiring, applying, and retaining the skills required to use digital health technologies, often due to limited prior exposure or memory-related difficulties
Interface complexity [33,34,39,41,42,45,46]	Obstacles users encounter when engaging with digital platforms due to poor design elements, confusing navigation, and unclear layouts
Technology anxiety [37,38,42,43,45]	Fear or discomfort experienced when using digital health technologies, often stemming from low confidence, mistrust in one's digital abilities, or intimidation by unfamiliar systems. This anxiety may lead to hesitation or complete avoidance, driven by concerns about making mistakes that could negatively impact one's health or care
Physical and sensory impairments [35,37,39,42]	Difficulties in using digital health technologies due to age-related sensory and motor impairments, such as reduced vision, hearing loss, or diminished fine motor control
Self-efficacy deficit [38,40,43,45]	A lack of confidence in one's ability to successfully use digital health tools or perform required technological tasks, often rooted in limited digital literacy, minimal prior experience, or insufficient training and support
Language and terminology complexity [39,42,44,46]	Difficulty using digital health tools due to complex medical, technical, or bureaucratic language, often compounded by limited proficiency in the language used by the platform
Value barriers	
Informality bias [32,40,41,43,45]	Reluctance to engage with digital health tools based on the perception that they lack legitimacy or necessity in medical care, accompanied by a belief that traditional health care methods are sufficient without digital augmentation
Limited use perception [32-34,37,39-43,45-48]	The belief that digital health tools offer little to no added value compared with traditional care methods, resulting in low motivation to adopt or engage with them
Risk barriers	
Diagnostic uncertainty [32,41-43,45,46]	Concerns about the accuracy and reliability of medical diagnosis due to the absence of physical examination, direct visual assessment, and potential miscommunication, which may increase the risk of medical errors
Missed diagnosis concern [32,35,41,43-45]	Fear that health care providers will miss critical patient information and that essential health issues may be overlooked due to the absence of physical exams, technical distractions, or miscommunication in digital health interactions
Technology misuse anxiety [32-35,37,38,41,46]	Uncertainty or fear about using digital health technologies incorrectly, driven by concerns about user error, system malfunctions, or communication failures that could negatively impact care delivery
Privacy and security concerns [33,34,37-39,43,47,48]	Concerns about the confidentiality, security, and accuracy of personal medical information in digital health care services, driven by fears of data breaches, unauthorized access, and unreliable IT systems
Tradition barrier	
In-person preference [32-37,39-47]	A strong preference for face-to-face health care interactions, rooted in trust in direct communication, perceived importance of physical examinations, and the belief that in-person care offers superior quality

XSL•FO
RenderX

Category and subcategory	Definition
Need for familiarity in care [39-41,46]	Preference for established health care routines and trusted provider relationships over digital health solutions, due to a desire for personalized care, continuity with known providers, and a reluctance to alter traditional in-person interactions
Image barrier	
Legitimacy gap [32-35,37,43,48]	Perception that digital health care is less effective and trustworthy than traditional in-person care, driven by concerns about depersonalization, bureaucratic complexity, and reduced reliability, leading to skepticism about its value and quality
Unsuitable for complex care [32,35,39,42,43,45]	Perception that digital health care services are insufficient for addressing complex medical conditions or cases requiring physical examination, due to concerns about thoroughness, accuracy, and the ability to provide a comprehensive diagnosis and care
Generational digital divide [37,39,41,45]	Perception that digital health care is designed for younger users and is difficult for older adults to adopt, due to differences in familiarity, confidence, and digital literacy

Table 3 details the barriers identified by each study, presenting a matrix that maps each study to the usage, value, risk, tradition, and image barriers. Table 4 defines each barrier subcategory and summarizes how these resistance themes were operationalized across the studies.

Functional Barriers

In the context of IRT, functional barriers refer to resistance stemming from the practical and objective attributes of the innovation itself, including its required usage, perceived value, and associated risks [19].

Usage Barriers

Usage Barriers were the most consistently reported resistance factor, found in 16 studies. Older adults face significant usage barriers to adopting digital health technologies, largely due to technical challenges, usability difficulties, and concerns about quality of care. A central theme across studies was interface complexity. Many participants described digital health platforms as confusing, unintuitive, and poorly designed. Common challenges included unclear layouts, unintuitive menus, and multi-step authentication processes requiring repetitive actions such as logging in, remembering passwords, and uploading medical documents [32-34,39,41,42,45,46]. These features increased cognitive load and made even basic digital interactions feel burdensome and prone to mistakes.

The difficulties were compounded by technology usability issues linked to age-related cognitive and sensory impairments. Older adults with a decline in vision, hearing loss, or memory difficulties and reduced fine motor skills struggled with small font sizes, poor audio quality, poorly structured information, and touchscreen sensitivity, which makes many applications inaccessible without assistance [35,37,42,44,48]. In addition, language and terminology complexity emerged as a significant obstacle. Technical jargon or unfamiliar medical terms often made it difficult for users to interpret instructions or understand the content presented on-screen, particularly among those with limited formal education or health literacy [39,42,44,46].

Another recurring issue was the digital learning curve. Older adults reported limited prior experience with digital health tools or services and found it challenging to adapt to new systems [32,34,39-43,45]. This often led to a self-efficacy deficit where individuals doubted their ability to complete digital health tasks independently. These doubts fueled hesitation and reinforced a sense of digital exclusion, leading to frustrations, avoidance behaviors, and a greater need for support before successfully adopting telemedicine tools [38-40,43,45]. Closely related to this was technology anxiety, the fear of making mistakes or causing harm through improper use, which discouraged many from engaging with telemedicine platforms.

Concerns about system reliability and uncertainty about using digital health care tools make older adults feel less confident in their technical abilities and unprepared [32,33,36,39,40,42,43,45], leading to avoidance behaviors, where they opt not to engage with digital health solutions to minimize the risk of errors [37,38,42].

Beyond usability concerns, preadoption resistance arises from changes in communication dynamics within digital health care. In contrast to traditional face-to-face consultations, which allow patients to express multiple health concerns in a single visit and rely on nonverbal cues, digital health platforms, particularly telemedicine services, alter this dynamic. Studies showed that when older adults use digital health services, they struggle to articulate their symptoms or find it difficult to understand medical terminology or provider explanations [39,45]. As a result, they hesitate to fully communicate medical concerns, whether typing them into digital platforms or discussing multiple health issues during digital visits. This contributes to a perception that digital care is less effective than in-person care [34-36], further discouraging older adults from fully embracing digital health technologies.

Value Barriers

Value barriers to adopting digital health solutions among older adults primarily stem from informality bias, the perceived lack of necessity of digital tools, concerns about care quality, and misalignment between the effectiveness of available digital health care services and patient expectations [40,43,45]. While many acknowledge that telemedicine may be appropriate for minor health issues and routine follow-ups, they often do not

XSL•FO
RenderX

view it as an adequate substitute for in-person consultations. This limited use perception is particularly strong when it comes to complex conditions that require physical examination or long-term management [32,34,41-43,45,46,48].

Skepticism about the effectiveness of remote consultations is a common concern. Many older adults feel that digital platforms fail to capture nonverbal cues, which are essential for accurate medical assessment and effective patient-provider communication. This concern is particularly pronounced among individuals managing chronic illnesses, who consider ongoing physical evaluations and in-person interactions with health care professionals to be vital components of proper care [33]. Moreover, older adults often emphasize the importance of relational continuity with their health care providers, an aspect they feel is disrupted and compromised in digital health environments. Telemedicine is frequently perceived as impersonal and transactional, lacking the trust and emotional support that typically characterize in-person visits, qualities that many older adults highly value in primary care settings [33,34,37,39]. As a result, some individuals refuse to see their providers outside of traditional clinical settings, which further reinforces resistance to digital health solutions [37,42].

Beyond concerns about quality of care, many older adults also question the necessity of digital health interventions, particularly when the current health care system meets their needs effectively [33,37]. Some dismissed telemedicine as a "solution for a nonexisting problem," believing that traditional in-person visits provide sufficient care without the added complexity of digital tools [33,37,40]. This skepticism is often exacerbated by low digital literacy or past negative experiences with digital health technology, leading many to view telemedicine and digital health apps as unnecessary, ineffective, or not worth the effort required to learn and adapt [37]. When the perceived benefits of digital health do not outweigh the effort and risks associated with adoption, resistance to these solutions remains strong.

Risk Barriers

Risk barriers to digital health adoption among older adults primarily revolve around concerns about diagnostic uncertainty and the potential of missed health issues due to the absence of physical examinations, body language, and other visual cues essential to accurate clinical assessment [32,41-46]. Many older individuals worry that the lack of hands-on evaluation in telemedicine could lead to overlooked symptoms or misinterpretations by health care providers. A prominent concern is technology misuse anxiety, which arises from fear of making errors during digital interactions. Participants described anxiety about technical distractions, errors in digital documentation, incomplete data entry, and uncertainty about whether submitted information, such as messages, forms, or test results, would be properly received and understood by their health care team [41,45,46]. These apprehensions are linked to fears of miscommunication with health care providers, incorrect medical decisions, or overlooked health conditions [32-35,37,40].

Beyond diagnostic concerns, older adults express privacy and security concerns. There is a common distrust of the integrity and security of digital health platforms [33,34,37-40,43,47,48], particularly related to fear that personal health data could be exposed to unauthorized access, fraud, or misuse. Some participants described concerns about scams that mimic legitimate digital services, increasing their reluctance to trust or engage with digital health tools [39]. This skepticism is further compounded by uncertainty around how health care institutions collect, store, and share data through electronic health records and patient portals [39].

Additionally, lack of confidence in digital skills was repeatedly cited as a major factor behind misuse anxiety. Older adults often lack confidence in their digital skills, particularly in navigating complex interfaces or troubleshooting technical issues. Common fears included accidentally deleting important information, misunderstanding medical results, or failing to complete critical health care tasks [34,40]. As a result, many preferred to avoid digital health services entirely rather than risk making mistakes that could negatively impact their care.

Another key source of resistance is the perceived loss of autonomy in health care decision-making. Some older adults expressed concerns that eHealth solutions shift decision-making control from patients to automated systems, reducing their ability to advocate for personalized care and communicate effectively with health care providers about their health care [37,41]. This fear is particularly prevalent among those unfamiliar with electronic health records or unaware of how to use digital clinical discussions.

Psychological Barriers

Psychological barriers refer to resistance stemming from subjective, cognitive, and emotional conflicts between the innovation and the individual's established traditions and self-image barriers [19].

Tradition Barriers

Traditional barriers to digital health adoption among older adults arise from long-established care routines, personal preferences for in-person interactions, and a strong need for familiarity in health care interactions. Many older adults have their health-seeking behaviors around face-to-face consultations, expressing satisfaction with traditional care models and questioning the necessity or value of digital alternatives [37,40,41]. They often perceive little incentive to switch to eHealth services when current systems already meet their expectations [37,44,45]. A central theme is the belief that *in-person interactions* offer superior quality of care, stronger provider-patient relationships, and greater emotional warmth. Digital platforms are often seen as impersonal, lacking the human touch and nonverbal communication cues that older adults consider essential for effective medical consultations [32,34-36,39,41,43,45-47]. This is especially concerning for individuals managing chronic conditions or complex health issues, where verbal-only communication may be insufficient for accurate symptom reporting and clinical assessment [32,33].

The need for familiarity in care also contributes to resistance toward digital health adoption. Many older adults prefer continuity with known health care professionals, such as physicians, nurses, or other health care professionals, and value personalized guidance and documentation, such as printed instructions or handwritten over generic digital content. Some

XSL•FO
RenderX

do not want all services to be transferred through digital platforms, especially when health care and social service issues are too complex to be handled without face-to-face contact [35,37,39,42,46].

Another common concern is the perceived legitimacy of telemedicine. Some older adults do not view phone or video consultations as valid medical encounters, describing them as informal and lacking the authority of traditional office visits [32]. This perception is heightened among individuals who had not used digital health before the COVID-19 pandemic and who experienced the rapid shift to telehealth as both disruptive and disorienting, owing to complex interfaces and limited user guidance [34]. For these individuals, digital health solutions interfere with familiar health care routines and pose significant adaptation challenges [39].

Image Barriers

Image barriers to digital health adoption among older adults arise from negative perceptions of technology, distrust in digital health solutions, and skepticism about their legitimacy and effectiveness in clinical care. Many older adults associate digital health technologies with lower quality of care and consider them as an unacceptable alternative to traditional in-person visits [35,42]. For some, these technologies are viewed as overly complex, impersonal, and rigid, contributing to a Legitimacy Gap, a perception that digital health care lacks the authenticity, reliability, and interpersonal value of conventional medical interactions [33,43,48]. This skepticism is reinforced by the belief that health care should be hands-on, personalized, and relational, the qualities they feel digital platforms fail to deliver.

Another central issue underlying this perception is the Generational Digital Divide. Many older adults view digital health tools as designed primarily for younger, digitally proficient users, and they report feeling excluded or disadvantaged by their limited experience with digital technologies [37,39-41,45]. This belief is often coupled with self-perceived technological inadequacy, where individuals feel “too old” to learn or incapable of using new systems effectively [39]. These psychological barriers are compounded by negative past encounters with health care bureaucracy or poorly designed interfaces, which foster the impression that digital health prioritizes efficiency over patient-centered care [37]. Additionally, difficulties navigating eHealth platforms often lead to a sense of powerlessness in managing their health, further alienating them from digital solutions.

Older adults also view telemedicine and digital health as unsuitable for both routine and complex care needs [32,42,43]. Many perceive these technologies as inferior to traditional, in-person medical consultations, citing concerns about their inability to provide thorough physical examinations, comprehensive assessments, and hands-on diagnostics [39]. Digital health is also associated with social isolation and reduced autonomy, as some fear that shifting toward digital health care may limit direct patient-provider interactions and diminish their role in medical decision-making [34]. This contributes to a strong preference for traditional care models, where in-person visits provide greater trust, familiarity, and perceived quality.

Evidence and Gap Map

Across the included studies, there was substantial variation in both the types of digital health technologies examined and the specific resistance factors reported. To strengthen the mapping component of this scoping review, we developed an evidence and gap map to summarize the distribution of evidence and identify gaps across digital health modalities and resistance constructs. Guided by IRT, we categorized studies by the type of digital health modality and by IRT-informed barrier subcategories derived from the extracted findings (Figure 2).

Figure 2. Evidence and gap map of digital health modalities by IRT-informed resistance subcategories in primary care among older adults. Bubble size and color intensity represent the number of included studies contributing to each intersection (n=17). IRT: innovation resistance theory.

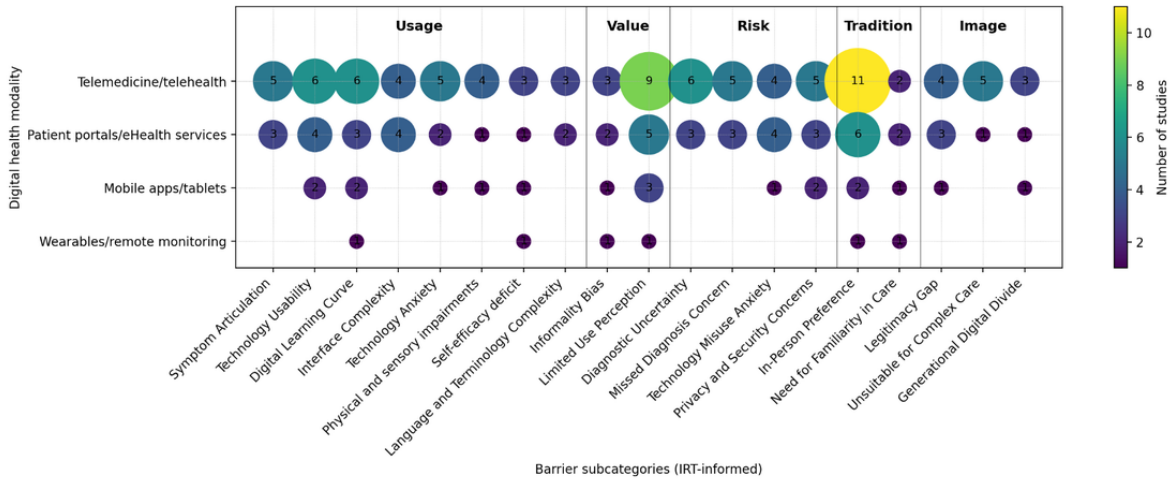

Specifically, the map highlights that evidence is concentrated in studies of telemedicine and patient portals or eHealth services, with fewer studies addressing mobile apps or tablets and minimal evidence on wearables or remote monitoring. Across modalities, frequently represented barriers included usability and interface complexity, self-efficacy and technology anxiety, and trust-related concerns such as privacy, data security, and perceived legitimacy of digital encounters. In contrast, several modalities-barrier intersections show limited or absent evidence, indicating that resistance to certain technologies, particularly wearables and app-based monitoring, remains underexplored in primary care contexts.

XSL•FO
RenderX

Conceptual Integration: Interconnected Barriers Leading to Digital Health Avoidance

Across the 17 included studies, usage barriers were the most consistently reported (16/17 studies). Risk barriers and tradition barriers were also prevalent (15/17 studies). Value barriers were common (13/17 studies), and image barriers were reported in a smaller, but still substantial subset (11/17 studies). Co-occurrence patterns were apparent across domains, and worked examples illustrate how linkages were derived. For example, one participant described limiting use to familiar functions and avoiding other features, indicating a usage barrier, accompanied by anxiety when stepping outside her comfort zone, suggesting an affective risk component and fear of making mistakes: "I never look over there, I just do everything I have learned… Outside of that, I become nervous." [38]. In another study, a participant noted that he did not grow up with technology, indicating a usage barrier related to limited digital familiarity, and expressed a tradition barrier by preferring to arrange appointments by phone and speak with the physician face-to-face rather than use digital channels: "But we did not grow up with the computer. I would rather make a phone call to arrange an appointment and prefer to talk face-to-face to the physician" [33]. Another participant questioned the adequacy of digital encounters for a proper clinical assessment, reflecting an image or quality concern that co-occurred with a tradition-related preference for face-to-face care and an implied need for greater diagnostic assurance (risk): "I would rather that the doctor can actually touch me, examine me with a stethoscope… I also think in-person communication is sometimes better…" [42]. Together, these patterns suggest that resistance is rarely attributable to a single factor; rather, studies frequently report clusters of functional and psychological barriers that co-occur. These recurring clusters informed the relational integration step; linkages were coded as explicit when directly stated in study results or participant quotes, inferential when implied through within-study co-occurrence and narrative context, and integrative when synthesized across multiple studies showing consistent patterns.

As part of the relational integration step of our synthesis (Stage 5), we developed a conceptual model that integrates the identified barriers into an interconnected structure (Figure 3). This conceptual integration was undertaken to move beyond listing individual barriers and to summarize recurring co-occurrence patterns observed across the studies. The interconnected nature of resistance barriers creates a self-reinforcing reaction cycle that leads to avoidance behaviors among older adults. Rather than operating in isolation, functional and psychological barriers interact dynamically, compounding resistance and entrenching disengagement from digital health platforms.

Figure 3. Conceptual model of interconnected resistance barriers leading to digital health avoidance among older adults in primary care, interacting co-occurrence patterns across included studies to illustrate directional relationships and feedback loops.

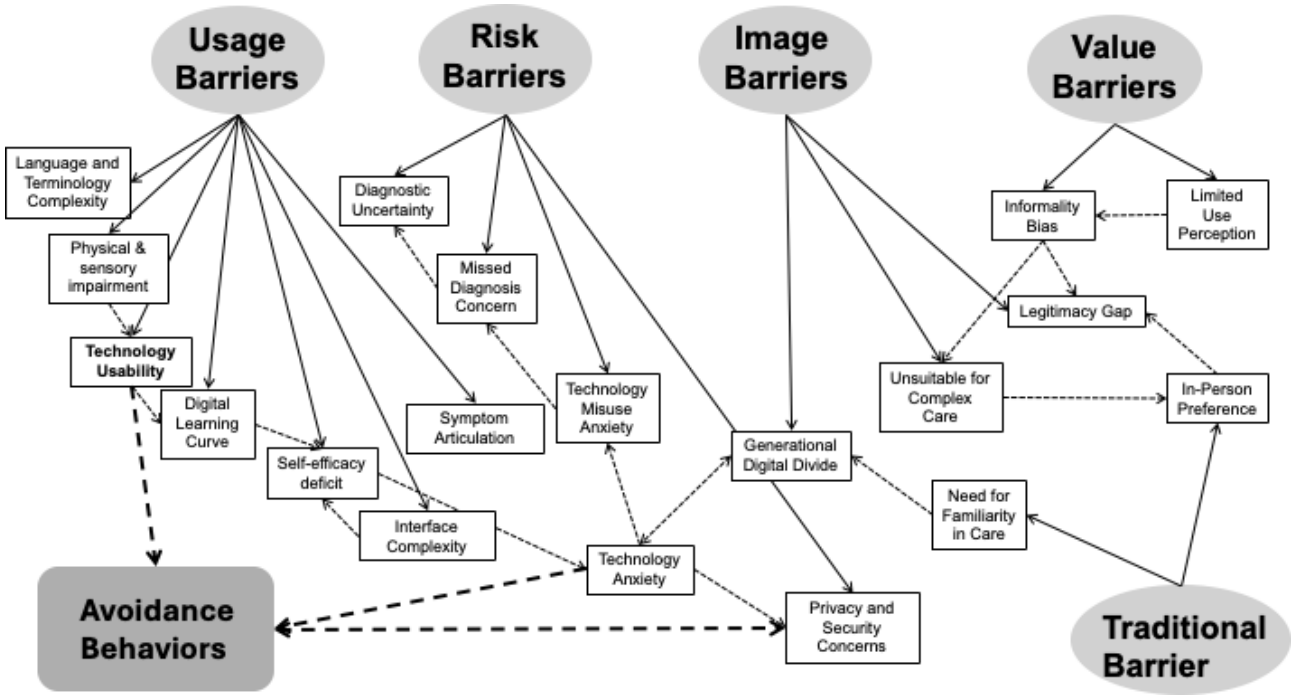

Technology usability challenges contribute to difficulties in the digital learning curve, which, along with interface complexity, results in a self-efficacy deficit and a lack of confidence in using digital health technologies. This diminished self-efficacy further fuels technology anxiety, increasing hesitation and discouraging engagement. Importantly, these usability issues do not just reduce confidence; they initiate a cascade of psychological barriers that elevate emotional discomfort and cognitive overload. Figure 3 illustrates this cascading effect: a feedback system where usability problems initiate low self-efficacy, which in turn escalates into technology anxiety. This psychological discomfort amplifies risk perceptions, including fear of misdiagnosis, privacy breaches, and technology misuse. These concerns reduce trust in digital health care solutions and reinforce avoidance behaviors. Privacy and security concerns and technology anxiety reinforce each other, creating a cycle of distrust. As the trust in the system diminishes, older adults become less likely to interact with digital platforms, which

XSL•FO
RenderX

limits exposure and impedes skill acquisition, further deepening their self-efficacy deficit. This cycle in Figure 3 is illustrated through closed feedback loops, where arrows between barriers represent how one resistance factor amplifies another (eg, Interface Complexity → Low Self-Efficacy → Technology Anxiety → Avoidance).

Traditional barriers, such as a strong preference for in-person care and the need for familiarity, also strengthen image barriers, including the legitimacy gap and the generational digital divide, further discouraging digital health adoption. As shown in Figure 3, these values-based preferences and generational perceptions reinforce internal skepticism with digital tools, particularly when technology is perceived as impersonal. The legitimacy gap reflects older adults' perception that digital tools lack the authenticity and authority of face-to-face care, while the generational divide reinforces feelings of exclusion from technologies perceived as designed for younger users. Figure 3 also highlights this convergence between identity-based resistance (eg, tradition/image) and capability-based resistance (eg, usability, anxiety). Together, these interrelated barriers form a self-reinforcing loop, where initial usability difficulties and emotional skepticism amplify resistance, which leads to withdrawal from digital health use entirely (Figure 3).

Discussion

Principal Findings

This scoping review applied the IRT to examine older adults' resistance to digital health technologies within primary care contexts. Across the included studies, we found consistent functional barriers (such as usability difficulties, interface complexity, and sensory or cognitive limitations) and recurrent psychological barriers (such as a preference for in-person care and concerns about the legitimacy of digital encounters), with value-related concerns (limited perceived benefit) and risk-related concerns (diagnostic uncertainty, privacy, and security worries) also prominent.

The findings suggest that resistance is not a static failure to adopt nor a passive disengagement, but rather a dynamic, emotionally embedded process. This process is shaped by the interaction of functional and psychological factors, including identity and value-related concerns, which do not operate in isolation but reinforce each other in feedback loops that entrench avoidance behaviors over time. The interplay between usability challenges, emotional discomfort, and value-based misalignment reflects the multifaceted nature of resistance in this population. Also, interrelationships indicated that capability-related barriers erode confidence and increase anxiety, while identity-related concerns reinforce distrust and preference for face-to-face care, together discouraging engagement. Linkages were categorized by evidentiary basis (explicit, inferential, integrative), supporting IRT as a useful framework for organizing and interpreting resistance patterns.

Functional barriers such as interface complexity, digital learning curves, and age-related sensory or cognitive limitations were among the most identified sources of resistance. However, their significance lies not only in their prevalence but in their role as catalysts: they often trigger negative psychological responses, including diminished self-efficacy, anxiety, and fear of error. These emotional reactions contributed to a broader sense of technological vulnerability and led to sustained disengagement, demonstrating how technical design and user experience are deeply interconnected.

Beyond usability, resistance was often rooted in symbolic and identity-related concerns. A preference for face-to-face interactions, generational beliefs regarding technology, and the desire for continuity with known providers were consistently linked to what can be described as symbolic distancing, a form of resistance grounded in perceived legitimacy and personal norms. Even where functionality improved, older adults continued to express skepticism, viewing digital tools as impersonal, exclusionary, or inappropriate for managing complex health needs. This suggests that emotional and symbolic dimensions may play a stronger influence on resistance than previously recognized.

These insights align with earlier theoretical work that repositions resistance as a dynamic, emotionally driven response process. The findings support an evolving theoretical perspective that frames resistance as an active process. Rather than being the inverse of adoption, resistance emerges from distinct cognitive and emotional pathways and may dominate decision-making even in the presence of positive attitudes [49]. Other research has also shown that tradition and identity-based concerns frequently outweigh usability considerations in shaping innovation rejection, particularly in service-oriented settings [20]. This review affirms that older adults' resistance is rarely due to a lack of awareness or rational evaluation alone, but rather reflects deeply embedded emotional and symbolic stances.

Breaking these loops requires targeted interventions that not only simplify interface design but also rebuild self-efficacy, trust, and the perceived legitimacy of digital care. Accordingly, programs should pair practical usability supports (eg, task simplification, assisted-digital options, scaffolded practice) with psychological strategies (eg, anxiety reduction, trust-building, culturally and linguistically responsive framing).

Comparison to Prior Work

The findings of this review both align with and challenge established models of technology acceptance. For instance, it complements the critiques of the extended UTAUT, which has been applied to prior studies involving older adults in health care settings. One study has highlighted effort expectancy, perceived usefulness, and trust in health care providers as primary predictors of adoption. While these factors remain relevant, this review suggests they are insufficient to fully account for persistent resistance observed in older populations. This resistance appears to stem not from a lack of understanding but from deeper emotional and symbolic misalignments between digital tools and the users' personal values, care routines, or generational identities [50]. In this context, resistance is not a knowledge deficit but a deliberate, emotionally grounded response to perceived risks, impersonality, or social exclusion. Our synthesis clarifies how such misalignments link to concrete pathways (eg, usability → low self-efficacy → anxiety → avoidance), adding a mechanism to prior critiques.

XSL•FO
RenderX

Reinhardt et al [51] claim in their study that resistance to innovation is not merely the opposite of adoption but a distinct phenomenon that operates through its own logic and dynamics, and thus warrants a separate theoretical approach. They proposed the concept of "adoption triggers," external events or contextual changes that interrupt entrenched resistance and enable eventual uptake. This finding aligns with the results of this review, where participants continued to resist engagement even after usability improvements, suggesting that design enhancements alone are insufficient [51]. Psychosocial catalysts such as trust in providers, alignment with identity, or significant life transitions may be necessary to shift deeply embedded resistance patterns.

Further support comes from the argument that TAM and UTAUT, widely used models, were not originally developed for health care but rather in organizational contexts. Like IRT, they were formulated outside the health domain and may require adaptation when applied in complex settings, such as digital health for older adults. In their original formulations, these models assume that perceived usefulness and ease of use directly predict technology acceptance. However, in health care, these assumptions are challenged, especially in the context of older adult users [52]. Health care studies often have to add context-specific variables such as computer anxiety, trust, or physician endorsement to increase explanatory power. This suggests that existing models may benefit from complementary perspectives that foreground resistance shaped by emotional discomfort and identity-related concerns, including symbolic dissonance around how digital health fits with older adults' roles and expectations. This review affirms the need to view resistance among older adults as socially embedded and identity-relevant, rather than reducible to issues of usability or cognitive evaluation.

Resistance constructs are not intended to replace established acceptance models such as TAM and UTAUT, but to extend them and provide a more complete account of older adults' technology use and nonuse patterns. Yu et al [53], in their research, also extend UTAUT with aging-specific variables such as perceived physical condition, self-actualization needs, and technology anxiety. Their empirical study among Chinese older adults found that while traditional UTAUT predictors (eg, performance and effort expectancy) remain significant, behavioral use was also shaped by perceived physical limitations and psychological needs for self-fulfillment. Notably, the effect of technology anxiety was nonsignificant, suggesting that usability alone does not explain resistance; rather, broader psychosocial and experiential factors must be considered [53]. These adaptations have introduced constructs such as perceived physical condition, self-actualization needs, and psychosocial well-being to better explain behavioral engagement with health care conversational agents among older adults. Our mapping complements these extensions by locating these constructs within the IRT domains and by indicating which inter-barrier links are explicitly supported by the literature.

Theoretical Implications

This review advances theory on digital health adoption and resistance among older adults in 2 main ways. First, it refines IRT for the context of aging and digital health by highlighting aging-specific resistance themes such as legitimacy gaps, generational digital divides, and anxiety about technology misuse as candidates for further conceptualization and measurement within the original IRT domains. Second, it points to resistance as a dynamic process in which these barriers interact in feedback patterns rather than operating as isolated categories. This mechanism-oriented view complements existing TAMs by underscoring that persistent nonuse reflects active, emotionally and symbolically shaped resistance, rather than merely weak adoption intentions.

Practical Implications

From a gerontechnology and age-inclusive design perspective, the IRT-based model translates the identified barriers and linkages into actionable design and implementation levers to reduce resistance among older adults in primary care. This review has important implications for digital health design, practice, and policy.

First, the disproportionate concentration of extant research within high-income Western countries necessitates a nuanced approach to global implementation, as resistance profiles are not homogenous but are contingent upon divergent socioeconomic structures, varying levels of digital literacy, and culture-specific perceptions of aging [54]. Addressing these complexities requires a paradigm shift from a reactive model, characterized by a narrow focus on technical troubleshooting and interface simplification, toward a proactive design. While mitigating interface complexity and accommodating sensory impairments remain fundamental requirements, such technical refinements in isolation are insufficient to resolve resistance that is fundamentally anchored in emotional and psychological factors. Consequently, proactive age-tech development should prioritize the alignment of digital interventions with users' long-standing traditions and the preservation of relational continuity in care [55]. By acknowledging traditional barriers and framing digital tools as seamless extensions of familiar, trusted care routines rather than disruptive innovations, developers can transition from delivering impersonal technical products to co-creating solutions that resonate with the core identities and values of older populations.

Building on the conceptual model in Figure 2, breaking the self-reinforcing cycle of resistance requires targeted interventions that address both practical usability barriers and underlying psychological resistance; focusing on interface design or digital literacy alone is unlikely to change deeply rooted patterns of nonuse. Designers need to focus not only on functionality but also on providing emotional reassurance and strengthening the perceived legitimacy and social meaning of digital care. Therefore, solutions should be co-designed with older adults not only to ensure they fit with their routines, communication styles, and cultural values, but also to directly address the specific IRT barriers identified in this review by incorporating strategies that reduce friction and promote confidence. These strategies may include simplifying high-friction tasks by using shorter flows, fewer required fields, larger tap targets, and accessible defaults. Also, designers can provide stepwise guidance and "practice mode," and offer assisted-digital options such as telephone call-back support,

XSL•FO
RenderX

shared on-screen navigation with staff, and on-site digital stations within clinics where staff can help patients complete digital tasks.

Privacy, risk perceptions, and distrust emerged as central barriers in our synthesis. Digital health platforms should incorporate trust-enhancing features, including sustained relationships with known providers, easy access to human support, and clear, simple explanations of data practices. To strengthen perceived legitimacy, systems should preserve care delivery choice (seamless switch to phone or in-person visits), display continuity cues (named clinician, photo, prior encounters), and surface concrete benefits (time saved, refill accuracy, faster appointments). Culturally and linguistically responsive content, combined with feedback that reinforces mastery, can further mitigate anxiety and improve self-efficacy, helping to disrupt the self-reinforcing loops that lead to avoidance. Together, these design-oriented recommendations translate our conceptual findings into practical guidance for technology designers and implementers seeking to reduce resistance among older adults.

Future Research Directions

Future research should investigate the temporal evolution of resistance, including how initial avoidance may shift or diminish over time, and under what conditions. There is a need to explore resistance dynamics among underrepresented populations, such as ethnic minorities, linguistically diverse groups, and individuals living in lower-resource settings. In line with Bevilacqua et al [56], emerging work on service-specific acceptance measures for older adults who developed the Robot-Era Inventory as a tailored acceptance scale for a social robotics platform, and called for customizable, context-specific tools tailored to specific technologies and services for older adults, future studies should develop and validate IRT-informed scales tailored to particular digital health modalities [56]. In addition, longitudinal and mixed-methods designs could provide deeper insight into how resistance is maintained or disrupted. Finally, the development and empirical testing of interventions grounded in IRT would help bridge the gap between theory and design strategies.

Strengths and Limitations

A key strength of this review is its structured, theory-driven synthesis across diverse empirical studies. By applying the IRT to various study designs and health care contexts, this review enhances the conceptual understanding of digital resistance among older adults. It was conducted according to best-practice guidelines for scoping reviews, which reflect established methodological standards.

Several limitations should be noted. First, the search was restricted to English-language publications, which might have excluded relevant studies published in other languages. Second, the review encompasses studies published between 2014 and 2025, a period characterized by rapid technological advancement. Improvements in device usability during this time may have influenced user experiences and patterns of resistance, potentially affecting cross-study comparability. Third, most of the included studies were conducted in high-income Western countries, and the patterns of resistance identified here may not fully capture experiences in lower-income or non-Western contexts, where digital infrastructures, health systems, and cultural norms around aging and technology may differ substantially. This concentration substantially reduces generalizability beyond high-income Western settings and limits the applicability of our findings to global contexts where digital literacy, socioeconomic factors, and cultural perceptions of aging and health care may create distinct resistance profiles. Fourth, none of the included studies reported participants' cognitive status or used standardized cognitive screening measures. As a result, we could not examine whether resistance barriers vary by cognitive integrity or distinguish attitudinal resistance from barriers related to cognitive impairment, which may influence learnability, confidence, and sustained use of digital health technologies. Finally, the proposed conceptual model has not yet been validated in practice and should be regarded as hypothesis-generating. Future research should operationalize the IRT domains and evaluate their factor structure, reliability, and predictive validity in empirical studies.

Conclusions

Applying IRT to older adults' experiences with digital health shifts the focus from "lack of readiness" or skills gaps to resistance mechanisms and how technologies are designed and integrated into primary care. Resistance emerges as an active, emotionally rooted process involving functional, psychological, and identity-based barriers to adoption, and this review integrates recurring co-occurrence patterns into a conceptual model, thereby moving beyond prior work that lists barriers in isolation. The synthesis clarifies how usability problems can undermine self-efficacy, increase technology anxiety, and amplify trust and legitimacy concerns, creating feedback loops that reinforce avoidance. Real-world implications: implementation strategies should go beyond technical usability by rebuilding emotional trust, supporting relational continuity, and aligning digital solutions with older adults' values and routines through meaningful channel choice and transparent communication about risks. In addition, IRT offers a structure for developing domain-specific measures and interventions that address usage, value, risk, tradition, and image barriers, supporting a more realistic and equitable digital transformation in primary care for aging populations.

Data Availability

The datasets generated and analyzed during this study are reported in the article and multimedia appendix.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Online search strategy.
[DOCX File, 18 KB - jmir_v28i1e75591_app1.docx]

Checklist 1
PRISMA-S Checklist.
[DOCX File, 17 KB - jmir_v28i1e75591_app2.docx]

Checklist 2
Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist.
[DOCX File, 56 KB - jmir_v28i1e75591_app3.docx]

References

1. Chen C, Ding S, Wang J. Digital health for aging populations. Nat Med 2023 Jul;29(7):1623-1630. [doi: 10.1038/s41591-023-02391-8] [Medline: 37464029]
2. Mitchell M, Kan L. Digital technology and the future of health systems. Health Syst Reform 2019;5(2):113-120. [doi: 10.1080/23288604.2019.1583040] [Medline: 30908111]
3. Stoumpos AI, Kitsios F, Talias MA. Digital transformation in healthcare: technology acceptance and its applications. Int J Environ Res Public Health 2023 Feb 15;20(4):3407. [doi: 10.3390/ijerph20043407] [Medline: 36834105]
4. Alruwaili MM, Shaban M, Elsayed Ramadan OM. Digital health interventions for promoting healthy aging: a systematic review of adoption patterns, efficacy, and user experience. Sustainability 2023 Dec 2;15(23):16503. [doi: 10.3390/su152316503]
5. Jokisch MR, Schmidt LI, Doh M, Marquard M, Wahl HW. The role of internet self-efficacy, innovativeness and technology avoidance in breadth of internet use: comparing older technology experts and non-experts. Comput Human Behav 2020 Oct;111:106408. [doi: 10.1016/j.chb.2020.106408]
6. Dhagarra D, Goswami M, Kumar G. Impact of trust and privacy concerns on technology acceptance in healthcare: an Indian perspective. Int J Med Inform 2020 Sep;141:104164. [doi: 10.1016/j.ijmedinf.2020.104164] [Medline: 32593847]
7. Zhou J, Salvendy G, Boot WR, et al. Grand challenges of smart technology for older adults. International Journal of Human–Computer Interaction 2025:1-43. [doi: 10.1080/10447318.2025.2457003]
8. Rahimi B, Nadri H, Lotfnezhad Afshar H, Timpka T. A systematic review of the technology acceptance model in health informatics. Appl Clin Inform 2018 Jul;9(3):604-634. [doi: 10.1055/s-0038-1668091] [Medline: 30112741]
9. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
10. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view1. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]
11. Rogers EM. Lessons for guidelines from the diffusion of innovations. Jt Comm J Qual Improv 1995 Jul;21(7):324-328. [doi: 10.1016/s1070-3241(16)30155-9] [Medline: 7581733]
12. Lee C, Coughlin JF. PERSPECTIVE: Older adults' adoption of technology: an integrated approach to identifying determinants and barriers. J of Product Innov Manag 2015 Sep;32(5):747-759 [FREE Full text] [doi: 10.1111/jpim.12176]
13. Wilson J, Heinsch M, Betts D, Booth D, Kay-Lambkin F. Barriers and facilitators to the use of e-health by older adults: a scoping review. BMC Public Health 2021 Aug 17;21(1):1556. [doi: 10.1186/s12889-021-11623-w] [Medline: 34399716]
14. Cajita MI, Hodgson NA, Lam KW, Yoo S, Han HR. Facilitators of and barriers to mHealth adoption in older adults with heart failure. Comput Inform Nurs 2018 Aug;36(8):376-382. [doi: 10.1097/CIN.0000000000000442] [Medline: 29742549]
15. Fox G, Connolly R. Mobile health technology adoption across generations: narrowing the digital divide. Information Systems Journal 2018 Nov;28(6):995-1019 [FREE Full text] [doi: 10.1111/isj.12179]
16. Russell TG, Gillespie N, Hartley N, Theodoros D, Hill A, Gray L. Exploring the predictors of home telehealth uptake by elderly Australian healthcare consumers. J Telemed Telecare 2015 Dec;21(8):485-489. [doi: 10.1177/1357633X15606264] [Medline: 26391512]
17. Mumtaz H, Riaz MH, Wajid H, et al. Current challenges and potential solutions to the use of digital health technologies in evidence generation: a narrative review. Front Digit Health 2023;5:1203945. [doi: 10.3389/fdgth.2023.1203945] [Medline: 37840685]
18. Hoque R, Sorwar G. Understanding factors influencing the adoption of mHealth by the elderly: an extension of the UTAUT model. Int J Med Inform 2017 May;101:75-84. [doi: 10.1016/j.ijmedinf.2017.02.002]
19. Ram S, Sheth JN. Consumer resistance to innovations: the marketing problem and its solutions. J Consum Mark 1989 Feb 1;6(2):5-14. [doi: 10.1108/EUM0000000002542]
20. Laukkanen T. Consumer adoption versus rejection decisions in seemingly similar service innovations: the case of the internet and mobile banking. J Bus Res 2016 Jul;69(7):2432-2439. [doi: 10.1016/j.jbusres.2016.01.013]

XSL•FO
RenderX

21. Laukkanen P, Sinkkonen S, Laukkanen T. Consumer resistance to internet banking: postponers, opponents and rejectors. International Journal of Bank Marketing 2008 Sep 5;26(6):440-455. [doi: 10.1108/02652320810902451]
22. Kaur P, Dhir A, Ray A, Bala PK, Khalil A. Innovation resistance theory perspective on the use of food delivery applications. Journal of Enterprise Information Management 2021 Nov 11;34(6):1746-1768. [doi: 10.1108/JEIM-03-2020-0091]
23. Castro CAB, Zambaldi F, Ponchio MC. Cognitive and emotional resistance to innovations: concept and measurement. JPBM 2019 Jul 4;29(4):441-455. [doi: 10.1108/JPBM-10-2018-2092]
24. Cieslak V, Valor C. Moving beyond conventional resistance and resistors: an integrative review of employee resistance to digital transformation. Cogent Business & Management 2025 Dec 12;12(1). [doi: 10.1080/23311975.2024.2442550]
25. Alshallaqi M, Al Halbusi H, Abbas M, Alhaidan H. Resistance to innovation in low-income populations: the case of university students' resistance to using digital productivity applications. Front Psychol 2022;13:961589. [doi: 10.3389/fpsyg.2022.961589] [Medline: 36275207]
26. Spinelli G, Gastaldi L, Van Hove L, Van Droogenbroeck E. Can cluster analysis enrich the innovation resistance theory? the case of mobile payment usage in Italy. Technol Soc 2024 Dec;79:102729. [doi: 10.1016/j.techsoc.2024.102729]
27. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
28. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5(1):69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
29. The Joanna Briggs Institute reviewers' manual. : The Joanna Briggs Institute; 2015 URL: https://reben.com.br/revista/wp-content/uploads/2020/10/Scoping.pdf [accessed 2026-01-28]
30. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
31. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
32. Khanassov V, Ilali M, Ruiz AS, Rojas-Rozo L, Sourial R. Telemedicine in primary care of older adults: a qualitative study. BMC Prim Care 2024 Jul 17;25(1):259. [doi: 10.1186/s12875-024-02518-x] [Medline: 39020277]
33. Vergouw JW, Smits-Pelser H, Kars MC, et al. Needs, barriers and facilitators of older adults towards eHealth in general practice: a qualitative study. Prim Health Care Res Dev 2020 Dec 2;21:e54. [doi: 10.1017/S1463423620000547] [Medline: 33263272]
34. Knotnerus HR, Ngo HTN, Maarsingh OR, van Vugt VA. Understanding older adults' experiences with a digital health platform in general practice: qualitative interview study. JMIR Aging 2024 Aug 30;7:e59168. [doi: 10.2196/59168] [Medline: 39212599]
35. Bhatia R, Gilliam E, Aliberti G, et al. Older adults' perspectives on primary care telemedicine during the COVID-19 pandemic. J Am Geriatr Soc 2022 Dec;70(12):3480-3492. [doi: 10.1111/jgs.18035] [Medline: 36169152]
36. Lam K, Lu AD, Shi Y, Covinsky KE. Assessing telemedicine unreadiness among older adults in the united states during the COVID-19 pandemic. JAMA Intern Med 2020 Oct 1;180(10):1389-1391. [doi: 10.1001/jamainternmed.2020.2671] [Medline: 32744593]
37. Nymberg VM, Bolmsjö BB, Wolff M, Calling S, Gerward S, Sandberg M. "Having to learn this so late in our lives…" Swedish elderly patients' beliefs, experiences, attitudes and expectations of e-health in primary health care. Scand J Prim Health Care 2019 Mar;37(1):41-52. [doi: 10.1080/02813432.2019.1570612] [Medline: 30732519]
38. van Houwelingen CT, Ettema RG, Antonietti MG, Kort HS. Understanding older people's readiness for receiving telehealth: mixed-method study. J Med Internet Res 2018 Apr 6;20(4):e123. [doi: 10.2196/jmir.8407] [Medline: 29625950]
39. Laukka E, Lakoma S, Harjumaa M, et al. Older adults' preferences in the utilization of digital health and social services: a qualitative analysis of responses to open-ended questions. BMC Health Serv Res 2024 Oct 4;24(1):1184. [doi: 10.1186/s12913-024-11564-1] [Medline: 39367429]
40. Rochmawati E, Kamilah F, Iskandar AC. Acceptance of e-health technology among older people: a qualitative study. Nurs Health Sci 2022 Jun;24(2):437-446. [doi: 10.1111/nhs.12939] [Medline: 35297152]
41. Hunsbedt Fjellså HM, Husebø AML, Braut H, Mikkelsen A, Storm M. Older adults' experiences with participation and eHealth in care coordination: qualitative interview study in a primary care setting. J Particip Med 2023 Oct 2;15:e47550. [doi: 10.2196/47550] [Medline: 37782538]
42. Mao A, Tam L, Xu A, et al. Barriers to telemedicine video visits for older adults in independent living facilities: mixed methods cross-sectional needs assessment. JMIR Aging 2022 Apr 19;5(2):e34326. [doi: 10.2196/34326] [Medline: 35438648]
43. Frishammar J, Essén A, Bergström F, Ekman T. Digital health platforms for the elderly? key adoption and usage barriers and ways to address them. Technol Forecast Soc Change 2023 Apr;189:122319. [doi: 10.1016/j.techfore.2023.122319]
44. Haimi M, Goren U, Grossman Z. Barriers and challenges to telemedicine usage among the elderly population in Israel in light of the COVID-19 era: a qualitative study. Digit Health 2024;10:20552076241240235. [doi: 10.1177/20552076241240235] [Medline: 38550265]
45. Landgren S, Cajander Å. Non-use of digital health consultations among Swedish elderly living in the countryside. Front Public Health 2021;9:588583. [doi: 10.3389/fpubh.2021.588583] [Medline: 34568247]

XSL•FO
RenderX

46. Ahmed N, Hall A, Poku B, McDermott J, Astbury J, Todd C. Experiences and views of older adults of South Asian, Black African, and Caribbean backgrounds about the digitalization of primary care services since the COVID-19 pandemic: qualitative focus group study. JMIR Form Res 2024 Dec 18;8:e57580. [doi: 10.2196/57580] [Medline: 39693146]
47. Ufholz K, Sheon A, Bhargava D, Rao G. Telemedicine preparedness among older adults with chronic illness: survey of primary care patients. JMIR Form Res 2022 Jul 27;6(7):e35028. [doi: 10.2196/35028] [Medline: 35896013]
48. Sproul A, Stevens J, Richard J. Older adults' use of and interest in technology and applications for health management: a survey study. Can J Hosp Pharm 2023;76(3):209-215. [doi: 10.4212/cjhp.3261] [Medline: 37409153]
49. Claudy MC, Garcia R, O'Driscoll A. Consumer resistance to innovation—a behavioral reasoning perspective. J of the Acad Mark Sci 2015 Jul;43(4):528-544. [doi: 10.1007/s11747-014-0399-0]
50. Cimperman M, Makovec Brenčič M, Trkman P. Analyzing older users' home telehealth services acceptance behavior-applying an extended UTAUT model. Int J Med Inform 2016 Jun;90:22-31. [doi: 10.1016/j.ijmedinf.2016.03.002] [Medline: 27103194]
51. Reinhardt R, Hietschold N, Gurtner S. Overcoming consumer resistance to innovations – an analysis of adoption triggers. R & D Management 2019 Mar;49(2):139-154 [FREE Full text] [doi: 10.1111/radm.12259]
52. Ammenwerth E. Technology acceptance models in health informatics: TAM and UTAUT. Stud Health Technol Inform 2019 Jul 30;263(264):64-71. [doi: 10.3233/SHTI190111] [Medline: 31411153]
53. Yu S, Chen T. Understanding older adults' acceptance of chatbots in healthcare delivery: an extended UTAUT model. Front Public Health 2024;12:1435329. [doi: 10.3389/fpubh.2024.1435329] [Medline: 39628811]
54. Badr J, Motulsky A, Denis JL. Digital health technologies and inequalities: a scoping review of potential impacts and policy recommendations. Health Policy 2024 Aug;146:105122. [doi: 10.1016/j.healthpol.2024.105122] [Medline: 38986333]
55. Ladds E, Khan M, Moore L, Kalin A, Greenhalgh T. The impact of remote care approaches on continuity in primary care: a mixed-studies systematic review. Br J Gen Pract 2023 May;73(730):e374-e383. [doi: 10.3399/BJGP.2022.0398] [Medline: 37105731]
56. Bevilacqua R, Di Rosa M, Riccardi GR, et al. Design and development of a scale for evaluating the acceptance of social robotics for older people: the robot era inventory. Front Neurorobot 2022;16:883106. [doi: 10.3389/fnbot.2022.883106] [Medline: 35874107]

Abbreviations:

IRT: innovation resistance theory
JBI: Joanna Briggs Institute
mHealth: mobile health
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension checklist
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
TAM: technology acceptance model
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by S Brini; submitted 07.Apr.2025; peer-reviewed by G Amabili, M Pinelli; revised version received 30.Dec.2025; accepted 30.Dec.2025; published 02.Feb.2026.

Please cite as:
Birati Y, Tzemah-Shahar R
Barriers to Digital Health Adoption in Older Adults: Scoping Review Informed by Innovation Resistance Theory
J Med Internet Res 2026;28:e75591
URL: https://www.jmir.org/2026/1/e75591
doi:10.2196/75591

© Yosefa Birati, Roy Tzemah-Shahar. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Machine Learning Techniques Used for the Identification of Sociodemographic Factors Associated With Cancer: Systematic Literature Review

Liz González-Infante[1,2*], MHRM; Gaston Marquez[2,3*], PhD; Solange Parra-Soto[2,4*], PhD; Mónica Cardona-Valencia[5*], PhD; Carla Taramasco[2,6*], PhD

[1]Facultad de Ciencias Empresariales, Universidad del Bío-Bío, Andrés Bello 720, Chillán, Chile

[2]Centro para la Prevención y el Control del Cáncer, Santiago, Chile

[3]Departamento de Ciencias de la Computación y Tecnologías de la Información, Facultad de Ciencias Empresariales, Universidad del Bío-Bío, Chillan, Chile

[4]Departamento de Nutrición y Salud Pública, Facultad Ciencias de la Salud y de los Alimentos, Universidad del Bío-Bío, Chillán, Chile

[5]Departamento Ciencias de la Rehabilitación en Salud, Facultad de Ciencias de la Salud y de los Alimentos, Universidad del Bío-Bío, Chillán, Chile

[6]ITISB, Facultad de Ingeniería, Universidad Andrés Bello, Viña del Mar, Chile

[*]all authors contributed equally

Corresponding Author:
Liz González-Infante, MHRM
Facultad de Ciencias Empresariales, Universidad del Bío-Bío, Andrés Bello 720, Chillán, Chile

Abstract

Background: Cancer remains one of the foremost global causes of mortality, with nearly 10 million deaths recorded by 2020. As incidence rates rise, there is a growing interest in leveraging machine learning (ML) to enhance prediction, diagnosis, and treatment strategies. Despite these advancements, insufficient attention has been directed toward the integration of sociodemographic variables, which are crucial determinants of health equity, into ML models in oncology.

Objective: This review aims to investigate how ML techniques have been used to identify patterns of predictive association between sociodemographic factors and cancer-related outcomes. Specifically, it seeks to map current research endeavors by detailing the types of algorithms used, the sociodemographic variables examined, and the validation methodologies used.

Methods: We conducted a systematic literature review in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. Searches were executed across 6 databases, focusing on the primary studies using ML to investigate the association between sociodemographic characteristics and cancer-related outcomes. The search strategy was informed by the PICO (population, intervention, comparison, and outcome) framework, and a set of predefined inclusion criteria was used to screen the studies. The methodological quality of each included paper was assessed.

Results: Out of the 328 records examined, 19 satisfied the inclusion criteria. The majority of studies used supervised ML techniques, with random forest and extreme gradient boosting being the most commonly used. Frequently analyzed variables include age, male or female or intersex, education level, income, and geographic location. Cross-validation is the predominant method for evaluating model performance. Nevertheless, the integration of clinical and sociodemographic data is limited, and efforts toward external validation are infrequent.

Conclusions: ML holds significant potential for discerning patterns associated with the social determinants of cancer. Nevertheless, research in this domain remains fragmented and inconsistent. Future investigations should prioritize the integration of contextual factors, enhance model transparency, and bolster external validation. These measures are crucial for the development of more equitable, generalizable, and actionable ML applications in cancer care.

(J Med Internet Res 2026;28:e79187) doi:10.2196/79187

KEYWORDS

cancer; health disparities; machine learning; predictive models; social determinants of health; sociodemographic factors; systematic review

XSL•FO
RenderX

Introduction

The use of machine learning (ML) in oncology has advanced significantly over the past decade, offering new opportunities for early detection, survival prediction, and treatment personalization. Models based on techniques such as random forests (RFs), extreme gradient boosting (XGBoost), and deep neural networks have demonstrated remarkable performance across different types of cancer, fueling enthusiasm for what has been termed digital precision oncology [1]. However, most of these applications rely almost exclusively on clinical and biomedical data, limiting their ability to capture the broader social and structural factors that shape health outcomes [2]. This gap raises important concerns, as it may compromise both the external validity and the equity of ML models. In this review, we consistently use the term sociodemographic factors to refer to variables such as age, male or female or intersex, educational attainment, income, ethnicity, rurality, and access to health care. These factors conceptually overlap with the broader category of social determinants of health (SDoH), but our focus is on those variables that are typically available in clinical and research datasets and are explicitly integrated into ML models. By doing so, we ensure clarity and terminological consistency throughout the paper.

Our review focuses on the most common sociodemographic variables in clinical and research datasets, such as age, male or female or intersex, education, income, and others, reflecting the current landscape of published ML studies rather than a deliberate theoretical choice. We recognize that these indicators only capture part of the social gradient influencing cancer outcomes. Therefore, we highlight the importance of future research integrating contextual and multilevel determinants, such as neighborhood characteristics, health care infrastructure, environmental exposures, and political factors, to promote an equity-centered approach to ML applications in oncology.

In parallel, the rise of explainable artificial intelligence (AI) has highlighted the importance of transparency and interpretability in clinical settings. Tools such as Shapley Additive Explanations and local interpretable model-agnostic explanations allow health care professionals to better understand ML models by identifying which variables are most relevant in predictions and how they interact with both clinical and sociodemographic factors [3]. These advances not only strengthen trust in ML-based systems but also enhance their potential for integration into clinical practice and public health policy [4]. The convergence of explainable AI and SDoH emerges as a promising pathway toward developing fairer and more actionable models.

Nevertheless, our review of the literature reveals that although research and reviews on ML in oncology are rapidly expanding, most have concentrated on methodological, genomic, or clinical aspects without adequately addressing sociodemographic factors. This omission limits the ability of the scientific community to develop robust guidelines for implementing models across diverse contexts and health systems. Against this backdrop, this study aimed to identify, characterize, and synthesize primary research that applied ML methods to analyze sociodemographic factors associated with cancer. The objective was to address both methodological and conceptual gaps while contributing to the development of fairer and more transparent models that can inform data-driven public health strategies. We present the results of a systematic literature review (SLR) examining how ML techniques have been used to identify and interpret sociodemographic factors in cancer-related studies. Of the 328 papers screened, 19 (5.8%) met the inclusion criteria. Rather than being a limitation, this number reflects the emerging nature of the field and highlights the value of conducting an early review to consolidate initial progress, make methodological and equity-related gaps more visible, and guide future research toward a stronger integration of sociodemographic factors in ML models applied to oncology.

Methods

Research Questions

Based on the main objective, we defined the following research questions:

1. What ML techniques have been applied in studies that analyze sociodemographic data of patients with cancer to identify factors associated with the disease?
2. What sociodemographic factors have been consistently identified as relevant to the diagnosis, progression, or treatment of cancer?

Identification

The SLR was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Checklist 1), which provide a rigorous framework for ensuring transparency and reproducibility in evidence synthesis [5]. To guide the construction of the search strategy, we also adopted the PICO (population, intervention, comparison, and outcome) model, as recommended by Petersen et al [6]. This framework allowed us to clearly define the target population, specify the type of intervention (ie, application of ML techniques), and focus the outcome on the identification of relevant sociodemographic factors associated with cancer (Table 1).

XSL•FO
RenderX

Table . Keywords used in the PICO (population, intervention, comparison, and outcome) structure.

Component	Description	Keywords
Population	Studies analyzing data from patients with cancer that include sociodemographic variables. These may encompass age, male or female or intersex, socioeconomic status, education, and residence among others.	"Sociodemographic factors," "social determinants," "sociodemographic characteristics," and "socio-demographic variables"
Intervention	Application of machine learning techniques to identify and analyze sociodemographic factors associated with cancer.	"Machine learning" and "artificial intelligence"
Comparison	No previous studies with similar scope and objectives were identified as suitable comparators. This review explores a novel approach.	Not applicable
Outcome	Identification of the most relevant sociodemographic variables associated with cancer outcomes, and assessment of the predictive performance of the applied machine learning models.	"Cancer," "oncology," variable importance, model accuracy, and AUC[a]

[a]AUC: area under the curve.

The search terms were combined using the Boolean operators AND and OR to ensure comprehensive retrieval of relevant literature. The final search string was as follows:

(["sociodemographic factors" OR "socio-demographic factors" OR "sociodemographic characteristics" OR "socio-demographic characteristics" OR "social determinants" OR "sociodemographic variables" OR "socio-demographic variables") AND ("machine learning" OR "artificial intelligence") AND ("cancer" OR "oncology"])

Screening

We conducted a comprehensive literature search across 6 major databases: PubMed (n=76), ACM Digital Library (n=85), ScienceDirect (n=7), IEEE Xplore (n=1), Web of Science Core Collection (n=80), and Scopus (n=79). Searches covered the period from database inception to October 14, 2024. PubMed was selected as the primary source for biomedical and oncology research. ScienceDirect was included to capture papers published in Elsevier journals not indexed elsewhere. ACM Digital Library and IEEE Xplore were used to retrieve computer science and engineering studies, where ML methods are often first reported. Web of Science facilitated interdisciplinary retrieval and citation tracking, while Scopus provided broad multidisciplinary coverage.

All records were exported, merged, and deduplicated prior to screening. To maximize comprehensiveness and minimize selection bias, we also applied forward and backward citation chasing on included studies. Full electronic search strategies for each database are provided in Multimedia Appendix 1.

Paper Selection

Eligibility Criteria

Primary studies were screened and selected based on predefined inclusion and exclusion criteria. The specific inclusion criteria applied are summarized in Textbox 1.

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Type of study: primary studies presenting original data or analysis. Quantitative studies applying machine learning techniques to analyze sociodemographic factors related to cancer, including experimental, observational (cohort, case-control, and cross-sectional), or methodological designs.
- Study area: application of machine learning in health, focused on the analysis of sociodemographic factors (eg, age, male or female or intersex, ethnicity, socioeconomic status, and health care access) and their association with any type of cancer (eg, breast, lung, prostate, and gastrointestinal).
- Machine learning techniques: use of supervised algorithms (eg, neural networks, decision trees, support vector machines, and logistic regression), unsupervised (eg, clustering), or semisupervised algorithms. Reporting of performance metrics such as accuracy, sensitivity, specificity, and receiver operating characteristic area under the curve.
- Sociodemographic factors: explicit analysis of sociodemographic variables related to cancer risk, prevalence, or progression, including age, male or female or intersex, ethnicity, income, education, occupation, geographic location, health care access, and other socioeconomic determinants.
- Publication period: studies published from 2014 onward.
- Language: publications in English or Spanish.
- Accessibility: full-text access or access to essential data and results enabling methodological evaluation.

Exclusion criteria

- Type of study: systematic reviews, narrative reviews, meta-analyses, or secondary studies.
- Study area: studies not analyzing the association between sociodemographic factors and cancer. Studies focused on other diseases (eg, diabetes and cardiovascular diseases).
- Machine learning techniques: studies relying solely on traditional statistical methods and not reporting model validation metrics.
- Sociodemographic factors: studies applying machine learning without including sociodemographic variables (eg, focused only on genetic, molecular, or biological data).
- Publication period: Studies published before 2014.
- Language: publications in other languages without available translation.
- Accessibility: abstracts or conference proceedings without access to the full paper.

Quality Assessment

The purpose of the quality assessment was to evaluate the relevance of each selected paper. Although quality assessment did not influence the selection of primary studies [7], we included it primarily to reflect the validity of the selected studies. Based on the response to each research question, we scored each paper with 2, 1, or 0 points. We then selected those papers that exceeded the 50% threshold. The studies chosen through this assessment ensure that our conclusions, drawn from the extracted data, are supported by adequately resourced evidence (Multimedia Appendix 1).

Study Selection and Resolution of Discrepancies

Each paper was independently screened by 2 reviewers according to predefined inclusion and exclusion criteria. Any disagreements regarding eligibility were addressed during consensus meetings, where reviewers jointly discussed the rationale for inclusion or exclusion. When consensus could not be reached, a third author was consulted to make the final decision. This procedure ensured transparency, reproducibility, and rigor throughout the study selection process.

Results

Overview

The SLR was conducted in accordance with the PRISMA guidelines, which provide a rigorous framework for ensuring transparency and reproducibility in evidence synthesis (Figure 1). Following the PRISMA methodology, a total of 15 primary studies published in peer-reviewed journals were identified. An additional 4 papers were included through forward snowballing, yielding a final sample of 19 studies. Among these, 58% (11/19) were conducted in the United States. Iran contributed 21% (4/19), followed by India with 11% (2/19), and South Korea with 5% (1/19). One study (5%) represented a collaborative effort between institutions in China and the United States (Table 2). The publication dates of the included studies ranged from 2018 to 2024. No eligible primary studies were found in workshop proceedings or book chapters.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of the selection of primary studies for the systematic literature review. N/A: not applicable.

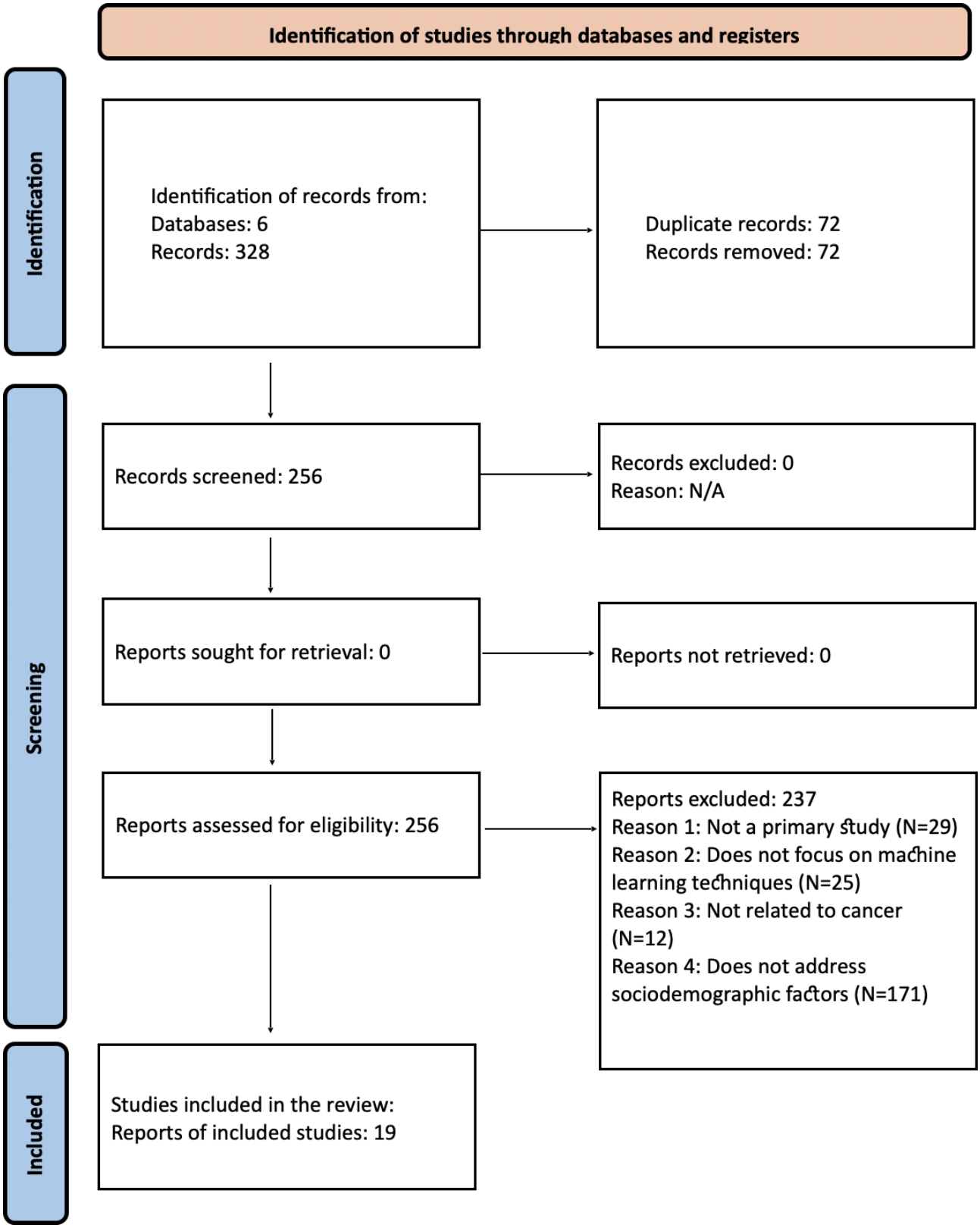

XSL•FO
RenderX

Table . Distribution of primary studies by country.

Country	Number of studies
United States	11
Iran	4
India	2
South Korea	1
China-US collaboration	1

Machine Learning Algorithms and Validation Strategies Reported

Across the studies analyzed, consistent patterns emerged in both the selection of ML algorithms and the validation methods used (Table 3).

XSL•FO
RenderX

Table . Summary of the machine learning algorithms and validation strategies reported across the 19 primary studies. Most studies applied ensemble methods such as random forest (RF) or gradient boosting, frequently combined with cross-validation schemes.

Study ID	Algorithms used	Validation strategy	Reference
S1	Lasso[a] LR[b], RF, gradient boosting, DT[c], SVM[d]	5-fold CV[e], ROC-AUC[f], accuracy, sensitivity, specificity	[8]
S2	XGBoost[g], LightGBM[h], CatBoost[i], RF, AdaBoost, Lasso regression	10-fold CV	[9]
S3	DT, RF	10-fold CV	[10]
S4	RF, artificial neural networks, bootstrap aggregating CART[j], XGBoost	10-fold CV	[11]
S5	XGBoost	10-fold CV	[12]
S6	LightGBM, XGBoost	10-fold CV	[13]
S7	RF, Neural networks, LR, XGBoost	CV, AUC, grid search	[14]
S8	RF, gradient boosting machine, SVM	5-fold CV, ROC	[15]
S9	Radiomics-signature model	No formal validation performed	[16]
S10	Multilayer perceptron, SVM, XGBoost	10-fold CV	[17]
S11	Max-p-regions, RF, Jenks natural breaks	RF VIMP[k] ranking	[18]
S12	CART, RF	Bootstrap sampling	[19]
S13	DT, RF, Boruta feature selection	Confusion matrix	[20]
S14	Bayesian additive, regression trees	Partial dependence plots, variable inclusion proportion	[21]
S15	LR, ridge classifier, SGD[l]classifier, KNN[m], DT, linear support vector classifier, support vector classifier with radial basis function kernel, Gaussian Naïve Bayes, AdaBoost classifier, RF, gradient boosting, QDA[n]	5-fold CV, LOOCV[o]	[22]
S16	Semiautomated segmentation + conditional LR	80/20 hold-out CV, ROC-AUC, Youden Index	[23]
S17	Random survival forest, Cox proportional hazards	Grid search, C-index[p]	[24]
S18	RF, SVM, gradient boosting machine	10-fold CV	[25]
S19	SVM, DT, naive Bayesian model, and KNN	10-fold CV	[26]

[a]Lasso: least absolute shrinkage and selection operator.
[b]LR: logistic regression.
[c]DT: decision tree.
[d]SVM: support vector machine.
[e]CV: cross-validation.
[f]ROC-AUC: receiver operating characteristic area under the curve.
[g]XGBoost: extreme gradient boosting.
[h]LightGBM: light gradient boosting machine.
[i]CatBoost: categorical boosting.
[j]CART: classification and regression tree.
[k]VIMP: variable importance.
[l]SGD: stochastic gradient descent.

XSL•FO
RenderX

[m]KNN: *K*-nearest neighbors.

[n]QDA: quadratic discriminant analysis.

[o]LOOCV: leave-one-out cross-validation.

[p]C-index: concordance index.

This review identified a wide array of ML algorithms applied to the analysis of sociodemographic and clinical data related to cancer. Each method presents distinct advantages and limitations, influencing its suitability depending on the specific research context and analytical goals. The most relevant algorithmic approaches are summarized below.

Tree-based methods, particularly RF, were the most frequently used, appearing in 13 of the included studies. RF is widely valued for its interpretability, robustness, and ability to process both categorical and continuous variables, making it especially well-suited to heterogeneous datasets.

Boosting techniques, such as XGBoost and light gradient boosting machine (LightGBM), featured prominently in studies aiming for high predictive accuracy. XGBoost, used in 7 studies, is noted for its computational efficiency and its capacity to manage imbalanced data, while LightGBM is often selected in contexts where large-scale data processing is prioritized.

A smaller subset of studies used Bayesian additive regression trees, which were particularly useful in modeling uncertainty and capturing complex non-linear associations. These features make Bayesian additive regression trees well-suited for analyzing disparities across ethnic and clinical subgroups.

Support vector machines (SVM) appeared in 5 studies and are recognized for their ability to handle high-dimensional data and to separate complex classes using nonlinear decision boundaries [27]. However, their performance is highly dependent on careful hyperparameter tuning, which can be challenging in the presence of large or noisy datasets [27]. Overall, SVM models remain a valuable choice for complex biomedical data when appropriately optimized and validated within diverse clinical contexts.

Artificial neural networks (ANNs) were applied in select studies and demonstrated strong performance in modeling nonlinear relationships and uncovering hidden patterns in complex datasets [28]. Despite their flexibility, the limited interpretability of ANNs often restricts their use in clinical contexts where transparency and explainability are required [28]. Their use, therefore, should be accompanied by complementary interpretability frameworks to ensure clinical reliability and trustworthiness.

Regression-based models, including the least absolute shrinkage and selection operator and ridge regression, were commonly used as baseline models or for feature selection. These methods are appreciated for their simplicity and interpretability, although they may underperform in settings involving nonlinear relationships or intricate interactions between variables [29]. Nevertheless, their transparency and ease of implementation make them a critical reference point for benchmarking more advanced ML models in oncology research.

Some studies also implemented bagged classification and regression tree models and ensemble methods such as stacking, reflecting a methodological interest in combining simplicity with predictive robustness. These strategies reduce model variance and enhance accuracy by integrating multiple base learners.

Overall, the analysis reveals a strong preference for tree-based algorithms, which offer an optimal balance between accuracy, interpretability, and adaptability to real-world clinical data. However, the choice of algorithm varied according to the nature of the dataset and the specific research objectives. More recent studies have increasingly adopted advanced methods such as boosting and neural networks, which provide enhanced predictive power but require greater expertise for interpretation and implementation.

Common Validation Methods

The reviewed studies showed a strong preference for cross-validation (CV) as the primary strategy to evaluate ML models applied to the identification of sociodemographic factors related to cancer. This approach is widely recognized for its ability to reduce overfitting and enhance the robustness of predictive performance. Several configurations of CV were used across studies, with 10-fold CV being the most commonly used. This method appeared in studies such as Dianati-Nasab et al [24], Stabellini et al [20], and Afrash et al [22], where it facilitated efficient partitioning of data into training and testing subsets, maximizing the use of available datasets.

In some cases, CV was complemented with repeated sampling to mitigate random variation and reinforce consistency. For instance, Wang et al [30] implemented repetitions alongside 10-fold CV to strengthen model reliability. A less frequently used configuration, 5-fold CV, was applied in studies like Kaushik et al [11], offering a computationally efficient alternative without substantially compromising model evaluation.

Several studies further enhanced reliability by incorporating multiple repetitions. A notable example is the work of He et al [9], who used 200 repetitions and evaluated model performance using metrics such as the concordance index and variable importance measures to ensure consistency and interpretability.

The choice of evaluation metrics reflected a balanced interest in both model discrimination and interpretability. The area under the receiver operating characteristic curve was one of the most frequently reported metrics, particularly valued for its ability to quantify discrimination capacity. It was prominently featured in studies such as Dehdar et al [19] and Niell et al [12]. Additionally, accuracy, sensitivity, and specificity were widely reported, especially in studies such as Galadima et al [25] and Lilhore et al [14], as they provided a detailed picture of false positive and false negative rates.

Some researchers adopted tailored interpretability metrics to better understand model behavior. For example, Niu et al [15] used variable inclusion proportions and partial dependence plots to explore the relative importance and marginal effect of

XSL•FO
RenderX

predictors, offering deeper insights into model mechanisms. Model optimization also played a critical role in the validation process. Techniques such as grid search were frequently used to fine-tune hyperparameters, as observed in the work of Dehdar et al [19]. In more specialized contexts, such as radiomics applications, validation using pretrained models was implemented, for example, in Dercle et al [21], focusing on metastatic colorectal cancer and highlighting the relevance of domain-specific strategies.

While most studies ensured strong internal validity, a common limitation was the lack of external validation. Although a few studies used unseen datasets or pretrained models to assess generalizability, the overall scarcity of external validation in heterogeneous populations restricts the broader applicability of findings. This underscores the importance of expanding validation practices to include more diverse datasets and real-world scenarios.

Analysis of Sociodemographic Variables

The reviewed studies demonstrate considerable variability in the types of sociodemographic variables incorporated into oncology research using ML techniques. Individual-level factors, such as age and male or female or intersex, were the most frequently included, underscoring their foundational role in the development and prognosis of various cancer types. For example, in breast cancer research, variables such as age at diagnosis and hormonal status appear consistently, as noted in the studies by Dianati-Nasab et al [24] and Niell et al [12]. Similarly, race and ethnicity were widely explored in studies addressing lung and colorectal cancer [9], highlighting disparities in health outcomes associated with these variables.

In addition to individual characteristics, several studies incorporated socioeconomic and access-related factors, which reflect broader SDH. Educational attainment and household income, often used as proxies for access to health resources and health-seeking behavior, featured prominently in studies on colorectal cancer [13] and advanced-stage breast cancer [13]. Other key access variables, such as transportation availability and type of health insurance, were also frequently considered to assess barriers to diagnosis and treatment, as shown in the works of Wang et al [30] and Afrash et al [22].

Some studies expanded their scope to include community- and environment-level variables, though these remain underrepresented overall. Galadima et al [25], for instance, investigated aspects of the built environment, such as crime rates and housing values, and their association with late-stage colorectal cancer diagnoses. Similarly, Dehdar et al [19] examined the influence of residence location, urban versus rural, on access to medical services, illustrating geographic disparities in health care delivery.

Regarding cancer types, breast cancer was the most frequently studied, followed by colorectal, lung, and gastric cancer. Research on breast cancer often focuses on the impact of delayed diagnosis and racial disparities, as seen in studies by Stabellini et al [20]. In contrast, studies on colorectal cancer emphasized socioeconomic factors and health care access, particularly in relation to late-stage detection [13,25]. Lung cancer studies primarily explored racial disparities and quality-of-life indicators in survival prediction [9,10].

A few studies adopted a broader, multicancer approach, examining sociodemographic patterns across different tumor types. For example, Stabellini et al [17] analyzed unplanned hospital readmissions in patients with solid tumors, integrating sociodemographic variables that have a direct influence on health outcomes. To provide a visual synthesis of these findings, Figure 2 presents a summary linking the ML algorithms used with the most frequently analyzed sociodemographic variables.

XSL•FO
RenderX

Figure 2. Association between machine learning techniques and sociodemographic variables. ANN: artificial neural network; BT: Bayesian tree; DT: decision tree; LASSO: least absolute shrinkage and selection operator; LightGBM: light gradient boosting machine; LR: logistic regression; RF: random forest; SVM: support vector machine; XGBoost: extreme gradient boosting.

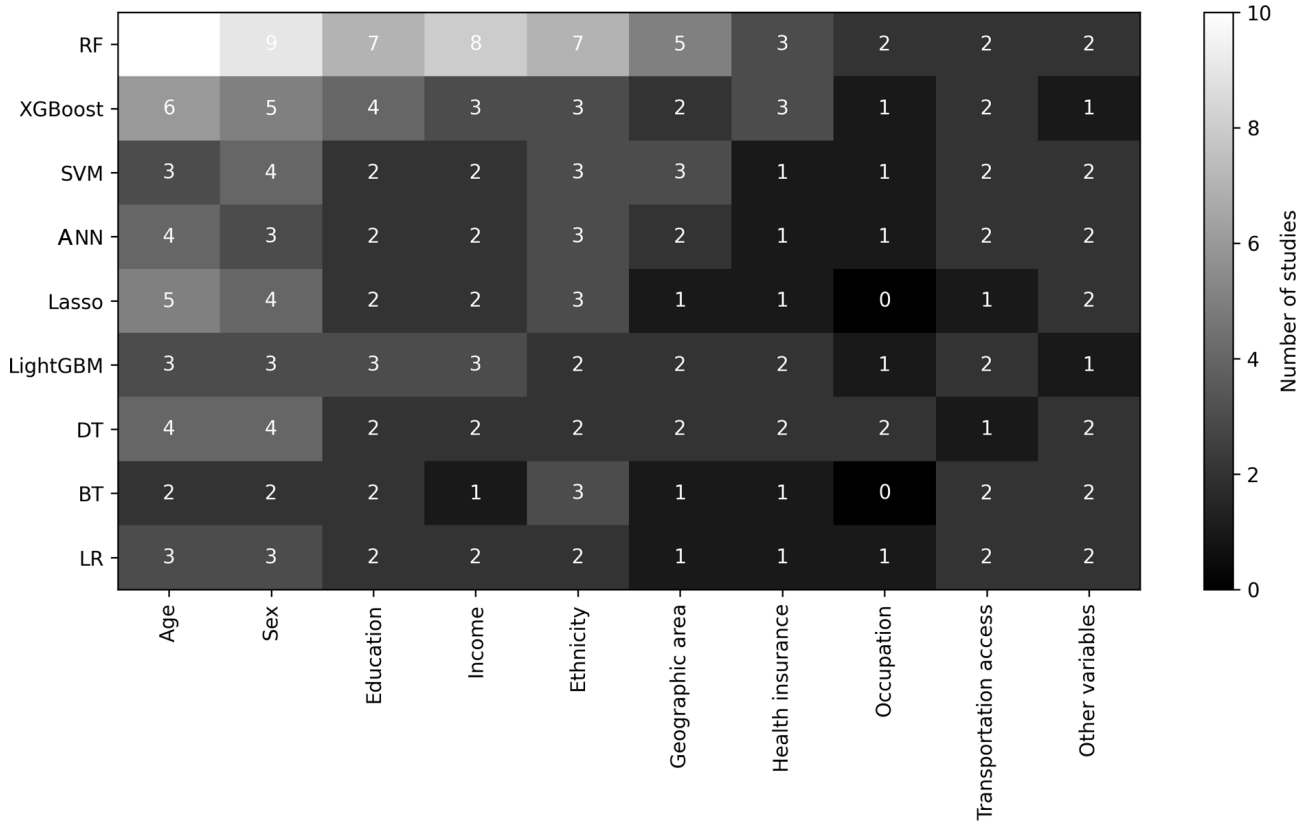

Discussion

Stratification of Findings

The reviewed studies confirm the potential of ML to identify patterns of predictive relevance of sociodemographic variables in relation to oncologic outcomes. However, the evidence remains fragmented and heterogeneous, with limited integration of contextual factors, reliance on predominantly internal validation, and little standardization in the reporting of performance and fairness. Overall, the findings suggest that ML can enhance risk stratification and the detection of disparities, but its real impact depends on methodological decisions that currently remain inconsistent.

In breast cancer, models most often prioritize age, race or ethnicity, and socioeconomic proxies to explain adverse events and late diagnosis. In colorectal cancer, income, insurance coverage, and geographic location are central for predicting advanced stage and survival. In lung cancer, studies more frequently explore ethnic disparities and quality-of-life measures associated with prognosis. This diversity suggests that the relevant set of SDoH is tumor-specific and linked to each care pathway.

Retrospective studies dominate; while they provide volume and feasibility, they limit causal inference and the ability to adapt to temporal social changes (eg, economic shocks, migration, or health system reforms). Prospective and longitudinal cohort designs would better capture the temporal variability of SDoH. Greater interpretative weight should be placed on studies with stronger control of confounding, explicit handling of missing data, subgroup analyses, and when available, external validation. In contrast, studies with incomplete reporting of variables and opaque pipelines should be viewed as exploratory signals rather than evidence ready for implementation.

Linking Inequities and ML Limitations

When sociodemographic factors are omitted or inconsistently defined, ML models often end up reflecting pre-existing inequities in access to and quality of care instead of uncovering or addressing them. This reflection of structural disparities undermines both the external validity and the generalizability of predictive models [31,32]. Evidence from recent reviews indicates that algorithmic bias in health care typically emerges from unbalanced data representation and the absence of systematic fairness assessments, highlighting the importance of transparency and interpretability in model design [33,34]. Although variable-importance analyses can reveal which sociodemographic features most influence predictions, they fall short of explaining underlying causal mechanisms. As Prosperi et al [35] and McCradden et al [36] emphasize, achieving fairness and accountability in ML-driven health applications requires methodological and ethical frameworks that move beyond conventional supervised learning. For this reason, throughout this review, the term “associated factors” is used exclusively in a predictive, not causal, sense.

To advance the field, it is essential to standardize the reporting of sociodemographic variables including age, male or female or intersex, race or ethnicity, education, income, rurality, and

XSL•FO
RenderX

health insurance as a minimum dataset to reduce heterogeneity and enable comparability across studies. Fairness metrics, such as demographic parity, equal opportunity, and subgroup calibration, should be applied alongside conventional measures like area under the curve and accuracy to explicitly assess model performance in vulnerable populations. Routine multicenter external validation is needed, testing models across diverse geographical and socioeconomic contexts. Incorporating neighborhood-level data (eg, area-level socioeconomic indices, transportation access, and housing conditions) can provide valuable context for individual predictors. Interdisciplinary collaboration between data scientists, oncologists, public health practitioners, and experts in social science and policy should be promoted to ensure that models achieve both technical precision and equity. Finally, transparent dissemination, including open-source code and model cards documenting limitations, is crucial to strengthen reproducibility and accountability.

Principal Findings

This systematic review synthesized evidence from 19 primary studies published between 2018 and 2024 that applied ML techniques to analyze sociodemographic factors associated with cancer. The analysis revealed consistent methodological patterns, frequently used variables, and prevalent validation strategies, while also identifying key implications for both academic research and professional practice.

From a methodological perspective, there was a strong preference for tree-based algorithms, particularly RF, which was the most frequently used due to its capacity to manage heterogeneous datasets while preserving a degree of interpretability. Boosting methods, notably XGBoost and LightGBM, were also prominent, especially in studies aiming for high predictive accuracy in high-dimensional or imbalanced data contexts. Less frequently, SVMs and ANNs were used to capture complex, nonlinear relationships, typically in specialized modeling scenarios. Regression-based approaches such as the least absolute shrinkage and selection operator and Ridge regression were primarily used for feature selection or as baseline models for comparative purposes.

Across the studies, a consistent set of core sociodemographic variables was identified. The most commonly included were age, male or female or intersex, educational level, income, ethnicity, and geographic location. These factors were primarily used to predict diagnostic timelines, disparities in access to treatment, and survival outcomes. However, only a limited number of studies incorporated broader structural or contextual variables—such as neighborhood characteristics, transportation access, or housing conditions—that could enrich model performance by capturing deeper dimensions of health inequity.

In terms of validation strategies, 10-fold CV was the most frequently implemented, followed by 5-fold validation in settings with limited computational resources. Most studies relied on standard evaluation metrics such as accuracy, area under the receiver operating characteristic curve, and sensitivity or specificity, reflecting a predominant focus on internal performance. However, the use of external validation with independent datasets was rare, limiting the generalizability of findings to broader, more diverse populations and real-world clinical environments.

From an applied perspective, the findings suggest that ML holds significant promise for identifying and quantifying structural health disparities in oncology. For the academic research community, this review highlights the importance of developing models that explicitly integrate SDoH, moving beyond individual-level data to encompass contextual and systemic influences. For clinicians and policymakers, predictive models incorporating sociodemographic factors offer a valuable complement to traditional clinical assessments, enabling the early identification of at-risk populations who might otherwise be overlooked.

Taken together, these findings underscore the transformative potential of ML when applied with methodological rigor, interpretability, and an explicit commitment to equity. Advancing this field will require not only continued technical innovation, but also interdisciplinary collaboration and a deliberate focus on addressing the social and structural dimensions of cancer prevention, diagnosis, and care.

Limitations

We critically assessed potential threats to the validity of our SLR based on the Wohlin classification, which provides clear guidelines for identifying and mitigating such threats [37].

Internal validity threats involve factors that could influence the reliability and accuracy of our study outcomes. A primary concern is selection bias, potentially stemming from limitations inherent in our search strategy and inclusion criteria. To minimize this risk, we carefully defined explicit and rigorous inclusion and exclusion criteria, conducting systematic searches across multiple reputable academic databases. Despite these measures, the relatively small final sample size (N=19) remains a limitation. To further reinforce internal validity, we conducted independent cross-checking and reviews with three domain experts, ensuring consistency and reliability in the selection and evaluation of studies.

External validity threats refer to the generalizability of our findings beyond the specific studies reviewed. A significant concern here is the representativeness of the primary studies regarding the broader application of ML to sociodemographic determinants of cancer. To mitigate this threat, we engaged external experts in data science and public health to provide critical insights and feedback on our findings, enhancing the relevance and applicability across different contexts [7].

Finally, construct validity threats pertain to the accurate interpretation and generalization of results in alignment with the study objectives. The primary concern here is potential subjectivity or bias in interpreting the findings. To address this, external collaborators participated in the analysis and classification phases, providing independent perspectives that strengthened the robustness and objectivity of our conclusions.

Comparison With Prior Work

Several systematic reviews have examined the application of ML techniques in oncology, but their scope differs significantly from this study. Adeoye et al [38] evaluated ML models in

XSL•FO
RenderX

oncology settings with limited resources, identifying gaps in external validation and clinical adoption, but without providing a detailed analysis of sociodemographic variables. Hossain Raju et al [26] reviewed the use of deep learning for breast cancer risk prediction, focusing mainly on imaging and genomic data. Kumar et al [39] offered a broad overview of AI in oncology, emphasizing technical innovation rather than social determinants. Zeinali et al [40] analyzed the application of ML in predicting cancer-related symptoms, again with a focus on clinical variables.

In addition, recent editorials and reviews have highlighted the need to move toward more interpretable and explainable models. For example, Hrinivich et al [4] warned about the risks associated with the lack of interpretability in ML models in oncology, noting that reliance on opaque systems may amplify biases and weaken clinical trust. However, while these works underscore the importance of technical transparency, they do not systematically address the incorporation of sociodemographic factors into predictive cancer models.

Our review differs from previous contributions in three main ways. First, we provide a systematic synthesis of primary studies in which sociodemographic factors are explicitly integrated into ML models applied to oncological outcomes, thereby moving beyond an exclusively clinical or technical lens. Second, we critically assess methodological limitations—such as the lack of external validation, limited interpretability, and absence of fairness metrics—specifically in relation to the inclusion of sociodemographic data. Third, we connect these findings to broader discussions of equity and public health, emphasizing that neglecting social determinants may inadvertently reinforce inequalities in cancer care. By placing sociodemographic factors at the center rather than at the periphery, this review addresses an underexplored yet essential dimension of the field.

Ultimately, our findings contribute meaningfully to the growing body of literature by illustrating how ML can be leveraged to deepen our understanding of social inequalities in cancer outcomes. Rather than treating sociodemographic variables as peripheral, this study brings them to the forefront of analysis, offering a more nuanced view of how structural and contextual factors shape cancer risk, access to care, and treatment outcomes. These insights can help guide the development of more inclusive health policies and inform interventions that are responsive to the realities of diverse and historically underserved populations.

Conclusions

This review indicates that the integration of sociodemographic factors into ML models for oncology is still an emerging field, with a modest evidence base that appears to be steadily growing. Only 19 primary studies met our inclusion criteria, yet their collective findings point to the potential benefits of embedding these variables within predictive frameworks. There is some evidence to suggest that explicitly accounting for sociodemographic factors could refine predictive accuracy and fairness, although these associations remain noncausal. That said, such conclusions remain tentative, as further research is needed to substantiate these observations. Looking ahead, researchers might prioritize enhancing the transparency of these models, exploring fairness metrics, and considering how such tools align with the broader goals of health policy. Advancing these aspects could prove vital in ensuring that ML supports both precision oncology and equitable public health outcomes. It is worth noting that, although the variables examined in this review are those most frequently reported in existing datasets, future research could benefit from incorporating contextual and structural determinants to strengthen both fairness and interpretability in ML-based cancer studies (Multimedia Appendices 2 and 3).

Acknowledgments

This study was conducted as part of the Doctoral Program in Economics and Information Management at the Universidad del Bío-Bío. The authors also acknowledge the support of the Center for Cancer Control and Prevention. Additionally, all the authors thank the Ñuble Health Hub of the Universidad del Bío-Bío for their valuable support. The authors did not use generative artificial intelligence technologies (such as ChatGPT or similar tools) in the generation of this manuscript.

Funding

This research was supported by the Agencia Nacional de Investigación y Desarrollo through the Fondo de Financiamiento de Centros de Investigación en Áreas Prioritarias (grant 152220002).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Quality assessment criteria and the assignment of scores.
[DOCX File, 16 KB - jmir_v28i1e79187_app1.docx]

Multimedia Appendix 2
Primary studies by year and publication type.
[PNG File, 323 KB - jmir_v28i1e79187_app2.png]

XSL•FO
RenderX

Multimedia Appendix 3
Primary studies description.
[DOCX File, 20 KB - jmir_v28i1e79187_app3.docx]

Checklist 1
PRISMA 2020 checklist.
[PDF File, 133 KB - jmir_v28i1e79187_app4.pdf]

References

1. Zhang J, Lu Y, Zhang N, et al. Global burden of female breast cancer and its association with socioeconomic development status, 1990-2044. Cancer Rep (Hoboken) 2023 Sep;6(Suppl 1):e1827. [doi: 10.1002/cnr2.1827] [Medline: 37095062]
2. Fountzilas E, Pearce T, Baysal MA, Chakraborty A, Tsimberidou AM. Convergence of evolving artificial intelligence and machine learning techniques in precision oncology. NPJ Digit Med 2025 Jan 31;8(1):75. [doi: 10.1038/s41746-025-01471-y] [Medline: 39890986]
3. Alelyani T, Alshammari MM, Almuhanna A, Asan O. Explainable artificial intelligence in quantifying breast cancer factors: Saudi Arabia context. Healthcare (Basel) 2024 May 15;12(10):1025. [doi: 10.3390/healthcare12101025] [Medline: 38786433]
4. Hrinivich WT, Wang T, Wang C. Editorial: Interpretable and explainable machine learning models in oncology. Front Oncol 2023;13:1184428. [doi: 10.3389/fonc.2023.1184428] [Medline: 37035194]
5. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097. [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
6. Petersen K, Vakkalanka S, Kuzniarz L. Guidelines for conducting systematic mapping studies in software engineering: an update. Inf Softw Technol 2015 Aug;64:1-18. [doi: 10.1016/j.infsof.2015.03.007]
7. Kitchenham B. Procedures for performing systematic reviews (technical report TR/SE-0401). : Keele University; 2004 URL: https://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf [accessed 2026-01-20]
8. Mortezagholi A, Khosravizadeh O, Menhaj MB, Shafigh Y, Kalhor R. Make intelligent of gastric cancer diagnosis error in Qazvin's medical centers: using data mining method. Asian Pac J Cancer Prev 2019 Sep 1;20(9):2607-2610. [doi: 10.31557/APJCP.2019.20.9.2607] [Medline: 31554353]
9. He J, Zhang JX, Chen CT, et al. The relative importance of clinical and socio-demographic variables in prognostic prediction in non-small cell lung cancer: a variable importance approach. Med Care 2020 May;58(5):461-467. [doi: 10.1097/MLR.0000000000001288] [Medline: 31985586]
10. Sim JA, Kim YA, Kim JH, et al. The major effects of health-related quality of life on 5-year survival prediction among lung cancer survivors: applications of machine learning. Sci Rep 2020 Jul 1;10(1):10693. [doi: 10.1038/s41598-020-67604-3] [Medline: 32612283]
11. Kaushik M, Joshi RC, Kushwah AS, et al. Cytokine gene variants and socio-demographic characteristics as predictors of cervical cancer: a machine learning approach. Comput Biol Med 2021 Jul;134:104559. [doi: 10.1016/j.compbiomed.2021.104559] [Medline: 34147008]
12. Niell BL, Abdalah M, Stringfield O, et al. Quantitative measures of background parenchymal enhancement predict breast cancer risk. AJR Am J Roentgenol 2021 Jul;217(1):64-75. [doi: 10.2214/AJR.20.23804] [Medline: 32876474]
13. Dong W, Bensken WP, Kim U, Rose J, Berger NA, Koroukian SM. Phenotype discovery and geographic disparities of late-stage breast cancer diagnosis across U.S. counties: a machine learning approach. Cancer Epidemiol Biomarkers Prev 2022 Jan;31(1):66-76. [doi: 10.1158/1055-9965.EPI-21-0838] [Medline: 34697059]
14. Lilhore UK, Poongodi M, Kaur A, et al. Hybrid model for detection of cervical cancer using causal analysis and machine learning techniques. Comput Math Methods Med 2022;2022:4688327. [doi: 10.1155/2022/4688327] [Medline: 35572826]
15. Niu L, Hu L, Li Y, Liu B. Correlates of cancer prevalence across census tracts in the United States: a Bayesian machine learning approach. Spat Spatiotemporal Epidemiol 2022 Aug;42:100522. [doi: 10.1016/j.sste.2022.100522] [Medline: 35934328]
16. Stabellini N, Dmukauskas M, Bittencourt MS, et al. Social determinants of health and racial disparities in cardiac events in breast cancer. J Natl Compr Canc Netw 2023 Jul;21(7):705-714. [doi: 10.6004/jnccn.2023.7023] [Medline: 37433439]
17. Stabellini N, Nazha A, Agrawal N, et al. Thirty-day unplanned hospital readmissions in patients with cancer and the impact of social determinants of health: a machine learning approach. JCO Clin Cancer Inform 2023 Jul;7:e2200143. [doi: 10.1200/CCI.22.00143] [Medline: 37463363]
18. Stone A, Kalahiki C, Li L, Hubig N, Iuricich F, Dunn H. Evaluation of breast tumor morphologies from African American and Caucasian patients. Comput Struct Biotechnol J 2023;21:3459-3465. [doi: 10.1016/j.csbj.2023.06.019] [Medline: 38213888]
19. Dehdar S, Salimifard K, Mohammadi R, et al. Applications of different machine learning approaches in prediction of breast cancer diagnosis delay. Front Oncol 2023;13:1103369. [doi: 10.3389/fonc.2023.1103369] [Medline: 36874113]

XSL•FO
RenderX

20. Stabellini N, Cullen J, Moore JX, et al. Social determinants of health data improve the prediction of cardiac outcomes in females with breast cancer. Cancers (Basel) 2023 Sep 19;15(18):4630. [doi: 10.3390/cancers15184630] [Medline: 37760599]
21. Dercle L, Yang M, Gönen M, et al. Ethnic diversity in treatment response for colorectal cancer: proof of concept for radiomics-driven enrichment trials. Eur Radiol 2023 Dec;33(12):9254-9261. [doi: 10.1007/s00330-023-09862-z] [Medline: 37368111]
22. Afrash MR, Shafiee M, Kazemi-Arpanahi H. Establishing machine learning models to predict the early risk of gastric cancer based on lifestyle factors. BMC Gastroenterol 2023 Jan 10;23(1):6. [doi: 10.1186/s12876-022-02626-x] [Medline: 36627564]
23. Dong W, Kim U, Rose J, et al. Geographic variation and risk factor association of early versus late onset colorectal cancer. Cancers (Basel) 2023 Feb 4;15(4):1006. [doi: 10.3390/cancers15041006] [Medline: 36831350]
24. Dianati-Nasab M, Salimifard K, Mohammadi R, et al. Machine learning algorithms to uncover risk factors of breast cancer: insights from a large case-control study. Front Oncol 2024;13:1276232. [doi: 10.3389/fonc.2023.1276232] [Medline: 38425674]
25. Galadima H, Anson-Dwamena R, Johnson A, Bello G, Adunlin G, Blando J. Machine learning as a tool for early detection: a focus on late-stage colorectal cancer across socioeconomic spectrums. Cancers (Basel) 2024 Jan 26;16(3):540. [doi: 10.3390/cancers16030540] [Medline: 38339293]
26. Raju MAH, Imam T, Islam J, Al Rakin A, Nayyem MN, Uddin MS. An ontological framework for lung carcinoma prognostication via sophisticated stacking and synthetic minority oversampling techniques. Presented at: 2024 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob); Nov 28-30, 2024; Bali, Indonesia p. 125-130. [doi: 10.1109/APWiMob64015.2024.10792946]
27. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition: Springer; 2009. URL: https://hastie.su.domains/ElemStatLearn/ [accessed 2025-01-20]
28. Goodfellow I, Bengio Y, Courville A. Deep Learning: MIT Press; 2016. URL: https://www.deeplearningbook.org/ [accessed 2026-01-20]
29. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol 2005 Apr 1;67(2):301-320. [doi: 10.1111/j.1467-9868.2005.00503.x]
30. Wang Z, Kim Y, Barbosa EJM. Demographics and socioeconomic determinants of health predict continued participation in a CT lung cancer screening program. Curr Probl Diagn Radiol 2024;53(5):552-559. [doi: 10.1067/j.cpradiol.2024.04.004] [Medline: 38658287]
31. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018 Dec 18;169(12):866-872. [doi: 10.7326/M18-1990] [Medline: 30508424]
32. Colacci M, Huang YQ, Postill G, et al. Sociodemographic bias in clinical machine learning models: a scoping review of algorithmic bias instances and mechanisms. J Clin Epidemiol 2025 Feb;178:111606. [doi: 10.1016/j.jclinepi.2024.111606] [Medline: 39532254]
33. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. JAMA 2019 Dec 24;322(24):2377-2378. [doi: 10.1001/jama.2019.18058] [Medline: 31755905]
34. Ning Y, Li S, Ng YY, et al. Variable importance analysis with interpretable machine learning for fair risk prediction. PLOS Digit Health 2024 Jul;3(7):e0000542. [doi: 10.1371/journal.pdig.0000542] [Medline: 38995879]
35. Prosperi M, Guo Y, Sperrin M, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nat Mach Intell 2020;2(7):369-375. [doi: 10.1038/s42256-020-0197-y]
36. McCradden MD, Joshi S, Mazwi M, Anderson JA. Ethical limitations of algorithmic fairness solutions in health care machine learning. Lancet Digit Health 2020 May;2(5):e221-e223. [doi: 10.1016/S2589-7500(20)30065-0] [Medline: 33328054]
37. Kitchenham B, Charters S. Guidelines for performing systematic literature reviews in software engineering. : Keele University and University of Durham; 2007 URL: https://legacyfileshare.elsevier.com/promis_misc/525444systematicreviewsguide.pdf [accessed 2025-01-20]
38. Adeoye J, Akinshipo A, Koohi-Moghadam M, Thomson P, Su YX. Construction of machine learning-based models for cancer outcomes in low and lower-middle income countries: a scoping review. Front Oncol 2022;12:976168. [doi: 10.3389/fonc.2022.976168] [Medline: 36531037]
39. Kumar Y, Gupta S, Singla R, Hu YC. A systematic review of artificial intelligence techniques in cancer prediction and diagnosis. Arch Comput Methods Eng 2022;29(4):2043-2070. [doi: 10.1007/s11831-021-09648-w] [Medline: 34602811]
40. Zeinali N, Youn N, Albashayreh A, Fan W, Gilbertson White S. Machine learning approaches to predict symptoms in people with cancer: systematic review. JMIR Cancer 2024 Mar 19;10:e52322. [doi: 10.2196/52322] [Medline: 38502171]

Abbreviations

AI: artificial intelligence
ANN: artificial neural network
LightGBM: light gradient boosting machine

XSL•FO
RenderX

ML: machine learning
PICO: population, intervention, comparison, and outcome
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RF: random forest
SDoH: social determinants of health
SLR: systematic literature review
SVM: support vector machine
XGBoost: extreme gradient boosting

Edited by J Sarvestan; submitted 17.Jun.2025; peer-reviewed by JCL Chow, SM Shaffi; revised version received 18.Oct.2025; accepted 30.Oct.2025; published 28.Jan.2026.

Please cite as:
González-Infante L, Marquez G, Parra-Soto S, Cardona-Valencia M, Taramasco C
Machine Learning Techniques Used for the Identification of Sociodemographic Factors Associated With Cancer: Systematic Literature Review
J Med Internet Res 2026;28:e79187
URL: https://www.jmir.org/2026/1/e79187
doi:10.2196/79187

© Liz González-Infante, Gaston Marquez, Solange Parra, Mónica Cardona, Carla Taramasco. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Phases of Living Evidence Synthesis Using AI: Living Evidence Synthesis (Version 1)

Xuping Song[1,2,3,4], PhD; Zhenjie Lian[1,2], MSc; Rui Wang[2,5], MSc; Ruixin Li[1,2], MSc; Zhenzhen Yang[1,2], MSc; Xufei Luo[4,6], PhD; Lei Feng[1], BSc; Zhiming Ma[1], BSc; Zhen Pu[1], BSc; Qi Wang[7], PhD; Long Ge[1,2,3,4], PhD; Caihong Li[8], PhD; Yaolong Chen[1,3,4,6], PhD; Kehu Yang[1,2,3,4], MSc; John Lavis[9], PhD

[1]School of Public Health, Lanzhou University, No. 222 South Tianshui Road, Lanzhou, Lanzhou, Gansu, China
[2]The Centre of Evidence-based Social Science, Lanzhou University, Lanzhou, Gansu, China
[3]Key Laboratory of Evidence Based Medicine & Knowledge Translation of Gansu Province, Lanzhou, Gansu, China
[4]WHO Collaborating Centre for Guideline Implementation and Knowledge Translation, Lanzhou, Gansu, China
[5]Dingxi Center for Disease Control and Prevention, Dingxi, Gansu, China
[6]Evidence-Based Medicine Center, School of Basic Medicine, Lanzhou University, Lanzhou, Gansu, China
[7]School of Public Health, University of Hong Kong, Hong Kong, China
[8]School of Information Science and Engineering, Lanzhou University, Lanzhou, Gansu, China
[9]Department of Health Research Methods, Evidence, and Impact, McMaster Health Forum, McMaster University, Hamilton, ON, Canada

Corresponding Author:
Kehu Yang, MSc
School of Public Health, Lanzhou University, No. 222 South Tianshui Road, Lanzhou, Lanzhou, Gansu, China

Abstract

Background: Living evidence (LE) synthesis refers to the method of continuously updating systematic evidence reviews to incorporate new evidence. It has emerged to address the limitations of the traditional systematic review process, particularly the absence of or delays in publication updates. The emergence of COVID-19 accelerated the progress in the field of LE synthesis, and currently, the applications of artificial intelligence (AI) in LE synthesis are expanding rapidly. However, in which phases of LE synthesis should AI be used remains an unanswered question.

Objective: This study aims to (1) document the phases of LE synthesis where AI is used and (2) investigate whether AI improves the efficiency, accuracy, or utility of LE synthesis.

Methods: We searched Web of Science, PubMed, the Cochrane Library, Epistemonikos, the Campbell Library, IEEE Xplore, medRxiv, COVID-19 Evidence Network to support Decision-making, and McMaster Health Forum. We used Covidence to facilitate the monthly screening and extraction processes to maintain the LE synthesis process. Studies that used or developed AI or semiautomated tools in the phases of LE synthesis were included.

Results: A total of 24 studies were included, including 17 on LE syntheses, with 4 involving tool development, and 7 on living meta-analyses, with 3 involving tool development. First, a total of 34 AI or semiautomated tools were involved, comprising 12 AI tools and 22 semiautomated tools. The most frequently used AI or semiautomated tools were machine learning classifiers (n=5) and the Living Interactive Evidence synthesis platform (n=3). Second, 20 AI or semiautomated tools were used for the data extraction or collection and risk of bias assessment phase, and only 1 AI tool was used for the publication update phase. Third, 3 studies demonstrated the improvement in efficiency achieved based on time, workload, and conflict rate metrics. Nine studies applied AI or semiautomated tools in LE synthesis, obtaining a mean recall rate of 96.24%, and 6 studies achieved a mean F_1-score of 92.17%. Additionally, 8 studies reported precision values ranging from 0.2% to 100%.

Conclusions: AI and semiautomated tools primarily facilitate data extraction or collection and risk of bias assessment. The use of AI or semiautomated tools in LE synthesis improves efficiency, leading to high accuracy, recall, and F_1-scores, while precision varies across tools.

Trial Registration: OSF Registries 87tp4; https://osf.io/4fvdq/overview

(J Med Internet Res 2026;28:e76130) doi:10.2196/76130

KEYWORDS

accuracy; artificial intelligence; efficiency; living evidence synthesis; phases; semiautomated tools; utility

XSL•FO
RenderX

Introduction

Evidence synthesis refers to an approach where data across studies are identified and combined to gain a clearer understanding of a body of research [1]. There is typically a significant gap between the time when a search is performed and the time when the results are published, often exceeding a year [2]. Furthermore, only a limited number of reviews are updated once they have been published [3]. This process can result in missing evidence, potentially affecting the accuracy of the findings. The approach of living evidence (LE) synthesis has been developed to address this challenge.

The method of constantly updating a systematic synthesis of evidence to incorporate newly available evidence is known as LE [4]. Elliott et al [5] developed the basis of the LE model in 2014, which effectively incorporates and summarizes new evidence. The LE synthesis process includes 4 phases: database searching and eligibility assessment, data extraction or collection and risk of bias assessment, synthesis and analysis, and publication update [6]. It has also been adapted in areas such as network meta-analysis and guidelines. The onset of COVID-19 increased the incentive to use LE [7]. Unlike traditional evidence synthesis, which requires the redeployment of significant resources for updates, the maintenance of an LE synthesis can require more modest resources [8]. However, LE synthesis that focuses on evolving topics may have a reduced reliability compared to traditional evidence synthesis. The incorporation of artificial intelligence (AI) techniques has the potential to enhance the reliability of LE synthesis by, for example, leveraging advanced algorithms to continuously assess and filter the most relevant and high-quality evidence [9].

The field of AI, which encompasses machine learning, deep learning, natural language processing, data mining, image recognition, and computer vision, to name a few, has the potential to enhance the efficiency of LE synthesis [10,11]. In 2013, Adams et al [11] indicated that leveraging AI to automate the LE synthesis procedures could simplify the regular updating and maintenance of evidence. The development of AI systems, particularly AI based on large language models (LLMs), such as the generative pretrained transformer, has significantly advanced natural generative language systems [12]. Various AI-driven tools have been developed for different phases of LE synthesis, such as crowdsourcing and task-sharing platforms like HDAS [13]. However, the performance of the AI techniques and the phases of LE synthesis where AI is used remain unclear.

Overall, the objectives of this review are (1) to conduct a review analyzing the phases of LE synthesis that use AI and (2) to explore whether AI can improve the efficiency, accuracy, or utility of LE synthesis.

Methods

This is the first version of an LE synthesis. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 statement for living systematic reviews (PRISMA-LSR; Checklist 1) was used as a guide for reporting this LE synthesis [14]. The review has been registered in the Open Science Forum [15].

Search Strategy

We systematically searched the Web of Science, PubMed, the Cochrane Library, Epistemonikos, the Campbell Library, IEEE Xplore, medRxiv, COVID-19 Evidence Network to support Decision-making, and McMaster Health Forum for publications up to April 2, 2025. The details of the search strategy used can be found in Table S1 in Multimedia Appendix 1. We subscribed to the Web of Science, PubMed, the Cochrane Library, the Campbell Library, and IEEE Xplore for monthly dynamic updates and used Covidence to facilitate the screening and extraction processes for maintaining an LE synthesis. We plan to conduct living updates for a 12-month period (from April 2025 to April 2026). The final update is scheduled for April 2, 2026, after which we will assess whether to retire the living mode based on the following established triggers: (1) evidence on “the AI application in LE synthesis” has reached conclusiveness, (2) the topic no longer holds decision-making value for the field, (3) no new eligible studies emerge during the 12-month update period, or (4) subsequent resource or funding support is unavailable [16,17].

Inclusion and Exclusion Criteria

First, the LE synthesis includes living systematic review, living meta-analysis, living network meta-analysis, living guideline, living scoping review, living overview, living umbrella review, and living mapping. In this review, the types of included studies were classified into 2 categories based primarily on whether a meta-analysis had been performed. These categories include the LE synthesis (without a meta-analysis) and living meta-analysis (with a meta-analysis conducted).

Second, the criteria for inclusion in this review are studies that use AI or semiautomated tools in the following phases of LE synthesis: (1) database searching and eligibility assessment, (2) data extraction or collection and risk of bias assessment, (3) synthesis and analysis, or (4) publication update [6]. The LE syntheses from any field were included. In addition, studies that developed AI or semiautomated tools for LE synthesis were also included. Textbox 1 provides further details.

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria for the study.

Inclusion criteria

- The studies using artificial intelligence (AI) or semiautomated tools in the following phases of living evidence (LE) synthesis: (1) database searching and eligibility assessment, (2) data extraction or collection and risk of bias assessment, (3) synthesis and analysis, or (4) publication update. A study can be any type of LE synthesis in any field, including but not limited to all scientific journals in the social sciences.
- Studies that developed AI or semiautomated tools for LE synthesis.

Exclusion criteria

- Studies that did not document the use of AI or semiautomated tools in LE synthesis.
- Protocol, commentaries, editorials, letters to the editor, and updating studies.

We excluded studies that did not document the use of AI or semiautomated tools in LE synthesis. In addition, protocols, commentaries, editorials, letters to the editor, and updating studies were also excluded, as shown in Textbox 1.

Third, AI tools are characterized by autonomous learning and end-to-end decision-making. They enable the independent execution of data collection, feature extraction, model training, and inference and generate output results without any human intervention. However, semiautomated tools incorporate human review or decision support at critical stages, using a "machine assistance and human oversight" collaborative paradigm [18,19]. Textbox 2 shows the types of AI or semiautomated tools, where AI or semiautomated tools were categorized by the application phases. First, the first segment of the AI or semiautomated tools for each phase is sourced from Bendersky et al [13]. Second, the subsequent segment is derived from the work of Khalil et al [20]. Third, for the final segment, AI or semiautomated tools were identified and summarized from relevant studies using a manual search. The AI techniques based on LLMs, such as the generative pretrained transformer, were also included.

Textbox 2. Artificial intelligence (AI) or semiautomated tools used in the 4 phases of living evidence (LE) synthesis.

Phase 1. Database searching and eligibility assessment

- Segment 1.1: Automatic, continuous database search with push notification, database aggregators (such as HDAS, Epistemonikos), notification from clinical trial registries, randomized clinical trial classifier, text mining technologies, and automatic retrieval of full-text papers
- Segment 1.2: RCT tagger, LitSuggest, Evidence mapping tool, SRA-Polyglot Search Translator, QuickClinical, HDAS, ROBOTsearch, SRA-word, frequency analyzer, The Search Refiner, Sherlock, SRA De-duplicate, Distiller, R package-rev tools, Rayyan, EPPI-reviewer, Abstrackr, SRA helper, LibSVM classifier, Bibot, Active Screener, RobotAnalyst, Swift-Review, Evidence Pipeline, JBI Sumari, EndNote, SARA, eSuRFr, ParsCit, and Citation searcher
- Segment 1.3: Natural language processing–assisted abstract screening tool, automatic text classifiers supported by deep learning–based language models, machine learning classifiers, Cochrane Crowd, Living Interactive Evidence (LIvE) synthesis platform, Cochrane RCT classifier, OpenAlex, Risklick AI, Bayesian classifier, Generative Pretrained Transformer models, and RobotReviewer LIVE

Phase 2. Data extraction or collection and risk of bias assessment

- Segment 2.1: Machine learning information-extraction systems, automated structured data extraction tools for PDFs, machine learning–assisted RoB tool, data repositories, and linked data
- Segment 2.2: RobotReviewer, DistelleR, JBI Sumari, in-house data extraction tool written in R, statistical package R, ExaCT, Revman, Raptor, ContentMine, Graph2Data, and Evidence mapping tool
- Segment 2.3: BioMart, MetaInsight COVID-19, LIvE synthesis platform, Open Science Framework (OSF), PsychOpen CAMA, and Generative Pretrained Transformer models

Phase 3. Synthesis and analysis

- Segment 3.1: Structured data extraction tools, which automatically provide data in a suitable format for statistical analysis; continuous analysis updating based on availability of structured extracted data; and statistical surveillance of key analysis results, with threshold set for potential conclusion change
- Segment 3.2: MetaPreg, MetaXL, NetMetaXL, Meta-analyst, Webplotdigitizer, Evidence mapping tool, PRISMA flow diagram generator, Evidence mapping tool, R package-rev tools
- Segment 3.3: Risklick AI, Web Source Processing Pipeline, LIvE synthesis platform, and generative pretrained transformer models

Phase 4. Publication update

- Segment 4.1: Templated reporting of some report items, automatic text generation tools for synthesis and writing, automatization in the identification of changes between LSR versions for peer review, and editorial process (such as Archie)
- Segment 4.2: Trial2rev, RevManHAL, DistelleR, SRA replicant writer, SRA-RevMan Replicant, and JBI Sumari
- Segment 4.3: Generative pretrained transformer models

XSL•FO
RenderX

Study Screening and Data Collection

Two reviewers independently screened the titles and abstracts of all selected studies, followed by a full-text review. Any disagreements regarding selection were resolved by a third researcher. Data were extracted using a predesigned Microsoft Excel sheet. Two reviewers independently extracted data from all included studies, including information such as title, first author, journal, year of publication, LE synthesis type, types of tool or technology, types of AI or semiautomated tools, phases of LE synthesis, outcomes, and so forth. Any disagreements were resolved by a third researcher. During data extraction, representative outcomes (such as means or ranges) were prioritized for synthesis, with the range of values considered subsequently when outcomes were similarly representative.

Methodological Quality Assessment

Given the lack of a standardized tool for assessing the methodological quality of AI-related studies, the 24 studies were categorized into 3 types by methodological characteristics and primary objective (diagnostic test, tool development, or—when neither applied—a general synthesis) and assessed for methodological quality using the modified version of the Quality Assessment of Diagnostic Accuracy Studies version 2 (QUADAS-2) tool, Joanna Briggs Institute (JBI) Critical Appraisal Checklist for Textual Evidence: Narrative, and AMSTAR 2 tool. First, 10 studies were assessed with the modified version of the QUADAS-2 tool: these studies specifically assessed the application of AI in the database searching and eligibility assessment phase, which aligns with a diagnostic test accuracy (DTA) framework. We adopted the modified version of the QUADAS-2 proposed by Rashid et al [21-23]. As QUADAS-2 is designed for DTA research contexts, this framework was only applicable to those studies where one of the objectives included the application of AI in the database searching and eligibility assessment phase [21,24,25]. The core elements of QUADAS-2 were revised to adapt it to AI-related research scenarios, as follows: "patient" was replaced with "study," "index test" with "AI," "reference standard" with "comparator," and "case-control design" with "DTA framework." We also constructed a 2×2 table, categorizing studies into "included" or "excluded" based on both "AI screening results" and "reference/original systematic review (SR) screening results," with counts denoted as a, b, c, and d, respectively. The details of the modified QUADAS-2 are provided in Table S2 in Multimedia Appendix 1. Second, 5 studies, which specifically developed AI or semiautomated tools for LE synthesis without DTA-related accuracy evaluation and were not designed as LE synthesis themselves, were assessed using the JBI Critical Appraisal Checklist for Textual Evidence: Narrative [26]. Third, 9 studies, which were designed as LE syntheses without DTA-related accuracy evaluation and not primarily focused on AI or semiautomated tool development (or tool development was only an auxiliary means), were assessed using the AMSTAR 2 tool [27,28]. The details are shown in Tables S3 and S4 in Multimedia Appendix 1. All of the included studies were evaluated independently by 2 reviewers (RL and ZY), and disagreement was resolved by a third reviewer (ZL). The LE synthesis did not involve a statistical combination of results (meta-analysis), as its aims were to document the phases of LE synthesis where AI is used and to investigate whether AI improves the efficiency, accuracy, or utility of LE synthesis. Therefore, several systematic review procedures—including sensitivity analyses, reporting bias assessment, certainty assessment, and investigations of heterogeneity—were not used.

Data Analysis

This review conducted 3 complementary analyses, as shown in Figure 1.

Figure 1. Road map for the use of artificial intelligence (AI): applications and extractable clinical outcomes across 4 phases of living evidence synthesis. LE: living evidence.

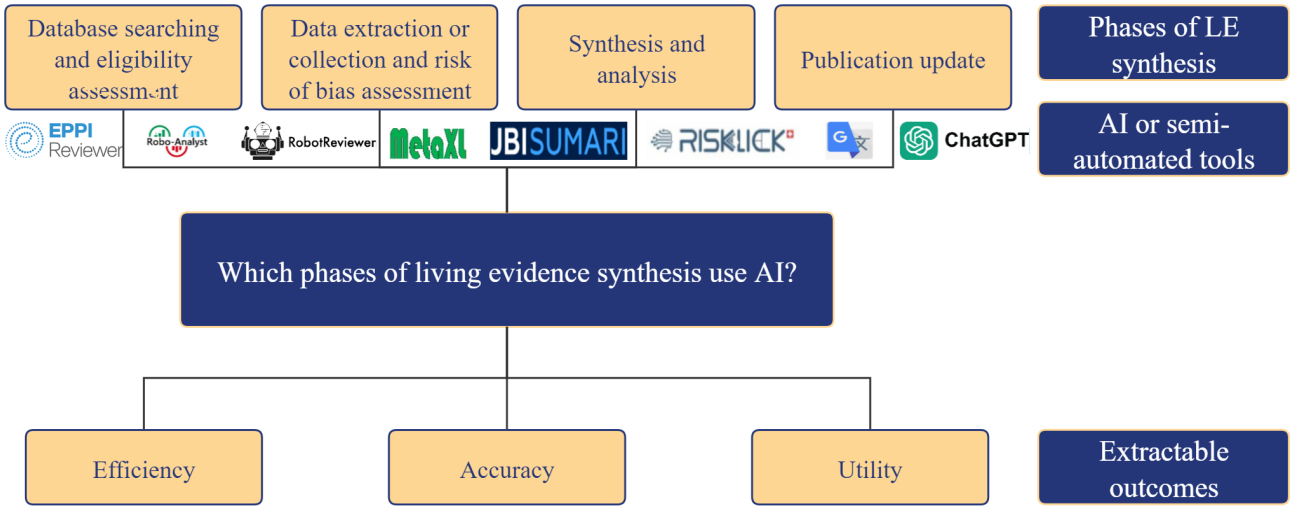

Analysis 1: Phases of LE Synthesis Utilizing AI or Semiautomated Tools

We analyzed the prevalence and distribution of AI or semiautomated tools across 4 phases of LE synthesis. Phase 1 is database searching and eligibility assessment. This process includes going through the databases, retrieving the results, importing them into the citation management software, removing any duplicate results, and assessing their eligibility individually. Phase 2 is data extraction or collection and risk of bias

XSL•FO
RenderX

assessment; once the eligibility of studies has been verified and they have been included in the review process, it becomes crucial to systematically extract and collect information about their main characteristics and results. Additionally, it is very important to assess the risk of bias associated with the conduct and methodology used in the studies. In phase 3—synthesis and analysis—the data that have been assessed to conform to the criteria are integrated, and the data are analyzed. In phase 4—publication update—after going through the aforementioned phases 1-3, sections of a review are generated based on their results, and conclusions are updated.

Analysis 2: AI or Semiautomated Tools Used in LE Synthesis

First, the types of AI or semiautomated tools applied in each LE synthesis phase were investigated. Second, the frequency of AI or semiautomated tools applied in the LE synthesis was analyzed.

Analysis 3: Primary Outcomes Investigating AI or Semiautomated Tools in LE Synthesis

The impact of applied AI or semiautomated tools in LE synthesis was analyzed across 3 outcomes [29]. First, efficiency, defined as the relationship between the time required to complete a workload and the workload itself, was evaluated to determine whether either the duration or workload was reduced with the use of AI or semiautomated tools. This outcome may be described as time reduction, workload reduction, and conflict rates with and without the tool.

Second, accuracy is used to assess performance with and without AI or semiautomated tools. It may be described as accuracy, recall, precision, F_1-score, area under the receiver operating characteristic curve, number needed to read, and study relevance. In addition, we calculated the overall mean recall and mean F_1-score using the following formula:

$$\bar{M}=1N \sum i=1NMi$$

where Mi is the representative value for study i, defined as the reported single value, if provided, or the midpoint of the reported range [L, U], calculated as (L+U)/2, if a range was provided. N is the number of studies reporting that metric [30,31].

Third, utility is used to assess whether user decisions align with those of AI or semiautomated tools, including user consistency, user satisfaction, perceived ease of use, and study quality.

Results

Search Results

Out of 9180 studies, 24 studies applied AI or semiautomated tools in LE synthesis, including 17 LE syntheses (4 developing tools) and 7 living meta-analyses (3 developing tools), as shown in Figure 2 [29,32-54]. In addition, 8 studies exclusively applied AI tools in LE synthesis, 11 studies exclusively applied semiautomated tools, and 5 studies utilized both AI and semiautomated tools. The basic characteristics of the included studies are shown in Table S5 in Multimedia Appendix 1. The details of the studies excluded at the full-text eligibility stage with reasons are shown in Table S6 in Multimedia Appendix 1 [5,9,55-75].

Figure 2. Database search flow diagram. LE: living evidence.

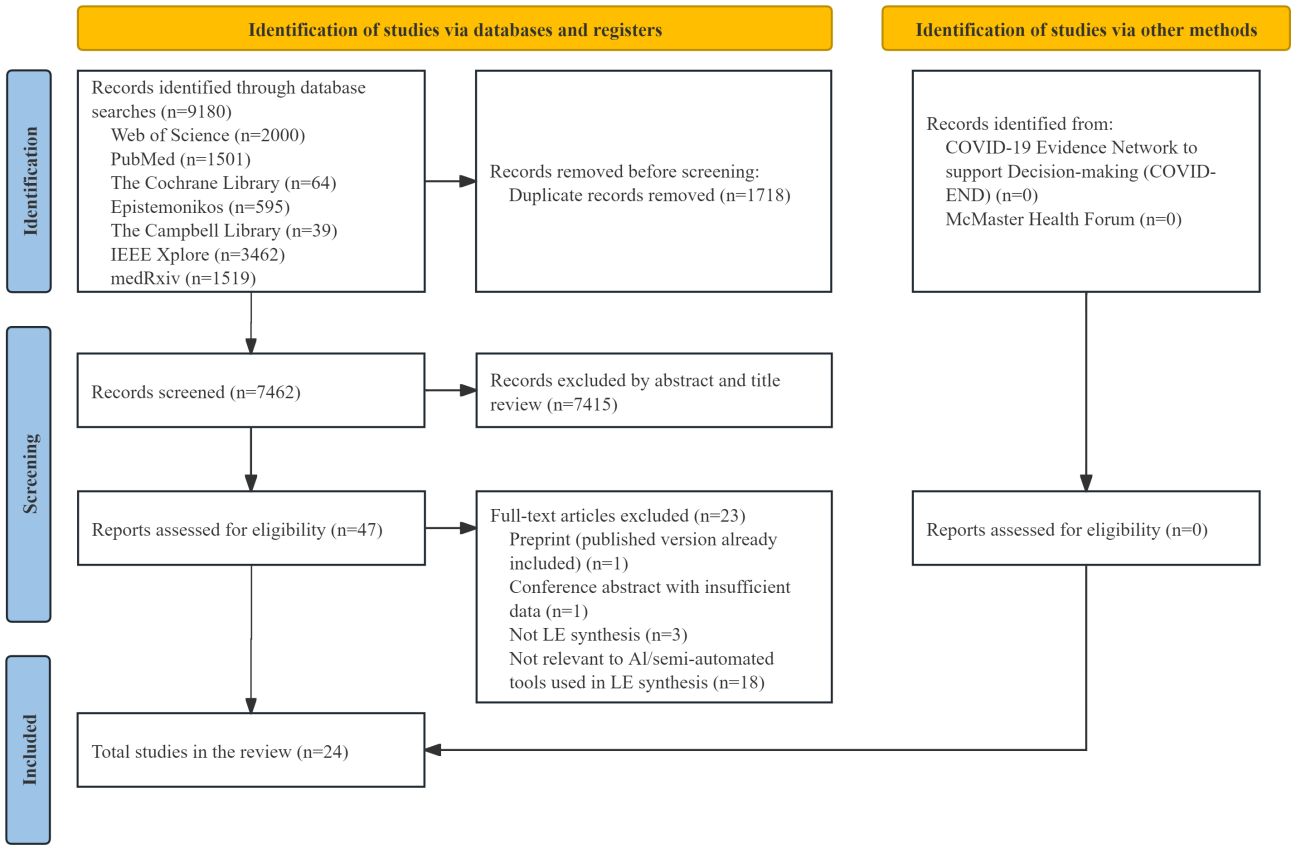

XSL•FO
RenderX

Methodological Quality of Included Studies

We conducted a methodological quality assessment of 10 studies using a revised QUADAS-2 tool within the DTA framework [29,32,35,36,42-44,51,52,54]. All studies were assessed as low-risk in the "Study selection," "Index test (AI)," and "Reference (comparator)" domains. While none of the studies specified the time interval between the task execution of AI and comparator-based analysis, all were determined as low-risk in the "Flow and timing" domain. Additionally, we did not identify any applicability concerns, as all studies were classified as low-risk in the "Applicability" domain (Table 1). Five studies were subjected to methodological quality assessment using the JBI Critical Appraisal Checklist for Textual Evidence: Narrative [41,46,48-50]. Four studies obtained a score of 5/6, with a narrative appraisal of "Exclude" owing to failure to meet the narrative classification criterion [41,46,48,49]. One study achieved a full score of 6/6 and was thus appraised as "Include" (Table S7 in Multimedia Appendix 1) [50]. In addition, we conducted a methodological quality assessment of 9 studies using AMSTAR 2 [33,34,37-40,45,47,53]. The methodological quality scores of the included studies ranged from 11 to 15. Overall, the methodological quality of eight studies [34,37-40,45,47,53] was rated as moderate, while only 1 study [33] was rated as low in methodological quality. The most common limitation was that the authors failed to provide a list of excluded studies (Table S8 in Multimedia Appendix 1).

Table . Summary of modified Quality Assessment of Diagnostic Accuracy Studies version 2 (QUADAS-2) assessments for studies using artificial intelligence (AI) or semiautomated tools in the database searching and eligibility phase of the living evidence (LE) synthesis process.

Author, year	Risk of bias				Applicability concern		
	Study selection	Index test (AI)	Reference (comparator)	Flow and timing	Study selection	Index test (AI)	Reference (comparator)
Knafou et al [32] (2023)	Low	Low	Low	Low	Low	Low	Low
Perlman-Arrow et al [29] (2023)	Low	Low	Low	Low	Low	Low	Low
Chou et al [35] (2020)	Low	Low	Low	Low	Low	Low	Low
Kamso et al [36] (2023)	Low	Low	Low	Low	Low	Low	Low
Marshall et al [42] (2023)	Low	Low	Low	Low	Low	Low	Low
Haas et al [43] (2021)	Low	Low	Low	Low	Low	Low	Low
Vaghela et al [44] (2021)	Low	Low	Low	Low	Low	Low	Low
Shemilt et al [51] (2024)	Low	Low	Low	Low	Low	Low	Low
Le-Khac et al [52] (2024)	Low	Low	Low	Low	Low	Low	Low
Hair et al [54] (2024)	Low	Low	Low	Low	Low	Low	Low

Types and Frequency of AI or Semiautomated Tools in LE Synthesis

A total of 34 AI or semiautomated tools were involved, including 12 (35.3%) AI tools and 22 (64.7%) semiautomated tools, as shown in Multimedia Appendix 2. The most frequently used AI or semiautomated tools were machine learning classifiers (n=5), followed by the Living Interactive Evidence (LIvE) synthesis platform (n=3), AD-SOLES (n=2), Covidence (n=2), and MAGICapp (n=2).

Phases of AI or Semiautomated Tools Application in LE Synthesis

There were 18 AI or semiautomated tools for database searching and eligibility assessment, 20 for data extraction or collection and risk of bias assessment, and 10 for synthesis and analysis. However, only 1 AI tool was used for publication updates. Out of all the tools, RobotReviewer LIVE can be used for all phases of LE synthesis, as shown in Textbox 3.

XSL•FO
RenderX

Textbox 3. Types of artificial intelligence (AI) or semiautomated tools applications in the 4 phases of living evidence (LE) synthesis.

Phase 1. Database searching and eligibility assessment

- LIvE platform, automatic text classifiers, machine learning ensemble classifier, Natural language processing–assisted abstract screening tool, machine learning classifiers, machine learning, PICO annotators, STAR tool, AD-SOLES, Covidence, rcrossref, openalexR, RISmed, RobotReviewer LIVE, Risklick AI, metaCOVID application, supervised text classification models, and text mining techniques

Phase 2. Data extraction or collection and risk of bias assessment

- LIvE platform, web-based interactive app, open-source living systematic review application, Covidence, AD-SOLES, Google Refine tool, script, REDASA, RobotReviewer LIVE, Risklick AI, Metainsight COVID-19, metaCOVID application, information extraction techniques, EndNote, semiautomated model, supervised text classification models, text mining techniques, GPT-4-turbo, Claude-3-Opus, and EPPI-Reviewer

Phase 3. Synthesis and analysis

- LIvE platform, MAGICapp, Trial sequential analysis (TSA) software, AD-SOLES, ODDPub, RobotReviewer LIVE, script, Metainsight COVID-19, metaCOVID application, and Dynameta

Phase 4. Publication update

- RobotReviewer LIVE

Impact of AI or Semiautomated Tools on LE Synthesis

Overview

A total of 10 (41.7%) studies reported on the impact of AI or semiautomated tools on LE synthesis in terms of efficiency, accuracy, or utility in the database searching and eligibility phase or the data extraction or collection and risk of bias assessment phase. Table 2 provides a description of the outcome metrics in the included studies.

XSL•FO
RenderX

Table . Summary of the indicator terms for outcome metrics in the included studies.

Metrics	Explanation
Efficiency	
Time	AI[a] or semiautomated tools were used to save time. Only 2 (8.3%) studies reported on time saving [29,35]. Specifically, Perlman-Arrow et al [29] reported a 45.9% reduction in screening time per abstract in the database searching and eligibility phase. Chou et al [35] estimated the time saving ranged from 2.0 to 13.2 hours in the database searching and eligibility phase.
Workload	Two (8.3%) studies reported on workload metrics related to the use of AI or semiautomated tools [29,42]. Perlman-Arrow et al [29] reported that the semiautomated tool completed 68% of the workload in the database searching and eligibility phase. Marshall et al [42] found that manual screening had an efficiency rate of 23% in obtaining 31 abstracts, whereas AI achieved a rate of 55%, demonstrating an efficiency improvement of approximately 140% in the database searching and eligibility phase.
Conflict rates with and without the tool	The efficiency of abstract screening decreases as the number of conflicting votes increases [29]. Perlman-Arrow et al [29] reported a reduction in conflict rates from 8.32% to 3.64% with the use of semiautomated tool in the database searching and eligibility phase.
Accuracy[b]	
Precision	Precision refers to the ratio of accurately categorized documents among all the documents that the model assigns to a particular class [32]. Eight (33.3%) studies reported on precision [29,32,35,42,43,51,53,54]. 1. Khan et al [53] reported a precision rate of even 100% using AI in the data extraction or collection and risk of bias assessment phase. 2. Perlman-Arrow et al [29] and Haas et al [43] reported precision rates of 92.10% and 96.07%, respectively, using AI or semiautomated tools in the database searching and eligibility phase. 3. Hair et al [54] reported that the average precision rate using AI is about 84.5% in the database searching and eligibility phase. 4. Shemilt et al [51] reported a precision rate of 50% - 86% using AI in the database searching and eligibility phase. 5. Marshall et al [42] reported a precision rate of 55% using AI in the database searching and eligibility phase. 6. Knafou et al [32] reported a precision rate of only 29.69% using AI in the database searching and eligibility phase. 7. However, Chou et al [35] reported a precision rate of only 0.2% - 8% using AI in the database searching and eligibility phase.
Recall[c]	Recall (also known as sensitivity) refers to the fraction of positive documents that have been accurately identified among all documents for the specified class [32]. Nine (37.5%) studies reported on recall [29,32,35,36,42,43,51,53,54]. All studies reported recall rates in excess of 87%. The average value was about 96.24%. 1. Perlman-Arrow et al [29], Chou et al [35], and Marshall et al [42] found recall rates of even 100% using AI or semiautomated tools in the database searching and eligibility phase. 2. Knafou et al [32], Haas et al [43], and Kamso et al [36] reported a recall rate of 89%, 99.25% and 99.3%, respectively, using AI in the database searching and eligibility phase. 3. Shemilt et al [51] reported a recall rate of 94% - 99% using AI in the database searching and eligibility phase. 4. Khan et al [53] reported a recall rate of 92% - 96% using AI in the data extraction or collection and risk of bias assessment phase. 5. Hair et al [54] reported that the average sensitivity rate using AI is about 95.1% in the database searching and eligibility phase.

XSL•FO
RenderX

Metrics	Explanation
F_1-score[c]	F_1-score refers to the balanced harmonic average between the model precision and recall [32]. Six (25%) studies reported on F_1-score [29,32,43,52-54]. All studies reported F_1-score between 80.47% and 99% after using AI. The average value was about 92.17%. 1. Knafou et al [32], Perlman-Arrow et al [29], and Haas et al [43] reported an F_1-score of 89.2%, 92.6%, and 97.59%, respectively, using AI or semiautomated tools in the database searching and eligibility phase. 2. Le-Khac et al [52] reported an F_1-score of 87% using AI in the data extraction or collection and risk of bias assessment phase. 3. Khan et al [53] reported F_1-scores between 96% and 98% after using AI in the data extraction or collection and risk of bias assessment phase. 4. Hair et al [54] reported that the average F_1-score using AI is about 89.6% in the database searching and eligibility phase.
Area under the receiver operating characteristic curve (AUC-ROC)	AUC-ROC calculates the area under the curve between the true positive rate and the false positive rate [32]. Knafou et al [32] reported higher AUC-ROC performance using AI in the database searching and eligibility phase and had an AUC-ROC performance of 94.25% - 94.77%.
Number needed to read (NNR)	NNR refers to the total number of literature considered within the search divided by the number of literature included from the search [35]. Only 2 (8.3%) studies reported on NNR [29,35]. Perlman-Arrow et al [29] reported an NNR between 1.086 and 1.125 after using a semiautomated tool in the database searching and eligibility phase. Chou et al [35] reported an NNR between 15 and 100 after using AI in the database searching and eligibility phase.
Article relevance	Vaghela et al [44] reported on studies included after searching using AI, and 50.49% were considered relevant to the query in the database searching and eligibility phase.
Utility	
User satisfaction	Perlman-Arrow et al [29] reported that the average satisfaction of users with the tool reached 4.2/5 in the database searching and eligibility phase.
Consistency	Kamso et al [36] reported that consistency in the use of AI between 2 reviewers was assessed using percentage agreement and Kappa scores, revealing a range of percentage agreement from 79.0% to 96.0%, and a variation in Kappa scores from moderate (0.40) to substantial (0.63) in the database searching and eligibility phase.
Article quality	Vaghela et al [44] reported that 64.53% of the included studies possess reliable quality in the database searching and eligibility phase.

[a]AI: artificial intelligence.

[b]Kamso et al [36] achieved an accuracy ranging from 75.9% to 96.9% in research classification using AI in the database searching and eligibility phase. Khan et al [53] reported that the collaborative large language models' accuracy, based on concordant responses in the prompt set, reached 99% in the data extraction or collection and risk of bias assessment phase.

[c]The overall mean recall (96.24%) and F_1-score (92.17%) are the simple averages of study-level values from Table S5 in Multimedia Appendix 1. For studies reporting a range, the midpoint was used as the study-level value.

Efficiency Enhancements Through AI or Semiautomated Tools in LE Synthesis

Three studies showed improved efficiency in the database searching and eligibility phase in terms of 3 indicator terms. A total of 2 (8.3%) studies [29,35] reported on time saving with AI or semiautomated tools, 2 (8.3%) studies [29,42] reported on workload metrics related to the use of AI or semiautomated tools, and 1 study [29] reported a reduction in conflict rates with the use of semiautomated tool, which consequently increases the efficiency.

Accuracy Improvements With AI or Semiautomated Tools in LE Synthesis

A total of 9 and 6 studies that applied AI or semiautomated tools in LE synthesis reported a mean recall rate and a mean F_1-score of 96.24% and 92.17%, respectively. While Khan et al [53] reported a precision rate of even 100% achieved using AI in the data extraction or collection and risk of bias assessment phase. However, in 7 studies, the reported precision rates varied significantly, ranging from 0.2% to 96.07% in the database searching and eligibility phase.

XSL•FO
RenderX

Utility of AI or Semiautomated Tools in LE Synthesis

Three studies reported on the utility of AI or semiautomated tools in the database searching and eligibility phase of LE synthesis, including user satisfaction, consistency, and study quality. Consistency in the use of AI between 2 reviewers was assessed using percentage agreement and Kappa scores [36].

Discussion

Principal Findings

AI or semiautomated tools are actively used to facilitate the process of LE synthesis. We conducted this review to identify the phases of LE synthesis that use AI and explore whether AI can improve the efficiency, accuracy, or utility of LE synthesis.

AI or semiautomated tools have been increasingly used in LE synthesis, particularly in living systematic review. This review discovered that AI or semiautomated tools are most commonly used for data extraction or collection and risk of bias assessment. However, only a few studies have addressed the use of AI or semiautomated systems for publication updates, highlighting the need for further development in this phase.

Diverse types of AI or semiautomated tools were identified in this study. These include the LIvE synthesis platform, AD-SOLES, metaCOVID application, and RobotReviewer LIVE, which are utilized in multiple phases of LE synthesis, indicating their versatility and potential for wider adoption [37,39,40,42,47,54]. The most frequently used AI or semiautomated tools were machine learning classifiers, the LIvE synthesis platform, Covidence, AD-SOLES, and MAGICapp. Furthermore, the rapid rise of AI tools involving LLM types, such as GPT-4-turbo and Claude-3-Opus, has led to their use in LE synthesis. These tools can be suitable for application in multiple or even all phases of LE synthesis, especially in the publication update phase. The application of LLMs to further enhance the efficiency, accuracy, and utility of LE synthesis remains a key focus for researchers and practitioners.

Governments worldwide, particularly those in leading AI nations such as China, the United States, Germany, the United Kingdom, France, and Canada, are especially emphasizing the transformative impact of AI on research and decision-making processes [76,77]. Funding from various sources, including the Economic and Social Research Council, reflects a strong financial commitment to advancing AI technologies in evidence synthesis. Furthermore, a growing number of AI guidance and organizations are emerging to embrace the opportunity that AI has taken in producing LE synthesis. For example, Responsible AI in Evidence SynthEsis has provided recommendations for the main roles of responsible AI in the evidence synthesis ecosystem that are involved in responsible AI use [78]. Furthermore, organizations such as ALIVE aim to improve societal outcomes by producing and utilizing timely, trustworthy, and affordable evidence.

Challenges remain in the application of AI in LE synthesis. Machine learning classifiers suffer from low precision and varying efficiency across different topics [35]. As an example, RobotReviewer LIVE faces challenges in performance variability for complex reviews, limited study types, and data source constraints [42]. Therefore, further research aimed at enhancing the adaptability and stability of AI across various research areas is urgently needed. In addition, ethical issues, data protection measures, and transparency in AI-driven LE synthesis are also key challenges that need to be addressed [79]. At the ethical level, AI is prone to selection bias due to the skewness of its training data, which impairs the inclusivity of evidence, and the mechanism of responsibility attribution remains unclear [80]. Data protection is another area that faces challenges, as research data required for AI training often contain sensitive information, and existing anonymization technologies cannot fully avoid the risk of privacy breaches [81]. Cost considerations in the implementation of AI tools, including initial investment, ongoing operational costs, training expenses, and requirements for hardware and software resources also constitute a significant issue [82].

Policymaking involves judgment, making it more of an art than a science, whereas science is primarily driven by evidence and shapes evidence-informed policymaking [83]. Study has indicated that relying solely on systematic reviews for policymaking is far from sufficient; instead, policymakers need to obtain a more diverse range of synthesized evidence to underpin decision-making [84]. The LE synthesis, especially by incorporating AI into evidence production, can deliver updated evidence to facilitate evidence-informed policymaking. AI could revolutionize policymaking by facilitating ongoing assessments, ensuring that the policies remain aligned with the latest evidence and evolve in response to new information as it emerges [2,5,85]. Furthermore, AI enables policymakers to continuously monitor and assess policies throughout their lifecycle, which allows adaptation to shifting circumstances and evolving societal needs in real time [86]. Furthermore, the advancement of AI capabilities, particularly through LLMs, adds a deeper analytical layer; LLMs can provide nuanced insights and help predict future research directions relevant to policymaking [87]. The application of AI in LE synthesis could transform policy decision-making, advancing policy formulation for policymakers.

Recent advances in AI provide researchers with new transformative capabilities [79]. Van Dijk et al [88] indicated that AI tools are a promising innovation in the current practice of systematic evaluation, and researchers have reported positive experiences with these tools. The use of AI enhances efficiency by significantly reducing researchers' time and workload [2,89]. Manion et al [90] indicated that natural language processing could enhance accuracy and reduce errors through a "human-in-the-loop" approach. The application of AI in LE synthesis has considerably benefited researchers, significantly enhancing their research capabilities.

This LE synthesis will retain its living mode beyond the present publication, consistent with the methodology. This decision is based on two key considerations: (1) the predefined retirement triggers have not been triggered and (2) the Safe and Responsible Use of AI Working Group (Working Group 3) and the Methods & Process Innovation Working Group (Working Group 4) of the Evidence Synthesis Infrastructure Collaborative will benefit from the continuous updates from this LE synthesis to support their future research initiatives [91-94].

XSL•FO
RenderX

Future Research Directions

In the above discussion, we have suggested the advancement of future work across multiple dimensions. From a technical point of view, efforts are needed to address limitations of existing AI tools, such as inadequate precision and poor adaptability, while deepening research into the LLM applications in the publication update phase of LE synthesis. In the realm of ethics and data governance, it is essential to establish responsibility attribution mechanisms and cross-regulatory data governance frameworks, as well as enhance evidence inclusivity and mitigate privacy risks through algorithmic optimization. Methodologically, we recommend the establishment of a standardized evaluation system for AI applications and refining research design and quality assessment protocols to strengthen the evidence base.

Strengths and Limitations

The strengths of this review include the following: (1) it systematically analyzes the types of AI and semiautomated tools used across the 4 phases of LE synthesis and (2) it provides insights into the opportunities and challenges of using AI or semiautomated tools in LE synthesis regarding efficiency, accuracy, and utility. However, this review still has a few limitations. First, study screening was based on whether the studies reported on the tools used in LE synthesis. Second, studies that did not document the use of AI or semiautomated tools in LE synthesis were excluded from this review, which may introduce bias. Third, the focus of our search strategy on "living evidence" terminology may have excluded studies describing AI tools for review updates that used different terminology.

Conclusion

Researchers are actively utilizing various AI and semiautomated tools in LE synthesis, primarily for data extraction or collection and risk of bias assessment, while their application in updating publications remains limited. The use of AI or semiautomated tools in LE synthesis improves efficiency in the database searching and eligibility phase and accuracy in the database searching and eligibility phase, as well as in the data extraction or collection and risk of bias assessment phase. The AI or semiautomated tools demonstrate high accuracy, recall, and F_1-scores, while precision varies across tools. AI or semiautomated tools also demonstrate good performance in terms of utility in the database searching and eligibility phase.

Funding

This work was supported by the Fundamental Research Funds for the Central Universities (lzujbky-2025 - 15) and Gansu Provincial Center for Disease Control and Prevention Research Program (GSJKKY2025-02).

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

XS designed the study, analyzed the data, and drafted the manuscript. ZL designed the study, analyzed the data, drafted the manuscript, and evaluated the quality of included studies. RW, QW, and XL developed the research design. RL and ZY drafted the manuscript and evaluated the quality of the included studies. LF, ZM, and ZP were in charge of data curation. CL, LG, YC, KY, and JL critically reviewed and revised the manuscript. All authors critically revised the study for important intellectual content and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies, included and excluded study information, and methodological quality assessment methods and results.
[DOCX File, 53 KB - jmir_v28i1e76130_app1.docx]

Multimedia Appendix 2
Frequency of artificial intelligence (AI) or semiautomated tools use.
[PNG File, 139 KB - jmir_v28i1e76130_app2.png]

Checklist 1
PRISMA-LSR checklist.
[DOCX File, 29 KB - jmir_v28i1e76130_app3.docx]

References

1. Elliott J, Lawrence R, Minx JC, et al. Decision makers need constantly updated evidence synthesis. Nature 2021 Dec;600(7889):383-385. [doi: 10.1038/d41586-021-03690-1] [Medline: 34912079]

XSL•FO
RenderX

2. Sampson M, Shojania KG, Garritty C, Horsley T, Ocampo M, Moher D. Systematic reviews can be produced and published faster. J Clin Epidemiol 2008 Jun;61(6):531-536. [doi: 10.1016/j.jclinepi.2008.02.004] [Medline: 18471656]
3. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med 2007 Aug 21;147(4):224-233. [doi: 10.7326/0003-4819-147-4-200708210-00179] [Medline: 17638714]
4. Turner T, Lavis JN, Grimshaw JM, Green S, Elliott J. Living evidence and adaptive policy: perfect partners? Health Res Policy Syst 2023 Dec 18;21(1):135. [doi: 10.1186/s12961-023-01085-4] [Medline: 38111030]
5. Elliott JH, Turner T, Clavisi O, et al. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. PLoS Med 2014 Feb;11(2):e1001603. [doi: 10.1371/journal.pmed.1001603] [Medline: 24558353]
6. Thomas J, Noel-Storr A, Marshall I, et al. Living systematic reviews: 2. Combining human and machine effort. J Clin Epidemiol 2017 Nov;91:31-37. [doi: 10.1016/j.jclinepi.2017.08.011] [Medline: 28912003]
7. Tendal B, Vogel JP, McDonald S, et al. Weekly updates of national living evidence-based guidelines: methods for the Australian living guidelines for care of people with COVID-19. J Clin Epidemiol 2021 Mar;131:11-21. [doi: 10.1016/j.jclinepi.2020.11.005] [Medline: 33188858]
8. Elliott JH, Synnot A, Turner T, et al. Living systematic review: 1. Introduction-the why, what, when, and how. J Clin Epidemiol 2017 Nov;91:23-30. [doi: 10.1016/j.jclinepi.2017.08.010] [Medline: 28912002]
9. Schmidt L, Sinyor M, Webb RT, et al. A narrative review of recent tools and innovations toward automating living systematic reviews and evidence syntheses. Z Evid Fortbild Qual Gesundhwes 2023 Sep;181:65-75. [doi: 10.1016/j.zefq.2023.06.007] [Medline: 37596160]
10. Yang Y, Qin J, Lei J, Liu Y. Research status and challenges on the sustainable development of artificial intelligence courses from a global perspective. Sustainability 2023;15(12):9335. [doi: 10.3390/su15129335]
11. Adams CE, Polzmacher S, Wolff A. Systematic reviews: work that needs to be done and not to be done. J Evid Based Med 2013 Nov;6(4):232-235. [doi: 10.1111/jebm.12072] [Medline: 24325416]
12. Suárez A, Jiménez J, Llorente de Pedro M, et al. Beyond the scalpel: assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery. Comput Struct Biotechnol J 2024 Dec;24:46-52. [doi: 10.1016/j.csbj.2023.11.058] [Medline: 38162955]
13. Bendersky J, Auladell-Rispau A, Urrútia G, Rojas-Reyes MX. Methods for developing and reporting living evidence synthesis. J Clin Epidemiol 2022 Dec;152:89-100. [doi: 10.1016/j.jclinepi.2022.09.020] [Medline: 36220626]
14. Akl EA, Khabsa J, Iannizzi C, et al. Extension of the PRISMA 2020 statement for living systematic reviews (PRISMA-LSR): checklist and explanation. BMJ 2024 Nov 19;387:e079183. [doi: 10.1136/bmj-2024-079183] [Medline: 39562017]
15. Which phases of living evidence synthesis use artificial intelligence (AI)? an living evidence synthesis. OSF. URL: https://doi.org/10.17605/OSF.IO/4FVDQ [accessed 2026-01-16]
16. Murad MH, Wang Z, Chu H, et al. Proposed triggers for retiring a living systematic review. BMJ Evid Based Med 2023 Oct;28(5):348-352. [doi: 10.1136/bmjebm-2022-112100] [Medline: 36889900]
17. Cochrane Living Systematic Reviews Network. Guidance for the production and publication of Cochrane living systematic reviews: Cochrane Reviews in living mode. : Cochrane; 2019 URL: https://community.cochrane.org/sites/default/files/uploads/inline-files/Transform/201912_LSR_Revised_Guidance.pdf [accessed 2026-01-13]
18. Kozhakhmetova A, Mamyrbayev A, Zhidebekkyzy A, Bilan S. Assessing the impact of artificial intelligence on project efficiency enhancement. Knowl Perform Manag 2024;8(2):109-126. [doi: 10.21511/kpm.08(2).2024.09]
19. Røhl UBU. Automated, administrative decision-making and good administration: friends, foes or complete strangers [Dissertation]. : Aalborg Universitetsforlag; 2022 URL: https://vbn.aau.dk/ws/files/549540893/PHD_UBUR.pdf [accessed 2026-01-19]
20. Khalil H, Ameen D, Zarnegar A. Tools to support the automation of systematic reviews: a scoping review. J Clin Epidemiol 2022 Apr;144:22-42. [doi: 10.1016/j.jclinepi.2021.12.005] [Medline: 34896236]
21. Rashid M, Yi CS, Lawin S, et al. MT14 role of generative artificial intelligence in assisting systematic review process in health research: a systematic review. Value Health 2025 Jul;28(6):S268. [doi: 10.1016/j.jval.2025.04.1124] [Medline: 40848037]
22. Validity assessment tools for evidence synthesis: your one-stop-shop. Latitudes Network. URL: https://www.latitudes-network.org [accessed 2026-01-13]
23. Equator Network - Enhancing the QUAlity and Transparency Of health Research. URL: http://equator-network.org [accessed 2025-08-06]
24. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
25. Santini A, Man A, Voidăzan S. Accuracy of diagnostic tests. J Crit Care Med (Targu Mures) 2021 Jul;7(3):241-248. [doi: 10.2478/jccm-2021-0022] [Medline: 34722928]
26. McArthur A, Klugarova J, Yan H, Florescu S. Chapter 4: systematic reviews of text and opinion. In: Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis JBI: JBI; 2020. [doi: 10.46658/JBIRM-17-04]

XSL•FO
RenderX

27. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017 Sep 21;358:j4008. [doi: 10.1136/bmj.j4008] [Medline: 28935701]
28. Welsh EJ, Normansell RA, Cates CJ. Assessing the methodological quality of systematic reviews. NPJ Prim Care Respir Med 2015 Mar 19;25:15019. [doi: 10.1038/npjpcrm.2015.19] [Medline: 25789911]
29. Perlman-Arrow S, Loo N, Bobrovitz N, Yan T, Arora RK. A real-world evaluation of the implementation of NLP technology in abstract screening of a systematic review. Res Synth Methods 2023 Jul;14(4):608-621. [doi: 10.1002/jrsm.1636] [Medline: 37230483]
30. Higgins JPT, Thomas J, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions Version 6.5: Cochrane; 2024. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current [accessed 2026-01-13]
31. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ 2020 Jan 16;368:l6890. [doi: 10.1136/bmj.l6890] [Medline: 31948937]
32. Knafou J, Haas Q, Borissov N, et al. Ensemble of deep learning language models to support the creation of living systematic reviews for the COVID-19 literature. Syst Rev 2023 Jun 5;12(1):94. [doi: 10.1186/s13643-023-02247-9] [Medline: 37277872]
33. Matl S, Brosig R, Baust M, Navab N, Demirci S. Vascular image registration techniques: a living review. Med Image Anal 2017 Jan;35:1-17. [doi: 10.1016/j.media.2016.05.005] [Medline: 27294558]
34. Schmidt L, Finnerty Mutlu AN, Elmore R, Olorisade BK, Thomas J, Higgins JPT. Data extraction methods for systematic review (semi)automation: update of a living systematic review. F1000Res 2021;10:401. [doi: 10.12688/f1000research.51117.3] [Medline: 34408850]
35. Chou R, Dana T, Shetty KD. Testing a Machine Learning Tool for Facilitating Living Systematic Reviews of Chronic Pain Treatments: Agency for Healthcare Research and Quality; 2020. [Medline: 33175480]
36. Kamso MM, Pardo JP, Whittle SL, et al. Crowd-sourcing and automation facilitated the identification and classification of randomized controlled trials in a living review. J Clin Epidemiol 2023 Dec;164:1-8. [doi: 10.1016/j.jclinepi.2023.10.007] [Medline: 37865299]
37. Riaz IB, Sipra Q, Naqvi SAA, et al. Quantifying absolute benefit for adjuvant treatment options in renal cell carcinoma: a living interactive systematic review and network meta-analysis. Crit Rev Oncol Hematol 2022 Jul;175:103706. [doi: 10.1016/j.critrevonc.2022.103706] [Medline: 35537621]
38. Butler AR, Hartmann-Boyce J, Livingstone-Banks J, Turner T, Lindson N. Optimizing process and methods for a living systematic review: 30 search updates and three review updates later. J Clin Epidemiol 2024 Feb;166:111231. [doi: 10.1016/j.jclinepi.2023.111231] [Medline: 38043829]
39. Riaz IB, He H, Ryu AJ, et al. A living, interactive systematic review and network meta-analysis of first-line treatment of metastatic renal cell carcinoma. Eur Urol 2021 Dec;80(6):712-723. [doi: 10.1016/j.eururo.2021.03.016] [Medline: 33824031]
40. Riaz IB, Siddiqi R, Islam M, et al. Adjuvant tyrosine kinase inhibitors in renal cell carcinoma: a concluded living systematic review and meta-analysis. JCO Clin Cancer Inform 2021 May;5:588-599. [doi: 10.1200/CCI.21.00035] [Medline: 34043431]
41. Hair K, Wilson E, Wong C, Tsang A, Macleod M, Bannach-Brown A. Systematic online living evidence summaries: emerging tools to accelerate evidence synthesis. Clin Sci (Lond) 2023 May 31;137(10):773-784. [doi: 10.1042/CS20220494] [Medline: 37219941]
42. Marshall IJ, Trikalinos TA, Soboczenski F, et al. In a pilot study, automated real-time systematic review updates were feasible, accurate, and work-saving. J Clin Epidemiol 2023 Jan;153:26-33. [doi: 10.1016/j.jclinepi.2022.08.013] [Medline: 36150548]
43. Haas Q, Alvarez DV, Borissov N, et al. Utilizing artificial intelligence to manage COVID-19 scientific evidence torrent with Risklick AI: a critical tool for pharmacology and therapy development. Pharmacology 2021;106(5-6):244-253. [doi: 10.1159/000515908] [Medline: 33910199]
44. Vaghela U, Rabinowicz S, Bratsos P, et al. Using a secure, continually updating, web source processing pipeline to support the real-time data synthesis and analysis of scientific literature: development and validation study. J Med Internet Res 2021 May 6;23(5):e25714. [doi: 10.2196/25714] [Medline: 33835932]
45. Karakülah G, Suner A, Adlassnig KP, Samwald M. A data-driven living review for pharmacogenomic decision support in cancer treatment. Stud Health Technol Inform 2012;180:688-692. [doi: 10.3233/978-1-61499-101-4-688] [Medline: 22874279]
46. Xin Y, Nevill CR, Nevill J, et al. Feasibility study for interactive reporting of network meta-analysis: experiences from the development of the Metainsight COVID-19 app for stakeholder exploration, re-analysis and sensitivity analysis from living systematic reviews. BMC Med Res Methodol 2022 Jan 22;22(1):26. [doi: 10.1186/s12874-022-01507-x] [Medline: 35065603]
47. Evrenoglou T, Boutron I, Seitidis G, Ghosn L, Chaimani A. metaCOVID: a web-application for living meta-analyses of COVID-19 trials. Res Synth Methods 2023 May;14(3):479-488. [doi: 10.1002/jrsm.1627] [Medline: 36772980]
48. Kaiser K, Miksch S. Versioning computer-interpretable guidelines: semi-automatic modeling of “living guidelines” using an information extraction method. Artif Intell Med 2009 May;46(1):55-66. [doi: 10.1016/j.artmed.2008.08.009] [Medline: 18950994]
49. Skinner G, Cooke R, Keum J, et al. Dynameta: a dynamic platform for ecological meta-analyses in R Shiny. SoftwareX 2023 Jul;23:101439. [doi: 10.1016/j.softx.2023.101439]

XSL•FO
RenderX

50. McDonald S, Hill K, Li HZ, Turner T. Evidence surveillance for a living clinical guideline: case study of the Australian stroke guidelines. Health Info Libr J 2023 Nov 9. [doi: 10.1111/hir.12515] [Medline: 37942888]
51. Shemilt I, Arno A, Thomas J, et al. Cost-effectiveness of Microsoft Academic Graph with machine learning for automated study identification in a living map of coronavirus disease 2019 (COVID-19) research. Wellcome Open Res 2024;6:210. [doi: 10.12688/wellcomeopenres.17141.2] [Medline: 38686019]
52. Le-Khac UN, Bolton M, Boxall NJ, Wallace SMN, George Y. Living review framework for better policy design and management of hazardous waste in Australia. Sci Total Environ 2024 May 10;924:171556. [doi: 10.1016/j.scitotenv.2024.171556] [Medline: 38458450]
53. Khan MA, Ayub U, Naqvi SAA, et al. Collaborative large language models for automated data extraction in living systematic reviews. J Am Med Inform Assoc 2025 Apr 1;32(4):638-647. [doi: 10.1093/jamia/ocae325] [Medline: 39836495]
54. Hair K, Wilson E, Maksym O, Macleod MR, Sena ES. A systematic online living evidence summary of experimental Alzheimer's disease research. J Neurosci Methods 2024 Sep;409:110209. [doi: 10.1016/j.jneumeth.2024.110209] [Medline: 38964475]
55. Grbin L, Nichols P, Russell F, Fuller-Tyszkiewicz M, Olsson CA. The development of a living knowledge system and implications for future systematic searching. J Aust Libr Inf Assoc 2022 Jul 3;71(3):275-292. [doi: 10.1080/24750158.2022.2087954]
56. Hearnden J, Dudoit K, Kim E, Tremblay G, Forsythe A. PMU118 use of computer-assisted methods to realize the concept of a living systematic review via an online platform. Value Health 2019 Nov;22:S729. [doi: 10.1016/j.jval.2019.09.1736]
57. Evrenoglou T, Boutron I, Chaimani A. metaCOVID: an R-Shiny application for living meta-analyses of COVID-19 trials. medRxiv. Preprint posted online on Sep 10, 2021. [doi: 10.1101/2021.09.07.21263207]
58. Stoll A, Wilms L, Ziegele M. Developing an incivility dictionary for German online discussions—a semi-automated approach combining human and artificial knowledge. Commun Methods Meas 2023 Apr 3;17(2):131-149. [doi: 10.1080/19312458.2023.2166028]
59. Meza N, Pérez-Bracchiglione J, Pérez I, et al. Angiotensin-converting-enzyme inhibitors and angiotensin II receptor blockers for COVID-19: a living systematic review of randomized clinical trials. Medwave 2021 Mar 3;21(2):e8105. [doi: 10.5867/medwave.2021.02.8105] [Medline: 33830976]
60. Verdejo C, Vergara-Merino L, Meza N, et al. Macrolides for the treatment of COVID-19: a living, systematic review. Medwave 2020 Dec 14;20(11):e8074. [doi: 10.5867/medwave.2020.11.8073] [Medline: 33361755]
61. Baladia E, Pizarro AB, Ortiz-Muñoz L, Rada G. Vitamin C for COVID-19: a living systematic review. Medwave 2020 Jul 28;20(6):e7978. [doi: 10.5867/medwave.2020.06.7978] [Medline: 32759894]
62. Rada G, Corbalán J, Rojas P, COVID-19 L·OVE Working Group. Cell-based therapies for COVID-19: a living, systematic review. Medwave 2020 Dec 17;20(11):e8079. [doi: 10.5867/medwave.2020.11.8078] [Medline: 33382060]
63. Gates M, Elliott SA, Gates A, et al. LOCATE: a prospective evaluation of the value of leveraging ongoing citation acquisition techniques for living evidence syntheses. Syst Rev 2021 Apr 19;10(1):116. [doi: 10.1186/s13643-021-01665-x] [Medline: 33875014]
64. Piechotta V, Iannizzi C, Chai KL, et al. Convalescent plasma or hyperimmune immunoglobulin for people with COVID-19: a living systematic review. Cochrane Database Syst Rev 2021 May 20;5(5):CD013600. [doi: 10.1002/14651858.CD013600.pub4] [Medline: 34013969]
65. Verdugo-Paiva F, Acuña MP, Solá I, Rada G, COVID-19 L·OVE Working Group. Remdesivir for the treatment of COVID-19: a living systematic review. Medwave 2020 Dec 9;20(11):e8080. [doi: 10.5867/medwave.2020.11.8080] [Medline: 33361753]
66. Shackelford GE, Martin PA, Hood ASC, Christie AP, Kulinskaya E, Sutherland WJ. Dynamic meta-analysis: a method of using global evidence for local decision making. BMC Biol 2021 Feb 17;19(1):33. [doi: 10.1186/s12915-021-00974-w] [Medline: 33596922]
67. Verdugo-Paiva F, Izcovich A, Ragusa M, Rada G. Lopinavir-ritonavir for COVID-19: a living systematic review. Medwave 2020 Jul 15;20(6):e7967. [doi: 10.5867/medwave.2020.06.7966] [Medline: 32678815]
68. Iannizzi C, Chai KL, Piechotta V, et al. Convalescent plasma for people with COVID-19: a living systematic review. Cochrane Database Syst Rev 2023 Feb 1;2(2):CD013600. [doi: 10.1002/14651858.CD013600.pub5] [Medline: 36734509]
69. Sommer I, Ledinger D, Thaler K, et al. Outpatient treatment of confirmed COVID-19: a living, rapid evidence review for the American College of Physicians (Version 2). Ann Intern Med 2023 Oct;176(10):1377-1385. [doi: 10.7326/M23-1626] [Medline: 37722115]
70. Verdugo-Paiva F, Vergara C, Ávila C, et al. COVID-19 living overview of evidence repository is highly comprehensive and can be used as a single source for COVID-19 studies. J Clin Epidemiol 2022 Sep;149:195-202. [doi: 10.1016/j.jclinepi.2022.05.001] [Medline: 35597369]
71. Paul D, Chakdar D, Saha S, Mathew J. Online research topic modeling and recommendation utilizing multiview autoencoder-based approach. IEEE Trans Comput Soc Syst 2024;11(1):1013-1022. [doi: 10.1109/TCSS.2023.3253502]
72. Vergara-Merino L, Verdejo C, Carrasco C, Vargas-Peirano M. Living systematic review: new inputs and challenges. Medwave 2020 Dec 23;20(11):e8092. [doi: 10.5867/medwave.2020.11.8092] [Medline: 33382391]
73. Elbers S, Wittink H, Kaiser U, et al. Living systematic reviews in rehabilitation science can improve evidence-based healthcare. Syst Rev 2021 Dec 7;10(1):309. [doi: 10.1186/s13643-021-01857-5] [Medline: 34876231]

XSL•FO
RenderX

74. Elvidge J, Hopkin G, Narayanan N, Nicholls D, Dawoud D. Diagnostics and treatments of COVID-19: two-year update to a living systematic review of economic evaluations. Front Pharmacol 2023;14:1291164. [doi: 10.3389/fphar.2023.1291164] [Medline: 38035028]
75. Winters M, Lyng KD, Holden S, et al. Infographic. Comparative effectiveness of treatments for patellofemoral pain: a living systematic review with network meta-analysis. Br J Sports Med 2021 Nov;55(22):1311-1312. [doi: 10.1136/bjsports-2021-104360] [Medline: 34244170]
76. Katanandov SL, Kovalev AA. Technological development of modern states: artificial intelligence in public administration. State and Municipal Management Scholar Notes 2023 Mar;1(1):174-182. [doi: 10.22394/2079-1690-2023-1-1-174-182]
77. Medaglia R, Gil-Garcia JR, Pardo TA. Artificial intelligence in government: taking stock and moving forward. Soc Sci Comput Rev 2021;41(1):123-140. [doi: 10.1177/08944393211034087]
78. Thomas J, Flemyng E, Noel-Storr A, et al. Responsible AI in Evidence SynthEsis (RAISE): guidance and recommendations. : Open Science Framework; 2024 URL: https://osf.io/fwaud/files/cn7x4 [accessed 2026-01-13]
79. Filetti S, Fenza G, Gallo A. Research design and writing of scholarly articles: new artificial intelligence tools available for researchers. Endocrine 2024 Sep;85(3):1104-1116. [doi: 10.1007/s12020-024-03977-z] [Medline: 39085566]
80. Ramnani S. Exploring ethical considerations of artificial intelligence in educational settings: an examination of bias, privacy, and accountability. International Journal of Novel Research and Development 2024;9(2):b173-b191. [doi: 10.1729/Journal.37869]
81. Li Y, Shao S, He Y, et al. Rethinking data protection in the (generative) artificial intelligence era. arXiv. Preprint posted online on Jul 3, 2025. [doi: 10.48550/arXiv.2507.03034]
82. Umeh II, Umeh KC. A comparative analysis of AI system development tools for improved outcomes. International Journal of Sustainability Management and Information Technologies 2025;11:1-20. [doi: 10.11648/j.ijsmit.20251101.11]
83. Ramírez G. Improving the health of populations—evidence for policy and practice action. J Evid Based Med 2009 Nov;2(4):216-219. [doi: 10.1111/j.1756-5391.2009.01044.x] [Medline: 21349019]
84. Manson H. Systematic reviews are not enough: policymakers need a greater variety of synthesized evidence. J Clin Epidemiol 2016 May;73:11-14. [doi: 10.1016/j.jclinepi.2015.08.032] [Medline: 26912122]
85. Berger-Tal O, Wong BBM, Adams CA, et al. Leveraging AI to improve evidence synthesis in conservation. Trends Ecol Evol 2024 Jun;39(6):548-557. [doi: 10.1016/j.tree.2024.04.007] [Medline: 38796352]
86. Jacob S. Artificial intelligence and the future of evaluation: from augmented to automated evaluation. Digit Gov Res Pract 2025 Mar 31;6(1):1-10. [doi: 10.1145/3696009]
87. Head CB, Jasper P, McConnachie M, Raftree L, Higdon G. Large language model applications for evaluation: opportunities and ethical implications. New Dir Eval 2023 Jun;2023(178-179):33-46. [doi: 10.1002/ev.20556]
88. van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, van der Palen J, Doggen CJM, Lenferink A. Artificial intelligence in systematic reviews: promising when appropriately used. BMJ Open 2023 Jul 7;13(7):e072254. [doi: 10.1136/bmjopen-2023-072254] [Medline: 37419641]
89. Thomas IN, Roche P, Grêt-Regamey A. Harnessing artificial intelligence for efficient systematic reviews: a case study in ecosystem condition indicators. Ecol Inform 2024 Nov;83:102819. [doi: 10.1016/j.ecoinf.2024.102819]
90. Manion FJ, Du J, Wang D, et al. Accelerating evidence synthesis in observational studies: development of a living natural language processing-assisted intelligent systematic literature review system. JMIR Med Inform 2024 Oct 23;12:e54653. [doi: 10.2196/54653] [Medline: 39441204]
91. Glanville J. The role of AI tools in developing search strategies and identifying evidence for systematic reviews. : Evidence Synthesis Ireland; 2025 URL: https://evidencesynthesisireland.ie/wp-content/uploads/2025/05/ESI-2025-AI-tools-Julie-Glanville.pdf [accessed 2026-01-13]
92. New Joint AI Methods Group: guiding responsible use of AI in evidence synthesis. Joanna Briggs Institute (JBI). URL: https://jbi.global/news/article/new-joint-ai-methods-group [accessed 2026-01-13]
93. Roadmap. Evidence Synthesis Infrastructure Collaborative. 2024. URL: https://evidencesynthesis.atlassian.net/wiki/spaces/ESE/pages/344817670/English [accessed 2026-01-13]
94. Scotcher S. Evidence Synthesis Infrastructure Collaborative. European Evaluation Society. 2025. URL: https://europeanevaluation.org/events/evidence-synthesis-infrastructure-collaborative [accessed 2026-01-13]

Abbreviations

AI: artificial intelligence
DTA: diagnostic test accuracy
JBI: Joanna Briggs Institute
LE: living evidence
LIvE: Living Interactive Evidence
LLM: large language model
PRISMA-LSR : Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 statement for living systematic reviews

XSL•FO
RenderX

QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies version 2

Edited by A Coristine; submitted 17.Apr.2025; peer-reviewed by C Bhagat, E Ting; accepted 15.Dec.2025; published 27.Jan.2026.

Please cite as:
Song X, Lian Z, Wang R, Li R, Yang Z, Luo X, Feng L, Ma Z, Pu Z, Wang Q, Ge L, Li C, Chen Y, Yang K, Lavis J
The Phases of Living Evidence Synthesis Using AI: Living Evidence Synthesis (Version 1)
J Med Internet Res 2026;28:e76130
URL: https://www.jmir.org/2026/1/e76130
doi:10.2196/76130

© Xuping Song, Zhenjie Lian, Rui Wang, Ruixin Li, Zhenzhen Yang, Xufei Luo, Lei Feng, Zhiming Ma, Zhen Pu, Qi Wang, Long Ge, Caihong Li, Yaolong Chen, Kehu Yang, John Lavis. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Patient Benefits in the Context of Sepsis-Related AI-Based Clinical Decision Support Systems: Scoping Review

Pascal Raszke[1], MA; Godwin Denk Giebel[1], Dr rer medic; Jürgen Wasem[1], Prof Dr; Michael Adamzik[2], Prof Dr Med; Hartmuth Nowak[3], PD, Dr med; Lars Palmowski[2], Dr med; Philipp Heinz[4], Dipl-Kfm; Nina Timmesfeld[5], Prof Dr; Marianne Tokic[5], MSc; Frank Martin Brunkhorst[6], Prof Dr Med; Nikola Blase[1], Dr med

[1]Institute for Health Care Management and Research, University of Duisburg-Essen, Thea-Leymann-Str. 9, Essen, Germany

[2]Department of Anesthesiology, Intensive Care Medicine and Pain Therapy, Ruhr University Bochum, Knappschaft Kliniken University Hospital Bochum, Bochum, Germany

[3]Department of Anesthesiology, Intensive Care Medicine and Pain Therapy, Center for Artificial Intelligence, Medical Informatics and Data Science, Ruhr-University Bochum, Knappschaft Kliniken University Hospital Bochum, Bochum, Germany

[4]Knappschaft Kliniken GmbH, Recklinghausen, Germany

[5]Department of Medical Informatics, Biometry and Epidemiology, Ruhr University Bochum, Bochum, Germany

[6]Institute of Infectious Diseases and Infection Control, Jena University Hospital, Jena, Germany

Corresponding Author:
Pascal Raszke, MA
Institute for Health Care Management and Research, University of Duisburg-Essen, Thea-Leymann-Str. 9, Essen, Germany

Abstract

Background: Global digitalization continues to advance, extending its influence into medicine and health care systems worldwide. In recent years, substantial advancements have been made in the research and development of artificial intelligence (AI), raising questions about its potential in medicine. The integration and application of AI in intensive care medicine, particularly in sepsis treatment, presents significant potential for advancing patient outcomes and enhancing patient-relevant benefits. However, a comprehensive and systematic overview of the full spectrum of patient-relevant benefits associated with AI-based clinical decision support systems (CDSS) remains lacking.

Objective: This scoping review aimed to identify and categorize evidence on patient-relevant benefits of AI-based CDSS in sepsis care.

Methods: Systematic research was conducted in 4 electronic databases: MEDLINE via PubMed, Embase, the ACM Digital Library, and IEEE Xplore. In addition, a comprehensive search on the websites of relevant international organizations, along with a citation search of the included articles, was conducted. Articles were included if they (1) focused on sepsis and (2) described patient-relevant benefits of AI-based CDSS. Articles published between January 1, 2008, and March 2, 2023, were considered for inclusion. Study selection was performed independently by 2 reviewers. The manuscript was drafted in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist. The analysis of the included articles was conducted using the program MAXQDA (VERBI Software GmbH), with systemization finalized in a consensus workshop.

Results: A total of 3368 records were identified across the 4 databases, of which 24 met the inclusion criteria and were included in the scoping review. The additional search on international websites and in reference lists identified 6 more relevant articles, resulting in 30 included studies. Of these, 20 were quantitative, comprising 7 prospective and 13 retrospective designs. In addition, 1 qualitative study, 1 mixed methods study, 6 review articles, and 2 articles from institutional websites were included. Patient-relevant benefits were systematized in six main categories: (1) prediction, (2) earlier treatment and prioritization of high-risk patients, (3) individualized therapy, (4) improved patient outcomes (including improved Sequential Organ Failure Assessment score, reduced length of stay, and reduced mortality), (5) general improvements in care, and (6) reduced readmission rate.

Conclusions: This scoping review underscores the potential of AI-based CDSS to positively impact patient-relevant benefits, particularly in sepsis care, where they demonstrate considerable promise for improving intensive care. However, the majority of the identified studies rely on retrospective database analyses. Future research should focus on validating these findings through prospective studies.

(J Med Internet Res 2026;28:e76772) doi:10.2196/76772

XSL•FO
RenderX

KEYWORDS

medical informatics; artificial intelligence; machine learning; computational intelligence; clinical decision support systems; CDSS; decision support; sepsis; bloodstream infection

Introduction

The treatment of infectious diseases has historically resulted in medical progress, exemplified by antibiotics and vaccines. Despite all medical advances, infections remain a major global cause of morbidity and mortality [1,2]. Sepsis, defined as "life-threatening organ dysfunction caused by a dysregulated host response to infection" [3], remains among the top contributors to worldwide mortality. It accounts for 30% - 50% of all hospital deaths in high-income countries, such as the United States [1], and approximately 11 million annual deaths worldwide [2]. Sepsis is a heterogeneous syndrome with variable phenotypes and outcomes. Thus, the interpretation of initial symptoms can be difficult for health care providers [3,4].

The effectiveness and accuracy of established rule-based scoring systems used for the assessment of patients in the intensive care unit (ICU), such as the systemic inflammatory response syndrome (SIRS) criteria, which were historically of importance, the sequential organ failure assessment (SOFA) score or the quickSOFA (qSOFA) score, the acute physiology and chronic health evaluation II (APACHE II) score, or the national early warning score 2 (NEWS2), is limited. This is partly because these scoring systems are not always specifically developed for sepsis patients and are therefore of limited use to health care providers in this context [5,6]. Nevertheless, timely identification and treatment are crucial to enhance patient outcomes [7-9], as untreated sepsis can progress to septic shock, exacerbating the patient's condition [10] and leading to multiple organ failure, which carries an even higher mortality rate than sepsis itself [11].

This is where recent developments in artificial intelligence (AI) become particularly relevant, as they are considered to hold substantial potential for improving sepsis diagnostics. Especially machine learning (ML), a branch of AI, has the ability to rapidly analyze vast amounts of data, exceeding human capacity to process. By evaluating numerous data points, ML can derive conclusions and recognize correlations that a human health care provider would be incapable of identifying. This is why ML is well-suited as a technological foundation for clinical decision support systems (CDSS), particularly in the complex clinical picture of sepsis [3]. The use of ML in the development of CDSS can make the sepsis diagnosis more reliable, with the prospect of long-term improvements in patient outcomes. Machine learning algorithms (MLAs) demonstrated potential to enhance patient-relevant benefits in distinct studies. Documented benefits include, for example, reductions in sepsis-related mortality and the average hospital length of stay (LOS). Additionally, MLAs facilitate earlier interventions, such as the timely administration of antibiotics [12-14].

Despite the high clinical relevance of sepsis and significant advancements in both the availability of digital patient data and in the field of ML, the real-world application of AI-based CDSS remains negligible. The majority of these algorithms remain in the prototype phase, with deployment limited to a single hospital or a single hospital operator. This gap is highlighted by an analysis of the Food and Drug Administration's database of medical devices using AI or ML. As of April 2025, none of the over 1000 listed products are specifically dedicated to intensive care [15], the medical field at the forefront of sepsis treatment. This illustrates the discrepancy between technological progress and its real-world implementation in the critical care environment. For AI-based CDSS to be successfully implemented in clinical practice, it is a necessary prerequisite that they demonstrate tangible added value. Accordingly, patient-relevant benefits should constitute a primary focus.

The research objective of the present study differs from those of previous scoping reviews on AI-based CDSS in sepsis care. Certain reviews focused specifically on neonatal [16] or pediatric [17] sepsis, whereas others concentrated on tasks for which MLAs were designed—such as risk assessment, treatment planning, or process support—and thus focused on the process of medical service delivery rather than on actual patient-relevant benefits [18] or on the actual design of the CDSS and its intended users [19]. Importantly, none of the aforementioned reviews [16-19] focused exclusively on patient-relevant benefits. Furthermore, several existing scoping reviews used narrow methodological approaches, for example, being restricted to a single ML method [17] or considering only antibiotic treatment of sepsis [16]. To the authors' knowledge, no other scoping review has explicitly examined the patient-relevant benefits of AI-based CDSS in the context of sepsis while applying a broad and exploratory methodological approach, without restrictions regarding the ML methods used or the types of patient-relevant benefits assessed. Accordingly, the objective of the present study is to identify patient-relevant benefits of AI-based CDSS in sepsis care compared with the current standard of care, thereby addressing this research gap, as patient-relevant benefits constitute a meaningful benchmark for evaluating the value of any medical innovation. In this context, a taxonomy of benefits comprising 6 main categories has been developed.

This scoping review was conducted within the framework of the KI@work (User-Oriented Requirements for AI-Based Clinical Decision Support Systems) project, which is funded by the German Federal Joint Committee (funding code: 01VSF22050). The research project is led by the Institute for Health Care Management and Research at the University of Duisburg-Essen. Consortium partners include the Department of Anesthesiology, Intensive Care Medicine and Pain Therapy at the University Hospital Knappschaftskrankenhaus Bochum, the Knappschaft Kliniken GmbH, the Department of Medical Informatics, Biometry and Epidemiology at the Ruhr University Bochum and the German Sepsis Society. This scoping review addressed 2 additional research questions. However, to ensure a coherent presentation of the findings, this article focuses exclusively on patient-relevant benefits.

XSL•FO
RenderX

Methods

Overview

This scoping review is based on the methodology framework of the Joanna Briggs Manual for evidence synthesis [20], a further development of the work of Arksey and O'Malley [21] and Levac et al [22]. The review process followed the five stages originally described by Arksey and O'Malley: (1) identifying the research question, (2) identifying relevant studies, (3) study selection, (4) charting the data, and (5) collating, summarizing, and reporting the results [21]. The manuscript was prepared according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist by Tricco et al [23] (Checklist 1). As scoping reviews encompass a broad range of study types in order to present a comprehensive overview of the research field [20-22], comparability between studies is limited. Consequently, no formal quality appraisal was conducted. Although no distinct protocol for the scoping review was published, the methodology was described in detail in a protocol for the overarching multimethod research project [24].

Search Strategy

The development of the search strategy commenced with an initial limited search in MEDLINE via PubMed and Embase to identify relevant search terms. Subsequently, the identified terms were discussed in recurring team discussions within the consortium. In a third step, the consented search terms were combined into search queries.

The electronic databases MEDLINE via PubMed, Embase, as well as the ACM Digital Library and IEEE Xplore, were searched for relevant literature on March 2, 2023. The databases were selected to ensure that the interdisciplinary research question could be adequately addressed from both a medical and a computer science perspective. The search string was developed using the PCC (population=persons with or at risk of sepsis, concept=CDSS, and context=AI) framework. The MEDLINE via PubMed search string was quality-assured by the chief librarian at the library of the University Medical Centre Essen before the database search was conducted. The other 3 search strings were developed based on the same quality assurance principles as the MEDLINE via PubMed search query. The individual search terms were limited to occurrences in title, abstract, and keyword searches but were supplemented by indexing terms (MeSH and Emtree) and truncations. The final search strategies for each database can be found in Multimedia Appendices 1-4.

In agreement with ML experts (NT, HN), the search was limited to articles published in the last 15 years. Further explanation for the time restriction is provided in the discussion of this article. The search was restricted to English and German. In cases of missing full texts, the interlibrary loan service of the University of Duisburg-Essen was used. If that approach was not successful, the reviewers contacted the respective authors of the papers of interest. The identified citations were imported into the reference management program Endnote 20 (Clarivate Analytics).

In addition to the systematic search of electronic databases, a structured search for gray literature (eg, working papers and guidelines) from various governmental and nongovernmental stakeholders was conducted via their websites. The selection of countries included in the search was based on the results of the Bertelsmann #SmartHealthSystem study, which examined the degree of digitalization of various health care systems in 2018. It was assumed that the prospect of identifying information on AI-based CDSS would be particularly high in countries with highly digitalized health care systems. According to the Bertelsmann study, this applies to the health care systems of Canada, Denmark, Estonia, Israel, and Spain. In addition, 3 large economies—Germany, the United Kingdom, and the United States—were included in the structured research. Alongside institutional websites from these countries, websites of relevant international stakeholders were also examined. These included the World Health Organization (WHO) and the Organisation for Economic Co-operation and Development (OECD), as well as websites of international sepsis, intensive care, and medical informatics associations. Further information about included websites can be found in Multimedia Appendix 5. To supplement further evidence, reference lists of articles identified through the systematic and structured search were screened, and the cited articles were subsequently assessed for eligibility. If eligible, the referenced articles were included in the scoping review.

Eligibility Criteria

Exploratory research and internal discussions contributed to the development of inclusion and exclusion criteria, which were refined iteratively during the initial stages of the research process. The search strategy was designed to address 3 different research questions. Studies were considered for inclusion if they described (1) patient-relevant benefits of AI-based CDSS in the context of sepsis as well as (2) problems in their development, implementation, or application, or (3) suggestions for improving these processes. Patient-relevant benefits were identified entirely exploratively and categorized independently of existing frameworks, allowing AI-based CDSS benefits to be classified without reliance on established definitions or patient-relevant endpoints. This approach provides a comprehensive and complete overview of the potential benefits of this emerging technology, without constraining the findings of this paper to predefined frameworks and definitions. Patient-relevant benefits were defined as the positive impact of an intervention on patients, irrespective of whether these comprise general qualitative observations or specific, measurable quantitative endpoints. Specific inclusion and exclusion criteria were developed for each research question to ensure a tailored approach to the unique scope of each question. AI was defined as ML-based algorithms that operate as a "black box" for the user (physician or caregiver), meaning their output is not directly interpretable for health care providers. Consequently, all ML-based technologies developed through data-driven training and sufficiently complex to preclude full comprehension by the user were eligible for inclusion. In contrast, rule-based algorithms, such as those relying on SIRS or SOFA criteria, did not meet this definition and were therefore excluded in this review. Moreover, earlier diagnosis facilitated by AI-based

XSL•FO
RenderX

CDSS was not considered a patient-relevant benefit, as earlier diagnosis itself has no impact on patient outcomes. It is the interventions that follow an earlier diagnosis—such as increased attention by health care providers to patients developing sepsis or earlier initiation of treatment—that positively influence patient-relevant benefits. Accordingly, these parameters are pertinent to the scope of this review. Articles were selected regardless of the research method used. The inclusion criteria are presented in Textbox 1.

Exclusion criteria for this review were not answering the research question, an exclusively technical description of the algorithms developed, or exclusively mathematical approaches not providing evidence for patient-relevant benefit. In addition, articles were excluded if they focused only on the evaluation of binary classifiers such as sensitivity, specificity, positive predictive value, or negative predictive value, as the superiority of AI-based algorithms over rule-based scores was considered a prerequisite for such systems. AI-based CDSS developed exclusively for neonates and/or children or for animals were also not included, because (1) the treatment of neonatal or pediatric sepsis patients differs significantly from the treatment of adult patients [25-27] and (2) the focus of the study is on human sepsis. Articles published before 2008 were also excluded, as were those written in languages other than English or German. Research protocols, conference abstracts, letters to the editor, and articles that were only expressions of opinions were also excluded. The exclusion criteria are listed in Textbox 1.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Articles focusing on sepsis and
- Involving AI-based CDSS, that
  - Describe patient-relevant benefits, or
  - Describe problems with development, implementation, or application, or
  - Describe strategies for success

Exclusion criteria

- Exclusively technical description of systems, or
- Focus on description of the evaluation of binary classifiers, or
- Articles describing AI-based CDSS for neonates and children or animals, or
- Not addressing any of the research questions in more detail, or
- Research protocols, conference abstracts, theses, letters to the editor, or expression of opinions, or
- Article published before 2008, or
- Language other than English or German

Evidence Screening, Selection, and Data Extraction

After identification and deletion of duplicates, title and abstract screening was conducted independently by 2 reviewers (PR and GDG) to decide whether an article was eligible for full-text screening. In a second step, the same 2 reviewers conducted a full-text screening of the included articles against the inclusion and exclusion criteria. In case of disagreement between the 2 reviewers during step 2 of the screening process, other members of the study team (NB, HN, and NT) were involved to decide whether an article was eligible for inclusion.

MAXQDA (VERBI Software GmbH) software was used to identify and tag relevant content in the included articles and to precategorize the patient-relevant benefit categories (PR) using an inductive coding approach. The preliminary categories were discussed and further refined in an in-person workshop based on the affinity mapping technique (PR, NB, and GDG). For this purpose, all relevant text passages were printed as snippets and physically assigned to the respective preliminary categories before being refined and finalized during the workshop. Each assignment was discussed in detail until full consensus among all 3 team members was reached. The results of the workshop were subsequently digitalized in Microsoft Excel. In addition to the patient-relevant benefits of AI-based CDSS, metadata, such as participating authors, year of publication, country of study, database for MLA, or study type, were extracted and summarized (see Multimedia Appendix 6).

Analysis and Presentation of Results

The results of the included studies were summarized descriptively, and analysis was conducted to derive implications for policy, practice, and research. The patient-relevant benefits were grouped into 6 main categories. The main categories were presented in tabular form in an Excel file and diagrammatically. The patient-relevant benefit categories are presented chronologically in Multimedia Appendix 7.

XSL•FO
RenderX

Results

Selection of Sources of Evidence

Selection Process

In the systematic search, a total of 3368 titles and abstracts were retrieved. After removing 850 duplicates, 2518 articles remained for screening (Figure 1). Of these, 141 articles were screened for full text, and 39 met the inclusion criteria. Among these, 24 provided statements on patient-relevant benefits [28-51]. In addition, reference lists of the articles identified through the systematic search were analyzed, resulting in the identification of 5 additional articles, 2 of which reported information on patient-relevant benefits [52,53]. A complementary search of institutional websites led to the inclusion of 5 additional articles, 4 of which contained relevant information on patient-relevant benefits [54-57]. In total, 30 articles were included in the scoping review about patient-relevant benefits. The full-text screening process, including a detailed account of the reasons for exclusion, is presented in Multimedia Appendix 8.

Figure 1. Flow diagram illustrating the selection process of evidence. CDSS: clinical decision support system.

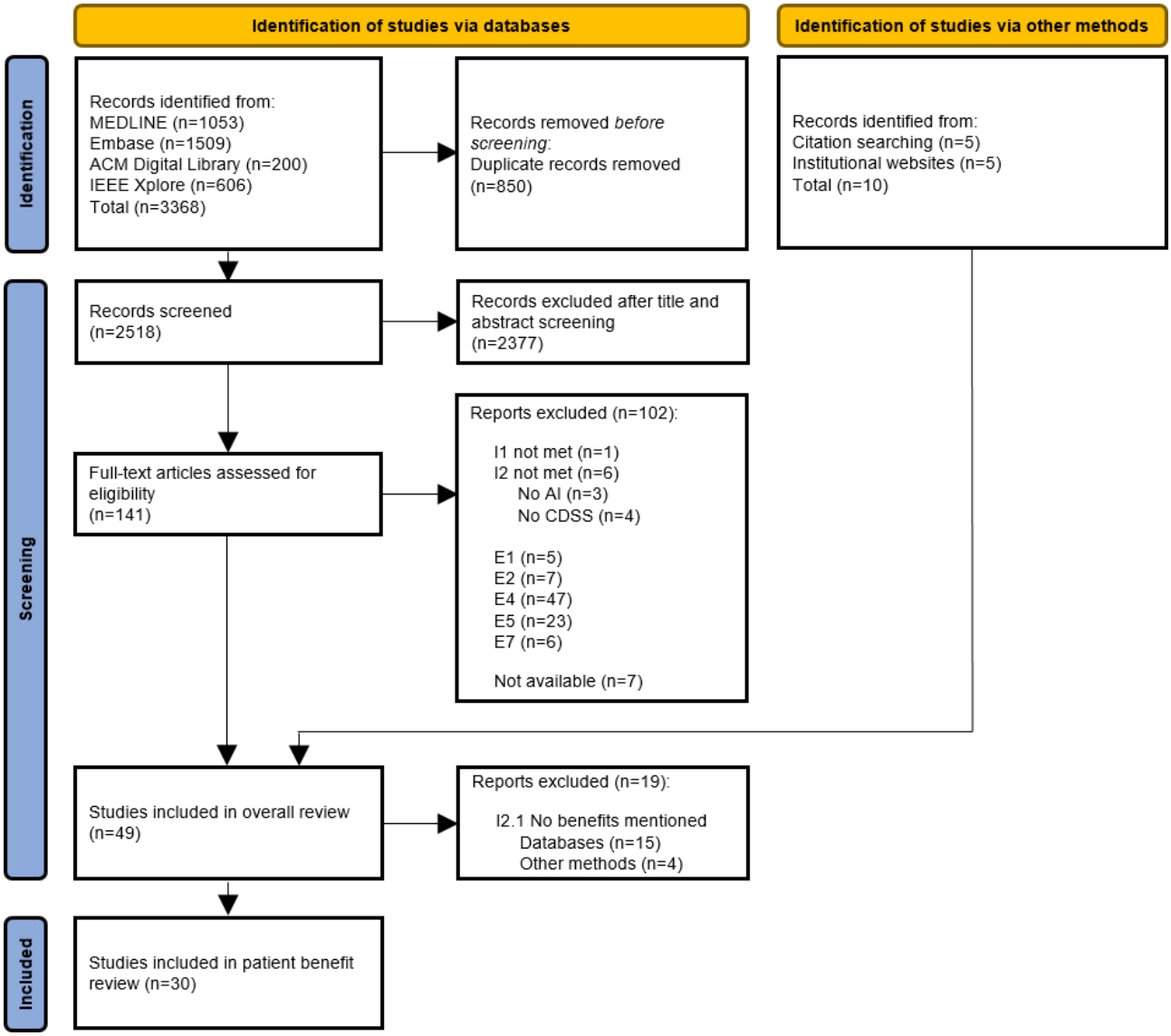

Included Studies

Of the 30 articles included, 16 originated from North America, all of which are from the United States (53.3%) [28,30-32,34,36-38,42,43,45,49,52,53,56,57]. Seven articles stem from Europe (23.3%); 3 from the Netherlands (10%) [46,48,50], 2 from Spain (6.6%) [33,44], 1 from Austria (3.3%) [29], and 1 from the United Kingdom (3.3%) [39]. Five articles are from Asia (16.7%), including 2 each from China [41,51] and Taiwan [40,54] (6.6% each), and 1 from Singapore (3.3%) [35]. There is 1 article from Australia (3.3%) [55] and 1 article from South America (Brazil) (3.3%) [47].

The study designs used in the included articles cover a wide range. Overall, 20 quantitative articles were identified. Of these, 7 used a prospective study design, of which 2 are multicenter studies [28,31] and 5 are single-center studies [34,43,49,52,54]. Thirteen of the quantitative studies used a retrospective approach, comprising 6 research database studies [29,30,39,40,46,53] and 7 electronic health record database studies [33,35,41,42,45,47,50]. In addition to the quantitative

XSL•FO
RenderX

articles, 1 article used a qualitative approach [37] and another applied a mixed methods approach [36]. Additionally, 6 review articles [32,38,44,48,51,55] and 2 articles from news sections of institutional websites were identified [56,57]. All articles are listed in Multimedia Appendix 6.

Synthesis of Results

In total, 6 main categories of patient-relevant benefit were identified. The 6 main categories identified reflect the patient pathway from pretreatment to posttreatment period. They include (1) prediction, (2) earlier treatment and prioritization of high-risk patients, (3) individualized therapy (which encompasses patient-centered care), (4) improved patient outcomes (which includes improved SOFA score, reduced length of stay, and reduced mortality), (5) general improvements in care, and (6) reduced readmission rate (see Figure 2). Multimedia Appendix 7 gives a detailed overview of the benefit categories addressed in each study.

Figure 2. Patient benefit categories related to artificial intelligence-based clinical decision support systems.

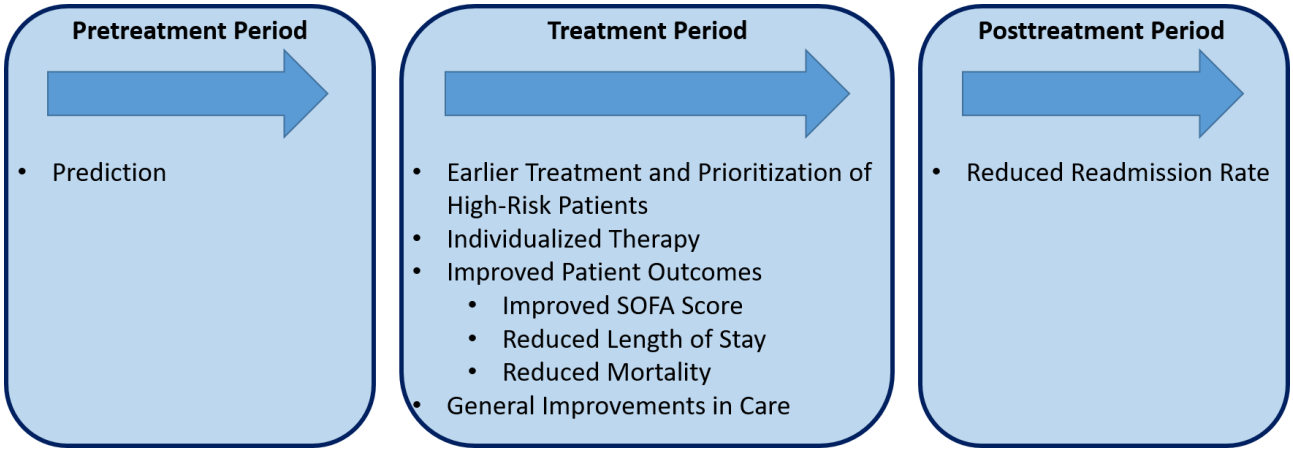

Pretreatment Period (Prediction)

Prediction of sepsis, septic shock, or sepsis-related organ dysfunction was addressed in 12 articles [31,33,35,37,38,40,42,45,48,51,53,56], comprising 7 quantitative studies (1 prospective [31] and 6 retrospective [33,35,40,42,45,53]), 1 qualitative study [37], 3 reviews [38,48,51], and 1 institutional news report [56]. The MLAs identified in this review indicate predictive capacity [48], which may be further optimized through algorithm fine-tuning [45]. Findings suggest that these models may predict sepsis between 4 and 48 hours prior to its onset [35,38,40,51,56], even before significant changes in vital or laboratory parameters become apparent [40]. MLAs were reported to support the identification of appropriate preventive measures [33]. Such predictions may have the potential to improve patient outcomes by providing timely warning of sepsis onset [31,37]. Beyond sepsis, the studies also reported the prediction of septic shock, with MLA predictions occurring between 4 and 7 hours before the onset of septic shock [42,51]. Compared with traditional rule-based routine screening protocols, predictive MLAs demonstrated superior early warning performance, identifying 58.6% more patients before organ dysfunction [53] and potentially contributing to a reduction in septic shock incidence [38]. Moreover, MLAs were shown to predict sepsis-related organ dysfunction approximately 7.5 hours earlier than rule-based routine screening protocols [53].

Treatment Period

Earlier Treatment and Prioritization of High-Risk Patients

Earlier treatment facilitated by MLAs was reported in 10 articles [28,31,33,36,37,40,49,52,53,57], including 7 quantitative studies (4 prospective [28,31,49,52] and 3 retrospective [33,40,53]), 1 mixed methods study [36], 1 qualitative study [37], and 1 institutional news report [57]. One retrospective study reported that the used MLA enabled earlier treatment up to 40 hours before the onset of sepsis [40], while another indicated that the use of ML may reduce the time to treatment, not providing a specific time reduction [33]. Earlier treatment was reported to enable intervention before or during clinical deterioration [53] and potentially prevent sepsis progression [52]. It may allow for early identification and intervention of patients at high risk for severe sepsis prior to clinical onset [31]. Additionally, the literature highlighted early identification and control of the pathogen causing sepsis [52]. Detecting patients before the onset of septic shock may facilitate earlier clinical assessment, diagnostic testing, therapeutic interventions, and transfer to appropriate levels of care [53]. Ultimately, earlier treatment may improve patient outcomes [37] and lead to an alteration in the prevalence of septic shock through timely intervention by health care providers from 5.3% in the control group to 1.5% in the experimental group (–71.7%) of the corresponding study [49]. MLAs were also associated with shorter times to obtain blood cultures (0.98 - 2.79 hours) [49,57], fluid administration (1.05 hours) [57] and earlier administration or adjustment of antibiotics (0.55 - 2.76) [28,36,49,57]. A positive correlation between timely evaluation of MLA alerts and quicker administration of antibiotics was reported, as earlier evaluation of alerts leads to faster use of antibiotics [36]. Finally, a quantitative prospective study reported that MLAs allow prioritization of high-risk patients, with the targeted real-time early warning system (TREWS) identifying who is most likely to benefit from timely treatment [28].

XSL•FO
RenderX

Individualized Therapy

MLAs were reported to support individualized, patient-centered therapy in 11 articles [29,30,33,39,41,44-47,50,54], comprising 10 quantitative studies (1 prospective [54] and 9 retrospective [29,30,33,39,41,45-47,50]) and 1 review [44]. Four main approaches for individualization were identified: (1) subgroup analyses and clustering, (2) optimized substance administration, (3) personalized nursing care, and (4) general statements. Three articles reported subgroup analyses and clustering of patients [41,45,46], which may enable hospitals to provide targeted treatments tailored to the specific needs of defined subgroups [45] and classify patients according to their diverging mortality risk due to factors such as fluid overload or norepinephrine overdose. Such classification might support the development of tailored resuscitation strategies for patients with septic shock [41]. Furthermore, subgroup analyses applied to populations with differing disease severity and progression allow MLAs to adjust the intensity of therapy [46]. The use of MLAs for optimal substance administration was reported in 6 articles [29,33,39,41,50,54]. Applications include personalized antibiotic dosing [33], faster adjustment to the most effective antibiotics, and drug resistance prediction. The comprehensive Intelligent Antimicrobial System demonstrated potentially faster drug resistance prediction times compared with conventional methods, requiring 39.8 hours for carbapenem-resistant Klebsiella pneumonia and 40.9 hours for methicillin-resistant Staphylococcus aureus, compared with 99.5 and 106.4 hours, respectively, using traditional methods [54]. MLA use was also associated with an 8% reduction in antibiotic resistance [50], may shorten the time to antimicrobial resistance detection by 37 hours [54], and was reported to reduce the duration of antibiotic treatment [50]. Furthermore, MLAs may support physicians in selecting appropriate antibiotic therapy [54], with AI-based antibiotic stewardship linked to decreased *Clostridium difficile* infections [50]. Beyond antibiotics, MLAs have demonstrated utility in optimizing dosing strategies for norepinephrine [41], vasopressors [39], corticosteroids [29], and fluid volume management [41]. MLAs also reported to enhance nursing competence and support more evidence-based, personalized nursing care [47]. General statements on individualized therapy were identified in 4 articles [30,33,44,54], including personalized treatment to support physicians in diagnosing and managing bacteremia [33], facilitation of shared decision-making through preoperative discussions [30], improved physician adherence [44], and more precise treatment tailored to individual patients [54].

Improved Patient Outcomes

Improved SOFA Score

Improved SOFA scores associated with the application of and timely response to MLAs were reported in 1 quantitative prospective study [28]. The SOFA score, the predominant measure for assessing the severity of organ dysfunction, is closely linked to the probability of mortality, with a higher score indicating an increased probability of death [3,58]. Using the TREWS algorithm, Adams et al [28] reported a SOFA score progression of –0.8 in their intervention group, compared to –0.4 in the control group. The article highlights a disproportionate reduction in the SOFA score for high-risk patients compared to nonhigh-risk patients. Additionally, timely evaluation and confirmation of the TREWS alerts is associated with improvements in SOFA score progression.

Reduced Length of Stay

Seven quantitative articles reported reductions in LOS [28,31,43,49,50,52,54], including 6 prospective [28,31,43,49,52,54] and 1 retrospective study [50]. Based on the identified literature, a distinction can be drawn between (1) specific reductions, reported in absolute or relative terms [28,31,43,49,52,54], and (2) general statements without precise quantification [28,49,50]. Reported specific reductions in hospital LOS ranged from 0.43 to 8.1 days [28,31,43,49,52], corresponding to decreases of 12.84%-45.25% [31,43,49,52]. Reported reductions in ICU LOS varied between 2.09 and 10.5 days [49,50]. One study also highlighted that shorter ICU stays may contribute to an overall reduction in hospital LOS, although LOS on the general ward increased by 2.4 days [50]. Another article reported a potential annual reduction of 1100 days in emergency department stays and the prevention of 34 ICU stays associated with MLA usage in the examined hospital [54]. General statements indicated a disproportionate, though not statistically significant, reduction in LOS among high-risk patients as well as reduced LOS when MLA-generated alarms were evaluated and confirmed timely [28]. AI-based antibiotic stewardship was also associated with shorter LOS [50] and MLAs were reported to significantly shorten hospital LOS compared to rule-based systems [49]. Furthermore, 1 study suggested that timely physician responses to MLA-generated alerts may contribute to reduced LOS [28].

Reduced Mortality

A reduction in mortality was reported in 14 articles [28,29,31-33,39,43,48,49,52,54-57], including 9 quantitative studies (6 prospective [28,31,43,49,52,54] and 3 retrospective [29,33,39]) as well as 3 reviews [32,48,55] and 2 institutional news reports [56,57]. Reported mortality reductions varied in type and presentation, encompassing (1) specific quantitative statements, expressed in relative or absolute terms [28,29,31,32,43,49,52,54,56,57], and (2) general statements without numerical specifications [28,29,33,39,48,49,54,55]. Relative reductions of mortality ranged from 13.19% to 74.94% [28,31,43,49,52,56,57], whereas absolute reductions ranged from 1.33% points to 26.4% points [28,29,31,43,49,52,54]. One study reported an increase in absolute survival rate of 11.7% and 23.7%, depending on the type of bacteria responsible for the sepsis [54]. Two articles provided reductions in natural numbers; one projected 22 potentially preventable annual deaths in the emergency department of the China Medical University Hospital [54], while another estimated several thousand preventable deaths in the United States alone [32]. General statements suggested that MLAs may disproportionately reduce mortality among high-risk patient cohorts, particularly when outputs are promptly evaluated and confirmed by physicians [28]. Improved survival rates may also be linked to the use of MLA-guided antibiotic recommendations [54] and the application of the 3PM (predictive, preventive, and personalized medicine) principles [33]. MLAs are associated with lower mortality compared to traditional physician assessments [29,39]

XSL•FO
RenderX

and predictions generated by rule-based tools [49]. Additionally, literature provided general statements, offering limited informational depth and indicating that the use of ML may contribute to reduced mortality [29,39,48,49,55].

General Improvements in Care

Eight articles reported improvements in care associated with MLAs [31,34,38,43,47-49,54], including 6 quantitative studies (5 prospective [31,34,43,49,54] and 1 retrospective [47]) as well as 2 reviews [38,48]. Reported benefits can be divided into 2 domains: (1) statements related to time and (2) statements on patient care enhancements. One study reported a reduced duration of septic shock [48]. Within the patient care enhancement category, MLAs were described as posing no risk to patients and offering potential benefits to patients and health care providers [31], reducing events of clinical deterioration [38], improving care accuracy [47,54], and increasing sepsis awareness among physicians [43,49]. Physicians and nurses also reported perceived improvements in care [34].

Posttreatment Period (Reduced Readmission Rate)

Predictive AI-based CDSS were associated with reduced 30-day readmission rates, as reported in 2 quantitative prospective studies [31,43]. In Burdick et al [31], implementation of an MLA reduced the 30-day readmissions from 36.4% to 28.12%, representing a 22.74% reduction compared to the baseline period. McCoy and Das [43] reported a decline from 46.19% (188/407) during the preimplementation baseline period to 29.8% (100/336) in a first postimplementation period and further to 25.2% (96/381) in a second postimplementation period. In a subsequent steady-state period, the 30-day readmission rate was further reduced to 7.84% (16/204). Across all surveyed months after implementation, the 30-day readmission rate was 23.03%, representing a 50.14% reduction in the sepsis-related 30-day readmission rate.

Discussion

Principal Findings

This scoping review presents the evidence on the patient-relevant benefits of AI-based CDSS in sepsis care. All articles focusing on sepsis and presenting the influence of AI-based CDSS on patient-relevant benefits, identified through the comprehensive search strategy, were included. In total, 30 articles were identified and integrated into the review. Investigating the literature, there is a number of AI-based CDSS for sepsis treatment developed in the past or currently under development. However, research typically has no or only limited reference to patient-relevant benefits and (1) mostly focuses on problems and/or success strategies [32,55,59-61] and/or (2) is indication-independent [59]. To the best of the authors' knowledge, this represents the first scoping review on this specific topic.

The findings of this scoping review, systematized into the 6 main categories, (1) prediction, (2) earlier treatment and prioritization of high-risk patients, (3) individualized therapy (which encompasses patient-centered care), (4) improved patient outcomes (which includes improved SOFA score, reduced length of stay, and reduced mortality), (5) general improvements in care, and (6) reduced readmission rate, underscore the potential patient-relevant benefits of AI-based CDSS in sepsis care across the entire inpatient pathway. The literature indicates that MLAs can potentially predict sepsis before its clinical onset [35,38,40,51,56]. Additionally, septic shock [42,51] and sepsis-related organ dysfunction [53] may be predicted in advance. These predictive capabilities can contribute to reducing the incidence of septic shock [38] and supporting decreased mortality rates among sepsis patients [12]. Sepsis prediction may facilitate timely treatment initiation through the use of MLAs [40]. This was associated with improved patient outcomes and a decreased prevalence of septic shock [49]. Furthermore, individualized therapy can potentially have a positive impact on patient-relevant benefits by reducing the time to treatment or LOS for each individual patient [33]. Moreover, the disproportionate reduction in the SOFA score through the use of ML compared to a control group whose treatment was not supported by MLAs should be mentioned. According to the Sepsis-3 definition, the level of the SOFA score positively correlates with the probability of death [3], and a SOFA score of ≥2 points corresponds to a mortality risk of over 10% in hospitalized patients outside the ICU [25]. The TREWS algorithm presented by Adams et al was able to reduce the SOFA score by 0.8 points, while a reduction of only 0.4 points was observed in the control group. Accordingly, the use of this MLA may contribute to the reduction in mortality. In general, the usage of MLAs was associated with a mortality reduction of up to 74.94% [28,31,43,49,52,56,57], with faster response times being associated with greater reductions in mortality [28]. This demonstrates the medical potential of ML in the treatment of sepsis, particularly when clinical recommendations are accepted and promptly implemented by physicians. With approximately 11 million deaths annually from sepsis according to the WHO [2], a corresponding reduction in mortality could translate into a substantial global health impact. Additionally, MLAs were linked to reduced hospital LOS [31,43,49,52] and ICU LOS [49,50]. Beyond their predictive capabilities, facilitation of timely treatment, mortality, and LOS reductions, AI-based CDSS in sepsis care provide further patient benefits, including shortened duration of septic shock [48], reduced antibiotic resistance, and reduced duration of antibiotic treatment [50]. MLAs may also contribute to a reduction of events of clinical deterioration [38] and increased physician awareness of sepsis [43,49]. Finally, the literature indicates that AI-based CDSS in sepsis care can contribute to reducing hospital readmission rates [43], further demonstrating their potential to improve patient-relevant benefits.

Comparison With Prior Work

While 6 reviews were included in this work, they primarily focused on other topics and predominantly used less systematic approaches [32,38,44,48,51,55]. Among the included reviews, 4 adopted a narrative review methodology [32,44,48,51]. By design, this approach is inherently less systematic than systematic reviews or scoping reviews, and this was evident in the search and selection process of the included narrative reviews. Two relied exclusively on limited, nonsystematic keyword searches, one using 8 keywords across 4 search engines [44] and another restricted to 3 keywords in a single database

XSL•FO
RenderX

[32]. Moreover, the review conducted by Ferreira et al [32] focused primarily on problems and success strategies related to AI-based CDSS, thereby addressing a different thematic focus than the present scoping review. Another narrative review applied a brief and partial search string without predefined inclusion and exclusion criteria and was limited to a single database [51], representing considerable methodological limitations relative to the present comprehensive scoping review. The narrative review conducted by Schinkel et al [48] adopted a more systematic approach, using a predefined search string and assessing the clinical value of AI-based systems by evaluating the AUROC as a criterion for article selection. While methodologically more robust, this review nonetheless differed from the present article, as it primarily evaluated the advantage of MLAs over rule-based scores, reflecting the status quo using a binary classifier. The advantage of MLAs over rule-based scores was considered a prerequisite for AI-based CDSS in the present study. With the exception of 1 review, where a manual search of reference lists was conducted [32], none of the narrative reviews [44,48,51] undertook a comprehensive search for gray literature or an analysis of the reference lists. Furthermore, only 1 narrative review reported a screening process conducted by 2 independent reviewers [48], whereas the other 3 reviews did not provide methodological detail [32,44,51]. By contrast, the present scoping review implemented a rigorous screening process with 2 independent reviewers to enhance objectivity, reliability, and reproducibility. Beyond these narrative reviews, 1 study followed an integrative review approach, explicitly focusing on predictive algorithms and embedding this narrow focus within a brief predefined search string [38]. In contrast, the present exploratory scoping review aimed to inductively derive patient benefit categories associated with AI-based CDSS in sepsis care. This integrative review relied on a single reviewer for screening [38], representing a methodological limitation in comparison with the dual-reviewer approach of the present scoping review. Finally, 1 systematic review included in this study used a largely rigorous and systematic methodology, with the notable exception of a gray literature search, which was not reported. In addition, this systematic review focused primarily on problems and success strategies [55], thereby diverging from the present scoping reviews' explicit focus on patient-relevant benefits. In sum, the present scoping review can be clearly distinguished from the included reviews both methodologically and thematically. By applying a comprehensive, exploratory design, centered on patient-relevant benefits, it makes a substantive and valuable contribution to closing the research gap regarding patient-relevant benefits of AI-based CDSS in sepsis care.

Implications and Recommendations

Patient-relevant benefits identified in the literature are not sufficient to ensure successful implementation of AI-based CDSS. Equally critical is the acceptance of the underlying technology by health care providers and their belief that its use possesses tangible benefits. The unified theory of acceptance and use of technology (UTAUT) provides a framework to understand factors influencing behavioral intention and use behavior using four constructs: (1) performance expectancy, (2) effort expectancy, (3) social influence, and (4) facilitating conditions. In this context, effective design of AI-based CDSS should ensure that providers perceive the system as both beneficial and easy to use, corresponding to the first 2 constructs of the UTAUT. Specifically, (1) users should believe that using AI-based CDSS enhances gains in job performance, and (2) the system is intuitive and easy to operate. Equally important are contextual factors: health care providers should perceive that (3) important others endorse system use, and (4) organizational and technical infrastructure is in place to support usage [61]. A meta-analysis by Dingel et al [62] applying the UTAUT to health care practitioners' intention to use AI-enabled CDSS confirms that implementation must address not only technical and organizational aspects but also psychological and social factors, particularly fostering user trust. Successful implementation of AI-based CDSS therefore depends only partly on system performance; it is largely contingent on user attitudes and framework conditions.

Beyond the 4 UTAUT constructs, specific barriers [63] and facilitators [64] must be considered when evaluating AI-based CDSS. A nuanced understanding of these factors is essential to accurately evaluate the potential impact of AI-based CDSS on sepsis care. The current evidence demonstrates a pronounced lack of prospective studies investigating the optimal integration of such systems [29]. This paucity of implementation-oriented research, coupled with limited clinician acceptance [37] and insufficient knowledge of AI among health care providers [65], constitutes a substantial barrier to clinical adoption. Concurrently, extant literature highlights pivotal facilitators, emphasizing the importance of prioritizing research on effective integration strategies [38]. For instance, low acceptance may be mitigated by involving health care providers directly in the design and development of CDSS [66-68], while targeted training and educational programs could address knowledge gaps among service providers and enhance trust in this technology [37,67]. These factors must therefore be carefully considered by all stakeholders involved in implementation (eg, caregivers, physicians, and researchers) before real-world adoption can occur. For clinicians, the findings provide insights into realistic benefits, current limitations, and evidence gaps that may guide expectations in clinical decision-making. For researchers, this review underscores the importance of conducting prospective studies and fostering user-centered development to ensure that CDSS effectively translate into clinical practice. In addition, although patient-relevant benefits—and not only measurable patient-relevant outcomes—have been investigated, the findings may contribute to the development of a consistent set of generic patient-relevant outcomes, as proposed by Kersting et al [69]. This could, in turn, facilitate a shared understanding and enhance comparability across studies targeting patient-relevant outcomes, particularly given the absence of a clear, widely accepted definition and standardized criteria for selecting such outcomes.

Strengths and Limitations

This scoping review was conducted by an interdisciplinary team comprising computer scientists, physicians, statisticians, and (health) economists. This diverse expertise facilitated a comprehensive examination of all relevant aspects across these fields, ensuring a thorough evaluation of the reviewed literature.

XSL•FO
RenderX

To address the interdisciplinary research question comprehensively, 4 databases focusing on medicine and informatics were included in the review. In addition, the structured search for gray literature targeted various institutions in 8 countries, including each country's Ministry of Health, diverse sepsis and intensive care associations for each country, and diverse health informatics associations for each country. The 8 countries were selected based on two criteria: (1) having highly digitalized health care systems and/or (2) holding the status of industrialized nations. These countries were presumed to have a higher likelihood of using AI-based systems. Furthermore, the search encompassed internationally active stakeholders, such as the WHO and OECD, alongside globally active health informatics organizations and sepsis and critical care associations.

Despite all efforts, this scoping review is not free of limitations. Given the exploratory nature of the methodology, publication bias must be considered a potential limitation [70]. Studies in which AI-based CDSS do not demonstrate improvements in patient-relevant benefits compared with conventional scores may not be submitted in peer-reviewed journals, potentially leading to an overestimation of their true patient benefit. Although no formal risk of bias assessment was conducted, the included studies demonstrated considerable heterogeneity in design. Moreover, 65% (13/20) of the quantitative studies relied solely on retrospective methodologies, in which evidence of patient benefits was demonstrated only theoretically. Consequently, the findings of this scoping review should be interpreted with caution, as the reported effects may be overestimated in the context of real-world care. The overall strength of evidence was limited by the predominance of retrospective study designs and the theoretical nature of reported benefits. In contrast, the included prospective studies provided more robust support for the identified benefit categories. Importantly, each benefit category has been substantiated in prospective studies, thereby affirming its validity in real-world clinical contexts rather than solely theoretically in retrospective or descriptive studies (Multimedia Appendix 9). Detailed information on the study designs of all included articles is provided in Multimedia Appendix 6. Multimedia Appendix 9 summarizes benefit categories identified across the respective study designs. Furthermore, the comparability of the reported MLA performance across articles is limited due to varying definitions of sepsis (eg, different causative pathogens, divergent sepsis definitions, and variations in the examined indications such as sepsis, septic shock, or sepsis-related organ dysfunction). The same limitation applies to the databases used for training and validation, which differed substantially in size. In addition, no assessment of the applied MLA methods was conducted, nor was the level of maturity of the individual MLAs explicitly considered. Furthermore, due to the heterogeneity of the included studies, no formal quality assessment was conducted. Rather, the present review was designed to exploratively map and comparatively present the entirety of available evidence in order to identify research gaps, without imposing methodological restrictions on the literature to be included [20-22]. Finally, a methodological limitation should be noted: Research conducted on institutional websites could only be partially conducted for Estonia, Denmark, and Spain due to language restrictions (English and German), as some stakeholder websites were available exclusively in the respective national languages. The utilization of translation tools was deliberately avoided, as the inclusion of material that none of the authors could fully comprehend and critically appraise in the original language was considered methodologically inappropriate.

The search restriction of 15 years should not be considered a limitation. The inclusion period was defined in consultation with ML experts (NT, HN), and algorithms developed prior to the review period (January 1, 2008-March 2, 2023) were predominantly anticipated to be (1) rule-based systems and/or nonblack-box systems for the users. Both types of algorithms are outside the scope of this review. Moreover, an initial limited search in the databases MEDLINE via PubMed and Embase, which accounted for approximately 75% of the screened literature (Figure 1), indicated that only a marginal proportion of articles relevant to the research question were published before 2008. Consequently, the time restriction is unlikely to have affected the identification of relevant literature.

Conclusion

The findings of this scoping review highlight the considerable medical relevance of AI-based CDSS in sepsis care. These systems offer benefits across the entire patient care pathway, from early detection and risk stratification to individualized therapy and various improved outcomes. AI-based CDSS has shown the ability to predict sepsis, septic shock, and sepsis-related organ dysfunction, enabling earlier initiation of treatment, prioritization of high-risk patients, and tailored therapeutic strategies. In addition to supporting earlier and more targeted interventions, AI-based CDSS contribute to better clinical outcomes, including improved SOFA scores, reduced LOS both in general wards and ICUs, and lower mortality rates. They may also help reduce readmission rates among sepsis patients, further enhancing long-term care quality. With their transformative potential, AI-based CDSS could fundamentally improve the global management of sepsis. However, further research is needed to optimize the development, implementation, and clinical application of these systems to maximize patient benefits and further improve outcomes for sepsis patients in the future. This is particularly important given the highly heterogeneous evidence base, with a substantial proportion of studies relying on retrospective data, as the results of the included studies cannot be directly generalized or applied without caution.

Acknowledgments

The authors acknowledge the support of the Open Access Publication Fund of the University of Duisburg-Essen and thank Mrs Katrin Wibker for her assistance in quality control during the development of the search string for the scoping review.

XSL•FO
RenderX

Funding

This research is part of a wider research project (Multimethod Research Project on Requirements for Artificial Intelligence-Based Clinical Decision Support Systems Using the Medical Example of Sepsis (="“KI@work”")) funded by the German Federal Joint Committee (funding code: 01VSF22050 –KI@work). The funders had no influence in the study design, conduct of the study, or the decision to publish or prepare the manuscript.

Authors' Contributions

Conceptualization: PR (lead), GDG (equal), NB (equal), MA (supporting), HN (supporting), LP (supporting), PH (supporting), NT (supporting), MT (supporting), FMB (supporting), JW (supporting)
Data curation: PR (lead), GDG (equal)
Formal analysis: PR (lead), GDG (supporting), NB (supporting)
Funding acquisition: GDG, JW, MA, HN, NT, MT, FMB, NB
Investigation: PR (lead), GDG (equal)
Methodology: PR (lead), GDG (supporting), NB (supporting), NT (supporting), HN (supporting)
Project administration: NB
Supervision: NB
Validation: PR (lead), GDG (equal), NB (equal)
Visualization: PR
Writing – original draft: PR
Writing – review & editing: PR (lead), GDG (supporting), NB (supporting), MA (supporting), HN (supporting), LP (supporting), PH (supporting), NT (supporting), MT (supporting), FMB (supporting), JW (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy – Medline via PubMed.
[DOCX File, 14 KB - jmir_v28i1e76772_app1.docx]

Multimedia Appendix 2
Search strategy – Embase.
[DOCX File, 14 KB - jmir_v28i1e76772_app2.docx]

Multimedia Appendix 3
Search strategy – ACM Digital Library.
[DOCX File, 14 KB - jmir_v28i1e76772_app3.docx]

Multimedia Appendix 4
Search strategy – IEEE Xplore.
[DOCX File, 15 KB - jmir_v28i1e76772_app4.docx]

Multimedia Appendix 5
Included websites in structured search.
[DOCX File, 20 KB - jmir_v28i1e76772_app5.docx]

Multimedia Appendix 6
Overview of included articles.
[DOCX File, 39 KB - jmir_v28i1e76772_app6.docx]

Multimedia Appendix 7
Patient benefits mentioned in articles.
[DOCX File, 15 KB - jmir_v28i1e76772_app7.docx]

Multimedia Appendix 8
Screened articles.
[DOCX File, 34 KB - jmir_v28i1e76772_app8.docx]

XSL•FO
RenderX

Multimedia Appendix 9
Used methods per benefit category.
[DOCX File, 13 KB - jmir_v28i1e76772_app9.docx]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 108 KB - jmir_v28i1e76772_app10.docx]

References

1. Rhee C, Jones TM, Hamad Y, et al. Prevalence, underlying causes, and preventability of sepsis-associated mortality in US acute care hospitals. JAMA Netw Open 2019 Feb 1;2(2):e187571. [doi: 10.1001/jamanetworkopen.2018.7571] [Medline: 30768188]
2. Global report on the epidemiology and burden of sepsis current evidence, identifying gaps and future directions. World Health Organization. 2020. URL: https://iris.who.int/server/api/core/bitstreams/d4ce3613-bf94-4205-85c8-14f3fc0609db/content [accessed 2025-12-04]
3. Singer M, Deutschman CS, Seymour CW, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 2016 Feb 23;315(8):801-810. [doi: 10.1001/jama.2016.0287] [Medline: 26903338]
4. Ericson O, Hjelmgren J, Sjövall F, Söderberg J, Persson I. The potential cost and cost-effectiveness impact of using a machine learning algorithm for early detection of sepsis in intensive care units in Sweden. J Health Econ Outcomes Res 2022;9(1):101-110. [doi: 10.36469/jheor.2022.33951] [Medline: 35620451]
5. Arriaga-Pizano LA, Gonzalez-Olvera MA, Ferat-Osorio EA, et al. Accurate diagnosis of sepsis using a neural network: Pilot study using routine clinical variables. Comput Methods Programs Biomed 2021 Oct;210:106366. [doi: 10.1016/j.cmpb.2021.106366] [Medline: 34500141]
6. Vincent JL, Martin GS, Levy MM. qSOFA does not replace SIRS in the definition of sepsis. Crit Care 2016 Jul 17;20(1):210. [doi: 10.1186/s13054-016-1389-z] [Medline: 27423462]
7. Dellinger RP, Levy MM, Rhodes A, et al. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2012. Crit Care Med 2013 Feb;41(2):580-637. [doi: 10.1097/CCM.0b013e31827e83af] [Medline: 23353941]
8. Kumar A, Roberts D, Wood KE, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Crit Care Med 2006 Jun;34(6):1589-1596. [doi: 10.1097/01.CCM.0000217961.75225.E9] [Medline: 16625125]
9. Brunkhorst FM, Weigand MA, Pletz M, et al. S3 Guideline Sepsis-prevention, diagnosis, therapy, and aftercare: long version. Med Klin Intensivmed Notfmed 2020 May;115(Suppl 2):37-109. [doi: 10.1007/s00063-020-00685-0] [Medline: 32356041]
10. Sakr Y, Jaschinski U, Wittebole X, et al. Sepsis in intensive care unit patients: worldwide data from the intensive care over nations audit. Open Forum Infect Dis 2018 Dec;5(12):ofy313. [doi: 10.1093/ofid/ofy313] [Medline: 30555852]
11. Arwyn-Jones J, Brent AJ. Sepsis. Surgery (Oxford) 2019 Jan;37(1):1-8. [doi: 10.1016/j.mpsur.2018.11.007]
12. Boussina A, Shashikumar SP, Malhotra A, et al. Impact of a deep learning sepsis prediction model on quality of care and survival. NPJ Digit Med 2024 Jan 23;7(1):14. [doi: 10.1038/s41746-023-00986-6] [Medline: 38263386]
13. Alanazi A, Aldakhil L, Aldhoayan M, Aldosari B. Machine learning for early prediction of sepsis in intensive care unit (ICU) patients. Med Bogota Colomb 2023 Jul 9;59(7):1276. [doi: 10.3390/medicina59071276]
14. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med 2018 Apr;46(4):547-553. [doi: 10.1097/CCM.0000000000002936] [Medline: 29286945]
15. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. US Food and Drug Administration. 2025. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices [accessed 2025-04-29]
16. O'Sullivan C, Tsai DHT, Wu ICY, et al. Machine learning applications on neonatal sepsis treatment: a scoping review. BMC Infect Dis 2023 Jun 29;23(1):441. [doi: 10.1186/s12879-023-08409-3] [Medline: 37386442]
17. Heneghan JA, Walker SB, Fawcett A, et al. The pediatric data science and analytics subgroup of the pediatric acute lung injury and sepsis investigators network: use of supervised machine learning applications in pediatric critical care medicine research. Pediatr Crit Care Med 2024 Apr 1;25(4):364-374. [doi: 10.1097/PCC.0000000000003425] [Medline: 38059732]
18. Susanto AP, Lyell D, Widyantoro B, Berkovsky S, Magrabi F. Effects of machine learning-based clinical decision support systems on decision-making, care delivery, and patient outcomes: a scoping review. J Am Med Inform Assoc 2023 Nov 17;30(12):2050-2063. [doi: 10.1093/jamia/ocad180] [Medline: 37647865]
19. Wan YKJ, Wright MC, McFarland MM, et al. Information displays for automated surveillance algorithms of in-hospital patient deterioration: a scoping review. J Am Med Inform Assoc 2023 Dec 22;31(1):256-273. [doi: 10.1093/jamia/ocad203] [Medline: 37847664]

XSL•FO
RenderX

20. Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, editors. JBI Manual for Evidence Synthesis JBI 2024. [doi: 10.46658/JBIMES-24-01]
21. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
22. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5(5):69. [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
23. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
24. Raszke P, Giebel GD, Abels C, et al. User-oriented requirements for artificial intelligence-based clinical decision support systems in sepsis: protocol for a multimethod research project. JMIR Res Protoc 2025 Jan 30;14:e62704. [doi: 10.2196/62704] [Medline: 39883929]
25. S3-leitlinie: sepsis – prävention, diagnose, therapie und nachsorge – update 2025 [Website in German]. AWMF-register der leitlinien der wissenschaftlichen medizinischen fachgesellschaften. 2025. URL: https://register.awmf.org/assets/guidelines/079-001l_S3_Sepsis-Praevention-Diagnose-Therapie-Nachsorge_2025-07.pdf [accessed 2025-12-12]
26. S2k-leitlinie 024-025 „sepsis bei kindern jenseits der neonatalperiode [Website in German]. AWMF-Register der Leitlinien der wissenschaftlichen medizinischen Fachgesellschaften. 2015. URL: https://register.awmf.org/assets/guidelines/024-025l_S2k_Sepsis_nach_Neonatalperiode_2016-04-abgelaufen.pdf [accessed 2025-01-20]
27. S2k-leitlinie 024-008: bakterielle infektionen bei neugeborenen [Website in German]. AWMF-Register der Leitlinien der wissenschaftlichen medizinischen Fachgesellschaften. 2021. URL: https://register.awmf.org/assets/guidelines/024-008l_S2k_Bakterielle_Infektionen_Neugeborene_2021-03-abgelaufne.pdf [accessed 2025-01-20]
28. Adams R, Henry KE, Sridharan A, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat Med 2022 Jul;28(7):1455-1460. [doi: 10.1038/s41591-022-01894-0] [Medline: 35864252]
29. Bologheanu R, Kapral L, Laxar D, et al. Development of a reinforcement learning algorithm to optimize corticosteroid therapy in critically ill patients with sepsis. J Clin Med 2023 Feb 14;12(4):1513. [doi: 10.3390/jcm12041513] [Medline: 36836046]
30. Bunn C, Kulshrestha S, Boyda J, et al. Application of machine learning to the prediction of postoperative sepsis after appendectomy. Surgery 2021 Mar;169(3):671-677. [doi: 10.1016/j.surg.2020.07.045] [Medline: 32951903]
31. Burdick H, Pino E, Gabel-Comeau D, et al. Effect of a sepsis prediction algorithm on patient mortality, length of stay and readmission: a prospective multicentre clinical outcomes evaluation of real-world patient data from US hospitals. BMJ Health Care Inform 2020 Apr;27(1):e100109. [doi: 10.1136/bmjhci-2019-100109] [Medline: 32354696]
32. Ferreira LD, McCants D, Velamuri S. Using machine learning for process improvement in sepsis management. J Healthc Qual Res 2023;38(5):304-311. [doi: 10.1016/j.jhqr.2022.09.006] [Medline: 36319584]
33. Garnica O, Gómez D, Ramos V, Hidalgo JI, Ruiz-Giardín JM. Diagnosing hospital bacteraemia in the framework of predictive, preventive and personalised medicine using electronic health records and machine learning classifiers. EPMA J 2021 Sep;12(3):365-381. [doi: 10.1007/s13167-021-00252-3] [Medline: 34484472]
34. Ginestra JC, Giannini HM, Schweickert WD, et al. Clinician perception of a machine learning-based early warning system designed to predict severe sepsis and septic shock. Crit Care Med 2019 Nov;47(11):1477-1484. [doi: 10.1097/CCM.0000000000003803] [Medline: 31135500]
35. Goh KH, Wang L, Yeow AYK, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nat Commun 2021 Jan 29;12(1):711. [doi: 10.1038/s41467-021-20910-4] [Medline: 33514699]
36. Henry KE, Adams R, Parent C, et al. Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. Nat Med 2022 Jul;28(7):1447-1454. [doi: 10.1038/s41591-022-01895-z] [Medline: 35864251]
37. Joshi M, Mecklai K, Rozenblum R, Samal L. Implementation approaches and barriers for rule-based and machine learning-based sepsis risk prediction tools: a qualitative study. JAMIA Open 2022 Jul;5(2):ooac022. [doi: 10.1093/jamiaopen/ooac022] [Medline: 35474719]
38. Kausch SL, Moorman JR, Lake DE, Keim-Malpass J. Physiological machine learning models for prediction of sepsis in hospitalized adults: an integrative review. Intensive Crit Care Nurs 2021 Aug;65:103035. [doi: 10.1016/j.iccn.2021.103035] [Medline: 33875337]
39. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 2018 Nov;24(11):1716-1720. [doi: 10.1038/s41591-018-0213-5] [Medline: 30349085]
40. Kuo YY, Huang ST, Chiu HW. Applying artificial neural network for early detection of sepsis with intentionally preserved highly missing real-world data for simulating clinical situation. BMC Med Inform Decis Mak 2021 Oct 22;21(1):290. [doi: 10.1186/s12911-021-01653-0] [Medline: 34686163]
41. Ma P, Liu J, Shen F, et al. Individualized resuscitation strategy for septic shock formalized by finite mixture modeling and dynamic treatment regimen. Crit Care 2021 Jul 12;25(1):243. [doi: 10.1186/s13054-021-03682-7]

XSL•FO
RenderX

42. Mao Q, Jay M, Hoffman JL, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. BMJ Open 2018 Jan 26;8(1):e017833. [doi: 10.1136/bmjopen-2017-017833] [Medline: 29374661]
43. McCoy A, Das R. Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. BMJ Open Qual 2017;6(2):e000158. [doi: 10.1136/bmjoq-2017-000158] [Medline: 29450295]
44. Ocampo-Quintero N, Vidal-Cortés P, del Río Carbajo L, Fdez-Riverola F, Reboiro-Jato M, Glez-Peña D. Enhancing sepsis management through machine learning techniques: a review. Medicina Intensiva (English Edition) 2022 Mar;46(3):140-156. [doi: 10.1016/j.medine.2020.04.015]
45. Rogers P, Boussina AE, Shashikumar SP, Wardi G, Longhurst CA, Nemati S. Optimizing the implementation of clinical predictive models to minimize national costs: sepsis case study. J Med Internet Res 2023 Feb 13;25:e43486. [doi: 10.2196/43486] [Medline: 36780203]
46. Roggeveen L, El Hassouni A, Ahrendt J, et al. Transatlantic transferability of a new reinforcement learning model for optimizing haemodynamic treatment for critically ill patients with sepsis. Artif Intell Med 2021 Feb;112:102003. [doi: 10.1016/j.artmed.2020.102003] [Medline: 33581824]
47. Scherer JDS, Pereira JS, Debastiani MS, Bica CG. Beyond technology: can artificial intelligence support clinical decisions in the prediction of sepsis? Rev Bras Enferm 2022;75(5):e20210586. [doi: 10.1590/0034-7167-2021-0586] [Medline: 35584427]
48. Schinkel M, Paranjape K, Nannan Panday RS, Skyttberg N, Nanayakkara PWB. Clinical applications of artificial intelligence in sepsis: a narrative review. Comput Biol Med 2019 Dec;115:103488. [doi: 10.1016/j.compbiomed.2019.103488] [Medline: 31634699]
49. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. BMJ Open Respir Res 2017;4(1):e000234. [doi: 10.1136/bmjresp-2017-000234] [Medline: 29435343]
50. Voermans AM, Mewes JC, Broyles MR, Steuten LMG. Cost-effectiveness analysis of a procalcitonin-guided decision algorithm for antibiotic stewardship using real-world U.S. hospital data. OMICS 2019 Oct;23(10):508-515. [doi: 10.1089/omi.2019.0113] [Medline: 31509068]
51. Wu M, Du X, Gu R, Wei J. Artificial intelligence for clinical decision support in sepsis. Front Med (Lausanne) 2021;8:665464. [doi: 10.3389/fmed.2021.665464] [Medline: 34055839]
52. Burdick H, Pino E, Gabel-Comeau D, et al. Evaluating a sepsis prediction machine learning algorithm in the emergency department and intensive care unit: a before and after comparative study. Clinical Trials. . [doi: 10.1101/224014]
53. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med 2015 Aug 5;7(299):299ra122. [doi: 10.1126/scitranslmed.aab3719] [Medline: 26246167]
54. Hsu KC, Cho DY, Hsueh PR, Yu J, Sun PR. Development of a comprehensive intelligent antimicrobial system: an epochal, fast, and digitally precise prediction of therapeutic antibiotics case study. China Medical University Hospital, Taiwan 2023 [FREE Full text]
55. van der Vegt AH, Scott IA, Dermawan K, Schnetler RJ, Kalke VR, Lane PJ. Deployment of machine learning algorithms to predict sepsis: systematic review and application of the SALIENT clinical AI implementation framework. J Am Med Inform Assoc 2023 Jun 20;30(7):1349-1361. [doi: 10.1093/jamia/ocad075] [Medline: 37172264]
56. Dascena pledges support for sepsis allianc clinical community. Sepsis alliance. 2021. URL: https://www.sepsis.org/news/dascena-pledges-support-for-sepsis-alliance-clinical-community/ [accessed 2025-01-20]
57. Saving Lives With Enhanced Sepsis Treatment Protocol. Healthcare information and management systems society (HIMSS). 2020. URL: https://oregon.himss.org/news/saving-lives-enhanced-sepsis-treatment-protocol [accessed 2025-01-20]
58. Vincent JL, de Mendonca A, Cantraine F, et al. Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units. Crit Care Med 1998 Nov;26(11):1793-1800. [doi: 10.1097/00003246-199811000-00016]
59. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17. [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
60. Moor M, Rieck B, Horn M, Jutzeler CR, Borgwardt K. Early prediction of sepsis in the ICU using machine learning: a systematic review. Front Med (Lausanne) 2021;8:607952. [doi: 10.3389/fmed.2021.607952] [Medline: 34124082]
61. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view1. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]
62. Dingel J, Kleine AK, Cecil J, Sigl AL, Lermer E, Gaube S. Predictors of health care practitioners' intention to use AI-enabled clinical decision support systems: meta-analysis based on the unified theory of acceptance and use of technology. J Med Internet Res 2024 Aug 5;26:e57224. [doi: 10.2196/57224] [Medline: 39102675]
63. Giebel GD, Raszke P, Nowak H, et al. Problems and barriers related to the use of AI-based clinical decision support systems: interview study. J Med Internet Res 2025 Feb 3;27:e63377. [doi: 10.2196/63377] [Medline: 39899342]

XSL•FO
RenderX

64. Giebel GD, Raszke P, Nowak H, et al. Improving AI-based clinical decision support systems and their integration into care from the perspective of experts: interview study among different stakeholders. JMIR Med Inform 2025 Jul 7;13:e69688. [doi: 10.2196/69688] [Medline: 40623684]
65. Sandhu S, Lin AL, Brajer N, et al. Integrating a machine learning system into clinical workflows: qualitative study. J Med Internet Res 2020 Nov 19;22(11):e22421. [doi: 10.2196/22421] [Medline: 33211015]
66. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. Manage Sci 1989 Aug;35(8):982-1003. [doi: 10.1287/mnsc.35.8.982]
67. Manetti S, Cumetti M, De Benedictis A, Lettieri E. Adoption of novel biomarker test parameters with machine learning-based algorithms for the early detection of sepsis in hospital practice. J Nurs Manag 2022 Nov;30(8):3754-3764. [doi: 10.1111/jonm.13807] [Medline: 36125938]
68. Sendak M, Elish MC, Gao M, et al. "The human body is a black box": supporting clinical decision-making with deep learning. 2020 Presented at: FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; Jan 27-30, 2020; Barcelona, Spain p. 99-109 URL: https://dl.acm.org/doi/10.1145/3351095.3372827 [accessed 2026-01-02]
69. Kersting C, Kneer M, Barzel A. Patient-relevant outcomes: what are we talking about? A scoping review to improve conceptual clarity. BMC Health Serv Res 2020 Jun 29;20(1):596. [doi: 10.1186/s12913-020-05442-9] [Medline: 32600321]
70. Nair AS. Publication bias - importance of studies with negative results!. Indian J Anaesth 2019 Jun;63(6):505-507. [doi: 10.4103/ija.IJA_142_19] [Medline: 31263309]

Abbreviations

3PM: predictive, preventive, and personalized medicine
AI: artificial intelligence
APACHE II: Acute Physiology and Chronic Health Evaluation II
CDSS: clinical decision support system
ICU: intensive care unit
LOS: length of stay
ML: machine learning
MLA: machine learning algorithm
NEWS2: National Early Warning Score 2
OECD: Organisation for Economic Co-operation and Development
PCC: Population, Concept, Context
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
qSOFA score: Quick Sequential Organ Failure Assessment Score
SIRS: systemic inflammatory response syndrome
SOFA score: Sequential Organ Failure Assessment Score
TREWS: targeted real-time early warning system
UTAUT: unified theory of acceptance and use of technology
WHO: World Health Organization

Edited by J Sarvestan; submitted 30.Apr.2025; peer-reviewed by AG Gad, A Staffini, S Meister; revised version received 25.Nov.2025; accepted 26.Nov.2025; published 26.Jan.2026.

Please cite as:
Raszke P, Giebel GD, Wasem J, Adamzik M, Nowak H, Palmowski L, Heinz P, Timmesfeld N, Tokic M, Brunkhorst FM, Blase N
Patient Benefits in the Context of Sepsis-Related AI-Based Clinical Decision Support Systems: Scoping Review
J Med Internet Res 2026;28:e76772
URL: https://www.jmir.org/2026/1/e76772
doi:10.2196/76772

© Pascal Raszke, Godwin Denk Giebel, Jürgen Wasem, Michael Adamzik, Hartmuth Nowak, Lars Palmowski, Philipp Heinz, Nina Timmesfeld, Marianne Tokic, Frank Martin Brunkhorst, Nikola Blase. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Diagnostic Performance of Deep Learning and Radiomics in Extracranial Carotid Plaque Detection: Systematic Review and Meta-Analysis

Lingjie Ju[1*], MM; Yongsheng Guo[1*], MM; Haiyong Guo[2], MM; Ruijuan Liu[1], MM; Yiyang Wang[3], MM; Siyu Wang[3], MD; Na Ma[3], MD; Junhong Ren[1], MM

[1]Department of Sonography, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, 1 Dahua Road, Dongdan, Dongcheng District, Beijing, China

[2]Department of Sonography, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Peking University Health Science Center, Beijing, China

[3]Department of Sonography, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Junhong Ren, MM
Department of Sonography, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, 1 Dahua Road, Dongdan, Dongcheng District, Beijing, China

Abstract

Background: Artificial intelligence–enhanced imaging techniques have demonstrated promising diagnostic potential for carotid plaques, a key cardiovascular and cerebrovascular risk factor. However, previous studies did not systematically synthesize their diagnostic accuracy.

Objective: This study aimed to quantitatively explore the diagnostic efficacy of deep learning (DL) and radiomics for extracranial carotid plaques and establish a standardized framework for improving plaque detection.

Methods: We searched the PubMed, Embase, Cochrane, Web of Science, and Institute of Electrical and Electronics Engineers databases to identify studies involving the use of radiomics or DL models to diagnose extracranial carotid artery plaques from inception up to September 24, 2025. The quality of the studies was determined using Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence (QUADAS-AI). A meta-analysis was conducted using StataMP (version 17.0; StataCorp) with a bivariate mixed-effects model to calculate pooled sensitivity and specificity, generate summary receiver operating characteristic (SROC) curves, assess Cochran Q statistic and I^2-based heterogeneity, and conduct subgroup analyses and regression analysis.

Results: Among 40 studies comprising 17,246 patients, 34 integrated independent test sets or validation sets in the quantitative statistical analysis. Among them, 24 focused on DL models, 10 on machine learning models based on radiomics. The combined sensitivity, specificity, and area under the SROC curve were 0.88 (95% CI 0.85 - 0.91; P<.001; I^2=93.58%), 0.89 (95% CI 0.85 - 0.92; P<.001; I^2=91.38%), and 0.95 (95% CI 0.92 - 0.96), respectively. Compared with the machine learning models based on radiomics algorithms, DL models achieved comparable improvements in specificity and area under the SROC curve. It was observed that transfer learning and a large sample size enhanced the diagnostic performance of models. Models used to identify plaque stability and presence had similar diagnostic performances, both of which were more effective in identifying symptomatic plaque models. A total of 7 studies demonstrated that the models that combined clinical features exhibited comparable diagnostic capability to pure DL and radiomics models. Additionally, 7 studies performed external validation, obtaining lower diagnostic performance than in testing groups. Limited regression analysis failed to identify significant sources of heterogeneity, and the limited number of eligible studies restricted more comprehensive subgroup analyses. The high heterogeneity in the study results may be due to different scanning parameters, model architecture, image segmentation, and algorithms.

Conclusions: Radiomics algorithms and DL models can effectively diagnose extracranial carotid plaque. However, there are concerns regarding irregularities in research design and the absence of multicenter studies and external validation. Future research should aim to reduce bias risk and enhance the generalizability and clinical orientation of the models.

(J Med Internet Res 2026;28:e77092) doi:10.2196/77092

KEYWORDS

extracranial carotid plaque; deep learning; radiomics; systematic review; diagnosis

XSL•FO
RenderX

Introduction

Extracranial carotid plaques are biomarkers of coronary artery disease and cerebral ischemic events, including ischemic heart disease and stroke. The global prevalence of carotid plaques among individuals aged 30 - 79 years is estimated at 21.1% (n=815.76 million) in 2020. This high prevalence reflects a growing global burden of cardiovascular and cerebrovascular diseases, posing a significant challenge to public health systems [1]. Therefore, early detection and management of carotid plaque can potentially reduce the risk of stroke and cardiovascular events [2-4], and thus, effective detection and classification technologies need to be prioritized.

Imaging methods for carotid plaque imaging, such as ultrasound, computed tomography angiography (CTA), magnetic resonance imaging (MRI), and digital subtraction angiography, facilitate detection, stenosis assessment, and plaque composition analysis [5]. Conventional ultrasound is the first-line screening method [6]. Studies show that periapical radiographs (PRs) can serve as a supplementary screening tool, demonstrating a 50% concordance with ultrasound or CTA [7-9]. Current imaging primarily identifies high-risk features, such as plaque neovascularity, lipid-rich necrotic cores, thin fibrous caps, and intraplaque hemorrhage plaque ulceration [4,10]. Among them, the contrast-enhanced ultrasound or superb microvascular imaging can accurately quantify neovascularization and correlates well with histopathology [11-14], offering rapid, noninvasive, and reliable quantification [15]. It is proficient in vascular imaging and ulcer detection [16], as well as stenosis assessment [17], but it faces challenges with small lipid cores and thin fibrous caps [18]. MRI remains the gold standard for assessing plaque composition, particularly for identifying lipid cores and intraplaque hemorrhage [19]. While digital subtraction angiography is the reference standard, its invasive nature limits its application. Notably, the accuracy of these diagnostic techniques largely relies on the expertise of imaging or clinical physicians, which causes inconsistencies in the assessment results of carotid atherosclerotic plaques—particularly in measuring carotid intima-media thickness, characterizing intraplaque components, and evaluating fibrous cap integrity.

The radiomics algorithms and deep learning (DL) models have demonstrated significant potential in medical image analysis [20]. Radiomics is a quantitative medical imaging analysis approach that aims to transform high-dimensional image features (such as texture heterogeneity, spatial topological relationships, and intensity distribution) into quantifiable digital biomarkers, thereby providing objective evidence to guide clinical decision-making. However, the characteristic dimensionality of radiomics data often far exceeds sample sizes, which renders the traditional statistical methods inadequate [21]. Machine learning (ML), with the potential to process large-scale, high-dimensional data and uncover deep correlations among these complex features [22]. Combining radiomics with ML to develop an ML model using radiomics can enhance the diagnostic performance of AI in large and complex datasets, exceeding the performance of models constructed through traditional statistical methods.

DL is also one of the important subbranches of artificial intelligence, which can automatically learn and layer from raw data without manual design of features, ultimately generating predictions via an output layer [23]. DL-driven image generation techniques have demonstrated remarkable effectiveness in cross-modality imaging and synthesis tasks across various sequences within the same modality. With the rapid development of computer technology, ML models based on radiomics and DL models based on radiomics have become important tools for cardiovascular disease research. Current evidence suggests that these methods can significantly improve the quantitative assessment accuracy of atherosclerotic plaque progression and enhance the diagnostic and predictive power of major adverse cardiovascular events [24-26]. In recent years, research on the application of these methods in the fields of plaque diagnosis, stability assessment, and symptomatic plaque identification has increased significantly. Although these advancements have significantly improved the diagnosis of carotid plaques, variations in data dependency and imaging configurations among different models create inconsistencies in diagnostic accuracy. Moreover, these models may become overly specialized in common imaging configurations, even when using radiomics data from identical sources. Currently, systematic evaluations of its clinical validity remain limited.

Therefore, this systematic review comprehensively assesses the applications of ML models based on radiomics algorithms and DL models in carotid plaques, while highlighting gray areas in the available literature.

Methods

Study Registration

The study was performed in line with the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy Studies) guidelines [27] and PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) standards [28,29] and was registered on the International Prospective Register of Systematic Reviews (PROSPERO CRD42025638492).

Data Sources and Search Strategy

Relevant articles were searched on PubMed, Embase, Web of Science, Cochrane Library, and Institute of Electrical and Electronics Engineers (IEEE) databases, focusing on English-language articles published up to September 24, 2025. The literature search was based on the PIO (population, intervention, and outcomes) principles: “P” represents carotid artery disease, carotid plaques, or atherosclerosis populations; “I” represents radiomics or DL as interventions; and “O” represents the outcomes of diagnosis and their subordinates and other keywords. Furthermore, we manually analyzed the reference lists of all included articles to identify additional relevant publications. The complete search strategy is outlined in Table S1 in Multimedia Appendix 1. The EndNote 20 software (Clarivate Analytics) was used to manage the included studies.

XSL•FO
RenderX

Eligibility Criteria

Inclusion Criteria

The inclusion criteria included:

1. Studies on patients with extracranial carotid plaques that aimed to detect or distinguish between unstable and symptomatic plaques, among other factors.
2. Studies using radiomics algorithms or DL models based on medical imaging techniques, such as ultrasound, CTA, or MRI, to diagnose carotid plaques.
3. Studies reported the diagnostic performance metrics, including confusion matrix, 2×2 diagnostic tables, accuracy, sensitivity, specificity, receiver operating characteristic (ROC) curves, F_1-score, precision, recall, etc.
4. Those that adopted the following designs: prospective or retrospective cohorts, diagnostic accuracy trials, model development or validation studies, and comparative studies (eg, AI models vs AI models combined with clinical features).
5. Only studies published in English and with extractable quantitative data were deemed eligible.

Exclusion Criteria

The exclusion criteria excluded:

1. Studies involving nonhuman subjects (animal experiments or in vitro models), those that explored intracranial or coronary plaques, enrolled pediatric populations (<18 years), or reported only generalized atherosclerosis without plaque-specific criteria (focal intima-media thickness ≥1.5 mm) or specific diagnostic metrics;
2. Those that did not adopt well-defined deep learning models or radiomics algorithms, focused only on image segmentation or texture analysis without diagnostic validation, or reported predictive models without providing a clear diagnostic relevance.
3. Studies that lacked a validated reference standard.
4. Studies that did not report diagnostic performance.
5. Informal publication types (eg, reviews, letters to the editor, editorials, and conference abstracts).
6. Studies that did not report validation or test sets.

Screening of Articles and Data Extraction

In the initial screening, duplicates were excluded followed by reading of full texts, and data were entered into a predefined extraction table, which included surnames of authors, source of data, publication year, algorithm architecture, type of internal validation, availability of open access data, external verification status, reference standard, transfer learning application, number of cases for training, test, internal, or external validation, study design, sample size, mean or median age, inclusion criteria, and model evaluation metrics. The contingency tables are derived from the models explicitly identified by the original authors as the best-performing ones. Data from external validation sets were prioritized. If there were no external validation set in the original studies, data from internal validation sets were used. If neither was available, the contingency tables corresponding to the test sets were selected. This process was performed by two researchers (LJ and YG), working independently, and any differences were resolved through discussion with a third researcher (HG).

Quality Assessment

Two blinded investigators (LJ and YG) systematically assessed the quality of studies using the Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence (QUADAS-AI) tool. Specifically, they evaluated the risk of bias and applicability concerns across 4 domains: flow and timing, reference standard, index test, and participant selection. Although the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) is extensively applied to assess the quality of diagnostic accuracy studies [30], it does not address the specific methodological choices, result analyses, and measurements related to diagnostic studies using AI. To address this gap, QUADAS-AI was developed as a consensus-based tool to aid readers in systematically examining the risk of bias and the usability of AI-related diagnostic accuracy studies (Table S6 in Multimedia Appendix 1) [31], thereby improving the quality assessment process [32,33]. Any evaluation discrepancies were resolved by a third investigator (HG).

Statistical Analysis

A meta-analysis was performed using STATA/MP software (version 17.0; Stata Corporation) with a bivariate random-effects model. For meta-analyses of the diagnostic accuracy of AI-based models, bivariate mixed-effects models can account for both within-study variability (random effects) and between-study heterogeneity (fixed effects), ensuring the robustness of the pooled estimates [34]. A contingency table was generated using data from the included literature, and then we calculated metrics such as the number of cases, the Youden index, sensitivity, specificity, and recall. The diagnostic efficacy of radiomics algorithms and DL models in evaluating carotid plaque was determined using a summary receiver operating characteristic (SROC) curve and area under the curve (AUC; 0.7≤AUC<0.8 fair; 0.8≤AUC<0.9 good; and AUC≥0.9 excellent). Publication bias was explored using Deeks funnel plot asymmetry test. The Fagan nomogram was developed to determine clinically pertinent posttest probabilities (P-post) and likelihood ratios (LRs). LRs were determined by comparing the probability of test results between diseased and nondiseased groups. The pretest probability was subsequently adjusted based on test results and LRs to obtain P-post [35]. The Cochran Q (P≤.05) and I^2 statistic were used to explore heterogeneity among the included studies, and regression analysis was conducted to assess sources of heterogeneity. I^2≤50% indicated mild heterogeneity, 50%<I^2<75% reflected moderate heterogeneity, and I^2≥75% indicated high heterogeneity.

The subgroup analysis encompassed the following factors: (1) model type (DL or ML model), (2) medical imaging modalities (PRs, ultrasound, MRI, or CTA), (3) application of transfer learning, (4) characteristics of carotid plaques (presence vs absence, stable vs vulnerable, and symptomatic vs asymptomatic), (5) comparison of the most effective ML model based on radiomics algorithm and DL models using the same dataset and clinicians' diagnoses, (6) different types of datasets (testing and validation), (7) low and high or unclear risk of bias

XSL•FO
RenderX

studies, (8) different sample sizes of model, and (9) models with different research designs (multicenter studies and single-center studies). To identify the sources of heterogeneity associated with nonthreshold effects, meta-regression was performed using the above-mentioned covariates.

Sensitivity analysis was performed to assess the stability of the results by several steps: (1) excluding specific articles one by one to determine the stability of the results, (2) excluding studies with extremely large sample sizes (N≥500; n=7 studies), (3) excluding studies with extremely small sample sizes (N≤50; n=4 studies), and (4) excluding studies with extreme effect sizes (sensitivity or specificity>0.95 or <0.7; n=11 studies).

Results

Study Selection

We obtained 5834 studies in the initial analysis, of which 1233 were excluded for duplication or redundancy. After screening titles and abstracts, 4507 publications were eliminated. After the full texts of the 94 articles were read, 40 studies were eligible for meta-analysis. The PRISMA flow diagram of the study showing the selection process is presented in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of study selection. IEEE: Institute of Electrical and Electronics Engineers.

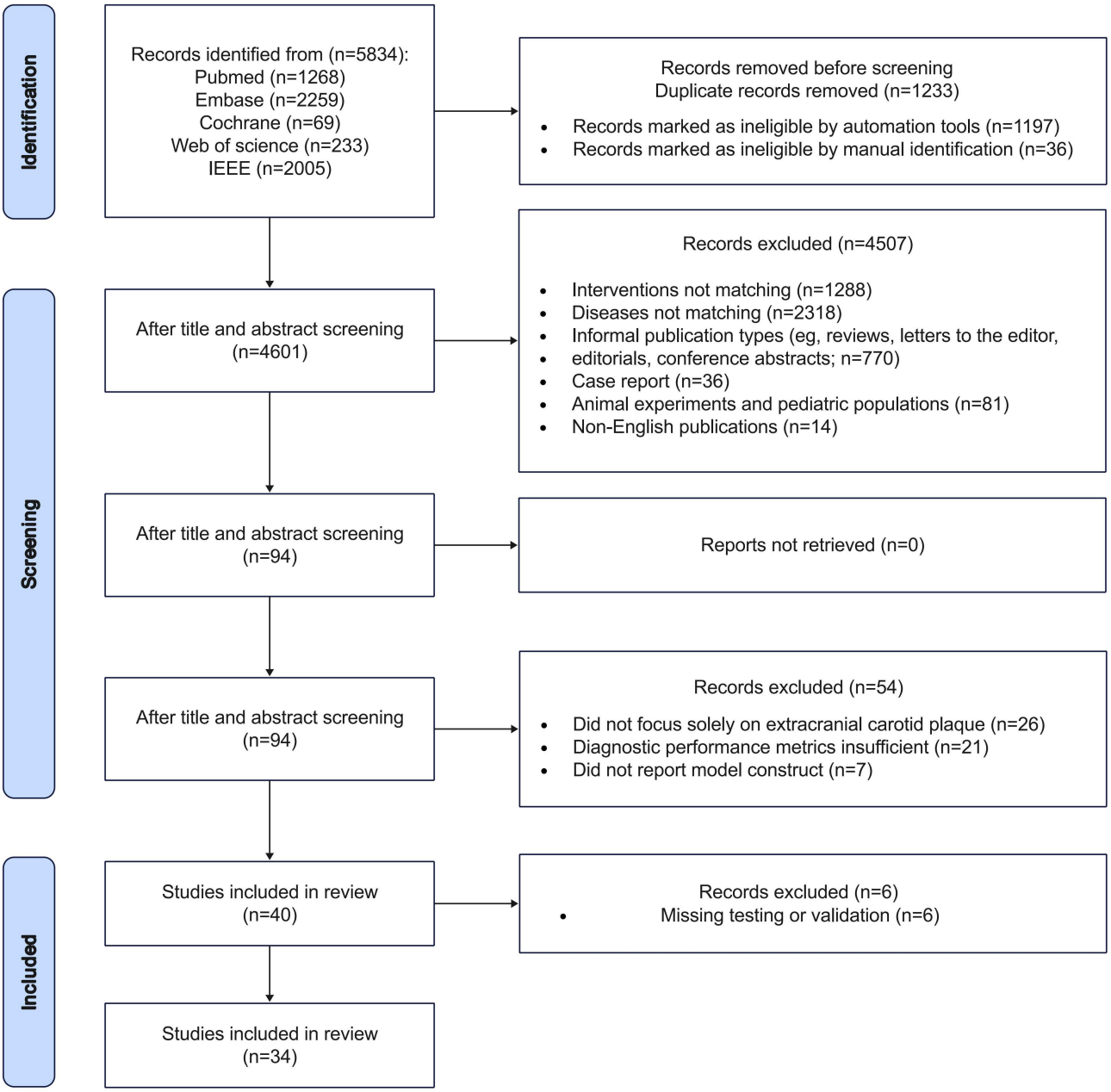

Study Characteristics

Among the 40 studies that fulfilled the systematic review's inclusion criteria, 34 provided sufficient quantitative data (contingency tables from validation or test sets) eligible for incorporation into the meta-analysis. The detailed characteristics of all 40 eligible studies are summarized in Tables S3 and S4 in Multimedia Appendix 1, while all subsequent quantitative

XSL•FO
RenderX

analyses were conducted based on the 34 studies with available quantitative data. Overall, 34 studies were included [36-69], among which 9 were multicenter studies [41,43,45,49,57,63-65,69], 3 used public databases [37,40,53], 13 provided open access to the data [37,40,45,48-50,53,57,59,63-66]. A total of 12 studies conducted internal validation [38,39,41,42,44,47,48,57,61,64,69,70] to confirm the reproducibility of the model development process and prevent overfitting. In addition, 7 studies conducted external validation [41,50,57,60,63,64,69] to assess the model's transportability and generalizability using unused datasets. Only 1 study conducted a comparative analysis of the diagnostic performance of DL models with that of clinicians [57]. The medical imaging modalities included PRs (n=5), ultrasound (n=16), MRI (n=5), and CTA (n=8). The core features of the 34 studies are presented in Tables 1 and 2, with further details provided in Tables S2 and S3 in Multimedia Appendix 1.

XSL•FO
RenderX

Table . Data sources, indicators, and algorithms of included studies.

Study, year	Data source				Validation type
	Source of data	Number of cases for training, test, internal, or external	Data range	Labels	
Su et al [51], 2023	China	322; 138; NR[a]; NR	NR	Stable or vulnerable plaque	No
Zhang et al [70], 2024	China	4064; NR; 1016; NR	NR	Stable or vulnerable plaque	Internal validation
Zhou et al [44], 2024	China	751; 261; 258; NR	NR	Stable or vulnerable plaque	Internal validation
Zhang et al [58], 2021	China	121; 41; NR; NR	NR	Symptomatic or asymptomatic	No
Zhai et al [41], 2024	NR	240; NR; 60; 100	January 2017-January 2022	Normal or abnormal	External validation
Yoo et al [39], 2024	South Korea	388; 130; 130; NR	2009 - 2022	Normal or abnormal	Internal validation
Xu et al [56], 2022	NR	NR	NR	Stable or vulnerable plaque	No
Xie et al [47], 2023	China	264; 75; 38; NR	2020 - 2021	Stable or vulnerable plaque	Internal validation
Wei et al [62], 2024	China	2725; 554; NR; NR	NR	Normal or abnormal	No
Ganitidis et al [60], 2021	Greece	46; 10; 18; NR	NR	Symptomatic or asymptomatic	Internal validation
Shi et al [50], 2023	China	134; 33; NR; NR	October 2019-July 2022	Symptomatic or asymptomatic	No
Gui et al [49], 2023	China	84; 20; NR; NR	NR	Symptomatic or asymptomatic	No
Ali et al [71], 2024	Italy	336; 84; NR; NR	NR	Symptomatic or asymptomatic	No
Amitay et al [48], 2023	Israel	371; 144; 144; NR	2016 - 2021	Normal or abnormal	Internal validation
Ayoub et al [72], 2023	China	136; 150; 69; NR	NR	Stable or vulnerable plaque	Internal validation
Cilla et al [55], 2022	Italy	NR	October 2015-October 2019	Stable or vulnerable plaque	No
Guang et al [57], 2021	China	136; NR; 69; NR	September 2017-September 2018	Stable or vulnerable plaque	Internal validation
He et al [69], 2024	China	3088; NR; 772; 1564	January 2021-March 2023	Normal or abnormal; stable or vulnerable plaque	Internal and external validation
Latha et al [73], 2021	India	NR	NR	Normal or abnormal	No
Ma et al [59], 2021	China	1169; 294; NR; NR	NR	A total of 3 types (echo-rich, intermediate, and echolucent)	No
Pisu et al [43], 2024	Italy	163; 106; NR; NR	March 2013-October 2019	Symptomatic or asymptomatic	No
Wang et al [74], 2024	China	154; 39; NR; NR	January 1, 2018-December 31, 2021	Symptomatic or asymptomatic	No
Gago et al [53], 2022	Spain	NR	2007 - 2010	Normal or abnormal	No
Omarov et al [40], 2024	The United Kingdom	577; 103; NR; NR	NR	Normal or abnormal	No
Wang et al [45], 2023	China	2619; 1122; NR; NR	NR	Stable or vulnerable plaque	No

XSL•FO
RenderX

Study, year	Data source				Validation type
	Source of data	Number of cases for training, test, internal, or external	Data range	Labels	
Vinayahalingam et al [42], 2024	Germany	280; 37; 37; NR	NR	Normal or abnormal	No
Singh et al [37], 2024	Cyprus; The United Kingdom; NR	3088; 772; NR; NR	NR	Stable or vulnerable plaque	No
Shan et al [46], 2023	China	52; 22; NR; NR	January 2018-December 2021	Stable or vulnerable plaque	No
Li et al [38], 2024	NR	4546; 1471; 1019; NR	NR	Normal or abnormal	Internal validation
Jain et al [54], 2021	NR	682; 76; NR; NR	July 2009-September 2010	Stable or vulnerable plaque	No
Molinari et al [36], 2018	Italy	NR	2004 - 2010	Symptomatic or asymptomatic	No
Kats et al [61], 2019	Israel	1946; 7; 12; NR	NR	Normal or abnormal	Internal validation
Chen et al [52], 2022	China	81; 34; NR; NR	July 2015-May 2021	Symptomatic or asymptomatic	No
Zhao et al [63], 2025	China	317; NR; NR; 328	January 2018-December 2023 (Center 1); Jan 2022-December 2023 (Center 2,3)	Symptomatic or asymptomatic	External validation
Hu et al [64], 2025	China	213; NR; 93; 110	January 2018-May 2023 (Center 1); January 2020-May 2023 (Center 2)	Symptomatic or asymptomatic	Internal and external validation
Li et al [75], 2025	China	2069; 887; NR; NR	October 2021-January 2022	normal or abnormal	No
Yu et al [66], 2025	China	146; 63; NR; NR	April 2022-August 2023	HIPs[b] or NHIPs[c]	No
Liapi et al [65], 2025	Cyprus, The United Kingdom, and Greece	168; 46; 22; NR	NR	Symptomatic or asymptomatic	Internal validation
Kuwada et al [67], 2025	Japan	Training and validation data: 500; Test data: 80	2008 - 2023	Normal or abnormal	No
Lao et al [68], 2025	China	76; 31; NR; NR	January 2017-October 2022	Stable or vulnerable plaque	No

[a]NR: not reported.

[b]HIP: highly inflammatory plaque.

[c]NHIP: non–highly inflammatory plaque.

XSL•FO
RenderX

Table . Data sources, indicators, and algorithms of all studies.

Study, year	Indicator definition		Algorithm		
	Device	Exclusion of poor quality cases	Algorithm architecture	ML[a] or DL[b]	Transfer learning applied
Su et al [51], 2023	Ultrasound	NR[c]	Inception V3; VGG-16[d]	DL	No
Zhang et al [70], 2024	Ultrasound	NR	Fusion-SSL	DL	No
Zhou et al [44], 2024	Ultrasound	NR	Tri-Correcting	DL	No
Zhang et al [58], 2021	MRI[e]	Yes	LASSO[f] MRI-based model (HRPMM[g])	ML models based on radiomics algorithms[h] (LASSO algorithm)	No
Zhai et al [41], 2024	CT	Yes	3D-UNet; ResUNet	DL	No
Yoo et al [39], 2024	PRs	Yes	CACSNet	DL	Yes
Xu et al [56], 2022	Ultrasound	NR	Multi-feature fusion method	DL	No
Xie et al [47], 2023	Ultrasound	NR	CPTV[i]	DL	No
Wei et al [62], 2024	Ultrasound	Yes	BETU[j]	DL	Yes
Ganitidis et al [60], 2021	Ultrasound	NR	CNNs[k]	DL	No
Shi et al [50], 2023	CT[l] and MRI	Yes	LASSO regression	ML models based on radiomics algorithms (LASSO algorithm)	No
Gui et al [49], 2023	MRI	Yes	3D-SE-DenseNet121[m]; ANOVA_spearman_LASSO and MLP[n]	ML models based on radiomics algorithms (LASSO, ANOVA_LASSO and ANOVA_spearman_LASSO) and DL	No
Ali et al [71], 2024	Ultrasound	No	CAROTIDNet[o]	DL	No
Amitay et al [48], 2023	PRs	Yes	InceptionResNetV2 (minimum-maximum)	DL	Yes
Ayoub et al [72], 2023	MRI	NR	HViT[p]	DL	No
Cilla et al [55], 2022	CT	Yes	SVM RBF[q] kernel	ML models based radiomics algorithms (logistic regression [LR]), support vector machine (SVM), and CART[r]	No
Guang et al [57], 2021	Ultrasound	Yes	DL-DCCP[s]	DL	Yes
He et al [69], 2024	Ultrasound	Yes	BCNN[t]-ResNet[u]	DL	No
Latha et al [73], 2021	Ultrasound	NR	CART; logistic regression; random forest; CNN; Mobilenet; Capsulenet	ML models based radiomics algorithms (CART, logistic regression, and random forest algorithm) and DL	Yes
Ma et al [59], 2021	Ultrasound	NR	MSP[v]-VGG	DL	Yes
Pisu et al [43], 2024	CT	Yes	GB-GAM[w]	ML models based radiomics algorithms (NR)	No
Wang et al [74], 2024	CT	Yes	SR[x]	DL	Yes
Gago et al [53], 2022	Ultrasound	NR	End-to-end framework	DL	No

XSL•FO
RenderX

Study, year	Indicator definition		Algorithm		
	Device	Exclusion of poor quality cases	Algorithm architecture	ML[a] or DL[b]	Transfer learning applied
Omarov et al [40], 2024	Ultrasound	Yes	YOLOv8[y]	DL	Yes
Wang et al [45], 2023	MRI	Yes	ResNet-50	DL	Yes
Vinayahalingam et al [42], 2024	PRs[z]	Yes	Faster R-CNN[aa] with Swin Transformer (Swin-T)	DL	Yes
Singh et al [37], 2024	Ultrasound	Yes	GoogLeNet[ab]	ML models based on radiomics algorithms (SVM algorithms) and DL	Yes
Shan et al [46], 2023	CT and ultrasound	Yes	LR[ac]; SVM[ad]; RF[ae]; LGBM[af]; daBoost; XGBoost[ag]; MLP	ML models based on radiomics algorithms (Pyradiomics package in Python software)	Yes
Li et al [38], 2024	Ultrasound	NR	U-Net; CNN	DL	No
Jain et al [54], 2022	Ultrasound	NR	SegNet-UNet[ah]	DL	No
Molinari et al [36], 2018	Ultrasound	NR	SVM	ML models based on radiomics algorithms (BEMD[ai])	No
Kats et al [61], 2019	PRs	NR	Faster R-CNN	DL	No
Chen et al [52], 2022	MRI	Yes	LASSO	ML models based on radiomics algorithms (mRMR[aj] algorithm and LASSO algorithm)	No
Zhao et al [63], 2025	CTA[ak]	Yes	XGBoost	ML models based on radiomics algorithms (XGBoost)	No
Hu et al [64], 2025	CTA	Yes	LASSO regression; SVM; logistic regression	ML models based on radiomics algorithms (LASSO algorithm) and classifier (SVM)	No
Li et al [75], 2025	Ultrasound	NR	XGBoost; RF; LASSO regression	ML models based on radiomics algorithms (XGBoost, RF, LASSO regression)	No
Yu et al [66], 2025	MRI	Yes	Plaque-R model; PVAT-R[al] model; ensemble model	ML models based on radiomics algorithms (LASSO algorithm) and ensemble learning	No
Liapi et al [65], 2025	Ultrasound	NR	Xception	DL	Yes
Kuwada et al [67], 2025	Ultrasound	NR	GoogLeNet; YOLOv7	DL	No

XSL•FO
RenderX

Study, year	Indicator definition		Algorithm		
	Device	Exclusion of poor quality cases	Algorithm architecture	ML[a] or DL[b]	Transfer learning applied
Lao et al [68], 2025	CTA	Yes	mRMR algorithm; LASSO regression	ML models based on radiomics algorithms (mRMR algorithm; LASSO algorithm)	No

[a]ML: machine learning.

[b]DL: deep learning.

[c]NR: not reported.

[d]VCG: VGG visual geometry group network.

[e]MRI: magnetic resonance imaging.

[f]LASSO: least absolute shrinkage and selection operator.

[g]HRPMM: high-risk plaque MRI-based model.

[h]Definition of ML models based on radiomics algorithms and deep learning (DL): ML models based on radiomics algorithms are models that rely on artificially designed features (such as texture and shape features) and use traditional algorithms (such as random forest, support vector machine, logistic regression, etc) to complete classification, without the need for DL algorithms to be in the core task. The DL model was defined as a model that automatically extracts features and completes classification through neural networks (such as convolutional neural network, ResNet, etc), regardless of whether the input contains a small number of artificial features, as long as the core task relies on the DL algorithm.

[i]CPTV: classification of plaque by tracking videos.

[j]BETU: be easy to use.

[k]CNN: convolutional neural network.

[l]CT: computed tomography.

[m]3D-SE-DenseNet121: 3D squeeze-and-excitation DenseNet with 121 layers.

[n]MLP: multilayer perceptron.

[o]CAROTIDNet: carotid symptomatic/asymptomatic plaque detection network.

[p]HViT: hybrid vision transformer.

[q]SVM RBF: kernel support vector machine with radial basis function kernel.

[r]CART: classification and regression tree.

[s]DL-DCCP: deep learning-based detection and classification of carotid plaque.

[t]BCNN: bilinear convolutional neural network.

[u]ResNet: deep residual network.

[v]MSP: multilevel strip pooling.

[w]GB-GAM: gradient-boosting generalized additive model.

[x]SR: super resolution.

[y]YOLOv8: you only look once version 8.

[z]PR: panoramic radiograph.

[aa]Faster R-CNN: faster region-based convolutional network.

[ab]GoogLeNet: Google network.

[ac]LR: logistic regression.

[ad]SVM: support vector machine.

[ae]RF: random forest.

[af]LGBM: light gradient boosting machine.

[ag]XGBoost: extreme gradient boosting.

[ah]SegNet-UNet: segmentation network-UNet.

[ai]BEMD: bidimensional empirical mode decomposition.

[aj]mRMR: minimum redundancy maximum relevance.

[ak]CTA: computed tomography angiography.

[al]PVAT: perivascular adipose tissue.

Meta-Analysis of Diagnostic Performance

Synthesized Results

The meta-analysis revealed pooled sensitivity, specificity, and an area under the SROC curve (SROC AUC) of 0.88 (95% CI 0.85 - 0.91; I^2=93.58%; P<.001; in Multimedia Appendix 2 [36-69]), 0.89 (95% CI 0.85 - 0.92; I^2=91.38%; P<.001; in Multimedia Appendix 2 [36-69]), and 0.95 (95% CI 0.92 - 0.96) for all 34 studies (Figure 2A); 0.88 (95% CI 0.84 - 0.92; I^2=93.70%; P<.001; Multimedia Appendix 3 [36-69]), 0.91

XSL•FO
RenderX

(95% CI 0.86 - 0.94; I^2=95.55%; P<.001; Multimedia Appendix 3 [36-69]), and 0.95 (95% CI 0.93 - 0.97) for all DL models (Figure 2B); 0.89 (95% CI 0.82 - 0.93; I^2=90.20%; P<.001; Multimedia Appendix 3 [36-69]), 0.83 (95% CI 0.76 - 0.88; I^2=78.92%; P<.001; Multimedia Appendix 3 [36-69]), and 0.92 (95% CI 0.89 - 0.94) for all ML models based on radiomics algorithms (Figure 2C), respectively. Notably, some studies used multiple diagnostic models; however, the diagnostic accuracy of certain models was not thoroughly assessed.

Figure 2. Receiver operating characteristic curves based on the overall performance of different algorithms. (A) All studies included in the meta-analysis (34 studies with 34 tables). (B) Deep learning (DL) models (24 studies with 24 tables). (C) Machine learning (ML) models based on radiomics algorithms (10 studies with 10 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

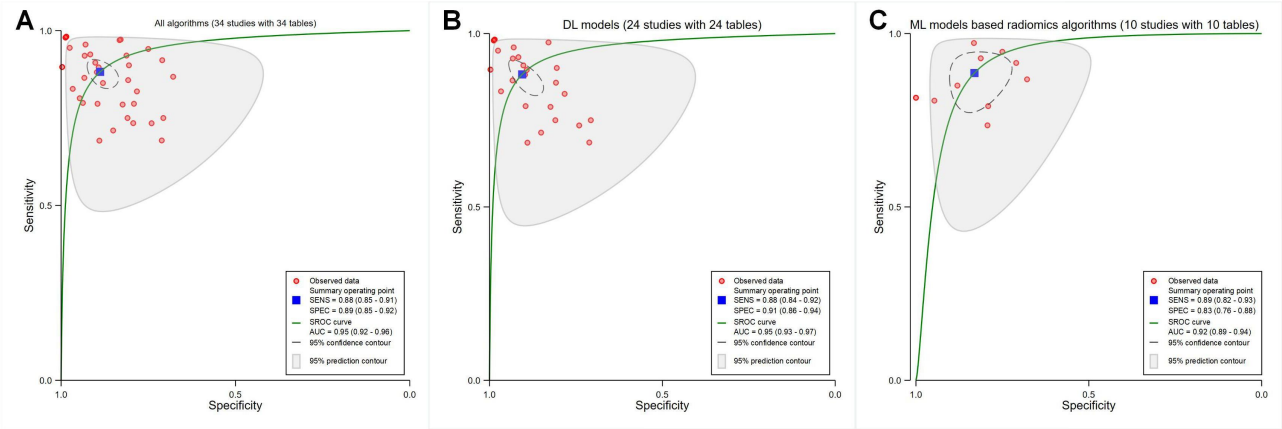

Subgroup Analysis

Medical Imaging Modalities

The pooled sensitivity, specificity, and SROC AUC were 0.91 (95% CI 0.80 - 0.96), 0.93 (95% CI 0.84 - 0.97), and 0.97 (95% CI 0.95 - 0.98) for the 5 studies using PRs (P<.001; with 5 contingency tables; Figure 3A); 0.89 (95% CI 0.84 - 0.93), 0.90 (95% CI 0.84 - 0.94), and 0.95 (95% CI 0.93 - 0.97) for the 16 studies using ultrasound images (P<.001with 16 contingency tables; Figure 3B); 0.87 (95% CI 0.87 - 0.92), 0.87 (95% CI 0.76 - 0.93), and 0.93 (95% CI 0.91 - 0.95) for the 5 studies using MRI images (P<.001; with 5 contingency tables; Figure 3C); 0.83 (95% CI 0.76 - 0.88), 0.83 (95% CI 0.75 - 0.89), and 0.90 (95% CI 0.87 - 0.92) for the 8 studies using CTA images (P<.001; with 8 contingency tables; Figure 3D), respectively. In addition, we conducted subgroup analyses using the same imaging modality based on differentiation. However, only subgroups of identifying the presence and stability of plaque had sufficient data for the ultrasound modality to perform statistical analyses and obtain pooled diagnostic performance metrics (Table S5 in Multimedia Appendix 1). The pooled sensitivity, specificity, and SROC AUC were 0.88 (95% CI 0.72 - 0.96), 0.91 (95% CI 0.80 - 0.96), and 0.95 (95% CI 0.93 - 0.97) for determining the presence of plaques (P<.001; with 5 contingency tables; Figure 3E), 0.90 (95% CI 0.84 - 0.94), 0.92 (95% CI 0.83 - 0.96), and 0.96 (95% CI 0.94 - 0.97) for distinguishing the stability of plaques (P<.001; with 8 contingency tables; Figure 3F).

XSL•FO
RenderX

Figure 3. Receiver operating characteristic curves for different medical imaging modalities. (A) Periapical radiographs (PRs) imaging models (5 studies with 5 tables). (B) Ultrasound imaging models (16 studies with 22 tables). (C) Magnetic resonance imaging (MRI) models (5 studies with 7 tables). (D) Computed tomography angiography (CTA) models (8 studies with 10 tables). (E) Models based on ultrasound modality for detecting the presence of carotid plaque (5 studies with 5 tables). (F) Models based on ultrasound modality for distinguishing the stability of carotid plaques (8 studies with 8 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

XSL•FO
RenderX

Use of Transfer Learning

The pooled sensitivity, specificity, and SROC AUC were 0.92 (95% CI 0.87 - 0.95), 0.93 (95% CI 0.88 - 0.96), and 0.97 (95% CI 0.95 - 0.96) for the 10 studies using transfer learning (*P*<.001; with 10 contingency tables; Figure 4A) and 0.86 (95% CI 0.82 - 0.90), 0.86 (95% CI 0.81 - 0.90), and 0.93 (95% CI 0.90 - 0.95) for the 24 studies without transfer learning (*P*<.001; with 24 contingency tables; Figure 4B), respectively.

Figure 4. Receiver operating characteristic curves demonstrating transfer learning application. (A) Models using transfer learning (10 studies with 10 tables). (B) Models without transfer learning (24 studies with 24 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

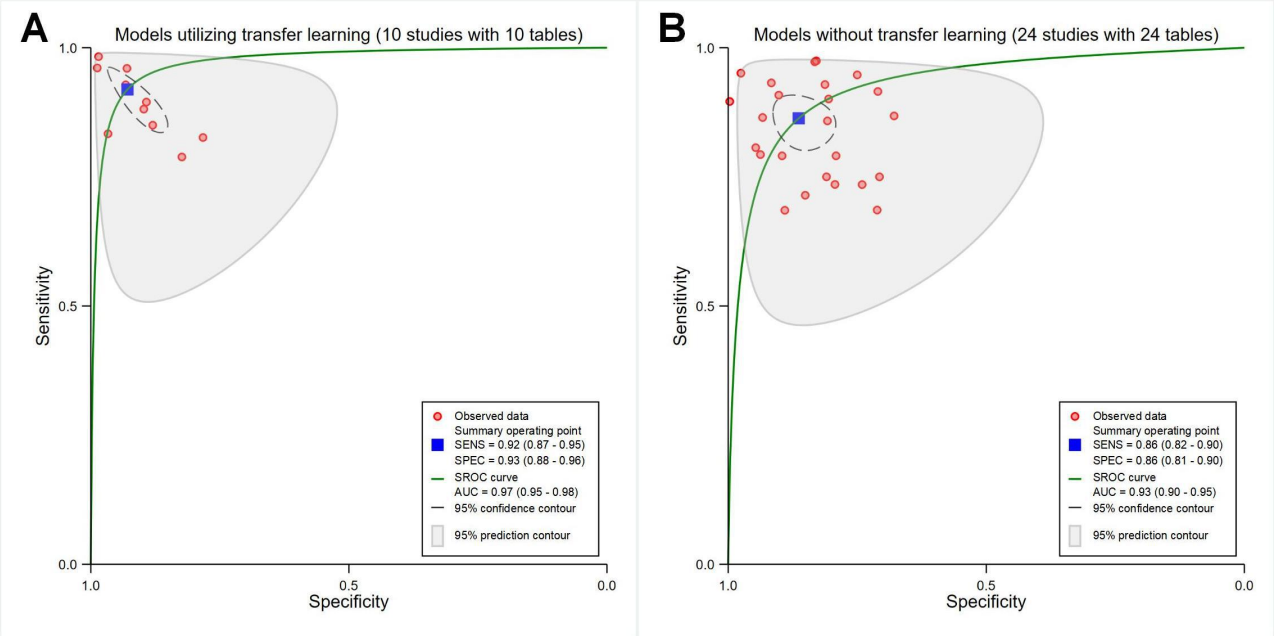

Carotid Plaque Type

The pooled sensitivity, specificity, and AUC were 0.89 (95% CI 0.81 - 0.94), 0.91 (95% CI 0.86 - 0.95), and 0.96 (95% CI 0.94 - 0.97) for the 11 studies identifying the presence or absence of carotid plaques (*P*<.001; with 11 contingency tables; Figure 5A); 0.90 (95% CI 0.85 - 0.94), 0.91 (95% CI 0.85 - 0.95), and 0.96 (95% CI 0.94 - 0.97) for the 12 studies identifying stable or vulnerable carotid plaques (*P*<.001; with 12 contingency tables), respectively (Figure 5B); and 0.86 (95% CI 0.78 - 0.91), 0.81 (95% CI 0.74 - 0.87), and 0.90 (95% CI 0.87 - 0.92) for the 10 studies identifying symptomatic or asymptomatic plaques (*P*<.001; with 10 contingency tables; Figure 5C), respectively.

Figure 5. Receiver operating characteristic curves for different carotid plaque types. (A) Presence versus absence of carotid plaques (11 studies with 11 tables). (B) Stable versus vulnerable carotid plaques (12 studies with 12 tables). (C) Symptomatic versus asymptomatic carotid plaques (10 studies with 10 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

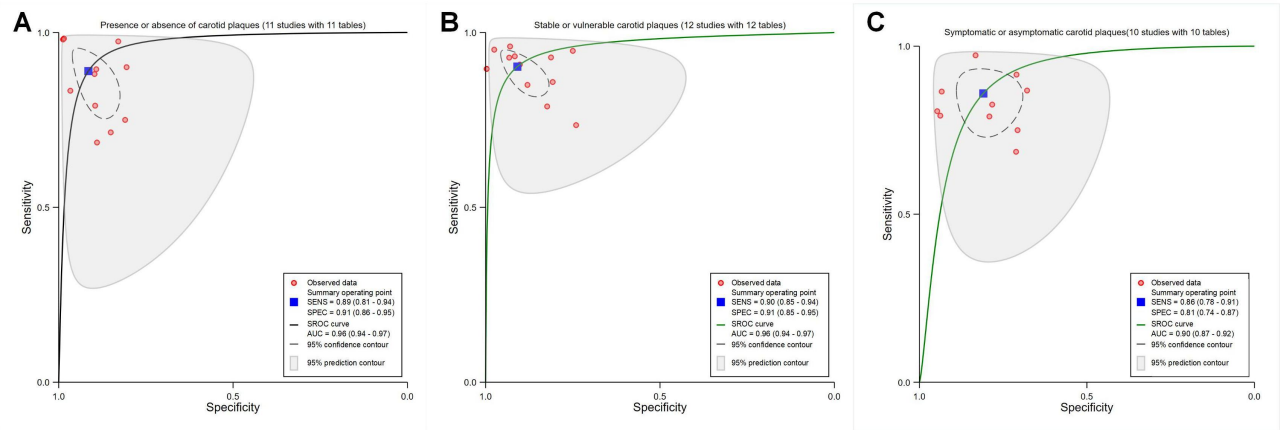

Pure Artificial Intelligence Models Versus Models Constructed by Combining Clinical Features

The pooled sensitivity, specificity, and SROC AUC were 0.82 (95% CI 0.74 - 0.88), 0.74 (95% CI 0.69 - 0.79), and 0.77 (95% CI 0.73 - 0.80) for the 7 studies involving pure artificial intelligence models meeting the inclusion criteria (*P*<.001; with 7 contingency tables; Figure 6A) and 0.85 (95% CI 0.76 - 0.92), 0.75 (95% CI 0.70 - 0.80), and 0.77 (95% CI 0.73 - 0.81) for models constructed by combining clinical features (*P*<.001; with 7 contingency tables; Figure 6B), respectively.

XSL•FO
RenderX

Figure 6. Receiver operating characteristic curves showing the diagnostic performance of pure artificial intelligence models or models constructed by combining clinical features. (A) Artificial intelligence models (7 studies with 7 tables). (B) Combined models (7 studies with 7 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

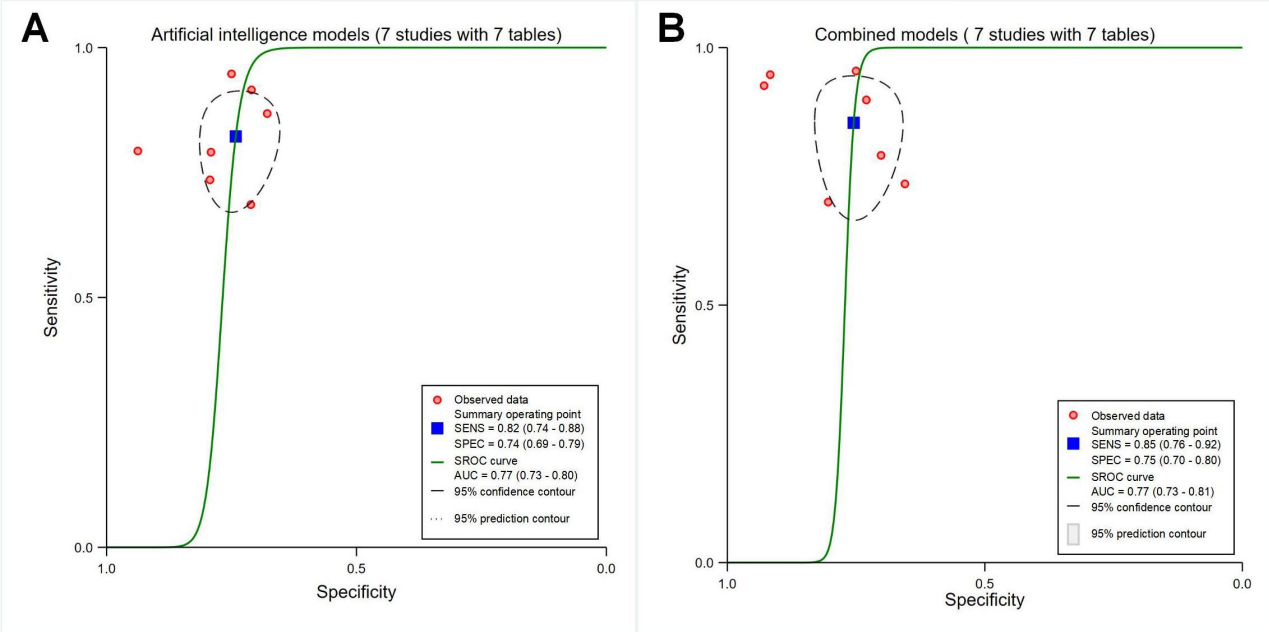

Different Sets of Datasets

The pooled sensitivity, specificity, and AUC were 0.90 (95% CI 0.87 - 0.93), 0.91 (95% CI 0.87 - 0.93), and 0.96 (95% CI 0.94 - 0.97) for testing sets (*P*<.001; with 27 contingency tables; Figure 7A); 0.78 (95% CI 0.71 - 0.83), 0.80 (95% CI 0.73 - 0.86), and 0.86 (95% CI 0.82 - 0.88) for external validation sets (*P*<.001; with 7 contingency tables; Figure 7B), respectively.

Figure 7. Receiver operating characteristic curves showing different sets of datasets. (A) Testing (27 studies with 27 tables). (B) External validation (7 studies with 7 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

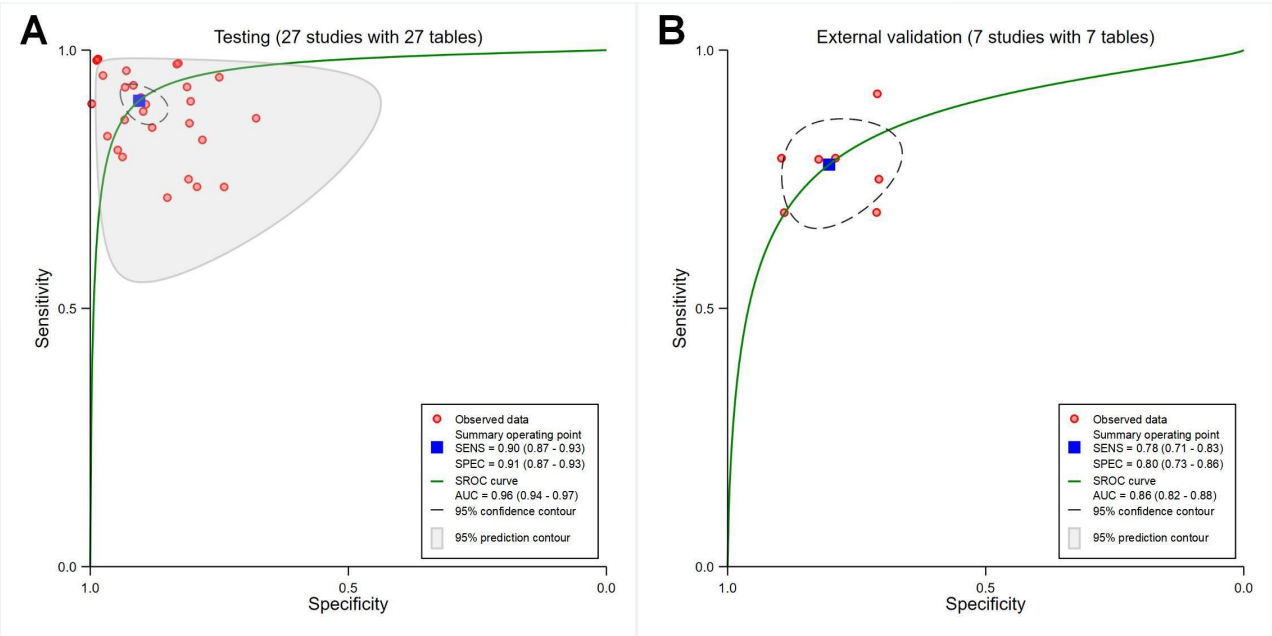

Low and High or Unclear Risk of Bias Studies

The pooled sensitivity, specificity, and AUC were 0.80 (95% CI 0.73 - 0.85), 0.80 (95% CI 0.71 - 0.87), and 0.86 (95% CI 0.83 - 0.89) for studies with a low risk of bias (*P*<.001; with 5 contingency tables; Figure 8A), and 0.89 (95% CI 0.86 - 0.92), 0.90 (95% CI 0.86 - 0.93), and 0.95 (95% CI 0.93 - 0.97) for studies with a high or unclear risk of bias (*P*<.001; with 29 contingency tables; Figure 8B), respectively.

XSL•FO
RenderX

Figure 8. Receiver operating characteristic curves showing studies with different risk of bias. (A) Studies with a low risk of bias (5 studies with 5 tables). (B) Studies with a high/unclear risk of bias (29 studies with 29 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

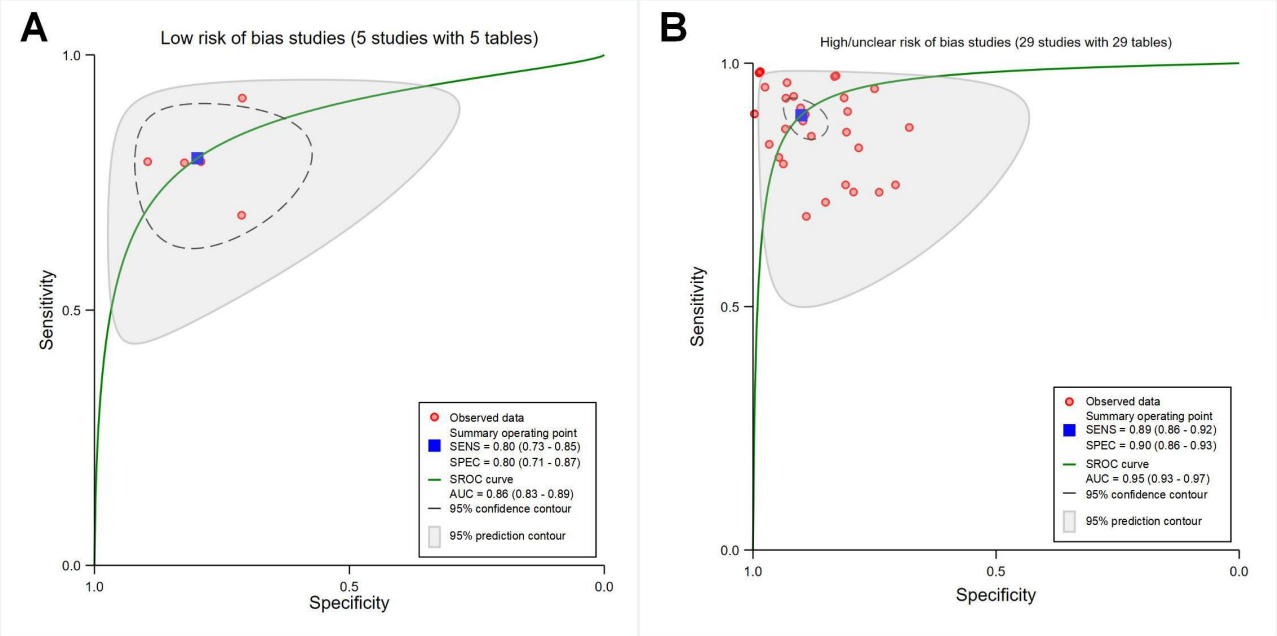

Different Sample Sizes of Model

The pooled sensitivity, specificity, and AUC were 0.91 (95% CI 0.86 - 0.94), 0.92 (95% CI 0.87 - 0.95), and 0.97 (95% CI 0.95 - 0.98) for sample size≥200 (*P*<.001; with 14 contingency tables) (Figure 9A), and 0.85 (95% CI 0.80 - 0.88), 0.86 (95% CI 0.80 - 0.90), and 0.91 (95% CI 0.89 - 0.94) for sample size<200 (*P*<.001; with 20 contingency tables; Figure 9B), respectively.

Figure 9. Receiver operating characteristic curves showing different sample sizes of model. (A) Sample size ≥200 (14 studies with 14 tables). (B) Sample size <200 (20 studies with 20 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

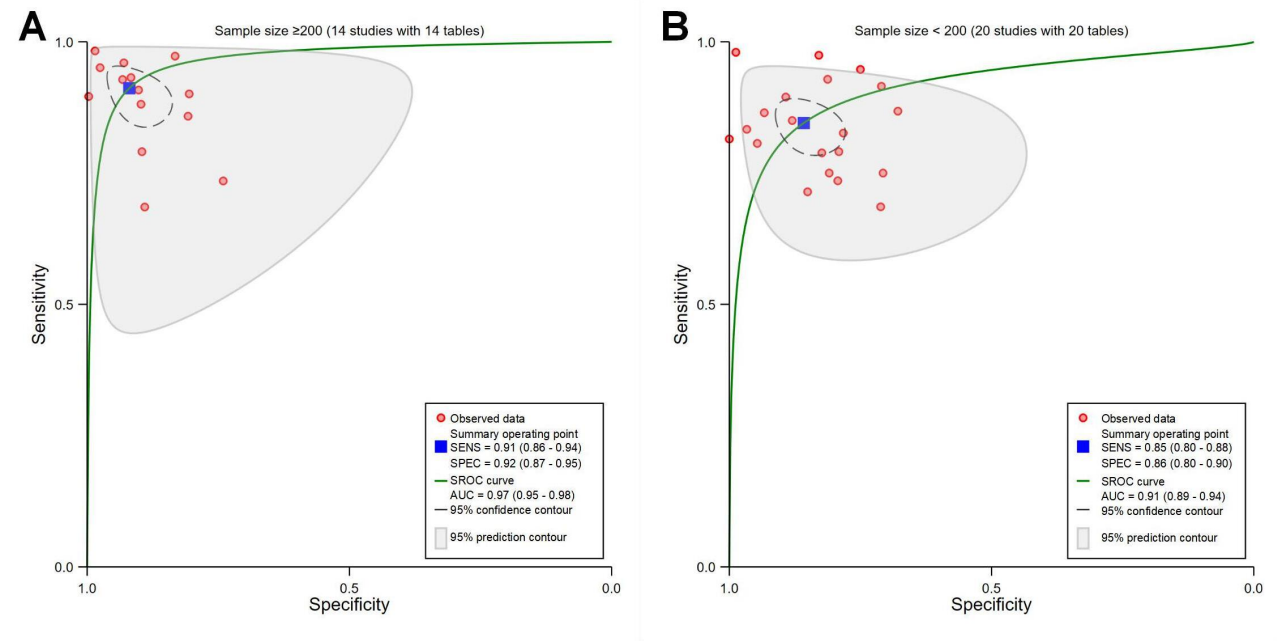

Models With Different Research Designs (Multicenter Studies and Single-Center Studies)

The pooled sensitivity, specificity, and AUC were 0.84 (95% CI 0.77 - 0.89), 0.87 (95% CI 0.81 - 0.91), and 0.92 (95% CI 0.90 - 0.94) for multicenter studies (*P*<.001; with 9 contingency tables; Figure 10A), and 0.89 (95% CI 0.84 - 0.92), 0.89 (95% CI 0.84 - 0.93), and 0.95 (95% CI 0.93 - 0.97) for single-center studies (*P*<.001; with 22 contingency tables; Figure 10B), respectively.

XSL•FO
RenderX

Figure 10. Receiver operating characteristic curves showing models with different research designs. (A) Multicenter studies (9 studies with 9 tables). (B) Single-center studies (22 studies with 22 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

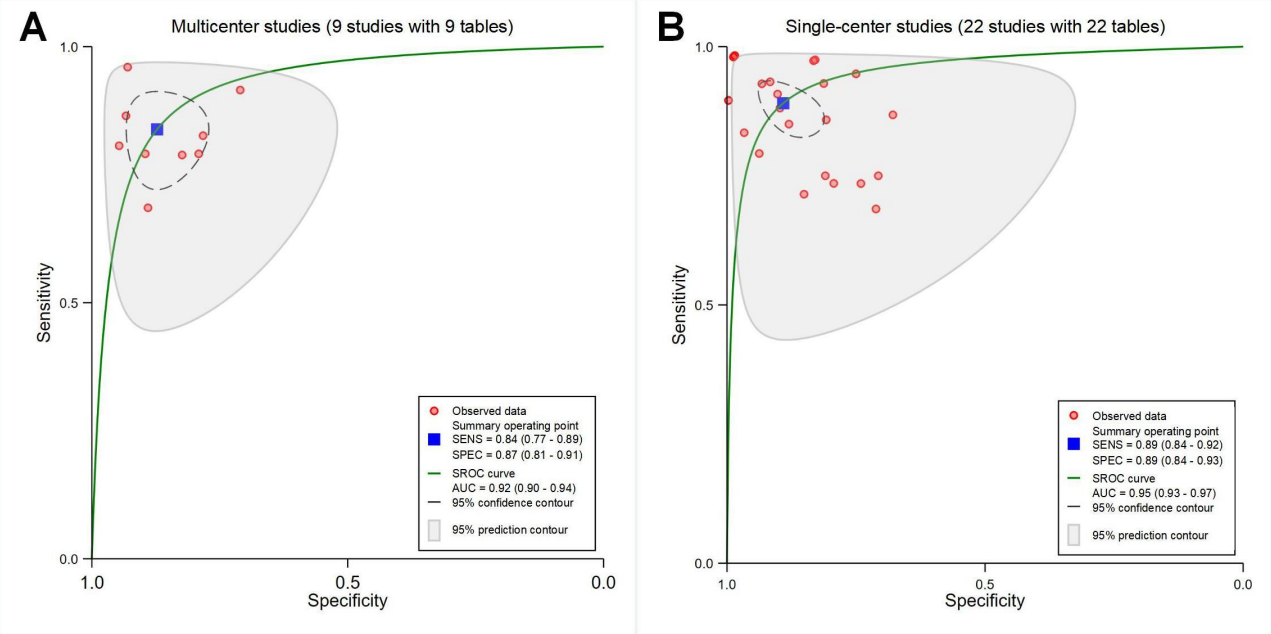

Heterogeneity Analysis and Meta-Regression Analysis

The Cochran Q test was used to indicate the presence of heterogeneity among subgroups (significance level $P \leq .05$) [15]. The I^2 index was used to assess the extent of heterogeneity among studies [15], revealing high sensitivity (I^2=93.58%) and specificity (I^2=91.38%; Multimedia Appendix 2). The Deek funnel plot asymmetry test, with P=.21, indicated no apparent publication bias (Multimedia Appendix 4). Subgroup analyses were performed using the random-effects models to identify the potential sources of heterogeneity, particularly when I^2 exceeded 50% [16]. Results were as follows:

1. AI model for carotid plaques: Both ML models based on radiomics algorithms and DL models exhibited high sensitivity, with an I^2 of 90.20% and 93.70%, and high specificity, with an I^2 of 78.92% and 95.55%, suggesting high performance and significant heterogeneity (Multimedia Appendix 3 [36-69]).
2. Medical imaging modalities: the sensitivity and specificity for PRs (sensitivity I^2=82.28%; specificity I^2=79.16%; Multimedia Appendix 5 [36-69]) and ultrasound (sensitivity I^2=96.92%; specificity I^2=94.98%; Multimedia Appendix 5 [36-69]). The sensitivity and specificity for MRI (sensitivity I^2=71.57%; specificity I^2=73.21%; Multimedia Appendix 5 [36-69]) and the sensitivity for CTA (I^2=56.80%) displayed moderate heterogeneity (Multimedia Appendix 5 [36-69]). The specificity of CTA (I^2=83.79%) was high (Multimedia Appendix 5 [36-69]). In the ultrasound modality, the sensitivity and specificity for determining the presence of plaques (sensitivity I^2=96.78%; specificity I^2=97.97%; Multimedia Appendix 5 [36-69]) and distinguishing the stability of plaques (sensitivity I^2=97.01%; sensitivity I^2=94.43%; Multimedia Appendix 5 [36-69]) were high.
3. Use of transfer learning: the specificity for models using transfer learning (specificity I^2=74.85%; Multimedia Appendix 6 [36-69]) displayed moderate heterogeneity. The sensitivity for models using transfer learning (sensitivity I^2=79.84%; Multimedia Appendix 6 [36-69]) and the sensitivity and specificity for the models without transfer learning (sensitivity I^2=94.12%; specificity I^2=87.35%; Multimedia Appendix 6 [36-69]) were high.
4. Carotid plaque type: all plaque types showed higher sensitivity and specificity; presence or absence of plaques (sensitivity I^2=94.08%; specificity I^2=97.60%; part A in Multimedia Appendix 7 [36-69]), stable or vulnerable plaques with (sensitivity I^2=95.19%; specificity I^2=91.29%; part B in Multimedia Appendix 7 [36-69]), and symptomatic or asymptomatic plaques (sensitivity I^2=93.28%; specificity I^2=84.67%; part C in Multimedia Appendix 7 [36-69]).
5. Both pure AI models and combined clinical features models did not exhibit high heterogeneity for AI models (sensitivity I^2=62.97%; specificity I^2=2.41%; part B in Multimedia Appendix 8 [50,52,58,63,64,66,68]) and combined models (sensitivity I^2=69.77%; specificity I^2=40.08%) for combined models (part A in Multimedia Appendix 8 [50,52,58,63,64,66,68]).
6. Different sets of datasets: both testing (sensitivity I^2=94.23%; specificity I^2=93.45%; part A in Multimedia Appendix 9 [36-69]) and external validation (specificity I^2=84.42%; part B in Multimedia Appendix 9 [36-69]) were high heterogeneity, except the sensitivity for external validation (I^2=66.67%; part B in Multimedia Appendix 9 [36-69]).

XSL•FO
RenderX

7. Different risk of bias studies: the sensitivity and specificity for high or unclear risk of bias studies (sensitivity I^2=94.61%; specificity I^2=92.59%; part B in Multimedia Appendix 10 [36-69]) and the specificity for low risk of bias studies (I^2=87.10%) were high (part A in Multimedia Appendix 10 [36-69]). The sensitivity for low risk of bias studies (I^2=62.20%) was moderate (part A in Multimedia Appendix 10 [36-69]).
8. Different sample sizes of model: The sensitivity and specificity for sample size ≥200 (sensitivity I^2=97.91%; specificity I^2=97.40%; part A in Multimedia Appendix 11 [36-69]) and the specificity for sample size <200 (I^2=78.02%; part B in Multimedia Appendix 11 [36-69]) were high. The sensitivity for sample size <200 (I^2=60.64%) was moderate (part B in Multimedia Appendix 11 [36-69]).
9. Models with different research designs: The sensitivity and specificity for multicenter studies (sensitivity I^2=81.36%; specificity I^2=80.24%; part A in Multimedia Appendix 12 [36,38,39,41-52,54-69]) and single-center studies (sensitivity I^2=95.07 %; specificity I^2=90.63%) were high (part B in Multimedia Appendix 12 [36,38,39,41-52,54-69]).

The meta-regression did not explore the factors contributing to heterogeneity (parts A-I in Multimedia Appendix 13 [36-69]). The results of all subgroups are depicted in Table S4 in Multimedia Appendix 1. The Fagan nomogram was used to evaluate the diagnostic performance of ML models based on radiomics algorithms and DL models for carotid plaques. The results showed a P-post of 89% and 12% for the positive and negative tests, respectively (Multimedia Appendix 14).

Sensitivity Analysis

Excluding the specific studies did not significantly change our research results (Table S7-S8 in Multimedia Appendix 1).

Quality Assessment

The quality of the 34 studies was evaluated using the QUADAS-AI tool (Multimedia Appendix 15). The QUADAS-AI specifically evaluates bias risk and applicability concerns in AI studies. Here, we observed that most studies had significant bias or applicability concerns, particularly regarding the selection of patients and index test. In the “patient selection” domain, 20 studies were classified as either high-risk or indeterminate due to reliance on closed-access data or failure to present the rationale and breakdown of its training, validation, and test sets. Only 7 externally validated studies were classified as low-risk in the “index test” category, while others showed elevated risks due to a lack of validation. In the “reference standard” assessment, the reference standard of all studies could be used to classify the target condition correctly. For the “flow and timing” assessment, 10 studies showed indeterminate risks due to insufficient justification for the timing between index and reference tests. Additionally, 20 studies presented significant concerns regarding applicability in the “patient selection” domain, receiving unclear ratings. In the “index test” domain, 7 studies were rated as having low applicability, while all studies received low applicability ratings in the “Reference Standard” domain.

Discussion

Principal Findings

This study represents the first systematic evaluation of ML models based on radiomics and DL models for the characterization of extracranial carotid plaques. Both approaches demonstrated robust diagnostic performance, with high SROC values of 0.95 and 0.92, respectively, highlighting their promising potential for clinical application in plaque detection and risk stratification.

Initially, the SP and SROC AUC of DL models were improved compared to ML models based on radiomics (0.91 vs 0.83; 0.95 vs 0.92), while their sensitivity was similar to that of ML (0.88). Moreover, we observed that radiomics and DL models used to identify the presence of plaques and stable plaques had similar diagnostic capabilities (SROC 0.96, 95% CI 0.94 - 0.97), and both were effective in identifying symptomatic plaques (SROC 0.90, 95% CI 0.87 - 0.92). Notably, these differences may not be simply due to model performance, but could result from a combination of different clinical objectives (simple exclusion diagnosis or differentiation of specific cases), imaging variations, and model techniques. By using knowledge gained from previous tasks, transfer learning enhances model performance on new datasets and minimizes data requirements. It has been successfully applied in various areas of cardiovascular disease to boost the performance of models [2,76,77]. In subgroup analyses, transfer learning significantly enhances model performance in data-limited scenarios and prevents overfitting. Large sample sizes can minimize sampling bias, decrease overfitting, and enhance the stability and reproducibility of the models. Moreover, we performed more detailed subgroup analyses based on the same imaging modality. Only the type of plaques in the ultrasound modality had sufficient data to perform statistical analysis and obtain summary diagnostic efficacy indicators. Results showed that ultrasound-based models have demonstrated excellent and similar performance in detecting the presence of plaques and assessing their stability. Considering the differences in equipment characteristics, patient demographics, and study design, these findings should be interpreted with caution. Nevertheless, these results provide valuable insights into the efficacy of radiomics algorithms and DL models in the diagnosis of carotid plaque.

Analysis of the Main Aspects

This meta-analysis demonstrates that radiomics-based models and DL models can diagnose extracranial carotid plaque, but the advantages of DL models in specificity and SROC should be interpreted with caution. A review of the included studies revealed that, among the 24 investigations using DL models, 20 primarily focused on plaque characterization (11 on the detection of plaques and 9 on plaque stability). Of these, 13 studies used ultrasound imaging to identify plaque-specific features such as echogenicity, morphology, and composition. In contrast, among the 10 studies using radiomics-based ML models, 6 were dedicated to identifying symptomatic plaques, predominantly using MRI (n=2) and CTA (n=3). The accuracy of symptomatic plaque identification was influenced not only

XSL•FO
RenderX

by intrinsic imaging characteristics but also by clinical indicators, including plaque rupture, thrombus formation, and the occurrence of cerebral hypoperfusion. The tasks were more complex, and model training seemed to focus on reducing false negatives to lower the risk of adverse outcomes such as stroke. In addition, traditional ML algorithms may rely on manual preprocessing and struggle to capture other subtle differences (such as the presence of tiny thrombi or fibrous cap thickness), which may introduce variability and additional costs. In contrast, the DL models (particularly convolutional neural networks) do not rely on artificially designed features; instead, they can directly process raw medical images, automatically filter noise, and automatically extract more meaningful image features (eg, slight echo attenuation behind plaques, differences in vascular wall elasticity, etc) [78]. It can also analyze the preset artificial extraction features, conduct independent learning, and uncover potential rules, thereby addressing the aforementioned challenges [23,79]. It is worth noting that a mismatch in the number of studies may also affect the interpretation of the results. Therefore, these differences may not be simply due to model performance, but could also be caused by multiple factors, which need to be further investigated.

Besides, the "black box" nature of AI algorithms, particularly DL models, raises concerns about the transparency and reliability of decision-making. Of the 34 studies reviewed, only 2 used explainable DL models, achieving an accuracy of 98.2% [37,65]. The explainable AI (XAI) approach leverages visualization techniques, feature attribution analysis, and both global and local explanations to clarify how models derive predictions from input data. By enhancing transparency, XAI fosters greater trust among medical professionals, strengthens model reliability and accountability, and helps mitigate concerns related to opaque decision-making [80]. The integration of XAI in medicine not only represents a technological advancement but also ensures safe, efficient, and robust medical decision-making, which needs to be further investigated. To realize this potential, a clinically oriented XAI implementation framework needs to be developed. First, the reporting criteria for interpretable techniques (including clinical applicability evaluation and operational guidelines) should be standardized to lower the threshold for physician use. Second, the design of algorithms should be optimized through collaborative efforts of medical professionals and engineers to improve the specificity of feature attribution methods based on real clinical needs. Further clinical validation studies are needed to evaluate the practical utility of XAI across diverse diagnostic settings—such as varying regions, hospital levels, and clinician experience—and to determine its true value in supporting clinical decision-making beyond algorithmic performance [28]. Furthermore, incomplete disclosure of model development processes in reports, selective presentation of results by investigators, and heterogeneity in diagnostic standard implementation across practitioners with different levels of experience may decrease the reliability and generalizability of findings. Therefore, we recommend the formulation of standardized imaging protocols, reporting procedures, and quality control measures for carotid plaque assessment and advocate for the establishment of specialized AI reporting guidelines for cardiovascular diseases.

Advances in imaging technology have now largely met the diagnostic requirements of current clinical practice, and current guidelines place heavy reliance on imaging tests for carotid plaque assessment. Among the 34 included studies, 27 constructed diagnostic models based only on imaging data. However, this should not be interpreted as rendering other clinical parameters irrelevant. Multidimensional diagnostic models combined with clinical features have been shown to achieve good diagnostic performance in identifying various diseases, such as pancreatic ductal adenocarcinoma [81], HCC recurrence after liver transplantation [82], hemorrhagic brain metastases [83], malignant BI-RADS 4 breast masses [84], and others. In our study, the diagnostic performance of combined models did not slightly improve, which may be due to the small sample size or some features could not provide more diagnostic information (for example, Hu et al [2] constructed a model relying only on indirect perivascular adipose tissue radiomic features and clinical features to identify symptomatic plaques, lacking direct imaging features). Considering this evidence, we strongly recommend that future research should aim to not only systematically incorporate laboratory tests, medical history, and other clinical parameters to develop multidimensional diagnostic models, but also to summarize the most meaningful features for specific types of plaques. This could address the limitations in current studies regarding single imaging modalities. This will also improve the precise classification of carotid plaques and personalized risk assessment.

This meta-analysis identified significant heterogeneity, while meta-regression and subgroup regression analysis did not identify the source, primarily attributable to the intrinsic challenges in regulating all potential confounding factors. Different imaging techniques can affect model performance based on the type of images used (static images vs dynamic videos), the equipment, and the operators. Guang et al [57] used a contrast-enhanced ultrasound video-based DL model to evaluate the diagnostic efficacy of a new carotid network structure for assessing carotid plaques, whereas other ultrasound studies consistently used static images. The sequence of MRI scans also influences diagnostic outcomes. Zhang et al [58] reported that a model incorporating a combination of T1-weighted, T2-weighted, dynamic contrast-enhanced, and postcontrast (POST) MRI sequences achieved a higher AUC for identifying high-risk carotid plaques compared to models using individual sequences or partial combinations. This enhanced performance is attributed to the complementary nature of these imaging sequences, each capturing distinct pathophysiological characteristics of the plaque, thereby improving diagnostic accuracy when used in combination. PRs have limited resolution, only detecting calcified components of carotid plaques and missing features such as lipid-rich necrotic cores or thin or ruptured fibrous caps. There are also notable differences in model architecture. Yoo et al [39] found performance variations among different convolutional neural network architectures within the CACSNet framework on the same dataset. Gui et al [49] compared multiple DL models (eg, 3D-DenseNet, 3D-SE-DenseNet) with 9 ML algorithms (including Decision Tree, Random Forest, SVM, etc) using identical datasets. They found that DL models generally performed better across key metrics like AUC and accuracy,

XSL•FO
RenderX

with significant performance differences between and within the two model types. These suggest that scanning parameters, model architectures, image segmentation, and algorithms may explain the heterogeneity in the research results. However, the small number of studies limits our ability to perform comprehensive subgroup analyses, which need to be further investigated.

The use of AI has significantly promoted the diagnosis of carotid plaque; however, its application requires cautious evaluation. Only 9 studies were multicenter (most used external validation), with diagnostic performance lower than single-center studies. Most studies (n=29) had a high risk of bias due to a lack of open-source data and external validation and failure to present the rationale and breakdown of its sets, which led to overestimation of the research results and affected the reproducibility and generalizability of the findings. Similar issues have been noted in previous reports, highlighting a broader deficiency in rigorous research standards within the field [85-87]. Furthermore, the contingency tables mostly come from the testing sets. Although the testing set achieved the best diagnostic performance, it had higher data quality or similar data distribution to the training, or overfitting noise, resulting in inaccurate performance estimation, and strong regularization may also decrease its performance, ultimately undermining clinical confidence in these models.

This study has certain clinical significance. We conducted an in-depth literature review and methodological quality evaluation, presenting the most current and comprehensive systematic review of AI-based diagnostic approaches for assessing carotid plaque. The findings reveal that AI technology shows considerable potential for diagnosing carotid plaque, but the findings need to be further validated by conducting more rigorous external validation using large-scale, high-quality independent datasets.

Limitations

This study has several limitations. First, the heterogeneity in model architectures and validation methods across studies prevents definitive conclusions regarding the most effective AI approaches. Second, many studies lack multicenter external validation, leading to a high risk of bias. The model overfitting and clinical applicability need to be carefully evaluated. Third, meta-regression and subgroup analysis did not identify the sources of high heterogeneity that existed in most of the included studies. We hypothesize that this heterogeneity may be caused by scanning parameters, model architectures, image segmentation, and algorithms. However, the overly scattered distribution of subgroups due to the limited number of studies restricts more in-depth subgroup analyses. Finally, although the Deeks test did not show significant publication bias, the included studies may have intentionally unreported negative results and omitted potentially relevant non-English literature.

Future studies should use a more comprehensive analytical methodology based on the current model. Researchers should strictly follow regulatory norms and standardized operating procedures. Prospective and multicenter studies and additional external validation are warranted to enhance the robustness and generalizability of the existing models. In the future, researchers should perform independent systematic reviews on specific subtopics—such as imaging modalities, lesion types, or model architectures—to facilitate targeted evaluations of AI performance across distinct clinical scenarios. In addition, studies on imaging modalities such as CT and MRI are advocated to generate more data, conduct subgroup analyses, and clarify the optimal matching of modality, plaque type, and algorithm. Future efforts should focus on identifying more meaningful features and building and evaluating the diagnostic performance of multidimensional diagnostic models. In parallel, establishing clinically oriented, XAI frameworks will be essential for enhancing transparency.

Conclusions

Current findings indicate that radiomics algorithms and DL models can effectively diagnose extracranial carotid plaque. However, the irregularities in research design and the lack of multicenter studies and external validation limit the robustness of the present findings. Future research should aim to reduce bias risk and enhance the generalizability and clinical orientation of the models.

Acknowledgments

The manuscript was written without the use of ChatGPT or other generative language models.

Funding

The conduct of this study, the writing of the manuscript, and its publication did not receive any external financial support or grants from any public, commercial, or nonprofit entities.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: LJ (lead), JR (equal)
Methodology: LJ (lead), YG (equal), HG (supporting)
Data curation: LJ (lead), YG (supporting)
Investigation: LJ (lead), YG (supporting), HG (supporting)

XSL•FO
RenderX

Software: LJ (lead), RL (equal), YW (supporting)
Supervision: JR (lead), NM (supporting)
Validation: LJ (lead), RL (supporting), YW (supporting)
Visualization: LJ (lead), YG (supporting), HG (supporting)
Writing – original draft: LJ (lead), YG (supporting)
Writing – review & editing: LJ (lead), RL (supporting), YW (supporting), SW (supporting), NM (supporting), JR (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete supplementary data tables for the systematic review and meta-analysis.
[DOC File, 259 KB - jmir_v28i1e77092_app1.doc]

Multimedia Appendix 2
Combined diagnostic performance estimates from the included studies (34 studies with 34 tables) [36-69].
[PNG File, 381 KB - jmir_v28i1e77092_app2.png]

Multimedia Appendix 3
Sensitivity and specificity of deep learning (DL) versus radiomics-based machine learning (ML) models. (A) DL models (24 studies with 24 tables). (B) ML models based on radiomics algorithms (10 studies with 10 tables) [36-69].
[JPG File, 601 KB - jmir_v28i1e77092_app3.jpg]

Multimedia Appendix 4
Publication bias.
[PNG File, 73 KB - jmir_v28i1e77092_app4.png]

Multimedia Appendix 5
Sensitivity and specificity for different medical imaging modalities. (A) Models based on periapical radiographs (PRs) imaging (5 studies with 5 tables). (B) Models based on ultrasound imaging (16 studies with 16 tables). (C) Models based on magnetic resonance imaging (MRI) imaging (5 studies with 5 tables). (D) Models based on computed tomography angiography (CTA) imaging (8 studies with 8 tables). (E) Models based on ultrasound modality for detecting the presence of carotid plaque (5 studies with 5 tables). (F) Models based on ultrasound modality for distinguishing the stability of carotid plaques (8 studies with 8 tables) [36-69].
[JPG File, 879 KB - jmir_v28i1e77092_app5.jpg]

Multimedia Appendix 6
Sensitivity and specificity of models for using transfer learning or not. (A) Models using transfer learning (10 studies with 10 tables). (B) Models without transfer learning (24 studies with 24 tables) [36-69].
[JPG File, 584 KB - jmir_v28i1e77092_app6.jpg]

Multimedia Appendix 7
Sensitivity and specificity for different diagnostic tasks. (A) Presence or absence of carotid plaques (11 studies with 11 tables). (B) Stable or vulnerable carotid plaques (12 studies with 12 tables). (C) Symptomatic or asymptomatic carotid plaques (10 studies with 11 tables) [36-69].
[JPG File, 743 KB - jmir_v28i1e77092_app7.jpg]

Multimedia Appendix 8
Sensitivity and specificity of pure artificial intelligence models or models constructed by combining clinical features for carotid plaques. (A) Combined models (7 studies with 7 tables). (B) Artificial intelligence models (7 studies with 7 tables) [50,52,58,63,64,66,68].
[JPG File, 416 KB - jmir_v28i1e77092_app8.jpg]

Multimedia Appendix 9
Sensitivity and specificity for different dataset types. (A) Testing (27 studies with 27 tables). (B) External validation (7 studies with 7 tables) [36-69].

XSL•FO
RenderX

[JPG File, 617 KB - jmir_v28i1e77092_app9.jpg]

Multimedia Appendix 10
Sensitivity and specificity for different risk of-bias studies. (A) Low risk of bias studies (5 studies with 5 tables). (B) High/unclear risk of bias studies (29 studies with 29 tables) [36-69].
[JPG File, 631 KB - jmir_v28i1e77092_app10.jpg]

Multimedia Appendix 11
Sensitivity and specificity of study using different sample sizes. (A) Sample size ≥200 (14 studies with 14 tables). (B) Sample size <200 (20 studies with 20 tables) [36-69].
[JPG File, 583 KB - jmir_v28i1e77092_app11.jpg]

Multimedia Appendix 12
Sensitivity and specificity for different research designs. (A) Multicenter studies (9 studies with 9 tables). (B) Single-center studies (22 studies with 22 tables) [36,38,39,41-52,54-69].
[JPG File, 578 KB - jmir_v28i1e77092_app12.jpg]

Multimedia Appendix 13
Exploration of potential sources of heterogeneity across multiple variables. (A) Different algorithms (deep learning models or machine learning models based radiomics algorithms). (B) Different medical imaging modalities. (C) Different carotid plaque type. (D) Utilizing transfer learning or not. (E) Different sets of datasets. (F) Different sample size. (G) Single-center or multicenter studies. (H) Pure artificial intelligence (AI) models or models constructed by combining clinical features. (I) Different risk of bias studies [36-69].
[JPG File, 1753 KB - jmir_v28i1e77092_app13.jpg]

Multimedia Appendix 14
The Fagan nomogram assesses the diagnostic ability of radiomics and deep learning to carotid plaques.
[PNG File, 186 KB - jmir_v28i1e77092_app14.png]

Multimedia Appendix 15
Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence (QUADAS-AI) for the assessment of the methodological qualities of all the enrolled studies.
[PNG File, 240 KB - jmir_v28i1e77092_app15.png]

Checklist 1
PRISMA 2020 checklist.
[PDF File, 99 KB - jmir_v28i1e77092_app16.pdf]

References

1. Song P, Fang Z, Wang H, et al. Global and regional prevalence, burden, and risk factors for carotid atherosclerosis: a systematic review, meta-analysis, and modelling study. Lancet Glob Health 2020 May;8(5):e721-e729. [doi: 10.1016/S2214-109X(20)30117-0] [Medline: 32353319]
2. Evans NR, Bhakta S, Chowdhury MM, Markus H, Warburton E. Management of carotid atherosclerosis in stroke. Pract Neurol 2024 Sep 13;24(5):382-386. [doi: 10.1136/pn-2023-003918] [Medline: 38589215]
3. Oushy SH, Essibayi MA, Savastano LE, Lanzino G. Carotid artery revascularization: endarterectomy versus endovascular therapy. J Neurosurg Sci 2021 Jun;65(3):322-326. [doi: 10.23736/S0390-5616.20.05207-8] [Medline: 33297612]
4. Kopczak A, Schindler A, Bayer-Karpinska A, et al. Complicated carotid artery plaques as a cause of cryptogenic stroke. J Am Coll Cardiol 2020 Nov 10;76(19):2212-2222. [doi: 10.1016/j.jacc.2020.09.532] [Medline: 33153580]
5. Abbott AL, Paraskevas KI, Kakkos SK, et al. Systematic review of guidelines for the management of asymptomatic and symptomatic carotid stenosisa. Stroke 2015 Nov;46(11):3288-3301. [doi: 10.1161/STROKEAHA.115.003390] [Medline: 26451020]
6. Brinjikji W, Huston J, Rabinstein AA, Kim GM, Lerman A, Lanzino G. Contemporary carotid imaging: from degree of stenosis to plaque vulnerability. JNS 2016 Jan;124(1):27-42. [doi: 10.3171/2015.1.JNS142452]
7. Paju S, Pietiäinen M, Liljestrand JM, et al. Carotid artery calcification in panoramic radiographs associates with oral infections and mortality. Int Endodontic J 2021 Apr;54(4):638-638. [doi: 10.1111/iej.13451]

XSL•FO
RenderX

8. Bengtsson VW, Persson GR, Renvert S. Assessment of carotid calcifications on panoramic radiographs in relation to other used methods and relationship to periodontitis and stroke: a literature review. Acta Odontol Scand 2014 Aug;72(6):401-412. [doi: 10.3109/00016357.2013.847489] [Medline: 24432815]
9. Schroder AGD, de Araujo CM, Guariza-Filho O, Flores-Mir C, de Luca Canto G, Porporatti AL. Diagnostic accuracy of panoramic radiography in the detection of calcified carotid artery atheroma: a meta-analysis. Clin Oral Investig 2019 May;23(5):2021-2040. [doi: 10.1007/s00784-019-02880-6] [Medline: 30923911]
10. Zeng P, Zhang Q, Liang X, Zhang M, Luo D, Chen Z. Progress of ultrasound techniques in the evaluation of carotid vulnerable plaque neovascularization. Cerebrovasc Dis 2024;53(4):479-487. [doi: 10.1159/000534372] [Medline: 37812915]
11. Guo Y, Wang X, Wang L, et al. The value of superb microvascular imaging and contrast-enhanced ultrasound for the evaluation of neovascularization in carotid artery plaques. Acad Radiol 2023 Mar;30(3):403-411. [doi: 10.1016/j.acra.2022.08.001] [Medline: 36123231]
12. Chen X, Wang H, Jiang Y, et al. Neovascularization in carotid atherosclerotic plaques can be effectively evaluated by superb microvascular imaging (SMI): Initial experience. Vasc Med 2020 Aug;25(4):328-333. [doi: 10.1177/1358863X20909992] [Medline: 32303154]
13. Li C, He W, Guo D, et al. Quantification of carotid plaque neovascularization using contrast-enhanced ultrasound with histopathologic validation. Ultrasound Med Biol 2014 Aug;40(8):1827-1833. [doi: 10.1016/j.ultrasmedbio.2014.02.010] [Medline: 24798387]
14. Hoogi A, Adam D, Hoffman A, Kerner H, Reisner S, Gaitini D. Carotid plaque vulnerability: quantification of neovascularization on contrast-enhanced ultrasound with histopathologic correlation. AJR Am J Roentgenol 2011 Feb;196(2):431-436. [doi: 10.2214/AJR.10.4522] [Medline: 21257897]
15. Zamani M, Skagen K, Scott H, Russell D, Skjelland M. Advanced ultrasound methods in assessment of carotid plaque instability: a prospective multimodal study. BMC Neurol 2020 Jan 29;20(1):39. [doi: 10.1186/s12883-020-1620-z] [Medline: 31996153]
16. Saba L, Caddeo G, Sanfilippo R, Montisci R, Mallarini G. Efficacy and sensitivity of axial scans and different reconstruction methods in the study of the ulcerated carotid plaque using multidetector-row CT angiography: comparison with surgical results. AJNR Am J Neuroradiol 2007 Apr;28(4):716-723. [Medline: 17416828]
17. Horev A, Honig A, Cohen JE, et al. Overestimation of carotid stenosis on CTA - real world experience. J Clin Neurosci 2021 Mar;85:36-40. [doi: 10.1016/j.jocn.2020.12.018] [Medline: 33581787]
18. Baradaran H, Gupta A. Carotid vessel wall imaging on CTA. AJNR Am J Neuroradiol 2020 Mar;41(3):380-386. [doi: 10.3174/ajnr.A6403] [Medline: 32029468]
19. Saba L, Yuan C, Hatsukami TS, et al. Carotid artery wall imaging: perspective and guidelines from the ASNR vessel wall imaging study group and expert consensus recommendations of the American Society of Neuroradiology. AJNR Am J Neuroradiol 2018 Feb;39(2):E9-E31. [doi: 10.3174/ajnr.A5488] [Medline: 29326139]
20. Xu HL, Gong TT, Song XJ, et al. Artificial intelligence performance in image-based cancer identification: umbrella review of systematic reviews. J Med Internet Res 2025 Apr 1;27:e53567. [doi: 10.2196/53567] [Medline: 40167239]
21. Scicolone R, Vacca S, Pisu F, et al. Radiomics and artificial intelligence: general notions and applications in the carotid vulnerable plaque. Eur J Radiol 2024 Jul;176:111497. [doi: 10.1016/j.ejrad.2024.111497] [Medline: 38749095]
22. Koçak B. Key concepts, common pitfalls, and best practices in artificial intelligence and machine learning: focus on radiomics. Diagn Interv Radiol 2022 Sep;28(5):450-462. [doi: 10.5152/dir.2022.211297] [Medline: 36218149]
23. Zhou T, Cheng Q, Lu H, Li Q, Zhang X, Qiu S. Deep learning methods for medical image fusion: a review. Comput Biol Med 2023 Jun;160:106959. [doi: 10.1016/j.compbiomed.2023.106959] [Medline: 37141652]
24. Ma Y, Li M, Wu H. The machine learning models in major cardiovascular adverse events prediction based on coronary computed tomography angiography: systematic review. J Med Internet Res 2025 Jun 13;27:e68872. [doi: 10.2196/68872] [Medline: 40513092]
25. Oikonomou EK, Williams MC, Kotanidis CP, et al. A novel machine learning-derived radiotranscriptomic signature of perivascular fat improves cardiac risk prediction using coronary CT angiography. Eur Heart J 2019 Nov 14;40(43):3529-3543. [doi: 10.1093/eurheartj/ehz592] [Medline: 31504423]
26. Pan J, Huang Q, Zhu J, et al. Prediction of plaque progression using different machine learning models of pericoronary adipose tissue radiomics based on coronary computed tomography angiography. Eur J Radiol Open 2025 Jun;14:100638. [doi: 10.1016/j.ejro.2025.100638] [Medline: 40034660]
27. Cohen JF, Deeks JJ, Hooft L, et al. Preferred reporting items for journal and conference abstracts of systematic reviews and meta-analyses of diagnostic test accuracy studies (PRISMA-DTA for Abstracts): checklist, explanation, and elaboration. BMJ 2021 Mar 15;372:n265. [doi: 10.1136/bmj.n265] [Medline: 33722791]
28. Bayor AA, Li J, Yang IA, Varnfield M. Designing Clinical Decision Support Systems (CDSS)-A user-centered lens of the design characteristics, challenges, and implications: systematic review. J Med Internet Res 2025 Jun 20;27:e63733. [doi: 10.2196/63733] [Medline: 40540451]
29. McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: the PRISMA-DTA statement. JAMA 2018 Jan 23;319(4):388-396. [doi: 10.1001/jama.2017.19163] [Medline: 29362800]

XSL•FO
RenderX

30. Lee J, Mulder F, Leeflang M, Wolff R, Whiting P, Bossuyt PM. QUAPAS: an adaptation of the QUADAS-2 tool to assess prognostic accuracy studies. Ann Intern Med 2022 Jul;175(7):1010-1018. [doi: 10.7326/M22-0276] [Medline: 35696685]
31. Sounderajah V, Ashrafian H, Rose S, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. Nat Med 2021 Oct;27(10):1663-1665. [doi: 10.1038/s41591-021-01517-0] [Medline: 34635854]
32. Guni A, Sounderajah V, Whiting P, Bossuyt P, Darzi A, Ashrafian H. Revised tool for the Quality Assessment of Diagnostic Accuracy Studies Using AI (QUADAS-AI): protocol for a qualitative study. JMIR Res Protoc 2024 Sep 18;13:e58202. [doi: 10.2196/58202] [Medline: 39293047]
33. Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. NPJ Digit Med 2020;3:126. [doi: 10.1038/s41746-020-00333-z] [Medline: 33043150]
34. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005 Oct;58(10):982-990. [doi: 10.1016/j.jclinepi.2005.02.022] [Medline: 16168343]
35. Samawi H, Alsharman M, Keko M, Kersey J. Post-test diagnostic accuracy measures under tree ordering of disease classes. Stat Med 2023 Dec 10;42(28):5135-5159. [doi: 10.1002/sim.9905] [Medline: 37720999]
36. Molinari F, Raghavendra U, Gudigar A, Meiburger KM, Rajendra Acharya U. An efficient data mining framework for the characterization of symptomatic and asymptomatic carotid plaque using bidimensional empirical mode decomposition technique. Med Biol Eng Comput 2018 Sep;56(9):1579-1593. [doi: 10.1007/s11517-018-1792-5] [Medline: 29473126]
37. Singh S, Jain PK, Sharma N, Pohit M, Roy S. Atherosclerotic plaque classification in carotid ultrasound images using machine learning and explainable deep learning. Intelligent Medicine 2024 May;4(2):83-95. [doi: 10.1016/j.imed.2023.05.003]
38. Li J, Huang Y, Song S, et al. Automatic diagnosis of carotid atherosclerosis using a portable freehand 3-D ultrasound imaging system. IEEE Trans Ultrason Ferroelectr Freq Control 2024 Feb;71(2):266-279. [doi: 10.1109/TUFFC.2023.3345740] [Medline: 38127609]
39. Yoo SW, Yang S, Kim JE, et al. CACSNet for automatic robust classification and segmentation of carotid artery calcification on panoramic radiographs using a cascaded deep learning network. Sci Rep 2024 Jun 17;14(1):13894. [doi: 10.1038/s41598-024-64265-4] [Medline: 38886356]
40. Omarov M, Zhang L, Jorshery SD, et al. Automated deep learning-based detection of early atherosclerotic plaques in carotid ultrasound imaging. medRxiv 2025 Sep 3:2024.10.17.24315675. [doi: 10.1101/2024.10.17.24315675] [Medline: 39484270]
41. Zhai D, Liu R, Liu Y, et al. Deep learning-based fully automatic screening of carotid artery plaques in computed tomography angiography: a multicenter study. Clin Radiol 2024 Aug;79(8):e994-e1002. [doi: 10.1016/j.crad.2024.04.015] [Medline: 38789330]
42. Vinayahalingam S, van Nistelrooij N, Xi T, et al. Detection of carotid plaques on panoramic radiographs using deep learning. J Dent 2024 Dec;151:105432. [doi: 10.1016/j.jdent.2024.105432] [Medline: 39461583]
43. Pisu F, Williamson BJ, Nardi V, et al. Machine learning detects symptomatic plaques in patients with carotid atherosclerosis on CT angiography. Circ Cardiovasc Imaging 2024 Jun;17(6):e016274. [doi: 10.1161/CIRCIMAGING.123.016274] [Medline: 38889214]
44. Zhou R, Gan W, Wang F, Yang Z, Huang Z, Gan H. Tri-correcting: label noise correction via triple CNN ensemble for carotid plaque ultrasound image classification. Biomed Signal Process Control 2024 May;91:105981. [doi: 10.1016/j.bspc.2024.105981]
45. Wang Y, Cai C, Du YM, et al. Assessment of stroke risk using MRI-VPD with automatic segmentation of carotid plaques and classification of plaque properties based on deep learning. J Radiat Res Appl Sci 2023 Sep;16(3):100630. [doi: 10.1016/j.jrras.2023.100630]
46. Shan D, Wang S, Wang J, et al. Computed tomography angiography-based radiomics model for predicting carotid atherosclerotic plaque vulnerability. Front Neurol 2023;14:1151326. [doi: 10.3389/fneur.2023.1151326] [Medline: 37396779]
47. Xie J, Li Y, Xu X, et al. CPTV: classification by tracking of carotid plaque in ultrasound videos. Comput Med Imaging Graph 2023 Mar;104:102175. [doi: 10.1016/j.compmedimag.2022.102175] [Medline: 36630795]
48. Amitay M, Barnett-Itzhaki Z, Sudri S, et al. Deep convolution neural network for screening carotid calcification in dental panoramic radiographs. PLOS Digit Health 2023 Apr;2(4):e0000081. [doi: 10.1371/journal.pdig.0000081] [Medline: 37043433]
49. Gui C, Cao C, Zhang X, Zhang J, Ni G, Ming D. Radiomics and artificial neural networks modelling for identification of high-risk carotid plaques. Front Cardiovasc Med 2023;10:1173769. [doi: 10.3389/fcvm.2023.1173769] [Medline: 37485276]
50. Shi J, Sun Y, Hou J, et al. Radiomics signatures of carotid plaque on computed tomography angiography: an approach to identify symptomatic plaques. Clin Neuroradiol 2023 Dec;33(4):931-941. [doi: 10.1007/s00062-023-01289-9] [Medline: 37195452]
51. Su SS, Li LY, Wang Y, Li YZ. Stroke risk prediction by color Doppler ultrasound of carotid artery-based deep learning using Inception V3 and VGG-16. Front Neurol 2023;14:1111906. [doi: 10.3389/fneur.2023.1111906] [Medline: 36864909]
52. Chen S, Liu C, Chen X, Liu WV, Ma L, Zha Y. A radiomics approach to assess high risk carotid plaques: a non-invasive imaging biomarker, retrospective study. Front Neurol 2022;13:35350403. [doi: 10.3389/fneur.2022.788652]

XSL•FO
RenderX

53. Gago L, Vila MDM, Grau M, Remeseiro B, Igual L. An end-to-end framework for intima media measurement and atherosclerotic plaque detection in the carotid artery. Comput Methods Programs Biomed 2022 Aug;223:106954. [doi: 10.1016/j.cmpb.2022.106954] [Medline: 35777216]
54. Jain PK, Sharma N, Saba L, et al. Automated deep learning-based paradigm for high-risk plaque detection in B-mode common carotid ultrasound scans: an asymptomatic Japanese cohort study. Int Angiol 2022 Feb;41(1):9-23. [doi: 10.23736/S0392-9590.21.04771-4] [Medline: 34825801]
55. Cilla S, Macchia G, Lenkowicz J, et al. CT angiography-based radiomics as a tool for carotid plaque characterization: a pilot study. Radiol Med 2022 Jul;127(7):743-753. [doi: 10.1007/s11547-022-01505-5] [Medline: 35680773]
56. Xu X, Huang L, Wu R, et al. Multi-feature fusion method for identifying carotid artery vulnerable plaque. IRBM 2022 Aug;43(4):272-278. [doi: 10.1016/j.irbm.2021.07.004]
57. Guang Y, He W, Ning B, et al. Deep learning-based carotid plaque vulnerability classification with multicentre contrast-enhanced ultrasound video: a comparative diagnostic study. BMJ Open 2021 Aug 27;11(8):e047528. [doi: 10.1136/bmjopen-2020-047528] [Medline: 34452961]
58. Zhang R, Zhang Q, Ji A, et al. Identification of high-risk carotid plaque with MRI-based radiomics and machine learning. Eur Radiol 2021 May;31(5):3116-3126. [doi: 10.1007/s00330-020-07361-z] [Medline: 33068185]
59. Ma W, Cheng X, Xu X, et al. Multilevel strip pooling-based convolutional neural network for the classification of carotid plaque echogenicity. Comput Math Methods Med 2021;2021:3425893. [doi: 10.1155/2021/3425893] [Medline: 34457035]
60. Ganitidis T, Athanasiou M, Dalakleidi K, Melanitis N, Golemati S, Nikita KS. Stratification of carotid atheromatous plaque using interpretable deep learning methods on B-mode ultrasound images. Annu Int Conf IEEE Eng Med Biol Soc 2021 Nov;2021:3902-3905. [doi: 10.1109/EMBC46164.2021.9630402] [Medline: 34892085]
61. Kats L, Vered M, Zlotogorski-Hurvitz A, Harpaz I. Atherosclerotic carotid plaque on panoramic radiographs: neural network detection. Int J Comput Dent 2019;22(2):163-169. [Medline: 31134222]
62. Wei Y, Yang B, Wei L, et al. Real-time carotid plaque recognition from dynamic ultrasound videos based on artificial neural network. Ultraschall Med 2024 Oct;45(5):493-500. [doi: 10.1055/a-2180-8405] [Medline: 38113893]
63. Zhao T, Lin G, Chen W, et al. Predicting symptomatic carotid artery plaques with radiomics-based carotid perivascular adipose tissue characteristics: a multicenter, multiclassifier study. BMC Med Imaging 2025 Aug 19;25(1):337. [doi: 10.1186/s12880-025-01876-x] [Medline: 40830841]
64. Hu W, Lin G, Chen W, et al. Radiomics based on dual-energy CT virtual monoenergetic images to identify symptomatic carotid plaques: a multicenter study. Sci Rep 2025 Mar 26;15(1):10415. [doi: 10.1038/s41598-025-92855-3] [Medline: 40140428]
65. Liapi GD, Loizou CP, Griffin M, Pattichis CS, Nicolaides A, Kyriacou E. Transfer learning with class activation maps in compositions driving plaque classification in carotid ultrasound. Front Digit Health 2025;7:1484231. [doi: 10.3389/fdgth.2025.1484231] [Medline: 40704367]
66. Yu F, Li X, Zhang Y, et al. MRI ensemble model of plaque and perivascular adipose tissue as PET-equivalent for identifying carotid atherosclerotic inflammation. EJNMMI Res 2025 Aug 6;15(1):103. [doi: 10.1186/s13550-025-01293-9] [Medline: 40768105]
67. Kuwada C, Mitsuya Y, Fukuda M, et al. Area detection improves the person-based performance of a deep learning system for classifying the presence of carotid artery calcifications on panoramic radiographs. Oral Radiol 2025 Jul 22;2025(1-10). [doi: 10.1007/s11282-025-00843-0] [Medline: 40694246]
68. Lao Q, Zhou R, Wu Y, et al. Predicting vulnerability status of carotid plaques using CTA-based quantitative analysis. J Cardiovasc Pharmacol 2025 Mar 1;85(3):217-224. [doi: 10.1097/FJC.0000000000001664] [Medline: 39739382]
69. He L, Yang Z, Wang Y, et al. A deep learning algorithm to identify carotid plaques and assess their stability. Front Artif Intell 2024;7:1321884. [doi: 10.3389/frai.2024.1321884] [Medline: 38952409]
70. Zhang Y, Gan H, Wang F, et al. A self-supervised fusion network for carotid plaque ultrasound image classification. Math Biosci Eng 2024 Jan 31;21(2):3110-3128. [doi: 10.3934/mbe.2024138] [Medline: 38454721]
71. Ali T, Pathan S, Salvi M, Meiburger KM, Molinari F, Acharya UR. CAROTIDNet: a novel carotid symptomatic/asymptomatic plaque detection system using CNN-based tangent optimization algorithm in B-mode ultrasound images. IEEE Access 2024;12:73970-73979. [doi: 10.1109/ACCESS.2024.3404023]
72. Ayoub M, Liao Z, Li L, Wong KKL. HViT: hybrid vision inspired transformer for the assessment of carotid artery plaque by addressing the cross-modality domain adaptation problem in MRI. Comput Med Imaging Graph 2023 Oct;109:102295. [doi: 10.1016/j.compmedimag.2023.102295] [Medline: 37717365]
73. Latha S, Muthu P, Lai KW, Khalil A, Dhanalakshmi S. Performance analysis of machine learning and deep learning architectures on early stroke detection using carotid artery ultrasound images. Front Aging Neurosci 2021;13:828214. [doi: 10.3389/fnagi.2021.828214] [Medline: 35153728]
74. Wang L, Guo T, Wang L, et al. Improving radiomic modeling for the identification of symptomatic carotid atherosclerotic plaques using deep learning-based 3D super-resolution CT angiography. Heliyon 2024 Apr 30;10(8):e29331. [doi: 10.1016/j.heliyon.2024.e29331] [Medline: 38644848]
75. Li YC, Zhang TR, Zhang F, et al. Development and validation of a carotid plaque risk prediction model for coal miners. Front Cardiovasc Med 2025;12:1490961. [doi: 10.3389/fcvm.2025.1490961] [Medline: 40416817]

XSL•FO
RenderX

76. Weimann K, Conrad TOF. Transfer learning for ECG classification. Sci Rep 2021 Mar 4;11(1):5251. [doi: 10.1038/s41598-021-84374-8] [Medline: 33664343]
77. Chiu IM, Cheng JY, Chen TY, et al. Using deep transfer learning to detect hyperkalemia from ambulatory electrocardiogram monitors in intensive care units: personalized medicine approach. J Med Internet Res 2022 Dec 5;24(12):e41163. [doi: 10.2196/41163] [Medline: 36469396]
78. van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal 2022 Jul;79:102470. [doi: 10.1016/j.media.2022.102470] [Medline: 35576821]
79. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol 2020 Feb 27;9(2):14. [doi: 10.1167/tvst.9.2.14] [Medline: 32704420]
80. de Vries BM, Zwezerijnen GJC, Burchell GL, van Velden FHP, Menke-van der Houven van Oordt CW, Boellaard R. Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review. Front Med (Lausanne) 2023;10:1180773. [doi: 10.3389/fmed.2023.1180773] [Medline: 37250654]
81. Huang Y, Zhang H, Ding Q, et al. Comparison of multiple machine learning models for predicting prognosis of pancreatic ductal adenocarcinoma based on contrast-enhanced CT radiomics and clinical features. Front Oncol 2024;14:1419297. [doi: 10.3389/fonc.2024.1419297] [Medline: 39605884]
82. Schindler P, von Beauvais P, Hoffmann E, et al. Combining radiomics and imaging biomarkers with clinical variables for the prediction of HCC recurrence after liver transplantation. Liver Transpl 2025 Oct 1;31(10):1226-1237. [doi: 10.1097/LVT.0000000000000603] [Medline: 40100771]
83. Cui L, Yu L, Shao S, et al. Improving differentiation of hemorrhagic brain metastases from non-neoplastic hematomas using radiomics and clinical feature fusion. Neuroradiology 2025 Jun;67(6):1455-1468. [doi: 10.1007/s00234-025-03590-5] [Medline: 40131431]
84. Zhang Q, Gao J, Agyekum EA, et al. A combined clinical-ultrasound radiomics model for differentiating benign and malignant BI-RADS category 4 breast masses. Am J Transl Res 2025;17(8):6370-6380. [doi: 10.62347/SBKU2090] [Medline: 40950304]
85. Fu Y, Huang Z, Deng X, et al. Artificial intelligence in lymphoma histopathology: systematic review. J Med Internet Res 2025 Feb 14;27:e62851. [doi: 10.2196/62851] [Medline: 39951716]
86. Wu Y, Chao J, Bao M, Zhang N. Predictive value of machine learning on fracture risk in osteoporosis: a systematic review and meta-analysis. BMJ Open 2023 Dec;13(12):e071430. [doi: 10.1136/bmjopen-2022-071430]
87. Mäkitie AA, Alabi RO, Ng SP, et al. Artificial intelligence in head and neck cancer: a systematic review of systematic reviews. Adv Ther 2023 Aug;40(8):3360-3380. [doi: 10.1007/s12325-023-02527-9] [Medline: 37291378]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
BI-RADS: Breast Imaging Reporting and Data System
CTA: computed tomography angiography
DL: deep learning
IEEE: Institute of Electrical and Electronics Engineers
LR: likelihood ratio
ML: machine learning
MRI: magnetic resonance imaging
P-post: posttest probability
PR: periapical radiograph
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy Studies
PROSPERO: International Prospective Register of Systematic Reviews
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2
QUADAS-AI: Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence
ROC: receiver operating characteristic
SROC: summary receiver operating characteristic
SROC AUC: area under the summary receiver operating characteristic curve
XAI: explainable AI

XSL•FO
RenderX

Edited by A Coristine; submitted 07.May.2025; peer-reviewed by MA Ashoobi, R Orozco, Z Fang; accepted 17.Nov.2025; published 22.Jan.2026.

Please cite as:
Ju L, Guo Y, Guo H, Liu R, Wang Y, Wang S, Ma N, Ren J
Diagnostic Performance of Deep Learning and Radiomics in Extracranial Carotid Plaque Detection: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e77092
URL: https://www.jmir.org/2026/1/e77092
doi:10.2196/77092

© Lingjie Ju, Yongsheng Guo, Haiyong Guo, Ruijuan Liu, Yiyang Wang, Siyu Wang, Na Ma, Junhong Ren. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evidence for Digital Health Tools Designed to Support the Triage of Musculoskeletal Conditions in Primary, Urgent, and Emergency Care Settings: Scoping Review

Linda K Truong[1,2], PT, PhD; James G Wrightson[3], PhD; Raphaël Vincent[4,5,6], PT, MSc; Eunice Lui[1], MPH; Jamon L Couch[7,8], PT; Ellen Wang[2,8], MSc; Cobie Starcevich[9,10], PT; Dean Giustini[11], PhD; Alex Haagaard[12], MA; Elena Lopatina[13,14], MD, PhD; Niels van Berkel[15], PhD; Michael Skovdal Rathleff[16,17], PT, PhD; Clare L Ardern[1,2,7], PT, PhD

[1]Centre for Aging SMART, University of British Columbia, Vancouver, BC, Canada
[10]Physiotherapy Department, Rockingham General Hospital, South Metropolitan Health Service, Perth, Australia
[11]Biomedical Branch Library, University of British Columbia, Vancouver, BC, Canada
[12]Pain BC, Vancouver, BC, Canada
[13]University of Calgary, Calgary, AB, Canada
[14]Primary Care Alberta, Calgary, AB, Canada
[15]Department of Computer Science, Aalborg University, Aalborg, Denmark
[16]Department of Health Science and Technology, Faculty of Medicine, Aalborg University, Aalborg, Denmark
[17]Center for General Practice at Aalborg University, Aalborg, Denmark
[2]Department of Physical Therapy, Faculty of Medicine, University of British Columbia, 13737 96 Avenue, Surrey, BC, Canada
[3]Department of Family Practice, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada
[4]School of Rehabilitation, Faculty of Medicine, Université de Montréal, Montreal, QC, Canada
[5]Hôpital Maisonneuve-Rosemont Research Center, Université de Montréal Affiliated Research Center, Montreal, QC, Canada
[6]Centre for Interdisciplinary Research in Rehabilitation of Greater Montreal, Institut Universitaire sur la Réadaptation en Déficience Physique de Montréal, Montreal, QC, Canada
[7]La Trobe Sport and Exercise Medicine Research Centre, La Trobe University, Melbourne, Australia
[8]Arthritis Research Canada, Vancouver, BC, Canada
[9]School of Allied Health, Faculty of Health Sciences, Curtin University, Perth, Australia

Corresponding Author:
Linda K Truong, PT, PhD
Centre for Aging SMART, University of British Columbia, Vancouver, BC, Canada

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2026/1/e92722

Abstract

Background: The digital health research field is growing rapidly, and a summary of the available digital tools for triaging musculoskeletal conditions is needed. Effective and safe digital triage tools for musculoskeletal conditions could support patients and clinicians in making informed care decisions and may contribute to reducing emergency department overcrowding and health care costs.

Objective: The aim of the study is to identify and describe digital health tools for use by adults to triage musculoskeletal conditions across primary, urgent, or emergency care settings.

Methods: Our scoping review was conducted following the Johanna Briggs Institute recommendations for scoping reviews and Arksey and O'Malley's framework. Systematic searches in MEDLINE (OVID), Embase (OVID), CENTRAL (OVID), CINAHL (EBSCO), Compendex, Web of Science, OpenGrey, Google Scholar, arXiv, medRxiv, and an extensive gray literature search were conducted with a librarian scientist from inception to September 18, 2025. Studies had to recruit adults (aged 18 years and older) with musculoskeletal conditions that identified a digital health tool designed to triage or diagnose in primary, urgent, or emergency care settings and report primary data to be included. In total, 2 reviewer pairs independently screened abstracts and full-text papers. Relevant data were extracted in duplicate, and results were summarized descriptively.

XSL•FO
RenderX

Results: The search yielded 5695 records, and we screened 189 full-text papers. In total, 34 studies (n=37,509 patients) met the inclusion criteria. The most common musculoskeletal conditions reported were rheumatoid or inflammatory arthritis (13/34, 38%). In total, 19 (19/34, 56%) studies reported on symptom checkers, 13 (13/34, 38%) studies on triage or diagnosis tools, and 2 (2/34, 6%) were studies of diagnostic predictor tools. There were 16 unique digital health tools. A total of 2 tools were built for triaging musculoskeletal conditions and were not publicly available outside the UK National Health Service. Most tools were generic tools designed to screen for general health problems, including musculoskeletal conditions. The most common approach to evaluating performance (eg, accuracy) of the tools was to compare the concordance of the tool to a clinician diagnosis or triage recommendation. Sensitivity and specificity ranged from 39% to 91% and 23% to 80%, respectively. The reported accuracy of the included tools ranged from 33% to 98%.

Conclusions: Musculoskeletal conditions remain a blind spot for people designing, implementing, and evaluating digital health for triage: few tools were specifically designed for musculoskeletal conditions, and most existing tools performed poorly when applied to musculoskeletal populations. We recommend health systems and clinicians use a multimodal approach, integrating both digital health tools and clinical decision-making to safely triage and diagnose until a more robust tool for musculoskeletal conditions is available. Future tool developers need to use transparent, standardized processes that prioritize tool safety, clinical value, and trustworthiness when designing for clinicians and patients.

(J Med Internet Res 2026;28:e81578) doi:10.2196/81578

KEYWORDS

musculoskeletal; digital health; emergency department; health system; triage; PRISMA; Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Introduction

Musculoskeletal conditions are one of the largest contributors to the global burden of disease and the sixth largest contributor to disability worldwide [1]. The global forecast predicts that the burden of musculoskeletal conditions will more than double in the decades between 2020 and 2050 [1]. Most musculoskeletal conditions can be effectively managed proactively in a primary care setting, not in the emergency department (ED) [2]. Our recent analysis of epidemiological data revealed that approximately 1 in 10 ED visits were related to musculoskeletal issues, and 6 in 10 of these cases could have been appropriately managed outside the ED [3]. This indicates a need to re-evaluate how people with musculoskeletal conditions access health care.

Effective and efficient triage processes are needed to help patients navigate health systems and find timely and high-quality care, avoiding inappropriate use of the ED [4]. The idea of triage was first applied in military settings to help allocate resources and timely care for the wounded [4]. In today's context, triage is often considered in the ED or the first point of contact for clinicians to help prioritize who needs attention first when patients present to the ED [4]. In regard to triaging musculoskeletal conditions, triage is often conducted by tele-triage, paper-based triage, and face-to-face triage [5]. However, patients and musculoskeletal experts have reported that these approaches are inefficient and ineffective in moving patients through the health system [5].

There is an increasing trend toward the use of digital triage, such as online symptom checkers by patients, to make an informed decision on the next and best course of action for their current problem [6-8]. More recently, the World Health Organization has launched a global strategy on digital health to help improve the health and well-being of all humans [9]. This includes defining digital health as "the use of information and communications technology in support of health and health-related fields," which encompasses eHealth, mobile health (mHealth), advanced computer sciences, such as big data and artificial intelligence (AI) or machine learning, and the broad scope of telehealth and telemedicine [10]. Digital health tools have the potential to tackle overcrowding in the ED and primary care settings by guiding patients to alternative services for musculoskeletal care that may be just as effective as the ED.

In the last decade, there has been a shift toward integrating digital health tools, such as symptom checkers, into the health system, as reflected in the volume of reviews evaluating such tools [6,8,11-14]. With the proliferation and widespread adoption of generative AI (eg, large language models [LLMs]), the public seems accepting of using digital health tools like AI to provide guidance and diagnoses for health conditions. This is despite generative AI not being specifically designed for health care use [15]. These findings reflect the growing demand for the integration of digital technology into health care.

Despite the advancement of digital health tools, there are no reviews currently available that have studied digital health tools for diagnosing and triaging musculoskeletal conditions [8,11-14]. Understanding the available tools, including their performance (eg, accuracy), will help researchers, policymakers, and clinicians tailor future digital health technologies for musculoskeletal conditions and make informed decisions about how to implement technology in health systems. Helping patients find the "right care at the right time" for musculoskeletal conditions may help reduce burden on the health system and allow the ED to do what it was created for: provide life-saving care.

The primary objective of this review was to identify and describe available digital health tools that can triage and diagnose musculoskeletal conditions in primary, urgent, and emergency settings. The secondary objective was to summarize the performance and accuracy of digital health tools.

XSL•FO
RenderX

Methods

Overview

This scoping review was conducted in accordance with the Johanna Briggs Institute methodology for scoping reviews [16,17] and reported following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews; Checklist 1) and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Search) [18,19]. We were guided by Arksey and O'Malley's [20] framework with the additions of Levac et al [21]. A protocol was prospectively registered on the Open Science Framework (OSF). Amendments to the protocol were updated and uploaded to the OSF [22].

Search Strategy

An electronic search was conducted in 6 databases (MEDLINE [OVID], Embase [OVID], CENTRAL [OVID], CINAHL [EBSCO], Compendex, and Web of Science) and 4 gray literature sites (OpenGrey, GoogleScholar, arXiv, and medRxiv) with the aid of a biomedical librarian and information specialist. Our search strategy was not peer-reviewed but was tested through an iterative process by the biomedical librarian to ensure that search strategies returned identified seed papers. Initial search strategies were adapted from previously published work [11,14]. We searched databases and gray literature from inception to September 18, 2025. Table 1 provides the population-concept-context framework for our search strategy and illustrates how we operationalized our search. The full MEDLINE search strategy is outlined in Multimedia Appendix 1, and all other search strategies are uploaded and available on OSF [22]. To supplement the search, we screened the reference lists of relevant reviews and included records. We also searched the Cochrane Database of Systematic Reviews, PROSPERO, OSF, and *JBI Evidence Synthesis* to identify any active systematic or scoping reviews on the topic. The search approach for identifying gray literature is detailed in Multimedia Appendix 2.

Table . Population-concept-context (P-C-C) framework.

P-C-C	Definition	Keywords
Population (people with musculoskeletal pain)	Acute (traumatic) or chronic injury related to muscles, bones, joints, tendons, or ligaments problems that cause regional or generalized pain[23] and musculoskeletal disease or conditions as defined by the Global Burden of Disease (rheumatoid arthritis, osteoarthritis, low back pain, neck pain, and gout) [1].	(Musculoskeletal or MSK) injur* or pain* or tear* or ligament* or sprain* or strain* or gout or arthritis or rheumatic arthritis
Concept (digital health)	"The use of information and communications technology in support of health and health-related fields" [10]. Digital health captures eHealth, mobile health, advanced computer sciences, such as big data and AI[a], and the broad scope of telehealth and telemedicine [10].	telemedic* or telehealth or teletriag* or teleconsult* or telecare or tele-care or virtual medicine or virtual care or virtual triage or digital health or digital tool or digital care or digital health technology or AI or artificial intelligence or deep learning or machine learning
Context (triage)	Triage guides the distribution of medical resources to patients when there is a scarcity of health care resources and often refers to a process to allocate, ration, or prioritize patient treatment and is considered first point-of-contact care [4]. Triage can be done by a clinician, patient, or technology (eg, AI) and may involve patients' self-assessment.	self-refer* or self-assess* or self-access* or tele-triage or triage or diagnosis or decision making or symptom checker*

[a]AI: artificial intelligence.

Inclusion Criteria

Studies of adults (aged 18 years and older) with musculoskeletal conditions (≥25% of the sample had to be musculoskeletal-related) that identified and reported a digital health tool designed specifically to triage or diagnose in primary, urgent, or emergency care settings were included. Textbox 1 describes the inclusion and exclusion criteria. We excluded studies that evaluated the effectiveness of virtual assessments. We also excluded studies that used digital health tools for secondary diagnoses (ie, patient had already seen a practitioner and given a diagnosis), as the tools were typically used to manage symptoms and not for primary triage or diagnosis.

XSL•FO
RenderX

Textbox 1. Overview of study selection criteria.

Inclusion criteria

- Adult participants (≥18 years) with a primary complaint of a musculoskeletal condition
- Sample has ≥25% musculoskeletal conditions
- Identifies and reports a digital health tool designed specifically for triage or diagnosis in primary care, urgent care, or emergency settings

Exclusion criteria

- Not English language
- Nonhuman data (eg, vignettes or simulated clinical cases)
- Study design (not original data, eg, review, opinion paper, commentaries, and guidelines)
- Not adult population (all participants aged at least 18 years)
- Not related to a digital health tool (instrument testing or replication or validation studies of clinician assessment to virtual assessment were excluded, no wearable or technology testing was excluded)

Study Selection

Records were collated and uploaded into EndNote (version 20.3; Clarivate Analytics), and duplicates were removed before uploading to Covidence (Veritas Health Innovation) for screening. Pairs of independent reviewers screened all records by title and abstracts, and a third reviewer (CLA) resolved any discrepancies if consensus could not be reached.

At the full-text stage, we first conducted pilot screening, where all reviewers assessed the same 5 full texts. If major discrepancies were identified, we met to review and discuss how to apply the screening criteria in a standardized manner. All full-text papers were reviewed by pairs of independent reviewers. Reasons for exclusion during full-text screening were recorded. Any disagreements between the reviewers at each stage of the selection process were resolved through consensus or by an additional reviewer (CLA) as required.

Data Extraction

Data were extracted from included records independently by pairs of reviewers using a custom data extraction tool designed in Microsoft Excel by the research team. Any disagreements were resolved via consensus. We extracted the following details where available: study characteristics (author, country, sample size, and study aim), participants' demographics (sex, age, and musculoskeletal pain or diagnosis), type of digital tool (name, purpose of tool, and target users), design and development process, platform of tool, tool delivery (eg, clinician or patient self-access), context or care setting in which the tool was used, assessment of performance or accuracy results, and key findings relevant to the review question. Where relevant, authors were contacted once via email to request missing data and to clarify details about the digital health tool.

Data Synthesis

Studies were summarized by study characteristics, digital health tool details, and performance assessment. Descriptive data were summarized as proportions when appropriate. Digital health tools were classified according to the World Health Organization digital health category [10] (eHealth, mHealth, and AI or machine learning), function (ie, triage, diagnosis, or both), care setting in which the tool was used (ie, ED, primary or urgent care, or mixed), research setting (ie, urban, rural, or both), how the tool was administered (eg, self-access or clinician-delivered), technology interface (eg, web-based or app), and intended user (patient-facing, clinician-facing, or both).

Digital tools were rated using the technology readiness level (TRL) and associated technology stage by the first author (LKT) and verified by a second rater [24]. The TRL ranges from 1 to 9, with 1 representing tool conception and 9 representing that the tool is ready to be used in real-world settings [24]. We classified TRL across technology stages (fundamental research, research and development, pilot and demonstration, early adoption, and commercially available) [24].

When available, the performance of the digital health tool was reported by identifying appropriate triage referrals or recommendations or diagnoses compared to a reference standard (eg, physician diagnosis). Measures of performance included diagnostic test accuracy (area under the receiver operator characteristic curve, sensitivity, specificity, positive predictive value, negative predictive value, and likelihood ratio) or reliability measures (internal consistency, test-retest reliability, and intra- or interrater reliability).

Deviations From Protocol

We sought to use the continuous active learning on Covidence to support title and abstract screening [25]. During attempts to calibrate the algorithm, the algorithm did not perform well in identifying relevant papers for this review. We were not confident in screening titles and abstracts with only 1 human reviewer (LKT). Instead, all titles and abstracts and full text records were screened in duplicate by 2 independent human reviewers.

As scoping reviews are an iterative process and aim to assess and evaluate the available evidence, a broad research question often results in a highly sensitive search and less specific records. We made pragmatic decisions and minor amendments to the selection criteria as the review progressed. At the full-text screening stage, studies including a general population had to report ≥25% of the sample being musculoskeletal-related to be included. This cutoff was determined based on studies that

XSL•FO
RenderX

indicated the prevalence of musculoskeletal presentations in ED was approximately 25% [26,27].

Results

Overview

The titles and abstracts of 5695 unique records were screened, and 189 papers were reviewed in full (Figure 1). In total, 34 studies met the inclusion criteria (n=37,509 participants across 33 studies, 12,470/37,509, 33% female). The median age was 50 (range 18-91) years. One study did not report the sample size. Sex and age data distribution were missing in 12 and 13 studies, respectively. A list of the studies that were excluded at the full-text stage is presented in Multimedia Appendix 3. Figure 2 charts the data of all 34 studies included by condition, publication year, and sample size of the study, if reported.

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) flowchart.

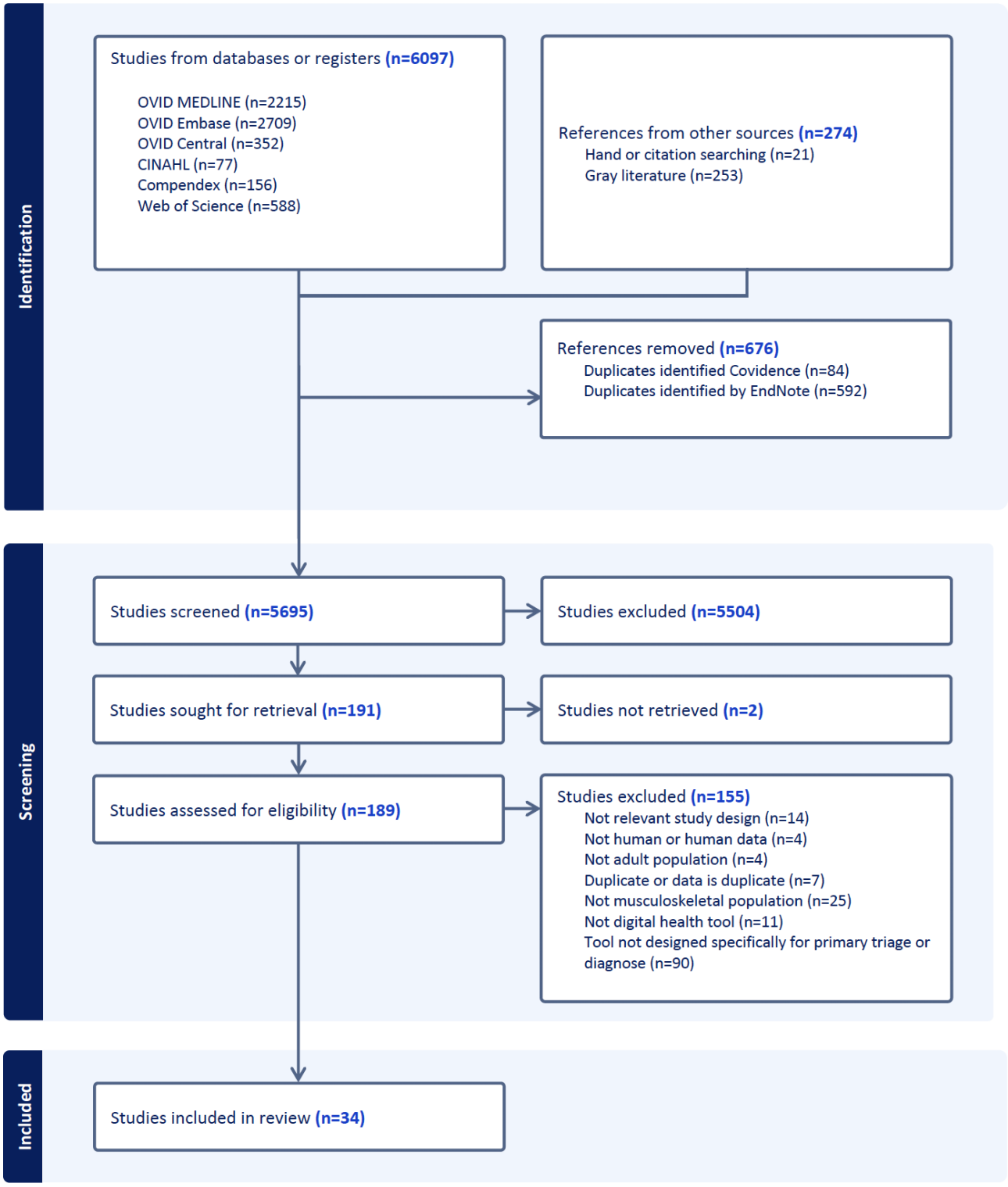

XSL•FO
RenderX

Figure 2. Sources of data charted by publication year and condition.

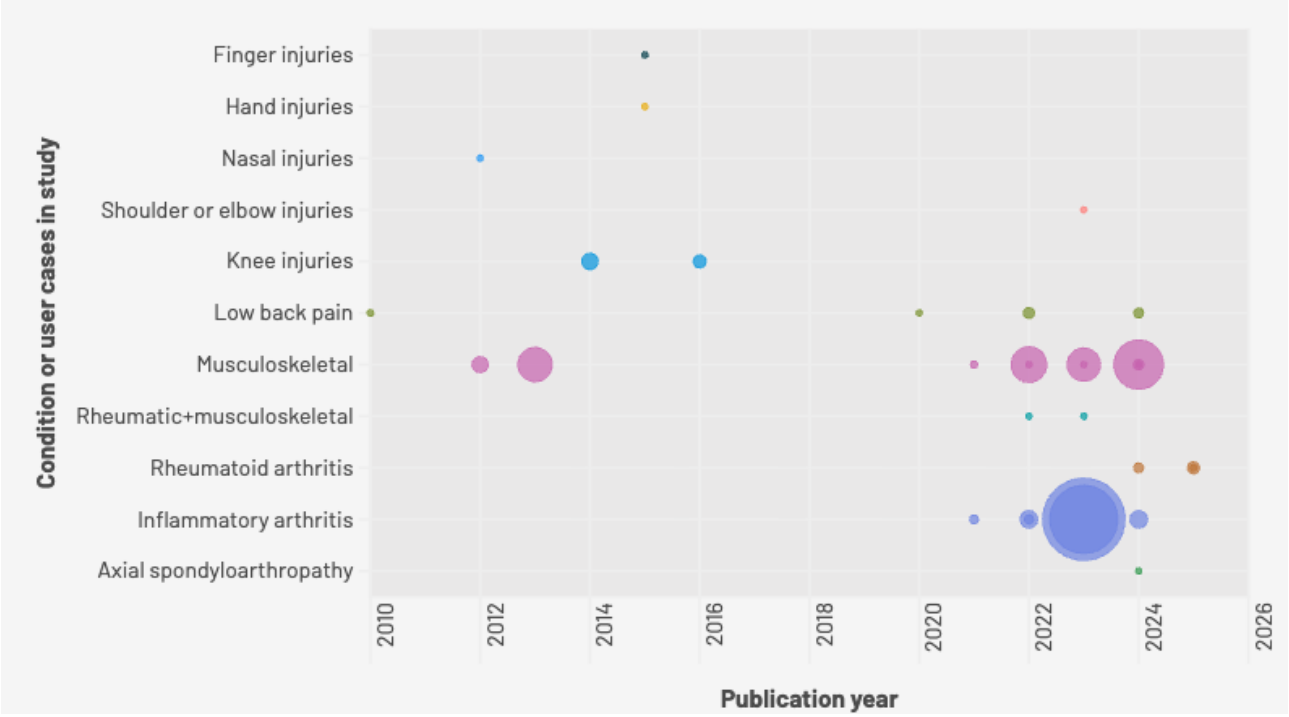

Study Characteristics

In total, 30 of 34 (88%) studies focused on primarily musculoskeletal conditions, and 4 of 34 (12%) studies focused on general populations with a subset of the sample being musculoskeletal conditions. Full study details can be found in Multimedia Appendix 4 [28-61].

In total, 25 of 34 (74%) studies were peer-reviewed, 7 of 34 (21%) records were conference abstracts, and 2 of 34 (6%) were industry case reports. Studies were published between 2010 and 2025. Cross-sectional (10/34, 29%) studies were most common, followed by nonrandomized or quasi-experimental studies (8/34, 24%), randomized controlled trials (5/34, 15%), retrospective cohort (4/34, 12%), mixed or multimethods (3/34), prospective observational cohort (2/34), and case report (2/34, 6%). All studies were conducted in high-income countries (n=1 country not reported), with the majority being from Europe (21/34, 62%) or the United States (7/34, 21%).

Digital health tools were evaluated across various care settings, including ED or urgent care (6/34, 18%), physician-led primary care (6/34, 18%), physiotherapist-led primary care (1/34, 3%), patient self-access (19/34, 56%), and mixed (eg, primary care and ED) settings (2/34, 6%).

Tool Identification and Characteristics

Overview

Inflammatory arthritis (eg, rheumatoid arthritis, gout, and spondyloarthropathy) and arthritis-related conditions were the most common (13/34, 38%) musculoskeletal conditions studied, followed by generic musculoskeletal conditions (11/34, 32%). Others tested digital tools for low back pain (4/34, 12%), knee (2/34, 6%), finger or hand (2/34, 6%), shoulder (1/34, 3%) conditions, and nasal fractures (1/34, 3%).

We identified 16 unique digital health tools (Table 2). In total, 7 studies did not report the name of the digital health tool that was studied or studied a bespoke tool designed for the study where we could not extract the name of the tool. Of the 34 tools, 13 (38%) tools reported that they were designed to "diagnose and triage," 12 (35%) tools were designed to diagnose, and 9 (26%) tools were designed to triage (Table 2). Only 2 tools (Phio [32,39] and Digital Assessment Routing Tool [DART] [51,52]) were designed specifically to triage musculoskeletal conditions. Overall, 3 tools (ADA [37,41,46-49], Buoy Health [33], and WebMD Symptom Checker [40]) were generic health tools that had integrated algorithms to screen musculoskeletal conditions (among other conditions). One tool used OpenAI or LLMs (ChatGPT [29,34,49]) and reported algorithms capable of diagnosing musculoskeletal conditions. Five tools (Rheumatic? [43,45,53,56], Rheport [46-48], RheumConnect [60], ReumAI [38], and Bechterew-check [41]) were condition-specific (ie, designed for rheumatological or inflammatory conditions) with capabilities to differentiate these conditions from other types of musculoskeletal conditions. Two tools were joint-specific (Therapha for low back pain [28] and Virtual Knee Doc for acute knee injuries [30,31]).

XSL•FO
RenderX

Table . Characteristics of digital health tools, summarizing purpose, use, technology development, and availability.

Digital health tool name	Purpose of tool[a]	Type of tool[b], intended user, and access level	Tool format	Digital health category	Technology readiness level[c]	Technology readiness assessment[d]	Tool processes[e]	Available to public
ADA [37,41,46-49]	Diagnosis and triage	Symptom checker (patient; self-access)	App	mHealth[f], AI[g] or ML[h]	9	Commercially available	AI	Yes
Bechterewcheck [41]	Diagnosis	Symptom checker (patient; self-access)	Web-based	mHealth	7	Pilot and demonstration	Clinical or decision support pathway	Yes (only in German)
Buoy Health [33]	Diagnosis and triage	Symptom checker (patient; self-access)	Web-based or app	eHealth, mHealth, AI or ML	9	Commercially available	AI	Yes
ChatGPT [29,34,49]	Diagnosis	Diagnostic predictor (patient; self-access)	Web-based	eHealth, AI or ML	9	Commercially available	AI	Yes
Digital Assessment Routing Tool [51,52]	Triage	Digital triage[i] (patient; self-access)	App	mHealth	8	Research and development	Clinical or decision support pathway	No
Phio [32,39]	Diagnosis and triage	Symptom checker[i] (patient; self-access)	App	mHealth, AI or ML	9	Commercially available	AI	No, Proprietary
Phone camera [42] (any built-in camera)	Triage	Tele-triage (clinician; clinician-administered)	Phone	mHealth	6	Pilot and demonstration	Clinical or decision support pathway	Yes
PhysioDirect [44]	Diagnosis and triage	Tele-triage (clinician; clinician-administered)	Phone	eHealth	6	Pilot and demonstration	Clinical or decision support pathway	No
Rheumatic? [43,45,53,56]	Diagnosis and triage	Symptom checker (patient; self-access)	Web-based	eHealth	8	Pilot and demonstration	Clinical or decision support pathway	Yes
Rheport[[37,46-48]	Diagnosis	Symptom checker (patient; self-access)	Web-based	eHealth	9	Early adoption	Clinical or decision support pathway	Yes (only in German)
Therapha [28]	Diagnosis and triage	Digital triage (clinician; clinician-administered)	Web-based	eHealth, AI or ML	9	Commercially available	Clinical or decision support pathway	No (propriety)
TriageXpert Dual Purpose [50]	Triage	Tele-triage (clinician; clinician-administered)	Phone	eHealth	7	Commercially available	Clinical or decision support pathway	No
RheumConnect [60]	Triage	Symptom checker (patient; self-access)	Web-based chatbot	eHealth, AI or ML	6	Pilot and demonstration	AI	No
ReumAI [38]	Triage	Tele-triage (clinician; clinician-administered)	Phone	eHealth, AI or ML	6	Pilot and demonstration	AI	No

XSL•FO
RenderX

Digital health tool name	Purpose of tool[a]	Type of tool[b], intended user, and access level	Tool format	Digital health category	Technology readiness level[c]	Technology readiness assessment[d]	Tool processes[e]	Available to public
Virtual Knee Doc [30,31]	Diagnosis	Symptom checker (patient; self-access)	Web-based	eHealth	6	Pilot and demonstration	Clinical or decision support pathway	No
WebMD Symptom Checker [40]	Diagnosis	Symptom checker (patient; self-access)	Web-based	eHealth, AI or ML	9	Commercially available	AI	Yes

[a]Triage: provide next steps for care based on symptoms and urgency, may provide preliminary diagnoses but not the objective of the tool. Diagnosis: provide a preliminary diagnosis based on symptoms, which aids to direct next steps in care.

[b]Type of tool: symptom checker: tool designed for patients to enter their symptom data; tele-triage: tool designed to triage using telephone interface; digital triage: tool designed to triage using eHealth or mHealth interface; diagnostic predictor: tool designed to use data to predict diagnosis or triage pathway.

[c]Based on Innovation Canada Technology Readiness Level: rated on a scale of 1-9, where 1=tool conception and 9=tool ready for real-world settings.

[d]Based on Innovation Canada Technology Readiness Stages: fundamental research, research and development, pilot and demonstration, early adoption, commercially available.

[e]Tool processes: AI: tool that uses big data to assign probability to allow for computer-driven decision-making; clinical decision support pathway: predefined decision tree or rule-based algorithms that support clinical decision-making.

[f]mHealth: mobile health.

[g]AI: artificial intelligence.

[h]ML: machine learning.

[i]Self-access within the UK National Health Service.

Intended Users

Most studies (24/34, 71%) reported on digital health tools designed for use by patients, 10 (10/34, 29%) studies targeted tools at clinicians. In total, 19 (19/34, 56%) studies reported on tools that were symptom checkers and were designed to be patient-facing. A total of 10 (29%) studies reported on clinician-facing tools for triage, diagnosis, or diagnostic prediction.

Patient-Facing

We classified ADA, Buoy Health, ChatGPT, Bechterew-check, Rheport, Rheumatic?, RheumConnect, Virtual Knee Doc, DART, and Phio as tools for patients. All used an app or a web-based interface. ADA, Buoy Health, and ChatGPT were tools used for generic health purposes, while the others were designed for specific groups of conditions (ie, rheumatological or musculoskeletal conditions). DART and Phio were designed to integrate with the UK National Health Service. Patients who used DART or Phio had their results forwarded to a primary care team or physiotherapist.

Clinician-Facing

In total, 7 tools were identified for clinicians. Therepha is a clinical decision support system designed for physiotherapists to diagnose and triage low back pain and was piloted in the ED [28]. ReumAI uses tele-triage where a nonphysician staff uses AI-guided telephone interviews to identify diagnoses and potential previsit tests [38]. Triage Xpert Dual Purpose [50] and PhysioDirect [58] were triage tools designed for implementation within specific health systems. One study examined triage of nasal fractures using a built-in camera to triage to the right hospital setting [42], and another used tele-triage to assess whether the ED could be avoided altogether for finger injuries [36]. Most clinician-facing triage tools used tele-triage (ie, phone call) as their interface, except for Therepha, which was a web-based tool. In total, 2 studies leveraged large datasets and AI to predict diagnoses, with 1 study using ChatGPT [29,60].

Performance and Usability

Of the 34 studies identified, 19 (56%) evaluated the performance of the digital health tool (Table 3). The most common definition and method to determine performance was measuring concordance to a clinician diagnosis or recommended triage pathway, often by evaluating sensitivity and specificity. The performance of these tools varied widely and was partly dependent on the context in which they were being used (eg, ED or primary care). Sensitivity ranged from 39% to 91%, and specificity from 23% to 80% [28,30,31,37,41,45,46,48]. The methods for measuring accuracy were poorly reported, often in the form of proportion of correct triage or diagnoses. Reported accuracy ranged from 33% to 98% across 12 unique tools (n=16 studies) [28,29,34,36-38,41,45,49,51,59,61]. The accuracy of tools used by patients in tertiary settings (eg, seeking care from a specialist such as an orthopedic surgeon or rheumatologist) was reported as higher than the accuracy of tools used in primary care settings.

For studies that compared digital health tools against each other, ADA was the most common tool used for comparison. When compared to rheumatologists and medical students, ADA was superior to clinician's diagnosis of rheumatic and nonrheumatic conditions, ChatGPT, and Bechterew-check [37,41,49]. ADA was comparable to Rheport for diagnostic accuracy of rheumatic conditions [48]. ChatGPT performed similar to experienced

XSL•FO
RenderX

rheumatologists for potential diagnostic accuracy for rheumatic conditions [49].

In total, 4 tools were available in multiple languages (ADA, ChatGPT, Rheumatic?, and WebMD Symptom Checker), and 8 tools were accessible to the public; however, 2 were designed for German speakers (Table 3). Based on the TRL, we classified 6 tools as being at the commercially available stage (ADA, Buoy Health, ChatGPT, Phio, Therapha, and WebMD Symptom Checker).

Figure 3 provides a visualization comparing TRL and performance evaluation for the identified digital health tools (Only 15 of the 16 identified digital tools are reported in this figure. The “phone built in camera” was not graphed.). If the tool did not complete a performance evaluation of the tool, a 0 was given for reported performance (ie, accuracy) on Figure 3. Despite some tools being commercially available, there was a discrepancy in reported performance findings for musculoskeletal conditions.

XSL•FO
RenderX

Table . Performance statistics of digital health tools.

Digital health tool and authors	Performance of tool evaluated (yes or no)	Definition used to define tool performance	Condition evaluated	Methods to evaluate performance	Sensitivity (%)	Specificity (%)	Accuracy of tool[a]	Other findings reported (%) (95% CI)
ADA[b]								
Knitza et al (2021) [46]	Yes	Concordance with physician diagnosis	Rheumatic	Sensitivity or specificity, PPV[c], NPV[d]	43	64	NR[e]	PPV 37 (26-48), NPV 69 (60-80)
Knitza et al (2024) [48]	Yes	Concordance with physician diagnosis	Rheumatic	Sensitivity or specificity, PPV, NPV	52	68	NR	PPV or NPV varied depending on whether ADA or Rheport was used first
Graf et al (2022) [37]	Yes	Concordance with identified diagnosis from clinical trial	Rheumatic	Sensitivity or specificity, accuracy	71	64	54% accurately diagnosed same condition	NR
Hannah et al (2024) [41]	Yes	Concordance with discharge summary report	Rheumatic	Sensitivity or specificity, accuracy	39	78	58% accurately diagnosed same condition	NR
Krusche et al (2024) [49]	Yes	Concordance with physician diagnosis	Rheumatic	Proportion	NR	NR	65% accurate for all cases; 71% accurate for cases with IRDs[f]; 61% accurate for non-IRD cases	NR
Bechterew-check								
Hannah et al (2024) [41]	Yes	Concordance with discharge summary report	Axial spondyloarthropathy	Sensitivity or specificity, accuracy	41	53	47% accurately diagnosed same condition	NR
Bespoke tool (no name reported)								
Demmelmaier et al (2010) [35]	No	NT[g]	Low back pain	NT	NT	NT	NT	NR
Martin and Payne (2020) [54]	No	NT	Low back pain	NT	NT	NT	NT	NR
Phillips et al (2012) [55]	No	NT	MSK[h]	NT	NT	NT	NT	NR
Ryan and Grinbergs (2024) [57]	No	NT	MSK	NT	NT	NT	NT	NR
Trivedi et al (2024) [61]	Yes	Concordance with nurse triage	MSK	Proportion	NR	NR	63% accurately triage	NR
Soin et al (2022) [59]	Yes	Concordance with physician diagnosis	Low back pain	NR	NR	NR	72% software predicted correct diagnosis	NR

XSL•FO
RenderX

Digital health tool and authors		Performance of tool evaluated (yes or no)	Definition used to define tool performance	Condition evaluated	Methods to evaluate performance	Sensitivity (%)	Specificity (%)	Accuracy of tool[a]	Other findings reported (%) (95% CI)
Buoy Health									
	Carmona et al (2022) [33]	No	NT	Generic MSK	NT	NT	NT	NT	NR
ChatGPT									
	Badsha et al (2024) [29]	Yes	Concordance with physician diagnosis	Rheumatic	NR	NR	NR	98% accurate with rheumatologist diagnosis	NR
	Daher et al (2023) [34]	Yes	Concordance with physician diagnosis	Shoulder or elbow injuries	NR	NR	NR	93% accurate with surgeon diagnosis; 83% accurate with surgeon management	NR
	Krusche et al (2024) [49]	Yes	Concordance with physician diagnosis	Rheumatic	Proportion	NR	NR	35% accurate for all cases; 71% accurate for cases with IRDs; 15% accurate for non-IRD cases	NR
Digital Assessment Routing Tool (DART)									
	Lowe et al (2022) [51]	Yes	Concordance with physiotherapist expert	MSK	Proportion	NR	NR	84% DART matched physiotherapist	NR
	Lowe et al (2024) [52]	Yes	Concordance with physiotherapist expert	MSK	Intraclass correlation coefficient (ICC)	NR	NR	NR	ICC 0.37 (0.16 - 0.55)
Phio									
	Bond et al (2024) [32]	No	NT	NT	NT	NT	NT	NT	NR
	Gymer et al (2023) [39]	No	NT	NT	NT	NT	NT	NT	NR
Phone Camera									
	Hara et al (2015) [42]	Yes	Accuracy of triage recommendations	Finger injuries	NR	NR	NR	NR	NR
PhysioDirect									
	Kelly et al (2021) [44]	No	NT	NT	NT	NT	NT	NT	NR
Rheport									
	Knitza et al (2021) [46]	Yes	Concordance with physician diagnosis	Rheumatic	Sensitivity or specificity, PPV, NPV	54	52	NR	PPV 35 (25-47); NPV 70 (58-79)

XSL•FO
RenderX

Digital health tool and authors		Performance of tool evaluated (yes or no)	Definition used to define tool performance	Condition evaluated	Methods to evaluate performance	Sensitivity (%)	Specificity (%)	Accuracy of tool[a]	Other findings reported (%) (95% CI)
	Knitza et al (2024) [48]	Yes	Concordance with physician diagnosis	Rheumatic	Sensitivity or specificity, PPV, NPV	62	47	NR	PPV or NPV varied depending on whether ADA or Rheport was used first
Rheumatic?[i]									
	Knevel et al (2022) [45]	Yes	Concordance with physician diagnosis/treatment recommendation	Rheumatic	Sensitivity or specificity, AUC-ROC[j]	67	72	AUC-ROC 75 (95% CI 62 - 89)	NR
	Qin et al (2024) [56]	No	NT	Rheumatic	NT	NT	NT	NT	NR
	Lundberg et al (2023) [53]	No	NT	Rheumatic	NT	NT	NT	NT	NR
	Jakobi et al (2025) [43]	No	NT	Rheumatic	NT	NT	NT	NT	NR
ReumAI									
	Gómez-Centeno et al (2025) [38]	Yes	Concordance with physician diagnosis	Rheumatic	NR	NR	NR	53% accurate with rheumatologists	NR
RheumConnect									
	Tan et al (2023) [60]	No	NT	Rheumatic	NT	NT	NT	NT	NT
Therapha									
	Badahman et al (2024) [28]	Yes	Concordance with MRI findings	Low back pain	Sensitivity or specificity, PPV, NPV, ROC[k]	88	80	ROC 0.84 (95% CI 0.6 - 1.0; *P*=.001)	PPV 99 (25-47); NPV 27 (58-79)
Triage Xpert Dual Purpose									
	Li et al (2023) [50]	No	NT	MSK	NT	NT	NT	NT	NT
Virtual Knee Doc									
	Bisson et al (2014) [30]	Yes	Concordance with physician diagnosis	Knee injuries	Sensitivity or specificity	89	27	NR	NT
	Bisson et al (2016) [31]	Yes	Concordance with physician diagnosis	Knee injuries	Sensitivity or specificity	91	23	NR	NT
WebMD Symptom Checker									

XSL•FO
RenderX

Digital health tool and authors	Performance of tool evaluated (yes or no)	Definition used to define tool performance	Condition evaluated	Methods to evaluate performance	Sensitivity (%)	Specificity (%)	Accuracy of tool[a]	Other findings reported (%) (95% CI)
Hageman et al (2015) [40]	Yes	Concordance with physician diagnosis	Hand injuries	Proportion	NR	NR	33% accurate with hand surgeon diagnosis	NT

[a]Reported values from the study and not interpretation of authors.
[b]Findings reported from ADA diagnosis 1 (D1) in study.
[c]PPV: positive predictive value.
[d]NPV: negative predictive value.
[e]NR: not reported.
[f]IRD: inflammatory rheumatic disease.
[g]NT: not tested.
[h]MSK: musculoskeletal.
[i]Findings reported from dataset A in study.
[j]AUC-ROC: area under the receiver operating curve.
[k]ROC: receiver operating curve.

Figure 3. Visualization comparing TRL and the highest reported performance evaluation across identified digital health tools. NR: not reported; TRL: technology readiness level.

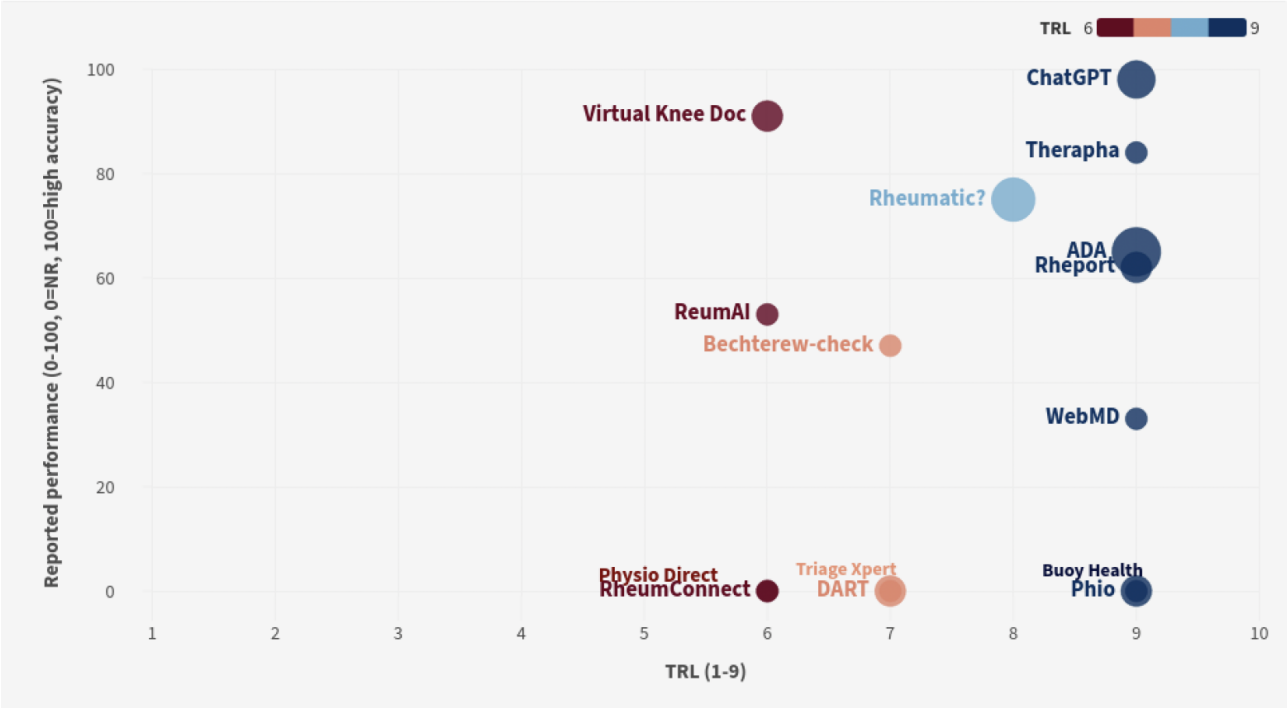

Discussion

Principal Findings

We aimed to identify and describe the available tools for triaging and diagnosing musculoskeletal conditions in primary, urgent, and emergency settings. Based on a synthesis of 34 studies and data from 16 different digital health tools, there were no digital health tools with sufficient evidence to support effective triage and diagnosis of musculoskeletal conditions across these settings. Approximately half of these tools were available to the public. Not all tools were available in English, with 2 tools only available in German (Bechterew-check and Rheport). The most frequently studied digital health tool was ADA (n=5), followed by Rheumatic? (n=4), then ChatGPT (n=3). Only 2 tools (DART and Phio) were purposely developed for screening musculoskeletal conditions. Both tools are not currently available outside of the UK's National Health Service. We were surprised to find so few digital health tools targeting musculoskeletal conditions, given the substantial global burden of musculoskeletal conditions [1]. Notably, rheumatological or inflammatory arthritis was the most prevalent musculoskeletal condition studied, despite low back pain being the most common musculoskeletal condition seen in ED and primary care settings [62]. We identified 4 studies that included digital health tools

XSL•FO
RenderX

targeting low back pain, but only 1 of these reported which tool was used (Therapha). Our findings reflect the discordance of research across digital health technology and the current health landscape. Many tools were inaccessible or not designed for practical use in managing musculoskeletal pain, the most burdensome conditions seen in primary care.

Our secondary objective was to summarize the performance and accuracy of the included digital health tools. Approximately 50% of the studies evaluated the performance of a digital health tool. Apart from ChatGPT, most generic digital health tools (eg, ADA and WebMD Symptom Checker) reported poor accuracy (often less than 50% accuracy in identifying the correct diagnosis compared to clinicians) for musculoskeletal conditions [37,40,48]. Despite the use of ChatGPT by the public as a symptom checker [15], ChatGPT's accuracy for diagnosing musculoskeletal and rheumatic conditions was variable, ranging from 33% to 98% [29,34,49]. We suggest that further research is needed before considering ChatGPT as an accurate diagnostic or screening tool. Tools that were designed to diagnose peripheral or spinal musculoskeletal conditions (eg, low back pain or knee injuries) appear to be more promising with high sensitivity (88% - 91%) [28,31]. Finally, tools designed specifically to triage (rather than diagnose) musculoskeletal conditions (ie, Phio and DART) demonstrated the best performance. Recent findings published on DART and Phio indicate that these tools have high agreement (>90%) with expert physiotherapist recommendations on next care pathways [63,64]. However, the heterogeneity across evaluation methods highlights the importance of standardized development and evaluation frameworks to ensure that digital triage tools for musculoskeletal conditions are accurate, transparent, and safe before integrated into clinical settings and workflows.

Not Yet Ready for Prime Time

One of the key findings of our review is that some tools are commercially available and integrated into health systems for musculoskeletal screening without robust methodological evaluation or reporting. Premature implementation raises concerns, particularly given the risk of misdirecting patients or delaying appropriate care. Before being adopted at scale, digital triage tools must demonstrate value in real-world settings and meet minimum standards for safety, accuracy, and usability. However, many studies evaluating these tools lack transparent reporting, making it difficult to assess how performance claims were derived. Studies reporting high accuracy of their digital tools often had poor transparency or a lack of details on their tool evaluation. We suggest caution with interpreting these digital health tools as ready for public use without further evaluation. It is also unclear how tools that use LLMs operate. These networks are often termed "black boxes" due to the inability to explain how these systems achieve their output [65].

Our findings have been confirmed by a recent study that evaluated diagnostic accuracy and clinical reasoning using 6 different generative AI (LLMs) for rheumatic diagnoses [66]. Despite the LLMs reporting high diagnostic accuracy (~80%), all models reported subpar clinical reasoning quality (eg, explaining reasons for supporting diagnoses) [66]. These findings underscore the importance of digital health tools requiring both high diagnostic accuracy alongside transparent algorithms to help to explain the logic behind the tool's decision. To improve transparency and enable reproducibility, it is important to establish standards for incorporating ethical AI in digital health. Without transparency in how tools were developed, or in the algorithms used, it is unclear whether the tools are safe for the public to use.

The only digital health tool with robust evaluation of its performance was the generic health app, ADA, which is a Conformité Européenne–certified medical product [46,48]. ADA's performance was inconsistent across the studies, and ADA correctly identified the musculoskeletal condition or triage option in fewer than half of the cases [37,41,46,48,49]. Condition-specific digital health tools (Rheport [46,48], Rheumatic? [45], Virtual Knee Doc [30,31], Therapha [28], and ReumAI [38]) performed slightly better. The reported accuracy was higher in these tools, especially if these tools were implemented in tertiary care settings (outside of the ED or primary care). We are not aware of an acceptable threshold for performance (ie, accuracy) for digital health tools. However, we recommend implementing tools that are at least more accurate than flipping a coin and provide consistent results across different study contexts or musculoskeletal conditions.

AI-driven tools, like ADA or ChatGPT, may perform better than clinician decision support systems or physicians or rheumatologists in diagnosing rheumatic conditions [49,67]. Integrating digital health tools in tandem with other nonspecialist professions (eg, general practitioners and allied health professionals) could help guide patients to their next care steps as they wait for specialists (eg, rheumatologists) or avoid unnecessary visits to specialists or other care providers. AI-driven tools that have included diagnostic findings (eg, imaging, clinical symptoms or signs, and bloodwork) have superior diagnostic accuracy to other AI models [41,67,68]. Until robust stand-alone digital health tools are developed (ie, a symptom checker that can be used independently by patients), combining digital health tools and clinician feedback may be the best method to streamline diagnosis and care in complex cases while providing timely care for common musculoskeletal conditions.

Several frameworks for evaluating digital health tools have been proposed [69]. A recent scoping review identified 12 key domains—ranging from tool description and content to safety, clinical effectiveness, and efficacy—across 95 frameworks that developers and researchers can draw on [69]. However, the heterogeneity reflects a broader challenge: many digital health tools span multiple categories (eg, eHealth or mHealth tools incorporating AI), making classification inconsistent and evaluation difficult. Advancing this field requires standardized terminology, harmonized testing and evaluation frameworks, and clear reporting guidelines—crucial steps to ensure both progress and patient safety.

Why Would a Digital Health Tool Do a Poor Job at Screening Musculoskeletal Conditions?

Through the process of screening studies for inclusion into our review, we found definitions of musculoskeletal conditions that were vague and varied widely. Definitions of "musculoskeletal"

XSL•FO
RenderX

are often limited to orthopedic conditions or pain related to musculoskeletal structures [1,23]. However, musculoskeletal conditions are a complex category involving heterogeneous conditions, such as rheumatological or inflammatory arthritis or gout, that are not typically grouped as musculoskeletal in clinical practice. We relied on a broad definition to capture specific musculoskeletal conditions (eg, rheumatological conditions, arthritis, and gout) and pain related to musculoskeletal structures (eg, sprains and strains).

There is nuance in how triage would be conducted for acute versus chronic musculoskeletal conditions, including screening questions related to condition pathophysiology, subjective history, pattern of symptoms, and disability (eg, red flags), which might explain some of the variability in performance metrics of different digital tools [70]. Early diagnosis and treatment planning is often iterative for those with musculoskeletal conditions and varies depending on the condition. For example, targeted medication plays a vital role in managing rheumatological conditions [71], whereas some orthopedic conditions are managed with exercise and minimal pharmacological interventions [2]. This complexity will impact triage algorithms by influencing treatment recommendations (eg, who the patient should see) and timing of care (eg, urgent or wait-and-see). Therefore, tools that have high accuracy (ie, good performance) for triaging and diagnosing general health conditions may not necessarily have the same effectiveness when applied to musculoskeletal conditions.

Digital health tools may perform poorly at screening because of user error relating to symptom data entry and patient interaction with the tool. One solution to this is adding more key information (eg, diagnostic tests) to an AI-driven model to improve the diagnostic accuracy of the model [67]. We also suggest future work to involve patient end users to develop and refine digital health tools. Most digital health tool algorithms are derived from clinicians' clinical reasoning, which may not follow the same thought process as a patient. In a recent qualitative study exploring how patients should be engaged in AI application to health care, patients felt that the priorities of researchers, particularly for AI tools, were to improve efficiency and effectiveness of care [72]. In contrast, patients were more interested in using AI to address issues related to accessing health care [72]. Patients should be involved early in the design and development phases to enhance the usability and understandability of digital health tools. However, patient perspectives are often included only after the digital health tool is designed. We argue that engaging patients early in the development process, such as developing the AI algorithms, may yield more acceptable and usable digital health tools.

It is unlikely that a "one size fits all" digital health tool can effectively diagnose and triage all musculoskeletal conditions. Most patient-facing tools in our review were web- or app-based tools in the form of generic symptom checkers. ChatGPT has an accessible interface and is relatively easy to use [15]. Clinician-facing tools may benefit from greater complexity or condition specificity, depending on the context in which the tools will be implemented. Instead of an either-or—general or condition-specific—we advocate for designers to consider their design goal (ie, triage or diagnosis) and intended user (ie, patient or clinician), which may improve accuracy in digital health tools for musculoskeletal conditions.

Move (Relatively) Fast, and Try Not to Break Things

The field of digital health is growing and changing rapidly. Many health systems have been forced to move toward implementing digital health, particularly AI-driven tools, without being afforded adequate time and resources to consider safety, effectiveness, or downstream consequences [13]. This may be in part due to social and political imperatives to set key performance (productivity) indicators, transition of health care services, and drive toward greater and faster innovation. We suggest that such a climate could be dangerous for health care, especially if digital health implementation continues without adequate evidence, as our findings highlight.

There is a place for digital health triage tools used by patients and clinicians in the current health care context. Self-referral and symptom checkers can be effective for musculoskeletal conditions and to support patients' access to care, particularly when patients do not have a consistent primary care team or provider [11]. Acute care clinics using a self-referral form found that patients with musculoskeletal conditions were accurate at self-referring, used less health care, and incurred fewer costs [73]. Emerging evidence also indicates that patients are using LLMs such as ChatGPT to make health care decisions, and it appears that the general public is accepting of using AI for health care advice and psychological support [15]. However, more research is needed to ensure that patients presenting with musculoskeletal conditions have a safe, accurate, and well-designed tool to direct them to the best care for their situation. Digital health tools also need to be designed to suit diverse populations, including those with low health literacy and limited digital literacy.

Future Considerations and Clinical Implications

While there is a breadth of studies available for digital health and digital triage, we identified the following knowledge gaps: (1) reporting and transparency on digital health tool development must improve, (2) evaluating digital health tools needs a standard approach, (3) studying the accuracy of triage recommendations requires robust prospective studies, and (4) implementing musculoskeletal-focused digital health tools for first point-of-contact care requires attention.

Despite the absence of digital health tools for triage of musculoskeletal conditions, we are aware of other tools in development, such as SupportPrim [74], which might fill some of the knowledge gaps for health care providers. Our findings do not provide conclusive evidence to support using digital health tools to accurately screen musculoskeletal conditions in many health settings. We recommend that clinicians use these digital health tools as an adjunct to help guide patients, particularly when used as a symptom checker, but to still defer to sound clinical judgment and help patients understand the limitations of the tools.

Limitations

Although we used a thorough search of published and unpublished data, it is possible that we have missed relevant

XSL•FO
RenderX

digital health tools or papers. We set a sample threshold of at least 25% of the sample population with musculoskeletal conditions, and this may have resulted in us missing some studies (eg, studies that were just below the threshold were excluded). The threshold was intended to maximize external validity [26,27]. Our goal was to identify tools that were primarily designed to triage or diagnose (vs manage) musculoskeletal conditions. Therefore, we excluded studies and tools that were designed for self-management, even if they included a symptom checker. This led us to exclude studies that used tools for secondary triage or diagnosis (ie, used by patients who had a diagnosis or had already been seen in a primary or emergency setting) as we wanted to capture tools that could be used at the first point-of-contact. We identified some potential musculoskeletal-specific digital health tools that could be used for secondary triage or diagnosis (Multimedia Appendix 5). While we attempted to report on the performance and accuracy of the tools identified, some tools pooled data from the entire population (ie, not musculoskeletal only). Therefore, the findings may under- or overestimate the accuracy of the tool for musculoskeletal conditions. This again points to the need to design musculoskeletal-specific tools and carefully evaluate their performance.

Conclusions

The rapid growth of AI and digital health solutions is transforming health care systems worldwide, with increasing interest in automating triage and diagnosis. However, our review shows that musculoskeletal conditions remain a blind spot: few tools were specifically designed for this purpose, and most performed poorly when applied to musculoskeletal populations. Despite commercial availability and implementation in some settings, the evidence base was weak, and tool performance was inconsistent and opaque. Health systems and clinicians should exercise caution before integrating these tools into care pathways. Musculoskeletal-specific digital tools developed through transparent, standardized processes are urgently needed to ensure safety, clinical value, and trustworthiness.

Acknowledgments

Generative artificial intelligence was not used to draft any portion of this manuscript.

Funding

LKT is a Mitacs Elevate Fellow and a 2025 Health Research BC Research Trainee recipient and is funded by Health Research BC (RT-2025-04847)

Data Availability

The datasets generated or analyzed during this study are available in the Open Science Framework repository [22].

Authors' Contributions

Conceptualization: LKT, CLA, JGW (equal)
Data curation (database searching): DG (lead), LKT, CLA (supporting)
Investigation: LKT, JGW, RV, EL, JLC, EW, CS, CLA
Methodology: LKT, CLA (equal)
Formal analysis: LKT (lead), CLA, JGW, RV (supporting)
Project administration: LKT (lead), CLA (supporting)
Visualization: LKT (lead), JGW, CLA (supporting)
Writing—original draft: LKT (lead), CLA (supporting)
Writing—review and editing: All authors

Conflicts of Interest

None declared.

Multimedia Appendix 1
OVID MEDLINE full search strategy.
[DOCX File, 49 KB - jmir_v28i1e81578_app1.docx]

Multimedia Appendix 2
Gray literature search strategy and results.
[DOCX File, 27 KB - jmir_v28i1e81578_app2.docx]

Multimedia Appendix 3
Studies excluded at full-text stage.
[DOCX File, 73 KB - jmir_v28i1e81578_app3.docx]

XSL•FO
RenderX

Multimedia Appendix 4
Characteristics of included studies, summarizing design, demographics, and digital tool features.
[DOCX File, 57 KB - jmir_v28i1e81578_app4.docx]

Multimedia Appendix 5
Digital health tools for secondary triage or diagnosis.
[DOCX File, 25 KB - jmir_v28i1e81578_app5.docx]

Checklist 1
PRISMA-ScR checklist.
[PDF File, 249 KB - jmir_v28i1e81578_app6.pdf]

References

1. Gill TK, Mittinty MM, March LM, et al. Global, regional, and national burden of other musculoskeletal disorders, 1990–2020, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. Lancet Rheumatol 2023 Nov;5(11):e670-e682. [doi: 10.1016/S2665-9913(23)00232-1]
2. Lin I, Wiles L, Waller R, et al. What does best practice care for musculoskeletal pain look like? Eleven consistent recommendations from high-quality clinical practice guidelines: systematic review. Br J Sports Med 2020 Jan;54(2):79-86. [doi: 10.1136/bjsports-2018-099878] [Medline: 30826805]
3. Wrightson J, Truong LK, Haagaard A, Ardern CL. Estimating the prevalence of low acuity musculoskeletal pain in the emergency department. Presented at: Canadian for Health Services and Policy Research (CHSPR) 2025 Conference Abstract; Mar 3-4, 2025.
4. Iserson KV, Moskop JC. Triage in medicine, part I: concept, history, and types. Ann Emerg Med 2007 Mar;49(3):275-281. [doi: 10.1016/j.annemergmed.2006.05.019] [Medline: 17141139]
5. Joseph C, Morrissey D, Abdur-Rahman M, Hussenbux A, Barton C. Musculoskeletal triage: a mixed methods study, integrating systematic review with expert and patient perspectives. Physiotherapy 2014 Dec;100(4):277-289. [doi: 10.1016/j.physio.2014.03.007] [Medline: 25242531]
6. Erku D, Khatri R, Endalamaw A, et al. Digital health interventions to improve access to and quality of primary health care services: a scoping review. Int J Environ Res Public Health 2023 Sep 28;20(19):19. [doi: 10.3390/ijerph20196854] [Medline: 37835125]
7. Ibrahim MS, Mohamed Yusoff H, Abu Bakar YI, Thwe Aung MM, Abas MI, Ramli RA. Digital health for quality healthcare: a systematic mapping of review studies. Digit Health 2022;8:20552076221085810. [doi: 10.1177/20552076221085810] [Medline: 35340904]
8. Chambers D, Cantrell AJ, Johnson M, et al. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. BMJ Open 2019 Aug 1;9(8):e027743. [doi: 10.1136/bmjopen-2018-027743] [Medline: 31375610]
9. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2025-10-24]
10. Recommendations on digital interventions for health system strengthening—executive summary. World Health Organization. 2019. URL: https://www.who.int/publications/i/item/WHO-RHR-19.8 [accessed 2025-01-05]
11. Babatunde OO, Bishop A, Cottrell E, et al. A systematic review and evidence synthesis of non-medical triage, self-referral and direct access services for patients with musculoskeletal pain. PLOS ONE 2020;15(7):e0235364. [doi: 10.1371/journal.pone.0235364] [Medline: 32628696]
12. Pairon A, Philips H, Verhoeven V. A scoping review on the use and usefulness of online symptom checkers and triage systems: how to proceed? Front Med (Lausanne) 2022;9:1040926. [doi: 10.3389/fmed.2022.1040926] [Medline: 36687416]
13. Tyler S, Olis M, Aust N, et al. Use of artificial intelligence in triage in hospital emergency departments: a scoping review. Cureus 2024 May;16(5):e59906. [doi: 10.7759/cureus.59906] [Medline: 38854295]
14. Wallace W, Chan C, Chidambaram S, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. NPJ Digit Med 2022 Aug 17;5(1):118. [doi: 10.1038/s41746-022-00667-w] [Medline: 35977992]
15. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors 2023 May 17;10:e47564. [doi: 10.2196/47564] [Medline: 37195756]
16. Peters MDJ, Godfrey C, McInerney P, et al. Best practice guidance and reporting items for the development of scoping review protocols. JBI Evid Synth 2022 Apr 1;20(4):953-968. [doi: 10.11124/JBIES-21-00242] [Medline: 35102103]
17. Pollock D, Peters MDJ, Khalil H, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. JBI Evid Synth 2023 Mar 1;21(3):520-532. [doi: 10.11124/JBIES-22-00123] [Medline: 36081365]
18. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]

XSL•FO
RenderX

19. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
20. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
21. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5:69. [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
22. Truong LK, Lui E, Wrightson J, et al. Digital health tools used to triage musculoskeletal pain in primary, urgent and emergency settings: a scoping review. Open Science Framework. 2024. URL: https://osf.io/y5rp7/overview [accessed 2026-01-05]
23. Smith E, Hoy DG, Cross M, et al. The global burden of other musculoskeletal disorders: estimates from the Global Burden of Disease 2010 study. Ann Rheum Dis 2014 Aug;73(8):1462-1469. [doi: 10.1136/annrheumdis-2013-204680] [Medline: 24590181]
24. Technology Readiness Level (TRL) Assessment Tool. Government of Canada. 2021. URL: https://ised-isde.canada.ca/site/clean-growth-hub/en/technology-readiness-level-trl-assessment-tool [accessed 2025-05-24]
25. Walton A. December 2022—title and abstract screening using machine learning. Covidence. 2022. URL: https://www.covidence.org/blog/release-notes-december-2022-machine-learning [accessed 2024-07-11]
26. Bird S, Thompson C, Williams KE. Primary contact physiotherapy services reduce waiting and treatment times for patients presenting with musculoskeletal conditions in Australian emergency departments: an observational study. J Physiother 2016 Oct;62(4):209-214. [doi: 10.1016/j.jphys.2016.08.005] [Medline: 27637771]
27. Gagnon R, Perreault K, Berthelot S, et al. Direct-access physiotherapy to help manage patients with musculoskeletal disorders in an emergency department: results of a randomized controlled trial. Acad Emerg Med 2021 Aug;28(8):848-858. [doi: 10.1111/acem.14237] [Medline: 33617696]
28. Badahman F, Alsobhi M, Alzahrani A, et al. Validating the accuracy of a patient-facing clinical decision support system in predicting lumbar disc herniation: diagnostic accuracy study. Diagnostics (Basel) 2024 Aug 26;14(17):1870. [doi: 10.3390/diagnostics14171870] [Medline: 39272655]
29. Badsha HM, Khan B, Harifi G, J A, Raman S. AB1488 is the future of rheumatology here? A study of a proprietary rule engine and artificial intelligence GPT4 (AI GPT4) for initial evaluation of rheumatology cases. Ann Rheum Dis 2024 Jun;83(Suppl 1):2112. [doi: 10.1136/annrheumdis-2024-eular.1942] [Medline: 644868572]
30. Bisson LJ, Komm JT, Bernas GA, et al. Accuracy of a computer-based diagnostic program for ambulatory patients with knee pain. Am J Sports Med 2014 Oct;42(10):2371-2376. [doi: 10.1177/0363546514541654] [Medline: 25073597]
31. Bisson LJ, Komm JT, Bernas GA, et al. How accurate are patients at diagnosing the cause of their knee pain with the help of a web-based symptom checker? Orthop J Sports Med 2016 Feb;4(2):2325967116630286. [doi: 10.1177/2325967116630286] [Medline: 26962542]
32. Bond C, Guard M, Grinbergs P. Case report: digital musculoskeletal triage and rehabilitation tools enhance accessibility, user experience and outcomes in mechanical knee pain. Physiotherapy 2024 Jun;123(Suppl 1):e115. [doi: 10.1016/j.physio.2024.04.143]
33. Carmona KA, Chittamuru D, Kravitz RL, Ramondt S, Ramírez AS. Health information seeking from an intelligent web-based symptom checker: cross-sectional questionnaire study. J Med Internet Res 2022 Aug 19;24(8):e36322. [doi: 10.2196/36322] [Medline: 35984690]
34. Daher M, Koa J, Boufadel P, Singh J, Fares MY, Abboud JA. Breaking barriers: can ChatGPT compete with a shoulder and elbow specialist in diagnosis and management? JSES Int 2023 Nov;7(6):2534-2541. [doi: 10.1016/j.jseint.2023.07.018] [Medline: 37969495]
35. Demmelmaier I, Denison E, Lindberg P, Asenlöf P. Physiotherapists' telephone consultations regarding back pain: a method to analyze screening of risk factors. Physiother Theory Pract 2010 Oct;26(7):468-475. [doi: 10.3109/09593980903433938] [Medline: 20649497]
36. Dias L, Maughan E, Kisha A, Moorthy R. Telephone triage in the management of patients with nasal injuries [Abstract]. Clin Otolaryngol 2012;37:45. [doi: 10.1111/j.1749-4486.2012.02517.x] [Medline: 71023181]
37. Gräf M, Knitza J, Leipe J, et al. Comparison of physician and artificial intelligence-based symptom checker diagnostic accuracy. Rheumatol Int 2022 Dec;42(12):2167-2176. [doi: 10.1007/s00296-022-05202-4] [Medline: 36087130]
38. Gómez-Centeno A, Sabaris-Vilas M, Garcia-Sancho F, Segura-Sanchez J. POS0883 Optimizing rheumatology consultations with artificial intelligence: insights from the ReumAI pilot study. Ann Rheum Dis 2025 Jun;84(Suppl 1):1018. [doi: 10.1016/j.ard.2025.06.238]
39. Gymer M, Guard M, Grinbergs P. A case report: digital musculoskeletal triage and rehabilitation tools improve outcomes and offer a positive experience for lower back pain. JMIR Bioinform Biotechnol. Preprint posted online on Oct 20, 2022. [doi: 10.2196/preprints.43686]
40. Hageman M, Anderson J, Blok R, Bossen JKJ, Ring D. Internet self-diagnosis in hand surgery. HAND (N Y) 2015 Sep;10(3):565-569. [doi: 10.1007/s11552-014-9707-x] [Medline: 26330798]

XSL•FO
RenderX

41. Hannah L, von Sophie R, Gabriella RM, et al. Stepwise asynchronous telehealth assessment of patients with suspected axial spondyloarthritis: results from a pilot study. Rheumatol Int 2024 Jan;44(1):173-180. [doi: 10.1007/s00296-023-05360-z] [Medline: 37316631]
42. Hara T, Nishizuka T, Yamamoto M, Iwatsuki K, Natsume T, Hirata H. Teletriage for patients with traumatic finger injury directing emergency medical transportation services to appropriate hospitals: a pilot project in Nagoya City, Japan. Injury 2015 Jul;46(7):1349-1353. [doi: 10.1016/j.injury.2015.02.022] [Medline: 25799472]
43. Jakobi S, Boy K, Wagner M, et al. Rheumatic? A diagnostic decision support tool for individuals suspecting rheumatic diseases: mixed-methods usability and acceptability study. BMC Rheumatol 2025 May 23;9(1):59. [doi: 10.1186/s41927-025-00507-w] [Medline: 40410901]
44. Kelly M, Higgins A, Murphy A, McCreesh K. A telephone assessment and advice service within an ED physiotherapy clinic: a single-site quality improvement cohort study. Arch Physiother 2021 Feb 8;11(1):4. [doi: 10.1186/s40945-020-00098-4] [Medline: 33550990]
45. Knevel R, Knitza J, Hensvold A, et al. Rheumatic?—A digital diagnostic decision support tool for individuals suspecting rheumatic diseases: a multicenter pilot validation study. Front Med 2022;9. [doi: 10.3389/fmed.2022.774945] [Medline: 2016520483]
46. Knitza J, Mohn J, Bergmann C, et al. Accuracy, patient-perceived usability, and acceptance of two symptom checkers (Ada and Rheport) in rheumatology: interim results from a randomized controlled crossover trial. Arthritis Res Ther 2021 Apr 13;23(1):112. [doi: 10.1186/s13075-021-02498-8] [Medline: 33849654]
47. Knitza J, Muehlensiepen F, Ignatyev Y, et al. Patient's perception of digital symptom assessment technologies in rheumatology: results from a multicentre study. Front Public Health 2022;10. [doi: 10.3389/fpubh.2022.844669] [Medline: 637486321]
48. Knitza J, Tascilar K, Fuchs F, et al. Diagnostic accuracy of a mobile AI-based symptom checker and a web-based self-referral tool in rheumatology: multicenter randomized controlled trial. J Med Internet Res 2024 Jul 23;26:e55542. [doi: 10.2196/55542] [Medline: 39042425]
49. Krusche M, Callhoff J, Knitza J, Ruffer N. Diagnostic accuracy of a large language model in rheumatology: comparison of physician and ChatGPT-4. Rheumatol Int 2024 Feb;44(2):303-306. [doi: 10.1007/s00296-023-05464-6] [Medline: 37742280]
50. Li KY, Kim PS, Thariath J, Wong ES, Barkham J, Kocher KE. Standard nurse phone triage versus tele-emergency care pilot on Veteran use of in-person acute care: an instrumental variable analysis. Acad Emerg Med 2023 Apr;30(4):310-320. [doi: 10.1111/acem.14681] [Medline: 36757685]
51. Lowe C, Browne M, Marsh W, Morrissey D. Usability testing of a digital assessment routing tool for musculoskeletal disorders: iterative, convergent mixed methods study. J Med Internet Res 2022 Aug 30;24(8):e38352. [doi: 10.2196/38352] [Medline: 36040787]
52. Lowe C, Sephton R, Marsh W, Morrissey D. Evaluation of a musculoskeletal Digital Assessment Routing Tool (DART): crossover noninferiority randomized pilot trial. JMIR Form Res 2024 Jul 30;8:e56715. [doi: 10.2196/56715] [Medline: 39078682]
53. Lundberg K, Qin L, Aulin C, van Spil WE, Maurits MP, Knevel R. Population-based user-perceived experience of Rheumatic?: a novel digital symptom-checker in rheumatology. RMD Open 2023 Apr;9(2):e002974. [doi: 10.1136/rmdopen-2022-002974] [Medline: 37094982]
54. Martin MJ, Payne KM. Using digital technology and user-centred design to develop a physiotherapy self-referral service for back pain. Physiotherapy 2020 May;107(Suppl 1):e139-e140. [doi: 10.1016/j.physio.2020.03.203]
55. Phillips CJ, Phillips Nee Buck R, Main CJ, et al. The cost effectiveness of NHS physiotherapy support for occupational health (OH) services. BMC Musculoskelet Disord 2012 Feb 23;13(1):29. [doi: 10.1186/1471-2474-13-29] [Medline: 22361319]
56. Qin L, Zegers F, Selani D, et al. Differentiation of immune mediated versus non immune mediated rheumatic diseases by online symptom checker in real-world patients—multiple diagnoses and particularly fibromyalgia is a stumbling block. Ann Rheum Dis 2024 Jun;83(Suppl 1):2082-2083. [doi: 10.1136/annrheumdis-2024-eular.5438] [Medline: 644868620]
57. Ryan K, Grinbergs P. Demographic analysis of users of a musculoskeletal physiotherapy self-referral digital triage tool in Bromley. Physiotherapy 2024 Jun;123:e210-e211. [doi: 10.1016/j.physio.2024.04.263]
58. Salisbury C, Montgomery AA, Hollinghurst S, et al. Effectiveness of PhysioDirect telephone assessment and advice services for patients with musculoskeletal problems: pragmatic randomised controlled trial. BMJ 2013 Jan 29;346(7893):f43. [doi: 10.1136/bmj.f43] [Medline: 23360891]
59. Soin A, Hirschbeck M, Verdon M, Manchikanti L. A pilot study implementing a machine learning algorithm to use artificial intelligence to diagnose spinal conditions. Pain Physician 2022 Mar;25(2):171-178. [Medline: 35322974]
60. Tan T, Santosa A, Roslan N, Li J. The development of an AI-based conversational agent for screening of rheumatic diseases [Abstract]. Int J Rheum Dis 2023;26(9). [doi: 10.1111/1756-185X.14505]
61. Trivedi SV, Batta R, Henao-Romero N, Mondal P, Wilson T, Stempien J. A comparison of self-triage tools to nurse driven triage in the emergency department. PLOS ONE 2024;19(8):e0297321. [doi: 10.1371/journal.pone.0297321] [Medline: 39196994]

XSL•FO
RenderX

62. Edwards J, Hayden J, Asbridge M, Gregoire B, Magee K. Prevalence of low back pain in emergency settings: a systematic review and meta-analysis. BMC Musculoskelet Disord 2017 Apr 4;18(1):143. [doi: 10.1186/s12891-017-1511-7] [Medline: 28376873]
63. Lowe C, Atherton L, Lloyd P, Waters A, Morrissey D. Improving safety, efficiency, cost, and satisfaction across a musculoskeletal pathway using the digital assessment routing tool for triage: quality improvement study. J Med Internet Res 2025 Apr 25;27:e67269. [doi: 10.2196/67269] [Medline: 40279646]
64. Burgess R, Tucker K, Smithson R, Dimbleby P, Casey C. Optimising musculoskeletal patient flow through digital triage and supported self-management: a service evaluation set within community musculoskeletal care. Musculoskelet Care 2024 Dec;22(4):e70013. [doi: 10.1002/msc.70013] [Medline: 39625285]
65. Xu H, Shuttleworth KMJ. Medical artificial intelligence and the black box problem: a view based on the ethical principle of “do no harm”. Intell Med 2024 Feb;4(1):52-57. [doi: 10.1016/j.imed.2023.08.001]
66. Mruthyunjaya P, Verma S, Agarwal A, Maharana U, Mandal M, Ahmed S. Right diagnoses but wrong reasoning: current large-language model-based agentic frameworks have flawed clinical reasoning despite high diagnostic accuracy. The Lancet. Preprint posted online on Jul 9, 2025. [doi: 10.2139/ssrn.5339074]
67. Kremer P, Schiebisch H, Lechner F, et al. Comparative analysis of large language models and traditional diagnostic decision support systems for rare rheumatic disease identification. EULAR Rheumatol Open 2025 Jun;1(2):51-59. [doi: 10.1016/j.ero.2025.04.007]
68. Zegers F, Qin L, Selani D, et al. POS1131 prediction models for rheumatic diseases: from clinical simplicity to data-driven complexity with patient-reported symptoms for an online symptom checker. Ann Rheum Dis 2025 Jun;84:1211-1212. [doi: 10.1016/j.ard.2025.06.481] [Medline: 19279015]
69. Segur-Ferrer J, Moltó-Puigmartí C, Pastells-Peiró R, Vivanco-Hidalgo RM. Methodological frameworks and dimensions to be considered in digital health technology assessment: scoping review and thematic analysis. J Med Internet Res 2024 Apr 10;26:e48694. [doi: 10.2196/48694] [Medline: 38598288]
70. Triage guidelines for orthopaedic optimisation pathway (based on musculoskeletal (MSK) referral) V6.0. South East London Integrated Care Board. 2024. URL: https://www.selondonics.org/wp-content/uploads/MSK-Triage-Guidelines-v6-Final.pdf [accessed 2025-07-14]
71. Fraenkel L, Bathon JM, England BR, et al. 2021 American College of Rheumatology Guideline for the treatment of rheumatoid arthritis. Arthritis Care Res (Hoboken) 2021 Jul;73(7):924-939. [doi: 10.1002/acr.24596] [Medline: 34101387]
72. Adus S, Macklin J, Pinto A. Exploring patient perspectives on how they can and should be engaged in the development of artificial intelligence (AI) applications in health care. BMC Health Serv Res 2023 Oct 26;23(1):1163. [doi: 10.1186/s12913-023-10098-2] [Medline: 37884940]
73. Lau BHF, Lafave MR, Mohtadi NG, Butterwick DJ. Utilization and cost of a new model of care for managing acute knee injuries: the Calgary Acute Knee Injury Clinic. BMC Health Serv Res 2012 Dec 5;12(1):445. [doi: 10.1186/1472-6963-12-445] [Medline: 23216946]
74. Lervik LCN, Vasseljen O, Austad B, et al. SupportPrim—a computerized clinical decision support system for stratified care for patients with musculoskeletal pain complaints in general practice: study protocol for a randomized controlled trial. Trials 2023 Apr 11;24(1):267. [doi: 10.1186/s13063-023-07272-6] [Medline: 37041631]

Abbreviations

AI: artificial intelligence
DART: Digital Assessment Routing Tool
ED: emergency department
LLM: large language model
mHealth: mobile health
OSF: Open Science Framework
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Search
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
TRL: technology readiness level

XSL•FO
RenderX

Edited by S Brini; submitted 31.Jul.2025; peer-reviewed by J Knitza; revised version received 08.Dec.2025; accepted 09.Dec.2025; published 14.Jan.2026.

Please cite as:
Truong LK, Wrightson JG, Vincent R, Lui E, Couch JL, Wang E, Starcevich C, Giustini D, Haagaard A, Lopatina E, van Berkel N, Rathleff MS, Ardern CL
Evidence for Digital Health Tools Designed to Support the Triage of Musculoskeletal Conditions in Primary, Urgent, and Emergency Care Settings: Scoping Review
J Med Internet Res 2026;28:e81578
URL: https://www.jmir.org/2026/1/e81578
doi:10.2196/81578

© Linda K Truong, James G Wrightson, Raphaël Vincent, Eunice Lui, Jamon L Couch, Ellen Wang, Cobie Starcevich, Dean Giustini, Alex Haagaard, Elena Lopatina, Niels van Berkel, Michael Skovdal Rathleff, Clare L Ardern. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Diagnostic Value of Image-Based Machine Learning for Osteoporosis: Systematic Review and Meta-Analysis

Rui Zhao[1], MM; Haolin Yang[1], MM; Yangbo Li[1], MM; Xiaoyun Li[1], MM; Zhijie Yang[1], MM; Yanping Lin[2], MM; Jiachun Huang[2], MD; Lei Wan[2], MD; Hongxing Huang[2], MD

[1]The Third Clinical College of Medicine, Guangzhou University of Traditional Chinese Medicine, Guangzhou, China

[2]The Third Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou University of Chinese Medicine, 261 Longxi Avenue, Liwan District, Guangzhou, China

Corresponding Author:
Hongxing Huang, MD
The Third Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou University of Chinese Medicine, 261 Longxi Avenue, Liwan District, Guangzhou, China

Abstract

Background: Osteoporosis (OP) is projected to be a major issue significantly impacting the well-being of middle-aged and old populations. Machine learning (ML) and deep learning (DL) models developed based on medical imaging have enhanced clinicians' diagnostic accuracy and work efficiency. However, the diagnostic performance of different types of medical imaging for OP has not been systematically assessed.

Objective: By summarizing related literature, this study aims to elucidate the role of DL models based on different medical imaging modalities in OP detection.

Methods: PubMed, Embase, the Cochrane Library, and Web of Science were systematically searched for studies using ML for the diagnosis of OP based on medical imaging. The final search was conducted on May 16, 2024. The risk of bias in the included studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 tool. A bivariate mixed-effects model was applied to perform meta-analyses of sensitivity (SEN) and specificity (SPC), stratified by imaging modality (x-ray, computed tomography [CT], magnetic resonance imaging [MRI]). In addition, subgroup analyses were carried out based on the type of ML algorithm, the method of validation dataset generation, and the anatomical site of assessment.

Results: A total of 60 studies comprising 66,195 participants were encompassed in this systematic review and meta-analysis. Among these, 22 studies used x-ray imaging, 37 applied CT imaging, and 3 used MRI for ML-based OP diagnosis. For x-ray–based models, the pooled SEN and SPC for studies focusing on the appendicular skeleton were 0.97 (95% CI 0.83 - 0.99) and 0.90 (95% CI 0.75 - 0.96), respectively. For studies using the mandible as the target site, SEN and SPC were 0.94 (95% CI 0.89 - 0.97) and 0.80 (95% CI 0.56 - 0.93), respectively. For those focusing on the lumbar spine, the pooled SEN and SPC were 0.87 (95% CI 0.77 - 0.93) and 0.82 (95% CI 0.75 - 0.87), respectively. For CT-based models, studies targeting the hip joint reported a pooled SEN and SPC of 0.87 (95% CI 0.83 - 0.90) and 0.92 (95% CI 0.81 - 0.96), respectively. For the thoracic spine, SEN and SPC were 0.91 (95% CI 0.86 - 0.94) and 0.94 (95% CI 0.92 - 0.95), respectively, while for the lumbar spine, they were 0.91 (95% CI 0.87 - 0.94) and 0.92 (95% CI 0.86 - 0.95), respectively.

Conclusions: ML based on medical imaging demonstrates high diagnosis accuracy for OP, particularly DL models using x-ray and CT modalities. However, this study included only a limited number of original studies using MRI-based ML, and there remains a lack of adequate external validation across studies, which poses interpretative limitations. Future research should aim to develop artificial intelligence tools with broader applicability and enhanced diagnostic precision.

(J Med Internet Res 2026;28:e75965) doi:10.2196/75965

KEYWORDS

osteoporosis; machine learning; artificial intelligence; systematic review; diagnostic imaging

Introduction

Osteoporosis (OP), a metabolic disorder, features a systemic reduction in bone mass and impaired bone microarchitecture and elevates the risk of fragility fractures. As the most prevalent chronic metabolic bone disease, it is strongly associated with advancing age, posing significant health threats. However, due to its insidious onset, prolonged disease course, and challenges in treatment, public awareness and attention toward OP prevention and management remain insufficient [1,2]. With the emerging global trend of population aging, OP is projected to become a major issue adversely affecting the quality of life of

XSL•FO
RenderX

middle-aged and older people. Epidemiological studies estimate that by 2050, the global population at high risk of fractures will surge to 6.26 million from 1.66 million in 1990 [3]. This escalation imposes immense social pressures and substantial economic burdens on early OP screening, prevention, and treatment.

At present, a variety of diagnostic methods are available for the clinical assessment of OP. Among them, dual-energy x-ray absorptiometry (DXA) for measuring *T* scores recommended by the World Health Organization is regarded as the authoritative and standardized technique [4]. Although DXA is widely used, it is unable to assess whole-body skeletal, fat, and lean mass, which restricts its utility in the routine diagnosis or evaluation of OP [5]. Moreover, due to disparities in socioeconomic development across different regions worldwide, DXA is not accessible in underdeveloped countries and regions. Therefore, some high-risk populations, such as postmenopausal women and older adults, are not detected and untreated. Medical imaging is crucial in clinical diagnosis and treatment. However, the hidden features within imaging techniques including x-rays, computed tomography (CT), and magnetic resonance imaging (MRI) are often overlooked due to low spatial resolution and high contrast resolution [6].

In the 1980s, computer-aided diagnosis (CAD) systems were developed to deeply interpret key features in medical images, providing radiologists with valuable insights into image interpretation [7]. Currently, CAD tools primarily include traditional machine learning (ML) models built on explainable clinical features and deep learning (DL) models developed using pathological or nuclear medicine images. They assist clinicians in disease diagnosis and prognostic prediction. Increasing evidence has demonstrated the utility of CAD in diagnosing conditions such as autism [8], pulmonary embolism [9], breast cancer [10], and bone metastases [11]. DL approaches based on medical imaging have attracted substantial research interest. Against this backdrop, ML models based on various imaging modalities such as x-rays, CT, and MRI have been constructed to diagnose OP [12]. However, the diagnostic performance of various imaging methods in OP is not supported by systematic evidence. This hindered the application of artificial intelligence (AI)–based CAD tools in OP and posed challenges for further systematic development.

Therefore, our study seeks to provide a comprehensive review of DL research in the diagnosis of OP based on medical imaging modalities, including x-ray, CT, and MRI. Furthermore, this study aims to analyze and evaluate the feasibility and accuracy of AI-driven DL in enhancing the screening and diagnostic rates of OP, thereby offering robust support for the prevention and management of the disease.

Methods

Study Registration

This study followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines and was prospectively registered on PROSPERO (CRD42024567736).

Eligibility Criteria

The eligible studies were (1) case-control, cohort, or cross-sectional studies; (2) papers with comprehensively developed image-based DL models for OP diagnosis; and (3) English publications. The following studies were excluded: (1) studies that only developed traditional ML models, (2) those that performed image segmentation without a complete DL model, and (3) those lacking outcome measures for evaluating the DL model's accuracy. Outcome measures must include at least 1 of the following: c-statistic, sensitivity (SEN), specificity (SPC), accuracy, recall, precision, confusion matrix, F_1-score, or calibration curve.

Data Sources and Search Strategy

PubMed, Cochrane, Embase, and Web of Science databases were thoroughly retrieved up to May 16, 2024. Both MeSH and free-text terms were used without restrictions on geographic location or study type. The search strategy is detailed in Multimedia Appendix 1.

Study Selection

The retrieved literature was uploaded to EndNote (Thomson Corporation), and duplicates were ostracized. Titles and abstracts were reviewed to identify potentially eligible studies. Full-text papers were subsequently screened to determine the eligible ones. Two researchers (RZ and HY) independently conducted the literature screening and cross-checked their results. Dissents were addressed by a third researcher (YL).

Data Extraction

The eligible papers were imported into EndNote, and data extraction was performed. A standard electronic data extraction form was developed beforehand to capture the following information: title, DOI, first author, publication year, author's country, study type, patient source, OP diagnosis criteria, medical imaging, background population, gender, age, use of image segmentation, number of OP cases, total cases, number of OP and total cases in the training or validation set, validation set generation method, model type, and comparison with clinical practitioners. Data were independently extracted by 2 researchers (RZ and HY), followed by cross-checking. There was a high level of agreement between the 2 reviewers in the screening process (Cohen κ=0.879). In cases of disagreement, a third reviewer (YL) would assist in addressing it.

Risk of Bias in Studies

The bias of risk in the eligible studies was assessed via Quality Assessment of Diagnostic Accuracy Studies-2, a tool for evaluating the collation risk of bias and clinical applicability of original diagnostic studies [13]. Quality Assessment of Diagnostic Accuracy Studies-2 covers 5 domains: case selection, trials to be evaluated, reference standard, case flow, and progress, with each involving a few specific questions. The answer of "Yes," "No," or "Uncertain" corresponds to a low, high, or uncertain risk of bias. The risk of bias was deemed low if all of the landmark questions within a range were answered with "Yes"; if one of the informative questions was answered with "No," bias may exist, and the evaluators must determine the risk of bias in line with the established guidelines. The risk

XSL•FO
RenderX

of bias must be judged by the evaluation authors as per the established criteria. An unclear risk indicated that the studies reported sufficient details. Therefore, evaluators could not make a definitive judgment.

The risk of bias in studies was independently conducted by 2 researchers (RZ and HY), followed by cross-checking. If any dissent arose, a third researcher (YL) would assist in addressing it.

Synthesis Methods

The meta-analysis was carried out via a bivariate mixed-effects model based on diagnostic 2×2 contingency tables. However, some of the original studies did not provide complete 2×2 diagnostic data. In such cases, the necessary information was derived using SEN, SPC, positive predictive value, negative predictive value, and accuracy, in conjunction with the corresponding sample sizes. The meta-analysis reported pooled estimates of SEN, SPC, positive likelihood ratio (PLR), negative likelihood ratio (NLR), diagnostic odds ratio (DOR), and the summary receiver operating characteristic (SROC) curve along with their corresponding 95% CIs. Publication bias across studies was assessed through Deeks' funnel plot, while the clinical utility of the predictive models was evaluated via Fagan's nomogram. Subgroup analyses were performed based on imaging modality (x-ray, CT, and MRI), modeling approach (traditional ML vs DL), and validation strategy. It is important to note that the bivariate mixed-effects model requires a minimum of four 2×2 diagnostic tables. As the ML models based on MRI images only provided 3 such tables, a narrative synthesis was performed for this subgroup instead. A 2-sided *P* value of <.05 denoted statistical significance.

Results

Study Registration

A total of 3427 papers were retrieved, including 685 from PubMed, 15 from Cochrane, 1942 from Embase, and 785 papers from Web of Science. Among them, 639 papers were duplicates and were excluded. After the title and abstract review, 2587 studies unrelated to the study topic were removed. Full texts of the rest were subsequently reviewed. In total, 23 conference abstracts without full-text publications and 68 that did not include medical imaging in the modeling process were ostracized. Ultimately, 60 studies were included in the analysis (Figure 1) [14-73]. This study was conducted in accordance with the PRISMA 2020 checklist (Checklist 1).

XSL•FO
RenderX

Figure 1. Literature screening process. RCT: randomized controlled trial.

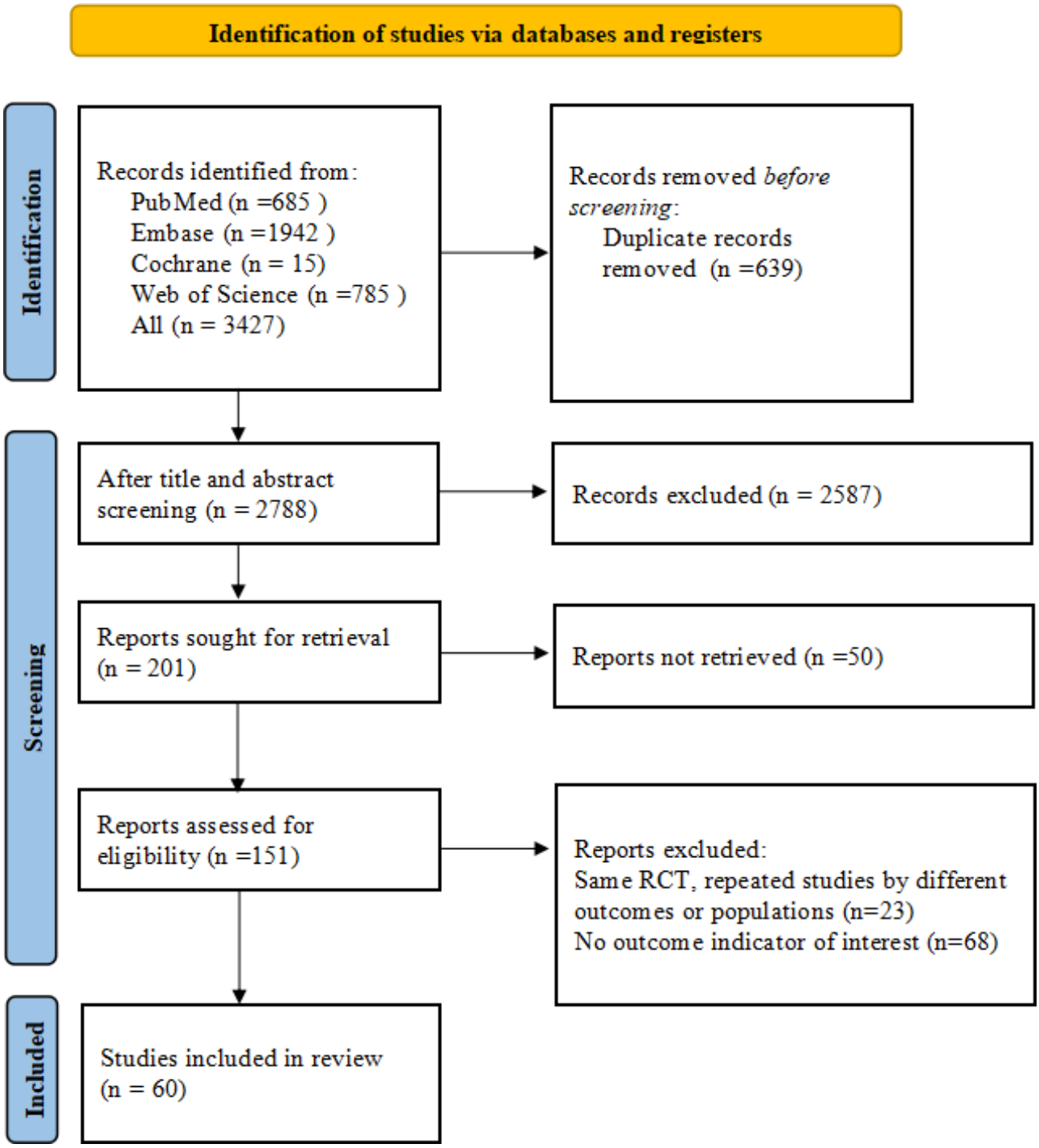

Study Characteristics

Among the 60 studies included in our analysis, 55 were case-control studies [14-21,25-35,37-67,69-73], and 5 were cohort studies [22-24,36,68]. These studies were predominantly published between 2012 and 2024 and involved 66,195 cases. These studies were published in 11 countries, including China (n=27), South Korea (n=10), the United States (n=8), and Japan (n=5) [15-17,20,22,23,26-33,35-48,50,51,53-60,62-73]. A smaller number of studies were from India (n=3), Saudi Arabia (n=2), Jordan (n=1), Latvia (n=1), Malaysia (n=1), Poland (n=1), and Switzerland (n=1) [14,18,19,21,24,25,34,39,52,61]. In total, 57 studies reported patient sources, of which 42 were single-center studies [14,15,17,19-23,26,29,31,32,36-42,47,48,50-60,62,64-67,69-73], 12 were multicenter studies [16,27,28,30,33,34,43-45,61,63,68], and 4 used database sources [24,41,46,49]. In terms of OP diagnosis, 47 studies explicitly provided diagnosis criteria [16,17,20-23,25,27-33,35-42,44,45,47,50-58,60-64,66-73]. Regarding medical imaging, 37 studies developed CT-based imaging models [15,16,18,19,21-23,25,26,29,31-33,38-41,43-45,47,48,50-54,56-58,65,67-69,72,73], 22 developed x-ray–based models [14,17,20,24,27,28,30,34-37,42,46,49,55,59-63,70,71], and 3 focused on MRI-based models [45,64,66]. Concerning the population, 4 studies specifically examined postmenopausal

XSL•FO
RenderX

women [16,35,36,52,71], and 1 study focused on men aged 50 years and older [16]. In terms of image processing, 48 studies used manual segmentation techniques to define regions of interest for analysis [14-19,21-23,25-36,38,39,41,42,44,45,47,49,52-57,59-66,68,70-73], while 12 did not define regions of interest [20,24,37,40,43,46,48,50,51,58,67,69]. Regarding the skeletal parts, 33 studies used lumbar vertebrae images [15-17,20,22,23,26,29-32,36,38-41,45,47,48,50,53-55,57,62,64-68,70,71,73], 9 used thoracic vertebrae images [23,25,31,38,44,48,53,55,56], 10 used hip images (including femoral neck, femoral head, and pelvis) [18,21,33,36,57,62,68,70,72], 7 used mandible images [17,36,42,52,60,63,71], and 10 used images of limb bones [14,28,34,35,43,51,58,59,61,69]. Regarding the generation of validation sets, 35 studies adopted random sampling [15-17,20-23,25,27,29-32,36,38,40-42,44,48,50-53,56,57,60-62,64,65,67,69,71,72], 12 used K-fold cross-validation [14,24,34,37,39,46,49,58,59,66,70,73], and 5 applied external validation [28,33,45,54,63]. In total, 9 studies compared their results with the screening results of clinicians [22,25,28,45-47,60,62,65]. In terms of model construction, 32 built DL models [18,21,23-31,33,37-42,44-47,52-56,59-62,66], and 28 constructed ML models [14-17,19,20,22,32,34-36,43,48-51,57,58,63-65,67-73] (Multimedia Appendix 2).

Risk of Bias in Studies

In all eligible studies, consecutive cases were included. Although most studies were case-control studies, 32 developed DL models, with variables derived from medical images. Therefore, these studies demonstrated a low risk of bias in case selection. In total, 28 studies applied ML models, where the process of variable generation might be influenced by the case-control study design, thereby leading to a higher risk of bias. Since this research is a meta-analysis of ML, whether or not the reference standards for OP diagnosis are known does not affect the results. Additionally, the criteria for determining positive results were pre-established, indicating that the trials under evaluation posed a low risk of bias. The implementation of a reference standard for OP diagnosis was considered reasonable, thus introducing a low risk of bias. Furthermore, there was a proper time interval between the trial and reference standard, and all patients in a given study followed the same diagnosis rules, with no cases omitted. Therefore, there was a low risk of bias in clinical applicability [14-73] (Figure 2 and Multimedia Appendix 3).

XSL•FO
RenderX

Figure 2. Risk of bias plot [14-73].

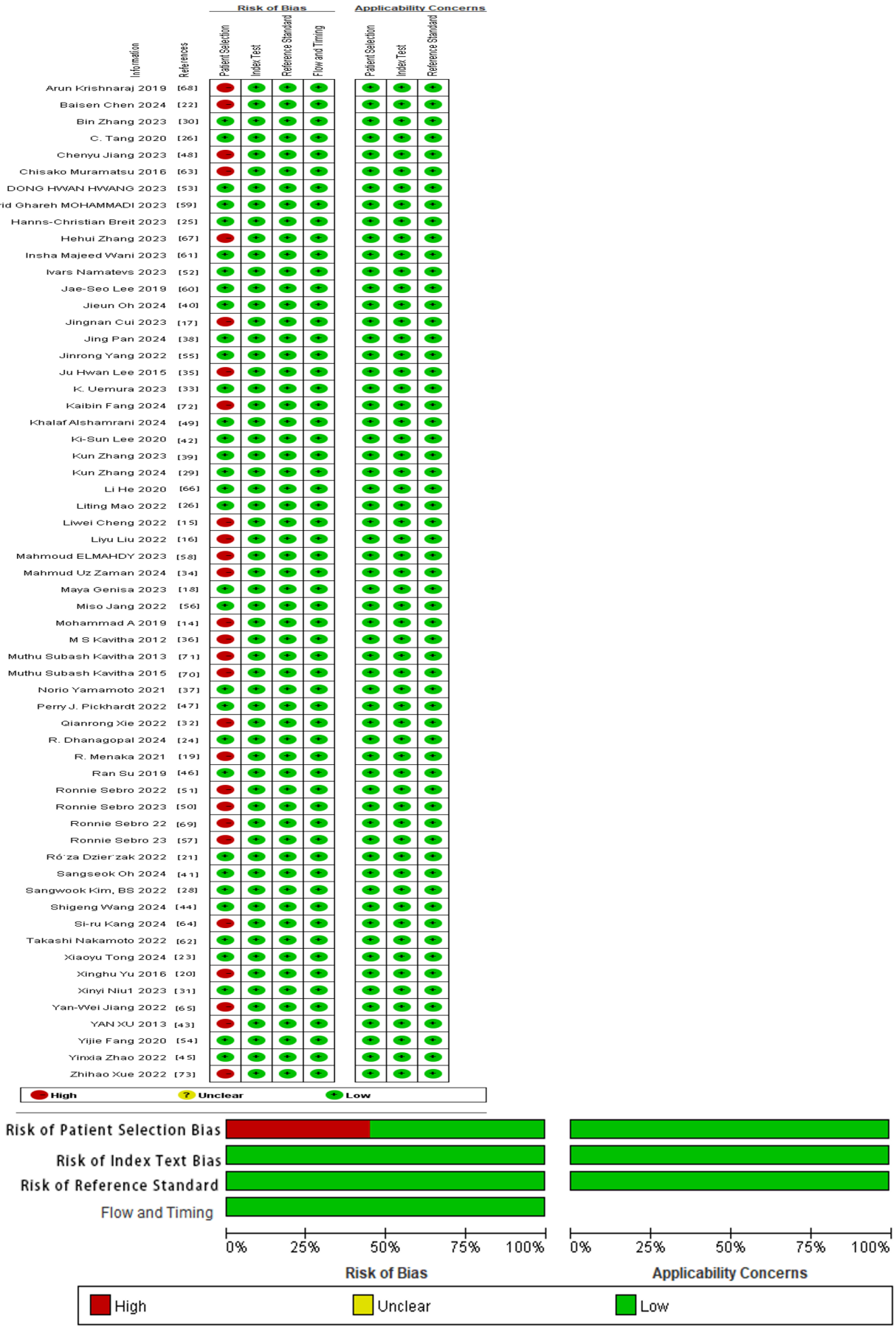

XSL•FO
RenderX

Meta-Analysis

ML Based on X-Ray

Synthesized Results

This validation set comprised 24 diagnostic 4-fold tables, which were used to verify the ML models based on x-rays for OP diagnosis. The results were summarized through the bivariate mixed-effects model. The pooled SEN, SPC, PLR, NLR, DOR, and SROC curves were 0.92 (95% CI 0.88 - 0.94), 0.83 (95% CI 0.76 - 0.88), 5.4 (95% CI 3.8 - 7.6), 0.10 (95% CI 0.07 - 0.15), 54 (95% CI 28 - 105), and 0.94 (95% CI 0.92 - 0.96), respectively (Figures 3 and 4) [14,17,20,24,27,28,30,34-37,42,46,49,55,59-63,70,71]. There was no discernible publication bias in the studies according to Deeks' funnel plot (Figure 5). In the included study participants, approximately 48.44% (n=6429) had OP. Assuming this as the prior probability, if the ML prediction result was OP, the actual probability of OP was .83. If the ML prediction result was non-OP, the actual probability of non-OP was .92 (Figure 6).

Figure 3. Forest plot of sensitivity and specificity for x-ray–based machine learning for diagnosing osteoporosis [14,17,20,24,27,28,30,34-36,42,46,49,56,60,62,63,70,71].

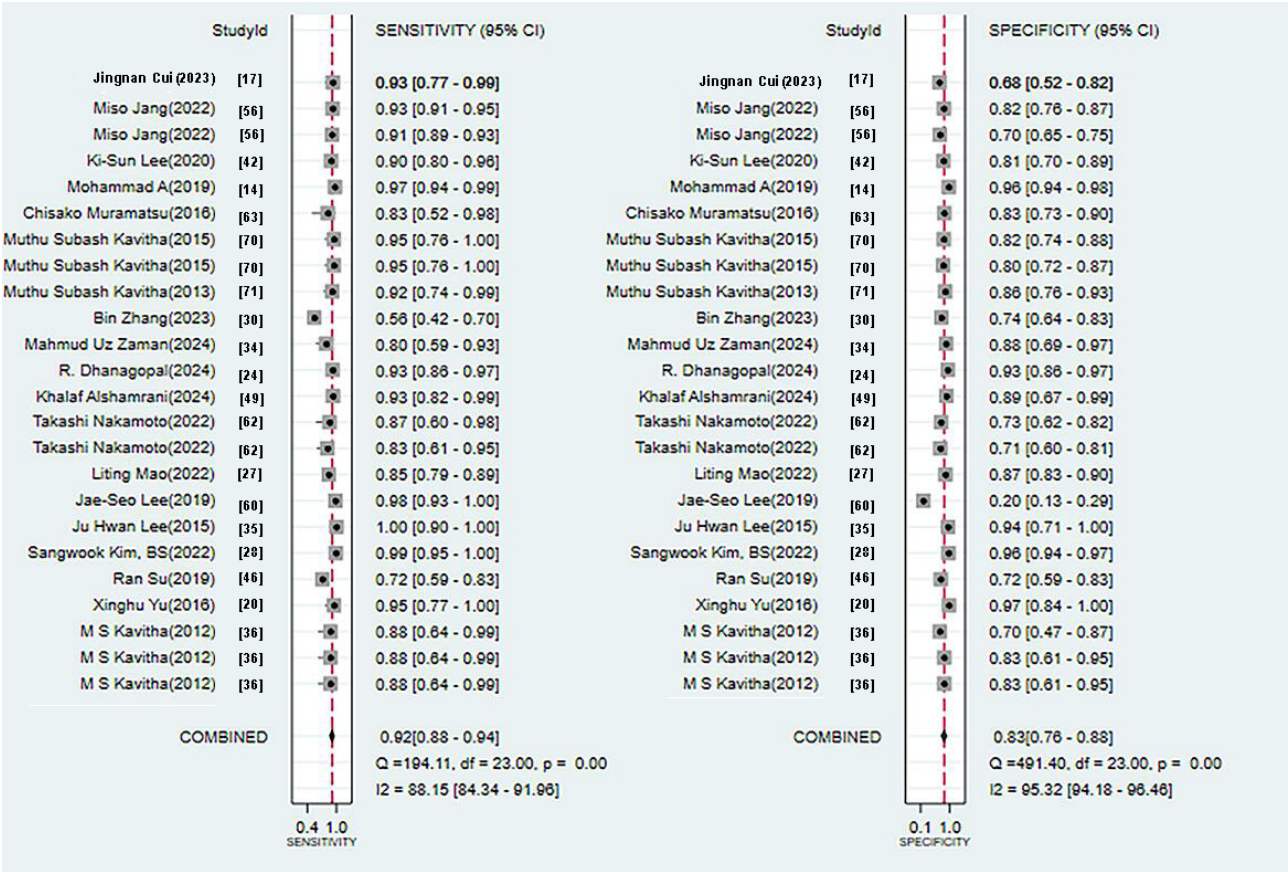

XSL•FO
RenderX

Figure 4. 95% Confidence contour for x-ray. AUC: Area underthe curve;SEN: Sensitivity; SPC: Specificity;SROC: Summary receiver operating characteristic.

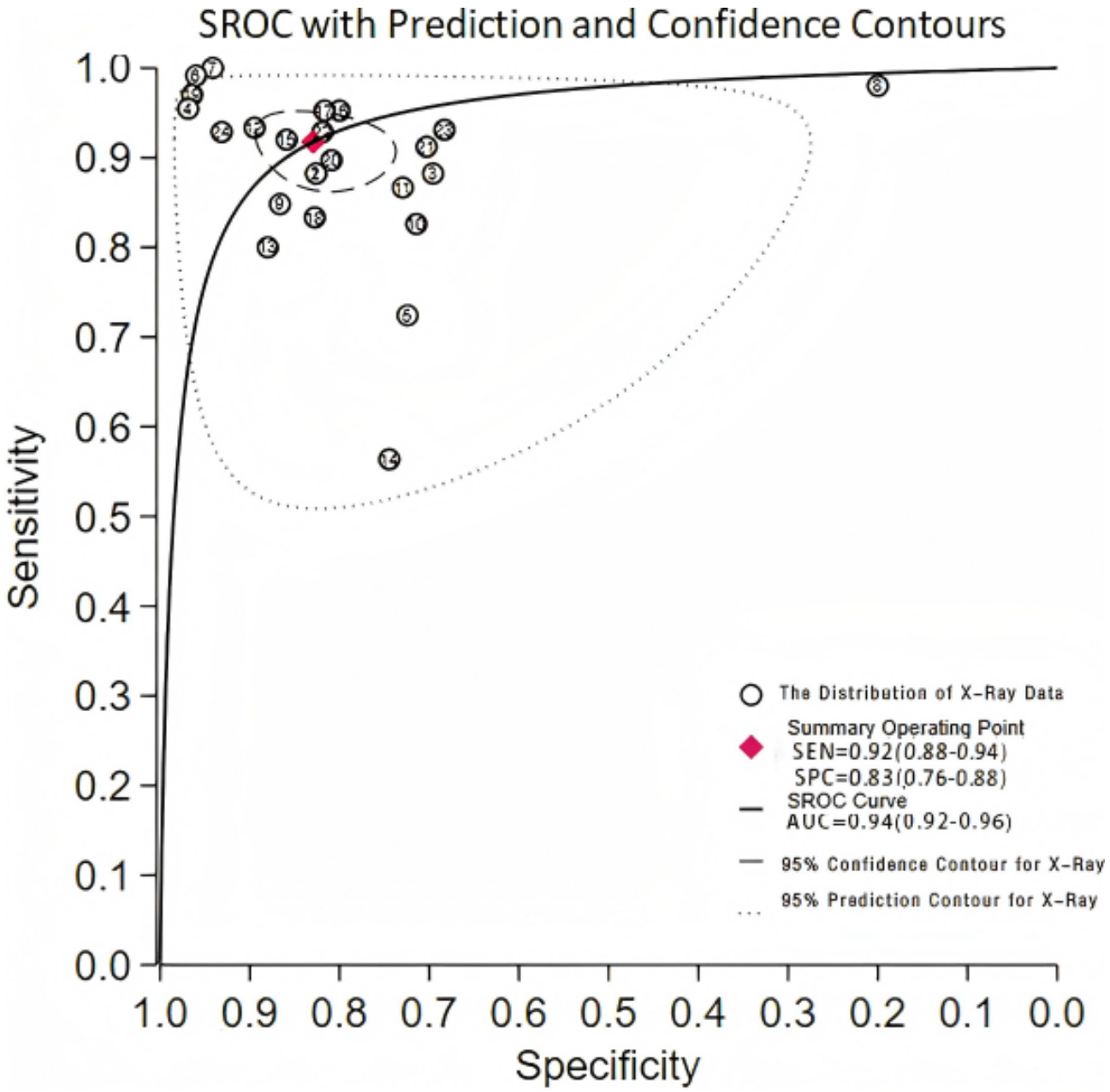

XSL•FO
RenderX

Figure 5. Diagnostic odds ratio for x-ray. ESS:Effective sample size.

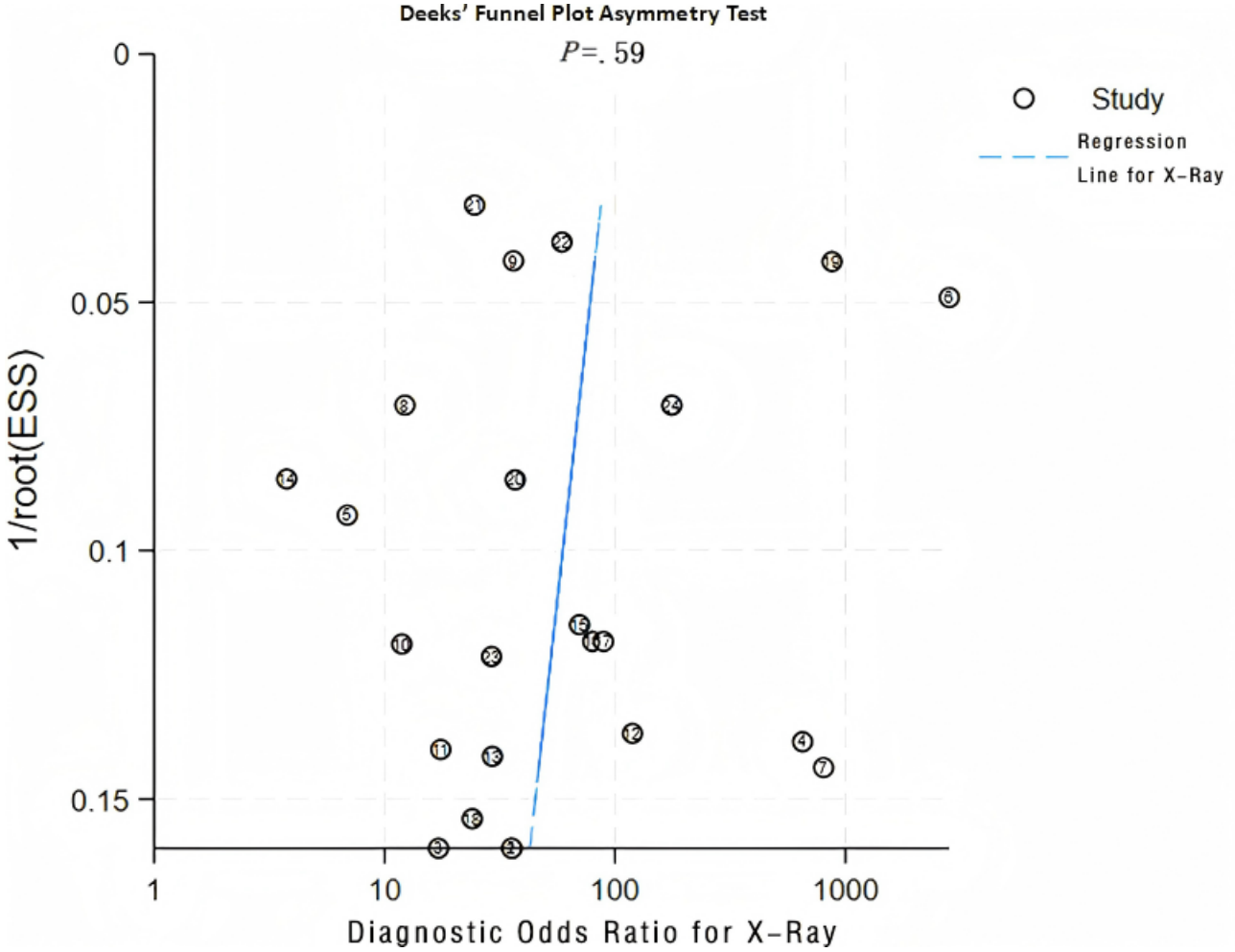

XSL•FO
RenderX

Figure 6. Pretest probability for x-ray. LR:Likelihood ratio.

XSL•FO
RenderX

Subgroup Analysis: Types of ML

Deep Learning

The validation set included 9 diagnostic 4-fold tables to assess the performance of DL models based on x-ray images for OP diagnosis. The results summarized from the bivariate mixed-effects model showed that SEN, SPC, PLR, NLR, DOR, and the SROC curve were 0.90 (95% CI 0.79 - 0.95), 0.79 (95% CI 0.62 - 0.89), 4.2 (95% CI 2.2 - 8.0), 0.13 (95% CI 0.06 - 0.29), 32 (95% CI 9 - 107), and 0.92 (95% CI 0.89 - 0.94), respectively (Figures S1 and S2 in Multimedia Appendix 4). Deeks' funnel plot revealed no marked publication bias (Figure S3 in Multimedia Appendix 4). In the encompassed studies, approximately 30% (n=2556) of the participants had OP. Therefore, assuming this as the prior probability, if the result from ML indicated OP, the actual probability of OP was .64. If the ML result indicated non-OP, the actual probability of non-OP was .95 (Figure S4 in Multimedia Appendix 4).

Traditional ML

The validation set encompassed 15 diagnostic 4-fold tables for validating the traditional ML models based on x-ray imaging for OP diagnosis. The bivariate mixed-effects model was leveraged. The pooled SEN, SPC, PLR, NLR, DOR, and SROC were 0.93 (95% CI 0.92 - 0.95), 0.85 (95% CI 0.79 - 0.89), 6.0 (95% CI 4.3 - 8.5), 0.08 (95% CI 0.06 - 0.10), 78 (95% CI 44 - 139), and 0.96 (95% CI 0.94 - 0.97), respectively (Figures S5 and S6 in Multimedia Appendix 4). Deeks' funnel plot showed no significant publication bias in studies (Figure S7 in Multimedia Appendix 4). Approximately 61% (n=3863) of the participants had OP. Therefore, assuming this as the prior probability, if the result from ML indicated OP, the actual probability of OP was .90. If the ML result indicated non-OP, the actual probability of non-OP was .81 (Figure S8 in Multimedia Appendix 4).

Generation Method of the Validation Set

K-Fold Cross-Validation

Among models constructed through x-ray for OP diagnosis, 7 diagnostic 4-fold tables used the K-fold cross-validation to generate the validation set. The results summarized by the bivariate mixed-effects model demonstrated that the SEN, SPC, PLR, NLR, DOR, and SROC curve were 0.90 (95% CI 0.83 - 0.95), 0.87 (95% CI 0.79 - 0.93), 7.2 (95% CI 4.0 - 12.7), 0.11 (95% CI 0.06 - 0.21), 64 (95% CI 20 - 204), and 0.95 (95% CI 0.93 - 0.96), respectively (Figures S9 and S10 in Multimedia Appendix 4). Deeks' funnel plot did not exhibit significant publication bias (Figure S11 in Multimedia Appendix 4). Among the participants in our included studies, approximately 42% (n=1287) had OP. Therefore, assuming this as the prior probability, if the ML result indicated OP, the actual probability of OP was .84. If the ML result indicated non-OP, the actual probability of non-OP was .92 (Figure S12 in Multimedia Appendix 4).

Random Sampling

In total, 14 diagnostic 4-fold tables used the random sampling method to generate the validation set. The results summarized by the bivariate mixed-effects model showed that the pooled SEN, SPC, PLR, NLR, DOR, and SROC curve were 0.90 (95% CI 0.84 - 0.93), 0.76 (95% CI 0.67 - 0.84), 3.8 (95% CI 2.7 - 5.4), 0.14 (95% CI 0.09 - 0.20), 28 (95% CI 16 - 48), and 0.91 (95% CI 0.88 - 0.93), respectively (Figures S13 and S14 in Multimedia Appendix 4). Significant publication bias was not noted in Deeks' funnel plot (Figure S15 in Multimedia Appendix 4). Among the participants in our included studies, approximately 60% (n=4049) had OP. Therefore, assuming this as the prior probability, if the ML result indicated OP, the actual probability of OP was .85. If the ML result showed non-OP, the actual probability of non-OP was .83 (Figure S16 in Multimedia Appendix 4).

Examination Parts

Limbs

In the OP diagnosis models constructed based on x-rays, 4 diagnostic 4-fold tables focused on the limb bones. The results summarized by the bivariate mixed-effects model showed a SEN of 0.97 (95% CI 0.83 - 0.99), SPC of 0.90 (95% CI 0.75 - 0.96), PLR of 9.6 (95% CI 3.5 - 25.9), NLR of 0.03 (95% CI 0.01 - 0.22), DOR of 277 (95% CI 20 - 3783), and the SROC curve of 0.98 (95% CI 0.96 - 0.99; Figures S17 and S18 in Multimedia Appendix 4). Deeks' funnel plot did not demonstrate significant publication bias (Figure S19 in Multimedia Appendix 4). In the included study participants, the proportion of OP cases was approximately 19% (n=1114). Assuming this as the prior probability, if the ML result indicated OP, the actual probability of OP was .69. If the ML result indicated non-OP, the actual probability of non-OP was <.001 (Figure S20 in Multimedia Appendix 4).

Mandible

In total, 6 diagnostic 4-fold tables focused on the mandible. The bivariate mixed-effects model was used. The pooled SEN, SPC, PLR, NLR, DOR, and SROC were 0.94 (95% CI 0.89 - 0.97), 0.80 (95% CI 0.56 - 0.93), 4.8 (95% CI 1.9 - 12.1), 0.07 (95% CI 0.04 - 0.14), 69 (95% CI 20 - 241), and 0.96 (95% CI 0.94 - 0.97), respectively (Figures S21 and S22 in Multimedia Appendix 4). Deeks' funnel plot indicated no significant publication bias (Figure S23 in Multimedia Appendix 4). In all included study participants, the proportion of OP cases was approximately 42% (n=1153). Assuming this as the prior probability, if the ML result indicated OP, the actual probability of OP was .78. If the ML result indicated non-OP, the actual probability of non-OP was .95 (Figure S24 in Multimedia Appendix 4).

Lumbar Vertebrae

In total, 8 diagnostic 4-fold tables focused on the lumbar vertebrae. The bivariate mixed-effects model was used to summarize data. The pooled SEN, SPC, PLR, NLR, DOR, and SROC were 0.87 (95% CI 0.77 - 0.93), 0.82 (95% CI 0.75 - 0.87), 4.8 (95% CI 3.4 - 6.7), 0.16 (95% CI 0.08 - 0.30), 31 (95% CI 12 - 77), and 0.90 (95% CI 0.87 - 0.92), respectively (Figures S25 and S26 in Multimedia Appendix 4). Significant publication bias was not found in Deeks' funnel plot (Figure S27 in Multimedia Appendix 4). In the included study participants, the proportion of OP cases was approximately 32% (n=1281). Assuming this as the prior probability, if the ML

XSL•FO
RenderX

result indicated OP, the actual probability of having OP was .69. If the ML result indicated non-OP, the actual probability of non-OP was .93 (Figure S28 in Multimedia Appendix 4).

ML Based on CT

Synthesized Results

The validation set consisted of 24 diagnostic 4-fold tables for validating CT-based ML models for diagnosing OP. The bivariate mixed-effects model was used to pool data. The pooled SEN, SPC, PLR, NLR, DOR, and SROC were 0.91 (95% CI 0.89 - 0.93), 0.92 (95% CI 0.89 - 0.94), 11.6 (95% CI 8.5 - 9.7), 0.09 (95% CI 0.07 - 0.12), 123 (95% CI 80 - 90), and 0.97 (95% CI 0.53 - 1.00), respectively (Figures 7 and 8) [15,16,18,19,21-23,25,26,29,31-33,38-41,43-45,47,48,50-54,56-58,65,67-69,72,73]. According to Deeks' funnel plot, there was no significant publication bias (Figure 9). Among the included research participants, the proportion of individuals with OP was approximately 50% (n=10,995). Therefore, assuming this as the prior probability, if the ML models predicted OP, the actual probability of OP was .92. If the ML models predicted no OP, the actual probability of non-OP was .91 (Figure 10).

Figure 7. Forestplot of sensitivity and specificity for computed tomography-based machine learning for diagnosing osteoporosis [15,16,18,19,21-23,25,26,29,31-33,38-41,43-45,47,48,50-54,56-58,65,67-69,72,73].

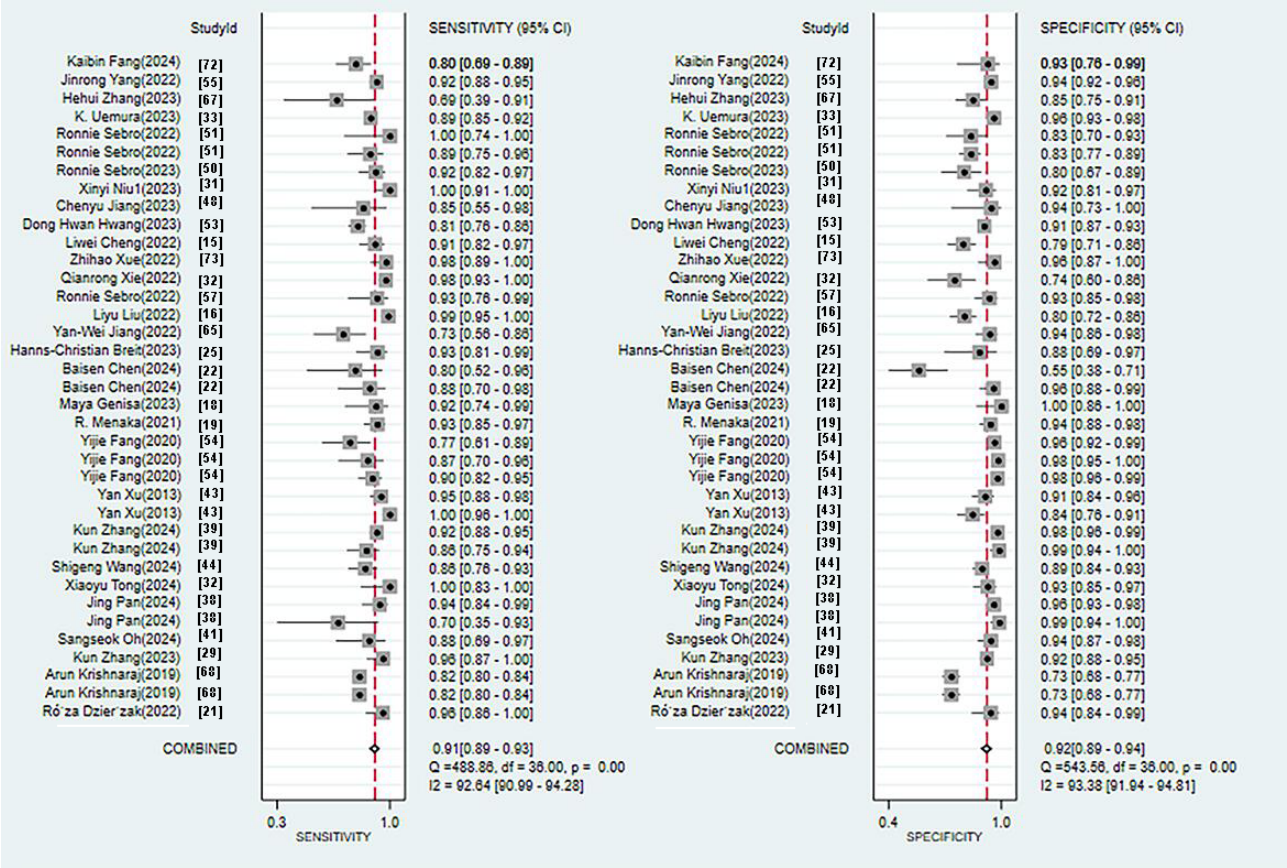

XSL•FO
RenderX

Figure 8. 95% Confidence contour for CT.AUC:Area under the curve;CT:Computed tomography;SEN: Sensitivity; SPC: Specificity;SROC: Summary receiver operating characteristic.

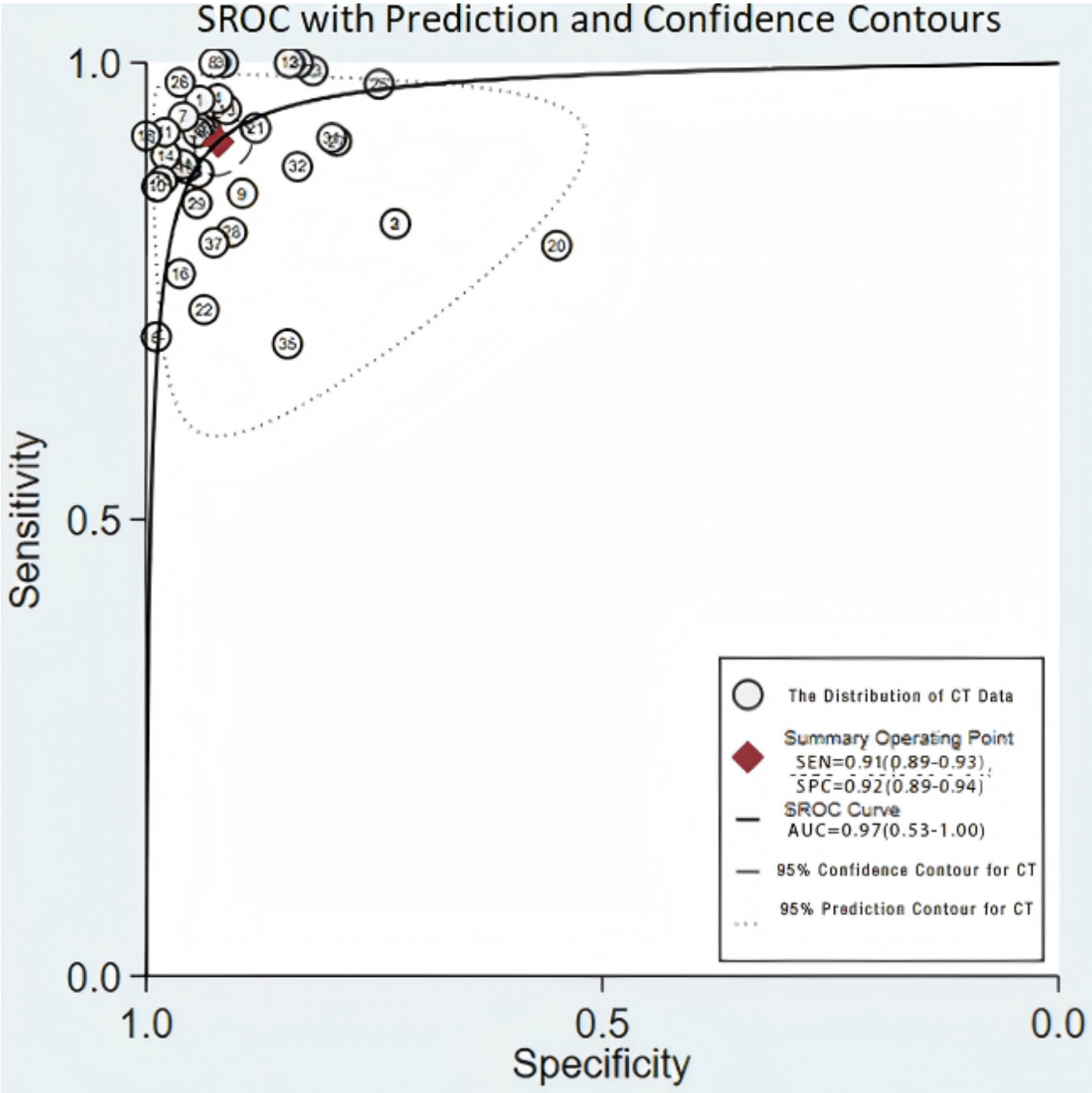

XSL•FO
RenderX

Figure 9. Diagnostic odds ratio for CT .CT: Computed tomography;ESS:Effective sample size.

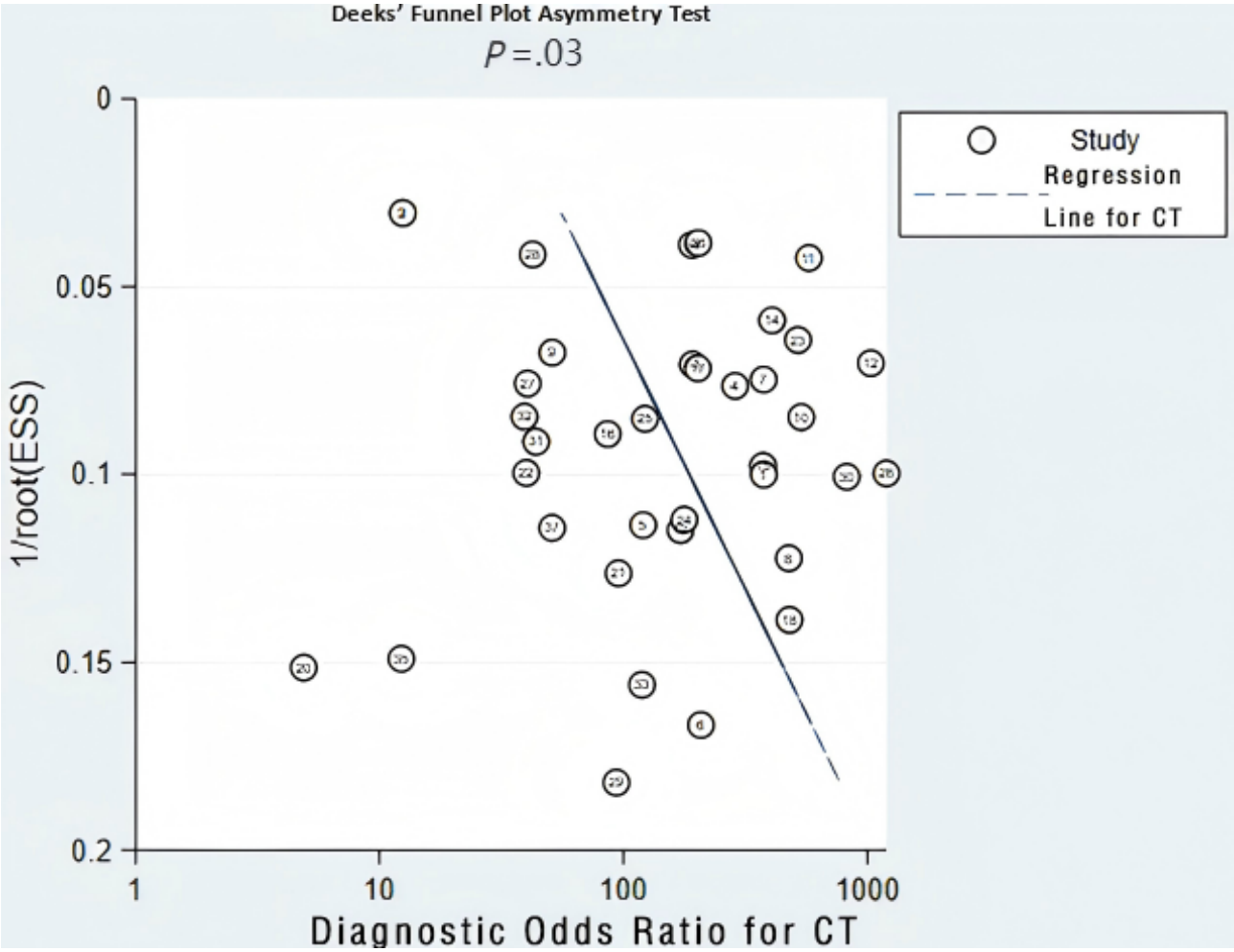

XSL•FO
RenderX

Figure 10. Pretest probability for CT. CT:Computed tomography;LR:Likelihood ratio.

XSL•FO
RenderX

Subgroup Analysis: Types of ML

Deep Learning

In the validation set, there were 15 diagnostic 4-fold tables for validating the CT-based DL models for diagnosing OP. The bivariate mixed-effects model was used. The pooled SEN, SPC, PLR, NLR, DOR, and SROC curve were 0.91 (95% CI 0.88 - 0.94), 0.94 (95% CI 0.92 - 0.96), 16.3 (95% CI 11.9 - 22.3), 0.09 (95% CI 0.06 - 0.13), 178 (95% CI 106 - 299), and 0.98 (95% CI 0.96 - 0.99), respectively (Figures S29 and S30 in Multimedia Appendix 4). Deeks' funnel plot exhibited no marked publication bias (Figure S31 in Multimedia Appendix 4). Among the research participants included, the proportion of individuals with OP was approximately 32% (n=3197). Therefore, assuming this as the prior probability, if the ML models predicted OP, the actual probability of OP was .88. If the ML models predicted no OP, the actual probability of non-OP was .96 (Figure S32 in Multimedia Appendix 4).

Traditional ML

In the validation set, there were 15 diagnostic 4-fold tables for validating traditional ML models based on CT for diagnosing OP. The bivariate mixed-effects model was used. The pooled SEN, SPC, PLR, NLR, DOR, and SROC curve were 0.92 (95% CI 0.88 - 0.95), 0.85 (95% CI 0.77 - 0.90), 6.1 (95% CI 4.0 - 9.4), 0.09 (95% CI 0.06 - 0.15), 67 (95% CI 35 - 128), and 0.95 (95% CI 0.93 - 0.97), respectively (Figures S33 and S34 in Multimedia Appendix 4). Deeks' funnel plot did not show notable publication bias (Figure S35 in Multimedia Appendix 4). Among the research participants, the proportion of individuals with OP was approximately 60% (n=6486). Therefore, assuming this as the prior probability, if the ML models predicted OP, the actual probability of OP was .90. If the ML models predicted no OP, the actual probability of non-OP was .88 (Figure S36 in Multimedia Appendix 4).

Validation Set Generation Method

External Validation

In the OP diagnosis models constructed based on CT, validation sets for 5 diagnostic 4-fold tables were generated through external validation. The bivariate mixed-effects model was leveraged to pool data. The pooled SEN, SPC, PLR, NLR, DOR, and SROC curve were 0.88 (95% CI 0.85 - 0.91), 0.97 (95% CI 0.96 - 0.98), 28.4 (95% CI 20.4 - 39.7), 0.12 (95% CI 0.10 - 0.16), 229 (95% CI 148 - 355), and 0.98 (95% CI 0.96 - 0.99), respectively (Figures S37 and S38 in Multimedia Appendix 4). Deeks' funnel plot indicated no discernible publication bias (Figure S39 in Multimedia Appendix 4). Among the included study participants, approximately 31% (n=1590) had OP. Assuming this as the prior probability, if ML predicted OP, the actual probability of OP was .93. Conversely, if ML predicted non-OP, the actual probability of non-OP was .95 (Figure S40 in Multimedia Appendix 4).

Random Sampling

Validation sets for 24 diagnostic 4-fold tables were generated using the random sampling method. The bivariate mixed-effects model was leveraged. The pooled SEN, SPC, PLR, NLR, DOR, and SROC were 0.91 (95% CI 0.87 - 0.94), 0.90 (95% CI 0.85 - 0.94), 9.4 (95% CI 6.0 - 14.9), 0.10 (95% CI 0.07 - 0.14), 96 (95% CI 57 - 161), and 0.96 (95% CI 0.94 - 0.97), respectively (Figures S41 and S42 in Multimedia Appendix 4). Deeks' funnel plot presented no significant publication bias (Figure S43 in Multimedia Appendix 4). Among the included study participants, approximately 36% (n=4175) had OP. Given this as the prior probability, when the ML models predicted OP, the actual probability of OP was .84. On the other hand, when the ML models predicted non-OP, the actual probability of non-OP was .95 (Figure S44 in Multimedia Appendix 4).

Examination Parts

Hip Joint

In the OP diagnostic models constructed based on CT, 6 diagnostic 4-fold tables focused on the hip joint. The bivariate mixed-effects model was used. The pooled SEN, SPC, PLR, NLR, DOR, and SROC were 0.87 (95% CI 0.83 - 0.90), 0.92 (95% CI 0.81 - 0.96), 10.4 (95% CI 4.4 - 24.7), 0.14 (95% CI 0.10 - 0.19), 76 (95% CI 24 - 239), and 0.92 (95% CI 0.90 - 0.94), respectively (Figures S45 and S46 in Multimedia Appendix 4). Deeks' funnel plot did not show any marked publication bias (Figure S47 in Multimedia Appendix 4). Among the included study participants, approximately 69% (n=2719) had OP. Assuming this as the prior probability, if ML predicted OP, the actual probability of OP was .96. If the models predicted non-OP, the actual probability of OP was .77 (Figure S48 in Multimedia Appendix 4).

Thoracic Vertebrae

In total, 9 diagnostic 4-fold tables focused on the thoracic vertebrae. The bivariate mixed-effects model was leveraged to pool data. The pooled SEN, SPC, PLR, NLR, DOR, and SROC were 0.91 (95% CI 0.86 - 0.94), 0.94 (95% CI 0.92 - 0.95), 14.4 (95% CI 10.7 - 19.3), 0.10 (95% CI 0.06 - 0.15), 150 (95% CI 75 - 300), and 0.97 (95% CI 0.95 - 0.98), respectively (Figures S49 and S50 in Multimedia Appendix 4). Significant publication bias was not observed in Deeks' funnel plot (Figure S51 in Multimedia Appendix 4). Among the encompassed study participants, approximately 29% (n=2523) had OP. Assuming this as the prior probability, if ML predicted OP, the actual probability of OP was .85. If ML predicted non-OP, the actual probability of non-OP was .96 (Figure S52 in Multimedia Appendix 4).

Lumbar Vertebrae

For OP diagnostic models using the lumbar vertebrae as the target part, 26 diagnostic 4-fold tables were analyzed. The bivariate mixed-effects model yielded a SEN of 0.91 (95% CI 0.87 - 0.94), SPC of 0.92 (95% CI 0.86 - 0.95), PLR of 10.7 (95% CI 6.7 - 17.2), NLR of 0.10 (95% CI 0.07 - 0.14), DOR of 110 (95% CI 63 - 191), and SROC curve of 0.96 (95% CI 0.94 - 0.98; Figures S53 and S54 in Multimedia Appendix 4). Deeks' funnel plot did not reflect discernible publication bias (Figure S55 in Multimedia Appendix 4). Among the encompassed study participants, approximately 42% (n=7327) had OP. Assuming this as the prior probability, if ML predicted OP, the probability of actual OP was .89. Conversely, if ML predicted non-OP, the actual likelihood of non-OP was .93 (Figure S56 in Multimedia Appendix 4).

XSL•FO
RenderX

ML Based on MRI

Only 3 studies have constructed diagnostic models for OP based on MRI [45,64,66], all of which used the lumbar vertebrae as the examination part. Due to the limited number of studies of this type and the substantial heterogeneity noted in the meta-analysis, the conclusions drawn lack sufficient reference significance. Therefore, this study presents only a narrative analysis of this part.

Among these, 2 studies used traditional ML models, while 1 used a DL model. The SEN of these models was 0.857, 0.872, and 0.892, and the SPC was 0.944, 0.688, and 0.892, respectively. The validation strategies used in these studies included external validation, K-fold cross-validation, and random sampling.

Discussion

Main Findings of This Study

Medical imaging is an indispensable tool in the diagnosis, treatment, and management of OP. Conventional imaging methods such as x-ray, CT, and MRI are pivotal clinically.

X-ray imaging enables clinicians to visually assess reductions in vertebral height, cortical bone thickness, and morphological changes in appendicular and mandibular bones, thus screening OP. However, as DXA is updated and improved, clinicians can more accurately know bone mineral density and structural parameters of the lumbar vertebrae and hip, thereby facilitating the diagnosis of OP. The World Health Organization has designated DXA as the gold standard for determining bone mineral density and diagnosing postmenopausal OP [74,75]. However, in low-resource environments and economically underdeveloped regions, the clinical application of DXA is limited due to factors such as insufficient medical knowledge and constrained health care infrastructure. In contrast, AI tools have the potential to maximize the extraction of clinically relevant information from various medical images, thereby enabling the early identification of the population with OP or low bone mass. This significantly supports the early prevention, diagnosis, and management of the disease.

Advantages of Different Imaging Modalities in the Diagnosis of OP

This shift has diminished the application of x-rays in quantitative analysis for OP. Nevertheless, ML and DL models have improved the diagnostic performance of x-ray imaging, providing significant impetus for its broader clinical application. CT, with its high resolution, enables clinicians to observe cortical and trabecular bone integrity, offering distinct advantages in evaluating spinal OP and changes in trabecular bone volume ratios in the hip [76]. Conventional CT generates images by measuring differences in the linear attenuation coefficients of x-ray beams as they pass through various biological tissues. However, when tissues possess similar densities, such as calcium and bone, conventional CT often yields comparable Hounsfield unit values due to the use of a single x-ray energy spectrum, limiting its ability to differentiate between such tissues. In contrast, spectral CT imaging, which is based on tissue-specific photoelectric effect weighting, offers enhanced resolution in distinguishing fine bone microarchitecture. This technological advancement holds significant potential for improving the diagnostic accuracy of OP. MRI is highly efficient in assessing bone microarchitecture [77]. However, MRI is not the first choice to detect OP because of its high cost, extended scan times, and obstacles faced by patients with metallic implants or claustrophobia. Our database search corroborated that most studies have focused on x-ray and CT imaging, while comparatively fewer have investigated MRI. Nevertheless, existing evidence supports the robust diagnostic performance of ML models based on imaging data. For example, the pooled SEN and SPC of ML models based on x-ray for OP diagnosis were 0.92 (95% CI 0.88 - 0.94) and 0.83 (95% CI 0.76 - 0.88), respectively. Similarly, ML models developed via CT achieved SEN and SPC of 0.91 (95% CI 0.89 - 0.93) and 0.92 (95% CI 0.89 - 0.94), respectively. These findings demonstrate the high accuracy of x-ray and CT in OP diagnosis. In addition, quantitative ultrasound is another commonly used modality for OP detection. Quantitative ultrasound relies on 2 primary parameters: speed of sound and broadband ultrasound attenuation, which assess the ability of ultrasound waves to propagate through bone both horizontally and longitudinally [78]. In summary, diverse imaging modalities and bone types provide flexible and enriched diagnostic options for OP. Furthermore, this variety brings ample opportunities for the development of advanced ML models tailored to different imaging techniques.

Status Quo of Research on ML

With advances in computer science, numerous researchers have sought to use these techniques in the prevention and treatment of OP. Compared with clinicians, who visually observe positive imaging features, AI-assisted tools significantly improve the efficiency and accuracy of diagnosing OP [46,60]. In addition, Yang et al [79] developed an ML-based predictive model using data from surveys on risk factors for OP, which is highly prospective for early screening and treating OP in the Hong Kong population. Similarly, ML models based on community health examinations and serum bone turnover markers have demonstrated a high area under the receiver operating characteristic curve, F_1-scores, and accuracy [80,81]. These findings highlight the efficiency of ML in the diagnosis and management of OP.

Mechanism of Image-Based ML

Image-based ML can broadly be categorized into traditional ML and DL. Traditional ML involves dividing data into a training set for model development and a test set for model validation. Through processes such as image segmentation, texture extraction, and feature selection, traditional ML models are constructed for predicting outcome events. However, the process of texture feature extraction and selection carries a significant risk of data loss. In contrast, DL incorporates feature extraction directly into the training process, thereby maximizing the retention of meaningful information within the image data. Convolutional neural networks, as a representative DL approach, can simultaneously extract and select features across multiple hidden layers to accomplish classification tasks. Moreover, DL-based models can correct image blurring in panoramic

XSL•FO
RenderX

x-rays caused by patient mispositioning and mitigate the impact of metal artifacts in CT images on feature extraction [82-84]. This study further demonstrates that ML models based on x-ray and CT outperform traditional ML models, suggesting that DL is more accurate than traditional ML approaches. Image analysis using DL can leverage AI to develop more efficient and user-friendly image interpretation tools, providing valuable insights into the development of medical imaging software.

The Impact of Validation Set Generation Methods on ML Performance

Validation methods are critical metrics for assessing the performance of ML models. These methods can be categorized into external validation and internal validation. Internal validation can be further subdivided into random sampling, leave-one-out validation, and K-fold cross-validation. External validation, which can accurately reflect the clinical applicability of ML, is widely preferred by researchers. In contrast, internal validation typically generates validation sets via random methods, which inherently carries a risk of similarity in features and distribution trends between the validation and training sets. This issue is prominent in image-based studies, where the application of ML is restricted in medical research due to the high similarity in images and parameters between internal validation sets. Although external validation offers a superior means of assessing model performance, conducting such validation requires access to independent research cohorts and often entails consideration of factors such as periods, geographical regions, populations, and health care institutions. These requirements inevitably lead to substantial increases in both the time and financial costs of research. This perspective provides an objective explanation for the limited external validation in this study.

Advantages and Limitations

This study is the first to summarize the evidence of the application of ML based on various imaging modalities in the diagnosis of OP. This study provides theoretical support for the subsequent development of clinical scoring systems and medical software. However, our research has the following limitations: first, despite a substantial number of included studies, only a small number of studies on MRI were encompassed in view of the practicability in clinical work. Therefore, in future research, our emphasis will be put on meta-analyses involving MRI studies, aiming to evaluate the utility of ML in the diagnosis of OP through medical imaging. As a result, only a narrative review was performed, without a direct evaluation of its diagnostic performance. Most of the included studies rely primarily on internal validation, with insufficient external validation, which imposes certain limitations on the interpretability and generalizability of our findings. This study encompassed only English publications, with the majority of research originating from countries where AI is more widely applied. In addition, the external validation conducted in this study was limited, constituting an objective constraint that may have influenced the outcomes of the meta-analysis. Future studies will endeavor to comprehensively incorporate globally available literature to enhance the authority and generalizability of the conclusions.

Conclusions

Image-based ML, particularly DL based on x-ray and CT images, is highly accurate in the diagnosis of OP. Future focus should be placed on developing AI-based software to expand its clinical applicability and enhance diagnostic precision.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (82274551) and the Guangdong Provincial Fundamental and Applied Basic Research Fund Program (2023B1515230001). All authors of this study hereby solemnly declare that no forms of artificial intelligence tools were used at any stage of the research process to ensure the originality and independence of this manuscript.

Data Availability

The datasets used and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

RZ was responsible for the formal analysis, data curation, and original manuscript writing of all the experiments. HY was responsible for software operation and data management. Y Li and XL were responsible for the formal analysis. ZY was in charge of the methodology. Y Lin and JH were responsible for the methodology and data validation. LW was in charge of project management. HH was responsible for project design, project management, and fund acquisition. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Literature search strategy.
[DOCX File, 25 KB - jmir_v28i1e75965_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Supplementary data sheet.
[XLSX File, 20 KB - jmir_v28i1e75965_app2.xlsx]

Multimedia Appendix 3
Quality Assessment of Diagnostic Accuracy Studies assessment process for included studies.
[XLSX File, 15 KB - jmir_v28i1e75965_app3.xlsx]

Multimedia Appendix 4
Supplementary materials.
[DOC File, 56031 KB - jmir_v28i1e75965_app4.doc]

Checklist 1
PRISMA checklist.
[PDF File, 99 KB - jmir_v28i1e75965_app5.pdf]

References

1. Consensus development conference: diagnosis, prophylaxis, and treatment of osteoporosis. Am J Med 1993 Jun;94(6):646-650. [doi: 10.1016/0002-9343(93)90218-E]
2. Wang Y, Tao Y, Hyman ME, Li J, Chen Y. Osteoporosis in china. Osteoporos Int 2009 Oct;20(10):1651-1662. [doi: 10.1007/s00198-009-0925-y] [Medline: 19415374]
3. Clynes MA, Harvey NC, Curtis EM, Fuggle NR, Dennison EM, Cooper C. The epidemiology of osteoporosis. Br Med Bull 2020 May 15;133(1):105-117. [doi: 10.1093/bmb/ldaa005] [Medline: 32282039]
4. Assessment of fracture risk and its application to screening for postmenopausal osteoporosis. Report of a WHO Study Group. World Health Organ Tech Rep Ser 1994;843(1-129):1-129. [Medline: 7941614]
5. Kanis JA, Cooper C, Rizzoli R, Reginster JY, on behalf of the Scientific Advisory Board of the European Society for Clinical and Economic Aspects of Osteoporosis (ESCEO) and the Committees of Scientific Advisors and National Societies of the International Osteoporosis Foundation (IOF). European guidance for the diagnosis and management of osteoporosis in postmenopausal women. Osteoporos Int 2019 Jan 18;30(1):3-44. [doi: 10.1007/s00198-018-4704-5]
6. Erickson BJ. Basic artificial intelligence techniques: machine learning and deep learning. Radiol Clin North Am 2021 Nov;59(6):933-940. [doi: 10.1016/j.rcl.2021.06.004] [Medline: 34689878]
7. Chan HP, Samala RK, Hadjiiski LM, Zhou C. Deep learning in medical image analysis. Adv Exp Med Biol 2020;1213(3-21):3-21. [doi: 10.1007/978-3-030-33128-3_1] [Medline: 32030660]
8. Oliveira JS, Franco FO, Revers MC, et al. Computer-aided autism diagnosis based on visual attention models using eye tracking. Sci Rep 2021 May 12;11(1):10131. [doi: 10.1038/s41598-021-89023-8] [Medline: 33980874]
9. Khan M, Shah PM, Khan IA, et al. IoMT-enabled computer-aided diagnosis of pulmonary embolism from computed tomography scans using deep learning. Sensors (Basel) 2023 Jan 28;23(3):1471. [doi: 10.3390/s23031471] [Medline: 36772510]
10. Loizidou K, Skouroumouni G, Nikolaou C, Pitris C. A review of computer-aided breast cancer diagnosis using sequential mammograms. Tomography 2022 Dec 6;8(6):2874-2892. [doi: 10.3390/tomography8060241] [Medline: 36548533]
11. Ceranka J, Wuts J, Chiabai O, Lecouvet F, Vandemeulebroucke J. Computer-aided diagnosis of skeletal metastases in multi-parametric whole-body MRI. Comput Methods Programs Biomed 2023 Dec;242:107811. [doi: 10.1016/j.cmpb.2023.107811] [Medline: 37742486]
12. Hong N, Cho SW, Shin S, et al. Deep-learning-based detection of vertebral fracture and osteoporosis using lateral spine X-ray radiography. J Bone Miner Res 2020 Dec 1;38(6):887-895. [doi: 10.1002/jbmr.4814]
13. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
14. Alzubaidi MA, Otoom M. A comprehensive study on feature types for osteoporosis classification in dental panoramic radiographs. Comput Methods Programs Biomed 2020 May;188(105301):105301. [doi: 10.1016/j.cmpb.2019.105301] [Medline: 31911333]
15. Cheng L, Cai F, Xu M, Liu P, Liao J, Zong S. A diagnostic approach integrated multimodal radiomics with machine learning models based on lumbar spine CT and X-ray for osteoporosis. J Bone Miner Metab 2023 Nov;41(6):877-889. [doi: 10.1007/s00774-023-01469-0] [Medline: 37898574]
16. Liu L, Si M, Ma H, et al. A hierarchical opportunistic screening model for osteoporosis using machine learning applied to clinical data and CT images. BMC Bioinformatics 2022 Feb 10;23(1):63. [doi: 10.1186/s12859-022-04596-z] [Medline: 35144529]

XSL•FO
RenderX

17. Cui J, Liu CL, Jennane R, Ai S, Dai K, Tsai TY. A highly generalized classifier for osteoporosis radiography based on multiscale fractal, lacunarity, and entropy distributions. Front Bioeng Biotechnol 2023;11:1054991. [doi: 10.3389/fbioe.2023.1054991] [Medline: 37274169]
18. Genisa M, Abdullah JY, Yusoff BM, Arief EM, Hermana M, Utomo CP. Adopting signal processing technique for osteoporosis detection based on CT scan image. Appl Sci (Basel) 2023;13(8):5094. [doi: 10.3390/app13085094]
19. Menaka R, Ramesh R, Dhanagopal R. Aggregation of region-based and boundary-based knowledge biased segmentation for osteoporosis detection from X-ray, dual X-ray and CT images. Curr Med Imaging 2021;17(2):288-295. [doi: 10.2174/1573405616999200730175526] [Medline: 32748751]
20. Yu X, Ye C, Xiang L. Application of artificial neural network in the diagnostic system of osteoporosis. Neurocomputing 2016 Nov;214:376-381. [doi: 10.1016/j.neucom.2016.06.023]
21. Dzierżak R, Omiotek Z. Application of deep convolutional neural networks in the diagnosis of osteoporosis. Sensors (Basel) 2022 Oct 26;22(21):8189. [doi: 10.3390/s22218189] [Medline: 36365886]
22. Chen B, Cui J, Li C, et al. Application of radiomics model based on lumbar computed tomography in diagnosis of elderly osteoporosis. J Orthop Res 2024 Jun;42(6):1356-1368. [doi: 10.1002/jor.25789] [Medline: 38245854]
23. Tong X, Wang S, Zhang J, Fan Y, Liu Y, Wei W. Automatic osteoporosis screening system using radiomics and deep learning from low-dose chest CT images. Bioengineering (Basel) 2024 Jan 2;11(1):50. [doi: 10.3390/bioengineering11010050] [Medline: 38247927]
24. Dhanagopal R, Menaka R, Suresh Kumar R, Vasanth Raj PT, Debrah EL, Pradeep K. Channel-boosted and transfer learning convolutional neural network-based osteoporosis detection from CT scan, dual X-ray, and X-ray images. J Healthc Eng 2024;2024(3733705):3733705. [doi: 10.1155/2024/3733705] [Medline: 38223259]
25. Breit HC, Varga-Szemes A, Schoepf UJ, et al. CNN-based evaluation of bone density improves diagnostic performance to detect osteopenia and osteoporosis in patients with non-contrast chest CT examinations. Eur J Radiol 2023 Apr;161:110728. [doi: 10.1016/j.ejrad.2023.110728] [Medline: 36773426]
26. Tang C, Zhang W, Li H, et al. CNN-based qualitative detection of bone mineral density via diagnostic CT slices for osteoporosis screening. Osteoporos Int 2021 May;32(5):971-979. [doi: 10.1007/s00198-020-05673-w]
27. Mao L, Xia Z, Pan L, et al. Deep learning for screening primary osteopenia and osteoporosis using spine radiographs and patient clinical covariates in a Chinese population. Front Endocrinol (Lausanne) 2022;13(971877):971877. [doi: 10.3389/fendo.2022.971877] [Medline: 36176468]
28. Kim S, Kim BR, Chae HD, et al. Deep radiomics-based approach to the diagnosis of osteoporosis using hip radiographs. Radiol Artif Intell 2022 Jul;4(4):e210212. [doi: 10.1148/ryai.210212] [Medline: 35923378]
29. Zhang K, Lin PC, Pan J, et al. DeepmdQCT: a multitask network with domain invariant features and comprehensive attention mechanism for quantitative computer tomography diagnosis of osteoporosis. Comput Biol Med 2024 Mar;170:107916. [doi: 10.1016/j.compbiomed.2023.107916] [Medline: 38237237]
30. Zhang B, Chen Z, Yan R, et al. Development and validation of a feature-based broad-learning system for opportunistic osteoporosis screening using lumbar spine radiographs. Acad Radiol 2024 Jan;31(1):84-92. [doi: 10.1016/j.acra.2023.07.002]
31. Niu X, Huang Y, Li X, et al. Development and validation of a fully automated system using deep learning for opportunistic osteoporosis screening using low-dose computed tomography scans. Quant Imaging Med Surg 2023 Aug 1;13(8):5294-5305. [doi: 10.21037/qims-22-1438] [Medline: 37581046]
32. Xie Q, Chen Y, Hu Y, et al. Development and validation of a machine learning-derived radiomics model for diagnosis of osteoporosis and osteopenia using quantitative computed tomography. BMC Med Imaging 2022 Aug 8;22(1):140. [doi: 10.1186/s12880-022-00868-5] [Medline: 35941568]
33. Uemura K, Otake Y, Takashima K, et al. Development and validation of an open-source tool for opportunistic screening of osteoporosis from hip CT images. Bone Joint Res 2023 Sep 20;12(9):590-597. [doi: 10.1302/2046-3758.129.BJR-2023-0115.R1] [Medline: 37728034]
34. Zaman MU, Alam MK, Alqhtani NR, et al. RETRACTED ARTICLE: Diagnosing osteoporosis using deep neural networkassisted optical image processing method. Opt Quant Electron 2024 Mar;56(3). [doi: 10.1007/s11082-023-06031-w]
35. Lee JH, Hwang YN, Park SY, Jeong JH, Kim SM. Diagnosis of osteoporosis by quantification of trabecular microarchitectures from hip radiographs using artificial neural networks. J Comp Theo Nano 2015 Jul 1;12(7):1115-1120. [doi: 10.1166/jctn.2015.3859]
36. Kavitha MS, Asano A, Taguchi A, Kurita T, Sanada M. Diagnosis of osteoporosis from dental panoramic radiographs using the support vector machine method in a computer-aided system. BMC Med Imaging 2012 Jan 16;12(1):22248480. [doi: 10.1186/1471-2342-12-1] [Medline: 22248480]
37. Yamamoto N, Sukegawa S, Yamashita K, et al. Effect of patient clinical variables in osteoporosis classification using hip X-rays in deep learning analysis. Medicina (Kaunas) 2021 Aug 20;57(8):846. [doi: 10.3390/medicina57080846] [Medline: 34441052]
38. Pan J, Lin PC, Gong SC, et al. Effectiveness of opportunistic osteoporosis screening on chest CT using the DCNN model. BMC Musculoskelet Disord 2024 Feb 27;25(1):176. [doi: 10.1186/s12891-024-07297-1] [Medline: 38413868]
39. Zhang K, Lin P, Pan J, et al. End to end multitask joint learning model for osteoporosis classification in CT images. Comput Intell Neurosci 2023;2023:3018320. [doi: 10.1155/2023/3018320] [Medline: 36970245]

XSL•FO
RenderX

40. Oh J, Kim B, Oh G, Hwangbo Y, Ye JC. End-to-end semi-supervised opportunistic osteoporosis screening using computed tomography. Endocrinol Metab (Seoul) 2024 Jun;39(3):500-510. [doi: 10.3803/EnM.2023.1860] [Medline: 38721637]
41. Oh S, Kang WY, Park H, et al. Evaluation of deep learning-based quantitative computed tomography for opportunistic osteoporosis screening. Sci Rep 2024 Jan 5;14(1):363. [doi: 10.1038/s41598-023-45824-7] [Medline: 38182616]
42. Lee KS, Jung SK, Ryu JJ, Shin SW, Choi J. Evaluation of transfer learning with deep convolutional neural networks for screening osteoporosis in dental panoramic radiographs. J Clin Med 2020 Feb 1;9(2):392. [doi: 10.3390/jcm9020392] [Medline: 32024114]
43. Xu Y, Li D, Chen Q, Fan Y. Full supervised learning for osteoporosis diagnosis using micro - CT images. Microsc Res Tech 2013 Apr;76(4):333-341. [doi: 10.1002/jemt.22171]
44. Wang S, Tong X, Cheng Q, et al. Fully automated deep learning system for osteoporosis screening using chest computed tomography images. Quant Imaging Med Surg 2024 Apr 3;14(4):2816-2827. [doi: 10.21037/qims-23-1617] [Medline: 38617137]
45. Zhao Y, Zhao T, Chen S, et al. Fully automated radiomic screening pipeline for osteoporosis and abnormal bone density with a deep learning-based segmentation using a short lumbar mDixon sequence. Quant Imaging Med Surg 2022 Feb;12(2):1198-1213. [doi: 10.21037/qims-21-587] [Medline: 35111616]
46. Su R, Liu T, Sun C, Jin Q, Jennane R, Wei L. Fusing convolutional neural network features with hand-crafted features for osteoporosis diagnoses. Neurocomputing 2020 Apr;385:300-309. [doi: 10.1016/j.neucom.2019.12.083]
47. Pickhardt PJ, Nguyen T, Perez AA, et al. Improved CT-based osteoporosis assessment with a fully automated deep learning tool. Radiol Artif Intell 2022 Sep;4(5):e220042. [doi: 10.1148/ryai.220042] [Medline: 36204542]
48. Jiang C, Jin D, Ni M, Zhang Y, Yuan H. Influence of image reconstruction kernel on computed tomography-based finite element analysis in the clinical opportunistic screening of osteoporosis—a preliminary result. Front Endocrinol 2023;14(1076990):36936156. [doi: 10.3389/fendo.2023.1076990]
49. Alshamrani K, Alshamrani HA. Lossless compression-based detection of osteoporosis using bone X-ray imaging. J Xray Sci Technol 2024;32(2):475-491. [doi: 10.3233/XST-230238] [Medline: 38393881]
50. Sebro R, Elmahdy M. Machine learning for opportunistic screening for osteoporosis and osteopenia using knee CT scans. Can Assoc Radiol J 2023 Nov;74(4):676-687. [doi: 10.1177/08465371231164743] [Medline: 36960893]
51. Sebro R, De la Garza-Ramos C. Machine learning for opportunistic screening for osteoporosis from CT scans of the wrist and forearm. Diagnostics (Basel) 2022 Mar 11;12(3):691. [doi: 10.3390/diagnostics12030691] [Medline: 35328244]
52. Namatevs I, Nikulins A, Edelmers E, et al. Modular neural networks for osteoporosis detection in mandibular cone-beam computed tomography scans. Tomography 2023 Sep 22;9(5):1772-1786. [doi: 10.3390/tomography9050141] [Medline: 37888733]
53. Hwang DH, Bak SH, Ha TJ, Kim Y, Kim WJ, Choi HS. Multi-view computed tomography network for osteoporosis classification. IEEE Access 2023;11:22297-22306. [doi: 10.1109/ACCESS.2023.3252361]
54. Fang Y, Li W, Chen X, et al. Opportunistic osteoporosis screening in multi-detector CT images using deep convolutional neural networks. Eur Radiol 2021 Apr;31(4):1831-1842. [doi: 10.1007/s00330-020-07312-8]
55. Yang J, Liao M, Wang Y, et al. Opportunistic osteoporosis screening using chest CT with artificial intelligence. Osteoporos Int 2022 Dec;33(12):2547-2561. [doi: 10.1007/s00198-022-06491-y]
56. Jang M, Kim M, Bae SJ, Lee SH, Koh JM, Kim N. Opportunistic osteoporosis screening using chest radiographs with deep learning: development and external validation with a cohort dataset. J Bone Miner Res 2022 Feb;37(2):369-377. [doi: 10.1002/jbmr.4477] [Medline: 34812546]
57. Sebro R, De la Garza-Ramos C. Opportunistic screening for osteoporosis and osteopenia from CT scans of the abdomen and pelvis using machine learning. Eur Radiol 2023 Mar;33(3):1812-1823. [doi: 10.1007/s00330-022-09136-0] [Medline: 36166085]
58. Elmahdy M, Sebro R. Opportunistic screening for osteoporosis using CT scans of the knee: a pilot study. Stud Health Technol Inform 2023 May 18;302:909-910. [doi: 10.3233/SHTI230305] [Medline: 37203533]
59. Mohammadi FG, Sebro R. Opportunistic screening for osteoporosis using hand radiographs: a preliminary study. Stud Health Technol Inform 2023 May 18;302:911-912. [doi: 10.3233/SHTI230306] [Medline: 37203534]
60. Lee JS, Adhikari S, Liu L, Jeong HG, Kim H, Yoon SJ. Osteoporosis detection in panoramic radiographs using a deep convolutional neural network-based computer-assisted diagnosis system: a preliminary study. Dentomaxillofac Radiol 2019 Jan;48(1):20170344. [doi: 10.1259/dmfr.20170344] [Medline: 30004241]
61. Wani IM, Arora S. Osteoporosis diagnosis in knee X-rays by transfer learning based on convolution neural network. Multimed Tools Appl 2023;82(9):14193-14217. [doi: 10.1007/s11042-022-13911-y] [Medline: 36185321]
62. Nakamoto T, Taguchi A, Kakimoto N. Osteoporosis screening support system from panoramic radiographs using deep learning by convolutional neural network. Dentomaxillofac Radiol 2022 Sep 1;51(6):20220135. [doi: 10.1259/dmfr.20220135] [Medline: 35816516]
63. Muramatsu C, Horiba K, Hayashi T, et al. Quantitative assessment of mandibular cortical erosion on dental panoramic radiographs for screening osteoporosis. Int J Comput Assist Radiol Surg 2016 Nov;11(11):2021-2032. [doi: 10.1007/s11548-016-1438-8] [Medline: 27289239]

XSL•FO
RenderX

64. Kang SR, Wang K. Radiomic nomogram based on lumbar spine magnetic resonance images to diagnose osteoporosis. Acta Radiol 2024 Aug;65(8):950-958. [doi: 10.1177/02841851241242052]
65. Jiang YW, Xu XJ, Wang R, Chen CM. Radiomics analysis based on lumbar spine CT to detect osteoporosis. Eur Radiol 2022 Nov;32(11):8019-8026. [doi: 10.1007/s00330-022-08805-4] [Medline: 35499565]
66. He L, Liu Z, Liu C, et al. Radiomics based on lumbar spine magnetic resonance imaging to detect osteoporosis. Acad Radiol 2021 Jun;28(6):e165-e171. [doi: 10.1016/j.acra.2020.03.046] [Medline: 32386949]
67. Zhang H, Wei W, Qian B, et al. Screening for osteoporosis based on IQon spectral CT virtual low monoenergetic images: comparison with conventional 120 kVp images. Heliyon 2023 Oct;9(10):e20750. [doi: 10.1016/j.heliyon.2023.e20750]
68. Krishnaraj A, Barrett S, Bregman-Amitai O, et al. Simulating dual-energy X-ray absorptiometry in CT using deep-learning segmentation cascade. J Am Coll Radiol 2019 Oct;16(10):1473-1479. [doi: 10.1016/j.jacr.2019.02.033]
69. Sebro R, De la Garza-Ramos C. Support vector machines are superior to principal components analysis for selecting the optimal bones' CT attenuations for opportunistic screening for osteoporosis using CT scans of the foot or ankle. Osteoporos Sarcopenia 2022 Sep;8(3):112-122. [doi: 10.1016/j.afos.2022.09.002] [Medline: 36268496]
70. Kavitha MS, An SY, An CH, et al. Texture analysis of mandibular cortical bone on digital dental panoramic radiographs for the diagnosis of osteoporosis in Korean women. Oral Surg Oral Med Oral Pathol Oral Radiol 2015 Mar;119(3):346-356. [doi: 10.1016/j.oooo.2014.11.009] [Medline: 25600978]
71. Kavitha MS, Asano A, Taguchi A, Heo MS. The combination of a histogram-based clustering algorithm and support vector machine for the diagnosis of osteoporosis. Imaging Sci Dent 2013;43(3):153. [doi: 10.5624/isd.2013.43.3.153]
72. Fang K, Zheng X, Lin X, Dai Z. Unveiling osteoporosis through radiomics analysis of hip CT imaging. Acad Radiol 2024 Mar;31(3):1003-1013. [doi: 10.1016/j.acra.2023.10.009]
73. Xue Z, Huo J, Sun X, et al. Using radiomic features of lumbar spine CT images to differentiate osteoporosis from normal bone density. BMC Musculoskelet Disord 2022 Dec;23(1):35395769. [doi: 10.1186/s12891-022-05309-6]
74. Chen M, Gerges M, Raynor WY, et al. State of the art imaging of osteoporosis. Semin Nucl Med 2024 May;54(3):415-426. [doi: 10.1053/j.semnuclmed.2023.10.008]
75. El Maghraoui A, Roux C. DXA scanning in clinical practice. QJM 2008 Aug;101(8):605-617. [doi: 10.1093/qjmed/hcn022] [Medline: 18334497]
76. Kessenich CR. Diagnostic imaging and biochemical markers of bone turnover. Nurs Clin North Am 2001 Sep;36(3):409-416. [Medline: 11532656]
77. Sollmann N, Löffler MT, Kronthaler S, et al. MRI - based quantitative osteoporosis imaging at the spine and femur. Magn Reson Imaging 2021 Jul;54(1):12-35. [doi: 10.1002/jmri.27260]
78. Oei L, Koromani F, Rivadeneira F, Zillikens MC, Oei EHG. Quantitative imaging methods in osteoporosis. Quant Imaging Med Surg 2016 Dec;6(6):680-698. [doi: 10.21037/qims.2016.12.13] [Medline: 28090446]
79. Yang Q, Cheng H, Qin J, et al. A machine learning-based Preclinical Osteoporosis Screening Tool (POST): model development and validation study. JMIR Aging 2023 Nov 8;6:e46791. [doi: 10.2196/46791] [Medline: 37986117]
80. Baik SM, Kwon HJ, Kim Y, Lee J, Park YH, Park DJ. Machine learning model for osteoporosis diagnosis based on bone turnover markers. Health Informatics J 2024;30(3):39115269. [doi: 10.1177/14604582241270778] [Medline: 39115269]
81. Ou Yang WY, Lai CC, Tsou MT, Hwang LC. Development of machine learning models for prediction of osteoporosis from clinical health examination data. Int J Environ Res Public Health 2021 Jul 18;18(14):7635. [doi: 10.3390/ijerph18147635] [Medline: 34300086]
82. Putra RH, Doi C, Yoda N, Astuti ER, Sasaki K. Current applications and development of artificial intelligence for digital dental radiography. Dentomaxillofac Radiol 2022 Jan 1;51(1):20210197. [doi: 10.1259/dmfr.20210197] [Medline: 34233515]
83. Du X, Chen Y, Zhao J, Xi Y. A convolutional neural network based auto-positioning method for dental arch in rotational panoramic radiography. Annu Int Conf IEEE Eng Med Biol Soc 2018 Jul;2018(2615-8):2615-2618. [doi: 10.1109/EMBC.2018.8512732] [Medline: 30440944]
84. Liang K, Zhang L, Yang H, Yang Y, Chen Z, Xing Y. Metal artifact reduction for practical dental computed tomography by improving interpolation-based reconstruction with deep learning. Med Phys 2019 Dec;46(12):e823-e834. [doi: 10.1002/mp.13644] [Medline: 31811792]

Abbreviations

AI: artificial intelligence
CAD: computer-aided diagnosis
CT: computed tomography
DL: deep learning
DOR: diagnostic odds ratio
DXA: dual-energy x-ray absorptiometry
ML: machine learning
MRI: magnetic resonance imaging
NLR: negative likelihood ratio

XSL•FO
RenderX

OP: osteoporosis
PLR: positive likelihood ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SEN: sensitivity
SPC: specificity
SROC: summary receiver operating characteristic

Edited by J Sarvestan; submitted 14.Apr.2025; peer-reviewed by M He, N Nanthasamroeng; revised version received 03.Jun.2025; accepted 04.Jun.2025; published 16.Jan.2026.

Please cite as:
Zhao R, Yang H, Li Y, Li X, Yang Z, Lin Y, Huang J, Wan L, Huang H
The Diagnostic Value of Image-Based Machine Learning for Osteoporosis: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e75965
URL: https://www.jmir.org/2026/1/e75965
doi:10.2196/75965

© Rui Zhao, Haolin Yang, Yangbo Li, Xiaoyun Li, Zhijie Yang, Yanping Lin, Jiachun Huang, Lei Wan, Hongxing Huang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Application of Mobile Health in Self-Management Among Patients Undergoing Dialysis: Scoping Review

Qin Xu, MNS; Yulin Xu, MNS; Xiaoqin Liu, MNS; Xiaolin Ma, BSN

Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, No. 1095 Jiefang Avenue, Wuhan, Hubei, China

Corresponding Author:
Xiaoqin Liu, MNS
Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, No. 1095 Jiefang Avenue, Wuhan, Hubei, China

Abstract

Background: The incidence of end-stage renal disease continues to rise annually, with dialysis currently serving as the primary replacement therapy. The effectiveness of dialysis treatment and patients' quality of life are highly dependent on their self-management. Mobile health (mHealth), which provides real-time medical support through portable devices, has become an essential tool for assisting patients undergoing dialysis in optimizing their self-management.

Objective: This study aimed to systematically explore the core elements of self-management in patients undergoing dialysis and clarify the primary applications of mHealth, including types of mHealth, relevant theories and models, mHealth-based interventions, and evaluation indicators.

Methods: This study was guided by Arksey and O'Malley's methodology, PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews), and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension). Databases, such as PubMed, Embase, CINAHL, PsycINFO, and Web of Science, were systematically searched from January 2010 until October 2025. The participants included in this study were patients undergoing dialysis, and the study design must incorporate quantitative research. Published protocols, reviews, editorials, conference papers, books, and non-English studies were excluded. The Mixed Methods Appraisal Tool was used to evaluate the quality of the included studies. Quantitative studies were extracted, mapped, and summarized. The results were collated and synthesized using a structured spreadsheet.

Results: Out of 1483 relevant studies, this scoping review ultimately selected 34 studies involving 2068 patients undergoing dialysis. Self-management among patients undergoing dialysis in this study included 6 major areas, including self-monitoring, diet and fluid management, medication management, disease-related knowledge, exercise management, and psychological management. Most studies used a single app (n=22) for management of patients undergoing dialysis, followed by 2 or more online interventions (n=6) and a remote patient monitoring system (n=3). The mHealth-based interventions in this study focused on self-monitoring, dietary and fluid management, and medication management. The transtheoretical model and stages of change (n=5), self-efficacy theory (n=4), and social cognitive theory (n=4) were the most commonly used theories. Among the evaluation indicators, interdialytic weight gain (n=12), serum potassium (n=14), serum phosphorus (n=20), and serum albumin (n=14) were the most commonly used objective indicators. Subjective indicators were assessed using scales, primarily covering adherence (n=17), self-efficacy (n=14), quality of life (n=12), knowledge (n=9), and diet and nutrition (n=9).

Conclusions: Although mHealth holds promise for improving self-management and outcomes among patients undergoing dialysis, there remains significant room for advancement. Future research in this field should focus on enhancing adaptive software development, deeply integrating artificial intelligence technologies, addressing the needs of special populations, and establishing a standardized self-management evaluation system. Our findings not only provide a theoretical framework for optimizing clinical management strategies for patients undergoing dialysis but also offer targeted guidance and practical insights for the subsequent development of apps.

(J Med Internet Res 2026;28:e76880) doi:10.2196/76880

KEYWORDS

kidney; self-management; mHealth; hemodialysis; peritoneal dialysis; mobile health

Introduction

End-stage renal disease (ESRD) refers to the end stage of various chronic kidney diseases (CKDs). The global prevalence of patients with renal failure receiving dialysis treatment continues to rise, with the latest estimate reaching 823 per million population [1,2]. Although kidney transplantation is the treatment of choice for patients with ESRD, the majority of

XSL•FO
RenderX

patients still rely on dialysis due to the shortage of donor kidneys [3]. Hemodialysis and peritoneal dialysis (PD) are the 2 most common types of dialysis. Although hemodialysis and PD have significantly improved survival rates among patients with ESRD [1,4], the invasive and long-term treatments also substantially increase the risk of dialysis-related complications or infections. Studies have shown that comorbidities, such as hypertension, diabetes mellitus, hyperkalemia, and hyperphosphatemia, and cardiovascular diseases are common in patients undergoing dialysis [1,5]. The quality of life and survival rate of patients undergoing dialysis also decline with increasing dialysis duration [5,6]. The quality of life and survival outcomes of patients undergoing dialysis are closely related to the quality of dialysis treatment, which in turn is directly dependent on the level of the patient's self-management [7].

Self-management encompasses multiple aspects of health management. According to Lorig et al [8], self-management involved medical management (special dietary adherence and medication adherence), role management, and emotion management. In patients undergoing dialysis, self-management refers to whether the patients perform self-monitoring, strict control of diet (sodium, potassium, phosphorus, and other micronutrients) and fluid intake, regular medication administration, and prevention and management of complications. Patients undergoing dialysis can reduce negative symptoms and improve the quality of life through self-management behaviors [7].

Disease management of patients undergoing dialysis is a difficult point in the current medical work. The results of several studies have shown that the overall self-management level of patients undergoing dialysis was low [9,10]. Patients undergoing dialysis generally lacked knowledge of the disease, and their self-management behaviors, such as dietary control, fluid intake, and treatment adherence, fell short of standards [10]. The proposal of mobile health (mHealth) provides new ideas and methods for the remote management of patients undergoing dialysis. mHealth is a medical and public health service initiative based on mobile communications technology delivered through mobile phones, monitoring devices, personal digital assistant devices, and other wireless devices [11]. Because of its unique convenience, it has significant advantages in monitoring diseases, controlling symptoms, and promoting healthy behaviors.

There is a growing body of research on the role of mHealth in improving self-management in patients undergoing dialysis [12-15]. Most studies concentrate on developing mobile apps specifically designed for patients undergoing dialysis, that is, specialized software tools running on mobile devices. Additionally, telephones and SMS text messaging are also commonly used tools. Despite the growing interest in mHealth, evidence on mHealth-based self-management among patients undergoing dialysis remains limited. Therefore, this review aims to provide an overview of the use of mHealth in the self-management of patients undergoing dialysis, examine existing interventions, and summarize existing evaluation tools. The goal is to empower patients undergoing dialysis through mHealth, improve their self-management to enhance prognosis, and provide a practical reference for subsequent app development.

Methods

Overview

A scoping review based on the 5-stage methodological framework of Arksey and O'Malley [16], involving (1) identifying the research question, (2) identifying relevant studies, (3) study selection, (4) charting the data, and (5) collating, summarizing, and reporting the results [16]. The PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) [17] and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) guidelines [18] (Checklist 1) were used as the protocol for this study. The study quality assessment was conducted using the 2018 version of the Mixed Methods Appraisal Tool, a tool specifically designed to evaluate the quality of qualitative, quantitative, and mixed methods studies [19].

Stage 1: Identifying the Research Questions

The study population was adult patients on dialysis (both hemodialysis and PD), and the type of intervention was a service using mHealth. The primary objective of this review was to explore the use of mHealth in the self-management of patients undergoing dialysis. The research questions were based on an initial literature search and were refined during discussions in the research team.

In accordance with the overall objectives of this review, we have refined the research questions as follows:

1. What are the types of mHealth in the included studies?
2. What are the attributes of self-management in the included studies?
3. What are the interventions for mHealth-based self-management?
4. What are the evaluation tools for self-management in the included studies?

Stage 2: Identifying Relevant Studies

A systematic literature search was conducted in PubMed (National Center for Biotechnology Information), Embase (Ovid), CINAHL (EBSCOhost), PsycINFO (EBSCOhost), and Web of Science (Clarivate Analytics) to identify studies relevant to the research objectives. We limited the search to studies that were published from January 2010 to October 2025. First, we identified the search strategy and search terms through group discussion. Then, a presearch was conducted in the database using the search strategy and search terms, and the search strategy and search terms were adjusted according to the search results. Subsequently, formal searches were conducted in 5 databases using the identified search strategies and search terms. Two keywords "mHealth" and "dialysis" were used in combination to cover the 2 main concepts of the research question. The specific search strategy and updated search methods can be found in Multimedia Appendix 1. The reference lists of all eligible studies were examined to identify any potentially relevant studies.

XSL•FO
RenderX

Stage 3: Study Selection

Included studies were required to fulfill the following criteria. (1) Participants: adult patients (age≥18 years) receiving long-term dialysis with no gender restrictions; (2) Concept: studies were included if they addressed mHealth and self-management. Self-management is the focus of this scoping review. Among patients undergoing dialysis, we defined self-management as knowledge, diet and fluids, dialysis treatments, medications, dialysis access, exercise, and psychology. mHealth refers to health and medical services (including remote monitoring, health education, online counseling, etc) that are delivered using mobile devices (eg, smartphones and tablets); (3) Context: dialysis occurs at home or in a hospital. Eligible study designs must include quantitative research. We excluded published research protocols, reviews, editorials, conference papers, books, and non-English studies. Finally, if the full text cannot be obtained, the study will be excluded.

EndNote (Clarivate Analytics) software was used to identify duplicates and manage literature. The literature was screened by 2 trained reviewers (QX and YX). In the first stage, 2 reviewers independently reviewed the titles and abstracts of studies based on inclusion and exclusion criteria. Then, the 2 reviewers continued to independently screen the full text. In the second stage, 2 reviewers assessed the quality of the included literature based on the Mixed Methods Appraisal Tool. When 2 reviewers disagree, a third reviewer (Xiaoqin Liu) will join the discussion until all reviewers reach consensus, ensuring the rigor of the selection process.

Stage 4: Charting the Data

The research team worked together to develop a data chart to guide the extraction of key information from each study. Descriptive chart information includes (1) a general description of the study, such as first author and year, country, study design, patient population, and purpose of the study; and (2) intervention-specific information, including type of mHealth, primary function, method of implementation, intervention time, and evaluation tools.

Stage 5: Collating, Summarizing, and Reporting the Results

The research team summarized the data iteratively. Descriptive analyses were used to summarize the types of mHealth and self-management evaluation tools, while thematic content analyses were used to summarize the attributes of self-management and the content of mHealth-based self-management. First, codes were developed and applied to analyze the data. Coded segments of the data chart were then created with color-coded quotations, and the coding results were summarized in an Excel (Microsoft Corp) sheet. The Excel sheet was sorted by code and density. Key themes were extracted by analyzing the studies in an overall iterative comparison.

Ethical Considerations

Ethical approval was not needed for this review.

Results

Basic Characteristics of the Included Studies

We retrieved 1483 records from PsychlNFO (n=67), Web of Science (n=516), PubMed (n=478), CINAHL (n=171), and Embase (n=251). In total, 359 studies were excluded due to duplication, and 913 studies were excluded after reading the title and abstract. There were still 211 studies that needed to be read in full. After reading the full text, a total of 34 studies [12-15,20-49] were included in this scoping review (Figure 1).

A total of 34 studies [12-15,20-49] were included in this study, involving 2068 patients undergoing dialysis. Among the included studies, 26 studies [12,13,20-26,28-44] involved patients undergoing hemodialysis, 6 studies [14,15,27,45-47] focused on patients undergoing PD, and 2 studies [48,49] encompassed patients undergoing both hemodialysis and PD. In terms of the regional distribution of the included studies, most of the studies were concentrated in Asia and North America. Among the 34 studies, Korea was dominated with 8 (23.53%) studies [13,15,23,27-29,35,46], followed by 5 (14.71%) studies [25,38,40,44,48] in the United States, 3 (8.82%) studies [12,24,47] in China, 3 (8.82%) studies [21,31,39] in Iran, 3 (8.82%) studies [26,30,49] in Australia, and 3 (8.82%) studies [14,37,42] in Thailand. Around 2 (5.88%) studies [20,22] in the Netherlands, 2 (5.88%) studies [32,45] in Japan, and 2 (5.88%) studies [34,43] in Indonesia each contributed to 2 studies, while 1 (2.94%) study [41] in Malaysia, 1 (2.94%) study [33] in Turkey, and 1 (2.94%) study [36] in Brazil each contributed 1 study. The studies were published mainly between 2019 and 2025, with a total of 31 studies [12-15,20-24,26-31,33-47,49]. Three studies were published in 2011, 2013, and 2017 (Figure 2A). The results of the quality assessment indicated that most studies demonstrated good quality. Specific details of the quality assessment can be found in Multimedia Appendix 2. Table 1 provides an overview of the key characteristics of the included studies. Additionally, Multimedia Appendix 3 presents the intervention type, core intervention contents, and evaluation indicators for each study.

XSL•FO
RenderX

Figure 1. Flowchart outlining the search process for studies across databases, following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines.

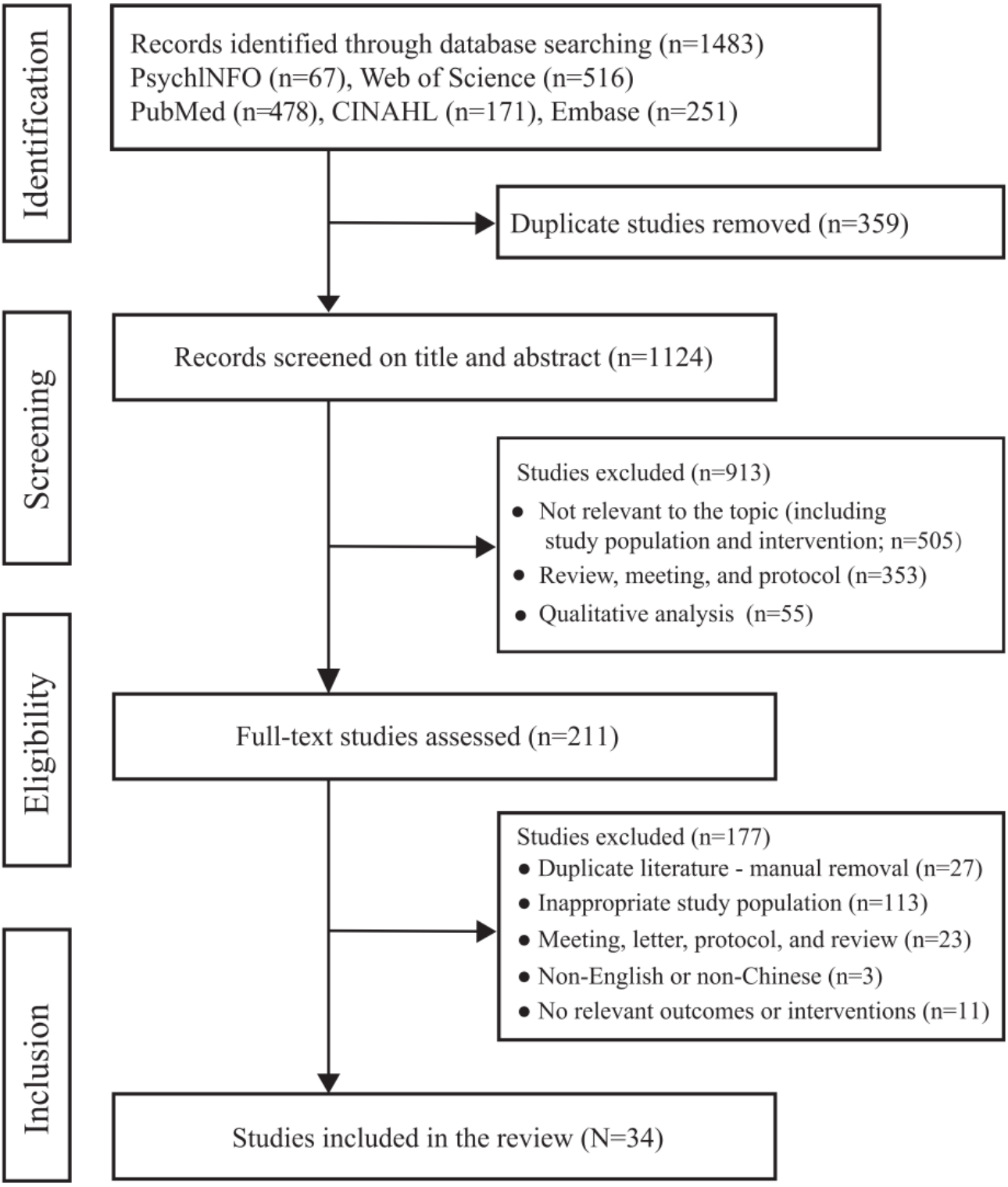

XSL•FO
RenderX

Figure 2. (A) The trend analysis of publication dates in the included studies. (B) The specific information on theories and models used to guide the content of mHealth self-management interventions. HD: hemodialysis; mHealth: mobile health; PD: peritoneal dialysis.

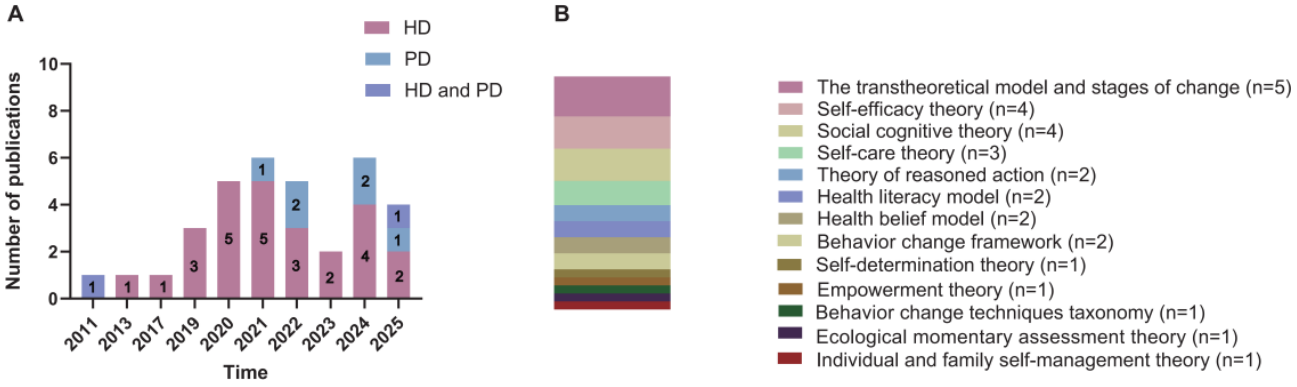

XSL•FO
RenderX

Table . Summary of basic characteristics of the included studies.

Author and year	Country	Aim of the study	Study design	Population	Total sample (experimental/control group)
Saadatifar et al [39], 2022	Iran	To explore the impact of mHealth[a] on treatment adherence in patients undergoing HD[b].	Quasi-experimental study	HD	80 (40/40)
Ren et al [24], 2019	China	To explore the impact of a WeChat-based health education program on the self-management of patients undergoing HD.	A longitudinal experimental intervention study	HD	85 (49/36)
Park and Kim [23], 2019	South Korea	To evaluate the impact of a program based on app and SMS text messaging for patients undergoing HD.	Quasi-experimental study	HD	84 (42/42)
Pack and Lee [35], 2021	South Korea	To develop a mobile app for dietary management and evaluate the impact on patients undergoing HD.	RCT[c]	HD	75 (37/38)
Fakih El Khoury et al [22], 2020	Netherlands	To evaluate the efficacy of the app-based dietary intervention for patients undergoing HD.	A pilot study: a before-and-after study	HD	23
Hanifi et al [31], 2019	Iran	To assess the impact of counseling and follow-up phone calls on patients undergoing HD.	Quasi-experimental study	HD	86 (43/43)
Cho and Park [28], 2020	South Korea	To assess the impact of a tablet-based self-management program on patients undergoing HD.	Quasi-experimental study	HD	46 (23/23)
Chiang et al [12], 2021	China	To evaluate whether the Assisted Care Program in the app can help patients better control their dietary phosphorus intake.	Quasi-experimental study	HD	60 (30/30)
Zwi et al [26], 2022	Australia	To evaluate the feasibility of the app and its impact on patients undergoing HD.	A mixed methods study	HD	61
Welch et al [25], 2013	United States	To evaluate the impact of mobile programs on diet and fluid intake in patients undergoing HD.	A pilot study	HD	33 (16/17)
Thongsunti et al [42], 2024	Thailand	To evaluate the effectiveness of telemedicine-based management of hyperphosphatemia in patients undergoing HD.	RCT	HD	80 (40/40)

XSL•FO
RenderX

Author and year	Country	Aim of the study	Study design	Population	Total sample (experimental/control group)
Chung et al [29], 2024	South Korea	To assess the impact of adaptive nutrition and education on patients undergoing HD using real electronic medical record data.	A decentralized clinical trial	HD	153 (42/40/34/37)[d]
Dawson et al [30], 2021	Australia	To assess the impact of SMS text messaging on patients undergoing HD.	A randomized feasibility study	HD	115 (78/37)
Fakih El Khoury et al [20], 2021	Netherlands	To assess the efficacy of an intervention using the app on phosphorus.	A pilot study: a before-and-after study	HD	23
Hosseini et al [21], 2023	Iran	To determine the effect of an app on self-efficacy and self-care of patients undergoing HD.	Quasi-experimental study (longitudinal single-group study)	HD	54
Min and Park [13], 2020	South Korea	To assess the impact of a mobile app–based self-management support program on older adults undergoing HD.	Quasi-experimental study	HD	56 (28/28)
Mollaoğlu et al [33], 2024	Turkey	To evaluate the impact of education and art therapy through a telemedicine approach in patients undergoing HD.	RCT	HD	60 (30/30)
Rocco et al [38], 2023	United States	To evaluate the impact of an app on self-monitoring of daily fluids in patients undergoing HD.	A pilot study: a before-and-after study	HD	18
St-Jules et al [40], 2021	United States	To assess the feasibility and acceptability of mHealth for managing hyperphosphatemia in patients undergoing HD.	A feasibility trial	HD	40 (13/14/13)[e]
Teong et al [41], 2022	Malaysia	To evaluate the effectiveness of an app for nutritional management in patients undergoing HD.	RCT	HD	66 (33/33)
Pungchompoo et al [37], 2024	Thailand	To evaluate the impact of a home telemedicine model on older patients undergoing HD.	A mixed methods study	HD	54 (24/30)
Nursalam et al [34], 2020	Indonesia	To evaluate the impact of the app on improving fluid restriction adherence in patients undergoing HD.	A mixed methods study	HD	60 (30/30)
Hayashil et al [32], 2017	Japan	To evaluate the usefulness of the self-management support system for self-monitoring in patients undergoing HD.	A pilot study	HD	18 (8/10)

XSL•FO
RenderX

Author and year	Country	Aim of the study	Study design	Population	Total sample (experimental/control group)
Pinto et al [36], 2020	Brazil	To evaluate the impact of the app on fluid restriction and dietary control in patients undergoing HD.	A randomized, single-center, self-controlled study	HD	48
Andriati [43], 2025	Indonesia	To evaluate the impact of the app on adherence and renal function of patients undergoing HD.	Quasi-experimental study	HD	55
Taguiam [44], 2025	United States	To evaluate the impact of the app on fluid intake management and body weight in patients undergoing HD.	A mixed methods study	HD	18
Lee and Kang [15], 2024	South Korea	Using a mobile instant messaging tool to customize diet plans for patients undergoing PD[f] and evaluate outcomes.	Quasi-experimental study	PD	43 (21/22)
Chae and Kim [27], 2024	South Korea	To develop the app for improved self-management and evaluate its impact on patients undergoing PD.	RCT	PD	53 (27/26)
Uchiyama et al [45], 2022	Japan	To assess the impact of using a remote patient monitoring system on patients undergoing PD.	A randomized crossover controlled trial	PD	15
Jung et al [46], 2021	South Korea	To evaluate the impact of remote patient monitoring on automated patients undergoing PD.	RCT	PD	50 (28/22)
Lukkanalikitkul et al [14], 2022	Thailand	To evaluate the availability and impact on patients undergoing PD for app.	User-centered design study	PD	9
Zeng et al [47], 2025	China	To evaluate the effectiveness of a PD management system in improving adherence and clinical outcomes.	A retrospective cohort study	PD	127
Stark et al [48], 2011	United States	To assess the effectiveness of a PDA[g]-based app for dietary management in patients undergoing dialysis.	RCT	PD, HD	HD:19 (9/10); PD: 21 (11/10)

XSL•FO
RenderX

Author and year	Country	Aim of the study	Study design	Population	Total sample (experimental/control group)
Beer et al [49], 2025	Australia	To evaluate the effectiveness of the app in controlling serum phosphorus levels in patients undergoing dialysis.	RCT	PD, HD	180 (90/90)

[a]mHealth: mobile health.

[b]HD: hemodialysis.

[c]RCT: randomized controlled trial.

[d]The study divided participants into four groups: (1) control (n=42), (2) education intervention (n=40), (3) meal intervention (n=34), and (4) education and meal interventions (n=37).

[e]The grouping is set up as follows: (1) educational videos and handouts (Education; n=13), (2) education intervention plus mobile self-monitoring with email feedback (Monitoring; n=14), or (3) education and monitoring interventions plus social cognitive theory-based behavioral videos (Combined; n=13).

[f]PD: peritoneal dialysis.

[g]PDA: personal digital assistant.

Theories or Models Involved

Among the studies included, 14 [14,15,29,31-33,36,37,39,41,43,45-47] did not explicitly mention the theories or models involved. Of these studies, 5 [14,15,45-47] were in patients undergoing PD, and 9 [29,31-33,36,37,39,41,43] involved patients undergoing hemodialysis. Nine [13,20-27] studies combined 2 or more theories or models, and 11 [12,28,30,34,35,38,40,42,44,48,49] studies involved only 1 theory or model.

Among mHealth-based self-management interventions for patients undergoing dialysis, the most frequently incorporated theories included the transtheoretical model (TTM) and stages of change (n=5 [20,22,24,26,42]), self-efficacy theory (n=4 [12,21,24,35]), social cognitive theory (n=4 [25,27,40,48]), and Orem's theory of self-care (n=3 [21,23,28]). Most studies have focused their attention on theories related to behavior change (such as the TTM, the theory of reasoned action [20,22], the health belief model [13,34], behavior change frameworks [30,49], and the behavioral change techniques taxonomy [38]). Additionally, Bandura self-efficacy theory and Oren self-care theory were often used in combination with other theories. Accordingly, we visualized the theories or models involved in this study (Figure 2B).

Attributes of Self-Management in Patients Undergoing Dialysis

We categorized the self-management included in this study into six main themes, which were (1) self-monitoring, (2) diet and fluid management, (3) medication management, (4) exercise management, (5) psychological management, and (6) disease-related knowledge. The components of self-management for patients undergoing dialysis and their corresponding explanations are detailed in Figure 3.

XSL•FO
RenderX

Figure 3. Components of self-management for patients undergoing dialysis. AVF: arteriovenous fistula; AVG: arteriovenous graft; IDWG: intradialytic weight gain; iPTH: intact parathyroid hormone; PTH: parathyroid hormone.

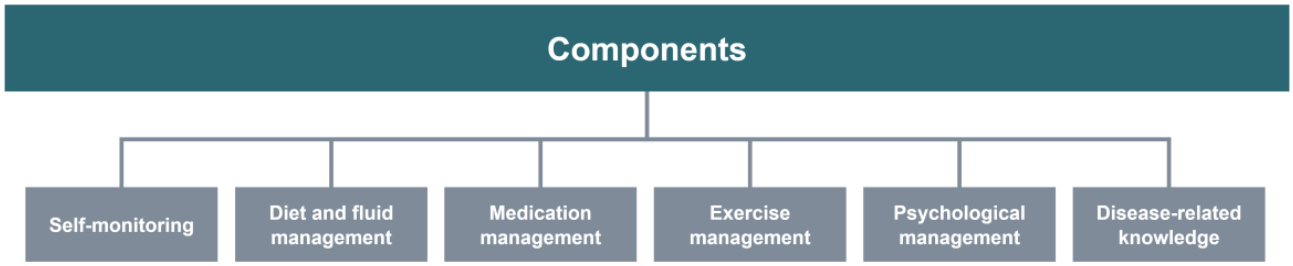

Components	Content	Meaning
Self-monitoring	Weight	Morning and evening weights, pre- and postdialysis weights, dry weights, IDWG, and BMI.
	Vital signs	Pay closer attention to changes in patients' blood pressure before and after dialysis.
	Body fluid	24-hour fluid intake and urine output.
	Symptom	Edema, chest tightness, or dyspnea etc.
	Dialysis indicators	Dialysis duration, number of dialyses, ultrafiltration volume, and adequacy of dialysis.
	Vascular access	The appearance, tremor, and murmur of the AVF or AVG.
	Laboratory tests	Blood glucose, serum potassium, serum sodium, serum phosphorus, serum calcium, serum albumin, iPTH, PTH, alkaline phosphatase, and hemoglobin.
Diet and fluid management	Diet	Control the intake of phosphorus, sodium, potassium, protein, and total calories in food.
	Fluid	Drink water up to the target values.
Medication management	Medication adherence	Accuracy of dosage and timing of administration.
	Drug reactions	Regularly record symptoms and adverse reactions after taking medication.
Exercise management	Exercise	Select the type, duration, and intensity of exercise based on their physical condition.
Psychological management	Psychology	Reduce negative emotions (depression and anxiety) and enhance emotional management skills.
Disease-related knowledge	Knowledge	Knowledge of the disease includes renal function, principles and treatment of dialysis, maintenance of dialysis access (eg, AVF, peritoneal dialysis tubing), prevention of complications, management of diet and fluids, effects and adverse reactions of medications, exercise and rehabilitation, and emotional adjustment.

mHealth-Based Self-Management Intervention Program

The mHealth-based self-management intervention program is focused on the type of mHealth, the content of the mHealth-based intervention, and the duration of the intervention.

The types of mHealth mentioned in this study included app (based on mobile device, tablet personal computer, or personal digital assistant), a remote patient monitoring system, SMS text messaging, and telephones. Apps were divided into 2 categories, such as dialysis-specific software and instant messaging software (eg, WeChat [Tencent Holdings Limited], Line [LY Corporation], Facebook [Meta Platforms, Inc], WhatsApp [Meta Platforms, Inc], and KakaoTalk [Kakao Corp]). Two studies [29,30] intervened via SMS text messaging only, 1 study [31] via telephone only, 4 studies [15,24,33,42] via instant messaging software only, and 18 studies [12-14,20-22,25,26,28,34-36,38,41,43,44,47,49] via dialysis-specific software only. Six studies [23,27,37,39,40,48] used 2 or more online interventions. There were only 3 studies [32,45,46] based on a remote patient monitoring system, 1 [32] for patients undergoing hemodialysis and the remaining 2 [45,46] for patients undergoing PD.

This scope review categorized the app's content into the following dimensions, including intelligent education hub, full-dimensional monitoring system, accurate nutritional management, behavioral interventions (fluid, exercise, and medication adherence), intelligent reminders and alerts, doctor-patient collaboration network, and social support system. Remote patient monitoring systems placed more emphasis on remote monitoring, alarms, and dynamic interventions for patients undergoing dialysis. The details could be found in Table 2. Interventions delivered via apps and remote patient monitoring systems were more comprehensive and satisfactory than those delivered via phone and SMS text messaging, even though they showed similar functionalities in certain aspects. Figure 4 illustrates the gap map of core self-management interventions across different mHealth categories for patients undergoing dialysis (blue circles indicate studies with hemodialysis as the research participants; purple circles indicate

XSL•FO
RenderX

studies with peritoneal dialysis as the research participants; red circles indicate studies with hemodialysis and peritoneal dialysis as research participants).

XSL•FO
RenderX

Table . Type of mobile health and mobile health–based interventions to improve self-management in patients undergoing dialysis.

Classification of mHealth[a]	Numbers	Function and intervention contents
App		

Dialysis-specific software	24	• Intelligent education hub
		  • Disease knowledge base: kidney function, dialysis principles, complication prevention, medications, and lifestyle guidance presented through animated videos, podcasts, manuals, and charts
		  • Operating system training: training for new patients (including equipment operation instruction) and periodic knowledge reinforcement.
		• Full-dimensional monitoring system
		  • Automatic acquisition and manual entry of data: physiological indicators, dialysis parameters, symptom logs, and medication records
		  • Visualization analysis: trend analysis and correlation analysis.
		• Accurate nutritional management
		  • Database support: 500+kinds of kidney disease-specific food nutrients, nutrition calculator (real-time display of phosphorus, sodium, potassium, and protein)
		  • Recording function: barcode scanning to enter food, manual recording of intake
		  • Analysis and feedback: nutritional value calculation for each meal, health scoring system (weekly or monthly summary), electrolyte excess warning (sodium, potassium, and phosphorus)
		  • Behavioral interventions: personalized dynamic recipe recommendations (adjusted based on lab data), daily water intake limits (residual urine volume algorithm)
		• Behavioral interventions (fluid, exercise, and medication adherence)
		  • Personalized goal management: goal setting, progress tracking (instant values + trend charts + health scores)
		  • Behavior shaping tools: badge reward mechanism (continuous recording of achievements)
		• Intelligent reminders and alerts
		  • Treatment reminder: dialysis time and follow-up appointment
		  • Medication reminder: medication time, dosage, and drug interaction
		  • Early warning system: set thresholds to trigger abnormal alerts and notify patients and health care at the same time.
		• Doctor-patient collaboration network
		  • Prescription cloud adjustment, test result synchronization (direct connection to electronic medical records), and equipment interconnection.
		• Social support system
		  • Patient community: experience sharing (recipes or exercise programs)
		  • Family linkage: caregiver collaborative recording function and family health data sharing.
		• System Infrastructure
		  • System compatibility: mobile (iOS/Android)+Web+PDA[b] compatibility, offline data caching, and synchronization
		  • Intelligent devices: support joint NFC[c], OCR[d], and PDA terminals
		  • Multimodal recording: support voice recording, image recognition, and manual supplementation.

XSL•FO
RenderX

Classification of mHealth[a]	Numbers	Function and intervention contents
Instant messaging software	5	• Knowledge delivery and personalized guidance: basic health manuals, video educational resources, and customized content delivery • Dietary interventions and guidance • Psychological intervention: structured facilitation (motivational interviewing), psychological facilitation (drawing healing-experts' video guidance), and emotional support (full psychological status tracking) • Instant interaction: online Q&A[e] with health care, regular phone consultations (psychological support and health guidance), and videoconference support (group discussions or personalized guidance).

A remote patient monitoring system	3	• Intelligent monitoring hub
		  • Device interconnection: bidirectional communication with automatic peritoneal dialysis devices through a cloud platform
		  • Full-dimensional data collection: real-time access to dialysis parameters, device alarm logs, and patient physiological indicators.
		• Remote dynamic intervention
		  • Intelligent alarms: yellow and red alarms, instantly triggering phone interventions;
		  • Prescription cloud adjustment: physicians remotely optimize automated peritoneal dialysis prescription parameters based on real-time data
		  • Physician-patient collaboration platform: support real-time treatment issues through system messages and phone calls.

Telephones	4	• Health education • Treatment adherence assessment • Personalized dietary recommendations • Psychological support • Problem solving
SMS text messaging	4	• Personalized health education • Regular collection of patients' feedback • Regular medication reminders • Positive motivational text messages

[a]mHealth: mobile health.

[b]PDA: personal digital assistant.

[c]NFC: near field communication.

[d]OCR: optical character recognition.

[e]Q&A: question-and-answer.

XSL•FO
RenderX

Figure 4. Gap map of core self-management interventions across different mobile health (mHealth) categories for patients undergoing dialysis.

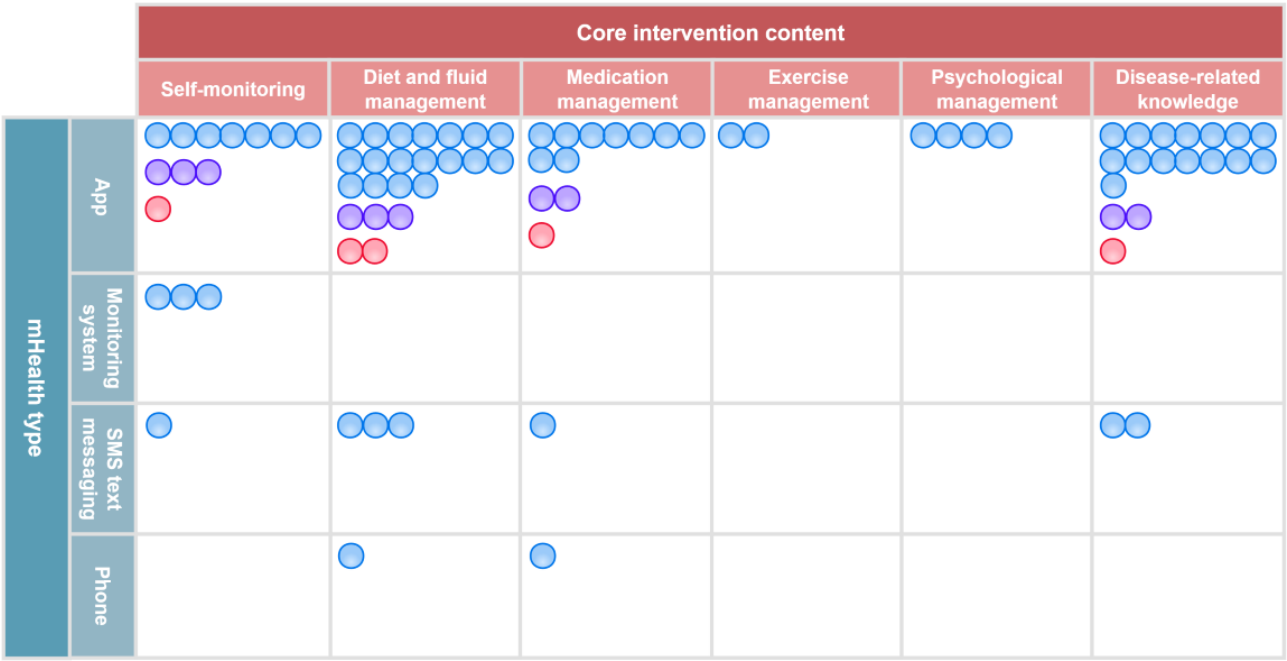

The duration of the intervention ranged from 2 to 24 weeks. The included studies preferred interventions for 3 months [12,24,36,39,43,49], 6 months [21,30,37,40], 12 weeks [26,31,33,41], 2 weeks [20,22,32], 8 weeks [23,29,35].

Key Indicators and Evaluation Tools on Self-Management

Indicators for assessing self-management in patients undergoing dialysis were divided into 2 categories, namely subjective and objective indicators. The subjective indicators mainly involved scales, such as measuring self-management, adherence, self-efficacy, literacy, depression, anxiety, perceived benefits, and quality of life of patients undergoing dialysis. Objective indicators included weight, blood pressure, and laboratory tests. Of these, intradialytic weight gain, serum albumin, serum potassium, and serum phosphorus were the most mentioned indicators in the included studies. Detailed information is provided in Table 3. Figure 5 illustrates the gap map of self-management assessment indicators across different mHealth categories for patients undergoing dialysis (blue circles indicate studies with hemodialysis as the research participants; purple circles indicate studies with peritoneal dialysis as the research participants; red circles indicate studies with hemodialysis and peritoneal dialysis as research participants).

XSL•FO
RenderX

Table . Mobile health–based indicators for assessing self-management in patients undergoing dialysis.

Items	Numbers	Tools
Self-management	4	• The self-management scale for patients undergoing hemodialysis [24,50] • The self-care performance questionnaire [21] • The Perceived Medical Condition Self-Management Scale (PKDSMS) [51] • The Patient Activation Measure-13 (PAM-13) [52]
Adherence	17	• The End-Stage Renal Disease Adherence Questionnaire (ESRD-AQ) [53] • Hemodialysis Compliance Questionnaire [43] • The Simplified Medication Adherence Questionnaire (SMAQ) [54] • The Dialysis Diet and Fluid Nonadherence Questionnaire (DDFQ) [55] • The Modified Morisky Scale (MMS) • The Compliance of Patient Role Behavior Tool [23] • The Sick-role Behavior Adherence [13] • The Adherence Questionnaire (self-developed or revised version)
Self-efficacy	14	• The 15-item dietary self-efficacy questionnaire [56] • The decision self-efficacy scale [57] • The 11-item Fluid Self-Efficacy Scale (FS-ES) [25] • The 6-item Chronic Disease Self-Efficacy Scale (CDSES) [24] • The Self-Efficacy for Appropriate Medication Use Scale (SEAMS) [58] • The strategies used by people to promote health • The Self-Efficacy Scale (self-developed or revised version) [59]
Perceived benefits	2	• The Benefits of Sodium Adherence (BSA) [25,60] • The 9-item Benefits of Fluid Adherence Scale [25,61]
Literacy	2	• The Media Health Literacy Questionnaire (MeHLit) [62] • The Health Literacy Questionnaire (HLQ) [63]
Knowledge	9	• Self-developed or revised knowledge-related questionnaires
Quality of life	12	• Kidney Disease Quality of Life Instrument-Short Form (KDQOL-SF) [64] • The Kidney Disease Quality of Life (KDQOL-36) [65] • Short Form 36 (SF-36) • EuroQol Five Dimensions Questionnaire (EQ-5D) • Health-Related Quality of Life (HRQoL) • 9-item Thai Health Status Assessment Instrument (9-THAI) [66]

XSL•FO
RenderX

Items	Numbers	Tools
Diet and nutrition	9	• 24-hour dietary recall method • 3-day dietary recall method • App automatic calculation • Food Frequency Questionnaires (FFQ) • Malnutrition Inflammation Score (MIS) • Healthy Eating Index (HEI-20)
Psychosocial	3	• Beck Depression Inventory (BDI) • Beck Anxiety Inventory (BAI)
Weight, dry weight	5	• Medical equipment
IDWG[a]	12	• A calculation formula
Blood pressure	3	• Medical equipment
Blood urea nitrogen	4	• Blood sample collection
Creatinine	5	• Blood sample collection
Urea	2	• Blood sample collection
KT/V, dialysis adequacy	5	• A calculation formula
Serum albumin	14	• Blood sample collection
Serum sodium	2	• Blood sample collection
Serum calcium	5	• Blood sample collection
Serum potassium	14	• Blood sample collection
Serum phosphorus	20	• Blood sample collection
Serum aluminum	1	• Blood sample collection
Serum iron	1	• Blood sample collection
Hemoglobin	6	• Blood sample collection
Glycosylated hemoglobin	1	• Blood sample collection
Hematocrit	1	• Blood sample collection
Bicarbonate	1	• Blood sample collection
Alkaline phosphatase	1	• Blood sample collection
iPTH[b]	2	• Blood sample collection
PTH[c]	4	• Blood sample collection
Brain natriuretic peptide	1	• Blood sample collection
C-reactive protein	1	• Blood sample collection

[a]IDWG: intradialytic weight gain.
[b]iPTH: intact parathyroid hormone.
[c]PTH: parathyroid hormone.

XSL•FO
RenderX

Figure 5. Gap map of self-management assessment indicators across different mobile health (mHealth) categories for patients undergoing dialysis.

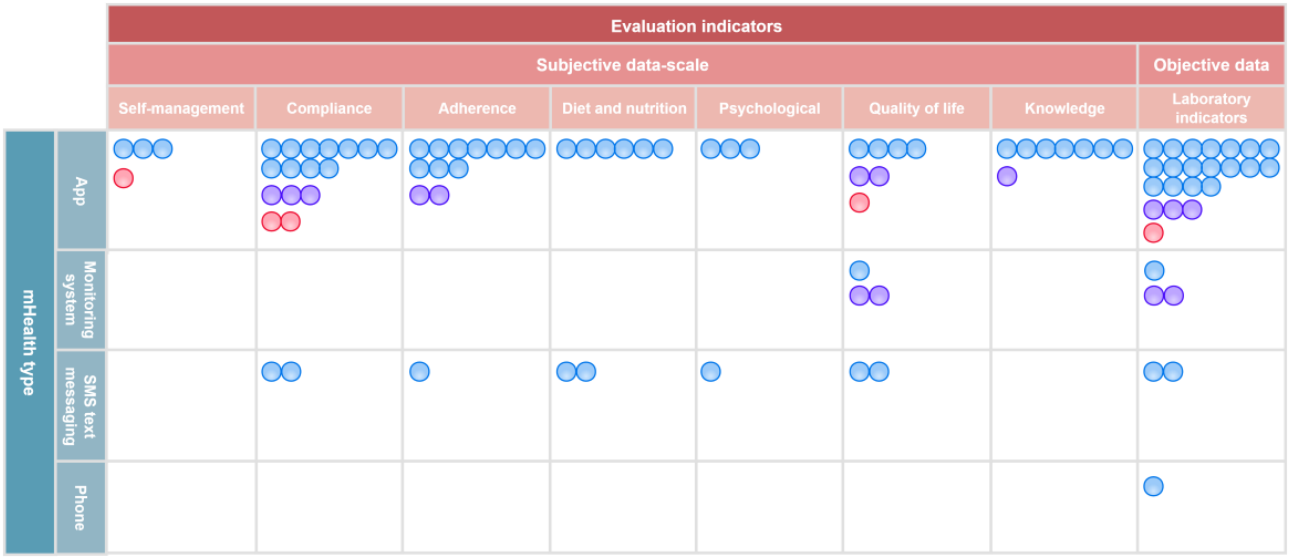

Discussion

Principal Findings

This scoping review provides the first comprehensive analysis and summary of self-management content for patients undergoing dialysis, types of mHealth, relevant theories and models, the content of mHealth-based interventions, and methods for evaluating their effectiveness. Most studies indicate that mHealth-based interventions significantly improve self-management of patients undergoing dialysis and enhance patient outcomes [23,27,37,39,40,48].

Compared to face-to-face training, mHealth-based interventions have a greater impact on patient adherence and laboratory outcomes [67]. In this study, apps and remote monitoring systems were more common. SMS text messaging and phone calls could be used as an aid to promote self-management in patients undergoing dialysis [68]. Due to differences in mHealth-based intervention content and evaluation indicators, quantitative analysis of apps, remote monitoring systems, phones, and SMS text messaging is extremely challenging. Overall, the vast majority of apps demonstrate strong potential for improving medication adherence, enhancing care efficiency, and increasing patient satisfaction and treatment outcomes [69,70].

Special Considerations for Specific Groups Among Patients Undergoing Dialysis

Currently, there are relatively few apps available on the market for patients undergoing dialysis. Existing research has identified 12 Android apps, 11 iOS apps, and 5 dual-platform apps closely associated with kidney disease [71]. In addition, middle-aged and older people are less receptive to apps than younger people, which is particularly reflected in lower usage rates and more negative attitudes toward apps [72]. Since middle-aged and older patients are less receptive to new things and have declining eyesight, the app should be designed to meet the special needs and usage habits of middle-aged and older patients. A survey of requirements for the app in middle-aged and older patients with CKD based on the Kano model showed that the app first had to protect privacy, followed by simplifying the data entry process, simplifying in-app navigation, and simplifying the function and number of buttons [73]. Navigation for beginners, appropriate text size, using color to distinguish different options or icons, and ensuring all buttons maintain consistency in size, labels, and spacing are also details that should be considered during software design [73]. It is also worth noting that there are also barriers that prevent the promotion of apps, such as information barriers, trustworthiness, security, compatibility, complexity, time constraints, and low mHealth literacy [68]. Therefore, in future research, user profiles can be incorporated based on patient age, dialysis modality, and digital literacy levels to automatically match interfaces [74]. During initial use, an interactive tutorial should be set up to guide patients. Beyond traditional touchscreen operations, additional features, such as voice commands, remote assistance, and offline mode, can be added to accommodate the diverse needs and usage habits of patients undergoing dialysis.

Key Areas for Self-Management in Patients Undergoing Dialysis Using mHealth

In terms of app self-management content, CKD disease-related knowledge, symptom management, medication management, provision of health insurance information, diet management, exercise guidance, and psychosocial support may be the content they need more (arranged according to the patient's needs) [73]. However, among the apps for patients undergoing dialysis in this study, the provision of health insurance information was what they lacked. Moreover, the app in this study focused more on disease knowledge, self-monitoring, diet, and medication management and lacked sufficient attention to exercise and psychological support. Compared to self-monitoring, dietary self-management becomes significantly more challenging for patients undergoing dialysis. Strict dietary restrictions, dynamic adjustments to meal plans (based on laboratory indicators), the need for specialized knowledge, lack of external supervision (eg, hospital and home), and the influence of long-standing dietary habits and psychological factors further compound the complexity of implementing dietary management for patients undergoing dialysis. Currently, mHealth's promotion offers

XSL•FO
RenderX

multiple solutions for dietary management among patients undergoing dialysis. The current app can automatically calculate the calories, protein, sodium, phosphorus, and potassium of foods consumed by patients undergoing dialysis, typically achieved through a nutritional database and barcode scanning functionality [25]. A study has also monitored the diet of patients undergoing dialysis by sending photos of food through an app and having them evaluated by the expert [15]. Food analysis and feedback also include the health system score, which is an easy-to-understand health score calculated by the app based on the phosphorus-to-protein ratio of the food [12]. Health care professionals provide personalized advice through dietary records and laboratory tests to promote changes in dietary behavior in patients undergoing dialysis.

Among the included studies, there were fewer studies for patients undergoing PD compared to patients undergoing hemodialysis. This may be related to the late start of PD. Self-management is an important influence on the quality of life and outcomes of patients undergoing PD [75]. Patients undergoing PD who undergo inappropriate operations face a higher risk of developing peritonitis [76]. If patients undergoing PD experience a technical failure, they must be treated with hemodialysis. Compared with PD, long-term regular hemodialysis significantly impacts patients' quality of life, hinders their social reintegration, and increases the risk of complications [77,78]. Therefore, for patients undergoing PD, self-management is necessary for them to master. The focus of self-management in patients undergoing PD is not the same as in patients undergoing hemodialysis. As patients undergoing PD need to perform peritoneal dialysis-related operations at home, training in relevant knowledge and operational skills is particularly important [76]. The training of skills mainly includes aseptic operation, change of peritoneal fluid and peritoneal dialysis catheter, and outlet care [79]. Patients undergoing PD also need to learn the calculation of ultrafiltration [79]. All of these can be learned and managed remotely based on mHealth. In the event of an emergency (eg, contamination or disconnection of the peritoneal dialysis tubing), patients should immediately contact a health care professional for on-site guidance. Thus, establishing emergency contact channels within the app is particularly crucial.

Theoretical or Model Support for mHealth Intervention Programs in Patients Undergoing Dialysis

Many theories and models have been used to guide practice in the self-management of patients undergoing dialysis. TTM, self-efficacy theory, and social cognitive theory are the most commonly used theories [12,20,40,42]. Currently, most studies use a single theory as their guiding framework, with fewer adopting 2 or more theories for guidance. Taking the TTM as an example, it emphasizes guiding patients through staged behavioral shifts but overlooks individual differences (such as cultural background and cognitive level) [20,22,24,26,42]. Self-care theory, on the other hand, emphasizes personalized care needs and can be combined with TTM. Additionally, the maintenance phase of TTM is prone to behavioral relapse, particularly among patients requiring lifelong treatment (eg, dialysis). A single theory cannot sustainably motivate patients. Integrating other theories (such as the PERMA [Positive Emotion, Engagement, Relationship, Meaning and Accomplishment] model [80]) during this phase can further reinforce patients' behavior. Therefore, future research should focus more on integrating multiple theories rather than relying on a single one. In addition, goal-setting theory, the information-motivation-behavioral skills model, the chronic care model, the ecological model of health behavior, and the theory of planned behavior can also be applied in subsequent research.

Implications for Future Research

Future research on self-management software for patients undergoing dialysis will focus on integrating the strengths of existing tools and promoting their synergistic use. Software development should break down barriers between tools, such as deeply integrating the real-time data collection capabilities of remote patient monitoring systems, the personalized analysis functions of intelligent decision support systems, and the instant communication advantages of phone or SMS text messaging, to provide patients with more comprehensive self-management support.

A combination of artificial intelligence can be considered for use in the remote management of patients undergoing dialysis (eg, wearable devices). Smart wristbands can monitor data, such as heart rate, blood pressure, activity, and sleep [81] and automatically transmit the data to the app for analysis and storage. Some researchers have implemented data linkage through apps using near field communication and optical character recognition. Data from measuring devices, such as sphygmomanometers and weight scales, can be automatically transmitted to the app via near field communication [14]. Alternatively, the numbers from the sphygmomanometer can be captured using the phone's camera and imported into the app [14]. These not only improve the efficiency of patients undergoing dialysis but also increase the accuracy of the records. In conclusion, there is still a lot of room to explore the use of artificial intelligence in combination with mHealth in patients undergoing dialysis.

Different studies have different insights into the evaluation criteria for self-management. Most of the studies used objective indicators as one of the evaluation criteria for self-management. For the evaluation of knowledge of patients undergoing dialysis, most studies have assessed it using self-developed questionnaires and lacked uniform criteria for judging. Various scales are currently available for assessing self-management, adherence, self-efficacy, and quality of life. The use of these assessment tools varies across different studies. Existing research lacks a gold standard for evaluating self-management in patients undergoing dialysis. Therefore, there is a need to standardize the criteria for evaluating self-management in patients undergoing dialysis in future studies.

This study provides guidance for the development of subsequent dialysis-related software for patients, including the design of functional modules, user experience optimization, and the integration of clinical indicators and assessment tools. Our findings will enhance the ability of patients undergoing dialysis to self-manage their health at home, improving both the effectiveness of their dialysis treatment and their quality of life.

XSL•FO
RenderX

In the future design of the app, attention should also be paid to the usage needs of special groups, strengthening adaptive software development, deeply integrating artificial intelligence technology, and establishing standardized self-management evaluation criteria.

Limitation

This scoping review also has some limitations. First, the broad scope of scoping reviews and the complexity of search strategies may lead to the omission of relevant studies. These challenges persist despite strict adherence to the PRISMA-ScR guidelines for greater rigor and transparency. Second, self-management assessment tools incorporate numerous subjective and objective indicators, which complicate data integration and comparability. Third, publication bias was one of the limitations of this study. This review only covered relevant studies published in English. Most of the included studies were limited to Asia and North America, which may have resulted in limited global generalizability. Future studies should use standardized, integrated measures to improve consistency and reliability. Additionally, these shortcomings can be remedied by improving search strategies, expanding database coverage, and removing language restrictions to include a more diverse patient population from different continents.

Conclusion

This study conducted a scoping review of the existing literature on mHealth-based self-management among patients undergoing dialysis. Although mHealth holds potential advantages for self-management in patients undergoing dialysis, it has not been widely adopted and integrated into standard renal care, requiring further optimization and refinement. This study provides theoretical and practical guidance for subsequent research, helping to enhance self-management of patients undergoing dialysis and improve their quality of life, ultimately offering insights for digital transformation in chronic disease management.

Multimedia Appendix 1
Search strategy
[DOCX File, 17 KB - jmir_v28i1e76880_app1.docx]
Multimedia Appendix 2
Critical appraisal of the selected studies using the Mixed Methods Appraisal Tool (MMAT).
[XLSX File, 72 KB - jmir_v28i1e76880_app2.xlsx] Multimedia Appendix 3
Types of mobile health, core intervention content, intervention time, and evaluation indicators for self-management in patients undergoing dialysis.
[DOCX File, 45 KB - jmir_v28i1e76880_app3.docx] Checklist 1
PRISMA-ScR and PRISMA-S checklists.
[PDF File, 315 KB - jmir_v28i1e76880_app4.pdf]

Acknowledgments

We are grateful to the editors and reviewers who helped with this study.

Funding

This work was supported by grants from the Tongji Hospital Research Fund Project (2024D10, 2022D22). The funder participated in the research design, data collection, analysis, interpretation, and manuscript writing.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Conflicts of Interest

None declared.

References

1. See E, Ethier I, Cho Y, et al. Dialysis outcomes across countries and regions: a global perspective from the International Society of Nephrology Global Kidney Health Atlas study. Kidney Int Rep 2024 Aug;9(8):2410-2419. [doi: 10.1016/j.ekir.2024.05.014] [Medline: 39156158]
2. Bello AK, Okpechi IG, Levin A, et al. An update on the global disparities in kidney disease burden and care across world countries and regions. Lancet Glob Health 2024 Mar;12(3):e382-e395. [doi: 10.1016/S2214-109X(23)00570-3] [Medline: 38365413]
3. Bastani B. The present and future of transplant organ shortage: some potential remedies. J Nephrol 2020 Apr;33(2):277-288. [doi: 10.1007/s40620-019-00634-x] [Medline: 31399908]
4. He Z, Hou H, Zhang D, et al. Effects of dialysis modality choice on the survival of end-stage renal disease patients in southern China: a retrospective cohort study. BMC Nephrol 2020 Sep 24;21(1):412. [doi: 10.1186/s12882-020-02070-7] [Medline: 32972378]
5. Chisavu L, Mihaescu A, Bob F, et al. Trends in mortality and comorbidities in hemodialysis patients between 2012 and 2017 in an East-European Country: a retrospective study. Int Urol Nephrol 2023 Oct;55(10):2579-2587. [doi: 10.1007/s11255-023-03549-6] [Medline: 36917413]

XSL•FO
RenderX

6. Al-Mansouri A, Al-Ali FS, Hamad AI, et al. Assessment of treatment burden and its impact on quality of life in dialysis-dependent and pre-dialysis chronic kidney disease patients. Res Social Adm Pharm 2021 Nov;17(11):1937-1944. [doi: 10.1016/j.sapharm.2021.02.010] [Medline: 33612446]
7. Pretto CR, Winkelmann ER, Hildebrandt LM, Barbosa DA, Colet CDF, Stumm EMF. Quality of life of chronic kidney patients on hemodialysis and related factors. Rev Lat Am Enfermagem 2020;28:e3327. [doi: 10.1590/1518-8345.3641.3327] [Medline: 32696925]
8. Lorig KR, Sobel DS, Ritter PL, Laurent D, Hobbs M. Effect of a self-management program on patients with chronic disease. Eff Clin Pract 2001;4(6):256-262. [Medline: 11769298]
9. Almutary H, Tayyib N. Evaluating self-efficacy among patients undergoing dialysis therapy. Nurs Rep 2021 Mar 23;11(1):195-201. [doi: 10.3390/nursrep11010019] [Medline: 34968324]
10. Escudero-Lopez M, Martinez-Andres M, Marcilla-Toribio I, Moratalla-Cebrian ML, Perez-Moreno A, Bartolome-Gutierrez R. Barriers and facilitators in self-care and management of chronic kidney disease in dialysis patients: a systematic review of qualitative studies. J Clin Nurs 2024 Oct;33(10):3815-3830. [doi: 10.1111/jocn.17193] [Medline: 38716807]
11. Yang Y, Chen H, Qazi H, Morita PP. Intervention and evaluation of mobile health technologies in management of patients undergoing chronic dialysis: scoping review. JMIR Mhealth Uhealth 2020 Apr 3;8(4):e15549. [doi: 10.2196/15549] [Medline: 32242823]
12. Chiang YC, Chang YP, Lin SC, et al. Effects of individualized dietary phosphate control program with a smartphone application in hemodialysis patients in Taiwan in China. Biol Res Nurs 2021 Jul;23(3):375-381. [doi: 10.1177/1099800420975504] [Medline: 33251815]
13. Min Y, Park M. Effects of a mobile-app-based self-management support program for elderly hemodialysis patients. Healthc Inform Res 2020 Apr;26(2):93-103. [doi: 10.4258/hir.2020.26.2.93] [Medline: 32547806]
14. Lukkanalikitkul E, Kongpetch S, Chotmongkol W, et al. Optimization of the chronic kidney disease-peritoneal dialysis app to improve care for patients on peritoneal dialysis in northeast Thailand: user-centered design study. JMIR Form Res 2022 Jul 6;6(7):e37291. [doi: 10.2196/37291] [Medline: 35793137]
15. Lee HJ, Kang HY. Effects of a customized diet education program using a mobile instant messenger for people undergoing peritoneal dialysis: a feasibility test. Asian Nurs Res (Korean Soc Nurs Sci) 2024 Oct;18(4):367-376. [doi: 10.1016/j.anr.2024.09.007] [Medline: 39284546]
16. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
17. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
18. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc 2021 Jan;109(2):34285662. [doi: 10.5195/jmla.2021.962]
19. Hong QN, Fàbregues S, Bartlett G, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018;34(4):285-291. [doi: 10.3233/EFI-180221]
20. Fakih El Khoury C, Crutzen R, Schols JM, Halfens RJ, Karavetian M. Adequate management of phosphorus in patients undergoing hemodialysis using a dietary smartphone app: prospective pilot study. JMIR Form Res 2021 Jun 1;5(6):e17858. [doi: 10.2196/17858] [Medline: 34061034]
21. Hosseini A, Jackson AC, Chegini N, et al. The effect of an educational app on hemodialysis patients' self-efficacy and self-care: A quasi-experimental longitudinal study. Chronic Illn 2023 Jun;19(2):383-394. [doi: 10.1177/17423953211073365] [Medline: 35179394]
22. Fakih El Khoury C, Crutzen R, Schols J, Halfens RJG, Karavetian M. A dietary mobile app for patients undergoing hemodialysis: prospective pilot study to improve dietary intakes. J Med Internet Res 2020 Jul 20;22(7):e17817. [doi: 10.2196/17817] [Medline: 32706698]
23. Park OL, Kim SR. Integrated self-management program effects on hemodialysis patients: a quasi-experimental study. Jpn J Nurs Sci 2019 Oct;16(4):396-406. [doi: 10.1111/jjns.12249] [Medline: 30669185]
24. Ren Q, Lian M, Liu Y, Thomas-Hawkins C, Zhu L, Shen Q. Effects of a transtheoretical model-based WeChat health education programme on self-management among haemodialysis patients: a longitudinal experimental intervention study. J Adv Nurs 2019 Dec;75(12):3554-3565. [doi: 10.1111/jan.14182] [Medline: 31441525]
25. Welch JL, Astroth KS, Perkins SM, et al. Using a mobile application to self-monitor diet and fluid intake among adults receiving hemodialysis. Res Nurs Health 2013 Jun;36(3):284-298. [doi: 10.1002/nur.21539] [Medline: 23512869]
26. Zwi S, Isautier J, Webster AC, et al. A feasibility study of a best practice health literacy app for Australian adults with chronic kidney disease. PEC Innov 2022 Dec;1:100047. [doi: 10.1016/j.pecinn.2022.100047] [Medline: 37213763]
27. Chae YJ, Kim HS. Effects of a mobile application on improving self-management of adult patients receiving peritoneal dialysis: a randomized controlled trial. Jpn J Nurs Sci 2024 Jan;21(1):e12555. [doi: 10.1111/jjns.12555] [Medline: 37589209]
28. Cho H, Park S. The effects of self‐performance management video program on patients receiving hemodialysis. Jpn J Nurs Sci 2020 Apr;17(2):e12303. [doi: 10.1111/jjns.12303] [Medline: 31746139]

XSL•FO
RenderX

29. Chung MK, Kim DH, Park JI, et al. Adaptive nutrition intervention stabilizes serum phosphorus levels in hemodialysis patients: a multicenter decentralized clinical trial using real-world data. J Ren Nutr 2024 Jan;34(1):47-57. [doi: 10.1053/j.jrn.2023.07.004] [Medline: 37586668]
30. Dawson J, Campbell KL, Craig JC, et al. A text messaging intervention for dietary behaviors for people receiving maintenance hemodialysis: a feasibility study of KIDNEYTEXT. Am J Kidney Dis 2021 Jul;78(1):85-95. [doi: 10.1053/j.ajkd.2020.11.015] [Medline: 33421456]
31. Hanifi N, Ezzat LS, Dinmohammadi M. Effect of consultation and follow-up phone calls on biochemical indicators and intradialytic weight gain in patients undergoing hemodialysis. Oman Med J 2019 Mar;34(2):137-146. [doi: 10.5001/omj.2019.26] [Medline: 30918608]
32. Hayashi A, Yamaguchi S, Waki K, et al. Testing the feasibility and usability of a novel smartphone-based self-management support system for dialysis patients: a pilot study. JMIR Res Protoc 2017 Apr 20;6(4):e63. [doi: 10.2196/resprot.7105] [Medline: 28428168]
33. Mollaoğlu M, Candan F, Solmaz G, Mollaoğlu S, Başer E, Yanmiş S. The effect of education and art therapy with telehealth method on diet-fluid restriction and anxiety in hemodialysis patients during the COVID-19 pandemic. Altern Ther Health Med 2024 Jul;30(7):58-64. [Medline: 39110042]
34. Nursalam N, Kurniawati ND, Putri IRP, Priyantini D. Automatic reminder for fluids management on confidence and compliance with fluid restrictions in hemodialysis patients. Sys Rev Pharm 2020 Jun 1;11(5):226-233. [doi: 10.31838/srp.2020.5.34]
35. Pack S, Lee J. Randomised controlled trial of a smartphone application-based dietary self-management program on haemodialysis patients. J Clin Nurs 2021 Mar;30(5-6):840-848. [doi: 10.1111/jocn.15627] [Medline: 33377565]
36. Pinto LCS, Andrade MC, Chaves RO, et al. Development and validation of an application for follow-up of patients undergoing dialysis: NefroPortátil. J Ren Nutr 2020 Jul;30(4):e51-e57. [doi: 10.1053/j.jrn.2019.03.082] [Medline: 32081517]
37. Pungchompoo W, Parinyachitta S, Pungchompoo S, Udomkhwamsuk W, Suwan P. The feasibility of integrating a home telehealth model for older persons living with hemodialysis. BMC Geriatr 2024 Apr 27;24(1):378. [doi: 10.1186/s12877-024-04981-8] [Medline: 38671357]
38. Rocco MV, Rigaud M, Ertel C, Russell G, Zemdegs J, Vecchio M. Fluid intake management in maintenance hemodialysis using a smartphone-based application: a pilot study. Kidney Med 2023 Sep;5(9):100703. [doi: 10.1016/j.xkme.2023.100703] [Medline: 37663954]
39. Saadatifar B, Sharifi S, Faghihi H, Sadeghi Googhary N. Effect of mHealth training on treatment adherence in hemodialysis patients. Med Surg Nurs J 2022;11(3):1-8. [doi: 10.5812/msnj-134851]
40. St-Jules DE, Woolf K, Goldfarb DS, et al. Feasibility and acceptability of mHealth interventions for managing hyperphosphatemia in patients undergoing hemodialysis. J Ren Nutr 2021 Jul;31(4):403-410. [doi: 10.1053/j.jrn.2020.07.009] [Medline: 33160812]
41. Teong LF, Khor BH, Ng HM, et al. Effectiveness of a nutritional mobile application for management of hyperphosphatemia in patients on hemodialysis: a multicenter open-label randomized clinical trial. J Pers Med 2022 Jun 12;12(6):961. [doi: 10.3390/jpm12060961] [Medline: 35743746]
42. Thongsunti A, Silpakit C, Rattananupong T, Kittanamongkolchai W, Sumethpimolchai W, Lohsoonthorn V. Effect of a transtheoretical model-based intervention and motivational interviewing on hyperphosphatemia management via telehealth (TMT program) among hemodialysis patients during the COVID-19 pandemic. Front Public Health 2024;12:1361778. [doi: 10.3389/fpubh.2024.1361778] [Medline: 39668955]
43. Andriati R. Innovation of the M-Nursing application to improve compliance with haemodialysis therapy and urea-creatinine levels in patients with chronic kidney disease. Malays J Nurs 2025;17(1) [FREE Full text] [doi: 10.31674/mjn.2025.v17i01.011]
44. Taguiam DJ. Optimizing fluid intake management in adult hemodialysis patients: the impact of the H2Overload mobile health app on interdialytic weight gain [Dissertation]. : William Paterson University; 2025 URL: http://hdl.handle.net/20.500.12164/3510 [accessed 2025-12-06]
45. Uchiyama K, Morimoto K, Washida N, et al. Effects of a remote patient monitoring system for patients on automated peritoneal dialysis: a randomized crossover controlled trial. Int Urol Nephrol 2022 Oct;54(10):2673-2681. [doi: 10.1007/s11255-022-03178-5]
46. Jung HY, Jeon Y, Kim YS, et al. Outcomes of remote patient monitoring for automated peritoneal dialysis: a randomized controlled trial. Nephron 2021;145(6):702-710. [doi: 10.1159/000518364] [Medline: 34515160]
47. Zeng Y, Yin Y, Deng J, et al. Effectiveness of a smart management system in improving adherence and clinical outcomes of patients receiving peritoneal dialysis: a retrospective cohort analysis. BMC Nurs 2025 Jul 7;24(1):860. [doi: 10.1186/s12912-025-03506-x] [Medline: 40624644]
48. Stark S, Snetselaar L, Piraino B, et al. Personal digital assistant-based self-monitoring adherence rates in 2 dialysis dietary intervention pilot studies: BalanceWise-HD and BalanceWise-PD. J Ren Nutr 2011 Nov;21(6):492-498. [doi: 10.1053/j.jrn.2010.10.026] [Medline: 21420316]
49. Beer J, Jacques A, Lambert K, Lim W, Howell M, Boudville N. TELEnutrition and KIdNey hEalth Study: protocol for a randomised controlled trial comparing the effect of digital health to standard care on serum phoSphate control in patients

XSL•FO
RenderX

on dIalysiS (TeleKinesis Study). BMJ Open 2025 May 2;15(5):e096381. [doi: 10.1136/bmjopen-2024-096381] [Medline: 40316358]
50. Song Y. Developing and testing of the hemodialysis self-management instrument. : Kaohsiung Medical University; 2009.
51. Wild MG, Wallston KA, Green JA, et al. The Perceived Medical Condition Self-Management Scale can be applied to patients with chronic kidney disease. Kidney Int 2017 Oct;92(4):972-978. [doi: 10.1016/j.kint.2017.03.018] [Medline: 28528132]
52. Hibbard JH, Mahoney ER, Stockard J, Tusler M. Development and testing of a short form of the patient activation measure. Health Serv Res 2005 Dec;40(6 Pt 1):1918-1930. [doi: 10.1111/j.1475-6773.2005.00438.x] [Medline: 16336556]
53. Kim Y, Evangelista LS, Phillips LR, Pavlish C, Kopple JD. The End-Stage Renal Disease Adherence Questionnaire (ESRD-AQ): testing the psychometric properties in patients receiving in-center hemodialysis. Nephrol Nurs J 2010;37(4):377-393. [Medline: 20830945]
54. Ortega Suárez FJ, Sánchez Plumed J, Pérez Valentín MA, et al. Validation on the simplified medication adherence questionnaire (SMAQ) in renal transplant patients on tacrolimus. Nefrologia 2011;31(6):690-696. [doi: 10.3265/Nefrologia.pre2011.Aug.10973] [Medline: 22130285]
55. Vlaminck H, Maes B, Jacobs A, Reyntjens S, Evers G. The dialysis diet and fluid non-adherence questionnaire: validity testing of a self-report instrument for clinical practice. J Clin Nurs 2001 Sep;10(5):707-715. [doi: 10.1046/j.1365-2702.2001.00537.x] [Medline: 11822521]
56. Seo AR, Park KS, Kim BK, Kim YL, Choi JY. A validation of dietary self-efficacy questionnaire in hemodialysis patients. Korean J Health Promot ;12(1):22-30 [FREE Full text]
57. O'Connor AM. User Manual - Decision Self-Efficacy Scale. Ottawa Hospital Research Institute. 1995. URL: https://decisionaid.ohri.ca/docs/develop/User_Manuals/UM_Decision_SelfEfficacy.pdf [accessed 2025-12-06]
58. Risser J, Jacobson TA, Kripalani S. Development and psychometric evaluation of the Self-efficacy for Appropriate Medication Use Scale (SEAMS) in low-literacy patients with chronic disease. J Nurs Meas 2007;15(3):203-219. [doi: 10.1891/106137407783095757] [Medline: 18232619]
59. Song MR, Kim MJ, Lee ME, Lee IB, Shu MR. A study on the correlation between self-efficacy and self-care in hemodialysis patients. J Korean Acad Nurs 1999;29(3):563. [doi: 10.4040/jkan.1999.29.3.563]
60. Welch JL, Bennett SJ, Delp RL, Agarwal R. Benefits of and barriers to dietary sodium adherence. West J Nurs Res 2006 Mar;28(2):162-180. [doi: 10.1177/0193945905282323] [Medline: 16513918]
61. Welch JL. Hemodialysis patient beliefs by stage of fluid adherence. Res Nurs Health 2001 Apr;24(2):105-112. [doi: 10.1002/nur.1013] [Medline: 11353458]
62. Nazarnia M, Zarei F, Rozbahani N. Development and psychometric properties of a tool to assess Media Health Literacy (MeHLit). BMC Public Health 2022 Oct 1;22(1):1839. [doi: 10.1186/s12889-022-14221-6] [Medline: 36180875]
63. Osborne RH, Batterham RW, Elsworth GR, Hawkins M, Buchbinder R. The grounded psychometric development and initial validation of the Health Literacy Questionnaire (HLQ). BMC Public Health 2013 Jul 16;13:658. [doi: 10.1186/1471-2458-13-658] [Medline: 23855504]
64. Hays RD, Kallich JD, Mapes DL, Coons SJ, Carter WB. Development of the kidney disease quality of life (KDQOL) instrument. Qual Life Res 1994 Oct;3(5):329-338. [doi: 10.1007/BF00451725] [Medline: 7841967]
65. Ricardo AC, Hacker E, Lora CM, et al. Validation of the Kidney Disease Quality of Life Short Form 36 (KDQOL-36) US Spanish and English versions in a cohort of Hispanics with chronic kidney disease. Ethn Dis 2013;23(2):202-209. [Medline: 23530302]
66. Cheawchanwattana A, Limwattananon C, Gross C, et al. The validity of A new practical quality of life measure in patients on renal replacement therapy. J Med Assoc Thai 2006 Aug;89 Suppl 2:S207-S217. [Medline: 17044474]
67. Torabikhah M, Farsi Z, Sajadi SA. Comparing the effects of mHealth app use and face-to-face training on the clinical and laboratory parameters of dietary and fluid intake adherence in hemodialysis patients: a randomized clinical trial. BMC Nephrol 2023 Jun 29;24(1):194. [doi: 10.1186/s12882-023-03246-7] [Medline: 37386428]
68. Shen H, van der Kleij R, van der Boog PJM, et al. Digital tools/eHealth to support CKD self-management: a qualitative study of perceptions, attitudes and needs of patients and health care professionals in China. Int J Med Inform 2022 Sep;165:104811. [doi: 10.1016/j.ijmedinf.2022.104811] [Medline: 35753175]
69. Paneerselvam GS, Lua PL, Chooi WH, Rehman IU, Goh KW, Ming LC. Effectiveness of mobile apps in improving medication adherence among chronic kidney disease patients: systematic review. J Med Internet Res 2025 Apr 16;27:e53144. [doi: 10.2196/53144] [Medline: 40239197]
70. Chao SM, Pan CK, Wang ML, Fang YW, Chen SF. Functionality and usability of mHealth apps in patients with peritoneal dialysis: a systematic review. Healthcare (Basel) 2024 Mar 5;12(5):593. [doi: 10.3390/healthcare12050593] [Medline: 38470704]
71. Singh K, Diamantidis CJ, Ramani S, et al. Patients' and Nephrologists' evaluation of patient-facing smartphone apps for CKD. Clin J Am Soc Nephrol 2019 Apr 5;14(4):523-529. [doi: 10.2215/CJN.10370818] [Medline: 30898873]
72. Baer NR, Vietzke J, Schenk L. Middle-aged and older adults' acceptance of mobile nutrition and fitness apps: a systematic mixed studies review. PLoS ONE 2022;17(12):e0278879. [doi: 10.1371/journal.pone.0278879] [Medline: 36520839]

XSL•FO
RenderX

73. Yan Y, Liu M, Duan DF, Yan LJ, Li L, Ma DY. Demand analysis of self-management mobile health applications for middle-aged and older patients with chronic kidney disease based on the Kano model. Nephron 2025;149(3):166-177. [doi: 10.1159/000541729] [Medline: 39396506]
74. Gosetto L, Falquet G, Ehrler F. Personalizing mobile applications for health based on user profiles: a preference matrix from a scoping review. PLOS Digit Health 2025 Aug;4(8):e0000978. [doi: 10.1371/journal.pdig.0000978] [Medline: 40828800]
75. Huang Y, Li S, Lu X, Chen W, Zhang Y. The effect of self-management on patients with chronic diseases: a systematic review and meta-analysis. Healthcare (Basel) 2024 Oct 29;12(21):2151. [doi: 10.3390/healthcare12212151] [Medline: 39517362]
76. Jaelani TR, Ibrahim K, Jonny J, et al. Peritoneal dialysis patient training program to enhance independence and prevent complications: a scoping review. Int J Nephrol Renovasc Dis 2023;16:207-222. [doi: 10.2147/IJNRD.S414447] [Medline: 37720493]
77. Ethier I, Hayat A, Pei J, et al. Peritoneal dialysis versus haemodialysis for people commencing dialysis. Cochrane Database Syst Rev 2024 Jun 20;6(6):CD013800. [doi: 10.1002/14651858.CD013800.pub2] [Medline: 38899545]
78. Chuasuwan A, Pooripussarakul S, Thakkinstian A, Ingsathit A, Pattanaprateep O. Comparisons of quality of life between patients underwent peritoneal dialysis and hemodialysis: a systematic review and meta-analysis. Health Qual Life Outcomes 2020 Jun 18;18(1):191. [doi: 10.1186/s12955-020-01449-2] [Medline: 32552800]
79. Figueiredo AE, Bernardini J, Bowes E, et al. A syllabus for teaching peritoneal dialysis to patients and caregivers. Perit Dial Int 2016;36(6):592-605. [doi: 10.3747/pdi.2015.00277] [Medline: 26917664]
80. Madeson M. Seligman's PERMA+ model explained: a theory of wellbeing. PositivePsychology.com. 2017 Feb 24. URL: https://positivepsychology.com/perma-model/ [accessed 2025-12-16]
81. Stauss M, Htay H, Kooman JP, Lindsay T, Woywodt A. Wearables in Nephrology: Fanciful Gadgetry or Prêt-à-Porter? Sensors (Basel) 2023 Jan 26;23(3):1361. [doi: 10.3390/s23031361]

Abbreviations

CKD: chronic kidney diseases
ESRD: end-stage renal disease
mHealth: mobile health
PD: peritoneal dialysis
PERMA: Positive Emotion, Engagement, Relationship, Meaning and Accomplishment
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
TTM: transtheoretical model

Edited by S Brini; submitted 03.May.2025; peer-reviewed by D Iryawati, Y Shi; accepted 17.Nov.2025; published 02.Jan.2026.

Please cite as:
Xu Q, Xu Y, Liu X, Ma X
The Application of Mobile Health in Self-Management Among Patients Undergoing Dialysis: Scoping Review
J Med Internet Res 2026;28:e76880
URL: https://www.jmir.org/2026/1/e76880
doi:10.2196/76880

© Qin Xu, Yulin Xu, Xiaoqin Liu, Xiaolin Ma. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Accuracy of Deep Learning in Diagnosing Chronic Obstructive Pulmonary Disease: Systematic Review and Meta-Analysis

Hui Yang[1,2]; Yijiu Wu[3]; Tong Wu[1]; Jingyan Ji[2]; Sitao Lei[2]; Weibin Xu[1]

[1]School of Management, Guizhou University, Huaxi District, Guiyang, Guizhou, China
[2]School of Forensic Medicine, Guizhou Medical University, Guiyang, Guizhou, China
[3]School of Anesthesiology, Guizhou Medical University, Guiyang, Guizhou, China

Corresponding Author:
Weibin Xu
School of Management, Guizhou University, Huaxi District, Guiyang, Guizhou, China

Abstract

Background: Chronic obstructive pulmonary disease (COPD) is a common chronic lung disease. Deep learning (DL), a data-driven machine learning approach, has gained attention in clinical practice, particularly for diagnosing COPD and grading its severity. However, systematic evidence of its diagnostic and grading accuracy remains limited, posing challenges for developing intelligent diagnostic tools.

Objective: This study aimed to systematically estimate the accuracy of DL models for diagnosing and grading COPD, providing up-to-date evidence for the design and clinical implementation of intelligent detection tools.

Methods: The Cochrane Library, Embase, Web of Science, and PubMed were systematically searched for studies on DL for diagnosing COPD and grading its severity published up to November 1, 2025. Risk of bias was assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 tool. Subgroup analyses by the validation set generation method and imaging data source were conducted, and meta-analyses were performed on the validation sets. For binary outcomes, diagnostic 2×2 tables were synthesized using a bivariate mixed effects model; for multiclass outcomes, accuracy estimates were pooled using random-effects models.

Results: In total, 56 studies comprising 886,753 participants were included. Inputs were computed tomography (CT) imaging (n=30), breath sounds or audio (n=12), conventional chest X-ray (n=2), X-ray film (n=2), and other modalities (n=10), including pulmonary function indices or curves or physiological waveforms, electrocardiograms, volumetric capnography maps, radiogenetic data, and clinical scores. For binary classification of COPD, DL models yielded a pooled sensitivity of 0.87 (95% CI 0.85 - 0.90), specificity of 0.88 (95% CI 0.84 - 0.92), diagnostic odds ratio (DOR) of 52 (95% CI 30 - 88), and the area under the summary receiver operating characteristic curve (AUC) of 0.93. For CT-based DL models, pooled sensitivity was 0.86 (95% CI 0.84 - 0.89), specificity was 0.87 (95% CI 0.82 - 0.90), DOR was 42 (95% CI 26 - 68), and AUC was 0.92. For respiratory sound–based models, sensitivity was 0.91 (95% CI 0.84 - 0.95), specificity was 0.96 (95% CI 0.91 - 0.98), DOR was 237 (95% CI 78 - 723), and AUC was 0.98. In multiclass classification, the DL models showed limited accuracy in discriminating Global Initiative for Chronic Obstructive Lung Disease (GOLD) stages: GOLD stage 0 (84.2%, 95% CI 60.5% - 98.2%), stage 1 (61.7%, 95% CI 40.7% - 80.8%), stage 2 (67.9%, 95% CI 37.6% - 91.7%), stage 3 (70.8%, 95% CI 16.3% - 100%), and stage 4 (70.8%, 95% CI 16.3% - 100%).

Conclusions: This study is the first systematic synthesis of DL applications for COPD detection and GOLD staging. DL models based on CT images and breath sounds show high accuracy for binary COPD detection, whereas multiclass GOLD grading remains concerning. These findings support the development and updating of artificial intelligence−assisted COPD screening tools; however, substantial heterogeneity and limited external validation warrant cautious interpretation. Future reproducible multicenter studies with standardized reporting are needed.

Trial Registration: PROSPERO CRD420251114195; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251114195

(J Med Internet Res 2026;28:e83459) doi:10.2196/83459

KEYWORDS

chronic obstructive pulmonary disease; COPD; deep learning; diagnosis; Global Initiative for Chronic Obstructive Lung Disease; GOLD grading; meta-analysis

XSL•FO
RenderX

Introduction

Chronic obstructive pulmonary disease (COPD) is a prevalent chronic respiratory illness characterized by persistent airflow limitation. It is irreversible and progressively worsens over time, severely affecting patients' quality of life and life expectancy [1]. According to the latest World Health Organization report, COPD is the fourth leading cause of death worldwide, responsible for over 3 million deaths each year, leading to a disproportionate burden in low- and middle-income countries [2]. China accounts for about one-quarter of the global burden of COPD, with an estimated 99.9 million people affected and a prevalence of 13.7% among adults aged ≥40 years [3]. Acute exacerbations are pivotal events in COPD, causing hospital admission and increasing the risk of mortality. The 5-year mortality rate after exacerbation is about 50% after hospitalization [4]. A real-world multicenter prospective cohort study in Japan has reported a 5-year survival rate of 85.4% among COPD patients, whereas those with very severe airflow limitation have a reduced 5-year survival rate of 66.1% [5]. Consequently, COPD not only represents a significant public health issue worldwide but has also become one of the main causes of disability and death.

In clinical practice, the gold standard for diagnosing COPD is pulmonary function testing (PFT), which primarily quantifies expiratory airflow limitation. Based on the Global Initiative for Chronic Obstructive Lung Disease (GOLD) guidelines, COPD is defined as the ratio of forced expiratory volume in 1 second to forced vital capacity <0.70 (or the lower limit of normal for individuals of the same age, sex, and height), measured prior to and following bronchodilator use [6]. However, it is challenging to implement PFT. It requires specialized spirometry equipment and trained personnel, and participants must repeatedly perform forceful exhalation maneuvers. Older adults or severely ill patients often produce false-negative results due to insufficient effort. In addition, the procedure may induce coughing, dizziness, or other discomforts and poses a risk of cross-infection under pandemic conditions or in poorly controlled environments. These factors limit the application of PFT in community and primary care settings [7]. Thus, relying solely on conventional PFT is insufficient for screening COPD. Developing simpler, non-invasive, and more scalable auxiliary diagnostic methods for early detection of COPD is, therefore, imperative.

In recent years, deep learning (DL) has attracted significant attention in clinical practice. DL is a complex neural network framework. Common DL models include convolutional neural networks, residual networks, densely connected networks, inception networks, and vision transformer models [8]. These models excel at feature extraction and classification, allowing the automatic learning of high-level semantic information from large datasets, thereby markedly improving the precision and efficiency of image processing and signal analysis [9]. Although PFT is recognized as the gold standard for the auxiliary diagnosis of COPD, researchers often employ chest imaging (including computed tomography [CT] scans and X-rays) or respiratory sounds to develop DL-based alternative or complementary tools for improving diagnostic efficiency and convenience. However, these traditional methods heavily rely on researchers' prior knowledge, and variations in diagnostic criteria and annotation practices across different teams result in significant heterogeneity, affecting the reproducibility and generalizability of diagnostic outcomes [10,11]. In this context, some studies have used DL for the automatic diagnosis of COPD, such as DL-based chest X-ray (CXR), for the classification of COPD [10] and DL-based cough sound signal analysis [11]. Nevertheless, systematic evidence of the actual performance and comparative advantages of different DL frameworks in the diagnosis of COPD is lacking.

Therefore, we conducted a systematic review and meta-analysis of diagnostic test accuracy studies on DL models for COPD. Our first objective was to describe the diagnostic performance of these models for identifying COPD across different data sources (such as CT images and respiratory sounds) in both internal and external validation sets. Our second objective was to assess the performance of DL models in classifying the severity of COPD, particularly GOLD stages. We hypothesized that DL models would show good accuracy for the diagnosis of COPD, whereas their performance for staging COPD would be more variable and less stable.

Methods

Study Registration

This systematic review and diagnostic test accuracy meta-analysis was conducted and reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement and the PRISMA-DTA (Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies) extension, and the search methods were reported following PRISMA-S (Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension) [12,13]. The PRISMA-S checklist is provided in Checklist 1. The protocol was prospectively registered in PROSPERO (International Prospective Register of Systematic Reviews; CRD420251114195 [14]).

Eligibility Criteria

The inclusion criteria were as follows: (1) original research that developed a DL model for diagnosing COPD or classifying COPD severity; (2) studies reported at least one of the following outcome measures for appraising the accuracy of DL model: concordance index, the receiver operating characteristic curve, specificity, sensitivity, precision rate, accuracy, recall rate, calibration curve, F_1-score, or confusion matrix; and (3) studies published in English.

Exclusion criteria were as follows: (1) conference abstracts without full-text publication; (2) studies limited to traditional machine learning, without the development of DL models; and (3) studies applying DL solely for image segmentation, without developing models for the diagnosis or classification of COPD. Although a very small number of the included studies may have used data from the same public database, we still included these studies because their DL models incorporated comparable

XSL•FO
RenderX

experimental designs, which helped us better understand the diagnostic performance of DL models for COPD.

Data Sources and Search Strategy

The search methods and reporting were guided by PRISMA-S [12]. Embase, Web of Science, the Cochrane Library, and PubMed were systematically searched from database inception to November 1, 2025. The search strategy was designed by combining medical subject headings and free-text keywords. To maximize the retrieval of relevant studies, no restrictions were applied on language or geographic location. The complete search strategies are provided in Table S1 in Multimedia Appendix 1.

We screened the reference lists of the included studies and relevant reviews; we did not search gray literature or conference proceedings and did not contact authors for additional data. No published search filters were used. Search strategies were developed de novo and were not adapted or reused from prior reviews. We did not conduct a formal peer review of the search strategy.

Study Selection

The retrieved studies were imported into EndNote. Duplicates were automatically and manually removed. Subsequently, the titles and abstracts of the remaining articles were independently reviewed by 2 authors (YH and YW) to identify potentially eligible studies. The full texts of these studies were then assessed to identify eligible studies. Any disagreements at any stage were resolved through discussion with a third reviewer (TW).

Data Extraction

Before data extraction, a standardized extraction form was developed. The collected data encompassed study title, publication year, DOI, country, authors, patient source, study design, task type, COPD diagnostic criteria, imaging modality used for modeling, number of COPD cases, total number of cases, number of COPD cases in the training set, total number of cases in the training set, method for validation set generation, external validation, number of COPD cases in the validation set, total number of cases in the validation set, and comparison with clinicians (yes or no). Two reviewers (HY and TW) independently extracted the data, followed by cross-checking. Any disagreements were addressed through consultation with a third reviewer (YW).

Risk of Bias in Studies

The QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2) tool was utilized to appraise the risk of bias (RoB) of the selected studies. The assessment covered 4 domains: reference standard, index test, patient selection, as well as flow and timing. Each domain included several specific questions, which were answered by “Yes (low RoB),” “No (high RoB),” or “Unclear (RoB uncertain).” The overall RoB for each domain was categorized as low, high, or unclear. The RoB assessment was independently performed by 2 reviewers (YW and TW), and disagreements were addressed through discussion with a third reviewer (HY).

Synthesis Methods

For binary classification tasks, a bivariate mixed effects model was used to pool diagnostic 2×2 contingency tables for DL for the diagnosis of COPD. In studies without complete contingency tables, specificity, sensitivity, negative and positive predictive values, accuracy, and the number of cases were used to estimate the contingency table. Sensitivity, specificity, negative likelihood ratio (NLR), positive likelihood ratio (PLR), diagnostic odds ratio (DOR), and the summary receiver operating characteristic curve with corresponding 95% CIs were pooled. Deeks’ funnel plot was applied to examine the small-study effects of the selected original studies, and clinical applicability was assessed through nomograms. Subgroup analyses by modality (CT, respiratory sounds, or CXR) were performed. All meta-analyses were based on validation set data. If a study reported multiple validation cohorts, each independent validation cohort was included in the analysis separately. If multiple models were evaluated on the same validation cohort, only 1 estimate (ie, the primary and final model reported) was extracted to avoid the nonindependence of the data.

For multiclass classification tasks, the accuracy across different severity grades was pooled. When the reported accuracy approached 99%, a double arcsine transformation was applied before meta-analysis. During the meta-analysis, we utilized the Hartung-Knapp-Sidik-Jonkman modified method [15]. Due to the potential heterogeneity, the 95% prediction intervals for the summary estimates were calculated using the confidence distribution approach proposed by Nagashima et al [16]. All analyses were carried out using STATA (version 15.0; StataCorp LLC) or R (version 4.4.3; R Foundation for Statistical Computing).

Results

Study Selection

Overall, 5194 records were retrieved from databases. After excluding 1958 duplicates, we removed 1695 studies unrelated to the study topic and 492 studies for other reasons. The titles and abstracts of 1049 studies were checked. Among them, 969 studies were removed due to irrelevant or unsuitable study design. The full texts of 80 articles were assessed for eligibility, among which 24 ineligible studies were further excluded. Ultimately, 56 studies [10,17-71] were included (Figure 1).

XSL•FO
RenderX

Figure 1. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart for the systematic review and diagnostic test accuracy meta-analysis, showing the entire process from the database search in Embase, Web of Science, Cochrane Library, and PubMed up to November 1, 2025, to the literature screening and final inclusion of 56 studies on chronic obstructive pulmonary disease (COPD) in adults.

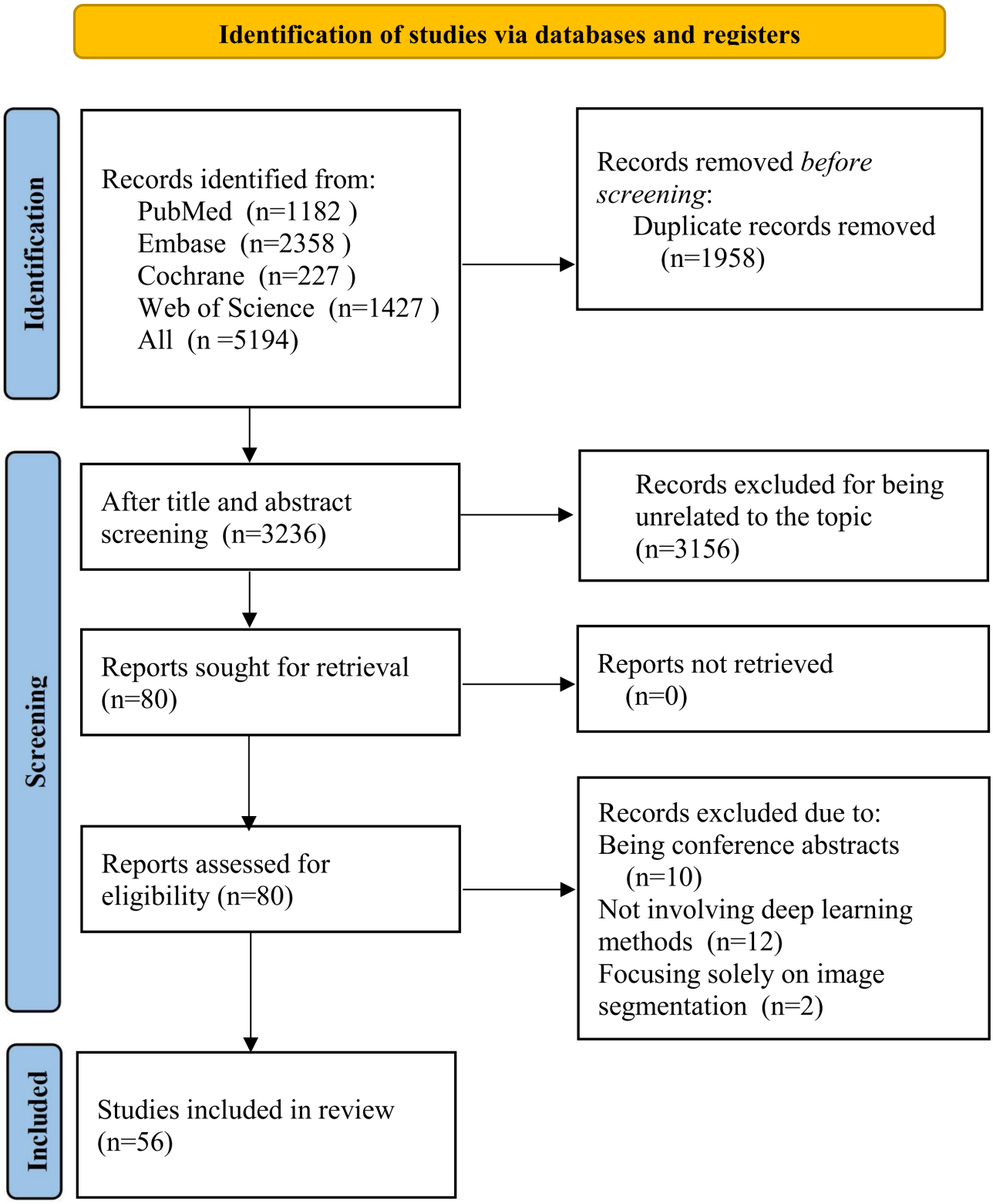

Study Characteristics

The 56 selected studies were published between 2019 and 2025 across 14 countries, with the majority conducted in China (n=21) and the United States (n=11). In terms of study design, there were 39 cohort studies (including retrospective cohort studies), 16 case-control studies, and 1 retrospective cross-sectional diagnostic study. Most datasets were derived from single-center (n=16) or multi-center (n=31) studies, while 9 studies utilized registry databases. Regarding task types, 23 studies focused solely on diagnosis, 17 studies solely on classification, and 16 studies on both diagnosis and classification (out of 16). All studies clearly reported the diagnostic criteria for COPD. The variables of the models primarily came from CT images (30

XSL•FO
RenderX

studies) and breath sound or audio data (12 studies); 4 studies used CXRs (including X-ray films in 2 studies); the remaining 10 studies used other input data (eg, pulmonary function indicators or curves or waveforms, electrocardiograms, volumetric carbon dioxide monitoring, clinical data, imaging-genetic data, or CT-based scores). The total number of cases was 886,753, with 272,881 in the validation sets and 1,352,782 in the training sets. The methods for generating the validation set were categorized as follows: only cross-validation used in 22 studies; only internal validation in 20 studies; external validation in 9 studies; a combination of internal and external validation in 3 studies; and a combination of cross-validation, internal validation, and external validation in 1 study. One study did not report its validation strategy (1 study; Table S2 in Multimedia Appendix 1).

RoB in Studies

In the patient selection domain, all studies employed consecutive or random case selection and applied appropriate exclusion criteria, thereby avoiding including inappropriate cases; therefore, RoB was judged to be low in this domain. For the index test domain, the included studies generally applied supervised DL methods with clearly defined decision rules, and RoB was judged to be mostly low. Regarding the reference standard, all studies used appropriate diagnostic criteria capable of effectively distinguishing COPD and its severity; however, if a study did not explicitly report whether the reference standard assessment was performed blinded to the index test, we rated this item as unclear, leading to an overall judgment of unclear RoB in the reference standard domain for those studies. For the flow and timing domain, RoB was generally low, although incomplete reporting of participant flow and timing resulted in some unclear judgments. In terms of applicability, patient selection was largely consistent with the review question, while a subset of studies raised applicability concerns related to the index test, the reference standard, or both. In addition, some studies reported only summary performance metrics (eg, accuracy) without complete 2×2 contingency tables, which limited transparency for evidence synthesis and introduced uncertainty when reconstructing contingency tables (Figures 2 and 3).

XSL•FO
RenderX

Figure 2. Detailed QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2) risk-of-bias assessment process for the included 56 diagnostic accuracy studies on deep learning (DL) models for chronic obstructive pulmonary disease (COPD) [10,17-19,21,23,24,26-29,31-35,37-41,43-61,63-70,72].

XSL•FO
RenderX

Figure 3. Summary of QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2) risk-of-bias and applicability assessments for the 56 diagnostic accuracy studies on deep learning (DL) models for chronic obstructive pulmonary disease (COPD).

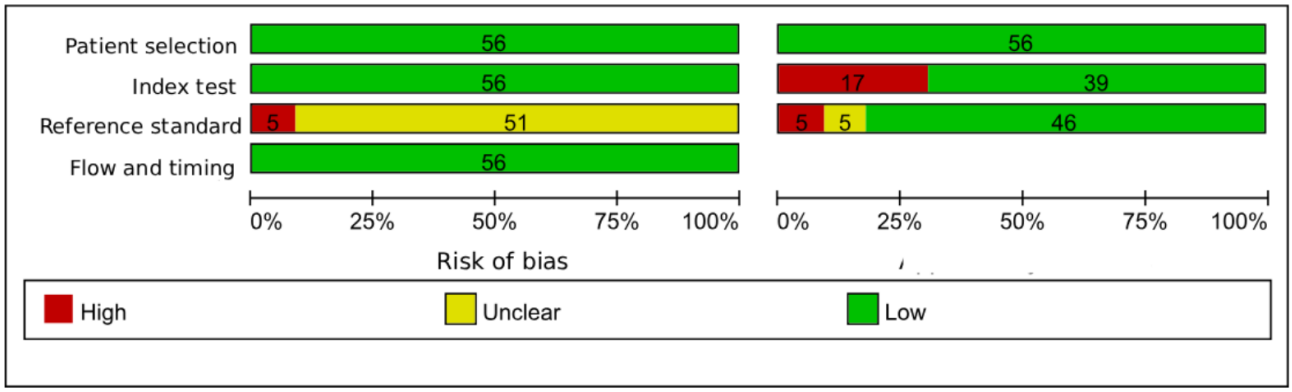

Meta-Analysis of Binary Classification Tasks

Overall

A total of 43 diagnostic 2×2 contingency tables were synthesized to appraise the diagnostic accuracy of DL models for COPD. The pooled results demonstrated that the DL models yielded a sensitivity of 0.87 (95% CI: 0.85 - 0.90), specificity of 0.88 (95% CI 0.84 - 0.92), PLR of 7.4 (95% CI 5.2 - 10.5), NLR of 0.14 (95% CI 0.11 - 0.18), DOR of 52 (95% CI 30 - 88), and the area under the summary receiver operating characteristic curve (AUC) of 0.93 (95% CI 0.18 - 1.00; Figures 4 and 5). Deeks' funnel plot demonstrated no significant small-study effects (*P*=.08; Figure 6). Assuming a pretest probability of 25%, the posttest probability rose to about 71% for a positive result and decreased to about 5% for a negative result, suggesting the potential clinical value of the models in the screening and diagnosis of COPD (Figure 7).

Figure 4. Forest plots of sensitivity and specificity of deep learning (DL) model for binary classification diagnosis of chronic obstructive pulmonary disease (COPD), summarizing 2×2 contingency table results from 43 validation cohorts in 14 countries from 2019 to 2025 [10,11,17,18,20,22,23,25-34,38,40,41,43-45,48-51,53,54,56,61,63,67,68,70,73].

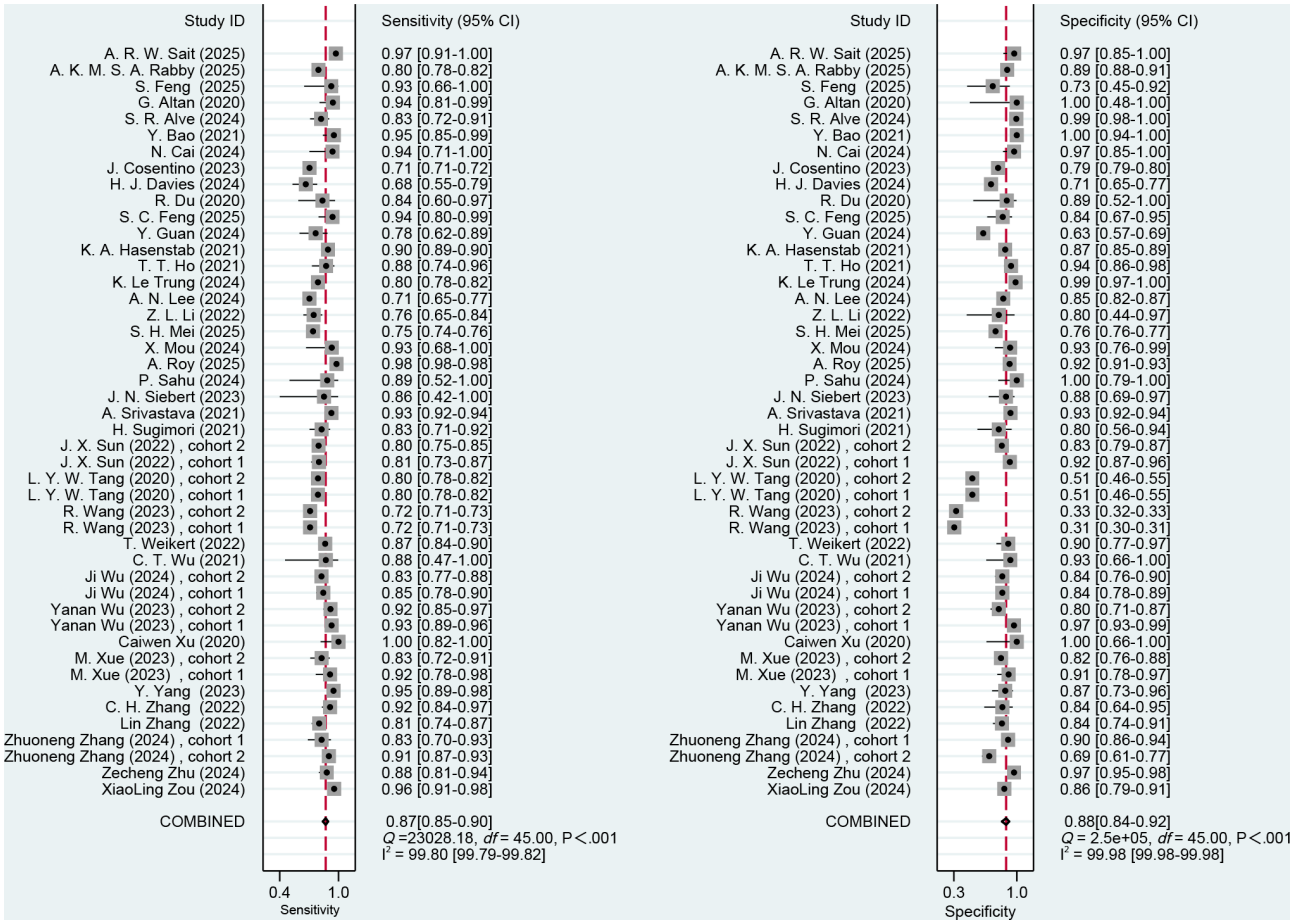

XSL•FO
RenderX

Figure 5. Summary receiver operating characteristic (SROC) curve for the meta-analysis of deep learning (DL) for the diagnosis of chronic obstructive pulmonary disease (COPD) in the validation sets. AUC: area under the summary receiver operating characteristic curve; SENS: sensitivity; SPEC: specificity.

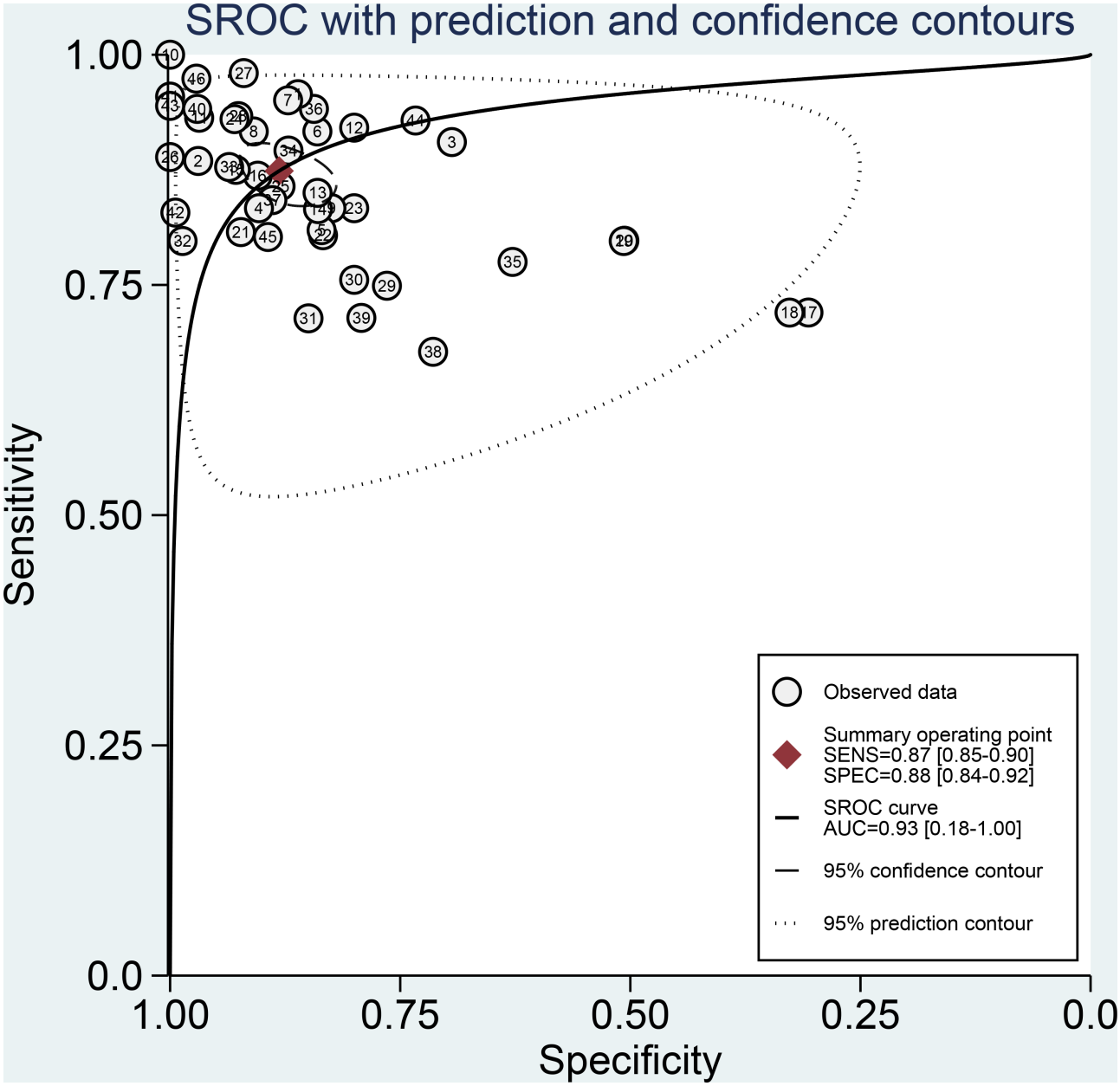

XSL•FO
RenderX

Figure 6. Deeks funnel plot of deep learning (DL) model in the binary classification diagnosis of chronic obstructive pulmonary disease (COPD), assessing publication bias and small-sample effect based on 43 validation cohorts. ESS: effective sample size.

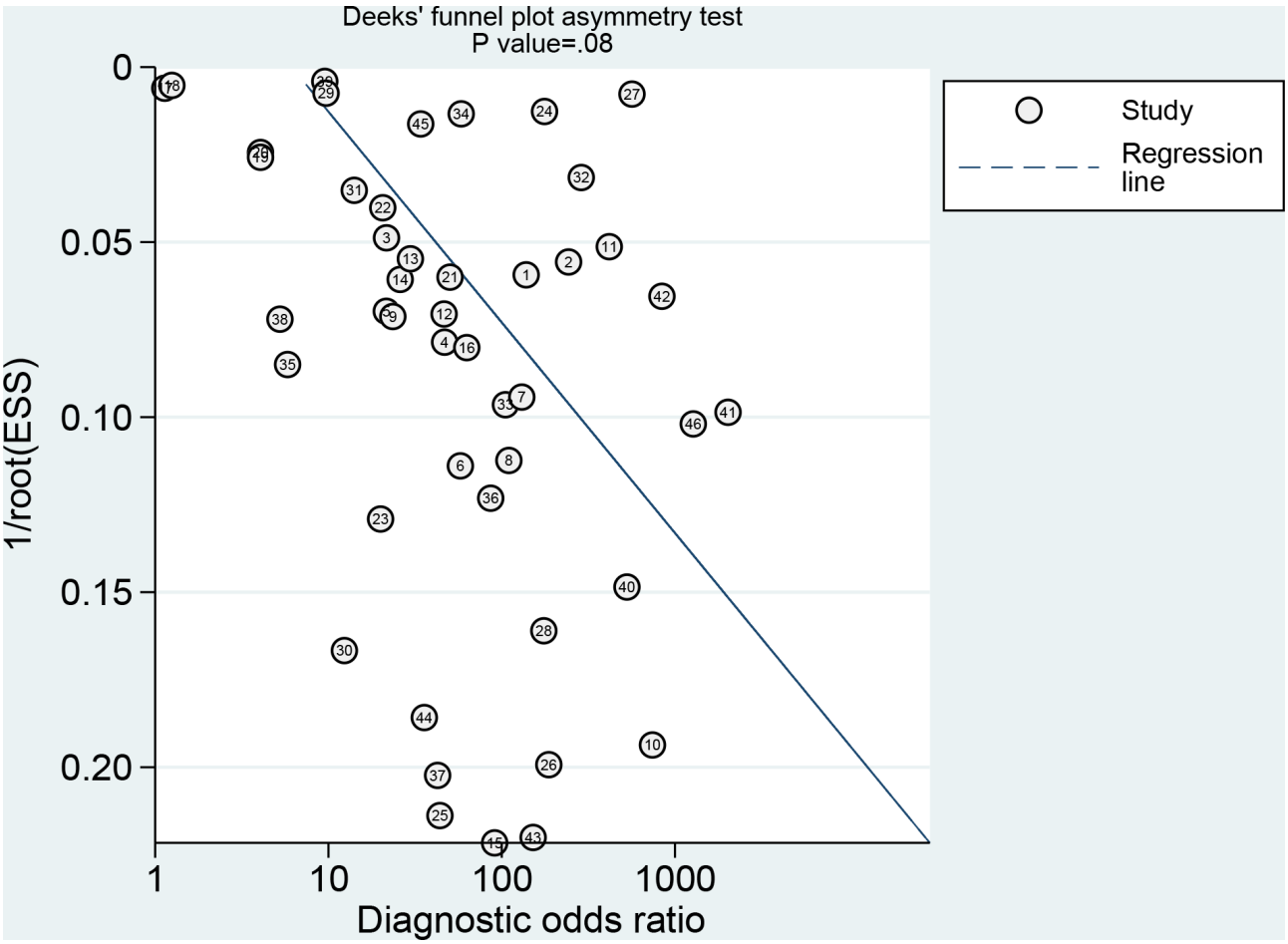

XSL•FO
RenderX

Figure 7. The Fagan cursor plot for deep learning (DL) model in the binary classification diagnosis of chronic obstructive pulmonary disease (COPD), providing the posterior probabilities of positive and negative results with a 25% predetection probability in individuals with suspected COPD based on the pooled likelihood ratio (43 validation cohorts).

XSL•FO
RenderX

DL Based on CT Images

A total of 30 contingency tables were included. The pooled sensitivity of the models was 0.86 (95% CI 0.84 - 0.89), specificity was 0.87 (95% CI 0.82 - 0.90), PLR was 6.6 (95% CI 4.8 - 9.1), NLR was 0.15 (95% CI 0.12 - 0.19), DOR was 42 (95% CI 26 - 68), and AUC was 0.92 (95% CI 0.90 - 0.94; Figures S1-S2 in Multimedia Appendix 1). Deeks' test indicated potential small-study effects (*P*=.02; Figure S3 in Multimedia Appendix 1). Assuming a pretest probability of 25%, the posttest probability rose to 69% for a positive result and decreased to 5% for a negative result (Figure S4 in Multimedia Appendix 1).

Among these, 24 contingency tables were derived from internal validation sets. The pooled sensitivity was 0.86 (95% CI 0.83 - 0.89), specificity was 0.88 (95% CI 0.83 - 0.92), PLR was 7.4 (95% CI 5.0 - 10.9), NLR was 0.16 (95% CI 0.12 - 0.20), DOR was 48 (95% CI 26 - 86), and AUC was 0.93 (95% CI 0.90 - 0.95; Figures S5–S6 in Multimedia Appendix 1). Deeks' funnel plot demonstrated small-study effects (*P*=.04; Figure S7 in Multimedia Appendix 1). Assuming a pretest probability of 25%, the posttest probability increased to about 71% for a positive result and decreased to about 5% for a negative result (Figure S8 in Multimedia Appendix 1). Among studies on CT-based DL, most models incorporated lung parenchymal attenuation patterns related to emphysema. Some additionally incorporated airway and bronchial wall morphology; gas trapping and small-airway abnormalities on inspiratory or expiratory CT; or combined radiomics features of lung parenchyma, airways, and pulmonary vessels.

A total of 8 contingency tables originated from external validation sets. The pooled sensitivity was 0.87 (95% CI 0.82 - 0.90), specificity was 0.83 (95% CI 0.72 - 0.90), PLR was 5.1 (95% CI 2.9 - 8.8), NLR was 0.16 (95% CI 0.11 - 0.23), DOR was 31 (95% CI 14 - 72), and AUC was 0.91 (95% CI 1.00 - 0.00; Figures S9-S10 in Multimedia Appendix 1). Deeks' funnel plot demonstrated no small-study effects (*P*=.06; Figure S11 in Multimedia Appendix 1). Assuming a pretest probability of 25%, the posttest probability rose to about 63% for a positive result and decreased to about 5% for a negative result (Figure S12 in Multimedia Appendix 1).

To further evaluate potential small-study effects, we additionally stratified the CT-based validation cohorts by the number of COPD cases in the validation set. A total of 15 cohorts were classified as a small-sample subgroup (COPD cases <100) and 17 cohorts as a large-sample subgroup (COPD cases ≥100). In the small-sample subgroup, the pooled sensitivity and specificity were 0.89 (95% CI 0.84 - 0.92) and 0.89 (95% CI 0.83 - 0.94), respectively, with an AUC of 0.94 (95% CI 0.92 - 0.96). In the large-sample subgroup, the pooled sensitivity and specificity were slightly lower at 0.85 (95% CI 0.82 - 0.88) and 0.85 (95% CI 0.78 - 0.90), respectively, with an AUC of 0.91 (95% CI 0.88 - 0.93; Figures S13-S16 in Multimedia Appendix 1). Assuming a pretest probability of 25%, the Fagan nomograms indicated that the posttest probability increased to 74% for a positive DL result in the small-sample studies and 65% in the large-sample studies, while it reduced to 4% and 6% for a negative result, respectively (Figures S17 and S18 in Multimedia Appendix 1). Deeks' funnel plot asymmetry tests for the small- and large-sample subgroups were not statistically significant (*P*=.34 and *P*=.15, respectively; Figures S19 and S20 in Multimedia Appendix 1), suggesting no strong evidence of small-study effects. However, given the consistently higher point estimates in the small-sample subgroup, some degree of small-study effects cannot be completely ruled out.

DL Based on Respiratory Sounds

A total of 10 contingency tables were included. The pooled sensitivity was 0.91 (95% CI 0.84 - 0.95), specificity was 0.96 (95% CI 0.91 - 0.98), PLR was 22.1 (95% CI 9.5 - 51.5), NLR was 0.09 (95% CI 0.05 - 0.18), DOR was 237 (95% CI 78 - 723), and AUC was 0.98 (95% CI 0.96 - 0.99; Figures S21-S22 in Multimedia Appendix 1). Deeks' funnel plot demonstrated no small-study effects (*P*=.32; Figure S23 in Multimedia Appendix 1). With a pretest probability of 25%, the posttest probability rose to about 88% following a positive result and decreased to about 3% following a negative result (Figure S24 in Multimedia Appendix 1). For respiratory sound–based DL models, lung sounds were recorded using electronic or digital stethoscopes at standard chest auscultation sites or obtained from open respiratory sound databases (eg, RespiratoryDatabase@TR and other multichannel lung sound datasets) and analyzed as single- or multichannel signals.

Summary of DL Based on CXR

Only 2 included studies evaluated DL models based on CXR for the diagnosis of COPD. In a multicenter study, Zou et al [10] constructed a DL model integrating CXR images and clinical parameters. This model achieved favorable performance in internal validation with a sensitivity of 0.96 and a specificity of 0.86. Conversely, Wang et al [29] constructed a model solely based on CXR images. Their model yielded a sensitivity of 0.72 and specificity of 0.31 in the MIMIC-CXR internal validation set and a sensitivity of 0.72 and specificity of 0.33 in the Emory-CXR external validation set. These findings suggest that combining clinical parameters with imaging data may substantially enhance diagnostic performance, whereas single-image models exhibit limited specificity.

Summary of DL Based on Externally Applied Airway Resistance

In the study by Davies [54], a physical simulation device was utilized to generate surrogate data for training a DL model. Tubes of varying diameters (3 - 25 mm) were installed in the respiratory tract of healthy participants to independently modulate inspiratory and expiratory resistance, thereby simulating COPD-related obstruction. Based on the generated photoplethysmography signals, a 1D convolutional neural network achieved an AUC of 0.75 in the binary classification of COPD and healthy controls. The accuracy of the model reached 40% - 88% for real COPD cases, with a 14% misdiagnosis rate in healthy participants. This approach may offer a low-cost alternative for data-scarce scenarios, particularly suitable for screening with wearable devices in primary care. However, since dynamic resistance simulation was limited, and the sample size for validation was small (only 4 patients), the model needs to be further optimized.

XSL•FO
RenderX

Multiclass DL for COPD Grading

A total of 6 studies [10,22,31,32,40,44] developed DL models for GOLD grading of COPD (multiclass classification). Among these studies, 5 developed models based on CT images, while Zou et al [10] used CXR images for modeling. Most studies applied different GOLD classification strategies. Several studies [10,31,40,44] implemented 5-class classification (GOLD 0 - 4). In another analysis by Zou [10], a 3-class strategy was applied (GOLD 0, GOLD 1 - 2, GOLD 3 - 4). Sugimori [32] and Yang [22] employed a 4-class strategy (GOLD 0, 1, 2, 3 - 4).

Overall analysis indicated considerable differences in the accuracy of the DL models for identifying each GOLD stage, reflecting substantial heterogeneity in model performance. The pooled results based on a random-effects model were as follows: the diagnostic accuracy was 0.842 (95% CI 0.605 - 0.982) for GOLD 0, 0.617 (95% CI 0.407 - 0.808) for GOLD 1, 0.679 (95% CI 0.376 - 0.917) for GOLD 2, 0.708 (95% CI 0.163 - 1.000) for GOLD 3, and 0.708 (95% CI 0.163 - 1.000) for GOLD 4 (Figure S25 in Multimedia Appendix 1). These findings demonstrated that the DL models were unstable in the identification of mild (GOLD 1) and very severe (GOLD 4) stages. Given the wide CIs, the diagnostic accuracy was still limited.

Discussion

Summary of the Main Findings

Current DL models for detecting COPD are primarily constructed based on CT imaging and respiratory sound data. The tasks are generally divided into binary and multiclass classifications. Our findings suggested that in binary classification tasks, the CT-based models performed well in internal validation cohorts, with a pooled sensitivity of 0.86 (95% CI 0.83 - 0.89) and specificity of 0.88 (95% CI 0.83 - 0.92). The models based on respiratory sounds yielded a sensitivity of 0.91 (95% CI 0.84 - 0.95) and specificity of 0.96 (95% CI 0.91 - 0.98), indicating a strong exclusion ability.

In multiclass classification tasks, the included studies mainly focused on the staging of GOLD. Overall analysis demonstrated that the DL models were unstable for discriminating between different GOLD stages. This finding supports our hypothesis that compared with binary diagnosis, the accuracy and reliability of the DL models for staging COPD still need to be improved.

Comparison With Previous Reviews

Prior studies have examined the application of CT and respiratory sounds in the diagnosis of COPD. The systematic review and network meta-analysis carried out by Balasubramanian et al [74] focuses on the diagnostic performance of CT-guided transthoracic biopsy or fine-needle aspiration in lung diseases, particularly lung cancer. Their study included 363 studies involving 79,519 patients and reported a pooled sensitivity of 88.9% but did not address the use of CT in the diagnosis of COPD. In addition, Arts et al [75] have evaluated the use of respiratory sounds for diagnosing acute pulmonary diseases. Their results demonstrate that respiratory sounds have a sensitivity of 37% (95% CI 30% - 47%) and specificity of 89% (95% CI 85% - 92%) for diagnosing COPD, based on approximately 12 relevant studies [75]. Willer et al [73] have examined the performance of X-ray dark-field imaging in detecting and evaluating emphysema in patients with COPD. Their study includes 77 patients and reports that this imaging modality exhibits high diagnostic performance for emphysema (correlation coefficient ρ=0.62, $P<.0001$) and is closely associated with microstructural changes in the lung. These findings suggest that dark-field chest imaging may be a rapid, low-dose, and sensitive tool for the screening and assessment of COPD. However, their study does not evaluate the diagnostic accuracy of conventional CXR for COPD.

In contrast, this meta-analysis reported higher diagnostic performance of the DL models based on CT imaging and respiratory sounds. The pooled results demonstrated that the DL models based on CT yielded a sensitivity of 0.86 (95% CI 0.84 - 0.89) and specificity of 0.87 (95% CI 0.82 - 0.90), while respiratory sound–based models yielded a sensitivity of 0.91 (95% CI 0.84 - 0.95) and specificity of 0.96 (95% CI 0.91 - 0.98). These results suggest that DL approaches might outperform traditional diagnostic methods. Earlier research has also investigated the role of artificial intelligence (AI) in COPD diagnosis. For instance, Wu et al [72] examined the potential of machine learning and DL in the detection, staging, and quantitative analysis of COPD using CT imaging. However, their review does not clearly differentiate between machine learning and DL, nor does it discuss in depth the advantages and limitations of image-based AI models for the diagnosis of COPD.

This study found that the included studies on DL for diagnosing COPD focused mainly on CT imaging, respiratory sounds, CXR, and externally applied airway resistance. Among these, CT, respiratory sounds, and CXR were the most frequently used data sources for model development and carried distinct clinical implications. Chest CT exerts a crucial role in diagnosing and phenotyping COPD, as it can identify structural abnormalities, such as airway narrowing and emphysema, and is recommended by current clinical guidelines. Our findings demonstrated that the CT-based DL models offered excellent specificity and sensitivity for the diagnosis of COPD, suggesting their potential as auxiliary diagnostic tools in clinical practice. The DL models based on respiratory sounds, as a non-invasive and portable modality, also had good diagnostic performance, particularly with high specificity, indicating potential value in primary screening. In contrast, the number of studies using CXR remains limited, and the existing evidence is insufficient to determine the stability and generalizability of CXR-based DL models for diagnosing COPD. It should be validated in the future. Moreover, although a few preliminary studies have explored the use of externally applied airway resistance to generate model inputs, the number of studies remains small, and reproducible, generalizable evidence is lacking. Thus, future studies are required to assess the utility and reliability of this approach in clinical practice.

Despite the promise of AI in the diagnosis of COPD, significant challenges need to be addressed before widespread clinical application, particularly in explainability and data integration. Although current research demonstrates encouraging diagnostic performance, a substantial gap persists between theoretical

XSL•FO
RenderX

development and real-world application. First, most included studies did not thoroughly examine how variations in imaging protocols, such as scan parameters or reconstruction algorithms, influence image features and the performance of DL models. Hence, a systematic evaluation of these factors is lacking. Second, as complex neural network frameworks, DL models rely on large-scale training datasets to improve robustness. However, most included studies developed models using limited samples, with only a few utilizing large datasets. The scarcity of data represents a core bottleneck in model development, constraining the generalizability of the models. Future studies should incorporate richer and more diverse imaging data. Third, the current model evaluation primarily relied on internal validation techniques, such as random sampling, cross-validation, or bootstrap methods. While internal validation sets share similar distributions with training data and often yield favorable results, they do not accurately reflect the generalizability of the models on heterogeneous datasets. Models should be rigorously externally validated before real-world application, particularly across institutions and using datasets obtained under different imaging protocols. Studies based on high-quality external validation remain scarce, and substantial differences in imaging protocols make it challenging to interpret model performance in external validation.

In clinical research and practice, grading disease severity is as crucial as diagnosing COPD. The widely applied GOLD classification, which stratifies COPD into 5 grades (0 and 1 - 4), reflects significant differences in clinical presentation, treatment strategies, and prognosis of COPD. Achieving early and precise grading is therefore of high clinical relevance. However, only 6 studies have attempted to develop DL models for grading the severity of COPD, providing limited evidence. These studies indicate that DL models generally perform suboptimally in multiclass classification tasks, with particularly low accuracy for GOLD 1, GOLD 2, and GOLD 4. These models achieve relatively higher accuracy only for GOLD 0 and GOLD 3, exceeding 70%. Nevertheless, their stability still needs to be enhanced. This suggests that multiclass classification itself represents a technical challenge for DL models. Moreover, under the current dataset size, label distribution, and model architecture, stable differentiation across all GOLD grades remains difficult. Future research should aim to enhance the discriminative ability of models, incorporate richer imaging data, and integrate clinical information to optimize training strategies, ultimately developing more accurate and adaptable intelligent tools for grading the severity of COPD to support clinical decision-making.

Strengths and Limitations of the Study

This meta-analysis systematically assessed the performance of DL in the detection of COPD for the first time, providing evidence to support the development of intelligent diagnostic tools. The findings indicate that DL models hold substantial potential for improving diagnostic accuracy, particularly through noninvasive and nonintrusive detection methods. This study provides valuable insights. However, some limitations must be noted. First, although a systematic literature search was carried out, the number of studies focusing on respiratory sounds remained relatively small. As respiratory sound analysis is an emerging diagnostic approach, the number and diversity of relevant studies remain far below those of CT imaging, which may limit a comprehensive assessment of this method. Second, most included studies relied primarily on internal validation, and only relatively few studies performed external validation. Although internal validation can provide some indication of diagnostic accuracy, limitations in sample size and validation methods may compromise the generalizability of the results. To further confirm the clinical utility of DL models, future studies should perform external validation. Third, research on the severity of COPD was relatively scarce, and some studies employed differing grading strategies. These variations may affect the reliability of classification models and the generalizability of their findings. Thus, this finding should be cautiously interpreted.

Heterogeneity and Clinical Applicability of DL Models

Although subgroup analyses were performed to explore the source of heterogeneity, significant heterogeneity still existed among the subgroups. This heterogeneity may stem from differences in DL frameworks used in different studies, such as 2D or 3D convolutional neural networks, multiview networks, multi-instance learning, and late fusion. The included studies used diverse DL models, which differed in network structure, input format, and parameter settings. Consequently, their model training and validation methods may also differ. Therefore, these differences in structure and parameters can lead to potential heterogeneity, which is a common challenge in current meta-analyses of DL models.

From the perspective of clinical practicality, DL still holds significant advantages over traditional radiomics. Traditional radiomics typically requires manual or semiautomatic image segmentation, followed by the extraction of a limited number of manual features, such as texture. An original image is compressed into a small number of quantitative features, then to a machine learning model. This multistep process is time-consuming, highly dependent on the researcher's experience, and may lose some original image information during dimensionality reduction and feature selection. DL, on the other hand, can directly train models end-to-end based on labeled (or segmented) images without additional feature engineering. It can preserve lesion-related image information to the greatest extent, potentially improving model performance and reducing manual operations and time costs. Therefore, given the relatively ideal diagnostic and grading accuracy of DL models, it is hoped that AI-assisted diagnostic DL tools should be developed to support, rather than replace, clinicians in screening and assessing the severity of COPD.

Future Perspectives

Most current studies are based on relatively limited imaging datasets and rely mainly on internal validation. Thus, the reported accuracy may not fully reflect the generalizability of models. Given substantial between-study heterogeneity and limited external validation, these findings should be interpreted cautiously. Future research should improve and update these DL models by using larger, multicenter imaging datasets from different geographical regions and scanners, and by

XSL•FO
RenderX

incorporating robust external validation and more rigorous model development strategies.

To our knowledge, this is the first systematic synthesis to quantify the diagnostic and grading performance of DL models across major data sources (eg, CT imaging and respiratory sounds), showing promising accuracy for binary COPD detection but suboptimal and less stable performance for multiclass GOLD staging.

In summary, our comprehensive study on DL provides an evidence base for guiding the development and external validation of AI-assisted screening tools for COPD, especially given the insufficient application of spirometry.

Conclusions

This study observed that DL models achieved promising accuracy in the detection of COPD. The models performed particularly well in binary classification tasks, exhibiting high sensitivity and specificity. However, its accuracy was suboptimal in multiclass tasks for grading the severity of GOLD. In addition, research on respiratory sound analysis and multiclass classification of COPD severity is still limited. Given the substantial heterogeneity and limited external validation, these results should be interpreted cautiously. Thus, future research should integrate larger and more diverse imaging datasets, particularly including images from different racial populations, to develop more robust and generalizable intelligent diagnostic tools. This approach would not only enhance the generalizability of models but also improve the accuracy of diagnosing COPD across diverse patient groups.

Funding

This research was not supported by any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: HY
Data curation: YW
Methodology: HY
Software: HY, JJ, SL
Supervision: TW
Validation: JJ, SL
Writing – original draft preparation: HY, YW
Writing – reviewing and editing: WX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials (search strategies, study characteristics, and additional analyses).
[DOCX File, 36906 KB - jmir_v28i1e83459_app1.docx]

Checklist 1
PRISMA-S checklist.
[PDF File, 66 KB - jmir_v28i1e83459_app2.pdf]

References

1. Agustí A, Celli BR, Criner GJ, et al. Global Initiative for Chronic Obstructive Lung Disease 2023 Report: GOLD executive summary. Eur Respir J 2023 Apr;61(4):2300239. [doi: 10.1183/13993003.00239-2023] [Medline: 36858443]
2. Al Wachami N, Guennouni M, Iderdar Y, et al. Estimating the global prevalence of chronic obstructive pulmonary disease (COPD): a systematic review and meta-analysis. BMC Public Health 2024 Jan 25;24(1):297. [doi: 10.1186/s12889-024-17686-9] [Medline: 38273271]
3. Xu J, Ji Z, Zhang P, Chen T, Xie Y, Li J. Disease burden of COPD in the Chinese population: a systematic review. Ther Adv Respir Dis 2023;17:17534666231218899. [doi: 10.1177/17534666231218899] [Medline: 38146618]
4. Agustí A, Celli BR, Criner GJ, et al. Global Initiative for Chronic Obstructive Lung Disease 2023 Report: GOLD executive summary. Am J Respir Crit Care Med 2023 Apr 1;207(7):819-837. [doi: 10.1164/rccm.202301-0106PP] [Medline: 36856433]

XSL•FO
RenderX

5. Takano T, Tsubouchi K, Hamada N, et al. Update of prognosis and characteristics of chronic obstructive pulmonary disease in a real-world setting: a 5-year follow-up analysis of a multi-institutional registry. BMC Pulm Med 2024 Nov 6;24(1):556. [doi: 10.1186/s12890-024-03347-5] [Medline: 39506773]
6. Singh D, Stockley R, Anzueto A, et al. GOLD Science Committee recommendations for the use of pre- and post-bronchodilator spirometry for the diagnosis of COPD. Eur Respir J 2025 Feb;65(2):2401603. [doi: 10.1183/13993003.01603-2024] [Medline: 39638416]
7. Baldomero AK, Kunisaki KM, Bangerter A, et al. Beyond access: factors associated with spirometry underutilization among patients with a diagnosis of COPD in urban tertiary care centers. Chronic Obstr Pulm Dis 2022 Oct 26;9(4):538-548. [doi: 10.15326/jcopdf.2022.0303] [Medline: 36040836]
8. Han K, Wang Y, Chen H, et al. A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell 2023 Jan;45(1):87-110. [doi: 10.1109/TPAMI.2022.3152247] [Medline: 35180075]
9. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017 Dec;42:60-88. [doi: 10.1016/j.media.2017.07.005] [Medline: 28778026]
10. Zou X, Ren Y, Yang H, et al. Screening and staging of chronic obstructive pulmonary disease with deep learning based on chest X-ray images and clinical parameters. BMC Pulm Med 2024 Mar 26;24(1):153. [doi: 10.1186/s12890-024-02945-7] [Medline: 38532368]
11. Zhang P, Swaminathan A, Uddin AA. Pulmonary disease detection and classification in patient respiratory audio files using long short-term memory neural networks. Front Med (Lausanne) 2023;10:1269784. [doi: 10.3389/fmed.2023.1269784] [Medline: 38020156]
12. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
13. McInnes MDF, Moher D, Thombs BD, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA 2018 Jan 23;319(4):388-396. [doi: 10.1001/jama.2017.19163] [Medline: 29362800]
14. Yang H, Wu Y, Wu T. Accuracy of deep learning in diagnosing COPD: a systematic review and meta-analysis. Centre for Reviews and Dissemination. URL: https://www.crd.york.ac.uk/PROSPERO/view/CRD420251114195 [accessed 2025-12-25]
15. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
16. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019 Jun;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]
17. Zhu Z, Zhao S, Li J, et al. Development and application of a deep learning-based comprehensive early diagnostic model for chronic obstructive pulmonary disease. Respir Res 2024 Apr 18;25(1):167. [doi: 10.1186/s12931-024-02793-3] [Medline: 38637823]
18. Zhang Z, Wu F, Zhou Y, et al. Detection of chronic obstructive pulmonary disease with deep learning using inspiratory and expiratory chest computed tomography and clinical information. J Thorac Dis 2024 Sep 30;16(9):6101-6111. [doi: 10.21037/jtd-24-367] [Medline: 39444883]
19. Zhang L, Jiang B, Wisselink HJ, Vliegenthart R, Xie X. COPD identification and grading based on deep learning of lung parenchyma and bronchial wall in chest CT images. Br J Radiol 2022 May 1;95(1133):20210637. [doi: 10.1259/bjr.20210637] [Medline: 35143286]
20. Zhang C, Liu J, Cao L, et al. Deep learning-based computed tomography features in evaluating early screening and risk factors for chronic obstructive pulmonary disease. Contrast Media Mol Imaging 2022;2022(1):5951418. [doi: 10.1155/2022/5951418] [Medline: 36051929]
21. Ying J, Dutta J, Guo N, et al. Classification of exacerbation frequency in the COPDGene Cohort Using Deep Learning With Deep Belief Networks. IEEE J Biomed Health Inform 2020 Jun;24(6):1805-1813. [doi: 10.1109/JBHI.2016.2642944] [Medline: 28026794]
22. Yang Y, Zeng N, Chen Z, et al. Multi-layer perceptron classifier with the proposed combined feature vector of 3D CNN Features and lung radiomics features for COPD stage classification. J Healthc Eng 2023;2023:3715603. [doi: 10.1155/2023/3715603] [Medline: 37953910]
23. Xue M, Jia S, Chen L, Huang H, Yu L, Zhu W. CT-based COPD identification using multiple instance learning with two-stage attention. Comput Methods Programs Biomed 2023 Mar;230. [doi: 10.1016/j.cmpb.2023.107356] [Medline: 36682106]
24. Xu C, Qi S, Feng J, et al. DCT-MIL: deep CNN transferred multiple instance learning for COPD identification using CT images. Phys Med Biol 2020 Jul 22;65(14). [doi: 10.1088/1361-6560/ab857d] [Medline: 32235077]
25. Wu Y, Du R, Feng J, et al. Deep CNN for COPD identification by multi-view snapshot integration of 3D airway tree and lung field. Biomed Signal Process Control 2023 Jan;79:104162. [doi: 10.1016/j.bspc.2022.104162]
26. Wu J, Lu Y, Dong S, Wu L, Shen X. Predicting COPD exacerbations based on quantitative CT analysis: an external validation study. Front Med (Lausanne) 2024;11:1370917. [doi: 10.3389/fmed.2024.1370917] [Medline: 38933101]

XSL•FO
RenderX

27. Wu CT, Li GH, Huang CT, et al. Acute exacerbation of a chronic obstructive pulmonary disease prediction system using wearable device data, machine learning, and deep learning: development and cohort study. JMIR Mhealth Uhealth 2021 May 6;9(5):e22591. [doi: 10.2196/22591] [Medline: 33955840]
28. Weikert T, Friebe L, Wilder-Smith A, et al. Automated quantification of airway wall thickness on chest CT using retina U-Nets—performance evaluation and application to a large cohort of chest CTs of COPD patients. Eur J Radiol 2022 Oct;155:110460. [doi: 10.1016/j.ejrad.2022.110460] [Medline: 35963191]
29. Wang R, Chen LC, Moukheiber L, et al. Enabling chronic obstructive pulmonary disease diagnosis through chest X-rays: a multi-site and multi-modality study. Int J Med Inform 2023 Oct;178:105211. [doi: 10.1016/j.ijmedinf.2023.105211] [Medline: 37690225]
30. Tang LYW, Coxson HO, Lam S, Leipsic J, Tam RC, Sin DD. Towards large-scale case-finding: training and validation of residual networks for detection of chronic obstructive pulmonary disease using low-dose CT. Lancet Digit Health 2020 May;2(5):e259-e267. [doi: 10.1016/S2589-7500(20)30064-9] [Medline: 33328058]
31. Sun J, Liao X, Yan Y, et al. Detection and staging of chronic obstructive pulmonary disease using a computed tomography-based weakly supervised deep learning approach. Eur Radiol 2022 Aug;32(8):5319-5329. [doi: 10.1007/s00330-022-08632-7] [Medline: 35201409]
32. Sugimori H, Shimizu K, Makita H, Suzuki M, Konno S. A comparative evaluation of computed tomography images for the classification of spirometric severity of the chronic obstructive pulmonary disease with deep learning. Diagnostics (Basel) 2021 May 21;11(6):929. [doi: 10.3390/diagnostics11060929] [Medline: 34064240]
33. Srivastava A, Jain S, Miranda R, Patil S, Pandya S, Kotecha K. Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease. PeerJ Comput Sci 2021;7:e369. [doi: 10.7717/peerj-cs.369] [Medline: 33817019]
34. Siebert JN, Hartley MA, Courvoisier DS, et al. Deep learning diagnostic and severity-stratification for interstitial lung diseases and chronic obstructive pulmonary disease in digital lung auscultations and ultrasonography: clinical protocol for an observational case-control study. BMC Pulm Med 2023 Jun 2;23(1):191. [doi: 10.1186/s12890-022-02255-w] [Medline: 37264374]
35. Sharma J, Vaid A, Nadkarni G, Kraft M. Diagnosis of chronic obstructive pulmonary disease using deep-learning on electrocardiograms. 2024 May Presented at: American Thoracic Society 2024 International Conference; May 17-22, 2024. [doi: 10.1164/ajrccm-conference.2024.209.1_MeetingAbstracts.A4522]
36. Seastedt KP, Litchman T, Moukheiber L, et al. Predicting chronic obstructive pulmonary disease from chest x-rays using deep learning. 2022 May Presented at: American Thoracic Society 2022 International Conference; May 13-18, 2022. [doi: 10.1164/ajrccm-conference.2022.205.1_MeetingAbstracts.A1078]
37. Sahu P, Kumar S, Behera AK. SOUNDNet: leveraging deep learning for the severity classification of chronic obstructive pulmonary disease based on lung sound analysis. Presented at: 2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT); Jul 12-14, 2024. [doi: 10.1109/CONECCT62155.2024.10677193]
38. Roy A, Gyanchandani B, Oza A, Singh A. TriSpectraKAN: a novel approach for COPD detection via lung sound analysis. Sci Rep 2025 Feb 21;15(1):6296. [doi: 10.1038/s41598-024-82781-1] [Medline: 39984500]
39. Nallanthighal VS, Harma A, Strik H. Detection of COPD exacerbation from speech: comparison of acoustic features and deep learning based speech breathing models. Presented at: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 23-27, 2022. [doi: 10.1109/ICASSP43922.2022.9747785]
40. Mou X, Wang P, Sun J, et al. A novel approach for the detection and severity grading of chronic obstructive pulmonary disease based on transformed volumetric capnography. Bioengineering (Basel) 2024 May 23;11(6):530. [doi: 10.3390/bioengineering11060530] [Medline: 38927766]
41. Mei S, Li X, Zhou Y, et al. Deep learning for detecting and early predicting chronic obstructive pulmonary disease from spirogram time series. NPJ Syst Biol Appl 2025 Feb 15;11(1):18. [doi: 10.1038/s41540-025-00489-y] [Medline: 39955293]
42. Makimoto K, Hague CJ, Hogg JC, Bourbeau J, Tan WC, Kirby M. Chronic obstructive pulmonary disease classification using sex-specific machine learning models and quantitative computed tomography imaging. 2024 May Presented at: American Thoracic Society 2024 International Conference; May 17-22, 2024. [doi: 10.1164/ajrccm-conference.2024.209.1_MeetingAbstracts.A5121]
43. Li Z, Huang K, Liu L, Zhang Z. Early detection of COPD based on graph convolutional network and small and weakly labeled data. Med Biol Eng Comput 2022 Aug;60(8):2321-2333. [doi: 10.1007/s11517-022-02589-x] [Medline: 35750976]
44. Lee AN, Hsiao A, Hasenstab KA. Evaluating the cumulative benefit of inspiratory CT, expiratory CT, and clinical data for COPD diagnosis and staging through deep learning. Radiol Cardiothorac Imaging 2024 Dec;6(6):e240005. [doi: 10.1148/ryct.240005] [Medline: 39665633]
45. Le Trung K, Nguyen Anh P, Han TT. A novel method in COPD diagnosing using respiratory signal generation based on CycleGAN and machine learning. Comput Methods Biomech Biomed Engin 2025 Jul;28(9):1538-1553. [doi: 10.1080/10255842.2024.2329938] [Medline: 38551327]
46. Doroodgar Jorshery S, Chandra J, Walia AS, et al. Leveraging deep learning of chest radiograph images to identify individuals at high risk for chronic obstructive pulmonary disease. medRxiv. 2024 Nov 15 p. 2024-11. [doi: 10.1101/2024.11.14.24317055] [Medline: 39606360]

XSL•FO
RenderX

47. Iturrioz Campo M, Nardelli P, Koc S, Bradford K, Krishnamurthy AK, San Jose Estepar R. COPD stage classification using deep learning on NHLBI biodata catalyst. 2022 May Presented at: American Thoracic Society 2022 International Conference; May 13-18, 2022. [doi: 10.1164/ajrccm-conference.2022.205.1_MeetingAbstracts.A1087]
48. Ho TT, Kim T, Kim WJ, et al. A 3D-CNN model with CT-based parametric response mapping for classifying COPD subjects. Sci Rep 2021 Jan 8;11(1):34. [doi: 10.1038/s41598-020-79336-5] [Medline: 33420092]
49. Hasenstab KA, Yuan N, Retson T, et al. Automated CT staging of chronic obstructive pulmonary disease severity for predicting disease progression and mortality with a deep learning convolutional neural network. Radiol Cardiothorac Imaging 2021 Apr;3(2):e200477. [doi: 10.1148/ryct.2021200477] [Medline: 33969307]
50. Guan Y, Zhang D, Zhou X, et al. Comparison of deep-learning and radiomics-based machine-learning methods for the identification of chronic obstructive pulmonary disease on low-dose computed tomography images. Quant Imaging Med Surg 2024 Mar 15;14(3):2485-2498. [doi: 10.21037/qims-23-1307] [Medline: 38545077]
51. Feng S, Zhang R, Zhang W, et al. Predicting acute exacerbation phenotype in chronic obstructive pulmonary disease patients using VGG-16 deep learning. Respiration 2025;104(1):1-14. [doi: 10.1159/000540383] [Medline: 39047695]
52. El Boueiz AR, Dy JG, Ross JC, et al. Deep learning prediction of COPD progression using enriched densitometry phenotypes. 2019 May Presented at: American Thoracic Society 2019 International Conference; May 17-22, 2019. [doi: 10.1164/ajrccm-conference.2019.199.1_MeetingAbstracts.A4054]
53. Du R, Qi S, Feng J, et al. Identification of COPD from multi-view snapshots of 3D lung airway tree via deep CNN. IEEE Access 2020;8:38907-38919. [doi: 10.1109/ACCESS.2020.2974617]
54. Davies HJ, Hammour G, Xiao H, et al. Physically meaningful surrogate data for COPD. IEEE Open J Eng Med Biol 2024;5:148-156. [doi: 10.1109/OJEMB.2024.3360688] [Medline: 38487098]
55. D Almeida S, Norajitra T, Lüth CT, et al. How do deep-learning models generalize across populations? Cross-ethnicity generalization of COPD detection. Insights Imaging 2024 Aug 7;15(1):198. [doi: 10.1186/s13244-024-01781-x] [Medline: 39112910]
56. Cosentino J, Behsaz B, Alipanahi B, et al. Inference of chronic obstructive pulmonary disease with deep learning on raw spirograms identifies new genetic loci and improves risk models. Nat Genet 2023 May;55(5):787-795. [doi: 10.1038/s41588-023-01372-4]
57. Christina Dally E, Banu Rekha B. Automated chronic obstructive pulmonary disease (COPD) detection and classification using Mayfly optimization with deep belief network model. Biomed Signal Process Control 2024 Oct;96:106488. [doi: 10.1016/j.bspc.2024.106488]
58. Chen J, Xu Z, Sun L, et al. Deep learning integration of chest computed tomography imaging and gene expression identifies novel aspects of COPD. Chronic Obstr Pulm Dis 2023 Oct 26;10(4):355-368. [doi: 10.15326/jcopdf.2023.0399] [Medline: 37413999]
59. Chaudhary MFA, Awan HA, Gerard SE, et al. Deep learning estimation of small airways disease from inspiratory chest CT is associated with FEV1 decline in COPD. medRxiv. Preprint posted online on Sep 11, 2024. [doi: 10.1101/2024.09.10.24313079] [Medline: 39314974]
60. Cai N, Xie Y, Cai Z, Liang Y, Zhou Y, Wang P. Deep learning assisted diagnosis of chronic obstructive pulmonary disease based on a local-to-global framework. Electronics (Basel) 2024;13(22):4443. [doi: 10.3390/electronics13224443]
61. Bao Y, Al Makady Y, Mahmoodi S. Automatic diagnosis of COPD in lung CT images based on multi-view DCNN. 2021 Presented at: 10th International Conference on Pattern Recognition Applications and Methods; Feb 4-6, 2021. [doi: 10.5220/0010296805710578]
62. Awan HA, Chaudhary MFA, Gerard SE, et al. Deep residual convolutional network predicts future severe exacerbations of COPD in SPIROMICS. 2023 May Presented at: American Thoracic Society 2023 International Conference; May 19-24, 2023. [doi: 10.1164/ajrccm-conference.2023.207.1_MeetingAbstracts.A3318]
63. Alve SR, Mahmud MZ, Islam S, Khan MM. Chronic obstructive pulmonary disease prediction using deep convolutional network. medRxiv. Preprint posted online on Dec 24, 2024. [doi: 10.1101/2024.12.22.24319500]
64. Altan G, Kutlu Y, Gökçen A. Chronic obstructive pulmonary disease severity analysis using deep learning on multi-channel lung sounds. Turk J Elec Eng & Comp Sci 2020;28(5):2979-2996. [doi: 10.3906/elk-2004-68]
65. Almeida SD, Norajitra T, Lüth CT, et al. Prediction of disease severity in COPD: a deep learning approach for anomaly-based quantitative assessment of chest CT. Eur Radiol 2024 Jul;34(7):4379-4392. [doi: 10.1007/s00330-023-10540-3] [Medline: 38150075]
66. Dorosti T, Schultheiss M, Hofmann F, et al. Optimizing convolutional neural networks for chronic obstructive pulmonary disease detection in clinical computed tomography imaging. Comput Biol Med 2025 Feb;185:109533. [doi: 10.1016/j.compbiomed.2024.109533] [Medline: 39705795]
67. Feng S, Zhang W, Zhang R, et al. The identification and severity staging of chronic obstructive pulmonary disease using quantitative CT parameters, radiomics features, and deep learning features. Respiration 2025 Sep 25;25:1-13. [doi: 10.1159/000548595] [Medline: 40996946]
68. Azad Rabby AS, Chaudhary MFA, Saha P, et al. Light convolutional neural network to detect chronic obstructive pulmonary disease (COPDxNET): a multicenter model development and external validation study. medRxiv. Preprint posted online on Aug 1, 2025. [doi: 10.1101/2025.07.30.25332459] [Medline: 40766163]

XSL•FO
RenderX

69. Rezvanjou S, Moslemi A, Peterson S, et al. Classifying chronic obstructive pulmonary disease status using computed tomography imaging and convolutional neural networks: comparison of model input image types and training data severity. J Med Imag (Bellingham) 2025;12(3):034502. [doi: 10.1117/1.JMI.12.3.034502] [Medline: 40415865]
70. Rahaman Wahab Sait A, Kumar Dutta A, Ahmed Shaikh M. Optimized Kolmogorov–Arnold networks-driven chronic obstructive pulmonary disease detection model. IEEE Access 2025;13:162947-162960. [doi: 10.1109/ACCESS.2025.3610633]
71. Sahu P, Prasad P, Verma VP, Kumar S. Deep learning framework for early diagnosis of COPD and respiratory diseases using lung sound analysis. 2025 Presented at: International Conference on Big Data Analytics; Dec 17-20, 2024; Hyderabad, India p. 295-304. [doi: 10.1007/978-3-031-81821-9_17]
72. Wu Y, Xia S, Liang Z, Chen R, Qi S. Artificial intelligence in COPD CT images: identification, staging, and quantitation. Respir Res 2024 Aug 22;25(1):319. [doi: 10.1186/s12931-024-02913-z] [Medline: 39174978]
73. Willer K, Fingerle AA, Noichl W, et al. X-ray dark-field chest imaging for detection and quantification of emphysema in patients with chronic obstructive pulmonary disease: a diagnostic accuracy study. Lancet Digit Health 2021 Nov;3(11):e733-e744. [doi: 10.1016/S2589-7500(21)00146-1] [Medline: 34711378]
74. Balasubramanian P, Abia-Trujillo D, Barrios-Ruiz A, et al. Diagnostic yield and safety of diagnostic techniques for pulmonary lesions: systematic review, meta-analysis and network meta-analysis. Eur Respir Rev 2024 Jul;33(173):240046. [doi: 10.1183/16000617.0046-2024] [Medline: 39293856]
75. Arts L, Lim EHT, van de Ven PM, Heunks L, Tuinman PR. The diagnostic accuracy of lung auscultation in adult patients with acute pulmonary pathologies: a meta-analysis. Sci Rep 2020 Apr 30;10(1):7347. [doi: 10.1038/s41598-020-64405-6] [Medline: 32355210]

Abbreviations

AI: artificial intelligence
AUC: area under the summary receiver operating characteristic curve
COPD: chronic obstructive pulmonary disease
CT: computed tomography
CXR: chest X-ray
DL: deep learning
DOR: diagnostic odds ratio
GOLD: Global Initiative for Chronic Obstructive Lung Disease
NLR: negative likelihood ratio
PFT: pulmonary function testing
PLR: positive likelihood ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension
PROSPERO: International Prospective Register of Systematic Reviews
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies-2
RoB: risk of bias

Edited by S Brini; submitted 03.Sep.2025; peer-reviewed by I Yang, Y Liao; revised version received 09.Dec.2025; accepted 09.Dec.2025; published 14.Jan.2026.

Please cite as:
Yang H, Wu Y, Wu T, Ji J, Lei S, Xu W
Accuracy of Deep Learning in Diagnosing Chronic Obstructive Pulmonary Disease: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e83459
URL: https://www.jmir.org/2026/1/e83459
doi:10.2196/83459

© Hui Yang, Yijiu Wu, Tong Wu, Jingyan Ji, Sitao Lei, Weibin Xu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessment of the Diagnostic Performance and Clinical Impact of AI in Hepatic Steatosis: Systematic Review and Meta-Analysis

Jiamei Song*, MM; Dan Liu*, MS; Jitong Li, MM; Haoru Cong, MM; Ruixue Deng, MM; Yihan Lu, MBBS; Jiayi Sun, MBBS; Jingzhou Zhang, MD

College of Traditional Chinese Medicine, Changchun University of Chinese Medicine, No 1035, Boshuo Road Jingyue National High-Tech Industrial Development Zone Changchun City, Changchun, Jilin, China

*these authors contributed equally

Corresponding Author:
Jingzhou Zhang, MD
College of Traditional Chinese Medicine, Changchun University of Chinese Medicine, No 1035, Boshuo Road Jingyue National High-Tech Industrial Development Zone Changchun City, Changchun, Jilin, China

Abstract

Background: The global rise of metabolic associated fatty liver disease reflects the urgent need for accurate, noninvasive diagnostic approaches. The invasive nature of liver biopsy and the limited sensitivity of ultrasound in detecting early steatosis highlight a critical diagnostic gap. Artificial intelligence (AI) has emerged as a transformative tool, enabling the automated detection and grading of hepatic steatosis (HS) from medical imaging data.

Objective: This review aims to quantitatively evaluate the diagnostic performance of AI models for HS, explore sources of interstudy heterogeneity, and provide an appraisal of their clinical applicability, translational potential, and the major barriers impeding widespread implementation.

Methods: PubMed, Cochrane Library, Embase, Web of Science, and IEEE Xplore databases were searched until September 24, 2025. Studies using AI for HS diagnosis, meeting predefined PIRT (Patient Selection, Index Test, Reference Standard, Flow and Timing) framework and providing extractable data were included. Diagnostic performance indicators, including sensitivity, specificity, and the area under the summary receiver operating characteristic curve (AUC), were extracted and quantitatively synthesized. Meta-analyses were conducted using a bivariate random effects model. The methodological quality and risk of bias were evaluated using the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2) tool. Heterogeneity was assessed through the I^2 statistic, bivariate box plots, 95% PIs, and threshold effect analysis. Clinical applicability was examined using the Fagan nomogram and likelihood ratio tests.

Results: A total of 36 eligible studies were identified, of which 33 (comprising 36 cohorts) were included in the subgroup analyses. Results demonstrated excellent diagnostic accuracy of AI models, with a summary sensitivity of 0.95 (95% CI 0.93-0.96), specificity of 0.93 (95% CI 0.91-0.94), and an AUC of 0.98 (95% CI 0.96-0.99). Clinical applicability analysis (positive likelihood ratio >10; negative likelihood ratio <0.1) supported AI's strong potential for both confirming and excluding HS. However, substantial heterogeneity was observed across studies (I^2 >75%). According to QUADAS-2, a high risk of bias, particularly in the Patient Selection domain (44.4%), may have contributed to the overestimation of real-world performance. Subgroup analyses showed that deep learning models significantly outperformed traditional machine learning approaches (AUC: 0.98 vs 0.94). Models using ultrasound or histopathology references, retrospective designs, transfer learning, and public datasets achieved the highest accuracy (AUC 0.98-0.99) but contributed to interstudy heterogeneity.

Conclusions: AI demonstrates remarkable potential for noninvasive screening and assessment of HS, especially in primary care. Nonetheless, clinical translation remains limited by performance variability, retrospective designs, lack of external validation, practical barriers such as data privacy and workflow integration. Future studies should prioritize prospective multicenter trials and standardized external validation to bridge the gap between current evidence and clinical application. The key innovation of this review lies in establishing a unified, modality-agnostic analytical framework that integrates evidence beyond single-modality evaluations.

Trial Registration: PROSPERO CRD420251046862; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251046862

(J Med Internet Res 2026;28:e78310) doi:10.2196/78310

KEYWORDS

artificial intelligence; AI; diagnostic performance; hepatic steatosis; meta-analysis; clinical impact

XSL•FO
RenderX

Introduction

Metabolic-associated fatty liver disease (MAFLD) has emerged as one of the most prevalent chronic liver diseases worldwide, with its pathophysiology intrinsically linked to metabolic syndrome. Affected individuals frequently exhibit concomitant metabolic abnormalities such as central obesity, type 2 diabetes mellitus, and insulin resistance. The disease spectrum of MAFLD represents a continuum ranging from simple hepatic steatosis (HS) to metabolic dysfunction-associated steatohepatitis (MASH), which may progress to hepatic fibrosis, cirrhosis, or hepatocellular carcinoma (HCC) [1]. Therefore, MAFLD is a significant and growing global public health threat [2].

In 2020, an international consensus panel proposed renaming "non-alcoholic fatty liver disease" to "MAFLD" to better reflect its metabolic foundation [3,4]. To ensure the present study focuses on the diagnostic performance of artificial intelligence (AI) for the core pathological feature of HS and to enhance its clinical generalizability, including to populations with mixed etiologies such as concomitant metabolic disorders and alcohol use, we adopted the broader term "MAFLD." This terminology aligns more closely with real-world clinical practice and provides a consistent framework for AI model training and validation. Epidemiological data estimate the global prevalence of MAFLD at approximately 38%, with substantial regional variation, the highest burdens observed in Latin America, the Middle East, and North Africa [5]. The disease is also increasingly recognized among pediatric and adolescent populations, particularly in individuals with obesity, where prevalence rates have been reported to range from 7% to 14% or higher [6].

Nevertheless, the reported MAFLD prevalence varies markedly across studies, from 5% to 46% [7], reflecting considerable heterogeneity. First, diagnostic methodologies differ. Although liver biopsy remains the histopathological gold standard, its invasiveness limits clinical use, shifting reliance toward multimodal imaging. Noninvasive modalities such as ultrasound and computed tomography (CT) are widely used due to accessibility and low cost, but they lack precision in quantifying HS. Quantitative imaging techniques, including magnetic resonance imaging–proton density fat fraction (MRI-PDFF) [8], controlled attenuation parameter-based transient elastography, and noninvasive analysis [9], offer superior accuracy but are constrained by cost and limited availability. Clinical prediction models such as the Fatty Liver Index [10], Hepatic Steatosis Index [11], and Liver Fat Equation [12] enable noninvasive diagnosis through integration of anthropometric and biochemical parameters. Nevertheless, they remain vulnerable to measurement variability and lack use for longitudinal monitoring. Second, the sensitivity of existing diagnostic modalities in detecting early-stage steatosis (hepatic fat content <5%) remains suboptimal. Conventional ultrasound, in particular, has a high false-negative rate when hepatic fat content falls below 20% [13], leading to underdiagnosis and misdiagnosis in subclinical populations.

Such diagnostic inaccuracies carry serious clinical implications. Patients erroneously classified as having "simple MASLD" but who also exhibit alcohol use disorder have been shown to experience mortality risks exceeding those of individuals with typical alcoholic liver disease [14]. As MAFLD incidence rises globally, associated cirrhosis and HCC cases are also increasing. Failure to achieve early and accurate diagnosis forfeits the therapeutic window during the reversible steatosis stage, allowing progression to MASH and fibrosis. Notably, MAFLD-related HCC may arise in noncirrhotic livers [15], challenging conventional surveillance strategies that primarily target cirrhotic patients. Moreover, MAFLD is an established independent risk factor for cardiovascular disease [16]. This elevated cardiovascular risk persists throughout the disease course and remains heightened even following liver transplantation [17], underscoring the necessity of lifelong risk management.

Recent advances in AI have revolutionized medical image analysis, and hepatology has been no exception. AI-based approaches have demonstrated strong diagnostic performance across multiple hepatic pathologies. For instance, Meng et al [18] developed a VGGNet-based multistage fibrosis classifier, achieving high accuracy across 3 fibrosis grades. Wang et al [19] introduced the Explainable Diagnosis Recommender intelligent diagnostic system, which uses deep learning (DL) to automatically detect hepatic echinococcosis and cysts from CT scans. Xiao et al [20] proposed a ResNet-101-based multimodal model that classified 6 hepatobiliary diseases using slit-lamp and fundus images, outperforming clinicians of varying experience levels. Calderaro et al [21] used a DL model to reclassify combined hepatocellular-cholangiocarcinoma into pure HCC or intrahepatic cholangiocarcinoma with high sensitivity and specificity, yielding predictions consistent with clinical and molecular profiles. Specifically for HS assessment, Yang et al [22] developed a 2-stage DL model that classified four steatosis grades with an overall accuracy of 76.3% and an area under the summary receiver operating characteristic (SROC) curve (AUC) of 0.88, surpassing traditional clinical indices. Similarly, Wang et al [23] employed DL to quantify hepatic fat content by inferring proton density fat fraction (PDFF) from routine T1-weighted magnetic resonance imaging (MRI) images, surpassing the performance of the conventional 2-point Dixon fat-fraction model.

Despite these promising developments, the application of AI in the diagnosis and grading of MAFLD or HS remains at an early stage [24]. Existing systematic reviews have primarily assessed AI performance within individual imaging modalities. A critical gap remains: a comprehensive evaluation of AI's overall diagnostic efficacy across diverse imaging platforms and a systematic analysis of the key technical and methodological determinants of performance are still lacking.

Therefore, this study, for the first time, uses a bivariate mixed effects model [25] to systematically assess the overall diagnostic performance of AI in imaging-based detection of HS. The primary objectives are: (1) to quantitatively determine the aggregate diagnostic accuracy of AI models in identifying HS; (2) to comprehensively explore the sources of heterogeneity, with particular emphasis on the influence of factors such as

XSL•FO
RenderX

algorithm type, reference standard, imaging modality, study design, and data accessibility; and (3) to evaluate the clinical applicability and translational potential of AI-based diagnostic systems, while identifying major barriers to their broad clinical adoption. Through these aims, the present study seeks to generate robust, high-level evidence that transcends the limitations of individual analytical approaches, thereby providing meaningful guidance for future research and clinical practice.

Methods

Research Design and Clinical Questions

This study protocol was registered with the PROSPERO International Prospective Register of Systematic Reviews (Registration: CRD420251046862). The research was conducted as per the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Diagnostic Test Accuracy) guidelines (Checklist 1) [26].

Search Strategy

This systematic search was independently designed and conducted by two researchers as per the PRISMA-DTA (Checklist 1) [26]. PubMed, the Cochrane Library, Embase, Web of Science, and IEEE Xplore were retrieved until September 24, 2025. The search strategy was structured around three core concepts: (1) the disease ("HS," "non-alcoholic fatty liver disease," and "MAFLD"); (2) the technology ("AI," "machine learning [ML]," "DL"); and (3) the diagnostic context ("diagnosis," "detection").

Keywords within each conceptual category were combined via the OR operator (eg, AI OR DL OR ML), whereas keywords across different categories were linked using the AND operator (eg, AI AND MAFLD AND diagnosis).

The authors of the identified studies were not contacted. Reference lists of all included studies were manually reviewed to identify any additional eligible publications. No restrictions on language or publication date were applied at the database level to maximize search sensitivity. However, non-English records were excluded during subsequent screening. Gray literature, preprints, and unpublished studies were not systematically searched. This decision was made a priori to focus on peer-reviewed, full-text articles that had undergone editorial review, thereby ensuring baseline methodological quality and the availability of sufficient details for data extraction. The complete, reproducible search strings for all databases are provided in Table S1 in Multimedia Appendix 1.

Screening Process

Two independent reviewers initially screened all retrieved titles and abstracts. After removing duplicate records, studies were deemed eligible for inclusion if they met the following criteria:

- Study content: the research conformed to the predefined PIRT (Patient Selection, Index Test, Reference Standard, Flow and Timing) framework:
  - Patient Selection (P): patients undergoing abdominal imaging or pathological examination for HS assessment.
  - Index Test (I): AI models based on DL or ML, using input images derived from ultrasound, CT, MRI, or pathology.
  - Reference Standard (R): defined by the original studies, including MRI-PDFF, liver biopsy pathology, or expert-graded ultrasound. These reference standards reflect real-world diagnostic diversity and were recognized as potential sources of heterogeneity.
  - Target Condition (T): Diagnosis and grading of HS according to the thresholds and criteria adopted in the included studies, allowing cross-comparison of AI performance across varying diagnostic definitions.
- Data availability: studies had to provide diagnostic contingency data, true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), or sufficient information to derive diagnostic performance metrics, such as AUC with 95% CIs, sensitivity, specificity, accuracy, and predictive values.

Exclusion criteria were: (1) language: non-English publications; (2) study type: letters, conference abstracts, reviews, or academic papers lacking original data; (3) study subjects: animal or nonhuman research, bioinformatics-based analyses, and predictive modeling studies focused on indices, risks, or associations rather than diagnosis; and (4) data sufficiency: studies without key contingency data or insufficient information to calculate diagnostic performance metrics.

Data Extraction

Two researchers independently extracted data based on the following domains: (1) study characteristics: first author, publication year, site of data collection, and duration of the study period; (2) study population: total sample size, and demographic characteristics (mean or median age); (3) methodological parameters: accessibility of clinical sample data, diagnostic reference standard, and validation strategy; (4) algorithmic architecture: type of algorithm, classifier employed, and application of transfer learning (TL); and (5) diagnostic efficacy: raw contingency table data, and aggregated diagnostic performance metrics.

Diagnostic Performance Evaluation and Quality Assessment

Pooled estimates of sensitivity, specificity, and AUC, together with their 95% CIs, were presented using forest plots. Heterogeneity was quantified using the I^2 statistic. The AUC was designated as the primary indicator for overall diagnostic accuracy, as it integrates performance across all thresholds and remains unaffected by any single cut-off point. A SROC curve was constructed following an assessment of the threshold effect using the Spearman correlation coefficient between the logit of sensitivity and the logit of (1-specificity). Heterogeneity and its implications were further visualized via 95% PIs and bivariate boxplots. Potential small-study effects were evaluated using the Deeks funnel plot asymmetry test. Additional diagnostic indicators, including the diagnostic odds ratio (DOR), positive likelihood ratio (LRP), and negative likelihood ratio (LRN), were calculated. Clinical applicability was further examined

XSL•FO
RenderX

using a Fagan nomogram, while the distribution of likelihood ratios across studies was illustrated via scatterplots.

Two investigators assessed the risk of bias via the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2) checklist (Checklist 2) in Rev-Man in terms of PIRT framework [27]. The QUADAS-2 tool, recommended by the Cochrane collaboration, was used to assess the methodological quality and risk of bias of included diagnostic accuracy studies.

Quality Assurance and Dispute Resolution

All screening, data extraction, and quality assessment procedures were independently conducted by 2 reviewers. Any discrepancies were first resolved through discussion to reach a consensus. When consensus was not achieved, a third senior investigator adjudicated the disagreement to make the final determination. This multilevel process ensured that all extracted data represented unanimous agreement within the research team.

Subgroup Analysis

The following independent meta-analyses were conducted:

1. AI type (ML versus DL): DL models were defined as those based on multilayer artificial neural networks, such as convolutional neural networks and recurrent neural networks (RNNs). This category included all studies that explicitly reported using DL or identified architectures such as VGG, ResNet, U-Net, DenseNet, or Inception. ML models referred to traditional algorithms that learn from data without relying primarily on deep neural architectures, including support vector machines (SVM), random forests, decision trees, and logistic regression. To explore the potential influence of different algorithmic approaches on diagnostic performance.
2. Reference standards for steatosis grading (MRI-PDFF, liver histopathology, or ultrasound): to determine whether there was a performance gap between models based on noninvasive imaging and those based on the pathological "gold standard."
3. Imaging modality (ultrasound, CT, or histopathology): to assess how differences in imaging principles, invasiveness, and the diagnostic information scale (macroscopic versus microscopic) affected model performance.
4. Application of TL: TL was used when a study explicitly reported the use of a model pretrained on a large-scale dataset (eg, ImageNet) as the initial framework for feature extraction or model fine-tuning. To evaluate whether this specific technique could improve model performance in small-sample medical datasets.
5. Study design (single-center versus multicenter): to assess the generalizability of models across different data distributions.
6. Study type (prospective vs retrospective): to explore the temporal relationship between data collection and model development and to evaluate the potential impact of selection bias on performance assessment.
7. Data accessibility: to evaluate the effect of study reproducibility and transparency on research outcomes.

Data Analysis

Given the substantial heterogeneity observed among included studies with respect to patient populations, imaging devices, and AI algorithms, a bivariate mixed effects model was used to derive more accurate and reliable pooled estimates [25]. To ensure the robustness of the meta-analytic results, quantitative synthesis (eg, subgroup analysis) was performed only when at least 3 independent studies, defined as studies conducted by different authors, using distinct experimental protocols, or involving separate participant cohorts, were available. Multiple effect estimates from the same publication were included when they originated from distinct participant cohorts (eg, multicenter datasets or independent validation sets). When multiple model outputs were reported, only the best-performing model or that validated using an independent dataset was retained. Subgroup analyses were not conducted when fewer than three independent studies were available for a given subgroup. All statistical analyses and visualizations were performed via Stata MP 18 (StataCorp LLC). A 2-tailed P value <.05 denoted statistical significance.

Results

Included Study Description

As of September 24, 2025, 2536 articles were retrieved. After removing 864 duplicates, the titles and abstracts of the rest were screened as per the predefined eligibility criteria, resulting in the exclusion of 1596 articles. Specifically, 9 were non-English publications, 884 were of other types, 673 involved inappropriate study subjects, and 30 used unsuitable research methods. The full texts of the remaining 76 articles were subsequently reviewed. Seventeen studies were excluded for incomplete data, 7 for being of other types, 7 for inappropriate methodologies, and 9 for being inaccessible. Ultimately, 36 studies were included in the final analysis (Figure 1). The characteristics of the included studies are summarized in Table 1, and the results of the subgroup analyses are presented in Table 2.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Search Extension) flowchart depicting the study selection process for the systematic review of artificial intelligence in diagnosing hepatic steatosis.

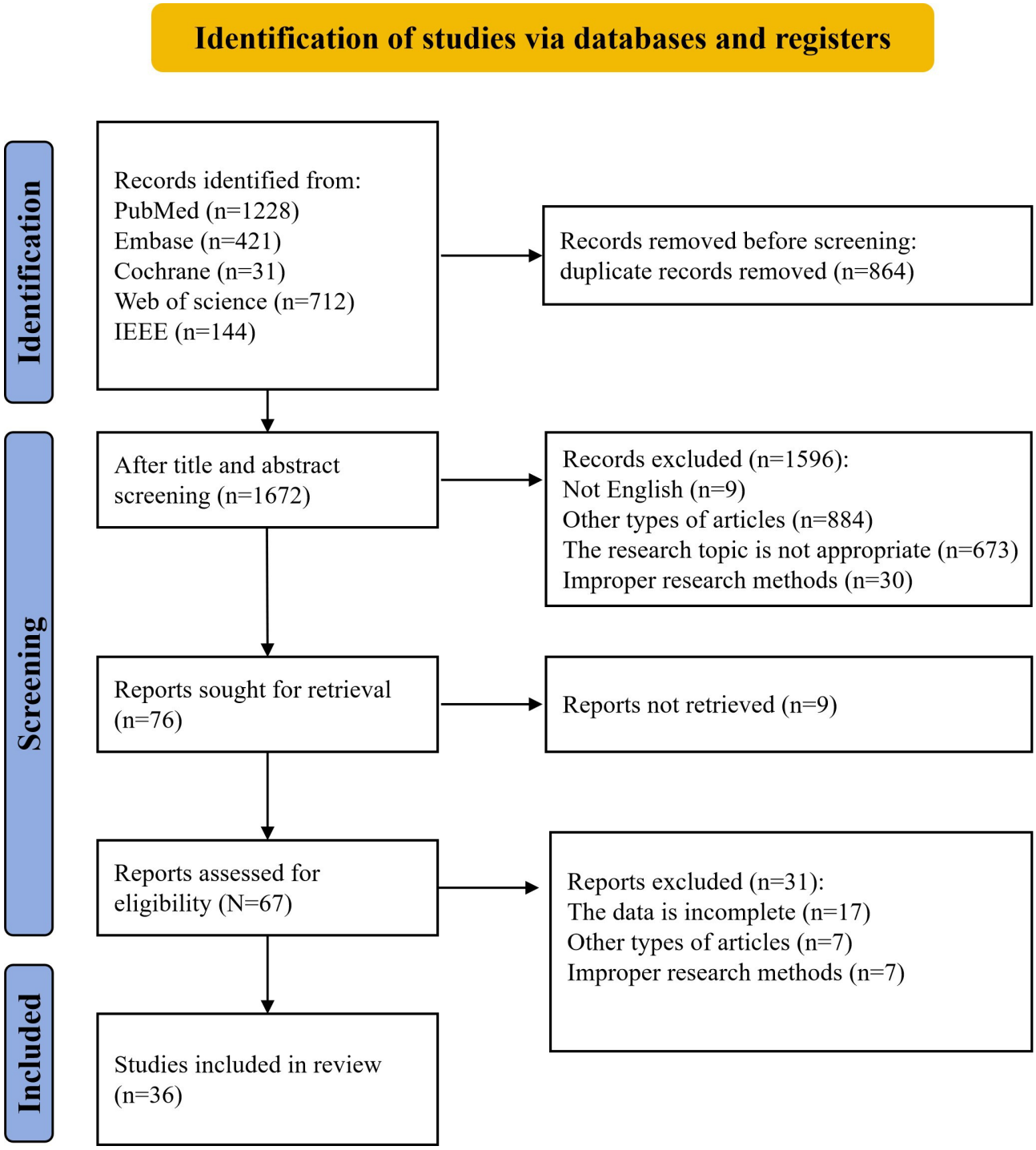

XSL•FO
RenderX

Table . Summary of study characteristics in the systematic review of artificial intelligence (AI)-assisted hepatic steatosis (HS) detection (N=36).

Number	Study	Device	DL[a] or ML[b]	TL[c]	Source of data	Data range	Sample size	Age, years	Open access data	Research type
1	Yang et al (2022) [28]	MRI[d]	DL	Yes	A hospital in Beijing	Jun - Jul 2020	50	• NR[e]	No	• Retrospective study • Single-center
2	Acharya et al (2016) [29]	Ultrasound	DL	NR	The University of Malaya Medical Centre, Malaysia	NR	100	• NR	No	• Retrospective study • Single-center
3	Jeon et al (2023) [30]	Ultrasound	DL	NR	Seoul National University Hospital	Jul 2020-Jun 2021	173	• mean (SD): 51 (14) • range: 19-74	No	• Prospective study • Single-center
4	Neogi et al (2018) [31]	Ultrasound	DL	NR	Chittaranjan National Cancer Hospital	NR (a span of 6 months)	51	• NR	No	• Retrospective study • Single-center
5	Chen et al (2020) [32]	Ultrasound	DL	NR	Chang Gung Memorial Hospital in Taiwan	2017 - 2020	205	• mean (SD): 55 (11.6)	No	• Retrospective study • Single-center
6	Dubois et al (2019) [33]	Ultrasound	DL	NR	Rennes University Hospital	Jun 2017 - Aug 2018	53	• median (IQR): 61 (28–72)	No	• Prospective study • Single-center
7	Shi et al (2019) [34]	Ultrasound	ML	NR	Shanghai Public Health Clinical Center	NR	60	• range: 19 - 69	No	• Retrospective study • Single-center
8	Jesper et al (2020) [35]	Ultrasound	ML	NR	Erlangen University Hospital	Oct 2018 - Sep 2019	27	• mean (SD): 50 (17)	No	• Prospective study • Single-center

XSL•FO
RenderX

Number	Study	Device	DL[a] or ML[b]	TL[c]	Source of data	Data range	Sample size	Age, years	Open access data	Research type
9	McHenry et al (2020) [36]	MRI	DL	NR	Dallas County	2000-2002; 2007-2009	2139	• median: 44	No	• Prospective study • Single-center
10	Roy et al (2021) [37]	Pathology	DL	NR	Children's Hospital of Atlanta and Emory University	2014 - 2016	36	• mean (SD): 14.9 (2.59)	Yes	• Retrospective study • Multi-center
11	Sun et al (2020) [38]	Pathology	DL	Yes	Washington University School of Medicine Transplant Pathology Service	Apr 2015 - Sep 2016	91	• NR	No	• Retrospective study • Multi-center
12	Chou et al (2021) [39]	Ultrasound	DL	Yes	Taipei Medical University Hospital	2016 - 2018	2070	• NR	No	• Retrospective study • Single-center
13	Constantinescu et al (2021) [40]	Ultrasound	DL	Yes	Outpatient clinic of a private healthcare network	NR	60	• range: 18 - 92	No	• Prospective study • Single-center
14	Pérez-Sanz et al (2021) [41]	Pathology	DL	NR	University Clinical Hospital Virgen de la Arrixaca-Biomedical Research Institute of Murcia	NR	20	• NR	No	• Retrospective study • Single-center
15	Pickhardt et al (2020) [42]	CT	DL	NR	A single academic medical center	Feb 2010 - Jan 2017	1204	• mean (SD): 45.2 (12.4)	No	• Retrospective study • Single-center
16	Rhyou et al (2021) [43]	Ultrasound	DL	Yes	Samsung Medical Center and Byra Dataset	NR	NR	• NR	Yes	• Retrospective study • Multi-center
17		Ultrasound	ML	NR			82		No	

XSL•FO
RenderX

Number	Study	Device	DL[a] or ML[b]	TL[c]	Source of data	Data range	Sample size	Age, years	Open access data	Research type
	Destrempes et al (2022) [44]				Center Hospitalier de l'Université de Montre´al and McGill University Health Center	Oct 2014 - Sep 2018		• mean (SD): 56 (12) • range: 23-78		• Retrospective study • Multicenter
18	Alshagathrh et al (2023) [45]	Ultrasound	DL	Yes	University of Warsaw, Poland	NR	55	• mean (SD): 40.1 (9.1)	Yes	• Retrospective study • Single-center
19	Podder et al (2023) [46]	Pathology	DL	Yes	Open Science Framework	NR	NR	• NR	Yes	• Retrospective study • Single-center
20	Ibrahim et al (2023) [47]	Ultrasound	DL	Yes	Beijing You'an Hospital in Beijing, China, and the National Hepatology and Tropical Medicine Research Institute in Cairo, Egypt	NR	478	• mean (SD): 40.97 (10.61)	No	• Prospective study • Single-center
21	Yao et al (2023) [48]	Ultrasound	DL	Yes	Byra dataset and the Health Service Center in the Chenghua District of Chengdu	2020 - 2022	1320	• NR	Yes	• Retrospective study • Multicenter
22	Byra et al (2018) [49]	Ultrasound	DL	Yes	Medical University of Warsaw, Poland	NR	55	• mean (SD): 40.1 (9.1)	Yes	• Prospective study • Multicenter
23	Torgersen et al (2024) [50]	CT	DL	NR	Philadelphia VA[f] Medical Center	01 Jan 2010 - 30 Dec 2017	120	• 61.1 (55.3 - 64.6)	No	• Retrospective study • Single-center
24		Ultrasound	DL	NR		NR	131		No	

XSL•FO
RenderX

Number	Study	Device	DL[a] or ML[b]	TL[c]	Source of data	Data range	Sample size	Age, years	Open access data	Research type
	Wang et al (2023) [51]				Chang Gung Memorial Hospital, Taiwan			• Mean (age range) • 60 (30-79) • 61 (30-79) • 62 (41-76) • 63 (28-79)		• Prospective study • Single-center
25	Jeon et al (2024) [52]	CT	ML	NR	Institute of Radiation Medicine, Seoul National University Medical Research Center, Seoul National University Hospital, Seoul, Korea	Dec 2018 - Dec 2021	252	• mean: 37.3 • range: 18-64	No	• Retrospective study • Single-center
26	Piella et al (2024) [53]	Mobile phones	ML	NR	Vall d'Hebron University Hospital	NR	192	• median (IQR): 62 (52.25 - 71.75)	No	• Retrospective study • Single-center
27	Yoo et al (2024) [54]	CT	DL	NR	Radiologic database in our institution	Jan 2017 - Jun 2021	362	• mean (SD): 37.3 (11.5) • range: 18-65	No	• Retrospective study • Single-center
28	Zhang et al (2024) [55]	CT	DL	NR		NR	986	• NR	Yes	• Retrospective study • Multicenter

XSL•FO
RenderX

Number	Study	Device	DL[a] or ML[b]	TL[c]	Source of data	Data range	Sample size	Age, years	Open access data	Research type
					LIDC-IDRI, NSCLC-Lung1, RIDER, VESSEL12, MIDRC-RICORD, COVID-19-Italy, and COVID-19-China					
29	Cherchi et al (2021) [56]	Pathology	DL	NR	The University Hospital of Udine	Jan 2018 - May 2019	33	• NR	No	• Retrospective study • Single-center
30	Wu et al (2023) [57]	Ultrasound	DL	NR	Chang Gung Memorial Hospital	NR	276	• NR	No	• Retrospective study • Single-center
31	Drazinos et al (2025) [58]	Ultrasound	DL	NR	The University of Texas MD Anderson Cancer Center	Jan 2018 - Jan 2019	112	• mean (SD): 51 (16.13)	No	• Retrospective study • Single-center
32	Kaffas et al (2025) [59]	Ultrasound	DL	NR	NR	01 Jan 2010 - 01 Jan 2022	403	• median (IQR): 53 (40 - 66)	No	• Retrospective study • Single-center
33	Kim et al (2025) [60]	CT	DL	NR	Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea	2001 - 2016	3620	• mean (SD): 31.7 (9.4)	No	• Retrospective study • Single-center
34	Zhang et al (2025) [61]	CT	DL	NR	The First Affiliated Hospital of Zhengzhou University	Jul 2022 - May 2023	840	• mean (SD): 49.1 (11.5)	Yes	• Retrospective study • Single-center
35		CT	DL	NR		NR	1740		Yes	

XSL•FO
RenderX

Number	Study	Device	DL[a] or ML[b]	TL[c]	Source of data	Data range	Sample size	Age, years	Open access data	Research type
	Derstine et al (2025) [62]				Michigan Medicine; VA			• mean (SD): 43.1 (12.8)		• Retrospective study • Single-center
36	Corso et al (2024) [63]	Ultrasound	DL	NR	The University of Texas MD Anderson Cancer Center	NR	186	• mean (SD): 51.95 (13.4)	No	• Prospective study • Multi-center

[a]DL: deep learning.

[b]ML: machine learning.

[c]TL: transfer learning.

[d]MRI: magnetic resonance imaging.

[e]NR: No report.

[f]VA: Veterans Administration.

XSL•FO
RenderX

Table . Summary of artificial intelligence (AI) performance for diagnosing hepatic steatosis (HS) based on different subgroups (number of studies or cohorts; pooled sensitivity; specificity; area under the curve [AUC]; heterogeneity [I^2, %]; Spearman correlation coefficient [for threshold effect]; and publication bias such as posttest probability positive or negative, positive likelihood ratio [LRP], and negative likelihood ratio [LRN]).

Subgroup and subgroup analysis (studies/datasets)	Pooled sensitivity (95% CI); I^2 (%)	Pooled specificity (95% CI); I^2 (%)	Summary AUC (95% CI)	Spearman correlation coefficient (*P* value)	Publication bias (*P* value)	Posttest probability positive (%)/ negative (%)	LRP	LRN
AI type								
DL[a] (29/32)	0.96 (0.94 - 0.98); 95.87	0.94 (0.91 - 0.95); 97.69	0.98 (0.97 - 0.99)	0.21 (*P*=.05)	.46	94/4	>10	<0.1
ML[b] (4/4)	0.87 (0.78 - 0.93); 32.63	0.88 (0.80 - 0.93); 63.45	0.94 (0.91 - 0.96)	–1 (*P*=.99)	.49	88/13	<10	>0.1
Reference standard								
MRI-PDFF[c] (7/7)	0.92 (0.86 - 0.95); 95.70	0.91 (0.86 - 0.94); 98.43	0.97 (0.95 - 0.98)	–0.36 (*P*=.13)	.15	91/8	>10	<0.1
Pathology (13/14)	0.97 (0.92 - 0.99); 97.86	0.92 (0.86 - 0.95); 85.60	0.98 (0.96 - 0.99)	0.12 (*P*=.02)	.29	92/3	>10	<0.1
Ultrasound (6/6)	0.98 (0.90 - 1.00); 94.32	0.96 (0.94 - 0.98); 85.23	0.98 (0.96 - 0.99)	1 (*P*=.99)	.21	96/2	>10	<0.1
Imaging modality								
Ultrasound (20/22)	0.96 (0.93 - 0.98); 94.86	0.93 (0.90 - 0.96); 96.16	0.98 (0.97 - 0.99)	0.21 (*P*=.4)	.50	94/4	>10	<0.1
CT[d] (8/9)	0.93 (0.86 - 0.96); 94.43	0.93 (0.87 - 0.96); 94.76	0.97 (0.95 - 0.98)	0.20 (*P*=.04)	.24	93/7	>10	<0.1
Pathology (4/4)	0.98 (0.91 - 1.00); 79.53	0.96 (0.86 - 0.99); 0.00	0.99 (0.98 - 0.99)	–1 (*P*=.99)	.00	96/2	>10	<0.1
TL[e]								
Used (9/9)	0.99 (0.96 - 1.00); 95.36	0.93 (0.88 - 0.97); 93.80	0.99 (0.98 - 1.00)	0.2 (*P*=.04)	.77	94/1	>10	<0.1
Not used (24/27)	0.93 (0.90 - 0.96); 84.22	0.93 (0.90 - 0.95); 96.71	0.98 (0.96 - 0.99)	0.22 (*P*=.05)	.53	93/7	>10	<0.1
Study design								
Single-center (25/27)	0.94 (0.91 - 0.96); 94.44	0.93 (0.91 - 0.95); 97.33	0.98 (0.96 - 0.99)	0.25 (*P*=.06)	.98	93/6	>10	<0.1
Multicenter (8/9)	0.99 (0.94 - 1.00); 95.26	0.92 (0.85 - 0.96); 82.33	0.97 (0.96 - 0.99)	0.47 (*P*=.22)	.30	92/1	>10	<0.1
Study type								
Retrospective (25/26)	0.95 (0.92 - 0.97); 96.58	0.95 (0.92 - 0.97); 98.15	0.98 (0.97 - 0.99)	0.39 (*P*=.15)	.53	95/5	>10	<0.1

XSL•FO
RenderX

Subgroup and subgroup analysis (studies/datasets)	Pooled sensitivity (95% CI); I^2 (%)	Pooled specificity (95% CI); I^2 (%)	Summary AUC (95% CI)	Spearman correlation coefficient (*P* value)	Publication bias (*P* value)	Posttest probability positive (%)/ negative (%)	LRP	LRN
Prospective (8/10)	0.97 (0.92 - 0.99); 82.76	0.87 (0.84 - 0.89); 53.89	0.90 (0.87 - 0.92)	1 (*P*=.99)	.87	88/4	<10	<0.1
Data availability								
Available (9/10)	0.99 (0.96 - 1.00); 97.06	0.95 (0.92 - 0.97); 73.17	0.99 (0.97 - 0.99)	−0.5 (*P*=.25)	.19	96/1	>10	<0.1
Unavailable (24/26)	0.92 (0.89 - 0.95); 81.89	0.92 (0.89 - 0.94); 96.62	0.97 (0.95 - 0.98)	0.09 (*P*=.01)	.30	92/8	>10	<0.1

[a]DL: deep learning.

[b]ML: machine learning.

[c]MRI-PDFF: magnetic resonance imaging–proton density fat fraction.

[d]CT: computed tomography.

[e]TL: transfer learning

Diagnostic Performance and Heterogeneity

Of the 36 included studies, 33 (comprising 36 datasets) satisfied the criteria for subgroup analysis. The pooled results (Figure 2) demonstrated a summary sensitivity of 0.96 (95% CI 0.93 - 0.97), a specificity of 0.93 (95% CI 0.91 - 0.95), and an AUC of 0.98 (95% CI 0.96 - 0.99), indicating excellent diagnostic discrimination by the AI models. Substantial heterogeneity was observed across studies (I^2>75%). The Spearman correlation coefficient (0.21, *P*=.05) suggested that threshold effects contributed minimally to overall heterogeneity. The broad 95% PI, however, indicated that differences in diagnostic thresholds were a major source of variability. No significant small-study effects were identified (*P*=.65). In terms of clinical applicability, at a pretest probability of 50%, a positive AI result increased the posttest probability to 93%, whereas a negative result reduced it to 4%. Likelihood ratio scattergram analysis confirmed that the pooled estimates were located within the “confirm and exclude” quadrant (LRP >10 and LRN <0.1), underscoring the strong clinical value of AI for both confirming and excluding HS.

XSL•FO
RenderX

Figure 2. Diagnostic performance of artificial intelligence (AI) models for hepatic steatosis (HS) detection across 33 studies comprising 36 datasets [28-35,37,39-52,54-63]. (A) Forest plots illustrating sensitivity and specificity for the subgroup of AI applications across 33 studies with 36 datasets. (B) Summary receiver operating characteristic (SROC) curve depicting diagnostic performance of AI across 33 studies with 36 datasets, with corresponding 95% CIs. The 95% prediction region reflects the expected range of true sensitivity and specificity in future studies. (C) Bivariate boxplot illustrating the distribution and heterogeneity of AI performance across 33 studies with 36 datasets. (D) The Deeks funnel plot for evaluation of potential publication bias. (E) The Fagan nomogram depicting posttest probabilities. (F) Clinical application plot showing positive likelihood ratio (LRP) and negative likelihood ratio (LRN). LLQ: lower-left quadrant; LUQ: upper-left quadrant; RLQ: lower-right quadrant; RUQ: upper-right quadrant; SROC: summary receiver operating characteristic.

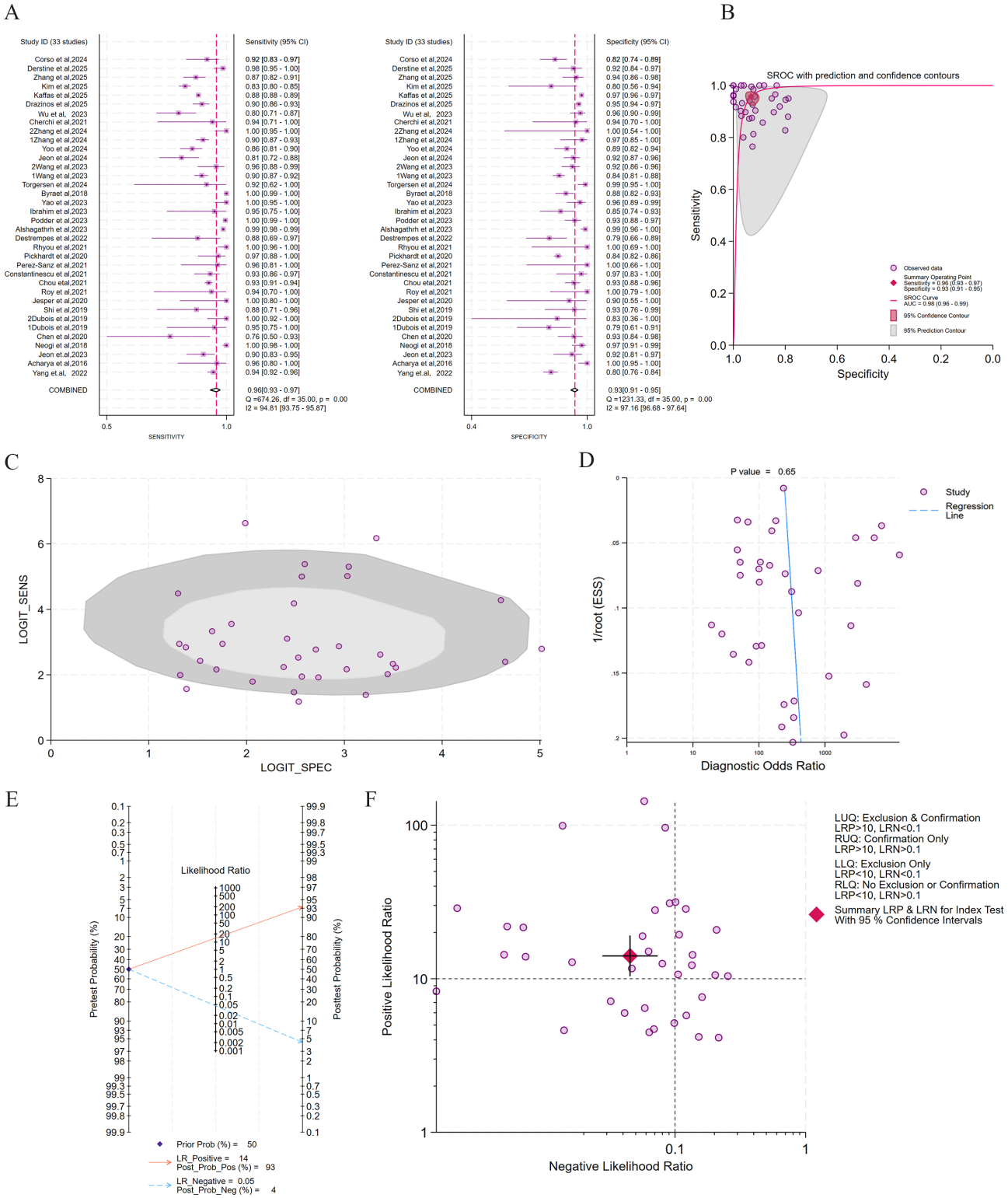

Risk of Bias Assessment

The QUADAS-2 quality assessment (Figure 3) revealed that 44% (16/36) of studies exhibited a high risk of bias in the patient selection domain, primarily due to selection bias, limited representativeness of study populations, and incomplete reporting of key clinical parameters. In the Index test domain, 6% (2/36) of studies were rated as high risk, largely attributable

XSL•FO
RenderX

to the absence of image quality control, subjective elements during image processing, and nonstandardized training or validation procedures. In the reference standard domain, 17% (6/36) of studies demonstrated a high risk of bias, most commonly due to deviations from gold-standard reference methods, unclear blinding procedures, or incomplete pathological sampling information. The flow and timing domain exhibited unclear risk in 36% (14/36) of studies, often due to insufficient reporting on patient inclusion pathways and the interval between image acquisition and diagnostic confirmation. These methodological limitations may contribute to an overestimation of AI model performance in real-world clinical practice.

Figure 3. Risk of bias assessment of the 36 included studies on artificial intelligence–based hepatic steatosis diagnosis using the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2) tool [28-63].

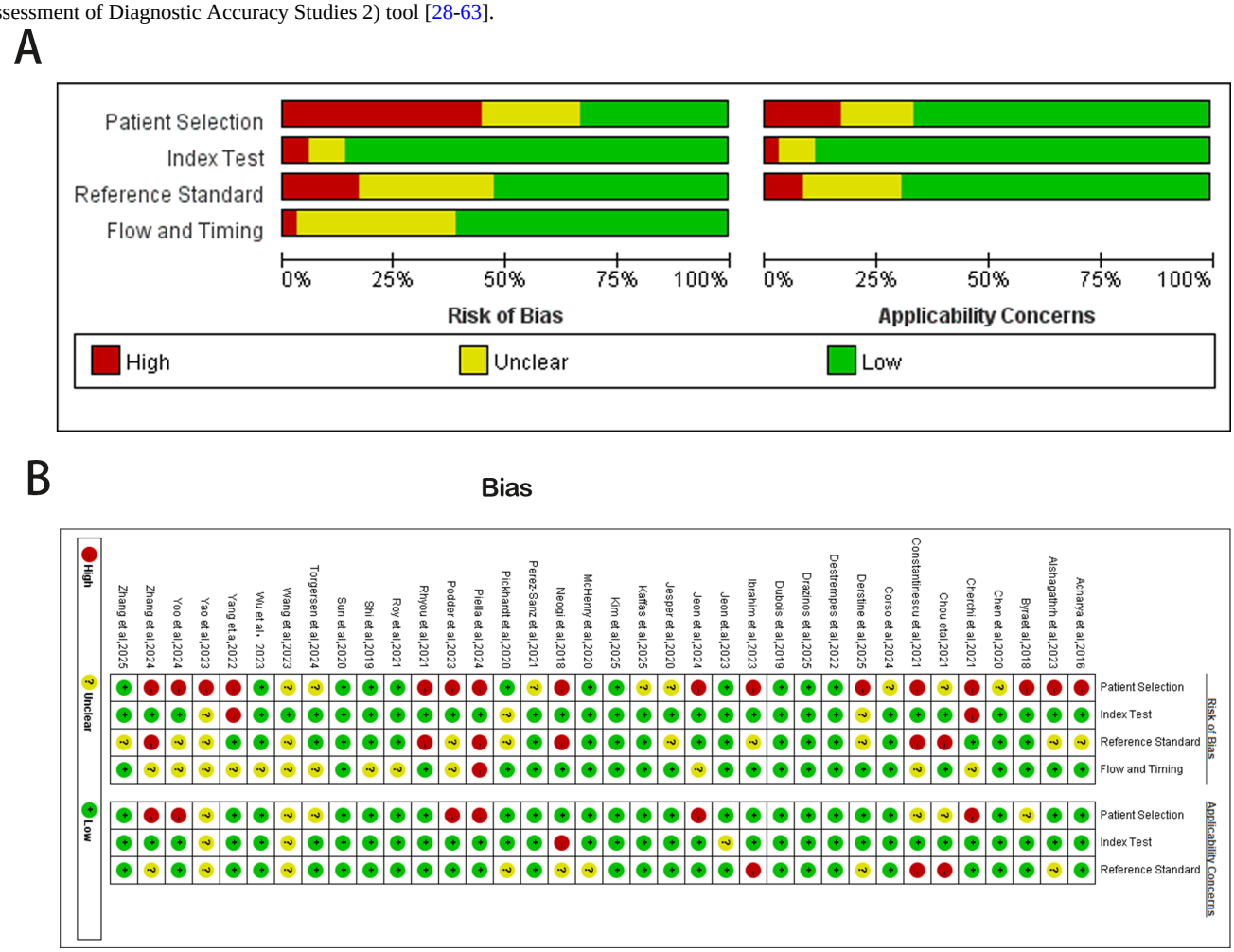

Subgroup Meta-Analyses

Subgroup analyses were conducted for 7 key variables (Figures S1-S16 in Multimedia Appendix 1), with diagnostic performance interpreted relative to clinical applicability thresholds (LRP >10 for strong rule-in capability; LRN <0.1 for strong rule-out capability).

Algorithm Type

As shown in Figure 4 A-B, the DL models demonstrated significantly higher diagnostic accuracy than ML models (AUC: 0.98 vs 0.94), exhibiting strong rule-in and rule-out performance. However, DL models displayed pronounced heterogeneity (I^2 >95%), likely influenced by threshold effects (Spearman=0.21, P=.05), suggesting that these findings should be generalized with caution. ML models showed lower heterogeneity (sensitivity I^2=32.63%; specificity I^2=63.45%) but weaker discriminatory power (LRP <10, LRN >0.1).

XSL•FO
RenderX

Figure 4. Diagnostic performance stratified by algorithm type for hepatic steatosis (HS) detection. (A) Summary receiver operating characteristic (SROC) curve for deep learning (DL) algorithms in 29 studies [28-33,37,39-43,45-51,54-63] comprising 32 datasets; (B) SROC curve for machine learning (ML) algorithms in 4 studies [34,35,44,52] comprising 4 datasets.

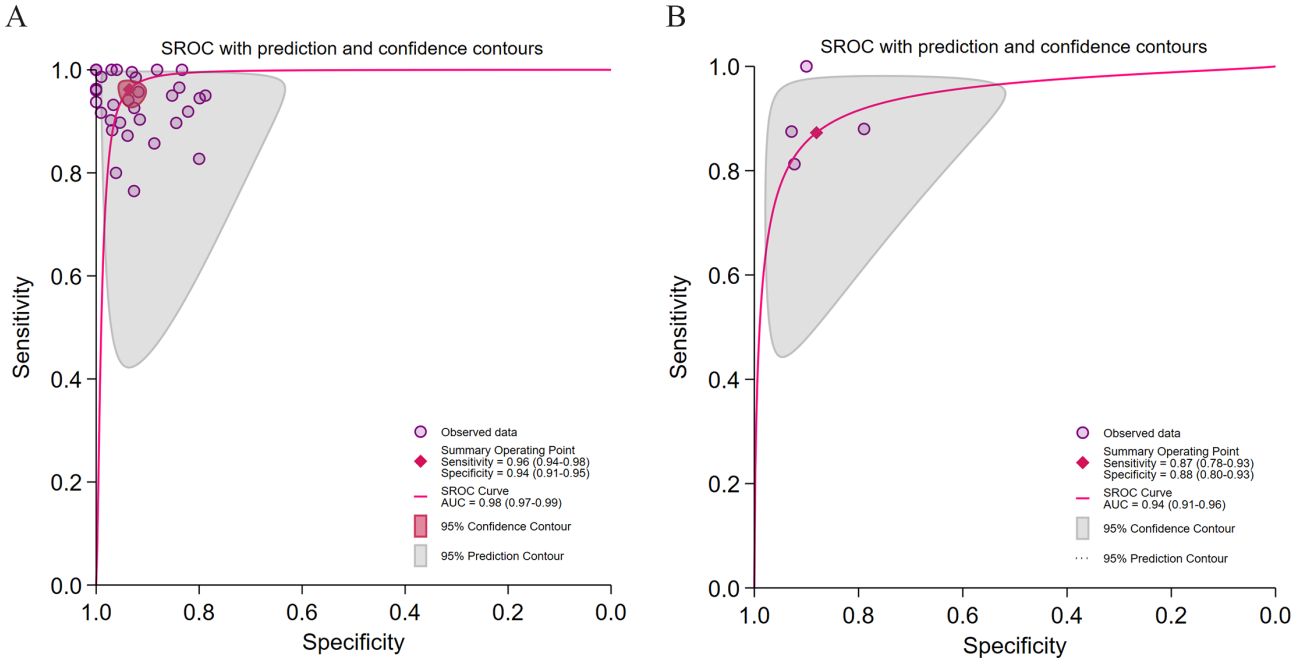

Reference Standard

As shown in Figure 5 A-C, studies using ultrasound and histopathology as reference standards achieved comparable AUCs (both were 0.98) with ideal likelihood ratios, although ultrasound-based models exhibited higher pooled sensitivity and specificity (0.98 and 0.96, respectively) than histopathology-based models (0.97 and 0.92). The ultrasound subgroup showed a perfect threshold effect (Spearman=1; *P*=.99), indicating well-defined diagnostic criteria that may be subjectively constrained. The histopathology subgroup exhibited a minimal threshold effect (Spearman=0.12; *P*=.02), suggesting that interstudy variations in sample handling and scoring could significantly influence model performance. The MRI-PDFF subgroup achieved a comparable AUC (0.97) but demonstrated very high heterogeneity (I^2 >95%), limiting result stability.

Figure 5. Diagnostic performance stratified by reference standard for hepatic steatosis (HS) detection. (A) Summary receiver operating characteristic (SROC) curve for magnetic resonance imaging–proton density fat fraction (MRI-PDFF) in 7 studies [30,42,52,54,59,62,63] comprising 7 datasets; (B) SROC curve for pathology in 13 studies [28,32,33,35,41,43-46,49,56,57,60] comprising 14 datasets; (C) SROC curve for ultrasound in 6 studies [29,31,39,40,48,58] comprising 6 datasets.

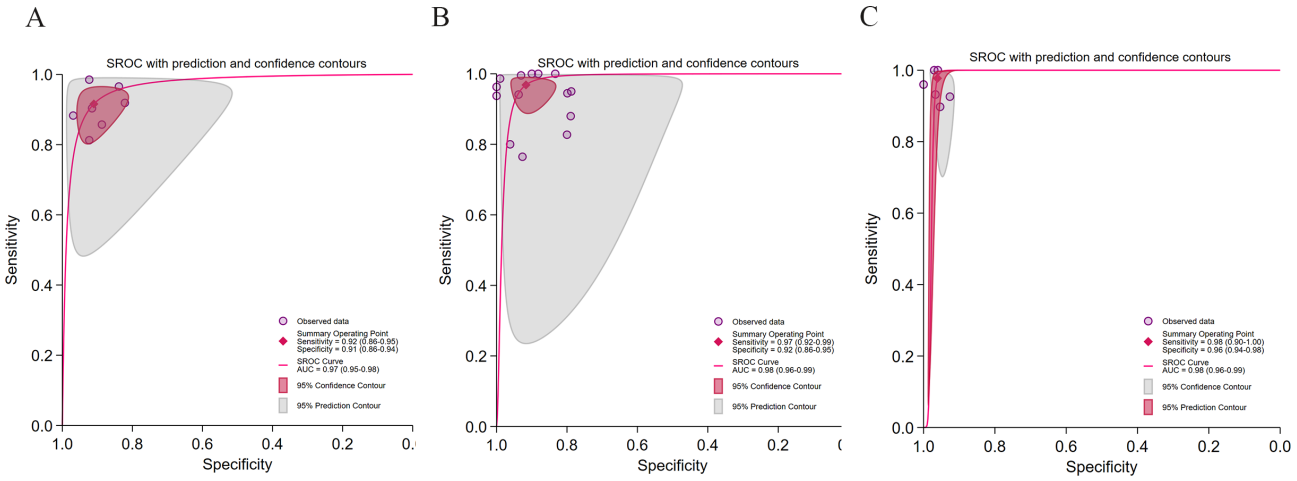

Imaging Modality

As shown in Figure 6 A-C, histopathology-based models achieved the highest diagnostic performance (AUC=0.99) with no detectable heterogeneity, suggesting robust and consistent results. However, significant publication bias was identified (*P*<.001), implying potential preferential publication of high-performing studies. AI models based on ultrasound and CT achieved comparable accuracy (AUC: 0.98 vs 0.97), though both exhibited marked heterogeneity (I^2 >94%). Only the CT subgroup showed a negligible threshold effect (Spearman=0.20; *P*=.04).

XSL•FO
RenderX

Figure 6. Diagnostic performance stratified by imaging modality for hepatic steatosis (HS) detection. (A) Summary receiver operating characteristic (SROC) curve for ultrasound imaging in 20 studies [29-35,39,40,43-45,47-49,51,57-59,63] comprising 22 datasets; (B) SROC curve for computed tomography (CT) imaging in 8 studies [42,50,52,54,55,60-62] comprising 9 datasets; (C) SROC curve for pathology imaging in 4 studies [37,41,46,56] comprising 4 datasets.

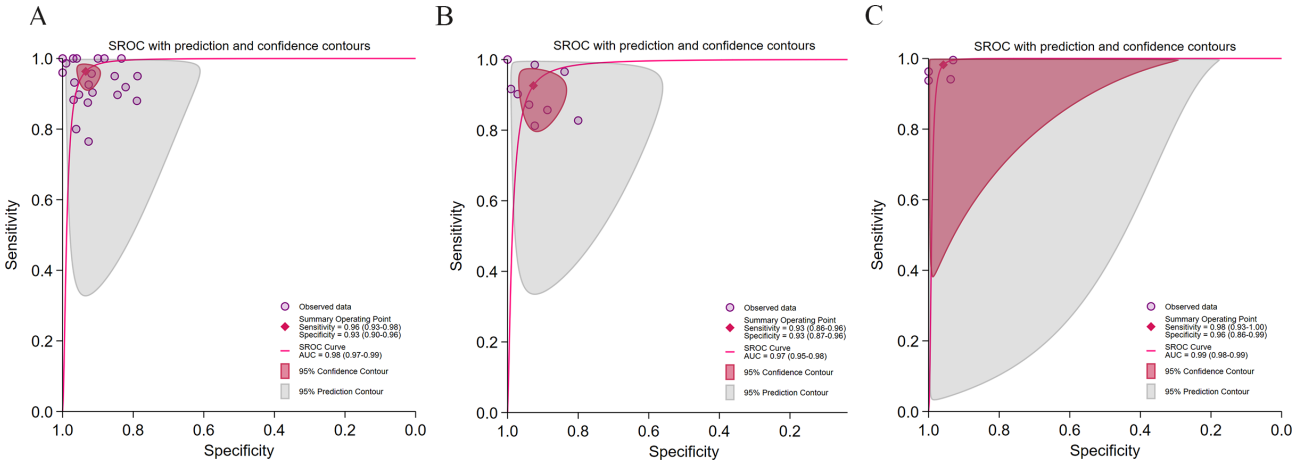

Application of TL

As shown in the figure Figure 7A-B models employing TL achieved higher pooled sensitivity (0.99 vs 0.93) and stronger rule-out capability (LRN: 0.01 vs 0.07). No significant publication bias was detected in either subgroup (P >.05). Nevertheless, both demonstrated considerable heterogeneity (I^2 >84%) and mild threshold effects (Spearman=0.20 vs 0.22; P=.04 vs .05), reflecting the influence of interdomain data discrepancies.

Figure 7. Diagnostic performance of transfer learning (TL) for hepatic steatosis (HS) detection. (A) Summary receiver operating characteristic (SROC) curve for studies employing TL in 9 studies [28,39,40,43,45-49] comprising 9 datasets; (B) SROC curve for studies not employing TL in 24 studies [29-35,37,41,42,44,50-52,54-63] comprising 26 datasets.

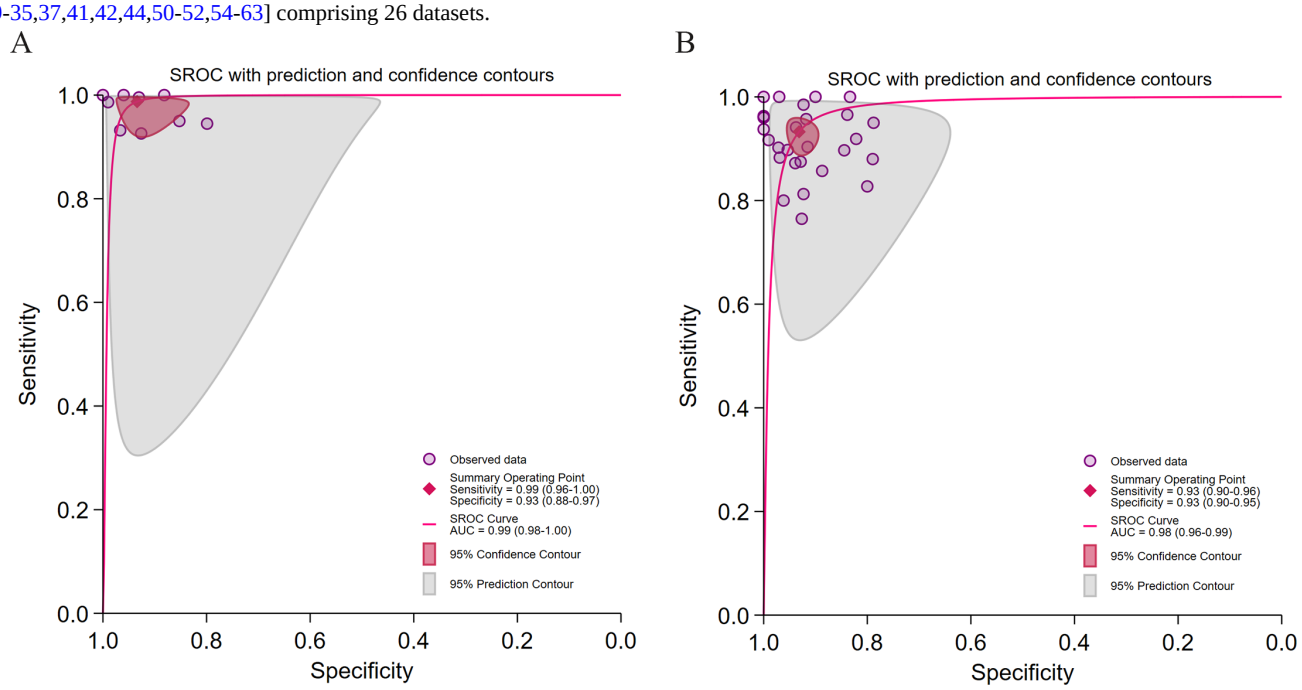

Study Design

As shown in Figure 8A-B, multicenter studies demonstrated superior sensitivity (0.99 vs 0.94) and lower heterogeneity (I^2=82.33%), indicating greater generalizability and stronger rule-out potential (LRN: 0.01 vs 0.06). In contrast, single-center studies exhibited marginally higher specificity (0.93 vs 0.92) but very high heterogeneity (I^2 >94%), suggesting limited external validity.

XSL•FO
RenderX

Figure 8. Diagnostic performance of different research designs for hepatic steatosis (HS) detection. (A) Summary receiver operating characteristic (SROC) curve for single-center studies in 25 studies [28-35,39-42,45-47,50-52,54,56-61] with 26 datasets; (B) SROC curve for multicenter studies in 8 studies [37,43,44,48,49,55,62,63] with 9 datasets.

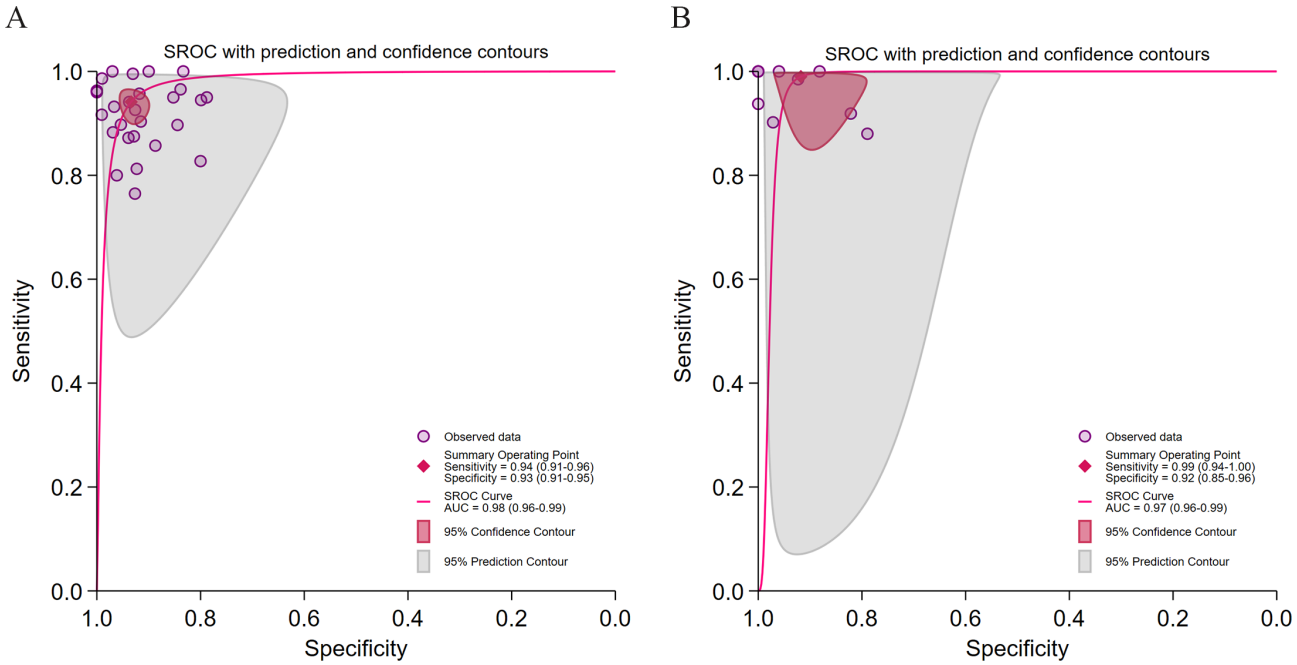

Study Type

As shown in Figure 9 A-B, retrospective studies achieved higher overall accuracy (AUC: 0.98 versus 0.90) and stronger rule-in ability (LRP: 9.5 versus 8.8), though with significant heterogeneity (I^2 >96%). Prospective studies, which can better reflect clinical reality, were affected by a perfect threshold effect (Spearman=1) and exhibited weaker rule-in performance (LRP<10).

Figure 9. Diagnostic performance of different research types for hepatic steatosis (HS) detection. (A) Summary receiver operating characteristic (SROC) curve for retrospective studies in 25 [28,29,31,32,34,37,41-46,48,50,52,54-62] studies with 26 datasets; (B) SROC curve for prospective studies in 8 studies with 9 datasets. AUC: area under curve.

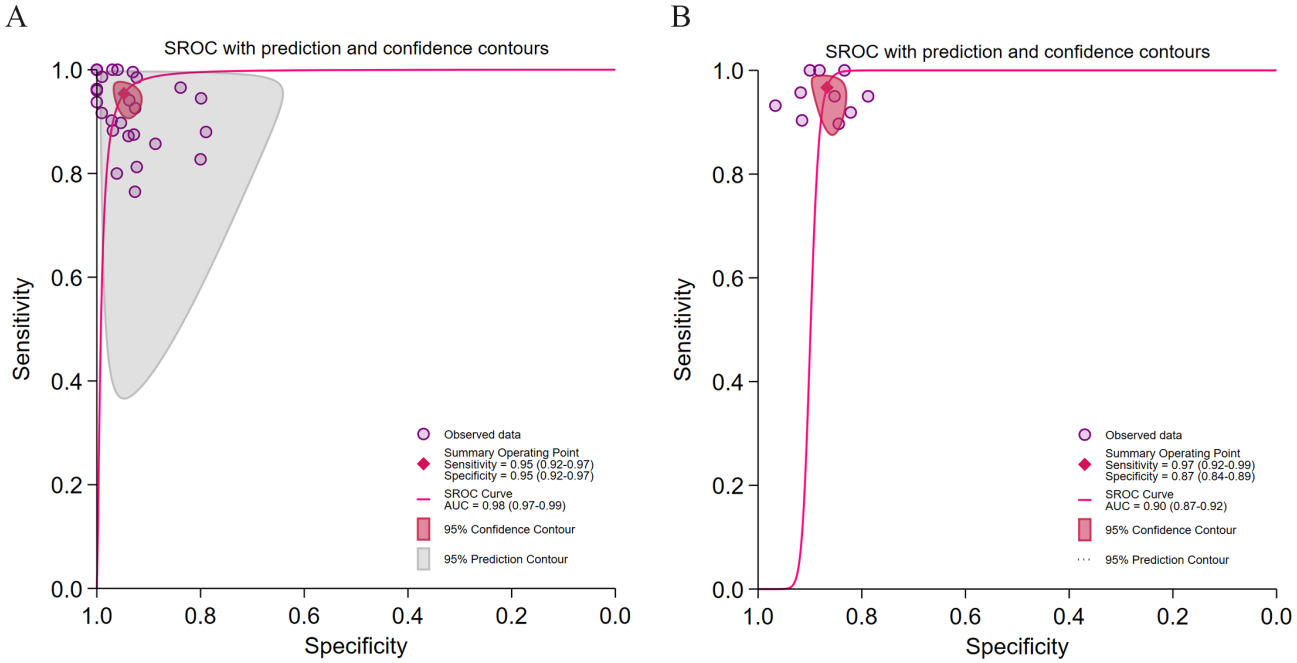

Data Accessibility

As shown in Figure 10A-B studies using publicly available datasets (n=9) achieved superior diagnostic accuracy (AUC=0.99) and stronger clinical applicability (LRP=9.6; LRN=0.01). In contrast, studies using nonpublic data performed comparably (AUC=0.97) but showed a significant threshold effect (Spearman=0.09; *P*=.01), indicating reduced result stability.

XSL•FO
RenderX

Figure 10. Diagnostic performance of data availability for hepatic steatosis (HS) detection. (A) Summary receiver operating characteristic (SROC) curve for studies with available data in 9 studies [37,43,45,46,48,49,55,61,62] with 10 datasets; (B) SROC curve for studies with unavailable data in 24 studies [28-35,38-42,44,47,50-52,54,56-60,63] with 25 datasets.

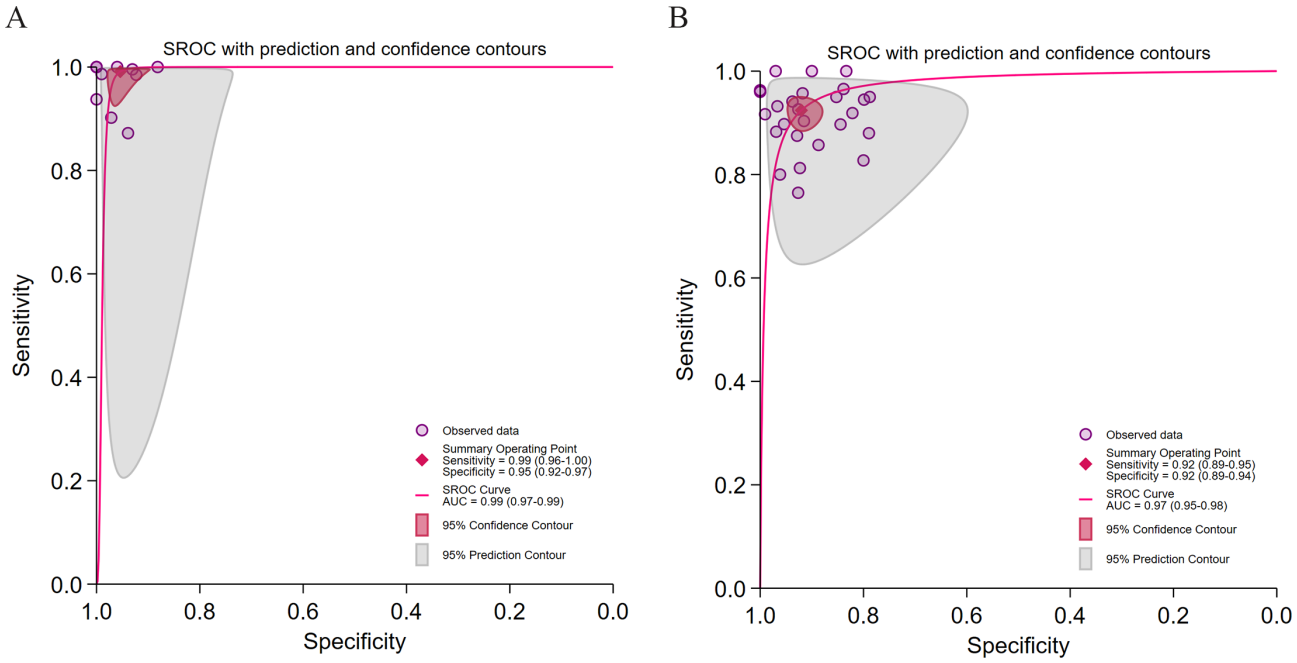

Subgroup definitions are detailed in the methods section. Heterogeneity was categorized as follows: I^2<50% low heterogeneity, 50% - 75% moderate heterogeneity, and >75% high heterogeneity.

Discussion

Principal Findings

This meta-analysis of 36 studies demonstrates the superior diagnostic performance of AI in identifying HS, yielding a pooled AUC of 0.98, surpassing that of conventional ultrasound, CT, and MRI, whose pooled AUCs were 0.93, 0.975, and 0.97, respectively [64-66]. These findings underscore AI's potential to overcome the inherent physical constraints of individual imaging techniques, thereby establishing it as a versatile and adaptive diagnostic approach. This capability carries considerable clinical significance, providing a strong rationale for further meta-analyses dedicated to AI-based diagnostic technologies and informing the development of flexible, context-specific clinical applications. By optimizing the use of the most accessible and cost-effective diagnostic resources, AI could markedly broaden the availability of early HS screening across diverse health care settings.

Although the encouraging performance of AI in diagnosing HS is promising, interpretation of these findings must be tempered by a critical appraisal of the methodological limitations underlying this research. Our analysis revealed substantial heterogeneity (I^2>75%) and a high overall risk of bias among included studies, particularly within the patient selection domain, where 44% (16/36) were judged to be at high risk. A major limitation lies in the predominance of retrospective, single-center designs (25/36, 69%). Such studies typically develop and validate models within controlled, idealized data environments, meaning that reported metrics may reflect "best-case scenarios" rather than true clinical performance across diverse devices, operators, and patient populations in routine practice. Moreover, independent external validation and multicenter prospective trials remain notably scarce, severely limiting the assessment of these models' generalizability. Therefore, while existing evidence underscores AI's considerable potential in HS diagnosis, the current body of research remains insufficient to justify its widespread clinical adoption. Bridging the translational gap from high-performing algorithms to reliable, universally applicable clinical tools thus remains a substantial challenge.

Subgroup analyses provide valuable insights for optimizing AI model design and informing clinical integration. DL-based models demonstrate exceptionally high specificity in diagnosing HS, offering a distinct clinical advantage by reducing unnecessary liver biopsies. However, these models require higher-quality data annotation and greater computational resources. Model performance was also closely associated with the reference standard and imaging modality used. Systems using histopathological images as input achieved the highest diagnostic accuracy. Nevertheless, their clinical applicability is restricted by procedural invasiveness and sampling error [67]. The AI-assisted whole-slide analysis model proposed by Roy et al [37] improves quantitative consistency but faces practical barriers related to cost and patient acceptance.

The choice of imaging modality inherently involves trade-offs between diagnostic accuracy, accessibility, and cost. MRI-PDFF, while providing precise, noninvasive quantification, is affected by confounders such as iron overload, edema, and concurrent pathologies [7], and its high cost limits use in primary care. Ultrasound remains the most accessible and economical option but suffers from operator dependency, reduced sensitivity for mild steatosis, limited penetration in obese individuals [68,69], and suboptimal accuracy in detecting fibrosis [70]. AI integration could mitigate these limitations by standardizing

XSL•FO
RenderX

acquisition and interpretation, though at the cost of increased system complexity and computational demand. CT achieves a sensitivity and specificity of 0.93 but is constrained by ionizing radiation exposure and potential interference from iodine-based contrast agents [71]. The dual-energy CT 3D nnU-Net model developed by Yoo et al [54] achieved an AUC of 0.97 for distinguishing steatotic from normal tissue, yet its clinical application is constrained by limited equipment availability.

To optimize model performance and data use, TL has been widely adopted, an especially valuable strategy given the substantial costs associated with medical data annotation. Nonetheless, the effectiveness of TL depends critically on the degree of similarity between the source and target domains; substantial domain discrepancies may lead to "negative transfer," as illustrated by sensitivity variations of up to 10% in the Inception-v3 model reported by Constantinescu et al [40]. To address this limitation, emerging approaches such as adversarial domain adaptation frameworks have achieved near-human classification accuracy on heterogeneous MRI datasets [72]. Similarly, hybrid pretraining strategies [73] and federated learning techniques have reached up to 99% of the performance attained through centralized training [74]. These approaches enhance model robustness while effectively addressing data privacy and heterogeneity.

Beyond algorithmic optimization, the real-world implementation of AI is profoundly influenced by study design and data governance. Retrospective studies, which constituted the majority (25/36, 69%) of the included reports, demonstrated significantly higher performance than prospective studies (AUC: 0.98 vs 0.94), likely reflecting the high-quality and well-curated imaging data typically available in retrospective cohorts. In contrast, prospective designs more faithfully capture real-world clinical workflows but are inherently subject to operational variability, such as inconsistent imaging protocols and unpredictable patient factors, thereby leading to attenuated performance.

Furthermore, data governance and accessibility are pivotal in determining model generalizability. Multicenter collaborations and data sharing can improve generalizability and reproducibility, though they require standardized imaging protocols, increased logistical coordination, and greater resource investment, posing feasibility challenges in resource-limited settings. Moreover, access to medical data for AI development remains hindered by privacy regulations, institutional policies, and technical interoperability barriers. Privacy-preserving strategies, such as federated learning, offer promising solutions by enabling multi-institutional collaboration without direct data exchange, albeit at the cost of increased computational demands and system complexity. It should also be noted that publicly available datasets may not fully represent the clinical heterogeneity encountered in real-world practice, thereby introducing potential selection bias. These factors, while critical for improving AI performance, also contribute substantially to heterogeneity, underscoring the necessity of comprehensive external validation and context-specific adaptation before large-scale clinical implementation.

Expanding Role in HS Management

The use of AI extends beyond diagnostic precision to encompass the comprehensive management of HS. Accumulating evidence indicates that AI not only enables accurate quantification of hepatic fat but also integrates radiomic, pathological, and clinical data to facilitate fibrosis staging, predict HCC risk, assess posttransplant survival, and stratify cardiovascular complications. For instance, a VGG16-based ultrasound model outperformed human interpretation in classifying borderline cases [75]. The integration of macrogenomic sequencing with ML has proven effective for the differential diagnosis of HS in obese pediatric populations [76]. Similarly, an ML model based on MRI-derived liver fat quantification markedly improved diagnostic accuracy for liver fibrosis [77]. AI-powered digital pathology platforms reduce the inherent subjectivity of conventional histological assessment [78], while DL-based radiomics facilitates the identification of critical pathological features such as microvascular invasion [79]. A DL algorithm demonstrated 99% accuracy in predicting postliver transplantation survival [80]. In the context of MAFLD-related complications, AI algorithms have been employed to accurately identify affected patients from electronic health records, revealing type 2 diabetes mellitus as a significant predictor of all-cause mortality (hazard ratio: 1.36) [81]. Moreover, a dual model combining tongue imaging with clinical indicators achieved precise prediction of coronary heart disease risk among patients with fatty liver [82]. The foregoing advances signal a diagnostic paradigm shift in HS management from a traditional "liver-centric" approach towards a "patient-centric" model of multi-system risk management, paving the way for early intervention and personalized therapy.

In summary, the advantages of AI in HS diagnosis are threefold as follows:

1. Enhanced early detection: DL models can detect subclinical pathological alterations, including hepatic fat infiltration below 5%, thereby reducing diagnostic subjectivity and improving reproducibility [43,45,83].
2. Standardized quantitative analysis: End-to-end, pixel-level segmentation enables automated calculation of HS, minimizing reliance on manual interpretation and potentially substituting for histopathological assessment in resource-constrained settings.
3. Longitudinal predictive modeling: The integration of time-series radiomic and metabolomic features facilitates the construction of individualized models predicting cirrhosis progression and MAFLD onset within 3 years, providing actionable insights for precision treatment planning.

Challenges and a Phased Implementation Framework

Despite its promising outlook, the widespread clinical adoption of AI in HS management faces multiple challenges. Technically, data heterogeneity, stemming from variations in imaging quality [84], scanner types, and reference standard thresholds, impedes the development of universally robust and generalizable models. Many high-performing algorithms are derived from single-center, retrospective datasets (eg, Yang et al [22], n=50, Beijing) with limited demographic diversity, thereby

XSL•FO
RenderX

compromising their external validity and real-world applicability. Moreover, most existing models primarily focus on imaging biomarkers for fat quantification without adequately elucidating the complex pathophysiological interplay among steatosis, metabolic comorbidities, and fibrosis, limiting both clinical interpretability and holistic disease assessment.

From a clinical integration perspective, the transition from algorithmic development to real-world deployment necessitates careful consideration of workflow compatibility, device dependency, and cost-effectiveness. Lightweight AI models hold promise for incorporation into primary care ultrasound systems, facilitating large-scale population screening, whereas more advanced MRI- or CT-based models may be more appropriately implemented in tertiary medical centers. The overarching objective is seamless integration into existing clinical workflows, ensuring that AI serves as an assistive, rather than disruptive, technology that streamlines radiological practice, conserves clinician time, and enhances diagnostic efficiency [85]. Furthermore, issues concerning data privacy [86], algorithmic bias [87], and accountability [88] lack clear regulatory frameworks.

From a global health perspective, the clinical use of AI in HS diagnosis varies according to resource availability. To promote both efficiency and equity in HS diagnosis and management, a phased implementation framework is proposed:

1. Tiered deployment in specific scenarios: in resource-limited settings, lightweight AI systems can be paired with portable ultrasound to enable cost-effective community screening and early detection. Suspected cases may then be referred to higher-level hospitals for precise stratified diagnosis (eg, MRI-PDFF), thereby optimizing resource allocation and minimizing unnecessary biopsies. In high-resource environments, AI-driven automated image processing facilitates accurate fat quantification and disease staging, forming a synergistic diagnostic–therapeutic feedback loop.
2. Establish cross-institutional collaborative data platforms: the adoption of federated learning and related technologies can enhance data diversity while ensuring privacy protection. Such approaches enable robust model development based on heterogeneous real-world data, mitigate model bias and validation gaps, eliminate the need for centralized storage of sensitive information, and provide the foundation for scalable, privacy-preserving deployment.
3. Transition from standalone tools to integrated management platforms: the ultimate objective is to advance AI from a single-function diagnostic aid to a comprehensive, multi-task management system. By synchronously quantifying steatosis, assessing fibrosis, and evaluating inflammatory markers through multimodal data integration. Incorporating these outputs directly into clinical decision-making workflows, AI could evolve from diagnostic assistance to intelligent, holistic disease management.

Limitations in the Literature

Several limitations warrant cautious interpretation. First, considerable methodological and clinical heterogeneity was observed across the included studies, constraining the reliability of the conclusions. Despite extensive subgroup analyses, variability arising from differences in patient characteristics, imaging equipment, and diagnostic thresholds could not be fully addressed. This residual heterogeneity undermines the robustness of pooled estimates and suggests the influence of unmeasured factors affecting AI performance.

Second, the analysis was limited by methodological shortcomings inherent in the primary studies. Inadequate reporting of key patient characteristics hindered subgroup analyses by disease etiology, particularly distinguishing pure MAFLD from mixed forms, a critical gap given the potential impact of comorbidities on diagnostic accuracy. Furthermore, wide variation in AI architectures and the limited number of comparable models precluded meaningful comparisons across technical approaches, leaving the effect of architectural design on diagnostic performance unclear.

Third, the generalizability and real-world applicability of the findings remain limited. Most studies were retrospective, single-center designs prone to selection bias, with scarce external or temporal validation. Thus, the high-performance metrics reported may represent an idealized best-case scenario rather than outcomes achievable in prospective clinical settings.

Additionally, although our restriction to peer-reviewed full-text publications ensured a baseline level of methodological rigor, the exclusion of relevant preprints and gray literature may have introduced publication bias. Such selective inclusion likely favored studies reporting positive outcomes, potentially leading to overestimated performance measures. Moreover, key practical factors, such as computational burden, workflow integration, and technical expertise, could not be quantitatively evaluated, despite their importance for real-world implementation.

Conclusions

This meta-analysis highlights the substantial diagnostic potential of AI, particularly DL, in assessing HS. Its key contribution lies in establishing a unified, imaging-modality-independent analytical framework that provides comprehensive evidence beyond the constraints of individual imaging techniques. Nonetheless, these results reflect technical promise rather than confirmed clinical use. The translation from high-performing algorithms to reliable clinical tools remains hindered by performance heterogeneity, retrospective study designs, and insufficient external validation. While the technological foundation of AI in HS is encouraging, clinical maturity has yet to be achieved. Bridging this translational gap will require prospective multicenter studies, standardized reporting protocols, and rigorous external validation. Ultimately, successful clinical adoption will depend on demonstrating not only algorithmic robustness but also tangible improvements in patient outcomes and workflow efficiency across real-world health care settings.

XSL•FO
RenderX

Funding

This work was supported by the Science and Technology Research Program of Jilin Provincial Department of Education (No. JKH20250664KJ). The sponsor participated in the preliminary design phase and guided the selection of research methodologies. However, the sponsor did not participate in subsequent data collection, analysis, interpretation of results, or the preparation of this manuscript.

Data Availability

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: JS
Data curation: JS and JL
Formal analysis: YL
Funding acquisition: DL
Investigation: JL and HC
Methodology: DL
Project administration: JZ
Resources: JZ
Software: JS
Supervision: DL and RD
Validation: JS
Visualization: HC
Writing – original draft: JS
Writing – review & editing: JZ
All authors commented on a previous version of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Retrieval strategy, and Forest plots, bivariate boxplots, Deeks' funnel plots, Fagan's nomogram plots, and clinical application plots for each subgroup analysis.
[DOCX File, 9782 KB - jmir_v28i1e78310_app1.docx]

Checklist 1
PRISMA-DTA checklist.
[PDF File, 138 KB - jmir_v28i1e78310_app2.pdf]

Checklist 2
QUADAS-2 checklist.
[XLSX File, 31 KB - jmir_v28i1e78310_app3.xlsx]

References

1. Unalp-Arida A, Ruhl CE. Prevalence of metabolic dysfunction-associated steatotic liver disease and fibrosis defined by liver elastography in the United States using national health and nutrition examination survey 2017-March 2020 and August 2021-August 2023 data. Hepatology 2025 Nov 1;82(5):1256-1273. [doi: 10.1097/HEP.0000000000001211] [Medline: 39724144]
2. Riazi K, Azhari H, Charette JH, et al. The prevalence and incidence of NAFLD worldwide: a systematic review and meta-analysis. Lancet Gastroenterol Hepatol 2022 Sep;7(9):851-861. [doi: 10.1016/S2468-1253(22)00165-0] [Medline: 35798021]
3. Eslam M, Newsome PN, Sarin SK, et al. A new definition for metabolic dysfunction-associated fatty liver disease: an international expert consensus statement. J Hepatol 2020 Jul;73(1):202-209. [doi: 10.1016/j.jhep.2020.03.039] [Medline: 32278004]
4. Gofton C, Upendran Y, Zheng MH, George J. MAFLD: how is it different from NAFLD? Clin Mol Hepatol 2023 Feb;29(Suppl):S17-S31. [doi: 10.3350/cmh.2022.0367] [Medline: 36443926]

XSL•FO
RenderX

5. Younossi ZM, Golabi P, Paik JM, Henry A, Van Dongen C, Henry L. The global epidemiology of nonalcoholic fatty liver disease (NAFLD) and nonalcoholic steatohepatitis (NASH): a systematic review. Hepatology 2023 Apr 1;77(4):1335-1347. [doi: 10.1097/HEP.0000000000000004] [Medline: 36626630]
6. Younossi ZM, Kalligeros M, Henry L. Epidemiology of metabolic dysfunction-associated steatotic liver disease. Clin Mol Hepatol 2025 Feb;31(Suppl):S32-S50. [doi: 10.3350/cmh.2024.0431] [Medline: 39159948]
7. Wong VWS, Ekstedt M, Wong GLH, Hagström H. Changing epidemiology, global trends and implications for outcomes of NAFLD. J Hepatol 2023 Sep;79(3):842-852. [doi: 10.1016/j.jhep.2023.04.036] [Medline: 37169151]
8. Starekova J, Hernando D, Pickhardt PJ, Reeder SB. Quantification of liver fat content with CT and MRI: state of the art. Radiology 2021 Nov;301(2):250-262. [doi: 10.1148/radiol.2021204288] [Medline: 34546125]
9. Leporq B, Ratiney H, Pilleul F, Beuf O. Liver fat volume fraction quantification with fat and water T1 and T 2* estimation and accounting for NMR multiple components in patients with chronic liver disease at 1.5 and 3.0 T. Eur Radiol 2013 Aug;23(8):2175-2186. [doi: 10.1007/s00330-013-2826-x] [Medline: 23588583]
10. Bedogni G, Bellentani S, Miglioli L, et al. The fatty liver index: a simple and accurate predictor of hepatic steatosis in the general population. BMC Gastroenterol 2006 Nov 2;6:33. [doi: 10.1186/1471-230X-6-33] [Medline: 17081293]
11. Lee JH, Kim D, Kim HJ, et al. Hepatic steatosis index: a simple screening tool reflecting nonalcoholic fatty liver disease. Dig Liver Dis 2010 Jul;42(7):503-508. [doi: 10.1016/j.dld.2009.08.002] [Medline: 19766548]
12. Kotronen A, Peltonen M, Hakkarainen A, et al. Prediction of non-alcoholic fatty liver disease and liver fat using metabolic and genetic factors. Gastroenterol 2009 Sep;137(3):865-872. [doi: 10.1053/j.gastro.2009.06.005] [Medline: 19524579]
13. Dasarathy S, Dasarathy J, Khiyami A, Joseph R, Lopez R, McCullough AJ. Validity of real time ultrasound in the diagnosis of hepatic steatosis: a prospective study. J Hepatol 2009 Dec;51(6):1061-1067. [doi: 10.1016/j.jhep.2009.09.001] [Medline: 19846234]
14. Cho Y. Hidden burden of alcohol use disorder in MASLD and MetALD: clinical and nomenclatural implications. Gut Liver 2025 Sep 15;19(5):637-638. [doi: 10.5009/gnl250414] [Medline: 40947954]
15. Lin H, Zhang X, Li G, Wong GLH, Wong VWS. Epidemiology and clinical outcomes of metabolic (dysfunction)-associated fatty liver disease. J Clin Transl Hepatol 2021 Dec 28;9(6):972-982. [doi: 10.14218/JCTH.2021.00201] [Medline: 34966660]
16. Stefan N, Yki-Järvinen H, Neuschwander-Tetri BA. Metabolic dysfunction-associated steatotic liver disease: heterogeneous pathomechanisms and effectiveness of metabolism-based treatment. Lancet Diabetes Endocrinol 2025 Feb;13(2):134-148. [doi: 10.1016/S2213-8587(24)00318-8] [Medline: 39681121]
17. Soldera J, Corso LL, Rech MM, et al. Predicting major adverse cardiovascular events after orthotopic liver transplantation using a supervised machine learning model: a cohort study. World J Hepatol 2024 Feb 27;16(2):193-210. [doi: 10.4254/wjh.v16.i2.193] [Medline: 38495288]
18. Meng D, Zhang L, Cao G, Cao W, Zhang G, Hu B. Liver fibrosis classification based on transfer learning and FCNet for ultrasound images. IEEE Access 2017;5:1-1. [doi: 10.1109/ACCESS.2017.2689058]
19. Wang Z, Bian H, Li J, et al. Detection and subtyping of hepatic echinococcosis from plain CT images with deep learning: a retrospective, multicentre study. Lancet Digit Health 2023 Nov;5(11):e754-e762. [doi: 10.1016/S2589-7500(23)00136-X] [Medline: 37770335]
20. Xiao W, Huang X, Wang JH, et al. Screening and identifying hepatobiliary diseases through deep learning using ocular images: a prospective, multicentre study. Lancet Digit Health 2021 Feb;3(2):e88-e97. [doi: 10.1016/S2589-7500(20)30288-0] [Medline: 33509389]
21. Calderaro J, Ghaffari Laleh N, Zeng Q, et al. Deep learning-based phenotyping reclassifies combined hepatocellular-cholangiocarcinoma. Nat Commun 2023 Dec 14;14(1):8290. [doi: 10.1038/s41467-023-43749-3] [Medline: 38092727]
22. Yang Y, Liu J, Sun C, et al. Nonalcoholic fatty liver disease (NAFLD) detection and deep learning in a Chinese community-based population. Eur Radiol 2023 Aug;33(8):5894-5906. [doi: 10.1007/s00330-023-09515-1] [Medline: 36892645]
23. Wang K, Cunha GM, Hasenstab K, et al. Deep learning for inference of hepatic proton sensity fat fraction from T1-weighted in-phase and opposed-phase MRI: retrospective analysis of population-based trial data. AJR Am J Roentgenol 2023 Nov;221(5):620-631. [doi: 10.2214/AJR.23.29607] [Medline: 37466189]
24. Zhou LQ, Wang JY, Yu SY, et al. Artificial intelligence in medical imaging of the liver. World J Gastroenterol 2019 Feb 14;25(6):672-682. [doi: 10.3748/wjg.v25.i6.672] [Medline: 30783371]
25. Reitsma JB, Glas AS, Rutjes AWS, Scholten R, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005 Oct;58(10):982-990. [doi: 10.1016/j.jclinepi.2005.02.022] [Medline: 16168343]
26. Cohen JF, Deeks JJ, Hooft L, et al. Preferred reporting items for journal and conference abstracts of systematic reviews and meta-analyses of diagnostic test accuracy studies (PRISMA-DTA for Abstracts): checklist, explanation, and elaboration. BMJ 2021 Mar 15;372:n265. [doi: 10.1136/bmj.n265] [Medline: 33722791]
27. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]

XSL•FO
RenderX

28. Yang R, Zhou Y, Liu W, Shang H. Study on the grading model of hepatic steatosis based on improved DenseNet. J Healthc Eng 2022;2022:9601470. [doi: 10.1155/2022/9601470] [Medline: 35340251]
29. Acharya UR, Fujita H, Bhat S, et al. Decision support system for fatty liver disease using GIST descriptors extracted from ultrasound images. Inf Fusion 2016 May;29:32-39. [doi: 10.1016/j.inffus.2015.09.006]
30. Jeon SK, Lee JM, Joo I, Yoon JH, Lee G. Two-dimensional convolutional neural network using quantitative US for noninvasive assessment of hepatic steatosis in NAFLD. Radiology 2023 Apr;307(1):e221510. [doi: 10.1148/radiol.221510] [Medline: 36594835]
31. Neogi N, Adhikari A, Roy M. Use of a novel set of features based on texture anisotropy for identification of liver steatosis from ultrasound images: a simple method. Multimed Tools Appl 2019 May;78(9):11105-11127. [doi: 10.1007/s11042-018-6675-0]
32. Chen JR, Chao YP, Tsai YW, et al. Clinical value of information entropy compared with deep learning for ultrasound grading of hepatic steatosis. Entropy (Basel) 2020 Sep 9;22(9):1006. [doi: 10.3390/e22091006] [Medline: 33286775]
33. Dubois M, Ronot M, Houssel-Debry P, et al. Performance of B-mode ratio and 2D shear wave elastography for the detection and quantification of hepatic steatosis and fibrosis after liver transplantation. Eur J Gastroenterol Hepatol 2020 Feb;32(2):222-230. [doi: 10.1097/MEG.0000000000001500] [Medline: 31464783]
34. Shi X, Ye W, Liu F, et al. Ultrasonic liver steatosis quantification by a learning-based acoustic model from a novel shear wave sequence. Biomed Eng Online 2019 Dec 21;18(1):121. [doi: 10.1186/s12938-019-0742-2] [Medline: 31864367]
35. Jesper D, Klett D, Schellhaas B, et al. Ultrasound-based attenuation imaging for the non-invasive quantification of liver fat - a pilot study on feasibility and inter-observer variability. IEEE J Transl Eng Health Med 2020;8:1800409. [doi: 10.1109/JTEHM.2020.3001488] [Medline: 32617199]
36. McHenry S, Park Y, Browning JD, Sayuk G, Davidson NO. Dallas steatosis index identifies patients with nonalcoholic fatty liver disease. Clin Gastroenterol Hepatol 2020 Aug;18(9):2073-2080. [doi: 10.1016/j.cgh.2020.01.020] [Medline: 31982611]
37. Roy M, Wang F, Vo H, et al. Deep-learning-based accurate hepatic steatosis quantification for histological assessment of liver biopsies. Lab Invest 2020 Oct;100(10):1367-1383. [doi: 10.1038/s41374-020-0463-y] [Medline: 32661341]
38. Sun L, Marsh JN, Matlock MK, et al. Deep learning quantification of percent steatosis in donor liver biopsy frozen sections. EBioMedicine 2020 Oct;60:103029. [doi: 10.1016/j.ebiom.2020.103029] [Medline: 32980688]
39. Chou TH, Yeh HJ, Chang CC, et al. Deep learning for abdominal ultrasound: a computer-aided diagnostic system for the severity of fatty liver. J Chin Med Assoc 2021 Sep 1;84(9):842-850. [doi: 10.1097/JCMA.0000000000000585] [Medline: 34282076]
40. Constantinescu EC, Udri toiu AL, Udri toiu C, et al. Transfer learning with pre-trained deep convolutional neural networks for the automatic assessment of liver steatosis in ultrasound images. Med Ultrason 2021 May 20;23(2):135-139. [doi: 10.11152/mu-2746] [Medline: 33626114]
41. Pérez-Sanz F, Riquelme-Pérez M, Martínez-Barba E, et al. Efficiency of machine learning algorithms for the determination of macrovesicular steatosis in frozen sections stained with sudan to evaluate the quality of the graft in liver transplantation. Sensors (Basel) 2021 Mar 12;21(6):1993. [doi: 10.3390/s21061993] [Medline: 33808978]
42. Pickhardt PJ, Blake GM, Graffy PM, et al. Liver steatosis categorization on contrast-enhanced CT using a fully automated deep learning volumetric segmentation tool: evaluation in 1204 healthy adults using unenhanced CT as a reference standard. AJR Am J Roentgenol 2021 Aug;217(2):359-367. [doi: 10.2214/AJR.20.24415] [Medline: 32936018]
43. Rhyou SY, Yoo JC. Cascaded deep learning neural network for automated liver steatosis diagnosis using ultrasound images. Sensors (Basel) 2021 Aug 5;21(16):16. [doi: 10.3390/s21165304] [Medline: 34450746]
44. Destrempes F, Gesnik M, Chayer B, et al. Quantitative ultrasound, elastography, and machine learning for assessment of steatosis, inflammation, and fibrosis in chronic liver disease. PLoS One 2022;17(1):e0262291. [doi: 10.1371/journal.pone.0262291] [Medline: 35085294]
45. Alshagathrh FM, Musleh S, Alzubaidi M, Schneider J, Househ MS. Efficient detection of hepatic steatosis in ultrasound images using convolutional neural networks: a comparative study. Trait du Signa 2023 Oct 30;40(5):1781-1794. [doi: 10.18280/ts.400501]
46. Podder S, Mallick A, Das S, Sau K, Roy A. Accurate diagnosis of liver diseases through the application of deep convolutional neural network on biopsy images. . 2023(4) p. 453-481. [doi: 10.3934/biophy.2023026]
47. Ibrahim MN, Blázquez-García R, Lightstone A, et al. Automated fatty liver disease detection in point-of-care ultrasound b-mode images. J Med Imaging (Bellingham) 2023 May;10(3):034505. [doi: 10.1117/1.JMI.10.3.034505] [Medline: 37284231]
48. Yao Y, Zhang Z, Peng B, Tang J. Bio-inspired network for diagnosing liver steatosis in ultrasound images. Bioengineering (Basel) 2023 Jun 26;10(7):768. [doi: 10.3390/bioengineering10070768] [Medline: 37508795]
49. Byra M, Styczynski G, Szmigielski C, et al. Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. Int J Comput Assist Radiol Surg 2018 Dec;13(12):1895-1903. [doi: 10.1007/s11548-018-1843-2] [Medline: 30094778]

XSL•FO
RenderX

50. Torgersen J, Akers S, Huo Y, et al. Performance of an automated deep learning algorithm to identify hepatic steatosis within noncontrast computed tomography scans among people with and without HIV. Pharmacoepidemiol Drug Saf 2023 Oct;32(10):1121-1130. [doi: 10.1002/pds.5648] [Medline: 37276449]
51. Wang Q, Lai MW, Bin G, et al. MBR-Net: a multi-branch residual network based on ultrasound backscattered signals for characterizing pediatric hepatic steatosis. Ultrasonics 2023 Dec;135:107093. [doi: 10.1016/j.ultras.2023.107093] [Medline: 37482038]
52. Jeon SK, Joo I, Park J, Yoo J. Automated hepatic steatosis assessment on dual-energy CT-derived virtual non-contrast images through fully-automated 3D organ segmentation. Radiol Med 2024 Jul;129(7):967-976. [doi: 10.1007/s11547-024-01833-8] [Medline: 38869829]
53. Piella G, Farré N, Esono D, et al. LiverColor: an artificial intelligence platform for liver graft assessment. Diagnostics (Basel) 2024 Jul 31;14(15):15. [doi: 10.3390/diagnostics14151654] [Medline: 39125531]
54. Yoo J, Joo I, Jeon SK, Park J, Yoon SH. Utilizing fully-automated 3D organ segmentation for hepatic steatosis assessment with CT attenuation-based parameters. Eur Radiol 2024 Sep;34(9):6205-6213. [doi: 10.1007/s00330-024-10660-4] [Medline: 38393403]
55. Zhang Z, Li G, Wang Z, et al. Deep-learning segmentation to select liver parenchyma for categorizing hepatic steatosis on multinational chest CT. Sci Rep 2024;14(1):11987. [doi: 10.1038/s41598-024-62887-2]
56. Cherchi V, Mea VD, Terrosu G, et al. Assessment of hepatic steatosis based on needle biopsy images from deceased donor livers. Clin Transplant 2022 Mar;36(3):e14557. [doi: 10.1111/ctr.14557] [Medline: 34890087]
57. Wu X, Lv K, Wu S, Tai DI, Tsui PH, Zhou Z. Parallelized ultrasound homodyned-K imaging based on a generalized artificial neural network estimator. Ultrasonics 2023 Jul;132:106987. [doi: 10.1016/j.ultras.2023.106987] [Medline: 36958066]
58. Drazinos P, Gatos I, Katsakiori PF, et al. Comparison of deep learning schemes in grading non-alcoholic fatty liver disease using b-mode ultrasound hepatorenal window images with liver biopsy as the gold standard. Phys Med 2025 Jan;129:104862. [doi: 10.1016/j.ejmp.2024.104862] [Medline: 39626614]
59. Kaffas AE, Bhatraju KC, Vo-Phamhi JM, et al. Development of a deep learning model for classification of hepatic steatosis from clinical standard ultrasound. Ultrasound Med Biol 2025 Feb;51(2):242-249. [doi: 10.1016/j.ultrasmedbio.2024.09.020] [Medline: 39537545]
60. Kim HY, Lee KJ, Lee SS, et al. Diagnosis of moderate-to-severe hepatic steatosis using deep learning-based automated attenuation measurements on contrast-enhanced CT. Abdom Radiol (NY) 2025 Sep;50(9):4139-4147. [doi: 10.1007/s00261-025-04872-5] [Medline: 40095018]
61. Zhang H, Liu J, Su D, et al. Diagnostic of fatty liver using radiomics and deep learning models on non-contrast abdominal CT. PLoS ONE 2025;20(2):e0310938. [doi: 10.1371/journal.pone.0310938] [Medline: 39946425]
62. Derstine BA, Holcombe SA, Chen VL, et al. Quantification of hepatic steatosis on post-contrast computed tomography scans using artificial intelligence tools. Abdom Radiol 2025. [doi: 10.1007/s00261-025-05137-x]
63. Del Corso G, Pascali MA, Caudai C, et al. ANN uncertainty estimates in assessing fatty liver content from ultrasound data. Comput Struct Biotechnol J 2024 Dec;24:603-610. [doi: 10.1016/j.csbj.2024.09.021] [Medline: 39421530]
64. Hernaez R, Lazo M, Bonekamp S, et al. Diagnostic accuracy and reliability of ultrasonography for the detection of fatty liver: a meta-analysis. Hepatology 2011 Sep 2;54(3):1082-1090. [doi: 10.1002/hep.24452] [Medline: 21618575]
65. Park HJ, Kim KW, Kwon HJ, et al. CT-based visual grading system for assessment of hepatic steatosis: diagnostic performance and interobserver agreement. Hepatol Int 2022 Oct;16(5):1075-1084. [doi: 10.1007/s12072-022-10373-0] [Medline: 35789473]
66. Azizi N, Naghibi H, Shakiba M, et al. Evaluation of MRI proton density fat fraction in hepatic steatosis: a systematic review and meta-analysis. Eur Radiol 2025 Apr;35(4):1794-1807. [doi: 10.1007/s00330-024-11001-1] [Medline: 39254718]
67. Tapper EB, Lok ASF. Use of liver imaging and biopsy in clinical practice. N Engl J Med 2017 Aug 24;377(8):756-768. [doi: 10.1056/NEJMra1610570] [Medline: 28834467]
68. Petroff D, Blank V, Newsome PN, et al. Assessment of hepatic steatosis by controlled attenuation parameter using the M and XL probes: an individual patient data meta-analysis. Lancet Gastroenterol Hepatol 2021 Mar;6(3):185-198. [doi: 10.1016/S2468-1253(20)30357-5] [Medline: 33460567]
69. Siddiqui MS, Vuppalanchi R, Van Natta ML, et al. Vibration-controlled transient elastography to assess fibrosis and steatosis in patients with nonalcoholic fatty liver disease. Clin Gastroenterol Hepatol 2019 Jan;17(1):156-163. [doi: 10.1016/j.cgh.2018.04.043] [Medline: 29705261]
70. Hepburn MJ, Vos JA, Fillman EP, Lawitz EJ. The accuracy of the report of hepatic steatosis on ultrasonography in patients infected with hepatitis C in a clinical setting: a retrospective observational study. BMC Gastroenterol 2005 Apr 13;5:14. [doi: 10.1186/1471-230X-5-14] [Medline: 15829009]
71. Fischer MA, Reiner CS, Raptis D, et al. Quantification of liver iron content with CT-added value of dual-energy. Eur Radiol 2011 Aug;21(8):1727-1732. [doi: 10.1007/s00330-011-2119-1] [Medline: 21472472]
72. Loizillon S, Bottani S, Maire A, et al. Automatic quality control of brain 3D FLAIR MRIs for a clinical data warehouse. Med Image Anal 2025 Jul;103:103617. [doi: 10.1016/j.media.2025.103617] [Medline: 40344945]

XSL•FO
RenderX

73. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. Nat Commun 2024 Jan 22;15(1):654. [doi: 10.1038/s41467-024-44824-z] [Medline: 38253604]
74. Sheller MJ, Edwards B, Reina GA, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep 2020 Jul 28;10(1):12598. [doi: 10.1038/s41598-020-69250-1] [Medline: 32724046]
75. Vianna P, Calce SI, Boustros P, et al. Comparison of radiologists and deep learning for US grading of hepatic steatosis. Radiology 2023 Oct;309(1):e230659. [doi: 10.1148/radiol.230659] [Medline: 37787678]
76. Zöggeler T, Kavallar AM, Pollio AR, et al. Meta-analysis of shotgun sequencing of gut microbiota in obese children with MASLD or MASH. Gut Microbes 2025 Dec;17(1):2508951. [doi: 10.1080/19490976.2025.2508951] [Medline: 40396204]
77. Hou M, Zhu Y, Zhou H, et al. Innovative machine learning approach for liver fibrosis and disease severity evaluation in MAFLD patients using MRI fat content analysis. Clin Exp Med 2025 Aug 5;25(1):275. [doi: 10.1007/s10238-025-01818-5] [Medline: 40762814]
78. Wei Y, Yang M, Zhang M, et al. Focal liver lesion diagnosis with deep learning and multistage CT imaging. Nat Commun 2024;15(1):7040. [doi: 10.1038/s41467-024-51260-6] [Medline: 39147767]
79. Dunn N, Verma N, Dunn W. Artificial Intelligence for predictive diagnostics, prognosis, and decision support in MASLD, hepatocellular carcinoma, and digital pathology. J Clin Exp Hepatol 2026;16(1):103184. [doi: 10.1016/j.jceh.2025.103184] [Medline: 41127419]
80. Raji CG, Chandra SSV, Gracious N, Pillai YR, Sasidharan A. Advanced prognostic modeling with deep learning: assessing long-term outcomes in liver transplant recipients from deceased and living donors. J Transl Med 2025 Feb 16;23(1):188. [doi: 10.1186/s12967-025-06183-1] [Medline: 39956905]
81. Guillot J, Williams CYK, Azzam S, et al. Risk prediction in patients with metabolic dysfunction-associated steatohepatitis using natural language processing. Gastro Hep Adv 2025;4(9):100701. [doi: 10.1016/j.gastha.2025.100701] [Medline: 40688387]
82. Zhang J, Feng S, Xue J, et al. AI-driven multimodal fusion of tongue images and clinical indicators for identifying MAFLD patients at risk of coronary artery disease: an exploratory study. ILIVER 2025 Sep;4(3):100181. [doi: 10.1016/j.iliver.2025.100181] [Medline: 41054419]
83. Li B, Tai DI, Yan K, et al. Accurate and generalizable quantitative scoring of liver steatosis from ultrasound images via scalable deep learning. World J Gastroenterol 2022 Jun 14;28(22):2494-2508. [doi: 10.3748/wjg.v28.i22.2494] [Medline: 35979264]
84. Determination of signal-to-noise ratio (SNR) in diagnostic magnetic resonance imaging. : NEMA; 2021 URL: https://www.nema.org/Standards/view/Determination-of-Signal-to-Noise-Ratio-in-Diagnostic-Magnetic-Resonance-Imaging [accessed 2021-03-31]
85. Lee B, Kramer P, Sandri S, et al. Early recalls and clinical validation gaps in artificial intelligence-enabled medical devices. JAMA Health Forum 2025 Aug 1;6(8):e253172. [doi: 10.1001/jamahealthforum.2025.3172] [Medline: 40844774]
86. Li YH, Li YL, Wei MY, Li GY. Innovation and challenges of artificial intelligence technology in personalized healthcare. Sci Rep 2024 Aug 16;14(1):18994. [doi: 10.1038/s41598-024-70073-7] [Medline: 39152194]
87. Bhandari M, Zeffiro T, Reddiboina M. Artificial intelligence and robotic surgery: current perspective and future directions. Curr Opin Urol 2020 Jan;30(1):48-54. [doi: 10.1097/MOU.0000000000000692] [Medline: 31724999]
88. Deshmukh AD, Wagner JK. FDA draft guidelines for AI and the need for ethical frameworks. JAMA Pediatr 2025 Sep 1;179(9):937-938. [doi: 10.1001/jamapediatrics.2025.1979] [Medline: 40622691]

Abbreviations

AI: artificial intelligence
AUC: area under curve
DL: deep learning
HCC: hepatocellular carcinoma
HS: hepatic steatosis
LRN: negative likelihood ratio
LRP: positive likelihood ratio
MAFLD: metabolic dysfunction associated fatty liver disease
ML: machine learning
MRI: magnetic resonance imaging
MRI-PDFF: magnetic resonance imaging-proton density fat fraction
PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Diagnostic Test Accuracy
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search Extension
SROC: summary receiver operating characteristic
TL: transfer learning

XSL•FO
RenderX

Edited by S Brini; submitted 30.May.2025; peer-reviewed by J Soldera, KH Lin, MBR Mahmoud; revised version received 11.Nov.2025; accepted 11.Nov.2025; published 13.Jan.2026.

Please cite as:
Song J, Liu D, Li J, Cong H, Deng R, Lu Y, Sun J, Zhang J
Assessment of the Diagnostic Performance and Clinical Impact of AI in Hepatic Steatosis: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e78310
URL: https://www.jmir.org/2026/1/e78310
doi:10.2196/78310

© Jiamei Song, Dan Liu, Jitong Li, Haoru Cong, Ruixue Deng, Yihan Lu, Jiayi Sun, Jingzhou Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI for Detecting and Predicting Postpartum Depression: Scoping Review

Mais Alkhateeb[1,2,3], PhD; Ajisha Nayeem[2]; Arfan Ahmed[2], PhD; Mohammed Alsahli[4], PhD; Javaid Sheikh[2], MD; Alaa Abd-Alrazaq[2], PhD

[1]College of Education and Art, Lusail University, Doha, Qatar

[2]AI Center for Precision Health, Weill Cornell Medicine-Qatar, Doha, Qatar

[3]Computer Science and Engineering, Hamad bin Khalifa University, Qatar Foundation, Doha, Qatar

[4]Health Informatics Department, College of Health Science, Saudi Electronic University, Riyadh, Saudi Arabia

Corresponding Author:
Mais Alkhateeb, PhD
College of Education and Art, Lusail University, Doha, Qatar

Abstract

Background: Postpartum depression (PPD) affects up to 20% of mothers globally. Early detection is vital for better outcomes, yet screening lacks scalability and predictive power. Artificial intelligence (AI)—through machine learning, deep learning, and natural language processing—enhances the early identification of mothers at risk with greater accuracy.

Objective: This study aims to systematically map the existing literature on AI-based methods for detecting and predicting PPD.

Methods: This scoping review was conducted in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines. We included empirical studies that applied AI techniques to detect or predict PPD and were published in peer-reviewed journals, conference proceedings, or dissertations. Studies were excluded if they were nonempirical (eg, reviews, editorials, and abstracts), not published in English, focused on general perinatal mental health without a specific emphasis on PPD, or used AI solely for monitoring or treatment rather than prediction or detection. We systematically searched 8 databases—MEDLINE, Embase, PsycINFO, CINAHL, Scopus, IEEE Xplore, ACM Digital Library, and Google Scholar—from inception through February 28, 2025. The search strategy was supplemented by backward and forward reference screening and biweekly alerts to capture newly published studies. Two independent (M [Alkhateeb] and A [Nayeem])reviewers (M [Alkhateeb] and A [Nayeem]) screened the retrieved studies, with disagreements resolved by a third reviewer (AA [Alrazaq]). Data were extracted by 2 independent reviewers using a standardized extraction form capturing study characteristics, AI model types, data sources, features, preprocessing, validation strategies, and performance metrics. A formal risk-of-bias assessment was not performed due to the scoping nature of the review. All extracted data were synthesized narratively.

Results: Out of 503 retrieved studies, 65 met the inclusion criteria. The United States contributed the largest proportion of studies (18/65, 27.7%). The highest number of publications occurred in 2024 (17/65, 26%). Most included studies were journal articles (46/65, 71%). Short-term postpartum outcomes (≤12 weeks) were most frequently assessed (20/65, 30.8%). Most included studies (52/65, 80%) applied AI models for predicting PPD, while 14 of 65 (22%) studies used them for detection. Sociodemographic data were most frequently used (49/65, 75.4%), followed by psychological data (44/65, 68%) and obstetric data (35/65, 55%). Data preprocessing mostly relied on basic scaling (51/65, 79%) and some missing data imputation (29/65, 44.6%). Machine learning dominated (57/65, 87.7%), especially random forest, support vector machines, and logistic regression. Internal validation (k-fold, hold-out) was standard, while external validation was scarce. Ensemble-based boosting models consistently demonstrated superior performance across key metrics, highlighting their potential for accurate and scalable PPD prediction. Current studies suffer from limited sample sizes, geographic bias, lack of standardized feature sets, minimal external validation, and inconsistent reporting of comprehensive model metrics.

Conclusions: This scoping review analyzes 65 studies on AI in PPD, highlighting dominant use of classical machine learning, limited deep learning adoption, underuse of advanced preprocessing, inconsistent validation, and reliance on structured, unimodal data—mainly sociodemographic, clinical, and obstetric features.

(J Med Internet Res 2026;28:e77376) doi:10.2196/77376

KEYWORDS

postpartum depression; maternal mental health; perinatal depression; artificial intelligence; machine learning; deep learning; natural language processing; prediction models; computer-aided diagnosis

XSL•FO
RenderX

Introduction

Background

Postpartum depression (PPD) is a common mental health issue affecting new mothers after they give birth. Its salient features include feelings of enduring melancholy, losing interest in hobbies and their everyday life (and, potentially, their baby), and reduced feelings of pleasure in activities.

Traditionally referred to as the "baby blues," PPD is a profound and serious condition undermining the activities of daily life and the psychosocial well-being of mothers. It affects up to a fifth of new mothers worldwide, albeit it is often undiagnosed; in any case, it is a major concern for public health [1].

Postpartum care is vital to ensure the best outcomes for both neonates and mothers. This includes creating a supportive environment with health-promoting activities and breastfeeding encouragement. It must also address each mother's individual mental health needs. [2].

Worldwide, studies of perinatal mental health have noted that PPD is increasingly evident [3]. According to the estimations of the National Institute of Mental Health, up to 15% of all women who experience pregnancy also experience related depression, whether during or after pregnancy. Prevalence is typically higher (ie, 18% - 25%) in low- and middle-income countries, where it is associated with socioeconomic issues and health care resource availability and access, as well as sociocultural factors [4].

The great variety in PPD prevalence underscores the requirement for efficacious strategies for screening women and delivering interventions catering to various needs. The mainstay for PPD detection now is dependent on women reporting classic symptoms and completing self-reported tools, of which the Edinburgh Postnatal Depression Scale (EPDS) [5] and the Patient Health Questionnaire-9 (PHQ-9) [6] are common and effective. However, such tools for screening women during and after pregnancy are typically not administered consistently. For instance, women are often screened once during the early stages of pregnancy, such as the second trimester. However, the same tools are rarely reapplied in later or postpartum periods [7,8].

Furthermore, the tools detect current depression, with no scope to anticipate future risk (based on current symptoms and feelings) [9]. Prediction and early diagnosis of PPD remain challenging, largely because qualitative narrative data are difficult to interpret and integrate alongside quantitative clinical metrics.

A detailed professional analysis is necessary to interpret data appropriately, which is costly, time-consuming, and potentially subjective [10]. PPD prevention and treatment interventions require improved screening solutions that can be delivered during early pregnancy and throughout the pregnancy journey and postpartum period.

Artificial intelligence (AI) can potentially address this impasse, with its capability to handle and process vast volumes of complex, high-dimensional, nonlinear data. Using machine learning (ML), large language models, and natural language processing (NLP) techniques, AI can detect subtle patterns inherent within data that could otherwise evade human analysis [10,11].

AI can enhance prediction accuracy by incorporating diverse data sources. These include electronic health records (EHRs), diagnostic indicators, self-reported feelings, and behavioral cues gathered from digital platforms, with appropriate safeguards [1,12]. Such possibilities render AI a highly useful clinical tool, offering real-time decision-making input for digital care delivery.

Research Problem and Aim

Many studies have developed AI models for detecting and predicting PPD, yet these studies offer fragmented insights into the full potential of AI methodologies. Several previous reviews attempted to summarize these insights [10,13-19], but they have notable limitations. Specifically, some prior reviews were traditional narrative reviews rather than systematic or scoping reviews and thus lacked rigorous, structured methodologies [13,14,16].

In addition, many earlier reviews used narrow search queries or omitted critical databases (eg, PsycINFO, ACM Digital Library, IEEE Xplore, Scopus, and Embase), potentially excluding relevant studies [10,13-19]. Furthermore, past reviews often broadly addressed general depression or women's mental health instead of specifically targeting PPD, limiting their direct relevance [10,15]. Also, the bibliographic searches of previous reviews mostly concluded before September 2022, omitting recent advancements in AI methodologies and multimodal data integration techniques. Importantly, most prior research emphasized traditional clinical and survey-based data, neglecting innovative data sources such as social media and wearable sensors. These novel data sources represent a promising opportunity to enhance AI model accuracy and predictive capabilities for PPD [10,13,14,16-19].

The primary aim of this review is to map the landscape of AI methodologies used in PPD detection and prediction and to identify key research trends, methodological features, and evidence gaps. Specifically, this review is guided by the following research subquestions:

- What types of AI models have been used to detect or predict PPD, and how do they differ in approach and complexity?
- What data modalities (eg, structured, unstructured, physiological, and digital) have been used in these studies?
- How have studies handled model development processes such as feature selection, validation, and interpretability?
- What are the key limitations, challenges, and future opportunities for applying AI to PPD detection in real-world clinical and community settings?

By addressing these questions, this review provides a structured, up-to-date, and integrative overview of AI in postpartum mental health—highlighting opportunities for innovation, responsible deployment, and policy translation in maternal care.

XSL•FO
RenderX

Methods

We conducted a scoping review in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines (see PRISMA-ScR checklist). The following sections detail the specific methods we used in this review.

Search Strategy

All-inclusive searches were done across the following 8 major electronic databases on November 18, 2024, to determine relevant studies: MEDLINE (via Ovid), PsycINFO (via Ovid), Embase (via Ovid), CINAHL (via EBSCO), IEEE Xplore, ACM Digital Library, Scopus, and Google Scholar. To keep our search up to date, we set up an automatic biweekly search alert for 24 weeks, ending on February 28, 2025. Given the vast number of results generated by Google Scholar, we focused on the first 100 results (10 pages), as they are ranked by relevance. In addition to database searches, we expanded our review by manually screening reference lists of included studies (backward reference checking) and identifying studies that cited them (forward reference checking). We also collected additional papers through automatic email alerts. To ensure that the search query was well-structured and effective, 3 experts in digital mental health were consulted and previous relevant literature was reviewed. Two main categories of terms were included in the final search query: AI-related terms (eg, artificial intelligence, machine learning, and deep learning) and PPD-related terms (eg, postpartum depression, postpartum depression, and postnatal depression). A detailed search query used for each database is shown in Multimedia Appendix 1.

Study Eligibility Criteria

This scoping review targeted studies that specifically applied AI to the detection or prediction of PPD. Eligible studies were empirical in nature, used AI methodologies, and were published in peer-reviewed journals, dissertations, or conference proceedings. There were no restrictions regarding publication year, country of origin, data type, study design, population, or outcome type.

Exclusion criteria encompassed nonempirical works such as reviews, abstracts, commentaries, and proposals, as well as studies lacking a specific focus on PPD (eg, addressing broader maternal or perinatal mental health). Studies that used AI solely for managing or monitoring PPD, rather than detecting or predicting it, were also excluded. In addition, only those papers published in English were considered.

Study Selection

The study selection process in this review involved 3 main steps. First, we used EndNote to remove any duplicate studies from our search results. Then, the titles and abstracts of the remaining studies were screened to determine their relevance. For studies that passed this initial screening, a full-text review was conducted, during which the entire paper, including any supplementary materials, was carefully read. To ensure accuracy, 2 independent reviewers (MA and AN) conducted the study selection process. In cases of disagreement during title or abstract screening or full-text review, a third reviewer (AAA) was consulted to resolve the conflict. In addition, we calculated Cohen κ statistic to assess interreviewer agreement, which yielded a value of 0.78-0.83 by title or abstract screening or full-text review—indicating a high level of consistency and reliability in the data selection process [20].

Data Extraction

To ensure a structured and consistent approach to data extraction, we developed a data extraction form, which was pilot-tested using 5 selected studies before full implementation. This form was designed to capture key details related to the study characteristics, datasets, features, and AI methodologies. The finalized data extraction form used in this review is shown in Multimedia Appendix 2. Two independent reviewers (MA and AN) used Microsoft Excel to extract data systematically. Any discrepancies between them were resolved through discussion.

Data Synthesis

We analyzed the extracted data using a narrative synthesis approach, summarizing key findings in descriptive text and tables to provide a clear overview of the research. First, we outlined the basic details of each study, including the year of publication and the country where the research was conducted. Subsequently, we characterized the datasets underpinning AI model development, cataloged the AI methodologies used in each study, and detailed the feature attributes used in model construction. To keep the process structured and ensure accuracy, we used Microsoft Excel to organize and synthesize the extracted data efficiently.

Results

Search Results

As illustrated in Figure 1, a total of 503 records were retrieved through searches across 9 databases: Ovid MEDLINE (n=64), Embase (n=48), PsycINFO (n=26), CINAHL (n=22), IEEE Xplore (n=16), ACM Digital Library (n=2), Scopus (n=145), Web of Science (n=80), and Google Scholar (n=100, limited to the top 100 results ranked by relevance). After removing 272 duplicate records using reference management software, 231 unique reports remained for screening. After reviewing the titles and abstracts, 145 records were excluded. The full texts of the remaining 86 records were retrieved for further assessment. Of these, 7 full-text papers were not available. After evaluating the 79 available full-text papers, 17 studies were excluded for the following reasons: they did not use AI (n=7); did not focus on PPD (n=1); were not journal papers, conference papers, or dissertations (n=8); or were not written in English (n=1). Three additional relevant studies were identified through both backward and forward reference list screening. Ultimately, 65 studies were included in this review [21-85].

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram illustrating the study selection process. AI: artificial intelligence; PPD: postpartum depression.

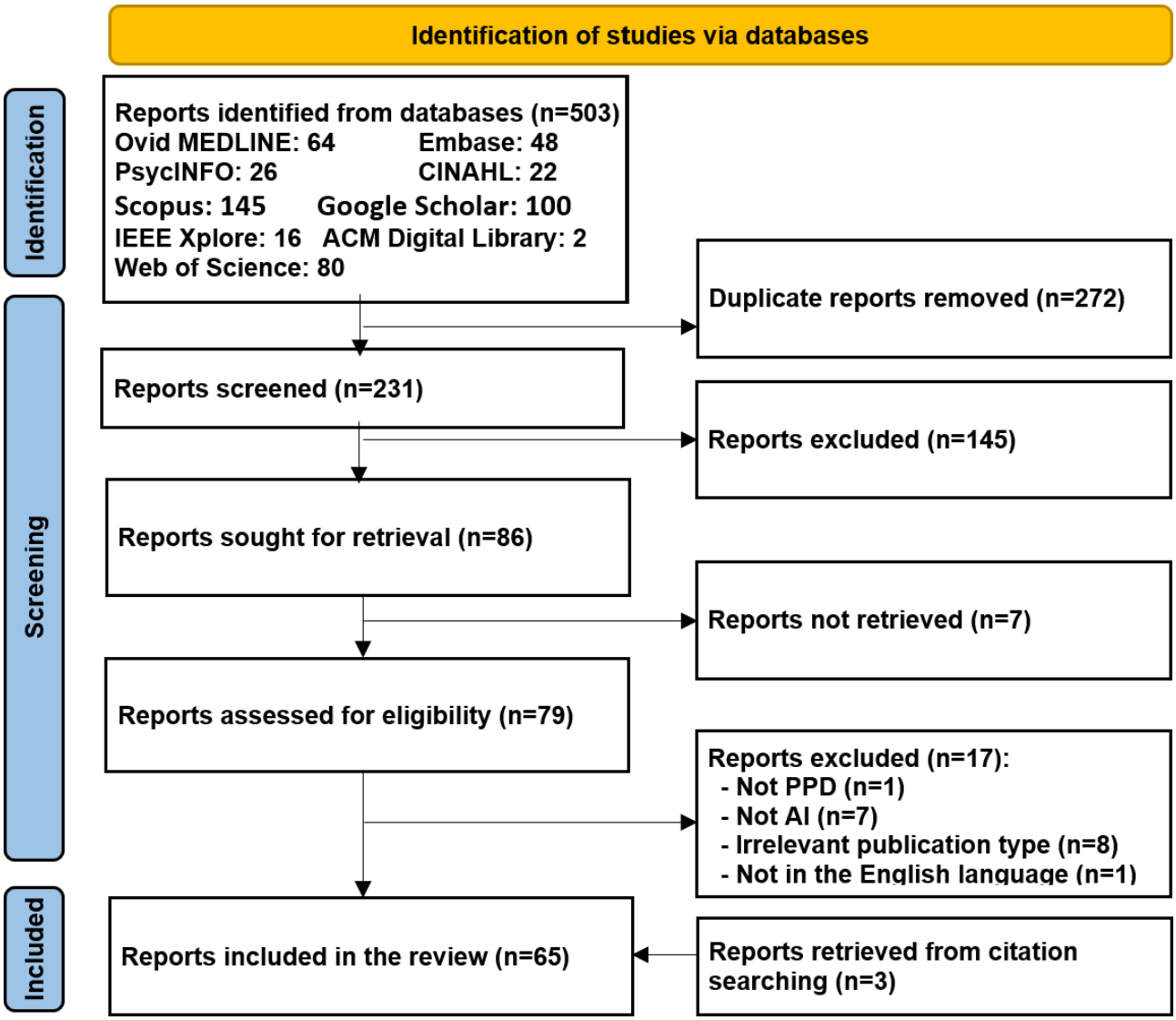

Characteristics of the Included Studies

As shown in Table 1, the included studies were published between 2009 and 2025, with the highest number of publications occurring in 2024 (26.1%). Regarding publication types, the majority were journal papers (70.8%). The United States contributed the largest proportion of studies (27.7%), followed by China (15.4%) and Bangladesh (13.9%). This review included 39 out of 65 (60%) retrospective studies and 27 out of 65 (41.5%) prospective studies. The number of participants in the included studies ranged from 11 to 573,634, with a mean of 18,187.4 (SD 72,933.8). Participant distribution was as follows: 23 studies had fewer than 500 participants, 27 studies included between 500 and 5000 participants, 8 studies had between 5001 and 50,000 participants, and 7 studies involved more than 50,000 participants. Among the 26 studies that reported mean participant age, values ranged from 26 to 44.5 years, with an overall mean average of 31.08 (SD 3.42) years. See Multimedia Appendix 3 for overall information about included studies.

XSL•FO
RenderX

Table . Characteristics of the included studies.

Key aspects	Studies	References
Year of publication, n/N (%)		
2025	3/65 (4.6)	[30,46,77]
2024	17/65 (26.1)	[21,24,35,37,40,41,43,45,51,52,65-67,69,78,83,85]
2023	14/65 (21.5)	[28,32,33,42,44,50,56,63,68,72-74,79,80]
2022	10/65 (15.4)	[34,36,39,47,58-62,81]
2021	5/65 (7.7)	[22,26,48,55,84]
2020	6/65 (9.2)	[23,31,57,64,71,82]
2019	4/65 (6.2)	[25,29,49,76]
2018	3/65 (4.6)	[27,54,75]
≤2017	3/65 (4.6)	[38,53,70]
Publication type, n/N (%)		
Journal paper	46/65 (70.8)	[22,23,25,28-31,33,34,37-44,46,47,49-52,54,55,57,60,62-65,67-70,74,76-85]
Conference paper	18/65 (27.7)	[21,24,26,27,32,35,45,48,53,56,58,59,61,66,71-73,75]
Dissertation	1/65 (1.5)	[36]
Country of publication, n/N (%)		
United States	18/65 (27.7)	[24,25,31,37,39,41,43,51,53,55,57,62,64,74,76,80,83,84]
China	10/65 (15.4)	[27,42,63,69,75,77-79,82,85]
Bangladesh	9/65 (13.9)	[21,35,40,45,52,61,65,67,73]
India	6/65 (9.2)	[32-34,48,56,66]
Japan	3/65 (4.6)	[46,47,81]
United Kingdom	2/65 (3.1)	[44,71]
Spain	2/65 (3.1)	[38,70]
Sri Lanka	2/65 (3.1)	[58,59]
Others[a]	13/65 (1 each) (20)	[22,23,26,28-30,36,49,50,54,60,68,72]
Research design[b], n/N (%)		
Retrospective	39/65 (60)	[21,23-27,30,32,35,37,40-45,47-49,51-56,60,61,64-66,71-76,80,83,84]
Prospective	27/65 (41.5)	[22,28,29,31,33,34,36,38,39,46,50,51,57-59,62,63,67-70,77-79,81,82,85]
Number of participants		
Mean (SD)	18,187.4 (72,933.8)	[21-85]
Range	11 - 573,634	[21-85]
<500	23/65 (35.4)	[24,26-28,30,31,33,34,37,48,49,53,54,57,61,66,68,71,72,77,79,81,85]
500 - 5000	27/65 (41.5)	[21,22,25,29,35,38-42,45,50,52,58-60,62,63,65,67,69,70,73,75,78,80,82]
5001 - 50,000	8/65 (12.3)	[32,36,47,51,56,64,74,76]
>50,000	7/65 (10.8)	[23,43,44,46,55,83,84]
Mean age (years)[c]		
Mean (SD)	31.08 (3.42)	[21-85]
Range	25.99 - 44.5	[21-85]

[a]Others include Australia, Brazil, Indonesia, Italy, Mexico, Nigeria, Norway, Pakistan, Palestine, Portugal, Saudi Arabia, Slovenia, and Sweden.

[b]The number of studies does not add up as one study used both retrospective and prospective designs.

[c]Mean age not reported: 39 studies (60%).

XSL•FO
RenderX

Characteristics of Datasets

As depicted in Table 2, the average dataset size was 37,338.5 (SD 160,309.6), with a range spanning from 16 to 1,170,446. The dataset size fell between 500 and 5000 in 28 out of 65 (43.1%) studies. Most studies (49/65, 75.4%) used closed-source data, while the remaining (16/65, 24.6%) relied on open-source datasets. The studies included various data formats, including textual, tabular, audio, video, and images. The majority (57/65, 87.7%) used a single data format (unimodal), while the remaining studies (8/65, 12.3%) integrated multiple data formats (multimodal). The survey was the most common data collection approach (50/65, 76.9% of studies), followed by data sourced from EHRs (25/65, 38.5% of studies). Most studies (45/65, 69.2%) were conducted in health care settings. Regarding the timing of outcome measurement for PPD, 20 out of 65 (30.8%) studies assessed outcomes within 12 weeks of delivery (short term), 14 out of 65 (21.5%) studies between 12 and 36 weeks (medium term), and 17 out of 65 (26.2%) studies after more than 36 weeks (long term). The most common reference standard used for labeling the data (outcomes) was the EPDS (37/65, 56.9% of studies). Further details on the characteristics of the datasets used in the included studies are provided in Multimedia Appendix 4.

XSL•FO
RenderX

Table . Characteristics of datasets used in the included studies.

Data summary	Studies, n (%)	References
Dataset size[a]		
Mean (SD)	37,338.5 (160,309.6)	[21-85]
Range	16 - 1,170,446	[21-85]
Dataset size categories, n/N (%)		
<500	20/65 (30.8)	[24,26,28,30,33,34,37,48,49,53,54,57,61,66,68,71,72,79,81,85]
500 - 5000	28/65 (43.1)	[21,22,25,29,31,35,38-42,45,50,52,58-60,62,63,65,67,69,70,73,77,78,80,82]
5001 - 50,000	8/65 (12.3)	[32,36,47,56,64,74-76]
>50,000	9/65 (13.9)	[23,27,43,44,46,51,55,83,84]
Data source		
Closed	49/65 (75.4)	[22-27,29-32,36-39,42-44,46-51,53-55,57-63,68-72,74-79,81-85]
Open	16/65 (24.6)	[21,28,33-35,40,41,45,52,56,64-67,73,80]
Data format[b]		
Unimodal	57/65 (87.7)	[21-26,29,30,32,34-40,42-50,52,54-66,68-76,79-85]
Multimodal	8/65 (12.3)	[27,28,31,33,41,51,53,67]
Data collection methodology[c]		
Survey	50/65 (76.9)	[21,22,24-28,30,31,33-36,38,40-42,44-46,50-54,56-66,67-73,75,78-82,85]
EHRs[d]	25/65 (38.5)	[23,37,41-44,46,47,49,50,55,67-70,73,74,76-79,81,83-85]
Social media	8/65 (12.3)	[27,29,32,33,51,53,66,67]
Sensor-based	5/65 (7.7)	[37-39,44,49]
Laboratory-based data	2/65 (3.1)	[79,81]
Setting[c]		
Health care	45/65 (69.2)	[21-23,28,30,33-35,37-43,45-47,49-52,54,55,57-59,63,65,67-71,73-79,82-85]
Community	18/65 (27.7)	[25,27,29,31,32,36,48,51,53,56,60-62,64,67,72,80,81]
Academic	5/65 (7.7)	[24,26,37,44,66]
Outcome measurement timing (weeks)[e]		
Short term (<12)	20/65 (30.8)	[34,38,42,46,47,50,52,55,57,62,63,67,68,71,74,77,78,81,82,85]
Medium term (12 - 36)	14/65 (21.5)	[36,46,50,57-59,62,63,70,72,78,80,81,85]
Long term (>36)	17/65 (26.2)	[22-24,36,37,43,48,53,57,63,64,69,76,80,83-85]
Reference standard		
EPDS[f]	37/65 (56.9)	[22,24,26-30,34,38-40,42,46,47,50,52,53,55-59,61-63,67-72,76-78,80-82]
ICD[g]	8/65 (12.3)	[2,23,29,41,42,44,49,76,83]
PHQ[h]	7/65 (10.8)	[3,34,43,60,61,64,67]
PDSS[i]	3/65 (4.6)	[4,25,34,48]
PPDS[j]	2/65 (3.1)	[5,32,67]

[a]Mean (SD) is calculated.

[b]The number of studies does not add up as some studies used multiple data collection methodologies.

[c]The number of studies does not add up as some studies are conducted in multiple settings.

[d]EHRs: electronic health records.

[e]The number of studies does not add up as the timing of outcome measurement varied across studies. Outcome measurement timing not reported: 27 (41.5%).

[f]EPDS: Edinburgh Postnatal Depression Scale.

XSL•FO
RenderX

[g]*ICD*: *International Classification of Diseases*.

[h]PHQ: Patient Health Questionnaire.

[i]PDSS: Postpartum Depression Screening Scale.

[j]PPDS: Postpartum Depression Scale.

Characteristics of Preprocessing Techniques

Table 3 summarizes the most frequently used preprocessing techniques identified across the reviewed studies. Across the reviewed literature, feature transformation overwhelmingly dominates preprocessing: Min-Max scaling and Z score standardization appear in 78.5% (51/65) of studies. In contrast, missing data strategies remain underutilized—only 44.6% (29/65) of papers applied any form of imputation, leaving 33.8% (22/65) to rely on case deletion or ignore the issue entirely. Class imbalance remedies are similarly rare: just 4.6% (3/65) of studies used stratified resampling or SMOTE variants, while cost-sensitive learning appeared in only 6.2% (4/65).

Categorical encoding methods vary in popularity: label encoding leads at 29.2% (19/65), one-hot encoding in 13.9% (9/65), binary encoding in 9.2% (6/65), dummy encoding in 6.2% (4/65), and target encoding in a mere 3.1% (2/65). For feature selection, tree-based importance (Gini impurity) featured in 18.5% (12/65) of studies and Pearson correlation filtering in 12.3%, with recursive feature elimination (5/65, 7.7%), information-gain ratio (6.2%), and SHAP-based methods (4/65, 6.2%) trailing behind.

Finally, dimensionality reduction and specialized feature extraction remain fringe techniques: sequential floating forward selection was used in only 7.7% (5/65) of papers and principal component analysis (PCA) in 6.2% (4/65), text vectorization methods (eg, N-grams, TF-IDF) in 4.6% (3/65), domain-specific statistical features in 3.1% (2/65), and acoustic‑signal processing (MFCC) in just 1 study (1/65, 1.5%). For a comprehensive overview of dataset characteristics used in the studies, refer to Multimedia Appendix 5.

XSL•FO
RenderX

Table . Characteristics of datasets used in the included studies.

Preprocessing techniques	Studies, n/N (%)	References
Dimensionality reduction techniques		
Sequential floating forward selection and SHAP[a]	5/65 (7.7)	[51,56,71,75,82]
Principal component analysis	4 /65 (6.2)	[36,52,77,81]
Others (each one 1)[b]	1/65 (1.5)	[33,57,64,66,67,72]
Feature extraction[c]		
Psycholinguistic and N-gram text vectorization	3/65 (4.6)	[29,33]
Domain-specific statistical	2/65 (3.1)	[51,56]
Acoustic signal feature extraction	1/65 (1.5)	[31]
Feature selection		
Regularization of the model		
Pearson correlation	8/65 (12.3)	[33,36,41,44,45,48,54,61]
Spearman rank filtering	2/65 (3.1)	[58,59]
Chi-square independence test	2/65 (3.1)	[21,65]
Cox proportional hazards and Kaplan-Meier survival analysis	1/65 (1.5)	[24]
Wrapper and tree-based selection		
Gini importance/mean decrease in impurity	12/65 (18.5)	[22,23,25,40,46,52,71,73,82]
Recursive feature elimination with cross-validation	5/65 (7.7)	[30,40,64,73,74]
Entropy-based information gain ratio	4/65 (6.2)	[25,30,41,46]
SHAP value-based importance via differential evolution	4/65 (6.2)	[37,43,62,63]
Other metaheuristic and ensemble filters[d]	1/65 (1.5)	[64,71,75,84]
Encoding approaches		
Label encoding	19/65 (29.2)	[23-25,29,31,33,38,42,48,50,51,53,61,63,64,66-68,73]
One-hot encoding	9/65 (13.9)	[26,27,32,36,41,43,52,56,65]
Binary encoding	6/65 (9.2)	[22,34,40,50,74,76]
Dummy encoding	4/65 (6.2)	[47,70,78,84]
Target encoding	2/65 (3.1)	[41,59]
Handling unbalanced data		
Manual resampling	15/65 (23.1)	[23,29,31,47,50,52,53,56,58,60,61,64,70,71,73]
Class weighting and cost-sensitive	4/65 (6.2)	[55,63,69]
Random oversampling (eg, SMOTE[e] variants)	3/65 (4.6)	[36,56,76]
Handling missing data		
Statistical imputation (mean/median/KNN[f]/MICE[g])	29/65 (44.6)	[22-24,31,33,34,36,38,41,43,44,46,51-54,58,60,62,65-67,70,73,78,80,82,84,85]
Complete case analysis (listwise deletion)	22/65 (33.8)	[23,27-29,34,36-38,40,45,47,50,51,54,64,74-79,81]
Feature transformation techniques		
Min–max scaling and Z-score standardization	51/65 (78.5)	[22-24,29-32,34,36-38,40,41,43,49-52,54,56,58,59,67-71,74-76,78-82,84,85]
Text tokenization, lemmatization, and stopword removal	23/65 (35.4)	[26,27,29,32,33,36,40,41,47,50-52,64,65,67,72]

XSL•FO
RenderX

Preprocessing techniques	Studies, n/N (%)	References
Log/power transforms (Box-Cox, Yeo-Johnson)	9/65 (13.8)	[25,31,40,43,45,46,66,70]
Polynomial and interaction feature generation	9/65 (13.8)	[25,31,40,43,45,46,66,70]

[a]SHAP: SHapley Additive exPlanations to interpret model.

[b]Others include linear discriminant analysis (LDA), t-SNE, latent semantic analysis (LSA), latent Dirichlet allocation (LDA), spatial feature extraction, and relief algorithm.

[c]Psycholinguistic text vectorization includes N-gram characteristics; linguistic inquiry and word count for emotion, cognition, social content; LDA topics; and TF-IDF. Acoustic signal feature extraction includes MFCC, spectral contrast, and chroma.

[d]Others include bagging-based selection-by-filter methods, sequential floating forward selection, sequential forward selection, relief algorithm, and Boruta algorithm.

[e]SMOTE: Synthetic Minority Oversampling Technique.

[f]KNN: K-Nearest neighbor algorithm.

[g]MICE: Multiple Imputation by Chained Equations to handle missing data.

Characteristics of Features Used in Included Studies

The reviewed studies incorporated 9 distinct categories of data in AI model development (Table 4). Sociodemographic data were most frequently used (49/65, 75.4% of studies), followed by psychological data (44/65, 67.7% of studies), obstetric data (36/65, 55.4% of studies), and behavioral data (23/65, 35.4% of studies). The number of features used varied significantly across studies, ranging from 2 to 988, with a mean average of 44.88 (SD 129.72). Notably, nearly two-thirds of the studies (43/65, 66.2%) used fewer than 26 features.

Within each data type, the most commonly used individual features were age for sociodemographic data (37/65, 56.9% of studies), mode of delivery for obstetric data (15/65, 23.1% of studies), maternal anxiety for psychological data (13/65, 20% of studies), breastfeeding status for behavioral data (11/65, 16.9% of studies), linguistic inquiry and word count (LIWC) features—such as positive emotions (“happy”), cognitive processes (“think”), and personal pronouns (“I” and “we”)—for linguistic data (11/65, 16.9% of studies), metabolic pathways and circulating markers for biomarker data (4/65, 6.2% of studies), newborn gender for neonatal data (11/65, 16.9% of studies), hypertensive disorders for medical history data (11/65, 16.9% of studies), and tweet metadata for sensor-based data (3/65, 4.6% of studies). Additional characteristics of the datasets used in the reviewed studies are shown in Multimedia Appendix 6.

XSL•FO
RenderX

Table . Characteristics of features used in the included studies.

Features characteristics	Studies, n/N (%)	References
Data type[a]		
Sociodemographic	49/65 (75.4)	[21-24,28,30-32,34-36,38,40-50,52,53,55,56,58-66,67-71,74,76-78,80,82-85]
Psychological	44/65 (67.7)	[21-24,26,28-30,34-36,38,40-46,48,50-54,57,60-66,67,70-74,76,78,80,82-84]
Obstetric	36/65 (55.4)	[22-24,30,34,36,41,43,44,46-49,51,55-61,63,64,67-71,74,76-78,80,83-85]
Behavioral	23/65 (35.4)	[22,25,32,34,36,42-44,47,50,51,53,55,56,58-61,63,64,67,80,82]
Medical history	17/65 (26.2)	[22,23,43,46,47,49,56,63,67,68,71,76,78,80,83-85]
Neonatal	16/65 (24.6)	[22,28,30,38,47,50,53,56,61,63,64,70-72,78,85]
Linguistic	9/65 (13.9)	[26,27,29,31,33,39,66,67,75]
Biomarkers	7/65 (10.8)	[46,57,68,69,77,79,81]
Sensor-based	5/65 (7.7)	[37,44,51,60,66]
Number of features[b]		
Mean (SD)	44.9 (129.7)	[21-85]
Range	2 - 988	[21-85]
Feature range		
≤25	43/65 (66.2)	[21,23-29,31-33,35-38,40-42,45,49-52,54-60,65-71,73,74,78,79,82,85]
26 - 50	16/65 (24.6)	[22,30,34,43,47,48,53,61,63,64,72,76,77,80,83,84]
>50	6/65 (9.2)	[39,44,46,62,75,81]
Data input features sociodemographic data[a]		
Age	37/65 (56.9)	[21-24,28,30-32,34,35,38,40,41,43,45,47-50,52,55,58-61,63,65,68-71,74,78,80,82,85]
Education level	21/65 (32.3)	[22,24,28,30,32,34,36,41,46,48,58-61,63,64,68,74,80,82,85]
Marital status	20/65 (30.8)	[22,30,34,41,43,46-48,53,58,59,61,63,64,68,70,76,83-85]
Monthly income	13/65 (20)	[30,34,36,38,41,50,60,61,64,70,78,82,85]
Employment status	11/65 (16.9)	[22,28,30,43,46,48,61,63,64,70,85]
Obstetric data[a]		
Mode of delivery	15/65 (23.1)	[30,32,34,41,47,48,55,61,68,74,78,80,83-85]
Parity	11/65 (16.9)	[22,30,36,43,46,63,64,68,78,80,85]
Gestational age	9/65 (13.9)	[24,30,34,47,49,61,68,71,78]
Gravida	7/65 (10.8)	[24,30,49,60,68,83,84]
Obstetric complications	6/65 (9.2)	[23,34,43,61,69,85]
Psychological data[a]		
Maternal anxiety	13/65 (20)	[21,25,27,35,40,45,48,52,62,65,71,76,83]
Depression history	12/65 (18.5)	[22,30,34,41,43,44,48,53,55,63,69,82]
Feeling of guilt	9/65 (13.9)	[21,25,27,35,40,45,52,65,73]
Feeling sad	8/65 (12.3)	[21,27,35,40,45,52,54,65]
Sleeping disorders	7/65 (10.8)	[25,27,46,47,53,54,62]
Behavioral data[a]		
Breastfeeding status	11/65 (16.9)	[23,28,34,47,48,53,56,61,64,78,85]
Problems bonding with baby	9/65 (13.9)	[21,35,40,45,48,52,61,65,73]
Planned pregnancy	8/65 (12.3)	[32,34,48,53,61,74,80,85]
Smoking status	7/65 (10.8)	[22,23,46,47,60,63,64]
Alcohol use	6/65 (9.2)	[22,36,46,47,63,64]

XSL•FO
RenderX

Features characteristics	Studies, n/N (%)	References
Linguistic data[a]		
LIWC[c] features	11/65 (16.9)	[27]
Speech and acoustic	8/65 (12.3)	[31]
Emotional and cognitive expression	4/65 (6.2)	[33]
Language models	2/65 (3.1)	[39]
Tweet attributes language	2/65 (3.1)	[66]
Biomarker data[a]		
Metabolic pathway	4/65 (6.2)	[81]
Circulating biomarkers	4/65 (6.2)	[46,57,68]
Neurological	3/65 (4.6)	[69,79]
Protein-related	2/65 (3.1)	[77]
Genetic/epigenetic	1/65 (1.5)	[57]
Neonatal data[a]		
Newborn gender	11/65 (16.9)	[30,31,38,47,50,58,59,61,64,70,85]
Birth weight	8/65 (12.3)	[30,34,41,47,64,71,77,78]
Preterm birth	6/65 (9.2)	[43,48,58,59,76,77]
Health of baby	4/65 (6.2)	[28,30,53,85]
Apgar scores	3/65 (4.6)	[47,71,78]
Medical history[a]		
Hypertension disorders	111/65 (16.9)	[22,43,47,49,56,63,76,78,80,83,84]
Gestational diabetes	44/65 (6.2)	[43,49,78,80]
Migraine	44/65 (6.2)	[22,63,83,84]
Preeclampsia	33/65 (4.6)	[43,83,84]
Hypothyroidism	33/65 (4.6)	[83-85]
Sensor-based[b]		
Tweet metadata	33/65 (4.6)	[66]
Activity intensity	33/65 (4.6)	[37,60]
Calories burned	11/65 (1.5)	[37]
Heart rate	11/65 (1.5)	[37]

[a]The number of studies does not add up as certain features are reported in multiple studies within each category, resulting in repeated counts.

[b]The sensor-based category includes only 4 features.

[c]LIWC: linguistic inquiry and word count.

Characteristics of AI Techniques

As shown in Table 5, the most included studies (52/65, 80%) used AI models for predicting PPD (ie, identifying women at risk of developing PPD in the future), while 14 out of 65 (21.5%) studies used them for detection (ie, identifying whether a woman is currently experiencing PPD). Most studies leveraged ML techniques (57/65, 87.7%), whereas DL techniques were applied in 11 out of 65 (16.9%) studies. The predominant application of AI models was in classification tasks (eg, identifying the presence, absence, or severity level of PPD). In contrast, 5 out of 65 (7.7%) studies used AI models for regression tasks (eg, detecting EPDS score). Various AI algorithms were used in the included studies, with random forest (RF) being the most common (29/65, 44.6%), followed by support vector machine (26/65, 4%) and logistic regression (LogR) (23/65, 35.4%).

The most frequently used optimization strategy among the included studies was stochastic gradient descent (9/65, 13.9%), followed by Adam (7/65, 10.8%) and learning rate scheduling (6/65, 9.2%). The most applied regularization and model stabilization techniques were L1/L2 regularization (9/65, 13.9%) and grid search (9/65, 13.9%). To validate AI model performance, both k-fold cross-validation and holdout validation were the most widely adopted approaches (32/65, 49.2%).

XSL•FO
RenderX

Accuracy was the most reported performance metric (49/65, 75.4%), followed by sensitivity (48/65, 73.9%) and area under the curve (AUC) (41/65, 63.1%). Additional characteristics of the Model of Characteristics of AI Techniques used in the reviewed studies are shown in Multimedia Appendix 7.

XSL•FO
RenderX

Table . Characteristics of artificial intelligence techniques used in the included studies.

Characteristics	Studies, n/N (%)	References
AI[a] algorithm aim		
Prediction	52/65 (80)	[21-27,29,30,32-36,39-47,49,50,52-55,57-60,62,64-69,72-74,76-78,80-85]
Detection	14/65 (21.5)	[28,31,37,38,48,51,56,61,63,69-71,75,79]
AI category		
Machine learning	57/65 (87.7)	[21-25,29-39,41-62,64,66-70,73-85]
Deep learning	11/65 (16.9)	[26-28,32,40,41,58,59,63,65,71]
Transfer learning	3/65 (4.6)	[26,40,41]
Natural language processing	3/65 (4.6)	[33,39,72]
Reinforcement learning	2/65 (3.1)	[63,69]
Problem solving approach		
Classification	60/65 (92.3%)	[21-35,37-43,45-53,55-67,69-76,78-85]
Regression	5/65 (7.7%)	[36,44,54,68,77]
AI model type		
Ensemble methods		
Bagging		
Random forest	29/65 (44.6)	[21,24,33,34,37,41,43,47,48,51,52,55,56,58,61,64,65,71,73,74,76,78,80,82,84,85]
Bagging	1/65 (1.5)	[48]
Extreme random trees	1/65 (1.5)	[73]
Extra trees	1/65 (1.5)	[22]
Boosting		
XGBoost	15/65 (23.1)	[21,24,30,36,41,43,45,55,61,65,68,76,80,84,85]
Gradient boosting	10/65 (15.4)	[22,23,30,51,56,60-62,68,73]
AdaBoost	9/65 (13.9)	[33,41,45,48,53,56,64,65,73]
CatBoost	6/65 (9.2)	[35,41,45,65,68,73]
LightGBM	4/65 (6.2)	[35,45,65,68]
Stacking		
Stacking ensemble	1/65 (1.5)	[21]
Stacking model	1/65 (1.5)	[52]
Nested stacking	1/65 (1.5)	[73]
Neural networks		
Multilayer perceptron	11/65 (16.9)	[43,52,56,67,70,84]
Recurrent neural	6/65 (9.2)	[26,27,40,41,66]
Convolutional neural	5/65 (7.7)	[28,31,36,64,80]
Natural language processing	1/65 (1.5)	[72]
Classification models		
Support vector machine	26/65 (40)	[21,26,29,32,33,37,38,43,47,49-51,53,56-61,64,75,76,78,79,82,85]
Decision tree	18/65 (27.7)	[21,24,25,30,41,46,48,49,51-53,56,60,65,76,78,84,85]
K-Nearest neighbors	9/65 (13.9)	[21,37,49,51,52,56,64,65,78]
Recursive partitioning	1/65 (1.5)	[64]
Probabilistic classification		
Logistic regression	23/65 (35.4)	[21,23,29,33,38,42-44,47,50,53,55,56,61,64,65,70,74,76,77,83-85]
Naive Bayes	7/65 (10.8)	[32,34,38,50,53,60,64]

XSL•FO
RenderX

Characteristics	Studies, n/N (%)	References
Linear regression models		
Ridge regression	6/65 (9.2)	[22,34,44,47,52,68]
LASSO regression	5/65 (7.7)	[22,34,39,44,77]
Elastic net	5/65 (7.7)	[23,36,44,47,68]
Support vector regression	2/65 (3)	[68]
Kernel regression	1/65 (1.55)	[68]
Optimization strategies (gradient-based optimization)		
Stochastic gradient descent	9/65 (13.9)	[27-29,32,36,40,56,64,66]
Adam	7/65 (10.8)	[27,32,36,41,48,56,66]
Learning rate scheduling	6/65 (9.2)	[23,27,29,32,41,56]
AdamW	2/65 (3.1)	[28,40]
Cosine Annealing	2/65 (3.1)	[40,48]
Momentum	1/65 (1.5)	[36]
Regularization and model stabilization		
L1/l2 regularization	9/65 (13.9)	[23,28,29,36,42,44,76,77,83]
Grid search	4/65 (6.2)	[43,65,79,81]
Dropout	8/65 (12.3)	[27,36,40,41,48,56,59,64]
Batch normalization	3/65 (4.6)	[36,40,41]
Early stopping	2/65 (3.1)	[28,36]
Weight decay	2/65 (3.1)	[36,56]
Osprey optimization	1/65 (1.5)	[67]
Validation techniques		
K-fold cross-validation	32/65 (49.2)	[22,23,25,29,30,35,37,39,43,47-49,52,53,61-65,68,69,71,73-77,79,82-84]
Holdout validation	32/65 (49.2)	[21,24,26-29,31-34,36,38-43,45-47,50,51,53-56,58-60,70,78,81]
Leave-one-out cross-validation	1/65 (1.5)	[57]
Nested cross-validation	1/65 (1.5)	[44]
ML[b] performance measures		
Accuracy	49/65 (75.4)	[21,22,24,26-35,38,40,41,43-46,48,49,52,54-67,69-71,73,74,76-80,82-85]
Sensitivity	48/65 (73.9)	[21,22,24-26,29-35,37,38,40-46,49-65,67,70,71,73,75,76,79,82-84]
AUC[c]	41/65 (63.1)	[22-25,31,34,37-51,53,55-57,59-61,64,70,74-84]
Precision	36/65 (55.4)	[21,22,24-26,29-35,37,40,41,43-45,50,53,56,57,59-62,64,65,67,73,78,82-85]
Specificity	23/65 (35.4)	[22,30,31,34,37,38,42,44-46,48,51,55,63,64,70,71,73,75,76,79,82,84]
Geometric mean	7/65 (10.8)	[38,50,51,61,62,69,70]
Negative predictive value	7/65 (10.8)	[22,32,34,44,45,82,84]
F_1-score	5/65 (7.7)	[21,35,58,69,73]
Root mean squared error/mean squared error	3/65 (4.6)	[36,58,68]

[a]AI: artificial intelligence.

[b]ML: machine learning.

[c]AUC: area under the curve, ROC-AUC (receiver operating characteristic) that plots true-positive rate against false-positive rate at different threshold settings.

Among the AI models evaluated in Table 6, ensemble methods emerged as the top performers, with an average accuracy of 93.4%, an F_1-score of 92.5%, and an AUC of 89.4%. Among gradient boosting techniques, CatBoost achieved the highest

XSL•FO
RenderX

AUC of 98.6%, alongside robust accuracy and F_1-score metrics. LightGBM also demonstrated strong performance, recording 92.6% accuracy, an F_1-score of 87.8%, and an AUC of 91.1%, highlighting its scalability and effectiveness. XGBoost delivered competitive results, with an accuracy of 89.1% and an AUC of 86.8%. Convolutional neural networks (CNNs) showed excellent performance as well—particularly in accuracy (92%) and F_1-score (95.1%)—although they were evaluated in a smaller subset of studies.

Traditional tree-based models, including RFs and recursive partitioning, also showed moderate to strong performance. RFs achieved an average accuracy of 80.5% and an AUC of 82.4%, while broader tree-based classifiers averaged 82.8% accuracy and 82.6% AUC. Recursive partitioning, however, showed lower accuracy (71.8%) and AUC (74.7%) across the few studies assessed.

Across all models included, the mean performance was 81.7% (SD 11.05) for accuracy, 80.51% (SD 15.44) for F_1-score, and 81.0% (SD 12.0) for AUC. Collectively, these findings underscore the strong predictive capabilities of DL architectures and ensemble-based approaches—especially boosting models—in detecting PPD, consistently outperforming conventional ML algorithms across most evaluation metrics or detailed information on performance metrics (accuracy, F_1-score, and AUC) (Multimedia Appendix 8).

XSL•FO
RenderX

Table . Accuracy, F_1-score, and area under the curve of artificial intelligence models used in postpartum depression prediction.

Metrics[a]	Accuracy			F_1-score			AUC[b]		
Model	Studies, n	Mean (SD)	Range	Studies, n	Mean (SD)	Range	Studies, n	Mean (SD)	Range
Random forest	23	80.5 (9.2)	59-96	19	80.9 (13.7)	39.3-95	25	82.4 (9.2)	65.1-98
Decision trees	16	82.8 (9.1)	68.8-98.1	12	83.9 (10.5)	66-98.58	13	82.6 (7.9)	69-97.6
XGBoost	13	89.1 (10.3)	67.6-100	8	91.8 (11.4)	66-100	7	86.8 (11.1)	73-100
LightGBM	6	92.6 (11)	70.2-98.4	6	87.8 (22.7)	41.6-98.6	4	91.1 (12.3)	72.7-98
AdaBoost	7	78.7 (9.6)	66-94	6	75.4 (11.6)	58-89	6	78.5 (5.7)	69-85.7
CatBoost	5	93.6 (9.5)	77-99.46	5	90.5 (11.1)	72-99.1	3	98.6 (0.5)	98-99
Gradient Boosting	8	79.2 (9.5)	67-79	7	76.7 (15.6)	45-92	10	86.8 (10.4)	70-97.3
Linear regressions	8	70.9 (4.9)	67-79	2	76.5 (0.7)	76-77	14	77.2 (5.3)	67-87
Logistic regressions	16	75.3 (7.3)	65.5-94.3	9	72.1 (16)	38.2-94.1	21	81.9 (9.6)	69.6-97
Naive Bayes	9	74 (6.6)	67.5-86.4	18	73.4 (13.2)	56-88.7	10	76.8 (7.1)	65.6-92
SVMs[c]	18	80 (8.4)	64-94.9	13	77.5 (14.1)	42.2 - 94	18	78.7 (6.9)	64.2-90.3
KNNs[d]	8	79.5 (13)	61.5 - 97	8	78.3 (13.9)	57 - 97	7	79.3 (9.5)	61.5-88.2
Neural networks	3	79.6 (14.4)	65-93.75	4	80.1 (15.1)	65.1-95.2	6	64.8 (24.8)	31.2-90.8
MLPs[e]	9	81.7 (8.3)	68-92	3	73.6 (24)	40.6-91.7	12	74.9 (17.3)	31.2-91.2
ANNs[f]	6	85.3 (10.8)	70.7-97.1	3	85.9 (12.3)	71.7-93	3	70.6 (6.3)	66-77.79
CNNs[g]	5	92 (8.1)	77.3-100	4	95.1 (3.9)	91.1-100	N/A[h]	N/A	N/A
Reinforcement learning	3	85.4 (4.1)	81-89.07	3	84.2 (4.7)	79.2-88.4	7	86.6 (3.8)	83-90.66
Ensemble models	9	93.4 (10.8)	65-99.84	9	92.5 (15.3)	52-99.21	7	89.4 (16)	56.5-98.95
Overall	174	81.7 (11.1)	59-100	141	80.5 (15.4)	38.2-100	176	81.0 (12)	31.2-100

[a]Models are grouped into tree-based, boosting, probabilistic, traditional machine learning, neural networks, and ensembles. Metrics are mean (SD). Study counts refer to the number of models reporting accuracy, F_1-score, or AUC—not the number of references.

[b]AUC:area under the curve.

[c]SVMs: Support Vector Machines.

[d]KNNs: K-Nearest Neighbor algorithm.

[e]MLPs: multilayer perceptrons.

[f]ANNs: Artificial Neural Network.

[g]CNNs: Convolutional Neural Networks.

[h]N/A: not applicable.

Discussion

Principal Findings

This scoping review examines the evolving application of AI in PPD research, with approximately 80% of studies prioritizing early prediction over detection. This reflects a growing awareness of AI's potential to enable proactive mental health interventions.

ML algorithms dominated (87.7%), suggesting a preference for structured data handling and model interpretability. Classical models such as RF (44.6%), LogR (35.4%), and XGBoost (23.1%) were especially prevalent, likely due to their ease of implementation, strong performance on tabular datasets, and alignment with the interpretability demands in health care as these models are well suited for structured and tabular clinical data (eg, demographics, EPDS scores, and EHR) and offer high interpretability, a core requirement in health care settings for clinical transparency and trust. In contrast, DL

XSL•FO
RenderX

approaches—while more capable of handling complex, high-dimensional inputs—were used in only 16.9% of studies, indicating underutilization of architectures such as convolutional or recurrent neural networks (RNNs). This aligns with the nature of DL methods (CNNs, RNNs, and transformers) that require large, high-dimensional datasets (eg, text, electroencephalogram, and sensor signals), which were rare in the reviewed studies; also, DL models are less interpretable, a major limitation in mental health where explainability is critical for clinician adoption. NLP and reinforcement learning (RL) were rarely used, despite their potential for analyzing unstructured clinical notes and dynamic decision-making, respectively. This aligns with the fact that NLP is suitable for analyzing unstructured clinical notes, patient narratives, or social media data, and few studies had access to such datasets. In addition, RL's strength in dynamic decision-making (eg, treatment adjustment over time) is difficult to apply in static, retrospective datasets common in PPD research.

More than 90% of studies focused on classification tasks—categorizing individuals as at risk or not for PPD—while only a few adopted regression models to estimate continuous risk levels. Although classification supports clinical decision-making, regression can offer more granular risk assessments, useful for personalized interventions and monitoring symptom trajectories. This aligns with real-world clinical workflows—binary screening tools are common. However, the underuse of regression models (predicting continuous risk) limits personalization and longitudinal risk tracking.

The optimization strategies such as Adam or stochastic gradient descent optimizers, learning rate scheduling, and dropout were rarely reported across the literature. This may reflect that most studies used classical ML, which does not require such parameters or limited technical expertise or a focus on classical methods over deep architectures. Regularization practices such as L1/L2 penalties, batch normalization, and early stopping were underused, despite their importance for model generalizability and performance stabilization. These techniques help prevent overfitting and improve generalizability. Their absence may reflect limited ML maturity or reliance on default model settings without tuning. Moreover, model validation techniques varied considerably. Although nearly half of the reviewed studies used k-fold cross-validation or holdout validation, external validation was seldom implemented. External validation requires access to independent datasets, which are often unavailable due to privacy constraints in mental health and raising concerns about generalizability. Performance evaluation also lacked consistency, with accuracy (75.4%) and sensitivity (73.9%) reported most frequently, while key metrics such as specificity, AUC, and F_1-score were less commonly disclosed, which are essential for imbalanced datasets such as PPD.

By considering the results of performance evaluation across accuracy, F_1-score, and AUC, they indicate that ensemble models, especially boosting techniques such as CatBoost, LightGBM, and XGBoost, consistently outperformed other AI methods in predicting PPD. Their high accuracy and AUC reflect strong generalization and robustness, owing to their ability to iteratively correct misclassifications and capture complex, nonlinear patterns—particularly valuable in noisy, imbalanced health care datasets.

CatBoost led with an AUC of 98.6%, benefiting from its advanced handling of categorical variables and built-in overfitting control, making it highly suited for structured health data. LightGBM followed closely, offering high accuracy (92.6%) and efficiency due to its gradient-based sampling and fast training, making it ideal for large-scale or real-time applications. XGBoost also performed competitively (89.1% accuracy and 86.8%) and remains popular for its transparency and feature importance tools.

In contrast, traditional decision tree–based classifiers such as RFs and generic tree–based models achieved moderate performance, with accuracy ranging from 80.5% to 82.8% and AUC values near 82%. While interpretable and computationally efficient, these models lacked the advanced learning mechanisms of boosting methods. Recursive partitioning methods, evaluated in only 2 studies, performed the weakest among tree-based approaches.

Overall, the mean model performance—accuracy: 81.7% (SD 11.05), F_1-score: 80.51% (SD 15.44), and AUC: 81.0% (SD 12.0)—shows variability likely due to differences in datasets, preprocessing, and validation strategies. This underscores the need for standardized evaluation and external validation to ensure model reproducibility and clinical reliability. Also, this suggests limited awareness or inconsistent standards in performance reporting and hinders meaningful comparisons and meta-analysis across studies.

This inconsistency complicates comparative assessments across models and highlights the need for standardized evaluation frameworks.

This scoping review offers the first comprehensive synthesis of both foundational and advanced preprocessing techniques used in AI-driven PPD studies. The high prevalence of basic normalization methods—such as Min-Max scaling and Z-score standardization, reported in 78.5% of studies—demonstrates a broad consensus on the need to standardize input features, particularly in ML models sensitive to feature magnitude. These practices are foundational for ensuring convergence stability and improving model performance, especially in algorithms such as LogR and k-nearest neighbors. However, more advanced preprocessing techniques were markedly underutilized, limiting the full potential of AI in PPD prediction.

For example, although missing data are ubiquitous in real-world health care datasets, only 44.6% of studies applied imputation techniques to address it. The remaining studies either dropped missing values or excluded incomplete cases—approaches that risk reducing sample size and introducing systematic bias, particularly in psychiatric populations where follow-up and self-report compliance can vary. Likewise, class imbalance, a well-documented issue in mental health data (eg, more controls than PPD cases), was insufficiently addressed: only 23.1% of studies used resampling methods such as SMOTE, and an even smaller fraction (6.2%) incorporated cost-sensitive learning,

XSL•FO
RenderX

which could improve model fairness and reduce false negatives—an important consideration in screening contexts.

In terms of categorical variable processing, label encoding (29.2%) and one-hot encoding (13.9%) were commonly used. While these methods are simple to implement, they may introduce ordinal bias or dimensionality explosion, respectively. More efficient encoding schemes (eg, target encoding and frequency encoding) that better preserve categorical relationships were rarely used, reflecting either limited awareness or concerns over interpretability.

Advanced feature engineering and selection techniques were also underexploited. Tree-based feature selection was used in only 18.5% of studies, despite its use in identifying nonlinear relationships and reducing overfitting. Dimensionality reduction methods such as PCA (6.2%) and interpretability tools such as SHAP (6.2%) were seldom implemented, limiting transparency and the ability to uncover key risk factors. Furthermore, multimodal or unstructured data processing techniques—including text or acoustic feature extraction—were applied in fewer than 5% of studies, despite their relevance in analyzing patient interviews, social media posts, or voice biomarkers.

In summary, while basic preprocessing steps have become standard practice, the limited adoption of more sophisticated strategies reflects a missed opportunity to enhance model robustness, generalizability, and interpretability. These gaps underscore the need for broader methodological literacy and the integration of more nuanced preprocessing pipelines tailored to the complexity and heterogeneity of PPD data.

Geographically, North America led the research landscape, with the United States contributing the largest share—just less than two-thirds. Asia also featured prominently, especially China, Bangladesh, and India. This wide participation demonstrates the global relevance and flexibility of AI solutions in diverse health care systems. However, it also reveals disparities in research capacity, underscoring the need for more contributions from underrepresented regions to ensure global equity. Bangladesh's notable presence is largely due to the use of public datasets, showing how open access data can significantly influence research output. The included studies span from 2009 to 2025, with publication volume rising steadily. A quarter of the studies were published in 2024 alone, reflecting growing global interest, better access to digital health data, and advances in AI. The lower count from 2025 is likely due to early-year data collection. Most studies appeared in peer-reviewed journals (more than two-thirds), while fewer were conference papers (less than one-third), and only 1 was a dissertation—illustrating both the academic rigor and fast-evolving nature of this field. Sample sizes varied widely, ranging from 11 to 5,73,634 participants (mean 18,187.4), yet nearly half of the studies had fewer than 5000 participants, raising concerns about generalizability and model performance. Among the 26 studies reporting participant age, the average was 31.08 years—consistent with the typical childbearing population—although the lack of demographic transparency in many studies limits comparability and clinical applicability. Among studies reporting participant age, the mean average was 31.08 (SD 3.42) years, aligning with the typical reproductive age. However, more than half of the studies did not report age, limiting comparability and model applicability across age groups.

Sample sizes varied greatly—from less than a dozen to more than half a million—with most studies enrolling fewer than 5000 participants. Smaller studies often used surveys or interviews, while larger ones relied on national registries. This underscores the importance of large datasets, especially for training DL models. Closed-source datasets dominated, with only about 25% of studies using open data. This limits reproducibility and hinders collaboration. Expanding open access datasets and standardized repositories would improve transparency and accelerate innovation. Most studies were retrospective, drawing on accessible surveys and EHR data. While more than two-thirds used surveys and many depended on structured clinical inputs, a recent shift toward prospective designs reflects growing interest in real-time, high-quality data for AI validation. Use of social media and sensor data is emerging, indicating a move toward passive, continuous monitoring. However, objective biomarkers—such as hormonal, genetic, or neuroimaging data—were underutilized, appearing in only 2 studies, underscoring a missed opportunity for clinical robustness. Moreover, nearly 90% of studies used unimodal inputs (eg, surveys and EHRs), with few incorporating multimodal data such as text, audio, or imaging. This limits the ability to capture the complex biopsychosocial nature of PPD. Research predominantly occurred in health care settings, reflecting strong clinical relevance. Around one-third took place in communities and fewer than 10% in academic contexts. Expanding into diverse settings could improve the inclusivity and generalizability of AI-based PPD interventions. Assessment timing for PPD varied widely, with 12-month evaluations being most common. Studies spanned short-, medium-, and long-term intervals, yet more than 40% failed to specify timing—revealing a major gap in methodological transparency. This variability underscores both evolving perspectives on PPD progression and the need for standardized follow-up periods to enhance comparability, reproducibility, and clinical relevance of AI models. The EPDS is the most widely used reference in PPD research, followed by *International Classification of Diseases* (*ICD*) codes and PHQ-9, highlighting its strong validation in postpartum populations. However, inconsistent use of diagnostic tools across studies hampers comparability and a unified understanding of PPD. While AI models show promise using varied data sources, few are benchmarked against tools such as EPDS or PHQ-9—limiting assessments of their real-world clinical use.

Feature counts varied widely (2-988), with an average of 44.9; nearly two-thirds of studies used fewer than 25 features, indicating a preference for simplicity and interpretability. These findings underscore the need for broader adoption of sophisticated feature selection and dimensionality reduction techniques (eg, PCA, SHAP, and recursive elimination) to enhance predictive performance and clinical relevance. The results of this scoping review underscore the central role of sociodemographic features, which were the most frequently used across included studies on PPD. Variables such as age

XSL•FO
RenderX

(56.9%), education level (32.3%), and marital status (30.8%) were among the most common predictors, highlighting the consistent reliance on structured patient-reported or administrative health records. Obstetric features were the second most common group, particularly mode of delivery (23.1%), parity (16.9%), and gestational age (13.9%). These findings align with literature suggesting that birth experience and maternal clinical history offer critical information in predicting PPD onset. Psychological indicators such as maternal anxiety (20%) and history of depression (18.5%) were also well represented. Their inclusion reflects growing interest in integrating mental health history and current affective symptoms into predictive frameworks. Similarly, behavioral factors such as breastfeeding status and bonding issues were frequently used to enhance emotional and functional contextualization of risk. Less frequently used were linguistic features (eg, LIWC metrics and emotional expression) and biomarkers (eg, epigenetic markers and neurological proteins), suggesting a growing but underutilized frontier. Notably, sensor-derived features (eg, tweet metadata, wearable-derived activity, or heart rate) appeared only in a handful of studies, despite the increasing ubiquity of digital health data. This spectrum of feature types illustrates a multidomain data integration trend, particularly among recent studies that incorporate EHRs, digital behavioral traces, and physiological data to enhance model robustness and precision.

Comparison With Previous Reviews

The findings of this review are broadly consistent with earlier literature, including reviews by Kwok et al [10], Fazraningtyas et al [15], Qi et al [17], Saqib et al [18], Fazraningtyas et al [30], and Acharya et al [13]. These prior studies similarly identified a reliance on retrospective study designs, structured demographic and clinical features (eg, age, parity, and psychiatric history), and traditional ML models such as RFs, support vector machines, and LogR. Most were applied to survey-based or EHR-derived datasets, reflecting the accessibility and interpretability of structured data in maternal mental health contexts.

However, this scoping review extends the current literature in several important ways. First, our review is not focused narrowly only on model types or performance metrics; it even systematically maps the entire AI modeling pipeline, from data characteristics and preprocessing techniques to model training, optimization, and validation strategies. For example, while earlier studies acknowledged preprocessing in general terms, our analysis quantifies the usage of basic methods (eg, scaling and label encoding) and highlights the underuse of advanced techniques such as SMOTE, SHAP, recursive feature elimination, and cost-sensitive learning.

Second, this review identifies the limited adoption of advanced AI methodologies, including DL, transformer-based NLP, and transfer learning, despite their growing success in related health care domains. While Fazraningtyas et al [15] and Qi et al [17] recognized these tools conceptually, few studies in their datasets applied them operationally to PPD detection tasks—an observation confirmed and quantified by our analysis.

Third, our review offers a granular classification of more than 45 features across 9 thematic domains, revealing persistent dependence on sociodemographic and self-reported data. This pattern, while accessible and interpretable, introduces potential biases and limits the generalizability of models. In contrast, passive and objective inputs—such as biosensors, electroencephalogram, speech signals, or real-time behavioral metrics—remain substantially underused, despite their promise for early and noninvasive detection of PPD.

Fourth, unlike previous studies that typically summarized trends descriptively, this review visualizes and quantitatively tracks the growth of literature over time, the global distribution of research output, and the evolution of study design types, using structured frameworks and stacked visualizations. For instance, we highlight that Bangladesh's growing presence is largely driven by the reuse of public datasets, illustrating how open data democratizes research participation.

Finally, this review distinguishes itself by its methodological scope and rigor. It includes studies published through February 2025 across 8 multidisciplinary databases, covering both prospective and retrospective designs, a wide range of countries, and diverse data sources. This comprehensive coverage enables a more nuanced understanding of current capabilities and persistent gaps in AI-based maternal mental health research. In particular, our audit of model regularization, hyperparameter tuning, and evaluation practices offers insight into areas often overlooked by earlier reviews. Taken together, these contributions provide a stronger foundation for the development of transparent, reproducible, and clinically relevant AI tools in PPD research—addressing both methodological blind spots and equity concerns raised in prior literature.

Implication and Further Works

To significantly advance the field of AI-driven PPD research, several key strategic priorities should be addressed. These priorities focus on improving methodological rigor, inclusivity of data, clinical applicability, and ethical implementation.

First, enhance the integration of multimodal and objective data sources. Currently, research predominantly relies on traditional sociodemographic and self-reported survey data. There is an urgent need to incorporate underutilized modalities such as linguistic data (eg, LIWC, sentiment analysis, and acoustic speech patterns), biosignals (eg, heart rate variability and activity monitoring), wearable technology outputs, and biological biomarkers (eg, hormonal, metabolic, genetic, or epigenetic markers). Leveraging these richer data types can significantly enhance the accuracy, personalization, and early detection capabilities of predictive models.

Second, expand and prioritize sharing of open access datasets. Only about 25% of studies included in our review used publicly available datasets, highlighting a substantial barrier to reproducibility, benchmarking, and international collaboration. Developing standardized, large-scale, anonymized datasets should become a priority. Techniques such as federated learning could facilitate collaborative research across different institutions while maintaining data privacy and security.

XSL•FO
RenderX

Third, increase the adoption of longitudinal and prospective research designs. Most existing AI models for PPD prediction are based on retrospective or immediate postpartum data, limiting their ability to capture evolving symptom patterns over time. Incorporating longitudinal data collection into future studies is essential to better understand symptom progression, delayed onset, and relapse scenarios, thus enhancing the clinical relevance and predictive accuracy of AI models.

Fourth, advance multimodal fusion frameworks. Given the complexity of PPD, future models must systematically integrate structured data (eg, EHRs) and unstructured inputs (eg, text, audio, and sensor signals). Developing robust multimodal fusion approaches that effectively combine diverse data sources will significantly enhance model interpretability, clinical effectiveness, and predictive power.

Fifth, standardize preprocessing and feature engineering pipelines. Variability and incomplete reporting in preprocessing methods currently limit model comparability and reproducibility. Adopting standardized protocols for data preprocessing—including feature extraction, transformation techniques, and class imbalance adjustments (eg, SMOTE and cost-sensitive learning)—is necessary. Transparent reporting of these processes should be enforced to enhance scientific rigor and validation.

Sixth, emphasize model explainability and ethical AI practices. Transparent and interpretable AI models are crucial for clinical adoption. Few studies currently apply advanced explainability methods such as SHAP, Local Interpretable Model-Agnostic Explanations, or counterfactual analyses. Integrating these interpretability techniques into AI pipelines will facilitate clinician trust and understanding. Moreover, ethical considerations—such as minimizing algorithmic bias (such as balanced datasets and resampling to correct imbalances in training data)—are considered in few studies, but preventing potential harms from false positives, safeguarding patient autonomy, and ensuring cultural sensitivity should be systematically addressed in all AI model developments.

Seventh, standardize evaluation and reporting metrics. While accuracy is often prioritized, metrics such as specificity, AUC, F_1-score, and precision must be consistently reported to enable comprehensive evaluation and meaningful comparisons across studies. Furthermore, systematic reviews and meta-analyses are required to identify existing methodological inconsistencies, biases, and underrepresented findings to refine future AI-based research approaches.

Eighth, shift from subjective screening tools to objective validation measures. Current studies heavily rely on subjective instruments (eg, EPDS, PHQ, and *ICD* codes), which, despite validation, vary significantly across studies. Future research should validate AI models using objective clinical measures such as physiological indicators and behavioral markers, thus improving reliability and facilitating clinical implementation.

Ninth, the superior performance of ensemble models—particularly boosting techniques such as CatBoost, LightGBM, and XGBoost—suggests that they are promising candidates for clinical implementation in PPD screening. Their ability to consistently achieve high accuracy, F_1-score, and AUC underscores their robustness in handling structured health data, including demographic and clinical features. Given the strong results of CatBoost in handling categorical variables and LightGBM's efficiency in large-scale settings, future research should prioritize evaluating these models in real-world clinical workflows and mobile health platforms, where scalability and interpretability are critical. In addition, since performance varied across studies due to differences in data characteristics and preprocessing strategies, future work should aim to establish benchmark datasets and standardized pipelines for fair comparison.

Efforts should also be made to assess model performance across different population subgroups, ensuring that these algorithms do not inadvertently introduce or amplify bias. Finally, comparative studies should continue to assess whether boosting models maintain their advantage as datasets grow and diversify, particularly in longitudinal or multisite contexts.

Finally, foster global and interdisciplinary collaboration. PPD research remains unevenly distributed globally. Encouraging cross-regional and interdisciplinary collaboration—particularly with underrepresented regions and diverse professional backgrounds such as computer science, psychiatry, public health, and ethics—will foster equitable research practices and drive innovation in maternal mental health care globally.

Limitations

Despite the comprehensive nature of this scoping review, several limitations should be acknowledged; first, we limited our inclusion to studies published in English. This language restriction may have resulted in the exclusion of relevant research published in other languages, particularly from non–English-speaking countries where maternal mental health may be a pressing issue.

Second, we prioritized peer-reviewed and indexed literature from 8 major databases and limited Google Scholar results to the first 100 entries ranked by relevance. Consequently, gray literature, including government reports, dissertations beyond ProQuest, and nonindexed conference proceedings, may have been underrepresented.

Third, our review focused specifically on studies using AI techniques for detection or prediction of PPD. As a result, studies that applied AI for monitoring, treatment delivery, or resource allocation in maternal mental health were excluded, which narrows the scope of applicability.

Finally, we did not conduct a quantitative meta-analysis or risk of bias assessment, as these are typically outside the scope of scoping reviews. Consequently, while we mapped methodological patterns and gaps, we did not evaluate effect sizes, statistical heterogeneity, or study-level quality in a standardized manner.

Conclusions

This scoping review comprehensively maps the application of AI in PPD research, analyzing 87 studies published between 2009 and 2025. The review identifies a predominant emphasis on early prediction (~80%) over detection, with ML

XSL•FO
RenderX

methods—particularly RF (44.6%), LogR (35.4%), and XGBoost (23.1%)—used in 87.7% of studies. These models were favored for their compatibility with structured clinical data and interpretability. DL approaches, including CNNs and RNNs, were underutilized (16.9%), reflecting data limitations and interpretability concerns. NLP and RL were rarely applied, mirroring limited access to unstructured or sequential data sources.

More than 90% of studies focused on classification tasks, aligning with standard clinical workflows, while regression approaches remained limited. Basic preprocessing practices, such as normalization, were widely adopted (78.5%), but advanced strategies—such as imputation (44.6%), resampling (23.1%), cost-sensitive learning (6.2%), and feature selection techniques such as PCA or SHAP—were inconsistently applied. Most models lacked detailed reporting of optimization strategies or regularization methods, and only half used internal validation. External validation was rarely reported, complicating model comparability.

The comparative analysis between accuracy, F_1-score, and AUC confirms that ensemble learning approaches, particularly boosting algorithms such as CatBoost and LightGBM, consistently outperform traditional models in predicting PPD, achieving superior accuracy, F_1-scores, and AUC values across studies.

Geographic trends showed research dominance by North America, particularly the United States, with notable contributions from Asia, driven by access to public datasets. Most studies used retrospective designs and unimodal inputs—mainly survey or EHR data—while multimodal and objective data (eg, biomarkers and sensor data) were rarely incorporated. Assessment timing, feature selection, and dataset transparency varied widely. Sociodemographic and obstetric features were the most frequently used predictors, while linguistic, behavioral, and physiological data were underrepresented. This review offers the first detailed synthesis of preprocessing workflows and feature domains in PPD-AI research, underscoring both progress and methodological gaps across the literature.

Acknowledgments

The authors declare the use of generative artificial intelligence in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing. The GAI tool used was ChatGPT-4. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by all authors. All intellectual content, study design, data collection, analysis, and final interpretations are the sole responsibility of the authors. Multimedia Appendix 9 shows prompts and responses used. No custom code or mathematical algorithm was used in this study.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

AN developed the protocol with guidance from and under the supervision of AAA. AAA searched the electronic databases and conducted backward and forward reference list checking. The study selection, data extraction, and data synthesis process were carried out by MA and AN under the supervision of AAA. MA wrote results, methods, discussion, and conclusion sections. M Alsahli wrote the background section. The paper was revised critically for important intellectual content by all authors under the supervision of AAA. All authors approved the manuscript for publication and agreed to be accountable for all aspects of the work.

Conflicts of Interest

AA-A is an associate editor of *JMIR Nursing* at the time of publication. All other authors declare no conflict of interest.

Multimedia Appendix 1
Search strategy.
[DOCX File, 26 KB - jmir_v28i1e77376_app1.docx]

Multimedia Appendix 2
Data extraction form.
[DOCX File, 24 KB - jmir_v28i1e77376_app2.docx]

Multimedia Appendix 3

XSL•FO
RenderX

Characteristics of each included study.
[DOCX File, 38 KB - jmir_v28i1e77376_app3.docx]

Multimedia Appendix 4
Characteristics of the dataset.
[DOCX File, 78 KB - jmir_v28i1e77376_app4.docx]

Multimedia Appendix 5
Characteristics of preprocessing.
[DOCX File, 33 KB - jmir_v28i1e77376_app5.docx]

Multimedia Appendix 6
Feature characteristics.
[DOCX File, 34 KB - jmir_v28i1e77376_app6.docx]

Multimedia Appendix 7
Characteristics of artificial intelligence.
[DOCX File, 40 KB - jmir_v28i1e77376_app7.docx]

Multimedia Appendix 8
Characteristics of machine learning performances.
[DOCX File, 101 KB - jmir_v28i1e77376_app8.docx]

Multimedia Appendix 9
Prompts and responses.
[DOCX File, 40 KB - jmir_v28i1e77376_app9.docx]

Checklist 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist.
[DOCX File, 110 KB - jmir_v28i1e77376_app10.docx]

References

1. Saharoy R, Potdukhe A, Wanjari M, Taksande AB. Postpartum depression and maternal care: exploring the complex effects on mothers and infants. Cureus 2023 Jul;15(7):e41381. [doi: 10.7759/cureus.41381] [Medline: 37546054]
2. Lopez-Gonzalez DM, Kopparapu AK. Postpartum care of the new mother. In: StatPearls: StatPearls Publishing; 2025. [Medline: 33351433]
3. Howard LM, Khalifeh H. Perinatal mental health: a review of progress and challenges. World Psychiatry 2020 Oct;19(3):313-327. [doi: 10.1002/wps.20769] [Medline: 32931106]
4. Roddy Mitchell A, Gordon H, Lindquist A, et al. Prevalence of perinatal depression in low- and middle-income countries. JAMA Psychiatry 2023 May 1;80(5):425. [doi: 10.1001/jamapsychiatry.2023.0069]
5. Levis B, Negeri Z, Sun Y, Benedetti A, Thombs BD, DEPRESsion Screening Data (DEPRESSD) EPDS Group. Accuracy of the Edinburgh Postnatal Depression Scale (EPDS) for screening to detect major depression among pregnant and postpartum women: systematic review and meta-analysis of individual participant data. BMJ 2020 Nov 11;371:m4022. [doi: 10.1136/bmj.m4022] [Medline: 33177069]
6. Levis B, Benedetti A, Thombs BD, DEPRESsion Screening Data (DEPRESSD) Collaboration. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. BMJ 2019 Apr 9;365:l1476. [doi: 10.1136/bmj.l1476] [Medline: 30967483]
7. Long MM, Cramer RJ, Bennington L, et al. Perinatal depression screening rates, correlates, and treatment recommendations in an obstetric population. Fam Syst Health 2020 Dec;38(4):369-379. [doi: 10.1037/fsh0000531] [Medline: 33119369]
8. Sidebottom A, Vacquier M, LaRusso E, Erickson D, Hardeman R. Perinatal depression screening practices in a large health system: identifying current state and assessing opportunities to provide more equitable care. Arch Womens Ment Health 2021 Feb;24(1):133-144. [doi: 10.1007/s00737-020-01035-x] [Medline: 32372299]
9. Cox J. Use and misuse of the Edinburgh Postnatal Depression Scale (EPDS): a ten point “survival analysis”. Arch Womens Ment Health 2017 Dec;20(6):789-790. [doi: 10.1007/s00737-017-0789-7] [Medline: 29101480]
10. Kwok WH, Zhang Y, Wang G. Artificial intelligence in perinatal mental health research: a scoping review. Comput Biol Med 2024 Jul;177:108685. [doi: 10.1016/j.compbiomed.2024.108685] [Medline: 38838557]

XSL•FO
RenderX

11. Mapari SA, Shrivastava D, Dave A, et al. Revolutionizing maternal health: the role of artificial intelligence in enhancing care and accessibility. Cureus 2024;16(9):e69555. [doi: 10.7759/cureus.69555]
12. Turchioe MR, Hermann A, Benda NC. Recentering responsible and explainable artificial intelligence research on patients: implications in perinatal psychiatry. Front Psychiatry 2023;14:1321265. [doi: 10.3389/fpsyt.2023.1321265] [Medline: 38304402]
13. Acharya A, Ramesh R, Fathima T, Lakhani T, S SK. Clinical tools to detect postpartum depression based on machine learning and EEG: a review. Presented at: 2023 2nd International Conference on Computational Systems and Communication (ICCSC); Mar 3-4, 2023. [doi: 10.1109/ICCSC56913.2023.10142970]
14. Cellini P, Pigoni A, Delvecchio G, Moltrasio C, Brambilla P. Machine learning in the prediction of postpartum depression: a review. J Affect Disord 2022 Jul;309:350-357. [doi: 10.1016/j.jad.2022.04.093]
15. Fazraningtyas WA, Rahmatullah B, Dwi Salmarini D, Arrieya Ariffin S, Ismail A. Recent advancements in postpartum depression prediction through machine learning approaches: a systematic review. Bull EEI 2024;13(4):2729-2737. [doi: 10.11591/eei.v13i4.7185]
16. Kimwomi G, Mgala M, Mwakondo F, Kimeto P. Machine learning prediction models for postpartum depression, a review of literature. IJCATR 2022;11(06):205-212. [doi: 10.7753/IJCATR1106.1005]
17. Qi W, Wang Y, Li C, et al. Predictive models for predicting the risk of maternal postpartum depression: a systematic review and evaluation. J Affect Disord 2023 Jul 15;333:107-120. [doi: 10.1016/j.jad.2023.04.026] [Medline: 37084958]
18. Saqib K, Khan AF, Butt ZA. Machine learning methods for predicting postpartum depression: scoping review. JMIR Ment Health 2021 Nov 24;8(11):e29838. [doi: 10.2196/29838] [Medline: 34822337]
19. Zhong M, Zhang H, Yu C, Jiang J, Duan X. Application of machine learning in predicting the risk of postpartum depression: a systematic review. J Affect Disord 2022 Dec 1;318:364-379. [doi: 10.1016/j.jad.2022.08.070] [Medline: 36055532]
20. Cohen JM. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr;20(1):37-46. [doi: 10.1177/001316446002000104]
21. Sharma Y, Jain V, Tarwani S. Ensemble machine learning model for predicting postpartum depression disorder. 2024 Presented at: 2024 IEEE Region 10 Symposium (TENSYMP); Mar 3-4, 2023; New Delhi, India p. 1-6. [doi: 10.1109/TENSYMP61132.2024.10752305]
22. Andersson S, Bathula DR, Iliadis SI, Walter M, Skalkidou A. Predicting women with depressive symptoms postpartum with machine learning methods. Sci Rep 2021 Apr 12;11(1):7877. [doi: 10.1038/s41598-021-86368-y] [Medline: 33846362]
23. Betts KS, Kisely S, Alati R. Predicting postpartum psychiatric admission using a machine learning approach. J Psychiatr Res 2020 Nov;130:35-40. [doi: 10.1016/j.jpsychires.2020.07.002] [Medline: 32771679]
24. Sona Ajay C, Juliet S, Alex A. Machine learning and survival analysis models for postpartum depression: a comprehensive risk factor analysis. 2024 Presented at: 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS); Mar 14-15, 2024; Coimbatore, India p. 1340-1345. [doi: 10.1109/ICACCS60874.2024.10717207]
25. Cai M, Wang Y, Luo Q, Wei G. Factor analysis of the prediction of the Postpartum Depression Screening Scale. Int J Environ Res Public Health 2019 Dec 10;16(24):5025. [doi: 10.3390/ijerph16245025] [Medline: 31835547]
26. Carneiro MB, Moreira MWL, Pereira SSL, Gallindo EL, Rodrigues J. Recommender system for postpartum depression monitoring based on sentiment analysis. Presented at: 2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM); Mar 1-2, 2020; Shenzhen, China p. 1-6. [doi: 10.1109/HEALTHCOM49281.2021.9398922]
27. Chen Y, Zhou B, Zhang W, Gong W, Sun G. Sentiment analysis based on deep learning and its application in screening for perinatal depression. Presented at: 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC); Jun 18-21, 2018; Guangzhou, China p. 451-456. [doi: 10.1109/DSC.2018.00073]
28. Fanos V, Dessì A, Deledda L, et al. Postpartum depression screening through artificial intelligence: preliminary data through the Talking About algorithm. J Pediatr Neonatal Individualized Med 2023;12(2):e120222. [doi: 10.7363/120222]
29. Fatima I, Abbasi BUD, Khan S, Al‐Saeed M, Ahmad HF, Mumtaz R. Prediction of postpartum depression using machine learning techniques from social media text. Expert Syst 2019 Aug;36(4). [doi: 10.1111/exsy.12409]
30. Fazraningtyas WA, Rahmatullah B, Naparin H, Basit M, Razak NA. A predictive model for postpartum depression: ensemble learning strategies in machine learning. Indones J Electrical Eng Comput Sci 2025;37(1):443. [doi: 10.11591/ijeecs.v37.i1.pp443-451]
31. Gabrieli G, Bornstein MH, Manian N, Esposito G. Assessing mothers' postpartum depression from their infants' cry vocalizations. Behav Sci (Basel) 2020;10(2):55. [doi: 10.3390/bs10020055]
32. Gopalakrishnan A, et al. A combined attribute extraction method for detecting postpartum depression using social media. Presented at: Health Information Science: 12th International Conference, HIS 2023; Oct 23-24, 2023; Melbourne, VIC, Australia p. 17-29. [doi: 10.1007/978-981-99-7108-4_2]
33. Gopalakrishnan A, Gururajan R, Venkataraman R, et al. Attribute Selection Hybrid Network Model for risk factors analysis of postpartum depression using social media. Brain Inform 2023 Oct 31;10(1):28. [doi: 10.1186/s40708-023-00206-7] [Medline: 37906324]
34. Gopalakrishnan A, Venkataraman R, Gururajan R, Zhou X, Zhu G. Predicting women with postpartum depression symptoms using machine learning techniques. Mathematics 2022;10(23):4570. [doi: 10.3390/math10234570]

XSL•FO
RenderX

35. Gupta V, Tripathi S, Singh D, Bansal A. Predictive algorithms for early postpartum depression detection: CatBoost vs. LightGBM. Presented at: 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO); Oct 14-15, 2024; Noida, India p. 1-4. [doi: 10.1109/ICRITO61523.2024.10522300]
36. Horgen ML. A machine learning approach to understanding depression and anxiety in new mothers. Predicting symptom levels using population-based registry data from a large Norwegian prospective study [Master's Thesis]. : Department of Physics, Faculty of Mathematics and Natural Sciences, University of Oslo; 2020 URL: https://github.com/marialinea/predicting-depression-and-anxiety-moba
37. Hurwitz E, Butzin-Dozier Z, Master H, et al. Harnessing consumer wearable digital biomarkers for individualized recognition of postpartum depression using the All of Us Research Program data set: cross-sectional study. JMIR Mhealth Uhealth 2024 May 2;12:e54622. [doi: 10.2196/54622] [Medline: 38696234]
38. Jiménez-Serrano S, Tortajada S, García-Gómez JM. A mobile health application to predict postpartum depression based on machine learning. Telemed J E Health 2015 Jul;21(7):567-574. [doi: 10.1089/tmj.2014.0113] [Medline: 25734829]
39. Krishnamurti T, Allen K, Hayani L, Rodriguez S, Davis AL. Identification of maternal depression risk from natural language collected in a mobile health app. Procedia Comput Sci 2022;206:132-140. [doi: 10.1016/j.procs.2022.09.092] [Medline: 36712815]
40. Lilhore UK, Dalal S, Faujdar N, et al. Unveiling the prevalence and risk factors of early stage postpartum depression: a hybrid deep learning approach. Multimed Tools Appl 2024;83(26):68281-68315. [doi: 10.1007/s11042-024-18182-3]
41. Lilhore UK, Dalal S, Varshney N, et al. Prevalence and risk factors analysis of postpartum depression at early stage using hybrid deep learning model. Sci Rep 2024 Feb 24;14(1):4533. [doi: 10.1038/s41598-024-54927-8] [Medline: 38402249]
42. Liu H, Dai A, Zhou Z, et al. An optimization for postpartum depression risk assessment and preventive intervention strategy based machine learning approaches. J Affect Disord 2023 May;328:163-174. [doi: 10.1016/j.jad.2023.02.028]
43. Liu Y, Joly R, Reading Turchioe M, et al. Preparing for the bedside-optimizing a postpartum depression risk prediction model for clinical implementation in a health system. J Am Med Inform Assoc 2024 May 20;31(6):1258-1267. [doi: 10.1093/jamia/ocae056] [Medline: 38531676]
44. Lyall LM, Sangha N, Zhu X, et al. Subjective and objective sleep and circadian parameters as predictors of depression-related outcomes: a machine learning approach in UK Biobank. J Affect Disord 2023 Aug 15;335:83-94. [doi: 10.1016/j.jad.2023.04.138] [Medline: 37156273]
45. Marshad I, Islam M, Samin AM, Shomyo M, Nishat MM, Faisal F. Optimizing maternal mental health: a study on boosting algorithms for suicidal tendencies prediction in postpartum depression. Presented at: 2024 International Conference on Inventive Computation Technologies (ICICT); Apr 24-26, 2024; Lalitpur, Nepal p. 1077-1081. [doi: 10.1109/ICICT60155.2024.10544607]
46. Matsumura K, Hamazaki K, Kasamatsu H, Tsuchida A, Inadera H. Decision tree learning for predicting chronic postpartum depression in the Japan Environment and Children's Study. J Affect Disord 2025 Jan 15;369:643-652. [doi: 10.1016/j.jad.2024.10.034] [Medline: 39389121]
47. Matsuo S, Ushida T, Emoto R, et al. Machine learning prediction models for postpartum depression: a multicenter study in Japan. J Obstet Gynaecol Res 2022 Jul;48(7):1775-1785. [doi: 10.1111/jog.15266] [Medline: 35438215]
48. Mazumder P, Baruah S. A community based study for early detection of postpartum depression using improved data mining techniques. 2021 Presented at: 2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS); Dec 16-18, 2021; Bangalore, India p. 1-7. [doi: 10.1109/CSITSS54238.2021.9682941]
49. Moreira MWL, Rodrigues J, Kumar N, Saleem K, Illin IV. Postpartum depression prediction through pregnancy data analysis for emotion-aware smart systems. Information Fusion 2019 May;47:23-31. [doi: 10.1016/j.inffus.2018.07.001]
50. Mustafa N. Use of m-health application to figure out post-natal depression, an evidence-based study. JAMMR 2023;35(24):81-90. [doi: 10.9734/jammr/2023/v35i245326]
51. Myneni S, Zingg A, Singh T, et al. Digital health technologies for high-risk pregnancy management: three case studies using Digilego framework. JAMIA Open 2024 Apr;7(1):ooae022. [doi: 10.1093/jamiaopen/ooae022] [Medline: 38455839]
52. Nasim S, Sami Al-Shamayleh A, Thalji N, et al. Novel meta learning approach for detecting postpartum depression disorder using questionnaire data. IEEE Access 2024;12:101247-101259. [doi: 10.1109/ACCESS.2024.3427685]
53. Natarajan S, Prabhakar A, Ramanan N, Bagilone A, Siek K, Connelly K. Boosting for postpartum depression prediction. 2017 Presented at: 2017 IEEE/ACM International Conference on Connected Health; Jul 17-19, 2017; Philadelphia, PA p. 232-240. [doi: 10.1109/CHASE.2017.82]
54. Osubor VI, Egwali AO. A neuro fuzzy approach for the diagnosis of postpartum depression disorder. Iran J Comput Sci 2018 Dec;1(4):217-225. [doi: 10.1007/s42044-018-0021-6]
55. Park Y, Hu J, Singh M, et al. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. JAMA Netw Open 2021 Apr 1;4(4):e213909. [doi: 10.1001/jamanetworkopen.2021.3909]
56. Paul A, Pragada SD, Murthy DN, Shruthi MLJ, Gurugopinath S. Performance comparison of machine learning techniques for early detection of postpartum depression using PRAMS dataset. 2023 Presented at: 2023 IEEE 15th International Conference on Computational Intelligence and Communication Networks (CICN); Dec 22-23, 2023; Bangkok, Thailand p. 310-315. [doi: 10.1109/CICN59264.2023.10402321]

XSL•FO
RenderX

57. Payne JL, Osborne LM, Cox O, et al. DNA methylation biomarkers prospectively predict both antenatal and postpartum depression. Psychiatry Res 2020 Mar;285:112711. [doi: 10.1016/j.psychres.2019.112711] [Medline: 31843207]
58. Prabhashwaree T, Wagarachchi NM. Predicting mothers with postpartum depression using machine learning approaches. 2022 Presented at: 2022 International Research Conference on Smart Computing and Systems Engineering (SCSE); Sep 1, 2022; Colombo, Sri Lanka p. 28-34. [doi: 10.1109/SCSE56529.2022.9905186]
59. Prabhashwaree T, Wagarachchi NM. Towards machine learning approaches for predicting risk level of postpartum depression. 2022 Presented at: 2022 6th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI); Dec 1-2, 2022; Colombo, Sri Lanka p. 1-6. [doi: 10.1109/SLAAI-ICAI56923.2022.10002477]
60. Qasrawi R, Amro M, VicunaPolo S, et al. Machine learning techniques for predicting depression and anxiety in pregnant and postpartum women during the COVID-19 pandemic: a cross-sectional regional study. F1000Res 2022;11:390. [doi: 10.12688/f1000research.110090.1] [Medline: 36111217]
61. Raisa JF, Kaiser MS, Mahmud M. A machine learning approach for early detection of postpartum depression in bangladesh. 2022 Presented at: Brain Informatics: 15th International Conference, BI 2022; Jul 15-17, 2022; Padua, Italy p. 241-252. [doi: 10.1007/978-3-031-15037-1_20]
62. Reps JM, Wilcox M, McGee BA, Leonte M, LaCross L, Wildenhaus K. Development of multivariable models to predict perinatal depression before and after delivery using patient reported survey responses at weeks 4-10 of pregnancy. BMC Pregnancy Childbirth 2022 May 26;22(1):442. [doi: 10.1186/s12884-022-04741-9] [Medline: 35619056]
63. Shen S, Qi S, Luo H. Automatic model for postpartum depression identification using deep reinforcement learning and differential evolution algorithm. IJACSA 2023;14(11):154-166. [doi: 10.14569/IJACSA.2023.0141115]
64. Shin D, Lee KJ, Adeluwa T, Hur J. Machine learning-based predictive modeling of postpartum depression. J Clin Med 2020 Sep 8;9(9):2899. [doi: 10.3390/jcm9092899] [Medline: 32911726]
65. Shivaprasad S, Chadaga K, Sampathila N, Prabhu S, Chadaga P R, K S S. Explainable machine learning methods to predict postpartum depression risk. Systems Science & Control Engineering 2024 Dec 31;12(1). [doi: 10.1080/21642583.2024.2427033]
66. Srivatsav P, Nanthini S. Detecting early markers of post-partum depression in new mothers: an efficient LSTM-CNN approach compared to logistic regression. 2024 Presented at: 2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT); Mar 15-16, 2024; Kottayam, India p. 1-6. [doi: 10.1109/ICITIIT61487.2024.10580321]
67. Suganthi D, Geetha A. Predicting postpartum depression with aid of social media texts using optimized machine learning model. IJIES 2024;17(3):417-427. [doi: 10.22266/ijies2024.0630.33]
68. Susič D, Bombač Tavčar L, Lučovnik M, Hrobat H, Gornik L, Gradišek A. Wellbeing forecasting in postpartum anemia patients. Healthcare (Basel) 2023 Jun 9;11(12):12. [doi: 10.3390/healthcare11121694] [Medline: 37372812]
69. Tang Y, Huang T, Yin X. Postpartum depression identification: integrating mutual learning-based artificial bee colony and proximal policy optimization for enhanced diagnostic precision. IJACSA 2024;15(6):332-347. [doi: 10.14569/IJACSA.2024.0150636]
70. Tortajada S, García-Gomez JM, Vicente J, et al. Prediction of postpartum depression using multilayer perceptrons and pruning. Methods Inf Med 2009;48(3):291-298. [doi: 10.3414/ME0562] [Medline: 19387507]
71. Valavani E, Doudesis D, Kourtesis I, et al. Data-driven insights towards risk assessment of postpartum depression. Presented at: Special Session on Mining Self-reported Outcome Measures, Clinical Assessments, and Non-invasive Sensor Data Towards Facilitating Diagnosis, Longitudinal Monitoring, and Treatment; Feb 24-26, 2020. [doi: 10.5220/0009369303820389]
72. Valdeolivar-Hernandez LI, Flores Quijano ME, Echeverria-Arjonilla JC, Perez-Gonzalez J, Piña-Ramírez O. Towards breastfeeding self-efficacy and postpartum depression estimation based on analysis of free-speech interviews through natural language processing. Presented at: 18th International Symposium on Medical Information Processing and Analysis (SIPAIM 2022); Nov 9-11, 2022. [doi: 10.1117/12.2669883]
73. Wagay FA. Comparing ensemble techniques for postpartum depression detection: a comprehensive analysis. Presented at: 2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI); Dec 29-30, 2023. [doi: 10.1109/ICAIIHI57871.2023.10489773]
74. Wakefield C, Frasch MG. Predicting patients requiring treatment for depression in the postpartum period using common electronic medical record data available antepartum. AJPM Focus 2023 Sep;2(3):100100. [doi: 10.1016/j.focus.2023.100100]
75. Wang J, Sui X, Hu B, Flint J, et al. Detecting postpartum depression in depressed people by speech features. In: Zu Q, Hu B, editors. Human Centered Computing: Springer, Cham; 2017, Vol. 10745. [doi: 10.1007/978-3-319-74521-3_46]
76. Wang S, Pathak J, Zhang Y. Using electronic health records and machine learning to predict postpartum depression. Stud Health Technol Inform 2019 Aug 21;264:888-892. [doi: 10.3233/SHTI190351] [Medline: 31438052]
77. Wang S, Xu R, Li G, Liu S, Zhu J, Gao P. A plasma proteomics-based model for identifying the risk of postpartum depression using machine learning. J Proteome Res 2025 Feb 7;24(2):824-833. [doi: 10.1021/acs.jproteome.4c00826] [Medline: 39772732]
78. Wang Y, Yan P, Wang G, et al. Trajectory on postpartum depression of Chinese women and the risk prediction models: a machine-learning based three-wave follow-up research. J Affect Disord 2024 Nov;365:185-192. [doi: 10.1016/j.jad.2024.08.074]

XSL•FO
RenderX

79. Xu J, Yu H, Lv H, et al. Consistent functional abnormalities in patients with postpartum depression. Behav Brain Res 2023 Jul 26;450:114467. [doi: 10.1016/j.bbr.2023.114467] [Medline: 37146719]
80. Xu W, Sampson M. Prenatal and childbirth risk factors of postpartum pain and depression: a machine learning approach. Matern Child Health J 2023 Feb;27(2):286-296. [doi: 10.1007/s10995-022-03532-0] [Medline: 36526882]
81. Yu Z, Matsukawa N, Saigusa D, et al. Plasma metabolic disturbances during pregnancy and postpartum in women with depression. iScience 2022 Dec 22;25(12):105666. [doi: 10.1016/j.isci.2022.105666] [Medline: 36505921]
82. Zhang W, Liu H, Silenzio VMB, Qiu P, Gong W. Machine learning models for the prediction of postpartum depression: application and comparison based on a cohort study. JMIR Med Inform 2020 Apr 30;8(4):e15516. [doi: 10.2196/15516] [Medline: 32352387]
83. Zhang Y, Joly R, Beecy AN, et al. Implementation of a machine learning risk prediction model for postpartum depression in the electronic health records. AMIA Jt Summits Transl Sci Proc 2024;2024:1057-1066. [Medline: 39444417]
84. Zhang Y, Wang S, Hermann A, Joly R, Pathak J. Development and validation of a machine learning algorithm for predicting the risk of postpartum depression among pregnant women. J Affect Disord 2021 Jan;279:1-8. [doi: 10.1016/j.jad.2020.09.113]
85. Zhu J, Ye Y, Liu X, et al. The incidence and risk factors of depression across six time points in the perinatal period: a prospective study in China. Front Med 2024;11:1407034. [doi: 10.3389/fmed.2024.1407034]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
CNN: convolutional neural network
DL: deep learning
EHR: electronic health record
EPDS: Edinburgh Postnatal Depression Scale
***ICD*:** *International Classification of Diseases*
LIWC: linguistic inquiry and word count
LogR: logistic regression
ML: machine learning
PCA: principal component analysis
PHQ-9: Patient Health Questionnaire-9
PPD: postpartum depression
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
RF: random forest
RL: reinforcement learning
RNN: recurrent neural network

Edited by A Coristine, T Leung; submitted 12.May.2025; peer-reviewed by S Tedesco, YA Moaiad; revised version received 24.Jul.2025; accepted 31.Aug.2025; published 08.Jan.2026.

Please cite as:
Alkhateeb M, Nayeem A, Ahmed A, Alsahli M, Sheikh J, Abd-Alrazaq A
AI for Detecting and Predicting Postpartum Depression: Scoping Review
J Med Internet Res 2026;28:e77376
URL: https://www.jmir.org/2026/1/e77376
doi:10.2196/77376

© Mais Alkhateeb, Ajisha Nayeem, Arfan Ahmed, Mohammed Alsahli, Javaid Sheikh, Alaa Abd-Alrazaq. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Face-to-Face Versus Digital, Telephone-Delivered, and Self-Help Cognitive Behavioral Therapy for Irritable Bowel Syndrome: Systematic Review and Bayesian Indirect Treatment Comparison Meta-Analysis

Qing-Feng Tao[1*], MD; Can Hua[2*], MD; Xiao Zhuo[3], MD; Jian-Jiao Mou[1], MD; Chao-Rong Xie[1], MD; Yu-Xin Zhang[1], MD; Bei Lv[1], MD; Xin-Ying Niu[1], MD; Min Chen[3], MD, PhD; Hui Zheng[1], MD, PhD

[1]Acupuncture and Tuina School, Chengdu University of Traditional Chinese Medicine, No.1166 Liutai Avenue, Wenjiang District, Chengdu, China

[2]Department of Traditional Chinese Medicine, Dazhou Dachuan District People's Hospital (Dazhou Third People's Hospital), Dazhou, China

[3]Department of Colorectal Diseases, Hospital of Chengdu University of Traditional Chinese Medicine, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Hui Zheng, MD, PhD
Acupuncture and Tuina School, Chengdu University of Traditional Chinese Medicine, No.1166 Liutai Avenue, Wenjiang District, Chengdu, China

Abstract

Background: Cognitive behavioral therapy (CBT) is recommended for irritable bowel syndrome (IBS). However, it remains unclear whether face-to-face CBT is as effective as digital, self-help, or telephone-delivered CBT for IBS.

Objective: This study aimed to estimate the relative effects of face-to-face CBT compared with digital, telephone-delivered, and self-help CBT for IBS and to assess whether there are adequate effective sample sizes to support the findings.

Methods: Ovid MEDLINE, Embase, and the Cochrane Library were searched up to September 27, 2025. Randomized controlled trials of face-to-face, digital, self-help, or telephone-delivered CBT for IBS in adults were included. The primary outcome was the IBS symptom severity scale. The secondary outcomes were IBS quality of life and abdominal pain intensity. A Bayesian random effects model was used for the meta-analysis. The effective and required sample sizes were calculated to estimate whether the sample sizes were adequate. The certainty of evidence was evaluated using the Confidence in Network Meta-Analysis Framework. The risk of bias of included studies was assessed using the Cochrane Collaboration's risk of bias tool (version 2).

Results: We analyzed 22 studies involving 3161 participants. The number of participants ranged between 28 and 558. The mean (SD) age of participants was 37.2 (10.6) years, and 78.6% (2485/3161) were women. These randomized controlled trials were published between 2003 and 2025. We found that face-to-face CBT had similar effects compared with digital CBT (mean difference [MD] −0.89, 95% credible interval [CrI] −20.78 to 18.73), self-help CBT (MD −1.73, 95% CrI −21.03 to 17.80), and telephone-delivered CBT (MD −0.76, 95% CrI −20.86 to 19.38) in improving IBS symptom severity scale scores. The comparison between face-to-face CBT and self-help CBT had sufficient effective sample sizes (375/140), whereas the effective sample sizes for comparisons with digital CBT (347/729) and telephone-delivered CBT (140/627) were insufficient. The certainty of evidence was moderate to low. Similarly, in improving quality of life and abdominal pain intensity, face-to-face CBT showed equal effect compared with digital and self-help CBT, with insufficient sample sizes and low to very low evidence certainty.

Conclusions: This is the first Bayesian meta-analysis to incorporate effective and required sample size calculations for comparisons among CBT modalities in IBS. We analyzed continuous data of the outcomes. Meanwhile, we computed the effective and required sample sizes, thereby quantifying the informational adequacy of each comparison. Our Bayesian meta-analysis demonstrated significant potential for digital, self-help, and telephone-delivered CBT for patients with IBS, but the effective sample sizes of most comparisons were inadequate. Digital, self-help, and telephone-delivered CBT can serve as important options for managing IBS in clinical practice. Given high heterogeneity, high risk of bias, and inadequate effective sample sizes, more high-quality studies are warranted.

Trial Registration: OSF Registries OSF.IO/ZW2HQ; https://doi.org/10.17605/OSF.IO/ZW2HQ

(J Med Internet Res 2026;28:e75833) doi:10.2196/75833

KEYWORDS

cognitive behavioral therapy; digital CBT; irritable bowel syndrome; indirect treatment comparison; meta-analysis

XSL•FO
RenderX

Introduction

Rationale

Irritable bowel syndrome (IBS) is a prevalent disorder of brain-gut interaction marked by recurrent abdominal pain and altered bowel habits [1]. It affects 5% to 10% of the global population [2], significantly reducing quality of life, productivity, and mental health [3-6]. IBS causes a huge burden with years lived with disability at 627 per 100,000 [7]. The condition imposes substantial economic burdens, with direct costs approximately US $1 billion and indirect costs reaching US $50 million [8].

Research has shown the effectiveness of brain-gut behavioral therapies in improving IBS symptoms and quality of life [9-11]. Cognitive behavioral therapy (CBT), which integrates cognitive and behavioral techniques, is one such approach for alleviating symptoms [12]. Previous randomized controlled trials (RCTs) and meta-analyses have demonstrated that face-to-face CBT effectively relieves gastrointestinal symptoms and enhances quality of life [13-15]. However, despite the robust evidence supporting face-to-face CBT, its implementation is limited by the need for skilled mental health providers and the time and financial burden it imposes on patients [16,17]. In regions with limited CBT availability, accessing face-to-face sessions can be challenging. Consequently, there is a need for effective, accessible, and cost-effective alternative treatments.

The advent of the internet and digital technologies has provided a promising solution to the challenges associated with traditional face-to-face CBT for IBS. Digital CBT, which incorporates interactive programs based on cognitive behavioral models specific to IBS, has been increasingly integrated into medical practice [18,19]. Additionally, telephone-delivered CBT and self-help CBT have emerged as alternative delivery methods for IBS treatment. While RCTs have explored the clinical efficacy of these digital and remote CBT approaches, most have compared them to usual care or waiting lists [19,20], with only a few studies directly comparing them to face-to-face CBT [18,21,22]. As a result, the relative effectiveness of face-to-face CBT versus digital, telephone-delivered, and self-help CBT in managing IBS remains unclear.

Black et al [23] have conducted a network meta-analysis to evaluate the efficacy of psychological therapies for IBS. They found that face-to-face, digital, telephone-delivered, and self-help CBT were all efficacious for global IBS symptoms or abdominal pain, but none was superior to the others. Similarly, Goodoory et al [15] have explored the brain-gut behavioral treatments for abdominal pain in patients with IBS. Their results also found that face-to-face, digital, telephone-delivered, and self-help CBT were all effective for overall abdominal pain, with no one approach superior to another. However, both studies dichotomized outcomes—abdominal pain or global IBS symptoms were classified as *improved* or *not improved*—a strategy that can result in loss of information, reduced statistical power, and an increased risk of false-positive findings [24]. Therefore, the relative effectiveness of these distinct CBT modalities for IBS management requires further validation.

Moreover, studies using active controls require a large number of participants to achieve sufficient statistical power, counteracting inherent expectation biases and biases introduced by blinding [16]. Previous meta-analyses reported comparative effect sizes without evaluating whether the sample sizes were sufficient to yield stable estimates.

Indirect treatment comparison (ITC) allows us to evaluate the relative effectiveness of different interventions by fully using existing studies when there is no or insufficient direct evidence [25]. In addition, calculating the effective sample sizes and required sample sizes for each comparison can further validate the robustness of the findings [26].

Objectives

Consequently, we conducted a Bayesian ITC meta-analysis with two primary objectives: (1) to estimate the relative effect of face-to-face CBT versus digital, telephone-delivered, and self-help CBT for IBS; and (2) to assess whether there are adequate effective sample sizes to support the findings.

Methods

Ethical Considerations

Our study was a systematic review and ITC meta-analysis of published studies and was exempt from review and approval by the research ethics committee. Ethical approval and consent to participate were acquired by each included study.

Protocol and Registration

The systematic review was designed, conducted, and reported following the PRISMA-NMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Network Meta-Analyses) guidelines [27] and the PRISMA-S (Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension) guidelines (Checklist 1) [28]. The systematic review had been previously registered on the Open Science Framework [29].

Eligibility Criteria

The inclusion criteria were as follows. (1) participants—adults (aged ≥18 y) diagnosed with IBS according to the Rome or Manning criteria; (2) interventions—at least one CBT intervention, categorized into four distinct delivery modalities by referring to a previous study [30]: (i) face-to-face CBT (therapist-guided, in-person sessions conducted individually or in groups), (ii) digital CBT (therapist guided, provided via web-based platforms or mobile apps), (iii) telephone-delivered CBT (therapist guided, administered via telephone), and (iv) self-help CBT (therapist guided or unguided, structured written or web-based self-help materials that patients use to implement CBT techniques independently; the definitions of different CBT interventions are presented in Table S4 in Multimedia Appendix 1); (3) comparisons—placebo, waiting list, usual care, or other active treatments; (4) outcomes—at least one of the following outcomes: IBS symptom severity scale (IBS-SSS), IBS quality of life (IBS-QOL), and abdominal pain intensity (API); and (4) study design—a parallel-design RCT or a crossover-design RCT with available first-period data.

XSL•FO
RenderX

The exclusion criteria were as follows: (1) full-text articles unavailable and (2) duplicate articles.

Information Sources and Search

MEDLINE, Embase, and the Cochrane Library were systematically searched from inception to September 29, 2024. A supplementary search was conducted on September 27, 2025. No single platform was used to search multiple databases. Medline was accessed via Ovid, whereas Embase and the Cochrane Library were searched through their respective official websites. There was no language restriction. Search terms encompassed *cognitive behavioral therapy* and *irritable bowel syndrome*. The search strategies are detailed in Table S1 - S3 in Multimedia Appendix 1. Our search strategies were not informed by previous reviews and were not peer-reviewed. Simultaneously, clinicaltrials.gov was also searched for potentially eligible studies. In addition, the reference lists of previous meta-analyses were screened for any eligible studies [15,23]. We did not search other online resources, browse websites, contact authors or experts, reach out to manufacturers, or use any other methods to obtain additional literature.

Study Selection

Search results were imported into Zotero (version 7.0.8; Corporation for Digital Scholarship). Duplicate records were removed manually. Then, two reviewers (QF-T and CH) independently scanned the title and abstract. Subsequently, they assessed the full texts of potential RCTs for eligibility. Any disagreements were resolved through discussion, with a third reviewer (HZ) consulted if necessary.

Outcome Assessments

The primary outcome was the change in IBS-SSS score. The secondary outcomes were the changes in IBS-QOL score and API score (measured by a visual analog scale or a gastrointestinal symptom diary for abdominal pain). Outcomes were evaluated at the end of treatment.

Data Collection Process and Data Items

A reviewer (QF-T) extracted data using a standardized form, and the data were checked by the second reviewer (CH) independently. The extracted items included characteristics of the included RCTs, specifics of the intervention and control groups, and outcomes data. Data extraction for the meta-analysis used within-group change scores and their corresponding SDs for continuous outcomes. Disagreements were addressed through discussion and were consulted with a third reviewer (HZ). For missing data, we attempted to obtain them by reviewing previous studies.

Risk of Bias Assessment

Risk of bias was assessed by two reviewers (JJ-M and XZ) independently using the Cochrane Collaboration's risk of bias tool (version 2) for randomized trials [31]. By evaluating the following five components of the problem, that is, randomization process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported result, each included RCT was assessed as having low risk, some concerns, or high risk of bias.

Statistical Analysis

Geometry of the Network

A network plot was generated in which the nodes represent the interventions and the edges represent the direct comparisons from RCTs. The size of each node was proportional to the number of participants receiving the intervention, and the thickness of the edges reflected the number of studies contributing to each direct comparison.

Summary Measures

As all the outcomes were continuous, the pooled effect size of the Bayesian analysis was estimated as mean difference (MD) with 95% credible intervals (CrIs). The effect size is the pooled estimate from a network meta-analysis that combines both direct and indirect evidence for each intervention. In addition, the surface under the cumulative rank curve (SUCRA) values were also calculated to determine the relative ranking of each intervention, with higher SUCRA values signifying a more favorable ranking for the intervention.

Planned Methods of Analysis

This systematic review was performed under a Bayesian framework with vague priors, which facilitates the integration of existing evidence with new data to update estimates of treatment effects and provide superior handling of uncertainty in small sample studies compared to traditional approaches. We assessed the fit of the random effects (REs) and fixed effects (FEs) models by examining their posterior total residual deviance, the deviance information criterion (DIC), and the number of unconstrained data points. A model was considered to have a better fit if it exhibited a posterior total residual deviance closer to the number of unconstrained data points and a lower DIC.

When two or more arms received the same intervention, we pooled their participants, means, and SD. Meanwhile, we pooled mindfulness, stress management, and education as an alternative psychotherapy. On the basis of the actual delivery methods used in the study, we categorized these interventions separately as alternative face-to-face psychotherapy, alternative self-help psychotherapy, or alternative digital psychotherapy. τ^2 was used to estimate the heterogeneity among RCTs, with a τ^2 of more than 0.36 indicating significant heterogeneity [23].

Assessment of Inconsistency

An unrelated mean effects model was fit by drawing the dev-dev plots to assess the inconsistency globally, and node-splitting was fit to assess the inconsistency locally for each potentially inconsistent comparison in turn.

Additional Analyses

Two sensitivity analyses were conducted to calculate the robustness of the results: (1) excluding the RCT that was rated as high risk of bias and (2) using frequentist framework as the statistical method; To explore the source of heterogeneity, we performed subgroup analyses according to treatment duration (<8 wk or ≥8 wk), the delivery format of face-to-face CBT (individual vs group), and the guidance level of self-help CBT (guided vs unguided) using the primary outcome data.

XSL•FO
RenderX

All the analyses were conducted in R version 4.3.1. The Bayesian analysis was performed using the *multinma* package, and the frequentist analysis used the *netmeta* package.

The Certainty of Evidence

The certainty of evidence of each outcome was graded using a web application—the Confidence in Network Meta-Analysis Framework [32]. This approach evaluated six domains, such as within-study bias, reporting bias, indirectness, imprecision, heterogeneity, and incoherence. The certainty of evidence of each outcome was graded as high, moderate, low, or very low.

The Estimate of Effective Sample Sizes and Required Sample Sizes

The effective number of trials and the effective sample sizes for the ITC were calculated using the method developed by Thorlund and Mills [26]. The effective number of trials represents the number of trials required in an ITC to achieve a comparable level of power and precision to that of a single direct head-to-head trial. The effective sample sizes denote the number of participants in the comparison that would provide the same degree and strength of evidence as that provided in an RCT. Additionally, the required sample sizes were estimated [33]. In this systematic review, a noninferiority design with 1:1 randomization was used to estimate the required sample sizes. If the effective sample sizes reach the required sample sizes, it would demonstrate the robustness of the findings; otherwise, further research is needed to confirm the findings.

Results

Study Selection

The initial database and registry searches identified 955 articles (Figure 1). After screening the titles and abstracts of 568 articles, 473 were excluded. Following a full-text review of 88 articles, 22 eligible RCTs were included [13,14,18-22,34-48].

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the literature search and study selection. IBS: irritable bowel syndrome; RCT: randomized controlled trial.

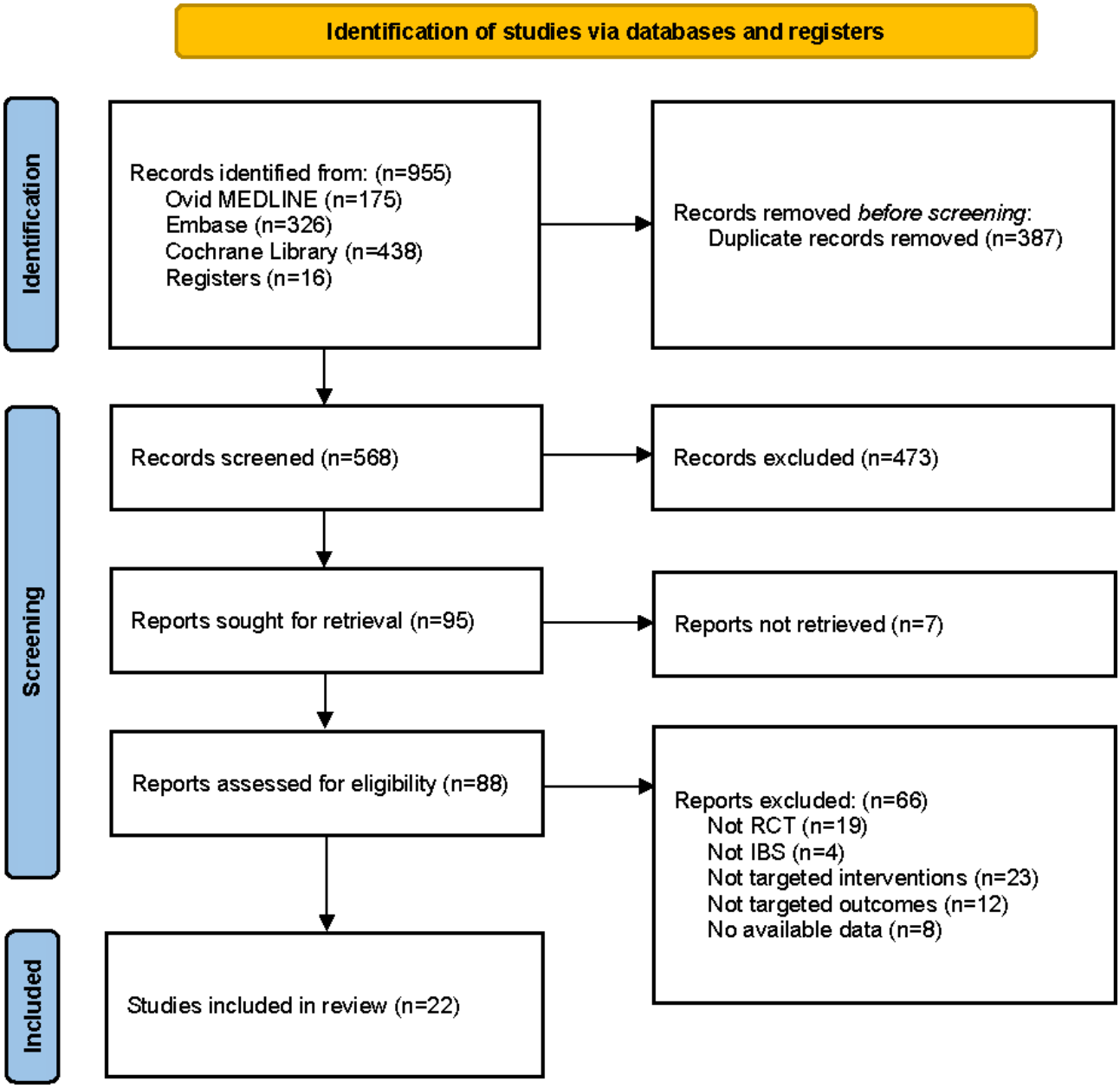

Study Characteristics

The characteristics of the eligible RCTs are presented in Table 1. A total of 3161 participants were enrolled in these RCTs, and the number of participants ranged between 28 and 558. The mean (SD) age of participants was 37.2 (10.6) years, and 78.6% were women. The RCTs included in our study were published between 2003 and 2025. Among these RCTs, 8 were from the United States (36.4%), 5 (22.7%) were from Sweden, 2 (9.1%) each from the United Kingdom, Korea, and Iran, and 1 (4.5%) each from the Netherlands, Canada, and Japan. In terms of intervention types, 7 (31.8%) RCTs reported face-to-face CBT, 6 (27.3%) RCTs reported digital CBT, 4 (18.2%) RCTs reported self-help CBT, 2 (9.1%) RCTs assessed both digital CBT and self-help CBT, 2 (9.1%) RCTs evaluated both self-help CBT and face-to-face CBT, and another (4.5%) examined both digital CBT and telephone-delivered CBT. All studies adopted the Rome criteria as the diagnostic standard. Two RCTs specifically focused on diarrhea-predominant IBS, while the remaining studies included all IBS subtypes.

XSL•FO
RenderX

Table . Characteristics of the included studies.

Study, year	Country	Study design	Disease	Diagnostic criteria	Interventions	Treatment duration (wk)	Population [sample size (n), female (%), mean ages (y)]	Results
Dehkordi and Solati, 2017 [34]	Iran	RCT[a]	IBS-D[b]	Rome III	Face-to-face CBT[c] versus usual care	8	64, 63, 33.7	IBS-QOL[d]: face-to-face CBT>usual care
Everitt et al, 2019 [35]	UK	RCT	IBS[e]	Rome III	Digital CBT versus telephone-delivered CBT versus usual care	48	558, 76, 43.1	IBS-SSS[f]: digital CBT=telephone-delivered CBT>usual care
Haghayegh et al, 2011 [36]	Iran	RCT	IBS-D	Rome II	Face-to-face CBT versus waiting list	8	32, 46, N/A[g]	IBS-QOL: face-to-face CBT>waiting list
Hunt et al, 2009 [38]	USA	RCT	IBS	Rome II	Digital CBT versus waiting list	6	54, 81, 38.5	IBS-QOL: digital CBT>waiting list
Hunt et al, 2015 [37]	USA	RCT	IBS	Rome III	Self-help CBT versus waiting list	6	60, 83, 36	IBS-QOL: self-help CBT>waiting list
Hunt et al, 2021 [20]	USA	RCT	IBS	Rome III	Self-help CBT versus waiting list	8	121, 75, 32	IBS-QOL: self-help CBT>waiting list
Hunt et al, 2025 [48]	USA	RCT	IBS	Rome IV	Self-help CBT versus alternative self-help psychotherapy	8	267, 72.3, 36.6	IBS-QOL: self-help CBT>alternative self-help psychotherapy
Jang et al, 2014 [39]	Korea	RCT	IBS	Rome III	Face-to-face CBT versus waiting list	8	90, 100, 21.6	IBS-QOL: face-to-face CBT>waiting list
Kennedy et al, 2005 [40]	UK	RCT	IBS	Rome I	Face-to-face CBT versus usual care	12	149, N/A, N/A	IBS-SSS: face-to-face CBT>usual care
Kikuchi et al, 2022 [13]	Japan	RCT	IBS	Rome III	Digital CBT versus face-to-face CBT versus usual care	10	114, 63, 39.7	IBS-SSS, IBS-QOL: digital CBT=face-to-face CBT>usual care

XSL•FO
RenderX

Study, year	Country	Study design	Disease	Diagnostic criteria	Interventions	Treatment duration (wk)	Population [sample size (n), female (%), mean ages (y)]	Results
Lackner et al, 2007 [14]	USA	RCT	IBS	Rome II	Face-to-face CBT versus alternative face-to-face psychotherapy versus waiting list	10	147, 82, 49.9	IBS-QOL: face-to-face CBT=alternative face-to-face psychotherapy=waiting list
Lackner et al, 2008 [21]	USA	RCT	IBS	Rome II	Face-to-face CBT versus self-help CBT versus waiting list	10	75, 86, 46.6	IBS-SSS, IBS-QOL: face-to-face CBT=self-help CBT>waiting list
Lackner et al, 2018 [22]	USA	RCT	IBS	Rome III	Face-to-face CBT versus self-help CBT versus alternative face-to-face psychotherapy	10	436, 80, 41.4	IBS-SSS: Face-to-face CBT=self-help CBT=alternative face-to-face psychotherapy
Lindfors 2020 [19]	Sweden	RCT	IBS	Rome IV	Digital CBT versus face-to-face CBT	3	141, 80, 37	IBS-SSS, IBS-QOL, API[h]: Digital CBT=face-to-face CBT
Ljótsson et al, 2010 [44]	Sweden	RCT	IBS	Rome III	Digital CBT versus waiting list	10	86, 85, 34.6	IBS-QOL, API: Digital CBT>waiting list
Ljótsson et al, 2011a [42]	Sweden	RCT	IBS	Rome III	Digital CBT versus alternative digital psychotherapy	10	195, 79, 38.9	IBS-QOL, API: Digital CBT=alternative digital psychotherapy
Ljótsson et al, 2011b [[41]	Sweden	RCT	IBS	Rome III	Digital CBT versus waiting list	10	61, 74, 34.9	IBS-QOL: Digital CBT>waiting list
Ljótsson et al, 2014 [43]	Sweden	RCT	IBS	Rome III	Digital CBT versus alternative digital psychotherapy	10	311, 80, 42.4	IBS-QOL: Digital CBT=alternative digital psychotherapy
Oerlemans et al, 2011 [19]	Netherlands	RCT	IBS	Rome III	Digital CBT versus usual care	4	76, 84, 38.3	IBS-QOL, API: Digital CBT>usual care

XSL•FO
RenderX

Study, year	Country	Study design	Disease	Diagnostic criteria	Interventions	Treatment duration (wk)	Population [sample size (n), female (%), mean ages (y)]	Results
Owusu et al, 2021 [45]	USA	RCT	IBS	Rome IV	Self-help CBT versus waiting list	12	36, 78, 39.2	IBS-SSS: self-help CBT>waiting list
Tkachuk et al, 2003 [46]	Canada	RCT	IBS	Rome II	Face-to-face CBT versus waiting list	9	28, 96, 39.5	API: face-to-face CBT=waiting list
Yang et al, 2022 [47]	Korea	RCT	IBS	Rome III	Face-to-face CBT versus waiting list	4	60, 88, 20.5	IBS-SSS, IBS-QOL: face-to-face CBT>waiting list

[a]RCT: randomized controlled trial.

[b]IBS-D: diarrhea-predominant irritable bowel syndrome.

[c]CBT: cognitive behavioral therapy.

[d]IBS-QOL: irritable bowel syndrome quality of life.

[e]IBS: irritable bowel syndrome.

[f]IBS-SSS: irritable bowel syndrome symptom severity scale.

[g]N/A: not available.

[h]API: abdominal pain intensity.

Risk of Bias

Risk of bias assessment is presented in Figure 2. Of the RCTs, 5 (22.7%) were adjudicated to be at high risk of bias, and 17 (77.3%) were adjudicated to have some concerns.

XSL•FO
RenderX

Figure 2. Risk of bias of included randomized controlled trials [13,14,18-22,34-48].

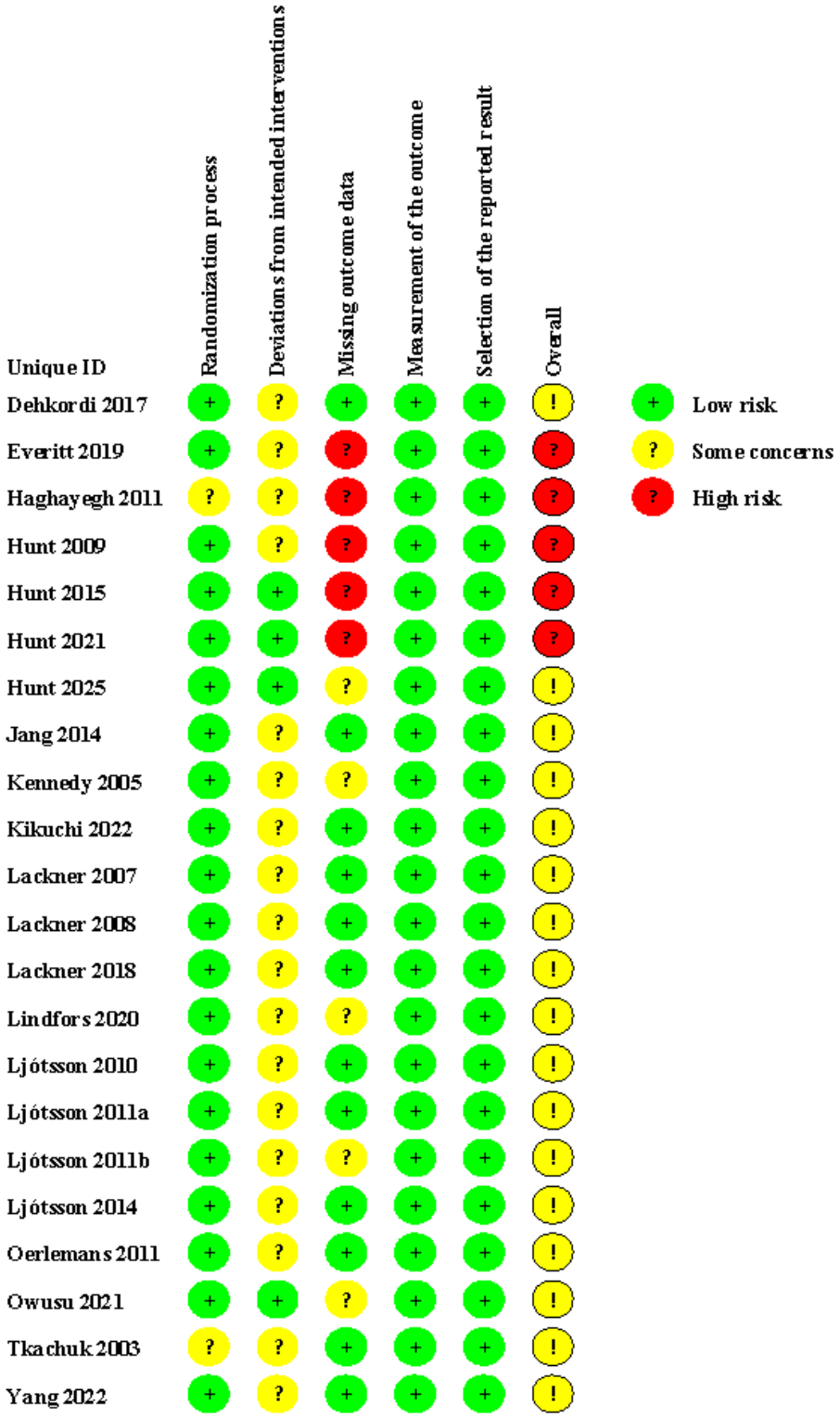

Exploration for Model Fit and Inconsistency

The results of the total posterior residuals deviance, the number of data points, and the DIC for both RE and FE models across all outcomes are presented in Table S5 in Multimedia Appendix 1. The RE model demonstrated a residual deviance closer to the number of data points and exhibited a lower DIC. Moreover, dev-dev plots indicated that all points were approximately aligned with the equality line, suggesting no evidence of global inconsistency (Figure S1 in Multimedia Appendix 1). Furthermore, the results of the node-splitting analysis indicated no local inconsistency (Figure S2 in Multimedia Appendix 1). Consequently, RE consistency models were selected for conducting the analyses.

Irritable Bowel Syndrome Symptom Severity Scale

A total of 8 RCTs, comprising 1562 participants, contributed data to the outcome analysis. The network plot is presented in

XSL•FO
RenderX

Figure S3 in Multimedia Appendix 1. Bayesian model results indicated that face-to-face CBT had similar effects compared to digital, telephone-delivered, and self-help CBT in reducing IBS-SSS (self-help CBT: MD −1.73, 95% CrI −21.03 to 17.80; digital CBT: MD −0.89, 95% CrI −20.78 to 18.73; telephone-delivered CBT: MD −0.76, 95% CrI −20.86 to 19.38; τ^2=24.61; Figure 3). The certainty of evidence was rated as moderate to low (Table S6 in Multimedia Appendix 1). The SUCRA rankings suggested that self-help CBT was the most effective (SUCRA 0.56, Table S7 in Multimedia Appendix 1). Furthermore, sensitivity analyses, which excluded high-risk RCTs (Figure S4 in Multimedia Appendix 1) and used a frequentist method (Figure S5 in Multimedia Appendix 1), yielded similar results, indicating the stability of the findings.

Figure 3. Effect of comparison between face-to-face CBT and other types of CBT for irritable bowel syndrome symptom severity scale [13,18,21,22,35,40,45,47]. The black vertical line corresponds to 0. CBT: cognitive behavioral therapy; CrI: credible interval; MD: mean difference.

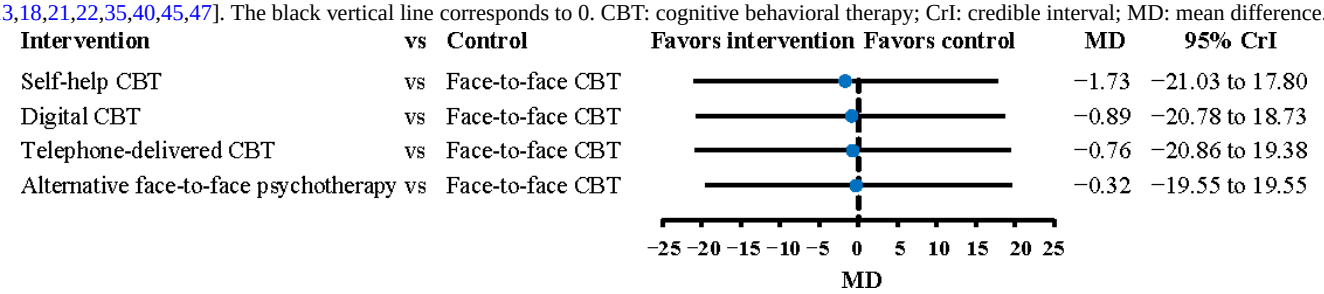

The results of the effective sample sizes and the required sample sizes are presented in Table 2 and Multimedia Appendix 2. Adequate effective sample sizes were observed for the comparison between face-to-face and self-help CBT (375/140, 267.9%; Table 2 and Multimedia Appendix 2). However, for the other comparisons, the effective sample sizes were insufficient, suggesting that more studies are needed in the future (digital CBT: 347/729, 47.6%; telephone-delivered CBT: 140/627, 22.3%; Table 2 and Multimedia Appendix 2).

Table . The effective sample sizes and required sample sizes estimation of outcomes.

Comparison (vs face-to-face CBT[a])	Effective number of trials (n/N)	Effective head-to-head sample sizes, n	Effective indirect sample sizes, n	Total effective sample sizes, n	Required sample sizes, n	Information fraction (n/N,%)	Network meta-analysis, MD[b] estimate (95% credible intervals)
IBS-SSS[c]							
Digital CBT	4/8	195	152	347	729	347/729, 47.6	−0.89 (−20.78 to 18.73)
Self-help CBT	4/8	339	36	375	140	375/140, 267.9	−1.73 (−21.03 to 17.80)
Telephone-delivered CBT	3/9	N/A[d]	140	140	627	140/627, 22.3	−0.76 (−20.86 to 19.38)
IBS-QOL[e]							
Digital CBT	8/9	195	107	302	3550	302/3550, 8.5	−4.29 (−19.10 to 10.59)
Self-help CBT	8/9	48	123	171	4610	171/4610, 3.7	−2.82 (−18.83 to 13.76)
API[f]							
Digital CBT	2/12	141	21	162	3764	162/3764, 4.3	−1.14 (−19.69 to 16.49)

[a]CBT: cognitive behavioral therapy.
[b]MD: mean difference.
[c]IBS-SSS: irritable bowel syndrome symptom severity scale.
[d]N/A: not applicable.
[e]IBS-QOL: irritable bowel syndrome quality of life.
[f]API: abdominal pain intensity.

Irritable Bowel Syndrome Quality of Life

A total of 17 RCTs, involving 1798 participants, contributed data to the outcome analysis. The network plot is presented in Figure S6 in Multimedia Appendix 1. Face-to-face CBT was found to be equally effective as digital CBT and self-help CBT in enhancing the quality of life for patients with IBS (digital CBT: MD −4.29, 95% CrI −19.10 to 10.59; self-help CBT: MD

XSL•FO
RenderX

−2.82, 95% CrI −18.83 to 13.76; τ^2=15.93; Figure 4). The Confidence in Network Meta-Analysis Framework evidence rating was very low (Table S6 in Multimedia Appendix 1). Digital CBT ranked higher than face-to-face CBT, with a SUCRA value of 0.71 (Table S7 in Multimedia Appendix 1). Sensitivity analyses confirmed the stability of these results after excluding high-risk RCTs (Figure S7 in Multimedia Appendix 1) and using a frequentist method (Figure S8 in Multimedia Appendix 1).

Figure 4. Effect of comparison between face-to-face CBT and other types of CBT for irritable bowel syndrome quality of life [13,14,18-21,34,36-39,41-44,47,48]. The black vertical line corresponds to 0. CBT: cognitive behavioral therapy; CrI: credible interval; MD: mean difference.

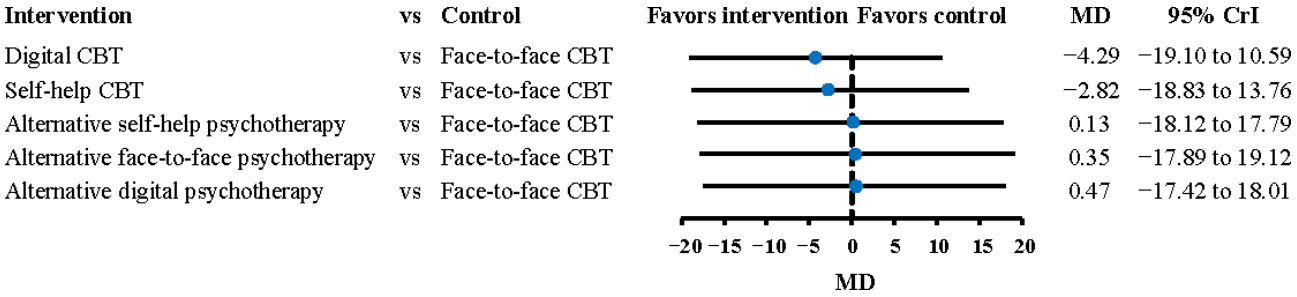

The effective sample sizes did not meet the required sample sizes, suggesting that more studies are warranted (digital CBT: 302/3550, 8.5%; self-help CBT: 171/4610, 3.7%; Table 2 and Multimedia Appendix 2).

Abdominal Pain Intensity

Four RCTs, involving 327 participants, contributed data to the outcome analysis. The network plot is presented in Figure S9 in Multimedia Appendix 1. Digital CBT showed similar effects compared with face-to-face CBT in reducing the API, with very low certainty of evidence (MD −1.14, 95% CrI −19.69 to 16.49, τ^2=23.66; Figure S10 in Multimedia Appendix 1, Table S6 in Multimedia Appendix 1). The SUCRA rankings indicated that digital CBT was ranked higher than face-to-face CBT (SUCRA 0.54 vs 0.49, Table S7 in Multimedia Appendix 1). The result was consistent with the sensitivity analysis using a frequentist method (Figure S11 in Multimedia Appendix 1).

As presented in Table 2 and Multimedia Appendix 2, the effective sample sizes did not meet the required sample sizes in the comparison between face-to-face CBT and digital CBT (141/3764, 4.3%), suggesting that more studies are needed.

Results of Subgroup Analysis

The results of the subgroup analysis of the primary outcome are presented in Figure S13 - S15 in Multimedia Appendix 1. We found that the duration of treatment (Figure S12 in Multimedia Appendix 1), the delivery method of face-to-face CBT (Figure S13 in Multimedia Appendix 1), or the guidance level of self-help CBT (Figure S14 in Multimedia Appendix 1) had no significant influence on the heterogeneity. Additionally, subgroup analysis suggested that the treatment duration (<8 wk or ≥8 wk; Figure S12 in Multimedia Appendix 1), the delivery method of face-to-face CBT (individual or group face-to-face CBT, Figure S13 in Multimedia Appendix 1), or the guidance level of self-help CBT (guided or unguided self-help CBT, Figure S14 in Multimedia Appendix 1) did not significantly modify the effect on global IBS symptoms.

Discussion

Summary of Evidence

In this systematic review, we assessed the relative effect of different delivery methods of CBT for patients with IBS and assessed the effective sample sizes and required sample sizes of each finding to evaluate the robustness. We included 22 RCTs involving 3161 adults with IBS. We found that face-to-face CBT presented a similar effect compared with self-help CBT in improving global IBS symptoms, with sufficient effective sample sizes. Face-to-face CBT showed an equal effect compared with digital CBT and telephone-delivered CBT in improving global IBS symptoms; however, the effective sample sizes were less than the required sample sizes. For quality of life and API, there were similar effects between face-to-face CBT and digital and self-help CBT, but the effective sample sizes were insufficient.

Face-to-face CBT is the primary mode of CBT delivery and is recognized as an effective intervention for various conditions, including major depressive disorder, insomnia, and IBS [1,49,50]. Our results showed that compared with face-to-face CBT, there were slight differences between digital, telephone-delivered, and self-help CBT in IBS-SSS, yet none reached the minimal clinically important difference value of 50 [51]. The similar effects might be explained by the fact that digital, telephone-delivered, and self-help CBT represent modifications of traditional face-to-face CBT in terms of delivery method. However, the core principles of CBT remain consistent across these modalities, ensuring efficacy while enhancing convenience and accessibility. Our findings are consistent with those of previous meta-analyses by Black et al [23] and Goodoory et al [15].

Compared to previous meta-analyses [15,23], their studies primarily used binary outcome measures and exclusively used frequentist analysis methods. In contrast, our study analyzed continuous outcomes for the IBS-SSS, IBS-QOL, and API. We used a Bayesian model for the primary analysis and a frequentist model for sensitivity analysis to confirm the stability of our findings. Additionally, our study is the first to incorporate effective and required sample size calculations for ITC within

XSL•FO
RenderX

CBT for IBS. We conducted a novel assessment of effective and required sample sizes, which has not been previously undertaken in CBT studies for IBS. Our results indicate that there are sufficient effective sample sizes to support that there is a similar effect between face-to-face CBT and self-help CBT in improving global IBS symptoms. However, other comparisons exhibited insufficient effective sample sizes.

Implication for Practice and Research

For clinical practice, the social and economic impact of IBS necessitates the demonstration of both clinical effectiveness and innovative direct-to-patient delivery systems to enhance patient access to appropriate therapeutic interventions [22]. Our findings indicate that digital, self-help, and telephone-delivered CBT have similar effects compared with face-to-face CBT. Meanwhile, a previous study has shown that a 10-week digital CBT treatment can reduce direct medical costs by US $358 and indirect costs by US $5014 [52]. Therefore, these alternative delivery methods deserve consideration in the clinical management of IBS. Notably, digital CBT has considerable potential for managing IBS because of its effectiveness [18], accessibility [38], and cost-effectiveness [52,53]. Additionally, our research found that both individualized face-to-face CBT and group face-to-face CBT demonstrated a similar effect in relieving the overall symptoms of patients with IBS compared to other treatments. Therefore, clinicians may choose between individual CBT and group CBT based on resource availability and patient preference in clinical practice.

Furthermore, for research, although our analysis suggests comparable effectiveness between digital CBT and face-to-face CBT, the effective sample size (n=347) represents only 47.6% of the required sample size (n=729) for this comparison. This indicates that while the current evidence is promising, implementation of digital CBT as a substitute for traditional face-to-face delivery in clinical practice would necessitate monitoring of larger patient cohorts to definitively confirm therapeutic equivalence and ensure consistent clinical outcomes.

Limitations

Several limitations should be considered when interpreting our findings. At the study and outcome level, first, none of the included studies were rated as low risk of bias, primarily due to the difficulty of blinding in RCTs of CBT. Moreover, five studies (22.7%) were rated as high risk of bias; hence, we excluded RCTs with high risk of bias for sensitivity analysis and found that the results were consistent with the main analysis. Second, there was significant heterogeneity among the included studies; therefore, we conducted subgroup analyses on treatment duration, the delivery method of face-to-face CBT, and the guidance level of self-help CBT. These analyses revealed that these factors did not significantly influence the heterogeneity. Initially, we intended to explore whether the source of heterogeneity was related to disease duration, disease subtype, and gender through subgroup analyses. However, this analysis could not be conducted due to the limited number of included studies, which is also attributed to the scarcity of high-quality RCTs on CBT for IBS, further emphasizing the need for more high-quality RCTs.

At the review level, first, we searched only Ovid MEDLINE, Embase, Cochrane Library, and ClinicalTrials.gov. We did not pursue additional literature via other online resources, correspondence with authors or experts, contact with manufacturers, or any other means; thus, some literature may have been missed. Nevertheless, we screened the reference lists of previous meta-analyses for any missed studies. Second, our search strategies did not consult an information scientist, which is another limitation of our study. Third, when estimating effective sample sizes, we used an estimation method that was not adjusted for heterogeneity, which may have overestimated the effective sample sizes [26]. Nonetheless, our results indicated that even if the effective sample sizes were overestimated, they still fell short of the required sample sizes for most comparisons. This finding further underscores the necessity for additional RCTs to validate our results in the future.

Conclusions

To our knowledge, this is the first Bayesian meta-analysis to incorporate effective and required sample size calculations for indirect treatment comparisons among CBT modalities in IBS. Compared with previous meta-analyses, we analyzed continuous outcomes of global symptoms, IBS-QOL, and API. For each comparison, we computed the effective sample size and the required sample size, thereby quantifying the informational adequacy. The meta-analysis suggested that digital, self-help, and telephone-delivered CBT had similar effects on global IBS symptoms, quality of life, and API as face-to-face CBT in patients with IBS. Therefore, digital CBT, self-help CBT, and telephone-delivered CBT can serve as important methods for managing IBS in clinical practice. However, our findings suggested that the effective sample sizes of most comparisons were insufficient. Given high heterogeneity, high risk of bias, and inadequate effective sample sizes, more high-quality studies are warranted in the future.

Acknowledgments

No generative artificial intelligence tools were used in any stage of the research, data analysis, or writing of this manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (2474658) and Chengdu Municipal Health Commission (WXLH202402020). These sponsors did not participate in the design of the study, the analyzed and interpretation the data, and the decision process to submit the manuscript for publication.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: HZ, MC
Data curation: QF-T, CH
Formal analysis: QF-T, CH
Funding acquisition: HZ, MC
Investigation: QF-T, CH
Methodology: QF-T, CH, XZ, JJ-M, CR-X, YX-Z
Project administration: MC
Software: QF-T, CH
Supervision: HZ
Validation: CH
Visualization: QF-T, CH, XZ, JJ-M
Writing – original draft: QF-T, CH
Writing – review & editing: XZ, JJ-M, CR-X, YX-Z, BL, XY-N, MC, HZ
All authors had full access to all the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis. All authors reviewed and approved the final manuscript.
QF-T and CH: co-first authors
HZ and MC are co-corresponding authors. MC can be contacted at Department of Colorectal Diseases, Hospital of Chengdu University of Traditional Chinese Medicine, No.39 Shi-er-qiao Road, Jinniu District, Chengdu 610000, China, cm@cdutcm.edu.cn

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies, definition of different cognitive behavioral therapy, model fit, certainty of evidence, consistency assessment, value of surface under the cumulative rank curve, network diagram, sensitivity analysis, and subgroup analysis.
[DOCX File, 3683 KB - jmir_v28i1e75833_app1.docx]

Multimedia Appendix 2
Assessment of effective number of trials, effective sample sizes, and required sample sizes.
[DOCX File, 2857 KB - jmir_v28i1e75833_app2.docx]

Checklist 1
PRISMA checklist.
[PDF File, 147 KB - jmir_v28i1e75833_app3.pdf]

References

1. Mearin F, Lacy BE, Chang L, et al. Bowel disorders. Gastroenterology 2016 Feb 18:S0016-5085(16)00222-5. [doi: 10.1053/j.gastro.2016.02.031] [Medline: 27144627]
2. Oka P, Parr H, Barberio B, Black CJ, Savarino EV, Ford AC. Global prevalence of irritable bowel syndrome according to Rome III or IV criteria: a systematic review and meta-analysis. Lancet Gastroenterol Hepatol 2020 Oct;5(10):908-917. [doi: 10.1016/S2468-1253(20)30217-X] [Medline: 32702295]
3. Peery AF, Crockett SD, Murphy CC, et al. Burden and cost of gastrointestinal, liver, and pancreatic diseases in the United States: update 2021. Gastroenterology 2022 Feb;162(2):621-644. [doi: 10.1053/j.gastro.2021.10.017] [Medline: 34678215]
4. Black CJ, Ford AC. Global burden of irritable bowel syndrome: trends, predictions and risk factors. Nat Rev Gastroenterol Hepatol 2020 Aug;17(8):473-486. [doi: 10.1038/s41575-020-0286-8] [Medline: 32296140]
5. Frändemark Å, Törnblom H, Jakobsson S, Simrén M. Work productivity and activity impairment in irritable bowel syndrome (IBS): a multifaceted problem. Am J Gastroenterol 2018 Oct;113(10):1540-1549. [doi: 10.1038/s41395-018-0262-x] [Medline: 30254230]
6. Kutschke J, Harris JR, Bengtson MB. The relationships between IBS and perceptions of physical and mental health-a Norwegian twin study. BMC Gastroenterol 2022 May 28;22(1):266. [doi: 10.1186/s12876-022-02340-8] [Medline: 35643443]
7. Porter CK, Thura N, Riddle MS. Quantifying the incidence and burden of postinfectious enteric sequelae. Mil Med 2013 Apr;178(4):452-469. [doi: 10.7205/MILMED-D-12-00510] [Medline: 23707833]

XSL•FO
RenderX

8. Everhart JE, Ruhl CE. Burden of digestive diseases in the United States part II: lower gastrointestinal diseases. Gastroenterology 2009 Mar;136(3):741-754. [doi: 10.1053/j.gastro.2009.01.015] [Medline: 19166855]
9. Henrich JF, Gjelsvik B, Surawy C, Evans E, Martin M. A randomized clinical trial of mindfulness-based cognitive therapy for women with irritable bowel syndrome-Effects and mechanisms. J Consult Clin Psychol 2020 Apr;88(4):295-310. [doi: 10.1037/ccp0000483] [Medline: 32134291]
10. Berry SK, Berry R, Recker D, Botbyl J, Pun L, Chey WD. A randomized parallel-group study of digital gut-directed hypnotherapy vs muscle relaxation for irritable bowel syndrome. Clin Gastroenterol Hepatol 2023 Nov;21(12):3152-3159. [doi: 10.1016/j.cgh.2023.06.015] [Medline: 37391055]
11. Gaylord SA, Palsson OS, Garland EL, et al. Mindfulness training reduces the severity of irritable bowel syndrome in women: results of a randomized controlled trial. Am J Gastroenterol 2011 Sep;106(9):1678-1688. [doi: 10.1038/ajg.2011.184] [Medline: 21691341]
12. Hayes SC, Hofmann SG. The third wave of cognitive behavioral therapy and the rise of process-based care. World Psychiatry 2017 Oct;16(3):245-246. [doi: 10.1002/wps.20442] [Medline: 28941087]
13. Kikuchi S, Oe Y, Ito Y, et al. Group cognitive-behavioral therapy with interoceptive exposure for drug-refractory irritable bowel syndrome: a randomized controlled trial. Am J Gastroenterol 2022 Apr 1;117(4):668-677. [doi: 10.14309/ajg.0000000000001664] [Medline: 35103022]
14. Lackner JM, Jaccard J, Krasner SS, Katz LA, Gudleski GD, Blanchard EB. How does cognitive behavior therapy for irritable bowel syndrome work? A mediational analysis of a randomized clinical trial. Gastroenterology 2007 Aug;133(2):433-444. [doi: 10.1053/j.gastro.2007.05.014] [Medline: 17681164]
15. Goodoory VC, Khasawneh M, Thakur ER, et al. Effect of brain-gut behavioral treatments on abdominal pain in irritable bowel syndrome: systematic review and network meta-analysis. Gastroenterology 2024 Oct;167(5):934-943. [doi: 10.1053/j.gastro.2024.05.010] [Medline: 38777133]
16. Staudacher HM, Black CJ, Teasdale SB, Mikocka-Walus A, Keefer L. Irritable bowel syndrome and mental health comorbidity-approach to multidisciplinary management. Nat Rev Gastroenterol Hepatol 2023 Sep;20(9):582-596. [doi: 10.1038/s41575-023-00794-z] [Medline: 37268741]
17. Camilleri M, Dilmaghani S. Update on treatment of abdominal pain in irritable bowel syndrome: a narrative review. Pharmacol Ther 2023 May;245:108400. [doi: 10.1016/j.pharmthera.2023.108400] [Medline: 37001737]
18. Lindfors P, Axelsson E, Engstrand K, et al. Online education is non-inferior to group education for irritable bowel syndrome: a randomized trial and patient preference trial. Clin Gastroenterol Hepatol 2021 Apr;19(4):743-751. [doi: 10.1016/j.cgh.2020.04.005] [Medline: 32289541]
19. Oerlemans S, van Cranenburgh O, Herremans PJ, Spreeuwenberg P, van Dulmen S. Intervening on cognitions and behavior in irritable bowel syndrome: a feasibility trial using PDAs. J Psychosom Res 2011 Mar;70(3):267-277. [doi: 10.1016/j.jpsychores.2010.09.018] [Medline: 21334498]
20. Hunt M, Miguez S, Dukas B, Onwude O, White S. Efficacy of zemedy, a mobile digital therapeutic for the self-management of irritable bowel syndrome: crossover randomized controlled trial. JMIR Mhealth Uhealth 2021 May 20;9(5):e26152. [doi: 10.2196/26152] [Medline: 33872182]
21. Lackner JM, Jaccard J, Krasner SS, Katz LA, Gudleski GD, Holroyd K. Self-administered cognitive behavior therapy for moderate to severe irritable bowel syndrome: clinical efficacy, tolerability, feasibility. Clin Gastroenterol Hepatol 2008 Aug;6(8):899-906. [doi: 10.1016/j.cgh.2008.03.004] [Medline: 18524691]
22. Lackner JM, Jaccard J, Keefer L, et al. Improvement in gastrointestinal symptoms after cognitive behavior therapy for refractory irritable bowel syndrome. Gastroenterology 2018 Jul;155(1):47-57. [doi: 10.1053/j.gastro.2018.03.063] [Medline: 29702118]
23. Black CJ, Thakur ER, Houghton LA, Quigley EMM, Moayyedi P, Ford AC. Efficacy of psychological therapies for irritable bowel syndrome: systematic review and network meta-analysis. Gut 2020 Aug;69(8):1441-1451. [doi: 10.1136/gutjnl-2020-321191] [Medline: 32276950]
24. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ 2006 May 6;332(7549):1080. [doi: 10.1136/bmj.332.7549.1080] [Medline: 16675816]
25. Song F, Altman DG, Glenny AM, Deeks JJ. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. BMJ 2003 Mar 1;326(7387):472. [doi: 10.1136/bmj.326.7387.472] [Medline: 12609941]
26. Thorlund K, Mills EJ. Sample size and power considerations in network meta-analysis. Syst Rev 2012 Sep 19;1:41. [doi: 10.1186/2046-4053-1-41] [Medline: 22992327]
27. Hutton B, Salanti G, Caldwell DM, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. Ann Intern Med 2015 Jun 2;162(11):777-784. [doi: 10.7326/M14-2385] [Medline: 26030634]
28. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
29. OSF. Digital versus face-to-face, telephone-delivered, and self-helped cognitive behavioral therapy for irritable bowel syndrome: an indirect treatment comparison meta-analysis. URL: https://osf.io/zw2hq/overview

XSL•FO
RenderX

30. Gao Y, Ge L, Liu M, et al. Comparative efficacy and acceptability of cognitive behavioral therapy delivery formats for insomnia in adults: a systematic review and network meta-analysis. Sleep Med Rev 2022 Aug;64:101648. [doi: 10.1016/j.smrv.2022.101648] [Medline: 35759820]
31. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
32. Nikolakopoulou A, Higgins JPT, Papakonstantinou T, et al. CINeMA: an approach for assessing confidence in the results of a network meta-analysis. PLoS Med 2020 Apr;17(4):e1003082. [doi: 10.1371/journal.pmed.1003082] [Medline: 32243458]
33. Zhong B. How to calculate sample size in randomized controlled trial? J Thorac Dis 2009 Dec;1(1):51-54. [Medline: 22263004]
34. Dehkordi AH, Solati K. The effects of cognitive behavioral therapy and drug therapy on quality of life and symptoms of patients with irritable bowel syndrome. J Adv Pharm Technol Res 2017;8(2):67-72. [doi: 10.4103/japtr.JAPTR-170-16]
35. Everitt HA, Landau S, O'Reilly G, et al. Assessing telephone-delivered cognitive-behavioural therapy (CBT) and web-delivered CBT versus treatment as usual in irritable bowel syndrome (ACTIB): a multicentre randomised trial. Gut 2019 Sep;68(9):1613-1623. [doi: 10.1136/gutjnl-2018-317805] [Medline: 30971419]
36. Haghayegh SA, Kalantari M, Molavi H, Talebi M, Alemohamad J. The efficacy of cognitive-behavior group therapy on health-related quality of life, health anxiety and depression in patients with diarrhea-predominant irritable bowel syndrome. Pak J Med Sci 2011;27(4):749-753. [doi: 10.1002/central/CN-00850913/full]
37. Hunt MG, Ertel E, Coello JA, Rodriguez L. Empirical support for a self-help treatment for IBS. Cogn Ther Res 2015 Apr;39(2):215-227. [doi: 10.1007/s10608-014-9647-3]
38. Hunt MG, Moshier S, Milonova M. Brief cognitive-behavioral internet therapy for irritable bowel syndrome. Behav Res Ther 2009 Sep;47(9):797-802. [doi: 10.1016/j.brat.2009.05.002] [Medline: 19570525]
39. Jang AL, Hwang SK, Kim DU. The effects of cognitive behavioral therapy in female nursing students with irritable bowel syndrome: a randomized trial. Eur J Gastroenterol Hepatol 2014 Aug;26(8):918-926. [doi: 10.1097/MEG.0000000000000140] [Medline: 24999797]
40. Kennedy T, Jones R, Darnley S, Seed P, Wessely S, Chalder T. Cognitive behaviour therapy in addition to antispasmodic treatment for irritable bowel syndrome in primary care: randomised controlled trial. BMJ 2005 Aug 20;331(7514):435. [doi: 10.1136/bmj.38545.505764.06] [Medline: 16093252]
41. Ljótsson B, Andersson G, Andersson E, et al. Acceptability, effectiveness, and cost-effectiveness of internet-based exposure treatment for irritable bowel syndrome in a clinical sample: a randomized controlled trial. BMC Gastroenterol 2011 Oct 12;11:110. [doi: 10.1186/1471-230X-11-110] [Medline: 21992655]
42. Ljótsson B, Hedman E, Andersson E, et al. Internet-delivered exposure-based treatment vs. stress management for irritable bowel syndrome: a randomized trial. Am J Gastroenterol 2011 Aug;106(8):1481-1491. [doi: 10.1038/ajg.2011.139] [Medline: 21537360]
43. Ljótsson B, Hesser H, Andersson E, et al. Provoking symptoms to relieve symptoms: a randomized controlled dismantling study of exposure therapy in irritable bowel syndrome. Behav Res Ther 2014 Apr;55:27-39. [doi: 10.1016/j.brat.2014.01.007] [Medline: 24584055]
44. Ljótsson B, Falk L, Vesterlund AW, et al. Internet-delivered exposure and mindfulness based therapy for irritable bowel syndrome--a randomized controlled trial. Behav Res Ther 2010 Jun;48(6):531-539. [doi: 10.1016/j.brat.2010.03.003] [Medline: 20362976]
45. Owusu JT, Sibelli A, Moss-Morris R, van Tilburg MAL, Levy RL, Oser M. A pilot feasibility study of an unguided, internet-delivered cognitive behavioral therapy program for irritable bowel syndrome. Neurogastroenterol Motil 2021 Nov;33(11):e14108. [doi: 10.1111/nmo.14108] [Medline: 33745228]
46. Tkachuk GA, Graff LA, Martin GL, Bernstein CN. Randomized controlled trial of cognitive–behavioral group therapy for irritable bowel syndrome in a medical setting. J Clin Psychol Med Settings 2003 Mar;10(1):57-69. [doi: 10.1023/A:1022809914863]
47. Yang YY, Jun S. The effects of cognitive behavioral therapy for insomnia among college students with irritable bowel syndrome: a randomized controlled trial. Int J Environ Res Public Health 2022 Oct 29;19(21):14174. [doi: 10.3390/ijerph192114174] [Medline: 36361052]
48. Hunt M, Dalvie A, Ipek S, Glinski S, Macks R. Efficacy of a CBT self-help app (Zemedy) versus an education, relaxation, and mindfulness app for IBS: results from post-treatment, 3-month, and 6-month follow-up. J Clin Gastroenterol 2025 Mar 12. [doi: 10.1097/MCG.0000000000002164] [Medline: 40071822]
49. Otte C, Gold SM, Penninx BW, et al. Major depressive disorder. Nat Rev Dis Primers 2016 Sep 15;2:16065. [doi: 10.1038/nrdp.2016.65] [Medline: 27629598]
50. Perlis ML, Posner D, Riemann D, Bastien CH, Teel J, Thase M. Insomnia. Lancet 2022 Sep 24;400(10357):1047-1060. [doi: 10.1016/S0140-6736(22)00879-0] [Medline: 36115372]
51. Francis CY, Morris J, Whorwell PJ. The irritable bowel severity scoring system: a simple method of monitoring irritable bowel syndrome and its progress. Aliment Pharmacol Ther 1997 Apr;11(2):395-402. [doi: 10.1046/j.1365-2036.1997.142318000.x] [Medline: 9146781]

XSL•FO
RenderX

52. Andersson E, Ljótsson B, Smit F, et al. Cost-effectiveness of internet-based cognitive behavior therapy for irritable bowel syndrome: results from a randomized controlled trial. BMC Public Health 2011 Apr 7;11(1):215. [doi: 10.1186/1471-2458-11-215] [Medline: 21473754]
53. Wallén H, Lindfors P, Andersson E, et al. Return on investment of internet delivered exposure therapy for irritable bowel syndrome: a randomized controlled trial. BMC Gastroenterol 2021 Jul 13;21(1):289. [doi: 10.1186/s12876-021-01867-6] [Medline: 34256715]

Abbreviations

API: abdominal pain intensity
CBT: cognitive behavioral therapy
CrI: credible interval
DIC: Deviance information criteria
FE: fixed effect
IBS: irritable bowel syndrome
IBS-QOL: irritable bowel syndrome quality of life
IBS-SSS: irritable bowel syndrome symptom severity scale
ITC: indirect treatment comparison
MD: mean difference
PRISMA-NMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Network Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension
RCT: randomized controlled trial
RE: random effect
SUCRA: Surface under the cumulative rank curve

Edited by S Brini; submitted 11.Apr.2025; peer-reviewed by E Sahker, P Sangvatanakul; revised version received 29.Nov.2025; accepted 01.Dec.2025; published 08.Jan.2026.

Please cite as:
Tao QF, Hua C, Zhuo X, Mou JJ, Xie CR, Zhang YX, Lv B, Niu XY, Chen M, Zheng H
Face-to-Face Versus Digital, Telephone-Delivered, and Self-Help Cognitive Behavioral Therapy for Irritable Bowel Syndrome: Systematic Review and Bayesian Indirect Treatment Comparison Meta-Analysis
J Med Internet Res 2026;28:e75833
URL: https://www.jmir.org/2026/1/e75833
doi:10.2196/75833

© Qing-Feng Tao, Can Hua, Xiao Zhuo, Jian-Jiao Mou, Chao-Rong Xie, Yu-Xin Zhang, Bei Lv, Xin-Ying Niu, Min Chen, Hui Zheng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Interventions Targeting Healthy and Sustainable Eating Behavior: Systematic Review and Meta-Analysis

Käbi Vanwinkelen, MA; Bram Spruyt, MA; Tim Smits, Prof Dr

Media, Information and Persuasion Lab, Department of Communication Science, KU Leuven, Parkstraat 45, Leuven, Belgium

Corresponding Author:
Käbi Vanwinkelen, MA
Media, Information and Persuasion Lab, Department of Communication Science, KU Leuven, Parkstraat 45, Leuven, Belgium

Abstract

Background: Current food consumption patterns contribute to the rising prevalence of obesity and noncommunicable diseases and exacerbate environmental degradation. Digital media offer promising opportunities to promote healthier and more sustainable eating; yet, evidence regarding their effectiveness remains fragmented.

Objective: The aim of this systematic review and meta-analysis is (1) to evaluate the effectiveness of digital interventions in improving healthy and sustainable food consumption and (2) to identify which participant and intervention characteristics are associated with greater effectiveness.

Methods: A systematic search was conducted in January 2024 and repeated in September 2025 across Web of Science, Embase, and Scopus, supplemented with forward and backward reference searching. Eligible studies were those with a quasi-experimental or longitudinal design evaluating digital interventions targeting nonclinical populations, with the aim of increasing plant-based food consumption or reducing animal-based food intake. Risk of bias was assessed using the Cochrane risk-of-bias tool. Included interventions were coded for behavior change techniques using the Behavior Change Taxonomy version 1. A random-effects meta-analysis with robust variance estimation was performed, and moderator analyses were conducted with participant and intervention characteristics.

Results: Eligibility screening led to the inclusion of 52 papers published between 2004 and 2025, with 24,652 participants in total. The meta-analysis revealed a small but statistically significant positive effect of digital interventions on food consumption outcomes (d=0.33, 95% CI 0.25 - 0.42; P<.001). However, substantial heterogeneity (I^2=86%, 95% prediction interval −0.21 to 0.87) indicates considerable variation in effectiveness across intervention characteristics. A moderator analysis showed no significant difference in effectiveness (P=.53) between interventions aimed at reducing meat consumption (d=0.38, 95% CI 0.20 - 0.57; P<.001) and those promoting plant-based eating (d=0.33, 95% CI 0.23 - 0.42; P<.001). Although digital interventions had the strongest effects among young adults (d=0.46, 95% CI 0.30 - 0.61; P<.001), age-related differences were not statistically significant. Intervention effectiveness differed significantly by platform (P=.03), with social media interventions (d=0.65, 95% CI 0.41 - 0.90; P<.001) yielding stronger effects than other modalities. Incorporating prompts or cues significantly enhanced effectiveness (d=0.58 vs d=0.30; P=.04). Although not statistically significant, interventions including social support or behavioral comparison (both d=0.39; P<.001) yielded larger effects. Few studies included adolescents or individuals from lower socioeconomic backgrounds.

Conclusions: This review underscores the innovative potential of digital interventions in improving eating behavior, highlighting how effectiveness varies by intervention design. Social media emerge as particularly promising, likely due to their unique social and interactive features. By pinpointing the contexts and types of digital interventions that most effectively promote plant-based eating, this study provides timely guidance for researchers and practitioners in increasingly digitalized food environments. Nonetheless, more high-quality studies are needed to confirm these insights and address the critical gap among adolescents and low socioeconomic groups.

Trial Registration: PROSPERO CRD42023487955; https://www.crd.york.ac.uk/PROSPERO/view/CRD42023487955

(J Med Internet Res 2026;28:e80821) doi:10.2196/80821

KEYWORDS

diet; healthy eating; digital media; social media; best practices; internet-based intervention; nutrition intervention; digital interventions; systematic review; meta-analysis

XSL•FO
RenderX

Introduction

Food plays a pivotal role in both human and planetary health. Current food consumption patterns are driving the steep increase in obesity and noncommunicable diseases such as diabetes or cancers [1]. Simultaneously, contemporary dietary habits contribute significantly to greenhouse gas emissions, deforestation, and water scarcity, thereby exacerbating environmental degradation [2]. As high intake of animal-based foods plays a substantial role herein, the EAT-Lancet Commission emphasizes the urgency of a worldwide shift to healthy and sustainable diets mainly characterized by a variety of plant-based foods (eg, fruits and vegetables, whole grains, and legumes) and low quantities of animal-based foods [2,3]. Adherence to the EAT-Lancet diet is associated with a lower risk of diabetes, cardiovascular disease, and cancer-related mortality while reducing greenhouse gas emissions by 29% [4-6]. The most profound health and environmental benefits can be achieved by reducing meat consumption and increasing the intake of fruit, vegetables, and legumes [7-9]. Nevertheless, global meat consumption has risen substantially over the past 5 decades and is projected to continue increasing, while intake of fruits, vegetables, and legumes remains inadequate [10-13]. Hence, it is crucial to explore effective strategies for promoting plant-based dietary patterns.

The widespread use of digital media and their capacity to influence consumption patterns through the promotion of unhealthy foods have sparked growing interest among researchers in leveraging them for health interventions [14,15]. Digital interventions can be delivered through a variety of platforms, including mobile apps, SMS text messaging, websites, and perhaps, most notably, social media. While digital intervention studies often still rely on custom-built or more traditional platforms, health researchers are increasingly exploring the potential of social media for dietary interventions and health promotion campaigns [16-19]. These developments highlight the importance of evaluating and comparing the effectiveness of different digital platforms for interventions [20,21].

To achieve dietary behavior change, interventions incorporate behavior change techniques (BCTs) [22]. BCTs are the smallest identifiable components of an intervention that can independently influence behavior [23,24]. Strategically implementing BCTs has been consistently emphasized as essential for developing effective interventions [22,25]. However, it remains unclear which techniques are most effective in improving healthy and sustainable eating and which combinations of BCTs and digital intervention platforms enhance the intervention's effectiveness. Hence, systematically reviewing and identifying the BCTs most strongly associated with the intervention's effects could enhance the design of future programs targeting food consumption [23].

Although digital interventions targeting eating behavior are gaining popularity, prior reviews report mixed conclusions regarding their effectiveness [18,19,26-29]. Notably, many of these reviews were narrow in scope: some focused on specific food categories, such as sugar-sweetened beverages or vegetables [19,30,31], targeted specific age groups [19,32,33], or selectively included specific digital platforms [18,19,33,34]. Moreover, previous systematic reviews that summarize the literature on digital interventions often overlook social media and focus on research-created platforms or (now) outdated digital tools [16,17,19,35]. Systematic reviews that do consider social media use limited keywords in their search strategy or focus exclusively on social media platforms, missing the opportunity to compare traditional digital interventions with social media interventions [18,28,36,37]. In addition, few reviews assess which BCTs are most successful in targeting dietary change through digital interventions, nor do they explore whether effectiveness differs across age groups or socioeconomic groups [38]. This is critical, as individuals with lower socioeconomic status (SES) are disproportionately affected by poor diets, and the interventions that succeed in high SES populations may not be equally effective for low-income groups [19,39].

While narrow reviews are valuable for exploring targeted questions in depth, there is a need for a comprehensive synthesis that compares intervention strategies across different digital media to gain a more profound understanding of the key components that contribute to a successful digital intervention. Therefore, the aim of this systematic review and meta-analysis is (1) to evaluate the overall effectiveness of digital interventions in terms of increased healthy or sustainable eating behavior and (2) to identify when interventions are (more) effective by considering the behavioral goal orientation (prevention vs promotion approach), intervention characteristics (ie, the digital platform and BCTs), and participant characteristics (ie, age and SES).

Methods

The reporting of this systematic review is in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) guidelines (Checklist 1) [40,41], and its protocol was prospectively registered in PROSPERO (CRD42023487955). This study was a secondary analysis of published literature and did not require ethics approval.

Data Sources and Search Strategy

Systematic searches were conducted on January 19, 2024, in Web of Science core collection, Embase, and Scopus, with an updated search run on September 23, 2025, to retrieve newly published papers. An experienced academic librarian was consulted to optimize the search strategy and to ensure that database-specific operators were used correctly. The final search string was limited to peer-reviewed papers published in English and contained keywords related to digital media, interventions, and healthy and sustainable eating behavior (Multimedia Appendix 1). No date restrictions or other published search filters were applied. The reference lists of all included papers were manually searched, and a citation search using Web of Science was conducted. Authors of all included studies were contacted for additional relevant papers. No other online resources or study registries were searched.

XSL•FO
RenderX

Eligibility Criteria

Studies were included if they met the following PICO-based criteria. Reviews, dissertations, books, and conference papers were excluded.

Population

Studies had to recruit a nonclinical population of participants. Studies exclusively including individuals with specific health conditions or specific nutritional needs (eg, diabetes, pregnancy, and eating disorders) were excluded. Similarly, populations primarily composed of individuals with overweight or obesity were not considered because of differences in appetite [42].

Intervention

Eligible studies had to test a digital intervention designed to promote healthy and sustainable eating behavior, specifically by increasing plant-based food consumption or reducing intake of animal-based products. An intervention was defined as "a coordinated set of activities designed to change specified behavior patterns" [43], thus excluding interventions with a one-time exposure, as these rarely allow for sustained behavioral changes. Multicomponent interventions with an offline element were not included, as effects (or lack thereof) cannot be specifically attributed to the digital component.

Comparison

Studies were required to include a form of baseline comparison or control group that could either be inactive (ie, no intervention and waitlist) or active (ie, alternative intervention). Studies had to use a quasi-experimental (eg, randomized controlled trial [RCT]) or longitudinal design.

Outcomes

Eligible studies had to provide an outcome measure related to eating behavior, such as food consumption, intention, choice, or purchase. These outcomes could be related to intake of plant-based foods (eg, fruits, vegetables, and legumes) and animal-based foods (eg, red or processed meat) or adherence to a healthy and sustainable diet (eg, Mediterranean diet).

Data Extraction

KV and BS independently extracted relevant data using the preregistered extraction book, covering study characteristics (eg, authors, year of publication, and country of origin), study design (eg, comparator and research question), population (eg, age, gender, and SES indicators), intervention characteristics (eg, digital medium, duration, BCTs, and approach), outcomes (eg, behavioral outcomes, metrics, and covariates), and results (eg, analysis technique and statistical results). The 3 reviewers collaborated closely to resolve any discrepancies during the data extraction process.

Coding of BCTs

BCTs that were not explicitly reported in the included studies were coded by 2 reviewers (KV and BS) using the Behavior Change Technique Taxonomy version 1 (BCTTv1) [24]. One reviewer (KV) coded all studies, and a second reviewer (BS) randomly double-coded 5, achieving 90% agreement. Both reviewers were certified coders, as they completed the online taxonomy training ("BCTTv1 Online Training"). Discrepancies were resolved via discussion between the 2 reviewers. For studies with multiple intervention groups (IG), BCTs were coded separately for each active arm. When insufficient detail was provided to code the BCTs, related documents (eg, protocols) were consulted.

Risk of Bias Assessment

Included studies were appraised for study quality using the Cochrane risk-of-bias tool [44]. For RCTs, the Revised RoB 2 tool was used; for cluster RCTs, an adapted version of RoB 2; and for nonrandomized studies (NRS), ROBINS-I [45-47]. The risk of bias assessment was conducted with consideration of best practices in food and communication research. While self-reported measures typically carry a moderate risk of bias, this was not considered problematic, as self-reporting is often the most appropriate or even the only feasible method for measuring eating behavior.

Risk of bias assessments were independently conducted by 2 reviewers (KV and BS), and discrepancies were resolved through consultation with a third reviewer (TS). One paper included 2 distinct experimental studies [48], resulting in a total of 53 separate studies evaluated for risk of bias. The majority of RCTs, including all 3 cluster RCTs, had a moderate risk of bias, primarily due to the use of self-report measures or lack of prespecified analysis plan (31/39, 79%; Figure 1) [49]. In total, 5 RCTs were rated as high risk and 3 as low risk. NRS were rated as moderate (8/14, 57%) or high (6/14, 43%) risk (Figure 2), commonly due to baseline confounding and issues similar to those in the RCTs (see Multimedia Appendix 2 [48,50-99] for a detailed overview of the risk of bias assessments).

Figure 1. Visualization of the risk of bias of randomized controlled trials (n=36).

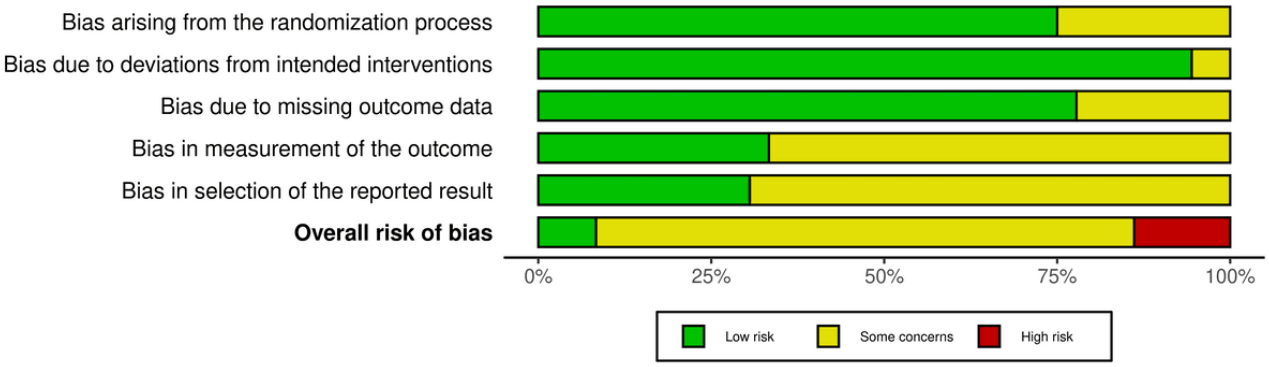

XSL•FO
RenderX

Figure 2. Visualization of the risk of bias of nonrandomized studies (n=14).

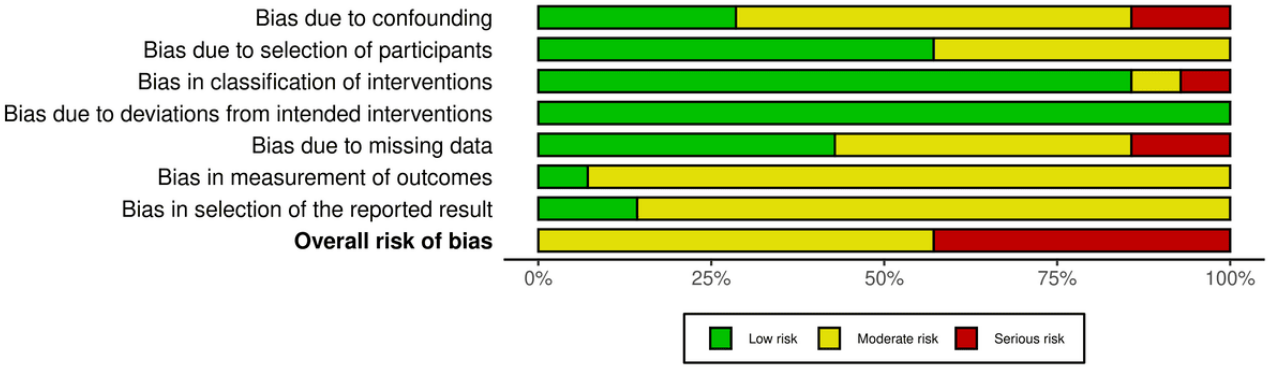

Data Synthesis

To narratively summarize the findings, IGs with differing content were treated as distinct, and identical interventions across studies were grouped [ie, 50-53]. A study was considered to have a consistent effect if it demonstrated a significant improvement in food consumption (ie, within-subjects effect) or if this change was significant compared to a control group (ie, between-subjects effect). For quantitative synthesis, standardized mean differences (Cohen *d*) were calculated. For pre- and postdesigns, Cohen *dav* was estimated, which is based on the mean difference and average SD of both sets of observations [100,101]. For studies that included both an intervention and control group, we calculated the effect size based on the mean pre- and postchange in the treatment group minus the mean pre- and postchange in the control group, divided by the pooled pretest SD [102]. This estimation has proven to be the most recommended effect size for repeated-measure designs in terms of bias and precision [103]. The variance of this effect size was calculated in R (R Foundation for Statistical Computing) with equation 25 of Morris [103], assuming a standard pre- to postintervention correlation of r=0.50. As an estimate of ρ (ie, the correlation of effect sizes) is required to develop efficient weights, several sensitivity analyses were conducted with correlation values of r=0.10 and r=0.90, which led to the standard size of r=0.5 [104]. Missing data were requested from authors and obtained for 2 papers.

A random-effects meta-analysis was performed, as we aimed to estimate the mean of a distribution of effects, and we anticipated heterogeneity in true effect sizes across studies due to variations in study characteristics [105]. Positive effect sizes reflect improvements in healthy or sustainable food intake (ie, increased plant-based consumption or reduced meat consumption), whereas negative values reflect declines. Effect sizes were dependent, as studies often contained multiple IGs compared to a common control, multiple types of food outcomes, or multiple follow-ups. Including effect sizes from the same study in a single model creates complications due to statistical dependence, which violates the assumption of independent sampling errors. To account for this dependency, robust variance estimation was performed in R with the packages *metafor*, *meta*, and *ClubSandwich* [106,107]. This method offers a robust solution to handle dependency, even when the nature of the dependence structure is unknown, by grouping effect sizes based on commonalities (ie, hierarchical clustering) and accounting for the correlation of sampling errors [107].

In total, 6 outliers and 1 intervention with inconsistent results were excluded, leaving 41 papers containing 57 interventions, with 82 effect sizes. Given that NRS represent an important part of the evidence base in digital intervention research and the lack of clear consensus on how to best integrate different types of study designs into meta-analyses [108,109], both RCTs and NRS were included in the meta-analyses. To assess the robustness of the findings, we conducted sensitivity analyses excluding all NRS. The modified method of Hartung-Knapp-Sidik-Jonkman was applied for greater accuracy [110]. Heterogeneity was evaluated using the I^2 statistic and prediction intervals (PIs). Although I^2 is a commonly reported metric, researchers have noted that it may not be the most informative indicator of heterogeneity; PIs are often preferred, as they reflect the distribution of true effects [111,112]. Funnel plots and Egger test were applied to examine small-study effects; however, the interpretation of funnel plots should be approached with caution due to their known limitations [113]. To identify the conditions under which digital interventions differ in effectiveness, we conducted a series of exploratory moderator (ie, subgroup) analyses. More specifically, subgroup analyses were conducted to examine (1) the stability of effects (ie, postintervention vs follow-up), as well as variations in effectiveness based on (2) behavioral goal orientation, (3) participant age, (4) digital medium, and (5) the presence of each distinct BCT cluster. Each moderator was examined in a separate model to avoid overfitting. Moderator analyses were also conducted for individual BCTs, but results are reported only in Multimedia Appendix 3 due to the limited number of studies per BCT and their similarity to the findings from BCT cluster analyses. More information on the meta-analysis can be found in Multimedia Appendix 4 [50,51,78,96,100-107,114].

Results

Study Selection

The database searches identified 9318 records in total. Following the removal of duplicates through the SR Accelerator software (Bond University) [115], 4765 papers were uploaded to Zotero for eligibility screening. Titles and abstracts were screened independently by the first author (KV) and 2 graduate students, followed by full-text screening. Papers that were commonly

XSL•FO
RenderX

included by all screeners received a definitive inclusion. Discrepancies were resolved through discussion with 2 researchers (BS and TS) who had not been involved in the initial screening process. The final database of included papers was checked again by 1 reviewer (BS) to ensure that each paper fulfilled eligibility criteria. The updated search conducted in September 2025 led to the identification of 2283 additional records. After deduplication, 1603 records remained and underwent the same screening procedure, resulting in 6 additional papers being included. This process resulted in 42 papers, with 10 additional studies identified through searching methods, totaling 52 papers. Figure 3 shows the study selection process and provides a more detailed summary of each screening stage, and Figure 4 provides an overview of the updated search.

XSL•FO
RenderX

Figure 3. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of study inclusion.

XSL•FO
RenderX

Figure 4. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of study inclusion for updated search (September 23, 2025).

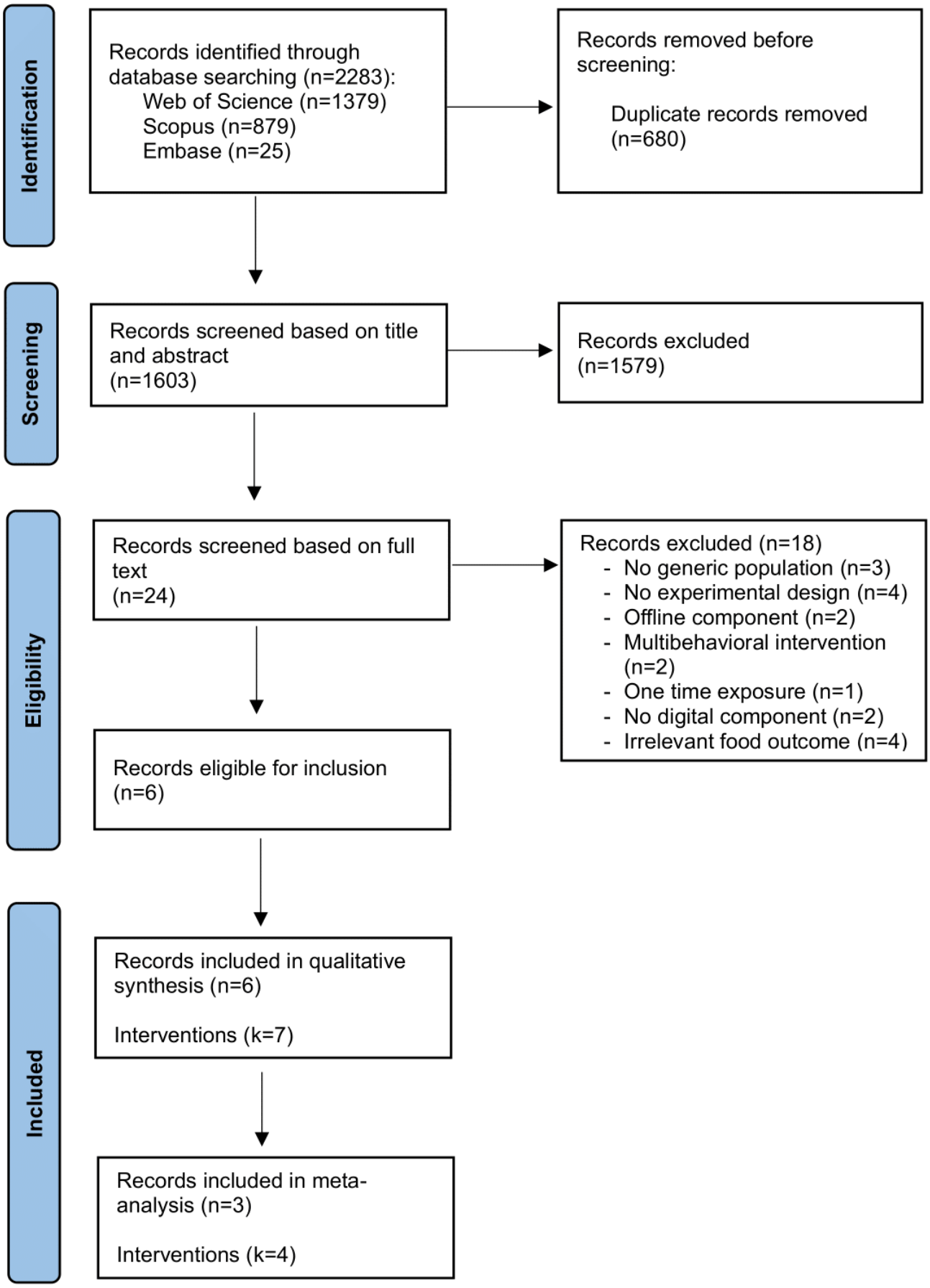

XSL•FO
RenderX

Intervention Characteristics

Of the 52 papers that met the inclusion criteria, 1 paper contained 2 separate experiments [48], while 2 pairs of papers reported findings from the same study conducted among the same sample [54-57]. Therefore, this review contained 51 unique studies, of which most were RCTs (n=38). The included studies were published across a wide range of disciplines, including public health, medical, communication, and psychological journals, reflecting the interdisciplinary nature of research on digital interventions and eating behavior. Most studies were conducted in developed countries, with the majority in the United States (n=10), followed by Italy (n=9), the United Kingdom (n=6), Australia (n=5), and other Western countries such as Denmark (n=2) and Belgium (n=2). Only 4 studies took place in a developing country, specifically in Mexico (n=2), China (n=1), and Saudi Arabia (n=1). All studies were published between 2004 and 2025 and were mostly conducted among adults (n=24), young adults (n=16), or adolescents (n=6). Some focused on children (n=1) or child-parent dyads (n=3), while 1 study did not disclose on population type or age [58]. Sample sizes varied considerably from 28 to 5062 participants at baseline and 24 to 1788 participants after the intervention. Altogether, the studies included 24,652 participants at baseline. More information on the study characteristics can be found in Multimedia Appendix 5 [48,50-99].

Socioeconomic characteristics were inconsistently reported, with only 56% (29/52) of the included studies providing relevant data. The most commonly used indicators were educational level (24/29, 83%), income (8/29, 28%), and occupation or working status (5/29, 17%). One study used food assistance as a proxy for income [54], and 3 incorporated area-based indicators [59-61]. Nearly all studies that included indicators were conducted among adults, and only 2 studies reported socioeconomic characteristics for adolescents. The majority demonstrated a predominance of participants with higher educational attainment, with some reporting up to 95% having tertiary education or International Standard Classification of Education levels 3 to 8. Similarly, among studies reporting income data, the proportion of participants in the lowest income category was small, ranging from 3.5% to 21%, with the exception of 1 study that reported an equal distribution between low- and middle-income groups [62]. While some studies controlled for socioeconomic indicators, few included them as moderators or created subgroups. Only 4 studies explicitly examined the role of SES in intervention effectiveness, of which 3 found significant effects on food intake among lower SES participants [56,57,62]. Additionally, Lim et al [63] found that income correlated with increased legume intake and decreased intention to consume animal-based foods. However, due to the limited and inconsistent reporting of SES as well as the underrepresentation of individuals from lower socioeconomic backgrounds, the planned moderator analysis based on SES was not feasible.

Across the 52 papers, a total of 69 digital interventions (k) were assessed. Intervention duration and intensity varied widely, spanning from 1 week to 6 months. While some were self-paced, most delivered content at intervals ranging from once a week to twice daily. The majority focused on promoting fruit and vegetable intake (FVI; 61%) as an indicator of healthy and sustainable eating behavior. A smaller proportion addressed the reduction of meat consumption (19%) or a broader healthy and sustainable diet (20%; eg, Mediterranean diet). Most interventions used SMS text messaging (41%) or websites (25%), followed by social media (13%), mobile apps (9%), games (7%), and emails (4%).

Use of BCTs

In total, 53 unique BCTs were identified belonging to 15 of the 16 hierarchical clusters of the BCTTv1. The most frequently used BCT clusters were “goals and planning” (65%), “natural consequences” (64%), “feedback and monitoring” (52%), and “comparison of behavior” (41%). The number of BCTs per intervention ranged from 1 to 21, with an average of 6 (SD 4.5). More information on the prevalence of each BCT can be found in Multimedia Appendix 6.

Narrative Summary of Findings by Digital Medium

SMS Text Messaging Interventions

Among the 28 SMS text messaging interventions, 22 (79%) showed significant effects. Most SMS text messaging interventions incorporated BCTs that aimed to inform participants about the consequences of unhealthy or unsustainable eating (BCT cluster 5; 86%). Some facilitated goal setting and planning (BCT cluster 1; 50%) or provided feedback and supported self-monitoring (BCT cluster 2; 25%). SMS text messaging interventions that draw on BCTs from clusters 5 and 2 appeared particularly promising, with significant effects in 79% and 71% of cases, respectively. Those based on cluster 1 were less consistently effective, with just over half (57%) yielding significant outcomes.

In total, 6 of 8 (75%) interventions targeting adults reported effects on eating behavior [59,64-66]. For instance, Carfora and Catellani [64] showed that a 2-week SMS text messaging intervention significantly increased legume intake and reduced meat consumption compared to a passive control group. The most effective messages leveraged dynamic norms that highlight an increase in people engaging in healthy and sustainable eating behavior. In contrast, 2 interventions aiming to improve adults’ FVI were unsuccessful [59,66]. Both studies might have had an insufficient intervention dose to foster behavior change, as participants only received messages 2 times per week or 5 times per month.

Among young adults (aged 18‑30 years), 16 SMS text messaging interventions were evaluated. Nearly all interventions (88%) demonstrated significant positive impacts on eating behavior, particularly in reducing meat intake. In total, 12 interventions decreased red or processed meat consumption, with effects lasting up to 8 weeks after the intervention [67-72]. Incorporating dynamic norms into SMS text messages about the environmental impact of meat seemed to enlarge the effects [71], which aligns with findings among adults [64]. Although most interventions successfully targeted meat reduction, those focusing on increasing intake of plant-based foods (k=4) showed mixed results [63,73,74].

XSL•FO
RenderX

A smaller subset of SMS text messaging interventions focused on adolescents (aged 14 - 19 years; k=4), with 75% reporting significant effects [75,76,116]. One study reported significant differences in FVI change between intervention and control, but this was primarily due to a sharp postintervention decline in the control group rather than an increase in FVI in the IG [75]. On the contrary, Pedersen et al [76] did not find significant improvements in FVI compared to the control group, likely due to low participant engagement, as positive changes were observed only among those who sent more than 50% of the SMS text messages.

Web-Based Interventions

Studies assessing the effectiveness of web-based interventions (k=16) reported limited significant results [50,51,56-62,77-83]: 7 (44%) interventions showed mixed effects due to inconsistent effects across time points or eating outcomes, and 5 found no significant effects. Almost all web-based interventions included BCTs related to goal setting and planning (BCT cluster 1; 94%) and feedback and monitoring (BCT cluster 2; 81%). Other commonly used techniques targeted knowledge (BCT cluster 4; 63%), comparison of behavior (BCT cluster 6; 56%), and social support (BCT cluster 3; 50%). Incorporating feedback and monitoring was effective in 85% of cases, and 80% of the interventions targeting knowledge demonstrated effects.

Among the 14 interventions targeting adults, the most promising were the web-based self-regulation interventions. The study of Plaete et al [50] reported significant improvements in FVI, but the feasibility study only noted an effect for fruit consumption [51]. Similarly, Frie et al [52] and Stewart et al [53] examined the effects of a self-regulation intervention on meat intake, observing reductions in meat consumption 1 week after the intervention but not at 1-month follow-up. Other interventions among adults showed limited success, as they either only had significant effects during the intervention that did not persist afterward [75] or because intake increased only for specific subgroups of the sample [62,81]. In total, 4 web-based interventions did not improve adults' eating behavior compared to the control group [56,57,60,78].

Only 2 interventions targeted the eating behavior of younger age groups. Chamberland et al [61] tested the impact of a web-based school intervention on adolescents aged 14 to 16 years, while Røed et al [80] developed a website for parents that focused on creating a healthy food environment to indirectly improve children's FVI. Both studies showed significant postintervention effects on FVI, but the effects did not persist at 3- to 6-month follow-up.

Social Media Interventions

In total, 9 interventions tested the effectiveness of a social media intervention, with 5 reporting significant effects. The most frequently incorporated BCT clusters were goals and planning (BCT cluster 1; 67%), social support (BCT cluster 3; 44%), natural consequences (BCT cluster 5; 56%), and comparison of behavior (BCT cluster 6; 56%). BCT clusters 3 and 6 appear most promising in social media intervention, as 75% and 80% of the interventions using them demonstrated significant effects.

In total, 4 interventions were conducted among young adults, 4 among adults, and 1 did not report the target group of the intervention. In 1 study of Kilb et al [48], young adults participated as dyads in an intervention, in which senders were asked to post about fruit and vegetables on Facebook, and network members were exposed to these messages. Neither senders nor network members significantly increased their FVI compared to control dyads. In contrast, the second study of this paper found that both private and public self-monitoring via social media increased FVI [48]. Similarly, Meng et al [84] showed that group-based self-tracking on a (researcher-developed) social network website led to a greater increase in FVI compared with individual self-tracking. A recent study of Hawkins et al [85] highlighted that mere exposure to healthy food content on Instagram can improve young adults' FVI.

Among adults, an intervention with support groups reported a significant increase in FVI during the intervention, but these effects were not maintained at follow-up [86]. In Ng et al [87], participants' FVI increased after completing a 4-week intervention containing recipes and videos delivered via Facebook. Carreño Enciso et al [88] tested an educational intervention delivered via Instagram or Facebook but found no significant effects on adherence to the Mediterranean diet. Weber and Nigg [58] tested an intervention containing motivational YouTube videos related to healthy eating. No changes were observed in FVI, which may be due to the low intervention dose (ie, only 6 exposures) and the requirement for participants to actively expose themselves to the intervention.

Mobile App Interventions

Mobile apps were used in 7 interventions. All incorporated feedback and monitoring techniques (BCT cluster 2), while 5 interventions also applied BCTs related to goals and planning (BCT cluster 1), and 4 addressed the consequences of unhealthy eating (BCT cluster 5). However, the effectiveness of these techniques within mobile app interventions appears limited, as only 40% - 57% of the interventions incorporating these clusters reported significant effects.

In total, 3 of 7 app-based interventions reported significant effects. Hendrie et al [89] tested an app that included different sections with recipes and feedback in a real-world setting and found an increase in vegetable intake. The PersuHabit app of Vázquez-Paz et al [90] effectively targeted parents in order to promote FVI among young children, and Liu et al [91] found that a food evaluation app significantly improved the animal-to-plant food ratio in employees' lunches. These studies were conducted among adults, while the 4 interventions showing no or mixed results targeted (pre or late) adolescents (aged 9 - 18 years) [92-94].

Interventions Using Games or Emails

While apps appear to be more effective among adults, games tend to yield better results in child populations. The FoodRateMaster game of Espinosa-Curiel et al [95] significantly increased FVI intake among children. Thompson et al [96] targeted both parents and children to improve children's FVI and found that games were effective in improving children's

XSL•FO
RenderX

diets, but only when they contained action planning. One study tested an intervention featuring 3 games among adults but found no improvement in their FVI [97]. However, it is important to note that this was one of the oldest studies included in this review; therefore, the gaming experience may have differed from those of more recent studies. Since 6 BCT clusters were applied in over 80% of the game-based interventions, it is difficult to explore the most used and promising BCTs in gaming interventions.

A limited number of studies tested email interventions (k=3). Rompotis et al [74] found that habit-based messages providing strategies to strengthen the automaticity of FVI were more effective in improving young adults' fruit intake compared to general nutrition information, regardless of whether the messages were sent via texting or email. Block et al [98] reported similar results for a worksite email intervention among adults. One study did not find significant effects on young adults' FVI [99], which could be explained by the limited intervention dose (30 days with emails every 3 days), compared to the 2 other interventions, which lasted 8 to 12 weeks [74,98]. All of the emailing interventions included BCTs related to goals and planning (BCT cluster 1).

Effectiveness of Digital Interventions

Overview

Of the 52 included studies, 41 provided sufficient data for inclusion in the meta-analysis. The results show a pooled effect size of 0.33 (95% CI 0.25 - 0.42; P<.001), indicating that digital interventions on average have a small, positive effect on healthy and sustainable eating behavior. The forest plot in Figure 5 presents the effect size for each intervention with its 95% CI. Outlier analysis identified 19 potential outliers; however, these were fairly uniformly distributed. Both the graphical representation and sensitivity analysis gave no clear indication of bias (Multimedia Appendix 7). Visual inspection of the funnel plot for small-study effects suggested limited asymmetry (Multimedia Appendix 7), which was supported by the nonsignificant value of the Egger test (P=.18). After removing high risk-of-bias studies, the analysis yielded similar results (d=0.36, 95% CI 0.27 - 0.45; P<.001). The sensitivity analysis excluding 8 effect sizes from NRS yielded results highly consistent with the primary analysis (d=0.32, 95% CI 0.23 - 0.41; P<.001), indicating that the findings are robust to study design. However, heterogeneity statistics revealed substantial heterogeneity (Q_{81}=559.40; P<.001; τ=0.16; τ^2=0.03; I^2=86%), which was supported by the 95% PI, which ranged from −0.21 to 0.87. Given the substantial heterogeneity, we cannot be confident that the positive effect is robust. True effects may vary considerably in settings, with both negative effects as well as strong positive effects being possible. Moderator analyses with categorical variables (ie, subgroup analyses) were conducted to explore between-study variance and provide a more nuanced understanding of the effectiveness of digital interventions. Forest plots for the subgroup analyses can be found in Figure 6.

XSL•FO
RenderX

Figure 5. Forest plot of standardized mean differences (Cohen *d*) for the effect of digital interventions on healthy and sustainable eating (ie, increased plant-based intake or reduced meat intake; for study details, see Multimedia Appendix 5) [48,51-56,58-60,62-71,73-77,79,80,82,84,85,87-90,92-94,96,98,99,116]. IG: intervention group; PI: prediction interval; SES: socioeconomic status; SMD: standardized mean difference.

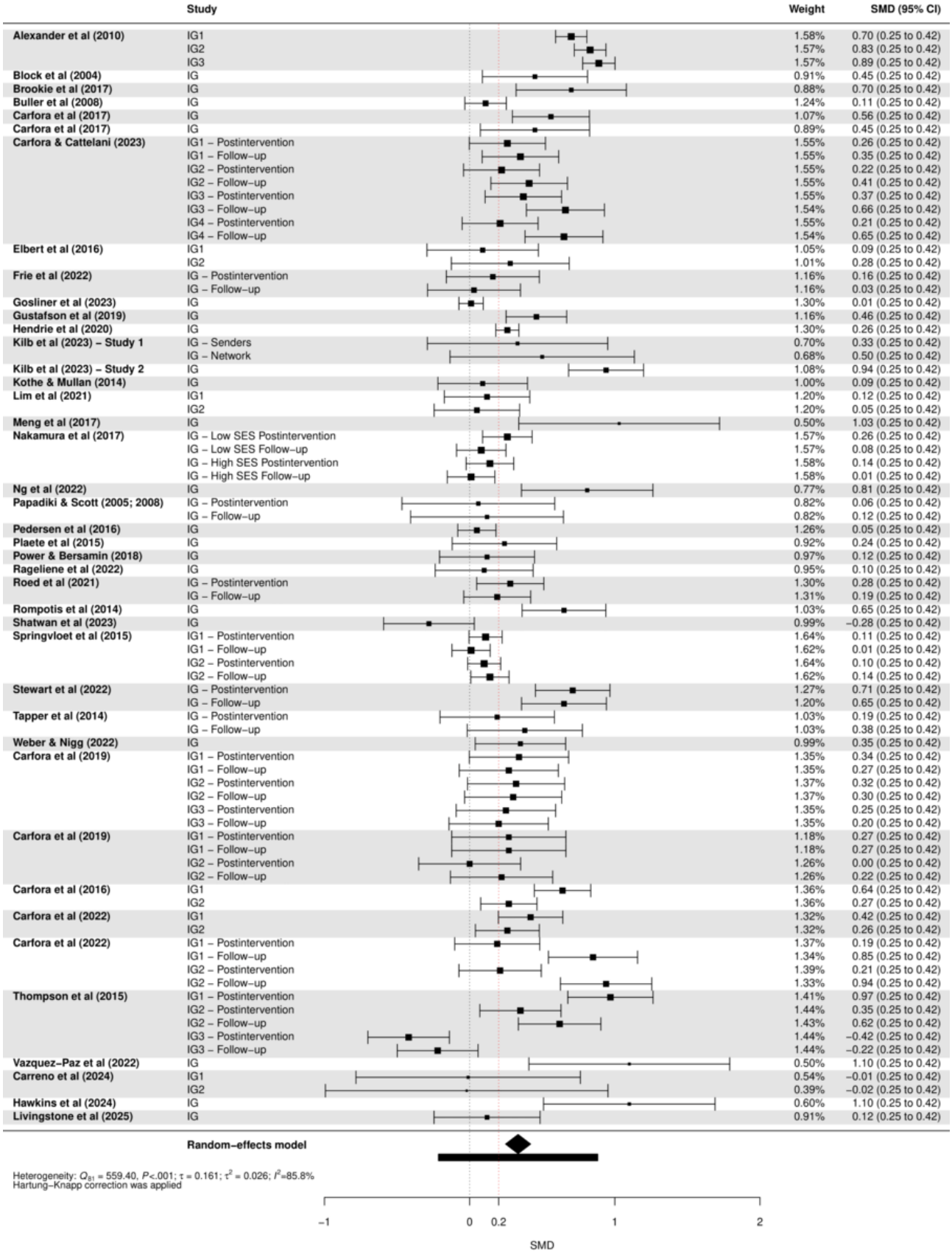

XSL•FO
RenderX

Figure 6. Forest plot of the pooled standardized mean differences (Cohen *d*) for each subgroup of the moderator analyses. SMD: standardized mean difference.

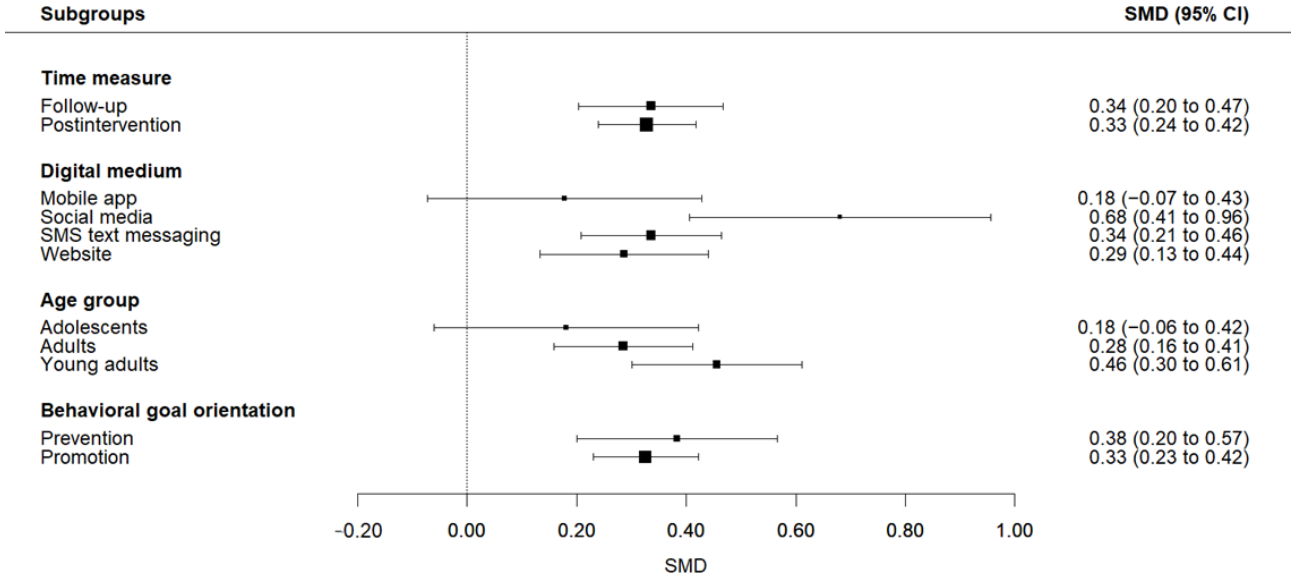

Postintervention and Follow-Up Effects

A meta-analysis with time as covariate yielded similar effect sizes for outcomes measured during or immediately after the intervention (d=0.33, 95% CI 0.24 - 0.42; P<.001; I^2=65%, 95% PI −0.17 to 0.85) and outcomes measured 1 to 6 months after the intervention (d=0.34, 95% CI 0.20 - 0.47; P<.001; I^2=77%, 95% PI −0.29 to 0.84). The test for subgroup differences shows that there are no significant differences between postintervention and follow-up effects (Q_M=0.01, df=1; P=.91). These results held when analyses were restricted to RCTs only (Multimedia Appendix 7). This suggests that, on average, the effectiveness of the digital interventions was quite stable over time, with no meaningful difference in effect between the postintervention and follow-up measures.

Behavioral Goal Orientation

The analysis with behavioral goal orientation as a moderator assessed whether promotion-focused and prevention-focused interventions differed significantly in effectiveness. Digital interventions with a promotion focus (k=47), aimed at encouraging intake of plant-based foods, yielded a statistically significant, small pooled effect size of d=0.33 (95% CI 0.23 - 0.42; P<.001). Heterogeneity was still substantial in this subgroup (I^2=70%, 95% PI −0.23 to 0.88). Prevention-focused interventions (k=12), all of which targeted reductions in red or processed meat consumption, demonstrated a slightly larger pooled effect size of d=0.38 (95% CI 0.20 - 0.57; P<.001) and lower heterogeneity (I^2=64%, 95% PI −0.08 to 0.85). The effect difference between promotion and prevention studies was not significant (Q_M=0.29, df=1; P=.59). The sensitivity analysis restricted to RCTs yielded comparable findings (Multimedia Appendix 7).

Age Group

Including participant age group as a moderator in the meta-analysis revealed that digital interventions had the largest effects among young adults (Table 1). In contrast, the smallest pooled effect size was found among adolescents; however, this finding is based on only 6 interventions. The test of moderators did not indicate a significant difference in pooled effect sizes across age groups (Q_M=4.51, df=2; P=.10). Effect estimates remained consistent when the analysis was restricted to RCTs (Multimedia Appendix 7).

Table . Meta-analysis for the effect of digital interventions with age group as moderator.

Age group	*k*	*d*	95% CI	I^2 (%)	95% PI[a]
Adults	26	0.28[b]	0.16 to 0.41	64	−0.19 to 0.75
Young adults	19	0.46[b]	0.30 to 0.61	65	−0.10 to 1.00
Adolescents	6	0.18	−0.06 to 0.42	86	−0.45 to 0.84

[a]PI: prediction interval.

[b]P<.001.

Digital Medium

The analysis with type of digital medium as moderator was significant (Q_M=8.49, df=3; P=.03), indicating that the mode of delivery modified the pooled effect. Interventions delivered via social media or SMS text messages yielded the largest effect sizes (Table 2). Despite the low number of social media interventions, pairwise comparison showed a significant subgroup difference with higher effectiveness for social media than for SMS text messaging (z=2.23; P=.03), website (z=2.56;

XSL•FO
RenderX

P=.01), and mobile app interventions (*z*=2.64; *P*=.01). The heterogeneity score for social media interventions was moderate, and the positive PI suggests that future social media interventions are likely to yield positive effects, although the lower bound is close to 0 (I^2=47%, 95% PI 0.01 - 1.31). Exclusion of NRS did not meaningfully alter the results, except that the subgroup difference between social media and SMS text messaging interventions was no longer statistically significant (*z*=1.72; *P*=.09). Additionally, the sensitivity analysis for mobile app interventions could not be conducted due to an insufficient number of studies (Multimedia Appendix 7).

Table . Meta-analysis for the effect of digital interventions with digital medium as moderator.

Digital medium	*k*	*d*	95% CI	I^2 (%)	95% PI[a]
SMS text messaging	24	0.34[b]	0.21 to 0.47	62	−0.05 to 0.73
Website	13	0.28[b]	0.13 to 0.43	77	−0.25 to 0.80
Social media	8	0.65[b]	0.41 to 0.90	47	0.01 to 1.31
Mobile app	5	0.18	−0.07 to 0.43	72	−0.53 to 0.93

[a]PI: prediction interval.
[b]*P*<.001.

BCT Clusters

Meta-analyses with the presence of each BCT cluster found the largest standardized mean difference for interventions incorporating prompts or cues (BCT cluster 7), with a pooled effect size of *d*=0.62 (95% CI 0.33 - 0.91; *P*<.001). The test of moderators indicated that interventions including this BCT cluster were significantly associated with a greater improvement in eating behavior than those without this cluster (*P*=.04). Interventions including social support (cluster 3) and comparison of behavior (cluster 6) demonstrated higher pooled effect sizes (*d*=0.39) compared to those that did not include these BCTs (*d*=0.31 and *d*=0.29, respectively). Although these differences were not statistically significant (*P*=.40 and *P*=.17, respectively), the numerically higher effect sizes suggest that these BCTs may still contribute positively to intervention outcomes. The lack of significance could be attributed to the limited number of studies incorporating these BCT clusters, underscoring the need for further research.

Several BCT clusters showed similar effect sizes regardless of whether they were present in interventions or not. For instance, feedback and monitoring (cluster 2), repetition and substitution (cluster 8), and antecedents (cluster 12) had comparable pooled effect sizes in both subgroups, suggesting no clear added value of including these BCTs (Table 3). Interventions that incorporated the BCT clusters “shaping knowledge” (cluster 4) and “natural consequences” (cluster 5) showed smaller pooled effect sizes compared to those in which these clusters were absent; however, these differences were not statistically significant (*P*=.07 and *P*=.35, respectively). Overall, restricting these moderator analyses to RCTs produced similar results (Multimedia Appendix 7). Moderator analyses were also conducted for individual BCTs, which generated similar findings (Multimedia Appendix 3).

XSL•FO
RenderX

Table . Meta-analysis for the effect of digital interventions with behavior change technique (BCT) cluster as moderator.

BCT cluster	BCT cluster present			BCT cluster absent			Test of moderators
	k	*d*	95% CI	*k*	*d*	95% CI	*P* value
1. Goals and planning	37	0.31[a]	0.19 to 0.43	20	0.37[a]	0.24 to 0.49	.56
2. Feedback and monitoring	32	0.33[a]	0.20 to 0.47	25	0.33[a]	0.22 to 0.44	>.99
3. Social support	16	0.39[a]	0.21 to 0.56	41	0.31[a]	0.22 to 0.40	.40
4. Shaping knowledge	27	0.25[b]	0.10 to 0.39	30	0.40[a]	0.29 to 0.52	.07
5. Natural consequences	39	0.30[a]	0.22 to 0.39	18	0.39[b]	0.18 to 0.59	.35
6. Comparison of behavior	27	0.39[a]	0.27 to 0.51	30	0.29[a]	0.18 to 0.39	.17
7. Associations	6	0.58[a]	−0.10 to 1.25	51	0.30[a]	0.23 to 0.38	.04
8. Repetition and substitution	12	0.30[b]	0.15 to 0.45	45	0.34[a]	0.24 to 0.44	.64
9. Comparison of outcomes	9	0.38	−0.06 to 0.83	48	0.32[a]	0.24 to 0.41	.60
10. Reward and threat	13	0.36[b]	0.13 to 0.60	44	0.32[a]	0.23 to 0.42	.72
12. Antecedents	6	0.33[b]	0.11 to 0.55	51	0.33[a]	0.24 to 0.42	.97
13. Identity	10	0.19[c]	0.07 to 0.30	47	0.36[a]	0.27 to 0.46	.09
14. Scheduled consequences	4	0.29	−0.74 to 1.32	53	0.34[a]	0.25 to 0.42	.78
15. Self-belief	7	0.29	−0.09 to 0.67	50	0.34[a]	0.24 to 0.43	.75

[a]$P<.001$.

[b]$P<.01$.

[c]$P<.05$.

Discussion

Principal Findings

While the meta-analysis found a moderate and statistically significant overall effect of digital interventions on healthy and sustainable food consumption, the substantial heterogeneity observed across studies suggests that these interventions do not have a consistent effect across all populations and contexts. Additionally, included studies mostly had a moderate risk of bias, and together with the inconsistency in effects, this may reduce the certainty of the evidence regarding the overall effectiveness of digital interventions. Both the narrative review and the moderator analyses showed that the effectiveness of digital interventions varied by multiple characteristics and settings.

Regarding the food outcomes, interventions preventing animal-based food consumption and those promoting plant-based food intake both yielded a significant, moderate effect. While interventions with a prevention focus yielded a slightly larger pooled effect size, moderator analyses showed no significant difference with interventions promoting plant-based food intake. This contrasts with prior research, suggesting that a promotion focus is more effective in encouraging healthy dietary patterns [26,117]. However, this might be explained by a difference in the food consumption studied. For example, the studies reviewed by Zlatevska et al [117] primarily targeted discretionary items such as sugary drinks and snacks, not meat consumption. Only 1 of the included interventions in this review combined a prevention and promotion strategy and found sustained effects at follow-up [64], highlighting a need for future research to test more integrative approaches that align closely to the EAT-Lancet diet.

Both the narrative synthesis and moderator analyses show that the effectiveness of digital interventions varies by modality. SMS text messaging interventions had a moderate effect on healthy and sustainable food intake, with over 75% demonstrating significant results in the narrative review. Despite the small number of social media interventions, they had the strongest effects on eating behavior, being significantly associated with larger effect sizes than other modes of delivery. While prior research associates social media use with unhealthy eating habits [118,119], these findings suggest that social media

XSL•FO
RenderX

can also be used to positively influence dietary behavior. Since many consumers engage with food content on social media, platforms such as Instagram or Facebook hold great potential to promote healthier and more sustainable eating [120,121]. Social media platforms present unique environments in which users can form large social networks, allowing them to seek information about others' behavior and receive positive reinforcements for their own behavior. Although the evidence base remains limited, social media interventions may promote stronger behavior change for several reasons: they expose users to peers' behaviors (social modeling), leverage participants' existing engagement with the platform to sustain intervention exposure and retention, and offer interactive features that facilitate social support [122,123]. Notably, none of the social media interventions targeted adolescents, despite this being a key demographic due to their high engagement with social media and high exposure to unhealthy food marketing [124,125].

These findings suggest that interventions delivered through accessible and familiar platforms (ie, SMS text messaging or social media) tend to be more effective than interventions requiring more intentional user engagement (ie, stand-alone websites or apps). This difference also links to the amount of agentic demand (ie, the degree to which participants are required to engage with the content in order to achieve the intended outcome) [126,127], though this requires further empirical testing. The common academic practice of developing digital interventions via research-created platforms, often used only briefly and actively throughout the study period, implies a misalignment with real-world digital behavior [35]. To enhance ecological validity and long-term behavior change, future studies could focus on accessible interventions using existing, familiar platforms [126,128]. While social media interventions were significantly more effective than interventions via other platforms, they also present specific challenges for researchers. As opposed to more controlled intervention environments such as apps or websites, social media constitute open and dynamic spaces that are saturated with unhealthy food marketing and health misinformation [129]. This may dilute or even counteract intervention effects and can lead to nutrition confusion [130].

A wide range of BCTs was incorporated across the interventions, and the choice of BCT varied depending on the type of digital medium used. BCTs targeting goals and planning were most frequently applied across all interventions, particularly in web-based interventions. Knowledge-related techniques, such as providing instructions, were frequently applied, especially in web-based interventions or games. Communicating the consequences of unhealthy eating behavior was also a commonly used cluster of techniques, mainly in SMS text messaging interventions. Yet, these information-based BCTs were not the most effective ones. Although definitive conclusions on the most effective BCT cluster cannot be made due to the limited number of studies incorporating certain clusters, moderator analyses [1] indicated that interventions incorporating prompts or cues (BCT cluster 7) were significantly more effective than those that did not, yielding a large effect size. Additionally, interventions targeting social mechanisms through social support (cluster 3) or comparison of behavior (cluster 6) demonstrated larger effects, though not significantly larger than those not including these BCTs. The narrative synthesis further suggests that social norm communication is an effective strategy, as several studies with significant effects incorporated descriptive norms to influence behavior. These findings suggest that future interventions may benefit from shifting emphasis away from purely informational strategies (eg, raising awareness or increasing knowledge) and instead testing techniques that leverage social influence, peer dynamics, and cues. Previous research on health interventions also suggests that providing social support strengthens the impact of digital interventions [131-133], and that norm communication can gradually reshape individuals' perceptions of others' behavior, prompting them to align with these evolving norms [134-136].

Finally, while age-related differences in the effect size of the interventions were not significant, moderator analyses yielded the strongest effect size for young adults. Due to the limited number of studies targeting adolescents, conclusions for this age group cannot be drawn. Regarding SES, most studies included SES indicators only as control variables rather than focusing on targeted recruitment, resulting in the underrepresentation of lower SES individuals. This limited our ability to conduct subgroup analyses for this population. This lack of research is concerning, given that food intake is socially structured, with unhealthy and unsustainable diets being more prevalent among lower SES individuals [137-139]. While financial constraints and educational disparities contribute to these patterns, modifiable psychosocial factors (eg, attitudes, literacy, desired identity, or social norms) also play a critical role [140-142]. Despite potential challenges related to digital access and literacy, it is crucial to reach lower SES populations in the current digital era to help mitigate health inequalities related to dietary behavior, especially adolescents, given that they are highly active on digital platforms. Leveraging platforms such as social media may help embed interventions within their daily routines and reduce barriers to participation. Future research should conduct both targeted testing of interventions within lower SES groups and exploration of SES moderation effects to understand intervention effectiveness across socioeconomic groups.

This review reveals some major gaps in the literature. First, individuals from lower socioeconomic backgrounds remain underrepresented in research testing digital interventions. None of the included studies used subjective measures of SES (ie, perceptions of own social status) [2] despite evidence that perceived SES is a strong predictor of health (behavior) even after controlling for objective indicators [143-145]. Future research should therefore not only systematically measure SES but also incorporate subjective indicators more frequently [146]. Second, while the included interventions targeted various age groups, few studies specifically focused on adolescents. The handful of studies that included adolescents as participant group primarily tested web-based or mobile app interventions, which were among the least effective digital modalities. Both the lack of research as well as the prominent focus on websites or apps may explain the smaller effect size for this demographic. Future research should prioritize low-agentic interventions for adolescents of different socioeconomic backgrounds, particularly through social media, as adolescents not only increasingly use

XSL•FO
RenderX

these platforms, but research also shows that exposure to social media food content influences their eating behavior [118,124]. Finally, most studies focused on specific food groups, particularly fruits and vegetables, framing them as key components of a healthy diet. These findings are similar to previous systematic reviews on interventions aiming to improve food consumption [33,147]. While fruits and vegetables are indeed a critical part of a healthy and sustainable diet, interventions should also highlight other important components, such as legumes and whole grain products. Furthermore, rather than solely focusing on reducing meat consumption, interventions can promote replacing meat with whole-food plant-based alternatives, as substituting certain products is often more achievable than complete elimination [18,64].

Implications

By identifying the conditions in which digital interventions are most effective for encouraging plant-based eating, this study offers timely guidance for practitioners in a rapidly evolving digital landscape. The results of this review suggest that digital interventions can effectively be implemented to support people in transitioning to healthier and more sustainable eating behavior. Given the variability in effectiveness following intervention characteristics, health care providers and practitioners should consider selecting digital platforms that best fit the preferences, literacy, and lifestyle of the target group. Social media may be particularly useful for delivering low-threshold interventions in community settings, targeting mechanisms of social influence for behavior change. Practitioners are encouraged to adopt evidence-based BCTs that leverage social influence (ie, social support and demonstration of behavior) and thus shift emphasis from purely informational strategies (eg, raising awareness or increasing knowledge). Collaboration between behavioral scientists and health care professionals can help ensure that interventions are evidence-based and aligned with participants' needs. Special attention is needed for the participation of underrepresented groups, more specifically, adolescents and individuals with lower SES. Future research should use specific strategies to recruit and engage these vulnerable populations, for example, through school-based programs or partnerships with community organizations. Moreover, it is important to ensure that intervention materials are accessible and inclusive to help reduce disparities in participation and retention. One way to ensure this is by actively involving members of the vulnerable group or trusted intermediaries in the design of the intervention [148].

Limitations

Despite using extensive search strategies, relevant research may have been overlooked if published in languages other than English or if they were inaccessible through the selected databases. The majority of included studies demonstrated moderate risk of bias, with measurement of outcome bias being the most prevalent concern due to reliance on self-reported dietary measures rather than objective assessments. This may lead to overestimation of intervention effects, as self-report measures are susceptible to social desirability bias and recall error. Future research should prioritize objective outcome measures to strengthen evidence for real-world applicability. In total, 11 studies could not be included in the meta-analysis due to insufficient data for effect size calculation. Additionally, the inconsistency in measuring SES and the scarcity of research involving participants from lower SES backgrounds hindered our ability to assess the differential effects of digital interventions across SES groups. Moreover, the small number of studies contributing data to certain moderators may have limited the statistical power to detect potential moderation effects [149] and restricted our ability to assess their combined influence. Ideally, a sufficiently large number of studies would be available to allow the modeling of joint effects, rather than examining each moderator separately. Furthermore, most evidence was derived from studies with a moderate risk of bias, and substantial heterogeneity was present. Although the subgroup analyses explained some of the observed variation, these factors may still affect the overall certainty of the evidence. We did not perform a formal Grading Recommendations Assessment, Development, and Evaluation assessment to systematically rate the certainty of the evidence, which limits our ability to evaluate the confidence in our findings [150].

Another potential limitation is the inclusion of both RCTs and NRS in the meta-analysis. The appropriateness of combining different study designs remains debated in the methodological literature, with studies having inconsistent conclusions about the risks of pooling RCTs and NRS (eg, achieving statistical significance only after inclusion of NRS) [109,151,152]. We retained NRS in our primary analyses for 2 reasons. First, sensitivity analyses excluding NRS demonstrated that our findings were robust, with effect estimates, CIs, and heterogeneity statistics remaining consistent. Second, NRS are increasingly recognized as valuable sources of evidence, also within digital intervention research, as ecological validity and real-world application are important criteria. In this context, NRS can provide complementary evidence alongside RCTs [153,154]. However, the pooled estimates should therefore be interpreted with awareness of some methodological considerations. RCTs and NRS are susceptible to different types of bias [151,152]. RCTs face risks related to randomization, allocation concealment, and blinding, while NRS are prone to confounding and selection bias, which is reflected in the different risk-of-bias assessment tools used for each design. The pooled estimates, therefore, represent weighted averages that incorporate these different types of bias, which may complicate direct translation of the results to practice [151,152]. Moreover, these study designs provide different types of evidence: RCTs answer questions of effectiveness under controlled conditions, while NRS can, for instance, address generalizability or real-world performance [151,153]. Consequently, the primary findings should be interpreted as reflecting the overall evidence base for digital interventions, rather than as a pure estimate of intervention effectiveness derived solely from controlled trials. Importantly, the consistency of the findings in the sensitivity analyses restricted to RCTs supports the robustness of the study's main conclusions regarding intervention effectiveness.

Other important limitations are with regard to the BCTs. Categorizing intervention components into BCTs based on study descriptions and protocols proved challenging, as studies used varying terminology and levels of detail on the intervention

XSL•FO
RenderX

content. Similar issues have been highlighted in previous systematic reviews [36,146]. Therefore, despite the use of a coding manual and following a training in BCT taxonomy, the presence of some techniques may have been overlooked. To enhance the replicability of interventions and accuracy of coding BCT presence for meta-analyses, future research should describe intervention content in greater detail and more systematically by following taxonomies or reporting guidelines [24,155]. It is also essential to recognize that BCTs are not exhaustive; they possess a certain level of superficiality. While BCTs can be implemented in various ways, the content was generalized to a certain type of BCT, potentially masking variations in their implementation. For instance, informing participants about the components of a healthy diet or providing personalized recipes both fall under the same BCT category (4.1 Instruction on how to perform the behavior) but represent distinct approaches that may yield different effects. Moreover, some characteristics of persuasive communication are overlooked by focusing on BCTs (eg, framing of information).

Conclusions

This review provides a comprehensive overview of digital interventions, suggesting that digital interventions can effectively improve eating behavior, though their success varies by intervention design and population targeted. Social media emerge as particularly promising, likely due to their unique social and interactive features. Importantly, the evidence base mainly consists of studies with a moderate risk of bias, highlighting the need for more high-quality studies to confirm current results. Moreover, the meta-analytic results have a broad PI, indicating that while the average effect is positive, individual interventions may range from highly effective to potentially ineffective, depending on context and design. To our knowledge, it is one of the first reviews to systematically code the characteristics of digital interventions, including their mode of delivery (ie, digital medium), content (ie, BCTs), behavioral goal orientation (prevention vs promotion), and targeted demographic (ie, age and SES), and to link these with the intervention effect size. Despite our aim to explore effects specifically among low SES groups, the limited available research restricted our ability to conduct subgroup analyses for this population. Our findings offer valuable insights for practitioners and researchers interested in leveraging digital media for behavior change by providing an evidence base on the contexts and types of digital interventions that most effectively promote plant-based eating.

Acknowledgments

The authors would like to thank Emma Sageot and Steffi Bellekens for their support with the title, abstract, and full-text screening. The authors would also like to express their appreciation to Veerle Tuerlinckx, who helped optimize the search string.

Funding

This work was supported by the FEAST project funded by the European Research Executive Agency (grant 101060536) and the HashTagToFork project funded by Internal Funds KU Leuven (grant C2M/23/007). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Data Availability

All data generated or analyzed during this study are included in Multimedia Appendix 5.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search string.
[DOCX File, 22 KB - jmir_v28i1e80821_app1.docx]

Multimedia Appendix 2
Risk of bias.
[PDF File, 174 KB - jmir_v28i1e80821_app2.pdf]

Multimedia Appendix 3
Moderator analyses—behavior change techniques.
[DOCX File, 27 KB - jmir_v28i1e80821_app3.docx]

Multimedia Appendix 4
Calculation of effect sizes and meta-analysis.
[DOCX File, 21 KB - jmir_v28i1e80821_app4.docx]

XSL•FO
RenderX

Multimedia Appendix 5
Study characteristics.
[XLSX File, 27 KB - jmir_v28i1e80821_app5.xlsx]

Multimedia Appendix 6
Prevalence of behavior change techniques.
[DOCX File, 24 KB - jmir_v28i1e80821_app6.docx]

Multimedia Appendix 7
Sensitivity analyses.
[PDF File, 215 KB - jmir_v28i1e80821_app7.pdf]

Checklist 1
PRISMA and PRISMA-S checklist.
[DOCX File, 280 KB - jmir_v28i1e80821_app8.docx]

References

1. Sun X, Yon DK, Nguyen TT, et al. Dietary and other lifestyle factors and their influence on non-communicable diseases in the Western Pacific region. Lancet Reg Health West Pac 2024 Feb;43:100842. [doi: 10.1016/j.lanwpc.2023.100842] [Medline: 38456094]
2. Willett W, Rockström J, Loken B, et al. Food in the anthropocene: the EAT–Lancet Commission on healthy diets from sustainable food systems. Lancet 2019 Feb;393(10170):447-492. [doi: 10.1016/S0140-6736(18)31788-4]
3. Clark M, Tilman D. Comparative analysis of environmental impacts of agricultural production systems, agricultural input efficiency, and food choice. Environ Res Lett 2017 Jun 1;12(6):064016. [doi: 10.1088/1748-9326/aa6cd5] [Medline: 37829169]
4. Bui LP, Pham TT, Wang F, et al. Planetary Health Diet Index and risk of total and cause-specific mortality in three prospective cohorts. Am J Clin Nutr 2024 Jul;120(1):80-91. [doi: 10.1016/j.ajcnut.2024.03.019] [Medline: 38960579]
5. Karavasiloglou N, Thompson AS, Pestoni G, et al. Adherence to the EAT-Lancet reference diet is associated with a reduced risk of incident cancer and all-cause mortality in UK adults. One Earth 2023 Dec 15;6(12):1726-1734. [doi: 10.1016/j.oneear.2023.11.002] [Medline: 38130482]
6. Knuppel A, Papier K, Key TJ, Travis RC. EAT-Lancet score and major health outcomes: the EPIC-Oxford study. Lancet 2019 Jul;394(10194):213-214. [doi: 10.1016/S0140-6736(19)31236-X]
7. Wallace TC, Bailey RL, Blumberg JB, et al. Fruits, vegetables, and health: a comprehensive narrative, umbrella review of the science and recommendations for enhanced public policy to improve intake. Crit Rev Food Sci Nutr 2020;60(13):2174-2211. [doi: 10.1080/10408398.2019.1632258] [Medline: 31267783]
8. Westhoek H, Lesschen JP, Rood T, et al. Food choices, health and environment: effects of cutting Europe's meat and dairy intake. Glob Environ Change 2014 May;26:196-205. [doi: 10.1016/j.gloenvcha.2014.02.004]
9. Yanni AE, Iakovidi S, Vasilikopoulou E, Karathanos VT. Legumes: a vehicle for transition to sustainability. Nutrients 2023 Dec 27;16(1):98. [doi: 10.3390/nu16010098] [Medline: 38201928]
10. Daily consumption of fruit and vegetables by sex, age and educational attainment level. Eurostat. 2022 Apr 4. URL: https://ec.europa.eu/eurostat/databrowser/view/HLTH_EHIS_FV3E__custom_1588514/default/table?lang=en [accessed 2025-12-17]
11. Gibbs J, Cappuccio FP. Plant-based dietary patterns for human and planetary health. Nutrients 2022 Apr 13;14(8):1614. [doi: 10.3390/nu14081614] [Medline: 35458176]
12. Hughes J, Pearson E, Grafenauer S. Legumes—a comprehensive exploration of global food-based dietary guidelines and consumption. Nutrients 2022 Jul 27;14(15):3080. [doi: 10.3390/nu14153080] [Medline: 35956258]
13. Vranken L, Avermaete T, Petalios D, Mathijs E. Curbing global meat consumption: emerging evidence of a second nutrition transition. Environ Sci Policy 2014 May;39:95-106. [doi: 10.1016/j.envsci.2014.02.009]
14. Evans WD, Abroms LC, Broniatowski D, et al. Digital media for behavior change: review of an emerging field of study. Int J Environ Res Public Health 2022 Jul 26;19(15):9129. [doi: 10.3390/ijerph19159129] [Medline: 35897494]
15. Folkvord F. The promotion of healthy foods: a review of the literature and theoretical framework. In: The Psychology of Food Marketing and Overeating: Routledge; 2019. [doi: 10.4324/9780429274404-8]
16. Beck Silva KB, Miranda Pereira E, Santana MD, Costa PRF, Silva RDC. Effects of computer-based interventions on food consumption and anthropometric parameters of adolescents: a systematic review and metanalysis. Crit Rev Food Sci Nutr 2024;64(6):1617-1631. [doi: 10.1080/10408398.2022.2118227] [Medline: 36062829]
17. Chen Y, Perez-Cueto FJA, Giboreau A, Mavridis I, Hartwell H. The promotion of eating behaviour change through digital interventions. IJERPH 2020;17(20):7488. [doi: 10.3390/ijerph17207488]

XSL•FO
RenderX

18. Hsu MSH, Rouf A, Allman-Farinelli M. Effectiveness and behavioral mechanisms of social media interventions for positive nutrition behaviors in adolescents: a systematic review. J Adolesc Health 2018 Nov;63(5):531-545. [doi: 10.1016/j.jadohealth.2018.06.009] [Medline: 30197198]
19. Livingstone KM, Rawstorn JC, Partridge SR, et al. Digital behaviour change interventions to increase vegetable intake in adults: a systematic review. Int J Behav Nutr Phys Act 2023 Mar 27;20(1):36. [doi: 10.1186/s12966-023-01439-9] [Medline: 36973716]
20. Al-Dhahir I, Reijnders T, Faber JS, et al. The barriers and facilitators of eHealth-based lifestyle intervention programs for people with a low socioeconomic status: scoping review. J Med Internet Res 2022 Aug 24;24(8):e34229. [doi: 10.2196/34229] [Medline: 36001380]
21. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010 Feb 17;12(1):e4. [doi: 10.2196/jmir.1376] [Medline: 20164043]
22. Michie S, West R, Sheals K, Godinho CA. Evaluating the effectiveness of behavior change techniques in health-related behavior: a scoping review of methods used. Transl Behav Med 2018 Mar 1;8(2):212-224. [doi: 10.1093/tbm/ibx019] [Medline: 29381786]
23. Carey RN, Connell LE, Johnston M, et al. Behavior change techniques and their mechanisms of action: a synthesis of links described in published intervention literature. Ann Behav Med 2018;10. [doi: 10.1093/abm/kay078]
24. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
25. Michie S, Abraham C, Whittington C, McAteer J, Gupta S. Effective techniques in healthy eating and physical activity interventions: a meta-regression. Health Psychol 2009 Nov;28(6):690-701. [doi: 10.1037/a0016136] [Medline: 19916637]
26. Kachwaha S, Kim SS, Das JK, et al. Behavior change interventions to address unhealthy food consumption: a scoping review. Curr Dev Nutr 2024 Mar;8(3):102104. [doi: 10.1016/j.cdnut.2024.102104] [Medline: 38482184]
27. Hedin B, Katzeff C, Eriksson E, Pargman D. A systematic review of digital behaviour change interventions for more sustainable food consumption. Sustainability 2019;11(9):2638. [doi: 10.3390/su11092638]
28. Klassen KM, Douglass CH, Brennan L, Truby H, Lim MSC. Social media use for nutrition outcomes in young adults: a mixed-methods systematic review. Int J Behav Nutr Phys Act 2018 Jul 24;15(1):70. [doi: 10.1186/s12966-018-0696-y] [Medline: 30041699]
29. Prowse R, Carsley S. Digital interventions to promote healthy eating in children: umbrella review. JMIR Pediatr Parent 2021 Nov 25;4(4):e30160. [doi: 10.2196/30160] [Medline: 34842561]
30. Vargas-Garcia EJ, Evans CEL, Prestwich A, Sykes-Muskett BJ, Hooson J, Cade JE. Interventions to reduce consumption of sugar-sweetened beverages or increase water intake: evidence from a systematic review and meta-analysis. Obes Rev 2017 Nov;18(11):1350-1363. [doi: 10.1111/obr.12580] [Medline: 28721697]
31. Wolfenden L, Barnes C, Lane C, et al. Consolidating evidence on the effectiveness of interventions promoting fruit and vegetable consumption: an umbrella review. Int J Behav Nutr Phys Act 2021 Jan 11;18(1):11. [doi: 10.1186/s12966-020-01046-y] [Medline: 33430879]
32. Ashton LM, Sharkey T, Whatnall MC, et al. Effectiveness of interventions and behaviour change techniques for improving dietary intake in young adults: a systematic review and meta-analysis of RCTs. Nutrients 2019 Apr 11;11(4):825. [doi: 10.3390/nu11040825] [Medline: 30979065]
33. Ghammachi N, Dharmayani PNA, Mihrshahi S, Ronto R. Investigating web-based nutrition education interventions for promoting sustainable and healthy diets in young adults: a systematic literature review. Int J Environ Res Public Health 2022 Feb 1;19(3):1691. [doi: 10.3390/ijerph19031691] [Medline: 35162714]
34. Curtin E, Green R, Brown KA, et al. The effectiveness of mobile app-based interventions in facilitating behaviour change towards healthier and more sustainable diets: a systematic review and meta-analysis. Int J Behav Nutr Phys Act 2025 Sep 30;22(1):122. [doi: 10.1186/s12966-025-01823-7] [Medline: 41035009]
35. Hingle M, Patrick H. There are thousands of apps for that: navigating mobile technology for nutrition education and behavior. J Nutr Educ Behav 2016 Mar;48(3):213-218. [doi: 10.1016/j.jneb.2015.12.009] [Medline: 26965099]
36. Ronteltap A, Bukman AJ, Nagelhout GE, et al. Digital health interventions to improve eating behaviour of people with a lower socioeconomic position: a scoping review of behaviour change techniques. BMC Nutr 2022 Dec 8;8(1):145. [doi: 10.1186/s40795-022-00635-3] [Medline: 36482430]
37. Tang H, Spreckley M, van Sluijs E, Ahern AL, Smith AD. The impact of social media interventions on eating behaviours and diet in adolescents and young adults: a mixed methods systematic review protocol. BMJ Open 2024 Apr;14(4):e083465. [doi: 10.1136/bmjopen-2023-083465]
38. Michie S, Jochelson K, Markham WA, Bridle C. Low-income groups and behaviour change interventions: a review of intervention content, effectiveness and theoretical frameworks. J Epidemiol Community Health 2009 Aug;63(8):610-622. [doi: 10.1136/jech.2008.078725] [Medline: 19386612]

XSL•FO
RenderX

39. Karimi N, Opie R, Crawford D, O'Connell S, Ball K. Digitally delivered interventions to improve nutrition behaviors among resource-poor and ethnic minority groups with type 2 diabetes: systematic review. J Med Internet Res 2023;26:e42595. [doi: 10.2196/42595] [Medline: 37490331]
40. Page MJ, Moher D, Bossuyt PM, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021 Mar 29;372:n160. [doi: 10.1136/bmj.n160] [Medline: 33781993]
41. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
42. Dykes J, Brunner EJ, Martikainen PT, Wardle J. Socioeconomic gradient in body size and obesity among women: the role of dietary restraint, disinhibition and hunger in the Whitehall II study. Int J Obes 2004 Feb;28(2):262-268. [doi: 10.1038/sj.ijo.0802523]
43. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Dec;6(1). [doi: 10.1186/1748-5908-6-42]
44. Cochrane handbook for systematic reviews of interventions (current version). Cochrane. URL: https://training.cochrane.org/handbook/current [accessed 2025-12-17]
45. Eldridge S, Campbell MK, Campbell MJ, et al. A revised Cochrane risk of bias tool for randomized trials (RoB 2). Additional considerations for cluster-randomized trials (RoB 2 CRT). 2021 Mar 18. URL: https://drive.google.com/file/d/1yDQtDkrp68_8kJiIUdbongK99sx7RFI-/view [accessed 2025-12-27]
46. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016 Oct 12;355:i4919. [doi: 10.1136/bmj.i4919] [Medline: 27733354]
47. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
48. Kilb M, Giese H, Mata J. How eating-related social media postings influence healthy eating in senders and network members: two field experiments with intensive longitudinal data. Appetite 2023 Mar 1;182:106430. [doi: 10.1016/j.appet.2022.106430] [Medline: 36549365]
49. McGuinness LA, Higgins JPT. Risk-of-bias VISualization (robvis): an R package and Shiny web app for visualizing risk-of-bias assessments. Res Synth Methods 2021 Jan;12(1):55-61. [doi: 10.1002/jrsm.1411] [Medline: 32336025]
50. Plaete J, Crombez G, Van der Mispel C, Verloigne M, Van Stappen V, De Bourdeaudhuij I. Effect of the web-based intervention MyPlan 1.0 on self-reported fruit and vegetable intake in adults who visit general practice: a quasi-experimental trial. J Med Internet Res 2016 Feb 29;18(2):e47. [doi: 10.2196/jmir.5252] [Medline: 26929095]
51. Plaete J, De Bourdeaudhuij I, Verloigne M, Crombez G. Acceptability, feasibility and effectiveness of an eHealth behaviour intervention using self-regulation: "MyPlan". Patient Educ Couns 2015 Jul 26;98(12):1617-1624. [doi: 10.1016/j.pec.2015.07.014] [Medline: 26277282]
52. Frie K, Stewart C, Piernas C, Cook B, Jebb SA. Effectiveness of an Online Programme to Tackle Individual's Meat Intake through SElf-regulation (OPTIMISE): a randomised controlled trial. Eur J Nutr 2022 Aug;61(5):2615-2626. [doi: 10.1007/s00394-022-02828-9] [Medline: 35244757]
53. Stewart C, Piernas C, Frie K, Cook B, Jebb SA. Evaluation of OPTIMISE (Online Programme to Tackle Individual's Meat Intake Through Self-regulation): cohort study. J Med Internet Res 2022 Dec 12;24(12):e37389. [doi: 10.2196/37389] [Medline: 36508245]
54. Papadaki A, Scott JA. The Mediterranean eating in Scotland experience project: evaluation of an internet-based intervention promoting the Mediterranean diet. Br J Nutr 2005 Aug;94(2):290-298. [doi: 10.1079/bjn20051476] [Medline: 16115365]
55. Papadaki A, Scott JA. Follow-up of a web-based tailored intervention promoting the Mediterranean diet in Scotland. Patient Educ Couns 2008 Nov;73(2):256-263. [doi: 10.1016/j.pec.2008.05.030] [Medline: 18640000]
56. Springvloet L, Lechner L, de Vries H, Candel MJJM, Oenema A. Short- and medium-term efficacy of a web-based computer-tailored nutrition education intervention for adults including cognitive and environmental feedback: randomized controlled trial. J Med Internet Res 2015 Jan 19;17(1):e23. [doi: 10.2196/jmir.3837] [Medline: 25599828]
57. Springvloet L, Lechner L, de Vries H, Oenema A. Long-term efficacy of a web-based computer-tailored nutrition education intervention for adults including cognitive and environmental feedback: a randomized controlled trial. BMC Public Health 2015 Apr 12;15(1):372. [doi: 10.1186/s12889-015-1707-4] [Medline: 25887891]
58. Weber J, Nigg CR. Promoting fruit and vegetable consumption during the COVID-19 pandemic—SportStudisMoveYou (SSMY): a randomized controlled trial. AIMS Public Health 2022;9(4):690-702. [doi: 10.3934/publichealth.2022048] [Medline: 36636149]
59. Power JM, Bersamin A. A text messaging intervention (Txt4HappyKids) to promote fruit and vegetable intake among families with young children: pilot study. JMIR Form Res 2018 Jul 6;2(2):e13. [doi: 10.2196/formative.8544] [Medline: 30684412]
60. Buller DB, Woodall WG, Zimmerman DE, et al. Randomized trial on the 5 a day, the Rio Grande Way Website, a web-based program to improve fruit and vegetable consumption in rural communities. J Health Commun 2008;13(3):230-249. [doi: 10.1080/10810730801985285] [Medline: 18569356]

XSL•FO
RenderX

61. Chamberland K, Sanchez M, Panahi S, Provencher V, Gagnon J, Drapeau V. The impact of an innovative web-based school nutrition intervention to increase fruits and vegetables and milk and alternatives in adolescents: a clustered randomized trial. Int J Behav Nutr Phys Act 2017 Oct 16;14(1):140. [doi: 10.1186/s12966-017-0595-7] [Medline: 29037203]
62. Nakamura S, Inayama T, Harada K, Arao T. Reduction in vegetable intake disparities with a web-based nutrition education intervention among lower-income adults in Japan: randomized controlled trial. J Med Internet Res 2017 Nov 24;19(11):e377. [doi: 10.2196/jmir.8031] [Medline: 29175810]
63. Lim TJ, Okine RN, Kershaw JC. Health- or environment-focused text messages as a potential strategy to increase plant-based eating among young adults: an exploratory study. Foods 2021 Dec 19;10(12):3147. [doi: 10.3390/foods10123147] [Medline: 34945698]
64. Carfora V, Catellani P. Legumes or meat? The effectiveness of recommendation messages towards a plant-based diet depends on people's identification with flexitarians. Nutrients 2023;15(1):15. [doi: 10.3390/nu15010015]
65. Carfora V, Morandi M, Catellani P. The effect of message framing in promoting the Mediterranean diet: the moderating role of eating self-efficacy. Foods 2022 May 17;11(10):1454. [doi: 10.3390/foods11101454] [Medline: 35627024]
66. Gosliner W, Felix C, Strochlic R, et al. Feasibility and response to the San Diego County, California, Supplemental Nutrition Assistance Program (SNAP) agency sending food and nutrition text messages to all participants: quasi-experimental web-based survey pilot study. J Med Internet Res 2023 Apr 19;25:e41021. [doi: 10.2196/41021] [Medline: 37074786]
67. Carfora V, Bertolotti M, Catellani P. Informational and emotional daily messages to reduce red and processed meat consumption. Appetite 2019 Oct 1;141:104331. [doi: 10.1016/j.appet.2019.104331] [Medline: 31276710]
68. Carfora V, Caso D, Conner M. Correlational study and randomised controlled trial for understanding and changing red meat consumption: the role of eating identities. Soc Sci Med 2017 Feb;175:244-252. [doi: 10.1016/j.socscimed.2017.01.005]
69. Carfora V, Caso D, Conner M. Randomised controlled trial of a text messaging intervention for reducing processed meat consumption: the mediating roles of anticipated regret and intention. Appetite 2017 Oct 1;117:152-160. [doi: 10.1016/j.appet.2017.06.025] [Medline: 28651971]
70. Carfora V, Catellani P, Caso D, Conner M. How to reduce red and processed meat consumption by daily text messages targeting environment or health benefits. J Environ Psychol 2019 Oct;65:101319. [doi: 10.1016/j.jenvp.2019.101319]
71. Carfora V, Zeiske N, van der Werff E, Steg L, Catellani P. Adding dynamic norm to environmental information in messages promoting the reduction of meat consumption. Environ Commun 2022 Oct 3;16(7):900-919. [doi: 10.1080/17524032.2022.2062019]
72. Wolstenholme E, Poortinga W, Whitmarsh L. Two birds, one stone: the effectiveness of health and environmental messages to reduce meat consumption and encourage pro-environmental behavioral spillover. Front Psychol 2020;11:577111. [doi: 10.3389/fpsyg.2020.577111] [Medline: 33117243]
73. Brookie KL, Mainvil LA, Carr AC, Vissers MCM, Conner TS. The development and effectiveness of an ecological momentary intervention to increase daily fruit and vegetable consumption in low-consuming young adults. Appetite 2017 Jan 1;108:32-41. [doi: 10.1016/j.appet.2016.09.015] [Medline: 27642037]
74. Rompotis CJ, Grove JR, Byrne SM. Benefits of habit‐based informational interventions: a randomised controlled trial of fruit and vegetable consumption. Aust N Z J Public Health 2014 Jun;38(3):247-252. [doi: 10.1111/1753-6405.12232]
75. Gustafson A, Jilcott Pitts SB, McQuerry K, Babtunde O, Mullins J. A mentor-led text-messaging intervention increases intake of fruits and vegetables and goal setting for healthier dietary consumption among rural adolescents in Kentucky and North Carolina, 2017. Nutrients 2019 Mar 11;11(3):593. [doi: 10.3390/nu11030593] [Medline: 30862118]
76. Pedersen S, Grønhøj A, Thøgersen J. Texting your way to healthier eating? Effects of participating in a feedback intervention using text messaging on adolescents' fruit and vegetable intake. Health Educ Res 2016 Apr;31(2):171-184. [doi: 10.1093/her/cyv104] [Medline: 26850061]
77. Alexander GL, McClure JB, Calvi JH, et al. A randomized clinical trial evaluating online interventions to improve fruit and vegetable consumption. Am J Public Health 2010 Feb;100(2):319-326. [doi: 10.2105/AJPH.2008.154468] [Medline: 20019315]
78. Dumas AA, Lemieux S, Lapointe A, Provencher V, Robitaille J, Desroches S. Effects of an evidence-informed healthy eating blog on dietary intakes and food-related behaviors of mothers of preschool- and school-aged children: a randomized controlled trial. J Acad Nutr Diet 2020 Jan;120(1):53-68. [doi: 10.1016/j.jand.2019.05.016] [Medline: 31519466]
79. Tapper K, Jiga-Boy G, Maio GR, Haddock G, Lewis M. Development and preliminary evaluation of an internet-based healthy eating program: randomized controlled trial. J Med Internet Res 2014 Oct 10;16(10):e231. [doi: 10.2196/jmir.3534] [Medline: 25305376]
80. Røed M, Medin AC, Vik FN, et al. Effect of a parent-focused eHealth intervention on children's fruit, vegetable, and discretionary food intake (Food4toddlers): randomized controlled trial. J Med Internet Res 2021 Feb 16;23(2):e18311. [doi: 10.2196/18311] [Medline: 33591279]
81. Eckert KF, Agostinelli J, Laila A, et al. Feasibility, acceptability, and preliminary impact of "Supper Heroes", a family-based sustainable diet intervention. Appetite 2025 Feb 1;206:107849. [doi: 10.1016/j.appet.2025.107849] [Medline: 39788349]
82. Livingstone KM, Rawstorn JC, Partridge SR, et al. Feasibility of a co-designed and personalised intervention to improve vegetable intake in rural-dwelling young adults. Int J Behav Nutr Phys Act 2025;22(1):97. [doi: 10.1186/s12966-025-01796-7]

XSL•FO
RenderX

83. Ricci M, Devecchi A, Migliavada R, Piochi M, Torri L. Effect of demographic characteristics and personality traits on eating patterns in the context of dietary intervention: the EATMED Case Study. Int J Environ Res Public Health 2025 Jul 10;22(7):1095. [doi: 10.3390/ijerph22071095] [Medline: 40724162]
84. Meng J, Peng W, Shin SY, Chung M. Online self-tracking groups to increase fruit and vegetable intake: a small-scale study on mechanisms of group effect on behavior change. J Med Internet Res 2017 Mar 6;19(3):e63. [doi: 10.2196/jmir.6537] [Medline: 28264793]
85. Hawkins L, Farrow C, Clayton M, Thomas JM. Can social media be used to increase fruit and vegetable consumption? A pilot intervention study. Digit Health 2024 Jan;10. [doi: 10.1177/20552076241241262]
86. Inauen J, Bolger N, Shrout PE, et al. Using smartphone‐based support groups to promote healthy eating in daily life: a randomised trial. Applied Psych Health Well 2017 Nov;9(3):303-323. [doi: 10.1111/aphw.12093]
87. Ng AH, ElGhattis Y, Biesiekierski JR, Moschonis G. Assessing the effectiveness of a 4‐week online intervention on food literacy and fruit and vegetable consumption in Australian adults: the online MedDiet challenge. Health Social Care Comm 2022 Nov;30. [doi: 10.1111/hsc.13909]
88. Carreño Enciso L, de Mateo Silleras B, de la Cruz Marcos S, Redondo Del Río P. Social media for nutrition education—a randomized controlled trial to promote fruit and vegetable intake in a university setting: “The University of Valladolid Community Eats Healthy” Study. Nutrients 2024 Apr 26;16(9):1308. [doi: 10.3390/nu16091308] [Medline: 38732555]
89. Hendrie GA, Hussain MS, Brindal E, James-Martin G, Williams G, Crook A. Impact of a mobile phone app to increase vegetable consumption and variety in adults: large-scale community cohort study. JMIR Mhealth Uhealth 2019;8(4):e14726. [doi: 10.2196/14726] [Medline: 31486407]
90. Vázquez-Paz AM, Michel-Nava RM, Delgado-Pérez EE, Lares-Michel M, Espinosa-Curiel IE. Parents’ mHealth app for promoting healthy eating behaviors in children: feasibility, acceptability, and pilot study. J Med Syst 2022 Sep 16;46(11):70. [doi: 10.1007/s10916-022-01860-w] [Medline: 36109423]
91. Liu H, Feng J, Shi Z, et al. Effects of a novel applet-based personalized dietary intervention on dietary intakes: a randomized controlled trial in a real-world scenario. Nutrients 2024;16(4):565. [doi: 10.3390/nu16040565]
92. Elbert SP, Dijkstra A, Oenema A. A mobile phone app intervention targeting fruit and vegetable consumption: the efficacy of textual and auditory tailored health information tested in a randomized controlled trial. J Med Internet Res 2016 Jun 10;18(6):e147. [doi: 10.2196/jmir.5056] [Medline: 27287823]
93. Ragelienė T, Aschemann-Witzel J, Grønhøj A. Efficacy of a smartphone application-based intervention for encouraging children’s healthy eating in Denmark. Health Promot Int 2022 Feb 17;37(1). [doi: 10.1093/heapro/daab081]
94. Shatwan IM, Alhefani RS, Bukhari MF, et al. Effects of a smartphone app on fruit and vegetable consumption among Saudi adolescents: randomized controlled trial. JMIR Pediatr Parent 2023 Feb 9;6:e43160. [doi: 10.2196/43160] [Medline: 36757770]
95. Espinosa-Curiel IE, Pozas-Bogarin EE, Lozano-Salas JL, Martínez-Miranda J, Delgado-Pérez EE, Estrada-Zamarron LS. Nutritional education and promotion of healthy eating behaviors among Mexican children through video games: design and pilot test of FoodRateMaster. JMIR Serious Games 2020 Apr 13;8(2):e16431. [doi: 10.2196/16431] [Medline: 32281539]
96. Thompson D, Bhatt R, Vazquez I, et al. Creating action plans in a serious video game increases and maintains child fruit-vegetable intake: a randomized controlled trial. Int J Behav Nutr Phys Act 2015 Dec;12(1). [doi: 10.1186/s12966-015-0199-z]
97. Buller MK, Kane IL, Dunn AL, Edwards EJ, Buller DB, Liu X. Marketing fruit and vegetable intake with interactive games on the internet. Soc Mar Q 2009 Mar;15(1_suppl):136-154. [doi: 10.1080/15245000903038316]
98. Block G, Block T, Wakimoto P, Block CH. Demonstration of an e-mailed worksite nutrition intervention program. Prev Chronic Dis 2004 Oct;1(4):A06. [Medline: 15670437]
99. Kothe EJ, Mullan BA. A randomised controlled trial of a theory of planned behaviour to increase fruit and vegetable consumption. Fresh facts. Appetite 2014 Jul;78:68-75. [doi: 10.1016/j.appet.2014.03.006] [Medline: 24656949]
100. Cumming G. Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis: Routledge; 2013. [doi: 10.4324/9780203807002]
101. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 2013 Nov 26;4:863. [doi: 10.3389/fpsyg.2013.00863] [Medline: 24324449]
102. Lenhard W. Computation of different effect sizes like d, f, r and transformation of different effect sizes. Psychometrica. URL: https://www.psychometrica.de/effect_size.html [accessed 2025-12-17]
103. Morris SB. Estimating effect sizes from pretest-posttest-control group designs. Organ Res Methods 2008 Apr;11(2):364-386. [doi: 10.1177/1094428106291059]
104. Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. Res Synth Methods 2010 Jan;1(1):39-65. [doi: 10.1002/jrsm.5] [Medline: 26056092]
105. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods 2010 Apr;1(2):97-111. [doi: 10.1002/jrsm.12] [Medline: 26061376]
106. Harrer M, Cuijpers P, Furukawa TA, Ebert DD. Doing Meta-Analysis with R: A Hands-on Guide: Chapman & Hall; 2021. URL: https://www.routledge.com/Doing-Meta-Analysis-with-R-A-Hands-On-Guide/Harrer-Cuijpers-Furukawa-Ebert/p/book/9780367610074 [accessed 2025-12-27]

XSL•FO
RenderX

107. Pustejovsky JE, Tipton E. Meta-analysis with robust variance estimation: expanding the range of working models. Prev Sci 2022 Apr;23(3):425-438. [doi: 10.1007/s11121-021-01246-3] [Medline: 33961175]
108. Sarri G, Patorno E, Yuan H, et al. Framework for the synthesis of non-randomised studies and randomised controlled trials: a guidance on conducting a systematic review and meta-analysis for healthcare decision making. BMJ Evid Based Med 2022 Apr;27(2):109-119. [doi: 10.1136/bmjebm-2020-111493] [Medline: 33298465]
109. Yao M, Mei F, Ma Y, et al. Including non-randomized studies of interventions in meta-analyses of randomized controlled trials changed the estimates in more than a third of the studies: evidence from an empirical analysis. J Clin Epidemiol 2025 Jul;183:111815. [doi: 10.1016/j.jclinepi.2025.111815] [Medline: 40334718]
110. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14(1):25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
111. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023 Dec;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
112. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta‐analysis: I2 is not an absolute measure of heterogeneity. Res Synth Methods 2017;8(1):5-18. [doi: 10.1002/jrsm.1230] [Medline: 28058794]
113. Lau J, Ioannidis JPA, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. BMJ 2006 Sep 16;333(7568):597-600. [doi: 10.1136/bmj.333.7568.597]
114. Hirst RJ, Cragg L, Allen HA. Vision dominates audition in adults but not children: a meta-analysis of the Colavita effect. Neurosci Biobehav Rev 2018 Nov;94:286-301. [doi: 10.1016/j.neubiorev.2018.07.012] [Medline: 30048672]
115. Forbes C, Greenwood H, Carter M, Clark J. Automation of duplicate record detection for systematic reviews: deduplicator. Syst Rev 2024 Aug 2;13(1):206. [doi: 10.1186/s13643-024-02619-9] [Medline: 39095913]
116. Carfora V, Caso D, Conner M. Randomized controlled trial of a messaging intervention to increase fruit and vegetable intake in adolescents: affective versus instrumental messages. Br J Health Psychol 2016 Nov;21(4):937-955. [doi: 10.1111/bjhp.12208] [Medline: 27374753]
117. Zlatevska N, Barton B, Dubelaar C, Hohberger J. Navigating through nutrition labeling effects: a second-order meta-analysis. J Public Policy Mark 2024 Jan;43(1):76-94. [doi: 10.1177/07439156231158115]
118. Qutteina Y, Hallez L, Raedschelders M, De Backer C, Smits T. Food for teens: how social media is associated with adolescent eating outcomes. Public Health Nutr 2022 Feb;25(2):290-302. [doi: 10.1017/S1368980021003116] [Medline: 34325764]
119. Ventura V, Cavaliere A, Iannò B. #Socialfood: virtuous or vicious? A systematic review. Trends Food Sci Technol 2021 Apr;110:674-686. [doi: 10.1016/j.tifs.2021.02.018]
120. Cuykx I, Decorte P, Teunissen L, et al. The magic is in the mix: a uses and gratifications approach to the cross-media use of food-related media content. Food Cult Soc 2024 Aug 7;27(4):1146-1170. [doi: 10.1080/15528014.2023.2263705]
121. Decorte P, Teunissen L, Cuykx I, et al. Media and personal socialization agents toward emerging adults' recipe choices: a cluster approach. J Foodserv Bus Res 2025:1-31. [doi: 10.1080/15378020.2024.2448642]
122. Laranjo L, Arguel A, Neves AL, et al. The influence of social networking sites on health behavior change: a systematic review and meta-analysis. J Am Med Inform Assoc 2015 Jan;22(1):243-256. [doi: 10.1136/amiajnl-2014-002841] [Medline: 25005606]
123. Laranjo L. Social media and health behavior change. In: Syed-Abdul S, Gabarron E, Lau AYS, editors. Participatory Health Through Social Media: Elsevier eBooks; 2016:83-111. [doi: 10.1016/B978-0-12-809269-9.00006-2]
124. Vogels EA, Gelles-Watnick R, Massarat N. Teens, social media and technology 2022. Pew Research Center. 2022. URL: https://www.pewresearch.org/internet/2022/08/10/teens-social-media-and-technology-2022/ [accessed 2025-12-27]
125. Qutteina Y, Hallez L, Mennes N, De Backer C, Smits T. What do adolescents see on social media? A diary study of food marketing images on social media. Front Psychol 2019;10:2637. [doi: 10.3389/fpsyg.2019.02637] [Medline: 31824391]
126. Adams J, Mytton O, White M, Monsivais P. Why are some population interventions for diet and obesity more equitable and effective than others? The role of individual agency. PLoS Med 2016 Apr;13(4):e1001990. [doi: 10.1371/journal.pmed.1001990] [Medline: 27046234]
127. Garrott K, Ogilvie D, Panter J, et al. Explaining differential socioeconomic effects in population health interventions: development and application of a new tool to classify intervention agentic demand. Lancet 2023 Nov;402:S3. [doi: 10.1016/S0140-6736(23)02056-1]
128. Heron KE, Smyth JM. Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. Br J Health Psychol 2010 Feb;15(Pt 1):1-39. [doi: 10.1348/135910709X466063] [Medline: 19646331]
129. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2020;23(1):e17187. [doi: 10.2196/17187]
130. Nagler RH. Adverse outcomes associated with media exposure to contradictory nutrition messages. J Health Commun 2014;19(1):24-40. [doi: 10.1080/10810730.2013.798384] [Medline: 24117281]
131. Mair JL, Salamanca-Sanabria A, Augsburger M, et al. Effective behavior change techniques in digital health interventions for the prevention or management of noncommunicable diseases: an umbrella review. Ann Behav Med 2023 Sep 13;57(10):817-835. [doi: 10.1093/abm/kaad041] [Medline: 37625030]

XSL•FO
RenderX

132. Ortiz R, Massar RE, McMacken M, Albert SL. Stronger together than apart: the role of social support in adopting a healthy plant-based eating pattern. Appetite 2024 Jul 1;198:107341. [doi: 10.1016/j.appet.2024.107341] [Medline: 38599245]
133. Seo DC, Niu J. Evaluation of internet-based interventions on waist circumference reduction: a meta-analysis. J Med Internet Res 2015 Jul 21;17(7):e181. [doi: 10.2196/jmir.3921] [Medline: 26199208]
134. Hawkins LK, Farrow C, Thomas JM. Do perceived norms of social media users' eating habits and preferences predict our own food consumption and BMI? Appetite 2020 Jun 1;149:104611. [doi: 10.1016/j.appet.2020.104611] [Medline: 31958481]
135. Robinson E, Fleming A, Higgs S. Prompting healthier eating: testing the use of health and social norm based messages. Health Psychol 2014 Sep;33(9):1057-1064. [doi: 10.1037/a0034213] [Medline: 24295025]
136. Stok FM, de Ridder DTD, de Vet E, de Wit JBF. Don't tell me what I should do, but what others do: the influence of descriptive and injunctive peer norms on fruit consumption in adolescents. Br J Health Psychol 2014 Feb;19(1):52-64. [doi: 10.1111/bjhp.12030] [Medline: 23406475]
137. Darmon N, Drewnowski A. Does social class predict diet quality? Am J Clin Nutr 2008 May;87(5):1107-1117. [doi: 10.1093/ajcn/87.5.1107]
138. De Irala-Estévez J, Groth M, Johansson L, Oltersdorf U, Prättälä R, Martínez-González M. A systematic review of socio-economic differences in food habits in Europe: consumption of fruit and vegetables. Eur J Clin Nutr 2000 Sep 1;54(9):706-714. [doi: 10.1038/sj.ejcn.1601080]
139. Malo JS, Schafer MH, Stull AJ. Healthy eating in life course context: asymmetric implications of socioeconomic origins and destinations. Soc Sci Med 2025 May;372:117936. [doi: 10.1016/j.socscimed.2025.117936]
140. Chan EY, Zlatevska N. Jerkies, tacos, and burgers: subjective socioeconomic status and meat preference. Appetite 2019 Jan 1;132:257-266. [doi: 10.1016/j.appet.2018.08.027] [Medline: 30172366]
141. Sawyer ADM, van Lenthe F, Kamphuis CBM, et al. Dynamics of the complex food environment underlying dietary intake in low-income groups: a systems map of associations extracted from a systematic umbrella literature review. Int J Behav Nutr Phys Act 2021 Dec;18(1). [doi: 10.1186/s12966-021-01164-1]
142. van der Heijden A, Te Molder H, Jager G, Mulder BC. Healthy eating beliefs and the meaning of food in populations with a low socioeconomic position: a scoping review. Appetite 2021 Jun 1;161:105135. [doi: 10.1016/j.appet.2021.105135] [Medline: 33493606]
143. Adler NE, Epel ES, Castellazzo G, Ickovics JR. Relationship of subjective and objective social status with psychological and physiological functioning: preliminary data in healthy, White women. Health Psychol 2000;19(6):586-592. [doi: 10.1037//0278-6133.19.6.586]
144. Cheon BK, Hong YY. Mere experience of low subjective socioeconomic status stimulates appetite and food intake. Proc Natl Acad Sci USA 2017 Jan 3;114(1):72-77. [doi: 10.1073/pnas.1607330114]
145. D'Hooge L, Achterberg P, Reeskens T. Mind over matter. The impact of subjective social status on health outcomes and health behaviors. PLoS ONE 2018;13(9):e0202489. [doi: 10.1371/journal.pone.0202489] [Medline: 30183731]
146. van den Bekerom L, van Gestel LC, Schoones JW, Bussemaker J, Adriaanse MA. Health behavior interventions among people with lower socio-economic position: a scoping review of behavior change techniques and effectiveness. Health Psychol Behav Med 2024;12(1):2365931. [doi: 10.1080/21642850.2024.2365931] [Medline: 38903803]
147. Taufik D, Verain MCD, Bouwman EP, Reinders MJ. Determinants of real-life behavioural interventions to stimulate more plant-based and less animal-based diets: a systematic review. Trends Food Sci Technol 2019 Nov;93:281-303. [doi: 10.1016/j.tifs.2019.09.019]
148. Eppes EV, Augustyn M, Gross SM, Vernon P, Caulfield LE, Paige DM. Engagement with and acceptability of digital media platforms for use in improving health behaviors among vulnerable families: systematic review. J Med Internet Res 2022;25:e40934. [doi: 10.2196/40934]
149. Richardson M, Garner P, Donegan S. Interpretation of subgroup analyses in systematic reviews: a tutorial. Clin Epidemiol Glob Health 2019 Jun;7(2):192-198. [doi: 10.1016/j.cegh.2018.05.005]
150. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-926. [doi: 10.1136/bmj.39489.470347.AD]
151. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev 2014;2014(4):MR000034. [doi: 10.1002/14651858.MR000034.pub2]
152. Golder S, Loke YK, Bland M. Meta-analyses of adverse effects data derived from randomised controlled trials as compared to observational studies: methodological overview. PLoS Med 2011 May;8(5):e1001026. [doi: 10.1371/journal.pmed.1001026] [Medline: 21559325]
153. Cuello-Garcia CA, Santesso N, Morgan RL, et al. GRADE guidance 24 optimizing the integration of randomized and non-randomized studies of interventions in evidence syntheses and health guidelines. J Clin Epidemiol 2022 Feb;142:200-208. [doi: 10.1016/j.jclinepi.2021.11.026]
154. Zhou Y, Yao M, Mei F, et al. Integrating randomized controlled trials and non-randomized studies of interventions to assess the effect of rare events: a Bayesian re-analysis of two meta-analyses. BMC Med Res Methodol 2024 Sep 27;24(1):219. [doi: 10.1186/s12874-024-02347-7] [Medline: 39333867]

XSL•FO
RenderX

155. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014 Mar 7;348(mar07 3):g1687. [doi: 10.1136/bmj.g1687] [Medline: 24609605]

Abbreviations

BCT: behavior change technique
BCTTv1: Behavior Change Technique Taxonomy version 1
FVI: fruit and vegetable intake
IG: intervention group
NRS: nonrandomized study
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
RCT: randomized controlled trial
SES: socioeconomic status

Edited by S Brini; submitted 17.Jul.2025; peer-reviewed by B Lam, ME Heidari, N Maye, O Oyetunji, S Narayan; revised version received 01.Dec.2025; accepted 02.Dec.2025; published 08.Jan.2026.

Please cite as:
Vanwinkelen K, Spruyt B, Smits T
Digital Interventions Targeting Healthy and Sustainable Eating Behavior: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e80821
URL: https://www.jmir.org/2026/1/e80821
doi:10.2196/80821

© Käbi Vanwinkelen, Bram Spruyt, Tim Smits. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Diet-Related Health Recommender Systems for Patients With Chronic Health Conditions: Scoping Review

Xiaolan Dong[1*], MD; Bei Yun[2*], MD; Anni Pakarinen[1,3], PhD; Zhuting Zheng[2], MSc; Hao Niu[2], PhD; Tian Jin[2], MD; Changrong Yuan[2], PhD; Jingting Wang[4], PhD

[1]Department of Nursing Science, University of Turku, Turku, Finland
[2]School of Nursing, Fudan University, Shanghai, China
[3]Faculty of Nursing and Physiotherapy, University of Lleida, Lleida, Spain
[4]School of Nursing, Naval Medical University, 800 Xiangyin Road, Shanghai, China
[*]these authors contributed equally

Corresponding Author:
Jingting Wang, PhD
School of Nursing, Naval Medical University, 800 Xiangyin Road, Shanghai, China

Abstract

Background: Diet-related Health Recommender Systems (HRSs) have gained attention for their potential to provide personalized dietary guidance, particularly for patients with chronic conditions. However, studies on diet-related HRSs in health care are relatively limited.

Objective: This scoping review aims to present the state of current research on diet-related HRSs for patients with chronic health conditions, identify existing gaps, and suggest future research directions.

Methods: The scoping review was conducted following the Arksey and O'Malley framework and was reported in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines. The literature search was conducted in October 2024 across 6 English databases (PubMed, Medline, Embase, Web of Science Core Collection, IEEE Xplore, and CINAHL) and 4 Chinese databases (SinoMed, CNKI, Wanfang, and VIP). Studies focusing on diet-related HRSs for patients with chronic conditions were included.

Results: Fifteen studies published between 2010 and 2024 from 9 countries were included. Diet-related HRSs mainly target adults with chronic diseases, with 9 systems (60%) including users with diabetes and 6 (40%) including users with hypertension. Nine studies (60%) described functional structures, which were categorized into 4 components: user information, food or diet recommendations, knowledge and decision support, and data management with additional functions. Recommended content was categorized into 5 types: food (n=6, 40%), recipes (n=4, 26.67%), diet plans or meal plans (n=3, 20%), recipes and food (n=1, 6.67%), and meals (n=1, 6.67%). Recommendation methods included constraint-based (n=6, 40%), focusing on patients' dietary restrictions; preference-based (n=5, 33.33%), considering patients' food preferences; and hybrid (n=4, 26.67%), combining both approaches. Of all recommendation technologies, most studies (n=13, 86.67%) applied hybrid approaches, enabling more robust personalization. For the data used for training, 13 studies (86.67%) explicitly mentioned the data sources, and 10 studies' (66.67%) data came from professional organizations and websites. The recommendation process followed a structured workflow. Twelve studies (80%) evaluated diet-related HRSs using either online or offline methods, while accuracy (n=9, 60%) has been the most common evaluation criterion. However, no studies went deeper into how these systems affected users' dietary behaviors over time.

Conclusions: Diet-related HRSs have the potential to deliver personalized dietary support for patients with chronic diseases, but current systems show key gaps. Future development must adopt user-centered design, provide practical and actionable dietary guidance, and use hybrid recommendation techniques to increase precision and clinical relevance. Standardized evaluation methods and real-world, long-term studies are essential to evaluate the impact of diet-related HRSs on dietary behavior and health outcomes. Addressing these needs will enable diet-related HRSs to become reliable tools for chronic disease management and patient-centered care.

Trial Registration: International Platform of Registered Systematic Review and Meta-analysis Protocols INPLASY202550081; https://inplasy.com/wp-content/uploads/2025/05/INPLASY-Protocol-7837.pdf

(J Med Internet Res 2026;28:e77726) doi:10.2196/77726

XSL•FO
RenderX

KEYWORDS

Health Recommender System; Diet-related Health Recommender System; chronic health conditions; diet; scoping review; PRISMA; HRS

Introduction

Dietary management is crucial to overall health, especially in the context of global trends where poor dietary habits have become a significant factor contributing to weight-related issues [1]. Statistics from the latest Global Nutrition Report indicated alarming rates of overweight and obesity among adults, with 40.8% of adult (18 years or older) women and 40.4% of adult men affected. Conversely, 9.1% of adult women and 8.1 % of adult men were underweight [2]. Improper dietary management not only affects body weight but also contributes to the development of diabetes [3,4], hypertension [5], cardiovascular diseases [6], chronic kidney disease [7], and inflammatory bowel disease [8], among others.

The Scientific Research Report on Dietary Guidelines for Chinese Residents (2021) highlighted that the overweight and obesity rates among children younger than 6 years old and those aged 6 - 17 years were 10.4% and 19.0%, respectively [9]. Among residents aged 18 years and older, the overweight rate was 34.3% and the obesity rate was 16.4%, with 50.7% of adults being overweight or obese [9]. Overweight and obesity are significant risk factors for cardiovascular diseases, diabetes, cancer, and other chronic health conditions [9].

Poor dietary habits contribute to both the onset and progression of these chronic conditions. Therefore, scientific and effective diet management is essential for ensuring proper nutrient intake, which directly enhances individual immunity, halts disease progression, impacts health status, and supports recovery from chronic health conditions, thereby improving overall health outcomes [7,10,11]. Moreover, good diet management regulates sleep and mood, reduces fatigue, and comprehensively enhances overall health [12,13]. In February 2024, China's National Health Commission released the "Dietary Guidelines for Adults with Hyperuricemia and Gout (2024 Edition)," "Dietary Guidelines for Adult Obesity (2024 Edition)," "Dietary Guidelines for Childhood and Adolescent Obesity (2024 Edition)," and "Dietary Guidelines for Adults with Chronic Kidney Disease (2024 Edition)" [14]. These guidelines aim to prevent and control the occurrence and progression of chronic diseases among the Chinese population through dietary management.

Patients with chronic health conditions often need to adjust their diets based on their specific health status to manage health conditions effectively and enhance the overall quality of life. When patients are required to follow dietary restrictions due to their chronic health condition, it is important for the patient to distinguish whether certain foods are permissible, ensure that the cooking methods meet the essential requirements, evaluate whether portion sizes are appropriate, and strive to maintain a balanced diet to reduce the risk of malnutrition and other issues caused by dietary limitations [15]. Therefore, these patients require targeted and personalized guidance to help them implement scientific and feasible dietary management practices.

Recommender Systems (RSs) are software tools that provide suggestions or recommendations of items to users, such as products, services, information, or content, based on their preferences, interests, and past behaviors [16], which have been used in several domains such as e-commerce, e-learning, e-tourism, or eHealth. To generate recommendations, RSs commonly rely on a set of foundational recommendation technologies. Collaborative filtering (CF) is one of the most widely applied recommendation technologies, which provides recommendations for users by using the known preferences of other users with similar behaviors [17]. In contrast, content-based (CB) methods recommend items similar to those a given user has previously liked [18]. Unlike CF and CB, which rely on large historical rating datasets, knowledge-based (KB) approaches produce suggestions by leveraging domain knowledge, expert rules, and explicit user constraints [19]. Knowledge graph (KG)-driven recommendation further enhances personalization by representing domain knowledge in graph structures to generate precise suggestions [20]. Finally, hybrid recommendation (HyR) integrates 2 or more strategies to maximize their strengths and mitigate individual limitations [21].

Health Recommender Systems (HRSs) are also one of the RS's important application scenarios. HRSs offer the potential to motivate and engage users to change their behavior by sharing better choices and practical knowledge based on observed user behavior [22]. HRSs have been applied in health care in recent years, in areas such as mental health [23], hearing aid usage [24], health education [25], physical activity [26], and diet-related health [27].

Diet-related HRSs are software tools that use personalized data to provide tailored food recommendations from a wide range of options [28]. Diet-related HRSs present promising solutions to the issues of information overload and limited food choices, which contribute significantly to diet-related health problems [29]. By using personalized data, diet-related HRSs offer tailored food recommendations that take into account users' taste preferences, dietary needs, and medical conditions [30]. Such systems can filter and sort food options [27], fostering a better understanding of dietary health and increasing user engagement [31]. Additionally, diet-related HRSs can enhance dietary outcomes by providing nutrition assessments and offering healthier meal plans and recipes [28].

While diet-related HRSs have been explored within the field of information and communication technology, research on their application in chronic disease dietary management is still limited, and there is a lack of comprehensive literature reviews in this area. To address these gaps, this study adopted a scoping review methodology to review the research status quo of diet-related HRSs for individuals with chronic conditions. In this review, diet-related HRSs are defined as systems specifically designed for patients with chronic health conditions that require disease-related dietary restrictions. The target populations,

XSL•FO
RenderX

function structures, recommendation content, implementation of recommendation features, and evaluation of the diet-related HRSs were analyzed to provide a reference for health care researchers seeking to design more effective and user-friendly systems.

Methods

Overview

Scoping reviews assess the extent of the research, range, and nature, identify gaps, determine systematic review value, and disseminate research findings [32]. A scoping review methodology is used to systematically map available research on the broad, complex, and emerging research question [32]. In emerging research fields, a lack of randomized controlled trials may impede formal systematic reviews or meta-analyses. Scoping studies can address diverse questions and incorporate a range of study designs, making them ideal to complement clinical trial findings. The scoping review was conducted following the Arksey and O'Malley framework [32] and was reported in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines [33] and was structured according to the following outlined steps.

Stage 1: Identification of the Research Question

The research question was identified after an initial review of the literature and a discussion within our research team. The key research questions that guided the review were as follows:

1. What is the current state of research on diet-related HRSs for patients with chronic health conditions?
2. What are the target users for diet-related HRSs for patients with chronic health conditions?
3. What are the function structures in diet-related HRSs for patients with chronic health conditions?
4. What types of recommendation content are provided by diet-related HRSs for patients with chronic health conditions?
5. How are recommendation features implemented in diet-related HRSs for patients with chronic health conditions?
6. How are diet-related HRSs for patients with chronic health conditions evaluated?

Stage 2: Identifying Relevant Studies

Before searching, the research team worked collaboratively in making decisions about inclusion and exclusion criteria and in planning the initial search strategies to comprehensively identify the relevant literature. The final search strategies were developed with the assistance of a health science librarian (Multimedia Appendix 1). The literature search was performed by 3 researchers in October 2024, covering 6 English databases (PubMed, Medline, Embase, Web of Science Core Collection, IEEE Xplore, and CINAHL) and 4 Chinese databases (SinoMed, CNKI, Wanfang, and VIPC). Relevant studies were searched from January 2010 to October 2024 to answer the above research questions (1-6). An example of the search strategy performed in Web of Science Core Collection is presented here: (recommender system* OR hybrid recommendation* OR collaborative filtering OR content based recommendation* OR recommendation* system* OR knowledge based recommendation*) AND (recipe* OR diet* OR food OR eat* OR nutrition*). As the search proceeded, additional terms were suggested by experts to potentially modify the question, but new research did not result in additional data, and the question did not change. Duplicate references were filtered out using EndNote. In addition to database search, hand searching was conducted by screening the reference lists of all included articles and relevant review papers to identify additional eligible studies. We also manually checked key journals in the field and conference proceedings to ensure comprehensive coverage.

Stage 3: Study Selection

All studies searched were independently assessed by 2 authors (XD and BY) based on the inclusion criteria listed below, and discrepancies were verified by a third author (JW). Studies were eligible for inclusion in this review if all the following criteria were met: (1) the recommended information was related to at least one of the following: food, meal plan, diet plan, and recipe; (2) the study applied personalized recommendation strategy; (3) the recommendations were generated using algorithmic and technological methods; (4) the study population comprised patients with chronic health conditions; and (5) the study was published in a peer-reviewed journal or conference proceeding. The exclusion criteria included: (1) the recommendation unrelated to human health (eg, study focusing on animal health [34]), (2) studies reporting the same RSs were considered duplicates; only the most recent or most comprehensive publication was retained (eg, the latest published study [35] was included, while the earlier one [36] was excluded), and (3) full-text articles not in English or Chinese language.

Stage 4: Charting the Data

Decisions regarding the information to be recorded from the primary studies were made through discussions within the research group. Subsequently, a structured chart was developed to collate, summarize, and share the extracted data. A descriptive-analytical narrative approach was used to extract and chart the data from the selected articles [32,37]. Three researchers (XD, BY, and ZZ) independently extracted the data and performed coding, with the other two authors (HN and JW) verifying accuracy. All discrepancies were resolved by consensus. The following details were documented for each included study to answer the research questions: (1) nationality and publication year (Multimedia Appendix 2); (2) target users; (3) function structures; (4) recommendation content; (5) recommendation method, recommendation technology, data of training set, and recommendation process; and (6) evaluation method, evaluation criteria, test set, or evaluation sample size.

Stage 5: Collating, Summarizing, and Reporting the Results

The scoping review methodology aimed to summarize the breadth and depth of the existing literature. At this stage, an overview of the characteristics of all included articles was collated, summarized, and reported. Initially, a basic numerical summary of the studies, including the extent, nature, and distribution of the articles, was presented. As this was a scoping

XSL•FO
RenderX

review, the critical appraisal of the quality of the included studies was not conducted. However, efforts were made to map the diversity and variety of diet-related HRSs based on factors such as the characteristics of target users, function structure, recommendation content, and other factors. This process facilitated researchers in reaching conclusions about the key characteristics of research in this field and provided insights for future studies.

Results

Overview

In total, 4492 published studies were identified in the searching process (Figure 1). EndNote X9 was used to exclude 350 duplicates, and 4142 studies were excluded based on a review of their titles and abstracts. The remaining 208 studies were searched for full text. Ultimately, 193 were excluded based on the exclusion criteria, and 15 studies [35,38-51] were included and analyzed.

Figure 1. Flow diagram according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines.

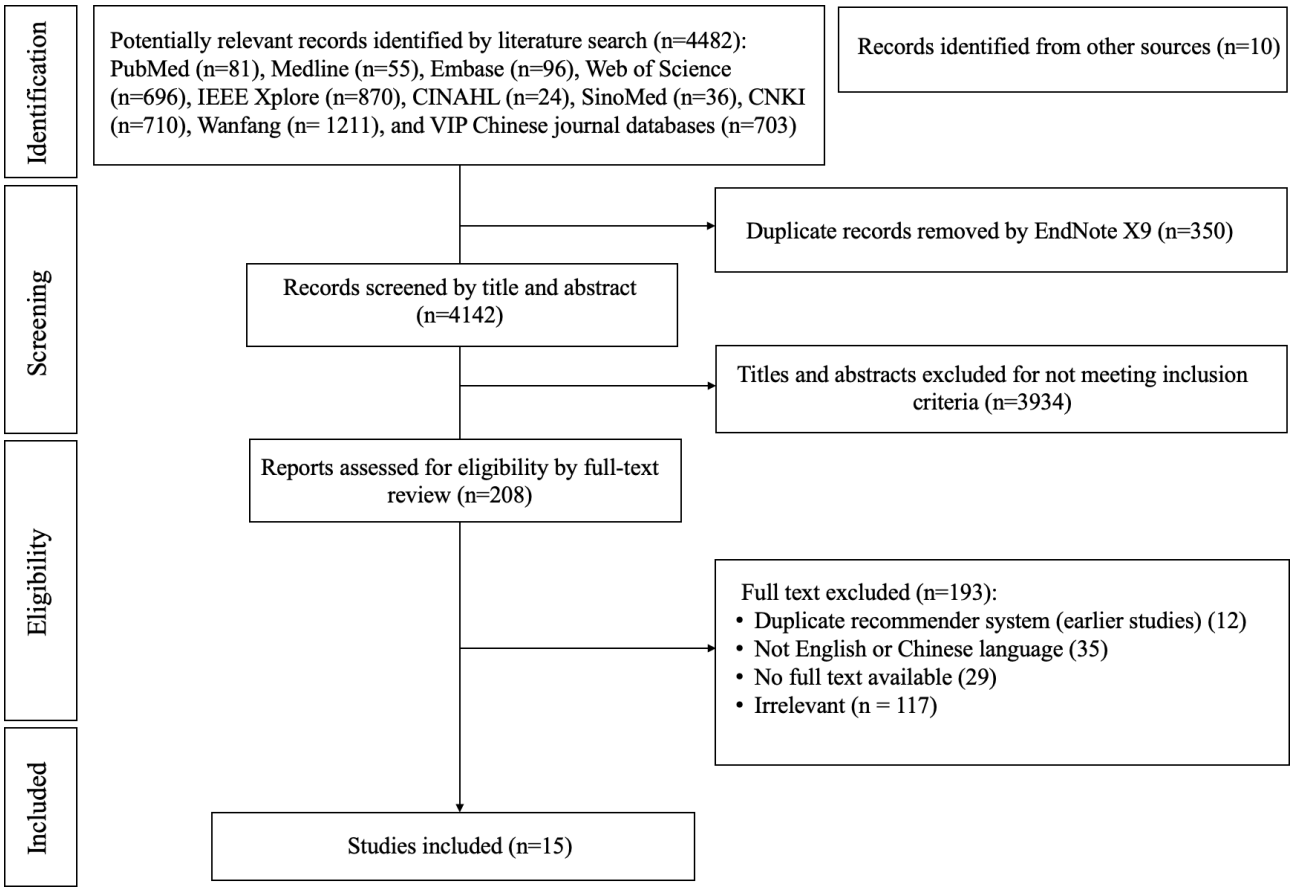

The Current State of Research on Diet-Related HRSs for Patients with Chronic Health Conditions

The publication years and the countries of the included studies are shown in Multimedia Appendix 2. The studies were published between 2010 and 2024, and there has not been a noticeable increase in their publication over the past 5 years. The articles originated from 9 countries (based on the first author's affiliations), with the top 3 being China (n=5, 33.33%) [39,40,43,44,49], India (n=2, 13.33%) [41,47], and Pakistan (n=2, 13.33%) [38,48].

Target Users for Diet-Related HRSs for Patients with Chronic Health Conditions

Among the 15 included studies [35,38-51], only 2 reported the age of the target users, which were 18 - 80 years [50] and older than 65 years [39], respectively. The target users and patients' chronic health conditions are shown in Table 1. Diet-related HRSs primarily target adults with chronic diseases. Among chronic diseases, diabetes and hypertension were the most commonly targeted by diet-related HRSs, with 9 systems (60%) including users with diabetes and 6 (40%) including users with hypertension.

XSL•FO
RenderX

Table . The characteristics of target users for diet-related Health Recommender Systems.

Variable	Studies, n (%)	References
Target users		
Patients with a single disease	6 (40)	[38,40,41,44,46,51]
Patients with complex diseases	4 (26.67)	[39,43,47,48]
Both healthy people and patients	5 (33.33)	[35,42,45,49,50]
Patients' chronic health conditions		
Diabetes	9 (60)	[35,38,42,45-49,51]
Hypertension	6 (40)	[35,40,45,47-49]
Cancer	3 (20)	[41,42,44]
Obesity	3 (20)	[42,47,50]
Kidney disease	3 (20)	[35,47,48]
High cholesterol	2 (13.33)	[47,49]
Metabolic syndrome	1 (6.67)	[43]
Osteoporosis	1 (6.67)	[42]
Iron deficiency	1 (6.67)	[48]
Cardiovascular diseases	1 (6.67)	[42]
Chronic dental problems	1 (6.67)	[42]
Unspecified geriatric diseases	1 (6.67)	[39]

The Function Structures in Diet-Related HRSs for Patients with Chronic Health Conditions

Among the 15 studies included [35,38-51], 9 (60%) [35,38-45] studies described function structures. The function structures described in these studies varied, but certain recurring components were consistently noted, which could be summarized into 4 major components: user information, food or diet recommendations, knowledge and decision support, and data management and additional functions. Detailed information is shown in Table 2.

XSL•FO
RenderX

Table . The function structures in diet-related Health Recommender Systems.

Variable	References
User information	
Basic user information	
Personal/user profile	[35,38,39]
User information	[40]
History	[35,41,42]
Personal health information	
Physical activity	[38]
Health management report	[43]
Diseases	[44]
Symptoms	[44]
Diagnoses	[41]
Food or diet recommendations	
Food filter and security	[35]
Food recommendations or suggestions	[38,45]
Weekly meal plans	[42]
Diet guides	[39,40]
Knowledge and decision support	
Chronic kidney disease calculator	[35]
Food	[38]
Nutrient search engines	[45]
News and related sites	[45]
Ingredients	[44]
Nutritional expert knowledge	[39]
Clinical practice and nutrition guidelines	[39]
Knowledge base	[39]
Diet charts	[41]
User ratings	[42]
Data management and additional functions	
Data entry	[38]
Data management	[40]
Reminders	[35]
Real-time interaction	[43]
Settings	[40]
Changing details	[41]

The Types of Recommendation Content Provided by Diet-Related HRSs for Patients with Chronic Health Conditions

The recommended content was divided into 5 categories. Detailed information is shown in Table 3.

XSL•FO
RenderX

Table . The types of recommendation content provided by the diet-related Health Recommender Systems.

Recommended content	Studies, n (%)	References
Food	6 (40)	[35,41,46-49]
Recipe	4 (26.67)	[38,42,44,45]
Diet plan or meal plan	3 (20)	[40,43,50]
Both recipe and food	1 (6.67)	[51]
Meal	1 (6.67)	[39]

The Implementation of Recommendation Features in Diet-Related HRSs for Patients with Chronic Health Conditions

Table 4 presents detailed information on the implementation of recommendation features in the reviewed diet-related HRSs. Recommendation methods were inductively classified into 3 categories: constraint-based, preference-based, and hybrid. In this review, constraint-based refers to recommendations based on the patient's condition-related dietary restrictions, while preference-based refers to recommendations based on the patient's dietary preferences. Hybrid refers to methods that incorporate both constraint-based and preference-based approaches. Among the 15 included studies [35,38-51], 6 (40%) used constraint-based methods [41,43,46,47,49,51], 5 (33.33%) applied preference-based methods [39,40,44,45,50], and 4 (26.67%) adopted hybrid methods [35,38,42,48].

XSL•FO
RenderX

Table . The implementation of recommendation features in diet-related Health Recommender Systems.

Author [reference], publication year	Recommendation method	Recommendation technology	Data of the training set	Recommendation process
Phanich et al [46], 2010	Constraint-based	Hybrid recommendation	Data were selected from the nutrition division, Ministry of Public Health, to form the training dataset (n=290[a])	• The dataset was grouped by the system and categorized according to food characteristics and nutrition for diabetes (normal food, limited food, and avoidable food). • Features were extracted by nutrient ranking. • Food clustering was analyzed. • Relevant food items were ranked based on the minimum distance, extracted, and recommended in ascending order from the ranking.
Arwan et al [51], 2013	Constraint-based	Hybrid recommendation	Data were selected from some references of foods and enriched with the information from nutrition experts to form the training dataset (n=NR[b]).	• The food ontology and calorie food ontology were developed. • The rule for classifying or grouping data categories within the ontology was developed. • The food menu search feature was built. • An experiment was conducted to test whether the system could recommend correctly.
Faiz et al [38], 2014	Preference-based and constraint-based	Hybrid recommendation	Data were selected from standard and well-known resources such as the US Department of Agriculture and MyFitnessPal to form the training dataset (n=NR[b]).	• The domain ontologies (personal health profile, food, and diseases) were built. • The ontologies were integrated and rule terms were defined. • The diet was recommended to the user at the predefined time.
Ting et al [45], 2014	Preference-based	Hybrid recommendation	Data were selected from the Food Composition Database and Japan Preventive Association of Lifestyle-related Disease to form the training dataset (n=NR[b]).	• A query with the user data was sent to the system. • Recipes were obtained from the database through a search. • Recipes that did not suit the current health status or match the user's preferences were filtered out. • Five recipes were recommended from the filtered set.

XSL•FO
RenderX

Author [reference], publication year	Recommendation method	Recommendation technology	Data of the training set	Recommendation process
Chen et al [49], 2015	Constraint-based	Hybrid recommendation	Data were collected from the Taiwan Area Food Nutrition Database website published by the US Food and Drug Administration to form the training dataset (n=NR[b]).	• The system collected personal information and classified the user's nutrients. • The patient's dietary records were imported into the nutrition expert knowledge system. • The personal information was imported into the personal disease ontology. • The nutrient data were analyzed by the expert knowledge system. • Suitable foods for the user were inferred.
Tseng et al [43], 2015	Constraint-based	Hybrid recommendation	__[c]	• The patient's vital signs were collected and transmitted. • The risk was evaluated and reported. • A diet plan was generated and recommended to the patient.
Elsweiler et al [50], 2015	Preference-based	Hybrid recommendation	Data were selected from a self-created food portal website, including users (n=148[d]) and recipes (n=957[a]) to form the training dataset.	• The user's nutritional requirements were established. • Recipe ratings were estimated based on each profile. • Recipes were combined. • Recipes were established if the combination met the requirements.
Rehman et al [48], 2017	Preference-based and constraint-based	Hybrid recommendation	Data were selected from the official website of the Composition of Foods Integrated Dataset (CoFID) to form the training dataset (n=3400[a]).	• A cloud-based food recommender system, named Diet-Right, was developed based on users' pathological reports. • The ant colony algorithm was used to generate an optimal food list. • Suitable foods were recommended according to the values of pathological reports.
Agapito et al [35], 2018	Preference-based and constraint-based	Hybrid recommendation	__[c]	

XSL•FO
RenderX

Author [reference], publication year	Recommendation method	Recommendation technology	Data of the training set	Recommendation process
				• The user profile was created by giving specific questions about clinical parameters. • All changes made by the user were saved, allowing the data to be used for monitoring the user's health status. • After the user was profiled, typical foods that could be consumed by the user were recommended.
Rathi et al [41], 2019	Constraint-based	Hybrid recommendation	Data were selected from the UC Irvine Machine Learning Repository (including the Liver Patient Dataset, Heart Disease Dataset, Diabetes Dataset, Breast Cancer Dataset, and Thyroid Dataset).	• The dataset was collected and preprocessed from different sources. • The system was built to detect diseases and recommend diets accordingly.
Manoharan et al [47], 2020	Constraint-based	Hybrid recommendation	Data were collected from 50 patients through the internet and hospitals to form the training dataset (n=50[d]).	• Data were collected from the internet and hospitals. • Features were sorted, preprocessed, and encoded, and the food data were segregated based on similarities. • Food was recommended. • The system was trained, tested, and cross-validated.
Qi et al [40], 2021	Preference-based	Hybrid recommendation	Data were selected from hospital meal history records and split into training and testing sets in a 7:3 ratio (n=718[a]).	• A MySQL database was used to implement data dictionary management. • The rule extraction module was implemented with the knowledge base management system. • A meal plan was generated and recommended.
Tang et al [44], 2023	Preference-based	Knowledge graph	Data were crawled from recipe websites and manually extracted from textbooks to form the training dataset (n=NR[b]).	• The embedded representation of items was enhanced through message-passing and update functions on node features. • The influence of time on users' taste preferences was considered. • Long Short-Term Memory (LSTM) networks were introduced to dynamically adjust users' personal taste preferences.

XSL•FO
RenderX

Author [reference], publication year	Recommendation method	Recommendation technology	Data of the training set	Recommendation process
Zioutos et al [42], 2024	Preference-based and constraint-based	Hybrid recommendation	Data were selected from a large real-world dataset of recipes to form the training dataset (n=2,774,676[a]).	• The user's health history was analyzed. • Similar users were identified. • Personalized recipe recommendations were provided. • Dynamically adaptable adjustments were made.
Xu et al [39], 2024	Preference-based	Knowledge graph	Data were selected from a community survey (n=96[d]) and a website of Chinese cuisine recipes and eating history (n=180[a]) to form the training dataset.	• The FoodKG was constructed. • User profiles related to older adults' dietary behaviors were built. • Personalized meal recommendation algorithms were developed, including candidate dish generation and combo meal recommendations.

[a]The sample size of subjects including a recipe, food, and diet plan or meal plan.

[b]NR: not reported.

[c]Not available.

[d]The sample size of the population.

While the recommendation technologies were coded according to terms reported in the studies. The recommendation technology included: KG (n=2, 13.33%) [39,44] and HyR (n=13, 86.67%) [35,38,40-43,45-51].

For the data of the training set, 13 studies (86.67%) [38-42,44-51] explicitly mentioned the data sources of the training set, while the other 2 studies (13.33%) [35,43] did not specify this information. The data sources used in the reviewed studies varied widely and can be categorized into 4 main types: authoritative government and institutional databases (n=5, 33.33%) [38,45,46,48,49], academic databases and publicly available datasets (n=2, 13.33%) [39,41], expert and hospital data (n=3, 20%) [40,47,51], and recipe and user-generated data (n=3, 20%) [42,44,50].

The recommendation process followed a structured workflow, integrating advanced technologies to generate more accurate, adaptive, and context-aware dietary recommendations. The structured workflow included user profiling, integration of structured knowledge (such as food databases or ontologies), personalized filtering and matching based on health conditions and preferences, and ranking of suitable options for final recommendation. While this core process was consistent, 6 systems incorporated advanced technologies to enable more accurate, adaptable, and context-aware dietary recommendations, such as ontology-based reasoning [51], optimization algorithms (eg, ant colony [48]), dynamic modeling using Long Short-Term Memory networks [44], expert rule systems [39,49], and the use of KGs such as FoodKG [39].

The Evaluation of Diet-Related HRSs for Patients with Chronic Health Conditions

Among the 15 included studies [35,38-51], 12 studies (80%) [35,39-44,46-49,51] evaluated the diet-related HRSs, while the remaining 3 studies (20%) [38,45,50] did not conduct an evaluation. Table 5 presents detailed information on the evaluation criteria, evaluation method, and test set or evaluation sample size of the 12 studies [35,39-44,46-49,51].

XSL•FO
RenderX

Table . The evaluation of diet-related Health Recommender Systems (n=12).

Author [reference], publication year	Evaluation criteria	Evaluation method	Test set or evaluation sample size
Phanich et al [46], 2010	Acceptance, usability, and accuracy	Online	Nutritionists (n=NR[a])
Arwan et al [51], 2013	Accuracy	Offline	Ontology patient instances (n=30[b])
Chen et al [49], 2015	Accuracy	Online	One older adult from a silver-haired home and dietitians (n=NR[a])
Tseng et al [43], 2015	Feasibility	Online	Patients (n=NR[a])
Rehman et al [48], 2017	Accuracy	Offline	Same as the training set
Agapito et al [35], 2018	Accuracy and usability	Online	Patients with CKD[c] (n=20[b]); healthy people (n=20[b])
Rathi et al [41], 2019	Accuracy	Offline	Same as the training set. The dataset was split into training:testing 75:25 (n=9326[d])
Manoharan et al [47], 2020	Accuracy	Offline	Same as the training set
Qi et al [40], 2021	Efficiency and satisfaction	Online	Patients (n=37[b]) and nutritionists (n=3[b])
Tang et al [44], 2023	Accuracy	Offline	Same as the training set
Zioutos et al [42], 2024	Accuracy and usability	Online	40 existing users of the food.com website: patients (n=20[b]) and healthy people (n=20[b])
Xu et al [39], 2024	Effectiveness	Online	Community-dwelling older adults (n=96[b]): tracked group (n=34[b]) and an untracked group (n=62[b]); A total of 91 participants (94.79%) were diagnosed with chronic conditions

[a]NR: not reported.

[b]The sample size of the population.

[c]CKD: chronic kidney disease.

[d]The sample size of subjects including a recipe, food, and diet plan or meal plan.

The evaluation criteria were extracted directly from the included studies, including: accuracy (n=9) [35,41,42,44,46-49,51] refers to the extent to which the system's predictions of patients' preferences or nutritional needs match their actual dietary preferences or intake; usability (n=3) [35,42,46] refers to the ease with which patients can, with little or no assistance, successfully operate the system to receive recommendations, navigate, and interact with its interface; acceptance (n=1) [46] refers to the degree to which patients are willing to receive and use the dietary recommendations provided by the system; satisfaction (n=1) [40] refers to patients' subjective sense of contentment or evaluation of the overall usefulness and experience of the recommendations, measured through participants' comprehensive comparison of system-generated meal plans with manually designed ones; efficiency (n=1) [40] refers to the amount of time, steps, or cognitive effort required for patients to obtain appropriate dietary recommendations from the system; feasibility (n=1) [43] refers to the practicality of deploying the recommendation system in real-life settings; and effectiveness (n=1) [39] refers to the actual positive impact of the system in real-life contexts on patients' dietary behavior changes or health outcomes. Four studies [35,40,42,46] included more than one evaluation criterion.

The evaluation methods included: online evaluation (n=7) [35,39,40,42,43,46,49] and offline evaluation (n=5) [41,44,47,48,51]. Among the 7 studies [35,38,42,44,45,48,51] that used online evaluation, 6 studies [35,39,40,42,43,49] evaluated the diet-related HRSs with target users of patients, while 3 studies evaluated the HRSs that relied on experts, such as nutritionists [40,46] and dietitians [49]; their main roles were to validate the recommendation results. Four of the 5 offline evaluation studies [39,40,43,46,47] used the same dataset for training and evaluation [41,44,47,48].

Discussion

Principal Findings

This scoping review revealed that diet-related HRSs for individuals with chronic conditions were still in their early stages, with limited patient-specific designs and significant room for improvement. The review highlighted substantial gaps in target users, system functions, recommendation content, recommendation feature implementation, and evaluation approaches, which need to be addressed to support the

XSL•FO
RenderX

development of more effective and patient-centered diet-related HRSs.

Comparison with Prior Work

Target Users of Diet-Related HRSs

The target users of diet-related HRSs were mainly patients with single or complex chronic health conditions. They not only need dietary recommendations tailored to their preferences [39,41,42,44,48,50], but most importantly, their health conditions [38,40,45,46,51], highlighting the need for more disease-specific solutions. Age groups are another factor that needs to be considered, for example, older adults often face challenges such as limited mobility, cognitive decline, and changes in appetite or food preferences [40,42], while sick children and their caregivers face unique dietary needs due to treatment-related restrictions and appetite loss [52,53]. Therefore, balancing disease-related requirements with personal needs is fundamental to developing high-quality diet-related HRS. It is also essential to support caregivers in preparing meals that are nutritionally appropriate and medically compliant.

Function Structures of Diet-Related HRSs

The user information component enables diet-related HRSs to collect basic demographic and health data, which supports the generation of tailored dietary recommendations [35,38-40]. By incorporating dynamic inputs such as food tracking and activity logs, the system can further provide personalized, context-sensitive advice with timely feedback and adjustments based on patients' health progress [43,54]. Knowledge and decision support functions served as a critical layer in enhancing the intelligence and reliability. For example, by integrating chronic kidney disease calculators [35], nutrient search engines [45], and expert knowledge bases [39], these systems move beyond simple food suggestions to provide evidence-based recommendations. This clarified how the nutritional rationale, which potentially enhances patients' understanding and trust, long-term adherence, and sustainable dietary behavior change [44,45]. The accuracy and relevance of food or diet recommendations depend on the robustness of the algorithm used to process the data and the diversity of food or recipes options integrated into the system [55-57]. Data management and additional interactive functions have enhanced the usability and clinical relevance of diet-related HRSs and supported user engagement, empowering patients to actively and continuously manage their dietary self-care.

Recommendation Content of Diet-Related HRSs

Patients' needs for the content of diet-related HRSs are multidimensional. Beyond simple food lists, an effective diet-related HRS must provide actionable guidance on cooking methods, ingredient combinations, and personalized meal plans [39,42-44,47,48,50]. Well-structured recipes emerge as a key feature, as they help users visualize meal preparation and improve understanding of the nutritional rationale behind recommendations [44]. Such transparency fosters engagement and adherence, particularly when supported by nutrition professionals [58,59]. Additionally, offering varied recipe options allows patients to personalize their meal plans according to their tastes, dietary restrictions, and the availability of ingredients [39]. This flexibility caters to users with complex health conditions and empowers them with a sense of control, which is vital for sustaining long-term dietary management [42].

Implementation of Recommendation Features in Diet-Related HRSs

Recommendation Methods

In diet-related HRSs, hybrid approaches appear particularly valuable for balancing clinical appropriateness with individual preferences to promote long-term adherence. In addition, adaptability is an essential feature, allowing systems to respond to changing health conditions, behaviors, and user needs. As patients' health status and preferences evolve, systems such as "SHARE" demonstrate the value of dynamic updates that allow users to refine their meal plans, making the system more user-centered and responsive [42]. However, a critical gap across the reviewed studies is the insufficient consideration of medication-food interactions. For patients with chronic conditions, clinically significant interactions such as warfarin-vitamin K [60], angiotensin-converting–enzyme inhibitors-potassium [61], or metformin-alcohol [62], carry substantial risks. Future diet-related HRSs should adopt hybrid approaches that integrate drug-nutrient knowledge and dynamic monitoring with preference learning, ensuring recommendations that are both personalized and clinically safe.

Recommendation Technology

Appropriate recommendation technology can transform users' vague needs into clear ones and filter out irrelevant information [63]. HyR that combines CF, CB, and KB methods is particularly effective. These methods overcome traditional limitations such as cold start and data sparsity by integrating multiple information sources, including nutritional knowledge and disease-specific data [64]. Notably, KB and KG-driven algorithms were especially well-suited for diet-related HRSs. By incorporating expert knowledge and medical guidelines, they ensure scientific accuracy, clinical reliability, and alignment with patients' dietary management goals, resulting in more comprehensive and robust systems [65].

Data Sources

The effectiveness of diet-related HRSs largely depends on the quality and appropriateness of data sources. Patient data mainly come from personal input, including medical records, treatment history, and sociodemographic information [35,38-45,47-49]. Incorporating dietary behavior data (eg, meal logs, food purchase records, and subjective reports) and clinical data (eg, diagnosis, treatment, and disease stage) helps ensure that dietary recommendations are both personalized and medically appropriate [66,67]. Although patient-specific data are essential for personalized recommendations [35,44], sensitive medical and behavioral data entail serious privacy, security, and ethical risks. Given patients' fragile health status and dietary restrictions, dietary data from professional resources, such as the Nutrition Division of the Ministry of Public Health (MOPH) [46], the official website of the composition of foods integrated dataset (CoFID) [48], and the Japan Preventive Association of Lifestyle-related Disease (JPALD) [45], are vital for ensuring the accuracy of recommendations. In contrast, nonprofessional

XSL•FO
RenderX

sources such as general cooking websites are not recommended in clinical contexts due to their limited reliability, particularly for patients with complex conditions.

Implementation Process

The Association for Computing Machinery has issued guidelines for designing and evaluating RSs [68]. However, no widely accepted standards have been established for implementing HRSs, especially diet-related HRSs. This gap highlights challenges arising from disease diversity, complex food preferences, and context-dependent factors that are difficult to acquire or simulate [28]. The reviewed systems generally followed a similar workflow, involving user profiling, structured knowledge integration, personalized filtering, and ranking, reflecting a growing understanding of how to align dietary recommendations with individual health needs and preferences. The integration of advanced recommendation technologies, such as ontology-based reasoning [51], optimization algorithms (eg, ant colony methods [48]), Long Short-Term Memory–based dynamic modeling [44], and KGs [39], has advanced more intelligent and context-aware systems. However, the adoption of these technologies remains limited, and few studies have systematically evaluated their impact on recommendation quality or patient outcomes.

Evaluation of Diet-Related HRSs

Evaluation Criteria

In diet-related HRSs, accuracy has been the primary evaluation criterion, typically evaluated by prediction score accuracy, prediction score correlation, classification accuracy, and sorting accuracy [69]. However, the rationality of the recommendations has received limited attention. None of the 15 reviewed studies [35,38-51] evaluated the scientific soundness of food or recipe recommendations. While health care providers were commonly involved in evaluating the recommendation results, official institutional input on content was rare, revealing a major gap in ensuring the credibility and reliability of recommendations.

Beyond accuracy and content rationality, the effectiveness of diet-related HRSs is closely tied to their ability to promote healthy eating behaviors. The key goal of the diet-related HRSs is to empower users to make informed dietary choices and motivate healthier behaviors [70]. Existing evaluations of behavior change have primarily examined improvements in diet quality and diversity, measured by the China Elderly Dietary Guidelines Index and the dietary diversity score [50]. Although user engagement and personalized feedback were recognized as essential for sustaining behavior change, few studies evaluated long-term effects on patients' dietary behaviors.

Moreover, clinical trials in real-world health care settings remain lacking; even the study [39] involving patients was an observational cohort study rather than a clinical trial, limiting statistical power and generalizability. Additionally, no studies examined the cost-effectiveness of implementing diet-related HRSs in health care settings, which is crucial for adoption in resource-constrained healthcare systems.

Evaluation Methods and Process

The evaluation of diet-related HRSs can be carried out using various methods, each with its advantages and limitations. Online evaluations, which assess system performance through real-time user feedback or surveys, provide valuable insights into user interactions but often involve higher costs [22]. Offline evaluations in controlled settings are more feasible but have limited external validity due to their inability to reflect real-world usage [22,71]. Emerging approaches such as online A or B testing enable rapid, cost-effective comparisons of different system versions, helping predict broader performance even with small samples [72,73].

Diet-related HRS evaluations generally follow a staged approach from development to postdeployment. During development, stakeholder involvement in usability evaluation is critical to ensure user-centered design [74]. After development, accurate assessments verify that recommendations are reliable and relevant to target users [35,41,42,44,46-49,51]. Postdevelopment, gathering user feedback on effectiveness and satisfaction is essential for refining the system and improving its responsiveness to user needs [35,39,40,42,43,46]. These stages form a comprehensive evaluation process ensuring the reliability and usability of diet-related HRSs.

Although several reviews have examined HRSs, few have focused specifically on diet-related HRSs. One scoping review of 36 studies identified only 8 diet-focused HRSs [75]. A systematic review of 73 studies identified 26 nutrition-related systems, most of which targeted populations without strict medical dietary requirements, thereby revealing the lack of disease-specific syntheses [22]. Existing diet-related HRS reviews have primarily addressed technical aspects, such as semantic interoperability [76], explainable artificial intelligence [77], and data-driven personalization methods [78]. They offer useful technical insights but rarely address disease-specific dietary needs or the related clinical and safety challenges. In contrast, our review focuses on diet-related HRSs for patients with chronic conditions. It highlights key requirements such as clinical data integration, dynamic health monitoring, tailored recipe-level recommendations, and persistent gaps in evaluating behavior change and clinical outcomes. These disease-focused insights extend prior work and inform the development of more personalized and clinically reliable diet-related HRSs.

Future Directions

Future research should prioritize developing diet-related HRSs based on a user-centered design framework. These systems must be tailored to patients' specific needs with a deep understanding of their cognitive abilities, digital literacy, cultural dietary preferences, and daily routines. Diet-related HRSs need to address clinical challenges through contextually relevant functional designs, customized for different patient groups, from older adults with multimorbidity to pediatric patients requiring caregiver-mediated support. Practical usability features, such as adaptive reminders, contextualized recipe guidance, and dynamic feedback loops, are essential for real-world effectiveness.

XSL•FO
RenderX

To support patients' sustained dietary behavior change, recommendations should move beyond static food lists toward evidence-based, actionable guidance using practical tools such as portion guides, ingredient lists, and meal preparation videos [79]. Future systems should adopt hybrid, adaptive recommendation mechanisms and integrate with electronic health records (EHRs) and mobile apps to enhance personalization, interpretability, clinical relevance, and engagement, which is essential for sustaining long-term adherence to dietary changes. Additionally, incorporating behavior change theories, such as Social Cognitive Theory [80], Theory of Planned Behavior [81], and the Fogg Behavior Model [82], can strengthen the dietary intervention design. Gamification elements informed by these theories, such as rewards and rankings, may further motivate users and increase adherence [83].

When managing sensitive health data, especially when integrating with EHRs and personal medical records, the systems must implement robust data governance, including secure storage, encryption, role-based access, and transparent consent. Compliance with regulations such as the US Health Insurance Portability and Accountability Act (HIPAA) [84], the European Union's General Data Protection Regulation (GDPR) [85], and national standards are essential to safeguard patient privacy and foster user trust. Finally, standardized evaluation methods and robust frameworks, combining clinical trials, behavior change evaluation, and usability evaluation, are essential to verify effectiveness, accuracy, and long-term impact in real-world settings, while ensuring cost-effectiveness and compliance (eg, US Food and Drug Administration approval and European Conformity marking) for safe and scalable healthcare implementation.

Implication for Clinical Practice

For clinicians and dietitians, diet-related HRSs can serve as supportive tools to deliver personalized, evidence-based, and adaptive dietary guidance aligned with patients' medical conditions, treatment regimens, cultural preferences, and daily routines. By incorporating medication-food interaction checks and actionable educational content, these systems can enhance patient safety, dietary adherence, and long-term health management. Future diet-related HRSs could integrate seamlessly into clinical workflows and demonstrate measurable health outcomes. Embedding diet-related HRSs within EHRs would ensure that dietary recommendations are consistent with medications, lab results, and care plans. By linking dietary adherence to physiological indicators such as hemoglobin A_{1c}, blood pressure, or lipid levels, systems can enable real-time feedback and timely clinical intervention. Beyond accuracy, future evaluations should predefine clinical end points (eg, hemoglobin A_{1c} reduction or blood pressure control) and validate effectiveness through randomized controlled trials and real-world studies.

Limitations

This study has several limitations. First, only studies published from January 2010 to October 2024 were included, which may not reflect recent developments. Second, only English and Chinese literature were reviewed, excluding studies in other languages. Third, the focus was primarily on chronic health conditions, with limited exploration of diet-related HRSs for acute conditions or general wellness, suggesting a need for broader research in future studies. In addition, the included studies themselves demonstrated certain methodological limitations, such as small sample sizes, reliance on offline testing, and limited evaluation of actual behavior change. These issues reflect not only the limitations of this review but also highlight gaps in the existing research that warrant more rigorous clinical evaluation.

Conclusions

Diet-related HRSs can offer personalized dietary recommendations for patients with chronic conditions. The analysis reveals significant gaps, demanding that future systems be grounded in user-centered design to meet patient-specific needs. These systems must recommend more practical and actionable dietary guidance. The adoption of hybrid recommendation techniques can enhance the precision and clinical relevance of dietary recommendations. Establishing standardized evaluation metrics and conducting real-world studies with long-term follow-up will be essential. This will verify the system's ability to positively change dietary behaviors and improve clinical outcomes. Addressing these issues will transform diet-related HRSs into trustworthy and impactful tools for managing chronic diseases and delivering patient-centered care.

Acknowledgments

We would like to thank Liping Fu for her assistance and guidance in developing the search strategy.

Generative artificial intelligence and artificial intelligence–assisted technologies in the writing process: during the preparation of this work, the authors used ChatGPT 4.0 to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Funding

This study was supported by the National Natural Science Foundation of China (grant number 72374204) and Soft Science Research Project of Shanghai (25692108400).

XSL•FO
RenderX

Authors' Contributions

Conceptualization: XD, JW
Data curation: XD, BY, HN, TJ, JW
Formal analysis: XD, BY, ZZ, JW
Project administration: XD, JW
Supervision: AP, CY, JW
Visualization: XD
Writing – original draft: XD, BY, ZZ, JW
Writing – review & editing: XD, BY, AP, JW

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[PDF File, 850 KB - jmir_v28i1e77726_app1.pdf]

Multimedia Appendix 2
The publication years and countries of the included studies.
[PNG File, 82 KB - jmir_v28i1e77726_app2.png]

Checklist 1
PRISMA-ScR Checklist.
[PDF File, 114 KB - jmir_v28i1e77726_app3.pdf]

References

1. Murray CJL, Aravkin AY, Zheng P, et al. Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct;396(10258):1223-1249. [doi: 10.1016/S0140-6736(20)30752-2]
2. The Global Nutrition Report. URL: https://globalnutritionreport.org/ [accessed 2025-12-31]
3. Ong KL, Stafford LK, McLaughlin SA, et al. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet 2023 Jul;402(10397):203-234. [doi: 10.1016/S0140-6736(23)01301-6]
4. Magkos F, Hjorth MF, Astrup A. Diet and exercise in the prevention and treatment of type 2 diabetes mellitus. Nat Rev Endocrinol 2020 Oct;16(10):545-555. [doi: 10.1038/s41574-020-0381-5] [Medline: 32690918]
5. Filippou CD, Tsioufis CP, Thomopoulos CG, et al. Dietary Approaches to Stop Hypertension (DASH) diet and blood pressure reduction in adults with and without hypertension: a systematic review and meta-analysis of randomized controlled trials. Adv Nutr 2020 Sep 1;11(5):1150-1160. [doi: 10.1093/advances/nmaa041] [Medline: 32330233]
6. Zhuang P, Wang F, Yao J, et al. Unhealthy plant-based diet is associated with a higher cardiovascular disease risk in patients with prediabetes and diabetes: a large-scale population-based study. BMC Med 2024 Oct 23;22(1):485. [doi: 10.1186/s12916-024-03683-7] [Medline: 39443972]
7. Lockwood G, Davey L, McFarlane C, Gray NA, Wright HH. Factors influencing meal provision and dietary support behaviour of caregivers of people with chronic kidney disease: a cross-sectional study. Nutrients 2024 Oct 14;16(20):3479. [doi: 10.3390/nu16203479] [Medline: 39458474]
8. Xiong H, Zhang X, Zeng H, Xie S, Yi S. Experience of diet in patients with inflammatory bowel disease: a thematic synthesis of qualitative studies. J Clin Nurs 2024 Aug;33(8):3283-3293. [doi: 10.1111/jocn.17186] [Medline: 38661241]
9. Scientific research report on dietary guidelines for Chinese residents [Website in Chinese]. : National Institute for Nutrition and Health, Chinese Center for Disease Control and Prevention; 2021 URL: https://www.chinanutri.cn/yyjkzxpt/yyjkkpzx/yytsg/zgjm/202103/t20210311_224598.html [accessed 2025-12-31]
10. Dahl C, Crichton M, Jenkins J, et al. Evidence for dietary fibre modification in the recovery and prevention of reoccurrence of acute, uncomplicated diverticulitis: a systematic literature review. Nutrients 2018 Jan 27;10(2):137. [doi: 10.3390/nu10020137] [Medline: 29382074]
11. Collins N, Belkaid Y. Control of immunity via nutritional interventions. Immunity 2022 Feb 8;55(2):210-223. [doi: 10.1016/j.immuni.2022.01.004] [Medline: 35139351]
12. Dietary Guidelines for Americans: US Department of Health and Human Services and US Department of Agriculture; 2015. URL: https://odphp.health.gov/sites/default/files/2019-09/2015-2020_Dietary_Guidelines.pdf [accessed 2025-11-05]

XSL•FO
RenderX

13. Flaskerud JH. Mood and food. Issues Ment Health Nurs 2015 Apr;36(4):307-310. [doi: 10.3109/01612840.2014.962677] [Medline: 25988966]
14. Notice on the issuance of dietary guidelines for the prevention and management of hyperuricemia and gout in adults (2024 edition) and three other dietary guidelines [Report in Chinese]. : National Health Commission of the People's Republic of China; 2024 URL: https://www.nhc.gov.cn/wjw/c100378/202402/afd5dda4bd6745fda10aad8d43a16369/files/1732845129551_37712.pdf [accessed 2025-12-31]
15. Rines J, Daley K, Loo S, et al. A patient-led, peer-to-peer qualitative study on the psychosocial relationship between young adults with inflammatory bowel disease and food. Health Expect 2022 Aug;25(4):1486-1497. [doi: 10.1111/hex.13488] [Medline: 35383400]
16. Singh PK, Pramanik PKD, Dey AK, Choudhury P. Recommender systems: an overview, research trends, and future directions. Int J Bus Syst Res 2021;15(1):14. [doi: 10.1504/IJBSR.2021.111753]
17. Su X, Khoshgoftaar TM. A survey of collaborative filtering techniques. Adv Artif Intell 2009 Oct 27;2009(1):1-19. [doi: 10.1155/2009/421425] [Medline: 41092928]
18. Lops P, de Gemmis M, Semeraro G. Content-based recommender systems: state of the art and trends. In: Recommender Systems Handbook: Springer; 2011:73-105. [doi: 10.1007/978-0-387-85820-3_3]
19. Burke R. Knowledge-based recommender systems. Encycl Libr Inf Syst 2000;69(Supplement 32) [FREE Full text]
20. Hou S, Wei D. Research on knowledge graph-based recommender systems. Presented at: 2023 3rd International Symposium on Computer Technology and Information Science (ISCTIS); Jul 7-9, 2023. [doi: 10.1109/ISCTIS58954.2023.10213083]
21. Muñoz EG, Parraga-Alava J, Meza J, Proaño Morales JJ, Ventura S. Housing fuzzy recommender system: a systematic literature review. Heliyon 2024 Mar 15;10(5):e26444. [doi: 10.1016/j.heliyon.2024.e26444] [Medline: 38439861]
22. De Croon R, Van Houdt L, Htun NN, Štiglic G, Vanden Abeele V, Verbert K. Health Recommender Systems: systematic review. J Med Internet Res 2021 Jun 29;23(6):e18035. [doi: 10.2196/18035] [Medline: 34185014]
23. Siriaraya P, Suzuki K, Nakajima S. Utilizing collaborative filtering to recommend opportunities for positive affect in daily life. HealthRecSys@ RecSys 2019 [FREE Full text]
24. Pasta A, Petersen MK, Jensen KJ, Larsen JE. Rethinking hearing aids as recommender systems. CEUR Workshop Proc 2019:11-17 [FREE Full text]
25. Sanchez Bocanegra CL, Sevillano Ramos JL, Rizo C, Civit A, Fernandez-Luque L. HealthRecSys: a semantic content-based recommender system to complement health videos. BMC Med Inform Decis Mak 2017 May 15;17(1):63. [doi: 10.1186/s12911-017-0431-7] [Medline: 28506225]
26. Agu E, Claypool M. Cypress: a cyber-physical recommender system to discover smartphone exergame enjoyment. Presented at: Proceedings of the ACM workshop on engendering health with recommender systems; Sep 15-16, 2016 URL: https://web.cs.wpi.edu/~claypool/papers/cypress-recsys/paper.pdf [accessed 2025-12-31]
27. Chavan P, Thoms B, Isaacs J. A recommender system for healthy food choices: building a hybrid model for recipe recommendations using big data sets. In: Chavan P, Thoms B, Isaacs J, editors. Presented at: Hawaii International Conference on System Sciences; Jan 5-8, 2021. [doi: 10.24251/HICSS.2021.458]
28. Mahajan P, Kaur PD. A systematic literature review of food recommender systems. SN Comput Sci 2024;5(1):1-26. [doi: 10.1007/s42979-023-02537-y]
29. Trattner C, Elsweiler D. Food recommender systems: important contributions, challenges and future research directions. arXiv. Preprint posted online on Nov 7, 2017. [doi: 10.48550/arXiv.1711.02760]
30. Islam T, Joyita AR, Alam M, Mehedi Hassan M, Hassan M, Gravina R. Human-behavior-based personalized meal recommendation and menu planning social system. IEEE Trans Comput Soc Syst 2022;10(4):2099-2110. [doi: 10.1109/TCSS.2022.3213506]
31. Vandeputte J, Herold P, Kuslii M, et al. Principles and validations of an artificial intelligence-based recommender system suggesting acceptable food changes. J Nutr 2023 Feb;153(2):598-604. [doi: 10.1016/j.tjnut.2022.12.022] [Medline: 36894251]
32. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
33. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
34. Song HS, Kim YA. A dog food recommendation system based on nutrient suitability. Expert Systems 2021 Mar;38(2). [doi: 10.1111/exsy.12623]
35. Agapito G, Simeoni M, Calabrese B, et al. DIETOS: a dietary recommender system for chronic diseases monitoring and management. Comput Methods Programs Biomed 2018 Jan;153:93-104. [doi: 10.1016/j.cmpb.2017.10.014] [Medline: 29157465]
36. Agapito G, Calabrese B, Guzzi PH, et al. DIETOS: a recommender system for adaptive diet monitoring and personalized food suggestion. Presented at: 2016 IEEE 12th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob); Oct 17-19, 2016. [doi: 10.1109/WiMOB.2016.7763190]
37. Wang J, Yuan C. Evidence-based scoping reviews: an introduction to the methodology. J Nurs Sci 2017;32(8):103-105 [FREE Full text] [doi: 10.3870/j.issn.1001-4152.2017.08.103]

XSL•FO
RenderX

38. Faiz I, Mukhtar H, Khan S. An integrated approach of diet and exercise recommendations for diabetes patients. Presented at: 2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom 2014); Oct 15-18, 2014. [doi: 10.1109/HealthCom.2014.7001899]
39. Xu Z, Gu Y, Xu X, et al. Developing a personalized meal recommendation system for Chinese older adults: observational cohort study. JMIR Form Res 2024 May 30;8:e52170. [doi: 10.2196/52170] [Medline: 38814702]
40. Qi M, Chen W, Li D, Yang F. Design of a knowledge-based nutritional meal recommendation system for patients. China Med Equip 2021;18(1). [doi: 10.3969/J.ISSN.1672-8270.2021.01.026]
41. Rathi M, Pareek V. Mobile based healthcare tool an integrated disease prediction & recommendation system. Int J Knowl Syst Sci 2019;10(1):38-62. [doi: 10.4018/IJKSS.2019010103]
42. Zioutos K, Kondylakis H, Stefanidis K. Healthy personalized recipe recommendations for weekly meal planning. Computers 2024;13(1):1. [doi: 10.3390/computers13010001]
43. Tseng JCC, Lin BH, Lin YF, et al. An interactive healthcare system with personalized diet and exercise guideline recommendation. Presented at: 2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI); Nov 20-22, 2015. [doi: 10.1109/TAAI.2015.7407106]
44. Tang J, Huang B, Xie M. Anticancer recipe recommendation based on cancer dietary knowledge graph. Eur J Cancer Care (Engl) 2023 Oct 18;2023:1-13. [doi: 10.1155/2023/8816960]
45. Ting YH, Zhao Q, Chen RC. Dietary recommendation based on recipe ontology. Presented at: 2014 IEEE 6th International Conference on Awareness Science and Technology (iCAST); Oct 29-31, 2014. [doi: 10.1109/ICAwST.2014.6981829]
46. Phanich M, Pholkul P, Phimoltares S. Food recommendation system using clustering analysis for diabetic patients. Presented at: 2010 International Conference on Information Science and Applications; Apr 21-23, 2010. [doi: 10.1109/ICISA.2010.5480416]
47. Manoharan D. Patient diet recommendation system using K clique and deep learning classifiers. J Artif Intell Capsule Netw 2020;2(2):121-130. [doi: 10.36548/jaicn.2020.2.005]
48. Rehman F, Khalid O, Haq N, Khan AUR, Bilal K, Madani S. Diet-Right: a smart food recommendation system. KSII TIIS 2017;11(6):2910-2925. [doi: 10.3837/tiis.2017.06.006]
49. Chen RC, Ting YH, Chen JK, Lo YW. The nutrients of chronic diet recommended based on domain ontology and decision tree. Presented at: 2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI); Dec 2-4, 2015. [doi: 10.1109/TAAI.2015.7407127]
50. Elsweiler D, Harvey M. Towards automatic meal plan recommendations for balanced nutrition. Presented at: RecSys '15; Sep 16-20, 2015; Vienna, Austria p. 313-316. [doi: 10.1145/2792838.2799665]
51. Arwan A, Sidiq M, Priyambadha B, Kristianto H, Sarno R. Ontology and semantic matching diabetic food recommendations. Presented at: 2013 International Conference on Information Technology and Electrical Engineering (ICITEE); Oct 7-8, 2013. [doi: 10.1109/ICITEED.2013.6676233]
52. Wang Y, Wang J, Yu S, Zhou F, Yuan C. Qualitative study on the needs of parents of children with newly diagnosed acute lymphoblastic leukemia. J Nurs PLA China 2016;33(4):6-10 [FREE Full text]
53. van den Brink M, Ter Hedde MM, van den Heuvel E, Tissing WJE, Havermans RC. The impact of changes in taste, smell, and eating behavior in children with cancer undergoing chemotherapy: a qualitative study. Front Nutr 2022;9:984101. [doi: 10.3389/fnut.2022.984101] [Medline: 36245523]
54. Zenun Franco R, Fallaize R, Weech M, Hwang F, Lovegrove JA. Effectiveness of web-based personalized nutrition advice for adults using the eNutri web app: evidence from the EatWellUK randomized controlled trial. J Med Internet Res 2022 Apr 25;24(4):e29088. [doi: 10.2196/29088] [Medline: 35468093]
55. Celis-Morales C, Livingstone KM, Marsaux CF, et al. Effect of personalized nutrition on health-related behaviour change: evidence from the Food4Me European randomized controlled trial. Int J Epidemiol 2017 Apr 1;46(2):578-588. [doi: 10.1093/ije/dyw186] [Medline: 27524815]
56. Ordovas JM, Ferguson LR, Tai ES, Mathers JC. Personalised nutrition and health. BMJ 2018 Jun 13;361:bmj.k2173. [doi: 10.1136/bmj.k2173] [Medline: 29898881]
57. Melese A. Food and restaurant recommendation system using hybrid filtering mechanism. North Am Acad Res (NAAR) 2021;4(4):268-281. [doi: 10.5281/zenodo.4712849]
58. Worthington A, Coffey T, Gillies K, Roy R, Braakhuis A. Exploring how researchers consider nutrition trial design and participant adherence: a theory-based analysis. Front Nutr 2024;11:1457708. [doi: 10.3389/fnut.2024.1457708] [Medline: 39742103]
59. Mattei J, Alfonso C. Strategies for healthy eating promotion and behavioral change perceived as effective by nutrition professionals: a mixed-methods study. Front Nutr 2020;7:114. [doi: 10.3389/fnut.2020.00114] [Medline: 32923451]
60. Holbrook AM, Pereira JA, Labiris R, et al. Systematic overview of warfarin and its drug and food interactions. Arch Intern Med 2005 May 23;165(10):1095-1106. [doi: 10.1001/archinte.165.10.1095] [Medline: 15911722]
61. Palmer BF. Managing hyperkalemia caused by inhibitors of the renin-angiotensin-aldosterone system. N Engl J Med 2004 Aug 5;351(6):585-592. [doi: 10.1056/NEJMra035279] [Medline: 15295051]
62. Bailey CJ, Turner RC. Metformin. N Engl J Med 1996 Feb 29;334(9):574-579. [doi: 10.1056/NEJM199602293340906] [Medline: 8569826]

XSL•FO
RenderX

63. Joy J, Pillai RVG. Review and classification of content recommenders in E-learning environment. J King Saud Univ, Comput Inf Sci 2022 Oct;34(9):7670-7685. [doi: 10.1016/j.jksuci.2021.06.009]
64. Ayemowa MO, Ibrahim R, Bena YA. A systematic review of the literature on deep learning approaches for cross-domain recommender systems. Decision Analytics Journal 2024 Dec;13:100518. [doi: 10.1016/j.dajour.2024.100518]
65. Abhari S. A systematic review of nutrition recommendation systems: with focus on technical aspects. J Biomed Phys Eng 2019 Dec;9(6). [doi: 10.31661/JBPE.V0I0.1248]
66. Arakaki M, Li L, Kaneko T, et al. Personalized nutritional therapy based on blood data analysis for malaise patients. Nutrients 2021 Oct 18;13(10):3641. [doi: 10.3390/nu13103641] [Medline: 34684641]
67. Shim JS, Oh K, Kim HC. Dietary assessment methods in epidemiologic studies. Epidemiol Health 2014;36:e2014009. [doi: 10.4178/epih/e2014009] [Medline: 25078382]
68. Zangerle E, Bauer C. Evaluating recommender systems: survey and framework. ACM Comput Surv 2023 Aug 31;55(8):1-38. [doi: 10.1145/3556536]
69. Zhu Y, Lyu L. Review of evaluation metrics for recommender systems. J Univ Electron Sci Technol China 2012;41(2):163-175. [doi: 10.3969/j.issn.1001-0548.2012.02.001]
70. Wang J, Dong X, Jin T, Yuan C. Research progress and prospects of health recommendation systems. J Med Intell 2024;45(1):70-76 [FREE Full text]
71. Rossetti M, Stella F, Zanker M. Contrasting offline and online results when evaluating recommendation algorithms. Presented at: RecSys ’16; Sep 15-19, 2016. [doi: 10.1145/2959100.2959176]
72. Kohavi R, Longbotham R. Online controlled experiments and a/b testing. In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning and Data Mining: Springer; 2017:922-929. [doi: 10.1007/978-1-4899-7687-1_891]
73. Georgiev GZ. Statistical Methods in Online A/B Testing: Statistics for Data-Driven Business Decisions and Risk Management in e-Commerce: Georgi Z: Georgiev; 2019.
74. Sun Y, Zhou J, Ji M, Pei L, Wang Z. Development and evaluation of health recommender systems: systematic scoping review and evidence mapping. J Med Internet Res 2023 Jan 19;25:e38184. [doi: 10.2196/38184] [Medline: 36656630]
75. Ananthakrishnan A, Milne-Ives M, Cong C, et al. The evaluation of health recommender systems: a scoping review. Int J Med Inform 2025 Mar;195:105697. [doi: 10.1016/j.ijmedinf.2024.105697] [Medline: 39608231]
76. Xhani D, Sedrakyan G, Gavai A, Guizzardi R, van Hillegersberg J. The role and applications of semantic interoperability tools and eXplainable AI in the development of smart food systems: findings from a systematic literature review. Intell Syst Appl 2025 Sep;27:200547. [doi: 10.1016/j.iswa.2025.200547]
77. Kalu KA, Ataguba G, Onifade O, Orji F, Giweli N, Orji R. Application of artificial intelligence technologies as an intervention for promoting healthy eating and nutrition in older adults: a systematic literature review. Nutrients 2025 Oct 14;17(20):3223. [doi: 10.3390/nu17203223] [Medline: 41156474]
78. Tsolakidis D, Gymnopoulos LP, Dimitropoulos K. Artificial intelligence and machine learning technologies for personalized nutrition: a review. Informatics (MDPI) 2024;11(3):62. [doi: 10.3390/informatics11030062]
79. Anderson HL, Moore JE, Millar BC. Comparison of innovative communication approaches in nutrition to promote and improve health literacy. Ulster Med J 2022 May;91(2):85-91. [Medline: 35722219]
80. Locke EA, Bandura A. Social foundations of thought and action: a social-cognitive view. Acad Manag Rev 1987 Jan;12(1):169. [doi: 10.2307/258004]
81. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
82. Fogg BJ. Persuasive technology: using computers to change what we think and do. Ubiquity 2002;2002. [doi: 10.1145/764008.763957]
83. Wu X, Xin X. The application of gamification design in persuading users’ behavior change [Article in Chinese]. Packag Eng 2017;38(20):194-198 [FREE Full text]
84. Cole LJ, Fleisher LD. Update on HIPAA privacy: are you ready? Genet Med 2003;5(3):183-186. [doi: 10.1097/01.GIM.0000068625.72823.86] [Medline: 12792427]
85. Voigt P, von dem Bussche A. The EU General Data Protection Regulation (GDPR): A Practical Guide: Springer Publishing Company, Incorporated; 2017. [doi: 10.1007/978-3-319-57959-7]

Abbreviations

CB: content-based
CF: collaborative filtering
CoFID: composition of foods integrated dataset
EHR: electronic health record
GDPR: General Data Protection Regulation
HIPAA: Health Insurance Portability and Accountability Act
HRS: Health Recommender System
HyR: hybrid recommendation

XSL•FO
RenderX

JPALD: Japan Preventive Association of Lifestyle-related Disease
KB: knowledge-based
KG: knowledge graph
MOPH: Ministry of Public Health
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
RS: Recommender System

Edited by N Cahill; submitted 19.May.2025; peer-reviewed by A Berihun, C Okolue, T Ojo; accepted 01.Dec.2025; published 14.Jan.2026.

Please cite as:
Dong X, Yun B, Pakarinen A, Zheng Z, Niu H, Jin T, Yuan C, Wang J
Diet-Related Health Recommender Systems for Patients With Chronic Health Conditions: Scoping Review
J Med Internet Res 2026;28:e77726
URL: https://www.jmir.org/2026/1/e77726
doi:10.2196/77726

© Xiaolan Dong, Bei Yun, Anni Pakarinen, Zhuting Zheng, Hao Niu, Tian Jin, Changrong Yuan, Jingting Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI-Supported Digital Microscopy Diagnostics in Primary Health Care Laboratories: Scoping Review

Joar von Bahr[1,2,3], MD; Antti Suutala[3], MSc; Vinod Diwan[1], MD, PhD; Andreas Mårtensson[2,4], MD, PhD; Johan Lundin[1,3*], MD, PhD; Nina Linder[2,3*], MD, PhD

[1]Department of Global Public Health, Karolinska Institutet, Tomtebodavägen 18 A, Solna, Sweden

[2]Global Health and Migration Unit, Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden

[3]Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland

[4]Department of Infectious Diseases, Uppsala University Hospital, Uppsala, Sweden

[*]these authors contributed equally

Corresponding Author:
Joar von Bahr, MD
Department of Global Public Health, Karolinska Institutet, Tomtebodavägen 18 A, Solna, Sweden

Abstract

Background: Digital microscopy combined with artificial intelligence (AI) is increasingly being implemented in health care, predominantly in advanced laboratory settings. However, AI-supported digital microscopy could be especially advantageous in primary health care settings, since such methods could improve access to diagnostics via automation and a decreased need for experts on-site. To our knowledge, no scoping or systematic review has previously examined the use of AI-supported digital microscopy in primary health care laboratories, and a scoping review could guide future research by providing insights into the challenges of implementing these novel methods.

Objective: This scoping review aimed to map published peer-reviewed studies on AI-supported digital microscopy in primary health care laboratories to generate an overview of the subject.

Methods: A systematic search of the databases PubMed, Web of Science, Embase, and IEEE was conducted on October 2, 2024. The inclusion criteria in the scoping review were based on 3 concepts: using digital microscopy, AI, and comparison of the results with a standard diagnostic system, and 1 context, being performed in primary health care laboratories. Additional inclusion criteria were peer-reviewed diagnostic accuracy studies published in English, performed on humans and achieving a sample-level diagnosis. The study selection and data extraction were performed by 2 independent researchers (JVB and AS), and cases of disagreement were resolved through discussion with a third researcher (NL). The methodology is in accordance with the Joanna Briggs Institute methodology for scoping reviews.

Results: A total of 3403 papers were screened during the paper identification process, of which 22 (0.6%) were included in the scoping review. The samples analyzed were as follows: blood (n=12) for blood cell and malaria detection, urine (n=4) for urinalysis and parasite detection, cytology of atypical oral (n=1) and cervical cells (n=2), stool (n=2) for parasite detection, and sputum (n=1) for ferning patterns indicating inflammation. Both conventional (n=15) and specifically developed methods (n=7) were used in sample preparation. The AI-supported digital microscopy achieved comparable diagnostic accuracy to the reference standard for complete blood counts, malaria detection, identification of stool and genitourinary parasites, screening for oral and cervical cellular atypia, detection of pulmonary inflammation, and urinalysis. Furthermore, AI-supported digital microscopy achieved higher sensitivity than manual microscopy in 6/7 (85.7%) studies that used a reference standard that allowed for this comparison.

Conclusions: AI-supported digital microscopy achieved comparable diagnostic accuracy to the reference standard for diagnosing multiple targets in primary health care laboratories and may be particularly advantageous for improving diagnostic sensitivity. With further research addressing challenges such as scalability and cost-effectiveness, AI-supported digital microscopy could improve access to diagnostics, especially in expert-scarce and resource-limited settings.

International Registered Report Identifier (IRRID): RR2-10.2196/58149

(J Med Internet Res 2026;28:e78500) doi:10.2196/78500

KEYWORDS

AI; artificial intelligence; convolutional neural network; deep learning; diagnosis; digital diagnostics; machine learning; pathology; primary health care; whole slide images

XSL•FO
RenderX

Introduction

Background

Artificial intelligence (AI) in the form of machine learning has successfully been applied to image-based diagnostics within several medical fields [1]. In parallel, manual microscopy remains a cornerstone of diagnostic practice in resource-limited settings and at the primary health care (PHC) level due to its low cost, versatility, and ability to provide direct visualization of pathogens and cellular changes. It is widely used for the diagnosis of infectious diseases such as malaria and intestinal parasitic infections, as well as for full blood counts and analysis of cervical and oral cytological samples and fine needle aspirates [2]. Despite its usefulness and broad applicability, microscopy is highly dependent on the availability of trained personnel and adequate infrastructure, which are often limited in such settings, leading to variability in diagnostic quality and coverage [3]. These limitations have motivated the development of AI-driven approaches, where deep learning methods can assist or automate microscopy-based diagnostics to improve accuracy and accessibility. Deep learning approaches, particularly convolutional neural networks (CNNs) and vision transformers, have become the dominant architectures for image classification and interpretation in medical imaging [4]. CNNs extract visual features, enabling recognition of complex structures such as cells, pathogens, and tissue patterns, while vision transformers can capture contextual relationships between distant structures [4,5].

Leveraging these methods for AI-based microscopy within laboratory workflows has the potential to automate processes, increase productivity, and improve diagnostic accuracy [6]. Multiple AI-based diagnostic systems have been approved for clinical use, for example, for cervical cancer screening and prostate cancer diagnostics [6-8]. Most of these AI-based diagnostic systems depend on expensive, high-end digital imaging instruments and require advanced laboratory infrastructure and are therefore not feasible for use in PHC laboratories [6,7]. However, the development of less expensive, portable digital microscope scanners has enabled research on the use of AI-supported diagnostic systems suitable for PHC laboratories [9-11].

A PHC laboratory, also known as a tier 1 laboratory, can be defined as a laboratory primarily serving outpatients by providing point-of-care (POC) tests and manual microscopy of specimens with simple preparations. An additional responsibility is preparing fine needle aspirations and other simple tissue specimens that are later dispatched to a tier 2 laboratory in a first-level hospital for analysis. The PHC laboratories work with a small budget compared with more advanced laboratories and are generally managed by a laboratory technician supervised by a pathologist from a distance [2].

The World Health Organization has emphasized the importance of providing diagnostics near the patient to enhance the accuracy and timeliness of diagnoses, improve clinical decision-making, and reduce the risk of diagnostic errors [12]. The implementation of AI-supported digital microscopy could help address these challenges at PHC laboratories. To begin with, since PHC laboratories lack access to pathology expertise, application of AI could enable more analyses on-site, consequently increasing both the availability and speed of diagnostics [2,13]. Increased speed and access to diagnostics through AI and telemedicine could reduce health inequities by strengthening diagnostic capacity, particularly in low- and middle-income countries (LMICs) and also in sparsely populated regions of high-income countries [11,14,15]. In addition, a systematic review showed that the implementation of AI-supported diagnostics for microscopy increased the effectiveness of laboratory personnel [6]. Although there is a global shortage of microscopy experts, the shortage of these specialists is more severe in LMICs; therefore, AI-supported digital microscopy may be especially advantageous in strengthening health systems and reducing the diagnostic gaps in these settings [11,16].

There are several diseases where AI-supported digital microscopy diagnostics in PHC laboratories could be advantageous, and studies have been performed on, for example, screening of oral and cervical cancer as well as targeting parasitic infections, such as schistosomiasis and infections caused by soil-transmitted helminths [9,17-19]. Although the targeted diseases differed in these studies, researchers often encountered similar challenges due to commonalities in the methodologies applied, and a review mapping these challenges could provide valuable insights.

A preliminary search of the databases PubMed and Cochrane was performed to investigate whether any scoping or systematic review had been performed on AI-supported digital microscopy in PHC laboratories. A few related reviews were found. One systematic review of AI diagnostics for oral cancer [20] overlaps to some extent with our review; however, since it focuses on a single disease, it does not provide an overview of the development of AI-supported digital microscopy in PHC laboratories. Another systematic review evaluating the application of AI to whole slide images of tissue samples stained with hematoxylin and eosin was also identified [21]. This paper presents the current state of knowledge on AI implementation in pathology within high-end laboratories.

While these reviews are similar to this scoping review, they do not provide an overview of which diseases have been investigated in AI-supported digital microscopy and the disease-agnostic challenges faced in PHC laboratories. Furthermore, the development of more affordable scanners and improved AI, along with persistent workforce and resource constraints, makes a scoping review timely. A scoping review performed on AI-supported digital microscopy in PHC laboratories would, therefore, provide a valuable overview of the subject and collate knowledge that could guide future implementation.

This scoping review aimed to systematically review published peer-reviewed studies that have been performed related to AI-supported digital microscopy in PHC laboratories and specifically address the following questions: (1) In which diseases and for which conditions and targets has AI-based microscopy been applied for diagnostics within PHC laboratories? (2) What methods have been used in acquiring microscopy images to train and analyze AI models for

XSL•FO
RenderX

diagnostics? (3) What AI models and training approaches have been applied? (4) How has the AI-supported diagnostic system performed compared with expert microscopists with regard to diagnostic accuracy?

Review Question

What peer-reviewed studies have been published on implementing AI-supported digital microscopy in PHC laboratories? What methods have been used, what issues have been faced, and what results have been achieved?

Methods

Study Design

The scoping review was conducted in accordance with the Joanna Briggs Institute methodology for scoping reviews updated in 2020 [22]. A PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) checklist is included [23]. A protocol was initially published in the Open Science Framework and later in the peer-reviewed journal JMIR Research Protocols [24,25]. The inclusion and exclusion criteria are shown in Table 1.

Table . Inclusion and exclusion criteria for identified studies.

Study characteristic	Inclusion criteria	Exclusion criteria
Language	• English	• Non-English
Study design	• Published peer-reviewed studies • Diagnostic test accuracy studies	• Non–peer reviewed studies • Not diagnostic test accuracy studies
Population	• Humans	• Studies performed on animals
Concept	• AI[a] techniques applied as a diagnostic tool on microscopy • Final slide-level diagnosis was performed and compared with a standard microscopist • Outcome valuable for clinicians	• Studies that applied AI models on images not conventionally analyzed in microscopy • No final slide diagnosis
Context	• Performed at primary health care laboratory (tier 1 laboratory) • No pathologist needed on site • Samples such as stool, urine, blood, cytology smears, and fine needle aspirations of superficial tissue (eg, from breast lumps) prepared with simple methods	• Studies performed in an advanced laboratory setting

[a]AI: artificial intelligence.

Eligibility Criteria

Participants

This scoping review considered studies on human participants. No exclusion was made based on age, sex, economic status, or nationality.

Concept

The studies included in this scoping review fulfilled 3 concept criteria. First, the studies needed to be performed on images obtained with an imaging instrument built to automatically capture microscopy sample areas large enough for diagnostic purposes. Furthermore, the imaging instrument used must be operated in a way that does not require human expertise to determine what areas of the slide should be captured. Microscopy was defined as deploying a light source, optical lenses, and a digital camera to acquire a magnified image of a biological sample, generating an image conventionally interpreted by a microscopist.

Second, the studies needed to use AI when analyzing the microscopy images. AI was defined as a computer system that is trained to perform a task that typically requires human intelligence. No exclusion was made based on the architecture of the AI model or the dataset used for training. This analysis of the microscopy images could be performed on-site or in a remote cloud environment.

Third, the studies needed to compare the AI-supported diagnostic system with a standard diagnostic system. A diagnostic system was defined as all the steps included in the diagnostic process, from sample collection to the acquisition of results. The result needed to be sufficient to reach a diagnosis at the subject level.

Context

The included studies needed to be performed in a PHC laboratory setting. To be defined as a PHC laboratory, also known as a tier 1 laboratory, the laboratory needed to fulfill 2 criteria. First, regarding staffing, the laboratory must be run by a laboratory technician, not requiring a pathologist on-site. Second, the sample preparations could not exceed the capabilities of a PHC laboratory. Acceptable samples collected included stool, urine, blood, cytology smears, and fine needle aspirations of superficial and easily accessible tissues (eg, from breast lumps and superficial lymph nodes). The sample staining procedure must be possible to perform manually without advanced laboratory equipment such as a microtome or tissue

XSL•FO
RenderX

processor [2]. Sample procedures that fulfill these criteria include Kato-Katz thick stool smears, blood smears, centrifuged urine samples, Papanicolaou-stained cervical or oral smears, and hematoxylin and eosin–stained fine needle cytology smears [2]. Since the context of PHC laboratories in this scoping review is based on human medicine, the exclusion criteria and initial search strategy were changed to exclude veterinary medicine, which was included in the initial protocol published on Open Science Framework [24]. This adjustment was made before submitting the protocol to JMIR Research Protocols to focus the scoping review specifically on challenges in implementing AI-supported microscopy in human health care [25].

Types of Sources

All types of diagnostic test accuracy studies were included. Because data collection in diagnostic test accuracy studies can be both retrospective and prospective, studies using either approach were included. In addition, studies using both paired and random designs for reference standards were included [26]. The included studies had to be published in English.

Search Strategy

The search strategy was designed to identify peer-reviewed published papers. An initial limited search of PubMed and Cochrane was undertaken to identify papers on the topic. Search blocks were created for the final search based on terms used in the identified papers. The search blocks were developed to find papers containing the 2 concepts, microscopy and AI, as well as the context specification of being in a PHC setting, with 1 block created for each. The databases searched were PubMed, Web of Science, Embase, and IEEE, and a detailed description of the search strategy is given in Multimedia Appendix 1. The search was performed on October 2, 2024. The reference lists and all the papers citing the included papers were gathered through the SpiderCite tool on December 3, 2024, and included in the review process [27].

Study Selection

Following the search, all identified papers were compiled in a reference management software system, Zotero (version 6.0.20, Digital Scholar; January 13, 2023, opensource) and duplicates removed. Following the pilot test, titles and abstracts were screened by 2 independent reviewers (JvB and AS) for assessment against the inclusion and exclusion criteria using Covidence systematic review software (Veritas Health Innovation, 2024) [28]. During this step, the Cohen κ agreement was 0.59. All disagreements between JvB and AS were resolved by NL, who provided the deciding vote and could consult the other screeners for their rationale. Thereafter, the full texts of the remaining papers were assessed in detail against the inclusion criteria by 2 independent reviewers (JvB and AS). During full text screening, the Cohen κ agreement was 0.75 for the database search and 0.36 for the citation search. Two reasons caused 17 out of 21 disagreements in the citation search and were resolved through discussions between JvB, AS, JL, and NL. The first issue concerned whether urine analyzers such as Iris iQ200 or Sysmex UF-100 fulfilled the PHC criteria: it was concluded that they did not, as these devices perform advanced sample preprocessing within the machine [29]. The second issue concerned whether handcrafted feature classification qualified as AI: it was concluded that it did not, as it does not involve AI training. With these 2 issues resolved, the citation search had a Cohen κ agreement of 0.79.

Data Charting and Synthesis

Data were extracted from the studies included in the scoping review by 2 reviewers (JvB and AS) using a data extraction tool developed with Covidence systematic review software. The predeveloped extraction tool can be found in Multimedia Appendix 2. Initially, the extraction was performed by JvB. Afterward, the extracted information was checked by AS. All disagreements were resolved through discussion between JvB and AS. When questions arose regarding an original paper, the corresponding author of that manuscript was contacted. The findings are presented narratively and additionally in a table format based on the extraction tool. The information from the extraction tool was split into 3 tables and 1 figure to increase readability. The figure contains a simplified overview of the studies, the first table summarizes the process from sample collection to scanning, the second table shows information on the AI analysis pipeline and training data, and the third table reports the study outcomes. Based on the information extracted to the tables, a narrative description was written to provide an overview of the mapped information. The studies were grouped based on the sample type investigated and the disease targeted as per the first objective of the study: to map in which diseases and for which conditions and targets has AI-based microscopy been applied for diagnostics within PHC laboratories.

Critical Appraisal of Results

The QUADAS-2 tool was applied to investigate the bias of the included studies. This tool was developed to assess the risk of bias for diagnostic accuracy studies in 4 areas: patient selection, index test, reference standard, and flow and timing [30]. The results are shown in the “Results” section, and the form used can be found in Multimedia Appendix 3.

Results

Overview

In total, 3403 papers were screened during the paper identification process, of which 22 (0.6%) were included in the scoping review. The results of the search and the study inclusion process are reported in full in a PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) flow diagram (Figure 1) [31].

XSL•FO
RenderX

Figure 1. Flowchart for study inclusion.

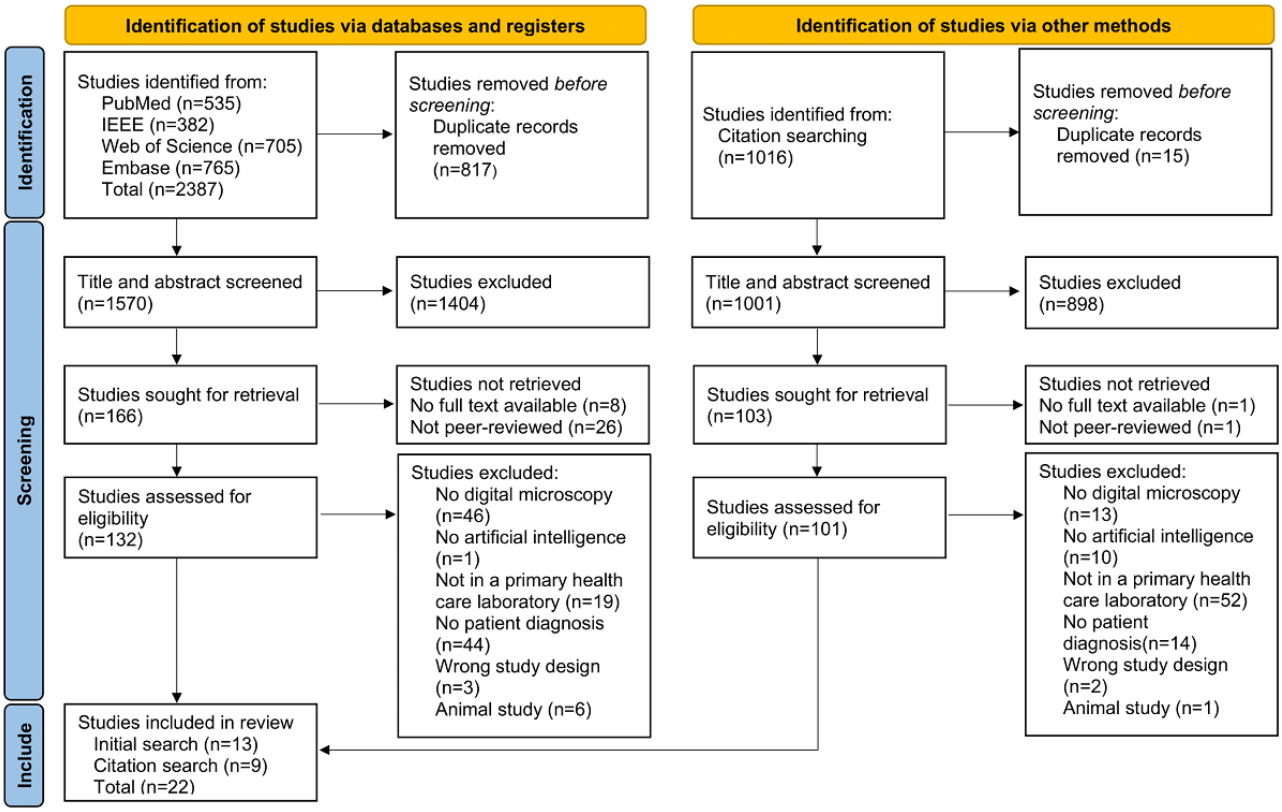

The oldest included study was published in 2014, while the remaining studies were published in 2018 or later, with 9 out of the 22 (40.9%) studies published in 2024. The papers were published in 15 different journals with the most common being *Malaria Journal* (n=4) and *PLOS One* (n=4). The most analyzed samples were blood (n=12), followed by urine (n=4), cytology (n=3), stool (n=2), and sputum (n=1). Different parasites (malaria, intestinal, or genitourinary parasites) were the most common targets (n=13), followed by blood cells (n=4), atypical cervical cells (n=2), atypical oral cells (n=1), and urine particles, such as cells (n=1) and crystalline ferning patterns in sputum (n=1). Detection of these targets was used for multiple diseases and conditions; complete blood counts (CBCs) and urinalysis were used for both organ-specific and systemic diseases, parasite detection for corresponding infections, atypical cells for screening and detection of cancer, and ferning patterns for identifying pulmonary inflammation in patients with COVID-19. An overview of all studies is shown in Figure 2.

XSL•FO
RenderX

Figure 2. Overview of the included studies. AI better than microscopy: White = No comparison, Yes = Higher/same sensitivity and specificity, Mix = Higher sensitivity and lower specificity, and No = Lower sensitivity and specificity. Number of samples in the test set: k=1000. Conditional formatting was applied to all numerical values, with high values shaded green and low values shaded yellow. AI: artificial intelligence; CBC: complete blood count; D: Downey cells; F: Ferning patterns indicative of inflammation; S: sputum; U: urinalysis.

Study Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Samples	Blood												Cytology			S	Stool		Urine			
Targets	CBC			D	Malaria parasite								Cell atypia			F	Parasite		U	Parasite		
Conventional prep.				✓		✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
Custom-built scanner						✓										✓	✓		✓	✓	✓	✓
Single AI algorithm		✓	✓	✓		✓					✓			✓		✓			✓	✓	✓	✓
Human verification		✓	✓		✓			✓			✓											
Reference: microscopy				✓			✓		✓		✓	✓	✓	✓				✓		✓	✓	
Sensitivity %				57	90		95	83	91	72	95	87		100	89	94	86	80	81	87	94	63
Specificity %				100	96		91	94	76	85	100	100		78	100	96	100	95	88	49	94	79
Pearson correlation	0.94	0.94	0.91			0.90																
AI better than microscopy					Mix			Mix		Mix					Yes		Yes		No			Mix
No. samples in test set	679	550	450	31	190	27	488	1.2k	2.3k	400	31	35	307	361	30	160	73	792	240	487	65	339

Sample Preparation and Scanning

Out of the 22 included studies, 12 relied solely on manual preparation methods and 3 used centrifuges. The remaining studies (n=7) used cartridges that simplified and eliminated manual steps. Both in-house–built and commercially available scanners such as Grundium, MiLab, and Motic EasyScan GO were used. The lowest numerical aperture used was 0.1 and the highest was 1.4. Several scanners used both autofocus algorithms and z-stacking to avoid out-of-focus areas. In the 5 studies reporting the time from sample collection to diagnosis using AI-supported digital microscopy, it was 20 - 40 minutes, but there was also a study reporting that it took more than 50 minutes for the scanning and AI analysis (Table 2) and a more detailed table in Multimedia Appendix 4.

XSL•FO
RenderX

Table . Time for analysis and sample processing for the included studies.

Study	Sample	Target	Sample preparation	Sample scanning	Time for analysis
Bachar et al (2021) [32]	Blood	CBC[a]	Cartridge with 2 stains	No retrievable magnification and resolution	No retrievable information
Gasparin et al (2023) [33]	Blood	CBC	Dual-chamber cartridge with 2 stains	No retrievable magnification and resolution	Total: 30 - 40 minutes
Gasparin et al (2022) [34]	Blood	CBC	Dual-chamber cartridge with 2 stains	No retrievable magnification and resolution	Total: 30 - 40 minutes
Akisin et al (2023) [35]	Blood	Downey cells	Manual blood smears stained with May-Grünwald and Giemsa	100× with oil immersion	No retrievable information
Hamid et al (2024) [36]	Blood	Malaria parasites	Cartridge with Giemsa staining	A resolution similar to 50× microscopy [37]	Total: <30 minutes
Holmström et al (2020) [38]	Blood	Malaria parasites	Manual blood smears stained with DAPI[b]	A resolution of 0.9 µm	No retrievable information
Bae et al (2024) [37]	Blood	Malaria parasites	Cartridge with Giemsa staining	A resolution similar to 50×	Total: <30 minutes [36,39]; Scanning: 7 - 10 minutes
Ewnetu et al (2024) [39]	Blood	Malaria parasites	Cartridge with Giemsa staining	A resolution similar to 50× [37]	Total: circa 20 minutes
Das et al (2022) [40]	Blood	Malaria parasites	Manual blood smears stained with Giemsa	40× (NA[c] 0.75)	Scanning and AI[d] analysis: 20 - 30 minutes
Torres et al (2018) [41]	Blood	Malaria parasites	Manual blood smears stained with Giemsa	100× with oil immersion (NA 1.25)	No retrievable information
Linder et al (2014) [42]	Blood	Malaria parasites	Thin blood smears stained with Giemsa	63× with oil immersion (NA 1.4)	No retrievable information
Horning et al (2021) [43]	Blood	Malaria parasites	Manual blood smears stained with Giemsa	40x (NA 0.75)	Scanning and AI analysis 54 minutes
Stegmüller et al (2024) [44]	Cervical cytology	Cellular atypia	SurePath procedure with Papanicolaou stain	40× (NA 0.75)	No retrievable information
Holmström et al (2021) [9]	Cervical cytology	Cellular atypia	Conventional Papanicolaou smears	20× (NA 0.4)	Scanning: 5 - 10 minutes; uploading 10 - 40 minutes
Sunny et al (2019) [19]	Oral cytology	Cellular atypia	Manual liquid-based cytology, stained with H&E[e] [45]	20× (NA 0.4) [45]	AI analysis: 10 minutes
Ghaderinia et al (2024) [46]	Sputum	Ferning patterns	Sedimented unstained sputum samples	40× magnification	No retrievable information
Soares et al (2024) [47]	Stool	Intestinal parasites	Fecal samples with centrifugation, flotation, and sedimentation [48]	No retrievable magnification and resolution	AI analysis: circa 3 minutes
Lundin et al (2024) [49]	Stool	Soil-transmitted helminths	Kato-Katz thick smears	20× (NA 0.4)	Scanning 5 - 10; uploading 10 - 20 minutes; AI analysis 5 minutes
Sahu et al (2024) [50]	Urine	Urinalysis	Cartridge that concentrates the urine through 5 minutes of sedimentation	40× (NA 0.65)	No retrievable information
Meulah et al (2022) [51]	Urine	Schistosoma	A membrane capturing filtered urine particles	4× (NA 0.1)	Scanning: 12 minutes; AI analysis: 5 minutes

XSL•FO
RenderX

Study	Sample	Target	Sample preparation	Sample scanning	Time for analysis
Oyibo et al (2022) [52]	Urine	Schistosoma	A membrane capturing filtered urine particles	4× (NA 0.1)	Scanning: 12 minutes; AI analysis: 10 - 12 minutes
Meulah et al (2024) [53]	Urine	Schistosoma	A membrane capturing filtered urine particles	4× (NA 0.1)	Scanning and AI analysis: 25 minutes

[a]CBC: complete blood count.

[b]DAPI: 4',6-diamidino-2-phenylindole.

[c]NA: numerical aperture.

[d]AI: artificial intelligence.

[e]H&E: hematoxylin and eosin.

Training Data and AI Analysis Pipeline

For training AI models, most studies used in-house collected and annotated datasets of varying sizes; some had hundreds of target objects in their dataset, whereas others had hundreds of thousands. Many studies reported using pretrained neural networks with different datasets such as COCOtrain2017 and ImageNet for training [47,53]. One study used unlabeled data from their collection for unsupervised pretraining and incorporated publicly available datasets [44].

The AI analysis pipeline for all included studies can be summarized as follows: a digitized microscopy sample was provided as input, fields-of-view (FOVs) were analyzed, FOV results were aggregated to produce a slide-level diagnosis, and this diagnosis served as the output (Figure 3). The digitized sample used as input could consist of either whole-slide images or multiple FOVs captured from the physical slide. The FOV analysis involved both the identification and the classification of specific targets; however, not all studies used this first identification of suspicious FOVs (regions of interest).

Figure 3. Visualization of the artificial intelligence analysis pipeline. As an illustrative case, the pipeline is applied to a digitized fecal smear with *Ascaris lumbricoides* parasite eggs.

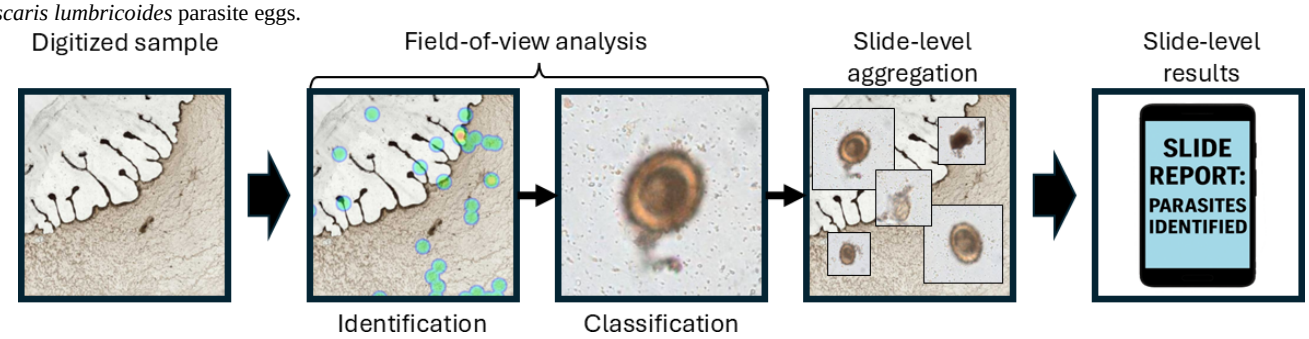

An initial detection of suspicious FOVs was described in half of the studies; when this was applied, algorithmic approaches, shallow CNNs, or support vector machines (SVMs) were used. The purpose of performing this initial identification of FOVs of interest was to reduce the number of FOVs that needed to be analyzed by more computationally expensive AI algorithms. An additional advantage of performing an initial detection of suspicious FOVs was that the FOVs were more homogeneous in content and quality, which can improve the accuracy of the AI classifier. All studies used AI for the FOV classification step, predominantly CNNs, except for the oldest study that used an SVM [42]. One approach was to use multiple classification steps, for example, by using SVM or shallow CNNs to classify targets and then reclassifying those with higher uncertainty with deeper CNNs [37,47].

To achieve the slide-level diagnosis from the FOV analysis results, multiple methods were used; for example, classifying slides with any number of positive targets as positive, using different cutoffs based on confidence or number of findings, or using AI-based methods such as SVMs and multiple instance learning (Table 3).

XSL•FO
RenderX

Table . Artificial intelligence model training and architecture for the included studies.

Study	Sample and target	Samples in training set	AI[a] model architecture and training
Bachar et al (2021) [32]	Blood and CBC[b]	No retrievable information	AI model with separate pipelines for platelets, RBCs[c], and WBCs[d] (1) algorithmically identifies candidates, and (2) candidates categorized by specialized CNNs[e] and machine learning algorithms
Gasparin et al (2023) [33]	Blood and CBC	Expert-verified training data gathered throughout development; no further retrievable information	AI model with CNN architecture using the YOLO[f] framework
Gasparin et al (2022) [34]	Blood and CBC	Expert-verified training data gathered throughout development; no further retrievable information [33]	AI model with CNN architecture using the YOLO framework
Akisin et al (2023) [35]	Blood and Downey cells	15,885 expert-annotated WBCs containing 172 Downey cells	AI model with YOLOv4-tiny-based framework with spatial attention using average and maximum pooling along the channel axis
Hamid et al (2024) [36]	Blood and malaria parasites	No retrievable information	AI model with (1) U-Net segmenting RBCs, (2) a 3-layer CNN removing normal RBCs, (3) a 23-layer CNN for detecting parasites, and (4) 1 positive object sufficient for slide positivity [37]
Holmström et al (2020) [38]	Blood and malaria parasites	25 thin blood smears with annotated trophozoites (n=5059) and other fluorescence signals (n=856)	AI model with (1) Circle Hough Transform identifying RBCs, (2) fluorescence signals from within the detected RBCs are used, and (3) RBCs with fluorescence signals were analyzed with a CNN (GoogLeNet)
Bae et al (2024) [37]	Blood and malaria parasites	No retrievable information	AI model with (1) U-Net segmenting RBCs, (2) a 3-layer CNN removing normal RBCs, (3) a 23-layer CNN for detecting parasites, and (4) 1 positive object sufficient for slide positivity
Ewnetu et al (2024) [39]	Blood and malaria parasites	No retrievable information	AI model with (1) U-Net segmenting RBCs, (2) a 3-layer CNN removing normal RBCs, (3) a 23-layer CNN for detecting parasites, and (4) 1 positive object sufficient for slide positivity [37]
Das et al (2022) [40]	Blood and malaria parasites	Subset of 1452 blood samples and 956,531 annotated parasite objects [54]	AI model analyzes only thick region with (1) potential parasites identified through dynamic thresholding and SVM[g], (2) CNN (VGG[h] architecture) classifies parasites, and (3) a predetermined threshold decides slide positivity [54]
Torres et al (2018) [41]	Blood and malaria parasites	Approximately 150 high-quality thick films with 75,000 parasites	AI model for thick region with (1) local thresholding and low-cost methods to identify potential parasites, (2) CNNs (VGG architecture) classifies parasites and stage, and (3) number and confidence of parasites determine slide-level diagnosis
Linder et al (2014) [42]	Blood and malaria parasites	A training set (n=10) with parasites (n=8329) and a validation set (n=6) parasites (n=569)	AI model with (1) thresholding algorithm segments potential parasites, and (2) mathematical feature extraction and classification with SVM

XSL•FO
RenderX

Study	Sample and target	Samples in training set	AI[a] model architecture and training
Horning et al (2021) [43]	Blood and malaria parasites	Thick model: Subset of 1452 blood samples with 956,531 parasite objects [54]. Thin model: 798 blood samples with more than 92,000 parasites [55]. Tuning slides: 48 slides	Separate AI models for thin and thick regions: Thick-AI model: (1) potential parasites identified through dynamic thresholding and SVM. (2) CNNs (VGG architecture) classify [54]. Thin AI model: (1) potential parasites detected with a gradient-boosted tree classifier. (2) CNNs for classifying parasite stages [55]
Stegmüller et al (2024) [44]	Cervical cytology and cellular atypia	A stratified 4-fold split approach to partition the 307 slides with 1228 tile-level annotations into training, validation, and test sets; 2 public datasets also used	AI model with (1) CNN (ResNet-50) with self-supervised training (DINO) and then supervised training with cell pasting, and (2) 8 most suspicious tiles used for slide classification with multiple instance learning (CLAM)
Holmström et al (2021) [9]	Cervical cytology and cellular atypia	350 WSIs[i] were used for training with 16,133 annotations made by a pathologist	AI model with (1) a CNN that segments slide into high- and low-grade atypia, and (2) a threshold that decides slide positivity
Sunny et al (2019) [19]	Oral cytology and cellular atypia	252 atypical and 280 normal cell images annotated (90% for training and 10% for validation)	AI model with (1) cells segmented to single cells, (2) a CNN (Inception V3) used for classification, and (3) cut-offs and SVMs based on percentage and mean score of atypical cells and mean cell score for slide diagnosis
Ghaderinia et al (2024) [46]	Sputum and ferning patterns (inflammation)	650 images (520 training and 130 validation) derived from 70 participants	AI model with (1) a CNN (EfficientNet-B0); and (2) CNN output used to classify sample
Soares et al (2024) [47]	Stool and intestinal parasites (both helminths and protozoans)	51,919 images containing 12,225 annotations of 15 parasite species (ranging from 83 to 3297 per species) [56]	AI model with (1) classification with extracted features and probabilistic SVM, and (2) uncertain objects analyzed with a CNN (Vgg-16) [56]
Lundin et al (2024) [49]	Stool and soil-transmitted helminths	388 samples with 15,058 annotations: *Ascaris lumbricoides* (n=2299), *Trichuris trichiura* (n=2727), hookworm (n=552), and artifacts (n=9480)	AI model with (1) YOLOv2 used to detect potential parasites, (2) a CNN (ResNet50) used for classification, and (3) 1 parasite sufficient for slide positivity
Sahu et al (2024) [50]	Urine and urinalysis	A dataset annotated by a pathologist	AI model with (1) a single CNN (YOLOX) to detect objects, and (2) object counts used to grade slide in tiers of positivity
Meulah et al (2022) [51]	Urine and Schistosoma	Both spiked laboratory samples and 33 field samples [17]	AI model with (1) a CNN segmentation model (U-Net architecture) [17]
Oyibo et al (2022) [52]	Urine and Schistosoma	17,799 annotated *Schistosoma haematobium* eggs in 2997 FOV images; dataset split into 80% training and 20% validation set	AI model with (1) a CNN (DeepLabv3-MobileNetV3), (2) egg-shaped ellipses fitted to segmented regions for counting, and (3) 1 parasite fulfilling criteria sufficient for slide positivity

XSL•FO
RenderX

Study	Sample and target	Samples in training set	AI[a] model architecture and training
Meulah et al (2024) [53]	Urine and Schistosoma	17,799 annotated *S. haematobium* eggs in 2997 FOV[j] images; dataset split into 80% training and 20% validation set [52]	AI model with (1) a CNN (DeepLabv3-MobileNetV3), (2) egg-shaped ellipses fitted to segmented regions for counting, and (3) 1 parasite fulfilling criteria sufficient for slide positivity [52]

[a]AI: artificial intelligence.
[b]CBC: complete blood count.
[c]RBCs: red blood cells.
[d]WBCs: white blood cells.
[e]CNNs: convolutional neural networks.
[f]YOLO: You Only Look Once.
[g]SVM: support vector machine.
[h]VGG: Visual Geometry Group.
[i]WSIs: whole slide images.
[j]FOV: field of view.

Study Outcomes

When the study outcomes were mapped, differences were observed in the reference standards, study sizes, and performance metrics used. For the 3 studies investigating CBCs, the Pearson correlation coefficient was compared with high-end analyzers and was above 0.9 for all cells except basophils, where the value ranged from 0.6 to 0.8 in all studies [32-34]. Nine studies used manual microscopy of the same samples as the reference standard and reported results with sensitivity and specificity (for malaria, soil-transmitted helminths, Schistosoma, and cervical cell atypia). Across these 9 studies, all reported a sensitivity and specificity of at least 80% except for 1 with a lower sensitivity of 57% [35] and 3 with lower specificity (75.6%, 78.4%, and 48.9%) [9,40,51]. One study included results with and without human expert verification: human verification of AI model findings increased specificity by 29.5% but conversely led to a sensitivity decrease of 0.9% for malaria detection [36]. Seven studies used reference standards such as polymerase chain reaction (PCR) or histology and included comparisons between AI-supported digital microscopy and manual microscopy; in 4 of these 7 studies, a higher sensitivity but lower specificity was reported for AI-supported digital microscopy. Of the remaining 3 studies, 1 evaluated urinalysis, in which manual analysis had higher sensitivity, specificity, or both across all targets [50], 1 for intestinal parasites where the AI had higher sensitivity and the same specificity [47], and 1 for oral atypia where the AI had both higher sensitivity and specificity [19]. The number of samples included in the diagnostic evaluations ranged from 27 to 2250. Most studies (n=15) achieved a low risk for bias according to QUADAS-2; however, some studies either lacked the information needed to properly evaluate bias or had methodological issues (n=7) (Table 4).

XSL•FO
RenderX

Table . Results for the included studies.

Study	Sample and target	Human verification	Outcome	Manual microscopy	Number of samples	Reference standard	QUADAS-2[a]
Bachar et al (2021) [32]	Blood and CBC[b]	No	*r*[c]≥0.94 (except basophils=0.6)	NR[d]	679	Hematology analyzer	Low
Gasparin et al (2023) [33]	Blood and CBC	Yes	*r*≥0.94 (except eosinophils/basophils=0.81)	NR	550	Hematology analyzer	Low
Gasparin et al (2022) [34]	Blood and CBC	Yes	*r*≥0.91 (except eosinophils/basophils=0.80)	NR	450	Hematology analyzer	Low
4: Akisin et al (2023) [35]	Blood and Downey cells	No	Se[e] 57%, Sp[f] 100%	NR	31	Manual microscopy	Mostly low
Hamid et al (2024) [36]	Blood and malaria	Yes	Se 90.2%, Sp 96.2%	Se 89.3%, Sp 100%	190	PCR[g]	Low
Holmström et al (2020) [38]	Blood and malaria	No	*r*=0.90 for parasite counts	NR	27	PCR	Mostly low
Bae et al (2024) [37]	Blood and malaria	No	Se 95.1%, Sp 91.4%	NR	488	Microscopy and RDTs[h]	Low
Ewnetu et al (2024) [39]	Blood and malaria	Yes	Se 83% - 93.9%, Sp 94% - 97.6%	Se 67% - 69.9%, Sp 97% - 98.7%	1165	PCR	Low
Das et al (2022) [40]	Blood and malaria	No	Se 91.1%, Sp 75.6%	NR	2250	Microscopy	Low
Torres et al (2018) [41]	Blood and malaria	No	Site 1: Se 72%, Sp 85% Site 2: Se 52%, Sp 70%	Site 1: Se 68%, Sp 100% Site 2: Se 42%, Sp 97%	Site 1: 400 Site 2: 300	PCR	Low
Linder et al (2014) [42]	Blood and malaria	Yes	Se 95%, Sp 100%	NR	31	Microscopy	Low
Horning et al (2021) [43]	Blood and malaria	No	Se 86.7%, Sp 100%	NR	35	Microscopy	Low
Stegmüller et al (2024) [44]	Cervical cytology and cellular atypia	No	Mean area under curve 77.5	NR	307 (4-fold split)	Microscopy	Low
Holmström et al (2021) [9]	Cervical cytology and cellular atypia	No	Se 100%, Sp 78.4%	NR	361	Microscopy	Mostly low
Sunny et al (2019) [19]	Oral cytology and cellular atypia	No	Se 89%, Sp 100%	Se 59%, Sp 67%	30	Histology	Low
Ghaderinia et al (2024) [46]	Sputum and ferning patterns (inflammation)	No	Se 94.3%, Sp 95.9%	NR	160	CT[i]	Mostly low
Soares et al (2024) [47]	Stool and intestinal parasites (both helminths and protozoans)	No	Se 86%, Sp 100%	Se 81%, Sp 100%	73	Manual and AI[j] microscopy	Mostly low
Lundin et al (2024) [49]	Stool and soil-transmitted helminths	No	Se 76.4% - 91.9% Sp 89.7% - 98.2%	NR	792	Microscopy	Low
Sahu et al (2024) [50]	Urine and urinalysis	No	Se ≥81% except for bacteria (76%) and casts (71%), Sp ≥88%	Se ≥94% and Sp ≥93%	240	Microscopy	Mostly low

XSL•FO
RenderX

Study	Sample and target	Human verification	Outcome	Manual microscopy	Number of samples	Reference standard	QUADAS-2[a]
Meulah et al (2022) [51]	Urine and Schistosoma	No	Se 87.3%, Sp 48.9%	NR	487	Microscopy	Low
Oyibo et al (2022) [52]	Urine and Schistosoma	No	Se 93.8%, Sp 93.9%	NR	65	Microscopy	Mostly low
Meulah et al (2024) [53]	Urine and Schistosoma	No	Se 62.9%, Sp 78.8 %	Se 61.9%, Sp 96.4%	339	PCR and particle lateral flow test	Low

[a]QUADAS-2: Quality assessment of diagnostic accuracy studies 2.

[b]CBC: complete blood count.

[c]*r*: Pearson correlation coefficient.

[d]NR: Not reported.

[e]Se: sensitivity.

[f]Sp: specificity.

[g]PCR: polymerase chain reaction.

[h]RDT: rapid diagnostic test.

[i]CT: computed tomographic scan.

[j]AI: artificial intelligence.

Discussion

Summary

This scoping review included 22 publications deploying AI-supported digital microscopy in PHC laboratories for multiple targets, published in 15 different journals. These studies fulfilled the concepts of using AI and digital microscopy to achieve a slide-level diagnosis in PHC laboratories. The number of included studies was low, given the extensive research on AI in medical imaging. The exclusion of 58 papers due to the absence of sample-level diagnoses and of 71 papers due to not being conducted in PHC laboratories suggests that most research has focused on target detection or advanced laboratory settings, rather than evaluating end-to-end diagnostic systems for PHC use. This is notable, given the potential benefits of such technologies in PHC laboratories. However, 9 of the 22 included studies were published in 2024 indicating an upward trend in studies focused on AI-supported digital microscopy at the PHC level.

The studies targeting specific diseases primarily focused on conditions that disproportionally affect vulnerable populations. The results from the included studies in this scoping review indicate that AI-supported digital microscopy can achieve accuracy comparable to that of standard microscopy for malaria, intestinal parasites, cell atypia, and urinalysis; to that of computed tomography for detecting pulmonary inflammation in patients with COVID-19; and to that of conventional hematology analyzers for CBC. Diagnostic accuracy comparable to the reference standard was defined as sensitivity and specificity of >80% or a Pearson correlation of >0.90. The reported results also indicate that AI-supported digital microscopy could be particularly advantageous for increasing sensitivity, as 6 out of 7 (85.7%) studies comparing it with manual microscopy reported higher sensitivity for AI-supported digital microscopy. Furthermore, the objective of the scoping review was to map target-agnostic challenges and solutions regarding sample preparation, scanning, AI methods, and human integration and discuss future implications for AI-supported digital microscopy in PHC laboratories.

Sample Preparation

Variability in target morphology and artifacts may reduce AI performance and can be introduced in all steps from sample collection to scanning. Manual steps in sample preparation are prone to introducing variability, and all included studies involved such steps, with 12 relying on entirely manual preparation. Decreased specificity due to sample variability was observed in one study, where poorly prepared samples led to the introduction of artifacts [41], and in another study where synthetically prepared samples in the training dataset lacked artifacts present in real-world samples [51]. Sensitivity can also be affected by variability in preparation, as demonstrated in one study on soil-transmitted helminths [49]. This indicates that variability introduced during sample preparation may be a major hurdle when developing AI-supported digital microscopy for PHC laboratories as more steps are performed manually.

There are possible solutions to sample variability. For example, improving consistency through good laboratory practices and standard operating procedures is one way to minimize sample variability; however, this requires system-specific training for personnel, good laboratory infrastructure, and quality controls, which might reduce the feasibility of implementation in PHC laboratories. The use of equipment such as cartridges to limit manual steps is another approach to minimize variability [33,36,50]. This may lower the demands on personnel; however, using disease-specific consumables may introduce issues, for example, increased costs. Another potential approach to minimize variability is to simplify sample preparation, for example, by removing staining or smearing steps [57]. Although this could reduce variability, it may also lead to a loss of valuable diagnostic information, in turn decreasing the AI model performance.

XSL•FO
RenderX

Scanning

The scanner needs to capture sufficient information to allow AI model classification of targets, but scanning large sample areas at high magnification is time-consuming. The scanning time can be decreased by analyzing a smaller sample area. However, this can lead to a reduced ability to detect low-density targets, highlighting a trade-off between faster diagnostics and high sensitivity for these cases. This is exemplified by one solution for malaria, where clinicians are able to increase the area analyzed to detect low-density infections [39]. Another solution to decrease scanning time is to use a lower magnification than what is conventionally used by microscopists: this approach achieved diagnostic accuracy comparable to manual microscopy for cytology [19], malaria [36], and parasitic infections [49,53]. Nonetheless, the use of lower magnification could result in information loss that reduces the AI model performance.

Training Data

All studies that specified the AI-training methods used variants of supervised learning which require annotated data. Annotating data is time-consuming and requires digitized samples that are rarely produced in PHC laboratories due to limited access to scanners. Therefore, many studies had to collect and annotate their own datasets rather than access existing data. One study also used laboratory-enriched samples to increase the number of targets [51]. In some cases, certain targets were underrepresented in the dataset, which caused the AI models to perform poorly on those [32,43], emphasizing the challenges of limited training data. To overcome limited datasets, approaches, such as using data augmentation, publicly available datasets, CNNs with pretrained weights, and unsupervised learning, were deployed [38,44,52]. Although there are many ways to limit the effect of small datasets, the improved diagnostic performance in studies, iteratively collecting larger datasets, highlights that insufficient training data remain a limiting factor when developing AI-supported digital microscopy for PHC laboratories [17,33,34,52]. Larger studies and collaborations that allow data sharing could provide solutions to the issue of limited training data.

AI Analysis Pipeline

The AI analysis pipelines used can be broadly divided into 3 main steps: FOV identification, FOV classification, and aggregation for sample-level diagnosis (Figure 3). Given that the analysis of a single sample took more than 30 minutes in some studies and that access to graphics processing units may be limited in PHC laboratories, efficiency becomes important. One strategy to minimize computational demands is to combine identification and classification, as implemented in the You Only Look Once framework, which uses a single CNN [34,50,58]. Another strategy is to first identify targets using fast and computationally efficient methods and subsequently feed the suspicious FOVs into more computationally intensive algorithms for classification. Using an initial object identification step may also enhance the uniformity of the data entering the classification stage, which may be particularly beneficial due to the variability in manually prepared samples that are commonly used in PHC laboratories [19].

For the third step, slide-level classification, different approaches were used: slides were classified as positive if a single positive target was detected; others applied cutoffs to reduce noise and false positives. In addition, certain studies used methods such as SVMs [19] or multiple instance learning [44] to aggregate slide-level results. While these methods may improve classification, they carry a risk of overfitting, especially since the number of training samples at the slide level is much smaller than at the object level.

Manual Verification

One study investigated AI-supported digital microscopy with and without human verification. In the study, human verification was performed on targets initially classified as positive by the AI models, which led to a 0.9% drop in sensitivity but a 29.5% increase in specificity [36]. This demonstrates that, with human intelligence, AI errors can be identified and removed without a substantial loss of sensitivity. This is in line with the high specificity presented in studies using human verification, which all showed specificity of >90% [39,42]. Expanding human verification to include borderline cases classified as negative may also be used to reduce false negatives and increase sensitivity.

Reported Diagnostic Performance

The reported diagnostic performance of the studies included in the scoping review indicates that AI-supported digital microscopy may achieve comparable diagnostic accuracy in PHC laboratories; however, it is important to account for methodological choices when interpreting the results and due to the heterogeneity in study designs, comparisons between studies and diseases become challenging. One example of methodological choices is that most studies used manual microscopy as the reference standard. Since microscopy itself is an imperfect diagnostic test, it can affect the performance of the index test and may result in over- or underestimation of the diagnostic accuracy of AI-supported microscopy. Two studies argued that this limitation may have reduced the apparent diagnostic accuracy of AI-supported digital microscopy [49,50]. Another aspect to consider is the number of samples on which the method was evaluated. For example, the study in which the AI-supported digital microscopy had higher sensitivity and specificity than manual microscopy analyzed 30 samples [19]. Generally, the QUADAS-2 tool indicated a low risk of bias. However, it did not capture the issue of AI models being trained on samples from the same collection, which is a potential source of poor generalizability for AI. This can occur even when the detection algorithms are trained on different datasets, for example, when thresholds or rules for deriving slide-level diagnoses are developed using the same slides on which diagnostic performance is later evaluated, leading to inflated estimates of diagnostic accuracy. However, some studies avoided training on the data from the same collection, included more than 100 test samples, had low risk of QUADAS-2 bias, and used a more advanced reference standard and still achieved comparable or better results than manual microscopy [33,36,39,53].

XSL•FO
RenderX

Limitations

A limitation of the extraction process was the lack of consistent terminology used in the field. This was exemplified in the search block aimed at identifying PHC. Terms such as "low-cost" and "PHC" were included but not "remote," which was used to describe one study that fulfilled the inclusion criteria [59]. Another limitation was the broad definition of PHC laboratories adopted from Fleming et al, which led to the inclusion of studies using relatively advanced methods, such as oil immersion scanning at 100× magnification and specially designed cartridges for sample preparation [2,33,41,42]. These methods may be difficult to implement in some PHC laboratories, but to achieve a more comprehensive overview of the field, the inclusion of these studies was deemed advantageous [25]. A third limitation stems from the lack of standardized methodological descriptions in the included studies. In some cases, key information, such as scanner magnification, was missing or reported inconsistently across studies, which complicated comparisons in the scoping review.

Steps Needed to Achieve Clinical Implementation

In this scoping review, we identified hurdles that were shared across several studies and that must be overcome before implementing AI-supported digital microscopy. Many developers have recognized a need to iteratively improve their AI-supported digital microscopy; thus, a framework that enables continuous improvements might be advantageous for supporting the development of more accurate AI models. This requires health policy guidelines and frameworks that give details on how these processes should be conducted [60]. Another hurdle in the implementation of AI-supported digital microscopy is cost. The development of lower-cost scanners has reduced expenses; however, most commercially available scanners remain more expensive than traditional microscopes [38,61]. Microscopes typically function reliably over long periods, and scanners may need comparable longevity for AI-supported digital microscopy to be cost-effective. One potential solution to this is modular scanner construction, which may improve its lifespan through component updates and thereby its sustainability. Some systems in this review were developed for specific diseases, which increases the cost of implementing them in PHC laboratories, as multiple systems would have to be acquired to replace microscopes. To make it economically feasible to implement AI-supported digital microscopy, it may, therefore, be necessary to adopt a multipurpose approach where systems are developed for multiple diseases. Some studies show that scanners can digitize different samples and similar approaches can be applied to different diseases [9,37,49]. Systems developed for specific diseases may instead be useful in large screening programs or epidemiological surveys, for example, for soil-transmitted helminths, malaria, or cancer screening. Cost-effectiveness trials could provide guidance for the feasibility of AI-supported digital microscopy and evaluate single-disease systems against multipurpose platforms.

Potential Implications for PHC Diagnostics

AI-supported digital microscopy has potential advantages compared with manual microscopy in PHC laboratories. First, it could improve diagnostic accuracy, especially sensitivity, and this may be further enhanced by incorporating human verification [39,53]. Second, it might increase the access and timeliness of diagnostics by allowing diagnostic procedures to be performed at the POC in PHC laboratories and eliminate the need to send samples elsewhere for analysis [9,19]. Third, it could alleviate the workload of personnel through task shifting. This could increase the productivity of experts and thereby access to image-based diagnosis [36,39].

Other POC diagnostic technologies, including rapid diagnostic tests (RDTs) and PCR-based methods, provide alternative means of diagnosing some conditions discussed in this scoping review [62,63]. One included study compared AI-supported digital microscopy with another POC-diagnostic method: a malaria study comparing it with RDTs [39]. The AI-supported digital microscopy had higher sensitivity and specificity than RDTs for some malaria species and settings but lower in others. One review proposed that AI-supported digital microscopy holds more promise than RDTs for malaria diagnosis [64]. However, other studies highlight the possibilities of other methods, such as RDTs and PCR, for POC diagnostics [62].

Implementation of AI-supported digital microscopy could strengthen health systems and increase health equity, particularly where resources are limited such as in scarcely populated areas and LMICs. This may be possible since, in addition to comparable diagnostic accuracy to microscopy, slides can be scanned and analyzed within approximately 30 minutes for multiple diseases [34,39,53]. As the process eliminates the need for microscopy expertise on-site, it could enable timely and accurate diagnostics in PHC laboratories that currently lack this capacity: even if manual verification is required, it can be performed remotely. Moreover, by decentralizing diagnostics, it may reduce referrals to higher-tier health care facilities, alleviating their work and minimizing the risk of referral-related dropouts [12].

Knowledge Gaps and Research Priorities

This scoping review identified evidence of the feasibility of AI-supported digital microscopy for multiple targets in PHC laboratories. Drawing on the evidence mapped here, future research should prioritize studying scalable and robust systems that can be transferred and implemented in new laboratories and settings. Achieving scalability requires research into AI-supported digital microscopy with an end-to-end perspective, where everything from sample preparation, scanning, and AI analysis until the final diagnosis is accounted for and easily reproducible. Adding to this, research that examines how predeveloped AI-based systems are transferred and implemented in new clinical settings would provide valuable insights into real-world robustness, which was done by some of the included studies. To enable this kind of research, large multisite collaborations are important, which could be facilitated by improved health policy guidelines and frameworks, as well as initiatives led by key stakeholders including governments and nongovernmental organizations. Furthermore, important research priorities include assessing cost-effectiveness and exploring perceived barriers to implementation among patients and health care professionals. Finally, the scoping review's screening process identified additional potential applications for

XSL•FO
RenderX

AI-supported digital microscopy, including tuberculosis, other parasitic diseases, respiratory cytology, sperm motility, and sickle cell anemia, which warrant further investigation in PHC settings [65-69].

Conclusions

This scoping review identified 22 studies deploying AI-supported digital microscopy in PHC laboratories. For multiple diagnostic purposes, AI-supported digital microscopy achieved comparable results to the reference standard and could be particularly advantageous for increasing sensitivity in diagnosis. Further research is needed on challenges such as generalizability, scalability, and cost-effectiveness. Such evidence is critical to stimulate product development, enable regulatory approval, and support reimbursement and adoption by health care authorities. If the methods can be demonstrated to be feasible in real-life clinical PHC settings, translated into medical device products, and carefully integrated into health care systems, they are likely to improve access to diagnostics, particularly in LMICs and scarcely populated regions.

Acknowledgments

The authors acknowledge the support of the Karolinska Institute's Library for guidance in developing a satisfactory search strategy. The authors declare the use of generative artificial intelligence in the research and writing process. According to the GAIDeT taxonomy (2025) [70], the following tasks were delegated to GAI tools under full human supervision: (1) Proofreading and editing, and (2) adapting and adjusting the tone and style. Paperpal Preflight and ChatGPT 4, 4.5, and 5 (GPT, OpenAI's large-scale language generation model) were used for these purposes.

Funding

This study was funded by the Erling-Persson Foundation and Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. In addition, it was supported by the Swedish Research Council, Finska Läkaresällskapet r.f., Wilhelm och Else Stockmanns stiftelse r.f and Medicinska Understödsföreningen Liv och Hälsa r.f. No funder had any role in the study design or decision to submit the paper for publication.

Data Availability

All data generated within this scoping review are available in the study, such as search details.

Authors' Contributions

JVB conceived and designed the study, wrote the manuscript, developed the search strategy, screened the studies for eligibility, and performed the data extraction. AS screened the studies for eligibility, performed the data extraction, and assisted in writing the manuscript. JL conceived and designed the study, assisted in writing the manuscript, and developing the search strategy. NL conceived and designed the study, assisted in writing the manuscript, developing the search strategy, and screened the studies for eligibility. VD conceived and designed the study and assisted in writing the manuscript. AM assisted in designing the study and in writing the manuscript.

Conflicts of Interest

JL reported receiving personal fees from Aiforia Technologies Oy and serving as cofounder and co-owner of Aiforia Technologies Oy outside the submitted work. JL and AS reported having a patent for Mobile Microscope pending (no. WO2017037334A1, the invention is related to the use of fluorescence imaging filters combined with inexpensive plastic lenses, and all rights are with the University of Helsinki) and JL having a patent for a slide holder for an optical microscope pending (no. WO2015185805A1; related to motorization of regular microscopes). All other authors have no conflicts of interest to declare.

Multimedia Appendix 1
Search strategy.
[DOCX File, 14 KB - jmir_v28i1e78500_app1.docx]

Multimedia Appendix 2
Original data extraction tool.
[DOCX File, 14 KB - jmir_v28i1e78500_app2.docx]

Multimedia Appendix 3
Quadas-2 tool.
[DOCX File, 40 KB - jmir_v28i1e78500_app3.docx]

Multimedia Appendix 4

XSL•FO
RenderX

Sample collection and scanning.
[DOCX File, 45 KB - jmir_v28i1e78500_app4.docx]

Checklist 1
PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File, 118 KB - jmir_v28i1e78500_app5.pdf]

References

1. Pinto-Coelho L. How artificial intelligence is shaping medical imaging technology: a survey of innovations and applications. Bioengineering (Basel) 2023 Dec 18;10(12):1435. [doi: 10.3390/bioengineering10121435] [Medline: 38136026]
2. Fleming KA, Naidoo M, Wilson M, et al. High-quality diagnosis: an essential pathology package. In: Jamison DT, Gelband H, Horton S, editors. Disease Control Priorities: Improving Health and Reducing Poverty, 3rd edition: International Bank for Reconstruction and Development/World Bank; 2017. [doi: 10.1596/978-1-4648-0527-1_ch11]
3. Wilson ML, Fleming KA, Kuti MA, Looi LM, Lago N, Ru K. Access to pathology and laboratory medicine services: a crucial gap. Lancet 2018 May 12;391(10133):1927-1938. [doi: 10.1016/S0140-6736(18)30458-6] [Medline: 29550029]
4. Shamshad F, Khan S, Zamir SW, et al. Transformers in medical imaging: a survey. Med Image Anal 2023 Aug;88:102802. [doi: 10.1016/j.media.2023.102802] [Medline: 37315483]
5. Mohammed FA, Tune KK, Assefa BG, Jett M, Muhie S. Medical image classifications using convolutional neural networks: a survey of current methods and statistical modeling of the literature. MAKE 2024;6(1):699-736. [doi: 10.3390/make6010033]
6. Rezende MT, Bianchi AGC, Carneiro CM. Cervical cancer: automation of Pap test screening. Diagn Cytopathol 2021 Apr;49(4):559-574. [doi: 10.1002/dc.24708] [Medline: 33548162]
7. da Silva LM, Pereira EM, Salles PG, et al. Independent real-world application of a clinical-grade automated prostate cancer detection system. J Pathol 2021 Jun;254(2):147-158. [doi: 10.1002/path.5662] [Medline: 33904171]
8. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): a comparative analysis. Lancet Digit Health 2021 Mar;3(3):e195-e203. [doi: 10.1016/S2589-7500(20)30292-2] [Medline: 33478929]
9. Holmström O, Linder N, Kaingu H, et al. Point-of-care digital cytology with artificial intelligence for cervical cancer screening in a resource-limited setting. JAMA Netw Open 2021 Mar 1;4(3):e211740. [doi: 10.1001/jamanetworkopen.2021.1740] [Medline: 33729503]
10. Ward P, Dahlberg P, Lagatie O, et al. Affordable artificial intelligence-based digital pathology for neglected tropical diseases: a proof-of-concept for the detection of soil-transmitted helminths and Schistosoma mansoni eggs in Kato-Katz stool thick smears. PLoS Negl Trop Dis 2022 Jun;16(6):e0010500. [doi: 10.1371/journal.pntd.0010500] [Medline: 35714140]
11. Linder N, Nyirenda D, Mårtensson A, Kaingu H, Ngasala B, Lundin J. AI supported diagnostic innovations for impact in global women's health. BMJ 2025 Oct 10;391:e086009. [doi: 10.1136/bmj-2025-086009] [Medline: 41073085]
12. Diagnostic errors. World Health Organization. 2016. URL: https://iris.who.int/handle/10665/252410 [accessed 2025-11-27]
13. Bogoch II, Lundin J, Lo NC, Andrews JR. Mobile phone and handheld microscopes for public health applications. Lancet Public Health 2017 Aug;2(8):e355. [doi: 10.1016/S2468-2667(17)30120-2]
14. Galvan P, Ortellado J, Rivas R, Grossling B, Hilario E. PP84 developing the network for the future of healthcare through telemedicine-driven diagnostic innovation. Int J Technol Assess Health Care 2024 Dec;40(S1):S89-S89. [doi: 10.1017/S0266462324002538]
15. Reschke P, Gruenewald LD, Koch V, et al. Radiology access in rural Germany: a nationwide survey on outpatient imaging and teleradiology. Diagnostics (Basel) 2025 Apr 10;15(8):962. [doi: 10.3390/diagnostics15080962] [Medline: 40310336]
16. Bychkov A, Fukuoka J. Evaluation of the global supply of pathologists [abstract]. In: Modern Pathology: Springer Nature; 2022, Vol. 35. [doi: 10.1038/s41379-022-01050-6]
17. Oyibo P, Jujjavarapu S, Meulah B, et al. Schistoscope: an automated microscope with artificial intelligence for detection of Schistosoma haematobium eggs in resource-limited settings. Micromachines (Basel) 2022 Apr 19;13(5):643. [doi: 10.3390/mi13050643] [Medline: 35630110]
18. Holmström O, Linder N, Ngasala B, et al. Point-of-care mobile digital microscopy and deep learning for the detection of soil-transmitted helminths and Schistosoma haematobium. Glob Health Action 2017 Jun;10(sup3):1337325. [doi: 10.1080/16549716.2017.1337325] [Medline: 28838305]
19. Sunny S, Baby A, James BL, et al. A smart tele-cytology point-of-care platform for oral cancer screening. PLoS ONE 2019;14(11):e0224885. [doi: 10.1371/journal.pone.0224885] [Medline: 31730638]
20. Khanagar SB, Naik S, Al Kheraif AA, et al. Application and performance of artificial intelligence technology in oral cancer diagnosis and prediction of prognosis: a systematic review. Diagnostics (Basel) 2021;11(6):1004. [doi: 10.3390/diagnostics11061004]
21. Rodriguez JPM, Rodriguez R, Silva VWK, et al. Artificial intelligence as a tool for diagnosis in digital pathology whole slide images: a systematic review. J Pathol Inform 2022;13:100138. [doi: 10.1016/j.jpi.2022.100138] [Medline: 36268059]

XSL•FO
RenderX

22. Peters MDJ, Godfrey C, McInerney P, Munn Z, Tricco AC, Khalil H. Chapter 11: scoping reviews (2020 version). In: Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis: JBI; 2020. [doi: 10.46658/JBIRM-20-01]
23. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
24. von Bahr J, Diwan V, Mårtensson A, Linder N, Lundin J. Artificial intelligence-supported digital microscopy diagnostics in primary health care laboratories: a scoping review protocol. OSF. 2024 Mar 7. URL: https://osf.io/yz67t/overview [accessed 2025-12-10]
25. von Bahr J, Diwan V, Mårtensson A, Linder N, Lundin J. AI-supported digital microscopy diagnostics in primary health care laboratories: protocol for a scoping review. JMIR Res Protoc 2024 Nov 1;13:e58149. [doi: 10.2196/58149] [Medline: 39486020]
26. Yang B, Olsen M, Vali Y, et al. Study designs for comparative diagnostic test accuracy: a methodological review and classification scheme. J Clin Epidemiol 2021 Oct;138:128-138. [doi: 10.1016/j.jclinepi.2021.04.013]
27. Clark J, Glasziou P, Del Mar C, Bannach-Brown A, Stehlik P, Scott AM. A full systematic review was completed in 2 weeks using automation tools: a case study. J Clin Epidemiol 2020 May;121:81-90. [doi: 10.1016/j.jclinepi.2020.01.008] [Medline: 32004673]
28. Covidence.: Veritas Health Innovation URL: https://www.covidence.org/ [accessed 2025-12-07]
29. Chien TI, Kao JT, Liu HL, et al. Urine sediment examination: a comparison of automated urinalysis systems and manual microscopy. Clinica Chimica Acta 2007 Sep;384(1-2):28-34. [doi: 10.1016/j.cca.2007.05.012]
30. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
31. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. [doi: 10.1136/bmj.n71]
32. Bachar N, Benbassat D, Brailovsky D, et al. An artificial intelligence-assisted diagnostic platform for rapid near-patient hematology. Am J Hematol 2021 Oct 1;96(10):1264-1274. [doi: 10.1002/ajh.26295] [Medline: 34264525]
33. Gasparin AT, Araujo CIF, Cardoso MR, et al. Hilab system device in an oncological hospital: a new clinical approach for point of care CBC test, supported by the internet of things and machine learning. Diagnostics (Basel) 2023 May 11;13(10):1695. [doi: 10.3390/diagnostics13101695] [Medline: 37238184]
34. Gasparin AT, Araujo CIF, Schmitt P, et al. Hilab system, a new point-of-care hematology analyzer supported by the Internet of Things and artificial intelligence. Sci Rep 2022 Jun 21;12(1):10409. [doi: 10.1038/s41598-022-13913-8] [Medline: 35729182]
35. Ardicoglu Akisin Y, Akar N, Burkay Cotelı M. Decision support system for the classification of Downey cells as a pre-diagnostic tool. Turkish Journal of Biochemistry 2024 Jan 2;48(6):634-640. [doi: 10.1515/tjb-2023-0035]
36. Hamid MMA, Mohamed AO, Mohammed FO, et al. Diagnostic accuracy of an automated microscope solution (miLab™) in detecting malaria parasites in symptomatic patients at point-of-care in Sudan: a case-control study. Malar J 2024 Jun 28;23(1):200. [doi: 10.1186/s12936-024-05029-3] [Medline: 38943203]
37. Bae CY, Shin YM, Kim M, et al. Embedded-deep-learning-based sample-to-answer device for on-site malaria diagnosis. Front Bioeng Biotechnol 2024;12:1392269. [doi: 10.3389/fbioe.2024.1392269] [Medline: 39100623]
38. Holmström O, Stenman S, Suutala A, et al. A novel deep learning-based point-of-care diagnostic method for detecting Plasmodium falciparum with fluorescence digital microscopy. PLoS One 2020;15(11):e0242355. [doi: 10.1371/journal.pone.0242355] [Medline: 33201905]
39. Ewnetu Y, Badu K, Carlier L, et al. A digital microscope for the diagnosis of Plasmodium falciparum and Plasmodium vivax, including P. falciparum with hrp2/hrp3 deletion. PLOS Glob Public Health 2024;4(5):e0003091. [doi: 10.1371/journal.pgph.0003091] [Medline: 38768243]
40. Das D, Vongpromek R, Assawariyathipat T, et al. Field evaluation of the diagnostic performance of EasyScan GO: a digital malaria microscopy device based on machine-learning. Malar J 2022 Apr 12;21(1):122. [doi: 10.1186/s12936-022-04146-1] [Medline: 35413904]
41. Torres K, Bachman CM, Delahunt CB, et al. Automated microscopy for routine malaria diagnosis: a field comparison on Giemsa-stained blood films in Peru. Malar J 2018 Sep 25;17(1):339. [doi: 10.1186/s12936-018-2493-0] [Medline: 30253764]
42. Linder N, Turkki R, Walliander M, et al. A malaria diagnostic tool based on computer vision screening and visualization of Plasmodium falciparum candidate areas in digitized blood smears. PLOS One 2014;9(8):e104855. [doi: 10.1371/journal.pone.0104855] [Medline: 25144549]
43. Horning MP, Delahunt CB, Bachman CM, et al. Performance of a fully-automated system on a WHO malaria microscopy evaluation slide set. Malar J 2021 Feb 25;20(1):110. [doi: 10.1186/s12936-021-03631-3] [Medline: 33632222]
44. Stegmüller T, Abbet C, Bozorgtabar B, et al. Self-supervised learning-based cervical cytology for the triage of HPV-positive women in resource-limited settings and low-data regime. Comput Biol Med 2024 Feb;169:107809. [doi: 10.1016/j.compbiomed.2023.107809] [Medline: 38113684]
45. Skandarajah A, Sunny SP, Gurpur P, et al. Mobile microscopy as a screening tool for oral cancer in India: a pilot study. PLoS One 2017;12(11):e0188440. [doi: 10.1371/journal.pone.0188440] [Medline: 29176904]

XSL•FO
RenderX

46. Ghaderinia M, Abadijoo H, Mahdavian A, et al. Smartphone-based device for point-of-care diagnostics of pulmonary inflammation using convolutional neural networks (CNNs). Sci Rep 2024 Mar 22;14(1):6912. [doi: 10.1038/s41598-024-54939-4] [Medline: 38519489]
47. Soares FA, Suzuki CTN, Sabadini E, et al. Laboratory validation of the automated diagnosis of intestinal parasites via fecal sample processing for the recovery of intestinal parasites through the dissolved air flotation technique. Parasit Vectors 2024 Aug 30;17(1):368. [doi: 10.1186/s13071-024-06434-y] [Medline: 39215369]
48. Carvalho JD, Santos BD, Gomes JF, et al. TF - Test modified: new diagnostic tool for human enteroparasitosis. Clinical Laboratory Analysis 2016 Jul;30(4):293-300. [doi: 10.1002/jcla.21854]
49. Lundin J, Suutala A, Holmström O, et al. Diagnosis of soil-transmitted helminth infections with digital mobile microscopy and artificial intelligence in a resource-limited setting. PLoS Negl Trop Dis 2024 Apr;18(4):e0012041. [doi: 10.1371/journal.pntd.0012041] [Medline: 38602896]
50. Sahu A, Kandaswamy S, Singh DV, et al. AI driven lab-on-chip cartridge for automated urinalysis. SLAS Technol 2024 Jun;29(3):100137. [doi: 10.1016/j.slast.2024.100137] [Medline: 38657705]
51. Meulah B, Oyibo P, Bengtson M, et al. Performance evaluation of the Schistoscope 5.0 for (semi-)automated digital detection and quantification of Schistosoma haematobium eggs in urine: a field-based study in Nigeria. Am J Trop Med Hyg 2022 Nov 14;107(5):1047-1054. [doi: 10.4269/ajtmh.22-0276] [Medline: 36252803]
52. Oyibo P, Meulah B, Bengtson M, et al. Two-stage automated diagnosis framework for urogenital schistosomiasis in microscopy images from low-resource settings. J Med Imag 2023;10(4):044005. [doi: 10.1117/1.JMI.10.4.044005]
53. Meulah B, Oyibo P, Hoekstra PT, et al. Validation of artificial intelligence-based digital microscopy for automated detection of Schistosoma haematobium eggs in urine in Gabon. PLoS Negl Trop Dis 2024 Feb;18(2):e0011967. [doi: 10.1371/journal.pntd.0011967] [Medline: 38394298]
54. Mehanian C, Jaiswal M, Delahunt C, et al. Computer-automated malaria diagnosis and quantitation using convolutional neural networks. Presented at: 2017 IEEE International Conference on Computer Vision Workshop (ICCVW); Oct 22-29, 2017; Venice, Italy p. 125. [doi: 10.1109/ICCVW.2017.22]
55. Delahunt CB, Jaiswal MS, Horning MP, et al. Fully-automated patient-level malaria assessment on field-prepared thin blood film microscopy images, including supplementary information. arXiv. Preprint posted online on Sep 11, 2022. [doi: 10.48550/arXiv.1908.01901]
56. Osaku D, Cuba CF, Suzuki CTN, Gomes JF, Falcão AX. Automated diagnosis of intestinal parasites: a new hybrid approach and its benefits. Comput Biol Med 2020 Aug;123:103917. [doi: 10.1016/j.compbiomed.2020.103917] [Medline: 32768052]
57. Rivenson Y, Wang H, Wei Z, et al. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. Nat Biomed Eng 2019 Jun;3(6):466-477. [doi: 10.1038/s41551-019-0362-y] [Medline: 31142829]
58. Ali ML, Zhang Z. The YOLO Framework: a comprehensive review of evolution, applications, and benchmarks in object detection. Computers 2024;13(12):336. [doi: 10.3390/computers13120336]
59. Cure-Bolt N, Perez F, Broadfield LA, et al. Artificial intelligence-based digital pathology for the detection and quantification of soil-transmitted helminths eggs. PLoS Negl Trop Dis 2024 Sep;18(9):e0012492. [doi: 10.1371/journal.pntd.0012492] [Medline: 39348405]
60. Regulatory considerations on artificial intelligence for health. World Health Organization. 2023. URL: https://www.who.int/publications/i/item/9789240078871 [accessed 2025-11-27]
61. Patel A, Balis UGJ, Cheng J, et al. Contemporary whole slide imaging devices and their applications within the modern pathology department: a selected hardware review. J Pathol Inform 2021;12:50. [doi: 10.4103/jpi.jpi_66_21] [Medline: 35070479]
62. Ofori B, Twum S, Nkansah Yeboah S, Ansah F, Amofa Nketia Sarpong K. Towards the development of cost-effective point-of-care diagnostic tools for poverty-related infectious diseases in sub-Saharan Africa. PeerJ 2024;12:e17198. [doi: 10.7717/peerj.17198] [Medline: 38915381]
63. Gupta R, Gupta S. Point-of-care tests for human papillomavirus detection in uterine cervical samples: a review of advances in resource-constrained settings. Indian J Med Res 2023 Nov 1;158(5 & amp; amp;6):509-521. [doi: 10.4103/ijmr.ijmr_1143_23] [Medline: 38236008]
64. Coro F, De Maria C, Mangano VD, Ahluwalia A. Technologies for the point-of-care diagnosis of malaria: a scoping review. Infect Dis Poverty 2025 Jun 23;14(1):54. [doi: 10.1186/s40249-025-01329-1] [Medline: 40551195]
65. Mandal S, Das D, Udutalapally V. mSickle: sickle cell identification through gradient evaluation and smartphone microscopy. J Ambient Intell Human Comput 2023 Oct;14(10):13319-13331. [doi: 10.1007/s12652-022-03786-0]
66. Aulia S, Suksmono AB, Mengko TR, Alisjahbana B. A novel digitized microscopic images of ZN-stained sputum smear and its classification based on IUATLD grades. IEEE Access 2024;12:51364-51380. [doi: 10.1109/ACCESS.2024.3386208]
67. Roberts J, Flenaugh E, Oprea-Ilies G. AI aided rapid on site evaluation of respiratory cytology. J Am Soc Cytopathol 2021 Sep;10(5):S2. [doi: 10.1016/j.jasc.2021.07.131]
68. Kanakasabapathy MK, Sadasivam M, Singh A, et al. An automated smartphone-based diagnostic assay for point-of-care semen analysis. Sci Transl Med 2017 Mar 22;9(382):eaai7863. [doi: 10.1126/scitranslmed.aai7863]
69. D'Ambrosio MV, Bakalar M, Bennuru S, et al. Point-of-care quantification of blood-borne filarial parasites with a mobile phone microscope. Sci Transl Med 2015 May 6;7(286):286re4. [doi: 10.1126/scitranslmed.aaa3480]

XSL•FO
RenderX

70. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (Generative AI Delegation Taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res 2025 Aug 8:1-27. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]

Abbreviations

AI: artificial intelligence
CBC: complete blood count
CNN: convolutional neural network
FDA: US Food and Drug Administration
FOV: fields-of-view
LMICs: low- and middle-income countries
PCR: polymerase chain reaction
PHC: primary health care
POC: point of care
PRISMA: Preferred Reporting Items for Systematic reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2
RDT: rapid diagnostic test
SVM: support vector machine
WHO: World Health Organization

Edited by J Sarvestan; submitted 05.Jun.2025; peer-reviewed by HH Kim, SM Shaffi; revised version received 16.Nov.2025; accepted 17.Nov.2025; published 05.Jan.2026.

Please cite as:
von Bahr J, Suutala A, Diwan V, Mårtensson A, Lundin J, Linder N
AI-Supported Digital Microscopy Diagnostics in Primary Health Care Laboratories: Scoping Review
J Med Internet Res 2026;28:e78500
URL: https://www.jmir.org/2026/1/e78500
doi:10.2196/78500

© Joar von Bahr, Antti Suutala, Vinod Diwan, Andreas Mårtensson, Johan Lundin, Nina Linder. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effectiveness of Machine Learning in Detecting Vessels Encapsulating Tumor Clusters in Hepatocellular Carcinoma: Systematic Review and Meta-Analysis

Huili Shui[1,2,3], MD; Wenyu Wu[1,2,3], MD; Zhenming Xie[1,2,3], MD; Bing Yang[1,2,3], MD; Jia Deng[1,2,3], MD; Dongxin Tang[1,2,3], MD

[1]Guizhou University of Traditional Chinese Medicine, Guiyang, Guizhou Province, China

[2]Clinical Medical Research Center, The First Affiliated Hospital of Guizhou University of Traditional Chinese Medicine, Guiyang, Guizhou Province, China

[3]Guizhou Traditional Chinese Medicine Oncology Heritage and Technology Innovation Talent Base, Guiyang, Guizhou Province, China

Corresponding Author:
Dongxin Tang, MD
Guizhou University of Traditional Chinese Medicine
71 Baoshan North Road, Yunyan District
Guiyang, Guizhou Province
China
Phone: 86 18608511204
Email: hemingankang@sina.com

Abstract

Background: Vessels encapsulating tumor clusters (VETC) are significantly associated with poor prognosis in hepatocellular carcinoma (HCC). However, identifying VETC early remains challenging. Recently, machine learning has shown promise for VETC detection, but their diagnostic accuracy lacks systematic validation.

Objective: This meta-analysis aimed to systematically evaluate the diagnostic accuracy of machine learning models for detecting VETC in patients with HCC.

Methods: The Cochrane Library, Embase, Web of Science, and PubMed were searched up to June 21, 2025. Eligible studies focused on machine learning models for HCC VETC diagnosis. Studies that merely analyzed risk factors or lacked outcome measures were excluded. The Prediction Model Risk of Bias Assessment Tool was used to evaluate the risk of bias. A bivariate mixed-effects model was used for a meta-analysis based on 2×2 diagnostic tables. Subgroup analyses were performed according to modeling variables (nonradiomic vs radiomic features) and model types (traditional machine learning vs deep learning).

Results: This meta-analysis included 31 studies comprising 6755 patients with HCC (2699 VETC-positive). Nineteen studies used machine learning models based on nonradiomic features, and 12 used radiomic features (including deep learning). In the validation set, the nonradiomic model demonstrated a pooled sensitivity of 0.72 (95% CI 0.66-0.78), specificity of 0.74 (95% CI 0.68-0.80), and an area under the summary receiver operating characteristic curve (SROC AUC) of 0.80 (95% CI 0.76-0.83). The radiomic model showed sensitivity of 0.81 (95% CI 0.73-0.87), specificity of 0.73 (95% CI 0.67-0.79), and SROC AUC of 0.84 (95% CI 0.80-0.87). Traditional machine learning achieved sensitivity of 0.84 (95% CI 0.71-0.92), specificity of 0.75 (95% CI 0.67-0.81), and SROC AUC of 0.83 (95% CI 0.80-0.86). Deep learning exhibited sensitivity of 0.77 (95% CI 0.69-0.84), specificity of 0.70 (95% CI 0.59-0.79), and SROC AUC of 0.81 (95% CI 0.77-0.85).

Conclusions: This meta-analysis is the first to quantitatively assess the efficacy of machine learning models in HCC VETC diagnosis, addressing an evidence gap in this field. Unlike previous descriptive reviews, this analysis provides the first quantitative evidence revealing the potential value of machine learning in detecting HCC VETC. The findings provide a foundation for developing and refining subsequent intelligent detection tools. Despite their promising prospects, machine learning models have not yet reached the maturity required for clinical translation, owing to methodological heterogeneity, limited validation, and a high risk of bias. Future research should focus on conducting multicenter, large-sample, standardized, prospective studies to advance clinical translation.

Trial Registration: PROSPERO CRD420251084894; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251084894

(J Med Internet Res 2026;28:e82839) doi:10.2196/82839

XSL•FO
RenderX

KEYWORDS

carcinoma; hepatocellular carcinoma; hepatocellular; machine learning; radiomics; sensitivity; specificity; vessels encapsulating tumor clusters; VETC

Introduction

Liver cancer is the sixth most frequently diagnosed malignancy and the third leading cause of cancer-related mortality worldwide. In 2022, there were approximately 865,000 new cases of liver cancer and 757,948 related deaths. Hepatocellular carcinoma (HCC) accounts for 75% to 85% of primary liver cancers. Higher incidence and mortality rates are predominantly observed in developing regions, including Mongolia, Cambodia, Laos, Thailand, Vietnam, and Egypt [1]. Consequently, HCC has become a significant global oncological concern. For early-stage HCC, curative interventions such as surgical resection, liver transplantation, and ablation are recommended. For intermediate stages, locoregional therapies are typically used, while systemic treatment is preferred for individuals with a significant intrahepatic tumor burden. Advanced HCC is primarily managed with immune checkpoint inhibitors [2]. While these treatments have prolonged the survival of some patients, others still have a poor prognosis, even after undergoing the same treatment regimen. Patients with HCC have an overall 5-year survival rate of less than 20% [3]. Adding to the clinical challenge, the postoperative recurrence rate for HCC is high, at around 70%, even after curative resection. This persistent risk is a primary factor in unfavorable long-term patient prognosis [4]. Several factors have been associated with poor HCC prognosis, including microvascular invasion [5], the macrotrabecular-massive subtype [6], and the coexpression of Ki-67 and cytokeratin 19 [7]. Thus, identifying the key factors that drive poor prognosis in HCC is crucial.

Recently, there has been an increase in attention directed toward a distinct microvascular pattern known as vessels encapsulating tumor clusters (VETC). First described by Fang et al [8] in 2015, VETC refers to a vascular network surrounding tumor clusters in a spiderweb configuration. This pattern has been characterized as an independent vascular morphology that is distinct from epithelial-mesenchymal transition. It facilitates the release of entire tumor clusters into the bloodstream and enables a noninvasive metastatic mechanism in HCC. Research indicates that the prevalence of VETC correlates with tumor stage and aggressiveness. VETC occurs in approximately 30%-40% of patients undergoing resection, 50%-55% of individuals with postresection recurrence, and up to 76% of patients with unresectable disease receiving liver transplants [9]. A positive VETC status substantially influences the long-term prognosis of patients with HCC [10]. Recent studies show that patients with VETC have significantly shorter overall and disease-free survival than patients without VETC [11]. The presence of VETC has been established as a robust predictor of aggressive HCC behavior [10]. A meta-analysis by Wang et al [12] further confirmed VETC as a significant predictor of overall survival and tumor recurrence, supporting its role as an effective prognostic biomarker. Furthermore, multiparameter prognostic models that incorporate VETC status demonstrate superior predictive capacity for disease-free and overall survival in patients with HCC compared to the conventional tumor-node-metastasis staging system. These models facilitate personalized temporal survival estimation and have the potential to enhance clinical decision-making regarding surveillance management and therapeutic strategies [13,14]. Concurrently, VETC status is valuable in guiding systemic therapy selection and predicting treatment response in HCC. Notably, patients with VETC experience greater survival benefits from therapies including sorafenib [15], lenvatinib [16], and transarterial chemoembolization [17] than their VETC-negative counterparts do. These observations suggest that VETC-based stratified treatment strategies may optimize patient outcomes further, providing an evidence base for clinical decision-making [9]. Therefore, early detection of VETC status is clinically relevant for improving HCC prognosis. Currently, a definitive diagnosis of VETC relies on histopathological examination of biopsy or resected tissue specimens. However, this approach has several limitations. Technical challenges include dependence on tumor size and needle gauge, as well as variability among clinicians and pathologists. Procedural risks encompass hemorrhage, seeding metastasis, sampling error, and uncertainty in tumor characterization [18]. Thus, noninvasive methods for identifying VETC status in HCC are urgently needed to circumvent the limitations of tissue acquisition. Recent advancements in image processing and artificial intelligence have sparked growing interest in clinical oncology in the development of predictive models that integrate computed tomography, magnetic resonance imaging (MRI), and contrast-enhanced ultrasound with machine learning algorithms. This methodology is increasingly being explored for the noninvasive diagnosis of HCC VETC [19-21]. Several studies have explored the potential for directly diagnosing VETC in HCC using images alone [22,23]. Furthermore, machine learning models that incorporate clinical and imaging features have been developed to noninvasively predict VETC status [20,24]. Despite these promising findings, there is a lack of systematic evidence substantiating the efficacy of machine learning–based approaches for VETC detection in HCC. This lack of evidence poses a significant challenge to the development and improvement of artificial intelligence–assisted diagnostic tools. To address this deficiency, this systematic review and meta-analysis were conducted to summarize the performance of machine learning in noninvasively detecting VETC in HCC. The aim is to provide evidence-based support for developing and optimizing future intelligent diagnostic tools.

Methods

Study Registration

This meta-analysis was conducted in strict accordance with the PRISMA (Preferred Reporting Items for a Systematic Review and Meta-Analysis; checklist provided in Multimedia Appendix 1) Diagnostic Test Accuracy Studies guidelines [25] and was prospectively registered with the International Prospective Register of Systematic Reviews (CRD420251084894).

XSL•FO
RenderX

Eligibility Criteria

Textbox 1 presents the eligibility criteria for studies.

Textbox 1. Eligibility criteria.

Inclusion criteria

- Participants diagnosed with hepatocellular carcinoma
- Cohort, case-control, or cross-sectional studies
- Studies that developed machine learning models for the diagnosis vessels encapsulating tumor clusters
- Publications reported in English

Exclusion criteria

- Reviews, guidelines, expert opinions, or conference abstracts
- Studies that only performed risk factor analyses without constructing machine learning models
- Studies lacking key metrics for assessing the accuracy of machine learning models
- Studies reporting only univariable predictive performance

Data Sources and Search Strategy

According to the PRISMA search guidelines, the PubMed, Embase, Cochrane Library, and Web of Science databases were searched up to June 21, 2025. The search combined Medical Subject Headings and free-text terms, with no restrictions on language, country, or publication date. The search strategy was developed independently for this analysis. It was not adapted from existing systematic reviews, nor did it incorporate additional information sources or use search filters. The strategy did not undergo peer review before its execution, and no updates were made to the search following the initial retrieval. Based on the existing literature, we manually examined the reference lists of selected studies and relevant reviews to identify additional articles. Conference proceedings were excluded, and no attempts were made to contact authors for additional information [26]. Details are presented in Table S1 in Multimedia Appendix 2.

Study Selection

All retrieved articles were imported into EndNote (version 21; Clarivate) to remove duplicates. Two researchers (HS and ZX) screened the titles and abstracts of the articles independently and excluded the irrelevant ones. Subsequently, the full texts of potentially eligible studies were acquired and assessed for final inclusion. The researchers then cross-checked their results. Any discrepancies were resolved through discussion or adjudication by a third researcher (WW).

Data Extraction

Prior to data extraction, a standardized spreadsheet was developed. The extracted data included the following: first author, number of VETC cases, patient source, total sample size, study design, detection method, number of VETC cases in the training set, total training set size, country, method of validation set generation, model type, publication year, total validation set size, number of VETC cases in the validation set, and modeling variables.

Risk of Bias

The Prediction Model Risk of Bias Assessment Tool was applied to evaluate the risk of bias across four domains: participants, predictors, analysis, and outcome. Each domain contained 2-9 signaling questions, which could be answered as “yes or probably yes,” “no or probably no,” or “no information.” Domain-specific judgments were categorized as low, high, or unclear risk of bias. A domain was judged as having a low risk of bias if all signaling questions were answered “yes or probably yes”; a high risk of bias if at least one was answered “no or probably no”; or an unclear risk of bias if at least 1 was answered “no information” while all others were answered “yes or probably yes.” Two researchers (HS and ZX) conducted the assessment independently. They then cross-checked their results. Any disagreements were settled by consensus or arbitration by a third researcher (WW).

Synthesis Methods

A bivariate random-effects model was used for the meta-analysis based on available 2×2 diagnostic tables (either reported directly or reconstructed from reported performance metrics and sample size). The following pooled estimates were derived with their corresponding 95% CIs, sensitivity, specificity, positive likelihood ratio (LR+), negative likelihood ratio (LR–), diagnostic odds ratio (DOR), and the area under the summary receiver operating characteristic curve (SROC AUC). Deeks’ funnel plot was used to evaluate small-study effects. Fagan’s nomogram was applied to evaluate clinical applicability. Subgroup analyses were conducted according to modeling variables (nonradiomic vs radiomic features) and model type (traditional machine learning vs deep learning). A P value of <.05 indicated statistical significance. Stata (version 15.1; StataCorp LLC) was used for all meta-analyses.

Results

Study Selection

The database search yielded 302 potentially relevant articles. Of these, 177 duplicates were excluded (117 identified by software and 60 manually). After screening the titles and

XSL•FO
RenderX

abstracts, 89 articles unrelated to the topic were removed. The full texts of the remaining 36 articles were assessed for eligibility. Among them, 5 records were excluded; 3 because they did not develop machine learning models, and 2 because they were conference abstracts without full-text publication. Ultimately, 31 eligible studies were included [20-24,27-52]. The specific process is depicted in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for a Systematic Review and Meta-Analysis) flow diagram of the selection process for studies applying machine learning to detect hepatocellular carcinoma vessels encapsulating tumor clusters.

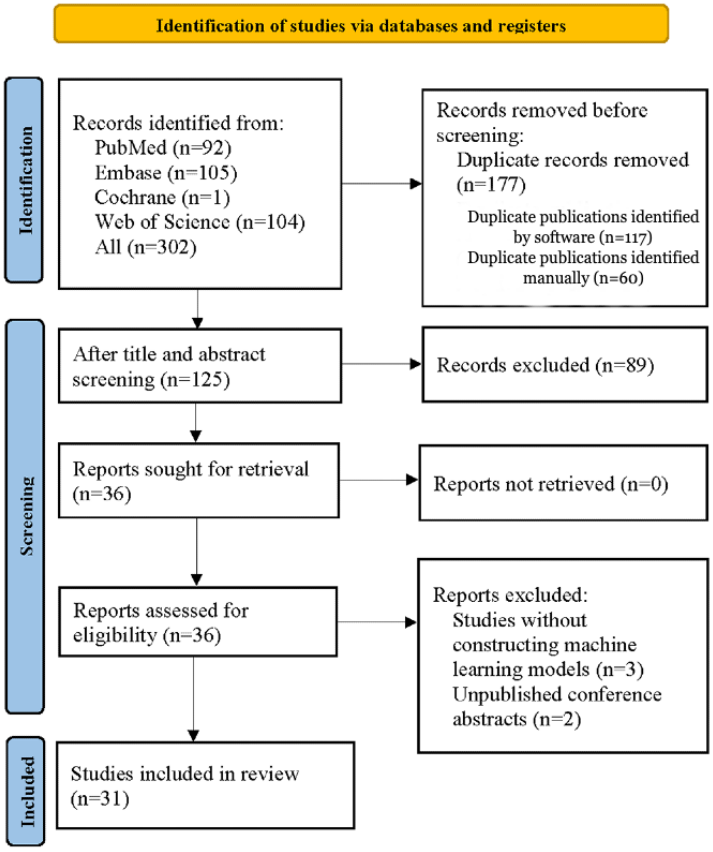

Study Characteristics

The 31 studies were published between 2021 and 2025. All were conducted in China and Japan. Of these, 8 studies used a case-control design, and 23 used a cohort design. Patient data were derived from multiple centers in 10 studies and from single centers in 21 studies. A total of 6755 participants with HCC were included, 2699 of whom were identified as VETC-positive. Regarding detection methods, 1 study used radiomic features based on computed tomography, 6 studies used MRI-based radiomics, 5 studies used deep learning, and 19 studies used traditional machine learning. The training sets collectively comprised 4411 participants with HCC, including 1714 with VETC. Internal validation was conducted in 14 studies, external validation in 3 studies, and both in 7 studies. The validation sets encompassed 2344 participants with HCC, 955 of whom were VETC-positive. The prediction models incorporated machine learning (n=5), logistic regression (n=24), least absolute shrinkage and selection operator regression (n=1), and random forest (n=1). Detailed characteristics are illustrated in Tables S2-S4 in Multimedia Appendix 2.

Risk of Bias

The Prediction Model Risk of Bias Assessment Tool was applied across 4 domains to assess the overall risk of bias. First, 8 of the 31 eligible studies in the participants domain used a case-control design, which introduced a high risk of bias due to potential differences in data sources and patient selection. Second, case-control studies were judged to carry a high risk of bias in the predictors domain because predictor assessment was influenced by knowledge of the outcome. Third, in the outcome domain, VETC status was consistently defined and confirmed via histopathological examination. Since the outcome definition, measurement, and classification were independent of predictor assessment and participant selection, this domain was assessed as having a low risk of bias. Fourth, in the analysis domain, 14 studies were judged to have a high risk of bias due to an insufficient sample size (including an events-per-variable ratio of <10 in model development, a validation set size of <100, or an absence of external validation). A total of 12 studies were rated as having an unclear risk of bias due to an inability to calculate the events-per-variable ratio. One study provided no explanation for missing values and was therefore judged to be at high risk of bias regarding missing data. Concerning model validation, 6 studies relied solely on random data splitting without cross-validation or mediator effect testing, resulting in a high risk of bias. Overall, 10 studies did not report the validation method used and were categorized as having an unclear risk of bias. Detailed assessment results are shown in Figure 2.

XSL•FO
RenderX

Figure 2. Risk of bias assessment for the included primary studies using the Prediction Model Risk of Bias Assessment Tool.

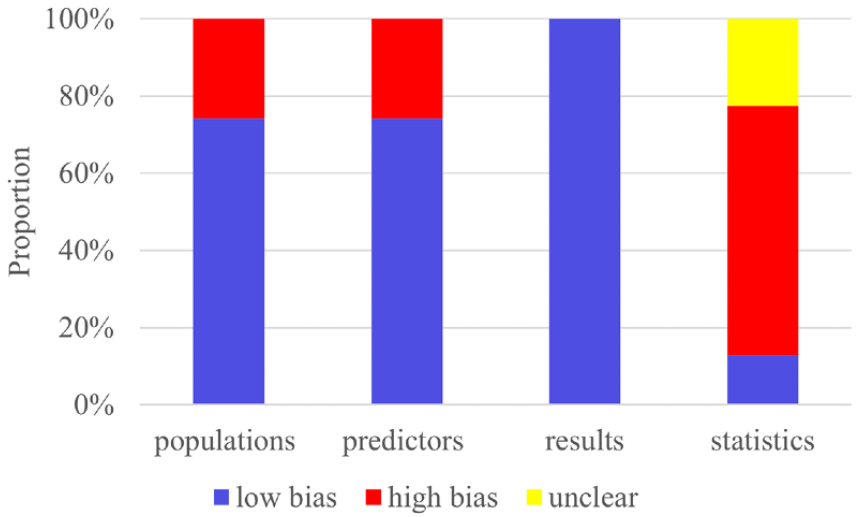

Meta-Analysis

Training Set-Overall

A total of 27 models from the training sets provided 2×2 diagnostic tables, with a 39% VETC-positive proportion. The pooled estimates were as follows: sensitivity 0.77 (95% CI 0.72-0.82), specificity 0.83 (95% CI 0.78-0.87), LR+ 4.5 (95% CI 3.5-5.8), LR– 0.27 (95% CI 0.22-0.34), DOR 16 (95% CI 11-24), and SROC AUC 0.87 (95% CI 0.84-0.89; Figures S1 and S2 in Multimedia Appendix 2). No significant small-study effect was illustrated via Deeks' funnel plot (*P*=.70; Figure S3 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 75%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 85% (Figure S4 in Multimedia Appendix 2).

Training Set-Nonradiomic Features

A total of 18 nonradiomic models from the training sets provided 2×2 diagnostic tables, with 38% representing VETC-positive cases. The pooled estimates were as follows: sensitivity 0.74 (95% CI 0.67-0.79), specificity 0.81 (95% CI 0.77-0.85), LR+ 3.9 (95% CI 3.2-4.7), LR– 0.33 (95% CI 0.27-0.40), DOR 12 (95% CI 9-16), and SROC AUC 0.85 (95% CI 0.81-0.88; Figures S5 and S6 in Multimedia Appendix 2). No significant small-study effect was detected via Deeks' funnel plot (*P*=.46; Figure S7 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 72%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 82% (Figure S8 in Multimedia Appendix 2).

Training Set-Radiomic Features

A total of 9 radiomic models from the training set provided 2×2 diagnostic tables, with a VETC-positive rate of 40%. The pooled estimates were as follows: sensitivity 0.83 (95% CI 0.75-0.90), specificity 0.86 (95% CI 0.71-0.94), LR+ 6.0 (95% CI 2.6-13.5), LR– 0.19 (95% CI 0.11-0.32), DOR 31 (95% CI 9-106), and SROC AUC 0.91 (95% CI 0.88-0.93; Figures S9 and S10 in Multimedia Appendix 2). No significant small-study effect was observed via Deeks' funnel plot (*P*=.40; Figure S11 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 80%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 89% (Figure S12 in Multimedia Appendix 2).

Of these, 6 traditional machine learning models provided 2×2 diagnostic tables, with a VETC-positive rate of 39%. The pooled estimates were as follows: sensitivity 0.88 (95% CI 0.70-0.96), specificity 0.85 (95% CI 0.67-0.94), LR+ 5.7 (95% CI 2.2-15.2), LR– 0.14 (95% CI 0.04-0.45), DOR 40 (95% CI 5-326), and SROC AUC 0.93 (95% CI 0.90-0.95; Figures S13 and S14 in Multimedia Appendix 2). No significant small-study effect was found via Deeks' funnel plot (*P*=.78; Figure S15 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 79%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 91% (Figure S16 in Multimedia Appendix 2).

Only 3 deep learning studies reported 2×2 diagnostic tables. Yu et al [35] developed an MRI-based deep learning model with a sensitivity of 0.87, a specificity of 0.54, and an area under the receiver operating characteristic curve (ROC AUC) of 0.83 (95% CI 0.83-0.84). Xu et al [49] reported a contrast-enhanced ultrasound–based model with sensitivity of 0.75, specificity of 0.92, and ROC AUC of 0.92 (95% CI 0.88-0.96). Yang et al [48] developed an MRI-based model with a sensitivity of 0.71, a specificity of 0.97, and an ROC AUC of 0.90 (95% CI 0.85-0.95).

Validation Set-Overall

A total of 27 models in the validation set provided complete 2×2 diagnostic tables, with a VETC-positive proportion of 41%. The pooled estimates were as follows: sensitivity 0.77 (95% CI 0.72-0.81), specificity 0.74 (95% CI 0.69-0.78), LR+ 2.9 (95% CI 2.5-3.3), LR– 0.32 (95% CI 0.26-0.38), DOR 9 (95% CI 7-12), and SROC AUC 0.82 (95% CI 0.78-0.85; Figures S17

XSL•FO
RenderX

and S18 in Multimedia Appendix 2). Deeks' funnel plot demonstrated no significant small-study effects (*P*=.09; Figure S19 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 66%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 83% (Figure S20 in Multimedia Appendix 2).

Validation Set-Nonradiomic Features

A total of 12 nonradiomic models in the validation set provided 2×2 diagnostic tables, with a VETC-positive proportion of 40%. The pooled estimates were as follows: sensitivity 0.72 (95% CI 0.66-0.78), specificity 0.74 (95% CI 0.68-0.80), LR+ 2.8 (95% CI 2.3-3.5), LR– 0.37 (95% CI 0.31-0.45), DOR 8 (95% CI 6-10), and SROC AUC 0.80 (95% CI 0.76-0.83; Figures S21 and S22 in Multimedia Appendix 2). No significant small-study effect was detected via Deeks' funnel plot (*P*=.98; Figure S23 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 65%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 80% (Figure S24 in Multimedia Appendix 2).

Validation Set-Radiomic Features

A total of 15 radiomic models in the validation set provided 2×2 diagnostic tables, with a VETC-positive rate of 41%. The pooled estimates were as follows: sensitivity 0.81 (95% CI 0.73-0.87), specificity 0.73 (95% CI 0.67-0.79), LR+ 3.0 (95% CI 2.5-3.7), LR– 0.26 (95% CI 0.19-0.36), DOR 12 (95% CI 8-17), and SROC AUC 0.84 (95% CI 0.80-0.87; Figures S25 and S26 in Multimedia Appendix 2). No significant small-study effect was observed via Deeks' funnel plot (*P*=.11; Figure S27 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 67%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 85% (Figure S28 in Multimedia Appendix 2).

Of these, 9 traditional machine learning models provided 2×2 diagnostic tables, with a VETC-positive rate of 41%. The pooled estimates were as follows: sensitivity 0.84 (95% CI 0.71-0.92), specificity 0.75 (95% CI 0.67-0.81), LR+ 3.3 (95% CI 2.6-4.3), LR– 0.21 (95% CI 0.11-0.39), DOR 16 (95% CI 8-32), and SROC AUC 0.83 (95% CI 0.80-0.86; Figure S29 and S30 in Multimedia Appendix 2). No significant small-study effect was shown via Deeks' funnel plot (*P*=.37; Figures S31 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 69%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 88% (Figure S32 in Multimedia Appendix 2).

Additionally, 6 deep learning models reported 2×2 diagnostic tables, with a VETC-positive proportion of 41%. The pooled estimates were as follows: sensitivity 0.77 (95% CI 0.69-0.84), specificity 0.70 (95% CI 0.59-0.79), LR+ 2.6 (95% CI 1.9-3.5), LR– 0.32 (95% CI 0.24-0.43), DOR 8 (95% CI 5-13), and SROC AUC 0.81 (95% CI 0.77-0.85; Figures S33 and S34 in Multimedia Appendix 2). Deeks' funnel plot suggested significant small-study effects (*P*=.04; Figure S35 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 63%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 82% (Figure S36 in Multimedia Appendix 2).

Discussion

Summary of Main Findings

This meta-analysis demonstrates that developing prediction models based on machine learning to detect HCC VETC status appears to be a feasible approach. Currently, these models are primarily constructed using nonradiomic and radiomic features. For nonradiomic machine learning models in the validation set, the pooled estimates were 0.72 (95% CI 0.66-0.78) for sensitivity and 0.74 (95% CI 0.68-0.80) for specificity. For radiomic machine learning models, the estimates were a sensitivity of 0.81 (95% CI 0.73-0.87) and a specificity of 0.73 (95% CI 0.67-0.79). For traditional machine learning models, the estimates were a sensitivity of 0.84 (95% CI 0.71-0.92) and a specificity of 0.75 (95% CI 0.67-0.81). For deep learning models, the estimates were a sensitivity of 0.77 (95% CI 0.69-0.84) and a specificity of 0.70 (95% CI 0.59-0.79).

Comparison With Previous Reviews

Previous research by Hyungjin Rhee et al [53] reviewed the angiodynamic changes in multistep HCC carcinogenesis. They introduced the typical pathological, clinical, and imaging features of HCC VETC and provided detailed guidance for VETC diagnosis. However, their study focused primarily on describing pathological mechanisms and typical features, lacking a quantitative assessment of different diagnostic methods. Ken Liu et al [9] investigated various methods for diagnosing VETC, including histopathology, imaging, and laboratory tests. They suggested that VETC could be predicted radiologically. While their research provided a comprehensive analysis of various diagnostic approaches, they did not quantitatively compare the sensitivity and specificity of different diagnostic methods. This omission limited a thorough evaluation of VETC diagnostic accuracy. Miaomiao Wang et al [54] explored the potential of machine learning in HCC VETC detection through a literature review and provided guidance for the auxiliary VETC diagnosis. While their review demonstrated the potential applications of machine learning in VETC detection, it lacked a direct comparison of different types of machine learning models, making it difficult to assess these models' actual application value in clinical practice. This study summarized nonradiomic (clinical features, image features, etc) and radiomic prediction models, and the diagnosis of current HCC VETC status appears to be an ideal noninvasive detection scheme that provides specific guidance for clinicians.

This study found that the model variables used to detect HCC VETC include both nonradiomic features (clinical features, image features, etc) and radiomic features. The clinical features

XSL•FO
RenderX

primarily consist of alpha-fetoprotein, carbohydrate antigen 19-9, aspartate aminotransferase, and indirect bilirubin. Image features mainly comprise intratumoral necrosis, low signal intensity around the tumor in the hepatobiliary phase, the tumor-to-liver signal intensity ratio on the hepatobiliary phase, and the tumor-to-liver apparent diffusion coefficient ratio. Various studies used different modeling variables. Most studies did not quantitatively present the association of modeling variables with VETC. Thus, a further summary of such correlations was not performed. Recently, radiomics has advanced the development and application of prediction models by converting images into repeatable quantitative data. Prediction models based on radiomic features have demonstrated significant clinical value in diagnosing and treating HCC. Studies have shown that radiomic features are effective in predicting HCC microvascular invasion [5], early recurrence [55], and Ki-67 and cytokeratin 19 expression [7].

In this meta-analysis, only a limited number of studies explored the diagnostic performance of radiomics for HCC VETC. While the studies demonstrated promising results, radiomics still faces significant challenges in practical application. For example, the quality of the image appears to change under different image parameters. Most studies in this meta-analysis did not discuss how such changes in image features affect radiomics results. Additionally, image segmentation is primarily divided into manual and deep learning automatic segmentation. The studies included in this meta-analysis primarily used manual segmentation. However, manual segmentation may be affected by the segmenter's prior knowledge. Although some researchers have attempted to summarize its repeatability through independent segmentation by multiple people, it is difficult to avoid the influence of the segmenter's experience on the region-of-interest area division. Therefore, future studies should consider developing and promoting a standardized radiomics analysis process manual to improve research repeatability. Many studies have demonstrated that models combining radiomics, clinical features, and imaging features perform better in disease diagnosis and prognosis prediction [56]. In this study, relatively few studies attempted to construct prediction models using a combination of clinical features and radiomics. Therefore, an effective quantitative analysis of the advantages of a combined model was difficult to perform. Future studies should explore and verify the value of radiomic models constructed from clinical features and imaging features in improving the diagnostic accuracy of HCC VETC.

The prediction models used in this study primarily encompassed logistic regression, random forest, deep learning, and least absolute shrinkage and selection operator regression. Due to the interpretability of its parameters, logistic regression allows for the development of simple and intuitive nomograms in clinical practice and appears to be favored by many researchers [57-59]. However, the interpretability of other machine learning models, such as random forest, support vector machines, and XGBoost, depends on analyses like Shapley additive explanations. Using them in clinical practice requires developing plugins, which increases the complexity of the application process [60-62]. Thus, from the perspectives of clinical simplicity and interpretability, logistic regression has relatively ideal advantages. Nonetheless, in many cases, logistic regression's predictive accuracy often appears no better than that of traditional machine learning models, such as random forest [46,63]. In radiomics, the core advantage of deep learning lies in its ability to efficiently process image data for disease diagnosis and prognosis prediction [64,65]. Relatively few studies in the radiomic feature literature included in this meta-analysis addressed deep learning models. Initial evidence suggested that deep learning models did not perform significantly better than traditional machine learning models. The primary reasons for this include the following. First, the study only incorporated 6 deep learning research projects, which is a relatively small sample size. Deep learning models typically require large-scale datasets to leverage their full advantages. Second, most studies lacked external validation, leaving the generalizability of the models inadequately tested. Third, variations in image acquisition parameters and quality across different research centers suggest that the design of deep learning model architectures and hyperparameter optimization may not yet be optimal. Therefore, future research developing intelligent tools to detect HCC VETC should attempt to integrate multicenter, large-sample medical image data to construct deep learning models for training and validation.

Advantages and Limitations

This meta-analysis is the first comprehensive summary of the performance of machine learning models in diagnosing HCC VETC. It provides evidence-based support for the subsequent development or updating of artificial intelligence systems. However, this study also has the following limitations. First, all 31 eligible studies originated from East Asia, and most relied primarily on internal validation. The lack of multicenter, multiethnic validation limited the assessment of the models' generalizability. Second, the best prediction model from each article was extracted, which covered a narrow range of machine learning types. The differences between different machine learning methods were not described. Third, the modeling variables were diverse. They were only presented without a quantitative description of their association with HCC VETC. Future research should adopt more transparent and interpretable modeling approaches to identify efficient predictors. Fourth, although deep learning can efficiently process image data, it does not have a significant advantage over traditional machine learning-based radiomics. However, the literature is limited, and the interpretation of the results may be subject to certain limitations. Fifth, HCC VETC is a novel mode of microvascular metastasis that has been proposed in recent years, and the associated research is in its initial stage. The positive definition has not yet been standardized.

Conclusions

This meta-analysis is the first to provide a systematic and quantitative assessment of machine learning for diagnosing HCC VETC, thereby addressing an evidence gap in this field. Unlike previous reviews, this study provides a quantitative evaluation of diagnostic performance. The findings demonstrate the feasibility and clinical potential of using machine learning to determine VETC status in patients with HCC. Notably, radiomics-based models exhibited significantly better

XSL•FO
RenderX

performance than nonradiomic models. While deep learning efficiently processes image data in radiomics, its performance is not significantly better than traditional machine learning-based radiomics. Despite their promising prospects, machine learning models have not yet reached the maturity required for clinical translation, owing to methodological heterogeneity, limited validation, and a high risk of bias. Future research should focus on conducting multicenter, large-sample, standardized, prospective studies to develop intelligent detection tools with higher performance. Validating the models across multiple regions and ethnic populations is essential to ensure their generalizability. This will ultimately enable the effective translation of research into clinical applications.

Funding

This research was supported by the Guizhou Provincial Engineering Research Center for Medical Transformation of Traditional Chinese Medicine and Ethnic Medicine in Cancer Prevention and Treatment (Qian-Jiao-Ji, 2023, No. 037), the Guizhou Provincial Science and Technology Program (Qian-Ke-He Platform Talents, 2020, No. 5013), the Talent Base for Traditional Chinese Medicine Oncology Inheritance and Technological Innovation of Guizhou Province (Qian-Ren-Ling-Fa, 2018, No. 3), and the Guizhou High-Level Innovative Talent Training Program (“Hundred-Level” Talents; Qian-Ke-He Ren-Cai, 2016, No. 4032).

The funding agency for this study did not participate in any aspect of the research design, data collection, analysis, interpretation of results, or manuscript preparation.

Data Availability

All data used in this study are included in the main text and supplementary materials. For further information or clarification, please contact the corresponding researchers.

Authors' Contributions

Conceptualization: HS
Methodology: HS, WW
Formal analysis: HS, JD, BY, WW
Investigation: HS
Data curation: HS, ZX, WW
Visualization: JD, BY
Writing—Original Draft: HS, JD, BY
Writing—Review & Editing: HS, DT
Supervision: DT
Project administration: DT
Funding acquisition: DT
All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-S checklist.
[DOCX File , 23 KB - jmir_v28i1e82839_app1.docx]

Multimedia Appendix 2
Supplementary documents.
[DOCX File , 36193 KB - jmir_v28i1e82839_app2.docx]

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74(3):229-263 [FREE Full text] [doi: 10.3322/caac.21834] [Medline: 38572751]
2. Hwang SY, Danpanichkul P, Agopian V, Mehta N, Parikh ND, Abou-Alfa GK, et al. Hepatocellular carcinoma: updates on epidemiology, surveillance, diagnosis and treatment. Clin Mol Hepatol 2025;31(Suppl):S228-S254 [FREE Full text] [doi: 10.3350/cmh.2024.0824] [Medline: 39722614]
3. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin 2022;72(1):7-33 [FREE Full text] [doi: 10.3322/caac.21708] [Medline: 35020204]

XSL•FO
RenderX

4. Herrero A, Toubert C, Bedoya JU, Assenat E, Guiu B, Panaro F, et al. Management of hepatocellular carcinoma recurrence after liver surgery and thermal ablations: state of the art and future perspectives. Hepatobiliary Surg Nutr 2024;13(1):71-88 [FREE Full text] [doi: 10.21037/hbsn-22-579] [Medline: 38322198]
5. Li L, Wu C, Huang Y, Chen J, Ye D, Su Z. Radiomics for the preoperative evaluation of microvascular invasion in hepatocellular carcinoma: a meta-analysis. Front Oncol 2022;12:831996 [FREE Full text] [doi: 10.3389/fonc.2022.831996] [Medline: 35463303]
6. Ziol M, Poté N, Amaddeo G, Laurent A, Nault J, Oberti F, et al. Macrotrabecular-massive hepatocellular carcinoma: a distinctive histological subtype with clinical relevance. Hepatology 2018;68(1):103-112. [doi: 10.1002/hep.29762] [Medline: 29281854]
7. Zhou L, Chen Y, Li Y, Wu C, Xue C, Wang X. Diagnostic value of radiomics in predicting Ki-67 and cytokeratin 19 expression in hepatocellular carcinoma: a systematic review and meta-analysis. Front Oncol 2023;13:1323534 [FREE Full text] [doi: 10.3389/fonc.2023.1323534] [Medline: 38234405]
8. Fang J, Zhou H, Zhang C, Shang L, Zhang L, Xu J, et al. A novel vascular pattern promotes metastasis of hepatocellular carcinoma in an epithelial-mesenchymal transition-independent manner. Hepatology 2015;62(2):452-454. [doi: 10.1002/hep.27760] [Medline: 25711742]
9. Liu K, Dennis C, Prince DS, Marsh-Wakefield F, Santhakumar C, Gamble JR, et al. Vessels that encapsulate tumour clusters vascular pattern in hepatocellular carcinoma. JHEP Rep 2023;5(8):100792. [doi: 10.1016/j.jhepr.2023.100792] [Medline: 37456680]
10. Renne SL, Woo HY, Allegra S, Rudini N, Yano H, Donadon M, et al. Vessels encapsulating tumor clusters (VETC) is a powerful predictor of aggressive hepatocellular carcinoma. Hepatology 2020;71(1):183-195. [doi: 10.1002/hep.30814] [Medline: 31206715]
11. Xu D, Li R, Shu C, Li Y, Tao R, Chen Y, et al. Association between vessels encapsulating tumor clusters and circulating tumor cells in hepatocellular carcinoma: clinical evidence and risk model development. Int J Med Sci 2025;22(12):2944-2955 [FREE Full text] [doi: 10.7150/ijms.111025] [Medline: 40657396]
12. Wang M, Cao L, Wang Y, Huang H, Tian X, Lei J. The prognostic value of vessels encapsulating tumor clusters (VETC) in patients with hepatocellular carcinoma: a systematic review and meta-analysis. Clin Transl Oncol 2024;26(8):2037-2046. [doi: 10.1007/s12094-024-03427-2] [Medline: 38523240]
13. Xiong S, Wang C, Zhang M, Yang X, Yun J, Liu L. A multi-parametric prognostic model based on clinicopathologic features: vessels encapsulating tumor clusters and hepatic plates predict overall survival in hepatocellular carcinoma patients. J Transl Med 2024;22(1):472 [FREE Full text] [doi: 10.1186/s12967-024-05296-3] [Medline: 38762511]
14. Wu M, Xiao Y, Wang Y, Deng L, Wang X, An T. Establishment of a clinical model based on vessels encapsulating tumour clusters that could efficiently predict recurrence of patients with hepatocellular carcinoma after curative hepatectomy. Pathology 2025;57(3):320-327. [doi: 10.1016/j.pathol.2024.08.014] [Medline: 39668071]
15. Fang J, Xu L, Shang L, Pan C, Ding J, Tang Y, et al. Vessels that encapsulate tumor clusters (VETC) pattern is a predictor of sorafenib benefit in patients with hepatocellular carcinoma. Hepatology 2019;70(3):824-839. [doi: 10.1002/hep.30366] [Medline: 30506570]
16. Zhang P, Ono A, Fujii Y, Hayes CN, Tamura Y, Miura R, et al. The presence of vessels encapsulating tumor clusters is associated with an immunosuppressive tumor microenvironment in hepatocellular carcinoma. Int J Cancer 2022;151(12):2278-2290. [doi: 10.1002/ijc.34247] [Medline: 36054900]
17. Wang J, Li X, Tang H, Fang R, Song J, Feng Y, et al. Vessels that encapsulate tumor clusters (VETC) pattern predicts the efficacy of adjuvant TACE in hepatocellular carcinoma. J Cancer Res Clin Oncol 2023;149(8):4163-4172. [doi: 10.1007/s00432-022-04323-4] [Medline: 36050540]
18. European Association for the Study of the Liver. EASL clinical practice guidelines on the management of hepatocellular carcinoma. J Hepatol 2025;82(2):315-374. [doi: 10.1016/j.jhep.2024.08.028] [Medline: 39690085]
19. Wei Y, Huang S, Huang L, Pei W, Zuo Y, Liao H. CT-based radiomics features combined with AFP for predicting vessels encapsulating tumor clusters and prognosis of hepatocellular carcinoma. J Hepatocell Carcinoma 2025;12:2069-2081 [FREE Full text] [doi: 10.2147/JHC.S542092] [Medline: 40969196]
20. Che F, Gao F, Li Q, Ren W, Tang H, Zaina G, et al. Fractal analysis based on Gd-EOB-DTPA-enhanced MRI for prediction of vessels that encapsulate tumor clusters in patients with hepatocellular carcinoma. Int J Surg 2025;111(7):4389-4399. [doi: 10.1097/JS9.0000000000002547] [Medline: 40441719]
21. Xu W, Huang B, Zhang R, Zhong X, Zhou W, Zhuang S, et al. Diagnostic and prognostic ability of contrast-enhanced unltrasound and biomarkers in hepatocellular carcinoma subtypes. Ultrasound Med Biol 2024;50(4):617-626. [doi: 10.1016/j.ultrasmedbio.2024.01.007] [Medline: 38281888]
22. Li C, Wen Y, Xie J, Chen Q, Dang Y, Zhang H, et al. Preoperative prediction of VETC in hepatocellular carcinoma using non-Gaussian diffusion-weighted imaging at high b values: a pilot study. Front Oncol 2023;13:1167209 [FREE Full text] [doi: 10.3389/fonc.2023.1167209] [Medline: 37305565]
23. Wang M, Cao L, Wang Y, Huang H, Cao S, Tian X, et al. Prediction of vessels encapsulating tumor clusters pattern and prognosis of hepatocellular carcinoma based on preoperative gadolinium-ethoxybenzyl-diethylenetriaminepentaacetic acid

XSL•FO
RenderX

magnetic resonance imaging. J Gastrointest Surg 2024;28(4):442-450. [doi: 10.1016/j.gassur.2024.02.004] [Medline: 38583894]
24. Qu Q, Liu Z, Lu M, Xu L, Zhang J, Liu M, et al. Preoperative gadoxetic acid-enhanced MRI features for evaluation of vessels encapsulating tumor clusters and microvascular invasion in hepatocellular carcinoma: creating nomograms for risk assessment. J Magn Reson Imaging 2024;60(3):1094-1110. [doi: 10.1002/jmri.29187] [Medline: 38116997]
25. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, the PRISMA-DTA Group, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA 2018;319(4):388-396 [FREE Full text] [doi: 10.1001/jama.2017.19163] [Medline: 29362800]
26. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc 2021;109(2):174-200 [FREE Full text] [doi: 10.5195/jmla.2021.962] [Medline: 34285662]
27. Chen H, Dong H, He R, Gu M, Zhao X, Song K, et al. Optimizing predictions: improved performance of preoperative gadobenate-enhanced MRI hepatobiliary phase features in predicting vessels encapsulating tumor clusters in hepatocellular carcinoma-a multicenter study. Abdom Radiol (NY) 2024;49(10):3412-3426. [doi: 10.1007/s00261-024-04283-y] [Medline: 38713432]
28. Li M, Zhang G, Li J, Ren Y, Jin X, Ke Q, et al. Intravoxel incoherent motion improves the accuracy of preoperative prediction of vessels encapsulating tumor clusters in hepatocellular carcinoma. J Hepatocell Carcinoma 2025;12:1177-1190 [FREE Full text] [doi: 10.2147/JHC.S519223] [Medline: 40524872]
29. Chu T, Zhao C, Zhang J, Duan K, Li M, Zhang T, et al. Application of a convolutional neural network for multitask learning to simultaneously predict microvascular invasion and vessels that encapsulate tumor clusters in hepatocellular carcinoma. Ann Surg Oncol 2022;29(11):6774-6783 [FREE Full text] [doi: 10.1245/s10434-022-12000-6] [Medline: 35754067]
30. Chen F, Du M, Qi X, Bian L, Wu D, Zhang S, et al. Nomogram estimating vessels encapsulating tumor clusters in hepatocellular carcinoma from preoperative gadoxetate disodium-enhanced MRI. J Magn Reson Imaging 2023;57(6):1893-1905. [doi: 10.1002/jmri.28488] [Medline: 36259347]
31. Chen H, He R, Gu M, Zhao X, Song K, Zou W, et al. Nomogram prediction of vessels encapsulating tumor clusters in small hepatocellular carcinoma ≤3 cm based on enhanced magnetic resonance imaging. World J Gastrointest Oncol 2024;16(5):1808-1820 [FREE Full text] [doi: 10.4251/wjgo.v16.i5.1808] [Medline: 38764811]
32. Fan Y, Yu Y, Wang X, Hu M, Du M, Guo L, et al. Texture analysis based on Gd-EOB-DTPA-enhanced MRI for identifying vessels encapsulating tumor clusters (VETC)-positive hepatocellular carcinoma. J Hepatocell Carcinoma 2021;8:349-359 [FREE Full text] [doi: 10.2147/JHC.S293755] [Medline: 33981636]
33. Zhang C, Zhong H, Zhao F, Ma Z, Dai Z, Pang G. Preoperatively predicting vessels encapsulating tumor clusters in hepatocellular carcinoma: machine learning model based on contrast-enhanced computed tomography. World J Gastrointest Oncol 2024;16(3):857-874 [FREE Full text] [doi: 10.4251/wjgo.v16.i3.857] [Medline: 38577448]
34. Fan Y, Yu Y, Hu M, Wang X, Du M, Guo L, et al. Imaging features based on Gd-EOB-DTPA-enhanced MRI for predicting vessels encapsulating tumor clusters (VETC) in patients with hepatocellular carcinoma. Br J Radiol 2021;94(1119):20200950 [FREE Full text] [doi: 10.1259/bjr.20200950] [Medline: 33417489]
35. Yu Y, Cao L, Shen B, Du M, Gu W, Gu C, et al. Deep learning radiopathomics models based on contrast-enhanced MRI and pathologic imaging for predicting vessels encapsulating tumor clusters and prognosis in hepatocellular carcinoma. Radiol Imaging Cancer 2025;7(2):e240213. [doi: 10.1148/rycan.240213] [Medline: 40084948]
36. Wang F, Numata K, Funaoka A, Liu X, Kumamoto T, Takeda K, et al. Establishment of nomogram prediction model of contrast-enhanced ultrasound and Gd-EOB-DTPA-enhanced MRI for vessels encapsulating tumor clusters pattern of hepatocellular carcinoma. Biosci Trends 2024;18(3):277-288 [FREE Full text] [doi: 10.5582/bst.2024.01112] [Medline: 38866488]
37. Ding Q, Deng X, Huang J, Zhang R, Liu T, Wang J, et al. Application value of enhanced CT imaging features in predicting vessels encapsulating tumor clusters (VETC) positivity in hepatocellular carcinoma. Curr Med Imaging 2025;21:e15734056361565. [doi: 10.2174/0115734056361565250530050624] [Medline: 40511650]
38. Zhang J, Liu M, Qu Q, Lu M, Liu Z, Yan Z, et al. Radiomics analysis of gadoxetic acid-enhanced MRI for evaluating vessels encapsulating tumour clusters in hepatocellular carcinoma. Front Oncol 2024;14:1422119 [FREE Full text] [doi: 10.3389/fonc.2024.1422119] [Medline: 39193385]
39. Feng Z, Li H, Zhao H, Jiang Y, Liu Q, Chen Q, et al. Preoperative CT for characterization of aggressive macrotrabecular-massive subtype and vessels that encapsulate tumor clusters pattern in hepatocellular carcinoma. Radiology 2021;300(1):219-229. [doi: 10.1148/radiol.2021203614] [Medline: 33973839]
40. Yu Y, Liang X, Hou G, Chen X, Hou W, Hou H, et al. Spectral computed tomography parameters for predicting vessels encapsulating tumor clusters (VETC) pattern in hepatocellular carcinoma: a pilot study. Quant Imaging Med Surg 2025;15(4):3285-3297 [FREE Full text] [doi: 10.21037/qims-24-2077] [Medline: 40235763]
41. Matsuda K, Ueno A, Tsuzaki J, Kurebayashi Y, Masugi Y, Yamazaki K, et al. Vessels encapsulating tumor clusters contribute to the intratumor heterogeneity of HCC on Gd-EOB-DTPA-enhanced MRI. Hepatol Commun 2025;9(1):e0593 [FREE Full text] [doi: 10.1097/HC9.0000000000000593] [Medline: 39670871]

XSL•FO
RenderX

42. Yang J, Dong X, Jin S, Wang S, Wang Y, Zhang L, et al. Radiomics model of dynamic contrast-enhanced mri for evaluating vessels encapsulating tumor clusters and microvascular invasion in hepatocellular carcinoma. Acad Radiol 2025;32(1):146-156 [FREE Full text] [doi: 10.1016/j.acra.2024.07.007] [Medline: 39025700]
43. Pan J, Huang H, Zhang S, Zhu Y, Zhang Y, Wang M, et al. Intraindividual comparison of CT and MRI for predicting vessels encapsulating tumor clusters in hepatocellular carcinoma. Eur Radiol 2025;35(1):61-72. [doi: 10.1007/s00330-024-10944-9] [Medline: 38992109]
44. Chernyak V. Editorial for "Deep learning radiomics model of dynamic contrast-enhanced MRI for evaluating vessels encapsulating tumor clusters and prognosis in hepatocellular carcinoma". J Magn Reson Imaging 2024;59(1):120-121. [doi: 10.1002/jmri.28775] [Medline: 37165916]
45. Li Z, Song W, Zhang J, Li Q, Song Z, Ren X, et al. Identification of vessels encapsulating tumor clusters in solitary hepatocellular carcinoma via imaging biomarkers in preoperative contrast-enhanced magnetic resonance imaging. Quant Imaging Med Surg 2024;14(12):8586-8600 [FREE Full text] [doi: 10.21037/qims-24-315] [Medline: 39698687]
46. Yu Y, Fan Y, Wang X, Zhu M, Hu M, Shi C, et al. Gd-EOB-DTPA-enhanced MRI radiomics to predict vessels encapsulating tumor clusters (VETC) and patient prognosis in hepatocellular carcinoma. Eur Radiol 2022;32(2):959-970. [doi: 10.1007/s00330-021-08250-9] [Medline: 34480625]
47. Guan R, Lin W, Zou J, Mei J, Wen Y, Lu L, et al. Development and validation of a novel nomogram for predicting vessels that encapsulate tumor cluster in hepatocellular carcinoma. Cancer Control 2022;29:10732748221102820 [FREE Full text] [doi: 10.1177/10732748221102820] [Medline: 35609265]
48. Yang J, Dong X, Wang F, Jin S, Zhang B, Zhang H, et al. A deep learning model based on MRI for prediction of vessels encapsulating tumour clusters and prognosis in hepatocellular carcinoma. Abdom Radiol (NY) 2024;49(4):1074-1083. [doi: 10.1007/s00261-023-04141-3] [Medline: 38175256]
49. Xu W, Zhang H, Zhang R, Zhong X, Li X, Zhou W, et al. Deep learning model based on contrast-enhanced ultrasound for predicting vessels encapsulating tumor clusters in hepatocellular carcinoma. Eur Radiol 2025;35(2):989-1000. [doi: 10.1007/s00330-024-10985-0] [Medline: 39066894]
50. Meng X, Qu X, Guo Y, Qi X, Bian L, Wu D, et al. Validation of proposed imaging criteria for estimating vessels encapsulating tumor clusters in hepatocellular carcinoma at CT and gadoxetic acid-enhanced MRI. Eur J Radiol 2025;183:111936. [doi: 10.1016/j.ejrad.2025.111936] [Medline: 39848126]
51. Wang Y, Wang M, Cao L, Huang H, Cao S, Tian X, et al. A nomogram for preoperative prediction of vessels encapsulating tumor clusters (VETC) pattern and prognosis of hepatocellular carcinoma. Am J Surg 2024;234:172-178. [doi: 10.1016/j.amjsurg.2024.05.004] [Medline: 38755026]
52. Ruan L, Yu J, Lu X, Numata K, Zhang D, Liu X, et al. A nomogram based on features of ultrasonography and contrast-enhanced CT to predict vessels encapsulating tumor clusters pattern of hepatocellular carcinoma. Ultrasound Med Biol 2024;50(12):1919-1929. [doi: 10.1016/j.ultrasmedbio.2024.08.020] [Medline: 39289116]
53. Rhee H, Park YN, Choi J. Advances in understanding hepatocellular carcinoma vasculature: implications for diagnosis, prognostication, and treatment. Korean J Radiol 2024;25(10):887-901 [FREE Full text] [doi: 10.3348/kjr.2024.0307] [Medline: 39344546]
54. Wang M, Wang Y, Shen Y, Cao L, Yan R, Lei J. Role of vessels encapsulating tumor clusters patterns in hepatocellular carcinoma: a literature review. Eur J Gastroenterol Hepatol 2025. [doi: 10.1097/MEG.0000000000003032] [Medline: 40631491]
55. Lu M, Wang C, Zhuo Y, Gou J, Li Y, Li J, et al. Preoperative prediction power of radiomics and non-radiomics methods based on MRI for early recurrence in hepatocellular carcinoma: a systemic review and meta-analysis. Abdom Radiol (NY) 2024;49(10):3397-3411. [doi: 10.1007/s00261-024-04356-y] [Medline: 38704783]
56. Wang B, Jiang B, Liu D, Zhu R. Early predictive accuracy of machine learning for hemorrhagic transformation in acute ischemic stroke: systematic review and meta-analysis. J Med Internet Res 2025;27:e71654 [FREE Full text] [doi: 10.2196/71654] [Medline: 40408765]
57. Qi Y, Yang S, Li J, Xing H, Su Q, Wang S, et al. Development and validation of a nomogram to predict impacted ureteral stones via machine learning. Minerva Urol Nephrol 2024;76(6):736-747 [FREE Full text] [doi: 10.23736/S2724-6051.24.05856-7] [Medline: 39093225]
58. Wu C, Zhu S, Wang Q, Xu Y, Mo X, Xu W, et al. Development, validation, and visualization of a novel nomogram to predict depression risk in patients with stroke. J Affect Disord 2024;365:351-358. [doi: 10.1016/j.jad.2024.08.105] [Medline: 39173927]
59. Bertens LCM, Moons KGM, Rutten FH, van Mourik Y, Hoes AW, Reitsma JB. A nomogram was developed to enhance the use of multinomial logistic regression modeling in diagnostic research. J Clin Epidemiol 2016;71:51-57. [doi: 10.1016/j.jclinepi.2015.10.016] [Medline: 26577433]
60. Nwanosike EM, Conway BR, Merchant HA, Hasan SS. Potential applications and performance of machine learning techniques and algorithms in clinical practice: a systematic review. Int J Med Inform 2022;159:104679. [doi: 10.1016/j.ijmedinf.2021.104679] [Medline: 34990939]

XSL•FO
RenderX

61. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020;20(1):310 [FREE Full text] [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
62. Qi X, Wang S, Fang C, Jia J, Lin L, Yuan T. Machine learning and SHAP value interpretation for predicting comorbidity of cardiovascular disease and cancer with dietary antioxidants. Redox Biol 2025;79:103470 [FREE Full text] [doi: 10.1016/j.redox.2024.103470] [Medline: 39700695]
63. Li D, Lu H, Wu J, Chen H, Shen M, Tong B, et al. Development of machine learning models for predicting depressive symptoms in knee osteoarthritis patients. Sci Rep 2024;14(1):28603 [FREE Full text] [doi: 10.1038/s41598-024-79601-x] [Medline: 39562701]
64. Zhou T, Cheng Q, Lu H, Li Q, Zhang X, Qiu S. Deep learning methods for medical image fusion: a review. Comput Biol Med 2023;160:106959. [doi: 10.1016/j.compbiomed.2023.106959] [Medline: 37141652]
65. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol 2019;20(5):e262-e273. [doi: 10.1016/S1470-2045(19)30149-4] [Medline: 31044724]

Abbreviations

DOR: diagnostic odds ratio
HCC: hepatocellular carcinoma
LR–: negative likelihood ratio
LR+: positive likelihood ratio
MRI: magnetic resonance imaging
PRISMA: Preferred Reporting Items for a Systematic Review and Meta-Analysis
ROC AUC: area under the receiver operating characteristic curve
SROC AUC: area under the summary receiver operating characteristic curve
VETC: vessels encapsulating tumor clusters

Edited by S Brini; submitted 22.Aug.2025; peer-reviewed by W Wang, X Liang; comments to author 31.Oct.2025; accepted 03.Dec.2025; published 14.Jan.2026.

Please cite as:
Shui H, Wu W, Xie Z, Yang B, Deng J, Tang D
Effectiveness of Machine Learning in Detecting Vessels Encapsulating Tumor Clusters in Hepatocellular Carcinoma: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e82839
URL: https://www.jmir.org/2026/1/e82839
doi:10.2196/82839
PMID:

©Huili Shui, Wenyu Wu, Zhenming Xie, Bing Yang, Jia Deng, Dongxin Tang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Machine Learning Prediction Models for Preeclampsia: Systematic Review and Meta-Analysis

Lu Liu[1,2], MM; Qixuan Zhu[3], BEng; Yichi Zong[2], MM; Xueyuan Chen[2], MM; Wei Zhang[2], MM; Jun Wang[4], PhD

[1]School of Public Health, China Medical University, Shenyang City, Liaoning Province, China

[2]Department of the Obstetrics and Gynecology, Shengjing Hospital Affiliated to China Medical University, Shenyang City, Liaoning Province China, China

[3]School of Engineering, University of Pennsylvania, Philadelphia, PA, United States

[4]Department of the Obstetrics and Gynecology, Shengjing Hospital Affiliated to China Medical University, Shenyang City, Liaoning Province, China

Corresponding Author:
Jun Wang, PhD
Department of the Obstetrics and Gynecology
Shengjing Hospital Affiliated to China Medical University
Heping District, No 36 Sanhao Street
Shenyang City, Liaoning Province, 110000
China
Phone: 86 18940254480
Email: wangj1@sj-hospital.org

Abstract

Background: Preeclampsia is a severe hypertensive disorder with rising global prevalence. While machine learning (ML) models for predicting preeclampsia are increasingly published, existing evidence shows high heterogeneity, and the distinction between internal performance and external transferability remains unclear.

Objective: This study aims to evaluate the performance of ML models in predicting preeclampsia through a systematic review and meta-analysis, while also exploring their potential clinical application value, in order to specifically enhance the quality of future research and the predictive capability of the models.

Methods: Following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines and PROSPERO registration, we searched PubMed, Web of Science, IEEE Xplore, and CNKI (China National Knowledge Infrastructure) for studies published through February 2025. We included studies using ML to predict preeclampsia in pregnant women. Bias was assessed using PROBAST (Prediction model Risk of Bias Assessment Tool). We calculated summary estimates using random-effects models and, crucially, computed 95% prediction intervals (PIs) to estimate performance in future clinical settings. Subgroup and meta-regression analyses were conducted to explore heterogeneity.

Results: In total, 26 studies comprising 31 ML models were included. While the pooled area under the receiver operating characteristic curve was high at 0.91 (95% CI 0.87-0.92), extreme heterogeneity was observed (I^2>99%). The 95% PI for sensitivity was wide (0.32-0.96), indicating that in some external settings, sensitivity could drop to 32%. Only 6 studies conducted external validation; in these, the pooled sensitivity decreased to 0.68, with a PI of 0.25-0.94.Subgroup analysis suggested that models incorporating laboratory biomarkers and neural networks outperformed others, though CIs overlapped.

Conclusions: Current evidence suggests that a high area under the curve in ML models is more likely to reflect the “performance” of the model on the internal development dataset rather than its universal “effectiveness” and clinical utility in independent, diverse populations. The apparent performance exhibits significant contextual dependence. Future studies should conduct multicenter, prospective external validation and recalibration research to enhance transferability and reliability.

Trial Registration: PROSPERO CRD420251005830;https://www.crd.york.ac.uk/PROSPERO/view/CRD420251005830

(J Med Internet Res 2026;28:e78714) doi:10.2196/78714

KEYWORDS

preeclampsia; machine learning; predictive models; meta-analysis; artificial intelligence; computer-assisted diagnosis

XSL•FO
RenderX

Introduction

Preeclampsia is a pregnancy-related hypertensive condition marked by the development of high blood pressure and protein in the urine after 20 weeks of gestation. Due to its multiple etiologies and complex pathogenesis, it poses significant risks to both maternal and perinatal health [1]. This specific condition negatively impacts maternal health and can also lead to serious complications for the fetus, including placental abruption and restricted fetal growth. According to global statistics, the incidence of preeclampsia ranges from 3% to 9%, with even higher rates observed in certain high-risk populations [2]. Furthermore, preeclampsia is one of the leading causes of maternal mortality worldwide, particularly in low- and middle-income countries. The prevalence of preeclampsia in China has increased from 5.79% in 2005 to 9.5% in 2019 [3], further underscoring the urgent need for early screening and management. To date, the etiology and pathogenesis of preeclampsia remain incompletely understood, and effective treatment measures are lacking. Consequently, early detection and enhanced management are essential clinical strategies.

Understanding the epidemiological characteristics of preeclampsia is essential for developing effective public health strategies. In the study of preeclampsia, traditional statistical methods primarily emphasize linear models and hypothesis testing, which are effective in uncovering singular relationships between variables. However, the pathological mechanisms underlying preeclampsia are highly complex, involving multiple interacting factors, and traditional methods may face limitations when addressing nonlinear and high-dimensional data. In contrast, machine learning (ML) technology has shown considerable promise in this domain.

A subset of artificial intelligence (AI), ML is a technology that allows computers to independently learn from data and make decisions or predictions using algorithms and models. Its application in clinical settings can effectively prevent and manage diseases. Currently, the usage of ML to develop predictive models for preeclampsia is becoming increasingly prevalent. For instance, Sylvain et al [4] noted that the implementation of ML methods has significantly improved the prediction accuracy of high-risk pregnancies, offering a novel perspective for the early identification of preeclampsia. Furthermore, Ranjbar et al [5] indicated that ML-based models surpass traditional regression models in predicting the incidence of preeclampsia. The multidimensional optimization capabilities of these models allow them to account for interactions among various clinical features and biomarkers, thereby enhancing diagnostic accuracy.

By leveraging ML, researchers can explore both linear and nonlinear relationships, as well as uncover deep-seated features and patterns within the data. This method establishes a scientific foundation for the prompt recognition and intervention of preeclampsia.

Compared with prior systematic reviews and protocols on pregnancy outcomes or preeclampsia, the incremental contributions of this study are as follows: (1) we prespecified and implemented subgroup analyses by outcome definition, gestational window, data source, and validation type to avoid indiscriminate pooling across highly heterogeneous models and populations; (2) we treated area under the curve (AUC) as the primary summary measure and applied robust univariate random-effects models (Hartung-Knapp-Sidik-Jonkman method) to pool sensitivity and specificity separately, accompanied by 95% prediction intervals (PIs) to estimate future performance; and 3) we clearly separated performance in internal vs external validation and documented whether decision-curve analysis was conducted. Taken together, these methodological enhancements aim to provide more interpretable evidence about where deployment may be appropriate and where it remains premature.

Methods

Research Design

This research was carried out in alignment with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 standards [6] (Multimedia Appendix 1 [7]). Specific details regarding the search keywords can be found in Textbox S1 of the Multimedia Appendix 2. Before the study began, the protocol received approval and was registered with the PROSPERO under the reference number CRD420251005830.

Literature Search Strategy

Comprehensive searches were executed in several prestigious databases, including PubMed, Web of Science, IEEE Xplore, and the CNKI (China National Knowledge Infrastructure). These searches focused on locating scholarly papers that were published in either English or Chinese. The time frame for this search encompassed works published until February 2025, ensuring that the most recent and relevant literature was included in the investigation. The search strategy was developed based on the PICO (Population, Intervention, Comparison, and Outcome) framework. In this study, “P” denotes the population with PE, “I” refers to ML methods as the intervention, “C” indicates the gold standard for comparison, and “O” encompasses outcomes, such as sensitivity, specificity, and accuracy for prediction and diagnosis (Table S1 in Multimedia Appendix 2). Additionally, the reference lists from each identified study underwent a manual review to uncover further relevant research. Zotero (Center for History and New Media at George Mason University) was used to organize the studies and remove any duplicates.

The study’s inclusion criteria were formulated to guarantee the rigor and relevance of the research. The criteria encompassed (1) research papers published in English or Chinese; (2) investigations involving pregnant women from the general population that explicitly defined the diagnosis of preeclampsia; (3) studies that used ML models for predicting preeclampsia, along with a thorough explanation of these models; and (4) investigations that showcased the performance of the ML models, offering adequate data to determine both sensitivity and specificity. These criteria aimed to strengthen the validity of the results and ensure a thorough assessment of the existing literature.

XSL•FO
RenderX

The exclusion criteria for this study are as follows: (1) studies that solely investigated risk factors without developing a predictive model; (2) papers published in languages other than English or of types other than original research, such as reports and reviews; (3) duplicate publications; (4) studies that included 2 or fewer predictors in the constructed model; and (5) studies for which the full text was not accessible.

Literature Screening and Data Extraction

Five researchers (LL, QZ, YZ, XC, and WZ) meticulously followed the established inclusion and exclusion criteria to screen the titles and abstracts of the literature. Studies that met these criteria advanced to the full-text reading phase, where all relevant studies were reviewed. Each article underwent a minimum of 2 rounds of screening. Both the title and abstract screening, as well as the full-text reading, were conducted independently by the 2 researchers (LL and QZ). In instances of disagreement between them, another researcher (JW) made the final decision.

In total, 26 studies [8-33] were chosen for analysis. Data extraction was independently performed by 2 researchers (LL and QZ) following the standardized protocol established by the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis), as outlined in the existing literature [34]. Data collected from each study included the following: (1) demographic details, such as the country of data collection, the study setting, the source of the data, the design of the study, and the definition of outcomes; (2) methods for data partitioning, feature selection algorithms, types of ML prediction models, model validation, and applications; (3) results of predictions, which involved accuracy, sensitivity, specificity, and the AUC; and (4) sources of funding and the approval of ethics. This study extracted sensitivity and specificity data from each research report, all based on the "optimal threshold" set in the respective original studies. This research did not standardize or adjust for the differences in thresholds among the various studies.

Bias and Applicability Assessment

Overview

We used PROBAST (Prediction Model Risk of Bias Assessment Tool) as the primary instrument to preserve comparability with prior preeclampsia meta-analyses (for detailed information, see Multimedia Appendix 3). Because many included studies predate PROBAST-AI and lack AI-specific reporting (eg, leakage safeguards, hyperparameter tuning, calibration, and thresholds), a full PROBAST-AI assessment would be dominated by underreporting rather than demonstrated bias. The PROBAST [35] was used to assess the risk of bias in the included studies across 4 domains, namely participants, predictors, outcomes, and analysis. Additionally, applicability assessments were conducted for the domains of population, predictors, and outcomes. Two researchers (LL and QZ) independently reviewed the studies, undergoing consistency training based on a preprepared and trialed scoring manual. The discrepancies were resolved through discussion, and if necessary, a third researcher (JW) acted as an adjudicator.

Bias Assessment

For all questions within a category, if the answers are "yes" or "possibly," the category is assessed as low risk. Conversely, if any answer is "no" or "possibly not," the category is classified as high risk. In cases where there is insufficient information, the category is deemed unclear. The overall risk of bias in the study is determined according to the PROBAST guidelines: (1) if all 4 domains are assessed as low risk, the overall risk of the study is low; (2) if one or more domains are assessed as high risk, the overall risk of the study is high; and (3) if one or more domains are assessed as unclear (and there are no high-risk domains), the overall risk of the study is unclear.

Applicability Assessment

The evaluation encompasses 3 categories, including study object, predictor, and outcome. Each category is assessed based on 3 levels of applicability, namely good applicability, poor applicability, and unclear applicability. If all 3 assessments are classified as good, the overall applicability is determined to be good. Conversely, if any one assessment is classified as poor, the overall applicability is deemed poor. In cases where one assessment is unclear while the other two are good, the overall applicability is classified as unclear.

Statistical Analysis

The methods described in the guidelines for conducting systematic reviews and meta-analyses concerning the performance of prediction models, along with previous meta-analyses of such models, indicate that the concordance index of a model is similar to the AUC [36]. This index indicates the diagnostic or prognostic discrimination ability, categorized as none (AUC≤0.6), poor (0.6<AUC<0.7), moderate (0.7<AUC<0.8), good (0.8<AUC<0.9), or excellent (0.9<AUC<1). Model calibration acts as an indicator of how well the model fits the data by evaluating the alignment between the actual and forecasted results, while also demonstrating the model's reliability via calibration graphs. Additionally, the diagnostic odds ratio (DOR) is calculated using the following formula:

$$DOR=PLR/NLR$$

In this study, we use the positive likelihood ratio (PLR) and the negative likelihood ratio (NLR) to evaluate the predictive performance of our model for preeclampsia. The equations used to calculate PLR and NLR express the frequency of preeclampsia in individuals who are predicted by the model to have preeclampsia compared to those who are predicted not to have preeclampsia:

$$PLR=Sensitivity/(1-Specificity)$$

$$NLR=(1-Sensitivity)/Specificity$$

Considering the diversity in populations, predictors, and algorithms across the included ML models, our objective was to generalize findings to broader clinical contexts. Therefore, following the recommendation of Borenstein et al [37], we a priori selected the random-effects model for all meta-analyses, irrespective of the magnitude of statistical heterogeneity (I^2). Specifically, we used the more robust

XSL•FO
RenderX

Hartung-Knapp-Sidik-Jonkman (HKSJ) method for final pooled estimates and interval calculations to ensure the robustness of statistical inferences [38]. The ML models included in this study exhibited substantial variations in sample size and population characteristics, with the I^2 statistic often approaching 100% in larger samples, potentially limiting their ability to effectively distinguish the actual clinical impact of heterogeneity. Therefore, in addition to reporting the 95% CI for pooled effect sizes, this study further calculated the 95% PI. Unlike CIs, which only reflect the precision of the average effect, PIs estimate the expected range of performance when the model is applied in a new, similar clinical setting in the future. This approach provides a more intuitive assessment of the model's clinical applicability and transferability [38]. Since the Meta-DiSC software (The developer is the clinical biostatistics team at Ramón y Cajal Hospital) cannot calculate PIs, we used the *meta* package (version 7.0) [39] in R software (R Foundation for Statistical Computing; version 4.4.2) with the HKSJ method to compute 95% PIs for area under the receiver operating characteristic curve (AUROC), sensitivity, and specificity. For AUROC values without reported SEs, we estimated them based on sample size using the Hanley & McNeil [40] method. External validation is regarded as the "gold standard" for assessing the transportability of models. Therefore, a separate evaluation of the performance of models that use external validation is conducted. Subsequently, the 4 predictive models with the highest and lowest values were excluded to conduct a sensitivity analysis aimed at evaluating the impact of outliers on the sensitivity and specificity of the summary. To reduce conceptual heterogeneity and enhance the interpretability of results, stratification is performed along the following dimensions: sample size (less than 2000 and greater than or equal to 2000); data source (electronic medical records; laboratory biomarkers; omics or imaging; mixed); gestational age window (early pregnancy; midpregnancy and late pregnancy or specific gestational weeks); and validation methods (internal validation and external validation); ML models (logistic regression [LR] and nonlogistic regression), followed by more detailed subgroup analysis (LR, extreme gradient boosting [XGBoost], random forest [RF], and support vector machine [SVM]) based on nonlogistic regression; types of predictive variables (demographic information; biological genetic markers; laboratory tests; demographic information and laboratory tests); and the number of predictive variables (less than 10 and greater than or equal to 10). Handling of missing data (extraction and synthesis). For each study, we recorded how missing data were handled and classified methods into 5 categories, namely listwise deletion, single-value imputation (eg, mean and median), multiple imputation, other (eg, random subset iterations), and not reported. When multiple approaches were mentioned, we coded the method used for the primary model. We summarize the overall distribution in the results of "Inclusion of Study Characteristics in the Paper" and discuss implications for comparability and generalizability. Subgroup analyses will be conducted on the included studies to evaluate the performance of ML methods in predicting preeclampsia across different clinical scenarios. Subgroup Analysis discusses the capabilities of different ML algorithms in predicting preeclampsia. Additionally, meta-regression was used to investigate the sources of heterogeneity. Given the extreme heterogeneity (I^2>99%) observed across studies and the lack of standardized threshold reporting (eg, fixed false-positive rates), hierarchical or bivariate models often fail to converge or yield unstable estimates. Therefore, we prioritized univariate random-effects models using the HKSJ adjustment for pooling sensitivity and specificity separately. This method is demonstrated to provide more robust coverage probabilities for CIs in the presence of substantial heterogeneity compared to standard DerSimonian-Laird [41] methods.

Results

Literature Screening

After removing duplicate entries, a total of 284 papers were evaluated. Of these, 284 papers were evaluated through abstract screening, which was subsequently followed by a full-text evaluation of 88 papers. This process culminated in the identification of 26 papers [8-33] that satisfied the overall inclusion criteria. The literature screening procedure and its outcomes are depicted in the related Figure 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for study selection. CNKI: China National Knowledge Infrastructure; PE: preeclampsia.

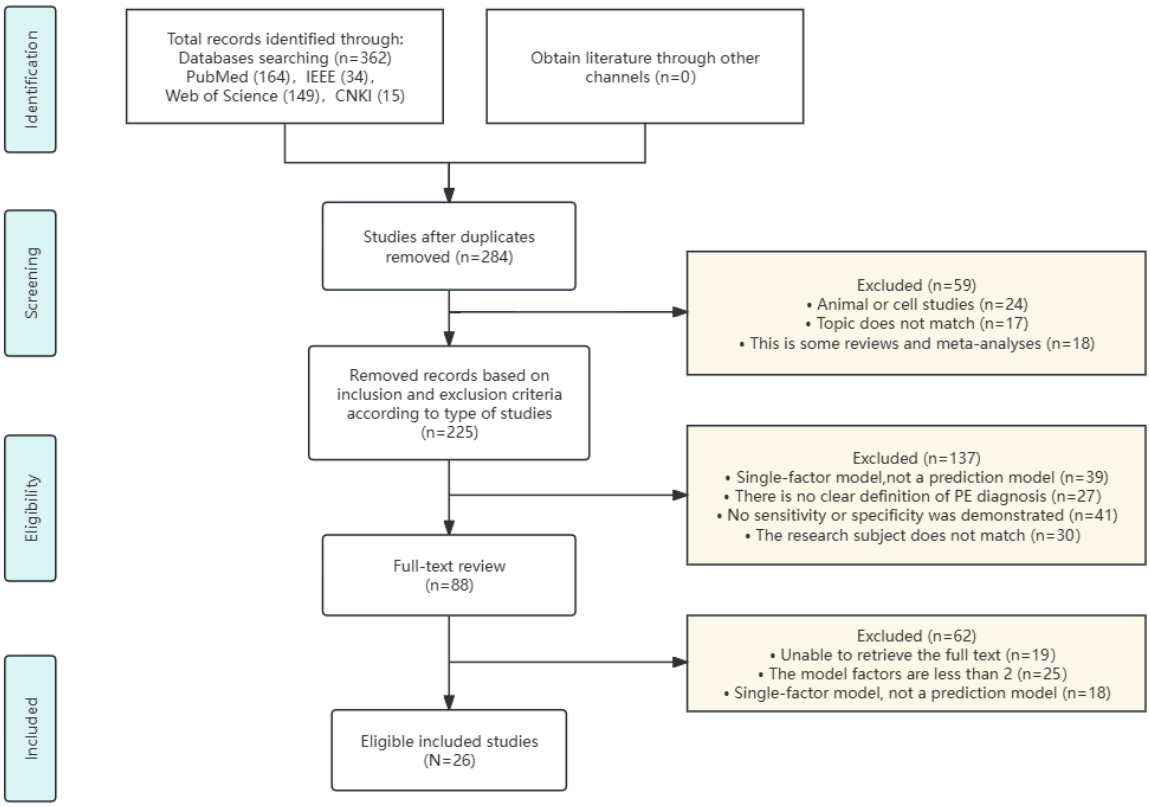

Inclusion of Study Characteristics in the Paper

The literature included in this study spans from 2019 to 2025 and consists of 23 English papers [8-10,12-26,28-30,32,33] and 3 Chinese papers [11,27,31]. When a study presented more than 2 models, the top 2 models demonstrating the best performance were selected based on a comprehensive evaluation of metrics, such as AUC, sensitivity, and specificity, culminating in the inclusion of 31 models from 26 papers [8-33]. The data sources for ML predominantly consisted of clinical electronic health records, community research cohorts, and self-administered questionnaires. The overall sample sizes in the studies examined showed considerable variation, fluctuating between 53 and 62,562 cases, while the count of predictors in the ultimate models ranged from 3 to 50. Among all the studies, 20 [8,11,13,14,16-18,20-32] conducted internal validation, while 6 [8,21,24,28,30,32] performed external validation. To assess model performance, the AUC, sensitivity, and specificity emerged as the most frequently used metrics. Among the 26 studies [8-33] reviewed, 5 (19.2%) studies [8,17,25,26,33] were prospective cohort studies, 17 (65.4%) studies [10,11,13-16,18,19,21-24,27-29,31,32] were retrospective cohort studies, 2 (7.7%) studies [9,20] were case-control studies, 1 (3.8%) study [30] was a retrospective case-control study, and 1 (3.8%) study [12]was a multicenter study. Regarding model approaches, of the 31 models included, 3 were LR. Among the remaining 28 models, there were 5 RF, 4 XGBoost, 4 Elastic-net, 3 neural network (NN), 3 SVM, 2 light gradient boosting, 2 AdaBoost, 1 k-nearest neighbor, 1 Naive Bayes, 1 stochastic gradient boosting, 1 CatBoost, and 1 voting classifier. In terms of handling missing data, 8 studies [11,18,22,24-27,29] opted to delete cases with missing data, 7 studies [9,12,14-16,19,23] used mean imputation to address the missing values, 3 studies [13,17,31] used multiple imputation techniques, 1 study [21] implemented random selection of data subsets for multiple iterative analyses, while the remaining 7 studies [8,10,20,28,30,32,33] did not explicitly report the presence of missing values. Such variation limits comparability and external transportability of performance metrics and increases uncertainty around calibration and threshold transfer. The specific details of the models are presented in Table 1.

XSL•FO
RenderX

Table 1. Construction of the risk prediction model for preeclampsia.

Literature and modeling method	Model performance			Sample size (modeling; internal validation; external validation)	Missing data		Predictors
	AUC[a]	Sensitivity	Specificity		Quantity (PCS[b])	Handling method	
Ansbacher et al [8]							
FfNN[c]	0.816	0.533	0.9	30437/10000/20352	—[d]	—	10 predictors: maternal age, maternal weight, maternal height, interpregnancy interval, ethnicity, medical history (such as chronic hypertension, diabetes, etc), uterine artery pulsatility index, mean arterial pressure, placental growth factor, and pregnancy-associated plasma protein-A.
Araújo et al [9]							
LGB[e]	0.9	0.95	0.79	132/—/—	—	Mean imputation	3 predictors: neutrophil count, mean corpuscular hemoglobin, and aggregate index of systemic inflammation.
Chen et al [10]							
SVM[f]	0.88	0.87	0.76	166/—/—	—	—	7 predictors: IL-17, IL-21, IL-22, IL-10, transforming growth factor-β, placental alkaline phosphatase, and lysosome-associated membrane protein 3.
Chen et al [11]							
CB[g]	0.983	0.8881	0.9848	1325/398/—	—	Delete	18 predictors: BMI, systolic blood pressure, diastolic blood pressure, number of pregnancies, mean corpuscular hemoglobin concentration, bacteria (urinalysis), glycocholic acid, high-density lipoprotein, potassium, sodium, phosphorus, uric acid, urine protein, creatinine, direct bilirubin, low-density lipoprotein, gestational age≥34 weeks, and family history of hypertension.
Giménez et al [12]				597/—/—	—	Mean imputation	6 predictors: gestational age, history of chronic hypertension, Soluble FMS-like Tyrosine Kinase-1, placental growth factor, N-terminal pro-brain natriuretic peptide, and uric acid.
PTB-RF[h]	0.901	0.796	0.91				
RF[i]	0.941	0.775	0.949				
Jhee et al [13]							
SGB[j]	0.924	0.603	0.991	7704/3302/—	25	Multiple Imputation	14 predictors: systolic blood pressure, serum urea nitrogen, serum creatinine, platelet count, serum potassium level, white blood cell count, serum calcium level, and urinary protein.
Kaya et al [14]							
XGBoost[k]	0.767	0.6	0.833	53/20/—	—	Mean imputation	8 predictors: maternal age, BMI, smoking status, history of diabetes, history of gestational diabetes, mean arterial pressure, and history of previous preeclampsia.
Kovacheva et al [15]				1125/—/—	—	Mean imputation	7 predictors: maternal age, BMI, systolic blood pressure, diastolic blood pressure, uric acid, history of kidney disease, and SBP PRS[m].

XSL•FO
RenderX

Literature and modeling method	Model performance			Sample size (modeling; internal validation; external validation)	Missing data		Predictors
	AUC[a]	Sensitivity	Specificity		Quantity (PCS[b])	Handling method	
LR[l]	0.83	0.85	0.66				
XGBoost	0.91	0.96	0.44				
Li et al [16]							
XGBoost	0.955	0.789	0.93	3759/191/—	—	Mean imputation	38 predictors: maternal age, BMI, mean blood pressure, abdominal circumference, gravidity, parity, history of preeclampsia, history of previous cesarean section, interpregnancy interval, primipara, multiple gestation, assisted reproductive technology, heart disease, pregestational diabetes, thyroid disease, kidney disease, autoimmune disease, mental illness, uterine fibroids, adenomyosis, uterine malformation, history of epilepsy, family history of hypertension, hemoglobin, white blood cell count, platelet count, creatinine, fasting blood glucose, total cholesterol, high-density lipoprotein, low-density lipoprotein, total protein, albumin, bile acids, uric acid, total bilirubin, direct bilirubin, and gamma-glutamyl transferase.
Li et al [17]							
VC[n]	0.831	0.77	0.769	3715/929/—		Multiple Imputation	16 predictors: maternal age, height, prepregnancy weight, primiparity, mode of conception, family history, smoking status, history of preeclampsia, history of chronic hypertension, history of chronic kidney disease, history of diabetes, history of systemic lupus erythematosus/antiphospholipid syndrome, mean arterial pressure, uterine artery pulsatility index, pregnancy-associated placental protein a, and placental growth factor.
Lv et al [18]							
XGBoost	0.963	0.917	0.894	832/208/—	—	Delete	6 predictors: prepregnancy BMI, gravidity, mean arterial pressure, smoking, alpha-fetoprotein, and conception method.
Marić et al [19]							

XSL•FO
RenderX

Literature and modeling method	Model performance			Sample size (modeling; internal validation; external validation)	Missing data		Predictors
	AUC[a]	Sensitivity	Specificity		Quantity (PCS[b])	Handling method	
EN[o]	0.79	0.452	0.919	5245/—/—	—	Mean imputation	55 predictors: maternal age, height, weight, ethnicity, number of fetuses, mean systolic blood pressure, mean diastolic blood pressure, maximum systolic blood pressure, maximum diastolic blood pressure, history of preeclampsia, chronic hypertension, type 1 and type 1 diabetes, gestational diabetes, obesity, assisted reproductive technology, diagnosis of autoimmune diseases, kidney disease, anemia, antiphospholipid syndrome, sexually transmitted diseases, hyperemesis gravidarum, headache, migraine, poor obstetric history, high-risk pregnancy, protein and glucose in urine, platelet count, red blood cells, white blood cells, creatinine, hemoglobin, hematocrit, monocytes, lymphocytes, eosinophils, neutrophils, basophils, Rh blood type, gastric acid, rubella, chickenpox, hepatitis B virus, syphilis, gonorrhea, aspirin, nifedipine, aldomet, labetalol, insulin, glyburide, prednisone, azathioprine, Plaquenil, heparin, levothyroxine, doxylamine, and acyclovir.
Melinte-Popescu et al [20]							
NB[p]	0.98	0.963	0.964	163/70/—	—	—	14 predictors: age, BMI, smoking status, interpregnancy interval, use of assisted reproductive technology, pregestational diabetes, chronic hypertension, history of kidney disease, personal or family history of preeclampsia, placental growth factor, pregnancy-associated plasma protein A, placental protein 13, uterine artery pulsatility index, and mean arterial pressure.
Munchel et al [21]							
AB[q]	0.964	0.88	0.92	113/11/448	—	Randomly select a subset of data for multiple iterative analyses.	49 predictors circulating transcripts in blood: immunomodulatory, fetal development, angiogenesis, and extracellular matrix remodeling.
Roque et al [22]							
LR	0.976	0.9	0.951	35706/8927/—	—	Delete	11 predictors: platelet count, white blood cell count, lymphocyte percentage, monocyte percentage, red blood cell count, red cell distribution width, platelet distribution width, band neutrophil percentage, red cell distribution width, hematocrit, and maternal age.
Sandström et al [23]							

XSL•FO
RenderX

Literature and modeling method	Model performance			Sample size (modeling; internal validation; external validation)	Missing data		Predictors
	AUC[a]	Sensitivity	Specificity		Quantity (PCS[b])	Handling method	
LR	0.67	0.282	0.9	62562/6256/—	—	Mean imputation	36 predictors: gestational age at first visit, maternal age, BMI, mean arterial pressure, capillary blood glucose level, urine protein, hemoglobin level, history of miscarriage, history of ectopic pregnancy, history of infertility treatment, family status, country of birth, smoking history, smoking status at registration, use of snuff in the first trimester of pregnancy, use of snuff during pregnancy, alcohol consumption in the 3 months before registration, alcohol consumption habits at the time of pregnancy registration, family history of preeclampsia, infertility, family history of hypertension, previous diabetes, chronic hypertension, chronic kidney disease, cardiovascular disease, endocrine disease, history of thrombosis, history of mental illness, history of epilepsy, Crohn/ulcerative colitis, lung disease or asthma, hepatitis, gynecological disease or surgery, recurrent urinary tract infections, and blood type.
Sufriyana et al [24]							
RF	0.86	0.7	0.89	23201/20975/GEV[r]:1322, TEV[s]: 90	301	Delete	13 predictors: age, family role, parity, type of work, infectious diseases, endocrine, nutritional and metabolic diseases, circulatory system diseases, immune-related diseases, ophthalmic diseases, urogenital diseases, skin and subcutaneous tissue–related diseases, breast-related diseases, digestive system–related diseases, and skin-related diseases.
Tiruneh et al [25]							
RF	0.84	0.76	0.79	33767/14475/—	66	Delete	13 predictors: maternal age, ethnicity, prepregnancy/early pregnancy BMI, history of preeclampsia in previous pregnancies, primiparity, history of gestational diabetes, pre-existing hypertension, diabetes, family history of hypertension and diabetes, family history of preeclampsia, renal disease, smoking history, and polycystic ovary syndrome.
Torres et al [26]				1068/914/—	78	Delete	13 predictors: placental growth factor, mean arterial pressure, uterine artery pulsatility index, BMI, antiphospholipid syndrome, previous preeclampsia, previous diabetes, smoking status, natural conception, Other drug use (such as cocaine and heroin), systemic lupus erythematosus, chronic hypertension, and maternal age.
all-EN	0.778	0.501	0.9				
EPE-EN[t]	0.963	0.882	0.9				
PPE-EN[u]	0.897	0.765	0.9				
Wang et al [27]							
KNN[v]	0.9	0.7142	0.926	516/172/—	—	Delete	7 predictors: urine protein, urine conductivity, alkaline phosphatase, serum uric acid, lactate dehydrogenase, mean corpuscular hemoglobin concentration, and amylase.

XSL•FO
RenderX

Literature and modeling method	Model performance			Sample size (modeling; internal validation; external validation)	Missing data		Predictors
	AUC[a]	Sensitivity	Specificity		Quantity (PCS[b])	Handling method	
Wang et al [28]							
AB	0.8775	0.7271	0.9	25709/77713/1760	—	—	20 predictors: maternal age, maternal BMI, regularity of maternal menstrual cycle, vomiting and nausea during pregnancy, previous miscarriages, preterm births, history of hypertension during pregnancy, hypertension, diabetes, chronic hypertension, history of drug allergies, maternal smoking history, previous delivery history, nutritional status during pregnancy, maternal ethnic background, history of hypertension, history of diabetes, glycated hemoglobin, and albumin.
Xue et al [29]							
SVM	0.93	0.67	0.999	800/160/—	—	Delete	50 predictors: diabetes mellitus, thrombotic diseases, systemic lupus erythematosus, antiphospholipid syndrome, renal diseases, assisted reproductive technology, obstructive sleep apnea syndrome, prepregnancy BMI>30 kg/m², age>35 years, multiple pregnancy, primipara, history of eclampsia or preeclampsia, Albumin, Alanine aminotransferase, Aspartate aminotransferase, Alkaline phosphatase, Complement C1q, Calcium, Creatinine, C-reactive protein, Cystatin C, Gamma-glutamyl transferase, Globulin, Triglycerides, Total cholesterol, High-density lipoprotein cholesterol, Low-density lipoprotein cholesterol, Lipoprotein(a), Apolipoprotein A1, Apolipoprotein B, Small dense low-density lipoprotein, Total protein, Total bile acid, Total bilirubin, Direct bilirubin, Uric acid, Urea, Phosphorus, Absolute Lymphocyte count, Absolute neutrophil count, Platelet count, NEU/LYM ratio, PLT/LYM ratio, Prothrombin time, Prothrombin activity, Activated partial thromboplastin time, Fibrinogen, D-Dimer, Fibrin degradation products, Thrombin time.
Yu et al [30]							
RF	0.96	0.87	0.91	404/1384/899	—	—	12 predictors: maternal age, BMI, parity, medical history (chronic hypertension, preeclampsia, systemic lupus erythematosus, antiphospholipid syndrome), mode of conception; cfDNA profile indicators: Fos-related antigen 2 (FOSL2), calcium/calmodulin-dependent protein kinase kinase 2 (CAMKK2), G1/S-specific cyclin-D1 (CCND1), Inositol 1,4,5-trisphosphate receptor type 1 (ITPR1), Protein kinase A catalytic subunit beta (PRKACB), Protein Wnt-7b (WNT7B), Voltage-dependent L-type calcium channel subunit beta-2(CACNB2), Nuclear respiratory factor 1 (NRF1), Fms-related tyrosine kinase 3 ligand (FLT3LG), Epidermal growth factor (EGF).
Zheng et al [31]							

XSL•FO
RenderX

Literature and modeling method	Model performance			Sample size (modeling; internal validation; external validation)	Missing data		Predictors
	AUC[a]	Sensitivity	Specificity		Quantity (PCS[b])	Handling method	
LGB	0.964	0.849	0.927	1609/483/—	—	Multiple imputation	12 predictors: urine specific gravity, uric acid, mean corpuscular hemoglobin concentration, globulin, platelet distribution width, potassium ion, age, family history of hypertension, systolic blood pressure, diastolic blood pressure, pulse, and gestational age≥34 weeks.
Zhou et al [32]				432/197/288	—	—	19 predictors: mRNA markers: Albumin, Fibrinogen Alpha Chain, Leptin, Insulin-Like Growth Factor Binding Protein 5, Alpha-1 Antitrypsin, S100 Calcium Binding Protein A9, Apolipoprotein A1, Thyroid Stimulating Hormone Beta Subunit, miRNA markers: MIR130A, MIR144, MIR19B1, MIR215, MIR376C, MIR27A, MIR106A, MIR33A, Inc ENA markers**:** Macrophage Migration Inhibitory Factor, Assisted Reproductive Technology, Mean Arterial Pressure.
AvNN[w]	0.91	0.63	0.93				
SVM	0.93	0.47	0.99				
Zhou et al [33]							
CNN[x]	0.883	0.722	0.934	1138/—/—	—	—	8 predictors: Retinal fundus image score, Prepregnancy BMI, maternal age, chronic hypertension, diabetes, history of gestational hypertension or preeclampsia, assisted reproductive technology, and autoimmune diseases.

[a]AUC: area under the curve.
[b]PCS: pieces.
[c]FfNN: feed-forward neural network.
[d]not reported.
[e]LGB: light gradient boosting.
[f]SVM: support vector machine.
[g]CB: CatBoost.
[h]PTB-RF: Premature birth - Random Forest.
[i]RF: random forest.
[f]KNN: k-nearest neighbor.
[j]SGB: stochastic gradient boosting.
[k]XGBoost: extreme gradient boosting.
[l]LR: logistic regression.
[m]SBP PRS: systolic blood pressure polygenic risk score.
[n]VC: Voting Classifier.
[o]EN: Elastic-net.
[p]NB: Naive Bayes.
[q]AB: AdaBoost.
[r]GEV: geographic external validation
[s]TEV: temporal external validation
[t]EPE-EN: early onset of preeclampsia Elastic-net.
[u]PPE-EN: Premature birth of preeclampsia Elastic-net.
[v]KNN: k-nearest neighbor.
[w]AvNN: Average Neural Network.
[x]CNN: Convolutional Neural Networks.

XSL•FO
RenderX

Research Quality

We evaluated the potential for bias and the relevance of the prediction models based on the PROBAST checklist, examining a total of 26 [8-33] studies. Among these, 3 (12%) studies [9,11,16] in the participant domain exhibited unclear risk of bias, primarily due to their case-control design, which is inherently associated with a higher risk of selection bias. In the predictor domain, 1 (4%) study [21] was identified as having unclear risk of bias because it used C-RNA transcriptome assays that depend on transcriptome enrichment and high-throughput sequencing, methods that are not typically used in routine clinical testing. In the analysis of bias domains, 8 (31%) studies [9,10,14,16,21,29,32,33] demonstrated unclear risk of bias, mainly due to insufficient sample sizes, unclear methodologies for addressing missing data, and uncertainties regarding the management of overfitting risks. Furthermore, 1 (4%) study [22] was classified with a high risk of bias as all data were sourced from a single hospital, despite the volume of data, failing to represent a multicenter or stratified analysis. Overall, the bias risk was determined to be unclear for 9 (35%) studies [9-11,14,16,21,29,32,33]. The applicability ratings were moderate for 4 (15%) studies [10,11,21,33], high for 1（4%) study [22], and low for the remaining studies [8,9,12-20,23-32], as detailed in Table 2. For the remaining details, see Table S2 in the Multimedia Appendix 2.

XSL•FO
RenderX

Table 2. Risk of bias and applicability assessment using PROBAST (Prediction Model Risk of Bias Assessment Tool).

Study and year	ROB[a]				Overall bias rating	Overall applicability rating	External validation
	Participants	Predictors	Outcome	Analysis			
Ansbacher et al [8], 2022	Low	Low	Low	Low	Low	Low	Yes
Araújo et al [9], 2024	Unclear	Low	Low	Unclear	Unclear	Low	No
Chen et al [10], 2022	Low	Low	Low	Unclear	Unclear	Unclear	No
Chen et al [11], 2023	Unclear	Low	Low	Low	Unclear	Unclear	No
Garrido-Giménez et al [12], 2023	Low	Low	Low	Low	Low	Low	No
Jhee et al [13], 2019	Low	Low	Low	Low	Low	Low	No
Kaya et al [14], 2024	Low	Low	Low	Unclear	Unclear	Low	No
Kovacheva et al [15], 2023	Low	Low	Low	Low	Low	Low	No
Li et al [16], 2021	Unclear	Low	Low	Unclear	Unclear	Low	No
Li et al [17], 2024	Low	Low	Low	Low	Low	Low	No
Lv et al [18], 2025	Low	Low	Low	Low	Low	Low	No
Marić et al [19], 2020	Low	Low	Low	Low	Low	Low	No
Melinte-Popescu et al [20], 2023	Low	Low	Low	Low	Low	Low	No
Munchel et al [21], 2020	Low	Unclear	Low	Unclear	Unclear	Unclear	Yes
Roque et al [22], 2024	Low	Low	Low	High	Low	High	No
Sandström et al [23], 2019	Low	Low	Low	Low	Low	Low	No
Sufriyana et al [24], 2020	Low	Low	Low	Low	Low	Low	Yes
Tiruneh et al [25], 2024	Low	Low	Low	Low	Low	Low	No
Torres et al [26], 2024	Low	Low	Low	Low	Low	Low	No
Wang et al [27], 2022	Low	Low	Low	Low	Low	Low	No
Wang et al [28], 2024	Low	Low	Low	Low	Low	Low	Yes
Xue et al [29], 2023	Low	Low	Low	Unclear	Unclear	Low	No
Yu et al [30], 2024	Low	Low	Low	Low	Low	Low	Yes
Zheng et al [31], 2021	Low	Low	Low	Low	Low	Low	No
Zhou et al [32], 2024	Low	Low	Low	Unclear	Unclear	Low	Yes
Zhou et al [33], 2023	Low	Low	Low	Unclear	Unclear	Unclear	No

[a]ROB: risk of bias.

The Performance of ML Models in Preeclampsia Prediction

A total of 26 (31 models) studies [8-33] were included. While the pooled estimates demonstrated high average discriminative potential of ML models, substantial between-study heterogeneity was observed, indicating significant context-dependency of model performance. The overall pooled AUROC was 0.91 (95% CI 0.87-0.92; Figure 2). However, its 95% PI ranged from 0.75 to 1.00, suggesting that AUC might decrease to 0.75 in some external validation settings. The pooled sensitivity was 0.81 (95% CI 0.70-0.83; P<.001; I^2=99.6%) In the Figure 3 [8-33], the first author of each study is listed along the Y-axis, the circles represent the point estimates of sensitivity for each model, with the size of the circles being proportional to the weight of the study; the horizontal lines indicate their 95% CIs. The letter Q represents the intersection point of the SROC curve with the inverse diagonal line where "Sensitivity = Specificity." The diamonds represent the aggregated sensitivity estimates of the models, with their width corresponding to the 95% CI of the aggregated values. The vertical red dashed line represents the 95% CI of the pooled sensitivity. However, this only

XSL•FO
RenderX

represents an average level; the wide 95% PI of 0.32-0.96] reveals potential clinical risks. In certain specific studies or future applications, the sensitivity may be as low as 32%, indicating a substantial risk of missed diagnoses. Similarly, although the pooled specificity was 0.88 (95% CI 0.84-0.94; P<.001; I^2=99.7%; Figure 4 [8-33]), its PI across different contexts was 0.49-0.99, demonstrating a similar lack of consistency in specificity. The other summary metrics were as follows: DOR was 37.67 (95% CI 23.46-60.48); PLR was 8.52 (95% CI 6.43-11.29); NLR was 0.24 (95% CI 0.18-0.34). Additionally, we calculated the Spearman correlation coefficient between the log of sensitivity and the log of (1-specificity), which yielded a result of 0.254 (P=.17), indicating no significant threshold effect in the included studies. This suggests that the observed high heterogeneity (as well as the broad PIs mentioned above) primarily stems from nonthreshold factors (such as differences in predictor selection or population characteristics), rather than merely from variations in cutoff value selection.

Figure 2. Summary Receiver Operating Characteristic (SROC) plot illustrating the dispersion of study results. AUC: area under the curve; SROC: Summary Receiver Operating Characteristic.

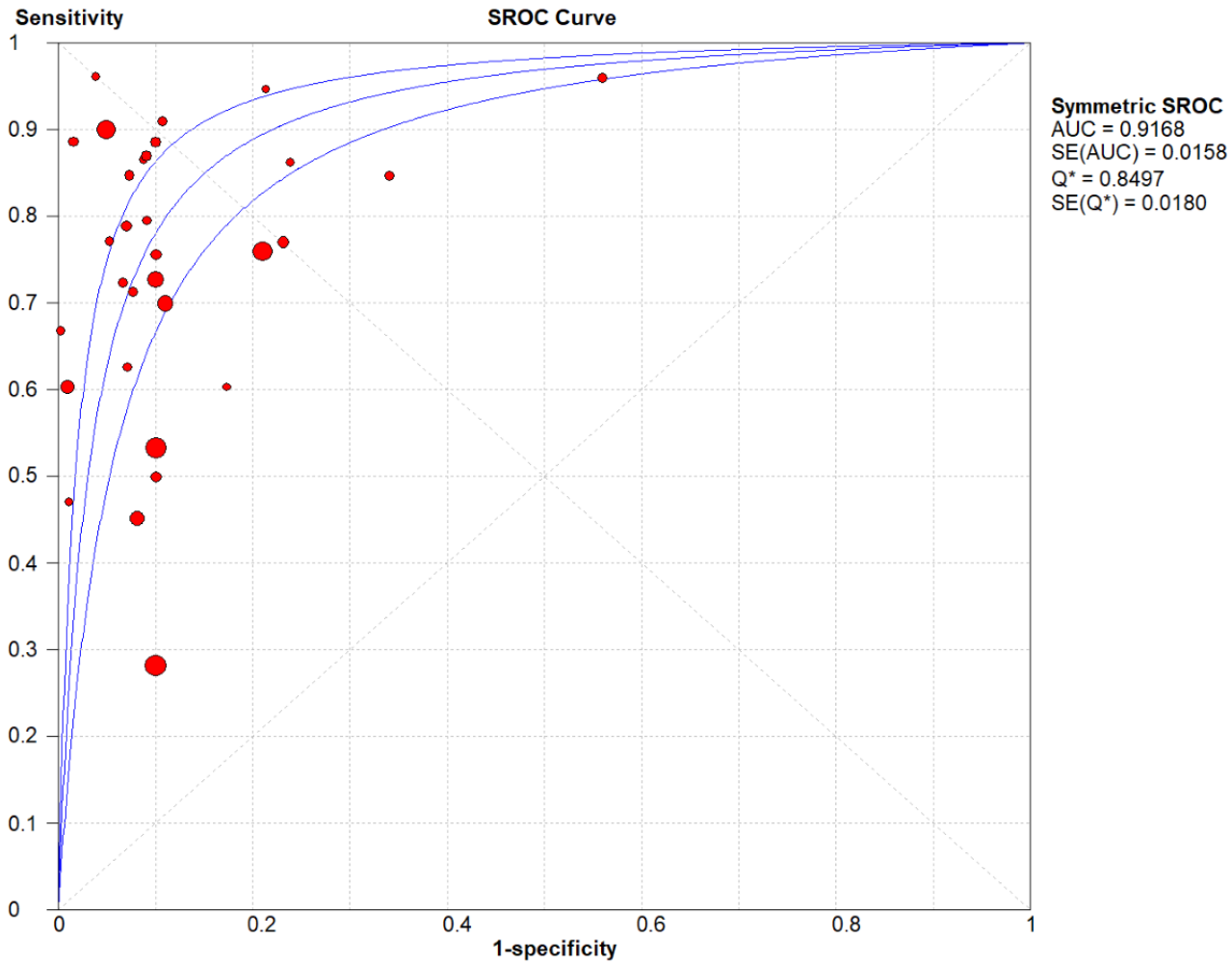

XSL•FO
RenderX

Figure 3. Overall sensitivity of machine learning models for the prediction of preeclampsia [13-17, 19-39].

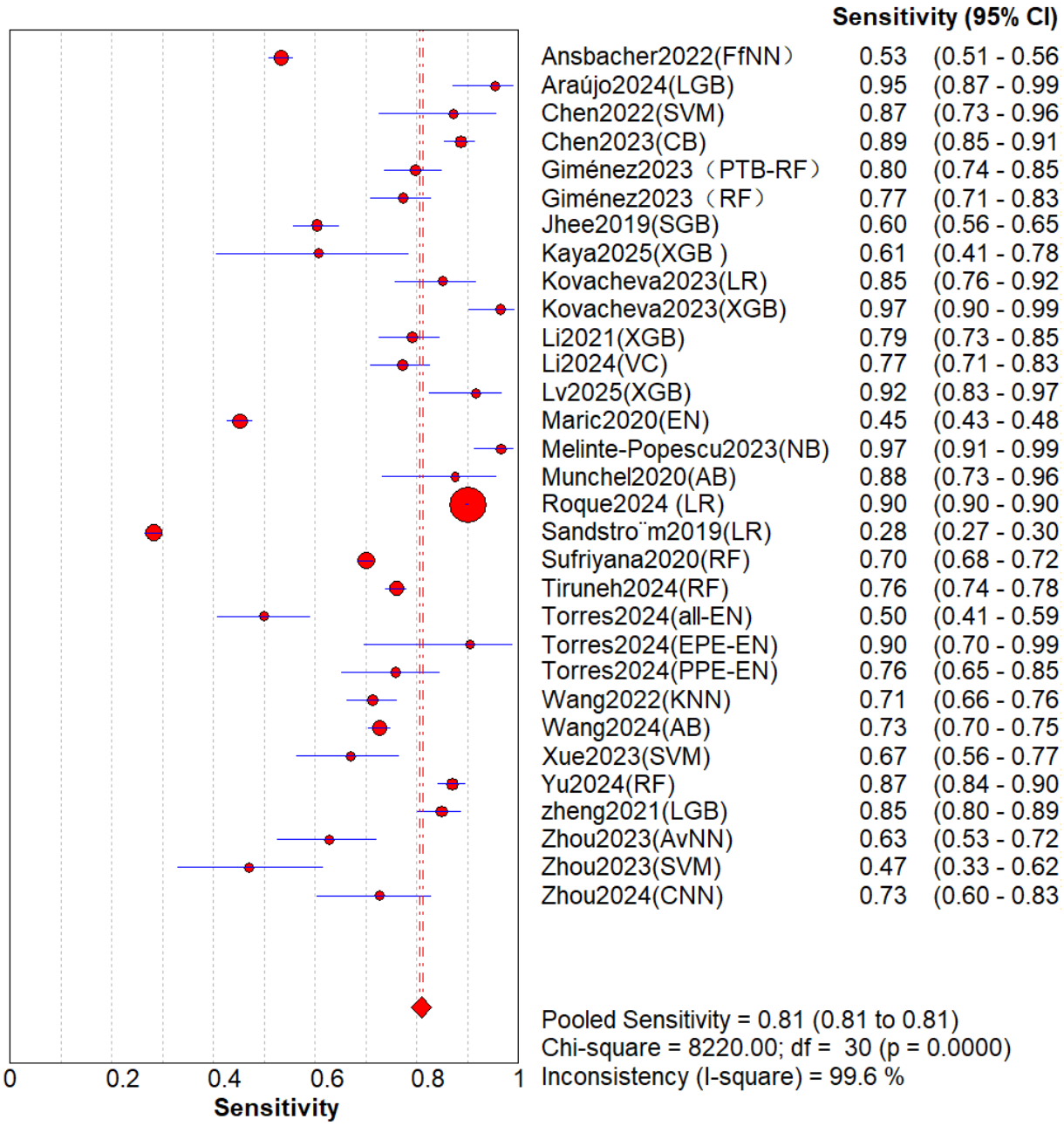

XSL•FO
RenderX

Figure 4. Overall summary specificity of machine learning for the prediction of preeclampsia. [13-17, 19-39].

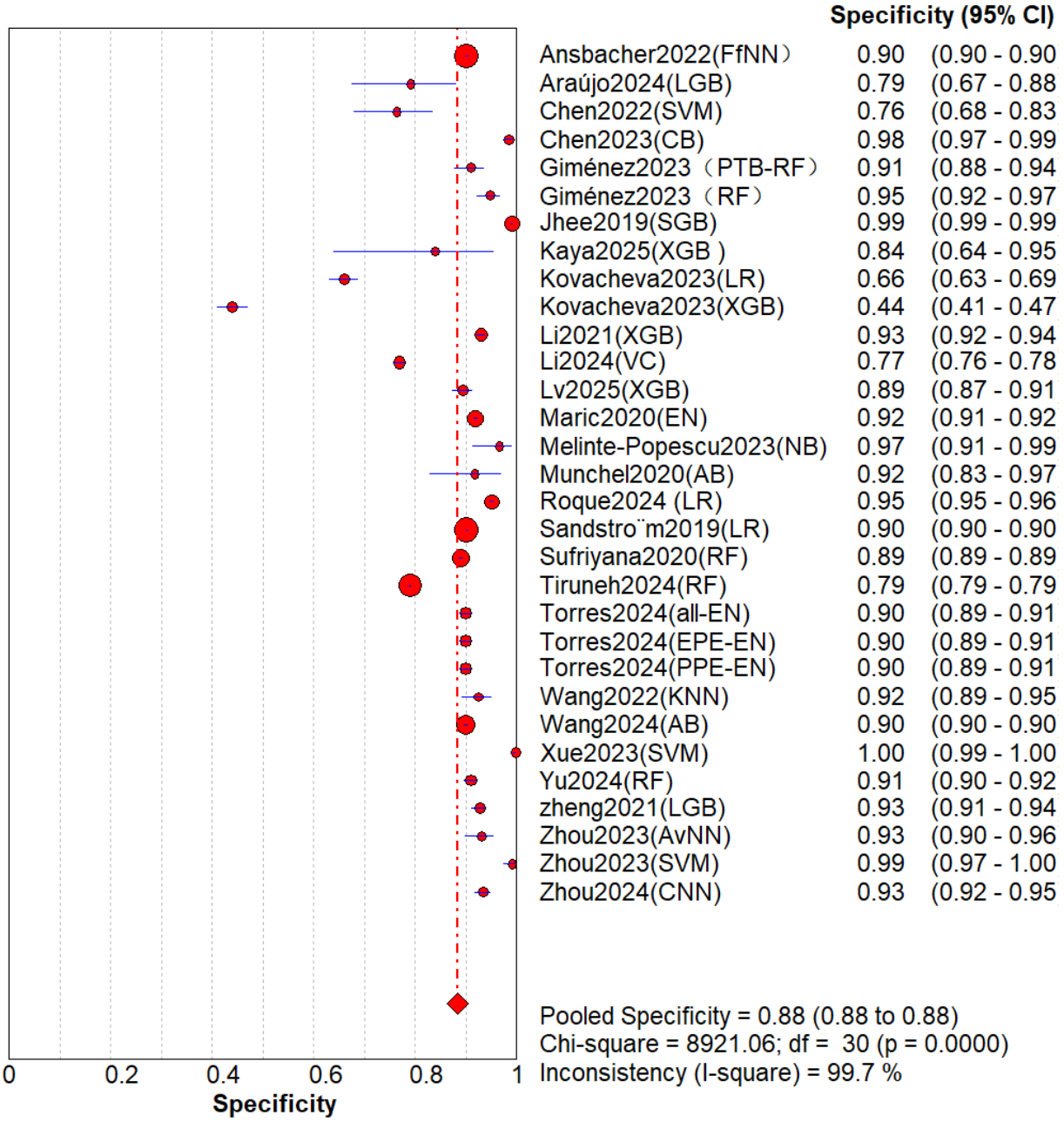

Performance Analysis of External Validation Models

A total of 6 (comprising 7 models) studies [8,17,24,28,30,32] underwent external validation. The analysis revealed that when applied to independent external populations, the models exhibited performance decline with persistent high heterogeneity. Specifically, the pooled AUC was 0.91 (95% CI 0.85-0.95; Figure 5). However, its 95% PI was 0.76-1.00, indicating that the model's discriminative ability might be suboptimal in certain external settings. The pooled sensitivity significantly decreased to 0.68 (95% CI 0.54-0.83; P<.001; I^2=99.6%; Figure 6 [8,21,24,28,30,32]), with a 95% PI of 0.25-0.94. The lower limit of 0.25 indicates that in the worst-case external validation scenario, the model may miss 75% (23/31) of patients, posing an extremely high risk of missed diagnosis. The pooled specificity was 0.90 (95% CI 0.86-0.96; P<.001; I^2=99.7%; Figure 7 [8,21,24,28,30,32]), with a 95% PI of 0.62-0.99. Other indicators included: DOR of 28.21 (95% CI 18.10-43.98; I^2=97.6%); PLR of 7.51; NLR of 0.32. The decrease in sensitivity (from 0.81 in the primary analysis to 0.68) and the extremely low limit of the PI (0.25) strongly confirmed the limited transportability of the model across populations, indicating that direct clinical application requires extreme caution.

XSL•FO
RenderX

Figure 5. Summary Receiver Operating Characteristic (SROC) plot for external validation models. AUC: area under the curve; SROC: Summary Receiver Operating Characteristic.

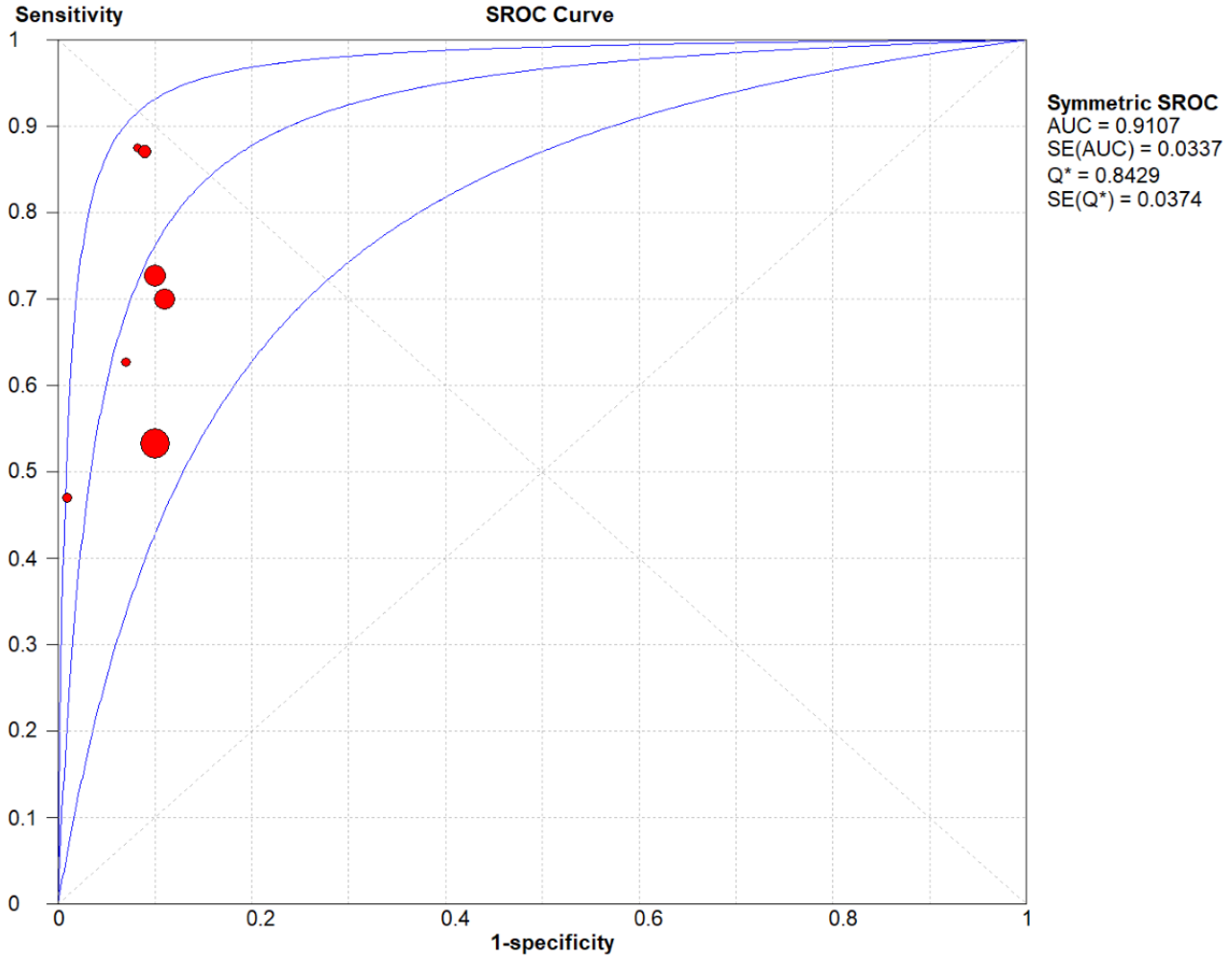

Figure 6. Summary sensitivity of machine learning models for predicting preeclampsia based on external validation [13,27,30,34,36,38].

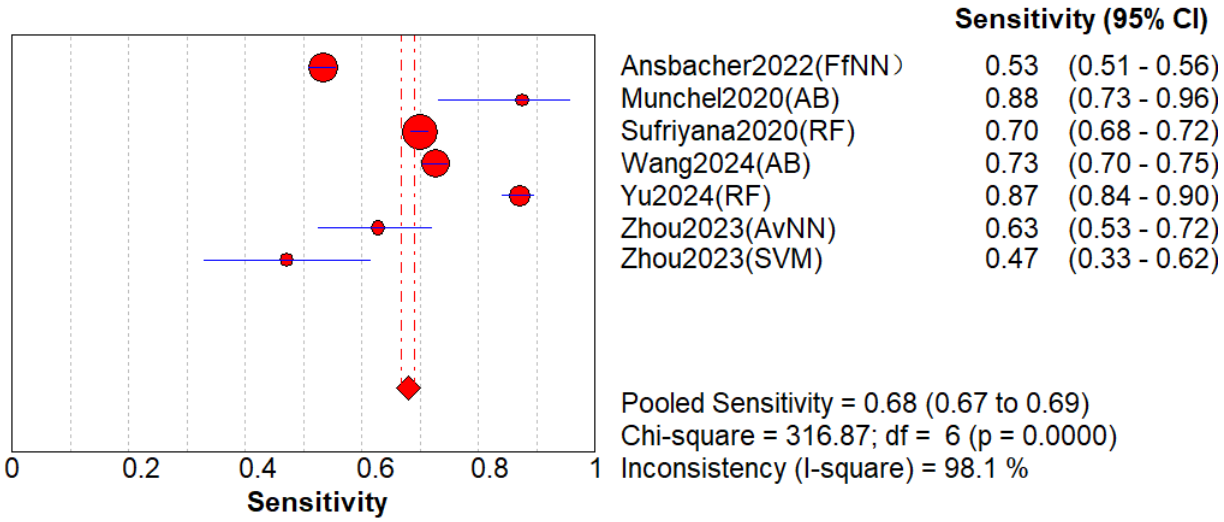

XSL•FO
RenderX

Figure 7. Overall summary specificity of machine learning for predicting preeclampsia [13,27,30,34,36,38].

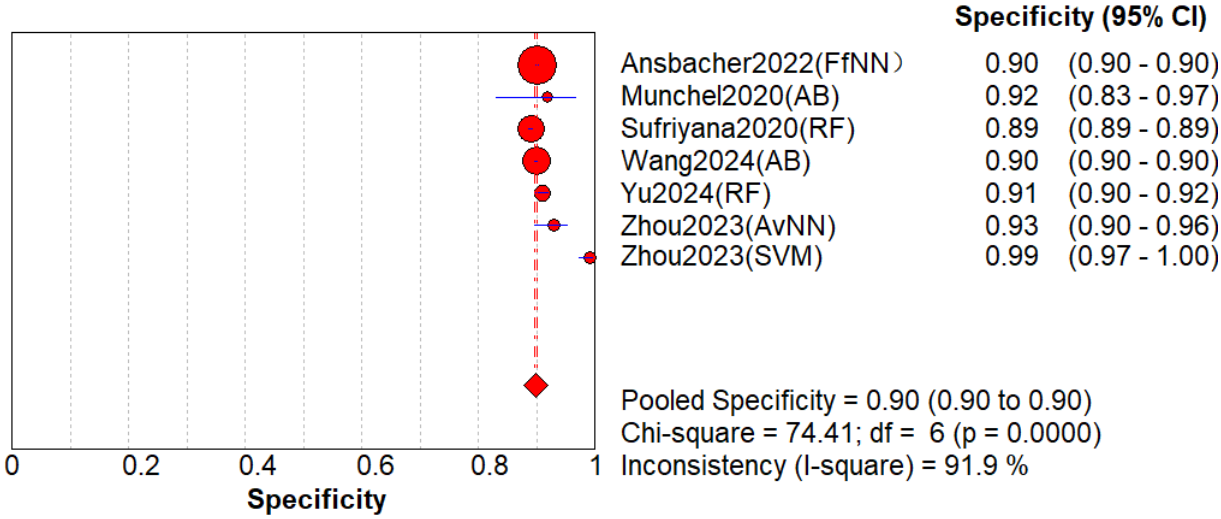

Sensitivity Analysis

After conducting a sensitivity analysis excluding case-control studies in a leave-one-domain-out with 4 (15%) models, the overall summary AUROC is 0.9109 (95% CI 0.8642-0.9390). The summary sensitivity estimate derived from the random-effects meta-analysis is 0.81 (95% CI 0.70-0.83; P<.001; I^2=99.7%), and the summary specificity is 0.88 (95% CI 0.84-0.94; P<.001; I^2=99.7%), as detailed in Figure 8 [8-33]. Consequently, it was concluded that the pooled estimates remained unaffected by the exclusion of outlier values. With an AUC>0.8, the model demonstrated good discriminative ability, but an I^2>75% indicated substantial heterogeneity within most subgroups. To address this issue and gain deeper insights, we undertook a subgroup analysis to investigate the potential sources of this heterogeneity across the studies that were included in our review. Accordingly, we do not interpret a single pooled estimate as "average clinical performance" and instead prioritize subgroup results. In addition, to eliminate the impact of multiple models (derived from the same population) within a single study on statistical independence (unit-of-analysis error), we conducted additional sensitivity analyses by retaining only the model with the highest AUROC from each study (N=26). The results showed that the pooled sensitivity after deduplication was 0.81 (95% CI 0.73-0.87), specificity was 0.88 (95% CI 0.83-0.91), and AUROC was 0.90 (95% CI 0.87-0.93). The above results were highly consistent with the primary analysis (N=31), with no significant differences observed in the CIs, indicating that incorporating different models from the same study did not lead to inflated results or underestimated variance. Therefore, we retained all models in the primary analysis to demonstrate the performance differences among various predictor combinations.

Figure 8. Forest plots of diagnostic performance [13-17, 19-39].

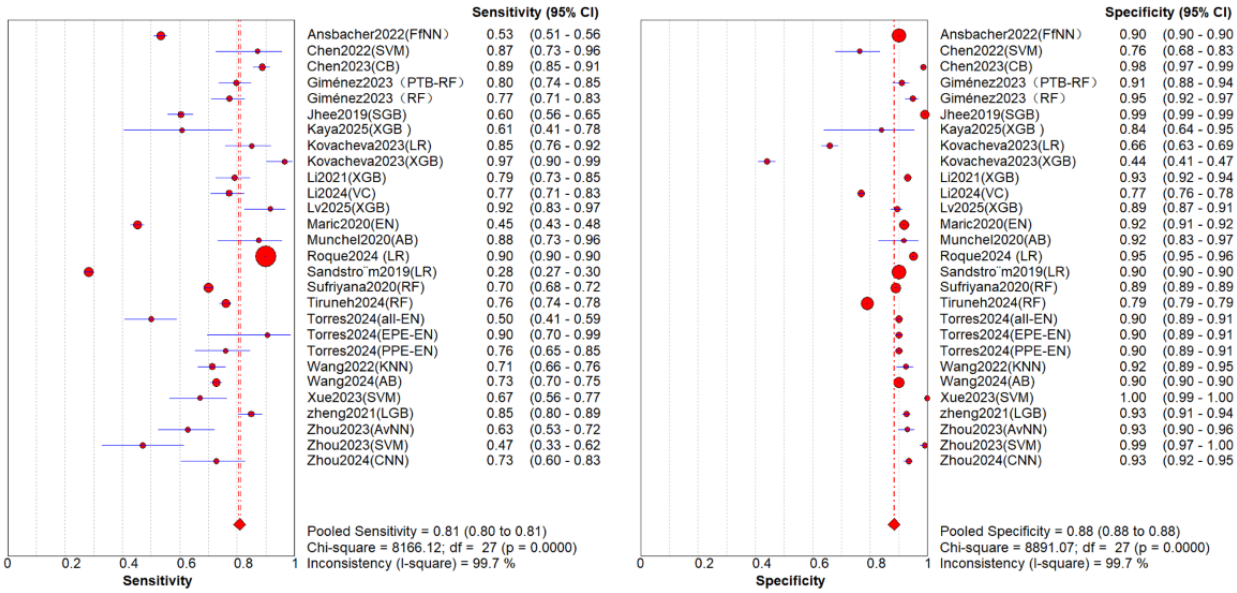

XSL•FO
RenderX

Subgroup Analysis

The comparative results of the subgroup analysis on preeclampsia prediction performance are presented in Table 3; types of ML models, forest plots are shown in Figures S1-S22 in Multimedia Appendix 2. The comparison between subgroups was determined by examining whether the 95% CI of the AUC overlapped. Nonoverlapping intervals indicated statistical significance while overlapping intervals indicated no statistical significance. Data were derived from electronic health records, high-throughput omics, and hybrid sources. Subgroup analysis indicated that models based on hybrid data demonstrated superior performance, followed by those using electronic health records and high-throughput omics. However, considerable heterogeneity was observed, and the 95% CIs extensively overlapped across the 3 data types, suggesting no statistically significant differences among them. The "pregnancy window" refers to the index timing window during which predictors were collected or model discrimination was performed. Models constructed using third-trimester data showed better performance with low heterogeneity. Nonetheless, overlapping 95% CIs across models indicated no statistically significant differences among pregnancy window subgroups. Regarding validation strategies, internally validated models outperformed externally validated ones, albeit with high heterogeneity. Subgroup analysis revealed overlapping 95% CIs between the 2 validation types, implying that the difference was not statistically significant. Regarding sample size, the subgroup analysis results showed that models with smaller sample sizes outperformed those with larger sample sizes, exhibiting lower heterogeneity. However, since the 95% CI overlapped, the differences between sample size subgroups were not statistically significant. Regarding the adopted model, nonlogistic regression prediction models outperformed logistic regression prediction models. Further analysis was conducted on nonlogistic regression models with 3 or more instances in each model category, revealing that neural networks exhibited the best predictive performance with an AUC of 0.9966 (95% CI 0.9772-1.0000) and the lowest heterogeneity. The difference in model performance was statistically significant when compared to elastic net models, but not statistically significant when compared to other models. Regarding the type of predictive variables, prediction models constructed solely using laboratory test indicators achieved the highest predictive performance with an AUC of 0.9463 (95% CI 0.9097-0.9820) and the lowest heterogeneity. Nevertheless, when compared to models built with alternative indicators, the difference in performance was not statistically significant. For the number of predictor variables used in model building, models with 10 or more variables exhibited higher predictive performance with an AUC of 0.9204 (95% CI 0.8671-0.9737), but the difference was not statistically significant compared to models with fewer than 10 variables.

XSL•FO
RenderX

Table 3. Subgroup analysis results.

Grouping	Number of prediction models (PCS[a])	AUC[b] (95% CI)	I^2 (%)	*P* value
Entire study	31	0.9168 (0.891-0.950)	99.6	<.001
Sample size				
<2000	16	0.9361 (0.9079-0.9643)	90.9	<.001
≥2000	15	0.9109 (0.8501-0.9717)	99.8	<.001
Data source				
Mixed	14	0.9154 (0.8713-0.9595)	99.6	<.001
EHR[c]	12	0.9126 (0.8430-0.982)	99.4	<.001
Omics	4	0.9406 (0.8898-0.9914)	95.3	<.001
Pregnancy window				
Early	10	0.9406 (0.7853-1.0000)	95.2	<.001
Mid	4	0.9304 (0.8965-0.9643)	77.2	.004
Late	3	0.9665 (0.9314-1.0000)	71.4	.03
Specific	14	0.9138 (0.8805-0.9471)	99.1	<.001
Machine learning model				
Logistic regression	3	0.9044 (0.6857-1.0000)	100.0	<.001
Nonlogistic regression	28	0.9171 (0.8871-0.9471)	97.6	<.001
RF[d]	5	0.8917 (0.7950-0.9884)	95.8	<.001
SVM[e]	3	0.9068 (0.7623-1.0000)	88.2	<.001
XGBoost[f]	4	0.9177 (0.8500-0.9854)	89.9	<.001
EN[g]	4	0.9419 (0.9125-0.9713)	93.9	<.001
NN[h]	3	0.9966 (0.9772-1.0000)	84.7	.001
Predictor variable type				
Demographic information	10	0.8754 (0.8315-0.9193)	99.4	<.001
Biological genetic marker	3	0.9300 (0.8375-1.0000)	96.8	<.001
Demographic information and laboratory tests	13	0.9275 (0.8665-0.9885)	98.4	<.001
Laboratory testing	5	0.9463 (0.9097-0.9820)	95.8	<.001
Number of predictor variables				
<10	10	0.9124 (0.8855-0.9393)	86.6	<.001
≥10	21	0.9196 (0.8665-0.9727)	99.8	<.001

[a]PCS: piece.
[b]AUC: area under the curve.
[c]EHR: electronic health record.
[d]RF: random forest.
[e]SVM: support vector machine.
[f]XGBoost: extreme gradient boosting.
[g]EN: elastic network.
[h]NN: neural network.

Meta-Regression Analysis

Due to the significant heterogeneity observed among the studies, a meta-regression analysis was conducted. The meta-analysis focused on various factors, including sample size, country of publication, type of ML model, year of publication, study design, study quality, and predictors, as detailed in Table 4. Variables were systematically removed based on the magnitude of their *P* values, and separate meta-regression analyses were performed

XSL•FO
RenderX

for each variable. The results indicated that the source of heterogeneity among the studies was primarily associated with the research quality, as illustrated in Table 5.

Table 4. Meta-regression analysis.

Variable	β coefficient (SE)	*P* value	RDOR[a](95% CI)
Constant	3.547 (1.3356)	.01	__[b]
Sample size	1.075 (0.5388)	.06	0.34 (0.11-1.05)
Country	–0.322 (0.4741)	.50	0.72 (0.27-1.94)
ML[c] method	0.588 (0.7387)	.43	1.80 (0.39-8.37)
Year	0.007 (0.4578)	.99	1.01 (0.39-2.61)
Design	–1.435 (0.8047)	.09	0.24 (0.04-1.27)
Quality	0.672 (0.4076)	.11	1.96 (0.84-4.57)
Predictive	0.773 (0.6075)	.22	2.17 (0.61-7.67)
Validation type	–0.318 (0.4797)	.51	0.73 (0.27-1.97)

[a]RDOR: relative diagnostic odds ratio.
[b]Not applicable.
[c]ML: machine learning.

Table 5. Meta-regression analysis after excluding *P* values from largest to smallest.

Variable	β coefficient (SE)	*P* value	RDOR[a] (95% CI)
Constant	2.398 (0.5879)	<.001	__[b]
Quality	0.800 (0.3951)	.05	2.23 (0.99-5.00)

[a]RDOR: relative diagnostic odds ratio.
[b]Not applicable.

Discussion

Principal Findings

This systematic review identified 31 ML models for preeclampsia prediction. Our primary finding highlights a critical paradox. While models demonstrate high average discriminative potential (pooled AUROC 0.91), they exhibit extreme heterogeneity (I^2>99%) and limited transportability. The wide 95% PI for sensitivity (0.32-0.96) warns that a model performing perfectly in development may miss nearly 70% of cases when applied to a new population. This "context dependence" is further confirmed by the performance drop in external validation studies (pooled sensitivity of 0.68), suggesting that current high AUROCs largely reflect internal fit rather than universal clinical effectiveness.

To investigate the sources contributing to this heterogeneity (as well as the wide PIs), our subgroup analysis revealed several key factors. In the subgroup analysis of all 31 models, we observed that their predictive performance was better when the sample size was small (less than 2000 cases), which contradicts the conventional understanding that "larger sample sizes lead to better predictive performance" [42]. The analysis may be significantly influenced by confounding factors, such as study design (eg, case-control studies) and research type—especially considering the very high AUC of the elastic net (AUC=0.963 for Torres et al [26]; AUC=0.96 for Yu et al [30]). Therefore, careful discernment is required, and one should not hastily interpret this as indicating superior predictive performance of models with smaller sample sizes. Regarding predictor types, laboratory test indicators exhibit superior predictive performance, as the core pathological mechanisms of preeclampsia include placental perfusion disorders, endothelial dysfunction, oxidative stress, and inflammatory responses [43]. Laboratory indicators can directly reflect pathological states, while demographic information provides only indirect risk assessments.

Among the ML models analyzed in this study, including RF, SVM, NN, and Elastic-net, the NN model demonstrated the highest predictive performance (AUC=0.99, 95% CI 0.98-1.00), surpassing traditional ML methods, such as LR, RF, and extreme gradient boosting. This analysis may be attributed to the complex etiology of preeclampsia, a pregnancy complication characterized by multiple pathological processes. The intricate, multidimensional interactions inherent in preeclampsia are challenging to capture comprehensively using linear models. In contrast, NN models are well-equipped to model nonlinear relationships and higher-order variable interactions, which more accurately reflect the pathological characteristics of preeclampsia [44]. Compared to traditional methods, NN can automatically extract features and assign weights to input variables without the need for extensive manual variable screening, demonstrating particular advantages in handling high-dimensional data [45]. Moreover, NN models can integrate

XSL•FO
RenderX

multisource heterogeneous data, such as demographic information, laboratory indicators, and biological genetic markers, thereby adapting to the increasingly complex trends in clinical data.

Higher predictive performance is observed when the number of predictors is equal to or greater than 10. This indicates that using a greater number of predictors helps to more comprehensively reflect disease status, significantly enhancing the model's predictive performance. This is especially true for nonlinear algorithms, which are better equipped to capture interaction effects and underlying patterns.

Nonstandardized handling of missing data means that AUC; concordance index and calibration may not be directly comparable across studies; in particular, listwise deletion or simple imputation combined with restricted case-mix and threshold tuning can inflate discrimination and understate uncertainty. We therefore recommend at minimum (1) transparent reporting of missingness (overall and by variable) and the primary imputation strategy; (2) preferential use of multiple imputation or model-based methods, with minimal recalibration (slope and Brier) and decision-curve analysis during external validation; and (3) reporting confusion matrices under fixed thresholds and top-N% triage plus subgroup robustness (GA window; outcome definitions and sites) to enhance interpretability for clinical and digital health use.

Strengths and Limitations

First, regarding methodological rigor and transparency, we strictly adhered to the PRISMA guidelines for reporting, and the research protocol has been preregistered in the international prospective systematic review registry PROSPERO (CRD420251005830). This ensures that the research objectives and methods are predetermined, thereby minimizing reporting bias. Second, concerning the comprehensiveness of the literature search, our search strategy exhibits significant interdisciplinary characteristics. We not only searched mainstream medical databases such as PubMed and CNKI, but also included IEEE Xplore and Web of Science to ensure a comprehensive capture of ML models published in the fields of engineering technology and computer science. This is critical for a topic that bridges clinical medicine and artificial intelligence, avoiding potential omissions of models that might occur if only medical databases were searched. Third, regarding the reliability of data processing, the entire process of literature screening and data extraction in this study was conducted independently by 2 researchers, with any discrepancies resolved through discussion or by involving a third researcher as an adjudicator. This "dual review" process is considered the gold standard for systematic reviews, ensuring the accuracy of data extraction. Fourth, in terms of the professionalism of quality assessment, we used the PROBAST tool, which is currently recommended by international authorities and specifically designed for predictive model research, rather than traditional diagnostic test evaluation tools, such as QUADAS-2 (Whiting and colleagues [46]). PROBAST enables us to thoroughly assess the risk of bias and applicability of the models across 4 key domains, including participants, predictive factors, outcomes, and analysis, which is more in-depth and relevant than previous reviews. Finally, regarding the prudence of analysis, this study recognizes the common pitfall of "performance overestimation" in meta-analyses of predictive models. Therefore, we clearly identified models lacking external validation and conducted an independent meta-analysis of studies that reported external validation. This approach allowed us to more accurately assess the transportability of the models in real-world applications, leading to the conclusion that they are "highly context-dependent," which is a more cautious and clinically realistic interpretation, avoiding overinterpretation of the aggregated AUROC.

Our study has several limitations that should be considered when interpreting the findings. First, and most critically, is the issue of threshold heterogeneity and optimistic bias. As detailed in the "Methods" section, the performance metrics were synthesized from study-specific "optimal thresholds." This precluded the use of threshold-independent summary measures from a bivariate model and means our pooled sensitivity and specificity are likely inflated compared to what would be achieved with a prespecified, clinically relevant cutoff. The wide PIs we report are, in part, a quantification of this inflation risk. Future primary studies should report performance at multiple, clinically justified thresholds to facilitate more meaningful meta-analysis. Second, related to the above, our statistical synthesis approach was necessitated by the data characteristics. The extreme heterogeneity and lack of threshold standardization made the preferred bivariate modeling approach unfeasible. While our use of univariate HKSJ models with PIs is a robust alternative that honestly communicates uncertainty, it does not model the correlation between sensitivity and specificity. Our subgroup and meta-regression analyses help explore sources of heterogeneity, but residual confounding is likely. Third, our search, though comprehensive, may have missed studies in other languages or in nonindexed repositories. Furthermore, we did not formally assess for publication bias using funnel plots or statistical tests, as these methods are less established and interpretable for diagnostic accuracy data with high heterogeneity. Therefore, our results may be influenced by the preferential publication of studies with positive or high-performance results.

Clinical Significance

The methodological choices in this meta-analysis directly inform its central message. The decision to extract data at study-specific "optimal thresholds" inherently captures the optimistic bias prevalent in ML model development. The strikingly wide 95% PI for sensitivity (0.32-0.96), calculated from these potentially inflated estimates, therefore represents a conservative and realistic warning. The true performance in a new setting, after necessary recalibration to a local threshold, could fall to clinically unacceptable levels. This finding powerfully reinforces the principle that external validation is not a mere formality but a fundamental requirement to bridge the gap between algorithmic promise and clinical utility.

Clinical implementation of these models requires a shift from "universal application" to "local adaptation." Given the wide PIs, hospitals should not adopt published models directly. Instead, we recommend a workflow of local validation and recalibration. Future research should prioritize multicenter

XSL•FO
RenderX

external validation over developing new models. Where data sharing is restricted, federated learning offers a promising pathway to train robust models across diverse populations without compromising privacy.

Conclusions

In summary, ML models demonstrate promising potential for predicting preeclampsia, rather than serving as ready-made universal solutions. While pooled analyses indicate high discriminative performance, the substantial heterogeneity (I^2>99%) and wide 95% PIs (sensitivity 0.32-0.96) reveal significant instability in model performance across different clinical contexts. This "context dependency" was further corroborated in external validation analyses. When applied to independent populations, the model not only exhibited decreased aggregate sensitivity but also the lower bound of its PI dropped to 0.25, quantifying the substantial transplantation risk encountered in cross-center applications. Current evidence therefore supports considering ML as a potential screening adjunct, but does not yet justify its use as a universal clinical diagnostic tool. Future research should shift focus from solely pursuing new models with high AUC values to conducting rigorous multicenter external validation and recalibration of existing models, in order to establish their applicable boundaries within real-world clinical pathways.

Acknowledgments

During the preparation of this work, the authors used Gemini (Google) to assist in refining the English language and structure of the manuscript, as well as to generate R code for the statistical analysis (specifically for the Hartung-Knapp-Sidik-Jonkman method and prediction intervals). After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

This work was supported by the grants from Liaoning Provincial Science and Technology Program Joint Initiative (Key Research and Development Program Project) and The general project of the Department of Education in Liaoning Province (JYTMS20230103).

Authors' Contributions

Methodology: LL
Formal analysis: LL
Investigation: LL, QZ, YZ, XC, WZ
Resources: QZ
Supervision: JW
Writing–original draft: LL

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA checklist.
[DOCX File , 18 KB - jmir_v28i1e78714_app1.docx]

Multimedia Appendix 2
Search strategies, baseline characteristics of included studies, and additional forest plots for subgroup analyses.
[DOCX File , 15347 KB - jmir_v28i1e78714_app2.docx]

Multimedia Appendix 3
Checklists from the Prediction Model Risk of Bias Assessment Tool (PROBAST).
[DOCX File , 24 KB - jmir_v28i1e78714_app3.docx]

Multimedia Appendix 4
Workflow of the meta-analysis on machine learning models for preeclampsia prediction.
[PNG File , 593 KB - jmir_v28i1e78714_app4.png]

References

XSL•FO
RenderX

1. Mol BWJ, Roberts CT, Thangaratinam S, Magee LA, de Groot CJM, Hofmeyr GJ. Thangaratinam S, Magee LA, De Groot CJM, Hofmeyr GJ. Pre-eclampsia. Lancet. Lancet 2016;387(10022):999-1011. [doi: 10.1016/S0140-6736(15)00070-7] [Medline: 26342729]
2. Roberts JM. Preeclampsia epidemiology(ies) and pathophysiology(ies). Best Pract Res Clin Obstet Gynaecol 2024;94:102480. [doi: 10.1016/j.bpobgyn.2024.102480] [Medline: 38490067]
3. Liu Y, Li N, Li Z, Zhang L, Li H, Zhang Y, et al. Impact of gestational hypertension and preeclampsia on fetal gender: a large prospective cohort study in China. Pregnancy Hypertens 2019;18:132-136. [doi: 10.1016/j.preghy.2019.09.020] [Medline: 31610399]
4. Sylvain MH, Nyabyenda EC, Uwase M, Komezusenge I, Ndikumana F, Ngaruye I. Prediction of adverse pregnancy outcomes using machine learning techniques: evidence from analysis of electronic medical records data in Rwanda. BMC Med Inform Decis Mak 2025;25(1):76 [FREE Full text] [doi: 10.1186/s12911-025-02921-z] [Medline: 39939998]
5. Ranjbar A, Taeidi E, Mehrnoush V, Roozbeh N, Darsareh F. Machine learning models for predicting pre-eclampsia: a systematic review protocol. BMJ Open 2023;13(9):e074705. [doi: 10.1136/bmjopen-2023-074705] [Medline: 37696628]
6. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009;339:b2700 [FREE Full text] [doi: 10.1136/bmj.b2700] [Medline: 19622552]
7. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
8. Ansbacher-Feldman Z, Syngelaki A, Meiri H, Cirkin R, Nicolaides KH, Louzoun Y. Machine-learning-based prediction of pre-eclampsia using first-trimester maternal characteristics and biomarkers. Ultrasound Obstet Gynecol 2022;60(6):739-745 [FREE Full text] [doi: 10.1002/uog.26105] [Medline: 36454636]
9. Araújo DC, de Macedo AA, Veloso AA, Alpoim PN, Gomes KB, Carvalho M, et al. Complete blood count as a biomarker for preeclampsia with severe features diagnosis: a machine learning approach. BMC Pregnancy Childbirth 2024 Oct 01;24(1):628 [FREE Full text] [doi: 10.1186/s12884-024-06821-4] [Medline: 39354367]
10. Chen H, Wang X, Lu K, Yu C, Su M, Kang L, et al. Maternal Th17/treg cytokines and small extracellular vesicles in plasma as potential biomarkers for preeclampsia. Int J Med Sci 2022;19(11):1672-1679 [FREE Full text] [doi: 10.7150/ijms.71047] [Medline: 36237987]
11. Chen Z, Chen Z, Huang Z, Cao Y, Wang H, Wang P. Construction and evaluation of machine learning-based predictive models for early-onset preeclampsia. Journal of Xuzhou Medical University 2023;43(8):571-576 [FREE Full text] [doi: 10.3969/j.issn.2096-3882.2023.08.005]
12. Garrido-Giménez C, Cruz-Lemini M, Álvarez FV, Nan MN, Carretero F, Fernández-Oliva A, et al. Predictive model for preeclampsia combining sFlt-1, PlGF, NT-proBNP, and uric acid as biomarkers. J Clin Med 2023;12(2):431 [FREE Full text] [doi: 10.3390/jcm12020431] [Medline: 36675361]
13. Jhee JH, Lee S, Park Y, Lee SE, Kim YA, Kang SW, et al. Prediction model development of late-onset preeclampsia using machine learning-based methods. PLoS One 2019 Aug;14(8):e0221202. [doi: 10.1371/journal.pone.0221202] [Medline: 31442238]
14. Kaya Y, Bütün Z, Çelik Ö, Salik EA, Tahta T. Risk assessment for preeclampsia in the preconception period based on maternal clinical history via machine learning methods. J Clin Med 2024;14(1):155 [FREE Full text] [doi: 10.3390/jcm14010155] [Medline: 39797241]
15. Kovacheva VP, Eberhard BW, Cohen RY, Maher M, Saxena R, Gray KJ. Preeclampsia prediction using machine learning and polygenic risk scores from clinical and genetic risk factors in early and late pregnancies. Hypertension 2024 Mar;81(2):264-272. [doi: 10.1161/HYPERTENSIONAHA.123.21053] [Medline: 37901968]
16. Li Y, Shen X, Yang C, Cao Z, Du R, Yu M, et al. Novelelectronic health records applied for prediction of pre-eclampsia: machine-learning algorithms. Pregnancy Hypertens 2021;26:102-109 [FREE Full text] [doi: 10.1016/j.preghy.2021.10.006] [Medline: 34739939]
17. Li T, Xu M, Wang Y, Wang Y, Tang H, Duan H, et al. Prediction model of preeclampsia using machine learning based methods: a population based cohort study in China. Front Endocrinol (Lausanne) 2024;15:1345573 [FREE Full text] [doi: 10.3389/fendo.2024.1345573] [Medline: 38919479]
18. Lv B, Wang G, Pan Y, Yuan G, Wei L. Construction and evaluation of machine learning-based predictive models for early-onset preeclampsia. Pregnancy Hypertens 2025;39:101198. [doi: 10.1016/j.preghy.2025.101198] [Medline: 39889366]
19. Marić I, Tsur A, Aghaeepour N, Montanari A, Stevenson DK, Shaw GM, et al. Early prediction of preeclampsia via machine learning. Am J Obstet Gynecol MFM 2020;2(2):100100. [doi: 10.1016/j.ajogmf.2020.100100] [Medline: 33345966]
20. Melinte-Popescu A, Vasilache I, Socolov D, Melinte-Popescu M. Predictive performance of machine learning-based methods for the prediction of preeclampsia-a prospective study. J Clin Med 2023;12(2):418 [FREE Full text] [doi: 10.3390/jcm12020418] [Medline: 36675347]
21. Munchel S, Rohrback S, Randise-Hinchliff C, Kinnings S, Deshmukh S, Alla N, et al. Circulating transcripts in maternal blood reflect a molecular signature of early-onset preeclampsia. Sci Transl Med 2020;12(550):eaaz0131. [doi: 10.1126/scitranslmed.aaz0131] [Medline: 32611681]

XSL•FO
RenderX

22. Roque ACA, Huamanzana J, Mauricio D. Forecasting risk pregnancies in peru using machine learning. In: IEEE.: IEEE; 2024 Presented at: 10th International Conference on Optimization and Applications (ICOA); Oct 17-18, 2024; Almeria, Spain p. 1-6 URL: https://ieeexplore.ieee.org/document/10753982 [doi: 10.1109/ICOA62581.2024.10753982]
23. Sandström A, Snowden JM, Höijer J, Bottai M, Wikström AK. Clinical risk assessment in early pregnancy for preeclampsia in nulliparous women: a population based cohort study. PLoS One 2019;14(11):e0225716 [FREE Full text] [doi: 10.1371/journal.pone.0225716] [Medline: 31774875]
24. Sufriyana H, Wu Y, Su EC. Artificial intelligence-assisted prediction of preeclampsia: development and external validation of a nationwide health insurance dataset of the BPJS Kesehatan in Indonesia. EBioMedicine 2020;54:102710 [FREE Full text] [doi: 10.1016/j.ebiom.2020.102710] [Medline: 32283530]
25. Tiruneh SA, Rolnik DL, Teede HJ, Enticott J. Prediction of pre-eclampsia with machine learning approaches: leveraging important information from routinely collected data. Int J Med Inform 2024;192:105645 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105645] [Medline: 39393122]
26. Torres-Torres J, Villafan-Bernal JR, Martinez-Portilla RJ, Hidalgo-Carrera JA, Estrada-Gutierrez G, Adalid-Martinez-Cisneros R, et al. Performance of machine-learning approach for prediction of pre-eclampsia in a middle-income country. Ultrasound Obstet Gynecol 2024 Mar;63(3):350-357 [FREE Full text] [doi: 10.1002/uog.27510] [Medline: 37774112]
27. Wang H, Wu Y, Guo Y, et al. Construction of a risk prediction model for pre-eclampsia based on routine clinical laboratory indicators. J Clin Lab 2022;40(10):731-736 [FREE Full text]
28. Wang L, Ma Y, Bi W, Meng C, Liang X, Wu H, et al. An early screening model for preeclampsia: utilizing zero-cost maternal predictors exclusively. Hypertens Res 2024;47(4):1051-1062 [FREE Full text] [doi: 10.1038/s41440-023-01573-8] [Medline: 38326453]
29. Xue Y, Yang N, Gu X, Wang Y, Zhang H, Jia K. Risk Prediction Model of Early-Onset Preeclampsia Based on Risk Factors and Routine Laboratory Indicators. Life (Basel) 2023;13(8) [FREE Full text] [doi: 10.3390/life13081648] [Medline: 37629504]
30. Yu Y, Xu W, Zhang S, Feng S, Feng F, Dai J, et al. Non-invasive prediction of preeclampsia using the maternal plasma cell-free DNA profile and clinical risk factors. Front Med (Lausanne) 2024;11:1254467 [FREE Full text] [doi: 10.3389/fmed.2024.1254467] [Medline: 38695016]
31. Zheng J, Zhu R, Yan Y, Zhou Y, Luo Y. Construction of a prediction model for preeclampsia based on machine learning algorithm. Front Endocrinol (Lausanne) 2022;47(8):802-808 [FREE Full text]
32. Zhou S, Li J, Yang W, Xue P, Yin Y, Wang Y, et al. Noninvasive preeclampsia prediction using plasma cell-free RNA signatures. Am J Obstet Gynecol 2023;229(5):553.e1-553.e16 [FREE Full text] [doi: 10.1016/j.ajog.2023.05.015] [Medline: 37211139]
33. Zhou T, Gu S, Shao F, Li P, Wu Y, Xiong J, et al. Prediction of preeclampsia from retinal fundus images via deep learning in singleton pregnancies: a prospective cohort study. J Hypertens 2024;42(4):701-710 [FREE Full text] [doi: 10.1097/HJH.0000000000003658] [Medline: 38230614]
34. Collins GS, Reitsma JB, Altman DG, Moons KGM, TRIPOD Group. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. The TRIPOD Group. Circulation 2015;131(2):211-219 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.114.014508] [Medline: 25561516]
35. Chen R, Wang SF, Zhou JC, Sun F, Wei WW, Zhan SY. [Introduction of the Prediction model Risk Of Bias ASsessment Tool: a tool to assess risk of bias and applicability of prediction model studies]. Zhonghua Liu Xing Bing Xue Za Zhi 2020;41(5):776-781. [doi: 10.3760/cma.j.cn112338-20190805-00580] [Medline: 32447924]
36. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010 Jan;21(1):128-138 [FREE Full text] [doi: 10.1097/EDE.0b013e3181c30fb2] [Medline: 20010215]
37. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta-analysis: I is not an absolute measure of heterogeneity. Res Synth Methods 2017 Mar;8(1):5-18. [doi: 10.1002/jrsm.1230] [Medline: 28058794]
38. IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open 2016;6(7):e010247 [FREE Full text] [doi: 10.1136/bmjopen-2015-010247] [Medline: 27406637]
39. Plana MN, Arevalo-Rodriguez I, Fernández-García S, Soto J, Fabregate M, Pérez T, et al. Meta-DiSc 2.0: a web application for meta-analysis of diagnostic test accuracy data. BMC Med Res Methodol 2022;22(1):306 [FREE Full text] [doi: 10.1186/s12874-022-01788-2] [Medline: 36443653]
40. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983 Sep;148(3):839-843. [doi: 10.1148/radiology.148.3.6878708] [Medline: 6878708]
41. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25 [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
42. Deo RC. Machine learning in medicine. Circulation 2015;132(20):1920-1930 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.115.001593] [Medline: 26572668]
43. Jung E, Romero R, Yeo L, Gomez-Lopez N, Chaemsaithong P, Jaovisidha A, et al. The etiology of preeclampsia. Am J Obstet Gynecol 2022;226(2S):S844-S866 [FREE Full text] [doi: 10.1016/j.ajog.2021.11.1356] [Medline: 35177222]

XSL•FO
RenderX

44. Can a neural network model using common clinical data predict preeclampsia? The ObG Project. 2024. URL: https://www.obgproject.com/2024/11/15/can-a-neural-network-model-using-common-clinical-data-predict-preeclampsia/ [accessed 2025-04-26]
45. Zhao Z, Liu X, Guan Y, Li C, Wang Z. Exploring the potential of cell-free RNA and Pyramid Scene Parsing Network for early preeclampsia screening. BMC Pregnancy Childbirth 2025 Apr 14;25(1):445 [FREE Full text] [doi: 10.1186/s12884-025-07503-5] [Medline: 40229739]
46. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536 [FREE Full text] [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
AUROC: area under the receiver operating characteristic curve
CNKI: China National Knowledge Infrastructure
DOR: diagnostic odds ratio
HKSJ: Hartung-Knapp-Sidik-Jonkman
LR: logistic regression
ML: machine learning
NLR: negative likelihood ratio
NN: neural network
PI: prediction interval
PICO: Population, Intervention, Comparison, and Outcome
PLR: positive likelihood ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST: Prediction Risk of Bias Assessment Tool
RF: random forest
SVM: support vector machine
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
XGBoost: extreme gradient boosting

Edited by J Sarvestan; submitted 08.Jun.2025; peer-reviewed by T Shi, P Wu; comments to author 01.Sep.2025; revised version received 20.Dec.2025; accepted 20.Dec.2025; published 19.Jan.2026.

Please cite as:
Liu L, Zhu Q, Zong Y, Chen X, Zhang W, Wang J
Machine Learning Prediction Models for Preeclampsia: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e78714
URL: https://www.jmir.org/2026/1/e78714
doi:10.2196/78714
PMID:

©Lu Liu, Qixuan Zhu, Yichi Zong, Xueyuan Chen, Wei Zhang, Jun Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Communication Strategies to Promote Patient Engagement in Telemedicine: Systematic Review

Yangna Hu[1], BA, MA, MEd; Cindy Sing Bik Ngai[1], BA, MPhil, PhD; Rui Jiang[1], BA, MA, DALS

Department of Language Science and Technology, Faculty of Humanities, Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)

Corresponding Author:
Cindy Sing Bik Ngai, BA, MPhil, PhD
Department of Language Science and Technology
Faculty of Humanities
Hong Kong Polytechnic University
PolyU Hung Hom Bay Campus, 8 Hung Lok Road
Hung Hom, Kowloon
Hong Kong, 999077
China (Hong Kong)
Phone: 852 27665111
Email: cindy.sb.ngai@polyu.edu.hk

Abstract

Background: The rapid growth of telemedicine offers convenience, flexibility, and accessibility for patients to have health care services worldwide. To succeed in telemedicine, health care practitioners and telemedicine tools must engage patients through effective communication. However, a research gap exists in understanding the communication strategies used in telemedicine and how they effectively engage patients.

Objective: This study aims to identify communication strategies influencing patient engagement in telemedicine with provider-patient interactions, as well as how included studies evaluate patient engagement through a systematic review.

Methods: We searched the literature comprehensively using 6 databases, Web of Science, PubMed, Scopus, MEDLINE, CINAHL, and Embase, from inception to October 2025. We included empirical, English-language studies that examined communication strategies affecting patient engagement in telemedicine with provider-patient interactions. Studies lacking actual patients or provider-patient interactions in telemedicine were excluded. We used content analysis to identify texts that were related to Theme 1: the communication strategies affecting patient engagement, and Theme 2: evaluation of patient engagement. Coded texts were analyzed to develop subthemes and themes of identified communication strategies. Methods for evaluating patient engagement were summarized. A narrative synthesis was conducted because of heterogeneity across study design and outcomes. We used the Mixed Methods Appraisal Tool to assess the quality of research included in this study.

Results: This study systematically reviewed 34 peer-reviewed articles, revealing 3 overarching themes of effective communication strategies that enhance patient engagement: interpersonal communication strategies, with 6 subthemes (building relationships, supportive attitude, interactive dialogic loop, nonverbal communication, professionalism and accuracy, and tailored communication); team-level communication strategies, with 3 subthemes (training and preparation, teamwork and care coordination, and cultural and linguistic sensitivity); and system-level communication strategies, with 3 subthemes (usefulness of information, ease of use, and data privacy and security). We also found that included studies predominantly used qualitative research methods, such as semistructured interviews and focus groups, to collect patient engagement data.

Conclusions: This review provides an innovative synthesis of communication strategies that promote patient engagement in telemedicine by integrating interpersonal (micro), team (meso), and system-level (macro) perspectives. Unlike previous reviews that focused on single aspects or levels of communication, this study offers a holistic framework that advances theoretical understanding of how multilevel communication strategies collectively shape patient engagement. Practically, the findings offer actionable guidance for health care professionals, telemedicine developers, and policymakers seeking to enhance the quality and sustainability of telemedicine services. In real-world settings, the identified strategies can inform professional training, platform design, and policy development to support patient-centered digital care. This review is the first to systematically bring together communication strategies for patient engagement in telemedicine across all 3 levels. Future research should build on this framework by developing and validating quantitative measures of patient engagement and examining the relationships between communication strategies and telemedicine outcomes.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e85456) doi:10.2196/85456

KEYWORDS

communication strategies; patient engagement; telemedicine; health care services; provider-patient interactions

Introduction

Background

Digitally accessed health care has accelerated globally, prompted not only by the advancement of communication technologies but also by the increasing demand for accessible and efficient care delivery [1,2]. Consequently, the global use of telemedicine services has grown substantially, with an estimated compound annual growth rate of around 24% between 2022 and 2032 [3]. Telemedicine involves the delivery of health care services via the use of ICTs to engage health care providers (HCPs), patients, and caregivers, and improve health care outcomes [4-6]. It offers convenience and flexibility for both patients and providers and reduces medical service costs and patient wait times [7-11]. Furthermore, it significantly contributes to medical resource allocation, improving patient access and helping health care departments in low-resource settings address resource shortages [6,12-14]. A study analyzed telemedicine consultations in a university-based outpatient telemedicine program and found that the average savings per consultation were 278 miles, 245 minutes, and US $156 [15]. Suzuki and colleagues' study [16] used principal component analysis and cluster analysis to identify countries in Asia and Africa with high potential for telemedicine development, such as Algeria, Egypt, Morocco, and Indonesia. It concluded that telemedicine could address the scarcity of medical resources in these countries.

Despite the great potential of telemedicine to enhance health care accessibility, its adoption remains relatively limited [12,17]. Studies reported that although there are over 300,000 mobile health (mHealth) apps, the user adoption of mHealth apps is low [18,19]. In China, statistics show that telemedicine services account for only 2% of total outpatient services, indicating the underuse of telemedicine services [10]. Except for technology-specific barriers [17,20], a significant factor contributing to this issue is the insufficient communication between patients and service providers, especially on telemedicine platforms where patients or users must initially visit to use these services [21]. Rosler [22] argues that intentional communication skills and tactics can overcome potential barriers to patient engagement within telemedicine and increase patients' connection with providers. Similarly, Fernández Coves and colleagues' study [21] revealed that established means of communication were the most prominent facilitators between patients and service providers at the organizational level of telemedicine adoption in primary care settings.

To succeed on telemedicine platforms, HCPs must effectively engage patients by addressing their needs and preferences [23]. Patient engagement refers to the multidimensional experiences that patients engage with their health management, including cognitive (think), emotional (feel), and behavioral (act) subdimensions of enactment [24,25]. Patient engagement is often used interchangeably with patient activation [26], a concept that focuses on the scenario where patients develop an incremental attitude and have cognitive and behavioral participation in their day-to-day health management [25,27,28]. While there are overlaps between these two concepts, patient engagement is seen as a more holistic consideration, which also includes the psychological involvement during patients' health management situations [25]. In telemedicine settings, patient engagement has been reported to be positively related to high levels of patient satisfaction, improved patient-provider relationships, and increased involvement in health care management [29-33]. For example, in a review study focusing on patient engagement in using hypertension telemedicine tools, Khanijahani et al [34] found that patients' engagement levels were associated with blood pressure reduction levels, their performance in follow-up consultations, and their interests in recording and monitoring their health data.

Despite the many benefits of patient engagement in telemedicine, current studies pay scant attention to the communication strategies used on telemedicine platforms and how they effectively engage users [23,35]. Costa and Serra [36] conducted one of the few review studies examining how communication influences patient engagement in telemedicine contexts. They found that effective communication serves as a cornerstone for improving patient adherence to treatment, whereas communication barriers, such as language barriers, can hinder patient participation in their own care. However, their review primarily focused on reviewing the general role of communication rather than identifying specific effective communication strategies, and it was limited to the field of chronic wound management. Understanding communication strategies is crucial for maximizing the potential of telemedicine, as effective communication in telemedicine is an essential prerequisite for its success, which not only fosters initial engagement but also maintains trust and cooperation and ensures the continued participation of telemedicine [37]. Specifically, communication in telemedicine with access to HCPs is argued to have high potential to stimulate patient engagement [38,39], which remains a favorable way to improve health care outcomes in telemedicine [40-42].

Objectives

Given the rapid growth of telemedicine in health care service delivery and the increasing significance of communication strategies for patient engagement in telemedicine systems [23,35,37], this paper aims to identify the communication strategies promoting patient engagement in telemedicine with HCP-patient interactions by conducting a systematic review of the existing telemedicine studies to explore the effective communication strategies discussed. As such, we propose the following research questions (RQs) to guide our study:

RQ1: What communication strategies have been found or hypothesized to contribute to patient engagement on telemedicine platforms with HCP-patient interactions?

XSL•FO
RenderX

RQ2: How has patient engagement in telemedicine been evaluated in the selected literature?

By synthesizing existing research on crucial communication strategies that enhance patient engagement in telemedicine, this review endeavors to provide HCPs, policymakers, telemedicine tool developers, and researchers with insights to inform the development of more effective telehealth strategies and policies.

Methods

Overview

This study was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [43]. We registered this systematic review on PROSPERO (International Prospective Register of Systematic Reviews; CRD420251053245). This study has been revised and updated from the originally registered PROSPERO protocol to incorporate methodological and reporting improvements based on editorial feedback.

Eligibility Criteria

We included studies if (1) they involved telemedicine using ICTs to deliver health care services, (2) they studied telemedicine tools including HCP-patient interactions, (3) they examined communication strategies influencing patient engagement, (4) they involved real patients or clinical populations who actively engaged with telemedicine, (5) they were peer-reviewed empirical studies, (6) they were published in English, and (7) they were available with full texts.

Articles were excluded if they did not include HCP-patient interactions and only included patients' health care management functions or health care education information in the telemedicine tool. We excluded studies that used standardized, virtual, or fictional patients without actual patient use with the telemedicine platform, as well as studies that focused on improving patient involvement and engagement in health care research. During the screening process, we excluded articles that were not empirical studies and were not published in a peer-reviewed journal, such as conference papers, editorial notes, and book chapters.

Search Strategy

We applied the PRISMA-S (Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension; Multimedia Appendix 1) to guide our search strategy [44] and searched Web of Science, PubMed, Scopus, MEDLINE (via EbscoHost), CINAHL (via EbscoHost), and Embase for relevant studies because these databases ensure that researchers can find comprehensive studies in a wide range of disciplines, including medicine, public health, and social sciences [45-49]. Two experienced librarians specializing in health, social science, and humanities provided professional consultation to help refine and enhance our search strategy. We summarized and searched key terms of "telemedicine," "patient engagement," and "HCP-patient interaction" in the title or abstract, or keywords as shown in Textbox 1. The search strategy combined these three concept blocks using Boolean operators (search strategy: Category 1 AND Category 2 AND Category 3). Apart from using three groups of key terms to identify relevant literature, no language or other restrictions were applied to the search, which was completed on October 31, 2025. The full research strategies applied to the 6 databases are summarized in Multimedia Appendix 2.

Textbox 1. Key terms and search strategy for studies on communication strategies influencing patient engagement in telemedicine involving health care provider (HCP)–patient interactions.

Category 1: telemedicine

eHealth OR e-health OR "electronic health" OR e-consultation OR econsultation* OR e-therapy OR mHealth OR "mobile health" OR telecare OR "tele care" OR telecardiology OR teleconsultation* OR teledentistry OR teledermatology OR telediagnosis OR telehealth OR "tele intensive care" OR "tele ICU" OR telemedicine OR telemonitoring OR telenephrology OR teleneurology OR telenursing OR telepathology OR telepharmacy OR telepsychiatry OR teleradiology OR teleradiotherapy OR telerehabilitation* OR tele-referral* OR "tele referral*" OR telesurgery OR teletherapy OR "virtual care" OR "remote care" OR "virtual medicine" OR "remote rehabilitation*" or "virtual rehabilitation*"

Category 2: patient engagement

"patient activation" OR "patient-centeredness" OR "patient engagement" OR "patient involvement" OR "patient participation"

Category 3: HCP-patient interaction

consultation* OR "online consultation*" OR "video consultation*" OR "video visit*" OR "virtual visit*" OR "remote visit*" OR "televisit*" OR "virtual appointment*" OR "remote appointment*" OR "clinician-patient interaction*" OR "clinician-patient communication*" OR "doctor-patient interaction*" OR "doctor-patient communication" OR "provider-patient interaction*" OR "provider-patient communication" OR "patient-provider interaction*" OR "patient-provider communication" OR "healthcare professional-patient communication" OR "healthcare professional-patient interaction*" OR "HCP-patient interaction*" OR "HCP-patient communication"

Selection Process

A total of 3 authors participated in the selection process. After removing the duplicates, the first reviewer (YH) and the second reviewer (RJ) independently screened all titles and abstracts for eligibility. Any discrepancies regarding study eligibility were resolved through discussion with a third reviewer (CSBN), who served as the adjudicator and made the final decision. During the full-text screening phase, the first reviewer (YH) and second reviewer (RJ) independently assessed all studies, and any disagreements were again resolved in consultation with the third reviewer (CSBN).

Data Collection Process

After the selection process, 2 reviewers (YH and RJ) independently extracted data from each included study using a standardized data extraction table [50] developed for this review. The extraction form was piloted on 7 studies to ensure clarity

XSL•FO
RenderX

and consistency. Extracted data included reference, study setting, country, type and number of participants, recruitment and sampling of participants, participant characteristics, enrollment time, telemedicine type, communication strategies influencing patient engagement, and patient engagement measures. Any discrepancies between reviewers were resolved through discussion. The data extraction table is presented in Multimedia Appendix 3 [39,51-83].

Study Outcomes

The primary outcome domains for this review were (1) communication strategies influencing patient engagement in telemedicine, and (2) methods used to evaluate patient engagement. Communication strategies were defined as any provider-, team-, or system-level communicative actions or decisions intended to enhance communicative effectiveness or compensate for communicative barriers [84-86], thereby shaping patients' cognitive, emotional, or behavioral engagement [25] during telemedicine encounters. Patient engagement measure was defined as any qualitative or quantitative approaches used to assess patients' cognitive, emotional, or behavioral engagement in telemedicine. All results that were compatible with these outcome domains were extracted regardless of the time frame of measurement.

The secondary outcomes extracted from each study included reference information, study setting, country, type and number of participants, recruitment and sampling of participants, participant characteristics, enrollment time, and telemedicine type. The extracted information provided contextual information necessary for interpreting outcome variability across studies.

Quality Assessment

The critical appraisal tool, Mixed Methods Appraisal Tool (MMAT), was used to assess the quality of research included in this study [87]. This tool provides a flexible framework for appraising qualitative, quantitative, and mixed methods studies included in a systematic review [87]. The first reviewer (YH) and the second reviewer (RJ) appraised all the included studies in quality assessment independently, and any disagreements were discussed and resolved with the third reviewer (CSBN) [88]. The product of the quality assessment can be found in the Methodological Quality subsection in the Results section.

Synthesis Methods

We conducted a deductive and inductive qualitative content analysis [89-91] to identify and analyze words, phrases, and texts extracted in the critical primary outcome domain, that is, the communication strategies influencing patient engagement. The extracted content was then examined through thematic analysis to develop sub-themes and overarching themes representing different types of communication strategies. Approaches used to assess patient engagement were also summarized.

An initial codebook for coding the primary outcome domains was developed based on 10 included studies, and new codes were added inductively as the analysis progressed. Multiple coding approaches were applied to ensure comprehensive analysis, since multicoding helps to reveal patterns and associations within the data, providing deeper insights [92,93]. The coding was conducted by two researchers, both with backgrounds in health communication and content analysis methodologies. The first coder (YH) and the second coder (RJ) performed 20% of the initial coding independently. The intercoder reliability was calculated using Cohen κ. The resulting κ=0.82 indicated almost perfect agreement [94]. The first coder (YH) then coded the rest of the included articles. Finally, the third coder (CSBN) reviewed a portion (4/34, 11.76%) of studies to further assess coding accuracy and ensure consistency. Any discrepancies were discussed and resolved through consensus.

A meta-analysis was not performed due to substantial methodological and contextual heterogeneity across studies. Meta-analysis requires sufficient homogeneity in study design, population, intervention, and outcome measures to ensure meaningful comparability of effect estimates [95]. Given the wide variation in health care contexts, forms of telemedicine, research methods, participant groups, as well as the limited number of comparable quantitative findings in the included research, a narrative synthesis was conducted instead. Consequently, quantitative effect measures (eg, risk ratios, odds ratios, and mean differences), methods to explore statistical heterogeneity (eg, subgroup analysis and meta-regression), sensitivity analyses, assessment of reporting bias due to missing results, and certainty or confidence assessment were not performed, as this review did not aim to statistically pool outcomes across studies. This synthesis approach emphasized thematic patterns in communication strategies and their reported influence on patient engagement.

Results

Study Selection and Study Characteristics

In total, 1726 articles were retrieved from 6 identified databases: Web of Science (n=269), PubMed (n=240), Scopus (n=663), MEDLINE (n=147), CINAHL (n=52), Embase (n=355). These studies were published between 1998 and 2025. After removing 857 duplicates, 869 studies remained to review titles and abstracts, and 126 studies were identified as potentially relevant documents. After the full-text review, 34 studies [39,51-83] were included in this systematic review (Figure 1). Included studies were published between 2015 and 2025, with 28/34 (82.35%) articles published after 2020, reflecting a growing scholarly focus on communication processes within rapidly evolving telemedicine practices. A list of included studies is provided in Multimedia Appendix 3 [39,51-83], and Table 1 presents primary outcomes of data extraction [50].

XSL•FO
RenderX

Figure 1. Flowchart of the literature search and screening process for studies on communication strategies influencing patient engagement in telemedicine involving health care provider–patient interactions (1998-2025).

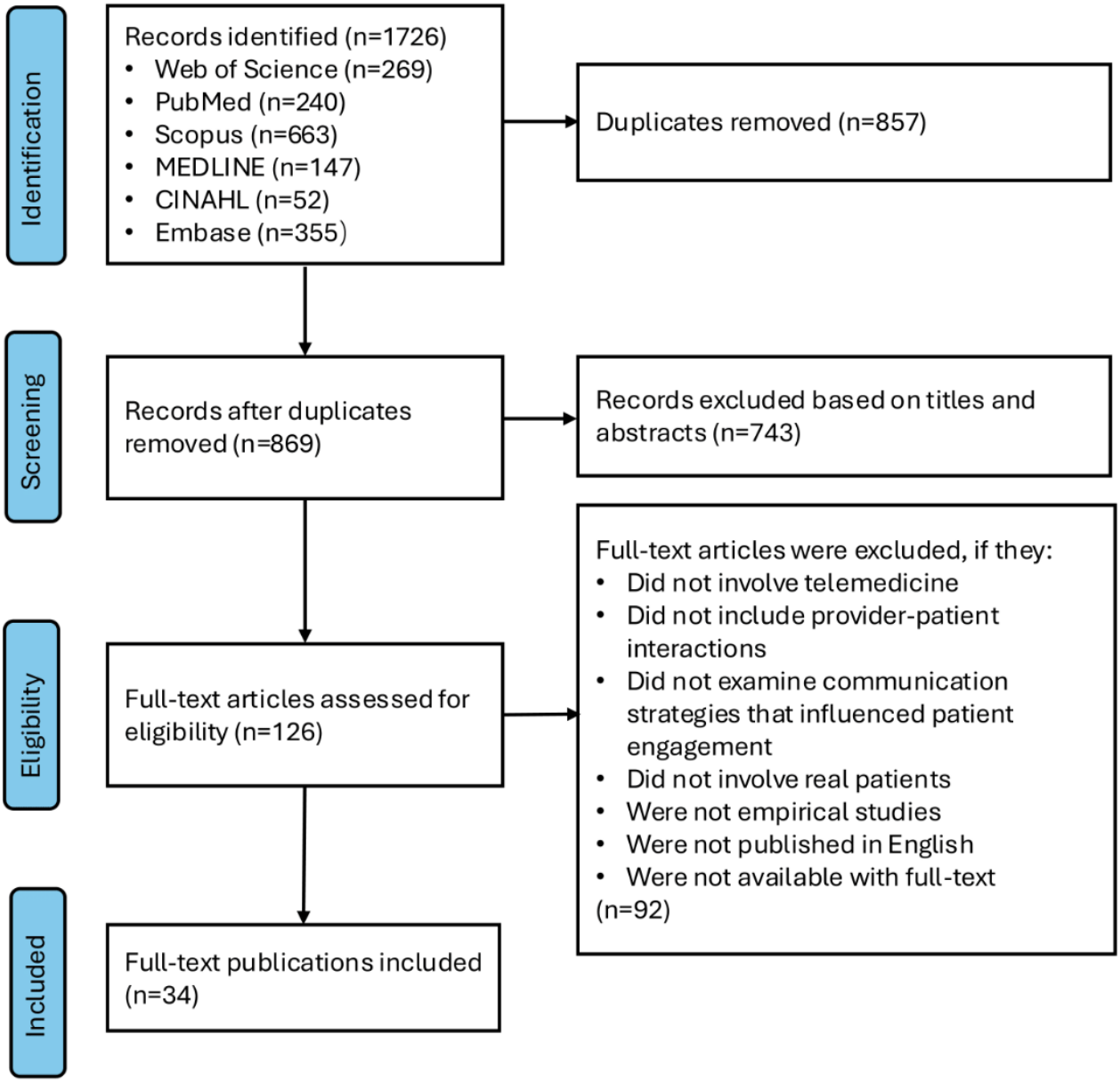

XSL•FO
RenderX

Table 1. Primary outcomes of data extraction on communication strategies influencing patient engagement and patient engagement measures.

Reference, year	Communication strategies influencing patient engagement	Patient engagement measures
Ackerman et al [51], 2020	• Trust-based communication between patients and primary care clinicians; • Using understandable language to provide clear explanations and updates about eConsult decisions.	• No standardized measure; patient engagement was reflected qualitatively via patient accounts of their acceptability of eConsult and feeling involved in care decisions.
Alpert et al [52], 2022	• Using a sincere, empathetic tone and plain language to communicate with patients; • Offering emotional support; • Encouraging patients' participation by using open-ended questions, validating patient input, and fostering dialogue; • Responding promptly to convey accessibility and approachability.	• No standardized measure; patient engagement was reflected qualitatively via clinician accounts of patients' participation, emotional responsiveness, and message interactivity.
Bavngaard et al [53], 2023	• Use of visuality in surroundings, such as showing the medicine bottles, facilitated communication; • Nonverbal communication through gaze direction and smartphone positioning signaled attentiveness and engagement; • Patients' gaze disengagement was interpreted as cognitive engagement in decision-making; • Showing rapport by permitting gaze disengagement from patients.	• No standardized measure; patient engagement was operationalized through the observation and thematic analysis of eight video-recorded consultations, focusing on exploring patients' verbal and non-verbal actions, including attending, contributing, clarifying, and signaling attentiveness.
Björndell [54], 2021	• Listening to patients' thoughts, concerns, and requests; • Guiding and trusting patients in self-examination during video consultations.	• No standardized measure; patient engagement was reflected qualitatively via physicians' accounts of patients' active participation in the consultation process, and patient involvement in decision-making.
Breton et al [55], 2021	• Using visual cues, such as seeing patients' facial expressions, during video visits to enhance communication; • Avoiding the issue of reduced confidentiality of consultations, such as conducting consultations with patients during the patient's grocery time.	• No standardized measure; patient engagement was reflected qualitatively via physicians' perceptions of patients' access, participation, and responsiveness during telemedicine consultations, including comfort, comprehension, follow-up adherence, and involvement in decision-making.
Brodar et al [56], 2022	• Teamwork between departments, including joint virtual visits, interdisciplinary "warm handoffs" between endocrinologists and psychology staff during virtual visits, educating the importance and relevance of consultation and care, and sharing important documents in electronic health records; • Encouraging the provider team to use creative and interactive methods to engage patients, such as playing an online game, using the Zoom Whiteboard feature, and sharing the screen to review materials; • Ensuring staff training in telemedicine.	• No standardized measure; patient engagement was recorded based on the psychosocial screener completion and consultation rates, as well as reflected qualitatively via team members' feedback about patient acceptability of online consults and patient participation.
Caffery et al [57], 2017	• Confusion around issues such as medical liability, privacy, and storage of images was identified as a barrier to patient engagement; • Communication issues, such as a language barrier, between the clinicians and patients hindered engagement.	• No standardized measure; patient engagement was reflected qualitatively via practitioners' perceptions of patient satisfaction, participation in teleconsultations, and continuity of care.
Davoust et al [58], 2025	• Building rapport and trust through open, honest communication; • Visual connections with providers; • Providing tailored communication, such as flexibility in visit modalities to accommodate patient preferences.	• No standardized measure; patient engagement was assessed qualitatively through participants' narratives about their experiences and perceived patient involvement in care.

XSL•FO
RenderX

Reference, year	Communication strategies influencing patient engagement	Patient engagement measures
Day et al [59], 2025	• A consistent, thorough, and mechanistic consultation structure helped engage patients; • Usefulness of information, such as appointment and treatment reminders, increased engagement.	• No standardized measure; patient engagement was reflected qualitatively through semistructured interviews.
Dong et al [60], 2023	• Established clinician-patient relationships influenced engagement.	• No standardized measure; patient engagement was assessed through self-report surveys on patients' engagement with tele-mental health sessions.
Esayed et al [61], 2025	• Building rapport through prior in-person contact and avoiding impersonal communication in telemedicine; • Providing interactive communication through facilitating dialogue and openness.	• No standardized measure; patient engagement was reflected qualitatively through patients' perceptions of their preference for telecare.
Gibson et al [62], 2016	• Interactive communication, such as speaking directly with consultants and getting feedback from them.	• No standardized measure; patient engagement was reflected qualitatively through patient accounts of their involvement in the process of teleconsultations and decision-making.
Grens et al [63], 2022	• Concerns about missing nonverbal cues in video consultations; • Concerns about impersonal telemedicine visits.	• No standardized measure; patient engagement was reflected qualitatively through participant accounts of their involvement in the process of teleconsultations.
Grove et al [64], 2023	• Providing feedback on patient-reported outcomes enhanced patient engagement; • The opportunity to initiate dialogue with providers; • Taking patients seriously and making them feel seen.	• No standardized measure; patient engagement was reflected qualitatively based on patients' perceptions and experiences, such as feelings of support, understanding of their condition, willingness to share information, and overall satisfaction with the remote follow-up.
Higa et al [65], 2021	• Tailoring communication style to meet patient preference; • Building trust-based relationships with patients, such as encouraging patients through text messaging.	• No standardized measure; patient engagement was reflected qualitatively based on multiple data sources, including participants' feedback interview data, answers to open-ended survey questions, the lead researcher's participant observations, and field notes from group meetings, telehealth sessions, and informal interactions with participants, text messages, emails, etc. Engagement outcomes included improvements in diabetes knowledge, frequency of blood glucose monitoring, self-care behaviors, and hemoglobin A1c levels.
Islind et al [66], 2019	• Interactive dialogic loop based on text and links shared via a text chat; • Explaining the reason why health care providers shifted sight and lost eye contact due to screen changes; • Understanding, acknowledging, caring, and trusting patients.	• No standardized measure; patient engagement was reflected qualitatively based on interview data and the researcher's observations.
James et al [67], 2021	• Concerns about missing nonverbal cues during teleconsultations; • Considering patients' multicultural backgrounds and allowing them to bring interpreters to facilitate communication.	• No standardized measure; patient engagement was reflected qualitatively based on nurses' perceptions of patients' need to be seen and respected with cultural sensitivity.
Jensen et al [68], 2023	• Establishing relationships with patients to engage in meaningful conversations.	• No standardized measure; patient engagement was reflected qualitatively through patient accounts of their engagement in care.
Jethwa et al [69], 2022	• Establishing relationships between patients and providers to engage in meaningful conversations; • Having trust and building rapport; • Concerns from patients who do not speak English as a first language; • Ensuring clarity in layman's terms; • Being emphatic when communicating with patients.	• No standardized measure; patient engagement was reflected qualitatively through answers to open-ended questions in a questionnaire, collecting patient preferences for telemedicine.

XSL•FO
RenderX

Reference, year	Communication strategies influencing patient engagement	Patient engagement measures
Jung et al [70], 2023	• Increasing interactions with patients to enhance both patient and staff engagement.	• No standardized measure; patient engagement was primarily observed through participation in daily symptom reporting via mobile/web apps and nurse call follow-ups.
Moore et al [71], 2022	• Prompt responses from providers to show care; • Maintaining established, ongoing patient-provider relationships to foster trust; • Provider knowledge and support regarding portal features; • Useful functions, such as written records to facilitate communication and engagement; • The user-friendly design impacted patients' decisions about how or to what extent they used the portal; • Concerns about the security of the portal.	• No standardized measure; patient engagement was assessed qualitatively through patients' perceptions of their willingness to use telemedicine tools and their preferences for these tools.
Morrison et al [72], 2021	• Ease of use regarding Near Me facilitated continued use of this tool.	• No standardized measure; patient engagement was assessed through feedback collected via surveys, informal verbal feedback during appointments, and participation in improvement cycles, contributing to iterative service refinement.
Olayiwola et al [73], 2018	• Establishing a trust-based provider-patient relationship; • Ensuring responsibilities and roles between clinicians were clearly communicated to patients; • Providing patients with clear explanations of referral processes and allowing communication for clarifications; • Coordination and communication between health care departments; • Cultural-linguistic alignment facilitated acceptance of the electronic consultation and referral; • Potential security and confidentiality concerns hindered engagement.	• No standardized measure; patient engagement was assessed qualitatively through focus groups and survey responses, focusing on perceptions, preferences, and attitudes toward involvement in electronic referral processes.
Osmundsen et al [74], 2015	• Increased knowledge and understanding of patients' disease improved patient engagement.	• No standardized measure; patient engagement was assessed qualitatively through participant questionnaires and interviews, focusing on perceptions of care involvement.
Rodkjær et al [75], 2022	• Using the information patients provide to increase patient engagement and focus on patients' needs.	• No standardized measure; patient engagement was assessed qualitatively through participant questionnaires and interviews, focusing on their perceptions of patient involvement in remote care.
Scruton et al [76], 2025	• Smooth communication between multiple health care providers; • Forming trusting and strong physician-patient relationships; • Giving patients time to process information and ask questions; • Providing emotional support; • Including useful functions or information, such as designing straightforward processes to obtain information, support, and care.	• No standardized measure; patient engagement was assessed qualitatively through patient perceptions of engagement, specifically feeling cared for and their desire to continue virtual options post pandemic.
Selick et al [77], 2023	• Using visual aids and assistive communication tools, choosing appropriate modalities (video over telephone) to support visual and nonverbal cues; • Using nonverbal communication, including body language and facial expressions, to support patient comprehension; • Establishing connections and building trusting relationships with providers.	• No standardized measure; patient engagement was assessed qualitatively based on participant reports of participation, comfort, and involvement during virtual encounters.

XSL•FO
RenderX

Reference, year	Communication strategies influencing patient engagement	Patient engagement measures
Spiess et al [78], 2023	• Concerns about "virtual inhibition," such as missing nonverbal cues and expressing empathy virtually to engage patients from the perspective of providers; • Using artwork to start a conversation and connect with patients meaningfully, helping them feel safe for self-disclosure.	• No standardized measure; patient engagement was primarily assessed through providers' perceptions of patient participation, such as self-disclosure, during virtual visits.
Trondsen et al [79], 2018	• Facilitating immediacy of assessment through real-time visual and verbal interaction; • Building trusting relationships; • Providing a sense of access to the "real" expert (psychiatrist), making patients feel seen and heard and invited to decision-making; • Showing professionalism in clearly justifying and clarifying assumptions and expectations.	• No standardized measure; patient engagement was qualitatively assessed based on participants' perceptions of patient involvement, the sense of being seen and heard, and the feeling of being involved in decision-making during video consultations.
Van Middelaar et al [39], 2018	• Building trusting relationships; • Providing useful information or functions, such as personal reminder and the measurement functionality; • User-friendliness design, such as the clear layout; • Timely and adequate response; • Using a positive and personal tone to motivate the use of telemedicine tools.	• No standardized measure; patient engagement was assessed qualitatively through interview themes addressing initial and sustained use and perceived usability.
Wei and Mao [80], 2023	• Using small talk; • Establishing doctor-patient connections.	• No standardized measure; patient engagement was assessed qualitatively through the analysis of patients' interactional behavior, such as initiation, avoidance, refusal, and topic shifting in the conversation excerpts.
White et al [81], 2024	• Asking questions and encouraging patient participation; • Providing clear explanations and checking for understanding; • Using visual aids; • Interactive communication, such as screen sharing (for video consultations) and sending links or additional resources, during the consultation; • Clarifying information and summarizing key points, engaging patients with health knowledge; • Using small talk to build rapport; • Building trusting relationships.	• No standardized measure; patient engagement was assessed by using multiple research methods, including discourse analysis and conversational analysis to study telehealth consultation recordings, interviewing patients and providers, and conducting patient surveys by asking patients to rate the engagement questions.
Wood et al [82], 2021	• Concerns about diminished rapport from clinicians.	• No standardized measure; patient engagement was primarily assessed qualitatively through participants' perceptions of engagement during telehealth visits.
Zainal et al [83], 2024	• Clear explanation of medical conditions and treatments, and doctors' efficiency was appreciated; • Maintaining eye contact during consultations (valued but not essential in telehealth); • Empathy and respectful communication, and doctors' abilities to address patient concerns patiently and compassionately.	• No standardized measure; patient engagement was assessed qualitatively based on participants' perceptions of patient involvement and participation during consultations.

Methodological Quality

The methodological quality assessment using the MMAT indicated generally high quality across the 34 included studies [39,51-83]. Of these, 22 used qualitative designs and 12 used mixed methods approaches. All studies presented clear research questions or objectives, and the collected data were appropriate for addressing them. Overall, the included studies demonstrated a low risk of bias. A summary of the quality assessment is provided in Table 2.

XSL•FO
RenderX

Table 2. Quality assessment of included studies on communication strategies influencing patient engagement in telemedicine with health care provider–patient interactions using the Mixed Methods Appraisal Tool.

Reference	Year of Publication	All studies		Qualitative studies					Mixed methods				
		S1[a]	S2[b]	1.1[c]	1.2[d]	1.3[e]	1.4[f]	1.5[g]	5.1[h]	5.2[i]	5.3[j]	5.4[k]	5.5[l]
Ackerman et al [51]	2020	✓	✓	✓	✓	✓	✓	✓					
Alpert et al [52]	2022	✓	✓	✓	✓	✓	✓	✓					
Bavngaard et al [53]	2023	✓	✓	✓	✓	✓	✓	✓					
Björndell et al [54]	2021	✓	✓	✓	✓	✓	✓	✓					
Breton et al [55]	2021	✓	✓	✓	✓	✓	✓	✓					
Brodar et al [56]	2022	✓	✓						✓	✓	✓	✓	✓
Caffery et al [57]	2017	✓	✓	✓	✓	✓	✓	✓					
Davoust et al [58]	2025	✓	✓	✓	✓	✓	✓	✓					
Day et al [59]	2025	✓	✓						✓	✓	✓	C[m]	✓
Dong et al [60]	2023	✓	✓						✓	✓	✓	C	✓
Esayed et al [61]	2025	✓	✓	✓	✓	✓	✓	✓					
Gibson et al [62]	2016	✓	✓	✓	✓	✓	✓	✓					
Grens et al [63]	2022	✓	✓						✓	✓	✓	✓	✓
Grove et al [64]	2023	✓	✓	✓	✓	✓	✓	✓					
Higa et al [65]	2021	✓	✓						✓	✓	✓	✓	✓
Islind et al [66]	2025	✓	✓	✓	✓	✓	✓	✓					
James et al [67]	2021	✓	✓	✓	✓	✓	✓	✓					
Jensen et al [68]	2023	✓	✓	✓	✓	✓	✓	✓					
Jethwa et al [69]	2023	✓	✓	✓	✓	✓	✓	✓					
Jung et al [70]	2022	✓	✓						✓	✓	✓	✓	✓
Moore et al [71]	2022	✓	✓	✓	✓	✓	✓	✓					
Morrison et al [72]	2021	✓	✓						✓	✓	✓	✓	✓
Olayiwola et al [73]	2018	✓	✓						✓	✓	✓	✓	✓
Osmundsen et al [74]	2015	✓	✓	✓	✓	✓	✓	✓					
Rodkjær et al [75]	2022	✓	✓						✓	✓	✓	✓	✓
Scruton et al [76]	2023	✓	✓	✓	✓	✓	✓	✓					
Selick et al [77]	2025	✓	✓						✓	✓	✓	✓	✓
Spiess et al [78]	2023	✓	✓	✓	✓	✓	✓	✓					
Trondsen et al [79]	2018	✓	✓	✓	✓	✓	✓	✓					
Van Middelaar et al [39]	2018	✓	✓	✓	✓	✓	✓	✓					
Wei and Mao [80]	2023	✓	✓	✓	✓	✓	✓	✓					
White et al [81]	2024	✓	✓						✓	✓	✓	✓	✓
Wood et al [82]	2021	✓	✓						✓	✓	✓	✓	✓
Zainal et al [83]	2024	✓	✓	✓	✓	✓	✓	✓					

[a]S1: Are there clear research questions?

[b]S2: Do the collected data allow addressing the research questions?

[c]1.1: Is the qualitative approach appropriate to answer the research question?

[d]1.2: Are the qualitative data collection methods adequate to address the research question?

[e]1.3: Are the findings adequately derived from the data?

XSL•FO
RenderX

[f]1.4: Is the interpretation of results sufficiently substantiated by data?

[g]1.5: Is there coherence between qualitative data sources, collection, analysis, and interpretation?

[h]5.1: Is there an adequate rationale for using a mixed methods design to address the research question?

[i]5.2: Are the different components of the study effectively integrated to answer the research question?

[j]5.3: Are the outputs of the integration of qualitative and quantitative components adequately interpreted?

[k]5.4: Are divergences and inconsistencies between quantitative and qualitative results adequately addressed?

[l]5.5: Do the different components of the study adhere to the quality criteria of each tradition of the methods involved?

[m]C: Can't tell.

Results of Syntheses

Based on 34 studies [39,51-83] included in this review, 3 themes of communication strategies were identified as associated with patient engagement: "interpersonal communication strategies," "team-level communication strategies," and "system-level communication strategies." Most studies used qualitative methods, including semistructured interviews and focus groups, to collect information about patient engagement. Studies also used mixed methods to collect patient engagement data, such as combining telemedicine tool use data with patients' qualitative feedback, to understand patient engagement.

Communication Strategies to Promote Patient Engagement in Telemedicine

Based on content analysis of included studies, three synthetic constructs were identified and synthesized; that is, interpersonal communication strategies, team-level communication strategies, and system-level communication strategies, which covered micro-, meso-, and macrolevels of communication strategies to enhance patient engagement in the environment of telemedicine. We developed Figure 2 to illustrate the conceptual framework, with the subsequent content explaining how the 3 levels of communication strategies contribute to patient engagement in telemedicine.

Figure 2. Communication strategies promoting patient engagement in telemedicine, identified in studies involving health care provider–patient interactions across various clinical contexts (2015-2025).

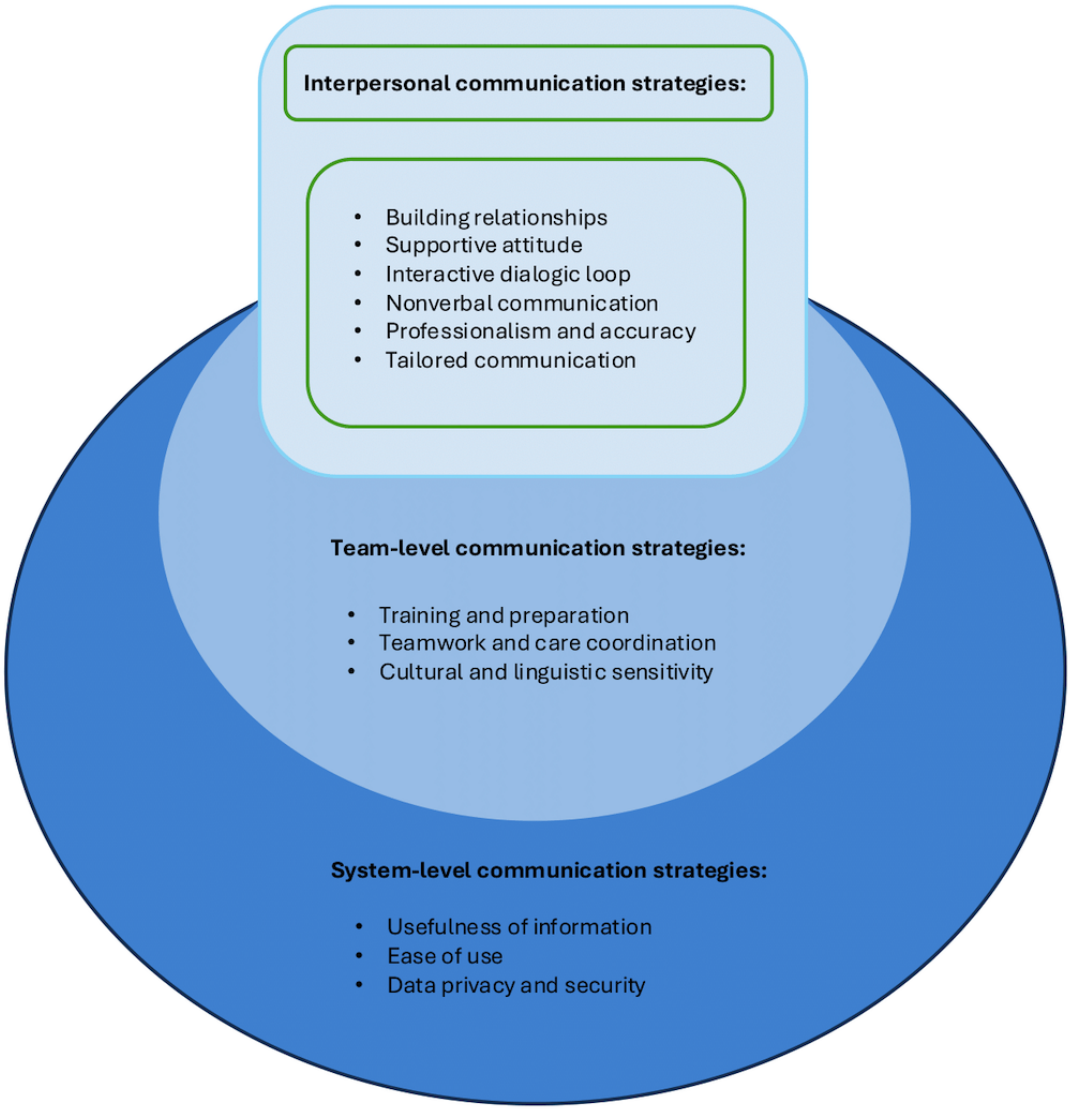

XSL•FO
RenderX

Microlevel: Interpersonal Communication Strategies

At the microlevel, included studies presented prominent interpersonal communication strategies in direct HCP-patient interactions that could improve patient engagement in telemedicine. Specifically, we synthesized 6 subthemes at this level, including building relationships, supportive attitude, interactive dialogic loop, nonverbal communication, professionalism and accuracy, and tailored communication.

The majority of included studies argued that building relationships between HCPs and patients positively impacted patient engagement on diverse telemedicine platforms. Studies found that when patients developed positive and trusting relationships with clinicians, nurses, or other HCPs, they were more likely to accept telemedicine, engage in meaningful conversations with providers, and complete consultation tasks [51,68,71,79,81]. Interpersonal relationship was not only the prerequisite for patients to share their health behaviors, medical concerns, and potential goals [39,78], but also the necessary condition to sustain engagement with telemedicine tools [39]. On the contrary, without established HCP-patient relationships, patients might have concerns about impersonal telemedicine visits [61,63]. Positive interpersonal relationships with HCPs could be built through previous in-person visits [61], visual cues during teleconsultations, such as seeing patients' facial expressions [54,55].

The second subtheme in interpersonal communication strategy is supportive attitude. During teleconsultations, providers were expected to demonstrate a supportive and sincere approach to enhance patient engagement [39,52]. When patients discussed their health behaviors, providers needed to take them seriously, actively listen, understand their concerns, and acknowledge their challenges [54,64,66]. Effective communication to engage patients also involved incorporating "emotional content," such as showing care [71,79], expressing empathy [52,83], and praising patients for positive health behaviors [76]. Such supportive attitudes and actions enabled patients to perceive rapport and genuine support from HCPs [52,79,82], which in turn encouraged greater participation in teleconsultations.

An interactive dialogic loop between HCPs and patients was identified as a crucial component of interpersonal communication strategies that enhanced patient engagement in telemedicine. Direct two-way communication with providers not only strengthened patients' cognitive engagement, such as improving their understanding of disease and increasing access to health knowledge [62,74,81], but also promoted behavioral engagement by encouraging active participation in treatment [52]. During teleconsultations, providers were expected to use a range of communication skills to sustain dialogue and foster engagement [52,61]. These included having small talk [80,81], finding common topics such as artwork to start a conversation and connect with patients [78], explaining the underlying causes of symptoms in detail [52,62], asking open-ended questions [52,53], checking patients' understanding [81], giving patients time to ask questions [76,81], and using chat functions to share screens and links for interactive exchanges [66,81]. In asynchronous communication, prompt and adequate responses to patient messages were essential for stimulating patient engagement, as patients felt reassured by sufficient access to HCPs [39,52]. Conversely, delays or lack of responses often led patients to discontinue platform use [39,64]. Across both synchronous and asynchronous consultations, clear and accessible communication in lay terms was consistently reported to encourage dialogues and strengthen provider-patient interactions [51,52,69].

Nonverbal communication was also found to play a critical role in patient engagement in telemedicine [77]. Studies noted that patients were concerned about the lack of nonverbal cues, such as being able to see what doctors were doing during telephone consultations [67] or missing body language during video consultations [63]. Islind et al [66] and Bavngaard et al [53] further highlighted the role of eye gaze in shaping patient engagement during teleconsultations. Islind et al [66] emphasized that explaining the reason why HCPs shifted their gaze or lost eye contact, often due to screen changes, was important for sustaining engagement. On the other hand, Bavngaard et al [53] underscored the value of allowing flexibility in patients' gaze directionality and even acknowledging momentary gaze disengagement, as brief breaks in eye contact could signal thoughtful and active involvement during consultations. They also highlighted that leveraging visual elements in the surroundings, such as showing the medicine bottles to convey accurate information, could facilitate patients' active participation [53]. Taken together, body language, eye gaze, and the use of visual objects were identified as key nonverbal communication strategies associated with patient engagement.

Within interpersonal communication strategies in telemedicine, patients emphasized the importance of both professionalism and accuracy, as well as tailored communication from HCPs. Zainal et al [83] found that although patients appreciated eye contact during teleconsultations, they placed great value on providers' efficiency and accuracy in communication to avoid errors. Conversely, when providers failed to justify or clearly clarify assumptions and expectations during teleconsultations, patient disengagement was evident [79]. Higa et al [65] highlighted that adapting communication according to patients' individual preferences was crucial for sustaining their engagement. For instance, while some patients responded positively to providers who gave nurturing and encouraging suggestions, others preferred a strict and relentless communication style. Similarly, Davoust et al [58] found that although patients valued a trusting relationship and positive rapport, their levels of comfort varied. Therefore, offering patients flexible options and implementing tailored approaches in telemedicine are essential to accommodate individual preferences and needs.

Mesolevel: Team-Level Communication Strategies

Included studies in this review also presented how communication strategies used by health care teams and organizations could influence patient engagement. A total of 3 subthemes, that is, training and preparation, teamwork and care coordination, and cultural and linguistic sensitivity in health care teams, were synthesized from the mesolevel of communication strategies in telemedicine.

XSL•FO
RenderX

Training and preparation in HCP teams was identified as crucial for patient engagement in telemedicine [56,73]. Patients who experienced difficulty in sustaining attention or "Zoom fatigue" during a remote visit might reduce engagement. To solve this issue, health care organizations should ensure that providers receive communication training in telemedicine, such as using the screen-sharing function to engage patients and playing an online game [56]. Members in provider teams should prepare and provide consistent and clear explanations of the teleconsultation process with patients to have their questions answered, which was reported to impact patients' acceptance of telemedicine tools [73]. Importantly, preparation in HCP teams extended beyond communication training to necessary patient education, particularly around confidentiality. Patients needed guidance on when and how to participate in teleconsultations appropriately, such as avoiding virtual meetings while at the grocery store or driving, so as to maintain privacy and reduce distraction and disengagement [55].

Teamwork and care coordination were identified as essential to influence patients' acceptance and use of telemedicine when they received care from multiple providers. Olayiwola et al [73] reported that clearly defined responsibilities and effective coordination among clinicians were prerequisites for patient acceptance of telemedicine. Similarly, Brodar et al [56] found that teamwork across departments and HCPs, such as joint virtual visits, warm handoff through visit summaries, and sharing key information in electronic health records, helped ensure continuity of care and strengthened patient engagement. Conversely, poor communication among multiple HCPs undermined continuity and reduced care quality, leaving patients feeling neglected and less willing to engage in teleconsultations [76].

For patients from multicultural backgrounds, cultural and linguistic sensitivity within health care teams was crucial to alleviating concerns about using telemedicine [57,67,69,73]. Teams needed to recognize potential cultural and language barriers, particularly when providers interacted with patients who were nonnative English speakers [57,69]. In such cases, involving interpreters during teleconsultations was recommended to help overcome these barriers and support patient engagement [67].

Macrolevel: System-Level Communication Strategies

In addition to identifying communication strategies involving individual HCPs and their teams, this review also examined system-level strategies within telemedicine that influenced patient engagement. A total of 3 key subthemes were identified within this category: usefulness of information, ease of use, and data privacy and security.

Patients reported that the perceived usefulness of information provided by telemedicine platforms, such as self-management tools, personal reminders, access to relevant health information, and a written record function that helped them recall providers' guidance and details from HCP-patient communication, facilitated their engagement [39,71,76]. Ease of use was another critical system-level factor influencing patients' adoption and continued use of telemedicine [71,72]. Platforms with a clear and simple layout and user-friendly features increased acceptability [39,71,72], whereas barriers, such as login difficulties, navigation challenges, or app freezing, discouraged patients from ongoing use and reduced the likelihood of recommending telemedicine tools [71].

Additionally, scholars reported that patients were sometimes hesitant to use telemedicine tools due to concerns about data privacy and security [57,71,73]. Given the sensitive nature of personal health information, some patients expressed worry about how their data were stored and protected [57,71]. Therefore, ensuring secure handling and safeguarding patient information on telemedicine platforms is essential to building trust and encouraging patient engagement.

Evaluation of Patient Engagement

The overwhelming majority of included studies (31/34, 91.18%) used qualitative methods, such as observations, one-on-one interviews, focused groups, asking open-ended questions, and collecting qualitative feedback, to investigate patient engagement from patients and HCPs. Researchers collected qualitative data about patient acceptability of telemedicine, user engagement, patient participation, attention during consultation, and involvement in decision making to evaluate patient engagement. For example, Bavngaard et al [53] conducted a qualitative observational study analyzing 8 video-recorded HCP-patient consultations to explore patient participation during teleconsultations. Van Middelaar et al [39] used semistructured interviews to investigate 20 patients' engagement experience on an online cardiovascular risk management tool. Olayiwola et al [73] collected patient engagement data from both patient focus groups and HCPs' perceptions about patient engagement from their open-ended feedback in an online survey.

Three studies [56,60,81] used mixed methods to evaluate patient engagement. Brodar et al [56] combined quantitative components, that is, health screener completion rate and consultation rate as indicators of engagement, with a qualitative component, that is, participants' feedback through open-ended responses and comments about their telehealth experiences. In Dong and colleagues' [60] telemental health study, patient engagement was measured through quantitative survey items, such as provider-reported ratings of patient engagement, as well as qualitative feedback from providers' open-ended responses describing types of patients that engaged or disengaged in tele-mental health services. White et al [81] used multiple research methods to evaluate patient engagement, including using discourse analysis and conversational analysis to study telehealth consultation recordings, interviewing patients and HCPs, and conducting patient surveys by asking patients to rate the engagement questions, which related to the patient's ability and comfort in communicating and participating in their care from the Telehealth Usability Questionnaire.

Discussion

Principal Findings

The objective of this systematic review was to identify communication strategies that influence patient engagement in telemedicine with the function of HCP-patient interactions. A total of 34 peer-reviewed studies were analyzed, revealing 3

XSL•FO
RenderX

overarching themes of effective communication strategies that enhance patient engagement: interpersonal communication strategies, with 6 subthemes (building relationships, supportive attitude, interactive dialogic loop, nonverbal communication, professionalism and accuracy, and tailored communication); team-level communication strategies, with 3 subthemes (training and preparation, teamwork and care coordination, and cultural and linguistic sensitivity); and system-level communication strategies, with 3 subthemes (usefulness of information, ease of use, and data privacy and security). Furthermore, this review found that qualitative research methods were the most commonly employed approach for assessing patient engagement in the included studies.

Implications Across Micro-, Meso-, and Macrolevel Communication Strategies

At the microlevel, interpersonal communication strategies between HCPs and patients emerged as a cornerstone of enhancing patient engagement in telemedicine. This finding is consistent with previous health care research. For example, Ngai et al [89] highlighted that communication strategies such as maintaining an interactive dialogic loop and demonstrating empathy during two-way HCP-patient communication were crucial for engaging users in health care settings. Similarly, Kwame and Petrucka [96] advanced a patient-centered model, arguing that person-centered communication fosters effective communication and contributes to positive health outcomes. Their model emphasized building meaningful relationships with patients, recognizing their concerns and needs, encouraging self-expression, explaining health conditions and care plans clearly, and engaging in empathetic communication—all of which align with the subthemes of interpersonal communication strategies identified in this review. These insights reinforce the approach of patient-centered communication. Rather than focusing solely on completing consultation tasks, HCPs should view patients as unique individuals with distinct care needs and as collaborators in the care process [65,66,83,96]. Such an approach facilitates effective communication and, ultimately, strengthens patient engagement in telemedicine.

This review identified communication strategies applied not only during synchronous or asynchronous consultations, but also in the form of adequate preparation, particularly at the team level. At the mesolevel, 3 key team-level communication strategies were identified, that is, training and preparation, teamwork and care coordination, and cultural and linguistic sensitivity, which resonate with relational coordination theory [97] and cultural competence model [98]. The relational coordination theory is widely discussed in organizational communication, which emphasizes shared goals, shared knowledge, and mutual respect among team members [97]. This aligns with evidence showing that coordinated teamwork, including team-level communication training in the environment of telemedicine, consistent and clear explanations of the teleconsultation processes, warm handoffs, and joint virtual visits, improved telemedicine acceptance and sustained patient engagement [56,73].

In addition, cultural and linguistic sensitivity emerged as a crucial dimension of team-level communication, consistent with the cultural competence model, which proposes a model of care that includes cultural awareness, knowledge, skills, encounters, and desire [98]. This framework underscores the importance of understanding patients' unique cultural backgrounds and needs, adapting communication styles, addressing language barriers, and involving interpreters where necessary to ensure equitable access and rapport with diverse patient populations [57,67,69,98]. Collectively, these strategies at the team level illustrate that patient engagement in telemedicine is not only an outcome of interpersonal interactions but also the product of well-prepared, well-coordinated, and culturally responsive health care teams.

The identified system-level communication strategies align with previous research on health-related communication on patient engagement. For example, many health communication studies have validated that providing useful content could improve the engagement of the targeted audience [89,99-101]. In addition, Xie and colleagues' [102] and Vasiloglou and colleagues' [103] studies reported that ease of use was a critical reason for users to choose a health app. The identified subthemes of usefulness of information and ease of use at the macrolevel resonate with the technology acceptance model, a leading model in technology acceptance, which argues that users' perceived usefulness and ease of use are primary factors influencing their adoption of new technologies [104].

Moreover, data privacy and security emerged as a critical system-level communication strategy in this review. Given the highly private and sensitive nature of health care data, it is understandable that some patients were reluctant to adopt telemedicine tools due to concerns about confidentiality [105,106]. To address these concerns, telemedicine developers must prioritize robust data protection measures. Suggested strategies include implementing an authentication mechanism [107] and providing patient telehealth "drop-in" kiosks with devices and soundproof space [82].

Advancing the Evaluation of Patient Engagement in Telemedicine

It is surprising to find that the included studies in this review predominantly used qualitative methods, such as semistructured interviews and qualitative feedback, to collect data about patient engagement. Research primarily using quantitative measurements of patient engagement was missing from the included studies. Although 3 studies [56,60,81] used surveys to collect participants' ratings of patient engagement-related items, none of the included studies measured patient engagement in the sense of quantifying engagement through standardized scales. In other words, the quantitative assessment tools for evaluating patient engagement were not unified and standardized. This might be due to a significant lack of clarity regarding the definition and conceptualization of patient engagement, as evidenced by the plethora of terms frequently used interchangeably in this field, as well as the lack of assessment instruments [25].

Not identified in this review, but in a worldwide context, the Patient Activation Measurement (PAM) scale [27] is one of the few assessment scales that have been used to evaluate patient engagement in telemedicine [40,108-110]. The PAM scale was

XSL•FO
RenderX

developed to quantify patients' knowledge, skills, and confidence in managing their health [27,111]. However, although the concepts of patient engagement and patient activation overlap, they differ in their conceptual breadth [25]. As discussed earlier, patient engagement represents a multidimensional psychosocial process in which individuals' cognitive, emotional, and behavioral actions collectively shape how they manage their health. In contrast, patient activation primarily emphasizes the cognitive and behavioral components of this process [25,31]. As such, the PAM scale could not capture the holistic nature of patient engagement. Another widely accepted patient engagement scale is the 5-item Patient Health Engagement (PHE) scale developed by Graffigna and colleagues [25]. The PHE scale assesses patients' perceived readiness for cognitive, emotional, and behavioral engagement. However, none of the studies included in this review used this instrument. In addition, patient engagement has been measured in previous research using other standardized tools, such as the observing patient involvement in decision making (OPTION) scale for measuring patient involvement [112], the Perceived Involvement in Care Scale [113], and the Patient Participation Scale [114], none of which were applied in the included studies. Nevertheless, these existing instruments hold potential for integration or adaptation to enable more consistent evaluation of engagement outcomes in future telemedicine research. We summarized available standardized tools for assessing patient engagement and their potential adaptations to telemedicine in Multimedia Appendix 4.

Limitations and Future Directions

This review has some limitations to note: First, it only included telemedicine studies with HCP-patient interactions. Although telemedicine tools with interactive support from providers have great potential to engage patients [38,39], other studies on telemedicine platforms that focus on patient education, health data management, or the dissemination of health-related information may also incorporate additional effective communication strategies that enhance patient engagement, which can be explored in future reviews. Second, the review did not include gray literature, which may have led to the omission of recent developments or emerging trends in the field first reported at conferences. Incorporating conference proceedings in future review could provide a more comprehensive and up-to-date understanding of the field. Third, this review only included peer-reviewed articles published in English, which may have excluded important research published in other languages that explored telemedicine in various contexts. Despite these limitations, this review serves as a foundational step in the field. It is hoped that future research will address these deficits by exploring the topic more comprehensively.

Future research can explore the following directions in studying effective communication strategies for promoting patient engagement with telemedicine tools. First, researchers should further clarify what patient engagement is by providing a rigorous conceptualization and exploring the dimensions of patient engagement, particularly in the telemedicine environment. Currently, studies have tested and collected data on usability, patient acceptability, patient participation, health condition management, and so on, to understand patient engagement. However, what the components of patient engagement are and how to measure them scientifically remain unclear. In addition to using explorative qualitative methods to ask questions about patients' attitudes and preferences toward telemedicine tools, validated assessment instruments for patient engagement in this field are expected to be developed. Second, future studies should examine and validate the relationships between 12 subthemes across the 3 overarching communication strategy themes identified in this review and patient engagement. Such efforts could contribute to the development of an integrated communication framework that fosters patient engagement with telemedicine tools. In particular, future studies may explore and empirically test the connections between specific communication subthemes and different dimensions of patient engagement. Third, future work can build on this study by exploring additional telemedicine contexts beyond HCP-patient interactions, integrating grey literature and conference proceedings, and including non-English publications to capture more comprehensive evidence, emerging trends, and broader cultural perspectives on communication strategies influencing patient engagement.

Conclusion

This systematic review underscores the critical role of various communication strategies in enhancing patient engagement in telemedicine with HCP-patient interactions. A total of 3 themes of communication strategies, namely interpersonal (micro), team (meso), and system (macro) level communication strategies, with 12 subthemes, were identified as important factors influencing patient engagement. This review offers an innovative and pioneering effort to systematically synthesize communication strategies that promote patient engagement in telemedicine. Unlike previous reviews that focused on isolated aspects or levels of communication, our review uniquely integrates strategies across all three levels to provide a holistic and comprehensive framework. Theoretically, it advances understanding of how micro-, meso-, and macrolevel communication strategies collectively influence patient engagement, filling a critical gap in existing literature. Practically, it provides actionable guidance for telemedicine developers, health care professionals, and policymakers. The identified strategies offer a comprehensive framework for improving the quality and sustainability of telemedicine practices. In real-world terms, these insights can inform training programs for health care professionals, guide platform design, and support policy initiatives that promote equitable, patient-centered digital care. We also found that the majority of included studies used qualitative research methods to assess patient engagement. Future studies can further explore, validate, and test quantitative methods to evaluate patient engagement and the relationships between different communication strategies and patient engagement in telemedicine.

XSL•FO
RenderX

Acknowledgments

This research was funded by Hong Kong Polytechnic University for the article processing fee (APF) payment support. The authors appreciate librarians (Queennie Ip and Emily Wu) for their professional assistance in developing the search strategies for this systematic review.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary files.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA 2020_S checklist.
[DOCX File , 58 KB - jmir_v28i1e85456_app1.docx]

Multimedia Appendix 2
Databases and search strategies for studies on communication strategies influencing patient engagement in telemedicine involving healthcare provider-patient interactions.
[DOCX File , 17 KB - jmir_v28i1e85456_app2.docx]

Multimedia Appendix 3
Data extraction table for included studies on communication strategies influencing patient engagement in telemedicine involving healthcare provider-patient interactions.
[XLSX File (Microsoft Excel File), 31 KB - jmir_v28i1e85456_app3.xlsx]

Multimedia Appendix 4
Summary of standardized instruments for measuring patient engagement and their relevance and potential to be adapted in telemedicine research.
[DOCX File , 21 KB - jmir_v28i1e85456_app4.docx]

References

1. Greenhalgh T, Wherton J, Shaw S, Morrison C. Video consultations for COVID-19. BMJ 2020;368:m998. [doi: 10.1136/bmj.m998] [Medline: 32165352]
2. Yeung AWK, Torkamani A, Butte AJ, Glicksberg BS, Schuller B, Rodriguez B, et al. The promise of digital healthcare technologies. Front Public Health 2023;11:1196596 [FREE Full text] [doi: 10.3389/fpubh.2023.1196596] [Medline: 37822534]
3. Lee NK, Kim JS. Status and trends of the digital healthcare industry. Healthc Inform Res 2024;30(3):172-183 [FREE Full text] [doi: 10.4258/hir.2024.30.3.172] [Medline: 39160777]
4. Cui F, He X, Zhai Y, Lyu M, Shi J, Sun D, et al. Application of telemedicine services based on a regional telemedicine platform in China From 2014 to 2020: longitudinal trend analysis. J Med Internet Res 2021;23(7):e28009 [FREE Full text] [doi: 10.2196/28009] [Medline: 34255686]
5. Kruse CS, Williams K, Bohls J, Shamsi W. Telemedicine and health policy: A systematic review. Health Policy and Technology 2021;10(1):209-229. [doi: 10.1016/j.hlpt.2020.10.006]
6. Telemedicine: Opportunities and developments in member state. WHO. 2010. URL: https://www.afro.who.int/publications/telemedicine-opportunities-and-developments-member-state [accessed 2025-07-10]
7. Caiata-Zufferey M, Abraham A, Sommerhalder K, Schulz PJ. Online health information seeking in the context of the medical consultation in Switzerland. Qual Health Res 2010;20(8):1050-1061. [doi: 10.1177/1049732310368404] [Medline: 20442347]
8. Moffatt JJ, Eley DS. Barriers to the up-take of telemedicine in Australia - a view from providers. RRH 2011;11(1):116-121. [doi: 10.22605/rrh1581]
9. Noceda AVG, Acierto LMM, Bertiz MCC, Dionisio DEH, Laurito CBL, Sanchez GAT, et al. Patient satisfaction with telemedicine in the Philippines during the COVID-19 pandemic: a mixed methods study. BMC Health Serv Res 2023;23(1):277 [FREE Full text] [doi: 10.1186/s12913-023-09127-x] [Medline: 36949479]
10. Wu X, Yang Y, Li Y, Li Y, Li H, Lyu Y, et al. Using theory-based frameworks to identify barriers and enablers of physicians' telemedicine adoption and develop intervention strategies in China: multicenter qualitative study. J Med Internet Res 2025;27:e73412 [FREE Full text] [doi: 10.2196/73412] [Medline: 40920450]

XSL•FO
RenderX

11. Craig A, Lawford H, Miller M, Chen-Cao L, Woods L, Liaw S, et al. Use of technology to support health care providers delivering care in low- and lower-middle-income countries: systematic umbrella review. J Med Internet Res 2025;27:e66288 [FREE Full text] [doi: 10.2196/66288] [Medline: 40533075]
12. Wootton R, Bonnardot L. Telemedicine in low-resource settings. Front Public Health 2015;3:3 [FREE Full text] [doi: 10.3389/fpubh.2015.00003] [Medline: 25654074]
13. Xiong D, Zhao L. Research on credit evaluation of mobile medical APP interactive online consultation service-take Haodaifu APP online payment service as an example. J Phys Conf Ser 2017;910(1):012055 [FREE Full text] [doi: 10.1088/1742-6596/910/1/012055]
14. Bonnechère B, Kossi O, Mapinduzi J, Panda J, Rintala A, Guidetti S, et al. Mobile health solutions: an opportunity for rehabilitation in low- and middle income countries? Front Public Health 2022;10:1072322. [doi: 10.3389/fpubh.2022.1072322] [Medline: 36761328]
15. Dullet NW, Geraghty EM, Kaufman T, Kissee JL, King J, Dharmar M, et al. Impact of a university-based outpatient telemedicine program on time savings, travel costs, and environmental pollutants. Value Health 2017;20(4):542-546 [FREE Full text] [doi: 10.1016/j.jval.2017.01.014] [Medline: 28407995]
16. Suzuki T, Hotta J, Kuwabara T, Yamashina H, Ishikawa T, Tani Y, et al. Possibility of introducing telemedicine services in Asian and African countries. Health Policy and Technology 2020;9(1):13-22. [doi: 10.1016/j.hlpt.2020.01.006]
17. Du Y, Zhou Q, Cheng W, Zhang Z, Hoelzer S, Liang Y, et al. Factors influencing adoption and use of telemedicine services in rural areas of China: mixed methods study. JMIR Public Health Surveill 2022;8(12):e40771 [FREE Full text] [doi: 10.2196/40771] [Medline: 36563026]
18. Giebel GD, Abels C, Plescher F, Speckemeier C, Schrader NF, Börchers K, et al. Problems and barriers related to the use of mhealth apps from the perspective of patients: focus group and interview study. J Med Internet Res 2024;26:e49982 [FREE Full text] [doi: 10.2196/49982] [Medline: 38652508]
19. An Q, Kelley MM, Hanners A, Yen P. Sustainable development for mobile health apps using the human-centered design process. JMIR Form Res 2023;7:e45694 [FREE Full text] [doi: 10.2196/45694] [Medline: 37624639]
20. Ftouni R, AlJardali B, Hamdanieh M, Ftouni L, Salem N. Challenges of telemedicine during the COVID-19 pandemic: a systematic review. BMC Med Inform Decis Mak 2022;22(1):207 [FREE Full text] [doi: 10.1186/s12911-022-01952-0] [Medline: 35922817]
21. Fernández Coves A, Yeung KHT, van der Putten IM, Nelson EAS. Teleconsultation adoption since COVID-19: comparison of barriers and facilitators in primary care settings in Hong Kong and the Netherlands. Health Policy 2022;126(10):933-944 [FREE Full text] [doi: 10.1016/j.healthpol.2022.07.012] [Medline: 36050194]
22. Rosler G. Pediatric telehealth experiences: myths and truths about video visits from a parent. J Patient Exp 2020;7(6):836-838 [FREE Full text] [doi: 10.1177/2374373520932724] [Medline: 33457506]
23. Bertoncello C, Colucci M, Baldovin T, Buja A, Baldo V. How does it work? Factors involved in telemedicine home-interventions effectiveness: A review of reviews. PLoS One 2018;13(11):e0207332 [FREE Full text] [doi: 10.1371/journal.pone.0207332] [Medline: 30440004]
24. Liu Z, Brandon-Jones A, Vasilakis C. Unpacking patient engagement in remote consultation. IJOPM 2024;44(13):157-194. [doi: 10.1108/ijopm-03-2023-0188]
25. Graffigna G, Barello S, Bonanomi A, Lozza E. Measuring patient engagement: development and psychometric properties of the patient health engagement (PHE) scale. Front Psychol 2015;6:274 [FREE Full text] [doi: 10.3389/fpsyg.2015.00274] [Medline: 25870566]
26. Harrington RL, Hanna ML, Oehrlein EM, Camp R, Wheeler R, Cooblall C, et al. Defining patient engagement in research: results of a systematic review and analysis: report of the ISPOR patient-centered special interest group. Value Health 2020;23(6):677-688. [doi: 10.1016/j.jval.2020.01.019] [Medline: 32540224]
27. Hibbard JH, Mahoney ER, Stockard J, Tusler M. Development and testing of a short form of the patient activation measure. Health Serv Res 2005;40(6 Pt 1):1918-1930 [FREE Full text] [doi: 10.1111/j.1475-6773.2005.00438.x] [Medline: 16336556]
28. Menichetti J, Libreri C, Lozza E, Graffigna G. Giving patients a starring role in their own care: a bibliometric analysis of the on-going literature debate. Health Expect 2016;19(3):516-526 [FREE Full text] [doi: 10.1111/hex.12299] [Medline: 25369557]
29. Kruse CS, Krowski N, Rodriguez B, Tran L, Vela J, Brooks M. Telehealth and patient satisfaction: a systematic review and narrative analysis. BMJ Open 2017;7(8):e016242 [FREE Full text] [doi: 10.1136/bmjopen-2017-016242] [Medline: 28775188]
30. Agha Z, Schapira RM, Laud PW, McNutt G, Roter DL. Patient satisfaction with physician-patient communication during telemedicine. Telemed J E Health 2009;15(9):830-839. [doi: 10.1089/tmj.2009.0030] [Medline: 19919189]
31. Barello S, Triberti S, Graffigna G, Libreri C, Serino S, Hibbard J, et al. eHealth for patient engagement: a systematic review. Front Psychol 2015;6:2013 [FREE Full text] [doi: 10.3389/fpsyg.2015.02013] [Medline: 26779108]
32. Inglis SC, Clark RA, Dierckx R, Prieto-Merino D, Cleland JGF. Structured telephone support or non-invasive telemonitoring for patients with heart failure. Cochrane Database Syst Rev 2015;2015(10):CD007228 [FREE Full text] [doi: 10.1002/14651858.CD007228.pub3] [Medline: 26517969]

XSL•FO
RenderX

33. Vicente MA, Fernández C, Guilabert M, Carrillo I, Martín-Delgado J, Mira JJ. Patient engagement using telemedicine in primary care during COVID-19 pandemic: a trial study. Int J Environ Res Public Health 2022;19(22):14682 [FREE Full text] [doi: 10.3390/ijerph192214682] [Medline: 36429402]
34. Khanijahani A, Akinci N, Quitiquit E. A systematic review of the role of telemedicine in blood pressure control: focus on patient engagement. Curr Hypertens Rep 2022;24(7):247-258. [doi: 10.1007/s11906-022-01186-5] [Medline: 35412188]
35. Meyer MA. COVID-19 pandemic accelerates need to improve online patient engagement practices to enhance patient experience. J Patient Exp 2020;7(5):657-664 [FREE Full text] [doi: 10.1177/2374373520959486] [Medline: 33294595]
36. Costa D, Serra R. The role of communication in managing chronic lower limb wounds. J Multidiscip Healthc 2025;18:3685-3708 [FREE Full text] [doi: 10.2147/JMDH.S533416] [Medline: 40589782]
37. Talal AH, Sofikitou EM, Jaanimägi U, Zeremski M, Tobin JN, Markatou M. A framework for patient-centered telemedicine: application and lessons learned from vulnerable populations. J Biomed Inform 2020;112:103622 [FREE Full text] [doi: 10.1016/j.jbi.2020.103622] [Medline: 33186707]
38. Cingi C, Yorgancioglu A, Cingi CC, Oguzulgen K, Muluk NB, Ulusoy S, et al. The "physician on call patient engagement trial" (POPET): measuring the impact of a mobile patient engagement application on health outcomes and quality of life in allergic rhinitis and asthma patients. Int Forum Allergy Rhinol 2015;5(6):487-497. [doi: 10.1002/alr.21468] [Medline: 25856270]
39. van Middelaar T, Beishuizen CRL, Guillemont J, Barbera M, Richard E, Moll van Charante EP, HATICE consortium. Engaging older people in an internet platform for cardiovascular risk self-management: a qualitative study among Dutch HATICE participants. BMJ Open 2018;8(1):e019683 [FREE Full text] [doi: 10.1136/bmjopen-2017-019683] [Medline: 29358447]
40. Clarke AL, Roscoe J, Appleton R, Parashar D, Muthuswamy R, Khan O, et al. Promoting integrated care in prostate cancer through online prostate cancer-specific holistic needs assessment: a feasibility study in primary care. Support Care Cancer 2020;28(4):1817-1827 [FREE Full text] [doi: 10.1007/s00520-019-04967-y] [Medline: 31338642]
41. Crotty BH, Hyun N, Polovneff A, Dong Y, Decker MC, Mortensen N, et al. Analysis of clinician and patient factors and completion of telemedicine appointments using video. JAMA Netw Open 2021;4(11):e2132917 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.32917] [Medline: 34735013]
42. Habbash F, Rabeeah A, Huwaidi Z, Abuobaidah H, Alqabbat J, Hayyan F, et al. Telemedicine in non-communicable chronic diseases care during the COVID-19 pandemic: exploring patients' perspectives. Front Public Health 2023;11:1270069 [FREE Full text] [doi: 10.3389/fpubh.2023.1270069] [Medline: 37818295]
43. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
44. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
45. Niyomyart A, Ruksakulpiwat S, Benjasirisan C, Phianhasin L, Nigussie K, Thorngthip S, et al. Current status of barriers to mhealth access among patients with stroke and steps toward the digital health era: systematic review. JMIR Mhealth Uhealth 2024;12:e54511 [FREE Full text] [doi: 10.2196/54511] [Medline: 39173152]
46. Chen S, Niu M, Ngai CSB. What is the next step of ICT development? The changes of ICT use in promoting elderly healthcare access: a systematic literature review. Heliyon 2024;10(3):e25197 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e25197] [Medline: 38371988]
47. Hu Y, Ngai CSB, Chen S. Automated approaches to screening developmental language disorder: a comprehensive review and future prospects. J Speech Lang Hear Res 2025;68(5):2478-2498. [doi: 10.1044/2025_JSLHR-24-00488] [Medline: 40228046]
48. Lu W, Ngai CSB, Yi L. A bibliometric review of constituents, themes, and trends in online medical consultation research. Health Commun 2024;39(2):229-243. [doi: 10.1080/10410236.2022.2163108] [Medline: 36581497]
49. Bjerkan J, Kane B, Uhrenfeldt L, Veie M, Fossum M. Citizen-patient involvement in the development of mhealth technology: protocol for a systematic scoping review. JMIR Res Protoc 2020;9(8):e16781 [FREE Full text] [doi: 10.2196/16781] [Medline: 32857061]
50. Li T, Higgins J, Deeks J. Chapter 5: collecting data. In: Cochrane Handbook for Systematic Reviews of Interventions Version 65. London: Wiley; 2024.
51. Ackerman SL, Dowdell K, Clebak KT, Quinn M, Shipman SA. Patients assess an econsult model's acceptability at 5 us academic medical centers. Ann Fam Med 2020;18(1):35-41 [FREE Full text] [doi: 10.1370/afm.2487] [Medline: 31937531]
52. Alpert JM, Hampton CN, Raisa A, Markham MJ, Bylund CL. Integrating patient-centeredness into online patient-clinician communication: a qualitative analysis of clinicians' secure messaging usage. Support Care Cancer 2022;30(12):9851-9857 [FREE Full text] [doi: 10.1007/s00520-022-07408-5] [Medline: 36260178]
53. Bavngaard MV, Lüchau EC, Hvidt EA, Grønning A. Exploring patient participation during video consultations: A qualitative study. Digit Health 2023;9:20552076231180682 [FREE Full text] [doi: 10.1177/20552076231180682] [Medline: 37325071]

XSL•FO
RenderX

54. Björndell C, Premberg Å. Physicians' experiences of video consultation with patients at a public virtual primary care clinic: a qualitative interview study. Scand J Prim Health Care 2021;39(1):67-76 [FREE Full text] [doi: 10.1080/02813432.2021.1882082] [Medline: 33650941]
55. Breton M, Sullivan EE, Deville-Stoetzel N, McKinstry D, DePuccio M, Sriharan A, et al. Telehealth challenges during COVID-19 as reported by primary healthcare physicians in Quebec and Massachusetts. BMC Fam Pract 2021;22(1):192 [FREE Full text] [doi: 10.1186/s12875-021-01543-4] [Medline: 34563113]
56. Brodar KE, Hong N, Liddle M, Hernandez L, Waks J, Sanchez J, et al. Transitioning to telehealth services in a pediatric diabetes clinic during COVID-19: an interdisciplinary quality improvement initiative. J Clin Psychol Med Settings 2022;29(4):727-738 [FREE Full text] [doi: 10.1007/s10880-021-09830-z] [Medline: 34708318]
57. Caffery LJ, Taylor M, North JB, Smith AC. Tele-orthopaedics: a snapshot of services in Australia. J Telemed Telecare 2017;23(10):835-841. [doi: 10.1177/1357633X17732800] [Medline: 28950754]
58. Davoust M, Bazzi AR, Blakemore S, Blodgett J, Cheng A, Fielman S, et al. Patient and clinician experiences with the implementation of telemedicine and related adaptations in office-based buprenorphine treatment during the COVID-19 pandemic: a qualitative study. Addict Sci Clin Pract 2025;20(1):21 [FREE Full text] [doi: 10.1186/s13722-025-00536-3] [Medline: 40033433]
59. Day S, Rae C, McOwan A, Wilkins R, Gray A, Harvey A, et al. Patient evaluation of Klick, a technology-enabled, nurse-delivered HIV outpatient pathway. HIV Med 2025;26(1):128-139. [doi: 10.1111/hiv.13710] [Medline: 39279040]
60. Dong F, Jumper MBE, Becker-Haimes EM, Vatza C, Miao L, Conroy C, et al. Tele-mental health transitions for Pennsylvania coordinated specialty care programs for early psychosis during the COVID-19 pandemic. Psychiatr Q 2023;94(2):89-102 [FREE Full text] [doi: 10.1007/s11126-023-10015-0] [Medline: 36820952]
61. Esayed S, Kim E, Sung HC, Al-Seraji A, Adeyemo S, Troutt H, et al. Hybrid telemedicine and in-person care for kidney transplant follow-up: a qualitative study. Clin Transplant 2025;39(2):e70106 [FREE Full text] [doi: 10.1111/ctr.70106] [Medline: 39945199]
62. Gibson J, Lightbody E, McLoughlin A, McAdam J, Gibson A, Day E, et al. 'It was like he was in the room with us': patients' and carers' perspectives of telemedicine in acute stroke. Health Expect 2016;19(1):98-111 [FREE Full text] [doi: 10.1111/hex.12333] [Medline: 25581591]
63. Grens H, de Bruin JP, Huppelschoten A, Kremer JAM. Fertility workup with video consultation during the COVID-19 pandemic: pilot quantitative and qualitative study. JMIR Form Res 2022;6(2):e32000. [doi: 10.2196/32000] [Medline: 34936981]
64. Grove BE, Valen Schougaard LM, Ivarsen P, Hjollund NH, de Thurah A, Mejdahl CT. Remote follow-up based on patient-reported outcomes in patients with chronic kidney disease: A qualitative study of patient perspectives. PLoS One 2023;18(2):e0281393 [FREE Full text] [doi: 10.1371/journal.pone.0281393] [Medline: 36763600]
65. Higa C, Davidson EJ, Loos JR. Integrating family and friend support, information technology, and diabetes education in community-centric diabetes self-management. J Am Med Inform Assoc 2021;28(2):261-275 [FREE Full text] [doi: 10.1093/jamia/ocaa223] [Medline: 33164074]
66. Islind AS, Snis UL, Lindroth T, Lundin J, Cerna K, Steineck G. The virtual clinic: two-sided affordances in consultation practice. Comput Supported Coop Work 2019;28(3-4):435-468. [doi: 10.1007/s10606-019-09350-3]
67. James S, Ashley C, Williams A, Desborough J, Mcinnes S, Calma K, et al. Experiences of Australian primary healthcare nurses in using telehealth during COVID-19: a qualitative study. BMJ Open 2021;11(8):e049095 [FREE Full text] [doi: 10.1136/bmjopen-2021-049095] [Medline: 34362804]
68. Jensen AL, Schougaard LMV, Laurberg T, Hansen TK, Lomborg K. Flexible patient-reported outcome-based telehealth follow-up for type 1 diabetes: A qualitative study. Scand J Caring Sci 2023;37(3):662-676. [doi: 10.1111/scs.13154] [Medline: 36775917]
69. Jethwa H, Brooke M, Parkinson A, Dures E, Gullick NJ. Patients' perspectives of telemedicine appointments for psoriatic arthritis during the COVID-19 pandemic: results of a patient-driven pilot survey. BMC Rheumatol 2022;6(1):13 [FREE Full text] [doi: 10.1186/s41927-021-00242-y] [Medline: 35189975]
70. Jung OS, Graetz I, Dorner SC, Hayden EM. Implementing a COVID-19 virtual observation unit in emergency medicine: frontline clinician and staff experiences. Med Care Res Rev 2023;80(1):79-91 [FREE Full text] [doi: 10.1177/10775587221108750] [Medline: 35815570]
71. Moore A, Chavez C, Fisher MP. Factors enhancing trust in electronic communication among patients from an internal medicine clinic: qualitative results of the RECEPT study. J Gen Intern Med 2022;37(12):3121-3127 [FREE Full text] [doi: 10.1007/s11606-021-07345-9] [Medline: 35048293]
72. Morrison C, Beattie M, Wherton J, Stark C, Anderson J, Hunter-Rowe C, et al. Testing and implementing video consulting for outpatient appointments: using quality improvement system thinking and codesign principles. BMJ Open Qual 2021;10(1):e001259 [FREE Full text] [doi: 10.1136/bmjoq-2020-001259] [Medline: 33674346]
73. Olayiwola JN, Knox M, Dubé K, Lu EC, Woldeyesus T, James IE, et al. Understanding the potential for patient engagement in electronic consultation and referral systems: lessons from one safety net system. Health Serv Res 2018;53(4):2483-2502 [FREE Full text] [doi: 10.1111/1475-6773.12776] [Medline: 28940495]

XSL•FO
RenderX

74. Osmundsen TC, Andreassen Jaatun EA, Heggem GF, Kulseng BE. Service innovation from the edges: enhanced by telemedicine decision support. Pers Ubiquit Comput 2015;19(3-4):699-708. [doi: 10.1007/s00779-015-0857-9]
75. Rodkjær L, Jeppesen M, Schougaard L. Management of cystic fibrosis during COVID-19: patient reported outcomes based remote follow-up among CF patients in Denmark - A feasibility study. J Cyst Fibros 2022;21(2):e106-e112 [FREE Full text] [doi: 10.1016/j.jcf.2021.10.010] [Medline: 34785157]
76. Scruton S, Wong G, Babinski S, Squires LR, Berlin A, Easley J, et al. Optimizing virtual follow-up care: realist evaluation of experiences and perspectives of patients with breast and prostate cancer. J Med Internet Res 2025;27:e65148 [FREE Full text] [doi: 10.2196/65148] [Medline: 39752659]
77. Selick A, Durbin J, Hamdani Y, Rayner J, Lunsky Y. "Can you hear me now?": a qualitative exploration of communication quality in virtual primary care encounters for patients with intellectual and developmental disabilities. BMC Prim Care 2023;24(1):105 [FREE Full text] [doi: 10.1186/s12875-023-02055-z] [Medline: 37081380]
78. Spiess ST, Gardner E, Turner C, Galt A, Fortenberry K, Ho T, et al. We cannot put this genie back in the bottle: qualitative interview study among family medicine providers about their experiences with virtual visits during the COVID-19 pandemic. J Med Internet Res 2023;25:e43877 [FREE Full text] [doi: 10.2196/43877] [Medline: 37651162]
79. Trondsen MV, Tjora A, Broom A, Scambler G. The symbolic affordances of a video-mediated gaze in emergency psychiatry. Soc Sci Med 2018;197:87-94. [doi: 10.1016/j.socscimed.2017.11.056] [Medline: 29222999]
80. Wei S, Mao Y. Small talk is a big deal: a discursive analysis of online off-topic doctor-patient interaction in traditional Chinese medicine. Soc Sci Med 2023;317:115632. [doi: 10.1016/j.socscimed.2022.115632] [Medline: 36584441]
81. White SJ, Nguyen AD, Roger P, Tse T, Cartmill JA, Hatem S, et al. Tailoring communication practices to support effective delivery of telehealth in general practice. BMC Prim Care 2024;25(1):232. [doi: 10.1186/s12875-024-02441-1] [Medline: 38937674]
82. Wood SM, Pickel J, Phillips AW, Baber K, Chuo J, Maleki P, et al. Acceptability, feasibility, and quality of telehealth for adolescent health care delivery during the COVID-19 pandemic: cross-sectional study of patient and family experiences. JMIR Pediatr Parent 2021;4(4):e32708 [FREE Full text] [doi: 10.2196/32708] [Medline: 34779782]
83. Zainal H, Hui XX, Thumboo J, Fong W, Yong FK. Patients' expectations of doctors' clinical competencies in the digital health care era: qualitative semistructured interview study among patients. JMIR Hum Factors 2024;11:e51972 [FREE Full text] [doi: 10.2196/51972] [Medline: 39190915]
84. Canale M. From communicative competence to communicative language pedagogy. In: Language and Communication. London: Longman Press Publishing; 1983:2-27.
85. Porter ME. What is strategy? Harvard business review 1996;74(6):61-78 [FREE Full text]
86. Porter ME, Lee TH. Why strategy matters now. N Engl J Med 2015;372(18):1681-1684. [doi: 10.1056/NEJMp1502419] [Medline: 25923546]
87. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The mixed methods appraisal tool (MMAT) version 2018 for information professionals and researchers. EFI 2018;34(4):285-291. [doi: 10.3233/EFI-180221]
88. Draganidis A, Fernando AN, West ML, Sharp G. Social media delivered mental health campaigns and public service announcements: A systematic literature review of public engagement and help-seeking behaviours. Soc Sci Med 2024;359:117231 [FREE Full text] [doi: 10.1016/j.socscimed.2024.117231] [Medline: 39278158]
89. Ngai CSB, Singh RG, Lu W, Koon AC. Grappling with the COVID-19 health crisis: content analysis of communication strategies and their effects on public engagement on social media. J Med Internet Res 2020;22(8):e21360 [FREE Full text] [doi: 10.2196/21360] [Medline: 32750013]
90. Ngai CSB, Singh RG, Yao L. Impact of COVID-19 vaccine misinformation on social media virality: content analysis of message themes and writing strategies. J Med Internet Res 2022;24(7):e37806 [FREE Full text] [doi: 10.2196/37806] [Medline: 35731969]
91. Hsieh H, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
92. Elliott V. Thinking about the coding process in qualitative data analysis. TQR 2018. [doi: 10.46743/2160-3715/2018.3560]
93. Ritchie J, Spencer L. Qualitative data analysis for applied policy research. In: Analyzing Qualitative Data Taylor & Francis e-Library. Thousand Oaks, CA: Sage Publications, Inc; 2002:173-194.
94. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310]
95. Borenstein M, Hedges LV, Higgins J, Rothstein H. When does it make sense to perform a meta'analysis? In: Introduction to meta?analysis. New Jerssey: John Wiley & Sons, Ltd; 2009:357-364.
96. Kwame A, Petrucka PM. A literature-based study of patient-centered care and communication in nurse-patient interactions: barriers, facilitators, and the way forward. BMC Nurs 2021;20(1):158 [FREE Full text] [doi: 10.1186/s12912-021-00684-2] [Medline: 34479560]
97. Gittell JH, Ali HN. The theory of relational coordination. In: Relational Analytics. United Kingdom: Routledge; 2021:16-38.
98. Campinha-Bacote J. The process of cultural competence in the delivery of healthcare services: a model of care. J Transcult Nurs 2002;13(3):181-4; discussion 200. [doi: 10.1177/10459602013003003] [Medline: 12113146]

XSL•FO
RenderX

99. Ngai CSB, Singh RG, Lu W. Exploring drivers for public engagement in social media communication with medical social influencers in China. PLoS One 2020;15(10):e0240303 [FREE Full text] [doi: 10.1371/journal.pone.0240303] [Medline: 33027269]
100. Sungur H, Yılmaz NG, Chan BMC, van den Muijsenbergh METC, van Weert JCM, Schouten BC. Development and evaluation of a digital intervention for fulfilling the needs of older migrant patients with cancer: user-centered design approach. J Med Internet Res 2020;22(10):e21238 [FREE Full text] [doi: 10.2196/21238] [Medline: 33104008]
101. Crafoord M, Fjell M, Sundberg K, Nilsson M, Langius-Eklöf A. Engagement in an interactive app for symptom self-management during treatment in patients with breast or prostate cancer: mixed methods study. J Med Internet Res 2020;22(8):e17058 [FREE Full text] [doi: 10.2196/17058] [Medline: 32663140]
102. Xie Z, Nacioglu A, Or C. Prevalence, demographic correlates, and perceived impacts of mobile health app use amongst Chinese adults: cross-sectional survey study. JMIR Mhealth Uhealth 2018;6(4):e103 [FREE Full text] [doi: 10.2196/mhealth.9002] [Medline: 29699971]
103. Vasiloglou MF, Christodoulidis S, Reber E, Stathopoulou T, Lu Y, Stanga Z, et al. Nutrients 2020;12(8):2214 [FREE Full text] [doi: 10.3390/nu12082214] [Medline: 32722339]
104. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. Management Science 1989;35(8):982-1003. [doi: 10.1287/mnsc.35.8.982]
105. Alzahrani A, Gay V, Alturki R. Exploring Saudi individuals' perspectives and needs to design a hypertension management mobile technology solution: qualitative study. Int J Environ Res Public Health 2022;19(19):12956 [FREE Full text] [doi: 10.3390/ijerph191912956] [Medline: 36232254]
106. Chu D, Lessard D, Laymouna MA, Engler K, Schuster T, Ma Y, et al. Understanding the risks and benefits of a patient portal configured for hiv care: patient and healthcare professional perspectives. J Pers Med 2022;12(2):314. [doi: 10.3390/jpm12020314] [Medline: 35207803]
107. Toni E, Pirnejad H, Makhdoomi K, Mivefroshan A, Niazkhani Z. Patient empowerment through a user-centered design of an electronic personal health record: a qualitative study of user requirements in chronic kidney disease. BMC Med Inform Decis Mak 2021;21(1):329 [FREE Full text] [doi: 10.1186/s12911-021-01689-2] [Medline: 34819050]
108. Chaudhry T, Ormandy P, Vasilica C. Using FLO text-messages to enhance health behaviours and self-management of long-term conditions in South-Asian patients. Digit Health 2024;10:20552076241242558 [FREE Full text] [doi: 10.1177/20552076241242558] [Medline: 38708186]
109. Jiang Y, Hwang M, Cho Y, Friese CR, Hawley ST, Manojlovich M, et al. The acceptance and use of digital technologies for self-reporting medication safety events after care transitions to home in patients with cancer: survey study. J Med Internet Res 2024;26:e47685 [FREE Full text] [doi: 10.2196/47685] [Medline: 38457204]
110. Shewchuk B, Green LA, Barber T, Miller J, Teare S, Campbell-Scherer D, et al. Patients' use of mobile health for self-management of knee osteoarthritis: results of a 6-week pilot study. JMIR Form Res 2021;5(11):e30495 [FREE Full text] [doi: 10.2196/30495] [Medline: 34842526]
111. Hibbard JH, Stockard J, Mahoney ER, Tusler M. Development of the patient activation measure (PAM): conceptualizing and measuring activation in patients and consumers. Health Serv Res 2004;39(4 Pt 1):1005-1026 [FREE Full text] [doi: 10.1111/j.1475-6773.2004.00269.x] [Medline: 15230939]
112. Elwyn G, Edwards A, Wensing M, Hood K, Atwell C, Grol R. Shared decision making: developing the OPTION scale for measuring patient involvement. Qual Saf Health Care 2003;12(2):93-99 [FREE Full text] [doi: 10.1136/qhc.12.2.93] [Medline: 12679504]
113. Lerman CE, Brody DS, Caputo GC, Smith DG, Lazaro CG, Wolfson HG. Patients' perceived involvement in care scale: relationship to attitudes about illness and medical care. J Gen Intern Med 1990;5(1):29-33. [doi: 10.1007/BF02602306] [Medline: 2299426]
114. Song M, Kim M. Development and validation of a patient participation scale. J Adv Nurs 2023;79(6):2393-2403. [doi: 10.1111/jan.15593] [Medline: 36814372]

Abbreviations

HCP: health care provider
mHealth: mobile health
MMAT: Mixed Methods Appraisal Tool
OPTION: observing patient involvement in decision making
PAM: Patient Activation Measurement
PHE: 5-item Patient Health Engagement
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension
PROSPERO: International Prospective Register of Systematic Reviews
RQ: research question

XSL•FO
RenderX

Edited by S Brini; submitted 08.Oct.2025; peer-reviewed by Y Li, E Afarikumah; comments to author 24.Oct.2025; revised version received 12.Dec.2025; accepted 22.Dec.2025; published 21.Jan.2026.

Please cite as:
Hu Y, Ngai CSB, Jiang R
Communication Strategies to Promote Patient Engagement in Telemedicine: Systematic Review
J Med Internet Res 2026;28:e85456
URL: https://www.jmir.org/2026/1/e85456
doi:10.2196/85456
PMID:

©Yangna Hu, Cindy Sing Bik Ngai, Rui Jiang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Products, Performance, and Technological Development of Ambulatory Oxygen Therapy Devices: Scoping Review

Shohei Kawachi[1,2*], BPT, PhD; Mariana Hoffman[1,3*], PhD; Lorena Romero[4], BA, MBIT; Magnus Ekström[5], MD, PhD; Jerry A Krishnan[6], MD, PhD; Anne E Holland[1,3,7*], BAppSc, PhD

[1]Respiratory Research@Alfred, School of Translational Medicine, Monash University, Melbourne, Victoria, Australia
[2]Department of Rehabilitation, Shinshu University Hospital, Matsumoto, Nagano, Japan
[3]Institute for Breathing and Sleep, Melbourne, Victoria, Australia
[4]Ian Potter Library, The Alfred Hospital, Melbourne, Victoria, Australia
[5]Respiratory Medicine, Allergology and Palliative Medicine, Department of Clinical Sciences Lund, Lund University, Lund, Sweden
[6]Breathe Chicago Center, University of Illinois Chicago, Chicago, IL, United States
[7]Departments of Physiotherapy and Respiratory Medicine, Alfred Health, Melbourne, Victoria, Australia
[*]these authors contributed equally

Corresponding Author:
Shohei Kawachi, BPT, PhD
Respiratory Research@Alfred
School of Translational Medicine
Monash University
Level 6 The Alfred Centre, 99 Commercial Rd
Melbourne, Victoria, 3004
Australia
Phone: 81 762652948
Email: shohei.kawachi@gmail.com

Abstract

Background: Ambulatory oxygen therapy is prescribed for patients with chronic lung diseases who experience exertional hypoxemia. However, available devices may not adequately meet user requirements, and their performance characteristics are heterogeneous.

Objective: This study aims to identify devices available for delivery of ambulatory oxygen therapy, the technologies that they use to generate oxygen, the performance characteristics of each device, and the development status.

Methods: We used medical and engineering databases to identify peer-reviewed papers (eg, MEDLINE, IEEE). Gray literature was used to identify additional descriptions of ambulatory oxygen devices in military medicine, space exploration, or patents. The last search was conducted in September 2025. Documents that described a device that can deliver oxygen in an ambulatory context (defined as weighing less than 10 kg) and were written in English were included. Search results were screened for inclusion by 2 independent reviewers. Data were synthesized by descriptively mapping the performance of each product, the technology used, and the development status of emerging technologies.

Results: From 9702 records identified, a total of 166 met eligibility criteria (106 scientific publications and 60 gray literature). We identified 33 portable oxygen concentrators (POCs; 29 commercially available), 10 oxygen cylinders, and 6 portable liquid oxygen (LOX) devices. The POC products showed a trade-off between portability and oxygen delivery capacity (maximum flow rate ranging from 2.0 to 6.0 L/min; device weight ranging from 1.0 to 9.1 kg). Pressure swing adsorption with zeolite was the most common oxygen generation technology in POCs on the market. The mean maximum continuous operating time of POCs was 3.8 hours. Two prototype POCs (maximum flow rate of 4-6 L/min and device weight of 8-9 kg) were developed for space exploration using modified adsorbents. LOX devices were the lightest and had the longest continuous operating time. Innovations in delivery included the downsizing of a POC by using nanozeolite as an adsorbent and pulse oximeter oxygen saturation (SpO_2)–targeted automatic titration of oxygen delivery based on the user's SpO_2.

Conclusions: This scoping review is the first study to integrate medical, engineering, and gray literature on ambulatory oxygen devices and their development. Although prior literature has narratively explained the products and technologies, no previous research has systematically investigated them. This review showed that POCs available to consumers may not meet the needs of

XSL•FO
RenderX

patients in terms of flow rate, portability, and operating time. LOX devices offered superior performance but are limited by high costs. Limitations of this review include the difficulty of comparing product performance across oxygen delivery settings and that the records were largely obtained from English-language sources. Innovation in ambulatory oxygen technology has been limited over the past decade, highlighting urgent need for research and development of new lightweight devices with higher oxygen delivery.

Trial Registration: OSF Registries 10.17605/OSF.IO/QS7FX; https://osf.io/qs7fx

(J Med Internet Res 2026;28:e81077) doi:10.2196/81077

KEYWORDS

ambulatory oxygen therapy; automatic titration of oxygen; home oxygen therapy; liquid oxygen; medical device innovation; oxygen cylinder; portable oxygen concentrator; pressure swing adsorption; zeolite

Introduction

Background

Supplemental oxygen may be prescribed to correct severe hypoxemia at rest, with exertion, and/or during sleep. Ambulatory oxygen therapy is defined as the use of supplemental oxygen during exertion [1]. The American Thoracic Society guideline conditionally recommends prescribing ambulatory oxygen therapy for patients who have severe exertional room air hypoxemia [1], based on acute improvements in exercise capacity and modest evidence of improvements in health-related quality of life [2,3]. Portable oxygen concentrators (POCs), oxygen cylinders, and liquid oxygen (LOX) can be used to deliver ambulatory oxygen therapy [1]. Oxygen cylinders for ambulatory use can deliver continuous oxygen flow up to 6 L per minute but run out quickly, especially when used at high flow rates, and need to be refilled. POCs supply concentrated oxygen by removing nitrogen from the air as long as they have a power source; however, they may provide oxygen only intermittently ("pulse flow" triggered by inspiration) and may have limited battery life. In continuous-flow settings, oxygen is delivered throughout both inhalation and exhalation, resulting in substantial oxygen waste during exhalation. Pulse-flow settings were developed to minimize this waste by supplying oxygen only during inspiration, thereby conserving both oxygen and battery power [4]. However, if the pulse is not synchronized well with inspiration, a portion of the oxygen bolus may be exhaled before reaching the alveoli [5]. LOX supplies oxygen by gradually evaporating LOX at cryogenic temperatures and can deliver higher flow rates for longer periods. However, LOX devices are costly, and availability may be limited [6].

In previous studies, users of portable oxygen devices reported a lack of physically manageable portable systems, a lack of devices capable of delivering higher oxygen-flow rates (>3 L/min), and an inability to leave their homes for more than 2-4 hours due to lack of reliable and enduring ambulatory oxygen supply [7-9]. In addition, users may face large out-of-pocket expenses for the ongoing costs of oxygen equipment [7]. Given these limitations, people using ambulatory oxygen reported that it was important to have a variety of device options that allowed a choice based on their individual needs [9]. However, there is little information on the types of devices available, the performance of each device, and the costs of ambulatory oxygen therapy devices [7]. It is possible that technologies to deliver ambulatory oxygen therapy are available in other fields (eg, military medicine) that are not yet available as commercial ambulatory devices.

A scoping review was deemed most appropriate to investigate the current status of portable oxygen devices, as the concepts of interest (technology, performance characteristics, and cost) extend beyond the medical field and may be published outside traditional peer-reviewed medical literature (eg, patent documents, technical reports) [10]. This scoping review aimed to identify the range of available portable oxygen devices, the technologies that are used to generate oxygen by the devices, the performance of each device, and the costs associated with its use. In addition, given the limitations of current portable oxygen devices, it is clinically important to clarify the status of technological innovation. Some patients require high oxygen flow rates that exceed the capacity of the currently available portable oxygen devices [11].

Furthermore, it is recognized that the weight of portable oxygen systems may limit their usability in clinical practice [12]. This may reduce adherence and contribute to inconsistent evidence for efficacy in exertional hypoxemia [11,13]. Therefore, we also aimed to identify innovations in the design of ambulatory oxygen therapy devices, such as improvements in weight or flow rates, and pulse oximeter oxygen saturation (SpO_2)–targeted automatic titration of oxygen delivery, which may not be available commercially [14].

Objectives

This scoping review aimed to identify the range of available portable oxygen devices, the technologies that are used to generate oxygen by the devices, the performance of each device, and the costs associated with its use. We also aimed to identify innovations in the design of ambulatory oxygen therapy devices, such as improvements in weight or flow rates and SpO_2-targeted automatic titration of oxygen delivery, which may not be available commercially.

Methods

Overview

This scoping review was conducted according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guideline (Multimedia Appendices 1-3; [15]), and the Joanna Briggs Institute (JBI) Manual for Evidence Synthesis [16]. The

XSL•FO
RenderX

protocol was registered prospectively with the Open Science Framework on May 29, 2024.

Eligibility Criteria

We included English language documents that described a device that can deliver oxygen in an ambulatory context, including scientific papers (including review articles) and gray literature (conference proceedings, patent documents, company websites, technical reports, and government documents). We excluded documents that only described stationary equipment, high-flow nasal oxygen therapy, hyperbaric oxygen therapy, positive pressure ventilation systems, heliotherapy, short-burst oxygen therapy, and wall-based high-flow nasal oxygen therapy, as well as those limited to over-the-counter medical devices.

Search Strategy

An initial search was performed on May 29, 2024, in MEDLINE (Ovid) and IEEE Xplore. After the initial search, the text words in the titles and abstracts of the retrieved articles and the index terms used to describe the articles were analyzed to refine the search terms. For the second search using updated terms, we searched scientific papers in 5 databases (MEDLINE [Ovid], Embase [Ovid], SCOPUS, CENTRAL [Wiley], and IEEE Xplore) using the search strings created with index terms (MeSH [Medical Subject Headings] terms or IEEE terms). Search strings are shown in Table S1 in Multimedia Appendix 4.

For searching gray literature, the following three strategies were used based on the guideline for gray literature and a previous study [17,18]: (1) targeted website browsing and searching, (2) gray literature database search, and (3) search engine searching. Specifically, technical reports, white papers including military medicine, and other gray literature on portable oxygen devices (manuals, company websites on product performance, etc) were searched using the appropriate websites respectively (National Technical Reports Library, World Health Organization, Defense Technical Information Center, International Health technology Assessment Database, and Google Advanced). As complex search strings were not allowed in some gray literature databases, we adjusted the search strings to match the database functionality. Gray literature search strings are also shown in Table S1 in Multimedia Appendix 4. Google Advanced searches were run on corporate, government, and military domains (.com, .gov, and .mil) to review the top 100 results from each. The latest patents of inventions and developments for the period January 1, 2022, to June 25, 2024, were searched in the World Intellectual Property Organization (WIPO) database using adjusted search strings. The search was repeated on September 29, 2025, for CENTRAL, MEDLINE (Ovid), and Embase (Ovid), on September 25, 2025, for all other databases including Scopus, IEEE Xplore, and gray literature sources, and on September 29, 2025, for WIPO.

A third search consisted of hand searching reference lists of all selected articles and review papers to identify any additional articles not found by the first two search methods. Additionally, Google Advanced searches were performed to obtain missing product performance information. The search period in each database was 2004 to the present to capture devices that were currently or recently available. We reported database searches and literature selection according to PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Search extension) [19].

Selection of Studies

Scientific papers were imported from the databases into the Covidence platform [20], and gray literature search results were imported into a spreadsheet. In the Covidence platform, duplicate records were automatically removed. Two independent reviewers (SK and MH) screened titles and abstracts of scientific papers for eligibility. Studies that met the inclusion criteria, or for which eligibility was unclear, underwent full-text review. Gray literature records were also screened for eligibility by 2 independent reviewers (SK and MH). Disagreements in study selection were resolved by consensus or by consulting a third reviewer (AEH). The reasons for exclusion at the full-text stage are reported.

Outcomes of Interest

Outcome selection was guided by patient priorities for oxygen devices, identified in previous studies and are shown in Table 1 [7-9].

XSL•FO
RenderX

Table 1. Outcomes of interest in the scoping review of ambulatory oxygen therapy devices extracted from scientific and gray literature published between 2004 and 2025.

Category	Outcome
Portable oxygen devices and their performance characteristics	• Type of device (POCs[a], portable oxygen cylinders, portable LOX[b]) • Product name • Manufacturer • Release date • Size of device • Weight of device (and battery weight where relevant) • Pulse or continuous flow • Method of transporting the device • Range of flow rates in the pulse flow setting • Range of flow rates in continuous flow setting • Pulse-dose bolus volume in maximum pulse flow setting (as much as possible at a respiratory rate of 20 breaths per minute) • Maximum continuous operating time (in the case of POC, battery duration, and in the case of oxygen cylinders or LOX, the duration until the oxygen stored in the tank runs out) • Concentration of supplied oxygen[c] • Operating noisec • Battery recharge time • Trigger sensitivity • Technology used to deliver oxygen (eg, nature of the sorbent in POC) • Product availability to consumers • FDA[d] and EMA[e] approved or equivalent • FAA[f] approved for air travel or equivalent • Frequency of product failures • Remote control function for changing flow rate
Cost of portable oxygen devices	• Description of publicly available supplier costs for each device (eg, purchase and rental costs of equipment, electricity costs for its use, etc)[g]

[a]POC: portable oxygen concentrator.

[b]LOX: liquid oxygen.

[c]Concentration of supplied oxygen and operating noise were included as they are important for the International Organization for Standardization (ISO) requirements to be approved as a medical device through the United States Food and Drug Administration and European Medicines Agency [21].

[d]FDA: Food and Drug Administration.

[e]EMA: European Medicines Agency.

[f]FAA: Federal Aviation Administration.

[g]The cost information, only officially disclosed by the company of each product, was extracted.

Data Charting Process

Data from the included sources of evidence were charted using a custom-designed form (SK and MH). Two review authors (SK and MH) independently charted the data from the eligible studies. Disagreements regarding the data charting between authors were resolved by discussion. If consensus could not be reached, a third author (AEH) reviewed the study and arbitrated.

The data chart included: types of publication (scientific paper, patent documents, technical report, etc), types of content (products, performances, costs, etc), author, year of publication, country where the information originated, information related to the products and performances outlined, and information related to the costs.

Risk of Bias (Quality Assessment)

Scoping reviews are conducted to provide an overview of the existing evidence regardless of methodological quality or risk of bias [22]. Therefore, the included sources of evidence are typically not critically appraised for scoping reviews. As such, we did not undertake a quality assessment of the included sources of evidence.

Data Synthesis

To synthesize data on products and their performance, we prioritized information published by the product company (including user manual), scientific papers, white papers, technical reports, and websites, in that order, if there was different information on the same item regarding one product. If multiple products were identified as having the same product name but differing only in numbers, the newest product was checked at the official website, and only the newest one was extracted (eg, Eclipse 1/2/3/4/5 Caire, Ball Ground, United States). As stationary oxygen concentrators are generally considered to weigh 10 kg or more, and POCs range from 1-9 kg [4], we defined ambulatory oxygen devices in this study as those weighing less than 10 kg and excluded heavier devices. Performance characteristics, the technology used, and the current status of development were reported descriptively. Costs for ambulatory oxygen therapy were mapped descriptively by country and type of device (portable oxygen cylinders, POCs,

XSL•FO
RenderX

and portable LOX). Descriptive statistics were used to describe all data. Categorical data were presented as frequency and percentages. In addition to descriptive synthesis, we developed a gap map to describe areas in need of future development. The gap map addressed the degree to which key patient needs for ambulatory oxygen devices (oxygen flow rate, device weight, operating time, and auto-titration) are met by characteristics of current products or those in development. Patients' needs extracted from previous studies were used as the framework [7-9].

Results

Principal Findings

A total of 9702 records were identified from scientific databases (medical database: n=3720; engineering database: n=133; multidisciplinary database: n=1842), and 4007 records from gray literature. After removing duplicates, a total of 3028 records from scientific databases and 4007 records from gray literature sources were screened (Figure 1). Of these, a total of 166 records related to ambulatory oxygen therapy devices were finally included in this review (n=106 from scientific databases; n=60 from gray literature sources). Among the 166 included records, a total of 81 (48.8%) originated from North America, 37 (22.3%) from Europe, 31 (18.7%) from Asia, 11 (6.6%) from Oceania, 2 (1.2%) from Africa, and 2 (1.2%) from South America (Table S2 in Multimedia Appendix 4). The scientific literature included 85 original research articles, 20 review articles (of which 1 was a systematic review and 19 were narrative reviews), and 1 clinical guideline. The gray literature comprised 7 technical reports, 3 white papers, 3 clinical trial registrations, 14 government-related websites (eg, military, energy, and space exploration), 4 company websites, 1 charitable organization website, and 28 patents (Tables 2 and 3). Although we identified 3 chemical oxygen generators in the scientific literature, they were excluded because chemical oxygen generators exhaust in 30 minutes or less and their output cannot be adjusted, making these devices unsuitable for delivery of ambulatory oxygen in clinical care [23].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) flow diagram showing the identification, screening, and inclusion of studies in this scoping review of ambulatory oxygen therapy devices. The search included scientific and gray literature from 2004 to 2025 across medical, engineering, military, and space exploration fields.

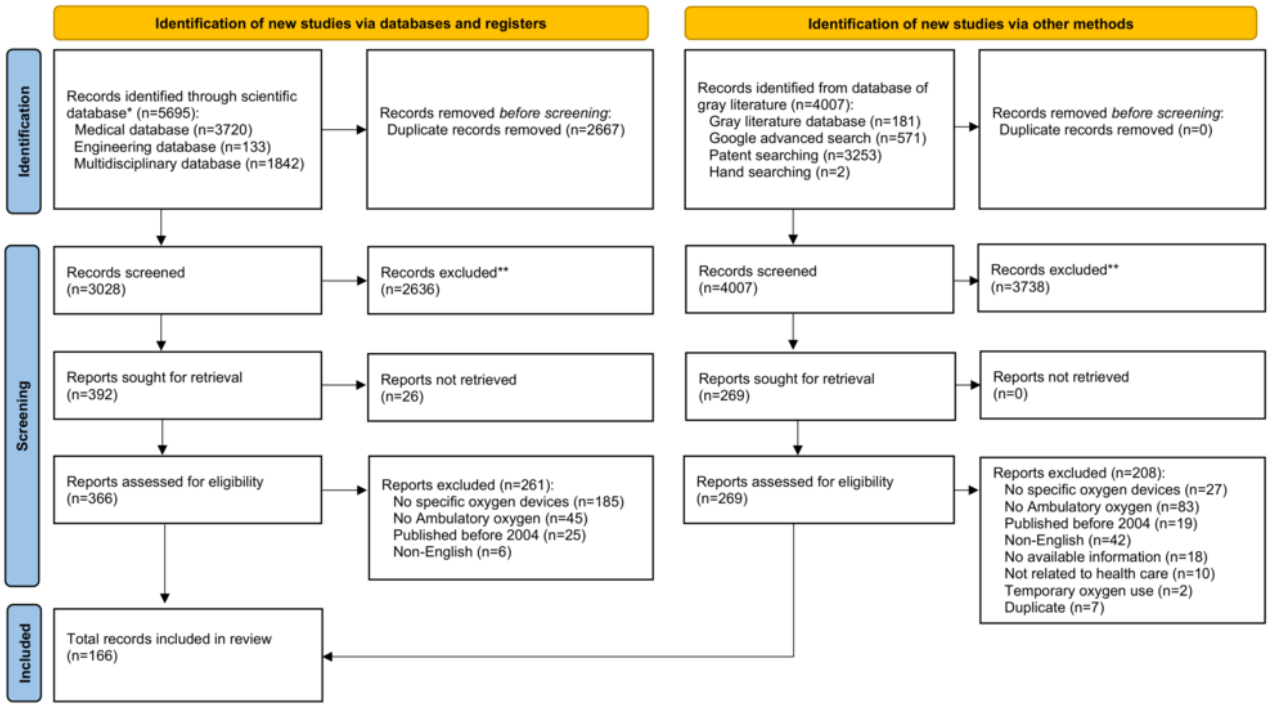

XSL•FO
RenderX

Table 2. Summary of literature on ambulatory oxygen therapy devices (n=166). Types of sources identified in this scoping review include scientific articles, technical reports, patents, government and military documents, and company websites published between 2004 and 2025.

Type of literature	Records, n (%)
Scientific literature (n=106)	
Original paper	85 (51)
Review paper	20 (12)
Systematic	1 (1)
Narrative	19 (11)
Guideline	1 (1)
Gray literature (n=60)	
Technical report	7 (4)
White paper	3 (2)
Clinical trial registry	3 (2)
Government website	14 (9)
Military	8 (5)
Energy	5 (3)
Space exploration	1 (1)
Company website	4 (2)
Charitable organization	1 (1)
Patent	28 (17)

Table 3. Summary of content from eligible literature (n=166) included in this scoping review of ambulatory oxygen therapy devices. Information sources span medicine, engineering, energy, military, and space exploration fields and describe portable oxygen concentrators, oxygen cylinders, and liquid oxygen systems, including innovations in portability, oxygen generation, and automatic titration technologies (2004-2025).

Type of contents	Records, n
Source of information	
Medicine	87
Engineering	63
Military	10
Energy	5
Space exploration	1
Device	
Oxygen concentrator (including devices not on market)	156
Oxygen cylinders	50
Liquid oxygen	16
New technology	
Downsize and portability enhancement	13
Improving or optimizing the oxygen generation process	15
Remote control (ie, automatic titration)	17
Other	12

Portable Oxygen Devices

We identified 33 different POCs in this study, of which 29 were on the market, 1 had discontinued production, and 3 were prototypes under development. The summary of identified POCs is shown in Table 4. The weight of the POCs products ranged from 1.0 to 9.1 kg, and size ranged from 15.7 × 11.7 × 6.1 cm (756 cm^3) to 51.3 × 27.7 × 20.3 cm (28,847 cm^3). Oxygen delivery capacity varied widely, with maximum flow rate in the continuous flow setting ranging from 2.0 to 6.0 L/min and maximum pulse-dose bolus volume of pulse flow setting ranging from 17.3 to 90 mL of oxygen. The pulse-dose bolus volume

XSL•FO
RenderX

varies widely across products, even within the same pulse flow setting (Figure 2). A trade-off between portability and oxygen delivery was shown in the identified POCs (Figure 3, Table S3 in Multimedia Appendix 4). Smaller POC products tended to be transported using bags, while larger POC products were transported using a dedicated cart (Table S3 in Multimedia Appendix 4). Smaller POC products also tended to have only pulse flow settings. The most recent POCs were the OxLife Liberty2 (25.4 × 22.9 × 8.9 cm), released in 2024, and the DISCOV-R (23.6 × 10.4 × 25.2 cm), released in 2023, which were capable of delivering oxygen in the continuous flow setting. This contrasts with earlier POCs of the same size (eg, Inogen Rove 6, 18.3 × 8.3 × 20.5 cm with single battery), which only had a pulse flow setting (Table S3 in Multimedia Appendix 4). The mean maximum continuous operating time in pulse flow setting was 3.8 hours, with a range of 1.3 to 6.3 hours, depending on flow settings and battery option (Table 4 and Figure 4). The maximum operating time tended to be shortened by half in the continuous flow setting compared to the same flow rate in the pulse flow setting in terms of equivalent volume (Table 4 and Table S3 in Multimedia Appendix 4). There were also POC products, such as the Simply Go mini, which extended the continuous operating time from 2 hours 40 minutes to 5 hours with an additional battery (Table S3 in Multimedia Appendix 4).

Pressure swing adsorption (PSA), vacuum pressure swing adsorption (VPSA), and vacuum pressure cycle (VPC) were the technologies used for oxygen generation (Table 4, Table S4 in Multimedia Appendix 4). Zeolites were identified as the adsorbent used in POCs. Air is composed of approximately 80% nitrogen, and PSA is a method to generate high concentrations of oxygen by selectively adsorbing nitrogen from air. In PSA, air pressurized by a compressor is sent through an adsorption column with zeolite to adsorb nitrogen and increase oxygen concentration. Then, the pressure is reduced to release the nitrogen. This adsorption and desorption cycle process is alternated in the multiple adsorption columns to continuously produce highly concentrated oxygen [24]. VPSA and VPC are adsorption methods that use a vacuum pump in combination with or instead of a compressor to reduce the pressure in the column [25]. We identified 2 POC prototypes and 1 POC product that were under development. The 4-SLPM prototype developed by TDA Research (Wheat Ridge, United States) weighed approximately 7.8 kg and measured 28 × 25 × 18 cm (Table S3 in Multimedia Appendix 4). The 4-SLPM has a maximum oxygen flow rate of 4 L/min in continuous flow setting, greater than most POCs on the market of comparable weight. The 4-SLPM used high lithium-exchanged X (LiLSX) as the adsorbent. The other PSA-based prototype, co-developed by NASA and Chart Industries (Ball Ground, United States), weighed less than 8.2 kg with a size of 35.6 × 30.5 × 20.3 cm and was capable of delivering up to 6 L/min in the continuous flow setting. There was no available information on commercially available POC innovations, such as remote control of flow settings in POCs. A POC product under development named JUNO (Roam Technologies Pty Ltd, Carlton NSW, Australia) was identified. According to the company's product information, JUNO is an ultraportable, tankless oxygen system currently under clinical development. Despite being designed to be carried with one hand, it has a continuous flow setting ranging from 1-3 L/min with a concentration of 91% [26]. Although the measurement conditions and detailed specification are not publicly disclosed (Table S3 in Multimedia Appendix 4), the recent patent on downsizing of POCs by Roam Technologies Pty Ltd, has been identified. The patent adopts a miniaturized PSA architecture with a compact arrangement of the adsorbent layer and internal gas pathway structure to shorten flow paths and minimize dead space [27].

Oxygen cylinders and LOX are summarized in Tables 5 and 6. Table 5 and Tables S3, S4, and S5 (shown in Multimedia Appendix 4) are also provided as xlsx files in Multimedia Appendix 5. The weight of cylinders ranged from 0.3-6.5 kg, and the size ranged from 14.9 × 6.4 cm to 86.5 × 10.2 cm. The maximum continuous operating time varied from 0.3 to 5.0 hours at 2 L/min in the continuous flow setting. The continuous operating time per continuous flow setting was comparable to the POC. (Figure 4). Smaller cylinders were transported using a shoulder bag, whereas larger cylinders required trolleys or hand-drawn carts for transport. Conventional cylinders were generally filled at approximately 150 bar, whereas the Ultra Lightweight Cylinder Oxygen System (IOSVR) uses a high-pressure filling of 300 bar. This allows nearly double the amount of oxygen to be filled compared with a comparably sized type M7 cylinder (Table 6). IOSVR is made possible by the high-strength aluminum alloy (L7X), which is reinforced with carbon fiber wrapping technology [28]. LOX products weighed from 1.6-3.7 kg (filled) and had gaseous oxygen capacities ranging from 275.0-1058 L (Table 6). Many oxygen cylinders require a pressure regulator, which is attached to the cylinder's top and works like a tap, allowing the safe adjustment of oxygen flow rate provided, in L/min [4]. In addition, some regulators support a pulse-dose delivery mode, which can extend the operating duration compared with continuous flow [4,29].

XSL•FO
RenderX

Table 4. Summary of identified portable oxygen concentrators (n=33) included in this scoping review of ambulatory oxygen therapy devices. Extracted characteristics include device weight and dimensions, flow settings, oxygen output, operating noise, trigger sensitivity, oxygen delivery technology, regulatory approval, and market availability based on records published between 2004 and 2025.

Variables	Statistic	Notes
Device volume (cm^3), mean (range)	8342.4 (755.9-28,846.5)	n=32
Weight (kg), mean (range)	3.8 (1.0-9.1)	n=16 (including battery); n=16 (unclear included)
Transport method, n	15	n=30; some devices fall into multiple categories
Carrying bag	2	
Carry case	2	
Backpack	6	
Cart	3	
Shoulder strap	2	
Waist	1	
Carry-on baggage	3	
PF[a] and CF[b] setting, n		
Only PF	19	N/A[c]
Both CF and PF	11	N/A
Not reported	2	N/A
Maximum number of PF setting, mean (range)	5 (1-10)	n=29
Maximum flow rates in CF setting (L/min), mean (range)	2.9 (2.0-6.0)	n=12 (only devices with CF settings)
Maximum pulse-dose bolus in PF setting (mL), mean (range)	53.7 (17.3-90.0)	n=15 (Only those at 20 bpm)
Maximum continuous operating time in PF setting (h)[d], mean (range)	3.8 (1.3-6.3)	n=25; depends on settings or respiratory rate
Maximum continuous operating time in CF setting (h)[d], mean (range)	1.9 (0.6-4.5)	n=9; depends on settings or respiratory rate
Concentration of supplied oxygen (%)[d], mean (range)	90.4 (82.0-96.0)	n=29; depends on settings or respiratory rate
Operating noise (dBA[e])[d], mean (range)	44.3 (35.0-59.0)	n=28; depends on settings or respiratory rate
Maximum battery recharge time (h), mean (range)	4.4 (1.5-12.8)	n=27
Trigger sensitivity (cm H_2O), mean (range)	–0.24 (–0.50 to –0.05)	n=23
Nature of the sorbent, n		
Molecular sieve	14	N/A
Zeolite	3	N/A
High lithium-exchanged X	1	N/A
Not reported	14	N/A
Technology used, n		
Pressure swing adsorption	7	N/A
Vacuum pressure swing adsorption	1	N/A
Vacuum pressure cycle	1	N/A
Not reported	23	N/A
Market availability, n		
On the market	29	N/A
Discontinued	1	N/A
Prototype	2	N/A

XSL•FO
RenderX

Variables	Statistic	Notes
FDA[f], n		
Approved	19	N/A
Not reported	13	N/A
FAA[g], n		
Approved	28	N/A
Not reported	4	N/A

[a]PF: pulse flow.

[b]CF: continuous flow.

[c]N/A: not applicable.

[d]JUNO (Roam Technologies Pty Ltd, Carlton NSW, Australia) was excluded from this analysis due to a significant lack of information.

[e]dBA: A-weighted decibels.

[f]FDA: Food and Drug Administration.

[g]FAA: Federal Aviation Administration.

Figure 2. Scatter plot showing the relationship between maximum pulse flow setting and pulse-dose bolus volume (mL) in portable oxygen concentrators identified in this scoping review of ambulatory oxygen therapy devices. Data were extracted from publicly available scientific and gray literature published between 2004 and 2025. Only products with publicly available data on pulse-dose bolus volume at 20 breaths per minute at the maximum pulse flow setting are shown.

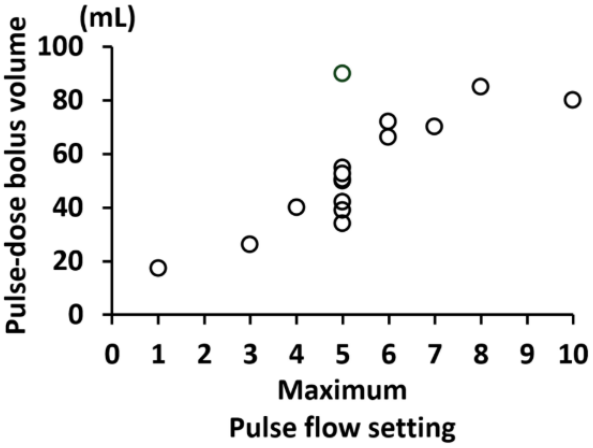

Figure 3. Scatter plot showing the relationship between product weight and pulse-dose bolus volume (mL) at the maximum pulse flow setting in portable oxygen concentrators identified in this scoping review of ambulatory oxygen therapy devices. Data were extracted from Lpublicly available scientific and gray literature published between 2004 and 2025. Only products with publicly available data on pulse-dose bolus volume at 20 breaths per minute at the maximum pulse flow setting are shown.

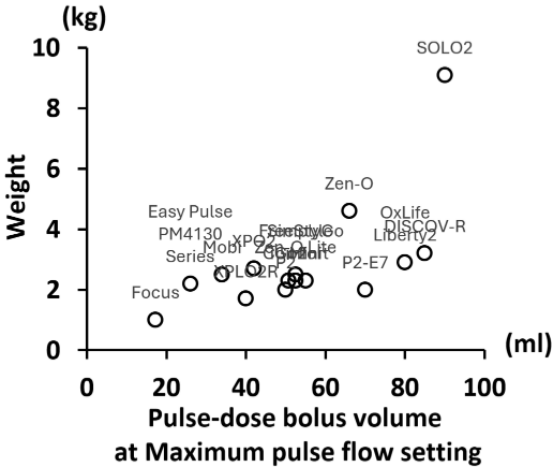

XSL•FO
RenderX

Figure 4. Continuous operating time at different continuous flow settings for (A) portable oxygen concentrators, (B) oxygen cylinders, and (C) liquid oxygen devices identified in this scoping review of ambulatory oxygen therapy devices. Data were extracted from publicly available product manuals and literature published between 2004 and 2025. Only products with publicly available data on continuous operating time for each continuous flow setting are shown.

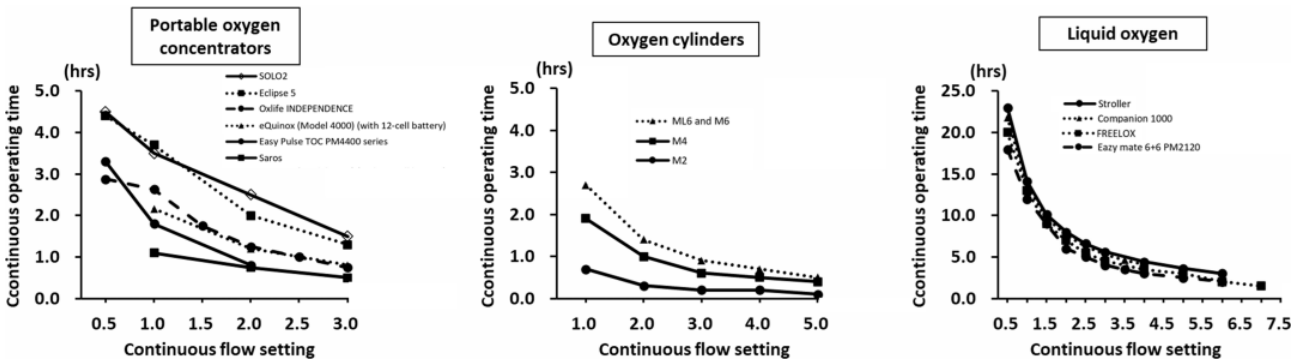

Table 5. Summary of identified oxygen cylinders included in this scoping review of ambulatory oxygen therapy devices. Data were extracted from patents, manuals, and technological reports published between 2004 and 2025 and include cylinder size, weight, oxygen capacity, nominal pressure, maximum operating time, and transport methods (1 bar=100 kPa).

Type	Size (cm)	Weight (kg)	Content (L)	Nominal pressure (bar)	Maximum continuous operating time	Transport method
M2	14.9 × 6.4	0.34	45	153[a]	0.3 h at 2 L/min[a]	Not reported
M4	22.1 × 8.1	1.0	113	153[a]	1 h at 2 L/min[a]	Not reported
ML6	20.0 × 11.1	1.3	170	139[a]	1.4 h at 2 L/min[a]	Shoulder bag
M6	30.0 × 8.2	1.3	170	153[a]	1.4 h at 2 L/min[b]	Shoulder bag
M7	23.1 × 11.1	1.5	198	139[a]	1.7 h at 2 L/min[a]	Shoulder bag
IOSVR	34.2 × 8.5	1.7	612	300[§]	8 h at 2 L/min with a standard conserver[c]	Handheld or bag
M9	27.7 × 11.1	1.7	255	139[a]	2.1 h at 2 L/min[a]	Shoulder bag
D	53.5 × 10.2	3.9	340	140[a]	3.6 h at 2 L/min[a]	Shoulder bag
CH/C	51.5 × 11.7	4.2	470	163 or 250[a]	Not reported	Medical trolleys or carry bags
E	86.5 × 10.2	6.5	680	137[a]	5 h at 2 L/min[c]	Hand-drawn cylinder cart

[a]Data obtained from nonmanufacturer websites.

[b]Data reported in the technical reports.

[c]Data reported in scientific reports.

LOX had a longer continuous operating time compared to POCs and cylinders, with a mean maximum continuous operating time of approximately 7.4 hours even at 2 L/min of continuous flow setting (Figure 4). There was no available information on the pulse-dose bolus volume of pulse flow setting in LOX.

XSL•FO
RenderX

Table 6. Summary of identified liquid oxygen systems included in this scoping review of ambulatory oxygen therapy devices. Data were extracted from manuals, patents, and technical reports published between 2004 and 2025 and include weight, oxygen capacity, flow settings, maximum operating duration, and regulatory approval.

Variables	Statistic	Notes
Device volume (cm³), mean (range)	7889.2 (3650.2-13879.1)	n=6
Weight (kg), mean (range)	3.00 (1.6-3.7)	n=6
Transport method, n		Some devices fall into multiple categories
Backpack	3	
Cart	3	
Carry bag	1	
Belt pack	3	
Handle	1	
Gaseous oxygen capacity (L), mean (range)	699.4 (275.0-1058.0)	N/A[a]
PF[b] or CF[c] setting, n		
Only PF	3	N/A
Both CF and PF	3	N/A
Maximum number of PF setting (n), mean (range)	5 (4-6)	n=3
Maximum flow rates in CF setting (L/min), mean (range)	4.3 (0.8-7)	n=6
Maximum continuous operating time in CF setting (h), mean (range)	7.4 (6.0-9.0)	n=5 (only those at 2 L/min of CF setting)
FDA[d], n		
Approved	5	N/A
Not approved	1	N/A

[a]N/A: not applicable.
[b]PF, pulse flow.
[c]CF, continuous flow.
[d]FDA, Food and Drug Administration.

Developments and New Technologies

Improving Portability

A patent was published in 2022 regarding the use of the vacuum swing adsorption method instead of the conventional PSA method in POCs to generate oxygen more efficiently [30]. By replacing the compressor in the PSA method with a fan or other air-moving device, air at atmospheric pressure instead of high pressure is passed through the zeolite adsorbent to adsorb nitrogen, and then the pressure is reduced to a vacuum level, which improves the efficiency of nitrogen desorption, resulting in improved oxygen generation rates [25]. This low-pressure operation makes it possible to use a small vacuum pump without a compressor in PSA, thus reducing the weight of the device and power consumption. Specifically, the POC in this patent had a device weight of 1.3 kg and a maximum oxygen flow rate of 1.5 L/min in the continuous flow setting. In addition, the POC in this patent uses LiLSX, which has recently been developed with improved selectivity and adsorption capacity for nitrogen [25,30,31]. A technology has been incorporated to maximize the efficiency of oxygen generation by monitoring the work of the zeolite adsorbent and the flow rate between the inlet and outlet of the adsorption column in real time to accurately detect the point when the adsorbent is saturated with nitrogen during the adsorption process (breakthrough point) and switch the cycle just before the saturated nitrogen is mixed with the concentrated oxygen. This allows oxygen to be separated without waste and maintains high oxygen purity while reducing the size of the device [30]. Another patent was published in 2023 for the design of a nasal-wearable POC entitled “Device and Method of Generating an Enriched Gas Within a Nasal Vestibule” [32]. However, few details were provided on the technical solutions used to achieve this small POC.

Improving or Optimizing the Oxygen Generation Process

A 2-bed rapid pressure swing adsorption (RPSA) system as an oxygen generation process was described in a scientific paper in 2021 [33]. The adsorption and desorption cycles performed in multiple adsorption columns in a conventional PSA system take a certain amount of time, which limits the amount of oxygen delivery. In this study, the cycle time was significantly reduced while maintaining a high oxygen concentration by optimizing the adsorbent and parameters such as pressure and cycle time related to adsorption. As a result, high oxygen

XSL•FO
RenderX

productivity was observed despite the much lower ratio of adsorbent volume required (bed size factor) [33]. Other scientific papers have shown that sensitivity analysis modelling of PSA systems has led to optimization of operating conditions such as adsorption pressure, cycle time and adsorption bed size, leading to their applicability as oxygen supply infrastructure in small medical facilities and in high-humidity environments [34,35].

Remote Control

A scientific report for SpO_2-targeted automatic titration of oxygen delivery during activities of daily living (ADL) testing based on the user's SpO_2 was identified [36]. This double-blinded randomized crossover trial evaluated the effect of SpO_2-targeted automatic titration during ADL testing in 31 patients with chronic obstructive pulmonary disease (COPD) on long-term oxygen therapy. In this trial, an oxygen cylinder was connected to the closed-loop device (O2matic, Herlev, Denmark), and this device was placed on a rollator during ADL testing. In the active arm, automatic titration of oxygen flow rate (0-8 L/min) was set to aim at keeping an SpO_2 target range of 90%-94% and those adjustments were done every second based on the average SpO_2 for the last 15 seconds. In the control arm, the flow rate was kept fixed according to each patient's medical prescription. As a result, automatic titration increased flow rates by triple and reduced ADL completion time, improved dyspnea, and reduced the number of events of severe desaturation compared to the control arm. Another technology for SpO_2-targeted adjustment of oxygen delivery based on the user's SpO_2 was identified [37]. An "intelligent oxygen concentrator" for patients with COPD has been developed as a system that measures physical activity levels and automatically adjusts oxygen delivery according to machine learning (eg, decision tree, probabilistic neuronal network, logistic regression) [38]. This machine learning model was trained using patient activity data to discriminate the patient's activity level (sedentary, light activity, and moderate activity). In a pilot study of 5 patients with COPD with long-term oxygen therapy, machine learning was used to automatically adjust the flow setting of pulse flow to match each patient's physical activity level. As a result, the cumulative time of SpO_2 below 90% when walking the circuit course was significantly reduced compared to manual changes [39].

Other Developments and New Technologies

A patent was identified for a new oxygen concentration control technique that delivers oxygen at low to moderate concentrations (eg, 30%-50%) by adjusting the pressure during the PSA cycle [40]. It was mentioned that the importance of this low-concentration oxygen delivery in clinical practice is unclear, but this patent could lead to reduced energy consumption of the POC, which has implications for extending battery life. In addition, a number of technologies for oxygen delivery control systems using sensors have been reported in recent years. One patent described sensors that use light to detect the timing of inhalation and valves that open and close electromagnetically at that timing to inject oxygen for a very short time (microbursts) [41]. This allows oxygen to be delivered with less waste of gas. In addition, a new delivery method was reported in which oxygen is generated and stored in advance and released in immediate response to inhalation [42].

Cost

There were no available records of the cost of POCs officially disclosed by POC companies. Two identified scientific papers referred to POC prices of approximately US $2000 and US $3000 as retail costs to end users [43,44]. There were also no records of specific costs for LOX and oxygen cylinders. However, an identified article reported that LOX was approximately 4 times as expensive as standard oxygen therapy using POCs or portable cylinders [45]. Another identified article indicated that LOX systems entail higher service costs due to the need for regular home refills [46].

Gap Map

The gap map (Figure 5) showed that only a limited range of oxygen cylinders met patient needs for high flow rates (>3 L/min) and light weight (<2.5 kg). Although current POCs did not meet this need, a one-hand carry device capable of delivering 3 L/min was under clinical development. In all device types, few products met patient needs for continuous operating time (>5-6 hours). For auto-titration, SpO_2-targeted automatic titration was available for POCs and cylinders as an interface, and the development of automatic titration capabilities embedded into the device itself has been reported only in POCs.

Figure 5. Gap map of patient needs and device availability in ambulatory oxygen therapy. The figure illustrates the extent to which current devices address 3 major patient needs: green indicates needs are met ("available"), yellow indicates limited devices meet patient needs ("limited availability"), orange indicates no current devices but technologies are under development ("in development"), and red indicates no available devices and no technologies under development ("not available").

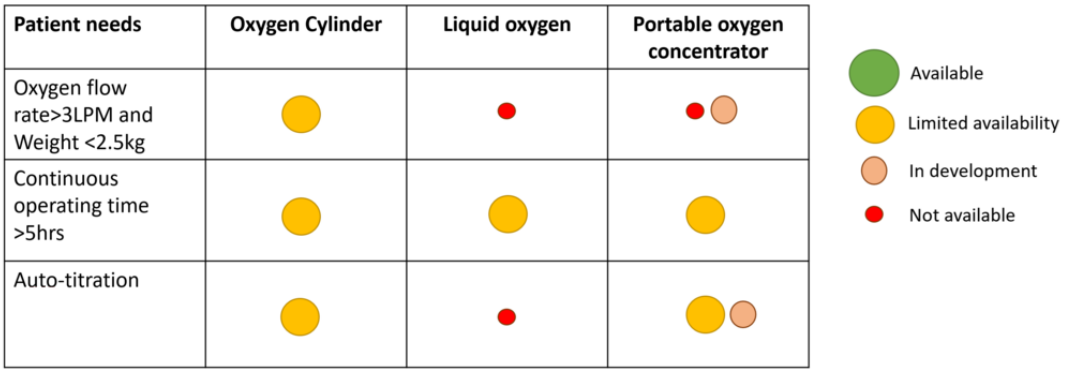

XSL•FO
RenderX

Discussion

Overview

This scoping review aimed to identify the range of portable oxygen devices, their oxygen generation technologies, performance characteristics, innovations in the design of ambulatory oxygen therapy devices, and cost. We identified 33 POCs, 10 oxygen cylinders, and 6 portable LOX systems. There was a trade-off between the portability (weight) and oxygen delivery capacity of POCs. The mean maximum continuous operating time of POCs was 3.8 hours. PSA with zeolite was the most common oxygen generation technology in POCs. Two POC prototypes were identified that had better oxygen delivery capacity despite being the same size as POCs on the market. LOX was the most portable and had the best continuous operating time among the devices. In terms of development and innovation, the downsizing of POCs using nanozeolite as the adsorbent, improving flow rate using RPSA and SpO_2-targeted automated oxygen titration based on to patient's SpO_2 and physical activity were identified. Costs were rarely disclosed by manufacturers, and published data indicated that LOX is substantially more expensive than POCs or cylinders.

This study showed that high-flow oxygen delivery of 3 L/min in the continuous flow setting in POCs is only possible from commercially available POCs weighing more than 7.2 kg (eg eQuinox, Table S3 in Multimedia Appendix 4), which need to be transported on a cart. This may be a burden for patients requiring high flow rates. Furthermore, the short continuous operating time of the POCs identified in this study may also limit daily activities and community engagement. Patients have previously identified a desirable operating time for portable oxygen devices of 5-6 hours [9]. In the identified POC products, the mean maximum continuous operating time was up to 3.8 hours in the pulse flow setting. As higher flow rates shorten the continuous operating time, it is likely that fewer devices will meet patients' needs. Although an additional battery can be used to double the operating time (eg, Simply Go), carrying the additional battery may be a burden for the patient. Furthermore, it should be noted that the numerical setting of a pulse flow does not correspond to the continuous flow rate and that the bolus volume varies between products even within the same pulse flow rate setting [47].

Among the devices, LOX had the best portability and continuous operating time, but the complicated maintenance was an obstacle. Previous studies have recognized LOX as a device with both better portability and continuous operating time [8], and have reported that LOX use is associated with better quality of life and physical activity compared to POCs and cylinders [48,49]. On the other hand, LOX requires refilling of liquid oxygen by a supplier at least 2-3 times a month, which imposes a high cost on the supplier and user [4,45]. As a result, the number of suppliers servicing LOX has been declining in the United States [8].

It was suggested that the trade-off between portability and oxygen delivery capacity is improving based on the latest products and prototypes identified. However, even in the latest POC products (DISCOV-R released in 2023 and OxLife Liberty2 released in 2024), the maximum oxygen flow rate of the continuous flow setting was still 2 L/min (Table S3 in Multimedia Appendix 4). JUNO, which is designed to be carried with one hand, is under clinical development and has a continuous flow setting ranging from 1-3 L/min with a concentration of 91% [26], and may be an innovative device, as the PSA Prototype was published in 2015 and the 4-SLPM in 2018 [50,51]. In terms of technological development, the nanozeolite (LilSX) had been of interest because of its use in 4-SLPM prototypes, patents, and basic research [30,52]. Although nanozeolites are characterized by higher nitrogen adsorption capacity, they are more expensive than conventional zeolites and have challenges such as long-term structural stability and degradation [53,54]. Although RPSA systems can improve oxygen productivity by shortening adsorption–desorption cycles, their application to POCs remains challenging because no clinical application studies have yet been reported [52]. Several recent clinical trials and patents described systems capable of adjusting flow rates based on the user's SpO_2 and physical activity. In a clinical trial with a small sample size, automated titration improved functional capacity in ADL, dyspnea, and the number of severe hypoxemia events [37]. However, these technologies have barriers, including sensor reliability or inaccuracy even during motion, requiring further development and clinical trials [37,55]. Identified patents in this study should be carefully considered because approximately half of the patents in general are not commercialized and launched on the market [56].

Limitations

Limitations of this scoping review include the following. First, it is difficult to compare performance such as pulse-dose bolus volumes and continuous operating time between POCs because they depend on the number of breaths and device settings [57]. Second, this scoping review included only information published in English, which may have underestimated information on devices developed and used in non–English-speaking countries. Third, the pulse-dose bolus volume of POC at maximum pulse flow setting (as much as possible at a respiratory rate of 20 breaths per minute) extracted in this study does not exactly indicate the product's oxygen delivery capacity. POC products vary in their algorithms for converting pulse-bolus volume and flow rate in response to changes in respiratory rate [57]. Fourth, available evidence regarding product performance relied almost on manufacturer-reported specifications, and no available evidence on device lifespan, failure rates, or long-term durability was available. There was no information available on the costs of POC products disclosed by manufacturers in this study. Lastly, the geographical distribution of the searched literature was biased toward the United States, as some gray literature was retrieved using domains such as ".gov," ".mil," and ".org" in Google Advanced Search.

Conclusions

This scoping review is the first study to integrate medical, engineering, and gray literature evidence on ambulatory oxygen devices and map current ambulatory oxygen devices and their development status. Although prior literature has narratively described products and technologies for ambulatory oxygen

XSL•FO
RenderX

therapy, no previous research has systematically investigated these products and new technologies. We identified 33 POCs, 10 oxygen cylinders, and 6 LOX systems. We showed the performance limitations of current devices and gaps in technological development in ambulatory oxygen therapy and suggested directions for future research and development. Specifically, POCs available to consumers may not meet patients' needs in terms of oxygen flow rate, portability, and operating time. LOX offered the best performance in terms of operating time and portability among the devices but is restricted by high costs and declining availability. Although POCs are the most widely developed devices, technological innovation to achieve high oxygen flow rates, better portability, and longer continuous operating time has been limited since the POC prototype published in 2015. Collaboration among device developers, researchers, health professionals, and patients is urgently required to develop new lightweight devices with greater oxygen delivery capacity. It is essential to incorporate consumer input from the early stages of design and testing to ensure that future portable oxygen devices meet patients' needs.

Acknowledgments

Generative artificial intelligence (ChatGPT) was used solely to assist in summarizing complex patent descriptions during the data extraction process. All outputs were reviewed and verified by the authors to ensure accuracy and to avoid hallucinations.

Funding

The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available in the Open Science Framework Storage from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: SK (lead), MH (equal), AEH (equal), ME (supporting), JAK (supporting)
Data curation: SK (lead), MH (equal), LR (supporting), AEH (supporting)
Formal analysis: SK (lead), MH (equal), LR (supporting), AEH (supporting)
Funding acquisition: AEH (lead), SK (equal)
Investigation: SK (lead), MH (equal), AEH (equal)
Methodology: SK (lead), MH (equal), AEH (equal)
Project administration: AEH (lead), SK (equal), MH (equal), ME (supporting), JAK (supporting)
Resources: SK (lead), MH (equal), AEH (equal)
Supervision: SK (lead), MH (equal), AEH (equal)
Validation: SK (lead), MH (equal), AEH (equal)
Visualization: SK (lead), MH (equal), AEH (equal)
Writing – original draft: SK (lead), AEH (equal), MH (supporting), ME (supporting), JAK (supporting)
Writing – review & editing: SK (lead), AEH (equal), MH (supporting), ME (supporting), JAK (supporting)

Conflicts of Interest

Unrelated to this work, 2 authors (JAK and ME) have relationships with Inogen and ResMed, respectively, which are manufacturers of portable oxygen concentrators (POCs). JAK reports receiving personal fees for Inogen to serve as a consultant on POCs, and ME reports receiving a research grant from ResMed. JAK also has received researchs grant from the National Institutes of Health, Patient-Centered Outcomes Research Institute, American Lung Association, COPD Foundation, and BioVie, and personal fees from AstraZeneca, DynaMed/EBSCO, Genentech, MedImmune, and Verona Pharmaceuticals. ME declares no conflicts of interest related to this work. Unrelated to this work, ME also has received personal fees from AstraZeneca, Boehringer Ingelheim, Novartis, and Roche. MH and AEH report receiving a grant from the National Health and Medical Research Council–Australia (GNT1139953). MH and AH report nonfinancial support from BOC Australia and Air Liquide Healthcare related to their research on ambulatory oxygen therapy for pulmonary fibrosis (PFOX trial).

Multimedia Appendix 1
PRISMA-ScR-Fillable-Checklist_revised.
[DOCX File , 86 KB - jmir_v28i1e81077_app1.docx]

Multimedia Appendix 2
PRISMA 2020 for Abstracts Checklist.
[DOCX File , 265 KB - jmir_v28i1e81077_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist.
[DOCX File , 21 KB - jmir_v28i1e81077_app3.docx]

Multimedia Appendix 4
Additional material.
[DOCX File , 106 KB - jmir_v28i1e81077_app4.docx]

Multimedia Appendix 5
The performance characteristics of each identified products.
[XLSX File (Microsoft Excel File), 25 KB - jmir_v28i1e81077_app5.xlsx]

References

1. Jacobs SS, Krishnan JA, Lederer DJ, Ghazipura M, Hossain T, Tan AM. Home oxygen therapy for adults with chronic lung disease: an official American Thoracic Society clinical practice guideline. Am J Respir Crit Care Med 2020;202(10):e121-e141 [FREE Full text] [doi: 10.1164/rccm.202009-3608ST] [Medline: 33185464]
2. Lacasse Y, Lecours R, Pelletier C, Bégin R, Maltais F. Randomised trial of ambulatory oxygen in oxygen-dependent COPD. Eur Respir J 2005;25(6):1032-1038 [FREE Full text] [doi: 10.1183/09031936.05.00113504] [Medline: 15929958]
3. Jarosch I, Gloeckl R, Damm E, Schwedhelm A, Buhrow D, Jerrentrup A. Short-term effects of supplemental oxygen on 6-min walk test outcomes in patients with COPD: a randomized, placebo-controlled, single-blind, crossover trial. Chest 2017;151(4):795-803. [doi: 10.1016/j.chest.2016.11.044] [Medline: 27940278]
4. Hardavella G, Karampinis I, Frille A, Sreter K, Rousalova I. Oxygen devices and delivery systems. Breathe (Sheff) 2019;15(3):e108-e116 [FREE Full text] [doi: 10.1183/20734735.0204-2019] [Medline: 31777573]
5. Chen J, Katz I, Pichelin M, Zhu K, Caillibotte G, Noga M. Comparison of pulsed versus continuous oxygen delivery using realistic adult nasal airway replicas. COPD 2017;Volume 12:2559-2571. [doi: 10.2147/copd.s141976]
6. Melani AS, Sestini P, Rottoli P. Home oxygen therapy: re-thinking the role of devices. Expert Rev Clin Pharmacol 2018;11(3):279-289. [doi: 10.1080/17512433.2018.1421457] [Medline: 29272974]
7. Tikellis G, Hoffman M, Mellerick C, Burge AT, Holland AE. Barriers to and facilitators of the use of oxygen therapy in people living with an interstitial lung disease: a systematic review of qualitative evidence. Eur Respir Rev 2023;32(169):230066 [FREE Full text] [doi: 10.1183/16000617.0066-2023] [Medline: 37611946]
8. Jacobs SS, Lederer DJ, Garvey CM, Hernandez C, Lindell KO, McLaughlin S. Optimizing home oxygen therapy: an official American Thoracic Society workshop report. Ann Am Thorac Soc 2018;15(12):1369-1381. [doi: 10.1513/AnnalsATS.201809-627WS] [Medline: 30499721]
9. Jacobs SS, Lindell KO, Collins EG, Garvey CM, Hernandez C, McLaughlin S. Patient perceptions of the adequacy of supplemental oxygen therapy: results of the American Thoracic Society Nursing Assembly Oxygen Working Group Survey. Ann Am Thorac Soc 2018;15(1):24-32. [doi: 10.1513/AnnalsATS.201703-209OC] [Medline: 29048941]
10. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
11. Johannson KA, Pendharkar SR, Mathison K, Fell CD, Guenette JA, Kalluri M. Supplemental oxygen in interstitial lung disease: an art in need of science. Ann Am Thorac Soc 2017;14(9):1373-1377. [doi: 10.1513/AnnalsATS.201702-137OI] [Medline: 28644693]
12. Ejiofor SI, Bayliss S, Gassamma A, Turner AM. Ambulatory oxygen for exercise-induced desaturation and dyspnea in chronic obstructive pulmonary disease (COPD): systematic review and meta-analysis. Chronic Obstr Pulm Dis 2016;3(1):419-434 [FREE Full text] [doi: 10.15326/jcopdf.3.1.2015.0146] [Medline: 28848863]
13. Long-Term Oxygen Treatment Trial Research Group, Albert RK, Au DH, Blackford AL, Casaburi R, Cooper JA. A randomized trial of long-term oxygen for COPD with moderate desaturation. N Engl J Med 2016;375(17):1617-1627 [FREE Full text] [doi: 10.1056/NEJMoa1604344] [Medline: 27783918]
14. Sanchez-Morillo D, Muñoz-Zara P, Lara-Doña A, Leon-Jimenez A. Automated home oxygen delivery for patients with COPD and respiratory failure: a new approach. Sensors (Basel) 2020;20(4):1178 [FREE Full text] [doi: 10.3390/s20041178] [Medline: 32093418]
15. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
16. JBI Manual for Evidence Synthesis.: JBI; 2024. URL: https://synthesismanual.jbi.global/ [accessed 2025-05-12]
17. Landerdahl Stridsberg S, Richardson MX, Redekop K, Ehn M, Wamala Andersson S. Gray literature in evaluating effectiveness in digital health and health and welfare technology: a source worth considering. J Med Internet Res 2022;24(3):e29307 [FREE Full text] [doi: 10.2196/29307] [Medline: 35319479]

XSL•FO
RenderX

18. Godin K, Stapleton J, Kirkpatrick SI, Hanning RM, Leatherdale ST. Applying systematic review search methods to the grey literature: a case study examining guidelines for school-based breakfast programs in Canada. Syst Rev 2015;4:138 [FREE Full text] [doi: 10.1186/s13643-015-0125-0] [Medline: 26494010]
19. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala A, Moher D, Page M. PRISMA-S: an extension to the PRISMA Statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
20. covidence. URL: https://www.covidence.org/ [accessed 2025-12-28]
21. WHO technical specifications for oxygen concentrators. World Health Organization. URL: https://www.who.int/publications/i/item/9789241509886 [accessed 2025-01-09]
22. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc 2015;13(3):141-146. [doi: 10.1097/XEB.0000000000000050] [Medline: 26134548]
23. Blakeman TC, Rodriquez D, Britton TJ, Johannigman JA, Petro MC, Branson RD. Evaluation of oxygen concentrators and chemical oxygen generators at altitude and temperature extremes. Mil Med 2016;181(5 Suppl):160-168. [doi: 10.7205/MILMED-D-15-00130] [Medline: 27168568]
24. Yadav VK, Choudhary N, Inwati GK, Rai A, Singh B, Solanki B. Recent trends in the nanozeolites-based oxygen concentrators and their application in respiratory disorders. Front Med (Lausanne) 2023;10:1147373 [FREE Full text] [doi: 10.3389/fmed.2023.1147373] [Medline: 37181347]
25. Ackley MW. Medical oxygen concentrators: a review of progress in air separation technology. Adsorption 2019;25(8):1437-1474. [doi: 10.1007/s10450-019-00155-w]
26. Tankless Oxygen Live Life Without Limits.: Roam Technologies URL: https://www.roamtech.ai/ [accessed 2025-12-24]
27. Shan-Shan WE, Sisi Z, Jay RF, Jeremy TK, Natalie EH, Gavin DM. Oxygen concentrator. WIPO.: Roam Technologies Pty Ltd; 2024. URL: https://patentscope2.wipo.int/search/en/detail.jsf?docId=US451301825 [accessed 2025-12-24]
28. Bruton A, Sinclair I, Arnold E, Hepples W, Kay F, Maisey G. The design and development of a new light-weight portable oxygen system. J Med Devices 2012;6(3):031007. [doi: 10.1115/1.4007180]
29. Palwai A, Skowronski M, Coreno A, Drummond C, McFadden ER. Critical comparisons of the clinical performance of oxygen-conserving devices. Am J Respir Crit Care Med 2010;181(10):1061-1071 [FREE Full text] [doi: 10.1164/rccm.200910-1638OC] [Medline: 20133925]
30. Jagger TW, Van Brunt NP, Kivisto JA, Lonnes PB. Removable cartridge for oxygen concentrator. WIPO. 2022. URL: https://patentscope2.wipo.int/search/en/detail.jsf?docId=US105472432&_cid=JP2-MIE7CN-34520-1 [accessed 2025-12-24]
31. Arora A, Hasan MMF. Flexible oxygen concentrators for medical applications. Sci Rep 2021;11(1):14317 [FREE Full text] [doi: 10.1038/s41598-021-93796-3] [Medline: 34253838]
32. Assani KD. Device and method of generating an enriched gas within a nasal vestibule. WIPO.: Worldwide Health Innovations LLC; 2023. URL: https://patentscope2.wipo.int/search/en/detail.jsf?docId=US321199054&_cid=JP1-MHZW3X-13119-1 [accessed 2025-12-24]
33. Qadir S, Li D, Gu Y, Yuan Z, Zhao Y, Wang S. Experimental and numerical analysis on the enhanced separation performance of a medical oxygen concentrator through two-bed rapid pressure swing adsorption. Ind Eng Chem Res 2021;60(16):5903-5913. [doi: 10.1021/acs.iecr.1c00420.s001]
34. Benkirane L, Samid A, Chafik T. Small-scale medical oxygen production unit using PSA technology: modeling and sensitivity analysis. J Med Eng Technol 2023;47(6):321-335. [doi: 10.1080/03091902.2024.2331693] [Medline: 38626001]
35. Prayoga GA, Husni E, Damar Jaya S. Design of an embedded controller and optimal algorithm of PSA for a novel medical oxygen concentrator. Int J Electr Eng Inform 2023;15(2):220-239 [FREE Full text]
36. Kofod LM, Hansen EF, Brocki BC, Kristensen MT, Roberts NB, Westerdahl E. Optimised oxygenation improves functional capacity during daily activities in patients with COPD on long-term oxygen therapy: a randomised crossover trial. Thorax 2025;80(11):803-809 [FREE Full text] [doi: 10.1136/thorax-2024-221883] [Medline: 40473413]
37. Prayoga GA. Design and implementation system of mobile oxygen concentrator and telemedicine for comprehensive treatment of SpO2. Int J Adv Technol Eng Explor 2023;10(106). [doi: 10.19101/ijatee.2023.10101547]
38. Lara-Doña A, Sánchez-Morillo D, Pérez-Morales M, Fernández-Granero M, León-Jiménez A. A prototype of intelligent portable oxygen concentrator for patients with COPD under oxygen therapy. 2019 Presented at: MEDICON 2019, IFMBE Proceedings; September 26-28, 2019; Porto. [doi: 10.1007/978-3-030-31635-8_55]
39. Sanchez-Morillo D, Muñoz-Zara P, Lara-Doña A, Leon-Jimenez A. Automated home oxygen delivery for patients with COPD and respiratory failure: a new approach. Sensors (Basel) 2020;20(4):1178 [FREE Full text] [doi: 10.3390/s20041178] [Medline: 32093418]
40. Warren J. Efficient enriched oxygen airflow systems and methods. Wearair Ventures Inc.: WIPO; 2022. URL: https://patentscope2.wipo.int/search/en/detail.jsf?docId=US390988244 [accessed 2025-11-15]
41. James JRJ, Zuzana EM, inventor SHL. Pulsed oxygen system and process. Seabeck Holdings LLC.: JUSTIA Patents; 2022. URL: https://patents.justia.com/patent/20220347418 [accessed 2025-11-15]
42. Wang Q, Liu Y. Integrated oxygen supply device. Telesair Inc. 2022. URL: https://patents.google.com/patent/US11517702B1/en?oq=+US11517702+-+Integrated+oxygen+supply+device [accessed 2025-11-15]

XSL•FO
RenderX

43. Mapel DW, Robinson SB, Lydick E. A comparison of health-care costs in patients with chronic obstructive pulmonary disease using lightweight portable oxygen systems versus traditional compressed-oxygen systems. Respir Care 2008;53(9):1169-1175. [Medline: 18718034]
44. Casaburi R, Hess M, Porszasz J, Clark W, Diesem R, Tal-Singer R, et al. Evaluation of over-the-counter portable oxygen concentrators utilizing a metabolic simulator. Respir Care 2023;68(4):445-451 [FREE Full text] [doi: 10.4187/respcare.10495] [Medline: 36400446]
45. Law S. Liquid Oxygen Therapy at Home. 2005. URL: https://database.inahta.org/article/4059 [accessed 2025-03-08]
46. Dunne PJ. The clinical impact of new long-term oxygen therapy technology. Respiratory care 2009;54(8):1100-1111 [FREE Full text] [Medline: 19650950]
47. Dunne PJ. Long-term oxygen therapy (LTOT) revisited: in defense of non-delivery LTOT technology. Rev Port Pneumol 2012;18(4):155-157 [FREE Full text] [doi: 10.1016/j.rppneu.2012.02.011] [Medline: 22575635]
48. Andersson A, Ström K, Brodin H, Alton M, Boman G, Jakobsson P. Domiciliary liquid oxygen versus concentrator treatment in chronic hypoxaemia: a cost-utility analysis. Eur Respir J 1998;12(6):1284-1289 [FREE Full text] [doi: 10.1183/09031936.98.12061284] [Medline: 9877478]
49. Su C, Lee C, Chen H, Feng L, Lin H, Chiang L. Comparison of domiciliary oxygen using liquid oxygen and concentrator in northern Taiwan. J Formos Med Assoc 2014;113(1):23-32 [FREE Full text] [doi: 10.1016/j.jfma.2012.03.013] [Medline: 24445009]
50. Alptekin G. Low power medical oxygen concentrators for space missions. 2018 Presented at: 48th International Conference on Environmental Systems; July 8-12, 2018; Albuquerque.
51. Gilkey KM, Olson SL. Evaluation of the Oxygen Concentrator Prototypes: Pressure Swing Adsorption Prototype and Electrochemical Prototype. 2015. URL: https://ntrs.nasa.gov/api/citations/20150011038/downloads/20150011038.pdf [accessed 2025-12-24]
52. Qadir S, Li D, Gu Y, Yuan Z, Zhao Y, Wang S. Experimental and numerical analysis on the enhanced separation performance of a medical oxygen concentrator through two-bed rapid pressure swing adsorption. Ind Eng Chem Res 2021;60(16):5903-5913. [doi: 10.1021/acs.iecr.1c00420]
53. Pan M, Omar H, Rohani S. Application of nanosize zeolite molecular sieves for medical oxygen concentration. Nanomaterials (Basel) 2017;7(8):195 [FREE Full text] [doi: 10.3390/nano7080195] [Medline: 28757586]
54. Yadav VK, Choudhary N, Inwati GK, Rai A, Singh B, Solanki B. Recent trends in the nanozeolites-based oxygen concentrators and their application in respiratory disorders. Front Med (Lausanne) 2023;10:1147373 [FREE Full text] [doi: 10.3389/fmed.2023.1147373] [Medline: 37181347]
55. Mondal A, Dutta D, Chanda N, Mandal N, Mandal S. RESPIPulse: machine learning assisted sensory device for pulsed mode delivery of oxygen bolus using surface electromyography (sEMG) signals. Sens Actuators A Phys 2024;369:115121. [doi: 10.1016/j.sna.2024.115121]
56. Svensson R. The scientific output of a database on commercialized patents. IFN Working Paper No. 1349.: Research Institute of Industrial Economics; 2020. URL: https://www.ifn.se/media/unzdc0gw/wp1349.pdf [accessed 2025-01-09]
57. Chatburn RL, Williams TJ. Performance comparison of 4 portable oxygen concentrators. Respir Care 2010;55(4):433-442. [Medline: 20406511]

Abbreviations

ADL: activities of daily living
COPD: chronic obstructive pulmonary disease
IOSVR: Ultra Lightweight Cylinder Oxygen System
JBI: Joanna Briggs Institute
LiLSX: high lithium-exchanged X
LOX: liquid oxygen
MeSH: Medical Subject Headings
POC: portable oxygen concentrator
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Search extension
PRISMA-ScR: Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews
PSA: pressure swing adsorption
RPSA: rapid pressure swing adsorption
SpO_2: pulse oximeter oxygen saturation
VPC: vacuum pressure cycle
VPSA: vacuum pressure swing adsorption
WIPO: World Intellectual Property Organization

XSL•FO
RenderX

Edited by S Brini; submitted 22.Jul.2025; peer-reviewed by MW Hess, S Narayan; comments to author 24.Sep.2025; accepted 22.Dec.2025; published 27.Jan.2026.

Please cite as:
Kawachi S, Hoffman M, Romero L, Ekström M, Krishnan JA, Holland AE
Products, Performance, and Technological Development of Ambulatory Oxygen Therapy Devices: Scoping Review
J Med Internet Res 2026;28:e81077
URL: https://www.jmir.org/2026/1/e81077
doi:10.2196/81077
PMID:41429116

©Shohei Kawachi, Mariana Hoffman, Lorena Romero, Magnus Ekström, Jerry A Krishnan, Anne E Holland. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Development and Use of AI Chatbots for Health Behavior Change: Scoping Review

Lingyi Fu[1], MS; Ryan Burns[1], PhD; Yuhuan Xie[1], MS; Jincheng Shen[2], PhD; Shandian Zhe[3], PhD; Paul Estabrooks[1], PhD; Yang Bai[1], PhD

[1]Department of Health and Kinesiology, College of Health, University of Utah, Salt Lake City, UT, United States

[2]Department of Internal Medicine, School of Medicine, University of Utah, Salt Lake City, UT, United States

[3]Kahlert School of Computing, University of Utah, Salt Lake City, UT, United States

Corresponding Author:
Yang Bai, PhD
Department of Health and Kinesiology
College of Health
University of Utah
HPER-North
Room 204
Salt Lake City, UT, 84102
United States
Phone: 1 8015870482
Email: Yang.Bai@utah.edu

Abstract

Background: Artificial intelligence (AI) chatbots are technologies that facilitate human-computer interaction through communication in a natural language format. By increasing cost-effectiveness, interaction, autonomy, personalization, and support, mobile health interventions can benefit health behavior change and make it more natural and intuitive.

Objective: This study aimed to provide an up-to-date and practical overview of how text-based AI chatbots are designed, developed, and evaluated across 8 health behaviors, including their roles, theoretical foundations, health behavior change techniques, technology development workflow, and performance validation framework.

Methods: In accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) framework, relevant studies published before March 2024 were identified from 9 bibliographic databases (ie, PubMed, CINAHL, MEDLINE, Embase, Web of Science, Scopus, APA PsycINFO, IEEE Xplore, and ACM Digital Library). Two stages (ie, title and abstract screening followed by full-text screening) were conducted to screen the eligibility of the papers via Covidence software. Finally, we extracted the data via Microsoft Excel software and used a narrative approach, content analysis, and evidence map to synthesize the reported results.

Results: Our systematic search initially identified 10,508 publications, 43 of which met our inclusion criteria. AI chatbots primarily served 2 main roles: routine coach (27/43, 62.79%) and on-demand assistant (12/43, 27.91%), while 4 studies (4/43, 9.30%) integrated both roles. Frameworks like cognitive behavioral therapy (13/24, 54.17%) and behavior change techniques, such as goal setting, feedback and monitoring, and social support, guided the development of theory-driven AI chatbots. Noncode platforms (eg, Google Dialogflow and IBM Watson) integrated with social messaging platforms (eg, Facebook Messenger) were commonly used to develop AI chatbots (23/43, 53.49%). AI chatbots have been evaluated across 4 domains: technical performance (17/43, 39.53%), usability (17/43, 39.53%), engagement (37/43, 86.05%), and health behavior change (33/43, 76.74%). Evidence for health behavior changes remains exploratory but promising. Among 33 studies with 120 comparisons, 81.67% (98/120) showed positive outcomes, though only 35.83% (43/120) had moderate or larger effects (Hedges g or odds ratio or Cohen d>0.5). Most involved nonclinical (36/43, 83.72%) and adults (23/43, 53.49%), and a few were randomized controlled trials (14/43, 32.56%). Benefits were mainly seen in physical activity, smoking cessation, stress management, and diet, with limited evidence for other behaviors. Findings were inconsistent regarding the influence of long-term effects, intervention duration, modality, and engagement on health behavior change outcomes.

XSL•FO
RenderX

Conclusions: The exploratory synthesis provides a roadmap for developing and evaluating AI chatbots in health behavior change, highlighting the need for further research on cost, implementation outcomes, and underexplored behaviors such as sleep, weight management, sedentary behavior, and alcohol use.

(J Med Internet Res 2026;28:e79677) doi:10.2196/79677

KEYWORDS

conversational agent; machine learning; health behavior change; physical activity; diet; sleep

Introduction

Health-related behavior (hereafter referred to as "health behaviors") is essential for improving population health worldwide [1]. Engaging in health-promoting behaviors, such as having a healthy diet, getting adequate physical activity (PA) and sleep, and avoiding health-risk behaviors, such as smoking, can substantially reduce chronic disease and all-cause mortality risk [1,2] as well as benefit mental health [3]. Despite these benefits, unhealthy behaviors remain a significant public health concern and place a substantial burden on health care systems [4]. Health coaching is one suggested intervention for promoting health behavior changes [5,6]. Health coaching is defined as "the practice of health education and health promotion within a coaching context, to increase the well-being of individuals and to facilitate the achievement of their health-related goals" [7]. Most health coaching interventions are delivered by humans in various ways, such as face-to-face, telephonically, or via email [5]. Effective communication between coaches and users can support user-centered care and shared decision-making [8]. However, human health coaching interventions are limited in their ability to reach everyone in need of support because of a lack of coaching practitioners, resources in resource-limited communities, and barriers for individuals in accessing coaching support, such as low income [9]. From the implementer's perspective, human health coaching often lacks consistent data collection, continuous monitoring, scalability, and long-term sustainability [10]. Therefore, finding resource-efficient, cost-effective, and easily implementable strategies to promote health behavior change can be helpful to alleviate an already burdened health care system.

A chatbot is a computer program designed to respond to conversational or informational replies with verbal (audio-based chatbot) or written (text-based chatbot) messages from users [11]. This technology can be another type of resource for delivering health coaching, complemented by traditional human health coaching [9]. Chatbots can be developed with and without artificial intelligence (AI) algorithms. Most prior chatbots for health behavior change were developed without AI algorithms ("Non-AI chatbots") used in template-, rule-, or retrieval-based dialogue systems [12]. These chatbots responded to users by selecting from a predefined list, allowing for a high degree of researcher control but lacking the conversational flexibility and personalization typically offered by human coaches [9,13]. Recently, with advances in AI in the health care field, some chatbots have been developed using AI algorithms [14], such as reinforcement learning, deep neural networks, and random forest. These AI chatbots can communicate with users in natural language [15-20], offering personalized support, multimodal reasoning, and greater conversational flexibility and intuitiveness [21]. In particular, AI chatbots can support health behavior change in various ways. Research has shown that motivational interviewing (MI)–based AI chatbot tends to be perceived as more empathetic and trustworthy than the directed intervention, and it significantly raises the participants' self-efficacy to overcome barriers and positively impacts intrinsic motivation and PA levels [22]. Additionally, AI chatbots can also help alleviate stress by enhancing perceived supportiveness through the provision of emotional support [23]. Although the benefits are numerous, several concerns remain, including privacy and security [16]; limited empathy, affect, and emotional support [24]; low engagement; and challenges in monitoring intervention fidelity. Therefore, it is important to provide evidence to address these concerns and ensure the efficacy and safety of chatbot interventions.

Previous review studies have reported varying evidence on the application of AI chatbots for health behavior change. For example, a meta-analysis demonstrated strong efficacy of chatbot-based interventions in increasing physical activity, fruit and vegetable consumption, and sleep duration and quality. The analysis also showed that the effects varied by intervention duration, intervention modality (chatbot-only vs multicomponent interventions), and chatbot characteristics (text-based vs audio-based and AI-driven vs non–AI-driven chatbots) [12]. Other systematic reviews have summarized the outcome in addition to efficacy, including engagement, acceptability, satisfaction, and safety [25], as well as feasibility, usability, and intervention characteristics [10]. However, there remains a lack of research exploring key topics, such as the role of AI chatbots in behavior change, the health behavior change techniques (BCTs) adopted by AI chatbots, comprehensive technology frameworks for chatbot development, and frameworks for performance validation. Therefore, to address this gap and complement existing review studies [10,12,25], it is necessary to conduct a scoping review to provide a comprehensive overview of existing research for both scholars and practitioners in this field. This scoping review aimed to provide an up-to-date and practical examination of the design (ie, roles, theories, and health BCTs), development (technology workflow), and use (ie, performance validation tool) of text-based AI chatbots for 8 health behaviors, including PA, diet, sleep, weight management, sedentary, stress management, smoking cessation, and alcohol. In particular, 4 specific research questions were proposed based on the indications from the scoping reviews [26]:

- Question 1: What are the most commonly targeted health behaviors in text-based AI chatbots?

XSL•FO
RenderX

- Question 2: What roles, theoretical foundations, and BCTs are applied in text-based AI chatbots, supporting health behavior change interventions?
- Question 3: What technologies are used to develop text-based AI chatbots for health behavior change?
- Question 4: How to validate text-based AI chatbot performance in health behavior change?
  - Question 4.1: What measures are used to assess technical performance, usability, engagement, and cost?
  - Question 4.2: What are the health behavior change outcomes?

Methods

Protocol and Registration

The scoping review process was designed following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analysis extension for Scoping Reviews) framework [27]. The PRISMA-ScR checklist is reported in Multimedia Appendix 1. The study has been registered in the Open Science Framework [28], the most common platform for deposit protocols for scoping reviews.

Search Strategy

Search Resources

Relevant studies published before March 2024 were identified through 9 bibliographic databases, including 4 widely used health science databases (ie, PubMed, CINAHL, MEDLINE, and Embase), 2 multidisciplinary databases (ie, Web of Science and Scopus), 1 behavior and social science database (ie, APA PsycINFO), and 2 technology databases (ie, IEEE Xplore and ACM Digital Library). The last search date for each database is March 13, 2024.

Search Terms

Our search strategy incorporated terms related to both the intervention (AI chatbots) and the outcomes (health behaviors; Table S1 in Multimedia Appendix 2). Given the 6 pillars of lifestyle medicine [29], this study focused on 8 health behaviors, including PA, diet, sleep, weight management, sedentary behavior, stress management, smoking cessation, and alcohol. We identified synonyms for AI chatbots and various health behaviors by searching for relevant terms in dictionaries and references. We subsequently generated search syntaxes adapted to the specific requirements of each database. The example of search syntax for PubMed is as follows: "(((lifestyle*[tiab] OR tobac*[tiab] OR cigarette*[tiab] OR cigar* [tiab] OR vap*[tiab] OR smok*[tiab] OR nico*[tiab] OR sleep*[tiab] OR bedtime[tiab] OR nap[tiab] OR insomnia[tiab] OR physical activ*[tiab] OR sport*[tiab] OR exercise*[tiab] OR diet*[tiab] OR nutriti*[tiab] OR eating [tiab] OR food*[tiab] OR appetite*[tiab] OR *weight*[tiab] OR obes*[tiab] OR sedentar*[tiab] OR screen time [tiab] OR stress* [tiab]) AND (Chatbot*[tiab] OR chat-bot*[tiab] OR chat bot*[tiab] OR chat robot*[tiab] OR virtual robot*[tiab] OR voice-bot[tiab] OR social bot*[tiab] OR social robot*[tiab] OR infobot*[tiab] OR health bot*[tiab] OR smartbot*[tiab] OR conversational bot*[tiab] OR artificial intelligence chatbot*[tiab] OR Ai agent*[tiab] OR conversational agent*[tiab] OR dialogue agent*[tiab] OR dialog agent* [tiab] OR interactive agent*[tiab] OR virtual agent*[tiab] OR automated agent*[tiab] OR relational agent*[tiab] OR AI assist*[tiab] OR conversational assistant*[tiab] OR digital assist*[tiab] OR intelligent assist*[tiab] OR virtual assist*[tiab] OR smart assist*[tiab] OR voice assist*[tiab] OR speech assist*[tiab] OR virtual health assist*[tiab] OR dialogue agent*[tiab] OR dialog agent*[tiab] OR AI advisor*[tiab] OR virtual advisor*[tiab] OR animated advisor*[tiab] OR smart advisor*[tiab] OR AI avatar*[tiab] OR virtual avatar*[tiab] OR animated avatar*[tiab] OR smart avatar*[tiab] OR AI coach*[tiab] OR virtual coach*[tiab] OR smart coach*[tiab] OR animated coach*[tiab] OR artificial conversation entit*[tiab] OR Assistance technolog*[tiab] OR conversational AI[tiab] OR conversational interface*[tiab] OR conversational system*[tiab] OR Dialog system*[tiab] OR dialogue system*[tiab] OR natural language interface*[tiab] OR automated conversation[tiab] OR virtual conversation[tiab] OR chatGPT[tiab])) AND (eng[la])) NOT (Systematic review[pt] OR meta-analysis[pt] OR review[pt])." Finally, the metadata of the identified papers were imported into the Covidence platform to eliminate duplication and screening.

Study Eligibility Criteria

Table S2 in Multimedia Appendix 2 outlines the study eligibility criteria for the study and publication characteristics. The study characteristics were designed based on the PICOS framework, including population, intervention, comparison, outcome, and study type [30]. The publication characteristics included publication date, language, and publication status (eg, full online).

Study Selection

The selection process had 2 stages: title and abstract screening, followed by full-text screening. Both were conducted independently by 2 reviewers (LF and YX), with conflicts resolved by a third reviewer (YB). We used Cohen κ metric to evaluate interrater agreement [31]. The reviewers achieved substantial agreement in stage 1, with a κ measure of 0.83, which is greater than the cutoff point of 0.81 [31]. The screening process and calculation of Cohen κ value were performed via Covidence software.

Study Quality Assessment

Given the diversity of study designs, we used the Mixed Methods Appraisal Tool (MMAT) to assess methodological quality [32] (Table S3 in Multimedia Appendix 2 [16-20,23,24,33-68]). The MMAT is a 21-item checklist covering 5 research designs: qualitative, quantitative randomized controlled trials (RCTs), quantitative nonrandomized studies, quantitative descriptive studies, and mixed methods studies. Interrater reliability for the MMAT has been reported to range from moderate to perfect [69]. Two reviewers (LF and YX) independently evaluated each paper, and any disagreements were resolved through discussion. In general, all studies clearly stated their research questions, and the collected data were sufficient to address them.

Assigning an overall numerical score based on MMAT ratings is discouraged, as a single number cannot capture specific

XSL•FO
RenderX

methodological issues [32]. Therefore, we presented detailed ratings for each criterion. All eligible studies were included in this review, regardless of their MMAT ratings, as excluding studies solely on the basis of low methodological quality is not recommended [32,70].

Data Items

We designed the initial elements of the charting form based on the research questions and the mobile health evidence reporting and assessment checklist [71]. The items were primarily developed by 1 investigator (LF) and subsequently verified by another investigator (YB). The items were refined throughout the process, resulting in a final set of 22 elements (Table S4 in Multimedia Appendix 2).

Charting Process and Data Synthesis

Regarding data charting, 1 reviewer (LF) was primarily responsible for data extraction. This process involved 2 stages: the first stage focused on extraction, and the second stage on confirmation and supplementation. Two additional reviewers (Conrad Ma and Shreya Sanghvi) then independently validated the extracted data. Clear instructions for data validation were provided to these reviewers. Any discrepancies were resolved through discussion among the 3 reviewers. Microsoft Excel software was used for data charting processes. Finally, we used a narrative approach, content analysis, and evidence map to synthesize the reported results. In particular, we conducted a deductive coding process to map each chatbot function to the existing BCT Taxonomy (version 1) [72], enabling cross-study comparisons. For the AI chatbot validation framework in health behavior change, we applied a combined deductive-inductive approach to map the measures from each study onto 5 domains (ie, technical performance, health behavior change, usability, cost, and engagement), drawing from the digital health scorecard framework [73] and engagement framework [74].

Results

Study Selection

Figure 1 summarizes the process of selecting the studies via the Covidence software (Veritas Health Innovation). A total of 10,508 studies were returned after searching the databases. After duplication removal and title and abstract screening, 229 studies remained. In total, 40 studies remained after the full-text screening phase, with 189 studies removed for the following reasons: not involving an AI chatbot (n=98), not being a health behavior change intervention or implementation study (n=81), or lacking sufficient information (n=10). We included 3 additional studies after forward and backward reference checking, bringing the total to 43 included studies.

Figure 1. Study selection process. AI: artificial intelligence.

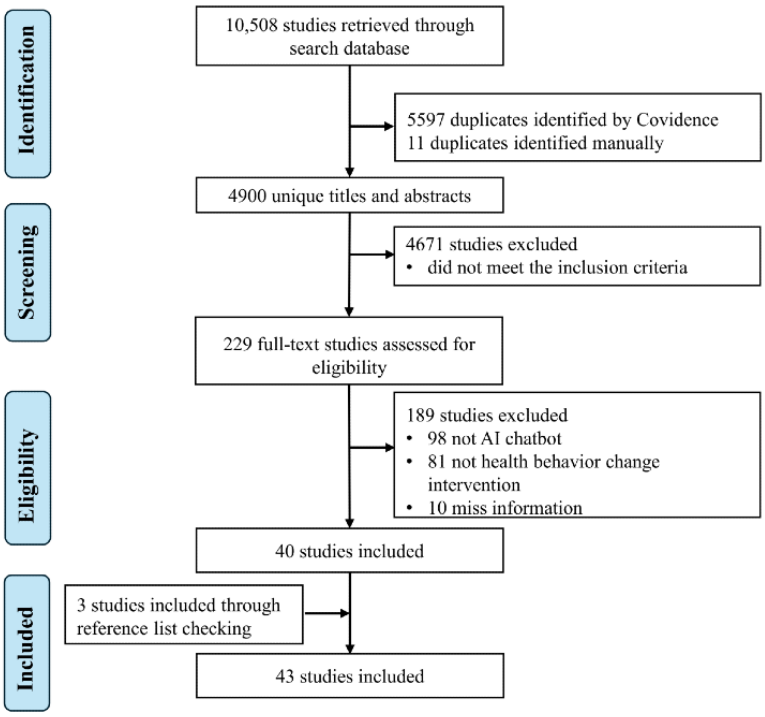

Study Overview

Overview

An overview of the studies included is presented in Table S5 in Multimedia Appendix 2 [16-20,23,24,33-68]. We further synthesized the findings across studies. Table S6 in Multimedia Appendix 2 [16-20,23,24,33-68] summarizes the publication characteristics of the studies included in the review. They were published between 2018 and 2024, with most published in 2023 (11/43, 25.58%). The studies were conducted across 15 countries or regions, with most of them conducted in Western countries (32/43, 74.42%), especially the United States (9/32, 28.13%), followed by the Netherlands (5/32, 15.63%), Italy (5/32, 15.63%), and the United Kingdom (5/32, 15.63%). Most research papers were published in journals (33/43, 76.74%), with the largest number published in *JMIR mHealth and uHealth* (6/33, 18.18%).

XSL•FO
RenderX

Methodological and Participant Characteristics

Table S7 in Multimedia Appendix 2 [16-20,23,24,33-68] summarizes the methodological and participant characteristics. Studies were classified based on the study types from Hong et al [32], including 14 RCTs, 15 quantitative nonrandomized trials (non-RCTs), 8 quantitative descriptive studies, 2 qualitative studies, and 4 mixed methods studies. The number of participants ranged from 7 [33] to 57,214 [34], with a primary focus on young and middle adulthood (aged 19-64 years; 23/43, 53.49%) based on age categories of Lindemann et al [75]. Most studies targeted nonclinical populations (36/43, 83.72%), such as individuals who are physically inactive, those with unhealthy diets, smokers, substance users, students (middle school, high school, and university), workers, and vulnerable groups, including low-income English- and Spanish-speaking individuals and residents of health professional shortage areas. In total, 7 studies targeted clinical populations (7/43, 16.28%), such as patients with colorectal cancer [35], patients with celiac [36], patients with cardiovascular problem [37], survivors of cancer [38], population with clinical eating disorder [39], as well as children [40] and youths [41] with obesity.

Intervention Characteristics

Table S7 in Multimedia Appendix 2 also provides a summary of the intervention characteristics. Regarding chatbot use, most studies used AI chatbots for 1-time use (10/43, 23.26%), followed by 8 weeks (7/43, 16.28%), 4 weeks (6/43, 13.95%), 12 weeks (5/43, 11.63%), 1 week (5/43, 11.63%), and 2 weeks (3/43, 6.98%). A small number of studies ranged from 3 days [42], 9 days [43], 3 weeks [44], 6 weeks [20], 24 weeks [19,45] to 48 weeks [46]. In terms of intervention modalities, most studies used a text-based chatbot alone (27/43, 62.79%; Table S5 [Study description] in Multimedia Appendix 2). Among them, some studies compared chatbots with other interventions, such as usual care by humans [19] and virtual humans or human teletherapists [24]. Design-focused studies explored personalization differences [47], reward structures [45], and emotional support and self-disclosure [23] to enhance chatbots' performance. In addition, a subset of studies (16/43, 37.21%) integrated AI chatbots with other components as part of multicomponent interventions, including digital tools, human interaction, and chatbot-implemented app functions (Figure 2 and Table S5 in Multimedia Appendix 2). Digital tools included exergames [37], the MedLiPal website [48], and wearable activity trackers (WATs). In particular, several studies (4/7, 57.14% [20,35,37,38]) technically integrated WATs with AI chatbots, enabling automatic data sharing. In others (3/7, 42.86% [46,48,49]), participants used WATs for self-monitoring to achieve goals set by the chatbots. AI chatbots were also integrated with human-delivered interventions, such as chemotherapy sessions [35], weight management programs in hospitals [41], physicians' clinic visits [45], the StudentBodies web-based program [39], family-based lifestyle modification programs [40], and remote traditional therapy [50,51]. Furthermore, several studies implemented AI chatbots within stand-alone applications that incorporated additional supportive features. These included a calendar and scoreboard [34], an online self-help library and e-book collection [33], an online smoking cessation diary [45], and self-care practice tools [18].

Figure 2. Architecture for the development of AI chatbots. This is a comprehensive picture derived from all the selected studies. Not every study reported the details of each module. AI: artificial intelligence; API: application programming interface.

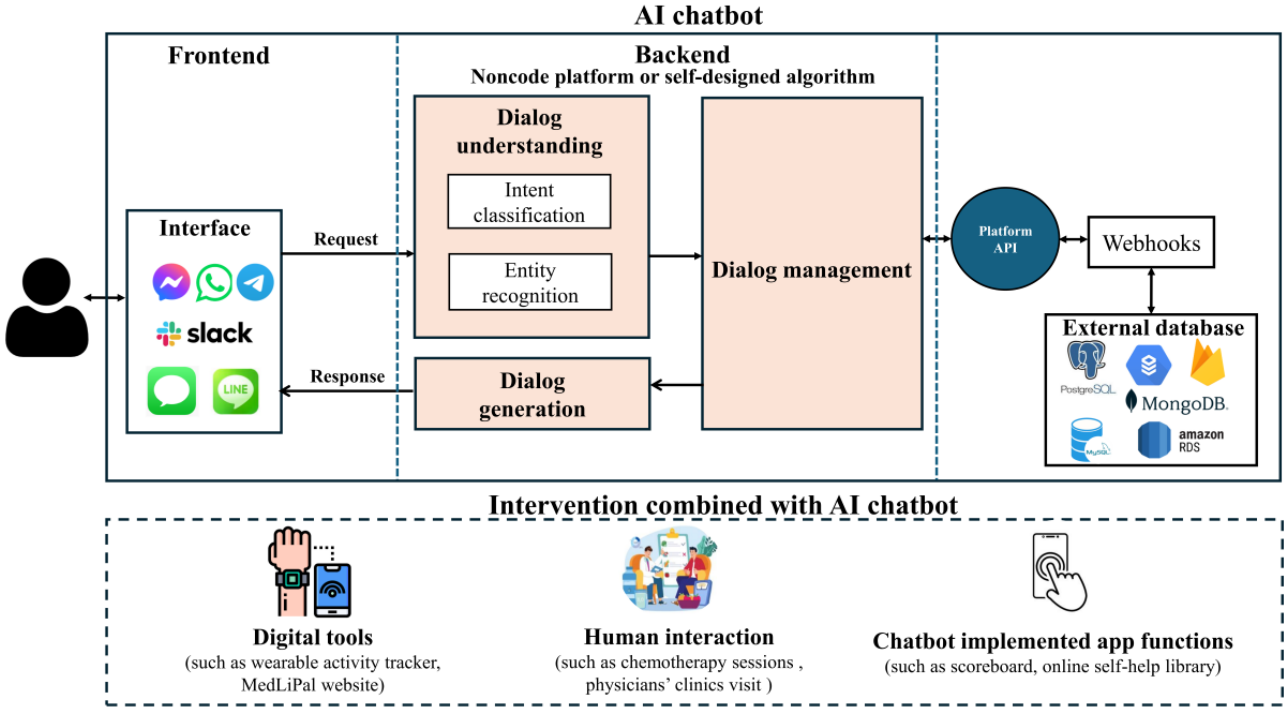

XSL•FO
RenderX

Question 1: What Are the Most Commonly Targeted Health Behaviors in Text-Based AI Chatbots?

Most studies focused on one health behavior (OHB; 29/43, 67.44%), and 14 studies addressed multiple health behaviors (MHBs; 14/43, 32.56%). PA was widely explored (18/43 with 8 OHB and 10 MHB), followed by stress management (16/43 with 10 OHB and 6 MHB), diet (11/43 with 1 OHB and 10 MHB), smoking cessation (7/43 with 6 OHB and 1 MHB), sleep (7/43 with 1 OHB and 6 MHB), weight management (6/43 with 2 OHB and 4 MHB), alcohol (1/43 OHB), and sedentary (1/43 MHB; Table 1).

Table 1. Targeted health behaviors (N=43).

Target behavior	References	Values, n (%)
Single behaviors (n=29, 67.44%)		
Stress management	[23,24,42,50-56]	10 (23.26)
Physical activity (PA)	[16,20,37,38,44,47,57,58]	8 (18.60)
Smoking cessation	[19,34,43,45,59,60]	6 (13.95)
Weight management	[17,41]	2 (4.65)
Diet	[36]	1 (2.33)
Sleep	[61]	1 (2.33)
Alcohol	[62]	1 (2.33)
Multiple behaviors (n=14, 32.56%)		
Diet and PA	[35,48,49,63]	4 (9.30)
Diet and weight management	[39,64]	2 (4.65)
Diet, PA, sleep, and stress management	[65,66]	2 (4.65)
Diet, PA, sleep, and weight management	[40]	1 (2.33)
Diet, PA, sleep, stress management, and sedentary	[67]	1 (2.33)
Sleep and stress management	[18,33]	2 (4.65)
PA and stress management	[68]	1 (2.33)
PA, smoking cessation, and weight management	[46]	1 (2.33)

Question 2: What Roles, Theoretical Foundations, and Behavior Change Techniques Are Applied in Text-Based AI Chatbots, Supporting Health Behavior Change Interventions?

Overview

We classified the AI chatbot for health behavior change into 2 roles and summarized theoretical foundations as well as corresponding functionalities (Table 2).

XSL•FO
RenderX

Table 2. Artificial intelligence chatbot roles, theoretical foundations, and health behavioral change techniques[a].

Paper	Theoretical foundation	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Routine coach and on-demand assistant (n=4)																	
Davis et al (2020) [49]	NR[b]	✓[c]	✓	✓	✓			✓									
Hassoon et al (2021) [38]	NR	✓	✓								✓						
Maher et al (2020) [48]	NR	✓	✓	✓	✓			✓									
Perski et al (2019) [34]	Mohr's model of supportive accountability					✓		✓									
Category sum	—[d]	3	3	2	2	1	0	3	0	0	1	0	0	0	0	0	0
Routine coach (n=27)																	
Albino de Queiroz et al (2023) [35]	CBT[e]		✓					✓									
Albers et al (2023) [47]	NR	✓				✓	✓										✓
Almusharraf et al (2020) [59]	MI[f]	✓			✓					✓							
Brown et al (2023) [60]	MI	✓		✓				✓		✓							
Cameron et al (2018) [33]	NR		✓	✓	✓												
Catellani et al (2023) [37]	Multiple theory 1[g]					✓											
Daley et al (2020) [52]	CBT and positive psychology		✓	✓	✓												
Dhinagaran et al (2021) [67]	COM-B[h]			✓	✓												
Figueroa et al (2021) [16]	Multiple theory 2[i]	✓															
Fitzsimmons - Craft et al (2022) [39]	CBT		✓	✓	✓				✓					✓			
Medeiros et al (2022) [42]	Emotion self-regulation			✓	✓											✓	
Legaspi et al (2022) [18]	NR	✓	✓		✓			✓						✓			
Karhiy et al (2023) [24]	NR				✓												
Meng and Dai (2021) [23]	NR		✓	✓													
Moore et al (2024) [57]	COM-B and the Theoretical Domains Framework	✓	✓		✓				✓							✓	
Piao et al (2020) [44]	HFM[j]	✓	✓					✓	✓		✓						
Piao et al (2020) [58]	HFM	✓	✓					✓	✓		✓						
Rahmanti et al (2022) [64]	COM-B	✓	✓	✓	✓	✓		✓			✓						
Sia et al (2021) [66]	NR	✓		✓	✓			✓			✓					✓	
Sun et al (2023) [68]	MI and graded exercise therapy	✓			✓												
Albers et al (2023) [43]	NR	✓	✓	✓	✓		✓	✓		✓	✓						
Aarts et al (2022) [61]	NR		✓														
Holmes et al (2019) [17]	NR		✓	✓													
Griol et al (2022) [63]	NR		✓	✓	✓	✓											
De Nieva et al (2020) [53]	CBT	✓	✓		✓				✓								
Carrasco-Hernandez et al (2020) [46]	CBT and MI		✓	✓		✓										✓	
Masaki et al (2019) [45]	NR	✓	✓	✓	✓												
Category sum	—	14	17	14	16	5	2	8	5	3	5	0	0	2	0	4	1
On-demand assistant (n=12)																	
Stephens et al (2019) [41]	Multiple theory 3[k]	✓	✓	✓													
Danieli et al (2021) [50]	CBT		✓	✓	✓												
Danieli et al (2022) [51]	CBT		✓	✓	✓												

XSL•FO
RenderX

Paper	Theoretical foundation	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Fadhil et al (2019) [65]	CBT		✓	✓	✓			✓									
Larizza et al (2023) [40]	NR	✓	✓	✓	✓												
Olano-Espinosa et al (2022) [19]	5A clinic practice guideline			✓	✓				✓		✓						
Alghamdi et al (2021) [36]	Multiple theory 4[l]		✓								✓	✓					
To et al (2021) [20]	COM-B	✓	✓	✓		✓		✓	✓		✓					✓	
Durden et al (2023) [54]	CBT	✓	✓		✓				✓								
Forman-Hoffman et al (2023) [55]	CBT	✓	✓	✓	✓				✓								
Hoffman et al (2023) [56]	CBT	✓	✓	✓	✓				✓								
Prochaska et al (2021) [62]	CBT	✓	✓	✓	✓				✓								
Category sum	—	7	11	10	9	1	0	2	6	0	3	1	0	0	0	1	0
Total number	—	24	31	26	27	7	2	13	11	3	9	1	0	2	0	5	1

[a]A: goals and planning; B: feedback and monitoring; C: social support; D: shaping knowledge; E: natural consequences; F: comparison of behavior; G: associations; H: repetition and substitution; I: comparison of outcomes; J: reward and threat; K: regulation; L: antecedents; M: identity; N: scheduled consequences; O: self-belief; P: covert learning.

[b]NR: not reported.

[c]✓: The study reported the results of this theme or subtheme.

[d]Not applicable.

[e]CBT: cognitive behavioral therapy.

[f]MI: motivational interviewing.

[g]Multiple theory 1: elaboration likelihood model, the self-regulatory model of message framing, the regulatory focus theory, and theories of emotions.

[h]COM-B: capability, opportunity, motivation—behavior.

[i]Multiple theory 2: MI, behavioral activation, acceptance and commitment therapy, and solution-focused brief therapy.

[j]HFM: habit formation model.

[k]Multiple theory 3: CBT, MI, and emotionally focused therapy.

[l]Multiple theory 4: chronic-disease extended model extending from the health belief model, the theory of planned behavior, diffusion of innovation theory, social norms theory, and the transtheoretical model.

Roles: Routine Coach and On-Demand Assistant

We classified the chatbot into 2 roles based on the intervention dosage (ie, use frequency and duration per interaction), specifically as a routine coach and an on-demand assistant. AI chatbots mostly played 1 role, such as routine coach (27/43, 62.79%) and on-demand assistant (12/43, 27.91%), whereas 4 studies integrated 2 roles (4/43, 9.30%; Table 2). Specifically, routine coaches delivered support in a defined use frequency and duration per interaction, such as 4 times per week, and each focused on 1 of 4 targeted health behaviors [67]. In contrast, on-demand assistants offer support with flexible frequency and intensity, allowing patients to contact the chatbot anytime and anywhere and determine the duration and frequency of interactions themselves [19].

Theoretical Foundation

Most studies incorporated a theoretical foundation to guide chatbot design strategies (28/43, 65.12%), with 20 studies applying a single theory and 8 studies using an integrated theoretical approach (Table 2 and Table S8 in Multimedia Appendix 2 [16-20,23,24,33-68]). Cognitive behavioral therapy (CBT; 13/28, 46.43%), MI (6/28, 21.53%), and the capability, opportunity, motivation—behavior framework (4/28, 14.29%) were primarily used, either individually or in combination with other theories.

CBT is a form of psychotherapy based on the concept that people's thinking influences their emotions and behaviors [76]. MI is a client-centered, directive therapeutic style to discover the client's own motivation for making changes by guiding clients to explore and resolve ambivalence [50,77]. Finally, capability, opportunity, motivation—behavior is a framework for understanding and changing behavior that posits capability, opportunity, and motivation.

Behavior Change Techniques

Because the functions of AI chatbots varied across studies, we mapped them onto the existing BCT taxonomy (version 1) [72] to enable cross-study comparisons (Table 2). The taxonomy included 16 themes and 93 techniques. We conducted a deductive coding process to code each chatbot function based on these 93 techniques and labeled a cluster if at least 1 technique within it was used. For example, if a chatbot provided automated, tailored feedback on reports and behavioral activity [35], we coded this as the technique "2.7 Feedback on outcome(s) of behavior," which falls under the "Feedback and Monitoring (B)" cluster. Accordingly, we indicated in the table that the study used at least 1 technique within that cluster. The coding was conducted by 1 primary extractor (LF) and validated by 2 additional reviewers (Conrad Ma and Shreya Sanghvi). Any discrepancies were resolved through discussion among all 3 reviewers. The interrater reliability was 81.70%.

XSL•FO
RenderX

Across all included studies, the most frequently applied clusters were goals and planning (A, 24/43, 55.81%), feedback and monitoring (B, 31/43, 72.09%), social support (C, 26/43, 60.47%), and shape knowledge (D, 27/43, 62.79%). No study used the techniques of antecedent (L) and scheduled consequences (N). This pattern shows that AI chatbots currently prioritize conversational, scalable, and digitally applicable techniques, while environmental restructuring and reinforcement-based strategies remain underused due to resource demands and limited environmental control.

BCTs by Roles

Goals and planning, social support, and shaping knowledge were commonly used in both routine coach and on-demand assistant (Table 2). The AI chatbot supported users in setting behavior goals and creating action plans [16,44,47,58], as well as reviewing behavior goals using historical data [20]. Additionally, AI chatbots provided social support (emotional or unspecified) in various ways, including personalized motivational dialogue [46], free-form responses [60], nurse contact information [33], and expressive elements such as sending emojis, icons, GIFs, gamification [52], and images [67]. Other strategies included providing a crisis hotline [36] and a 24-hour on-call [19]. Finally, shaping knowledge primarily involved providing clear instructions on how to perform the behavior [18,24,33,38,42,57,66,68]. For example, chatbots prompted users to share stressful experiences and then offered personalized suggestions, such as planning their day or reframing their mindset [18,42]. The main difference between routine coaches and on-demand assistants was the use of feedback and monitoring techniques (B), applied by 91.67% (11/12) of on-demand assistants and 62.96% (17/27) of routine coaches. This may be explained by the characteristics of on-demand assistants, which allow for continuous interaction, real-time behavior tracking, immediate feedback delivery, and progress reinforcement at any time. Routine coaches typically delivered feedback based on 1-time interactions, such as mood reflection [52], behavioral reporting [17,63], sleep diary [61], and gratitude journaling [53]. In contrast, on-demand assistants provided personalized feedback through active monitoring of health behavior level [20,40,65], mood status [41,55,56,62], and overall health behavior change progress [19,35,36].

Question 3: What Technologies Are Used to Develop Text-Based AI Chatbots for Health Behavior Change?

Workflow of AI Chatbots

Figure 2 summarizes the workflow of AI chatbots, including the frontend module, backend module, and external service module. The process generally involves the following steps: (1) the user sends messages through the frontend interface (eg, social messaging platform, web-based interface, or SMS); (2) the frontend interface forwards the message to the backend (eg, noncode platforms and self-designed algorithms) hosted on the research team's server; and (3) the backend system processes the messages, including dialogue understanding, management, and generation. The function of dialogue understanding is to extract meaning from the user input, such as intents and entities. Dialogue management involves domain-specific knowledge for tailoring replies. The text generation provides output to the user [78]; (4) the generated replies are sent and delivered back to the user through the frontend interface. If the selected intent requires additional operations, such as retrieving data from the external database, the platform sends a request to the webhook through its application programming interface (API). The webhook then processes the intents and returns a response to the platform, which subsequently delivers it to the user in an understandable format. Researchers can create custom webhooks to handle more complex intents.

Based on the development workflow, we classified AI chatbot technologies into 3 types: noncode platforms (23/43, 53.49%), self-designed algorithms (8/43, 18.60%), and stand-alone apps (12/43, 27.91%). These approaches can inform and guide researchers in developing AI chatbots for future use (Table S9 in Multimedia Appendix 2 [16-20,23,24,33-68]).

Noncode Platform

Noncode platforms offer built-in AI algorithms, allowing researchers to develop chatbots without writing code. Among the 23 studies using noncode platforms, the most commonly used platforms were Google Dialogflow (8/23, 34.78%) and IBM Watson (8/23, 34.78%), followed by X2AI (2/23, 8.70%) and Chatfuel (2/23, 8.70%). These AI chatbots were primarily accessed through social messaging platforms, such as Facebook Messenger [23,35,42,67], Slack [48,49], and SMS [16,39,41]. Several studies connected the chatbot to external systems to customize historical data through APIs. Examples include Google Cloud Functions and databases [35], PostgreSQL [40], and MongoDB [16,17,42,63].

Self-Designed Algorithms

Self-designed algorithms require researchers to develop their own AI models, allowing for greater customization of chatbot capabilities. Among 8 studies, some applied Bayes' theorem to assess individual needs and used natural language processing to enhance the chatbot's understanding and response capabilities [19]. Other approaches included combining GPT-2 XL with natural language processing for dialogue understanding and generation [60], deep reinforcement learning [37,43], supervised goal-based models [50,51], self-learning algorithms [46], and the Microsoft Bot Framework [33]. These custom backend systems were integrated with frontend interfaces through APIs, including Telegram [19], web-based interfaces, and SMS text messaging [38], as well as connected with external databases, such as MySQL database [33].

Stand-Alone Applications

Stand-alone apps encapsulate the entire workflow of AI chatbots, as illustrated in Figure 2. In total, 12 studies used stand-alone apps, such as Woebot [53-56,62], Wysa [18], Vitalk [52], Smoke-Free [34], and PsyMe [37]. Each app offered unique features tailored to specific functions. For example, Woebot emphasized helping users develop emotional regulation skills and support mood monitoring and management through conversations. Wysa was not only an AI chatbot but also a comprehensive mental health app, offering additional features such as access to a human talk therapist, stress management techniques, a journal for gratitude, and an international distress signal feature to seek help.

XSL•FO
RenderX

Question 4: How to Validate Text-Based AI Chatbot Performance in Health Behavior Change?

Performance Validation Framework

We mapped the measurements from all included studies onto the digital health scorecard framework [73] and engagement framework [74] to identify existing evidence and current gaps in validating the performance of AI chatbots in health behavior change. The digital health scorecard framework included 4 domains: technical, clinical, usability, and cost. Technical refers to testing if the AI chatbots actually perform to their self-proclaimed functionality with accuracy and precision. Clinical was operationalized as the critical appraisal of evidence demonstrating whether the AI chatbots impact the defined health behavior change outcomes. Usability refers to how easily an AI chatbot can be used for its intended purpose and the minimal effort required to complete tasks. Cost refers to the price for user access, technology lifecycle expenses, and integration costs within clinical workflows. However, these domains primarily assess how well a chatbot performs in influencing changes in health behaviors. Therefore, we incorporated engagement into the framework to capture user engagement between human-AI interaction, reflecting motivational and relational aspects that the other domains do not address. Engagement with digital behavior change interventions includes (1) the extent of use, such as amount, frequency, duration, and depth; and (2) a subjective experience characterized by attention, interest, and affect [74].

In terms of the coding process, first, we conducted deductive coding to map the measures from each study onto 5 domains: technical performance, health behavior change, usability, cost, and engagement. For example, if a study measured "technical feedback from users, ease of use, ease of learning, perceived usefulness, and satisfaction," we mapped "technical feedback from users" to the technical domain; "ease of use" and "ease of learning" to the usability domain; and "perceived usefulness" and "satisfaction" to the engagement domain, based on the definitions of each domain [65]. This process was conducted by 1 primary extractor (LF) and independently validated by 2 additional reviewers (Conrad Ma and Shreya Sanghvi). Any discrepancies were resolved through discussion among all 3 reviewers. The interrater reliability was 90.18%. In the next step, within each domain, we conducted inductive coding to group similar measurements and identify representative metrics, as shown in Figure 3.

Across 43 studies, 17 assessed technical performance, 33 evaluated health behavior change outcomes, 17 examined usability outcomes, and 37 measured engagement outcomes. None of the studies reported cost-related evidence (Figure 3).

Figure 3. Artificial intelligence chatbots on the health behavior change validation framework. MVPA: moderate-to-vigorous physical activity; PA: physical activity.

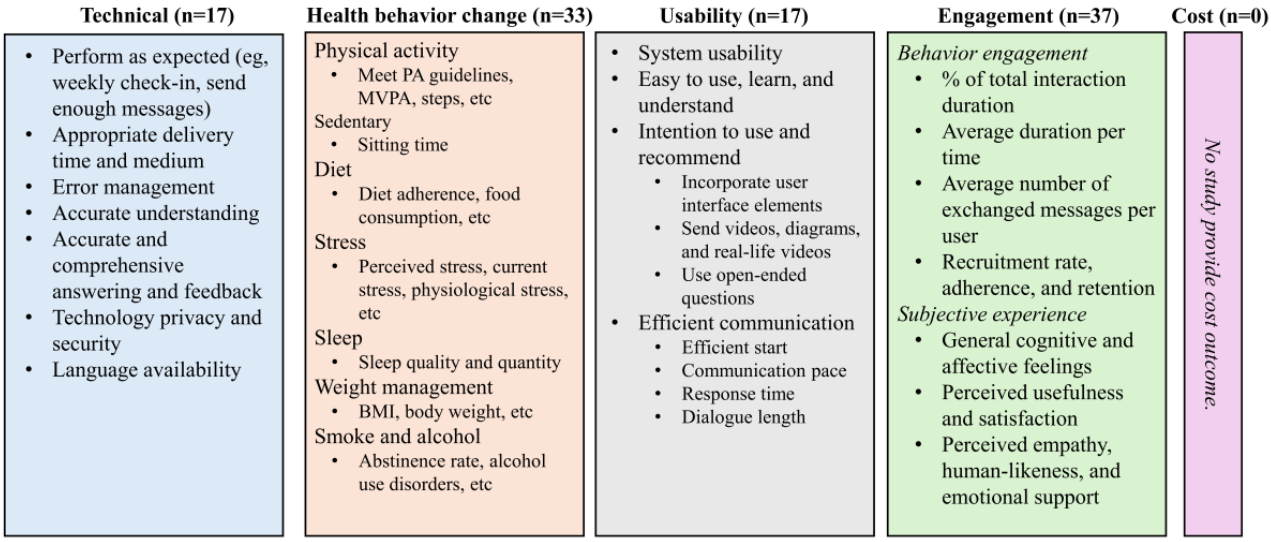

Question 4.1: What Measures Are Used to Assess Technical Performance, Usability, and Engagement?

Technical

In Figure 3, the technical performance of AI chatbots for health behavior change was evaluated across several aspects. The first metric was performance as intended, such as supporting weekly check-ins [49] and delivering a sufficient number of messages [20]. Delivery time and medium were evaluated to determine whether chatbots provided information promptly and through appropriate channels [67]. Error management focused on how effectively chatbots handled unexpected issues [17,20,33]. Several studies also assessed the chatbot's ability to accurately understand user input [33,46,49,53,63,65,66], as well as the provision of accurate and comprehensive information and feedback [40,63,68]. Language availability emerged as a key metric influencing chatbot performance, including the need for additional language options [63] and maintaining language consistency to generate appropriate, user-aligned responses [61,66]. Studies also highlighted that the use of local languages enhanced human connection and personalization [65], while simple and clear language improved user interaction and accessibility [67]. Finally, privacy and security concerns related to the technology were also important metrics to consider when adopting AI chatbots for user interventions [16].

XSL•FO
RenderX

Usability

In terms of usability, 4 studies measured general usability using the System Usability Scale, which includes items such as "I thought the system was easy to use." In total, 3 of those reported above-average industry scores (>68), including 88.2 [33], 84.8 [17], and 79.6 [35], while 1 reported a below-average score of 61.6 [20]. Beyond the System Usability Scale, some studies included items measuring ease of use, ease of learning, and ease of understanding [16,17,35,44,64,65,67,68], as well as intention to use [16,57,58,67,68] and recommendations to others [62]. Usability was also assessed through efficient communication, including smooth onboarding [17,33], suitable interaction pace and response time [17,33,57,64,67], and appropriate conversation length to maintain user engagement [61]. Researchers also identified several features that enhance usability, including human touch and user interface elements [65]; using multimedia such as videos, diagrams, and real-life examples; posing open-ended questions [57]; and allowing free-text input for communication [43].

Engagement

In Figure 3, behavioral engagement refers to engagement intensity and longevity. Most studies reported moderate interaction duration, including 50% [55], 62.5% [35], 66.78% [45] of the total intervention period. The average interaction duration per time was typically less than 30 minutes, such as 5.1 (SD 7.4) minutes [41], 12.5 (SD 15.62) minutes [59], and 21.3 (SD 14.0) minutes [20]. The average number of exchanged messages per user throughout the entire interventions varied between 245.1 [52] and 547.3 [59]. In addition, engagement at different stages was measured through recruitment (how many new individuals are added to a project within a specific time frame), adherence (the extent to which a person's behavior corresponds with agreed-upon recommendations from a care provider [79,80]), and retention (the extent to which the participants completed the study). The recruitment rate ranged from a high of 82% among inactive community-dwelling adults aged 45-75 years [48] to lower rates of 60% among young adults with eating disorders aged 18-30 years [39] and 55.1% among healthy adults aged 21 years and older [67]. Additionally, low adherence (<70% [48]) was reported in 2 studies, with participants completing an average of 63% (6.9/11) of weekly check-ins in one study [48] and 61% (6.7/11) in another [49]. A slight decline in weekly adherence was also observed, decreasing from 77% in week 1 to 69% in week 4 [62]. In contrast, higher retention rates were reported in the other 2 studies (>70% [48]), including 90% [48] and 93% [67].

In terms of subjective experience, several trials reported positive attitudes and acceptance [43,47,53,57], as well as feelings of low frustration, enjoyment [59], interesting [37], attractiveness, stimulation, novelty [17,57], and openness [18]. In addition, other metrics, such as feelings of helpfulness [39,41,51,59] and satisfaction [16,51,65,67], were also commonly reported. Furthermore, several studies highlighted negative perceptions of relational quality, such as complications [16], lack of empathy [18,60], limited human likeness [16,18,66], low affective support [53,57,66], robotic or unfriendly [17], lack of authenticity [47], and low motivational, as well as low emotional support [23].

Question 4.2: What Are the Health Behavior Change Outcomes?

Behavior Change Outcome Overview

Figure 3 illustrates the primary health behavior change outcomes, and Table S10 in Multimedia Appendix 2 [18-20,23,24,34-39,41-43,45-58,60,62,66-68] provides exploratory findings on the efficacy of AI chatbots for each outcome. In total, 33 of the included studies reported health behavior–related outcomes, yielding a total of 120 comparisons. To quantify the magnitude of change across interventions or pre- and postassessments, effect sizes were expressed as either Hedges *g*, odds ratios (ORs), or Cohen *d*. Hedges *g* was calculated when means and SDs were available, whereas ORs were used when only categorical data were reported. Cohen *d* from the original study was used when insufficient information was available to calculate Hedges *g*. Studies reporting Cohen *d* are indicated in Table S10 in Multimedia Appendix 2. According to Cohen conventions, a medium effect of 0.5 is visible to the naked eye of a careful observer [81].

Positive Changes

Among 33 studies with 120 comparisons, 81.67% (98/120) reported positive changes in promoting health behaviors. Positive changes refer to either statistically significant or nonsignificant improvements. However, only 35.83% (43/120) of these comparisons demonstrated observable positive changes with a moderate or larger effect size (Hedges *g* or OR or Cohen *d*>0.5). Moreover, it should be noted that most positive findings were observed in PA, smoking, stress management, and diet, indicating the need for more evidence in weight management, sleep, alcohol use, and sedentary behavior. Additionally, only a small portion of studies were RCTs (14/33, 42.42%), and the populations were primarily nonclinical adults (21/33, 63.64%).

Effectiveness in Real-World Settings

Only 4 of 33 (12.12%) studies evaluated the effectiveness of AI chatbots in real-world settings, all of which were non-RCTs focusing on stress reduction. Among them, the strongest clinically significant decrease in stress was *g*=–0.90 (95% CI –0.97 to –0.83) [52]. When comparing effectiveness by location, use patterns, and emotional status, studies found no significant differences between medically underserved areas and nonmedically underserved areas (t_{253}=0.30; *P*=.77; *d*=0.04, 95% CI –0.23 to 0.30) or between mental health provider shortage areas and nonmental health provider shortage areas (t_{253}=–1.39; *P*=.17; *d*=–0.18, 95% CI –0.44 to 0.07) [55]. Efficient users, those with lower behavioral engagement but stronger therapeutic alliance, achieved greater stress reductions (*g*=–0.60, 95% CI –0.86 to –0.33) than typical users (*g*=–0.25, 95% CI –0.47 to –0.03) and early users (*g*=–0.44, 95% CI –0.71 to –0.17) [56]. Participants with elevated mood symptoms at baseline experienced the greatest stress reduction (*g*=–0.68, 95% CI –0.93 to –0.44) compared with those with low mood symptoms (*g*=–0.28, 95% CI –0.53 to –0.02) [54].

Long-Term Efficacy

A total of 5 (5/33, 15.15%) studies evaluated follow-up efficacy after the intervention, including smoking cessation [45,60], stress [50,51], and diet-related outcomes [39]. Continuous

XSL•FO
RenderX

smoking-related improvements were observed. One study reported confidence (*g*=0.56, 95% CI 0.27-0.84), importance (*g*=0.24, 95% CI –0.03 to 0.52), and readiness to quit smoking (*g*=0.17, 95% CI –0.10 to 0.45) from baseline to 1-week follow-up compared to a single postsession measurement [60]. There was also strong and sustained smoking cessation across multiple follow-up points, with large effects at 12 weeks (*g*=1.40, 95% CI 1.03-1.78), 24 weeks (*g*=1.74, 95% CI 1.31-2.17), and 52 weeks (*g*=1.24, 95% CI 0.88-1.61) compared to 9 weeks after the intervention [45]. However, 2 RCTs did not find a consistent reduction in stress up to 12 weeks after an 8-week intervention in either the chatbot-only group [51] or the multicomponent intervention integrated with the chatbot group [50,51]. Similarly, for nonclinical eating disorder symptoms, the effect size between the intervention and control groups declined over time after the 4-week intervention (12-week: *g*=–0.41, 95% CI –0.63 to –0.20; 24-week: *g*=–0.20, 95% CI –0.41 to 0.01) [39].

Intervention Duration

Intervention duration appears to be an important factor influencing the efficacy of AI chatbots on health behavior change (4/33, 12.12%). A pre- and poststudy found that longer intervention duration (>6 weeks) yielded small but additional benefits across multiple behaviors among middle-aged and older adults, including weight loss (6 weeks: *g*=–0.06; 12 weeks: *g*=–0.07), waist circumference (6 weeks: *g*=–0.06; 12 weeks: *g*=–0.13), diet adherence (6 weeks: *g*=2.04; 12 weeks: *g*=2.06), and PA (6 weeks: *g*=0.32; 12 weeks: *g*=0.39) [48]. Similarly, another RCT reported that the percentage of participants increasing their metabolic equivalent of task scores rose from mid-intervention (*g*=–0.14, 95% CI –1.35 to 1.07) to the end of the 48-week intervention (*g*=0.06, 95% CI –0.87 to 0.98) [46]. Differences in smoking cessation outcomes related to intervention duration were also observed across 2 RCTs. A longer intervention duration of 48 weeks was associated with higher odds of biochemically validated abstinence in the chatbot group compared with the control group (OR 1.01, 95% CI 0.18-1.84) [46], whereas a shorter 24-week duration showed lower odds (OR 0.84, 95% CI 0.31-1.37) [19].

Intervention Modalities

In total, 9 of 33 (27.27%) studies have examined the impact of chatbot modalities on health behavior change outcomes, with all of them being RCTs. First, text-based chatbots performed worse than other modalities, such as video-based chatbots, virtual humans, and human coaches. A text-based AI chatbot showed a smaller increase in PA (*g*=0.35, 95% CI –0.39 to 1.10) compared with a video-based chatbot (*g*=1.14, 95% CI 0.34-1.94) after a 4-week intervention [38]. Similarly, the text-based chatbot demonstrated the smallest effect size in stress reduction (Cohen *d*=0.36) compared with the virtual human (Cohen *d*=0.52) and teletherapy (Cohen *d*=0.54) groups [24]. A text-based chatbot-only intervention (*g*=–0.34, 95% CI –1.33 to 0.65) also performed worse in reducing stress compared with traditional therapy (*g*=–0.71, 95% CI –1.44 to 0.03) [51]. Furthermore, participants who believed that they were interacting with a bot experienced a smaller reduction in stress than those who knew they were interacting with a human [42]. In addition, multicomponent interventions that combined text-based chatbots performed better than traditional therapy or other digital tools alone. For example, traditional therapy plus an AI chatbot led to greater stress reduction after an 8-week intervention compared with traditional therapy alone [50,51]. Similarly, combining psychopharmacological therapy with a digital therapeutic solution including an AI chatbot produced better stress reduction than psychopharmacological therapy alone (*g*=0.13, 95% CI –0.28 to 0.53) [46]. Finally, regarding chatbot design features, chatbots incorporating cues and intrinsic or extrinsic rewards significantly increased PA compared with a control chatbot without intrinsic rewards (t_{104}=2.12; *P*=.04) [58]. Personalized examples were linked to a significant increase in motivation (*g*=0.98, 95% CI 0.60-1.36) but a significant decrease in self-efficacy for PA engagement (*g*=–2.57, 95% CI –3.22 to –1.92) [47].

Engagement

Engagement, encompassing both behavioral engagement and subjective experience, emerged as a significant factor in promoting health behavior change (6/33, 18.18%). Most studies found that strong engagement was associated with positive outcomes, with the exception of one study [52]. High engagers (≥8 weekly check-ins) demonstrated greater increases in PA (high: *g*=0.65, 95% CI –0.14 to 1.44; low: *g*=0.51, 95% CI –0.15 to 1.18) but lower improvements in diet adherence (high: *g*=2.60, 95% CI 1.55-3.64; low: *g*=3.66, 95% CI 2.54-4.78) compared with low engagers [49]. Efficient engagers, those with lower behavioral engagement but stronger therapeutic alliance, had significantly greater reductions in stress than other user groups (*g*=–0.60, 95% CI –0.86 to –0.33) [56]. Intensive users (>4 contacts and >30 minutes of total interaction time) achieved higher quit rates than nonintensive users in both the chatbot intervention group (*g*=1.12, 95% CI 0.55-1.68) and usual care group (*g*=0.52, 95% CI 0.02-1.02) [19]. In addition, a causal mediation analysis explained that higher message involvement positively influenced PA intention through increased feelings of calmness (β=.07; *P*=.003) and greater hope (β=.44; *P*<.001) [37]. Finally, subjective feelings, such as the emotional support provided by AI chatbots, significantly reduced perceived stress through perceived supportiveness, underscoring the importance of subjective engagement experiences [23].

Sleep, Alcohol, Sedentary, and Weight Management

There were a small number of studies examining the efficacy of AI chatbots on sleep (n=2), weight management (n=4), alcohol use (n=1), and sedentary behavior (n=1). All of these were non-RCTs, except for the study by Carrasco-Hernandez et al [46]. First, there was no consistent evidence that AI chatbots effectively improved sleep quality or sleep quantity. One study found no significant effects on sleep quality (*g*=0.02, 95% CI –0.34 to 0.38) or sleep duration, with the proportion of short sleepers increasing by 6% after a 4-week intervention [67]. In contrast, another study reported a modest 3% improvement in sleep quality after a 1-week intervention [66]. Additionally, weight management appeared to be more challenging to change through chatbot interventions. A pre- and poststudy observed only small effects on weight loss (*g*=–0.07, 95% CI –0.57 to 0.43) and waist circumference reduction (*g*=–0.13, 95% CI –0.63 to 0.37) after 12 weeks among

XSL•FO
RenderX

middle-aged and older adults [48]. Similarly, other studies found no significant changes in BMI at 6-week postintervention (g=–0.01, 95% CI –0.27 to 0.24) [20], 24-week mid-intervention (g=–0.05, 95% CI –0.45 to 0.35) [46], and 48-week postintervention (g=0.13, 95% CI –0.28 to 0.53) [46]. Furthermore, the findings for alcohol use and sedentary behavior were relatively positive, showing a significant reduction in alcohol use disorder symptoms (g=–0.42, 95% CI –0.81 to –0.03) [62] and a 32 minutes per day decrease in sitting time [67].

Discussion

Principal Findings

The rapid advancement of AI and increased computational power have significantly expanded the potential applications and advantages of AI chatbots in facilitating health behavior change. This study aimed to provide an up-to-date overview of AI chatbot applications in this domain, along with practical guidance for their development and implementation. Consistent with prior research reviews [10,25,82], PA has emerged as a prominent focus area. This might be due to the need for a scalable intervention to solve the pandemic PA problems [83]. Health behavior change chatbots were classified as routine coaches (predefined frequency and intensity) and on-demand assistants (no specific frequency and intensity). Routine coaching offers a low-cost alternative that can supplement human therapists by providing guidance during their unavailable periods. On-demand assistants allow users to self-monitor and provide timely feedback. The 2 roles address key limitations of conventional interventions by providing more timely, low-cost, and personalized support while also reducing the resource burden on the traditional health care system. Considering theoretical foundations, most AI chatbots have been developed based on CBT and use BCTs such as goal setting and planning, feedback and monitoring, social support, and shaping knowledge. Notably, compared with routine couches, on-demand support chatbots rely more heavily on CBT as well as feedback and monitoring techniques. To achieve these functions, 3 main approaches have been used to develop AI chatbots: noncode platforms, self-designed algorithms, and stand-alone applications. Most studies used noncode platforms, such as Google Dialogflow and IBM Watson, which were then integrated into popular social messaging interfaces, including Facebook Messenger. These noncode platforms are particularly feasible for health behavior researchers who might lack programming expertise. Thus, it significantly improved accessibility and promoted wider adoption of chatbot interventions in health behavior change (across 262 health care centers [19] and up to 57,214 participants [34]).

We refined the validated digital framework [73] by adding engagement elements [74]. The updated framework, which includes technical, health behavior change, usability, engagement, and cost, captures all major measures of assessing the performance of AI chatbots in supporting health behavior change. The findings revealed a notable gap in cost-related evidence and highlighted the need for standardized approaches to calculate a global performance score. Such a standardized benchmark would help distinguish between high- and low-performing AI chatbots and enable cross-study comparisons. Moreover, the exploratory efficacy findings indicated that, although existing studies generally show positive effects of AI chatbots on health behavior change, evidence supporting clinically observable outcomes remains limited. Additionally, most studies have been conducted with nonclinical adult populations (aged 19-64 years), using nonrandomized or short-term trials (≤4 weeks), and have primarily focused on PA, stress management, smoking, and diet. Therefore, researchers should be cautious when applying these findings to clinical settings.

There is also a lack of evidence on the effectiveness of AI chatbots in real-world settings and their long-term efficacy in supporting health behavior change. With regard to intervention design and efficacy, a recent meta-analysis found no significant differences in chatbot effectiveness for increasing moderate-to-vigorous PA, daily steps, or fruit and vegetable consumption by intervention duration or intervention components [12]. In contrast, our exploration scoping review identified consistent findings that the longer intervention duration provides additional benefits across multiple behaviors, such as PA, diet, stress management, and weight management. Multicomponent interventions appeared more effective for stress management and food intake than chatbot-only interventions, though findings for PA were inconsistent. Regarding chatbot modalities, a previous meta-analysis reported that text-based chatbots were more efficacious than audio-based chatbots for fruit and vegetable consumption [12]. In contrast, our exploratory scoping review consistently found that text-based chatbots did not outperform other modalities, such as audio-based chatbots, human therapy, and virtual humans, in terms of PA and perceived stress management. This confirmed the statements that AI chatbots are not intended to replace health care professionals or provide treatment, but rather to complement existing care [52]. The inconsistent findings between this scoping review and the previous meta-analysis [12] underscore the need for additional systematic reviews and meta-analyses to provide more up-to-date and definitive conclusions. The exploratory findings also showed that higher engagement with AI chatbots was associated with greater improvements in health behavior outcomes, including increased PA, better diet adherence, lower perceived stress, and higher quit rates. These findings support previous research, indicating that engagement is a key factor in promoting health behavior change [83]. Finally, the minimal effects on weight management outcomes found in this scoping review were consistent with findings from the broader digital health intervention literature [84,85]. This likely reflects the physiological constraints of weight loss [46] and the fact that most chatbot interventions have targeted activity-related outcomes rather than weight outcomes. Despite these insightful findings, researchers should interpret this conclusion with caution, as it is exploratory and drawn from a broad scoping review rather than a rigorous systematic review and meta-analysis. Moreover, the evidence is based on small and fragmented samples across diverse health behaviors, which limits the strength of conclusions for any single behavior. Future systematic reviews and meta-analyses

XSL•FO
RenderX

covering a wider range of health behaviors are needed to provide stronger and more definitive evidence.

Implications

Practical Implications

AI chatbot shows benefits in promoting health behaviors among nonclinical adult populations, including PA, smoking, stress management, and diet. The chatbot can be strategically leveraged to facilitate health behavior change either as a stand-alone tool or by integrating it into existing programs, serving 2 primary roles: routine coaching and on-demand assistance. Establishing a clear distinction between these roles is critical for determining the appropriate frequency, intensity, and structure of user use. Moreover, researchers can design AI chatbot functionalities based on the synthesized evidence from health behavior change theories and BCTs identified in this scoping review. However, the most effective functionalities remain to be fully explored, and the underlying mechanisms are not yet well understood. Additionally, an accessible approach for health behavior scientists is to use no-code platforms (eg, IBM Watson and Google Dialogflow) or consumer-facing applications (eg, Woebot and Wysa) to develop and deploy AI chatbots for health behavior change interventions. Engagement is a critical factor that requires careful consideration, given the well-documented challenges of sustaining long-term engagement in AI chatbot interventions [34,39]. To address this issue, researchers should develop strategic approaches to maintain user engagement throughout the intervention period. Such strategies may include ensuring high response quality, optimizing interaction length [39], and incorporating visual elements, such as icons and graphs, to enhance user experience and promote sustained participation. It should also be noted that designing such chatbots requires careful consideration of participant characteristics (eg, age, gender, and clinical vs nonclinical populations) and contextual factors (eg, socioeconomic status, digital literacy, cultural norms, and technological environments) to ensure relevance and accessibility, thereby enhancing long-term engagement and achieving targeted outcomes. Finally, future research should incorporate a comprehensive set of evaluation measures encompassing 5 key domains, including technical performance, usability, health behavior changes outcomes, user engagement, and cost, to enable a more rigorous and holistic validation of AI chatbot efficacy.

Research Implications

This review summarized only the BCTs and theoretical foundations, underscoring the need for future research to identify the most influential BCTs and to examine how specific techniques (eg, rewards and graphical feedback) influence health behavior change outcomes within particular theoretical frameworks. In terms of technologies, most studies rely heavily on noncode platforms and conventional AI models. This approach might result in limited natural language communication capabilities and several well-documented issues, including insufficient human-like interaction, a lack of affect, empathy, and emotional support. To address these challenges, future research should consider integrating more advanced AI algorithms, such as generative models (eg, GPTs). Examining whether variations in these technologies influence the overall performance of AI chatbots is also important [86]. Additionally, most of the studies included in this review were conducted among Western, educated, industrialized, rich, and democratic populations [87] and nonclinical healthy adults. This limits the generalizability of the findings and practical guidance, as AI chatbot performance may be moderated by factors such as age differences [18,61], digital literacy, app familiarity, linguistic and cultural differences [16], as well as underserved settings [55]. Therefore, future research should focus on designing AI chatbots tailored to diverse demographic groups (eg, clinical populations, youths, and older adults) and contextual factors (eg, digital health equity) to achieve better outcomes across a broader range of populations.

Regarding AI chatbot validation outcomes, more evidence is needed on cost, weight management, sleep, sedentary behavior, and alcohol use. Additionally, more RCTs involving diverse populations, including younger and older adults, clinical populations, and individuals from varied social, economic, and cultural backgrounds, are needed to provide stronger and more comprehensive evidence. There is also a need to establish a gold standard to standardize scoring across different framework domains, including technical, usability, health behavior change, engagement, and cost. For example, a benchmark can be used to determine that when ≥75% of people think the chatbot is useful, it can be regarded as high accuracy (10/10). This enables the aggregation of individual domain assessments into a Global Digital Health Score, which can help validate the quality of AI chatbots and identify effective digital solutions. It can also highlight areas for improvement and inform stakeholders about potential gaps prior to product deployment. In addition to AI chatbot intervention outcomes, implementation outcomes such as reach, adoption, cost-effectiveness, fidelity, maintenance, scalability, and effectiveness also need to be explored. This would enhance the practical relevance of AI chatbots for digital health practitioners, supporting their implementation in real-world settings and improving scalability. Furthermore, more systematic reviews and meta-analyses need to explore the influence of intervention duration, multicomponent designs, and dose-response factors (eg, duration, frequency, and intensity) on AI chatbot performance, particularly given the variations across different health behaviors. Finally, the associations among different measures within the 5 clusters, including technical, usability, health behavior change, engagement, and cost, require further investigation. For example, usability, measured by willingness to continue, was associated with motivation to engage in activities and smoking quitter self-identity [43]. This can help optimize chatbots to better align with user needs, ultimately leading to improved health behavior change outcomes.

Strengths and Limitations

Strengths

This scoping review contributed to previous research in 5 key ways [10,12,25]. First, unlike prior reviews that focused on narrow behavioral domains, this study encompassed a comprehensive range of health behaviors, including PA, diet, sleep, weight management, sedentary behavior, stress

XSL•FO
RenderX

management, smoking, and alcohol consumption. We also used an extensive search strategy incorporating synonyms and part-of-speech variations. This approach yielded a substantially larger pool of eligible papers, providing a more holistic understanding of AI chatbots' role in health behavior change. Second, this review presents a detailed technology workflow for developing AI chatbots, which spans from frontend interfaces, backend architecture, and integration with external systems. By presenting this framework, we offered health behavior researchers, particularly those without computer science expertise, clear guidance on the technical foundations of chatbot implementation. Third, we classified AI chatbots on the predefined frequency and intensity, offering practical insights for researchers who sought to integrate this technology into health behavior change intervention studies. Fourth, we mapped chatbot functionalities onto the health BCT framework to help practitioners select appropriate BCTs for AI chatbots. Finally, we refined the digital validation tools by incorporating engagement measures, providing future intervention studies with clearer guidance on assessing chatbot performance comprehensively.

Limitations and Future Studies

Several limitations should be noted. First, we excluded studies that integrated audio-based chatbots, embodied conversational agents, humanoid coach virtual reality, augmented reality virtual coach, therapeutic robots, etc. This occurred because our focus was on the communication characteristics of AI chatbots in health behavior change rather than visual, action, or simulated environments. These additional characteristics add another layer of complexity to the deployment of AI chatbots. Future reviews could explore different AI chatbots that include these technologies. Second, we limited our study to publications in English, which might exclude relevant chatbots developed in other languages. Future reviews could consider including studies published in other languages. We also strongly encourage researchers conducting studies in non-English–speaking countries to publish their findings in English to enable cross-cultural comparisons. Finally, we included all types of studies to provide a more comprehensive synthesis, even though some were of relatively low quality. However, the heterogeneity in study designs and methodologies may limit the comparability of findings and the overall strength of the conclusions. We encourage future systematic reviews and meta-analyses to draw more robust insights by focusing on high-quality studies only.

Conclusions

This scoping review offers a comprehensive synthesis of AI chatbots as health behavior change interventions. The analysis revealed that PA was mostly targeted. When designing an AI chatbot, it is important to clearly define its roles (ie, routine coach or on-demand assistant or the combination) as well as to specify its theoretical foundation (eg, CBT), BCTs (eg, goals and planning), and technology workflow (eg, Google Dialogflow integrated with Facebook Messenger). The performance of AI chatbots can be evaluated across 5 clusters: technical, health behavior change, usability, engagement, and cost. Future studies should explore more on cost, sleep, weight management, sedentary behavior, and alcohol use to provide more comprehensive evidence. Additionally, they should also examine implementation outcomes to enhance the scalability of AI chatbot interventions. Moreover, rigorous RCTs in diverse populations are needed to generate robust and generalizable findings. Finally, the sustainability of AI chatbot effects on health behavior change along with factors such as intervention duration, modality, and engagement (eg, use duration, frequency, and intensity), as well as the interactions among the 5 evaluation clusters, warrant further exploration.

Acknowledgments

The authors especially thank Conrad Ma and Shreya Sanghvi for validating the data charting and synthesis for this scoping review. There is "nonhuman" assistance, such as machine learning, artificial intelligence, language models, and similar technology or tools, for the purposes of information gathering, analysis, content creation, manuscript writing, and editing.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Authors' Contributions

Conceptualization: LF (lead), YB (supporting)
Methodology: LF (lead), RB (supporting), YB (supporting)
Data curation: LF (lead), YX (supporting)
Formal analysis: LF (lead)
Writing—original draft: LF (lead)
Writing—review and editing: LF (lead), RB (supporting), YX (supporting), JS (supporting), SZ (supporting), PE (supporting), YB (supporting)

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
PRISMA-ScR checklist.
[DOCX File , 42 KB - jmir_v28i1e79677_app1.docx]

Multimedia Appendix 2
Detailed information on search terms, study eligibility criteria, quality assessment, data extraction items, overview of included studies, study description, characteristics of included studies, theoretical foundation, techniques for developing artificial intelligence chatbots, and efficacy and effectiveness of artificial intelligence chatbots on health behavior change.
[DOCX File , 954 KB - jmir_v28i1e79677_app2.docx]

References

1. Glanz K, Rimer BK, Viswanath K. Health Behavior: Theory, Research, and Practice. 5th Edition. San Francisco, CA: Jossey-Bass; 2015.
2. Rippe JM. Lifestyle medicine: the health promoting power of daily habits and practices. Am J Lifestyle Med 2018;12(6):499-512 [FREE Full text] [doi: 10.1177/1559827618785554] [Medline: 30783405]
3. Merlo G, Vela A. Mental health in lifestyle medicine: a call to action. Am J Lifestyle Med 2022;16(1):7-20 [FREE Full text] [doi: 10.1177/15598276211013313] [Medline: 35185421]
4. Cecchini M, Sassi F, Lauer JA, Lee YY, Guajardo-Barron V, Chisholm D. Tackling of unhealthy diets, physical inactivity, and obesity: health effects and cost-effectiveness. Lancet 2010;376(9754):1775-1784. [doi: 10.1016/S0140-6736(10)61514-0] [Medline: 21074255]
5. Olsen JM, Nesbitt BJ. Health coaching to improve healthy lifestyle behaviors: an integrative review. Am J Health Promot 2010;25(1):e1-e12. [doi: 10.4278/ajhp.090313-LIT-101] [Medline: 20809820]
6. Kivelä K, Elo S, Kyngäs H, Kääriäinen M. The effects of health coaching on adult patients with chronic diseases: a systematic review. Patient Educ Couns 2014;97(2):147-157. [doi: 10.1016/j.pec.2014.07.026] [Medline: 25127667]
7. Palmer S, Tubbs I, Whybrow A. Health coaching to facilitate the promotion of healthy behaviour and achievement of health-related goals. Int J Health Promot Educ 2003;41(3):91-93. [doi: 10.1080/14635240.2003.10806231]
8. Barry MJ, Edgman-Levitan S. Shared decision making—pinnacle of patient-centered care. N Engl J Med 2012;366(9):780-781. [doi: 10.1056/NEJMp1109283] [Medline: 22375967]
9. Mitchell EG, Maimone R, Cassells A, Tobin JN, Davidson P, Smaldone AM, et al. Automated vs. human health coaching: exploring participant and practitioner experiences. Proc ACM Hum Comput Interact 2021;5(CSCW1):1-37 [FREE Full text] [doi: 10.1145/3449173] [Medline: 36304916]
10. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: Systematic review. J Med Internet Res 2023;25:e40789 [FREE Full text] [doi: 10.2196/40789] [Medline: 36826990]
11. Chatbot. Dictionary.com. URL: https://www.dictionary.com/browse/chatbot [accessed 2024-06-25]
12. Singh B, Olds T, Brinsley J, Dumuid D, Virgara R, Matricciani L, et al. Systematic review and meta-analysis of the effectiveness of chatbots on lifestyle behaviours. NPJ Digit Med 2023;6(1):118 [FREE Full text] [doi: 10.1038/s41746-023-00856-1] [Medline: 37353578]
13. Rutjes H, Willemsen MC, Jsselsteijn IWA. Beyond behavior: the coach's perspective on technology in health coaching. 2019 Presented at: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; May 4-9, 2019; Glasgow, Scotland, United Kingdom p. 1-14. [doi: https://doi.org/10.1145/3290605.3300900]
14. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019;6(2):94-98 [FREE Full text] [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
15. Adamopoulou E, Moussiades L. An overview of chatbot technology. Artif Intell Appl Innov 2020;584:373-383. [doi: 10.1007/978-3-030-49186-4_31]
16. Figueroa CA, Luo TC, Jacobo A, Munoz A, Manuel M, Chan D, et al. Conversational physical activity coaches for Spanish and English speaking women: a user design study. Front Digit Health 2021;3:747153 [FREE Full text] [doi: 10.3389/fdgth.2021.747153] [Medline: 34713207]
17. Holmes S, Moorhead A, Bond R, Zheng H, Coates V, Mctear M. Usability testing of a healthcare chatbot: can we use conventional methods to assess conversational user interfaces? 2019 Presented at: Proceedings of the 31st European Conference on Cognitive Ergonomics; September 10-13, 2019; Belfast, United Kingdom p. 207-214. [doi: 10.1145/3335082.3335094]
18. Legaspi CM, Pacana TR, Loja K, Sing C, Ong E. User perception of Wysa as a mental well-being support tool during the COVID-19 pandemic. 2022 Presented at: Asian HCI Symposium '22. ACM; April 29-May 5, 2022; New Orleans, LA, United States p. 52-57. [doi: 10.1145/3516492.3559064]
19. Olano-Espinosa E, Avila-Tomas JF, Minue-Lorenzo C, Matilla-Pardo B, Serrano Serrano ME, Martinez-Suberviola FJ, et al. Effectiveness of a conversational chatbot (Dejal@bot) for the adult population to quit smoking: pragmatic, multicenter, controlled, randomized clinical trial in primary care. JMIR Mhealth Uhealth 2022;10(6):e34273 [FREE Full text] [doi: 10.2196/34273] [Medline: 35759328]

XSL•FO
RenderX

20. To QG, Green C, Vandelanotte C. Feasibility, usability, and effectiveness of a machine learning-based physical activity chatbot: quasi-experimental study. JMIR Mhealth Uhealth 2021;9(11):e28577 [FREE Full text] [doi: 10.2196/28577] [Medline: 34842552]
21. Jörke M, Sapkota S, Warkenthien L. Supporting physical activity behavior change with LLM-based conversational agents. ArXiv Preprint posted online on May 9, 2024. [doi: 10.48550/arXiv.2405.06061]
22. Chauvin R, Clavel C, Ravenet B, Sabouret N. A virtual coach with more or less empathy: impact on older adults' engagement to exercise. 2023 Presented at: 23rd International Conference on Intelligent Virtual Agents (ACM IVA 2023); September 19-22, 2023; Würzburg, Germany. [doi: 10.1145/3570945.3607338]
23. Meng J, Dai YN. Emotional support from AI chatbots: should a supportive partner self-disclose or not? J Comput-Mediat Commun 2021;26(4):207-222. [doi: 10.1093/jcmc/zmab005]
24. Karhiy M, Sagar M, Antoni M, Loveys K, Broadbent E. Mindfulness based stress reduction: a randomised trial of a virtual human, teletherapy, and a chatbot. 2023 Presented at: 2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW); September 10-13, 2023; Cambridge, MA, United States p. 1-7. [doi: 10.1109/aciiw59127.2023.10388195]
25. Oh YJ, Zhang J, Fang M, Fukuoka Y. A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss. Int J Behav Nutr Phys Act 2021;18(1):160 [FREE Full text] [doi: 10.1186/s12966-021-01224-6] [Medline: 34895247]
26. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018;18(1):143 [FREE Full text] [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
27. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
28. Fu L, Xie Y, Bai Y. The development and use of artificial intelligence chatbots for health behavior change—a scoping review. Open Science Framework. 2024. URL: https://osf.io/wcspx/overview [accessed 2026-01-04]
29. Gray ID, Kross AR, Renfrew ME, Wood P. Precision medicine in lifestyle medicine: the way of the future? Am J Lifestyle Med 2019;14(2):169-186 [FREE Full text] [doi: 10.1177/1559827619834527] [Medline: 32231483]
30. Methley AM, Campbell S, Chew-Graham C, McNally R, Cheraghi-Sohi S. PICO, PICOS and SPIDER: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews. BMC Health Serv Res 2014;14:579 [FREE Full text] [doi: 10.1186/s12913-014-0579-0] [Medline: 25413154]
31. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012:276-282. [doi: 10.11613/BM.2012.031]
32. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018;34(4):285-291. [doi: 10.3233/EFI-180221]
33. Cameron G, Cameron D, Megaw G. Assessing the usability of a chatbot for mental health care. In: Lecture Notes in Computer Science. 2018 Presented at: INSCI 2018 International Workshop; October 24-26, 2018; St Petersburg, Russia p. 121-132. [doi: 10.1007/978-3-030-17705-8_11]
34. Perski O, Crane D, Beard E, Brown J. Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study. Digit Health 2019;5:2055207619880676 [FREE Full text] [doi: 10.1177/2055207619880676] [Medline: 31620306]
35. Albino de Queiroz D, Silva Passarello R, Veloso de Moura Fé V, Rossini A, Folchini da Silveira E, Aparecida Isquierdo Fonseca de Queiroz E, et al. A wearable chatbot-based model for monitoring colorectal cancer patients in the active phase of treatment. Healthc Anal 2023;4:100257. [doi: 10.1016/j.health.2023.100257]
36. Alghamdi E, Alnanih R. Chatbot design for a healthy life to celiac patients: a study according to a new behavior change model. Int J Adv Comput Sci Appl 2021;12(10):698-707. [doi: 10.14569/ijacsa.2021.0121077]
37. Catellani P, Biella M, Carfora V, Nardone A, Brischigiaro L, Manera MR, et al. A theory-based and data-driven approach to promoting physical activity through message-based interventions. Front Psychol 2023;14:1200304 [FREE Full text] [doi: 10.3389/fpsyg.2023.1200304] [Medline: 37575427]
38. Hassoon A, Baig Y, Naiman DQ, Celentano DD, Lansey D, Stearns V, et al. Randomized trial of two artificial intelligence coaching interventions to increase physical activity in cancer survivors. NPJ Digit Med 2021;4(1):168 [FREE Full text] [doi: 10.1038/s41746-021-00539-9] [Medline: 34887491]
39. Fitzsimmons-Craft EE, Chan WW, Smith AC, Firebaugh M, Fowler LA, Topooco N, et al. Effectiveness of a chatbot for eating disorders prevention: a randomized clinical trial. Int J Eat Disord 2022;55(3):343-353. [doi: 10.1002/eat.23662] [Medline: 35274362]
40. Larizza C, Bosoni P, Quaglini S, Chasseur M, Bevolo V, Zuccotti G, et al. V-care: an application to support lifestyle improvement in children with obesity. Int J Med Inform 2023;177:105140. [doi: 10.1016/j.ijmedinf.2023.105140] [Medline: 37463558]
41. Stephens TN, Joerin A, Rauws M, Werk LN. Feasibility of pediatric obesity and prediabetes treatment support through Tess, the AI behavioral coaching chatbot. Transl Behav Med 2019;9(3):440-447. [doi: 10.1093/tbm/ibz043] [Medline: 31094445]

XSL•FO
RenderX

42. Medeiros L, Bosse T, Gerritsen C. Can a chatbot comfort humans? Studying the impact of a supportive chatbot on users' self-perceived stress. IEEE Trans Human-Mach Syst 2022;52(3):343-353. [doi: 10.1109/THMS.2021.3113643]
43. Albers N, Neerincx MA, Aretz NL, Ali M, Ekinci A, Brinkman WP. Attitudes toward a virtual smoking cessation coach: relationship and willingness to continue. In: Meschtscherjakov A, Midden C, Ham J, editors. Persuasive Technology. PERSUASIVE 2023. Lecture Notes in Computer Science, vol 13832. Switzerland: Springer; 2023:265-274.
44. Piao M, Kim J, Ryu H, Lee H. Development and usability evaluation of a healthy lifestyle coaching chatbot using a habit formation model. Healthc Inform Res 2020;26(4):255-264 [FREE Full text] [doi: 10.4258/hir.2020.26.4.255] [Medline: 33190459]
45. Masaki K, Tateno H, Kameyama N, Morino E, Watanabe R, Sekine K, et al. Impact of a novel smartphone app (CureApp smoking cessation) on nicotine dependence: prospective single-arm interventional pilot study. JMIR Mhealth Uhealth 2019;7(2):e12694 [FREE Full text] [doi: 10.2196/12694] [Medline: 30777848]
46. Carrasco-Hernandez L, Jódar-Sánchez F, Núñez-Benjumea F, Moreno Conde J, Mesa González M, Civit-Balcells A, et al. A mobile health solution complementing psychopharmacology-supported smoking cessation: randomized controlled trial. JMIR Mhealth Uhealth 2020;8(4):e17530 [FREE Full text] [doi: 10.2196/17530] [Medline: 32338624]
47. Albers N, Hizli B, Scheltinga BL, Meijer E, Brinkman WP. Setting physical activity goals with a virtual coach: Vicarious experiences, personalization and acceptance. J Med Syst 2023;47(1):15 [FREE Full text] [doi: 10.1007/s10916-022-01899-9] [Medline: 36710276]
48. Maher CA, Davis CR, Curtis RG, Short CE, Murphy KJ. A physical activity and diet program delivered by artificially intelligent virtual health coach: proof-of-concept study. JMIR Mhealth Uhealth 2020;8(7):e17558 [FREE Full text] [doi: 10.2196/17558] [Medline: 32673246]
49. Davis CR, Murphy KJ, Curtis RG, Maher CA. A process evaluation examining the performance, adherence, and acceptability of a physical activity and diet artificial intelligence virtual health assistant. Int J Environ Res Public Health 2020;17(23):9137 [FREE Full text] [doi: 10.3390/ijerph17239137] [Medline: 33297456]
50. Danieli M, Ciulli T, Mousavi SM, Riccardi G. A conversational artificial intelligence agent for a mental health care app: evaluation study of its participatory design. JMIR Form Res 2021;5(12):e30053 [FREE Full text] [doi: 10.2196/30053] [Medline: 34855607]
51. Danieli M, Ciulli T, Mousavi SM, Silvestri G, Barbato S, Di Natale L, et al. Assessing the impact of conversational artificial intelligence in the treatment of stress and anxiety in aging adults: randomized controlled trial. JMIR Ment Health 2022;9(9):e38067 [FREE Full text] [doi: 10.2196/38067] [Medline: 36149730]
52. Daley K, Hungerbuehler I, Cavanagh K, Claro HG, Swinton PA, Kapps M. Preliminary evaluation of the engagement and effectiveness of a mental health chatbot. Front Digit Health 2020;2:576361 [FREE Full text] [doi: 10.3389/fdgth.2020.576361] [Medline: 34713049]
53. De Nieva JO, Joaquin JA, Tan CB, Marc Te RK, Ong E. Investigating students' use of a mental health chatbot to alleviate academic stress. 2021 Presented at: 6th International ACM In-Cooperation HCI and UX Conference; October 21-23, 2020; Jakarta & Bandung, Indonesia p. 1-10. [doi: 10.1145/3431656.3431657]
54. Durden E, Pirner MC, Rapoport SJ, Williams A, Robinson A, Forman-Hoffman VL. Changes in stress, burnout, and resilience associated with an 8-week intervention with relational agent "Woebot". Internet Interv 2023;33:100637 [FREE Full text] [doi: 10.1016/j.invent.2023.100637] [Medline: 37635948]
55. Forman-Hoffman VL, Pirner MC, Flom M, Kirvin-Quamme A, Durden E, Kissinger JA, et al. Engagement, satisfaction, and mental health outcomes across different residential subgroup users of a digital mental health relational agent: exploratory single-arm study. JMIR Form Res 2023;7:e46473 [FREE Full text] [doi: 10.2196/46473] [Medline: 37756047]
56. Hoffman V, Flom M, Mariano TY, Chiauzzi E, Williams A, Kirvin-Quamme A, et al. User engagement clusters of an 8-week digital mental health intervention guided by a relational agent (Woebot): exploratory study. J Med Internet Res 2023;25:e47198 [FREE Full text] [doi: 10.2196/47198] [Medline: 37831490]
57. Moore R, Al-Tamimi AK, Freeman E. Investigating the potential of a conversational agent (Phyllis) to support adolescent health and overcome barriers to physical activity: co-design study. JMIR Form Res 2024;8:e51571 [FREE Full text] [doi: 10.2196/51571] [Medline: 38294857]
58. Piao M, Ryu H, Lee H, Kim J. Use of the healthy lifestyle coaching chatbot app to promote stair-climbing habits among office workers: exploratory randomized controlled trial. JMIR Mhealth Uhealth 2020;8(5):e15085 [FREE Full text] [doi: 10.2196/15085] [Medline: 32427114]
59. Almusharraf F, Rose J, Selby P. Engaging unmotivated smokers to move toward quitting: design of motivational interviewing-based chatbot through iterative interactions. J Med Internet Res 2020;22(11):e20251 [FREE Full text] [doi: 10.2196/20251] [Medline: 33141095]
60. Brown A, Kumar AT, Melamed O, Ahmed I, Wang YH, Deza A, et al. A motivational interviewing chatbot with generative reflections for increasing readiness to quit smoking: iterative development study. JMIR Ment Health 2023;10:e49132 [FREE Full text] [doi: 10.2196/49132] [Medline: 37847539]
61. Aarts T, Markopoulos P, Giling L, Vacaretu T, Pillen S. Snoozy: a chatbot-based sleep diary for children aged eight to twelve. 2022 Presented at: IDC '22: Interaction Design and Children; June 27-30, 2022; Braga, Portugal p. 297-307. [doi: https://doi.org/10.1145/3501712.3529718]

XSL•FO
RenderX

62. Prochaska JJ, Vogel EA, Chieng A, Kendra M, Baiocchi M, Pajarito S, et al. A therapeutic relational agent for reducing problematic substance use (Woebot): development and usability study. J Med Internet Res 2021;23(3):e24850 [FREE Full text] [doi: 10.2196/24850] [Medline: 33755028]
63. Griol D, Callejas Z, Fernandez-Martinez F, Esposito A. An application of conversational systems to promote healthy lifestyle habits. 2022 Presented at: 2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech); September 12-15, 2022; Falerna, Italy p. 1-6. [doi: 10.1109/dasc/picom/cbdcom/cy55231.2022.9927835]
64. Rahmanti AR, Yang H, Bintoro BS, Nursetyo AA, Muhtar MS, Syed-Abdul S, et al. SlimMe, a chatbot with artificial empathy for personal weight management: system design and finding. Front Nutr 2022;9:870775 [FREE Full text] [doi: 10.3389/fnut.2022.870775] [Medline: 35811989]
65. Fadhil A, AbuRa'ed A. OlloBot—towards a text-based Arabic health conversational agent: evaluation and results. 2019 Presented at: Proceedings—Natural Language Processing in a Deep Learning World; September 2-4, 2019; Shoumen, Bulgaria p. 295-303. [doi: 10.26615/978-954-452-056-4_034]
66. Sia DE, Yu MJ, Daliva JL, Montenegro J, Ong E. Investigating the acceptability and perceived effectiveness of a chatbot in helping students assess their well-being. 2021 Presented at: Asian CHI Symposium 2021; May 8-13, 2021; Yokohama, Japan p. 34-40. [doi: 10.1145/3429360.3468177]
67. Dhinagaran DA, Sathish T, Soong A, Theng Y, Best J, Tudor Car L. Conversational agent for healthy lifestyle behavior change: web-based feasibility study. JMIR Form Res 2021;5(12):e27956 [FREE Full text] [doi: 10.2196/27956] [Medline: 34870611]
68. Sun X, Casula D, Navaratnam A. Virtual support for real-world movement: using chatbots to overcome barriers to physical activity. 2023 Presented at: 2nd International Conference on Hybrid Human-Artificial Intelligence; June 26-30, 2023; Munich, Germany p. 201-214. [doi: 10.3233/FAIA230084]
69. Pace R, Pluye P, Bartlett G, Macaulay AC, Salsberg J, Jagosh J, et al. Testing the reliability and efficiency of the pilot Mixed Methods Appraisal Tool (MMAT) for systematic mixed studies review. Int J Nurs Stud 2012;49(1):47-53. [doi: 10.1016/j.ijnurstu.2011.07.002] [Medline: 21835406]
70. Luo TC, Aguilera A, Lyles CR, Figueroa CA. Promoting physical activity through conversational agents: mixed methods systematic review. J Med Internet Res 2021;23(9):e25486. [doi: 10.2196/25486] [Medline: 34519653]
71. Lefevre AE, Agarwal S, Zeller K, Vasudevan L, Healey K, Tamrat T, et al. Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment. World Health Organization. 2016. URL: https://www.who.int/publications/i/item/9789241511766 [accessed 2024-03-08]
72. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
73. Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB, Cohen AB. Digital health: a path to validation. NPJ Digit Med 2019;2(1):38 [FREE Full text] [doi: 10.1038/s41746-019-0111-3] [Medline: 31304384]
74. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017;7(2):254-267 [FREE Full text] [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
75. Lindemann EA, Chen ES, Wang Y, Skube SJ, Melton GB. Representation of social history factors across age groups: a topic analysis of free-text social documentation. AMIA Annu Symp Proc 2017:1169-1178 [FREE Full text] [Medline: 29854185]
76. Beck JS. Cognitive behavior therapy. 3rd Edition. In: Basics and Beyond. New York, NY: Guilford Publications; 2020.
77. Miller WR, Rollnick S. Motivational Interviewing: Helping People Change. New York, NY: Guilford Press; 2012.
78. Safi Z, Abd-Alrazaq A, Khalifa M, Househ M. Technical aspects of developing chatbots for medical applications: scoping review. J Med Internet Res 2020;22(12):e19127 [FREE Full text] [doi: 10.2196/19127] [Medline: 33337337]
79. Persch AC, Page SJ. Protocol development, treatment fidelity, adherence to treatment, and quality control. Am J Occup Ther 2013;67(2):146-153 [FREE Full text] [doi: 10.5014/ajot.2013.006213] [Medline: 23433268]
80. Salema NEM, Elliott RA, Glazebrook C. A systematic review of adherence-enhancing interventions in adolescents taking long-term medicines. J Adolesc Health 2011;49(5):455-466. [doi: 10.1016/j.jadohealth.2011.02.010] [Medline: 22018559]
81. Sullivan GM, Feinn R. Using effect size—or why the P value is not enough. J Grad Med Educ 2012;4(3):279-282 [FREE Full text] [doi: 10.4300/JGME-D-12-00156.1] [Medline: 23997866]
82. Chew HSJ. The use of artificial intelligence-based conversational agents (chatbots) for weight loss: scoping review and practical recommendations. JMIR Med Inform 2022;10(4):e32578 [FREE Full text] [doi: 10.2196/32578] [Medline: 35416791]
83. Kohl HW, Craig CL, Lambert EV, Inoue S, Alkandari JR, Leetongin G, et al. The pandemic of physical inactivity: global action for public health. Lancet 2012;380(9838):294-305 [FREE Full text] [doi: 10.1016/S0140-6736(12)60898-8] [Medline: 22818941]

XSL•FO
RenderX

84. Beleigoli AM, Andrade AQ, Cançado AG, Paulo MN, Diniz MDFH, Ribeiro AL. Web-based digital health interventions for weight loss and lifestyle habit changes in overweight and obese adults: systematic review and meta-analysis. J Med Internet Res 2019;21(1):e298 [FREE Full text] [doi: 10.2196/jmir.9609] [Medline: 30622090]
85. Cavero-Redondo I, Martinez-Vizcaino V, Fernandez-Rodriguez R, Saz-Lara A, Pascual-Morena C, Álvarez-Bueno C. Effect of behavioral weight management interventions using lifestyle mHealth self-monitoring on weight loss: a systematic review and meta-analysis. Nutrients 2020;12(7):1977 [FREE Full text] [doi: 10.3390/nu12071977] [Medline: 32635174]
86. Albers N, Neerincx MA, Brinkman WP. Addressing people's current and future states in a reinforcement learning algorithm for persuading to quit smoking and to be physically active. PLoS One 2022;17(12):e0277295 [FREE Full text] [doi: 10.1371/journal.pone.0277295] [Medline: 36454782]
87. Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? Behav Brain Sci 2010;33(2-3):61-83 [FREE Full text] [doi: 10.1017/s0140525x0999152x]

Abbreviations

AI: artificial intelligence
API: application programming interface
BCT: behavior change technique
CBT: cognitive behavioral therapy
MHB: multiple health behavior
MI: motivational interviewing
MMAT: Mixed Methods Appraisal Tool
OHB: one health behavior
OR: odds ratio
PA: physical activity
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RCT: randomized controlled trial
WAT: wearable activity tracker

Edited by J Sarvestan; submitted 25.Jun.2025; peer-reviewed by N Acharya, S Hooker; comments to author 29.Jul.2025; revised version received 09.Dec.2025; accepted 10.Dec.2025; published 28.Jan.2026.

Please cite as:
Fu L, Burns R, Xie Y, Shen J, Zhe S, Estabrooks P, Bai Y
The Development and Use of AI Chatbots for Health Behavior Change: Scoping Review
J Med Internet Res 2026;28:e79677
URL: https://www.jmir.org/2026/1/e79677
doi:10.2196/79677
PMID:

©Lingyi Fu, Ryan Burns, Yuhuan Xie, Jincheng Shen, Shandian Zhe, Paul Estabrooks, Yang Bai. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Efficacy of Brain-Computer Interface Therapy for Upper Limb Rehabilitation in Chronic Stroke: Systematic Review and Meta-Analysis of Randomized Controlled Trials

HongJie Chen[1,2], MSc; GuoJun Yun[1], MD

[1]Department of Rehabilitation Medicine, Shenzhen Children's Hospital, Shenzhen City, Guangdong Province, China
[2]College of Rehabilitation Medicine, Jiamusi University, Jiamusi, Heilongjiang, China

Corresponding Author:
GuoJun Yun, MD
Department of Rehabilitation Medicine
Shenzhen Children's Hospital
7019 Yitian Road, Futian District
Shenzhen City, Guangdong Province, 518026
China
Phone: 86 18938691526
Email: 103872187@qq.com

Abstract

Background: Over 50% of people with chronic stroke experience persistent upper limb dysfunction. Brain-computer interface (BCI) therapy, creating a sensorimotor loop via neural feedback, is a promising alternative; yet, its optimal application remains unclear.

Objective: This meta-analysis evaluates BCI's efficacy on motor function, tone, and activities of daily living (ADL) in chronic stroke and identifies optimal feedback modalities and intervention parameters.

Methods: We systematically searched Cochrane Library, Embase, PubMed, Scopus, Web of Science, and Wanfang Data from inception to October 2025 for randomized controlled trials (RCTs) comparing BCI-based training to control interventions in adults with chronic stroke. Primary outcomes were upper limb motor function (Fugl-Meyer Assessment for upper extremity [FMA-UE], Action Research Arm Test [ARAT]), muscle tone (Modified Ashworth Scale [MAS]), and ADL (Modified Barthel Index [MBI], Motor Activity Log [MAL]). Screening, data extraction, and risk-of-bias assessment were performed independently. Meta-analysis used a random-effects model with Hartung-Knapp-Sidik-Jonkman adjustment. Pooled mean differences (MDs) with 95% CIs and 95% prediction intervals (PIs) were calculated. Subgroup analyses examined feedback modalities, intervention intensity, and follow-up effects. Sensitivity analysis was also conducted.

Results: From 3529 records, 21 RCTs (650 participants) were included. BCI training significantly improved motor function (FMA-UE: MD 2.50, 95% CI 0.60-4.40; *P*=.01; 95% PI –2.52 to 7.22) and ADL performance (MBI: MD 8.38, 95% CI 2.23-14.53; *P*=.02; 95% PI –3.92 to 20.53; MAL: MD 2.09, 95% CI 0.42-3.76; *P*=.03; 95% PI –0.69 to 4.54). No significant effects were observed for fine motor skills (ARAT: MD 0.18, 95% CI –0.27 to 0.62; *P*=.30; 95% PI –3.64 to 3.99) or muscle tone (MAS: MD –0.48, 95% CI –1 to 0.03; *P*=.06; 95% PI –1.27 to 0.35). Subgroup analyses revealed that BCI-functional electrical stimulation (FES) yielded the greatest improvement in motor recovery (FMA-UE: MD 5, 95% CI 1.86-8.13; *P*=.01). The optimal intervention protocol was identified as 30-minute sessions, administered 4-5 times per week over 2 weeks (total of 10-12 sessions). However, benefits were not sustained at follow-up.

Conclusions: Low- to moderate-certainty evidence suggests that BCI training, particularly the BCI-FES paradigm, can improve upper limb motor function and ADL in people with chronic stroke on average. However, wide prediction intervals indicate the effect may vary substantially across settings, ranging from negligible to beneficial. Subgroup analyses suggested a potential optimal protocol of 30-minute sessions, 4-5 times per week for 2 weeks, but these findings are limited by the small number of studies in each subgroup and the high risk of bias in several included trials. Therefore, this proposed protocol should be viewed as preliminary and requires validation in future, high-quality RCTs. Future research should also focus on identifying patient subgroups most likely to benefit and on strategies to sustain long-term gains.

Trial Registration: PROSPERO CRD420251063808; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251063808

XSL•FO
RenderX

(J Med Internet Res 2026;28:e79132) doi:10.2196/79132

KEYWORDS

brain-computer interface; chronic disease; upper limb function; stroke; rehabilitation; meta-analysis

Introduction

Stroke, a neurological disorder resulting from cerebrovascular rupture or obstruction, leads to motor, speech, and cognitive impairments, consequently compromising performance in activities of daily living (ADL) [1]. Upper limb motor dysfunction represents one of the most prevalent sequelae of stroke [2,3], affecting a substantial proportion of survivors. Contemporary evidence indicates conventional rehabilitation strategies achieve optimal therapeutic gains predominantly within the first 6 months poststroke [4,5]. However, a substantial proportion of patients miss this critical window. For those in the chronic phase, current interventions demonstrate limited efficacy [6]. Notably, comparative studies reveal that stand-alone exoskeleton-assisted training or functional electrical stimulation (FES) provides comparable therapeutic benefits to conventional rehabilitation in chronic stroke cohorts [7]. Given the persistent, unmet rehabilitation needs in this population, brain-computer interface (BCI)–based training merits serious consideration. By enabling heightened patient engagement in volitional motor tasks [8], BCI represents a paradigm-shifting approach with significant potential to enhance recovery trajectories.

BCI technology establishes a direct communication pathway between the brain and an external device, bypassing damaged neural pathways. Fundamentally, BCI systems acquire and decode characteristic patterns of neural activity associated with user intent, such as motor imagery [9]. These decoded signals are then translated into commands to operate external feedback devices. BCI training represents an emerging neurorehabilitation technology based on this principle. This approach collects and decodes characteristic brain activity patterns, translating them into computerized commands to operate external feedback devices. These devices include FES [10], robotic exoskeletons [11], and visual feedback systems [12]. BCI systems are broadly categorized as invasive or noninvasive based on signal acquisition methodology. Due to safety concerns and practical limitations associated with invasive techniques [13], noninvasive BCIs are predominantly favored in current rehabilitation practice [14]. Primary noninvasive signal acquisition modalities encompass electroencephalography (EEG), magnetoencephalography, functional near-infrared spectroscopy (fNIRS), and functional magnetic resonance imaging. Among these, EEG stands as the predominant modality for signal acquisition in clinical rehabilitation settings [15]. The level of control exerted over external devices is contingent upon the specific neural signal source used. Integrating diverse neural signals within BCI frameworks enables more refined and efficient operation of feedback apparatus [16]. In contrast to conventional rehabilitation methods, the BCI paradigm establishes a “central-peripheral-central” closed-loop model. This approach holds promise for facilitating more timely movement adjustments and compensation strategies in people who have experienced a stroke [17], potentially offering greater alignment with personalized rehabilitation requirements.

Since the initial report on EEG-based BCI training in 2009 [18], numerous studies have demonstrated the efficacy of BCI interventions for improving upper limb motor outcomes in people who have experienced a stroke, including muscle strength, motor function, and ADL [12,19,20]. While some meta-analyses have addressed stroke stages in subgroup analyses, recent systematic reviews and meta-analyses primarily focus on comparing the magnitude of improvement across different stroke stages for a single primary outcome measure. For instance, subgroup analyses in studies by Xie et al [21] and Yang et al [22] solely compared BCI efficacy on upper limb function across stroke stages, without specifically evaluating the long-term therapeutic effects of BCI in people with chronic stroke. Furthermore, the influence of critical clinical parameters—such as intervention duration and session frequency—remains inadequately examined. A systematic analysis of BCI-based training protocols optimized for chronic stroke populations has yet to be conducted. Addressing these aspects is crucial for developing tailored rehabilitation protocols to enhance upper limb recovery, ADL performance, and overall quality of life in people with chronic stroke.

Therefore, this meta-analysis was conducted with three specific aims:

1. To systematically evaluate the efficacy of BCI-based training on upper limb motor function, muscle tone, and ADL exclusively in people with chronic stroke;
2. To perform in-depth subgroup analyses of key moderating factors—including BCI feedback modalities, intervention intensity parameters (session duration, frequency, and total sessions), and follow-up effects—which have been inadequately addressed in prior syntheses;
3. To attempt to propose optimized BCI intervention protocols tailored to the chronic stroke population based on current evidence.

By addressing these gaps, this review seeks to provide clearer guidance for clinical practice and future research in BCI-based stroke rehabilitation.

Methods

Protocol and Registration

This systematic review was registered with PROSPERO (International Prospective Register of Systematic Reviews), the international systematic review database, bearing identifier CRD420251063808. This meta-analysis followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines published in 2020 [23] (Multimedia Appendix 1).

XSL•FO
RenderX

Search Strategy

The systematic literature search was designed, conducted, and reported in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses – Search Extension) guideline [24]. An experienced information specialist (HJC) developed the search strategy in collaboration with the review team.

We executed a comprehensive search across 6 electronic databases from their inception until October 16, 2025, including PubMed (via the National Library of Medicine), Embase (via Ovid), Cochrane Central Register of Controlled Trials (via Wiley), Scopus (via Elsevier), Web of Science Core Collection (via Clarivate Analytics), and Wanfang Data. These platforms were selected to ensure extensive coverage of both international and Chinese literature. All databases were searched individually; no multidatabase searching was performed on a single platform.

The search strategy used a combination of controlled vocabulary (eg, MeSH [Medical Subject Headings] in PubMed and Emtree in Embase) and keywords related to the core concepts of "brain-computer interfaces," "stroke," "upper extremity," and "rehabilitation." The complete search strategies for all databases are provided in Multimedia Appendix 2. To maximize sensitivity, no restrictions were placed on language, publication date, or study design during the search. Similarly, no published search filters were used, and the strategy was developed de novo for this review, not adapted from prior work.

To enhance the robustness of the search, the PubMed strategy underwent peer review by an information specialist prior to execution, following the Peer Review of Electronic Search Strategies (PRESS) guideline framework [25]. Additionally, in attempts to obtain missing or incomplete data, we contacted the corresponding authors of studies via email.

Beyond the methods above, we did not systematically search study registries, websites, or gray literature, nor did we use citation searching.

The total number of records retrieved from each database is documented in the PRISMA flow diagram. All identified records were imported into EndNote X9 (Clarivate) for management, where duplicates were removed using the software's automated deduplication feature, followed by a manual verification process conducted independently by 2 reviewers (HJC and GJY).

Inclusion criteria comprised (1) population: adults (>18 years) diagnosed with chronic stroke (>6 months poststroke), exhibiting stable vital signs and alert consciousness; (2) intervention: receiving any form of BCI-based training; (3) control: receiving either sham BCI interventions or conventional rehabilitation therapy (eg, physical therapy, occupational therapy, and treadmill training); (4) outcomes: assessment using the Fugl-Meyer Assessment for upper extremity (FMA-UE), Action Research Arm Test (ARAT), Modified Ashworth Scale (MAS), Modified Barthel Index (MBI), and Motor Activity Log (MAL); (5) study design: randomized controlled trials (RCTs) published in English or Chinese.

Exclusion criteria were (1) nonprimary research publications (eg, reviews, meta-analyses, systematic reviews, and conference abstracts); (2) studies reporting outcome measures inconsistent with those prespecified for analysis; and (3) publications with incomplete or irretrievable data.

Selection Process

All records identified through the database searching were imported into EndNote X9 (Clarivate Analytics) for management. The total number of records retrieved from each database and information source was documented. Duplicates were removed using a 2-step process: first, automatically using EndNote's built-in deduplication feature, followed by a manual verification conducted independently by 2 investigators (HJC and GJY). The screening process was then carried out in 2 stages; first, titles and abstracts were screened against the inclusion criteria; second, the full texts of potentially eligible records were retrieved and assessed. Both investigators independently evaluated the studies at each stage. Any disagreements regarding study selection were resolved through discussion.

Data Extraction

Two investigators (HJC and GJY) independently reviewed the full texts of included studies. Data extraction was performed independently by both reviewers, capturing key details, including first author's name, the age of the participants, time after stroke, the number of participants by different hemiplegic sides and stroke types, publication year, BCI signal acquisition method, feedback device, sample size, intervention details, intervention duration, outcome measures, and follow-up period. For studies reporting outcomes solely as median and IQR, these values were converted to estimated mean and SD using the methods recommended by the Cochrane Handbook [26]: mean ≈ median; SD ≈ IQR / 1.35 [9]. Any discrepancies between reviewers were initially resolved through discussion. Persistent disagreements were resolved through discussion.

Quality Assessment

According to the PRISMA guidelines [23], the risk of bias assessment was conducted for the included RCTs. Two investigators (HJC and GJY) independently assessed the methodological quality and risk of bias for included studies using the Cochrane Risk of Bias tool (RoB 2.0) [27]. The RoB 2.0 evaluates five domains: (1) bias arising from the randomization process; (2) bias due to deviations from intended interventions; (3) bias due to missing outcome data; (4) bias in outcome measurement; and (5) bias in selection of the reported result. Studies were rated as superior if all domains were judged at low risk, indicating minimal bias concerns. Studies with some domains at low risk but others raising concerns were rated good, reflecting a moderate risk of bias. Studies where no domain achieved low risk were rated poor, indicating substantial bias concerns. Disagreements between assessors were resolved through discussion.

Additionally, the overall quality of evidence for each primary outcome was assessed using the GRADE (Grading of Recommendations Assessment, Development and Evaluation) framework [28]. Two reviewers (HJC and GJY) independently evaluated the evidence quality for the following outcomes. The GRADE approach considers 5 domains, including risk of bias,

XSL•FO
RenderX

inconsistency, indirectness, imprecision, and publication bias. Evidence quality was categorized as high, moderate, low, or very low. Any discrepancies in ratings were resolved through discussion.

Statistical Analyses

Statistical analysis was performed using Stata 18 (StataCorp LLC) software. Data were pooled using mean differences (MDs) with 95% CIs to assess BCI training efficacy. The extent of heterogeneity was quantified using the tau-square (τ^2) statistic, which estimates the variance of true effect sizes across studies. The I^2 statistic is reported as a supplementary measure, representing the percentage of total variability in effect estimates that is due to heterogeneity rather than sampling error [29,30]. The Cochran Q test (chi-square test) was used to test the null hypothesis of homogeneity; a significant P value ($P<.05$) was taken as evidence of the presence of heterogeneity [31]. Given the anticipated clinical and methodological diversity among the included studies, all meta-analyses were performed using the random-effects model. To ensure more robust and conservative estimates, especially given the varying number of studies included in different analyses, we applied the Hartung-Knapp-Sidik-Jonkman adjustment for calculating the 95% CIs around the pooled MDs [32]. This model provides a more conservative estimate of the effect size and its CI by accounting for both within-study and between-study variability. To quantify the implications of heterogeneity and estimate the range within which the true effect size is likely to fall in future similar scenarios, we calculated 95% prediction intervals (PIs) for the meta-analyses of all outcome measures included in this systematic review so as to ensure the reliability of the estimation of between-study variance [33].

The outcome measures of our meta-analysis included upper limb motor function, muscle tone, and ADL. The primary outcome measure was the FMA-UE, which is a commonly used indicator for clinical assessment of upper limb dysfunction in people who have experienced a stroke [34]. Specifically, the ARAT served as the assessment scale for upper limb function, the MAS was used for muscle tone evaluation, and both the MBI and MAL were used to assess ADL in people who have experienced a stroke.

Subgroup analyses were performed to assess the influence of the following potential factors on upper limb functional outcomes:

1. Feedback modality: BCI-FES, BCI-robot, or BCI-visual feedback;
2. Intervention intensity: stratified by session duration (20 min vs 30 min vs 60 min), weekly frequency (2-3 days vs 4-5 days), total intervention duration (2 weeks vs 4-5 weeks vs 8 weeks), and total number of sessions (10-12 times vs 20-24 times);
3. Follow-up time point: short-term (≤3 months) vs long-term (>3 months).

The thresholds for these subgroup classifications were defined based on the most frequently reported intervention parameters across the included studies and common dosing regimens in prior clinical research [35,36]. This categorization allows for a direct comparison of the different training intensities most commonly encountered in the current evidence base.

Sensitivity analysis was performed to investigate potential sources of heterogeneity. Assessment of small-study effects was performed using funnel plots and the Egger test for outcomes that included a sufficient number of studies (typically n>10). A P value <.05 in the Egger test was considered indicative of potential small-study effects [37].

Results

Search Results

The systematic literature search yielded a total of 3529 records from the 6 databases Cochrane Library (n=1173), Embase (n=622), PubMed (n=378), Scopus (n=398), Web of Science (n=546), and Wanfang (n=412). Following the deduplication process as prespecified in the methods, 1534 duplicate publications were removed. The remaining 1995 unique records underwent initial title and abstract screening. Subsequently, 152 articles were selected for full-text assessment, from which 21 studies met the predefined inclusion criteria and were included in the final meta-analysis. The study selection process is detailed in the PRISMA flow diagram (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the study selection process.

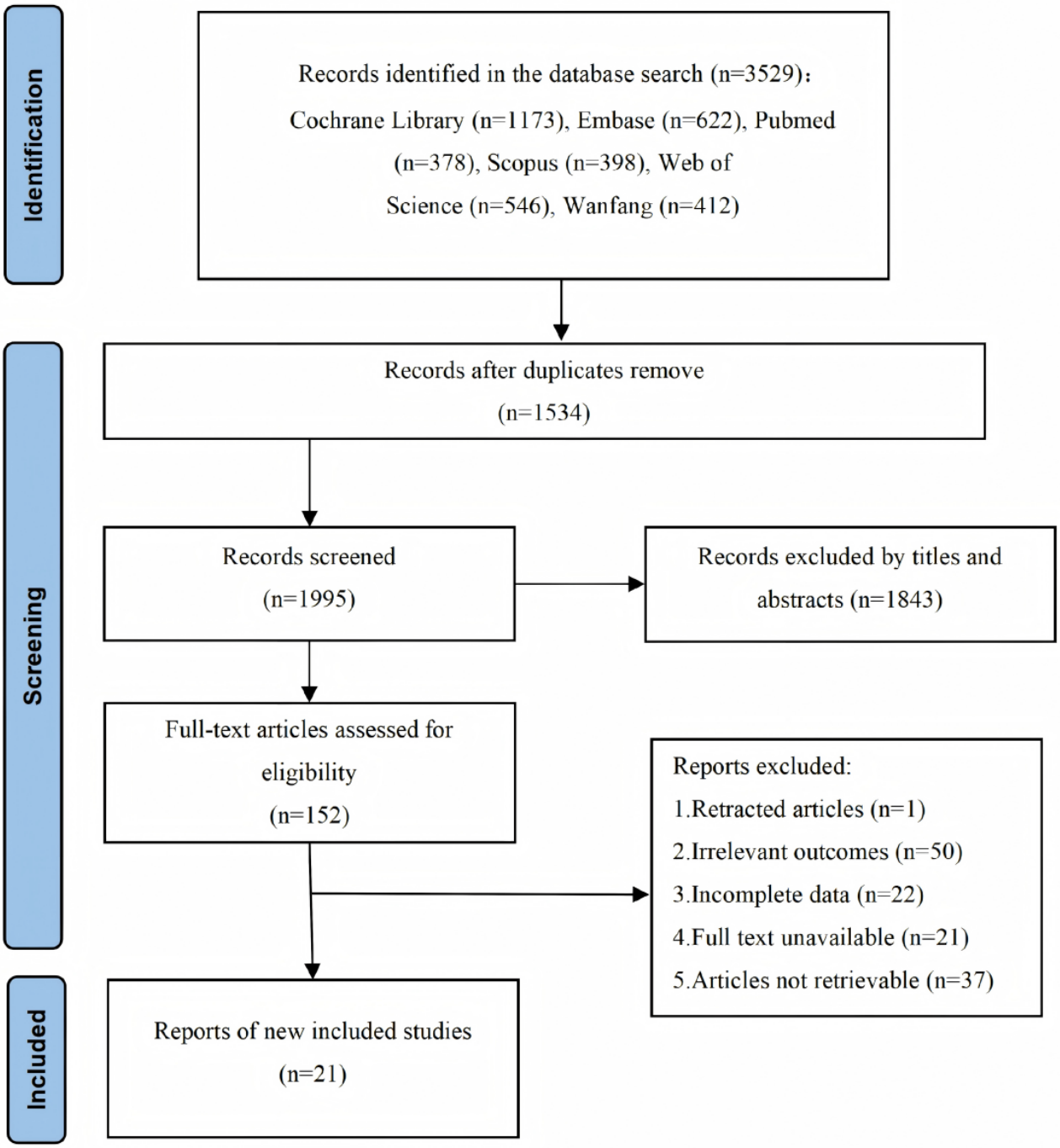

Quality Evaluation

The methodological quality of the 21 included RCTs was assessed using the revised RoB 2. Based on this assessment, 6 RCTs were classified as "Superior," 6 as "Good," and 9 as "Poor." In several studies, the substantial differences in interventions between the experimental and control groups made blinding of participants and intervention providers not feasible. Consequently, these studies were judged to carry a high risk of bias. The specific evaluation results are shown in Figure 2 [11,19,35,38-55].

XSL•FO
RenderX

Figure 2. Risk of bias assessment for included randomized controlled trials (RoB 2.0) [11,19,35,38-55].

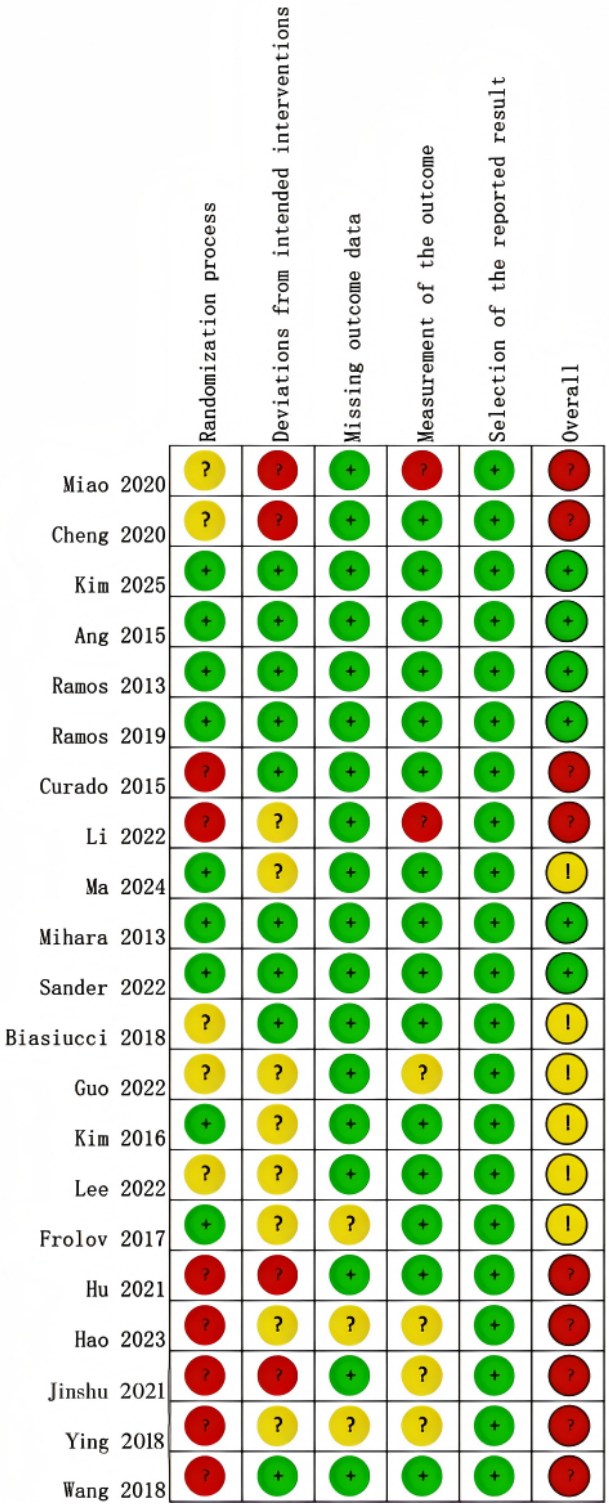

The overall quality of evidence for the primary outcomes, as assessed by the GRADE approach, ranged from low to moderate. The summary of findings and the detailed GRADE evidence profile for each outcome are available in Multimedia Appendix 3.

Characteristics of the Included Literature

A total of 21 studies [11,19,35,38-55], published between 2013 and 2025, were included in this meta-analysis. These studies comprised 337 participants in experimental groups and 313 in control groups. Individual intervention sessions ranged from 20 minutes to 120 minutes, while the total intervention duration varied from 3 days to 10 weeks, encompassing 6-70 sessions in total. Follow-up assessments were performed in 8 studies [11,19,39,42,45-47,55]. Regarding the BCI feedback modality, FES was used in 7 studies [19,38,40,48,49,52,54], exoskeleton devices in 10 studies [11,39,41-44,47,50,53,55], and visual feedback in 4 studies [35,45,46,51]. Specific methods for

XSL•FO
RenderX

random sequence generation were described in 15 studies [11,19,35,40-42,44-46,48-50,52-54]. Allocation concealment was implemented in 8 studies [11,19,40-42,45,46,48], and outcome assessors were blinded in 15 studies [11,19,35,39-43,45,46,48-51,55]. The detailed characteristics of the included studies are presented in Multimedia Appendix 4.

Results of Meta-Analysis

Effect of BCI on Overall Motor Function (FMA-UE)

As illustrated in Figure 3, sixteen [11,19,35,38,40,41,43-49,51,54,55] studies reported the effects of BCI training on FMA-UE scores in people with chronic stroke. The meta-analysis demonstrated a statistically significant improvement in FMA-UE scores following BCI training (MD 2.50, 95% CI 0.60-4.40; *P*=.01). The test for heterogeneity was not statistically significant (Q=18.72; *P*=.23), and the I^2 was 45.18%. The estimated between-study variance was tau^2=6.06. The 95% PI was (–2.52 to 7.22), indicating that the effect of BCI on FMA-UE in a future similar study could range from a clinically irrelevant decline of 2.52 points to a clinically meaningful improvement of 7.22 points.

Figure 3. Forest plot for the effect of brain-computer interface (BCI) training on Fugl-Meyer Assessment for upper extremity (FMA-UE) scores [11,19,35,38,40,41,43-49,51,54,55].

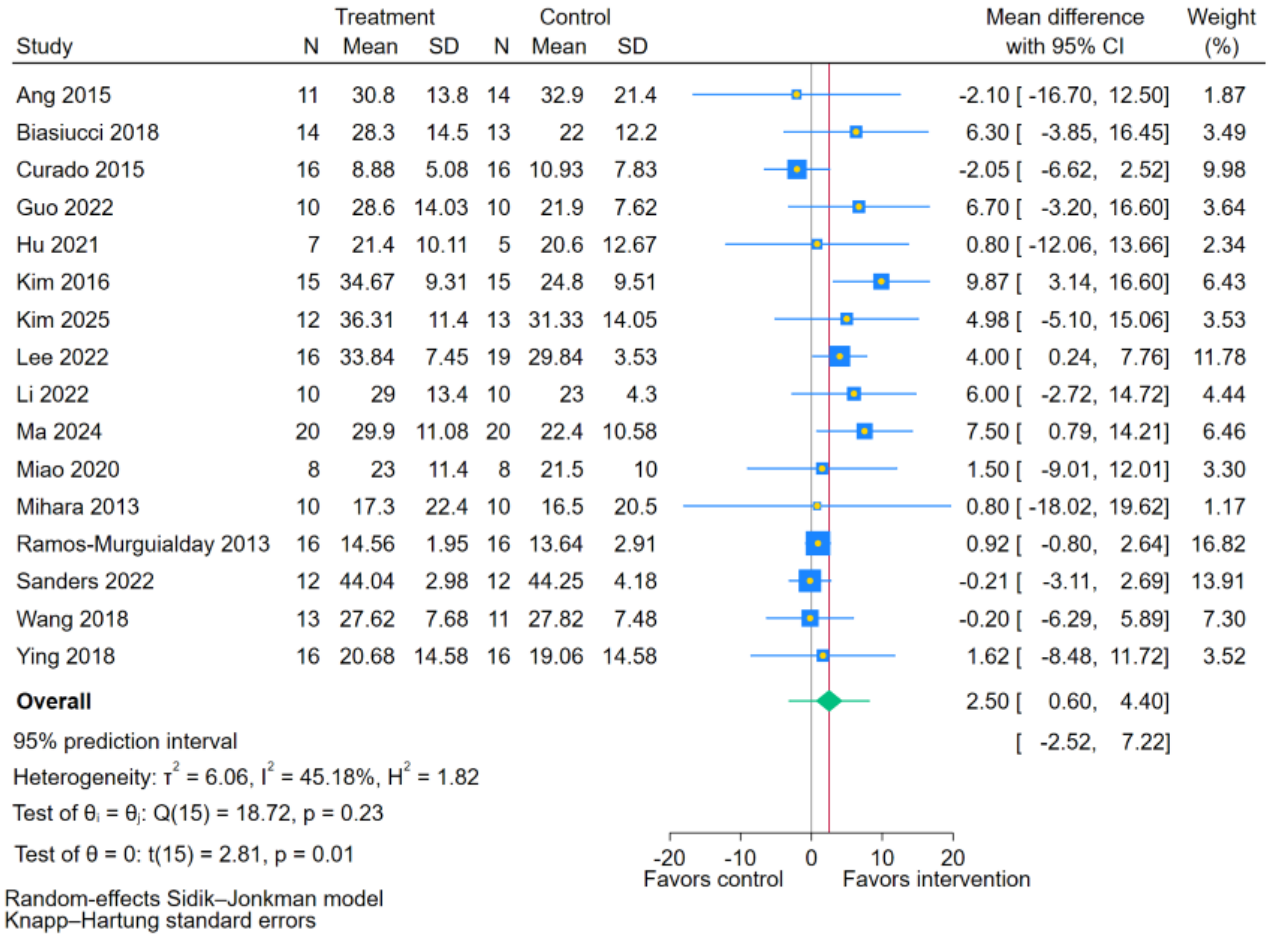

Effect of BCI on Fine Motor Skills (ARAT)

As depicted in Figure 4, four [39,45,50,51] studies evaluated the impact of BCI training on ARAT scores in people with chronic stroke. The meta-analysis revealed no statistically significant difference in ARAT scores between the intervention and control groups (MD=0.18, 95% CI –0.27 to 0.62; *P*=.30). The test for homogeneity was not statistically significant (Q=0.04; *P*=.99). The estimated between-study variance was tau^2=0. The I^2 statistic was 0.03%. Furthermore, the 95% PI was (–3.64 to 3.99). This wide interval, which spans both clinically negligible negative and positive effects, underscores the considerable uncertainty regarding the true effect of BCI on fine motor skills and indicates that in future settings, the outcome could range from a slight worsening to a modest improvement.

XSL•FO
RenderX

Figure 4. Forest plot for the effect of brain-computer interface (BCI) training on Action Research Arm Test (ARAT) scores [39,45,50,51].

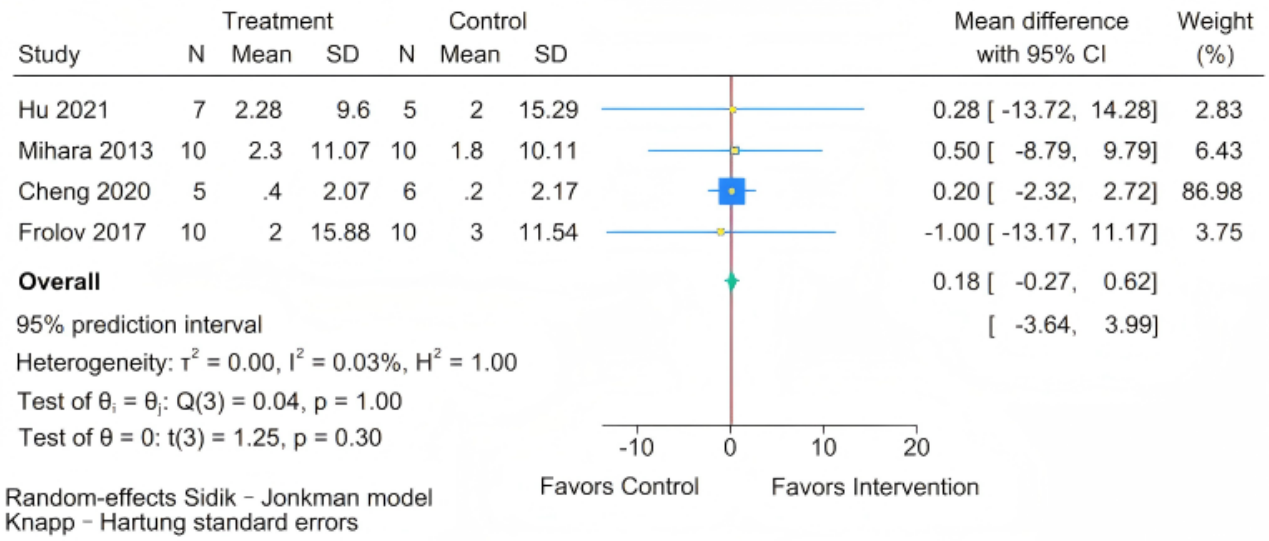

Effect of BCI on Muscle Tone (MAS)

As presented in Figure 5, four [19,41,47,52] studies assessed the effect of BCI training on upper limb muscle tone in people with chronic stroke. The meta-analysis demonstrated no statistically significant difference in muscle tone outcomes between the BCI and control groups (MD –0.48, 95% CI –1 to 0.03; *P*=.06). The test for homogeneity was not statistically significant (Q=4.69; *P*=.20). The estimated between-study variance was tau²=0.05. The I^2 statistic was 42.28%. However, the 95% PI was (–1.27 to 0.35), offering a more nuanced interpretation. As lower scores on the MAS indicate reduced spasticity, this interval suggests that in future clinical settings, the effect of BCI on muscle tone is predicted to range from a small but potentially meaningful reduction to a negligible change.

Figure 5. Forest plot for the effect of brain-computer interface (BCI) training on Modified Ashworth Scale (MAS) scores [19,41,47,52].

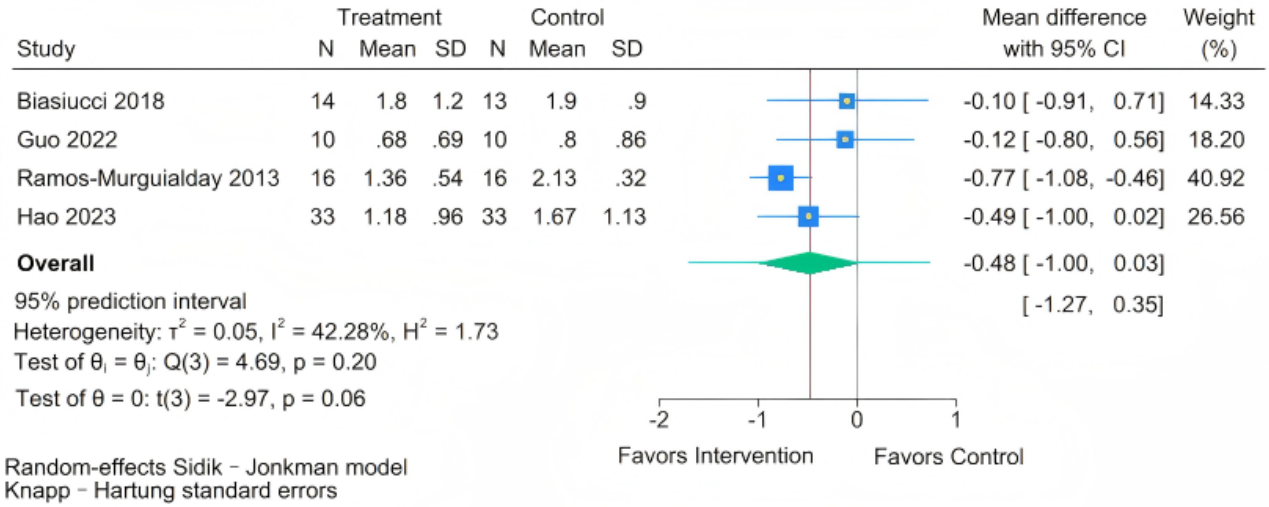

Effects on ADL

Effect of BCI on Activities of Daily Living (MBI)

As illustrated in Figure 6, six [48,49,51-54] studies evaluated the effects of BCI training on MBI scores in people with chronic stroke. The meta-analysis demonstrated a statistically significant improvement in MBI scores following BCI intervention (MD 8.38, 95% CI 2.23-14.53; *P*=.02). The test for homogeneity was not statistically significant (Q=9.85; *P*=.08). However, the I^2 statistic was 63.63%, and the estimated between-study variance was tau²=22.55, indicating substantial heterogeneity in the magnitude of the effect across studies. The 95% PI was (–3.92 to 20.53), suggesting substantial uncertainty in the magnitude of the effect across different clinical settings, with effects potentially ranging from negligible to substantially beneficial.

XSL•FO
RenderX

Figure 6. Forest plot for the effect of brain-computer interface training on Modified Barthel Index (MBI) scores [48,49,51-54].

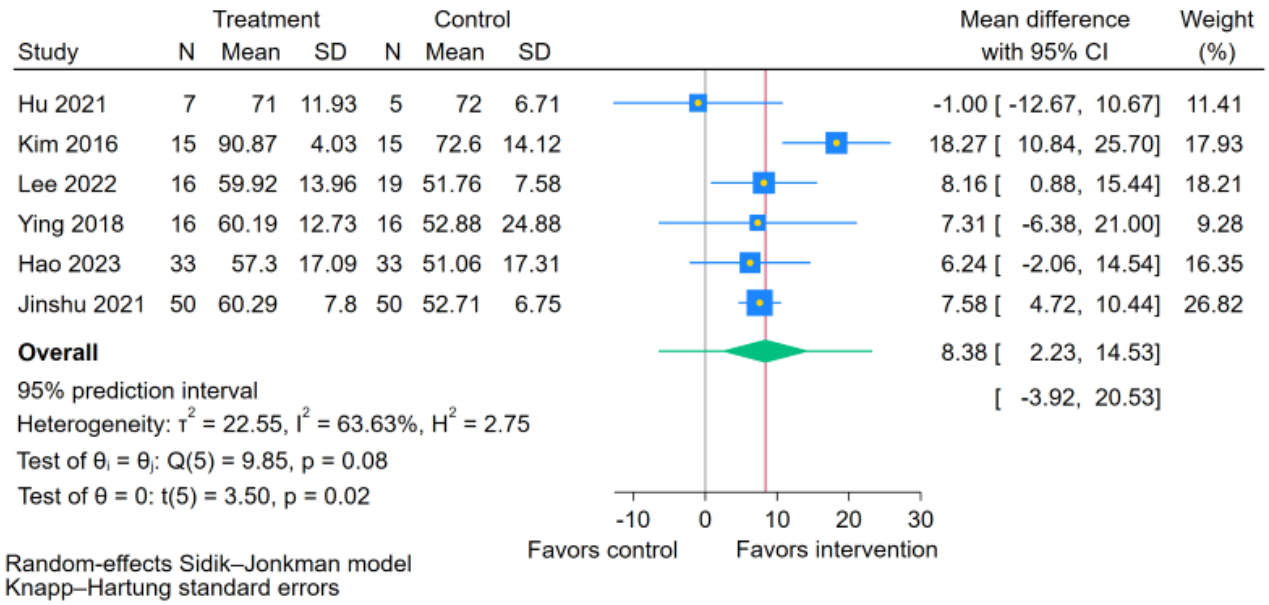

Effect of BCI on Self-Reported Arm Use (MAL)

As shown in Figure 7, five [41,42,45,48,49] studies examined the effect of BCI training on MAL scores in people with chronic stroke. The meta-analysis showed a statistically significant improvement in MAL scores following BCI intervention (MD 2.09, 95% CI 0.42-3.76; *P*=.03). The test for homogeneity was not statistically significant (Q=3.86; *P*=.42). The estimated between-study variance was tau²=1.43. The I^2 statistic was 45.7%. The 95% PI was (–0.69 to 4.54). This indicates that while the average effect is positive, the true effect in a new setting could range from a negligible or slightly negative impact to a substantial improvement in patient-perceived arm use during daily activities. The fact that the majority of the interval lies above zero strengthens the evidence for a likely beneficial effect, albeit of variable magnitude.

Figure 7. Forest plot for the effect of brain-computer interface (BCI) training on Motor Activity Log (MAL) scores [41,42,45,48,49].

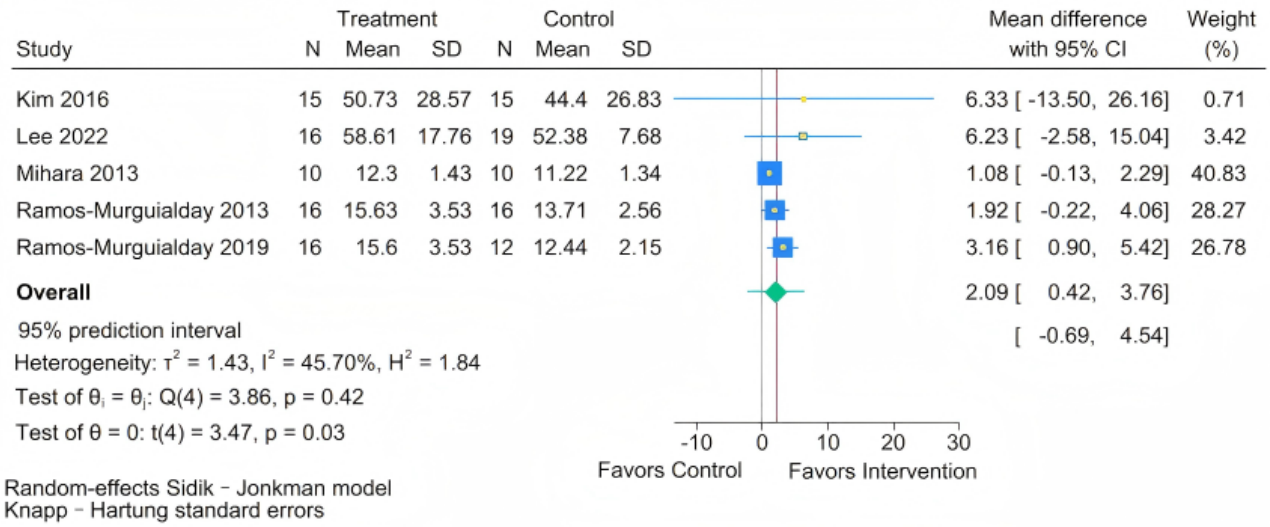

Subgroup Analysis: Feedback

In the subgroup analysis stratified by BCI feedback modality (Figure 8 [11,19,35,38,40,41,43-49,51,54,55]), the test for subgroup differences indicated no statistically significant difference between the modalities (*P*=.21). However, within-subgroup analyses revealed that only the BCI-FES paradigm showed a significant improvement in FMA-UE scores compared to routine rehabilitation therapy (RRT: MD 5, 95% CI 1.86-8.13; *P*=.01). In contrast, neither the BCI-Exoskeleton (MD 0.92, 95% CI –2.31 to 4.15; *P*=.50) nor the BCI-Visual (Beijing Intelligent Brain Science and Technology Co, Ltd) feedback (MD 2.01, 95% CI –4.14 to 8.16; *P*=.37) subgroups demonstrated significant effects. Although the differences between feedback modalities were not statistically significant, BCI-FES may be associated with greater motor recovery relative to RRT than the other modalities.

XSL•FO
RenderX

Figure 8. Subgroup analysis of Fugl-Meyer Assessment for upper extremity (FMA-UE) scores by brain-computer interface (BCI) feedback modality [11,19,35,38,40,41,43-49,51,54,55].

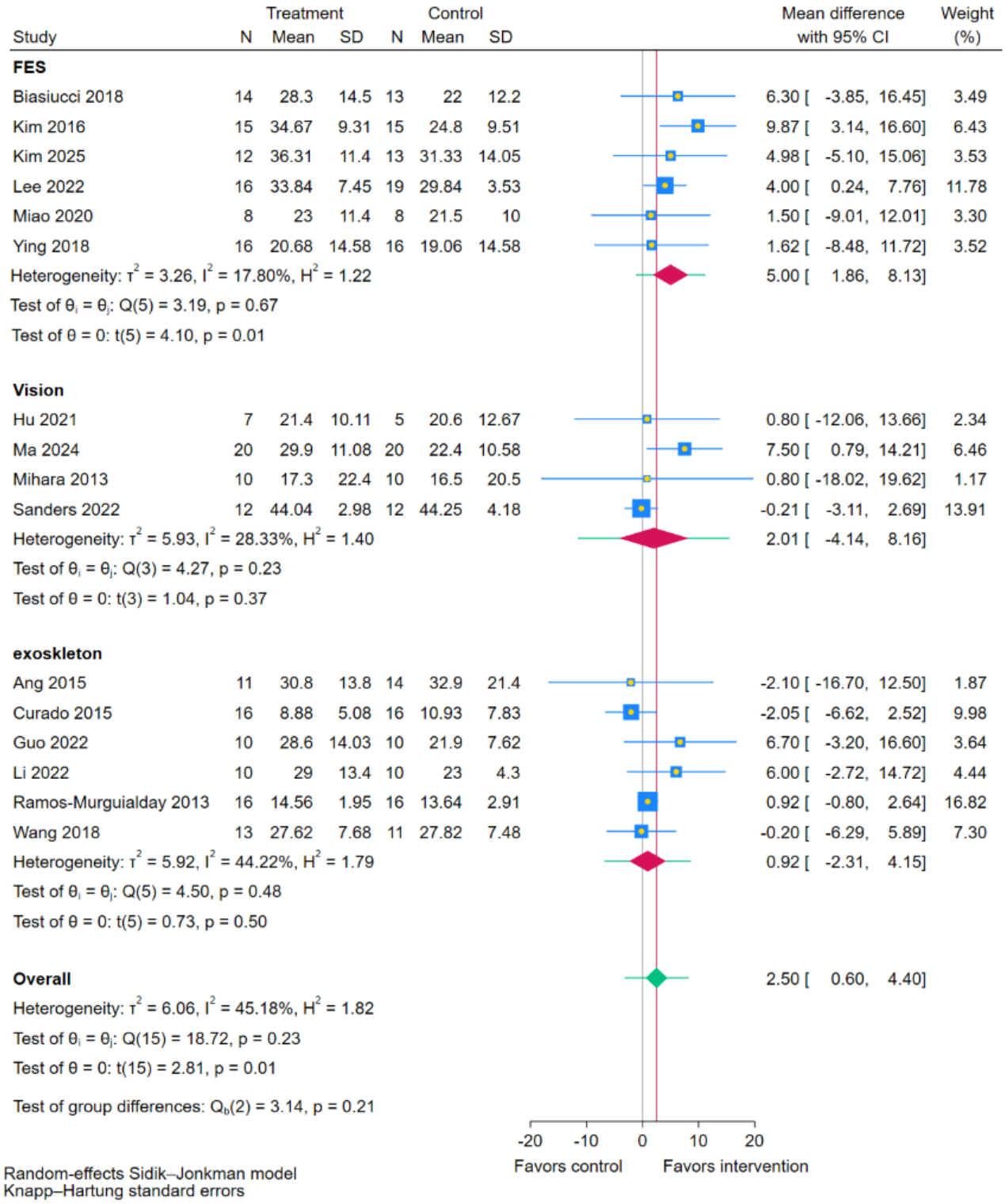

Intervention Intensity: Session Duration

Subgroup analysis based on single-session intervention duration (Figure 9 [11,19,35,38,40,41,43,44,47-49,51,54,55]) demonstrated no statistically significant difference between session duration subgroups (*P*=.16). Within the subgroups, a regimen of 30-minute sessions elicited a significant improvement in FMA-UE scores compared to RRT (MD 4.36, 95% CI 0.28-8.44; *P*=.04), whereas sessions lasting 20 minutes (MD –0.07, 95% CI –1.45 to 1.32; *P*=.85) or 60 minutes (MD 1.90, 95% CI –1.36 to 5.17; *P*=.20) did not. These results suggest that while the difference between session durations was not statistically significant, a 30-minute session may be associated with optimal outcomes.

XSL•FO
RenderX

Figure 9. Subgroup analyses of Fugl-Meyer Assessment for upper extremity (FMA-UE) scores by session duration [11,19,35,38,40,41,43-49,51,54,55].

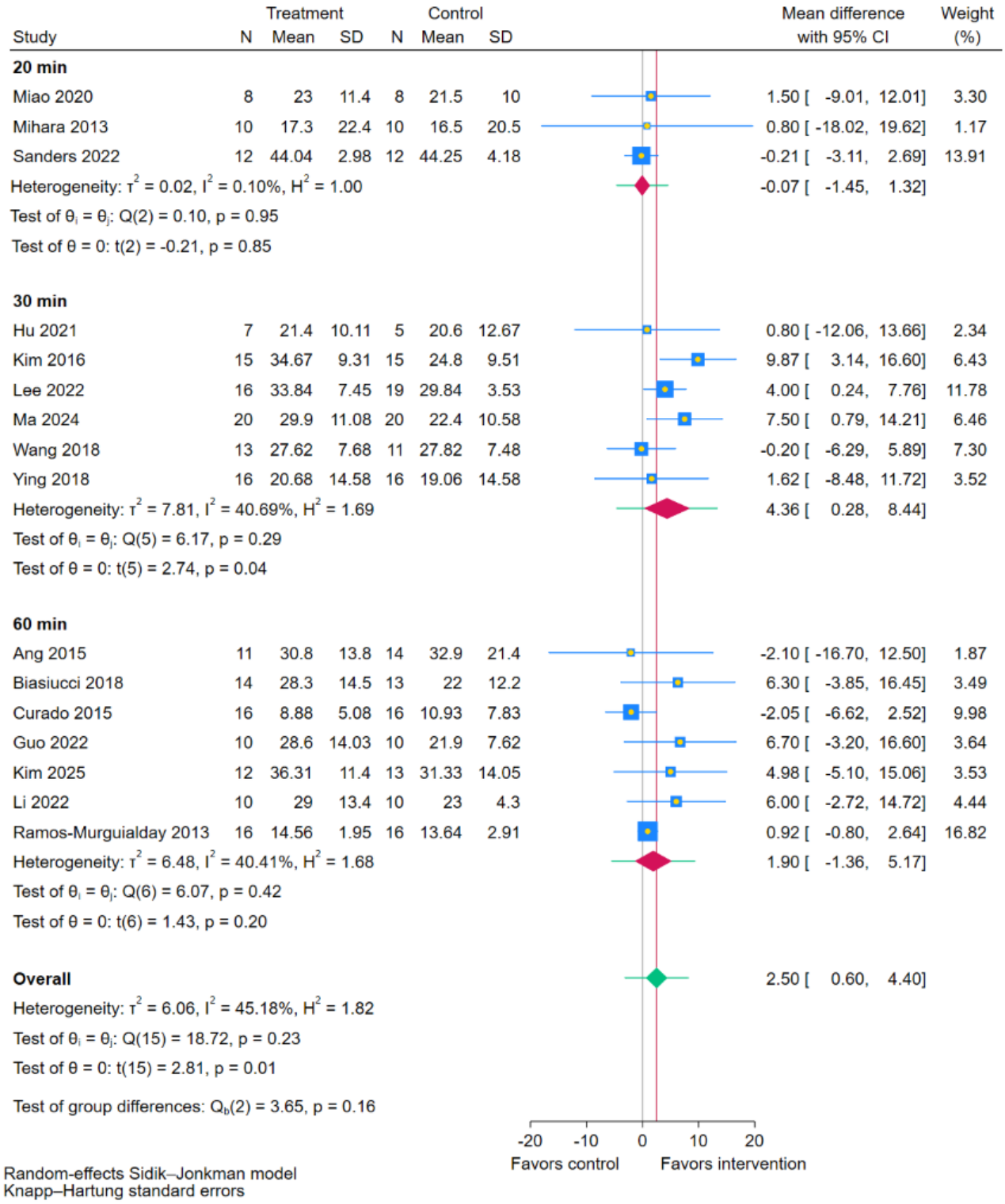

Training Sessions per Week

Subgroup analysis based on weekly intervention frequency (Figure 10 [11,19,35,38,40,41,43,44,47-49,51,54,55]) found no statistically significant difference between the frequency subgroups (*P*=.22). The higher-frequency regimen (4-5 sessions per week) was associated with a significant improvement in FMA-UE scores compared to RRT (MD 3.20, 95% CI 0.42-5.97; *P*=.03). The lower-frequency regimen (2-3 sessions per week) did not show a significant effect (MD 0.61, 95% CI –1.62 to 2.85; *P*=.51). This indicates that while the difference between weekly frequencies was not statistically significant, a higher frequency of 4-5 sessions per week may be linked to better motor recovery.

XSL•FO
RenderX

Figure 10. Subgroup analyses of Fugl-Meyer Assessment for upper extremity (FMA-UE) scores by training sessions per week [11,19,35,38,40,41,43-49,51,54,55].

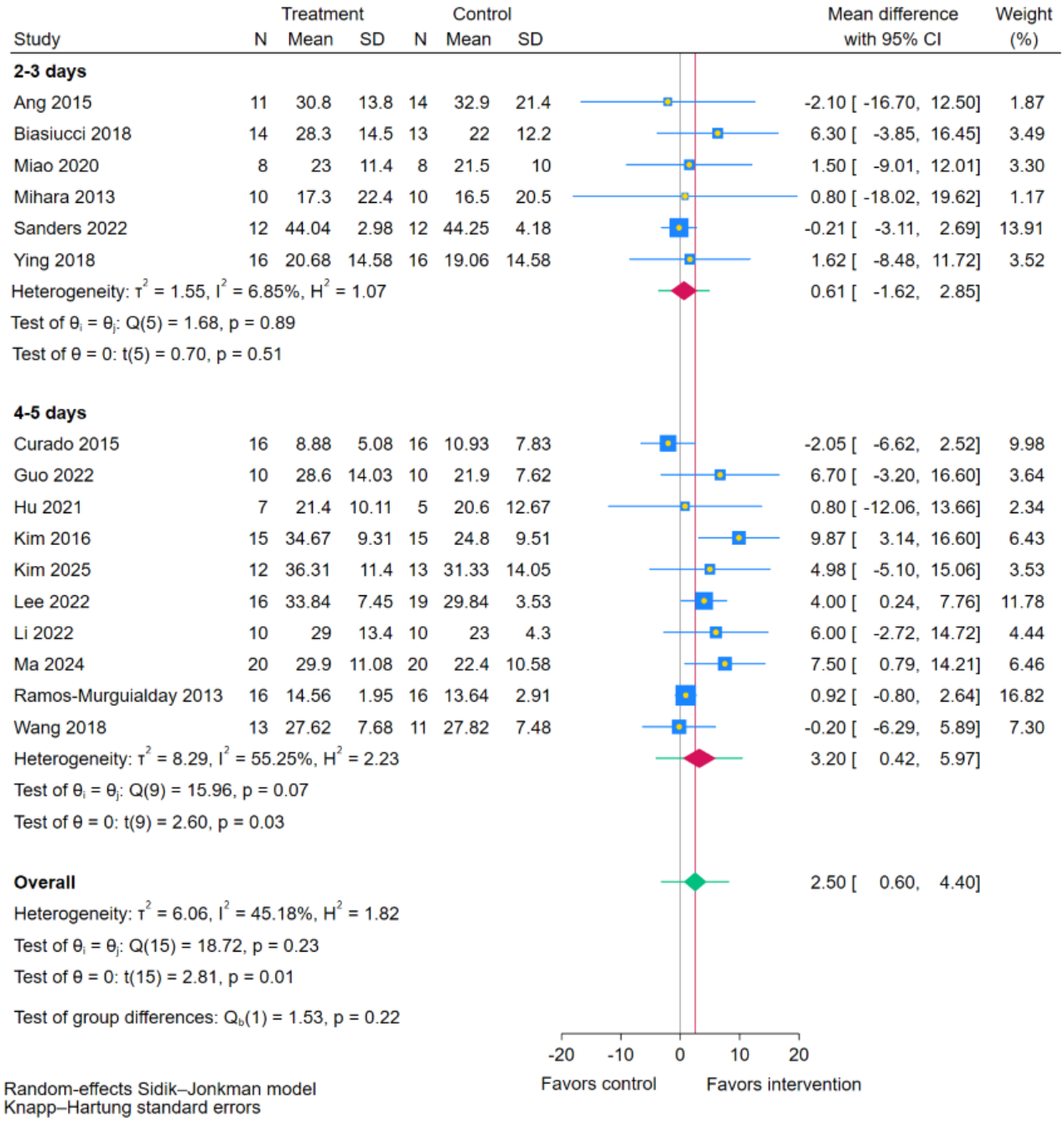

Duration of Intervention

Subgroup analysis based on the total intervention duration (Figure 11 [11,19,35,38,40,41,43-45,47-49,51,54,55]) showed no statistically significant difference between the duration subgroups (*P*=.39). Within the subgroups, a shorter-duration regimen of 2 weeks elicited a significant improvement in FMA-UE scores compared to RRT (MD 6.67, 95% CI 1.04-12.31; *P*=.04). Interventions lasting 4-5 weeks (MD 2.46, 95% CI 0.01-4.92; *P*=.05) or 8 weeks (MD 1.62, 95% CI –8.48 to 11.72) did not show significant effects. This suggests that although the difference between intervention durations was not statistically significant, a shorter, more concentrated 2-week period may be associated with superior efficacy.

XSL•FO
RenderX

Figure 11. Subgroup analyses of Fugl-Meyer Assessment for upper extremity (FMA-UE) scores by duration of intervention [11,19,35,38,40,41,43-45,47-49,51,54,55].

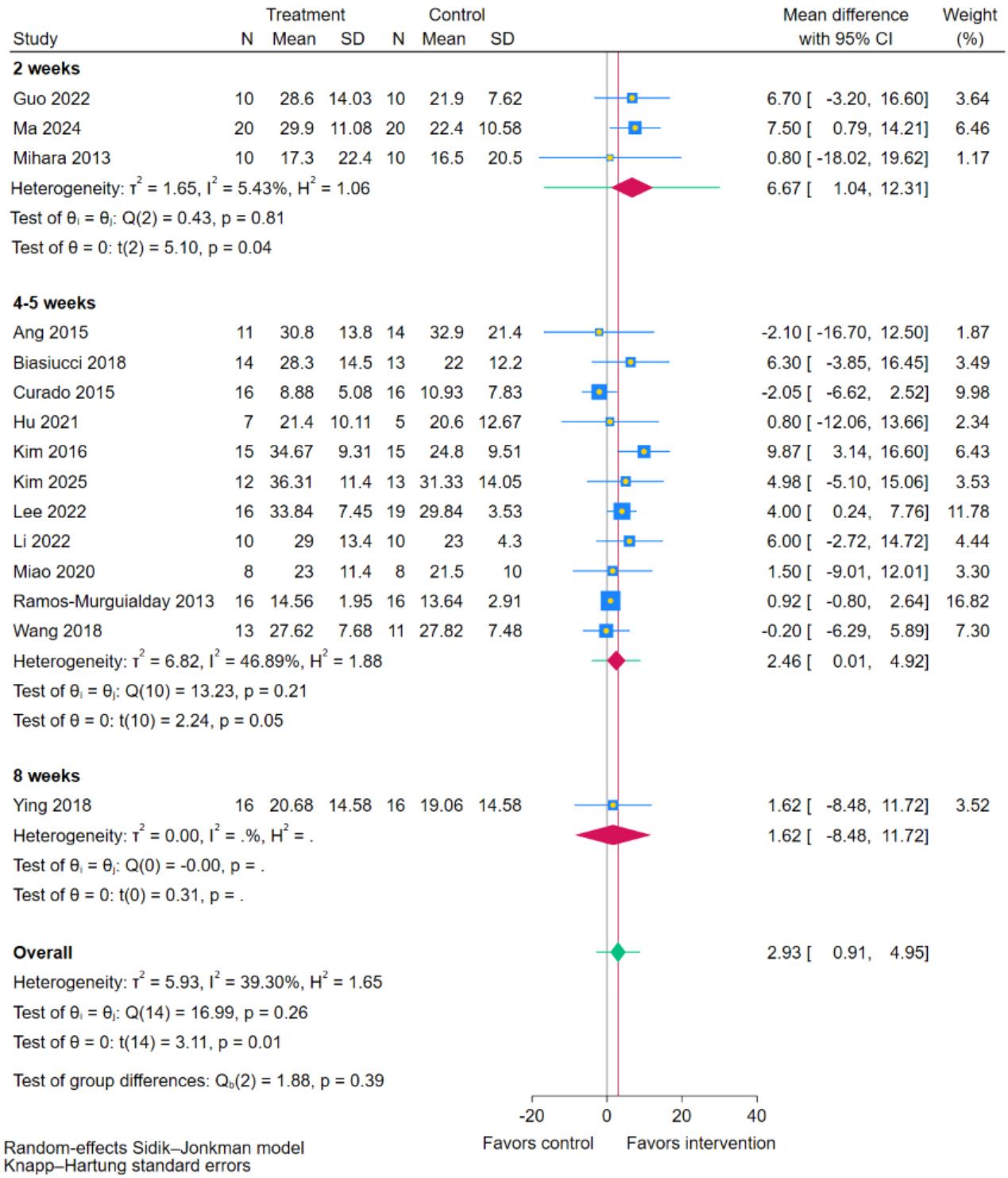

Total Number of Sessions

Subgroup analysis stratified by the total number of intervention sessions (Figure 12 [11,19,35,38,40,41,43,44,47-49,51,54,55]) indicated no statistically significant difference between the session count subgroups (*P*=.32). A lower total session count (10-12 sessions) was associated with a significant improvement in FMA-UE scores compared to RRT (MD 5.16, 95% CI 0.76-9.56; *P*=.03). A higher session count (20-24 sessions) did not demonstrate a significant effect (MD=2.40, 95% CI –0.39 to 5.20; *P*=.08). These findings imply that while the difference between total session numbers was not statistically significant, a protocol comprising 10-12 sessions may be linked to more favorable outcomes.

XSL•FO
RenderX

Figure 12. Subgroup analyses of Fugl-Meyer Assessment for upper extremity (FMA-UE) scores by total number of sessions [11,19,35,38,40,41,43,44,47-49,51,54,55].

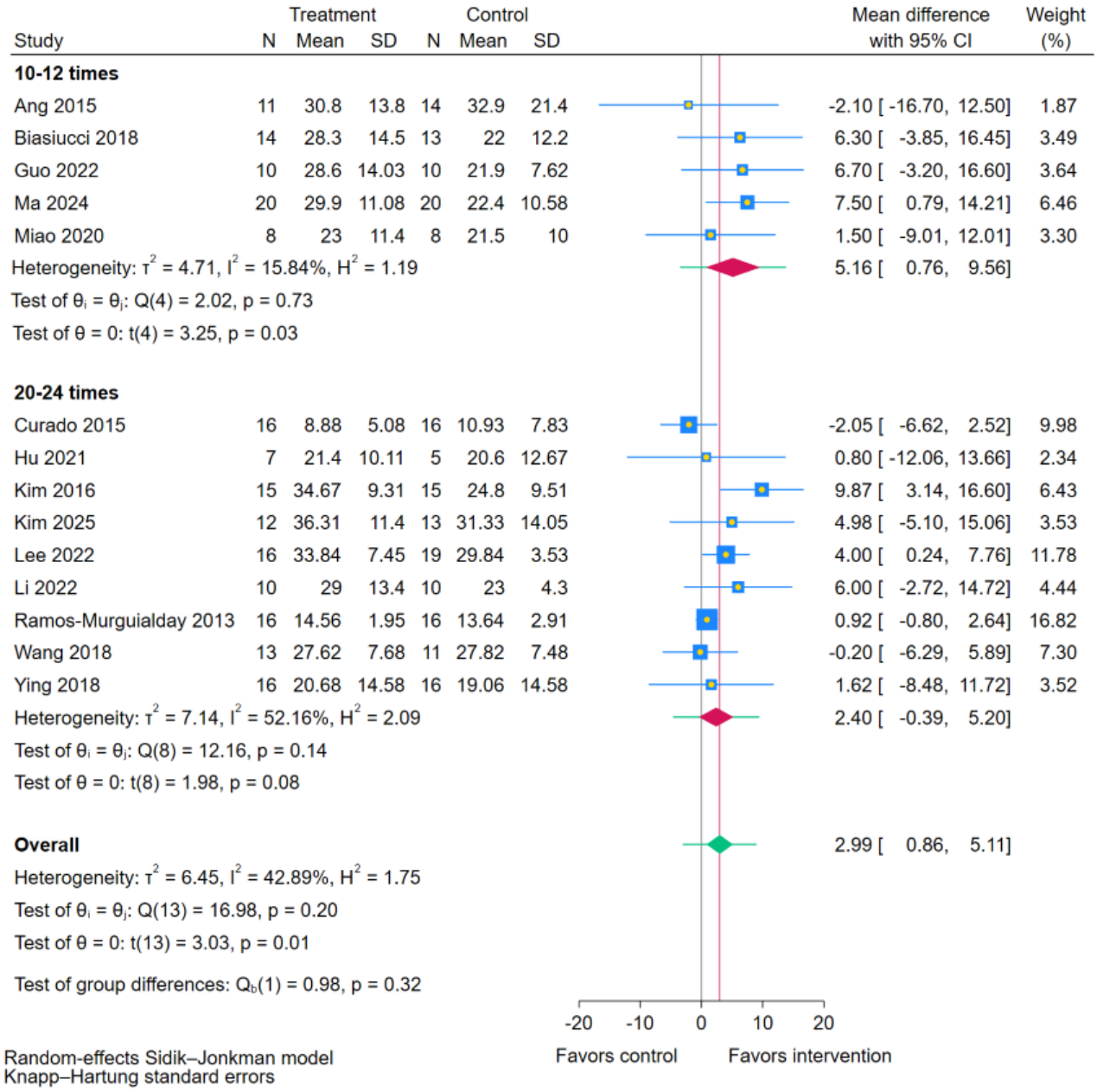

Follow-Up

Analysis of follow-up outcomes (Figure 13 [11,19,41,45,47,55]) revealed no significant differences in long-term FMA-UE improvement between BCI-based training and RRT at any follow-up interval, whether assessed at short-to-medium term (≤3 months; MD 3.23, 95% CI –8.75 to 15.22; *P*=.37) or long-term (>3 months; MD 1.19, 95% CI –4.06 to 6.43; *P*=.43).

XSL•FO
RenderX

Figure 13. Subgroup analysis of Fugl-Meyer Assessment for upper extremity (FMA-UE) scores by follow-up period [11,19,41,45,47,55].

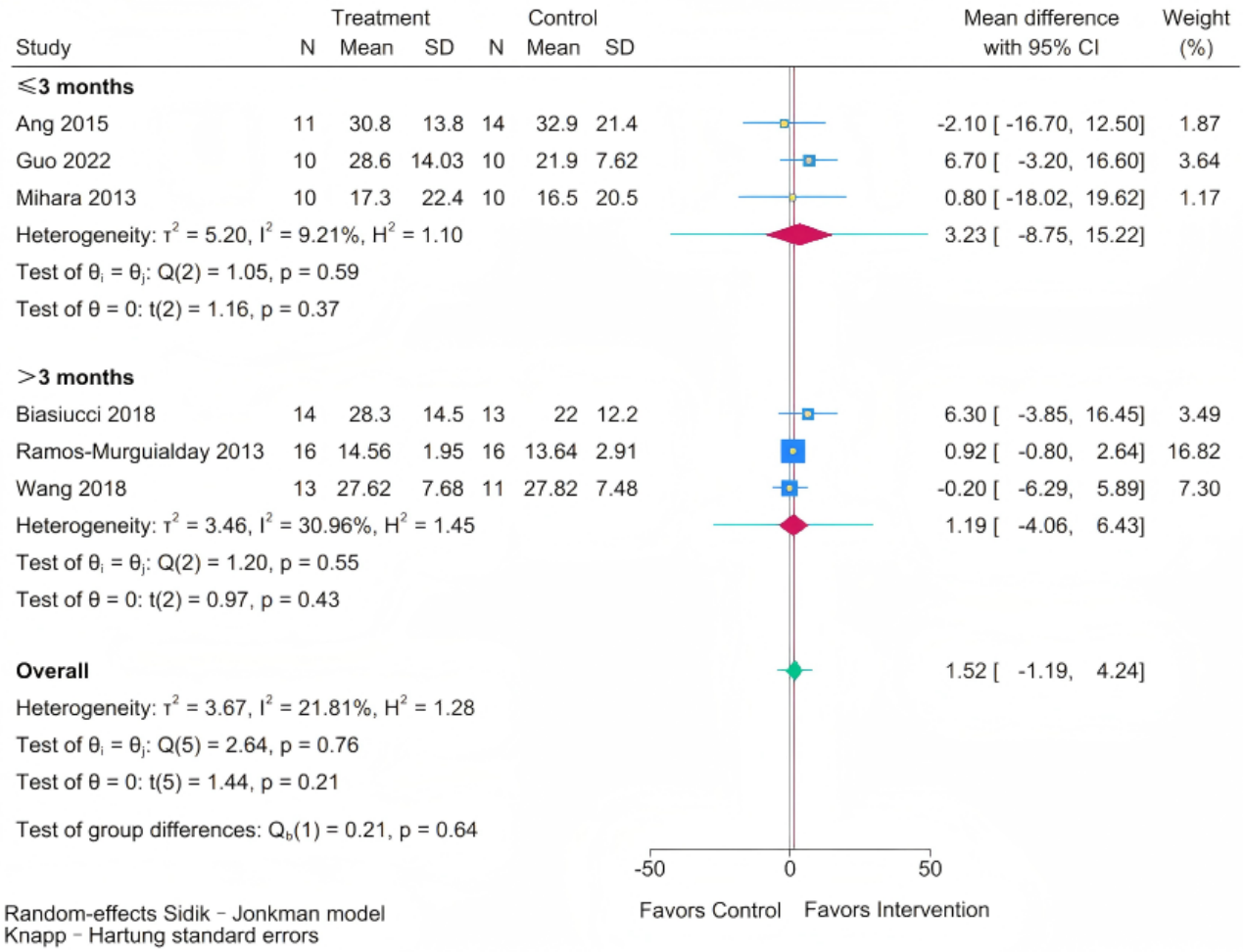

Sensitivity Analysis

A leave-one-out sensitivity analysis was performed to assess the robustness of the pooled MBI results and to investigate the influence of individual studies on the substantial observed heterogeneity (initial I^2=63.63%; τ^2=22.55).

As shown in Figure 14 [49,51-54], the sequential exclusion of each study revealed that the findings were robust overall. However, the exclusion of a single study—Kim et al [48]—led to a marked reduction in heterogeneity, with the I^2 statistic decreasing from 63.63% to 22.66% and the tau² value dropping from 22.55 to 4.03. Crucially, the pooled estimate remained statistically significant and the CI narrowed, indicating increased precision (MD 6.77, 95% CI 3.45-10.09; P<.001), compared to the original analysis (MD 8.38, 95% CI 2.23-14.53; P=.02). This suggests that while the study by Kim et al [48] was a major contributor to the statistical heterogeneity, the conclusion that BCI training improves ADL is robust.

XSL•FO
RenderX

Figure 14. Sensitivity analysis for the effect of brain-computer interface (BCI) training on Modified Barthel Index (MBI) scores [49,51-54].

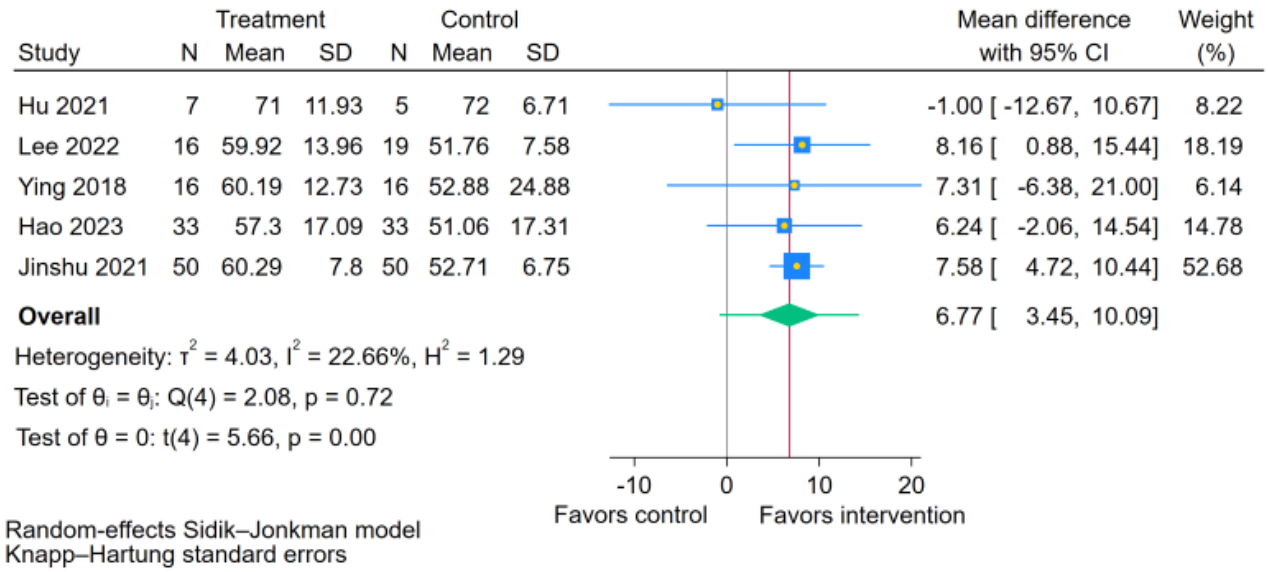

Small-Study Effects Analysis

Within this analysis, only the FMA-UE outcome pooled a sufficient number of studies (n>10) for the assessment of small-study effects. Accordingly, a funnel plot was constructed for the FMA-UE outcome (Figure 15). Visual inspection revealed no substantial asymmetry. Furthermore, the Egger regression test yielded a nonsignificant result (*P*=.20), suggesting no strong evidence for small-study effects for this outcome.

Figure 15. Funnel plot assessing small-study effects for the Fugl-Meyer Assessment for upper extremity (FMA-UE) outcome.

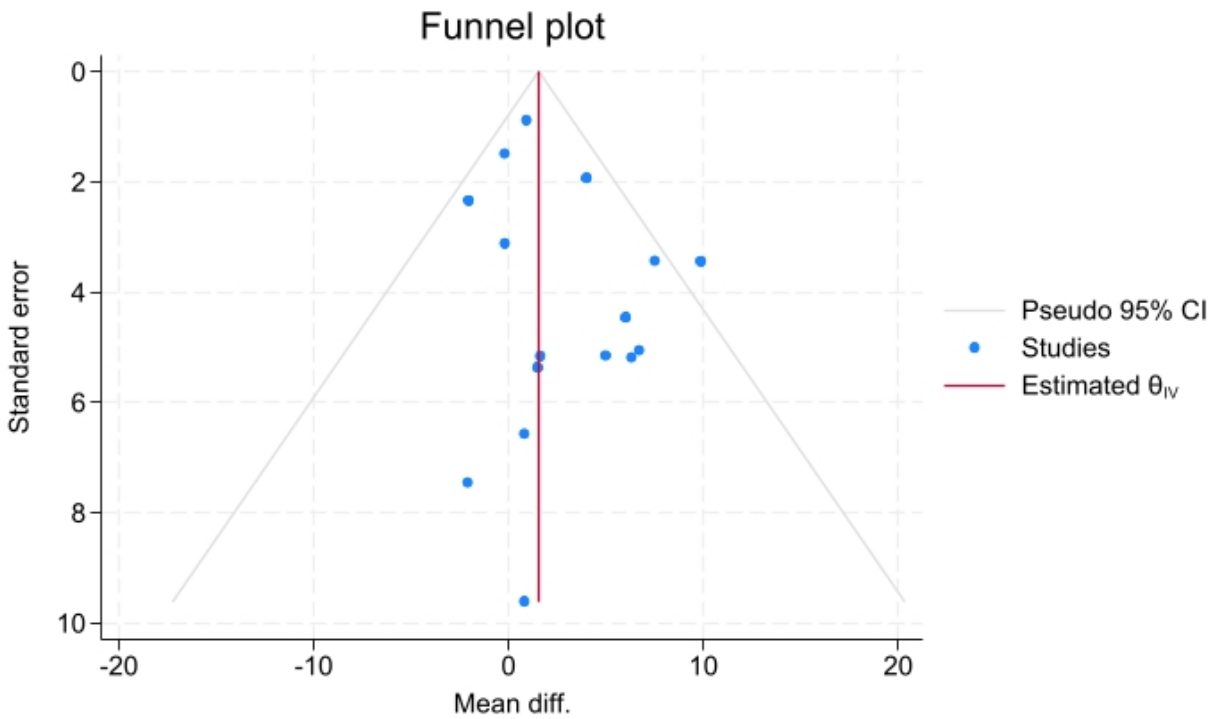

Discussion

Overview

This meta-analysis differs from the general efficacy evaluation of BCI in the treatment of chronic stroke. By conducting an in-depth analysis of how treatment parameters influence prognosis, it has deepened the understanding of this therapeutic approach. While the overall benefits of BCI for upper limb motor function and ADLs are confirmed, our meta-analysis attempts to offer greater clinical utility by identifying the specific feedback modality (BCI-FES) and a distinct, time-efficient training schedule associated with optimal recovery. These findings contribute to optimizing BCI intervention protocols and promoting their clinical

XSL•FO
RenderX

implementation, advancing the translation of this technology between experimental applications and real-world clinical practice.

Summary of Main Findings

The meta-analysis findings demonstrate that BCI training yields benefits in improving overall upper limb motor function and ADLs in patients with chronic stroke. Specifically, BCI interventions led to significant improvements in the primary outcome of motor impairment (FMA-UE) and in patient-reported and performance-based measures of daily function (MBI and MAL). In contrast, BCI training did not yield significantly superior effects on fine motor function (ARAT) or muscle tone (MAS) compared to control interventions.

Clinical Significance and Heterogeneity of Effects

Building on previous meta-analyses [9,56], our findings reveal a statistically significant incremental improvement in FMA-UE scores over conventional therapy. While this gain alone falls below the minimal clinically important difference threshold of approximately 5 points [57], it represents a meaningful augmentation to the foundational improvements from standard care. This is particularly noteworthy given the significant enhancements in MBI and MAL scores. This convergence suggests that BCI's closed-loop methodology, integrating central neural signals with peripheral feedback, may provide a synergistic effect that more effectively translates into functional gains in daily activities.

Furthermore, while the 95% CIs for the FMA-UE, MBI, and MAL confirm a positive average benefit of BCI over control interventions, the prediction intervals (FMA-UE: –2.52 to 7.22; MBI: –3.92 to 20.53; MAL: –0.69 to 4.54) reveal a more complex scenario. These intervals suggest that in future clinical settings, the effect of BCI compared to RRT could range from negligible or even slightly adverse to substantial improvements meeting the minimal clinically important difference. This indicates that current BCI training may be insufficient to yield reliable therapeutic effects for all patients in the chronic phase. The considerable heterogeneity observed suggests that the efficacy of BCI training is not uniform and may be influenced by individual patient conditions or differences in BCI treatment protocols. In line with this, a study by Guo et al [47] also emphasizes that future research should prioritize identifying patient subgroups most likely to benefit from this therapy, specifically, those with effect sizes at the upper end of the prediction interval, such as patients retaining partial integrity of the corticospinal tract.

Lack of Effect on Fine Motor Control and Spasticity

The absence of significant improvement in fine motor function, as measured by the ARAT, and in muscle tone, assessed by the MAS, warrants further mechanistic consideration. The ARAT primarily evaluates distal upper limb functions, such as grasp, grip, and pinch. The nonsignificant findings may stem from a fundamental limitation of current BCI paradigms, which often decode neural correlates of gross motor imagery (eg, whole-arm reaching or hand opening/closing) rather than the finely graded, individuated movements required for dexterous tasks. The feedback provided, particularly via exoskeleton or visual modalities, may lack the specificity necessary to engage and reinforce the delicate cortical representations of the hand.

The PI for the ARAT (–3.64-3.99) provides a deeper perspective on this null result. This interval is not only symmetrically distributed around the null effect but also entirely excludes the possibility of any large, clinically meaningful positive effects. However, given that only 6 of the included studies were rated as having a low risk of bias and the GRADE assessment indicated moderate-quality evidence for the ARAT outcome, this conclusion must be interpreted with caution.

Consequently, clinicians should be cautious about prioritizing the improvement of fine motor function as a primary goal when applying current mainstream BCI paradigms. Future optimization of BCI systems should focus on enhancing the decoding resolution of finer motor intentions, potentially by using high-density EEG or hybrid BCI approaches, and by integrating hand-specific training adjuncts such as virtual reality environments with object manipulation or wearable devices providing tactile or proprioceptive feedback to the distal limb.

Similarly, the lack of a significant effect of BCI on muscle tone suggests that its primary mechanism of action likely involves facilitating active motor control and cortical reorganization, rather than directly modulating the spinal reflex pathways underlying spasticity. In chronic stroke, hypertonia is often well-established, necessitating targeted interventions. While BCI can promote Hebbian plasticity through the associative pairing of motor intention and movement execution, this effect may be insufficient to reverse impaired supraspinal inhibitory control over the spinal motor pool.

Nevertheless, the 95% PI for the MAS (–1.27 to 0.35) provides valuable clinical insight. As lower scores indicate reduced spasticity, this interval—spanning from "no change" to "improvement"—suggests that BCI therapy is unlikely to exacerbate spasticity in future applications. Moreover, its lower bound of –1.27 indicates that under specific conditions, such as when combined with certain forms of FES, BCI may yield meaningful reductions in muscle tone. This potentially "non-harmful" profile, particularly when considered alongside the moderate quality of the existing evidence, represents an important factor for clinical decision-making and offers a preliminary rationale for exploring BCI as a component of comprehensive spasticity management protocols.

Superiority of BCI-FES and the Role of Feedback Modality

Subgroup analyses revealed that only the BCI-FES paradigm demonstrated significantly greater improvement in FMA-UE compared to control. This superiority can be explained through the lens of neuroplasticity and sensorimotor integration. The BCI-FES paradigm creates a closed-loop system that tightly couples motor intention with peripheral afferent feedback. According to Hebbian learning principles [58], which posit that "neurons that fire together, wire together," the synchronous activation of the motor cortex (during attempted movement imagery) and the somatosensory cortex (via FES-induced limb movement and proprioceptive input) strengthens the synaptic

XSL•FO
RenderX

connections within the sensorimotor network, aligning with findings from the meta-analysis by Li et al [9].

The lack of significant benefit with BCI-Exoskeleton may relate to factors such as device complexity, comfort limitations, and suboptimal anatomical fit. Furthermore, excessive robotic assistance could potentially reduce patient engagement and diminish the crucial training effect driven by active neural effort [59]. The negative results with BCI-Vision indicate that visual feedback alone, in the absence of concomitant somatosensory input and actual limb movement, may have limited efficacy in driving neural reorganization and functional recovery, particularly in patients with chronic deficits or more severe functional impairments.

Sustainability of Benefits and the Need for Maintenance

Subgroup analysis based on follow-up duration revealed no significant sustained advantage of BCI over control after the active treatment phase ceased. This implies that the functional gains may lack long-term stability without ongoing application.

In chronic patients, BCI may effectively induce neuroplasticity during the intensive training period, potentially by reinforcing specific neural pathways or facilitating compensatory mechanisms. However, these newly formed connections or patterns might lack sufficient stability or robustness. Following intervention cessation, without ongoing functional application or specific maintenance training, these acquired neural adaptations may gradually regress or weaken. Conversely, standard RRT protocols often inherently incorporate recommendations for continued activity.

These findings highlight the necessity for systematic maintenance strategies—such as telerehabilitation, behavioral incentive programs, or continued use of assistive technologies—to be integrated postintervention. Addressing this challenge of sustained efficacy represents a crucial future research direction.

Toward an Optimal and Efficient Intervention Protocol

Although tests for subgroup differences were not statistically significant, the within-subgroup comparisons revealed a consistent and clinically meaningful pattern favoring specific parameters. The most robust finding is the efficacy of a short-term, high-density protocol.

This paradigm—comprising sessions of approximately 30 minutes each, delivered 4-5 times per week over a total of 10-12 sessions (approximately 2 weeks)—yielded optimal FMA-UE outcomes compared to control interventions. While broader intensity subgroup comparisons (eg, session duration or total weeks) were not statistically significant, this specific, condensed protocol was the only intensity paradigm that consistently demonstrated a significant within-subgroup effect. This finding underscores the potential primacy of training density—the concentration of practice within a shorter timeframe—over the total intervention duration [50]. Furthermore, a higher frequency of 4-5 sessions per week proved superior to regimens of 2-3 sessions per week, indicating that more frequent exposure facilitates sharper motor patterns and stronger memory traces [60].

Notably, completing 10-12 sessions within 2 weeks was more effective than protocols delivering 20-24 sessions over 4-5 weeks. This observation strongly supports the established concept that maximizing neuroplasticity in people with chronic stroke often requires intensive, repetitive, and focused training to overcome neural inhibition and promote synaptic strengthening [61].

Therefore, based on the current evidence, a protocol of 30-minute sessions, administered 4-5 times per week over 2 weeks (totaling 10-12 sessions), emerges as a promising and efficient model for BCI intervention. It is crucial to emphasize that this proposal is not intended as a definitive guideline but rather highlights a potentially optimal treatment paradigm derived from the existing, albeit limited, data. This model warrants prioritization and validation in future rigorous research.

Research Prospects

Although subgroup analyses favor short-term, high-frequency protocols, this intensity may be insufficient to induce lasting neuroplastic reorganization. This observation aligns with the dissipation of functional gains at follow-up. Future high-quality RCTs are required to (1) delineate the dose-response relationship of BCI training, (2) analyze the synergistic effects of BCI combined with complementary therapies to optimize rehabilitation protocols, and (3) evaluate efficacy differentials based on lesion characteristics and upper-limb impairment severity in patients with chronic disease, thereby identifying responsive subpopulations. These investigations aim to inform evidence-based rehabilitation strategies and research priorities for stroke recovery.

Limitations

Several limitations warrant consideration in this meta-analysis. First, while the included 21 studies underwent rigorous quality assessment using the Cochrane RoB 2 tool, only 6 were rated as having “Low” risk of bias. This relatively low proportion of high-quality studies may limit the robustness of our findings. Second, insufficient studies (fewer than 5) were included in some subgroup analyses, potentially compromising the reliability of conclusions drawn for those specific comparisons. Third, the subgroup analyses, particularly for intervention intensity, were likely underpowered to detect statistically significant differences between parameters due to the limited number of studies in each category. Therefore, the identified optimal protocol should be viewed as the most evidence-based recommendation from the current data, rather than a definitively proven superior approach. Future research should prioritize incorporating a greater number of high-quality RCTs. Furthermore, greater emphasis is needed on exploring the impact of BCI training intervention intensity and focusing on outcomes such as improvements in muscle tone and standardized assessments like the ARAT. Investigating these areas represents promising avenues for advancing stroke rehabilitation.

XSL•FO
RenderX

Conclusion

Low- to moderate-certainty evidence suggests that BCI training, particularly the BCI-FES paradigm, can improve upper limb motor function and ADL in people with chronic stroke on average. However, wide prediction intervals indicate the effect may vary substantially across settings, ranging from negligible to beneficial. Subgroup analyses suggested a potential optimal protocol of 30-minute sessions, 4-5 times per week for 2 weeks, but these findings are limited by the small number of studies in each subgroup and the high risk of bias in several included trials. Therefore, this proposed protocol should be viewed as preliminary and requires validation in future, high-quality RCTs. Future research should also focus on identifying patient subgroups most likely to benefit and on strategies to sustain long-term gains.

Data Availability

The datasets generated or analyzed during this systematic review are available from the corresponding author on reasonable request.

Authors' Contributions

HJC and GJY were responsible for conceptualization. HJC completed database searching. HJC and GJY completed article screening, data extraction, and critical appraisal. HJC was responsible for data curation and analysis. HJC and GJY contributed to data interpretation. HJC wrote the original draft. All authors contributed to reviewing and editing the final manuscript and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA checklist.
[DOCX File , 30 KB - jmir_v28i1e79132_app1.docx]

Multimedia Appendix 2
Search strategy.
[DOCX File , 19 KB - jmir_v28i1e79132_app2.docx]

Multimedia Appendix 3
GRADE assessment.
[DOCX File , 36 KB - jmir_v28i1e79132_app3.docx]

Multimedia Appendix 4
The characteristics of included literatures.
[DOCX File , 31 KB - jmir_v28i1e79132_app4.docx]

References

1. Gibbons EM, Thomson AN, de Noronha M, Joseph S. Are virtual reality technologies effective in improving lower limb outcomes for patients following stroke - a systematic review with meta-analysis. Top Stroke Rehabil 2016;23(6):440-457. [doi: 10.1080/10749357.2016.1183349] [Medline: 27237336]
2. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the global burden of disease study 2019. Lancet Psychiatry 2022;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
3. Ahmed I, Mustafaoglu R, Erhan B. The effects of low-intensity resistance training with blood flow restriction versus traditional resistance exercise on lower extremity muscle strength and motor functionin ischemic stroke survivors: a randomized controlled trial. Top Stroke Rehabil 2024;31(4):418-429. [doi: 10.1080/10749357.2023.2259170] [Medline: 37724785]
4. O'Dell MW, Jaywant A, Frantz M, Patel R, Kwong E, Wen K, et al. Changes in the activity measure for post-acute care domains in persons with stroke during the first year after discharge from inpatient rehabilitation. Arch Phys Med Rehabil 2021;102(4):645-655. [doi: 10.1016/j.apmr.2020.11.020] [Medline: 33440132]
5. Heller A, Wade DT, Wood VA, Sunderland A, Hewer RL, Ward E. Arm function after stroke: measurement and recovery over the first three months. J Neurol Neurosurg Psychiatry 1987;50(6):714-719 [FREE Full text] [doi: 10.1136/jnnp.50.6.714] [Medline: 3612152]

XSL•FO
RenderX

6. Duncan PW, Goldstein LB, Horner RD, Landsman PB, Samsa GP, Matchar DB. Similar motor recovery of upper and lower extremities after stroke. Stroke 1994;25(6):1181-1188. [doi: 10.1161/01.str.25.6.1181] [Medline: 8202977]
7. McCabe J, Monkiewicz M, Holcomb J, Pundik S, Daly JJ. Comparison of robotics, functional electrical stimulation, and motor learning methods for treatment of persistent upper extremity dysfunction after stroke: a randomized controlled trial. Arch Phys Med Rehabil 2015;96(6):981-990 [FREE Full text] [doi: 10.1016/j.apmr.2014.10.022] [Medline: 25461822]
8. Remsik A, Young B, Vermilyea R, Kiekhoefer L, Abrams J, Evander Elmore S, et al. A review of the progression and future implications of brain-computer interface therapies for restoration of distal upper extremity motor function after stroke. Expert Rev Med Devices 2016;13(5):445-454. [doi: 10.1080/17434440.2016.1174572] [Medline: 27112213]
9. Li D, Li R, Song Y, Qin W, Sun G, Liu Y, et al. Effects of brain-computer interface based training on post-stroke upper-limb rehabilitation: a meta-analysis. J Neuroeng Rehabil 2025;22(1):44 [FREE Full text] [doi: 10.1186/s12984-025-01588-x] [Medline: 40033447]
10. Li M, Liu Y, Wu Y, Liu S, Jia J, Zhang L. Neurophysiological substrates of stroke patients with motor imagery-based brain-computer interface training. Int J Neurosci 2014;124(6):403-415. [doi: 10.3109/00207454.2013.850082] [Medline: 24079396]
11. Ang KK, Chua KSG, Phua KS, Wang C, Chin ZY, Kuah CWK, et al. A randomized controlled trial of EEG-based motor imagery brain-computer interface robotic rehabilitation for stroke. Clin EEG Neurosci 2015;46(4):310-320. [doi: 10.1177/1550059414522229] [Medline: 24756025]
12. Pichiorri F, Morone G, Petti M, Toppi J, Pisotta I, Molinari M, et al. Brain-computer interface boosts motor imagery practice during stroke recovery. Ann Neurol 2015;77(5):851-865. [doi: 10.1002/ana.24390] [Medline: 25712802]
13. Miller KJ, Schalk G, Fetz EE, den Nijs M, Ojemann JG, Rao RPN. Cortical activity during motor execution, motor imagery, and imagery-based online feedback. Proc Natl Acad Sci U S A 2010;107(9):4430-4435 [FREE Full text] [doi: 10.1073/pnas.0913697107] [Medline: 20160084]
14. Casimo K, Weaver KE, Wander J, Ojemann JG. BCI use and its relation to adaptation in cortical networks. IEEE Trans Neural Syst Rehabil Eng 2017;25(10):1697-1704 [FREE Full text] [doi: 10.1109/TNSRE.2017.2681963] [Medline: 28320670]
15. Lebedev MA, Nicolelis MAL. Brain-machine interfaces: from basic science to neuroprostheses and neurorehabilitation. Physiol Rev 2017;97(2):767-837 [FREE Full text] [doi: 10.1152/physrev.00027.2016] [Medline: 28275048]
16. Ma T, Li H, Deng L, Yang H, Lv X, Li P, et al. The hybrid BCI system for movement control by combining motor imagery and moving onset visual evoked potential. J Neural Eng 2017;14(2):026015. [doi: 10.1088/1741-2552/aa5d5f] [Medline: 28145274]
17. Cervera MA, Soekadar SR, Ushiba J, Millán JDR, Liu M, Birbaumer N, et al. Brain-computer interfaces for post-stroke motor rehabilitation: a meta-analysis. Ann Clin Transl Neurol 2018;5(5):651-663 [FREE Full text] [doi: 10.1002/acn3.544] [Medline: 29761128]
18. Daly JJ, Cheng R, Rogers J, Litinas K, Hrovat K, Dohring M. Feasibility of a new application of noninvasive brain computer interface (BCI): a case study of training for recovery of volitional motor control after stroke. J Neurol Phys Ther 2009;33(4):203-211. [doi: 10.1097/NPT.0b013e3181c1fc0b] [Medline: 20208465]
19. Biasiucci A, Leeb R, Iturrate I, Perdikis S, Al-Khodairy A, Corbet T, et al. Brain-actuated functional electrical stimulation elicits lasting arm motor recovery after stroke. Nat Commun 2018;9(1):2421 [FREE Full text] [doi: 10.1038/s41467-018-04673-z] [Medline: 29925890]
20. Liu X, Zhang W, Li W, Zhang S, Lv P, Yin Y. Effects of motor imagery based brain-computer interface on upper limb function and attention in stroke patients with hemiplegia: a randomized controlled trial. BMC Neurol 2023;23(1):136 [FREE Full text] [doi: 10.1186/s12883-023-03150-5] [Medline: 37003976]
21. Xie Y, Yang Y, Jiang H, Duan X, Gu L, Qing W, et al. Brain-machine interface-based training for improving upper extremity function after stroke: a meta-analysis of randomized controlled trials. Front Neurosci 2022;16:949575 [FREE Full text] [doi: 10.3389/fnins.2022.949575] [Medline: 35992923]
22. Peng Y, Wang J, Liu Z, Zhong L, Wen X, Wang P, et al. The application of brain-computer interface in upper limb dysfunction after stroke: a systematic review and meta-analysis of randomized controlled trials. Front Hum Neurosci 2022;16:798883 [FREE Full text] [doi: 10.3389/fnhum.2022.798883] [Medline: 35422693]
23. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
24. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
25. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc 2021;109(2):174-200 [FREE Full text] [doi: 10.5195/jmla.2021.962] [Medline: 34285662]
26. Cumpston M, Li T, Page MJ, Chandler J, Welch VA, Higgins JP, et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane handbook for systematic reviews of interventions. Cochrane Database Syst Rev 2019;10(10):ED000142. [doi: 10.1002/14651858.ED000142] [Medline: 31643080]

XSL•FO
RenderX

27. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
28. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336(7650):924-926 [FREE Full text] [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
29. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
30. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta-analysis: I is not an absolute measure of heterogeneity. Res Synth Methods 2017;8(1):5-18. [doi: 10.1002/jrsm.1230] [Medline: 28058794]
31. Michaelis R, Tang V, Wagner JL, Modi AC, LaFrance WC, Goldstein LH, et al. Cochrane systematic review and meta-analysis of the impact of psychological treatments for people with epilepsy on health-related quality of life. Epilepsia 2018;59(2):315-332 [FREE Full text] [doi: 10.1111/epi.13989] [Medline: 29313968]
32. IntHout J, Ioannidis JPA, Borm GF. The hartung-knapp-sidik-jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard dersimonian-laird method. BMC Med Res Methodol 2014;14:25 [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
33. Gomes C, Geels J, Debray TPA, Malekzadeh A, Asselbergs FW, Linschoten M. Risk prediction models for cancer therapy related cardiac dysfunction in patients with cancer and cancer survivors: systematic review and meta-analysis. BMJ 2025;390:e084062 [FREE Full text] [doi: 10.1136/bmj-2025-084062] [Medline: 40987514]
34. Fugl-Meyer AR, Jääskö L, Leyman I, Olsson S, Steglind S. The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. Scand J Rehabil Med 1975;7(1):13-31. [Medline: 1135616]
35. Ma Z, Wu J, Cao Z, Hua X, Zheng M, Xing X, et al. Motor imagery-based brain-computer interface rehabilitation programs enhance upper extremity performance and cortical activation in stroke patients. J Neuroeng Rehabil 2024;21(1):91 [FREE Full text] [doi: 10.1186/s12984-024-01387-w] [Medline: 38812014]
36. Lu R, Pang Z, Gao T, He Z, Hu Y, Zhuang J, et al. Multisensory BCI promotes motor recovery via high-order network-mediated interhemispheric integration in chronic stroke. BMC Med 2025;23(1):380. [doi: 10.1186/s12916-025-04214-8] [Medline: 40598460]
37. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011;343:d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
38. Miao Y, Chen S, Zhang X, Jin J, Xu R, Daly I, et al. BCI-based rehabilitation on the stroke in sequela stage. Neural Plast 2020;2020:8882764 [FREE Full text] [doi: 10.1155/2020/8882764] [Medline: 33414824]
39. Cheng N, Phua KS, Lai HS, Tam PK, Tang KY, Cheng KK, et al. Brain-computer interface-based soft robotic glove rehabilitation for stroke. IEEE Trans Biomed Eng 2020;67(12):3339-3351. [doi: 10.1109/TBME.2020.2984003] [Medline: 32248089]
40. Kim MS, Park H, Kwon I, An K, Kim H, Park G, et al. Efficacy of brain-computer interface training with motor imagery-contingent feedback in improving upper limb function and neuroplasticity among persons with chronic stroke: a double-blinded, parallel-group, randomized controlled trial. J Neuroeng Rehabil 2025;22(1):1 [FREE Full text] [doi: 10.1186/s12984-024-01535-2] [Medline: 39757218]
41. Ramos-Murguialday A, Broetz D, Rea M, Läer L, Yilmaz O, Brasil FL, et al. Brain-machine interface in chronic stroke rehabilitation: a controlled study. Ann Neurol 2013;74(1):100-108 [FREE Full text] [doi: 10.1002/ana.23879] [Medline: 23494615]
42. Ramos-Murguialday A, Curado MR, Broetz D, Yilmaz, Brasil FL, Liberati G, et al. Brain-machine interface in chronic stroke: randomized trial long-term follow-up. Neurorehabil Neural Repair 2019;33(3):188-198. [doi: 10.1177/1545968319827573] [Medline: 30722727]
43. Curado MR, Cossio EG, Broetz D, Agostini M, Cho W, Brasil FL, et al. Residual upper arm motor function primes innervation of paretic forearm muscles in chronic stroke after brain-machine interface (BMI) training. PLoS One 2015;10(10):e0140161 [FREE Full text] [doi: 10.1371/journal.pone.0140161] [Medline: 26495971]
44. Li X, Wang J, Cao X, Huang W, Hu Y. Soft robotic glove with alpha band brain computer interface for post-stroke hand function rehabilitation. : IEEE; 2022 Presented at: 14th Biomedical Engineering International Conference (BMEiCON); 2022 November 10-13; Songkhla, Thailand. [doi: 10.1109/bmeicon56653.2022.10012103]
45. Mihara M, Hattori N, Hatakenaka M, Yagura H, Kawano T, Hino T, et al. Near-infrared spectroscopy-mediated neurofeedback enhances efficacy of motor imagery-based training in poststroke victims: a pilot study. Stroke 2013;44(4):1091-1098. [doi: 10.1161/STROKEAHA.111.674507] [Medline: 23404723]
46. Sanders Z, Fleming MK, Smejka T, Marzolla MC, Zich C, Rieger SW, et al. Self-modulation of motor cortex activity after stroke: a randomized controlled trial. Brain 2022;145(10):3391-3404 [FREE Full text] [doi: 10.1093/brain/awac239] [Medline: 35960166]
47. Guo N, Wang X, Duanmu D, Huang X, Li X, Fan Y, et al. SSVEP-based brain computer interface controlled soft robotic glove for post-stroke hand function rehabilitation. IEEE Trans Neural Syst Rehabil Eng 2022;30:1737-1744. [doi: 10.1109/TNSRE.2022.3185262] [Medline: 35731756]

XSL•FO
RenderX

48. Kim T, Kim S, Lee B. Effects of action observational training plus brain-computer interface-based functional electrical stimulation on paretic arm motor recovery in patient with stroke: a randomized controlled trial. Occup Ther Int 2016;23(1):39-47 [FREE Full text] [doi: 10.1002/oti.1403] [Medline: 26301519]
49. Lee S, Kim SS, Lee B. Action observation training and brain-computer interface controlled functional electrical stimulation enhance upper extremity performance and cortical activation in patients with stroke: a randomized controlled trial. Physiother Theory Pract 2022;38(9):1126-1134. [doi: 10.1080/09593985.2020.1831114] [Medline: 33026895]
50. Frolov AA, Mokienko O, Lyukmanov R, Biryukova E, Kotov S, Turbina L, et al. Post-stroke rehabilitation training with a motor-imagery-based brain-computer interface (bci)-controlled hand exoskeleton: a randomized controlled multicenter trial. Front Neurosci 2017;11:400 [FREE Full text] [doi: 10.3389/fnins.2017.00400] [Medline: 28775677]
51. Hu Y, Gao T, Li J, Tao J, Bai Y, Lu R. Motor imagery-based brain-computer interface combined with multimodal feedback to promote upper limb motor function after stroke: a preliminary study. Evid Based Complement Alternat Med 2021;2021:1116126 [FREE Full text] [doi: 10.1155/2021/1116126] [Medline: 34777531]
52. Hao M, Fang Q, Wu B, Liu L, Tang H, Tian F, et al. Rehabilitation effect of intelligent rehabilitation training system on hemiplegic limb spasms after stroke. Open Life Sci 2023;18(1):20220724 [FREE Full text] [doi: 10.1515/biol-2022-0724] [Medline: 37791058]
53. Jinshu Z, Mingming W, Yuan Z, Xuanxiang S. Clinical research on rehabilitation treatment of hemiplegia after stroke by rehabilitation robot based on brain-computer interaction technology. J External Ther Tradit Chin Med 2021;30(3):3-5. [doi: 10.3969/j.issn.1006-978X.2021.03.001]
54. Ying X, Yanyun J, Jie J, Xiaomei W. Efficacy observation of brain - computer interface combined with functional electrical stimulation training on upper limb function and cognition in elderly stroke patients. Chin J Geriatr Heart Brain Vessel Dis 2018;20(9):988-990. [doi: 10.3969/j.issn.1009-0126.2018.09.023]
55. Wang X, Wong W, Sun R, Chu WC, Tong K. Differentiated effects of robot hand training with and without neural guidance on neuroplasticity patterns in chronic stroke. Front Neurol 2018;9:810 [FREE Full text] [doi: 10.3389/fneur.2018.00810] [Medline: 30349505]
56. Ren C, Li X, Gao Q, Pan M, Wang J, Yang F, et al. The effect of brain-computer interface controlled functional electrical stimulation training on rehabilitation of upper limb after stroke: a systematic review and meta-analysis. Front Hum Neurosci 2024;18:1438095 [FREE Full text] [doi: 10.3389/fnhum.2024.1438095] [Medline: 39391265]
57. Hung J, Wu W, Chen Y, Pong Y, Chang K. Predictors of clinically important improvements in motor function and daily use of affected arm after a botulinum toxin a injection in patients with chronic stroke. Toxins (Basel) 2021;14(1):13 [FREE Full text] [doi: 10.3390/toxins14010013] [Medline: 35050990]
58. Lazari A, Salvan P, Cottaar M, Papp D, Rushworth MFS, Johansen-Berg H. Hebbian activity-dependent plasticity in white matter. Cell Rep 2022;39(11):110951 [FREE Full text] [doi: 10.1016/j.celrep.2022.110951] [Medline: 35705046]
59. Veerbeek JM, Langbroek-Amersfoort AC, van Wegen EEH, Meskers CGM, Kwakkel G. Effects of robot-assisted therapy for the upper limb after stroke. Neurorehabil Neural Repair 2017;31(2):107-121. [doi: 10.1177/1545968316666957] [Medline: 27597165]
60. Kruse A, Suica Z, Taeymans J, Schuster-Amft C. Effect of brain-computer interface training based on non-invasive electroencephalography using motor imagery on functional recovery after stroke - a systematic review and meta-analysis. BMC Neurol 2020;20(1):385 [FREE Full text] [doi: 10.1186/s12883-020-01960-5] [Medline: 33092554]
61. Park H, Kim S, Winstein CJ, Gordon J, Schweighofer N. Short-duration and intensive training improves long-term reaching performance in individuals with chronic stroke. Neurorehabil Neural Repair 2016;30(6):551-561 [FREE Full text] [doi: 10.1177/1545968315606990] [Medline: 26405046]

Abbreviations

ADL: activities of daily living
ARAT: Action Research Arm Test
BCI: brain-computer interface
EEG: electroencephalography
FES: functional electrical stimulation
FMA-UE: Fugl-Meyer Assessment for upper extremity
fNIRS: functional Near-Infrared Spectroscopy
GRADE: Grading of Recommendations Assessment, Development and Evaluation
MAL: Motor Activity Log
MAS: Modified Ashworth Scale
MBI: Modified Barthel Index
MD: mean difference
MeSH: Medical Subject Headings
PI: prediction interval
PRESS: Peer Review of Electronic Search Strategies

XSL•FO
RenderX

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses – Search Extension
PROSPERO: International Prospective Register of Systematic Reviews
RCT: randomized controlled trial
ROB: Risk of Bias tool
RRT: routine rehabilitation therapy

Edited by S Brini; submitted 16.Jun.2025; peer-reviewed by J Liang, Y Fan; comments to author 14.Oct.2025; revised version received 30.Nov.2025; accepted 04.Dec.2025; published 28.Jan.2026.

Please cite as:
Chen H, Yun G
Efficacy of Brain-Computer Interface Therapy for Upper Limb Rehabilitation in Chronic Stroke: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e79132
URL: https://www.jmir.org/2026/1/e79132
doi:10.2196/79132
PMID:

©HongJie Chen, GuoJun Yun. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Behavioral Determinants and Effectiveness of Digital Behavior Change Interventions for the Prevention of Sexually Transmitted Infections and HIV: Overview of Systematic Reviews

Giuliano Duarte-Anselmi[1,2*], MPH; Susana Sanduvete-Chaves[3*], PhD; Salvador Chacón-Moscoso[3,4*], PhD; Daniel López-Arenas[3*], MPR

[1]Facultad de Psicología, Universitat de Barcelona, Barcelona, Spain
[2]Facultad de Ciencias Médicas, Universidad de Santiago de Chile, Santiago, Chile
[3]Departamento de Psicología Experimental, Facultad de Psicología, Universidad de Sevilla, Sevilla, Spain
[4]Universidad Autónoma de Chile, Santiago, Chile
[*]all authors contributed equally

Corresponding Author:
Susana Sanduvete-Chaves, PhD
Departamento de Psicología Experimental
Facultad de Psicología
Universidad de Sevilla
Campus Ramón y Cajal. C/ Camilo José Cela, s/n
Sevilla, 41018
Spain
Phone: 34 954557676
Email: sussancha@us.es

Abstract

Background: Unsafe sexual practices remain a major contributor to global morbidity, premature mortality, and health care burden. More than 1 million people acquire a sexually transmitted infection (STI) daily, including HIV. Although biomedical innovations such as pre-exposure prophylaxis have expanded prevention options, consistent condom use and regular HIV and STI testing remain essential behavioral strategies. Adherence to these behaviors remains uneven, underscoring the need for complementary digital and behavioral approaches. Digital behavior change interventions (DBCIs), technology-based programs designed to support health-related behavior change, offer scalable and personalized tools for safer-sex promotion. However, evidence regarding their behavioral components and effectiveness remains fragmented across systematic reviews (SRs).

Objective: This study aims to synthesize and critically appraise evidence on the effectiveness of DBCIs for preventing STIs and HIV, and to identify which behavior change techniques (BCTs) and theoretical domains framework (TDF) have been used to improve safe-sex behaviors.

Methods: A search was conducted in MEDLINE, Cochrane Database of SRs, Epistemonikos, and PsycINFO for all publications up to November 12, 2025, without language or date restrictions. Eligible SRs examined DBCIs targeting STI and HIV prevention or reduction of risky sexual behaviors. Two reviewers (GDA and DLA) independently screened, extracted data, and appraised methodological quality using the AMSTAR-2 tool. The reporting followed the PRIOR (Preferred Reporting Items for Overviews of Reviews) and PRISMA-S (Preferred Reporting Items for SRs and Meta-Analyses Literature Search Extension) recommendations.

Results: Overall, 23 SRs, comprising 514 primary studies and 129,481 participants, met the inclusion criteria. Most interventions were SMS-based, mobile app–based, or web-delivered. Digital interventions consistently improved STI and HIV testing uptake and engagement with sexual health services. Evidence for condom use and biological outcomes was mixed. Improvements in cognitive determinants, such as HIV-related knowledge, motivation, and self-efficacy, were frequently reported. Only 4 reviews explicitly applied BCT or TDF taxonomies, identifying goal setting, feedback on behavior, and prompts and cues as commonly used techniques. Research predominantly originated from high-income settings, with limited evidence from low- and middle-income countries and minimal reporting of sex- or gender-disaggregated outcomes.

Conclusions: DBCIs show promise for strengthening STI/HIV prevention, particularly by increasing testing behaviors and supporting cognitive determinants of risk reduction. However, sustained condom use and biological outcomes remain inconsistent,

XSL•FO
RenderX

and reporting of behavioral mechanisms is limited. This overview is the first to integrate effectiveness evidence with a systematic, mechanism-focused mapping of BCTs and TDF constructs, providing an innovation not present in earlier reviews. Clarifying which active components of digital interventions are most consistently linked to beneficial outcomes offers concrete guidance for designing culturally tailored, theory-driven, and equity-focused digital strategies. These insights have direct implications for researchers, clinicians, and policymakers seeking to develop digital prevention programs that more effectively address behavioral determinants of STI and HIV risk.

Trial Registration: PROSPERO CRD42023485887; https://www.crd.york.ac.uk/PROSPERO/view/CRD42023485887

International Registered Report Identifier (IRRID): RR2-10.5867/medwave.2025.02.3020

(J Med Internet Res 2026;28:e74201) doi:10.2196/74201

KEYWORDS

behavioral change; behavioral design; sexually transmitted diseases; HIV; digital behavior change intervention (DBCI)

Introduction

Unsafe sexual practices are major contributors to global morbidity and premature mortality, representing one of the leading behavioral risk factors worldwide [1,2]. Among young people, these practices significantly increase the risk of sexually transmitted infections (STIs) including syphilis, gonorrhea, chlamydia, and HIV, as well as human papillomavirus (HPV)–related cancers [3,4]. A recent report by the World Health Organization highlights an alarming decrease in condom use among adolescents in Europe, leading to higher rates of unprotected sex and, consequently, an increased risk of STIs, particularly among adolescents from low-income families [5]. Although biomedical innovations such as pre-exposure prophylaxis (PrEP) have expanded prevention options, consistent condom use and regular HIV and STI testing remain essential behavioral strategies for reducing infection risk [6]. Together, these measures form a complementary prevention framework; yet, adherence remains uneven across populations. HIV and AIDS continues to be a leading cause of death globally, with more than 1 million people acquiring an STI daily and nearly 39 million living with HIV [7,8]. Behavioral determinants, including motivation, self-regulation, risk perception, social and cultural norms, and structural barriers, shape whether individuals engage in STI and HIV prevention behaviors, yet they are often insufficiently addressed or poorly defined in existing prevention strategies.

Given the scale of these challenges, effective prevention strategies must emphasize sociocultural and behavioral changes, such as increasing awareness, reducing stigma, and promoting safe sex practices like consistent condom use and regular STI and HIV testing [9-11]. Widespread access to the Internet and mobile phones presents a unique opportunity to leverage digital interventions as private and effective methods for improving sexual health, particularly in regions with varying levels of literacy [11-14].

Recent evidence underscores the growing use of digital technologies in HIV and STI prevention. A 2024 umbrella review found that eHealth interventions, ranging from mobile apps and websites to telemedicine and social media programs, were generally effective in supporting HIV prevention, testing, and clinical management, although the methodological quality of many reviews was low [15]. However, the Shi et al [15] review included both prevention and treatment interventions, whereas the present overview focuses exclusively on preventive strategies and their behavioral mechanisms. Evidence also suggests that the inclusion of behavior change techniques (BCTs) in digital tools enhances user engagement and intervention effectiveness [16]. Additionally, recent work has highlighted the expanding role of interactive digital tools in partner notification and sexual health engagement [17]. Together, these findings demonstrate the rapid evolution of digital health approaches and emphasize the need to systematically map their behavioral components to guide the design of effective prevention programs. However, most existing systematic reviews (SRs) and umbrella reviews describe the effects of digital interventions without examining their mechanisms of action, that is, the specific theoretical pathways and BCTs through which interventions influence prevention behaviors. Without identifying these mechanisms of action, it is difficult to understand why some digital programs succeed while others do not, and which components should be replicated or scaled.

Digital interventions can be broadly defined as health-promoting programs delivered through digital platforms such as websites, mobile apps, text messaging, or social media [18]. Within this broad category, digital behavior change interventions (DBCIs) are those that explicitly incorporate theoretical frameworks and structured BCTs to influence health-related behaviors [19]. In other words, while all DBCIs are digital interventions, not all digital interventions qualify as DBCIs. This conceptual distinction underpins our search strategy and synthesis approach, focusing on interventions that use digital delivery to achieve behavioral outcomes through identifiable active ingredients. These definitions are provided upfront to reduce conceptual ambiguity, as emphasized by recent critiques in digital behavior change research.

DBCIs offer multiple advantages over traditional prevention approaches: they can deliver tailored, interactive, and adaptive content; are cost-effective and scalable; and can integrate technological features such as automated feedback and passive sensing [20-22]. However, information alone is insufficient to drive behavior change: integrating BCTs to these digital platforms is essential for achieving meaningful health outcomes [23,24].

XSL•FO
RenderX

In this context, the theoretical domains framework (TDF) is presented as a widely used proposal to systematically identify barriers and facilitators to change specific behaviors, helping design more effective interventions in health, among other fields. It is useful to understand why people do or do not do something, and highlight factors needing intervention [25,26]. It synthesizes 33 behavior change theories into 14 core domains, such as cognitive (knowledge, skills, beliefs about capabilities and consequences), affective and emotional (emotions and reinforcement), social and environmental (social and professional role, social influences, environmental context, and resources), and beliefs and intentions (optimism, intentions, goals, memory, attention, decision processes, behavioral regulation) [27].

Given the volume of SRs assessing digital interventions for STI and HIV prevention, an overview of reviews enables synthesis and comparison across multiple bodies of evidence rather than relying on a single set of primary studies. This approach provides a broader understanding of intervention effectiveness, identifying patterns, strengths, methodological gaps, and avenues for future research [28,29]. Despite their promise, many digital interventions lack clear descriptions of the BCTs and theoretical domains they use, hindering replicability and practical translation [30,31]. Identifying the most effective BCTs, especially those that successfully promote safe sex and reduce STI and HIV transmission, is crucial for public health initiatives [32-35]. To date, no overview has systematically integrated BCTs, theoretical domains, and prevention outcomes to produce a mechanism-focused synthesis of digital interventions for STI and HIV prevention. Existing reviews also provide limited up-to-date evidence and do not incorporate studies published through 2025. Addressing this gap is essential for identifying which behavioral components drive meaningful changes in prevention behaviors and for informing the development of digital strategies that are theoretically grounded, culturally responsive, and scalable.

Accordingly, this overview aims to synthesize current evidence on the use of BCTs and the TDF in digital interventions designed to prevent STIs and HIV. By examining how these behavioral components are implemented and how they influence prevention-related outcomes, this research seeks to inform the development of more effective, theory-driven digital interventions and strengthen future public health strategies. This overview therefore provides not only an updated assessment of the evidence but also a behavioral mapping that has been largely absent from previous syntheses. It further advances the field by incorporating SRs published through 2025, which have not yet been integrated in any prior synthesis.

Methods

Study Design

This study is an overview of SRs and adheres to the Cochrane Handbook for SRs of Interventions [29] and the PRIOR (Preferred Reporting Items for Overviews of Reviews) statement [36]. The PRIOR checklist is reported in Multimedia Appendix 1 and the Sex and Gender Equity in Research (SAGER) guidelines [37], in Multimedia Appendix 2 (section 2). In addition, the search strategy and reporting follow the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) [38] to ensure full transparency and reproducibility of search methods (see Multimedia Appendix 3).

Protocol and Registration

The protocol for this overview was prospectively registered in PROSPERO (CRD42023485887) on December 5, 2023 (Multimedia Appendix 2 section 3), and later published in full [39]. The methods adhered to the predefined protocol and incorporated SAGER guidance where applicable. Reporting also followed PRISMA 2020 recommendations for SRs and overviews [40]. The full protocol can be downloaded from the Open Science Framework [41,42].

Patient and Public Involvement

Neither patients nor the public were involved in designing or conducting this study. Therefore, no ethical approval was required for this overview. The analyzed data were open access.

Eligibility Criteria

The eligibility criteria were reported in detail in our protocol [39]. The inclusion criteria for this overview were based on the population, intervention, comparison, outcome, study type (PICOS) framework (Table 1).

XSL•FO
RenderX

Table 1. Eligibility criteria for elements of a comprehensive search strategy.

Element	Inclusion and exclusion criteria
Population	• Included: SRs[a] that have evaluated the effect of digital behavior change interventions in any population and that have described BCTs[b], mechanisms of action, or any behavioral model or framework that takes into account how the digital intervention influences the behavior change used to reduce the risk or prevent the transmission of STIs[b], including HIV. • Excluded: studies that did not focus on prevention or that focused on treatment and adherence to antiretroviral therapy and self-care of people living with HIV were excluded, as the focus of the question research is risk reduction and prevention of STIs and HIV.
Intervention	• Included: SRs that evaluated digital and mobile health behavior change interventions focusing on modifying unsafe sexual behaviors or preventing STIs and HIV, that is, interventions carried out using a digital or mobile platform as a direct interface with the participants. • Excluded: studies that did not report the use of digital intervention, that is, did not incorporate digital technology such as smartphones, computers, tablets, multimedia, and social networks.
Comparator	• As this was an overview of SRs, a comparator or control group was not an inclusion criterion for this study. However, given our interest in effectiveness of interventions and BCTs, we included reviews where evidence from primary experimental studies with an appropriate comparator was available.
Outcome	• Included: during the selection process, we do not consider concrete results as an inclusion criterion. All studies that assessed short- or long-term behavior change concerning the following primary outcomes were included: reduction in risky sexual behaviors, such as condom use (last sexual encounter, frequency, consistency) and increased STI and HIV testing; and prevention (vaccination against HPV[d] and hepatitis A and B, and HIV pre-exposure prophylaxis). • As secondary outcomes, the use of behavior change theories or techniques was analyzed, using standardized classifications if available, such as the taxonomy of BCTs. • Excluded: any study that did not include a primary and/or secondary outcome.
Study design	• Included: SRs only • Excluded: studies other than SRs (eg, primary studies, commentary articles, and conferences) were excluded.

[a]SR: systematic review.

[b]BCT: behavior change technique.

[c]STI: sexually transmitted infection.

[d]HPV: human papillomavirus.

Summary of inclusion and exclusion criteria used to identify SRs of digital behavior change interventions for the prevention of STIs and HIV. Criteria were defined following the Cochrane Handbook and the PRIOR guideline. All SRs were operationally defined as secondary research including primary clinical studies with explicit search strategies in ≥2 databases [29,43].

Information Sources

In this overview, a comprehensive search strategy was used, leveraging multiple key sources. Primary searches were conducted on the leading international databases MEDLINE (via PubMed), the Cochrane Database of SRs, Epistemonikos, and PsycINFO. Each database was searched independently using its native platform; no multidatabase platform (eg, EBSCOhost and Ovid) was used. To enhance the scope of our search, we conducted supplementary searches to identify any studies potentially overlooked by the primary search strategy or absent from the indexed databases. These supplementary efforts included (1) meticulous manual reviews of the references cited in the included studies, (2) examination of related SRs that shared at least one study with the reviews included in our study, and (3) additional records identified through various websites (n=7), organizations (n=1), citation searching (n=321), references provided by authors (n=1), and consultations with experts (n=28; Figure 1). Citation searching involved both backward citation screening (reference list checking) and forward citation tracking using Google Scholar. No additional search methods such as automated alerts, web-scraping tools, or application programming interface–based search retrieval were used.

XSL•FO
RenderX

Figure 1. Study selection flow diagram for the overview of systematic reviews on digital behavior change interventions (DBCIs) for STI/HIV prevention, following the PRIOR reporting guideline. PRIOR: Preferred Reporting Items for Overviews of Reviews.

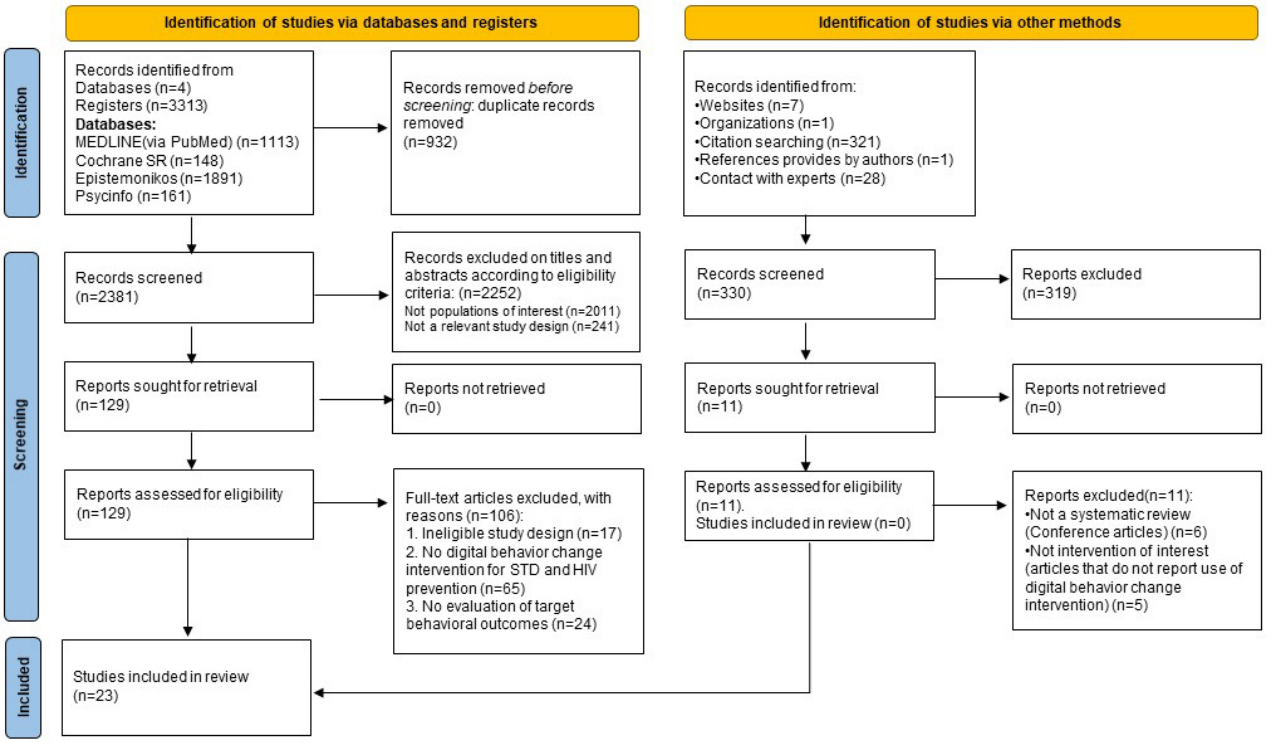

Since this overview synthesizes published SRs rather than primary studies, study and trial registries (eg, ClinicalTrials.gov, International Clinical Trials Registry Platform) were not searched. Each source was rigorously scrutinized, with the date of the last search or consultation carefully documented to ensure the currency and relevance of the findings. Database-specific yields were MEDLINE (via PubMed; n=1113), Cochrane Library (n=148), PsycINFO (n=161), and Epistemonikos (n=1891), as detailed in Multimedia Appendix 2. Detailed methodologies and search strategies are available in Multimedia Appendix 2 (section 4).

Search Strategy

The electronic search strategy (sections 5 and 6 in Multimedia Appendix 2) was developed and conducted under the supervision of an experienced librarian. The strategy did not undergo formal PRESS (Peer Review of Electronic Search Strategies) peer review. The search strategy was newly developed for this overview and was not adapted from any previous review. The first author performed the electronic search from the databases' inception up to November 12, 2025, with no restrictions on publication date, language, or country of origin, following PRISMA-S recommendations to ensure transparent and reproducible reporting of search methods. No methodological search filters (eg, SR filters, randomized trial filters, and human-only filters) were applied beyond the predefined eligibility criteria. An earlier search completed on August 31, 2024, was rerun and updated in accordance with PRISMA-S guidance [38], to capture the most recent SRs before resubmission. Additionally, we manually searched the bibliographies of relevant reviews and the articles initially retrieved. Letters were also sent to authors and experts identified in the included and excluded studies during the screening stage to identify additional eligible studies. The eligibility criteria are listed in Table 1. Full search dates for each database, including the initial search (August 31, 2024) and rerun (November 12, 2025), are reported in Multimedia Appendix 2. Full search strategies for all databases, copied verbatim as executed, are provided in Multimedia Appendix 2.

Study Selection Process

Following deduplication and a pilot test of the inclusion and exclusion criteria, 2 independent reviewers (GDA and DLA) screened all titles, abstracts, and full-text articles for eligibility without knowledge of each other's decisions. For the deduplication process, all records retrieved from database searches and supplementary sources were imported into Collaboratron (Epistemonikos). The software's automated similarity-detection algorithm was used to identify duplicate entries, followed by manual verification by 2 independent reviewers (GDA and DLA) to ensure accuracy. This hybrid deduplication approach, combining automated and manual procedures, aligns with PRISMA-S recommendations for transparent management of search records. Records from electronic and bibliographic searches were stored and full text screening was conducted using Collaboratron by Epistemonikos [44]. Differences between the 2 reviewers (GDA and DLA) were resolved through discussion, and a third reviewer (SSC) was consulted when necessary. The list of studies excluded after the full text review, along with reasons for exclusions, is detailed in section 7 in Multimedia Appendix 2.

We reached out to 28 authors (up to 3 email attempts) to request additional information, particularly on gray literature such as conference presentations and reports (see the list of experts in

XSL•FO
RenderX

the Contacting Experts section of section 4 in Multimedia Appendix 2).

The interrater reliability was assessed using Cohen kappa coefficient [45]. Two independent reviewers (GDA and DLA) evaluated the full-texts (section 8 in Multimedia Appendix 2).

Data Collection Process

We developed a data extraction tool in Microsoft Excel to obtain various study data recommended by Cochrane [46,47]. The data extracted from the SRs selected for the study was tested and calibrated by the team (section 9 in Multimedia Appendix 2). For this purpose, 1 author (GDA) created the spreadsheet and then extracted the data from 1 SR. Subsequently, 2 authors (GDA and DLA) independently extracted data from 3 SRs, and all authors provided feedback on whether the data elements were complete, and the extracted data were unambiguous. Once a consensus was reached through discussion, 1 author (GDA) created a data extraction manual for the spreadsheet, which can be found in Multimedia Appendix 2 (section 10).

Data Items

Data items included SR characteristics, PICOS criteria, and variables related to DBCI (Textbox 1).

XSL•FO
RenderX

Textbox 1. Data items in this overview of systematic reviews.

Data items

- Bibliographic information (author, year of publication, title, and aim of the SR).

Population characteristics:

- Participants (total number of participants included in the studies)
- Population age
- Specific population (men who have sex with men, lesbian, gay, bisexual, transgender, queer, and other sexual orientations and gender identities (LGBTQI+) individuals, people with a diagnosis of sexually transmitted disease [STD] without HIV)

Study characteristics

- Total number of studies included in the review
- Number of randomized controlled trials (RCTs) included
- Type of studies (only RCTs, only non-RCTs, including experimental nonrandomized and observational studies-, both RCTs and non-RCTs)
- Review period (range of publication years of the primary studies included in each systematic review, reflecting the temporal coverage of the synthesized evidence)
- Period or specific date range of the literature search (years during which databases were searched by the authors of each SR);
- Country or geographic location of the studies

Intervention details

- Target population
- Target behavior
- Explicit mention or extraction of theoretical frameworks used.
- Behavioral outcomes (condom use, frequency of unprotected sexual intercourse, number of sexual partners, STD and HIV testing, uptake of medical male circumcision, HIV counseling, vaccination)
- Cognitive outcomes (self-efficacy, STD and HIV-related knowledge, attitudes toward condom use, and pre-exposure prophylaxis awareness)
- Biological outcomes (acquisition of HIV or sexually transmitted infection)

Intervention acceptability and feasibility

- Acceptability: participants' acceptance of the intervention
- Practicability: ease of implementation in the real world
- Effectiveness: achievement of the intervention objectives
- Affordability: cost-effectiveness of the intervention
- Spill-over effects: unintended consequences
- Equity: impact on health equity

Technology and delivery methods

- Mobile devices, desktop computers, digital billboards, wearable accessories, digital objects, and projection and holograms
- Mode of delivery of digital contents (audio calls and messages, video calls and messages, text and instant messages, emails, audio broadcasts and podcasts, websites and computer programs and apps, eBooks, virtual or augmented reality, artificial intelligence; for example, use of artificial intelligence–based chatbots to promote safe sex or other sexual behaviors)

Other characteristics

- Number of primary studies included and overlap, tools used to assess the risk of bias in primary studies, whether a meta-analysis was conducted, and certainty of evidence (Grading of Recommendations, Assessment, Development, and Evaluation)

Quality Appraisal of the SRs

We performed critical appraisals of SRs using AMSTAR-2 [48], as outlined in our published protocol [39]. AMSTAR-2 consists of 16 items that assess the thoroughness of various aspects of a SR, such as the preparation process, literature search, study selection, data extraction, and analysis, as well as potential biases (eg, risk of bias, publication bias, or funding sources). Based on the type and number of weaknesses identified (ie, unmet items), the reviews were assigned a confidence rating:

XSL•FO
RenderX

high, moderate, low, or critically low. Two authors (GDA and DLA) independently assessed all SRs using a spreadsheet (Microsoft Excel 2010) and reached consensus through discussion (section 11 in Multimedia Appendix 2).

Overlap in Primary Studies Included in Reviews

To ensure the accuracy of the primary study outcome data and avoid overlap, we checked whether the included SRs shared overlapping primary studies. This was done by creating a citation matrix and calculating the overall corrected covered area (CCA) using Graphical Representation of Overlap for Overviews (GROOVE) [49]. The CCA quantifies the degree of overlap between primary studies included across SRs, calculated as the number of repeated primary studies (numerator) divided by the product of the total number of unique primary studies and the number of reviews, minus the total number of unique studies (denominator). The resulting value represents the proportion of shared evidence between reviews. A CCA of 0% to 5% indicates a slight overlap, 6% to 10% a moderate overlap, 11% to 15% a high overlap, and greater than 15% a very high overlap."

Data Synthesis Methods

Overview

Across the included SRs, the types of control or comparison groups varied substantially. In most cases, DBCIs were compared against nondigital or usual-care conditions, such as standard health education, printed materials, or no intervention controls. A smaller number of reviews included comparators that were themselves digital but lacked explicit behavior change components (eg, informational websites or SMS reminders without BCTs). This heterogeneity makes it difficult to disentangle whether observed effects are attributable to the digital delivery mode, the behavioral content, or both. Therefore, comparator conditions reported by each review were documented, and findings were interpreted with caution, emphasizing the combined influence of digital and behavioral mechanisms [50].

Outcome Definitions

Behavioral outcomes were classified according to the definitions provided in the included SRs. "STI and HIV prevention behaviors" encompassed all behavioral actions aimed at reducing infection risk, including but not limited to condom use, STI and HIV testing, vaccination, and adherence to treatment. "Safe sex behaviors" referred specifically to sexual practices such as consistent condom use, partner reduction, and negotiation of safer sex. This hierarchical approach was adopted to maintain consistency with the terminology used in the original reviews while avoiding redundancy between overlapping categories. A clarifying note was also added to Table 2 indicating that "safe sex behaviors" represent a subset of "STI and HIV prevention behaviors.

XSL•FO
RenderX

Table 2. Main characteristics of systematic reviews published between 2014 and 2015. "Safe sex behaviors" are a subset of "sexually transmitted infection and HIV prevention behaviors. The table summarizes study design, population, intervention features, behavioral and cognitive outcomes, and country income level based on World Bank classification.

Characteristics (N=23[a])	Results
SR[b] design	
Number of studies included	514[c]
Participants (total)	129,481
Range of years of the studies included	(2014-2025)
Place (country: geographic location)	
High-income countries	204
Upper-middle-income countries	22
Lower-middle-income countries	17
Only RCT[d] (n=23)	9
Only non-RCT (includes nonrandomized experimental and observational studies) (n=23)	2
RCT and non-RCT (n=23)	12
Total RCTs (N=514)	410
Population type (n=23)	
Adolescents (population age 10 to 19 years)	14
Youth (20 to 29 years)	17
Adults (>29 years)	15
Men who have sex with men	14
LGBTQI+[e] individuals	10
People diagnosed with an STD[f] (without HIV)	10
Intervention characteristics (n=23)	
Target of behavior (prevention of STI[g] and HIV)	19
Target behavior (safe sex)	16
Target behavior (STI and HIV testing)	17
Target behavior (treatment-related, ie, attend the appointment or getting treatment for an STI)	10
Environment: schools	10
Environment: universities (where it was implemented)	4
Environment: health centers (where it was implemented)	10
Explicit framework of theoretical domains of behavior change from the studies included in the SR	10
Description of behavior change techniques according to BCTTv1	2
Description according to Theoretical Domains Framework (TDF)	1
Behavioral outcomes (n=23)	
Condom use (internal or external)	19
Frequency in unprotected sexual intercourse	11
Number of sexual partners	7
STD and HIV testing	18
Uptake of medical male circumcision	4
HIV counseling	8
Get vaccinated against (VPH- HEP A y B)	7
Cognitive outcomes (mediators of prevention) (n=23)	

XSL•FO
RenderX

Characteristics (N=23[a])	Results
Self-efficacy	10
STD- and HIV-related knowledge	10
Attitudes toward condom use	8
Pre-exposure prophylaxis awareness	4
Biological outcomes	
HIV or STI acquisition (n=23)	8
Technology delivered (n=23)	
Mobile device	19
Desktop computer	12
Digital billboard	0
Wearable (clothing and accessory)	0
Digital object	0
Projection and hologram	0
Digital content type (n=23)	
Audio call and message	10
Video call and message	14
Text and instant message	18
Email	9
Video game	6
Audio broadcast and podcast	1
Website, computer, program and app	18
eBook	1
Virtual or augmented reality	2
Other descriptions	
Artificial intelligence-based chatbots for promoting safe sex or other sexual behaviors	0
Meta-analysis and certainty of evidence (n=23)	
Meta-analysis	7
Certainty of evidence (GRADE)[h]	4
APEASE[i] (n=23)	
Acceptability	6
Practicability	0
Effectiveness	19
Affordability	4
Spill-over effects	3

XSL•FO
RenderX

Characteristics (N=23[a])	Results
Equity	2

[a]Total number of SRs included in this overview.

[b]SR: systematic review.

[c]Total number of primary studies included in the SRs of this overview.

[d]RCT: randomized controlled trial.

[e]LGBTQI+: lesbian, gay, bisexual, transgender, queer, intersex, and plus.

[f]STD: sextually transmitted disease.

[g]STI: sexually transmitted infection.

[h]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.

[i]APEASE: Acceptability, Practicability, Effectiveness, Affordability, Spill-over effects, and Equity.

A written description was generated on the results of each SR on digital interventions for the prevention of STIs and HIV, and a table detailed the characteristics and outcomes of each review. The extracted data were synthesized using a predefined and team-approved template, identifying common themes and mapping them according to the stated objectives (section 9 in Multimedia Appendix 2).

Subgroup analyses were done to evaluate various factors, including the aim of the SR, the target population, behavioral outcomes, cognitive outcomes, biological outcomes, and the type of digital content and intervention. The mode of delivery (MoD) framework by Marques et al [51] was used to categorize interventions into types such as text and instant messages, video calls and messages, websites, computer programs and apps, emails, video games, audio broadcasts and podcasts, e-books, and virtual or augmented reality.

The effectiveness of digital interventions for STI and HIV prevention was assessed based on SRs that explicitly described the use of BCTs or the TDF. Subsequently, the impact of BCTs on each specific outcome was analyzed following the approach of Michie et al [52,53], allowing for the identification of the most effective techniques. Findings were synthesized into structured tables to visualize the impact of digital interventions across behavioral, cognitive, and biological domains.

Finally, a 5-level classification system was applied: "▼" indicated strong evidence of a negative effect; "O," mixed or null evidence; "O−" and "O+," a negative or positive effect with limited evidence; and "▲," strong evidence of a positive effect. The strength of evidence was determined according to three criteria: (1) the number of SRs reporting consistent results in the same direction, (2) the methodological quality of these reviews based on AMSTAR-2, and (3) the presence of meta-analytic data when available. Strong evidence ("▲" or "▼") required consistent findings across at least 2 high-quality reviews or meta-analyses, while limited evidence ("O−" or "O+") was assigned when findings were reported in only one review or when quality or consistency was lower. Mixed or null evidence ("O") indicated conflicting or inconclusive findings. This approach aligns with recent overviews applying the same evidence-grading framework for BCTs [53,54].

This system aims to provide clear guidance on which BCTs are most effective in digital interventions for STI and HIV prevention. The BCTs were coded according to the BCT Taxonomy version 1 (BCTTv1) [31,55], in studies that reported interventions using TDF, while the original descriptions were retained for studies that explicitly reported BCTs. This methodological approach ensured that conclusions were systematic, evidence-based, and aligned with established behavior change frameworks.

The SAGER [37] guidelines were used to ensure the consideration of sex and gender variables during the data extraction. These guidelines aim to prevent bias and improve the relevance and validity of findings by promoting the clear distinction between "sex" (biological differences) and "gender" (social and cultural factors), and its purpose is for these distinctions to be accurately reflected in study design, data analysis, and results reporting (sections 2 and 12 in Multimedia Appendix 2).

Results

Overview

From the 3643 records identified (3313 from electronic databases and 330 from bibliographic sources), 129 full-text articles were assessed for eligibility. Of these, 106 were excluded, leading to the inclusion of 23 SRs in this overview [56-78] (Figure 1). The interrater reliability for full-text screening of the initial 122 studies was robust (κ=0.88). An updated search conducted on November 12 yielded 378 additional records, of which 7 were assessed in full-text and 4 met the inclusion criteria. Additional studies were identified through supplementary methods, including website searches, citation searching, and expert consultations, ensuring a comprehensive review (section 4 in in Multimedia Appendix 2).

Characteristics of SRs

The 23 SRs included in the overview were published between 2014 and 2025, and covered 514 primary studies, including 410 randomized controlled trials (RCTs). Of these reviews, 9 focused exclusively on RCTs [56-58,66,69,70,75,77,78], 2 on non-RCTs [72,73], and 12 included both study types [59-65,67,68,71,74,76]. These reviews encompassed 129,481 participants, with individual studies ranging from 2662 to 27,704 participants. A detailed appraisal of methodological quality and overlap across the included SRs is presented at the end of this section to contextualize confidence in the synthesized evidence. The targeted populations included adolescents (n=14), youth

XSL•FO
RenderX

(n=17), adults (n=15), men who have sex with men (n=14), LGBTQ+ individuals (n=10), and people diagnosed with sexually transmitted diseases other than HIV (n=10; sections 13, 14a, and 14b in Multimedia Appendix 2).

Most studies were conducted in high-income countries, including the United States, Portugal, and Chile, followed by upper-middle-income countries such as China and South Africa, and lower-middle–income countries such as India and Kenya. Country income levels were classified according to the World Bank Country and Lending Groups (FY2025, Atlas method, USD), which use Gross National Income per capita as the defining criterion. Classifications were verified as of November 10, 2025, based on the latest publicly available dataset [79]. All data are reported, and detailed characteristics of the included SRs are presented in section 13 in in Multimedia Appendix 2. The synthesis of study characteristics is summarized in Table 2.

SAGER Application

The SAGER guidelines revealed that most studies did not include sex- or gender-disaggregated data, and significant gender differences were generally not reported. Only one study, Kamitani et al [63], mentioned transgender participants, but without a detailed analysis (section 12 in in Multimedia Appendix 2).

Characteristics of the Interventions

Overview

Across the included SRs, digital interventions targeted multiple prevention-related behaviors, including STI and HIV prevention, safer sex practices, and engagement with testing and sexual health services. Most SRs focused on STI and HIV prevention behaviors (19/23, 82.6%) safe-sex promotion (16/23, 69.6%), and STI and HIV testing (17/23, 73.9%). Additionally, 10 SRs (43.5%) incorporated strategies to enhance treatment adherence, such as attending medical appointments or completing syphilis treatment (Table 2 and section 15a in in Multimedia Appendix 2).

The research study settings varied, with health centers being the most common (10/23, 43.5%), followed by schools (10/23, 43.5%), while universities (4/23, 17.4%) were the least frequent. Regarding the use of theoretical frameworks, 10 SRs (43.5%) applied a behavior change framework, yet only 2 SRs (8.7%) explicitly described techniques based on the BCTTv1, and just one (4.3%) used the TDF (section 15a in in Multimedia Appendix 2).

In terms of behavioral outcomes, out of 23 SRs analyzed, condom use was assessed in 19 reviews (82.6%), while STI and HIV testing was reported in 18 reviews (78.3%). Other relevant outcomes included the frequency of unprotected sexual intercourse (47.8%), the number of sexual partners (30.4%), and vaccination against HPV or hepatitis (30.4%). Finally, the most analyzed cognitive outcomes were self-efficacy (43.5%), STI and HIV-related knowledge (43.5%), and attitudes toward condom use (34.8%), whereas PrEP awareness was examined in only 4 studies (17.4%; section 15b and 15c in in Multimedia Appendix 2).

MoD

According to the MoD classification by Marques et al [51], most interventions were delivered via mobile devices (n=19) and desktop computers (n=12), with text and instant messaging being the most common digital content type (n=18). Other content types included video calls and messages (n=14), emails (n=9), and video games (n=6). Notably, no studies used emerging delivery methods such as digital billboards, wearable accessories, digital objects, or projection and holograms (Table 2; section 15d in in Multimedia Appendix 2).

Theoretical Frameworks and BCTs

Of the 23 SRs analyzed, 10 (43.5%) incorporated behavioral theories, with the information-motivation-behavioral skills model (10 of 71 framework mentions, 14.3%), health belief model, and social cognitive theory being the most common. However, 13 reviews (56.5%) lacked any theoretical framework, reflecting inconsistent application of behavior change science. Only 4 (17.4%) reviews explicitly reported the identification or coding of BCTs or TDF [56,58,59,78] (section 15e in in Multimedia Appendix 2).

Overall, there was limited variability in the reporting of theoretical and behavioral frameworks across the 23 SRs. Explicit descriptions of BCTs or TDF mapping were rare, with most reviews indicating “not reported.” This pattern reflects heterogeneity in reporting practices across digital interventions for STI and HIV prevention and highlights that only a minority of reviews provided systematic or detailed descriptions of behavioral frameworks.

Effectiveness and Implementation of BCTs in Digital Interventions: Subset Analysis

This subsection focuses specifically on the subset of SRs (4 out of 23) that explicitly identified, coded, or analyzed BCTs within digital interventions for STI and HIV prevention. These reviews (Bailey et al [56]; Burns et al [58]; Clarke et al [59]; and Mo et al [78]) provided sufficient methodological detail to enable comparison of BCT use, frequency, and effectiveness. The remaining reviews, which did not report BCT coding or implementation frameworks, are synthesized in the previous sections that address broader behavioral, cognitive, and biological outcomes. This clarification ensures transparency and maintains consistency with the overview’s comprehensive scope.

DBCIs for STI and HIV prevention have incorporated various BCTs; however, their explicit classification using standardized frameworks such as the BCTTv1 or TDF remains limited. Among the studies analyzed, Burns et al [58] and Clarke et al [59] reported interventions explicitly coded using BCTTv1, whereas Bailey et al [56] used domains associated with TDF to describe behavioral determinants. Mo et al [78] expanded the evidence base by systematically identifying and mapping BCTs across digital HIV prevention interventions for adolescents and young people (Table 3).

XSL•FO
RenderX

Table 3. Summary of behavioral determinants, frequently reported behavioral change techniques (BCTs), and observed effectiveness of digital interventions for the prevention of sexually transmitted infections (STIs) and HIV based on 4 reviews (Bailey et al [56]; Burns et al [58]; Clarke et al [59]; and Mo et al [78]). Includes intervention design, target population, mode of delivery, and AMSTAR-2 quality rating. BCT codes correspond to Behavior Change Technique Taxonomy version 1.

Reference and review type	Review period	Mode of delivery	N Primary studies and study designs	Target population and number of participants	Target of behavior	Intervention effectiveness and observed behavior change	Determinants of behavior	Most frequently used BCTs and components	Statistical significance and effect size (*P* value, Cohen *d*, r)	AMSTAR 2 Rating
Bailey et al [56] SR[a] and meta-analysis	Searches: from 2014 to June 2017 Publication of primary studies included: 1991-2017	Web-based programs, mobile apps, online modules	31 studies RCTs[b]	Population: young people, men who have sex with men (MSM), HIV-positive people, at-risk adults, African American women. Total: 11,293 participants - IDI vs. minimal intervention: 10,423 participants - IDI vs. face-to-face intervention: 870 participants	Prevention of HIV and other STIs, promotion of safe sex behaviors, adherence to testing and treatment. Condom use, partner reduction, and safe sex negotiation	Increased HIV-related knowledge (moderate effect) - small improvement in behavioral intention - Positive effect on HIV prevention behaviors - No clear impact on self-efficacy - No significant effect on biological outcomes (STI and HIV acquisition, viral load)	Goals, behavioral regulation, knowledge, emotion, optimism, beliefs about capabilities	Goal setting (1.1), commitment (1.9), feedback on behavior (2.2), biofeedback (2.6), information about antecedents (4.2), re-attribution (4.3), information about health consequences (5.1), salience of consequences (5.2), verbal persuasion (15.1), self-talk (15.4)	HIV-related knowledge: SMD[c]=0.56 (95% CI 0.33-0.80) - HIV prevention self-efficacy: SMD=0.13 (95% CI 0.00-0.27) - HIV prevention intention: SMD=0.16 (95% CI 0.06-0.26) - HIV prevention behaviors: OR[d] 1.28 (95% CI 1.04-1.57) - Biological outcomes (STI and HIV acquisition, viral load): OR 1.48 (95% CI 0.96-2.28), *P*=.08 (not significant)	High No critical flaws One noncritical weakness (no information on funding sources)
Burns et al [58] SR of RCTs	Searches: January 1999 and July 2014 Publication of primary studies included:: 2006-2014	Mobile phone-based interventions, SMS reminders, mobile apps, video messages	10 studies RCTs	Population: General population, at-risk adults, young people, (MSM) Total: 16,773 participants	Promotion of sexual health services uptake, reduction of risky sexual behaviors, reduction of recall bias in self-reported sexual activity	Two trials showed significant increases in clinic attendance with SMS reminders. - One trial improved sexual health knowledge. - No trials showed significant increases in condom use. - One trial found mobile technology acceptable for sexual health data collection	Goals, intentions, behavioral regulation, knowledge, social influences, environmental context and resources, reinforcement	Goal Setting (1.1), feedback on Behavior (2.2), information about health consequences (5.1), demonstration of behavior (6.1), social comparison (6.2), prompts/cues (7.1), material incentives (10.1)	Clinic attendance: SMS reminders significantly increased attendance (RR[e] 0.86, 95% CI 0.74-1.00) - Chlamydia retesting: SMS reminders increased retesting (RR 4.5, 95% CI 1.05-19.22) - HIV testing uptake: no significant effect (RR 0.94, 95% CI 0.81-1.09) - Sexual health knowledge: SMS improved knowledge (RR 1.75, 95% CI 1.11-2.77) - Condom use: no significant changes (RR 0.87, 95% CI 0.62-1.24)	Moderate No critical flaws More than one noncritical weakness (no information on funding sources, publication bias reported but not discussed)

XSL•FO
RenderX

Reference and review type	Review period	Mode of delivery	N Primary studies and study designs	Target population and number of participants	Target of behavior	Intervention effectiveness and observed behavior change	Determinants of behavior	Most frequently used BCTs and components	Statistical significance and effect size (*P* value, Cohen *d*, r)	AMSTAR 2 Rating
Clarke et al [59] SR	Searches: 1 January 2000 to 1 September 2021. Publication of primary studies included:: 2011-2018	Digital interventions including SMS (text messages), social media, and app-based messaging	13 studies (RCTs=5 Non RCTs before-after=8)	Population: Adolescents, youth, adults, MSM, LGBTQI+[f]. Total: not reported	Increasing attendance at scheduled sexual health appointments	Behavioral interventions increased attendance at scheduled sexual health appointments. Text messages were the most frequently used MoD[g]. Some interventions were effective, while others had mixed results	Beliefs about consequences, environmental context and resources, emotion, reinforcement, social influence, optimism	Credible source (9.1), prompts/cues (7.1), social support (3.2), social reward (10.4), self-incentive (10.5), restructuring of the environment (12.1, 12.2), focus on past success (15.3), vicarious consequences (16.3)	Some interventions significantly increased attendance while others had mixed results	Critically Low More than one critical flaw (no justification for excluding individual studies, no consideration of bias when interpreting results) More than one noncritical weakness (study selection not done in duplicate, no information on funding sources)
Mo et al [78] SR	Searches: January 2008 to November 2024. Publication of primary studies included: 2008-2023	Mobile apps, SMS text messaging, web-based modules, computer-based digital game (IYG-Tech), online educational platforms	34 studies (RCTs, quasi-experimental, and observational designs)	Population: adolescents and young people (10-29 y). Total: Not reported	Prevention of HIV; promotion of safer sex practices; HIV testing; risk reduction behaviors	Narrative synthesis indicated consistent improvements in cognitive determinants (HIV-related knowledge, self-efficacy, perceived risk). Small, inconsistent effects were reported for condom use. Some interventions showed increases in HIV testing motivation or intentions, but behavioral outcomes were heterogeneously measured and rarely pooled	Knowledge, beliefs about consequences, behavioral skills, self-efficacy, environmental context and resources, motivation	Information about health consequences (5.1), feedback on behavior (2.2), prompts/cues (7.1), goal setting (1.1), problem solving (1.2), demonstration of behavior (6.1), social support (3.1), self-monitoring (2.3)	No pooled effect sizes reported. Several primary studies demonstrated significant improvements in HIV knowledge and self-efficacy but effects could not be meta-analyzed	Critically Low More than one critical flaw (no justification for excluding individual studies, no consideration of bias when interpreting results) More than one noncritical weakness (study selection not done in duplicate, no information on funding sources)

[a]SR: systematic review.

[b]RCT: randomized controlled trial.

[c]SMD: standardized mean difference.

[d]OR: odds ratio.

[e]RR: relative risk.

[f]LGBTQI+: lesbian, gay, bisexual, transgender, queer, and other sexual orientations and gender identities.

[g]MoD: mode of delivery.

XSL•FO
RenderX

In total, 26 BCTs were reported. The most frequently used in these interventions included goal setting (1.1), feedback on behavior (2.2), and the use of prompts and cues (7.1), commonly delivered through mobile apps, text messaging (SMS), online modules, and digital games. Other approaches included commitment strategies (1.9), biofeedback (2.6), social support (3.1, 3.2), and providing information about health consequences (5.1). These techniques targeted key behavioral determinants such as knowledge, behavioral regulation, social influence, and reinforcement strategies, aiming to enhance self-efficacy, motivation, and risk awareness (Table 3).

Table 3 summarizes behavioral determinants, commonly used BCTs, and intervention effectiveness. Figure 2 synthesizes the impact of specific BCTs across behavioral, cognitive, and biological outcomes. Findings on BCT effectiveness were drawn from the 3 SRs that reported effect estimates (Bailey et al [56]; Burns et al [58]; Clarke et al [59]). Mo et al [78] contributed descriptive evidence on BCT implementation but did not report quantitative effect estimates. This section also highlights how BCTs were implemented across SMS-based strategies, mobile apps, online platforms, and digital learning modules.

Figure 2. Summary of the effectiveness of behavior change techniques in digital interventions for sexually transmitted infection and HIV prevention.

	Author	AMSTAR-2 Rating	1.1 Goal setting (behavior)	1.9 Commitment	2.2 Feedback on behavior	2.6 Biofeedback	4.2 Information about antecedents	4.3 Re-attribution	5.1 Information about health consequences	5.2 Salience consequences	7.1 Prompts/Cues	9.1 Credible source	10.1 Material incentive (behavior)	10.4 Social reward	10.5 Self-Incentive	11.2 Reduce negative emotions
Overall effect			▲		▲		▲		O		▲	O	▲	O+	O+	O+
Behavioral outcome																
Condom use	Bailey 2021	🟢	▲	▲	▲	-	▲	▲	▲	-	-	-	-	-	-	-
	Burns 2016	🟡	O	-	-	-	-	-	O	-	O	-	-	-	-	-
Sexual health services uptake	Burns 2016	🟡	O+	-	▲	-	-	-	▲	-	▲	-	▲	-	-	-
STI and HIV testing	Burns 2016	🟡	O	O+	▲	-	-	-	▲	-	▲	-	O	-	-	-
	Clarke 2022	🔴	-	-	-	-	-	-	O	-	▲	O	O	O+	O+	O+
Cognitive outcome																
HIV knowledge	Bailey 2021	🟢	▲	▲	▲	▲	-	-	▲	-	-	-	-	-	-	-
HIV prevention self-efficacy	Bailey 2021	🟢	▲	▲	▲	-	-	-	▲	-	-	-	-	-	-	-
HIV prevention intentions	Bailey 2021	🟢	O+	O+	O+	-	-	-	O	-	-	-	-	-	-	-
Attitudes toward condom use	Bailey 2021	🟢	O+	O+	O+	-	-	-	O	▲	-	-	-	-	-	-
Biological outcome																
STI and HIV acquisition	Bailey 2021	🟢	O+	O+	O	-	-	-	▲	-	-	-	-	-	-	-

Behavioral Outcomes

Condom Use

Digital interventions showed mixed effectiveness. Bailey et al [56] found a significant increase in condom use (odds ratio [OR] 1.28, 95% CI 1.04-1.57, ▲ evidence), supporting the role of goal setting (1.1), commitment (1.9), and feedback (2.2). Conversely, Burns et al [58], reported no significant effect (relative risk [RR] 0.87, 95% CI 0.62-1.24, O evidence), suggesting that effectiveness varied by population and intervention design.

Regarding mobile apps for risk awareness [56], digital interventions focusing on health consequences (5.1) significantly improved HIV-related knowledge (standardized mean difference [SMD]=0.56, 95% CI 0.33-0.80), reinforcing the role of interactive digital tools in promoting condom use behaviors.

XSL•FO
RenderX

Sexual Health Services Uptake

To ensure conceptual clarity, in this overview STI and HIV testing refers to diagnostic testing behaviors, whereas sexual health services uptake encompasses broader engagement with preventive or clinical services, including retesting when reported as part of general service use [58].

Mixed results (O+ to ▲) were found regarding the effectiveness of goal setting (1.1) and Feedback (2.2) in increasing engagement with sexual health services. SMS reminders significantly increased clinic attendance (RR 0.86, 95% CI 0.74-1) and chlamydia retesting rates (RR 4.5, 95% CI 1.05-19.22) [58]. SMS reminders providing health-related information (5.1) and material incentives (10.1) were effective in some cases, particularly for STI retesting, though results varied across populations.

Regarding SMS feedback and goal setting [58], participants received weekly SMS messages inquiring about risky sexual behaviors, followed by feedback and goal-setting prompts to encourage health service use.

STI/HIV Testing

Evidence from Burns et al [58] was mixed (O to ▲), with Feedback (2.2) and prompts and cues (7.1) being somewhat effective in encouraging STI testing. Clarke et al [59] found that personalized SMS reminders significantly improved STI retesting rates (56% vs. 33%; $P<.01$). However, material incentives (BCT 10.1) yielded mixed effects (O), as financial rewards increased short-term attendance but did not sustain engagement.

Regarding personalized SMS reminders and incentives [59], tailored SMS messages encouraged STI retesting, with financial incentives increasing attendance (29.17% in the incentive group vs. 0% in the control group).

Cognitive Outcomes

HIV-Related Knowledge

Digital interventions incorporating Information about Health Consequences (5.1) and Biofeedback (2.6) significantly improved HIV-related knowledge (SMD=0.56, 95% CI 0.33-0.80, ▲ evidence). These interventions relied on mobile apps to enhance risk awareness through educational modules and interactive tools [56].

HIV Prevention Self-Efficacy

Strong evidence (▲) supported Goal Setting (1.1), Commitment (1.9), and Feedback (2.2) in improving self-efficacy for HIV prevention. Self-monitoring and digital education modules were key to reinforcing behavior change [56].

HIV Prevention Intentions and Attitudes Toward Condom Use

For both outcomes, evidence was weaker (O+ to O), indicating that Goal Setting (1.1) and Feedback (2.2) may contribute to positive attitudes toward condom use, but their long-term impact remains uncertain (SMD=0.16, 95% CI 0.06-0.26). Health Consequences (5.1) showed inconsistent effects, suggesting that behavior change may require additional reinforcement. [56].

Regarding mobile apps for risk awareness [56]: digital interventions focusing on HIV prevention knowledge were designed to improve risk awareness and motivation, though long-term behavior adoption remained a challenge.

Complementary evidence was provided by Mo et al [78], who systematically reviewed digital HIV prevention interventions for adolescents and young adults and identified a broad set of BCTs mapped across the included programs. Although this review did not quantify behavioral or biological outcomes associated with specific BCTs, it reported consistent improvements in key cognitive determinants—particularly HIV-related knowledge, prevention self-efficacy, and attitudes toward condom use. These findings reinforce the role of information-based strategies (eg, providing information about health consequences, 5.1), feedback mechanisms (2.2), and prompts and cues (7.1) as foundational techniques supporting cognitive readiness for behavior change in digital sexual health interventions.

Biological Outcomes: STI/HIV Acquisition

Weak positive evidence (O+ to O) was assigned to goal setting (1.1), commitment (1.9), and feedback (2.2), with ▲ (strong evidence) for health consequences (5.1). Despite improvements in knowledge and behavioral determinants, these interventions did not significantly reduce STI and HIV acquisition rates (OR 1.48, 95% CI 0.96-2.28; $P=.08$) [56].

In Figure 2, summary of the effectiveness of individual BCTs reported in the 3 reviews included in the BCT subset analysis. Outcomes are grouped into behavioral (eg, condom use, STI and HIV testing), cognitive (eg, knowledge, self-efficacy), and biological domains. Effect strength was graded using a 5-level evidence classification system adapted from Michie et al. (2018) [9] and Mair et al. (2023) [53]: "▲" denotes strong positive. "O" denotes mixed or null, "O+/O−" denotes limited positive or negative, and "▼" denotes strong negative. AMSTAR-2 quality ratings are indicated by colored circles as follows: Green = High, Yellow = Moderate, Orange = Low, Red = Critical Low Classification of effectiveness from this systematic reviews and meta-analyses. ▲ = Positive effect of BCT based on good evidence such as subgroup or regression analyses. O+ = Positive effect of BCT based on low evidence such as frequency of individual BCTs within effective interventions. O = Mixed evidence or no effect. O− = Negative effect of BCT based on low level of evidence such as frequency of individual BCTs within effective interventions. ▼ = Negative effect of BCT based on good evidence such as subgroup or regression analyses.

Quality Appraisal of the SRs

Confidence levels were determined according to the AMSTAR-2 criteria described in the Methods section, based on the number and severity of critical and noncritical weaknesses. Overall, confidence in the results was high in 21.7% (5/23) of the SRs, moderate in 13% (3/23), low in 17.4% (4/23), and critically low in 47.8% (11/23). The most common weaknesses identified were the absence of reported funding sources for primary studies (22/23, 95.7%), the lack of a full list of excluded studies (13/23, 56.5%), and the omission of a review protocol (10/23, 43.5%; section 16 in in Multimedia Appendix 2).

XSL•FO
RenderX

Overlap in Primary Studies Cited in Reviews

The overlap assessment revealed a minimal overlap in the 321 primary studies, most of which were cited only once in the 23 SRs, with a CCA of 2.70. Comparisons between pairs of SRs indicated that 82.6% (209 out of 253) had a low overlap (<5%), 8.7% (22 out of 253) had a moderate overlap (5% to <10%), 4% (10 out of 253) had a high overlap (10% to <15%), and 4.7% (12 out of 253) had a very high overlap (≥15%).

Given that 4 reviews (Bailey et al [56]; Burns et al [58]; Clarke et al [59]; and Mo et al [78]) provided explicit coding or mapping of BCTs, these were analyzed as a separate methodological subset (Table 3). The overlap assessment indicated consistently slight overlap (<5%) among them, confirming that this BCT-focused evidence base draws on distinct sets of primary studies. These findings are illustrated in sections 17 and 18 in in Multimedia Appendix 2.

Discussion

Principal Results

This overview synthesized evidence from 23 SRs to evaluate the behavioral determinants, theoretical foundations, and effectiveness of DBCIs for preventing STIs and HIV. Consistent with the study aims, the findings indicate that DBCIs, particularly SMS reminders, mobile apps, and interactive web-based programs, can enhance engagement with sexual health services and increase STI and HIV testing. Cognitive determinants, such as HIV-related knowledge, motivation, and self-efficacy, also improved across numerous interventions. The most frequently identified and effective BCTs included goal setting, feedback on behavior, and prompts and cues. However, the high heterogeneity of intervention formats and the inconsistent reporting of BCTs across reviews limit the ability to draw definitive conclusions regarding the independent contribution of specific components. Taken together, these findings offer a clear synthesis of behavioral mechanisms across digital interventions and strengthen confidence in the patterns observed across reviews.

Only 4 reviews, Bailey et al [56], Burns et al [58], Clarke et al [59], and Mo et al [78], explicitly coded or mapped BCTs using standardized frameworks (BCTTv1 or TDF), allowing for a focused subset analysis. For the remaining reviews, behavioral mechanisms could only be inferred from narrative descriptions. The GROOVE overlap analysis confirmed minimal overlap across the 321 primary studies, indicating that the synthesized evidence draws from distinct and independent datasets. This methodological combination strengthens the validity of the synthesized findings and responds directly to recent calls for more mechanism-oriented evidence synthesis in digital health [80].

Interpretation of Findings

Overall, DBCIs show considerable promise in supporting STI and HIV prevention, particularly through timely reminders, personalized feedback, and interactive educational content. SMS-based interventions consistently improved clinic attendance and STI retesting, reinforcing the well-established role of prompts and cues in facilitating preventive behaviors. Web-based programs and mobile apps contributed to enhanced knowledge, motivation, and behavioral skills, aligning with established behavioral models emphasizing the importance of cognitive determinants. This overview highlights that cognitive determinants represent the most consistent pathways of change in DBCIs, clarifying how digital components influence prevention behaviors.

Behavioral outcomes such as condom use demonstrated mixed effects: Bailey et al [56] reported significant improvements, whereas Burns et al [58] found no notable impact. These differences likely reflect variations in populations, intervention content, MoDs, and follow-up durations, as well as the extent to which interventions incorporated explicit BCTs or theoretical frameworks. Findings from Mo et al [78] further highlight the inconsistency in how interventions implement and report behavior change strategies, particularly among adolescent populations. These discrepancies underscore that digital modalities may be more effective when they target motivational and self-regulatory determinants rather than complex interpersonal behaviors, a distinction that has been underexplored in prior reviews.

The limited and inconsistent application of behavioral theory across reviews is a central issue. Nearly half of the reviews did not reference any behavioral framework, despite strong evidence from broader health literature showing that theoretically grounded digital interventions are more effective and more interpretable. This gap constrains the field's ability to identify mechanisms of action and optimize intervention design. By systematically examining these gaps, our overview provides conceptual clarity on how limited theoretical integration constrains interpretability, scalability, and optimization of digital interventions. Addressing these gaps is essential for moving the field toward theory-driven digital prevention, where mechanisms of action are explicitly linked to intervention components and outcomes.

The updated search added 4 recent SRs [75-78], which expand the evidence base with up-to-date findings on digital PrEP adherence support, smartphone-based HIV prevention tools, and BCT-coded adolescent interventions. However, their conclusions mirror earlier patterns: digital tools consistently improve testing and cognitive outcomes, whereas sustained behavioral and biological impacts remain inconsistent. By incorporating the most recent evidence available, including 4 SRs published in 2024-2025, this overview offers a timely and comprehensive picture of current digital prevention strategies and how they engage (or fail to engage) key behavioral mechanisms.

Comparison With Prior Work

Our findings align with prior work demonstrating that digital interventions can effectively promote preventive health behaviors when grounded in behavioral theory and equipped with active engagement strategies. Studies by Simoni et al [23], Albarracín et al [34], and Thomas Craig et al [24] underscore the importance of behavioral mechanisms, self-regulation processes, and contextual tailoring, elements only partially reflected across the reviews included in this overview. Unlike previous syntheses, this overview integrates behavioral

XSL•FO
RenderX

mechanisms, intervention content, and methodological quality to provide a more coherent understanding of how and why DBCIs work. This integration allows for a more granular understanding of intervention pathways than outcome-only syntheses can provide.

A key divergence from other health domains (eg, mental health and chronic disease management) is the limited integration of emerging digital technologies, such as AI-driven personalization, conversational agents, and virtual or augmented reality, in STI and HIV prevention. Similarly, sex- and gender-disaggregated analyses remain rare, with only 1 review explicitly addressing transgender populations. This raises concerns about equity and generalizability in digital sexual health research. This highlights a broader limitation in the STI and HIV prevention literature: the field has been slower than other digital health domains to adopt advanced technologies and rigorous behavioral frameworks, reducing its capacity to generate equitable and generalizable impact. Future digital prevention efforts must bridge this technological and conceptual gap to ensure that innovation translates into equitable public health impact. Extending beyond earlier reviews, this overview examines not only whether digital interventions work but also how and why they work, through explicit mapping of behavioral mechanisms across reviews.

Strengths and Limitations

This overview offers the most comprehensive behavioral synthesis to date of digital interventions for STI and HIV prevention, following the PRISMA-S (Multimedia Appendix 3). The use of AMSTAR-2 enabled robust appraisal of methodological quality, while GROOVE allowed quantification of overlap across primary studies. Together, the minimal overlap and explicit behavioral coding enhance confidence in the synthesized evidence.

However, several limitations should be acknowledged. First, conclusions depend on the quality and reporting of the included SRs, nearly half of which were rated critically low by AMSTAR-2. Second, BCT coding was absent or insufficient in most reviews, restricting the depth of mechanistic synthesis. Third, high intervention heterogeneity precluded meta-analytic pooling and limits comparability. Finally, reliance on SRs means that relevant primary studies not captured in those reviews may have been missed.

Despite these limitations, the overview provides a clear, methodologically grounded synthesis of how digital tools contribute to STI and HIV prevention and where future research should focus. This multilayered approach responds to recent calls for more rigorous, mechanism-oriented evidence synthesis capable of informing real-world decision-making and aligns with emerging frameworks such as evidence-based X, which emphasize the integration of mechanisms, context, and methodological rigor in digital health research [80].

Conclusions

DBCIs represent a promising and scalable strategy for strengthening STI and HIV prevention, particularly for improving testing behaviors and key cognitive determinants. Interventions incorporating goal setting, feedback on behavior, and prompts and cues show the most consistent positive effects, whereas outcomes related to condom use and biological measures remain mixed. The inconsistent application and reporting of behavioral theory across reviews limits both interpretability and scalability.

Building on prior work, this overview provides a novel contribution by integrating effectiveness evidence with a systematic mapping of BCTs and theoretical mechanisms, a perspective that has been largely absent from earlier syntheses. By identifying which active components are most consistently associated with beneficial outcomes and highlighting persistent reporting and methodological gaps, this study offers actionable guidance for designing next-generation digital interventions. These findings, which incorporate evidence updated through 2025, have pragmatic real-world implications for researchers, clinicians, and policymakers seeking to develop scalable, culturally responsive, and equity-focused digital prevention programs that meaningfully address the behavioral pathways that drive STI and HIV risk across diverse populations. Collectively, these insights offer a pathway for accelerating the development of digital public health tools that are both evidence-based and behaviorally informed.

Funding

This work was funded by the ANID, the Chilean government, the National Scientific and Technological Development Fund (FONDECYT) research project number 1250316. The funders had no role in the study design; data collection, analysis, or interpretation; the drafting of the report; or the decision to submit the paper for publication.

Data Availability

Data sharing is not applicable to this article as no new datasets were generated or analyzed during this study. All data used in this overview were extracted from previously published systematic reviews and are included in the manuscript and its supplementary materials.

Acknowledgments

The authors acknowledge the University of Santiago of Chile for granting the sabbatical under University Decree No. 372 and Memorandum No. 43663 in support of GD's PhD dissertation. This work would not have been possible without the help of the PhD Program in the Psychology of Communication and Change at Universitat de Barcelona. The authors would also like to thank the Spanish Ministry of Science, Innovation, and Universities (MICIU/AEI/10.13039/501100011033) under research project PID2020-115486GB-I00 for their support. No generative artificial intelligence tools (such as ChatGPT or other large language models) were used to write, edit, or analyze any portion of this manuscript.

XSL•FO
RenderX

Authors' Contributions

The study design and literature search strategy were developed by GD, SSC, and SCM. Articles were located, identified, and evaluated by GD and DL, who also check the data extraction. All authors were responsible for the initial dimension and criteria list. The manuscript was drafted and edited by GD with close revision and feedback from SSC and SCM as well as review and feedback from the other coauthors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Reporting guideline for overviews of reviews of healthcare interventions: development of the PRIOR statement.
[DOCX File , 18 KB - jmir_v28i1e74201_app1.docx]

Multimedia Appendix 2
Behavioral Determinants and Effectiveness of Digital Behavior Change Interventions for STI/HIV Prevention: An Overview of Systematic Reviews.
[DOCX File , 1900 KB - jmir_v28i1e74201_app2.docx]

Multimedia Appendix 3
PRISMA-S Checklist for Reporting Literature Searches.
[DOCX File , 16 KB - jmir_v28i1e74201_app3.docx]

References

1. Murray C, Aravkin AY, Zheng P, Abbafati C, Abbas KM, Abbasi-Kangevari M, et al. Global burden of 87 risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020;396(10258):1223-1249 [FREE Full text] [doi: 10.1016/S0140-6736(20)30752-2] [Medline: 33069327]
2. Fishbein M, Triandis H, Kanfer F, Becker M, Middlestadt S, Eichler A. Factors influencing behavior and behavior change. In: Handbook of Health Psychology. Mahwah, NY: Erlbaum; 2001.
3. Tran KB, Lang JJ, Compton K, Xu R, Acheson AR, Henrikson HJ, et al. The global burden of cancer attributable to risk factors, 2010-19: a systematic analysis for the global burden of disease study 2019. Lancet 2022;400(10352):563-591 [FREE Full text] [doi: 10.1016/S0140-6736(22)01438-6] [Medline: 35988567]
4. Zheng Y, Yu Q, Lin Y, Zhou Y, Lan L, Yang S, et al. Global burden and trends of sexually transmitted infections from 1990 to 2019: an observational trend study. Lancet Infect Dis 2022;22(4):541-551. [doi: 10.1016/S1473-3099(21)00448-5] [Medline: 34942091]
5. Alarming decline in adolescent condom use, increased risk of sexually transmitted infections and unintended pregnancies, reveals new WHO report. World Health Organization. URL: https://tinyurl.com/r49f98yc [accessed 2024-09-03]
6. Eisingerich AB, Wheelock A, Gomez GB, Garnett GP, Dybul MR, Piot PK. Attitudes and acceptance of oral and parenteral HIV preexposure prophylaxis among potential user groups: a multinational study. PLoS One 2012;7(1):e28238 [FREE Full text] [doi: 10.1371/journal.pone.0028238] [Medline: 22247757]
7. Infecciones de transmisión sexual (ITS) [Web page in Spanish]. World Health Organization. 2024. URL: https://www.who.int/es/news-room/fact-sheets/detail/sexually-transmitted-infections-(stis) [accessed 2024-08-31]
8. Ritchie H, Spooner F, Roser M. Causes of death. Our World in Data Internet. URL: https://ourworldindata.org/causes-of-death [accessed 2022-11-30]
9. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
10. Johnson BT, Michie S, Snyder LB. Effects of behavioral intervention content on HIV prevention outcomes: a meta-review of meta-analyses. J Acquir Immune Defic Syndr 2014;66 Suppl 3(Suppl 3):S259-S270 [FREE Full text] [doi: 10.1097/QAI.0000000000000235] [Medline: 25007195]
11. Duarte-Anselmi G, Okan Y, Johnston M, Ortiz L, Villalobos Dintrans P, Armayones M. Introduction to behavioral science and its practical applications in public health. Medwave 2025;25(1):e3017-e3017. [doi: 10.5867/medwave.2025.01.3017] [Medline: 39836869]
12. Michie S, Johnston M. Theories and techniques of behaviour change: developing a cumulative science of behaviour change. Health Psychol Rev 2012;6(1):1-6. [doi: 10.1080/17437199.2012.654964]
13. How to prevent STIs. U.S. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/sti/prevention/index.html [accessed 2024-09-04]

XSL•FO
RenderX

14. Duarte-Anselmi G, Crane SM, Ruiz MA, Dintrans PV. Behavioral science meets public health: a scoping review of the fogg behavior model in behavior change interventions. BMC Public Health 2025;25(1):3468 [FREE Full text] [doi: 10.1186/s12889-025-24525-y] [Medline: 41088011]
15. Shi H, Du J, Jin G, Yang H, Guo H, Yuan G, et al. Effectiveness of eHealth interventions for HIV prevention, testing and management: an umbrella review. Int J STD AIDS 2024;35(10):752-774. [doi: 10.1177/09564624241252457] [Medline: 38733263]
16. Milne-Ives M, Homer SR, Andrade J, Meinert E. Potential associations between behavior change techniques and engagement with mobile health apps: a systematic review. Front Psychol 2023;14:1227443 [FREE Full text] [doi: 10.3389/fpsyg.2023.1227443] [Medline: 37794916]
17. Woodward C, Bloch S, McInnes-Dean A, Lloyd KC, McLeod J, Saunders J, et al. Digital interventions for STI and HIV partner notification: a scoping review. Sex Transm Infect 2024;100(4):242-250 [FREE Full text] [doi: 10.1136/sextrans-2023-056097] [Medline: 38754986]
18. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016;51(5):843-851 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
19. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. J Med Internet Res 2017;19(6):e232 [FREE Full text] [doi: 10.2196/jmir.7126] [Medline: 28663162]
20. Hollands GJ, Shemilt I, Marteau TM, Jebb SA, Kelly MP, Nakamura R, et al. Altering micro-environments to change population health behaviour: towards an evidence base for choice architecture interventions. BMC Public Health 2013;13:1218 [FREE Full text] [doi: 10.1186/1471-2458-13-1218] [Medline: 24359583]
21. Hekler EB, Michie S, Pavel M, Rivera DE, Collins LM, Jimison HB, et al. Advancing models and theories for digital behavior change interventions. Am J Prev Med 2016;51(5):825-832 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.013] [Medline: 27745682]
22. West R, Michie S. Guide to Development and Evaluation of Digital Behaviour Change Interventions in Healthcare. United Kingdom: Silverback Publishing; 2016.
23. Simoni JM, Ronen K, Aunon FM. Health behavior theory to enhance ehealth intervention research in HIV: rationale and review. Curr HIV/AIDS Rep 2018;15(6):423-430 [FREE Full text] [doi: 10.1007/s11904-018-0418-8] [Medline: 30511186]
24. Thomas Craig KJ, Morgan LC, Chen C, Michie S, Fusco N, Snowdon JL, et al. Systematic review of context-aware digital behavior change interventions to improve health. Transl Behav Med 2021;11(5):1037-1048 [FREE Full text] [doi: 10.1093/tbm/ibaa099] [Medline: 33085767]
25. Cane J, O'Connor D, Michie S. Validation of the theoretical domains framework for use in behaviour change and implementation research. Implement Sci 2012;7:37 [FREE Full text] [doi: 10.1186/1748-5908-7-37] [Medline: 22530986]
26. Atkins L, Francis J, Islam R, O'Connor D, Patey A, Ivers N, et al. A guide to using the theoretical domains framework of behaviour change to investigate implementation problems. Implement Sci 2017;12(1):77 [FREE Full text] [doi: 10.1186/s13012-017-0605-9] [Medline: 28637486]
27. Michie S, Atkins L, West R. The Behaviour Change Wheel Book - A Guide To Designing Interventions. United Kingdom: Silverback Publishing; 2014.
28. Smith V, Devane D, Begley CM, Clarke M. Methodology in conducting a systematic review of systematic reviews of healthcare interventions. BMC Med Res Methodol 2011;11(1):15 [FREE Full text] [doi: 10.1186/1471-2288-11-15] [Medline: 21291558]
29. Pollock M, Fernandes RM, Becker LA, Pieper D, Hartling L. Chapter V: overviews of reviews. In: Cochrane Handbook for Systematic Reviews of Interventions version 6.3. Chichester, UK: John Wiley & Sons, Ltd; 2022.
30. Glasziou P, Meats E, Heneghan C, Shepperd S. What is missing from descriptions of treatment in trials and reviews? BMJ 2008;336(7659):1472-1474 [FREE Full text] [doi: 10.1136/bmj.39590.732037.47] [Medline: 18583680]
31. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
32. Cornet VP, Holden RJ. Systematic review of smartphone-based passive sensing for health and wellbeing. J Biomed Inform 2018;77:120-132 [FREE Full text] [doi: 10.1016/j.jbi.2017.12.008] [Medline: 29248628]
33. Armayones Ruiz M, Pinto EL, Figueroa O, Robles N, Prehn DL, Riquelme FV, et al. Barriers and facilitators for safe sex behaviors in students from universidad de santiago de Chile (USACH) through the COM-B model. BMC Public Health 2023;23(1):677 [FREE Full text] [doi: 10.1186/s12889-023-15489-y] [Medline: 37041528]
34. Albarracín D, Fayaz-Farkhad B, Granados Samayoa JA. Determinants of behaviour and their efficacy as targets of behavioural change interventions. Nat Rev Psychol 2024;3(6):377-392. [doi: 10.1038/s44159-024-00305-0] [Medline: 40909436]
35. Duarte-Anselmi G, Leiva-Pinto E, Vanegas-López J, Thomas-Lange J. Experiences and perceptions on sexuality, risk and STI/HIV prevention campaigns by university students. Designing a digital intervention. Cien Saude Colet 2022;27(3):909-920 [FREE Full text] [doi: 10.1590/1413-81232022273.05372021] [Medline: 35293468]

XSL•FO
RenderX

36. Gates M, Gates A, Pieper D, Fernandes RM, Tricco AC, Moher D, et al. Reporting guideline for overviews of reviews of healthcare interventions: development of the PRIOR statement. BMJ 2022;378:e070849 [FREE Full text] [doi: 10.1136/bmj-2022-070849] [Medline: 35944924]
37. Heidari S, Babor TF, De Castro P, Tort S, Curno M. Sex and gender equity in research: rationale for the SAGER guidelines and recommended use. Res Integr Peer Rev 2016;1:2 [FREE Full text] [doi: 10.1186/s41073-016-0007-6] [Medline: 29451543]
38. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
39. Duarte G, Sanduvete-Chaves S, López-Arenas D, Chacón-Moscoso S. Digital strategies and behavior change techniques for preventing sexually transmitted infections: protocol for an overview of systematic reviews. Medwave 2025;25(2):e3020. [doi: 10.5867/medwave.2025.02.3020] [Medline: 40063926]
40. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009;339:b2535 [FREE Full text] [doi: 10.1136/bmj.b2535] [Medline: 19622551]
41. Duarte G, Sanduvete-Chaves S, Chacón-Moscoso S. Specifying the content of digital interventions for prevention of sexually transmitted infections /HIV using the behaviour change technique taxonomy v1 (BCTTv1): protocol for an overview of systematic reviews. PROSPERO International prospective register of systematic reviews Internet. URL: https://www.crd.york.ac.uk/PROSPERO/view/CRD42023485887 [accessed 2024-07-18]
42. Duarte Anselmi G. Digital behaviour change interventions to prevent sexually transmitted infections (STIs) including HIV: evidence reviews and integrated report on the quantitative and qualitative evidence. Open Science Framework. 2024. URL: https://osf.io/u9gz7/ [accessed 2024-04-12]
43. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions. Cochrane. 2022. URL: https://handbook-5-1.cochrane.org/ [accessed 2022-11-24]
44. Collaboratron™ [Software]. Epistemonikos. URL: https://www.epistemonikos.org/en/documents/647cdbf675135cca583bcfeca7dadc0560f80e02 [accessed 2024-05-06]
45. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005;37:360-363. [Medline: 15883903]
46. Ryan R, Synnot A, Hill MP. Cochrane Consumers and Communication. Data extraction template. Australia: La Trobe University; 2018.
47. Ryan R, Hill S. Supporting implementation of Cochrane methods in complex communication reviews: resources developed and lessons learned for editorial practice and policy. Health Res Policy Syst 2019;17(1):32 [FREE Full text] [doi: 10.1186/s12961-019-0435-0] [Medline: 30922338]
48. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017;358:j4008 [FREE Full text] [doi: 10.1136/bmj.j4008] [Medline: 28935701]
49. Pérez-Bracchiglione J, Meza N, Bangdiwala SI, Niño de Guzmán E, Urrútia G, Bonfill X, et al. Graphical representation of overlap for OVErviews: GROOVE tool. Res Synth Methods 2022;13(3):381-388. [doi: 10.1002/jrsm.1557] [Medline: 35278030]
50. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014;348(mar07 3):g1687-g1687 [FREE Full text] [doi: 10.1136/bmj.g1687] [Medline: 24609605]
51. Marques MM, Carey RN, Norris E, Evans F, Finnerty AN, Hastings J, et al. Delivering behaviour change interventions: development of a mode of delivery ontology. Wellcome Open Res 2020;5:125. [doi: 10.12688/wellcomeopenres.15906.1]
52. Michie S, West R, Sheals K, Godinho CA. Evaluating the effectiveness of behavior change techniques in health-related behavior: a scoping review of methods used. Transl Behav Med 2018;8(2):212-224 [FREE Full text] [doi: 10.1093/tbm/ibx019] [Medline: 29381786]
53. Mair JL, Salamanca-Sanabria A, Augsburger M, Frese BF, Abend S, Jakob R, et al. Effective behavior change techniques in digital health interventions for the prevention or management of noncommunicable diseases: an umbrella review. Ann Behav Med 2023;57(10):817-835 [FREE Full text] [doi: 10.1093/abm/kaad041] [Medline: 37625030]
54. Duarte-Anselmi G, de Oliveira Barbosa AA, Ferrari G, Fernandes ACP, Araneda-Flores J, González DA, et al. Overview of systematic reviews on behavioral determinants of physical activity and healthy eating in schoolchildren. Sci Rep 2025;15(1):35379 [FREE Full text] [doi: 10.1038/s41598-025-19332-9] [Medline: 41068188]
55. Castro O, Fajardo G, Johnston M, Laroze D, Leiva-Pinto E, Figueroa O, et al. Translating the behaviour change technique taxonomy version 1 into Spanish: methodology and validation. Wellcome Open Res 2024;9:298 [FREE Full text] [doi: 10.12688/wellcomeopenres.21388.1] [Medline: 39323609]
56. Bailey JV, Wayal S, Aicken CRH, Webster R, Mercer CH, Nazareth I, et al. Interactive digital interventions for prevention of sexually transmitted HIV. AIDS 2021;35(4):643-653 [FREE Full text] [doi: 10.1097/QAD.0000000000002780] [Medline: 33259345]

XSL•FO
RenderX

57. Berendes S, Gubijev A, McCarthy OL, Palmer MJ, Wilson E, Free C. Sexual health interventions delivered to participants by mobile technology: a systematic review and meta-analysis of randomised controlled trials. Sex Transm Infect 2021;97(3):190-200. [doi: 10.1136/sextrans-2020-054853] [Medline: 33452130]
58. Burns K, Keating P, Free C. A systematic review of randomised control trials of sexual health interventions delivered by mobile technologies. BMC Public Health 2016;16(1):778 [FREE Full text] [doi: 10.1186/s12889-016-3408-z] [Medline: 27514851]
59. Clarke R, Heath G, Ross JDC, Farrow C. Increasing attendance at pre-booked sexual health consultations: a systematic review. Sex Health 2022;19(4):236-247. [doi: 10.1071/SH21245] [Medline: 35615776]
60. Conserve DF, Jennings L, Aguiar C, Shin G, Handler L, Maman S. Systematic review of mobile health behavioural interventions to improve uptake of HIV testing for vulnerable and key populations. J Telemed Telecare 2017;23(2):347-359 [FREE Full text] [doi: 10.1177/1357633X16639186] [Medline: 27056905]
61. Ilskens K, Wrona KJ, Dockweiler C, Fischer F. An evidence map on serious games in preventing sexually transmitted infections among adolescents: systematic review about outcome categories investigated in primary studies. JMIR Serious Games 2022;10(1):e30526 [FREE Full text] [doi: 10.2196/30526] [Medline: 35107438]
62. Jones K, Eathington P, Baldwin K, Sipsma H. The impact of health education transmitted via social media or text messaging on adolescent and young adult risky sexual behavior: a systematic review of the literature. Sex Transm Dis 2014;41(7):413-419. [doi: 10.1097/OLQ.0000000000000146] [Medline: 24922099]
63. Kamitani E, Peng Y, Hopkins D, Higa DH, Mullins MM, Community Preventive Services Task Force. A community guide systematic review: digital HIV Pre-exposure prophylaxis interventions. Am J Prev Med 2024;67(2):303-310. [doi: 10.1016/j.amepre.2024.02.009] [Medline: 38367928]
64. Khuwaja SS, Peck JL. Increasing HPV vaccination rates using text reminders: an integrative review of the literature. J Pediatr Health Care 2022;36(4):310-320. [doi: 10.1016/j.pedhc.2022.02.001] [Medline: 35288016]
65. Knight R, Karamouzian M, Salway T, Gilbert M, Shoveller J. Online interventions to address HIV and other sexually transmitted and blood-borne infections among young gay, bisexual and other men who have sex with men: a systematic review. J Int AIDS Soc 2017;20(3):e25017 [FREE Full text] [doi: 10.1002/jia2.25017] [Medline: 29091340]
66. Manby L, Aicken C, Delgrange M, Bailey JV. Effectiveness of eHealth interventions for HIV prevention and management in sub-saharan Africa: systematic review and meta-analyses. AIDS Behav 2022;26(2):457-469 [FREE Full text] [doi: 10.1007/s10461-021-03402-w] [Medline: 34427813]
67. Nguyen LH, Tran BX, Rocha LEC, Nguyen HLT, Yang C, Latkin CA, et al. A systematic review of ehealth interventions addressing HIV/STI prevention among men who have sex with men. AIDS Behav 2019;23(9):2253-2272 [FREE Full text] [doi: 10.1007/s10461-019-02626-1] [Medline: 31401741]
68. Ou L, Chen AC, Amresh A. The effectiveness of mhealth interventions targeting parents and youth in human papillomavirus vaccination: systematic review. JMIR Pediatr Parent 2023;6:e47334 [FREE Full text] [doi: 10.2196/47334] [Medline: 37988155]
69. Palmer M, Henschke N, Villanueva G, Maayan N, Bergman H, Glenton C, et al. Targeted client communication via mobile devices for improving sexual and reproductive health. Cochrane Database Syst Rev 2020;8(8):CD013680 [FREE Full text] [doi: 10.1002/14651858.CD013680] [Medline: 32779730]
70. Saragih ID, Imanuel Tonapa S, Porta CM, Lee B. Effects of telehealth interventions for adolescent sexual health: a systematic review and meta-analysis of randomized controlled studies. J Telemed Telecare 2024;30(2):201-214. [doi: 10.1177/1357633X211047762] [Medline: 34903065]
71. Sewak A, Yousef M, Deshpande S, Seydel T, Hashemi N. The effectiveness of digital sexual health interventions for young adults: a systematic literature review (2010-2020). Health Promot Int 2023;38(1):38. [doi: 10.1093/heapro/daac104] [Medline: 36757346]
72. Veronese V, Ryan KE, Hughes C, Lim MS, Pedrana A, Stoové M. Using digital communication technology to increase HIV testing among men who have sex with men and transgender women: systematic review and meta-analysis. J Med Internet Res 2020;22(7):e14230 [FREE Full text] [doi: 10.2196/14230] [Medline: 32720902]
73. Schnall R, Travers J, Rojas M, Carballo-Diéguez A. eHealth interventions for HIV prevention in high-risk men who have sex with men: a systematic review. J Med Internet Res 2014;16(5):e134 [FREE Full text] [doi: 10.2196/jmir.3393] [Medline: 24862459]
74. Xin M, Viswanath K, Li AY, Cao W, Hu Y, Lau JT, et al. The effectiveness of electronic health interventions for promoting HIV-preventive behaviors among men who have sex with men: meta-analysis based on an integrative framework of design and implementation features. J Med Internet Res 2020;22(5):e15977 [FREE Full text] [doi: 10.2196/15977] [Medline: 32449685]
75. Du J, Jin G, Zhang H, Don O, Shi H, Wang S, et al. Effectiveness of digital health interventions in promoting the pre-exposure prophylaxis (PrEP) care continuum among men who have sex with men (MSM): a systematic review of randomized controlled trials. Curr HIV/AIDS Rep 2025;22(1):25. [doi: 10.1007/s11904-025-00733-4] [Medline: 40100510]
76. Huang W, Stegmueller D, Ong JJ, Wirtz SS, Ning K, Wang Y, et al. Technology-based HIV prevention interventions for men who have sex with men: systematic review and meta-analysis. J Med Internet Res 2025;27:e63111 [FREE Full text] [doi: 10.2196/63111] [Medline: 40293786]

XSL•FO
RenderX

77. Li F, Xie C, Xiang F. Impact of mHealth on enhancing pre-exposure prophylaxis adherence and strengthening the HIV prevention cascade among key populations: a systematic review and meta-analysis. Front Public Health 2025;13:1600773 [FREE Full text] [doi: 10.3389/fpubh.2025.1600773] [Medline: 40642253]
78. Mo PK, Xie L, Lee TC, Li AYC. Use of behavior change techniques in digital HIV prevention programs for adolescents and young people: systematic review. JMIR Public Health Surveill 2025;11:e59519 [FREE Full text] [doi: 10.2196/59519] [Medline: 40293783]
79. World bank country and lending groups. World Bank Data Help Desk. URL: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups [accessed 2025-11-08]
80. Stewart G. Is realising evidence-based X the future of evidence synthesis? Campbell Syst Rev 2025;21(2):e70037. [doi: 10.1002/cl2.70037] [Medline: 40124218]

Abbreviations

BCT: behavior change technique
BCTTv1: behavior change technique taxonomy version 1
CCA: corrected covered area
DBCI: digital behavior change intervention
GROOVE: Graphical Representation of Overlap for Overviews
HPV: human papillomavirus
MoD: mode of delivery
OR: odds ratio
PICOS: Population, Intervention, Comparison, Outcome, Study type
PrEP: pre-exposure prophylaxis
PRESS: Peer Review of Electronic Search Strategies
PRIOR: Preferred Reporting Items for Overviews of Reviews
PRISMA-S: Preferred Reporting Items for SRs and Meta-Analyses Literature Search Extension
RCT: randomized controlled trial
RR: relative risk
SAGER: Sex and Gender Equity in Research
SMD: standardized mean difference
SR: systematic review
STD: sexually transmitted disease
STI: sexually transmitted infection
TDF: theoretical domains framework

Edited by S Brini; submitted 17.Apr.2025; peer-reviewed by C Hao, A Eisingerich; comments to author 31.Oct.2025; accepted 21.Dec.2025; published 29.Jan.2026.

Please cite as:
Duarte-Anselmi G, Sanduvete-Chaves S, Chacón-Moscoso S, López-Arenas D
Behavioral Determinants and Effectiveness of Digital Behavior Change Interventions for the Prevention of Sexually Transmitted Infections and HIV: Overview of Systematic Reviews
J Med Internet Res 2026;28:e74201
URL: https://www.jmir.org/2026/1/e74201
doi:10.2196/74201
PMID:

©Giuliano Duarte-Anselmi, Susana Sanduvete-Chaves, Salvador Chacón-Moscoso, Daniel López-Arenas. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effect of Digital Health Interventions on College Students' Lifestyle Behaviors: Systematic Review

Qingyuan Zhou[1*], BEd; Jiajun Jiang[1*], BEd; Zhihua Yin[1], PhD; Ruishi Fan[1], BEd

College of Physical Education and Health, East China Normal University, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Zhihua Yin, PhD
College of Physical Education and Health
East China Normal University
500 Dongchuan Road
Shanghai, 200241
China
Phone: 86 1 512 104 3880
Email: yzhkj86888@sina.com

Abstract

Background: College students undergo a critical transition from adolescence to adulthood, during which lifestyle behaviors such as physical activity, sedentary behavior, diet, and sleep are key determinants of long-term health. Digital health interventions (DHIs) are increasingly recognized as a promising strategy for improving these behaviors among college students.

Objective: This systematic review aims to evaluate the effectiveness and applicability of DHIs targeting lifestyle behaviors among college students by analyzing intervention objectives, modalities, functionalities, outcomes, and other key characteristics.

Methods: In accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines, multiple scientific databases, including Scopus, Web of Science, PubMed, MEDLINE, PsycINFO, SPORTDiscus, ProQuest Central, APA PsycArticles, ERIC, and Academic Search Premier, were searched for studies published between January 2010 and December 2025 (initial search: August 5, 2025; updated search: December 27, 2025). The inclusion criteria were original empirical studies on DHIs targeting lifestyle behaviors (physical activity, sedentary behavior, diet, and sleep) among college students, published in English. Studies focusing on nondigital interventions, lacking sufficient methodological details, or not reporting lifestyle behavior–related outcomes were excluded. Quality assessment was conducted in 2 stages: all studies were first evaluated using the Mixed Methods Appraisal Tool (2018 version), followed by Risk of Bias 2 for randomized controlled trials and Joanna Briggs Institute critical appraisal tools for nonrandomized studies. A narrative synthesis was used to present and synthesize the findings.

Results: A total of 2998 records were retrieved, of which 46 publications met the inclusion criteria. These included 30 (65%) studies related to physical activity, 26 (57%) studies to diet, 10 (22%) studies related to sedentary behavior, and 6 (13%) studies related to sleep. This review enabled an examination of the effects of DHIs on college students' lifestyle behaviors. DHIs primarily used mobile apps, web-based platforms, and mobile communication technologies, with core functionalities such as education, guidance, monitoring, and prompting. DHIs were more effective in improving physical activity and diet; however, evidence for reducing sedentary behavior and improving sleep remained limited. Of the 46 studies, 31 (67%) reported positive effects, with larger sample sizes and intervention durations of 8-16 weeks being associated with more favorable outcomes.

Conclusions: This review focuses on college students, addressing a gap in the literature that often centers on general adult populations. Unlike previous reviews that focus on a single behavior, this study integrates multiple lifestyle behaviors and evaluates DHIs across diverse modalities and functionalities. These contributions help refine future DHIs for college students and inform health promotion strategies in higher education. Although DHIs show potential for improving lifestyle behaviors, evidence of their long-term effectiveness remains limited. Future interventions should prioritize multibehavior integration, interactivity, and population-differentiated design to enhance precision, sustainability, and equity. This study has several limitations, including issues related to sample representativeness, intervention refinement, and methodological rigor.

Trial Registration: PROSPERO CRD420251119078; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251119078

(J Med Internet Res 2026;28:e82192) doi:10.2196/82192

XSL•FO
RenderX

KEYWORDS

digital health interventions; college students; lifestyle behaviors; systematic review; physical activity; sedentary behavior; diet; sleep

Introduction

College students are in a critical developmental stage characterized by the transition from adolescence to adulthood, during which they encounter multiple challenges, including increased academic demands and evolving social roles. Evidence suggests that college students often exhibit insufficient self-management capacity related to healthy lifestyle behaviors [1], with inadequate physical activity, prolonged sedentary behavior, irregular diet patterns, and sleep disturbances being particularly prevalent. Previous research has demonstrated that health behaviors established during this developmental period tend to exhibit substantial stability and continuity over time [2]. The adoption of unhealthy lifestyle behaviors during this stage has been shown to significantly increase the risk of chronic diseases, depression, and anxiety later in adulthood [3,4]. Therefore, the implementation of early and effective interventions targeting these 4 key lifestyle behaviors among college students is of substantial public health significance [1].

With the rapid advancement of digital technologies and the widespread adoption of smart devices, digital health interventions (DHIs) have emerged as an innovative approach to health promotion and are increasingly recognized as an important means of improving lifestyle behaviors among college students [5,6]. Particularly in the post–COVID-19 era, DHIs have demonstrated greater adaptability and broader application potential than traditional face-to-face health intervention models [7-9]. In recent years, a growing body of empirical evidence has shown that DHIs are effective in promoting physical activity among college students [10-12], reducing sedentary time [13,14], improving diet behaviors [15,16], and enhancing sleep quality [17,18]. These interventions—encompassing mobile apps, wearable devices, online platforms, and social media—offer several advantages, including low cost, high scalability, and a high degree of personalization [19,20], and have been shown to enhance user engagement and facilitate sustained behavior change [21,22]. Concurrently, advancements in emerging technologies, such as artificial intelligence, continue to drive the refinement of DHI implementation strategies and further enhance intervention effectiveness [23].

However, the existing body of research on DHIs targeting lifestyle behaviors among college students remains subject to several limitations. On the one hand, the majority of original intervention studies have focused on single lifestyle behaviors or specific technological modalities, with a relative lack of comprehensive designs that integrate multiple behaviors and intervention approaches. At the same time, key intervention dimensions—such as functional characteristics, intervention duration, participant demographics, and adherence—have yet to reach unified standards or methodological consensus [24,25]. On the other hand, existing systematic reviews and meta-analyses in this field also demonstrate limitations in terms of specificity and methodological rigor. First, systematic syntheses that specifically target the college student population remain relatively scarce, with insufficient attention paid to lifestyle behaviors such as sedentary behavior and sleep. Second, existing analyses have not adequately synthesized the combined effects of multiple lifestyle behaviors across diverse DHI intervention formats [26].

In light of the current research context and identified limitations, this review is guided by the following research questions: (1) What is the current state of the literature on DHIs targeting 4 key lifestyle behaviors among college students (physical activity, sedentary behavior, diet, and sleep)? (2) What are the specific implementation strategies and modalities of DHIs addressing these behaviors? (3) To what extent are DHIs effective in influencing these 4 target lifestyle behaviors among college students? Through a comprehensive synthesis and analysis of relevant primary research evidence, this review will explicitly consider the characteristics of college students as “digital natives” [27]. The review will systematically examine the forms, functions, and key components of different DHIs, and comprehensively evaluate their effects on the 4 target behaviors that are closely related to college student health. This review aims to clarify the applicability and effectiveness of DHIs within this population, thereby providing evidence-based recommendations for optimizing DHI tools, informing health promotion strategies in higher education settings, and guiding future research.

Methods

Search Strategy

This systematic review was prospectively registered in PROSPERO (International Prospective Register of Systematic Reviews) on August 4, 2025 (registration number: CRD420251119078), and the reporting of the review findings adheres to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines (see Multimedia Appendix 1). A comprehensive literature search was conducted across 10 major English-language electronic databases, including Scopus, Web of Science, PubMed, ProQuest Central, and 6 databases accessed via the EBSCOhost platform (MEDLINE, PsycINFO, SPORTDiscus, APA PsycArticles, ERIC, and Academic Search Premier), with Google Scholar used as a supplementary search source. In addition, the reference lists of relevant articles were screened to identify potentially missed studies. The initial search was completed on August 5, 2025, covering studies published between January 1, 2010, and June 1, 2025, for primary study identification, and an updated search was conducted on December 27, 2025, to capture studies published within the most recent 6 months; the same search strategy was applied consistently across both searches. No published search filters were used, and the search strategy was neither adapted from nor reused, in whole or in part, from previous reviews. The search strategy was initially developed by the authors and subsequently peer reviewed by an experienced searcher with

XSL•FO
RenderX

expertise in scientific information retrieval. Beyond these approaches, no study registries were searched, no purposeful searching or browsing (eg, table of contents screening, print conference proceedings, or website searches) was conducted, and no additional information was sought by contacting authors, experts, manufacturers, or other relevant parties.

The literature search strategy was systematically developed in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Search Extension) guidelines to ensure transparency and reproducibility of the search process. The strategy combined Medical Subject Headings terms with free-text terms and constructed keyword combinations around 3 core concepts: (1) intervention formats (eg, digital health, digital intervention, eHealth, and mobile health [mHealth]); (2) target behaviors (eg, physical activity, sedentary behavior, sleep, and diet); and (3) study populations (eg, university students, college students, and undergraduate students). Boolean operators (AND and OR) were applied to balance search sensitivity and specificity. Using Scopus as an example, the search query was as follows: TITLE-ABS-KEY ("digital health" OR "eHealth" OR "mHealth" OR "mobile health" OR "digital intervention" OR "health app") AND TITLE-ABS-KEY ("college students" OR "university students" OR "undergraduate students" OR "young adults") AND TITLE-ABS-KEY ("lifestyle behavior" OR "health behavior" OR "physical activity" OR "exercise" OR "diet" OR "nutrition" OR "sleep" OR "sedentary behavior"). The complete English-language search terms used across all databases are provided in Multimedia Appendix 2. To ensure comprehensive coverage, no geographical restrictions were applied during the literature search, allowing for the inclusion of relevant studies from diverse global regions.

Inclusion and Exclusion Criteria

The study inclusion criteria were developed in accordance with the PICOS (Population-Intervention-Comparison-Outcome -Study Design) framework, as outlined in Textbox 1.

The exclusion criteria were as follows: (1) The study population was not explicitly identified as "college students," "university students," or individuals enrolled in higher education institutions. (2) The nondigital components constituted the majority of the intervention (≥50%), or the study relied solely on wearable devices for passive behavioral monitoring without incorporating feedback mechanisms or active intervention strategies. (3) The study did not implement a behavioral intervention, or the intervention description lacked sufficient detail to determine its content and implementation procedures. (4) Studies that did not report any lifestyle behavior–related outcome measures. (5) Conference abstracts, theses, unpublished manuscripts, and other forms of gray literature. (6) Full-text articles were unavailable, or the publication was not in English.

Textbox 1. Study inclusion criteria.

1. Population

- Participants were required to be aged ≥18 years and explicitly identified as "college students," "university students," or "young adults enrolled in higher education."

2. Intervention

- Studies were required to evaluate at least one health intervention primarily delivered through digital health technologies and targeting lifestyle-related behaviors. Digital health interventions included, but were not limited to, mobile apps, web-based platforms, SMS text message reminders, online courses, virtual coaches, digital gamification strategies, social media, and other eHealth/mHealth tools.

3. Comparison

- The presence of a control group was not mandatory; all original studies reporting intervention effects were eligible for inclusion.

4. Outcomes

- The primary outcomes included lifestyle behavior indicators, specifically physical activity, sedentary behavior, diet, and sleep.
- Secondary outcomes included physical and mental health indicators, such as weight, waist circumference, and self-efficacy.

5. Study design

- Original empirical studies targeting 1 or more of the 4 lifestyle behavior domains among college students and implementing digital health interventions were included.
- No restrictions were placed on study design; however, intervention content, participant characteristics, and relevant outcome measures were required to be clearly reported.

Study Selection

All retrieved records were imported into EndNote 20 (Clarivate Plc) reference management software for duplicate removal and standardized record numbering. Subsequently, 2 reviewers (QYZ and JJJ) independently screened titles and abstracts for initial eligibility. Records that passed the initial screening were subjected to full-text assessment to determine final eligibility for inclusion. To ensure standardization and consistency in the screening process, all reviewers received standardized training on the predefined inclusion and exclusion criteria. Interrater reliability between the 2 reviewers was assessed using the Cohen κ coefficient, yielding a value of 0.86, which indicates a high level of screening agreement. In cases of disagreement regarding

XSL•FO
RenderX

individual records, a third reviewer (ZHY) was consulted to facilitate discussion and achieve a final consensus.

Data Extraction and Synthesis

To ensure standardization and consistency in the data extraction process, the research team developed a structured data extraction form in advance, covering the study title, first author, publication year, study region, study design, intervention population characteristics, intervention protocol characteristics, outcome measures, intervention effectiveness, and study conclusions. The data extraction form was pilot-tested using 5 studies to assess its feasibility. During the formal data extraction process, 2 reviewers (QYZ and JJJ) independently extracted the data. In cases of missing data or discrepancies in interpretation, a third reviewer (ZHY) was consulted to resolve disagreements. The final extracted data were consolidated into a standardized table and are presented in Multimedia Appendix 3.

Quality Assessment

All included studies were initially assessed for methodological quality using the Mixed Methods Appraisal Tool (MMAT, 2018 version) to obtain an overall preliminary appraisal of study quality. The MMAT is designed to evaluate 5 categories of study designs: qualitative research (QR), quantitative randomized controlled trials (QRCTs), quantitative nonrandomized studies (QNRSs), quantitative descriptive studies (QDSs), and mixed methods studies (MMSs), each comprising 5 appraisal criteria [28]. To enhance specificity and methodological rigor, the Risk of Bias 2 (RoB 2) tool was further applied to assess the risk of bias in QRCTs. For all other study designs, the Joanna Briggs Institute (JBI) critical appraisal tools were applied. This 2-stage quality assessment approach was intended to balance breadth and depth in methodological evaluation. Quality assessments were conducted independently by 2 reviewers (QYZ and JJJ), with discrepancies resolved through discussion. To ensure consistency, both reviewers received standardized training on the MMAT, RoB 2, and JBI critical appraisal tools and completed pilot scoring exercises before the formal assessment.

Results

Screening and Inclusion Results

Search and Screening Results

In this study, a total of 2998 records were retrieved from 10 major English-language databases. After deduplication and initial title and abstract screening, 273 articles were selected for full-text review. Based on the predefined exclusion criteria, exclusions were made for the following reasons: nonuniversity samples (n=68); interventions not primarily digital-based (n=20); wearable devices only, without active intervention components (n=5); absence of behavioral interventions (n=83); lack of relevant behavioral outcomes (n=35); and protocol or abstract only (n=19). Additionally, 3 more articles were identified through manual reference tracing of relevant review papers. Ultimately, 46 publications met the inclusion criteria and were included in the final analysis, as depicted in Figure 1.

Figure 1. PRISMA flowchart of the study selection process.

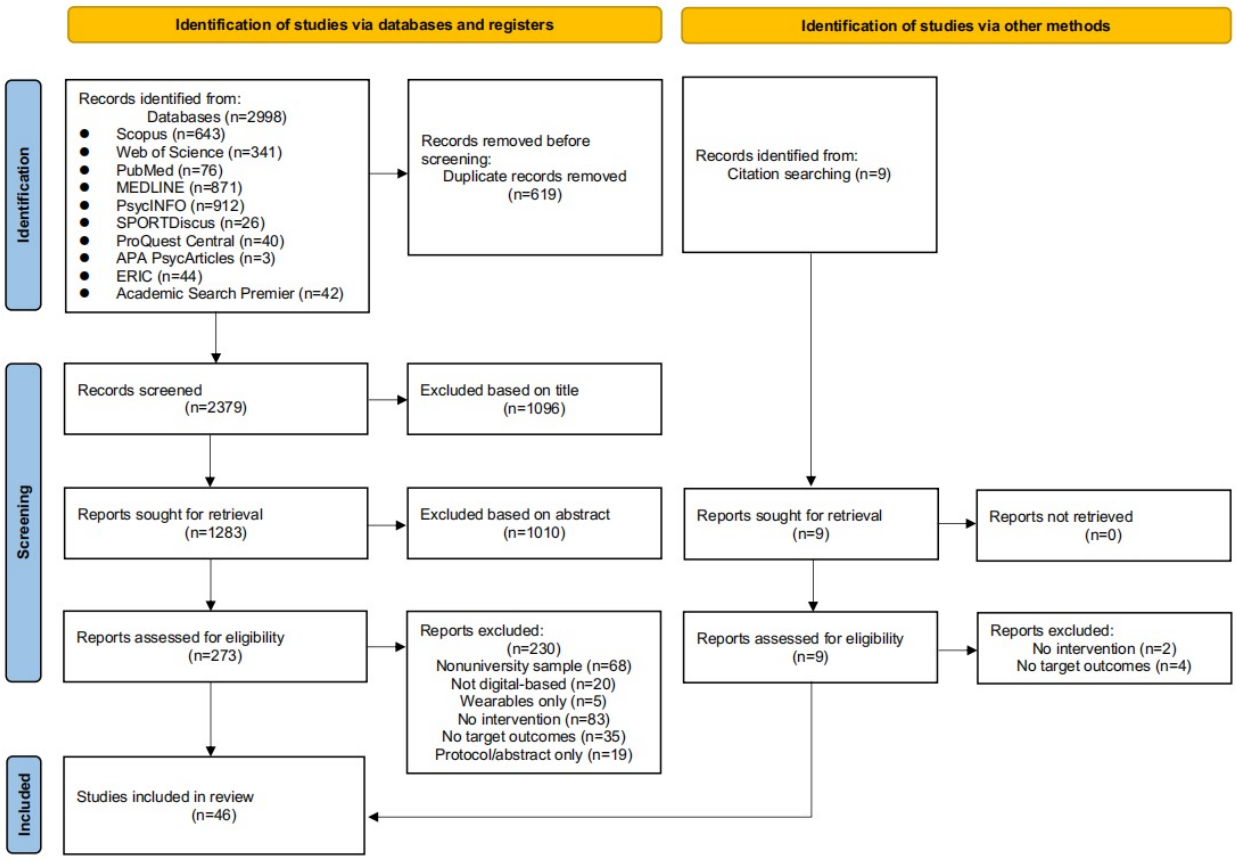

XSL•FO
RenderX

Quality Assessment Results

Following 2 rounds of quality assessment, the first-round MMAT evaluation indicated that the 46 included studies demonstrated an overall high level of methodological quality. Specifically, 28 (61%) studies were rated as high quality, 14 (30%) as moderate quality, and 4 (9%) as low quality (see Figure 2; also see [10,11,13-15,17,20,26,29-66]). Major methodological concerns identified during the assessment were primarily concentrated in MMAT items C4 and C5. Item C4 was primarily related to the implementation of blinding procedures, the adequacy of outcome interpretation, and the control of risk of bias, whereas item C5 reflected issues such as insufficient intervention adherence and the lack of rigorous statistical analyses. In the second round of assessment, the RoB 2 tool was applied to evaluate 30 QRCTs, indicating that the primary sources of bias were related to outcome measurement, deviations from intended interventions, and the handling of missing outcome data (see Figure 3; see also [13,14,17,29-31,33-35,37,38,42-48,50-52,54,56,58,59,61-65]). Concurrently, the JBI critical appraisal of the remaining 16 studies indicated that key factors influencing study quality primarily included sample representativeness, intervention adherence, and the objectivity of outcome measurement (see Figure 4; see also [10,11,15,20,26,32,36,39-41,49,53,55,57,60,66]).

The emergence of these methodological issues can be primarily attributed to 2 factors. On the one hand, the behavioral nature of DHIs makes the implementation of blinding inherently challenging, and several key behavioral outcomes rely on participant self-report measures. On the other hand, relatively high dropout rates associated with DHIs contribute to issues such as low intervention adherence and elevated loss-to-follow-up rates in some studies. When combined with insufficiently rigorous statistical analyses, these challenges may result in suboptimal handling of missing data or deviations from intended interventions. Although studies rated as moderate to low quality constitute a notable proportion of the included literature, it is important to recognize that many of their methodological limitations are closely related to the inherent characteristics of DHIs. Moreover, many of these studies primarily aimed to explore the feasibility and applicability of DHIs rather than to provide definitive evidence of intervention efficacy. Therefore, these studies retain substantial value for informing future research and intervention development. Given these considerations, no studies were excluded from this review solely based on methodological quality. Instead, all eligible studies were included, and findings from risk-of-bias assessments were systematically incorporated into the narrative synthesis. This approach allows for a comprehensive presentation of the current evidence landscape while explicitly identifying both its strengths and limitations.

XSL•FO
RenderX

Figure 2. Quality assessment results of the Mixed Methods Appraisal Tool.

Num.	*Study*	*Study design*	*Screening Q1*	*Screening Q2*	*C1*	*C2*	*C3*	*C4*	*C5*	*Results*
1	Hebden et al [35]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🟢	High
2	Nour et al [66]	Quantitative descriptive study	YES	YES	🟢	🟢	🟢	🟢	🟢	High
3	Schweitzer et al [54]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🟢	High
4	Allman-Farinelli et al [29]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🟢	High
5	Walsh et al [45]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🟢	High
6	Hutchesson et al [53]	Quantitative nonrandomized study	YES	YES	🟢	🟢	🟡	🟡	🟢	Moderate
7	O'Brien and Palfai [42]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🟢	High
8	Partridge et al [30]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🟢	High
9	Cotton and Prapavessis [43]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🟢	High
10	Morris et al [33]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🔴	Low
11	Xian et al [60]	Quantitative nonrandomized study	YES	YES	🟢	🟢	🟡	🟡	🟢	Moderate
12	Sarcona et al [55]	Quantitative descriptive study	YES	YES	🟢	🟢	🟢	🟢	🟢	High
13	Ashton et al [46]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🟢	High
14	Chung et al [49]	Quantitative nonrandomized study	YES	YES	🟢	🟢	🟡	🟡	🟢	Moderate
15	Inauen et al [34]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🟢	High
16	Hershner and O'Brien [38]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🟢	High
17	Fitzsimmons-Craft et al [32]	Quantitative nonrandomized study	YES	YES	🟢	🟢	🟡	🟡	🟢	Moderate
18	Whatnall et al [56]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🟢	High
19	Nour et al [50]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟡	🟡	🟢	Moderate
20	Lee and Park [20]	Quantitative nonrandomized study	YES	YES	🟢	🟢	🟡	🟡	🟢	Moderate
21	Napolitano et al [36]	Quantitative nonrandomized study	YES	YES	🟢	🟢	🟢	🟡	🟢	High
22	Roure et al [51]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🟢	High
23	Napolitano et al [59]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🟢	High
24	Hahn et al [63]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🟢	High
25	Figueroa et al [44]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟡	🟡	🟢	Moderate
26	Stork et al [61]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🟢	High
27	Muntaner-Mas et al [31]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🟢	High
28	Pope and Gao [14]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟡	🟡	🟢	Moderate
29	Smith and Volkwyn [39]	Quantitative descriptive study	YES	YES	🟢	🟢	🟢	🟢	🟢	High
30	Al-Nawaiseh et al [62]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🟢	High
31	Cantisano et al [41]	Quantitative nonrandomized study	YES	YES	🟢	🟢	🟡	🟡	🟢	Moderate
32	Haslam et al [65]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🔴	Low
33	Belogianni et al [58]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟡	🟡	🟢	Moderate
34	Kellner et al [52]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🟢	High
35	Floyd and Vargas [37]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🔴	Low
36	Kaneda et al [47]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🟢	High
37	Åsberg et al [26]	Qualitative research	YES	YES	🟢	🟢	🟢	🟡	🟢	High
38	Rajan and Muthunarayanan [11]	Quantitative descriptive study	YES	YES	🟢	🟢	🟢	🟢	🟢	High
39	Khatri and Sharma [40]	Quantitative nonrandomized study	YES	YES	🟢	🟢	🟡	🟡	🟢	Moderate
40	Malloy et al [48]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🟢	High
41	Kim et al [17]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🟢	High
42	Wittmar et al [57]	Mixed methods study	YES	YES	🟢	🟢	🟡	🟡	🟢	Moderate
43	Andargeery and El-Rafey [13]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🟢	High
44	Olatona et al [15]	Quantitative nonrandomized study	YES	YES	🟢	🟢	🟡	🟡	🟢	Moderate
45	Fucito et al [64]	Quantitative randomized controlled trial	YES	YES	🟢	🟢	🟢	🟡	🔴	Low
46	Gao et al [10]	Quantitative nonrandomized study	YES	YES	🟢	🟢	🟡	🟡	🟢	Moderate

XSL•FO
RenderX

Figure 3. Quality assessment results of the Risk of Bias 2 tool.

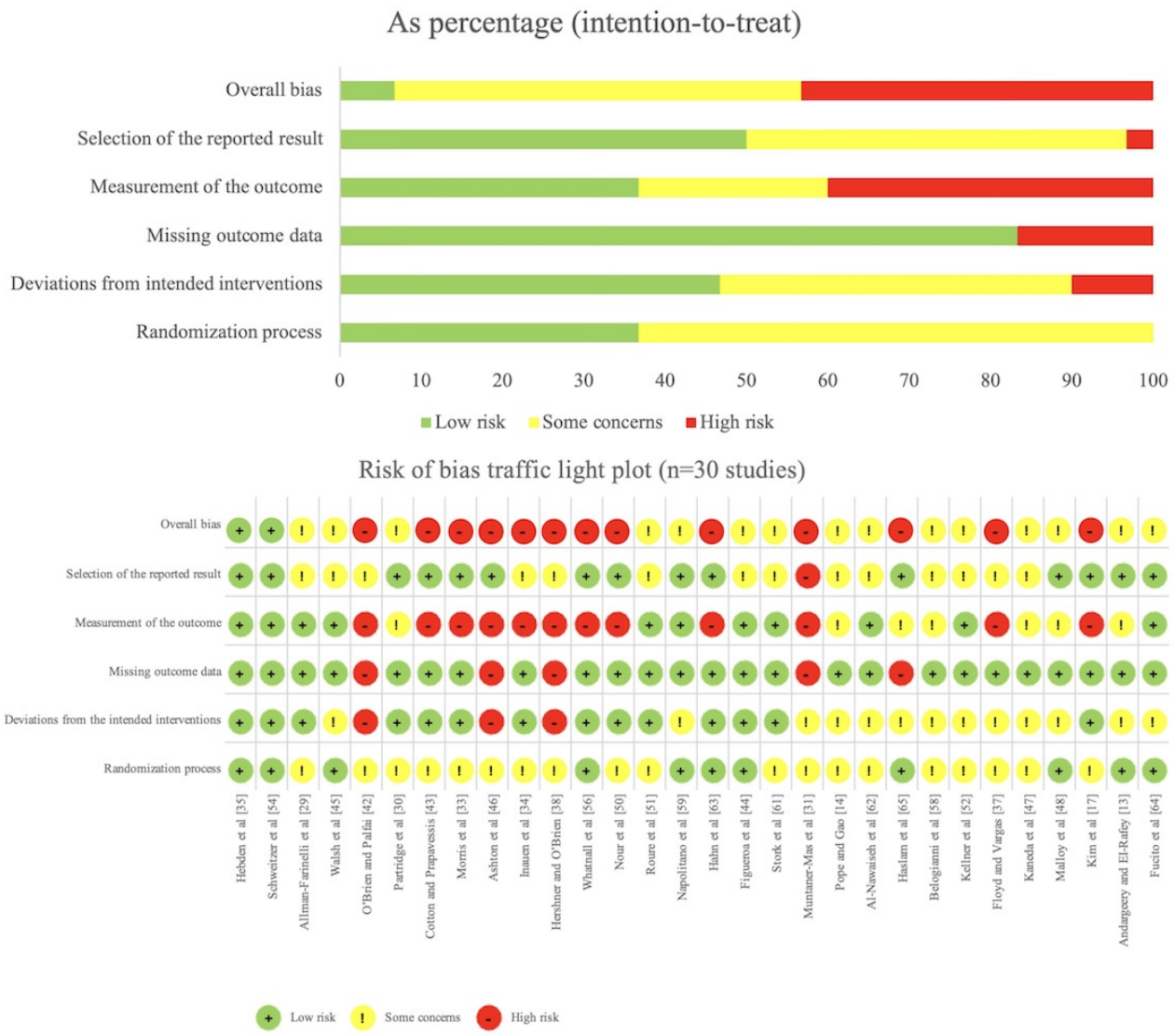

XSL•FO
RenderX

Figure 4. Quality assessment results of Joanna Briggs Institute critical appraisal tools. N/A: not applicable.

Study	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Overall appraisal
Checklist for quasi-experimental studies 2023 (9 questions)											
Hutchesson et al [53]	Yes	No	N/A	N/A	Yes	Yes	Yes	No	Yes	N/A	Include
Xian et al [60]	Yes	No	N/A	N/A	Yes	Yes	Yes	Unclear	Yes	N/A	Include
Chung et al [49]	Yes	No	N/A	N/A	Yes	Yes	Yes	Unclear	Yes	N/A	Include
Fitzsimmons-Craft et al [32]	Yes	No	N/A	N/A	Unclear	Yes	Yes	No	Yes	N/A	Include
Lee and Park [20]	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	N/A	Include
Napolitano et al [36]	Yes	No	N/A	N/A	Yes	Yes	Unclear	Yes	Yes	N/A	Include
Cantisano et al [41]	Yes	N/A	Unclear	Yes	No	Yes	N/A	No	Yes	N/A	Include
Khatri and Sharma [40]	Yes	No	Yes	Yes	Yes	N/A	Yes	Yes	Yes	N/A	Include
Olatona et al [15]	Yes	N/A	Unclear	Yes	Yes	Yes	N/A	No	Yes	N/A	Include
Gao et al [10]	N/A	N/A	Yes	Unclear	No	Yes	Yes	Yes	Yes	N/A	Include
Checklist for analytical cross sectional studies 2020 (8 questions)											
Nour et al [66]	Yes	Yes	Yes	Yes	Unclear	Unclear	Yes	Yes	N/A	N/A	Include
Sarcona et al [55]	Yes	Yes	No	Unclear	Yes	Yes	Unclear	Yes	N/A	N/A	Include
Smith and Volkwyn [39]	Yes	Yes	Unclear	Yes	No	No	Unclear	Yes	N/A	N/A	Include
Rajan and Muthunarayanan [11]	Yes	Yes	Yes	Yes	Yes	No	Unclear	Yes	N/A	N/A	Include
Wittmar et al [57]	Yes	Yes	Yes	Yes	Unclear	Yes	Yes	Yes	N/A	N/A	Include
Checklist for qualitative research 2020 (10 questions)											
Åsberg et al [26]	No	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Include
Wittmar et al [57]	Unclear	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Include

Study design	Checklist	Study	Total questions	"N/A" count	"Yes" count	Compliance rate (%)	Quality grade	Overall appraisal
Quantitative nonrandomized study (n=10)	Checklist for quasi-experimental studies 2023	Hutchesson et al [53]	9	2	5	71.4%	Moderate	Include
		Xian et al [60]	9	2	5	71.4%	Moderate	Include
		Chung et al [49]	9	2	5	71.4%	Moderate	Include
		Fitzsimmons-Craft et al [32]	9	2	4	57.1%	Low	Include
		Lee and Park [20]	9	0	7	77.8%	High	Include
		Napolitano et al [36]	9	2	5	71.4%	Moderate	Include
		Cantisano et al [41]	9	2	4	57.1%	Low	Include
		Khatri and Sharma [40]	9	1	7	87.5%	High	Include
		Olatona et al [15]	9	2	5	71.4%	Moderate	Include
		Gao et al [10]	9	2	5	71.4%	Moderate	Include
Quantitative descriptive study (n=4)	Checklist for analytical cross sectional studies 2020	Nour et al [66]	8	0	6	75.0%	High	Include
		Sarcona et al [55]	8	0	5	62.5%	Moderate	Include
		Smith and Volkwyn [39]	8	0	4	50.0%	Moderate	Include
		Rajan and Muthunarayanan [11]	8	0	7	87.5%	High	Include
Qualitative research (n=1)	Checklist for qualitative research 2020	Åsberg et al [26]	10	0	7	70.0%	Moderate	Include
mixed methods study (n=1)	Checklist for analytical cross sectional studies 2020 + Checklist for qualitative research 2020	Wittmar et al [57]	18	0	14	77.8%	High	Include

Data Extraction Results

This review included a total of 46 studies. The basic characteristics of the included studies are summarized in Table 1, with detailed data extraction results provided in Multimedia Appendix 3. Given the substantial heterogeneity among the included studies with respect to study design, target behaviors, intervention formats, core functions, and primary outcome measures, as well as variations in methodological quality, a meta-analysis was not conducted. Instead, a comprehensive analysis was performed using descriptive synthesis and comparative approaches. By systematically organizing and describing key characteristics of DHIs—including intervention targets, participant characteristics, sample sizes, formats, functions, durations, outcomes, and effects—this review delineates the overall patterns and heterogeneity within the field. Specific details are elaborated in the subsequent sections and illustrated through relevant tables, charts, and figures.

XSL•FO
RenderX

Table 1. Summary of data extraction from included studies.

Study design and relevant studies	Total completed, N	Participant age (years), mean (SD)	Intervention(s)	Target behavior(s)	Function	Effectiveness
Quantitative randomized controlled trial						
Hebden et al [35]	46	22.8 (4.6)	SMS text messages, emails, smartphone apps, and internet forums	• Physical activity • Sedentary behavior • Diet	• Prompting • Education • Guidance	Limited[a]
Schweitzer et al [54]	106	19.7 (0.73)	Email	• Physical activity • Diet	• Guidance • Education • Prompting	Yes[b]
Allman-Farinelli et al [29]	202	27.7 (4.9)	Coaching calls, SMS text messages, emails, apps, and downloadable website resources	• Physical activity • Diet	• Guidance • Prompting • Education	Yes
Walsh et al [45]	55	20.55 (2.07)	Smartphone app	• Physical activity	• Monitoring • Feedback	Yes
O'Brien and Palfai [42]	148	19.24 (1.16)	Web and SMS text messages	• Diet	• Education • Prompting • Guidance	Limited
Partridge et al [30]	248	27.0 (4.0)	Coaching calls, SMS text messages, emails, smartphone apps, and website	• Physical activity • Diet	• Education • Guidance	Yes
Cotten and Prapavessis [43]	56	21.19 (4.19)	SMS text messages	• Sedentary behavior	• Prompting • Guidance	Limited
Morris et al [33]	112	20.5 (1.95)	Internet	• Sleep	• Education • Guidance	Yes
Ashton et al [46]	47	22.1 (2.0)	Website, wearable device, and Facebook support group	• Physical activity • Diet	• Guidance • Education • Interaction	Limited
Inauen et al [34]	141	27.5 (8.6)	App	• Diet	• Interaction • Monitoring	Limited
Hershner and O'Brien [38]	358	21.9 (4.1)	Website	• Sleep	• Education	Yes
Whatnall et al [56]	90	22.4 (4.0)	Website	• Diet	• Education • Guidance	Limited
Nour et al [50]	47	24.8 (3.4)	Self-monitoring app, gamified app, and social media (Facebook)	• Diet	• Monitoring • Interaction	Limited
Roure et al [51]	60	20.8 (1.3)	Exergame	• Physical activity	• Immersion • Engagement	Yes
Napolitano et al [59]	283	23.3 (4.4)	Facebook and SMS text messages	• Physical activity • Diet	• Prompting • Feedback	Yes
Hahn et al [63]	192	20.2 (2.4)	App	• Physical activity • Diet	• Monitoring	No[c]
Figueroa et al [44]	93	20.2 (2.47)	App and SMS text messages	• Physical activity	• Prompting • Feedback • Monitoring	Yes

XSL•FO
RenderX

Study design and relevant studies	Total completed, N	Participant age (years), mean (SD)	Intervention(s)	Target behavior(s)	Function	Effectiveness
Stork et al [61]	46	24.0 (5.0)	App	• Physical activity	• Guidance • Monitoring	Yes
Muntaner-Mas et al [31]	66	23.1 (4.0)	App	• Physical activity	• Guidance	Yes
Pope and Gao [14]	42	21.6 (NR)[d]	App and Facebook	• Physical activity • Sedentary behavior	• Monitoring • Education • Prompting	Yes
Al-Nawaiseh et al [62]	114	21.1 (2.2)	App	• Physical activity	• Monitoring • Feedback	Yes
Haslam et al [65]	141	21.7 (2.0)	Website	• Diet	• Feedback • Education	No
Belogianni et al [58]	65	23.01 (3.82)	Website	• Physical activity • Sedentary behavior • Diet	• Education • Immersion	No
Kellner et al [52]	34	22.31 (2.59)	SMS text messages	• Sedentary behavior	• Prompting	Yes
Floyd and Vargas [37]	55	19.9 (0.97)	App	• Sleep	• Guidance • Education	Yes
Kaneda et al [47]	46	20.8 (1.2)	E-learning and exercise video	• Physical activity • Sedentary behavior	• Education • Guidance	No
Malloy et al [48]	46	21.34 (2.02)	Social media	• Physical activity • Diet	• Education • Prompting • Guidance	Limited
Kim et al [17]	60	21.9 (1.43)	Virtual reality	• Sleep	• Immersion • Guidance	Yes
Andargeery and El-Rafey [13]	220	19.97 (2.61)	Mobile health tools and videos	• Physical activity • Diet • Sleep	• Education • Guidance • Monitoring	Yes
Fucito et al [64]	98	21.16 (1.75)	Wearable devices, website, and smartphone	• Sleep	• Monitoring • Guidance • Feedback	Yes
Quantitative nonrandomized study						

XSL•FO
RenderX

Study design and relevant studies	Total completed, N	Participant age (years), mean (SD)	Intervention(s)	Target behavior(s)	Function	Effectiveness
Hutchesson et al [53]	12	22.8 (3.2)	Website, emails, online forum, smartphone app, and SMS text messages	• Physical activity • Sedentary behavior • Diet	• Feedback • Education • Interaction	Yes
Xian et al [60]	167	25.0 (4.0)	Reality game	• Physical activity	• Immersion • Prompting	Yes
Chung et al [49]	12	19.8 (1.0)	Fitbit, Twitter, and gamification	• Physical activity • Diet	• Monitoring • Interaction • Prompting	Yes
Fitzsimmons-Craft et al [32]	2454	22.89 (6.59)	Online platforms	• Diet	• Screening • Guidance	Yes
Lee and Park [20]	59	22.0 (2.0)	Apps and wearable devices	• Physical activity • Diet	• Monitoring • Guidance	Yes
Napolitano et al [36]	20	18.3 (0.72)	Digital learning modules	• Physical activity • Sedentary behavior • Diet	• Monitoring • Feedback	Limited
Cantisano et al [41]	16	20.69 (1.74)	eHealth tools	• Physical activity • Diet	• Education	Limited
Khatri and Sharma [40]	500	20.74 (1.77)	App	• Sedentary behavior	• Monitoring • Feedback • Guidance	Yes
Olatona et al [15]	1182	Unclear	Social media	• Diet	• Education • Guidance	Yes
Gao et al [10]	456	21.5 (1.4)	Artificial intelligence–powered gamification	• Physical activity	• Interaction	Yes
Quantitative descriptive study						
Nour et al [66]	401	27.7 (4.9)	Telephone, website, smartphone app, and SMS text messages	• Diet	• Guidance • Education	Yes
Sarcona et al [55]	230	22.0 (3.0)	Mobile health apps	• Physical activity • Diet	• Monitoring	Yes
Smith and Volkwyn [39]	192	22.7 (3.7)	App	• Physical activity	• Monitoring • Feedback	Yes
Rajan and Muthunarayanan [11]	680	23.82 (1.62)	App	• Physical activity • Diet	• Monitoring • Education • Screening	Yes
Qualitative research						
Åsberg et al [26]	50	31.3 (6.4)	SMS text messages	• Physical activity • Sedentary behavior • Diet	• Guidance • Education • Feedback	Limited
Mixed methods study						

XSL•FO
RenderX

Study design and relevant studies	Total completed, N	Participant age (years), mean (SD)	Intervention(s)	Target behavior(s)	Function	Effectiveness
Wittmar et al [57]	142	24.0 (4.0)	Web application	• Physical activity	• Education • Interaction	Yes

[a]Limited: limited evidence of effectiveness, based on reported effect measures, CIs, and authors' conclusions (see Multimedia Appendix 3).

[b]Yes: evidence of effectiveness, based on reported effect measures, CIs, and authors' conclusions (see Multimedia Appendix 3).

[c]No: no evidence of effectiveness, based on reported effect measures and authors' conclusions (see Multimedia Appendix 3).

[d]NR: not reported.

In terms of annual distribution (see Figure 5), the number of studies during the early period (2014-2015) was low, with only 1 publication per year. Since 2016, the number of publications increased markedly, reaching a first minor peak in 2016 (n=8), possibly associated with the rapid adoption of smartphones and mobile apps among college students. From 2017 to 2020, the number of studies fluctuated between 1 and 5 annually, maintaining an overall moderate level. The number increased again and stabilized in 2021-2022, declined slightly in 2023, reached a second peak in 2024 (n=8), and remained high in 2025 (n=4). Publications from the last 5 years accounted for more than half of all studies identified.

Figure 5. Annual and cumulative publication counts of the included studies.

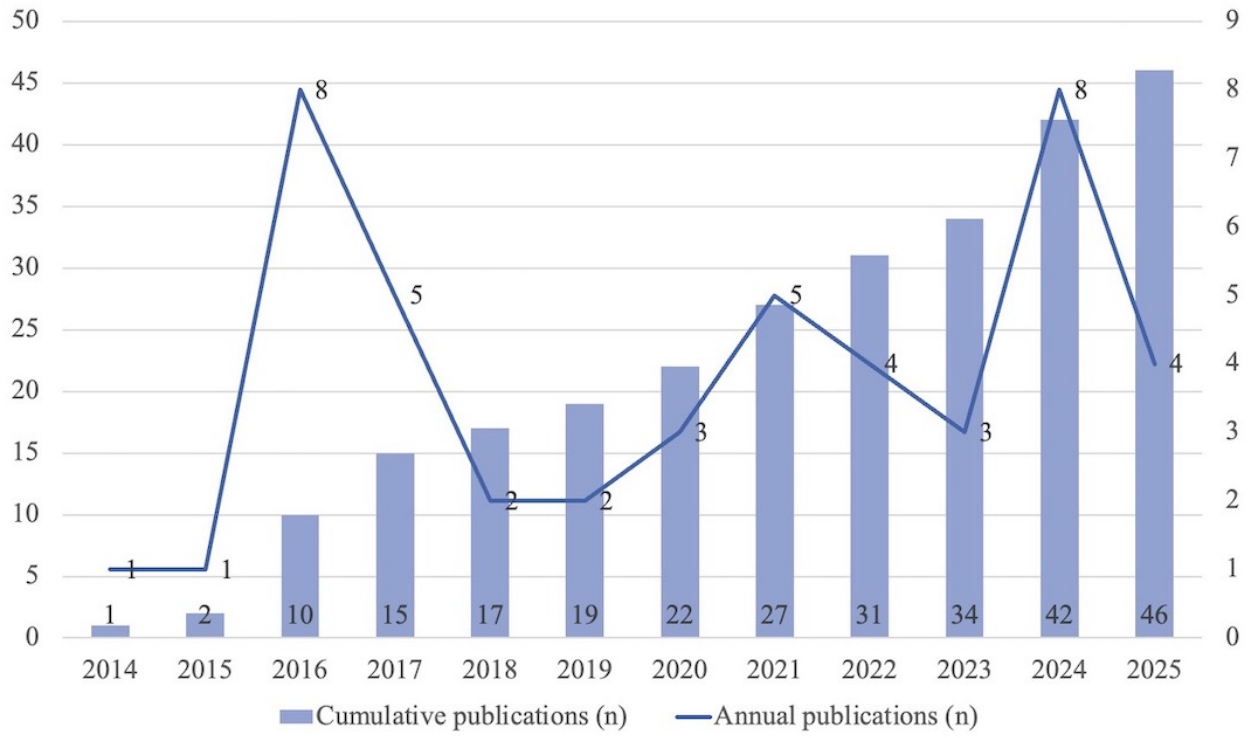

The regional and country distribution of the included studies demonstrates a clear geographical concentration. At the regional level, most studies were conducted in North America (n=18, 39%), followed by Oceania (n=10, 22%) and Europe (n=9, 20%). Asia accounted for 6 (13%) studies, while Africa contributed the smallest share with 3 (7%) studies. At the country level, the United States recorded the highest number of publications (n=15, 33%), followed by Australia (n=9, 20%). The United Kingdom, Germany, Canada, South Korea, and India each contributed 2 studies. The remaining countries were represented by a single study, indicating a relatively dispersed distribution beyond the leading contributors.

The distribution of study design types among the included studies exhibited a clear structural pattern. The largest proportion comprised QRCTs (n=30, 65%). This was followed by QNRSs (n=10) and QDSs (n=4), which were primarily used for exploratory analyses and descriptive accounts of phenomena. By contrast, QR and MMS were represented by only 1 article each, accounting for less than 2% of the total. Overall, DHI studies addressing college students' lifestyle behaviors are predominantly quantitative, with a marked preference for QRCTs.

With respect to ethical compliance, all included studies adhered to relevant ethical guidelines, with all 46 (100%) explicitly reporting informed consent procedures and ethics committee approval or review status. Regarding privacy protection and data security, 24 (52%) studies explicitly reported the implementation of protective measures, including secure server storage compliant with data safety standards, encrypted data transmission, data deidentification, and strict access control mechanisms. With respect to adverse events and intervention-related risks, no serious adverse events were

XSL•FO
RenderX

reported across the included studies. Only a small number of studies reported minor negative issues related to technology use, such as fluctuations in intervention engagement, higher dropout rates, or reduced compliance attributable to participants' competing academic or personal commitments. No health risks were identified that were directly attributable to the DHIs.

Intervention Design and Implementation Results

Intervention Objectives

Among the intervention objectives examined in the included studies, 30 addressed physical activity, 26 addressed diet, 10 targeted sedentary behavior, and 6 targeted sleep. Single-behavior interventions accounted for a large proportion of the studies; however, multibehavior crossover interventions were also substantial, with combined physical activity and diet interventions being the most common (n=18). Notably, physical activity was both the most frequent single-behavior intervention target and the primary entry point for multibehavior combined interventions, whereas sleep was relatively underemphasized in intervention design.

Intervention Participants

Based on the PROGRESS-Plus (Place of Residence, Race/Ethnicity, Occupation, Gender/Sex, Religion, Education, Socioeconomic Status, Plus Other Relevant Factors) framework, a synthesis of sociodemographic characteristics from 46 DHI studies identified 10 primary participant categories (see Multimedia Appendix 3), including health status (n=46), age (n=45), gender/sex (n=45), education (n=41), occupation (n=39), place of residence (n=36), race/ethnicity (n=28), socioeconomic status (n=14), social capital (n=8), and religion (n=1). The analysis revealed the following: (1) All participants were college students, predominantly aged 18-30 years, which is consistent with typical college student demographics and showed no substantial deviation across studies. (2) Most interventions targeted students with generally healthy status, whereas 14 out of 46 (30%) focused on subpopulations with specific health risks or special needs, such as overweight or obesity, sleep disorders, psychological stress, or disordered eating behaviors. (3) Gender/sex distribution was relatively balanced across studies, whereas education and occupation exhibited limited variability owing to the homogeneity of the study population. (4) By contrast, PROGRESS-Plus dimensions such as race/ethnicity, socioeconomic status, social capital, and religion received notably limited attention, with a lack of systematic analysis from a health equity perspective.

Intervention Sample

The sample sizes of the included studies varied considerably. Histograms indicated that most studies had sample sizes concentrated below 200 participants, with a median of approximately 95, whereas a few studies had small (<50) or extremely large (>400) samples. As shown in Figure 6, box-and-whisker plots further revealed an uneven distribution with long-tailed characteristics. Variations in sample size were closely associated with study design. Rigorous QRCTs typically require larger samples to ensure statistical power and therefore tend to employ medium- to large-scale sample sizes. By contrast, QDSs and QR are more inclined toward small-sample explorations, sometimes recruiting only a few dozen participants, and are more susceptible to selection bias.

Figure 6. Sample size distribution of the included studies.

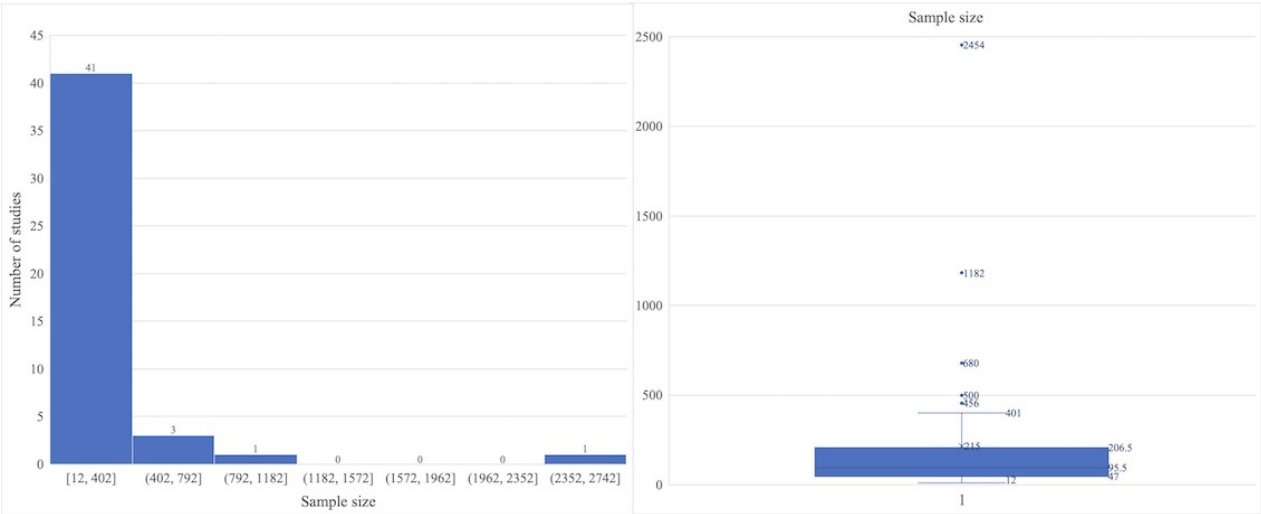

Intervention Modalities

The intervention formats in the included studies fell into 3 main categories. The first category, *single*, referred to interventions employing only 1 digital health technology (n=29), such as mobile apps. The second category, *multiple*, involved combining multiple digital health technologies within the same intervention (n=10). For example, the TXT2BFiT program integrated phone calls, websites, apps, and SMS text messaging simultaneously to achieve intervention goals. The third category, *combined* (n=7), compared the effectiveness of different combinations of digital health technologies, such as a "web-based nutrition intervention only" versus a "web-based intervention combined with daily SMS text message reminders." Regarding the types of intervention technologies, these could be categorized into 7 groups: (1) mobile apps, used 21 times; (2) web-based platforms, including websites (13 times), online forums (3 times), and digital learning or eHealth tools (4 times); (3) mobile

XSL•FO
RenderX

communications, including SMS text messages (11 times), emails (5 times), and phone calls (3 times); (4) social media (7 times); (5) wearable devices (4 times); (6) gamification and multimedia, including gamification and exergames (5 times), videos (2 times), and virtual reality (1 time); and (7) intelligent technologies, represented only by artificial intelligence (1 time). Overall, mobile apps and web-based platforms were the most frequently used technologies.

Intervention Functionalities

The technological functions of the DHIs included in this review exhibited distinct patterns of emphasis. Educational and guidance-related functions predominated across most interventions, followed by monitoring and prompting functions; by contrast, feedback and interactive functions were used less frequently, while immersive, screening, and engagement-related functions were rarely incorporated. Coding these interventions using the Behavior Change Technique Taxonomy version 1 (BCTTv1) indicated that the most frequently employed techniques were "4.1 Instruction on how to perform the behavior" and "5.1 Information about health consequences," suggesting that current DHIs primarily emphasize foundational behavioral support functions. Further frequency analysis of BCT coding among effective intervention studies (see Table 2) showed that BCTTv1 codes 4.1 (16/87, 18%), 5.1 (14/87, 16%), and 2.3 (13/87, 15%) constituted the core set of techniques, collectively accounting for nearly half of all techniques used in effective interventions.

Table 2. Frequency distribution of codes in effective intervention studies (N=87).

Behavior Change Technique Taxonomy version 1 code	Description	Frequency, n (%)
4.1	Instruction on how to perform behavior	16 (18)
5.1	Information about health consequences	14 (16)
2.3	Self-monitoring of behavior	13 (15)
2.2	Feedback on behavior	8 (9)
7.1	Prompts/cues	8 (9)
6.1	Demonstration of behavior	4 (5)
2.1	Monitoring by others (no feedback)	3 (3)
3.1	Social support (unspecified)	3 (3)
5.3	Social/environmental consequences	3 (3)
6.2	Social comparison	3 (3)
12.1	Restructuring physical environment	3 (3)
1.2	Problem solving	2 (2)
1.1	Goal setting (behavior)	1 (1)
1.6	Discrepancy between current behavior and goal	1 (1)
2.4	Self-monitoring of outcomes	1 (1)
2.6	Biofeedback	1 (1)
2.7	Feedback on outcomes	1 (1)
5.6	Emotional consequences	1 (1)
9.1	Credible source	1 (1)

Intervention Duration

The duration of interventions varied considerably across the included studies (see Figure 7; see also [10,11,13-15,17,20,26,29-66]), with the majority concentrated in the short- to medium-term range (1-16 weeks). Studies involving long-term interventions (>16 weeks) were relatively scarce, with only 4 studies identified. Among these studies, most incorporated follow-up periods, and medium- to long-term interventions were typically associated with more systematic follow-up protocols. With respect to study design, randomized controlled trials predominantly employed interventions of medium duration (8-16 weeks). Among the QDSs (n=4) and MMS (n=1) analyzed, some studies employed longer intervention durations to observe behavioral maintenance; however, these accounted for a relatively small proportion of the evidence base. Subgroup analysis demonstrated a progressive increase in the proportion of studies classified as "effective" with increasing intervention duration (see Table 3): 2 out of 4 (50.0%) for ultra-short-term (<1 week), 10 out of 16 (63%) for short-term (>1 and <8 weeks), 12 out of 18 (67%) for medium-term (8-16 weeks), and 3 out of 4 (75%) for long-term (>16 weeks). Notably, medium-duration interventions (8-16 weeks) not only represented the largest proportion of the existing evidence but also demonstrated both a relatively high "effective" rate (12/18, 67%) and a low "ineffective" rate (1/18, 6%). These findings indicate that current DHI research remains skewed toward short- and medium-term interventions, with the

XSL•FO
RenderX

8-16-week category standing out in terms of evidence volume and the apparent stability of intervention effects.

Figure 7. Chart of intervention duration and follow-up duration.

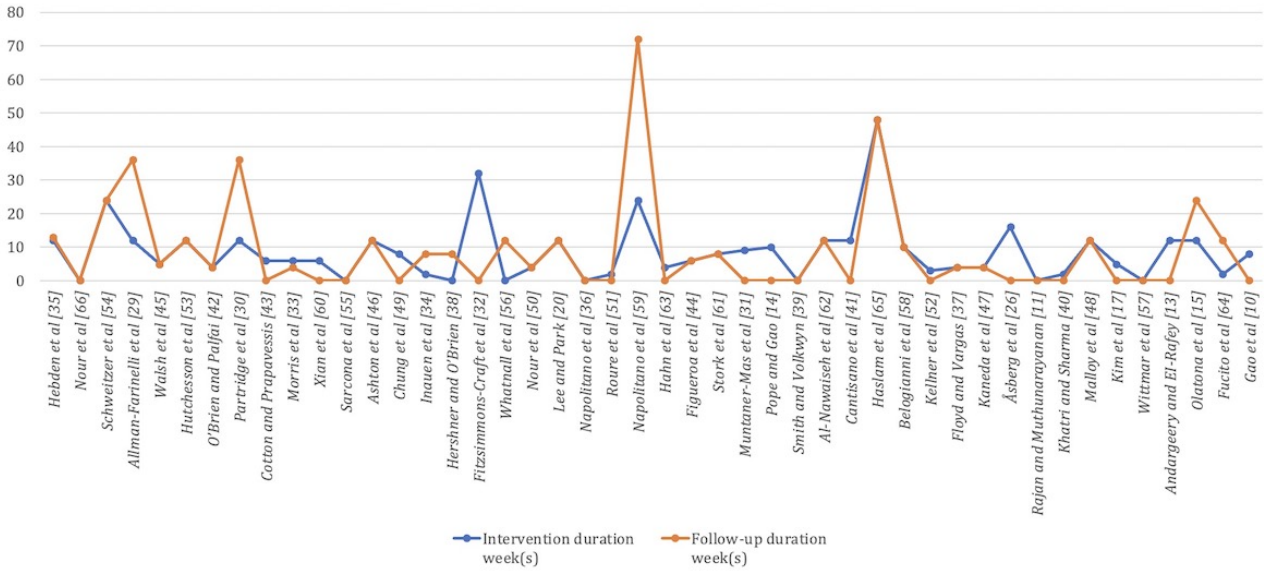

Table 3. Subgroup analysis of intervention duration.

Duration group (weeks)	Number, n	Effective (yes), n (%)	Limited effect, n (%)	Not effective (no), n (%)
Ultrashort (≤1)	4	2 (50.0)	2 (50.0)	0 (0)
Short (>1 and <8)	16	10 (63)	4 (25)	2 (13)
Medium (8-16)	18	12 (67)	5 (28)	1 (6)
Long (>16)	4	3 (75)	0 (0)	1 (25)
Subtotal (analyzed)	42	27 (64)	11 (26)	4 (10)
Excluded: not reported	4	N/A[a]	N/A	N/A

[a]N/A: not applicable.

Intervention Outcomes

As a result of substantial heterogeneity among the included studies with respect to outcome measurement instruments, outcome definitions, and assessment time points, it was not feasible to define a unified primary outcome or to conduct a statistically valid meta-analysis. Accordingly, this review adopted a descriptive synthesis framework to summarize and integrate the relevant outcomes. The outcome metrics in the included studies were classified into 2 main categories. The primary outcomes focused on lifestyle behaviors, including physical activity (eg, activity level, step count, and activity intensity), sedentary behaviors (eg, total sedentary time and frequency of breaks from sitting or resting), diet (eg, dietary quality; intake of fruits, vegetables, and sugar-sweetened beverages; energy intake; and nutritional knowledge), and sleep (eg, sleep quality, duration, efficiency, and severity of insomnia). These indicators directly reflect changes in core health behaviors resulting from the intervention and serve as a key basis for evaluating its effectiveness. Secondary outcomes, serving as supplementary indicators, were more diverse and encompassed physical health status and psychosocial dimensions, such as weight and body composition (eg, weight, BMI, waist circumference, and body fat percentage), physical fitness indicators (eg, flexibility, muscle strength, and cardiorespiratory fitness), cardiometabolic indicators (eg, blood pressure, blood glucose, and blood lipid profiles), and psychological and self-perception measures (eg, self-efficacy, body image, and life satisfaction). Overall, current studies remain primarily focused on primary outcomes, while secondary outcomes have expanded but continue to exhibit limited coverage.

Intervention Effectiveness

Based on the reported effect measure types, effect estimates, confidence levels (%), and CIs across the included studies, together with a comprehensive assessment of the authors' conclusions (see Multimedia Appendix 3), the results indicated that 31 (67%) studies demonstrated evidence of intervention effectiveness, suggesting that DHIs are generally associated with positive outcomes in improving lifestyle behaviors among college students. Four studies reported no statistically significant effects, with limitations primarily attributed to small sample sizes or short intervention durations. The remaining 11 studies demonstrated limited effectiveness, with improvements observed only in selected secondary outcomes or during short-term follow-up periods.

XSL•FO
RenderX

Based on a comprehensive assessment of each behavioral domain using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework, the certainty of evidence for the physical activity and diet domains was rated as "moderate," whereas the evidence for the sedentary behavior and sleep domains was rated as "low." With respect to evidence credibility, this review indicates a moderate level of confidence in the overall estimate that DHIs are effective in improving lifestyle behaviors among college students. The certainty of evidence in some domains was downgraded due to methodological limitations in the existing primary studies, including small sample sizes, challenges in implementing blinding, and inconsistencies in outcome assessment tools. Nevertheless, these GRADE assessments provide an accurate reflection of the current state of the evidence and its overall strength for DHIs among college students, thereby offering valuable guidance for interpretation and future research.

Discussion

Principal Findings

Discussion on Current Research Status

In terms of temporal trends, research on DHIs targeting college students' lifestyle behaviors has gradually emerged since 2014, expanded rapidly after 2016, and reached a peak in the past 5 years [6]. This trend has been driven primarily by 4 categories of factors. First, technological advances have laid a solid foundation for DHIs, with the proliferation of smartphones, wearable devices, and app ecosystems significantly enhancing their accessibility and operability [67]. Second, conceptual advancements have accelerated theoretical and methodological innovations in DHIs, underscoring their distinctive advantages in facilitating behavioral improvement [68]. Third, demand has increased substantially, particularly during the COVID-19 pandemic, with DHIs gaining broad recognition as viable alternatives when traditional approaches were constrained [14]. Fourth, resource investment has continued to expand, with funding, supportive policy frameworks, and interdisciplinary collaboration creating a favorable environment for research. Overall, future research is expected to shift from assessing short-term feasibility to evaluating long-term effectiveness, scalability, and the capacity to accommodate personalization [29,30].

In terms of spatial distribution, research on DHIs is predominantly concentrated in high-income countries, particularly in the United States and Australia. This concentration is primarily driven by a combination of technological infrastructure, research resources, and supportive policy environments. On the one hand, North America and Oceania initiated mHealth development relatively early, benefiting from substantial technological and financial advantages [31]. On the other hand, colleges in these regions generally possess mature health promotion systems and well-established ethical review mechanisms, facilitating the implementation of intervention trials. In addition, higher levels of health awareness and greater digital acceptance in Western cultures further contribute to this pattern. However, the generalizability of these findings may be limited when extrapolated to low- and middle-income countries. For example, resource-constrained settings may encounter infrastructural and hardware-related barriers, such as uneven network coverage and low rates of digital device ownership. Furthermore, substantial cross-cultural variations exist in perceptions of privacy, the role of family involvement, and prevailing health communication practices. In the future, cross-cultural validation and localized adaptation of DHIs should be strengthened [69,70], particularly in resource-constrained settings. Moreover, the development of low-cost, low-threshold DHI models should be explored to advance global health equity [11,71].

In terms of population structure, current research on DHIs has predominantly focused on generally healthy college students, a focus attributable to this group's modifiable health behaviors and susceptibility to environmental influences. However, some studies have extended to special populations, including college students with overweight or obesity, individuals at risk for eating disorders [15,32], and students experiencing sleep disorders or psychological stress [33,34]. This differentiation strategy is partly motivated by the fact that special populations face higher health risks, thereby increasing the potential benefits and clinical value of interventions [72,73]. It also aligns with the need for precision interventions and stratified management. However, existing research has not yet sufficiently examined variations in engagement levels among students from diverse sociodemographic backgrounds. Limited attention to factors such as socioeconomic status and access to digital devices may result in disproportionate benefits for students with greater financial or digital resources, while those experiencing economic constraints or limited device access may be marginalized in the intervention process. Accordingly, future research is likely to advance along 2 complementary directions: first, continuing large-scale studies targeting general undergraduate populations to assess the generalizability of interventions; and second, strengthening targeted interventions for high-risk groups while prioritizing the reduction of participation barriers among students from diverse backgrounds [74]. Such efforts may drive the development of DHIs toward greater refinement, equity, and personalization.

Discussion on Intervention Implementation

In terms of intervention objectives, physical activity and diet are the 2 lifestyle behavior categories receiving the most research attention [35], whereas sedentary behavior and sleep are relatively underrepresented. Both single-behavior and multibehavior combined interventions coexist. This pattern is primarily influenced by several factors. First, physical activity and diet are directly associated with weight management, energy balance, and metabolic health—core variables that affect college students' physical fitness and chronic disease risk [36]. Related measures (eg, step count, energy intake) are more easily quantifiable and standardized, making them more likely targets for intervention. By contrast, sedentary behavior and sleep, despite their recognized importance [37], pose technical and operational challenges for DHIs, including measurement complexity and delayed feedback on intervention effects [38], contributing to a relative paucity of research [39,40]. Current DHIs demonstrate limited effectiveness in reducing sedentary time among college students [75], whereas sleep interventions,

XSL•FO
RenderX

although promising, remain understudied and predominantly focus on insomnia relief [18,76]. Second, intervention strategies reflect researchers' assessment of behavioral variability: physical activity and diet exhibit a wide window for controllability and modification, whereas sedentary behavior typically occurs in academic or leisure contexts, complicating immediate adjustment via a single technique. The prevalence of single-behavior interventions is attributed to their suitability for early exploratory phases, allowing easier control of variables and validation of intervention effects. Conversely, the increase in multibehavior interventions reflects the aggregation of lifestyle risks among college students, which complicates achieving sufficient health benefits through changes in a single behavior. Notably, the combination of physical activity and diet is the most frequent, reflecting the necessity for integrated interventions targeting weight management and energy metabolism [41]. Overall, future research is expected to increasingly adopt multibehavior approaches, integrating behavioral science theories and technological tools to develop synergistic interventions that address the complexity of lifestyle risks.

In terms of intervention modalities, an evolutionary trend is evident, progressing from single to multiple formats and from low to high levels of interaction, driven by the combined forces of technological advancement, user demand, and intervention science. In early studies, SMS text messages and emails were the predominant forms of DHIs [42], owing to their low technological threshold, ease of deployment, and minimal cost, which made them suitable for rapid implementation in resource-limited contexts [43,44]. However, these approaches primarily involved 1-way information delivery, lacked personalization and real-time interaction, and were insufficient in maintaining user engagement. With the widespread adoption of smartphones and the maturation of the app ecosystem, mobile apps have gradually become the mainstream form of DHIs. These apps are highly integrated and interactive, capable of incorporating multiple functions such as goal-setting, feedback, reminders, and data tracking [13], aligning with college students' high-frequency mobile usage habits and significantly enhancing the intervention experience and engagement [45]. Web platforms retain advantages in scalability but are somewhat less user-friendly and less effective in delivering push notifications compared with apps [46,47]. In addition, the integration of social media and wearable devices increases the interactivity and contextual adaptability of DHIs [48], further enhancing behavioral monitoring and the provision of immediate feedback [49-51]. Future trends are expected to emphasize technological convergence and intelligent development. On the one hand, combinations of multiple formats (eg, apps, social platforms, and gamification) will become increasingly prevalent to address the multidimensional needs of behavioral interventions [52]. On the other hand, personalized interventions leveraging artificial intelligence, virtual coaching, and immersive experiences (eg, augmented reality/virtual reality) are anticipated to emerge as key research directions [17,53], shifting DHIs from being information-driven to experience-driven and ultimately facilitating sustained behavior change.

In terms of intervention functions, current DHIs are predominantly characterized by education, guidance, monitoring, and prompting components [54], indicating that these interventions primarily emphasize information delivery and basic behavior management. This design approach is partly driven by the substantial demand for health knowledge and skills among college students, with education and guidance functions facilitating improvements in cognition and self-efficacy. Concurrently, monitoring and prompting functions leverage technology to enable data recording and behavioral reinforcement, thereby promoting the initiation of target behaviors in the short term [55]. However, high-engagement features such as feedback, interaction, and gamification-based incentives remain underutilized [56,57], suggesting that DHIs often lack deep personalization and social support components [58], which may be a critical factor limiting long-term user engagement and intervention effectiveness. In terms of future trends, the convergence of behavior change theories (eg, Capability-Opportunity-Motivation-Behavior [COM-B], behavioral economics) with intelligent algorithmic applications is expected to drive the evolution of DHI functionality toward greater personalization, interactivity, and emotional engagement [59]. For example, artificial intelligence–driven real-time feedback could enhance intervention adaptability, virtual communities could strengthen social support, and gamification mechanisms coupled with reward systems could foster intrinsic motivation. Such advancements are likely to not only increase intervention engagement but also substantially improve behavioral maintenance, fostering a gradual shift from information delivery–oriented DHIs to approaches that place greater emphasis on user experience and social interaction.

Discussion on Intervention Effectiveness

Overall, 31 of 46 (67%) studies reported effective outcomes (yes), indicating the high feasibility and considerable potential of DHIs in improving the lifestyle behaviors of college students [60-62,77]. However, a subset of studies yielded insignificant (no) or limited (limited) effects, which can be examined from several analytical dimensions. First, insufficient refinement and lack of theoretical underpinning in intervention design represent key factors constraining effectiveness. In several cases, interventions lacked explicit theoretical frameworks for behavior change, relying predominantly on information delivery. Such approaches often failed to sufficiently stimulate participant motivation or reinforce behavior maintenance, leading to short-term gains that were difficult to sustain [26]. Second, existing intervention studies generally lack robust validation of long-term effects. Most studies are limited to durations of 8-16 weeks and include insufficient follow-up, which constrains the ability to verify the sustainability and stability of behavioral changes [63]. As a result, the long-term value and durability of DHIs remain difficult to assess adequately. Third, intervention effectiveness appears to be strongly influenced by participant adherence. Analyses of engagement-related metrics indicate that higher levels of user engagement, compliance, and intervention consistency are generally associated with more favorable behavioral and clinical outcomes. By contrast, studies characterized by high dropout rates often rely predominantly on 1-way information delivery, with limited opportunities for

XSL•FO
RenderX

feedback and interaction. Fourth, the type of target behavior and associated measurement challenges also contribute to these outcomes. Compared with physical activity, the intervention effects on diet, sedentary behavior, and sleep were more vulnerable to external environmental influences (eg, academic workload, dietary contexts), and measurement tools relied predominantly on self-reporting, thereby increasing bias and uncertainty. Taken together, variations in intervention design, technological application, and behavioral characteristics collectively contribute to the substantial heterogeneity observed in intervention outcomes [64].

To gain a deeper understanding of variations in intervention effectiveness, the COM-B framework can be applied as a systematic analytical tool [78]. (1) Within the "Capability" dimension, most interventions primarily enhanced college students' health-related knowledge through educational content and guidance materials. Examples included the provision of diet guidelines, exercise plans, and sleep regulation strategies designed to increase participants' awareness of the importance of healthy behaviors. However, these improvements often remained at the cognitive level, with limited emphasis on the development of practical behavioral skills. Specific components, such as diet substitution options, situational coping strategies, or flexible exercise planning, were frequently absent. In addition, some studies did not provide adequate support for data interpretation, which limited participants' ability to translate behavioral monitoring data into actionable steps [79]. (2) Within the "Opportunity" dimension, DHIs generally rely on virtual platforms to create enabling behavioral conditions, such as goal tracking, reminder functions, and online resource sharing, which may theoretically reduce psychological barriers to behavior enactment. However, the structuring of opportunities within real-world contexts remains insufficiently optimized. Some interventions do not adequately account for the distinctive time pressures and contextual constraints experienced by college students on campus. For example, strategies aimed at reducing sedentary behavior often remain limited to generic standing reminders, without adaptation to classroom environments or common study spaces, thereby constraining opportunities for sustained behavior change. Furthermore, although some interventions attempt to incorporate social support mechanisms (eg, community interactions or peer challenges), the depth and quality of participant engagement are generally limited. These interactions frequently involve 1-way information transmission, with limited capacity to foster emotional connection or effective behavioral modeling. (3) Within the "Motivation" dimension, existing interventions primarily emphasize the stimulation of extrinsic motivation through short-term incentives, such as point-based rewards and task completion reminders. While such strategies may promote initial engagement, they generally lack mechanisms for the sustained cultivation of intrinsic motivation. Specifically, many interventions have not effectively supported college students in developing a sense of self-worth derived from continued engagement in healthy behaviors. In addition, strategies aimed at enhancing positive emotional experiences are rarely incorporated. For example, gamification designs often remain confined to superficial point-based systems, with limited capacity to stimulate participants' sense of exploration, mastery, or accomplishment. Additionally, insufficient personalization of feedback appears to substantially constrain the maintenance of motivation over time. Participants often receive generic informational messages rather than timely, individualized feedback closely aligned with their actual behavioral performance.

In summary, current DHIs predominantly adopt a "technology-driven" or "utility-oriented" design logic, with a primary emphasis on functional implementation and surface-level engagement metrics. Because of the limited integration of behavior change theory, such interventions tend to exhibit constrained effectiveness in sustaining long-term outcomes. By contrast, theory-driven interventions—such as those grounded in the COM-B framework—extend beyond short-term behavior initiation, emphasizing the synergistic development and dynamic support of capability, opportunity, and motivation. Through structured and phased behavioral support strategies, such interventions may facilitate the establishment of enduring foundations for sustained change across cognitive, skill-based, environmental, and emotional dimensions [10]. As a result, long-term behavior maintenance may become more attainable [65]. Future research should further position behavior change theory as a central guiding principle in intervention design, moving beyond the view of technology as a standalone tool and instead embedding it organically within support systems centered on behavior change mechanisms.

Strengths and Limitations

This study is among the first English-language reviews to systematically integrate multiple forms of DHIs and multiple lifestyle behavior domains within a core population of college students, and it presents the following strengths. First, the study design strictly adheres to PRISMA 2020 and was preregistered on PROSPERO. The systematic search spanned 10 major international databases, ensuring the comprehensiveness and representativeness of the evidence base. Second, by focusing on college students as "digital natives," this study systematically analyzes intervention characteristics across 4 health behavior domains—physical activity, sedentary behavior, diet, and sleep—thereby addressing limitations of prior reviews that emphasized a single behavior or tool. Third, drawing on the COM-B framework, this study examines the mechanisms of DHIs across the Capability, Opportunity, and Motivation dimensions and identifies key bottlenecks in intervention strategies—such as limited technological functionality, suboptimal ecological adaptability, and insufficient motivational activation—thereby providing both theoretical support and practical guidance for the future design and optimization of DHIs for college students.

Although this review endeavored to incorporate the existing literature as comprehensively as possible, several limitations remain. First, the geographical distribution of the included studies was uneven, with a heavy concentration in high-income countries—particularly North America and Australia—which constrains the global applicability of the findings; specifically, their generalizability to college students in low- and middle-income countries requires empirical verification. Second, many studies employed small samples, short intervention durations, and limited follow-up, and some lacked robust control

XSL•FO
RenderX

groups or adequate randomization, thereby weakening the stability of effect estimates and the strength of causal inference. In addition, DHIs were often relatively homogeneous, with limited multidimensional interactivity and personalization; blinding procedures were difficult to implement; and risks of bias arose in adherence assessment and outcome measurement. Therefore, future research should strengthen sample representativeness, enhance intervention refinement, and improve methodological rigor to increase the external validity and practical utility of the findings.

Implications and Recommendations

Recommendations for Policy and Practice

To fully realize the potential of DHIs while ensuring the sustainability and broad accessibility of intervention effects, systematic improvements in policy design and implementation pathways are required. First, college students should be explicitly incorporated into national and regional digital health strategies to facilitate a shift from traditional health education toward integrated digital platforms, and higher-education institutions should be encouraged to develop or adopt scientifically grounded, standardized tools with clearly articulated mechanisms of action. Second, localized development of intervention content and functionality should be supported, with attention to adaptability across behavioral domains, cultural contexts, and student needs, thereby advancing refined, human-centered design with respect to technological thresholds, data security, and personalized recommendations. Third, intervention practice should strengthen students' active engagement and establish feedback-driven, behavior-reinforcing, and peer-support mechanisms to enhance sustained use and intrinsic motivation. In parallel, cross-departmental cooperation mechanisms should be established at the college level, and health interventions should be embedded within curricula, psychological support systems, and campus service resources to form a synergistic support network. Finally, at the policy level, ethical oversight and effectiveness evaluation of DHIs programs should be strengthened, and an evidence-based evaluation framework for DHIs should be established to ensure fairer, more adaptable, and more effective interventions for college students.

Recommendations for Future Research

This study indicates that current research on DHIs for college students remains constrained by unrepresentative samples, single-focus intervention content, and unclear technological mechanisms; future work should be refined and deepened in the following respects. First, geographical and cultural diversity should be expanded, prioritizing studies from low- and middle-income countries, varied higher-education institution types, and diverse social groups to enhance the external validity of the findings. Second, the design and evaluation of multibehavior-integrated interventions should be strengthened by moving beyond single-behavior paradigms and examining behavioral synergies and optimal combinations of intervention components. Third, higher-quality study designs—such as QRCTs, MMSs, and long-term follow-up—should be employed to strengthen causal inference and the sustainability of intervention effects. Fourth, theoretical development and empirical testing of intervention mechanisms should be strengthened by grounding analyses in behavior change theory to clarify how technology enhances Capability, Opportunity, and Motivation, and to advance DHIs from merely providing technical functions to creating a supportive ecosystem conducive to sustained behavior change. Finally, future studies should emphasize the assessment of intervention equity, systematically account for potential moderators such as gender, socioeconomic status, and psychological status, and identify subgroups with limited responsiveness, thereby providing a robust evidence base for constructing a more inclusive and adaptive DHI model for college students.

Conclusions

This review addresses a gap in the literature by focusing specifically on college students, a group often overlooked in research that typically centers on broader adult populations. Unlike prior reviews that mainly examine a single lifestyle behavior, this study adopts a more holistic approach by integrating multiple behaviors and evaluating a range of DHIs with diverse modalities and functionalities. These findings provide valuable insights for refining future DHIs targeting college students and contribute to the development of more effective health promotion strategies in higher education. Although DHIs show potential for improving lifestyle behaviors, their long-term effectiveness remains uncertain. Current interventions face several limitations, including a narrow behavioral focus, basic technological functionality, and limited adaptability to diverse contexts, all of which may restrict long-term engagement and personalized responsiveness. Moreover, many interventions do not fully account for variations in resource access and individual behavior change pathways, potentially limiting their applicability and equity. Future research should prioritize integrating multiple behaviors, enhancing user engagement, improving contextual adaptability, and expanding technological accessibility. Long-term studies and equity-focused evaluations are essential for strengthening the evidence base and ensuring the sustainability and inclusivity of health behavior change among college students.

Acknowledgments

We are grateful to all participants in this study for their time, effort, and dedication. The authors declare that no generative artificial intelligence was used in the preparation of this manuscript.

XSL•FO
RenderX

Funding

This research was funded by the Post-Funded Project of China National Social Science Foundation (grant 25FTYB012), the General Project of Shanghai Philosophy and Social Science Foundation (grant 2025BTY002), and the Youth Project of Shanghai Eastern Talent Plan (grant QNJY2025162).

Data Availability

Data are provided within the manuscript or multimedia appendices; data/materials can be shared upon reasonable request to the corresponding author (ZHY).

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 28 KB - jmir_v28i1e82192_app1.docx]

Multimedia Appendix 2
Search strategy used in this review.
[DOCX File , 19 KB - jmir_v28i1e82192_app2.docx]

Multimedia Appendix 3
Comprehensive data extraction table of included studies.
[XLSX File (Microsoft Excel File), 44 KB - jmir_v28i1e82192_app3.xlsx]

References

1. Peng S, Yuan F, Othman AT, Zhou X, Shen G, Liang J. The effectiveness of e-Health interventions promoting physical activity and reducing sedentary behavior in college students: a systematic review and meta-analysis of randomized controlled trials. Int J Environ Res Public Health 2022 Dec 25;20(1):318-318 [FREE Full text] [doi: 10.3390/ijerph20010318] [Medline: 36612643]
2. Huang T, Zheng K, Li S, Yang Y, Kong L, Zhao Y. Screen-based sedentary behaviors but not total sedentary time are associated with anxiety among college students. Front Public Health 2022 Oct 20;10:994612-994612 [FREE Full text] [doi: 10.3389/fpubh.2022.994612] [Medline: 36339232]
3. Feng J, Lau PWC, Shi L, Huang WY. Longitudinal shifts in movement behaviors during the COVID-19 pandemic: relations to posttraumatic stress disorder among university students. Int J Environ Res Public Health 2022 Oct 18;19(20):13449-13449 [FREE Full text] [doi: 10.3390/ijerph192013449] [Medline: 36294027]
4. Barbosa BCR, Menezes-Júnior LAAD, de Paula W, Chagas CMDS, Machado EL, de Freitas ED, et al. Sedentary behavior is associated with the mental health of university students during the Covid-19 pandemic, and not practicing physical activity accentuates its adverse effects: cross-sectional study. BMC Public Health 2024 Jul 11;24(1):1860 [FREE Full text] [doi: 10.1186/s12889-024-19345-5] [Medline: 38992623]
5. Bi S, Yuan J, Wang Y, Zhang W, Zhang L, Zhang Y, et al. Effectiveness of digital health interventions in promoting physical activity among college students: systematic review and meta-analysis. J Med Internet Res 2024 Nov 20;26:e51714-e51714 [FREE Full text] [doi: 10.2196/51714] [Medline: 39566049]
6. Taylor ME, Liu M, Abelson S, Eisenberg D, Lipson SK, Schueller SM. The reach, effectiveness, adoption, implementation, and maintenance of digital mental health interventions for college students: a systematic review. Curr Psychiatry Rep 2024 Oct 11;26(12):683-693 [FREE Full text] [doi: 10.1007/s11920-024-01545-w] [Medline: 39392547]
7. Gunasekeran DV, Tseng RMWW, Tham Y, Wong TY. Applications of digital health for public health responses to COVID-19: a systematic scoping review of artificial intelligence, telehealth and related technologies. NPJ Digit Med 2021 Feb 26;4(1):40-40 [FREE Full text] [doi: 10.1038/s41746-021-00412-9] [Medline: 33637833]
8. Keesara S, Jonas A, Schulman K. Covid-19 and health care's digital revolution. N Engl J Med 2020 Jun 04;382(23):e82. [doi: 10.1056/NEJMp2005835] [Medline: 32240581]
9. Tan SY, Sumner J, Wang Y, Wenjun Yip A. A systematic review of the impacts of remote patient monitoring (RPM) interventions on safety, adherence, quality-of-life and cost-related outcomes. NPJ Digit Med 2024 Jul 18;7(1):192-192 [FREE Full text] [doi: 10.1038/s41746-024-01182-w] [Medline: 39025937]
10. Gao Y, Zhang J, He Z, Zhou Z. Feasibility and usability of an artificial intelligence-powered gamification intervention for enhancing physical activity among college students: quasi-experimental study. JMIR Serious Games 2025 Mar 24;13:e65498-e65498 [FREE Full text] [doi: 10.2196/65498] [Medline: 40127464]

XSL•FO
RenderX

11. Rajan R, Muthunarayanan L. A comparative study on the prevalence of lifestyle risk factors among users and non-users of any health-based mobile app among university students in Chennai. Cureus 2024 Mar 15;16(3):e56203-e56203 [FREE Full text] [doi: 10.7759/cureus.56203] [Medline: 38618332]
12. Remskar M, Ainsworth B, Maynard OM, Malkowski OS, Birch A, Burd AK, et al. Getting active through mindfulness: randomised controlled trial of a digital mindfulness-based intervention promoting physical activity engagement and enjoyment. Mental Health and Physical Activity 2025 Mar 08;28:100680-100680. [doi: 10.1016/j.mhpa.2025.100680]
13. Andargeery SY, El-Rafey DS. A randomized controlled trial of the effectiveness of the mHealth program in improving the lifestyle of nursing students. Sci Rep 2025 Mar 21;15(1):9765-9765 [FREE Full text] [doi: 10.1038/s41598-024-80982-2] [Medline: 40118869]
14. Pope ZC, Gao Z. Feasibility of smartphone application- and social media-based intervention on college students' health outcomes: a pilot randomized trial. J Am Coll Health 2022 Jan 09;70(1):89-98. [doi: 10.1080/07448481.2020.1726925] [Medline: 32150514]
15. Olatona FA, Van Onselen A, Kolanisi U. The effect of online nutrition and lifestyle education on body image dissatisfaction, body mass index, and disordered eating among female university undergraduate students in Lagos, Nigeria. South African Journal of Clinical Nutrition 2025 Mar 05;38(1):15-22. [doi: 10.1080/16070658.2024.2439752]
16. Seid A, Fufa DD, Bitew ZW. The use of internet-based smartphone apps consistently improved consumers' healthy eating behaviors: a systematic review of randomized controlled trials. Front Digit Health 2024 Jan 12;6:1282570-1282570 [FREE Full text] [doi: 10.3389/fdgth.2024.1282570] [Medline: 38283582]
17. Kim K, Hur M, Kim W. Effects of virtual reality (VR)-based meditation on sleep quality, stress, and autonomic nervous system balance in nursing students. Healthcare (Basel) 2024 Aug 08;12(16):1581-1581 [FREE Full text] [doi: 10.3390/healthcare12161581] [Medline: 39201140]
18. Lu Y, Lin H, Tsai P. Effects of digital sleep interventions on sleep among college students and young adults: systematic review and meta-analysis. J Med Internet Res 2025 May 12;27:e69657-e69657 [FREE Full text] [doi: 10.2196/69657] [Medline: 40354636]
19. Kechagias EP, Papadopoulos GA, Rokai I. Evaluating the impact of digital health interventions on workplace health outcomes: a systematic review. Administrative Sciences 2024 Jun 20;14(6):131-131. [doi: 10.3390/admsci14060131]
20. Lee S, Park J. Systematic review and meta analysis of standalone digital behavior change interventions on physical activity. NPJ Digit Med 2025 Jul 14;8(1):436-436 [FREE Full text] [doi: 10.1038/s41746-025-01827-4] [Medline: 40659761]
21. Milne-Ives M, Homer S, Andrade J, Meinert E. Associations between behavior change techniques and engagement with mobile health apps: protocol for a systematic review. JMIR Res Protoc 2022 Mar 29;11(3):e35172-e35172 [FREE Full text] [doi: 10.2196/35172] [Medline: 35348460]
22. Milne-Ives M, Homer SR, Andrade J, Meinert E. Potential associations between behavior change techniques and engagement with mobile health apps: a systematic review. Front Psychol 2023 Sep 18;14:1227443-1227443 [FREE Full text] [doi: 10.3389/fpsyg.2023.1227443] [Medline: 37794916]
23. Gabarron E, Larbi D, Rivera-Romero O, Denecke K. Human factors in AI-driven digital solutions for increasing physical activity: scoping review. JMIR Hum Factors 2024 Jul 03;11:e55964-e55964 [FREE Full text] [doi: 10.2196/55964] [Medline: 38959064]
24. Ferrari M, Allan S, Arnold C, Eleftheriadis D, Alvarez-Jimenez M, Gumley A, et al. Digital interventions for psychological well-being in university students: systematic review and meta-analysis. J Med Internet Res 2022 Sep 28;24(9):e39686-e39686 [FREE Full text] [doi: 10.2196/39686] [Medline: 36169988]
25. Matos Fialho PM, Wenig V, Heumann E, Müller M, Stock C, Pischke CR. Digital public health interventions for the promotion of mental well-being and health behaviors among university students: a rapid review. BMC Public Health 2025 Jul 18;25(1):2500-2500 [FREE Full text] [doi: 10.1186/s12889-025-23669-1] [Medline: 40681968]
26. Åsberg K, Eldh AC, Löf M, Bendtsen M. "Simply complicated": uncovering the processes of lifestyle behavior change among college and university students with access to a digital multiple lifestyle intervention. Digit Health 2024 Apr 09;10:20552076241245905 [FREE Full text] [doi: 10.1177/20552076241245905] [Medline: 38601184]
27. Ta A, Salgin N, Demir M. Real-time stress monitoring, detection, and management in college students: a wearable technology and machine-learning approach. arXiv. Preprint posted online on May 21, 2025 [FREE Full text] [doi: 10.48550/arXiv.2505.15974]
28. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. EFI 2018 Dec 18;34(4):285-291. [doi: 10.3233/EFI-180221]
29. Allman-Farinelli M, Partridge SR, McGeechan K, Balestracci K, Hebden L, Wong A, et al. A mobile health lifestyle program for prevention of weight gain in young adults (TXT2BFiT): nine-month outcomes of a randomized controlled trial. JMIR Mhealth Uhealth 2016 Jun 22;4(2):e78-e78 [FREE Full text] [doi: 10.2196/mhealth.5768] [Medline: 27335237]
30. Partridge SR, McGeechan K, Bauman A, Phongsavan P, Allman-Farinelli M. Improved eating behaviours mediate weight gain prevention of young adults: moderation and mediation results of a randomised controlled trial of TXT2BFiT, mHealth program. Int J Behav Nutr Phys Act 2016 Apr 02;13:44-44 [FREE Full text] [doi: 10.1186/s12966-016-0368-8] [Medline: 27039178]

XSL•FO
RenderX

31. Muntaner-Mas A, Sanchez-Azanza VA, Ortega FB, Vidal-Conti J, Borràs PA, Cantallops J, et al. The effects of a physical activity intervention based on a fatness and fitness smartphone app for university students. Health Informatics J 2021;27(1):1460458220987275-1460458220987275 [FREE Full text] [doi: 10.1177/1460458220987275] [Medline: 33446036]
32. Fitzsimmons-Craft EE, Firebaugh M, Graham AK, Eichen DM, Monterubio GE, Balantekin KN, et al. State-wide university implementation of an online platform for eating disorders screening and intervention. Psychol Serv 2018 Nov 08;16(2):239-249 [FREE Full text] [doi: 10.1037/ser0000264] [Medline: 30407047]
33. Morris J, Firkins A, Millings A, Mohr C, Redford P, Rowe A. Internet-delivered cognitive behavior therapy for anxiety and insomnia in a higher education context. Anxiety Stress Coping 2015 Jul 20;29(4):415-431. [doi: 10.1080/10615806.2015.1058924] [Medline: 26079158]
34. Inauen J, Bolger N, Shrout PE, Stadler G, Amrein M, Rackow P, et al. Using smartphone-based support groups to promote healthy eating in daily life: a randomised trial. Appl Psychol Health Well Being 2017 Sep 25;9(3):303-323. [doi: 10.1111/aphw.12093] [Medline: 28948690]
35. Hebden L, Cook A, van DPHP, King L, Bauman A, Allman-Farinelli M. A mobile health intervention for weight management among young adults: a pilot randomised controlled trial. J Hum Nutr Diet 2013 Aug 29;27(4):322-332. [doi: 10.1111/jhn.12155] [Medline: 23992038]
36. Napolitano MA, Lynch SB, Mavredes M, Shambon B, Posey L. Evaluating an interactive digital intervention for college weight gain prevention. J Nutr Educ Behav 2020 May 23;52(9):890-897. [doi: 10.1016/j.jneb.2020.04.007] [Medline: 32456988]
37. Floyd V, Vargas I. Evaluating a mobile application based intervention for insomnia in college students: a preliminary study. J Am Coll Health 2024 Nov 08;73(9):3540-3548. [doi: 10.1080/07448481.2024.2423225] [Medline: 39514819]
38. Hershner S, O'Brien LM. The impact of a randomized sleep education intervention for college students. J Clin Sleep Med 2018 Mar 15;14(3):337-347 [FREE Full text] [doi: 10.5664/jcsm.6974] [Medline: 29510791]
39. Smith L, Volkwyn C. Promotion of physical activity through health applications among students of selected universities: a preliminary study. Cent Eur J Sport Sci Med 2022 Feb 01;38(2):5-12. [doi: 10.18276/cej.2022.2-01]
40. Khatri S, Sharma R. Effective management of sedentary behavior among Indian university students: an empirical exploration into health-related behavior. J Educ Health Promot 2024 Apr 29;13:131-131 [FREE Full text] [doi: 10.4103/jehp.jehp_1489_23] [Medline: 38784278]
41. Cantisano LM, Gonzalez-Soltero R, Blanco-Fernández A, Belando-Pedreño N. ePSICONUT: An e-health programme to improve emotional health and lifestyle in university students. Int J Environ Res Public Health 2022 Jul 28;19(15):9253-9253 [FREE Full text] [doi: 10.3390/ijerph19159253] [Medline: 35954601]
42. O'Brien LM, Palfai TP. Efficacy of a brief web-based intervention with and without SMS to enhance healthy eating behaviors among university students. Eat Behav 2016 Sep 01;23:104-109. [doi: 10.1016/j.eatbeh.2016.08.012] [Medline: 27619174]
43. Cotten E, Prapavessis H. Increasing nonsedentary behaviors in university students using text messages: randomized controlled trial. JMIR Mhealth Uhealth 2016 Aug 19;4(3):e99 [FREE Full text] [doi: 10.2196/mhealth.5411] [Medline: 27543317]
44. Figueroa CA, Deliu N, Chakraborty B, Modiri A, Xu J, Aggarwal J, et al. Daily motivational text messages to promote physical activity in university students: results from a microrandomized trial. Ann Behav Med 2022 Feb 11;56(2):212-218. [doi: 10.1093/abm/kaab028] [Medline: 33871015]
45. Walsh JC, Corbett T, Hogan M, Duggan J, McNamara A. An mHealth intervention using a smartphone app to increase walking behavior in young adults: a pilot study. JMIR Mhealth Uhealth 2016 Sep 22;4(3):e109 [FREE Full text] [doi: 10.2196/mhealth.5227] [Medline: 27658677]
46. Ashton LM, Morgan PJ, Hutchesson MJ, Rollo ME, Collins CE. Feasibility and preliminary efficacy of the 'HEYMAN' healthy lifestyle program for young men: a pilot randomised controlled trial. Nutr J 2017 Jan 13;16(1):2 [FREE Full text] [doi: 10.1186/s12937-017-0227-8] [Medline: 28086890]
47. Kaneda K, Maeda N, Fukui K, Tashiro T, Komiya M, Urabe Y. Impact of the simultaneous distribution of e-learning and exercise videos on the health literacy and lifestyle of college students during the COVID-19 pandemic: a randomized controlled trial. J Phys Ther Sci 2024 Nov 01;36(11):703-710 [FREE Full text] [doi: 10.1589/jpts.36.703] [Medline: 39493688]
48. Malloy JA, Partridge SR, Kemper JA, Braakhuis A, Roy R. Feasibility and preliminary efficacy of co-designed and co-created healthy lifestyle social media intervention programme the daily health coach for young women: a pilot randomised controlled trial. Nutrients 2024 Dec 18;16(24):4364 [FREE Full text] [doi: 10.3390/nu16244364] [Medline: 39770984]
49. Chung AE, Skinner AC, Hasty SE, Perrin EM. Tweeting to health: a novel mHealth intervention using Fitbits and Twitter to foster healthy lifestyles. Clin Pediatr (Phila) 2016 Jul 16:26-32. [doi: 10.1177/0009922816653385] [Medline: 27317609]
50. Nour M, Chen J, Allman-Farinelli M. Young adults' engagement with a self-monitoring app for vegetable intake and the impact of social media and gamification: feasibility study. JMIR Form Res 2019 May 10;3(2):e13324 [FREE Full text] [doi: 10.2196/13324] [Medline: 31094322]
51. Roure C, Pasco D, Benoît N, Deldicque L. Impact of a design-based bike exergame on young adults' physical activity metrics and situational interest. Res Q Exerc Sport 2019 Nov 13;91(2):309-315. [doi: 10.1080/02701367.2019.1665621] [Medline: 31718499]

XSL•FO
RenderX

52. Kellner M, Dold C, Lohkamp M. Objectively assessing the effect of a messenger-based intervention to reduce sedentary behavior in university students: a pilot study. J Prev 2023 May 12;44(5):521-534 [FREE Full text] [doi: 10.1007/s10935-023-00735-1] [Medline: 37171555]
53. Hutchesson MJ, Morgan PJ, Callister R, Pranata I, Skinner G, Collins CE. Be positive be healthe: development and implementation of a targeted e-Health weight loss program for young women. Telemed J E Health 2015 Dec 24;22(6):519-528. [doi: 10.1089/tmj.2015.0085] [Medline: 26701611]
54. Schweitzer AL, Ross JT, Klein CJ, Lei KY, Mackey ER. An electronic wellness program to improve diet and exercise in college students: a pilot study. JMIR Res Protoc 2016 Feb 29;5(1):e29 [FREE Full text] [doi: 10.2196/resprot.4855] [Medline: 26929118]
55. Sarcona A, Kovacs L, Wright J, Williams C. Differences in eating behavior, physical activity, and health-related lifestyle choices between users and nonusers of mobile health apps. American Journal of Health Education 2017 Jul 11;48(5):298-305. [doi: 10.1080/19325037.2017.1335630]
56. Whatnall MC, Patterson AJ, Chiu S, Oldmeadow C, Hutchesson MJ. Feasibility and preliminary efficacy of the Eating Advice to Students (EATS) brief web-based nutrition intervention for young adult university students: a pilot randomized controlled trial. Nutrients 2019 Apr 23;11(4):905 [FREE Full text] [doi: 10.3390/nu11040905] [Medline: 31018565]
57. Wittmar S, Frankenstein T, Timm V, Frei P, Kurpiers N, Wölwer S, et al. User experience with a personalized mHealth service for physical activity promotion in university students: mixed methods study. JMIR Form Res 2025 Mar 28;9:e64384 [FREE Full text] [doi: 10.2196/64384] [Medline: 40153787]
58. Belogianni K, Ooms A, Lykou A, Nikoletou D, Jayne Moir H. An online game-based intervention using quizzes to improve nutrition and physical activity outcomes among university students. Health Education Journal 2023 Jun 09;82(6):636-650. [doi: 10.1177/00178969231179032]
59. Napolitano MA, Whiteley JA, Mavredes M, Tjaden AH, Simmens S, Hayman LL, et al. Effect of tailoring on weight loss among young adults receiving digital interventions: an 18 month randomized controlled trial. Transl Behav Med 2021 Apr 26;11(4):970-980 [FREE Full text] [doi: 10.1093/tbm/ibab017] [Medline: 33739422]
60. Xian Y, Xu H, Xu H, Liang L, Hernandez AF, Wang TY, et al. An initial evaluation of the impact of Pokémon GO on physical activity. J Am Heart Assoc 2017 May 16;6(5):e005341 [FREE Full text] [doi: 10.1161/JAHA.116.005341] [Medline: 28512111]
61. Stork MJ, Bell EG, Jung ME. Examining the impact of a mobile health app on functional movement and physical fitness: pilot pragmatic randomized controlled trial. JMIR Mhealth Uhealth 2021 May 28;9(5):e24076 [FREE Full text] [doi: 10.2196/24076] [Medline: 34047704]
62. Al-Nawaiseh HK, McIntosh WA, McKyer LJ. An-m-Health intervention using smartphone app to improve physical activity in college students: a randomized controlled trial. Int J Environ Res Public Health 2022 Jun 13;19(12):7228 [FREE Full text] [doi: 10.3390/ijerph19127228] [Medline: 35742477]
63. Hahn SL, Kaciroti N, Eisenberg D, Weeks HM, Bauer KW, Sonneville KR. Introducing dietary self-monitoring to undergraduate women via a calorie counting app has no effect on mental health or health behaviors: results from a randomized controlled trial. J Acad Nutr Diet 2021 Aug 20;121(12):2377-2388 [FREE Full text] [doi: 10.1016/j.jand.2021.06.311] [Medline: 34427188]
64. Fucito LM, Ash GI, Wu R, Pittman B, Barnett NP, Li CR, et al. Wearable intervention for alcohol use risk and sleep in young adults: a randomized clinical trial. JAMA Netw Open 2025 May 01;8(5):e2513167. [doi: 10.1001/jamanetworkopen.2025.13167] [Medline: 40445615]
65. Haslam RL, Baldwin JN, Pezdirc K, Truby H, Attia J, Hutchesson MJ, et al. Efficacy of technology-based personalised feedback on diet quality in young Australian adults: results for the advice, ideas and motivation for my eating (Aim4Me) randomised controlled trial. Public Health Nutr 2023 Feb 09;26(6):1293-1305. [doi: 10.1017/S1368980023000253] [Medline: 36755380]
66. Nour MM, McGeechan K, Wong AT, Partridge SR, Balestracci K, Roy R, et al. Diet quality of young adults enrolling in TXT2BFiT, a mobile phone-based healthy lifestyle intervention. JMIR Res Protoc 2015 May 27;4(2):e60 [FREE Full text] [doi: 10.2196/resprot.4484] [Medline: 26018723]
67. Oliveira C, Pereira A, Vagos P, Nóbrega C, Gonçalves J, Afonso B. Effectiveness of mobile app-based psychological interventions for college students: a systematic review of the literature. Front Psychol 2021 May 11;12:647606-647606 [FREE Full text] [doi: 10.3389/fpsyg.2021.647606] [Medline: 34045994]
68. Konlan KD, Ibrahim ZA, Lee J, Lee H. The inclusion of implementation outcomes in digital health interventions for young adults: a scoping review. Digit Health 2025 Mar 28;11:20552076251330194 [FREE Full text] [doi: 10.1177/20552076251330194] [Medline: 40162162]
69. Spanhel K, Balci S, Feldhahn F, Bengel J, Baumeister H, Sander LB. Cultural adaptation of internet- and mobile-based interventions for mental disorders: a systematic review. NPJ Digit Med 2021 Aug 25;4(1):128-128 [FREE Full text] [doi: 10.1038/s41746-021-00498-1] [Medline: 34433875]
70. Whitehead L, Talevski J, Fatehi F, Beauchamp A. Barriers to and facilitators of digital health among culturally and linguistically diverse populations: qualitative systematic review. J Med Internet Res 2023 Feb 28;25:e42719-e42719 [FREE Full text] [doi: 10.2196/42719] [Medline: 36853742]

XSL•FO
RenderX

71. Kim J, Aryee LMD, Bang H, Prajogo S, Choi YK, Hoch JS, et al. Effectiveness of digital mental health tools to reduce depressive and anxiety symptoms in low- and middle-income countries: systematic review and meta-analysis. JMIR Ment Health 2023 Mar 20;10:e43066-e43066 [FREE Full text] [doi: 10.2196/43066] [Medline: 36939820]
72. Hayes JF, Darling KE, Tomashek H, Elwy AR, Wing RR. Behavioral weight loss interventions in college health centers: a qualitative analysis of barriers and facilitators to implementation. Obes Sci Pract 2024 Dec 14;10(6):e70021-e70021. [doi: 10.1002/osp4.70021] [Medline: 39544506]
73. Hernández-Jaña S, Huber-Pérez T, Palma-Leal X, Guerrero-Ibacache P, Campos-Nuñez V, Zavala-Crichton JP, et al. Effect of a single nutritional intervention previous to a critical period of fat gain in university students with overweight and obesity: a randomized controlled trial. Int J Environ Res Public Health 2020 Jul 16;17(14):5149-5149 [FREE Full text] [doi: 10.3390/ijerph17145149] [Medline: 32708831]
74. Berman AH, Topooco N, Lindfors P, Bendtsen M, Lindner P, Molander O, et al. Transdiagnostic and tailored internet intervention to improve mental health among university students: research protocol for a randomized controlled trial. Trials 2024 Mar 01;25(1):158-158 [FREE Full text] [doi: 10.1186/s13063-024-07986-1] [Medline: 38429834]
75. Iwakura M, Ozeki C, Jung S, Yamazaki T, Miki T, Nohara M, et al. An umbrella review of efficacy of digital health interventions for workers. NPJ Digit Med 2025 Apr 14;8(1):207-207 [FREE Full text] [doi: 10.1038/s41746-025-01578-2] [Medline: 40229460]
76. Vollert B, Müller L, Jacobi C, Trockel M, Beintner I. Effectiveness of an app-based short intervention to improve sleep: randomized controlled trial. JMIR Ment Health 2023 Mar 21;10:e39052-e39052 [FREE Full text] [doi: 10.2196/39052] [Medline: 36943337]
77. Lee J, Kang M, Lee S. Effects of the e-Motivate4Change program on metabolic syndrome in young adults using health apps and wearable devices: quasi-experimental study. J Med Internet Res 2020 Jul 30;22(7):e17031 [FREE Full text] [doi: 10.2196/17031] [Medline: 32729838]
78. Daryabeygi-Khotbehsara R, Dunstan DW, Shariful Islam SM, Rhodes RE, Hojjatinia S, Abdelrazek M, et al. A control system model of capability-opportunity-motivation and behaviour (COM-B) framework for sedentary and physical activity behaviours. Digit Health 2024 Jun 07;10:20552076241255658 [FREE Full text] [doi: 10.1177/20552076241255658] [Medline: 38854921]
79. Paterson S, Dawes H, Winward C, Bartram E, Dodds E, McKinon J, et al. Use of the Capability, Opportunity and Motivation Behaviour model (COM-B) to understand interventions to support physical activity behaviour in people with stroke: an overview of reviews. Clin Rehabil 2024 Jan 09;38(4):543-557 [FREE Full text] [doi: 10.1177/02692155231224365] [Medline: 38192225]

Abbreviations

BCTTv1: Behavior Change Technique Taxonomy version 1
COM-B: Capability-Opportunity-Motivation-Behavior
DHI: digital health intervention
GRADE: Grading of Recommendations Assessment, Development and Evaluation
JBI: Joanna Briggs Institute
mHealth: mobile health
MMAT: Mixed Methods Appraisal Tool
MMS: mixed methods study
PICOS: Population-Intervention-Comparison-Outcome-Study Design
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Search Extension
PROGRESS-Plus: Place of Residence, Race/Ethnicity, Occupation, Gender/Sex, Religion, Education, Socioeconomic Status, Plus Other Relevant Factors
PROSPERO: International Prospective Register of Systematic Reviews
QDS: quantitative descriptive study
QNRS: quantitative nonrandomized study
QR: qualitative research
QRCT: quantitative randomized controlled trial
RoB 2: Risk of Bias 2

XSL•FO
RenderX

Edited by S Brini; submitted 11.Aug.2025; peer-reviewed by GLP Bodagala, F Owoseje, Y Elsanousi, M Ekwueme, H Kitiabi; comments to author 27.Oct.2025; accepted 05.Jan.2026; published 04.Feb.2026.

Please cite as:
Zhou Q, Jiang J, Yin Z, Fan R
Effect of Digital Health Interventions on College Students' Lifestyle Behaviors: Systematic Review
J Med Internet Res 2026;28:e82192
URL: https://www.jmir.org/2026/1/e82192
doi:10.2196/82192
PMID:

©Qingyuan Zhou, Jiajun Jiang, Zhihua Yin, Ruishi Fan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Key Components and Barriers in Web-Based Suicide Prevention Gatekeeper Training: Systematic Narrative Review

Olivier Ferlatte[1,2], PhD; Emmanuelle Gareau[1,2], MSc; Keven Lee[1,2], PhD; Kinda Wassef[1,2], MPH; John Lindsay Oliffe[3,4], PhD; Hannah Kia[5], PhD; Brock Dumville[6], MPH

[1]École de santé publique de l'Université de Montréal, Montréal, QC, Canada
[2]Centre de recherche en santé publique, Montréal, QC, Canada
[3]School of Nursing, University of British Columbia, Vancouver, BC, Canada
[4]Department of Nursing, University of Melbourne, Melbourne, Australia
[5]School of Social Work, University of British Columbia, Vancouver, QC, Canada
[6]Suicide Prevention Centre of Montreal, Montréal, QC, Canada

Corresponding Author:
Olivier Ferlatte, PhD
École de santé publique de l'Université de Montréal
7101 Av. du Parc
Montréal, QC, H3N 1X9
Canada
Phone: 1 514 343 6111 ext 44090
Email: olivier.ferlatte@umontreal.ca

Abstract

Background: Gatekeeper training programs (GTPs) are a key component of contemporary suicide prevention strategies, equipping community members and non–mental health professionals with the skills to identify, engage with, and refer individuals at risk of suicide. Increasingly, these programs are delivered via the web, offering a compelling alternative to in-person training through greater scalability, flexibility, and cost-effectiveness. However, little consensus exists regarding the design, modes of delivery, and implementation strategies of web-based GTPs. Further, there is a limited understanding of which components affect their usability and engagement.

Objective: This systematic narrative review aims to identify the key components—including facilitators and barriers—of web-based GTPs.

Methods: We systematically searched web-based databases (CINAHL, Embase, MEDLINE, PsycINFO, and Web of Science) to identify peer-reviewed articles published between 2000 and 2025 that involved web-based GTPs. After screening, 59 studies met the inclusion criteria and were analyzed using content analysis to identify key components and barriers affecting the delivery and receipt of web-based GTPs.

Results: Results were organized under 3 categories: design, content, and pedagogy. Key design considerations emphasized accessibility for diverse learning styles and digital literacy levels, customizability for different user groups, privacy protection, and the long-term sustainability of training content and delivery platforms. Core training content covered four domains: (1) suicide-related knowledge (eg, prevalence, myths, and at-risk groups), (2) gatekeeping skills (eg, understanding risk factors, recognizing warning signs, problem-solving and safety planning), (3) resource awareness (eg, available local resources and referral procedures), and (4) general mental health education (eg, mental fitness, mindfulness, and self-care strategies for gatekeepers). In terms of pedagogy, the reviewed studies used a wide range of strategies that comprised interactive learning activities (eg, simulation, practice exercises), periodic knowledge checks (eg, quizzes), and reinforcement mechanisms (eg, booster sessions). Additionally, fostering a sense of community (eg, online support spaces or discussion forums) and promoting trainees' autonomy (eg, self-paced training) were highlighted as key components of training delivery.

Conclusions: Web-based GTPs represent a promising avenue for expanding access to suicide prevention training. Their effectiveness may be strengthened through the integration of frameworks tailored to web-based learning environments, as well as interactive and user-centered design elements that support learning and retention. Future research should examine the acceptability, feasibility, and sustainability of these programs, while also refining their adaptation for diverse populations. In this regard, co-design approaches could facilitate the tailoring of such programs to the needs and specificities of their target populations.

XSL•FO
RenderX

Overall, enhancing the design and delivery of web-based GTPs may ultimately improve their contribution to suicide prevention efforts.

(J Med Internet Res 2026;28:e81572) doi:10.2196/81572

KEYWORDS

narrative review; suicide prevention; suicide; online; web-based; gatekeeper training

Introduction

Suicide is a critical public health issue worldwide, with significant social, emotional, and economic impacts on individuals, families, and communities [1]. As one of the leading causes of preventable death [2], suicide requires multifaceted approaches including awareness raising, reducing mental illness stigmas, and enhancing timely interventions [3]. Among these strategies, gatekeeper training programs (GTPs) have emerged as a cornerstone in equipping individuals to identify, approach, and support those at risk of suicide [4]. Gatekeepers are non–mental health professionals who may have contact with individuals at risk of suicide (ie, educators, parents, peers, or other community members), and are trained to recognize warning signs, initiate conversations about suicide, and connect individuals to appropriate professional help [1,5]. These programs, which are endorsed by the World Health Organization (WHO) [6], have demonstrated effectiveness in various settings and populations [7], highlighting their potential key role in suicide prevention.

Traditionally, GTPs have been delivered through in-person workshops and seminars [5], offering opportunities for direct interaction, role-playing, and immediate feedback. However, advances in technology and the increasing digitization of health education have led to the development and adoption of web-based GTPs [8,9]. Web-based formats provide significant advantages, including scalability, accessibility for geographically dispersed participants, and the ability to tailor training to diverse populations [7,10,11]. These programs can not only overcome logistical barriers (eg, physical attendance or availability of training) [12] but can also be particularly valuable for populations where confidentiality and anonymity are essential, such as stigmatized communities, including migrants and lesbian, gay, bisexual, trans, queer, and other sexual and gender minorities (LGBTQ+) populations [13]. Indeed, confidentiality can encourage more meaningful engagement with sensitive topics such as mental health and suicide among communities already affected by cumulative stigmas [14]. In addition, web-based training allows participants to learn at their own pace, accommodating busy schedules and varying levels of prior knowledge [15].

Recent evaluations indicate that web-based and in-person GTPs have similar effectiveness [8,9,11,14]. However, there are implementation challenges for web-based programs, including limited internet access, technological difficulties, varying levels of digital literacy, and user engagement issues [15]. Moreover, the design and content of web-based training are critical, and poorly structured or overly generic programs may fail to meet the complex and varied needs of participants [16].

Nonetheless, recent advancements in interactive technologies present unique opportunities to innovate and scale suicide prevention with GTPs. However, effectively harnessing web-based technologies requires a clear understanding of both the factors that contribute to program success and the challenges that hinder implementation. This systematic narrative review synthesizes the current evidence for web-based GTPs, addressing two key questions: (1) What are the key components of promising and successful web-based suicide prevention GTPs? (2) What are the barriers to delivery and usability of these programs? In summarizing the existing literature, this review provides recommendations for the development, implementation, and scaling of web-based GTPs, with the overall goal of contributing to global efforts for reducing suicide.

Methods

Overview

This systematic narrative review was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Multimedia Appendix 1) [17] and has been registered with PROSPERO (Registration ID: CRD42023462414). Since this study is a review of published, peer-reviewed articles, ethical approval was not required.

Search Strategy

A comprehensive search strategy was developed and conducted under the guidance of a specialized university librarian. The strategy combined 3 main concepts: suicide prevention, gatekeeping intervention, and web-based training. The full search strategy is available in Multimedia Appendix 2. The search was initially launched on October 1, 2023, and conducted across 5 databases: CINAHL, Embase, MEDLINE, PsycINFO, and Web of Science. Additionally, we searched the first 20 pages of Google Scholar, as results beyond the first 20 pages were not related to our 3 main concepts. The reference lists of all subsequently included articles were manually reviewed to identify further relevant sources. To ensure the inclusion of the most recent literature, the complete search strategy was relaunched on June 19, 2025.

Eligibility Criteria

We included peer-reviewed research articles on web-based GTPs, published in English or in French from 2000 onwards. We included all types of studies and designs, except for articles that did not discuss the web-based components of the training and research protocols. Articles focusing on general mental health programs not specific to suicide prevention were excluded. In addition, articles targeting mental health professionals (eg, psychotherapists, psychiatrists, and mental health nurse practitioners) were excluded, as their educational

XSL•FO
RenderX

background and professional experiences could result in significantly different training components, facilitators, and barriers compared with training programs for laypersons [18]. Articles targeting health care professionals outside specialized mental health—such as pharmacists, pharmacy staff, health care lecturers, and police officers—were eligible for inclusion. Studies published before 2000 were also excluded due to significant technological advancements since that time.

Screening and Study Selection

Using Covidence, EG, KW, and KL independently screened all titles and abstracts. Any records deemed potentially relevant by at least one reviewer were then subjected to a full-text examination, which followed the same independent screening process. In case of disagreement after full-text screening, EG, KW, and KL discussed their interpretations of the articles until they reached consensus. If consensus could not be achieved, OF served as a third-party mediator.

Data Extraction

Data extraction was conducted using Covidence. KW and EG independently extracted the following descriptive data items:

1. Methodological information: Geographical context, study goals, research approach, and study design.
2. Gatekeeper training program information: Program name, training setting, format, if training was adapted from an existing program, development process, training supporting platform, training objectives, training description, organizing institution, trainees' characteristics, number of trainees, components of training, topics covered by training, effectiveness of the training, target population of the gatekeeping intervention.

Conflicts in data extraction were resolved jointly by KW and EG using the Covidence Consensus tool. These data were used to characterize the included studies and inform Table 1, but were not analyzed further in the identification of barriers or facilitators. The analysis of facilitators and barriers was conducted independently of the data extraction process and is detailed in the following section.

XSL•FO
RenderX

Table 1. A brief overview of the characteristics of the training programs included in the systematic narrative review of web-based gatekeeper training programs (GTPs).

Author	Name	Components of training	Topics covered[a,b]			
			Information about suicide	Information about being a gatekeeper	Information about resources/referrals	Information about general mental health
Afsharnejad et al [19]	Talk-to-me MOOC[c]	Videos, quiz/tests	✓	✓		
Albritton et al [20]	Be present	Videos, e-tool box, homework assignments, social media posting	✓			
Bartgis and Albright [21]	Kognito gatekeeper simulations	Role-play/simulation		✓	✓	
Brown et al [22]	Indigenous network suicide intervention skills training (INSIST) program	N/S[d]	N/S	N/S	N/S	N/S
Bryant et al [23]	Kogito "At Risk Primary Care"	Role-play/simulation		✓	✓	
Canady [24]	Signs matter: early detection	Role-play/simulation quiz/tests, resources	✓			
Canady [24]	At-risk for high school educators	Role-play/simulation quiz/tests, resources	✓	✓	✓	
Canady [24]	At-risk for middle school educators	Role-play/simulation quiz/tests, resources	✓	✓	✓	
Carpenter et al [25]	Online veteran administration's (VA) suicide prevention GTP (S.A.V.E./SAVE [signs, ask, validate, encourage/expedite])	Videos, role-play/simulation	✓	✓		
Caughlan et al [26]	Mind4Health	Videos, readings, resources		✓	✓	
Cohen et al [27]	Israeli gatekeeper training	Role-play/simulation	✓		✓	
Colder Carras et al [28]	Stack-Up overwatch program	N/S	N/S	N/S	N/S	N/S
Coleman et al [29]	Kognito at risk for college students	Role-play/simulation		✓	✓	
Colucci et al [30]	Suicide first aid guidelines training	Videos, quiz/tests, infographics, homework assignments, reflective journaling	✓	✓	✓	
Ghoncheh et al [8]	MHO	PowerPoint presentation, audio features, graphs, quiz/tests, reading material, discussion board	✓	✓	✓	
Ghoncheh et al [8]	Children and family court advisory and support service program	N/S	N/S	N/S	N/S	N/S
Ghoncheh et al [8]	Question persuade and respond (QPR) online gatekeeper training	PowerPoint presentation, role-play/simulation, quiz/tests, reading material, videos, audio features	N/S	N/S	N/S	N/S
Ghoncheh et al [8]	Hollywood homeless youth partnership (HHYP) program	PowerPoint presentation, audio features, quiz/tests	N/S	N/S	N/S	N/S
Ghoncheh et al [8]	In the line of duty	Videos, audio features	N/S	N/S	N/S	N/S
Ghoncheh et al [15]	MHO	PowerPoint presentation, quiz/tests, discussion board	✓	✓	✓	

XSL•FO
RenderX

Author	Name	Components of training	Topics covered[a,b]			
			Information about suicide	Information about being a gatekeeper	Information about resources/referrals	Information about general mental health
Hawley et al [31]	Not applicable	Didactic content, video and audio clips, and reflection questions	✓	✓	✓	✓
Hill and McCray [32]	The Texas ask about suicide to save a life (AS + K?) suicide GTP	N/S	✓	✓	✓	
Hill et al [33]	ASK about suicide to save a life (AS + K?)	Videos	✓	✓		
Hofmann et al [34]	COPS (coping with suicide)	Videos, reading material, worksheets, quiz/tests	✓	✓	✓	
Hofmann and Wagner [35]	N/S	Videos, audio plays, manual	✓	✓	✓	✓
Holmes et al [14]	Start	N/S	N/S	N/S	N/S	N/S
Kawashima et al [36]	N/S	PowerPoint presentation, videos	✓	✓	✓	
Kimbrel et al [37]	Safety planning intervention (SPI)	PowerPoint presentation, videos, role-play/simulation, reading material, worksheets		✓		
Kingi-Ulu'av et al [38]	LifeKeepers booster session	Readings	✓	✓	✓	
Kingi-Ulu'av et al [7]	QPR	N/S	N/S	N/S	N/S	N/S
Kingi-Ulu'av et al [7]	MHO	N/S	N/S	N/S	N/S	N/S
Kingi-Ulu'av et al [7]	I CARE	N/S	N/S	N/S	N/S	N/S
Kingi-Ulu'av et al [7]	Act on FACTS: making educators partners (MEP)	N/S	N/S	N/S	N/S	N/S
Kingi-Ulu'av et al [7]	Kognito gatekeeper simulations	N/S	N/S	N/S	N/S	N/S
Kreuze and Ruggiero [39]	Kognito at-risk for high school educators	Videos, role-play/simulation		✓	✓	
Kreuze and Ruggiero [39]	QPR	Videos	✓	✓	✓	
Kreuze and Ruggiero [39]	MEP in youth suicide prevention: ACT on FACTS	Videos, role-play/simulation, audio features	✓	✓	✓	
Kreuze et al [10]	QPR	Videos, testimonials, narration, bulleted lists, mnemonics, pocket cards, role-play/simulation, self - audit checklist	✓	✓	✓	
Kreuze et al [10]	MEP in youth suicide prevention	Videos lectures, expert content, conversations example, role-play, testimonies, activities related to videos	✓	✓	✓	
Lamis et al [40]	MEP in youth suicide prevention: ACT on FACTS	Lecture, question and answers, digital vignettes/interactive activities	✓	✓	✓	
Lancaster et al [9]	Web-based QPR	Videos, text, pictures, audio features	N/S	N/S	N/S	N/S
Lee-Tauler et al [41]	Chaplains-CARE online program	Didactic lectures, videos, reading material, quiz/tests, interactive activities	✓	✓		✓

XSL•FO
RenderX

Author	Name	Components of training	Topics covered[a,b]			
			Information about suicide	Information about being a gatekeeper	Information about resources/referrals	Information about general mental health
Liu et al [42][e]	Various	Various	N/S			
MacDonald Hart et al [43]	Suicide intervention first aid (SIFA)	Didactic lectures, interactive, discussions, role-play/simulation exercise, skills practice	✓	✓	✓	
Manning and Van Deusen [44]	Western Michigan University suicide prevention program online course	Videos, photographs, and graphics	✓	✓	✓	
Marley et al [45]	Pharm-SAVES training	N/S		✓	✓	
McKay et al [46]	Living works start	Videos, reading material	✓	✓	✓	
Mirick [47]	SOS (signs of suicide) for school staff	Role-play/simulation with both child and adolescent	✓	✓	✓	
Mishkind et al [48]	VitalCog: suicide prevention in the workplace (formerly known as Working Minds)	Videos, role-play/simulation, group discussion, workbook	✓	✓	✓	
Osteen et al [49]	QPR for law enforcement	N/S		✓		
Perepezko et al [50]	Stack-Up overwatch program	N/S	N/S	N/S	N/S	N/S
Pilbrow et al [51]	Advanced suicide prevention training for pharmacists	Video, role-play/simulation, digital workbook, group discussion	✓	✓	✓	
Postuvan et al [52]	IAlive (iˇZiv in Slovenian)	Videos lectures, animated examples, interactive images/graphics with pop-ups	✓	✓		
Quinnett [53]	QPR pathfinder training	Video, role-play/simulation, reading	✓	✓	✓	✓
Reifegerste et al [54]	Help for relatives	Videos, audio-recordings, manual, text content	✓	✓	✓	✓
Rein et al [55]	Kognito	Role-play/simulation		✓		
Robinson-Link et al [56]	Kognito	Role-play/simulation	✓	✓	✓	
Roslan et al [57]	Online advanced C.A.R.E suicide prevention GTP (AdCARE)	Role-play/simulation	N/S	N/S	N/S	N/S
Ross et al [58]	Suicide prevention for college student gatekeepers program	Role-play/simulation, skills practice, discussions, peer cofacilitation	✓	✓		
Ross et al [59]	Suicide prevention for college student gatekeepers program	Role-play/simulation, skills practice, discussions	✓	✓		
Schmeckenbecher et al[e] [60]	Various	N/S	N/S	N/S	N/S	N/S
Seabury [61]	Crisis counseling: I Am Chipper!	Interactive PowerPoint presentation, videos, role-play/simulation, reading material	✓	✓		
Seabury [61]	Suicide assessment: rube farmer	Interactive PowerPoint presentation, videos, role-play/simulation, reading material	✓	✓		

XSL•FO
RenderX

Author	Name	Components of training	Topics covered[a,b]			
			Information about suicide	Information about being a gatekeeper	Information about resources/referrals	Information about general mental health
Seabury [62]	Crisis counseling: I Am Chipper!	Interactive PowerPoint presentation, videos, role-play/simulation, reading material, quiz/tests	✓	✓		
Seabury [62]	Suicide assessment: rube farmer	Interactive PowerPoint presentation, videos, role-play/simulation, reading material, quiz/tests	✓	✓		
ShantaBridges et al [63]	Suicide prevention and awareness for depression	N/S	✓	✓	✓	
Smith-Millman et al [64]	Kognito	Role-play/simulation		✓	✓	
Stone et al [65]	Youth suicide prevention: an introduction to gatekeeping	PowerPoint presentation, quiz/tests, resources, worksheet, audio files	✓	✓	✓	✓
Stover et al [66]	Pharm - SAVES	Videos, reading material, resources	✓	✓	✓	
Sun et al [67]	Chinese life gatekeeper training program	Videos, role-play/simulation, contextual understanding, group discussion, Q and A session	✓	✓	✓	
Teo et al [68]	VA S.A.V.E.	Videos	✓		✓	
Teo et al [69]	VA S.A.V.E.	Videos, vignettes	✓	✓	✓	
Timmons-Mitchell et al [70]	Kognito At-Risk for Middle School Educators	Role-play/simulation	✓	✓	✓	
Wislocki et al [71]	Multiple (n=506)[f]	Videos	✓	✓	✓	

[a]We categorized topics in four categories: (1) Information about suicide (including information about suicide prevention, suicidal or self-injury behaviors, suicide myths, suicide prevalence and statistics, risk factors for suicide, protective factors against suicide, and signs of mental distress/suicidal ideation/warning signs), (2) Information about being a gatekeeper (including intervention skills and identification of at-risk individuals), (3) Information about resources and referrals, and (4) Information about general mental health (including mental fitness and self-care).

[b]We exclusively reported the topics explicitly mentioned in the article, but we acknowledge that the training programs might cover additional topics not mentioned in the article.

[c]MOOC: mass open online course.

[d]N/S: not specified.

[e]These systematic reviews and meta-analyses are included for thoroughness; however, data from the primary studies were not re-extracted in this table since most were already included individually, and the remaining did not meet our eligibility criteria.

[f]This scoping review included 506 training videos. For the sake of conciseness, we did not include the program names in this table.

Data Analysis

We used an inductive content analysis approach [72,73] to identify the barriers and facilitators of web-based GTPs. Unlike the data extraction process, which aimed to summarize descriptive characteristics of the studies, the content analysis involved a separate, in-depth examination of the full texts of the included articles. EG and KW used Dedoose (version 9.0.107; Socio Cultural Research Consultants, LLC) to independently and inductively code full-text PDFs of the articles to reduce bias in interpretation. Coding began with line-by-line open coding, assigning initial code to relevant segments of text without predetermined categories. After 10% of the articles were coded, the two coders met to compare their coding frameworks and ensure consistency. Joined by KL, they then proceeded to independently code the remaining articles. This process allowed for a systematic and thorough examination of the data and minimized the risk of missing important nuances. Once all articles were coded, EG aggregated codes into higher-order headings, which were reviewed by the research team to develop broader categories through the process referred to as abstraction [72].

Results

Results of the Search

Our search yielded 4662 entries, from which 1365 duplicates were removed. After title and abstract screening, 3143 articles were excluded, and 95 articles were further removed after full-text screening (Figure 1). The final sample included 59 articles published between 2003 and 2025.

XSL•FO
RenderX

Figure 1. PRISMA flow diagram.

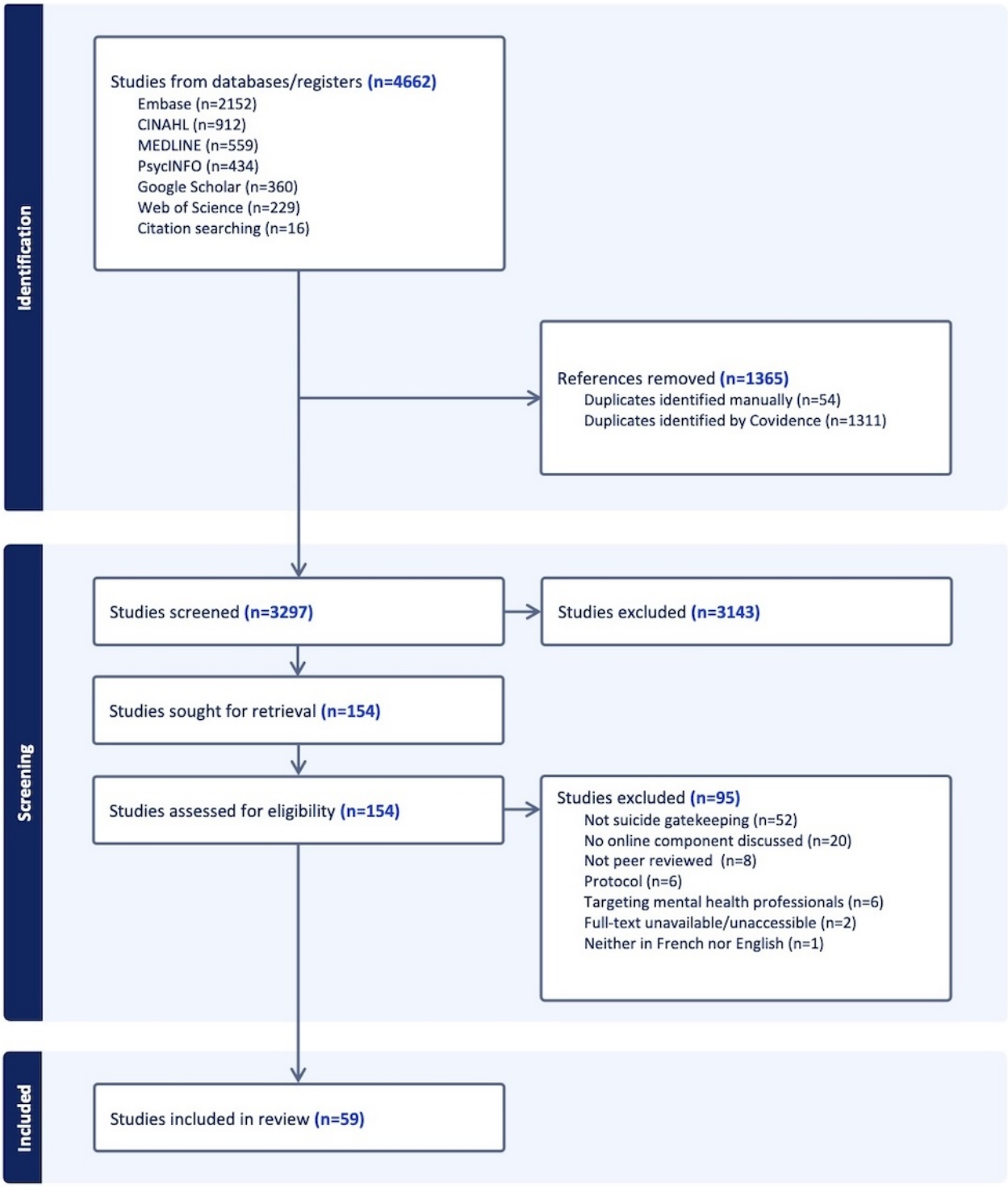

Reports Included in the Synthesis

Among the 59 articles reviewed, 36 used quantitative methodologies, including 11 randomized controlled trials. Additionally, 10 articles discussed multiple training [7,8,10,24,39,42,60-62,71]. The most common training programs discussed were Kognito (n=7); S.A.V.E./SAVE (signs, ask, validate, and encourage/expedite) or pharm-SAVE (n=5); and question, persuade, refer (n=4; excluding counts from review articles). The program settings varied, with schools/universities (n=21), military or law enforcement environments (n=6), and health care/ clinical settings (n=4) being the most common. Training duration ranged widely, from 20 minutes [25] to 32 hours [50] or multiple weeks [26,63,65]. Most training programs were aimed at youth or students (n=13), school staff (n=9), health care providers (n=6), or parents/caregivers (n=3). A brief overview of the characteristics of the GTPs included in this review is presented in Table 1, while a complete version is provided in Multimedia Appendix 3.

XSL•FO
RenderX

Design

We identified several critical design elements for web-based GTPs. Foremost, authors recommend that programs be grounded in theory and evidence-based practices [23,26,31,43,52,54,67,71]. For instance, Timmons-Mitchell et al [70] recommended evidence-based techniques such as motivational interviewing, and Afsharnejad et al [19] emphasized that mental health education principles, specifically the PERMA framework, should be used to enhance the efficacy of training. In addition, it was recommended that programs align with existing available suicide prevention initiatives to leverage and integrate resources [27,53,58]. This design specificity underlines web-based GTPs as a connector between community members and existing prevention efforts rather than being a stand-alone solution [22,43,54]. Indeed, it was emphasized by several authors that web-based GTPs should complement and be purposefully integrated rather than aspire to replace other suicide prevention initiatives [22,30,40,54,56]. To ensure usefulness and promote uptake, several programs integrated a co-design approach, involving interdisciplinary experts (including website developers), stakeholders, and potential end users participating in the design and implementation processes to better tailor the training programs [19,22,26,28,34,37,50,66-68,71]. In addition, the need for adequate financial support to sustain, maintain, and improve the programs to ensure the integration of new knowledge and best practices was mentioned [8,21].

Accessibility was also an essential design element. Authors highlighted that technical issues, such as glitches in program software, can hinder accessibility by disrupting the flow of the training [10,34,41,61]. Despite assumptions of widespread internet access, real-world barriers such as low bandwidth, limited availability of internet or technologies (eg, computers, smartphones), and insufficient media skills persist, potentially limiting users' access [10,21,22,30,35,61]. Authors suggested that the training platforms should be more accessible by reducing the bandwidth requirements and ensuring 24/7 availability [8,53]. They also emphasized simplifying navigation [54] through avoiding registration [52], the use of specific software (eg, Adobe), or requiring specific web browsers [10]. Further, the variability of individual trainee competencies and preferences in terms of technology use necessitates that training platforms be user-friendly and ideally use technologies familiar to learners for a better user experience [22,48,52,57,70]. Brown et al [22] recommended the adaptability of training programs across multiple technological devices (eg, computer, smartphones, and tablets) and suggested incorporating diverse web-based supports (mobile app, Facebook group, messenger feature, etc). Accessibility was also contingent on delivering content through diverse formats, including audio, video, and text, and ensuring these vehicles could be adjusted to suit individual learning speeds and accommodate trainees with disabilities and learning or attention deficits. Practically, this means adding audio components to written material [8] with a sufficiently large font [10], being able to control the speed of audio and video content [30], and adding embedded text explanations and captions to videos [66]. Multiple studies [10,54,61] suggested adding simple, nongeneric visual components to text as well as color-coding sections to create a visual learning structure, avoiding long sentences to describe concepts and prioritizing bullet points or synthesized information. Authors also highlighted the importance of accessibility and program duration, noting that web-based training programs should be as brief as possible while still meeting learning objectives [10,20,25,34,37,41,44,46,54,57,69]. For example, trainees in the study by Carpenter et al [25] preferred the theoretical part of the training to last 30 minutes or less, and trainees in the study by Reifegerste et al [54] rated videos averaging 7-8 minutes to be too lengthy. Still, Liu et al [42] recommend that training sessions last more than 2 hours to induce a significant change in trainees' attitudes and behaviors. Last, Carpenter et al [25] and Wislocki et al [71] also emphasized the importance of cost for the training programs, and making them financially feasible for a wide range of trainees.

Customizability was another critical element wherein the program design needed to be adaptable to meet trainees' unique needs, personal characteristics, and professional backgrounds [8,22,26,27,31,35,41,43,53,54,57,64,71]. This approach comprised providing personalized feedback tailored to each trainee's progress [10,21,57] and developing customized learning experiences for trainees based on their prior gatekeeping and suicide prevention experience and knowledge [8,22,32,37,54]. This included adding complementary modules, supplemental material, and additional content or offering flexible options for practice sessions [10,47]. Moreover, the customization of web-based training should encompass a variety of gatekeeping examples, scenarios, and practice exercises, thereby facilitating the alignment of content with diverse trainee preferences [10,25,37,41,66,70]. Furthermore, to facilitate customization that accommodates trainees' variable schedules and pacing, authors recommend enhancing program flexibility through multiple learning modes and having the option to segment learning sessions [8,10,31,35,48,70].

Another design consideration for web-based training was standardization. Emphasized was the importance of implementing mechanisms to ensure consistency in the delivery of training content [31]. In web-based synchronous programs led by human instructors, standardization largely depends on the instructor's expertise and familiarity with the material and the target audience, which was not specified in the reviewed articles. While instructors can adapt to trainees' needs and pace, maintaining consistency remains crucial, especially given potential fatigue, which can affect delivery. In asynchronous programs, standardization was built into the training itself. To support this, authors have recommended incorporating elements such as standardized scenarios, automated feedback systems, and prerecorded videos to ensure uniformity across sessions [21,62]. These standardization mechanisms promote consistency and create a sense of psychological safety for trainees, as they reduce the risk of judgment from instructors or peers [21,62,70]. Bartgis et al [21] have recommended standardization in delivery mechanisms, including web-based role-play modalities using avatars, to ensure consistency in content delivery, regardless of trainees' personal characteristics.

XSL•FO
RenderX

Another design consideration was privacy [10,22,28,42,46]. Reifegerste et al [54] noted that web-based formats may offer a greater sense of anonymity compared with in-person settings. While complete anonymity cannot be guaranteed in web-based training programs, Lancaster et al [9], Caughlan et al [26], and Reifegerste et al [54] argue that it remains an important feature, as it may reduce trainees' anxiety, thereby encouraging more open engagement with sensitive topics. Some authors raised concerns about privacy risks associated with using third-party internet platforms. For example, Brown et al [22] noted that while using a private Facebook group can foster support and connections among trainees, it may also pose risks to participants' privacy and confidentiality. Further, confidentiality concerns extended beyond the training itself to include program evaluation. Thus, it is essential that all data collected during the training, including during role-playing exercises, is stored on a secure server [21].

The final key element to consider in the design phase was sustainability. Specifically, maintaining and updating technological aspects were significant challenges for the use and longevity of web-based GTPs [8]. Brown et al [22] suggested that program designers allocate resources to ensure the maintenance and updates of the training in the initial development plan. For example, they specified that while web-based forums can be found valuable among trainees, they require substantial effort in terms of moderation and maintenance, which can be resource-intensive. Ghoncheh et al [15] emphasized the importance of developing a sustainability plan that minimizes maintenance and cost while preserving the integrity of the program.

Content

A consistent content recommendation was the use of clear and concise terminology and vocabulary familiar to trainees [10,51,54,69]. In programs designed for nonprofessionals, avoiding technical terminologies or clinical jargon, which may act as a barrier, was a consistent recommendation [10,37,54]. Authors recommended balancing testimonials ("emotional content") and practical information ("informative content") to better engage trainees and support the destigmatization of suicide prevention [10,54]. In addition, the importance of activities that help trainees feel comfortable using the word "suicide" [25] and support them in discussing suicidality in a nonstigmatizing manner was emphasized [19,26,27,31,57]. At the same time, caution was advised against including graphic details of suicide death [57] or downplaying the gravity of the topic [47]. Authors of the included studies recommended providing a manageable amount of clear, straightforward, and easy-to-follow information to avoid overwhelming trainees [10,69] while still offering sufficiently rich content [47,54].

In terms of didactic topics to cover, trainees endorsed four different topics: (1) Information about suicide including suicide definition, epidemiology, statistics and prevalence [10,25,27,31,43,52,66], and suicide myths and beliefs [10,25,32,52]; contextual factors connected with suicide [26,30], content about at-risk subgroups [27,30,35,52,54], legal requirements or policies [10,25], and broader community concerns regarding suicide prevention [22,58]; (2) Information about how to be a gatekeeper including how to identify at-risk individuals [22] and the differentiation between warning signs and risk factors according to various settings [10,27,30,31,35,43,52,54,58,69,71], clear and memorable steps to follow for gatekeeping interventions [10,25,51,52,66], ways to initiate conversations about suicide [25,26,35,51,52,58,69], what to say and topics to avoid when talking with suicidal individuals [41,47,54,69], safety planning and problem-solving [35,41,71], including with dealing with nonreceptive individuals [50], and follow-up strategies postintervention and postvention care [22,41,66,71]; (3) Information about the broad range of existing services including referral guidance [10,22,25,27,30,35,42,43,45,50,52,54,69,71] according to existing local resources [10,21,25,26,66]; and (4) Information about general mental health for gatekeepers [42,54] including mental fitness [19] or mindfulness [41], the challenges of being a [22,25,30,31] and gatekeepers' well-being and self-care strategies [22,35,41,57-60]. Importantly, these topics should be adapted to trainees' backgrounds and accompanied by relevant examples to which trainees could relate [10,25,41,42,45,51,66] as well as the specific context of the intervention [10,22,30,43,47]. The importance of ensuring that the content was culturally relevant to the intended audience was described as a factor bolstering the effectiveness and the inclusivity of training programs [7,21,22,26,27,30,38,42,53,63,67]. Notably, 4 studies underscored the limited diversity in training, pointing to a lack of content and examples specifically addressing the needs of LGBTQ+ populations and women [28,41,54,71]. In contrast, topics perceived as less relevant by trainees included procedural guidance on reporting suicide cases, professional assessment practices that were not directly applicable to their roles or lived experiences, and research data from unrelated contexts [22,30].

In addition to the topics above, several core skills and competencies to be acquired by trainees were highlighted. These include the ability to establish a strong rapport and build a trusting relationship with individuals experiencing suicidal ideation [22,30]. Brown et al [22] and Hawley et al [31] emphasized the inclusion of interpersonal "soft skills"—including active listening, compassion, patience, and nonjudgmental attitudes—as foundational elements of effective training. Similarly, Bartgis et al [21] underscored the importance of motivational interviewing skills, including the use of open-ended questions, providing affirmation, reflective listening, and summarizing. The development of advocacy skills was also identified as a particularly important skill when supporting individuals living in marginalizing conditions who face significant structural barriers to services [22].

Pedagogy

Several pedagogical elements were identified as critical for the effective delivery of web-based gatekeeper training. First, the inclusion of interactive learning activities—including role-plays, hands-on activities, practice, and scenario-based exercises—was consistently emphasized as essential to skills development and enhancing learning outcomes [8,10,24,25,36,41,47,53,57,63,64]. Trainees highly valued these activities [10,22,23,25,26,30,37,41,45,51,62,69]. For example, participants in 4 studies [10,41,47,51] emphasized the importance of having

XSL•FO
RenderX

more time allocated for practice during the web-based training through hands-on exercises and interactive learning opportunities to enhance motivation and information retention. Ross et al [59] also suggested having small group sizes to increase participation and knowledge gains in the case of synchronous training. Participants in 5 studies [10,25,47,57,66] endorsed role-plays that were concrete, realistic, relatable, and applicable to their contexts. Although some authors noted challenges for implementing role-plays via web [15,30,51], Seabury [62] and Liu et al [42] encouraged leveraging the use of innovative technologies. Examples of such technologies include the use of avatars [21,23,29,55,60,64], virtual reality [60], interactive videos or video demonstrations [41,51,66], live videoconferencing role-play practice sessions [10], or mathematical behavioral models and algorithms to create realistic simulations where trainees can practice gatekeeper skills [70]. Knowledge checks, such as tests and quizzes, were identified as another vital component of interactive learning [8,10,34,41], providing trainees with immediate feedback on their understanding of the material covered [61]. However, some authors cautioned that web-based training might lack the interactive and adaptable learning environments synonymous with traditional in-person training programs where instructors could respond to trainees in real-time [62,71]. Offering clarifications and real-time feedback dispersed throughout the training, including via knowledge checks, could address this limitation [8,10,41,62].

A second key pedagogical element was the autonomy for trainees regarding the pace and style of learning [8,9,19,21,34,35,62]. Inversely, several authors cautioned against over-reliance on trainees' intrinsic motivation [9,20], which can limit engagement [10,19] and even increase attrition [65]. Thus, authors emphasized that it is crucial to implement ongoing guidance, learner incentives (eg, raffle for a gift certificate [44]), and motivational strategies (eg, email reminders to finish the training [37]), given that intrinsic motivation may not be sufficient to complete the training [19,41,63]. Such mechanisms could take the form of a time limit for training completion [20] or reminders to encourage trainees to complete the training [37,41].

Building a community of practice emerged as another key recommendation. Unlike face-to-face training, web-based training programs often limit direct interactions with trainers as well as with other trainees. This lack of human interactions was mentioned as hindering skills practice and development [19,21]. Afsharnejad et al [19] highlighted the importance of fostering emotional connection with trainees. Establishing a community of practice, as suggested by other authors, could support connectivity by enabling trainees to access expert insights and feedback [8,41,53,57]. Some also highlighted that such communities of practice could provide trainees with opportunities to be paired with gatekeeping buddies for peer support [22], offering continuing networking and debriefing opportunities [22,38], offering support [59], and contributing to shared learning [51]. Examples of implemented approaches include messaging platforms [57], moderated forums with discussion boards and threads [8,65], chatrooms [28], and digital coaches providing direct and personalized feedback [21].

Finally, reinforcement strategies were recommended to consolidate skills and knowledge acquired during web-based training. This could include incorporating multiple repetitions of the learning material throughout the training [8]. While repetition was mentioned to maximize learning, participants in the studies by Kimbrel et al [37] and Kreuze et al [10] indicated that repetition was irritating, distracting, and even useless. Other recommendations included follow-up training sessions spaced in time to reinforce and sustain knowledge [31,42,51], training refresher sessions [15,21,22,37,42,51,53,56,69]—although Kingi-Ulu'ave et al [38] concluded that passive boosters were not impacting knowledge retention and trainees' self-efficacy—and complementary material like digital workbooks to write notes, take-home training summaries, and complementary resources [10,22,34,35,41,51].

Overall, Kreuze et al [10] recommended having a variety of teaching and evaluating strategies to address diverse learning styles and needs, promote critical thinking, and incorporate learning across the cognitive, affective, and psychomotor domains (eg, through role-plays).

An integrated list of dos and don'ts, organized from the results, is provided as a one-page checklist in Multimedia Appendix 4 [8-10,14,15,19-54,56-71,74].

Discussion

Principal Findings

This narrative review systematically synthesized findings drawn from published literature focused on web-based suicide prevention GTPs to make recommendations for key design, content, and delivery. While there is increasing enthusiasm for web-based education broadly [75,76], including in the field of suicide prevention [9,77], there is a lack of clear guidance on best practices for the development and implementation of these programs [14]. The findings of this review contribute to addressing this knowledge gap by detailing critical elements across 3 key areas: design, content, and pedagogy. These insights offer a conceptual foundation for future research and offer practical guidance for the development and implementation of effective web-based gatekeeper training initiatives.

A central finding of this review is the necessity of grounding program design in evidence and theory, both in terms of the GTP content and its pedagogy. Integrating evidence-based practices in its content not only enhances the credibility of gatekeeper training but is also essential for ensuring its effectiveness. Yet, this is not without challenges. Evidence-based suicide prevention initiatives remain in early stages of translation [78], and a key critique of GTPs is the frequent lack of rigorous evaluation that can demonstrate their effectiveness [79]. In terms of pedagogy, most programs to date draw from social cognitive theory [80] and the theory of planned behavior [81], which highlight psychosocial determinants of behavior change and draw attention to contextual factors that influence how new behaviors are learned and sustained. However, these theories focus on individual behaviors and may fall short of accounting for the specific affordances and constraints of digital learning environments [82]. The inclusion

XSL•FO
RenderX

of instructional design theories tailored to web-based modalities—such as the community of inquiry framework for web-based learning [83] and the cognitive theory of multimedia learning [84]—could significantly strengthen the pedagogical foundation of these programs, taking full advantage of web-based platforms.

Moreover, research and existing guidelines suggest that effective gatekeeper training should be designed and situated within a broader, multilevel suicide prevention strategy [1]. This aligns with recommendations from the WHO [1] and multiple reviews of national suicide prevention strategies [85-87], which recognize the potential of GTPs and their limitations as a stand-alone intervention (Zalsman et al [79]). For example, GTPs are not meant to replace suicide specific interventions delivered by trained mental health professionals. In addition, the effectiveness of GTPs may be limited in environments where suicide is highly stigmatized [88]. Indeed, stigma has consistently been identified as a key factor contributing to gatekeepers' reluctance to intervene [4]. It can also play a complicit role in silencing individuals with suicidal thoughts [12,89]. The threat of stigma is especially relevant for marginalized populations, including LGBTQ+ and racialized populations, who already experience significant barriers to safe and appropriate mental health care [90,91]. Moreover, while these programs can help connect individuals experiencing suicidal thoughts to appropriate support, there is little evidence on how they foster hope or promote a meaningful life [12,92]. In addition to building core gatekeeper competencies—identifying warning signs, talking about suicide, referring—future GTPs should incorporate upstream approaches that actively promote mental health well-being and reduce stigma.

To better tailor GTPs to current needs, several articles recommended co-design approaches in the development of web-based GTPs. Co-design approaches have been described in the health literature as an effective approach to enhance program relevance and user engagement by involving end users in the development process [93,94]. However, as identified by Qasim et al [95], important knowledge gaps remain on how to best mobilize co-design approaches in maximizing outcomes. A useful starting point to optimize the inclusion of end users' insights and needs in the co-design process could be to draw from participatory and community-based research principles [96]. Despite the espoused widespread use of participatory and community-based methods with vulnerable populations, there is still very little consensus on best practices [97,98]. Nevertheless, emerging key components such as fostering collaborative spaces [99,100], building capacities [98,101], and the balancing of power [102-104] could inform co-design practices of future GTPs.

In terms of content and pedagogy, this review reinforces the need for flexibility and adaptability to accommodate the varied needs and learning styles of trainees, aligning with broader findings on adult education and training programs [105]. Customizability—including tailoring scenarios, examples, and resources to specific user groups—has been shown to enhance engagement and practical application in education [106]. Dreier et al [16] suggest that customization of the training should also include the "desired degree of confrontation"—that is, text, images, videos, testimonials, representation of suicidality—not only to improve the engagement and satisfaction of the learners, but also to provide a sense of agency regarding the possible emotional distress in reaction to sensitive content. Interactive elements like role-plays, simulations, and real-time feedback are frequently cited as best practices in the education literature and are particularly effective in improving skill retention and building confidence [107-109]. While the adaptability and flexibility of web-based training have been extensively acknowledged as a strength in logistical terms (eg, pace and access) and range of possible content [110], they have often been criticized for their lack of interpersonal interactions, a key component to enhancing learning outcomes [111]. Among the reviewed articles, incorporating ways to foster a sense of connection and community among trainees in a web-based format (eg, creating or mobilizing already existing peer support networks and collaborative learning opportunities) has been emphasized and is consistent with research on social learning theory [112].

As a possible alternative to actual interpersonal interaction through the building of peer networks for collaborative learning, some authors have suggested the use of avatars or models—whose machine learning algorithm was not specified in the current reviewed articles—that would provide in vivo feedback in response to the trainees' interactions. Such use of recent technologies provides an interesting avenue for interactive learning when actual human interaction is not possible. Although technologies based on artificial intelligence (AI) have been adopted in education—whether through students' initiative or within the actual teaching curriculum—the absence of guidelines and best practices limits their adoption. In their systematic review of the use of AI in higher education, Ouyang et al [113] discussed the potential of AI technologies for improving the learner's experiences in terms of engagement and providing accurate predictions (ie, real-time feedback). Yet, they further claim that the mere use of AI technologies does not necessarily lead to positive educational outcomes, emphasizing the importance of integrative frameworks or theories to support and enhance their ethical use. Another key ethical consideration in the use of AI technologies and open-source machine learning models is related to issues of confidentiality and privacy [114,115]. Such precautions become even more important in the context of vulnerable and sensitive topics such as suicide. Concerns about the privacy of information may also limit trainees in their willingness to participate in real-time AI-based role-play or feedback sessions. To mitigate risks and barriers to engagement, GTPs should explicitly disclose their use of such technologies, ensure secure data storage (eg, encrypted servers and restricted access), and implement clear protocols for handling data from role-play or training recordings [116].

Key barriers to the successful implementation of web-based GTPs for suicide prevention were identified in this review, many of which are inherent to web-based education in general. In addition to the lack of interpersonal and customizable human interactions, accessibility—while a strength—remains a significant challenge in terms of limited internet bandwidth, lack of access to the necessary technology, and varying levels

XSL•FO
RenderX

of digital literacy impeding program reach and effectiveness. While the digital divide is narrowing globally [117], it continues to be a barrier for some populations due to persisting structural inequities flowing to and from the social determinants of health, including income, geographic location (ie, underdeveloped areas), and educational disparities [117,118]. Further, technical difficulties, including platform glitches and inefficiencies, not only hinder user experience and engagement but also the retention of learnings [119].

Sustainability is a critical challenge for web-based suicide GTPs. Without dedicated long-term funding, government, or commercial support, updates to keep content relevant and evidence-based are not possible. In fact, some of the interventions reviewed in this study could not be accessed at the time of writing, suggesting potential challenges to sustainability or ongoing availability. Ensuring long-term success requires more than just the initial development and implementation of an intervention. Sustainability should be taken into consideration in the initial steps of program development, given that it demands ongoing engagement, adaptation to evolving needs, and seamless integration into broader suicide prevention frameworks. The successful implementation of GTPs and their sustainability could benefit from network interventions at the public health level that focus attention on forging and solidifying intersectoral partnerships towards a more distributive model of responsibilities and accountability for suicide prevention [120].

Future Directions for Research on Web-Based GTPs

Multiple calls have been made for more robust research into the evaluation of GTPs [4,79,121], and while we agree with this need, we also emphasize that understanding the most effective ways to deliver GTPs in web-based formats is important. Future studies should explore questions such as: “What are the most effective web-based approaches for training gatekeepers?” “What specific features of web-based training (eg, interactive modules, live simulations, peer discussions) can enhance skill acquisition and retention?” “How does web-based GTPs impact participants’ ability to recognize and respond to suicidal crises compared to in-person training?” Additionally, process evaluations and user feedback are essential to assess the quality of training delivery, user engagement, and practical application of skills learned via the web. Such evaluations can help to better understand implementation barriers and facilitators to improve the overall design of web-based GTPs. Another aspect to consider in the evaluation and design of GTPs is the impact and best delivery options of refreshers and booster sessions. While there is growing attention in tailoring GTPs to the needs of at-risk populations (eg, indigenous youth, migrant youth, gamers, and veteran as seen in this review), future research should also explore how web-based formats can better reach underrepresented and marginalized populations—such as LGBTQ+, Black, Indigenous and Racialized communities, or geographically isolated communities—more effectively and assess the training’s scalability and sustainability. Mixed methods studies and comparative trials are critical to advancing knowledge about how web-based GTPs can contribute to comprehensive suicide prevention strategies.

Limitations

This review is limited by its focus on peer-reviewed articles published in English and French, potentially excluding relevant studies in other languages or formats (gray literature), which may introduce selection bias. The heterogeneity of included studies, such as variations in design, target populations, and program settings, made it difficult to draw generalizable conclusions. Many studies lacked rigorous evaluation methods, limiting our understanding of the impacts of the GTPs included. Additionally, insufficient reporting on user feedback, contextual factors (such as cultural and technological differences), and the implementation process restricts the applicability of findings across diverse settings. Notably, none of the studies reviewed specifically addressed our research question, which is, “what works” and “does not work” in web-based GTP. This gap underscores the need for further research that directly explores effective strategies to deliver web-based suicide prevention training to provide definitive guidance and improve training outcomes.

Conclusions

Web-based GTPs hold significant promise for suicide prevention, yet much of the existing research has focused on determining whether these programs are effective [7], with evidence inconclusive as stand-alone interventions [88,122,123]. To advance the field, it is crucial to enhance evaluation efforts [123] while simultaneously exploring what works best in web-based formats for building gatekeeper skills. While there is a call for randomized controlled trials to build evidence of suicide prevention interventions such as GTPs to prove their efficacy, the next logical step in the evaluation and development of effective web-based GTPs should focus attention on the process of their implementation, including acceptability and feasibility. This review highlights a wide range of key considerations regarding web-based pedagogy that require closer attention to determine which might represent best practices to implement in the development of future GTPs. This involves a commitment to continuous improvement by leveraging technological advancements to enhance accessibility, engagement, and adaptability for diverse populations. By aligning program design with the latest innovations, including emerging tools such as AI [124,125], and rigorously evaluating their impact, web-based GTPs can evolve into a vital component of comprehensive suicide prevention strategies. Ultimately, these efforts will not only improve the quality and reach of interventions but also have the power to save lives.

Acknowledgments

The authors would like to thank Sylvie Fontaine and Travis Salway for their valuable assistance with the development of the search strategy. Generative artificial intelligence assistance (ChatGPT, OpenAI) was used solely to proofread and improve the grammar and clarity of text drafted entirely by the authors. The tool was not used to generate content, analyze or interpret data,

XSL•FO
RenderX

or produce tables or figures. All artificial intelligence–assisted edits were reviewed and verified by the authors, who remain fully responsible for the scientific integrity of the manuscript.

Funding

This work was supported by the Canadian Institutes of Health Research (CIHR-grant number: 183605). OF is supported by a salary award from the Fonds de Recherche du Québec.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during this study.

Authors' Contributions

OF contributed to the conceptualization of the study and was responsible for writing the original draft, as well as reviewing and editing the manuscript. OF also provided resources, acquired funding, and supervised the project. EG contributed to the investigation, methodology, data curation, and formal analysis, and participated in writing the original draft and reviewing and editing the manuscript. KL contributed to writing the original draft and to reviewing and editing the manuscript. KW contributed to formal analysis, software development, validation, and reviewing and editing the manuscript. JLO, HK, and BD each contributed to reviewing and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA checklist.
[PDF File (Adobe PDF File), 322 KB - jmir_v28i1e81572_app1.pdf]

Multimedia Appendix 2
Search strategy.
[DOCX File , 34 KB - jmir_v28i1e81572_app2.docx]

Multimedia Appendix 3
Complete table of characteristics of the online gatekeeper training programs included in the systematic narrative review.
[DOCX File , 90 KB - jmir_v28i1e81572_app3.docx]

Multimedia Appendix 4
Online gatekeeper training programs (GTP): Do’s and Don’ts checklist.
[DOCX File , 18 KB - jmir_v28i1e81572_app4.docx]

References

1. Preventing suicide: a global imperative. World Health Organization. 2014. URL: https://www.who.int/publications/i/item/9789241564779 [accessed 2025-12-13]
2. Suicide worldwide in 2019: global health estimates. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240026643 [accessed 2025-12-13]
3. Caine ED, Reed J, Hindman J, Quinlan K. Comprehensive, integrated approaches to suicide prevention: practical guidance. Inj Prev 2018 Jun;24(Suppl 1):i38-i45. [doi: 10.1136/injuryprev-2017-042366] [Medline: 29263088]
4. Burnette C, Ramchand R, Ayer L. Gatekeeper Training for Suicide Prevention: A Theoretical Model and Review of the Empirical Literature. Rand Health Q 2015 Jul 15;5(1):16 [FREE Full text] [Medline: 28083369]
5. Isaac M, Elias B, Katz LY, Belik S, Deane FP, Enns MW, Swampy Cree Suicide Prevention Team. Gatekeeper training as a preventative intervention for suicide: a systematic review. Can J Psychiatry 2009 Apr;54(4):260-268. [doi: 10.1177/070674370905400407] [Medline: 19321032]
6. Public health action for the prevention of suicide: a framework. World Health Organization. 2012 Jan 01. URL: https://www.who.int/publications/i/item/9789241503570 [accessed 2025-12-13]
7. Kingi-Uluave D, Taufa N, Tuesday R, Cargo T, Stasiak K, Merry S, et al. A review of systematic reviews: gatekeeper training for suicide prevention with a focus on effectiveness and findings. Arch Suicide Res 2025;29(2):329-346 [FREE Full text] [doi: 10.1080/13811118.2024.2358411] [Medline: 38884349]
8. Ghoncheh R, Koot HM, Kerkhof AJFM. Suicide prevention e-learning modules designed for gatekeepers: a descriptive review. Crisis 2014;35(3):176-185. [doi: 10.1027/0227-5910/a000249] [Medline: 24901058]

XSL•FO
RenderX

9. Lancaster PG, Moore JT, Putter SE, Chen PY, Cigularov KP, Baker A, et al. Feasibility of a web-based gatekeeper training: implications for suicide prevention. Suicide Life Threat Behav 2014;44(5):510-523. [doi: 10.1111/sltb.12086] [Medline: 24571612]
10. Kreuze E, York J, Lamis DA, Jenkins C, Quinnett P, Mueller M, et al. Gatekeeper training for youth suicide prevention: a mixed method comparative analysis of two online programs. Psychol Sch 2024;62(2):492-511. [doi: 10.1002/pits.23335]
11. Torok M, Calear AL, Smart A, Nicolopoulos A, Wong Q. Preventing adolescent suicide: a systematic review of the effectiveness and change mechanisms of suicide prevention gatekeeping training programs for teachers and parents. J Adolesc 2019;73:100-112. [doi: 10.1016/j.adolescence.2019.04.005] [Medline: 31054373]
12. Baril A. Undoing Suicidism: A Trans, Queer, Crip Approach to Rethinking (Assisted) Suicide. Philadelphia, Pennsylvania: Temple University Press; 2023.
13. Ferlatte O, Salway T, Oliffe JL, Kia H, Rice S, Morgan J, et al. Sexual and gender minorities' readiness and interest in supporting peers experiencing suicide-related behaviors. Crisis 2020;41(4):273-279. [doi: 10.1027/0227-5910/a000632] [Medline: 31657638]
14. Holmes G, Clacy A, Hamilton A, Kõlves K. Online versus in-person gatekeeper suicide prevention training: comparison in a community sample. J Ment Health 2024;33(5):605-612 [FREE Full text] [doi: 10.1080/09638237.2024.2332811] [Medline: 38602188]
15. Ghoncheh R, Gould MS, Twisk JW, Kerkhof AJ, Koot HM. Efficacy of adolescent suicide prevention E-learning modules for gatekeepers: a randomized controlled trial. JMIR Ment Health 2016;3(1):e8 [FREE Full text] [doi: 10.2196/mental.4614] [Medline: 26825006]
16. Dreier M, Ludwig J, Härter M, von dem Knesebeck O, Rezvani F, Baumgardt J, et al. Evaluation of an online suicide prevention program to improve suicide literacy and to reduce suicide stigma: a mixed methods study. PLoS One 2023;18(4):e0284944 [FREE Full text] [doi: 10.1371/journal.pone.0284944] [Medline: 37115766]
17. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 2015;4(1):1 [FREE Full text] [doi: 10.1186/2046-4053-4-1] [Medline: 25554246]
18. Holmes G, Clacy A, Hermens DF, Lagopoulos J. The long-term efficacy of suicide prevention gatekeeper training: a systematic review. Arch Suicide Res 2021;25(2):177-207. [doi: 10.1080/13811118.2019.1690608] [Medline: 31809659]
19. Afsharnejad B, Milbourn B, Hayden-Evans M, Baker-Young E, Black MH, Thompson C, et al. The efficacy of the "Talk-to-Me" suicide prevention and mental health education program for tertiary students: a crossover randomised control trial. Eur Child Adolesc Psychiatry 2023;32(12):2477-2489 [FREE Full text] [doi: 10.1007/s00787-022-02094-4] [Medline: 36194311]
20. Albritton T, Ford KL, Elsbernd K, Santodomingo M, Juzang I, Weddington P, et al. Implementing a peer advocate mental health digital intervention program for Ohio youth: descriptive pilot study. JMIR Ment Health 2021;8(4):e24605 [FREE Full text] [doi: 10.2196/24605] [Medline: 33890868]
21. Bartgis J, Albright G. Online role-play simulations with emotionally responsive avatars for the early detection of native youth psychological distress, including depression and suicidal ideation. Am Indian Alsk Native Ment Health Res 2016;23(2):1-27. [doi: 10.5820/aian.2302.2016.1] [Medline: 27115130]
22. Brown K, Toombs M, Nasir B, Kisely S, Ranmuthugala G, Brennan-Olsen SL, et al. How can mobile applications support suicide prevention gatekeepers in Australian indigenous communities? Soc Sci Med 2020;258:113015. [doi: 10.1016/j.socscimed.2020.113015] [Medline: 32559573]
23. Bryant R, Vogt M, Miller C. Effects of an avatar-based simulation on family nurse practitioner students' self-evaluated suicide prevention knowledge and confidence. Nurs Educ Perspect 2025;46(3):179-181. [doi: 10.1097/01.NEP.0000000000001321] [Medline: 39167360]
24. Canady VA. National, local efforts train educators on suicide prevention, MH support. Mental Health Weekly 2015 Jun;25(22):1-7. [doi: 10.1002/mhw.30206]
25. Carpenter DM, Roberts CA, Lavigne JE, Cross WF. Gatekeeper training needs of community pharmacy staff. Suicide Life Threat Behav 2021;51(2):220-228. [doi: 10.1111/sltb.12697] [Medline: 33876495]
26. Caughlan C, Kakuska A, Manthei J, Galvin L, Martinez A, Kelley A, et al. Mind4Health: decolonizing gatekeeper trainings using a culturally relevant text message intervention. Front Public Health 2024;12:1397640 [FREE Full text] [doi: 10.3389/fpubh.2024.1397640] [Medline: 39286750]
27. Cohen E, Pomerance Y, Touati Ohayon L, Brunstein Klomek A. Efficacy of suicide prevention gatekeeper training in Israel: exploring diverse at-risk populations, gender differences, and comparisons between online and in-person training. Death Stud 2025:1-9 [FREE Full text] [doi: 10.1080/07481187.2025.2510482] [Medline: 40442900]
28. Colder Carras M, Bergendahl M, Labrique AB. Community case study: stack up's overwatch program, an online suicide prevention and peer support program for video gamers. Front Psychol 2021;12:575224 [FREE Full text] [doi: 10.3389/fpsyg.2021.575224] [Medline: 33776826]
29. Coleman D, Black N, Ng J, Blumenthal E. Kognito's avatar-based suicide prevention training for college students: results of a randomized controlled trial and a naturalistic evaluation. Suicide Life Threat Behav 2019;49(6):1735-1745. [doi: 10.1111/sltb.12550] [Medline: 30957909]

XSL•FO
RenderX

30. Colucci E, Jaroudy S, Rossmann M. Piloting of a suicide first aid gatekeeper training (online) for children and young people in conflict affected areas in Syria. Int Rev Psychiatry 2022;34(6):640-648 [FREE Full text] [doi: 10.1080/09540261.2022.2100245] [Medline: 36695207]
31. Hawley SR, Skinner T, Young M, St Romain T, Provines J. Suicide prevention across the community: evaluation of mental health training for multiple gatekeeper groups. Kans J Med 2024;17(6):127-132. [doi: 10.17161/kjm.vol17.22524] [Medline: 39758535]
32. Hill RM, McCray CL. Suicide-related stigma and social responsibility moderate the effects of an online suicide prevention gatekeeper training program. Arch Suicide Res 2024;28(2):706-715. [doi: 10.1080/13811118.2023.2199802] [Medline: 37209132]
33. Hill RM, Picou P, Hussain Z, Vieyra BA, Perkins KM. Randomized Controlled Trial of an Online Suicide Prevention Gatekeeper Training Program. Göttingen, Germany: Hogrefe Publishing; 2024:57-64.
34. Hofmann L, Glaesmer H, Przyrembel M, Wagner B. An evaluation of a suicide prevention e-learning program for police officers (COPS): improvement in knowledge and competence. Front Psychiatry 2021;12:770277 [FREE Full text] [doi: 10.3389/fpsyt.2021.770277] [Medline: 34966307]
35. Hofmann L, Wagner B. Efficacy of an online gatekeeper program for relatives of men at risk of suicide - a randomized controlled trial. BMC Public Health 2024;24(1):2693 [FREE Full text] [doi: 10.1186/s12889-024-20193-6] [Medline: 39358752]
36. Kawashima D, Koga Y, Yoshioka M. Feasibility of brief online gatekeeper training for Japanese university students: a randomized controlled trial. Death Stud 2023;47(5):531-540. [doi: 10.1080/07481187.2022.2101076] [Medline: 35855580]
37. Kimbrel NA, Aho NA, Neal LC, Bernes SA, Beaver TA, Hertzberg JA, et al. Development and implementation of web-based safety planning intervention training for firefighter peer support specialists. Crisis 2024;45(2):108-117. [doi: 10.1027/0227-5910/a000924] [Medline: 37727969]
38. Kingi-Ulu'ave D, Framptom C, Cargo T, Stasiak K, Hetrick S. Evaluating the effectiveness of a post-training enhancement to the lifekeepers suicide prevention gatekeeper training. Crisis 2025;46(3):157-165. [doi: 10.1027/0227-5910/a001001] [Medline: 40235266]
39. Kreuze E, Ruggiero KJ. Technology-oriented suicide prevention interventions for adolescents and adolescent gatekeepers: a qualitative review. Adolescent Res Rev 2017;3(2):219-233. [doi: 10.1007/s40894-017-0060-5]
40. Lamis DA, Underwood M, D'Amore N. Outcomes of a suicide prevention gatekeeper training program among school personnel. Crisis 2017;38(2):89-99. [doi: 10.1027/0227-5910/a000414] [Medline: 27561223]
41. Lee-Tauler SY, Grammer J, LaCroix JM, Walsh AK, Clark SE, Holloway KJ, et al. Pilot evaluation of the online 'Chaplains-CARE' program: enhancing skills for united states military suicide intervention practices and care. J Relig Health 2023;62(6):3856-3873. [doi: 10.1007/s10943-023-01882-9] [Medline: 37612485]
42. Liu H, Zheng C, Cao Y, Zeng F, Chen H, Gao W. Gatekeeper training for suicide prevention: a systematic review and meta-analysis of randomized controlled trials. BMC Public Health 2025;25(1):1206 [FREE Full text] [doi: 10.1186/s12889-025-21736-1] [Medline: 40165179]
43. Hart SM, Colucci E, Marzano L. Evaluating suicide prevention gatekeeper training designed to identify and support people from asylum-seeking and refugee backgrounds. BMC Public Health 2024;24(1):2959 [FREE Full text] [doi: 10.1186/s12889-024-20304-3] [Medline: 39455999]
44. Manning J, Vandeusen K. Suicide prevention in the dot com era: technological aspects of a university suicide prevention program. J Am Coll Health 2011;59(5):431-433. [doi: 10.1080/07448480903540507] [Medline: 21500064]
45. Marley G, Lavigne JE, Cross W, Gamble A, Zhang Z, Carpenter DM. Comparing three methods to assess learning outcomes for a suicide prevention training program for pharmacy staff. PEC Innov 2024;5:100348 [FREE Full text] [doi: 10.1016/j.pecinn.2024.100348] [Medline: 39444544]
46. McKay S, Byrne S, Clarke A, Lamblin M, Veresova M, Robinson J. Parent education for responding to and supporting youth with suicidal thoughts (PERSYST): An evaluation of an online gatekeeper training program with Australian parents. Int J Environ Res Public Health 2022;19(9):5025 [FREE Full text] [doi: 10.3390/ijerph19095025] [Medline: 35564419]
47. Mirick RG. Acceptability and feasibility of a brief online suicide prevention training for school staff. J Technol Hum Serv 2025;43(1):34-48. [doi: 10.1080/15228835.2024.2447692]
48. Mishkind MC, Yannacone A, Lopez A, Jortberg BT, Sherrill A, Mescher T. Virtual versus in-person suicide prevention training in the workplace: evaluation of the vitalcog program. J Technol Behav Sci 2023:1-8 [FREE Full text] [doi: 10.1007/s41347-023-00301-w] [Medline: 36742417]
49. Osteen PJ, Ohme K, Morris R, Arciniegas J, Frey JJ, Woods M, et al. Suicide intervention training with law enforcement officers. Suicide Life Threat Behav 2021;51(4):785-794. [doi: 10.1111/sltb.12763] [Medline: 33998030]
50. Perepezko K, Bergendahl M, Kunz C, Labrique A, Carras M, Colder Carras M. "Instead, You're Going to a Friend": evaluation of a community-developed, peer-delivered online crisis prevention intervention. Psychiatr Serv 2024;75(12):1267-1275. [doi: 10.1176/appi.ps.20230233] [Medline: 39054853]
51. Pilbrow S, Staniland L, Uren HV, Shand F, McGoldrick J, Thorp E, et al. Evaluation of an online advanced suicide prevention training for pharmacists. Int J Clin Pharm 2023;45(5):1203-1211 [FREE Full text] [doi: 10.1007/s11096-023-01636-3] [Medline: 37702959]

XSL•FO
RenderX

52. Poštuvan V, Gomboc V, Čopič Pucihar K, Kljun M, Vičič J, Tančič Grum A, et al. Development and evaluation of online suicide preventive tool ialive to increase competences in engaging with a suicidal person. Crisis 2024;45(3):187-196. [doi: 10.1027/0227-5910/a000934] [Medline: 38140805]
53. Quinnett PG. The certified QPR pathfinder training program: A description of a novel public health gatekeeper training program to mitigate suicidal ideation and suicide deaths. J Prev (2022) 2023;44(6):813-824 [FREE Full text] [doi: 10.1007/s10935-023-00748-w] [Medline: 37740846]
54. Reifegerste D, Wagner AJM, Huber L, Fastuca M. Formative evaluation of suicide prevention websites for men: qualitative study with men at risk of suicide and with potential gatekeepers. JMIR Form Res 2025;9:e59829 [FREE Full text] [doi: 10.2196/59829] [Medline: 40009838]
55. Rein BA, McNeil DW, Hayes AR, Hawkins TA, Ng HM, Yura CA. Evaluation of an avatar-based training program to promote suicide prevention awareness in a college setting. J Am Coll Health 2018;66(5):401-411. [doi: 10.1080/07448481.2018.1432626] [Medline: 29461940]
56. Robinson-Link N, Hoover S, Bernstein L, Lever N, Maton K, Wilcox H. Is gatekeeper training enough for suicide prevention? School Mental Health 2019;12(2):239-249. [doi: 10.1007/s12310-019-09345-x]
57. Roslan AF, Pheh KS, Mahadevan R, Bujang SM, Subramaniam P, Yahya HF, et al. Effectiveness of online advanced C.A.R.E suicide prevention gatekeeper training program among healthcare lecturers and workers in national university of Malaysia: a pilot study. Front Psychiatry 2023;14:1009754 [FREE Full text] [doi: 10.3389/fpsyt.2023.1009754] [Medline: 36741120]
58. Ross SG, Pazienza R, Rosa JD. The suicide prevention for college student (SPCS) gatekeepers program. Crisis 2024;45(1):41-47. [doi: 10.1027/0227-5910/a000914] [Medline: 37322902]
59. Ross SG, Pazienza R, Rosa JD. The suicide prevention for college students (SPCS) gatekeepers program: comparing in-person and online training outcomes. J Am Coll Health 2025;73(9):3322-3325. [doi: 10.1080/07448481.2024.2423237] [Medline: 39514812]
60. Schmeckenbecher J, Lentner S, Emilian CA, Plener PL, Baran A, Kapusta ND. E-learning as a tool of suicide prevention training: a meta-analysis and systematic review. Death Stud 2024;48(9):962-974. [doi: 10.1080/07481187.2023.2297058] [Medline: 38133538]
61. Seabury B. On-line, computer-based, interactive simulations: bridging classroom and field. J Technol Hum Serv 2003;22(1):29-48. [doi: 10.1300/j017v22n01_04]
62. Seabury BA. An evaluation of on-line, interactive tutorials designed to teach practice concepts. J Teach Soc Work 2005;25(1-2):103-115. [doi: 10.1300/j067v25n01_07]
63. Shanta Bridges L, Sharma M, Lee JHSH, Bennett R, Buxbaum SG, Reese-Smith J. Using the PRECEDE-PROCEED model for an online peer-to-peer suicide prevention and awareness for depression (SPAD) intervention among African American college students: experimental study. Health Promot Perspect 2018;8(1):15-24 [FREE Full text] [doi: 10.15171/hpp.2018.02] [Medline: 29423358]
64. Smith-Millman M, Bernstein L, Link N, Hoover S, Lever N. Effectiveness of an online suicide prevention program for college faculty and students. J Am Coll Health 2022;70(5):1457-1464. [doi: 10.1080/07448481.2020.1804389] [Medline: 32813627]
65. Stone DM, Barber CW, Potter L. Public health training online: the national center for suicide prevention training. Am J Prev Med 2005;29(5 Suppl 2):247-251. [doi: 10.1016/j.amepre.2005.08.019] [Medline: 16376726]
66. Stover AN, Lavigne JE, Shook A, MacAllister C, Cross WF, Carpenter DM. Development of the pharm-SAVES educational module for gatekeeper suicide prevention training for community pharmacy staff. Health Expect 2023;26(3):1246-1254 [FREE Full text] [doi: 10.1111/hex.13741] [Medline: 36852881]
67. Sun Y, Zhang Q, Wu W, Lin J, Sun S, An J, et al. Efficacy of a localized caregiver gatekeeper training program for suicide prevention among Chinese adolescents: a pilot study. Asian J Psychiatr 2025;109:104555. [doi: 10.1016/j.ajp.2025.104555] [Medline: 40449414]
68. Teo AR, Call AA, Hooker ER, Fong C, Karras E, Dobscha SK. Feasibility of recruitment and retention in a remote trial of gatekeeper training for close supports of military veterans: mixed methods study. Contemp Clin Trials Commun 2022;30:100993 [FREE Full text] [doi: 10.1016/j.conctc.2022.100993] [Medline: 36159001]
69. Teo AR, Hooker ER, Call AA, Dobscha SK, Gamble S, Cross WF, et al. Brief video training for suicide prevention in veterans: a randomized controlled trial of VA S.A.V.E. Suicide Life Threat Behav 2024;54(1):154-166 [FREE Full text] [doi: 10.1111/sltb.13028] [Medline: 38095049]
70. Timmons-Mitchell J, Albright G, McMillan J, Shockley K, Cho S. Virtual role-play: middle school educators addressing student mental health. Health Behav Policy Rev 2019;6(6):546-557. [doi: 10.14485/hbpr.6.6.1]
71. Wislocki K, Jager-Hyman S, Brady M, Weiss M, Schaechter T, Khazanov G, et al. Freely available training videos for suicide prevention: scoping review. JMIR Ment Health 2023;10:e48404 [FREE Full text] [doi: 10.2196/48404] [Medline: 37921847]
72. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
73. Mikkonen K, Kääriäinen M. Content Analysis in Systematic Reviews. Cham: Springer; 2020:105-115.

XSL•FO
RenderX

74. Kingi-Uluave D, Taufa N, Tuesday R, Cargo T, Stasiak K, Merry S, et al. A review of systematic reviews: gatekeeper training for suicide prevention with a focus on effectiveness and findings. Arch Suicide Res 2025;29(2):329-346 [FREE Full text] [doi: 10.1080/13811118.2024.2358411] [Medline: 38884349]
75. Martin F, Sun T, Westine CD. A systematic review of research on online teaching and learning from 2009 to 2018. Comput Educ 2020;159:104009 [FREE Full text] [doi: 10.1016/j.compedu.2020.104009] [Medline: 32921895]
76. Greenhow C, Graham CR, Koehler MJ. Foundations of online learning: challenges and opportunities. Educ Psychol 2022;57(3):131-147. [doi: 10.1080/00461520.2022.2090364]
77. Kingi-Ulu'ave D, Framptom C, Cargo T, Stasiak K, Hetrick S. Evaluating the impact and cultural relevance of lifekeepers gatekeeper training across three training modalities. Crisis 2024;45(6):417-424. [doi: 10.1027/0227-5910/a000977] [Medline: 39545404]
78. Melhem N, Moutier CY, Brent DA. Implementing evidence-based suicide prevention strategies for greatest impact. Focus (Am Psychiatr Publ) 2023;21(2):117-128 [FREE Full text] [doi: 10.1176/appi.focus.20220078] [Medline: 37201145]
79. Zalsman G, Hawton K, Wasserman D, van Heeringen K, Arensman E, Sarchiapone M, et al. Suicide prevention strategies revisited: 10-year systematic review. Lancet Psychiatry 2016;3(7):646-659. [doi: 10.1016/S2215-0366(16)30030-X] [Medline: 27289303]
80. Bandura A. Social cognitive theory: an agentic perspective. Annu Rev Psychol 2001;52:1-26. [doi: 10.1146/annurev.psych.52.1.1] [Medline: 11148297]
81. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
82. Brieger E, Arghode V, McLean G. Connecting theory and practice: reviewing six learning theories to inform online instruction. Eur J Train Dev 2020;44(4/5):321-339. [doi: 10.1108/ejtd-07-2019-0116]
83. Shea P, Richardson J, Swan K. Building bridges to advance the community of inquiry framework for online learning. Educ Psychol 2022;57(3):148-161. [doi: 10.1080/00461520.2022.2089989]
84. Mayer RE. The past, present, and future of the cognitive theory of multimedia learning. Educ Psychol Rev 2024;36(1):8. [doi: 10.1007/s10648-023-09842-1]
85. van der Feltz-Cornelis CM, Sarchiapone M, Postuvan V, Volker D, Roskar S, Grum AT, et al. Best practice elements of multilevel suicide prevention strategies: a review of systematic reviews. Crisis 2011;32(6):319-333 [FREE Full text] [doi: 10.1027/0227-5910/a000109] [Medline: 21945840]
86. Beautrais A, Fergusson D, Coggan C, Collings C, Doughty C, Ellis P, et al. Effective strategies for suicide prevention in New Zealand: a review of the evidence. N Z Med J 2007;120(1251):U2459. [Medline: 17384687]
87. Mann JJ, Apter A, Bertolote J, Beautrais A, Currier D, Haas A, et al. Suicide prevention strategies: a systematic review. JAMA 2005;294(16):2064-2074. [doi: 10.1001/jama.294.16.2064] [Medline: 16249421]
88. Spafford SG, Silverman MM, Gutierrez PM. What is known about suicide prevention gatekeeper training and directions for future research. Suicide Life Threat Behav 2025;55(1):e13130. [doi: 10.1111/sltb.13130] [Medline: 39351789]
89. Creighton G, Oliffe J, Ogrodniczuk J, Frank B. "You've Gotta Be That Tough Crust Exterior Man": depression and suicide in rural-based men. Qual Health Res 2017;27(12):1882-1891. [doi: 10.1177/1049732317718148] [Medline: 28936927]
90. Farvid P, Vance TA, Klein SL, Nikiforova Y, Rubin LR, Lopez FG. The health and wellbeing of transgender and gender non-conforming people of colour in the United States: a systematic literature search and review. J Community Appl Soc Psychol 2021;31(6):703-731. [doi: 10.1002/casp.2555]
91. Clark KD, Lunn MR, Bosse JD, Sevelius JM, Dawson-Rose C, Weiss SJ, et al. Societal stigma and mistreatment in healthcare among gender minority people: a cross-sectional study. Int J Equity Health 2023;22(1):162 [FREE Full text] [doi: 10.1186/s12939-023-01975-7] [Medline: 37620832]
92. Wexler L, White J, Trainor B. Why an alternative to suicide prevention gatekeeper training is needed for rural indigenous communities: presenting an empowering community storytelling approach. Crit Public Health 2015;25(2):205-217 [FREE Full text] [doi: 10.1080/09581596.2014.904039] [Medline: 36779086]
93. Arnold D, Glässel A, Böttger T, Sarma N, Bethmann A, Narimani P. "What Do You Need? What Are You Experiencing?" relationship building and power dynamics in participatory research projects: critical self-reflections of researchers. Int J Environ Res Public Health 2022;19(15):9336 [FREE Full text] [doi: 10.3390/ijerph19159336] [Medline: 35954692]
94. Mulvale A, Miatello A, Hackett C, Mulvale G. Applying experience-based co-design with vulnerable populations: lessons from a systematic review of methods to involve patients, families and service providers in child and youth mental health service improvement. Patient Exp J 2016;3(1):117-129. [doi: 10.35680/2372-0247.1104]
95. Qasim R, Farooqui WA, Rahman A, Haroon R, Saleem M, Rafique M, et al. Community centred co-design methodology for designing and implementing socio-behavioural interventions to counter COVID-19 related misinformation among marginalized population living in the squatter settlements of Karachi, Pakistan: a methodology paper. BMC Proc 2023;17(Suppl 7):15 [FREE Full text] [doi: 10.1186/s12919-023-00265-y] [Medline: 37438805]
96. Kral MJ, Kidd S. Community-based participatory research and community empowerment for suicide prevention. In: A Positive Psychological Approach to Suicide : Theory, Research, and Prevention. Cham: Springer International Publishing; 2018:285-299.

XSL•FO
RenderX

97. Chudyk AM, Stoddard R, Duhamel TA, Patient Engagement in Research Partners, Schultz ASH. Future directions for patient engagement in research: a participatory workshop with Canadian patient partners and academic researchers. Health Res Policy Syst 2024;22(1):24 [FREE Full text] [doi: 10.1186/s12961-024-01106-w] [Medline: 38350974]
98. Thomas E, Benjamin-Thomas TE, Sithambaram A, Shankar J, Chen S. Participatory action research among people with serious mental illness: a scoping review. Qual Health Res 2024;34(1-2):3-19 [FREE Full text] [doi: 10.1177/10497323231208111] [Medline: 37929751]
99. Wallerstein N, Muhammad M, Sanchez-Youngman S, Rodriguez Espinosa P, Avila M, Baker EA, et al. Power dynamics in community-based participatory research: a multiple-case study analysis of partnering contexts, histories, and practices. Health Educ Behav 2019;46(1_suppl):19S-32S. [doi: 10.1177/1090198119852998] [Medline: 31549557]
100. Haarmans M, Nazroo J, Kapadia D, Maxwell C, Osahan S, Edant J, et al. The practice of participatory action research: complicity, power and prestige in dialogue with the 'racialised mad'. Sociol Health Illn 2022;44 Suppl 1(Suppl 1):106-123 [FREE Full text] [doi: 10.1111/1467-9566.13517] [Medline: 36001350]
101. Brush BL, Mentz G, Jensen M, Jacobs B, Saylor KM, Rowe Z, et al. Success in long-standing community-based participatory research (CBPR) partnerships: a scoping literature review. Health Educ Behav 2020;47(4):556-568 [FREE Full text] [doi: 10.1177/1090198119882989] [Medline: 31619072]
102. Park M, Lee K, Estein O. “That was sexy” to “hope for change”: resituating an ethics of power in the capacity to aspire in public and patient-involved research. In: Meaningful and Safe: The Ethics and Ethical Implications of Patient and Public Involvement in Health and Medical Research. Suffolk, UK: Ethics Press; 2024:89-119.
103. Popay J, Whitehead M, Ponsford R, Egan M, Mead R. Power, control, communities and health inequalities I: theories, concepts and analytical frameworks. Health Promot Int 2021;36(5):1253-1263 [FREE Full text] [doi: 10.1093/heapro/daaa133] [Medline: 33382890]
104. Rose D, Kalathil J. Power, privilege and knowledge: the untenable promise of co-production in mental "Health". Front Sociol 2019;4:57 [FREE Full text] [doi: 10.3389/fsoc.2019.00057] [Medline: 33869380]
105. Flexible adult learning provision: what it is, why it matters, and how to make it work. OECD. 2023. URL: https://www.oecd.org/content/dam/oecd/en/topics/policy-sub-issues/adult-learning/booklet-flexibility-2023.pdf [accessed 2025-12-13]
106. Aladar A, Ilg D, King J, Hajja A. EduKona: A customizable, mobile-friendly platform for enhanced educational engagement and collaboration. : IEEE; 2024 Presented at: IEEE Frontiers in Education Conference (FIE); 2024 October 13-16; Washington, DC, USA p. 13-16. [doi: 10.1109/fie61694.2024.10892814]
107. Chernikova O, Heitzmann N, Stadler M, Holzberger D, Seidel T, Fischer F. Simulation-based learning in higher education: a meta-analysis. Rev Educ Res 2020;90(4):499-541. [doi: 10.3102/0034654320933544]
108. Elendu C, Amaechi DC, Okatta AU, Amaechi EC, Elendu TC, Ezeh CP, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore) 2024;103(27):e38813 [FREE Full text] [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
109. Lavanya KM, Somu LK, Mishra SK. Effectiveness of scenario-based roleplay as a method of teaching soft skills for undergraduate medical students. Int J Appl Basic Med Res 2024;14(2):78-84 [FREE Full text] [doi: 10.4103/ijabmr.ijabmr_431_23] [Medline: 38912358]
110. Salloum SA, Salloum A, Alfaisal R, Basiouni A, Shaalan K. Predicting student adaptability to online education using machine learning. In: Breaking Barriers with Generative Intelligence Using GI to Improve Human Education and Well-Being. Cham: Springer Nature Switzerland; 2024.
111. Mojtahedzadeh R, Hasanvand S, Mohammadi A, Malmir S, Vatankhah M. Students' experience of interpersonal interactions quality in e-Learning: a qualitative research. PLoS One 2024;19(3):e0298079 [FREE Full text] [doi: 10.1371/journal.pone.0298079] [Medline: 38530814]
112. Hill JR, Song L, West RE. Social learning theory and web-based learning environments: a review of research and discussion of implications. Am J Distance Educ 2009;23(2):88-103. [doi: 10.1080/08923640902857713]
113. Ouyang F, Zheng L, Jiao P. Artificial intelligence in online higher education: a systematic review of empirical research from 2011 to 2020. Educ Inf Technol 2022;27(6):7893-7925. [doi: 10.1007/s10639-022-10925-9]
114. Nel JM. The The implications of artificial intelligence in online education: a critical examination in the context of philosophy and ethics. In: Towards a Hybrid, Flexible and Socially Engaged Higher Education. Cham: Springer Nature Switzerland; 2024.
115. Mougiakou E, Papadimitriou S, Chrysafiadi K, Virvou M. Artificial intelligence in educational games and consent under general data protection regulation. Int Decis Technol 2025;19(2):670-686. [doi: 10.1177/18724981251322884]
116. Holmes G, Tang B, Gupta S, Venkatesh S, Christensen H, Whitton A. Applications of large language models in the field of suicide prevention: scoping review. J Med Internet Res 2025;27:e63126 [FREE Full text] [doi: 10.2196/63126] [Medline: 39847414]
117. Buchholz BA, DeHart J, Moorman G. Digital citizenship during a global pandemic: moving beyond digital literacy. J Adolesc Adult Lit 2020;64(1):11-17 [FREE Full text] [doi: 10.1002/jaal.1076] [Medline: 32834710]
118. Withers EM. The Digital Divide and Health: Examining Digital Access as a Social Determinant of Health. Downtown Portland: Portland State University; 2021.

XSL•FO
RenderX

119. Sitzmann T, Ely K, Bell BS, Bauer KN. The effects of technical difficulties on learning and attrition during online training. J Exp Psychol Appl 2010;16(3):281-292. [doi: 10.1037/a0019968] [Medline: 20853987]
120. Heeren T, Ward C, Sewell D, Ashida S. Applying network analysis to assess the development and sustainability of multi-sector coalitions. PLoS One 2022;17(10):e0276114 [FREE Full text] [doi: 10.1371/journal.pone.0276114] [Medline: 36256640]
121. Hawgood J, Woodward A, Quinnett P, Leo DD. Gatekeeper Training and Minimum Standards Competency: Essentials for the Suicide Prevention Workforce. Newburyport, Massachusetts: Hogrefe Publishing; 2022:516-522.
122. Morton M, Wang S, Tse K, Chung C, Bergmans Y, Ceniti A, et al. Gatekeeper training for friends and family of individuals at risk of suicide: a systematic review. J Community Psychol 2021;49(6):1838-1871. [doi: 10.1002/jcop.22624] [Medline: 34125969]
123. Spafford SG, McWhirter Boisen MR, Tanner-Smith EE, Rodriguez G, Muruthi JR, Seeley JR. The effects of suicide prevention gatekeeper training on behavioral intention and intervention behavior: a systematic review and meta-analysis. Prev Sci 2024;25(6):978-988. [doi: 10.1007/s11121-024-01710-w] [Medline: 39023720]
124. Atmakuru A, Shahini A, Chakraborty S, Seoni S, Salvi M, Hafeez-Baig A, et al. Artificial intelligence-based suicide prevention and prediction: a systematic review (2019–2023). Inf Fusion 2025;114:102673. [doi: 10.1016/j.inffus.2024.102673]
125. Li X, Chen F, Ma L. Exploring the potential of artificial intelligence in adolescent suicide prevention: current applications, challenges, and future directions. Psychiatry 2024;87(1):7-20. [doi: 10.1080/00332747.2023.2291945] [Medline: 38227496]

Abbreviations

AI: artificial intelligence
GTP: gatekeeper training program
LGBTQ+: lesbian, gay, bisexual, trans, queer, and other sexual and gender minorities
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
S.A.V.E./SAVE: signs, ask, validate, encourage/expedite
WHO: World Health Organization

Edited by A Stone; submitted 31.Jul.2025; peer-reviewed by A Shivanna, G Holmes; comments to author 20.Aug.2025; revised version received 25.Nov.2025; accepted 25.Nov.2025; published 05.Feb.2026.

Please cite as:
Ferlatte O, Gareau E, Lee K, Wassef K, Oliffe JL, Kia H, Dumville B
Key Components and Barriers in Web-Based Suicide Prevention Gatekeeper Training: Systematic Narrative Review
J Med Internet Res 2026;28:e81572
URL: https://www.jmir.org/2026/1/e81572
doi:10.2196/81572
PMID:

©Olivier Ferlatte, Emmanuelle Gareau, Keven Lee, Kinda Wassef, John Lindsay Oliffe, Hannah Kia, Brock Dumville. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Interventions That Use Highly Visual Social Media Platforms to Tackle Unhealthy Body Image in Adolescents and Young Adults: Systematic Review of Randomized Controlled Trials and Quasi-Experimental Studies

Anthony Copez-Lonzoy[1,2], MSc; Janina Bazalar-Palacios[3], MSc; Nikol Mayo-Puchoc[4], BS; Gustavo Calderon de la Cruz[5], MSc; Liliana Cruz-Ausejo[6], MSc; Evelyn Goicochea-Ríos[7], PhD; Juan Carlos Bazo-Alvarez[8,9], PhD

[1]Bibliometrics Unit, Universidad San Ignacio de Loyola, Lima, Lima Province, Peru
[2]Research Unit, PSYCOPERU Peruvian Institute for Psychological and Psychosocial Research, Ancash, Chimbote, Peru
[3]Facultad de Ciencias de la Salud, Universidad Tecnológica del Perú, Lima, Lima Province, Peru
[4]Mental Health Unit, Instituto Peruano de Orientación Psicológica, Lima, Peru
[5]School of Psychology, Peruvian University of Applied Sciences, Lima, Peru
[6]School Medicine, National University of San Marcos, Lima, Peru
[7]Grupo MedFam, Escuela de Medicina, Universidad César Vallejo, Trujillo, Peru
[8]Research Department of Primary Care and Population Health, University College London, London, United Kingdom
[9]Grupo MedFam, Escuela de Psicología, Universidad Cesar Vallejo, Trujillo, Peru

Corresponding Author:
Juan Carlos Bazo-Alvarez, PhD
Research Department of Primary Care and Population Health
University College London
Gower Street
London, WC1E 6BT
United Kingdom
Phone: 1 +44 7376 076260
Email: juan.alvarez.16@ucl.ac.uk

Abstract

Background: Highly visual social media (HVSM) platforms such as Facebook (Meta Platforms, Inc), Instagram (Meta Platforms, Inc), TikTok (ByteDance Ltd), and Snapchat (Snap Inc) have become central to the digital lives of adolescents and young adults. While these platforms have been linked to body dissatisfaction, they are also increasingly used as vehicles for health promotion. However, the evidence on interventions delivered through HVSM to address body image issues remains fragmented.

Objective: This review aimed to synthesize available evidence on interventions using HVSM platforms to reduce negative body image in adolescents and young adults.

Methods: We conducted a systematic search across 5 electronic databases (Scopus, MEDLINE, APA Psynet, Embase, and Web of Science) for studies published between January 2012 and October 2025. Eligible studies included experimental or quasi-experimental designs evaluating the effect of an HVSM-based intervention on body image outcomes in individuals aged 13 to 35 years. Risk of bias was assessed using the Risk of Bias Tool 2.0 (Cochrane) and was conducted independently by 2 researchers.

Results: Eight studies met the inclusion criteria with 4975 participants (2612 in intervention groups and 2363 in control groups). Most studies were conducted in high-income countries and had predominantly female participants. The interventions varied widely in format, duration, and theoretical basis. Microinterventions, brief interactive strategies such as gamified chatbots or short videos, were the most common and had moderate effects. Stimulus-based interventions using content with a positive body image or that did not focus on appearance were also identified, achieving moderate effects ($\eta_p^2<.07$), as well as combined approaches that integrated digital and face-to-face components to reduce negative body image ($P<.001$). The use and functionality of interventions using social media platforms were also compared by gender.

XSL•FO
RenderX

Conclusions: Body image management platforms offer an emerging avenue for implementing body image interventions in adolescents and young adults. While current evidence suggests modest benefits, the high heterogeneity among presentation formats and the variability in duration make comparisons between these studies difficult. This review synthesizes social media–delivered interventions for body image disturbance, going beyond broader digital approaches centered on websites or apps. It identifies cross-platform, putative mechanisms of action and common intervention formats, highlighting the potential of brief interventions for scalable reach and user empowerment via content curation. These findings define targets for optimization and underscore the need for platform safeguards and supportive policy and regulatory frameworks to enable safe real-world implementation, particularly for adolescents.

(J Med Internet Res 2026;28:e80141) doi:10.2196/80141

KEYWORDS

adolescent; body image; digital intervention; microinterventions; social media; systematic review; young adult

Introduction

Body image (BI) is a person's perception of their physical appearance and the potential positive or negative thoughts about self-worth based on that perception [1]. The problems associated with this negative perception of BI can affect various levels of health, such as eating disorders, risky behaviors, and excessive dieting [2]. This generates a public health problem affecting diverse age and gender populations [3]. Among the various mental health problems, negative BI remains a major concern for young people, especially adolescent girls [4]. According to research, between 20% and 40% of adolescent girls experience severe body dissatisfaction, which often persists or worsens during adulthood [4,5]. However, this prevalence of negative BI can reach 35% to 81% in adolescent girls and 16% to 55% in adolescent boys in low- and middle-income countries [5].

There is a strong correlation between BI and self-esteem. Adolescents and young adults frequently develop erroneous self-perceptions as a result of unrealistic physique ideals promoted by social media platforms [6]. According to research, excessive use of social media increases body dissatisfaction, promotes unfavorable comparisons, and lowers self-esteem [7]. This combination of low self-esteem and distorted BI increases the risk of developing eating disorders such as anorexia and bulimia nervosa [8]. Systematic reviews show that body dissatisfaction predicts the beginning of eating disorders and is linked to long-term mental health difficulties such as sadness and anxiety [9].

Adolescents and young adults are among the most active social media users, with more than 90% reporting frequent use of at least one site such as Instagram, TikTok, or Snapchat (Snap Inc) [10]. Globally, social media users reach 60% of the population, with adolescents and young adults frequently having accounts on various platforms and switching between them depending on age, cultural influences, and geographic location [11]. These platforms are intentionally designed to increase user engagement with features such as tailored content feeds, targeted notifications, and tools for modifying and sharing multimedia material [12]. Highly visual social media (HVSM) platforms, such as Instagram and TikTok, have transformed the digital environment by emphasizing photos and videos, making visual appeal an important driver of user retention [13,14]. These systems' algorithms continuously adjust content delivery to match user preferences, increasing interactivity and encouraging long-term engagement [15]. This dynamic produces a digital environment that captivates adolescents and young adults while also influencing their actions and beliefs, especially those connected to self-image and mental health [16].

Though highly visually appealing social media (HVSM) platforms are designed to enhance user engagement and dependency, they also have great potential as tools for mental health screening and intervention among teenagers and young adults. Recent research shows how feasible and successful digital mental health interventions through social media (eg, including HVSM platforms) [17-20]. These types of interventions overcome the accessibility and affordability limitations of traditional mental health services [21]. They can be administered at scale, are cost-effective to implement, and are easy to train, improving access to mental health services for individuals, thereby reducing the workload of health personnel [22].

Despite these developments, there are still important gaps in methodically summarizing treatments aimed at poor BI using HVSM. For example, little data exist on which particular platforms best handle BI issues and the special characteristics of interventions catered to these platforms [23]. Nevertheless, some alternatives use affordances of HVSM, such as specific algorithms, visual material, and interaction, to assess and minimize negative concerns about BI among adolescents and young adults. Nonetheless, further investigation is required to find best practices for using HVSM-based therapies, so that they are accessible, culturally relevant, and able to generate quantifiable changes in BI and general mental health [24,25], ultimately leading to the creation of creative and scalable solutions for urgent problems.

We aim to summarize evidence on the effectiveness of interventions that use HVSM platforms to tackle unhealthy BI in adolescents and young adults.

Methods

Overview

The systematic review was conducted according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Search Extension) guidelines [26-28] (Multimedia Appendices 1 and 2). The study protocol was published in open access [29].

XSL•FO
RenderX

Eligibility Criteria

Only studies included (1) adolescent and/or young adult populations (only groups with an age range from 13 to 35 years were included), (2) randomized clinical trials or quasi-experimental studies that evaluated BI interventions, (3) studies that fully or partially used social media as an intervention (eg, Facebook [Meta Platforms, Inc], Instagram [Meta Platforms, Inc], and TikTok [ByteDance Ltd)], or other interactive social media), and (4) studies that evaluated outcomes directly related to BI (eg, body dissatisfaction, body satisfaction, self-esteem, disordered eating, and appearance satisfaction). Studies that meet any of these criteria were excluded (1) studies that are not original journal articles (letters, book chapters, systematic reviews, and meta-analyses), (2) studies that cannot be considered interventions (eg, without pre or postevaluation and random assignment), or that are comparison conditions (eg, stimuli only), (3) studies that include a higher age range than established (eg, university students more than 35 years old at the time of the intervention), (4) studies that use social media to build the intervention, and (5) studies before 2012 (the first study that used social media as part of the intervention dates back to 2012) [30].

Information Sources

The initial search was conducted during 2024 and the final search was conducted up to October 10, 2025. Five databases were consulted (Scopus, Embase, Web of Science, Medline, and PsycINFO) in their entirety (ie, only studies that are listed in the databases). Only the information published in the studies was taken into account. We examined MeSH (Medical Subject Headings) terms for relevance and frequency of potentially included studies. We used validated filters to locate adolescent-based and young adult–based HVSM interventions where appropriate (eg, title, keyword, or abstract). AC-L and JB-P conducted citation tracking of initially included studies included in Scopus, identifying previously excluded studies and some of the studies included in the final review. Throughout the search process, standardized terms (eg, MeSH for MEDLINE) and Boolean operators (AND, OR) were applied to ensure a replicable process for each database.

Search Strategy

The following search strategy was used, based on the literature on BI and social media interventions in adolescents and young adults. In addition, publication year filters were included (eg, pubyear >2011). We include search strategies on BI, Mahon and Seekis [31], and adapt them to the needs of this study. The search strategies were peer reviewed by 2 team members (ACL and JBP) and an external expert (senior information specialist) prior to execution. All identified fields (eg, social media AND intervention) were unified to generate the search strategy. BI terms (appearance OR “body anxiety” OR “body attitude” OR “body attitudes” OR “body checking” OR “body concern*” OR “body esteem” OR “body evaluation” OR “body satisfaction” OR “body dissatisfaction” OR “body image” OR “body functionality” OR “body surveillance” OR “body shame” OR “body positive” OR “body positivity” OR “body acceptance” OR “body pride” OR “body preoccupation” OR “body awareness” OR “body anxiety” OR “body appreciation” OR “body neutrality” OR “physical self-perception” OR “physical self-perception” OR “physical self-concept” OR “physical self-concept” OR “shape concern” OR “shape concerns” OR “weight concern*” OR “shape satisfaction” OR “shape dissatisfaction” OR interocep* OR “weight dissatisfaction” OR “weight satisfaction” OR “body ideal*” OR “appearance ideal*” OR muscularity OR “thin ideal” OR “physical appearance” OR muscular* OR “social physique anxiety” OR “self-objectification” OR self-objectification OR “objectified body consciousness” OR thinness OR “cognitive restraint” OR “body identity” OR “body schema” OR “body representation*” OR “self-esteem” OR “eating disorders”), social media terms (internet OR “social media” OR “social network site*” OR instagram OR facebook OR “tik tok” OR “TikTok” OR twitter OR reddit OR telegram OR whapsapp OR weibo OR snapchat OR youtube OR pinterest), population terms (Child* OR adolescent* OR young OR youth OR teen* OR college OR universit* OR “Young Adult*”) randomized controlled trial (RCT) and quasi-experimental terms (intervention* OR “controlled trial” OR “clinical trial” OR “major clinical trial” OR “quasi-experimental” OR “quasi experimental”).

We limited the results to studies from 2012 onward, as this year saw the proliferation of large-scale studies on social media (eg, Facebook, Instagram, TikTok, or other HVSMs) [13]. The search strategy for each database is in Zenodo [32] or refer to Multimedia Appendix 3.

Selection Process and Data Collection

After eliminating duplicates, 2 researchers (ACL and JBP) independently assessed the titles and abstracts of all articles based on established criteria using Rayyan software (Qatar Computing Research Institute) [33]. We removed results that did not provide sufficient information from the full-text review. We repeated this process for the full-text phase. Data were extracted using a digital form predesigned by the research team. The form contains information on (1) general characteristics of the studies: first author, journal, country, study design (eg, RCT and quasi-experimental), participants or target population (Table 1); (2) social media: social media components and features used for the intervention or evaluation; (3) interventions: type of intervention, level of intervention (eg, country and hospital), duration of intervention, methodological details (eg, details of the design for the effectiveness evaluation, sample size, and instruments), effectiveness estimate and/or its estimation (eg, mean difference); and (4) assessment tools: construct or object to be assessed (eg, image, video, chatbot, or others), self-reports used one or more dimensions [34,35] (scales or inventories; Table 2).

XSL•FO
RenderX

Table 1. Description of intervention characteristics and highly visual social media platform features in 8 studies [24,25,36-41] to improve body image.a

Author (year)	Country	Body image outcomes	Social media platform	Affordances	Duration	Delivered by	Intervention description
Sampson et al [25] (2020)	United Kingdom	Body and facial dissatisfaction	Instagram	Photos	2 weeks	Researchers	They explored the assigned images on specific individual devices (eg, iPads), comparing the viewing of idealized content (eg, attractive smiles) and neutral images in Instagram users for 5 minutes.
Seekis et al [40] (2020)	Australia	Body dissatisfaction, drive for thinness, body appreciation, and self-compassion	Facebook	Facebook group	2 weeks	Researchers: workshop and Facebook group moderator	Attendance at 50-minute in-person workshops (6 workshops) held for 2 days, with an average of 12 participants each. These workshops are based on 6 strategies from the Mindful Self-Compassion manual, delivered as psychoeducation on self-compassion for body image. In addition, discussions are held in a private Facebook group (3 times per 2 weeks) to share experiences and apply what was learned in the workshop.
Fioravanti et al [38] (2023)	Italy	Body satisfaction, appearance comparison	Instagram	Images	28 days	Researchers	Researchers created Instagram profiles based on "body positivity," "encompassing fitspiration," and "neutral content" images, creating images for each profile type (approximately 112 images). Participants were randomly assigned to each group to follow their respective profiles. Content was delivered daily (1 post + 3 Instagram stories) for 28 days. In addition, participants received 2 messages per day asking questions about the content they received on Instagram.
Garbett et al [41] (2023)	Indonesia	Body satisfaction, attitudes toward appearance	Facebook and Instagram	Video	30 minutes	Self-guided	Self-guided web-based activities on 6 sequential videos (~ 5 minutes each) on body image issues related to Indonesian youth, complemented by social media activities.
Pilot et al [39] (2023)	United States	Body dissatisfaction	Instagram	Images	Not clear	Researchers	Progressive viewing of 20 pairs of fitspiration images on Instagram contrasted with reading and generating a list of 12 personal values and identifying the most important value.
Lewis-Smith et al [37] (2023)	United States	Body satisfaction, acceptance of appearance diversity and appearance-related internalized racism	YouTube	Video	5 minutes	Researchers	A theoretically grounded microintervention in the form of a dramatic miniseries (first episode only) aimed at improving body satisfaction, internalized appearance-related racism, and acceptance of appearance diversity (specifically, acceptance of larger-bodied people of Black people) among Black and non-Black adolescents (American Indian and Alaska Native, Asian, Native Hawaiian and Other Pacific Islands or White) in the United States.
Matheson et al [24] (2023)	Brazil	Body satisfaction, positive and negative affect	Facebook	Chatbot	3 days	Avatar	A chatbot operated through Facebook Messenger that provided microinterventions for improving body image and mental well-being based on psychoeducational techniques, interactive activities, and coping strategies.
Fardouly et al [36] (2023)	Australia	Body dissatisfaction and appreciation, self-objectification, appearance comparison tendency, and body activism	Facebook	Facebook group	14 days	Social media	Microintervention created in private Facebook groups was based on the exposure of body-positive content. Two different conditions were administered (positive and neutral body awareness). Participants were dosed with posts of content relevant to each condition for 14 consecutive days (3 times a day).

[a]Affordances = set of tools, characteristics, and options of social media platforms for interaction, creation, management, and sharing between users (eg, photos, private or open groups, text, video, etc).

XSL•FO
RenderX

Table 2. Essential characteristics and key findings from randomized controlled trials on the effectiveness of social media interventions to improve body image[a].

Author (year)	Sample size	Age	Study design	Follow-up	Comparison group	Body images measures	Key findings	Did the intervention achieve its goals?
Sampson et al [25] (2020)	132	Mean 20.50 (2.21), range 18-30 years	Randomized controlled trial	N/A[b]	Yes, images of neutral nature	BSS[c]; FSS[d]; SACS[e]; SDI[f]	Exposure to "ideal" facial images on social media decreases facial satisfaction (P<.001).	Partially
Seekis et al [40] (2020)	76	Mean 18.04 (0.90), range 17-21 years	Cluster randomized trial	1-3 months	Yes, waitlist	Subscale EDI[g]; SAAS[h]; UPACS[i]; BAS-2[j]; SCS-SF[k]	The combined strategy (workshop + Facebook group) based on self-compassion reduced body dissatisfaction (β=–.74; P<.001).	Yes
Garbett et al [41] (2023)	1847	Intervention mean 16.94 (SD 1.40) years; control: mean 17 (SD 1.41), range 15-19 years	Randomized controlled trial	1 month	Yes, waitlist	BESAA[l]; SATAQ-3[m]; PANAS-C[n]	Warna-Warni Waktu is an effective eHealth intervention to reduce body dissatisfaction among adolescent girls and young women (P=.005, η^2_p=0.05).	Yes
Lewis-Smith et al [37] (2023)	686	Mean 15.72 (SD 1.60), range 13-18 years	Parallel randomized controlled trial	N/A	Yes, non–body image–related video	VAS[o]; subscale IROS[p]	No changes in body satisfaction were observed. The changes identified were only in conditions of racism and acceptance of diversity (P<.001; Cohen d>0.17).	No
Matheson et al [24] (2023)	1715	Range 13-18 years (52.54% girls)	Randomized controlled trial	1 week; 1 month	Yes, assessment-only control	BESAA; PANAS; BIS-ES[q]	Body self-esteem identified improvements in positive (Cohen d=0.13; P=.02) and negative (Cohen d=0.11; P=.047) appearance.	Yes
Fardouly et al [36] (2023)	159	Mean 20.92 (SD 2.21), 18-25 years	Parallel randomized controlled trial	4 weeks	Yes, assessment-only control	Subscale EDI; BAS-2; PANAS-SF[r]; SOBBS[s]; PACS[t]; SAM[u] modified	Body dissatisfaction was reduced in the body positive Facebook group condition (P=.03; η_p^2=0.05) with small effect.	Yes
Pilot et al [39] (2023)	238	Mean 19.89 (SD 1.25), 18-22 years	Randomized controlled trial	N/A	Yes, travel images	Subscale EDEQ[v]; WREQ[w]; VAS	The intervention based on Values Affirmation + Fitspiration does not show significant changes in body image (P=.73) or in negative mood (P=.60) compared to the control.	No

[a]Relevant results for BI = The information was assessed by the team as "no" (no favorable results for body image), "partially" (favorable results for some body image outcomes), and "yes" (favorable results for body image outcomes).

[b]N/A: not applicable.

[c]BSS: Body Satisfaction Scale.

[d]FSS: Facial Satisfaction Scale.

[e]SACS: State Appearance Comparison Scale.

[f]SDI: Self-Discrepancy Index.

[g]EDI-3: Eating Disorder Inventory.

[h]SAAS: Social Appearance Anxiety Scale.

[i]UPACS: Upward Physical Appearance Comparison Scale.

[j]BAS-2: Body Appreciation Scale-2.

[k]SCS-SF: Self-Compassion Scale–Short-Form.

[l]BESAA: Body Esteem Scale for Adolescents and Adults.

[m]SATAQ-3 Sociocultural Attitudes Towards Appearance Questionnaire.

[n]PANAS-C: Positive and Negative Affect Schedule for Children.

[o]VAS: Visual Analogue Scales.

XSL•FO
RenderX

[p]IROS: Internalized Racial Oppression Scale (Alteration of Physical Appearance and Hair Change subscales).

[q]BISES: Body Image Self-Efficacy Scale.

[r]PANAS-SF: Positive and Negative Affect Scale-Short Form.

[s]SOBBS: Self-Objectification Beliefs and Behaviors Scale.

[t]PACS: Physical Appearance Comparison Scale.

[u]SAM: modified version of the social activism measure.

[v]EDEQ: Eating Disorder Examination Questionnaire (weight or shape subscales).

[w]WREQ: Weight-related eating questionnaire.

Risk of Bias

Methodological quality was assessed using the Risk of Bias Tool (RoB 2; Cochrane Collaboration) established by the Cochrane System of Interventions Assessment Handbook [42]. This tool evaluates 5 domains: the randomization process, deviations from planned interventions, missing data, outcome measurement, and selection of reported outcomes. We classified the judgment rating for each domain as "yes," "probably yes," "no," "probably no," and "unreported." Furthermore, the combination of the 5 domains provides an overall score that classifies the risk as "low," "high," or "some concerns." Two reviewers (AC-L and JB-P) conducted independent assessments of the studies included in this paper. We resolved any disagreements through discussions with a third reviewer (JCB-A).

Synthesis Methods

The information was systematically synthesized into tables containing descriptive information on the selected results. Tables were generated for the descriptive results and separated by specific design type (eg, RCT or quasi-experimental). Figures were generated based on the specific characteristics of the included studies and the methodological quality assessment by domain and overall (risk of bias). The information was grouped taking into account 4 intervention conditions that use social media as microinterventions, stimulus comparison, combined interventions, and self-guided videos. A narrative synthesis was chosen due to two critical conditions that prevented a meta-analysis: (1) the wide heterogeneity previously identified in the studies such as the different populations (eg, adolescents, young adults, or both), social media platforms (Facebook, Instagram, YouTube, or combined), affordances (ie, photos, images, videos, Facebook groups, post or chatbots), duration of interventions (eg, 5 minutes, 3 days, and up to 14 days) and follow-up (1 week to 1 month) and (2) the small number of studies included (k=3) per condition assessed (eg, body satisfaction or dissatisfaction) would not allow a subgroup analysis to understand the possible source of heterogeneity that could affect clinical decision-making as mentioned in the Cochrane guide [43].

Ethical Considerations

The research ethics committee from Universidad Cesar Vallejo, with a code of 100-CEI-EPM-UCV-2022, reviewed and approved this study. No additional permissions were required as the data were based on articles published in different databases accessible to the authors (AC-L, JB-P, and JCB-A).

Results

Study Selection

A total of 2571 references were identified across all databases reviewed in the literature search. After identifying duplicates, 1761 articles were retained for review by title and abstract. Sixty-one full-text articles were included. Of these 61 articles, 53 were excluded, the following exclusion characteristics were outside the age range (3/53, 5.7%) were not an intervention (4/53, 7.5%), did not use social media (4/53, 7.5%) did not have RCTs or quasi-experimental (40/53, 75.5%) and the outcome was not related to BI (2/53, 3.8%) leaving 8 articles for this review (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analysis) flow diagram describing the literature review process. RCT: randomized controlled trial.

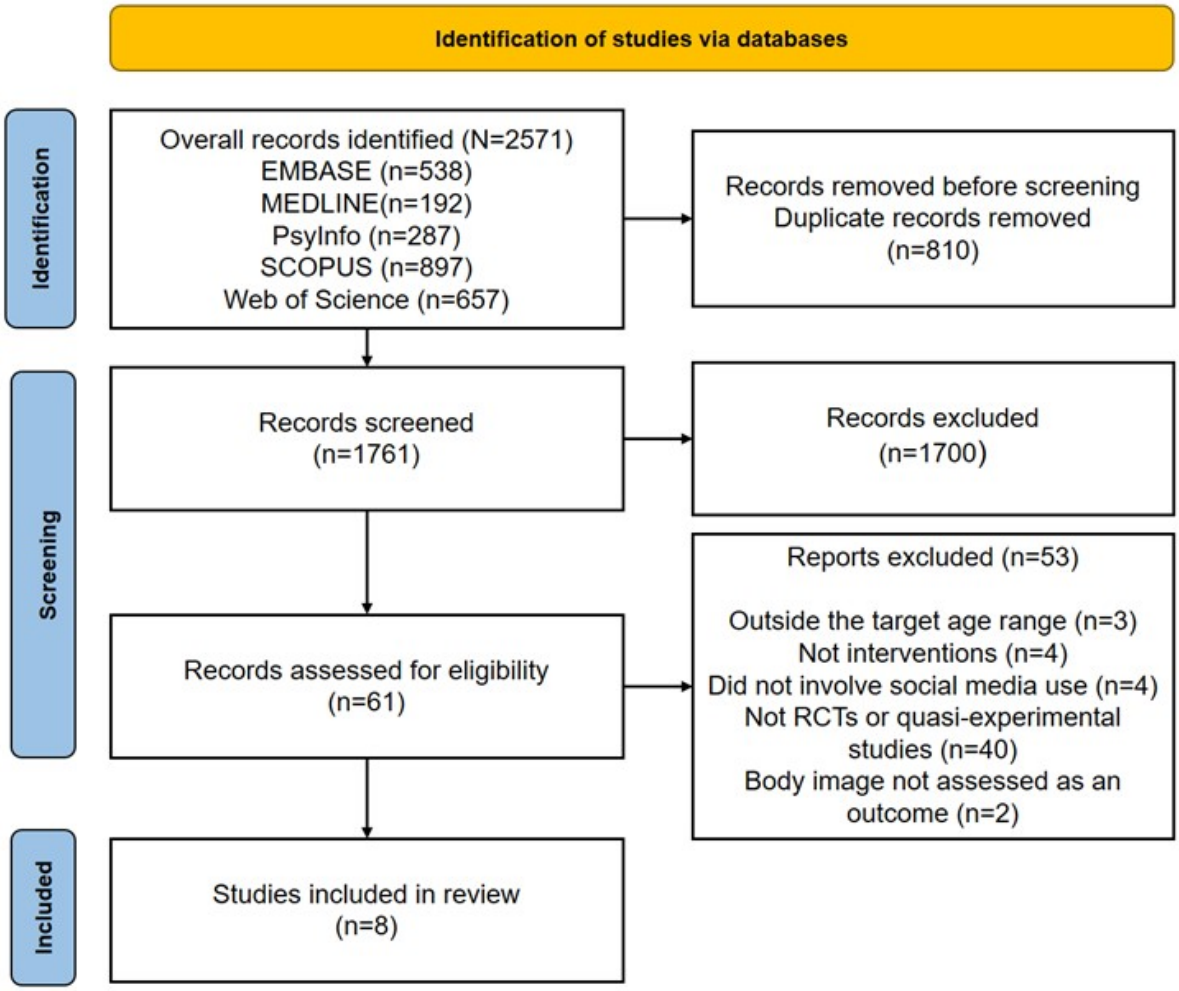

Descriptive Characteristics of the Studies

Table 1 shows an overview of the studies identified by social media year, country, and social media use. All identified studies were published between 2020 and 2023, with a majority based in the American Continent (Brazil and the United States). The majority of studies were conducted on female participants (6/8, 75%), and only 2 studies were conducted for mixed-gender groups [24,25]. The main platform for intervention (partial or total) was Facebook (4/8, 50%). The intervention duration ranged from 5 minutes to 28 days, and in most cases, the intervention was delivered by the researchers (6/8, 75%). The characteristics of the identified interventions were microinterventions to improve body satisfaction (n=3) [24,36,37], combined interventions (workshop + online group; n=1) [40], stimulus-based interventions with the presentation of socially "ideal" and "attractive" images (n=3) [25,38,39], and a self-guided model based on sequential videos to combat appearance-related pressures (n=1) [41].

Characteristics of the Interventions

The total number of study participants across all studies was 4975 (approximately 17.7% boys [24,25] and 0.3% others [13]) with a sample size range from 76 to 1847 participants per study. The age range of participants was 13 to 30 years. There were 7 RCTs and 1 quasi-experimental study. Furthermore, 50% (4/8) of the studies included follow-up, ranging from 1 week to 3 months [24,36,40,41]. Control group conditions varied, including placebo (3/8, 42.8%) [25,37,39], waitlist (2/8, 28.6%) [40,41], and assessment-only control (2/8, 28.6%) [24,36] (Table 2; Figure 2). The instruments to assess BI were very diverse, including the visual analogical scale (n=3) [37-39], positive or negative affect (n=2) [24,41], and subscales of the eating disorders inventory (n=3) [36,39,40] (Table S1 in Multimedia Appendix 4) [24,25,36-41].

XSL•FO
RenderX

Figure 2. Graphical summary of the main characteristics and outcomes identified.

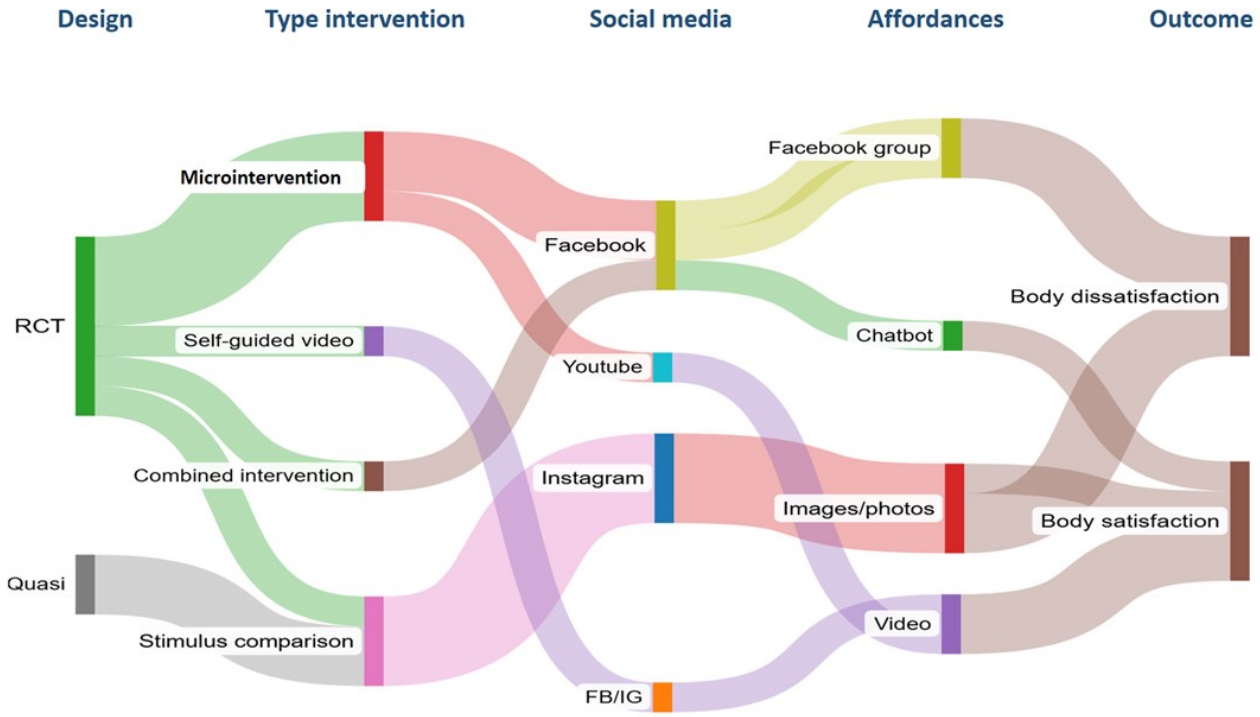

The Effectiveness of Microintervention for Improving BI

Microinterventions, due to their ease of administration (ie, single or multiple), can generate an effective and positive sequence on the immediate impact of specific symptoms (eg, body satisfaction or dissatisfaction) [44]. The effectiveness of these interventions for improving BI is centered on reducing body comparison, developing strategies to mitigate criticism of media beauty standards, and cognitive reevaluation of body-related thoughts [45,46]. These types of interventions show positive effects when using gamified and interactive self-guided models or edutainment strategies for healthy behavior change [47,48]. Three studies included microinterventions to improve body satisfaction and/or dissatisfaction [24,36,37]. They offered easily understandable content to adolescents and young adults on a single occasion or over a specific period, the latter usually daily, with different degrees of success. For example, African American girls did not show any improvement in body satisfaction after watching content on social media for comparison and pressure in a drama miniseries format (P=.11) [37]. In another study, gamified chatbot-based models did improve the body self-esteem of boys and girls, although with a small effect size (P<.02) [24]. A third study exposed adolescents and young adults to body-positive content on Facebook in different presentations (eg, access to nonidealized social beauty content presented in photos, images, external links, or access to other social media) in specific Facebook groups, resulting in reduced body dissatisfaction (P=.03) and reduced tendency to compare appearance (P<.01) [36].

The Effectiveness of Comparative Stimulus Interventions

Three studies assessed how specific social media features (eg, posting photos or images) can change some aspects of BI in adolescents and young adult women [25,38,39]. Exposure to idealized images (eg, attractive smiles of Instagram users) was found to reduce facial satisfaction (P<.01) [25]. In another study, creating profiles on social media platforms (eg, Instagram) with body-positive content increased positive mood and body satisfaction (ie, with the neutral image group) [38]. Strategies based on identifying self-affirming values (ie, selecting and reflecting on the most important value presented in the intervention) did not produce significant changes in BI and mood (P>.60) [39].

The Effectiveness of a Combined Intervention

One study combined a workshop with group interactions on Facebook [40]. The brief workshop was focused on self-compassion competencies. Then, workshop participants shared previous workshop experiences and learnings in a private Facebook group. The evaluation showed a favorable reduction in body dissatisfaction (P<.01).

The Effectiveness of Self-Guided Video Activities

One study used a self-guided educational model consisting of short, sequential videos. These videos were shared on social media (Facebook or Instagram) to explore BI risk factors among young Indonesian women [41]. These activities were effective in reducing body dissatisfaction in girls and young women (P<.01), although with a small effect size (η^2=0.05).

XSL•FO
RenderX

Efficacy of HVSM Interventions Comparison Between Adolescents and Young Adults

The studies targeting adolescents were mostly conducted in the Americas (ie, the United States or Brazil) with a variety of HVSM administration (eg, Facebook, Instagram, YouTube, or a combination of both HVSM) [24,37]. Overall, the time of the interventions in the adolescent group was shorter (ie, from 5 minutes to 3 days) compared to young adults (14 to 28 days) [24,25,36-40]. Only the adolescent group used video and chatbot interventions. Two of the 3 interventions for adolescents did improve BI, with small effects for reducing body dissatisfaction or improving body self-esteem (Cohen d=0.13) [24,41], while the interventions that achieved positive effects on BI in young adults focused on reducing body dissatisfaction.

Risk of Bias

A summary of the risk of bias assessment of the RCT studies is visible in Figure 3. Three studies did not report enough details on randomization [25,37,39,40], leading to internal validity uncertainty. One study raised concerns about bias resulting from potential overestimated intervention effects [25,39]. All studies presented low risk in terms of missing data, outcome measurement validity, and outcome selection [24,25,36,37,39,40,41].

Figure 3. Risk of bias graph for randomized controlled trials.

Discussion

Principal Findings

This is the first systematic review of interventions that use HVSM platforms to tackle unhealthy BI in adolescents and young adults. While the number of available studies remains limited, the data suggest that these interventions—despite their variability—can support modest improvements in body-related outcomes. We classified these programs into 4 types: microinterventions, interventions based on comparative stimulus, combined interventions, and self-guided video activities. All intervention types showed improvements in BI, albeit with small effect sizes. Some risk of bias was detected, particularly a lack of randomization in half of the studies, which can compromise the validity of intervention effect estimates. Overall, the studies identified in the systematic review addressed both the increase in body satisfaction and the reduction in dissatisfaction.

Microinterventions can be effective under certain conditions, such as clarity in the actions directed to the target audience (eg, formative or interventionist), a combination of specific strategies (ie, how and when to implement the actions), including interactive strategies (eg, completing activities). Chatbot-based microintervention techniques (eg, social media content management, cognitive-behavioral models, and psychoeducational techniques) showed positive effects on body satisfaction [24]. Microinterventions using gamification for BI may be advantageous due to their narrow applicability (ie, even up to a single session), short duration (eg, videos less than 5 minutes long), and greater specificity (eg, adolescents in specific age ranges) [47,49]. Brief experiences followed by quick feedback—typical in social media—can produce rewarding experiences that promote engagement and adherence [50,51]. Similarly, microinterventions using a sequenced dosage of body-positive content (eg, photo posts, image posts, and links) in social media groups (eg, Facebook groups) were also beneficial in improving body satisfaction [36]. Thus, repeated and sequenced exposure to content that promotes body acceptance or is aesthetically neutral on a daily basis (eg, ~2 weeks and at least once a day) could reduce the internalization of beauty ideals and improve BI, acceptance, and body diversity by counteracting the effect of feeds of idealized images [52].

Exposure to positive content on social media could increase cognitive flexibility regarding the internalization of socially acceptable beauty standards [53]. Even viewing neutral content

XSL•FO
RenderX

(eg, travel, landscapes, art, etc) can motivate young women to decentralize their BI (ie, interests beyond appearance) [54].

However, a microintervention based solely on a single online dramatization episode did not reduce body dissatisfaction [37]. This may be due to the indirect and unrealistic narrative (ie, presented in a single short video) of body dissatisfaction-related messages, as opposed to strategies based on successive videos (eg, miniseries or sequential videos) that can provide feedback on specific and direct topics [55]. The lack of human interaction may decrease the appeal of these interventions, especially for adolescents and young adults [12,47]. Likewise, having main characters that resemble thin-ideal standards (eg, in online series produced for an intervention) may encourage internalized appearance comparison, generating the indirect effect of body dissatisfaction [52].

Comparison stimuli based on positive images or photographs (eg, more realistic bodies) or neutral content (eg, nature, landscapes, or trips) can reduce the effects of body dissatisfaction [25,38,39]. Exposure to positive visual content can significantly reduce the evaluative determinants of body dissatisfaction (eg, increased stress and decreased subjective well-being) [56]. Neutral content (ie, related to nature) can generate subjective benefits through the sensation and appreciation of a healthy body [57,58]. However, images with thinness standards or fitspiration have been shown to increase thoughts directed toward social comparison with people subjectively perceived as superior (eg, artists, actors, influencers, or other stereotypes), increasing body dissatisfaction [59].

Exposure to sequential educational series (ie, more than one chapter) that included interactive activities (eg, completing a comic or identifying the message of a specific episode) reduced body dissatisfaction [41]. The use of entertainment series (eg, drama series designed with educational content) may have a positive effect on attitude change [60]. Thus, the familiarity that the characters can generate and the interaction with the quality of the storyline (ie, the anticipatory effect) play a crucial role in the incorporation of the desired messages about thought control and the elimination of personal comparisons that promote risk factors in body dissatisfaction [61].

The use of Facebook groups in combination with self-compassion workshops shows favorable results in reducing body dissatisfaction [40]. Combined strategies of conscious self-compassion through healthy self-talk, positive changes in self-worth, and physical warmth foster tolerance of one's own physical appearance [62,63]. Added to the online component (eg, Facebook group), they can reduce appearance comparison ideals and body concerns. A combined intervention that includes active participation in Facebook groups (eg, reading others' experiences and compassionate writing) may increase the benefits of joint feedback and how this can help recognize negative emotions related to corporality [64,65].

Despite the variability of the interventions, they share common elements such as (1) repeated dose and sequence (eg, microinterventions), (2) specific conceptual focus, that is, they must be centered on outcomes close to BI, and (3) adaptable, with plausible and culturally clear messages about body acceptance. Exposure to positive images on social media (eg, inspirational images that do not conform to hegemonic beauty standards) can foster self-improvement motives related to body satisfaction [38,66]. Furthermore, self-compassionate writing as a core component in interventions using private social media (eg, Facebook) can enhance interventions or workshops aimed at reducing body-related concerns [40,65]. Sequenced videos (eg, entertainment-education) primarily rely on psychoeducational components (eg, appearance-based comparisons) and elements of cognitive behavioral therapy, such as cognitive dissonance (ie, redirecting thoughts about physical changes), combined with interactive models that can generate motivation and adherence [24,41].

Our review identified differential effectiveness in BI interventions that align with the activities and needs of adolescents and young adults groups with a majority of women. In adolescent populations, interactive resources (eg, videos or game strategies) that incorporate cognitive-behavioral and coping components can help manage external validation, regulate feelings of BI self-worth, and address peer influence [24,41]. In contrast, young adults respond effectively to positive visual content (ie, photos, posts, or links), generating positive upward comparison (ie, motivation for self-improvement) [67]. Furthermore, the use of private groups can facilitate self-compassionate feedback to reduce preoccupation with rigid beauty ideals [65]. This suggests the need to develop differentiated strategies for both channels and content to optimize the effectiveness of interventions.

Overall, there is a methodological limitation in half of the studies, which may be considered unreliable results. Specifically, the studies by Sampson et al [25], Seekis et al [40], Pilot and Stutts [39], and Lewis-Smith et al [37] presented some risk of bias due to a lack of randomization and 2 studies also included issues regarding deviations from planned interventions [25,39]. Regardless of the intervention resource administered in HVSM, selection bias (D1) could exist, generating differences in potentially important prognostic characteristics (eg, related to baseline BI) [68,69]. This lack of comparability to overestimate the potential effects (eg, in a combined intervention workshop + Facebook group) means that the observed benefits could be attributed to the pre-existing characteristics of the participants, rather than the intervention itself [68]. Additionally, implementation bias (D2) could introduce further biases, such as potential contamination of the control group (eg, in interventions using HVSM), thus reducing the actual effects of the intervention group [69]. Despite this, another group of studies achieved a low risk of bias and identified high feasibility and acceptability with favorable results for HVSM interventions in different age groups (eg, adolescents and young adults) and intervention formats (ie, video, chat, and visual content) [24,36,41]. This suggests that platforms such as Facebook and Instagram (or others) could be useful for improving BI once the methodological limitations identified in this study are overcome, potentially providing a strong rationale for future interventions.

While other reviews have explored various digital interventions (eg, apps or websites), this research is innovative because it focuses exclusively on social media as a core component of interventions to improve BI. To our knowledge, this is the first review to synthesize evidence specifically on social media–based

XSL•FO
RenderX

interventions for BI disturbance (rather than broader digital or eHealth approaches), thereby expanding our understanding of the potential use of social media as an intervention delivery channel and active therapeutic ingredient. This review extends prior work by moving beyond platform descriptions to examine how interventions operate, identifying putative mechanisms of action used regardless of platform (eg, social comparison, cognitive dissonance, self-compassion, and behavioral modeling) and the intervention format used (eg, video-based content, stimulus comparison, and microinterventions). The contributions to the field include evidence suggesting that brief interventions can still be effective and the potential for large-scale reach and scalability. Although studies of social media–based interventions remain limited, our findings highlight clear targets for optimization that future intervention developers can incorporate. In real-world terms, social media–based tools could be embedded within platforms or delivered alongside routine social media use to help prevent or reduce symptoms of BI disturbance and counterbalance exposure to unhealthy beauty standards among groups at higher risk, including adolescents [24,70]. Realizing this potential may require platform-level design changes, clearer safeguarding standards, and supportive policy and regulatory frameworks, alongside partnerships with civil society organizations working to protect young people's mental health [15].

Strengths and Limitations

This study offers the first systematic evaluation of BI interventions implemented through HVSM platforms among youth and emerging adults. Furthermore, our study identifies the content and digital affordances that can improve body satisfaction in this vulnerable population. However, in the search strategy, other terms that encompass the full meaning of BI may not have been captured. To mitigate this, the research team comprehensively mapped search terms that could potentially capture BI-related words. The small number of studies identified precluded a meta-analysis, which would help to provide cumulative evidence for informed recommendations and clinical decision-making. Nevertheless, our study is useful in summarizing the updated information on this type of cost-effective intervention with a high adherence rate in improving mental health problems.

Conclusions

Interventions delivered through HVSM platforms to improve BI still require further development and research due to the preliminary evidence of their findings. Some formats—especially those that are interactive, brief, and visually engaging—appear to foster small but meaningful changes in how adolescents and young adults perceive their appearance. The identified methodological heterogeneity implies difficulty in obtaining a single estimate, which currently prevents a formal comparison of effects across included studies. This suggests that the interventions were designed exclusively for specific contexts and may not generate the same expected effects. However, this variety of intervention formats allows us to identify functional principles and adaptable conditions related to the dosage and mechanisms of healthy BI improvement, as well as the contextual sensitivity of success or failure in specific populations (eg, adolescents and young adults). Despite this challenge, social media remains a space with enormous reach and influence. If thoughtfully designed, interventions embedded in these platforms could become an important part of broader efforts to address BI issues and promote psychological well-being in vulnerable populations.

Acknowledgments

This review is independent research supported by the National Institute for Health and Care Research Applied Research Collaboration (ARC) North Thames. The views expressed in this publication are those of the author(s) and not necessarily those of the National Institute for Health and Care Research or the Department of Health and Social Care. The authors did not use generative artificial intelligence in the research and writing process.

Funding

This study was partially funded by Universidad Cesar Vallejo (project number 894). JCB-A was funded by the National Institute for Health Research (NIHR) 3 Research Schools' Mental Health Programme (Grant Reference Number: MHF012).

Data Availability

The protocol generated or analyzed during this study is available in Zenodo [71].

Authors' Contributions

AC-L, JB-P, EG-R, and JCB-A conceived the study idea and prepared and published the protocol. AC-L and JB-P performed the search, review, and synthesis supervised by JCB-A. NMP, LCA, and GCDC carried out both phases of the extraction process. AC-L and JB-P drafted the manuscript, receiving iterative edits and feedback from JCB-A and EG-R. All authors provided final approval for the submitted version. AC-L is the guarantor for this submission.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
PRISMA checklist.
[PDF File (Adobe PDF File), 41 KB - jmir_v28i1e80141_app1.pdf]

Multimedia Appendix 2
PRISMA-S checklist.
[PDF File (Adobe PDF File), 472 KB - jmir_v28i1e80141_app2.pdf]

Multimedia Appendix 3
Search strategy.
[DOCX File , 16 KB - jmir_v28i1e80141_app3.docx]

Multimedia Appendix 4
Measurement tools.
[DOCX File , 14 KB - jmir_v28i1e80141_app4.docx]

References

1. Grogan S. Body Image: Understanding Body Dissatisfaction in Men, Women and Children. 4th ed. London: Routledge; 2021.
2. Hong W, Ahmad NS. Body image dissatisfaction, depression and anxiety in a sample of college students. Open Psychol J 2024;17(1):e18743501303541. [doi: 10.2174/0118743501303541240521104133]
3. Mallaram GK, Sharma P, Kattula D, Singh S, Pavuluru P. Body image perception, eating disorder behavior, self-esteem and quality of life: a cross-sectional study among female medical students. J Eat Disord 2023;11(1):225 [FREE Full text] [doi: 10.1186/s40337-023-00945-2] [Medline: 38102717]
4. McLean SA, Rodgers RF, Slater A, Jarman HK, Gordon CS, Paxton SJ. Clinically significant body dissatisfaction: prevalence and association with depressive symptoms in adolescent boys and girls. Eur Child Adolesc Psychiatry 2022;31(12):1921-1932. [doi: 10.1007/s00787-021-01824-4] [Medline: 34129105]
5. Moehlecke M, Blume CA, Cureau FV, Kieling C, Schaan BD. Self-perceived body image, dissatisfaction with body weight and nutritional status of Brazilian adolescents: a nationwide study. J Pediatr (Rio J) 2020;96(1):76-83 [FREE Full text] [doi: 10.1016/j.jped.2018.07.006] [Medline: 30098939]
6. Jiménez-García AM, Arias N, Hontanaya EP, Sanz A, García-Velasco O. Impact of body-positive social media content on body image perception. J Eat Disord 2025;13(1):153 [FREE Full text] [doi: 10.1186/s40337-025-01286-y] [Medline: 40713881]
7. Holland G, Tiggemann M. A systematic review of the impact of the use of social networking sites on body image and disordered eating outcomes. Body Image 2016;17:100-110. [doi: 10.1016/j.bodyim.2016.02.008] [Medline: 26995158]
8. Karakuş Aydos Y, Dövencioğlu D, Karlı Oğuz K, Özdemir P, Pehlivantürk Kızılkan M, Kanbur N, et al. Neural correlates of distorted body images in adolescent girls with anorexia nervosa: How is it different from major depressive disorder? J Neuropsychol 2024;18(1):154-172. [doi: 10.1111/jnp.12340]
9. Stice E. Risk and maintenance factors for eating pathology: a meta-analytic review. Psychol Bull 2002;128(5):825-848. [doi: 10.1037/0033-2909.128.5.825] [Medline: 12206196]
10. Anderson M, Auxier B. Social media use in 2021. Pew Research Center. URL: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ [accessed 2025-02-19]
11. Social media. Statista. URL: https://www.statista.com/topics/1164/social-networks/ [accessed 2025-02-19]
12. Kruzan KP, Williams KD, Meyerhoff J, Yoo DW, O'Dwyer LC, De Choudhury M, et al. Social media-based interventions for adolescent and young adult mental health: a scoping review. Internet Interv 2022;30:100578. [doi: 10.1016/j.invent.2022.100578]
13. Sherlock M, Wagstaff DL. Exploring the relationship between frequency of Instagram use, exposure to idealized images, and psychological well-being in women. Psychol Pop Media Cult 2019;8(4):482-490. [doi: 10.1037/ppm0000182]
14. Herriman Z, Taylor AM, Roberts RM. Interventions to reduce the negative impact of online highly visual social networking site use on mental health outcomes: a scoping review. Psychol Pop Media 2024;13(1):111-139. [doi: 10.1037/ppm0000455]
15. Harriger JA, Evans JA, Thompson JK, Tylka TL. The dangers of the rabbit hole: reflections on social media as a portal into a distorted world of edited bodies and eating disorder risk and the role of algorithms. Body Image 2022;41:292-297. [doi: 10.1016/j.bodyim.2022.03.007] [Medline: 35378338]
16. Ridout B, Campbell A. The use of social networking sites in mental health interventions for young people: systematic review. J Med Internet Res 2018;20(12):e12244 [FREE Full text] [doi: 10.2196/12244] [Medline: 30563811]
17. Valentine L, McEnery C, D'Alfonso S, Phillips J, Bailey E, Alvarez-Jimenez M. Harnessing the potential of social media to develop the next generation of digital health treatments in youth mental health. Curr Treat Options Psych 2019;6(4):325-336. [doi: 10.1007/s40501-019-00184-w]

XSL•FO
RenderX

18. Ballesteros-Holmes J, Teague M, Derouin A. Decreasing social media use through motivational interviewing: a pediatric primary care quality improvement project. J Pediatr Health Care 2024;38(6):903-911. [doi: 10.1016/j.pedhc.2024.05.006] [Medline: 39365218]
19. Yu DJ, Wing YK, Li TMH, Chan NY. The impact of social media use on sleep and mental health in youth: a scoping review. Curr Psychiatry Rep 2024;26(3):104-119 [FREE Full text] [doi: 10.1007/s11920-024-01481-9] [Medline: 38329569]
20. Valentine L, McEnery C, O'Sullivan S, Gleeson J, Bendall S, Alvarez-Jimenez M. Young people's experience of a long-term social media-based intervention for first-episode psychosis: qualitative analysis. J Med Internet Res 2020;22(6):e17570 [FREE Full text] [doi: 10.2196/17570] [Medline: 32384056]
21. Marcu G, Ondersma SJ, Spiller AN, Broderick BM, Kadri R, Buis LR. The perceived benefits of digital interventions for behavioral health: qualitative interview study. J Med Internet Res 2022;24(3):e34300 [FREE Full text] [doi: 10.2196/34300] [Medline: 35353045]
22. Zhang Q, Huang Z, Sui Y, Lin F, Guan H, Li L, et al. Social-media-based mental health interventions: meta-analysis of randomized controlled trials. J Med Internet Res 2025;27:e67953 [FREE Full text] [doi: 10.2196/67953] [Medline: 40811685]
23. Marengo D, Longobardi C, Fabris M, Settanni M. Highly-visual social media and internalizing symptoms in adolescence: the mediating role of body image concerns. Comput Hum Behav 2018;82:63-69. [doi: 10.1016/j.chb.2018.01.003]
24. Matheson EL, Smith HG, Amaral ACS, Meireles JFF, Almeida MC, Linardon J, et al. Using chatbot technology to improve brazilian adolescents' body image and mental health at scale: randomized controlled trial. JMIR Mhealth Uhealth 2023;11:e39934 [FREE Full text] [doi: 10.2196/39934] [Medline: 37335604]
25. Sampson A, Jeremiah HG, Andiappan M, Newton JT. The effect of viewing idealised smile images versus nature images via social media on immediate facial satisfaction in young adults: a randomised controlled trial. J Orthod 2020;47(1):55-64. [doi: 10.1177/1465312519899664]
26. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
27. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39. [doi: 10.1186/s13643-020-01542-z]
28. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021 Mar 29;372:n160 [FREE Full text] [doi: 10.1136/bmj.n160] [Medline: 33781993]
29. Copez-Lonzoy A, Palacios J, Goicochea-Ríos E, Bazo-Alvarez J. Interventions that use highly visual social media platforms to tackle unhealthy body image in adolescents and young adults: a systematic review protocol. BMJ Open 2025;15(7):e095133. [doi: 10.1136/bmjopen-2024-095133]
30. Cavallo DN, Tate DF, Ries AV, Brown JD, DeVellis RF, Ammerman AS. A social media-based physical activity intervention: a randomized controlled trial. Am J Prev Med 2012;43(5):527-532 [FREE Full text] [doi: 10.1016/j.amepre.2012.07.019] [Medline: 23079176]
31. Mahon C, Seekis V. Systematic review of digital interventions for adolescent and young adult women's body image. Front Glob Womens Health 2022;3:832805. [doi: 10.3389/fgwh.2022.832805]
32. body image search strategy. Zenodo. URL: https://zenodo.org/records/18111594 [accessed 2026-01-20]
33. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
34. Clark LA, Watson D. Constructing validity: new developments in creating objective measuring instruments. Psychol Assess 2019;31(12):1412-1427 [FREE Full text] [doi: 10.1037/pas0000626] [Medline: 30896212]
35. Gorrall BK, Curtis J, Little TD, Panko P. Innovations in measurement: visual analog scales and retrospective pretest self-report designs. Actualidades en Psicología 2016;30(120):1-6 [FREE Full text] [doi: 10.15517/ap.v30i120.22932]
36. Fardouly J, Slater A, Parnell J, Diedrichs PC. Can following body positive or appearance neutral Facebook pages improve young women's body image and mood? Testing novel social media micro-interventions. Body Image 2023;44:136-147. [doi: 10.1016/j.bodyim.2022.12.008] [Medline: 36608433]
37. Lewis-Smith H, Pegram G, White P, Ward LM, Diedrichs PC. A short-form drama series created for the digital media environment: a randomised controlled trial exploring effects on girls' body satisfaction, acceptance of appearance diversity, and appearance-related internalised racism. Body Image 2023;47:101610. [doi: 10.1016/j.bodyim.2023.08.002] [Medline: 37659248]
38. Fioravanti G, Svicher A, Ceragioli G, Bruni V, Casale S. Examining the impact of daily exposure to body-positive and fitspiration Instagram content on young women’s mood and body image: an intensive longitudinal study. New Media Soc 2021;25(12):3266-3288. [doi: 10.1177/14614448211038904]
39. Pilot IG, Stutts LA. The impact of a values affirmation intervention on body dissatisfaction and negative mood in women exposed to fitspiration. Body Image 2023;44:36-42. [doi: 10.1016/j.bodyim.2022.11.005] [Medline: 36455512]
40. Seekis V, Bradley GL, Duffy AL. Does a Facebook-enhanced mindful self-compassion intervention improve body image? An evaluation study. Body Image 2020;34:259-269. [doi: 10.1016/j.bodyim.2020.07.006] [Medline: 32717627]

XSL•FO
RenderX

41. Garbett KM, Haywood S, Craddock N, Gentili C, Nasution K, Saraswati LA, et al. Evaluating the efficacy of a social media-based intervention (Warna-Warni Waktu) to improve body image among young indonesian women: parallel randomized controlled trial. J Med Internet Res 2023;25:e42499 [FREE Full text] [doi: 10.2196/42499] [Medline: 37010911]
42. Chandler J, Cumpston M, Li T, Page M, Welch V. Cochrane Handbook for Systematic Reviews of Interventions. In: Higgins JPT, Thomas J. editors. Newark. University of Bristol (UK) and Thomas at UCL Institute of Education (UK): John Wiley & Sons; 2019.
43. Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, editors. Cochrane Handbook for Systematic Reviews of Interventions. 2nd Edition. Chichester (UK): John Wiley & Sons; 2019.
44. Persson DR, Ramasawmy M, Khan N. Framework for microintervention software technology in digital health: critical interpretive synthesis. J Med Internet Res 2025;27(1):e72658.
45. Atkinson MJ, Diedrichs PC. Examining the efficacy of video-based microinterventions for improving risk and protective factors for disordered eating among young adult women. Int J Eat Disord 2021;54(5):708-720. [doi: 10.1002/eat.23460] [Medline: 33415790]
46. Cohen R, Irwin L, Newton-John T, Slater A. #bodypositivity: a content analysis of body positive accounts on Instagram. Body Image 2019;29:47-57. [doi: 10.1016/j.bodyim.2019.02.007] [Medline: 30831334]
47. Baumel A, Fleming T, Schueller SM. Digital micro interventions for behavioral and mental health gains: core components and conceptualization of digital micro intervention care. J Med Internet Res 2020;22(10):e20631 [FREE Full text] [doi: 10.2196/20631] [Medline: 33118946]
48. Yim SH, Schmidt U. Experiences of computer-based and conventional self-help interventions for eating disorders: a systematic review and meta-synthesis of qualitative research. Int J Eat Disord 2019;52(10):1108-1124. [doi: 10.1002/eat.23142] [Medline: 31343088]
49. Sardi L, Idri A, Fernández-Alemán JL. A systematic review of gamification in e-Health. J Biomed Inform 2017;71:31-48 [FREE Full text] [doi: 10.1016/j.jbi.2017.05.011] [Medline: 28536062]
50. Corepal R, Best P, O'Neill R, Tully MA, Edwards M, Jago R, et al. Exploring the use of a gamified intervention for encouraging physical activity in adolescents: a qualitative longitudinal study in Northern Ireland. BMJ Open 2018;8(4):e019663 [FREE Full text] [doi: 10.1136/bmjopen-2017-019663] [Medline: 29678971]
51. Kristoffersen M, Johnson C, Atkinson MJ. Feasibility and acceptability of video-based microinterventions for eating disorder prevention among adolescents in secondary schools. Int J Eat Disord 2022;55(11):1496-1505 [FREE Full text] [doi: 10.1002/eat.23781] [Medline: 35851686]
52. Rodgers RF, McLean SA, Paxton SJ. Longitudinal relationships among internalization of the media ideal, peer social comparison, and body dissatisfaction: implications for the tripartite influence model. Dev Psychol 2015;51(5):706-713. [doi: 10.1037/dev0000013] [Medline: 25751099]
53. Cowles E, Guest E, Slater A. Imagery versus captions: The effect of body positive Instagram content on young women's mood and body image. Body Image 2023;44:120-130. [doi: 10.1016/j.bodyim.2022.12.004] [Medline: 36563473]
54. Rutter H, Campoverde C, Hoang T, Goldberg SF, Berenson KR. Self-compassion and women's experience of social media content portraying body positivity and appearance ideals. Psychol Pop Media 2024;13(1):12-22. [doi: 10.1037/ppm0000453]
55. Selensky JC, Carels RA. Weight stigma and media: An examination of the effect of advertising campaigns on weight bias, internalized weight bias, self-esteem, body image, and affect. Body Image 2021;36:95-106. [doi: 10.1016/j.bodyim.2020.10.008] [Medline: 33217717]
56. Bratman GN, Olvera‐Alvarez HA, Gross JJ. The affective benefits of nature exposure. Social & Personality Psych 2021;15(8):e12630. [doi: 10.1111/spc3.12630]
57. Swami V, Barron D, Todd J, Horne G, Furnham A. Nature exposure and positive body image: (re-)examining the mediating roles of connectedness to nature and trait mindfulness. Body Image 2020;34:201-208 [FREE Full text] [doi: 10.1016/j.bodyim.2020.06.004] [Medline: 32604023]
58. Naor L, Mayseless O. The therapeutic value of experiencing spirituality in nature. Spiritual Clin Pract 2020;7(2):114-133. [doi: 10.1037/scp0000204]
59. Tiggemann M, Zaccardo M. 'Strong is the new skinny': a content analysis of #fitspiration images on Instagram. J Health Psychol 2018;23(8):1003-1011. [doi: 10.1177/1359105316639436]
60. Singhal A, Cody M, Rogers E, Sabido M. Entertainment-Education and Social Change: History, Research, and Practice. New York: Routledge; 2003.
61. Shen F, Han J(. Effectiveness of entertainment education in communicating health information: a systematic review. Asian J Commun 2014 Jul 15;24(6):605-616 [FREE Full text] [doi: 10.1080/01292986.2014.927895]
62. Nightingale BA, Cassin SE. Self-Compassion may have benefits for body image among women with a higher body mass index and internalized weight bias. Healthcare (Basel) 2023;11(7):970 [FREE Full text] [doi: 10.3390/healthcare11070970] [Medline: 37046897]
63. Cadena SBR, Iyer A, Webb TL, Millings A. Understanding the relationship between self‐compassion and body dissatisfaction. J Applied Social Pyschol 2024;54(11):669-682. [doi: 10.1111/jasp.13065]

XSL•FO
RenderX

64. Barbeau K, Guertin C, Boileau K, Pelletier L. The effects of self-compassion and self-esteem writing interventions on women's valuation of weight management goals, body appreciation, and eating behaviors. Psychol Women Q 2021;46(1):82-98. [doi: 10.1177/03616843211013465]
65. Stern NG, Engeln R. Self-Compassionate writing exercises increase college women's body satisfaction. Psychol Women Q 2018;42(3):326-341. [doi: 10.1177/0361684318773356]
66. Cohen R, Fardouly J, Newton-John T, Slater A. #BoPo on Instagram: an experimental investigation of the effects of viewing body positive content on young women's mood and body image. New Media Soc 2019;21(7):1546-1564. [doi: 10.1177/1461444819826530]
67. Festinger L. A theory of social comparison processes. Hum Relat 1954;7(2):117-140. [doi: 10.1177/001872675400700202]
68. Nguyen HQ, Cuenco D, Wolpin S, Benditt J, Carrieri-Kohlman V. Methodological considerations in evaluating eHealth interventions. Can J Nurs Res 2007;39(1):116-134. [Medline: 17450708]
69. Moore THM, Higgins JPT, Dwan K. Ten tips for successful assessment of risk of bias in randomized trials using the RoB 2 tool: early lessons from Cochrane. Cochrane Evid Synth Methods 2023;1(10):e12031. [doi: 10.1002/cesm.12031]
70. Shah J, DePietro B, D'Adamo L, Firebaugh M, Laing O, Fowler LA, et al. Development and usability testing of a chatbot to promote mental health services use among individuals with eating disorders following screening. Int J Eat Disord 2022;55(9):1229-1244 [FREE Full text] [doi: 10.1002/eat.23798] [Medline: 36056648]
71. highly visual social media platforms to tackle unhealthy body image in adolescents and young adults: a systematic review protocol. Zenodo. URL: https://doi.org/10.5281/zenodo.15660997 [accessed 2026-01-20]

Abbreviations

BI: body image
HVSM: highly visual social media
MeSH: Medical Subject Headings
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Search Extension
RCT: randomized controlled trial

Edited by S Brini; submitted 04.Jul.2025; peer-reviewed by J Willoughby, B Jiménez-Alonso; comments to author 07.Oct.2025; accepted 11.Jan.2026; published 09.Feb.2026.

Please cite as:
Copez-Lonzoy A, Bazalar-Palacios J, Mayo-Puchoc N, Calderon de la Cruz G, Cruz-Ausejo L, Goicochea-Ríos E, Bazo-Alvarez JC Interventions That Use Highly Visual Social Media Platforms to Tackle Unhealthy Body Image in Adolescents and Young Adults: Systematic Review of Randomized Controlled Trials and Quasi-Experimental Studies
J Med Internet Res 2026;28:e80141
URL: https://www.jmir.org/2026/1/e80141
doi:10.2196/80141
PMID:

©Anthony Copez-Lonzoy, Janina Bazalar-Palacios, Nikol Mayo-Puchoc, Gustavo Calderon de la Cruz, Liliana Cruz-Ausejo, Evelyn Goicochea-Ríos, Juan Carlos Bazo-Alvarez. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Text-Based Depression Estimation Using Machine Learning With Standard Labels: Systematic Review and Meta-Analysis

Shengming Zhang[1], MSc, BSc; Chaohai Zhang[1], BSc; Jiaxin Zhang[1,2], PhD

[1]School of Automation and Intelligent Manufacturing, Southern University of Science and Technology, Shenzhen, Guangdong, China

[2]Guangdong Provincial Key Laboratory of Fully Actuated System Control Theory and Technology, School of Automation and Intelligent Manufacturing, Southern University of Science and Technology, Shenzhen, Guangdong, China

Corresponding Author:
Jiaxin Zhang, PhD
School of Automation and Intelligent Manufacturing
Southern University of Science and Technology
1088 Xueyuan Avenue, Nanshan District
Shenzhen, Guangdong, 518055
China
Phone: 86 13416141690
Email: zhangjx@sustech.edu.cn

Abstract

Background: Depression affects people's daily lives and even leads to suicidal behavior. Text-based depression estimation using natural language processing has emerged as a feasible approach for early mental health screening. However, most existing reviews often included studies with weak depression labels, which affected the reliability of the results and further limited the practical application of the automatic depression estimation models.

Objective: This review aimed to evaluate the predictive performance of text-based depression models that used standard labels, and to identify text resources, text representation, model architecture, annotation source, and reporting quality contributing to performance heterogeneity.

Methods: Following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines, we systematically searched 4 main databases (PubMed, Scopus, IEEE Xplore, and Web of Science) for studies published between 2014 and 2025. The eligible studies were included: machine learning models were developed based on the text generated by the participants and used validated scales or clinical diagnoses as depression labels. Pooled effect sizes (r) were calculated using random-effects meta-analysis with Hartung-Knapp-Sidik-Jonkman correction, and subgroup and meta-regression analyses were conducted to explore potential moderators.

Results: We scanned 3067 articles and finally filtered 15 models from 11 studies for the meta-analysis. The overall pooled effect size was 0.605 (95% CI 0.498-0.693), indicating a large strength of association. Subgroup analyses showed that models using embedding-based text representations achieved higher performance than those using traditional features (r=0.741, 95% CI 0.648-0.812 vs r=0.514, 95% CI 0.385-0.623; P<.001 for subgroup difference), and deep learning architectures outperformed shallow models (r=0.731, 95% CI 0.660-0.789 vs r=0.486, 95% CI 0.352-0.599; P<.001). Models trained with clinician diagnoses also outperformed better than those relying on self-report scales (r=0.688, 95% CI 0.554-0.787 vs r=0.500, 95% CI 0.340-0.631; P=.03). Reporting quality was positively associated with model performance (β=0.085, 95% CI 0.050-0.119; P<.001). Begg–Mazumdar and Egger tests provided no evidence of small-study effects. Begg–Mazumdar test (Kendall τ=0.17143, P=.37) and the Egger test (t_{14}=1.13401, 2-tailed P=.28) indicated no evidence of small-study effects.

Conclusions: Text-based depression estimation models trained with standard depression labels demonstrate solid predictive performance, with embedding features, deep model architectures, and clinician diagnosis labels showing significantly higher performance. Transparent reporting is also positively associated with model performance. This study highlights the importance of standard labels, feature representation, and reporting quality for improving model reliability. Unlike prior reviews that included weak or heterogeneous depression labels, this study offers more clinically reliable and comparable evidence. Moreover, this review provides clearer methodological guidance for developing more consistent and practically informative text-based depression screening models.

Trial Registration: PROSPERO CRD420251056902; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251056902

XSL•FO
RenderX

(J Med Internet Res 2026;28:e82686) doi:10.2196/82686

KEYWORDS

depression; natural language processing; standard labels; text; TRIPOD; Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

Introduction

Background

Depression, a type of mental disorder, is clinically characterized by persistent and significant low mood [1]. Individuals with severe depression often experience impaired social functioning and may even exhibit suicidal behaviors [2,3]. Currently, the diagnosis of depression relies on face-to-face consultation by psychiatrists and refers to the patients' self-reports. Although the clinical diagnosis by psychiatrists is regarded as the gold standard of depression detection [4], due to the time-constrained consultations and subjective information in psychiatry [5], the lack of clinical resources has prevented the gold standard from being widely used. To alleviate clinical burden, standardized self-reported scales have been developed based on psychiatrists' clinical experience [6], and several of them have been demonstrated to have comparable performance to clinical diagnosis [7,8]. However, these tools partially address the shortage of clinical resources; limitations such as subjective bias and time-consuming assessments remain [9-11]. With the development of artificial intelligence technologies, the limitations of traditional depression diagnosis methods facilitate the emergence of automatic depression estimation (ADE) models based on various multimodal data sources [12,13].

Among these multimodal data, text-based features have become a popular target in depression estimation studies [14-17]. Unlike other features used in ADE models [18-20], text (language) serves as a natural medium of human communication, conveying not only semantic content but also affective states [21,22]. The ability to capture emotional cues from text aligns closely with the depression diagnostic and may provide references for estimation [23]. For example, Cariola et al [24] found that mothers with depression used more first-person singular pronouns and present-focused words in mother-child dialogues, possibly reflecting heightened self-focus and introspection. Another study demonstrated that the ratio of negative to positive language played a vital role in establishing emotional tone [25]. Additionally, text data can be sourced from various contexts, including social media posts [26,27], SMS messages [28], chat logs [17], clinical transcripts [29], and even electronic health records [30]. Wang et al [27] achieved 91% classification accuracy based on social media posts, while Liu et al [28] reported an area under the curve of 76% on SMS messages. These studies achieved notable performance and supported the feasibility of text-based ADEs. However, differences in the reported contents (eg, text sources, model construction and validation, and depression labels) across studies may lead to substantial variability in model performance, thereby affecting the reliability of the results and the further practical application.

The quality of training data is critical to model performance, and the difference in text sources affects the performance of the ADE models [31]. The datasets of existing studies could be broadly categorized into two types: (1) public datasets, such as the Distress Analysis Interview Corpus – Wizard-of-Oz (DAIC-WOZ) [32], the Audio/Visual Emotion Challenge 2014 dataset [33], the Extended Audio-Transcript Depression Corpus [34], and the Chinese Multimodal Depression Corpus [35], which are collected under standardized protocols and provide a unified benchmark for comparing algorithm performance. For example, text-only models trained on DAIC-WOZ have shown gradual improvements in recent years [36,37]. The others belong to (2) self-constructed datasets, which are more diverse compared to public datasets in collection, and are usually used for exploring model feasibility in specific populations or contexts rather than comparing model algorithms. They include transcripts of clinical interviews or therapeutic dialogues, personal essays, diaries, questionnaires, etc. (see the previous paragraph for details). These datasets tend to better reflect natural language use in real-world contexts, which enhances the practical application ability of the ADE models [38]. For example, Cheng et al [39] analyzed social media (Weibo [Sina Corporation]) texts to predict depression and found that linguistic and cultural factors can affect the model performance. While Dogrucu et al [40] used text data via smartphones for developing Moodable, an Android-based depression sensing application developed at Worcester Polytechnic Institute, and validated this application for depression estimation. Overall, these examples illustrated the importance of self-constructed datasets, while the differences in text sources and contexts may contribute to heterogeneity in model performance.

After determining the dataset, model construction and validation are the core stages of ADE studies. Previous studies mainly relied on statistical methods to manually extract linguistic features. Rude et al [41] noted that individuals with depression frequently used negations and first-person singular pronouns, while another research examined the frequency of affective words such as "anger" or "sadness" to derive participants' emotion [42]. However, with the rapid development of natural language processing (NLP), embedding-based representations (eg, Word2Vec [43], Global Vector [44]) and pretrained models (eg, BERT [Bidirectional Encoder Representations from Transformers] [45], GPT [46]) have enhanced the semantic modeling capacity of ADE models. Niu et al [47] applied graph attention networks to capture hierarchical contextual semantics, and some studies used BERT to encode transcribed speech into context-aware sentence embeddings [48,49]. In model architecture, traditional shallow models, such as support vector machines, random forests, and decision trees, were commonly used in previous works [50], while deep learning models have also performed well in the ADE tasks recently [51-53]. Dinkel et al [51] built a multitask model based on a bidirectional gated recurrent unit, achieving an F_1-score of 0.84 on DAIC-WOZ. Martinez et al [52] further improved classification performance using RoBERTa. Furthermore, the validation strategies also vary: while single-holdout testing is commonly used when

XSL•FO
RenderX

datasets are large [54-56], small-sample medical texts often use repeated bootstrapping [57] or k-fold cross-validation [30,58] to ensure robustness. Overall, differences in feature representation, model architecture, and validation strategies all contribute to potential differences in model performance and generalization.

Another potential factor affecting the model performance is the quality of depression labels. Clinical diagnosis by certified psychiatrists is widely considered the gold standard of depression [1,59]. However, clinical diagnosis is costly and subject to the experience of the psychiatrists [60]. As a substitute, many text-based ADE studies used standardized self-report depression scales. Common scales include the Patient Health Questionnaire, 8 or 9 items (PHQ-8 or 9) [61], Beck Depression Inventory-II (BDI-II) [62], Zung Self-Rating Depression Scale (SDS) [63], 21-item Depression, Anxiety, and Stress Scale (DASS-21) [64], and Center for Epidemiological Studies Depression Scale (CES-D) [65]. The ADE tasks can be equal to predicting the self-reported scores. For example, Li et al [66] achieved an F_1-score of 78.3% on the DAIC-WOZ dataset using PHQ-8, and another study reported a classification accuracy of 69% using PHQ-9 labels [67]. However, structured scales demonstrate good reliability and validity, but their use is often limited by ethical constraints and annotation costs. Consequently, many studies resort to weak depression labels, such as keyword matching [68,69], sentiment analysis tools [70], or self-declared depressive status from social media [71]. These approaches facilitate large-scale data collection and exploration, but there may be deviations in the reliability and validity of practical applications. For example, keyword-based labels may ignore semantic context [72], and sentiment analysis tools may confuse sadness with clinical depression [73]. Therefore, increasing emphasis has been placed on adopting gold-standard labels for model training and validation [16,31,74]. In this study, we included only those studies that use either clinical diagnosis or the PHQ-9 scale as depression labels: Other scales, such as BDI-II and CES-D, though commonly used for screening, differ in target populations and sensitivity, which may impair cross-study comparability [75]. The potential impact of annotation source on model performance is examined in our Results section.

In recent years, several systematic reviews have investigated NLP-based depression estimation. Mao et al [31] provided a multimodal overview covering facial expression, speech, and text models, highlighting the lack of model transparency and limited practical application as critical challenges. Some reviews focus on specific methods. For example, Yao et al [16] summarized depression-related studies using social media text and found that machine learning (ML) and statistical analysis methods were more prevalent. Tahir et al [76] provided a comprehensive review of ML and deep learning approaches for ADE tasks based on social media data. Nanggala et al [77] conducted a systematic review of the performance of the transformer structure in ADE tasks and noted the competitive advantages of the text modalities. Many studies have examined the use of social media text for depression estimation, analyzing feature design and training strategies [16,78]. While these previous works provided valuable insights into this field, 2 major limitations remain: First, most reviews concentrate on specific text sources (eg, social media and public datasets), which may not generalize well to practical applications [79]. Second, annotation sources are often overlooked, and weak depression labels, such as self-declared or keyword matching, are often confused without discussing the impact of their differences from standard labels on model performance [72]. Therefore, this study uses label quality as a core inclusion criterion. We systematically assessed performance and sources of heterogeneity in text-based ADE models.

Research Aims and Structure

This review aims to evaluate the performance of text-based depression estimation models that use standard labels and to identify the potential moderators that may account for performance differences. We specifically analyzed variations in text sources, model architectures, validation strategies, annotation standards, as well as the quality of study reporting, to examine their potential influence on depression predictive performance. The structure of this review is as follows: the Introduction section outlines the background and research aims, the Methods section presents the study selection criteria, data extraction, and meta-analysis protocol, the Results section presents the outcomes of the meta-analysis, and the Discussion and Conclusion section interprets the findings and provides implications for future research and practical applications.

Methods

Study Design

This systematic review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) 2020 and the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) for reporting literature searches guidelines [80] (Multimedia Appendix 1). We also registered our review on PROSPERO (CRD20251056902). There were no deviations from the registered protocol.

Information Sources and Search Strategy

To include as many studies of text-based ADE as possible at the beginning, we systematically searched 4 datasets: Scopus, IEEE Xplore, Web of Science, and PubMed. The search strategy followed the PRISMA-S reporting extension [81] and was developed based on common model architectures, text features, and outcome formats used in ADE research. The search strategy included combinations of the following terms: (“depression” OR “depressive” OR “clinical depression” OR “major depressive disorder” OR “MDD”) AND (“assessment” OR “measur*” OR “diagnos*” OR “predict*” OR “estimat*”) AND (“automated” OR “automatic” OR “AI” OR “machine learning” OR “deep learning” OR “large language model” OR “natural language processing” OR “NLP”) AND (“text” OR “linguistic analys*” OR “sentiment analys*” OR “semantic analys*” OR “lexical feature*” OR “transcribed speech” OR “written” OR “textual*”). The initial search was conducted on February 24, 2025, covering studies published from January 2014 onward, and was updated on December 3, 2025. Full database-specific search strings are provided in Multimedia Appendix 2. Additional eligible studies

XSL•FO
RenderX

were identified by manually searching the reference lists of the included studies and previous reviews.

Eligibility Criteria

We included studies that fulfill the following four criteria: (1) used texts to investigate depression, (2) used ML to establish an ADE model, (3) reported the data collection process and the depression label sources, and (4) reported the metrics of the model for effect size calculation. Our exclusion criteria are the studies that (1) did not use the standard depression label (details in paragraph 5 of the introduction); (2) were not written in English; (3) published before 2024 and having fewer citations than years of publish were excluded (eg, articles published in 2021 with 3 citations, or articles published in 2019 with 5 citations would be included) [31], to ensure that full-text screening focused on studies that had been acknowledged and referenced within the field; and (4) were developed exclusively on public datasets, as these models often iterate rapidly, which could introduce potential bias.

Data Collection Process

Two independent researchers (SZ and JZ) excluded duplicates from the relevant studies retrieved, then screened and selected the relevant studies by inclusion and exclusion criteria. Disagreements were solved through discussion with a third author (CZ) when necessary. The relevant studies were evaluated through title and abstract screening. Thereafter, the remaining studies were screened through full texts to identify those for further analysis. Studies were managed using EndNote and Microsoft Excel. No automation tools were used.

Data Items

Two researchers (SZ and JZ) independently coded the included studies in Microsoft Excel. The following information was extracted from each article: (1) study characteristics (first author and publication year), (2) annotation source (clinician diagnosis and self-report scales [PHQ-9]), (3) population characteristics (sample size and positive rate), (4) modeling strategy (text sources [eg, interviews, social media, writing tasks, and diaries], text representation [eg, RoBERTa, transformer, TF-IDF, LIWC, and Empath], model architecture [deep learning vs shallow ML] and validation strategies [eg, cross-validation, hold-out, and external validation]), and (5) predictor values [sensitivity, specificity, F_1-score or data used to impute these values]. Furthermore, for studies that used different algorithms based on the same text dataset, only the model with the highest F_1-score was included in the meta-analysis. If a study reported results based on different text sources (eg, clinical transcripts vs written documents), the best-performing model from each dataset was included separately. For multimodal studies, only the performance metrics from the text modality were included.

In addition, we assessed the reporting quality of each included study using the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) tool [82]. TRIPOD was adapted to be more specific to NLP studies: The modifications were based on previously published research of NLP studies [83,84]. Two independent researchers (SZ and JZ) independently assessed each study based on TRIPOD items; disagreements were resolved through a third author (CZ).

Effect Measures

For each study, the sensitivity, specificity, and sample sizes were used to calculate the effect size (ES). Specifically, the odds ratio (OR) was first computed by sensitivity and specificity, followed by a log transformation to obtain log (OR) [85]. The SE of log (OR) was then calculated based on the sensitivity, specificity, and sample size. Log (OR) and SE (log OR) were converted into ES by using Comprehensive Meta-Analysis (version 4; Biostat Inc). To direct comparison of model performance across studies, Pearson correlation coefficient (r), transformed via Fisher z for analysis and subsequently back-transformed for interpretation, was selected as the ES indicator [86]. Furthermore, the Pearson correlation coefficient contributes to a general understanding of the relationship between text-based models' predictions and true labels. Values of r around 0.1, 0.3, and 0.5 are generally interpreted as indicating small, moderate, and large effect sizes, respectively. These thresholds have been widely applied in psychological and behavioral research, including prior reviews of depression prediction models [78,87].

Data Synthesis and Analysis

Meta-analysis was initially performed using Comprehensive Meta-Analysis (version 4; Biostat Inc). Between-study heterogeneity was assessed using the chi-square Q statistic, which tests the null hypothesis of homogeneity [88]. The extent of heterogeneity was further described using the inconsistency index (I^2), along with the between-study variance (τ^2) and its SD (τ) [89]. In addition, 95% prediction intervals were reported to reflect the expected range of effects in future studies [90]. We adopted the random-effects model to assess the heterogeneity among studies and estimate the pooled effect size from each study. Given the limited number of included studies, the Hartung-Knapp-Sidik-Jonkman (HKSJ) method was applied to adjust the SEs [91]. Because CMA 4.0 does not implement the HKSJ procedure, we reestimated the pooled effects using the Meta package in R (version 4.3.2), and the 95% CI values reported in the main text were derived from the HKSJ-adjusted models.

To explore potential sources of heterogeneity, we performed a moderator analysis with a random effects model [92]. According to the extracted information from each study, four predefined groups were included: (1) text representation (embedding-based vs traditional features), (2) annotation source (clinician diagnosis vs self-reported scale), (3) model architecture (deep vs shallow), and (4) text source (documentation vs transcribed speech). For each subgroup, pooled ES (r) were estimated under a random-effects model, and between-group differences were assessed using the Q-test for heterogeneity. In addition, univariate meta-regression analyses were conducted to examine the potential moderating effects of 3 continuous covariates: reporting quality (TRIPOD score), sample size (log-transformed), and positive rate. Each covariate was analyzed separately with a random-effects model, and an analog R^2 indicated its explanatory power for study variance. Finally,

XSL•FO
RenderX

we performed a sensitivity analysis using the leave-one-out method to assess the influence of each study on the overall results. We also conducted a cumulative meta-analysis to explore whether more recent studies have contributed to increased consistency.

Small-Study Effects

We assessed potential small-study effects, including the possibility of publication bias, by visually inspecting a funnel plot of the SE Fisher *z* [93]. Statistical evaluation of funnel plot asymmetry was conducted using the Begg–Mazumdar rank correlation test and the Egger regression test, with a *P* value <.05 indicating significant small-study effects [94]. In addition, we applied the Duval and Tweedie trim-and-fill procedure to explore the potential impact of missing studies on the pooled effect size [95].

Quality and Certainty Assessment

The methodological risk of bias of each included study was independently assessed by 2 authors (SZ and CZ) across five domains relevant to text-based ML research: (1) participant selection, (2) outcome labeling, (3) text acquisition and preprocessing, (4) model development and validation, and (5) reporting completeness. Each domain was rated as "low risk," "unclear risk," or "high risk" and an overall judgment was derived accordingly. The certainty of evidence for the primary pooled effect was evaluated using the GRADE (Grading of Recommendations Assessment, Development and Evaluation) framework [96], considering risk of bias, inconsistency, indirectness, imprecision, and publication bias.

Results

Study Selection Process

The flow diagram of studies screening and selection is presented in Figure 1. The initial database search was conducted on February 24, 2025, and was updated on December 3, 2025.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram of studies included in this review.

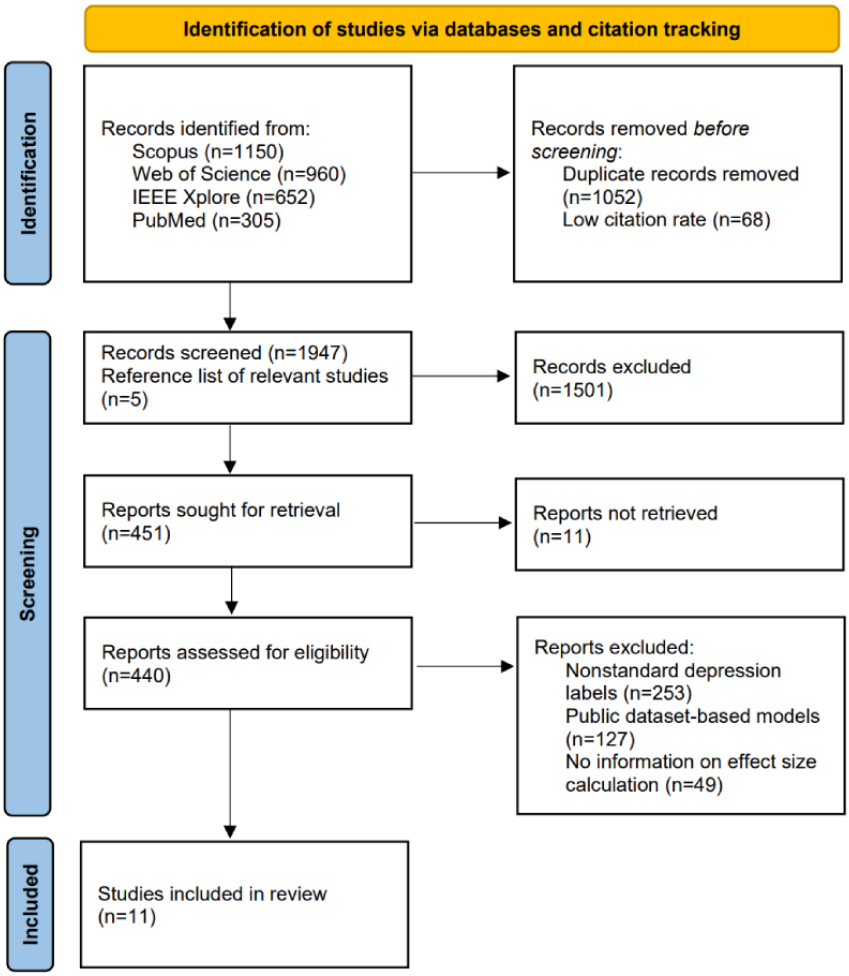

A total of 3067 records were identified through database searching, including Web of Science (n=960), PubMed (n=305), Scopus (n=1150), and IEEE Xplore (n=652). After removing 1052 duplicates and 68 low-citation records, 1947 records remained for title and abstract screening, together with 5 additional records identified through manual citation tracking. Among them, 1501 records were excluded because they were not relevant to text-based depression estimation, including that the model inputs were not text data and no quantitative results were reported. The remaining 451 records were included in the full-text screening, and 11 studies that were inaccessible or retracted were first excluded. Next, we excluded studies that

XSL•FO
RenderX

used nonstandard depression labels (n=253), studies based on public datasets (n=127), and studies lacking sufficient information to compute effect sizes (n=49). Finally, 11 studies [17,24,26,67,97-103] were included in the meta-analysis.

Characteristics of Included Studies

A total of 11 studies were included in the final meta-analysis [17,24,26,67,97-103], contributing 15 independent text-based depression estimation models. A summary of included studies and model characteristics is presented in Table 1, and the detailed data extraction table is provided in Multimedia Appendix 3. The sample size ranged from 77 to 749 participants, with a median of 110. Positive rates (proportion of participants labeled as depressed) varied substantially across studies, ranging from 9.2% to 77.9%, reflecting differences in population selection and depression labeling strategies.

Table 1. Summary of study and model characteristics included in the meta-analysis (n=15).

Author (year)	Sample size[a]	Positive rate	TRIPOD	Text representation	Annotation source	Model architecture	Text source	Sensitivity	Specificity
Geraci et al (2017) [97]	366	0.243	17	Traditional features	Clinician diagnosis	Deep	Documentation	0.94	0.68
Ricard et al (2018) [26]	749	0.0921	13	Traditional features	Self-report scale	Shallow	Documentation	0.57	0.77
Tlachac et al (2020) [98]	162	0.3395	12	Traditional features	Self-report scale	Shallow	Documentation	0.93	0.63
Zhao et al (2021) [67]	110	0.4	11	Traditional features	Self-report scale	Shallow	Documentation	0.33	0.86
Zhao et al (2021) [67]	114	0.614	11	Traditional features	Self-report scale	Shallow	Documentation	0.81	0.53
Zhao et al (2021) [67]	341	0.463	11	Traditional features	Self-report scale	Shallow	Documentation	0.51	0.86
Shin et al (2022) [99]	166	0.5	16	Traditional features	Clinician diagnosis	Shallow	Transcribed speech	0.7	0.97
Cariola et al (2022) [24]	140	0.514	9	Traditional features	Clinician diagnosis	Shallow	Transcribed speech	0.68	0.66
Munthuli et al (2023) [100]	80	0.5	15	Embedding-based	Clinician diagnosis	Deep	Transcribed speech	0.83	0.85
Munthuli et al (2023) [100]	80	0.5	15	Embedding-based	Clinician diagnosis	Deep	Transcribed speech	0.88	0.93
Munthuli et al (2023) [100]	80	0.5	15	Embedding-based	Clinician diagnosis	Deep	Transcribed speech	0.9	0.83
Tlachac et al (2023) [101]	88	0.602	13	Traditional features	Self-report scale	Shallow	Documentation	0.79	0.74
Jihoon et al (2024) [102]	77	0.779	17	Traditional features	Clinician diagnosis	Shallow	Transcribed speech	0.96	0.25
Shin et al (2024) [17]	91	0.171	14	Embedding-based	Self-report scale	Deep	Documentation	0.929	0.761
Xu et al (2025) [103]	100	0.55	19	Embedding-based	Clinical diagnosis	Deep	Transcribed speech	0.955	0.933

[a]Validation strategies include k-fold cross-validation, leave-one-out, and nested validation.

Sample size refers to the total number of participants included in each model.

Positive rate refers to the proportion of participants labeled as depressed.

TRIPOD score was based on a modified 27-item checklist adapted for NLP-based studies.

Text representations include traditional features (eg, Lexical, TF-IDF, LWIC, and Emotional Dictionary) and

XSL•FO
RenderX

embedding-based features (eg, BERT, Word2Vec, and RoBERTa).

Annotation source indicates whether the depression label was derived from clinician diagnosis or self-report scales (Patient Health Questionnaire-9, PHQ-9).

Model architecture refers to shallow learning (eg, SVM, Logistic Regression, and Random Forests) versus deep learning (eg, GRU, BERT, Transformers).

Text sources were classified as documentation (eg, Writing, Social media, and Messages) or transcribed speech (eg, Clinical Interviews and Psychological Conversations).

Regarding some characteristics of the modeling, 8 (53.3%) models used clinician diagnosis as depression labels, and the remaining 7 (46.7%) models relied on self-report scales such as the PHQ-9. In total, 10 (66.7%) models used traditional features to text representation, such as TF-IDF, LIWC, or emotional dictionaries, while 5 (33.3%) models used embedding-based techniques, including BERT and Word2Vec. 9 (60%) models used shallow learning models (eg, SVM, random forest, and logistic regression), and the remaining 6 (40%) models adopted deep learning architectures (eg, GRU and BERT-based classifiers). Text sources were classified as either documentation (n=8, 53.3%) or transcribed speech (n=7, 46.7%), the latter often derived from clinical interviews or therapeutic dialogues. Validation strategies also varied: 11 models implemented some form of k-fold cross-validation, while others applied nested or leave-group-out cross-validation.

The reporting quality (mean 13.9, SD 2.8; range 9-19) of all included models was assessed using an adapted 21-item TRIPOD checklist for NLP-based predictive modeling. Detailed scoring criteria and individual scores are provided in Multimedia Appendix 4.

Results of Meta-Analysis

A random-effects model was applied to synthesize the effect sizes of 15 text-based depression estimation models extracted from 11 eligible studies [17,24,26,67,97-103]. In studies reporting multiple models, each model was derived from a distinct text dataset with nonoverlapping participant samples and was therefore treated as a point estimate [104]. The pooled effect size (r) was 0.605 (95% CI 0.498-0.693) [87], indicating an overall high predictive ability in models using standard depression labels (Figure 2). Substantial between-model heterogeneity was observed (Q=99.02, P<.001; I^2=85.9%). The estimated between-study variance was τ^2=0.062 (τ=0.249). The 95% prediction interval ranged from 0.140 to 0.851, indicating considerable variability in the magnitude of effects expected across future studies. The forest plot revealed that each model achieved significant correlations (P<.05), the effect sizes varied across models (ranging from 0.292 to 0.776), further supporting the observed heterogeneity. The HKSJ-adjusted forest plot generated in R is provided in Multimedia Appendix 5. This variation highlights the influence of study-level characteristics such as model architecture, annotation source, and text representation on model performance, which were further analyzed in subsequent moderator and meta-regression analyses.

Figure 2. Forest plot presenting the pooled effect size (correlation r) of machine-learning models for text-based depression estimation trained with gold-standard labels [17,24,26,67,97-103]. For studies appearing more than once in the forest plot (eg, Zhao et al, 2021 [67]; Munthuli et al, 2023 [100]), indicate models trained and evaluated on independent text datasets derived from nonoverlapping participant samples.

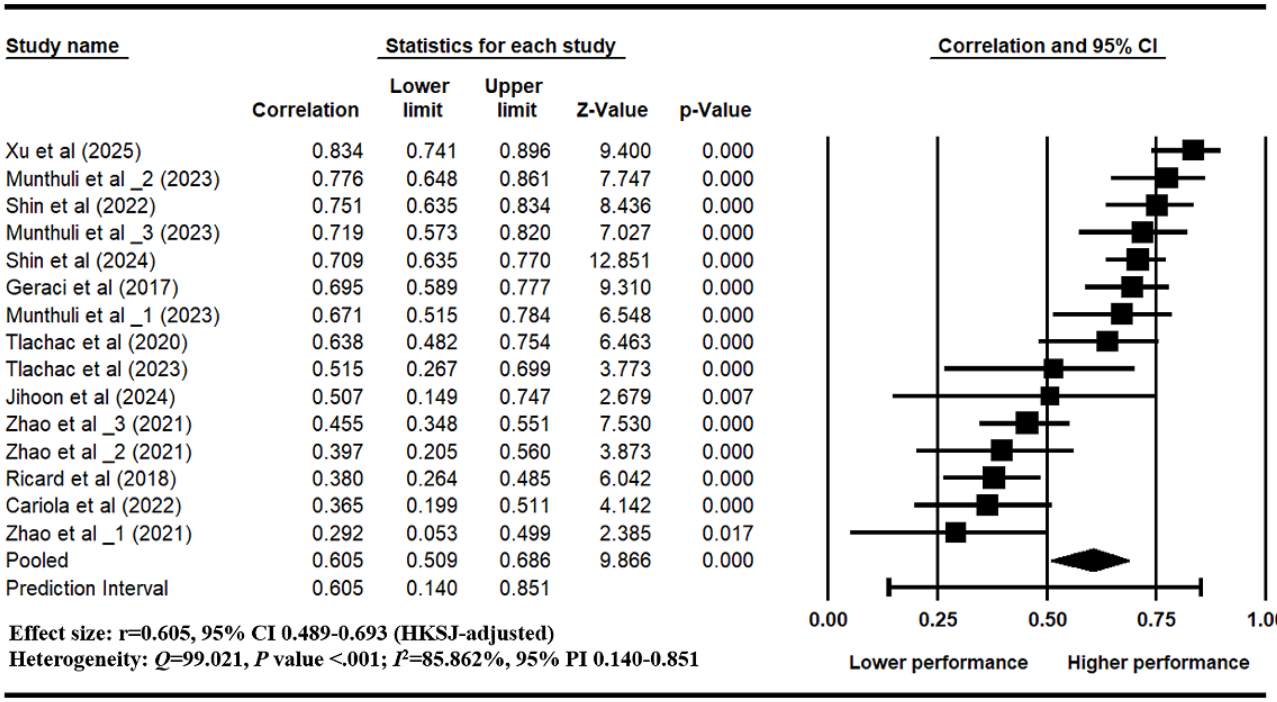

Results of Subgroup Analysis for Moderators

Subgroup analyses were conducted to examine the moderating effects of text representation, annotation source, model architecture, and text source (Table 2). Significant moderation effects were observed for text representation (Q=16.47, P<.001) and model architecture (Q=22.60, P<.001). Models using embedding-based features yielded higher performance (r=0.741,

XSL•FO
RenderX

95% CI 0.648-0.812) compared with those using traditional features (*r*=0.514, 95% CI 0.385-0.623). Deep learning architectures also outperformed shallow models (*r*=0.731, 95% CI 0.660-0.789 vs *r*=0.486, 95% CI 0.352-0.599). Annotation source showed a statistically significant moderation effect (*Q*=5.00, *P*=.03), with models using clinician diagnoses achieving higher pooled performance (*r*=0.688, 95% CI 0.554-0.787) than those using self-report scales (*r*=0.500, 95% CI 0.340-0.631). Text source did not reach statistical significance (*Q*=3.00, *P*=.08), although models using transcribed speech tended to outperform those using documentation (r=0.687 vs 0.529). Forest plots of these subgroup analyses are presented in Figure 3, and HKSJ-adjusted subgroup plots generated in R are provided in Multimedia Appendix 5.

Table 2. Subgroup analyses examining the influence of text representation, annotation source, model architecture, and text source on model performance in text-based depression estimation.

Moderators	Models, n (%)	Point estimation (95% CI)	*Q*-value (*df*)	*P* value
Text representation	15 (100)	—[a]	16.472 (1)	<.001
Embedding-based	5 (33.3)	0.741 (0.648-0.812)	—[a]	<.001
Traditional features	10 (66.7)	0.514 (0.385-0.623)	—*	<.001
Annotation source	15 (100)	—[a]	4.996 (1)	.03
Clinician diagnosis	8 (53.3)	0.688 (0.554-0.787)	—[a]	<.001
Self-report scale	7 (46.7)	0.500 (0.340-0.631)	—[a]	<.001
Model architecture	15 (100)	—[a]	22.595 (1)	<.001
Deep	6 (40)	0.731 (0.660-0.789)	—[a]	<.001
Shallow	9 (60)	0.486 (0.352-0.599)	—[a]	<.001
Text source	15 (100)	—[a]	3.003 (1)	.08
Documentation	8 (53.3)	0.529 (0.381-0.650)	—[a]	<.001
Transcribed speech	7 (46.7)	0.687 (0.521-0.803)	—[a]	<.001

[a]Not applicable.

XSL•FO
RenderX

Figure 3. Forest plots of subgroup analyses by (A) text representation, (B) annotation source, (C) model architecture, and (D) text source generated using CMA4.0 [17, 24, 26, 67, 97, 98, 99, 100, 101, 102, 103]. High-resolution versions of all subgroup forest plots are provided in Multimedia Appendix 5.

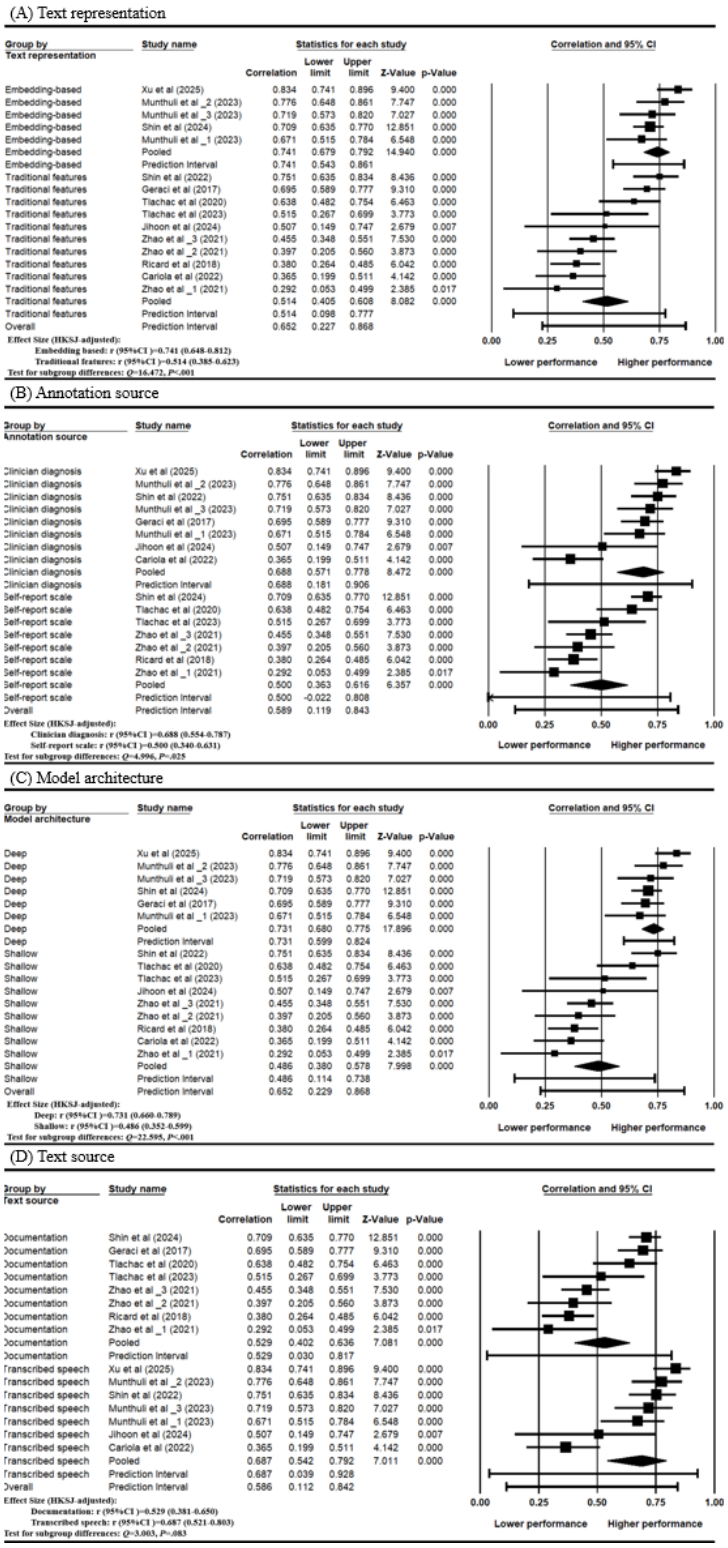

Univariable meta-regression analysis was used to test 3 continuous moderators: TRIPOD score, the positive rates, and the log-transformed sample size (log n). All meta-regressions were conducted using Fisher *z*-transformed correlation coefficients under a random-effects model. As shown in Table 3, only the TRIPOD score was significantly associated with model performance (β=0.085, 95% CI 0.050-0.119; *P*<.001), indicating a positive association between reporting quality and effect size. This model explained 71% of the between-study variance (R^2 analog=0.71). In contrast, neither the positive rate (β=–0.027, 95% CI –0.830 to 0.884; *P*=.95) nor the sample size (β=–0.012, 95% CI –0.214 to 0.190; *P*=.91) showed significant associations with effect size. Corresponding regression plots

XSL•FO
RenderX

are included in Multimedia Appendix 6 (CMA-based) and Multimedia Appendix 5 (HKSJ-adjusted).

Table 3. Univariable meta-regression assessing reporting quality (TRIPOD), positive rate, and sample size as moderators of model performance.

Moderators	N	β (95% CI)	SE	R^2	*P* value
TRIPOD[a] NLP[b]	15	0.085 (0.050-0.119)	0.018	0.71	<.001
Positive Rate	15	–0.027 (–0.830 to 0.884)	0.437	<0.2	.95
Log n	15	–0.012 (–0.214-0.190)	0.103	<0.2	.91

[a]TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

[b]NLP: natural language processing.

Risk of Bias and Certainty Assessment

Risk-of-bias assessment showed that most included models demonstrated low methodological concerns. Of the 15 models, 7 were rated as low risk, 6 as unclear risk, and 2 as high risk, with the latter typically related to concerns in outcome measurement or reporting (Multimedia Appendix 7). The overall certainty of evidence for the primary pooled effect was rated as moderate under the GRADE framework. This rating was driven mainly by the substantial heterogeneity observed across models, while concerns regarding imprecision and publication bias were minimal. Detailed domain ratings and justification for GRADE decisions are provided in Supplementary Table S4 (Multimedia Appendix 7).

Sensitivity Analysis and Cumulative Analysis

Leave-one-out analysis indicated that the pooled effect size remained stable across all iterations (*r*=0.605, 95% CI 0.498-0.693), suggesting that no single study had a disproportionate impact on the overall result (Figure 4). Cumulative meta-analysis showed that the effect estimates gradually stabilized over time, indicating consistency across publication years (Figure 5).

Figure 4. Leave-one-out sensitivity analysis [17,24,26,67,97-103].

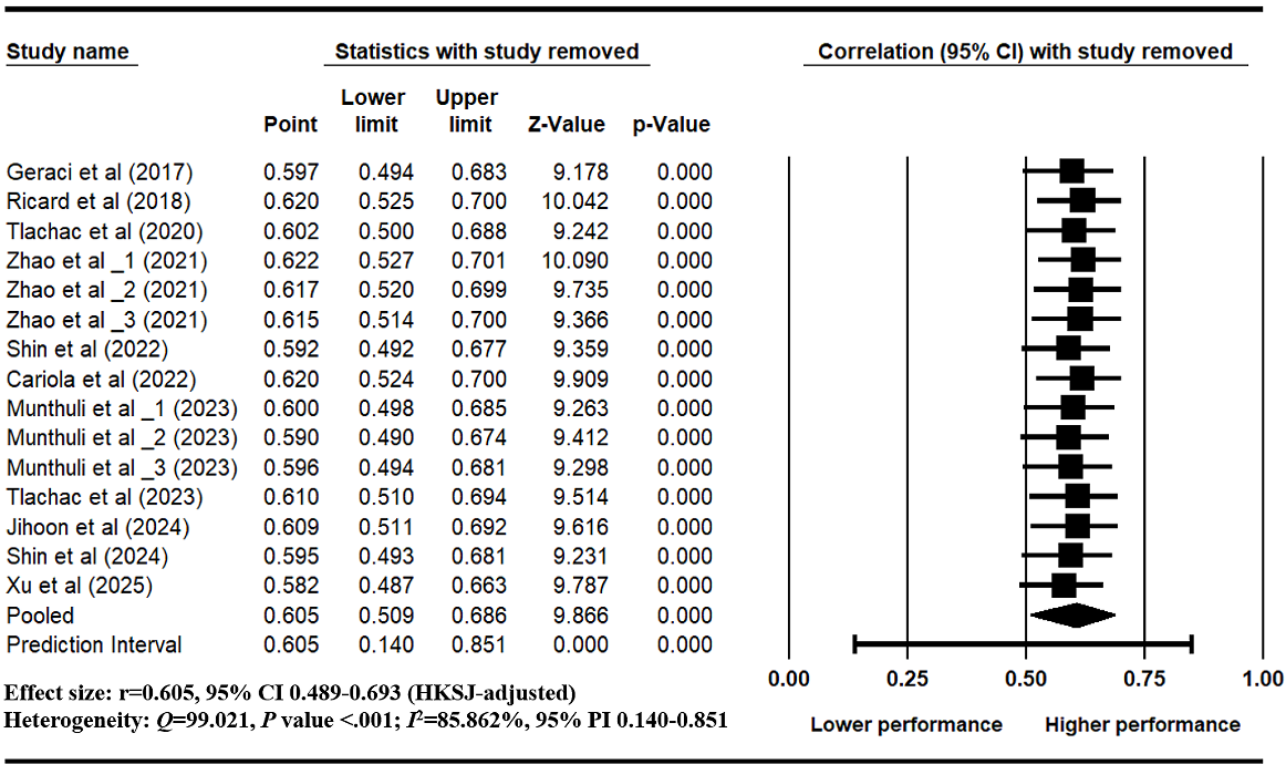

XSL•FO
RenderX

Figure 5. Cumulative meta-analysis [17,24,26,67,97-103].

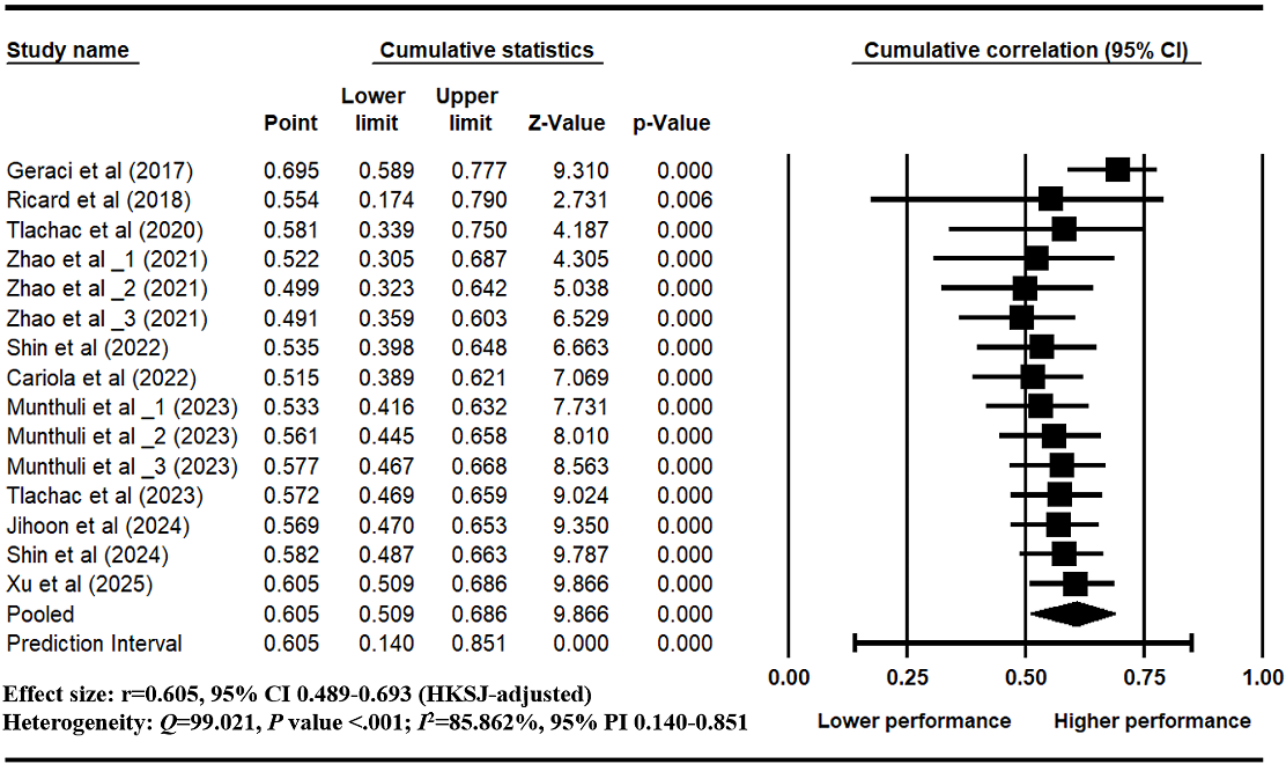

Small-Study Effects

The funnel plot (Figure 6) showed a generally symmetrical distribution of effect sizes. Both the Begg–Mazumdar test (Kendall τ=0.17143, P=.37) and the Egger regression test (t_{14}=1.13401, 2-tailed P=.28) were nonsignificant, indicating no evidence of small-study effects. The trim-and-fill procedure did not impute any additional studies, and the adjusted pooled effect was identical to the original estimate.

Figure 6. Funnel plot.

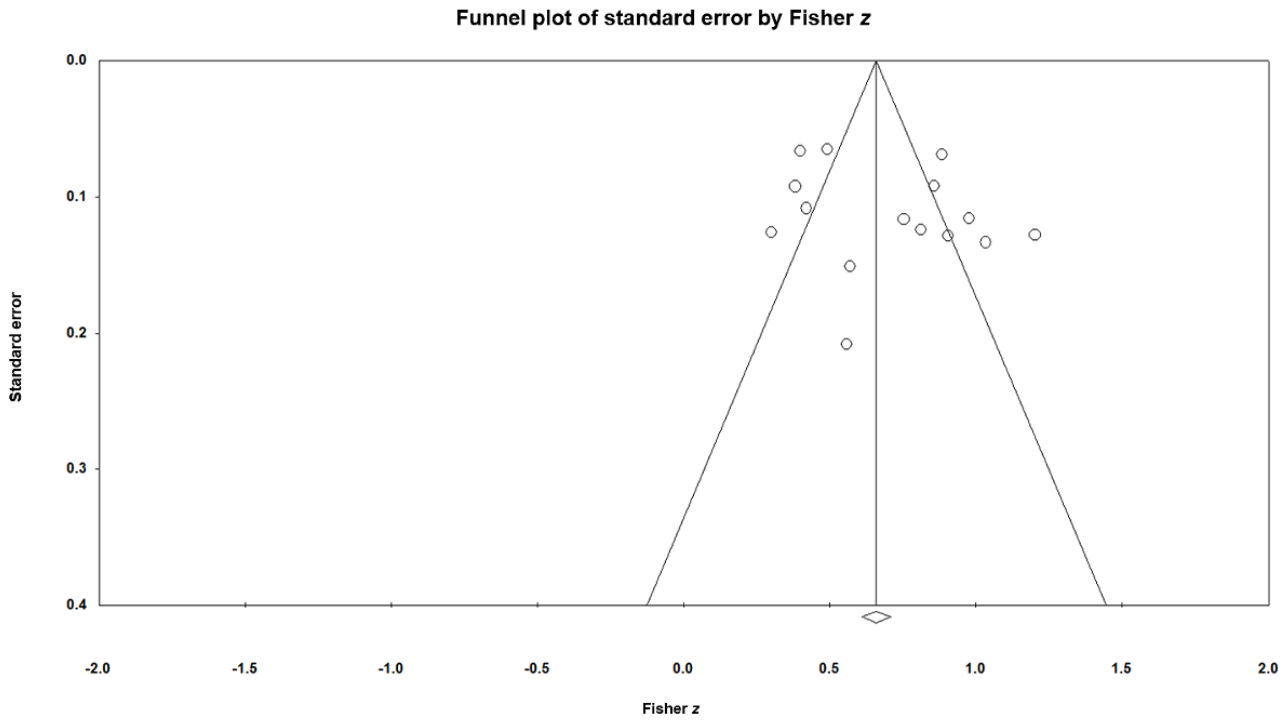

XSL•FO
RenderX

Discussion

Principal Findings

This study evaluated the performance of text-based depression estimation models that were trained and validated using standard depression labels. Our findings indicate that text-based ADE models demonstrate promising predictive performance. A previous review study mentioned the importance of a uniform standard label when evaluating the consistency of model performance across studies [78]. In this study, we focused on the studies using clinical diagnoses or PHQ-9 and excluded weak depression labels datasets (eg, keyword matching and self-declared), thereby enhancing the clinical reliability of the results [72,73]. These findings support the potential of text, especially that in transcribed speech and personal documentation, as informative signals for depression detection [23-25]. Nonetheless, the high between-study heterogeneity implies the differences in study-level characteristics, including modeling approaches, corpus sources, and reporting quality.

For the text representation and model architecture, the results of subgroup analyses showed that models using embedding-based text representations achieved higher performance than those using traditional features, and deep learning architectures outperformed shallow models. Consistent results have been reported on commonly used public benchmarks [105,106]. On the DAIC-WOZ dataset, the high-performing models often use embedding-based representation and deep model architecture [37,107]. This result aligns with the development of context-aware models like BERT [45] and GPT [46], which have superior semantic representation capabilities. A previous review study on emotion detection also reported similar conclusions [22]. However, the traditional linguistic features remain critical in the field of ADEs [16]. Linguistic features can be extracted from depressive texts by topic modeling techniques [108], which provide insights into underlying cognitive and emotional states and contribute to clinical understanding [109,110]. Overall, embedding-based and deep model architectures are superior in prediction performance, and the traditional features and shallow models remain valuable for enhancing interpretability and depression understanding.

For the annotation sources, the results of subgroup analyses indicated that models trained with clinician diagnoses also outperformed better than those relying on self-report scales. This supported the view expressed by Mao et al [31] that self-reported measures may not align with clinical diagnoses, potentially leading to inconsistency in model performance. Another study also highlighted the heterogeneity of depression estimation caused by subjectivity [111]. In addition, the results indicate that while models trained with clinically validated scales like PHQ-9 remain usable, their performance is generally lower than models trained with clinician diagnoses. For the text sources, models using transcribed speech were slightly higher than those using personal documentation, but this difference did not reach a statistically meaningful level. This may reflect the more information and affective cues embedded in spoken language [38], but some studies have pointed out that there are also noises such as stop words that are not directly related to emotion [112,113]. Overall, these results suggest that the annotation source demonstrated a determinative impact on model performance, with clinician diagnoses outperforming PHQ-9, whereas the text source showed only directional effects. We suggest constructing the ADE models based on the gold standard and natural environment-related text data. However, when clinician diagnoses are unavailable, validated self-report scales represent feasible alternatives.

It is worth noting that there is a significant positive correlation between the model report quality (TRIPOD score) and the model performance. This indicates that studies with higher report quality exhibit better performance in the ADE model. The TRIPOD checklist has been previously adopted in other automatic estimation domains [114,115]. This result further emphasized the necessity for comprehensive reporting of study-level characteristics, including modeling approaches and corpus sources, to enhance reliability and validity in ADE models [82]. In contrast, sample size and positive rates were not significantly associated with model performance. It is consistent with the previous review studies' results [76,78], suggesting that in the ADE tasks, the data quality, annotation, and methodological transparency may be more crucial than the quantity of the samples. This finding underscores the importance of rigorous methodological reporting for ADE studies.

Taken together, the findings of this meta-analysis highlight several methodological and reporting considerations that are relevant for the development and evaluation of text-based ADEs. Specifically, models using embedding-based features and deep architectures are superior in predictive performance, suggesting that future studies could prioritize richer text representations combined with more sophisticated model architectures. Adopting standardized depression labels, particularly clinician diagnoses or widely validated scales, may further facilitate comparability across studies. Notably, we found that there is a significant positive correlation between the TRIPOD score and the model performance. This indicates that improving adherence to established reporting guidelines could enhance the transparency and reproducibility of ADE models. The overall certainty of evidence was rated as moderate using the GRADE approach, reflecting acceptable confidence in the pooled findings despite substantial between-study heterogeneity.

Strengths and Limitations

The primary strength of this review lies in its strict focus on text-based ADE models trained and validated using standard depression labels. By excluding studies based on weak or inconsistent annotation sources, such as keyword matching or self-declared labels, we reduced label-related heterogeneity and improved the clinical interpretability of the synthesized findings. In addition, this meta-analysis systematically examined multiple methodological sources of heterogeneity, including text representation, model architecture, annotation source, and reporting quality, using a unified analytical framework combining subgroup analyses and meta-regression.

Nevertheless, this study also has some limitations. First, our inclusion criteria strictly require the use of standard depression labels. Although this enhanced the reliability of the labels, it

XSL•FO
RenderX

also reduced the number of eligible studies (57.5% of the literature was excluded in the full-text assessment). Some of the latest studies were also not included, including large language models [116] and the prompt learning strategy [36]. Studies on conversational agents with NLP-based automatic depression detection were also not included [117]. Second, although the Egger test indicated no significant publication bias, the exclusion of non-English and low-visibility studies raises the possibility of study omission. Future work may cautiously broaden the inclusion scope to better capture emerging methodological trends, including hybrid approaches that integrate text with other modalities, with the aim of improving the practical applicability of ADE models.

Implications

The results of this review indicate that multiple methodological factors, including text source, text representation, model architecture, annotation source, and reporting quality, influence the performance of text-based depression estimation models. These findings not only provide methodological guidance for future text-based ADE research but also offer important insights into the sources of variation observed in existing studies. To our knowledge, this study is among the first systematic reviews to explicitly restrict inclusion to models trained and validated using standard depression labels. Previous systematic reviews in this field have largely focused on aspects such as model architecture or the use of social media data, with little systematic attention given to the impact of depression label quality. The inclusion of weak depression labels in prior research may have affected the reliability of reported results and further limited their practical applicability. By treating standard depression labels as a core inclusion criterion, this review demonstrates that annotation source is one of the key determinants of model performance. This observation suggests that performance differences across studies may reflect variations in labeling standards and reporting quality, rather than differences in intrinsic model capability alone. More broadly, the present findings support a shift in text-based ADE research toward evaluation frameworks that emphasize label rigor, transparent reporting, and methodological consistency, thereby enabling more meaningful comparison across studies and ultimately facilitating the clinical translation of ADE models.

Conclusions

In summary, this systematic review and meta-analysis surveyed the last decade of text-based ADE using standard depression labels. The overall pooled effect size (r=0.605) suggests that ADE models perform well and have the potential for practical application. However, substantial heterogeneity across studies was observed. Models using embedding-based features and deep architectures generally achieved superior performance, whereas the influence of annotation source and text source was comparatively limited. Models trained with clinical diagnoses and transcribed speech tended to outperform those using self-report scales and documentation, though the difference was not statistically significant. Moreover, reporting quality, as assessed by TRIPOD, was positively associated with performance, highlighting the need for transparent reporting. Future studies should consider richer text representations, standardized labels, integration with other modalities, and transparent reporting to enhance the reliability and practical applicability of ADE models. In our future work, we will also aim to capture emerging approaches and cautiously broaden the inclusion scope, thereby providing a more comprehensive view of the ADE field.

Acknowledgments

The authors would like to thank the editor and reviewers for their constructive comments, which strengthened the quality of this manuscript. The authors declare the use of generative artificial intelligence (GenAI) to assist with language editing, proofreading, and improving clarity of expression. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing, and adapting and adjusting emotional tone. The GenAI tool used was ChatGPT (OpenAI). Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Data Availability

All data and search strategies are available in the additional files accompanying this article. Further details in this study are available from the corresponding author upon reasonable request.

Funding

This study was funded by the Second Batch of the 2023 Ministry of Education of the People's Republic of China Industry–University Collaborative Education Program (Program number 230905329052215; Kingfar-CES "Human Factors and Ergonomics" Program); the Guangdong Provincial Key Laboratory of Fully Actuated System Control Theory and Technology (Grant number 2024B1212010002); the Startup Fund of Southern University of Science and Technology (grant numbers Y01966117 and Y01966223).

Authors' Contributions

SZ contributed to the conceptualization, literature screening, data extraction, meta-analysis, and writing of the original draft. CZ contributed to the refinement of the review methodology, quality assessment, preparation of supplementary materials, and revision

XSL•FO
RenderX

of the manuscript. JZ provided guidance on methodological design, supervised the review process, and contributed to conceptualization, data extraction, and manuscript revision. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA 2020 checklist.
[DOCX File , 274 KB - jmir_v28i1e82686_app1.docx]

Multimedia Appendix 2
Search strategies.
[DOCX File , 17 KB - jmir_v28i1e82686_app2.docx]

Multimedia Appendix 3
Detailed data extraction table of included studies.
[DOCX File , 27 KB - jmir_v28i1e82686_app3.docx]

Multimedia Appendix 4
TRIPOD-AI_checklist.
[DOCX File , 86 KB - jmir_v28i1e82686_app4.docx]

Multimedia Appendix 5
Hartung–Knapp–Sidik–Jonkman corrected analyses using the R meta package.
[DOCX File , 1457 KB - jmir_v28i1e82686_app5.docx]

Multimedia Appendix 6
CMA4.0-based subgroup and meta-regression analyses.
[DOCX File , 1513 KB - jmir_v28i1e82686_app6.docx]

Multimedia Appendix 7
Risk of Bias and GRADE Assessment Tables.
[DOCX File , 141 KB - jmir_v28i1e82686_app7.docx]

References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. In: Diagnostic and Statistical Manual of Mental Disorders (DSM-5-TR). United States: American Psychiatric Association Publishing DSM-5, 5th ed; 2013.
2. Tolentino JC, Schmidt SL. DSM-5 criteria and depression severity: implications for clinical practice. Front Psychiatry 2018;9:450 [FREE Full text] [doi: 10.3389/fpsyt.2018.00450] [Medline: 30333763]
3. Hawton K, Casañas I C, Haw C, Saunders K. Risk factors for suicide in individuals with depression: a systematic review. J Affect Disord 2013;147(1-3):17-28 [FREE Full text] [doi: 10.1016/j.jad.2013.01.004] [Medline: 23411024]
4. Gabbard GO, Crisp-Han H. The early career psychiatrist and the psychotherapeutic identity. Academic Psychiatry 2017;41(1):30-34. [doi: 10.1007/s40596-016-0627-7] [Medline: 27882522]
5. Beck A, Alford B. Depression: Causes and Treatment, 2nd ed. US: University of Pennsylvania Press; 2009.
6. Breedvelt JJF, Zamperoni V, South E, Uphoff EP, Gilbody S, Bockting CLH, et al. A systematic review of mental health measurement scales for evaluating the effects of mental health prevention interventions. Eur J Public Health 2020;30(3):539-545. [doi: 10.1093/eurpub/ckz233] [Medline: 32236548]
7. Maust D, Cristancho M, Gray L, Rushing S, Tjoa C, Thase M. In: Aminoff MJ, Boller F, Swaab D, editors. Chapter 13 - Psychiatric rating scales. Cambridge University Press, Springer: Indian Psychiatric Society; 2012.
8. Nuevo R, Lehtinen V, Reyna-Liberato PM, Ayuso-Mateos JL. Usefulness of the beck depression inventory as a screening method for depression among the general population of finland. Scand J Public Health 2009;37(1):28-34. [doi: 10.1177/1403494808097169] [Medline: 19141552]
9. Stockings E, Degenhardt L, Lee Y, Mihalopoulos C, Liu A, Hobbs M, et al. Symptom screening scales for detecting major depressive disorder in children and adolescents: a systematic review and meta-analysis of reliability, validity and diagnostic utility. Journal of Affective Disorders 2015;174:447-463 [FREE Full text] [doi: 10.1016/j.jad.2014.11.061] [Medline: 25553406]

XSL•FO
RenderX

10. Gilbody S, Sheldon T, House A. Screening and case-finding instruments for depression: a meta-analysis. Canadian Medical Association Journal 2008;178(8):997-1003 [FREE Full text] [doi: 10.1503/cmaj.070281] [Medline: 18390942]
11. Ren Y, Yang H, Browning C, Thomas S, Liu M. Performance of screening tools in detecting major depressive disorder among patients with coronary heart disease: a systematic review. Med Sci Monit 2015;21:646-653 [FREE Full text] [doi: 10.12659/MSM.892537] [Medline: 25725615]
12. Ray A, Bhardwaj A, Malik Y, Singh S, Gupta R. Artificial intelligence and Psychiatry: an overview. Asian J Psychiatr 2022;70:103021 [FREE Full text] [doi: 10.1016/j.ajp.2022.103021] [Medline: 35219978]
13. Sun J. Artificial intelligence in psychiatry research, diagnosis, and therapy. Asian Journal of Psychiatry 2023:103705 [FREE Full text] [doi: 10.1016/j.ajp.2023.103705]
14. Wu C, Kuo C, Su C, Wang S. Using text mining to extract depressive symptoms and to validate the diagnosis of major depressive disorder from electronic health records. Journal of Affective Disorders 2020:617-623 [FREE Full text] [doi: 10.1016/j.jad.2019.09.044]
15. Yao X, Yu G, Tang J, Zhang J. Extracting depressive symptoms and their associations from an online depression community. Computers in Human Behavior 2021 Jul;120:106734 [FREE Full text] [doi: 10.1016/j.chb.2021.106734]
16. Yao S, Wang F, Chen J, Lu Q. Utilizing health-related text on social media for depression research: themes and methods. Library Hi Tech 2023;43(1):274-294. [doi: 10.1108/lht-02-2023-0076]
17. Shin D, Kim H, Lee S, Cho Y, Jung W. Using large language models to detect depression from user-generated diary text data as a novel approach in digital mental health screening: instrument validation study. J Med Internet Res 2024;26:e54617 [FREE Full text] [doi: 10.2196/54617] [Medline: 39292502]
18. Haque A, Guo M, Miner A, Fei-Fei L. Measuring depression symptom severity from spoken language and 3D facial expressions. arXiv 2018 Nov 21. [doi: 10.48550/arXiv.1811.08592]
19. Zhang B, Wei D, Yan G, Lei T, Cai H, Yang Z. Feature-level fusion based on spatial-temporal of pervasive EEG for depression recognition. Comput Methods Programs Biomed 2022;226:107113. [doi: 10.1016/j.cmpb.2022.107113] [Medline: 36103735]
20. Zhang Z, Huang P, Li S, Liu Z, Zhang J, Li Y, et al. Neural mechanisms underlying the processing of emotional stimuli in individuals with depression: an ALE meta-analysis study. Psychiatry Res 2022;313:114598. [doi: 10.1016/j.psychres.2022.114598] [Medline: 35544984]
21. Berger J, Packard G. Using natural language processing to understand people and culture. American Psychologist 2022;77(4):525-537. [doi: 10.1037/amp0000882] [Medline: 34914405]
22. Maruf AA, Khanam F, Haque MM, Jiyad ZM, Mridha MF, Aung Z. Challenges and opportunities of text-based emotion detection: a survey. IEEE Access 2024;12:18416-18450. [doi: 10.1109/access.2024.3356357]
23. Jackson J, Watts J, List J, Drabble R, Lindquist K. From text to thought: how analyzing language can advance psychological science. PsyArXiv 2020 Jul 01. [doi: 10.31234/osf.io/nws35]
24. Cariola L, Hinduja S, Bilalpur M, Sheeber L, Allen N, Morency L, et al. Language use in mother-adolescent dyadic interaction: preliminary results. Int Conf Affect Comput Intell Interact Workshops 2022;2022. [doi: 10.1109/acii55700.2022.9953886] [Medline: 39161704]
25. Olson G, Damme K, Cowan HR, Alliende LM, Mittal VA. Emotional tone in clinical high risk for psychosis: novel insights from a natural language analysis approach. Front Psychiatry 2024;15:1389597 [FREE Full text] [doi: 10.3389/fpsyt.2024.1389597] [Medline: 38803678]
26. Ricard B, Marsch L, Crosier B, Hassanpour S. Exploring the utility of community-generated social media content for detecting depression: an analytical study on instagram. JMIR Preprints 2018 Aug 06 [FREE Full text] [doi: 10.2196/preprints.11817]
27. Wang L, Zhang Y, Zhou B, Cao S, Hu K, Tan Y. Automatic depression prediction via cross-modal attention-based multi-modal fusion in social networks. Comput Electr Eng 2024;118:109413 [FREE Full text] [doi: 10.1016/j.compeleceng.2024.109413]
28. Liu T, Meyerhoff J, Eichstaedt J, Karr C, Kaiser S, Kording K, et al. The relationship between text message sentiment and self-reported depression. J Affect Disord 2022;302:7-14 [FREE Full text] [doi: 10.1016/j.jad.2021.12.048] [Medline: 34963643]
29. Tao Y, Yang M, Shen H, Yang Z, Weng Z, Hu B. Classifying anxiety and depression through LLMs virtual interactions: a case study with ChatGPT. 2023 Presented at: Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 5-8, 2023; Istanbul, Turkiye. [doi: 10.1109/bibm58861.2023.10385305]
30. Feng W, Wu H, Ma H, Yin Y, Tao Z, Lu S, et al. Deep learning based prediction of depression and anxiety in patients with type 2 diabetes mellitus using regional electronic health records. Int J Med Inform 2025;196:105801 [FREE Full text] [doi: 10.1016/j.ijmedinf.2025.105801] [Medline: 39889672]
31. Mao K, Wu Y, Chen J. A systematic review on automated clinical depression diagnosis. Npj Ment Health Res 2023;2(1):20 [FREE Full text] [doi: 10.1038/s44184-023-00040-z] [Medline: 38609509]
32. Gratch J, Arstein R, Lucas G, Stratou G, Morency L. The distress analysis interview corpus of human and computer interviews. 2014 Presented at: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14); May 28-30, 2014; Paris, France.

XSL•FO
RenderX

33. Valstar M, Schuller B, Smith K, Almaev T, Eyben F, Krajewski J, et al. AVEC 2014: 3D dimensional affect and depression recognition challenge. 2014 Presented at: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge; Nov 7, 2014; Orlando Florida p. 3. [doi: 10.1145/2661806.2661807]
34. Shen Y, Yang H, Lin L. Automatic depression detection: an emotional audio-textual corpus and a Gru/Bilstm-based model. 2022 Presented at: Proceedings of the ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 23-27, 2022; Singapore, Singapore. [doi: 10.1109/icassp43922.2022.9746569]
35. Zou B, Han J, Wang Y, Liu R, Zhao S, Feng L, et al. Semi-structural interview-based Chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders. IEEE Trans. Affective Comput 2023;14(4):2823-2838. [doi: 10.1109/taffc.2022.3181210]
36. Guo Y, Liu J, Wang L, Qin W, Hao S, Hong R. A prompt-based topic-modeling method for depression detection on low-resource data. IEEE Trans Comput Soc Syst 2024;11(1):1430-1439. [doi: 10.1109/tcss.2023.3260080]
37. Senn S, Tlachac M, Flores R, Rundensteiner E. Ensembles of BERT for depression classification. 2022 Presented at: Proceedings of the 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); May 11-15, 2022; New York City p. 11-15. [doi: 10.1109/embc48229.2022.9871120]
38. Sikström S, Valavičiūtė I, Kuusela I, Evors N. Question-based computational language approach outperforms rating scales in quantifying emotional states. Commun Psychol 2024;2(1):45. [doi: 10.1038/s44271-024-00097-2] [Medline: 39242812]
39. Cheng Q, Li T, Kwok C, Zhu T, Yip P. Assessing suicide risk and emotional distress in Chinese social media: a text mining and machine learning study. J Med Internet Res 2017;19(7):e243 [FREE Full text] [doi: 10.2196/jmir.7276] [Medline: 28694239]
40. Dogrucu A, Perucic A, Isaro A, Ball D, Toto E, Rundensteiner E, et al. Moodable: on feasibility of instantaneous depression assessment using machine learning on voice samples with retrospectively harvested smartphone and social media data. Smart Health 2020;17:100118 [FREE Full text] [doi: 10.1016/j.smhl.2020.100118]
41. Rude S, Gortner E, Pennebaker J. Language use of depressed and depression-vulnerable college students. Cognition & Emotion 2004;18(8):1121-1133. [doi: 10.1080/02699930441000030]
42. Gong Y, Poellabauer C. Topic modeling based multi-modal depression detection. 2017 Presented at: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge; Oct 23, 2017; Mountain View California USA. [doi: 10.1145/3133944.3133945]
43. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. Computer Science 2020. [doi: 10.3126/jiee.v3i1.34327]
44. Pennington J, Socher R, Manning C. Glove: global vectors for word representation. 2014 Presented at: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); Oct 26-28, 2014; Doha, Qatar. [doi: 10.3115/v1/d14-1162]
45. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv 2019 May 24. [doi: 10.48550/arXiv.1810.04805]
46. Radford A, Narasimhan K. Improving Language Understanding by Generative Pre-Training. San Francisco, California: OpenAI; 2018.
47. Niu M, Chen K, Chen Q, Yang L. HCAG: a hierarchical context-aware graph attention model for depression detection. 2021 Presented at: Proceedings of the ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP);; June 06-11, 2021; Toronto, ON, Canada p. 6-11. [doi: 10.1109/icassp39728.2021.9413486]
48. Makiuchi M, Warnita T, Uto K, Shinoda K. Multimodal fusion of BERT-CNN and gated CNN representations for depression detection. : Association for Computing Machinery; 2019 Presented at: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop; Nice, France: Association for Computing Machinery; Oct 21, 2019; Nice, France p. 55-63. [doi: 10.1145/3347320.3357694]
49. Fan W, He Z, Xing X, Cai B, Lu W. Multi-modality depression detection via multi-scale temporal dilated CNNs. : Association for Computing Machinery; 2019 Presented at: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop; Oct 19, 2019; Nice, France p. 73-80. [doi: 10.1145/3347320.3357695]
50. Yang L, Jiang D, He L, Pei E, Oveneke M, Sahli H. Decision tree based depression classification from audio video and language information. : Association for Computing Machinery; 2016 Presented at: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge; Amsterdam, The Netherlands: Association for Computing Machinery; Oct 16, 2016; Amsterdam, The Netherlands. [doi: 10.1145/2988257.2988269]
51. Dinkel H, Wu M, Yu K. Text-based depression detection: what triggers an alert. arXiv 2019 Apr 08. [doi: 10.48550/arXiv.1904.05154]
52. Martinez AE, Cuadrado J, Peña D, Martinez SJ, Puertas E. Automated depression detection in text data. Leveraging lexical features, phonesthemes embedding, and RoBERTa transformer model? 2023.
53. Wu Y, Liu Z, Yuan J, Chen B, Cai H, Liu L. PIE: a personalized information embedded model for text-based depression detection. Inf Process Manag 2024:103830 [FREE Full text] [doi: 10.1016/j.ipm.2024.103830]
54. Dai H, Su C, Lee Y, Zhang Y, Wang C, Kuo C, et al. Deep learning-based natural language processing for screening psychiatric patients. Front Psychiatry 2020;11:533949 [FREE Full text] [doi: 10.3389/fpsyt.2020.533949] [Medline: 33584354]

XSL•FO
RenderX

55. Jarvers I, Ecker A, Donabauer P, Kampa K, Weißenbacher M, Schleicher D, et al. M.I.N.I.-KID interviews with adolescents: a corpus-based language analysis of adolescents with depressive disorders and the possibilities of continuation using Chat GPT. Front Psychiatry 2024;15:1425820 [FREE Full text] [doi: 10.3389/fpsyt.2024.1425820] [Medline: 39748903]
56. Zhang Z, Zhang S, Ni D, Wei Z, Yang K, Jin S, et al. Multimodal sensing for depression risk detection: integrating audio, video, and text data. Sensors (Basel) 2024;24(12):3714 [FREE Full text] [doi: 10.3390/s24123714] [Medline: 38931497]
57. He Z. Prediction mechanism of depression tendency among college students under computer intelligent systems. J Intell Syst 2024;33(1). [doi: 10.1515/jisys-2023-0209]
58. Liu Z, Wu Y, Zhang H, Li G, Ding Z, Hu B. Stimulus-response patterns: the key to giving generalizability to text-based depression detection models. IEEE J Biomed Health Inform 2024;28(8):4925-4936. [doi: 10.1109/JBHI.2024.3393244] [Medline: 38656850]
59. He L, Niu M, Tiwari P, Marttinen P, Su R, Jiang J, et al. Deep learning for depression recognition with audiovisual cues: a review. Inf Fusion 2022;80:56-86 [FREE Full text] [doi: 10.1016/j.inffus.2021.10.012]
60. Freedman R, Lewis DA, Michels R, Pine DS, Schultz SK, Tamminga CA, et al. The initial field trials of DSM-5: new blooms and old thorns. Am J Psychiatry 2013;170(1):1-5. [doi: 10.1176/appi.ajp.2012.12091189] [Medline: 23288382]
61. Kroenke K, Spitzer R. The PHQ-9: A new depression diagnostic and severity measure. San Francisco, California: US: SLACK; 2002.
62. Beck AT, Steer RA, Ball R, Ranieri WF. Comparison of beck depression inventories -IA and -II in psychiatric outpatients. J Pers Assess 1996;67(3):588-597. [doi: 10.1207/s15327752jpa6703_13] [Medline: 8991972]
63. Zung W. A self-rating depression scale. Arch Gen Psychiatry 1965;12:63-70. [doi: 10.1001/archpsyc.1965.01720310065008] [Medline: 14221692]
64. Lovibond S, Lovibond P. Manual for the Depression Anxiety Stress Scales. Australia: S. H. Lovibond, Peter F. Lovibond; 1995.
65. Radloff L. The CES-D Scale: a self-report depression scale for research in the general population. US: Sage Publications; 1977:385-401.
66. Li M, Xu H, Liu W, Liu J. Bidirectional LSTM and attention for depression detection on clinical interview transcripts. 2022 Presented at: Proceedings of the 2022 IEEE 10th International Conference on Information, Communication and Networks (ICICN); Aug 23-24, 2022; Zhangye, China. [doi: 10.1109/icicn56848.2022.10006532]
67. Zhao T, Tlachac M. Optimization with tree ensembles to improve depression screening on textual datasets. IEEE Transactions on Affective Computing 2025;16(2):573-585. [doi: 10.1109/taffc.2024.3442557]
68. Deshpande M, Rao V. Depression detection using emotion artificial intelligence. 2018 Presented at: Proceedings of the 2017 International Conference on Intelligent Sustainable Systems (ICISS); June 21, 2018; Palladam, India p. 858-862. [doi: 10.1109/iss1.2017.8389299]
69. Amanat A, Rizwan M, Javed AR, Abdelhaq M, Alsaqour R, Pandya S, et al. Deep learning for depression detection from textual data. Electronics 2022;11(5):676. [doi: 10.3390/electronics11050676]
70. Raza G, Butt Z, Latif S, Wahid A. Sentiment analysis on COVID tweets: an experimental analysis on the impact of count vectorizer and TF-IDF on sentiment predictions using deep learning models. 2021 Presented at: Proceedings of the 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2); May 20-21, 2021; Islamabad, Pakistan. [doi: 10.1109/icodt252288.2021.9441508]
71. Kour H, Gupta MK. Predicting the language of depression from multivariate twitter data using a feature - rich hybrid deep learning model. Concurr Comput 2022 Jul 22;34(24):107363. [doi: 10.1002/cpe.7224]
72. Milintsevich K, Sirts K. Your model is not predicting depression well and that is why: a case study of PRIMATE dataset. arXiv 2024 Mar 01. [doi: 10.48550/arXiv.2403.00438]
73. Arias J, Williams C, Raghvani R, Aghajani M, Baez S, Belzung C, et al. The neuroscience of sadness: a multidisciplinary synthesis and collaborative review. Neurosci Biobehav Rev 2020;111:199-228 [FREE Full text] [doi: 10.1016/j.neubiorev.2020.01.006] [Medline: 32001274]
74. Kaushik P, Bansal K, Kumar Y, Changela A. Mental disorders prognosis and predictions using artificial intelligence techniques: a comprehensive study. SN Comput Sci 2024 Nov 14;5(8):1048. [doi: 10.1007/s42979-024-03416-w]
75. Levis B, Benedetti A, Ioannidis J, Sun Y, Negeri Z, He C, et al. Patient Health Questionnaire-9 scores do not accurately estimate depression prevalence: individual participant data meta-analysis. J Clin Epidemiol 2020:115-128 [FREE Full text] [doi: 10.1016/j.jclinepi.2020.02.002] [Medline: 32105798]
76. Tahir WB, Khalid S, Almutairi S, Abohashrh M, Memon SA, Khan J. Depression detection in social media: a comprehensive review of machine learning and deep learning techniques. IEEE Access 2025;13:12789-12818. [doi: 10.1109/access.2025.3530862]
77. Nanggala K, Elwirehardja GN, Pardamean B. Systematic literature review of transformer model implementations in detecting depression. 2023 Presented at: Proceedings of the 6th International Conference of Computer and Informatics Engineering (IC2IE) 2023; Sep 14-15, 2023; Lombok, Indonesia. [doi: 10.1109/ic2ie60547.2023.10331448]
78. Phiri D, Makowa F, Amelia VL, Phiri YVA, Dlamini LP, Chung M. Text-based depression prediction on social media using machine learning: systematic review and meta-analysis. J Med Internet Res 2025;27:e59002 [FREE Full text] [doi: 10.2196/59002] [Medline: 40215481]

XSL•FO
RenderX

79. Kelley SW, Mhaonaigh CN, Burke L, Whelan R, Gillan CM. Machine learning of language use on Twitter reveals weak and non-specific predictions. NPJ Digit Med 2022;5(1):35 [FREE Full text] [doi: 10.1038/s41746-022-00576-y] [Medline: 35338248]
80. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev 2021;10(1):89 [FREE Full text] [doi: 10.1186/s13643-021-01626-4] [Medline: 33781348]
81. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
82. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:e078378 [FREE Full text] [doi: 10.1136/bmj.q902] [Medline: 38636956]
83. Lam B, Chrysafi P, Chiasakul T, Khosla H, Karagkouni D, McNichol M, et al. Machine learning natural language processing for identifying venous thromboembolism: systematic review and meta-analysis. Blood advances 2024;8(12):2991-3000 [FREE Full text] [doi: 10.1182/bloodadvances.2023012200] [Medline: 38522096]
84. Casey A, Davidson E, Poon M, Dong H, Duma D, Grivas A, et al. A systematic review of natural language processing applied to radiology reports. BMC Med Inform Decis Mak 2021;21(1):179 [FREE Full text] [doi: 10.1186/s12911-021-01533-7] [Medline: 34082729]
85. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. Journal of Clinical Epidemiology 2003;56(11):1129-1135. [doi: 10.1016/s0895-4356(03)00177-x]
86. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science. A practical primer for t-tests and ANOVAs 2013;4:863 [FREE Full text] [doi: 10.3389/fpsyg.2013.00863] [Medline: 24324449]
87. Cohen J. Statistical power analysis for the behavioral sciences (Rev. ed). Hillsdale, NJ. US: Lawrence Erlbaum Associates, Inc; 1977.
88. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21(11):1539-1558. [doi: 10.1002/sim.1186] [Medline: 12111919]
89. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta-analysis: I is not an absolute measure of heterogeneity. Res Synth Methods 2017;8(1):5-18. [doi: 10.1002/jrsm.1230] [Medline: 28058794]
90. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023:101014 [FREE Full text] [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
91. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014;14:25 [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
92. Bown MJ, Sutton AJ. Quality control in systematic reviews and meta-analyses. Eur J Vasc Endovasc Surg 2010;40(5):669-677 [FREE Full text] [doi: 10.1016/j.ejvs.2010.07.011] [Medline: 20732826]
93. Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 2004;56(2):455-463 [FREE Full text] [doi: 10.1111/j.0006-341x.2000.00455.x] [Medline: 10877304]
94. Sterne JA, Egger M, Smith GD. Systematic reviews in health care: investigating and dealing with publication and other biases in meta-analysis. BMJ 2001;323(7304):101-115 [FREE Full text] [doi: 10.1136/bmj.323.7304.101] [Medline: 11451790]
95. Duval S. The trim and fill method. In: Publication Bias in Meta-Analysis. Chichester, UK: John Wiley & Sons, Ltd; 2005:127-144.
96. Guyatt G, Oxman A, Akl E, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;64(4):383-394 [FREE Full text] [doi: 10.1016/j.jclinepi.2010.04.026] [Medline: 21195583]
97. Geraci J, Wilansky P, de Luca V, Roy A, Kennedy JL, Strauss J. Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression. Evid Based Ment Health 2017;20(3):83-87 [FREE Full text] [doi: 10.1136/eb-2017-102688] [Medline: 28739578]
98. Tlachac ML, Rundensteiner E. Screening for depression with retrospectively harvested private versus public text. IEEE J Biomed Health Inform 2020;24(11):3326-3332. [doi: 10.1109/JBHI.2020.2983035] [Medline: 32224470]
99. Shin D, Kim K, Lee S, Lee C, Bae YS, Cho WI, et al. Detection of depression and suicide risk based on text from clinical interviews using machine learning: possibility of a new objective diagnostic marker. Frontiers in Psychiatry 2022;13:801301. [doi: 10.3389/fpsyt.2022.801301] [Medline: 35686182]
100. Munthuli A, Pooprasert P, Klangpornkun N, Phienphanich P, Onsuwan C, Jaisin K, et al. Classification and analysis of text transcription from Thai depression assessment tasks among patients with depression. PLoS One 2023;18(3):e0283095 [FREE Full text] [doi: 10.1371/journal.pone.0283095] [Medline: 36996118]

XSL•FO
RenderX

101. Tlachac M, Shrestha A, Shah M, Litterer B, Rundensteiner EA. Automated construction of lexicons to improve depression screening with text messages. IEEE J Biomed Health Inform 2023;27(6):2751-2759. [doi: 10.1109/JBHI.2022.3203345] [Medline: 36044503]
102. Oh J, Lee T, Chung ES, Kim H, Cho K, Kim H, et al. Development of depression detection algorithm using text scripts of routine psychiatric interview. Front Psychiatry 2023;14:1256571 [FREE Full text] [doi: 10.3389/fpsyt.2023.1256571] [Medline: 38239906]
103. Xu C, Chen Y, Tao Y, Xie W, Liu X, Lin Y. Deep learning-based detection of depression by fusing auditory, visual and textual clues. Journal of Affective Disorders 2025:119860 [FREE Full text] [doi: 10.1016/j.jad.2025.119860]
104. Senn SJ. Overstating the evidence: double counting in meta-analysis and related problems. BMC Med Res Methodol 2009;9(1):10 [FREE Full text] [doi: 10.1186/1471-2288-9-10] [Medline: 19216779]
105. Zhang Y, He Y, Rong L, Ding Y. A hybrid model for depression detection with transformer and bi-directional long short-term memory. 2022 Presented at: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 6-8, 2022; Las Vegas, NV, USA. [doi: 10.1109/bibm55620.2022.9995184]
106. Kaynak EB, Dibeklio?lu H. Systematic Analysis of Speech Transcription Modeling for Reliable Assessment of Depression Severity. Sakarya University Journal of Computer and Information Sciences. 2024 April 2024;7(1):77-91. [doi: 10.35377/saucis...1381522]
107. Yadav U, Sharma A. A novel automated depression detection technique using text transcript. Int J Imaging Syst Tech 2022;33(1):108-122 [FREE Full text] [doi: 10.1002/ima.22793]
108. Avasthi S, Chauhan R, Acharjya D. Topic modeling techniques for text mining over a large-scale scientific and biomedical text corpus. IJACI 2022;13(1):1-18. [doi: 10.4018/ijaci.293137]
109. Du X, Sun Y. Linguistic features and psychological states: a machine-learning based approach. Front Psychol 2022;13:955850 [FREE Full text] [doi: 10.3389/fpsyg.2022.955850] [Medline: 35936260]
110. Yang K, Kim J, Chun M, Ahn MS, Chon E, Park J, et al. Factors to improve distress and fatigue in cancer survivorship; further understanding through text analysis of interviews by machine learning. BMC Cancer 2021;21(1):741 [FREE Full text] [doi: 10.1186/s12885-021-08438-8] [Medline: 34176470]
111. Ghosh CC, McVicar D, Davidson G, Shannon C, Armour C. What can we learn about the psychiatric diagnostic categories by analysing patients' lived experiences with machine-learning? BMC Psychiatry 2022;22(1):427 [FREE Full text] [doi: 10.1186/s12888-022-03984-2] [Medline: 35751077]
112. Tang SX, Kriz R, Cho S, Park SJ, Harowitz J, Gur RE, et al. Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. NPJ Schizophr 2021;7(1):25 [FREE Full text] [doi: 10.1038/s41537-021-00154-3] [Medline: 33990615]
113. Teleki M, Dong X, Caverlee J. Quantifying the impact of disfluency on spoken content summarization. 2024 Presented at: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024); 2024 May; Torino, Italia.
114. Shang J, Xue K, Yang C, Shi H, Pan L, Zeng Y. Prediction models for health care workers' exposure to type II workplace violence: a systematic review and meta-analysis. J Nurs Care Qual 2026;41(1):29-36. [doi: 10.1097/NCQ.0000000000000897] [Medline: 40658938]
115. Zhang M, Nieuwenhuys A, Zhang Y. Posture prediction models in digital human modeling for ergonomic design: a systematic review. Med Eng Phys 2025;143:104391 [FREE Full text] [doi: 10.1016/j.medengphy.2025.104391] [Medline: 40835359]
116. Gu Z, Zhu Q. MENTALER: toward professional mental health support with LLMs via multi-role collaboration. 2024 Presented at: IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 3-6, 2024; Lisbon, Portugal p. 3-6. [doi: 10.1109/bibm62325.2024.10822643]
117. Otero-González I, Pacheco-Lorenzo M, Fernández-Iglesias MJ, Anido-Rifón LE. Conversational agents for depression screening: a systematic review. Int J Med Inform 2024;181:105272 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.105272] [Medline: 37979500]

Abbreviations

ADE: automatic depression estimation
BDI-II: Beck Depression Inventory-II
BERT: Bidirectional Encoder Representations from Transformers
CES-D: Center for Epidemiological Studies Depression Scale
DAIC-WOZ: Distress Analysis Interview Corpus – Wizard-of-Oz
DASS-21: 21-item Depression, Anxiety, and Stress Scale
ES: effect size
GRADE: Grading of Recommendations Assessment, Development and Evaluation
HKSJ: Hartung-Knapp-Sidik-Jonkman
ML: machine learning
NLP: natural language processing

XSL•FO
RenderX

OR: odds ratio
PHQ-8: Patient Health Questionnaire-8
PHQ-9: Patient Health Questionnaire-9
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
SDS: Self-Rating Depression Scale
TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

Edited by S Brini; submitted 20.Aug.2025; peer-reviewed by M Ibrahim, U Pasupulety; comments to author 28.Nov.2025; accepted 10.Jan.2026; published 11.Feb.2026.

Please cite as:
Zhang S, Zhang C, Zhang J
Text-Based Depression Estimation Using Machine Learning With Standard Labels: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e82686
URL: https://www.jmir.org/2026/1/e82686
doi:10.2196/82686
PMID:

©Shengming Zhang, Chaohai Zhang, Jiaxin Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

AI and Wearables for Early Detection of Cognitive Impairment and Dementia: Systematic Review

Ander Cejudo[1,2], MSc; Markel Arrojo[1], MSc; Cristina Martín[1,2,3], PhD; Aitor Almeida[2], PhD

[1]Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Donostia-San Sebastián, Spain

[2]Faculty of Engineering, University of Deusto, Bilbao, Spain

[3]Biodonostia Health Research Institute, e-Health Group, San Sebastián, Spain

Corresponding Author:
Ander Cejudo, MSc
Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
Mikeletegi 57
Donostia-San Sebastián, 20009
Spain
Phone: 34 943 309 230
Email: acejudo@vicomtech.org

Abstract

Background: Traditional cognitive screening relies on episodic clinical assessments and may miss early changes preceding cognitive impairment and dementia. Wearable and mobile health technologies enable continuous monitoring of sleep, physical activity, and circadian rhythms, generating digital biomarkers that may support scalable early detection and prevention. However, current evidence remains fragmented across devices, analytic approaches, and cognitive outcomes.

Objective: This study synthesizes and critically evaluates recent evidence on wearable devices for early detection and prevention of cognitive impairment and dementia, focusing on device categories, cognitive outcomes, analytic approaches, and prevention relevance.

Methods: We searched PubMed, Scopus, ACM Digital Library, and SpringerLink for peer-reviewed studies published between January 2020 and December 1, 2025. Eligible studies included human participants with a mean age ≥50 years, continuous wearable-derived data collected for ≥24 hours, and validated cognitive outcomes; reviews, protocols, smartphone-only studies, and pharmacological interventions were excluded. Two reviewers independently screened studies, extracted data, and assessed risk of bias using the Appraisal Tool for Cross-Sectional Studies, Newcastle-Ottawa Scale, Cochrane Risk of Bias tool, and Quality Assessment of Diagnostic Accuracy Studies-2. Owing to substantial heterogeneity in devices, outcomes, and analytic methods, quantitative meta-analysis was not feasible; a structured narrative synthesis was conducted in accordance with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidance. This study was not prospectively registered.

Results: We included 49 studies, with sample sizes ranging from 14 to 91,948 participants (>200,000 total) and a median sample size of 145. Most used research-grade actigraphy (43/49, 87.8%), while fewer used commercial wearables (7/49, 14.3%). Cognitive outcomes most frequently relied on global screening instruments, including the Mini-Mental State Examination (18/49, 36.7%), followed by *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*)–based clinical diagnoses (7/49, 14.3%) and the Montreal Cognitive Assessment (7/49, 14.3%). Analytic approaches were predominantly statistical (36/49, 73.5%), with fewer studies applying machine learning (7/49, 14.3%) or deep learning methods (6/49, 12.2%). Statistical analyses linked disrupted sleep, circadian rhythm fragmentation, and irregular activity patterns to worse cognitive outcomes, with modest-to-moderate effect sizes. Machine learning and deep learning approaches reported classification performance with area under the curve values between approximately 0.70 and 0.95. Approximately one-quarter of the studies (13/49, 26.5%) addressed early detection or prevention through longitudinal risk estimation or predictive modeling. Key limitations included small sample sizes, short monitoring durations, and limited external validation.

Conclusions: Wearable-derived behavioral markers show promise for early risk stratification. This review advances the field by shifting from descriptive associations toward a digital phenotyping framework evaluating artificial intelligence–driven prediction in the preclinical window. Unlike prior reviews focused on established dementia, it differentiates direct predictive evidence from indirect correlational findings and critically assesses methodological maturity. Continuous, passive monitoring may enable scalable detection of subtle behavioral changes, supporting earlier and more personalized risk reduction strategies.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e86262) doi:10.2196/86262

KEYWORDS

actigraphy; artificial intelligence; circadian rhythms; cognitive impairment; dementia; machine learning; mHealth; mobile health; prevention; sleep; wearable devices

Introduction

Background

Cognitive decline and dementia represent a major public health concern, with significant global and regional prevalence. In 2021, approximately 57 million people were affected by dementia worldwide, with more than 60% residing in low- and middle-income countries [1]. As the global population ages, dementia is expected to impact 78 million individuals by 2030 and 139 million by 2050 [2]. This increasing prevalence is attributed to both the rising number of older adults and the fact that dementia disproportionately affects the older population. Furthermore, the severity of the problem is anticipated to rise more sharply in regions with significant demographic changes, particularly low- and middle-income countries, where health care resources are often limited [3].

Dementia severely affects the quality of life, as it is a leading cause of disability and dependency in older individuals [4]. In addition to the physical and cognitive impairments experienced by individuals, dementia imposes a heavy emotional and economic burden on families and caregivers. Caregivers, often family members, are responsible for a substantial amount of care, with an average of 5 hours of daily care and supervision per individual affected. The global cost of dementia is estimated to exceed US $1.3 trillion, with informal caregiving accounting for approximately half of this expenditure. Health care and social systems face immense strain in responding to the needs of individuals with dementia, highlighting the urgent need for preventive measures and early intervention [2].

There are several modifiable risk factors for cognitive decline and dementia, including physical inactivity, unhealthy diets, smoking, and cardiovascular conditions such as hypertension, diabetes, and obesity [5]. Early identification of cognitive decline is crucial for enabling timely interventions. Traditional screening methods, such as the Mini-Mental State Examination (MMSE) and Montreal Cognitive Assessment (MoCA) [6], have limitations due to their episodic nature and reliance on periodic visits to health care providers. These methods are often influenced by situational factors [7,8] and may not detect early signs of decline or track changes in cognitive function over time [9,10]. In addition, performance on these instruments can be affected by educational level, language, and cultural background, and they provide only a cross-sectional snapshot of cognitive status [11]. Therefore, there is a growing need for continuous and more sensitive methods of monitoring cognitive health [12]. At the same time, the use of artificial intelligence (AI)–driven wearable technologies for continuous cognitive monitoring raises ethical and data privacy considerations, as such systems may involve the collection of sensitive behavioral and health-related data, underscoring the importance of appropriate data governance and responsible implementation [13].

Wearable devices present significant potential for the continuous and passive monitoring of key physiological variables, including activity levels, sleep patterns, heart rate, and gait. These devices can detect subtle, longitudinal changes in behavior that may occur before clinical symptoms emerge [14], allowing for the early identification of cognitive decline. AI enhances this process by analyzing data from wearables to identify patterns that indicate potential risks. AI algorithms can process large volumes of data, detecting even small shifts in physiological and behavioral patterns, which can inform personalized interventions. By leveraging AI for predictive analytics, it is possible to improve early detection capabilities, facilitating timely interventions and supporting more effective management of cognitive health [15]. This combination of wearables and AI offers a promising approach [16] to enhancing dementia prevention and care.

Objective

Previous reviews have mainly examined associations between wearable-derived measures and dementia outcomes [16], often focusing on physiological or behavioral differences in individuals already diagnosed with the disease. Other reviews have emphasized specific domains, such as gait and mobility assessment using wearables [17] or the relationship between physical activity and cognitive decline [18]. In contrast, this review focuses on how wearable-derived data are analyzed, with particular attention to advanced analytic approaches, including machine learning and deep learning, used to support early detection and prevention across the cognitive aging spectrum, from subjective cognitive decline to clinical dementia. By synthesizing evidence within a digital phenotyping framework, this review highlights how longitudinal, multimodal wearable data combined with contemporary analytic methods can move beyond descriptive associations toward individualized risk characterization and prevention-oriented applications.

Thus, the main objective of this systematic review is to synthesize and critically evaluate the current evidence base and methodological maturity of wearable devices for the early detection and prevention of cognitive impairment and dementia. We structure our synthesis around 4 specific objectives:

- To evaluate the usage trends and application contexts of wearable technologies, specifically comparing the deployment of research-grade actigraphy vs consumer-grade devices.
- To synthesize the strength of statistical associations between wearable-derived digital biomarkers (eg, sleep and circadian rhythms) and standard clinical cognitive assessments.
- To critically assess the performance and robustness of analytic approaches, comparing the outcomes of conventional statistical modeling, machine learning, and deep learning techniques.

XSL•FO
RenderX

- To assess how early detection is addressed in wearable-based studies by differentiating between studies that directly implement early detection and those that frame their findings as potentially relevant to early detection.

By addressing these specific objectives, the review aims to clarify the current state of knowledge and identify priorities for advancing wearable-based approaches toward clinical and public health application.

Methods

Overview

The reporting of this systematic review was guided by the standards of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement [19] (Multimedia Appendix 1). The literature search strategy was developed and reported in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) [20] to enhance transparency and reproducibility of the search process (Multimedia Appendix 2). The aim was to evaluate the association between wearable device–based sleep and activity measures and the onset or progression of cognitive impairment, including mild cognitive impairment (MCI), Alzheimer disease, and cognitive decline. The review process was carried out in multiple stages: identification of relevant articles, screening for eligibility, data extraction, and assessment of risk of bias. Studies that meet predefined inclusion criteria were included, and their findings were synthesized to provide insights into the role of wearable devices in monitoring cognitive health. No review protocol was prepared or registered before conducting this study.

Information Sources and Search Strategy

A comprehensive search strategy was used across multiple electronic databases, including PubMed, Scopus, ACM Digital Library, and SpringerLink, to identify relevant studies published after 2020. The search was conducted using a combination of MeSH (Medical Subject Headings) terms and keywords related to cognitive impairment, sleep and activity measures, wearable devices, and their associations with cognitive outcomes. Databases were searched individually using their native interfaces. Duplicate records were removed, and studies published before 2020 were discarded to align the scope of this systematic review with contemporary research questions, analytic frameworks, and the use of wearable devices representative of current practice in digital phenotyping and early detection and prevention of cognitive decline. The search strategy was designed to capture a broad range of studies, including randomized controlled trials, cohort studies, and observational studies, to provide a comprehensive overview of the existing evidence.

The search strategy integrates key concepts relevant to studies on the early detection of cognitive impairment and dementia using wearable devices that monitor activity and sleep. Table 1 lists the search terms, organized into categories capturing cognitive impairment, sleep and activity measurements, wearable devices, and their application in early detection. Within each group, similar terms were combined using the OR operator, and the 4 groups were combined using the AND operator. Conceptually, the search followed the structure: (HC_1 OR HC_2 OR ...) AND ($Data_1$ OR $Data_2$ OR ...) AND ($Wearable_1$ OR $Wearable_2$ OR ...) AND ($EarlyDetection_1$ OR $EarlyDetection_2$ OR ...). Full search queries are included in Multimedia Appendix 3.

Alzheimer disease is specifically included under dementia, as it accounts for 60%-70% of all dementia cases [1]. Data capture must include information on sleep, activity, or physiological signals. Wearable devices may be commercial (eg, smartwatches) or noncommercial (ad hoc or research-grade devices) and are typically worn on the body for extended periods. This review focuses on studies advancing early detection and prevention, excluding research that examines features of the condition without the potential to identify cognitive decline before clinical diagnosis or progression to a more advanced stage.

To consider an article in this study, at least 1 search term of each group must be present in either the title or the abstract. This ensures that the manuscript has a clear focus on continuous data collection on sleep, activity, or physiological signals and links these measurements to cognitive impairment outcomes. The search was limited to studies published after 2020 and in English to ensure accessibility to full-text articles.

No additional search methods were used, and no study registries were searched. No targeted website searching or manual browsing of conference proceedings was conducted. Study authors and experts were not contacted for additional data. No published search filters were used. Search strategies were developed specifically for this review and were not adapted from prior reviews. No search updates were performed after the initial search, and the search strategy was not peer reviewed. Reference lists of included studies were screened to identify additional relevant records.

XSL•FO
RenderX

Table 1. Groups of search terms and representative keywords used to identify studies in the systematic review, organized by health condition, data type, wearable technology, and early detection.

Group	Search terms
Health condition	Cognitive impairment, mild cognitive impairment, dementia, cognitive decline, Alzheimer, neurocognitive disorder, and memory impairment
Data	Sleep, sleep duration, sleep quality, circadian rhythm, rest-activity rhythm, physical activity, actigraphy, accelerometry, heart rate, heart rate variability, HRV, respiratory rate, body temperature, respiration, skin temperature, step count, steps, and distance
Wearable	Wearable, wearable device, wearable devices, wearable sensor, wearable sensors, body-worn sensor, body-worn sensors, fitness tracker, smartwatch, smart band, wristband, wristbands, accelerometers, consumer-grade wearable, actigraph, actigraphy, ambulatory monitor, and digital biomarker
Early detection	Risk factor, onset, association, predictor, correlation, early detection, screening, and prevention

Eligibility Criteria

The inclusion and exclusion criteria that guided the selection of studies for this systematic review are presented in Textbox 1. The criteria are designed to ensure that the review captures only the most relevant, recent, and methodologically sound research.

For inclusion, studies must be published between 2020 and 2025 to reflect the most recent developments in the field. Only publications identified by the predefined search queries were considered. Eligible studies must present either new statistical outcomes or AI-based results that provide original contributions. Because the review focuses on prevention and early detection, manuscripts must directly address these topics. Furthermore, included studies were required to involve human participants with a mean age of 50 years or older, ensuring relevance to an aging population. Another important condition was the continuous collection of data from wearable devices for at least 24 hours, which allows for robust and objective measurements. Only peer-reviewed journal articles were considered, discarding gray literature and preprints.

Exclusion criteria further refine the selection. Publications not written in English or not available in full text, for example, when behind an inaccessible paywall, were excluded. Review articles, study protocols, and books were not considered to maintain a focus on original research. Retracted publications were excluded. In addition, studies were not eligible if they did not apply validated measures of cognitive impairment or dementia, focused on pharmacological interventions, relied exclusively on self-reported data or smartphone-based sensing, or included only healthy participants without relevant clinical characteristics. No minimum sample-size threshold was applied during study selection. As a result, included studies span a wide range of sample sizes, including small-sample investigations (eg, n<30).

Textbox 1. Inclusion and exclusion criteria applied in the systematic review.

Inclusion criteria:

- Published between 2020 and 2025
- Identified by the search queries
- Articles presenting new statistical outcomes or artificial intelligence results
- Manuscripts closely related to prevention and early detection
- Human participants must be involved in the study with average age of 50 years or older
- Continuous data captured from wearable devices for 24 hours or more

Exclusion criteria:

- Publication not in English
- Publication behind a paywall that cannot be retrieved
- Review papers, protocols, and books
- Retracted publications
- Validated measures of cognitive impairment and dementia not included
- Focus on drug or pharmacological interventions
- Only used self-reported data or smartphone-based sensing
- Only included healthy participants

XSL•FO
RenderX

Risk of Bias Assessment

The risk of bias in included studies was assessed using the Appraisal Tool for Cross-Sectional Studies, the Newcastle-Ottawa Scale for cohort studies, the Cochrane Risk of Bias tool for randomized controlled trials, and the Quality Assessment of Diagnostic Accuracy Studies-2 for diagnostic studies [21]. These instruments evaluate domains such as selection, performance, detection, attrition, and reporting bias, as well as other potential sources of bias. Each study was independently evaluated by 2 coauthors (AC and MA), with disagreements resolved through discussion or consultation with a third coauthor (CM). The overall risk of bias was classified as low, moderate, or high, and these assessments were considered when interpreting the results of the review.

Study Selection and Data Collection Process

Search results from each database were exported as BibTeX files and managed in Zotero (Corporation for Digital Scholarship), then converted to comma-separated values for automated preprocessing using a custom Python script. This process harmonized metadata across databases, removed duplicates, applied publication year filters, and verified the presence of predefined search terms. Automated preprocessing was conducted by the first author and independently verified by the second author. The following data items were extracted from each eligible study: bibliographic details (author and year), study design, participant characteristics (sample size, age, and diagnosis), wearable device specifications (type, brand, and sensor modality), monitoring duration, cognitive assessment tools, analytic methods (statistical vs AI), and primary outcomes (statistical associations or predictive performance metrics). Study selection followed a 3-stage screening process, with independent title, abstract, and full-text reviews performed by the first and second authors. Discrepancies were resolved through consensus or adjudication by a third author. Eligible studies were synthesized using a structured narrative approach aligned with predefined review objectives, grouping studies by device type, cognitive outcomes, analytic methods, and prevention-oriented contributions.

Analytic approaches were classified according to the overall modeling paradigm rather than individual algorithms. Studies were categorized as AI-based when wearable-derived data were analyzed within predictive modeling frameworks emphasizing automated pattern learning, such as machine learning or deep learning methods, including regression-based models implemented within machine learning pipelines (eg, automated feature selection or cross-validation). In contrast, studies relying on predefined statistical models for inferential or association analyses were classified as statistical model–based approaches. In this review, “direct” evidence was defined for studies demonstrating individual-level early detection capability, including both cross-sectional screening or discriminative analyses and longitudinal or interventional designs enabling risk estimation or prevention. Cross-sectional studies were therefore considered direct when they demonstrated the ability to distinguish preclinical or prodromal disease states at the time of assessment.

Synthesis Methods

Quantitative synthesis was planned only for outcomes estimating a common construct using comparable analytical objectives and metrics. When outcomes cannot be mapped to a shared estimand for early detection, quantitative pooling was not conducted. In such cases, results were synthesized using a structured narrative approach aligned with the predefined review objectives. This approach grouped studies according to analytic strategy, cognitive assessment, and wearable measurement characteristics, without applying formal qualitative synthesis methodologies.

Results

Study Selection

The database search identified 7175 records (SpringerLink, n=3247; Scopus, n=937; ACM Digital Library, n=2471; and PubMed, n=520). After removal of 705 duplicates, 458 records with missing information, and 422 reviews, a total of 5590 records remained for screening. Of these, 1493 were excluded based on publication year (<2020, n=1493), title (n=3527), or abstract (n=452). The remaining 118 reports were sought for full-text retrieval, but 15 could not be obtained. A total of 103 full-text articles were assessed for eligibility, and 54 were excluded for not meeting the inclusion criteria. Finally, a total of 49 studies were included in this systematic review (Figure 1). Figure 2 shows the number of included studies by publication year. At the full-text screening stage, most studies were excluded because they did not include continuous wearable-based data collected over at least 24 hours, did not use wearable-derived data as part of the analysis, or did not report results relevant to the early detection of cognitive impairment or dementia. Table 2 shows the relevant information extracted from the included studies, and additional details are provided in Multimedia Appendix 4.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the study selection process.

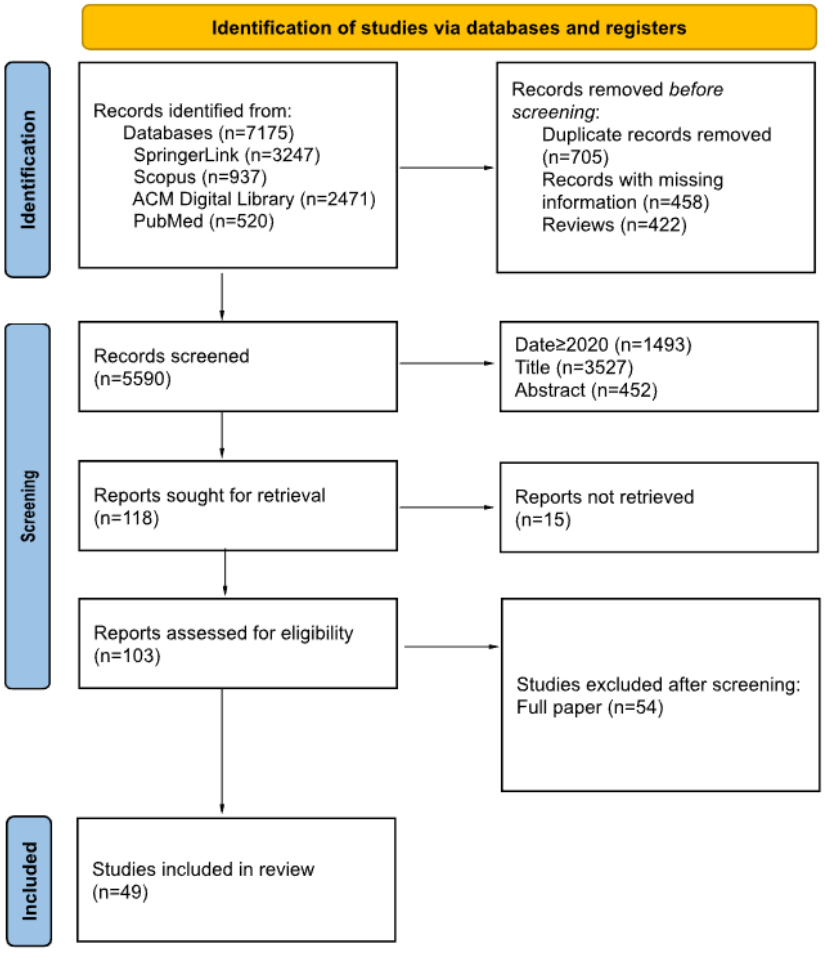

Figure 2. Temporal distribution of included studies by publication year (2020-2025). The bar chart and overlaid trend line illustrate the progression of research interest in the field. Data show a substantial increase in publication volume beginning in 2021, with research activity remaining consistently elevated through the 2024-2025 period. Data for 2025 are incomplete by 1 month.

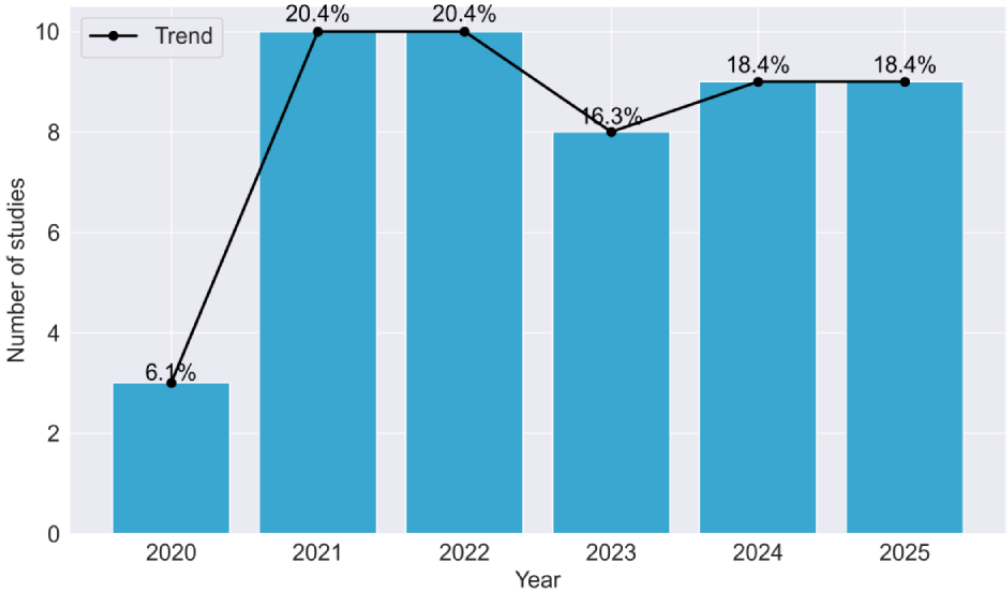

XSL•FO
RenderX

Table 2. Summary of included cross-sectional studies investigating wearable-derived measures in relation to dementia outcomes. The table reports study source, cross-sectional study design, participant groups and mean ages, and assessed risk of bias.

Source	Objective	Groups, n	Age (years), mean (SD)	Risk of bias
Lentzen et al [22]	Early detection	HC[a]=69; DEM[b]=160	69.2 (7.7)	Low
Jones et al [23]	Early detection	NCD[c]=21; MCI[d]=19	67.2 (8.4)	Low
Satomi et al [24]	Early detection	HC=25; MCI=63; DEM=21	79.3 (6.3)	Low
Sauers et al [25]	Risk assessment	HC=196; MCI=47	74	Low
Baril et al [26]	Risk assessment	HC=159; MCI=44	68.25	Low
Seo et al [27]	Early detection	HC=1519; MCI=353; DEM=87	70.1 (6.1)	Low
Huang et al [28]	Preventive intervention	HC=225; MCI=133; DEM=36	71.3(7.8); 72.9 (8.0)	Low
Palmer et al [29]	Progression monitoring	SCI[e]=29; MCI=79	68.8 (8.9)	Low
Lim [30]	Early detection	HC=14; DEM=4	≥65	Low
Boa Sorte Silva et al [31]	Early detection	MCI=113	73.1 (5.7)	Low
Kim et al [32]	Progression monitoring	HC=41; NCD=54	Median 60	Low
Basta et al [33]	General insight	MCI=40; DEM=49	78.3 (6.5)	Low
Roh et al [34]	Progression monitoring	MCI=70; DEM=30	Median 73	Low
Liu et al [35]	Early detection	HC=31; MCI=68	66.5	Low
Espinosa et al [36]	Early detection	SCI=44; MCI=99	67.4 (7.9)	Medium
Manning et al [37]	Risk assessment	HC=7; MCI=2; DEM=5	65.8	Medium
Ghosal et al [38]	Early detection	HC=54; DEM=38	73.4 (7.1)	Medium
Park et al [39]	Early detection	MCI/DEM=177; HC=14,482	74.1 (7); 39 (2.3)	Medium
Palmer et al [40]	Early detection	SCI=23; MCI=44	65.8 (7.9)	Medium
Wei et al [41]	Early detection	HC=20; DEM=8	Mean 59.2	Medium
Corbi and Burgos [42]	Early detection	HC=37; DEM=74	80.5	Medium

[a]HC: healthy controls.

[b]DEM: dementia.

[c]NCD: neurocognitive disorder.

[d]MCI: mild cognitive impairment.

[e]SCI: subjective cognitive impairment.

Characteristics of Included Studies

A total of 49 studies were included in the review (summarized in Tables 3-8). Sample sizes ranged from 14 to 91,948 participants, with a median of 145. A total of 5 large cohorts heavily influenced the mean values (ranging between 47,371 and 91,948 participants). When stratified by analytic approach, studies using traditional statistical methods (36/49, 73.5%) had a median of 100 participants (mean 2597, SD 13,574), while those applying machine learning (7/49, 14.3%) had a median of 99 participants (mean 114.9, SD 76.4). Studies using deep learning (6/49, 12.2%) had a median of 108 participants (mean 397.6, SD 465.4), largely driven by population-scale datasets. Most studies were rated as low risk (34/49, 69.4%), nearly one-third (13/49, 26.5%) were classified as medium risk, and 2/49 (4.1%) as high risk. Participant populations were predominantly composed of individuals with MCI (29/49, 59.2%), dementia (35/49, 71.4%), or healthy controls (36/49, 73.5%). A smaller subset included participants with subjective cognitive decline (4/49, 8.2%) or other at-risk groups (2/49, 4.1%). Regarding study design, half of the studies were cross-sectional (21/49, 42.9%), followed by longitudinal cohorts (24/49, 49%), with fewer randomized controlled trials or diagnostic studies. Most studies (28/49, 57.1%) focused on early detection, while others assessed risk factors (8/49, 19.5%), monitored disease progression (7/49, 14.3%), or preventive interventions (5/49, 10.2%). Risk of bias (complete tables are provided in Multimedia Appendix 5) was generally low (35/49, 71.4%), though higher risk ratings were associated with small sample sizes and lack of external validation.

XSL•FO
RenderX

Table 3. Summary of included cohort studies investigating wearable-derived measures in relation to dementia outcomes. The table reports study source, study objective, cohort study design, participant groups and mean ages, and assessed risk of bias.

Source	Objective	Groups, n	Age (years), mean (SD)	Risk of bias
Baril et al [43]	Progression monitoring	HC[a]=159; MCI[b]=44	68.2 (5.4)	Low
Sun et al [44]	Early detection	HC=807; DEM[c]=270	80.9 (7.3)	Low
Winer et al [45]	Early detection	HC=82,377; DEM=452	62.0 (7.8)	Low
Haghayegh et al [46]	Early detection	HC=90,962; MCI or DEM=555	62.4 (7.8)	Low
Basta et al [47]	Progression monitoring	HC=29; MCI=49; DEM=32	80.3 (6.6)	Low
Milton et al [48]	Risk assessment	HC=476; MCI=164; DEM=93	82.5 (2.9)	Low
Skourti et al [49]	Progression monitoring	HC=146; MCI=231; DEM=128	72.8 (6.7)	Low
Gao et al [50]	Early detection	HC=409; MCI=61; DEM=81	58 (8)	Low
Jeon et al [51]	Early detection	HC=100; MCI=29	69.3 (7.7)	Low
Yi Lee et al [52]	Preventive intervention	HC=123; MCI=51	75.6 (6.9)	Low
Lysen et al [53]	General insight	HC=1262; DEM=60	66.1 (7.6)	Low
Agudelo et al [54]	Early detection	DEM=1035	55.2 (2.5)	Low
Ning et al [55]	Risk assessment	HC=91,212; DEM=736	Median 63	Low
Ning et al [56]	Risk assessment	HC=89,619; DEM=710	Median 64.9	Low
Zhao et al [57]	Risk assessment	HC=87,857; DEM=735	61.9 (7.9)	Low
Lu et al [58]	Early detection	HC=303; MCI=130; DEM=277	81.1 (5.2)	Low
Chan et al [59]	Early detection	HC=46,984; DEM=387	67.0 (4.0)	Low
Shi et al [60]	Early detection	HC=346; DEM=344	70.1 (6.9)	Low
Xiao et al [61]	Progression monitoring	HC=570; MCI=120; DEM=73	84.1	Low
Hoepel et al [62]	Preventive intervention	HC=1849; DEM=50	71.3 (9.26)	Low
Cho et al [63]	Early detection	DEM=222	80.4 (7.4)	Medium
Plotogea et al [64]	Early detection, preventive intervention	HC=25; SCI[d]=7; MCI=17; DEM=25	58.9 (9.8)	Medium
Targa et al [65]	Early detection	DEM=100	Median 76	Medium
Cho et al [66]	Risk assessment	DEM=145	81.2 (6)	Medium

[a]HC: healthy controls.

[b]MCI: mild cognitive impairment.

[c]DEM: dementia.

[d]SCI: subjective cognitive impairment.

Table 4. Summary of included studies with other study designs investigating wearable-derived measures in relation to dementia outcomes. The table reports study source, study objective (early detection or progression monitoring), study design (diagnostic study, randomized controlled trial, or quasi-experimental study), participant groups and ages, and assessed risk of bias.

Source	Objective	Study design	Groups, n	Age (years)	Risk of bias
Rykov et al [67]	Early detection	Quasi-experimental study	MCI[a]=17	60.3 (SD 4.5)	Medium
David et al [68]	Progression monitoring	Randomized controlled trial	DEM[b]=38	70 (SD 7)	Medium
Khosroazad et al [69]	Early detection	Diagnostic study	HC=45; MCI[c]=50	73.6	High
Hossain et al [70]	Early detection	Diagnostic study	HC=22; MCI=6; DEM=5	57.5	High

[a]HC: healthy controls.

[b]DEM: dementia.

[c]MCI: mild cognitive impairment.

XSL•FO
RenderX

Table 5. Wearable devices, recording parameters, and analytic methods used across the included studies. The table lists device type, monitoring duration, product and model, analytic approach (statistical, machine learning, deep learning, or clustering), primary task, and cognitive assessments applied.

Source	Wearable category	Wearable	Product (company name)	Duration (days), n	Model	Cognitive test
Espinosa et al [36]	Research	Actigraph	Actiwatch Spectrum (Philips)	13.7	Statistical	MMSE[a]
Lentzen et al [22]	Commercial research	Smartwatch	Fitbit Charge 3 (Google); AX3 (Axivity); Physilog (Gait Up)	56	Machine learning	Amsterdam IADL[b]
Manning et al [37]	Research	Actigraph	Actiwatch (Philips)	14	Statistical	MoCA[c]
Jones et al [23]	Commercial	Smartwatch	Apple Watch 8 (Apple)	6.9	Statistical	HVLT[d]
Satomi et al [24]	Research	Actigraph	ActTrust (Condor Instruments)	7	Statistical	CDR[e]
Baril et al [43]	Research	Actigraph	Actiwatch (Philips)	6.5	Statistical	RBANS[f]
Sauers et al [25]	Research	Actigraph	Sleep Profiler (Advanced Brain Monitoring); Actiwatch2 (Philips); Alice PDx (Philips)	6	Statistical	CDR
Sun et al [44]	Research	Actigraph	Actical (Philips)	14	Deep learning	NINCDS-ADRDA[g]
Winer et al [45]	Research	Accelerometer	AX3 (Axivity)	7	Statistical	TMT[h]
[67]	Multimodal research	Wrist-worn device	E4 wristband (Empatica)	10	Machine learning	NTB[i]
Haghayegh et al [46]	Research	Accelerometer	AX3 (Axivity)	6	Statistical	*ICD-10* [j]
Basta et al [47]	Ad hoc	Actigraph	Not specified	7	Statistical	*ICD-10*
Baril et al [26]	Research	Actigraph	Actiwatch (Philips)	6.5	Statistical	MMSE
Milton et al [48]	Multimodal research	Actigraph	SleepWatch-O (Ambulatory Monitoring)	3	Statistical	MMSE
Cho et al [63]	Research	Actigraph	ActiGraph wGT3X-BT (ActiGraph)	14	Machine learning	MMSE
Seo et al [27]	Research	Accelerometer	HW-100 (Kao Corporation)	28	Statistical	MMSE
Huang et al [28]	Research	Actigraph	GENEActiv Original (Activinsights Company)	6	Statistical	MMSE
Khosroazad et al [69]	Research	Actigraph	Actiwatch (Philips)	7	Deep learning	HVLT
Skourti et al [49]	Research	Actigraph	Actigraph GT3XP (ActiGraph)	3	Statistical	RAVLT[k]
David et al [68]	Commercial	Smartwatch	Fitbit Charge 2 (Google)	13	Statistical	MoCA
Gao et al [50]	Research	Accelerometer	AX3 (Axivity)	7	Statistical	*ICD-10*
Ghosal et al [38]	Research	Accelerometer	Actigraph GT3x+ (ActiGraph)	7	Machine learning	CDR
Palmer et al [29]	Research	Actigraph	Actiwatch Spectrum (Philips)	10.5	Deep learning	MMSE
Jeon et al [51]	Research	Actigraph	Actiwatch 2 (Philips)	5.5	Statistical	MMSE
Lim et al [30]	Ad hoc	Smartwatch	Not specified	7	Deep learning	MMSE
Plotogea et al [64]	Research	Actigraph	Actiwatch Spectrum Pro (Philips)	7	Statistical	PHES[l]
Boa Sorte Silva et al [31]	Research	Actigraph	MotionWatch8 (CamNtech)	7	Statistical	MoCA
Park et al [39]	Commercial	Accelerometer	Fitmeter (FitNLife); ActiGraph AM-7164 (ActiGraph)	24.6 7	Deep learning	MMSE; CDR; GDS[m]
Palmer et al [40]	Research	Actigraph	Actiwatch Spectrum (Philips)	7	Statistical	MMSE
Kim et al [32]	Research	Actigraph	Spectrum Plus (Philips); Actiwatch Spectrum (Philips)	10.5	Statistical	NIH[n]

XSL•FO
RenderX

Source	Wearable category	Wearable	Product (company name)	Duration (days), n	Model	Cognitive test
Targa et al [65]	Research	Actigraph	Actiwatch 2 (Philips)	14	Statistical	MMSE
Wei et al [41]	Commercial	Smartwatch	Mi Band 2 (Xiaomi)	14	Statistical	MMSE
Hossain et al [70]	Ad hoc	Smartwatch	Not specified	182.5	Machine learning	MMSE
Cho et al [66]	Ad hoc	Actigraph	wGT3X-BT (ActiGraph)	11.5	Statistical	MMSE
Corbi and Burgos [42]	Research	Accelerometer	AX3 (Axivity)	4	Machine learning	GDS
Basta et al [33]	Research	Actigraph	Actigraph GT3XP (ActiGraph)	3	Statistical	MMSE
Yi Lee et al [52]	Research	Actigraph	GENEActiv Original (ActivInsights)	7	Statistical	MoCA
Roh et al [34]	Commercial	Accelerometer	Fitmeter (FitNLife)	4	Statistical	SNSB[o]
Lysen et al [53]	Research	Actigraph	ActiWatch AW4 (CamNtech)	6	Statistical	MMSE
Agudelo et al [54]	Research	Actigraph	Actiwatch Spectrum (Philips)	7	Statistical	B-SEVLT-Sum[p]
Ning et al [55]	Research	Accelerometer	AX3 (Axivity)	7	Statistical	*ICD-10*
Ning et al [56]	Research	Accelerometer	AX3 (Axivity)	7	Statistical	*ICD-10*
Zhao et al [57]	Research	Accelerometer	AX3 (Axivity)	7	Statistical	*ICD-10*
Lu et al [58]	Research	Actigraph	wGT3X-BT (Ametris)	7	Statistical	MoCA
Chan et al [59]	Research	Smartwatch	Not specified	7	Statistical	*ICD-10*
Shi et al [60]	Research	Actigraph	wGT3X-BT (Ametris); Actiwatch Spectrum (Philips)	7; 3	Deep learning	MoCA
Liu et al [35]	Commercial	Accelerometer	W180 (Shenzhen Fitfaith)	14	Machine learning	MoCA
Xiao et al [61]	Research	Actigraph	Actigraph GT3XP (Ametris)	7	Statistical	TMT
Hoepel et al [62]	Research	Actigraph	GENEActiv Original (ActivInsights)	4	Statistical	MMSE

[a]MMSE: Mini-Mental State Examination.
[b]Amsterdam IADL: Amsterdam Instrumental Activities of Daily Living.
[c]MoCA: Montreal Cognitive Assessment.
[d]HVLT: Hopkins Verbal Learning Test-Revised.
[e]CDR: Clinical Dementia Rating.
[f]RBANS: Repeatable Battery for the Assessment of Neuropsychological Status.
[g]NINCDS-ADRDA: National Institute of Neurologic, Communicative Disorders and Stroke and Alzheimer's Disease and Related Disorders Association.
[h]TMT: Trail-Making Test.
[i]NTB: Neuropsychological Test Battery.
[j]ICD-10: International Statistical Classification of Diseases, Tenth Revision.
[k]RAVLT: Rey Auditory Verbal Learning Test.
[l]PHES: Psychometric Hepatic Encephalopathy Score.
[m]GDS: Global Deterioration Scale.
[n]NIH: National Institutes of Health.
[o]SNSB: Seoul Neuropsychological Screening Battery.
[p]B-SEVLT-Sum: Brief Spanish-English Verbal Learning Test.

XSL•FO
RenderX

Table 6. Summary of machine learning used across studies, including model type, main performance metrics, analytical task (classification, regression, or clustering), and the corresponding feature selection or explainable artificial intelligence (XAI) techniques applied. Only the best-performing metric for each model is reported.

Source	Model	Metrics (best results)	Task	Feature selection/XAI
Lentzen et al [22]	LR[a]; DT[b]; RF[c]; XGBoost[d]	AUC[e]=0.73	Binary classification	SHAP[f]
Rykov et al [67]	LR; RF; XGBoost	R^2=0.690; ρ=0.700; MAE[g]=0.460	Regression	Correlation-based feature selection
Cho et al [63]	LR; RF; GBM[h]; SVM[i]	AUC=0.929; accuracy=0.935; precision=0.800; sensitivity=0.956; F_1-score (0.819)	Binary classification	Permutation feature importance
Ghosal et al [38]	GLM[j]; SOFR[k]; SOTDR[l]; SOTDR-L[m]	AUC=0.811; R^2=0.333	Binary classification; regression	LASSO[n]/GEL[o] penalties + functional coefficients
Hossain et al [70]	GBM; SVM; LR + LASSO; CTGAN[p]	Accuracy=0.948	Multiclass classification; regression	Correlation-based feature selection; wrapper feature selection
Corbi and Burgos [42]	Expectation-maximization clustering	Accuracy=0.910	Unsupervised clustering	Feature relevance via variable testing
Liu et al [35]	GBDT[q]; XGBoost[r]	Accuracy=0.757; recall=0.952; AUC=0.628	Binary classification	Permutation importance

[a]LR: logistic regression.
[b]DT: decision tree.
[c]RF: random forest.
[d]XGBoost: Extreme Gradient Boosting.
[e]AUC: area under the curve.
[f]SHAP: Shapley additive explanations.
[g]MAE: mean absolute error.
[h]GBM: gradient boosting machine.
[i]SVM: support vector machine.
[j]GLM: generalized linear model.
[k]SOFR: scalar-on-function regression.
[l]SOTDR: scalar on time-by-distribution regression.
[m]SOTDR-L: SOTDR via time varying L moments.
[n]LASSO: least absolute shrinkage and selection operator.
[o]GEL: group exponential LASSO.
[p]CTGAN: conditional tabular generative adversarial network.
[q]GBDT: gradient boosting decision tree.
[r]XGBoost: Extreme Gradient Boosting.

XSL•FO
RenderX

Table 7. Summary of deep learning used across studies, including model type, main performance metrics, analytical task (classification, regression, or survival analysis), and the corresponding feature selection or explainable artificial intelligence (XAI) techniques applied. Only the best-performing metric for each model is reported.

Source	Model	Metrics (best results)	Task	Feature selection/XAI
Sun et al [44]	CNN[a]; ElasticNet; RSF[b]	C-index=0.840; AUC[c]=0.880	Survival analysis (time to Alzheimer disease onset)	Gradient and Gini feature importance; hazard ratio interpretability
Khosroazad et al [69]	Neural network	AUC=0.880; sensitivity=0.870; specificity=0.890	Binary classification	Intrinsic via time-latency
Palmer et al [29]	MS-GAN[d]; Bayesian LR[e]	Dice=0.730; OR[f]=1.830	Segmentation; regression	Bayesian coefficients (odds ratio and CI)
Lim et al [30]	Neural network + PCA[g]	AUC=0.990	Binary classification	Correlation-based interpretability
Park et al [39]	1D convolutional autoencoder + LR	R^2=0.979	Regression	Backward feature elimination
Shi et al [60]	CDPred[h]; XGBoost[i]	Hip: accuracy=0.84 and AUC=0.86; Wrist: accuracy=0.69 and AUC=0.73	Binary classification	Nonzero predictor relative importance ranking

[a]CNN: convolutional neural network.

[b]RSF: random survival forest.

[c]AUC: area under the curve.

[d]MS-GAN: multispecies generative adversarial network.

[e]LR: logistic regression.

[f]OR: odds ratio.

[g]PCA: principal component analysis.

[h]CDPred: cognitive decline predictor.

[i]XGBoost: Extreme Gradient Boosting.

XSL•FO
RenderX

Table 8. Summary of statistical approaches applied across studies, including model type, reporting metrics, covariates or adjustment factors, and main associations between wearable-derived variables and cognitive impairment or dementia indicators. Only the most relevant and statistically significant results for each study are presented. For wearable-derived metrics, arrows indicate the direction of association: an upward arrow (↑) denotes that higher values of the variable are associated with greater cognitive impairment, whereas a downward arrow (↓) indicates the opposite.

Source	Model	Metrics	Score
Espinosa et al [36]	General linear models	CWP[a]	IV[b] (CWP<.001, ↑); IS[c] (CWP<.001, ↓)
Manning et al [37]	Pearson correlation analysis	Pearson CC[d], *P* value	Activity counts (CC=–0.829; *P*=.041, ↓)
Jones et al [23]	Linear regression models	β coefficient, 95% CI	IV (β≈.40, q=.022, ↑); sleep onset (β=–.28, 95% CI –0.55 to –0.02, ↑); IS (β=–.27, 95% CI –0.54 to 0.00, ↓)
Satomi et al [24]	Multinomial logistic regression	RR[e], 95% CI, *P* value	IV (RR 3.21, 95% CI 1.40-7.34, *P*=.006, ↑); IS (RR 0.44, 95% CI 0.21-0.93, *P*=.03, ↓); M10 (RR 0.40, 95% CI 0.18-0.89, *P*=.02, ↓)
Baril et al [43]	Linear regression models	Standardized β coefficient, *P* value	Sleep duration (β=.384, *P*=.001, ↑)
Sauers et al [25]	Linear regression models	Estimate (β), *P* value	Sleep efficiency (β=–6.026, *P*<.001, ↓); sleep latency (β=11.302, *P*<.001, ↑); number of awakenings (β=6.585, *P*=.001, ↑)
Winer et al [45]	Cox proportional hazards models	HR[g], 95% CI, *P* value	IS (HR 1.25, 95% CI 1.050-1.480; *P*=.007, ↑); amplitude (HR 0.79, 95% CI 0.650-0.960, *P*=.02, ↓); M10 (HR 0.75, 95% CI 0.610-0.940, *P*=.01, ↓); MESOR[h] (HR 0.78, 95% CI 0.590-0.998, *P*=.048, ↓)
Haghayegh et al [46]	Cox proportional hazard models	HR, 95% CI	Amplitude (HR 1.32, 95% CI 1.17-1.49, ↑); M10 (HR 1.28, 95% CI 1.14-1.44, ↑); L5 (HR 1.15, 95% CI 1.10-1.21, ↑); IV (HR 1.14, 95% CI 1.05-1.24, ↑); rest-activity rhythm (HR 1.23, 95% CI 1.16-1.29, ↑)
Basta et al [47]	ANCOVA	*P* value	Sleep duration (night TST[j], *P*<.001, ↑); TiB[k] (night TiB, *P*<.001, ↑)
Baril et al [26]	Linear regression models	*P* value	Sleep duration (*P*<.05, ↑); activity counts (*P*<.05, ↑); circadian rhythm (*P*<.05, ↑)
Milton et al [48]	Multinomial logistic regression	OR[l], 95% CI	Wake after sleep onset (OR 2.26, 95% CI 1.12-4.55, ↑); sleep efficiency (OR 2.15, 95% CI 1.03-4.47, ↓)
Seo et al [27]	Two-way ANCOVA	*P* value	Movement/acceleration (*P*=.03, ↓)
Huang et al [28]	Unconditional multivariable logistic regression	AOR[m], 95% CI	MESOR (AOR 1.99, 95% CI 1.04-3.81, ↑)
Skourti et al [49]	Path models (analysis of moment structures)	Standardized β coefficient, *P* value	Sleep efficiency (direct β=.266, *P*=.001, ↓); wake after sleep onset (direct β=–.211, *P*=.001, ↑); TiB (24-hour TiB, indirect β=–.079, *P*<.001, ↑)
David et al [68]	Spearman rank correlation	Partial correlation coefficient (ρ), *P* value	Moderate-to-vigorous physical activity (ρ=0.558, *P*=.02, ↓)
Gao et al [50]	Cox proportional hazards	HR, 95% CI	Amplitude (HR 1.94, 95% CI 1.53-2.46, *P*<.001, ↑); IV (HR 1.49, 95% CI 1.18-1.88, *P*<.001, ↑); sleep duration (HR 1.28, 95% CI 1.06-1.55, *P*=.01, ↑)
Jeon et al [51]	Multivariate linear regressions	β coefficient, *P* value	Acrophase (β=–.256, *P*=.004, ↑)
Plotogea et al [64]	Multivariate logistic regression	OR, 95% CI, *P* value	Sleep efficiency (OR 0.803, 95% CI 0.711-0.904, *P*=.001, ↓); sleep latency (OR 1.212, 95% CI 1.063-1.383, *P*=.004, ↑)
Boa Sorte Silva et al [31]	Linear regression models	Unstandardized β (β), *P* value	Fragmentation index (β=.004, *P*=.046, ↑)
Palmer et al [40]	Fixel-wise linear regression; Bayesian multiple linear regression	β coefficient, *P* value	L5 (β=.29, *P*<.050, ↑)
Kim et al [32]	Multivariable linear models	Partial rank correlation (ρ), *P* value	IV (ρ=–0.44, *P*<.001, ↑); M10 (ρ=0.45, *P*<.001, ↓); IS (ρ=0.40, *P*=.009, ↓)
Targa et al [65]	Linear regression models	Effect size, *P* value	IV (effect size=–0.715, *P*=.013, ↑)
Wei et al [41]	Descriptive statistics; 1-tailed *t* test	Mean (SD), *P* value	Amplitude (0.93, SD 0.59, *P*=.030, ↓); IS (0.32, SD 0.19, *P*=.02, ↓); acrophase (44, SD 145, *P*<.001, ↓)
Cho et al [66]	Generalized linear mixed model	OR, 95% CI, *P* value	Sleep duration (OR 0.9, 95% CI 0.8-1.0, *P*<.001, ↓); activity counts (OR 0.02, 95% CI 0.0-0.5, *P*=.02, ↓)

XSL•FO
RenderX

Source	Model	Metrics	Score
Basta et al [33]	ANCOVA	Mean (SD), *P* value	Sleep duration (night TST=7.7-hour vs 7.2-hour, *P*=.011, ↑); sleep duration (24-hour TST=8.5-hour vs 7.8-hour, *P*=.012, ↑)
Yi Lee et al [52]	Multivariate logistic regression; multinomial logistic regression	AOR, 95% CI	Percent rhythm (AOR 0.26, 95% CI 0.08-0.79, ↓)
Roh et al [34]	Multiple linear regression	Estimate (β), SE, *P* value	MESOR (β=1.17, SE=0.37, *P*<.001, ↓); L5 (β=3.77, SE=1.22, *P*<.001, ↓)
Lysen et al [53]	Cox proportional hazards	HR, 95% CI	Sleep latency (HR 1.44, 95% CI 1.13-1.83, ↑); TiB (HR 1.40, 95% CI 1.04-1.88, ↑); sleep efficiency (HR 0.72, 95% CI 0.55-0.93, ↓); wake after sleep onset (HR 1.38, 95% CI 1.10-1.74, ↑)
Agudelo et al [54]	Survey linear regression models	β coefficient, *P* value	Sleep latency (β=–.003, *P*<.001, ↑); sleep duration (β=–.070, *P*<.05, ↑)
Ning et al [55]	Cox proportional hazards	HR, 95% CI	Moderate-to-vigorous physical activity (HR 0.69, CI 0.54-0.87, *P*<.001, ↓)
Ning et al [56]	Cox proportional hazards	HR, 95% CI	Moderate-to-vigorous physical activity (HR 0.60, 95% CI 0.40-0.90, ↓)
Zhao et al [57]	Cox proportional hazards	HR, 95% CI	Sleep duration (HR 0.801, 95% CI 0.717-0.893, ↓)
Lu et al [58]	Logistic regression models ANCOVA	OR, 95% CI	Relative amplitude (OR 1.68, 95% CI 1.12-2.50, ↑)
Chan et al [59]	Cox proportional hazards	HR, 95% CI	Bedtime (HR 1.52, 95% CI 1.22-1.85, ↑)
Xiao et al [61]	Cox proportional hazards	HR, 95% CI	MESOR (HR 2.45, 95% CI 1.52-3.94, ↓)
Hoepel et al [62]	Cox proportional hazards	HR, 95% CI	Sedentary behavior (HR 0.36, 95% CI 0.24-0.55, ↓)

[a]CWP: clusterwise *P* value.
[b]IV: intradaily variability.
[c]IS: interdaily stability.
[d]CC: correlation constant.
[e]RR: relative risk.
[f]M10: most active 10-hour.
[g]HR: hazard ratio.
[h]MESOR: midline estimated statistic of rhythm.
[i]L5: least active 5-hour.
[j]TST: total sleep time.
[k]TiB: time in bed.
[l]OR: odds ratio.
[m]AOR: adjusted odds ratio.

Feasibility of Quantitative Synthesis

The feasibility of quantitative meta-analysis for early detection depends on whether the included studies estimate a common outcome construct or estimand. Quantitative synthesis requires studies to measure the same underlying construct. In the current literature, early detection is not operationalized as a single measurable outcome but is addressed through multiple, nonequivalent analytical objectives and metrics. The evidence base is characterized by the following dimensions:

Reported statistical metric: of the 49 included studies, 13 (26.5%) report performance metrics from AI-based approaches, such as accuracy or area under the curve. The remaining 36 studies report heterogeneous statistical effect measures, including hazard ratios for time-to-event analyses (n=16), odds ratios (ORs) for binary risk (n=8), and beta coefficients or correlation measures for continuous associations (n=12). These metrics reflect distinct inferential targets and are not interchangeable.

Device and sensor characteristics: most studies rely on research-grade actigraphy with access to raw accelerometry data (43/49, 87.8%), whereas a smaller subset uses consumer-grade wearables integrating additional sensors and proprietary processing (7/49, 14.3%). Differences in device type, sensor modality, and preprocessing lead to wearable-derived measures that do not map onto a common estimand.

Study design: study designs were divided between cross-sectional analyses assessing concurrent associations (21/49, 42.8%) and longitudinal cohort studies estimating prospective risk or disease onset (24/49, 49%). These designs address different research questions and operate over distinct temporal frameworks.

XSL•FO
RenderX

Clinical endpoint: cognitive outcomes included both categorical clinical diagnoses and continuous screening scores. The MMSE was the most frequently used outcome (18/49, 36.7%), followed by *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*)–based diagnoses and the MoCA (each 7/49, 14.3%). These endpoints differ in scale, sensitivity, and clinical interpretation.

Because early detection is not defined as a single measurable construct across studies, no common estimand can be specified for quantitative pooling. A structured narrative synthesis is therefore adopted to examine how analytical approaches, cognitive assessments, and wearable measurement strategies are used to demonstrate early detection capability across the cognitive continuum.

Wearable Categories

Wearable devices were categorized into 4 groups according to their primary intended use. Research devices were by far the most common, reported in 87.8% (43/49) of studies, and included systems such as ActiGraph, Actiwatch, and accelerometer loggers (eg, Axivity AX3). These devices were purpose-built and validated for assessing activity, sleep, and circadian rhythms in research and clinical contexts. Commercial everyday wearables, such as Fitbit, Apple Watch, and Xiaomi Mi Band, were used in 14.3% (7/49) of studies; these multipurpose consumer devices are affordable and widely accessible but often provide limited access to raw data. Ad hoc prototypes, developed specifically for individual studies, accounted for 8.2% (4/49) of studies. Finally, multimodal research devices were used in 4.1% (2/49) of studies, typically integrating physiological sensing beyond accelerometry, such as heart rate, electrodermal activity, or peripheral arterial tonometry. Counts are nonexclusive because some studies used multiple device types in the same cohort [19].

Across all categories, wear time was typically about 1 week (median, 7 days), with durations ranging from 3-4 days to several months (up to 182.5 days in 1 longitudinal cohort). Most of the studies (29/49, 59.2%) relied on actigraphy devices to capture rest–activity rhythms and sleep, followed by accelerometer devices (12/49, 25.5%) and smartwatches (7/49, 14.3%).

Cognitive Outcomes

The most frequently used cognitive measures were global screening tests, with their distribution shown in Figure 3. The MMSE was the most frequently used measure, applied in 36.7% (18/49) of studies. Clinical diagnoses based on *ICD-10* dementia or MCI codes were reported in 14.3% (7/49) of studies. The MoCA appeared in 14.3% (7/49) of studies and Clinical Dementia Rating was used in 8.2% (7/49) of studies. The Hopkins Verbal Learning Test and the Global Deterioration Scale were each used in 4.1% (2/49) of studies. A wide range of other neuropsychological and functional instruments, including the Amsterdam Instrumental Activities of Daily Living scale, Repeatable Battery for the Assessment of Neuropsychological Status, National Institute of Neurologic, Communicative Disorders and Stroke and Alzheimer's Disease and Related Disorders Association criteria, Trail Making Test, Neuropsychological Test Battery, Rey Auditory Verbal Learning Test, Psychometric Hepatic Encephalopathy Score, National Institutes of Health toolbox, Seoul Neuropsychological Screening Battery, and Brief Spanish-English Verbal Learning Test-Sum, were each applied in 2% (1/49) of studies. The overall distribution of cognitive measures is shown in Figure 3.

One study applied AI-based approaches to directly predict cognitive test scores [67], while others relied on predefined cut-off points for classification [23,36,41,46,55,57]. Some studies used both direct score estimation and classification approaches [43,54]. A separate line of work focused on predicting survival outcomes, which were not directly comparable to cognitive test scores [43]. Finally, 2 studies addressed related regression tasks by estimating white matter characteristics in the brain from wearable-derived data [44,49].

Across the 49 studies, the most frequently reported wearable-derived variables (complete list is provided in Multimedia Appendix 6) used to estimate cognitive assessments (Figure 4) were predominantly sleep-related. Sleep duration was the most frequently reported measure (22/49, 44.9%), followed by wake after sleep onset (16/49, 32.7%), sleep efficiency (16/49, 32.7%), and circadian rhythm metrics (16/49, 32.7%). Activity-related variables were reported less often overall, including movement/acceleration (12/49, 24.5%), activity counts (9/49, 18.4%), step count (7/49, 14.3%), moderate-to-vigorous physical activity (6/49, 12.2%), active minutes (5/49, 10.2%), and sedentary behavior (5/49, 10.2%). Physiological variables were least commonly reported, with skin temperature appearing in a small subset of studies (5/49, 10.2%).

XSL•FO
RenderX

Figure 3. Frequency of cognitive assessment instruments across included studies. The bar chart illustrates the distribution of specific cognitive tests and diagnostic criteria used. Frequencies and percentages represent the prevalence of each tool within the total sample of studies. Amsterdam IADL: Amsterdam Instrumental Activities of Daily Living; B-SEVLT-Sum: Brief Spanish-English Verbal Learning Test; CDR: Clinical Dementia Rating; GDS: Global Deterioration Scale; HVLT: Hopkins Verbal Learning Test-Revised; ICD-10: International Classification of Diseases, Tenth Revision; MMSE: Mini-Mental State Examination; MoCA: Montreal Cognitive Assessment; NIH: National Institutes of Health; NINCDS-ADRDA: National Institute of Neurologic, Communicative Disorders and Stroke and Alzheimer's Disease and Related Disorders Association; NTB: Neuropsychological Test Battery; PHES: Psychometric Hepatic Encephalopathy Score; RAVLT: Rey Auditory Verbal Learning Test; RBANS: Repeatable Battery for the Assessment of Neuropsychological Status; SNSB: Seoul Neuropsychological Screening Battery; TMT: Trail-Making Test.

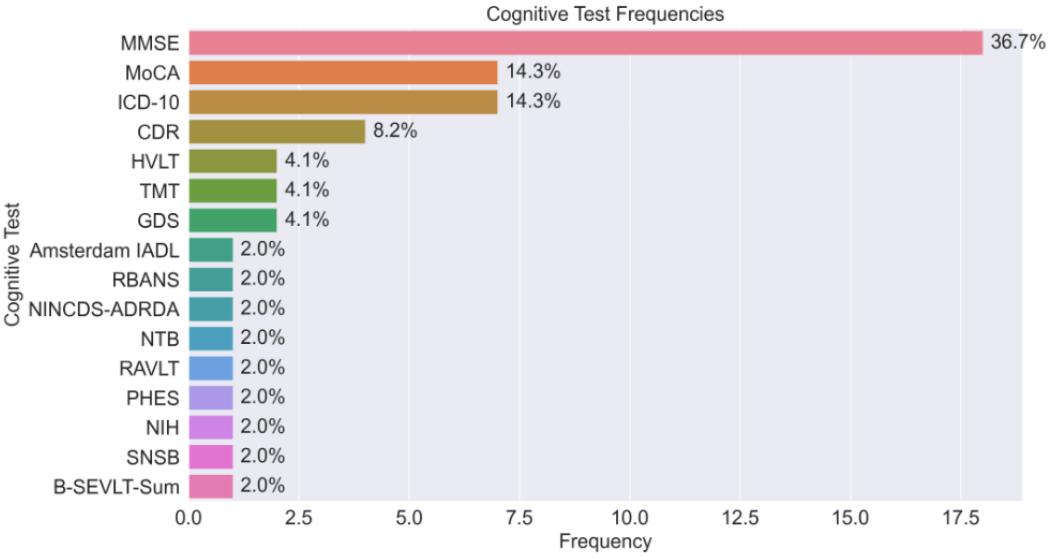

Figure 4. Top 20 most frequently used wearable-derived variables across studies. Variables are grouped by domain: sleep (blue), activity (green), and physiological (red). Frequencies represent the number of studies reporting each variable after standardization and filtering for wearable compatibility.

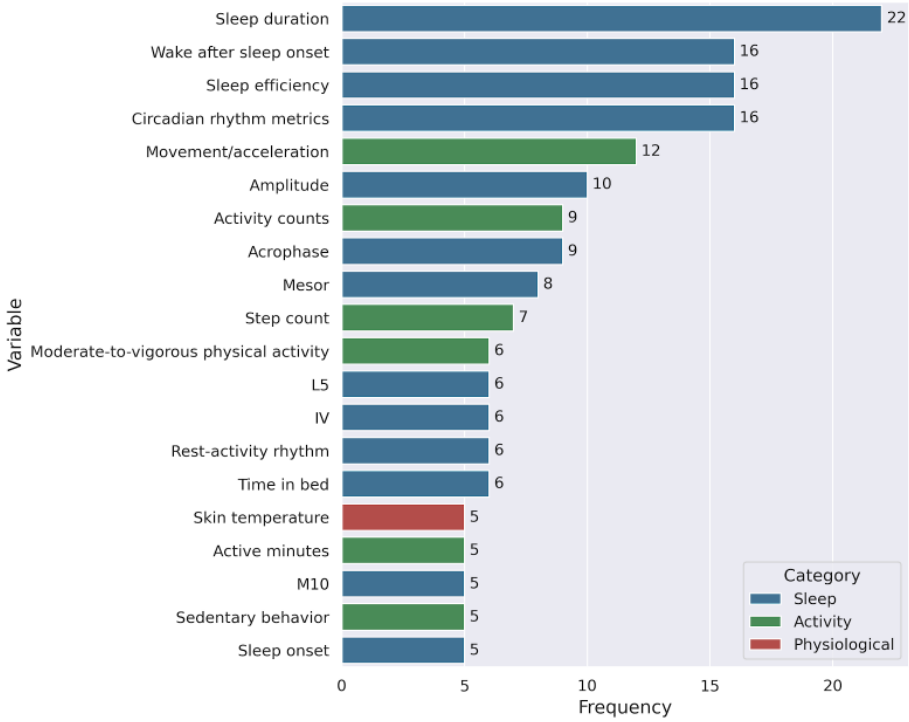

Wearable Data Analysis Approaches

Traditional statistical methods were the most frequently applied, reported in 73.5% (36/49) of studies (Table 3). These included group comparisons, correlation analyses, and regression models evaluating associations between wearable-derived activity or sleep metrics and cognitive outcomes. Commonly reported features were total daily activity counts, sleep duration and efficiency, circadian rhythm indices (intradaily variability, interdaily stability, and relative amplitude), and timing markers such as acrophase or sleep midpoint. The list of wearable-derived features used across these studies, with a description per feature, is included in Multimedia Appendix 6.

XSL•FO
RenderX

Most studies relying on conventional statistical approaches reported significant associations between wearable-derived variables and cognitive outcomes, although the magnitude of reported effects varied with sample size. In smaller studies (n<100), standardized regression coefficients and odds ratios tended to be larger, typically ranging from β≈.35-.55 or OR 2.0-3.1 [24,29,43]. In larger cohort studies (n≥100), effect estimates were generally lower, around β≈.10-.25 or OR 1.3-1.8 [45,46,49,54].

According to established benchmarks for social and behavioral sciences, standardized coefficients are typically categorized as small (β≈.20), medium (β≈.50), or large (β≈.80) [71]. While the estimates from larger cohorts in this review fall within the small-to-moderate range, they must be interpreted within the specific context of cognitive research. Given that cognitive performance is a multifactorial outcome influenced by a broad array of genetic, environmental, and demographic variables, the observation of stable effect sizes between .10 and .25 is highly meaningful. Such values represent robust contributions to the variance that, despite their conservative magnitude, suggest that wearable-derived metrics are reliable markers of cognitive health at a population level. While most associations reached statistical significance (P<.05), only a limited number of studies reported fully adjusted models or external validation [72].

Machine learning methods were applied in 14.3% (7/49) of studies, most often for classification tasks such as distinguishing MCI or dementia from healthy controls. Machine learning models included logistic regression, decision trees, support vector machines, random forests, and gradient-boosting approaches such as Extreme Gradient Boosting (XGBoost). Deep learning approaches were reported in 12.2% (6/49) of studies and were typically applied in larger datasets, where models captured temporal patterns of activity or sleep. Deep learning architectures included neural networks, convolutional neural networks, generative adversarial network–based approaches, and autoencoders. In addition, a small number of studies (2/49, 4.1%) applied clustering techniques to identify subgroups with distinct rest–activity profiles.

Across machine learning and deep learning studies, reported performance metrics varied widely depending on task, cohort size, and study design. Classification performance ranged from moderate discrimination in large, pragmatic settings (eg, area under the receiver operating characteristic curve [AUROC]≈0.73; n=229; dementia=160 and healthy controls=69 [23]) to higher values in smaller or more constrained cohorts (eg, AUROC≈0.93; n=222; affective symptoms=126 and not affective symptoms=96 [63]). Deep learning models reported metrics such as a C-index of approximately 0.84 (n=1077; dementia=270 and healthy controls=807) for long-horizon prediction of Alzheimer disease onset [44] and R^2 values approaching 0.98 (n=14,659; dementia=177 and healthy controls=14,482) for regression tasks estimating brain structural characteristics from wearable-derived data [40]. Several high-performing models were trained on small or imbalanced datasets, including studies with limited case numbers or highly unequal case–control ratios [28,31,40,70].

In addition, machine learning approaches applied to wearable-derived features achieved high discriminative performance across multiple tasks. Reported results included classification accuracies up to 93.5% [63] (n=222; sleep and nighttime behaviors=81 and not sleep and nighttime behaviors=141) for identifying high-risk individuals, accuracies of approximately 94%-95% (n=33; dementia=5, MCI=6, and healthy controls=22) across multiple cognitive impairment levels [70], area-under-the-curve values ranging from approximately 0.73 (n=229; dementia=160 and healthy controls=69) [23] to 0.99 (n=18; dementia=4 and healthy controls=14) [31] for early or prodromal Alzheimer disease detection, sensitivities and specificities exceeding 85% (n=95; MCI=50 and healthy controls=45) [69] for distinguishing MCI from normal cognition, and C-index values up to 0.84 (n=1077; dementia=207 and healthy controls=807) [44] for long-horizon Alzheimer disease onset prediction.

Interpretability approaches were reported in a limited subset of machine learning and deep learning studies. Explicit explainability techniques were applied in 46.1% (6/13) of AI-based studies, including Shapley additive explanations values or permutation feature importance in 2 machine learning studies [23,49], functional coefficients in 1 statistical–machine learning hybrid approach [31], and Bayesian coefficient estimation in 1 deep learning framework [32]. One additional study reported feature relevance through statistical testing in a clustering-based approach [39]. The remaining AI-based studies relied on feature-selection heuristics or dimensionality reduction methods, such as correlation-based selection, wrapper methods, or principal component analysis, without reporting interpretable model outputs [33,35,67,70].

Prevention-Oriented Findings

The included studies were divided into 2 categories according to the type of contribution reported. A total of 22.4% (11/49) of studies explicitly provided quantitative results relevant to the early detection or prevention of cognitive decline (direct evidence), defined as studies demonstrating or validating predictive or preventive applications. The remaining 38 (77.6%) studies contributed indirect evidence, identifying associations between wearable-derived features and cognitive outcomes that may inform, but do not yet constitute, preventive or predictive applications.

Among studies providing direct evidence (n=11), longitudinal or interventional designs were commonly used. Specifically, 45.5% (5/11) of studies used longitudinal follow-up or randomized designs, while 54.5% (6/11) of studies relied on cross-sectional analyses. Direct-evidence studies included randomized intervention trials reporting outcomes such as slower progression to dementia, improved gait speed, or increased adherence to preventive programs, as well as longitudinal cohort studies showing that disrupted rest–activity rhythms, sleep fragmentation, and physical activity patterns predicted incident dementia or MCI over follow-up periods extending up to 8-15 years. Several studies further demonstrated that wearable-derived motor activity features could forecast clinical Alzheimer disease onset, with reported C-index values ranging from approximately 0.80 to 0.84.

XSL•FO
RenderX

In contrast, studies contributing indirect evidence (n=38) more frequently adopted cross-sectional designs. Of these, 52.6% (20/38) of studies were cross-sectional, whereas 47.4% (18/38) of studies used longitudinal designs. These studies primarily focused on identifying associations between wearable-derived sleep, circadian rhythm, activity, or physiological measures and cognitive outcomes, including global cognitive scores, neuropsychological performance, biomarkers, or clinical diagnoses, rather than on prediction or intervention.

Across all included studies, longitudinal designs accounted for 44.9% (22/49) of studies, while 55.1% (27/49) of studies used cross-sectional designs. This distribution reflects differing study aims, with direct evidence studies more often incorporating longitudinal follow-up or intervention components, and indirect evidence studies emphasizing associative analyses.

Overall, wearable technologies contributed to prevention-oriented research both directly, by demonstrating predictive and early detection capabilities in longitudinal and interventional settings, and indirectly, by identifying behavioral and physiological markers associated with future cognitive decline. This dual contribution reflects the role of wearables as both measurement tools for preventive interventions and sources of early behavioral risk markers in dementia research.

Discussion

Principal Findings

This systematic review synthesized 49 studies published since 2020 on wearable devices for the early detection and prevention of cognitive impairment and dementia. The findings are presented in 4 dimensions corresponding to the subobjectives described earlier.

Wearable Devices and Measurement Contexts

Research-grade actigraphy and accelerometry devices were used in most of the included studies (43/49, 87.8%), reflecting both a strong methodological emphasis on validated access to raw accelerometry data and the relative maturity of actigraphy-based approaches in this field [73,74]. Evidence from these studies was derived predominantly from observational designs, including cross-sectional and longitudinal cohorts, and was frequently assessed as having low risk of bias, particularly among cohort studies with large samples [48,53,60]. Across actigraphy-based studies, sample sizes were heterogeneous but predominantly large, with nearly four-fifths enrolling 100 participants or more, and large samples were strongly associated with low risk of bias. While research-grade devices enable robust investigation of behavioral markers associated with cognitive impairment, their limited scalability and accessibility may constrain broader clinical and public health deployment [75].

In contrast, consumer-grade wearable devices were comparatively underrepresented despite their widespread adoption and potential for large-scale, longitudinal monitoring. Evidence from consumer-wearable studies was derived predominantly from cross-sectional designs and exhibited greater heterogeneity in sample size and methodological quality, with smaller studies more frequently assessed as having moderate risk of bias, reflecting restricted access to raw data and reliance on proprietary signal processing [41,68,76]. However, consumer-wearable evidence also included a small number of large cross-sectional studies assessed as having low risk of bias, indicating that methodological rigor is driven more by study scale and design than by device class alone [22,34,35]. Taken together, these findings suggest that different wearable device classes capture complementary aspects of behavior and physiology rather than interchangeable measurements, supporting the potential value of multimodal integration or purpose-driven device selection aligned with specific research or clinical objectives [76].

Wearable-Derived Features and Cognitive Outcomes

Across included studies, cognition was frequently assessed using brief global screening instruments or standardized clinical diagnostic classifications, reflecting a pragmatic choice aligned with clinical practice and the feasibility requirements of observational research. Evidence derived from commonly used cognitive measures was drawn primarily from cross-sectional and cohort designs and was frequently assessed as having low risk of bias, particularly in larger cohort studies [46,50,55,56]. While global cognitive measures facilitate standardized comparisons across heterogeneous cohorts, their limited sensitivity to subtle or domain-specific changes constrains interpretation of associations with wearable-derived features, especially in preclinical populations [77].

Wearable devices were not used to measure cognition directly but to capture continuous behavioral and physiological signals that were statistically associated with cognitive outcomes. Sleep-related and circadian rhythm measures were examined most frequently and were reported across both cross-sectional and longitudinal studies, predominantly in cohort designs with large samples, and were most often assessed as having low risk of bias when derived from validated actigraphy devices and predefined metrics [33,47,59,62]. In contrast, physiological signals beyond accelerometry were examined in only a small subset of studies, which used heterogeneous, primarily noncohort designs, often involved small to medium sample sizes, and showed higher proportions of moderate-to-high–risk-of-bias assessments [35,41] limiting the strength of inferences that can currently be drawn from these measures and highlighting priorities for future methodological development.

Analytical Approaches and Methodological Considerations

Conventional statistical methods were the predominant analytical approach, used in 73.5% (36/49) of studies, reflecting their interpretability and familiarity in clinical and epidemiological research. These methods were applied mainly in cross-sectional and longitudinal observational designs and were frequently assessed as having low risk of bias, particularly in large cohort studies using multivariable or survival models [45,46,50,55-62]. Across these studies, associations between wearable-derived features and cognitive outcomes were generally modest and sensitive to covariate adjustment, especially in large and heterogeneous samples, underscoring

XSL•FO
RenderX

the importance of adequate sample size, rigorous model specification, and replication for reliable inference.

Machine learning and deep learning approaches were applied in a smaller subset of studies (13/49, 26.5%) and were used primarily for classification or prediction tasks. These studies were frequently exploratory, relied on small or imbalanced samples, and were more often assessed as having moderate to high risk of bias, largely due to limited external validation and optimistic performance estimates [67,69,70]. Across AI-based studies, higher predictive performance was more commonly reported in cross-sectional or nonlongitudinal designs using constrained datasets, whereas applications in larger or longitudinal cohorts generally yielded more modest performance estimates but were more often assessed as having low or moderate risk of bias, reflecting improved robustness and generalizability [35,44]. Overall, this pattern suggests a trade-off between performance and methodological robustness, in which performance gains observed in small or exploratory samples may reflect overfitting or cohort-specific structure rather than generalizable predictive signal.

Implications for Early Detection and Prevention

Evidence supporting early detection or prevention was derived from a limited subset of studies providing direct prevention evidence, which more often used cohort-based or predictive modeling designs, enrolled larger samples, and were predominantly assessed as having low risk of bias. Representative examples include large longitudinal or predictive studies demonstrating that wearable-derived behavioral measures, particularly sleep–wake organization and circadian regularity, can precede clinical cognitive impairment by several years [39,46,56]. These findings support the potential role of wearables in early risk stratification rather than post hoc characterization.

In contrast, most of the included studies contributed indirect prevention evidence and relied primarily on cross-sectional or association-focused designs. Although these studies were frequently assessed as having low to moderate risk of bias and consistently reported associations between wearable-derived markers and cognitive status, their typically limited temporal resolution and lack of predictive validation constrained their ability to establish preventive relevance [24,27,33,40]. As a result, their contribution to prevention remains inferential rather than actionable.

Overall, the current evidence supports wearable devices as tools for monitoring and early risk identification rather than as stand-alone preventive interventions [78]. Across both direct and indirect evidence, disrupted sleep–wake patterns and circadian irregularity consistently emerged as markers of elevated cognitive risk [33,39,55]. Translating these associations into effective prevention strategies will require confirmation in larger, longitudinal, and externally validated studies [78] that integrate wearable-based risk stratification with clinical decision-making and behavioral intervention frameworks.

Limitations

Several limitations should be considered when interpreting these findings. The included studies were heterogeneous in design, target populations, device types, and outcome definitions. Devices ranged from consumer-grade wearables to research actigraphy and multimodal physiological instruments, and the metrics derived from them were inconsistently defined. Cognitive outcomes also varied: global screening tools such as the MMSE and MoCA were most frequent, but other studies used different neuropsychological tests, functional measures, or clinical diagnoses. This heterogeneity precluded formal meta-analysis and necessitated a structured narrative synthesis. We did not formally assess the certainty of evidence (eg, using Grading of Recommendations Assessment, Development, and Evaluation), given the methodological heterogeneity of included studies and the predominance of observational designs.

The evidence base is also constrained by small sample sizes and short monitoring periods. In total, 8.2% (4/49) of studies enrolled 30 participants or fewer, while an additional 24.5% (12/49) included between 30 and 100 participants, indicating that more than one-third of the evidence relies on small cohorts. Furthermore, 67.3% (33/49) of studies monitored participants for only 1 week or less, limiting the ability to capture long-term variability in activity, sleep, or circadian rhythms. Only 18.4% (9/49) of studies systematically assessed real-world feasibility, including adherence, comfort, or long-term usability. Moreover, external validation was performed in just 6.1% (3/49) of studies, restricting the generalizability of reported associations or predictive models beyond the original cohorts. Thus, a key limitation of the current evidence is the risk of overinterpretation. Most studies were cross-sectional or small cohorts, and only few have undergone external validation [29,39,63]. In several studies, high accuracy or AUROC values were reported in the context of small or imbalanced datasets, conditions that increase the risk of overfitting and optimism bias. As a result, these performance estimates should be interpreted cautiously in the absence of external validation or evaluation in independent cohorts. While models often report high accuracy, these findings may not generalize. Preventive claims remain speculative until tested in large prospective cohorts or intervention trials. In addition, research-grade devices typically provide access to raw, high-resolution data and validated measurement pipelines, whereas consumer devices often rely on proprietary algorithms and offer limited transparency, which may affect comparability across studies.

The risk of bias was present to varying degrees. Although most studies were rated as low risk (34/49, 69.4%), nearly one-third (13/49, 26.5%) were classified as medium risk, and 4.1% (2/49) as high risk. The high-risk studies exemplify common problems such as selective sampling, inadequate reference standards, and small or restricted populations, all of which likely led to overestimation of diagnostic performance and limited the generalizability of their findings. Publication bias is also possible, as studies reporting positive or novel results are more likely to be published. Together, these issues highlight that the field remains fragmented, with methodological and reporting inconsistencies that restrict the strength and reproducibility of current evidence.

Limitations of the review process itself must also be acknowledged. The restriction to English-language publications and peer-reviewed journals may have introduced language or

XSL•FO
RenderX

publication bias, potentially excluding relevant gray literature or local studies. Additionally, while we searched for 4 major interdisciplinary databases, the exclusion of specialized clinical indices (eg, PsycINFO or Embase) could have resulted in the omission of some relevant records.

Implications for Future Research and Practice

This review highlights several priorities for future research and practice. A key challenge is bridging the gap between research-grade actigraphy and consumer wearables. While actigraphy remains the most validated tool for assessing sleep, activity, and circadian rhythms, limited use of consumer devices constrains scalability [79]. Evidence suggests that some multisensor consumer wearables can achieve accuracy comparable to research-grade actigraphs under controlled conditions [80,81], but direct comparisons across diverse populations and real-world settings remain needed. Where sufficient accuracy and reliability are demonstrated, consumer wearables could support large-scale, longitudinal monitoring for research and preventive care [82].

Another priority is the expanded use of advanced analytics and digital phenotyping. Longitudinal wearable data enable modeling of dynamic behavioral and physiological changes, supporting early and personalized prevention. However, machine learning and deep learning methods remain underused, particularly sequence-based models such as recurrent neural networks, temporal convolutional models, and transformer-based approaches (eg, Informer and TimesFM) that are well suited for time-series health data [83]. In addition, large language model–based frameworks (eg, HealthLLM) offer opportunities to integrate wearable-derived time series with clinical and contextual information [84]. Progress in this area will depend on access to larger, well-annotated datasets and multicenter collaboration, as well as the incorporation of explainable AI methods to ensure interpretability and clinical trust [85].

Finally, more prevention-oriented research is needed. Only 2 randomized controlled trials tested wearable-assisted interventions, and a small subset of observational studies examined prevention-relevant behaviors. Yet prevention is central to dementia care given the absence of disease-modifying treatments. Wearables could play a dual role, both by detecting early signs of decline and by monitoring modifiable lifestyle factors such as sleep, circadian rhythms, and physical activity [86,87]. Positioning wearables within preventive frameworks could therefore be among the most clinically meaningful directions for future research.

From a translational perspective, ethical and regulatory considerations will be critical for the clinical adoption of AI-based dementia risk prediction. Issues such as data privacy, transparency of algorithms, potential psychological impact of early risk labeling, and compliance with regulatory frameworks must be carefully addressed. At present, research-grade wearable devices and analytically transparent models appear more suitable for clinical research settings, whereas many consumer-grade devices and advanced AI approaches remain primarily experimental. Further validation, standardization, and regulatory oversight will be necessary before these methods can be routinely integrated into clinical practice.

Overall, this review advances the field by framing the transition from descriptive statistics to a predictive digital phenotyping framework. Distinct from prior reviews limited to established dementia, it synthesizes evidence specifically for the preclinical window, distinguishing findings with direct predictive utility from those offering only indirect correlational insights. It contributes a critical assessment of methodological maturity, identifying the heavy reliance on research-grade devices and the lack of external validation as primary barriers to implementation. Ultimately, these findings suggest that shifting to continuous, passive monitoring offers a scalable method to detect subtle behavioral deviations, creating opportunities for earlier, personalized risk reduction strategies [88].

Acknowledgments

The authors thank the Vicomtech Foundation for supporting this research. The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing, reformatting, and quality assessment. The GAI tool used was: ChatGPT-5.1 (OpenAI). Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by: “collective responsibility.”

Funding

This research was funded by the Vicomtech Foundation. The funder had no role in the study design, data collection, analysis, interpretation, or writing of the manuscript.

Data Availability

No new primary data were generated in this systematic review. Data extracted from the included studies and the curated list of screened articles are available from the corresponding author upon reasonable request. The Python code used to support automated screening and study management is not publicly available, as it was developed for internal use; however, it can be shared upon reasonable request for purposes of methodological transparency.

XSL•FO
RenderX

Authors' Contributions

Conceptualization was carried out by AC. Methodology was developed by AC and MA, and software was prepared by AC. Validation was conducted by CM and AA. Data curation was performed by AC and MA. The original draft of the manuscript was written by AC and MA, and review and editing were undertaken by AC, MA, CM, and AA. Supervision was provided by CM and AA, while project administration was managed by CM. Funding acquisition was secured by CM. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA 2020 checklist.
[DOCX File , 40 KB - jmir_v28i1e86262_app1.docx]

Multimedia Appendix 2
PRISMA-S checklist.
[DOCX File , 17 KB - jmir_v28i1e86262_app2.docx]

Multimedia Appendix 3
Exact search queries used across all databases to identify records included in this review.
[PDF File (Adobe PDF File), 53 KB - jmir_v28i1e86262_app3.pdf]

Multimedia Appendix 4
Extended study characteristics and principal findings of all the studies included in this systematic review.
[XLSX File (Microsoft Excel File), 16 KB - jmir_v28i1e86262_app4.xlsx]

Multimedia Appendix 5
Risk of bias assessments of included studies using validated tools appropriate to study design.
[PDF File (Adobe PDF File), 276 KB - jmir_v28i1e86262_app5.pdf]

Multimedia Appendix 6
List and definitions of sleep, circadian rhythm, physical activity, and physiological variables extracted from wearable devices across included studies.
[PDF File (Adobe PDF File), 111 KB - jmir_v28i1e86262_app6.pdf]

References

1. Dementia. World Health Organization. 2025. URL: https://www.who.int [accessed 2025-08-30]
2. Chowdhary N, Barbui C, Anstey KJ, Kivipelto M, Barbera M, Peters R. Reducing the risk of cognitive decline and dementia: WHO recommendations. Front Neurol 2021;12:765584 [FREE Full text] [doi: 10.3389/fneur.2021.765584] [Medline: 35082745]
3. Risk Reduction of Cognitive Decline and Dementia: WHO guidelines. Geneva: World Health Organization; 2019:96.
4. Chong TWH, Macpherson H, Schaumberg MA, Brown BM, Naismith SL, Steiner GZ. Dementia prevention: the time to act is now. Med J Aust 2021;214(7):302-304. [doi: 10.5694/mja2.50972] [Medline: 33715184]
5. Jones A, Ali MU, Kenny M, Mayhew A, Mokashi V, He H. Potentially modifiable risk factors for dementia and mild cognitive impairment: an umbrella review and meta-analysis. Dement Geriatr Cogn Disord 2024;53(2):91-106 [FREE Full text] [doi: 10.1159/000536643] [Medline: 38346414]
6. Hardcastle C, Taylor B, Price C. Global Deterioration Scale. In: Encyclopedia of Gerontology and Population Aging. Cham: Springer; 2022:2198-2201.
7. Baron-Shahaf D, Shahaf G. Markers of too little effort or too much alertness during neuropsychological assessment: demonstration with perioperative changes. Brain Behav 2024;14(8):e3649 [FREE Full text] [doi: 10.1002/brb3.3649] [Medline: 39169455]
8. Holding BC, Ingre M, Petrovic P, Sundelin T, Axelsson J. Quantifying cognitive impairment after sleep deprivation at different times of day: a proof of concept using ultra-short smartphone-based tests. Front Behav Neurosci 2021;15:666146. [doi: 10.3389/fnbeh.2021.666146] [Medline: 33927603]
9. Salis F, Costaggiu D, Mandas A. Mini-mental state examination: optimal cut-off levels for mild and severe cognitive impairment. Geriatrics (Basel) 2023;8(1):12 [FREE Full text] [doi: 10.3390/geriatrics8010012] [Medline: 36648917]

XSL•FO
RenderX

10. Arevalo-Rodriguez I, Smailagic N, Roqué-Figuls M, Ciapponi A, Sanchez-Perez E, Giannakou A. Mini-mental state examination (MMSE) for the early detection of dementia in people with mild cognitive impairment (MCI). Cochrane Database Syst Rev 2021;7(7):CD010783. [doi: 10.1002/14651858.CD010783.pub3] [Medline: 34313331]
11. Chithiramohan T, Santhosh S, Threlfall G, Hull L, Mukaetova-Ladinska EB, Subramaniam H. Culture-fair cognitive screening tools for assessment of cognitive impairment: a systematic review. J Alzheimers Dis Rep 2024;8(1):289-306. [doi: 10.3233/adr-230194]
12. Sakal C, Li T, Li J, Li X. Predicting poor performance on cognitive tests among older adults using wearable device data and machine learning: a feasibility study. NPJ Aging 2024;10(1):56 [FREE Full text] [doi: 10.1038/s41514-024-00177-x] [Medline: 39587119]
13. Soumma SB, Mamun A, Ghasemzadeh H. AI-powered wearable sensors for health monitoring and clinical decision making. Curr Opin Biomed Eng 2025;36:100628. [doi: 10.1016/j.cobme.2025.100628]
14. Keshet A, Reicher L, Bar N, Segal E. Wearable and digital devices to monitor and treat metabolic diseases. Nature metabolism 2023;5(4):563-571. [doi: 10.1038/s42255-023-00778-y]
15. Zaman Q, others. The role of artificial intelligence in early disease detection: transforming diagnostics and treatment. Multidiscip J Healthc (MJH) 2024:1-54 [FREE Full text]
16. Cote AC, Phelps RJ, Kabiri NS, Bhangu JS, Thomas. Evaluation of wearable technology in dementia: a systematic review and meta-analysis. Front Med (Lausanne) 2021;7:501104 [FREE Full text] [doi: 10.3389/fmed.2020.501104] [Medline: 33505979]
17. Weizman Y, Tirosh O, Beh J, Fuss FK, Pedell S. Gait assessment using wearable sensor-based devices in people living with dementia: a systematic review. Int J Environ Res Public Health 2021;18(23):12735. [doi: 10.3390/ijerph182312735] [Medline: 34886459]
18. Iso-Markku P, Aaltonen S, Kujala UM, Halme H, Phipps D, Knittle K. Physical activity and cognitive decline among older adults: a systematic review and meta-analysis. JAMA Netw Open 2024;7(2):e2354285. [doi: 10.1001/jamanetworkopen.2023.54285] [Medline: 38300618]
19. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
20. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
21. Ma LL, Wang YY, Yang ZH, Huang D, Weng H, Zeng X. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? Mil Med Res 2020;7(1):7. [doi: 10.1186/s40779-020-00238-8] [Medline: 32111253]
22. Lentzen M, Vairavan S, Muurling M, Alepopoulos V, Atreya A, Boada M. RADAR-AD: assessment of multiple remote monitoring technologies for early detection of Alzheimer's disease. Alzheimers Res Ther 2025;17(1):29. [doi: 10.1186/s13195-025-01675-0] [Medline: 39865315]
23. Jones CD, Wasilko R, Zhang G, Stone KL, Gujral S, Rodakowski J. Detecting sleep/wake rhythm disruption related to cognition in older adults with and without mild cognitive impairment using the myRhythmWatch platform: feasibility and correlation study. JMIR Aging 2025;8:e67294-e67294. [doi: 10.2196/67294] [Medline: 40194310]
24. Satomi E, Apolinário D, Magaldi RM, Busse AL, Vieira Gomes GC, Ribeiro E. Beyond sleep: rest and activity rhythm as a marker of preclinical and mild dementia in older adults with less education. Neurobiol Sleep Circadian Rhythms 2025;18:100110 [FREE Full text] [doi: 10.1016/j.nbscr.2024.100110] [Medline: 39834590]
25. Sauers SC, Toedebusch CD, Richardson R, Spira AP, Morris JC, Holtzman DM. Midpoint of sleep is associated with sleep quality in older adults with and without symptomatic Alzheimer's disease. Sleep Adv 2024;5(1):zpae023. [doi: 10.1093/sleepadvances/zpae023] [Medline: 38711547]
26. Baril AA, Picard C, Labonté A, Sanchez E, Duclos C, Mohammediyan B. Day-to-day sleep variability with Alzheimer's biomarkers in at-risk elderly. Alzheimers Dement Diagn Assess Dis Monit 2024;16(1):e12521 [FREE Full text] [doi: 10.1002/dad2.12521] [Medline: 38371359]
27. Seo K, Takayanagi N, Sudo M, Yamashiro Y, Chiba I, Makino K. Association between daily gait speed patterns and cognitive impairment in community-dwelling older adults. Sci Rep 2023;13(1):2783. [doi: 10.1038/s41598-023-29805-4] [Medline: 36797381]
28. Huang B, Liao G, Lee PMY, Chan CK, Tai L, Tsang CYJ. Association of circadian rhythm with mild cognitive impairment among male pneumoconiosis workers in Hong Kong: a cross-sectional study. Sci Rep 2023;13(1):1650. [doi: 10.1038/s41598-023-28832-5] [Medline: 36717700]
29. Palmer JR, Wang C, Kong D, Cespedes M, Pye J, Hickie IB. Rest-activity rhythms and tract specific white matter lesions in older adults at risk for cognitive decline. Mol Psychiatry 2022;27(8):3410-3416 [FREE Full text] [doi: 10.1038/s41380-022-01641-4] [Medline: 35764707]
30. Lim J. A smart healthcare-based system for classification of dementia using deep learning. Digit Health 2022;8:20552076221131667 [FREE Full text] [doi: 10.1177/20552076221131667] [Medline: 36312848]

XSL•FO
RenderX

31. Boa Sorte Silva NC, Falck RS, Chan PCY, Tai D, Backhouse D, Stein R. The association of sleep and cortical thickness in mild cognitive impairment. Experimental Gerontology 2022;167:111923. [doi: 10.1016/j.exger.2022.111923]
32. Kim M, Liotta EM, Maas MB, Braun RI, Garcia-Canga B, Ganger DR. Rest-activity rhythm disturbance in liver cirrhosis and association with cognitive impairment. Sleep 2021;44(6):zsaa288. [doi: 10.1093/sleep/zsaa288] [Medline: 33367862]
33. Basta M, Zaganas I, Simos P, Koutentaki E, Dimovasili C, Mathioudakis L. Apolipoprotein E ε4 allele is associated with long sleep duration among elderly with cognitive impairment. J Alzheimers Dis 2021;79(2):763-771. [doi: 10.3233/jad-200958]
34. Roh HW, Choi J, Kim NR, Choe YS, Choi JW, Cho SM. Associations of rest-activity patterns with amyloid burden, medial temporal lobe atrophy, and cognitive impairment. EBioMedicine 2020;58:102881. [doi: 10.1016/j.ebiom.2020.102881] [Medline: 32736306]
35. Liu Z, Zhang L, Wu J, Zheng Z, Gao J, Lin Y. Machine learning-based classification of circadian rhythm characteristics for mild cognitive impairment in the elderly. Front Public Health 2022;10:1036886. [doi: 10.3389/fpubh.2022.1036886] [Medline: 36388285]
36. Espinosa N, Hoyos CM, McKinnon AC, Almgren H, Duffy SL, Naismith SL. Rest-activity rhythm fragmentation and synchronization are linked with reduced cortical thickness in older adults "at risk" for dementia. Sleep 2025;48(5):zsaf017. [doi: 10.1093/sleep/zsaf017] [Medline: 40052961]
37. Manning CA, Youngkin A, Quigg M. Sleepiness, sleep time, and depression of caregivers are linked with sleep and behaviors of their paired partners with dementia. J Alzheimers Dis Rep 2025;9:25424823241300981 [FREE Full text] [doi: 10.1177/25424823241300981] [Medline: 40034516]
38. Ghosal R, Varma VR, Volfson D, Urbanek J, Hausdorff JM, Watts A. Scalar on time-by-distribution regression and its application for modelling associations between daily-living physical activity and cognitive functions in Alzheimer's Disease. Sci Rep 2022;12(1):11558. [doi: 10.1038/s41598-022-15528-5] [Medline: 35798763]
39. Park B, Choi BJ, Lee H, Jang JH, Roh HW, Kim EY. Modeling brain volume using deep learning-based physical activity features in patients with dementia. Front Neuroinform 2022;16:795171 [FREE Full text] [doi: 10.3389/fninf.2022.795171] [Medline: 35356447]
40. Palmer JR, Duffy SL, Meares S, Pye J, Calamante F, Cespedes M. Rest-activity functioning is related to white matter microarchitecture and modifiable risk factors in older adults at-risk for dementia. Sleep 2021;44(7):zsab007. [doi: 10.1093/sleep/zsab007] [Medline: 33428761]
41. Wei J, Boger J. Sleep detection for younger adults, healthy older adults, and older adults living with dementia using wrist temperature and actigraphy: prototype testing and case study analysis. JMIR Mhealth Uhealth 2021;9(6):e26462 [FREE Full text] [doi: 10.2196/26462] [Medline: 34061038]
42. Corbi A, Burgos D. Connection between sleeping patterns and cognitive deterioration in women with alzheimer's disease. Sleep Breath 2022;26(1):361-371. [doi: 10.1007/s11325-021-02327-x] [Medline: 33792886]
43. Baril A, Picard C, Labonté A, Sanchez E, Duclos C, Mohammediyan B. Longer sleep duration and neuroinflammation in at-risk elderly with a parental history of Alzheimer's disease. Sleep 2024;47(6):zsae081 [FREE Full text] [doi: 10.1093/sleep/zsae081] [Medline: 38526098]
44. Sun H, Li P, Gao L, Yang J, Yu L, Buchman AS. Altered motor activity patterns within 10-minute timescale predict incident clinical Alzheimer's disease. J Alzheimers Dis 2024;98(1):209-220 [FREE Full text] [doi: 10.3233/JAD-230928] [Medline: 38393904]
45. Winer JR, Lok R, Weed L, He Z, Poston KL, Mormino EC. Impaired 24-h activity patterns are associated with an increased risk of Alzheimer's disease, Parkinson's disease, and cognitive decline. Alzheimers Res Ther 2024;16(1):35 [FREE Full text] [doi: 10.1186/s13195-024-01411-0] [Medline: 38355598]
46. Haghayegh S, Gao C, Sugg E, Zheng X, Yang HW, Saxena R. Association of rest-activity rhythm and risk of developing dementia or mild cognitive impairment in the middle-aged and older population: prospective cohort study. JMIR Public Health Surveill 2024;10:e55211 [FREE Full text] [doi: 10.2196/55211] [Medline: 38713911]
47. Basta M, Bouloukaki I, Skourti E, Zampetakis A, Alexopoulou C, Ganiaris A. Long objective sleep duration is a marker of cognitive impairment in older adults: findings from the Cretan Aging Cohort. J Alzheimers Dis Rep 2024;8(1):927-934 [FREE Full text] [doi: 10.3233/ADR-230203] [Medline: 38910938]
48. Milton S, Cavaillès C, Ancoli-Israel S, Stone KL, Yaffe K, Leng Y. Five-year changes in 24-hour sleep-wake activity and dementia risk in oldest old women. Neurology 2025;104(8):e213403. [doi: 10.1212/WNL.0000000000213403] [Medline: 40106755]
49. Skourti E, Simos P, Zampetakis A, Koutentaki E, Zaganas I, Alexopoulou C. Long-term associations between objective sleep quality and quantity and verbal memory performance in normal cognition and mild cognitive impairment. Front Neurosci 2023;17:1265016. [doi: 10.3389/fnins.2023.1265016] [Medline: 37928739]
50. Gao L, Li P, Gaykova N, Zheng X, Gao C, Lane JM. Circadian rest-activity rhythms, delirium risk, and progression to dementia. Ann Neurol 2023;93(6):1145-1157 [FREE Full text] [doi: 10.1002/ana.26617] [Medline: 36808743]
51. Jeon SY, Byun MS, Yi D, Jung G, Lee J, Kim YK. Circadian rest-activity rhythm and longitudinal brain changes underlying late-life cognitive decline. Psychiatry Clin Neurosci 2023;77(4):205-212 [FREE Full text] [doi: 10.1111/pcn.13521] [Medline: 36527292]

XSL•FO
RenderX

52. Yi Lee PM, Ling Kwok BH, Ting Ma JY, Tse LA. A population-based prospective study on rest-activity rhythm and mild cognitive impairment among Hong Kong healthy community-dwelling older adults. Neurobiol Sleep Circadian Rhythms 2021;10:100065 [FREE Full text] [doi: 10.1016/j.nbscr.2021.100065] [Medline: 33997474]
53. Lysen TS, Luik AI, Ikram MK, Tiemeier H, Ikram MA. Actigraphy-estimated sleep and 24-hour activity rhythms and the risk of dementia. Alzheimers Dement 2020;16(9):1259-1267 [FREE Full text] [doi: 10.1002/alz.12122] [Medline: 32558256]
54. Agudelo C, Tarraf W, Wu B, Wallace DM, Patel SR, Redline S. Actigraphic sleep patterns and cognitive decline in the hispanic community health study/study of latinos. Alzheimers Dement 2021;17(6):959-968. [doi: 10.1002/alz.12250] [Medline: 33350583]
55. Ning Y, Chen M, Yang H, Jia J. Accelerometer-derived "weekend warrior" physical activity, sedentary behavior, and risk of dementia. Alzheimers Res Ther 2025;17(1):67 [FREE Full text] [doi: 10.1186/s13195-024-01657-8] [Medline: 40121491]
56. Ning Y, Chen M, Yang H, Jia J. Accelerometer-measured physical activity timing with incident dementia. Alzheimers Dement 2025;21(2):e14452. [doi: 10.1002/alz.14452] [Medline: 39868637]
57. Zhao B, Zhou S, Chang J, Li A, Geng C, Wei T. Association between weekend recovery sleep and risk of incident dementia: a prospective cohort study in the UK Biobank. J Neurol 2025;272(9):612. [doi: 10.1007/s00415-025-13363-y] [Medline: 40911069]
58. Lu Z, Leung JCS, Feng H, Zhang J, Wing YK, Kwok TCY. Circadian rest-activity rhythms and cognitive decline and impairment in older Chinese adults: a multicohort study with prospective follow-up. Arch Gerontol Geriatr 2023;116:105215. [doi: 10.1016/j.archger.2023.105215] [Medline: 39491073]
59. Chan LLY, Espinoza Cerda MT, Brodie MA, Lord SR, Taylor ME. Daily-life walking speed, running duration and bedtime from wrist-worn sensors predict incident dementia: a watch walk - UK biobank study. Int Psychogeriatr 2025;37(3):100031 [FREE Full text] [doi: 10.1016/j.inpsyc.2024.100031] [Medline: 39827009]
60. Shi C, Babiker N, Urbanek JK, Grossman RL, Huisingh-Scheetz M, Rzhetsky A. Free-living wrist and hip accelerometry forecast cognitive decline among older adults without dementia over 1- or 5-years in two distinct observational cohorts. NPJ Aging 2022;8(1):7. [doi: 10.1038/s41514-022-00087-w] [Medline: 35927250]
61. Xiao Q, Shadyab AH, Rapp SR, Stone KL, Yaffe K, Sampson JN. Rest-activity rhythms and cognitive impairment and dementia in older women: results from the women's health initiative. J Am Geriatr Soc 2022;70(10):2925-2937 [FREE Full text] [doi: 10.1111/jgs.17926] [Medline: 35708069]
62. Hoepel SJW, Ikram MK, Ikram MA, Voortman T, Luik AI. Substitutions of sleep, physical activity, and sedentary behaviour and the risk of dementia and stroke. J Sleep Res 2026;35(1):e70166. [doi: 10.1111/jsr.70166] [Medline: 40793962]
63. Cho E, Kim S, Heo S, Shin J, Hwang S, Kwon E. Machine learning-based predictive models for the occurrence of behavioral and psychological symptoms of dementia: model development and validation. Sci Rep 2023;13(1):8073 [FREE Full text] [doi: 10.1038/s41598-023-35194-5] [Medline: 37202454]
64. Plotogea OM, Diaconu CC, Gheorghe G, Stan-Ilie M, Badea M, Prelipcean CC. The prevalence and association of cognitive impairment with sleep disturbances in patients with chronic liver disease. Brain Sci 2022;12(4):444. [doi: 10.3390/brainsci12040444] [Medline: 35447976]
65. Targa ADS, Benítez ID, Dakterzada F, Fontenele-Araujo J, Minguez O, Zetterberg H. The circadian rest-activity pattern predicts cognitive decline among mild-moderate alzheimer's disease patients. Alzheimers Res Ther 2021;13(1):161 [FREE Full text] [doi: 10.1186/s13195-021-00903-7] [Medline: 34563258]
66. Cho E, Kim S, Hwang S, Kwon E, Heo S, Lee JH, et al. Factors associated with behavioral and psychological symptoms of dementia: prospective observational study using actigraphy. J Med Internet Res 2021;23(10):e29001. [doi: 10.2196/29001] [Medline: 34714244]
67. Rykov YG, Patterson MD, Gangwar BA, Jabar SB, Leonardo J, Ng KP. Predicting cognitive scores from wearable-based digital physiological features using machine learning: data from a clinical trial in mild cognitive impairment. BMC Med 2024;22(1):36 [FREE Full text] [doi: 10.1186/s12916-024-03252-y] [Medline: 38273340]
68. David S, Hohenfeld C, Haberl L, Pahl J, Costa AS, Kilders A. Physical activity monitoring in Alzheimer's disease during sport interventions: a multi-methodological perspective. Front Neurol 2023;14:1195694 [FREE Full text] [doi: 10.3389/fneur.2023.1195694] [Medline: 37808485]
69. Khosroazad S, Gilbert CF, Aronis JB, Daigle KM, Esfahani M, Almaghasilah A. Sleep movements and respiratory coupling as a biobehavioral metric for early Alzheimer's disease in independently dwelling adults. BMC Geriatr 2023;23(1):252 [FREE Full text] [doi: 10.1186/s12877-023-03983-2] [Medline: 37106470]
70. Hossain M, Uddin M, Thapa K, Al Zubaer MA, Islam M, Lee J. Detecting cognitive impairment status using keystroke patterns and physical activity data among the older adults: a machine learning approach. J Healthc Eng 2021;2021:1302989. [doi: 10.1155/2021/1302989] [Medline: 34966518]
71. Kallogjeri D, Spitznagel EL, Piccirillo JF. Importance of defining and interpreting a clinically meaningful difference in clinical research. JAMA Otolaryngol Head Neck Surg 2020;146(2):101-102. [doi: 10.1001/jamaoto.2019.3744] [Medline: 31804662]
72. Rykov YG, Patterson MD, Gangwar BA, Jabar SB, Leonardo J, Ng KP. Predicting cognitive scores from wearable-based digital physiological features using machine learning: data from a clinical trial in mild cognitive impairment. BMC Med 2024;22(1):36 [FREE Full text] [doi: 10.1186/s12916-024-03252-y] [Medline: 38273340]

XSL•FO
RenderX

73. Berger A, Wielgus KK, Young-McCaughan S, Fischer P, Farr L, Lee KA. Methodological challenges when using actigraphy in research. J Pain Symptom Manage 2008;36(2):191-199 [FREE Full text] [doi: 10.1016/j.jpainsymman.2007.10.008] [Medline: 18400460]
74. Patterson MR, Nunes AAS, Gerstel D, Pilkar R, Guthrie T, Neishabouri A. 40 years of actigraphy in sleep medicine and current state of the art algorithms. NPJ Digit Med 2023;6(1):51 [FREE Full text] [doi: 10.1038/s41746-023-00802-1] [Medline: 36964203]
75. Morris AC, Seker A, Telesia L, Wickersham A, Ching BC, Roy R. Adherence to actigraphic devices in elementary school-aged children: systematic review and meta-analysis. J Med Internet Res 2025;27:e79718 [FREE Full text] [doi: 10.2196/79718] [Medline: 41183377]
76. Dobson R, Stowell M, Warren J, Tane T, Ni L, Gu Y. Use of consumer wearables in health research: issues and considerations. J Med Internet Res 2023;25:e52444 [FREE Full text] [doi: 10.2196/52444] [Medline: 37988147]
77. D'Ignazio G, Carlucci L, Sergi MR, Palumbo R, Dattilo L, Terrei M. Is the MMSE enough for MCI? A narrative review of the usefulness of the MMSE. Front Psychol 2025;16:1727738 [FREE Full text] [doi: 10.3389/fpsyg.2025.1727738] [Medline: 41356018]
78. Gu B, Kim HS, Kim H, Yoo J. Advancements in wearable sensor technologies for health monitoring in terms of clinical applications, rehabilitation, and disease risk assessment: systematic review. JMIR Mhealth Uhealth 2026;14:e76084. [doi: 10.2196/76084] [Medline: 41511829]
79. Kainec KA, Caccavaro J, Barnes M, Hoff C, Berlin A, Spencer RMC. Evaluating accuracy in five commercial sleep-tracking devices compared to research-grade actigraphy and polysomnography. Sensors (Basel) 2024;24(2):635 [FREE Full text] [doi: 10.3390/s24020635] [Medline: 38276327]
80. Smith MT, McCrae CS, Cheung J, Martin JL, Harrod CG, Heald JL, et al. Use of actigraphy for the evaluation of sleep disorders and circadian rhythm sleep-wake disorders: an American academy of sleep medicine systematic review, meta-analysis, and GRADE assessment. J Clin Sleep Med 2018;14(7):1209-1230. [doi: 10.5664/jcsm.7228] [Medline: 29991438]
81. Kahawage P, Jumabhoy R, Hamill K, de Zambotti M, Drummond SPA. Validity, potential clinical utility, and comparison of consumer and research-grade activity trackers in Insomnia Disorder I: in-lab validation against polysomnography. J Sleep Res 2020;29(1):e12931. [doi: 10.1111/jsr.12931] [Medline: 31626361]
82. Huhn S, Axt M, Gunga HC, Maggioni MA, Munga S, Obor D. The impact of wearable technologies in health research: scoping review. JMIR Mhealth Uhealth 2022;10(1):e34384 [FREE Full text] [doi: 10.2196/34384] [Medline: 35076409]
83. Ahmed S, Nielsen IE, Tripathi A, Siddiqui S, Ramachandran RP, Rasool G. Transformers in time-series analysis: a tutorial. Circuits Syst Signal Process 2023;42(12):7433-7466. [doi: 10.1007/s00034-023-02454-8]
84. Kim Y, Xu X, McDuff D, Breazeal C, Park HW. Healthllm: large language models for health prediction via wearable sensor data. ArXiv Preprint posted online on January 12, 2024. [doi: 10.48550/arXiv.2401.06866]
85. Saraswat D, Bhattacharya P, Verma A, Prasad VK, Tanwar S, Sharma G. Explainable AI for healthcare 5.0: opportunities and challenges. IEEE Access 2022;10:84486-84517. [doi: 10.1109/access.2022.3197671]
86. Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. Lancet 2020;396(10248):413-446 [FREE Full text] [doi: 10.1016/S0140-6736(20)30367-6] [Medline: 32738937]
87. Kivipelto M, Mangialasche F, Ngandu T. Lifestyle interventions to prevent cognitive impairment, dementia and Alzheimer disease. Nat Rev Neurol 2018;14(11):653-666. [doi: 10.1038/s41582-018-0070-3] [Medline: 30291317]
88. Bufano P, Laurino M, Said S, Tognetti A, Menicucci D. Digital phenotyping for monitoring mental disorders: systematic review. J Med Internet Res 2023;25:e46778 [FREE Full text] [doi: 10.2196/46778] [Medline: 38090800]

Abbreviations

AI: artificial intelligence
ICD-10: International Statistical Classification of Diseases, Tenth Revision
MCI: mild cognitive impairment
MeSH: Medical Subject Headings
MMSE: Mini-Mental State Examination
MoCA: Montreal Cognitive Assessment
OR: odds ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
XGBoost: Extreme Gradient Boosting

XSL•FO
RenderX

Edited by S Brini; submitted 22.Oct.2025; peer-reviewed by B Howell, S Ajayi, S Ye; comments to author 27.Nov.2025; revised version received 23.Jan.2026; accepted 26.Jan.2026; published 23.Feb.2026.

Please cite as:
Cejudo A, Arrojo M, Martín C, Almeida A
AI and Wearables for Early Detection of Cognitive Impairment and Dementia: Systematic Review
J Med Internet Res 2026;28:e86262
URL: https://www.jmir.org/2026/1/e86262
doi:10.2196/86262
PMID:

©Ander Cejudo, Markel Arrojo, Cristina Martín, Aitor Almeida. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Design Practices for Data Dashboards in Health Care: Scoping Review

Heike Vornhagen[1*], MA, PhD; Stephen Barrett[2*], BSc, MSc; Ciara Carroll[2], BCh, BAO, MB; Lydia Kavochi Iladiva[3], MBBCHIR, MPH, MBA; Gregory Martin[2], MBBCHIR, MBA, MPH; Declan McKeown[2], MB, BCh, BAO, MPH; Jennifer Martin[2], MB, BCh, BAO, MPH

[1]Insight Research Ireland Centre for Data Analytics, Data Science Institute, Ollscoil na Gaillimhe – University of Galway, Galway, Ireland

[2]Health Service Executive, Dublin, Ireland

[3]Department of Epidemiology & Public Health, University College Cork, Cork, Ireland

[*]these authors contributed equally

Corresponding Author:
Heike Vornhagen, MA, PhD
Insight Research Ireland Centre for Data Analytics
Data Science Institute
Ollscoil na Gaillimhe – University of Galway
IDA Business Park
Lower Dangan
Galway, H91 AEX4
Ireland
Phone: 353 91 495053
Email: heike.vornhagen@universityofgalway.ie

Related Article:

Comment in: https://www.jmir.org/2026/1/e98272

Abstract

Background: Health care dashboards have the potential to enhance understanding, decision-making, and communication. However, their design, implementation, and evaluation are often hindered by the absence of standardized guidelines. This scoping review synthesizes international evidence to identify common practices for health care dashboard design, providing a foundation for application in the Irish context.

Objective: This study aimed to identify existing guidelines and common practices for health care dashboard design to inform future development and implementation within the Irish health care system.

Methods: A scoping review using an evidence summary approach was conducted. PubMed, Embase, Scopus, and IEEE Xplore (2014-2024) were searched. Practices were extracted and analyzed using reflexive thematic analysis and then grouped into 4 main pillars: approach (engagement of end users and stakeholders), content (data quality, effective insights, and presentation), behavior (usability and accessibility), and adoption (sustainability).

Results: From 1644 initially identified studies, 18 (1.1%) met the inclusion criteria. Most were hospital focused (13/18, 72.2%), with few community- or public-facing dashboards. Only 4 of 18 (22.2%) studies described structured guidelines; most implementations (14/18, 77.8%) were ad hoc. Common practices included user involvement, actionable metrics, data quality, usability, and workflow integration. Divergences were observed: hospitals prioritized clinical indicators, public dashboards emphasized transparency, and community dashboards were underrepresented. Conflicting findings included debate over interactivity vs static simplicity.

Conclusions: Dashboard design remains fragmented, with limited guidance for structured design or implementation. The 4 pillars provide a practical synthesis of best practices, highlighting pathways for evidence-informed, user-centered design. These pillars will inform future consensus building and co-design of health care dashboards in Ireland and can serve as a foundation for broader application in primary care, community, and public health settings.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e77361) doi:10.2196/77361

KEYWORDS

data visualization; dashboard design; health care analytics; evidence-based design; health informatics

Introduction

This review explores how dashboards are developed within the health care sector and which guidelines, if any, steer such development. Drawing insights from international research, this review ultimately aims to identify best practices and demonstrate how visual tools can enhance understanding, decision-making, and communication.

Dashboards are electronic displays that usually combine multiple visualizations to enable users to simultaneously access and interact with information on related topics [1]. Ideally, dashboards should support rapid understanding and reduce cognitive load, a key consideration in high-stress health care environments where time-sensitive decisions are paramount [2]. As visual dashboards can significantly improve decision-making [3], they should enable health care professionals to efficiently and effectively identify care challenges and progress. Health organizations use dashboards for different purposes, such as management and oversight, as well as for monitoring and improving direct patient care. Dashboards can also help improve teamwork and communication among people with different roles, expertise, and schedules [1].

Data visualization refers to the process of converting complex datasets into visual formats, such as charts and graphs. By facilitating clearer and more immediate insights, effective data visualizations have the potential to significantly improve information sharing among health care professionals, leading to enhanced collaborative decision-making [4]. This is supported by Lurie and Mason [5], who highlight that appropriately designed data visualizations not only improve comprehension but also enhance retention of critical information. Similarly, international studies have found that well-designed data visualization tools can reduce medical errors and support more accurate clinical decisions, particularly when used to communicate complex patient data [6].

Several frameworks for data visualization and dashboard design exist outside of health care and provide useful starting points. The nested model by Munzner [7] describes levels of design decision-making, from domain problems to visual encoding, emphasizing alignment between user tasks and visualization choices. The dashboard design principles by Few [8] offer practical guidance on clarity, simplicity, and information density, primarily targeted at business contexts. Human-computer interaction (HCI) standards, such as ISO 9241-210 [9], similarly emphasize usability, accessibility, and iterative user-centered design. These frameworks are valuable, but their application to health care contexts is limited. They often neglect the complexities of health care data, such as interoperability, sensitive patient information, and the unique cognitive demands of clinical environments.

Previous reviews of health care dashboards have advanced the field but remain constrained in several ways. For example, a scoping review by Helminski et al [10] found that few included studies used usability testing or formal theoretical frameworks for design or evaluation. The systematic review of hospital dashboards by Rabiei and Almasi [11] identified both functional and nonfunctional requirements but reported that many studies lacked empirical evaluation and had limited detail on implementation or integration with existing systems. As a result, existing evidence has not yet been consolidated into a structured, theoretically grounded framework that can guide the design of health care dashboards across settings.

To address this gap, we undertook an international scoping review of health care dashboard literature. Our aim was not only to map and catalog existing practices but also to organize them into a cohesive framework that bridges theory and practice. Building on previous work, our contribution is to consolidate existing evidence into 4 core pillars of dashboard design (approach, content, behavior, and adoption), which provide a foundation for more consistent, evidence-informed dashboard development in health care while remaining adaptable to diverse contexts.

Methods

Study Design

This study was conducted as a scoping review, following the framework proposed by Arksey and O'Malley [12] and enhanced by Levac et al [13]. Scoping reviews are particularly suited to mapping the existing literature in areas where guidelines are not well established, identifying key concepts, and synthesizing recurring practices. Our objective was to explore the international literature on health care dashboard design and identify common practices. As such, we did not perform formal quality appraisals, which is consistent with scoping review methodology. We report this review in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist [14] (Multimedia Appendix 1).

Needs Assessment

The initial proposal emerged from discussions between health professionals working on health service improvement initiatives and a data visualization expert. On the basis of a need to develop regional population profiles, a lack of clarity on effective data presentation was identified as an area that needed to be addressed.

Identify the Research Question

The objective of this scoping review was to explore current international health care literature to identify existing guidelines or, in the absence of such guidelines, to identify common practices for health care dashboard design. The following research questions guided the search strategy: (1) What

XSL•FO
RenderX

dashboards exist within (national) health systems that support decision-making? (2) Are there good practice guidelines for the creation of such dashboards? and (3) If not, are the design practices sufficiently described to enable the creation of dashboard design guidelines?

Proposal Development and Approval

The questions and the overall time plan were discussed with the project team. Concurrently, discussions with other stakeholders involved in the development of health dashboards allowed the team to gather feedback regarding the potential impact of developing an evidence summary.

Systematic Literature Search

We conducted an electronic search using PubMed, Embase, Scopus, and IEEE Xplore. The search strategy combined terms for dashboards and data visualization (eg, dashboard, data visualization, and visual analytics) with terms for health care and health systems (eg, health care and health infrastructure). Boolean operators and truncation were used to capture variations. Controlled vocabulary (eg, Medical Subject Headings [MeSH] terms such as data visualization and decision support systems) was included where applicable. These keywords and associated terms were based on the research questions and manual searches for papers about the topic. Searches were limited to English-language articles published between January 2014 and April 2024. The complete reproducible search strings for each database are provided in Multimedia Appendix 2.

Screen and Select Relevant Literature

On completion of the electronic searches, the results were imported into Microsoft Excel. After removal of duplicates, titles and abstracts were screened by 2 independent reviewers. Conflicts or disagreements about selections were resolved by a third independent reviewer.

Eligibility criteria were defined to ensure that the review captured evidence most relevant to contemporary health care dashboard design. We limited the search to publications from 2014 onward, reflecting significant advances in data visualization technologies, health care information systems, and the use of dashboards in health service decision-making over the past decade. Only English-language publications were included due to resource constraints for translation. We included peer-reviewed research papers, systematic reviews, and dissertations or theses, as these sources often provide detailed methodological and design insights. In contrast, conference abstracts, commentaries, and opinion papers were excluded because they generally provide insufficient methodological detail or practical guidance. The specific inclusion and exclusion criteria are summarized in Table 1.

Table 1. Eligibility criteria used for the screening of studies.

	Inclusion criteria	Exclusion criteria
Population	Data visualization and data dashboards	Clinical visualizations and clinical visual aids
Context	All health care settings and any country or health system	None
Types of evidence sources	Research papers, systematic reviews, and dissertations or theses (including empirical findings and author-reported practices)	Conference abstracts and commentaries or opinion papers
Languages	English	Other than English
Time	2014 to 2024	Publications before 2014

Consistent with scoping review methodology and PRISMA-ScR guidance, no formal risk of bias or quality appraisal of included studies was undertaken, as our aim was to map the characteristics and range of existing evidence rather than to assess study validity.

After the initial screening of titles and abstracts, the selected studies were reviewed in full by 2 authors, and any disagreements were resolved by a third reviewer.

A data charting table summarizing key characteristics of the included studies is provided in Multimedia Appendix 3. The table also indicates the extent to which each study described guidelines and specifies which sections of the paper were used for reflexive thematic analysis (RTA) [15].

Identify Pillars From Included Studies

After the full-text screening was complete, the guidelines and practices described were extracted into an Excel sheet. Using RTA, the first author (HV) coded each described practice with a descriptive keyword (eg, *user engagement* or *data quality*) and iteratively grouped related codes into broader conceptual clusters. These clusters were reviewed and discussed by all team members (HV, SB, and CC), leading to the identification of four overarching themes (pillars):

1. Approach—the *methods* and *mindset* that shape the design process
2. Content—the *type of information* that should be provided and displayed
3. Behavior—the *manner of interaction* with the information (access, user-friendliness, and navigation)
4. Adoption—the *practices* that support adoption and ongoing engagement

Multimedia Appendix 4 lists the second-level keywords that informed each pillar. The analysis was conducted in line with the interpretive and reflexive orientation of RTA, emphasizing depth of understanding over coder agreement or reproducibility.

XSL•FO
RenderX

Produce Reports

This paper presents the first report arising from this search. Future papers will detail how the identified pillars and associated practices have been contextualized in the Irish context and describe the overall impact of this initiative.

Ongoing Follow-Up and Dialogue

The pillars and associated practices will inform a series of workshops with users and stakeholders to communicate and contextualize the findings.

Results

After removing duplicates, a total of 1644 potentially relevant reports were identified. Two reviewers (HV and SB) screened the abstracts of the selected papers, while a third reviewer (CC) resolved disagreements between the 2 reviewers. A total of 1471 papers were excluded, while 173 were selected for full-text review. A citation search of the selected papers led to a further 39 records, which were screened as outlined earlier. This added another 13 papers to the 173 papers to be included in the full-text review. After assessing against the eligibility criteria, a total of 18 reports were included in the synthesis of evidence. The entire selection process is represented in a PRISMA-ScR flow diagram (Figure 1; full study characteristics and extracted data are provided in Multimedia Appendix 3) [16].

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension) flow diagram showing selection of relevant papers [16].

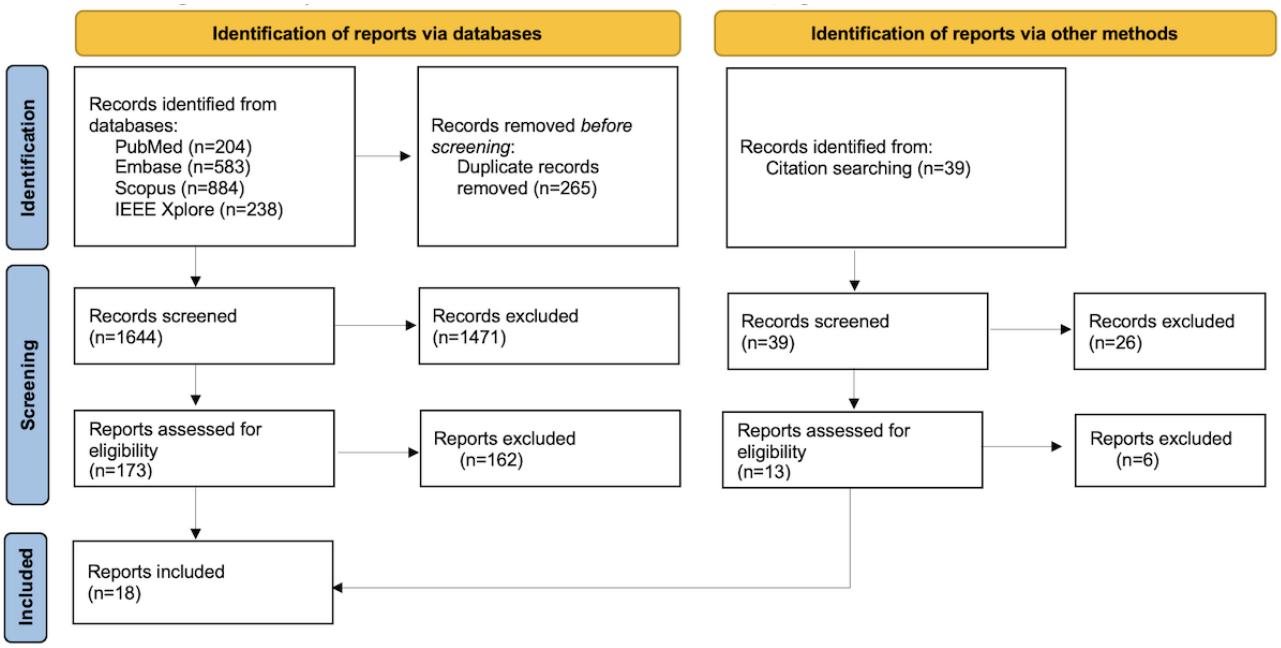

A total of 18 studies were included, with the majority focused on hospital settings (n=13, 72.2%). Few studies addressed national or public dashboards, and dashboards implemented at the community level were notably underrepresented, indicating a gap in evidence for nonhospital contexts. Two papers were review articles. Three papers focused on the development of national dashboards or the integration of a mobile app with a dashboard. Studies based in UK hospitals were the focus in 4 papers, while studies in the United States and the Netherlands were the setting for 2 papers each. The remainder included 1 study each in Australia, Iran, Portugal, Somalia, South Africa, South Korea, and Switzerland. Three papers did not specify the country in which the study was conducted.

In total, 4 papers [17-20] explicitly described comprehensive guidelines for dashboard development from start to finish, highlighting structured and replicable approaches, while an additional 9 papers [11,21-28] presented practices addressing specific aspects, requirements, or phases of dashboard design. The remaining 5 papers [29-33] either outlined specific design goals or described how dashboards were developed in practice but did not offer development practices that could inform broader or generalizable approaches to dashboard design. Overall, most implementations appeared ad hoc rather than following a structured framework, which may limit reproducibility and generalizability. Notably, half of the studies (9/18, 50%) focused primarily on evaluating dashboards after their creation, rather than explicitly detailing robust design principles during development.

Common practices identified across studies included user involvement in design and evaluation [17,28]; use of relevant and actionable metrics [25]; attention to data quality, including data accuracy and timeliness [30]; usability and interactivity [33]; and integration into existing workflows to avoid overburdening staff [20].

To illustrate, van de Baan et al [17] in the Netherlands developed quality and safety dashboards through a 5-stage participatory process with 13 health professionals across 2 care pathways. Their structured, clinician-led approach demonstrated how iterative co-design can enhance ownership and ensure dashboards address real-world needs (pillar: approach). Randell

XSL•FO
RenderX

et al [25], drawing on National Health Service clinical audit data, highlighted challenges in selecting indicators and ensuring their accuracy, translating these into clear requirements for dashboards that can support decision-making "from ward to board" (pillar: content). Loorak et al [33] introduced TimeSpan, a dashboard for stroke care pathways, where interactivity enabled clinicians to generate new insights and explore patterns beyond static reporting (pillar: behavior). Finally, Weggelaar-Jansen et al [20] showed that sustainable dashboard adoption required embedding dashboards into governance routines and review meetings, not just technical development (pillar: adoption).

Contrasts and conflicts across settings were notable: hospital dashboards prioritized clinical indicators and immediate operational utility, whereas public dashboards emphasized transparency, accessibility, and stakeholder accountability. Community dashboards were minimally represented, suggesting that practical guidance for these contexts is lacking. Furthermore, divergences in design philosophy emerged: some studies emphasized highly interactive dashboards with rich visualizations, while others favored static simplicity to reduce cognitive load. Similarly, governance and sustainability considerations varied: certain implementations benefited from strong institutional buy-in, whereas others struggled to maintain dashboards over time. These differences highlight potential tensions between usability, engagement, and sustainability and underscore that no single approach fits all contexts.

Methodologically, studies differed in both their analytic rigor and their use of visualization techniques. Hospital-based dashboards typically adopted iterative or co-design methods, enabling clinical validation and rapid refinement of metrics [17,18], whereas national or public-facing dashboards tended to rely on top-down technical builds using aggregated data [26]. These methodological differences shaped the visualization choices observed: co-designed hospital dashboards favored targeted, high-resolution displays with limited indicator sets, while public health dashboards prioritized breadth, using simplified or static visuals to maximize accessibility. Studies using participatory design generally reported stronger usability and contextual fit, while purely technical implementations often struggled with user engagement or indicator relevance.

Context also influenced implementation and sustainability. Hospital dashboards were typically embedded within established governance structures, allowing integration into performance review meetings and continuous data validation [20]. In contrast, public dashboards faced challenges of data standardization, cross-agency coordination, and long-term ownership, which limited ongoing updates. Methodological variation, therefore, reflected contextual constraints: where clinical workflows supported routine data collection, interactive and iterative designs thrived; where data pipelines were diffuse, design flexibility was restricted. These contrasts illustrate that methodological rigor and stakeholder proximity are decisive factors in determining dashboard longevity and real-world impact.

More than half of the papers (11/18, 61.1%) identified evaluation, appropriate visual encoding [7,34], and user involvement as key to developing a dashboard. Other key practices included modularity (the ability to add or delete components) and data quality (including timeliness and accuracy). Interactivity and the need to avoid adding to staff's existing workload were mentioned by 6 papers (33.3%).

The secondary analysis (RTA) resulted in 4 key themes (pillars) being defined: "approach," "content," "behavior," and "adoption." Most practices supported the "approach" pillar (33 unique mentions) (Table 2), while practices for "behavior" were comparatively few (15 unique mentions; Figure 2).

XSL•FO
RenderX

Figure 2. Key terms grouped by pillar: approach, content, behaviour and adoption.

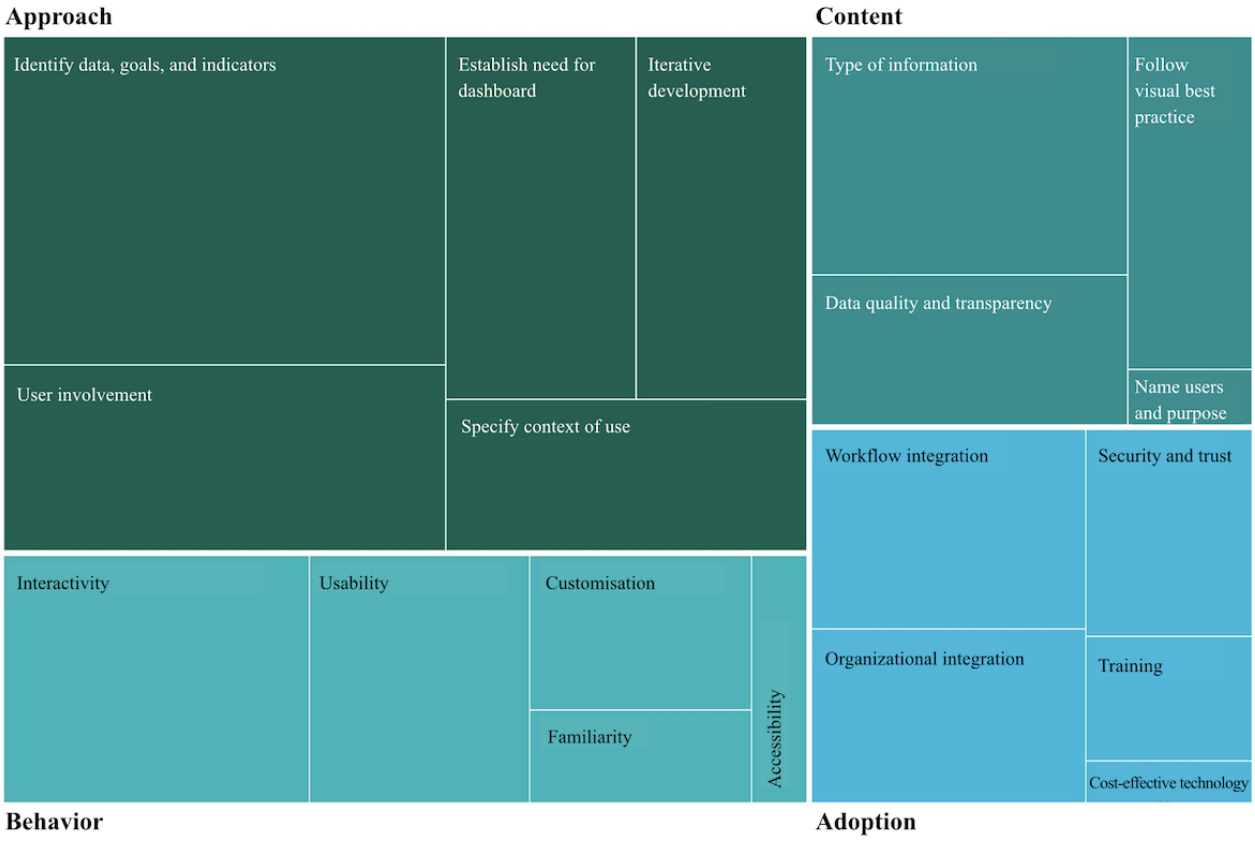

Table 2. Illustrative contributions of included studies to each pillar.

Pillar	Illustrative contributions (examples)
Approach	Participatory and iterative design approaches were reported in several studies [17,18,28], emphasizing co-design and ownership.
Content	Indicator selection and data quality were central concerns, as highlighted by Randell et al [25] on National Health Service clinical audits, Christen et al [32] on modular EMS dashboards, and Ivanković et al [26] on transparency in global COVID-19 dashboards.
Behavior	Usability and interactivity were highlighted in dashboards such as TimeSpan for stroke care [33], population health frameworks [16], and productivity dashboards [31].
Adoption	Sustainable uptake depended on workflow integration and governance structures, as described by Weggelaar-Jansen et al [20], Ahn et al [30], and Kamadjeu and Gathenji [27].

This finding underscores that while dashboard design is frequently addressed, critical attention to adoption, behavior change, and context-specific implementation is limited, suggesting gaps in translating design principles into practical impact.

Discussion

Overview

This review provides an insight into current practices in health care dashboard design. Our findings reveal a significant gap in the use of established guidelines or frameworks during the dashboard design process. While there is broad agreement on some core practices such as evaluation and user involvement, other areas such as data security or staff training are rarely mentioned. These gaps may arise from the reactive and ad hoc nature of dashboard implementation in response to immediate operational pressures, rather than from a lack of awareness of best practices. Moreover, our synthesis highlights how design priorities differ across health care contexts, reflecting the specific needs and constraints of each setting. Understanding these practices provides a foundation for mapping observed patterns to established visualization and HCI frameworks, which underpins the subsequent synthesis of the 4 pillars: approach, content, behavior, and adoption.

Anchoring the 4 Pillars in Established Theory

Theoretical Alignment With Visualization

Each of the 4 pillars aligns with recognized frameworks in visualization and HCI. Approach reflects principles of user-centered and participatory design, emphasizing stakeholder involvement and iterative cocreation (ISO 9241-210 [9]). Content is consistent with the nested model by Munzner [7] and the semiology by Bertin [34], underscoring the importance of selecting actionable metrics and applying effective visual encoding. Behavior draws on HCI principles of usability [35] and the concept of affordances by Norman [36], highlighting interactivity, customization, and accessibility. Finally, adoption

XSL•FO
RenderX

resonates with sociotechnical theories such as the technology acceptance model [37] and normalization process theory [38]. By mapping inductively derived themes to these theories, the 4 pillars provide a bridge between conceptual models and applied health care practice, ensuring the framework is both empirically grounded and theoretically anchored.

These theoretical frameworks were not used to generate the pillars but were applied retrospectively as an interpretive lens to examine their coherence and alignment with established concepts in visualization, usability, and sociotechnical theory. This approach ensures that the 4 pillars remain inductively grounded while being conceptually validated against recognized theoretical models.

Pillar 1: Approach

This pillar includes the methods and mindsets that shape the dashboard design process. It focuses strongly on involving and engaging end users and stakeholders in defining needs, iterative cocreation, and a review of existing systems to ensure that the best solution is put forward for development.

Pillar 2: Content

This pillar describes the type of information a dashboard should provide and how it should be displayed. It includes the need for relevant and actionable metrics and an examination of data accuracy and quality. Content should be presented following visual best practices (eg, using percentages instead of raw numbers, eliminating visual noise, and highlighting key signals) to ensure clarity and support decision-making.

Pillar 3: Behavior

Behavior includes practices related to how users interact with the dashboard. This includes interactivity to support engagement and exploration, as well as usability and accessibility considerations to accommodate diverse levels of data literacy.

Pillar 4: Adoption

This pillar focuses on practices that support adoption both by immediate users and the wider organization, thereby ensuring longevity and sustainability. Practices such as the ability to customize the dashboard, workflow integration to minimize additional staff workloads, broader integration into the organizational culture through stakeholder engagement, and continuous improvements help to ensure that the dashboard becomes part of daily operations. Attention to cost-effective tools, data security, privacy, and continuous evaluation further supports sustained adoption.

Principal Findings

A key insight from our review is that 50% (9/18) of studies focused on evaluating dashboards after their creation rather than incorporating robust design principles during development. Evaluation efforts often highlighted issues with usability, data integration, and user engagement, underscoring the importance of proactive design strategies to address these challenges upfront. Ad hoc development and inconsistent application of design principles were common, contributing to persistent gaps in effective dashboard implementation. Adoption-related factors, such as governance, training, and sustainability, were rarely addressed, despite their known importance for long-term use.

Our analysis also demonstrates that context shapes dashboard priorities: hospitals often emphasize operational and clinical metrics to support immediate decision-making, while publicly accessible dashboards tend to prioritize transparency and stakeholder communication. Community and population health dashboards were scarce, which may explain the limited guidance on integrating diverse data sources or engaging broader stakeholder groups. Conflicting design priorities were also noted: some studies emphasized highly interactive dashboards to enhance engagement, while others favored simpler, static dashboards to reduce cognitive load or resource requirements. Technological, governance, and sustainability challenges were frequently overlooked, particularly in nonhospital or community settings.

Furthermore, while concepts such as user involvement and iterative refinement were mentioned, they were not consistently applied, leaving gaps in achieving effective, user-centered dashboards.

Taken together, the 4 pillars summarize a practical set of considerations for dashboard design and adoption. They provide a lens for assessing both technical and sociotechnical aspects, from user-centered approaches and data quality to usability, governance, and sustainability. For practice, this highlights embedding dashboards into workflows with training and stakeholder engagement. For policy, investment is required in the sociotechnical infrastructure to ensure dashboards are effective and sustainable beyond immediate returns.

Strengths and Limitations

This review contributes to the understanding of best practices in health care dashboard design by synthesizing existing literature and identifying key themes. The integration of theory strengthens the framework and allows critical evaluation of divergences and gaps. Limitations include scope and selection bias, lack of empirical validation, and the overrepresentation of hospital settings. Future research should test the 4 pillars in underrepresented contexts, such as primary care, community health, and public dashboards, and evaluate impacts on usability, adoption, and decision-making.

Implications for Ireland

Although international in scope, these findings provide an evidence base for the next phase of work in the Irish health care system. The 4 pillars will serve as the starting point for a consensus building process with clinicians, public health professionals, service planners, policymakers, IT specialists, and patient or community representatives. This process will contextualize international practices to Irish data infrastructures, governance, and regulatory frameworks, ensuring that future dashboards are both evidence informed and locally relevant.

Future Work

This review represents an initial step toward establishing structured, evidence-based guidelines for health care dashboard design, synthesizing international best practices into the 4 pillars: approach, content, behavior, and adoption. Building on the

XSL•FO
RenderX

findings of this international review, the next phase of this work focuses on consensus building and co-design within the Irish health care system. Specifically, the 4 pillars will provide the starting point for structured discussions with Irish stakeholders, ensuring that international evidence is explicitly adapted to local infrastructure, governance, and regulatory requirements. By engaging a diverse range of stakeholders—including public health professionals; service planners; clinicians; IT specialists; policymakers; and public, patient, and community partners—we will refine the practices under each of the 4pillars into actionable, context-specific guidelines. For example, the practice of ensuring data quality (pillar: content) will be mapped to current data sources within Ireland, their interoperability, and Ireland-specific data protection principles (eg, General Data Protection Regulation [39]), illustrating how international guidance informs practical national implementation.

Through this iterative process, the resulting guidelines will be globally informed; locally applicable; and co-designed to foster equity, accessibility, and usability.

Several areas also warrant further investigation to strengthen the evidence base supporting the 4 pillars. First, empirical evaluation of the 4 pillars across diverse health care settings, including primary care, community health, and public-facing dashboards, could test their impact on usability, adoption, and decision-making. Second, the 4 pillars could be translated into practical tools, such as checklists, design templates, or evaluation criteria, for use by health care organizations and developers. Third, comparative studies with other sectors (eg, business intelligence and education) may provide insights into transferable practices and highlight areas requiring health care–specific adaptation.

Conclusions

This review is an essential first step in understanding the global landscape of health care dashboards and provides a foundation for contextualization in Ireland. As outlined earlier, we have identified a significant gap, namely, a lack of standardized principles that could guide the design of dashboards in the health care sector. Arising from this review, we have defined 4 main pillars with associated key practices. These 4 pillars, derived from the international literature, will guide the upcoming consensus building process in Ireland and ensure that international evidence informs locally relevant guidelines. They will serve as the foundation for creating a comprehensive framework for health care data visualization in Ireland.

Ultimately, this phased approach from international review to consensus building and framework development ensures a solid foundation of evidence and co-design for transforming the design of data dashboards in the Irish health care system. By enabling dashboards that are not only functional but also user-friendly and impactful, this work supports improved decision-making, resource allocation, and health outcomes for all stakeholders.

Acknowledgments

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the Generative AI Delegation Taxonomy (GAIDeT) [40], the following task was delegated to GenAI tools under full human supervision: publication support. The GAI tool used was GPT-5-mini. The tool was used solely to assist with responding to reviewer comments, including rewriting text for clarity and consistency, providing clarification, and formulating responses to reviewers. GenAI was not used for ideation, study design, data analysis, interpretation of results, or drafting of the original manuscript. All content generated with AI assistance was reviewed, edited, and approved by the authors. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This publication has emanated from research jointly funded by Taighde Éireann—Research Ireland under grant number 12/RC/2289_P2—Insight Research Ireland Centre for Data Analytics, and supported by a Science Foundation Ireland Public Fellowship Award (project: HSE2 Dashboard Design for Health Service Executive Quality Improvement Reports). For the purpose of open access, the authors have applied a CC BY public copyright license to any author-accepted manuscript version arising from this submission.

Authors' Contributions

HV, SB, and JM developed the initial research objectives. HV, SB, and CC reviewed and selected relevant papers. HV and SB carried out the reflexive thematic analysis of the methodology sections of all papers and, together with CC and LKI, synthesized the themes into pillars. HV and SB drafted the initial manuscript. CC, LKI, GM, DM, and JM reviewed the manuscript and provided suggestions for improvement.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR checklist.
[DOC File , 119 KB - jmir_v28i1e77361_app1.doc]

XSL•FO
RenderX

Multimedia Appendix 2
Electronic search report.
[DOCX File , 23 KB - jmir_v28i1e77361_app2.docx]

Multimedia Appendix 3
Characteristics of reports included.
[DOCX File , 37 KB - jmir_v28i1e77361_app3.docx]

Multimedia Appendix 4
Overview of coding for reflexive thematic analysis.
[DOCX File , 33 KB - jmir_v28i1e77361_app4.docx]

Multimedia Appendix 5
Reports mapped to pillars.
[DOCX File , 26 KB - jmir_v28i1e77361_app5.docx]

References

1. Sarikaya A, Correll M, Bartram L, Tory M, Fisher D. What do we talk about when we talk about dashboards? IEEE Trans Vis Comput Graph 2018 Aug 21:682-692. [doi: 10.1109/TVCG.2018.2864903] [Medline: 30136958]
2. Sweller J. Cognitive load during problem solving: effects on learning. Cogn Sci 2010 Feb 11;12(2):257-285 [FREE Full text] [doi: 10.1207/s15516709cog1202_4]
3. Bateman S, Mandryk RL, Gutwin C, Genest A, McDine D, Brooks C. Useful junk? The effects of visual embellishment on comprehension and memorability of charts. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2010 Presented at: CHI '10; April 10-15, 2010; Atlanta, GA URL: https://doi.org/10.1145/1753326.1753716 [doi: 10.1145/1753326.1753716]
4. Wanderer JP, Nelson SE, Ehrenfeld JM, Monahan S, Park S. Clinical data visualization: the current state and future needs. J Med Syst 2016 Dec;40(12):275. [doi: 10.1007/s10916-016-0643-x] [Medline: 27787779]
5. Lurie NH, Mason CH. Visual representation: implications for decision making. Journal of Marketing 2007 Jan 01;71(1):160-177 [FREE Full text] [doi: 10.1509/jmkg.71.1.160]
6. Abudiyab NA, Alanazi AT. Visualization techniques in healthcare applications: a narrative review. Cureus 2022 Nov;14(11):e31355 [FREE Full text] [doi: 10.7759/cureus.31355] [Medline: 36514654]
7. Munzner T. Visualization Analysis and Design. Boca Raton, FL: CRC Press; 2014.
8. Few S. Information Dashboard Design: The Effective Visual Communication of Data. Sebastopol, CA: O'Reilly Media; 2006.
9. ISO 9241-210:2019 Ergonomics of human-system interaction part 210: human-centred design for interactive systems. International Organization for Standardization. 2019. URL: https://www.iso.org/standard/77520.html [accessed 2025-10-20]
10. Helminski D, Sussman JB, Pfeiffer PN, Kokaly AN, Ranusch A, Renji AD, et al. Development, implementation, and evaluation methods for dashboards in health care: scoping review. JMIR Med Inform 2024 Dec 10;12:e59828 [FREE Full text] [doi: 10.2196/59828] [Medline: 39656991]
11. Rabiei R, Almasi S. Requirements and challenges of hospital dashboards: a systematic literature review. BMC Med Inform Decis Mak 2022 Nov 08;22(1):287 [FREE Full text] [doi: 10.1186/s12911-022-02037-8] [Medline: 36348339]
12. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. International Journal of Social Research Methodology 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
13. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
14. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
15. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Jun 13;11(4):589-597 [FREE Full text] [doi: 10.1080/2159676x.2019.1628806]
16. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev 2021 Mar 29;10(1):89 [FREE Full text] [doi: 10.1186/s13643-021-01626-4] [Medline: 33781348]
17. van de Baan FC, Lambregts S, Bergman E, Most J, Westra D. Involving health professionals in the development of quality and safety dashboards: qualitative study. J Med Internet Res 2023 Jun 12;25:e42649 [FREE Full text] [doi: 10.2196/42649] [Medline: 37307058]

XSL•FO
RenderX

18. Elshehaly M, Randell R, Brehmer M, McVey L, Alvarado N, Gale CP, et al. QualDash: adaptable generation of visualisation dashboards for healthcare quality improvement. IEEE Trans Vis Comput Graph 2021 Feb;27(2):689-699 [FREE Full text] [doi: 10.1109/TVCG.2020.3030424] [Medline: 33048727]
19. Concannon D, Herbst K, Manley E. Developing a data dashboard framework for population health surveillance: widening access to clinical trial findings. JMIR Form Res 2019 Apr 04;3(2):e11342 [FREE Full text] [doi: 10.2196/11342] [Medline: 30946016]
20. Weggelaar-Jansen AM, Broekharst DS, de Bruijne M. Developing a hospital-wide quality and safety dashboard: a qualitative research study. BMJ Qual Saf 2018 Dec;27(12):1000-1007 [FREE Full text] [doi: 10.1136/bmjqs-2018-007784] [Medline: 29950323]
21. Lee K, Jung SY, Hwang H, Yoo S, Baek HY, Baek RM, et al. A novel concept for integrating and delivering health information using a comprehensive digital dashboard: an analysis of healthcare professionals' intention to adopt a new system and the trend of its real usage. Int J Med Inform 2017 Jan;97:98-108. [doi: 10.1016/j.ijmedinf.2016.10.001] [Medline: 27919400]
22. Alvarado N, McVey L, Elshehaly M, Greenhalgh J, Dowding D, Ruddle R, et al. Analysis of a web-based dashboard to support the use of national audit data in quality improvement: realist evaluation. J Med Internet Res 2021 Nov 23;23(11):e28854 [FREE Full text] [doi: 10.2196/28854] [Medline: 34817384]
23. Bunting Jr RF, Siegal D. Developing risk management dashboards using risk and quality measures: a visual best practices approach. J Healthc Risk Manag 2017 Oct;37(2):8-28. [doi: 10.1002/jhrm.21287] [Medline: 28960593]
24. Fazaeli S, Khodaveisi T, Vakilzadeh AK, Yousefi M, Ariafar A, Shokoohizadeh M, et al. Development, implementation, and user evaluation of COVID-19 dashboard in a third-level hospital in Iran. Appl Clin Inform 2021 Oct;12(5):1091-1100 [FREE Full text] [doi: 10.1055/s-0041-1740188] [Medline: 34879405]
25. Randell R, Alvarado N, McVey L, Ruddle RA, Doherty P, Gale C, et al. Requirements for a quality dashboard: lessons from National Clinical Audits. AMIA Annu Symp Proc 2019;2019:735-744 [FREE Full text] [Medline: 32308869]
26. Ivanković D, Barbazza E, Bos V, Brito Fernandes Ó, Jamieson Gilmore K, Jansen T, et al. Features constituting actionable COVID-19 dashboards: descriptive assessment and expert appraisal of 158 public web-based COVID-19 dashboards. J Med Internet Res 2021 Feb 24;23(2):e25682 [FREE Full text] [doi: 10.2196/25682] [Medline: 33577467]
27. Kamadjeu R, Gathenji C. Designing and implementing an electronic dashboard for disease outbreaks response - case study of the 2013-2014 Somalia Polio outbreak response dashboard. Pan Afr Med J 2017;27(Suppl 3):22 [FREE Full text] [doi: 10.11604/pamj.supp.2017.27.3.11062] [Medline: 29296157]
28. Khanbhai M, Symons J, Flott K, Harrison-White S, Spofforth J, Klaber R, et al. Enriching the value of patient experience feedback: web-based dashboard development using co-design and heuristic evaluation. JMIR Hum Factors 2022 Feb 03;9(1):e27887 [FREE Full text] [doi: 10.2196/27887] [Medline: 35113022]
29. Ratwani RM, Fong A. 'Connecting the dots': leveraging visual analytics to make sense of patient safety event reports. J Am Med Inform Assoc 2015 Mar;22(2):312-317. [doi: 10.1136/amiajnl-2014-002963] [Medline: 25336592]
30. Ahn E, Liu N, Parekh T, Patel R, Baldacchino T, Mullavey T, et al. A mobile app and dashboard for early detection of infectious disease outbreaks: development study. JMIR Public Health Surveill 2021 Mar 09;7(3):e14837 [FREE Full text] [doi: 10.2196/14837] [Medline: 33687334]
31. Pestana M, Pereira R, Moro S. Improving health care management in hospitals through a productivity dashboard. J Med Syst 2020 Mar 12;44(4):87. [doi: 10.1007/s10916-020-01546-1] [Medline: 32166499]
32. Christen OM, Mösching Y, Müller P, Denecke K, Nüssli S. Dashboard visualization of information for emergency medical services. Stud Health Technol Inform 2020 Nov 23;275:27-31. [doi: 10.3233/SHTI200688] [Medline: 33227734]
33. Loorak MH, Perin C, Kamal N, Hill M, Carpendale S. TimeSpan: using visualization to explore temporal multi-dimensional data of stroke patients. IEEE Trans Visual Comput Graphics 2016 Jan;22(1):409-418. [doi: 10.1109/tvcg.2015.2467325]
34. Bertin J. Sémiologie Graphique: Les Diagrammes, Les Réseaux, Les Cartes. Berlin, Germany: Mouton Publishers; 1973.
35. Nielsen J. Enhancing the explanatory power of usability heuristics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 1994 Presented at: CHI '94; April 24-28, 1994; Boston, MA URL: https://doi.org/10.1145/191666.191729 [doi: 10.1145/191666.191729]
36. Norman D. Affordances and design. jnd.org. 2008. URL: https://jnd.org/affordances-and-design/ [accessed 2025-08-25]
37. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
38. May C, Finch T. Implementing, embedding, and integrating practices: an outline of normalization process theory. Sociology 2009 Jun 15;43(3):535-554 [FREE Full text] [doi: 10.1177/0038038509103208]
39. Data Protection Commission. URL: https://www.dataprotection.ie/en/organisations [accessed 2025-01-15]
40. GAIDeT declaration generator. GitHub. URL: https://panbibliotekar.github.io/gaidet-declaration/index.html [accessed 2025-11-13]

Abbreviations

MeSH: Medical Subject Headings

XSL•FO
RenderX

PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RTA: reflexive thematic analysis

Edited by J Sarvestan; submitted 16.May.2025; peer-reviewed by S Guan, G Adewumi, O Nguyen; comments to author 20.Aug.2025; accepted 18.Nov.2025; published 25.Feb.2026.

Please cite as:
Vornhagen H, Barrett S, Carroll C, Iladiva LK, Martin G, McKeown D, Martin J
Design Practices for Data Dashboards in Health Care: Scoping Review
J Med Internet Res 2026;28:e77361
URL: https://www.jmir.org/2026/1/e77361
doi:10.2196/77361
PMID:

©Heike Vornhagen, Stephen Barrett, Ciara Carroll, Lydia Kavochi Iladiva, Gregory Martin, Declan McKeown, Jennifer Martin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Effects of Digital Health Interventions for Family Members in Intensive Care Units: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Chuchu Zhang[1], BSN; Weijing Sui[1], MSN; Weilin Jiang[1], BSN; Yibing Shen[1], BSN; Zhenzhen Huang[1], BSN; Ran Yan[1], BSN; Jia Yi[1], BSN; Jiayu Zhang[1], MSN; Chaoyi Zhang[1], BSN; Yiyu Zhuang[1], MSN; Xiaoyan Gong[1], MSN

Department Nursing, Zhejiang University School of Medicine Sir Run Run Shaw Hospital, Hangzhou, Zhejiang, China

Corresponding Author:
Xiaoyan Gong, MSN
Department Nursing
Zhejiang University School of Medicine Sir Run Run Shaw Hospital
No.3 Qingchun East Road, Shangcheng District
Hangzhou, Zhejiang, 310016
China
Phone: 86 0571 86006616
Email: 3203071@zju.edu.cn

Related Article:

Comment in: https://www.jmir.org/2026/1/e95876/

Abstract

Background: Family members of intensive care units (ICUs) often experience psychological difficulties like anxiety and depression, which affect family functioning and their ability to support patient recovery. Digital health interventions (DHIs) offer accessible and timely support; yet, evidence regarding their effectiveness for ICU families demonstrates inconsistent.

Objective: This systematic review and meta-analysis aimed to explore the impact of DHIs on psychological outcomes, quality of life (QoL), and quality of communication (QOC) among ICU family members.

Methods: A total of 9 databases (CNKI [China Knowledge Resource Integrated Database], Wanfang, VIP [Weipu Database], SinoMed, Cochrane Library, PubMed, Embase, CINAHL, and Web of Science) were searched for randomized controlled trials (RCTs) published up to October 11, 2025. Eligible studies evaluated DHIs targeting adult ICU patients' family members and reported psychological outcomes, QoL, or QOC; interventions used solely for monitoring or tracking were excluded. Furthermore, 2 reviewers independently screened studies, extracted data, and assessed the risk of bias using the Cochrane Risk of Bias Tool. Random-effects meta-analyses were performed using the Hartung-Knapp-Sidik-Jonkman method. Prediction intervals (PIs) were calculated to reflect between-study heterogeneity. Subgroup analyses were performed by patients' primary diagnosis, relationship with patients, and type of digital health technology. Evidence quality was assessed using the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) approach.

Results: In total, 17 RCTs involving 1864 participants were included. Interrater agreement was high across study selection and assessment processes (κ=0.55-1.00). Most studies were judged to have some concerns about the risk of bias. Meta-analyses showed no statistically significant pooled effects of DHIs on anxiety (standardized mean difference [SMD] –0.34, 95% CI –0.68 to 0.00; 95% PI –1.46 to 0.79; *P*=.05), depression (SMD –0.26, 95% CI –0.52 to 0.01; 95% PI –1.11 to 0.60; *P*=.06), posttraumatic stress disorder (SMD –0.21, 95% CI –0.49 to 0.06; 95% PI –1.03 to 0.60; *P*=.11), QoL (SMD 0.09, 95% CI –0.10 to 0.28; 95% PI –0.41 to 0.60; *P*=.36), or QOC (SMD 0.14, 95% CI –0.03 to 0.31; 95% PI –0.26 to 0.55; *P*=.10). Wide PIs indicated substantial variability in intervention effects across settings. The certainty of evidence ranged from very low to moderate.

Conclusions: This review emphasizes significant uncertainty in estimated effects and provides the first thorough synthesis of RCT evidence on DHIs for ICU family members. The inherently high heterogeneity and low certainty of the underlying evidence limit the conclusions of this review, despite its methodological rigor and use of PIs. Compared with previous reviews, this review concentrates on the ICU family members and includes more recent research. Future studies should focus on high-risk subgroups,

XSL•FO
RenderX

use mixed methods designs, and develop theory-informed, personalized, and interactive DHIs. DHIs may enable proactive, customized communication techniques for families of ICU patients, even when there are few formal clinical recommendations.

Trial Registration: PROSPERO CRD420251044704; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251044704

(J Med Internet Res 2026;28:e83294) doi:10.2196/83294

KEYWORDS

digital health interventions; intensive care units; family members; psychology; systematic review; meta-analysis; randomized controlled trials

Introduction

Patients admitted to the intensive care units (ICUs) frequently experience varying degrees of organ or system failure, which can be life-threatening [1]. Their families may endure enormous burdens and suffer from psychological disorders like anxiety, depression, and posttraumatic stress disorder (PTSD). The Society of Critical Care Medicine's 2024 Clinical Practice Guidelines on Family-Centered Care for Adult ICUs underscore the need for delivering patient-centered and family-centered care in the ICU to uphold clinical excellence [2]. Nonetheless, various psychological challenges for family members, including anxiety, depression, acute stress disorder, PTSD, and complicated grief, are induced by the uncertainty of disease trajectories, the complexity of medical decisions, and the urgent demand for time. The term "postintensive care syndrome family" (PICS-F) encompasses these mental health disorders [3,4]. In addition to undermining family functioning and integrity, these problems may hinder the family's ability to facilitate the patient's recovery in the ICU [5,6].

A retrospective study [6] showed that the prevalence of psychological disorders related to PICS-F among family members of ICU patients within 6 months post admission was 15.1%, with spouses, as primary caregivers, exhibiting a heightened risk of psychological disorders. Furthermore, 1 research indicated that 10% to 69% of family members of ICU patients experienced moderate to severe psychological distress, including symptoms of anxiety, depression, and PTSD, potentially persisting for up to a year [7,8]. Family members endure not only emotional distress, such as anxiety and depression, but also are influenced by several biopsychosocial factors, including sleep deprivation, diminished quality of life (QoL) and mental health, and insufficient social support. These factors make it difficult for them to provide adequate family support to facilitate the patient's recovery [5]. Common contributors to PICS-F among family members include female sex, relatively younger age, younger patient age, lower educational level, unmarried status, inadequate communication skills of ICU medical personnel, and the degree of family participation in decision-making [3,9]. Several strategies, such as ICU diaries [10], information support, and peer support, have been implemented to mitigate PICS-F but often focus on the patient's hospitalization problems [11,12], and due to the family dynamics, geography, family resources, and family values, the effect of related intervention on PTSD, anxiety, and depression among family members is unclear [13]. Traditional educational tools are rarely implemented in the ICU environment due to high service costs, limited resources, and a lack of personalization [14,15]. With the rapid development of internet technology, mobile health care solutions have become popular, providing innovative psychological empowerment interventions [16].

Digital health technology (DHT) was initially introduced by the World Health Organization in the Global Strategy for Digital Health (2020-2024) and is defined as a comprehensive technical practice that facilitates clinical care, health-related education for patients and professionals, and public health through the extensive usage of electronic information and communication technologies [17,18]. Compared with conventional interventions, digital health interventions (DHIs) amalgamate computing platforms, connectivity, software, networked devices, and sensors—including smartphone apps, wearable devices, web-based communication platforms, videoconferencing, chatbots, artificial intelligence, and virtual reality (VR) technologies—leveraging technological advancements to transcend temporal and spatial limitations [19,20]. DHIs provide family members with greater accessibility, affordability, and convenience in receiving timely support and information, which is critical in the ICU context. In addition, DHIs help reduce health care spending [21,22]. DHIs have been shown to improve the psychological outcomes and QoL of caregivers, such as depressive symptoms, perceived stress, anxiety, and stress in other high-stress situations, such as caregivers of patients with cancer [23], dementia [24,25], and mental illness [15]. This study highlights the potential value of its application to family members of ICU patients. However, other studies have suggested that DHIs have little effect on alleviating the anxiety and depression of these family members [26]. Therefore, there are ambiguities in the existing research results on whether DHIs can effectively alleviate the psychological distress symptoms of family members of ICU patients [26].

After searching relevant databases, no meta-analysis of these intervention outcomes has been found. Randomized controlled trials (RCTs) continue to serve as the gold standard for delivering the most rigorous scientific findings [27]. Therefore, a systematic review and meta-analysis of RCTs focusing on the effects of DHIs on ICU family members is essential. This meta-analysis will assess the effects of DHIs on alleviating psychological symptoms, improving QoL, and enhancing communication quality among ICU family members, thereby providing critical evidence for the formulation of a comprehensive and effective psychological support program.

XSL•FO
RenderX

Methods

Research Design

This systematic review and meta-analysis were planned, conducted, and reported following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) [28] (Multimedia Appendix 1). A systematic literature search, screening, study selection, data extraction, and risk-of-bias assessment were conducted according to the Cochrane methodology [29].

Search Strategy

The search strategy was reported in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-analyses literature search extension) checklist (Multimedia Appendix 2). In total, 2 investigators (Chuchu Z and WJ) independently systematically searched 9 databases, including CNKI [China Knowledge Resource Integrated Database], Wanfang, VIP (Weipu Database), SinoMed, Cochrane Library, PubMed, Embase, CINAHL, and Web of Science, from inception to October 11, 2025. The search terms included "digital health," "intensive care units," "family," and "randomized controlled trial," which were pilot-tested and validated by independent researchers. No published search filters for study design were used. To identify additional studies, we manually searched the reference lists of included articles and relevant websites, including PROSPERO clinical trial registries. Before the final data analysis, the database searches were updated. Gray literature, such as conference abstracts and academic dissertations, was excluded. The whole search strategy is available in Multimedia Appendix 3.

Eligibility Criteria

Studies were selected based on the following inclusion criteria: (1) population: adult family members (≥18 y) of adult ICU patients (a family member could be a spouse, partner, friend, or relative [blood-related or not]); (2) interventions: all DHIs, which may include one or more methods, such as telephone, video, websites, or mobile apps, wearable devices, and others; (3) control: a comparison group that received either standard care, routine care or a blank control; (4) outcomes: the primary outcomes associated with the psychological health of family members included anxiety, depression, PTSD, and other related outcomes like QoL; and (5) study type: RCTs.

Studies were excluded if they met one of the following criteria: (1) articles were excluded if the DHI was used for monitoring or tracking purposes only; (2) duplicate publications, or (3) full text was not available.

Study Selection

The search results were imported into NoteExpress (Aegean Software) for review. After removing duplicates, 2 investigators (Chuchu Z and RY) independently screened the titles and abstracts, then reviewed the full texts of the remaining records. Any disagreements were resolved by discussion with the third investigator (WS). The κ coefficients were used to specify the consistency of independent work. The degree of consistency was identified as slight, fair, moderate, substantial, and almost perfect when the κ coefficients were calculated as 0.00-0.20, 0.21-0.40, 0.41-0.60, 0.61-0.80, and 0.81-1.00, respectively [30].

Data Extraction

Two authors (Chuchu Z and JY) independently extracted data from each of the included studies using a structured data extraction form, including (1) study: author, year, country, and design; (2) participants: patients' diagnosis, sample size, female, age, relationship with patients, occupation, marital status, education level, and financial burden; (3) control; (4) experimental: intervention measurement, intervention delivery (categorized as mobile app, internet-based, and others [eg, telephone calls, VR, and emails]), intervention duration, and data collection time points; (5) outcomes and measurements; and (6) results. If necessary, the corresponding authors were contacted by email and asked for the missing information. A third reviewer (XG) verified the extracted data.

Study Risk of Bias Assessment

The Cochrane's risk-of-bias tool for randomized trials, version 2 (RoB 2), was used by 2 independent reviewers (JZ and ZH) to assess [31,32]. The RoB 2 tool consists of 5 domains: randomization process, deviations from intended intervention, missing outcome data, measurement of the outcome, and selection of the reported result. The risk of bias was assessed as low, some concerns, or high in the RoB 2 based on responses to signal questions in each domain. The κ coefficient was calculated for each item to identify the degree of consistency.

Quality of Evidence

The Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) approach was used to assess the certainty of the body of evidence [33]. According to the approach, the evidence quality of RCTs was initially graded as high and downgraded to moderate, low, or very low when any limitations were identified in risk of bias, inconsistency, directness, imprecision, or publication bias. However, the evidence could be upgraded by a significant effect and dose-response gradient. Furthermore, 2 authors (Chaoyi Z and YS) independently assessed the quality of evidence using the GRADE handbook, and the interrater agreement was evaluated using the κ coefficient. Disagreements were resolved through discussion, or help was sought from a third author (YZ).

Statistical Analysis

For continuous outcomes, the Hartung-Knapp-Sidik-Jonkman random-effects method [34] was used to obtain standardized mean differences (SMDs) with 95% CIs when studies used different psychometric scales. In studies with follow-up lasting more than 1 month, data from the shortest follow-up period were used. A meta-analysis was undertaken if more than 2 studies reported outcomes in the same format; otherwise, the studies' findings were described narratively [35]. We applied random-effects models because of suspected heterogeneity across studies to provide a relatively conservative estimate of the combined results [36]. Moreover, the prediction intervals (PI) were reported to indicate heterogeneity in the same metric as SMD and to reflect the variation in actual treatment effects expected in future studies, which aids clinical decision-making

XSL•FO
RenderX

[37]. A leave-one-out sensitivity analysis was performed to assess the robustness of the synthesized results. Subgroup analysis was conducted to explore the potential sources of heterogeneity. These analyses were based on patient primary diagnosis (medical vs surgical), relationship with patients (spouse vs children), and type of DHT (education-based vs communication-based vs mental health–related). Finally, if sufficient studies were available for the outcome (n>10), a funnel plot was generated to assess potential small-study effects [38]. The Egger test was performed to further quantify funnel plot asymmetry, with a *P* value <.05 considered statistically significant [39]. A 2-tailed *P*<.05 was considered statistically significant. All analyses and visualizations were conducted using the meta package in R software (version 4.5.1; R Core Team).

Ethical Considerations

This study and paper used published data, so no ethics approval was needed.

Results

Study Selection

The PRISMA flow diagram in Figure 1 summarizes the decision pathway for final study inclusion. In total, 2 authors (Chuchu Z and WJ) independently conducted the study selection process, and the κ coefficients for title or abstract screening and full-text reading were 0.82 (*P*<.001) and 0.84 (*P*<.001), respectively, indicating almost perfect consistency in these steps. We identified 2624 records through database searches and 2 records through citation searching. After title and abstract screening, we excluded 1944 records and identified 55 records for full-text evaluation. After the full-text screening, we excluded 38 records that did not meet the inclusion criteria and included 17 studies [14,16,26,40-53].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) flow diagram. CNKI: China Knowledge Resource Integrated Database; VIP: Weipu Database.

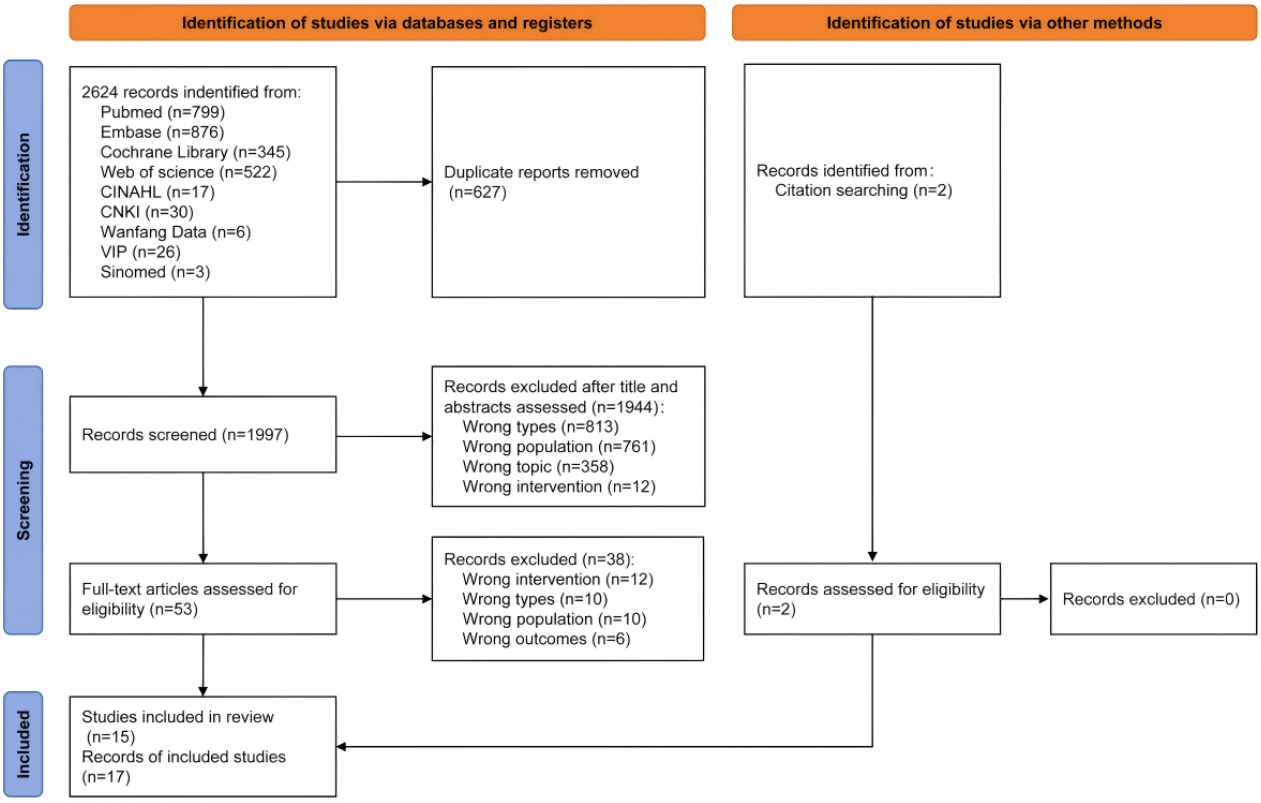

Study Characteristics

Overview

The characteristics of all the included studies are summarized in Table 1. Among the 17 RCTs, 2 were cluster RCTs [41,43], and 4 were pilot RCTs [26,45,47,53]. Studies were conducted in the United States (n=8) [26,41,42,45,47,51-53], China (n=3) [16,40,50], the Netherlands (n=1) [43], Norway (n=1) [48], Germany (n=1) [44], Austria (n=1) [14], Iran (n=1) [46], and Korea (n=1) [49], which were published from 2017 to 2025.

XSL•FO
RenderX

Table 1. Characteristics of the studies and participants in the included studies.

Study (year, country)	Design	Patient's diagnosis	Sample size, n (I/C[a]), % of female	Age (y), mean (SD)	Relationship with patients, n (%)	Occupation, n (%)	Marital status, n (%)	Education level, n (%)	Financial burden, n (%)
Carlozzi et al [51] (2025, United States)	RCT[b]	• Traumatic brain injury 100%	257 (129/128), 79.1	52.25 (14.48)	• Partner 104 (40.9) • Parent 83 (32.7) • Child 36 (14.2) • Sibling 20 (7.9) • Other 6 (2.4)	• Employed 157 (61.8) • Retired 58 (22.8) • Unemployed 15 (5.9) • Disabled 8 (3.1) • Other 16 (6.3)	—[c]	—	—
Chiang et al [40] (2017, China)	RCT	—	74 (39/35), 62.2	—	• Children 32 (43.2) • Parent 21 (28.4) • Spouse 9 (12.2) • Sibling 5 (6.8) • Other 7 (9.5)	—	• Married 45 (60.8) • Single 22 (29.7) • Divorced 7 (9.5)	• Below university 46 (62.2) • University or above 26 (35.1)	• High 32 (43.2) • Low 31 (41.9)
Cox et al [42] (2019, United States)	RCT	• Medical 50.9% • Neurological 22% • Surgical 13.4% • Trauma 11.9% • Cardiology 1.8%	275 (137/138), 73.1	51.20 (12.60)	• Spouse 128 (46.5) • Children 65 (23.6) • Parent 54 (19.6) • Sibling 22 (8) • Other 6 (2.2)	• Employed 184 (66.9) • Retired 54 (19.6) • Disabled 18 (6.5) • Unemployed 17 (6.2) • Other 2 (0.7)	• Married 210 (76.4) • Single 28 (10.2) • Divorced 37 (13.4)	• Below university 157 (57.1) • University or above 118 (42.9)	• Low 130 (42.3) • No 80 (29.1) • Moderate 45 (16.4) • High 20 (7.3)
Cox et al [41] (2024, United States)	Cluster RCT	• Medical 51.6% • Neurological 22% • Shock 14% • Trauma 13%	111 (55/56), 83	51.00 (15.00)	• Spouse 50 (45) • Parent 23 (20.7) • Children 22 (19.8) • Sibling 11 (9.9) • Other 5 (4.5)	• Employed 81 (73) • Retired 19 (17.1) • Disabled 7 (6.3) • Unemployed 4 (3.6)	• Married 67 (60.4) • Single 21 (18.9) • Divorced 23 (20.7)	• Below university 20 (18) • University or above 91 (82)	• No 45 (40.5) • Low 44 (39.6) • Moderate 13 (11.7) • High 9 (8.1)
Cox et al [52] (2025, United States)	RCT	• Medical 37.1% • Neurological 27.8% • Cardiac 24.5% • Surgical 8% • Community medical-surgical 2.7%	151 (76/75), 72.9	57.40 (12.90)	• Spouse 80 (60.6) • Child 36 (27.3) • Parent 12 (9.1) • Sibling 2 (1.5) • Other 2 (1.5)	• Employed 82 (62.1) • Retired 36 (27.3) • Disabled 11 (8.3) • Unemployed 3 (2.3)	• Married 107 (81.1) • Single 13 (9.9) • Divorced 12 (9.1)	—	—
Yuan et al [50] (2023, China)	RCT	• Surgical 100%	98 (50/48)	—	• Children 66 (67.3) • Spouse 24 (24.5)	—	—	—	—

XSL•FO
RenderX

Study (year, country)	Design	Patient's diagnosis	Sample size, n (I/C[a]), % of female	Age (y), mean (SD)	Relationship with patients, n (%)	Occupation, n (%)	Marital status, n (%)	Education level, n (%)	Financial burden, n (%)
Drop et al [43] (2025, Netherlands)	Cluster RCT	• Medical 60.9% • Surgical 31.7% • Trauma 7.5%	189 (100/89), 53.4	47.35 (12.36)	• Children 71 (37.6) • Spouse 62 (32.8) • Sibling 11 (5.8) • Parent 9 (4.8) • Other 7 (3.7)	—	—	• Below university 142 (75.1) • University or above 18 (9.5)	—
Gawlytta et al [44] (2022, Germany)	RCT	• Severe sepsis 100%	25 (12/13), 64	54.00 (11.01)	• Spouse 25 (100)	—	• Married 21 (84)	• Below university 15 (60) • University or above 10 (40)	—
Hoffmann et al [14] (2023, Austria)	RCT	• Nonsurgical 69.7% • Surgical 30.3%	89 (46/43), 66.3	47.00 (13.00)	• Spouse 36 (40.4) • Children 31 (34.8) • Sibling 9 (10.1) • Parent 8 (9) • Other 5 (5.6)	• Employed 67 (75.3) • Unemployed 22 (24.7)	—	• Below university 55 (61.8) • University or above 34 (38.2)	—
Petrinec et al [45] (2023, United States)	Pilot RCT	• Infectious disease 31.7% • Neurological 13.3% • Cardiovascular 11.7% • Gastrointestinal 8.3% • Trauma 8.3%	60 (30/30), 61.7	46.39 (13.04)	• Children 23 (38.3) • Spouse 16 (26.7) • Parent 8 (13.3) • Sibling 7 (11.7) • Other 6 (10)	—	—	—	—
Shariati et al [46] (2021, Iran)	RCT	• Patient with COVID-19	67 (33/34), 46.3	38.68 (11.18)	• Children 37 (55.2) • Spouse 21 (31.3) • Parent 4 (6) • Sibling 1 (1.5) • Other 4 (6)	• Employed 62 (92.5) • Retired 2 (3) • Student 3 (4.5)	• Married 41 (61.2) • Single 26 (38.8)	• Below university 4 (6) • University or above 63 (94)	—
Shin et al [26] (2025, United States)	Pilot RCT	• Respiratory 42.9% • Cardiovascular 21.4% • Sepsis 17.9% • Surgical 17.9%	28 (14/14), 85.7	50.93 (16.05)	• Spouse 19 (67.9) • Parent 4 (14.3) • Sibling 3 (10.7) • Children 2 (7.1)	• Employed 12 (42.9) • Unemployed 6 (21.4) • Retired 8 (28.6) • Disabled 2 (7.1)	—	• Below university 14 (50) • University or above 14 (50)	—

XSL•FO
RenderX

Study (year, country)	Design	Patient's diagnosis	Sample size, n (I/C[a]), % of female	Age (y), mean (SD)	Relationship with patients, n (%)	Occupation, n (%)	Marital status, n (%)	Education level, n (%)	Financial burden, n (%)
Suen et al [47] (2021, United States)	Pilot RCT	• Medical 100%	48 (23/25), 62.5	56.37 (12.95)	• Spouse 16 (33.3) • Children 16 (33.3) • Sibling 4 (8.3) • Parent 4 (8.3) • Other 8 (16.7)	—	—	• Below university 24 (50) • University or above 24 (50)	—
Vranceanu et al [53] (2020, United States)	Pilot RCT	• Neurological 100%	58 (29/29), 67.2	52.25 (14.48)	• Spouse 46 (79.3) • Parent 7 (12.1) • Sibling 2 (3.4) • Other 3 (5.2)	• Employed 41 (70.7) • Retired 10 (17.2) • Other 7 (12.1)	—	• Below university 29 (50) • University or above 29 (50)	—
Watland et al [48] (2025, Norway)	RCT	• Respiratory 37.7% • Cardiovascular 20.4% • Accident 18.4% • Neurological 13.3% • Sepsis 12.7%	196 (101/95), 64.8	46.80 (13.51)	• Children 79 (40.3) • Spouse 45 (23) • Parent 32 (16.3) • Sibling 26 (13.3) • Other 14 (7.1)	—	—	• Below university 73 (37.2) • University or above 123 (62.8)	—
Woo et al [49] (2024, Korea)	RCT	• Surgery 71.1% • Shock 18.4% • Multiple traumas 7.9% • Respiratory 2.6%	38 (18/20), 68.4	51.60 (11.80)	• Spouse 20 (52.6) • Children 8 (21.1) • Parent 5 (13.2) • Sibling 4 (10.5) • Other 1 (2.6)	—	—	—	—
Xiong et al [16] (2025, China)	RCT	• Patients who underwent HVR[d]	100 (50/50), 61	—	• Children 48 (48) • Spouse 42 (42) • Parent 8 (8) • Sibling 1 (1) • Other 1 (1)	• Employed 92 (92) • Retired 8 (8)	• Married 93 (93) • Not married 7 (7)	• Below university 75 (75) • University or above 25 (25)	• Moderate 53 (53) • Low 41 (41) • High 6 (6)

[a]I/C: intervention/control.

[b]RCT: randomized controlled trial.

[c]Not applicable.

[d]HVR: heart valve replacement.

Participants

The sample sizes ranged from 25 to 275, yielding a total of 1864 participants (intervention: 942, 50.5% and control: 922, 49.5%). Regarding patient types, only 1 study [40] did not report; 11 studies [14,26,41-48,52] focused on medical patients; and 5 studies [16,49-51,53] focused on surgical patients. Furthermore, 7 studies [41,42,44,49,50,52,53] recruited not only family

XSL•FO
RenderX

members but also ICU patients or ICU health care staff, whereas the remaining 10 studies recruited only family members. In addition, 14 studies reported the mean ages of family members, ranging from 38.68 (SD 11.18) years to 57.40 (SD 12.90) years. Except for the study by Yuan et al [50], all studies reported that 46.3% to 85.7% of participants were female. Concerning the relationship to patients, 7 studies [16,40,43,45,46,48,50] reported that the main participants were children, and 10 studies reported that they were spouses. Moreover, 9 studies [14,16,26,41,42,46,51-53] reported participants' occupation, indicating most of them were employed; 10 studies [14,26,43,45,47-51,53] did not report marital status; the other 7 did, and the majority of participants were married. Regarding educational level, 5 studies [45,49-52] did not report it; 8 reported that most participants had nontertiary education; and 4 reported that most participants had university education or higher. Financial burden was not reported in 13 studies; 2 studies [41,42] reported no or low burden, and 2 studies [16,40] reported moderate or high burden (Table 1).

Intervention

In the included studies, various types of DHT were used. In total, 9 studies [16,26,40,41,45,49-53] used mobile apps, 5 studies [14,42,44,46,47] used internet-based technology, and 3 used other technologies (1 was a video tool [49], 1 was a digital assessment tool [48], and 1 was VR technology [43]). Furthermore, 10 studies [41-46,49,51-53] reported the specific duration of the intervention, and the remaining studies indicated that family members could use digital intervention tools during the ICU or within months after discharge. Regarding data collection timing, in addition to pre- and postintervention assessments, 6 studies [14,16,26,43,45,51] collected data at 1 month post intervention; 1 study [45] at 2 months post enrollment; 10 studies [14,26,41-43,47,48,51-53] at 3 months post enrollment or discharge; 4 studies [26,42,43,51] at 6 months post discharge; and 2 studies [14,51] at 12 months post admission. No adverse outcomes or events were reported across studies (Table 2).

XSL•FO
RenderX

Table 2. Characteristics of the outcomes and measurements in the included studies.

Study (year, country)	Control	Experimental				Outcomes and measurements	Results
		Intervention measurement	Intervention delivery	Intervention duration	Data collection time points		
Carlozzi et al [51] (2025, United States)	Self-monitoring only	Self-monitoring plus personalized self-care from the CareQOL app	Mobile app	10-day run-in period followed by a 6-month home monitoring period	baseline; 1-12 months	• Caregiver appraisal: CAS[a] • Adaptability: MPAI-4[b] • PTSD[c]: PCL-5[d] • Caregiver-specific anxiety: PROMIS[e] • Sleep-related impairment: PROMIS • Sepression: PROMIS • Fatigue: PROMIS • Perceived stress: PROMIS • Self-efficacy general: PROMIS • Overall health: PROMIS	• Although we did not observe improvements in HRQOL[f], physical activity, or sleep, we found that across all measures, approximately one-third of participants showed clinically meaningful improvements, one-third remained the same, and one-third worsened. • Care partners who reported engaging in the intervention were more likely to show improvement than those who did not. • There was preliminary support for factors that being male, caring for a person with post-traumatic stress symptoms, living in the same household as the person with TBI[g], being a spousal care partner, working, and being diagnosed with COVID-19 during the study were associated with increased risk for adverse outcomes.
Chiang et al [40] (2017, China)	EF-R[h]	EF-T[i]	Interactive mobile technology	During ICU[j]	Baselines; the posttest one day after EF-T or EF-R	• Depression: C-DASS[k] • Anxiety: C-DASS • Stress: C-DASS • Satisfaction: CPCS[l]	• The EF-R group had a slightly lower stress score than the EF-T group at posttest, but the difference was statistically insignificant. • There was also no significant improvement in anxiety after EF-T. • The EF-T group had a significantly better reduction of depression score than the EF-R group. • Information need satisfaction was not significantly different between intervention and control groups.

XSL•FO
RenderX

Study (year, country)	Control	Experimental				Outcomes and measurements	Results
		Intervention measurement	Intervention delivery	Intervention duration	Data collection time points		
Cox et al [42] (2019, United States)	Usual care practices followed by a family meeting	A web-based decision aid	Web-based	2 days	In person at enrollment but before randomization; on study day 3; via telephone at 3 months and 6 months	• Clinician surrogate concordance: CSCS[m] • Comprehension of diagnosis, treatment, and prognosis: MCS[n] • Clinician communication satisfaction: QOC[o] • Anxiety: HADS[p] • Depression: HADS • PTSD: PTSS[q] • Decision-making uncertainty: DCS[r] • Patient-centeredness of care: PPPC[s]	• Concordance improvement did not differ between intervention and control groups. • Intervention surrogates had a greater reduction in decisional conflict than control surrogates. • Intervention and control primary surrogates did not differ in medical comprehension or communication. • Surrogates' total scores on anxiety, depression, and PTSD also did not differ at 3 and 6 months.
Cox et al [41] (2024, United States)	Usual ICU care without protocolized family meetings	ICUconnect uses text messages and emails	Mobile app	7-10 days	Baselines; study days 1, 3, and 7; 3 months	• Needs at the End-of-Life Screening Tool: NEST[t] • Goal concordance of care • Quality of communication: QOC • Patient-centeredness of decision-making, eliciting concerns, discrimination: IPC[u] • Depression: PHQ-9[v] • Anxiety: GAD-7[w] • PTSD: PTSS	• There was greater improvement in NEST scores among intervention recipients between baseline and both day 3 and day 7. • There were no treatment group differences at 3 months in psychological distress symptoms. • Change from baseline to study day 3 was not statistically significantly different between intervention and control groups in IPC-Elicited Concerns, and IPC-Decision Making subscales, as well as QOC scores. • Although intervention family members reported a numerically greater improvement in goal-concordant care compared with control family members, this finding was not statistically significant. • White family members experienced a greater reduction in NEST scores compared with Black family members at day 3 and day 7.
		Intervention measurement	Intervention delivery	Intervention duration	Data collection time points		

XSL•FO
RenderX

Study (year, country)	Control	Experimental				Outcomes and measurements	Results
		Intervention measurement	Intervention delivery	Intervention duration	Data collection time points		
Cox et al [52] (2025, United States)	Usual ICU care	PC planner	Mobile app-based platform	7 days	Baseline; day1; day3; day 7; and 3 month	• Needs at the End-of-Life: NEST • Goal concordant of care: GCC[x] • Quality of communication: QOC • Patient-centeredness: PPPC • Depression: PHQ-9 • Anxiety: GAD-7 • PTSD: PTSS	• Treatment group differences in estimated mean NEST scores were similar at 3 days and at 7 days. • Goal concordance, communication quality, and psychological distress symptoms did not differ. • 29 (38.2%) intervention participants had palliative care consultations, compared with only 3 (4%) among controls (P<.001).
Yuan et al [50] (2023, China)	Routine care	Video visitation	WeChat video call	During the ICU, every afternoon	Admission to the ICU, and discharge from the ICU	• Anxiety: SAS[y] • Satisfaction: questionnaire in the hospital	• There were no statistically significant differences between the groups in patients' anxiety and depression scores, and no statistically significant differences in delirium incidence between the groups. • There were no significant differences in changes in family members' anxiety scores. • A statistically significant difference in satisfaction was found between the two groups of patients, and the result of family members' satisfaction was also statistically significant.
Drop et al [43] (2025, Netherlands)	Standard care	Standard care plus ICU-VR[z]	VR	14 min	Immediately after enrollment at ICU; admission at ICU; discharge at 1, 3, and 6 months post-ICU discharge	• PTSD: IES-R[aa] • Anxiety: HADS • Depression: HADS • Quality of life: SF-36[ab] • Relatives' understanding: CQI[ac]-relatives ICU • Perceived stress and appreciation of ICU-VR: family satisfaction with ICU care tools	• Relatives who received ICU-VR did not experience a decrease in PTSD, anxiety, or depression. • There was no significant difference between the median mental quality of life, physical quality of life, or understanding of ICU care. • Patients in the intervention group highly endorsed ICU-VR, favoring it over traditional informational brochures, and the majority stated it improved their understanding of ICU treatment.
Gawlytta et al [44] (2022, Germany)	5-week waiting period	iCBT[ad]: 10 writing assignments	Internet-based	Two 50-minute sessions per week over 5 weeks	Start of intervention or waiting; end of intervention or waiting; end of intervention in WL[ae] control group		

XSL•FO
RenderX

Study (year, country)	Control	Experimental				Outcomes and measurements	Results
		Intervention measurement	Intervention delivery	Intervention duration	Data collection time points		
						• PTSD: PCL-5 • Depression: BSI-18[af] • Anxiety: BSI-18 • Somatization: BSI-18 • Relationship satisfaction: RAS[ag] • Health-related quality of life: EQ-5D-5L (five-level EuroQol five-dimensional questionnaire)	• There was no evidence of a difference in PCL-5 pre-post change between iCBT and WL. • iCBT led to a larger RAS change than waiting, with effects in favor of waiting. • For all other secondary efficacy outcomes, we did not observe evidence for an association between score changes and iCBT in the primary analysis set.
Hoffmann et al [14] (2023, Austria)	Control website	Intervention website	Online	Not reported	Within 2 days of the patient's admission to the ICU; 30 days after T0; 90, and 365 days after admission	• PTSD: IES • Anxiety: HADS • Depression: HADS • Frequency and times of website use	• An informative website did not result in reduced PTSD and depression symptoms.
Petrinec et al [45] (2023, United States)	Usual care	A mental health app (Sanvello)	Smartphone app	Each module contains 6 to 9 lessons (15 min per lesson)	Upon enrollment (within 5 days of ICU admission); 30 and 60 days after enrollment	• Anxiety: HADS • Depression: HADS • PTSD: PCL-5 • Quality of life: SF-12[ah] • self-efficacy: MHSES[ai] • Use of the app • Satisfaction with the app • Selection of the app	• Anxiety and depression symptom severity decreased significantly over time in the intervention group but not in the control group. • Family members logged in to the app a mean of 11.4 times and spent a mean of 50.16 minutes using the app. • Participants' mean rating of their satisfaction with the mental health app was 4.19 out of 5.
Shariati et al [46] (2021, Iran)	All common interventions except web-based communication	Web-based communication, audio or video (WhatsApp)	Web-based	4 consecutive days for 10 to 15 minutes per day	Before intervention; within 24 hours after the intervention in both groups	• Stress: PSS[aj]	• After the intervention, the mean PSS-14 in the intervention group was significantly lower than that of the control group.
Shin et al [26] (2025, United States)	Standard Android tablet loaded with MyChart	The tablet also contained the MyChart Bedside app and selected games as standard in the hospital	Tablet computer app	During ICU	Baseline; extubation or ICU discharge; 1, 3, and 6 months follow-up	• Anxiety: HADS • Depression: HADS • PTSD: IES-R • Communication difficulty: FCS[ak]	

XSL•FO
RenderX

Study (year, country)	Control	Experimental				Outcomes and measurements	Results
		Intervention measurement	Intervention delivery	Intervention duration	Data collection time points		
							• No statistically significant difference was found between groups in changes in family psychological outcomes. The VidaTalk was associated with a slight to medium improvement in anxiety symptoms at 1 month. • The VidaTalk group had lower PTSD-related symptoms than the AC[al] group with a medium effect size at 1 month and a medium-to-large effect size at 3 months. • Compared with the perceived communication difficulty, although not statistically significant, the mean change in FCS score was larger for the intervention group than for AC, with a small effect size.
Suen et al [47] (2021, United States)	iPad without the tool	iPad preloaded with the Family Support Tool	Interactive website	During ICU	Baseline; 48 h of study enrollment; 5-9 days of study enrollment; 3 months	• Tool usability and acceptability • Perceived effectiveness: internally generated 11-item questionnaire • Quality of communication: QOC	• Surrogates reported that the tool was highly usable, acceptable, and effective. • Surrogates who used the tool reported higher overall quality of communication and higher quality in shared decision-making, but the difference did not reach statistical significance.
Vranceanu et al [53] (2020, United States)	Mimic the dose and duration of RT[am] without teaching RT skills	RT is an active intervention. The first 2 sessions were taught at the bedside, and the next 4 were taught after discharge	Web-based app on a smartphone	30 minutes per session	Baseline; 6 weeks; and 12 weeks	• Feasibility of recruitment and intervention delivery, credibility, and satisfaction • Depression and anxiety: HADS • PTSD: PCL-C • Mindfulness: CAMS-R[an] • Coping: MOCS-A[ao] • Dyadic interactions: DRS[ap]	• Feasibility, adherence, fidelity, satisfaction, credibility, and expectancy exceeded benchmarks set a priori. • Participation in RT was associated with statistically and clinically significant improvement between baseline and postintervention in symptoms of depression, anxiety, and PTSD. • Improvements sustained through the 12-week follow-up visit. • RT-dependent improvement in dyadic interpersonal interactions for survivors.
Watland et al [48] (2025, Norway)	Standard care	The Caregiver Pathway 4-step model	Digital assessment tool	From admission to 3 months after discharge	Baseline; 3 months	• PTSD: IES-R • Anxiety: HADS • Depression: HADS • Quality of life: SF-12 • Self-efficacy: GSE[aq] • Hope: Herth Hope Index	• The Caregiver Pathway is associated with reduced symptoms of PICS-F[ar] in terms of reduced symptoms of PTSD, anxiety, and depression for caregivers of surviving patients.

XSL•FO
RenderX

Study (year, country)	Control	Experimental				Outcomes and measurements	Results
		Intervention measurement	Intervention delivery	Intervention duration	Data collection time points		
Woo et al [49] (2024, Korea)	Daily phone calls	Daily virtual visits	Zoom app	Began on the first day after ICU admission until ICU discharge, lasting for a maximum of 7 days (each limited to 15 min)	At study enrollment after ICU admission; at the end of the study	• Satisfaction concerning ICU care and decision-making processes: FS-ICU 24[as] • Anxiety: HADS • Depression: HADS	• The FS-ICU 24 survey score was significantly higher in the virtual visitation group. • The reduction in HADS-A and HADS-D scores between the 2 time points was significantly larger in the virtual visitation group.
Xiong et al [16] (2025, China)	Standard care	Wab-WPPEP	WeChat applet	Patient's admission to the ICU until 1 month after discharge	Preoperative baseline; before transfer from the ICU; before discharge from the hospital; 1 month after discharge	• Anxiety: HADS • Depression: HADS • PTSD: IES-R • Quality of life: SF-36	• The intervention group exhibited greater improvements in anxiety, depression, and quality of life compared with the control group. • PTSD scores were also significantly lower in the intervention group.

[a]CAS: caregiver appraisal scale.
[b]MPAI-4: Mayo-Portland Adaptability Inventory-Fourth Edition.
[c]PTSD: posttraumatic stress disorder.
[d]PCL-5: PTSD (posttraumatic stress disorder) Checklist for DSM-5 (Diagnostic and Statistical Manual of Mental Disorders, [Fifth Edition]).
[e]PROMIS: Patient-Reported Outcomes Measurement Information System.
[f]HRQOL: health-related quality of life.
[g]TBI: traumatic brain injury.
[h]EF-R: education of families by routine.
[i]EF-T: education of families by interactive tabs.
[j]ICU: intensive care unit.
[k]C-DASS: Depression Anxiety Stress Scale-Chinese.
[l]CPCS: Society of Critical Care Medicine's Family Needs Assessment Questionnaire's Communication and Physical Comfort Scale.
[m]CSCS: clinician-surrogate concordance scale.
[n]MCS: medical comprehension scale.
[o]QOC: quality of communication questionnaire.
[p]HADS: Hospital Anxiety And Depression Scale.
[q]PTSS: posttraumatic stress symptom inventory.
[r]DCS: decision conflict scale.
[s]PPPC: Patient-Perceived Patient-Centeredness scale.
[t]NEST: Needs at the End-of-Life Screening Tool.
[u]IPC: interpersonal processes of care scales.
[v]PHQ-9: patient health questionnaire-9 scale.
[w]GAD-7: generalized anxiety disorder 7 scale.
[x]GCC: goal concordant of care.
[y]SAS: Self-Rating Anxiety Scale.
[z]VR: virtual reality.
[aa]IES-R: Impact of Event Scale-revised.
[ab]SF-36: 36-Item Short-Form Health Survey.
[ac]CQI: Consumer Quality Index.
[ad]iCBT: internet-based cognitive behavioral writing therapy.
[ae]WL: wait list.
[af]BSI-18: Brief Symptom Inventory-18.

XSL•FO
RenderX

[ag]RAS: Relationship Satisfaction Scale.
[ah]SF-12: 12-Item Short-Form General Health Survey.
[ai]MHSES: Mental Health Self-Efficacy Scale.
[aj]PSS: Perceived Stress Scale.
[ak]FCS: Family Communication Scale.
[al]AC: attention control.
[am]RT: Recovering Together.
[an]CAMS-R: Cognitive and Affective Mindfulness Scale-Revised.
[ao]MOCS-A: Measure of Current Status-Part A.
[ap]DRS: Dyadic Relationship Scale.
[aq]GSE: General Self-Efficacy Scale.
[ar]PICS-F: Post-Intensive Care Syndrome Family.
[as]FS-ICU: Family Satisfaction-Intensive Care Unit Questionnaire.

Control

Control groups varied in different studies; 12 studies [16,40-43,45-50,52] adapted usual or routine care, including regular education on the ICU environment, basic intensive care processes, the use of monitoring devices, ICU rules, spontaneous conversations with the ICU nurses and meetings with the physicians, effective communication with critically ill patients, and basic principles of patient care. In addition, 4 studies [14,26,51,53] used a standard app or website that provided only basic information, and 1 study [44] was a blank control (Table 2).

Outcomes and Assessment Tools

Table 2 summarizes the outcomes and measurement tools of each study. All the assessment scales used in the outcome measurement were validated. In total, 5 scales were used to assess the anxiety and depression levels of ICU patients' families. Of the total, 9 of the included studies [14,16,26,42,43,45,48,49,53] adopted the Hospital Anxiety and Depression Scale to assess anxiety and depression. Moreover, 2 studies [41,52] assessed anxiety using the Generalized Anxiety Disorder 7 scale and depression using the Patient Health Questionnaire-9 Scale. Another study [40] assessed anxiety and depression using the Depression Anxiety Stress Scale-Chinese. One study [51] used the Patient-Reported Outcomes Measurement Information System questionnaire. One study [50] used the Self-Rating Anxiety Scale to assess anxiety. One study [44] used the Brief Symptom Inventory-18. The measurement tools used for PTSD varied among studies. In total, 3 studies [41,42,52] measured the Post-Traumatic Stress Symptom Inventory, 3 studies [44,45,53] used the PTSD Checklist for *DSM-5* (Diagnostic and Statistical Manual of Mental Disorders [Fifth Edition]), and 5 studies [14,16,26,43,48] used the Impact of Event Scale-Revised. In addition, 5 studies measured QoL with 3 scales. 1 study [44] measured the Health Questionnaire of the EQ-5D-5L; 2 studies [45,48] measured the 12-Item Short-Form Health Survey, and 2 studies [16,43] measured the 36-Item Short-Form Health Survey.

Risk of Bias

Two authors (JZ and ZH) used the Cochrane Risk of Bias Assessment Tool version 2 independently to assess the methodological quality of the included studies, and the weighted κ coefficients of domains ranged between 0.55 (P=.002) and 1.00 (P<.001), indicating substantial to almost perfect interrater agreement of the independent work in this part. The results are presented in Figure 2 [14,16,26,40-53]. Specifically, all included studies were reported as randomized, but selective bias was evident because 3 studies [45,48,53] did not report allocation concealment. Due to the nature of DHIs, blinding of participants and investigators was not possible; only 1 study [14] used a double-blind design for the intervention because the structure was the same across both websites. However, almost all studies reported using intention-to-treat analysis to estimate the effect of intervention assignment, thereby reducing bias. A total of 5 studies [40,42,45,50,53] reported deviations from the intended intervention arising from the trial context, and are therefore rated as high-bias. In addition, 3 studies [26,43,52] reported incomplete data, which increased the risk of attrition bias. Furthermore, 12 studies were blinded to outcome assessors, and the other 5 studies [26,44,48,49,53] addressed bias in outcome measurement. Moreover, 2 studies [14,45] provided insufficient information to permit judgment and might have led to overestimated effects and introduced selection bias into the reported results. Finally, 4 studies were assessed as "low-risk bias," 10 as "some concerns," and 3 as "high-risk bias." However, in accordance with the consensus agreement of all the authors, the studies with a high risk of overall bias were ultimately reserved.

XSL•FO
RenderX

Figure 2. The results of the risk of bias assessment of included studies. (A) Risk of bias summary, (B) risk of bias graph [14,16,26,40-53].

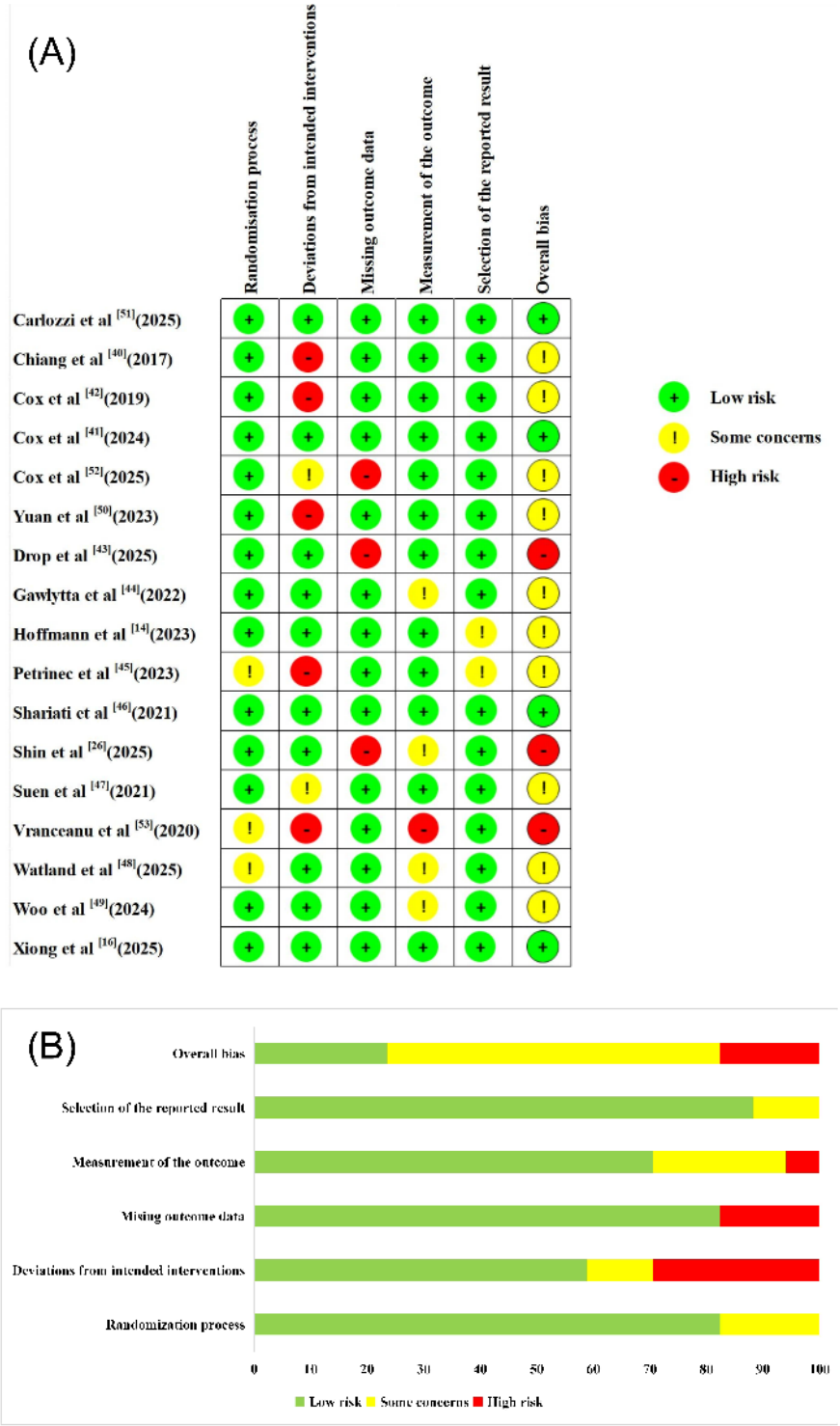

Quality of Evidence

The GRADE approach [33] was used to assess the overall certainty of the evidence. The κ coefficient in this part was 0.88 (P<.001), indicating almost perfect interrater consistency. Among the 7 outcomes analyzed, the certainty of evidence was downgraded to moderate for QoL and quality of communication (QOC). The certainty for anxiety, depression, PTSD, and self-efficacy was downgraded to low. Stress was downgraded to very low certainty of evidence (Table 3).

XSL•FO
RenderX

Table 3. GRADE[a] evidence profile for digital health intervention compared with standard care for intensive care unit family members.

Outcomes	Follow-up		Certainty of the evidence[b] (GRADE)	Relative effect (95% CI)[c]	Anticipated absolute effects	
	Participants[d], n	Studies, n			Risk with standard care	Risk difference with digital health intervention
Anxiety	1209	12 RCTs[e]	• • Low[f,g,h]	—[i]	—	SMD[j] 0.34 SD lower (0.68 lower to 0)
Depression	1181	12 RCTs	• • Low[f,g,h]	—	—	SMD 0.26 SD lower (0.52 lower to 0.01 higher)
PTSD[k]	1077	11 RCTs	• • Low[f,g,h]	—	—	SMD 0.21 SD lower (0.49 lower to 0.06 higher)
Stress	141	2 RCTs	• • Very low[g,h,l,m]	—	—	SMD 1.17 lower (3.53 lower to 1.19 higher)
Self-efficacy	204	2 RCTs	• • Lowf,h,n,o	—	—	SMD 0.01 lower (0.28 lower to 0.27 higher)
QoL[p]	424	5 RCTs	• • Moderate[f,h,n]	—	—	SMD 0.09 SD higher (0.17 lower to 0.35 higher)
Quality of communication	555	4 RCTs	• • Moderate[f,h,n]	—	—	SMD 0.15 SD higher (0.06 lower to 0.35 higher)

[a]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.

[b]High certainty (we are very confident that the true effect lies close to that of the estimate of the effect); moderate certainty (we are moderately confident in the effect estimate: the true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different); low certainty (our confidence in the effect estimate is limited: the true effect may be substantially different from the estimate of the effect); and very low certainty (we have very little confidence in the effect estimate: the true effect is likely to be substantially different from the estimate of effect).

[c]The risk in the intervention group (and its 95% CI) is based on the assumed risk in the comparison group and the relative effect of the intervention (and its 95% CI).

[d]Patient or population: intensive care unit family members; setting: intensive care unit; intervention: digital health intervention; and comparison: standard care.

[e]RCT: randomized controlled trial.

[f]Most of the included studies were assessed as some concerns or high-risk bias.

[g]$I^2 \geq 50\%$.

[h]Direct participants, interventions and outcomes.

[i]Not applicable.

[j]SMD: standardized mean difference.

[k]PTSD: posttraumatic stress disorder.

[l]Most of the included studies were assessed as low-risk or some concerns bias.

[m]Sample size is small and CI was wide.

[n]$I^2 < 50\%$.

[o]CI was wide.

[p]QoL: quality of life.

Effect of DHIs

Psychological Outcomes

In total, 12 studies assessed the impact of DHIs on anxiety among ICU family members. The pooled results showed that DHIs did not lead to a statistically significant reduction in anxiety levels (SMD –0.34, 95% CI –0.68 to 0.00; P=.05). The heterogeneity among studies was substantial (I^2=80.19%; τ=0.49; τ^2=0.24), and the certainty of the evidence was rated as low. The 95% PI of the effect of DHIs on anxiety was –1.46 to 0.79, which indicates that we can be 95% confident that the true effect size from a future individual study would fall within this range (Figure 3 [16,26,40-43,45,48-50,52,53]). Similarly, 12 studies investigated the impact of DHIs on depression among ICU family members. The results revealed no significant effect (SMD –0.26, 95% CI –0.52 to 0.01; P=.06), with moderate heterogeneity (I^2=67.80%; τ=0.37; τ^2=0.13). The certainty of this evidence was also rated as low. A wide 95% PI (–1.11 to 0.60) suggests that future studies could show effects ranging

XSL•FO
RenderX

from a large reduction to a modest increase in depression levels (Figure 4 [14,16,26,40-43,45,48,49,52,53]). In total, 11 studies evaluated the effects of DHIs on PTSD among ICU family members. The meta-analysis demonstrated no significant difference in PTSD scores between groups (SMD –0.21, 95% CI –0.49 to 0.06; *P*=.11), with substantial heterogeneity (I^2=69.98%; τ=0.34; τ^2=0.12). The certainty of this evidence was also rated as low. The 95% PI (–1.03 to 0.60) indicates that the actual effect of DHIs on PTSD could vary widely, indicating potential for both beneficial and harmful effects (Figure 5 [14,16,26,41-45,48,52,53]). Furthermore, regarding stress and self-efficacy, which were evaluated in only 2 studies each, we discussed them qualitatively. 2 studies reported different results regarding stress: 1 study [40] found no effect of DHI on stress. In contrast, another [46] reported a significant effect of DHI in reducing stress among ICU family members. Neither study [45,48] on self-efficacy found a significant benefit from DHIs. The certainty of the 2 pieces of evidence was rated as very low and low.

Figure 3. Forest plot of the effectiveness of digital health technology on anxiety. SMD: standardized mean difference [16,26,40-43,45,48-50,52,53].

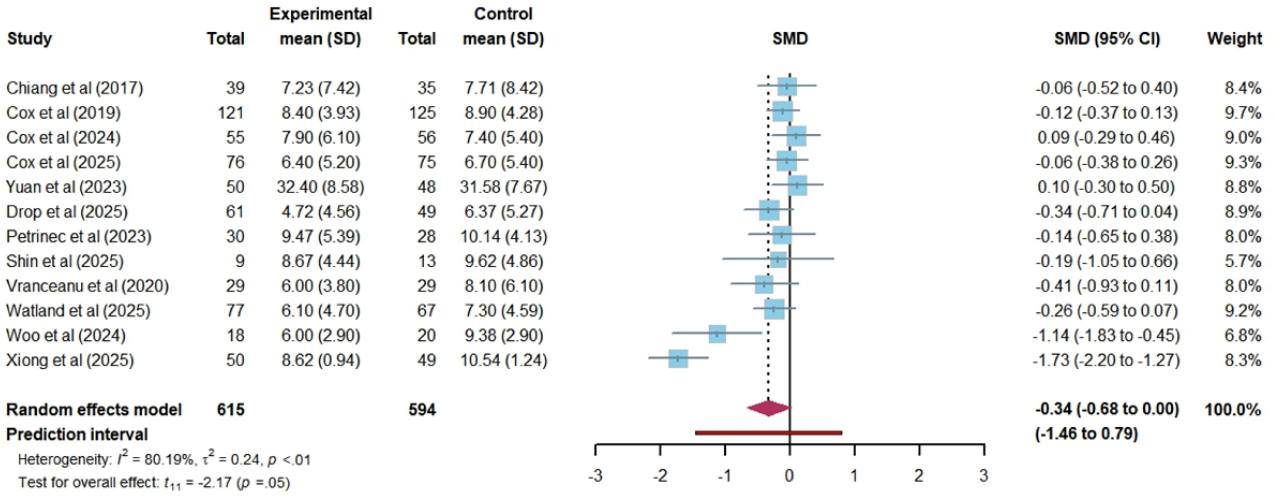

Figure 4. Forest plot of the effectiveness of digital health technology on depression. SMD: standardized mean difference [14,16,26,40-43,45,48,49,52,53].

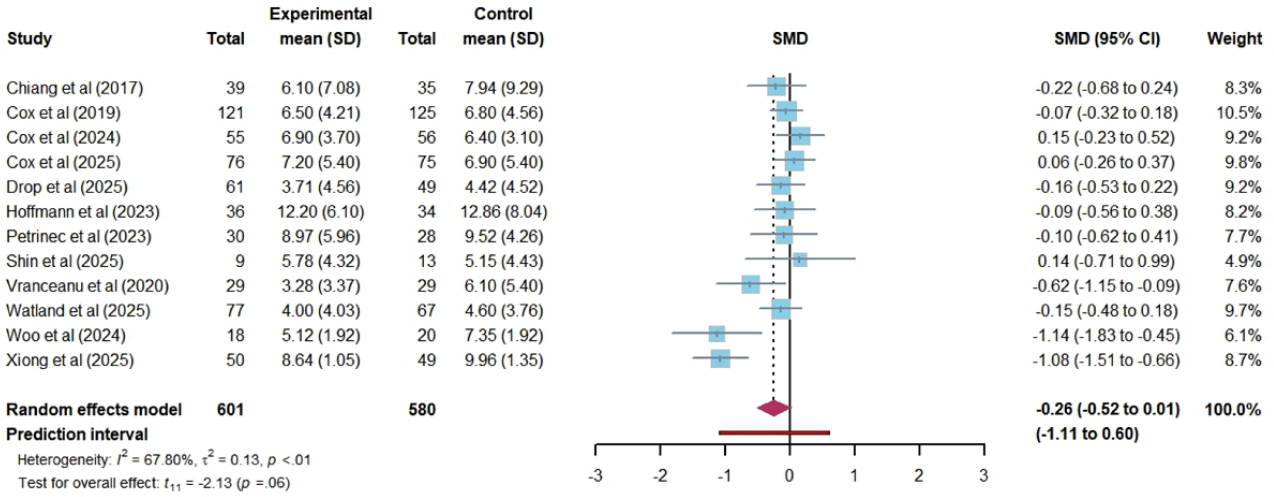

XSL•FO
RenderX

Figure 5. Forest plot of the effectiveness of digital health technology on post-traumatic stress disorder. SMD: standardized mean difference [14,16,26,41-45,48,52,53].

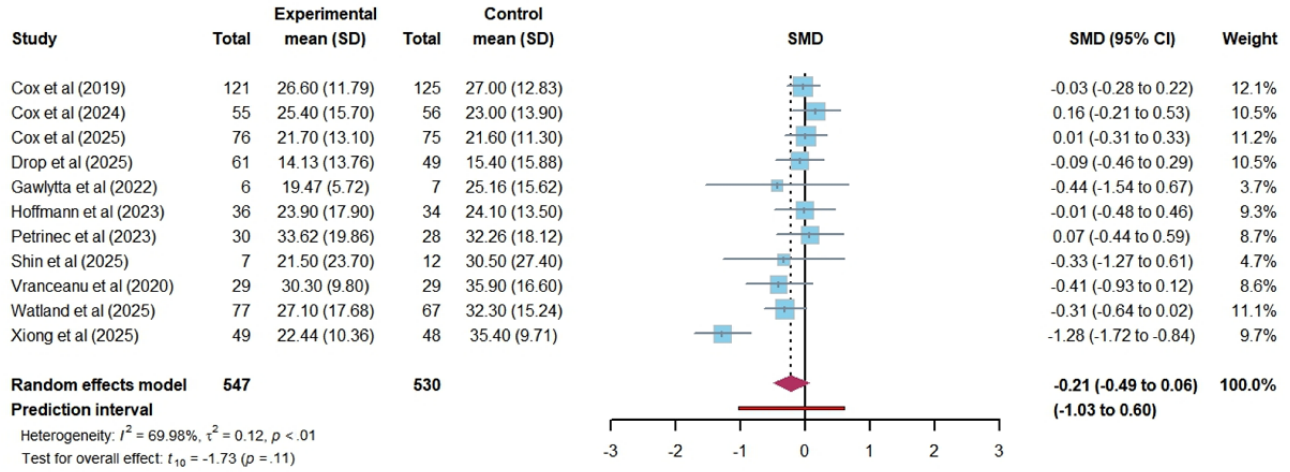

QoL

In total, 5 studies assessed the impact of DHIs on QoL among ICU family members. A random-effect meta-analysis indicated no statistically significant improvement in QoL (SMD 0.09, 95% CI –0.10 to 0.28; *P*=.36). Despite the absence of heterogeneity (I^2=0.00%; τ=0.14; τ^2=0.02), the actual effect of DHIs on QoL in the future falls within the broader PI range (–0.41 to 0.60), suggesting that future outcome could vary significantly across different populations and settings (Figure 6 [16,43-45,48]). The certainty of this evidence was rated as moderate.

Figure 6. Forest plot of the effectiveness of digital health technology on quality of life. SMD: standardized mean difference [16,43-45,48].

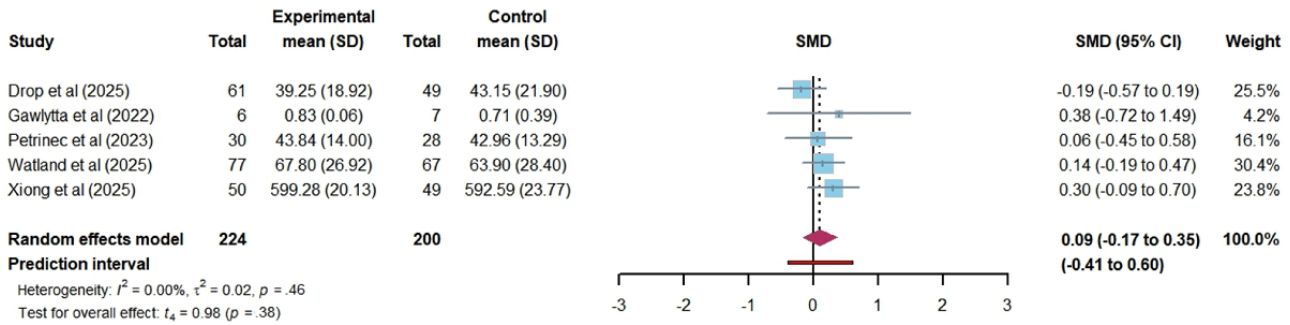

QOC

In total, 4 studies evaluated the effect of DHIs on the QOC among ICU family members. The pooled analysis showed no statistically significant difference between groups (SMD 0.14; 95% CI –0.03 to 0.31; *P*=.10; Figure 7 [41,42,47,52]). Despite the absence of heterogeneity (I^2=0.00%; τ=0.08; τ^2=0.01), the actual effect of DHIs on QoL in the future falls within the broader PI range (–0.26 to 0.55), suggesting that future outcomes could vary significantly across different populations and settings. The certainty of this evidence was rated as moderate.

XSL•FO
RenderX

Figure 7. Forest plot of the effectiveness of digital health technology on the quality of communication. SMD: standardized mean difference [41,42,47,52].

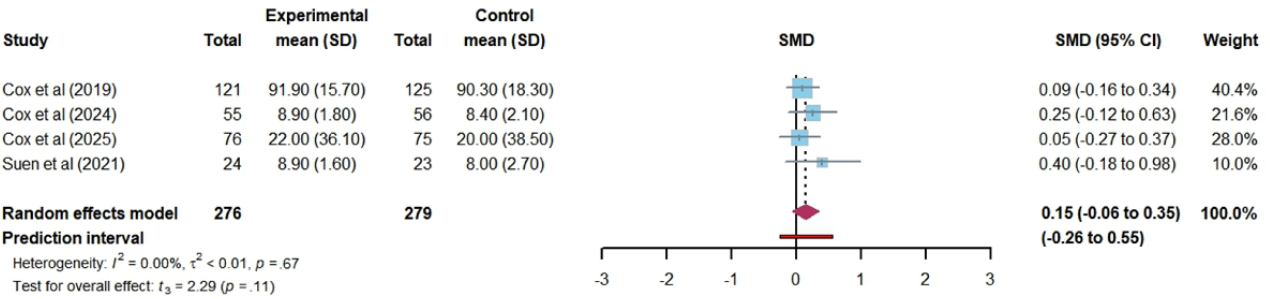

Subgroup Analysis

We conducted subgroup analyses for the three outcome indicators of anxiety, depression, and PTSD based on potential covariates, including patients' primary diagnosis, relationship with patients, and type of DHT (Table S1 and Figures S1-S3 in Multimedia Appendix 4). Regarding anxiety, the results of subgroup analyses showed that there was no significant difference (P>.05) in the patients' primary diagnosis (χ^2_2=2.60; P=.27), relationship with patients (χ^2_1=0.23; P=.63), and type of DHT (χ^2_2=1.43; P=.49), suggesting that these factors may not be related to the change of intervention effect. Within the subgroup, family members of medical patients showed a slight improvement trend (SMD –0.14, 95% CI –0.26 to –0.01; P=.04). From the perspective of heterogeneity sources, the patients' primary diagnosis and the type of DHT may be potential sources of heterogeneity. Regarding depression, subgroup analysis showed significant differences between subgroups with different patients' primary diagnosis (χ^2_2=29.17; P<.01). Among the surgical patients' family members (3 studies), depression scores in the intervention group were significantly reduced (SMD –0.94, 95% CI –1.64 to –0.24, t_2=–5.80; P=.03). In studies focusing on medical patients, no significant effect of the intervention on family members depressive symptoms was observed (SMD –0.04, 95% CI –0.14 to 0.05, t_7=–1.11; P=.30). There was no statistical significance (P>.05) in the subgroup differences in relationship with patients (χ^2_1=0.38; P=.54) and type of DHT (χ^2_2=3.11; P=.21). Among the subgroups of type of DHT, the effect size of the education-based DHT group (4 studies) had a more minor statistical significance (SMD –0.11, 95% CI –0.21 to –0.01, t_3=–3.47; P=.04). Patients' primary diagnosis and type of DHT may be significant sources of heterogeneity. Regarding PTSD, the differences between subgroups in the patients' primary diagnosis (χ^2_1=3.29; P=.07), relationship with patients (χ^2_1=1.30; P=.25), and type of DHT (χ^2_2=2.75; P=.25) did not reach statistical significance (P>.05). However, looking at the sources of heterogeneity, these factors may be the primary drivers. Due to the limited number of studies in some subgroups, these results should be interpreted with caution.

Sensitivity Analysis

For anxiety, it had no impact on the statistical significance of the results after removing the study conducted by Xiong et al [16] (SMD –0.18, 95% CI –0.38 to 0.01; P=.06; Figure S4 in Multimedia Appendix 4), which decreased heterogeneity to 27.36%. Regarding depression, removing the study by Xiong et al [16] (SMD –0.17, 95% CI –0.38 to 0.05; P=.11; Figure S5 in Multimedia Appendix 4) decreased heterogeneity to 36.09%, but it did not alter the statistical significance of the results. For PTSD, sensitivity analyses demonstrated that removing the study by Xiong et al [16] (SMD –0.08, 95% CI –0.21 to 0.05; Figure S6 in Multimedia Appendix 4) resolved heterogeneity, although it did not alter the statistical significance of the results. For QoL, statistical significance was observed only after removing the study by Drop et al [43] (SMD 0.19, 95% CI 0.01 to –0.37; P=.04; Figure S7 in Multimedia Appendix 4). Regarding QOC, the effect of DHIs remained stable across studies, indicating robust findings.

Bias Assessment

Visual inspection of the funnel plots for anxiety, depression, and PTSD showed approximate symmetry (Figures S1-S3 in Multimedia Appendix 5). Egger tests did not provide statistically significant evidence of small-study effects for anxiety (P=.24), depression (P=.20), and PTSD (P=.44; Table S1 in Multimedia Appendix 5). However, we note that small-study effects can arise from various sources beyond publication bias, such as heterogeneity or methodological differences. For other outcomes, there were insufficient numbers of systematic reviews (n<10) to create funnel plots.

Discussion

Principal Findings

In total, 17 studies were included, revealing that compared with control groups, DHIs may not confer a significant benefit in reducing anxiety, depression, PTSD, or stress among ICU family members. Moreover, they also did not notably improve

XSL•FO
RenderX

self-efficacy, QoL, or communication quality. Given the low certainty of the evidence for the proposed association, the substantial study heterogeneity, and the overall risk of bias in the included studies, the findings should be interpreted with caution. Subgroup analyses indicated that patients' primary diagnosis may influence the intervention's effect, with the benefits being more pronounced for families of surgical patients. Patients' primary diagnosis, relationship with patients, and type of DHT may be a significant source of heterogeneity. Given the PI value, adverse effects of the intervention cannot be ruled out in future studies, and more RCTs are needed to provide robust results.

The observed effect of DHIs on anxiety aligns with previous research concerning improvements in anxiety among caregivers of patients with cancer [23]. The ICU environment is characterized by high stress, restricted access, asymmetric information, and uncertain prognosis, which can significantly increase the psychological burden of family members [54,55]. DHI compensates for the absence of in-person support by using mobile apps, network platforms, and virtual visits [56]. It provides timely updates on diseases, health education, and emotional support for family members, thereby reducing uncertainty and equipping them with knowledge and skills that, in turn, alleviate anxiety symptoms [57]. However, the clinical significance of this finding is tempered by its statistical marginality (P=.05), a low certainty of the evidence, and risk of bias in some studies, which significantly reduces the credibility of this potential effect. It is critical to underscore that substantial heterogeneity (I^2=80.19%) yields a wide 95% PI (–1.46 to 0.79), suggesting that the distribution of effects on anxiety is unpredictable. This finding should only be considered a preliminary exploratory conclusion and requires further validation through additional research. No significant effects were observed on depression, PTSD, stress, self-efficacy, QoL, and QOC. It is important to note that the certainty of the evidence for these outcomes is low, with an associated risk of bias, so the reasons for the lack of significant observed effects should be considered from multiple perspectives. From a clinical view, the onset of PTSD is often delayed and complex, typically manifesting months to a year post discharge, making it challenging for short-term interventions and follow-ups to capture long-term effects [14,44]. Additionally, stress, self-efficacy, QOC, and QoL are inherently multidimensional indicators influenced by confounding factors, including economic conditions and social support, beyond psychological state [45,58,59]. From a methodological perspective, existing studies have limitations, including risk of bias, low certainty of the evidence, and imprecision—particularly given that only 2 studies were included for stress and self-efficacy. Current evidence is insufficient to identify the potential benefits or harms of DHIs for these outcomes, and subsequent research is very likely to alter these estimates.

The observed significant heterogeneity across outcomes prompted extensive subgroup analyses. While these analyses pointed to several potential moderators—such as patients' primary diagnosis and type of DHT for depression—their interpretability is limited. For example, the notable reduction in depression among surgical patients (SMD –0.94, 95% CI –1.64 to –0.24), derived from only 3 studies, may reflect a type I error or chance variation rather than an actual differential effect. Similarly, although traditional factors such as family members' sex and relationship with patients did not reach statistical significance in our analyses, existing literature consistently indicates that female caregivers, especially spouses, often endure greater caregiving stress and more persistent psychological distress [14,60,61]. The overall heterogeneity remained largely unexplained by subgrouping, further underscoring the robustness of these comparisons. Thus, these analyses should not be viewed as confirming specific moderators but rather as highlighting priority variables for future, adequately powered studies.

The limited effectiveness of DHIs may also be attributed to the design of the interventions. Most existing interventions are information-centric, providing unidirectional education rather than incorporating structured psychotherapy, skills training, or facilitating two-way communication [42,62]. Such a format appears insufficient for translating into actual skill enhancement, family interaction, or overall well-being improvement [43]. Furthermore, some studies reported positive subjective feedback from families despite nonsignificant quantitative outcomes, indicating that current measurement tools may lack sensitivity to nuanced experiential gains [40,43,44].

Moving forward, future research should use mixed methods designs to capture subjective experience and identify needs of high-risk subgroups; tailor interventions according to patients' diagnosis, cultural context, and characteristics of caregiver; develop more interactive and personalized DHIs formats, such as artificial intelligence–based scenario simulations, two-way communication platforms, and longitudinal support programs; extend follow-up duration and use multimodal outcome measures that combine patients and family members reported experiences with clinically relevant indicators; and explore hybrid online-offline interventions to support continuity of care across different phases of the ICU trajectory.

Strengths and Limitations

This systematic review has several strengths. First, to our knowledge, this is the first meta-analysis to synthesize and appraise the evidence about the effects of DHIs on psychological and related outcomes among family members in the ICU. Second, we strictly followed PRISMA guidelines to strengthen transparency and clarity of reporting. Third, we searched across 9 databases from inception to October 11, 2025, reflecting the most recent advancements in the field of DHI. Moreover, we used the Cochrane RoB 2 and GRADE methodology for quality assessment. Finally, subgroup analyses, sensitivity analyses, and bias assessment were conducted to validate the stability of the results and explore potential sources of heterogeneity.

However, there were several limitations to this review. First, the considerable heterogeneity among studies, as evidenced by the wide PIs for our primary outcomes, indicates uncertainty about the intervention effect. Discrepancies in participants' characteristics and intervention protocols could lead to inconsistencies and bias in the study findings. Second, the interventions included in the review ranged widely, from simple information websites and SMS text messages to VR- and

XSL•FO
RenderX

app-based psychotherapy, and intervention duration varied. Pooling across these fundamentally different modalities may dilute or obscure which components work. Third, while blinding interventionists poses difficulties in DHT care settings, the absence of blinding for study participants and interventionists in most of the included studies introduced risks of bias, resulting in a methodological error. Fourth, the search approach has certain limitations as it exclusively incorporated papers published in English or Chinese, rendering findings from articles in other languages unknown and constraining the global generalizability of the results. Fifth, while no significant small-study effects were detected, the ability of these tests may be constrained by the limited number of included studies. Hence, we cannot conclusively dismiss potential publication bias. Sixth, the majority of the selected trials were conducted in developed countries. This bias may result from the restricted access to technology and the low practicality of technology-based interventions in low-income countries. Seventh, the results regarding the effect of DHIs on psychological outcomes, QoL, and communication quality are inconclusive because of the limited number of studies and the wide PIs. Negligible effects are likely from poor statistical power, not a genuine absence of an intervention effect.

Implications of Nursing Practice

Given the current limitations and low certainty of the evidence, our findings have limited direct implications for clinical practice but provide valuable insights for future research. ICU nurses, as the medical personnel with the most frequent interactions with patients and their families, are ideally positioned to lead DHIs [23]. They can facilitate online meetings involving physicians, patients, and families to dismantle information barriers, guided by shared decision-making theory. More importantly, nurses can leverage digital platforms to enable personalized communication. This nurse-driven, tailored digital communication strategy not only provides adequate informational support to mitigate their negative emotions but also significantly improves communication quality and overall QoL [63]. The subgroup analyses suggest that DHIs' effectiveness may not be universal but can be enhanced by tailoring interventions to specific family or patient characteristics. In the future, using technologies such as videoconferencing, virtual simulations, and advanced language models, we can deliver more comprehensive and personalized psychological support interventions led by nurses via the internet and mobile devices. Simultaneously, it is necessary to advocate large-sample, multicenter, high-quality, and reproducible RCTs to provide more robust data to support DHI's findings on the psychological effects and QoL of ICU family members in China.

Conclusions

In conclusion, this study provides the first comprehensive synthesis of RCT evidence on DHIs for ICU family members, revealing significant uncertainty in the estimated effects. Despite its methodological rigor and use of PIs, the conclusions of this review are constrained by the intrinsically heterogeneous and low certainty of the existing evidence. Unlike previous reviews, this work integrates more recent studies and focuses explicitly on the family members of ICU patients. Future research should prioritize the development of theory-informed, personalized, and interactive DHIs. Incorporating mixed methods designs and targeting high-risk subgroups would be particularly valuable. Although direct clinical recommendations are limited, DHIs hold potential for enabling more proactive and individualized communication strategies for families of ICU patients.

Acknowledgments

We extend our gratitude to the librarians for their invaluable assistance and expertise in conducting the literature search for this study.

We essentially wrote the paper ourselves. Since the authors are not native speakers, we used ChatGPT (OpenAI) to brainstorm verbs and adjectives to express our opinions better.

Funding

This work was supported by the Zhejiang Provincial Natural Science Foundation of China under Grant LTGY24H090006, and the General Program of the Zhejiang Provincial Health Commission under Grant 2025KY904. The funding sources had no role in the design and conduct of the study, the collection, management, analysis, and interpretation of the data, the preparation, review, or approval of the manuscript, or the decision to submit the manuscript for publication. The contents of this paper are solely the responsibility of the authors and do not necessarily represent those of the Zhejiang Provincial Department of Science and Technology or the Zhejiang Provincial Health Commission.

Data Availability

All data used in this study are available in the article and supplementary material. Further inquiries could be directed to the first author.

Authors' Contributions

Chuchu Z: writing - original draft, validation, software, methodology, formal analysis, data curation, reviewing and editing
WS: writing - review and editing, validation, software

XSL•FO
RenderX

WJ: writing - review & editing, software, methodology, data curation
YS: writing - conceptualization, software, methodology, data curation
ZH: writing - formal analysis, data curation
RY: writing - methodology
JY: writing - methodology
JZ: writing - methodology
Chaoyi Z: writing - data curation
YZ: writing - conceptualization, funding acquisition
XG: writing - review & editing, supervision, funding acquisition

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA 2020 checklist.
[DOC File , 124 KB - jmir_v28i1e83294_app1.doc]

Multimedia Appendix 2
PRISMA-S checklist.
[DOCX File , 17 KB - jmir_v28i1e83294_app2.docx]

Multimedia Appendix 3
Search strategy.
[DOCX File , 35 KB - jmir_v28i1e83294_app3.docx]

Multimedia Appendix 4
Subgroup analysis by potential covariates.
[DOCX File , 3412 KB - jmir_v28i1e83294_app4.docx]

Multimedia Appendix 5
Publication bias.
[DOCX File , 33 KB - jmir_v28i1e83294_app5.docx]

References

1. Jiang W, Ni Q, Zhang C, Dong Y, Yi J, Yan R, et al. Factors influencing quality of life after intensive care: A systematic review and meta-analysis. Crit Care Med 2025;53(9):e1803-e1814. [doi: 10.1097/CCM.0000000000006770] [Medline: 40662879]
2. Hwang D, Oczkowski S, Lewis K, Birriel B, Downar J, Farrier C, et al. Society of critical care medicine guidelines on family-centered care for adult ICUs: 2024. Crit Care Med 2025;53(2):e465-e482. [doi: 10.1097/CCM.0000000000006549] [Medline: 39982184]
3. Shirasaki K, Hifumi T, Nakanishi N, Nosaka N, Miyamoto K, Komachi M, et al. Postintensive care syndrome family: A comprehensive review. Acute Med Surg 2024;11(1):e939 [FREE Full text] [doi: 10.1002/ams2.939] [Medline: 38476451]
4. Sui W, Gong X, Qiao X, Zhang L, Cheng J, Dong J, et al. Family members' perceptions of surrogate decision-making in the intensive care unit: A systematic review. Int J Nurs Stud 2023;137:104391. [doi: 10.1016/j.ijnurstu.2022.104391] [Medline: 36442321]
5. Lin Z, Huang Q, Ya R, Huang H, Bai Y, Lin W. Survey on sleep, mental health, and quality of life among family members of critically ill patients in rehabilitation hospitalization. Chin J Rehabil Med 2025;40(7):1071-1075.
6. Shirasaki K, Ohbe H, Hifumi T, Okajima M, Otani N, Yasunaga H. Association between patients' admission to the ICU and psychological disorders in their families: a retrospective matched-pair cohort study. Crit Care Med 2025;53(10):e1963-e1972. [doi: 10.1097/CCM.0000000000006793] [Medline: 40736355]
7. Abdul Halain A, Tang LY, Chong MC, Ibrahim NA, Abdullah KL. Psychological distress among the family members of intensive care unit (ICU) patients: A scoping review. J Clin Nurs 2022;31(5-6):497-507. [doi: 10.1111/jocn.15962] [Medline: 34254377]
8. de Souza JMB, Miozzo AP, da Rosa Minho Dos Santos R, Mocellin D, Rech GS, Trott G, et al. Long-term effects of flexible visitation in the intensive care unit on family members' mental health: 12-month results from a randomized clinical trial. Intensive Care Med 2024;50(10):1614-1621. [doi: 10.1007/s00134-024-07577-3] [Medline: 39172240]

XSL•FO
RenderX

9. Davidson JE, Jones C, Bienvenu OJ. Family response to critical illness: postintensive care syndrome-family. Crit Care Med 2012;40(2):618-624. [doi: 10.1097/CCM.0b013e318236ebf9] [Medline: 22080636]
10. Huang W, Gao Y, Zhou L, Xiao X, Xu H, Lu L, et al. Effects of ICU diaries on psychological disorders and sleep quality in critically ill patients and their family members: A systematic review and meta-analysis. Sleep Med 2024;122:84-91. [doi: 10.1016/j.sleep.2024.08.002] [Medline: 39137664]
11. Nielsen AH, Angel S. How diaries written for critically ill influence the relatives: a systematic review of the literature. Nurs Crit Care 2016;21(2):88-96. [doi: 10.1111/nicc.12158] [Medline: 26487425]
12. McPeake J, Hirshberg E, Christie L, Drumright K, Haines K, Hough C, et al. Models of peer support to remediate post-intensive care syndrome: a report developed by the society of critical care medicine thrive international peer support collaborative. Crit Care Med 2019;47(1):e21-e27 [FREE Full text] [doi: 10.1097/CCM.0000000000003497] [Medline: 30422863]
13. Bohart S, Møller AM, Andreasen AS, Waldau T, Lamprecht C, Thomsen T. Effect of patient and family centred care interventions for adult intensive care unit patients and their families: a systematic review and meta-analysis. Intensive Crit Care Nurs 2022;69:103156. [doi: 10.1016/j.iccn.2021.103156] [Medline: 34753631]
14. Hoffmann M, Jeitziner M, Riedl R, Mueller G, Peer A, Bachlechner A, et al. Effects of an online information tool on post-traumatic stress disorder in relatives of intensive care unit patients: a multicenter double-blind, randomized, placebo-controlled trial (ICU-Families-Study). Intensive Care Med 2023;49(11):1317-1326. [doi: 10.1007/s00134-023-07215-4] [Medline: 37870597]
15. Münchenberg PS, Yessimova D, Panteli D, Kurth T. Digital health interventions for informal family caregivers of people with first-episode psychosis: systematic review on user experience and effectiveness. JMIR Ment Health 2024;11:e63743 [FREE Full text] [doi: 10.2196/63743] [Medline: 39607998]
16. Xiong J, Wu T, Cai H, He J, Wang T, Lin X, et al. Effects of wab-WPPEP on post-intensive care syndrome-family in family caregivers of heart valve replacement patients: A randomized controlled trial. Intensive Crit Care Nurs 2025;86:103817 [FREE Full text] [doi: 10.1016/j.iccn.2024.103817] [Medline: 39426206]
17. NA. Global Strategy on Digital Health 2020-2025. Geneva: World Health Organization; 2021.
18. Zhao Q, Chen X, Zhang X, Wang S. Research progress and hotspot analysis in the digital health field. China Digit Med 2022;17(4):37-44.
19. Dong M, Huang Y, Zhang X, Wu L, Ghazal L, Cao F. Effects of psychological intervention based on digital health technology on psychological outcomes in cancer survivors: systematic review and meta-analysis. Psychooncology 2025;34(5):e70166. [doi: 10.1002/pon.70166] [Medline: 40325624]
20. Guo G, Lan M, Zeng F, Liang J, Guo L, Cai P, et al. Effects of remote interventions with digital health technologies in lung transplant patients:a meta-analysis. Chin J Nurs 2024;59(11):1389-1396.
21. Shaffer KM, Turner KL, Siwik C, Gonzalez BD, Upasani R, Glazer JV, et al. Digital health and telehealth in cancer care: a scoping review of reviews. Lancet Digit Health 2023;5(5):e316-e327 [FREE Full text] [doi: 10.1016/S2589-7500(23)00049-3] [Medline: 37100545]
22. Lau CKY, Saad A, Camara B, Rahman D, Bolea-Alamanac B. Acceptability of digital mental health interventions for depression and anxiety: systematic review. J Med Internet Res 2024;26:e52609 [FREE Full text] [doi: 10.2196/52609] [Medline: 39466300]
23. Zhang Y, Flannery M, Zhang Z, Underhill-Blazey M, Bobry M, Leblanc N, et al. Digital health psychosocial intervention in adult patients with cancer and their families: systematic review and meta-analysis. JMIR Cancer 2024;10:e46116 [FREE Full text] [doi: 10.2196/46116] [Medline: 38315546]
24. Shin Y, Kim SK, Kim Y, Go Y. Effects of app-based mobile interventions for dementia family caregivers: a systematic review and meta-analysis. Dement Geriatr Cogn Disord 2022;51(3):203-213. [doi: 10.1159/000524780] [Medline: 35609526]
25. Leng M, Zhao Y, Xiao H, Li C, Wang Z. Internet-based supportive interventions for family caregivers of people with dementia: systematic review and meta-analysis. J Med Internet Res 2020;22(9):e19468 [FREE Full text] [doi: 10.2196/19468] [Medline: 32902388]
26. Shin JW, Tan A, Tate J, Balas M, Dabelko-Schoeny H, Happ MB. Preliminary efficacy of the vidatalk communication application on family psychological symptoms in the intensive care unit: A pilot study. Heart Lung 2025;70:14-22. [doi: 10.1016/j.hrtlng.2024.11.001] [Medline: 39546958]
27. Hariton E, Locascio JJ. Randomised controlled trials - the gold standard for effectiveness research: Study design: randomised controlled trials. BJOG 2018;125(13):1716 [FREE Full text] [doi: 10.1111/1471-0528.15199] [Medline: 29916205]
28. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
29. Julian H, James T, Jacqueline C, Miranda C, Tianjing L, Matthew P, et al. Cochrane Handbook for Systematic Reviews of Interventions (Current Version). 2024. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current [accessed 2025-07-27]
30. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159-174. [Medline: 843571]

XSL•FO
RenderX

31. RoB 2: A revised cochrane risk-of-bias tool for randomized trials. Cochrane bias. 2019. URL: https://methods.cochrane.org/bias/resources/rob-2-revised-cochrane-risk-bias-tool-randomized-trials [accessed 2025-07-27]
32. Sterne JAC, Savovi J, Page MJ, Elbers RG. RoB 2: A revised tool for assessing risk of bias in randomised trials. BMJ. 2019. URL: https://www.bmj.com/content/366/bmj.l4898.long [accessed 2025-07-27]
33. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;64(4):383-394. [doi: 10.1016/j.jclinepi.2010.04.026] [Medline: 21195583]
34. IntHout J, Ioannidis JPA, Borm GF. The hartung-knapp-sidik-jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard dersimonian-laird method. BMC Med Res Methodol 2014;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
35. Ng AYM, Takemura N, Xu X, Smith R, Kwok JY, Cheung DST, et al. The effects of advance care planning intervention on nursing home residents: A systematic review and meta-analysis of randomised controlled trials. Int J Nurs Stud 2022;132:104276. [doi: 10.1016/j.ijnurstu.2022.104276] [Medline: 35667145]
36. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods 2010;1(2):97-111. [doi: 10.1002/jrsm.12] [Medline: 26061376]
37. Borenstein M. How to understand and report heterogeneity in a meta-analysis: The difference between I-squared and prediction intervals. Integr Med Res 2023;12(4):101014 [FREE Full text] [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
38. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011;343:d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
39. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997;315(7109):629-634 [FREE Full text] [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
40. Chiang VCL, Lee RLP, Ho FM, Leung CK, Tang YP, Wong WS, et al. Fulfilling the psychological and information need of the family members of critically ill patients using interactive mobile technology: A randomised controlled trial. Intensive Crit Care Nurs 2017;41:77-83. [doi: 10.1016/j.iccn.2017.03.006] [Medline: 28438476]
41. Cox CE, Ashana DC, Riley IL, Olsen MK, Casarett D, Haines KL, et al. Mobile application-based communication facilitation platform for family members of critically ill patients: a randomized clinical trial. JAMA Netw Open 2024;7(1):e2349666 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.49666] [Medline: 38175648]
42. Cox CE, White DB, Hough CL, Jones DM, Kahn JM, Olsen MK, et al. Effects of a personalized web-based decision aid for surrogate decision makers of patients with prolonged mechanical ventilation: a randomized clinical trial. Ann Intern Med 2019;170(5):285-297 [FREE Full text] [doi: 10.7326/M18-2335] [Medline: 30690645]
43. Drop DLQ, Vlake JH, Wils E, Van Bommel J, Jung C, Hilling DE, et al. Effect of an intensive care unit virtual reality intervention on relatives´ mental health distress: a multicenter, randomized controlled trial. Crit Care 2025;29(1):62 [FREE Full text] [doi: 10.1186/s13054-025-05281-2] [Medline: 39910674]
44. Gawlytta R, Kesselmeier M, Scherag A, Niemeyer H, Böttche M, Knaevelsrud C, et al. Internet-based cognitive-behavioural writing therapy for reducing post-traumatic stress after severe sepsis in patients and their spouses (REPAIR): results of a randomised-controlled trial. BMJ Open 2022;12(3):e050305 [FREE Full text] [doi: 10.1136/bmjopen-2021-050305] [Medline: 35264337]
45. Petrinec AB, Wilk C, Hughes JW, Zullo MD, George RL. Self-care mental health app intervention for post-intensive care syndrome-family: A randomized pilot study. Am J Crit Care 2023;32(6):440-448. [doi: 10.4037/ajcc2023800] [Medline: 37907376]
46. Shariati E, Dadgari A, Talebi SS, Mahmoodi Shan GR, Ebrahimi H. The effect of the web-based communication between a nurse and a family member on the perceived stress of the family member of patients with suspected or confirmed covid-19: a parallel randomized clinical trial. Clin Nurs Res 2021;30(7):1098-1106. [doi: 10.1177/10547738211017688] [Medline: 34044625]
47. Suen AO, Butler RA, Arnold RM, Myers B, Witteman HO, Cox CE, et al. A pilot randomized trial of an interactive web-based tool to support surrogate decision makers in the intensive care unit. Ann Am Thorac Soc 2021;18(7):1191-1201 [FREE Full text] [doi: 10.1513/AnnalsATS.202006-585OC] [Medline: 33326348]
48. Watland S, Solberg Nes L, Ekeberg Ø, Rostrup M, Hanson E, Ekstedt M, et al. The caregiver pathway intervention can contribute to reduced post-intensive care syndrome among family caregivers of icu survivors: a randomized controlled trial. Crit Care Med 2025;53(3):e555-e566. [doi: 10.1097/CCM.0000000000006546] [Medline: 39718436]
49. Woo H, Oh S, Lim L, Im H, Lee H, Ha E, et al. Efficacy of virtual visitation in ICU during COVID-19 pandemic: The ICU visits randomized controlled trial. Crit Care Med 2024;52(12):e593-e603. [doi: 10.1097/CCM.0000000000006429] [Medline: 39283178]
50. Yuan C, Xiao Y, Wang F, Wang Y, Wang Y, Lin F. The effect of video visitation on intensive care unit patients and family members outcomes during the COVID-19 pandemic: A randomised controlled trial. Intensive Crit Care Nurs 2023;76:103394 [FREE Full text] [doi: 10.1016/j.iccn.2023.103394] [Medline: 36731263]

XSL•FO
RenderX

51. Carlozzi NE, Troost JP, Sen S, Choi SW, Wu Z, Miner JA, et al. Improving outcomes for care partners of individuals with traumatic brain injury: results for a mhealth randomized control trial of the careqol app. Arch Phys Med Rehabil 2025;106(4):548-561. [doi: 10.1016/j.apmr.2024.12.022] [Medline: 39798892]
52. Cox CE, Ashana DC, Dempsey K, Olsen MK, Parish A, Casarett D, et al. Mobile app-facilitated collaborative palliative care intervention for critically ill older adults: a randomized clinical trial. JAMA Intern Med 2025;185(2):173-183. [doi: 10.1001/jamainternmed.2024.6838] [Medline: 39680398]
53. Vranceanu A, Bannon S, Mace R, Lester E, Meyers E, Gates M, et al. Feasibility and efficacy of a resiliency intervention for the prevention of chronic emotional distress among survivor-caregiver dyads admitted to the neuroscience intensive care unit: a randomized clinical trial. JAMA Netw Open 2020;3(10):e2020807 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.20807] [Medline: 33052404]
54. Wen J, Hu Z, Yang S, Huang Y, Zhang L, Zhong C, et al. A qualitative study on experience of post-intensive care syndrome-family symptoms. Med Philos 2024;45(19):41-46.
55. Kean S, Donaghy E, Bancroft A, Clegg G, Rodgers S. Theorising survivorship after intensive care: A systematic review of patient and family experiences. J Clin Nurs 2021;30(17-18):2584-2610. [doi: 10.1111/jocn.15766] [Medline: 33829568]
56. Aghaie B, Anoosheh M, Foroughan M, Mohammadi E, Kazemnejad A. A whirlpool of stress in families of intensive care unit patients: a qualitative multicenter study. Crit Care Nurse 2021;41(3):55-64. [doi: 10.4037/ccn2021322] [Medline: 34061191]
57. Anders S, Albert R, Miller A, Weinger MB, Doig AK, Behrens M, et al. Evaluation of an integrated graphical display to promote acute change detection in ICU patients. Int J Med Inform 2012;81(12):842-851 [FREE Full text] [doi: 10.1016/j.ijmedinf.2012.04.004] [Medline: 22534099]
58. Garcia ACM, Camargos Junior JB, Sarto KK, Silva Marcelo CAD, Paiva EMDC, Nogueira DA, et al. Quality of life, self-compassion and mindfulness in cancer patients undergoing chemotherapy: A cross-sectional study. Eur J Oncol Nurs 2021;51:101924. [doi: 10.1016/j.ejon.2021.101924] [Medline: 33610930]
59. Zhang S, Li J, Zhang Y, Hu X. Effects of advance care planning for patients with advanced cancer: A meta-analysis of randomized controlled studies. Int J Nurs Stud 2025;168:105096. [doi: 10.1016/j.ijnurstu.2025.105096] [Medline: 40378812]
60. Liu S, Yang X, Xie W, Huang S, Xiao W, Deng Y, et al. Identifying family experiences and perspectives on involvement in care of adult ICU patients under family-centered care: A systematic review of qualitative studies. Intensive Crit Care Nurs 2025;90:104059. [doi: 10.1016/j.iccn.2025.104059] [Medline: 40300243]
61. Putowski Z, Rachfalska N, Majewska K, Megger K, Krzych K. Identification of risk factors for post-intensive care syndrome in family members (PICS-F) among adult patients: a systematic review. Anaesthesiol Intensive Ther 2023;55(3):168-178 [FREE Full text] [doi: 10.5114/ait.2023.130831] [Medline: 37728444]
62. Reifarth E, Böll B, Köpke S, Altenrath L, Garcia Borrega J, Kochanek M, et al. Multicomponent communication intervention to support family members of the critically ill: a controlled pre-post study. Crit Care Med 2025;53(7):e1437-e1448. [doi: 10.1097/CCM.0000000000006714] [Medline: 40387470]
63. Wolff JL, Aufill J, Echavarria D, Heughan J, Lee KT, Connolly RM, et al. Sharing in care: engaging care partners in the care and communication of breast cancer patients. Breast Cancer Res Treat 2019;177(1):127-136 [FREE Full text] [doi: 10.1007/s10549-019-05306-9] [Medline: 31165374]

Abbreviations

CNKI: China Knowledge Resource Integrated Database
DSM-5: Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition)
ICU: intensive care unit
DHI: digital health intervention
DHT: digital health technology
GRADE: Grading of Recommendations, Assessment, Development, and Evaluation
PI: prediction interval
PICS-F: Post-intensive Care Syndrome-Family
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-analyses literature search extension
PTSD: posttraumatic stress disorder
QOC: quality of communication
QoL: quality of life
RCT: randomized controlled trial
ROB 2: Cochrane's risk-of-bias tool for randomized trials, version 2
SMD: standardized mean difference
VIP: Weipu Databse
VR: virtual reality

XSL•FO
RenderX

Edited by S Brini; submitted 31.Aug.2025; peer-reviewed by H Maheshwari, I Orimolade; comments to author 07.Oct.2025; accepted 22.Dec.2025; published 25.Feb.2026.

Please cite as:
Zhang C, Sui W, Jiang W, Shen Y, Huang Z, Yan R, Yi J, Zhang J, Zhang C, Zhuang Y, Gong X
The Effects of Digital Health Interventions for Family Members in Intensive Care Units: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e83294
URL: https://www.jmir.org/2026/1/e83294
doi:10.2196/83294
PMID:

©Chuchu Zhang, Weijing Sui, Weilin Jiang, Yibing Shen, Zhenzhen Huang, Ran Yan, Jia Yi, Jiayu Zhang, Chaoyi Zhang, Yiyu Zhuang, Xiaoyan Gong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

AI-Enabled Wearables for Motor Function Assessment and Rehabilitation in Parkinson Disease: Scoping Review

Shengting Li[1,2*], MSc; Siqi Chen[1,2*], MSc; Xiaosong Yu[1,2*], MSc; Huixiang Shang[1,2], MSc; Tingting Tu[1,2], MSc; Mingtao Quan[1,2], MSc

[1]Department of Nursing, Affiliated Hospital of Zunyi Medical College, Zunyi, Guizhou, China

[2]School of Nursing, Zunyi Medical University, Zunyi, Guizhou, China

[*]these authors contributed equally

Corresponding Author:
Mingtao Quan, MSc
Department of Nursing
Affiliated Hospital of Zunyi Medical College
149 Dalian Road, Huichuan District
Zunyi, Guizhou, 563000
China
Phone: 86 18185066688
Email: 865768818@qq.com

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2026/1/e97418

Abstract

Background: Artificial intelligence (AI)–enabled wearable devices are rapidly emerging in rehabilitation and motor function assessment for patients with Parkinson disease (PD). However, evidence remains fragmented, integration into nursing practice is limited, and comprehensive synthesis is lacking.

Objective: This study aimed to summarize studies on AI-enabled wearable devices for PD rehabilitation and motor function assessment, describing device types, monitored indicators, algorithms, and application characteristics, and identifying research gaps and barriers to clinical translation.

Methods: Guided by the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) framework, 9 databases (China National Knowledge Infrastructure, Wanfang Data, SinoMed, Cochrane Library, PubMed, Web of Science, CINAHL, Scopus, and Embase) were searched from inception to December 2025. Eligible studies were published in English or Chinese from January 1, 2020, onward and enrolled people with PD using noninvasive, body-worn AI-enabled wearable devices for rehabilitation, assessment, or monitoring. Dissertations and full conference papers were included, whereas preprints and conference abstracts were excluded. Methodological quality was appraised using the Mixed Methods Appraisal Tool, 2018 tool. Results were synthesized narratively and mapped to characterize devices, sensing modalities, algorithms, and evaluation methods.

Results: A total of 66 studies involving approximately 3579 participants were included. Wearable devices mainly comprised multisensor modules, smart insoles, and wrist-worn devices, with accelerometers being the most frequently used sensors. Data collection was predominantly passive, and most studies were conducted in laboratory or clinical settings using single- or short-term sessions. Internal validation approaches, particularly leave-one-out and k-fold cross-validation, were common, whereas external validation was rare, and reporting of calibration and clinical decision thresholds was limited. Sensitivity and accuracy were the most frequently reported performance metrics, highlighting substantial heterogeneity in analytical methods and outcome reporting.

Conclusions: This scoping review systematically synthesized evidence on AI-enabled wearable devices for motor function assessment and rehabilitation in PD, complemented by an evidence map and guided by a rehabilitation- and nursing-oriented perspective, and identified key translational gaps between proof-of-concept studies and real-world rehabilitation workflows. Compared with previous reviews that primarily focused on monitoring functions or device performance, this review places greater emphasis on rehabilitation applications and nurse-led translation into practice, and proposes a conceptual “challenges and

XSL•FO
RenderX

opportunities" framework to inform the design, evaluation, and reporting of devices and algorithms, while further highlighting key considerations for workflow integration and the implementation of decision-support systems. These findings have practical relevance for advancing continuity of rehabilitation across clinical, home, and community settings, and may help guide nurses in delivering continuous monitoring, personalized follow-up, and timely intervention, thereby improving the efficiency and accessibility of rehabilitation management.

(J Med Internet Res 2026;28:e85596) doi:10.2196/85596

KEYWORDS

artificial intelligence; Parkinson disease; rehabilitation; scoping review; wearable devices

Introduction

Background

Parkinson disease (PD) is a common neurodegenerative disorder characterized by tremor, bradykinesia, rigidity, and postural or gait disturbances [1,2]. Its high prevalence and disability rates profoundly impair patients' quality of life and impose substantial burdens on families and health care systems. Since 1990, the number of patients with PD has more than doubled—from approximately 2.8 million to over 6.2 million—and is projected to exceed 12 million by 2040 [3]. In China, accelerated population aging has led to a rising prevalence, with an estimated 1.37% among adults aged 60 years and older and more than 3.62 million patients nationwide [4,5]. Rehabilitation remains essential for improving motor function and slowing functional decline [6-8]. However, conventional face-to-face rehabilitation is time- and resource-intensive, often inaccessible in remote regions, and frequently associated with poor adherence [9,10].

In recent years, the rapid development and application of wearable devices have provided significant advantages for assessment and intervention in neurological and psychiatric disorders, such as stroke and depression [11,12]. Wearable devices encompass a wide range of intelligent electronic devices, including smartwatches, wristbands, insoles, exoskeletons, multisensor wearable devices, and electronic textiles [13-16]. These devices support continuous, real-time, and multidimensional monitoring during daily activities or rehabilitation training, capturing key motor indicators such as gait, tremor, joint mobility, postural control, motor coordination, and overall physical activity levels [17-19].

Against this backdrop, the rapid emergence of artificial intelligence (AI)–enabled wearable devices has created new opportunities for the rehabilitation of PD. In this review, AI-enabled wearable devices refer to noninvasive wearable technologies that incorporate machine learning (ML) or deep learning (DL) methods for data-driven analysis beyond rule-based approaches. Unlike traditional approaches that rely on manual observation or intermittent scale-based assessments, AI-enabled wearable devices enable the capture of large volumes of fine-grained data. Supported by ML and DL techniques, these data can be efficiently processed and intelligently interpreted within a data-driven rehabilitation framework, highlighting the importance of AI model interpretability for clinical and nursing decision-making [20-22]. Moreover, AI-enabled wearable devices facilitate continuous remote monitoring and long-term follow-up, making it possible to track disease progression and changes in motor function in real time within home-based rehabilitation settings. These data can then be leveraged by clinicians and rehabilitation teams to design more precise training programs [23,24]. Such an approach enhances the continuity and personalized rehabilitation, offers a feasible pathway for remote and home-based rehabilitation, and holds promising potential for optimizing health care resource allocation and improving patient adherence.

Research Questions and Objectives

Although numerous studies and reviews have explored the use of wearable devices in PD rehabilitation, several limitations remain, including (1) most studies involve small sample sizes and short follow-up periods, providing insufficient longitudinal evidence to evaluate long-term rehabilitation outcomes [17,25,26]; (2) current reviews mainly focus on device performance or symptom monitoring, with limited attention to AI data-processing workflows, algorithmic applications, and their potential value for clinical and nursing practice [18,19]; (3) studies differ substantially in device types, monitoring indicators, data-processing strategies, and AI algorithm selection. The absence of a unifying framework to synthesize these variations hinders a comprehensive understanding of the research landscape [27-29]; and (4) evidence in nursing practice remains limited; despite nurses' central role in rehabilitation, research on AI-enabled tools for nursing decision-making, health education, and adherence support is still scarce [30,31]. Collectively, these issues result in fragmented and unsystematic evidence, limiting its ability to guide clinical rehabilitation nursing practice and theoretical development.

To address these needs, we conducted a scoping review and developed an evidence map to systematically present the current applications of AI-enabled wearable devices in PD rehabilitation and motor function assessment. The scoping review summarized the breadth, depth, and nature of existing evidence [32], while the evidence map visualizes research distribution, patterns, and emerging trends across domains [33]. By integrating these 2 approaches, this study aimed to map and visualize the current evidence on AI-enabled wearable devices for PD rehabilitation and motor function assessment. The objectives were to summarize device types, monitoring indicators, AI algorithms, and application characteristics, and to identify research gaps and future directions to support nursing practice and further digital-health implementation in PD rehabilitation.

XSL•FO
RenderX

Methods

Protocol and Registration

The review followed the methodological guidance outlined in the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [34], and the completed checklist is provided in Multimedia Appendix 1. No protocol was registered for this review.

Eligibility Criteria

The inclusion criteria were defined using the population, concept, and context framework. Eligible studies enrolled participants with a confirmed diagnosis of PD and used noninvasive, body-worn wearable devices integrated with AI-based analytical methods. The focus of included studies was on rehabilitation, motor function assessment, or monitoring. Gray literature such as dissertations and full conference papers was included, whereas preprints and conference abstracts were excluded to ensure methodological consistency and to prioritize evidence that had undergone formal peer review. Only studies published in English or Chinese on or after January 1, 2020, were eligible for inclusion. Detailed inclusion and exclusion criteria are presented in Multimedia Appendix 2.

Information Sources

A comprehensive search was conducted on August 12, 2025, across 9 databases: CNKI (China National Knowledge Infrastructure), Wanfang Data, SinoMed, Cochrane Library, PubMed, Web of Science, CINAHL, Scopus, and Embase. The literature coverage spanned from database inception to the search date. No websites or other nondatabase information sources were searched. Reference lists and forward citations of the included studies were additionally screened to identify further potentially relevant records. To ensure the currency and completeness of the evidence, an updated search using the same strategy was conducted on December 10, 2025.

Search Strategy

The reporting of the search process adhered to the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) [35]. The search strategy was developed by a researcher trained in systematic literature retrieval and combined MeSH (Medical Subject Headings) terms with free-text keywords across four core concepts: (1) wearable or body-worn mobile devices; (2) rehabilitation, assessment, or monitoring; (3) PD; and (4) AI or ML. No limits or restrictions, including language, publication date, or study design, were applied at the search level. All eligibility criteria were applied manually during title/abstract screening and full-text review. The search strategy did not undergo peer review. Complete search strategies for each database and information source are provided in Multimedia Appendix 3.

Selection of Sources of Evidence

The selection process comprised two stages: (1) title/abstract screening and (2) full-text review. In the first stage, 1 reviewer imported all records into EndNote X9 (version 12062; Clarivate), removed duplicates, and excluded studies outside the eligible date range. Two reviewers then independently screened titles and abstracts to determine provisional inclusion. In the second stage, the same 2 reviewers independently assessed the full texts to establish the final set of included studies. Disagreements were resolved through discussion; unresolved cases were adjudicated by a third reviewer. Cohen κ was calculated to measure the interrater agreement [36]. Agreement was high for title and abstract screening (κ=0.86), whereas it was moderate for full-text review (κ=0.57).

Data Charting Process

Data charting was conducted independently by 1 reviewer using a standardized extraction form developed a priori. The form included key study characteristics, wearable device details, AI methods, outcomes, and findings relevant to the review questions. Following Joanna Briggs Institute guidance, charting was iterative, and the form was refined as needed. A second reviewer verified all data, with disagreements resolved by a third reviewer. The form was pilot-tested on 5 studies before full implementation.

Data Items

The data extraction form was developed by the research team based on the objectives of this review and included three core modules: (1) study details, (2) wearable device details, and (3) AI details. When the relevant information was not reported, it was recorded as not available. Each module contained specific data variables, as detailed in Multimedia Appendix 4.

Critical Appraisal of Individual Sources of Evidence

The methodological quality of all included studies was evaluated using the Mixed Methods Appraisal Tool, 2018 version [37]. Two reviewers conducted the assessments independently and resolved discrepancies through discussion; cases in which consensus could not be reached were adjudicated by a third reviewer. The appraisal was used descriptively to identify methodological strengths and limitations of the included studies and was not used as a criterion for study inclusion. The detailed appraisal results are provided in Multimedia Appendix 5 [38-103].

Synthesis of Results

A narrative synthesis approach was used. Extracted data were summarized thematically and presented in descriptive text, tables, and an evidence map to illustrate research trends, methodological characteristics, and application domains of AI-enabled wearable devices in PD.

Ethical Considerations

This scoping review used only published, peer-reviewed literature and therefore did not require approval from an ethics committee or institutional review board. Findings will be submitted to an open access, peer-reviewed journal and presented at relevant medical and engineering conferences.

Results

Study Selection

A total of 1467 records were initially retrieved. Of these, 475 (32.38%) were removed using EndNote X9 due to duplication,

XSL•FO
RenderX

publication date, or language incompatibility. The remaining 992 (67.62%) records underwent title and abstract screening, during which 877 (88.41%) were excluded. Among the 115 records reviewed in full text, 47 (40.87%) were unrelated to AI, 7 (6.09%) were unavailable in full text, and 2 (1.74%) lacked extractable data. An additional 7 relevant studies were identified through reference tracing. In total, 66 studies were included in this review [38-103]. The detailed selection process is illustrated in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram (Figure 1).

Figure 1. Flowchart of the study selection process.

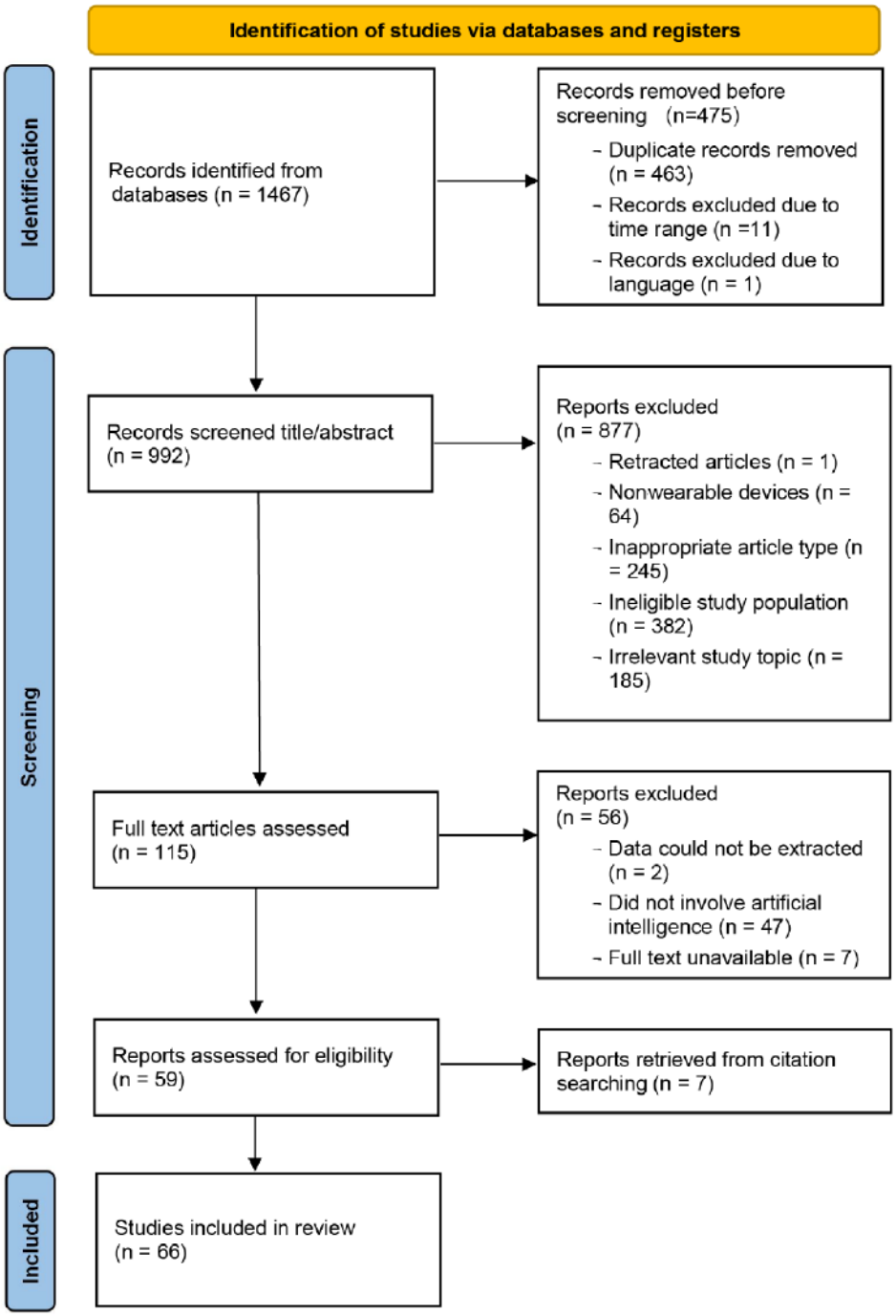

Characteristics of Included Studies

All included studies were published between 2020 and 2025, and the publication volume remained relatively stable over this 6-year period. The highest number of studies was published in 2025 (16/66, 24.24%; Table 1). The included studies originated from multiple countries and regions (Figure 2), with Switzerland

XSL•FO
RenderX

contributing the largest proportion (27/66, 40.91%). With respect to publication type, the vast majority were journal articles (61/66, 92.42%).

XSL•FO
RenderX

Table 1. Characteristics of the included studies using artificial intelligence–enabled wearable devices for Parkinson disease (PD; N=66).

Features	Values	References
Year of publication, n (%)		
2020	8 (12.12)	[38,47,50,53,61,63,66,73]
2021	11 (16.67)	[44,49,55,56,59,62,72,82,86,94,96]
2022	9 (13.64)	[39,46,51,58,64,68,80,81,83]
2023	11 (16.67)	[43,57,60,71,74,76,77,84,85,89,91]
2024	11 (16.67)	[48,54,65,67,69,70,78,87,90,93,95]
2025	16 (24.24)	[40-42,45,52,75,79,88,92,97-103]
Type of publication, n (%)		
Journal article	61 (92.42)	[38,40-45,47-69,71-84,86-91,93-103]
Conference paper	4 (6.06)	[39,46,70,92]
Thesis	1 (1.52)	[85]
Country/region of publication, n (%)		
Switzerland	27 (40.91)	[41-44,47,49,55,57,58,62,65-68,71,75,79,80,88,90,91,96,97,99,101-103]
United States	19 (28.79)	[39,46,48,51,54,59,61,64,69,70,73,74,76,78,81,86,89,92,94]
United Kingdom	11 (16.67)	[50,52,53,56,60,63,72,87,93,95,98]
Netherlands	5 (7.58)	[38,40,45,82,100]
China	3 (4.55)	[83-85]
Germany	1 (1.52)	[77]
Number of participants		
Mean (SD; range)	55.06 (132.78; 2-1057)	[38-99,101-103]
1-30, n (%)	41 (62.12)	[38-42,44,46-49,52,55-64,67,69-75,82,85,86,90,91,93-97,99,103]
31-60, n (%)	8 (12.12)	[43,68,76-79,81,92]
61-100, n (%)	9 (13.64)	[45,50,51,53,65,66,83,89,101]
101-500, n (%)	6 (9.09)	[54,84,87,88,98,102]
>500, n (%)	1 (1.52)	[80]
Not reported	1 (1.52)	[100]
Age distribution of participants[a] (years)		
Mean[b] (SD; range)	64.59 (10.03; 25.00-76.10)	[38-47,49-53,56-61,63-76,80,81,83-85,89,91-93,95,96,98,99,101-103]
<60, n (%)	3 (4.55)	[39,40,55]
60-69, n (%)	39 (59.09)	[42-45,47,49,51,53,56,63,65-67,69-77,79-81,83,85,86,88,89,91-93,95,96,98,99,102,103]
70-79, n (%)	14 (21.21)	[38,41,46,50,52,57-61,64,68,84,101]
≥80, n (%)	2 (3.03)	[54,97]
Not reported	8 (12.12)	[48,62,78,82,87,90,94,100]
Sex distribution[c]		
Female (%), mean (SD; range)	35.88 (18.08; 0–100)	[38-41,45-48,50-68,70-73,75,79-81,83-92,94-96,98,99,101,102]
Inclusion details n (%)		
PD only	47 (71.21)	[38,40,41,43,46-49,51,52,54-72,76,81,82,84-88,90,93-99,102,103]
Mixed	18 (27.27)	[39,42,44,45,50,53,73-75,77-80,83,89,91,92,101]
Not reported	1 (1.52)	[100]

XSL•FO
RenderX

Features	Values	References
Application objective[d], n (%)		
Motor function assessment	54 (81.82)	[38,39,41-47,49,50,52-55,57-80,82-89,91,92,95,99-102]
Disease progression/symptom monitoring	26 (39.39)	[39,46,51-53,55,56,65,68,69,71,74,76,79,87,89,90,93,94,96-100,102,103]
Efficacy evaluation	11 (16.67)	[46,54,56,61,70,71,77,78,81,90,103]
Rehabilitation training	8 (12.12)	[40,48,54,55,60,67,77,96]
Others	1 (1.52)	[88]

[a]Reported as mean (SD), median (IQR), or range, depending on the study.

[b]Age was summarized from studies reporting mean (SD); studies with only median or without age were excluded. For mixed populations, only PD patient data were used.

[c]Based only on studies that reported sex data.

[d]Application objectives were classified based on the primary functional aim of the wearable system. Motor function assessment refers to quantitative evaluation of motor performance at specific time points; disease progression/symptom monitoring refers to continuous or longitudinal tracking of disease states; efficacy evaluation refers to the assessment of responses to interventions or treatments; and rehabilitation training refers to systems explicitly designed to support training or feedback to improve motor function. The number of studies does not add up, as several studies have used >1 application objective.

Figure 2. Geographical distribution of the included studies using artificial intelligence–enabled wearable devices for Parkinson disease.

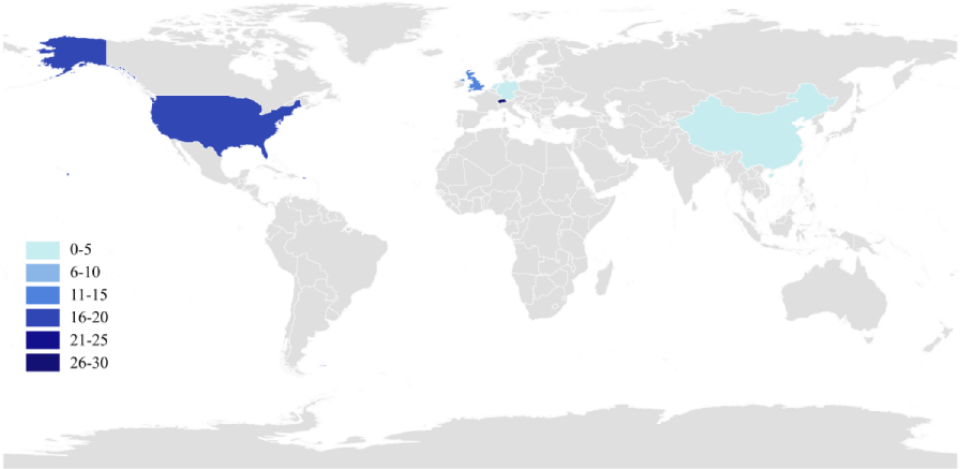

Regarding study populations, sample sizes ranged from 2 to 1057 participants, with a mean sample size of 55.06 (SD 132.78), yielding a total of approximately 3579 participants. Overall, most studies involved relatively small samples with substantial heterogeneity, and the majority included only patients with PD (47/66, 71.21%). In terms of the application objectives of wearable devices, most studies focused on motor function assessment (54/66, 81.82%), whereas rehabilitation-oriented applications were relatively scarce, being reported in only 8 (12.12%) studies. Overall, this distribution suggests that intervention-focused research remains limited. Detailed characteristics of the included studies are presented in Table 1 and Multimedia Appendix 6 [38-103]. For categories in which studies could contribute to more than 1 classification, percentages may exceed 100%; detailed coding rules are provided in the table footnotes.

Characteristics of Wearable Devices

A total of 5 major types of wearable devices were identified across the included studies. Sensor module was the predominant device type, accounting for 84.85% (56/66) of studies, followed by smart insoles (7/66, 10.61%) and smartwatches (5/66, 7.58%; Table 2). Regarding device origin, commercially available devices were more common (41/66, 62.12%) than noncommercial devices (21/66, 31.82%); nevertheless, a substantial proportion of studies still relied on customized or research-grade devices. A total of 14 distinct wearing locations were reported (Figure 3). The most common placements were the shank (25/66, 37.88%) and the wrist (23/66, 34.85%), followed by the foot (19/66, 28.79%) and the ankle (16/66, 24.24%). Many studies used multiple wearing locations within a single study, reflecting the diversity of measurement strategies adopted.

XSL•FO
RenderX

Table 2. Features of artificial intelligence–enabled WDsa used in the included Parkinson disease studies (N=66).

Features	Values, n (%)	References
Type of WD[a,b]		
Sensor module	56 (84.85)	[38-45,47,48,50-75,78-82,84,86-89,91-98,102,103]
Smart insole	7 (10.61)	[39,46,55,57,59,100,101]
Smartwatch	5 (7.58)	[49,63,85,97,99]
Smart wristband	3 (4.55)	[83,85,90]
Stimulator	2 (3.03)	[48,103]
Others (adhesive electrodes, ankle band, or smart glove)	3 (4.55)	[42,76,77]
Status of WD[a,b]		
Commercial	41 (62.12)	[38,42,44,46,48,49,51-59,63-67,70-72,74,75,80,81,84-89,91-93,97-99,101,102]
Noncommercial	21 (31.82)	[39-41,43,45,47,48,50,60,67-69,73,76-78,83,84,90,100,103]
Not reported	8 (12.12)	[61,62,79,82,85,94-96]
Company of WD[a,b]		
APDM Inc	6 (9.09)	[53,62,64,66,86,87]
Shimmer Sensing	5 (7.58)	[38,55,59,85,93]
Activinsights Ltd	4 (6.06)	[43,44,85,97]
MC10 Inc	3 (4.55)	[53,63,75]
Moticon ReGo AG	3 (4.55)	[46,57,101]
Noraxon USA Inc	2 (3.03)	[42,74]
PD Neurotechnology Ltd	2 (3.03)	[57,71]
Great Lakes NeuroTechnologies Inc	2 (3.03)	[56,70]
STMicroelectronics	2 (3.03)	[68,69]
Tekscan Inc	2 (3.03)	[55,59]
Self-developed	13 (19.70)	[39,51,60,61,67,73,76,77,83,84,89,100,103]
Others	16 (24.24)	[43,44,48,50,54,58,63,65,88-92,96,99,102]
Not reported	16 (24.24)	[40,41,45,47,49,52,72,78-82,94,95,97,98]
Placement[b]		
Shank	25 (37.88)	[40,42-44,48,52,55,59,60,67,69,71,74,75,80-82,84,86,88,91,93-95,102]
Wrist	23 (34.85)	[38,44,49,53,56-58,63,69-71,79,83,85,87-92,97,99,102]
Foot	19 (28.79)	[39,40,46,48,55,57,59,62,72-74,87,88,91,96,100-103]
Ankle	16 (24.24)	[38,39,41,45,51,52,61,64,66,70,76,83,93,98,99,103]
Waist	13 (19.70)	[44,47,52,54,57,65,69,71,86,88,89,97,102]
Trunk	12 (18.18)	[45,66,74,75,78,81,84,87,88,92,98,102]
Arm	9 (13.64)	[45,74,75,78,81,84,85,97,98]
Thigh	7 (10.61)	[55,59,68,74,88,95,102]
Hip	6 (9.09)	[62,74,78,82,93,98]
Hand	3 (4.55)	[63,77,91]
Others (Neck, knees, back, or head)	4 (6.06)	[50,51,74,95]
Compatibility with OS[c,d]		
Local logger	16 (24.24)	[38,40,43,44,46,50,54,57,58,60,61,69,73,87,88,93]
Android	6 (9.09)	[64,65,76,85,90,97]

XSL•FO
RenderX

Features	Values, n (%)	References
iOS	5 (7.58)	[48,51,63,65,90]
Windows	3 (4.55)	[68,83,84]
Not applicable	7 (10.61)	[39,52,67,71,94,95,103]
Not reported	31 (46.97)	[41,42,45,47,49,53,55,56,59,62,66,70,72,74,75,77-82,86,89,91,92,96,98-102]
Gateway[e]		
PC[f]	10 (15.15)	[52,68,73,74,76,77,80,86,90,102]
Smartphone	6 (9.09)	[39,48,64,76,85,90]
Tablet	3 (4.55)	[51,80,92]
IoT[g] Gateway	2 (3.03)	[82,103]
Not reported	48 (72.73)	[38,40-47,49,50,53-63,65-67,69-72,75,78,79,81,83,84,87-89,91,93-101]
Host[h]		
PC[f]	35 (53.03)	[38,41,43,44,52,54-59,61,66,68,69,73-77,80,82-84,86-89,91,93,94,96,98,102,103]
Server	6 (9.09)	[39,48,58,71,90,97]
Smartphone	5 (7.58)	[48,64,65,76,85]
On-device	4 (6.06)	[40,60,65,67]
Tablet	3 (4.55)	[51,80,92]
Not reported	18 (27.27)	[42,45-47,49,50,53,62,63,70,72,78,79,81,95,99-101]
Mode of data transfer[i]		
Bluetooth	15 (22.73)	[39,40,48,51,52,64,65,68,76,80,84,86,88,89,102]
Internet	8 (12.12)	[39,48,58,71,83,85,90,103]
Removable media	6 (9.09)	[38,40,43,48,61,69]
Wired	5 (7.58)	[40,44,73,77,103]
Not reported	39 (59.09)	[41,42,45-47,49,50,53-57,59,60,62,63,66,67,70,72,74,75,78,79,81,82,84,87,91-101]
Sensors in the wearables[j]		
Accelerometer	61 (92.42)	[38-54,57-59,61-99,102,103]
Gyroscope	50 (75.76)	[38-43,45-49,51,52,54-57,59-64,66-74,76,78,80,81,83,84,86-89,92-94,96,98,99,102,103]
Magnetometer	11 (16.67)	[45,52,53,66,71,76,78,83,86,87,94]
Pressure sensor	9 (13.64)	[39,40,46,55,57,59,77,100,101]
sEMG[k] sensor	3 (4.55)	[40,42,78]
Flex sensor	2 (3.03)	[40,77]
Measured biosignals[l]		
Acceleration	61 (92.42)	[38-54,57-59,61-99,102,103]
Angular velocity	50 (75.76)	[38-43,45-49,51,52,54-57,59-64,66-74,76,78,80,81,83,84,86-89,92-94,96,98,99,102,103]
Magnetic field signals	11 (16.67)	[45,52,53,66,71,76,78,83,86,87,94]
Pressure/mechanical signals	9 (13.64)	[39,40,46,55,57,59,77,100,101]
EMG[m] signals	3 (4.55)	[40,42,78]
Bending/Flex sensing	2 (3.03)	[40,77]
Sensing type		
Passive	52 (78.79)	[38,40,41,44-60,62,65,71-73,75-79,82-103]
Active	14 (21.21)	[39,42,43,61,63,64,66-70,74,80,81]
Application scenario		

XSL•FO
RenderX

Features	Values, n (%)	References
Clinical	31 (46.97)	[38,48,51,54,55,57,59-61,63,66-68,74,76-78,80,81,83,84,86,88,89,92-96,98,99]
Home	12 (18.18)	[44,47,58,65,71,72,79,82,97,100-102]
Clinical and Home	23 (34.85)	[39-43,45,46,49,50,52,53,56,62,64,69,70,73,75,85,87,90,91,103]
Duration of monitoring/intervention		
Single session	32 (48.48)	[38,40-44,46,48-50,52,55,64,67-69,73,74,76,77,79,83-85,89-91,96,99-102]
Multiple sessions	20 (30.30)	[47,51,53,54,56,57,60,61,63,66,70,75,80,92-95,97,98,103]
Long-term monitoring	9 (13.64)	[39,45,58,62,65,71,72,81,86]
Not reported	5 (7.58)	[59,78,82,87,88]

[a]WD: wearable device.

[b]The number of studies does not add up, as several studies have used >1 wearable device.

[c]OS: operating system.

[d]The number of studies does not add up, as several studies have used >1 wearable device, and many wearable devices are compatible with >1 operating system.

[e]The number of studies does not add up, as several studies used >1 wearable device, and many wearable devices used >1 gateway.

[f]PC: personal computer.

[g]IoT: Internet of Things.

[h]The number of studies does not add up, as several studies used >1 wearable device, and many wearable devices used >1 host.

[i]The number of studies does not add up, as several studies used >1 wearable device, and many wearable devices used >1 mode of data transfer.

[j]The number of studies does not add up, as several studies used >1 wearable device, and most wearable devices have >1 sensor.

[k]sEMG: surface electromyography.

[l]The number of studies does not add up, as several studies used >1 wearable device, and most wearable devices assess >1 biosignal.

[m]EMG: electromyography.

Figure 3. Placement of wearable sensors in studies using artificial intelligence–enabled wearable devices for Parkinson disease.

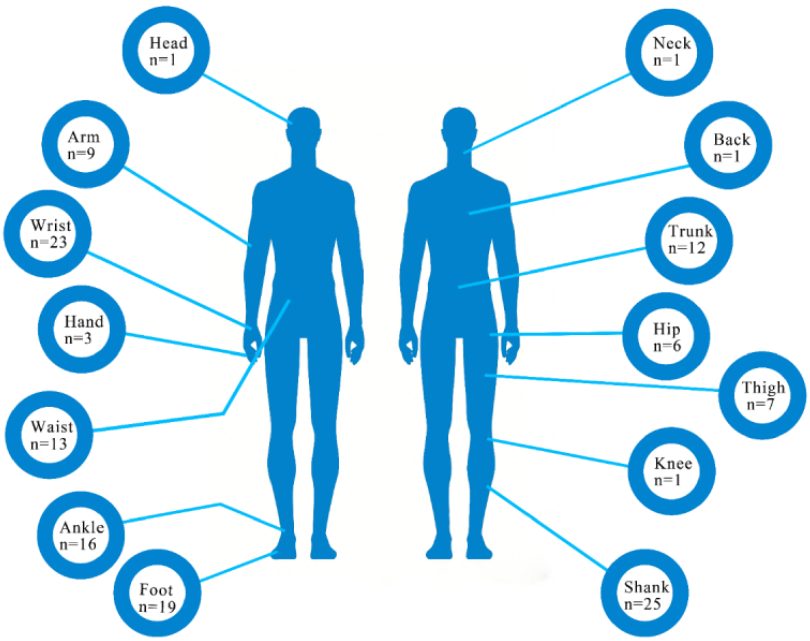

With regard to sensor configuration, devices integrating multiple sensors predominated, accounting for 75.76% (50/66) of the included studies. The collected biosignals were predominantly inertial signals. Accelerometer data were acquired in the vast majority of studies (61/66, 92.42%), followed by gyroscope data (50/66, 75.76%), whereas magnetic field, pressure, and electromyography signals were used less frequently (Table 2).

With respect to application settings, wearable devices were more frequently deployed in clinical environments (31/66, 46.97%), with an additional 34.85% (23/66) of studies spanning both clinical and home environments, whereas studies conducted exclusively in home settings were relatively uncommon (12/66, 18.18%). Regarding the duration of monitoring or intervention, most studies used either single-session assessments (32/66, 48.48%) or repeated short-term testing designs (20/66, 30.30%),

XSL•FO
RenderX

while studies involving long-term continuous monitoring were comparatively limited (9/66, 13.64%). Overall, wearable device research in PD is dominated by sensor module devices, with devices most commonly worn on the lower limbs and wrist. Data collection is primarily passive, and studies largely focus on short-term applications conducted in clinical settings. Additional technical characteristics of the wearable devices are provided in Table 2 and Multimedia Appendix 7 [38-103].

AI Algorithm Characteristics of Wearable Devices

The AI applications in the included studies were categorized into four purposes: (1) monitoring or assessment (35/66, 53.03%), (2) state recognition or functional screening (17/66, 25.76%), (3) prediction (8/66, 12.12%), and (4) rehabilitation and feedback (6/66, 9.09%; Table 3). Among the studies, 62.12% (41/66) used ML algorithms only, 28.79% (19/66) used DL algorithms only, and 9.09% (6/66) combined ML and DL approaches. These studies used algorithms to address classification problems (59/66, 89.39%), regression problems (15/66, 22.73%), and clustering problems (3/66, 4.55%).

XSL•FO
RenderX

Table 3. Features of artificial intelligence (AI) algorithms used in the included Parkinson disease wearable-device studies (N=66).

Features	Values, n (%)	References
AI category		
ML[a]	41 (62.12)	[38,39,43,45,48,50,53-55,57-63,65,66,68,71-77,80-84,86-88,92,96,98,99,101-103]
DL[b]	19 (28.79)	[41,42,44,46,49,51,52,56,64,67,69,70,78,79,89-91,94,100]
ML and DL	6 (9.09)	[40,47,85,93,95,97]
Task type[c]		
Classification	59 (89.39)	[39-55,57-60,62-67,69,71,72,74-79,81-103]
Regression	15 (22.73)	[38,43-45,53,56,61,68-71,77,80,98,102]
Clustering	3 (4.55)	[73,77,98]
AI algorithm[d]		
Support vector machine	28 (42.42)	[38,40,43,45,47,50,53,57,61,65,66,68,74,76,78,81-85,87,88,92,93,96,101-103]
Convolutional neural network	23 (34.85)	[40-42,44,46,47,49,51,52,54,56,67,69,70,78,85,89-91,93,95,97,100]
Random forest	19 (28.79)	[39,45,47,53,57,60,63,79,80,82,83,86,87,92,95,97-99,101]
Logistic regression	15 (22.73)	[39,43,53,61,62,65,71,78-80,87,88,92,98,101]
Decision tree	14 (21.21)	[45,47-50,55,57,59,61,62,76,82,86,88]
Long short-term memory	11 (16.67)	[42,47,52,56,58,69,70,85,93,94,100]
Multilayer perceptron	8 (12.12)	[40,41,44,77,78,88,92,100]
k-nearest neighbors	7 (10.61)	[50,76,83,88,92,93,101]
Naive Bayes	5 (7.58)	[43,50,76,88,101]
Extreme gradient boosting	4 (6.06)	[49,87,93,102]
Gradient boosting	3 (4.55)	[39,57,86]
Self-attention mechanism	3 (4.55)	[42,69,85]
Hidden Markov model	3 (4.55)	[58,72,73]
k-means clustering	3 (4.55)	[75,77,98]
Elastic net	3 (4.55)	[80,87,98]
AdaBoost	3 (4.55)	[76,92,96]
Radial basis function	2 (3.03)	[78,79]
Others	12 (18.18)	[43,45,47,64,67,78-80,82,86,97,102]
Aim of AI algorithm		
Monitoring or assessment	35 (53.03)	[38,39,42-47,50,53,56-59,61,63,65,70-73,79,81,83-85,88,90-93,97-99,102]
State recognition or functional screening	17 (25.76)	[41,49,51,54,62,66,74-78,82,89,94,100,101,103]
Prediction	8 (12.12)	[52,55,64,68,69,80,87,95]
Rehabilitation and feedback	6 (9.09)	[40,48,60,67,86,96]
Validation approach[e]		
Leave-one-out cross-validation	37 (56.06)	[39,41,42,45-47,49,50,53-55,57-63,65-68,70,73,74,76,78,79,81,83,86,88,90,91,93,102,103]
k-fold cross-validation	27 (40.91)	[38-41,47,49-52,54,57,69,72,76,77,84,85,87-92,99,101,103]
Hold-out validation	18 (27.27)	[40,44,48,49,56,61,64,66,76,78,80,82,88,92,94,100-102]
External validation	4 (6.06)	[43,60,69,71]
Not reported	5 (7.58)	[75,95-98]
Performance measures[f]		

XSL•FO
RenderX

Features	Values, n (%)	References
Sensitivity	43 (65.15)	[39-42,46,47,49-55,57-60,62,64-67,69,72,73,75-79,82-85,87,90,94-97,99,101,103]
Accuracy	41 (62.12)	[39-43,45,46,48,50-55,57,58,62,65,67,69,74-77,82-87,89-92,95-98,100-102]
Specificity	27 (40.91)	[46,47,49,50,52-55,59,60,62,64-67,69,76,78,79,82,83,94-96,99,101,103]
F_1-score	27 (40.91)	[39-41,49,51-55,57,67,73,75-78,82,84,85,87,92,93,95,99,101-103]
Area under the curve	25 (37.88)	[39-41,47,52,55,57,58,62-65,69,74,78,79,82,86-88,92,94,96,98,101]
Precision	24 (36.36)	[39-42,46,51,53,54,58,59,62,65-67,75-78,84,85,87,92,95,97]
Correlation coefficient	14 (21.21)	[38,44,53,56,58,61,68,70,71,79-81,88,102]
Receiver operating characteristic	7 (10.61)	[60,63,66,74,89,94,98]
Mean absolute error	7 (10.61)	[43,44,56,68,70,98,102]
Root-mean-square error	6 (9.09)	[38,53,61,68,80,102]
Intraclass correlation coefficient	4 (6.06)	[38,54,77,93]
Cohen κ	3 (3.03)	[77,95,102]
Equal error rate	2 (3.03)	[47,52]
G-mean	2 (3.03)	[51,52]
Negative predictive value	2 (3.03)	[52,65]
Coefficient of determination	2 (3.03)	[70,98]
Others	9 (13.64)	[38,52,54,59,66,67,82,88,99]

[a]ML: machine learning.

[b]DL: deep learning.

[c]The number of studies does not add up, as many studies have used >1 task type.

[d]The number of studies does not add up, as many studies have used >1 AI algorithm.

[e]The number of studies does not add up, as many studies have used >1 validation approach.

[f]The number of studies does not add up, as most studies used >1 performance measure.

A total of 30 distinct AI algorithms were identified across the included studies. The most frequently used were support vector machines (SVMs; 28/66, 42.42%), followed by convolutional neural networks (CNNs; 23/66, 34.85%), random forests (19/66, 28.79%), logistic regression (15/66, 22.73%), decision trees (14/66, 21.21%), long short-term memory (LSTM) networks (11/66, 16.67%), and multilayer perceptrons (8/66, 12.12%; Table 3).

As shown in Figure 4, SVMs and CNNs were predominantly applied in the “monitoring or assessment” and “state recognition or functional screening” categories. Random forests, decision trees, and logistic regression models were distributed across application purposes but were mainly used for monitoring- or assessment-related tasks. Temporal and probabilistic models, such as LSTM networks and hidden Markov models, were more often observed in state recognition or functional screening tasks, although their overall adoption remained limited. In contrast, algorithmic applications in the “prediction” and “rehabilitation or feedback” categories were comparatively sparse, indicating that prospective and closed-loop applications remain at an early stage of development.

XSL•FO
RenderX

Figure 4. Distribution of artificial intelligence (AI) algorithms and application objectives in the included Parkinson disease wearable-device studies.

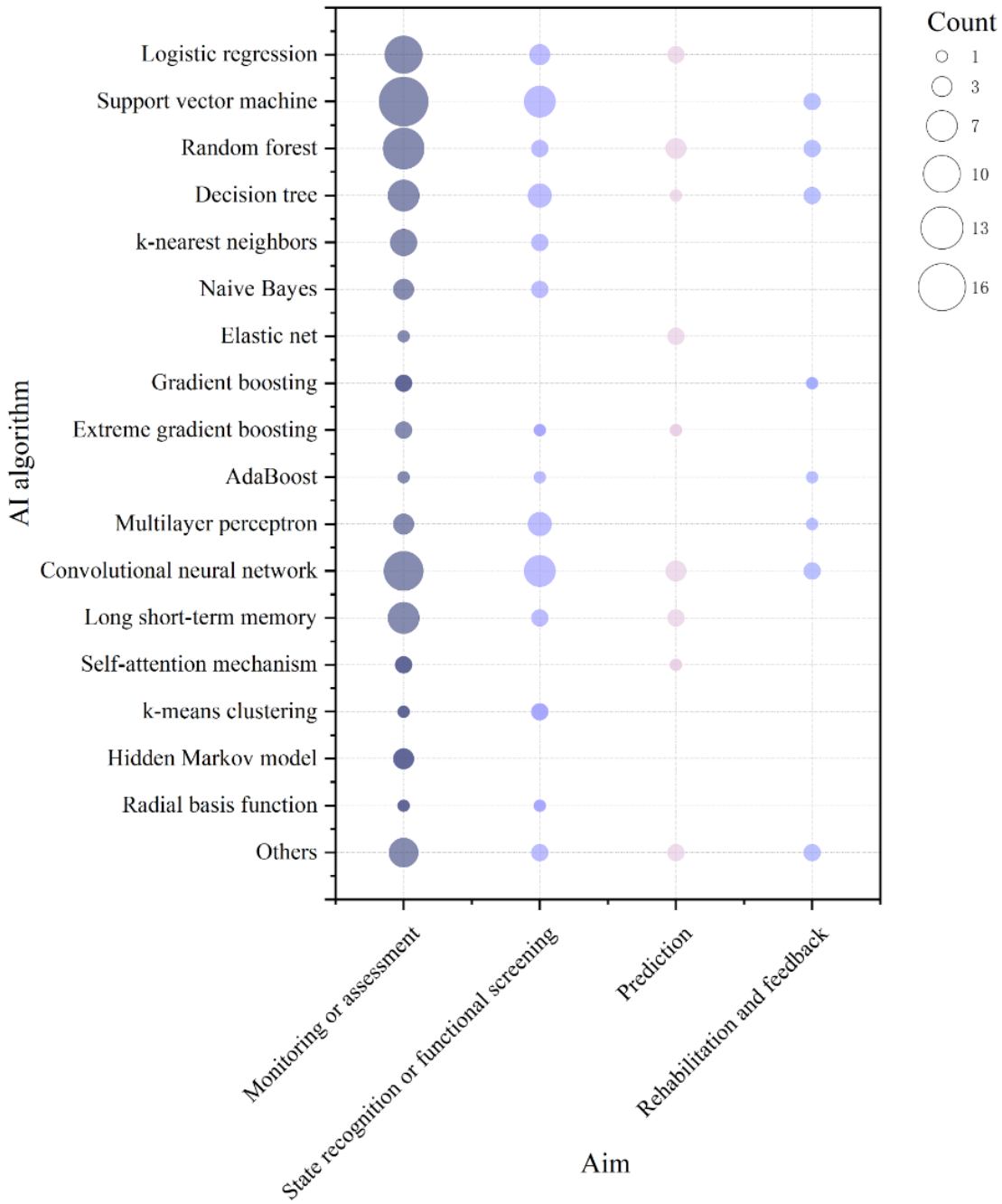

A total of 4 different model validation approaches were used across the included studies, with approximately 31.82% (21/66) using more than 1 method (Table 3). Leave-one-out cross-validation was the most frequently used approach (37/66, 56.06%), followed by k-fold cross-validation (27/66, 40.91%), hold-out validation (18/66, 27.27%), and external validation (4/66, 6.06%). A total of 25 performance metrics were used to evaluate model performance. The most commonly reported metrics were sensitivity (43/66, 65.15%), accuracy (41/66, 62.12%), specificity (27/66, 40.91%), F_1-score (27/66, 40.91%), area under the curve (25/66, 37.88%), precision (24/66, 36.36%), and correlation coefficient (14/66, 21.21%). Detailed characteristics of the AI algorithms applied in each study are provided in Multimedia Appendix 8 [38-103].

Challenges and Opportunities of Wearable Devices

Based on a synthesis of the included evidence, we developed a conceptually integrated "challenges and opportunities" framework grounded in the evidence base. This framework categorizes the challenges and opportunities of AI-enabled wearable devices in clinical applications into six dimensions: (1) evidence and data quality; (2) technical limitations; (3) usability, adherence, and equity; (4) economic and policy barriers; (5) privacy, security, and data governance; and (6)

XSL•FO
RenderX

clinical translation and workflow integration. The framework is intended to inform future research and practice rather than to represent an empirically validated implementation pathway, as illustrated in Figure 5 and detailed in Multimedia Appendix 9.

Figure 5. From monitoring to closed-loop rehabilitation: key opportunity domains for artificial intelligence (AI)–enabled wearable devices in Parkinson disease.

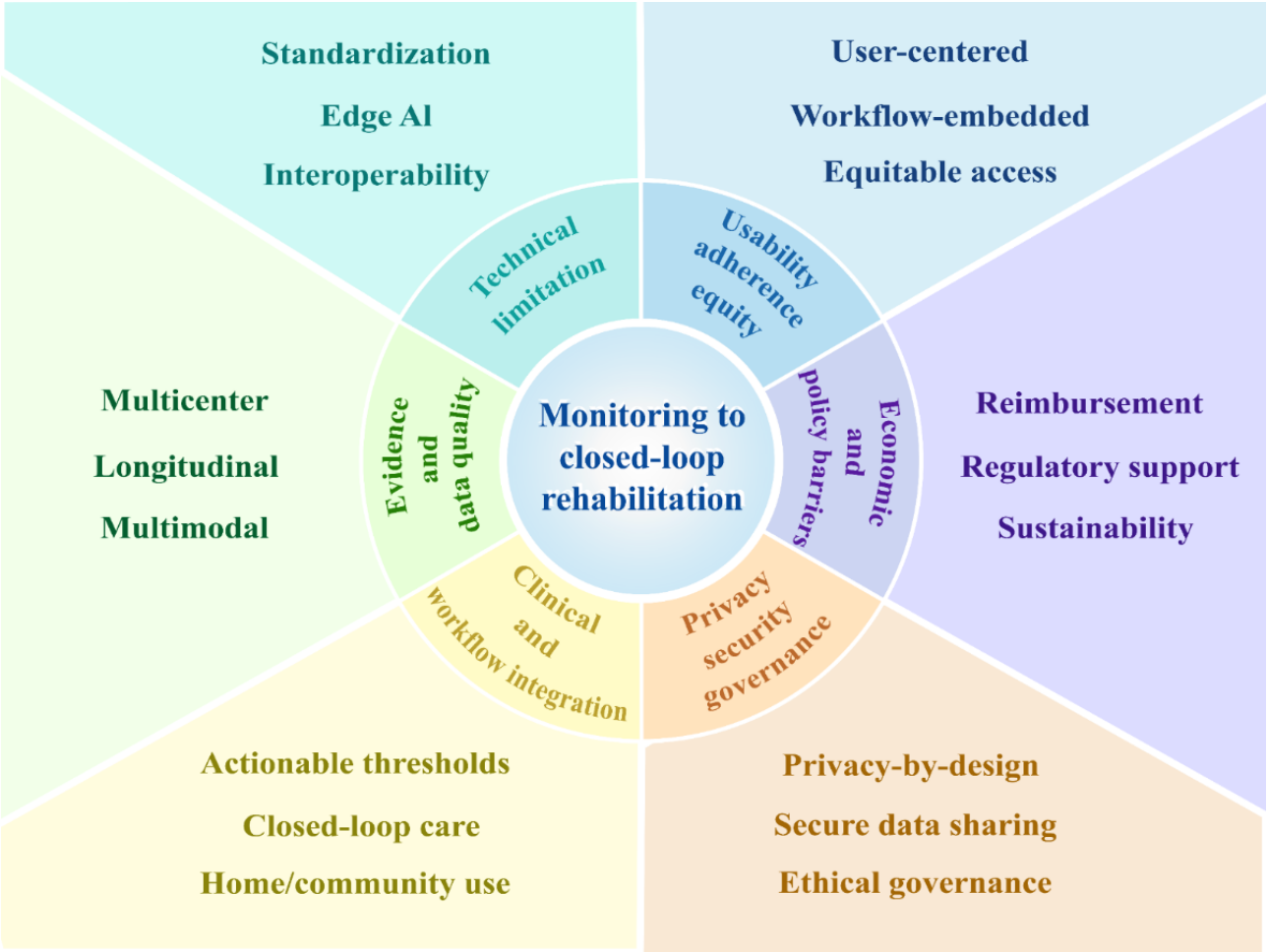

Discussion

Principal Findings

Overview

We synthesized the evidence on AI-enabled wearable devices for PD rehabilitation and motor function assessment, focusing on device types, monitoring indicators, algorithmic approaches, and application characteristics. Overall, the available evidence remains predominantly monitoring- and assessment-oriented, with relatively limited rehabilitation interventions and workflow-integrated applications. Key translational gaps persist, including limited evidence from real-world home/community settings, scarce external validation, underreporting of model calibration and clinical use, and underrepresentation of diverse populations.

Most prior reviews were published between 2020 and 2023 [104-106], and may not fully capture more recent developments, including the wider adoption of remote monitoring devices [107,108], rapid advances in commercial multisensor systems [109], and the evolution of AI models [110]. In contrast to earlier reviews that primarily focused on diagnosis or device technical performance, the present review adopts a rehabilitation- and nursing-oriented perspective and foregrounds evidence relevant to real-world implementation and nurse-led translation. Building on these findings, we propose a conceptual closed-loop framework linking monitoring, assessment, intervention, feedback, and reevaluation, which can be tested and refined in future research.

Study Characteristics

This review covered the 2020-2025 publication window and found that the volume of studies remained consistently high in recent years, suggesting that the field has entered a relatively stable phase of development rather than a period of rapid expansion, consistent with previous reviews [106]. Nevertheless, notable limitations in methodological rigor and sampling design persist: (1) sample sizes varied substantially, with mean values often influenced by outliers [48,80]; (2) sampling was typically based on single-center convenience cohorts [41,64,96]; and (3) subgroup analyses by sex, age, or special populations (such as early-onset or comorbid PD) were largely absent [41,83,84,96]. Many studies enrolled only people with PD, with relatively few including healthy control groups. Outcome reporting tended to prioritize discrimination metrics (eg, accuracy), while model calibration, uncertainty quantification, and clinically actionable thresholding or decision-curve analysis were rarely reported

XSL•FO
RenderX

[57,63,64,72,94-96]. In addition, research has been concentrated in Europe and North America (Table 1), with comparatively limited evidence from low- and middle-income countries and insufficient attention to home/community settings, thereby constraining generalizability and real-world applicability. Overall, persistent methodological limitations—including small, single-center samples, predominantly internal validation, and limited evidence from home/community settings—suggest that growth in study quantity has outpaced advances in translational value.

Application Characteristics and Signal Focus

In terms of application characteristics, AI-enabled wearable devices in PD show a marked imbalance between assessment-oriented and rehabilitation-oriented work, consistent with prior systematic reviews describing the field as "detection-heavy but prediction- and intervention-light" [104-106]. Although several representative studies have demonstrated the feasibility of continuous sensor-based quantification of tremor and bradykinesia [53], objective quantification of medication response [38], and sensor-based early detection of freezing of gait [54,55], most remain at a proof-of-concept stage, with limited evidence for real-world, workflow-integrated, or actionable use in routine care [104,105]. This disparity likely reflects inherent differences in methodological complexity. Assessment-focused studies typically rely on passive data capture and brief, structured tasks that are easier to implement in controlled settings, whereas rehabilitation studies require longer follow-up, personalized feedback, adherence management, and integration with clinical workflows—substantially raising both technical and methodological demands [13,111]. Moreover, most studies adopt motion capture–oriented wearable devices centered on inertial and/or pressure signals, with sensors preferentially placed on the lower limbs or the wrist. This design choice reduces data-collection burden and improves task reproducibility, thereby reinforcing assessment paradigms built around controlled tasks; however, it also structurally steers research toward motor outcomes (eg, gait, freezing of gait, tremor, and bradykinesia), making sleep, mood, and cognition more difficult to capture continuously and comprehensively [38,46,53]. Meanwhile, many applications still rely on short, structured assessments, which further constrain the continuous capture of nonmotor domains in real-world contexts [16,19]. Taken together, the evidence continues to prioritize motor-domain tasks, with insufficient coverage of nonmotor symptoms and their continuous assessment in real-world contexts. As a result, current applications fall short of meeting the comprehensive needs of long-term chronic disease management.

AI and Algorithmic Characteristics

In our review, SVMs and CNNs are the most frequently used approaches, while decision trees and logistic regression are commonly used as interpretable baseline models. Currently, the field reflects a pattern of "traditional ML dominance, parallel advances in convolutional DL, and early-stage development of sequential models and emerging paradigms" [38,40,50,57,68,74], which aligns with the ML/DL distribution observed in our results. This distribution may be attributable to differences in how algorithms adapt to data scale and structure. SVMs tend to provide more stable performance in small-sample, feature-engineered wearable data settings, whereas CNNs are better suited to end-to-end feature learning from high-frequency inertial signals, thereby reducing reliance on manual feature engineering [112,113]. Sequential approaches (eg, LSTM networks and a small number of HMMs) have been explored for gait-phase and state-recognition tasks, but remain relatively uncommon. This suggests that temporal dependencies and state-transition dynamics are still underutilized [36,52,72,73]. Consistent with these constraints, validation remains largely internal (eg, cross-validation), while independent external validation is rare, limiting confidence in generalizability [105]. More complex architectures (eg, self-attention mechanisms and extreme gradient boosting) also appear underused, potentially reflecting feasibility constraints for on-device or real-world deployment [39,75,80,92-96,98].

In summary, AI-enabled wearable devices for PD rehabilitation and assessment remain characterized by a strong emphasis on monitoring, with comparatively weak integration into closed-loop systems. From the perspective of the *JMIR mHealth and uHealth* taxonomy, most devices included in this review are better characterized as mobile health tools, primarily focused on individual-level monitoring and assessment [114]. Although these functions provide valuable objective data, there is still insufficient evidence that they have been systematically embedded into routine nurse-led rehabilitation workflows or broader health care delivery systems as mature eHealth solutions. Building on these findings, future research should move beyond stand-alone monitoring applications toward workflow-integrated eHealth systems with decision-support functions that can be tested in nurse-led rehabilitation pathways.

Practical Recommendations

Implications for Clinical and Nursing Practice

In both clinical and home-based rehabilitation contexts, nurses increasingly require continuous and objective patient information to enhance the precision of assessment and monitoring, and strengthen patient engagement in the rehabilitation process [115]. AI-enabled wearable devices represent a promising approach to addressing this need. Within clinical rehabilitation settings, continuous monitoring of gait, tremor, sleep, and activity levels can augment routine nursing surveillance, facilitate earlier detection of functional deterioration, and support timely intervention or prioritization—moving beyond the inherently fragmented nature of episodic in-clinic observations [25,115]. In remote or home-based rehabilitation, the continuous and objective data generated by wearables can be used to evaluate therapeutic response and adherence, while personalized feedback, symptom prompts, medication-taking support, and nurse-led follow-up via telerehabilitation platforms promote patient self-management [26,115]. Specifically, when algorithmic detection identifies abnormal changes in patients' daily activity levels or tremor severity that exceed an individualized threshold, an alert can be triggered on the nursing interface. This may prompt nurses to rapidly assess symptom changes, medication adherence, and

XSL•FO
RenderX

potential adverse effects and, as appropriate, arrange earlier review, reinforce medication and safety guidance, or adjust follow-up plans [97,115]. Likewise, when long-term monitoring trend reports indicate sustained deterioration in motor function–related indicators or marked fluctuations in motor symptoms, these data can provide objective evidence for the rehabilitation team to deliver targeted health education, adjust the frequency of follow-up, and, when needed, coordinate involvement of rehabilitation therapists or a multidisciplinary team [56,89,99]. These functions align closely with core nursing responsibilities in patient education, symptom monitoring, and long-term chronic disease management, and may help reinforce the full nursing care cycle of assessment, diagnosis, planning, implementation, and evaluation [116].

Priorities for Future Research and Translation

To advance the field, future research may consider five priority areas. (1) Strengthening the evidence base through multicenter, long-term real-world studies with preregistration, appropriate sample size estimation, and attrition management, alongside reporting of external validation, model calibration, uncertainty, and clinical use [117-120]. (2) Enhancing model feasibility through multimodal signal fusion and data-efficient learning strategies to support practical real-world deployment [104]. (3) Exploring conceptual closed-loop rehabilitation frameworks that integrate assessment and intervention, while examining the feasibility of actionable thresholds and their potential integration into routine clinical workflows [120]. (4) Improving device comfort and usability to balance patient adherence with caregiver and clinician workload, address the digital divide, and enhance accessibility [119,120]. (5) Strengthening privacy and data security frameworks, alongside improving insurance reimbursement mechanisms and regulatory support, to promote responsible cross-institutional data sharing [121,122], including explicitly addressing equity, responsible data use, and related ethical considerations in real-world nursing and rehabilitation contexts.

Limitations

This study has several limitations. First, no prospective protocol was registered; however, this is an acceptable practice for scoping reviews, and we clarify this point to enhance methodological transparency. Second, although a systematic search and screening process was conducted, the scope of databases, search strategies, and inclusion criteria may have led to the omission of eligible studies, posing a risk of incompleteness or potential bias. In addition, most included studies had small sample sizes, single-center designs, short follow-up durations, and limited external validation, alongside substantial heterogeneity across study populations, task paradigms, signal sources, algorithmic approaches, and outcome measures. This heterogeneity may limit cross-study comparability and the generalizability of findings. Given the predominantly exploratory and proof-of-concept nature of the included studies, the closed-loop rehabilitation framework and related recommendations proposed in this review should be interpreted as conceptual and practice-informing rather than empirically validated implementation pathways. Finally, the search was restricted to English and Chinese peer-reviewed publications, with limited inclusion of gray literature or other languages, which may introduce time-lag, linguistic, and publication biases. In particular, studies reporting null or negative findings, early-stage failures, or limited model performance may be less likely to be published and therefore underrepresented, potentially leading to an overestimation of the maturity, effectiveness, and translational readiness of AI-enabled wearable technologies.

Conclusions

This study systematically summarized recent evidence on AI-enabled wearable devices for motor function assessment and rehabilitation in people with PD. By combining a scoping review with an evidence map and adopting a rehabilitation-oriented perspective, we synthesized evidence across multiple dimensions, including application aims, sensing modalities, AI methods, and validation practices, thereby identifying key translational gaps between proof-of-concept studies and real-world rehabilitation workflows. Compared with previous reviews that primarily focused on monitoring functions or device performance, this study places greater emphasis on rehabilitation applications and nurse-led translation into practice, and further clarifies the key actionable gaps that limit real-world implementation. On this basis, a conceptually integrated “challenges and opportunities” framework is proposed to inform the design, evaluation, and reporting of devices and algorithms, and to highlight considerations for developing workflow-integrated, decision-support wearable systems. From a real-world perspective, these findings may support continuity of rehabilitation across clinical, home, and community settings by enabling nurses to deliver continuous monitoring, personalized follow-up, and timely intervention, ultimately improving the efficiency and accessibility of PD rehabilitation management.

Acknowledgments

Generative artificial intelligence tools (ChatGPT [OpenAI] and DeepSeek [Hangzhou DeepSeek Artificial Intelligence Co, Ltd]) were used solely under full human supervision to assist with proofreading/editing and translation, including terminology conversion, textual clarification, and translation of reviewer comments. These tools were not used for study design, data processing, analysis, or the generation of scientific content. All AI-assisted outputs were reviewed and revised by the authors, who take full responsibility for the accuracy and integrity of the manuscript.

XSL•FO
RenderX

Data Availability

This scoping review did not generate any primary data. All data extracted and synthesized in this review were obtained from previously published studies, which are publicly available. No additional datasets were created or analyzed.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR checklist.
[DOCX File , 71 KB - jmir_v28i1e85596_app1.docx]

Multimedia Appendix 2
Eligibility criteria.
[DOCX File , 19 KB - jmir_v28i1e85596_app2.docx]

Multimedia Appendix 3
Search strategy.
[DOCX File , 30 KB - jmir_v28i1e85596_app3.docx]

Multimedia Appendix 4
Extracted data form.
[DOCX File , 19 KB - jmir_v28i1e85596_app4.docx]

Multimedia Appendix 5
Quality assessment of the included studies.
[DOCX File , 72 KB - jmir_v28i1e85596_app5.docx]

Multimedia Appendix 6
Detailed characteristics of included studies using artificial intelligence–enabled wearable devices in Parkinson disease research.
[DOCX File , 50 KB - jmir_v28i1e85596_app6.docx]

Multimedia Appendix 7
Detailed characteristics of artificial intelligence–enabled wearable devices used in the included Parkinson disease studies.
[DOCX File , 59 KB - jmir_v28i1e85596_app7.docx]

Multimedia Appendix 8
Detailed characteristics of artificial intelligence algorithms used in the included wearable device studies for Parkinson disease.
[DOCX File , 50 KB - jmir_v28i1e85596_app8.docx]

Multimedia Appendix 9
Challenges and opportunities of artificial intelligence–enabled wearable devices in Parkinson disease research.
[DOCX File , 20 KB - jmir_v28i1e85596_app9.docx]

References

1. Kouli A, Torsney KM, Kuan WL. Parkinson’s disease: etiology, neuropathology, and pathogenesis. In: Stoker TB, Greenland JC, editors. Parkinson’s Disease: Pathogenesis and Clinical Aspects Internet. Brisbane (AU): Codon Publications; 2018.
2. Abusrair AH, Elsekaily W, Bohlega S. Tremor in Parkinson's disease: from pathophysiology to advanced therapies. Tremor Other Hyperkinet Mov (NY) 2022;12:29 [FREE Full text] [doi: 10.5334/tohm.712] [Medline: 36211804]
3. Dorsey ER, Sherer T, Okun MS, Bloem BR. The emerging evidence of the Parkinson pandemic. J Parkinsons Dis 2018;8(s1):S3-S8 [FREE Full text] [doi: 10.3233/JPD-181474] [Medline: 30584159]
4. Qi S, Yin P, Wang L, Qu M, Kan GL, Zhang H, et al. Prevalence of Parkinson's disease: a community-based study in China. Mov Disord 2021;36(12):2940-2944. [doi: 10.1002/mds.28762] [Medline: 34390510]
5. Zhang ST, Jia XF, Zhang JG. Research progress on the association between dietary nutrition, lifestyle, and Parkinson's disease [in Chinese]. J Environ Occup Med 2025;40(2):149-155 [FREE Full text]
6. Garg D, Dhamija RK. Rehabilitation in Parkinson's disease: current status and future directions. Ann Mov Disord 2020;3(2):79-85. [doi: 10.4103/aomd.aomd_1_20]

XSL•FO
RenderX

7. Yoon SY. Update on Parkinson's disease rehabilitation. Brain Neurorehabil 2022;15(2):e15 [FREE Full text] [doi: 10.12786/bn.2022.15.e15] [Medline: 36743207]
8. Zhang L, Liu Y, Han W. Advance of rehabilitation for Parkinson's disease [in Chinese]. Chin J Contemp Neurol Neurosurg 2017;17(5):328-332 [FREE Full text]
9. Pastana Ramos LF, Vilacorta-Pereira TDCS, Duarte JDS, Yamada ES, Santos-Lobato BL. Feasibility and effectiveness of a remote individual rehabilitation program for people with Parkinson's disease living in the Brazilian Amazon: a randomized clinical trial. Front Neurol 2023;14:1244661 [FREE Full text] [doi: 10.3389/fneur.2023.1244661] [Medline: 37693755]
10. Sevcenko K, Lindgren I. The effects of virtual reality training in stroke and Parkinson's disease rehabilitation: a systematic review and a perspective on usability. Eur Rev Aging Phys Act 2022;19(1):4 [FREE Full text] [doi: 10.1186/s11556-022-00283-3] [Medline: 35078401]
11. Fedor S, Lewis R, Pedrelli P, Mischoulon D, Curtiss J, Picard RW. Wearable technology in clinical practice for depressive disorder. N Engl J Med 2023;389(26):2457-2466. [doi: 10.1056/NEJMra2215898] [Medline: 38157501]
12. Maceira-Elvira P, Popa T, Schmid AC, Hummel FC. Wearable technology in stroke rehabilitation: towards improved diagnosis and treatment of upper-limb motor impairment. J Neuroeng Rehabil 2019;16(1):142 [FREE Full text] [doi: 10.1186/s12984-019-0612-y] [Medline: 31744553]
13. Porciuncula F, Roto AV, Kumar D, Davis I, Roy S, Walsh CJ, et al. Wearable movement sensors for rehabilitation: a focused review of technological and clinical advances. PM R 2018;10(9 Suppl 2):S220-S232 [FREE Full text] [doi: 10.1016/j.pmrj.2018.06.013] [Medline: 30269807]
14. Rodríguez-Fernández A, Lobo-Prat J, Font-Llagunes JM. Systematic review on wearable lower-limb exoskeletons for gait training in neuromuscular impairments. J Neuroeng Rehabil 2021;18(1):22 [FREE Full text] [doi: 10.1186/s12984-021-00815-5] [Medline: 33526065]
15. Meena JS, Choi SB, Jung S, Kim J. Electronic textiles: new age of wearable technology for healthcare and fitness solutions. Mater Today Bio 2023;19:100565 [FREE Full text] [doi: 10.1016/j.mtbio.2023.100565] [Medline: 36816602]
16. Huang X, Xue Y, Ren S, Wang F. Sensor-based wearable systems for monitoring human motion and posture: a review. Sensors (Basel) 2023;23(22):9047 [FREE Full text] [doi: 10.3390/s23229047] [Medline: 38005436]
17. Adams JL, Dinesh K, Snyder CW, Xiong M, Tarolli CG, Sharma S, et al. A real-world study of wearable sensors in Parkinson's disease. NPJ Parkinsons Dis 2021;7(1):106 [FREE Full text] [doi: 10.1038/s41531-021-00248-w] [Medline: 34845224]
18. Lu R, Xu Y, Li X, Fan Y, Zeng W, Tan Y, et al. Evaluation of wearable sensor devices in Parkinson's disease: a review of current status and future prospects. Parkinsons Dis 2020;2020:4693019 [FREE Full text] [doi: 10.1155/2020/4693019] [Medline: 33029343]
19. Bougea A. Application of wearable sensors in Parkinson’s disease: state of the art. JSAN 2025;14(2):23. [doi: 10.3390/jsan14020023]
20. Sotirakis C, Su Z, Brzezicki MA, Conway N, Tarassenko L, FitzGerald JJ, et al. Identification of motor progression in Parkinson's disease using wearable sensors and machine learning. NPJ Parkinsons Dis 2023;9(1):142 [FREE Full text] [doi: 10.1038/s41531-023-00581-2] [Medline: 37805655]
21. Abujrida H, Agu E, Pahlavan K. Machine learning-based motor assessment of Parkinson's disease using postural sway, gait and lifestyle features on crowdsourced smartphone data. Biomed Phys Eng Express 2020;6(3):035005. [doi: 10.1088/2057-1976/ab39a8] [Medline: 33438650]
22. Benyoucef Y, Melliti I, Harmouch J, Asadi B, Del Mastro A, Lapuente-Hernández D, et al. Wearable sensors and artificial intelligence for the diagnosis of Parkinson's disease. J Clin Med 2025;14(12):4207 [FREE Full text] [doi: 10.3390/jcm14124207] [Medline: 40565952]
23. Farabolini G, Baldini N, Pagano A, Andrenelli E, Pepa L, Morone G, et al. Continuous movement monitoring at home through wearable devices: a systematic review. Sensors (Basel) 2025;25(16):4889 [FREE Full text] [doi: 10.3390/s25164889] [Medline: 40871752]
24. Islam MS, Rahman W, Abdelkader A, Lee S, Yang PT, Purks JL, et al. Using AI to measure Parkinson's disease severity at home. NPJ Digit Med 2023;6(1):156 [FREE Full text] [doi: 10.1038/s41746-023-00905-9] [Medline: 37608206]
25. Moreau C, Rouaud T, Grabli D, Benatru I, Remy P, Marques A, et al. Overview on wearable sensors for the management of Parkinson's disease. NPJ Parkinsons Dis 2023;9(1):153 [FREE Full text] [doi: 10.1038/s41531-023-00585-y] [Medline: 37919332]
26. Del Din S, Kirk C, Yarnall AJ, Rochester L, Hausdorff JM. Body-worn sensors for remote monitoring of Parkinson's disease motor symptoms: vision, state of the art, and challenges ahead. J Parkinsons Dis 2021;11(s1):S35-S47 [FREE Full text] [doi: 10.3233/JPD-202471] [Medline: 33523020]
27. Li H, Zecca M, Huang J. Evaluating the utility of wearable sensors for the early diagnosis of Parkinson disease: systematic review. J Med Internet Res 2025;27:e69422 [FREE Full text] [doi: 10.2196/69422] [Medline: 40690754]
28. Gutowski T, Stodulska O, Ćwiklińska A, Gutowska K, Kopeć K, Betka M, et al. Machine learning-based assessment of Parkinson's disease symptoms using wearable and smartphone sensors. Sensors (Basel) 2025;25(16):4924 [FREE Full text] [doi: 10.3390/s25164924] [Medline: 40871788]

XSL•FO
RenderX

29. Wu X, Ma L, Wei P, Shan Y, Chan P, Wang K, et al. Wearable sensor devices can automatically identify the ON-OFF status of patients with Parkinson's disease through an interpretable machine learning model. Front Neurol 2024;15:1387477 [FREE Full text] [doi: 10.3389/fneur.2024.1387477] [Medline: 38751881]
30. Tischendorf T, Heitmann-Möller A, Ruppert S, Marchwacka M, Schaffrin S, Schaal T, et al. Sustainable integration of digitalisation in nursing education-an international scoping review. Front Health Serv 2024;4:1344021 [FREE Full text] [doi: 10.3389/frhs.2024.1344021] [Medline: 38665930]
31. AlDossary MBD, Al-Enezi MAH, Aldosary MM. The new frontier in care: a systematic review of nursing implications of remote patient monitoring (RPM) for managing chronic diseases. Saudi J Med Public Health 2024;1(2):582-589 [FREE Full text] [doi: 10.64483/jmph-167]
32. Tricco AC, Antony J, Zarin W, Strifler L, Ghassemi M, Ivory J, et al. A scoping review of rapid review methods. BMC Med 2015;13:224 [FREE Full text] [doi: 10.1186/s12916-015-0465-6] [Medline: 26377409]
33. Miake-Lye IM, Hempel S, Shanman R, Shekelle PG. What is an evidence map? A systematic review of published evidence maps and their definitions, methods, and products. Syst Rev 2016;5:28 [FREE Full text] [doi: 10.1186/s13643-016-0204-x] [Medline: 26864942]
34. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
35. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
36. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions. Chichester (UK): John Wiley & Sons; 2008.
37. Hong QN, Gonzalez-Reyes A, Pluye P. Improving the usefulness of a tool for appraising the quality of qualitative, quantitative and mixed methods studies, the Mixed Methods Appraisal Tool (MMAT). J Eval Clin Pract 2018;24(3):459-467. [doi: 10.1111/jep.12884] [Medline: 29464873]
38. Aghanavesi S, Westin J, Bergquist F, Nyholm D, Askmark H, Aquilonius SM, et al. A multiple motion sensors index for motor state quantification in Parkinson's disease. Comput Methods Programs Biomed 2020;189:105309. [doi: 10.1016/j.cmpb.2019.105309] [Medline: 31982667]
39. Li Y, Bai Q, Yang X, Zhou X, Sun Y, Yao Z. An abnormal gait monitoring system for patients with Parkinson's disease based on wearable devices. Piscataway, NJ: IEEE; 2022 Presented at: 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI); 2022 November 05-07; Beijing, China p. -15. [doi: 10.1109/cisp-bmei56279.2022.9980005]
40. Ghayvat H, Awais M, Geddam R, Quasim MT, Khowaja SA, Dev K. AiCareGaitRehabilitation: multi-modalities sensor data fusion for AI-IoT enabled realtime electrical stimulation device for pre-fog and post-fog to person with Parkinson's disease. Information Fusion 2025;122:103155. [doi: 10.1016/j.inffus.2025.103155]
41. Delgado-Terán JD, Hilbrants K, Mahmutović D, Silva de Lima AL, Wezel RJAV, Heida T. Ankle sensor-based detection of freezing of gait in Parkinson's disease in semi-free living environments. Sensors (Basel) 2025;25(6):1895 [FREE Full text] [doi: 10.3390/s25061895] [Medline: 40293010]
42. Liu X, Zhang X, Li J, Pan W, Sun Y, Lin S, et al. Automated UPDRS gait scoring using wearable sensor fusion and deep learning. Bioengineering (Basel) 2025;12(7):686 [FREE Full text] [doi: 10.3390/bioengineering12070686] [Medline: 40722378]
43. Han Y, Liu X, Zhang N, Zhang X, Zhang B, Wang S, et al. Automatic assessments of Parkinsonian gait with wearable sensors for human assistive systems. Sensors (Basel) 2023;23(4):2104 [FREE Full text] [doi: 10.3390/s23042104] [Medline: 36850705]
44. Peraza LR, Kinnunen KM, McNaney R, Craddock IJ, Whone AL, Morgan C, et al. An automatic gait analysis pipeline for wearable sensors: a pilot study in Parkinson's disease. Sensors (Basel) 2021;21(24):8286 [FREE Full text] [doi: 10.3390/s21248286] [Medline: 34960379]
45. Sánchez-Fernández LP, Sánchez-Pérez LA, Martínez-Hernández JM. Computer model for gait assessments in Parkinson's patients using a fuzzy inference model and inertial sensors. Artif Intell Med 2025;160:103059. [doi: 10.1016/j.artmed.2024.103059] [Medline: 39732050]
46. Channa A, Popescu N, Faisal M. Parkinson's disease gait evaluation leveraging wearable insoles and a deep learning approach. : IEEE; 2022 Presented at: 8th International Conference on Control, Decision and Information Technologies (CoDIT); 2022 May 17-20; Istanbul, Turkey p. 543-549. [doi: 10.1109/codit55151.2022.9804064]
47. Sigcha L, Costa N, Pavón I, Costa S, Arezes P, López JM, et al. Deep learning approaches for detecting freezing of gait in Parkinson's disease patients through on-body acceleration sensors. Sensors (Basel) 2020;20(7):1895 [FREE Full text] [doi: 10.3390/s20071895] [Medline: 32235373]
48. Dvorani A, Wiesener C, Salchow-Hommen C, Jochner M, Spieker L, Skrobot M, et al. On-demand gait-synchronous electrical cueing in Parkinson's disease using machine learning and edge computing: a pilot study. IEEE Open J Eng Med Biol 2024;5:306-315 [FREE Full text] [doi: 10.1109/OJEMB.2024.3390562] [Medline: 38766539]

XSL•FO
RenderX

49. Bikias T, Iakovakis D, Hadjidimitriou S, Charisis V, Hadjileontiadis LJ. DeepFoG: an IMU-based detection of freezing of gait episodes in Parkinson's disease patients via deep learning. Front Robot AI 2021;8:537384 [FREE Full text] [doi: 10.3389/frobt.2021.537384] [Medline: 34113654]
50. Aich S, Pradhan PM, Chakraborty S, Kim HC, Kim HT, Lee HG, et al. Design of a machine learning-assisted wearable accelerometer-based automated system for studying the effect of dopaminergic medicine on gait characteristics of Parkinson's patients. J Healthc Eng 2020;2020:1823268 [FREE Full text] [doi: 10.1155/2020/1823268] [Medline: 32148741]
51. Shi B, Tay A, Au WL, Tan DML, Chia NSY, Yen S. Detection of freezing of gait using convolutional neural networks and data from lower limb motion sensors. IEEE Trans Biomed Eng 2022;69(7):2256-2267. [doi: 10.1109/TBME.2022.3140258] [Medline: 34986092]
52. Ezhilarasi J, Senthil Kumar T. Develop a novel, faster mask region-based convolutional neural network model with leave-one-subject-out to predict freezing of gait abnormalities of Parkinson’s disease. Neural Comput Applic 2025;37(7):5441-5457. [doi: 10.1007/s00521-024-10832-9]
53. Mahadevan N, Demanuele C, Zhang H, Volfson D, Ho B, Erb MK, et al. Development of digital biomarkers for resting tremor and bradykinesia using a wrist-worn wearable device. NPJ Digit Med 2020;3:5 [FREE Full text] [doi: 10.1038/s41746-019-0217-7] [Medline: 31970290]
54. Chan LLY, Yang S, Aswani M, Kark L, Henderson E, Lord SR, et al. Development, validation, and limits of freezing of gait detection using a single waist-worn device. IEEE Trans Biomed Eng 2024;71(10):3024-3031. [doi: 10.1109/TBME.2024.3407059] [Medline: 38814761]
55. Pardoel S, Shalin G, Nantel J, Lemaire ED, Kofman J. Early detection of freezing of gait during walking using inertial measurement unit and plantar pressure distribution data. Sensors (Basel) 2021;21(6):2246 [FREE Full text] [doi: 10.3390/s21062246] [Medline: 33806984]
56. Hssayeni MD, Jimenez-Shahed J, Burack MA, Ghoraani B. Ensemble deep model for continuous estimation of unified Parkinson's disease rating scale III. Biomed Eng Online 2021;20(1):32 [FREE Full text] [doi: 10.1186/s12938-021-00872-w] [Medline: 33789666]
57. Tsakanikas V, Ntanis A, Rigas G, Androutsos C, Boucharas D, Tachos N, et al. Evaluating gait impairment in Parkinson's disease from instrumented insole and IMU sensor data. Sensors (Basel) 2023;23(8):3902 [FREE Full text] [doi: 10.3390/s23083902] [Medline: 37112243]
58. Brand YE, Schwartz D, Gazit E, Buchman AS, Gilad-Bachrach R, Hausdorff JM. Gait detection from a wrist-worn sensor using machine learning methods: daily living study in older adults and people with Parkinson's disease. Sensors (Basel) 2022;22(18):7094 [FREE Full text] [doi: 10.3390/s22187094] [Medline: 36146441]
59. Pardoel S, Shalin G, Lemaire ED, Kofman J, Nantel J. Grouping successive freezing of gait episodes has neutral to detrimental effect on freeze detection and prediction in Parkinson's disease. PLoS One 2021;16(10):e0258544 [FREE Full text] [doi: 10.1371/journal.pone.0258544] [Medline: 34637473]
60. Li D, Hallack A, Gwilym S, Li D, Hu MT, Cantley J. Investigating gait-responsive somatosensory cueing from a wearable device to improve walking in Parkinson's disease. Biomed Eng Online 2023;22(1):108 [FREE Full text] [doi: 10.1186/s12938-023-01167-y] [Medline: 37974260]
61. Aghanavesi S, Bergquist F, Nyholm D, Senek M, Memedi M. Motion sensor-based assessment of Parkinson's disease motor symptoms during leg agility tests: results from levodopa challenge. IEEE J Biomed Health Inform 2020;24(1):111-119. [doi: 10.1109/JBHI.2019.2898332] [Medline: 30763248]
62. Atrsaei A, Hansen C, Elshehabi M, Solbrig S, Berg D, Liepelt-Scarfone I, et al. Effect of fear of falling on mobility measured during lab and daily activity assessments in Parkinson's disease. Front Aging Neurosci 2021;13:722830 [FREE Full text] [doi: 10.3389/fnagi.2021.722830] [Medline: 34916920]
63. Shawen N, O'Brien MK, Venkatesan S, Lonini L, Simuni T, Hamilton JL, et al. Role of data measurement characteristics in the accurate detection of Parkinson's disease symptoms using wearable sensors. J Neuroeng Rehabil 2020;17(1):52 [FREE Full text] [doi: 10.1186/s12984-020-00684-4] [Medline: 32312287]
64. Naghavi N, Wade E. Towards real-time prediction of freezing of gait in patients with Parkinson's disease: a novel deep one-class classifier. IEEE J Biomed Health Inform 2022;26(4):1726-1736. [doi: 10.1109/JBHI.2021.3103071] [Medline: 34375292]
65. Zampogna A, Borzì L, Rinaldi D, Artusi CA, Imbalzano G, Patera M, et al. Unveiling the unpredictable in Parkinson's disease: sensor-based monitoring of dyskinesias and freezing of gait in daily life. Bioengineering (Basel) 2024;11(5):440 [FREE Full text] [doi: 10.3390/bioengineering11050440] [Medline: 38790307]
66. Reches T, Dagan M, Herman T, Gazit E, Gouskova NA, Giladi N, et al. Using wearable sensors and machine learning to automatically detect freezing of gait during a FOG-provoking test. Sensors (Basel) 2020;20(16):4474 [FREE Full text] [doi: 10.3390/s20164474] [Medline: 32785163]
67. Slemenšek J, Geršak J, Bratina B, van Midden VM, Pirtošek Z, Šafarič R. Wearable online freezing of gait detection and cueing system. Bioengineering (Basel) 2024;11(10):1048 [FREE Full text] [doi: 10.3390/bioengineering11101048] [Medline: 39451423]
68. Borzì L, Mazzetta I, Zampogna A, Suppa A, Irrera F, Olmo G. Predicting axial impairment in parkinson's disease through a single inertial sensor. Sensors (Basel) 2022;22(2):412 [FREE Full text] [doi: 10.3390/s22020412] [Medline: 35062375]

XSL•FO
RenderX

69. Wang W, Lin J, Le X, Li Y, Liu T, Pan L, et al. Addressing multiple challenges in early gait freezing prediction for Parkinson's disease: a practical deep learning approach. IEEE J Biomed Health Inform 2025;29(9):6251-6262. [doi: 10.1109/JBHI.2024.3522664] [Medline: 40030782]
70. Shuqair M, Jimenez-Shahed J, Ghoraani B. Advancing Parkinson's disease management through multi-shared-task self-supervised signal processing. : IEEE; 2024 Presented at: 58th Asilomar Conference on Signals, Systems, and Computers; 2024 October 27-30; Pacific Grove, CA, USA p. 957-961. [doi: 10.1109/ieeeconf60004.2024.10942713]
71. Kanellos FS, Tsamis KI, Rigas G, Simos YV, Katsenos AP, Kartsakalis G, et al. Clinical evaluation in Parkinson's disease: is the golden standard shiny enough? Sensors (Basel) 2023;23(8):3807 [FREE Full text] [doi: 10.3390/s23083807] [Medline: 37112154]
72. Roth N, Küderle A, Ullrich M, Gladow T, Marxreiter F, Klucken J, et al. Hidden Markov model based stride segmentation on unsupervised free-living gait data in Parkinson's disease patients. J Neuroeng Rehabil 2021;18(1):93 [FREE Full text] [doi: 10.1186/s12984-021-00883-7] [Medline: 34082762]
73. Perez-Ibarra JC, Siqueira AAG, Krebs HI. Identification of gait events in healthy subjects and with Parkinson's disease using inertial sensors: an adaptive unsupervised learning approach. IEEE Trans Neural Syst Rehabil Eng 2020;28(12):2933-2943. [doi: 10.1109/TNSRE.2020.3039999] [Medline: 33237863]
74. Meng L, Pang J, Yang Y, Chen L, Xu R, Ming D. Inertial-based gait metrics during turning improve the detection of early-stage Parkinson's disease patients. IEEE Trans Neural Syst Rehabil Eng 2023;31:1472-1482. [doi: 10.1109/TNSRE.2023.3237903] [Medline: 37021993]
75. Dattola S, Ielo A, Quartarone A, De Cola MC. Integrating wearable sensor signal processing with unsupervised learning methods for tremor classification in Parkinson's disease. Bioengineering (Basel) 2025;12(1):37 [FREE Full text] [doi: 10.3390/bioengineering12010037] [Medline: 39851311]
76. Zhao Y, Liu Y, Lu W, Li J, Shan P, Lian C, et al. Intelligent IoT anklets for monitoring the assessment of Parkinson's diseases. IEEE Sensors J 2023;23(24):31523-31536. [doi: 10.1109/jsen.2023.3331277]
77. Li Y, Yin J, Liu S, Xue B, Shokoohi C, Ge G, et al. Learning hand kinematics for Parkinson's disease assessment using a multimodal sensor glove. Adv Sci (Weinh) 2023;10(20):e2206982 [FREE Full text] [doi: 10.1002/advs.202206982] [Medline: 37150855]
78. Yan F, Gong J, Zhang Q, He H. Learning motion primitives for the quantification and diagnosis of mobility deficits. IEEE Trans Biomed Eng 2024;71(12):3339-3349. [doi: 10.1109/TBME.2024.3404357] [Medline: 38776195]
79. Evers LJW, Raykov YP, Heskes TM, Krijthe JH, Bloem BR, Little MA. Passive monitoring of Parkinson tremor in daily life: a prototypical network approach. Sensors (Basel) 2025;25(2):366 [FREE Full text] [doi: 10.3390/s25020366] [Medline: 39860736]
80. Greene BR, Premoli I, McManus K, McGrath D, Caulfield B. Predicting fall counts using wearable sensors: a novel digital biomarker for Parkinson's disease. Sensors (Basel) 2021;22(1):54 [FREE Full text] [doi: 10.3390/s22010054] [Medline: 35009599]
81. Ricci M, Lazzaro GD, Errico V, Pisani A, Giannini F, Saggio G. The impact of wearable electronics in assessing the effectiveness of levodopa treatment in Parkinson's disease. IEEE J Biomed Health Inform 2022;26(7):2920-2928. [doi: 10.1109/JBHI.2022.3160103] [Medline: 35316198]
82. Ghosh N, Banerjee I. IoT-based freezing of gait detection using grey relational analysis. Internet Things 2021;13:100068. [doi: 10.1016/j.iot.2019.100068]
83. Lin ZR. Quantitative assessment of Parkinson's motor symptoms based on machine learning. Sci Technol Innov Appl 2022;12(34):50-54. [doi: 10.19981/j.CN23-1581/G3.2022.34.013]
84. Sun Y, Rong Z, Wang F. Automatic quantitative study of motor function in patients with Parkinson's disease based on wearable inertial sensors [in Chinese]. China Med Equip 2023;38(10):27-32 [FREE Full text]
85. Li YM. Quantitative Assessment of Parkinson's Motor Symptoms Based on Deep Learning [dissertation]. Lanzhou, China: Lanzhou Jiaotong University; 2023.
86. Donisi L, Cesarelli G, Balbi P, Provitera V, Lanzillo B, Coccia A, et al. Positive impact of short-term gait rehabilitation in Parkinson patients: a combined approach based on statistics and machine learning. Math Biosci Eng 2021;18(5):6995-7009 [FREE Full text] [doi: 10.3934/mbe.2021348] [Medline: 34517568]
87. Sotirakis C, Brzezicki MA, Patel S, Conway N, FitzGerald JJ, Antoniades CA. Predicting future fallers in Parkinson's disease using kinematic data over a period of 5 years. NPJ Digit Med 2024;7(1):345 [FREE Full text] [doi: 10.1038/s41746-024-01311-5] [Medline: 39638907]
88. Hong G, Mao F, Zhang M, Zhang F, Wang X, Ren K, et al. Modeling and validation of wearable sensor-based gait parameters in Parkinson's disease patients with cognitive impairment. Front Aging Neurosci 2025;17:1590224 [FREE Full text] [doi: 10.3389/fnagi.2025.1590224] [Medline: 40787670]
89. Chen M, Sun Z, Xin T, Chen Y, Su F. An interpretable deep learning optimized wearable daily detection system for Parkinson's disease. IEEE Trans Neural Syst Rehabil Eng 2023;31:3937-3946. [doi: 10.1109/TNSRE.2023.3314100] [Medline: 37695969]

XSL•FO
RenderX

90. Channa A, Ruggeri G, Ifrim RC, Mammone N, Iera A, Popescu N. Cloud-connected bracelet for continuous monitoring of Parkinson's disease patients: integrating advanced wearable technologies and machine learning. Electronics 2024;13(6):1002. [doi: 10.3390/electronics13061002]
91. Singh M, Prakash P, Kaur R, Sowers R, Brašić JR, Hernandez ME. A deep learning approach for automatic and objective grading of the motor impairment severity in Parkinson's disease for use in tele-assessments. Sensors (Basel) 2023;23(21):9004 [FREE Full text] [doi: 10.3390/s23219004] [Medline: 37960703]
92. Weikert T, Li Y, Paez-Granados D, Easthope CA. Automated prediction of item-level ARAT scores from wearable sensors. : IEEE; 2025 Presented at: Proceedings of the 2025 IEEE International Conference on Rehabilitation Robotics (ICORR); 2025 May 12–16; Chicago, IL, USA p. 1239-1244. [doi: 10.1109/icorr66766.2025.11063162]
93. Yang PK, Filtjens B, Ginis P, Goris M, Nieuwboer A, Gilat M, et al. Freezing of gait assessment with inertial measurement units and deep learning: effect of tasks, medication states, and stops. J Neuroeng Rehabil 2024;21(1):24 [FREE Full text] [doi: 10.1186/s12984-024-01320-1] [Medline: 38350964]
94. Esfahani AH, Dyka Z, Ortmann S, Langendorfer P. Impact of data preparation in freezing of gait detection using feature-less recurrent neural network. IEEE Access 2021;9:138120-138131. [doi: 10.1109/access.2021.3117543]
95. Sun H, Ye Q, Xia Y. Predicting freezing of gait in patients with Parkinson's disease by combination of manually-selected and deep learning features. Biomed Signal Process Control 2024;88:105639. [doi: 10.1016/j.bspc.2023.105639]
96. Dvorani A, Waldheim V, Jochner MCE, Salchow-Hömmen C, Meyer-Ohle J, Kühn AA, et al. Real-time detection of freezing motions in Parkinson's patients for adaptive gait phase synchronous cueing. Front Neurol 2021;12:720516 [FREE Full text] [doi: 10.3389/fneur.2021.720516] [Medline: 34938252]
97. Khan U, Riaz Q, Hussain M, Zeeshan M, Krüger B. Towards effective Parkinson's monitoring: movement disorder detection and symptom identification using wearable inertial sensors. Algorithms 2025;18(4):203. [doi: 10.3390/a18040203]
98. Park H, Youm C, Cheon SM, Kim B, Choi H, Hwang J, et al. Using machine learning to identify Parkinson's disease severity subtypes with multimodal data. J Neuroeng Rehabil 2025;22(1):126 [FREE Full text] [doi: 10.1186/s12984-025-01648-2] [Medline: 40457336]
99. Goubault E, Martin C, Duval C, Daneault JF, Boissy P, Lebel K. Enhanced detection and segmentation of sit phases in patients with Parkinson's disease using a single smartwatch and random forest algorithms. Sensors (Basel) 2025;25(19):6104 [FREE Full text] [doi: 10.3390/s25196104] [Medline: 41094927]
100. Liu T, Li Z, Ji X, Xu M, Wang L, Chi Y. Friction nanogenerators based on asymmetrically adherent conductive hydrogels for multifunctional sensing and Parkinsonian gait diagnostics. Chem Eng J 2025;524:169026. [doi: 10.1016/j.cej.2025.169026]
101. Nanayakkara T, Herath HMKKMB, Malekroodi HS, Madusanka N, Yi M, Lee BI. Multi-domain CoP feature analysis of functional mobility for Parkinson's disease detection using wearable pressure insoles. Sensors (Basel) 2025;25(18):5859 [FREE Full text] [doi: 10.3390/s25185859] [Medline: 41013102]
102. Ma L, Lin S, Jin J, Wang Z, Wang X, Chen Z, et al. Objective assessment of gait and posture symptoms in Parkinson's disease using wearable sensors and machine learning. Front Aging Neurosci 2025;17:1618764 [FREE Full text] [doi: 10.3389/fnagi.2025.1618764] [Medline: 40860216]
103. Lopes T, Reis Carneiro M, Morgadinho A, Reis Carneiro D, Tavakoli M. ParCuR-a novel AI-enabled gait cueing wearable for patients with Parkinson's disease. Sensors (Basel) 2025;25(22):7077 [FREE Full text] [doi: 10.3390/s25227077] [Medline: 41305286]
104. Channa A, Popescu N, Ciobanu V. Wearable solutions for patients with Parkinson's disease and neurocognitive disorder: a systematic review. Sensors (Basel) 2020;20(9):2713 [FREE Full text] [doi: 10.3390/s20092713] [Medline: 32397516]
105. Mughal H, Javed AR, Rizwan M, Almadhor AS, Kryvinska N. Parkinson's disease management via wearable sensors: a systematic review. IEEE Access 2022;10:35219-35237. [doi: 10.1109/access.2022.3162844]
106. Sigcha L, Borzì L, Amato F, Rechichi I, Ramos-Romero C, Cárdenas A, et al. Deep learning and wearable sensors for the diagnosis and monitoring of Parkinson's disease: a systematic review. Expert Syst Appl 2023;229:120541. [doi: 10.1016/j.eswa.2023.120541]
107. Joo JH, Lieu N, Tang Y, Browne DS, Agusala B, Liao JM. Trends in utilization of remote monitoring in the United States. Health Aff Sch 2025;3(6):qxaf115. [doi: 10.1093/haschl/qxaf115] [Medline: 40575760]
108. Tan SY, Sumner J, Wang Y, Wenjun Yip A. A systematic review of the impacts of remote patient monitoring (RPM) interventions on safety, adherence, quality-of-life and cost-related outcomes. NPJ Digit Med 2024;7(1):192 [FREE Full text] [doi: 10.1038/s41746-024-01182-w] [Medline: 39025937]
109. Hsueh WT, Yu CX, Cheng HC, Chen MY, Wang HM, Fu LM. A comprehensive review of wearable devices for non-invasive biosensing. Trends Anal Chem 2025;193:118425. [doi: 10.1016/j.trac.2025.118425]
110. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature 2023;616(7956):259-265. [doi: 10.1038/s41586-023-05881-4] [Medline: 37045921]
111. Smuck M, Odonkor CA, Wilt JK, Schmidt N, Swiernik MA. The emerging clinical role of wearables: factors for successful implementation in healthcare. NPJ Digit Med 2021;4(1):45 [FREE Full text] [doi: 10.1038/s41746-021-00418-3] [Medline: 33692479]
112. Xu H, Liu J, Hu H, Zhang Y. Wearable sensor-based human activity recognition method with multi-features extracted from Hilbert-Huang transform. Sensors (Basel) 2016;16(12):2048 [FREE Full text] [doi: 10.3390/s16122048] [Medline: 27918414]

XSL•FO
RenderX

113. Ordóñez FJ, Roggen D. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. Sensors (Basel) 2016;16(1):115 [FREE Full text] [doi: 10.3390/s16010115] [Medline: 26797612]
114. Fiordelli M, Diviani N, Schulz PJ. Mapping mHealth research: a decade of evolution. J Med Internet Res 2013;15(5):e95 [FREE Full text] [doi: 10.2196/jmir.2430] [Medline: 23697600]
115. Evers LJW, Peeters JM, Bloem BR, Meinders MJ. Need for personalized monitoring of Parkinson's disease: the perspectives of patients and specialized healthcare providers. Front Neurol 2023;14:1150634 [FREE Full text] [doi: 10.3389/fneur.2023.1150634] [Medline: 37213910]
116. Toney-Butler TJ, Thayer JM. Nursing Process Internet. Treasure Island (FL): StatPearls Publishing; 2023.
117. Kwon SH, Park JK, Koh YH. A systematic review and meta-analysis on the effect of virtual reality-based rehabilitation for people with Parkinson's disease. J Neuroeng Rehabil 2023;20(1):94 [FREE Full text] [doi: 10.1186/s12984-023-01219-3] [Medline: 37475014]
118. Chuang CS, Chen YW, Zeng BY, Hung CM, Tu YK, Tai YC, et al. Effects of modern technology (exergame and virtual reality) assisted rehabilitation vs conventional rehabilitation in patients with Parkinson's disease: a network meta-analysis of randomised controlled trials. Physiotherapy 2022;117:35-42. [doi: 10.1016/j.physio.2022.07.001] [Medline: 36242929]
119. Yau CE, Ho ECK, Ong NY, Loh CJK, Mai AS, Tan EK. Innovative technology-based interventions in Parkinson's disease: a systematic review and meta-analysis. Ann Clin Transl Neurol 2024;11(10):2548-2562 [FREE Full text] [doi: 10.1002/acn3.52160] [Medline: 39236299]
120. Bowman T, Gervasoni E, Arienti C, Lazzarini SG, Negrini S, Crea S, et al. Wearable devices for biofeedback rehabilitation: a systematic review and meta-analysis to design application rules and estimate the effectiveness on balance and gait outcomes in neurological diseases. Sensors (Basel) 2021;21(10):3444 [FREE Full text] [doi: 10.3390/s21103444] [Medline: 34063355]
121. Giannakopoulou KM, Roussaki I, Demestichas K. Internet of things technologies and machine learning methods for Parkinson's disease diagnosis, monitoring and management: A systematic review. Sensors (Basel) 2022;22(5):1799 [FREE Full text] [doi: 10.3390/s22051799] [Medline: 35270944]
122. Mathias R, McCulloch P, Chalkidou A, Gilbert S. Digital health technologies need regulation and reimbursement that enable flexible interactions and groupings. NPJ Digit Med 2024;7(1):148 [FREE Full text] [doi: 10.1038/s41746-024-01147-z] [Medline: 38890404]

Abbreviations

AI: artificial intelligence
CNKI: China National Knowledge Infrastructure
CNN: convolutional neural network
DL: deep learning
LSTM: long short-term memory
MeSH: Medical Subject Headings
ML: machine learning
PD: Parkinson disease
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
SVM: support vector machine

Edited by S Brini; submitted 10.Oct.2025; peer-reviewed by O Oyetunji, S Jiang, S Mohanadas, S Ajayi; comments to author 01.Dec.2025; accepted 04.Feb.2026; published 26.Feb.2026.

Please cite as:
Li S, Chen S, Yu X, Shang H, Tu T, Quan M
AI-Enabled Wearables for Motor Function Assessment and Rehabilitation in Parkinson Disease: Scoping Review
J Med Internet Res 2026;28:e85596
URL: https://www.jmir.org/2026/1/e85596
doi:10.2196/85596
PMID:

©Shengting Li, Siqi Chen, Xiaosong Yu, Huixiang Shang, Tingting Tu, Mingtao Quan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution,

XSL•FO
RenderX

and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effectiveness of Telerehabilitation Interventions for Self-Management of Tinnitus: Update of a Systematic Review

Sara Demoen[1,2,3,4], PhD; Elise Van Kerchove[1], MSc; Annick Timmermans[1], Prof Dr; Vincent Van Rompaey[2,3], Prof Dr Med; Sarah Michiels[1,3*], Prof Dr; Annick Gilles[2,3,5*], Prof Dr

[1]REVAL Rehabilitation Research Center, Faculty of Rehabilitation Sciences, Hasselt University, Hasselt, Belgium

[2]Department of Translational Neurosciences, Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Flanders, Belgium

[3]Department of Otorhinolaryngology and Head and Neck Surgery, Faculty of Medicine and Health Sciences, Antwerp University hospital, Antwerp, Belgium

[4]Rehabilitation Research Group (RERE), Faculty of Health Sciences, Vrije Universiteit Brussel, Brussels, Brussels Capital, Belgium

[5]Department of Education, Humanities and Social Work, Faculty of Human and Social Welfare, University College Ghent, Ghent, Flanders, Belgium

[*]these authors contributed equally

Corresponding Author:
Sarah Michiels, Prof Dr
REVAL Rehabilitation Research Center
Faculty of Rehabilitation Sciences
Hasselt University
Martelarenlaan 42, 3500
Hasselt
Belgium
Phone: 32 011269307
Email: sarah.michiels@uhasselt.be

Abstract

Background: Approximately 14% of the adult population has tinnitus, and current treatments are often costly and time-consuming. Telerehabilitation might reduce treatment costs without compromising effectiveness.

Objective: Telerehabilitation is a quickly evolving research topic. Therefore, this systematic review update aims to give an overview of the research concerning the effectiveness of telerehabilitation interventions for self-management of tinnitus published between 2022 and 2025.

Methods: This systematic review adheres to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020) guidelines. PubMed, ScienceDirect, Scopus, Web of Science, and Cochrane Library were consulted for eligible studies concerning a study intervention of any possible form of self-management or telerehabilitation for adult patients with subjective tinnitus as a primary complaint. The risk of bias (RoB) and certainty of all included studies were assessed respectively by the Cochrane RoB2-tool and GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) framework.

Results: In total, 24 papers were included, of which 6 studied multiple telerehabilitation forms. Internet-based cognitive behavioral therapy with guidance by a psychologist or audiologist was examined in 5 studies (n=619), self-help manuals in 1 study (n=10), technological self-help devices in 3 studies (n=286), smartphone apps in 13 studies (n=23,788), and other internet-based interventions in 5 studies (n=442). These rehabilitation categories were proven to be effective in decreasing tinnitus severity and relieving tinnitus distress as measured by tinnitus questionnaires.

Conclusions: The strength of this review is the gathering of recent studies on the very evolving topic of telerehabilitation for tinnitus. An important limitation of all included studies is that they raised some to great concerns of RoB. As a result, it is necessary to acknowledge that the overall certainty of the evidence ranged from low to moderate certainty. In addition, some crucial confounding parameters, such as the presence of hearing loss, hyperacusis, anxiety, depression, or sleeping problems, were not taken into consideration by all studies. This review gives an indication of the use of different telerehabilitation and self-management interventions for real-world clinical use, stating not only their possibilities but also their limitations. Overall, telerehabilitation was found to be effective in reducing tinnitus severity and distress. It forms a possible tool to improve the self-management capacities of the patient and the accessibility of tinnitus care as a replacement or an addition to in-person care. Nevertheless, barriers such as a lack of time, engagement, motivation, and openness of the patient, causing high dropout, should be taken into

XSL•FO
RenderX

consideration. This review accentuated the shift from internet-based cognitive behavioral therapy to the growing interest in the use of smartphone apps, increasing the accessibility of the treatments even more.

Trial Registration: PROSPERO CRD 42021285450; https://www.crd.york.ac.uk/PROSPERO/view/CRD42021285450

(J Med Internet Res 2026;28:e83529) doi:10.2196/83529

KEYWORDS

tinnitus; telerehabilitation; self-management; mobile phone; smartphone application

Introduction

Background

Tinnitus is defined as the perception of a constant or intermittent sound without the presence of an external auditory source [1,2]. It is a highly prevalent symptom affecting approximately 14% of the adult population [3]. The daily functioning and health-related quality of life of tinnitus patients can directly or indirectly be disrupted by the distress they experience due to the tinnitus sound [4]. Therefore, the most commonly used and evidence-based treatments are tinnitus retraining therapy (TRT) and cognitive behavioral therapy (CBT) [5,6]. The high costs and time consumption of these treatments, combined with the repeated need for care of these patients, resulted in a quest to find a more cost-effective alternative [7-9]. As an answer, telerehabilitation is on the rise to promote self-management in tinnitus patients and reduce their dependency on in-person care. Telerehabilitation, meaning delivery of care from a distance, became a necessity rather than an opportunity during the COVID-19 pandemic [10-12].

The previous systematic review published in 2023 on the topic of telerehabilitation for chronic subjective tinnitus patients indicated that low-contact treatment provided from a distance through several telerehabilitation forms could be used as a substitution for or an addition to in-person clinical care for tinnitus [13]. It can potentially reduce the costs of treatment without sacrificing the level of effectiveness. Telerehabilitation treatment forms are easy to implement in clinical practice and form accessible treatment options [13]. Overall, telerehabilitation in the form of internet-based cognitive behavioral therapy (iCBT) with or without guidance, self-help manuals, self-help devices, smartphone apps, and other internet-based interventions was found to be effective in reducing tinnitus severity and distress [13]. Nevertheless, barriers such as lack of time, engagement, motivation, and openness of the patient need to be considered, as they can cause high dropout rates. The previous review also concluded that future research should further explore which factors are most likely to cause the lack of compliance and how clinicians can counteract these [13].

Due to the advances in technology made every day, telerehabilitation will keep evolving and, therefore, remain a trending topic to follow up on. The application of digital technology in tinnitus therapy is rising stronger than ever. For example, a variety of smartphone apps for people with tinnitus are available, promising relief from the condition. Yet, it is difficult to estimate their potential [14,15]. After the COVID-19 pandemic, the potential benefits of telerehabilitation forms became undeniable, and recently, these new tinnitus treatment forms were even implemented in the guidelines for tinnitus care. Since then, telehealth has been a highly dynamic research field, and several new studies on telerehabilitation for tinnitus have been published in the last couple of years [16-18]. The high number of patients with tinnitus, their repeated need for care, long waiting lists for regular care, and the high costs related to gold-standard treatments are some of the reasons why the need to explore telerehabilitation options is so high. It might form the key to making treatments more cost-effective and more accessible without losing on a level of effectiveness. It can even benefit the effectiveness when it is used as a tool to improve tailoring of the treatment approach to the patient's needs and as a tool to increase adherence and self-management skills of the patient, reducing the need for repeated care.

Objectives

In the past decades, various low-contact and self-management telerehabilitation treatments were developed for patients with tinnitus to search for an alternative or additional tool for the existing treatment options that are time-consuming and costly. Due to the rise in scientific research and the high number of new publications in recent years on this topic, this systematic review aimed to give an updated overview of the research concerning the effectiveness of telerehabilitation interventions for self-management of tinnitus published between 2022 and 2025.

Methods

Study Design

This systematic review is an update of an earlier systematic review published in February 2023 [13]. It adheres to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [19]. The predefined protocol registered in PROSPERO from the earlier review, with registration number CRD42021285450, was repeated identically.

Eligibility Criteria

The eligibility criteria can be found in Textbox 1. These criteria remained the same for the updated review. However, a publication date restriction was added as an exclusion criterion. Papers published before October 18, 2021, were not taken into consideration because they were already included in the previous review. The search ended on February 10, 2025.

XSL•FO
RenderX

Textbox 1. Eligibility criteria.

Inclusion criteria

- Adult patients (>18 y).
- Subjective tinnitus as a primary complaint.
- A study intervention comprising any possible form of self-management or telerehabilitation.
- No publication date restrictions.

Exclusion criteria

- No tinnitus.
- Objective tinnitus.
- Tinnitus as a secondary complaint.
- Full text not available in Dutch, French, German, Greek, or English.

Information Sources

On February 10, 2025, a total of 5 databases were consulted: PubMed, ScienceDirect, Scopus, Web of Science, and Cochrane Library.

Search Strategy

The search queries were composed according to the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) guidelines [20] (checklist provided in Multimedia Appendix 1). The queries comprise free terms and MeSH (Medical Subject Headings) terms for tinnitus (population) and self-management, telerehabilitation, smartphone application (intervention), and their synonyms. All terms and synonyms were controlled for value to the search query. If a term did not result in an additional number of studies found, it was eliminated to keep the search query as concise as possible while maintaining the highest level of efficiency. The search queries were not adapted from the previous review and were rerun identically, because the addition of terms did not result in more relevant eligible studies. The complete search queries can be consulted in Multimedia Appendix 2. All searches were concluded on February 10, 2025. Besides the date restriction that papers had to be from 2022 or later, no additional search filters or restrictions were added to the search strategy. Rayyan was used to eliminate duplicates from the searches.

Selection Process

Double screening of the gathered papers was performed by 2 reviewers (Sara Demoen and EVK) independently and blindly. A first screening on the abstract and title was performed, followed by a screening on the full text. The Rayyan web platform was used to screen all papers efficiently. The final decision on the inclusion of the papers was obtained after a consensus meeting, discussing potential conflicts or papers that were labelled as maybe by at least one of the reviewers.

Data Collection

Both reviewers collected the data independently, and the data extraction tables were composed during a consensus meeting with the reviewers and the supervisors. In the earlier review, 6 subcategories were made: iCBT with guidance (feedback, monitoring, and support by a psychologist or audiologist), iCBT without guidance, self-help devices, self-help manuals, smartphone apps, and other internet-based interventions. For all of these categories, new studies were included during this updated search, except for the subcategory iCBT without guidance. For each subcategory, a data-extraction table was composed, comprising the collected study information according to the applicable items of the subdivisions: PICO (population, intervention, comparison, and outcome).

Data Items

The study and population characteristics that were listed are the author, the year of publication, the study design, the sample size (n), the gender distribution (%), the mean age and the SD (y), the type of tinnitus, the mean duration of tinnitus, and the presence of hearing loss.

Intervention is the first aspect summarized in the table of evidence. This section discusses the following items: (+ if applicable comparison: control intervention, ie, if there is a comparison with a control group [CG], these items are also discussed for the control intervention and not solely the experimental intervention): the used intervention form and specifics, such as guidance, duration, and follow-up. For the next section, Outcome, outcome measures, and dropout were presented, and finally, under the Results section, data results, GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) score, and the conclusions were listed. Each of the 5 subcategories of telerehabilitation included in this review had a table of evidence giving an overview of the information on the intervention, outcome, results, and conclusions of the included study.

Risk of Bias Assessment

The risk of bias (RoB) assessment was performed using the Cochrane RoB2 tool, a tool for assessing RoB in randomized controlled trials (RCTs) [21]. However, it can also be used for other study designs because the domains are, in general, also applicable [21]. The study design of the included studies is diverse, and for this reason, following a separate RoB assessment for each study design would not give a clear overview for comparison [21]. Alternative tools such as ROBINS-I (Risk of Bias in Non-Randomized Studies of

XSL•FO
RenderX

Interventions) were considered, but using a different instrument in the current update would introduce heterogeneity in assessment methods. We therefore applied the Cochrane RoB tool across all designs to ensure methodological consistency with the previous version of this review, and because this tool provides a more conservative, strict assessment of bias for nonrandomized studies rather than an overly lenient one. RoBvis, an RoB visualization tool by Cochrane, was used to summarize the results in a graphic manner [22]. The RoB screening was performed blinded by 2 reviewers (Sara Demoen and EVK), and conflicts were resolved after a consensus meeting.

Data Synthesis

The heterogeneity in reported interventions and outcome measures precluded meta-analysis. A meta-analysis was not performed in the present review update for several methodological reasons. First, this work serves as an update of a previously published systematic review, which also did not pool results due to substantial heterogeneity. Maintaining consistency with the earlier review ensures comparability across iterations. Second, the included studies used a wide range of outcome measures that were operationalized and reported in noncomparable ways, limiting the feasibility of quantitative synthesis. Finally, the statistical reporting across studies varied considerably, with insufficiently detailed or inconsistent data formats that prevented reliable extraction of effect sizes. In combination, these factors precluded a valid and meaningful meta-analysis, and therefore, it was decided to summarize the collected data in a systematic review. This was not different in the current update.

Certainty Assessment

The GRADE framework is a tool to assess the level of certainty by examining RoB, imprecision, inconsistency, indirectness, and publication bias of the included studies [23]. The GRADE assessment was completed according to the GRADE handbook guidelines [23].

Results

Study Selection

The initial database searches on February 10, 2025, resulted in 1192 potentially relevant results (Figure 1 [19]). After removal of the duplicates and of the papers published before October 2021, a total of 496 papers remained for primary screening on title and abstract. During the primary screening, 441 papers were excluded; the reasons for exclusion are specified in Figure 1. In total, 55 papers were sought for retrieval for the screening of full text. Subsequently, the screening of the full text was completed for 46 papers because the full text of 9 papers could not be retrieved. The reason for exclusion is again specified in Figure 1. A total of 24 papers were found to be eligible [24-47]. These eligible papers were divided into the same subgroups used in the previous systematic review [13]. The subgroups were based on the form of telerehabilitation as an intervention: iCBT with guidance, iCBT without guidance, self-help devices, self-help manuals, smartphone apps, and other internet-based interventions.

XSL•FO
RenderX

Figure 1. Flow diagram for study inclusion (PRISMA 2020). Studies can belong to multiple of the listed study objective categories. iCBT: internet-based cognitive behavioral therapy; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses. S

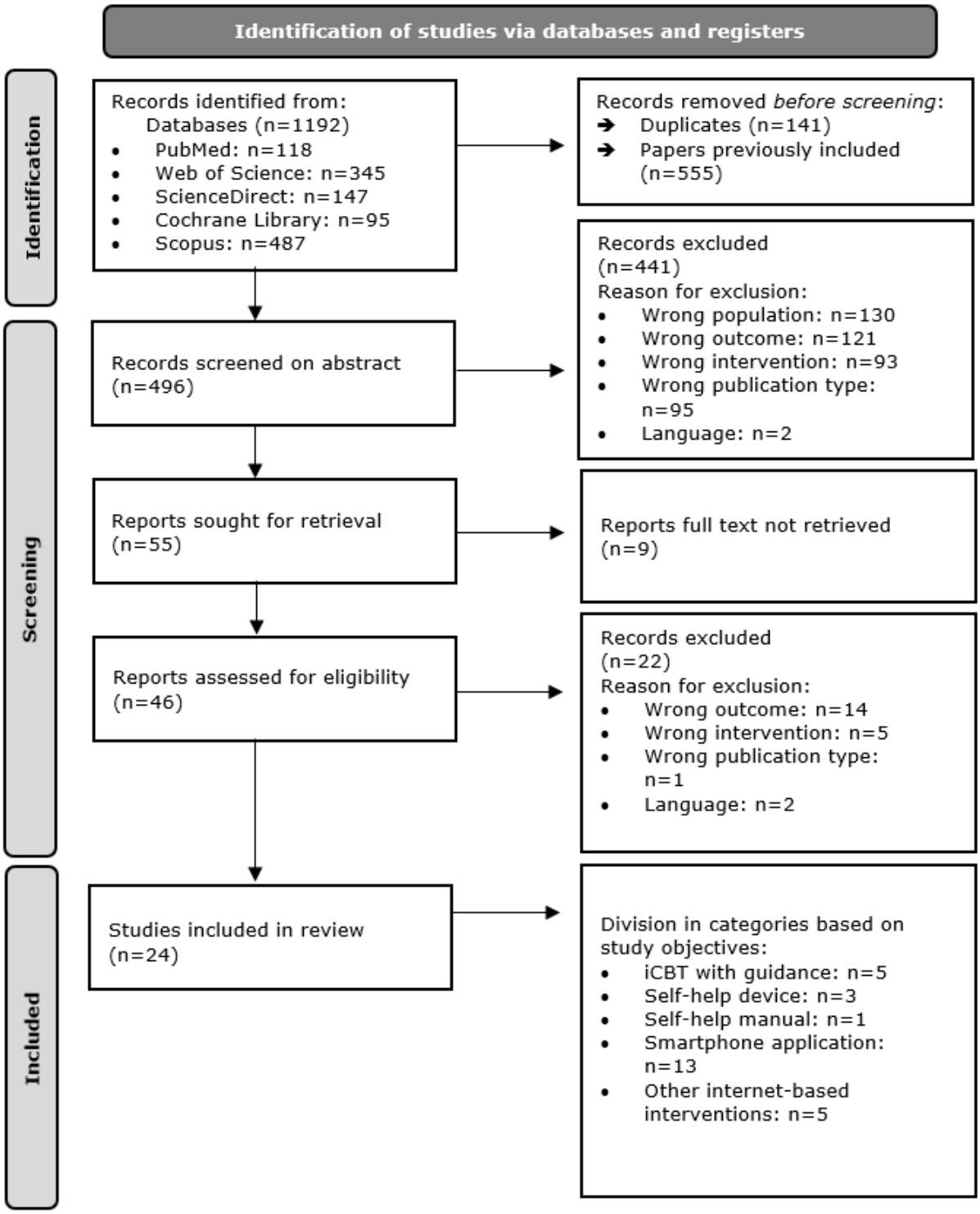

About RoB

Controlled Trials

In total, 11 studies were controlled trials [24,25,30-32,35,37,41,44,45,47], of which 6 studies were RCTs [25,30,31,35,41,47]. Five RCTs had a low RoB arising from the randomization process, except for Altissimi et al [25], who showed a higher RoB due to the lack of information given concerning the randomization method (D1) [25,30,31,35,41,47]. All controlled trials showed some reasons for concern of RoB due to deviations from the intended interventions (D2), except for Goshtasbi et al [30]. RoB, due to missing outcome data (D3), was scored as "high concerns" and "some concerns" for studies with respectively moderate to high dropout rates or no clearly specified plan on how to handle these missing data. All controlled studies were rated "low concerns for RoB" in the

XSL•FO
RenderX

measurement of outcome (D4), except Ravi et al [37], a study with a sample size of only 2 patients and only 1 questionnaire as an outcome measure. Most studies had a full, clear, correct, and specified plan for statistical analysis, meaning the concern for RoB in the selection of the reported result (D5) was low. However, Alonso-Valerdi et al [24] and Ravi et al [37] had a nondetailed protocol, missing information on the exact premised data analysis plan. Overall, studies were rated "some concerns" to "high concerns" of RoB. See Figure 2 [24,25,30-32,35,37,41,44,45,47] for the RoB assessment results of the controlled trials. Additionally, a summarization of RoB results can be found in Multimedia Appendix 3.

Figure 2. Risk of bias in the controlled trials scoring per domain [25,26,31,32,33,36,38,42,45,46,48].

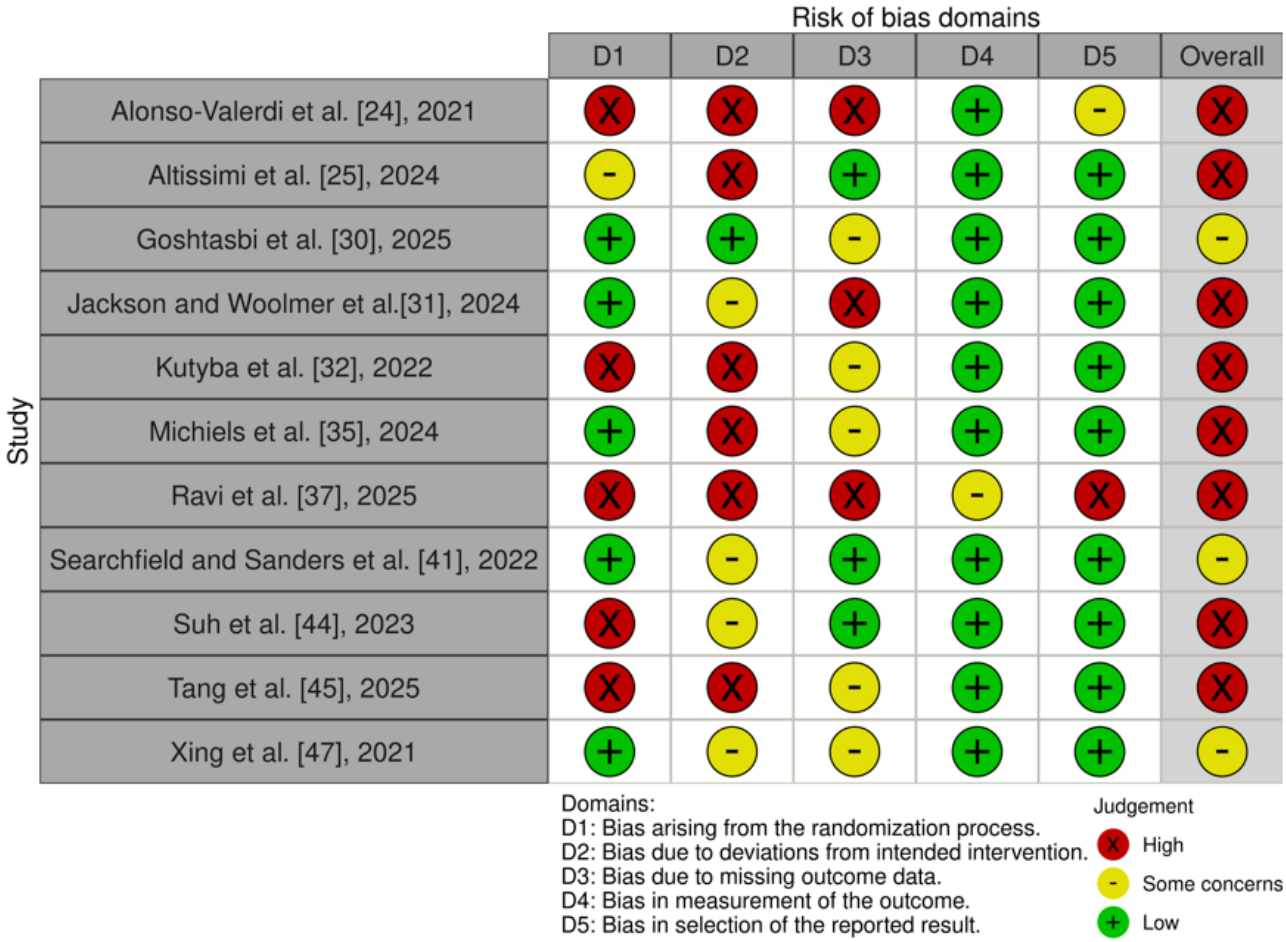

Noncontrolled Trials

A total of 13 studies were noncontrolled trials [26-29,33,34,38-40,42,43,46]. They all show some concerns about RoB due to a lack of randomization, blinding, and CGs (D1 and D2). RoB due to missing outcome data (D3) was again scored high, and there were some concerns for studies with moderate to high dropout rates. The majority of the studies specified their statistical analysis beforehand, resulting in low RoB in the selection of the reported results (D5). See Figure 3 [26-29,32,34,36,38-40,42,43,46] for the RoB assessment results of the noncontrolled trials.

XSL•FO
RenderX

Figure 3. Risk of Bias in the noncontrolled trials per domain [27,28,29,30,33,35,37,39,40,41,43,44,47].

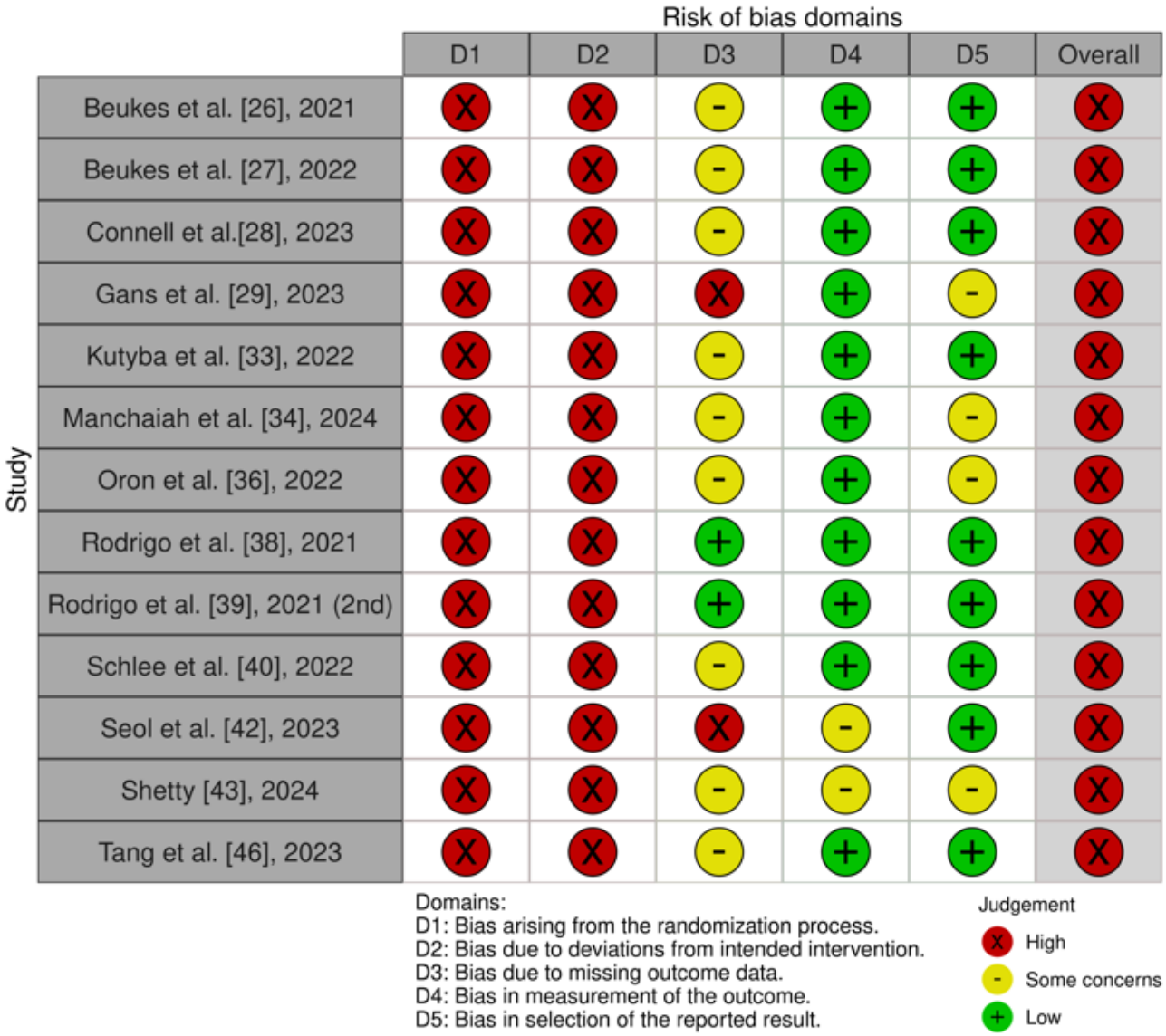

Certainty

In general, the included studies had a low to moderate certainty level according to the GRADE framework [23].

iCBT with Guidance

Study and Patient Characteristics

A total of 3 studies looked into the use of an iCBT program with guidance provided by an audiologist in the form of feedback, monitoring, and support through an online messaging system within the platform [26,27,34]. All 3 studies were cohort studies with an adult study population with chronic tinnitus. Further, 2 studies additionally specified that tinnitus symptoms had to be present for more than 3 months, and it had to be at least a mild severity of tinnitus. Beukes et al [26] stated that patients were only eligible if they scored at least 25 of 100 on the Tinnitus Functional Index (TFI). The mean tinnitus duration ranged from 11.64 to 13.91 years between the studies. Beukes et al [26] was the only study controlling for the presence of hearing loss, while Beukes et al [27] was the only one specifying the tinnitus characteristics, namely the tinnitus location. The total sample is 391 patients receiving iCBT with guidance. For further specifications, see Table S1 in Multimedia Appendix 4. Two studies examined an iCBT program with minimal guidance [38,39]. However, it needs to be noted that both these published studies were from the same author and concerned the same sample data. Both studies used data from a cohort of 228 patients with long-standing chronic tinnitus with a duration of 17.68 (SD 19.42). For further specifications, see Table S1 in Multimedia Appendix 4.

Results of Individual Studies

All 3 studies used a similar iCBT program consisting of an interactive e-learning program with 22 modules based on the original CBT self-help program developed by Andersson and Kaldo [48] (2004) [26,27,34]. The interventions were guided by an audiologist who monitored progress, provided feedback on worksheets that were completed, outlined the content of new modules, and answered questions. The guidance by the audiologist was offered through a 2-way encrypted online

XSL•FO
RenderX

messaging system. The primary tinnitus outcome of all studies was tinnitus severity or distress measured by the TFI. Various secondary outcome measures, using self-report questionnaires, were appended. All studies used data from previous trials. See Table 1 for further details. All studies concluded that the iCBT with guidance provided by an audiologist trained in iCBT principles was effective in the treatment of chronic tinnitus. Beukes et al [26] demonstrated that iCBT has long-term positive effects, and Manchaiah et al [34] added that it can help individuals with tinnitus to think more positively by changing unhelpful patterns. Both studies from Rodrigo et al [38,39] used the same data that was gathered from multiple trials and looked into the use of an iCBT program with minimal guidance through an encrypted messaging system solely for support. This iCBT program was again based on the original CBT self-help program developed by Andersson and Kaldo [48] (2004). However, it was adapted to a set of only 21 modules, of which 5 were optional. These 2 studies also found a decrease in TFI after completing the iCBT program and additionally investigated predictors of iCBT outcomes. For more information, see Table 1.

XSL•FO
RenderX

Table 1. Table of evidence for the studies concerning iCBT[a] with guidance.

Study	Intervention			Outcome			Results				
Author and year	Intervention	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[b]
Beukes et al [26], 2022	Interactive e-learning iCBT consisting of 22 modules and guidance by an audiologist	8 weeks	FU1[c]: postintervention assessment; FU2[d]: 2-months postintervention; FU3[e]:1-year postintervention	TFI[f]	THS[g].	Effect sizes, LMM[h], and the RCI[i].	Significant BL[j]-FU3 TFI mean improvement of 24.3 (SD 22.5; T0: 54.23, SD 20.7; T3: 30.15, SD 24.51). There was a clinically significant change for 56 (42%) of the participants using the reliable change criterion of 26.98 at T3.	Statistically significant BL-FU3 difference with a large effect size (Cohen *d*=1.06, CI 0.80 to 1.32).	68/200 (34%) did not complete FU1, 83/200 (43%) FU2, and 68/200 (34%) FU3.	The study provides evidence of long-term efficacy for an audiologist-guided iCBT tinnitus intervention.	⊕⊕⊕◯
Beukes et al [27], 2022	Interactive e-learning iCBT consisting of 22 modules and guidance by an audiologist	8 weeks	FU[k]=2 months	TFI	GAD-7[l], PHQ-9[m], ISI[n], EQ-5D-5L, THS, TCQ[o], tinnitus and hearing survey.	Pretest and posttest effect size for all primary and secondary outcomes; 2-sided *P* values using α=.05.	A significant, large effect size TFI change postintervention (mean 1.60, 0.91-2.23), which was maintained at FU.	The RCI indicated a prepost score difference of 19.51 on the TFI would be a clinically significant change.	4/27 did not complete FU1 and 9/27 did not complete FU2.	This pilot study has indicated the feasibility of iCBT for tinnitus in the United States.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention			Outcome			Results				
Author and year	Intervention	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[b]
Manchaiah et al [34], 2024	iCBT with guidance by an audiologist	8 weeks	FU: posttreatment	TFI	GAD-7, PHQ-9, ISI, EQ-5D-5L.	Mixed method approach, paired sample *t* test as the data met the assumption of normality using the Shapiro-Wilk test (*P*>.05).	More participants reported positive experiences following the intervention in the overall sample (preintervention mean 0.78, postintervention mean 1.63, *P*<.001).	53/164 (32.3%) provided at least 1 positive experience in both pre and postintervention (group 1), 19/164 (9.1%) provided positive experiences only during preintervention, 49/164 (29.9%) provided positive experiences only during postintervention, and 28.7 participants did not provide any positive experience in both pre- and postintervention.	NR[p].	iCBT can help individuals with tinnitus to think more positively by changing unhelpful thought patterns.	⊕⊕⊕◯
Rodrigo et al [38], 2022	Interactive e-learning iCBT consisting of 21 modules and minimal guidance by an audiologist	8 weeks (2-3 modules a week)	FU1=postintervention assessment; FU2=2-months postintervention	TFI	ISI, PHQ[q], HHIA-S[r], HQ[s], CFQ[t], SWLS[u].	Paired *t* test.	The mean BL tinnitus severity and tinnitus severity following the iCBT intervention were 57.93 (SD 19.17) and 34.22 (SD 22.76), respectively, indicating a significant decrease (*P*<.001).	65% had a successful iCBT outcome, meaning they had a 13-point or higher reduction in TFI after the intervention.	NR.	Some predictors can be noted for a better outcome of iCBT, such as, for example, level of education, BL tinnitus, workload, and disability allowance.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention			Outcome			Results				
Author and year	Intervention	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[b]
Rodrigo et al [39], 2021	iCBT	8 weeks (2-3 modules a week)	FU: postintervention	TFI	ISI, PHQ, HHIA-S, HQ, CFQ, SWLS.	Paired sample *t* tests.	Undertaking iCBT significantly reduced tinnitus severity scores (t_{227}=16.37; P<.001) from a mean BL severity score of 57.93 (SD 19.17) to a mean post-iCBT severity score of 34.22 (SD 22.78), as measured by the TFI.	A clinically significant 13-point change in TFI scores was achieved by 150 of the 228 participants after the intervention.	NR.	Decision tree models, especially the CART[v] and gradient boosting models, appeared to be promising in predicting iCBT outcomes.	⊕⊕⊕◯

[a]iCBT: internet-based cognitive behavioral therapy.
[b]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.
[c]FU1: first follow-up moment.
[d]FU2: second follow-up moment.
[e]FU3: third follow-up moment.
[f]TFI: tinnitus functional index.
[g]THS: tinnitus handicap scale.
[h]LMM: linear mixed effects models.
[i]RCI: Reliable Change Index.
[j]BL: baseline.
[k]FU: follow-up moment.
[l]GAD-7: Generalized Anxiety Disorder-7.
[m]PHQ-9: Patient Health Questionnaire-9.
[n]ISI: Insomnia Severity Index.
[o]TCQ: Tinnitus Cognitions Questionnaire.
[p]NR: not reported.
[q]PHQ: Physical Health Questionnaire.
[r]HHIA-S: Hearing Handicap Inventory for Adults-Short Form.
[s]HQ: Hyperacusis Questionnaire.
[t]CFQ: Cognitive Failures Questionnaire.
[u]SWLS: Satisfaction With Life Scale.
[v]CART: Communication Access Realtime Translation.

XSL•FO
RenderX

Self-Help Devices

Study and Patient Characteristics

Three controlled trials, of which 2 RCTs, studied the use of self-help devices for the treatment of chronic tinnitus [24,25,41]. The studies had a total sample size of 286 patients. The specifications of all 3 studies can be consulted in Table S2 in Multimedia Appendix 4.

Results of Individual Studies

All 3 studies looked into the use of a self-help device delivering a form of sound therapy. For Altissimi et al [25], this was the use of a sound generator for TRT treatment, and this was the CG. The intervention group received TRT treatment using a smartphone app. Searchfield and Sanders [41] also used a smartphone app in combination with the self-help devices for sound therapy, and the CG did as well use an app, but without a self-help device. All studies concluded that the use of a self-help device was effective in reducing tinnitus distress according to the TFI or Tinnitus Handicap Inventory [24,25,41]. Further details can be found in Table 2.

XSL•FO
RenderX

Table 2. Table of evidence for the studies concerning self-help devices.

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Alonso-Valerdi et al [24], 2021	(1) TRT[b]; (2) TEAE[c]; (3) ADT[d]; (4) BBT[e]; (5) music	No treatment.	2 months (60 min daily)	FU1[f]: after 1 week; FU2[g]: after 5 weeks; FU3[h]: after 8 weeks	THI[i]	HADS[j], audiometry, tinnitus characteristics analysis, and BL[k] neurophysiological analysis	Identification of ± or 0 effects *t* test	All negative differences between BL and FU[l] were statistically significant (*P*=.009)	29.4% of the TRT group reported diminution of their tinnitus perception, followed by those in TEAE (18.8%), BBT or ADT (14.3%), and music (7.7%) groups.	Dropout: music 6/16, BBT[m] 4/18, TRT 3/18, TEAE 5/18, ADT 5/18, and control 3/14	TRT most effective sound-based therapy to reduce tinnitus perception for individuals without anxiety. BBT, ADT, and TEAE produced very similar effects.	⊕⊕⊕◯
Altissimi et al [25], 2024	TRT with mobile app	TRT with a traditional sound generator.	6 months	FU1: after 3 months of treatment; FU2: after 6 months of treatment	THI	TRT interview: with questions on tinnitus, hearing loss, hyperacusis, classifying gravity, the annoyance, and the effect on the QoL[n] of tinnitus, using a VAS[o].	Descriptive data by counts of a mixed-effect model Tukey test	A significant difference in the THI scores was found in category 1 patients between those treated with a sound generator vs app (–14.529, *P*<.001)	Improvement was found in category 0 patients, both treated with the app and sound generator; conversely, in category 1, significant improvement was seen only for the group using a sound generator.	No dropout	TRT can be delivered to patients with mild tinnitus (class 0) through mobile apps with results comparable to traditional sound generators.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Searchfield and Sanders [41], 2022	The digital therapeutic consisted of an app for iPhone or Android smartphone, Bluetooth bone conduction headphones, a neck pillow speaker, and a cloud-based clinician dashboard to enable messaging and app personalization	The control app was a popular self-help passive sound therapy app called WN[p] Lite.	12 weeks	FU1: mid-trial=6 weeks; FU2: postinterevention	TFI[q]	COSIT[r], SUS[s], MAUQ[t]	ANOVA with 2 groups and 4 repeated measures, intent-to-treat and completer (per protocol) analyses	Mean changes TFI for USL[u] at 6 (16.36, SD 17.96 points) and 12 weeks (17.83, SD 19.87 points) were clinically meaningful (>13 reduction), the mean changes in WN group were not (6 wk 10.77, SD 18.53 points; 12 wk 10.12, SD 21.36 points)	A statistically higher proportion of USL participants achieved meaningful TFI change at 6 weeks (55%) and 12 weeks (65%) than the WN group at 6 weeks (33%) and 12 weeks (43%).	4/35 USL and 6/36 WN	The USL group demonstrated a higher proportion of responders than the WN group. The usability of the USL therapeutic was similar to the established WN app. The digital polytherapeutic demonstrated significant benefit for tinnitus reduction, supporting further development.	⊕⊕⊕◯

[a]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.
[b]TRT: tinnitus retraining therapy.
[c]TEAE: therapy for an enriched acoustic environment.
[d]ADT: auditory discrimination therapy.
[e]BBT: broad band therapy.
[f]FU1: first follow-up moment.
[g]FU2: second follow-up moment.
[h]FU3: third follow-up moment.
[i]THI: Tinnitus Handicap Inventory.
[j]HADS: Hospital Anxiety and Depression Scale.
[k]BL: baseline.
[l]FU: follow-up moment.
[m]BBT: Broad band therapy.
[n]QoL: quality of life.
[o]VAS: visual analog scale.
[p]WN: white noise.
[q]TFI: Tinnitus Functional Index.

XSL•FO
RenderX

[r]COSIT: Client-Oriented Scale of Improvement in Tinnitus.

[s]SUS: System Usability Scale.

[t]MAUQ: mHealth App Usability Questionnaire.

[u]USL: UpSilent app.

Self-Help Manual

Study and Patient Characteristics

Only 1 cohort with a sample size of 10 adult patients with tinnitus and with bilateral mild sloping sensorineural hearing loss who wore a hearing aid used a self-help manual as an intervention [43]. Further details can be found in Table S3 in Multimedia Appendix 4.

Results of Individual Studies

Shetty [43] examined the use of a self-management manual for tinnitus with content such as (1) hearing aids; (2) listening to songs or soothing sounds; (3) tips for food intake; (4) exercise, yoga, and meditation; (5) locking negative thinking; (6) distraction techniques; (7) progressive muscle relaxation; (8) slow breathing; and (9) attention control and imaginary techniques. The self-help manual was prepared using the health literacy recommendations by adapting the United States Department of Health and Human Services. It was concluded that the self-management manual effectively reduces tinnitus perception and associated problems [43]. See Table 3 for further information.

Table 3. Table of evidence for the studies concerning the self-help manual.

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Shetty et al [43], 2024	Self-management manual for tinnitus	N/A[b]	30 days	FU[c]: postintervention	THI[d], TFI[e]	N/A	Independent samples *t* test.	The results revealed a significant reduction in THI after 30 days of manual use (t_9=1.99, *P*=.07). A similar result was found in TFI, in which functional impairment from tinnitus was significantly reduced between visits one and two (t_9=6.13, *P*=.001)	The means of THI and TFI scores in visit 1 were 31 (SD 5.16) and 58 (SD 5.92), respectively. Furthermore, at visit 2, the THI and TFI scores were 29 (SD 4.59) and 45 (SD 7.05), respectively.	No dropout	Patients who have used the manual reported relaxing and coping with tinnitus. It is concluded that the self-management manual effectively reduces tinnitus perception and associated problems.	⊕⊕⊕◯

[a]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.

[b]N/A: not applicable.

[c]FU: follow-up moment.

[d]THI: Tinnitus Handicap Inventory.

[e]TFI: Tinnitus Functional Index.

XSL•FO
RenderX

Smartphone Apps

Study and Patient Characteristics

A total of 13 articles, with a combined sample size of 23,637 participants, discussed the use of a smartphone app as part of the treatment of tinnitus patients [25,30,32,33,35-37,40-42,44-46]. Clarification on the study and patient characteristics of the studies examining the use of a smartphone app is noted in Table S4 in Multimedia Appendix 4.

Results of Individual Studies

The intervention used a smartphone app in 13 studies [25,30,32,33,35-37,40-42,44-46]. First, in 8 studies, this smartphone app was used to deliver a form of sound therapy to the patients. These studies showed that sound therapy through a smartphone app can be beneficial in reducing tinnitus distress. In 2 of these studies, Goshtasbi et al [30] and Searchfield and Sanders [41], the smartphone app combined both sound therapy and CBT or psychoeducation, respectively. Second, 5 studies look into TRT, CBT, or psychoeducational treatment delivered through a smartphone app. TRT with a mobile app appeared to be as effective as traditional TRT with or without a sound generator, and daily training using a CBT-based app targeting maladaptive beliefs may decrease tinnitus intrusiveness and handicap. Psycho-education through a smartphone app gave promising results in reducing tinnitus distress, but did not result in changes in tinnitus loudness. Finally, 1 study investigated the use of a smartphone app in the treatment of a specific subpopulation of tinnitus patients, namely patients with somatic tinnitus [35]. This pilot study provides evidence that an app-based physiotherapy intervention with a cervical spine exercise program could be an effective option for treating somatic tinnitus [35]. See Table 4 for an overview of the detailed information for all studies concerning treatments using smartphone apps.

XSL•FO
RenderX

Table 4. Table of evidence for the studies concerning smartphone apps.

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Altissimi et al [25], 2024	TRT[b] with mobile app	TRT with a traditional sound generator	6 months	FU1[c]: after 3 months of treatment; FU2[d]: after 6 months of treatment	THI[e]	TRT interview: with questions on tinnitus, hearing loss, hyperacusis, classifying gravity, the annoyance, and the effect on the QoL[f] of tinnitus, using a VAS[g].	Descriptive data by counts, a mixed-effect model Tukey test.	A significant difference in the THI scores was found in category 1 patients between those treated with a sound generator vs app (–14.529, *P*<.001), while no significance was found in category 0 patients (–1.186, *P*=.78).	Improvement was found in category 0 patients, both treated with the app and sound generator; conversely, in category 1, significant improvement was seen only for the group using a sound generator.	No dropout.	TRT can be delivered to patients with mild tinnitus (class 0) through mobile apps with results comparable to traditional sound generators.	⊕⊕⊕◯
Goshtasbi et al [30], 2025	Smartphone-based CBT[h] and customized sound therapy.	Waitlisted	8 weeks (1 weekly CBT module and 2 h daily sound therapy)	Posttreatment	TFI[i]	PHQ-9[j], GAD-7[k], PSQI[l], PSS[m].	Independent sample *t* test and chi-square tests of independence.	Subjects in the IG[n] showed greater improvements on their TFI (16.7, SD 14.9 versus 1.9, SD 10.8 (*P*<.001) and on 7 of 8 subscales (*P* <.001-.002).	18 (38.3%) IG patients showed a TFI improvement of 20 or more points posttreatment.	15 patients dropped out between week 2 and week 8 and were not included in the IG.	Smartphone-based CBT, providing CBT and customized sound therapy, was effective in reducing symptom severity and improving anxiety, sleep, and mood.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Kutyba et al [32], 2022	Free ReSound Tinnitus Relief app	N/A[o]	6 months (≥30 min/day)	FU1: after 3 months of treatment FU2: after 6 months of treatment	THI, TFI	User survey.	Repeated measures ANOVA with Bonferroni correction.	THI: $F_{2,102}=6.24$; $P<.001$; $e2=0.550$, TFI: $F_{2,102}=2.89$; $P<.001$; $e2=0.362$.	In the study, the general severity decreased after the first 3 months (THI: –6.5 in mean score in comparison to BL[p]; TFI: –6.3 in mean score) and again in the following 3 months of using the application (THI: –19.4; TFI: –14.3).	NR.	The sound therapy through the use of a smartphone app appeared to be beneficial in reducing subjective tinnitus distress.	⊕⊕⊕◯
Kutyba et al [33], 2022	Free ReSound Tinnitus Relief app	Participants not willing to participate in the intervention group	6 months (≥30 min/day)	FU1: after 3 months of treatment FU2: after 6 months of treatment	THI	VAS, user survey.	Independent (unpaired) *t* test or a chi-square test, and a mixed-design ANOVA with Bonferroni adjustment for multiple comparisons.	After 3 months of using the app, the THI global score significantly decreased ($P<.001$) in the study group, decreasing again at 6 months ($P<.001$).	A clinically important change (>20) in the THI was reported by 39% (17/44) of the study group and in none of the controls.	Dropout rate: 53% (50/94) in the study group and 55% (29/53) in the control group.	Use of an app that generates background sounds appears to be an effective way of reducing tinnitus severity.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Michiels et al [35], 2024	App-based cervical spine exercise program	Waiting list	9 weeks	FU: post-treatment	THI miniTQ[q]	N/A.	Independent sample *t* test and chi-square tests of independence, and mixed model ANOVA.	Reduction in THI and miniTQ for the IG, not for CG[r]. Strong significant decrease in THI (from mean 51.33, SD 23.77 to mean 38.96, SD 21.66) and miniTQ (from mean 13.27, SD 5.81 to mean 9.14, SD 5.27).	Effect sizes for THI were d=1.71 (95% CI 0.85 to 2.56) and for miniTQ d=1.02 (95% CI 0.24 to 1.80).	CG: 3 dropouts; IG: 4 dropouts.	This pilot study provides evidence that an app-based physiotherapy intervention could be an effective option for treating somatic tinnitus.	⊕⊕⊕◯
Oron et al [36], 2022	Brief cognitive behavioral training for tinnitus relief using a mobile GGTI[s] app	N/A	16 days (3 min/day)	FU1: at level 24 half way through the app; FU2: level 48 final level of the app	H-THI[t]	VAS, user survey.	Repeated measures ANOVA of completers.	A significant linear trend was present, $F_{1,13}$=14.5, P=.002, d=2.12. Oron et al [36] indicated 5 was a steady decrease in the H-THI score of 28.57, 22.14, and 22.0 for BL, FU1, and FU2, respectively.	50% of completers have shown reliable change (indicated by their RCI[u]).	12/26 (46%).	Several minutes a day of training using a CBT-based app targeting maladaptive beliefs may decrease tinnitus intrusiveness and handicap.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Ravi et al [37], 2025	Mobile-based TRT	Conventional face-to-face TRT	3 months (2×/week)	FU1: after 1 month of treatment; FU2: after 2 months of treatment; FU3: after 3 months of treatment	THI	N/A.	N/A.	Patient A: BL=40, FU1=40; FU2=36; FU3=24. Patient B: BL=36, FU1=24; FU2=16; FU3=14.	N/A.	N/A.	Mobile-based TRT might be a good option for patients who cannot attend in-person sessions.	⊕⊕◯◯
Schlee et al [40], 2022	Educational smartphone app	N/A	4 months	FU[v]: postintervention	THI, TS[w] NRS[x]	N/A.	Wilcoxon Signed Rank Test.	Statistical significance between the pre- and postintervention phase (W=345, *P*=.02).	Improvements on the tinnitus severity NRS reached an effect size of 0.408, while the improvements on the THI[y] were much smaller with an effect size of 0.168.	26 of 52 (50%) participants dropped out.	Educational training on a smartphone app without the guidance of a medical person can result in small to medium improvements in tinnitus distress measures, but not for tinnitus loudness.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Searchfield et al [41], 2022	The digital therapeutic consisted of an app for iPhone or Android smartphones, Bluetooth bone conduction headphones, a neck pillow speaker, and a cloud-based clinician dashboard to enable messaging and app personalization	The control app was a popular self-help passive sound therapy app called WN[z] Lite.	12 weeks	FU1: mid-trial=6 weeks FU2: postinterevention	TFI	COSIT[aa], SUS[ab], MAUQ[ac]	ANOVA with 2 groups and 4 repeated measures intent-to-treat and completer (per protocol) analyses	Mean changes in TFI for the USL[ad] group at 6 (16.36, SD 17.96 points) and 12 weeks (17.83, SD 19.87 points) were clinically meaningful (>13 points reduction), the mean change in WN scores were not clinically meaningful (6 weeks 10.77, SD 18.53 points; 12 weeks 10.12, SD 21.36 points).	A statistically higher proportion of USL participants achieved meaningful TFI change at 6 (55%) weeks and 12 (65%) weeks than the WN group at 6 (33%) weeks and 12 (43%) weeks.	4/35 USL and 6/36 WN.	The USL group demonstrated a higher proportion of responders than the WN group. The usability of the USL therapeutic was similar to the established WN app. The digital polytherapeutic demonstrated significant benefit for tinnitus reduction supporting further development.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Seol et al [42], 2023	Sound therapy smartphone app	N/A	6 months	FU1: mid-trial=3 months; FU2: postinterevention	THI	BDI[ae], SSL[af], VAS stress assessment, hearing tests.	1-way RM[ag] ANOVA as it passed the normality test (Shapiro-Wilk).	The THI scores, on the other hand, significantly decreased over the 6-month period (*P*=.02); when the Bonferroni correction method was applied for all *P* values, statistical significance was observed for THI scores between visits 1 and 3 (*P*=.006).	After 6 months of listening to the sound stimuli, participants' perceived tinnitus handicap severity significantly decreased.	NR[ah].	The findings of this study demonstrate the potential benefit of the tinnitus application for tinnitus improvement.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Suh et al [44], 2023	Smart device-based TRT with smart pad app	Conventional TRT	3 months (3 sessions 30-60 min) with 1 month interval	FU1=postintervention assessment; FU2=2-months postintervention	THI	VAS, STAI[ai], BDI, PSQI, pure tone audiometry.	*t* tests and chi-square tests.	The greatest THI improvement at 3 months was –23.3 (95% CI –33.41 to –13.4), and at 2 months –16.8 (95% CI –30.8 to –2.8), and the reductions were significant, *P*<.001.	There was no difference between IG and CG (*P*<.001).	NS[aj].	TRT based on smart devices can be an effective alternative to conventional in-person TRT.	⊕⊕⊕◯
Tang et al [46], 2022	FTRS[ak] mobile app	Cohort	2 h/d	FU every 3 months	THI	HADS[al], AIS[am], VAS, FTQ[an], TCS[ao].	Descriptive statistics.	N/A.	N/A.	Only 22.9% of patients adhered to the application (29,864 patients were excluded because of incomplete records of personal and tinnitus-related information).	FTRS greatly reduced medical costs in comparison to traditional in-person care and enabled patients with tinnitus to arrange their own treatment times.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Tang et al [45], 2025	DCRS[ap] through a smartphone app	Unmodified music	3 months (2 h/d)	FU1: 1 month of treatment; FU2: 2 months of treatment; FU3: 3 months of treatment	THI	HADS, AIS, FTQ, TCS.	Linear mixed models and a multiclass logistic model with a stepwise function.	DCRS-treated patients after 3 months of treatment had significant tinnitus relief compared to those in the CG group ($P<.001$).	At 3 months, 92.5% of patients undergoing DCRS reported tinnitus relief or disappearance.	NS.	DCRS showed to be a promising noninvasive therapy for chronic tinnitus that can be delivered through a mobile app.	⊕⊕⊕◯

[a]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.
[b]TRT: tinnitus retraining therapy.
[c]FU1: first follow-up moment.
[d]FU2: second follow-up moment.
[e]THI: Tinnitus Handicap Inventory.
[f]QoL: quality of life.
[g]VAS: visual analog scale.
[h]CBT: cognitive behavioral therapy.
[i]TFI: Tinnitus Functional Index.
[j]PHQ-9: Patient Health Questionnaire-9.
[k]GAD-7: Generalized Anxiety Disorder-7.
[l]PSQI: Pittsburgh Sleep Quality Index.
[m]PSS: Perceived Stress Scale.
[n]IG: intervention group.
[o]N/A: not applicable.
[p]BL: baseline.
[q]miniTQ: Mini Tinnitus Questionnaire.
[r]CG: control group.
[s]GGTI: GG Tinnitus.
[t]H-THI: Hearing-Tinnitus Handicap Inventory.
[u]RCI: Reliable Change Index.
[v]FU: follow-up moment.
[w]TS: Tinnitus Scale.
[x]NRS: Numeric Rating Scale.
[y]THI: Tinnitus Handicap Inventory.
[z]WN: white noise.
[aa]COSIT: Client-Oriented Scale of Improvement in Tinnitus.
[ab]SUS: System Usability Scale.
[ac]MAUQ: mHealth App Usability Questionnaire.
[ad]USL: UpSilent app.
[ae]BDI: Beck Depression Index.
[af]SSL: Subjective Stress Level.
[ag]RM: repeated measure.
[ah]NR: not reported.

XSL•FO
RenderX

[ai]STAI: State Trait Anxiety Inventory.
[aj]NS: not specified.
[ak]FTRS: Fudan Tinnitus Relieving System.
[al]HADS: Hospital Anxiety and Depression Scale.
[am]AIS: Acceptance of Illness Scale.
[an]FTQ: Fear of Tinnitus Questionnaire.
[ao]TCS: Tinnitus Catastrophizing Scale.
[ap]DCRS: digital customized relieving sound.

Other Internet-Based Interventions

Study and Patient Characteristics

A total of 5 studies looked into the use of other internet-based interventions for chronic tinnitus, ranging from sound therapy, mindfulness, meditation, and intensive auditory cognitive training [24,28,29,31,47]. See Table S5 in Multimedia Appendix 4 for more information.

Results of Individual Studies

In total, 5 studies examined other internet-based interventions, not CBT-related [24,28,29,31,47]. All 5 studies looked into a different treatment, some more related to sound therapy, others focusing more on mindfulness or cognitive training. All studies reported promising results concerning the effectiveness of the treatment, except for Xing et al [47], who concluded that auditory intensive cognitive training was not associated with changes in self-reported tinnitus. Further details can be found in Table 5.

XSL•FO
RenderX

Table 5. Table of evidence for the studies concerning other internet-based interventions.

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Alonso-Valerdi et al. [24], 2021	(1) TRT[b], (2) TEAE[c], (3) ADT[d], (4) BBT[e], (5) music	No treatment	2 months (60 min daily)	FU1[f]: after 1 week; FU2[g]: after 5 weeks; FU3[h]: after 8 weeks	THI[i]	HADS[j], audiometry, tinnitus characteristics analysis, and BL[k] neurophysiological analysis.	Identification of ± or 0 effects *t* test.	All negative differences between BL and FU[l] were statistically significant (*P*=.009).	29.4% of tinnitus sufferers in the TRT group reported diminution of their tinnitus perception, followed by those in TEAE (18.8%), BBT or ADT (14.3%), and music (7.7%) groups.	Dropout: music 6/16; BBT[m] 4/18; TRT 3/18; TEAE 5/18; ADT 5/18; control 3/14.	TRT most effective sound-based therapy to reduce tinnitus perception for individuals without anxiety. BBT, ADT, and TEAE produced very similar effects.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Connell et al [28], 2023	Customized acoustic therapy delivered through a web-based platform	N/A[n]	6 weeks (2 h daily)	Every week of treatment	FiveQ tinnitus symptoms severity scores	N/A.	Wilcoxon signed-rank tests, Mann-Whitney U-tests, linear mixed models, Tukey method *P* value adjustment for multiple variables, and a Kenward-Roger degrees of freedom estimation.	The mean difference between pre- and posttreatment of –9.3 (SD 15.5) was statistically significant (*P*<.001).	At treatment end point, 39/58 (67.2%) patients demonstrated an overall improvement in FiveQ, 4/58 (6.9%) had no change, and 15/58 (25.9%) declined.	51% (57/115) dropout.	The findings of this study support a customized acoustic therapy program as an efficacious tinnitus treatment modality that can improve symptom severity. Patients with low-frequency tinnitus and mild or above symptom severity at baseline demonstrated a greater treatment response.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Gans et al [29], 2023	i-MBT-SR[o] focusing on applying mindfulness to tinnitus	N/A	8 weeks (2 h lesson/wk and 30 min daily instructor-led meditation practice)	FU1: posttreatment; FU2: 6 months posttreatment	TFI[p]	PSS[q] TIQ[r].	Repeated-measures 2 (group) × 3 (time) repeated-measures (using Wilks' Lambda because of heterogeneity of variance) MANOVA[s], ANOVA with the Greenhouse-Geisser method.	Post hoc contrast comparisons with Bonferroni adjustments were used to examine the differences between the levels of each main effect. Contrasts revealed that the greatest gains were made in TFI and PSS between time 1 and either time 2 or time 3 time periods.	31 (72%) patients experienced clinically meaningful improvements of 13 or more on the TFI.	534/577 (94%) did not complete FU1, and 555/577 (98%) FU2.	The iMBT-SR course appears to be a viable and effective treatment modality for the tinnitus population.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Jackson and Woolmer [31], 2024	Online body scan meditation intervention	Waiting list	8 weeks (3-4 times a week)	FU1: posttreatment; FU2: 6 months posttreatment	TFI	TCQ[t] MAAS[u].	1-way ANOVA design.	A significant main effect of condition on self-reported tinnitus distress measured by the TFI ($F_{1,100}$=18.318; P<.001; 2P=.15) was present.	The mean TFI change over 8 weeks of body scan meditation was –9.52 in comparison to –0.58 for patients on the waiting list. 29.63% of the IG[v] experienced a clinically significant change on the TFI of more than 14/100.	15/48 lost to follow-up in the control group and 9/54 in IG ≥LOCF[w].	These results evidence that those in the body scan meditation group experienced a significant reduction in tinnitus distress, an increase in mindful thought, and a decrease in negative tinnitus cognitions.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Xing et al [47], 2021	Brain HQ[x] auditory intensive training	Brain HQ nonauditory intensive training	8 weeks (5 d/wk, 20 min/d)	FU1: posttreatment; FU2: 1 month posttreatment	TFI	TGBS[y], CFQ[z], BAI[aa], PHQ-9[ab].	Spearman correlation mixed effects model with an unrestricted variance-covariance matrix.	Marginal mean differences for IG and CG[ac] were at 1 month (0.24, 95% CI –11.20 to 10.7), at 2 months 2.17 (95% CI –8.50 to 13.3), and at 3 months 3.36 (95% CI –7.91 to 14.6).	At 3 months, 13% of the CG and 19% of the IG experienced a clinically relevant reduction in the TFI of 13 or more.	>20%	Auditory intensive cognitive training was not associated with changes in self-reported tinnitus.	⊕⊕⊕◯

[a]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.
[b]TRT: tinnitus retraining therapy.
[c]TEAE: therapy for an enriched acoustic environment.
[d]ADT: auditory discrimination therapy.
[e]BBT: Broad Band Therapy.
[f]FU1: first follow-up moment.
[g]FU2: second follow-up moment.
[h]FU3: third follow-up moment.
[i]THI: Tinnitus Handicap Inventory.
[j]HADS: Hospital Anxiety and Depression Scale.
[k]BL: baseline.
[l]FU: follow-up moment.
[m]BBT: broad band therapy.
[n]N/A: not applicable.
[o]iMBTSR: internet-based mindfulness-based tinnitus stress reduction.
[p]TFI: Tinnitus Functional Index.
[q]PSS: Perceived Stress Scale.
[r]TIQ: Tinnitus Impact Questionnaire.
[s]MANOVA: multivariate ANOVA.
[t]TCQ: Tinnitus Cognitions Questionnaire.
[u]MAAS: Mindful Attention Awareness Scale.
[v]IG: intervention group.
[w]LOCF: Last Observation Carried Forward.
[x]HQ: Hyperacusis Questionnaire.
[y]TGBS: Tinnitus Global Bothersome Scale.
[z]CFQ: Cognitive Failures Questionnaire.
[aa]BAI: Beck Anxiety Inventory.
[ab]PHQ-9: Patient Health Questionnaire-9.
[ac]CG: control group.

XSL•FO
RenderX

Discussion

Principal Findings

The previous systematic review published in 2023 on the topic of telerehabilitation for patients with chronic subjective tinnitus indicated that low-contact treatment provided from a distance through several telerehabilitation forms could be used as a substitution for or an addition to in-person clinical care for tinnitus [13]. It can potentially reduce the costs of treatment without sacrificing the level of effectiveness. Due to the advances in technology made every day, telerehabilitation is a quickly evolving research topic, and several new studies on telerehabilitation for tinnitus have been published in the last couple of years. Therefore, the aim of this systematic review was to give an updated overview of the research concerning the effectiveness of telerehabilitation interventions for self-management of tinnitus published between 2022 and 2025. The recent studies included in this review confirm the earlier findings that telerehabilitation shows promising results in effectively reducing tinnitus distress in a less costly and more accessible way. However, the recent findings also demonstrate some changes within the field of telerehabilitation for tinnitus.

In the last decade, a lot of research has been published concerning the development and effectiveness of iCBT programs. These studies formed the vast majority of the preceding review. In the last 3 years, a shift in telerehabilitation research has been noticeable. The interest in iCBT has abated, and the use of smartphone apps for telerehabilitation for tinnitus is a growing topic. Not only in the field of tinnitus, but in medicine and rehabilitation in general, the use of smartphone apps to offer treatment from a distance is on the rise and forms the newest trending topic in the field of medicine [49,50]. The implementation of smartphone apps offers several advantages over conventional internet-based interventions that typically rely on desktop or browser-based platforms. First, it ensures greater accessibility and reach, increasing the likelihood of consistent patient engagement and adherence. They also enable more flexible and context-sensitive delivery of rehabilitation tasks. Push notifications, reminders, and just-in-time prompts can support adherence and promote habitual use, increasing spontaneous use. Smartphone apps also tend to perform better in terms of user friendliness and patient satisfaction [51-53]. It is not just the use of apps, but also the content of the researched apps that has shifted. While earlier research investigated the use of apps specifically for sound therapy for tinnitus patients, the newer research also shows other treatments in the form of an app [54,55]. The majority of the researched apps still gravitate more toward the use of sound therapy, but also, CBT, TRT, and psycho-education are made more accessible in this way. It is, however, noticeable that while research concerning sound therapy in app form has scaled up its sample size in the last couple of years to strengthen its power, this is not yet the case for the other treatments in app form. CBT, TRT, and psycho-education without sound therapy were still investigated in smaller sample sizes and pilot studies. Additionally, 1 study looked into the use of an app with a cervical exercise program for patients with somatic tinnitus. These patients experienced neck complaints in addition to their tinnitus, and these complaints can influence the tinnitus sound and loudness due to somatosensory interference. As a tailored care approach detecting present influencing factors is recommended, it is good that telerehabilitation is expanding and not only focusing on the psychosocial aspect, but also on specific other potential factors that might be of influence, such as neck complaints.

Another trend noticeable from the research of the last couple of years is that internet-based programs are also expanding beyond iCBT with or without guidance. This review included studies looking into several forms of sound therapy, customized acoustic therapy, mindfulness, meditation, and even cognitive training through a web-based platform. These treatments all demonstrated promising results for reducing tinnitus distress. Except for the cognitive training program with acoustic cognitive exercises, which was not reported with self-reported changes in tinnitus experiences. However, given the promising results concerning the link between cognition and tinnitus and the potential use of cognitive training for tinnitus, future studies on cognitive training telerehabilitation forms for tinnitus remain relevant [56-58]. Sound therapy, meditation, and mindfulness have been widely researched in a nontelerehabilitation context for tinnitus, making it a great avenue to explore how these treatment forms can be further improved by increasing accessibility.

Important notes from the preceding review were the high dropout rates from iCBT programs and the lack of reporting dropout in studies concerning the use of smartphone apps [13]. The same note can still be made in the current review. For iCBT, 2 of 5 studies reported dropout, which was around 34% at postintervention follow-up. This is in line with previous similar studies concerning iCBT in this population. According to Beukes et al [26], patients withdraw from treatment due to time constraints or lack of engagement. In accordance, the preceding review stated that earlier research indicated that barriers such as lack of time, engagement, motivation, and openness of the patient cause high dropout and should be taken into consideration. The previous review also concluded that future research should further explore which factors are most likely to cause the lack of compliance and how clinicians can counteract these. Therefore, it is crucial that studies analyze and report where the dropout is coming from, as Beukes et al [26] did extensively in their discussion. For the smartphone apps, a change could be noted, namely, the dropout rates were reported in 8 of 13 studies and ranged from no dropout to 53%. In other self-management intervention forms discussed, the rate ranged up to 94%, but on average was between 34% and 63%. The reasons for dropout noted were related to practical obstacles or treatment preferences from the patients. Oron et al [36] offered that monitoring programs might enhance the retention of treatment. In other treatment forms, the dropout rate and reasons for dropout were often not listed. As stated in the preceding review, to enhance the effectiveness of telerehabilitation, the limited current understanding of predictors of dropout should be tackled. Additionally, potential ways to improve adherence, such as therapist guidance levels, motivational feedback, automated reminders, etc, should be explored further [59].

XSL•FO
RenderX

Limitations

The included studies of this systematic review presented similar limitations as earlier research concerning this topic summarized in the previous review. All included studies scored some to great concerns of RoB. As a result, it needs to be acknowledged that the overall certainty of evidence ranged from low to moderate certainty. In addition, some crucial confounding parameters, such as the presence of hearing loss, hyperacusis, anxiety, depression, or sleeping problems, were not taken into consideration by all studies. However, it needs to be noted that concerning the presence of hearing loss, improvements were made on this level. A total of 15 of 24 studies reported how large the proportion of participants with hearing loss was or how many patients used hearing aids. Yet, how they defined the presence of hearing loss was often not clarified. An audiometric examination of tinnitus patients is therefore of value when researching patients with tinnitus because hearing loss is an important risk factor for developing tinnitus [1]. Due to the additional distress of hearing loss, tinnitus patients with hearing loss might present with worse tinnitus distress levels in comparison to tinnitus patients without hearing loss, according to earlier research [60].

While hearing loss is an important confounding factor to keep in mind, other factors also need to be considered: gender, age, mental state, hyperacusis, and neck and jaw complaints [61-71]. These factors might play a role in how patients perceive their tinnitus and how they react to treatment forms [61-71]. Additionally, the link between tinnitus and comorbidities such as anxiety, depression, stress, sleeping problems, hyperacusis, etc, are already widely researched [66,72-74]. The more serious these comorbidities are, the worse the prognosis of the tinnitus [73]. All the factors mentioned above might have an impact on the effectiveness of a treatment and should, therefore, be considered in future research. Additionally, for this reason, there is an increasing need for personalization of the treatment approach. Currently, often 1 factor is taken into consideration; however, a multidisciplinary and tailored approach is recommended. Therefore, the generalizability of these results is low, tinnitus mechanisms can greatly differ individually, and the limited tailoring options of these self-management interventions cause the evidence concerning the efficacy to remain rather limited. Due to a lack of high-quality and homogeneous studies concerning this topic, no confident conclusions can be made concerning the generalizability of self-management interventions for patients with tinnitus. Telerehabilitation has the potential to make treatment tailored to the needs of the patients more accessible; however, future research should invest more into using this potential. Future studies should focus on delivering high-quality study designs with homogenous outcomes to build the evidence of the strengths of telerehabilitation, which are now considered potential benefits.

Conclusions

This updated systematic review concerning the effectiveness of telerehabilitation for patients with chronic subjective tinnitus confirms the conclusion stated in the preceding review. The results of this systematic review indicate that telerehabilitation in the form of iCBT with guidance, self-help manuals, self-help devices, smartphone apps, and other internet-based interventions (except for cognitive training) are effective in relieving tinnitus distress. Telerehabilitation might form an alternative or additional tool to the recommended in-person care that patients currently receive. This review accentuated the shift from iCBT to the growing interest in the use of smartphone apps, increasing the accessibility of the treatments even more. However, the most significant barrier to the success of telerehabilitation remains the lack of compliance with the treatment. The great expansion of research within the field of telerehabilitation for tinnitus treatment in recent years indicates that it remains an evolving topic to keep track of.

Acknowledgments

The authors gratefully acknowledge Fonds Wetenschappelijk Onderzoek for its financial support. No generative artificial intelligence tools (including large language models such as ChatGPT, OpenAI) were used in the writing, editing, data analysis, interpretation of results, or preparation of tables and figures for this paper. All content was produced solely by the authors.

Funding

This study was supported by the Research Foundation – Flanders (Fonds Wetenschappelijk Onderzoek). The funding agency had no role in the design of this study; in the collection, analysis, and interpretation of data; or in writing the paper.

Data Availability

All data underlying the findings of this systematic review are derived from the studies included in the review and are publicly available through the original publications, which are fully referenced in this paper. No additional primary data were generated or collected for this work. Extracted datasets and materials used for the review process are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

PRISMA checklist for systematic review.
[PDF File (Adobe PDF File), 48 KB - jmir_v28i1e83529_app1.pdf]

Multimedia Appendix 2
Search strategy updated.
[DOC File , 33 KB - jmir_v28i1e83529_app2.doc]

Multimedia Appendix 3
Risk of bias summarization of newly included papers.
[DOCX File , 91 KB - jmir_v28i1e83529_app3.docx]

Multimedia Appendix 4
Tables of study and patient characteristics of newly included papers.
[DOCX File , 29 KB - jmir_v28i1e83529_app4.docx]

References

1. Baguley D, McFerran D, Hall D. Tinnitus. Lancet 2013;382(9904):1600-1607. [doi: 10.1016/s0140-6736(13)60142-7]
2. De Ridder D, Schlee W, Vanneste S, Londero A, Weisz N, Kleinjung T, et al. Tinnitus and tinnitus disorder: theoretical and operational definitions (an international multidisciplinary proposal). Prog Brain Res 2021;260:1-25. [doi: 10.1016/bs.pbr.2020.12.002] [Medline: 33637213]
3. Jarach CM, Lugo A, Scala M, van den Brandt PA, Cederroth CR, Odone A, et al. Global prevalence and incidence of tinnitus: a systematic review and meta-analysis. JAMA Neurol 2022;79(9):888-900. [doi: 10.1001/jamaneurol.2022.2189] [Medline: 35939312]
4. Demoen S, Cardon E, Jacquemin L, Timmermans A, Van Rompaey V, Gilles A, et al. Health-related quality of life in subjective, chronic tinnitus patients: a scoping review. J Assoc Res Otolaryngol 2024;25(2):103-129. [doi: 10.1007/s10162-024-00926-5] [Medline: 38253898]
5. Soni A, Dubey A. Chronic primary tinnitus: a management dilemma. Audiol Res 2020;10(2):55-66. [doi: 10.3390/audiolres10020010] [Medline: 33255533]
6. Han BI, Lee HW, Ryu S, Kim J. Tinnitus update. J Clin Neurol 2021;17(1):1-10. [doi: 10.3988/jcn.2021.17.1.1] [Medline: 33480192]
7. Maes IH, Cima RFF, Vlaeyen JW, Anteunis LJC, Joore MA. Tinnitus: a cost study. Ear Hear 2013;34(4):508-514. [doi: 10.1097/AUD.0b013e31827d113a] [Medline: 23411656]
8. Trochidis I, Lugo A, Borroni E, Cederroth CR, Cima R, Kikidis D, et al. Systematic review on healthcare and societal costs of tinnitus. Int J Environ Res Public Health 2021;18(13):6881. [doi: 10.3390/ijerph18136881] [Medline: 34206904]
9. Simões JP, Neff PKA, Langguth B, Schlee W, Schecklmann M. The progression of chronic tinnitus over the years. Sci Rep 2021;11(1):4162. [doi: 10.1038/s41598-021-83068-5] [Medline: 33602995]
10. Bokolo AJ. Exploring the adoption of telemedicine and virtual software for care of outpatients during and after COVID-19 pandemic. Ir J Med Sci 2021;190(1):1-10. [doi: 10.1007/s11845-020-02299-z] [Medline: 32642981]
11. Scaglione T, Kuzbyt B. Tinnitus management: the utilization of a hybrid telehealth and in-person delivery model. Semin Hear 2021;42(2):115-122. [doi: 10.1055/s-0041-1731692] [Medline: 34381295]
12. Beukes EW, Lourenco MPCG, Biot L, Andersson G, Kaldo V, Manchaiah V, et al. Suggestions for shaping tinnitus service provision in Western Europe: lessons from the COVID-19 pandemic. Int J Clin Pract 2021;75(7):e14196. [doi: 10.1111/ijcp.14196] [Medline: 33837639]
13. Demoen S, Chalimourdas A, Timmermans A, Van Rompaey V, Vanderveken OM, Jacquemin L, et al. Effectiveness of telerehabilitation interventions for self-management of tinnitus: systematic review. J Med Internet Res 2023;25:e39076. [doi: 10.2196/39076] [Medline: 36757768]
14. Searchfield GD, Sanders PJ, Doborjeh Z, Doborjeh M, Boldu R, Sun K, et al. A state-of-art review of digital technologies for the next generation of tinnitus therapeutics. Front Digit Health 2021;3:724370 [FREE Full text] [doi: 10.3389/fdgth.2021.724370] [Medline: 34713191]
15. Langguth B, Kleinjung T, Schlee W, Vanneste S, De Ridder D. Tinnitus guidelines and their evidence base. J Clin Med 2023;12(9):3087 [FREE Full text] [doi: 10.3390/jcm12093087] [Medline: 37176527]
16. Thomas EE, Haydon HM, Mehrotra A, Caffery LJ, Snoswell CL, Banbury A, et al. Building on the momentum: sustaining telehealth beyond COVID-19. J Telemed Telecare 2020;28(4):301-308. [doi: 10.1177/1357633x20960638]
17. Saharkhiz M, Rao T, Parker-Lue S, Borelli S, Johnson K, Cataife G. Telehealth expansion and medicare beneficiaries' care quality and access. JAMA Netw Open 2024;7(5):e2411006 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.11006] [Medline: 38739388]

XSL•FO
RenderX

18. Morelli S, Daniele C, D'Avenio G, Grigioni M, Giansanti D. Optimizing telehealth: leveraging key performance indicators for enhanced telehealth and digital healthcare outcomes (telemechron study). Healthcare (Basel) 2024;12(13):1319 [FREE Full text] [doi: 10.3390/healthcare12131319] [Medline: 38998854]
19. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
20. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
21. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
22. McGuinness LA, Higgins JPT. Risk-of-bias visualization (robvis): an R package and Shiny web app for visualizing risk-of-bias assessments. Res Synth Methods 2021;12(1):55-61. [doi: 10.1002/jrsm.1411] [Medline: 32336025]
23. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336(7650):924-926. [doi: 10.1136/bmj.39489.470347.ad]
24. Alonso-Valerdi LM, González-Salazar JI, Ibarra-Zarate DI. Neuropsychological monitoring of current acoustic therapies as alternative treatment of chronic tinnitus. Am J Otolaryngol 2021;42(6):103109. [doi: 10.1016/j.amjoto.2021.103109] [Medline: 34175772]
25. Altissimi G, Musacchio A, Pace A, Greco A, Magliulo G, Iagnocco G, et al. Sound therapy in patients with tinnitus: traditional sound generators vs. mobile apps. Eur Rev Med Pharmacol Sci 2024;28(11):3781-3786. [doi: 10.26355/eurrev_202406_36384] [Medline: 38884513]
26. Beukes EW, Andersson G, Fagelson M, Manchaiah V. Audiologist-supported internet-based cognitive behavioral therapy for tinnitus in the United States: a pilot trial. Am J Audiol 2021;30(3):717-729. [doi: 10.1044/2021_aja-20-00222]
27. Beukes EW, Andersson G, Manchaiah V. Long-term efficacy of audiologist-guided internet-based cognitive behaviour therapy for tinnitus in the United States: a repeated-measures design. Internet Interv 2022;30:100583. [doi: 10.1016/j.invent.2022.100583] [Medline: 36353148]
28. Connell JT, Bassiouni A, Harrison E, Laden S, O'Brien S, Sahota R, et al. Customised acoustic therapy delivered through a web-based platform-an innovative approach to tinnitus treatment. Clin Otolaryngol 2023;48(2):226-234. [doi: 10.1111/coa.14027] [Medline: 36550768]
29. Gans JJ, Holst J, Holmes C, Hudock D. Healing from home: examination of an online mindfulness-based tinnitus stress reduction course during the 2020 COVID pandemic. Am J Audiol 2023;32(1):160-169. [doi: 10.1044/2022_aja-22-00063]
30. Goshtasbi K, Tawk K, Khosravi P, Abouzari M, Djalilian HR. Smartphone-based cognitive behavioral therapy and customized sound therapy for tinnitus: a randomized controlled trial. Ann Otol Rhinol Laryngol 2025;134(2):125-133. [doi: 10.1177/00034894241297594] [Medline: 39506291]
31. Jackson JG, Woolmer CD. An 8-week online body scan meditation intervention for tinnitus: accessibility, adherence, and rates of clinically meaningful success. Mindfulness 2024;15(5):1176-1187. [doi: 10.1007/s12671-024-02357-y]
32. Kutyba J, Gos E, Jędrzejczak WW, Raj-Koziak D, Karpiesz L, Niedziałek I, et al. Effectiveness of tinnitus therapy using a mobile application. Eur Arch Otorhinolaryngol 2022;279(3):1257-1267 [FREE Full text] [doi: 10.1007/s00405-021-06767-9] [Medline: 33783597]
33. Kutyba JJ, Jędrzejczak WW, Gos E, Raj-Koziak D, Skarzynski PH. Chronic tinnitus and the positive effects of sound treatment via a smartphone app: mixed-design study. JMIR mHealth uHealth 2022;10(4):e33543. [doi: 10.2196/33543] [Medline: 35451975]
34. Manchaiah V, Beukes EW, Andersson G, Bateman E, Swanepoel DW, Uhler K, Vinay. Individuals with tinnitus report more positive experiences following internet-based cognitive behavioral therapy. Clin Pract 2024;14(4):1615-1624 [FREE Full text] [doi: 10.3390/clinpract14040130] [Medline: 39194934]
35. Michiels S, Wölflick S, Simões J, Schlee W. Exploring app-based physiotherapy for somatic tinnitus: results from a pilot study. J Clin Med 2024;13(23):7203 [FREE Full text] [doi: 10.3390/jcm13237203] [Medline: 39685670]
36. Oron Y, Ben David BM, Doron G. Brief cognitive-behavioral training for tinnitus relief using a mobile application: a pilot open trial. Health Inf J 2022;28(1):14604582221083483 [FREE Full text] [doi: 10.1177/14604582221083483] [Medline: 35349777]
37. Ravi P, Samayan K, Sivagnanapandian D. Tinnitus retraining therapy: in person versus mobile application based - two case scenarios. Indian J Otolaryngol Head Neck Surg 2025;77(2):1142-1146. [doi: 10.1007/s12070-024-05300-8] [Medline: 40070760]
38. Rodrigo H, Beukes EW, Andersson G, Manchaiah V. Internet-based cognitive-behavioural therapy for tinnitus: secondary analysis to examine predictors of outcomes. BMJ Open 2021;11(8):e049384 [FREE Full text] [doi: 10.1136/bmjopen-2021-049384] [Medline: 34417217]
39. Rodrigo H, Beukes EW, Andersson G, Manchaiah V. Exploratory data mining techniques (decision tree models) for examining the impact of internet-based cognitive behavioral therapy for tinnitus: machine learning approach. J Med Internet Res 2021;23(11):e28999 [FREE Full text] [doi: 10.2196/28999] [Medline: 34726612]

XSL•FO
RenderX

40. Schlee W, Neff P, Simoes J, Langguth B, Schoisswohl S, Steinberger H, et al. Smartphone-guided educational counseling and self-help for chronic tinnitus. J Clin Med 2022;11(7):1825 [FREE Full text] [doi: 10.3390/jcm11071825] [Medline: 35407432]
41. Searchfield GD, Sanders PJ. A randomized single-blind controlled trial of a prototype digital polytherapeutic for tinnitus. Front Neurol 2022;13:958730. [doi: 10.3389/fneur.2022.958730] [Medline: 35989940]
42. Seol HY, Jo M, Moon IJ. The effectiveness of a smartphone application for tinnitus relief. Healthcare (Basel) 2023;11(17):2368. [doi: 10.3390/healthcare11172368] [Medline: 37685402]
43. Shetty HN. Development and validation of self-management manual for acute tinnitus. Indian J Otolaryngol Head Neck Surg 2024;76(3):2654-2659. [doi: 10.1007/s12070-024-04504-2] [Medline: 38883543]
44. Suh MW, Park MK, Kim Y, Kim YH. The treatment outcome of smart device-based tinnitus retraining therapy: prospective cohort study. JMIR mHealth uHealth 2023;11:e38986 [FREE Full text] [doi: 10.2196/38986] [Medline: 36633890]
45. Tang D, Peng Y, Gu D, Wu Y, Li H. Digital frequency customized relieving sound for chronic subjective tinnitus management: prospective controlled study. J Med Internet Res 2025;27:e60150 [FREE Full text] [doi: 10.2196/60150] [Medline: 39823588]
46. Tang D, Wang H, Gu D, Ye L, Sun S, Li H. The fudan tinnitus relieving system application for tinnitus management. BMC Med Inform Decis Mak 2023;23(1):76. [doi: 10.1186/s12911-023-02164-w] [Medline: 37085904]
47. Xing M, Kallogjeri D, Piccirillo JF. Investigating the impact of cognitive training for individuals with bothersome tinnitus: a randomized controlled trial. Otolaryngol Head Neck Surg 2021;165(6):854-861. [doi: 10.1177/0194599821994742] [Medline: 33650921]
48. Andersson G, Kaldo V. Internet-based cognitive behavioral therapy for tinnitus. J Clin Psychol 2004;60(2):171-178. [doi: 10.1002/jclp.10243] [Medline: 14724924]
49. Nussbaum R, Kelly C, Quinby E, Mac A, Parmanto B, Dicianno BE. Systematic review of mobile health applications in rehabilitation. Arch Phys Med Rehabil 2019;100(1):115-127. [doi: 10.1016/j.apmr.2018.07.439] [Medline: 30171827]
50. Moral-Munoz JA, Zhang W, Cobo MJ, Herrera-Viedma E, Kaber DB. Smartphone-based systems for physical rehabilitation applications: a systematic review. Assist Technol 2021;33(4):223-236. [doi: 10.1080/10400435.2019.1611676] [Medline: 31112461]
51. Lee C, Ahn J, Lee BC. A systematic review of the long-term effects of using smartphone- and tablet-based rehabilitation technology for balance and gait training and exercise programs. Bioengineering (Basel) 2023;10(10):1142. [doi: 10.3390/bioengineering10101142] [Medline: 37892872]
52. Arensman R, Kloek C, Pisters M, Koppenaal T, Ostelo R, Veenhof C. Patient perspectives on using a smartphone app to support home-based exercise during physical therapy treatment: qualitative study. JMIR Hum Factors 2022;9(3):e35316 [FREE Full text] [doi: 10.2196/35316] [Medline: 36098993]
53. Simmich J, Ross MH, Russell T. Real-time video telerehabilitation shows comparable satisfaction and similar or better attendance and adherence compared with in-person physiotherapy: a systematic review. J Physiother 2024;70(3):181-192 [FREE Full text] [doi: 10.1016/j.jphys.2024.06.001] [Medline: 38879432]
54. Mehdi M, Dode A, Pryss R, Schlee W, Reichert M, Hauck FJ. Contemporary review of smartphone apps for tinnitus management and treatment. Brain Sci 2020;10(11):867 [FREE Full text] [doi: 10.3390/brainsci10110867] [Medline: 33212928]
55. Mehdi M, Stach M, Riha C, Neff P, Dode A, Pryss R, et al. Smartphone and mobile health apps for tinnitus: systematic identification, analysis, and assessment. JMIR mHealth uHealth 2020;8(8):e21767 [FREE Full text] [doi: 10.2196/21767] [Medline: 32808939]
56. Kallogjeri D, Piccirillo JF, Spitznagel E, Hale S, Nicklaus JE, Hardin FM, et al. Cognitive training for adults with bothersome tinnitus: a randomized clinical trial. JAMA Otolaryngol Head Neck Surg 2017;143(5):443-451. [doi: 10.1001/jamaoto.2016.3779] [Medline: 28114646]
57. Tegg-Quinn S, Bennett RJ, Eikelboom RH, Baguley DM. The impact of tinnitus upon cognition in adults: a systematic review. Int J Audiol 2016;55(10):533-540. [doi: 10.1080/14992027.2016.1185168] [Medline: 27240696]
58. Henry JL, Wilson PH. An evaluation of two types of cognitive intervention in the management of chronic tinnitus. Scand J Behav Ther 1998;27(4):156-166. [doi: 10.1080/02845719808408510]
59. Simon N, Robertson L, Lewis C, Roberts NP, Bethell A, Dawson S, et al. Internet-based cognitive and behavioural therapies for post-traumatic stress disorder (PTSD) in adults. Cochrane Database Syst Rev 2021;5(5):CD011710 [FREE Full text] [doi: 10.1002/14651858.CD011710.pub3] [Medline: 34015141]
60. Mahafza N, Zhao F, El Refaie A, Chen F. A comparison of the severity of tinnitus in patients with and without hearing loss using the tinnitus functional index (TFI). Int J Audiol 2021;60(3):220-226. [doi: 10.1080/14992027.2020.1804081] [Medline: 32787604]
61. Seydel C, Haupt H, Olze H, Szczepek AJ, Mazurek B. Gender and chronic tinnitus: differences in tinnitus-related distress depend on age and duration of tinnitus. Ear Hear 2013;34(5):661-672. [doi: 10.1097/AUD.0b013e31828149f2] [Medline: 23439056]
62. Han TS, Jeong J, Park S, Kim JJ. Gender differences affecting psychiatric distress and tinnitus severity. Clin Psychopharmacol Neurosci 2019;17(1):113-120. [doi: 10.9758/cpn.2019.17.1.113] [Medline: 30690946]

XSL•FO
RenderX

63. Hiller W, Goebel G. Factors influencing tinnitus loudness and annoyance. Arch Otolaryngol Head Neck Surg 2006;132(12):1323-1330. [doi: 10.1001/archotol.132.12.1323] [Medline: 17178943]
64. Lugo A, Trpchevska N, Liu X, Biswas R, Magnusson C, Gallus S, et al. Sex-specific association of tinnitus with suicide attempts. JAMA Otolaryngol Head Neck Surg 2019;145(7):685-687. [doi: 10.1001/jamaoto.2019.0566] [Medline: 31046059]
65. Van der Wal A, Luyten T, Cardon E, Jacquemin L, Vanderveken OM, Topsakal V, et al. Sex differences in the response to different tinnitus treatment. Front Neurosci 2020;14:422. [doi: 10.3389/fnins.2020.00422] [Medline: 32477049]
66. Boecking B, Biehl R, Brueggemann P, Mazurek B. Health-related quality of life, depressive symptoms, anxiety, and somatization symptoms in male and female patients with chronic tinnitus. J Clin Med 2021;10(13):2798 [FREE Full text] [doi: 10.3390/jcm10132798] [Medline: 34202097]
67. Al-Swiahb J, Park SN. Characterization of tinnitus in different age groups: a retrospective review. Noise Health 2016;18(83):214-219 [FREE Full text] [doi: 10.4103/1463-1741.189240] [Medline: 27569409]
68. Shargorodsky J, Curhan SG, Curhan GC, Eavey R. Change in prevalence of hearing loss in US adolescents. JAMA 2010;304(7):772-778. [doi: 10.1001/jama.2010.1124] [Medline: 20716740]
69. Michiels S, Cardon E, Gilles A, Goedhart H, Vesala M, Schlee W. Somatosensory tinnitus diagnosis: diagnostic value of existing criteria. Ear Hear 2022;43(1):143-149. [doi: 10.1097/AUD.0000000000001105] [Medline: 34261856]
70. Michiels S, Sanchez TG, Oron Y, Gilles A, Haider HF, Erlandsson S, et al. Diagnostic criteria for somatosensory tinnitus: a Delphi process and face-to-face meeting to establish consensus. Trends Hear 2018;22:2331216518796403. [doi: 10.1177/2331216518796403] [Medline: 30213235]
71. Michiels S, Naessens S, Van de Heyning P, Braem M, Visscher CM, Gilles A, et al. The effect of physical therapy treatment in patients with subjective tinnitus: a systematic review. Front Neurosci 2016;10:545 [FREE Full text] [doi: 10.3389/fnins.2016.00545] [Medline: 27965530]
72. Li YL, Hsu Y, Lin C, Wu J. Sleep disturbance and psychological distress in adult patients with tinnitus. J Formosan Med Assoc 2022;121(5):995-1002. [doi: 10.1016/j.jfma.2021.07.022] [Medline: 34366185]
73. Hébert S. Psychological comorbidities of tinnitus. In: Searchfield GD, Zhang J, editors. The Behavioral Neuroscience of Tinnitus. Cham: Springer International Publishing; 2021:349-359.
74. Jacquemin L, Cardon E, Michiels S, Luyten T, Van der Wal A, De Hertogh W, et al. Hyperacusis: demographic, audiological, and clinical characteristics of patients at the ENT department. Eur Arch Oto-Rhino-Laryngol 2022;279(10):4899-4907. [doi: 10.1007/s00405-022-07336-4] [Medline: 35298688]

Abbreviations

CBT: cognitive behavioral therapy
CG: control group
GRADE: Grading of Recommendations, Assessment, Development, and Evaluation
iCBT: internet-based cognitive behavioral therapy
MeSH: Medical Subject Headings
PICO: population, intervention, comparison, and outcome
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
RCT: randomized controlled trial
RoB: risk of bias
ROBINS-I: Risk of Bias in Non-Randomized Studies of Interventions
TFI: Tinnitus Functional Index
TRT: tinnitus retraining therapy

Edited by S Brini; submitted 05.Sep.2025; peer-reviewed by A Chalimourdas, AA Danesh; comments to author 06.Nov.2025; revised version received 07.Jan.2026; accepted 07.Jan.2026; published 27.Feb.2026.

Please cite as:
Demoen S, Van Kerchove E, Timmermans A, Van Rompaey V, Michiels S, Gilles A
Effectiveness of Telerehabilitation Interventions for Self-Management of Tinnitus: Update of a Systematic Review
J Med Internet Res 2026;28:e83529
URL: https://www.jmir.org/2026/1/e83529
doi:10.2196/83529
PMID:

XSL•FO
RenderX

©Sara Demoen, Elise Van Kerchove, Annick Timmermans, Vincent Van Rompaey, Sarah Michiels, Annick Gilles. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Digital Health Interventions to Promote Physical Activity Among Adolescents: Systematic Review

Rui Shi Fan[1,2*], BEd; Jia Jun Jiang[1], BEd; Qing Yuan Zhou[1], BEd; Xin Yue Zhang[1,2*], MEd; Zhou Hang Wu[1,2], MEd; Liu Ji[1,2], PhD

[1]College of Physical Education and Health, East China Normal University, ShangHai, Shanghai, China

[2]Key Laboratory of Adolescent Health Assessment and Exercise Intervention of Ministry of Education, East China Normal University, ShangHai, China

[*]these authors contributed equally

Corresponding Author:
Liu Ji, PhD
College of Physical Education and Health
East China Normal University
500 Dongchuan Road
ShangHai, 200241
China
Phone: 86 13901817299
Email: lji@tyxx.ecnu.edu.cn

Abstract

Background: Insufficient physical activity among adolescents is a major global public health concern. Digital health interventions (DHIs) have gained increasing attention as a promising approach to promoting physical activity in adolescents. However, existing systematic reviews predominantly focus on single-intervention formats or specific study designs, while reviews that integrate multiple DHIs and diverse study designs remain scarce.

Objective: This systematic review aims to synthesize evidence from diverse DHIs and multiple study designs to assess their effectiveness in promoting physical activity among adolescents.

Methods: The review protocol was registered in PROSPERO (International Prospective Register of Systematic Reviews; CRD420251117923). This systematic review searched literature published between January 1, 2014, and June 30, 2025, across Web of Science, PubMed, EBSCO, Scopus, Embase, the Cochrane Library, ProQuest, and Google Scholar. The final search was completed on August 3, 2025. Using the PICOS (population, intervention, comparator, outcomes, and study design) framework, the review included adolescents aged 10-19 years and focused on evidence-based research promoting physical activity through DHIs. The review was limited to peer-reviewed English-language literature and excluded studies solely focused on measurement tools, those not evaluating intervention effectiveness, or those not involving adolescents. Two reviewers independently screened studies and extracted data. Research quality was assessed using the Joanna Briggs Institute tool. Findings were synthesized through narrative synthesis and qualitative content analysis.

Results: A total of 24 studies were included, involving approximately 12,183 adolescents. Study designs comprised 10 randomized controlled trials, 4 quasi-experimental studies, 3 quantitative research studies, 3 cross-sectional studies, and 4 mixed methods studies. Overall, 7 (29%) studies were of high quality, 16 (67%) were of moderate quality, and 1 (4%) was of low quality. Study populations included general adolescents as well as subgroups with specific health risks: insufficient physical activity (1/24, 4%), obesity or overweight (4/24, 17%), attention-deficit/hyperactivity disorder (1/24, 4%), cancer survivors (1/24, 4%), and at-risk youth (1/24, 4%). DHIs were categorized into 3 types: single-driver interventions (14/24, 58%), multimodal integrated interventions (7/24, 29%), and interaction-enhanced interventions (3/24, 13%). Most studies reported positive outcomes, including direct effectiveness (15/24, 63%), indirect effectiveness (8/24, 33%), and unclear effectiveness (1/24, 4%).

Conclusions: This systematic review synthesizes evidence from diverse research designs and multiple types of DHIs, offering a more comprehensive perspective than previous reviews focused on single designs or technological formats. The results indicate that DHIs generally enhance adolescent physical activity levels, although their effectiveness varies considerably across intervention types and study designs. The review fills key research gaps and highlights the critical role of intervention adaptability and implementation context. It also addresses practical concerns, including adolescents with special health conditions, digital health inequalities, and technology dependency. Despite limitations related to methodological quality and insufficient follow-up, this

XSL•FO
RenderX

review provides important evidence to inform practical application, policy development, and the equitable promotion of DHIs to enhance adolescent physical activity. Against the backdrop of rising global adolescent physical inactivity and widening health disparities, it also outlines directions for future high-quality research.

Trial Registration: PROSPERO CRD420251117923; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251117923

(J Med Internet Res 2026;28:e82395) doi:10.2196/82395

KEYWORDS

adolescents; physical activity; digital health interventions; systematic review; health promotion; behavior change; youth health promotion

Introduction

Background

Adolescence represents a pivotal developmental period for the establishment of long-term healthy lifestyle behaviors. Engaging in moderate physical activity during this stage contributes to physical development, improves cardiorespiratory fitness, supports mental health regulation, and shapes future health trajectories, including the prevention and management of chronic diseases. However, the prevalence of insufficient physical activity among adolescents has increased markedly in recent years and is now widely recognized as a significant global public health concern affecting youth populations [1].

Against this backdrop, digital health interventions (DHIs), an emerging approach to health promotion, have been increasingly implemented among adolescent populations [2]. These interventions typically utilize digital technologies—such as mobile health (mHealth) apps, wearable devices, and online fitness platforms—to facilitate behavioral change among adolescents. They offer advantages such as high accessibility and real-time feedback, rendering them particularly suitable for this demographic group [3,4]. For example, within the domain of gamified mobile apps, fitness platforms such as “Nike Training Club” enhance daily exercise participation through mechanisms including instructional video tutorials and task-based check-ins. Wearable devices, such as Fitbit and Garmin, deliver real-time feedback on physical activity indicators, including step count, heart rate, and energy expenditure, thereby supporting users in enhancing self-regulation of physical activity [5]. Overall, DHIs exhibit growing diversity and adaptability in promoting physical activity among adolescents, while simultaneously presenting novel research opportunities and methodological challenges within the field.

As DHIs evolve, understanding the underlying behavioral mechanisms is key to optimizing intervention design and improving effectiveness. The Behavior Change Wheel (BCW) explains how DHIs influence behavior change and provides a foundation for developing targeted strategies. Research suggests that the COM-B (capability, opportunity, motivation—behavior) system is a robust framework for designing DHIs. The COM-B system elucidates the mechanisms underlying behavior change [6], explaining how different technologies and strategies operate across various contexts. Proposed by Michie et al [6], this model is central to the BCW and has been widely applied in health behavior research. It explains how multimechanism strategies work together to promote health behavior change. The system proposes that behavior requires 3 elements: (1) capability, (2) opportunity, and (3) motivation. These elements interact to form a behavioral system; for example, enhanced capability can increase motivation, favorable opportunities can facilitate the use of capability, and higher motivation can drive further pursuit and development.

In recent years, DHIs have undergone continuous innovation in content strategy. The content of such interventions has evolved from single-dimensional information delivery to multidimensional models that integrate behavioral incentives, motivational guidance, and social support. Similarly, intervention modalities have transitioned from single-medium approaches to integrated, multicomponent systems that combine mobile apps, wearable technologies, and social networking platforms. Despite these advancements, research on DHIs aimed at promoting physical activity among adolescents remains substantially underdeveloped. At the level of primary research, existing studies targeting physical activity promotion among children and adolescents demonstrate considerable diversity in intervention modalities; however, a standardized classification framework remains absent. Moreover, substantial variation is observed across studies with respect to intervention tools, implementation frequency, and intervention duration, contributing to pronounced heterogeneity in reported outcomes [7]. Simultaneously, methodological weaknesses persist in some studies—such as the absence of control groups, small sample sizes, short intervention periods, or a lack of long-term follow-up—thereby undermining the interpretability and external validity of the findings [8].

Existing reviews primarily focus on specific DHIs, with few systematic reviews encompassing multiple intervention types and diverse study designs. Previous reviews, such as that by Wang et al [9], focused on mHealth app interventions using randomized controlled trials (RCTs) and reported benefits for total physical activity, sedentary behavior, BMI, agility, and muscle strength in children and adolescents. He et al [10] examined smartphone-based interventions using RCTs, showing increases in total physical activity and step counts among children and adolescents, with app-based interventions proving more effective. Sequí-Domínguez et al [11] examined eHealth interventions in RCTs and found that they reduced sedentary time in children and adolescents.

The scope of these previous systematic reviews was relatively narrow, focusing solely on specific DHIs or intervention types. The included studies were limited to a single study design,

XSL•FO
RenderX

resulting in incomplete interpretation of the evidence. Consequently, there remains a lack of comprehensive integration and comparison across multiple DHIs, as well as a paucity of reviews and further analyses encompassing diverse study designs.

Objectives

Against this backdrop, this systematic review aims to conduct a focused analysis of the existing literature targeting adolescent populations. It seeks to identify and examine the capabilities of different research designs and various types of DHIs in promoting physical activity among adolescents, thereby addressing current evidence gaps. The review is intended to provide theoretical support and an empirical foundation for optimizing subsequent intervention design and research.

Methods

Review Design and Protocol Registration

This systematic review adopts a systematic methodology to comprehensively examine the current landscape, intervention types, and implementation effectiveness of DHIs aimed at promoting physical activity among adolescents in recent years. In addition, it evaluates and synthesizes the methodological quality and practical outcomes reported in the existing literature. The review strictly adheres to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 reporting guidelines [12], encompassing key steps including literature retrieval, screening, data extraction, and quality appraisal. To ensure scientific rigor and interstudy comparability, a structured risk-of-bias assessment was performed using the Joanna Briggs Institute (JBI) Critical Appraisal Tool [13]. The review protocol was prospectively registered with the international systematic review registry PROSPERO (International Prospective Register of Systematic Reviews; registration number CRD420251117923), and the study was conducted and reported in strict accordance with the preestablished protocol. The study protocol can be accessed via the corresponding registration record on the PROSPERO platform.

Eligibility Criteria

Inclusion Criteria

The systematic review utilized the PICOS (population, intervention, comparator, outcomes, and study design) framework to define the following inclusion criteria.

- Population: Studies must include adolescent participants aged 10-19 years, with a primary focus on physical activity behaviors during this developmental stage.
- Intervention: Studies must examine DHIs aimed at promoting physical activity among adolescents. Eligible intervention formats include, but are not limited to, mHealth, wearable technologies, eHealth platforms, fitness tracking devices, virtual fitness programs, and gamified health apps.
- Comparison: Acceptable control conditions may include no intervention, traditional physical activity programs, or other nondigital health promotion approaches.
- Outcomes: Included studies must report outcomes related to adolescent physical activity behaviors, including but not limited to physical activity levels, exercise frequency, intensity of participation, sports involvement, and engagement in fitness-related activities.
- Study design: Eligible study designs include RCTs, quasi-experimental studies (QESs), quantitative research (QR), cross-sectional studies (CSSs), and mixed methods studies (MMSs).
- All included studies must be peer-reviewed publications written in English and published no later than June 30, 2025.

Exclusion Criteria

- Systematic reviews that exclusively examined physical activity assessment tools or indicators without evaluating intervention effectiveness.
- Systematic reviews involving participants who were not adolescents or whose data did not allow clear differentiation of adolescent subgroups.
- Systematic reviews classified as dissertations (master's or doctoral), review articles, conference papers, or other non–peer-reviewed publications.
- Systematic reviews for which the full text was not accessible.
- Systematic reviews with unclear research designs or markedly low methodological quality.

Information Sources

This systematic review retrieved relevant studies from multiple databases and information sources. Primary databases included PubMed (National Library of Medicine), Web of Science (Clarivate Analytics), EBSCO (EBSCOhost), Scopus (Elsevier), Embase (Elsevier), the Cochrane Library (Wiley Online Library), and ProQuest (ProQuest Platform), with Google Scholar serving as a supplementary search source. All databases were searched independently on their respective platforms, with no simultaneous searches conducted on a single platform.

To identify potentially overlooked studies, the research team conducted manual supplementary searches. This process involved reviewing the reference lists of included studies and examining references from relevant systematic reviews and meta-analyses. No additional studies or data were obtained from authors, experts, or other individuals. The review comprised 2 rounds of searching: the initial search was completed on June 10, 2025, for primary study identification, and a second, updated search was completed on August 3, 2025, to capture newly published studies. Both searches employed consistent search strategies.

Search Strategy

The literature search strategy was constructed in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Search Extension) guidelines to ensure transparency and reproducibility [14]. The search strategy was organized around 3 core concept categories. First, physical activity behaviors: "physical activity," "exercise," "physical fitness," "sports participation," and "fitness activity." Second, DHIs: "digital health," "mobile health," "wearable device," "electronic health," and "fitness tracking device." Third, study

XSL•FO
RenderX

participants: “adolescent” and “teenager.” Search terms within the same concept category were combined using the Boolean operator “OR,” while different concepts were linked with “AND.” Search fields and expressions were adjusted according to database-specific syntax and subject heading systems (eg, Medical Subject Headings [MeSH]).

The search time frame spanned from January 1, 2014, to June 30, 2025, and was limited to English-language original empirical research. This systematic review did not utilize published standardized search filters; the retrieval strategy was specifically designed for this study and was not adapted or reused from previous systematic reviews or meta-analyses. The complete search terms and database-specific parameters are detailed in Multimedia Appendix 1. Beyond these methods, this systematic review did not search study registries, conduct purposeful browsing of conference proceedings, government websites, or other online resources; employ additional complementary search methods; or undergo a dedicated peer review of the search strategy.

Selection Process

All search results were first imported into the EndNote reference management software (Clarivate Plc) for preliminary deduplication. Subsequently, the Rayyan online reference management platform was used for literature screening. The screening process was conducted independently by 2 researchers (RSF and JJJ) and comprised 2 stages: title and abstract screening, followed by full-text screening. Disagreements during screening were resolved through discussion and negotiation between the 2 researchers. If consensus could not be reached, a third researcher (XYZ) intervened to make the final decision, ensuring the objectivity and consistency of the screening process.

Data Collection Process

After study inclusion, data were collected using preestablished standardized data extraction forms. Two researchers (RSF and JJJ) independently extracted data on study characteristics, including basic study information, methodological features, forms of DHIs, and intervention outcomes. When critical information was missing or unclear, the researchers contacted the corresponding author to obtain supplementary data. Discrepancies identified during data extraction were resolved through discussion and consensus, and unresolved disagreements were referred to a third researcher (XYZ) for deliberation and final adjudication.

Study Outcomes (Data Items)

Data extraction was performed independently by 2 researchers using predefined standardized data extraction forms. Extracted data included primary and secondary outcomes. Primary outcomes comprised basic study information (eg, authors, study population characteristics, sample size), study design and methodological features, and specific forms of DHIs along with their effects on adolescent physical activity–related outcomes, summarized in conjunction with the reported outcome measures and overall intervention results.

Secondary outcomes included supplementary information supporting outcome classification and interpretation, such as geographic region, intervention duration, and follow-up or sample attrition rates. Disagreements during the process were resolved by a third researcher. Complete data extraction forms and the classification framework are presented in Tables 1 and 2.

XSL•FO
RenderX

Table 1. Overview of included studies (N=24).

Study	Age, mean (SD); grade; gender; condition; population	Region	Study design	Intervention duration	Digital health interventions	Effects of digital health interventions
Soltero et al [15]	• 14.9 (0.91); 53% female; obese/overweight	Spain	Quantitative research	12 weeks	Wearable device (Fitbit Watch) + SMS text messaging	High engagement in self-monitoring behavior and perceived increase in activity
Goodyear et al [16]	• 13-14; 53% female; general adolescents	The United Kingdom	Quantitative research	8 weeks	Wearable device (Fitbit) + virtual platform (with activity goal setting)	Encouraged youth to engage in more physical exercise
Van Dyck et al [17]	• 12-14; 49% female; general adolescents	Belgium	Quantitative research	1 week	SMS text messaging/chatbot + social media (Facebook) + SMS	Facebook and SMS text messaging were considered promising methods for physical activity interventions among vocational school adolescents
Dinç et al [18]	• 14.27 (0.44); 56% female; general adolescents	Turkey	Cross-sectional study	N/A[a]	mHealth[b]: smartphone app	Effectively improves physical activity levels and health awareness
Mojica et al [19]	• 11-14; 100% female; general adolescents	The United States	Cross-sectional study	N/A	Social media: cell phones, computers, game consoles, and internet	Increased physical activity and more frequent participation in daily physical education classes after the intervention
Ng et al [20]	• 11-15; 42% female; general adolescents	Finland	Cross-sectional study	N/A	Wearable device: physical activity tracker (smartwatches and heart rate monitors)	Positive correlation between physical activity behavior and use of wearables/apps; further research needed to verify relationship
Mendoza et al [21]	• • Intervention: 16.9 (1.5); 41.4% female; cancer survivors • Control: 16.3 (SD 1.5); 40% female; cancer survivors	The United States	Randomized controlled trial	10 weeks	mHealth: Fitbit + Facebook	Demonstrates good feasibility and acceptability
Larsen et al [22]	• 14.7 (2.1); insufficient physical activity	The United States	Randomized controlled trial	12 weeks	Virtual platform: intervention website	Shows good feasibility and acceptability, with a significant increase in self-reported physical activity levels
Thompson et al [23]	• 14-17; 51.88% female; general adolescents (group: control, pedometer only, pedometer + goal prompts, and pedometer + goal prompts + self-determination theory texts)	The United States	Randomized controlled trial	12 weeks	mHealth: smartphone (SMS text messaging) + pedometer	Moderate increase in average daily steps and moderate-to-vigorous physical activity
Guthrie et al [24]	• 12-14; 55% female; general adolescents	The United States	Randomized controlled trial	6 weeks	Virtual platform: online intervention system (Zamzee)	Positive impact on moderate-to-vigorous physical activity levels
Direito et al [25]	• 14-17; general adolescents	New Zealand	Randomized controlled trial	8 weeks	mHealth: smartphone (immersive application)	Demonstrated feasibility
Chen et al [26]	• 14.9 (1.7); 42% female; obesity or overweight	The United States	Randomized controlled trial	6 months	mHealth: + wearable device + virtual platform: SMS text messaging, Fitbit Flex, and online educational program	Increased number of days per week engaged in physical activity

XSL•FO
RenderX

Study	Age, mean (SD); grade; gender; condition; population	Region	Study design	Intervention duration	Digital health interventions	Effects of digital health interventions
Caillaud et al [27]	• Intervention: 10.9 (0.7); 51% female; general adolescents • Control: 10.4 (0.5); 62% female; general adolescents	Australia	Randomized controlled trial	5 weeks	Virtual platform: app (iEngage Program)	Improved physical activity goals, academic performance, and moderate-to-vigorous physical activity levels
Egilsson et al [28]	• 15.6 (0.26); 41% female; general adolescents	Iceland	Randomized controlled trial	6 weeks	mHealth: smartphone (Mobile Health Program)	Feasible and usable
Staiano et al [29]	• 11.2 (0.8); 46% female; obesity or overweight	The United States	Randomized controlled trial	24 weeks	Game: Squad Intervention (gaming console)	Improved physical activity levels
Stasinaki et al [30]	• PathMate2: 12.6 (range 11.4-16.9); 38.9% female; obesity or overweight • Control: 13.7 (range 10.9-16.8); 46.2% female; obesity or overweight	Switzerland	Randomized controlled trial	5.5 months	mHealth: smartphone (PathMate2)	Significant and sustained improvements in physical capacity and body composition
Ortega and Cushing [31]	• 13-18; 45% female; general adolescents	The United States	Mixed methods study	27 days	SMS text messaging/chatbot: text bot (temporally augmented goal setting)	Increased participation, with observable changes in physical activity before and after the intervention
Willinger et al [32]	• 12.6 (1.7);27% female; general adolescents	Germany	Mixed methods study	4 months	mHealth: smartphone (the KIJANI app)	Helps oneself and others to be more active in daily life
Schoenfelder et al [33]	• 15.5 (1.4); 54% female; with attention-deficit/hyperactivity disorder	The United States	Mixed methods study	4 weeks	mHealth + Fitbit Flex + Facebook	Average weekly step count increased; intervention shows promise for promoting physical activity in adolescents with attention-deficit/hyperactivity disorder
Koorts et al [34]	• 13.7 (0.4); 49%female; general adolescents	Australia	Mixed methods study	12 weeks	mHealth + wearable device (wrist-worn Fitbit)	Perceived short-term positive effect on exercise motivation
Glaser et al [35]	• Ninth to eleventh grade; 28% female; at-risk youth	Israel	Quasi-experimental study	8 months	Virtual platform: Friends Online Intervention Program—Zoom/video chat with online physical activities and dialogue	The program effectively increased physical activity among adolescents and reduced risky behaviors
Garde et al [36]	• 8-13; general adolescents	Canada	Quasi-experimental study	2 weeks	mHealth: smartphone game (Mobile Kids Monster Manor)	After the intervention, more physical activity was observed; the game's role in promoting activity needs further validation
Mateo-Orcajada et al [37]	• 13.96 (1.21); 48% female; general adolescents	Spain	Quasi-experimental study	10 weeks	mHealth: mobile app	Postintervention improvements in physical activity level, body composition, and physical fitness quality
Cushing et al [38]	• 13-18; 75% female; general adolescents	The United States	Quasi-experimental study	3 weeks	SMS/chatbot: tailored SMS text messaging intervention	Increased physical activity and reduced sedentary behavior

[a]N/A: not applicable.

[b]mHealth: mobile health.

XSL•FO
RenderX

Table 2. Effectiveness of DHIs[a] to promote PA[b] in adolescents: a classification and ranking by intervention type.

Study	Measurement	Key findings	Result	Effect
Single-driver intervention				
Dinç et al [18]	Questionnaire	• PA-themed (40.5%) and healthy nutrition–themed (33.3%) apps were the most frequently used. Adolescents who preferred PA apps had significantly higher healthy lifestyle belief scores (mean 61.82, SD 0.70) than those who preferred Ministry of Health apps (mean 43.54, SD 2.91; P<.001). Similarly, users of self-monitoring health apps scored higher (mean 68.20, SD 0.28) than those using sleep-monitoring apps (mean 55.35, SD 1.24; P<.001).	Intrinsic motivation increased	I[c]
Mojica et al [19]	App data	• Girls who owned a mobile phone were more likely than those who did not to engage in PA for more than 5 days in the past week (odds ratio 5.5, 95% CI 2.1-14) and to participate in daily physical education classes (odds ratio 2.6, 95% CI 1.1-5.9).	Enhanced PA	D[d]
Mendoza et al [21]	Measurable PA + questionnaire	• Exploratory analyses found no significant differences in MVPAe (4.4 vs 5.0 minutes/day; P=.92) or sedentary time (–4.5 vs 1.0 minutes/day; P=.73) between the intervention and control groups. Mean differences were reported for these outcomes, with P values indicating no significant effects. Some modest differences were observed for select subscales of quality of life and motivation for PA.	Intervention feasible/acceptable	I
Larsen et al [22]	Questionnaire + semistructured interview + accelerometer measurement	• According to physical activity recall measurements, participants' MVPA time increased to an average of 79.4 (SD 46.8) minutes at follow-up, representing a per capita increase of 58.8 (SD 46.3) minutes. This difference was statistically significant (P<.001). Overall, PA levels showed a significant improvement from baseline to 12 weeks (P<.001).	Enhanced PA	D
Thompson et al [23]	Measurable PA + questionnaire	• Program satisfaction: The mean score was 17.47/20, with no significant differences between groups (P=.068). The feasibility criterion—a program satisfaction score of ≥15 out of 20 points—was met. • Internal consistency: All measures exceeded the feasibility criterion (α≥.70), with the lowest α being .74 for social desirability. • PA trends: The all-intervention group showed the greatest increases in steps (+317.8 steps/day) and MVPA (+1.73 minutes/day) compared with other groups. However, no significant group-by-time interactions were observed (P>.05).	Enhanced PA	D
Guthrie et al [24]	Zamzee data	• Participants receiving the Zamzee intervention exhibited average MVPA levels 54% higher than the passive control group (P<.0001) and 68% higher than the active control group using commercially available active video games (P<.0001).	Enhanced PA compared with the control group	D
Direito et al [25]	Self-report + accelerometer measurement	• N/Af	The intervention will further validate its effectiveness.	I
Caillaud et al [27]	Accelerometer measurement + Physical Activity Questionnaire for Children + fitness test + iEngage app recording	• Participants increased their daily step count by 30% (+2647 steps/day, P<.001), and the proportion of days exceeding 11,000 steps rose to 48%.	Enhanced PA	D
Egilsson et al [28]	Scale + app data	• Self-reported daily PA increased by 19.61% in the intervention group but decreased by 26.21% in the control group. Self-efficacy levels increased by 8.23% in the intervention arm, compared with a 3.03% decrease in the control group.	Enhanced PA	D

XSL•FO
RenderX

Study	Measurement	Key findings	Result	Effect
Ortega and Cushing [31]	Accelerometer + questionnaire	• The temporally augmented goal setting intervention was generally feasible, acceptable, and technically and functionally reliable. Adolescents demonstrated adequate levels of engagement. Changes in MVPA from pre- to postintervention were small, approximately a 2-minute increase.	Enhanced PA	D
Willinger et al [32]	Semistructured interviews + questionnaires	• Most participants (n=16) believed that using KIJANI would increase their PA levels.	Enhanced PA	D
Garde et al [36]	PA monitor	• During the intervention week, both groups showed significant increases in PA levels (1191 and 796 steps/day, respectively; P=.01). • In the game group, children with higher BMI z scores exhibited greater PA levels, with an average increase of 964 steps/day for each unit increase in BMI z score (P=.03; 95% CI 98-1829).	Further validation is required.	U[g]
Mateo-Orcajada et al [37]	Questionnaire + physical fitness test + anthropometric measurements	• Compared with the control group, adolescents in the experimental group demonstrated higher levels of PA, along with improvements in body composition and physical fitness following the intervention.	Enhanced PA compared with the control group and improvement in health indicators	D
Cushing et al [38]	Questionnaire + accelerometer	• Compared with the attention control group, the intervention group spent an average of 20.84 minutes more per day in MVPA (β=20.84, SE 8.19).	Enhanced PA compared with the control group	D
Multimodal integrated intervention				

XSL•FO
RenderX

Study	Measurement	Key findings	Result	Effect
Soltero et al [15]	Interview	• Young adults reported increases in both PA and self-efficacy for PA.	Intervention feasible and acceptable	I
Goodyear et al [16]	Interview	• Daily goals of 10,000 steps and calorie burn motivate young people to engage in more PA.	Exercise goals encourage PA	I
Van Dyck et al [17]	Questionnaire + interview	• Participants indicated that SMS text messages provide a simple way to receive information about PA, while Facebook group pages are ideal for sharing information with others.	The intervention shows promise as a method for increasing PA	I
Ng et al [20]	Scale	• Compared with nonusers, app users had higher odds of engaging in daily MVPA (males: ORh 1.27, 95% CI 1.04-1.55; females: OR 1.49, 95% CI 1.20-1.85) and of being members of sports clubs (males: OR 1.37, 95% CI 1.15-1.62; females: OR 1.25, 95% CI 1.07-1.50). These associations were even stronger among wearable device users, with MVPA ORs up to 2.25. Male wearable users also had higher odds of engaging in active travel (OR 1.39, 95% CI 1.04-1.86).	Compared with the control group, the intervention showed a positive correlation with PA.	I
Chen et al [26]	Anthropometric measurements + questionnaire + Fitbit Flex data + interview	• Compared with participants in the control group, those in the mobile phone–based intervention showed significant improvements in BMI (z=–4.37, P=.001), diastolic blood pressure (z=–3.23, P=.001), PA days/week (z=2.58, P=.01), TV and computer time (z=–3.34, P=.001), servings of fruits and vegetables per day (z=2.74, P=.006), servings of soda and sweetened drinks (z=–3.19, P=.001), PA self-efficacy (z=2.75, P=.006), and dietary self-efficacy (z=5.05, P=.001). Medium to large effect sizes were observed for these outcomes.	Enhanced PA and increased self-efficacy.	D
Schoenfelder et al [33]	Quantitative measures + Fitbit Flex data + questionnaires + interview	• Participants significantly increased their average weekly step count and showed improvements in attention-deficit/hyperactivity disorder inattention symptoms.	Enhanced PA	D
Koorts et al [34]	Questionnaire + interview	• Adolescents perceived the intervention content as easy to understand (100/120, 83.3%) and the Fitbit as easy to use (112/120, 93.3%). Half of the adolescents found the SMS text messages useful (61/120, 50.8%), while 47.5% (57/120) liked the weekly challenges, and 38.3% (46/120) liked the Facebook videos. • Teenagers reported that the Fitbit Flex increased their awareness of PA and their likelihood of being more active in the short term.	Enhanced PA compared with the control group.	I
Interaction-enhanced intervention				

XSL•FO
RenderX

Study	Measurement	Key findings	Result	Effect
Staiano et al [29]	Anthropometry + full-body scanner + accelerometer + questionnaire	• The intervention group showed a significant reduction in BMI z score, with a marginal difference in the intent-to-treat analysis (mean −0.06, SE 0.03 for the intervention group vs mean 0.02, SE 0.03 for the control group, P=.065). Compared with the control group, the intervention group also showed improvements in systolic blood pressure, diastolic blood pressure, total cholesterol, low-density lipoprotein-cholesterol, and MVPA (all P<.05).	Enhanced PA and improvement in health indicators	D
Stasinaki et al [30]	Physical fitness test (Dordel-Koch test) + questionnaire + application data	• Both groups showed significant improvements in muscle mass, strength, and agility at T2, whereas only the PathMate2 group demonstrated a significant reduction in body fat at both T1 and T2. The average daily PathMate2 app usage rate was 71.5%.	Enhanced PA and improvement in health indicators	D
Glaser et al [35]	Questionnaire	• Compared with the control group, the intervention group experienced a 57% increase in PA.	Enhanced PA	D

[a]DHI: digital health intervention.

[b]PA: physical activity.

[c]DHIs are indirectly effective in promoting PA among adolescents (indirectly effective).

[d]DHIs are directly effective in promoting PA among adolescents (directly effective).

[e]MVPA: moderate-to-vigorous physical activity.

[f]N/A: not applicable.

[g]U: The effectiveness of DHIs in promoting PA among adolescents remains unclear (unclear).

[h]OR: odds ratio.

Study Risk-of-Bias Assessment

To ensure methodological rigor and assess internal validity, structured risk-of-bias assessments were conducted using JBI critical appraisal tools, each aligned with the corresponding study design. This included assessment checklists applicable to CSSs, RCTs, QESs, QR, and MMSs [13].

Risk of bias was assessed independently by 2 researchers using a 3-color coding system: green for “met,” yellow for “partially met,” and red for “not met.” Based on these assessments, each study’s methodological quality was categorized as high, moderate, or low. Disagreements were resolved through discussion and consensus, and if consensus was not reached, a third researcher adjudicated. Studies with reliability concerns—such as bias, unclear criteria, incomplete reporting, inappropriate methods, or insufficient justification—were flagged for further discussion and potential exclusion. No studies were excluded solely based on quality scores. Risk of bias was incorporated into the narrative synthesis to contextualize the strengths and limitations of the evidence. Detailed assessment results and scoring criteria are presented in Figures 1-4.

XSL•FO
RenderX

Figure 1. Quality assessment tool: cross-sectional studies. JBI: Joanna Briggs Institute.

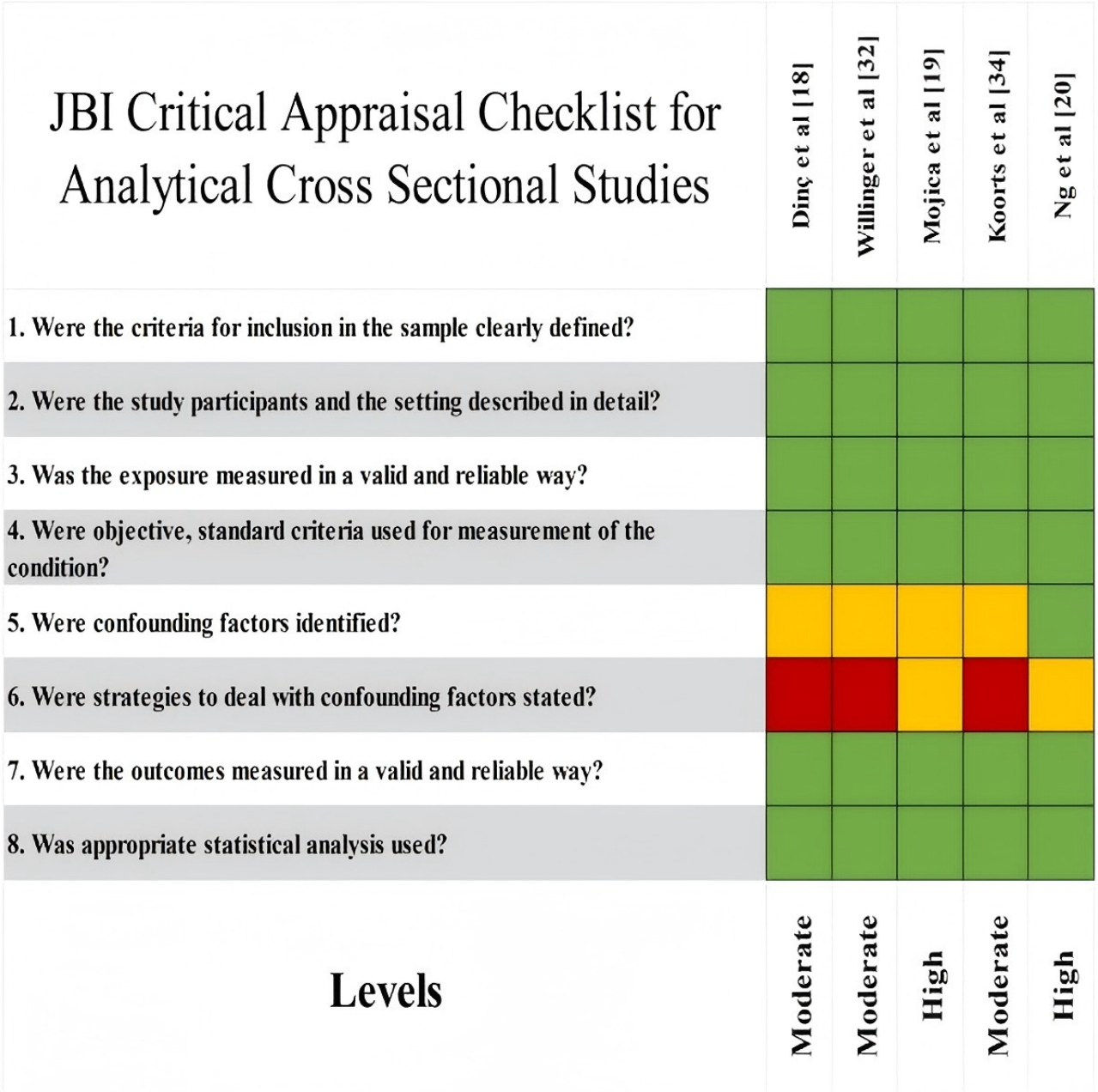

XSL•FO
RenderX

Figure 2. Quality assessment tool: qualitative research. JBI: Joanna Briggs Institute.

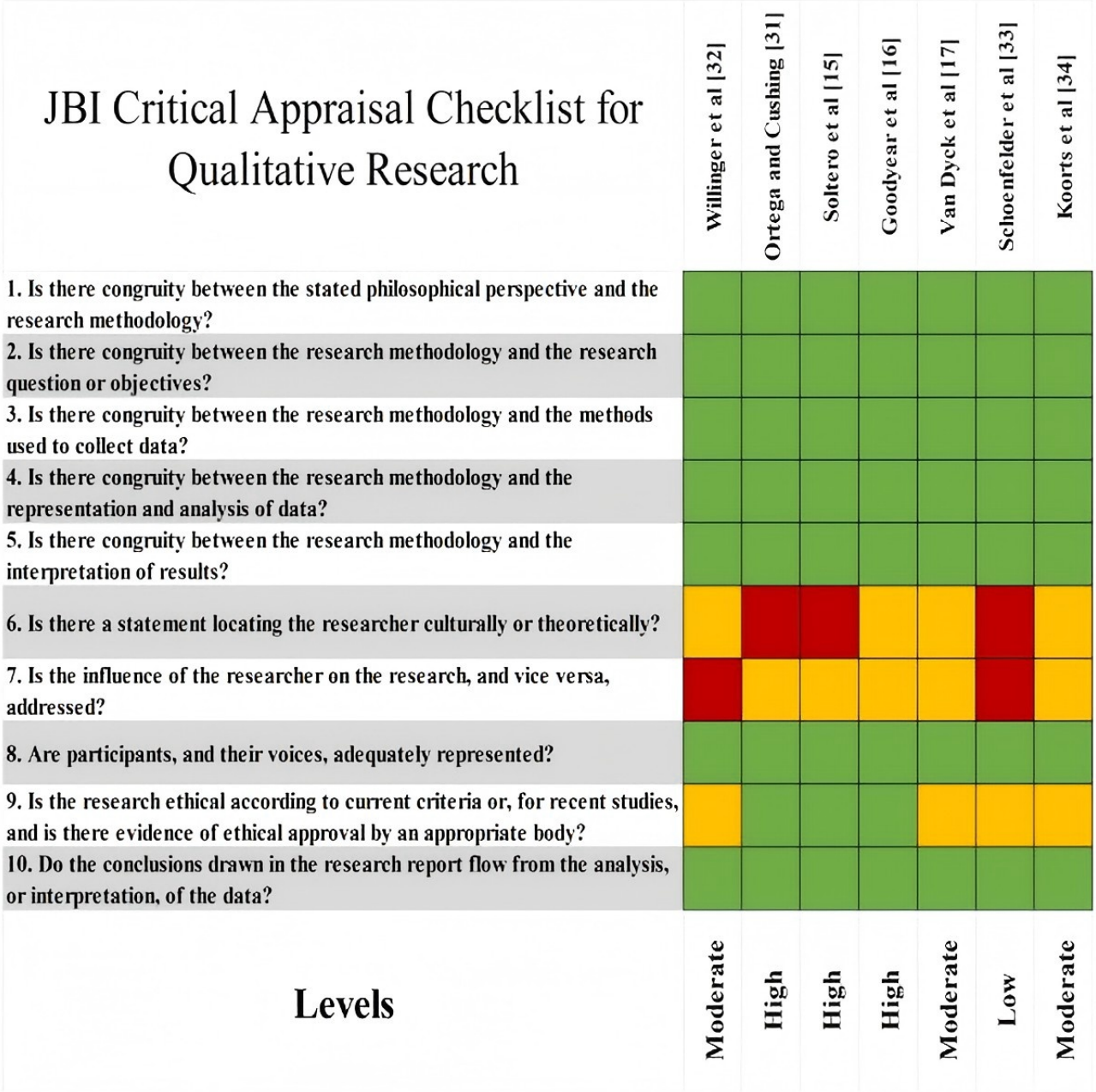

XSL•FO
RenderX

Figure 3. Quality assessment tool: randomized controlled trials. JBI: Joanna Briggs Institute.

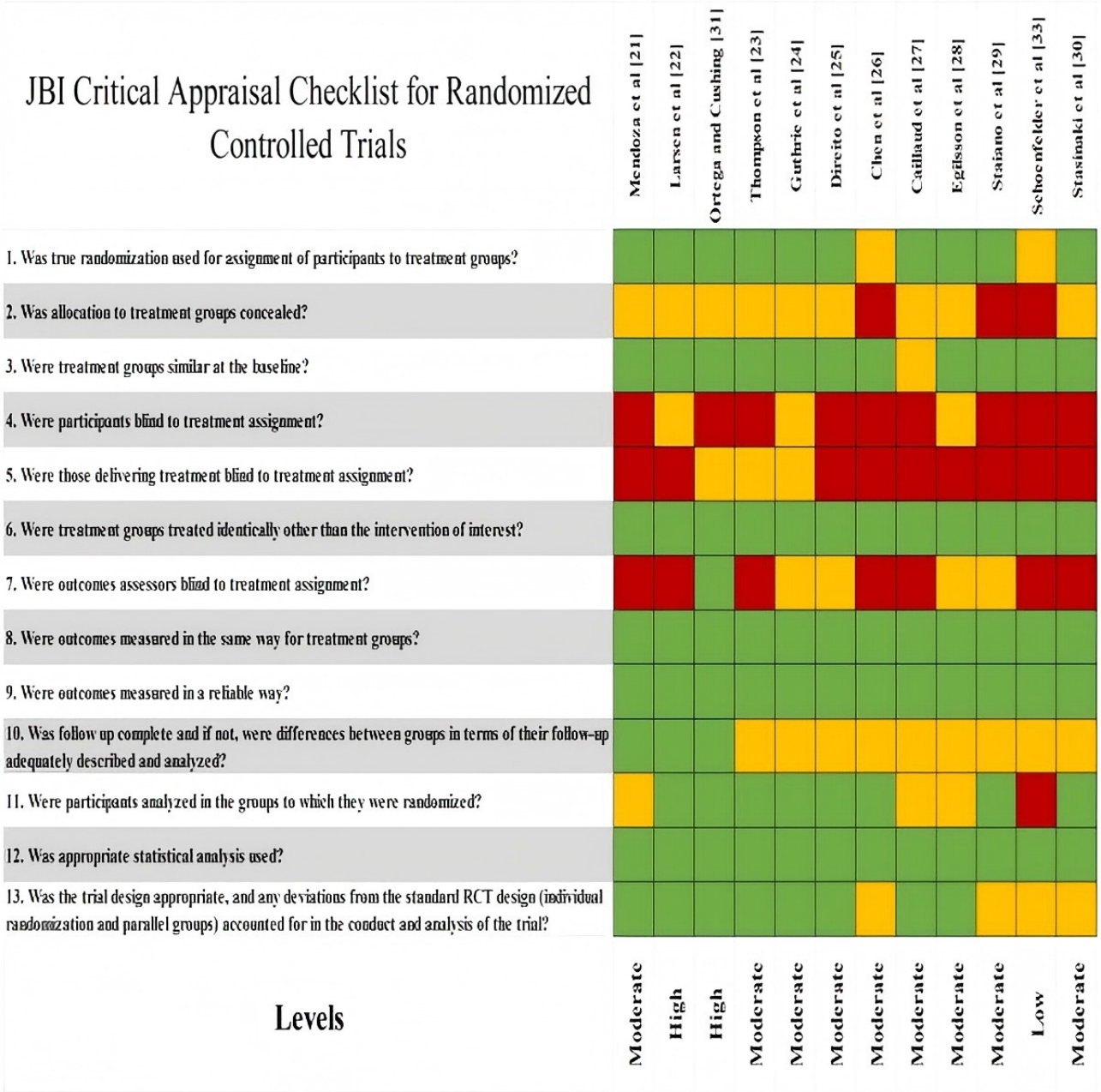

XSL•FO
RenderX

Figure 4. Quality assessment tool: quasi-experimental studies. JBI: Joanna Briggs Institute.

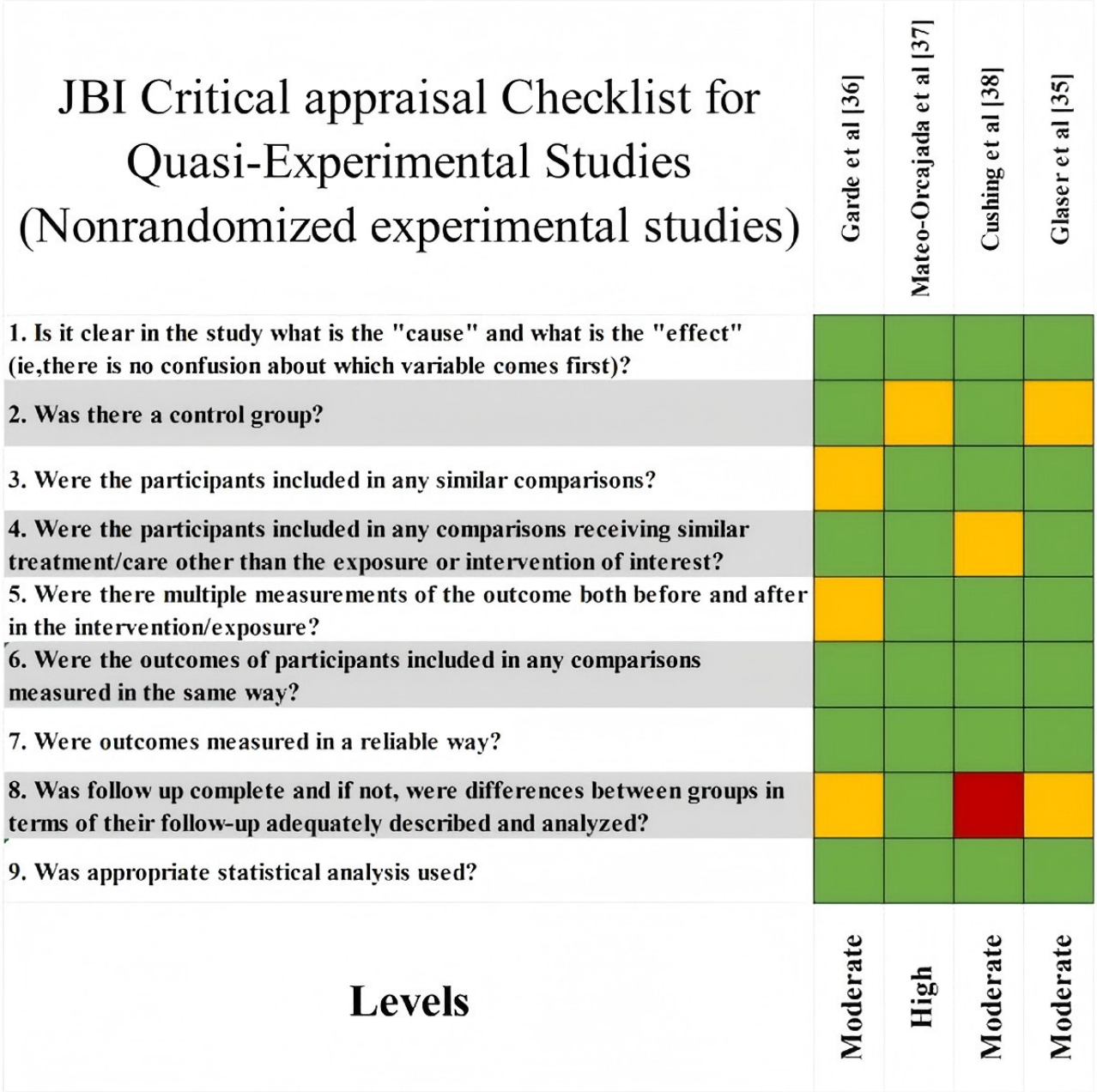

Effect Measures

Effect measures reported in the studies were diverse, encompassing both quantitative metrics and qualitative outcomes. Quantitative studies primarily presented intervention effects through comparisons of mean values or change values (eg, physical activity duration, step count, and their changes before and after intervention or between intervention and control groups), odds ratios, and regression model estimates, supplemented by statistical indicators such as CIs and *P* values to support conclusions. Some studies reported only the direction of change or statistical significance of outcomes without providing uniformly convertible effect sizes.

Additionally, some studies reported intervention effects through qualitative descriptions (eg, user experience, satisfaction, or perceived changes). Given the inconsistencies in effect measure formats and statistical reporting methods, this review did not uniformly convert or pool effect sizes. Instead, it focused on comprehensively describing the directional trends and overall performance of intervention effects.

Synthesis Methods

Given the substantial differences among included studies in research design, intervention modalities, outcome measurement, and statistical analysis methods, and considering that many studies did not report effect size parameters suitable for uniform conversion or pooling, this systematic review did not conduct a quantitative meta-analysis to avoid producing statistically unstable or potentially misleading pooled results. Prerequisites for meta-analysis include comparability or reasonable convertibility of outcome measurements and effect size definitions across studies. However, the included studies employed both self-reported and objectively measured outcomes,

XSL•FO
RenderX

with inconsistent measurement time points, making these conditions unattainable.

The research team therefore employed a narrative synthesis and qualitative content analysis approach to integrate the findings. During synthesis, a classification framework based on core forms of DHIs and their effect characteristics was developed to determine study applicability across synthesis units and to facilitate group comparisons. This classification and synthesis were performed independently by 2 researchers, with disagreements resolved through discussion and, when necessary, adjudicated by a third researcher. Findings were primarily presented using structured tables and typological narratives (Tables 1 and 2), with interpretations of outcome variations grounded in intervention formats, study designs, and implementation characteristics. As no statistical synthesis was conducted, this review did not perform formal heterogeneity testing or sensitivity analyses.

Reporting Bias Assessment

This systematic review employed a narrative synthesis and did not perform statistical tests for publication bias. Given the absence of a quantitative meta-analysis and the substantial heterogeneity in study design and outcome reporting across included studies, methods such as funnel plots were considered inapplicable. During evidence synthesis and result interpretation, the research team conducted a qualitative assessment of potential reporting bias. By comparing the consistency between study objectives, methods, and reported outcomes, and by incorporating study registration information (where available) and author explanations, the team cautiously discussed the potential impact of missing results on study conclusions.

Certainty Assessment

This systematic review conducted a qualitative synthesis of the overall credibility of the evidence, considering the design types of included studies, the results of the JBI methodological quality assessment, and the consistency of study conclusions across different studies. Given that no quantitative meta-analysis was performed and the included studies exhibited substantial heterogeneity in design types and outcome measures, quantitative evidence grading tools such as GRADE (Grading of Recommendations, Assessment, Development and Evaluation) were not applied. The assessment of evidence certainty was conducted independently by 2 researchers (RSF and JJJ). The reliability and limitations of the evidence were interpreted in the “Results” and “Discussion” sections in conjunction with specific study findings.

Ethical Approval

Ethics approval was not required for this systematic review.

Results

Study Selection

A total of 2127 English-language articles were initially retrieved from 7 major databases, including Web of Science (n=601), PubMed (n=518), EBSCO (n=50), Scopus (n=305), Embase (n=164), Cochrane Library (n=128), ProQuest (n=78), and Google Scholar (n=283). After initial deduplication using EndNote, 1288 duplicate records were removed. Subsequently, an additional 164 duplicates were removed using the Rayyan online literature management platform, resulting in 675 records retained for title and abstract screening. Following this initial screening, 213 articles remained for full-text assessment. Of these, 189 articles were excluded based on predefined exclusion criteria, including study population, intervention characteristics, and research design. The specific reasons for exclusion were as follows: studies not focused on adolescents (n=36), absence of DHIs (n=14), insufficient clarity in intervention or outcome measures (n=21), conference abstracts (n=8), studies focused solely on instrument validation (n=49), review articles (n=35), and studies with unrelated topics (n=26). Ultimately, 24 English-language studies met the inclusion criteria and were included in the final qualitative synthesis of this systematic review (see Figure 5 and Multimedia Appendix 2).

XSL•FO
RenderX

Figure 5. PRISMA Flow Diagram of the Study Selection Process.

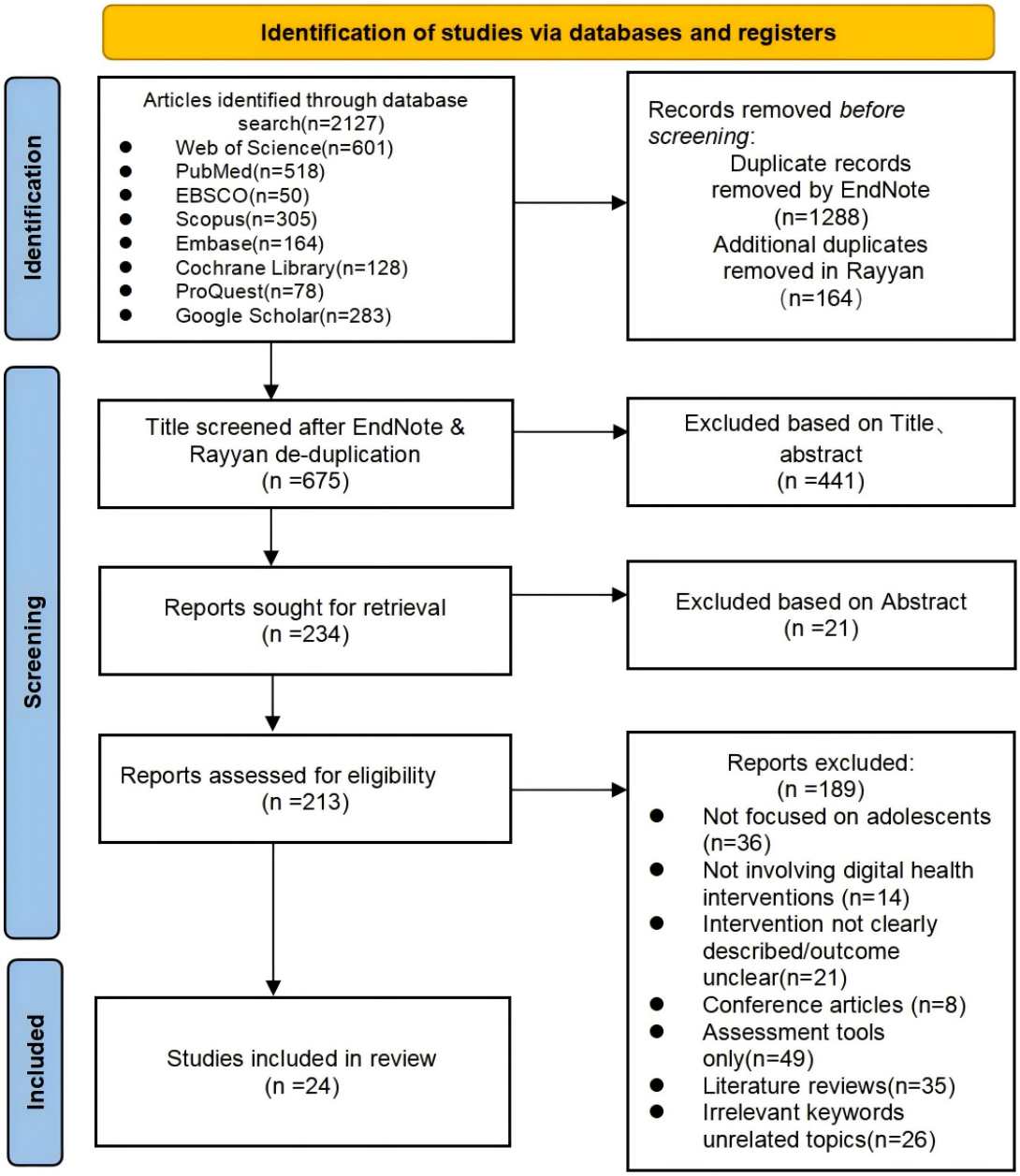

Study Characteristics

Overview

This systematic review included 24 peer-reviewed articles published between 2014 and 2025, spanning multiple countries and regions, thereby reflecting the global uptake of DHIs among adolescents. The included studies employed a range of methodological designs, including RCTs, MMS, QR, and other empirical approaches. All interventions targeted adolescents aged 10-19 years, including the general population, individuals with insufficient physical activity, and those with specific health conditions. Sample sizes ranged from 10 to 9940 participants, with the majority of studies involving fewer than 500 individuals. Intervention duration ranged from 2 weeks to 8 months. The included studies employed diverse intervention formats, utilizing either single-modality or multicomponent digital technology approaches. Detailed study characteristics are summarized in Table 1.

Participant Characteristics

This systematic review encompassed approximately 12,183 adolescent participants. Sample sizes varied widely, ranging from 10 to 9,940 participants. The overall mean sample size was approximately 507, whereas the median was substantially lower at 46. Most studies employed small-to-medium sample sizes, with a few large-scale studies exerting a disproportionate influence on the overall mean. Among studies that reported age data, participants' age ranged from 8 to 19 years, with the highest concentration in the 13-15 age group, representing over one-third of the total sample. The study populations included both general adolescent cohorts and specific subgroups characterized by distinct health risks or behavioral profiles. Of

XSL•FO
RenderX

the total 24 studies, 8 (33%) explicitly targeted specific subgroups, including 4 (17%) on adolescents with obesity or overweight, 1 (4%) on adolescents with insufficient physical activity, 1 (4%) on individuals with attention-deficit/hyperactivity disorder, 1 (4%) on cancer survivors, and 1 (4%) on at-risk adolescents. Additional details are provided in Multimedia Appendix 3.

Geographic Distribution

The 24 included studies were conducted across multiple countries and regions. North America accounted for the highest concentration, contributing 11 (46%) studies of the total, significantly exceeding other regions. Europe followed with 8 (33%) studies, while the Middle East and Oceania contributed 2 (8%) and 3 (13%) studies, respectively. At the country level, the United States emerged as the leading research site, accounting for 10 (42%) studies—the highest among all countries. Spain and Australia each contributed 2 studies. Other countries—namely, Switzerland, Turkey, Germany, Israel, Belgium, New Zealand, Iceland, Canada, the United Kingdom, and Finland—each contributed 1 study. Regional variations were also observed in the selection and application of intervention technologies. North American countries (eg, the United States and Canada) generally favored the combination of wearable devices and mobile apps, providing integrated data tracking and personalized feedback. By contrast, several European countries primarily integrated traditional intervention models with digital platforms, including web-based systems and educational software. Although studies from Middle Eastern and Oceanic countries featured smaller sample sizes, their intervention formats were diverse, incorporating tools such as SMS text messaging and interactive platforms (see Multimedia Appendix 3).

Intervention Duration and Frequency

Considerable variation was observed in the duration of interventions. The shortest intervention lasted 1 week (excluding 3 CSSs), while the longest extended to 8 months. The most commonly adopted duration was 12 weeks (3 months), reported in 4 (17%) studies, followed by durations of 6, 8, and 10 weeks, each reported in 2 (8%) studies. Other durations included 1 week, 2 weeks, and 8 months, each reported in a single study (4%). Overall, approximately one-third of the studies fell within the 6- to 12-week range. Based on duration, short-term interventions (≤4 weeks) were reported in 5 (21%) studies; medium-term interventions (5-12 weeks) were the most prevalent, appearing in 11 (46%) studies; and long-term interventions (>12 weeks) were implemented in 5 (21%) studies. Long-term interventions were predominantly associated with high-quality RCTs, whereas short-term interventions were distributed across a variety of study designs. In summary, the majority of studies employed medium-term interventions, typically lasting 2-3 months.

Risk of Bias in Studies

Among the 24 included studies, research designs comprised 10 (42%) RCTs, 4 (17%) QES, 3 (13%) QR, 3 (13%) CSS, and 4 (17%) MMS. Quality appraisal results indicated that 7 (29%) studies were rated as high quality, 16 (67%) as moderate quality, and 1 (4%) as low quality. Further analysis revealed that most RCTs achieved high or moderate quality ratings, while QES, CSS, and MMS studies generally met moderate or higher quality standards. Notably, QR studies demonstrated outstanding performance, with 2 projects achieving high quality (see Figure 2).

Classification and Outcomes of DHIs

The classification and corresponding outcomes of DHIs are presented in Table 2. The types of interventions employed across the included studies exhibited considerable diversity.

These interventions were broadly categorized into 3 types: single-driver interventions, multimodal integrated interventions, and interaction-enhanced interventions. Based on the combined findings from Tables 1 and 2, single-driver interventions were the most commonly reported, appearing in 14 (58%) studies. Interventions in this category typically relied on a single technological medium, most commonly smartphones as the primary device. mHealth interventions were the most prevalent, followed by virtual platform–based interventions and robotic systems delivering personalized SMS text messages. Multimodal integrated interventions were identified in 7 (29%) studies. These interventions combined 2 or more digital components, including mHealth apps, wearable devices, and web-based platforms. Interaction-enhanced interventions were relatively limited, reported in only 3 (13%) studies. Characterized by active adolescent engagement—such as promoting physical activity through video chats—these interventions primarily involved gamified strategies, social media platforms, and SMS text message–based messaging systems.

Overall, all 3 categories of DHIs demonstrated positive outcomes. As shown in Table 2, the majority of studies were classified as “D” (directly effective; 15/24, 63%), while smaller proportions were categorized as “I” (indirectly effective; 8/24, 33%) or “U” (unclear; 1/24, 4%).

Among the “D” interventions, most studies reported statistically significant improvements in physical activity outcomes [19,22-24,26-33,35,37,38], including increased step counts, longer durations of moderate-to-vigorous physical activity, and greater participation in sports. Several studies also reported improvements in physical health indicators related to physical activity [15-18,20,21,25,34], including reductions in BMI, improved body composition and physical fitness, and decreased body fat percentage. Studies classified as “I” did not report direct increases in physical activity levels but demonstrated positive effects on psychological mechanisms, such as enhanced exercise motivation and self-efficacy. These effects included strengthened beliefs in healthy lifestyles, the establishment of exercise goals that promote physical activity, and increased awareness of the benefits of physical activity [36]. Studies categorized as “U” lacked conclusive evidence, making it difficult to determine the effectiveness of the DHIs under investigation.

Results of Syntheses

A comprehensive analysis of the included studies was conducted based on the form of DHIs and their effectiveness characteristics. The synthesis identified significant variation in

XSL•FO
RenderX

research designs (eg, QR, CSS, RCT, QES, MMS), sample sizes (10-9940 participants), outcome measurement methods, and methodological quality. Overall study quality was predominantly moderate to high, with most interventions lasting short to medium durations. As a result of substantial heterogeneity in study design, outcome definitions, and statistical analysis methods, as well as the absence of standardized effect size parameters for pooling across studies, a quantitative meta-analysis was not performed. Instead, results were synthesized using narrative synthesis and qualitative content analysis based on intervention type. The findings are presented in structured tables and categorized descriptions (Tables 1 and 2) to compare the effectiveness of different intervention formats and study designs in promoting physical activity among adolescents.

Reporting Biases and Certainty of Evidence

As this systematic review did not perform a quantitative meta-analysis and due to considerable variation in study designs and outcome reporting methods, no formal testing for publication bias was conducted using funnel plots or statistical methods. Instead, a qualitative assessment of potential reporting bias was undertaken during evidence synthesis and interpretation. Evidence certainty was not formally assessed using a quantitative tool; rather, the credibility of the evidence was judged by integrating study design type, methodological quality assessment results, and the consistency of study conclusions. Overall, the included studies were of moderate-to-high quality (Figures 1-4). However, owing to study heterogeneity and methodological limitations, the conclusions should be interpreted within the context of specific research settings.

Discussion

Principal Findings

This systematic review synthesizes evidence from diverse DHIs and study designs to evaluate their role in promoting physical activity among adolescents and to explore their adaptability across different research designs and settings. Results from the 24 included studies indicate that DHIs can enhance adolescents' physical activity through direct or indirect mechanisms. Substantial heterogeneity was observed among the studies in terms of research design types, resulting in inconsistent evidence strength and limited generalizability; this heterogeneity represents a key finding of the review. A systematic analysis of intervention formats further revealed that different types of DHIs (single-driver, multimodal integrated, and interaction-enhanced) exhibit distinct effect profiles and are applicable in different contexts. In addition, factors such as target population characteristics and intervention duration significantly influence the interpretation and application of outcomes. Accordingly, this review emphasizes the need for cautious interpretation of research findings, with careful consideration of the scope of inference supported by the study designs and implementation contexts.

This review included 3 QR and 3 CSS. As shown in Tables 1 and 2, the evidence derived from these study types differs. QR evidence reflects the meaning or experience of behaviors or events rather than quantitative measurements [39]. For example, Soltero et al [15] used qualitative methods to assess the feasibility and acceptability of DHIs, identifying potential pathways for promoting physical activity among adolescents with obesity. CSSs measure exposure and outcomes at a single time point, revealing correlations between DHI use, individual characteristics, and physical activity levels [40]. For example, Ng et al [20] examined the association between ownership and use of personal activity trackers and physical activity behaviors. Consequently, this evidence is better suited to suggesting potential mechanisms and real-world usage scenarios rather than to drawing strong causal conclusions.

This review observes the indirect effectiveness of DHIs on adolescent physical activity within QR and CSS evidence. Although such interventions may not immediately lead to significant behavioral changes, these study types more readily capture positive psychological shifts, such as enhanced motivation, increased behavioral intent, strengthened self-efficacy, improved health cognition, or greater interest in participation [41]. Compared with RCTs, QR emphasizes contextual experiences and subjective perceptions, making it more effective in identifying positive psychological shifts, such as heightened motivation. This provides crucial insights into the indirect mechanisms of intervention. CSSs excel at revealing structural relationships between psychological or behavioral characteristics and DHI use at a single time point, offering key clues for understanding these indirect mechanisms.

According to self-determination theory [42], the persistence of adolescent behaviors depends not only on external reminders or incentives but also on the fulfillment of fundamental psychological needs, such as autonomy, competence, and relatedness. Among adolescents in the early stages of behavior change or with unstable motivation, such psychological improvements are particularly important [43] and may gradually translate into behavioral changes through positive experiences, goal identification, and social support [44]. This mechanism of psychological-behavioral conversion helps explain the delayed behavioral improvements observed in some follow-up studies; however, further longitudinal research is needed to validate these findings.

While this evidence provides valuable insights into underlying mechanisms and the overall methodological quality is high, risks of bias remain. Some QR studies did not adequately disclose the researcher's stance [15-17], and some CSSs did not sufficiently control for confounding factors [18,20]. These issues may compromise the robustness of the findings and therefore warrant cautious interpretation.

QR studies and CSSs have distinct characteristics. The former illuminates indirect effects and underlying mechanisms of DHIs, whereas the latter reveals associations among variables but cannot establish causality. Future research should prioritize high-quality study designs to validate the efficacy and long-term effects of DHIs.

This systematic review included 10 RCTs, which provided evidence with varying focuses. RCTs are considered the gold standard for evaluating intervention effectiveness, as their rigorous design allows for systematic assessment of intervention

XSL•FO
RenderX

outcomes over time [45]. The RCTs included in this review involved both general adolescent populations and adolescents with specific health conditions, primarily aiming to validate the feasibility or efficacy of DHIs for these groups [21].

Collectively, these studies indicate that DHIs have potential for promoting physical activity. For populations with specific health conditions, effectiveness often depends more on personalized and sustained support than on technological features [46-48]. Although subgroups share certain commonalities, they also differ in their intervention needs. Commonalities include challenges in behavior change, strong demands for psychological and emotional support, and the need for long-term follow-up and sustained support [10,32,49-53]. Differences are reflected in personalized preferences. Adolescents with attention-deficit/hyperactivity disorder may require more structured and highly interactive tools, whereas cancer survivors may prioritize privacy and self-paced control. Adolescents with obesity or overweight issues may benefit from integrated solutions that combine physical activity, dietary adjustments, and psychological support [54-56]. These differences arise from variations in behavioral mechanisms, cognitive characteristics, and motivational responses across populations. Even with robust technology, failure to align interventions with users' cognition, motivation, and behavior can hinder effective engagement and behavior change. Future research should prioritize methodologically rigorous RCTs and tailored designs that match population characteristics with intervention protocols to enhance evidence interpretability and real-world applicability.

Of the follow-up studies included in this systematic review, 3 were RCTs [22,23,26]. The findings suggest that the presence or absence of follow-up may introduce bias in assessing DHI outcomes. Follow-up is crucial for evaluating the sustainability and long-term impact of interventions, as extended follow-up periods help identify both enduring effects and potential side effects [57]. For example, Larsen et al [22] reported that a 12-week follow-up showed the intervention promoted diverse physical activities. Similarly, Thompson et al [23] found that personalized SMS text message interventions helped adolescents establish long-term exercise habits, with follow-up feedback further optimizing intervention effectiveness.

In this systematic review, studies with follow-up periods allow assessment of the sustainability of intervention effects, whereas studies without follow-up can evaluate only short-term outcomes. Evidence suggests that a 12-week intervention can yield favorable results for adolescents [23], and such feedback helps refine intervention strategies. RCTs provide relatively strong evidence for this review; however, the limited number of studies primarily supports short-term effects and does not adequately validate intervention sustainability or stability. Consequently, robust evidence for the long-term sustainability of DHIs in real-world settings, such as schools and communities, remains insufficient, which may hinder practical implementation and policy support.

The included RCTs indicate that a 3-month (12-week) intervention duration is the most common. This cycle facilitates gradual behavior establishment and consolidation while minimizing negative effects such as fatigue, a finding widely recognized [23]. Research suggests that shorter intervention durations are insufficient to produce significant changes, whereas excessively long durations may increase complexity and reduce adherence [58]. However, no studies specifically examined the impact of varying intervention durations on adolescent populations.

The impact of sample attrition on outcomes cannot be overlooked. Three of the included studies reporting attrition were RCTs [23,28,30]. Our analysis found that attrition typically occurs in studies requiring sustained participation and long-term engagement, making it a critical factor influencing outcomes. Studies with high attrition rates often face challenges such as insufficient intervention personalization, limited feedback, and inadequate technical support, all of which can weaken participants' motivation and commitment.

From a behavioral science perspective, based on the COM-B framework, if interventions fail to sustainably enhance individuals' sense of competence, adolescents are prone to burnout and disengagement. Egilsson et al [28] confirmed this, suggesting that high attrition rates during 6-week follow-ups may be related to insufficient intervention personalization, monotonous formats, or technical barriers. Conversely, other studies found that lower attrition rates are closely associated with adolescent cocreation of intervention content [23]. Thus, enhancing intervention personalization, diversity, and interactivity can reduce attrition, thereby improving research stability and the generalizability of findings.

As shown in Table 2, among the studies demonstrating direct effectiveness, 8 were RCTs. RCTs minimize confounding bias through randomization, which helps clarify intervention effects. Analysis indicates that these interventions promote short-term behavioral changes (eg, increased daily step counts) through goal setting, task execution, and feedback design [59]. This effect depends on clear goals and immediate feedback, enabling adolescents to master behavioral cognition through tracking or outcome comparisons [60]. Such interventions drive strong short-term behavioral change, particularly among participants with initial motivation. The core mechanism enhances individual competence and execution capacity, fosters a sense of accomplishment, and lays the foundation for long-term behavioral strategies.

In summary, RCT designs provided key evidence for this review, although some biases remain. Many studies, due to the nature of DHIs, lacked blinding and faced challenges in follow-up and assessment, which may have affected the reliability of their findings. Future research should employ high-quality RCTs with robust follow-up designs to assess long-term effects and address sample attrition. Given that over half of existing studies used RCTs to evaluate direct effects, future studies should select study designs that align with specific objectives and research contexts.

This systematic review included 4 MMS and 4 QES. MMS combines quantitative and qualitative approaches, enabling simultaneous evaluation of effects and underlying mechanisms, whereas QES assesses impacts but is prone to selection bias due to the absence of randomization or control groups [61]. Given the limited number of MMS and QES, the evidence is

XSL•FO
RenderX

scattered. These study types are suitable for identifying feasible strategies and potential mechanisms in real-world contexts, but should not be overgeneralized into universal conclusions. For example, one study from each design type explored the effectiveness and feasibility of DHIs for promoting physical activity in adolescents with specific health conditions. As shown in Table 2, the majority of studies within both research designs demonstrated the direct effectiveness of DHIs, supporting the applicability of these approaches. However, the overall quality was moderate, likely due to small sample sizes and selection bias. Given the limited number of studies, the evidence lacks representativeness and should be interpreted with caution.

Beyond the evidence provided by different study designs, we identified additional relevant findings. Compared with previous systematic reviews [9-11], this review included multiple types of DHIs and categorized them thematically. Results indicate that most studies employed single-driver interventions, while a minority utilized multimodal integrated or interaction-enhanced interventions. Each intervention type exhibits distinct characteristics, tailored to different target populations and intervention phases (see Multimedia Appendix 3 and Table 2).

First, single-driver interventions, such as smartphone apps, promote behavior change through goal setting and reminders and are particularly effective during the initiation phase, leading to significant increases in daily step counts [10,62]. However, their effects are limited by the lack of continuous feedback, restricting their applicability to short-term behavior promotion [63]. Second, multimodal integrated interventions combine wearable devices with mHealth technologies. These interventions are well-suited for the long-term management of high-risk populations, consistently increasing activity frequency and reducing sedentary behavior [64,65]. Nevertheless, their complexity and user burden may hinder acceptance and adherence [66], highlighting the need for simplified workflows to enhance participation.

Third, interaction-enhanced interventions leverage gamification and social interaction to stimulate adolescents' emotional and social motivation, thereby increasing participation enthusiasm and self-efficacy [67]. Their strength lies in enhancing behavioral appeal and fostering sustained engagement, making them potent drivers of short-term behavioral change. The core mechanism involves strengthening individuals' sense of competence and execution capacity, enabling them to experience accomplishment and receive positive feedback upon goal attainment, thereby laying the foundation for long-term behavioral strategies. However, effectiveness may vary among individuals with weaker social motivation or limited digital access [32,49]. Current research predominantly focuses on single-driver interventions due to their operational simplicity. Nonetheless, distinct characteristics and limitations emerge across intervention types when applied to diverse adolescent groups and developmental stages. Consequently, the selection of intervention type can significantly influence outcomes, particularly for specific populations and developmental phases.

On a broader scale, this systematic review highlights 2 critical real-world challenges in DHIs. First, these interventions reflect regional disparities and health inequalities [68]. North America, particularly the United States, leads in research and technological applications, emphasizing personalized feedback and data tracking [69,70]. In Europe, although fewer studies are conducted, there is a focus on educational and universal approaches tailored to diverse sociocultural contexts [71]. Meanwhile, the Middle East and Oceania explore a variety of technological pathways within the constraints of limited resources [72].

However, these regional disparities not only reflect differences in technological adoption but also highlight digital health inequalities across domains. Digital health disparities are becoming increasingly prominent worldwide, particularly in low- and middle-income countries [73], where significant gaps persist in the accessibility and effectiveness of DHIs. These disparities are largely influenced by factors such as infrastructure development and digital literacy [74]. Therefore, in promoting global DHIs, greater focus must be placed on addressing these regional inequalities. Measures should be taken to allocate resources appropriately and enhance technical support to ensure the inclusivity and equity of technology.

Second, the technological dependency of DHIs is both a crucial factor influencing outcomes and a determinant of long-term adherence. When interventions rely on digital platforms, apps, or wearable devices, participants' sustained engagement may be disrupted by technical issues, thereby affecting intervention stability and the assessment of outcomes [75-77].

Mechanistically, the effectiveness of system design in technological interventions largely determines participants' long-term adherence, as supported by previous studies. For example, Jakob et al [78] found that sustained use of mHealth apps is influenced by design factors such as technical stability and personalized push notifications. Similarly, Kelders et al [79] demonstrated that digital interventions incorporating persuasive design significantly enhance user adherence. Therefore, addressing the negative impacts of technological dependency represents a critical future development pathway for DHIs.

Limitations

This systematic review has several key limitations. First, the included studies exhibited considerable heterogeneity in design type, intervention format, feedback mechanisms, duration, and behavioral assessment methods. This heterogeneity limited evidence consistency and hindered the integration and generalizability of the findings. Second, many original studies had brief evaluation periods and lacked medium- to long-term follow-up data, complicating the assessment of intervention sustainability and long-term impact. Additionally, some studies had methodological limitations and insufficient reporting transparency. Although most studies were of moderate to high quality, these issues compromised the reliability and generalizability of the evidence. Finally, despite conducting multidatabase systematic searches, relevant literature may have been omitted due to limitations in search coverage and potential publication bias.

XSL•FO
RenderX

Therefore, future research can be enhanced in several areas. First, the study design should be strengthened to elevate overall evidence quality. Conducting more high-quality RCTs would improve the scientific rigor and feasibility of intervention studies. Standardization of intervention content and behavioral indicators should be promoted, establishing a unified assessment system to enhance comparability and integration of results across studies. Simultaneously, methodological design and reporting transparency should be improved to provide a more reliable evidence base for systematic reviews.

Second, future research should integrate discussions of intervention mechanisms and effects into study methodologies. The subjective experiences and underlying mechanisms of DHIs hold significant research value. Combining subjective experience data with objective behavioral data can generate more comprehensive and representative evidence. A deeper understanding of the transformation processes underlying these mechanisms will provide a foundation for developing personalized intervention models. Finally, continuous research should address practical issues, such as strengthening follow-up. Study designs should incorporate more follow-up components to systematically document postintervention behavioral maintenance and motivational changes. Concurrently, attention should be given to populations with specific health conditions, digital health inequalities, and technology dependency challenges, to enhance the interpretive power and practical applicability of research in real-world contexts.

Conclusions

Against the backdrop of increasing physical inactivity and sedentary behavior among adolescents globally, this systematic review synthesizes evidence from 24 studies encompassing diverse research designs and multiple types of DHIs. Compared with previous systematic reviews that focused exclusively on specific study designs or single technological formats, this review provides a more holistic and comparative perspective on emerging evidence. The integrated findings indicate that DHIs generally contribute to enhancing adolescents' physical activity levels. However, their effectiveness varies considerably across intervention types, research designs, and application contexts. This finding addresses gaps in existing research, suggesting that both the adaptability of interventions and the implementation setting play crucial roles in shaping outcomes. Distinct strengths were observed across different research designs: RCTs provided relatively reliable evidence for direct and long-term intervention effects, while QR and CSS contributed to understanding underlying mechanisms and associations. QES and some MMS offered relatively limited representativeness due to scattered evidence. This finding supplements and refines the existing understanding of the evidence structure for DHIs. Further analysis indicates that the suitability of different DHIs varies across adolescent populations and practice settings, with the choice of intervention type significantly influencing outcomes. Concurrent attention to adolescents with specific health conditions, digital health inequities, and technology dependency challenges provides new perspectives for equitable implementation and sustainable scaling of DHIs in real-world settings. Despite limitations in the existing evidence—including methodological bias, inconsistent study quality, and insufficient follow-up data—the findings of this review remain valuable for understanding the role of DHIs in promoting adolescent physical activity. They provide useful insights for practical application, global health promotion, health equity, and innovation in cross-cultural intervention strategies. Future research should integrate the strengths of diverse study designs to conduct high-quality, long-term follow-up studies. Such studies should comprehensively consider target population characteristics, intervention duration, and regional variations to enhance the effectiveness and scalability of DHIs in real-world settings.

Acknowledgments

We are also grateful to all participants in this systematic review for their time, effort, and dedication. The authors declare that no generative artificial intelligence was used in the writing of this manuscript.

Funding

The authors received no financial support for the research, authorship, or publication of this article.

Data Availability

All data generated or analyzed during this systematic review are included in this published article and its multimedia appendices. Additional information is available from the corresponding author (LJ) on reasonable request.

Authors' Contributions

Conceptualization: RSF (lead), JJJ (equal), LJ (supporting)
Data curation: QYZ (lead), XYZ (supporting)
Formal analysis: RSF
Funding acquisition: LJ
Investigation: RSF (lead), JJJ (supporting)
Methodology: RSF (lead), LJ (equal), QYZ (supporting)
Project administration: RSF (lead), XYZ (supporting)

XSL•FO
RenderX

Resources: RSF (lead), JJJ (supporting), ZHW (supporting)
Supervision: LJ (lead), XYZ (supporting)
Validation: RSF (lead), QYZ (supporting), XYZ (supporting)
Visualization: RSF (lead), JJJ (supporting), QYZ (supporting)
Writing – original draft: RSF (lead), JJJ (supporting), QYZ (supporting)
Writing – review and editing: RSF (lead), JJJ (supporting), XYZ (supporting), LJ (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 28 KB - jmir_v28i1e82395_app1.docx]

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 28 KB - jmir_v28i1e82395_app2.docx]

Multimedia Appendix 3
Additional analysis.
[DOCX File , 516 KB - jmir_v28i1e82395_app3.docx]

References

1. Chaput J, Willumsen J, Bull F, Chou R, Ekelund U, Firth J, et al. 2020 WHO guidelines on physical activity and sedentary behaviour for children and adolescents aged 5-17 years: summary of the evidence. Int J Behav Nutr Phys Act 2020 Nov 26;17(1):141 [FREE Full text] [doi: 10.1186/s12966-020-01037-z] [Medline: 33239009]
2. Rose T, Barker M, Maria Jacob C, Morrison L, Lawrence W, Strömmer S, et al. A systematic review of digital interventions for improving the diet and physical activity behaviors of adolescents. J Adolesc Health 2017 Dec;61(6):669-677 [FREE Full text] [doi: 10.1016/j.jadohealth.2017.05.024] [Medline: 28822682]
3. Sampasa-Kanyinga H, Colman I, Goldfield GS, Janssen I, Wang J, Podinic I, et al. Combinations of physical activity, sedentary time, and sleep duration and their associations with depressive symptoms and other mental health problems in children and adolescents: a systematic review. Int J Behav Nutr Phys Act 2020 Jun 05;17(1):72 [FREE Full text] [doi: 10.1186/s12966-020-00976-x] [Medline: 32503638]
4. Düking P, Tafler M, Wallmann-Sperlich B, Sperlich B, Kleih S. Behavior change techniques in wrist-worn wearables to promote physical activity: content analysis. JMIR Mhealth Uhealth 2020 Nov 19;8(11):e20820 [FREE Full text] [doi: 10.2196/20820] [Medline: 33211023]
5. Au WW, Recchia F, Fong DY, Wong SHS, Chan DKC, Capio CM, et al. Effect of wearable activity trackers on physical activity in children and adolescents: a systematic review and meta-analysis. Lancet Digit Health 2024 Sep;6(9):e625-e639 [FREE Full text] [doi: 10.1016/S2589-7500(24)00139-0] [Medline: 39112110]
6. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
7. Collin C, Eyraud C, Martin P, Michel M, Le Roux E, Alberti C. A scoping review of outcome selection and accuracy of conclusions in complex digital health interventions for young people (2017-2023): methodological proposals for population health intervention research. BMC Med 2025 Jul 02;23(1):400 [FREE Full text] [doi: 10.1186/s12916-025-04245-1] [Medline: 40604855]
8. Partridge SR, Knight A, Todd A, McGill B, Wardak S, Alston L, et al. Addressing disparities: a systematic review of digital health equity for adolescent obesity prevention and management interventions. Obes Rev 2024 Dec;25(12):e13821. [doi: 10.1111/obr.13821] [Medline: 39256958]
9. Wang J, Zhu Z, Shuling Z, Fan J, Jin Y, Gao Z, et al. Effectiveness of mHealth app-based interventions for increasing physical activity and improving physical fitness in children and adolescents: systematic review and meta-analysis. JMIR Mhealth Uhealth 2024 Apr 30;12:e51478 [FREE Full text] [doi: 10.2196/51478] [Medline: 38687568]
10. He Z, Wu H, Yu F, Fu J, Sun S, Huang T, et al. Effects of smartphone-based interventions on physical activity in children and adolescents: systematic review and meta-analysis. JMIR Mhealth Uhealth 2021 Feb 01;9(2):e22601 [FREE Full text] [doi: 10.2196/22601] [Medline: 33522980]
11. Sequí-Domínguez I, Cavero-Redondo I, Álvarez-Bueno C, López-Gil JF, Martínez-Vizcaíno V, Pascual-Morena C. Effectiveness of eHealth interventions promoting physical activity in children and adolescents: systematic review and meta-analysis. J Med Internet Res 2024 Feb 21;26:e41649 [FREE Full text] [doi: 10.2196/41649] [Medline: 38381490]

XSL•FO
RenderX

12. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Rev Esp Cardiol (Engl Ed) 2021 Sep;74(9):790-799 [FREE Full text] [doi: 10.1016/j.rec.2021.07.010] [Medline: 34446261]
13. Barker T, Stone J, Sears K, Klugar M, Tufanaru C, Leonardi-Bee J, et al. The revised JBI critical appraisal tool for the assessment of risk of bias for randomized controlled trials. JBI Evid Synth 2023 Mar 01;21(3):494-506. [doi: 10.11124/JBIES-22-00430] [Medline: 36727247]
14. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
15. Soltero EG, Musaad SM, O'Connor TM, Thompson D, Norris K, Beech BM. Feasibility of Fit24, a digital diabetes prevention program for Hispanic adolescents: qualitative evaluation study. JMIR Form Res 2024 May 17;8:e54595 [FREE Full text] [doi: 10.2196/54595] [Medline: 38758584]
16. Goodyear VA, Kerner C, Quennerstedt M. Young people's uses of wearable healthy lifestyle technologies; surveillance, self-surveillance and resistance. Sport, Education and Society 2017 Sep 22;24(3):212-225. [doi: 10.1080/13573322.2017.1375907]
17. Van Dyck D, D'Haese S, Plaete J, De Bourdeaudhuij I, Deforche B, Cardon G. Opinions towards physical activity interventions using Facebook or text messaging: focus group interviews with vocational school-aged adolescents. Health Soc Care Community 2019 May;27(3):654-664. [doi: 10.1111/hsc.12679] [Medline: 30398291]
18. Dinç F, Kurt A, Yıldız D. The use of mobile health applications in the development of a healthy lifestyle of adolescents: a cross-sectional study. J Pediatr Nurs 2025;83:1-6. [doi: 10.1016/j.pedn.2025.04.012] [Medline: 40273677]
19. Mojica CM, Parra-Medina D, Yin Z, Akopian D, Esparza LA. Assessing media access and use among Latina adolescents to inform development of a physical activity promotion intervention incorporating text messaging. Health Promot Pract 2014 Jul 19;15(4):548-555 [FREE Full text] [doi: 10.1177/1524839913514441] [Medline: 24357863]
20. Ng K, Kokko S, Tammelin T, Kallio J, Belton S, O'Brien W, et al. Clusters of Adolescent Physical Activity Tracker Patterns and Their Associations With Physical Activity Behaviors in Finland and Ireland: Cross-Sectional Study. J Med Internet Res 2020 Sep 01;22(9):e18509-e18555 [FREE Full text] [doi: 10.2196/18509] [Medline: 32667894]
21. Mendoza JA, Baker KS, Moreno MA, Whitlock K, Abbey-Lambertz M, Waite A, et al. A Fitbit and Facebook mHealth intervention for promoting physical activity among adolescent and young adult childhood cancer survivors: a pilot study. Pediatr Blood Cancer 2017 Dec;64(12):e26660. [doi: 10.1002/pbc.26660] [Medline: 28618158]
22. Larsen B, Benitez T, Cano M, Dunsiger SS, Marcus BH, Mendoza-Vasconez A, et al. Web-based physical activity intervention for Latina adolescents: feasibility, acceptability, and potential efficacy of the Niñas Saludables study. J Med Internet Res 2018 May 09;20(5):e170 [FREE Full text] [doi: 10.2196/jmir.9206] [Medline: 29743151]
23. Thompson D, Cantu D, Ramirez B, Cullen KW, Baranowski T, Mendoza J, et al. Texting to increase adolescent physical activity: feasibility assessment. Am J Health Behav 2016 Jul;40(4):472-483 [FREE Full text] [doi: 10.5993/AJHB.40.4.9] [Medline: 27338994]
24. Guthrie N, Bradlyn A, Thompson SK, Yen S, Haritatos J, Dillon F, et al. Development of an accelerometer-linked online intervention system to promote physical activity in adolescents. PLoS One 2015;10(5):e0128639 [FREE Full text] [doi: 10.1371/journal.pone.0128639] [Medline: 26010359]
25. Direito A, Jiang Y, Whittaker R, Maddison R. Smartphone apps to improve fitness and increase physical activity among young people: protocol of the Apps for IMproving FITness (AIMFIT) randomized controlled trial. BMC Public Health 2015 Jul 11;15:635 [FREE Full text] [doi: 10.1186/s12889-015-1968-y] [Medline: 26159834]
26. Chen J, Guedes CM, Cooper BA, Lung AE. Short-term efficacy of an innovative mobile phone technology-based intervention for weight management for overweight and obese adolescents: pilot study. Interact J Med Res 2017 Aug 02;6(2):e12 [FREE Full text] [doi: 10.2196/ijmr.7860] [Medline: 28768612]
27. Caillaud C, Ledger S, Diaz C, Clerc G, Galy O, Yacef K. iEngage: a digital health education program designed to enhance physical activity in young adolescents. PLoS One 2022;17(10):e0274644 [FREE Full text] [doi: 10.1371/journal.pone.0274644] [Medline: 36197890]
28. Egilsson E, Bjarnason R, Njardvik U. Usage and weekly attrition in a smartphone-based health behavior intervention for adolescents: pilot randomized controlled trial. JMIR Form Res 2021 Feb 17;5(2):e21432 [FREE Full text] [doi: 10.2196/21432] [Medline: 33481750]
29. Staiano AE, Beyl RA, Guan W, Hendrick CA, Hsia DS, Newton RL. Home-based exergaming among children with overweight and obesity: a randomized clinical trial. Pediatr Obes 2018 Nov;13(11):724-733 [FREE Full text] [doi: 10.1111/ijpo.12438] [Medline: 30027607]
30. Stasinaki A, Büchter D, Shih CI, Heldt K, Güsewell S, Brogle B, et al. Effects of a novel mobile health intervention compared to a multi-component behaviour changing program on body mass index, physical capacities and stress parameters in adolescents with obesity: a randomized controlled trial. BMC Pediatr 2021 Jul 09;21(1):308 [FREE Full text] [doi: 10.1186/s12887-021-02781-2] [Medline: 34243738]

XSL•FO
RenderX

31. Ortega A, Cushing CC. Design of a temporally augmented text messaging bot to improve adolescents' physical activity and engagement: proof-of-concept study. JMIR Form Res 2024 Oct 10;8:e60171 [FREE Full text] [doi: 10.2196/60171] [Medline: 39388222]
32. Willinger L, Schweizer F, Böhm B, Scheller DA, Jonas S, Oberhoffer-Fritz R, et al. Evaluation of the gamified application KIJANI to promote physical activity in children and adolescents: a multimethod study. Digit Health 2024;10:20552076241271861 [FREE Full text] [doi: 10.1177/20552076241271861] [Medline: 39161345]
33. Schoenfelder E, Moreno M, Wilner M, Whitlock KB, Mendoza JA. Piloting a mobile health intervention to increase physical activity for adolescents with ADHD. Prev Med Rep 2017 Jun;6:210-213 [FREE Full text] [doi: 10.1016/j.pmedr.2017.03.003] [Medline: 28373931]
34. Koorts H, Salmon J, Timperio A, Ball K, Macfarlane S, Lai SK, et al. Translatability of a wearable technology intervention to increase adolescent physical activity: mixed methods implementation evaluation. J Med Internet Res 2020 Aug 07;22(8):e13573 [FREE Full text] [doi: 10.2196/13573] [Medline: 32763872]
35. Glaser M, Green G, Barak S, Bord S, Levi S, Jakobovich R, et al. The effects of the Friendship Online Intervention Program on physical activity, substance abuse, psychosomatic symptoms, and well-being among at-risk youth. J Adolesc 2024 Feb;96(2):251-265. [doi: 10.1002/jad.12272] [Medline: 37985148]
36. Garde A, Umedaly A, Abulnaga SM, Robertson L, Junker A, Chanoine JP, et al. Assessment of a mobile game ("MobileKids Monster Manor") to promote physical activity among children. Games Health J 2015 Apr;4(2):149-158. [doi: 10.1089/g4h.2014.0095] [Medline: 26181809]
37. Mateo-Orcajada A, Abenza-Cano L, Albaladejo-Saura MD, Vaquero-Cristóbal R. Mandatory after-school use of step tracker apps improves physical activity, body composition and fitness of adolescents. Educ Inf Technol (Dordr) 2023 Jan 24:1-32 [FREE Full text] [doi: 10.1007/s10639-023-11584-0] [Medline: 36714445]
38. Cushing C, Bejarano C, Ortega A, Sayre N, Fedele D, Smyth J. Adaptive mHealth intervention for adolescent physical activity promotion. J Pediatr Psychol 2021 Jun 03;46(5):536-546. [doi: 10.1093/jpepsy/jsaa125] [Medline: 33484137]
39. Moffatt S, White M, Mackintosh J, Howel D. Using quantitative and qualitative data in health services research - what happens when mixed method findings conflict? [ISRCTN61522618]. BMC Health Serv Res 2006 Mar 08;6:28 [FREE Full text] [doi: 10.1186/1472-6963-6-28] [Medline: 16524479]
40. Wang X, Cheng Z. Cross-sectional studies: strengths, weaknesses, and recommendations. Chest 2020 Jul;158(1S):S65-S71. [doi: 10.1016/j.chest.2020.03.012] [Medline: 32658654]
41. Domhardt M, Schröder A, Geirhos A, Steubl L, Baumeister H. Efficacy of digital health interventions in youth with chronic medical conditions: a meta-analysis. Internet Interv 2021 Apr;24:100373 [FREE Full text] [doi: 10.1016/j.invent.2021.100373] [Medline: 33732626]
42. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000 Jan;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
43. Zhang G, Feng W, Zhao L, Zhao X, Li T. The association between physical activity, self-efficacy, stress self-management and mental health among adolescents. Sci Rep 2024 Mar 06;14(1):5488 [FREE Full text] [doi: 10.1038/s41598-024-56149-4] [Medline: 38448518]
44. Teixeira PJ, Silva MN, Mata J, Palmeira AL, Markland D. Motivation, self-determination, and long-term weight control. Int J Behav Nutr Phys Act 2012 Mar 02;9:22 [FREE Full text] [doi: 10.1186/1479-5868-9-22] [Medline: 22385818]
45. Collin C, Eyraud C, Martin P, Michel M, Le Roux E, Alberti C. A scoping review of outcome selection and accuracy of conclusions in complex digital health interventions for young people (2017-2023): methodological proposals for population health intervention research. BMC Med 2025 Jul 02;23(1):400 [FREE Full text] [doi: 10.1186/s12916-025-04245-1] [Medline: 40604855]
46. Stavric V, Kayes NM, Rashid U, Saywell NL. The effectiveness of self-guided digital interventions to improve physical activity and exercise outcomes for people with chronic conditions: a systematic review and meta-analysis. Front Rehabil Sci 2022;3:925620 [FREE Full text] [doi: 10.3389/fresc.2022.925620] [Medline: 36188933]
47. Michalsen H, Wangberg SC, Hartvigsen G, Jaccheri L, Muzny M, Henriksen A, et al. Physical activity with tailored mHealth support for individuals with intellectual disabilities: protocol for a randomized controlled trial. JMIR Res Protoc 2020 Jun 29;9(6):e19213 [FREE Full text] [doi: 10.2196/19213] [Medline: 32437328]
48. Ma J, Floegel T, Li L, Leese J, De Vera MA, Beauchamp M, et al. Tailored physical activity behavior change interventions: challenges and opportunities. Transl Behav Med 2021 Dec 14;11(12):2174-2181 [FREE Full text] [doi: 10.1093/tbm/ibab106] [Medline: 34424344]
49. Shameli A, Althoff T, Saberi A, Leskovec J. How gamification affects physical activity: large-scale analysis of walking challenges in a mobile application. Proc Int World Wide Web Conf 2017 Apr;2017:455-463 [FREE Full text] [doi: 10.1145/3041021.3054172] [Medline: 28990011]
50. Bandura A. Self-Efficacy: The Exercise of Control. New York, NY: W. H. Freeman & Co; 1997.
51. Flølo TN, Tørris C, Riiser K, Almendingen K, Chew HSJ, Fosså A, et al. Digital health interventions to treat overweight and obesity in children and adolescents: an umbrella review. Obes Rev 2025 Jun;26(6):e13905. [doi: 10.1111/obr.13905] [Medline: 39972996]

XSL•FO
RenderX

52. van Baak MA, Pramono A, Battista F, Beaulieu K, Blundell JE, Busetto L, et al. Effect of different types of regular exercise on physical fitness in adults with overweight or obesity: systematic review and meta-analyses. Obes Rev 2021 Jul;22 Suppl 4(Suppl 4):e13239 [FREE Full text] [doi: 10.1111/obr.13239] [Medline: 33939229]
53. Fismen A, Galler M, Klepp K, Chatelan A, Residori C, Ojala K, et al. Weight status and mental well-being among adolescents: the mediating role of self-perceived body weight. a cross-national survey. J Adolesc Health 2022 Aug;71(2):187-195. [doi: 10.1016/j.jadohealth.2022.02.010] [Medline: 35428562]
54. Kollins SH, DeLoss DJ, Cañadas E, Lutz J, Findling RL, Keefe RSE, et al. A novel digital intervention for actively reducing severity of paediatric ADHD (STARS-ADHD): a randomised controlled trial. Lancet Digit Health 2020 Apr;2(4):e168-e178 [FREE Full text] [doi: 10.1016/S2589-7500(20)30017-0] [Medline: 33334505]
55. Wang W, Cheng J, Song W, Shen Y. The effectiveness of wearable devices as physical activity interventions for preventing and treating obesity in children and adolescents: systematic review and meta-analysis. JMIR Mhealth Uhealth 2022 Apr 08;10(4):e32435 [FREE Full text] [doi: 10.2196/32435] [Medline: 35394447]
56. Stamatis CA, Farlow DN, Mercaldi C, Suh M, Maple A, Savarese A, et al. Two single arm trials of AKL-T01, a digital therapeutic for adolescents and adults with ADHD. Npj Ment Health Res 2024 Jun 19;3(1):30 [FREE Full text] [doi: 10.1038/s44184-024-00075-w] [Medline: 38898133]
57. Cuzick J. The importance of long-term follow up of participants in clinical trials. Br J Cancer 2023 Feb;128(3):432-438 [FREE Full text] [doi: 10.1038/s41416-022-02038-4] [Medline: 36456713]
58. Kahwati L, Viswanathan M, Golin CE, Kane H, Lewis M, Jacobs S. Identifying configurations of behavior change techniques in effective medication adherence interventions: a qualitative comparative analysis. Syst Rev 2016 May 04;5:83 [FREE Full text] [doi: 10.1186/s13643-016-0255-z] [Medline: 27209092]
59. Bi S, Yuan J, Wang Y, Zhang W, Zhang L, Zhang Y, et al. Effectiveness of digital health interventions in promoting physical activity among college students: systematic review and meta-analysis. J Med Internet Res 2024 Nov 20;26:e51714 [FREE Full text] [doi: 10.2196/51714] [Medline: 39566049]
60. Kassim PSJ, Muhammad NA, Rahman NFA, Sidik SM, Essau CA, Shah SA. Digital behaviour change interventions to promote physical activity in overweight and obese adolescents: a systematic review protocol. Syst Rev 2022 Sep 05;11(1):188 [FREE Full text] [doi: 10.1186/s13643-022-02060-w] [Medline: 36064617]
61. Barker T, Habibi N, Aromataris E, Stone J, Leonardi-Bee J, Sears K, et al. The revised JBI critical appraisal tool for the assessment of risk of bias for quasi-experimental studies. JBI Evid Synth 2024 Mar 01;22(3):378-388. [doi: 10.11124/JBIES-23-00268] [Medline: 38287725]
62. Gómez-Cuesta N, Mateo-Orcajada A, Meroño L, Abenza-Cano L, Vaquero-Cristóbal R. A mobile app-based intervention improves anthropometry, body composition and fitness, regardless of previous active-inactive status: a randomized controlled trial. Front Public Health 2024;12:1380621 [FREE Full text] [doi: 10.3389/fpubh.2024.1380621] [Medline: 39193194]
63. Silva AG, Simões P, Queirós A, P Rocha N, Rodrigues M. Effectiveness of mobile applications running on smartphones to promote physical activity: a systematic review with meta-analysis. Int J Environ Res Public Health 2020 Mar 27;17(7):2251 [FREE Full text] [doi: 10.3390/ijerph17072251] [Medline: 32230769]
64. Howard-Wilson S, Ching J, Gentile S, Ho M, Garcia A, Ayturk D, et al. Efficacy of a multimodal digital behavior change intervention on lifestyle behavior, cardiometabolic biomarkers, and medical expenditure: protocol for a randomized controlled trial. JMIR Res Protoc 2024 Oct 30;13:e50378 [FREE Full text] [doi: 10.2196/50378] [Medline: 39475852]
65. Böhm B, Karwiese SD, Böhm H, Oberhoffer R. Effects of mobile health including wearable activity trackers to increase physical activity outcomes among healthy children and adolescents: systematic review. JMIR Mhealth Uhealth 2019 Apr 30;7(4):e8298 [FREE Full text] [doi: 10.2196/mhealth.8298] [Medline: 31038460]
66. Singh B, Ahmed M, Staiano AE, Gough C, Petersen J, Vandelanotte C, et al. A systematic umbrella review and meta-meta-analysis of eHealth and mHealth interventions for improving lifestyle behaviours. NPJ Digit Med 2024 Jul 05;7(1):179 [FREE Full text] [doi: 10.1038/s41746-024-01172-y] [Medline: 38969775]
67. Bas-Sarmiento P, Julián-López C, Fernández-Gutiérrez M, Poza-Méndez M, Marín-Paz A. Gamified eHealth interventions for health promotion and disease prevention in children and adolescents: a scoping review. Humanit Soc Sci Commun 2025 Mar 20;12(1):397. [doi: 10.1057/s41599-025-04670-w]
68. Badr J, Motulsky A, Denis J. Digital health technologies and inequalities: a scoping review of potential impacts and policy recommendations. Health Policy 2024 Aug;146:105122 [FREE Full text] [doi: 10.1016/j.healthpol.2024.105122] [Medline: 38986333]
69. Ghadi YY, Shah SFA, Waheed W, Mazhar T, Ahmad W, Saeed MM, et al. Integration of wearable technology and artificial intelligence in digital health for remote patient care. J Cloud Comp 2025 Jul 21;14(1):39. [doi: 10.1186/s13677-025-00759-4]
70. Grand VR, Inc. U.S. Digital Health Market Size, Share. Grand View Research. 2024. URL: https://www.grandviewresearch.com/industry-analysis/us-digital-health-market-report [accessed 2025-11-10]
71. Majcherek D, Hegerty SW, Kowalski AM, Lewandowska MS, Dikova D. Opportunities for healthcare digitalization in Europe: comparative analysis of inequalities in access to medical services. Health Policy 2024 Jan;139:104950 [FREE Full text] [doi: 10.1016/j.healthpol.2023.104950] [Medline: 38061175]

XSL•FO
RenderX

72. El-Jardali F, Bou-Karroum L, Jabbour M, Bou-Karroum K, Aoun A, Salameh S, et al. Digital health in fragile states in the Middle East and North Africa (MENA) region: a scoping review of the literature. PLoS One 2023;18(4):e0285226 [FREE Full text] [doi: 10.1371/journal.pone.0285226] [Medline: 37115778]
73. Sylla B, Ismaila O, Diallo G. 25 years of digital health toward universal health coverage in low- and middle-income countries: rapid systematic review. J Med Internet Res 2025 May 29;27:e59042 [FREE Full text] [doi: 10.2196/59042] [Medline: 40440696]
74. ew GSMA report shows mobile internet connectivity continues to grow globally but barriers for 3.45 billion unconnected people remain. GSMA. 2024 Oct 23. URL: https://www.gsma.com/newsroom/press-release/new-gsma-report-shows-mobile-internet-connectivity-continues-to-grow-globally-but-barriers-for-3-45-billion-unconnected-people-remain/ [accessed 2025-11-10]
75. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: a systematic and meta-analytic review. J Consult Clin Psychol 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
76. Pratap A, Neto EC, Snyder P, Stepnowsky C, Elhadad N, Grant D, et al. Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. NPJ Digit Med 2020;3:21 [FREE Full text] [doi: 10.1038/s41746-020-0224-8] [Medline: 32128451]
77. Pathiravasan CH, Zhang Y, Wang X, Trinquart L, Benjamin EJ, Borrelli B, et al. Factors associated with long-term use of digital devices in the electronic Framingham Heart Study. NPJ Digit Med 2022 Dec 27;5(1):195 [FREE Full text] [doi: 10.1038/s41746-022-00735-1] [Medline: 36572707]
78. Jakob R, Harperink S, Rudolf AM, Fleisch E, Haug S, Mair JL, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371 [FREE Full text] [doi: 10.2196/35371] [Medline: 35612886]
79. Kelders SM, Kok RN, Ossebaard HC, Van Gemert-Pijnen JE. Persuasive system design does matter: a systematic review of adherence to web-based interventions. J Med Internet Res 2012 Nov 14;14(6):e152 [FREE Full text] [doi: 10.2196/jmir.2104] [Medline: 23151820]

Abbreviations

BCW: Behavior Change Wheel
COM-B: capability, opportunity, motivation—behavior
CSS: cross-sectional study
DHI: digital health intervention
GRADE: Grading of Recommendations, Assessment, Development and Evaluation
JBI: Joanna Briggs Institute
MeSH: Medical Subject Headings
mHealth: mobile health
MMS: mixed methods study
PICOS: population, intervention, comparator, outcomes, and study design
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Search Extension
PROSPERO: International Prospective Register of Systematic Reviews
QES: quasi-experimental study
QR: quantitative research
RCT: randomized controlled trial

Edited by S Brini; submitted 14.Aug.2025; peer-reviewed by C Ahmadu, O Akhadelor; comments to author 08.Sep.2025; accepted 02.Jan.2026; published 27.Feb.2026.

Please cite as:
Fan RS, Jiang JJ, Zhou QY, Zhang XY, Wu ZH, Ji L
Digital Health Interventions to Promote Physical Activity Among Adolescents: Systematic Review
J Med Internet Res 2026;28:e82395
URL: https://www.jmir.org/2026/1/e82395
doi:10.2196/82395
PMID:

XSL•FO
RenderX

©Rui Shi Fan, Jia Jun Jiang, Qing Yuan Zhou, Xin Yue Zhang, Zhou Hang Wu, Liu Ji. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Digital Technologies That Support Meaningful Connections in Care Homes: Scoping Review

Deborah Muldrew[1], PhD; Rosemary Bradley[2], MSc; Kelly Conway[1], MSc

[1]Institute of Nursing and Health Research, University of Ulster, Belfast, United Kingdom

[2]Institute of Nursing and Health Research, University of Ulster, Londonderry, United Kingdom

Corresponding Author:
Deborah Muldrew, PhD
Institute of Nursing and Health Research
University of Ulster
Ulster University Belfast Campus
2-24 York Street
Belfast, BT15 1AP
United Kingdom
Phone: 44 2895365301
Email: d.muldrew@ulster.ac.uk

Abstract

Background: Meaningful connections in which people feel valued, seen, and heard are essential for social health and well-being. However, individual, systemic, and structural barriers exist within care homes that exacerbate risks for social isolation in this population. Leveraging digital technology to promote meaningful connections has the potential to affect positive health outcomes; however, the evidence base within the care home environment is not clear.

Objective: This scoping review aimed to identify how digital technologies are used in long-term care settings to facilitate meaningful connections.

Methods: Following Joanna Briggs Institute and PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) reporting guidelines, 6 online databases (CINAHL, MEDLINE, PsycINFO, Scopus, IEEE Explore, and ACM) were searched on December 12, 2025. Studies were included if they were published in English; included care home residents, relatives, or staff; and directly discussed the application of a digital technology that focused on building meaningful connections. No date limiters were applied. Gray literature was searched through Google Scholar and conference proceedings, and forward and backward citation searching was also undertaken. Papers were managed in Covidence and reviewed against inclusion criteria by 3 reviewers (DM, RB, and KC). Results were presented using tabular, graphical, and narrative summaries to describe how the findings relate to the review objectives.

Results: In total, 72 studies were included. More than half of the studies were published since 2022 (38/72, 53%). Australia (15/72, 21%) and the Netherlands (12/72, 17%) had the highest frequency of publications. Technologies including robotics, virtual reality, mobile or tablet apps, digital devices, and online programs have been used in care homes. Factors underpinning the development of meaningful connections include getting to know the person, increases autonomy and choice, source of enjoyment and fun, facilitates communication, and builds community. Engagement, well-being or satisfaction, emotional response, quality of life, purpose and meaning, social closeness, loneliness, depression and anxiety, and psychosocial capacity were the main indicators of meaningful connections. Limitations include English-language restriction, potential missed studies on social connection, exclusion of passive technologies, and heterogeneous outcome measures.

Conclusions: This review is the first to explore how digital technologies are applied in care homes to facilitate meaningful connections and identify how digital technology is a catalyst for meaningful human connection rather than a replacement. The application of generative artificial intelligence technology is currently a gap in the evidence base and needs to be considered in conjunction with key stakeholders to ensure that future developments meet the social needs of the entire community. Future research needs to consider how new evaluation metrics can be developed that combine digital engagement data with validated measures of meaningful connection to assess the person-centered impact of generative artificial intelligence technologies in care homes.

(J Med Internet Res 2026;28:e88181) doi:10.2196/88181

XSL•FO
RenderX

KEYWORDS

scoping review; digital technology; meaningful connection; social connection; care homes; nursing homes; older people; aged

Introduction

Social connection describes how individuals relate to each other, encompassing both observable aspects such as engagement with social networks and the experiences of loneliness or social connectedness [1]. People who are socially connected feel an overall sense of belonging to a group or community. The World Health Organization (WHO) recognizes social connection as a global public health priority. Unfortunately, societal trends over a number of decades indicate that, as a population, we have become less socially connected, and a high percentage of the population is lonely [2]. Current global estimates suggest that 1 in 4 older adults experience social isolation [3]. This concern is likely to become more prevalent, given the speed of population aging, which is significantly faster than in previous decades. Current population predictions by the WHO estimate that by 2050, the number of individuals aged 60 years and older will double to 2.1 billion globally, and the number of individuals aged 80 years and older will triple to 426 million [4]. An aging population, coupled with societal shifts in how care is delivered, is predicted to place a substantial rise in the demand on long-term care facilities [5]. Long-term care consists of a range of medical or nursing care, personal care (help with activities of daily living), assistance services (help with instrumental activities of daily living), and support to maintain social relationships and activities [6]. The term "care homes" encompasses 2 types of long-term care settings: residential care homes (for people requiring personal care) and nursing homes (for people requiring personal and nursing care) [7]. Individual, systemic, and structural barriers exist within care homes that exacerbate the risks for social isolation in this population [8].

While the goal toward social connection is well documented, meaningful connections are the first step toward creating an enriching life and can have a significant impact on health and well-being [9]. Meaningful connections are characterized by deeper, more intimate relationships that make the person feel seen, heard, and valued. For the purpose of this research, we have defined meaningful connections as "interpersonal interactions, whether one-off or repeated, that foster feelings of being seen or heard, empathy, or enhanced social belonging; and/or lead to indicators such as reduced loneliness, increased engagement, or improvements in psychosocial outcomes." An avenue to such meaningful connection is meaningful activity. National Institute for Health and Care Excellence guidelines recommend that people should be offered opportunities to participate in meaningful activity that promotes health and mental well-being [10]. Meaningful activities are defined as "all activities or occupations that are significant or meaningful for the person and reflect someone's current and past interests, routines, habits, and roles and are adjusted to someone's abilities" [11]. Engagement in personally meaningful activities leads to improved psychological well-being [12], and research into people with dementia living in care homes reported improvements in behavioral and psychological symptoms through the provision of meaningful activities [13]. However, when the focus becomes the activity itself and not the meaningful connection formed between the people involved, this approach could lead to task-oriented approaches to care, currently being challenged within health care [14]. Furthermore, activities viewed as something separate from everyday routine, and require additional staffing to coordinate in a busy environment where resources and staffing are already overstretched, are less likely to be meaningful or person-centered [15]. Instead, considerations should be given to the meaningful connections that can form between people at any time, not just during planned or purposeful meaningful activity.

According to Kemp et al [16], meaningful connections are essential for well-being and may serve as effective nonpharmacological approaches to mitigate adverse behavioral and psychological manifestations such as anxiety, depression, apathy, agitation, and aggression. Gonnord et al [17] also demonstrated that engagement could positively affect a number of aspects of well-being, such as social connection, personal autonomy, physical health, and mental stimulation. Social connectedness and participation in meaningful activities are important sources of joy, which have been shown in the literature as important to the health and social care of older people [18]. In the nursing home context, meaningful social relationships with peers and families foster a sense of belonging, purpose, achievement, continuity, and significance, contributing to their psychosocial well-being [19]. Within nursing homes, social connectedness may be an important modifiable risk factor for adverse health outcomes [20]. However, barriers to quality social participation exist at an individual level, for example, personal motivations and preexisting social networks, and at a community level, including accessibility and neighborhood cohesion [21]. Although these findings come from research into older adults more widely, their relevance to older adults living in care homes is clear. As highlighted by Nygaard et al [22], there are additional obstacles and challenges to achieving person-centered meaningful connections in care homes including a regimented structure to care home life and a lack of time for staff to spend quality time with individual residents, as well as a lack of connectedness within meaningful social relationships.

Leveraging digital technology to promote social connectedness has the potential to affect positive health outcomes [23]. The introduction of digital technologies was accelerated into care homes as a result of global restrictions during the COVID-19 pandemic, despite reported systemic and structural barriers [24]. Technology-supported interventions can improve old-age social well-being among older adults living in social isolation [25]; however, the evidence base is typically condition-specific (eg, dementia) or focused on treating a clinical outcome (eg, depression), rather than exploring the role of technology to facilitate meaningful connections among residents, relatives, and staff. Furthermore, much of the evidence based around digital technology considers technology solely for passive monitoring purposes such as digital movement sensors or blood pressure monitors. Therefore, a review of the evidence is required to fill the current evidence gap on how digital

XSL•FO
RenderX

technologies can be used to form meaningful connections in care homes. Given the overall limited focus within the evidence base on application of digital technology to support meaningful connections among the wider care home population, formulating a precise question for a systematic review was not possible. Instead, a scoping review supported the mapping of all digital technologies applied to the care home setting with the aim of supporting meaningful connections [26]. A preliminary search of CINAHL, MEDLINE, and *JBI Evidence Synthesis* was conducted, and no current or underway systematic reviews or scoping reviews on the topic were identified.

The aim of this scoping review is to identify how digital technologies are used in long-term care settings to facilitate meaningful connections. The objectives related to this aim are (1) to identify the scope of digital technologies being applied within the care home setting to build meaningful connections, (2) to understand the factors through which digital technologies are able to support meaningful connections, (3) to identify which outcomes are measured to evidence a change in meaningful connections, and (4) to identify gaps in the evidence base to inform future research, policy, and practice.

Methods

Design

A scoping review methodology following guidance from the Joanna Briggs Institute (JBI) and PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews; Multimedia Appendix 1) [27] reporting guidelines was selected as the most appropriate evidence synthesis approach, based on the guidance from Munn et al [26]. The aim was to scope and map the existing literature on digital technologies used in long-term care homes to support meaningful connections, including technologies used, factors showing how the digital technologies promote meaningful connections, and outcome measures of such connections. The systematic approach outlined by the JBI [28] was chosen for this review to provide a high degree of rigor and transparency, to ensure that the review can be replicated easily, and to increase its value to health care decision-makers. Adherence to the items reported in the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) checklist was used to strengthen the transparency of reporting the search [29]. Ethics approval was not required because scoping reviews synthesize existing publicly available literature without involving human participants or identifiable data.

Protocol

A protocol for the scoping review was developed and subject to peer review within the first author's (DM) higher education institution. However, this protocol was not registered externally. Following the review of the protocol, changes to the final review included the addition of "aged care" as a concept, the addition of 2 databases (ACM Digital Library and IEEE Xplore), and the expansion of the gray literature searches to include a range of relevant conference proceedings from 2025.

Eligibility Criteria

As the scoping review question guides and directs the development of the inclusion criteria, the JBI recommends the use of the population, concept, and context mnemonic in ensuring the eligibility of the primary research question [28]. The research questions were developed through consultation with the research team and key stakeholders focusing on the overall aim of the scoping review. Textbox 1 presents the identified population, concept, and context and the primary research question.

Textbox 1. Population, concept, and context framework for defining the inclusion criteria of studies.

Population

Participants include care home residents, relatives, and staff. Care homes include residential and nursing homes. Studies were excluded if they were undertaken in nonresidential (day care) services, supported housing, facilities for people with learning disabilities or enduring mental health conditions, or other community care settings such as in the person's own home.

Concept

Meaningful connections, activities, or engagement as a goal of the digital technology. Studies were excluded if they focused solely on clinical health or other outcomes without data related specifically to meaningful connections.

Context

Digital technologies used actively with or by people in care homes. Studies were excluded if the technology was solely for passive monitoring purposes, such as digital movement sensors or blood pressure monitors, or did not prioritize the creation of meaningful connections through digital means.

Research question

How do digital technologies facilitate meaningful connections in care homes?

Studies published in English were included due to practical limitations (translator not available to the research team). No date limitations were applied to the searches to allow for an exploration of how digital technologies have changed over time in type and frequency of application in care homes. No published search filters were used. This scoping review included all primary research, including qualitative, quantitative, and mixed methods approaches. Opinion papers and editorials were also considered for inclusion in this scoping review; however, they were required to report some primary data.

Information Sources

CINAHL Ultimate (EBSCOhost), MEDLINE (Ovid), Scopus, ACM Digital Library, IEEE XPlore, and PsycINFO (Ovid) were independently searched on December 12, 2025. Gray literature

XSL•FO
RenderX

was searched using Google Scholar (first 100 hits) and websites including the WHO and National Institute for Health and Care Excellence. Published conference proceedings from 2025 were hand-searched to identify any new publications that may not be published and accessible through traditional database searches. Conferences included the CHI Conference on Human Factors in Computing Systems, Good IT Proceedings of the 2025 International Conference on Information Technology for Social Good, COMPASS '25: Proceedings of the 2025 ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies, C&C '25: Proceedings of the 2025 Conference on Creativity and Cognition, MuC '25: Proceedings of the Mensch und Computer 2025, WWW '25: Companion Proceedings of the ACM on Web Conference 2025, and VRST '25: Proceedings of the 2025 31st ACM Symposium on Virtual Reality Software and Technology.

Search Strategy

The search strategy aimed to locate both published and unpublished studies. A 3-step search strategy was used in this review. First, an initial limited search of MEDLINE Ovid and CINAHL Ultimate was undertaken to identify papers on the topic in conjunction with a university librarian and subject specialist. This search was based on the subject expertise of the team and the database searching skills of the librarian. The text words contained in the titles and abstracts of relevant papers and the index terms used to describe the papers were used to develop a full search strategy (Multimedia Appendix 2).

The search strategy, including all identified keywords and index terms, was adapted for each included database and information source. A broad range of keywords and MeSH (Medical Subject Headings) headings were used to reduce the likelihood of missing relevant literature, and then, the inclusion and exclusion criteria were applied strictly to ensure that the papers included were relevant to the objectives. Forward and backward citation searching on the included sources of evidence was undertaken to identify additional studies. No authors were contacted, as all papers identified through the search strategy were available. Email alerts were set up for each database to provide weekly updates of new literature.

Study Selection

Following the search, all identified citations were collated and uploaded into Covidence, and duplicates were removed. Duplication was managed within Covidence by the in-built duplication feature and thorough review from the authors. Studies were sorted alphabetically by author to allow for identification of additional duplicates missed by Covidence. Following a pilot test, titles and abstracts were screened by multiple independent reviewers (DM, KC, and RB) for assessment against the inclusion criteria for the review. Potentially relevant sources were retrieved in full and stored in Covidence. The full text of selected citations was assessed in detail against the inclusion criteria by multiple independent reviewers (DM, KC, and RB). Reasons for exclusion of sources of evidence at full text that did not meet the inclusion criteria were recorded and reported in the scoping review. Any disagreements that arose between the reviewers at each stage of the selection process were resolved through discussion.

Data Charting

Data extraction was undertaken to develop a logical and descriptive summary of the results, aligned to the objectives of the review. Data were extracted from papers included in the scoping review by multiple independent reviewers (DM, KC, and RB) using a data extraction tool developed by the research team. It was based on the JBI data extraction template to meet the objectives of the review. Each researcher (DM, KC, and RB) piloted the data extraction tool on 5 papers before meeting to discuss the utility of the tool. Originally, all findings were collected in one column; however, this was leading to confusion, and key data being overlooked. The data extraction tool was updated to have 3 separate results columns: one for the type of digital technology discussed, one for the factors that create meaningful connections, and one for the outcome measures identified. Any disagreements that arose between the reviewers were resolved through discussion. Once data extraction was completed for the included studies, a second reviewer (RB) repeated the data extraction process for 10% (7/72) of the included studies to ensure that all relevant information was extracted accurately. See Multimedia Appendix 3 for the completed data extraction table.

Data Items

The data extracted included specific details about the author, year of publication, country of origin, population and sample size, methods, type of technology used, factors supporting meaningful connection, and outcome measures.

Quality Appraisal

Quality appraisal was not undertaken, aligned to methodological norms for scoping reviews. However, a descriptive assessment of each study design, based on the levels of evidence identified by Melnyk and Fineout-Overholt [30], was included to support the development of more robust policy and practice recommendations. The rating scale ranged from 1=strong (systematic review or meta-analysis of randomized controlled trials) to 7=weak (opinion pieces and narrative literature review). In total, 65 studies were categorized as level 6, representing evidence from a single qualitative or descriptive study.

Synthesis of Results

An overview of included studies was used to summarize bibliographic information. A narrative summary was developed to describe how the findings relate to the core objectives of the review: the digital technologies used, the factors through which meaningful connections are established, and the outcome measures reported. Three independent researchers (DM, KC, and RB) read and reread the included studies and extracted data relevant to the core objectives. The extracted data were then synthesized to ensure that a comprehensive account of relevant information was presented in the final data extraction table. A narrative summary accompanying the data presented was developed by the first author (DM) and reviewed and edited by the second and third authors (RB and KC).

XSL•FO
RenderX

Results

Selection of Sources of Evidence

In total, 12,742 papers were identified through the database searches, citation searching, and gray literature. Following the removal of duplicates, 12,219 papers were screened against the inclusion and exclusion criteria. The results of the search are shown in the PRISMA flow diagram (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

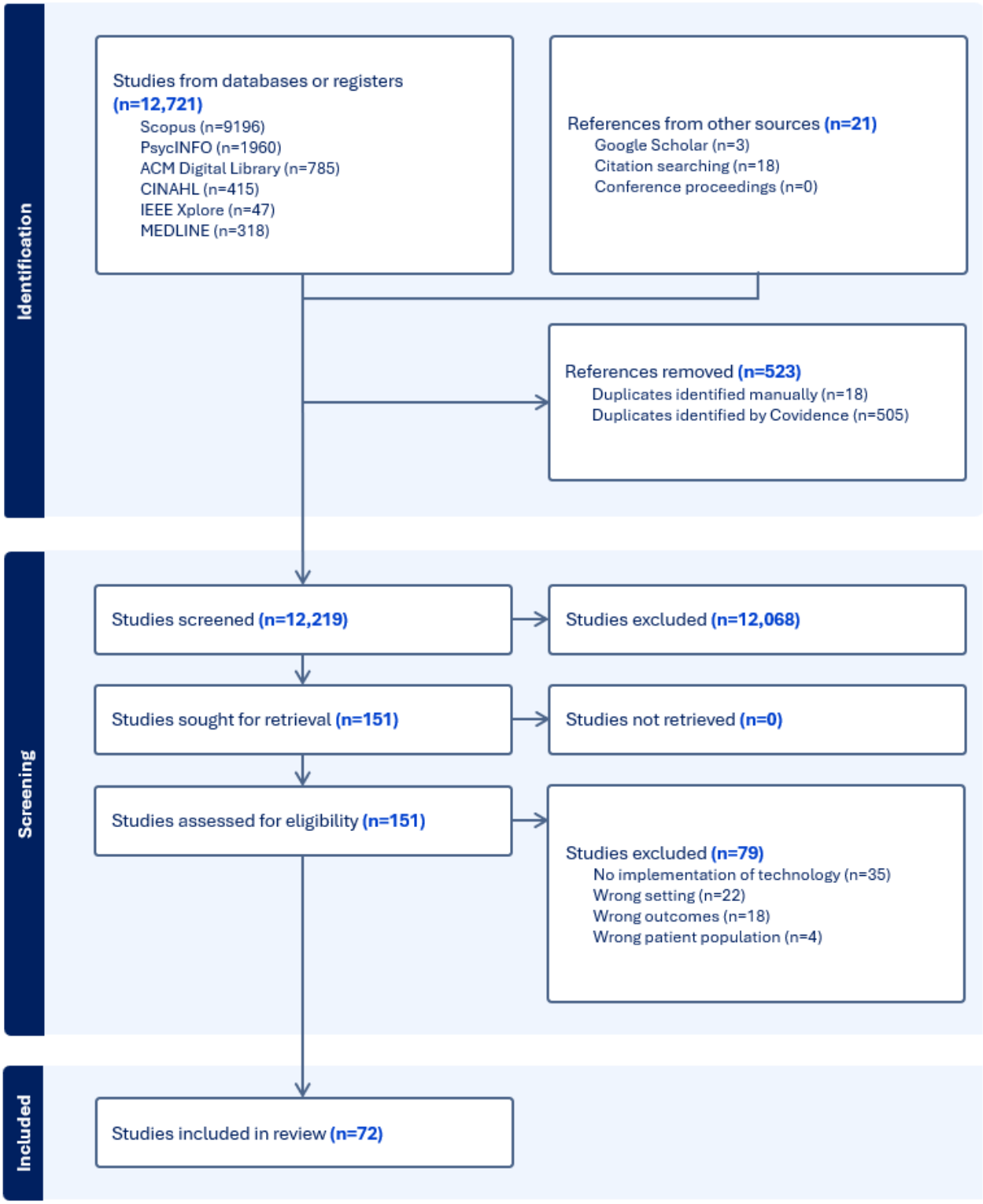

In total, 151 full-text studies were assessed for eligibility, and 79 were excluded at this point. Reasons for exclusion included no digital technologies being implemented in the paper (35/79, 44%), papers focused on clinical outcomes instead of meaningful connection (18/79, 23%), research was undertaken in the wrong setting (22/79, 28%), or papers included a range of settings or populations and did not have a separate analysis available for the focus of the review (4/79, 5%). In total, 72 studies were

XSL•FO
RenderX

identified for inclusion in the final review (Multimedia Appendix 3).

Characteristics of Included Studies

The first study that reported on digital technologies to support meaningful connections in care homes was published in 2002; however, more than half of the studies were published since 2022 (38/72, 53%; Table 1). While older studies focused more on film screenings and videophones, virtual reality (VR) and digital apps were the most popular digital technologies in 2024-2025, showing a change over time in the types of digital technologies explored.

Table 1. Year of publication.

Year	Values, n (%)
2002	1 (1)
2003	1 (1)
2009	1 (1)
2010	1 (1)
2013	1 (1)
2014	1 (1)
2016	2 (3)
2017	1 (1)
2018	3 (4)
2019	4 (6)
2020	9 (13)
2021	10 (14)
2022	11 (15)
2023	7 (10)
2024	14 (19)
2025	5 (7)

In total, 15 studies were undertaken in Australia, 12 studies in the Netherlands, 8 studies in Canada, 5 studies in Norway and the United Kingdom, respectively, 4 studies in the United States, 3 studies across multiple countries, 3 studies in Germany, Italy, and Taiwan, respectively, 2 studies in Denmark, Finland, and Sweden, respectively, and 1 study in Egypt, Ireland, Japan, Jordan, and Switzerland, respectively. Study settings were identified as care homes (17/72, 24%), residential homes (17/72, 24%), nursing homes (13/72, 18%), long-term care settings (16/72, 22%), and other care facilities including “aged care residences” (9/72, 13%). Participants included care home residents (47 studies, 1432 residents), professional caregivers (30 studies, 811 people), and friends and family (18 studies, 407 people) primarily. Some studies included other key stakeholders alongside residents, families, and staff, including high school students, expert older adults, and volunteers. In total, 17 studies specifically included patients with dementia in their sample, including a total of 269 residents with a dementia diagnosis.

Types of Digital Technologies

Types of digital technologies used in care homes to support meaningful connections were grouped into 5 main categories: mobile and tablet apps, VR and immersive technology, robotics, digital devices and ambient technology, and online or computer-assisted programs (Table 2).

XSL•FO
RenderX

Table 2. Summary of intervention types and functions.

Intervention and function	Examples
Mobile and tablet apps (n=18)	
Communication	• Ticket To Talk app • iPad communication app (unnamed) • Video calling (n=4; eg, FaceTime) • Teleconferencing (n=2)
Reminiscence	• MemoryKeeper app • StoryTiling app
Social activity	• Connecting Today app (remote visiting) • A Better Visit app (shared activity) • Cognitive stimulation therapy app (n=2) • Virtual visits to places of personal significance (n=2) • iBeacons and audio server with iOS applications (virtual museum visit; n=2)
Devices and ambient technology (n=21)	
Wearables and sensory stimulation	• MEMENTO wearable reminiscence technology • Hepatic devices during video calls (warm hand, vibrating music strap, weighted collar, warm hug robot, Hey bracelet) • Interactive sound cushion “VITA” (n=3) • Alternative controller musical instrument
Creative engagement	• VENSTER interactive art (n=2) • ABBY interactive touchscreen • AmbientEcho interactive media system • Vodafone smart photo frame
Social activity	• Computer kiosks • Digital television for reminiscence • Interactive table • Motitech bikes • KOMPa “one button computer connecting generations” (n=4) • SAMb companions
Smart home and voice-controlled assistants	• Google Nest Hub Max • Google Home Assistant
Online and computer-assisted programs (n=8)	
Reminiscence	• Videotape (Remembering When) • Film screenings • Computer Interactive Reminiscence and Communication Aid
Social activity	• SitDance (interactive touchscreen with exercise program) • Virtual creative aging program (museum and music) • ArtOnTheBrain (museum)
Communication	• Online social networking • Internet-based videoconferencing
Virtual reality (VR) and immersive technology (n=10)	
Social activity	• VR treadmill • VR Take A Cycle With Me • VR cycling with pedal exercisers and GoPro video footage • VR Vacation HomeImmersive • 3D VR with hands-on horticultural therapy • VR companion Kiera
Reminiscence	• 3D virtual worlds (reminiscence room, virtual tours, gardening) • VR reminiscence with VR photo albums or videos • SENSE-Garden (n=2)
Robotics (n=14)	

XSL•FO
RenderX

Intervention and function	Examples
Social activity	• LiveNature (sheep) robotic social sheep (n=3) • Robotic pets (cat and dog; n=2) • Paro (baby harp seal; n=2) • Zora humanoid • Adam humanoid • Matilda • Pepper semihumanoid • Rebecca social robot • Zoomorphic cleaning robot • Robot-facilitated dance therapy

[a]KOMP: a communication tool developed by Norwegian company No Isolation, designed to bridge the gap between older adults and their digitally experienced families. It consists of a screen with one single button that turns it on and off and an app that is connected to the screen.

[b]SAM: an interactive companion that plays sounds and displays colors in response to touch-based inputs and can mimic or react to the input of a paired device to stimulate social engagement between care home residents.

Mobile apps were discussed in 18 studies [31-48]. The primary aim of these interventions was to increase communication (eg, Ticket to Talk, video calling, and FaceTime), support reminiscence (eg, MemoryKeeper and StoryTiling), increase access to shared activities (eg, A Better Visit, Virtual Museum visiting app paired with iBeacons and audio servers, cognitive stimulation therapy, and virtual visits), and deliver cognitive stimulation therapy. Digital devices and ambient-assisted living technologies were reported in 21 studies [49-69]. New devices included wearables and sensory stimulation such as MEMENTO wearable reminiscence technology and hepatic devices to wear during video calls, creative engagement devices such as VENSTER interactive artwork and AmbientECHO interactive media system, social activities such as Motitech bikes and interactive tables, and smart home and voice-controlled assistants such as Google Nest Hub Max. Online and computer-assisted programs were reported in 8 studies [70-77]. These programs supported reminiscence such as videotapes, film screenings, and computer-interactive reminiscence-specific programs, social activity such as interactive exercise videos (SitDance), creative aging programs (museum visits and music), and communication such as online social networking. VR and immersive experiences were reported in 10 studies [78-87], including social activities such as VR treadmills, VR cycling (Take a Cycle with Me), VR Vacation Home, a VR companion (Kiera), an immersive personalized sensory experience (SENSE-Garden), and reminiscence activities such as 3D virtual words with a reminiscence room and VR reminiscence through virtual photo albums and videos. The application of robotics in care homes was reported in 14 studies [88-101], including robotic animals (sheep, cats, dogs, and seals), humanoids and semihumanoids (Adam, Matilda, Pepper, Rececca, and Zora), a zoomorphic cleaning robot with a social role, and robot-facilitated dance therapy. One study reported on the application of multiple digital technologies including digital music therapy, mobile apps, video games, videoconference, social robots, VR, and other technologies [102].

Outcomes to Measure Meaningful Connection

In total, 38 studies primarily used qualitative analysis techniques such as thematic analysis to assess outcomes, and another 12 studies combined qualitative analysis techniques with quantitative analysis. Observation was also a commonly used technique to observe changes in meaningful connections such as observing verbal and nonverbal responses, enjoyment, ease of responding to stimuli, and eye contact [35,72,88,99]. Engagement, well-being or satisfaction, emotional response, quality of life, purpose and meaning, social closeness, loneliness, depression and anxiety, and psychosocial capacity were the main themes identified from the scales used to indicate whether meaningful connections were present (Table 3). Studies also considered technology acceptance and clinical outcomes such as cognitive function, activities of daily living, and neuropsychiatric symptoms.

XSL•FO
RenderX

Table 3. Summary of outcome measurement instruments.

Instrument	Studies	Focus	Commentary on the tool from the authors
Engagement			
Engagement of a Person With Dementia Scale	[94]	Evaluates user engagement through activity participation (affective, visual, verbal, behavioral, and social).	Multiple raters can be a strength; however, when staff are unfamiliar with using the tool, and different staff complete the tool due to shift changes, this may lead to inconsistencies in the results.
Observational Measurement of Engagement	[88,94,95]	Direct observation measure of activity engagement for persons with dementia (attention and attitude).	Rated based on direct observations on-site by 2 different raters.
Music in Dementia Assessment Scale (adapted)	[70]	Trained raters observe and quantify behaviors (interest, response, initiation, involvement, and enjoyment).	No comment.
Engagement in Meaningful Activities Survey	[70]	Measures the impressions of the meaningfulness of daily activities.	Some of the items focus on domains typically impacted by aging and may influence responses.
Video Coding Incorporating Observed Emotion	[51]	Overview of emotional responses through 6 dimensions of engagement (emotional, verbal, visual, behavioral, collective engagement, and agitation).	A second coder is required to enhance the reliability and validity of the analysis. Automatic interpretation of affective facial expression and qualitative measures of connectedness and social inclusion should be included.
Ethographic and Laban-Inspired Coding System of Engagement	[88]	Measures engagement of a person with dementia through observational behaviors (head, torso, and hands or arms).	Instructions are needed to ensure consistency in the measurement procedures (eg, reduce personal conversations irrelevant to the study).
Visual Focus of Attention	[103]	OpenVINO used to autonomously track 2D positions of users' eyes, ears, and nose.	No comment.
Well-being and satisfaction			
Warwick-Edinburgh Mental Well-Being Scale	[59]	Unidimensional self-report instrument (14 items) assessing well-being administered either in person or via telephone.	Previous studies demonstrated the feasibility of the tool's use with vulnerable populations.
World Health Organization-Five Well-Being Index	[81]	Self-report measure of mental well-being.	Previous studies demonstrated validity in screening for depressive mood.
Satisfaction With Life Scale	[75]	Self-report 5-item scale assessing global satisfaction with life.	No comment.
Ryff's Psychological Well-Being Scale	[85]	Six dimensions—autonomy, environmental mastery, personal development, supportive relationships, purpose in life, and acceptance of oneself.	No comment.
Emotional response			
Observed Emotional Rating Scale	[94]	Measures the extent of emotional expressions (pleasure, anger, anxiety, sadness, and general alertness).	Rated based on direct observations on-site by 2 different raters. Evidence supports interrater reliability for each state and concurrent and discriminant validity for use by trained raters.
Four items measuring enjoyment from the Game Experience Questionnaire	[40]	Self-report 4 items measuring enjoyment in the experience.	No comment.
People Environment Apathy Scale—Apathy subscale	[94]	Four-item Likert scale assessing apathy-related behaviors through facial expression, eye contact, physical engagement, purposeful activity, verbal tone, and verbal expression.	Valuable tool to assess the extent of intrinsic motivation to behave despite being positively or negatively engaged.
Quality of life			

XSL•FO
RenderX

Instrument	Studies	Focus	Commentary on the tool from the authors
EuroQOL[a] EQ-5D-5L	[80]	A 2-page instrument with a descriptive system and visual analog scale, comprising 5 dimensions (mobility, self-care, usual activities, pain or discomfort, and anxiety or depression).	Valid measure of health-related quality of life in people living in residential care.
Multicultural Quality of Life Index	[70]	Captures subjective quality of life over 10 dimensions such as physical well-being and spiritual fulfillment.	Two domains did not apply to long-term care residents, so were removed.
Quality of Life in Late Stage Dementia scale	[70]	Rates observed resident behaviors on 11 items such as crying or appearing uncomfortable over the last week.	No comment.
DEMQOL[b] Proxy instrument	[38]	A 31-item scale completed by proxy to measure quality of life—social relationships items selected (getting in touch with people, having enough company, and being able to help other people).	Evidence supports the DEMQOL Proxy's internal consistency and test-retest reliability when used with people living with mild to moderate and severe dementia, and its content, convergent, and discriminate validity.
Quality of Life—Alzheimer's Disease	[67]	A 13-item questionnaire used to provide a self-reported indication of quality of life.	Multiple assessors may have introduced sources of variability.
Purpose and meaning			
Chinese Health Questionnaire	[86]	A 12-item scale measuring both participant health and mattering (the degree of being valued by significant others), meaning of life, loneliness, and depression.	Valid tool for use in older adults.
Purpose in Life Survey	[86]	Nine items (Likert scales).	Cronbach α coefficient was 0.87 in this study.
Social closeness			
Patient Reported Outcomes Measurement Information System Social Isolation Short Form 4a	[70]	Self-report 4-item scale captures social relationships.	Some of the items focus on domains typically impacted by institutional residence and may influence responses.
Inclusion of Other in Self Scale	[39,40]	Single-item graphical measure that shows 2 circles for self and others at various levels of distance until they substantially overlap. Participants indicate which one best represents the perceived interaction with the on-site companions.	No comment.
Quality of the Caregiving Relationship Questionnaire	[67]	Fourteen items with two subscales, warmth and absence of conflict, measuring quality of relationship between participant and relative.	Ceiling effect noted.
Social Support Behavior Scale	[77]	Three subscales—social support network, quantity of social support, and satisfaction with social support.	No comment.
Lubben Social Networking Scale	[63]	Self-report measure of the size of social networks with friends and family.	Recommended measure for assessing social networks among older adults.
Loneliness			
UCLA[c] Loneliness Scale	[77,80,86,96]	Self-report 3-item scale measuring loneliness.	Previous studies have shown this to be a reliable and valid measure of loneliness.
Agitation			

XSL•FO
RenderX

Instrument	Studies	Focus	Commentary on the tool from the authors
Cohen-Mansfield's Agitation Scale	[86]	Eight problematic behaviors were coded for frequency of occurrence on a scale of 0=did not occur to 3=occurred almost constantly. These included leaving one's seat, cursing or verbal aggression, requests for attention, asking repetitive questions or sentences, screaming, shouting or making strange noises, complaining, engaging in physical aggression to self or others, and exhibiting general restlessness.	Interrater reliability was established at >90% agreement among 3 independent coders. Mood prior to the study was observed to have an impact on agitation observed during the study, potentially skewing the results. Baseline agitation levels should be collected as a baseline to allow for comparison.
Depression and anxiety			
Center for Epidemiologic Studies Depression Scale	[75]	Self-report 8-item scale measuring depressive symptoms in the last week.	No comment.
Cornell Scale of Depression in Dementia	[63]	Clinician rated 19-item scale indexing mood and related signs, behavior disturbance, cyclic function, ideational disturbance, and physical signs of depression.	No comment.
Geriatric Depression Scale	[63,77,96]	Thirty items with a yes or no response.	The validity of the scale decreases when older adults have cognitive impairment.
Geriatric Depression Scale—Residential	[67]	A 12-item depression scale relevant to nursing and residential home populations.	No comment.
Chinese Geriatric Depression Scale	[86]	A 15-item scale measuring feelings of depression over the prior week.	Cronbach α coefficient was 0.90 in this study.
Geriatric Anxiety Inventory	[63]	A 29-item scale with yes or no responses relating to anxiety over the prior week.	Strong utility in detecting anxiety disorders.
Psychosocial capacity			
Mini-ICF[d] Rating for Impairment in psychological activities and capacities	[81]	Assesses psychosocial capacity in nursing homes through (1) adherence to regulations, (2) planning and structuring of tasks, (3) flexibility, (4) competence and knowledge application, (5) capacity to make decisions and judgments, (6) proactivity and spontaneous activities, (7) endurance, (8) self-assertiveness, (9) contact with others, (10) group integration, (11) intimate relationships, (12) self-care, and (13) mobility.	The tool is recommended in social medicine guidelines. Proven to be manageable, reliable, and valid in clinical practice.

[a]EuroQOL: EuroQOL consists of a Research Foundation and Group Association, which provides tools for measuring and valuing health that aim to improve decisions about health and health care globally.

[b]DEMQOL: Dementia Quality of Life is a patient-reported outcome measure, which is designed to enable the assessment of health-related quality of life of people with dementia.

[c]UCLA: University of California, Los Angeles.

[d]ICF: International Classification of Functioning, Disability and Health.

Engagement was measured using the Engagement of a Person With Dementia Scale [94], Observational Measurement of Engagement [94], Music in Dementia Assessment Scale [70], Engagement in Meaningful Activities Survey [70], Video Coding Incorporating Observed Emotion protocol [51], the Ethographic and Laban-Inspired Coding System of Engagement [88], and Visual Focus of Attention [103]. Well-being and satisfaction were measured through the Warwick-Edinburgh Mental Well-Being Scale [59], World Health Organization-Five Well-Being Index [81], the Satisfaction With Life Scale [75], and Ryff's Psychological Well-Being Scale [85]. Emotional response was assessed through the Observed Emotional Rating Scale [38,88], 4 items measuring enjoyment from the Game Experience Questionnaire [39,40], and the People Environment Apathy Scale [94]. Quality of life was measured using the EuroQOL EQ-5D-5 L [80], Multicultural Quality of Life Index [70], Quality of Life in Late Stage Dementia scale [70], the Dementia Quality of Life Proxy instrument [38], and the Quality of Life—Alzheimer's Disease scale [67]. Purpose and meaning were measured by the Chinese Health Questionnaire [86] and the Purpose in Life Survey [86]. Social closeness was measured using the Patient Reported Outcomes Measurement Information System Social Isolation Short Form 4a [70], the Inclusion of Other in Self Scale [39,40], the Quality of the Caregiving Relationship Questionnaire [67], the Social Support Behavior Scale [77], and the Lubben Social Networking Scale [63]. Loneliness was measured using the University of California, Los Angeles 3-Item Loneliness Scale [80]. Agitation was assessed through the Cohen-Mansfield's Agitation Scale [51,72]. Depression was measured using the Center for Epidemiologic

XSL•FO
RenderX

Studies Depression Scale [75], the Cornell Scale of Depression in Dementia [63], the Geriatric Depression Scale [63,77,96], the Geriatric Depression Scale—Residential [67], and the Chinese Geriatric Depression Scale [86]. Anxiety was measured using the Geriatric Anxiety Inventory [63]. Psychosocial capacity was measured using the Mini-ICF Rating for Limitations of Activities and Participation in Psychological Disorders [81], which includes subscales assessing group integration and intimate relationships.

Instruments included a mixture of self-report scales and observation of behavior. Observational tools were more commonly seen when residents also had a diagnosis of dementia. While some authors did not justify their choice of instruments, others provided justification through the provision of evidence on their validity and reliability in similar contexts. Challenges to the use of the tools were also recorded, including potential inconsistencies between different raters [35,67], the impact of the long-term care setting on the accuracy of the results [70], and the conceptual gaps around interpreting affective facial expressions, connectedness, and social inclusion [51].

Factors That Support Meaningful Connection

Overview

A narrative summary was developed to explore how the included studies reported the factors underpinning meaningful connections during the use of digital technology. Five key themes were identified: getting to know the person, increases autonomy and choice, source of enjoyment and fun, facilitates communication, and builds community. Table 4 identifies the studies associated with each theme.

Table 4. Summary of papers associated with factors that support meaningful connections.

Theme and subtheme	Papers
Getting to know the person	
Reminiscence or life story work	[33-36,41,47,49,53,55,57,67,70,71,75,77,78,81-84,87,88,93,94]
Seeing the patient as a person	[36-38,43,46,55,79,84,86,90]
Increases autonomy and choice	
Autonomy and control	[34,52,56,64,70]
Personal choice valued	[34,48,58,74,79,81,84,104]
Source of enjoyment and fun	
Technology was fun	[39-41,52,53,55,56,59,64,65,70,73,74,81,88,89,91,92,104]
Space for creativity	[34,56,70,73,79]
Facilitates communication	
Communicating with a facilitator or HCP[a]	[31,37,46,50,52-54,56-58,63,68,81,88,90,93,96]
Communicating with family	[32,33,36,38,41,43-47,49,50,52,60-62,67,75-78,93]
Builds community	
Shared social activities within the care home	[39,40,51,52,56,63,64,66,71,74,76,78,81,87,89,93,99-103]
Intergenerational and wider community connection	[39,40,42,49,50,71,74,75,80,104]

[a]HCP: health care professional.

Getting to Know the Person

Reminiscence and life story work were identified as a key mechanism through which meaningful connections were formed by providing an opportunity to get to know the person as an individual without focusing on their medical condition. Reminiscence was enabled through a range of activities from virtual cycling by reminiscing on experiences of cycling earlier in life [83] to VR photo albums and videos [85]. Technology supported less pressured nonverbal interactions, allowing personal preferences to emerge, helping practitioners to get to know individuals more effectively [54], and offering multiple channels through which people could express their identity [52]. Shifting from patient to person was seen not only in reminiscence-specific technologies such as MEMENTO, a spontaneous reminiscence therapy smartwatch [58], but also in the use of robotics such as the social sheep [88,94,95] and in the development of personalized multimedia spaces through life story work such as SENSE-Garden [79]. Residents enjoyed when activities led to an increased feeling of value and a sense of being treated seriously, for example, when combining 3D VR with hands-on horticultural therapy [86]. Getting to know the individual resident led to a change in the care culture, for example, gaining deeper insights into residents' life stories to facilitate SENSE-Garden fostered a more profound personal connection and shift in focus to a person-centered approach.

Increases Autonomy and Choice

Digital technologies provided an opportunity for residents to feel in control, empowered, and independent while having their autonomy and personal choice respected, creating spaces for meaningful connections to form. Many of the technologies, for example, virtual talks [34], SENSE-Garden [79], VR Vacation Home [81], and Virtual Creative Ageing Program [70], all promoted a sense of control and autonomy over the activity, which was reported as a key element that promoted meaningful

XSL•FO
RenderX

connections. Computer kiosks provided an opportunity for personal development while being in control, which helped to increase residents' feelings of self-worth and esteem as well as connecting socially [64]. Lifestyle enrichment was also noted in the study by Webber et al [48], which used technology to facilitate virtual visits to places of personal significance. Residents could also meaningfully connect with their families without needing staff as gatekeepers [50]. Promoting ability instead of focusing on limitations created positive experiences for residents and staff.

Source of Enjoyment and Fun

To build meaningful connections, this theme identified the importance of digital technologies as a source of enjoyment and fun for the users. Digital technology was described as playful [53,55,59,89], pleasurable [65], fun [39,40], and enjoyable [52,88,103]. The digital technology created a space to express creativity [70], engage imagination [56], socialize [43], and break routine [34], which relieved stress and boredom for residents. Robotics in particular was identified as a source of shared joy, such as Pepper, the social semihumanoid robot [91], and Paro, the baby harp seal [92].

Facilitates Communication

Digital technologies were often implemented with facilitation from a health care professional or study-specific facilitator. The process of engaging with the facilitator was identified as a key opportunity to connect meaningfully, for example, during the delivery of digital cognitive stimulation therapy [31] or the introduction of the interactive sound cushion VITA [53]. Spending more time with staff while engaging with technology was reported positively by residents [68]. Technology supported meaningful communication with the facilitator, enabling trust and reducing misunderstandings [37]. Robotics such as pet cats and dogs provided something positive to talk about between the residents and health care professionals, opening up a channel of communication [96]. In addition to increasing communication with a facilitator, some technologies increased communication with families, for example, teleconferencing [44,45] and video calling relatives [33,41,47,77]. Apps such as Memory Keeper acted as a conversation starter [36], and new devices, such as a digital television for reminiscence [67], VENSTER interactive artwork, encouraged conversation about a present activity [57]. KOMP supported the maintenance of social bonds through family interaction [60,61], and a digital photo frame improved social connectiveness with family [62].

Builds Community

Opportunities were created to engage in a social activity, which supported residents to find others with shared interests and develop a sense of community. Meaningful social relationships and small group interactions were enabled through the use of online social networking in the study by Ysseldyk et al [75] and the use of a Computer Interactive Reminiscence and Communication Aid in the study by Samuelsson et al [76]. The introduction of new devices such as an interactive table [51], PARO the baby harp seal [100], virtual visits to museums [39,40], technology-assisted music making [66], and shared activities such as SitDance [74] encouraged interaction between residents, staff, and relatives through a shared social activity, reducing loneliness. VR Take a Cycle With Me supported intergenerational rapport and cohesiveness, which fed into an increased sense of community [80], and adaptive Motitech bikes provided a purpose and challenge while building a sense of community with other residents in an inclusive way [63]. Technology improved participation for people living with dementia, with robotic cats and dogs promoting an inclusive community for all abilities in care [90], zoomorphic robots initiating conversations [99], robot-led activities improving engagement [101], and robot-facilitated dance sessions improving group valence [103].

Discussion

What This Study Adds

This scoping review explored the findings from 72 papers that identified the active digital technologies used in care homes to help build meaningful connections, as well as the factors that underpin the formation of meaningful connections and the range of outcome measures used to assess their impact. Based on our search of CINAHL, MEDLINE, and *JBI Evidence Synthesis* up to December 12, 2025, we did not identify any previous scoping review that looks at meaningful connections across the care home, without focusing on a specific condition or clinical outcome. Five main types of digital technology were identified, including mobile apps, VR and immersive technology, robotics, digital devices and ambient technology, and online or computer-assisted programs. A narrative summary described the factors through which meaningful connections were developed, and these included getting to know the person, increases sense of autonomy and choice, source of enjoyment and fun, facilitates communication, and builds community. While most studies relied on some form of qualitative analysis or observation to identify changes in meaningful connections, a wide variety of standardized instruments were also used, which looked for changes in engagement, well-being or satisfaction, emotional response, quality of life, purpose and meaning, social closeness, loneliness, depression and anxiety, and psychosocial capacity.

Comparison of Findings With Existing Literature

The findings from this scoping review concluded that meaningful connections were facilitated through the creation of opportunities for the residents to be known and heard as a person, and the development of group and shared spaces where people could share an enjoyable experience and an increased sense of community. Effective digital technologies amplified human interactions instead of replacing them, enabling increased interaction with family, friends, and community outside of the care home, shared activities and hobbies with other residents in the care home, and opportunities for staff and facilitators to get to know the resident on a deeper level. These findings align with a community-based implementation project of social robots with older adults, which found that social robots facilitated person-person interactions and inspired participants to share experiences with their family and friends after the engagement session [105]. Fostering and maintaining meaningful connections is at the core of social health, which is defined as an adequate

XSL•FO
RenderX

quantity and quality of relationships in a particular context to meet an individual's need for meaningful human connection [3]. These findings align with wider research into the impact of digital technology on patient experience, which reported improvements in patient experience and improved quality of care [106].

While digital technology plays a key role in addressing social health by acting as a catalyst for forming meaningful human connections, caution is also due when considering the potentially harmful effects of overreliance on artificial intelligence companions that could replace social human interaction [107]. Warm technology, which aims to enhance human potential and social connectedness [108], has been developed primarily within the field of dementia care, but could be further developed and tested for application across the wider care home population. Furthermore, focus should be given to ensuring that both staff and residents are familiar with and comfortable using such technologies [109]. However, it is also important to remember that while many of the technologies included a social element, not all meaningful connections were inherently social. Some technologies such as the MEMENTO smartwatch [58] encouraged companionship through spontaneous meaningful conversations and enhanced staff access to residents' stories through reminiscence. SENSE-Garden created space for interactions between staff and residents, during which they engaged with life stories, focused on abilities more than limitations, and supported a culture shift from patient to person [79]. This highlights the fundamental importance of technology as a facilitator of individual meaningful connection, as well as a facilitator of social connection and community. While some studies support the integration of technology-driven solutions into caregiving procedures to promote social health among nursing home residents with dementia [102], it is also important to consider the potential practical, clinical, and ethical limitations of the introduction of such technologies. From a practical perspective, research has shown that the use of technology for social connection in aged care requires careful facilitation from staff, better resourcing and infrastructural support, collaborations with volunteers, and more attention to individual needs [110]. Clinically guided, emotional warmth, complex social interactions, and empathy are essential in the care of older adults [111]. This may be reduced or removed if the care sector becomes overreliant on technology to meet social health needs. Ethically, the use of social robots should be guided by the same ethical guidelines followed in all therapeutic interventions [112]; however, it was noted in a recent qualitative analysis of research into robotics in health care of older adults that limited attention is given to ethical and legal aspects of care robots [113].

The application of such active digital technologies to social health sits in contrast to the use of passive technologies, which improve safety and efficiency but may inadvertently reinforce a surveillance-oriented model and add little to the social well-being of residents in care homes [114]. A scoping review of digital technologies supporting aged care services more broadly found 7 key roles of digital technology, including health monitoring and assessment, remote health care services, assistive technology to support treatment, self-care management, social technology to facilitate interaction, clinical decision support, and aged care quality measurement [115]; however, only social robots to provide better communication were discussed. When looking at policies focused on digital health, the core messages around technology use in health care move toward clinical applications, such as diagnosis, decision-making, virtual care, remote monitoring, remote data capture, and clinical trials within the Global Strategy on Digital Health 2020-2025 [116]. The focus is once again on more passive clinical monitoring than the application of active digital technology to social well-being. This trend continues in new research on generative artificial intelligence (GenAI), which, despite the capacity to address social health concerns, is currently focused on increasing task efficiency such as scheduling and care planning [117]. Alternatively, active technologies have a greater potential to promote engagement and connection [118,119]. While their effectiveness is often dependent on family or professional caregivers to help set up and use the technology to ensure meaningful use rather than superficial engagement [120], the involvement of caregivers is, in itself, a mechanism through which more meaningful connections can be achieved.

The scoping review showed how the types of digital technologies used for meaningful connection have changed over time and have exponentially increased in the last 3 years. Smart Home technologies have been shown to significantly enhance safety and monitoring capabilities for adults living independently [121]; however, this overlooks the social health of older adults. A current area for future exploration is the application of GenAI in care homes to promote meaningful connections. GenAI, while originally appearing in physical form through the chatbot Eliza in the 1960s, has recently boomed in popularity in the public sphere through chatbots such as ChatGPT, and more recently again through its application to health care [122]. Currently, the primary application in clinical practice is to support health care, planning, and information searching [117]. For example, one study successfully applied GenAI to reduce emergency incidents and help daily activities run more smoothly within assisted living environments [123]. Another study explored how GenAI models can improve decision accuracy among emergency nurses; however, they noted the lack of contextual sensitivity in emergency care [124]. Within nursing practice, GenAI is being embedded in patient care, supporting workflow integration, and providing operational support [125]. In gerontological nursing education, GenAI has been integrated with art therapy to support the development of empathy in students [125]. In community care, artificial intelligence–powered reflection tools have been shown to improve communication between older adults and caregivers [126], to support older adults with health information tasks [127,128], and to provide empathetic responses to online queries related to Alzheimer disease and related dementias [129].

However, GenAI with the goal of building meaningful connections has received little attention, specifically in the long-term care setting, which has additional barriers to meaningful connections. The HARMONEE (Harnessing Artificial Intelligence Resources for Mental Well-Being for Older Adults and Nurturing Empathy in Education) project is a novel example of how GenAI can be used to support

XSL•FO
RenderX

reminiscence with older adults in long-term care to improve their mood [119]; however, its dependence on a research team to facilitate generates issues with the sustainability. Technology must be designed so it can be used by the resident or care team independently. Additionally, to ensure that technology is fit for purpose and suited to the unique setting of the care home, care home residents must be part of decision-making when it comes to co-designing and implementing new technologies in order to defy techno-solutionist scripting of dehumanizing images of fragility, dependency, and technical ineptitude embedded in product designs [130]. Research has indicated an openness of older adults to new technologies for physical assistive functions [131]; however, further exploration is needed to understand their attitudes toward technologies aimed at supporting meaningful connections.

This review identified that the measurement of meaningful connection was inconsistent across studies. This may be attributable to the varying operational definitions of meaningful connections, activities, and engagement across these studies. Qualitative analysis and observation were most frequently used to assess concepts associated with meaningful connections. Studies that used scales and instruments considered a wide variety of measures including engagement, well-being or satisfaction, emotional response, quality of life, purpose and meaning, social closeness, loneliness, depression and anxiety, and psychosocial capacity. This highlights a broader challenge around defining and measuring meaningful connections in the literature. It could be argued that meaningful connections in themselves are not measured specifically, but instead outcome measures focused on indicators of meaningful connections (eg, well-being, loneliness, and quality of life). When exploring social connection, Holt-Lunstad [2] discusses similar terminological challenges; for example, loneliness is often used to measure social connection; however, it is defined as a distressing experience resulting from perceived inadequate meaningful connections; therefore, it does not fully capture the breadth of social connection. Similarly with meaningful connections, a clear consensus has not been reached on a definition or approach to measurement. Measurement of social connection is being progressed, for example, through the development of the Connection During Conversations Scale [132]; however, with the focus being on social connection during social interactions, it is possible that it overlooks some of the meaningful connections occurring outside social spaces.

Strengths and Limitations

By adopting a scoping review approach, a comprehensive overview of the broad field of digital technology within care homes could be provided, helping to explore the scope of existing knowledge and identify gaps in the evidence base to highlight for future research. The scoping review was conducted with strict adherence to the JBI methodology and included multiple authors at each stage to minimize the risk of individual bias. However, limitations are also noted. While the language of meaningful connections was left as broad as possible within the search strategy, it is possible that papers with relevant information around social connection or person-centered care delivery were missed. Given the time and resource restraints within the project, the decision was made to limit the search to the English language. However, language bias may result in missing important cultural contexts in non-English–speaking settings, limiting the relevance of findings in some geographical locations. The main focus of this review was on active digital technologies, excluding passive monitoring technologies. However, these technologies contribute to care home life through their influence on care practices, staff time and availability, and safety perceptions that directly impact opportunities for connection. It is possible that passive monitoring technologies could increase staff availability and create a safer care environment, which would lead to increased opportunities for meaningful connections. Through their omission from the review, only a partial understanding of the digital landscape within care homes has been explored. Finally, the heterogeneity of outcome measures reported across the studies limits the comparability of studies and restricts the synthesis of these findings. Clear conclusions regarding the measurement of meaningful connections could not be drawn; however, this did highlight a gap for future research.

Recommendations and Real-World Implications

At present, technologies used within aged care seem to not be meeting end user needs [133]. While technology design for older adults is often focused for family members or caregivers to allow passive monitoring, giving control back to older adults nurtures their sense of autonomy [134]. Future technologies must also be designed with social health in mind, considering both the engagement of the individual and the community. This review has highlighted how technology can facilitate both individual meaningful connections as well as social connections and a sense of community. However, to ensure technology is fit for purpose and suited to the unique setting of the care home, care home residents must be part of decision-making when it comes to co-designing and implementing new technologies in order to defy the dehumanizing images of fragility, dependency, and technical ineptitude embedded in product designs [130]. Furthermore, future research needs to consider the cultural environment of the care home when embedding new active digital technologies with the goal of building meaningful connections, as research has shown that how the cultural environment has the most significant influence on nursing home residents' leisure and social occupational choice [135].

Future research needs to consider the development and psychometric testing of a standardized tool to measure meaningful connection more broadly, as well as in relation to the digital technologies implemented. Consideration needs to be given to both individual and social connection indicators, as well as emotional response, engagement, satisfaction, loneliness, anxiety, depression, and quality of life. There is also a gap specifically around meaningful connections within care homes. The challenge raised with some of the instruments used to collect outcome measures was the uniqueness of the care home setting. Challenges in fostering meaningful engagement in care homes include a lack of understanding of residents' values and interests, staff needing to manage multiple requirements simultaneously, and staff not receiving support to involve residents [136]. Some of the items in the scales were not reflective of, or relevant to, long-term care settings, leading to the need for modifications or the acknowledgment that this may

XSL•FO
RenderX

have impacted on the findings of the study. Future research needs to consider the development of a bespoke instrument that considers meaningful connections within the context of the care home environment.

Research in this space is rapidly expanding, aligned with advances in digital technology availability. However, few studies that focused on meaningful connections considered an implementation design. The reliance on qualitative and observational methods may link to the challenges in measurement or the infancy of the technology; however, future research needs to consider the development of implementation studies, considering broader factors such as cost-effectiveness and organizational readiness. The TIDieR (Template for Intervention Description and Replication) checklist [137] may structure the reporting of pre- and postinterventions from other settings, allowing them to be replicated in the care home context.

Conclusions

This scoping review has presented a detailed exploration of how digital technologies have been used within care homes to facilitate the development of meaningful connections between residents, families, staff, and communities. While the wider literature indicates that many digital technologies overlook social health and well-being in care homes, this review highlights how a range of technologies including mobile apps, VR and immersive technology, robotics, digital devices and ambient technology, and online or computer-assisted programs can be applied to facilitate meaningful connections within care homes. Digital technology is a catalyst for meaningful human connection, which is a key element of social health, rather than a replacement. Further research is needed to explore how digital technologies can support individual meaningful connections with residents as well as enhance the sense of community through shared social activities and spaces. The application of GenAI to the development of meaningful connections and social health in care homes is a current gap in the evidence base. Moving forward, the development of GenAI-based technologies needs to be undertaken in conjunction with care home residents, staff, and communities to ensure that they meet the social needs of the entire community. An ongoing challenge is found in the measurement of the effectiveness of digital technologies to enhance meaningful connections. Future research needs to consider how new evaluation metrics can be developed that combine digital engagement data with validated measures of social and meaningful connection to assess the person-centered impact of digital technologies in care homes.

Acknowledgments

The authors thank the participants of the My Home Life Northern Ireland Leadership Support Programme who helped identify the gap that led to this scoping review and the librarian Kelly Coogan who supported the team during database searching. This manuscript has been prepared without assistance from generative artificial intelligence tools.

Funding

No funding was received to undertake this review. The paper processing fee was paid through the Ulster-JMIR Open Access deal.

Data Availability

All data generated and analyzed during this study are included in this published paper and Multimedia Appendix 3.

Authors' Contributions

Conceptualization: DM, RB
Data curation: DM, RB, KC
Formal analysis: DM (lead), RB (supporting), KC (supporting)
Methodology: DM
Project administration: DM
Supervision: DM
Validation: DM
Visualization: DM, RB
Writing—original draft: DM (lead), RB (supporting)
Writing—review and editing: DM (lead), RB (supporting), KC (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA ScR checklist.
[PDF File (Adobe PDF File), 172 KB - jmir_v28i1e88181_app1.pdf]

XSL•FO
RenderX

Multimedia Appendix 2
Search strings per database from search on December 12, 2025.
[DOCX File , 18 KB - jmir_v28i1e88181_app2.docx]

Multimedia Appendix 3
Data summary table.
[XLSX File (Microsoft Excel File), 35 KB - jmir_v28i1e88181_app3.xlsx]

References

1. Bethell J, Aelick K, Babineau J, Bretzlaff M, Edwards C, Gibson JL, et al. Social connection in long-term care homes: a scoping review of published research on the mental health impacts and potential strategies during COVID-19. J Am Med Dir Assoc 2021;22(2):228-237.e25 [FREE Full text] [doi: 10.1016/j.jamda.2020.11.025] [Medline: 33347846]
2. Holt-Lunstad J. Social connection as a critical factor for mental and physical health: evidence, trends, challenges, and future implications. World Psychiatry 2024 Oct;23(3):312-332 [FREE Full text] [doi: 10.1002/wps.21224] [Medline: 39279411]
3. From loneliness to social connection: charting a path to healthier societies. World Health Organization. 2025. URL: https://www.who.int/publications/i/item/978240112360 [accessed 2025-09-18]
4. Ageing and health. World Health Organization. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health [accessed 2025-09-19]
5. Ageing and long-term care. Organisation for Economic Co-operation and Development. 2024. URL: https://www.oecd.org/en/topics/sub-issues/ageing-and-long-term-care.html [accessed 2025-09-19]
6. Llena-Nozal A, Barszczewski J, Rauet-Tejeda J. How do countries compare in their design of long-term care provision? A typology of long-term care systems. OECD Health Working Papers. 2025 Jun 26. URL: https://www.oecd.org/en/publications/how-do-countries-compare-in-their-design-of-long-term-care-provision_44f5453a-en.html [accessed 2026-02-11]
7. Ettelt S, Williams L, Damant J, Perkins M, Wittenberg R. What kind of home is your care home? A typology of personalised care provided in residential and nursing homes. Ageing Soc 2020;42(5):993-1013. [doi: 10.1017/s0144686x20001142]
8. Boamah SA, Weldrick R, Lee TJ, Taylor N. Social isolation among older adults in long-term care: a scoping review. J Aging Health 2021;33(7-8):618-632 [FREE Full text] [doi: 10.1177/08982643211004174] [Medline: 33779366]
9. Swinton J, Vilapakkam Nagarajan S, Atee M, Morris T, Poulos CJ. Being present and meaningful engagement for aged care residents living with dementia: a mixed-methods evaluation of Australian care worker's experiences. Gerontologist 2024;64(4):gnad131 [FREE Full text] [doi: 10.1093/geront/gnad131] [Medline: 37777877]
10. QS50: Mental wellbeing of older people in care homes: quality standards. National Institute for Health and Care Excellence. 2013. URL: https://www.nice.org.uk/guidance/qs50/chapter/quality-statement-1-participation-in-meaningful-activity [accessed 2025-09-19]
11. Groenendaal M, Loor A, Trouw M, Achterberg WP, Caljouw MA. Perspectives of healthcare professionals on meaningful activities for persons with dementia in transition from home to a nursing home: an explorative study. Healthcare (Basel) 2019;7(3):98 [FREE Full text] [doi: 10.3390/healthcare7030098] [Medline: 31430855]
12. Hooker SA, Masters KS, Vagnini KM, Rush CL. Engaging in personally meaningful activities is associated with meaning salience and psychological well-being. J Posit Psychol 2019;15(6):821-831. [doi: 10.1080/17439760.2019.1651895]
13. Travers C, Brooks D, Hines S, O'Reilly M, McMaster M, He W, et al. Effectiveness of meaningful occupation interventions for people living with dementia in residential aged care: a systematic review. JBI Database System Rev Implement Rep 2016;14(12):163-225. [doi: 10.11124/JBISRIR-2016-003230] [Medline: 28009676]
14. Framework on integrated, people-centred health services. World Health Organization. 2016. URL: https://apps.who.int/gb/ebwha/pdf_files/WHA69/A69_39-en.pdf?ua=1%26ua=1 [accessed 2025-09-19]
15. Smith N, Towers AM, Palmer S, Beecham J, Welch E. Being occupied: supporting ‘meaningful activity’ in care homes for older people in England. Ageing Soc 2017;38(11):2218-2240. [doi: 10.1017/s0144686x17000678]
16. Kemp CL, Bender AA, Morgan JC, Burgess EO, Epps FR, Hill AM, et al. Understanding capacity and optimizing meaningful engagement among persons living with dementia. Dementia (London) 2023;22(4):854-874 [FREE Full text] [doi: 10.1177/14713012231162713] [Medline: 36913646]
17. Gonnord T, Clarys D, Boucard G, Esnard C. Positive impact of social relationships fostered by physical and/or cognitive group activity on older people's quality of life: PRISMA systematic review. Front Psychol 2023;14:1166072 [FREE Full text] [doi: 10.3389/fpsyg.2023.1166072] [Medline: 37767210]
18. Goodwin VA, Hall AJ, Bethel A, Taylor E, Tatnell L, Turner M, et al. Understanding joy amongst older people: a scoping review. Arch Gerontol Geriatr 2023;115:105118 [FREE Full text] [doi: 10.1016/j.archger.2023.105118] [Medline: 37516061]
19. Kang B, Scales K, McConnell ES, Song Y, Lepore M, Corazzini K. Nursing home residents' perspectives on their social relationships. J Clin Nurs 2020;29(7-8):1162-1174 [FREE Full text] [doi: 10.1111/jocn.15174] [Medline: 31889360]
20. Lim E, Nielsen N, Lapane L, Barooah A, Xu S, Qu S, et al. Health effects of social connectedness in older adults living in congregate long-term care settings: a systematic review of quantitative and qualitative evidence. Int J Older People Nurs 2023;18(6):e12577. [doi: 10.1111/opn.12577] [Medline: 37803996]

XSL•FO
RenderX

21. Townsend BG, Chen JTH, Wuthrich VM. Barriers and facilitators to social participation in older adults: a systematic literature review. Clin Gerontol 2021;44(4):359-380. [doi: 10.1080/07317115.2020.1863890] [Medline: 33393443]
22. Nygaard A, Halvorsrud L, Grov EK, Bergland A. 'What matters to you?'—a qualitative study on the views of nursing home residents with dementia regarding the health care they receive. J Clin Nurs 2022;31(1-2):262-274. [doi: 10.1111/jocn.15904] [Medline: 34096110]
23. Wright PJ, Raynor PA, Bowers D, Combs EM, Corbett CF, Hardy H, et al. Leveraging digital technology for social connectedness among adults with chronic conditions: a systematic review. Digit Health 2023;9:20552076231204746 [FREE Full text] [doi: 10.1177/20552076231204746] [Medline: 37799504]
24. Chu CH, Yee A, Stamatopoulos V. Poor and lost connections: essential family caregivers' experiences using technology with family living in long-term care homes during COVID-19. J Appl Gerontol 2022;41(6):1547-1556 [FREE Full text] [doi: 10.1177/07334648221081850] [Medline: 35416076]
25. Ibarra F, Baez M, Cernuzzi L, Casati F. A systematic review on technology-supported interventions to improve old-age social wellbeing: loneliness, social isolation, and connectedness. J Healthc Eng 2020;2020:2036842 [FREE Full text] [doi: 10.1155/2020/2036842] [Medline: 32765823]
26. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018;18(1):143 [FREE Full text] [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
27. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
28. Peters M, Godfrey C, McInerney P, Munn Z, Trico A, Khalil H. Chapter 11: Scoping reviews. In: Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis. Adelaide: JBI; 2020.
29. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
30. Melnyk BM, Fineout-Overholt E. Evidence-Based Practice in Nursing & Healthcare: A Guide to Best Practice. Philadelphia, PA: Wolters Kluwer Health; 2015.
31. Abdalrahim A, ALBashtawy M, Alkhawaldeh A, Ayed A. Examining the feasibility and acceptability of digital cognitive stimulation therapy for dementia care in Jordan: a qualitative study. SAGE Open Nurs 2024;10:23779608241272599. [doi: 10.1177/23779608241272599] [Medline: 39314644]
32. Barbosa Neves B, Franz R, Judges R, Beermann C, Baecker R. Can digital technology enhance social connectedness among older adults? A feasibility study. J Appl Gerontol 2019;38(1):49-72. [doi: 10.1177/0733464817741369] [Medline: 29166818]
33. Chapman H, Bethell J, Dewan N, Liougas MP, Livingston G, McGilton KS, et al. Social connection in long-term care homes: a qualitative study of barriers and facilitators. BMC Geriatr 2024;24(1):857 [FREE Full text] [doi: 10.1186/s12877-024-05454-8] [Medline: 39438797]
34. Gleeson E, Early E. Bringing the outside in through facilitated communication technology in long term care: how do facilitated, virtual communication interventions influence peer interaction, social engagement and participant well-being? Irish Association for Applied Linguistics. 2021. URL: https://hdl.handle.net/10468/11698 [accessed 2025-09-19]
35. Hird N, Osaki T, Ghosh S, Palaniappan SK, Maeda K. Enabling personalization for digital cognitive stimulation to support communication with people with dementia: pilot intervention study as a prelude to AI development. JMIR Form Res 2024;8:e51732 [FREE Full text] [doi: 10.2196/51732] [Medline: 38227357]
36. McAllister M, Dayton J, Oprescu F, Katsikitis M, Jones CM. Memory Keeper: a prototype digital application to improve engagement with people with dementia in long-term care (innovative practice). Dementia (London) 2020;19(4):1287-1298. [doi: 10.1177/1471301217737872] [Medline: 29096546]
37. Munk S, Toohey R, Remtilla A, D'Cunha NM, Gibson D, Isbel S, et al. "I forgot she used to make chocolate cake": digital storytelling supporting person-focussed dementia care: a qualitative thematic analysis. Dementia (London) 2025;24(4):666-696 [FREE Full text] [doi: 10.1177/14713012251317761] [Medline: 39876485]
38. O'Rourke HM, Swindle J, Chacinski D, Dal Pizzol FLF, Lee H, Hoben M, et al. Connecting Today: feasibility and acceptability of a remote visiting program for people living with dementia in long-term care homes. Dementia (London) 2023;22(7):1321-1347 [FREE Full text] [doi: 10.1177/14713012231176858] [Medline: 37341515]
39. Pisoni G. Mediating distance: new interfaces and interaction design techniques to follow and take part in remote museum visits. J Syst Inform Technol 2020;22(4):329-350. [doi: 10.1108/jsit-03-2020-0038]
40. Pisoni G, Daniel F, Casati F, Callaway C, Stock O. Interactive remote museum visits for older adults: an evaluation of feelings of presence, social closeness, engagement, and enjoyment in an social visit. 2019 Presented at: IEEE International Symposium on Multimedia (ISM); December 9-11, 2019; San Diego, CA, United States p. 99-102 URL: https://ieeexplore.ieee.org/document/8959064 [doi: 10.1109/ISM46123.2019.00023]
41. Sävenstedt S, Brulin C, Sandman PO. Family members' narrated experiences of communicating via video-phone with patients with dementia staying at a nursing home. J Telemed Telecare 2003;9(4):216-220. [doi: 10.1258/135763303322225544] [Medline: 12952692]

XSL•FO
RenderX

42. Welsh D, Morrissey K, Foley S, McNaney R, Salis C, McCarthy J, et al. Ticket to Talk: supporting conversation between young people and people with dementia through digital media. 2018 Presented at: CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems; April 21-26, 2018; Montreal, QC, Canada p. 1-14. [doi: 10.1145/3173574.3173949]
43. Muñoz D, Pedell S, Sterling L. Evaluating engagement in technology-supported social interaction by people living with dementia in residential care. ACM Trans Comput-Hum Interact 2022;29(5):1-31. [doi: 10.1145/3514497]
44. Hardy MS, Fanaki C, Savoie C, Dallaire C, Wilchesky M, Gallani MC, et al. Acceptability of videoconferencing to preserve the contact between cognitively impaired long-term care residents and their family caregivers: a mixed-methods study. Geriatr Nurs 2022;48:65-73 [FREE Full text] [doi: 10.1016/j.gerinurse.2022.09.006] [Medline: 36155311]
45. Hardy MS, Fanaki C, Savoie C, Wilchesky M, Gagnon MP, Laberge M, et al. Long-term care staffs' experience in facilitating the use of videoconferencing by cognitively impaired long-term care residents during the COVID-19 pandemic: a mixed-methods study. BMC Health Serv Res 2024;24(1):646 [FREE Full text] [doi: 10.1186/s12913-024-11095-9] [Medline: 38769512]
46. Munoz D, Favilla S, Pedell S, Murphy A, Beh J, Petrovich T. Evaluating an app to promote a beter visit through shared activities for people living with dementia and their families. 2021 Presented at: CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; May 8-13, 2021; Yokohama, Japan p. 1-13. [doi: 10.1145/3411764.3445764]
47. Thach KS, Lederman R, Waycott J. Key considerations for the design of technology for enrichment in residential aged care: an ethnographic study. 2023 Presented at: CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems; April 23-28, 2023; Hamburg, Germany p. 1-16. [doi: 10.1145/3544548.3581176]
48. Webber S, Baker S, Waycott J. Virtual visits: reminiscence in residential aged care with digital mapping technologies. Australas J Ageing 2021;40(3):293-300. [doi: 10.1111/ajag.12902] [Medline: 33538405]
49. Akhtar SA. Socially connected while apart: the use of technology to increase social connection between nursing home residents and their relatives. Front Public Health 2024;12:1296524 [FREE Full text] [doi: 10.3389/fpubh.2024.1296524] [Medline: 38333742]
50. Badawy A, Solberg M, Obstfelder AU, Alnes RE. Improvised use of a digital tool for social interaction in a Norwegian care facility during the COVID-19 pandemic: an exploratory study. BMC Health Serv Res 2022;22(1):136 [FREE Full text] [doi: 10.1186/s12913-022-07526-0] [Medline: 35105344]
51. Feng Y, van RR, Yu S, Hu J, Rauterberg M, Barakova E. Using observational engagement assessment method VC-IOE for evaluating an interactive table designed for seniors with dementia. 2017 Presented at: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); June 26-27, 2017; Hong Kong, China p. 26-37. [doi: 10.1007/978-3-319-67964-8_3]
52. Grigorovich A, Marcotte AA, Colobong R, Szabo M, MacNeill C, Blais D, et al. Using voice-activated technologies to enhance well-being of older adults in long-term care homes. Innov Aging 2024;8(12):igae102 [FREE Full text] [doi: 10.1093/geroni/igae102] [Medline: 39664605]
53. Houben M, Brankaert R, Dhaeze E, Kenning G, Bongers I, Eggen B. Enriching everyday lived experiences in dementia care. 2022 Presented at: TEI '22: Proceedings of the Sixteenth International Conference on Tangible, Embedded, and Embodied Interaction; February 13-16, 2022; Daejeon, Republic of Korea p. 1-13. [doi: 10.1145/3490149.3501326]
54. Houben M, Brankaert R, Bakker S, Kenning G, Bongers I, Eggen B. The role of everyday sounds in advanced dementia care. 2020 Presented at: Conference on Human Factors in Computing Systems, CHI 2020; April 25-30, 2020; Honolulu p. 14. [doi: 10.1145/3313831.3376577]
55. Houben M, Brankaert R, Kenning G, Eggen B, Bongers I. The perspectives of professional caregivers on implementing audio-based technology in residential dementia care. Int J Environ Res Public Health 2020;17(17):6333 [FREE Full text] [doi: 10.3390/ijerph17176333] [Medline: 32878116]
56. Jamin G, Luyten T, Delsing R, Braun S. The process of co-creating the interface for VENSTER, an interactive artwork for nursing home residents with dementia. Disabil Rehabil Assist Technol 2018;13(8):809-818. [doi: 10.1080/17483107.2017.1385102] [Medline: 29037109]
57. Luyten T, Braun S, Jamin G, van Hooren S, de Witte L. How nursing home residents with dementia respond to the interactive art installation 'VENSTER': a pilot study. Disabil Rehabil Assist Technol 2018;13(1):87-94 [FREE Full text] [doi: 10.1080/17483107.2017.1290701] [Medline: 28287047]
58. Paay J, Kjeldskov J, Aaen I, Bank M. User-centred iterative design of a smartwatch system supporting spontaneous reminiscence therapy for people living with dementia. Health Informatics J 2022;28(2):14604582221106002 [FREE Full text] [doi: 10.1177/14604582221106002] [Medline: 35658692]
59. van Wingerden E, Vacaru SV, Holstege L, Sterkenburg PS. Hey Google! Intelligent personal assistants and well-being in the context of disability during COVID-19. J Intellect Disabil Res 2023;67(10):973-985. [doi: 10.1111/jir.13064] [Medline: 37317950]
60. Akhtar SA. Together through technology: next of kin’s use of technology to maintain social connection with nursing home residents. Ageing Soc 2025;45(11):2339-2355. [doi: 10.1017/s0144686x25000054]

XSL•FO
RenderX

61. Badawy A, Solberg M, Obstfelder AU, Alnes RE. Together, at a distance: experiences with a novel technology for social contact among older people and their relatives in Norway during the COVID-19 pandemic. BMC Geriatr 2023;23(1):218 [FREE Full text] [doi: 10.1186/s12877-023-03869-3] [Medline: 37024822]
62. Biemans M, Van Dijk B. Food for talk: photo frames to support social connectedness for elderly people in a nursing home. 2009 Presented at: ECCE '09: European Conference on Cognitive Ergonomics: Designing beyond the Product --- Understanding Activity and User Experience in Ubiquitous Environments; September 30-October 2, 2009; Helsinki, Finland p. 147-153 URL: https://dl.acm.org/doi/10.5555/1690508.1690528 [doi: 10.5555/1690508.1690528]
63. Brookman R, Hulm Z, Hearn L, Siette J, Mathew N, Deodhar S, et al. Evaluation of an exercise program incorporating an international cycling competition: a multimodal intervention model for physical, psychological, and social wellbeing in residential aged care. BMC Geriatr 2024;24(1):435 [FREE Full text] [doi: 10.1186/s12877-024-05033-x] [Medline: 38755554]
64. Hasan H, Linger H. Enhancing the wellbeing of the elderly: social use of digital technologies in aged care. Educ Gerontol 2016;42(11):749-757. [doi: 10.1080/03601277.2016.1205425]
65. Lignos N, McCloskey R, Donovan C, Ellis K, Herrington M, Kanik M. Use of an ambient activity technology for long-term care residents with dementia. J Gerontol Nurs 2022;48(1):35-41. [doi: 10.3928/00989134-20211206-05] [Medline: 34978492]
66. MacRitchie J, Breaden M, Taylor JR, Milne AJ. Exploring older adult needs and preferences for technology-assisted group music-making. A qualitative analysis of data collected during the participatory user-centred design process. Disabil Rehabil Assist Technol 2024;19(5):1935-1944 [FREE Full text] [doi: 10.1080/17483107.2022.2077461] [Medline: 35658719]
67. Subramaniam P, Woods B. Digital life storybooks for people with dementia living in care homes: an evaluation. Clin Interv Aging 2016;11:1263-1276 [FREE Full text] [doi: 10.2147/CIA.S111097] [Medline: 27698556]
68. Thoolen M, Brankaert R, Lu Y. AmbientEcho: exploring interactive media experiences in the context of residential dementia care. 2020 Presented at: DIS '20: Proceedings of the 2020 ACM Designing Interactive Systems Conference; July 6-10, 2020; Eindhoven, Netherlands p. 1495-1508. [doi: 10.1145/3357236.3395432]
69. Vasseur D, Ipakchian Askari S, Suijkerbuijk S, Nap HH, IJsselsteijn W. Sensory, affective, and social experiences with haptic devices in intramural care practice. Nurs Rep 2024;14(1):230-253 [FREE Full text] [doi: 10.3390/nursrep14010019] [Medline: 38391064]
70. Belgrave M, Palmer K, Luger Motyka TM. Creative aging in virtual spaces: using museum content and music therapy to explore cultural diversity. Front Med (Lausanne) 2023;10:1273000 [FREE Full text] [doi: 10.3389/fmed.2023.1273000] [Medline: 38098846]
71. Breckenridge JP, Kroll T, Wylie G, Salzberg A. Remembering the past, enhancing the present and sharing the future: a qualitative study of the impact of film screenings in care home communities. Ageing Soc 2020;41(11):2615-2640. [doi: 10.1017/s0144686x20000501]
72. Caserta MS, Lund DA. Video Respite® in an Alzheimer's care center. Act Adapt Aging 2008;27(1):13-28. [doi: 10.1300/j016v27n01_02]
73. Howard EP, Retalic T, Rogan J, Murphy K, Swaminathan S, Altschuler A. ArtontheBrain: results of a pilot project among long-term care residents. Res Gerontol Nurs 2021;14(5):235-243. [doi: 10.3928/19404921-20210825-03] [Medline: 34542348]
74. Juul A, Wilding R, Baldassar L. The best day of the week: new technology enhancing quality of life in a care home. Int J Environ Res Public Health 2019;16(6):1000 [FREE Full text] [doi: 10.3390/ijerph16061000] [Medline: 30893945]
75. Ysseldyk R, Morton TA, Haslam C, Haslam SA, Boger J, Giau E, et al. You've Got E-Mail: a pilot study examining the feasibility and impact of a group-based technology-training intervention among older adults living in residential care. Can J Aging 2024;43(1):45-56. [doi: 10.1017/S0714980823000375] [Medline: 37501571]
76. Samuelsson C, Ferm U, Ekström A. "It's Our Gang"—promoting social inclusion for people with dementia by using digital communication support in a group activity. Clin Gerontol 2021;44(4):418-429 [FREE Full text] [doi: 10.1080/07317115.2020.1795037] [Medline: 32791946]
77. Tsai HH, Tsai YF, Wang HH, Chang YC, Chu HH. Videoconference program enhances social support, loneliness, and depressive status of elderly nursing home residents. Aging Ment Health 2010;14(8):947-954. [doi: 10.1080/13607863.2010.501057] [Medline: 21069600]
78. Bradwell HL, Cooper L, Edwards KJ, Baxter R, Tomaz SA, Ritchie J, et al. Staff perceptions towards virtual reality-motivated treadmill exercise for care home residents: a qualitative feedback study with key stakeholders and follow-up interview with technology developer. BMJ Open 2023;13(11):e073307 [FREE Full text] [doi: 10.1136/bmjopen-2023-073307] [Medline: 37996232]
79. Goodall G, Taraldsen K, Granbo R, Serrano JA. Towards personalized dementia care through meaningful activities supported by technology: a multisite qualitative study with care professionals. BMC Geriatr 2021;21(1):468 [FREE Full text] [doi: 10.1186/s12877-021-02408-2] [Medline: 34418973]
80. Isbel S, Frost J, Fearon A, Kellett J, D'Cunha NM. Take a Cycle with Me: an intergenerational group activity with older adults and adolescents. J Intergener Relatsh 2023;23(1):1-11. [doi: 10.1080/15350770.2023.2282499]

XSL•FO
RenderX

81. Li Y, Wilke C, Shiyanov I, Muschalla B. Impact of virtual reality-based group activities on activity level and well-being among older adults in nursing homes: longitudinal exploratory study. JMIR Serious Games 2024;12:e50796 [FREE Full text] [doi: 10.2196/50796] [Medline: 38551635]
82. Sheehy L, Bouchard S, Kakkar A, El Hakim R, Lhoest J, Frank A. Development and initial testing of an artificial intelligence-based virtual reality companion for people living with dementia in long-term care. J Clin Med 2024;13(18):5574 [FREE Full text] [doi: 10.3390/jcm13185574] [Medline: 39337060]
83. D'Cunha NM, Isbel ST, Frost J, Fearon A, McKune AJ, Naumovski N, et al. Effects of a virtual group cycling experience on people living with dementia: a mixed method pilot study. Dementia (London) 2021;20(5):1518-1535. [doi: 10.1177/1471301220951328] [Medline: 32820955]
84. Goodall G, André L, Taraldsen K, Serrano JA. Supporting identity and relationships amongst people with dementia through the use of technology: a qualitative interview study. Int J Qual Stud Health Well-being 2021;16(1):1920349 [FREE Full text] [doi: 10.1080/17482631.2021.1920349] [Medline: 33955324]
85. Khirallah Abd El Fatah N, Abdelwahab Khedr M, Alshammari M, Mabrouk Abdelaziz Elgarhy S. Effect of immersive virtual reality reminiscence versus traditional reminiscence therapy on cognitive function and psychological well-being among older adults in assisted living facilities: a randomized controlled trial. Geriatr Nurs 2024;55:191-203. [doi: 10.1016/j.gerinurse.2023.11.010] [Medline: 38007908]
86. Lin TY, Huang CM, Hsu HP, Liao JY, Cheng VYW, Wang SW, et al. Effects of a combination of three-dimensional virtual reality and hands-on horticultural therapy on institutionalized older adults' physical and mental health: quasi-experimental design. J Med Internet Res 2020;22(11):e19002 [FREE Full text] [doi: 10.2196/19002] [Medline: 33135666]
87. Siriaraya P, Ang CS. Recreating living experiences from past memories through virtual worlds for people with dementia. 2014 Presented at: CHI '14: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; April 26-May 1, 2014; Toronto, Ontario, Canada p. 3977-3986. [doi: 10.1145/2556288.2557035]
88. Feng Y, Perugia G, Yu S, Barakova EI, Hu J, Rauterberg GWM. Context-enhanced human-robot interaction: exploring the role of system interactivity and multimodal stimuli on the engagement of people with dementia. Int J Soc Robot 2021;14(3):807-826. [doi: 10.1007/s12369-021-00823-4]
89. Melkas H, Hennala L, Pekkarinen S, Kyrki V. Impacts of robot implementation on care personnel and clients in elderly-care institutions. Int J Med Inform 2020;134:104041 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.104041] [Medline: 31838291]
90. Paluch R, Müller C. "That's Something for Children": an ethnographic study of attitudes and practices of care attendants and nursing home residents towards robotic pets. 2022 Presented at: Proceedings of the ACM on Human-Computer Interaction, Volume 6, Issue GROUP; January 14, 2022; New York, NY, United States p. 1-35. [doi: 10.1145/3492850]
91. Blindheim K, Solberg M, Hameed IA, Alnes RE. Promoting activity in long-term care facilities with the social robot Pepper: a pilot study. Inform Health Soc Care 2023;48(2):181-195 [FREE Full text] [doi: 10.1080/17538157.2022.2086465] [Medline: 35702818]
92. Chen SC, Davis BH, Kuo CY, Maclagan M, Chien CO, Lin MF. Can the Paro be my Buddy? Meaningful experiences from the perspectives of older adults. Geriatr Nurs 2022;43:130-137. [doi: 10.1016/j.gerinurse.2021.11.011] [Medline: 34883391]
93. De Carolis B, Macchiarulo N, La Forgia A, Melone G, Nardelli F. Designing and evaluating Rebecca: a telepresence robot in the shoes of a social assistant for cognitive stimulation therapy. 2025 Presented at: CHItaly '25: Proceedings of the 16th Biannual Conference of the Italian SIGCHI Chapter; October 6-10, 2022; Salerno, Italy p. 1-8. [doi: 10.1145/3750069.3750119]
94. Feng Y, Barakova EI, Yu S, Hu J, Rauterberg GWM. Effects of the level of interactivity of a social robot and the response of the augmented reality display in contextual interactions of people with dementia. Sensors (Basel) 2020;20(13):3771 [FREE Full text] [doi: 10.3390/s20133771] [Medline: 32635640]
95. Feng Y, Yu S, Van DMD, Barakova E, Hu J, Rauterberg M. LiveNature: ambient display and social robot-facilitated multi-sensory engagement for people with dementia. 2019 Presented at: DIS '19: Proceedings of the 2019 on Designing Interactive Systems Conference; June 23-28, 2019; San Diego, CA, United States p. 1321-1333. [doi: 10.1145/3322276.3322331]
96. Fogelson DM, Rutledge C, Zimbro KS. The impact of robotic companion pets on depression and loneliness for older adults with dementia during the COVID-19 pandemic. J Holist Nurs 2022;40(4):397-409 [FREE Full text] [doi: 10.1177/08980101211064605] [Medline: 34889146]
97. Khosla R, Chu MT. Embodying care in matilda: an affective communication robot for emotional wellbeing of older people in Australian residential care facilities. ACM Trans Manag Inf Syst 2013;4(4):1-33. [doi: 10.1145/2544104]
98. Chang CY, Wang PL, Li CJ, Hwang GJ. From empathy to quality long-term care: a generative AI-based art therapy approach based on the self-directed learning model. Interact Learn Environ 2024;33(5):3333-3353. [doi: 10.1080/10494820.2024.2443072]
99. Marchetti E, Grimme S, Hornecker E, Kollakidou A, Graf P. Pet-robot or appliance? Care home residents with dementia respond to a zoomorphic floor washing robot. 2022 Presented at: CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems; April 29-May 5, 2022; New Orleans, LA, United States p. 1-21. [doi: 10.1145/3491102.3517463]

XSL•FO
RenderX

100. Wangmo T, Duong V, Felber NA, Tian YJA, Mihailov E. No playing around with robots? Ambivalent attitudes toward the use of Paro in elder care. Nurs Inq 2024;31(3):e12645. [doi: 10.1111/nin.12645] [Medline: 38812242]
101. Wu D, Pu L, Jo J, Hexel R, Moyle W. Deploying robot-led activities for people with dementia at aged care facilities: a feasibility study. J Am Med Dir Assoc 2024;25(7):105028. [doi: 10.1016/j.jamda.2024.105028] [Medline: 38772526]
102. Hoel V, Seibert K, Domhoff D, Preuß B, Heinze F, Rothgang H, et al. Social health among German nursing home residents with dementia during the COVID-19 pandemic, and the role of technology to promote social participation. Int J Environ Res Public Health 2022;19(4):1956 [FREE Full text] [doi: 10.3390/ijerph19041956] [Medline: 35206143]
103. Li Y, Liang N, Nejat G. How emotional dances with social robots influence group HRI for older adults. ACM Trans Hum Robot Interact 2025;15(2):1-33. [doi: 10.1145/3773894]
104. Waycott J, Zhao W, Kelly RM, Robertson E. Technology-mediated enrichment in aged care: survey and interview study. JMIR Aging 2022;5(2):e31162 [FREE Full text] [doi: 10.2196/31162] [Medline: 34975014]
105. Ito H, Banh H, Wong KLY, Wong L, Allen S, Hung L. A critical reflection of an intergenerational, student-led team bringing social robots and research to older adults in the community. Gerontol Geriatr Educ 2025;46(4):587-602. [doi: 10.1080/02701960.2025.2507413] [Medline: 40380793]
106. Dodson P, Haase AM, Jeffreys M, Hales C. Capturing patient experiences of care with digital technology to improve service delivery and quality of care: a scoping review. Digit Health 2024;10:20552076241282900 [FREE Full text] [doi: 10.1177/20552076241282900] [Medline: 39449855]
107. Jecker NS, Sparrow R, Lederman Z, Ho A. Digital humans to combat loneliness and social isolation: ethics concerns and policy recommendations. Hastings Cent Rep 2024;54(1):7-12. [doi: 10.1002/hast.1562] [Medline: 38390675]
108. Brankaert R. Inclusive designs in healthcare: warm technology for people with dementia. In: Applied Design Research. Boca Raton, FL: CRC Press; 2022:137-146.
109. Hicks B, Karim A, Jones E, Burgin M, Cutler C, Tang W, et al. Care home practitioners' perceptions of the barriers and facilitators for using off-the-shelf gaming technology with people with dementia. Dementia (London) 2022;21(5):1532-1555 [FREE Full text] [doi: 10.1177/14713012221085229] [Medline: 35427198]
110. Zhao W, Kelly RM, Waycott J. Technology deployment for social connection in residential aged care: care and technology providers' experiences during the COVID-19 pandemic. 2023 Presented at: CHI '23: CHI Conference on Human Factors in Computing Systems; April 23-28, 2023; Hamburg, Germany. [doi: 10.1145/3544549.3585662]
111. Mihailov E, Wangmo T. Does humanness matter? An ethical evaluation of sharing care work with social robots. Sci Eng Ethics 2025;31(4):20. [doi: 10.1007/s11948-025-00547-y] [Medline: 40721552]
112. Ware C, Rigaud AS, Blavette L, Damnée S, Dacunha S, Lenoir H, et al. Development of an ethical framework for the use of social robots in the care of individuals with major neurocognitive disorders: a qualitative study. BMC Geriatr 2025;25(1):260 [FREE Full text] [doi: 10.1186/s12877-025-05891-z] [Medline: 40264012]
113. Khaksar W, Lindblom DS, Bygrave LA, Torresen J. Robotics in elderly healthcare: a qualitative analysis of 20 recent European research projects. ACM Trans Hum Robot Interact 2025;14(2):1-38. [doi: 10.1145/3711936]
114. Hall A, Wilson CB, Stanmore E, Todd C. Implementing monitoring technologies in care homes for people with dementia: a qualitative exploration using normalization process theory. Int J Nurs Stud 2017;72:60-70 [FREE Full text] [doi: 10.1016/j.ijnurstu.2017.04.008] [Medline: 28494333]
115. Theopilus Y, Davis H, Pedell S. Digital technology for supporting aged care services: a scoping review. 2025 Presented at: CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems; April 26-May 1, 2025; Yokohama, Japan p. 1-12. [doi: 10.1145/3706598.3713498]
116. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240020924 [accessed 2025-09-18]
117. Tischendorf T, Hinsche L, Hasseler M, Schaal T. GenAI in nursing and clinical practice: a rapid review of applications and challenges. J Public Health 2025:1-11. [doi: 10.1007/s10389-025-02523-z]
118. Pardini S, Calcagno R, Genovese A, Salvadori E, Ibarra OM. Exploring virtual reality-based reminiscence therapy on cognitive and emotional well-being in people with cognitive impairments: a scoping review. Brain Sci 2025;15(5):500 [FREE Full text] [doi: 10.3390/brainsci15050500] [Medline: 40426671]
119. Reed JM, Dodson T, Petrinec A, Hughes J, Miller RD. The HARMONEE project: using genAI images for reminiscence with older adults in long-term care. Geriatr Nurs 2025;65:103448 [FREE Full text] [doi: 10.1016/j.gerinurse.2025.103448] [Medline: 40618723]
120. Bradley L, Shanker S, Murphy J, Fenge LA, Heward M. Effectiveness of digital technologies to engage and support the wellbeing of people with dementia and family carers at home and in care homes: a scoping review. Dementia (London) 2023;22(6):1292-1313 [FREE Full text] [doi: 10.1177/14713012231178445] [Medline: 37235791]
121. Vrančić A, Zadravec H, Orehovački T. The role of smart homes in providing care for older adults: a systematic literature review from 2010 to 2023. Smart Cities 2024;7(4):1502-1550. [doi: 10.3390/smartcities7040062]
122. Abbasian M, Azimi I, Rahmani AM, Jain R. Conversational health agents: a personalized LLM-powered agent framework. JAMIA Open 2025;8(4):ooaf067 [FREE Full text] [doi: 10.1093/jamiaopen/ooaf067]

XSL•FO
RenderX

123. Khan MA, Din IU, Khan NA, Hassan S, Almogren A. Adaptive generative AI for elderly-assisted living environments: a proactive approach. 2024 Presented at: 7th International Conference on Internet Applications, Protocols, and Services (NETAPPS); November 6-7, 2024; Kuala Lumpur, Malaysia. [doi: 10.1109/netapps63333.2024.10823634]
124. Levin C, Zaboli A, Turcato G, Saban M. Nursing judgment in the age of generative artificial intelligence: a cross-national study on clinical decision-making performance among emergency nurses. Int J Nurs Stud 2025;172:105216 [FREE Full text] [doi: 10.1016/j.ijnurstu.2025.105216] [Medline: 40975905]
125. Bautista JR, Westendorf R, Goth O. Using a socio-technical strategy to identify the use and implications of generative artificial intelligence tools on nursing education and practice. J Adv Nurs 2025. [doi: 10.1111/jan.70401] [Medline: 41288087]
126. Johnson J, Jamshed H, Zuppke R, Leggett A, Provost EM, Brewer RN. “I Felt Listened to”: evaluating an AI-powered reflection tool for care partners. ACM Trans Access Comput 2025;18(3):1-35. [doi: 10.1145/3731562]
127. Karimi P, Martin-Hammond A. Finding the right balance: user control and automation in AI tools for supporting older adults' health information tasks. 2025 Presented at: CHI EA '25: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems; April 26-May 1, 2025; Yokohama, Japan p. 1-10. [doi: 10.1145/3706599.3719773]
128. Karimi P, Martin-Hammond A. Designing intelligent voice assistants for older adults' collaborative care: exploring supportive and non-supportive interactions. Proc ACM Hum-Comput Interact 2025;9(7):1-35. [doi: 10.1145/3757465]
129. Saha K, Jain Y, Liu C, Kaliappan S, Karkar R. AI vs. humans for online support: comparing the language of responses from LLMs and online communities of alzheimer's disease. 2025 Presented at: ACM Transactions on Computing for Healthcare; March 11, 2025; New York, NY, United States. [doi: 10.1145/3709366]
130. Katz S. Aging together-with: The growing older of humans, non-humans and more-than-humans. A commentary. J Aging Stud 2024;71:101280 [FREE Full text] [doi: 10.1016/j.jaging.2024.101280] [Medline: 39608904]
131. Vandemeulebroucke T, Dzi K, Gastmans C. Older adults' experiences with and perceptions of the use of socially assistive robots in aged care: a systematic review of quantitative evidence. Arch Gerontol Geriatr 2021;95:104399. [doi: 10.1016/j.archger.2021.104399] [Medline: 33813208]
132. Okabe-Miyamoto K, Walsh LC, Ozer DJ, Lyubomirsky S. Measuring the experience of social connection within specific social interactions: the Connection During Conversations Scale (CDCS). PLoS One 2024;19(1):e0286408 [FREE Full text] [doi: 10.1371/journal.pone.0286408] [Medline: 38236933]
133. Moyle W, Pu L, Murfield J, Sung B, Sriram D, Liddle J, et al. Consumer and provider perspectives on technologies used within aged care: an Australian qualitative needs assessment survey. J Appl Gerontol 2022;41(12):2557-2565 [FREE Full text] [doi: 10.1177/07334648221120082] [Medline: 35948942]
134. Bhowmick P, Stolterman BE. Exploring tangible user interface design for social connection among older adults: a preliminary review. 2023 Presented at: CHI EA '23: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems; April 23-28, 2023; Hamburg, Germany p. 1-9. [doi: 10.1145/3544549.3585722]
135. Keane J, Ryan C, Usher R. Leisure and social occupational choice within nursing home facilities in Ireland: residents perspectives. Scand J Occup Ther 2024;31(1):2259222 [FREE Full text] [doi: 10.1080/11038128.2023.2259222] [Medline: 37918009]
136. Yuan S, Coghlan S, Lederman R, Waycott J. Meaningful engagement, ethical care, and design opportunities: an ethnographic study on social activities in long-term care. 2025 Presented at: CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems; April 26-May 1, 2025; Yokohama, Japan p. 1-15. [doi: 10.1145/3706598.3713755]
137. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014;348:g1687 [FREE Full text] [doi: 10.1136/bmj.g1687] [Medline: 24609605]

Abbreviations

GenAI: generative artificial intelligence
HARMONEE: Harnessing Artificial Intelligence Resources for Mental Well-Being for Older Adults and Nurturing Empathy in Education
JBI: Joanna Briggs Institute
MeSH: Medical Subject Headings
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
TIDieR: Template for Intervention Description and Replication
VR: virtual reality
WHO: World Health Organization

XSL•FO
RenderX

Edited by S Brini; submitted 26.Nov.2025; peer-reviewed by Z Su, W Ghulam Hussain; comments to author 09.Dec.2025; accepted 26.Jan.2026; published 03.Mar.2026.

Please cite as:
Muldrew D, Bradley R, Conway K
Digital Technologies That Support Meaningful Connections in Care Homes: Scoping Review
J Med Internet Res 2026;28:e88181
URL: https://www.jmir.org/2026/1/e88181
doi:10.2196/88181
PMID:

©Deborah Muldrew, Rosemary Bradley, Kelly Conway. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Association Between Digital Biomarkers of Health and Anxiety: Systematic Review and Meta-Analysis

Yolanda Lau[1], MSc; Natalia Chemas[2], MSc; Heema Ajeet Gokani[1], MSc; Rachel Morrell[1], MSc; Harisd Phannarus[1,3], MD; Claudia Cooper[4], MD; Zuzana Walker[1,5], MD; Harriet Demnitz-King[1,4], PhD; Natalie L Marchant[1], PhD

[1]Division of Psychiatry, University College London, London, United Kingdom

[2]Centre for Preventive Neurology, Wolfson Institute of Population Health, Queen Mary University of London, London, United Kingdom

[3]Department of Preventive and Social Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand

[4]Centre for Psychiatry and Mental Health, Wolfson Institute of Population Health, Queen Mary University of London, London, United Kingdom

[5]Essex Partnership University NHS Foundation Trust, Essex, United Kingdom

Corresponding Author:
Natalie L Marchant, PhD
Division of Psychiatry
University College London
Division of Psychiatry, 4th Floor Maple House
149 Tottenham Court Road
London, W1T 5AU
United Kingdom
Phone: 44 20 3108 7961
Email: n.marchant@ucl.ac.uk

Abstract

Background: Digital biomarkers are gaining interest as proxy markers for mental health, as they enable passive and continuous data collection. However, the association between digital biomarkers of health and anxiety, both generalized anxiety disorder and anxiety symptoms, remains unknown.

Objective: This systematic review and meta-analysis examined the association between digital biomarkers of health obtained from wrist-worn wearables and anxiety in adults.

Methods: Systematic literature searches were conducted across 6 databases, including unpublished gray literature. The final search was done on September 21, 2025. Cross-sectional or longitudinal studies investigating the association between digital biomarkers from wrist-worn wearables and anxiety were eligible. Studies using inferential statistics or machine learning methods were both eligible. Studies were excluded if participants received diagnoses of neurodegenerative disorders or physical health conditions. Two risk-of-bias tools were used: the National Heart, Lung, and Blood Institute assessment tool for inferential statistical studies, and the modified version of the Quality Assessment of Diagnostic Accuracy Studies-2 for machine learning studies. Whenever possible, effect sizes were combined across studies, for each digital biomarker of health separately, using random-effects meta-analyses. Sensitivity analyses were performed to assess whether results differed according to anxiety type (state or trait) and age group. Otherwise, studies were synthesized narratively.

Results: A total of 44 studies from 42 articles were eligible. Among these, 36 studies used inferential statistical approaches for analysis (21 reporting sleep characteristics, 8 reporting physical activity, 2 reporting heart rate variability, and 5 reporting more than 1 type), and 8 studies used machine learning approaches. Sample size ranged from 17 to 170,320. Meta-analyses on 4 sleep metrics found no associations: sleep efficiency (Fisher z=–0.07, 95% CI –0.14 to 0.002; P=.06; PI –0.19 to 0.05), wake after sleep onset (Fisher z=0.13, 95% CI –0.04 to 0.30; P=.11; PI –0.15 to 0.41), total sleep time (Fisher z=0.009, 95% CI –0.01 to 0.03; P=.28; PI –0.02 to 0.03), and sleep onset latency (Fisher z=0.04, 95% CI –0.07 to 0.15; P=.08; PI –0.19 to 0.27). Qualitative syntheses revealed that lower physical activity levels and higher heart rate were associated with greater anxiety symptoms. Machine learning studies using wrist-worn wearable data alone showed varied performance, with predictive performance improving when wearable data were combined with other data sources.

Conclusions: This is the first review to synthesize evidence from inferential statistical (mostly fair quality) and machine learning studies examining association between wearable-derived digital biomarkers and anxiety. Meta-analyses found no associations between sleep metrics and anxiety. Although based on limited studies, lower physical activity levels and elevated heart rate were

XSL•FO
RenderX

associated with greater anxiety symptoms. Digital biomarkers may be more useful when integrated with other data sources (eg, self-report and clinical data) rather than used as stand-alone screening tools.

Trial Registration: PROSPERO CRD42023409995; https://www.crd.york.ac.uk/PROSPERO/view/CRD42023409995

(J Med Internet Res 2026;28:e73812) doi:10.2196/73812

KEYWORDS

anxiety; digital; heart rate; machine learning; meta-analysis; physical activity; sleep; systematic review; wearable

Introduction

Anxiety is the most prevalent mental health condition globally, impacting 301 million people in 2019 [1]. Symptoms can include worry, restlessness, fatigue, difficulty concentrating, and sleep disturbances [2]. Among older adults, anxiety disorders and anxiety symptoms are common, with 1.2% to 28% experiencing anxiety disorders (generalized anxiety disorder [GAD] being the most common type) and 15% to 56% experiencing anxiety symptoms in community and clinical populations [3]. In older adults, both GAD and anxiety symptoms have been associated with increased risk of dementia [4-8]. Given the increasing global prevalence of dementia [9] and the lack of effective disease-modifying treatments, it is crucial to prioritize the identification and treatment of modifiable risk factors associated with dementia to reduce dementia cases. Therefore, early identification of anxiety in both the general population and older adults could have significant clinical implications.

Predominant methods to assess anxiety include the use of self-report questionnaires, clinical diagnoses, or a combination of both. These methods have several limitations. Self-report questionnaires measuring trait anxiety are retrospective and rely on an individual's ability, insight, and willingness to respond accurately. Evidence suggests that self-report questionnaires can also be affected by recall bias and social desirability, potentially leading to misreporting of anxiety symptoms [10]. Clinical diagnoses require clinical expertise and are, therefore, time-consuming and resource-heavy. In addition, in older adults, anxiety is often underdiagnosed due to overlapping symptoms with physical health conditions [11]. Both methods only provide a snapshot of an individual's symptomology unless assessed repeatedly, which can increase demands on both individuals and clinicians.

Leveraging digital biomarkers, defined as objective, quantifiable, physiological, and behavioral measures collected using digital devices [12], as proxy markers of anxiety has the potential to overcome the limitations outlined above. Digital biomarkers collect continuous objective data on various aspects of behavior and physiology.

While digital biomarkers can be measured using different technologies (eg, contactless sensors or smartphones), this review focuses on wrist-worn wearable devices, which are widely used by the general public and in existing research [13,14]. Wrist-worn wearables can measure physical activity (eg, different physical intensities), sleep (eg, sleep efficiency [SE]), and heart rate (eg, heart rate variability [HRV]). These variables have been linked to anxiety; for example, lower physical activity levels and more sleep disturbances have been associated with more anxiety symptoms and GAD [15,16]. There is also emerging evidence supporting the effectiveness of wearable-derived metrics in detecting mental health symptoms [17,18]. Studies also found that digital biomarkers collected from other devices can detect mental health symptoms. For instance, a study found that metrics from smartphones (including heart rate metrics) were able to predict changes in anxiety symptoms [19]. Similarly, studies using the Oura Ring also reported associations between sleep metrics and anxiety [20].

Two recent meta-analyses of machine learning studies examined the association between metrics collected from wearable devices and depression and anxiety, both of which are modifiable psychosocial risk factors of dementia [17,18]. They found that the models accurately classified individuals with and without depression in 89% of cases [17] and correctly identified individuals with and without anxiety in 82% of cases [18]. Similar wearable-derived metrics were used as predictors in anxiety and depression models (eg, activity, sleep, and heart rate). These meta-analyses only focused on machine learning studies; existing research also used inferential statistical methods.

Inferential statistical methods and machine learning models offer distinct advantages. Inferential statistical models provide interpretable results, allowing for hypothesis testing and quantifying associations between digital biomarkers of health and anxiety. In contrast, machine learning models often prioritize predictive accuracy and excel at capturing complex, nonlinear relationships that may not be easily captured through inferential statistical models. Together, these 2 approaches offer complementary insights.

Given the growing interest in using wearable-derived metrics in mental health and dementia research, this review aims to contribute to these research fields by systematically reviewing existing studies that have investigated the association between metrics collected from wrist-worn wearables and anxiety using inferential and machine learning statistical approaches.

Methods

Design

This review was conducted in line with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) recommendation [21] and registered with PROSPERO (International Prospective Register of Systematic Reviews) in March 2023 (CRD42023409995). The PRISMA 2020 checklist [22] (Multimedia Appendix 1), the PRISMA Abstract checklist [22] (Multimedia Appendix 2), and the PRISMA-S (Preferred

XSL•FO
RenderX

Reporting Items for Systematic Reviews and Meta-Analyses–Search Extension) checklist [23] (Multimedia Appendix 3) were used.

Search Strategy

A total of 5 databases (Medline, Embase, PsycINFO, Web of Science, and CINAHL) were systematically searched. Further, unpublished gray literature searches were conducted with ProQuest Dissertations and Theses Global. Following published recommendations [24], we also checked the first 300 papers on Google Scholar to supplement the gray literature search. Finally, to supplement the electronic search, references of eligible papers were reviewed to identify any additional eligible studies.

The search strategy combined 2 search term strings with "AND." Each string reflected a concept related to the research question: (1) digital biomarkers and (2) psychosocial risk factors of dementia (Multimedia Appendix 4). The search strategy was informed by previous reviews in related research areas [17], but all search terms and strategies were developed specifically for this review and adapted for each database. There were no restrictions imposed on the date of publication. This review did not restrict the database searches to human studies and studies published in English, and no filters were used. The initial searches were conducted on April 27, 2023; a subsequent search was done on September 24, 2024, and a final search focusing only on anxiety was done on September 21, 2025. While the PROSPERO registration included multiple psychosocial risk factors, this systematic review focused specifically on anxiety, due to the large number of articles that made it past full-text screening. As all eligible anxiety studies in earlier searches were identified via anxiety-specific terms, a final search restricted to anxiety-related search terms was conducted in 2025.

Study Selection

The web platform Covidence (Veritas Health Innovation, 2020) was used for deduplication and to coordinate multiuser title-abstract and full-text screening. Two reviewers independently screened titles and abstracts of all the identified studies (YL screened 100% of the articles, and NC, HAG, RM, and HD-K collectively screened 100%), followed by full-text screening. Any disagreements were resolved by a third reviewer (NM or HD-K), with HD-K only resolving conflicts for articles she had not screened at the title-abstract screening stage. Interrater agreement was assessed using the Cohen κ coefficient.

In accordance with the PROSPERO registration, our search strategies encompassed a range of psychosocial risk factors for dementia. However, given the large number of eligible articles that made it through to the full-text screening stage, we decided to categorize these articles into different psychosocial risk factors before the data extraction stage.

This systematic review focused on anxiety, both clinical anxiety and anxiety symptoms. The PECOS (Population, Exposure, Control, Outcome, and Study design) framework was used to structure the study selection. Studies were selected if they (1) involved adults with a mean age of >18 years (Population); (2) reported data from research- or consumer-grade wrist-worn devices (Exposure); (3) did not require a control group, although studies including participants without anxiety were considered as comparators when applicable (Control); (4) assessed GAD or anxiety symptoms using standardized assessments (eg, self-report questionnaires or clinical diagnoses [Outcome]); and (5) used cross-sectional or longitudinal designs (Study design). Studies published in an English-language, peer-reviewed journal or thesis/dissertation, and studies that examined the relationship between anxiety and digital biomarkers via inferential statistics or machine learning methods were also eligible. Studies were excluded if they (1) focused on a sample that received diagnoses of neurodegenerative disorders or physical health conditions (however, studies were eligible if participants without these conditions were analyzed separately), and (2) primarily focused on participants with a psychiatric comorbidity, neurological disorder, or stroke (>50% of the sample).

Data Extraction

A standardized form was created to extract the following data from eligible inferential statistical studies: (1) authors and year of publication; (2) study characteristics, including type of study (ie, cross-sectional or longitudinal); (3) demographic information; (4) characteristics of digital biomarkers, including digital biomarker type and devices used for data collection; (5) anxiety measurement, and (6) data required for meta-analysis (eg, mean difference and correlation coefficients). For studies using machine learning approaches, another standardized form was created to extract the following details: (1) authors and year of publication, (2) demographic information, (3) ground truth assessment, (4) characteristics of digital biomarkers (5) machine learning task, (6) predictor features, (7) algorithms, (8) validation method, and (9) model performance.

Where necessary, authors were contacted to request additional information.

Quality Assessments

Inferential Statistical Studies

The National Heart, Lung, and Blood Institute assessment tool was used [25]. This quality assessment tool comprises 14 criteria, which are used to evaluate the validity and reliability of studies. For cross-sectional studies, 4 criteria that are relevant to longitudinal study designs are not applicable and were therefore omitted [26]. Scores were averaged, and each study was categorized as having a "Good" (≥80), "Fair" (50%-80%), or "Poor" (50%) rating, as specified in the guidelines.

Machine Learning Studies

Currently, there is no validated quality assessment tool for machine learning studies. We used a modified version of the validated Quality Assessment of Diagnostic Accuracy Studies 2 [27], which evaluates the risk of bias of studies across 4 domains (participants, index test [artificial intelligence algorithms], reference standard [ground truth], and analysis), and evaluates applicability concerns across 3 domains (excluding the analysis domain) [17].

Two independent reviewers assessed the quality of each eligible study. Any disagreements were resolved by a third reviewer.

XSL•FO
RenderX

Synthesis and Analysis

Effect sizes were combined across studies, separately for each type of digital biomarker of health, using random-effects meta-analyses. When effect sizes were unavailable, other available data were used for calculations (eg, *t* test statistic). Random-effects meta-analyses were used as they take into account between-study heterogeneity [28]. For meta-analysis based on a limited number of studies (<20 studies [29]), we applied the Hartung-Knapp-Sidik-Jonkman adjustments when calculating CIs and *P* values, which are recommended for small meta-analyses [30]. Prediction intervals for random-effects meta-analyses were also reported [31].

To ensure only 1 effect size from each included study was used in the primary meta-analysis, studies were selected based on an a priori determined hierarchy. Specifically, we prioritized (1) the largest sample size, (2) unadjusted over adjusted estimates, and (3) trait (chronic/ongoing) over state (present moment) anxiety, when more than 1 study used the same cohort. Trait anxiety refers to the general tendency to respond with anxiety across various situations, whereas state anxiety is a temporary response reflecting the present moment [32]. Since trait anxiety reflects a chronic condition with long-term implications and offers a more stable measure of general anxiety levels, it was the focus of our analysis.

Sensitivity analyses were conducted to examine whether results differed according to anxiety type (trait and state anxiety) and age group (young, middle, and older).

For each meta-analysis, heterogeneity was assessed using τ^2 (between-study variance) and I^2 (proportion of total variation in study estimates that is due to heterogeneity [33]). When there were 10 or more studies included in a meta-analysis, funnel plots would be used to assess small-study effects [34].

All analyses were conducted using the *metafor* package in R software (version 4.2.2; R Foundation for Statistical Computing).

Qualitative synthesis was conducted for studies using inferential statistical approaches that were not included in the meta-analysis. These studies were synthesized based on the type of digital biomarkers of health examined (ie, sleep metrics and activity metrics). On the other hand, studies using machine learning approaches were not meta-analyzed and were synthesized qualitatively, ordered by performance metrics (ie, accuracy, sensitivity, precision, and F_1-score).

Results

Study Selection

After the removal of duplicates, title/abstract screening, and full-text review, 365 articles were eligible (Figure 1). Of these, 47 articles focused on anxiety. Interrater reliability was high, with Cohen κ of 0.73 (substantial agreement) and 0.83 (almost perfect agreement) at the title/abstract screening and full-text review stages, respectively.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram illustrating the systematic process.

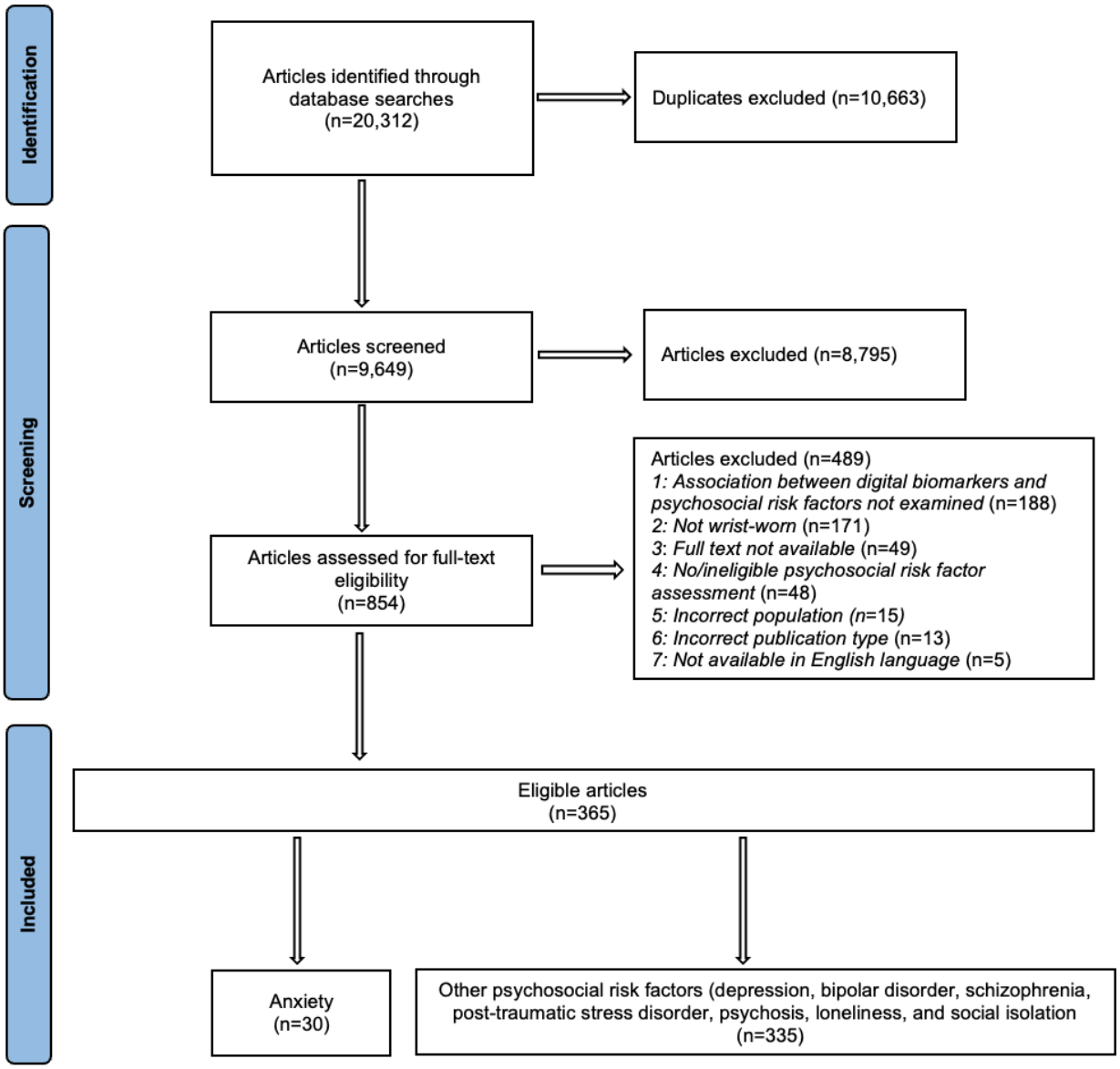

Study and Participant Characteristics

Overview

When an article reported outcomes from different cohorts or separately analyzed different populations (eg, male and female), each result was treated as a different study derived from the same article. This resulted in 44 studies from 42 articles. Of these, 36 (81.8%) studies used inferential statistical approaches for analysis, while 8 (18.2%) studies used machine learning approaches.

Studies Using Inferential Statistical Approaches

Characteristics of studies that used inferential statistical approaches are presented in Table 1. There was a total of 36 studies from 35 articles (cross-sectional: k=21, 58.3%; longitudinal: k=11, 30.6%; both cross-sectional and longitudinal: k=4, 11.1%). The study sizes ranged from 14 to 170,320 participants (median 189, IQR 79-2276). Most of the studies assessed trait anxiety or clinically diagnosed anxiety (k=19, 52.8%), 10 (27.8%) assessed state anxiety, and 7 (19.4%) assessed both state and trait anxiety. The majority of studies were conducted in young adults (k=18, 50%), 8 (22.2%) in middle-aged adults, 8 (22.2%) in older adults, and 2 (5.6%) did not report mean age data.

XSL•FO
RenderX

Table 1. Characteristics of samples eligible for systematic review and/or meta-analysis of studies using inferential statistical approaches. Most studies were cross-sectional (k=21), with sample sizes ranging from 14 to 170,320 (median 189, IQR 79-2276); most assessed trait or clinically diagnosed anxiety (k=21) and were conducted in young adults (k=18).

Study (location)	Study design (follow-up time)	Participants, N	Population type	Age (years), mean (SD) Sex (female), % Education, mean (SD) or categories % Ethnicity (White), %	Anxiety measure (type)	Wearable device (research/consumer-grade)	Digital biomarker type (subtype)
Ahn et al, 2024 (United States) [35]	Cross-sectional	33	Family caregivers	Age: 61.27 (11.2) Female: 84.8% Education: NR[a] White: 90.9%	GAD-7[b] (trait)	ActiGraph GT9X Link (research)	Sleep (SE[c])
Bullis, 2016 (United States) [36][d]	Cross-sectional	30	Clinical samples and healthy controls	Age: 28 (9.0) Female: 70% Education: high school graduate 6.7%; partial college 16.7%; college graduate 46.7%; graduate 30% White: 73.3%	Clinical diagnosis (*DSM-5*[e]—trait)	Actiwatch (research)	Sleep (SOL[f], TST[g], SE, and WASO[h])
Bullis, 2016 (United States) [36][c]	Longitudinal (14 days)	30	Clinical samples and healthy controls	Age: 28 (9.0) Female: 70% Education: high school graduate 6.7%; partial college 16.7%; college graduate 46.7%; graduate 30% White: 73.3%	Clinical diagnosis (*DSM-5*—trait) and daily symptom ratings (Ecological Momentary Assessment—state)	Actiwatch (research)	Sleep (SOL, TST, SE, and WASO)
Casanova et al, 2023 (United Kingdom) [37]	Cross-sectional	95,776	General population	Age: NR Female: NR Education: NR White: NR	GAD-7 (trait)	Axivity AX3 (research)	Activity (physical activity and SB[i])
Cox et al, 2018 (United States) [38]	Cross-sectional	138	Undergraduate students and community adults	Age: 22.5 (9.2) Female: 75.5% Education: NR White: 57.6%	Momentary anxiety (state)	ActiGraph wGT3X-BT (research)	Sleep (TST)
Cox et al, 2018 (United States) [38]	Longitudinal (9 days)	138	Undergraduate students and community adults	Age: 22.5 (9.2) Female: 75.5% Education: NR White: 57.6%	Momentary anxiety (state)	ActiGraph wGT3X-BT (research)	Sleep (TST)
D'Aurizio et al, 2023 (Italy) [39][c]	Cross-sectional	60	Prisoner and healthy control	Age: 36 (8.46) Female: 0% Education: 5-8 years White: NR	STAI[j] (trait and state)	Actiwatch Spectrum Plus (research)	Sleep (SOL, SE, TST, and WASO)
Dillon et al, 2018 (Ireland) [40]	Cross-sectional	397	Patients (population-representative random sample, from a primary care center)	Age: 59.6 (5.5) Female: 53.9% Education: primary 26.7%, secondary/tertiary 73.3% White: NR	HADS-A[k] (state)	GENEActiv accelerometer (research)	Activity (SB, LPA[l], and MVPA[m])
Doane et al, 2015 (United States) [41][c]	Cross-sectional	82	Adolescent transition to college	Age: 18.1 (0.4) Female: 76% Education: NR White: 54%	DASS-A[n] (state)	Actiwatch Score (research)	Sleep (sleep start time variability, SOL, SE, and TST)

XSL•FO
RenderX

Study (location)	Study design (follow-up time)	Participants, N	Population type	Age (years), mean (SD) Sex (female), % Education, mean (SD) or categories % Ethnicity (White), %	Anxiety measure (type)	Wearable device (research/consumer-grade)	Digital biomarker type (subtype)
Doane et al, 2015 (United States) [41][c]	Longitudinal (1 year)	Follow-up 1: 76 Follow-up 2: 71	Adolescent transition to college	NR	DASS-A (state)	Actiwatch Score (research)	Sleep (sleep start time variability, SOL, SE, and TST)
El-Sheikh et al, 2013 (United States; female) [42][c]	Cross-sectional	135	Cohabiting couples	Age: 36.5 (5.9) Female: 100% Education: NR White: NR	BAI[o] (trait)	Octagonal Basic Motionloggers (research)	Sleep (TST, SE, and SOL)
El-Sheikh et al, 2013 (United States; male) [42][c]	Cross-sectional	135	Cohabiting couples	Age: 39.4 (7.3) Female: 0% Education: NR White: NR	BAI (trait)	Octagonal Basic Motionloggers (research)	Sleep (TST, SE, and SOL)
Facer-Childs et al, 2022 (Australia) [43]	Cross-sectional	68	Elite football league athletes	Age: 23.3 (3.4) Female: 0% Education: NR White: NR	GAD-7 (trait) and DASS-A (state)	GENEActiv actigraphy devices (research)	Sleep (snooze time, SE, TST, and TIB[p])
Feng et al, 2021 (United States) [44]	Cross-sectional	14	Ophthalmology residents	Age: 30.2 (2.5) Female: 57.1% Education: NR White: NR	DASS-A (state)	Fitbit Alta HR (consumer)	Sleep (lower average sleep on call) and activity (average daily exercise)
Fuller-Rowell et al, 2024 (United States) [45][c]	Cross-sectional	874	Small towns and semirural communities	Age: 41.0 (7.0) Female: 67% Education: NR White: 67%	BAI (trait)	Octagonal Basic Motionloggers (research)	Sleep (TST, night-to-night variability, SE, and SOL)
Hofman et al, 2022 (Netherlands) [46]	Cross-sectional	1943	General population	Age: 70.9 (9.3) Female: 51.6% Education: primary 6%, lower 36.7%, further/intermediate 30.6%, higher 25.4% White: NR	HADS-A (state)	GENEActiv (research)	Sleep (TST) and activity (SB, LPA, and MVPA)
Huang et al, 2022 (China) [47]	Cross-sectional	321	College students	Age: 19.7 (1.2) Female: 55.8% Education: NR White: NR	SAS[q] (state)	Axivity AX3 (research)	Activity (SB-total, weekend, and weekdays)
Lee et al, 2021 (United States, Australia, South Korea) [48][c]	Cross-sectional	17	Elite esports athletes	Age: 20 (3.5) Female: 0% Education: NR White: NR	STAI (state)	Readiband V5 (research)	Sleep (TST, SOL, WASO, SOT[r], and WT[s])

XSL•FO
RenderX

Study (location)	Study design (follow-up time)	Participants, N	Population type	Age (years), mean (SD) Sex (female), % Education, mean (SD) or categories % Ethnicity (White), %	Anxiety measure (type)	Wearable device (research/consumer-grade)	Digital biomarker type (subtype)
Montanari et al, 2025 (France) [49]	Cross-sectional	211	Older adults	Age: 60-70 years: 105 (50%); 71-80 years: 89 (42%); >80 years: 17 (8.1%) Female: 36% Education: lower education 31%, secondary education 32%, higher education (37%) White: NR	STAI (trait and state)	Actigraph GT3X+ (research)	Sleep (TST, WASO, and SE)
Nguyen et al, 2023 (Australia) [50][c]	Cross-sectional	101	Newly recruited paramedics	Age: 26.0 (5.5) Female: 52% Education: NR White: NR	GAD-7 (trait)	Actiwatch Spectrum or Spectrum PRO (research)	Sleep (TST, WASO, SOL, and SE)
Nguyen et al, 2023 (Australia) [50][c]	Longitudinal (6 months)	101	Newly recruited paramedics	Age: 26.0 (5.5) Female: 52% Education: NR White: NR	GAD-7 (trait)	Actiwatch Spectrum or Spectrum PRO (research)	Sleep (TST, WASO, SOL, and SE)
Spira et al, 2008 (United States) [51][c]	Cross-sectional	59	Adults with primary insomnia	Age: 63.8 (6.9) Female: 74.6% Education: mean 16.3 (SD 2.3) years White: 78%	STAI (trait and state)	Actiwatch-L (research)	Sleep (WASO, SE, and SOL)
Spira et al, 2009 (United States) [52][c]	Cross-sectional	3040	Community-dwelling older women	Age: 83.6 (3.8) Female: 100% Education: mean 12.9 (SD 2.6) White: 89.3%	Goldberg Anxiety Scale (trait)	SleepWatch-O (research)	Sleep (TST, SE, time awake after sleep onset, SOL, and napping time)
Stremler et al, 2017 (Canada) [53]	Cross-sectional	118	Parents of critically ill hospitalized children	Age: 34.1 (7.2) Female: 62.7% Education: <high school 5.1%, high school 22%, university/college 53.4%, postgraduate 17.8%, unknown 1.7% White: 69.4%	STAI (state)	Octagonal Basic Motionlogger actigraphs (research)	Sleep (TST, number of nocturnal awakenings, and individual sleep variability)
Swanson et al, 2023 (United States) [54][c]	Cross-sectional	1197	Postmenopausal women	Age: 65.5 (2.6) Female: 100% Education: less than high school 5.2%, high school 15.3%, some college 30.7%, college 22.6%, post-college 26.2% White: 45.7%	GAD-7 (trait)	Actiwatch-2 (research)	Sleep (sleep irregularity and sleep midpoint outside 2-4 AM)
Uchida and Kurosawa, 2025 (Japan) [55]	Cross-sectional	50	University students	Age: 20.58 (1.34) Female: 64% Education: NR White: NR	Japanese version of STAI (trait and state)	Fitbit Charge 5 (consumer)	Sleep (TST, bedtime, and wake time)
Windmill et al, 2024 [56]	Cross-sectional	83	General population	Age: 24 (5.1) Female: 65.7% Education: NR White: NR	PANAS[t] (state and trait), STAI (state and trait), BAI (state)	GENEActiv, ActivInsights (research)	Sleep (SE and TST)

XSL•FO
RenderX

Study (location)	Study design (follow-up time)	Participants, N	Population type	Age (years), mean (SD) Sex (female), % Education, mean (SD) or categories % Ethnicity (White), %	Anxiety measure (type)	Wearable device (research/consumer-grade)	Digital biomarker type (subtype)
Walters et al, 2020 (Australia) [57]	Cross-sectional	110	Without sleep disorders	Age: 32.3 (12.4) Female: 50.9% Education: NR White: NR	BAI (trait)	Respironics Actiwatch Spectrum Pro (research)	Sleep (SE)
Zhang et al, 2025 (United States) [58]	Cross-sectional	10,129	General population	Age: median 48.0 (IQR 36.0-58.0) Female: 69.3% Education: NR White: 91.4%	GAD-7 (trait)	Fitbit (consumer)	Sleep (eg, TST, wake time, sleep start, percentage of long sleep days, and percentage of short sleep days) Physical activity (eg, steps, sedentary duration, light activity duration, and total active duration) Heart rate metrics (eg, average heart rate and SD)
Zheng et al, 2024 (United States) [59]	Longitudinal (4.5 years)	6785	General population	Age: median 50.2 (IQR 35.7-61.5) Female: 70.8% Education: NR White: 83.6%	Electronic health records	Fitbit (consumer)	Sleep (TST, sleep stages, and sleep regularity)
Bai et al, 2024 (United States) [60]	Longitudinal (16 weeks)	167	University students	Age: NR Female: 79% Education: NR White: 92.2%	Daily survey measuring anxiety	Apple Watch (consumer)	Physical activity (MVPA and steps)
Jo et al, 2024 (Korea) [61]	Longitudinal (4 weeks)	47	Young adults	Age: 28.7 (5.9) Female: 48.9% Education: NR White: NR	GAD-7 (trait)	Samsung Galaxy Active 2 (consumer)	HRV[u] (time-domain measures: RMSSD[v], SDNN[w], SDSD[x], and PNN50[y]; frequency-domain measures: LF[z], HF[aa], and LF/HF[ab])
Kandola et al, 2021 (United Kingdom) [62]	Longitudinal (2 years)	60,235	General population	Age: 55.9 (7.7) Female: 56% Education: college/university degree 48%; a-level/as levels/equivalent 14%; GCSEs[ac]/O-levels/equivalent 19%; CSEs[ad]/equivalent 3.4%; NVQ[ae]/HND[af]/equivalent 4.9%; other professional qualifications 4.9%; none (6.1%) White: 98%	GAD-7 (trait)	Axivity AX3 (research)	Activity (SB, LPA, and MVPA)

XSL•FO
RenderX

Study (location)	Study design (follow-up time)	Participants, N	Population type	Age (years), mean (SD) Sex (female), % Education, mean (SD) or categories % Ethnicity (White), %	Anxiety measure (type)	Wearable device (research/consumer-grade)	Digital biomarker type (subtype)
Liu et al, 2025 (China) [63]	Longitudinal (7-9 years)	84,570	General population	Age: 56.2 (7.8) Female: 55.5% Education: college/university degree 34.8%, other levels of education (64%), unknown (1.2%) White: 96.8%	Patient Health Questionnaire, *ICD-10*[ag] codes (trait)	Axivity AX3 (research)	Physical activity (MVPA)
McDuff et al, 2025 (United States) [64]	Longitudinal (4 weeks)	237	General population	Age: 45.1 (11.8) Female: 67.5% Education: NR White: 87.7%	GAD-7 (trait)	Fitbit Sense 2TM (consumer)	Skin conductance level, heart rate, skin temperature, and step count
Okawara et al, 2024 (Japan) [65]	Longitudinal (60 days)	279	Office workers	Age: 42.0 (10.0) Female: 16.1% Education: NR White: NR	STAI trait subscale (trait)	Apple Watch (consumer)	HRV (HF, LF, LF/HF ratio, and LnccL/H[ah])
Presby et al, 2025 (Switzerland, United States, Australia) [66]	Longitudinal (13 months)	170,320	General population	Age: 37.39 (10.36) Female: 33% Education: NR White: 33.4%	Generalized Anxiety Disorder 2-item (trait)	WHOOP strap version 3.0 and 4.0 (consumer)	Sleep (TST, SE, SD of TST, sleep consistency, sleep start, and wake start) Heart rate (HRV and resting heart rate) Physical activity (total activity)
Straus et al, 2023 (United States) [67]	Longitudinal (8 weeks)	2021	Trauma survivors from the emergency department	Age: 35.8 (13.0) Female: 62.2% Education: <high school 12%, high school graduate 25.6%, some college 41.3%, college graduate 20.7% White: 34.3%	Smartphone-based questionnaires (state—measured 6 times over the 8-week period)	Verily Life Sciences (research)	Sleep (number of transitions between sleep and wake)
Wang et al, 2025 (China) [68]	Longitudinal (12.6 years)	91,800	General population	Age: 61.8 (7.9) Female: 56.1% Education: college above 43.7%, college below 56.3% White: 97%	*ICD-10* (trait)	Axivity AX3 (research)	Physical activity (LPA, moderate intensity, vigorous intensity, and MVPA)

XSL•FO
RenderX

Study (location)	Study design (follow-up time)	Participants, N	Population type	Age (years), mean (SD) Sex (female), % Education, mean (SD) or categories % Ethnicity (White), %	Anxiety measure (type)	Wearable device (research/consumer-grade)	Digital biomarker type (subtype)
Yu et al, 2025 (Japan) [69]	Longitudinal (2 years)	71,556	Healthy adults aged 40-69 years	Age: 62.11 (7.83) Female: 54.4% Education: college or higher 44.4% White: 97.3%	GAD-7 (trait)	Axivity AX3 (research)	Physical activity (MVPA, LPA, and SB)

[a]NR: not reported.
[b]GAD-7: Generalized Anxiety Disorder-7.
[c]SE: sleep efficiency.
[d]Studies included in meta-analyses.
eDSM-5: Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition).
[f]SOL: sleep onset latency.
[g]TST: total sleep time.
[h]WASO: wake after sleep onset.
[i]SB: sedentary behavior.
[j]STAI: State-Trait Anxiety Inventory.
[k]HADS-A: Hospital Anxiety and Depression Scale-Anxiety section.
[l]LPA: light-intensity physical activity.
[m]MVPA: moderate to vigorous physical activity.
[n]DASS-A: Depression, Anxiety, and Stress Scale-Anxiety Subscale.
[o]BAI: Beck Anxiety Inventory.
[p]TIB: time in bed.
[q]SAS: Self-Rating Anxiety Scale.
[r]SOT: sleep onset time.
[s]WT: wake time.
[t]PANAS: Positive and Negative Affect Schedule.
[u]HRV: heart rate variability.
[v]RMSSD: root mean square of successive differences between RR intervals (intervals between successive R peaks) interval differences.
[w]SDNN: SD of NN intervals (intervals between successive normal heartbeats).
[x]SDSD: SD of RR interval intervals.
[y]PNN50: percentage of successive RR intervals that differ by more than 50 ms.
[z]LF: absolute power of the low-frequency band.
[aa]HF: absolute power of the high-frequency band.
[ab]LF/HF: ratio of LF to HF power.
[ac]GCSE: General Certificate of Secondary Education.
[ad]CSES: Certificate of Secondary Education.
[ae]NVQ: National Vocational Qualification.
[af]HND: Higher National Diploma.
agICD-10: International Statistical Classification of Diseases, Tenth Revision.
[ah]LnccL/H: log-transformed coefficient of component variance of the low-frequency component to high-frequency component power ratio.

Most studies assessed only sleep (k=21, 58.3%), 8 (22.2%) studies assessed only physical activity, 2 (5.6%) studies assessed only HRV, and 5 (13.9%) studies assessed more than one type. The majority of sleep studies assessed the following 4 sleep metrics: total sleep time (TST; ie, total time spent in bed), wake after sleep onset (WASO; ie, time spent awake after initially falling asleep), SE (ie, ratio of TST to the total time spent in bed), and sleep onset latency (SOL; ie, time it takes to fall asleep after going to bed). Regarding physical activity, sedentary behavior, light physical activity, and moderate to vigorous physical activity (MVPA) were most commonly assessed.

Studies Using Machine Learning Approaches

Characteristics of studies using machine learning approaches are presented in Table 2. A total of 8 studies from 7 articles were included. The study sizes ranged from 40 to 727 participants (median 326, IQR 66-410). All studies used algorithms trained with a labeled dataset to select predictors for predicting anxiety (ie, supervised learning). Included studies used a variety of digital biomarkers of health as predictors in their models, for example, HRV, sleep, and activity.

XSL•FO
RenderX

Table 2. Characteristics of samples for studies using machine learning approaches. All studies used supervised learning, with sample sizes ranging from 40 to 727 (median 326, IQR 66-410).

Study	Population type	Participants, N	Age (years), mean (SD) Sex (female), %	Ground truth assessment, anxiety measure (type)	Wearable device (consumer or research-grade)	Machine learning task	Predictors	Algorithms	Validation method
Coutts et al, 2020 (United Kingdom) Trial 1 [70]	University students	68	Age: 21 (NR[a]) Female: 64%	DASS-A[b] (state) and STAI[c] (trait and state)	Biobeam band (research)	Binary classification (supervised)	Heart rate variability	Deep neural networks (LSTM[d])	Train-test split: training (80%), test (10%), validation (10%)
Coutts et al, 2020 (United Kingdom) Trial 2 [70]	University students	584	Age: NR Female: NR	DASS-A (state) and STAI (trait and state)	Biobeam band (research)	Binary classification (supervised)	Heart rate variability	Deep neural networks (LSTM)	Train-test split: training (80%), test (10%), validation (10%)
Fukuda et al, 2020 (Japan) [71]	Office workers	60	Age: NR Female: NR	DAMS[e] (state)	Fitbit Charge 3 (Consumer)	Binary classification (supervised)	Sleep actigraphy data (13 features, eg, total sleep time and number of wake)	RF[f]	Leave-one-person-out cross-validation
He et al (2025) [72]	University students	40	Age: 21.7 (3.11) Female: 52.5%	STAI Y6 (state)	E4 Empatica Hexoskin smart shirt	Binary classification (unsupervised)	Wrist-worn wearable: skin temprature and EDA[g] Smart shirt: ECG[h]	SVM[i], RF, KNN[j], naïve Bayes	Leave-one-participant-out
Lee et al, 2022 (Korea) [73]	Older adults with mild cognitive impairment	352	Age: 72.48 (5.9) Female: 73%	Clinical diagnosis (Trait)	Fitbit Alta HR2 (Consumer)	Binary classification (supervised)	Model 1: 24-hour activity rhythms and sleep pattern[k] Model 2: Model 1+ minimal K-GAI[l] (5 items)	Logistic regression, SVM, RF, GBM[m]	10-fold stratified cross-validation
Lee et al, 2024 (Korea) [74]	Older adults with mild cognitive impairment	352	Age: 72.5 (5.9) Female: 73%	Clinical diagnosis (Trait)	Fitbit Alta HR2 (Consumer)	Binary classification (supervised)	Model 1: activity, sleep Model 2: activity + minimal K-GAI	Convolutional neural network, LSTM, residual network	Cross-validation
Saylam et al, 2023 (United States) [75]	Office workers	727	Age: NR Female: NR	Single item[n] (state)	Garmin smartwatch (consumer)	Regression	Activity, stress, sleep, heart rate	RF, XGBoost[o], LSTM	80% training, 20% testing
Saylam et al, 2024 (United States) [76]	College students	Baseline: ~700 Follow-up: 300	Age: NR Female: NR	BAI[p]	Fitbit (consumer)	Regression	Activity and sleep	RF, XGBoost, LSTM	Training: 5 semesters Validation: 2 semesters Testing: 2 semesters

XSL•FO
RenderX

[a]NR: not reported.

[b]DASS-A: Depression, Anxiety, and Stress Scale-Anxiety Subscale.

[c]STAI: State Trait Anxiety Inventory

[d]LSTM: Long Short-Term Memory networks.

[e]DAMS: Depression and Anxiety Mood Scale.

[f]RF: random forest.

[g]EDA: electrodermal activity.

[h]ECG: electrocardiogram.

[i]SVM: Support Vector Machine.

[j]KNN: k-nearest neighbors.

[k]Model 1: 24-hour activity rhythms (interdaily stability [the stability of an activity rhythm of a daily pattern], intra-daily variability [the variability of the activity rhythms throughout the day], dominant rest phase onset [the start time of the 5-hour period with the least activity within 24 hours], and sleep patterns (total sleep time, sleep onset latency, wake after sleep onset, and sleep quality).

[l]K-GAI: Korean version of Geriatric Anxiety Inventory.

[m]GBM: gradient boosting machine.

[n]Single-item question "Please select the response that shows how anxious you feel at the moment. Response scale: 1 (not at all anxious), 2 (a little anxious), 3 (moderately anxious), 4 (very anxious), and 5 (extremely anxious)."

[o]XGBoost: extreme gradient boosting.

[p]BAI: Beck Anxiety Inventory.

Quality Assessment

Studies Using Inferential Statistical Approaches

For studies that conducted both cross-sectional and longitudinal analyses, quality assessment was based on the longitudinal studies. Of the 36 studies, 12 (33.3%) received a "Good" quality rating; the remainder received a "Fair" rating (k=24, 66.7%; Multimedia Appendix 5).

Studies Using Machine Learning Approaches

The quality assessment tool for studies using machine learning approaches assesses the risk of bias across 4 domains: participants, index test, reference standard, and analysis. A total of 5 studies were rated as having a high risk of bias in the participants' domain. For the index test, ground truth, and analysis, all studies were rated as either low risk or unclear. There were either unclear or low concerns regarding the applicability of the studies to our review question across the participants, index test, and reference standard domains (Multimedia Appendix 5 [70-76]).

Qualitative Synthesis of Inferential Statistical Studies

Sleep

A total of 8 studies investigated the association between sleep metrics not included in the meta-analyses and anxiety. Sleep start variability [41], sleep onset time [48], wake time [48,55,58], bedtime [55,58], number of nocturnal awakenings [53], sleep variability [53], and sleep timing [54] were not associated with anxiety symptoms. Lower average sleep among doctors who were on call was negatively associated with anxiety symptoms [44]. Other studies found that greater sleep irregularity [54], greater sleep duration variability [45,58], higher sleep start variability [58], greater wake time variability [58], later wake time [58], and a higher percentage of long (>10 hours) and short (<6 hours) sleep days [58] were associated with more anxiety symptoms.

A total of 7 studies reported the longitudinal association between sleep and anxiety [36,38,41,50,59,66,67]. While 1 study found no association between baseline sleep metrics (ie, SE, WASO, SOL, and TST) and anxiety at 6 months follow-up in paramedics [50], the majority of studies did observe an association over varying periods of time. Lower TST compared to an individual's personal TST average was associated with higher anxiety levels on the next day [38]. Anxious mood also predicted next day's poorer sleep (ie, lower SE, longer WASO, and less TST) in individuals with clinical anxiety, but not in healthy controls [30]. Worsening sleep continuity (measured by the number of transitions between sleep and wake) was associated with worsening anxiety over an 8-week period [67], lower SE and longer SOL were associated with more anxiety symptoms at 1 year follow-up [41], and sleep patterns (increased sleep irregularity, decreased deep and rapid eye movement (REM) sleep percentage, and increased light sleep percentage) captured through continuous monitoring over a median period of 4.5 (IQR 2.5-6.5) years were associated with increased odds of GAD [59]. Additionally, they identified a nonlinear, J-shaped relationship between average daily sleep duration and GAD—compared with the median average daily sleep of 6.8 hours, individuals with an average daily sleep duration of 5 hours or 10 hours had increased odds of GAD. One study found that more consistent wake and sleep times were associated with lower anxiety symptoms [66].

Activity

A total of 6 studies investigated the association between sedentary behavior and anxiety [37,40,46,47,58,62]. While 1 study found no association between sedentary behavior and anxiety symptoms [47], the remainder did observe associations. Individuals with moderate to severe anxiety symptoms in primary care settings spent more time in sedentary activity per day than participants with no significant anxiety [40]. This finding was supported by a large study in the UK general population [58]. A study using isotemporal substitution models (which assess the effect of hypothetically replacing 1 activity with another while keeping the total time constant) found similar

XSL•FO
RenderX

findings; they found that replacing MVPA with sedentary behavior was associated with higher odds of having clinically relevant anxiety symptoms [46]. Another study found that while sedentary time was not associated with current GAD and anxiety symptoms [37], it was associated with lower odds of lifetime GAD, and in males when stratified by sex [37]. The association between sedentary behavior and anxiety symptoms was also supported in a longitudinal study, with more baseline sedentary behavior associated with more anxiety symptoms at follow-up [62].

A total of 5 studies examined the association between light activity, MVPA, and anxiety [40,46,58,60,62] and reported mixed findings. Regarding light physical activity, 1 study found that lower light activity duration was associated with more anxiety symptoms. Consistent with this finding, 1 study found that individuals with moderate to severe anxiety symptoms spent less time in light physical activity than those without significant levels of anxiety [40]. Using isotemporal substitution models, 1 study found that replacing sedentary behavior with light physical activity was associated with a decrease in anxiety symptoms [40], while another study found no association [46], and a longitudinal study (UK Biobank) found that replacing sedentary behavior with light physical activity was associated with lower anxiety symptoms at follow-up [62]. Regarding MVPA, 1 study found that replacing sedentary behavior with MVPA was associated with lower odds of clinically relevant anxiety symptoms [46], but 1 study found no association [40]. One longitudinal study found that replacing baseline sedentary behavior with MVPA was associated with lower anxiety symptoms at follow-up [62], and another found that more anxiety symptoms (measured using a daily survey) were associated with lower MVPA [60].

A total of 3 studies examined associations between physical activity and anxiety using data from the UK Biobank [63,68,69]. One study reported an L-shaped association, where higher levels of moderate-intensity physical activity and vigorous-intensity physical activity were associated with lower incident anxiety [68]. Another study also identified a nonlinear relationship, with higher levels of MVPA and total physical activity associated with reduced anxiety risk, whereas light physical activity and sedentary behavior were not associated with anxiety [69]. The third study found that, compared with inactive individuals, those who were regularly active during the week or active only during weekends had lower risks of anxiety [63].

A total of 3 studies examined the associations between step count and anxiety. One study found that lower daily step count and lower SD of daily step count were associated with more anxiety symptoms [58]. A longitudinal study found that higher step count was associated with lower anxiety symptoms [60]; another longitudinal study found no association [64].

Of the 5 studies investigating other activity metrics (average daily exercise duration, total sedentary time on weekdays and weekends, and overall physical activity) [37,44,47,58,66], only 2 did not find associations with anxiety symptoms [44,47]. More physical activity was associated with lower odds of current and lifetime GAD [37] and lower anxiety symptoms [37,66].

Heart Rate

A total of 5 studies investigated the association between heart rate metrics and anxiety symptoms [58,61,64-66]. Four studies investigated the association between HRV metrics and anxiety [61,64-66]. HRV metrics can be examined in either the time-domain (measuring the amount of variability in intervals between heartbeats over time) or the frequency domain (measuring how HRV is distributed across different frequency bands). Within the frequency domain, the high-frequency component has been linked with parasympathetic nervous system (PNS) activation, and the low-frequency component has been linked with sympathetic nervous system (SNS) activation. Both central tendency and variability of metrics can be assessed in the HRV time and frequency domains.

One study focusing on the frequency domain found that, compared to a low trait anxiety group, the high trait anxiety group showed greater variability in HRV frequency metrics associated with PNS activation and the consistency of this activation, SNS activation, and the balance of PNS and SNS activation (eg, low-frequency/high-frequency HRV) in addition to daily changes of this balance [65]. Additionally, a higher mean and greater variance in daily changes in the balance of PNS and SNS activation were associated with increased odds of having trait anxiety [65]. In contrast, a separate study observed that higher mean low frequency (SNS activation) was associated with lower anxiety scores, and that mean high-frequency HRV (PNS activation) and the low frequency/high frequency ratio were not associated with anxiety symptoms [61]. This study also examined the association between HRV time-domain metrics and anxiety at 2 different time points, and found that higher mean deviation from the average HRV and short-term variations in heart rate were associated with lower anxiety scores [61]. One study found no association between HRV root mean square of successive differences and anxiety, a time-domain HRV metric that measures beat-to-beat changes in heart rate [64]. A study examined HRV but did not specify whether it was calculated in the time or frequency domain and found that higher HRV was associated with lower anxiety symptoms [66].

A total of 2 studies examined associations between average heart rate and anxiety, all reporting consistent findings. One study found that higher average heart rate and lower SD of heart rate were associated with more anxiety symptoms [58]. Another study found that individuals with anxiety had elevated heart rate compared with those without [64]. One study examined resting heart rate and found that higher levels were associated with more anxiety symptoms [66].

Other Metrics

One study found that individuals who were anxious had an elevated skin conductance response and skin temperature compared with those without anxiety [64].

Qualitative Synthesis of Machine Learning Studies

Of the 8 studies included in this review, some studies used multiple algorithms within the same study, each with multiple performance metrics. Accuracy measures correct classifications (ie, presence/absence of anxiety). Sensitivity is the ability to

XSL•FO
RenderX

detect true positives (ie, correct identification of individuals with anxiety), while precision is the correct classification among predicted positives (ie, how many individuals classified as having anxiety actually have anxiety). F_1-score balances precision and sensitivity, and mean absolute error (MAE) measures the average prediction error, with lower MAE signifying better performance.

A total of 3 studies reported 13 accuracy estimates, ranging from 56.3% to 94.6% [70,71,73]. Two studies reported 9 sensitivity estimates (range 56.3%-94.6%), 9 precision estimates (range 56.3%-94.8%), and 9 F_1-score estimates (range 55.9%-94.6%) [71,73]. Overall, 56 MAE estimates were reported from 2 studies [75,76], ranging from 0.05 to 0.63.

One article using 2 cohorts found better accuracy in predicting anxiety using frequency-domain HRV and nighttime HRV (range 63.9%-69.4%) data than time-domain HRV and daytime (range 63.2%-63.4%) data [70]. They also found that nighttime HRV data (maximum accuracy: 69.4%) performed better than daytime data (maximum accuracy: 63.2%) [70].

One study with older adults who had mild cognitive impairment found that combining activity (24-hour activity rhythm), sleep, and anxiety questionnaire items improved model performance across accuracy (91.0%-94.6% vs 56.3%-59.9%), precision (91.4%-94.8% vs 56.4%-61.0%), recall (91.0%-94.6% vs 56.3%-59.9%), and F_1-scores (90.9%-94.6% vs 55.9%-59.0%) compared to models that excluded anxiety questionnaire data [73]. Similarly, another study using the same sample found that combining activity (step count) and sleep stages achieved a predictive success rate for participants without anxiety (98%) but had a low predictive success rate identifying those with anxiety (18%) [74]. Models including anxiety questionnaire items improved the predictive success rate for identifying participants with anxiety (82%) and maintained a good predictive success rate at identifying those without anxiety (95%). These studies did not report feature importance in their models.

A study with university students used digital biomarkers of health—HRV, electrodermal activity, and skin temperature—collected from wrist-worn wearables and smart shirts to predict low versus high anxiety levels. The best classifier (support vector machine) achieved an overall accuracy of 90.9%, with a precision of 0.90-0.92, recall of 0.90-0.92, and F_1-score of 0.91 [72].

Two studies predicted anxiety in office workers [71,75]. One study, using 13 sleep metrics, achieved an F_1-score of 75.6% and identified that the top 5 important predictors were REM sleep time ratio, REM sleep minutes, light sleep time ratio, deep sleep time ratio, and total wake time ratio [71]. Another study, using data from wearable devices (activity, stress, sleep, and heart rate data) and smartphones, found that the best non–time-based model (ie, without incorporating temporal data) predicted anxiety with an MAE of 0.6316. Time-based models showed improved performance, with an MAE of 0.3729 when using data from the previous 3 days to predict the next day's anxiety [75]. They found that the top 20 predictors included measures of sleep (bedtime, wake time, and sleep duration), stress (low stress duration), and activity (highly active duration).

Finally, 1 study comparing 4 modeling approaches—conventional, multitask learning approach (model trained to solve multiple tasks simultaneously), time-based approach, and time-based multitask approach—found that time-based models using data from the previous 15 days to predict next day's anxiety achieved the lowest MAE (0.0047), followed by the time-based multitask model (MAE 0.0083), the conventional model (MAE 0.1277), and the multitask model (MAE 0.1347) [76]. While activity, sleep, and heart rate were considered, the top 20 predictors did not include any of these metrics.

Quantitative Synthesis of Cross-Sectional Studies

Random-effects meta-analyses were performed for studies that provided combinable effect sizes. All studies were cross-sectional and assessed associations between sleep metrics and anxiety symptoms.

No association between SOL and anxiety symptoms was observed (k=9, N=3643; Fisher z=0.04, 95% CI –0.07 to 0.15; P=.08; Figure 2 [36,39,41,42,48,50-52] and Multimedia Appendix 6), with evidence of moderate to substantial heterogeneity between studies (I^2=49.77%). The prediction interval ranged from –0.19 to 0.27. Around half of the included studies were rated as fair quality (k=5, 55.6%). The results remained unchanged in sensitivity analyses based on anxiety type (trait and state) and young adults (Multimedia Appendix 6). One study was not eligible for meta-analysis but found that higher SOL was associated with more anxiety symptoms [45].

XSL•FO
RenderX

Figure 2. Forest plot of the associations between sleep onset latency and anxiety symptoms. Effect sizes are Fisher z with corresponding 95% CIs. No association between sleep onset latency and anxiety symptoms was observed (k=9, N=3643; Fisher z=0.04, 95% CI –0.07 to 0.15; *P*=.08). ANX: Goldberg Anxiety Scale; BAI: Beck Anxiety Inventory; DASS: Depression, Anxiety, and Stress Scale-Anxiety Subscale; GAD-7: Generalized Anxiety Disorder-7; HKSJ: Hartung-Knapp-Sidik-Jonkman; STAI: State-Trait Anxiety Inventory [36,39,41,42,48,50,51,52].

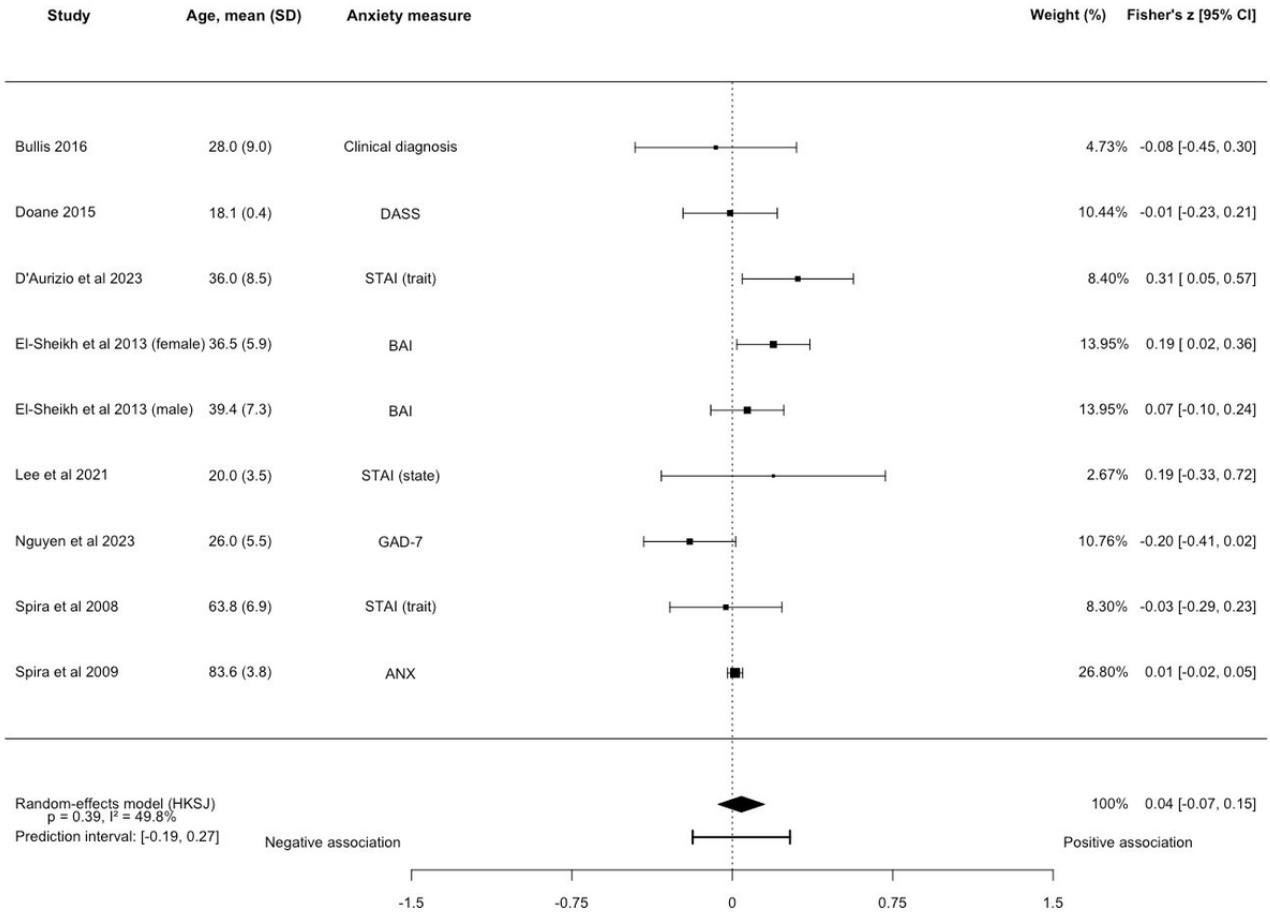

No association between TST and anxiety symptoms was observed (k=12, N=14,721; Fisher *z*=0.009, 95% CI –0.01 to 0.03; *P*=.28; Figure 3 [36,39,41,42,45,48,50,52,55,56,58] and Multimedia Appendix 6), with very low heterogeneity between studies (I^2=3.89%). The prediction interval ranged from –0.02 to 0.03. The majority of the studies were rated as fair quality (k=9, 75%). The results remained unchanged in sensitivity analyses on anxiety type and young adults (Multimedia Appendix 6). Five studies were not eligible for quantitative synthesis. Findings from 4 studies aligned with those from the meta-analysis [38,46,53,66]. One study found a positive association between poor sleep—as measured by a composite including TST and other sleep metrics—and anxiety symptoms [43]. One study reported no association [49].

XSL•FO
RenderX

Figure 3. Forest plot of the association between total sleep time and anxiety symptoms. Effect sizes are Fisher z with corresponding 95% CIs. No association between total sleep time and anxiety symptoms was observed (k=12, N=14,721; Fisher z=0.009, 95% CI –0.01 to 0.03; *P*=.28). ANX: Goldberg Anxiety Scale; BAI: Beck Anxiety Inventory; DASS: Depression, Anxiety, and Stress Scale-Anxiety Subscale; GAD-7: Generalized Anxiety Disorder-7; HKSJ: Hartung-Knapp-Sidik-Jonkman; STAI: State-Trait Anxiety Inventory [36,39,41,42,45,48,50,52,55,56,58].

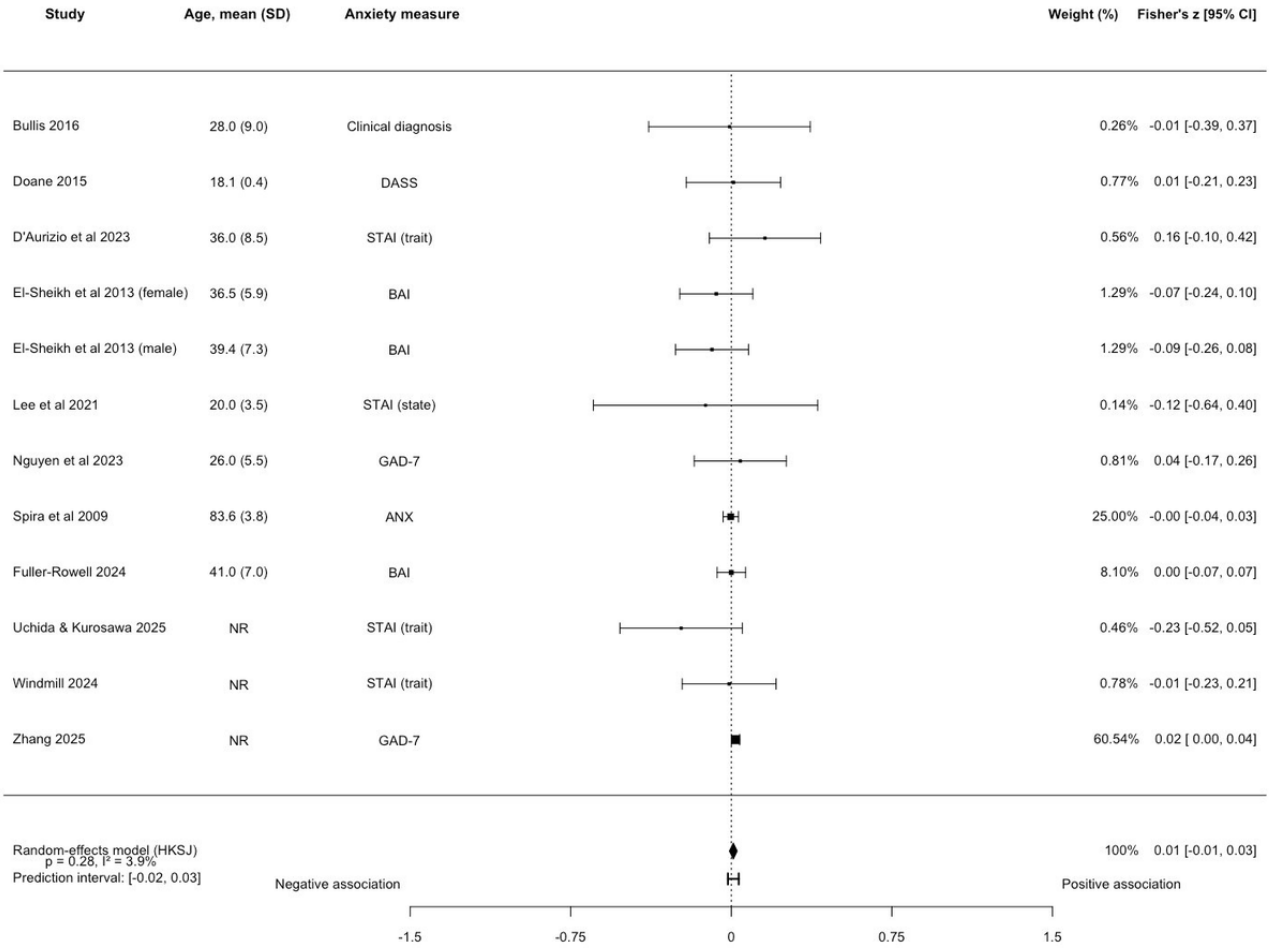

No association between SE and anxiety symptoms was observed (k=9, N=3710; Fisher *z*=–0.07, 95% CI –0.14 to 0.002; *P*=.06; Figure 4 [36,39,41,42,50-52,56] and Multimedia Appendix 6), with low heterogeneity between studies (I^2=20.97%). The prediction interval ranged from –0.19 to 0.05. The majority of studies were rated as fair quality (k=7, 77.8%). These results remained unchanged in sensitivity analyses on anxiety type and young adults (Multimedia Appendix 6). Of the 5 studies ineligible for quantitative synthesis, 3 found no association between SE and anxiety symptoms [45,49,57], 1 found a positive association [66], while 2 studies reported negative associations [43,77].

XSL•FO
RenderX

Figure 4. Forest plot of the association between sleep efficiency and anxiety symptoms. Effect sizes are Fisher z with corresponding 95% CIs. No association between sleep efficiency and anxiety symptoms (k=9, N=3710; Fisher z=-0.07, 95% CI –0.14 to 0.002; *P*=.06). BAI: Beck Anxiety Inventory; DASS: Depression, Anxiety, and Stress Scale-Anxiety Subscale; GAD-7: Generalized Anxiety Disorder-7; HKSJ: Hartung-Knapp-Sidik-Jonkman; STAI: State-Trait Anxiety Inventory [36,39,41,42,50,51,52,56].

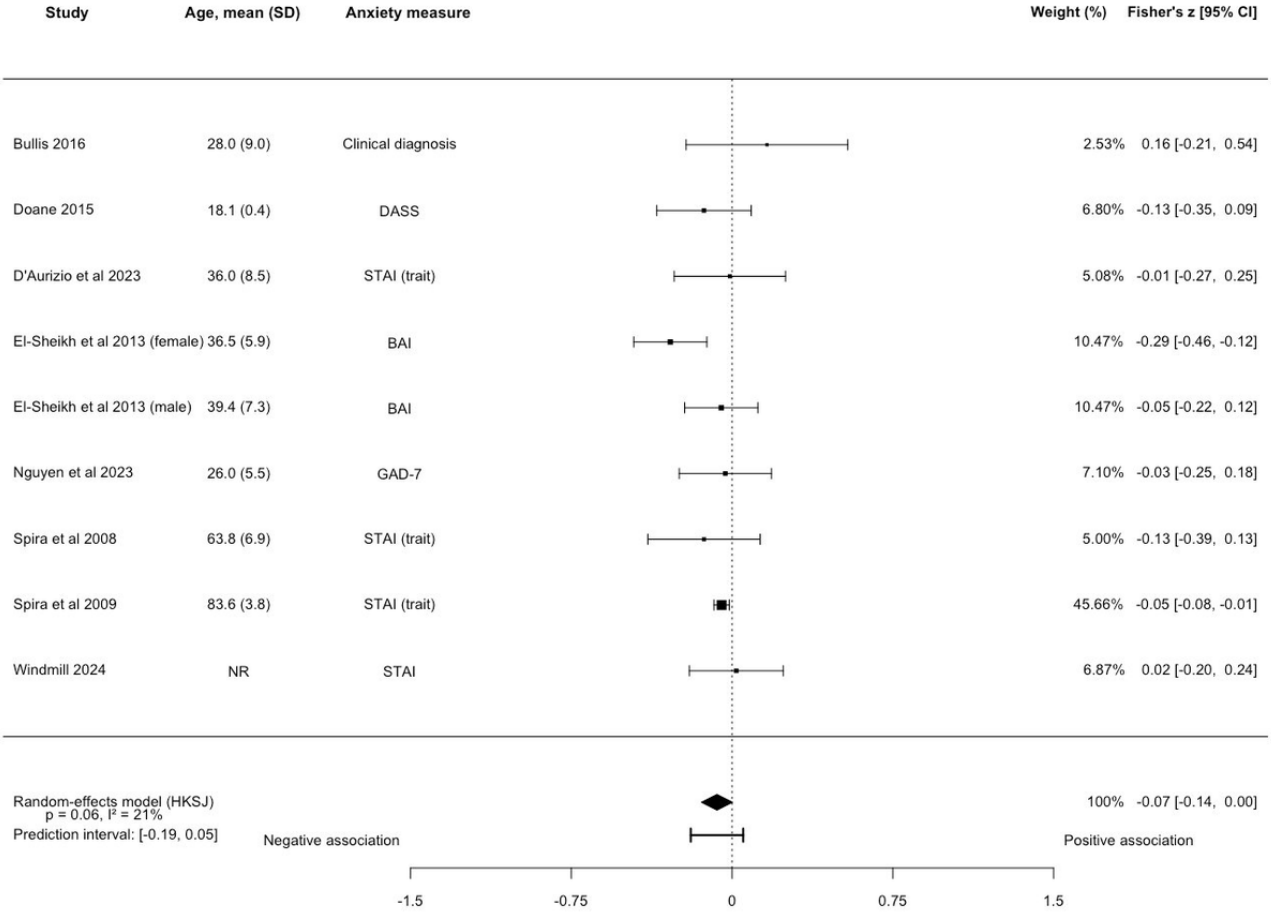

No association between WASO and anxiety symptoms was observed (k=6, N=3291; Fisher *z*=0.13, 95% CI –0.04 to 0.30; *P*=.11; Figure 5 [36,39,48,50-52] and Multimedia Appendix 6), with low to moderate heterogeneity between studies (I^2=40.62%). The prediction interval ranged from –0.15 to 0.41. Most studies were rated as fair quality (k=4, 66.7%). These results remained unchanged in sensitivity analyses on anxiety type and young adults (Multimedia Appendix 6). One study that was not eligible for quantitative synthesis also found no association [49].

XSL•FO
RenderX

Figure 5. Forest plot of the association between wake after sleep onset and anxiety symptoms. Effect sizes are Fisher z with corresponding 95% CIs. No association between wake after sleep onset and more anxiety symptoms (k=6, N=3291; Fisher z=0.13, 95% CI –0.04 to 0.30; *P*=.11). ANX: Goldberg Anxiety Scale; GAD-7: Generalized Anxiety Disorder-7; HKSJ: Hartung-Knapp-Sidik-Jonkman; STAI: State-Trait Anxiety Inventory [36,39,48,50,51,52].

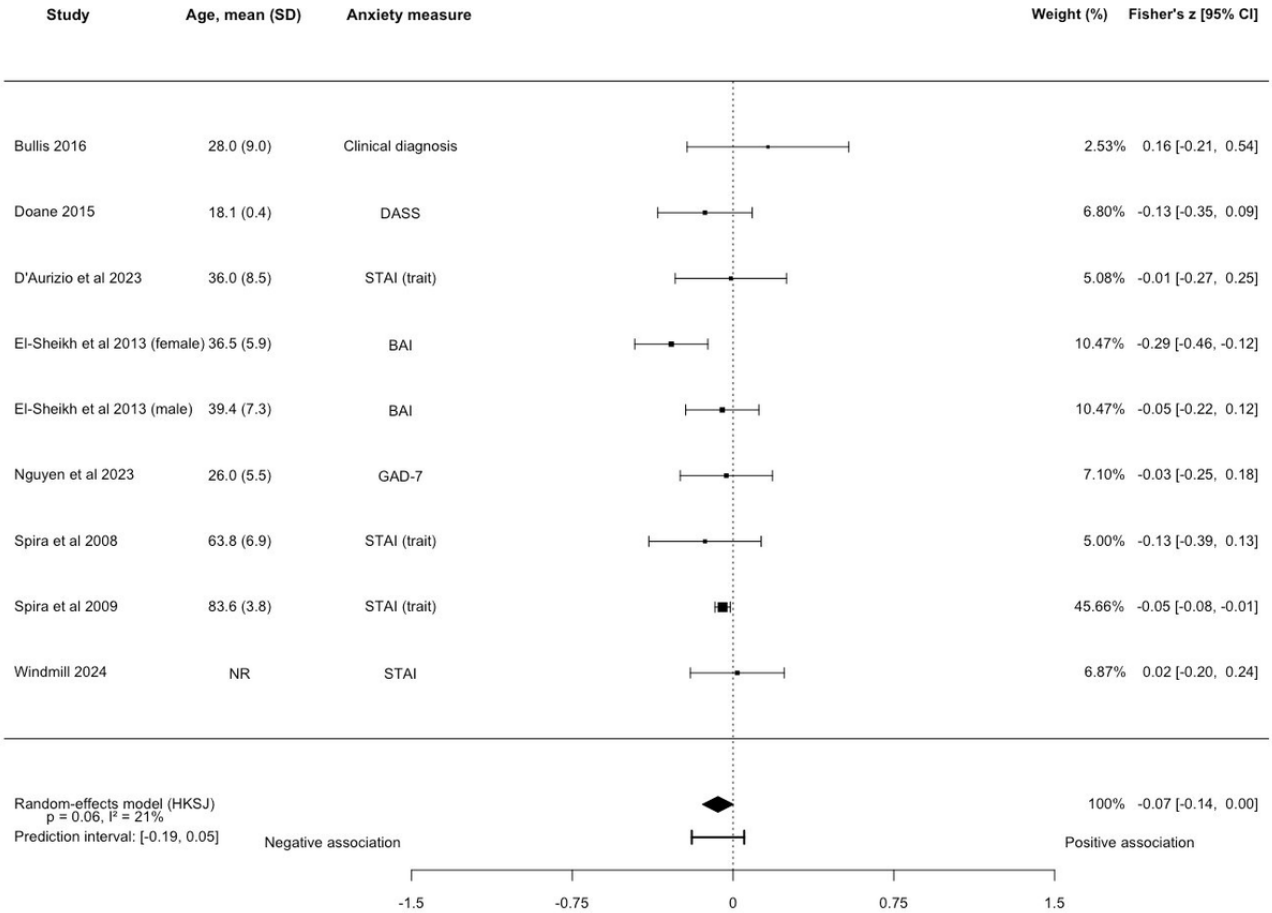

The Hartung-Knapp-Sidik-Jonkman–adjusted CIs reflect uncertainty around the average associations across studies, whereas the prediction intervals estimate the range of effects that may be observed in future studies conducted in different populations or settings. Across the sleep metrics examined, prediction intervals were generally similar to, but slightly wider than, the CIs, suggesting that effects observed in future studies are likely to be close to the average pooled estimates.

Discussion

Overview

This review aimed to advance scientific understanding of the association between digital biomarkers of health collected from wrist-worn devices and anxiety. Specific interest centered on determining which digital biomarkers of health have been the focus of existing research, and which demonstrate the strongest evidence of an association with anxiety. A total of 36 studies were included in this systematic review. Meta-analyses were conducted for 4 sleep metrics (SE, WASO, TST, and SOL), and none of these metrics were associated with anxiety. It was not possible to conduct meta-analyses for physical activity and heart rate metrics; however, qualitative synthesis suggests that lower physical activity levels and elevated heart rate were associated with higher levels of anxiety.

Sleep

Our meta-analytic findings showed that SE, WASO, TST, and SOL were not statistically associated with anxiety. However, the association for SE was close to the significance threshold, suggesting a possible relationship. Heterogeneity was low for TST and SE, indicating that studies included in these meta-analyses found relatively consistent findings. In contrast, heterogeneity was moderate for WASO and SOL, suggesting that results varied across studies, potentially due to differences in populations, devices used to measure sleep metrics, or anxiety measurement. Subgroup analyses were not conducted as this was not prespecified in the protocol, and the number of studies contributing to each meta-analysis was small [78]. Finally, these findings should also be interpreted in the context of study quality. Most studies were rated as fair quality, reflecting some methodological limitations such as the lack of sample size calculation and no adjustment for key potential confounding variables.

The Pittsburgh Sleep Quality Index, commonly used to assess subjective sleep, comprises 7 components (sleep quality, sleep latency, sleep duration, SE, sleep disturbance, use of sleep medication, and daytime dysfunction) [79]. Existing research using the Pittsburgh Sleep Quality Index generally reports that shorter sleep duration, poor sleep quality, lower SE, and longer SOL are associated with anxiety disorders in both young and older adults [80-83]. In addition, worse subjective sleep quality

XSL•FO
RenderX

has also been associated with more nonclinical levels of anxiety symptoms [84,85].

Research examining associations between anxiety and sleep metrics derived from polysomnography, the gold standard measurement of objective sleep, has been limited. One review identified 6 studies investigating polysomnographic sleep metrics, and found decreased SE and TST, and increased WASO in individuals with GAD compared with healthy controls [86]. Another review examined sleep metrics derived from electroencephalography and found that individuals with GAD had decreased TST, and increased WASO and SOL compared with healthy controls; however, no differences in SE were observed [16].

The lack of statistically significant associations observed in our meta-analyses contrasts with some findings from existing literature and may reflect differences in devices and methodological approaches. For instance, only 2 studies included in our meta-analysis asked participants to press an event marker to indicate when they began to fall asleep [36,87]. Without these event markers, estimating SOL may be less accurate [88], as different sleep behaviors can influence wrist movements. The majority of the studies (12/13, 92.3%) in our meta-analyses assessed anxiety symptoms, whereas the studies using polysomnography/electroencephalography focused on individuals with clinical anxiety. The lack of associations in our meta-analyses could also be attributed to differences in sample characteristics. Associations between sleep metrics and anxiety may emerge in clinical anxiety, but not in subthreshold anxiety symptoms.

Physical Activity

Our review suggests that higher levels of physical activity are associated with lower anxiety symptoms. These findings appear to be largely consistent across different aspects of physical activity, including step count, total physical activity, and physical activity intensities. These findings align with existing reviews showing the protective effects of physical activity on reducing anxiety symptoms [89,90].

Our review found that sedentary behavior was associated with greater anxiety symptoms. The observed positive association between sedentary behavior and anxiety symptoms aligns with existing findings [91,92]. A meta-analysis primarily based on studies that used self-report questionnaires found a small positive association between sedentary behavior and anxiety [92]. Similar to our review, their study populations were largely composed of healthy individuals rather than those with anxiety disorders. It may be that stronger associations emerge with more severe anxiety. For example, 1 study included in our review observed a significant association only when assessing clinical vs nonclinical anxiety symptoms [46]. No association was observed when analyzing anxiety as a continuous variable (with a mean anxiety score below the clinical threshold) [46]. This potential threshold effect is further supported by another study included in our review, which found that individuals with more severe anxiety symptoms engaged in more sedentary behavior than those with no significant anxiety symptoms [40].

Regarding physical intensities, the majority of studies examining light physical activity (k=3, 75%) and MVPA (k=3, 75%) found that spending more time in these intensities was associated with lower anxiety symptoms. Similarly, a meta-analysis of physical activity interventions found that these interventions reduced anxiety symptoms [93]. Although the number of studies within each subgroup is limited, subgroup analyses of exercise intensity found that both moderate- and high-intensity exercise reduced anxiety symptoms [93]. Taken together, our review's findings on physical activity intensities align with existing research showing that higher-intensity physical activity is effective in reducing anxiety symptoms.

HRV

Our review found inconclusive evidence on the association between HRV and anxiety, as only 4 studies included in this review examined HRV. An existing meta-analysis on HRV found that individuals with GAD had reduced high-frequency and time-domain HRV [94], which means they may have less flexibility in adapting to stress. This partially aligns with the findings from 2 studies included in this review: 1 found that higher HRV (domain not specified) was associated with anxiety symptoms [66], while another found that higher mean time-domain HRV was associated with lower anxiety symptoms [61]. This study also found that greater high-frequency HRV variability was associated with more anxiety symptoms [61]. The discrepancy in findings may be due to the focus of the existing meta-analysis on clinical anxiety, whereas the 2 studies in our review examined anxiety symptoms.

Elevated resting or average heart rate reflects chronic SNS activation (SNS regulates the body's fight-or-flight response) [95]. Compared to HRV, far fewer studies have examined average heart rate or resting heart rate in relation to anxiety. Existing evidence suggests that a higher resting heart rate is linked to greater anxiety symptoms and anxiety disorders [96,97]. Consistent with this literature, our review found that higher average heart rate and resting heart rate were associated with more anxiety symptoms.

Studies Using Machine Learning Approaches

Machine learning approaches emphasize building models that generalize to new data. These approaches offer advantages over inferential statistical methods [98]. For example, machine learning approaches can capture both linear and nonlinear relationships between variables without strong assumptions about variable relationships. However, machine learning models can be challenging to interpret due to their "black box" nature, where decision-making processes are not transparent [99].

In this review, machine learning models predicting anxiety had accuracies ranging from 56.3% to 90.9%. The accuracy reported here is somewhat comparable to the 2 reviews on machine learning models using any wearable devices to predict depression (70%-89%) [17] and anxiety (82%) [18], which used a range of predictors (eg, heart rate, activity, sleep, audio, and skin temperature). The majority of studies included in this review did not include demographic predictors, which might further enhance model accuracy. Indeed, some studies in the depression and anxiety machine learning reviews used

XSL•FO
RenderX

demographic data as predictors, further supporting the potential use of these predictors to improve the prediction of anxiety. The model with the highest performance (accuracy of 90.9%) combined digital biomarkers of health from more than 1 type of digital device (wrist-worn wearable and smart shirts) [72]. This aligns with existing research showing that models integrating digital biomarkers of health from multiple devices achieved better predictive performance than those relying on a single device [20]. For studies using only digital biomarkers of health from wrist-worn wearables in this review, the models with the highest accuracy included anxiety questionnaires as a predictor to predict clinically diagnosed anxiety. This likely inflates model performance due to circularity. When excluding models that incorporated anxiety questionnaires as inputs, the studies using only digital biomarkers of health from wrist-worn wearables reported accuracies ranging from 56.3% to 69.4%.

Wrist-Worn Wearable Devices

Digital biomarkers from wrist-worn wearables offer several benefits over traditional assessments of anxiety, as they allow for continuous data collection over long periods without the need for repeated administration. In addition, these largely affordable and minimally intrusive devices can be used in natural settings, providing valuable insights into sleep and activity habits. While both research-grade and consumer-grade devices are less precise than polysomnography in measuring sleep [100-102], this may be due to their indirect measurement (eg, based on movement). Despite this limitation, they are generally considered valid tools for sleep assessment [100,103]. Moreover, polysomnography is neither cost-effective nor practical for large-scale real-world studies. Given that the majority of the studies in our review used research-grade devices, a key remaining question is whether these findings generalize to consumer-grade wrist-worn devices, which are more widely accessible to the general population.

There are also benefits to using self-report data. Self-report questionnaires can capture subjective experiences, such as perceived levels of exercise or sleep quality, which wearables cannot measure. Further, they have the potential to record underlying reasons for poor sleep, for example, bad dreams or pain. Therefore, while wrist-worn wearables offer advantages, incorporating self-report measures can provide a more comprehensive understanding of both objective and subjective experience.

Strengths, Limitations, and Future Directions

To the best of our knowledge, this is the first review to examine the association between digital biomarkers of health from wrist-worn wearables and anxiety. We used double screening for all studies to ensure a rigorous selection process and to minimize bias. We also included gray literature to reduce potential publication bias and provide a more comprehensive review of existing studies.

This review also has several limitations. First, most studies focused on young adults. Part of the impetus for this review was to explore whether metrics from wrist-worn wearables could eventually be used to screen for dementia risk, as indicated by changes in anxiety symptoms. This is because anxiety is a potentially modifiable factor associated with dementia risk. However, the number of studies conducted in older adults is scarce. Future studies involving older adults would enrich our understanding of the association between digital biomarkers of health and anxiety in the context of dementia. Second, despite previous literature showing associations between physical activity and anxiety [104,105], few studies in our review focused on activity-related metrics. This precluded a meta-analytic synthesis of these findings. Wrist-worn devices generally provide more objective measurements of activity compared to self-report, which can be influenced by different biases [106], highlighting the need for further research. Third, although our search strategy and eligibility criteria allowed inclusion of any metrics measured by wrist-worn wearables, the majority of studies focused on sleep metrics. The search strategy was not peer-reviewed, and although it is possible that consulting with an information scientist might have led to the inclusion of additional relevant studies, this gap highlights an opportunity for future research to explore a wider range of metrics that could be associated with anxiety. Fourth, while machine learning has potential in predicting anxiety, the quality of these studies needs improvement. For example, future studies need to carefully consider the predictors used. None of the included machine learning studies considered demographics as potential predictors, despite existing research indicating sex differences in anxiety [107]. In addition, future studies should increase the interpretability of machine learning models by routinely reporting Shapley additive explanations values or feature importance scores. Additionally, most of the studies focused on a single metric (eg, sleep), despite the potential for the models to include multiple predictors. More research using machine learning approaches is needed, especially to compare whether this approach is superior to traditional approaches that use inferential statistics in predicting anxiety.

Wrist-worn wearable devices offer a scalable and noninvasive method to passively and continuously monitor anxiety symptoms. Metrics collected from these devices have the potential to serve as a screening or detection tool for anxiety, allowing individuals at risk of developing clinical anxiety to be identified before consulting a clinician. If these metrics and models have high predictive accuracy, they could support clinicians in identifying at-risk individuals and provide timely assessments and interventions. Before such an approach can be implemented in real-world clinical practice, more research is essential. In order to develop a clinically meaningful prediction model, interdisciplinary collaboration is crucial (eg, clinicians or machine learning experts). This will ensure the models are accurate, interpretable, clinically meaningful, and integrated into the health care system rather than remaining a research-focused tool.

Conclusions

This review is the first to comprehensively synthesize evidence on the association between digital biomarkers of health derived from wrist-worn wearable devices and anxiety, integrating findings from both inferential statistical and machine learning studies. By systematically synthesizing the literature, this review identified gaps in evidence for future research.

XSL•FO
RenderX

Our meta-analyses found no associations between SOL, TST, SE, and WASO and anxiety. Although the number of studies was limited, there was evidence that lower physical activity levels and elevated heart rate were associated with more anxiety symptoms. Existing studies on light physical activity, MVPA, and HRV present inconsistent or inconclusive findings, highlighting the need for more research. Further, a comprehensive review of the literature yielded few machine learning studies in this area, and their quality was variable. While machine learning is increasingly used to analyze wearable data, its application for predicting anxiety remains in the early stages.

Machine learning and digital biomarkers could become powerful tools for early detection of anxiety, symptom monitoring and management, and prevention, particularly when used as part of broader assessment frameworks. Continuous data collection could support long-term monitoring of behavioral and physiological patterns relevant to anxiety, and may complement self-report and clinical assessments rather than being used as stand-alone screening tools. Given that anxiety is a potentially modifiable psychological factor associated with dementia risk, wearable-derived digital biomarkers may also offer future opportunities to support early identification and monitoring of dementia risk. However, further high-quality research is required before such approaches can be meaningfully translated into clinical or public health practice.

Funding

YL is a PhD student funded by the Economic & Social Research Council's London (UBEL) Doctoral Training Partnership (ES/P000592/1), embedded within the APPLE-TREE programme, which was funded by an Economic and Social Research Council/National Institute for Health Research programme grant (ES/S010408/1). The views expressed in this publication are those of the authors and not necessarily those of the National Institute for Health Research or the Department of Health and Social Care.

Data Availability

The dataset generated and analyzed during this study is available from the corresponding author on reasonable request.

Authors' Contributions

Protocol development: YL, HD-K, NM
Database search: YL
Screening, data extraction, and risk of bias assessment: YL, NC, HAG, RM
Analysis and interpretation of data: YL, HD-K, NLM
Drafting of manuscript: YL
Critical revision of the manuscript for important intellectual content: YL, NC, HAG, RM, CC, ZW, HD-K, NLM
Study supervision: ZW, HD-K, NLM

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA 2020 Checklist.
[DOCX File , 271 KB - jmir_v28i1e73812_app1.docx]

Multimedia Appendix 2
PRISMA 2020 for abstracts checklist.
[DOCX File , 264 KB - jmir_v28i1e73812_app2.docx]

Multimedia Appendix 3
PRISMA-S Checklist.
[DOCX File , 17 KB - jmir_v28i1e73812_app3.docx]

Multimedia Appendix 4
Overview of search terms.
[DOCX File , 18 KB - jmir_v28i1e73812_app4.docx]

Multimedia Appendix 5
Risk of bias studies.

XSL•FO
RenderX

[DOCX File , 23 KB - jmir_v28i1e73812_app5.docx]

Multimedia Appendix 6
Meta-analytic results for sleep metrics.
[DOCX File , 23 KB - jmir_v28i1e73812_app6.docx]

References

1. Global burden of disease Collaborative Network. Global Burden of Disease Study 2019. Seattle, United States: Institute for Health Metrics and Evaluation (IHME); 2020.
2. Diagnostic and Statistical Manual of Mental Disorders: DSM-5TM, 5th ed. Arlington, VA, US: American Psychiatric Publishing, Inc; 2013:xliv-x947.
3. Bryant C, Jackson H, Ames D. The prevalence of anxiety in older adults: methodological issues and a review of the literature. J Affect Disord 2008;109(3):233-250. [doi: 10.1016/j.jad.2007.11.008] [Medline: 18155775]
4. Gulpers B, Ramakers I, Hamel R, Köhler S, Voshaar RO, Verhey F. Anxiety as a predictor for cognitive decline and dementia: a systematic review and meta-analysis. Am J Geriatr Psychiatry 2016;24(10):823-842. [doi: 10.1016/j.jagp.2016.05.015] [Medline: 27591161]
5. Ford E, Greenslade N, Paudyal P, Bremner S, Smith HE, Banerjee S, et al. Predicting dementia from primary care records: a systematic review and meta-analysis. PLoS One 2018;13(3):e0194735 [FREE Full text] [doi: 10.1371/journal.pone.0194735] [Medline: 29596471]
6. Santabárbara J, Villagrasa B, López-Antón R, Olaya B, Bueno-Notivol J, de la Cámara C, et al. Clinically relevant anxiety and risk of Alzheimer's disease in an elderly community sample: 4.5 years of follow-up. J Affect Disord 2019;250:16-20. [doi: 10.1016/j.jad.2019.02.050] [Medline: 30825716]
7. Santabárbara J, Lipnicki DM, Olaya B, Villagrasa B, Bueno-Notivol J, Nuez L, et al. Does anxiety increase the risk of all-cause dementia? An updated meta-analysis of prospective cohort studies. J Clin Med 2020;9(6):1791 [FREE Full text] [doi: 10.3390/jcm9061791] [Medline: 32526871]
8. Kuring JK, Mathias JL, Ward L. Risk of dementia in persons who have previously experienced clinically-significant depression, anxiety, or PTSD: a systematic review and meta-analysis. J Affect Disord 2020;274:247-261. [doi: 10.1016/j.jad.2020.05.020] [Medline: 32469813]
9. GBD 2019 Dementia Forecasting Collaborators. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. Lancet Public Health 2022;7(2):e105-e125 [FREE Full text] [doi: 10.1016/S2468-2667(21)00249-8] [Medline: 34998485]
10. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. J Multidiscip Healthc 2016;9:211-217 [FREE Full text] [doi: 10.2147/JMDH.S104807] [Medline: 27217764]
11. Balsamo M, Cataldi F, Carlucci L, Fairfield B. Assessment of anxiety in older adults: a review of self-report measures. Clin Interv Aging 2018;13:573-593 [FREE Full text] [doi: 10.2147/CIA.S114100] [Medline: 29670342]
12. Babrak LM, Menetski J, Rebhan M, Nisato G, Zinggeler M, Brasier N, et al. Traditional and digital biomarkers: two worlds apart? Digit Biomark 2019;3(2):92-102 [FREE Full text] [doi: 10.1159/000502000] [Medline: 32095769]
13. Nelson BW, Low CA, Jacobson N, Areán P, Torous J, Allen NB. Guidelines for wrist-worn consumer wearable assessment of heart rate in biobehavioral research. NPJ Digit Med 2020;3(1):90 [FREE Full text] [doi: 10.1038/s41746-020-0297-4] [Medline: 32613085]
14. Ahmed A, Aziz S, Alzubaidi M, Schneider J, Irshaidat S, Abu Serhan H, et al. Comput Methods Programs Biomed Update 2023;3:100095 [FREE Full text] [doi: 10.1016/j.cmpbup.2023.100095] [Medline: 36743720]
15. McDowell CP, Dishman RK, Gordon BR, Herring MP. Physical activity and anxiety: a systematic review and meta-analysis of prospective cohort studies. Am J Prev Med 2019;57(4):545-556. [doi: 10.1016/j.amepre.2019.05.012] [Medline: 31542132]
16. Cox RC, Olatunji BO. A systematic review of sleep disturbance in anxiety and related disorders. J Anxiety Disord 2016;37:104-129. [doi: 10.1016/j.janxdis.2015.12.001] [Medline: 26745517]
17. Abd-Alrazaq A, AlSaad R, Shuweihdi F, Ahmed A, Aziz S, Sheikh J. Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression. NPJ Digit Med 2023;6(1):84 [FREE Full text] [doi: 10.1038/s41746-023-00828-5] [Medline: 37147384]
18. Abd-Alrazaq A, AlSaad R, Harfouche M, Aziz S, Ahmed A, Damseh R, et al. Wearable artificial intelligence for detecting anxiety: systematic review and meta-analysis. J Med Internet Res 2023;25(1):e48754 [FREE Full text] [doi: 10.2196/48754] [Medline: 37938883]
19. Jacobson NC, Bhattacharya S. Digital biomarkers of anxiety disorder symptom changes: personalized deep learning models using smartphone sensors accurately predict anxiety symptoms from ecological momentary assessments. Behav Res Ther 2022;149:104013 [FREE Full text] [doi: 10.1016/j.brat.2021.104013] [Medline: 35030442]
20. Moshe I, Terhorst Y, Opoku Asare K, Sander LB, Ferreira D, Baumeister H, et al. Predicting symptoms of depression and anxiety using smartphone and wearable data. Front Psychiatry 2021;12:625247 [FREE Full text] [doi: 10.3389/fpsyt.2021.625247] [Medline: 33584388]

XSL•FO
RenderX

21. Moher D, Liberati A, Tetzlaff J, Altman DG, Group TP. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
22. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
23. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
24. Haddaway NR, Collins AM, Coughlin D, Kirk S. The role of Google Scholar in evidence reviews and its applicability to grey literature searching. PLoS One 2015;10(9):e0138237 [FREE Full text] [doi: 10.1371/journal.pone.0138237] [Medline: 26379270]
25. Study quality assessment tools. NHLBI, NIH. URL: https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools [accessed 2023-11-18]
26. Berner K, Morris L, Baumeister J, Louw Q. Objective impairments of gait and balance in adults living with HIV-1 infection: a systematic review and meta-analysis of observational studies. BMC Musculoskelet Disord 2017;18(1):325 [FREE Full text] [doi: 10.1186/s12891-017-1682-2] [Medline: 28764704]
27. Jayakumar S, Sounderajah V, Normahani P, Harling L, Markar SR, Ashrafian H, et al. Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: a meta-research study. NPJ Digit Med 2022;5(1):11 [FREE Full text] [doi: 10.1038/s41746-021-00544-y] [Medline: 35087178]
28. A basic introduction to fixed-effect and random-effects models for meta-analysis - Borenstein - 2010 - research synthesis methods -. Wiley Online Library. URL: https://onlinelibrary.wiley.com/doi/10.1002/jrsm.12 [accessed 2024-01-05]
29. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014;14:25 [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
30. Korevaar E, Turner SL, Forbes AB, Karahalios A, Taljaard M, McKenzie JE. Evaluation of statistical methods used to meta-analyse results from interrupted time series studies: a simulation study. Res Synth Methods 2023;14(6):882-902 [FREE Full text] [doi: 10.1002/jrsm.1669] [Medline: 37731166]
31. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]
32. Endler NS, Kocovski NL. State and trait anxiety revisited. J Anxiety Disord 2001;15(3):231-245. [doi: 10.1016/s0887-6185(01)00060-3] [Medline: 11442141]
33. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21(11):1539-1558. [doi: 10.1002/sim.1186] [Medline: 12111919]
34. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011;343:d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
35. Ahn J, Kim J, Park Y, Kim R, Choi H. Nature-based virtual reality relaxation to improve mental health and sleep in undergraduate students: a randomized controlled trial. Digit Health 2025;11:20552076251365140 [FREE Full text] [doi: 10.1177/20552076251365140] [Medline: 40799837]
36. Bullis J. Temporal patterns of sleep disturbance, anxiety, and depressed mood in generalized anxiety disorder. OpenBU. 2016. URL: https://open.bu.edu/handle/2144/19590 [accessed 2023-08-15]
37. Casanova F, O'Loughlin J, Karageorgiou V, Beaumont RN, Bowden J, Wood AR, et al. Effects of physical activity and sedentary time on depression, anxiety and well-being: a bidirectional Mendelian randomisation study. BMC Med 2023;21(1):501 [FREE Full text] [doi: 10.1186/s12916-023-03211-z] [Medline: 38110912]
38. Cox RC, Sterba SK, Cole DA, Upender RP, Olatunji BO. Time of day effects on the relationship between daily sleep and anxiety: an ecological momentary assessment approach. Behav Res Ther 2018;111:44-51 [FREE Full text] [doi: 10.1016/j.brat.2018.09.008] [Medline: 30300779]
39. D'Aurizio G, Tosti B, Tempesta D, Avvantaggiato L, Splendiani A, Sacco S, et al. Reduced sleep amount and increased sleep latency in prisoners: a pilot study in an Italian jail. Brain Sci 2023;13(1):132 [FREE Full text] [doi: 10.3390/brainsci13010132] [Medline: 36672113]
40. Dillon CB, McMahon E, O'Regan G, Perry IJ. Associations between physical behaviour patterns and levels of depressive symptoms, anxiety and well-being in middle-aged adults: a cross-sectional study using isotemporal substitution models. BMJ Open 2018;8(1):e018978 [FREE Full text] [doi: 10.1136/bmjopen-2017-018978] [Medline: 29358436]
41. Doane LD, Gress-Smith JL, Breitenstein RS. Multi-method assessments of sleep over the transition to college and the associations with depression and anxiety symptoms. J Youth Adolesc 2015;44(2):389-404. [doi: 10.1007/s10964-014-0150-7] [Medline: 25034248]
42. El-Sheikh M, Kelly R, Rauer A. Quick to berate, slow to sleep: interpartner psychological conflict, mental health, and sleep. Health Psychol 2013;32(10):1057-1066 [FREE Full text] [doi: 10.1037/a0031786] [Medline: 23544995]

XSL•FO
RenderX

43. Facer-Childs ER, Mascaro L, Hoffman D, Mansfield D, Drummond SPA, Rajaratnam SMW. Sleep as a major determinant for mental health outcomes in elite Australian Football League (AFL) athletes. Med Sci Sports Exerc 2022;54(4):665-672. [doi: 10.1249/MSS.0000000000002825] [Medline: 34772903]
44. Feng S, Yi JS, Deitz G, Ding L, Van Gelder RN, Menda S. Relationships between sleep, activity, and burnout in ophthalmology residents. J Surg Educ 2021;78(3):1035-1040. [doi: 10.1016/j.jsurg.2020.09.003] [Medline: 32967802]
45. Fuller-Rowell TE, Zeringue MM, Saini EK, Yip T, El-Sheikh M. Do sleep problems exacerbate the mental health consequences of discrimination among adults? Psychosom Med 2024;86(4):324-333. [doi: 10.1097/psy.0000000000001305]
46. Hofman A, Voortman T, Ikram MA, Luik AI. Substitutions of physical activity, sedentary behaviour and sleep: associations with mental health in middle-aged and elderly persons. J Epidemiol Community Health 2022;76(2):175-181 [FREE Full text] [doi: 10.1136/jech-2020-215883] [Medline: 34301796]
47. Huang T, Zheng K, Li S, Yang Y, Kong L, Zhao Y. Screen-based sedentary behaviors but not total sedentary time are associated with anxiety among college students. Front Public Health 2022;10:994612 [FREE Full text] [doi: 10.3389/fpubh.2022.994612] [Medline: 36339232]
48. Lee S, Bonnar D, Roane B, Gradisar M, Dunican IC, Lastella M, et al. Sleep characteristics and mood of professional esports athletes: a multi-national study. Int J Environ Res Public Health 2021;18(2):664 [FREE Full text] [doi: 10.3390/ijerph18020664] [Medline: 33466788]
49. Montanari A, Fancello G, Sueur C, Kestens Y, van Lenthe FJ, Chaix B. A sensor-based study on the environmental determinants of sleep in older adults. Environ Res 2025;274:120874. [doi: 10.1016/j.envres.2025.120874] [Medline: 39855412]
50. Nguyen E, Meadley B, Harris R, Rajaratnam SMW, Williams B, Smith K, et al. Sleep and mental health in recruit paramedics: a 6-month longitudinal study. Sleep 2023;46(8):zsad050 [FREE Full text] [doi: 10.1093/sleep/zsad050] [Medline: 36861384]
51. Spira AP, Friedman L, Aulakh JS, Lee T, Sheikh JI, Yesavage JA. Subclinical anxiety symptoms, sleep, and daytime dysfunction in older adults with primary insomnia. J Geriatr Psychiatry Neurol 2008;21(2):149-153. [doi: 10.1177/0891988707317120] [Medline: 18474724]
52. Spira AP, Stone K, Beaudreau SA, Ancoli-Israel S, Yaffe K. Anxiety symptoms and objectively measured sleep quality in older women. Am J Geriatr Psychiatry 2009;17(2):136-143 [FREE Full text] [doi: 10.1097/JGP.0b013e3181871345] [Medline: 19155746]
53. Stremler R, Haddad S, Pullenayegum E, Parshuram C. Psychological outcomes in parents of critically ill hospitalized children. J Pediatr Nurs 2017;34:36-43. [doi: 10.1016/j.pedn.2017.01.012] [Medline: 28274664]
54. Swanson LM, Hood MM, Hall MH, Avis NE, Joffe H, Colvin A, et al. Sleep timing, sleep regularity, and psychological health in early late life women: findings from the Study of Women's Health Across the Nation (SWAN). Sleep Health 2023;9(2):203-210 [FREE Full text] [doi: 10.1016/j.sleh.2022.11.001] [Medline: 36509657]
55. Uchida Y, Kanamori M, Fukushima S, Takemura K. Interdependent culture and older adults' well-being: health and psychological happiness in Japanese communities. Curr Opin Psychol 2024;55:101729 [FREE Full text] [doi: 10.1016/j.copsyc.2023.101729] [Medline: 38096782]
56. Windmill H, Dyer A, Smith AD, Roser ME, Dhanda A, Hall SD. Characterising the inconsistency between perceived and actual sleep and its impact upon cognition and mood. Sci Rep 2024;14(1):29342 [FREE Full text] [doi: 10.1038/s41598-024-80683-w] [Medline: 39592657]
57. Walters EM, Phillips AJ, Hamill K, Norton PJ, Drummond SP. Anxiety predicts dyadic sleep characteristics in couples experiencing insomnia but not in couples without sleep disorders. J Affect Disord 2020;273:122-130. [doi: 10.1016/j.jad.2020.04.031] [Medline: 32421592]
58. Zhang Y, Stewart C, Ranjan Y, Conde P, Sankesara H, Rashid Z, et al. Large-scale digital phenotyping: identifying depression and anxiety indicators in a general UK population with over 10,000 participants. J Affect Disord 2025;375:412-422 [FREE Full text] [doi: 10.1016/j.jad.2025.01.124] [Medline: 39892753]
59. Zheng NS, Annis J, Master H, Han L, Gleichauf K, Ching JH, et al. Sleep patterns and risk of chronic disease as measured by long-term monitoring with commercial wearable devices in the All of Us Research Program. Nat Med 2024;30(9):2648-2656. [doi: 10.1038/s41591-024-03155-8] [Medline: 39030265]
60. Bai Y, Burns R, Ma C, Curl A, Hudziak J, Copeland WE. Tracking well-being: a comprehensive analysis of physical activity and mental health in college students across COVID-19 phases using ecological momentary assessment. Scand J Med Sci Sports 2024;34(10):e14738. [doi: 10.1111/sms.14738] [Medline: 39410823]
61. Jo YT, Lee SW, Park S, Lee J. Association between heart rate variability metrics from a smartwatch and self-reported depression and anxiety symptoms: a four-week longitudinal study. Front Psychiatry 2024;15:1371946 [FREE Full text] [doi: 10.3389/fpsyt.2024.1371946] [Medline: 38881544]
62. Kandola AA, Del Pozo Cruz B, Osborn DPJ, Stubbs B, Choi KW, Hayes JF. Impact of replacing sedentary behaviour with other movement behaviours on depression and anxiety symptoms: a prospective cohort study in the UK Biobank. BMC Med 2021;19(1):133 [FREE Full text] [doi: 10.1186/s12916-021-02007-3] [Medline: 34134689]
63. Liu Y, Huang C, Wei K, Lin R, Yao Y, Qin G, et al. Accelerometer-derived "weekend warrior" physical activity, genetic susceptibility, and risk of depression and anxiety: the UK Biobank study. J Affect Disord 2025;380:55-62. [doi: 10.1016/j.jad.2025.03.110] [Medline: 40122250]

XSL•FO
RenderX

64. McDuff D, Galatzer-Levy I, Thomson S, Barakat A, Heneghan C, Abdel-Ghaffar S, et al. Evidence of differences in diurnal electrodermal, temperature and heart rate patterns by mental health status in free-living data. BMJ Ment Health 2025;28(1):e301307 [FREE Full text] [doi: 10.1136/bmjment-2024-301307] [Medline: 40819881]
65. Okawara H, Shiraishi Y, Sato K, Nakamura M, Katsumata Y. Visually assessing work performance using a smartwatch via day-to-day fluctuations in heart rate variability. Digit Health 2024;10:20552076241239240 [FREE Full text] [doi: 10.1177/20552076241239240] [Medline: 38533307]
66. Presby D, Jasinski S, Capodilupo E, Holmes KE, von Hippel W, Grosicki GJ, et al. Inter- and intrapersonal associations between physiology and mental health: a longitudinal study using wearables and mental health surveys. J Med Internet Res 2025;27:e64955 [FREE Full text] [doi: 10.2196/64955] [Medline: 40700646]
67. Straus LD, An X, Ji Y, McLean SA, Neylan TC. AURORA Study Group. Utility of wrist-wearable data for assessing pain, sleep, and anxiety outcomes after traumatic stress exposure. JAMA Psychiatry 2023;80(3):220-229 [FREE Full text] [doi: 10.1001/jamapsychiatry.2022.4533] [Medline: 36630119]
68. Wang Z, Cao Z, Min J, Duan T, Xu C. Associations between device-measured and self-reported physical activity and common mental disorders: findings from a large-scale prospective cohort study. BMJ Evid Based Med 2025;30(1):45-54. [doi: 10.1136/bmjebm-2024-112933] [Medline: 38997150]
69. Yu K, Yang Q, Wang J, Zeng B. Accelerometer-derived physical activity, sedentary behavior, and the risk of depression and anxiety in middle-aged and older adults: a prospective cohort study of 71,556 UK Biobank participants. Am J Geriatr Psychiatry 2025;33(5):512-523. [doi: 10.1016/j.jagp.2024.10.015] [Medline: 39532554]
70. Coutts LV, Plans D, Brown AW, Collomosse J. Deep learning with wearable based heart rate variability for prediction of mental and general health. J Biomed Inform 2020;112:103610 [FREE Full text] [doi: 10.1016/j.jbi.2020.103610] [Medline: 33137470]
71. Fukuda S, Matsuda Y, Tani Y, Arakawa Y, Yasumoto K. Predicting depression and anxiety mood by wrist-worn sleep sensor. 2020 Presented at: 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops); 2020 March 23-27; Austin, TX, USA p. 1-6. [doi: 10.1109/percomworkshops48775.2020.9156176]
72. He M, Cerna J, Mathew R, Zhao J, Zhao J, Espina E, et al. Objective anxiety level classification using unsupervised learning and multimodal physiological signals. Smart Health 2025;36:100572. [doi: 10.1016/j.smhl.2025.100572]
73. Lee TR, Kim GH, Choi MT. Identification of geriatric depression and anxiety using activity tracking data and minimal geriatric assessment scales. Appl Sci 2022;12(5):2488. [doi: 10.3390/app12052488]
74. Lee TR, Kim GH, Choi MT. Geriatric depression and anxiety screening via deep learning using activity tracking and sleep data. Int J Geriatr Psychiatry 2024;39(2):e6071. [doi: 10.1002/gps.6071] [Medline: 38372966]
75. Saylam B, İncel Ö. Quantifying digital biomarkers for well-being: stress, anxiety, positive and negative affect via wearable devices and their time-based predictions. Sensors (Basel) 2023;23(21) [FREE Full text] [doi: 10.3390/s23218987] [Medline: 37960685]
76. Saylam B, İncel Ö. Multitask learning for mental health: depression, anxiety, stress (DAS) using wearables. Diagnostics (Basel) 2024;14(5):501 [FREE Full text] [doi: 10.3390/diagnostics14050501] [Medline: 38472973]
77. Ahn S, Lobo JM, Davis EM, Howie-Esquivel J, Chung ML, Logan JG. Characterization of sleep efficiency transitions in family caregivers. J Behav Med 2024;47(2):308-319. [doi: 10.1007/s10865-023-00461-3] [Medline: 38017251]
78. Spineli LM, Pandis N. Problems and pitfalls in subgroup analysis and meta-regression. Am J Orthod Dentofacial Orthop 2020;158(6):901-904. [doi: 10.1016/j.ajodo.2020.09.001] [Medline: 33250104]
79. Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res 1989;28(2):193-213. [Medline: 2748771]
80. van den Berg JF, Luijendijk HJ, Tulen JHM, Hofman A, Neven AK, Tiemeier H. Sleep in depression and anxiety disorders: a population-based study of elderly persons. J Clin Psychiatry 2009;70(8):1105-1113. [doi: 10.4088/JCP.08m04448] [Medline: 19607762]
81. van Mill JG, Hoogendijk WJG, Vogelzangs N, van Dyck R, Penninx BWJH. Insomnia and sleep duration in a large cohort of patients with major depressive disorder and anxiety disorders. J Clin Psychiatry 2010;71(3):239-246. [doi: 10.4088/JCP.09m05218gry] [Medline: 20331928]
82. Potvin O, Lorrain D, Belleville G, Grenier S, Préville M. Subjective sleep characteristics associated with anxiety and depression in older adults: a population-based study. Int J Geriatr Psychiatry 2014;29(12):1262-1270. [doi: 10.1002/gps.4106] [Medline: 24733621]
83. Ramsawh HJ, Stein MB, Belik SL, Jacobi F, Sareen J. Relationship of anxiety disorders, sleep quality, and functional impairment in a community sample. J Psychiatr Res 2009;43(10):926-933. [doi: 10.1016/j.jpsychires.2009.01.009] [Medline: 19269650]
84. Zhu Y, Meng R, Jiang C, Yang N, Huang M, Wang X, et al. Sleep quality and subjective well-being in healthcare students: examining the role of anxiety and depression. Front Public Health 2023;11:1281571 [FREE Full text] [doi: 10.3389/fpubh.2023.1281571] [Medline: 38213643]
85. Gould CE, Spira AP, Liou-Johnson V, Cassidy-Eagle E, Kawai M, Mashal N, et al. Association of anxiety symptom clusters with sleep quality and daytime sleepiness. J Gerontol B Psychol Sci Soc Sci 2018;73(3):413-420 [FREE Full text] [doi: 10.1093/geronb/gbx020] [Medline: 28379498]

XSL•FO
RenderX

86. Monti JM, Monti D. Sleep disturbance in generalized anxiety disorder and its treatment. Sleep Med Rev 2000;4(3):263-276. [doi: 10.1053/smrv.1999.0096] [Medline: 12531169]
87. Doane LD, Thurston EC. Associations among sleep, daily experiences, and loneliness in adolescence: evidence of moderating and bidirectional pathways. J Adolesc 2014;37(2):145-154. [doi: 10.1016/j.adolescence.2013.11.009] [Medline: 24439620]
88. Fekedulegn D, Andrew ME, Shi M, Violanti JM, Knox S, Innes KE. Actigraphy-based assessment of sleep parameters. Ann Work Expo Health 2020;64(4):350-367 [FREE Full text] [doi: 10.1093/annweh/wxaa007] [Medline: 32053169]
89. Wanjau MN, Möller H, Haigh F, Milat A, Hayek R, Lucas P, et al. Physical activity and depression and anxiety disorders: a systematic review of reviews and assessment of causality. AJPM Focus 2023;2(2):100074 [FREE Full text] [doi: 10.1016/j.focus.2023.100074] [Medline: 37790636]
90. Singh B, Olds T, Curtis R, Dumuid D, Virgara R, Watson A, et al. Effectiveness of physical activity interventions for improving depression, anxiety and distress: an overview of systematic reviews. Br J Sports Med 2023;57(18):1203-1209 [FREE Full text] [doi: 10.1136/bjsports-2022-106195] [Medline: 36796860]
91. Teychenne M, Costigan SA, Parker K. The association between sedentary behaviour and risk of anxiety: a systematic review. BMC Public Health 2015;15(1):513 [FREE Full text] [doi: 10.1186/s12889-015-1843-x] [Medline: 26088005]
92. Allen MS, Walter EE, Swann C. Sedentary behaviour and risk of anxiety: a systematic review and meta-analysis. J Affect Disord 2019;242:5-13. [doi: 10.1016/j.jad.2018.08.081] [Medline: 30170238]
93. Lin Y, Gao W. The effects of physical exercise on anxiety symptoms of college students: a meta-analysis. Front Psychol 2023;14:1136900 [FREE Full text] [doi: 10.3389/fpsyg.2023.1136900] [Medline: 37063553]
94. Chalmers JA, Quintana DS, Abbott MJA, Kemp AH. Anxiety disorders are associated with reduced heart rate variability: a meta-analysis. Front Psychiatry 2014;5:80 [FREE Full text] [doi: 10.3389/fpsyt.2014.00080] [Medline: 25071612]
95. Kishi T. Heart rate is the clinical indicator of sympathetic activation and prognostic value of cardiovascular risks in patients with hypertension. Hypertension 2020;76(2):323-324. [doi: 10.1161/hypertensionaha.120.14898]
96. Kemp AH, Brunoni AR, Santos IS, Nunes MA, Dantas EM, Carvalho de Figueiredo R, et al. Effects of depression, anxiety, comorbidity, and antidepressants on resting-state heart rate and its variability: an ELSA-Brasil cohort baseline study. Am J Psychiatry 2014;171(12):1328-1334. [doi: 10.1176/appi.ajp.2014.13121605] [Medline: 25158141]
97. Latvala A, Kuja-Halkola R, Rück C, D'Onofrio BM, Jernberg T, Almqvist C, et al. Association of resting heart rate and blood pressure in late adolescence with subsequent mental disorders: a longitudinal population study of more than 1 Million men in Sweden. JAMA Psychiatry 2016;73(12):1268-1275. [doi: 10.1001/jamapsychiatry.2016.2717] [Medline: 27784035]
98. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nat Methods 2018;15(4):233-234 [FREE Full text] [doi: 10.1038/nmeth.4642] [Medline: 30100822]
99. Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, et al. Interpreting black-box models: a review on explainable artificial intelligence. Cogn Comput 2024;16(1):45-74. [doi: 10.1007/s12559-023-10179-8]
100. Lehrer HM, Yao Z, Krafty RT, Evans MA, Buysse DJ, Kravitz HM, et al. Comparing polysomnography, actigraphy, and sleep diary in the home environment: the Study of Women's Health Across the Nation (SWAN) sleep study. Sleep Adv 2022;3(1):zpac001 [FREE Full text] [doi: 10.1093/sleepadvances/zpac001] [Medline: 35296109]
101. Blackwell T, Yaffe K, Ancoli-Israel S, Schneider JL, Cauley JA, Hillier TA, Study of Osteoporotic Fractures Group. Poor sleep is associated with impaired cognitive function in older women: the study of osteoporotic fractures. J Gerontol A Biol Sci Med Sci 2006;61(4):405-410. [doi: 10.1093/gerona/61.4.405] [Medline: 16611709]
102. Chinoy ED, Cuellar JA, Huwa KE, Jameson JT, Watson CH, Bessman SC, et al. Performance of seven consumer sleep-tracking devices compared with polysomnography. Sleep 2021;44(5):zsaa291 [FREE Full text] [doi: 10.1093/sleep/zsaa291] [Medline: 33378539]
103. Marino M, Li Y, Rueschman MN, Winkelman JW, Ellenbogen JM, Solet JM, et al. Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. Sleep 2013;36(11):1747-1755 [FREE Full text] [doi: 10.5665/sleep.3142] [Medline: 24179309]
104. De Mello MT, Lemos VDA, Antunes HKM, Bittencourt L, Santos-Silva R, Tufik S. Relationship between physical activity and depression and anxiety symptoms: a population study. J Affect Disord 2013;149(1-3):241-246. [doi: 10.1016/j.jad.2013.01.035] [Medline: 23489405]
105. Azevedo Da Silva M, Singh-Manoux A, Brunner EJ, Kaffashian S, Shipley MJ, Kivimäki M, et al. Bidirectional association between physical activity and symptoms of anxiety and depression: the Whitehall II study. Eur J Epidemiol 2012;27(7):537-546 [FREE Full text] [doi: 10.1007/s10654-012-9692-8] [Medline: 22623145]
106. Prince SA, Adamo KB, Hamel ME, Hardt J, Connor Gorber S, Tremblay M. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. Int J Behav Nutr Phys Act 2008;5:56 [FREE Full text] [doi: 10.1186/1479-5868-5-56] [Medline: 18990237]
107. McLean CP, Asnaani A, Litz BT, Hofmann SG. Gender differences in anxiety disorders: prevalence, course of illness, comorbidity and burden of illness. J Psychiatr Res 2011;45(8):1027-1035 [FREE Full text] [doi: 10.1016/j.jpsychires.2011.03.006] [Medline: 21439576]

XSL•FO
RenderX

Abbreviations

GAD: generalized anxiety disorder
HRV: heart rate variability
MAE: mean absolute error
MVPA: moderate to vigorous physical activity
PECOS: Population, Exposure, Control, Outcome, and Study design
PNS: parasympathetic nervous system
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search Extension
PROSPERO: International Prospective Register of Systematic Reviews
REM: rapid eye movement
SE: sleep efficiency
SNS: sympathetic nervous system
SOL: sleep onset latency
TST: total sleep time
WASO: wake after sleep onset

Edited by S Brini; submitted 12.Mar.2025; peer-reviewed by EDA Bacil, R Chigevenga; comments to author 12.Sep.2025; revised version received 06.Feb.2026; accepted 09.Feb.2026; published 09.Mar.2026.

Please cite as:
Lau Y, Chemas N, Ajeet Gokani H, Morrell R, Phannarus H, Cooper C, Walker Z, Demnitz-King H, Marchant NL
Association Between Digital Biomarkers of Health and Anxiety: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e73812
URL: https://www.jmir.org/2026/1/e73812
doi:10.2196/73812
PMID:

©Yolanda Lau, Natalia Chemas, Heema Ajeet Gokani, Rachel Morrell, Harisd Phannarus, Claudia Cooper, Zuzana Walker, Harriet Demnitz-King, Natalie L Marchant. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effects of Digital Health Interventions on Functional and Psychological Outcomes in Older Patients With Hip Fractures: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Wei Fan[1], MD; Qi Zhang[2*], PhD; Qunfeng Lu[2*], PhD

[1]School of Medicine, Tianjin Tianshi College, Tianjin, China

[2]Department of Nursing, Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Xuhui District, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Qunfeng Lu, PhD
Department of Nursing
Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine
No.600 Yishan Road
Xuhui District, Shanghai, 200233
China
Phone: 86 18917128006
Email: luluroom2004@163.com

Abstract

Background: Hip fractures in older adults increasingly challenge public health, making traditional rehabilitation very challenging. Digital health interventions (DHIs) have emerged as a promising solution for postoperative rehabilitation. However, evidence on DHIs' effects on functional and psychological outcomes remains insufficient.

Objective: This systematic review aimed to comprehensively examine the effects of DHIs on functional and psychological outcomes in older adults with hip fractures.

Methods: Following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, we searched 9 databases (PubMed, Embase, CENTRAL, APA PsycINFO, Web of Science, PEDro, CNKI, WANFANG, and SinoMed) from inception to November 13, 2025. Included studies enrolled adults aged 60 years and older with hip fractures, delivered DHIs, assessed functional and psychological outcomes, set usual care or no intervention as the control, and had a randomized controlled trial design. Studies were excluded if they enrolled nonhospitalized patients in the emergency department, patients discharged to nonhome settings, or had inaccessible full text or insufficient data. Study quality was evaluated using the Cochrane Risk of Bias tool 2.0 (Cochrane Collaboration), and evidence certainty was assessed using GRADE (Grading of Recommendations, Assessment, Development and Evaluation). The literature screening, data extraction, and quality assessment were independently conducted by 2 researchers, and any disputes were resolved by the third researcher. We performed analysis using R version 4.0.3 (R Foundation for Statistical Computing) with a random-effects model.

Results: Of 17,723 studies screened, 13 met the inclusion criteria. DHIs, compared to the control, significantly improved hip function (standardized mean difference [SMD] 0.80, 95% CI 0.33-1.26; 95% prediction interval [PI] –0.24 to 1.83; P=.007) and functional independence (SMD 1.23, 95% CI 0.34-2.11; 95% PI –0.98 to 3.34; P=.02). Despite favorable pooled effects, a wide 95% PI spanning positive or negative values signals substantial heterogeneity. No significant difference was observed in balance function, risk of falling, and quality of life. Only a single available study reported a 70% adherence rate in the DHIs group. Subgroup analyses stratified by intervention duration revealed no significant intersubgroup differences for hip function (χ_1^2=0.1; P=.75) or functional independence (χ_1^2=2.93; P=.09). For hip function, the point estimate favored the 3 months subgroup (SMD 0.89, 95% CI 0.36-1.41; I^2=7%; P=.41) over the <3 months subgroup. Conversely, for functional independence, the point estimate favored shorter intervention duration (SMD 0.67, 95% CI 0.12-1.23; I^2=0%; P=.72).

XSL•FO
RenderX

Conclusions: This review incorporates the latest randomized controlled trials and comprehensively assesses functional and psychological outcomes of DHIs in older patients with hip fractures, distinct from prior studies focusing solely on functional outcomes. While the 95% CI supports the potential of DHIs to improve hip function and functional independence, the wide 95% PI indicating substantial real-world response variability, which calls for cautious interpretation, informs the design of targeted DHI-based rehabilitation regimens, warranting further research into optimal techniques and dosages in clinical practice.

Trial Registration: PROSPERO CRD42024626186; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024626186

(J Med Internet Res 2026;28:e79563) doi:10.2196/79563

KEYWORDS

digital health; older patients; hip fractures; systematic review; meta-analysis

Introduction

Hip fractures, particularly prevalent in older adults, present a significant and escalating public health challenge, characterized by well-documented morbidity, mortality, and health care burden [1-3]. By 2050, the annual global incidence of hip fractures is expected to reach 4.5 million [4], exerting ongoing pressure on existing health care and welfare systems [2,5]. With the general increase in human life expectancy, the systemic medical challenges posed by this condition are expected to intensify further. As per existing literature, the 30-day mortality rate after hip fractures can reach up to 10%, escalating further to 25% within the first year [6], while most surviving patients experience a significant decline in quality of life (QoL) [7]. Effective postsurgical rehabilitation is essential to mitigating the adverse outcomes of older adults with hip fractures, such as functional decline [8,9], psychological distress [10,11], and reduced QoL [9,12]. The majority of rehabilitation must now take place outside of hospitals due to shorter hospital stays brought on by a lack of health care resources and the need to increase hospital bed turnover rates [13].

Traditional rehabilitation programs face significant challenges in older adults, including limited access to health care services, low intervention adherence to prescribed rehabilitation programs, and insufficient psychological support. Adherence to rehabilitation programs remains a major issue, with many patients failing to complete recommended programs, thereby compromising functional recovery and increasing the risk of complications [14]. In addition, older adults may face specific barriers, such as limited mobility, geographic constraints, and resource shortages, which further reduce the feasibility and success of traditional rehabilitation methods [15-19]. Digital health interventions (DHIs) have emerged as a promising solution. DHIs, defined as the application of digital technologies in health care, leverage both conventional and innovative forms of information and communication technologies, such as mobile apps, telemedicine, wearable devices, and online platforms, to enhance health care delivery and patient engagement [20].

By bridging gaps in distance and accessibility, DHIs offer innovative solutions to overcome traditional health care barriers, particularly for populations with limited mobility, such as older patients with hip fractures [21]. Moreover, as digital technologies continue to advance, DHIs are evolving in both scope and sophistication, capable of providing more personalized, scalable, and effective approaches for the care of most patients whose rehabilitation training needs to be carried out outside the hospital [20,22]. DHIs have been used in the rehabilitation of older patients with hip fractures, and their potential effects have also been preliminarily demonstrated [13,21,23-26]. A study carried out by Gilboa et al [27] investigating a home-based tele-rehabilitation program, where occupational therapists remotely guided patients through personalized goal-setting and daily activity planning, demonstrated significant improvements in activities of daily living, health-related QoL, and mobility, while also reducing caregiver burden. In contrast to studies supporting DHIs, some researchers have argued that there are no or only minor effects detected in functional and psychological outcomes [28,29].

While the above findings highlight the potential impact of DHIs on functional recovery and psychological well-being in older adults with hip fractures, the evidence remains limited. Additionally, most studies focus primarily on functional recovery without adequately addressing psychological well-being and intervention adherence. Considering that the rehabilitation of older patients with hip fractures is an overall concept, it not only involves the recovery of function but also the restoration of psychological outcomes [30]. At the same time, studying the influence of intervention on intervention adherence is essential for the development of successful interventions in the rehabilitation of older patients with hip fractures [31]. The current research gaps underscore the need for a more comprehensive evaluation of DHIs in the context of the rehabilitation of hip fractures. With ongoing advancements in DHIs, a plethora of original studies focused on functional and psychological outcomes in older adults with hip fractures have been recently published. This makes it possible to review DHIs' applications and guide future design and implementation of DHIs technology in clinical practice.

Thus, our systematic review aimed to comprehensively examine the effects of DHIs on functional and psychological outcomes, as well as intervention adherence in older adults with hip fractures. Compared with previous systematic reviews, this systematic review incorporates more recent research, includes a greater number of studies addressing psychological outcomes, and assesses the quality of the evidence [21].

Methods

Overview

This systematic review and meta-analysis was reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement

XSL•FO
RenderX

(Multimedia Appendix 1) and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) to strengthen transparency (Multimedia Appendix 2) [32,33]. The review protocol was registered on the PROSPERO International Prospective Register of Systematic Reviews (CRD42024626186).

Search Strategy

This systematic review was systematically searched in 9 scientific databases, including PubMed, Embase, CENTRAL, APA PsycINFO, Web of Science, PEDro, CNKI, WANFANG, and SinoMed. The literature search was updated by re-executing the predefined search string in the same 9 databases on November 13, 2025, and newly retrieved records were screened following the identical inclusion and exclusion criteria. Search strategies were originally designed by our research team and underwent a peer review process conducted by an independent medical doctor with extensive evidence-based medicine expertise, who verified the comprehensiveness of search terms, database selection, and eligibility criteria prior to the initiation of retrieval. No substantive modifications to the strategy were required following this review. No published search filters were used in our review. No search strategies from previously published literature reviews were adapted or reused for any part of our literature retrieval process.

Reference lists of eligible studies and cited studies (identified via Google Scholar) were manually screened to retrieve additional relevant literature. The ClinicalTrials.gov (United States Clinical Trials Registry) and the Chinese Clinical Trial Registry were searched for eligible studies.

Search engines and citation chasers were also screened to capture all relevant studies. For studies where the full text could not be obtained, corresponding authors were contacted to attempt to acquire the full text. Considering the availability of academic resources and the language proficiency of the research team, the included studies were limited to those published in English and Chinese with no date restriction. The full search strategies are available in Multimedia Appendix 3.

Eligibility Criteria

We included the studies that related to (1) older adults (including patients aged 60 years or older) with hip fractures defined as any type of hip fractures, such as femoral neck, intertrochanteric, subtrochanteric fractures, and no specified type of hip fractures; (2) the intervention consisted of DHIs according to classifications of the World Health Organization (WHO) and must be provided or supported through DHIs for patients rather than trials targeting health care providers or other stakeholders [34]; (3) the primary outcome in this systematic review including functional outcomes (eg, hip function, walking capacity, and balance ability) and secondary outcomes including psychological outcomes (eg, depression and QoL) and intervention adherence; (5) the comparator was usual care, control group or no intervention; and (6) the study design limited to randomized controlled trials (RCTs). The following studies were excluded: (1) patients enrolled in the emergency department but not admitted to the hospital, (2) patients discharged from the hospital to a setting other than home, (3) studies involving DHIs that did not directly intervene with patients (including database management), and (4) studies with inaccessible full text or insufficient data for evaluating the characteristics of DHIs or quantitative synthesis.

Study Selection and Data Extraction

After automatic and manual removal of duplicate records using the reference management software EndNote 21 (Clarivate Analytics), 2 review authors (WF and QZ) independently screened titles and abstracts and then full texts against predefined eligibility criteria. A senior author (QL) resolved conflicts.

Two authors (WF and QZ) independently but in parallel collected relevant data in a standardized data collection sheet, including but not limited to characteristics of the included studies, such as the authors, publication year, country, and study design, as well as participant characteristics, including sample size, sex ratio, average age, and the details of the intervention and control groups.

For studies with multiple intervention arms, data from the relevant arms were combined to create a single pairwise comparison, as recommended in the Cochrane Handbook, to avoid double-counting participants in the control group. For studies reporting multiple time points for the same outcome, data at the end of the intervention were extracted for meta-analysis. Any discrepancies between groups were resolved through discussion and consensus, and when necessary, a senior reviewer (QL) was consulted to adjudicate unresolved differences.

Risk-of-Bias and Grading of Recommendations Assessment, Development, and Evaluation

Risk of bias in each study was assessed independently by 2 authors (WF and QZ). The RCTs were evaluated using the Cochrane Risk of Bias tool 2.0 (Cochrane Collaboration), which assesses random sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment, incomplete outcome data, selective reporting, and other bias [35]. The reviewers made a judgment of “yes” (low bias), “no” (high bias), or “unclear” (lack of relevant information or uncertainty of bias) for each item. Included studies were also subjected to scoring for certainty of evidence using GRADE (Grading of Recommendations, Assessment, Development and Evaluation) criteria independently by 2 authors (WF and QZ). Any discrepancies between groups were resolved through discussion and consensus, and when necessary, a third reviewer (QL) was consulted to adjudicate unresolved differences.

Meta-Analysis Methods

R (version 4.0.3; R Foundation for Statistical Computing) was used to conduct the meta-analyses. The data analysis was based on postintervention means and SDs, with conversion performed using established formulas when only change values were reported in the original studies [36]. The effect measures were estimated through the standardized mean difference (SMD) based on Hedges *g*, along with the 95% CI. To assess statistical heterogeneity, τ^2 was calculated. The Cochran Q test and I^2 were also derived, which together quantify the extent and significance

XSL•FO
RenderX

of the observed heterogeneity. Considering the expected significant heterogeneity among studies, we adopted a random-effects model to combine the effect sizes. The estimation of the CI was performed by the HKSJ (Hartung-Knapp-Sidik-Jonkman) method. For outcomes with enough studies (n3), we calculated 95% prediction intervals (PIs) to estimate the range of true effects in future settings. Forest plots are used to present the results of the merger.

Subgroup analyses were conducted to explore sources of heterogeneity, stratified by (1) type of DHIs (mobile health [mHealth]: telephone calls, text messages, and apps; eHealth: internet based, websites, and emails; or digital device: wearables and Bluetooth-connected devices) [37]; and (2) the duration of intervention (<3 months vs ≥3 months). The duration threshold was chosen based on common clinical rehabilitation milestones. If a prespecified subgroup analysis is not feasible due to a small number (<3 studies) of included studies, we will use narrative synthesis to summarize the included studies.

In addition, sensitivity analyses were conducted by sequentially excluding one study at a time to assess the reliability and stability of the main outcome (hip function). A funnel plot and Egger test were used to assess the small study effect. All tests were 2-tailed with a significance level of α=.05.

Results

Search Findings

The initial database search yielded 17,708 records, and an additional 15 records were identified through other sources, resulting in a total of 17,723 records. After removing 6105 duplicates, 11,618 records were excluded by title and abstract screening. Subsequently, the remaining 43 studies were screened for full text. Finally, a subset of 13 studies [29,38-49] was included. The PRISMA flow diagram is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Flowchart of the study selection process.

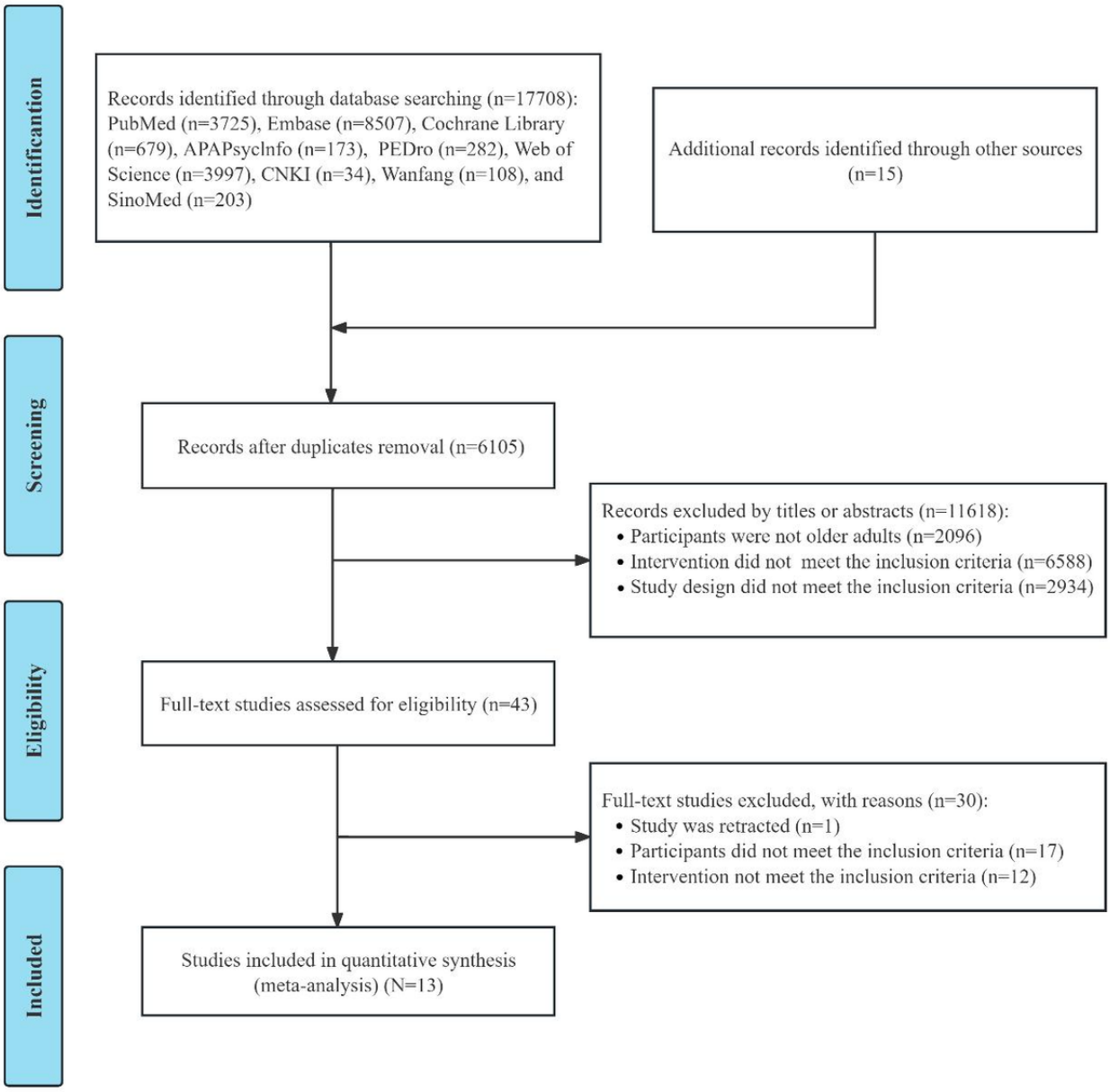

Characteristics of Included Studies

This systematic review included 13 studies [29,38-49] with 737 participants in total and published between 2015 and 2025 (Table 1). Eight studies were implemented in China [29,38-44], and one each in Israel [45], Sweden [46], Spain [47], Thailand [48], and Canada [49]. The age of participants ranged from 65.7 (SD 7.8) to 86.0 (SD 16.0) years, and the percentage of males ranged from 27.2% (38/138) to 80.6% (25/31). The intervention lasted between 3 weeks and 4 months, and the duration of follow-up was from 3 weeks to 9 months.

XSL•FO
RenderX

Table 1. Characteristics of included studies.

Study, year, country, and study design	Sample size (IG[a]:CG[b]); sex (M[c]:F[d]); age (IG:CG)	Intervention type and provider	Contents of the intervention group	Control group	Outcome and outcome measurement tools
Kalron et al [45], 2018, Israel: RCT[e] (FS[f]), Blind (participants, investigators)	32 (15:17); 20:12; 65.7(7.8):67.3(9.5)	mHealth[g] (app): physiotherapist	• Video lesson (includes health education, rehabilitation training standards, and key points of training movements) and progressive rehabilitation training. • Frequency: 5 times per week. • Duration: 4 months.	Regular manage	• Balance function and risk of falling: TUG[h]
Li et al [38], 2022, China, Hong Kong: RCT, Blind (outcome assessors)	31 (15:16); 25:6; 76.5 (8.6):82.1 (9.7)	eHealth (internet-based): physiotherapist	• Telerehabilitation was delivered through the Caspar Health e-system (CASPAR Health) to initiate falls prevention information, coaching to remain active, problem-solving skills, mobility goal setting, and advice to help participants maintain and increase their prescribed home exercises. • Frequency: personalized. • Duration: 8 weeks.	Usual care	• Functional independence: BI[i] • Balance function and risk of falling: TUG
Howell et al [46], 2023, Sweden: RCT (FS), Blind (participants)	21 (10:11); 15:6; 86.0 (16.0):83.0 (13.0)	Digital device (wearables): physiotherapist	• Participant wore a wearable device to measure body position and movements; the therapist receives feedback and can ensure that the rehabilitation program is tailored to the needs of the individual. • Frequency: no mention. • Duration: 4 weeks.	Standard rehabilitation	• Functional independence: BI • Balance function and risk of falling: FBG[j] • QoL[k]: EQ-5D
Langford et al [49], 2015, Canada: RCT, Blind (investigators, outcome assessors)	30 (15:15); 19:11; 83.0 (8.0):82.0 (10.0)	mHealth (app): physician	• Exercises relating to movement, strength of the lower limbs, and balance performance were included by using a mobile app. • Frequency: personalized. • Duration: 3 weeks.	Usual care	• QoL: EQ-5D
Cheng et al [29], 2022, China, Hong Kong: RCT, Open label	39 (19:20); 20:12; 75.8 (7.2):79.0 (8.8)	mHealth (app): physiotherapist	• Exercises with embedded videos in the app, receive a home-based rehabilitation program, and caregiver education. • Frequency: once per day. • Duration: 2 months.	Standard rehabilitation	• Functional independence: MFAC[l]
Prieto-Moreno et al [47], 2024, Spain and Belgium: RCT, Blind (outcome assessors)	105 (51:54); 30:75; 79.6 (7.1):80.1 (7.7)	eHealth: (internet-based); no mention	• Health knowledge, education, and telerehabilitation guidance, including physical exercise and an occupational therapy session. • Frequency: twice a week. • Duration: 12 weeks.	Routine rehabilitation	• Functional independence: FIM[m] • QoL: EQ-5D

XSL•FO
RenderX

Study, year, country, and study design	Sample size (IG[a]:CG[b]); sex (M[c]:F[d]); age (IG:CG)	Intervention type and provider	Contents of the intervention group	Control group	Outcome and outcome measurement tools
Li [44], 2024, China: RCT, No mention	80 (38:42); 47:33; 68.5 (5.6):69.4 (5.0)	mHealth (app): physician	• Remote rehabilitation guidance to the patients and their family caregivers via WeChat (Tencent) video, answered the patients' questions, and guided functional exercises. • Frequency: once each in the second, fourth, and eighth weeks. • Duration: 12 weeks.	Routine health education	• Hip Function: HHS[n] • QoL: SF-12[o]
Zhang [42], 2024, China: RCT, Blind (outcome assessors)	61 (30:31); 36:25; 69.0 (4.5):70.0 (7.4)	mHealth (app): physiotherapist	• The experimental group provided exercise guidance and training feedback to the subjects through a mobile app (Joymotion software, Shanghai Fudong Medical Management Co, Ltd), and conducted real-time 2-way audio and video interactive consultations with physical therapists. • Frequency: 5 times a week. • Duration: 4 weeks.	Routine rehabilitation	• Hip function: HOOS[p] • QoL: SF-12
Gao et al [39], 2021 China: RCT (No mention)	80 (40:40); 30:50; 77.8(5.9):78.1(5.6)	mHealth (app): physician	• Doctors transmitted text, pictures, voice, and video to the group to guide and urge patients to perform rehabilitation exercises through WeChat. For special patients, doctors would provide personalized and targeted guidance. • Frequency: no mention. • Duration: 2 months.	Routine rehabilitation	• Hip function: HHS
Zhang (1) [40], 2021, China: RCT, Blind (participants, outcome assessors)	51 (27:24); 18:33; 77.0 (7.9):75.2 (7.7)	eHealth (internet-based): physiotherapist	• The experimental group adopted the "Family-Oriented Postoperative Rehabilitation Management System for Elderly Hip Fractures" developed by Tianjin Hospital of Tianjin, which was used for remote home rehabilitation. This system is based on technologies such as Internet Plus, the Internet of Things, cloud computing, mobile Internet, and big data. It adopts the approach of "integrated application of technologies and rapid combination of modules" to establish a family rehabilitation treatment platform. • Frequency: no mention. • Duration: 3 months.	Usual care	• Hip function: HHS • Functional independence: FIM • Balance function and risk of falling: TUG • QoL: SF-36[q]
Zhang (2) [41], 2021, China: RCT (No mention)	36 (18:18); 15:21; 72.4 (8.4):68.8 (12.5)	mHealth (app): rehabilitation team		Usual care	• Hip Function: HHS • QoL: SF-36

XSL•FO
RenderX

Study, year, country, and study design	Sample size (IG[a]:CG[b]); sex (M[c]:F[d]); age (IG:CG)	Intervention type and provider	Contents of the intervention group	Control group	Outcome and outcome measurement tools
			• Participants received remote rehabilitation based on network media support, mainly using 2 social software programs, WeChat and TikTok (ByteDance). First, establish a WeChat group composed of rehabilitation physicians, surgeons, physical therapists, head nurses, patients, and their families. Make full use of their spare time to provide real-time rehabilitation guidance to patients and answer their questions and their families' questions. • Frequency: no mention. • Duration: 10 weeks.		
Dajpratham et al [48], 2025, Thailand: RCT, Blind (investigators)	33 (16:17); 6:27; 75.1 (9.5):78.3 (7.7)	mHealth (app): physiotherapist	• Real-time videoconference exercises with a physiotherapist via the LINE app. • Frequency: 3 times per week, with each session lasting approximately 45 minutes. • Duration: 12 weeks.	Usual care	• Balance function and risk of falling: SPPB[r] • Intervention adherence: percentage format
Shui [43], 2025, China: RCT (No mention)	138 (69:69); 38:100; 76.0 (5.7):75.2 (5.6)	mHealth (app): rehabilitation team	• Family remote rehabilitation program training: the attending physician, rehabilitation physician, and responsible nurse established an online communication group through the WeChat platform to strengthen patient management and guide the training content. • Frequency: 3 times per month. • Duration: 3 months.	Usual care	• Hip function: HHS • Functional independence: FIM • QoL: EQ-5D

[a]IG: intervention group.
[b]CG: control group.
[c]M: male.
[d]F: female.
[e]RCT: randomized controlled trial.
[f]FS: feasibility study.
[g]mHealth: mobile health.
[h]TUG: Timed Up and Go test.
[i]BI: Barthel Index.
[j]FBG: Functional Balance for Geriatric Patients.
[k]QoL: quality of life.
[l]MFAC: Modified Functional Ambulatory Category.
[m]FIM: Functional Independence Measure.
[n]HHS: Harris Hip Score.
[o]SF-12: Short Form–12 Health Survey.
[p]HOOS: Hip Disability and Osteoarthritis Outcome Score.
[q]SF-36: Medical Outcomes Study 36-Item Short Form Health Survey.
[r]SPPB: Short Physical Performance Battery.

XSL•FO
RenderX

In the 13 included studies [29,38-49], the type of DHIs could be divided into 3 categories, which included mHealth (n=9 [29,39,41-45,48,49]), eHealth (n=3 [38,40,47]), and digital device (n=1 [46]). The involvement of health care providers in DHIs varied across studies: a rehabilitation team was involved in 2 studies [41,43], physiotherapists in 7 studies [29,38,40,42,45,46,48], physicians in 3 studies [39,44,49], and 1 study [47] did not specify the intervention provider.

Of the 13 included trials [29,38-49], outcomes of our interest included hip function (6 studies [39-44]), functional independence (6 studies [29,38,40,43,46,47]), balance function and risk of falling (5 studies [38,40,45,46,48]), QoL (8 studies [40-44,46,47,49]), and intervention adherence (1 study [48]). Although the instruments used for outcome evaluation varied across studies, all were valid scales, and the process of data collection was carried out. Six studies [39-44] that reported hip function were assessed using the Harris Hip Score (HHS) [39-41,43,44] and Hip Disability and Osteoarthritis Outcome Score (HOOS) [42]. Six trials [29,38,40,43,46,47] evaluated functional independence using the Functional Independence Measure (FIM) [40,43,47], the Modified Functional Ambulatory Category (MFAC) [29], and the Barthel index [38,46]. Instruments used to evaluate balance function and risk of falling included the Timed Up and Go test (TUG) [38,40,45], Short Physical Performance Battery (SPPB) [48], and Functional Balance for Geriatric Patients (FBG) [46]. Eight studies [40-44,46,47,49] that reported QoL were assessed using the Medical Outcomes Study (MOS) 36-Item Short Form Health Survey (SF-36) [40,41], the EQ-5D scale [43,46,47,49], and the Short Form-12 Health Survey (SF-12) [42,44]. Intervention adherence was described using the percentage format in a single study [48].

Risk of Bias

Overall, 5 studies [40,42,45,47,48] had low risk of bias, and 5 studies [38,39,41,43,44] had some concerns. Another 3 studies [29,46,49] had a high risk of bias, 1 study [29] had a high risk in terms of the randomization process, 1 study [46] had a high risk in terms of the measurement of the outcome, and 1 study [49] had a high risk in terms of the selection of the reported result. Detailed results are presented in Figure 2.

Figure 2. Risk of bias.

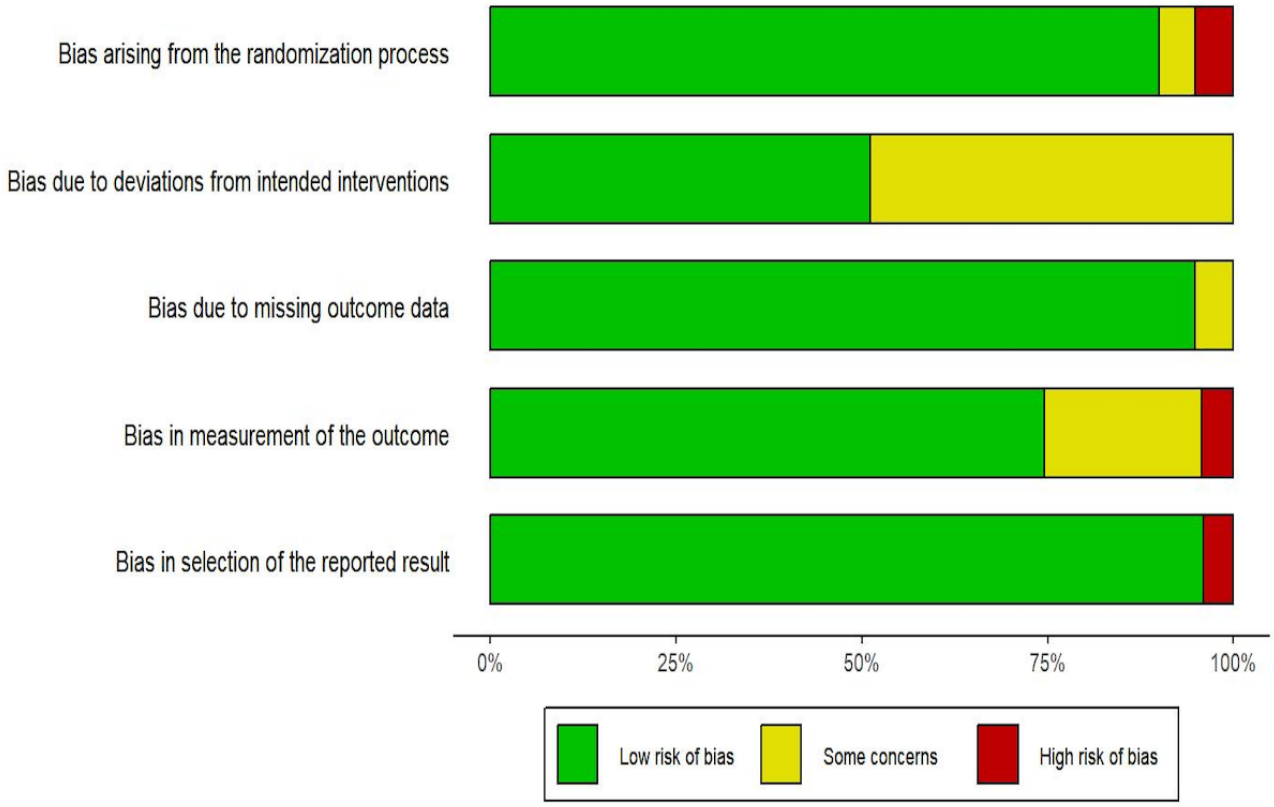

Certainty of Evidence

The GRADE assessment scores ranged from low to moderate (Table 2). Evidence was downgraded if (1) the risk of bias was apparent; (2) inconsistency was demonstrated, with I^2>75%; (3) there was indirectness in participants or comparators; (4) there was evidence of imprecision; or (5) there was publication bias. An overall GRADE rating was agreed upon (WF, QZ, and QL) for each outcome at 4 levels, including very low, low, moderate, and high.

XSL•FO
RenderX

Table 2. GRADE (Grading of Recommendations, Assessment, Development and Evaluation) summary of findings.

Outcomes	Number of participants (studies)	Quality of evidence (GRADE[a])	Relative effect (SMD[b], 95% CI)	Anticipated absolute effect	
				Control group, mean (SD)	DHIs[c] group, mean (95% CI)
Hip function	435 (6 RCTs[d])	(⊠) Moderate[e,f] (due to imprecision)	0.80 (0.33-1.26)	HHS[g]: 56.71 (14.71)	68.48 (61.56-75.24)
Functional independence	369 (6 RCTs)	(⊠) Low[e,f] (due to inconsistency, imprecision)	1.23 (0.34-2.11)	FIM[h]: 82.67 (5.69)	89.67 (84.60-94.68)
Balance function and risk of fall	163 (5 RCTs)	(⊠) Moderate[f] (due to imprecision)	–0.09 (–0.99 to 0.81)	TUG[i]: 34.16 (15.07)	32.80 (19.24-46.37)
Quality of life	501 (8 RCTs)	(⊠) Low[e,f] (due to inconsistency, imprecision)	0.65 (–0.21 to 1.50)	SF-36[j]: 56.71 (14.71)	66.27 (53.62-78.78)
Intervention adherence	33 (1 RCT)	__[k]	—	—	—

[a]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.
[b]SMD: standard mean difference.
[c]DHI: digital health intervention.
[d]RCT: randomized controlled trial.
[e]Inconsistency.
[f]Imprecision.
[g]HHS: Harris Hip Score.
[h]FIM: Functional Independence Measure.
[i]TUG: Timed Up and Go.
[j]SF-36: MOS 36-Item Short Form Health Survey.
[k]Not available.

Findings of Meta-Analysis

Effects of DHIs on Hip Function

As shown in Figure 3, 6 trials [39-44] with 435 patients were eligible for pooling for the analysis of hip function outcome. The pooled results from the random-effects model indicated that DHIs significantly improved hip function compared to the control group, with an SMD of 0.80 (95% CI 0.33-1.26; P=.007). The 95% PI (–0.24 to 1.83) further quantified the range of potential treatment effects across individual patients and clinical settings. Although pooled results suggest that DHIs yield a favorable average effect, some patients or subgroups might experience negligible or even marginal negative responses to the intervention, which aligns with the substantial heterogeneity observed (χ_5^2=15.1; P=.01; τ^2=0.13; I^2=67%).

XSL•FO
RenderX

Figure 3. Forest plot for the effect of digital health interventions on hip function [39-44]. DHI: digital health intervention; SMD: standardized mean difference.

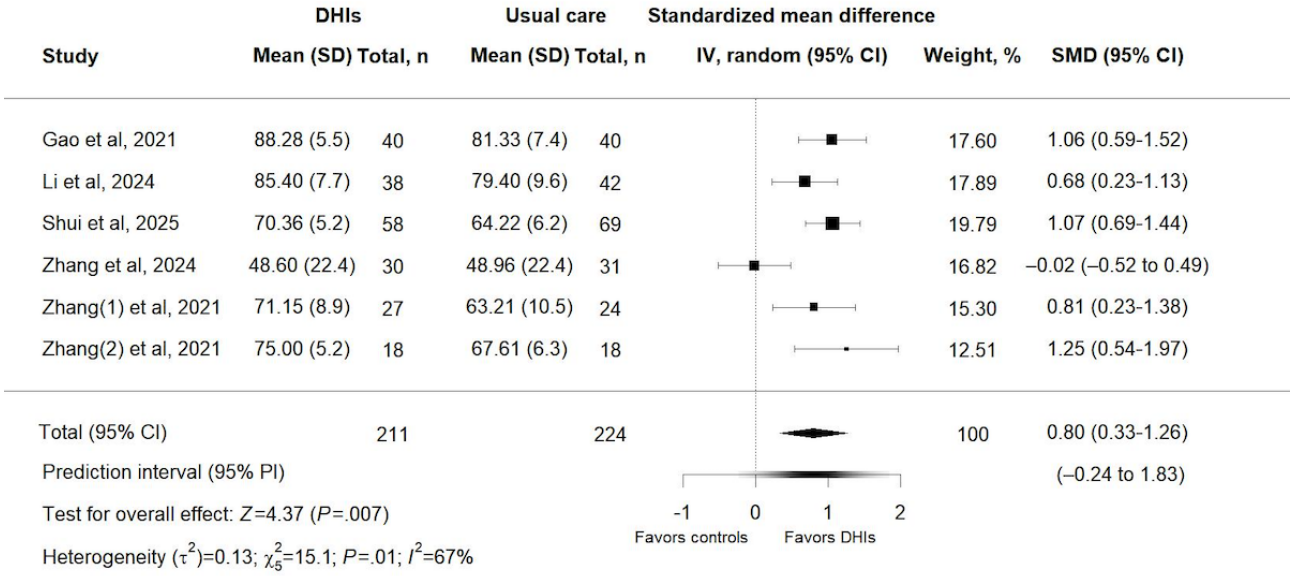

Effects of DHIs on Functional Independence

The effects of DHIs on functional independence were measured in 6 trials [29,38,40,43,46,47] involving 369 participants (Figure 4). Based on a random-effects model, pooled results showed a statistically significant beneficial effect of DHIs on enhancing functional independence in older patients with hip fractures, with an SMD of 1.23 (95% CI 0.34-2.11; *P*=.02). In stark contrast, the 95% PI (–0.98 to 3.34) reflects the wide spectrum of individual responses to DHIs, indicating that while DHIs yield a favorable mean effect, some patients may experience no improvement or even a slight decline in functional independence, whereas others could achieve marked benefits. This pronounced variability aligns with the extremely high heterogeneity (χ_5^2=39.7; *P*=.02; τ^2=0.62; I^2=87%) observed in the meta-analysis, highlighting the critical role of patient-specific and trial-related factors in modulating the effects of DHIs.

XSL•FO
RenderX

Figure 4. Forest plot for the effect of digital health interventions on functional independence [29,38,40,43,46,47]. DHI: digital health intervention; SMD: standardized mean difference.

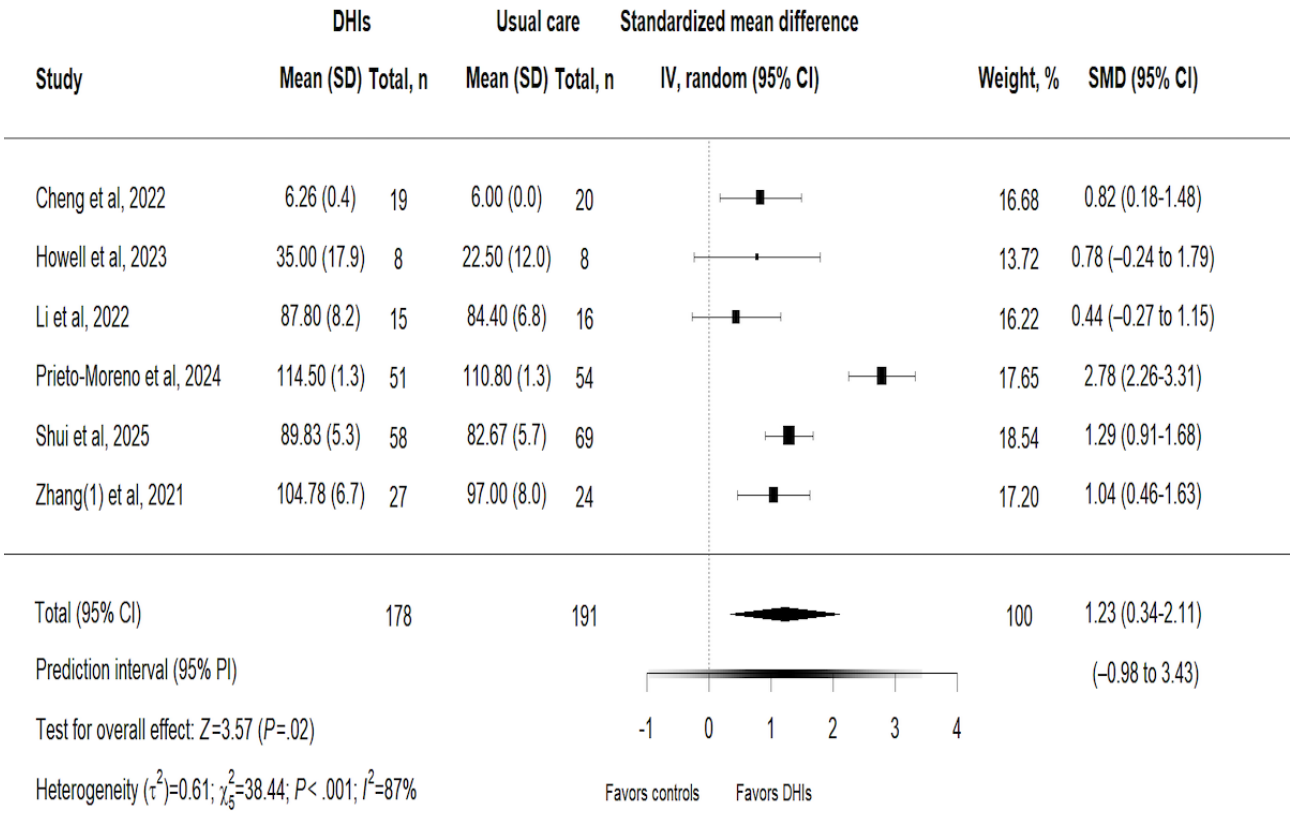

Effects of DHIs on Balance Function and Risk of Falling

In terms of balance function and risk of falling, it was measured in 5 trials [38,40,45,46,48] involving 163 participants. Based on a random-effects model, the results showed that compared to usual care, DHIs had no significant effects on improving balance function and reducing risk of falling (SMD –0.09, 95% CI –0.99 to 0.81; *P*=.80). The 95% PI (–2.07 to 1.89) was further estimated to delineate the variability of individual treatment responses and indicated that while the mean effect is negligible, some patients might experience meaningful improvements in balance function with DHIs, whereas others could exhibit worse outcomes. This pronounced variability in individual responses corresponds well with the high heterogeneity (χ_4^2=16.7; *P*=.002;τ^2=0.40; I^2=75%; Figure 5).

XSL•FO
RenderX

Figure 5. Forest plot for the effect of digital health interventions on balance function and risk of falling [38,40,45,46,48]. DHI: digital health intervention; SMD: standardized mean difference.

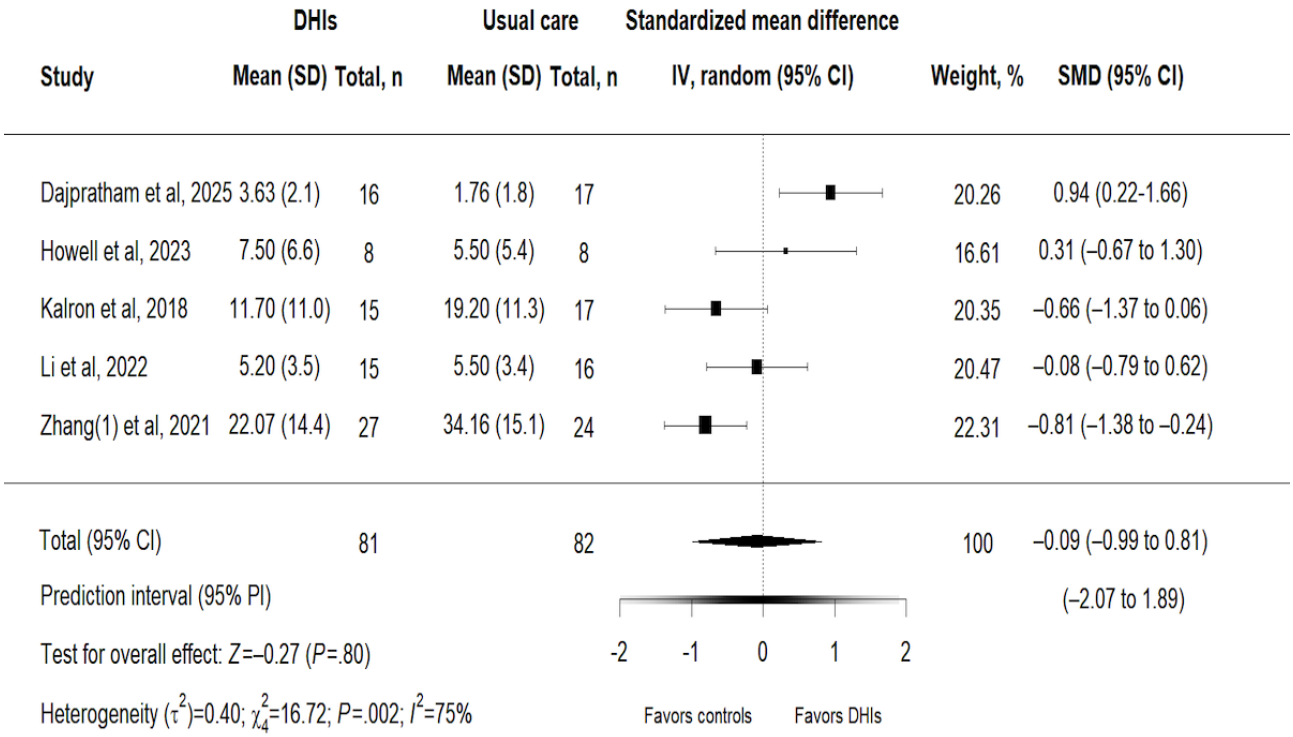

Effects of DHIs on Quality of Life

The effect of DHIs on QoL was measured in 8 trials [40-44,46,47,49] involving 501 participants. This meta-analysis adopted a random-effects model for data pooling. The overall results indicated that there was no statistically significant difference in QoL between groups, with an SMD of 0.65 (95% CI –0.21 to 1.50; P=.12). The 95% PI (–1.83 to 3.13) was further calculated, and substantial statistical heterogeneity was observed among the included trials (χ_7^2=89.5; P<.001; τ^2=0.97; I^2=93%; Figure 6).

XSL•FO
RenderX

Figure 6. Forest plot for the effect of digital health interventions on quality of life [46,49,44,47,43,42,41,40]. DHI: digital health intervention; SMD: standardized mean difference.

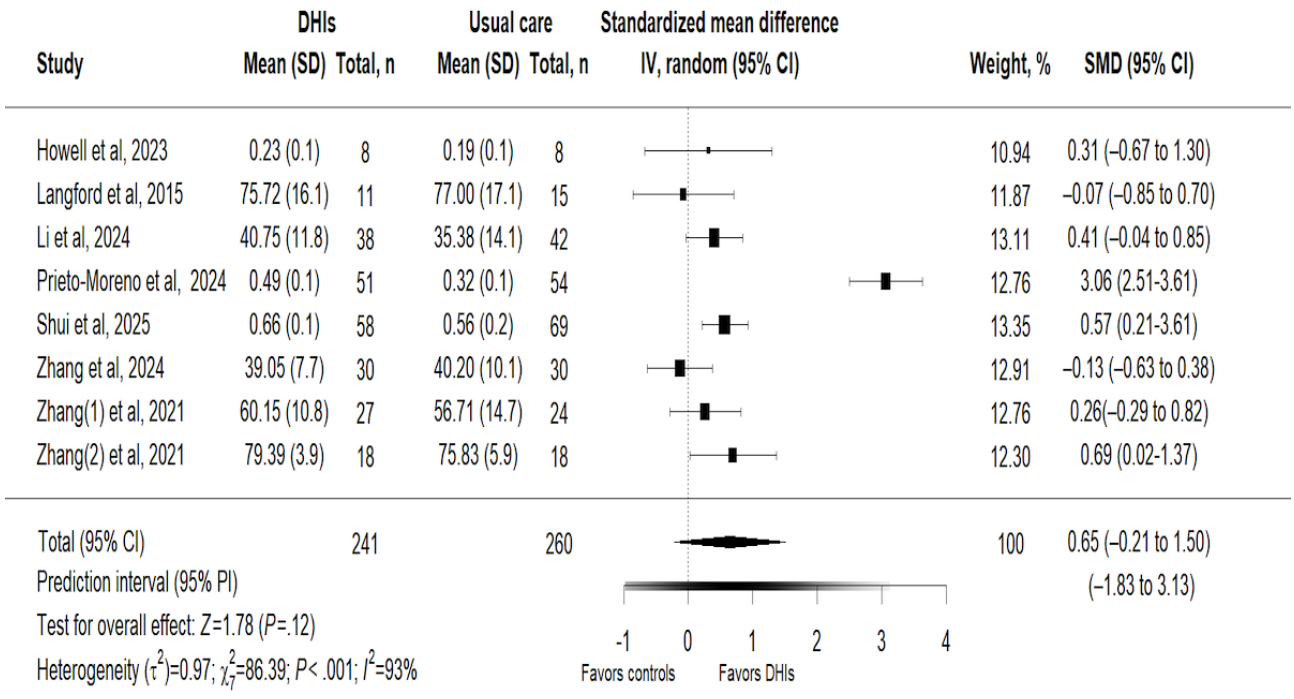

Effects of DHIs on Intervention Adherence

Due to the limited reporting of intervention adherence data across studies, a meta-analysis was not possible; only a single available study reported a 70% intervention adherence rate in the intervention group [48].

Subgroup Analyses

Subgroup Analysis by Type of Intervention

A subgroup analysis based on the type of DHIs was planned. However, the number of studies in each predefined intervention category (mHealth; n=9 [29,39,41-45,48,49], eHealth; n=3 [38,40,47], and digital device; n=1 [46]) fell below the threshold of 3 for all outcomes, and the planned subgroup analysis based on the type of DHIs was therefore forfeited due to the inability to provide meaningful quantitative estimates.

Subgroup Analysis by Intervention Duration

Subgroup analysis was conducted based on the intervention duration (<3 months vs ≥3 months) to explore the effects of DHIs on hip function. The subgroup analysis for hip function revealed no statistically significant difference between subgroups (χ_1^2=0.1; *P*=.75), with interventions lasting ≥3 months showing a greater effect (SMD 0.89, 95% CI 0.36-1.41; I^2=7%; *P*=.41) compared to those lasting <3 months (Figure 7). The test for subgroup differences in functional independence suggested a marginally significant variation in effects (χ_1^2=2.93; *P*=.09), indicating that intervention duration may be a potential contributor to the overall heterogeneity of the meta-analysis, albeit not reaching formal statistical significance, while the point estimate still favored shorter intervention duration (SMD 0.67, 95% CI 0.12-1.23; I^2=0%; *P*=.72; Figure 8).

XSL•FO
RenderX

Figure 7. Subgroup analysis on hip function [39,42,41,44,43,40]. DHI: digital health intervention; RE: Random effect; SMD: standardized mean difference.

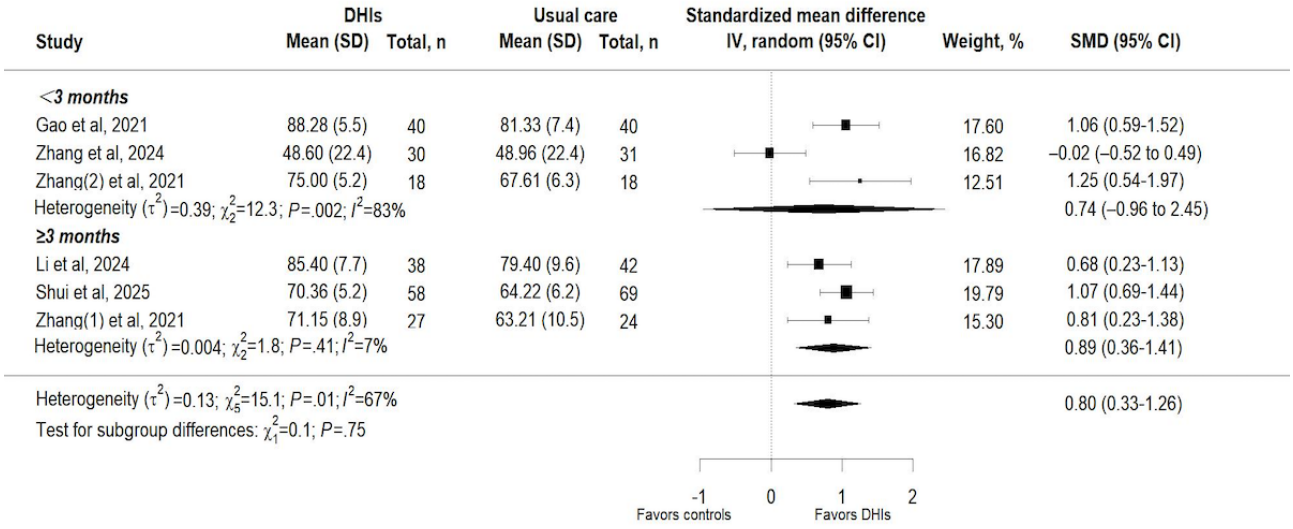

Figure 8. Subgroup analysis on functional independence [29,38,40,43,46,47]. DHI: digital health intervention; SMD: standardized mean difference.

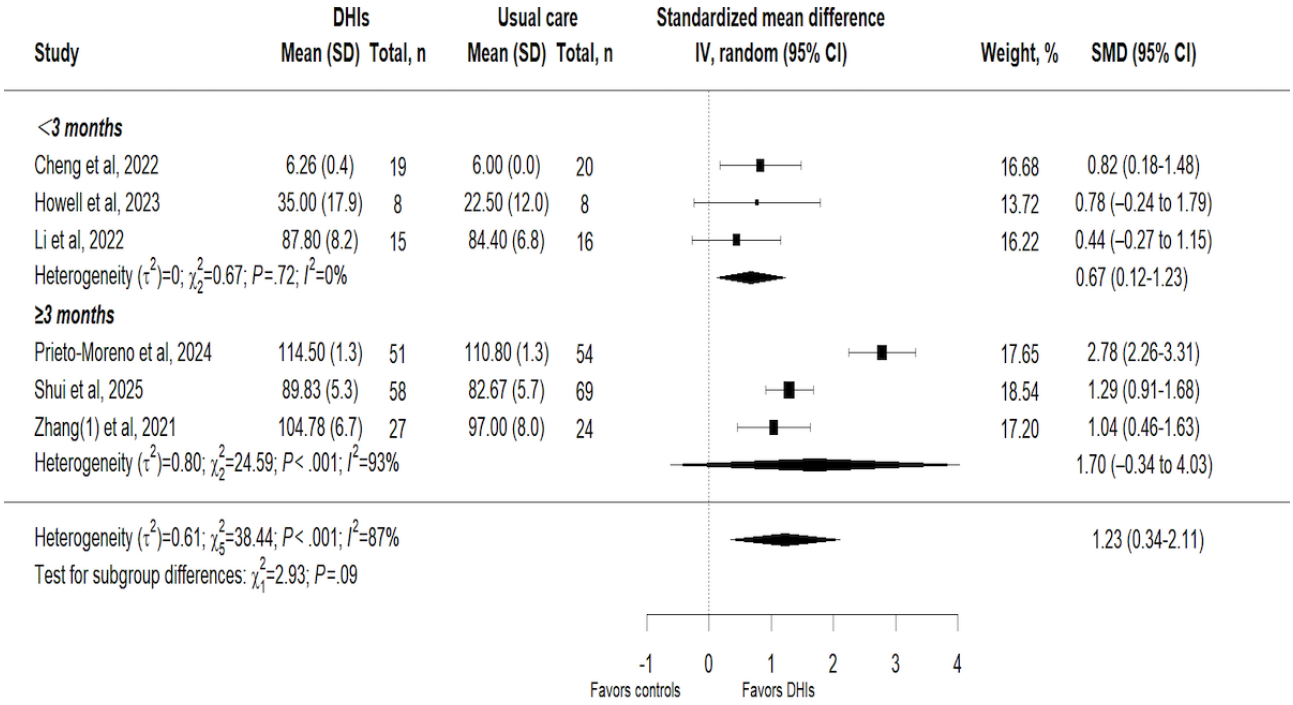

Small-Study Effects and Sensitivity Analysis

The funnel plot for hip function underlying the meta-analyses was symmetrical, and the Egger test was conducted to assess potential small-study effects, and no statistically significant evidence of such effects was identified in this meta-analysis (z-score=0.84; P=.40; Multimedia Appendix 4). Sensitivity analysis, performed by sequentially removing each study, confirmed that no single study unduly influenced the overall pooled estimate for hip function, demonstrating the robustness of this finding (Multimedia Appendix 4).

XSL•FO
RenderX

Discussion

Principal Findings

This meta-analysis confirms that DHIs confer statistically reliable average benefits for hip function and functional independence in older adults with hip fractures, whereas their effects on balance and QoL remain inconclusive. A wide 95% PI, however, indicates substantial real-world variability in individual DHI responses, restricting broad clinical deployment. Subgroup analyses also reflect this inconsistency—with intervention duration providing partial explanatory power—but the limited number of included studies necessitates further research to elucidate the underlying sources of such heterogeneity. Additionally, intervention adherence was reported in only 1 study [48], precluding quantitative assessment of DHI-related effects on this outcome. These results underscore that the potential of DHIs to promote functional and psychological recovery in this population requires further validation.

Low-to-moderate-certainty evidence, as determined by the GRADE assessment, suggests that DHIs are superior to control conditions in improving hip function and functional independence among older adults with hip fractures. This finding aligns with previous research supporting the role of telerehabilitation interventions in enhancing physical function following joint surgery and has suggested that this is related to the ability of DHIs to provide effective communication and feedback, as well as an educational component [21,50,51]. Health education during the postoperative rehabilitation stage is crucial for patients' recovery. Postoperative patients need to receive numerous recommendations, including early ambulation, exercise and rehabilitation programs, preventing postoperative complications, modifying home hazards, and taking nutrition supplements [21,52]. DHIs mostly adopt mobile apps as the mode of delivery, which can provide real-time or asynchronous communication and feedback. It can break down the boundaries of time and space, providing an effective communication platform between physiotherapists and patients, which is conducive to accurately transmitting rehabilitation knowledge and ultimately achieving positive functional recovery effects [53,54].

In addition, most DHIs systems incorporate exercise training, allowing physiotherapists to provide exercise guidance through voice, video, and other means, while receiving feedback from patients [55]. This helps patients correct improper movements promptly and ensures training quality [56]. Studies have shown that appropriate exercise after orthopedic surgery can counteract muscle weakness and functional disorders that may be caused by limited limb usage and also improve the range of joint motion, which is conducive to the recovery of the hip joint and physical function [21,57]. Currently, DHIs predominantly use combined exercises, such as resistance or strength training paired with balance training, to maximize functional improvement. Resistance training and strength training can not only enhance muscle strength but also lead to the upregulation of bone turnover and subsequent bone deposition, which has a positive regulatory effect on bone health [58-60]. Balance training is expected to improve balance function and prevent falls.

However, it is critical to emphasize that the effects of DHIs exhibited substantial real-world variability. For hip function, the 95% PI spanned from marginally negative to large positive treatment responses, which quantifies the true heterogeneity across populations and settings far more pragmatically than I^2 alone. For functional independence, the 95% PI covered an even broader range of individual patient responses, consistent with extremely high between-study heterogeneity. This pattern indicates inconsistent DHIs' self-efficacy across studies and patient subgroups in real-world settings. In addition, variations in study design (eg, differences in individual characteristics, DHIs delivery modes, and control group interventions) may have introduced bias, potentially confounding the pooled effect estimates. The low-to-moderate certainty of evidence reflects limitations, including imprecision in effect estimates and inconsistency across trials. Collectively, these factors mandate cautious clinical translation of the findings.

Moderate-certainty evidence suggested that the effects of DHIs on balance function and fall risk were uncertain, which contrasts with previous studies. Previous studies have reported the positive effects of DHIs on improving balance ability in populations, such as older adults and patients with neurological diseases [61-63]. A study investigating the mechanism of balance improvement in older adults after balance training found that balance training can increase the duration of cocontraction of the soleus or tibialis anterior, which can enhance joint stability and thereby improve the balance control ability of older adults [64]. However, the effects of balanced training also seem to have a dose-response relationship [65]. Earlier studies showed that a high practice intensity and an increased session frequency led to significant retention over the longer term [66,67]. However, increasing the intensity, frequency, and duration of such protocols could also have disadvantages, such as activity-induced fatigue and reduced participation [68]. It has also been found that the plasticity of the human brain provides physiological preparation for the restoration of balance function, but this still requires long-term persistence [69]. Therefore, the fact that no significant improvement in balance function was observed in this systematic review might be due to the lack of a suitable balance training program for older patients with hip fractures, or the duration of intervention was not sufficient to restore balance function.

It should also be noted that the wide 95% PI for balance function and fall risk reflects pronounced variability in individual patient responses, aligning with the high between-study heterogeneity observed. Additionally, many trials lacked standardized assessment tools for these outcomes and failed to implement outcome assessor blinding, and these methodological limitations likely introduced bias, confounded pooled effect precision, and contributed to the observed heterogeneity. The evidence for effects of DHIs on balance function and fall risk is rated as moderate certainty, owing to imprecise effect sizes. Collectively, these factors indicate that the null average effect should not be interpreted as definitive proof of DHIs' ineffectiveness, and instead, the broad PI underscores the potential for

XSL•FO
RenderX

subgroup-specific benefits or harms that aggregate analyses cannot capture.

Contrary to our expectations, low-certainty evidence showed no effect of DHIs on QoL. Molina-Garcia et al [70] indicated that telerehabilitation interventions can significantly improve QoL in patients with musculoskeletal disorders. This difference may be attributed to the fact that Molina-Garcia et al [70] focused on a broader range of musculoskeletal diseases, while our meta-analysis specifically targeted patients with hip fractures—a population typically characterized by advanced age, multiple diseases, and slow recovery [71]. Additionally, the intensity and duration of DHIs conducted in existing studies may not be sufficient to overcome the profound physical and psychological challenges associated with this injury and restore the QoL of older patients to the level before the fracture [12,72]. A cohort study also showed that social support and pain while walking can predict QoL among older adults 3 months after discharge from rehabilitation following a surgical hip fracture repair [73]. It indicates that the failure to effectively address pain may indirectly lead to a lack of statistically significant improvement in QoL outcomes. Therefore, in order to improve the psychological outcomes of older patients with hip fractures, it is best to integrate pain management and social support measures into DHIs.

The wide 95% PI also reflects pronounced variability in QoL in individual patient responses to DHIs, which aligns closely with the extremely high between-study heterogeneity observed. Included trials lacked a uniform QoL instrument and failed to account for confounding factors, such as baseline mental health status and social support levels, introducing methodological bias that may have inflated heterogeneity and compromised the precision of pooled estimates. Consequently, the evidence for effects of DHIs on QoL is rated as low certainty, constrained by imprecise effect sizes and inconsistent reporting of long-term QoL trajectories. Collectively, these findings indicate that the null average effect should not be equated with a lack of DHIs' efficacy, and instead, the broad PI highlights the potential for substantial subgroup-specific benefits or null effects that aggregate analyses cannot fully delineate.

Although a formal meta-analysis on intervention adherence was not feasible due to the lack of quantifiable data, the role of adherence in improving prognosis cannot be overlooked. In the included studies, Dajpratham et al [48] reported that teleresistance exercise showed an adherence rate of 70%, higher than that of the control group [48]. A recently published meta-analysis also indicates that telerehabilitation is associated with higher levels of medical adherence compared to conventional care [74]. All these reflect the potential of DHIs in improving intervention adherence. Theoretically, the convenience, reminders, and engaging nature of DHIs are posited to enhance adherence. However, future research must prioritize the standardized measurement and reporting of adherence metrics (eg, login frequency and exercise completion rates) to quantitatively validate this claim and understand its relationship with clinical outcomes.

Subgroup Analyses

Given the high heterogeneity, we performed subgroup analyses based on the type and duration of the intervention. Although only 1 study [46] used digital device-based DHIs (wearable devices), which prevented subgroup analysis by the type of intervention, the potential of digital device-based DHIs, especially wearable devices, to provide objective and continuous data on physical activity and movement patterns is immense [75,76]. It is reported that wearable devices have been used to collect body posture and movement data, providing real-time feedback to therapists to ensure rehabilitation programs are precisely tailored to individual needs [46]. It is evident that their role in enabling precise dosing of rehabilitation and detecting early signs of functional decline is a promising frontier [76]. The emerging field of wearable technology warrants dedicated investigation in the future, with more trials to isolate its specific contribution to the effectiveness of DHIs for hip fracture recovery [75].

Subgroup analyses stratified by intervention duration were performed to explore sources of heterogeneity in DHIs' effects on hip function, with findings interpreted in light of heterogeneity and risk of bias. The test for subgroup differences revealed no statistically significant effects, indicating that intervention duration was not a contributor to the overall heterogeneity observed in the meta-analysis. The 3 months subgroup yielded a significant pooled treatment effect, which suggests that prolonged intervention duration may mitigate outcome variability by enabling cumulative rehabilitation gains and sustained behavior modification. Additionally, more uniform long-term intervention protocols across included trials may have reduced methodological heterogeneity. This result may also be related to the dose-response relationship of rehabilitation [77,78]. A longer intervention period allows for a greater cumulative "dose" of therapy, which is necessary to counteract the severe physical deconditioning post fracture.

Interestingly, the subgroup analysis on functional independence based on intervention duration yielded inconsistent findings, with interventions lasting <3 months appearing more effective than those lasting 3 months. Notably, however, each of the 3 individual studies [3,40,43] in the 3 months subgroup demonstrated a statistically significant improvement in functional independence for the DHIs group relative to controls. When pooled, these results became statistically insignificant—a discrepancy attributable to marked between-study heterogeneity within this subgroup. These 3 studies used the same measurement tools, and the degree of improvement evaluated by Prieto Moreno et al [47] was much higher than that of the other 2 studies [40,43], which might be related to their higher frequency, intervention, and usage rate. Prieto Moreno et al [47] had an intervention frequency of twice a week and reported a relatively high app usage rate. The other 2 studies, however, had no reports on usage rate or adherence rate and had a lower intervention frequency.

Collectively, these findings underscore the complexity of optimizing DHI protocols for older adults with hip fractures. While prolonged intervention duration conferred consistent benefits for hip function, its effects on functional independence

XSL•FO
RenderX

appear contingent on ensuring sufficient intervention frequency and monitoring adherence. Future research should therefore prioritize standardized reporting of intervention dose and adherence metrics, alongside rigorous protocol design that specifies both duration and frequency of DHIs delivery, to clarify and replicate these subgroup-specific effects.

Comparison with Prior Work

With the increasing popularity of DHIs, scholars are paying more and more attention to their effects in the field of bone and joint rehabilitation. Several similar systematic reviews and meta-analyses have been published. While Molina-Garcia et al [70] provided a broad overview of DHIs across musculoskeletal disorders (including hip fractures), our study focused specifically on older patients with hip fractures, which made the evidence we provided direct and specifically applicable to clinicians and rehabilitation experts managing older patients with hip fractures, enabling them to make more precise decisions regarding DHIs.

Two published systematic reviews and meta-analyses have focused on the older population with hip fractures, as we have. One of the studies was published by Pliannuom et al [21] in 2024, focusing only on the impact of DHIs on functional outcomes, while we simultaneously evaluated functional and psychological outcomes and intervention adherence [21]. Furthermore, their paper was limited by the number of studies (only 6 RCTs [39-44] were included), and they were unable to conduct subgroup analyses to further clarify which subgroups could benefit the most from DHIs. Compared with another meta-analysis by Xiao et al [74], our population mainly focuses on older patients with hip fractures rather than all patients after hip fractures and replacement surgeries. Furthermore, compared with Xiao et al [74], our analysis included more of the latest published studies and conducted a GRADE assessment on the included studies to ensure the quality of the studies and the reliability of the results of the meta-analysis [47,48]. Our study and the study by Xiao et al [74] both performed subgroup analyses of hip function based on intervention duration, and both found that a longer intervention period was associated with a greater improvement effect on hip function [74]. However, in the study by Xiao et al [74], the 6-month subgroups showed extremely high heterogeneity (I^2=95%), indicating instability. In our study, the heterogeneity of the ≥3 months subgroup was low (I^2=7%), which indicates that our findings are more robust and reliable.

However, our review indicates that only low-to-moderate certainty evidence supports the effects of DHIs on our outcomes of interest. This suggests that while there may be some positive impact, the evidence is not strong enough to draw definitive conclusions. The variability in study quality, sample sizes, and methodologies may contribute to the uncertainty, highlighting the need for more robust and well-conducted trials to strengthen the evidence base. Therefore, the results should be interpreted carefully, as future research might yield divergent findings.

Strengths and Limitations

This meta-analysis followed the guidelines for performing rigorous systematic reviews. We also proposed a rigorous screening and search strategy to identify the most comprehensive literature. Sensitivity analyses revealed robust findings. Consequently, the results of our review support a widespread belief. However, there are also limitations in this systematic review. First, high statistical heterogeneity was observed in some results, which could not be fully explained by the subgroup analysis we prespecified. This may stem from the diversity of intervention components (such as types, dose, and intervention providers), control groups, and patient characteristics. Second, the limited number of studies on certain outcomes has restricted the statistical power and accuracy of estimates. Third, some studies had high risk or some concerns, particularly in terms of randomization and selection of reported results. This may have influenced the overall findings. Fourth, only one study [48] included in our review reported intervention adherence, which limits the quantitative evaluation for adherence of DHIs. Finally, as most of the research (61.5%) was conducted in China, the universality of the research results may be limited, and their applicability to other health care systems and cultures needs to be verified.

Further Research and Clinical Implications

For further research, the sources of heterogeneity should be identified based on patient characteristics (eg, prefracture functional level), intervention dose (eg, frequency and duration), as our meta-analysis observed significant heterogeneity. The high heterogeneity observed across studies also underscores the need for standardized protocols of DHIs. Future research should also aim to identify the most effective components of DHIs and develop standardized intervention frameworks to reduce variability and improve the generalizability of findings. Second, as the existing RCTs lack assessment of psychological outcomes and intervention adherence, future RCTs should consider including these outcomes, which are particularly important for a more comprehensive evaluation of the effect of DHIs. Third, due to the limited number of studies using digital devices for DHIs, subgroup analysis based on intervention types was not achieved in this systematic review. Future research needs to explore the effects of digital devices to further clarify the impact of intervention types on improving the functional and psychological outcomes of patients with hip fractures. Fourth, future research should further explore the dose-response of exercise rehabilitation on the recovery of physical function after hip fractures to achieve the best intervention effect. Clinically, our research results indicate that DHIs have potential value in improving hip function and functional independence in older patients with hip fractures. Clinicians should consider incorporating DHIs into clinical practice and using them as routine care after hip fractures in older adults. This is of great significance for optimizing the allocation of medical resources and expanding the scope of rehabilitation services.

Conclusions

This review incorporates the latest RCTs and comprehensively assesses both functional and psychological outcomes of DHIs in older patients with hip fractures, distinct from prior studies focusing solely on functional outcomes. While 95% CI supports the potential of DHIs to improve hip function and functional independence, the wide 95% PI indicating substantial real-world response variability calls for cautious interpretation, informing

XSL•FO
RenderX

the design of targeted DHI-based rehabilitation regimens, and warrants further research into optimal techniques and dosages in clinical practice.

Acknowledgments

No artificial intelligence tools were used in any aspect of this manuscript, from study implementation to manuscript preparation. All content reflects the authors' original work. We also sincerely acknowledge Dr Shang Wenru from the Hospital Management Center of Shanghai Sixth People's Hospital for peer-reviewing our search strategy by verifying search terms, database selection, and eligibility criteria prior to retrieval.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets generated or analyzed during this systematic review are available from the corresponding author on reasonable request.

Authors' Contributions

QZ and QFL were responsible for conceptualization, methodology, supervision, validation, and writing–original draft. QFL handled project administration and resources. WF and QZ completed data curation, formal analysis, software, and visualization. WF, QZ, and QL contributed to writing–review & editing. QZ and QFL are considered co-corresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISIMA 2020 checklist.
[DOCX File , 28 KB - jmir_v28i1e79563_app1.docx]

Multimedia Appendix 2
PRISMA-S checklist.
[DOCX File , 16 KB - jmir_v28i1e79563_app2.docx]

Multimedia Appendix 3
Search strategy.
[DOCX File , 19 KB - jmir_v28i1e79563_app3.docx]

Multimedia Appendix 4
Sensitivity analysis.
[DOCX File , 88 KB - jmir_v28i1e79563_app4.docx]

References

1. Neuman MD, Silber JH, Magaziner JS, Passarella MA, Mehta S, Werner RM. Survival and functional outcomes after hip fracture among nursing home residents. JAMA Intern Med 2014;174(8):1273-1280 [FREE Full text] [doi: 10.1001/jamainternmed.2014.2362] [Medline: 25055155]
2. Baji P, Patel R, Judge A, Johansen A, Griffin J, Chesser T, REDUCE Study Group. Organisational factors associated with hospital costs and patient mortality in the 365 days following hip fracture in England and Wales (REDUCE): a record-linkage cohort study. Lancet Healthy Longev 2023;4(8):e386-e398 [FREE Full text] [doi: 10.1016/S2666-7568(23)00086-7] [Medline: 37442154]
3. Sing C, Lin TC, Bartholomew S, Bell JS, Bennett C, Beyene K, et al. Global epidemiology of hip fractures: secular trends in incidence rate, post-fracture treatment, and all-cause mortality. J Bone Miner Res 2023;38(8):1064-1075 [FREE Full text] [doi: 10.1002/jbmr.4821] [Medline: 37118993]
4. Gullberg B, Johnell O, Kanis J. World-wide projections for hip fracture. Osteoporos Int 1997;7(5):407-413. [doi: 10.1007/pl00004148] [Medline: 9425497]

XSL•FO
RenderX

5. Feng J, Zhang C, Li B, Zhan S, Wang S, Song C. Global burden of hip fracture: The Global Burden of Disease Study. Osteoporos Int 2024;35(1):41-52. [doi: 10.1007/s00198-023-06907-3] [Medline: 37704919]
6. Nordström P, Gustafson Y, Michaëlsson K, Nordström A. Length of hospital stay after hip fracture and short term risk of death after discharge: a total cohort study in Sweden. BMJ 2015;350:h696 [FREE Full text] [doi: 10.1136/bmj.h696] [Medline: 25700551]
7. Branca Vergano L, Monesi M, Vicenti G, Bizzoca D, Solarino G, Moretti B. Posterior approaches in malleolar fracture: when, why and how. J Biol Regul Homeost Agents 2020;34(3 Suppl. 2):89-95. [Medline: 32856446]
8. Araiza-Nava B, Méndez-Sánchez L, Clark P, Peralta-Pedrero ML, Javaid MK, Calo M, et al. Short- and long-term prognostic factors associated with functional recovery in elderly patients with hip fracture: a systematic review. Osteoporos Int 2022;33(7):1429-1444. [doi: 10.1007/s00198-022-06346-6] [Medline: 35247062]
9. Yang T, Kuo Y, Huang S, Chen Y. Minimal short-term decline in functional performance and quality of life predicts better long-term outcomes for both in older Taiwanese adults after hip fracture surgery: a prospective study. J Orthop Surg Res 2023;18(1):791 [FREE Full text] [doi: 10.1186/s13018-023-04278-3] [Medline: 37872535]
10. Nightingale S, Holmes J, Mason J, House A. Psychiatric illness and mortality after hip fracture. Lancet 2001;357(9264):1264-1265 [FREE Full text] [doi: 10.1016/S0140-6736(00)04421-4] [Medline: 11418154]
11. Milton-Cole R, Ayis S, Lambe K, O'Connell MDL, Sackley C, Sheehan KJ. Prognostic factors of depression and depressive symptoms after hip fracture surgery: systematic review. BMC Geriatr 2021;21(1):537 [FREE Full text] [doi: 10.1186/s12877-021-02514-1] [Medline: 34627160]
12. Alexiou K, Roushias A, Varitimidis S, Malizos K. Quality of life and psychological consequences in elderly patients after a hip fracture: a review. Clin Interv Aging 2018;13:143-150 [FREE Full text] [doi: 10.2147/CIA.S150067] [Medline: 29416322]
13. Wang J, Tong Y, Jiang Y, Zhu H, Gao H, Wei R, et al. The effectiveness of extended care based on Internet and home care platform for orthopaedics after hip replacement surgery in China. J Clin Nurs 2018;27(21-22):4077-4088. [doi: 10.1111/jocn.14545] [Medline: 29851157]
14. Rosas Hernández AM, Alarcón T, Menéndez-Colino R, Martín Maestre I, González-Montalvo JI. Factors affecting exercise program adherence in patients with acute hip fracture and impact on one-year survival. Braz J Phys Ther 2020;24(6):479-487 [FREE Full text] [doi: 10.1016/j.bjpt.2019.07.008] [Medline: 31378633]
15. Taylor N, Rimayanti MU, Peiris CL, Snowdon DA, Harding KE, Semciw AI, et al. Hip fracture has profound psychosocial impacts: a systematic review of qualitative studies. Age Ageing 2024;53(9):afae194. [doi: 10.1093/ageing/afae194] [Medline: 39238124]
16. Crotty M. Rehabilitation interventions for improving physical and psychosocial functioning after hip fracture in older people. Cochrane Database Syst Rev 2010:20091644. [doi: 10.1002/14651858.cd007624]
17. Handoll H, Cameron ID, Mak J, Panagoda CE, Finnegan TP. Multidisciplinary rehabilitation for older people with hip fractures. Cochrane Database Syst Rev 2021;11(11):Cd007125 [FREE Full text] [doi: 10.1002/14651858.CD007125.pub3] [Medline: 34766330]
18. Blackburn J, Yeowell G. Patients' perceptions of rehabilitation in the community following hip fracture surgery. a qualitative thematic synthesis. Physiotherapy 2020;108:63-75. [doi: 10.1016/j.physio.2020.02.001] [Medline: 32711229]
19. Lewis S, Macey R, Lewis J, Stokes J, Gill J, Cook JA, et al. Surgical interventions for treating extracapsular hip fractures in older adults: a network meta-analysis. Cochrane Database Syst Rev 2022;2(2):CD013405 [FREE Full text] [doi: 10.1002/14651858.CD013405.pub2] [Medline: 35142366]
20. Recommendations on digital interventions for health system strengthening. World Health Organization.: World Health Organization; 2019. URL: https://www.who.int/publications/i/item/9789241550505 [accessed 2026-01-11]
21. Pliannuom S, Pinyopornpanish K, Buawangpong N, Wiwatkunupakarn N, Mallinson PAC, Jiraporncharoen W, et al. Characteristics and effects of home-based digital health interventions on functional outcomes in older patients with hip fractures after surgery: systematic review and meta-analysis. J Med Internet Res 2024;26:e49482. [doi: 10.2196/49482]
22. Backman C, Papp S, Harley A, Skidmore B, Green M, Shah S, et al. Patient-clinician digital health interventions for the hip fracture population: a scoping review. BMC Health Serv Res 2023;23(1):1052 [FREE Full text] [doi: 10.1186/s12913-023-09784-y] [Medline: 37784118]
23. Zhang J, Yang M, Ge Y, Ivers R, Webster R, Tian M. The role of digital health for post-surgery care of older patients with hip fracture: a scoping review. Int J Med Inform 2022;160:104709 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104709] [Medline: 35144101]
24. Chan JKY, Klainin-Yobas P, Chi Y, Gan JKE, Chow G, Wu XV. The effectiveness of e-interventions on fall, neuromuscular functions and quality of life in community-dwelling older adults: a systematic review and meta-analysis. Int J Nurs Stud 2021;113:103784. [doi: 10.1016/j.ijnurstu.2020.103784] [Medline: 33120138]
25. Morita PP, Sahu KS, Oetomo A. Health monitoring using smart home technologies: scoping review. JMIR Mhealth Uhealth 2023;11:e37347. [doi: 10.2196/37347]
26. Oyibo K, Wang K, Morita PP. Using smart home technologies to promote physical activity among the general and aging populations: scoping review. J Med Internet Res 2023;25:e41942 [FREE Full text] [doi: 10.2196/41942] [Medline: 37171839]

XSL•FO
RenderX

27. Gilboa Y, Maeir T, Karni S, Eisenberg ME, Liebergall M, Schwartz I, et al. Effectiveness of a tele-rehabilitation intervention to improve performance and reduce morbidity for people post hip fracture - study protocol for a randomized controlled trial. BMC Geriatr 2019;19(1):135 [FREE Full text] [doi: 10.1186/s12877-019-1141-z] [Medline: 31109289]
28. Pfeiffer K, Kampe K, Klenk J, Rapp K, Kohler M, Albrecht D, et al. Effects of an intervention to reduce fear of falling and increase physical activity during hip and pelvic fracture rehabilitation. Age Ageing 2020;49(5):771-778. [doi: 10.1093/ageing/afaa050] [Medline: 32832985]
29. Cheng KC, Lau KMK, Cheng AS, Lau TSK, Lau FOT, Lau MCH, et al. Use of mobile app to enhance functional outcomes and adherence of home-based rehabilitation program for elderly with hip fracture: a randomized controlled trial. Hong Kong Physiother J 2022;42(2):99-110 [FREE Full text] [doi: 10.1142/S101370252250010X] [Medline: 37560168]
30. Câmara SMA, Falvey JR, Orwig D, Gruber-Baldini AL, Auais M, Feng Z, et al. Associations between living alone, social interactions, and physical performance differ by sex: results from the Baltimore Hip Studies. J Am Geriatr Soc 2023;71(9):2788-2797. [doi: 10.1111/jgs.18403] [Medline: 37171145]
31. Shams F, Tai AM, Kim J, Boyd M, Meyer M, Kazemi A, et al. Adherence to e-health interventions for substance use and the factors influencing it: systematic Review, meta-analysis, and meta-regression. Digit Health 2023;9:20552076231203876 [FREE Full text] [doi: 10.1177/20552076231203876] [Medline: 37780062]
32. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
33. Rethlefsen M, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
34. Classification of digital health interventions v1.0.: World Health Organization; 2018. URL: https://www.who.int/publications/i/item/WHO-RHR-18.06 [accessed 2026-01-11]
35. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Cochrane Bias Methods Group, Cochrane Statistical Methods Group. The cochrane collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928 [FREE Full text] [doi: 10.1136/bmj.d5928] [Medline: 22008217]
36. Xu YP, Yu XY, Fu WJ, Tao Y, Wu JY, He Q, et al. Introduction to the methods of missing standard deviation conversion and standardized mean difference estimation in Meta-Analysis. Chinese journal of evidence-based cardiovascular medicine, 2017 Feb 1;8(12):1412-1415.
37. Zangger G, Bricca A, Liaghat B, Juhl CB, Mortensen SR, Andersen RM, et al. Benefits and harms of digital health interventions promoting physical activity in people with chronic conditions: systematic review and meta-analysis. J Med Internet Res 2023;25:e46439 [FREE Full text] [doi: 10.2196/46439] [Medline: 37410534]
38. Li CT, Hung GK, Fong KN, Gonzalez PC, Wah S, Tsang HW. Effects of home-based occupational therapy telerehabilitation via smartphone for outpatients after hip fracture surgery: a feasibility randomised controlled study. J Telemed Telecare 2022;28(4):239-247. [doi: 10.1177/1357633X20932434] [Medline: 32594832]
39. Gao S, Wang Y, Zhang G, Zhang W. Rehabilitation guidance for hip fracture patients during the COVID-19 pandemic using chat software: a new model. J Back Musculoskelet Rehabil 2021;34(3):337-342. [doi: 10.3233/BMR-200324] [Medline: 33896813]
40. Zhang Y. Research on the Application of Internet + Rehabilitation Management System in Family Rehabilitation after Hip Fracture Surgery in the Elderly [dissertation]. Tianjin, China: Tianjin University of Sport; 2021.
41. Zhang GZ, Ding L, Lin QJ, Gu LD, Wang X. Evaluation of the remote rehabilitation effect supported by network media for patients after total hip arthroplasty. Chinese Journal of Bone and Joint Injury 2021 Oct;36(10):1053-1055.
42. Zhang X, Lu YT, Zhang C, Li H, Lu SD, Zhang ZA. Comparison of the therapeutic effects of remote home rehabilitation and inpatient rehabilitation after total hip replacement. Journal of Tongji University(Medical Science) 2024 Feb;45(1):57-65.
43. Shui XM, Xu Q, Sun YT, Xu MP, Kong HX. Clinical study on the impact of family remote rehabilitation training programs on postoperative rehabilitation outcomes of hip fractures in the elderly. Chinese Journal of Bone and Joint 2025 Jul 30;14(7):640-644.
44. Li ZW, Fan JJ, Wang TR, Yuan Z. Evaluation of the effect of remote rehabilitation on elderly patients and their family caregivers after intertrochanteric fracture surgery. Journal of Air Force Medical University 2024;45(11):1262-1266.
45. Kalron A, Tawil H, Peleg-Shani S, Vatine JJ. Effect of telerehabilitation on mobility in people after hip surgery: a pilot feasibility study. Int J Rehabil Res 2018;41(3):244-250. [doi: 10.1097/MRR.0000000000000296] [Medline: 29794545]
46. Howell DF, Malmgren Fänge A, Rogmark C, Ekvall Hansson E. Rehabilitation outcomes following hip fracture of home-based exercise interventions using a wearable device-A randomized controlled pilot and feasibility study. Int J Environ Res Public Health 2023;20(4):3107 [FREE Full text] [doi: 10.3390/ijerph20043107] [Medline: 36833801]
47. Prieto-Moreno R, Mora-Traverso M, Estévez-López F, Molina-Garcia P, Ortiz-Piña M, Salazar-Graván S, et al. Effects of the activeHip+ mHealth intervention on the recovery of older adults with hip fracture and their family caregivers: a multicentre open-label randomised controlled trial. EClinicalMedicine 2024;73:102677 [FREE Full text] [doi: 10.1016/j.eclinm.2024.102677] [Medline: 38911836]

XSL•FO
RenderX

48. Dajpratham P, Komas J, Yamthed R, Chanthon P, Kovintaset K, Claikhem T, et al. Feasibility and efficacy of real-time teleresistance exercise programs for physical function in elderly patients after hip fracture surgery: a randomized controlled trial. BMC Geriatr 2025;25(1):647 [FREE Full text] [doi: 10.1186/s12877-025-06230-y] [Medline: 40836222]
49. Langford D, Fleig L, Brown KC, Cho NJ, Frost M, Ledoyen M, et al. Back to the future - feasibility of recruitment and retention to patient education and telephone follow-up after hip fracture: a pilot randomized controlled trial. Patient Prefer Adherence 2015;9:1343-1351 [FREE Full text] [doi: 10.2147/PPA.S86922] [Medline: 26491262]
50. Suso-Martí L. Effectiveness of telerehabilitation in physical therapist practice: an umbrella and mapping review with meta-meta-analysis. Phys Ther 2021;101(5):pzab075. [doi: 10.1093/ptj/pzab075]
51. Pastora-Bernal JM, Martín-Valero R, Barón-López FJ, Estebanez-Pérez MJ. Evidence of benefit of telerehabitation after orthopedic surgery: systematic review. J Med Internet Res 2017;19(4):e142 [FREE Full text] [doi: 10.2196/jmir.6836] [Medline: 28455277]
52. Freitag MH, Magaziner J. Post-operative considerations in hip fracture management. Curr Rheumatol Rep 2006;8(1):55-62. [doi: 10.1007/s11926-006-0027-1] [Medline: 16515767]
53. Wu W, Zhang Y, Zhang Y, Peng B, Xu W. Clinical effectiveness of home-based telerehabilitation program for geriatric hip fracture following total hip replacement. Orthop Surg 2023;15(2):423-431 [FREE Full text] [doi: 10.1111/os.13521] [Medline: 36448261]
54. Vincent C, Wegier P, Chien V, Kurahashi AM, Ginsburg S, Molla Ghanbari H, et al. Qualitative evaluation of a Novel Educational tool to communicate individualized hip fracture prognostic information to patients and surrogates: my hip fracture (My-HF). Geriatr Orthop Surg Rehabil 2021;12:21514593211050513 [FREE Full text] [doi: 10.1177/21514593211050513] [Medline: 34707914]
55. Pak SS, Janela D, Freitas N, Costa F, Moulder R, Molinos M, et al. Comparing digital to conventional physical therapy for chronic shoulder pain: randomized controlled trial. J Med Internet Res 2023;25:e49236 [FREE Full text] [doi: 10.2196/49236] [Medline: 37490337]
56. Tore NG, Oskay D, Haznedaroglu S. The quality of physiotherapy and rehabilitation program and the effect of telerehabilitation on patients with knee osteoarthritis. Clin Rheumatol 2023;42(3):903-915 [FREE Full text] [doi: 10.1007/s10067-022-06417-3] [Medline: 36279075]
57. Brown MC, Lin CQ, Jin C, Rohde M, Rocos B, Belding J, et al. Effect of a digital health exercise program on the intention for spinal surgery in adult spinal deformity: exploratory cross-sectional survey. JMIR Form Res 2025;9:e66889 [FREE Full text] [doi: 10.2196/66889] [Medline: 40300159]
58. Salehian F, Mahmoudzadeh-Sagheb Z, Yoosefinejad AK, Zakerabasali S. A home-based tele-rehabilitation exercise system for patients after knee replacement surgery. BMC Musculoskelet Disord 2024;25(1):605 [FREE Full text] [doi: 10.1186/s12891-024-07731-4] [Medline: 39085850]
59. Chang X, Xu S, Zhang H. Regulation of bone health through physical exercise: mechanisms and types. Front Endocrinol (Lausanne) 2022;13:1029475 [FREE Full text] [doi: 10.3389/fendo.2022.1029475] [Medline: 36568096]
60. Troy K. Exercise Early and Often: Effects of Physical Activity and Exercise on Women's Bone Health. Int J Environ Res Public Health 2018;15(5).
61. Park C, Lee B. A Systematic review of the effects of interactive telerehabilitation with remote monitoring and guidance on balance and gait performance in older adults and individuals with neurological conditions. Bioengineering (Basel) 2024;11(5):460 [FREE Full text] [doi: 10.3390/bioengineering11050460] [Medline: 38790328]
62. He Z, Wu H, Zhao G, Zhang Y, Li C, Xing Y, et al. The effectiveness of digital technology-based Otago Exercise Program on balance ability, muscle strength and fall efficacy in the elderly: a systematic review and meta-analysis. BMC Public Health 2025;25(1):71 [FREE Full text] [doi: 10.1186/s12889-024-21251-9] [Medline: 39773392]
63. Gaspar AGM, Lapão LV. eHealth for addressing balance disorders in the elderly: systematic review. J Med Internet Res 2021;23(4):e22215 [FREE Full text] [doi: 10.2196/22215] [Medline: 33908890]
64. Alizadehsaravi L, Koster RA, Muijres W, Maas H, Bruijn SM, van Dieën JH. The underlying mechanisms of improved balance after one and ten sessions of balance training in older adults. Hum Mov Sci 2022;81:102910 [FREE Full text] [doi: 10.1016/j.humov.2021.102910] [Medline: 34864610]
65. McCrum C, Bhatt TS, Gerards MHG, Karamanidis K, Rogers MW, Lord SR, et al. Perturbation-based balance training: principles, mechanisms and implementation in clinical practice. Front Sports Act Living 2022;4:1015394 [FREE Full text] [doi: 10.3389/fspor.2022.1015394] [Medline: 36275443]
66. Pai Y, Yang F, Bhatt T, Wang E. Learning from laboratory-induced falling: long-term motor retention among older adults. Age (Dordr) 2014;36(3):9640 [FREE Full text] [doi: 10.1007/s11357-014-9640-5] [Medline: 24668268]
67. Bhatt T, Yang F, Pai Y. Learning to resist gait-slip falls: long-term retention in community-dwelling older adults. Arch Phys Med Rehabil 2012;93(4):557-564. [doi: 10.1016/j.apmr.2011.10.027]
68. McCrum C, Vaes AW, Delbressine JM, Koopman M, Liu W, Willems P, et al. A pilot study on the feasibility and effectiveness of treadmill-based perturbations for assessing and improving walking stability in chronic obstructive pulmonary disease. Clin Biomech (Bristol) 2022;91:105538 [FREE Full text] [doi: 10.1016/j.clinbiomech.2021.105538] [Medline: 34823220]

XSL•FO
RenderX

69. Liu H, Yin H, Yi Y, Liu C, Li C. Effects of different rehabilitation training on balance function in stroke patients: a systematic review and network meta-analysis. Arch Med Sci 2023;19(6):1671-1683 [FREE Full text] [doi: 10.5114/aoms/167385] [Medline: 38058731]
70. Molina-Garcia P, Mora-Traverso M, Prieto-Moreno R, Díaz-Vásquez A, Antony B, Ariza-Vega P. Effectiveness and cost-effectiveness of telerehabilitation for musculoskeletal disorders: a systematic review and meta-analysis. Ann Phys Rehabil Med 2024;67(1):101791 [FREE Full text] [doi: 10.1016/j.rehab.2023.101791] [Medline: 38128150]
71. Peeters CM, Visser E, Van de Ree CL, Gosens T, Den Oudsten BL, De Vries J. Quality of life after hip fracture in the elderly: a systematic literature review. Injury 2016;47(7):1369-1382. [doi: 10.1016/j.injury.2016.04.018]
72. Soukkio PK, Suikkanen SA, Sintonen H, Kukkonen-Harjula KT, Kautiainen H, Kääriä SM, et al. Health-related quality of life after hip fracture: effects of a 12-month home-based exercise intervention-secondary analyses of an RCT. Qual Life Res 2024;33(2):541-550. [doi: 10.1007/s11136-023-03545-4] [Medline: 37932555]
73. Gilboa Y, Maeir T, Weber A, Maeir A, Rotenberg S. Predictors of community reintegration and quality of life after hip fracture among community-dwelling older adults. Int J Rehabil Res 2019;42(3):234-239. [doi: 10.1097/MRR.0000000000000355] [Medline: 31394559]
74. Xiao H, Zeng W, Lu L, Yuan J, Yan Z, Wang J. Effectiveness of telerehabilitation in postoperative outcomes in patients on hip fracture surgery: a meta-analysis of randomized controlled trials. BMC Sports Sci Med Rehabil 2025;17(1):130 [FREE Full text] [doi: 10.1186/s13102-025-01169-4] [Medline: 40420240]
75. Marmor MT, Grimm B, Hanflik AM, Richter PH, Sivananthan S, Yarboro SR, et al. Use of wearable technology to measure activity in orthopaedic trauma patients: a systematic review. Indian J Orthop 2022;56(7):1112-1122 [FREE Full text] [doi: 10.1007/s43465-022-00629-0] [Medline: 35813536]
76. Choi S, Lee J, Bae KJ, Kim DK, Lee YH, Lee Y. Clinical effect of rehabilitation after distal radius fracture surgery using a wearable device: a comparative prospective cohort study. Hand Surg Rehabil 2024;43(6):101779. [doi: 10.1016/j.hansur.2024.101779] [Medline: 39349296]
77. Nuzzo JL, Pinto MD, Kirk BJC, Nosaka K. Resistance exercise minimal dose strategies for increasing muscle strength in the general population: an overview. Sports Med 2024;54(5):1139-1162 [FREE Full text] [doi: 10.1007/s40279-024-02009-0] [Medline: 38509414]
78. Spiering B, Mujika I, Sharp MA, Foulis SA. Maintaining physical performance: The minimal dose of exercise needed to preserve endurance and strength over time. J Strength Cond Res 2021;35(5):1449-1458. [doi: 10.1519/JSC.0000000000003964] [Medline: 33629972]

Abbreviations

DHI: digital health intervention
FBG: Functional Balance for Geriatric Patients
FIM: Functional Independence Measure
GRADE: Grading of Recommendations, Assessment, Development and Evaluation
HHS: Harris Hip Score
HKSJ: Hartung-Knapp-Sidik-Jonkman
HOOS: Hip Disability and Osteoarthritis Outcome Score
MFAC: Modified Functional Ambulatory Category
mHealth: mobile health
MOS: Medical Outcomes Study
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
QoL: quality of life
RCT: randomized controlled trial
SF-12: Short Form-12 Health Survey
SF-36: MOS 36-Item Short Form Health Survey
SMD: standardized mean difference
SPPB: Short Physical Performance Battery
TUG: Timed Up and Go Test
WHO: World Health Organization

XSL•FO
RenderX

Edited by S Brini; submitted 01.Jul.2025; peer-reviewed by F Ye, J Zhang; comments to author 30.Oct.2025; accepted 05.Jan.2026; published 12.Mar.2026.

Please cite as:
Fan W, Zhang Q, Lu Q
Effects of Digital Health Interventions on Functional and Psychological Outcomes in Older Patients With Hip Fractures: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e79563
URL: https://www.jmir.org/2026/1/e79563
doi:10.2196/79563
PMID:

©Wei Fan, Qi Zhang, Qunfeng Lu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Use of Health and Welfare Technology in Palliative Care: State-of-the-Art Review

Viktoria Zander[1*]; Maja Holm[2,3*]; Monir Mazaheri[2,4*]; Christine Gustafsson[1,2*]; Sara Landerdahl Stridsberg[5*]; Ragnhild Hedman[2,6*]

[1]School of Health, Care and Social Welfare, Mälardalen University, Eskilstuna, Sweden
[2]Department of Nursing Science, Sophiahemmet University College, Stockholm, Stockholm, Sweden
[3]Department of Health Care Sciences, Palliative Research Centre, Marie Cederschiöld University, Stockholm, Sweden
[4]Department of Neurobiology Care Sciences and Society, Nursing Division, Karolinska Institutet, Stockholm, Sweden
[5]University Library, Mälardalen University, Västerås, Sweden
[6]Stockholm Gerontology Research Center, Stockholm, Sweden
[*]all authors contributed equally

Corresponding Author:
Viktoria Zander
School of Health, Care and Social Welfare
Mälardalen University
SE-Box 325
Eskilstuna, 631 05
Sweden
Phone: 46 0736620515
Email: viktoria.zander@mdu.se

Abstract

Background: As more individuals live longer with complex conditions, the need for effective palliative care (PC) grows. It has been stated that access to PC should be integrated early and delivered in a timely manner to patients with life-threatening illnesses. Health and welfare technologies (HWTs) offer tools to enhance care delivery, particularly in home and rural settings. Although there is a profound lack of evidence regarding the impact when used in PC, it is necessary to critically assess the current state of knowledge regarding impacts and consequences of technologies, ensuring that their integration considers broader implications for patients, caregivers, and health care systems in PC.

Objective: This review explores health and welfare technology used in PC, aiming to inform practice and improve care quality.

Methods: This state-of-the-art review included empirical studies describing the use of HWT in PC for adult patients. We used a thematic synthesis approach to compare studies and provide a synthesis of the key points.

Results: Based on the inclusion criteria, 94 studies were included. PC is both a clinical specialty and an overall approach to care that focuses on improving quality of life and relieving suffering for patients and families facing serious illness, based on needs and not prognosis. HWT shows potential to increase access and continuity of care, for symptom management to support patients to remain at home and prevent frequent emergency visits. It can have the potential to build and remain relationships between patients, their families, and the health care team, as well as for interprofessional collaboration and support. However, there are challenges to overcome that might affect the quality of care when using technology.

Conclusions: HWT shows potential as a complement to usual PC. Our findings point toward the importance of caution in choosing when to use HWT in PC, and for which patients.

(J Med Internet Res 2026;28:e79637) doi:10.2196/79637

KEYWORDS

digital technology; health and welfare technology; palliative care; state-of-the-art review; systematic review

XSL•FO
RenderX

Introduction

As more individuals live longer with complex conditions, the need for effective palliative care (PC) grows. It has been stated that access to PC should be integrated early and delivered in a timely manner to patients with life-threatening illnesses. However, research has demonstrated that health care systems fail in this regard. Home-based PC is gaining attention, reflecting patient preferences to remain at home and potential cost savings from reduced hospitalizations. Health and welfare technologies (HWTs) offer tools to enhance care delivery, particularly in home and rural settings. However, questions about their feasibility and acceptability remain.

Today, an increasing number of individuals are living to older age with complex conditions and a great need for PC, including symptom relief and family support [1,2]. However, reports show that less than 15 % of patients requiring PC receive it in a timely and accurate manner [3]. PC is a holistic approach to the care of people with life-limiting conditions and near the end of life. It aims to acknowledge and attend to all aspects of the patients' and family members' needs when they are facing severe illness, including physiological, psychological, existential, and social issues. Symptom management, support of family members, and interprofessional teamwork are emphasized as crucial elements [3]. In recent decades, PC has evolved from primarily being offered to patients with incurable cancer, to currently being recommended to all patients with chronic and life-limiting illness [4]. Nonspecialized PC can be delivered by all care professionals, whereas specialized PC is provided in hospices, hospital PC units, and by specialized PC teams in home care [3]. Further, it has been highlighted that PC should be delivered early and integrated with specific illness treatment [5]. However, it has been demonstrated that access to PC varies and that patients with cancer still receive it more often than patients with other diagnoses [6,7].

Lately, there has been an increased focus on home-based care, especially since many patients wish to be cared for and die in their homes [8,9]. Home-based PC services offer many benefits, such as a sense of normalcy, choice, and comfort [10]. The prospect of dying at home is regarded as a more comfortable and dignified experience than dying in a hospital [10]. However, challenges exist in providing an optimal service. Unmet needs, uncoordinated care, and insufficient communication with health care professionals [11,12], as well as the demanding collaboration between specialists and home care professionals, make this challenging [13]. Moreover, home-based PC relies on the contribution of family caregivers [14], who often find themselves in a situation of managing multiple responsibilities and often forget their own needs to attend to those of the patient [10]. This informal care work often goes unnoticed and unaddressed [10,15]. From a societal point of view, home care has been associated with lower costs, as repeated hospital admissions are a major driver of expenditure in PC [16]. Although, the evidence for this is uncertain [17].

The COVID-19 pandemic has accelerated the demand for technologies in health care [18]. HWT comprises technology-based interventions that aim to maintain or promote health, well-being, quality of life, and/or increase efficiency in the operational delivery of welfare, social, and health care services, while improving working conditions for staff [19]. There are high expectations for HWT as solutions to challenges such as aging populations and limited resources [20-22]. HWT is expected to enhance the delivery and accessibility of PC, particularly in home settings [23] and in rural areas [24]. For instance, videoconferencing systems enable the remote delivery of multispecialty care, and artificial intelligence–driven wearable and nonwearable technologies facilitate remote assessments of patients' conditions in their home environments, leading to more comprehensive clinical evaluations and empowering patients to monitor their own health [25]. Additionally, digital care plans can streamline home care delivery, potentially reducing avoidable hospitalizations [26].

Previous research has sought to gather evidence on the use of HWT in PC with a focus on telehealth (describing the provision of health care remotely by means of a variety of telecommunication tools and video consultations) [27-32], telemedicine (the use of remote technology and telecommunications) [33,34], eHealth (the use of information and communication technology for health care provision) [35], or with a focus on video consultations [36]. There is also specific focus on populations or settings, such as older adults [23], the professional perspective [30], or the patient perspective [29,35], home-based PC [23,32,37], or PC in rural areas [24,27,31,33]. However, there is a lack of literature reviews that have included the overall variety of HWT for all patients in PC and focused on impacts on patients as well as informal and formal caregivers.

Overall, HWT is frequently met with expectations, often based on biases, such as optimistic assumptions about what it could achieve and the belief that a technology will be as good as or better than a human, regardless of the task [38]. While this optimism emphasizes the potential for HWT, it often overlooks the complexities and challenges of implementing technology in sensitive areas such as PC [39]. There is also a profound lack of evidence of benefits and harms, as well as of impacts, when HWT is used in PC [40]. Other than telehealth, videoconferencing, or after-hours telephone support, there is little evidence for HWT used in PC [37]. Moreover, it is still unclear whether PC delivered remotely or with support of HWT is equivalent to more resource-intensive methods of in-person care [41]. Therefore, it is necessary to critically assess the current state of knowledge regarding the use of technologies for patients, caregivers, and health care systems in PC. This review explores HWT used in PC, aiming to inform practice and improve care quality.

This systematic review responds to the following research questions: (1) Which HWT is used in PC? (2) What impact does the use of HWT in PC have on patients and informal and formal caregivers? (3) What knowledge gaps and research needs are identified related to the use of HWT in PC? (The last research question is reported elsewhere).

XSL•FO
RenderX

Methods

Study Design

We conducted a state-of-the-art review. This type of review is appropriate to shape a comprehensive understanding of the current state of knowledge in a specific area [42]. The review followed PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [43]. The PRISMA-ScR checklist for this review is presented in Multimedia Appendix 1. A protocol was prospectively registered in the Open Science Framework on October 10, 2024 [44].

Eligibility Criteria

Eligibility criteria were formulated in dialogue within the research group, based on the aim of the review. In line with Radbruch et al [4], PC was defined as holistic care of patients with severe illness, especially near the end of life, regardless of diagnosis and care place. Studies involving patients receiving curative treatment, or with the possibility of receiving curative treatment such as, kidney transplantation, were excluded.

Studies focusing on technology-based interventions for safety, activity, participation, and independence and/or increased efficiency in PC, and working conditions for professional caregivers were included. Further inclusion criteria were empirical studies (qualitative, quantitative, mixed methods, and case studies) describing the use of HWT in PC of adult patients (aged ≥18 years). The search was limited to studies published after 2012.

Studies that used technologies only for data collection were excluded, as were studies not reporting technology use in the intended setting. Study protocols, reviews, and studies describing the development of technologies without involvement of patients, informal caregivers, or health care professionals were also excluded, as were studies in languages other than English and studies that could not be obtained in full text.

Information Sources

The following electronic databases were searched: PubMed, APA PsycINFO, Cochrane Library, CINAHL Plus, Scopus, and Web of Science Core Collection.

Searches were conducted on October 27, 2022. Additional update searches were completed on November 23, 2023.

Search Strategy

Searches were conducted by an academic librarian (SLS). The search terms were organized to match the review's goals, covering concepts such as PC, digital health and welfare technology, and home monitoring. Free-text words were searched in article titles and abstracts, along with database-specific subject headings, like MeSH (Medical Subject Headings) terms, when relevant. The search targeted articles published after 2012, excluding topics related to neonatal and pediatric care using the Boolean operator NOT to avoid content on children's PC. Textbox 1 contains an example of search terms. The complete search strategies for all databases are available in Multimedia Appendix 2.

Textbox 1. Search strategy PubMed.

PubMed search strategy:

- "Palliative Care"[MeSH Terms] OR "Palliative Medicine"[MeSH Terms] OR "Hospice and Palliative Care Nursing"[MeSH Terms] OR "Terminally ill"[MeSH Terms] OR "Terminal Care"[MeSH Major Topic] OR "Hospice Care"[MeSH Terms]
- "palliati*"[Title/Abstract] OR "hospice care"[Title/Abstract] OR "hospice nursing"[Title/Abstract] OR "terminal care"[Title/Abstract] OR "supportive care"[Title/Abstract] OR "terminal stage"[Title/Abstract] OR "terminal disease"[Title/Abstract] OR "terminally ill"[Title/Abstract] OR "end stage"[Title/Abstract] OR "end of life"[Title/Abstract] OR "hospice program*"[Title/Abstract] OR "advanced illness"[Title/Abstract]
- "eHealth"[Title/Abstract] OR "e-health"[Title/Abstract] OR "telemedicine"[Title/Abstract] OR "telehealth"[Title/Abstract] OR "mhealth"[Title/Abstract] OR "m-health"[Title/Abstract] OR "mobile health"[Title/Abstract] OR "e-homecare"[Title/Abstract] OR "digital health"[Title/Abstract]
- "Telemedicine"[MeSH Terms]
- "welfare technolog*"[Title/Abstract] OR "ambient assisted living*"[Title/Abstract] OR "ambient intelligence*"[Title/Abstract]
- "Ambient Intelligence"[MeSH Terms]
- "home monitoring"[Title/Abstract] OR "distance monitoring"[Title/Abstract] OR "distance care"[Title/Abstract] OR "distance nursing"[Title/Abstract] OR "distance medicine"[Title/Abstract] OR "environmental control*"[Title/Abstract] OR "remote sensing"[Title/Abstract] OR "Distance Counseling"[Title/Abstract] OR "Internet-Based Intervention"[Title/Abstract] OR "ambulatory monitoring"[Title/Abstract] OR "remote consultation"[Title/Abstract] OR "telecommunication*"[Title/Abstract] OR "E-Counseling"[Title/Abstract] OR "e therapy"[Title/Abstract] OR "distance spanning"[Title/Abstract] OR "health informatics"[Title/Abstract] OR "health information technology"[Title/Abstract] OR "medical information science"[Title/Abstract]
- "Remote Sensing Technology"[MeSH Terms] OR "Internet-Based Intervention"[MeSH Terms] OR "monitoring, ambulatory"[MeSH Terms] OR "Remote Consultation"[MeSH Terms] OR "Telecommunications"[MeSH Terms]

Selection of Sources of Evidence

After the search, references were uploaded to Covidence, a web-based collaboration software for literature reviews [45]. Following the automatic removal of duplicates by Covidence and the manual removal of duplicates by the reviewers, a total of 3662 studies were examined for eligibility. Five of the authors (CG, VZ, MH, MM, and RH) conducted the abstract and full

XSL•FO
RenderX

text screening. To start with, 40 abstracts were screened by 2 or 3 of the authors in various constellations, and their decisions to include or exclude were compared and discussed to ensure consistency. Subsequently, the abstract screening was equally divided among the authors. After the title and abstract screening, the remaining 601 articles were read in full and assessed for eligibility. A subset of 10 articles was initially read by all 5 authors separately and then compared. The remaining studies were divided among the authors and discussed when needed. All articles that matched the inclusion criteria were included in the review. In accordance with established practice for state-of-the-art reviews, no formal quality assessment of the included studies was undertaken [42].

The screening process is described in a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. HWT: health and welfare technology; PC: palliative care.

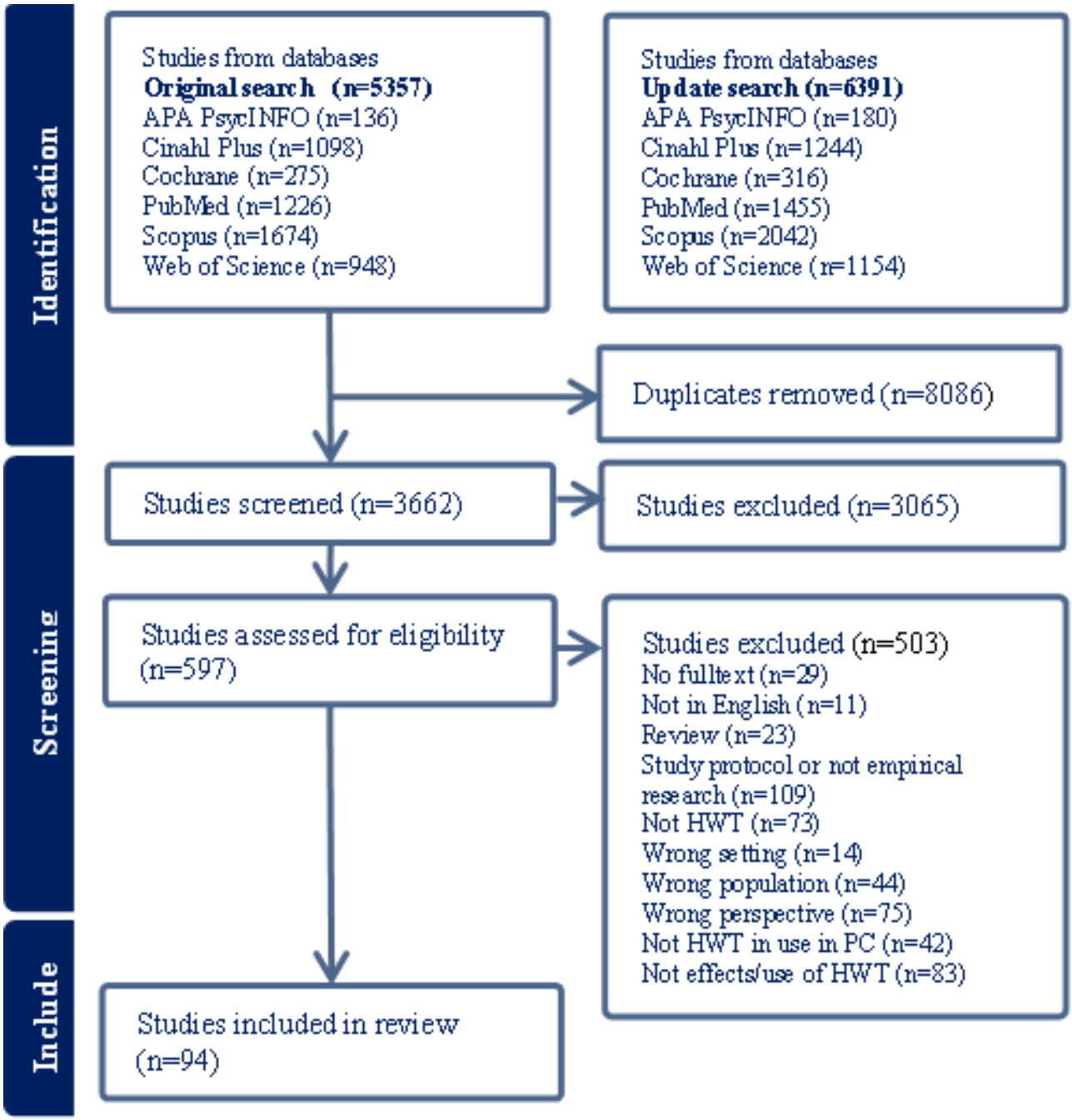

Data Charting Process

After reviewing eligible studies, a data extraction template was generated to guide the data extraction process, including citation, study location, study aim and design, sample characteristics, intended target group, PC description, characteristic of technology, impacts of technology use in PC, knowledge gaps and research needs, and conclusions (knowledge gaps and research needs have been reported elsewhere).

XSL•FO
RenderX

Synthesis of Results

Study characteristics were entered into an Excel (Microsoft Corp) spreadsheet. A thematic synthesis approach, led by VZ, supported by RH and MH, and discussed within the author group, was used to compare the studies and provide a synthesis of the key points. The process of the synthesis was guided by the first 2 steps described by Thomas and Harden [46]. These steps were (1) coding of the text—identification of themes across the included studies. Each study's findings were identified and put into a metric. The findings were then coded line by line. Each sentence was read inductively to identify underpinning themes or concepts, which were labeled with a code; (2) developing descriptive themes—the descriptive codes were compared and organized into descriptive themes. Descriptive themes across the different categories of technology were compared and organized into one overall synthesis of the use of HWT in PC.

Results

Description of Included Studies

The selection process resulted in the inclusion of 94 articles describing the use of HWT in PC [47-140]. The included studies were published between 2012 and 2023. They were conducted in North America (Canada and United States; n=44), South America (Brazil and Chile; n=3), Australia (n=14), Africa (Kenya, sub-Saharan Africa, and Tanzania; n=4), Asia (China, India, Indonesia, and Taiwan; n=5), and Europe (Austria, Belgium, Finland, Georgia, Germany, Netherlands, Norway, Italy, Sweden, and United Kingdom; n=26). Some were cross-national. The review included studies using quantitative (n=50), qualitative (n=19), as well as mixed methods (n=25). The sample size of the studies ranged from 1 to 3178. The most common target group was patients, who were involved in 64 studies, followed by informal caregivers (involved in 31 studies) and health care professionals (involved in 21 studies). Several studies involved more than one target group. The most common diagnoses were cancer (n=38), followed by a few articles focusing on a variety of illnesses such as heart failure (n=2), chronic obstructive pulmonary disease (COPD, n=1), cirrhosis (n=1), dementia (n=1), and cognitive impairments in amyotrophic lateral sclerosis (n=1). Besides diagnoses, articles focused on older adults suffering from chronic illnesses (n=3), terminal illness and the end-of-life phase (n=12), and palliative care regardless of diagnosis (n=24). The characteristics of the studies included are presented in Multimedia Appendix 3.

Health and Welfare Technology in Palliative Care

The HWT in the included studies were categorized based on function (Table 1). In some of the studies, HWT from more than 1 category were included, such as video consultations based on symptom monitoring.

Table 1. Categorization of health and welfare technology (HWT) in included studies.

Category of HWT	Examples of type of HWT
Symptom monitoring (n=25)	• Systems to alert clinicians of changes in patient's symptoms • Symptom reporting systems and applications • Systems for wireless data transfer • Wearable monitoring devices
Telehealth consultations and conferences (n=51)	• Videoconferencing • Teleconsultations • Telehealth-delivered care
Sharing of patient information (n=2)	• Electronic care coordination systems • Needs rounds via telehealth
Remote therapy and treatment interventions (n=7)	• VR[a] headset • Therapeutic interventions in mobile apps • Web-based therapeutic interventions • Therapy delivered via email
Education and support (n=27)	• Mobile apps and web-based education platforms for patients, caregivers, or health care personnel • Teleconferencing technology to support and train health care personnel remotely • Virtual webinar sessions • Facebook support groups

[a]VR: virtual reality.

XSL•FO
RenderX

The Use of Health and Welfare Technology in Palliative Care

Overview

The included studies described the use of HWT in PC on an individual level (including patients, informal caregivers, and health care professionals) and on an organizational level. The synthesis resulted in 7 themes, which are presented in the following subsections (Table 2 provides an overview).

Table 2. Descriptive themes and codes.

Descriptive themes	Codes
Symptom control and disease progression (n=26)	• Impact on disease progression • Impact on physical and mental symptoms • Management of symptoms • Disease-related stress
Quality of life and death (n=10)	• Patients' and caregivers' quality of life • Support in the end-of-life phase and death
Competence and palliative literacy (n=9)	• Knowledge and understanding • Self-efficacy • Insight into norms and values
Palliative care provision (n=44)	• Care satisfaction • Medication use and compliance with therapies • Prevention of excessive care • Care coordination • Accelerating advanced care planning • Access to care
Multidimensional care relationships (n=18)	• Partnership between patients and the health care team • Companionship with and among significant others • Interprofessional collaboration
Facilitating a comprehensive support system (n=33)	• Professionals' understanding • Patient empowerment • Support for the caring role
Organizational outcome (n=8)	• Cost-effectiveness • Time and travel • Job satisfaction and distress

Symptom Control and Disease Progression

This theme focuses on the potential of HWT to support and prevent disease progression, physical and mental symptoms, and to relieve disease-related stress [47-72]. Types of HWT involved in symptom control and disease progression were technology for symptom monitoring [47,50,52-55,60,64,65,69,71], telehealth consultations and conferences [47,48,52,55-57,60,61,63,64,66-68,70], remote therapy and treatment interventions [58,59,62,72], and technology for education and support [49,51,56].

Among the included studies, the potential of HWT to support and prevent disease progression was mainly reported and discussed by 2 studies. Kazankov et al [47] described how a digital health solution with monitoring of heart rate, blood pressure, weight, body water, and cognitive function, for individuals with cirrhosis could facilitate timely intervention to prevent disease progression. Jiang et al [48] described how a model with a communication platform resulted in less functional decline. A total of 3 studies explored the effects of HWT on mortality, none of which could show any significant effects [49-51].

The impact of HWT on physical and mental symptoms was described by several studies. Positive impact on physical symptoms related to cancer illness and treatment was described by Tumeh et al [52], Mooney et al [53], Besse et al [54], and Cornetta et al [55]. Ngoma et al [50] reported no differences in overall symptoms when using remote symptom assessment and care coordination, although there were effects on symptom severity. Mark et al [56] reported a positive impact on dyspnea with an online teaching intervention for persons with COPD. Rafter [57] showed that an eHealth system allowed hospice professionals to deliver more effective care, with improvement in pressure ulcers for end-of-life patients. Impact on mental symptoms, such as anxiety and depression, was reported in 7 of the included studies. Positive effects were reported in cancer care studies using symptom monitoring [52], a virtual reality headset [58], a mindfulness intervention [59], telemedicine [60], videoconferencing [61], and web-based psychotherapy for informal caregivers [62]. On the other hand, one study using teleconsultations showed no differences in depression scores and increased anxiety scores [63]. Eight of the included studies showed the potential of HWT to contribute support in symptom management [64-71].

XSL•FO
RenderX

HWT was also used to relieve disease-related stress. For example, studies showed that an online support system in the care of patients with cancer has the potential to reduce symptom distress [49], mindfulness via video could reduce cancer-related stress [59], telehealth music therapy had positive affective effects [72], and teleconsultations resulted in lower distress symptom scores [68]. Maguire et al [69] showed that using a remote symptom monitoring system provided reassurance about symptom experience and the feeling of being listened to. However, Hoek et al [63], on the other hand, reported that adding weekly teleconsultations to usual PC led to a higher total distress score among home-dwelling patients with advanced cancer.

Quality of Life and Death

This theme describes patients' and caregivers' quality of life [52,56,60,69,73,74] and support in end-of-life phase and death [75-78]. It involves HWT for symptom monitoring [52,60,69], telehealth consultations and conferences [52,56,60,74,75,77,78], sharing of patient information [76], remote therapy and treatment interventions [73], and education and support [56,74].

Five of the included studies showed some effects on patients' and informal caregivers' quality of life with the use of HWT for symptom monitoring [52,60], training delivered over Skype for patients with COPD [56], a supportive care mobile app intervention [74], and a cognitive-behavioral therapy mobile app for anxiety [73]. Maguire et al [69] found that a system with daily symptom reports provided reassurance, but it did not lead to changes in quality of life.

Outcomes regarding support in the end-of-life phase and death were reported in 4 of the included studies. The potential to support dying at home or in another community setting was reported using an electronic PC coordination system [76], home care delivery with point-of-care technology and remotely located health care professionals [78], and a PC after-hours telephone number [75]. Johnston et al [77] reported the potential of Skype calls with dying persons and family members to bring closure and reconciliation, inclusion in the dying process, and healthy grieving.

Competence and Palliative Literacy

Competence and palliative literacy refer to knowledge and understanding, self-efficacy, and insight into norms and values [74,75,79-85]. This theme includes HWT for symptom monitoring [80,81], telehealth consultations and conferences [74,83], remote therapy and treatment interventions [85], and education and support [74,75,79,82-84].

Different HWT interventions have been shown to increase knowledge and understanding of health, diagnosis, symptoms, and concerns among patients [74,79], informal caregivers [80,81], and health care professionals [82,83]. An e-learning intervention for nurses to increase knowledge of PC and attitudes toward dying patients and death showed positive effects [82]. Teleconferencing technology to support and train hospice nurses remotely was perceived to improve knowledge as well as self-efficacy in caring of patients in PC [83]. The intervention had given them access to education that otherwise would have been difficult to obtain due to geography. Self-efficacy to manage their own condition was reported by patients with advanced cancer using a digital support app [75]. Digital information and education were also shown useful for addressing insights into one's own thoughts and values related to advanced dementia among patients [84] and related to caring for a person with advanced cancer among informal caregivers, including accepting experiences of negative thoughts and feelings and being more aware of personal values [85].

Palliative Care Provision

PC provision includes subthemes such as care satisfaction [50,87], medication use and compliance with therapies [72,73,81,82,87,88], prevention of excessive care [48,51,64,65,68,69,76,86,89-98], care coordination [75,83,94,102-110], accelerating advanced care planning [51,84,87,93,94,97,99-101], and access to care [68,74,96,106,108,111-117]. This was mainly related to use of technology for telehealth consultations and conferences [48,64,68,74,75,83,84,87,88,91-96,98,100,102-104,106,108,109,111-116], but also technology for symptom monitoring [50,64,65,69,81,89,96,98,106,107,110,117], education and support [51,74,82-84,86,90,97,99,101,117], sharing of patient information [76,100], and remote therapy and treatment interventions [72,105].

The use of HWT was shown to be promising in the provision of care, including care satisfaction [50,87] and medication use and compliance with therapies [73,82,87,88]. Zeiser et al [72] reported that music therapy telehealth services increased compliance with other therapies. Other studies reported decreased polypharmacy and potentially decreased adverse drug events [88], but no effects on analgesic adherence [82] and no differences in anticipatory medication prescribing [87].

Some studies reported positive impacts on care coordination [102], care provision [83], increased efficiency [103-108], increased care resource use [75], and positive outcomes related to quality of care [94,106]. For example, Groothuizen et al [109] reported that virtual team meetings resulted in increased flexibility, reduced travel time, and easier real-time access to patient information for health care professionals. Health care professionals in the study by Oelschlägel et al [110], on the other hand, found that organizational challenges made it difficult to obtain and share the information necessary to provide seamless and optimal service to patients.

Prevention of excessive care is a core purpose of technology in PC and relates to the use of hospice care, emergency care, and hospitalization. Among the included studies, no effects were shown on hospice enrollment or length of stay [51], nor on time until entry into hospice [98]. Emergency department visits and admission could be avoided or decreased in several studies [65,69,76,86,89,90,92]. The HWT used in these studies varied. Technology to support patients (symptom monitoring and means to communicate needs to health care professionals) showed positive effects [65,69,89,92], as did technology used to support health care professionals in their work [76,86,90]. Manz et al [51] showed that opt-out text messages to prompt serious illness conversations decreased end-of-life systemic therapy relative to controls, but there was no effect on hospice enrollment or length of stay, inpatient death, or end-of-life intensive care unit

XSL•FO
RenderX

use. Nor did a teleconsultation service with a triage system change the number of emergency visits [91]. HWT to support patients in communication with health care professionals [48,91,94,96] showed potential to decrease the need for hospital referrals. Unnecessary hospital transfers could also be prevented by improved collaboration among physicians using telemedicine [95]. Moreover, an advance care planning video program intervention increased documented "Do Not Hospitalize" orders among nursing home residents with advanced illness, but did not significantly reduce hospitalizations [97].

Technology was also used to accelerate advance care planning, with promising results in 7 studies [51,84,94,97,99-101], but no results in 2 studies [87,93].

Other positive effects reported were increased access to care [69,96,106,108,111-114] and increased continuity of care [113,115,116]. HWT was shown to enable health care professionals to respond quickly to patients' care needs [96,106], allowing them to reach more patients [68,111], especially during epidemics or when out of town. For patients and informal caregivers, the technology increased the sense of security related to the ability to contact the clinic when needed [96,117].

Multidimensional Care Relationships

Multidimensional care relationships concern the ability to use HWT to form partnership between patients and health care teams, companionship with significant others, and interprofessional collaboration [65,69,77,96,104,110,114,116,118-127]. HWT used in the studies related to this theme included technology for symptom monitoring [65,69,96,110,119,120], telehealth consultations and conferences [77,96,104,110,114,116,118,119,122,127], and education and support [121,123-126].

Whether remote care can affect the partnership between patients and health care teams varies between studies and within studies. Rosa et al [118] described positive as well as negative impacts on the quality of relationships with patients, families, and between health care professionals. For health care professionals, increased availability to patients and their caregivers was positive, but there was a perceived loss of nonverbal cues, which made communication more difficult. Some health care professionals also experienced difficulties being supportive remotely during difficult times. It also put a strain on relationships among coworkers. Other studies reported no negative effects on relationships [114], enhanced connectivity with the care team [70,96], and contributions of unique insight into the daily lives of patients [116]. Over time, care delivered with support of HWT, such as telehealth consultations, can result in trustful relationships [116]. The introduction of technologies has the potential to alter the dynamic of relationships between patients, families, and community PC clinicians, serving as a means to complement in-person care [119]. Among the included studies, there were also reports of positive impacts on patient–caregiver communication regarding symptom management [66] and perceptions that it was easier to discuss psychological and care needs remotely [120].

Moreover, HWT can be used to facilitate companionship with and among significant others and informal caregivers [67,78,104,121-125]. The use of technology, such as video calls, was shown to be able to connect patients and family members at the end of life [78]. Another area of use is online support groups for informal caregivers, with the aim of forming companionship between peers [121].

Other studies have shown the ability of HWT to increase professional collaboration [126,127], offering an understanding of each other's professions [126].

Facilitating a Comprehensive Support System

This theme focuses on the ability to use HWT to support professionals' understanding, patient empowerment, and support for the caring role among informal caregivers. Telehealth consultation and conference technology [56,57,67,91,96,102,103,110-112,114,117,118,127,129,133] and technology for education and support [56,86,125,126,128,130-135,137-139] were the most common HWT, followed by technology for symptom monitoring [96,110,117,138,139] and remote therapy and treatment interventions [62].

Among the included studies, HWT was used to support professionals' understanding. Oelschlägel et al [110] described how remote home care helped municipal health care professionals shift their perspective toward patients' priorities. In the study by Rosa et al [118], remote care used as an alternative to in-person care during the COVID-19 pandemic led to changes in perceived self-efficacy among health care professionals in managing their job responsibilities. HWT, such as videoconferences, can also be used by senior professionals to support and mentor the junior workforce [86].

According to the included studies, HWT can play a role in empowering patients, for example by increasing involvement and enabling patients to take an active part in their care [96,117], feel in control of treatment [117], and manage everyday life [110]. HWT was useful in preparing patients before care visits [128], providing support [91,102,112,129], facilitating comfort, safety, and independence for patients [103,114], and increasing patient satisfaction [68].

Technologies were also used to support informal caregivers. Examples of technologies addressing their needs included interventions for mental and emotional support [63,130,131]. These were shown to have potential for preparing individuals for the caring role [86] and addressing unmet needs [63]. Digital technology interventions were used to support informal caregivers in managing everyday challenges and stress [57,132-135]. For example, an education intervention delivered through a smartphone app was effective in improving family readiness and quality of life among family members [135]. However, other studies were not able to show this potential. According to Dionne-Odom et al [130] a telehealth intervention to educate and support informal caregivers of patients with heart failure did not provide any significant effects on quality of life, mood, or caregiver burden.

Other technologies, which primarily focus on patient's needs, may also be used to support the caregiver role, with increased caregiver involvement [136], increased connection between caregivers and patients [125], reduced caregiver burden

XSL•FO
RenderX

[64,137], and effects such as less negative mood and emotional distress [138], as well as reduced loneliness [134]. Access to hospice personnel provided by HWT might be perceived as comforting [68]. Among the included studies, one study was not able to show any significant effects on caregiver burden, self-efficacy, or quality of life with the use of an eHealth self-management application for caregivers of patients with incurable cancer [139].

Organizational Outcomes

Organizational outcomes reported were related to cost-effectiveness [61,111], time and travel [96,108,109,114,140], and job satisfaction and distress [62,68,111,118]. The most common HWT related to the theme was telehealth consultations and conferences [61,62,68,96,108,109,111,114,118,140], although one study also involved HWT for symptom monitoring [96].

Two studies reported results related to cost-effectiveness. Compared with traditional in-person service, telehealth services resulted in equivalent costs but greater efficiency by allowing PC to reach more patients [111]. Video consultations with the PC team for rural patients were found feasible and resulted in travel and cost savings for patients [62]. The potential to minimize travel is clear, as reported by several studies. According to the case report by Morgan et al [96], telehealth-supported care was an effective adjunct to routine clinical care, Groothuizen [109] reported that virtual PC team meetings reduced travel time. Others reported that video-based consultations reduced the burden and expense of travel for patients, families, and consultants [108,114], and patient-reported satisfaction with telehealth in oncology was mainly attributed to advantages in travel and time savings [140]. Using telehealth services and eHealth systems increased job satisfaction due to the patient-centered nature of the care service, increased peer support, and increased professional development [111]; having 24/7 access to hospice triage personnel [68]; and effective care and good patient experiences [58]. On the other hand, a multidisciplinary PC team delivering telecare for hospitalized patients with cancer during the COVID-19 pandemic expressed distress related to competing loyalties (such as institutional obligations, ethical obligations to patients, resentment, and distrust of leadership) and feelings of disempowerment (due to guilt in providing subpar support, decisional regret, and loss of identity as a provider) [118].

Discussion

Principal Results

According to the findings, HWT has the potential to facilitate interventions to support symptom management. Different HWT interventions have shown usefulness to increase knowledge and understanding of health, diagnosis, symptoms, and concerns among patients, informal caregivers, and health care professionals. There were a variety among the reports regarding impact on quality of life among patients and caregivers; although, HWT can sometimes be useful to support dying persons and family members to bring closure and reconciliation. Related to PC provision, the use of HWT has shown promising results on care satisfaction and on medication use and compliance with therapies. Technology has also been used to accelerate advance care planning, with promising results and positive effects on the use of emergency and hospital care. One large benefit is the increased access to care using HWT, which means comfort and security for patients and families. HWT can be useful to include and involve significant others. On the other hand, research has shown both positive and negative impacts on the quality of partnership between patients, families, and health care teams when using HWT, suggesting situations when physical meetings are preferred and others when online meetings are suitable to build and maintain good quality care. From an organizational perspective, HWT has the potential to save time and costs due to decreased travelling by using digital means to consultations. The use of digital tools can increase job satisfaction but also contribute to perceived job strains and job distress.

Comparison With Prior Work

PC is a holistic approach to care for people with life-limiting conditions and near end of life. It aims to acknowledge and attend to all aspects of the patients' and family members' needs, including physiological, psychological, existential, and social issues [3]. Symptom management is identified as one of the critical areas in PC [141]. According to the current review, HWT can have a positive impact on physical and mental symptoms and support symptom management. Patients in PC suffer from advanced and chronic conditions, often with a negative progression. There was some evidence for the potential of using remote monitoring and care support to prevent disease progression or functional decline, at least for some time. Not very surprisingly, there were no studies presenting any significant effects on mortality.

For symptom management, timely and close contact with the health care team is essential [141]. It has long been recognized that access to and communication with the PC team are vitally important for patients as well as informal caregivers [142,143]. Based on the findings, HWT provides new ways of access and continuity of care, with improved opportunities to follow patients' needs without travelling. The support of patients to remain and continue to receive care at home [31,32] is especially important in cases when clinical conditions or geographic location prevent patients from accessing conventional care [33].

Although there is a concern that the partnership between patients, informal caregivers, and the PC team might be suffering. According to included studies, digital means of communication could be perceived as difficult due to loss of nonverbal cues and limited proximity to the patient when not being in the same room. Meeting online instead of in person can affect the possibilities for the health care professional to be supportive and comforting during difficult situations. Other reviews, on the other hand, have concluded that technology has the potential to not only facilitate communication through the inherent flexibility provided by technology [35] but also to build and enhance relationships [30,31]. These contradictory findings point toward the importance of caution in choosing when to use HWT in PC and for what patients.

Informal caregivers have an important role and take great responsibility for everyday symptom and medicine management

XSL•FO
RenderX

[143]. As having insight into the patient's well-being and needs, they are an important link between patient and the care team, relaying concerns, managing medication, and generally advocating for the patient when they are unable or unwilling to do so [143]. However, they also carry a heavy burden, while they must manage their own health and grief. It is, therefore, vital that they are adequately supported. Among the included studies, there was HWT addressing the caring role, aimed at supporting informal caregivers. Moreover, HWT focusing on the patients' needs were also shown to have an impact on the situation for informal caregivers. PC should take a holistic view of the patient and the informal caregiver, the concerns of both being intertwined and interdependent [143]. Based on the findings, there is research supporting the potential for HWT to complement the ordinary PC to increase support for the caregiving role.

Besides symptom management, access to the health care team, and support of informal caregivers, interprofessional teamwork is emphasized as a crucial element in PC [3]. From a professional perspective, the use of HWT means both possibilities and challenges. According to the included studies, the use of HWT may increase the possibility for professionals to reach the patients and their families when needed, for example, to follow progression and adjust treatment. On the other hand, according to some of the studies, this means yet another responsibility, which may contribute to increased job strain among health care professionals. Previous research suggests that HWT can improve the coordination of care and help build and enhance personal and professional relationships [30]. Other potential effects are time saving and reduction of no-show rates [27].

Limitations and Future Directions

The methodology of this study involved a comprehensive, state-of-the-art review of the existing literature on HWT in PC. With the purpose of enhancing reliability and contributing to the generalizability of findings, the search was conducted in a variety of electronic databases—APA, PsycINFO, Cochrane Library, CINAHL Plus, PubMed, Scopus, and Web of Science Core Collection. A supplementary search ensured up-to-date coverage and inclusion of recent research developments in the rapidly evolving field of HWT. In the search, several concepts were used to describe technology-based interventions in health care, often used interchangeably [144]. HWT is a common term in Nordic countries, but not as commonly used elsewhere. To ensure broad inclusion of relevant studies, broader terms like mHealth, telemedicine, and ambient intelligence were used, although specific technologies such as alarms or monitoring were avoided [145]. This approach may have limited our results to more general articles, excluding those focused on specific technologies.

Notably, among the included studies, there was a predominance of studies focusing on patients with cancer, often neglecting other palliative conditions. Furthermore, the limited diversity in participant demographics—with studies often focused on specific age groups, genders, or cultural contexts—highlights a gap in the representativeness of the findings. Future research should aim to include a wider range of diagnoses and settings to ensure that the findings are applicable to the broader PC population [36]. Moreover, there is a need for more rigorous research related to patient outcomes and evidence regarding effectiveness [29], health system outcomes (eg, usage and costs) [146], and best practices for quality remote PC [27]. New literature reviews should be conducted in a few years to update the state of the art.

Conclusions

PC is both a clinical specialty and an overall approach to care that focuses on improving quality of life and relieving suffering for patients and families facing serious illnesses, based on need and not prognosis. HWT shows potential as a complement to usual PC to increase access and continuity of care, for symptom management, to support patients to remain at home and prevent frequent emergency visits. In some situations, it has the potential to build and maintain relationships between patients, their families, and the health care team, as well as serve as a means for increased interprofessional collaboration and support. However, there are challenges to overcome that might affect the quality of care using HWT. It is unclear whether PC delivered remotely or with support of HWT is equivalent to the more resource-intensive in-person care. Our findings point toward the importance of caution in choosing when to use HWT in PC and for which patients.

Acknowledgments

The authors wish to express their gratitude to Sophiahemmet University and Mälardalen University for their support and provision of resources, which were instrumental in conducting this state-of-the-art review.

Funding

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

XSL•FO
RenderX

Authors' Contributions

SLS completed database searching. CG, MH, MM, RH, and VZ completed article screening and data extraction. MH, RH, and VZ were responsible for data curation and analysis and data interpretation. MH, RH, and VZ wrote the original draft. CG, MM, and SLS provided advice. All authors contributed to reviewing and editing the final manuscript and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR checklist.
[DOCX File , 85 KB - jmir_v28i1e79637_app1.docx]

Multimedia Appendix 2
Documentation database searches.
[DOCX File , 42 KB - jmir_v28i1e79637_app2.docx]

Multimedia Appendix 3
Characteristics of included studies.
[DOCX File , 61 KB - jmir_v28i1e79637_app3.docx]

References

1. Centeno C, Garralda E, Carrasco JM, den Herder-van der Eerden M, Aldridge M, Stevenson D. The palliative care challenge: analysis of barriers and opportunities to integrate palliative care in Europe in the view of national associations. J Palliat Med 2017 Nov;20(11):1195-1204. [doi: 10.1089/jpm.2017.0039] [Medline: 28509657]
2. Knaul FM, Farmer PE, Krakauer EL, De Lima L, Bhadelia A, Jiang Kwete X. Alleviating the access abyss in palliative care and pain relief-an imperative of universal health coverage: the Lancet Commission report. Lancet 2018;391(10128):1391-1454. [doi: 10.1016/S0140-6736(17)32513-8] [Medline: 29032993]
3. Palliative care. World Health Organization. 2018. URL: https://www.who.int/news-room/fact-sheets/detail/palliative-care [accessed 2026-01-16]
4. Radbruch L, De Lima L, Knaul F, Wenk R, Ali Z, Bhatnaghar S. Redefining palliative care—a new consensus-based definition. J Pain Symptom Manage 2020;60(4):754-764 [FREE Full text] [doi: 10.1016/j.jpainsymman.2020.04.027] [Medline: 32387576]
5. Kaasa S, Loge JH, Aapro M, Albreht T, Anderson R, Bruera E. Integration of oncology and palliative care: a Lancet Oncology Commission. Lancet Oncol 2018;19(11):e588-e653 [FREE Full text] [doi: 10.1016/S1470-2045(18)30415-7] [Medline: 30344075]
6. Hansen MB, Rojas-Concha L, Petersen MA, Adsersen M, Groenvold M. Differences in palliative care needs between cancer patients and non-cancer patients at the start of specialized palliative care: a nationwide register-based study. Palliat Med 2024;38(9):1021-1032. [doi: 10.1177/02692163241269705] [Medline: 39194071]
7. Donath C, Ostgathe C, Heckel M. Dementia patients in palliative care according to data from the German National Hospice and Palliative Care Register (2009-2021). BMC Palliat Care 2024;23(1):185 [FREE Full text] [doi: 10.1186/s12904-024-01509-0] [Medline: 39054518]
8. Arnold E, Finucane AM, Oxenham D. Preferred place of death for patients referred to a specialist palliative care service. BMJ Support Palliat Care 2015;5(3):294-296 [FREE Full text] [doi: 10.1136/bmjspcare-2012-000338] [Medline: 24644165]
9. Ali M, Capel M, Jones G, Gazi T. The importance of identifying preferred place of death. BMJ Support Palliat Care 2019;9(1):84-91. [doi: 10.1136/bmjspcare-2015-000878] [Medline: 26408428]
10. Hudson P. Home-based support for palliative care families: challenges and recommendations. Med J Aust 2003;179(S6):S35-S37. [doi: 10.5694/j.1326-5377.2003.tb05576.x] [Medline: 12964935]
11. Mack JW, Jacobson J, Frank D, Cronin AM, Horvath K, Allen V. Evaluation of patient and family outpatient complaints as a strategy to prioritize efforts to improve cancer care delivery. Jt Comm J Qual Patient Saf 2017;43(10):498-507. [doi: 10.1016/j.jcjq.2017.04.008] [Medline: 28942774]
12. Ventura AD, Burney S, Brooker J, Fletcher J, Ricciardelli L. Home-based palliative care: a systematic literature review of the self-reported unmet needs of patients and carers. Palliat Med 2014;28(5):391-402. [doi: 10.1177/0269216313511141] [Medline: 24292156]
13. Lundereng ED, Dihle A, Steindal SA. Nurses' experiences and perspectives on collaborative discharge planning when patients receiving palliative care for cancer are discharged home from hospitals. J Clin Nurs 2020;29(17-18):3382-3391. [doi: 10.1111/jocn.15371] [Medline: 32533726]

XSL•FO
RenderX

14. Aranda SK, Hayman-White K. Home caregivers of the person with advanced cancer: an Australian perspective. Cancer Nurs 2001;24(4):300-307. [doi: 10.1097/00002820-200108000-00011] [Medline: 11502039]
15. Sharpe L, Butow P, Smith C, McConnell D, Clarke S. The relationship between available support, unmet needs and caregiver burden in patients with advanced cancer and their carers. Psychooncology 2005;14(2):102-114. [doi: 10.1002/pon.825] [Medline: 15386783]
16. Nze Ossima A, Szfetel D, Denoyel B, Beloucif O, Texereau J, Champion L. End-of life medical spending and care pathways in the last 12 months of life: a comprehensive analysis of the national claims database in France. Medicine (Baltimore) 2023;102(31):e34555 [FREE Full text] [doi: 10.1097/MD.0000000000034555] [Medline: 37543784]
17. Shepperd S, Gonçalves-Bradley DC, Straus S, Wee B. Hospital at home: home-based end-of-life care. Cochrane Database Syst Rev 2021;3(3):CD009231 [FREE Full text] [doi: 10.1002/14651858.CD009231.pub3] [Medline: 33721912]
18. Budd J, Miller BS, Manning EM, Lampos V, Zhuang M, Edelstein M. Digital technologies in the public-health response to COVID-19. Nat Med 2020;26(8):1183-1192. [doi: 10.1038/s41591-020-1011-4] [Medline: 32770165]
19. Andersson SW, Richardson MX, Cozza M, Lindén M, Redekop K. Addressing evidence in health and welfare technology interventions from different perspectives. Health Policy Technol 2021;10(2):100519. [doi: 10.1016/j.hlpt.2021.100519]
20. Danaher J. Techno-optimism: an analysis, an evaluation and a modest defence. Philos. Technol 2022;35(2):54 [FREE Full text] [doi: 10.1007/s13347-022-00550-2]
21. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
22. Kuoppamäki S. The application and deployment of welfare technology in Swedish municipal care: a qualitative study of procurement practices among municipal actors. BMC Health Serv Res 2021;21(1):918 [FREE Full text] [doi: 10.1186/s12913-021-06944-w] [Medline: 34488740]
23. Basile I, Consolo L, Colombo S, Rusconi D, Rampichini F, Lusignani M. Technology to support older adults in home palliative care: a scoping review. Am J Hosp Palliat Care 2024;41(6):673-690 [FREE Full text] [doi: 10.1177/10499091231189502] [Medline: 37473720]
24. Marshall C, Virdun C, Phillips JL. Evidence-based models of rural palliative care: a systematic review. Palliat Med 2023;37(8):1129-1143. [doi: 10.1177/02692163231183994] [Medline: 37537971]
25. Luis-Martínez R, Monje MHG, Antonini A, Sánchez-Ferro Á, Mestre TA. Technology-enabled care: integrating multidisciplinary care in Parkinson's disease through digital technology. Front Neurol 2020;11:575975 [FREE Full text] [doi: 10.3389/fneur.2020.575975] [Medline: 33250846]
26. Windle N, Alam A, Patel H, Street JM, Lathwood M, Farrington T. A retrospective cohort study evaluating the association between implementation of a digital care plan and hospitalization rates for home care residents in the United Kingdom. Home Health Care Manage Pract 2022;35(2):123-128 [FREE Full text] [doi: 10.1177/10848223221135560]
27. Gordon B, Mason B, Smith SLH. Leveraging telehealth for delivery of palliative care to remote communities: a rapid review. J Palliat Care 2022;37(2):213-225 [FREE Full text] [doi: 10.1177/08258597211001184] [Medline: 33730904]
28. Hancock S, Preston N, Jones H, Gadoud A. Telehealth in palliative care is being described but not evaluated: a systematic review. BMC Palliat Care 2019;18(1):114 [FREE Full text] [doi: 10.1186/s12904-019-0495-5] [Medline: 31835998]
29. Hutchinson R, Chiu EJ, Belin SC, Klein-Fedyshin M, Impagliazzo CR, Costanza L. How is telehealth used to increase access to specialty palliative care? A systematic review. J Pain Symptom Manage 2025;69(4):e303-e314. [doi: 10.1016/j.jpainsymman.2024.12.017] [Medline: 39755286]
30. Lundereng ED, Nes AAG, Holmen H, Winger A, Thygesen H, Jøranson N. Health care professionals' experiences and perspectives on using telehealth for home-based palliative care: scoping review. J Med Internet Res 2023;25:e43429 [FREE Full text] [doi: 10.2196/43429] [Medline: 36989024]
31. Namasivayam P, Bui DT, Low C, Barnett T, Bridgman H, Marsh P. The use of telehealth in the provision of after-hours palliative care services in rural and remote Australia: a scoping review. PLoS One 2022;17(9):e0274861 [FREE Full text] [doi: 10.1371/journal.pone.0274861] [Medline: 36156089]
32. Steindal SA, Nes AAG, Godskesen TE, Holmen H, Winger A, Österlind J. Advantages and challenges of using telehealth for home-based palliative care: systematic mixed studies review. J Med Internet Res 2023;25:e43684 [FREE Full text] [doi: 10.2196/43684] [Medline: 36912876]
33. Sánchez-Cárdenas MA, Iriarte-Aristizábal MF, León-Delgado MX, Rodríguez-Campos LF, Correa-Morales JE, Cañón-Piñeros A. Rural palliative care telemedicine for advanced cancer patients: a systematic review. Am J Hosp Palliat Care 2023;40(8):936-944. [doi: 10.1177/10499091221130329] [Medline: 36331174]
34. Tang M, Reddy A. Telemedicine and its past, present, and future roles in providing palliative care to advanced cancer patients. Cancers (Basel) 2022;14(8):1884 [FREE Full text] [doi: 10.3390/cancers14081884] [Medline: 35454791]
35. Widberg C, Wiklund B, Klarare A. Patients' experiences of eHealth in palliative care: an integrative review. BMC Palliat Care 2020;19(1):158 [FREE Full text] [doi: 10.1186/s12904-020-00667-1] [Medline: 33054746]
36. Jess M, Timm H, Dieperink KB. Video consultations in palliative care: a systematic integrative review. Palliat Med 2019;33(8):942-958. [doi: 10.1177/0269216319854938] [Medline: 31187688]

XSL•FO
RenderX

37. Disalvo D, Agar M, Caplan G, Murtagh FE, Luckett T, Heneka N. Virtual models of care for people with palliative care needs living in their own home: a systematic meta-review and narrative synthesis. Palliat Med 2021;35(8):1385-1406. [doi: 10.1177/02692163211024451] [Medline: 34169759]
38. Greenhalgh T. Five biases of new technologies. Br J Gen Pract 2013;63(613):425 [FREE Full text] [doi: 10.3399/bjgp13X670741] [Medline: 23972179]
39. Mormina M. Science, technology and innovation as social goods for development: rethinking research capacity building from Sen's capabilities approach. Sci Eng Ethics 2019;25(3):671-692 [FREE Full text] [doi: 10.1007/s11948-018-0037-1] [Medline: 29497970]
40. Hofmann B. Progress bias versus status quo bias in the ethics of emerging science and technology. Bioethics 2020;34(3):252-263. [doi: 10.1111/bioe.12622] [Medline: 31617222]
41. Finucane AM, O'Donnell H, Lugton J, Gibson-Watt T, Swenson C, Pagliari C. Digital health interventions in palliative care: a systematic meta-review. NPJ Digit Med 2021;4(1):64 [FREE Full text] [doi: 10.1038/s41746-021-00430-7] [Medline: 33824407]
42. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Info Libr J 2009;26(2):91-108 [FREE Full text] [doi: 10.1111/j.1471-1842.2009.00848.x] [Medline: 19490148]
43. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
44. Gustafsson C, Zander V, Stridsberg S, Holm M, Mazaheri M, Hedman R. Health and welfare technology in palliative care. A state-of-the-art review. Open Science Framework. 2022. URL: https://osf.io/yr85w [accessed 2026-01-23]
45. Covidence systematic review software. Veritas Health Innovation. URL: https://www.covidence.org/ [accessed 2026-01-16]
46. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008;8:45 [FREE Full text] [doi: 10.1186/1471-2288-8-45] [Medline: 18616818]
47. Kazankov K, Novelli S, Chatterjee DA, Phillips A, Balaji A, Raja M. Evaluation of CirrhoCare—a digital health solution for home management of individuals with cirrhosis. J Hepatol 2023;78(1):123-132. [doi: 10.1016/j.jhep.2022.08.034] [Medline: 36087864]
48. Jiang B, Bills M, Poon P. Integrated telehealth-assisted home-based specialist palliative care in rural Australia: a feasibility study. J Telemed Telecare 2023;29(1):50-57. [doi: 10.1177/1357633X20966466] [Medline: 33079611]
49. Gustafson D, DuBenske LL, Namkoong K, Hawkins R, Chih MY, Atwood AK. An eHealth system supporting palliative care for patients with non-small cell lung cancer: a randomized trial. Cancer 2013;119(9):1744-1751 [FREE Full text] [doi: 10.1002/cncr.27939] [Medline: 23355273]
50. Ngoma M, Mushi B, Morse RS, Ngoma T, Mahuna H, Lambden K. mPalliative Care Link: examination of a mobile solution to palliative care coordination among Tanzanian patients with cancer. JCO Glob Oncol 2021;7:1306-1315 [FREE Full text] [doi: 10.1200/GO.21.00122] [Medline: 34406856]
51. Manz CR, Zhang Y, Chen K, Long Q, Small DS, Evans CN. Long-term effect of machine learning–triggered behavioral nudges on serious illness conversations and end-of-life outcomes among patients with cancer. JAMA Oncol 2023;9(3):414-418 [FREE Full text] [doi: 10.1001/jamaoncol.2022.6303] [Medline: 36633868]
52. Tumeh IBRG, Bergerot CD, Lee D, Philip EJ, Freitas-Júnior R. mHealth program for patients with advanced cancer receiving treatment in a public health hospital in Brazil. Psychooncology 2023;32(1):125-132. [doi: 10.1002/pon.6059] [Medline: 36284459]
53. Mooney K, Whisenant MS, Wilson CM, Coombs LA, Lloyd J, Alekhina N. Technology-assisted mHealth caregiver support to manage cancer patient symptoms: a randomized controlled trial. J Pain Symptom Manage 2023;66(1):33-43. [doi: 10.1016/j.jpainsymman.2023.02.320] [Medline: 36889453]
54. Besse KTC, Faber-te Boveldt ND, Janssen GHP, Vernooij-Dassen M, Vissers KCP, Engels Y. Pain assessment with short message service and interactive voice response in outpatients with cancer and pain: a feasibility study. Pain Pract 2016;16(3):320-326. [doi: 10.1111/papr.12278] [Medline: 25581306]
55. Cornetta K, Nyariki S, Manji I, Kiplimo I, Korir M, Muinga E. Telehospice for cancer patients discharged from a tertiary care hospital in Western Kenya. J Pain Symptom Manage 2023;65(5):378-387 [FREE Full text] [doi: 10.1016/j.jpainsymman.2023.01.027] [Medline: 36773732]
56. Mark DD, Ikehara C, Matsuura C, Hara K, Li D. Validating the impact of teaching pursed-lips breathing with Skype. J Hosp Palliat Nurs 2013;15(8):424-432. [doi: 10.1097/njh.0000000000000015]
57. Rafter L. Employing e-health in the palliative care setting to manage pressure ulcers. Wounds UK 2016;12(1):114-123. [doi: 10.1093/oso/9780198508434.003.0028]
58. Moscato S, Sichi V, Giannelli A, Palumbo P, Ostan R, Varani S. Virtual reality in home palliative care: brief report on the effect on cancer-related symptomatology. Front Psychol 2021;12:709154 [FREE Full text] [doi: 10.3389/fpsyg.2021.709154] [Medline: 34630217]
59. Milbury K, Li Y, Durrani S, Liao Z, Tsao AS, Carmack C. A mindfulness-based intervention as a supportive care strategy for patients with metastatic non-small cell lung cancer and their spouses: results of a three-arm pilot randomized controlled trial. Oncologist 2020;25(11):e1794-e1802 [FREE Full text] [doi: 10.1634/theoncologist.2020-0125] [Medline: 32621630]

XSL•FO
RenderX

60. Nemecek R, Huber P, Schur S, Masel EK, Baumann L, Hoeller C. Telemedically augmented palliative care: empowerment for patients with advanced cancer and their family caregivers. Wien Klin Wochenschr 2019;131(23-24):620-626. [doi: 10.1007/s00508-019-01562-3] [Medline: 31667588]
61. Watanabe SM, Fairchild A, Pituskin E, Borgersen P, Hanson J, Fassbender K. Improving access to specialist multidisciplinary palliative care consultation for rural cancer patients by videoconferencing: report of a pilot project. Support Care Cancer 2013;21(4):1201-1217. [doi: 10.1007/s00520-012-1649-7] [Medline: 23161339]
62. Applebaum AJ, Buda KL, Schofield E, Farberov M, Teitelbaum ND, Evans K. Exploring the cancer caregiver's journey through web-based meaning-centered psychotherapy. Psychooncology 2018;27(3):847-856 [FREE Full text] [doi: 10.1002/pon.4583] [Medline: 29136682]
63. Hoek PD, Schers HJ, Bronkhorst EM, Vissers KCP, Hasselaar JGJ. The effect of weekly specialist palliative care teleconsultations in patients with advanced cancer -a randomized clinical trial. BMC Med 2017;15(1):119 [FREE Full text] [doi: 10.1186/s12916-017-0866-9] [Medline: 28625164]
64. Ho YX, Morse RS, Lambden K, Mushi BP, Ngoma M, Mahuna H. How a digital case management platform affects community-based palliative care of sub-Saharan African cancer patients: clinician-users' perspectives. Appl Clin Inform 2022;13(5):1092-1099 [FREE Full text] [doi: 10.1055/s-0042-1758223] [Medline: 36384234]
65. LeBaron V, Alam R, Bennett R, Blackhall L, Gordon K, Hayes J. Deploying the Behavioral and Environmental Sensing and Intervention for Cancer Smart Health System to support patients and family caregivers in managing pain: feasibility and acceptability study. JMIR Cancer 2022;8(3):e36879 [FREE Full text] [doi: 10.2196/36879] [Medline: 35943791]
66. Graven LJ, Abbott L, Schluck G. The coping in heart failure (COPE-HF) partnership intervention for heart failure symptoms: implications for palliative care. Prog Palliat Care 2022;31(3):169-178. [doi: 10.1080/09699260.2022.2124144]
67. Cameron P. Caregiver comfort with telehospice. Home Healthc Now 2021;39(1):32-38. [doi: 10.1097/NHH.0000000000000928] [Medline: 33417360]
68. Hennemann-Krause L, Lopes AJ, Araújo JA, Petersen EM, Nunes RA. The assessment of telemedicine to support outpatient palliative care in advanced cancer. Palliat Support Care 2015;13(4):1025-1030. [doi: 10.1017/S147895151400100X] [Medline: 25159308]
69. Maguire R, Connaghan J, Arber A, Klepacz N, Blyth KG, McPhelim J. Advanced Symptom Management System for patients with malignant pleural mesothelioma (ASyMSmeso): mixed methods study. J Med Internet Res 2020;22(11):e19180 [FREE Full text] [doi: 10.2196/19180] [Medline: 33180025]
70. Palma A, Rojas V, Ihl F, Ávila C, Plaza-Parrochia F, Estuardo N. Implementation of a palliative hospital-centered spiritual and psychological telehealth system during COVID-19 pandemic. J Pain Symptom Manage 2021;62(5):1015-1019 [FREE Full text] [doi: 10.1016/j.jpainsymman.2021.04.016] [Medline: 33957254]
71. Gustafson DH, DuBenske LL, Atwood AK, Chih M, Johnson RA, McTavish F. Reducing symptom distress in patients with advanced cancer using an e-alert system for caregivers: pooled analysis of two randomized clinical trials. J Med Internet Res 2017;19(11):e354 [FREE Full text] [doi: 10.2196/jmir.7466] [Medline: 29138131]
72. Zeiser SR. Quality improvement evaluation of a Veterans' palliative care telehealth music therapy program. Music Ther Perspect 2023;41(1):54-62. [doi: 10.1093/mtp/miac017]
73. Greer J, Jacobs J, Pensak N, MacDonald JJ, Fuh CX, Perez GK. Randomized trial of a tailored cognitive-behavioral therapy mobile application for anxiety in patients with incurable cancer. Oncologist 2019;24(8):1111-1120. [doi: 10.1634/theoncologist.2018-0536] [Medline: 30683710]
74. Merz A, Mohamed A, Corbett C, Herring K, Hildenbrand J, Locke SC. A single-site pilot feasibility randomized trial of a supportive care mobile application intervention for patients with advanced cancer and caregivers. Support Care Cancer 2022;30(10):7853-7861. [doi: 10.1007/s00520-022-07224-x] [Medline: 35718794]
75. Baird-Bower D, Roach J, Andrews M, Onslow F, Curnin E. Help is just a phone call away: after-hours support for palliative care patients wishing to die at home. Int J Palliat Nurs 2016;22(6):286-291. [doi: 10.12968/ijpn.2016.22.6.286] [Medline: 27349847]
76. Finucane AM, Davydaitis D, Horseman Z, Carduff E, Baughan P, Tapsfield J. Electronic care coordination systems for people with advanced progressive illness: a mixed-methods evaluation in Scottish primary care. Br J Gen Pract 2020;70(690):e20-e28 [FREE Full text] [doi: 10.3399/bjgp19X707117] [Medline: 31848198]
77. Johnston J, Hillier LM, Purdon M, Pears D, Robson M. Opportunities for improving the quality of life of dying persons: use of technology to connect people at the end of life. J Palliat Care 2018;28(3):178-180. [doi: 10.1177/082585971202800309]
78. Royackers AR, Regan S, Donelle L. The eShift model of care: informal caregivers' experience of a new model of home-based palliative care. Prog Palliat Care 2016;24(2):84-92. [doi: 10.1179/1743291x15y.0000000006]
79. LoCastro M, Sanapala C, Mendler JH, Norton S, Bernacki R, Carroll T. Adaptation of Serious Illness Care Program to be delivered via telehealth for older patients with hematologic malignancy. Blood Adv 2023;7(9):1871-1884 [FREE Full text] [doi: 10.1182/bloodadvances.2022008996] [Medline: 36521100]
80. Harding R, Carrasco JM, Serrano-Pons J, Lemaire J, Namisango E, Luyirika E. Design and evaluation of a novel mobile phone application to improve palliative home-care in resource-limited settings. J Pain Symptom Manage 2021;62(1):1-9 [FREE Full text] [doi: 10.1016/j.jpainsymman.2020.09.045] [Medline: 33246073]

XSL•FO
RenderX

81. Wilkie DJ, Yao Y, Ezenwa MO, Suarez ML, Dyal BW, Gill A. A stepped-wedge randomized controlled trial: effects of eHealth interventions for pain control among adults with cancer in hospice. J Pain Symptom Manage 2020;59(3):626-636 [FREE Full text] [doi: 10.1016/j.jpainsymman.2019.10.028] [Medline: 31711969]
82. Hao Y, Zhan L, Huang M, Cui X, Zhou Y, Xu E. Nurses' knowledge and attitudes towards palliative care and death: a learning intervention. BMC Palliat Care 2021;20(1):50 [FREE Full text] [doi: 10.1186/s12904-021-00738-x] [Medline: 33765995]
83. White C, McIlfatrick S, Dunwoody L, Watson M. Supporting and improving community health services-a prospective evaluation of ECHO technology in community palliative care nursing teams. BMJ Support Palliat Care 2019;9(2):202-208. [doi: 10.1136/bmjspcare-2015-000935] [Medline: 26628535]
84. Chiu Wu CH, Perng S, Shi C, Lai H. Advance care planning and advance directives: a multimedia education program in community-dwelling older adults. J Appl Gerontol 2020;39(8):811-819. [doi: 10.1177/0733464819831596] [Medline: 30795711]
85. Köhle N, Drossaert CHC, Jaran J, Schreurs KMG, Verdonck-de Leeuw IM, Bohlmeijer ET. User-experiences with a web-based self-help intervention for partners of cancer patients based on acceptance and commitment therapy and self-compassion: a qualitative study. BMC Public Health 2017;17(1):225 [FREE Full text] [doi: 10.1186/s12889-017-4121-2] [Medline: 28245794]
86. Ariss SMB, Taylor P, Fitzsimmons D, Kyeremateng S, Mawson S. Mobile technology and delegated work in specialist community services: the EnComPaSS Integration project. BMJ Support Palliat Care 2021:e1144-e1151 [FREE Full text] [doi: 10.1136/bmjspcare-2020-002288] [Medline: 34916238]
87. Samara JL, Liu W, Kroon W, Harvie B, Hingeley R, Johnston N. Telehealth palliative care needs rounds during a pandemic. The Journal for Nurse Practitioners 2021;17(3):335-338. [doi: 10.1016/j.nurpra.2020.12.009]
88. Shirley LD, DiScala S, Brooks A, Vartan CM, Silverman M. Pilot of a pharmacist-integrated interprofessional team to optimize prescribing in a telemedicine palliative care clinic. J Am Coll Clin Pharm 2021;4(9):1093-1099. [doi: 10.1002/jac5.1478]
89. Bhargava R, Keating B, Isenberg SR, Subramaniam S, Wegier P, Chasen M. RELIEF: a digital health tool for the remote self-reporting of symptoms in patients with cancer to address palliative care needs and minimize emergency department visits. Curr Oncol 2021;28(6):4273-4280 [FREE Full text] [doi: 10.3390/curroncol28060363] [Medline: 34898539]
90. Marco D, McMillan-Drendel E, Philip JAM, Williamson T, Le B. Establishment of the first Australian public and health-professional palliative care advice service: exploring caller needs and gaps in care. Aust Health Rev 2023;47(5):569-573. [doi: 10.1071/AH23108] [Medline: 37516935]
91. Atreya S, Kumar G, Samal J, Bhattacharya M, Banerjee S, Mallick P. Patients'/caregivers' perspectives on telemedicine service for advanced cancer patients during the COVID-19 pandemic: an exploratory survey. Indian J Palliat Care 2020;26(Suppl 1):S40-S44 [FREE Full text] [doi: 10.4103/IJPC.IJPC_145_20] [Medline: 33088085]
92. Teter CM, Whitted K, Vardaman SA. Reducing hospital visit rates in hospice patients using telemedicine. Comput Inform Nurs 2021;39(11):682-688. [doi: 10.1097/CIN.0000000000000769] [Medline: 34145206]
93. Bange EM, Li Y, Kumar P, Doucette A, Gabriel P, Parikh R. The association between telemedicine, advance care planning, and unplanned hospitalizations among high-risk patients with cancer. Cancer 2024;130(4):636-644 [FREE Full text] [doi: 10.1002/cncr.35116] [Medline: 37987207]
94. Baxter KE, Kochar S, Williams C, Blackman C, Himmelvo J. Development of a palliative telehealth pilot to meet the needs of the nursing home population. J Hosp Palliat Nurs 2021;23(5):478-483. [doi: 10.1097/NJH.0000000000000784] [Medline: 34225341]
95. Bückmann A, Bernhardt F, Eveslage M, Storck M, Thölking G, Buss H. Telemedical consultations in palliative care: benefits through knowledge exchange and intercollegiate collaboration—findings from the German oVID project. Cancers (Basel) 2023;15(9):2512 [FREE Full text] [doi: 10.3390/cancers15092512] [Medline: 37173978]
96. Morgan DD, Swetenham K, To THM, Currow DC, Tieman JJ. Telemonitoring via self-report and video review in community palliative care: a case report. Healthcare (Basel) 2017;5(3):51 [FREE Full text] [doi: 10.3390/healthcare5030051] [Medline: 28858221]
97. McCreedy EM, Yang X, Mitchell SL, Gutman R, Teno J, Loomer L. Effect of advance care planning video on do-not-hospitalize orders for nursing home residents with advanced illness. BMC Geriatr 2022;22(1):298 [FREE Full text] [doi: 10.1186/s12877-022-02970-3] [Medline: 35392827]
98. Takahashi PY, Hanson GJ, Thorsteinsdottir B, Van Houten HK, Shah ND, Naessens JM. The impact of telemonitoring upon hospice referral in the community: a randomized controlled trial. Clin Interv Aging 2012;7:445-451 [FREE Full text] [doi: 10.2147/CIA.S36461] [Medline: 23152674]
99. Hayek S, Nieva R, Corrigan F, Zhou A, Mudaliar U, Mays D. End-of-life care planning: improving documentation of advance directives in the outpatient clinic using electronic medical records. J Palliat Med 2014;17(12):1348-1352. [doi: 10.1089/jpm.2013.0684] [Medline: 24988497]
100. Rainsford S, Hall Dykgraaf S, Phillips C. Effectiveness of telehealth palliative care Needs Rounds in rural residential aged care during the COVID-19 pandemic: a hybrid effectiveness-implementation study. Aust J Rural Health 2022;30(1):108-114 [FREE Full text] [doi: 10.1111/ajr.12789] [Medline: 34757687]

XSL•FO
RenderX

101. Taubert M, Norris J, Edwards S, Snow V, Finlay IG. Talk CPR - a technology project to improve communication in do not attempt cardiopulmonary resuscitation decisions in palliative illness. BMC Palliat Care 2018;17(1):118 [FREE Full text] [doi: 10.1186/s12904-018-0370-9] [Medline: 30340632]
102. Beesley VL, Turner J, Chan RJ, Yates P, Gordon LG, Burge M. Supporting patients and carers affected by pancreatic cancer: a feasibility study of a counselling intervention. Eur J Oncol Nurs 2020;46:101729. [doi: 10.1016/j.ejon.2020.101729] [Medline: 32339910]
103. Alcalde Castro MJ, Zaig S, Nissim R, O'Connor B, Lau J, Mak E. Telehealth outpatient palliative care in the COVID-19 pandemic: patient experience qualitative study. BMJ Support Palliat Care 2024;14:e1512-e1520. [doi: 10.1136/spcare-2023-004189] [Medline: 36828625]
104. Bandini JI, Scherling A, Farmer C, Fratkin M, Mistler B, Ast K. Experiences with telehealth for outpatient palliative care: findings from a mixed-methods study of patients and providers across the United States. J Palliat Med 2022;25(7):1079-1087. [doi: 10.1089/jpm.2021.0545] [Medline: 35506994]
105. Bentley B, O'Connor M, Williams A, Breen LJ. Dignity therapy online: piloting an online psychosocial intervention for people with terminal illness. Digit Health 2020;6:2055207620958527 [FREE Full text] [doi: 10.1177/2055207620958527] [Medline: 33014409]
106. Bonsignore L, Bloom N, Steinhauser K, Nichols R, Allen T, Twaddle M. Evaluating the feasibility and acceptability of a telehealth program in a rural palliative care population: TapCloud for palliative care. J Pain Symptom Manage 2018;56(1):7-14 [FREE Full text] [doi: 10.1016/j.jpainsymman.2018.03.013] [Medline: 29551433]
107. Lind LK, Karlsson D. Electronic patient-reported symptom assessment in palliative end-of-life home care. Health Syst 2017;2(3):171-180. [doi: 10.1057/hs.2013.4]
108. Read Paul L, Salmon C, Sinnarajah A, Spice R. Web-based videoconferencing for rural palliative care consultation with elderly patients at home. Support Care Cancer 2019;27(9):3321-3330. [doi: 10.1007/s00520-018-4580-8] [Medline: 30613908]
109. Groothuizen JE, Aroyewun E, Zasada M, Harris J, Hewish M, Taylor C. Virtually the same? Examining the impact of the COVID-19 related shift to virtual lung cancer multidisciplinary team meetings in the UK National Health Service: a mixed methods study. BMJ Open 2023;13(6):e065494 [FREE Full text] [doi: 10.1136/bmjopen-2022-065494] [Medline: 37328174]
110. Oelschlägel L, Christensen VL, Moen A, Heggdal K, Österlind J, Dihle A. Patients' experiences with a welfare technology application for remote home care: a longitudinal study. J Clin Nurs 2023;32(17-18):6545-6558. [doi: 10.1111/jocn.16592] [Medline: 36437490]
111. Haydon HM, Snoswell CL, Thomas EE, Broadbent A, Caffery LJ, Brydon J. Enhancing a community palliative care service with telehealth leads to efficiency gains and improves job satisfaction. J Telemed Telecare 2021;27(10):625-630. [doi: 10.1177/1357633X211048952] [Medline: 34726990]
112. Hutchinson RN, Anderson EC, Ruben MA, Manning N, John L, Daruvala A. A formative mixed-methods study of emotional responsiveness in telepalliative care. J Palliat Med 2022;25(8):1258-1267 [FREE Full text] [doi: 10.1089/jpm.2021.0589] [Medline: 35417249]
113. Pasanen L, Le Gautier R, Wong A, Wawryk O, Collins A, Schwetlik S. Telehealth in outpatient delivery of palliative care: a qualitative study of patient and physician views. Palliat Support Care 2023;21(6):980-987. [doi: 10.1017/S1478951522000670] [Medline: 35818898]
114. Tasneem S, Kim A, Bagheri A, Lebret J. Telemedicine video visits for patients receiving palliative care: a qualitative study. Am J Hosp Palliat Care 2019;36(9):789-794. [doi: 10.1177/1049909119846843] [Medline: 31064195]
115. Soliman AA, Akgün KM, Coffee J, Kapo J, Morrison LJ, Hopkinson E. Quality of telehealth-delivered inpatient palliative care during the early COVID-19 pandemic. J Pain Symptom Manage 2023;65(1):6-15 [FREE Full text] [doi: 10.1016/j.jpainsymman.2022.09.014] [Medline: 36206949]
116. van Gurp J, van Selm M, Vissers K, van Leeuwen E, Hasselaar J. How outpatient palliative care teleconsultation facilitates empathic patient-professional relationships: a qualitative study. PLoS One 2015;10(4):e0124387 [FREE Full text] [doi: 10.1371/journal.pone.0124387] [Medline: 25902263]
117. Viitala A, Åstedt-Kurki P, Lehto JT, Palonen M. Online follow-up with a mobile device improves incurable cancer patients' coping - a qualitative study. Eur J Oncol Nurs 2021;55:102047 [FREE Full text] [doi: 10.1016/j.ejon.2021.102047] [Medline: 34688053]
118. Rosa WE, Lynch KA, Hadler RA, Mahoney C, Parker PA. "Doing palliative care with my hands tied behind my back": telepalliative care delivery for oncology inpatients during a COVID-19 surge. Transl Behav Med 2022;12(7):816-824 [FREE Full text] [doi: 10.1093/tbm/ibac044] [Medline: 35716077]
119. Collier A, Morgan DD, Swetenham K, To TH, Currow DC, Tieman JJ. Implementation of a pilot telehealth programme in community palliative care: a qualitative study of clinicians' perspectives. Palliat Med 2016;30(4):409-417. [doi: 10.1177/0269216315600113] [Medline: 26290500]
120. Misplon S, Marneffe W, Himpe U, Hellings J, Demedts I. Evaluation of the implementation of value-based healthcare with a weekly digital follow-up of lung cancer patients in clinical practice. Eur J Cancer Care (Engl) 2022;31(6):e13653. [doi: 10.1111/ecc.13653] [Medline: 35819130]

XSL•FO
RenderX

121. Benson JJ, Oliver DP, Washington KT, Rolbiecki AJ, Lombardo CB, Garza JE. Online social support groups for informal caregivers of hospice patients with cancer. Eur J Oncol Nurs 2020;44:101698 [FREE Full text] [doi: 10.1016/j.ejon.2019.101698] [Medline: 31816508]
122. Elma A, Cook D, Howard M, Takaoka A, Hoad N, Swinton M. Use of video technology in end-of-life care for hospitalized patients during the COVID-19 pandemic. Am J Crit Care 2022;31(3):240-248. [doi: 10.4037/ajcc2022722] [Medline: 35118491]
123. Halton-Hernandez E, Abrams R, Cooney G, Ali K. Exploring audio recording in terminally ill patients receiving hospice care: “Stories for Life Charity” model. J Hosp Palliat Nurs 2023;25(5):271-276. [doi: 10.1097/NJH.0000000000000958] [Medline: 37347958]
124. Kuntz JG, Kavalieratos D, Esper GJ, Ogbu N, Mitchell J, Ellis CM. Feasibility and acceptability of inpatient palliative care e-family meetings during COVID-19 pandemic. J Pain Symptom Manage 2020;60(3):e28-e32 [FREE Full text] [doi: 10.1016/j.jpainsymman.2020.06.001] [Medline: 32505643]
125. Guo Q, Cann B, McClement S, Thompson G, Chochinov HM. Keep in Touch (KIT): feasibility of using internet-based communication and information technology in palliative care. BMC Palliat Care 2017;16(1):29 [FREE Full text] [doi: 10.1186/s12904-017-0203-2] [Medline: 28477630]
126. Potter G, Pesut B, Hooper BP, Erbacker L. Team-based education in a palliative approach for rural nurses and unlicensed care providers. J Contin Educ Nurs 2015;46(6):279-288. [doi: 10.3928/00220124-20150518-04] [Medline: 26057165]
127. van Gurp J, van Selm M, van Leeuwen E, Vissers K, Hasselaar J. Teleconsultation for integrated palliative care at home: a qualitative study. Palliat Med 2016;30(3):257-269. [doi: 10.1177/0269216315598068] [Medline: 26269323]
128. Kamal AH, Wolf S, Nicolla JM, Friedman F, Xuan M, Bennett AV. Usability of PCforMe in patients with advanced cancer referred to outpatient palliative care: results of a randomized, active-controlled pilot trial. J Pain Symptom Manage 2019;58(3):382-389 [FREE Full text] [doi: 10.1016/j.jpainsymman.2019.05.007] [Medline: 31163259]
129. Middleton-Green L, Gadoud A, Norris B, Sargeant A, Nair S, Wilson L. ‘A friend in the corner’: supporting people at home in the last year of life via telephone and video consultation—an evaluation. BMJ Support Palliat Care 2019;9(4):e26. [doi: 10.1136/bmjspcare-2015-001016] [Medline: 26850118]
130. Dionne-Odom JN, Ejem DB, Wells R, Azuero A, Stockdill ML, Keebler K. Effects of a telehealth early palliative care intervention for family caregivers of persons with advanced heart failure: The ENABLE CHF-PC randomized clinical trial. JAMA Netw Open 2020;3(4):e202583 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.2583] [Medline: 32282044]
131. Parker D, Hudson P, Tieman J, Thomas K, Saward D, Ivynian S. Evaluation of an online toolkit for carers of people with a life-limiting illness at the end-of-life: health professionals' perspectives. Aust J Prim Health 2021;27(6):473-478 [FREE Full text] [doi: 10.1071/PY21019] [Medline: 34802508]
132. Cruz-Oliver DM, Abshire Saylor M, Nelson KE, Milner GE, Blinka MD, Durkin N. Hospice family caregiver perceptions of benefits and challenges of a telenovela educational intervention. J Palliat Med 2022;25(6):945-951 [FREE Full text] [doi: 10.1089/jpm.2021.0628] [Medline: 35446674]
133. Koufacos NS, Gottesman EM, Dorisca E, Howe JL. Supporting caregivers of veterans with dementia. J Soc Work End Life Palliat Care 2023;19(1):12-22. [doi: 10.1080/15524256.2023.2184597] [Medline: 36856029]
134. Olesen LK, la Cour K, With H, Mahoney AF, Handberg C. A cross-sectional evaluation of acceptability of an online palliative rehabilitation program for family caregivers of people with amyotrophic lateral sclerosis and cognitive and behavioral impairments. BMC Health Serv Res 2022;22(1):697 [FREE Full text] [doi: 10.1186/s12913-022-07986-4] [Medline: 35610609]
135. Rochmawati EP, Putranto D. Mobile application-based education to improve family caregivers’ readiness: feasibility study. J Nurse Pract 2022;18(10):1097-1101. [doi: 10.1016/j.nurpra.2022.08.023]
136. Schuler T, King C, Matsveru T, Back M, Clark K, Chin D. Wearable-triggered ecological momentary assessments are feasible in people with advanced cancer and their family caregivers: feasibility study from an outpatient palliative care clinic at a cancer center. J Palliat Med 2023;26(7):980-985. [doi: 10.1089/jpm.2022.0535] [Medline: 37134212]
137. DuBenske LL, Gustafson DH, Namkoong K, Hawkins RP, Atwood AK, Brown RL. CHESS improves cancer caregivers' burden and mood: results of an eHealth RCT. Health Psychol 2014;33(10):1261-1272 [FREE Full text] [doi: 10.1037/a0034216] [Medline: 24245838]
138. Chih M, DuBenske LL, Hawkins RP, Brown RL, Dinauer SK, Cleary JF. Communicating advanced cancer patients’ symptoms via the internet: a pooled analysis of two randomized trials examining caregiver preparedness, physical burden, and negative mood. Palliat Med 2013;27(6):533-543 [FREE Full text] [doi: 10.1177/0269216312457213] [Medline: 22988042]
139. Schuit AS, Rienks MM, Hooghiemstra N, Jansen F, Lissenberg-Witte BI, Cuijpers P. Reach and efficacy of the eHealth application Oncokompas, facilitating partners of incurably ill cancer patients to self-manage their caregiver needs: a randomized controlled trial. Support Care Cancer 2022;30(12):10191-10201 [FREE Full text] [doi: 10.1007/s00520-022-07441-4] [Medline: 36367563]
140. Hamilton E, Van Veldhuizen E, Brown A, Brennan S, Sabesan S. Telehealth in radiation oncology at the Townsville Cancer Centre: service evaluation and patient satisfaction. Clin Transl Radiat Oncol 2019;15:20-25 [FREE Full text] [doi: 10.1016/j.ctro.2018.11.005] [Medline: 30582017]

XSL•FO
RenderX

141. Wilkie DJ, Ezenwa MO. Pain and symptom management in palliative care and at end of life. Nurs Outlook 2012;60(6):357-364 [FREE Full text] [doi: 10.1016/j.outlook.2012.08.002] [Medline: 22985972]
142. Murray CD, McDonald C, Atkin H. The communication experiences of patients with palliative care needs: a systematic review and meta-synthesis of qualitative findings. Palliat Support Care 2015;13(2):369-383. [doi: 10.1017/S1478951514000455] [Medline: 24784479]
143. Chapman EJ, Paley CA, Pini S, Ziegler LE. Correction: updating a conceptual model of effective symptom management in palliative care to include patient and carer perspective: a qualitative study. BMC Palliat Care 2024;23(1):237 [FREE Full text] [doi: 10.1186/s12904-024-01571-8] [Medline: 39367358]
144. Otto L, Harst L, Schlieter H, Wollschlaeger B, Richter P. Towards a unified understanding of ehealth and related terms: proposal of a consolidated terminological basis. 2018 Presented at: Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018) – HEALTHINF; January 18-20, 2018; Funchal. [doi: 10.5220/0006651005330539]
145. Borg J, Gustafsson C, Landerdahl Stridsberg S, Zander V. Implementation of welfare technology: a state-of-the-art review of knowledge gaps and research needs. Disabil Rehabil Assist Technol 2023;18(2):227-239 [FREE Full text] [doi: 10.1080/17483107.2022.2120104] [Medline: 36103349]
146. Kavalieratos D, Gelfman LP, Tycon LE, Riegel B, Bekelman DB, Ikejiani DZ. Palliative care in heart failure: rationale, evidence, and future priorities. J Am Coll Cardiol 2017;70(15):1919-1930 [FREE Full text] [doi: 10.1016/j.jacc.2017.08.036] [Medline: 28982506]

Abbreviations

COPD: chronic obstructive pulmonary disease
HWT: health and welfare technology
MeSH: Medical Subject Headings
PC: palliative care
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by K Liew; submitted 25.Jun.2025; peer-reviewed by WJ She, SHX Ma; comments to author 19.Sep.2025; revised version received 18.Dec.2025; accepted 10.Jan.2026; published 12.Mar.2026.

Please cite as:
Zander V, Holm M, Mazaheri M, Gustafsson C, Landerdahl Stridsberg S, Hedman R
Use of Health and Welfare Technology in Palliative Care: State-of-the-Art Review
J Med Internet Res 2026;28:e79637
URL: https://www.jmir.org/2026/1/e79637
doi:10.2196/79637
PMID:

©Viktoria Zander, Maja Holm, Monir Mazaheri, Christine Gustafsson, Sara Landerdahl Stridsberg, Ragnhild Hedman. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Effects of Digital Health Interventions on Motor Symptoms, Nonmotor Symptoms, and Quality of Life in Patients With Parkinson Disease: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Ruwen Liu[1,2*], MS; Sirui Zhang[1,2*], MS; Yi Xiao[1,2], MD; Yangfan Cheng[1,2], MD; Huifang Shang[1,2], MD

[1]Department of Neurology, Laboratory of Neurodegenerative Disorders, National Clinical Research Center for Geriatric, West China Hospital of Sichuan University, Chengdu, China

[2]Rare Disease Center, West China Hospital of Sichuan University, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Huifang Shang, MD
Department of Neurology, Laboratory of Neurodegenerative Disorders, National Clinical Research Center for Geriatric
West China Hospital of Sichuan University
37 Guoxue Alley
Chengdu, 610041
China
Phone: 1 028 85422114
Fax: 1 028 85582944
Email: shanghuifang@scu.edu.cn

Abstract

Background: Parkinson disease (PD) is a progressive neurodegenerative disorder with increasing global prevalence, necessitating innovative management. Digital health interventions (DHIs) offer potential advantages for PD care; yet, a comprehensive systematic review and synthesis across all DHI types and core outcomes is still lacking.

Objective: This review aimed to assess the effectiveness of DHIs for improving motor symptoms, nonmotor symptoms, and quality of life in patients with PD and to summarize the reach, uptake, and feasibility.

Methods: We searched PubMed, Ovid Embase, Web of Science, CINAHL, Cochrane Central Register of Controlled Trials, and APA PsycINFO up to November 2025. Pooled standardized mean differences (SMDs) were calculated using random-effects models. We calculated 95% prediction intervals (PIs) to estimate the true effects. The revised Cochrane Risk of Bias 2 tool was used to assess risk of bias. Heterogeneity was assessed using I^2, τ^2, and 95% PI. Subgroup analyses, meta-regression, and sensitivity analyses were conducted to address heterogeneity and potential bias. The quality of evidence was assessed using GRADE (Grading of Recommendations Assessment, Development, and Evaluation).

Results: The review included 112 randomized controlled trials involving 5594 participants. Significant postintervention improvements were identified in motor symptoms (SMD=–0.39, 95% CI –0.60 to –0.18, 95% PI –1.75 to 0.99; I^2=80.3%) and overall nonmotor symptoms (SMD=–0.26, 95% CI –0.49 to –0.03, 95% PI –0.56 to 0.03; I^2=13.8%), including cognitive function (SMD=0.47, 95% CI 0.22 to 0.72, 95% PI –0.41 to 1.35; I^2=63.5%) and psychiatric symptoms (SMD=–0.42, 95% CI –0.74 to –0.09, 95% PI –1.82 to –0.99; I^2=85.4%); however, there was no significant enhancement in quality of life (SMD=–0.19, 95% CI –0.47 to 0.09, 95% PI –1.50 to 1.12; I^2=81.2%). The certainty of evidence was very low for quality of life, motor, and psychiatric symptoms and low for cognitive function and overall nonmotor symptoms. Improvements in motor symptoms and cognitive function remained stable at follow-up. Meta-regression analysis indicated that age, percentage of female participants, and supervision mode were possible sources of heterogeneity. Overall, 94 studies reported reach (median 37.5%), 38 reported fidelity (95.7%), and 105 reported dropout rates (9.1%).

Conclusions: In contrast to previous reviews focused on single technologies or outcomes, this review provided the first comprehensive synthesis across all DHI types on multiple outcomes and indicated their potential as nonpharmacological interventions for PD management. However, current evidence is of low to very low certainty, and wide 95% PIs, together with

XSL•FO
RenderX

high risk of bias and substantial heterogeneity, indicate considerable uncertainty regarding the true effect in future implementations. Therefore, findings should be interpreted with caution. These findings provide integrated evidence to guide the design and prioritization of future research. The results have important real-world implications, supporting cautious implementation while underscoring the need for more robust trials, particularly in resource-limited settings.

Trial Registration: PROSPERO CRD42023492123; https://www.crd.york.ac.uk/PROSPERO/view/CRD42023492123

(J Med Internet Res 2026;28:e79935) doi:10.2196/79935

KEYWORDS

digital; eHealth; meta-analysis; motor symptoms; non-motor symptoms; Parkinson disease; Parkinson's disease; quality of life

Introduction

Parkinson disease (PD) is the most common serious movement disorder, characterized clinically by bradykinesia, rest tremor, rigidity, and postural and gait abnormalities [1,2]. As the global population ages, the global prevalence of PD is projected to exceed 12 million by 2040 [3,4], representing a significant public health challenge. In addition to motor symptoms, nearly all patients experience complex nonmotor manifestations, including hyposmia, autonomic dysfunction, psychiatric symptoms, cognitive decline, and sleep disorders [1,5,6], which also impair the quality of life of patients with PD [7]. PD management mainly relies on dopamine replacement therapy and incorporates a range of medical and surgical treatments for complex symptoms [5].

However, PD management still faces many obstacles that must be overcome. It remains largely limited to symptomatic treatment and is challenged by long-term motor complications as well as many nonmotor symptoms, which increase the complexity of medication regimens. Crucially, there are still no clinically established disease-modifying therapies to slow or halt neurodegeneration [2,8]. Consequently, long-term, individualized pharmacological optimization and multidisciplinary care are essential for people with PD; yet, delivering sustained, high-quality care in routine outpatient and community settings remains challenging. This challenge may underscore the potential value of digital health approaches.

In addition, nonpharmacological interventions such as physical and cognitive training have shown promise for improving both motor and nonmotor symptoms of PD [5,9], but access to these interventions can be challenging for patients. Patients often need to visit medical institutions in person to receive rehabilitation training, which requires sufficient time and financial resources. Furthermore, the disease may make it difficult or cause reluctance for patients to go out. Moreover, in some economically underdeveloped regions, few medical institutions offer such rehabilitation training. In addition, patient adherence also poses a significant challenge. The combination of high prevalence, multifaceted symptomatology, and absence of disease-modifying treatments makes PD management particularly challenging; therefore, strategies that improve symptoms and quality of life are essential to reduce the burden on patients and health care systems.

Digital technology is now revolutionizing PD management, creating new possibilities across the entire spectrum of patient care, from disease identification and diagnosis [10,11] to treatment and prognosis [12,13]. It offers opportunities to enhance and extend nonpharmacological interventions [14,15]. Emerging evidence highlights the potential of digital health interventions (DHIs) in the long-term management of PD [16]. DHIs, defined as the “discrete functionality of digital technology that is applied to achieve health objectives” [17-19], may be especially valuable for long-term PD management.

A growing number of randomized controlled trials (RCTs) have demonstrated that various types of DHIs, including computer-based cognitive training [20] and telerehabilitation [21], may benefit patients with PD. These digital interventions use diverse digital tools, spanning from wearable sensors [22] to smartphone apps [23], to improve both motor and nonmotor outcomes. DHIs offer several potential advantages (1) they may reduce the frequency of in-person visits and thus relieve pressure on medical services and regional resource imbalances in medical resources through information and communication technology [24], (2) they may augment and enhance conventional rehabilitation and pharmacotherapy to improve engagement and outcomes [25], and (3) they may address patients’ individualized demands in real time by leveraging adaptive systems that analyze real-time health data [26].

Current DHIs for PD management include robot-assisted physical therapy [27], exergaming [28], videoconferencing [29], smartphone apps [30], wearable sensors [31], and others. Although these studies demonstrate the considerable potential of DHIs in PD care, no study has summarized current evidence systematically and comprehensively in this field, as previous reviews primarily focused on single digital intervention types (eg, virtual reality) [32] or specific outcomes (eg, motor symptoms) [33]. A comprehensive synthesis across DHI modalities and health-related outcomes in PD is still lacking.

To fill this gap, we performed a large-scale, updated systematic review and meta-analysis to evaluate the effectiveness of diverse DHIs across multiple health outcomes in patients with PD. We additionally summarized the reported reach, fidelity, and feasibility of these interventions. Our findings aim to inform clinicians in selecting management strategies for PD and to guide priorities for future research.

Methods

Protocol and Registration

This systematic review and meta-analysis were preregistered on PROSPERO (CRD42023492123) and were conducted according to the PRISMA (Preferred Reporting Items for

XSL•FO
RenderX

Systematic Reviews and Meta-Analyses) guidelines for transparent and comprehensive reporting of methodology and results [34]. The PRISMA 2020 checklist and the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) checklist are provided in Multimedia Appendices 1 and 2, respectively. Any deviations from the preregistered protocol are outlined and explained in Multimedia Appendix 3.

Search Strategy and Study Selection

Searches were first performed in PubMed (via NCBI), Ovid Embase (via Ovid), Web of Science (via Clarivate), CINAHL (via EBSCOhost), Cochrane Central Register of Controlled Trials (via Wiley), and APA PsycINFO (via EBSCOhost) from inception to November 21, 2023. Updated database searches were performed in the above databases via the same platforms from November 21, 2023, to January 5, 2025. Final database searches were conducted on November 22, 2025, for articles published in 2025, and email alerts were maintained to capture newly published studies until final data analysis. No study registries, websites, or other non–peer-reviewed online resources were systematically searched to ensure the reliability of the included data. The reference lists of included articles and other relevant systematic reviews were examined. Authors of potentially eligible studies were contacted via email to inquire about additional data or unpublished results.

Three groups of search terms were used: terms related to PD, terms related to DHIs (eg, digital, technology, telemedicine, internet, robotics, virtual reality, and computers), and terms related to RCTs. Complete search strategies and search limits are detailed in Multimedia Appendix 4. Search results were imported into EndNote 20 (Clarivate), and duplicates were removed using the software's automatic deduplication feature, followed by manual review. The titles and abstracts of the involved studies were first screened, and then the full texts of potentially eligible studies were reviewed according to the inclusion criteria.

The inclusion criteria were developed based on the PICOS (Population, Intervention, Comparison, Outcomes, and Study Type) framework. First, for population, we included studies involving patients with a confirmed diagnosis of PD. Second, for intervention, interventions provided through the computer, smartphone, tablet, virtual reality, wearable sensor, robot, or any other digital technology for physical rehabilitation, care aid, or cognitive training were defined as DHIs in this review. DHIs were eligible regardless of setting (eg, used in medical institution vs at home), mode of interaction and supervision, intervention content, or intervention duration. Although digital technology also contributes to the application of deep brain stimulation, transcranial direct current stimulation, and repetitive transcranial magnetic stimulation, these interventions with specific therapeutic aims were not in the scope of this review. DHIs solely used for monitoring parameters related to the disease symptoms were not included. Third, for comparison, comparators were categorized into passive control and active control. Passive control refers to waitlist control or no intervention. Active control included traditional, nontechnological caregiving methods, such as conventional physical rehabilitation, standard care aid, or drug treatments. Studies comparing the effectiveness of different modes of DHIs were excluded. Fourth, for outcome, we considered any outcomes related to motor symptoms, nonmotor symptoms, and quality of life in patients with PD as eligible to provide a comprehensive summary. Studies that only reported the reach, fidelity, or feasibility of the DHIs in PD were also included. Fifth, for study type, only RCT studies were included in this review. Studies with fewer than 5 participants in each group were excluded from the quantitative meta-analysis to ensure the robustness of the results.

The entire process of study selection was performed by 2 researchers independently, and disagreements were resolved with the involvement of a third researcher. Interrater reliability was assessed using Cohen κ [35], which was 0.831 for full-text review, indicating a high agreement between the 2 researchers.

Data Extraction

One reviewer extracted data using a comprehensive data extraction form: publication and author details (title, year, first author's name, and country), study details (study design, study setting, sample size, dropout, and follow-up duration), participants' characteristics (age, sex, and disease severity), intervention information (type of intervention, purpose of the intervention, intervention duration, and supervision mode), comparison information (type of comparator), and outcomes (scales used for assessing outcomes, reach, fidelity, and feasibility of interventions). The extracted data were checked by a second reviewer. Discrepancies in data extraction between the 2 researchers were resolved through discussion with a third reviewer.

Types of DHIs

DHIs were classified into three categories: (1) digital databases, (2) online classes, and (3) technology-based rehabilitation devices, based on the interaction modalities influenced by the technical implementation and the level of supervision inherent to the intervention, following the classification framework recommended by a previous review [36].

Digital databases are platforms that store health- or disease-related information in the form of videos or textual materials and offer self-paced access to predesigned resources, characterized by asynchronous interaction and a minimal level of supervision. Online classes involve health management and telerehabilitation sessions delivered in real time through telephone or video conferencing, offering synchronous interaction and direct supervision by professionals remotely and allowing for immediate feedback and guidance. Technology-based rehabilitation devices encompass a broad range of technologies, including exergaming, virtual reality, robotics, and mobile apps, which support physical rehabilitation or cognitive training. These devices involve close interactions with users, who respond to guidance and instructions provided by the DHIs, which in turn provide cues and feedback based on the users' actions. The level of supervision in these interventions varies depending on the technical features of the devices (eg, sensor integration) and the study design (eg, use under professional supervision).

XSL•FO
RenderX

Additionally, DHIs were categorized into physical rehabilitation, cognitive training, and care aid according to the intervention purpose.

Outcome Measures

The primary outcomes of the review were changes in symptom severity, including motor symptoms, cognitive function, psychiatric symptoms, overall nonmotor symptoms, and quality of life from baseline to postintervention and last follow-up. The scales used for assessing the severity of each symptom and quality of life depended on the original study. When one study utilized multiple assessment scales for one type of syndrome (eg, Parkinson's Disease Questionnaire-8 and Parkinson's Disease Questionnaire-39 for quality of life), the most frequently used one among all included studies was selected for analysis to reduce potential heterogeneity between studies. The assessment scales for each outcome used in each study are summarized in Multimedia Appendix 5.

The secondary outcomes included percentage reach, fidelity, dropout rate, and feasibility. According to the Medical Research Council process evaluation framework and Proctor implementation outcomes, we used the more common term reach to replace penetration, which is defined as the integration of a practice within a service setting and its subsystems [37,38]; therefore, we calculated the proportion of people who come into contact with the intervention. Fidelity was defined as the degree to which an intervention was implemented as prescribed in the original protocol or as intended by the program developers [38], and we extracted the fulfillment of the intervention. Feasibility was defined as the extent to which a new treatment or innovation can be successfully used or carried out within a given agency or setting [38] and can be measured by retention, acceptability, adherence, satisfaction, preference, safety, and cost-effectiveness [39]. Feasibility was summarized according to the definition of each study, and the aforementioned items were extracted even if the study did not discuss feasibility.

Quality Assessment

The revised Cochrane Risk of Bias 2 tool [40] was used to assess study quality across the following domains: randomization process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported result. Studies were categorized as low risk, some concerns, or high risk. We used the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) criteria [41] to assess the certainty of the overall evidence, which incorporates 5 key considerations: study limitations, inconsistency of effects, indirectness, imprecision, and publication bias. Any deviation across these 5 domains resulted in downgrading of the quality of evidence. The overall certainty of evidence was classified as high, moderate, low, or very low. All quality assessments were independently conducted by 2 reviewers, and any discrepancies were resolved through discussion.

Meta-Analysis Methods

Pairwise meta-analyses were performed to evaluate the effect of DHIs on the primary outcomes when more than 5 effect estimates were available in the analysis. The means and SDs of within-group differences from baseline to postintervention and follow-up assessments were used. If means and SDs were not reported in a study, the corresponding author was contacted via email to obtain the data. Otherwise, means and SDs were calculated based on available data using recommended formulas [42,43].

Standardized mean differences (SMDs; Hedges g) at postintervention and the last follow-up assessment were used as the effect measure, allowing comparison of data from different assessment scales. In studies with more than 2 intervention groups, the intervention group demonstrating the most favorable effect was selected for the main analysis as it represented the most well-designed DHIs. In studies featuring multiple control groups, the most active control group was chosen to adopt a more conservative analytical approach.

Considering that different assessment scales may reflect symptom severity in opposite directions, data from the original studies for the same outcome were standardized based on the most commonly used scale. For instance, while higher scores on the Unified Parkinson's Disease Rating Scale Part 3 indicate more severe motor symptoms, higher scores on the 6-Minute Walk Test suggest milder symptoms. Therefore, the results of the 6-Minute Walk Test were inverted (multiplied by –1) to align with the Unified Parkinson's Disease Rating Scale Part 3 for the analysis.

Due to the presence of potential heterogeneity, the random-effects model and the DerSimonian-Laird estimator were applied to calculate pooled estimates across studies. The Hartung-Knapp-Sidik-Jonkman method was applied to increase the robustness of the results [44]. Forest plots were generated to visually display the effect size and 95% CIs for individual studies and the overall pooled estimate. To present the expected range of true effects in similar studies, the 95% prediction intervals (PIs) were also calculated and displayed in the forest plot [45]. When the number of studies was fewer than 10, the method proposed by Nagashima et al [46] was applied to adjust the 95% PI. The 95% CI reflects the precision of the estimated overall average effect, whereas the 95% PI estimates the distribution of effects within which the true effect size is expected to fall in future similar studies. These 2 intervals complement each other to provide comprehensive information for result interpretation.

Subgroup analyses were conducted based on the intervention type (technology-based rehabilitation devices, online classes, and digital databases), considering the different inherent characteristics of various DHIs. Pairwise comparisons between subgroups were performed using random-effects meta-analysis with Hartung-Knapp adjustment and were limited to subgroups with 3 or more studies. Funnel plots and Egger regression intercept were used to assess small-study effects, which may indicate the potential publication bias. A sensitivity analysis was conducted to evaluate whether the overall results were statistically and significantly affected by 1 individual study using the leave-one-out analysis method. In addition, another sensitivity analysis restricting the synthesis to studies with a low risk of bias was performed for outcomes for which at least 10 such studies were available. All meta-analyses were

XSL•FO
RenderX

conducted using R (version 4.2.3; R Foundation for Statistical Computing).

Heterogeneity and Meta-Regression

The presence of heterogeneity among studies was assessed using Cochran Q test and the I^2 statistic, whereas 95% PI, τ, and τ^2 were calculated to evaluate the magnitude of heterogeneity [45]. By estimating the expected range of true effects for future studies, the 95% PI directly evaluates the variability of the intervention effect across diverse settings [47]. τ represents the estimated SD of the between-study effects, and τ^2 denotes the variance, which quantifies the spread of the true study effects around the mean [48].

Because substantial heterogeneity was anticipated to result from various types of DHIs in the analysis, univariate meta-regression analyses were performed using the mixed-effects model with the DerSimonian-Laird estimator for between-study variance and the Hartung-Knapp-Sidik-Jonkman method for SE adjustment to identify potential moderators that could account for heterogeneity. These potential moderators included individual-level variables (mean age, sex, and disease severity) and study-level variables (publication years, country income level, intervention type, intervention purpose, intervention setting, type of control, intervention duration, and supervision mode). In addition, bubble plots were provided as visual representation of the meta-regressions.

Results

Study Search and Selection

The search initially yielded 30,918 related study records from 6 databases in November 2023, among which 12,909 duplicate records were first removed during screening. The titles and abstracts of the remaining 18,009 records were screened, and the full texts of 284 studies were further assessed. A total of 90 studies from 91 articles evaluating the effect of DHI in patients with PD were included.

A supplementary search was conducted on January 5, 2025, and 4179 new publications were identified using the same keywords and databases as the initial search. After removing duplicates, a total of 3019 unique records were screened based on title and abstract, and the full texts of 66 potentially eligible studies were then reviewed. A total of 11 studies meeting the inclusion criteria were included.

The final search was performed on November 22, 2025, to ensure the inclusion of the most recent evidence and identified 3980 new publications. Following deduplication, a total of 2481 unique records were screened by title and abstract. Subsequently, a total of 44 studies underwent full-text review, resulting in the inclusion of 11 studies that met the eligibility criteria. During the full-text screening process, no studies were identified that appeared to meet the inclusion criteria but were ultimately excluded; all exclusions were straightforward decisions based on eligibility criteria.

Ultimately, a total of 112 unique studies from 115 articles [27,30,31,49-158] were eligible for the systematic review, and 97 unique studies from 98 articles were included in the quantitative meta-analysis. Of these 112 studies, 73 evaluated motor symptoms, 8 evaluated overall nonmotor symptoms, 26 evaluated cognitive function, 30 evaluated psychiatric symptoms, and 41 evaluated quality of life. The PRISMA flow diagram of the study search and selection is presented in Figure 1.

Figure 1. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for study selection.

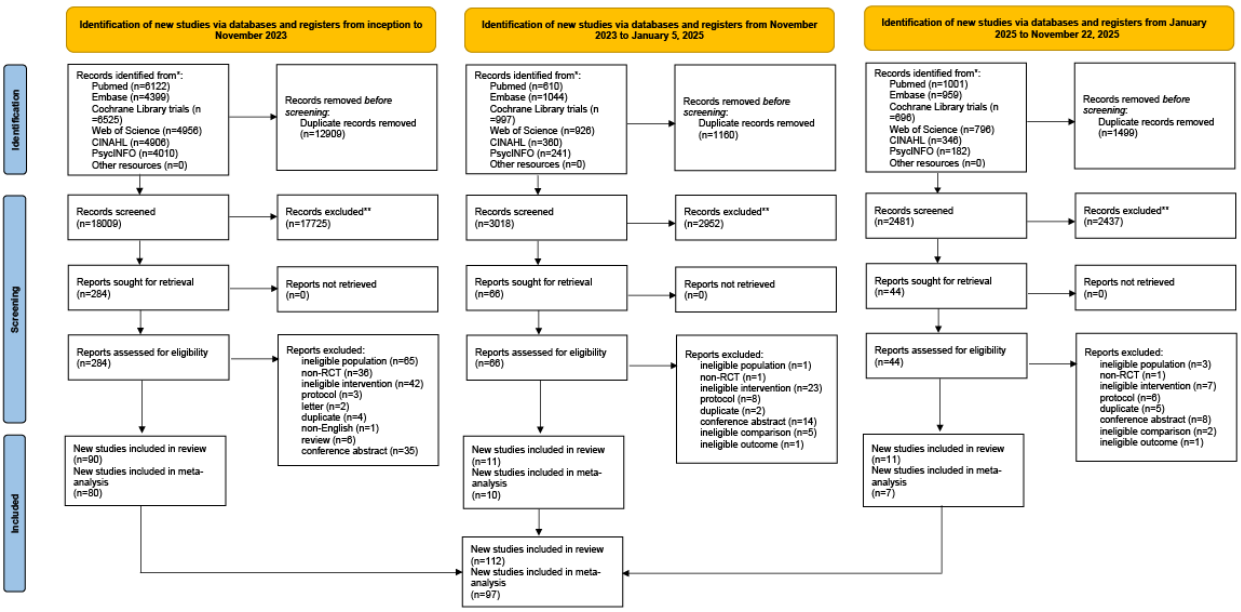

XSL•FO
RenderX

Study Characteristics

A total of 112 included studies comprising 5594 participants were identified, among which 84 studies (4658 participants) were performed in high-income countries, whereas only 28 studies (936 participants) were performed in low- and middle-income countries (LMICs). According to the predetermined classification criteria, the DHIs used in 91 studies were categorized as technology-based rehabilitation devices, 17 as online classes, and 8 as digital databases. The intervention purposes of 74 studies were physical rehabilitation, 28 were cognitive training, and 17 were care aid.

Active comparators (eg, traditional and nontechnological caregiving methods) were used in 91 trials, whereas others used passive comparators (eg, waitlist control or no intervention). The study setting of 51 studies was at home, and that of 55 studies was in medical institutions. Participants in the study by Halpern et al [88] first received intervention in the clinic for 9 sessions and then at home.

Regarding delivery mode, most interventions were delivered through computers and virtual reality devices; other modes included wearable sensors, robots, tablets, smartphones, and others. Most interventions lasted for at least 4 weeks, whereas 4 did not [51,54,92,115]. Among the 39 (34.8%) studies that reported follow-up assessments, the follow-up duration ranged from 4 to 24 weeks in 38 studies, except for the study by Yang et al [143], which implemented a shorter 2-week follow-up period.

A total of 97 of 112 studies were included in the meta-analysis, with the remaining 15 retained in the systematic review because of unavailable outcome data. A comprehensive summary of the characteristics of the included studies is presented in Multimedia Appendix 6.

Quality of Studies

The methodological limitations assessed by the Cochrane Risk of Bias 2 tool revealed that ≥20% of the included studies exhibited some concerns or high risk in the following domains (Figure 2): randomization process (postintervention: 53/180, 29% and follow-up: 11/55, 20%), deviations from intended interventions (postintervention: 95/180, 53% and follow-up: 19/55, 34%), missing outcome data (postintervention: 46/180, 26% and follow-up: 13/55, 24%), and selection of reported results (postintervention: 96/180, 53% and follow-up: 33/55, 60%). The main sources of bias were inadequate reporting of randomization methods (eg, lack of explicit allocation concealment) and failure to blind participants or caregivers in trials. Regarding motor symptoms, 12 studies were assessed as having a low risk of bias, 43 studies as having some concerns, and 18 studies as having a high risk of bias. Regarding cognitive function, 6 studies were assessed as having a low risk of bias, 16 studies as having some concerns, and 4 studies as having a high risk of bias. Regarding psychiatric symptoms, 4 studies were assessed as having a low risk of bias, 15 studies as having some concerns, and 10 studies as having a high risk of bias. Regarding overall nonmotor symptoms, 5 studies were assessed as having some concerns and 3 as having a high risk of bias. Regarding quality of life, 5 studies were assessed as having a low risk of bias, 22 studies as having some concerns, and 14 studies as having a high risk of bias (Multimedia Appendix 7).

XSL•FO
RenderX

Figure 2. Risk of bias across effect estimates at postintervention and follow-up assessments based on the Cochrane Risk of Bias 2 tool for randomized trials.

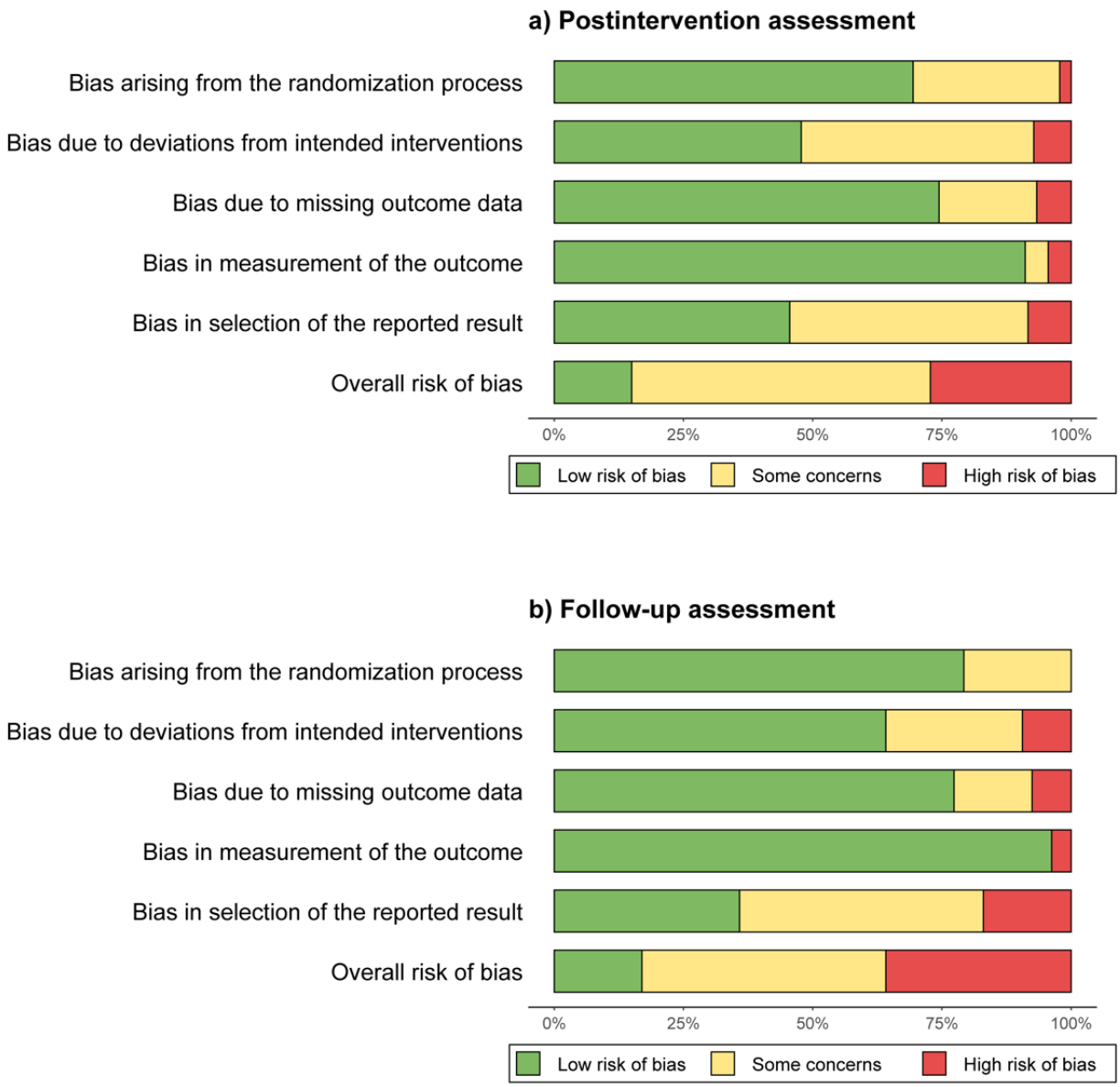

Effect of DHIs on PD Across Different Symptoms

Motor Symptoms

DHIs significantly improve motor symptoms (SMD=–0.39, 95% CI –0.60 to –0.18), while the 95% PI (–1.75 to 0.99) indicates that the true effect is uncertain and could range from a substantial benefit to being inferior to the control condition in future implementations (Figure 3 [27,31,49-53,56-59,61,62,65,68,69,71,72,74-79,82,84-87,89-97,99,100,103,106,110-113,115-117,119,120,122-124,126-132,135,136,141-146,151-153,156]). Although the intervention group had fewer symptoms on average (SMD=–0.39), the PI suggests considerable variability in future outcomes. DHIs may improve symptoms (SMD=–1.75) in some populations, while in other populations DHIs may be less effective than control groups (SMD=0.99). Also, DHIs may not affect other populations since the PI contains the 0 value. However, only 1 of 73 included studies suggested that DHIs were inferior to standard training, while both led to improvements in motor symptoms [89]. Subgroup analyses revealed no statistically significant differences in pairwise comparisons among the 3 intervention types (Multimedia Appendix 8). Only technology-based rehabilitation devices significantly improve motor symptoms (SMD=–0.36, 95% CI –0.59 to –0.13, 95% PI –1.80 to 1.07). Neither online classes nor digital databases exhibited favorable effects, with SMD=–0.19 (95% CI –0.87 to 0.48) for online classes, and SMD=–0.76 (95% CI –1.85 to 0.33) for digital databases. However, the subgroup analysis revealed no statistically significant difference in effects across intervention types (P=.39). Significant heterogeneity was found (I^2=80.3%, P<.001), and the heterogeneity was substantial (τ=0.680 and

XSL•FO
RenderX

τ^2=0.4623). The intervention types did not completely explain the significant heterogeneity, with I^2=81.3% for technology-based rehabilitation devices (P<.001) and I^2=83% for digital databases (P<.001). In the follow-up analysis, DHIs also exhibited significant improvement in the motor symptoms of patients with PD (Multimedia Appendix 9). The leave-one-out analysis obtained a consistent result. The sensitivity analysis of 12 low-risk-of-bias studies showed a significant but modest benefit of DHIs on motor symptoms (SMD=–0.22, 95% CI –0.40 to –0.04), with negligible heterogeneity (I^2=0.6%; τ=0.025; τ^2=0.0006; 95% PI –0.43 to –0.01). The 95% PI indicated a high certainty that the intervention will yield at least a small beneficial effect in future similar study settings. GRADE ratings indicated that the certainty of the evidence for the effect of DHIs on motor symptoms of patients with PD was very low owing to the high risk of bias of included studies, high heterogeneity, and potential publication bias (Figure 4).

Figure 3. Forest plot of the meta-analysis of digital health interventions' efficacy on postintervention motor symptoms [27,31,49-53,56-59,61,62,65,68,69,71,72,74-79,82,84-87,89-97,99,100,103,106,110-113,115-117,119,120,122-124,126-132,135,136,141-146,151-153,156]. Standardized mean differences (SMDs) with 95% CIs were calculated using a random-effects model. Negative SMD values favor the experimental intervention. Subgroup analyses were performed by intervention type (technology-based rehabilitation devices, online classes, digital databases). The square represents the individual study estimate. The rhombus shape represents the pooled estimates of the lowest accuracy in all studies. P values are shown as p.

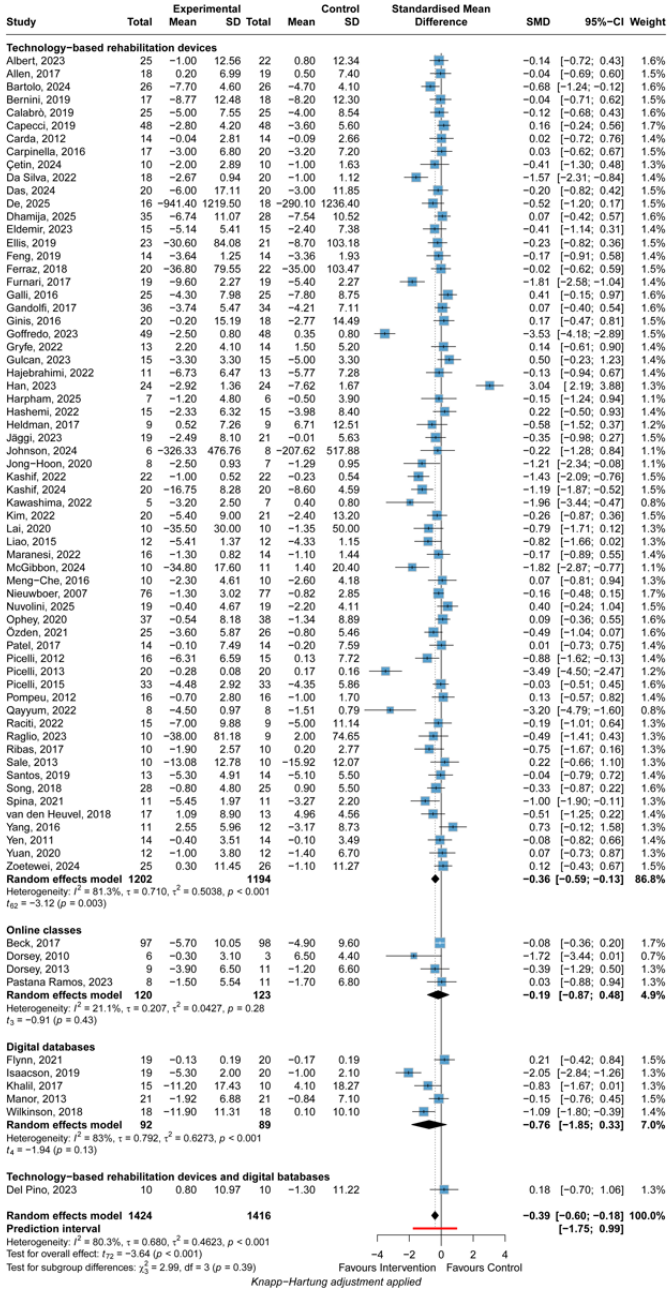

XSL•FO
RenderX

Figure 4. GRADE (Grading of Recommendations Assessment, Development, and Evaluation) ratings at postintervention assessments.

Digital health intervention compared to convetional intervention/usual care/waitlist group for people with Parkinson's disease
Bibliography:

Certainty assessment							Summary of findings				
							Study event rates (%)			Anticipated absolute effects	
Participants (studies) Follow-up	Risk of bias	Inconsistency	Indirectness	Imprecision	Publication bias	Overall certainty of evidence	With convetional intervention/usual care/waitlist group	With digital health intervention	Relative effect (95% CI)	Risk with convetional intervention/usual care/waitlist group	Risk difference with digital health intervention
Motor symptoms											
2840 (73 RCTs)	serious[a]	very serious[b]	not serious	not serious	publication bias strongly suspected[c]	⊕◯◯◯ Very low[a,b,c]	1416	1424	-	-	SMD **0.39 SD lower** (0.6 lower to 0.18 lower)
Psychiatric symptoms											
1741 (30 RCTs)	serious[d]	very serious[e]	not serious	not serious	none	⊕◯◯◯ Very low[d,e]	871	870	-	-	SMD **0.42 SD lower** (0.74 lower to 0.09 lower)
Cognitive function											
1148 (26 RCTs)	not serious	serious[f]	not serious	not serious	publication bias strongly suspected[g]	⊕⊕◯◯ Low[f,g]	568	580	-	-	SMD **0.47 SD higher** (0.22 higher to 0.72 higher)
Overall Non-motor symptoms											
572 (8 RCTs)	serious[h]	not serious	not serious	serious[i]	none	⊕⊕◯◯ Low[h,i]	294	278	-	-	SMD **0.26 SD lower** (0.49 lower to 0.03 higher)
Quality of life											
1929 (41 RCTs)	serious[j]	very serious[k]	not serious	serious[l]	none	⊕◯◯◯ Very low[j,k,l]	970	959	-	-	SMD **0.19 SD lower** (0.47 lower to 0.09 higher)

CI: confidence interval; **SMD:** standardised mean difference

Explanations

a. 18 of 73 studys are rated at overall high risk with some limitations for multiple criteria.
b. I^2 = 80.3%, τ^2 = 0.462, P<0.001, 95% prediction interval [-1.75, 0.99]
c. Egger's test indicated potential small-study effect. (P = 0.013)
d. 9 of 30 studys are rated at overall high risk with some limitations for multiple criteria.
e. I^2 = 85.4%, τ^2 = 0.451 P<0.001, 95% prediction interval [-1.82, 0.99]
f. I^2 = 63.5%, τ^2 = 0.172, P<0.001, 95% prediction interval [-0.41, 1.35]
g. Egger's test indicated potential small-study effect. (P = 0.002)
h. 3 of 8 studys are rated at overall high risk, with only one study being rated at overall low risk.
i. Small sample size.
j. 14 of 41 studys are rated at overall high risk with some limitations for multiple criteria.
k. I2 = 81.2%, τ^2 = 0.405, p <0.001, 95% prediction interval [-1.50; 1.12]
l. 95% Confidence Interval crosses null effect

Overall Nonmotor Symptoms

DHIs significantly improved overall nonmotor symptoms (SMD=–0.26, 95% CI –0.49 to –0.03, 95% PI –0.56 to 0.03; I^2=0; Figure 5 [30,52,65,91,100,120,134,152]). No significant heterogeneity was observed for the analysis (I^2=13.8%; P=.32). The leave-one-out analysis did not obtain a consistent result after omitting the studies by Dhamija et al [152], and So et al [134], suggesting the instability of this finding. GRADE ratings indicated that the certainty of the evidence for the effect of DHIs on overall nonmotor symptoms of patients with PD was low owing to the high risk of bias of included studies and serious imprecision (Figure 4).

XSL•FO
RenderX

Figure 5. Forest plot of the meta-analysis of digital health interventions' efficacy on postintervention overall nonmotor symptoms [30,52,65,91,100,120,134,152]. Standardized mean differences (SMDs) with 95% CIs were calculated using a random-effects model. Negative SMD values favor the experimental intervention. Subgroup analyses were performed by intervention type (technology-based rehabilitation devices, online classes, digital databases). The square represents the individual study estimate. The rhombus shape represents the pooled estimates of the lowest accuracy in all studies. *P* values are shown as p.

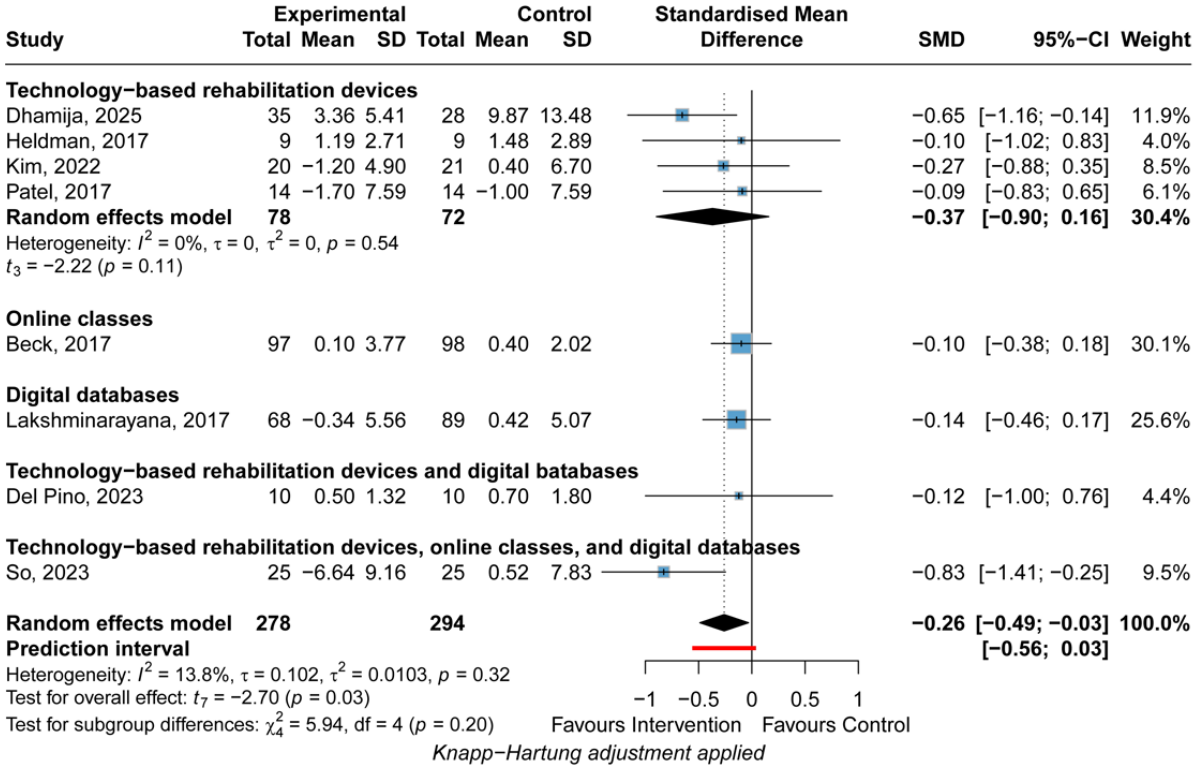

Psychiatric Symptoms

DHIs significantly improved psychiatric symptoms (SMD=–0.42, 95% CI –0.74 to –0.09), while the 95% PI (–1.82 to 0.99) indicates that the true effect is uncertain and could range from a substantial benefit to being inferior to the control condition in future implementations (Figure 6 [30,52,53,55,57,63,66-68,73,75,77,82,85,87,102,107,109,111,117,118,120,125,137,139,141,142,145,157,158]). Although the intervention group had fewer symptoms on average (SMD=–0.42), the PI suggests considerable variability in future outcomes. DHIs may improve symptoms (SMD=–1.82) in some populations, while in other populations DHIs may be less effective than control groups (SMD=0.99). Also, DHIs may not affect other populations since the PI contains the 0 value. However, only 1 of 30 included studies suggested that DHIs were significantly harmful to psychiatric symptoms. Subgroup analyses revealed no statistically significant differences in pairwise comparisons among the intervention types (Multimedia Appendix 8). No types of DHIs exhibited significantly favorable effects in the subgroup analysis, with SMD=–0.49 (95% CI –0.98 to 0.01) for technology-based rehabilitation devices, SMD=–0.05 (95% CI –1.89 to 1.79) for digital databases, and SMD=–0.42 (95% CI –0.98 to 0.13) for online classes. Significant heterogeneity was found (I^2=85.4%; P<.001), and the heterogeneity was substantial (τ=0.671; τ^2=0.4506; 95% PI –1.82 to 0.99). The intervention types could not completely explain the significant heterogeneity, with I^2=88.1% for technology-based rehabilitation devices (P<.001) and I^2=80% for online classes (P<.001). In the follow-up analysis, DHIs did not exhibit significant improvement in the psychiatric symptoms of patients with PD (Multimedia Appendix 9). The leave-one-out analysis obtained a consistent result. GRADE ratings indicated that the certainty of the evidence for the effect of DHIs on psychiatric symptoms of patients with PD was very low owing to the high risk of bias of included studies and high heterogeneity (Figure 4).

XSL•FO
RenderX

Figure 6. Forest plot of the meta-analysis of digital health interventions' efficacy on postintervention psychiatric symptoms [30,52,53,55,57,63,66-68,73,75,77,82,85,87,102,107,109,111,117,118,120,125,137,139,141,142,145,157,158]. Standardized mean differences (SMDs) with 95% CIs were calculated using a random-effects model. Negative SMD values favor the experimental intervention. Subgroup analyses were performed by intervention type (technology-based rehabilitation devices, online classes, digital databases). The square represents the individual study estimate. The rhombus shape represents the pooled estimates of the lowest accuracy in all studies. *P* values are shown as p.

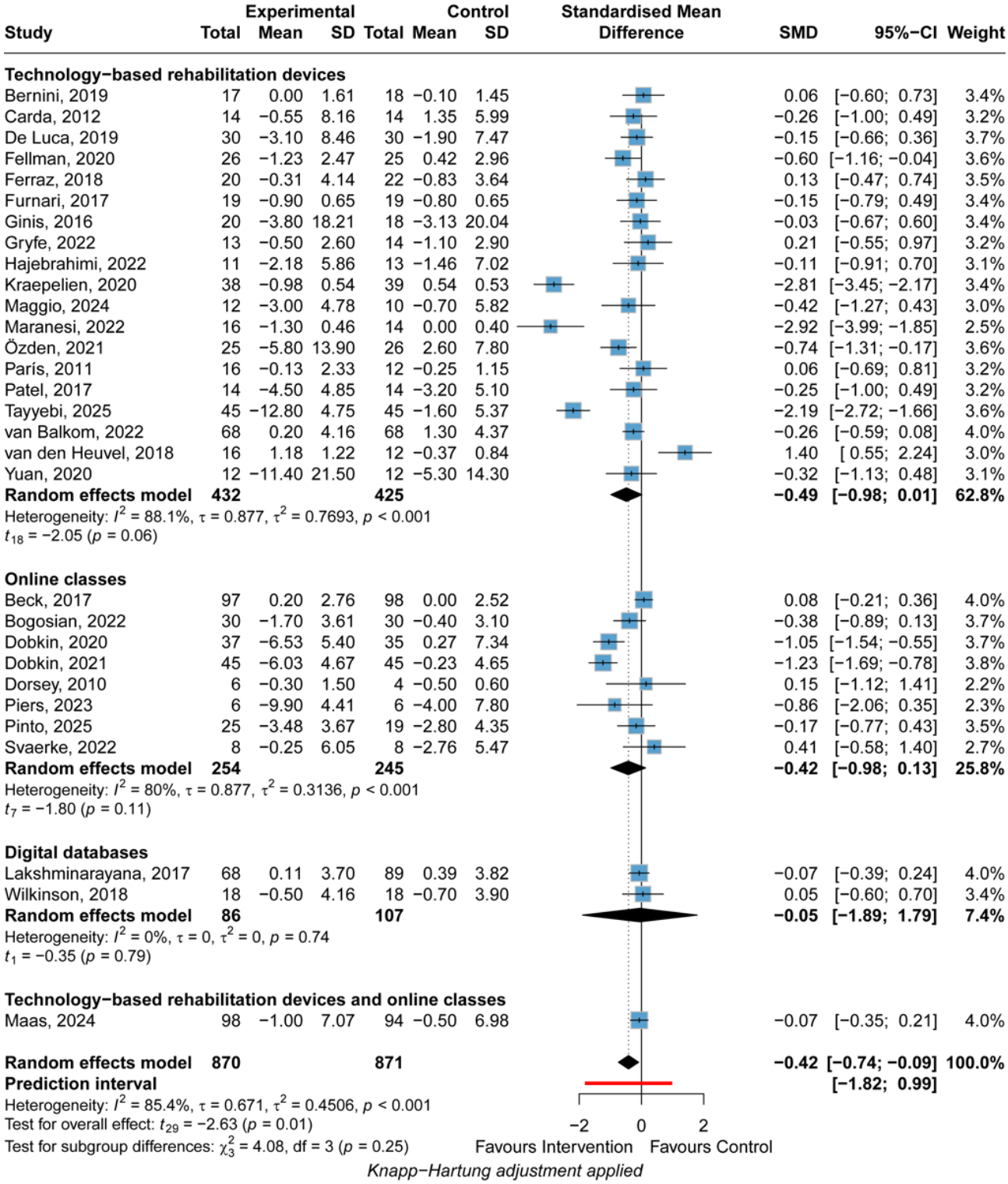

Cognitive Function

DHIs significantly improve cognitive function (SMD=0.47, 95% CI 0.22 to 0.72), while the 95% PI (–0.41 to 1.35) indicates that the true effect is uncertain and could range from a substantial benefit to being inferior to the control condition in future implementations (Figure 7 [50,52-54,59,62,65,68,73,80,82,85,87,92,108,109,112,116,118,126,135,137,139,140,155,156]). Although the intervention group had fewer symptoms on average (SMD=0.47), the PI suggests considerable variability in future outcomes. DHIs may improve symptoms (SMD=1.35) in some populations, while in other populations DHIs may be less effective than control groups (SMD=–0.41). Also, DHIs may

XSL•FO
RenderX

not affect other populations since the PI contains the 0 value. However, no included studies suggested that DHIs were significantly harmful to cognitive function. Technology-based rehabilitation devices demonstrated significantly greater improvement in cognitive function compared with online classes (SMD=0.39, 95% CI 0.01 to 0.77; *P*=.049; Multimedia Appendix 8). Technology-based rehabilitation devices exhibited significant improvement in cognitive functions of patients with PD (SMD=0.53, 95% CI 0.24 to 0.82), but not online classes (SMD=0.14, 95% CI –0.44 to 0.72). Significant heterogeneity was found (I^2=63.5%; *P*<.001), and the heterogeneity was substantial (τ=0.415 and τ^2=0.1723). The intervention types could not explain the significant heterogeneity, with I^2=68.2% for technology-based rehabilitation devices (*P*<.001). In the follow-up assessment, DHIs also exhibited significant improvement in the cognitive function of patients with PD (Multimedia Appendix 9). The leave-one-out analysis obtained a consistent result. GRADE ratings indicated that the certainty of the evidence for the effect of DHIs on the cognitive function of patients with PD was low owing to high heterogeneity and potential publication bias (Figure 4).

Figure 7. Forest plot of the meta-analysis of digital health interventions' efficacy on postintervention cognitive function [50,52-54,59,62,65,68,73,80,82,85,87,92,108,109,112,116,118,126,135,137,139,140,155,156]. Standardized mean differences (SMDs) with 95% CIs were calculated using a random-effects model. Negative SMD values favor the experimental intervention. Subgroup analyses were performed by intervention type (technology-based rehabilitation devices, online classes, digital databases). The square represents the individual study estimate. The rhombus shape represents the pooled estimates of the lowest accuracy in all studies. *P* values are shown as p.

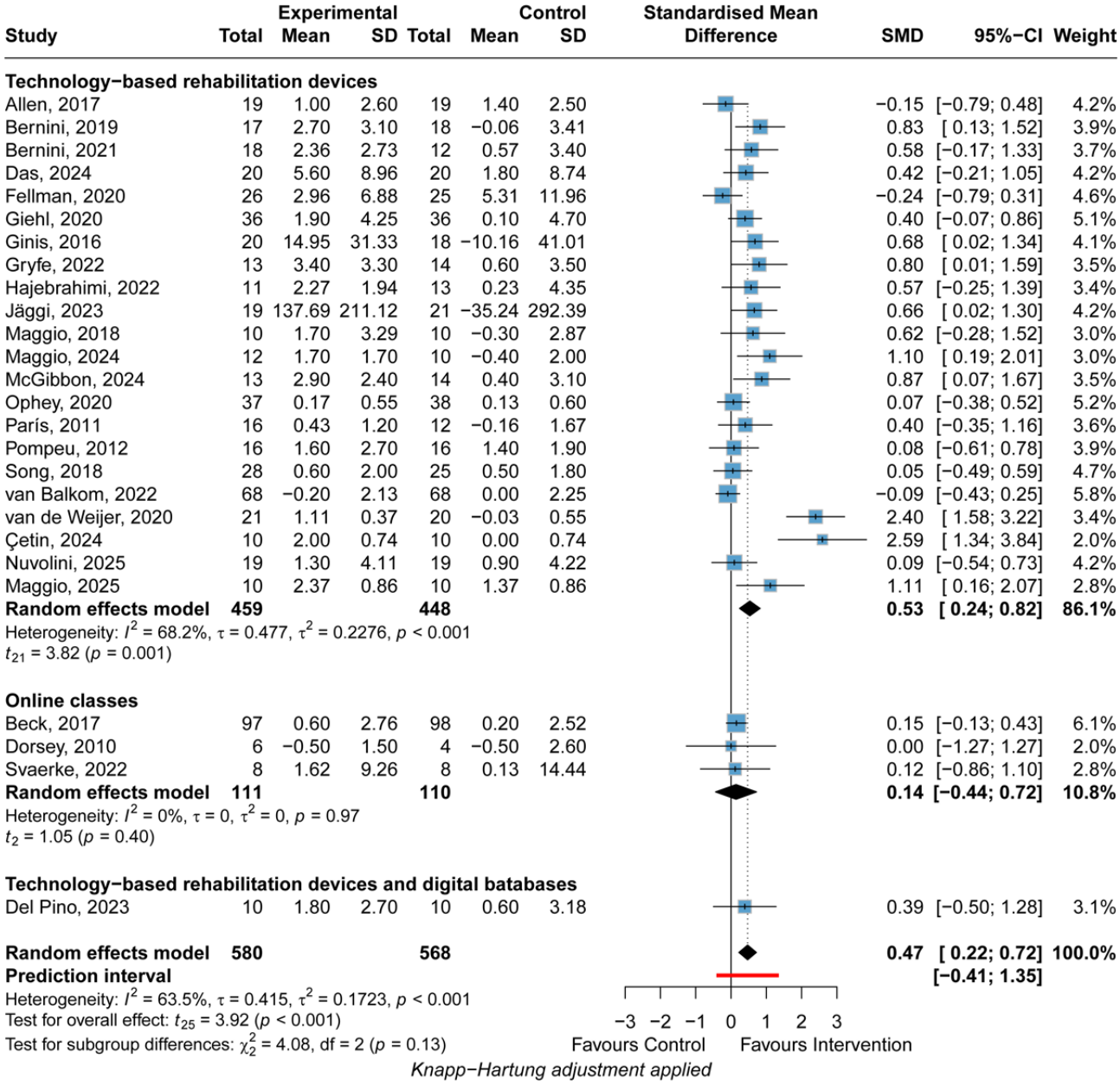

XSL•FO
RenderX

Quality of Life

DHIs did not improve quality of life in patients with PD (SMD=–0.19, 95% CI –0.47 to 0.09), and the 95% PI (–1.50 to 1.12) also indicates the uncertain true effect (Figure 8 [27,30,31,50,52,58,59,61,62,65,68,69,71,72,75,79,83,85,87,91,93,97,102,107,111,115,118-121,129,132,134,136-138,141,143,152,155,157]). Subgroup analyses revealed no statistically significant differences in pairwise comparisons among the intervention types (Multimedia Appendix 8). The analysis of each intervention type obtained a consistent result, with SMD=–0.13 (95% CI –0.52 to 0.25) for technology-based rehabilitation devices, SMD=–0.24 (95% CI –0.77 to 0.29) for online classes, and SMD=–0.19 (95% CI –9.86 to 9.49) for digital databases. Significant heterogeneity was found (I^2=81.2%; P<.001), and the heterogeneity was substantial (τ=0.634; τ^2=0.4053; 95% PI –1.50 to 1.12). The intervention types could not completely explain the heterogeneity, with I^2=84.1% for technology-based rehabilitation devices (P<.001) and I^2=53.6% for online classes (P=.04). The follow-up analysis indicated a consistent result (Multimedia Appendix 9). The leave-one-out analysis obtained a consistent result. GRADE ratings indicated that the certainty of the evidence for the effect of DHIs on quality of life of patients with PD was very low owing to the high risk of bias of included studies, high heterogeneity, and serious imprecision (Figure 4).

XSL•FO
RenderX

Figure 8. Forest plot of the meta-analysis of digital health interventions' efficacy on postintervention quality of life [27,30,31,50,52,58,59,61,62,65,68,69,71,72,75,79,83,85,87,91,93,97,102,107,111,115,118-121,129,132,134,136-138,141,143,152,155,157]. Standardized mean differences (SMDs) with 95% CIs were calculated using a random-effects model. Negative SMD values favor the experimental intervention. Subgroup analyses were performed by intervention type (technology-based rehabilitation devices, online classes, digital databases). The square represents the individual study estimate. The rhombus shape represents the pooled estimates of the lowest accuracy in all studies. *P* values are shown as p.

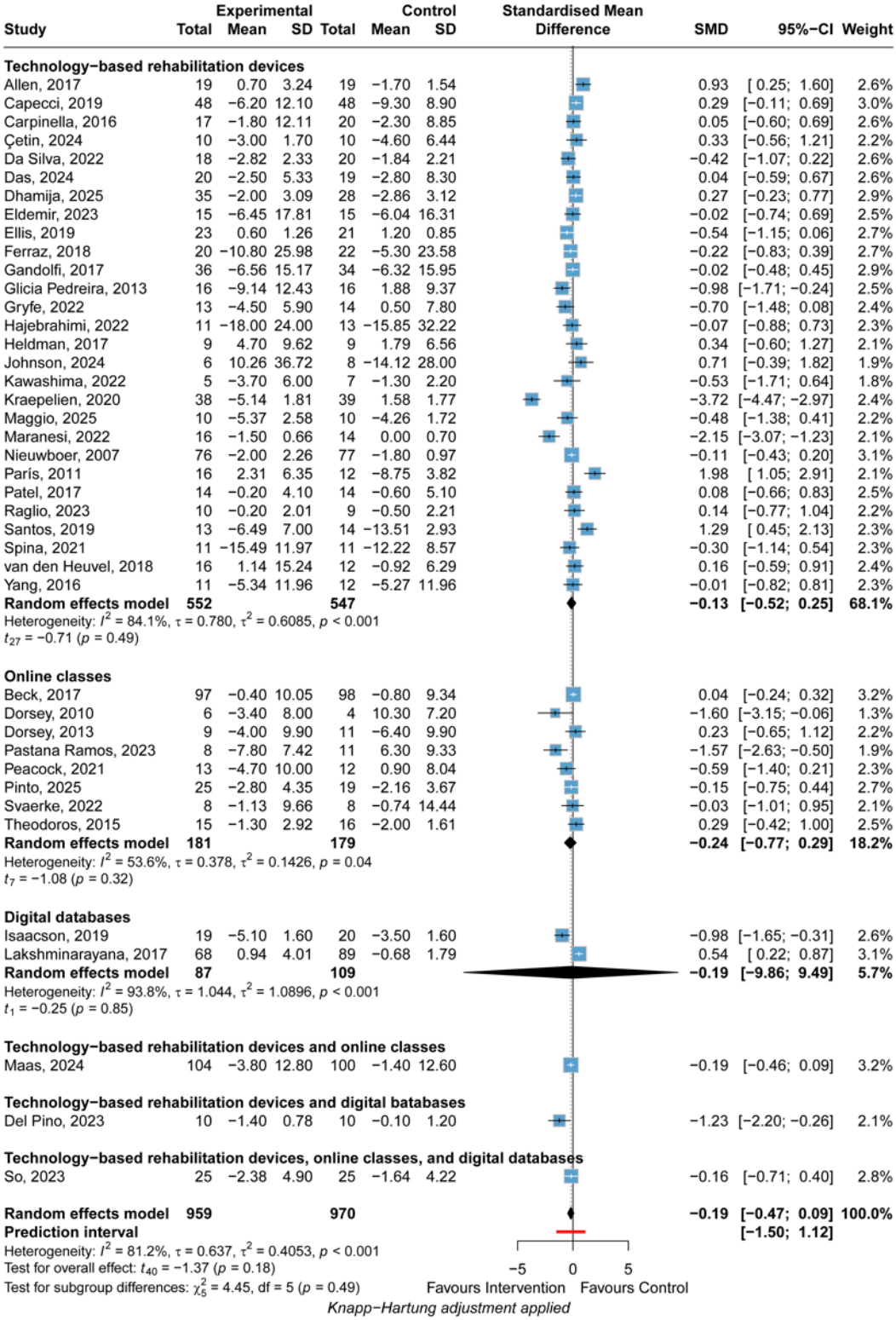

Publication Bias

Small-study effect was assessed via visual inspection of a funnel plot and Egger test (Multimedia Appendix 10). The Egger test indicated a significant small-study effect in the analysis of the effect of DHIs on motor (intercept=–2.07, 95% CI –3.69 to –0.45; *P*=.02) and cognitive function (intercept=2.54, 95% CI –1.12 to 3.97; *P*=.002) of patients with PD. Further, trim-and-fill analysis (Multimedia Appendix 10) was performed, and we observed that the effects were no longer significant for both motor symptoms (SMD=–0.04, 95% CI –0.29 to 0.21) and cognitive function (SMD=0.16, 95% CI –0.14 to 0.47).

XSL•FO
RenderX

Meta-Regressions

To explore potential sources of significant heterogeneity, we performed univariate meta-regression analyses examining 3 individual-level and 8 study-level variables (Multimedia Appendix 11), and we also present bubble plots to visualize the relationship involving the continuous variables (Multimedia Appendix 12). We conducted multivariable meta-regressions using the variables mentioned above as predictors of between-study variance when 2 or more variables accounted for R^2. In the analysis of cognitive functions, 2 of these study variables accounted for R^2: mean age of participants (model Q statistic [QM]=0.08; P=.78; R^2=0.49%) and intervention setting (QM=1.16; P=.29; R^2=8.88%), with a total R^2=1.38% (P=.53). There are 4 variables associated with the improvement of psychiatric symptoms with a total R^2=9.83% (P=.35): the percentage of female participants (QM=2.73; P=.11; R^2=7.42%), publication year (QM=3.39; P=.08; R^2=4.93%), country income level (QM=0.38; P=.54; R^2=2.64%), and intervention purpose (QM=1.44; P=.25; R^2=11.19%). As for motor symptoms, mean age of participants (QM=4.22; P=.04; R^2=2.96%) and supervision mode (QM=1.50; P=.22; R^2=3.96%) may explain part of the heterogeneity, with a total R^2=9.42% (P=.04). Supervision mode (QM=1.41; P=.25; R^2=10.38%) and mean age (QM=0.90; P=.35; R^2=.64%) moderated the effect of DHIs on quality of life with a total R^2 of 14.56% (P=.14). These meta-regression results should be interpreted as exploratory. The low R^2 values indicate that the factors account for only a small proportion of the total heterogeneity. It reminds us that the variability in intervention effects is likely due to a complex combination of clinical and methodological factors not fully captured here. Nevertheless, these findings generate valuable hypotheses for future research.

Reach, Fidelity, and Feasibility of DHIs

Most studies reported their eligible population and participant enrollment (Multimedia Appendix 13). A total of 94 studies reported on intervention reach with a median reach of 37.5% (range 4.5%-94.7%), and all included studies reported the randomly assigned populations, with a median of 50% (range 30.6%-100%) randomly assigned to the intervention groups. A total of 38 studies reported intervention fidelity, revealing a high degree of fidelity across a diverse range, of which 34 reported percentage fidelity with a median of 95.7% (range 26.7%-100%), and another 4 studies provided relevant information regarding fidelity. Dropout rates were reported in 105 studies, with a median of 9.1% (range 0%-61.1%).

A total of 32 studies reported that the interventions were feasible to deliver with high satisfaction, adoption, preference for interventions, high interest, high retention rates, convenience, high acceptance, reliability, safety, comfort, flexibility, high adherence, cost-effectiveness, user-friendliness, high recruitment, and high fidelity. Only 5 of these reports were from LMICs. A total of 35 studies did not mention feasibility directly, but they reported a high rate of usability, no serious adverse events, favorable satisfaction ratings, willingness to recommend to others, low cost, high participant interactivity, and low dropout rate. The remaining 44 studies did not report any data on feasibility.

Discussion

In contrast to previous reviews focusing on single DHI modalities or specific symptoms, this is the first systematic review and meta-analysis to comprehensively evaluate the effectiveness of all existing DHIs across motor symptoms, nonmotor symptoms, and quality of life in patients with PD, while providing a concurrent analysis of implementation feasibility. In this review, we systematically evaluated the effectiveness and implementation of currently reported DHIs in patients with PD, collecting data from 97 RCTs, including 4404 participants from 24 countries. Favorable postintervention effects of DHIs were found over control groups on motor symptoms and specific nonmotor domains (cognitive function, psychiatric symptoms, and overall burden), with the improvement of motor symptoms and cognitive function stable at follow-up assessments. However, substantial and unexplained heterogeneity was detected, and studies with a moderate to high risk of bias accounted for a high proportion; hence, the GRADE approach rated the overall certainty of evidence as low to very low. The wide 95% PIs indicate substantial heterogeneity, suggesting that the true effect remains uncertain. Future implementations might find DHIs produce strong benefits in some populations, yet be ineffective or even worse than control in others, with the results spanning from no effect to a clear positive or negative outcome. In addition, a high proportion of the included studies were assessed as having moderate to high risk of bias, which calls for more rigorously designed studies in the future. Therefore, clinicians should interpret these findings with caution. Additionally, we summarized the reach, fidelity, and feasibility of DHIs, which is valuable for the design of future trials as well as clinical decision-making regarding the application of DHIs. This review brings to the field an updated and integrative evidence that can inform clinical decision-making and guide the design of future DHIs. In real-world implications, our findings highlight that while DHIs hold promise for scalable and accessible PD management, their clinical application should be cautious, personalized, and supported by further high-quality evidence, especially in underserved regions.

Previously reported meta-analyses of the use of DHIs in PD mostly focus on the effects of virtual reality technology [159-167], and others focused on telemedicine [168-170], computerized cognitive training [20], robot-assisted training [112,171-174], mobile app [109,175], and wearable technology [51,176,177]. A recent umbrella review including 8 meta-analyses showed that virtual reality training significantly improved motor performance, mainly including balance ability and stride length in patients with PD [178]. A meta-analysis found that robot-assisted gait training was helpful for the improvement of motor function and balance function [179], while another showed that the efficacy was very uncertain [180]. All of the above meta-analyses were based on a specific type of technology, and their outcomes focused primarily on improvements in motor symptoms. A recent network

XSL•FO
RenderX

meta-analysis compared 4 technology-based interventions (internet-based, proprioceptive, robot, and virtual reality) and found that the virtual reality–based intervention fares the best in terms of improving motor symptoms and quality of life [181]. Notably, no study in the field of PD has included all RCTs that met the definition of DHIs and conducted a systematic synthesis. With technological advancements leading to a rapid expansion of DHI applications in PD, a comprehensive analysis of existing DHI modalities in PD is essential to establish an evidence-based framework for guiding future development.

Our review included all types of DHIs, systematically categorizing them by type, purpose, and implementation characteristics while focusing on the 5 key health outcomes in PD. Regarding motor symptoms and cognitive function, our findings are consistent with most previous meta-analyses that DHIs are effective compared with control groups. In addition, we identified significant small-study effects, a critical issue that has not been reported in prior studies. This may reflect publication bias, whereby smaller studies with null results remain unpublished; genuine heterogeneity between smaller and larger trials is also one of the important sources of small-study effects. We used the trim-and-fill method to estimate and adjust for the number and outcomes of missing studies, and the postintervention effects on both motor symptoms and cognitive function became statistically insignificant. This implies that small-study effects might have an impact on these outcomes, and the wide 95% PIs emphasize the uncertainty of the benefits of DHIs. These results highlight the need for further high-quality studies with larger sample sizes and rigorous designs to provide more definitive evidence for updated meta-analyses. For psychiatric symptoms, although the result was statistically significant, it must be interpreted with caution due to the very low certainty of evidence according to the GRADE framework. This low rating may be driven by several factors: the wide 95% PIs, substantial heterogeneity, and a high risk of bias across the included studies.

As for overall nonmotor symptoms, although the results were statistically significant, they should be considered with caution due to the small sample size. Notably, So et al [134] and Dhamija et al [152] were the only 2 trials among 8 to report significant improvements in nonmotor symptoms, and omitting either of them would result in a null effect of the result. It is potentially attributable to its multimodal design (face-to-face education, telecounseling, and smartphone or wearable tools) and early-stage PD cohort. Hence, we did not use the results as a point of reference. In terms of quality of life, unlike previous studies, we did not find a significant improvement. It may be because the majority of scales measuring quality of life are self-reported, such as the most widely used Parkinson's Disease Questionnaire-39. It lacks high sensitivity for detecting mild to moderate improvements in quality of life, particularly for patients with PD with milder symptoms (Hoehn and Yahr stage ≤3) [182]. In addition, most studies tend to consider quality of life as a secondary outcome, which may affect the design of DHIs. Moreover, complex or long-term interventions may impose an additional burden on patients. Standardized quality of life instruments evaluate a multidimensional construct including emotional well-being, social support, and communication abilities, but the most widely discussed applications of DHIs in PD were motor or cognitive training. Therefore, improvements in isolated domains (eg, mobility or cognitive function) may not substantially enhance overall quality of life. Future DHIs should adopt a user-centered design approach to improve accessibility, reduce treatment-related burdens, and optimize patient experience.

Of the included studies, the majority reported reach or fidelity, and some researchers further reported the feasibility of the intervention. Nearly half of the researchers mentioned that the intervention was feasible and provided quantitative data and qualitative interpretation. This implies that researchers are consciously valuing the feasibility of their interventions. However, the existing problem is the lack of a widely accepted framework for evaluating them, which makes comparing the feasibility of different DHIs across studies challenging. In this review, when analyzing feasibility, we extracted data on reach, fidelity, and dropout rates and summarized the authors' definitions and explanations related to the feasibility of DHIs as mentioned in the study. For those not directly mentioning feasibility, we also extracted information that may be relevant, including satisfaction, adverse effects, user-friendliness, and so on. Some studies reported high safety with few adverse events [61,85,114,119,133,149], as DHIs that provide real-time surveillance of patients can detect unexpected conditions promptly. Furthermore, physical therapy that incorporates digital technology can help prevent falls and freezing of gait [22] during the training process, such as through the use of robot-assisted gait training [179,183]. The reasons for patients dropping out may include the complexity of the intervention system, and face-to-face guidance before DHIs may increase adherence [184]. Future studies should take into account the aforementioned factors to design more feasible intervention protocols.

The included studies exhibited substantial heterogeneity. The primary source of heterogeneity may be that all DHIs were comprehensively summarized rather than a specific type. The types of DHIs range from the simplest intervention designs of only watching relevant videos to more complex designs involving specifically developed robotic devices and even using multiple digital tools simultaneously. Therefore, we conducted subgroup analyses according to the intervention type, yet substantial heterogeneity persisted within subgroups. Although only the technology-based rehabilitation devices subgroup of the 3 subgroups showed a statistically significant pooled effect on motor symptoms, the test for subgroup differences was not significant; therefore, these findings should not be interpreted as evidence that this subgroup is superior to the other 2 subgroups. Moreover, while the original RCTs may demonstrate that a specifically designed DHI is superior to the control group, they do not imply that individuals using different DHIs across trials were randomized. Therefore, such cross-trial comparisons are invalid for establishing causal relationships between different types of DHIs. Consequently, observed subgroup effects may reflect differences in study populations, intervention intensity, comparators, durations, or outcome instruments across studies, rather than genuine differences between the DHIs themselves. Meta-regression revealed that the percentage of female

XSL•FO
RenderX

participants, mean age, intervention setting, publication year, country income level, intervention purpose, and supervision mode might explain a modest portion of the I^2 heterogeneity. While we cannot fully account for all heterogeneity, this finding suggests that the effectiveness of DHIs may vary depending on individual patient characteristics and supervision mode. Therefore, comprehensive consideration of these factors is essential when selecting DHIs for patients. Moreover, the 95% PIs for every primary outcome were wide and consistently crossed the null value. This finding shows substantial heterogeneity and indicates that the effect of DHIs in future settings could range from being less effective than the control condition to being clearly beneficial due to this clinical and methodological diversity. It reminds us that the same intervention may not work equally well for every patient. Hence, clinicians should consider patient responses individually, and future research could aim to identify which types of DHIs are most suitable for specific patient groups.

Clinicians should interpret our findings and their applicability to local settings with caution, as 75% of participants were recruited from high-income countries. The effectiveness and feasibility of these DHIs may differ substantially in LMICs due to variations in resources, infrastructure, and cultural context. Future research in LMICs is critically needed to provide evidence that is generalizable to these settings. As a majority of the world's population has access to mobile technologies [185], the implementation of digital technology in LMICs could potentially help to close the treatment gap [186,187].

Strengths of this review include that only RCTs were included, which are of high evidence quality levels; the review was conducted strictly by the methodology outlined in the Cochrane guidelines; a large number of studies were included, enrolling a large population from 24 countries; the review encompassed a broad range of outcomes, including motor symptoms, overall nonmotor symptoms, cognitive functions, psychiatric symptoms, and quality of life; the Hartung-Knapp-Sidik-Jonkman method was applied to obtain more conservative results; and the 95% PI was provided to estimate the potential true effect.

This review also has limitations. First, this review comprehensively included various types of DHIs for analysis, leading to high heterogeneity in the analysis. We explored heterogeneity through prespecified subgroup analyses and meta-regression and identified 3 variables accounting for the heterogeneity significantly, although to a small degree. Second, we included a large number of studies, which resulted in a longer period from literature search to manuscript preparation. Therefore, secondary and third screenings were conducted to ensure that as many recent studies as possible were comprehensively included. In addition, we observed possible publication bias in the analyses; more rigorous studies are needed to confirm these findings.

In summary, this systematic review provided the first comprehensive synthesis of evidence across all DHI types for PD, which distinguishes it from previous reviews that focused on single technologies or specific outcomes. DHIs were associated with improvement in motor symptoms, overall nonmotor symptoms, cognitive function, and psychiatric symptoms, but not with quality of life, compared with the non-DHI group. This study supports the potential of DHIs as a promising nonpharmacological intervention for PD. It brings to the field an integrated evidence base that can guide the design of future research and inform the careful implementation of DHIs in clinical practice. However, in real-world implications, the conclusions should be interpreted with caution given the wide 95% PIs, relatively high risk of bias, small-study effects, and the very-low GRADE ratings. Therefore, more well-designed DHIs and high-quality research are urgently needed, especially in underresourced regions.

Acknowledgments

We would like to thank the authors of the primary studies for providing the data used in this meta-analysis.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Funding

This research was funded by the Sichuan Science and Technology Program (grant number 2022ZDZX0023) and Chengdu Science and Technology Project Key Research and Development Support Program (grant number 2024-YF08-00038-GX).

Authors' Contributions

RL and SZ contributed to conceptualization, methodology, investigation, data curation, formal analysis, and writing of the original draft. YX contributed to investigation, data curation, formal analysis, validation, and writing—review and editing. YC and HS contributed to supervision, project administration, and writing—review and editing. All authors approved the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA 2020 checklist.

XSL•FO
RenderX

[DOCX File , 32 KB - jmir_v28i1e79935_app1.docx]

Multimedia Appendix 2
PRISMA-S checklist.
[DOCX File , 17 KB - jmir_v28i1e79935_app2.docx]

Multimedia Appendix 3
Deviations from the original protocol in the final review.
[DOCX File , 21 KB - jmir_v28i1e79935_app3.docx]

Multimedia Appendix 4
Search strategies per database.
[DOCX File , 24 KB - jmir_v28i1e79935_app4.docx]

Multimedia Appendix 5
Outcome measurement tools utilized in individual studies.
[DOCX File , 38 KB - jmir_v28i1e79935_app5.docx]

Multimedia Appendix 6
Characteristics of included studies.
[DOCX File , 68 KB - jmir_v28i1e79935_app6.docx]

Multimedia Appendix 7
Risk of bias for included effect estimates based on the Cochrane risk-of-bias tool for randomized trials (RoB2).
[DOCX File , 44 KB - jmir_v28i1e79935_app7.docx]

Multimedia Appendix 8
Pairwise Comparisons of Intervention Type Subgroups.
[DOCX File , 21 KB - jmir_v28i1e79935_app8.docx]

Multimedia Appendix 9
Meta-analysis of digital health interventions' efficacy on follow-up motor symptoms, psychiatric symptoms, cognitive function, and quality of life.
[DOCX File , 448 KB - jmir_v28i1e79935_app9.docx]

Multimedia Appendix 10
Publication bias analyses – Funnel plots and trim and fill analyses at post-intervention assessments on motor symptoms, psychiatric symptoms, cognitive function, overall non-motor symptoms and quality of life.
[DOCX File , 176 KB - jmir_v28i1e79935_app10.docx]

Multimedia Appendix 11
Meta-regression analyses of the effect on motor symptoms, quality of life, cognitive function and psychiatric symptoms.
[DOCX File , 28 KB - jmir_v28i1e79935_app11.docx]

Multimedia Appendix 12
Bubble plots for the continuous variables in meta-regression.
[DOCX File , 365 KB - jmir_v28i1e79935_app12.docx]

Multimedia Appendix 13
Reach, uptake, dropout rate, and feasibility of included interventions.
[DOCX File , 36 KB - jmir_v28i1e79935_app13.docx]

References

1. Bloem BR, Okun MS, Klein C. Parkinson's disease. Lancet 2021;397(10291):2284-2303. [doi: 10.1016/S0140-6736(21)00218-X] [Medline: 33848468]

XSL•FO
RenderX

2. Tanner CM, Ostrem JL. Parkinson's disease. N Engl J Med 2024;391(5):442-452. [doi: 10.1056/NEJMra2401857] [Medline: 39083773]
3. Dorsey ER, Sherer T, Okun MS, Bloem BR. The emerging evidence of the parkinson pandemic. J Parkinsons Dis 2018;8(s1):S3-S8 [FREE Full text] [doi: 10.3233/JPD-181474] [Medline: 30584159]
4. GBD 2016 Parkinson's Disease Collaborators. Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol 2018;17(11):939-953 [FREE Full text] [doi: 10.1016/S1474-4422(18)30295-3] [Medline: 30287051]
5. Foltynie T, Bruno V, Fox S, Kühn AA, Lindop F, Lees AJ. Medical, surgical, and physical treatments for Parkinson's disease. Lancet 2024;403(10423):305-324. [doi: 10.1016/S0140-6736(23)01429-0] [Medline: 38245250]
6. Leite Silva ABR, Gonçalves de Oliveira RW, Diógenes GP, de Castro Aguiar MF, Sallem CC, Lima MPP. Premotor, nonmotor and motor symptoms of Parkinson's disease: a new clinical state of the art. Ageing Res Rev 2023;84:101834. [doi: 10.1016/j.arr.2022.101834] [Medline: 36581178]
7. Li S, Jia C, Li T, Le W. Hot topics in recent Parkinson's disease research: where we are and where we should go. Neurosci Bull 2021;37(12):1735-1744 [FREE Full text] [doi: 10.1007/s12264-021-00749-x] [Medline: 34313916]
8. The Lancet. What next in Parkinson's disease? Lancet 2024;403(10423):219. [doi: 10.1016/S0140-6736(24)00094-1] [Medline: 38245237]
9. Langeskov-Christensen M, Franzén E, Grøndahl Hvid L, Dalgas U. Exercise as medicine in Parkinson's disease. J Neurol Neurosurg Psychiatry 2024;95(11):1077-1088. [doi: 10.1136/jnnp-2023-332974] [Medline: 38418216]
10. Schalkamp AK, Peall KJ, Harrison NA, Sandor C. Wearable movement-tracking data identify Parkinson's disease years before clinical diagnosis. Nat Med 2023;29(8):2048-2056. [doi: 10.1038/s41591-023-02440-2] [Medline: 37400639]
11. Rábano-Suárez P, Del Campo N, Benatru I, Moreau C, Desjardins C, Sánchez-Ferro Á. Digital outcomes as biomarkers of disease progression in early Parkinson's disease: a systematic review. Mov Disord 2025;40(2):184-203. [doi: 10.1002/mds.30056] [Medline: 39613480]
12. Hyeon B, Shin J, Lee J, Kim W, Kwon J, Lee H. Integrating artificial intelligence and optogenetics for Parkinson's disease diagnosis and therapeutics in male mice. Nat Commun 2025;16(1):7797 [FREE Full text] [doi: 10.1038/s41467-025-63025-w] [Medline: 40841722]
13. Qi W, Shen S, Dong C, Zhao M, Zang S, Zhu X. Digital biomarkers for Parkinson disease: bibliometric analysis and a scoping review of deep learning for freezing of gait. J Med Internet Res 2025;27:e71560 [FREE Full text] [doi: 10.2196/71560] [Medline: 40392578]
14. Park Y, Kim EJ, Park S, Lee M. Digital health intervention effect on older adults with chronic diseases living alone: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2025;27:e63168 [FREE Full text] [doi: 10.2196/63168] [Medline: 40163849]
15. Bai X, Zhang H, Jiao Y, Yuan C, Ma Y, Han L. Digital health interventions for chronic wound management: a systematic review and meta-analysis. J Med Internet Res 2024;26:e47904 [FREE Full text] [doi: 10.2196/47904] [Medline: 39012684]
16. Ellis TD, Earhart GM. Digital therapeutics in Parkinson's disease: practical applications and future potential. J Parkinsons Dis 2021;11(s1):S95-S101 [FREE Full text] [doi: 10.3233/JPD-202407] [Medline: 33646177]
17. Xiong S, Lu H, Peoples N, Duman EK, Najarro A, Ni Z. Digital health interventions for non-communicable disease management in primary health care in low-and middle-income countries. NPJ Digit Med 2023;6(1):12 [FREE Full text] [doi: 10.1038/s41746-023-00764-4] [Medline: 36725977]
18. Boima V, Doku A, Agyekum F, Tuglo LS, Agyemang C. Effectiveness of digital health interventions on blood pressure control, lifestyle behaviours and adherence to medication in patients with hypertension in low-income and middle-income countries: a systematic review and meta-analysis of randomised controlled trials. EClinicalMedicine 2024;69:102432 [FREE Full text] [doi: 10.1016/j.eclinm.2024.102432] [Medline: 38333367]
19. Zangger G, Bricca A, Liaghat B, Juhl CB, Mortensen SR, Andersen RM. Benefits and harms of digital health interventions promoting physical activity in people with chronic conditions: systematic review and meta-analysis. J Med Internet Res 2023;25:e46439 [FREE Full text] [doi: 10.2196/46439] [Medline: 37410534]
20. Gavelin HM, Domellöf ME, Leung I, Neely AS, Launder NH, Nategh L. Computerized cognitive training in Parkinson's disease: a systematic review and meta-analysis. Ageing Res Rev 2022;80:101671. [doi: 10.1016/j.arr.2022.101671] [Medline: 35714854]
21. Darbandsari P, Pescatello LS, Piscitelli D, Smith JM, Ugolini A, Colón-Semenza C. Effect of telerehabilitation in Parkinson disease: a systematic review and meta-analysis. Phys Ther 2025;105(11):376-383. [doi: 10.1093/ptj/pzaf121] [Medline: 41065729]
22. Kim J, Porciuncula F, Yang HD, Wendel N, Baker T, Chin A. Soft robotic apparel to avert freezing of gait in Parkinson's disease. Nat Med 2024;30(1):177-185. [doi: 10.1038/s41591-023-02731-8] [Medline: 38182783]
23. Schootemeijer S, de Vries NM, Darweesh SKL, Ascherio A, Schwarzschild MA, Macklin EA. Promoting physical activity in people with Parkinson's disease through a smartphone app: a pilot study. J Neurol Phys Ther 2025;49(2):74-81. [doi: 10.1097/NPT.0000000000000507] [Medline: 39763021]

XSL•FO
RenderX

24. Hinman RS, Campbell PK, Kimp AJ, Russell T, Foster NE, Kasza J. Telerehabilitation consultations with a physiotherapist for chronic knee pain versus in-person consultations in Australia: the PEAK non-inferiority randomised controlled trial. Lancet 2024;403(10433):1267-1278. [doi: 10.1016/S0140-6736(23)02630-2] [Medline: 38461844]
25. Rawson TM, Zhu N, Galiwango R, Cocker D, Islam MS, Myall A. Using digital health technologies to optimise antimicrobial use globally. Lancet Digit Health 2024;6(12):e914-e925. [doi: 10.1016/S2589-7500(24)00198-5] [Medline: 39547912]
26. Knitza J, Gupta L, Hügle T. Rheumatology in the digital health era: status quo and quo vadis? Nat Rev Rheumatol 2024;20(12):747-759. [doi: 10.1038/s41584-024-01177-7] [Medline: 39482466]
27. Capecci M, Pournajaf S, Galafate D, Sale P, Le Pera D, Goffredo M. Clinical effects of robot-assisted gait training and treadmill training for Parkinson's disease. A randomized controlled trial. Ann Phys Rehabil Med 2019;62(5):303-312 [FREE Full text] [doi: 10.1016/j.rehab.2019.06.016] [Medline: 31377382]
28. Hajebrahimi F, Velioglu HA, Bayraktaroglu Z, Helvaci Yilmaz N, Hanoglu L. Clinical evaluation and resting state fMRI analysis of virtual reality based training in Parkinson's disease through a randomized controlled trial. Sci Rep 2022;12(1):8024 [FREE Full text] [doi: 10.1038/s41598-022-12061-3] [Medline: 35577874]
29. Dorsey ER, Venkataraman V, Grana MJ, Bull MT, George BP, Boyd CM. Randomized controlled clinical trial of "virtual house calls" for Parkinson disease. JAMA Neurol 2013;70(5):565-570 [FREE Full text] [doi: 10.1001/jamaneurol.2013.123] [Medline: 23479138]
30. Lakshminarayana R, Wang DL, Burn D, Chaudhuri KR, Galtrey C, Guzman NV. Erratum: author correction: using a smartphone-based self-management platform to support medication adherence and clinical consultation in Parkinson's disease. NPJ Parkinsons Dis 2017;3(1):32 [FREE Full text] [doi: 10.1038/s41531-017-0034-0] [Medline: 29152558]
31. Isaacson SH, Boroojerdi B, Waln O, McGraw M, Kreitzman DL, Klos K. Effect of using a wearable device on clinical decision-making and motor symptoms in patients with Parkinson's disease starting transdermal rotigotine patch: a pilot study. Parkinsonism Relat Disord 2019;64:132-137. [doi: 10.1016/j.parkreldis.2019.01.025] [Medline: 30948242]
32. Sarasso E, Gardoni A, Tettamanti A, Agosta F, Filippi M, Corbetta D. Virtual reality balance training to improve balance and mobility in Parkinson's disease: a systematic review and meta-analysis. J Neurol 2022;269(4):1873-1888. [doi: 10.1007/s00415-021-10857-3] [Medline: 34713324]
33. Elena P, Demetris S, Christina M, Marios P. Differences between exergaming rehabilitation and conventional physiotherapy on quality of life in Parkinson's disease: a systematic review and meta-analysis. Front Neurol 2021;12:683385 [FREE Full text] [doi: 10.3389/fneur.2021.683385] [Medline: 34434157]
34. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
35. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
36. Herold F, Theobald P, Gronwald T, Kaushal N, Zou L, de Bruin ED. Alexa, let's train now! - a systematic review and classification approach to digital and home-based physical training interventions aiming to support healthy cognitive aging. J Sport Health Sci 2024;13(1):30-46 [FREE Full text] [doi: 10.1016/j.jshs.2023.01.004] [Medline: 36736727]
37. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W. Process evaluation of complex interventions: medical research council guidance. BMJ 2015;350:h1258 [FREE Full text] [doi: 10.1136/bmj.h1258] [Medline: 25791983]
38. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011;38(2):65-76 [FREE Full text] [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
39. Moschonis G, Siopis G, Jung J, Eweka E, Willems R, Kwasnicka D. Effectiveness, reach, uptake, and feasibility of digital health interventions for adults with type 2 diabetes: a systematic review and meta-analysis of randomised controlled trials. Lancet Digit Health 2023;5(3):e125-e143 [FREE Full text] [doi: 10.1016/S2589-7500(22)00233-3] [Medline: 36828606]
40. Chapter 8: assessing risk of bias in a randomized trial. Cochrane Handbook for Systematic Reviews of Interventions. 2019. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-08 [accessed 2019-10-01]
41. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336(7650):924-926 [FREE Full text] [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
42. Luo D, Wan X, Liu J, Tong T. Optimally estimating the sample mean from the sample size, median, mid-range, and/or mid-quartile range. Stat Methods Med Res 2018;27(6):1785-1805. [doi: 10.1177/0962280216669183] [Medline: 27683581]
43. Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. BMC Med Res Methodol 2014;14:135 [FREE Full text] [doi: 10.1186/1471-2288-14-135] [Medline: 25524443]
44. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014;14:25 [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
45. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023;12(4):101014 [FREE Full text] [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]

XSL•FO
RenderX

46. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]
47. IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open 2016;6(7):e010247 [FREE Full text] [doi: 10.1136/bmjopen-2015-010247] [Medline: 27406637]
48. Ruppar T. Meta-analysis: how to quantify and explain heterogeneity? Eur J Cardiovasc Nurs 2020;19(7):646-652. [doi: 10.1177/1474515120944014] [Medline: 32757621]
49. Alberts JL, Kaya RD, Penko AL, Streicher M, Zimmerman EM, Davidson S. A randomized clinical trial to evaluate a digital therapeutic to enhance gait function in individuals with Parkinson's disease. Neurorehabil Neural Repair 2023;37(9):603-616. [doi: 10.1177/15459683231184190] [Medline: 37465959]
50. Allen NE, Song J, Paul SS, Smith S, O'Duffy J, Schmidt M. An interactive videogame for arm and hand exercise in people with Parkinson's disease: a randomized controlled trial. Parkinsonism Relat Disord 2017;41:66-72. [doi: 10.1016/j.parkreldis.2017.05.011] [Medline: 28528804]
51. Bartolo M, Castelli A, Calabrese M, Buttacchio G, Zucchella C, Tamburin S. A wearable system for visual cueing gait rehabilitation in Parkinson's disease: a randomized non-inferiority trial. Eur J Phys Rehabil Med 2024;60(2):245-256 [FREE Full text] [doi: 10.23736/S1973-9087.24.08381-3] [Medline: 38483335]
52. Beck CA, Beran DB, Biglan KM, Boyd CM, Dorsey ER, Schmidt PN. National randomized controlled trial of virtual house calls for Parkinson disease. Neurology 2017;89(11):1152-1161 [FREE Full text] [doi: 10.1212/WNL.0000000000004357] [Medline: 28814455]
53. Bernini S, Alloni A, Panzarasa S, Picascia M, Quaglini S, Tassorelli C. A computer-based cognitive training in mild cognitive impairment in Parkinson’s disease. Neuro Rehabilitation 2019;44(4):555-567. [doi: 10.3233/nre-192714]
54. Bernini S, Panzarasa S, Barbieri M, Sinforiani E, Quaglini S, Tassorelli C. A double-blind randomized controlled trial of the efficacy of cognitive training delivered using two different methods in mild cognitive impairment in Parkinson's disease: preliminary report of benefits associated with the use of a computerized tool. Aging Clin Exp Res 2021;33(6):1567-1575. [doi: 10.1007/s40520-020-01665-2] [Medline: 32895890]
55. Bogosian A, Hurt CS, Hindle JV, McCracken LM, Vasconcelos E Sa DA, Axell S. Acceptability and feasibility of a mindfulness intervention delivered via videoconferencing for people with Parkinson's. J Geriatr Psychiatry Neurol 2022;35(1):155-167 [FREE Full text] [doi: 10.1177/0891988720988901] [Medline: 33504245]
56. Calabrò RS, Naro A, Filoni S, Pullia M, Billeri L, Tomasello P. Walking to your right music: a randomized controlled trial on the novel use of treadmill plus music in Parkinson's disease. J Neuroeng Rehabil 2019;16(1):68 [FREE Full text] [doi: 10.1186/s12984-019-0533-9] [Medline: 31174570]
57. Carda S, Invernizzi M, Baricich A, Comi C, Croquelois A, Cisari C. Robotic gait training is not superior to conventional treadmill training in parkinson disease: a single-blind randomized controlled trial. Neurorehabil Neural Repair 2012;26(9):1027-1034. [doi: 10.1177/1545968312446753] [Medline: 22623206]
58. Carpinella I, Cattaneo D, Bonora G, Bowman T, Martina L, Montesano A. Wearable sensor-based biofeedback training for balance and gait in Parkinson disease: a pilot randomized controlled trial. Arch Phys Med Rehabil 2017;98(4):622-630.e3 [FREE Full text] [doi: 10.1016/j.apmr.2016.11.003] [Medline: 27965005]
59. Çetin B, Kılınç M, Çakmaklı GY. The effects of exergames on upper extremity performance, trunk mobility, gait, balance, and cognition in Parkinson's disease: a randomized controlled study. Acta Neurol Belg 2024;124(3):853-863. [doi: 10.1007/s13760-023-02451-3] [Medline: 38182919]
60. Constantinescu G, Theodoros D, Russell T, Ward E, Wilson S, Wootton R. Treating disordered speech and voice in Parkinson's disease online: a randomized controlled non-inferiority trial. Int J Lang Commun Disord 2011;46(1):1-16. [doi: 10.3109/13682822.2010.484848] [Medline: 21281410]
61. Da Silva KG, Nuvolini RA, Bacha JMR, De Freitas TB, Doná F, Torriani-Pasin C. Comparison of the effects of an exergame-based program with conventional physiotherapy protocol based on core areas of the European guideline on postural control, functional mobility, and quality of life in patients with Parkinson's disease: randomized clinical trial. Games Health J 2023;12(3):228-241. [doi: 10.1089/g4h.2022.0039] [Medline: 36206023]
62. Das J, Barry G, Walker R, Vitorio R, Celik Y, McDonald C. The feasibility of a visuo-cognitive training intervention using a mobile application and exercise with stroboscopic glasses in Parkinson's: findings from a pilot randomised controlled trial. PLOS Digit Health 2024;3(12):e0000696. [doi: 10.1371/journal.pdig.0000696] [Medline: 39693278]
63. De Luca R, Latella D, Maggio MG, Di Lorenzo G, Maresca G, Sciarrone F. Computer assisted cognitive rehabilitation improves visuospatial and executive functions in Parkinson's disease: preliminary results. Neuro Rehabilitation 2019;45(2):285-290. [doi: 10.3233/NRE-192789] [Medline: 31498141]
64. de Melo GEL, Kleiner AFR, Lopes JBP, Dumont AJL, Lazzari RD, Galli M. Effect of virtual reality training on walking distance and physical fitness in individuals with Parkinson's disease. Neuro Rehabilitation 2018;42(4):473-480. [doi: 10.3233/NRE-172355] [Medline: 29660956]
65. Del Pino R, de Echevarría AO, Díez-Cirarda M, Ustarroz-Aguirre I, Caprino M, Liu J. Virtual coach and telerehabilitation for Parkinson's disease patients: vCare system. J Public Health (Berl.) 2023;33(7):1583-1596. [doi: 10.1007/s10389-023-02082-1]

XSL•FO
RenderX

66. Dobkin RD, Mann SL, Gara MA, Interian A, Rodriguez KM, Menza M. Telephone-based cognitive behavioral therapy for depression in Parkinson disease. Neurology 2020;94(16):e1764-e1773. [doi: 10.1212/wnl.0000000000009292]
67. Dobkin RD, Mann SL, Weintraub D, Rodriguez KM, Miller RB, St Hill L. Innovating Parkinson's care: a randomized controlled trial of telemedicine depression treatment. Mov Disord 2021;36(11):2549-2558. [doi: 10.1002/mds.28548] [Medline: 33710659]
68. Dorsey ER, Deuel LM, Voss TS, Finnigan K, George BP, Eason S. Increasing access to specialty care: a pilot, randomized controlled trial of telemedicine for Parkinson's disease. Mov Disord 2010;25(11):1652-1659. [doi: 10.1002/mds.23145] [Medline: 20533449]
69. Dorsey ER, Venkataraman V, Grana MJ, Bull MT, George BP, Boyd CM. Randomized controlled clinical trial of "virtual house calls" for Parkinson disease. JAMA Neurol 2013;70(5):565-570 [FREE Full text] [doi: 10.1001/jamaneurol.2013.123] [Medline: 23479138]
70. Edwards JD, Hauser RA, O'Connor ML, Valdés EG, Zesiewicz TA, Uc EY. Randomized trial of cognitive speed of processing training in Parkinson disease. Neurology 2013;81(15):1284-1290 [FREE Full text] [doi: 10.1212/WNL.0b013e3182a823ba] [Medline: 24014503]
71. Eldemir S, Guclu-Gunduz A, Eldemir K, Saygili F, Yilmaz R, Akbostancı MC. The effect of task-oriented circuit training-based telerehabilitation on upper extremity motor functions in patients with Parkinson's disease: a randomized controlled trial. Parkinsonism Relat Disord 2023;109:105334. [doi: 10.1016/j.parkreldis.2023.105334] [Medline: 36917914]
72. Ellis TD, Cavanaugh JT, DeAngelis T, Hendron K, Thomas CA, Saint-Hilaire M. Comparative effectiveness of mHealth-supported exercise compared with exercise alone for people with Parkinson disease: randomized controlled pilot study. Phys Ther 2019;99(2):203-216. [doi: 10.1093/ptj/pzy131] [Medline: 30715489]
73. Fellman D, Salmi J, Ritakallio L, Ellfolk U, Rinne JO, Laine M. Training working memory updating in Parkinson's disease: a randomised controlled trial. Neuropsychol Rehabil 2020;30(4):673-708. [doi: 10.1080/09602011.2018.1489860] [Medline: 29968519]
74. Feng H, Li C, Liu J, Wang L, Ma J, Li G. Virtual reality rehabilitation versus conventional physical therapy for improving balance and gait in Parkinson's disease patients: a randomized controlled trial. Med Sci Monit 2019;25:4186-4192 [FREE Full text] [doi: 10.12659/MSM.916455] [Medline: 31165721]
75. Ferraz DD, Trippo KV, Duarte GP, Neto MG, Bernardes Santos KO, Filho JO. The effects of functional training, bicycle exercise, and exergaming on walking capacity of elderly patients with Parkinson disease: a pilot randomized controlled single-blinded trial. Arch Phys Med Rehabil 2018;99(5):826-833. [doi: 10.1016/j.apmr.2017.12.014] [Medline: 29337023]
76. Flynn A, Preston E, Dennis S, Canning CG, Allen NE. Home-based exercise monitored with telehealth is feasible and acceptable compared to centre-based exercise in Parkinson's disease: a randomised pilot study. Clin Rehabil 2021;35(5):728-739. [doi: 10.1177/0269215520976265] [Medline: 33272025]
77. Furnari A, Calabrò RS, De Cola MC, Bartolo M, Castelli A, Mapelli A. Robotic-assisted gait training in Parkinson's disease: a three-month follow-up randomized clinical trial. Int J Neurosci 2017;127(11):996-1004. [doi: 10.1080/00207454.2017.1288623] [Medline: 28132574]
78. Galli M, Cimolin V, De Pandis MF, Le Pera D, Sova I, Albertini G. Robot-assisted gait training versus treadmill training in patients with Parkinson's disease: a kinematic evaluation with gait profile score. Funct Neurol 2016;31(3):163-170 [FREE Full text] [doi: 10.11138/fneur/2016.31.3.163] [Medline: 27678210]
79. Gandolfi M, Geroin C, Dimitrova E, Boldrini P, Waldner A, Bonadiman S. Virtual reality telerehabilitation for postural instability in Parkinson's disease: a multicenter, single-blind, randomized, controlled trial. Biomed Res Int 2017;2017:7962826 [FREE Full text] [doi: 10.1155/2017/7962826] [Medline: 29333454]
80. Giehl K, Ophey A, Hammes J, Rehberg S, Lichtenstein T, Reker P. Working memory training increases neural efficiency in Parkinson's disease: a randomized controlled trial. Brain Commun 2020;2(2):fcaa115 [FREE Full text] [doi: 10.1093/braincomms/fcaa115] [Medline: 32954349]
81. Giehl K, Ophey A, Reker P, Rehberg S, Hammes J, Barbe MT. Effects of home-based working memory training on visuo-spatial working memory in Parkinson's disease: a randomized controlled trial. J Cent Nerv Syst Dis 2020;12:1179573519899469 [FREE Full text] [doi: 10.1177/1179573519899469] [Medline: 32002011]
82. Ginis P, Nieuwboer A, Dorfman M, Ferrari A, Gazit E, Canning CG. Feasibility and effects of home-based smartphone-delivered automated feedback training for gait in people with Parkinson's disease: a pilot randomized controlled trial. Parkinsonism Relat Disord 2016;22:28-34 [FREE Full text] [doi: 10.1016/j.parkreldis.2015.11.004] [Medline: 26777408]
83. Pedreira G, Prazeres A, Cruz D, Gomes I, Monteiro L, Melo A. Virtual games and quality of life in Parkinson's disease: a randomised controlled trial. Advances Parkinson's Disease 2013;02(04):97-101. [doi: 10.4236/apd.2013.24018]
84. Goffredo M, Baglio F, De Icco R, Proietti S, Maggioni G, Turolla A. Efficacy of non-immersive virtual reality-based telerehabilitation on postural stability in Parkinson's disease: a multicenter randomized controlled trial. Eur J Phys Rehabil Med 2023;59(6):689-696 [FREE Full text] [doi: 10.23736/S1973-9087.23.07954-6] [Medline: 37847247]
85. Gryfe P, Sexton A, McGibbon CA. Using gait robotics to improve symptoms of Parkinson's disease: an open-label, pilot randomized controlled trial. Eur J Phys Rehabil Med 2022;58(5):723-737 [FREE Full text] [doi: 10.23736/S1973-9087.22.07549-9] [Medline: 35708047]

XSL•FO
RenderX

86. Gulcan K, Guclu-Gunduz A, Yasar E, Ar U, Sucullu Karadag Y, Saygili F. The effects of augmented and virtual reality gait training on balance and gait in patients with Parkinson's disease. Acta Neurol Belg 2023;123(5):1917-1925 [FREE Full text] [doi: 10.1007/s13760-022-02147-0] [Medline: 36443623]
87. Hajebrahimi F, Velioglu HA, Bayraktaroglu Z, Helvaci Yilmaz N, Hanoglu L. Clinical evaluation and resting state fMRI analysis of virtual reality based training in Parkinson's disease through a randomized controlled trial. Sci Rep 2022;12(1):8024 [FREE Full text] [doi: 10.1038/s41598-022-12061-3] [Medline: 35577874]
88. Halpern AE, Ramig LO, Matos CEC, Petska-Cable JA, Spielman JL, Pogoda JM. Innovative technology for the assisted delivery of intensive voice treatment (LSVT®LOUD) for Parkinson disease. Am J Speech Lang Pathol 2012;21(4):354-367. [doi: 10.1044/1058-0360(2012/11-0125)] [Medline: 23071195]
89. Han T, Liu Q, Hu Y, Wang Y, Xue K. Effect of Pro-kin visual feedback balance training on balance function of individuals with early Parkinson's disease: a randomized controlled pilot trial. Afr Health Sci 2023;23(2):582-588 [FREE Full text] [doi: 10.4314/ahs.v23i2.67] [Medline: 38223579]
90. Hashemi Y, Taghizadeh G, Azad A, Behzadipour S. The effects of supervised and non-supervised upper limb virtual reality exercises on upper limb sensory-motor functions in patients with idiopathic Parkinson's disease. Hum Mov Sci 2022;85:102977. [doi: 10.1016/j.humov.2022.102977] [Medline: 35932518]
91. Heldman DA, Harris DA, Felong T, Andrzejewski KL, Dorsey ER, Giuffrida JP. Telehealth management of Parkinson's disease using wearable sensors: an exploratory study. Digit Biomark 2017;1(1):43-51 [FREE Full text] [doi: 10.1159/000475801] [Medline: 29725667]
92. Jäggi S, Wachter A, Adcock M, de Bruin ED, Möller JC, Marks D. Feasibility and effects of cognitive-motor exergames on fall risk factors in typical and atypical Parkinson's inpatients: a randomized controlled pilot study. Eur J Med Res 2023;28(1):30 [FREE Full text] [doi: 10.1186/s40001-022-00963-x] [Medline: 36647177]
93. Johnson JK, Longhurst JK, Gevertzman M, Jefferson C, Linder SM, Bethoux F. The use of telerehabilitation to improve movement-related outcomes and quality of life for individuals with Parkinson disease: pilot randomized controlled trial. JMIR Form Res 2024;8:e54599 [FREE Full text] [doi: 10.2196/54599] [Medline: 39083792]
94. Jong-Hoon M, Jin-Hwa J, Hwi-Young C. Effects of balance training using a wii fit balance board on balance, gait and activities of daily living in patients with Parkinson disease: a pilot, randomized controlled trial. Medico-Legal 2020;20(1):1799-1803. [doi: 10.37506/v20/i1/2020/mlu/194564] [Medline: 162448856]
95. Kashif M, Ahmad A, Bandpei MAM, Syed HA, Raza A, Sana V. A randomized controlled trial of motor imagery combined with virtual reality techniques in patients with Parkinson's disease. J Pers Med 2022;12(3):450 [FREE Full text] [doi: 10.3390/jpm12030450] [Medline: 35330450]
96. Kashif M, Albalwi AA, Zulfiqar A, Bashir K, Alharbi AA, Zaidi S. Effects of virtual reality versus motor imagery versus routine physical therapy in patients with parkinson's disease: a randomized controlled trial. BMC Geriatr 2024;24(1):229 [FREE Full text] [doi: 10.1186/s12877-024-04845-1] [Medline: 38443801]
97. Kawashima N, Hasegawa K, Iijima M, Nagami K, Makimura T, Kumon A. Efficacy of wearable device gait training on Parkinson's disease: a randomized controlled open-label pilot study. Intern Med 2022;61(17):2573-2580 [FREE Full text] [doi: 10.2169/internalmedicine.8949-21] [Medline: 35135928]
98. Kegelmeyer DA, Minarsch R, Kostyk SK, Kline D, Smith R, Kloos AD. Use of a robotic walking device for home and community mobility in Parkinson disease: a randomized controlled trial. J Neurol Phys Ther 2024;48(2):102-111. [doi: 10.1097/NPT.0000000000000467] [Medline: 38441461]
99. Khalil H, Busse M, Quinn L, Nazzal M, Batyha W, Alkhazaleh S. A pilot study of a minimally supervised home exercise and walking program for people with Parkinson's disease in Jordan. Neurodegener Dis Manag 2017;7(1):73-84. [doi: 10.2217/nmt-2016-0041] [Medline: 28097926]
100. Kim H, Kim E, Yun SJ, Kang M, Shin HI, Oh B. Robot-assisted gait training with auditory and visual cues in Parkinson's disease: a randomized controlled trial. Ann Phys Rehabil Med 2022;65(3):101620 [FREE Full text] [doi: 10.1016/j.rehab.2021.101620] [Medline: 34896605]
101. Kluger BM, Katz M, Galifianakis NB, Pantilat SZ, Hauser JM, Khan R. Patient and family outcomes of community neurologist palliative education and telehealth support in Parkinson disease. JAMA Neurol 2024;81(1):39-49. [doi: 10.1001/jamaneurol.2023.4260] [Medline: 37955923]
102. Kraepelien M, Schibbye R, Månsson K, Sundström C, Riggare S, Andersson G. Individually tailored internet-based cognitive-behavioral therapy for daily functioning in patients with Parkinson's disease: a randomized controlled trial. J Parkinsons Dis 2020;10(2):653-664. [doi: 10.3233/jpd-191894]
103. Lai B, Bond K, Kim Y, Barstow B, Jovanov E, Bickel CS. Exploring the uptake and implementation of tele-monitored home-exercise programmes in adults with Parkinson's disease: a mixed-methods pilot study. J Telemed Telecare 2020;26(1-2):53-63. [doi: 10.1177/1357633X18794315] [Medline: 30134777]
104. Lau J, Regis C, Burke C, Kaleda MJ, McKenna R, Muratori LM. Immersive technology for cognitive-motor training in Parkinson's disease. Front Hum Neurosci 2022;16(1):863930-863936 [FREE Full text] [doi: 10.3389/fnhum.2022.863930] [Medline: 35615742]

XSL•FO
RenderX

105. Li KP, Zhang ZQ, Zhou ZL, Su JQ, Wu XH, Shi BH. Effect of music-based movement therapy on the freezing of gait in patients with Parkinson's disease: a randomized controlled trial. Front Aging Neurosci 2022;14:924784 [FREE Full text] [doi: 10.3389/fnagi.2022.924784] [Medline: 36337701]
106. Liao YY, Yang YR, Wu YR, Wang RY. Virtual reality-based wii fit training in improving muscle strength, sensory integration ability, and walking abilities in patients with Parkinson's disease: a randomized control trial. Int J Gerontol 2015;9(4):190-195. [doi: 10.1016/j.ijge.2014.06.007]
107. Maas JJL, de Vries NM, IntHout J, Bloem BR, Kalf JG. Effectiveness of remotely delivered speech therapy in persons with Parkinson's disease - a randomised controlled trial. E Clinical Medicine 2024;76:102823 [FREE Full text] [doi: 10.1016/j.eclinm.2024.102823] [Medline: 39309726]
108. Maggio MG, De Cola MC, Latella D, Maresca G, Finocchiaro C, La Rosa G. What about the role of virtual reality in Parkinson disease's cognitive rehabilitation? Preliminary findings from a randomized clinical trial. J Geriatr Psychiatry Neurol 2018;31(6):312-318. [doi: 10.1177/0891988718807973] [Medline: 30360679]
109. Maggio MG, Luca A, Cicero CE, Calabrò RS, Drago F, Zappia M. Effectiveness of telerehabilitation plus virtual reality (Tele-RV) in cognitive e social functioning: a randomized clinical study on Parkinson's disease. Parkinsonism Relat Disord 2024;119:105970. [doi: 10.1016/j.parkreldis.2023.105970] [Medline: 38142630]
110. Manor Y, Mootanah R, Freud D, Giladi N, Cohen JT. Video-assisted swallowing therapy for patients with Parkinson's disease. Parkinsonism Relat Disord 2013;19(2):207-211. [doi: 10.1016/j.parkreldis.2012.10.004] [Medline: 23131836]
111. Maranesi E, Casoni E, Baldoni R, Barboni I, Rinaldi N, Tramontana B. The effect of non-immersive virtual reality exergames versus traditional physiotherapy in Parkinson's disease older patients: preliminary results from a randomized-controlled trial. Int J Environ Res Public Health 2022;19(22):14818 [FREE Full text] [doi: 10.3390/ijerph192214818] [Medline: 36429537]
112. McGibbon CA, Sexton A, Gryfe P. Exercising with a robotic exoskeleton can improve memory and gait in people with Parkinson's disease by facilitating progressive exercise intensity. Sci Rep 2024;14(1):4417 [FREE Full text] [doi: 10.1038/s41598-024-54200-y] [Medline: 38388571]
113. Shih MC, Wang RY, Cheng SJ, Yang YR. Effects of a balance-based exergaming intervention using the Kinect sensor on posture stability in individuals with Parkinson's disease: a single-blinded randomized controlled trial. J Neuroeng Rehabil 2016;13(1):78 [FREE Full text] [doi: 10.1186/s12984-016-0185-y] [Medline: 27568011]
114. Mirelman A, Rochester L, Maidan I, Del Din S, Alcock L, Nieuwhof F. Addition of a non-immersive virtual reality component to treadmill training to reduce fall risk in older adults (V-TIME): a randomised controlled trial. Lancet 2016;388(10050):1170-1182. [doi: 10.1016/S0140-6736(16)31325-3] [Medline: 27524393]
115. Nieuwboer A, Kwakkel G, Rochester L, Jones D, van Wegen E, Willems AM. Cueing training in the home improves gait-related mobility in Parkinson's disease: the RESCUE trial. J Neurol Neurosurg Psychiatry 2007;78(2):134-140 [FREE Full text] [doi: 10.1136/jnnp.200X.097923] [Medline: 17229744]
116. Ophey A, Giehl K, Rehberg S, Eggers C, Reker P, van Eimeren T, et al. Effects of working memory training in patients with Parkinson's disease without cognitive impairment: a randomized controlled trial. Parkinsonism Relat Disord 2020;72:13-22. [doi: 10.1016/j.parkreldis.2020.02.002] [Medline: 32078917]
117. Özden F. Letter to the editor: "comparison of virtual reality rehabilitation and conventional rehabilitation in Parkinson's disease: a randomised controlled trial". Physiotherapy 2021;110:87. [doi: 10.1016/j.physio.2020.04.001] [Medline: 32736882]
118. París AP, Saleta HG, de la Cruz Crespo Maraver M, Silvestre E, Freixa MG, Torrellas CP. Blind randomized controlled study of the efficacy of cognitive training in Parkinson's disease. Mov Disord 2011;26(7):1251-1258. [doi: 10.1002/mds.23688] [Medline: 21442659]
119. Pastana Ramos LF, Vilacorta-Pereira TDCS, Duarte JDS, Yamada ES, Santos-Lobato BL. Feasibility and effectiveness of a remote individual rehabilitation program for people with Parkinson's disease living in the Brazilian Amazon: a randomized clinical trial. Front Neurol 2023;14:1244661 [FREE Full text] [doi: 10.3389/fneur.2023.1244661] [Medline: 37693755]
120. Patel S, Ojo O, Genc G, Oravivattanakul S, Huo Y, Rasameesoraj T. A computerized cognitive behavioral therapy randomized, controlled, pilot trial for insomnia in Parkinson disease (ACCORD-PD). J Clin Mov Disord 2017;4(1):16 [FREE Full text] [doi: 10.1186/s40734-017-0062-2] [Medline: 28852567]
121. Peacock D, Yoneda J, Thomson V, Wile D. Tailoring the use of wearable systems and telehealth for Parkinson's disease. Parkinsonism Relat Disord 2021;89:111-112. [doi: 10.1016/j.parkreldis.2021.07.004] [Medline: 34274619]
122. Picelli A, Melotti C, Origano F, Waldner A, Gimigliano R, Smania N. Does robotic gait training improve balance in Parkinson's disease? A randomized controlled trial. Parkinsonism Relat Disord 2012;18(8):990-993. [doi: 10.1016/j.parkreldis.2012.05.010] [Medline: 22673035]
123. Picelli A, Melotti C, Origano F, Neri R, Waldner A, Smania N. Robot-assisted gait training versus equal intensity treadmill training in patients with mild to moderate Parkinson's disease: a randomized controlled trial. Parkinsonism Relat Disord 2013;19(6):605-610. [doi: 10.1016/j.parkreldis.2013.02.010] [Medline: 23490463]
124. Picelli A, Melotti C, Origano F, Neri R, Verzè E, Gandolfi M. Robot-assisted gait training is not superior to balance training for improving postural instability in patients with mild to moderate Parkinson's disease: a single-blind randomized controlled trial. Clin Rehabil 2015;29(4):339-347 [FREE Full text] [doi: 10.1177/0269215514544041] [Medline: 25082957]

XSL•FO
RenderX

125. Piers RJ, Farchione TJ, Wong BN, Rosellini AJ, Cronin-Golomb A. Telehealth transdiagnostic cognitive behavioral therapy for depression in Parkinson's disease: a pilot randomized controlled trial. Mov Disord Clin Pract 2023;10(1):79-85 [FREE Full text] [doi: 10.1002/mdc3.13587] [Medline: 36704072]
126. Pompeu JE, Mendes FADS, Silva KGD, Lobo AM, Oliveira TDP, Zomignani AP. Effect of Nintendo Wii™-based motor and cognitive training on activities of daily living in patients with Parkinson's disease: a randomised clinical trial. Physiotherapy 2012;98(3):196-204. [doi: 10.1016/j.physio.2012.06.004] [Medline: 22898575]
127. Qayyum S, Hashmi Z, Waqas S, Tariq M, Mughal MW, Tanvir M. Effects of exer–gaming on balance and gait in Parkinson's patients. Pak J Med Health Sci 2022;16(12):213-214. [doi: 10.53350/pjmhs20221612213]
128. Raciti L, Pignolo L, Perini V, Pullia M, Porcari B, Latella D. Improving upper extremity bradykinesia in Parkinson's disease: a randomized clinical trial on the use of gravity-supporting exoskeletons. J Clin Med 2022;11(9):2543 [FREE Full text] [doi: 10.3390/jcm11092543] [Medline: 35566669]
129. Raglio A, De Maria B, Parati M, Giglietti A, Premoli S, Salvaderi S. Movement sonification techniques to improve balance in Parkinson's disease: a pilot randomized controlled trial. Brain Sci 2023;13(11):1586 [FREE Full text] [doi: 10.3390/brainsci13111586] [Medline: 38002546]
130. Ribas CG, Alves da Silva L, Corrêa MR, Teive HG, Valderramas S. Effectiveness of exergaming in improving functional balance, fatigue and quality of life in Parkinson's disease: a pilot randomized controlled trial. Parkinsonism Relat Disord 2017;38(3):13-18. [doi: 10.1016/j.parkreldis.2017.02.006] [Medline: 28190675]
131. Sale P, De Pandis MF, Le Pera D, Sova I, Cimolin V, Ancillao A. Robot-assisted walking training for individuals with Parkinson's disease: a pilot randomized controlled trial. BMC Neurol 2013;13(3):50-57 [FREE Full text] [doi: 10.1186/1471-2377-13-50] [Medline: 23706025]
132. Santos P, Machado T, Santos L, Ribeiro N, Melo A. Efficacy of the nintendo wii combination with conventional exercises in the rehabilitation of individuals with Parkinson's disease: a randomized clinical trial. NeuroRehabilitation 2019;45(2):255-263. [doi: 10.3233/NRE-192771] [Medline: 31498138]
133. Sekimoto S, Oyama G, Hatano T, Sasaki F, Nakamura R, Jo T. A randomized crossover pilot study of telemedicine delivered via iPads in Parkinson's disease. Parkinsons Dis 2019;2019:9403295 [FREE Full text] [doi: 10.1155/2019/9403295] [Medline: 30723541]
134. So HY, Kim SR, Kim S, Park YS, Jo S, Park KW. Effect of home-based self-management intervention for community-dwelling patients with early Parkinson's disease: a feasibility study. J Community Health Nurs 2023;40(2):133-146. [doi: 10.1080/07370016.2022.2133566] [Medline: 36920114]
135. Song J, Paul SS, Caetano MJD, Smith S, Dibble LE, Love R. Home-based step training using videogame technology in people with Parkinson's disease: a single-blinded randomised controlled trial. Clin Rehabil 2018;32(3):299-311. [doi: 10.1177/0269215517721593] [Medline: 28745063]
136. Spina S, Facciorusso S, Cinone N, Armiento R, Picelli A, Avvantaggiato C. Effectiveness of robotic balance training on postural instability in patients with mild Parkinson's disease: a pilot, single blind, randomized controlled trial. J Rehabil Med 2021;53(2):jrm00154 [FREE Full text] [doi: 10.2340/16501977-2793] [Medline: 33585943]
137. Svaerke K, Faerk AK, Riis A, Stiegnitz von Ehrenfels SEM, Mogensen J, Lokkegaard A. Effects of computer-based cognitive rehabilitation on attention, executive functions, and quality of life in patients with Parkinson's disease: a randomized, controlled, single-blinded pilot study. Dement Geriatr Cogn Disord 2021;50(6):519-528. [doi: 10.1159/000520591] [Medline: 34986487]
138. Theodoros DG, Hill AJ, Russell TG. Clinical and quality of life outcomes of speech treatment for Parkinson's disease delivered to the home via telerehabilitation: a noninferiority randomized controlled trial. Am J Speech Lang Pathol 2016;25(2):214-232. [doi: 10.1044/2015_ajslp-15-0005]
139. van Balkom TD, Berendse HW, van der Werf YD, Twisk JWR, Peeters CFW, Hoogendoorn AW. Effect of eight-week online cognitive training in Parkinson's disease: a double-blind, randomized, controlled trial. Parkinsonism Relat Disord 2022;96(1):80-87 [FREE Full text] [doi: 10.1016/j.parkreldis.2022.02.018] [Medline: 35248830]
140. van de Weijer SCF, Duits AA, Bloem BR, de Vries NM, Kessels RP, Köhler S. Feasibility of a cognitive training game in Parkinson's disease: the randomized Parkin'Play study. Eur Neurol 2020;83(4):426-432 [FREE Full text] [doi: 10.1159/000509685] [Medline: 32756067]
141. van den Heuvel MRC, Kwakkel G, Beek PJ, Berendse HW, Daffertshofer A, van Wegen EEH. Effects of augmented visual feedback during balance training in Parkinson's disease: a pilot randomized clinical trial. Parkinsonism Relat Disord 2014;20(12):1352-1358 [FREE Full text] [doi: 10.1016/j.parkreldis.2014.09.022] [Medline: 25283070]
142. Wilkinson JR, Spindler M, Wood SM, Marcus SC, Weintraub D, Morley JF. High patient satisfaction with telehealth in Parkinson disease: a randomized controlled study. Neurol Clin Pract 2016;6(3):241-251 [FREE Full text] [doi: 10.1212/CPJ.0000000000000252] [Medline: 27347441]
143. Yang WC, Wang HK, Wu RM, Lo CS, Lin KH. Home-based virtual reality balance training and conventional balance training in Parkinson's disease: a randomized controlled trial. J Formos Med Assoc 2016;115(9):734-743 [FREE Full text] [doi: 10.1016/j.jfma.2015.07.012] [Medline: 26279172]

XSL•FO
RenderX

144. Yen CY, Lin KH, Hu MH, Wu RM, Lu TW, Lin CH. Effects of virtual reality-augmented balance training on sensory organization and attentional demand for postural control in people with Parkinson disease: a randomized controlled trial. Phys Ther 2011;91(6):862-874. [doi: 10.2522/ptj.20100050] [Medline: 21474638]
145. Yuan RY, Chen SC, Peng CW, Lin YN, Chang YT, Lai CH. Effects of interactive video-game-based exercise on balance in older adults with mild-to-moderate Parkinson's disease. J Neuroeng Rehabil 2020;17(1):91 [FREE Full text] [doi: 10.1186/s12984-020-00725-y] [Medline: 32660512]
146. Zoetewei D, Herman T, Ginis P, Palmerini L, Brozgol M, Thumm PC. On-demand cueing for freezing of gait in Parkinson's disease: a randomized controlled trial. Mov Disord 2024;39(5):876-886. [doi: 10.1002/mds.29762] [Medline: 38486430]
147. Tagliente S, Minafra B, Aresta S, Santacesaria P, Buccoliero A, Palmirotta C. Effectiveness of a home-based computerized cognitive training in Parkinson's disease: a pilot randomized cross-over study. Front Psychol 2024;15:1531688 [FREE Full text] [doi: 10.3389/fpsyg.2024.1531688] [Medline: 39850970]
148. Asadiof F, Hasanpour GZ, Naeim M, Mirzai F, Mohammadi Y. Effectiveness of virtual reality-based cognitive behavioral group therapy on reducing anxiety and depression in Parkinson's patients: a randomized controlled trial. Ann Med Surg (Lond) 2025;87(3):1318-1323. [doi: 10.1097/ms9.0000000000003035]
149. Cubo E, Rohani M, Eissazade N, Garcia-Bustillo Á, Ramírez-Sanz JM, Garrido-Labrador JL. Cost-utility analysis of a coadjutant telemedicine intervention for fall prevention in Parkinson's disease. Eur J Neurol 2025;32(1):e16561. [doi: 10.1111/ene.16561] [Medline: 39838709]
150. Das J, Barry G, Vitorio R, Walker R, McDonald C, Morris R. "It was testing me, but not distressing me." Stroboscopic visual training in Parkinson's disease: a qualitative study. BMC Digit Health 2025;3(1):39. [doi: 10.1186/s44247-025-00182-4]
151. De Vleeschhauwer J, Nackaerts E, D'Cruz N, Zhang Y, Janssens L, Vandenberghe W. The effects of task-specific home-based touchscreen training in people with Parkinson's disease: a pilot randomized controlled trial. J Neurol 2025;272(5):328. [doi: 10.1007/s00415-025-13065-5] [Medline: 40204991]
152. Dhamija RK, Saluja A, Garg D, Chauhan S, Majumdar R, Bhardwaj SB. Teleneurorehabilitation and motor and nonmotor symptoms and quality of life in Parkinson disease: the TELEPARK randomized clinical trial. JAMA Neurol 2025;82(4):376-383. [doi: 10.1001/jamaneurol.2024.5387] [Medline: 39992667]
153. Harpham C, Gunn H, Marsden J, Bescos-Garcia R, Murgatroyd C, Connolly L. Home-based high-intensity interval training for people with Parkinson's: a randomized, controlled, feasibility trial. Health Sci Rep 2025;8(7):e71024 [FREE Full text] [doi: 10.1002/hsr2.71024] [Medline: 40661734]
154. Kratz E, Scheffer J, Volc D, Seebacher B. Comparing face-to-face and online LSVT®LOUD speech training using LSVT®Coach in patients with Parkinson's disease: a pilot randomised controlled trial. BMC Neurol 2025;25(1):183 [FREE Full text] [doi: 10.1186/s12883-025-04161-0] [Medline: 40287683]
155. Maggio MG, Rizzo A, Benenati A, Giambò Mauro F, Cannavò A, De Pasquale P. Efficacy and feasibility of virtual reality-based cognitive tele-rehabilitation in Parkinson's disease: a pilot randomized controlled trial on patients and caregivers. Digit Health 2025;11:20552076251376534 [FREE Full text] [doi: 10.1177/20552076251376534] [Medline: 41018508]
156. Nuvolini RA, Silva KGD, Freitas TBD, Doná F, Torriani-Pasin C, Pompeu JE. Exergame-based program and conventional physiotherapy based on core areas of the European Guideline similarly improve gait and cognition in people with Parkinson's disease: randomized clinical trial. Games Health J 2025;14(5):358-368. [doi: 10.1089/g4h.2024.0116] [Medline: 40160128]
157. Pinto C, Norton S, Cubi-Molla P, Hurt C, Chowdhury S, Dooley N. Evaluating a digital mental health intervention for people with Parkinson's (PACT): acceptability and feasibility randomised controlled trial. Aging Ment Health 2025;29(9):1694-1702 [FREE Full text] [doi: 10.1080/13607863.2025.2478508] [Medline: 40135828]
158. Tayyebi G, Asadiof F, Hashempour B, Lotfi M, Taheri M, Naeim M. Efficacy of virtual reality-based cognitive behavioral group therapy in enhancing emotional well-being and quality of life in Parkinson's disease: a randomized controlled trial. Clin Park Relat Disord 2025;12:100316 [FREE Full text] [doi: 10.1016/j.prdoa.2025.100316] [Medline: 40161512]
159. Chen Y, Gao Q, He CQ, Bian R. Effect of virtual reality on balance in individuals with Parkinson disease: a systematic review and meta-analysis of randomized controlled trials. Phys Ther 2020;100(6):933-945. [doi: 10.1093/ptj/pzaa042] [Medline: 32157307]
160. Lina C, Guoen C, Huidan W, Yingqing W, Ying C, Xiaochun C. The effect of virtual reality on the ability to perform activities of daily living, balance during gait, and motor function in Parkinson disease patients: a systematic review and meta-analysis. Am J Phys Med Rehabil 2020;99(10):917-924. [doi: 10.1097/PHM.0000000000001447] [Medline: 32304383]
161. Wang B, Shen M, Wang YX, He ZW, Chi SQ, Yang ZH. Effect of virtual reality on balance and gait ability in patients with Parkinson's disease: a systematic review and meta-analysis. Clin Rehabil 2019;33(7):1130-1138. [doi: 10.1177/0269215519843174] [Medline: 31016994]
162. Triegaardt J, Han TS, Sada C, Sharma S, Sharma P. The role of virtual reality on outcomes in rehabilitation of Parkinson's disease: meta-analysis and systematic review in 1031 participants. Neurol Sci 2020;41(3):529-536 [FREE Full text] [doi: 10.1007/s10072-019-04144-3] [Medline: 31808000]
163. Chuang CS, Chen YW, Zeng BY, Hung CM, Tu YK, Tai YC. Effects of modern technology (exergame and virtual reality) assisted rehabilitation vs conventional rehabilitation in patients with Parkinson's disease: a network meta-analysis of randomised controlled trials. Physiotherapy 2022;117:35-42. [doi: 10.1016/j.physio.2022.07.001] [Medline: 36242929]

XSL•FO
RenderX

164. Parra AG, Gonzalez-Medina G, Perez-Cabezas V, Casuso-Holgado MJ, Vinolo-Gil MJ, García-Muñoz C. Dropout rate of participants in randomized clinical trials that use virtual reality to train balance and gait in Parkinson's disease. A systematic review with meta-analysis and meta-regression. J Med Syst 2023;47(1):46. [doi: 10.1007/s10916-023-01930-7] [Medline: 37010723]
165. Kwon SH, Park JK, Koh YH. A systematic review and meta-analysis on the effect of virtual reality-based rehabilitation for people with Parkinson's disease. J Neuroeng Rehabil 2023;20(1):94 [FREE Full text] [doi: 10.1186/s12984-023-01219-3] [Medline: 37475014]
166. Sarasso E, Gardoni A, Tettamanti A, Agosta F, Filippi M, Corbetta D. Virtual reality balance training to improve balance and mobility in Parkinson's disease: a systematic review and meta-analysis. J Neurol 2022;269(4):1873-1888. [doi: 10.1007/s00415-021-10857-3] [Medline: 34713324]
167. Cortés-Pérez I, Zagalaz-Anula N, Montoro-Cárdenas D, Lomas-Vega R, Obrero-Gaitán E, Osuna-Pérez MC. Leap motion controller video game-based therapy for upper extremity motor recovery in patients with central nervous system diseases. A systematic review with meta-analysis. Sensors (Basel) 2021;21(6):2065 [FREE Full text] [doi: 10.3390/s21062065] [Medline: 33804247]
168. León-Salas B, González-Hernández Y, Infante-Ventura D, de Armas-Castellano A, García-García J, García-Hernández M. Telemedicine for neurological diseases: a systematic review and meta-analysis. Eur J Neurol 2023;30(1):241-254. [doi: 10.1111/ene.15599] [Medline: 36256522]
169. Chen YY, Guan BS, Li ZK, Yang QH, Xu TJ, Li HB, et al. Application of telehealth intervention in Parkinson's disease: a systematic review and meta-analysis. J Telemed Telecare 2018;26(1-2):3-13. [doi: 10.1177/1357633x18792805]
170. Alnajjar AZ, Abouelmagd ME, Krayim A, AbdelMeseh M, Bushara N, Nabil Y. Efficacy of cognitive behavioral therapy for anxiety and depression in Parkinson's disease patients: an updated systematic review and meta-analysis. Neurol Sci 2024;45(11):5277-5290. [doi: 10.1007/s10072-024-07659-6] [Medline: 38958793]
171. Tao Y, Luo J, Tian J, Peng S, Wang H, Cao J. The role of robot-assisted training on rehabilitation outcomes in Parkinson's disease: a systematic review and meta-analysis. Disabil Rehabil 2024;46(18):4049-4067. [doi: 10.1080/09638288.2023.2266178] [Medline: 37818694]
172. Jiang X, Zhou J, Chen Q, Xu Q, Wang S, Yuan L. Effect of robot-assisted gait training on motor dysfunction in Parkinson's patients: a systematic review and meta-analysis. J Back Musculoskelet Rehabil 2024;37(2):253-268. [doi: 10.3233/BMR-220395] [Medline: 37955075]
173. Alwardat M, Etoom M, Al Dajah S, Schirinzi T, Di Lazzaro G, Sinibaldi Salimei P. Effectiveness of robot-assisted gait training on motor impairments in people with Parkinson's disease: a systematic review and meta-analysis. Int J Rehabil Res 2018;41(4):287-296. [doi: 10.1097/MRR.0000000000000312] [Medline: 30119060]
174. Kegelmeyer DA, Minarsch R, Kostyk SK, Kline D, Smith R, Kloos AD. Use of a robotic walking device for home and community mobility in Parkinson disease: a randomized controlled trial. J Neurol Phys Ther 2024;48(2):102-111. [doi: 10.1097/NPT.0000000000000467] [Medline: 38441461]
175. Özden F. The effect of mobile application-based rehabilitation in patients with Parkinson's disease: a systematic review and meta-analysis. Clin Neurol Neurosurg 2023;225:107579. [doi: 10.1016/j.clineuro.2022.107579] [Medline: 36603336]
176. Barrachina-Fernández M, Maitín AM, Sánchez-Ávila C, Romero JP. Wearable technology to detect motor fluctuations in Parkinson's disease patients: current state and challenges. Sensors (Basel) 2021;21(12):4188 [FREE Full text] [doi: 10.3390/s21124188] [Medline: 34207198]
177. Zhang T, Meng DT, Lyu DY, Fang BY. The efficacy of wearable cueing devices on gait and motor function in Parkinson disease: a systematic review and meta-analysis of randomized controlled trials. Arch Phys Med Rehabil 2024;105(2):369-380. [doi: 10.1016/j.apmr.2023.07.007] [Medline: 37532166]
178. Yu J, Wu J, Lu J, Wei X, Zheng K, Liu B. Efficacy of virtual reality training on motor performance, activity of daily living, and quality of life in patients with Parkinson's disease: an umbrella review comprising meta-analyses of randomized controlled trials. J Neuroeng Rehabil 2023;20(1):133 [FREE Full text] [doi: 10.1186/s12984-023-01256-y] [Medline: 37777748]
179. Xue X, Yang X, Deng Z. Efficacy of rehabilitation robot-assisted gait training on lower extremity dyskinesia in patients with Parkinson's disease: a systematic review and meta-analysis. Ageing Res Rev 2023;85:101837. [doi: 10.1016/j.arr.2022.101837] [Medline: 36634871]
180. Lazzarini SG, Mosconi B, Cordani C, Arienti C, Cecchi F. Effectiveness of robot-assisted training in adults with Parkinson's disease: a systematic review and meta-analysis. J Neurol 2024;272(1):22. [doi: 10.1007/s00415-024-12798-z] [Medline: 39666104]
181. Yau CE, Ho ECK, Ong NY, Loh CJK, Mai AS, Tan E. Innovative technology-based interventions in Parkinson's disease: a systematic review and meta-analysis. Ann Clin Transl Neurol 2024;11(10):2548-2562 [FREE Full text] [doi: 10.1002/acn3.52160] [Medline: 39236299]
182. Hagell P, Nygren C. The 39 item Parkinson's disease questionnaire (PDQ-39) revisited: implications for evidence based medicine. J Neurol Neurosurg Psychiatry 2007;78(11):1191-1198 [FREE Full text] [doi: 10.1136/jnnp.2006.111161] [Medline: 17442762]

XSL•FO
RenderX

183. Xu H, Merryweather A, Foreman KB, Zhao J, Hunt M. Dual-task interference during gait on irregular terrain in people with Parkinson's disease. Gait Posture 2018;63:17-22. [doi: 10.1016/j.gaitpost.2018.04.027] [Medline: 29702370]
184. Kwok JYY, Lee JJ, Choi EPH, Chau PH, Auyeung M. Stay mindfully active during the coronavirus pandemic: a feasibility study of mHealth-delivered mindfulness yoga program for people with Parkinson's disease. BMC Complement Med Ther 2022;22(1):37 [FREE Full text] [doi: 10.1186/s12906-022-03519-y] [Medline: 35130894]
185. Li H, Lewis C, Chi H, Singleton G, Williams N. Mobile health applications for mental illnesses: an Asian context. Asian J Psychiatr 2020;54:102209 [FREE Full text] [doi: 10.1016/j.ajp.2020.102209] [Medline: 32623190]
186. Araya R, Menezes PR, Claro HG, Brandt LR, Daley KL, Quayle J. Effect of a digital intervention on depressive symptoms in patients with comorbid hypertension or diabetes in Brazil and Peru: two randomized clinical trials. JAMA 2021;325(18):1852-1862 [FREE Full text] [doi: 10.1001/jama.2021.4348] [Medline: 33974019]
187. Ouedraogo I, Some BMJ, Oyibo K, Benedikter R, Diallo G. Using serious mobile games to improve health literacy in rural Sub-Saharan Africa: a literature review. Front Public Health 2022;10:768252 [FREE Full text] [doi: 10.3389/fpubh.2022.768252] [Medline: 36466518]

Abbreviations

DHI: digital health intervention
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
LMIC: low- and middle-income country
PD: Parkinson disease
PI: prediction interval
PICOS: Population, Intervention, Comparison, Outcomes, and Study Type
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
QM: model Q statistic
RCT: randomized controlled trial
SMD: standardized mean difference

Edited by S Brini; submitted 01.Jul.2025; peer-reviewed by I Adefolaju, X Liang; comments to author 11.Sep.2025; accepted 07.Jan.2026; published 12.Mar.2026.

Please cite as:
Liu R, Zhang S, Xiao Y, Cheng Y, Shang H
The Effects of Digital Health Interventions on Motor Symptoms, Nonmotor Symptoms, and Quality of Life in Patients With Parkinson Disease: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e79935
URL: https://www.jmir.org/2026/1/e79935
doi:10.2196/79935
PMID:

©Ruwen Liu, Sirui Zhang, Yi Xiao, Yangfan Cheng, Huifang Shang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effects of Wearable Devices on Parkinson Disease: Systematic Review and Meta-Analysis of Randomized Controlled Trials Within the International Classification of Functioning, Disability, and Health Framework

Jiarong Wu[1], MSc; Wanli Zang[1], PhD; Mingqing Fang[2], MM; Ningkun Xiao[3], PhD; Xianzuo Zhang[4], PhD; Su Wang[5], PhD; Qiuxia Zhang[1], PhD

[1]School of Physical Education, Soochow University, Suzhou, Jiangsu, China

[2]Xiangya School of Medicine, Central South University, Changsha, Hunan, China

[3]Laboratory for Brain and Neurocognitive Development, Department of Psychology, Institution of Humanities, Ural Federal University, Yekaterinburg, Russian Federation

[4]Department of Orthopedics, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, China

[5]College of Sports Science and Health, University of Harbin Sport, Harbin, Heilongjiang, China

Corresponding Author:
Qiuxia Zhang, PhD
School of Physical Education
Soochow University
No. 50 East Ring Road
Suzhou, Jiangsu, 215021
China
Phone: 86 13915542576
Email: qxzhang@suda.edu.cn

Abstract

Background: Parkinson disease (PD) impairs gait, balance, and quality of life, and wearable devices have been proposed to support rehabilitation, but evidence for their clinical efficacy remains uncertain.

Objective: This study aimed to evaluate, within the International Classification of Functioning, Disability, and Health (ICF) framework, the effects of wearable-device interventions on gait performance, balance, and health-related quality of life in people with PD by conducting a systematic review and meta-analysis of randomized controlled trials (RCTs).

Methods: We searched PubMed, Web of Science, Cochrane Library, Embase, and ClinicalTrials.gov from inception to November 18, 2025, for RCTs in people with PD comparing wearable-device interventions with control conditions. We used Hartung-Knapp random-effects models to pool mean differences (MDs) or standardized mean differences (SMDs) and reported 95% prediction intervals when ≥3 studies were pooled. Risk of bias was assessed using the Cochrane Risk of Bias (RoB) tool, and certainty of evidence was rated using Grading of Recommendations Assessment, Development, and Evaluation (GRADE).

Results: Nine RCTs involving 260 participants were included. Wearable devices produced a small improvement in stride length (MD 0.10 meter, 95% CI 0.03-0.17), but there was no clear benefit for the 10-Meter Walk Test time (MD 0.04 second, 95% CI –0.06 to 0.15). Double support time showed no reduction (MD –1.59% gait cycle, 95% CI –3.79 to 0.61). Freezing of gait (Freezing of Gait Questionnaire [FOG-Q] and New Freezing of Gait Questionnaire [NFOG-Q]) did not significantly improve (SMD –0.24, 95% CI –0.72 to 0.24). Motor severity (Unified Parkinson Disease Rating Scale Part III [UPDRS III]) showed a small, nonsignificant trend favoring wearable devices (MD –2.16 points, 95% CI –4.39 to 0.07). For balance, pooled results from the Berg Balance Scale (BBS), Mini Balance Evaluation Systems Test (Mini-BESTest), and Performance-Oriented Mobility Assessment Balance Subscale (POMA balance) suggested a borderline effect (SMD 0.48, 95% CI –0.02 to 0.98). Wearable devices did not meaningfully improve Parkinson Disease Questionnaire (PDQ) scores (SMD –0.28, 95% CI –0.74 to 0.17), EQ-5D utility (MD 0.10, 95% CI –0.24 to 0.44), or Falls Efficacy Scale–International (FES-I) scores (MD –0.04, 95% CI –1.10 to 1.02). Prediction intervals frequently crossed the null, suggesting effects may vary by setting and population.

XSL•FO
RenderX

Conclusions: Wearable device interventions for Parkinson disease produced a small improvement on average in stride length, with no consistent benefits for other gait outcomes, balance, or patient-centered outcomes. By integrating ICF mapping with Hartung-Knapp meta-analysis, prediction intervals, and GRADE, and avoiding pooling of conceptually distinct gait measures used in prior reviews, this review clarifies where evidence is most consistent, supports using wearables as adjuncts to rehabilitation, and underscores the need for larger, longer RCTs with standardized outcomes to determine who benefits and how to implement them.

Trial Registration: PROSPERO CRD42024585686; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024585686

(J Med Internet Res 2026;28:e85914) doi:10.2196/85914

KEYWORDS

gait performance; balance function; wearable devices; meta-analysis; International Classification of Functioning, Disability and Health (ICF)

Introduction

Parkinson disease (PD) is the second most common neurodegenerative disorder, characterized by the loss of dopaminergic neurons in the substantia nigra and the presence of Lewy bodies containing α-synuclein [1]. Globally, the estimated number of patients with PD has increased from over 6 million in 1990 to more than 6 million in 2016 [2], and it is expected to double, reaching over 12 million by 2040 [3]. PD poses significant health challenges to patients and places considerable burdens on health care systems and the economy, primarily due to productivity loss and rising health care costs. As the disease progresses, motor symptoms such as bradykinesia, muscle rigidity, resting tremor, and postural instability become more pronounced [4-7]. These changes lead to secondary issues, including diminished walking ability (eg, reduced speed, stride length, and step frequency), increased fall risk, and severe limitations in community participation, which significantly impair patients' quality of life [8-10]. Therefore, restoring walking and balance functions is a crucial aspect of PD rehabilitation, contributing to improved quality of life and independence for patients. While many studies have shown that levodopa can increase stride length and walking speed, treatment options become limited as the disease progresses. Often, after 3-5 years of levodopa use, the duration of its effect decreases, resulting in a worsening of symptoms before the next dose [11,12]. Moreover, long-term use of dopaminergic medications may lead to reduced efficacy and adverse effects such as dyskinesia [13], underscoring the need for alternative and promising PD treatments, such as wearable devices [14-16].

Wearable cueing and sensor-based devices provide a lightweight and portable means of delivering external cues or monitoring motor symptoms in everyday environments, without requiring the continuous presence of health care professionals [17-20]. In PD, these technologies have been used to provide auditory, visual, or somatosensory cues during gait, to quantify tremor and bradykinesia, and to track gait abnormalities, balance, activity levels, and sleep patterns in real time [21-23]. Among these technologies, inertial measurement units (IMUs)—typically combining triaxial accelerometers and gyroscopes into body-worn sensors—have become the most widely used platform for quantitative mobility assessment in both in-clinic and home-based settings [24]. Recent clinical and experimental studies suggest that such devices can acutely improve selected gait parameters and support more individualized adjustment of pharmacological or stimulation therapies [25]. The growing number of clinical and research applications has also prompted international initiatives, such as Movement Disorder Society task forces and national guidance (eg, the 2023 National Institute for Health and Care Excellence recommendations on device-based monitoring in PD), which highlight both the potential and the current uncertainties surrounding wearable technologies in routine care [26]. In addition, qualitative research indicates that long-term use is often limited by issues, such as physical discomfort, interface complexity, privacy concerns, and fluctuating motivation, which may reduce adherence and attenuate real-world effectiveness [27]. Overall, current evidence points to wearable devices as a promising but still incompletely understood adjunct in the management and rehabilitation of people with PD.

Despite this rapid technological development, the specific benefits of wearable devices in improving gait, balance, and quality of life for people with PD have not been fully confirmed. Many studies use observational or small pilot designs, involve heterogeneous device types and training regimens, and focus primarily on laboratory-based gait measures, with relatively less attention to balance, fear of falling, and broader patient-reported outcomes [28,29]. Existing reviews and meta-analyses have often combined wearable and nonwearable cueing interventions or have not applied contemporary random-effects methods, prediction intervals, or Grading of Recommendations Assessment, Development, and Evaluation (GRADE) to formally appraise the certainty of evidence [30]. Moreover, sensor-derived mobility metrics can be influenced by psychological, cognitive, environmental, and technical factors and may differ substantially between supervised laboratory assessments and unsupervised home recordings [30,31], further complicating interpretation of existing data [26,31]. As a result, the overall impact of wearable-device interventions on gait performance, balance, and health-related quality of life in PD remains uncertain.

Therefore, the aim of this study was to conduct a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)-aligned systematic review and meta-analysis of randomized controlled trials (RCTs) to evaluate the effects of wearable-device interventions on gait performance, balance, and health-related quality of life in people with PD, interpreted within the International Classification of Functioning, Disability,

XSL•FO
RenderX

and Health (ICF) framework. We specifically sought to (1) quantify the effects of wearable devices on key gait and motor outcomes (eg, stride length, short-distance gait speed, double support time (DST), freezing of gait, and Unified Parkinson Disease Rating Scale Part III (UPDRS III)), (2) examine their impact on balance and patient-reported outcomes such as quality of life and fear of falling, and (3) appraise the overall certainty of evidence using the GRADE approach. We hypothesized that wearable devices would yield modest improvements in specific gait parameters, whereas evidence for consistent benefits on balance and quality of life would be more limited.

Methods

Overview

This study is based on the ICF framework [32], analyzing the intervention effects of wearable devices on gait, balance, and quality of life in patients with PD. The Population, Intervention, Comparison, and Outcome (PICO) structure is detailed in Table 1. This study has been registered in PROSPERO (International Prospective Register of Systematic Reviews) [33], with registration number CRD42024585686. The review protocol is available in the PROSPERO record.

Table 1. Population, Intervention, Comparison, and Outcome (PICO) framework of wearable device interventions on gait, balance, and quality of life in patients with Parkinson disease under the International Classification of Functioning, Disability, and Health (ICF) framework.

Population and intervention	Comparison	Outcome
Patients with Parkinson disease		
• Intervention settings	Comparison between the wearable device group and the control group	• Gait performance (10MWT[a], Stride length, DST[b], FOG-Q[c], UPDRS III[d])
Hoehn and Yahr stages 1-3		
• Hospital	__[e]	• d450 Walking
• Home and community	—	• d450 Walking
• Laboratory	—	• b730 Muscle power functions
• Intervention personnel	—	• Balance function (BBSf, Mini-BESTestg)
• Physiotherapist	—	• b235 Vestibular functions
• Clinical researchers	—	• d410 Changing basic body position
• Professional instructors • Interventions	—	• Quality of life (PDQh, EQ-5D-utility, FES-Ii)
• Use of wearable devices	—	• b130 Energy and drive functions
• Conventional rehabilitation	—	• d920 Recreation and leisure
• Intervention protocol	—	• d570 Looking after one's health
• Type	—	• d760 Family relationships
• Frequency	—	—
• Duration	—	—

[a]10MWT: 10-Meter Walk Test.
[b]DST: double support time.
[c]FOG-Q: Freezing of Gait Questionnaire.
[d]UPDRS III: Unified Parkinson's Disease Rating Scale Part III.
[e]Not applicable.
[f]BBS: Berg Balance Scale.
[g]Mini-BESTest: Mini Balance Evaluation Systems Test.
[h]PDQ: Parkinson Disease Questionnaire.
[i]FES-I: Falls Efficacy Scale–International.

XSL•FO
RenderX

Information Sources

We conducted a comprehensive literature search in PubMed (National Library of Medicine), Embase (Elsevier, Embase.com), the Cochrane Library (Wiley), and Web of Science Core Collection (Clarivate) from inception to November 18, 2025. We also searched ClinicalTrials.gov on November 18, 2025, to identify ongoing or unpublished RCTs evaluating wearable-device interventions in PD.

Databases were searched separately via their native interfaces (ie, no multidatabase searching on a single platform). No search updates were performed after the final search date, and no email alerts were set. No additional online or print sources (eg, journal tables of contents, conference proceedings, or websites) were purposefully searched or browsed.

We screened the reference lists of relevant systematic reviews and all included trials (backward citation searching) to identify additional eligible studies. Reference list screening was performed during full-text assessment and was completed on November 18, 2025. We did not perform forward citation tracking (citing-reference searches) or set up citation alerts. When necessary, we contacted corresponding authors of included studies to request clarification or missing outcome data. No additional information sources or search methods (eg, gray literature databases, preprint servers, or Google Scholar) were used beyond those described above.

Search Strategy

The search strategy combined controlled vocabulary terms (eg, MeSH [Medical Subject Headings] and Emtree) and free-text keywords related to PD, wearable or sensor-based devices, and RCTs. We did not use published search filters (eg, validated RCT filters); trial design was captured using a combination of controlled vocabulary and free-text terms. Search strategies were developed de novo and were not adapted or reused from previous literature reviews.

Searches were limited to English-language publications involving human adults; no restrictions were placed on publication status. These restrictions were applied to align with the eligibility criteria (adult PD populations and RCTs evaluating intervention efficacy) and for feasibility of screening and data extraction.

The search strategy and its reporting followed PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension). The full search strings for each database and information source (including all search terms, Boolean operators, limits, and the date last searched) are provided in S1 in Multimedia Appendix 1. The search strategy was not formally peer-reviewed.

Eligibility Criteria

Inclusion Criteria

Studies were eligible if they met all of the following criteria:

(1) Study design: RCTs evaluating wearable-device interventions in people with PD.

(2) Participants: adults (≥18 years) with a clinical diagnosis of PD based on internationally accepted criteria. Participants were required to be clinically stable; where pharmacotherapy was used (eg, levodopa or other anti-Parkinsonian medications), medication regimens were stable during the trial period (ie, no major medication changes). No restrictions were applied with respect to sex.

(3) Interventions and comparators: the experimental group received a wearable-device intervention (eg, wearable cueing or feedback or sensor-based device, assisted rehabilitation), either as a standalone intervention or as an adjunct to usual care (eg, conventional training and/or pharmacotherapy). Control conditions included usual care without an active wearable-device component and could include conventional training and/or pharmacotherapy, with or without sham or placebo wearable devices.

(4) Outcomes: studies were eligible if they reported at least one prespecified outcome in any of the following domains: gait and mobility, motor severity, balance, and health-related quality of life outcomes. Detailed outcome definitions and measurement instruments are provided in Section 2.6 (Data items—outcomes).

Exclusion Criteria

Studies were excluded if they:

(1) Were not RCTs (eg, nonrandomized studies, observational designs, case series, and qualitative studies);

(2) Enrolled participants younger than 18 years or did not clearly involve a PD population;

(3) Did not evaluate an eligible wearable-device intervention and/or lacked an appropriate control condition;

(4) Did not report any outcomes of interest and/or did not provide sufficient data for extraction (eg, missing summary statistics), and the necessary data could not be obtained from the report or study authors; or

(5) Had unavailable full texts.

For duplicate publications or multiple reports from the same trial, the most complete or most recent report was retained, and companion reports were used to supplement missing details where applicable.

Selection Process

Two reviewers (JW and WZ) independently screened all retrieved records. In the first stage, titles and abstracts were assessed for relevance and study design. In the second stage, full texts were evaluated against the predefined inclusion and exclusion criteria. Any discrepancies were resolved by discussion; if consensus could not be reached, a third reviewer adjudicated. The study selection process was documented using a PRISMA 2020 flow diagram. The completed PRISMA checklist is provided in Multimedia Appendix 2. No automation tools were used for study selection. As only English-language reports were eligible, no translation was required for screening.

XSL•FO
RenderX

Data Collection Process

Data were extracted independently by 2 reviewers (MF and NX) using a piloted, standardized extraction form. Discrepancies in data extraction were resolved by discussion, with adjudication by a third reviewer when necessary. For each prespecified outcome, we extracted data at baseline and at the end of the intervention (postintervention values and/or change scores, as reported), prioritizing the end-of-intervention assessment for the primary synthesis. Risk-of-bias information for each included study was extracted in parallel to support risk-of-bias assessments and GRADE judgments. When required data were missing or unclear, attempts were made to contact study authors; otherwise, available data were analyzed as reported. No automation tools were used for data extraction. As only English-language reports were eligible, no translation was required for data collection.

Data Items (Outcomes)

Outcome data were extracted for the prespecified outcome domains and measurement instruments, with scale directions checked to ensure consistent interpretation. Outcomes of interest included:

1. Gait and mobility: 10-Meter Walk Test (10MWT) time, stride length, DST, and freezing of gait assessed by the Freezing of Gait Questionnaire (FOG-Q) or New Freezing of Gait Questionnaire (NFOG-Q).
2. Motor severity: UPDRS III.
3. Balance: Berg Balance Scale (BBS), Mini Balance Evaluation Systems Test (Mini-BESTest), and Performance-Oriented Mobility Assessment—Balance subscale (POMA balance).
4. Health-related quality of life outcomes: Parkinson Disease Questionnaire (PDQ-39 or PDQ-8), EQ-5D utility index, and Falls Efficacy Scale-International (FES-I).

For each outcome, we extracted the assessment time points and baseline and end-of-intervention data (postintervention values and/or change scores as reported), including the mean and SD or SE and sample size per group.

Data Items (Other Variables)

The following additional data items were extracted:

1. Study characteristics: first author, publication year, country, study design (restricted to RCTs), recruitment source, sample size per arm, follow-up duration, trial registration, ethics approval, funding sources, and declared conflicts of interest.
2. Participant baseline characteristics: age, sex distribution, disease duration, PD severity or stage, medication status (ON or OFF or stable medication), and baseline comparability between groups.
3. Intervention and comparator details: type and modality of wearable device (eg, auditory, visual or vibrotactile cueing, feedback features, stimulation system, and sensor placement), training context (clinic, home, or laboratory), intervention dose (frequency, session duration, or total intervention period), cointerventions, adherence, compliance, dropouts, and adverse events.

Study Risk of Bias Assessment

Risk of bias was assessed at the study level for included RCTs using the Cochrane Risk of Bias tool (RoB 1), following the Cochrane Handbook (version 5.1.0). The following domains were evaluated, including random sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment, incomplete outcome data, selective reporting, and other sources of bias. Each domain was judged as low, high, or unclear risk of bias. An overall risk-of-bias judgment was derived as follows: studies were rated low overall if all domains were low risk, high overall if one or more domains were high risk, and unclear overall otherwise. Two reviewers (JW and WZ) independently performed the assessments, and disagreements were resolved through discussion, with adjudication by a third reviewer (QZ) when necessary. Where required, study authors were contacted to clarify information relevant to risk-of-bias judgments. No automation tools were used for risk-of-bias assessment.

Effect Measures

All prespecified outcomes were continuous measures. For outcomes assessed using the same instrument and reported on the same scale across studies, we synthesized effects as mean differences (MDs) with 95% CIs (eg, 10MWT time [seconds], stride length [meters], double support time [% gait cycle], UPDRS III [points], EQ-5D utility, FES-I [points]). When the same construct was assessed using different instruments, we used standardized mean differences (SMDs; Hedges *g*) with 95% CIs (eg, pooling balance outcomes across BBS, Mini-BESTest, and POMA balance; pooling quality-of-life outcomes across PDQ-39 and PDQ-8). Effect directions were checked before synthesis to ensure consistent interpretation across scales (eg, lower scores indicating improvement for PDQ, FOG-Q and NFOG-Q, UPDRS III, and 10MWT time; higher scores indicating improvement for balance scales).

For each outcome, we prioritized end-of-intervention assessments. We preferentially extracted post-intervention values; when only change-from-baseline data were reported, these were used as reported. Where change scores were not reported but could be derived from available pre- and postintervention summary statistics, they were calculated using standard formulas; otherwise, postintervention values were used. For all meta-analyses, effect estimates were presented with 95% CIs and, when 3 or more studies were available, 95% prediction intervals. We did not prespecify universal thresholds for “small, moderate, or large” effects; interpretation emphasized clinical relevance on the original scale (for MDs) and certainty of evidence (GRADE).

Synthesis Methods

Eligibility for Each Synthesis

For each prespecified outcome, studies were eligible for quantitative synthesis if they were randomized controlled trials and reported the relevant outcome with extractable data at baseline and at the end of the intervention for at least one wearable-device versus control comparison. When fewer than 2 studies were available for a given outcome, results were summarized narratively without pooling.

XSL•FO
RenderX

Preparing for Synthesis

Prior to synthesis, outcome definitions, units, and scale directions were checked to ensure consistent interpretation of effect estimates (eg, higher scores consistently indicated better or worse status as appropriate). Where required summary statistics were not directly reported, we attempted to derive them from available information (eg, SEs or 95% CIs) using standard formulas; when this was not possible, study authors were contacted. If key statistics remained unavailable after these attempts, the outcome was not pooled.

Methods Used to Tabulate and Visually Display Results

Study characteristics and extracted outcome data were summarized in tables. Meta-analysis results were presented using forest plots. Risk-of-bias judgments were summarized both graphically (eg, traffic-light plots) and in tabular form.

Methods Used to Synthesize Results

Studies were grouped for synthesis by outcome domain and measurement instrument. We conducted separate meta-analyses for (1) gait performance (10MWT, stride length, double support time, FOG-Q, and NFOG-Q), (2) motor severity (UPDRS III), (3) balance (BBS, Mini-BESTest, and POMA balance), and (4) health-related quality of life outcomes (PDQ-39, PDQ-8, EQ-5D utility, and FES-I). When multiple time points were reported, we prioritized the end-of-intervention assessment for the primary synthesis.

Meta-analyses were conducted in R (meta package; R Core Team). For continuous outcomes, pooled effects were expressed as MD when studies used the same measurement scale or SMD (Hedges *g*) when different instruments were used to assess the same construct, both with 95% CIs. Given expected clinical and methodological variability across interventions, all meta-analyses used a random-effects model [34]. Between-study variance (τ^2) was estimated using restricted maximum likelihood (REML) [35], and the Hartung-Knapp-Sidik-Jonkman method was applied to obtain more robust CIs [36]. Statistical heterogeneity was quantified using I^2 and τ/τ^2 statistics. Where at least three studies were available, 95% prediction intervals were calculated using the bootstrap approach proposed by Nagashima et al [37], (method.predict="NNF" in the meta package; requires the pimeta package) to improve performance in meta-analyses with few studies [38].

Methods Used to Explore Possible Causes of Heterogeneity

Formal investigations of heterogeneity (eg, subgroup analyses, meta-regression, or statistical tests for interaction) were not undertaken because most syntheses included few studies and statistical heterogeneity was generally low, making effect-modifier analyses unreliable.

Sensitivity Analyses

To examine robustness to measurement heterogeneity, prespecified sensitivity analyses were conducted when feasible (typically when 3 or more studies were available) by restricting analyses to studies using the same instrument for a given construct (eg, pooling FOG-Q only when most studies used FOG-Q rather than NFOG-Q; pooling BBS only when multiple balance instruments were used). Sensitivity analysis results were reported alongside the primary analyses.

Reporting Bias Assessment

Because the number of included studies contributing to each synthesis was small, formal assessment of reporting bias due to missing results (eg, funnel plot asymmetry and Egger regression test for small-study effects) was planned only when 10 or more studies were available for a given outcome. When fewer than 10 studies were available, formal statistical and graphical assessments were not undertaken because of limited power and interpretability. In such cases, the potential impact of missing results was considered qualitatively when judging certainty of evidence within the GRADE framework (publication bias domain).

Certainty Assessment

The certainty (confidence) in the body of evidence for each prespecified outcome was assessed using the GRADE approach, considering five domains: risk of bias, inconsistency, imprecision, indirectness, and publication bias [39]. Evidence certainty was categorized as high, moderate, low, or very low. Two reviewers (JW and WZ) independently performed the GRADE assessments, and disagreements were resolved by discussion, with adjudication by a third reviewer (QZ) when necessary.

Results

Study Selection

A total of 2300 relevant articles were identified. After the initial screening, 866 duplicate articles were excluded. Based on the titles and abstracts, a further 1394 articles were excluded. Following a full-text review according to the inclusion and exclusion criteria, 31 additional articles were excluded. Ultimately, 9 studies were included in the analysis [25,40-47]. The literature screening process is shown in Figure 1. Full-text reports that appeared potentially eligible but were excluded, together with reasons for exclusion, are listed in S2 in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram of study selection.

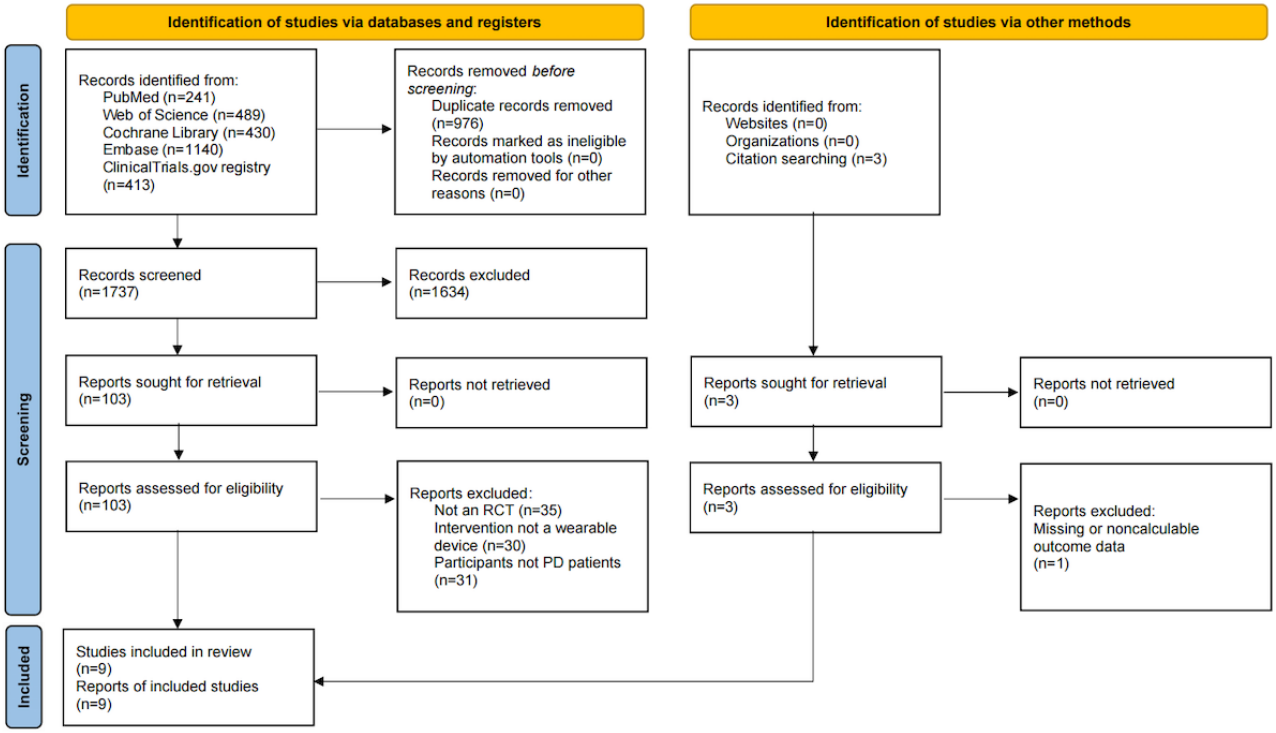

Risk of Bias in Studies

Figures 2A and 2B summarize the risk-of-bias assessment [25,40-47]. Overall, methodological quality was mixed. Most trials reported adequate random sequence generation, whereas details of allocation concealment were frequently insufficient or not reported, leading to many domains being judged as having an unclear risk of bias. Because blinding of participants and personnel was largely not feasible given the nature of wearable-device interventions, several studies were rated at high risk of performance bias. In contrast, blinding of outcome assessors was implemented in approximately half of the trials. Incomplete outcome data and selective reporting were generally judged to be at low risk of bias, with most studies reporting attrition and prespecified outcomes adequately. Other potential sources of bias were usually rated as unclear.

XSL•FO
RenderX

Figure 2. Quality assessment of included studies. (A) Risk of bias summary. (B) Risk of bias graph: percentage summary across all included studies for each type of bias [25,40-46,48].

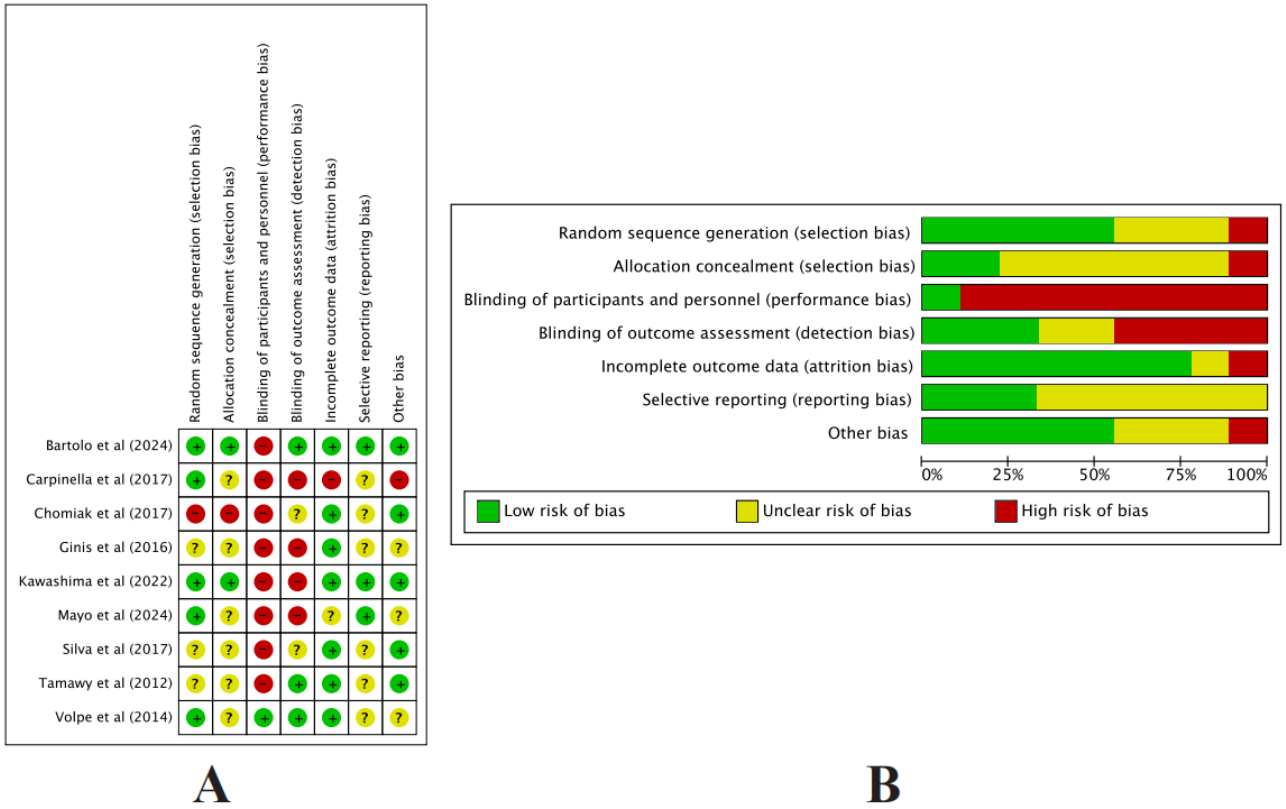

Study Characteristics

A total of 9 studies were included [25,40-47], all of which were published in English, involving a total of 260 PD patients. In most studies, the average age of patients was over 60 years, and the average disease duration exceeded 3 years. Two studies [42,45] conducted 10MWT analysis, 4 studies [40,41,43,46] analyzed stride length, 3 studies [41,43,46] analyzed DST, 4 studies [41,42,44,45] analyzed FOG-Q, 5 studies [41,42,45-47] analyzed UPDRS III, 5 studies [41,42,45-47] analyzed balance function, 3 studies [42,45,47] analyzed PDQ-39, 2 studies [25,45] analyzed EQ-5D utility, and 2 studies [41,44] analyzed FES-I. In most of these studies, patients maintained stable medication regimens before and during the treatment period. Detailed information on the patients and wearable device variables used in these studies is shown in Table 2.

XSL•FO
RenderX

Table 2. Characteristics of included studies

Included studies	T[a]/C[b] (n)	Age in years (T/C), mean (SD)	Hoehn and Yahr stage (T/C)	Disease duration (T/C) years	Wearable prompt device	Feedback characteristics	Stimulation system	Intervention method, (T/C)	Training environment	Intervention duration	Outcome measures
El-Tamawy et al (2012) [40]	15/15	61.4 (7.28)/63.2 (5.6)	2.8 (0.5)/2.6 (0.4)	4.0 (0.9)/3.8 (0.9)	Vibration device placed in the shoes	Proprioception	Open-loop system	Proprioceptive neuromuscular facilitation technique and vibration stimulation + physical therapy program / conventional physical therapy	Laboratory	8 weeks, 3 times/week, 45 minutes/session	Stride length
Volpe et al (2014) [47]	20/20	66.5 / 69.	3.0 / 3.0	6.0 / 6.5	Equistasi nanotech device (Equistasi Srl) placed on C7 and both soleus tendons	Proprioception	Open-loop system	Proprioceptive stimulation (Equistasi) + balance training / balance training only	Laboratory setting (hospital)	8 weeks, 5 times/week, 60 minutes/session	UPDRS III (Unified Parkinson Disease Rating Scale Part III); DST (double support time); BBS (Berg Balance Scale); PDQ (Parkinson Disease Questionnaire)

XSL•FO
RenderX

Included studies	T[a]/C[b] (n)	Age in years (T/C), mean (SD)	Hoehn and Yahr stage (T/C)	Disease duration (T/C) years	Wearable prompt device	Feedback characteristics	Stimulation system	Intervention method, (T/C)	Training environment	Intervention duration	Outcome measures
Ginis et al (2016) [41]	20/18	67.3 (8.13)/ 66.11 (8.07)	2.3 (0.44)/2.2 (0.39)	10.65 (5.39)/11.67 (7.63)	CuPiD system	Auditory feedback	Open-loop system	Real-time gait feedback training provided by the CuPiD system / personalized gait advice	Home	6 weeks, 3 times/week, 30 minutes/session	Stride length; UPDRS III (Unified Parkinson Disease Rating Scale Part III); FOG-Q (Freezing of Gait Questionnaire); DST (double support time); FES-I (Falls Efficacy Scale - International); Mini-BESTest (Mini Balance Evaluation Systems Test)
Carpinella et al (2017) [42]	17/20	73 (7.1)/ 75.6 (8.2)	2.7 (0.7)/2.9 (0.5)	7.5 (3.2)/ 10.3 (5.7)	Gamepad system	Visual and auditory feedback	Open-loop system	Gait and balance training using the Gamepad system / personalized physical therapy	Hospital	20 sessions, 3 times/week, 45 minutes/session	10MWT (10-Meter Walk Test); UPDRS III (Unified Parkinson Disease Rating Scale Part III); FOG-Q (Freezing of Gait Questionnaire); BBS (Berg Balance Scale); PDQ (Parkinson Disease Questionnaire)

XSL•FO
RenderX

Included studies	T[a]/C[b] (n)	Age in years (T/C), mean (SD)	Hoehn and Yahr stage (T/C)	Disease duration (T/C) years	Wearable prompt device	Feedback characteristics	Stimulation system	Intervention method, (T/C)	Training environment	Intervention duration	Outcome measures
Lirani-Silva et al (2017) [43]	10/9	70.4 (6.87)/72 (6.2)	2 (0.5)/ 1.9 (0.4)	N/A[c]	Textured insoles	Proprioceptive feedback	Open-loop system	Using textured insoles / using conventional insoles	Home	1 week	Stride length; DST (double support time)
Chomiak et al (2017) [44]	5/6	70.8 (5.6)/ 69 (5.7)	2.5 (0.50/2.7 (0.41)	15.4 (5.4)/ 11.2 (5)	Ambulosono platform	Auditory feedback	Closed-loop system	Feedback through music playback / Feedback through CBC podcasts	Home	4 weeks, 3 times/week, 10-20 minutes/session	FOG-Q (Freezing of Gait Questionnaire); FES-I (Falls Efficacy Scale - International)
Kawashima et al (2022) [45]	5/7	76.6 (5.3)/ 75.4 (5.7)	2.4 (0.55)/2.4 (0.79)	11.2 (5.8)/ 12.4 (4.6)	Step Management Assist exoskeleton	Proprioceptive feedback	Open-loop + closed-loop system	Gait training using the SMA / conventional gait training	Home-based	3 months, 10 sessions, 30 minutes per session	10MWT (10-Meter Walk Test); stride length; UPDRS III (Unified Parkinson Disease Rating Scale Part III); FOG-Q (Freezing of Gait Questionnaire); BBS (Berg Balance Scale); PDQ (Parkinson Disease Questionnaire); EQ-5D-utility

XSL•FO
RenderX

Included studies	T[a]/C[b] (n)	Age in years (T/C), mean (SD)	Hoehn and Yahr stage (T/C)	Disease duration (T/C) years	Wearable prompt device	Feedback characteristics	Stimulation system	Intervention method, (T/C)	Training environment	Intervention duration	Outcome measures
Bartolo et al (2024) [46]	26/26	73.0 (7.3)/ 70.3 (11)	2 (0.6)/2.1 (0.7)	9.4 (3.1)/ 9.8 (3.9)	Q-Walk system	Visual feedback	Closed-loop system	Gait training using the Q-Walk system / conventional gait training	Clinical rehabilitation	2 weeks, 5 sessions per week, 90 minutes per session	10MWT (10-Meter Walk Test); UPDRS III (Unified Parkinson Disease Rating Scale Part III); DST (double support time); POMA Balance (Performance-Oriented Mobility Assessment Balance Subscale)
Mayo et al (2024) [25]	14/7	70.2 (8.5)/ 70.7 (8.8)	2～3	N/A	Heel2Toe sensor	Auditory feedback	Closed-loop system	Training with Heel2Toe sensor / training according to the manual	Home-based	3 months, twice daily, 5 minutes per session	PDQ (Parkinson Disease Questionnaire); EQ-5D-utility

[a]T: trials (experimental group).

[b]C: controls (control group).

[c]N/A: not applicable.

Results of Syntheses

Gait Performance

10MWT

Two studies [42,45] evaluated the 10MWT outcome. Between-study heterogeneity was negligible (τ^2 0, τ=0; I^2=0%; Q test P=.92). The pooled analysis showed no clear evidence of a difference between the wearable-device and control groups in 10MWT (MD 0.04, 95% CI –0.06 to 0.15; P=.12; Figure 3A). Prediction intervals were not calculated because fewer than 3 studies were available.

XSL•FO
RenderX

Figure 3. Meta-analyses of gait performance and motor severity outcomes for wearable-device interventions versus control in Parkinson disease. (A) 10-Meter Walk Test (10MWT) time (seconds). (B) Stride length (meters). (C) Double support time (DST, % gait cycle). (D) Freezing of gait, pooled across the Freezing of Gait Questionnaire (FOG-Q) and the New Freezing of Gait Questionnaire (NFOG-Q), expressed as standardized mean differences (SMDs). (E) Sensitivity analysis for freezing of gait restricted to studies using the FOG-Q only, expressed as mean differences (points). (F) Unified Parkinson Disease Rating Scale Part III (UPDRS III) score (points) [40-46,48].

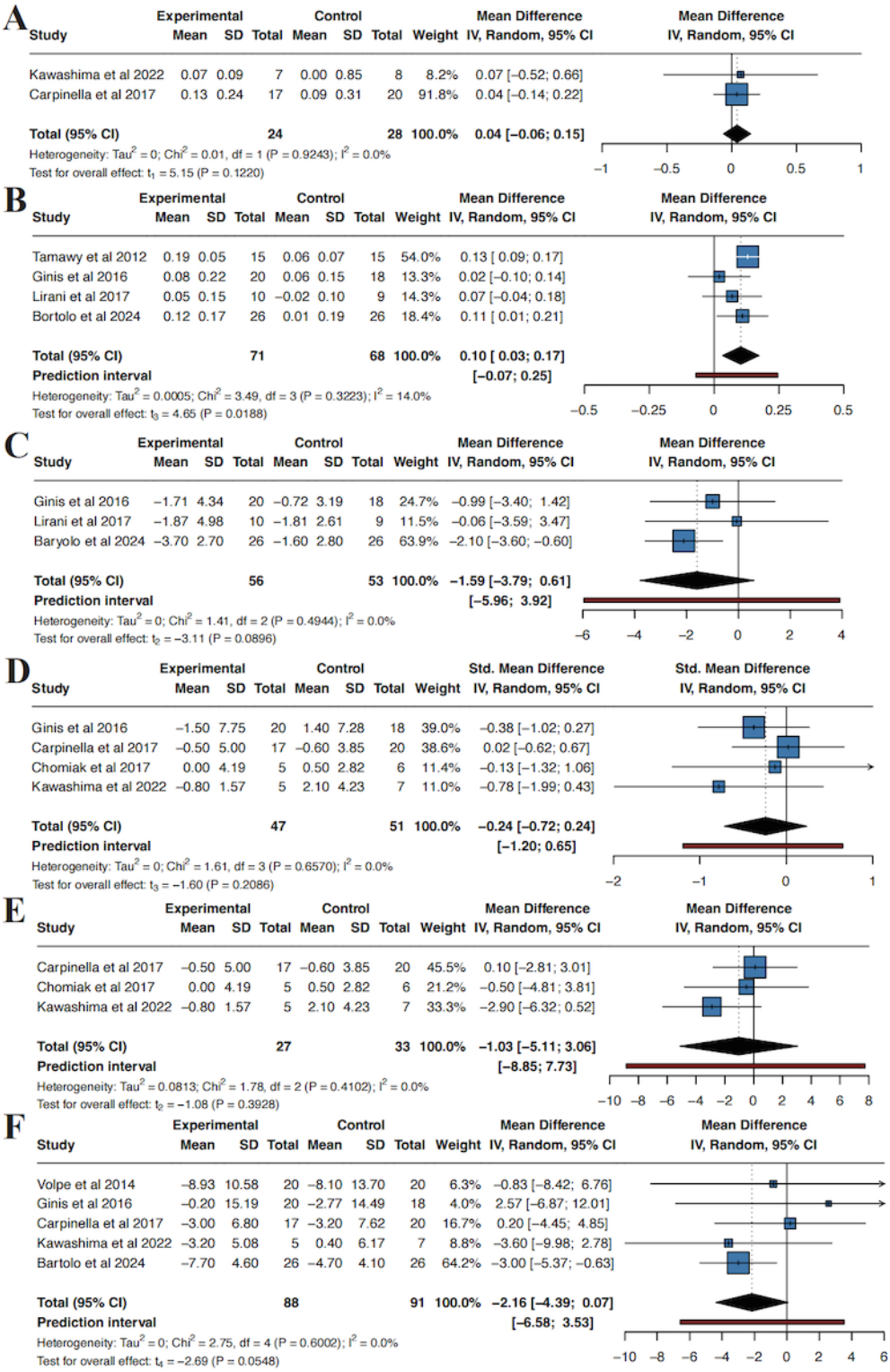

Stride Length

Four studies [40,41,43,46] evaluated stride length (wearable-device group n=71; control group n=68). Between-study heterogeneity was low (τ^2=0.0005; τ=0.02; I^2=14.0%; Q test *P*=.32). Based on our prespecified random-effects model (REML with Hartung–Knapp adjustment), wearable-device interventions were associated with a small but statistically significant improvement on average in stride length compared with controls (MD 0.10 m, 95% CI 0.03-0.17; *P*=.02). The 95% prediction interval was –0.07 to 0.25 m (Figure 3B).

XSL•FO
RenderX

DST

Three studies [41,43,46] evaluated DST (wearable-device group n=56; control group n=53). Between-study heterogeneity was negligible (τ^2=0; τ=0; I^2=0.0%; Q test P=.49). Based on our prespecified random-effects model (REML with Hartung–Knapp adjustment), the pooled analysis showed no clear evidence of a difference in DST between wearable-device interventions and controls (MD –1.59% gait cycle, 95% CI –3.79 to 0.61; P=.09). The 95% prediction interval ranged from –5.96 to 3.92% gait cycle (Figure 3C).

FOG-Q

Four studies [41,42,44,45] evaluated freezing of gait using either the FOG-Q or NFOG-Q (wearable-device group n=47; control group n=51). Between-study heterogeneity was negligible (τ^2=0; τ=0; I^2=0.0%; Q test P=.66). Using our prespecified random-effects model (REML with Hartung–Knapp adjustment), the pooled analysis showed no clear evidence of a difference in freezing of gait between wearable-device interventions and controls (SMD –0.24, 95% CI –0.72 to 0.24; P=.21). The 95% prediction interval was –1.20 to 0.65 (Figure 3D). In sensitivity analyses restricted to studies using the same instrument (FOG-Q only; 3 studies [42,44,45], wearable-device group n=27; control group n=33), heterogeneity remained low (τ^2=0.0813; τ=0.29; I^2=0.0%; Q test P=.41). The pooled effect again showed no statistically significant between-group difference (MD –1.03 points, 95% CI –5.11 to 3.06; P=.39). The 95% prediction interval ranged from –8.85 to 7.73 points (Figure 3E).

UPDRS III

Five studies [41,42,45-47] evaluated UPDRS III (wearable-device group n=88; control group n=91). Between-study heterogeneity was negligible (τ^2=0; τ=0; I^2=0.0%; Q test P=.60). Using the prespecified random-effects model (REML with Hartung–Knapp adjustment), the pooled analysis found no statistically significant difference in UPDRS III between wearable-device interventions and controls (MD –2.16 points, 95% CI –4.39 to 0.07; P=.06). The 95% prediction interval was –6.85 to 3.53 points (Figure 3F).

Balance Function

Three studies [42,45,47] reported results using the BBS, one study [41] reported using the Mini-BESTest, and one study [46] reported using the POMA balance (wearable-device group n=88; control group n=91). Because different scales were used to measure the same construct, results were pooled using SMD. Between-study heterogeneity was low (τ^2=0.0510; I^2=25.5%; Q test P=.25). Using a random-effects model (REML with Hartung–Knapp adjustment), the pooled analysis showed no statistically significant difference in balance between wearable-device interventions and controls (SMD 0.48, 95% CI –0.02 to 0.98; P=.06). The 95% prediction interval ranged from –0.57 to 1.57 (Figure 4A).

XSL•FO
RenderX

Figure 4. Meta-analyses of balance, health-related quality of life, and fear-of-falling outcomes for wearable-device interventions versus control in Parkinson disease. (A) Balance performance expressed as standardized mean differences (SMDs), pooling the Berg Balance Scale (BBS), Mini Balance Evaluation Systems Test (Mini-BESTest), and the balance subscale of the Performance-Oriented Mobility Assessment (POMA balance). (B) Sensitivity analysis for balance restricted to studies using the BBS only, expressed as mean differences (MDs, points). (C) Health-related quality of life assessed with Parkinson Disease Questionnaire (PDQ) instruments, pooling PDQ-39 and PDQ-8 as SMDs. (D) Sensitivity analysis for PDQ restricted to PDQ-39 only, expressed as MDs (points). (E) EQ-5D utility index scores. (F) Fear of falling assessed with the Falls Efficacy Scale–International (FES-I, points) [25,41,42,44-46,48].

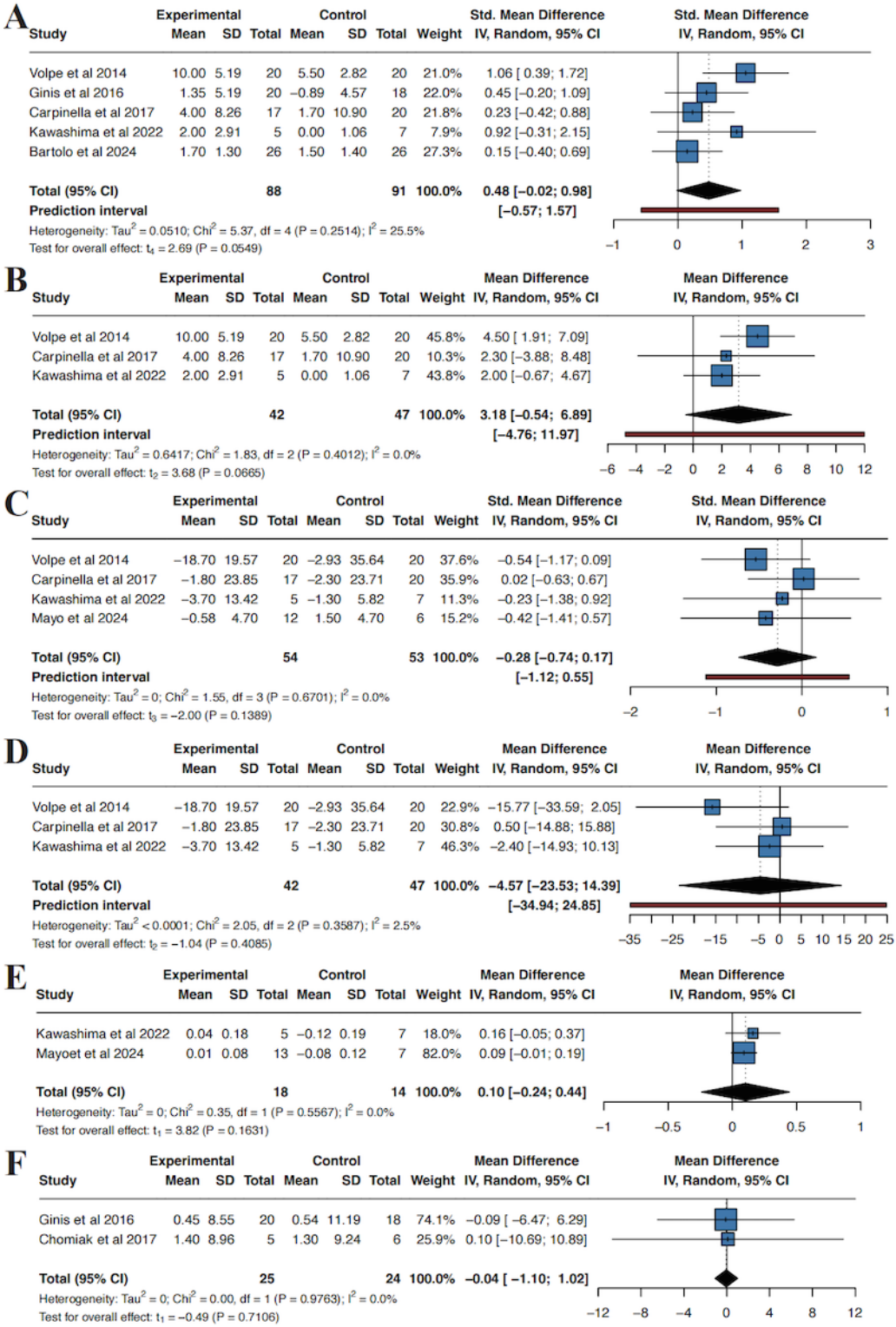

To examine the robustness of this SMD result to measurement heterogeneity, we performed a sensitivity analysis restricted to the three studies using the same balance instrument (BBS; wearable-device group n=42; control group n=47). Heterogeneity remained negligible (τ^2=0.6417; I^2=0%; Q test P=.40). The pooled effect was not statistically significant (MD 3.18 BBS points, 95% CI –0.54 to 6.89; P=.07). The 95%

XSL•FO
RenderX

prediction interval ranged from –4.76 to 11.97 BBS points (Figure 4B).

Quality of Life

PDQ

Four studies [25,42,45,47] assessed quality of life using PDQ instruments. Three studies used the PDQ-39 [42,45,47], whereas 1 study [25] used the PDQ-8 (wearable-device group n=54; control group n=53). Therefore, the primary meta-analysis pooled these outcomes using SMD (Hedges *g*). Between-study heterogeneity was negligible (τ^2=0; τ=0; I^2=0%; Q test *P*=.67). Based on our prespecified random-effects model (REML with Hartung–Knapp adjustment), the pooled analysis showed no statistically significant difference in PDQ scores between wearable-device interventions and controls (SMD –0.28, 95% CI –0.74 to 0.17; *P*=.14). The 95% prediction interval ranged from –1.12 to 0.55 (Figure 4C). To examine the robustness of the pooled effect to measurement heterogeneity, we repeated the analysis after excluding the PDQ-8 study and pooling only PDQ-39 outcomes using MD (wearable-device group n=42; control group n=47). Heterogeneity remained low (τ^2<0.0001; I^2=2.5%; Q test *P*=.36). The pooled result showed no statistically significant difference between groups (MD –4.57, 95% CI –23.53 to 14.39; *P*=.41). The 95% prediction interval was –34.94 to 24.85 (Figure 4D).

EQ-5D Utility

Two studies [25,45] evaluated EQ-5D utility (wearable-device group n=18; control group n=14). Between-study heterogeneity was negligible (τ^2=0; τ=0; I^2=0.0%; Q test *P*=.56). Using the prespecified random-effects model (REML with Hartung–Knapp adjustment), the pooled analysis showed no statistically significant difference in EQ-5D utility between wearable-device interventions and controls (MD 0.10, 95% CI –0.24 to 0.44; *P*=.16; Figure 4E). Prediction intervals were not calculated because fewer than three studies were available.

FES-I

Two studies [41,44] evaluated fear of falling using the FES-I (wearable-device group n=25; control group n=24). Between-study heterogeneity was negligible (τ^2=0; τ=0; I^2=0.0%; Q test *P*=.98). Using the prespecified random-effects model (REML with Hartung–Knapp adjustment), the pooled analysis showed no statistically significant difference in FES-I scores between wearable-device interventions and controls (MD –0.04 points, 95% CI –1.10 to 1.02; *P*=.71; Figure 4F). Prediction intervals were not calculated because fewer than three studies were available.

Certainty of Evidence

According to GRADE, the overall certainty of evidence ranged from low to very low for all outcomes. For stride length, the certainty was rated as low, reflecting consistent direction of effect across four randomized trials and reasonably precise estimates, but downgrading for very serious risk of bias related to unclear allocation concealment, lack of blinding of outcome assessors, and incomplete outcome data. For all other prespecified outcomes—including 10MWT, DST, freezing of gait, UPDRS III, balance scales, PDQ, EQ-5D utility, and FES-I—the certainty of evidence was judged very low. These ratings were primarily driven by very serious risk of bias and very serious imprecision, with small sample sizes per outcome and wide confidence and prediction intervals that encompassed both no effect and potentially important benefit or harm. Inconsistency and indirectness were generally not considered serious (Table 3). Because the number of included studies was small, we could not formally evaluate funnel plot asymmetry; therefore, the possibility of small-study effects could not be ruled out [48].

XSL•FO
RenderX

Table 3. Certainty of evidence assessment of included studies.

Certainty assessment							Number of patients		Effect			
Number of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Intervention	Comparison	Relative (95% CI)	Absolute (95% CI)	Certainty	Importance
10-Meter Walk Test												
2	Randomized trials	Very serious[a]	Not serious	Not serious	Very serious[b]	None	24	28	—[c]	MD[d] 0.04s (95% CI –0.06 to 0.15)	Very low[a,b]	Important
Stride length												
4	Randomized trials	Very serious[a]	Not serious	Not serious	Not serious	None	71	68	—	MD 0.10 m, (95% CI 0.03-0.17)	Low[a]	Important
Double support time												
3	Randomized trials	very serious[a]	Not serious	Not serious	Very serious[b]	None	56	53	—	MD 1.59 % gait cycle, (95% CI –3.79 to 0.61)	Very low[a,b]	Important
Freezing of Gait Questionnaire												
4	Randomized trials	Very serious[a]	Not serious	Not serious	Very serious[b]	None	47	51	—	SMD[e] –0.24, (95% CI –0.72 to 0.24)	Very low[a,b]	Important
Unified Parkinson Disease Rating Scale, Part III												
5	Randomized trials	Very serious[a]	Not serious	Not serious	Serious[b]	None	88	91	—	MD –2.16 points, (95% CI –4.39 to 0.07)	Very low[a,b]	Important
Balance function												
5	Randomized trials	Very serious[a]	Not serious	Not serious	Serious[b]	None	88	91	—	SMD 0.48, (95% CI –0.02 to 0.98)	Very low[a,b]	Important
Parkinson Disease Questionnaire												
4	Randomized trials	Very serious[a]	Not serious	Not serious	Serious[b]	None	55	53	—	SMD –0.28, (95% CI –0.74 to 0.17)	Very low[a,b]	Important
EQ-5D questionnaire												
2	Randomized trials	Very serious[a]	Not serious	Not serious	Very serious[b]	None	18	14	—	MD 0.10, (95% CI –0.24 to 0.44)	Very low[a,b]	Important
Falls Efficacy Scale–International												

XSL•FO
RenderX

Certainty assessment							Number of patients		Effect			
Number of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Intervention	Comparison	Relative (95% CI)	Absolute (95% CI)	Certainty	Importance
2	Randomized trials	Very serious[a]	Not serious	Not serious	Very serious[b]	None	25	24	—	MD –0.04, (95% CI –1.10 to 1.02)	Very low[a,b]	Important

[a]Risk of bias: downgraded 1 level when most contributing trials had unclear or high risk of bias in one or more key domains (eg, allocation concealment, blinding of outcome assessment, incomplete outcome data); downgraded 2 levels when the pooled evidence was dominated by trials at high risk of bias across multiple key domains.

[b]Imprecision: downgraded 1 level when the total information size was small and the 95% CI included the null effect; downgraded 2 levels when the total information size was very small and the 95% CI was wide, spanning both clinically important benefit and no effect (or harm).

[c]not applicable.

[d]MD: mean difference.

[e]SMD: standardized mean difference.

Discussion

Interpretation

This systematic review and meta-analysis synthesized RCTs evaluating the effects of wearable-device interventions on gait, balance, and quality of life in people with PD, interpreting the findings within the ICF framework [49]. We found a small but statistically significant improvement on average in stride length; however, most other outcomes—including short-distance gait speed, double support time, freezing of gait, balance scales, health-related quality of life, and fear of falling—did not show clear between-group differences. These results suggest that current wearable technologies may have only modest effects on specific gait parameters at the level of body functions and activities, with insufficient evidence for consistent improvement in balance or overall quality of life. Importantly, prediction intervals for several outcomes crossed the null, indicating that effects may vary across populations and implementation contexts and may be minimal or absent in some settings.

Regarding gait and motor outcomes, wearable devices produced a small improvement on average in stride length with low statistical heterogeneity, although the prediction interval suggests that effects may be minimal in some settings, consistent with the idea that cueing and feedback can help increase step amplitude in PD. In contrast, we did not observe clear between-group differences in 10-meter walk performance, double support time, freezing-of-gait scores, or UPDRS III, and the corresponding prediction intervals generally spanned both small benefits and no effect. Overall, these results point to at most modest and uncertain benefits of wearable-device interventions on gait and motor severity. Our findings on short-distance gait performance differ from those of a previous meta-analysis by Zhang et al [30], which reported a statistically significant improvement in gait speed with wearable cueing devices. One likely explanation is that Zhang et al [30] pooled several conceptually different outcomes (10MWT, 6-minute walk test, and treadmill-based speed measures) under a single “gait velocity” construct, whereas we analyzed the 10-Meter Walk Test separately and did not combine it with endurance or treadmill outcomes [30]. In addition, we used a more conservative random-effects approach (REML with Hartung–Knapp adjustment and prediction intervals [36]), which yields wider uncertainty intervals and may partly explain why the apparent speed benefit did not remain statistically significant or robust across settings in our analyses. The small but statistically significant improvement in stride length, in the absence of clear effects on 10MWT or freezing measures, may reflect that step amplitude is more directly modulated by external cueing and feedback than more complex phenomena such as gait initiation, freezing episodes, or sustained walking speed; moreover, many participants were in Hoehn and Yahr stages 1-3 with relatively preserved gait, which may limit the observable incremental benefit of wearable devices on standard clinical tests [50].

For balance-related outcomes, this meta-analysis is, to our knowledge, the first to synthesize randomized evidence on the effects of wearable-device interventions on balance function in people with PD. We pooled BBS, Mini-BESTest, and POMA balance subscores using SMDs, as they reflect a common construct of postural control within the ICF domain of body functions and activities. The combined analysis suggested a borderline improvement in balance favoring wearable devices, but the confidence and prediction intervals were wide and included no effect, indicating substantial uncertainty. A key contextual factor is that, among the included trials, only one study explicitly designed the wearable intervention as a balance-focused training program, whereas the others primarily targeted gait initiation or step regulation; thus, the cueing content and training priorities may not have been optimal for producing measurable changes on balance scales. In contrast, a meta-analysis in a nonspecific older-adult population by Mao et al [51] reported clearer balance gains from sensor-based interventions, suggesting that differences in underlying pathology, intervention content, and training dose between PD cohorts and general older adults may be important.

Quality-of-life and fear-of-falling outcomes map more closely to the ICF domains of participation and contextual or personal factors [49]. Across trials, wearable-device interventions did

XSL•FO
RenderX

not result in statistically significant improvements in PDQ scores, EQ-5D utility, or FES-I scores, and, where calculated, prediction intervals were wide or included no effect. This pattern suggests that short-term, device-focused interventions may not be sufficient to translate modest improvements in gait or motor scores into perceived gains in overall quality of life or reduced fear of falling. Nonmotor symptoms, psychosocial and environmental factors, and real-world participation constraints were rarely targeted explicitly in the included trials, which may further limit the impact of wearable devices on these broader patient-centered outcomes. From a clinical perspective, these findings underscore that wearable devices are unlikely to replace comprehensive multidisciplinary rehabilitation; rather, they may serve as adjunct tools within a broader program that also addresses cognition, mood, balance confidence, and environmental adaptation [52].

When the findings were interpreted in light of heterogeneity, risk of bias, and GRADE assessments, the overall certainty of evidence for most outcomes was low to very low. Although we restricted inclusion to RCTs, several studies had unclear or high risk of bias in domains such as allocation concealment, blinding of outcome assessment, and incomplete outcome data. Many meta-analyses included only 2-5 trials, leading to wide confidence and, where applicable, prediction intervals that encompassed both no effect and potentially clinically relevant benefit or harm. This imprecision, together with some inconsistency in results and indirectness arising from heterogeneous devices and protocols, resulted in downgrading of GRADE ratings for multiple outcomes [39]. Consequently, our conclusions should be interpreted with caution and considered hypothesis-generating rather than definitive.

Limitations of Evidence

This study has several limitations that should be considered when interpreting the results. First, the number of eligible trials and the total sample size for each outcome were modest, limiting statistical power and precision, especially for balance, quality-of-life, and fear-of-falling measures. Second, there was substantial diversity in the technical characteristics and application methods of the wearable devices (eg, cueing modality, feedback algorithms, and sensor placement) and in training protocols (setting, frequency, and duration), which may have diluted or masked device-specific effects. Third, most participants were in Hoehn and Yahr stages 1-3, with relatively preserved motor function and independence, restricting the generalizability of our findings to individuals with more advanced disease. Fourth, intervention periods and follow-up durations were relatively short, so the long-term sustainability of any observed benefits remains unknown. Finally, the number of studies per outcome was too small to formally evaluate funnel plot asymmetry; therefore, small-study effects cannot be ruled out.

Limitations of Review Processes

Several limitations of our review processes should be acknowledged. First, we restricted eligibility to English-language reports, which may have introduced language bias and could have led to missed trials published in other languages. Second, our search did not include gray literature sources (eg, preprint servers, theses, conference proceedings, or general search engines), and we did not perform forward citation tracking; consequently, relevant unpublished or difficult-to-index studies may not have been identified. Third, because all outcomes were informed by fewer than 10 studies, formal assessments of small-study effects (eg, funnel plots and Egger test) were not feasible; therefore, reporting bias due to missing results cannot be excluded. Finally, although we contacted study authors when key information was unclear or missing, incomplete responses may have limited data availability for some syntheses, potentially contributing to imprecision.

Implications

This review provides an updated and methodologically rigorous synthesis of RCT evidence on wearable devices for gait, balance, and quality of life in Parkinson disease. Innovatively, we integrated the ICF framework with contemporary random-effects meta-analysis (Hartung–Knapp adjustment and prediction intervals) and GRADE to contextualize both the pooled average effects and the expected range of effects across real-world settings. Compared with prior reviews, we avoided pooling conceptually different gait outcomes into a single construct, enabling a more implementation-relevant interpretation of heterogeneity and generalizability. This work contributes to the field by clarifying where evidence is most consistent (eg, stride length) and where effects remain uncertain across settings, particularly for balance and patient-centered outcomes. In real-world practice, wearable technologies are best viewed as adjuncts rather than stand-alone rehabilitation strategies and may need to be integrated with task-specific gait and balance training, behavioral strategies, and environmental modifications to translate modest gait gains into meaningful improvements in confidence and quality of life. Future trials should enroll larger and more diverse populations, extend intervention and follow-up periods, standardize ICF-aligned outcomes, and transparently report device characteristics and implementation strategies to clarify for whom, under what conditions, and how wearables can be optimally integrated into routine care.

Conclusion

This systematic review and meta-analysis of randomized controlled trials suggests that wearable-device interventions may provide modest benefits on average for specific gait parameters in PD, particularly stride length, but do not consistently improve short-distance gait speed, double support time, freezing of gait, balance performance, health-related quality of life, or fear of falling. Within the ICF framework, current evidence indicates that wearable devices may support selected body functions and activities, yet there is insufficient and uncertain evidence that they reliably enhance broader participation or overall quality of life.

When interpreted alongside heterogeneity estimates, prediction intervals, risk-of-bias assessments, and GRADE ratings, the certainty of evidence for most outcomes is low to very low. Many meta-analyses were based on a small number of heterogeneous trials with modest sample sizes and relatively short intervention and follow-up durations, leading to wide ranges of plausible effects across settings, including minimal or no benefit. Accordingly, our findings should be viewed as

XSL•FO
RenderX

cautious and hypothesis-generating rather than definitive and do not currently justify the use of wearable devices as a stand-alone rehabilitation strategy in PD.

From a clinical perspective, wearable technologies are best considered as adjuncts to comprehensive multidisciplinary rehabilitation. Combining wearable cueing and feedback with task-specific gait and balance training, education, and psychosocial and environmental interventions may be required to translate small gains in gait parameters into meaningful improvements in balance confidence, participation, and quality of life. Future trials should recruit larger and more diverse patient populations, including those with more advanced disease; standardize outcome measures across ICF domains; extend intervention and follow-up periods; and transparently report device characteristics and implementation strategies to clarify which patients benefit most and how wearable devices can be optimally integrated into routine care.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The analytic code used for the meta-analyses and figure generation and the extracted data underlying the analyses are available from the corresponding author upon reasonable request. The data collection forms and additional materials are available upon request.

Acknowledgments

Generative artificial intelligence (AI; OpenAI; ChatGPT) was used solely to support language editing (eg, grammar checking and rewording for clarity and consistency). It was not used to generate study data, perform analyses, interpret results, or draw conclusions. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the manuscript.

Authors' Contributions

JW, WZ, MF, NX, SW, and QZ contributed to the conceptualization and methodology of the study. JW, WZ, MF, NX, SW, and QZ conducted the investigation. JW, WZ, MF, NX, and QZ performed the formal analysis. WZ, SW, and QZ provided supervision. All authors reviewed and approved the final manuscript.

Conflicts of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Multimedia Appendix 1
Full electronic search strategies for all databases and registers, excluded full-text articles with reasons, and PRISMA-S checklists.
[DOC File , 117 KB - jmir_v28i1e85914_app1.doc]

Multimedia Appendix 2
The PRISMA 2020 statement.
[DOCX File , 22 KB - jmir_v28i1e85914_app2.docx]

References

1. Fu Y, Zhou L, Li H, Hsiao JT, Li B, Tanglay O, et al. Adaptive structural changes in the motor cortex and white matter in Parkinson's disease. Acta Neuropathol 2022;144(5):861-879 [FREE Full text] [doi: 10.1007/s00401-022-02488-3] [Medline: 36053316]
2. GBD 2016 Parkinson's Disease Collaborators. Global, regional, and national burden of Parkinson's disease, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol 2018;17(11):939-953 [FREE Full text] [doi: 10.1016/S1474-4422(18)30295-3] [Medline: 30287051]
3. Dorsey ER, Sherer T, Okun MS, Bloem BR. The emerging evidence of the Parkinson pandemic. J Parkinsons Dis 2018;8(s1):S3-S8 [FREE Full text] [doi: 10.3233/JPD-181474] [Medline: 30584159]
4. Simon DK, Tanner CM, Brundin P. Parkinson disease epidemiology, pathology, genetics, and pathophysiology. Clin Geriatr Med 2020;36(1):1-12 [FREE Full text] [doi: 10.1016/j.cger.2019.08.002] [Medline: 31733690]
5. Foltynie T, Bruno V, Fox S, Kühn AA, Lindop F, Lees AJ. Medical, surgical, and physical treatments for Parkinson's disease. Lancet 2024;403(10423):305-324. [doi: 10.1016/S0140-6736(23)01429-0] [Medline: 38245250]
6. Morris HR, Spillantini MG, Sue CM, Williams-Gray CH. The pathogenesis of Parkinson's disease. Lancet 2024;403(10423):293-304. [doi: 10.1016/S0140-6736(23)01478-2] [Medline: 38245249]
7. Lei J, Tang L, You H. Pathological pain: non-motor manifestations in Parkinson disease and its treatment. Neurosci Biobehav Rev 2024;161:105646. [doi: 10.1016/j.neubiorev.2024.105646] [Medline: 38569983]

XSL•FO
RenderX

8. Pang MY. Physiotherapy management of Parkinson's disease. J Physiother 2021;67(3):163-176 [FREE Full text] [doi: 10.1016/j.jphys.2021.06.004] [Medline: 34154949]
9. Jost ST, Aloui S, Evans J, Ashkan K, Sauerbier A, Rizos A, et al. Neurostimulation for advanced parkinson disease and quality of life at 5 years: a nonrandomized controlled trial. JAMA network open 2024;7:e2352177. [doi: 10.1101/2023.06.07.23291019]
10. Tosin MH, Goetz CG, Stebbins GT. Patient with Parkinson disease and care partner perceptions of key domains affecting health-related quality of life: systematic review. Neurology 2024;102(3):e208028 [FREE Full text] [doi: 10.1212/WNL.0000000000208028] [Medline: 38215353]
11. Poewe W, Espay AJ. Long duration response in Parkinson's disease: levodopa revisited. Brain 2020;143(8):2332-2335. [doi: 10.1093/brain/awaa226] [Medline: 32844192]
12. Ramesh V, Bilal E. Detecting motor symptom fluctuations in Parkinson's disease with generative adversarial networks. NPJ Digit Med 2022;5(1):138 [FREE Full text] [doi: 10.1038/s41746-022-00674-x] [Medline: 36085350]
13. Boelens Keun JT, Arnoldussen IA, Vriend C, van de Rest O. Dietary approaches to improve efficacy and control side effects of levodopa therapy in Parkinson's disease: a systematic review. Adv Nutr 2021;12(6):2265-2287 [FREE Full text] [doi: 10.1093/advances/nmab060] [Medline: 34113965]
14. Bowman T, Pergolini A, Carrozza MC, Lencioni T, Marzegan A, Meloni M, et al. Wearable biofeedback device to assess gait features and improve gait pattern in people with Parkinson's disease: a case series. J Neuroeng Rehabil 2024;21(1):110 [FREE Full text] [doi: 10.1186/s12984-024-01403-z] [Medline: 38926876]
15. Sapienza S, Tsurkalenko O, Giraitis M, Mejia AC, Zelimkhanov G, Schwaninger I, et al. Assessing the clinical utility of inertial sensors for home monitoring in Parkinson's disease: a comprehensive review. NPJ Parkinsons Dis 2024;10(1):161 [FREE Full text] [doi: 10.1038/s41531-024-00755-6] [Medline: 39164257]
16. Kim J, Porciuncula F, Yang HD, Wendel N, Baker T, Chin A, et al. Soft robotic apparel to avert freezing of gait in Parkinson's disease. Nat Med 2024;30(1):177-185. [doi: 10.1038/s41591-023-02731-8] [Medline: 38182783]
17. Babu M, Lautman Z, Lin X, Sobota MH, Snyder MP. Wearable devices: implications for precision medicine and the future of health care. Annu Rev Med 2024;75(1):401-415 [FREE Full text] [doi: 10.1146/annurev-med-052422-020437] [Medline: 37983384]
18. Olsen RJ, Hasan SS, Woo JJ, Nawabi DH, Ramkumar PN. The fundamentals and applications of wearable sensor devices in sports medicine: a scoping review. Arthroscopy 2025;41(2):473-492. [doi: 10.1016/j.arthro.2024.01.042] [Medline: 38331364]
19. Ginsburg GS, Picard RW, Friend SH. Key issues as wearable digital health technologies enter clinical care. N Engl J Med 2024;390(12):1118-1127. [doi: 10.1056/NEJMra2307160] [Medline: 38507754]
20. Longhini J, Marzaro C, Bargeri S, Palese A, Dell'Isola A, Turolla A, et al. Wearable devices to improve physical activity and reduce sedentary behaviour: an umbrella review. Sports Med Open 2024;10(1):9 [FREE Full text] [doi: 10.1186/s40798-024-00678-9] [Medline: 38219269]
21. Ma Y, Zhang Y, Li R, Cheng W, Wu F. The experience and perception of wearable devices in Parkinson's disease patients: a systematic review and meta-synthesis of qualitative studies. J Neurol 2025;272(5):350. [doi: 10.1007/s00415-025-13085-1] [Medline: 40252116]
22. Caroppo A, Manni A, Rescio G, Carluccio AM, Siciliano PA, Leone A. Movement disorders and smart wrist devices: a comprehensive study. Sensors (Basel) 2025;25(1):266 [FREE Full text] [doi: 10.3390/s25010266] [Medline: 39797057]
23. Chatzaki C, Skaramagkas V, Tachos N, Christodoulakis G, Maniadi E, Kefalopoulou Z, et al. The smart-insole dataset: gait analysis using wearable sensors with a focus on elderly and Parkinson's patients. Sensors (Basel) 2021;21(8):2821 [FREE Full text] [doi: 10.3390/s21082821] [Medline: 33923809]
24. Espay AJ, Hausdorff JM, Sánchez-Ferro Á, Klucken J, Merola A, Bonato P, Movement Disorder Society Task Force on Technology. A roadmap for implementation of patient-centered digital outcome measures in Parkinson's disease obtained using mobile health technologies. Mov Disord 2019;34(5):657-663 [FREE Full text] [doi: 10.1002/mds.27671] [Medline: 30901495]
25. Mayo NE, Mate KKV, Fellows LK, Morais JA, Sharp M, Lafontaine A, et al. Real-time auditory feedback for improving gait and walking in people with Parkinson's disease: a pilot and feasibility trial. Pilot Feasibility Stud 2024;10(1):115 [FREE Full text] [doi: 10.1186/s40814-024-01542-z] [Medline: 39192343]
26. Moreau C, Rouaud T, Grabli D, Benatru I, Remy P, Marques A, et al. Overview on wearable sensors for the management of Parkinson's disease. NPJ Parkinsons Dis 2023;9(1):153 [FREE Full text] [doi: 10.1038/s41531-023-00585-y] [Medline: 37919332]
27. Warmerdam E, Hausdorff JM, Atrsaei A, Zhou Y, Mirelman A, Aminian K, et al. Long-term unsupervised mobility assessment in movement disorders. Lancet Neurol 2020;19(5):462-470. [doi: 10.1016/S1474-4422(19)30397-7] [Medline: 32059811]
28. Maddumage Dona SP, Chand PB, Haden C, Kalyani N, Lehn A, Sullivan K, et al. Effectiveness of haptic cues on gait in people with Parkinson's disease - a systematic review and meta-analysis. Gait Posture 2025;122:358-374 [FREE Full text] [doi: 10.1016/j.gaitpost.2025.08.069] [Medline: 40850056]

XSL•FO
RenderX

29. Rovini E, Maremmani C, Cavallo F. How wearable sensors can support Parkinson's disease diagnosis and treatment: a systematic review. Front Neurosci 2017;11:555 [FREE Full text] [doi: 10.3389/fnins.2017.00555] [Medline: 29056899]
30. Zhang T, Meng D, Lyu D, Fang B. The efficacy of wearable cueing devices on gait and motor function in Parkinson disease: a systematic review and meta-analysis of randomized controlled trials. Arch Phys Med Rehabil 2024;105(2):369-380. [doi: 10.1016/j.apmr.2023.07.007] [Medline: 37532166]
31. Sigcha L, Borzì L, Amato F, Rechichi I, Ramos-Romero C, Cárdenas A, et al. Deep learning and wearable sensors for the diagnosis and monitoring of Parkinson's disease: a systematic review. Expert Systems with Applications 2023;229:120541. [doi: 10.1016/j.eswa.2023.120541]
32. Jaiswal A, Paramasivam A, Budhiraja S, Santhakumaran P, Gravel C, Martin J, et al. The international classification of functioning, disability and health (ICF) core sets for deafblindness, part II of the systematic review: linking data to the ICF categories. Eur J Phys Rehabil Med 2024;60(5):893-902 [FREE Full text] [doi: 10.23736/S1973-9087.24.07984-X] [Medline: 39235255]
33. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. bmj 2021;372:n71-n71 [FREE Full text] [doi: 10.31222/osf.io/jb4dx] [Medline: 33782057]
34. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods 2010;1(2):97-111. [doi: 10.1002/jrsm.12] [Medline: 26061376]
35. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023;12(4):101014 [FREE Full text] [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
36. IntHout J, Ioannidis JP, Borm GF. The hartung-knapp-sidik-jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014;14:25 [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
37. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]
38. IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open 2016;6(7):e010247 [FREE Full text] [doi: 10.1136/bmjopen-2015-010247] [Medline: 27406637]
39. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336(7650):924-926 [FREE Full text] [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
40. El-Tamawy M, Darwish M, Khallaf M. Effects of augmented proprioceptive cues on the parameters of gait of individuals with Parkinson's disease. Ann Indian Acad Neurol 2012;15(4):267-272 [FREE Full text] [doi: 10.4103/0972-2327.104334] [Medline: 23349591]
41. Ginis P, Nieuwboer A, Dorfman M, Ferrari A, Gazit E, Canning CG, et al. Feasibility and effects of home-based smartphone-delivered automated feedback training for gait in people with Parkinson's disease: a pilot randomized controlled trial. Parkinsonism Relat Disord 2016;22:28-34 [FREE Full text] [doi: 10.1016/j.parkreldis.2015.11.004] [Medline: 26777408]
42. Carpinella I, Cattaneo D, Bonora G, Bowman T, Martina L, Montesano A, et al. Wearable sensor-based biofeedback training for balance and gait in Parkinson disease: a pilot randomized controlled trial. Arch Phys Med Rehabil 2017;98(4):622-630.e3 [FREE Full text] [doi: 10.1016/j.apmr.2016.11.003] [Medline: 27965005]
43. Lirani-Silva E, Vitório R, Barbieri FA, Orcioli-Silva D, Simieli L, Gobbi LTB. Continuous use of textured insole improve plantar sensation and stride length of people with Parkinson's disease: a pilot study. Gait Posture 2017;58:495-497. [doi: 10.1016/j.gaitpost.2017.09.017] [Medline: 28946113]
44. Chomiak T, Watts A, Meyer N, Pereira FV, Hu B. A training approach to improve stepping automaticity while dual-tasking in Parkinson's disease: a prospective pilot study. Medicine (Baltimore) 2017;96(5):e5934 [FREE Full text] [doi: 10.1097/MD.0000000000005934] [Medline: 28151878]
45. Kawashima N, Hasegawa K, Iijima M, Nagami K, Makimura T, Kumon A, et al. Efficacy of wearable device gait training on Parkinson's disease: a randomized controlled open-label pilot study. Intern Med 2022;61(17):2573-2580 [FREE Full text] [doi: 10.2169/internalmedicine.8949-21] [Medline: 35135928]
46. Bartolo M, Castelli A, Calabrese M, Buttacchio G, Zucchella C, Tamburin S, et al. A wearable system for visual cueing gait rehabilitation in Parkinson's disease: a randomized non-inferiority trial. Eur J Phys Rehabil Med 2024;60(2):245-256. [doi: 10.23736/S1973-9087.24.08381-3] [Medline: 38483335]
47. Volpe D, Giantin MG, Fasano A. A wearable proprioceptive stabilizer (Equistasi®) for rehabilitation of postural instability in Parkinson's disease: a phase II randomized double-blind, double-dummy, controlled study. PLoS One 2014;9(11):e112065 [FREE Full text] [doi: 10.1371/journal.pone.0112065] [Medline: 25401967]
48. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011;343:d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]

XSL•FO
RenderX

49. World Health Organization. International classification of functioning, disability and health: ICF. In: International Classification of Functioning, Disability and Health. Geneva: World Health Organization; 2001.
50. Hoehn MM, Yahr MD. Parkinsonism: onset, progression and mortality. Neurology 1967;17(5):427-442. [doi: 10.1212/wnl.17.5.427] [Medline: 6067254]
51. Mao Q, Zhang J, Yu L, Zhao Y, Luximon Y, Wang H. Effectiveness of sensor-based interventions in improving gait and balance performance in older adults: systematic review and meta-analysis of randomized controlled trials. J Neuroeng Rehabil 2024;21(1):85. [doi: 10.1186/s12984-024-01375-0] [Medline: 38807117]
52. Nielsen C, Siersma V, Ghaziani E, Beyer N, Magnusson SP, Couppé C. Health-related quality of life and physical function in individuals with Parkinson's disease after a multidisciplinary rehabilitation regimen-a prospective cohort feasibility study. Int J Environ Res Public Health 2020;17(20):7668 [FREE Full text] [doi: 10.3390/ijerph17207668] [Medline: 33096677]

Abbreviations

10MWT: 10-Meter Walk Test
BBS: Berg Balance Scale
DST: double support time
FES-I: Falls Efficacy Scale-International
FOG-Q: Freezing of Gait Questionnaire
GRADE: Grading of Recommendations Assessment, Development and Evaluation
ICF: International Classification of Functioning, Disability and Health
IMU: inertial measurement unit
MD: mean difference
MeSH: Medical Subject Headings
Mini-BESTest: Mini Balance Evaluation Systems Test
NFOG-Q: New Freezing of Gait Questionnaire
PD: Parkinson disease
PDQ: Parkinson Disease Questionnaire
PDQ-8: 8-item Parkinson Disease Questionnaire
PDQ-39: 39-item Parkinson Disease Questionnaire
PICO: Population, Intervention, Comparison, and Outcome
POMA balance: Performance-Oriented Mobility Assessment balance subscore
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
PROSPERO: International Prospective Register of Systematic Reviews
RCT: randomized controlled trial
REML: restricted maximum likelihood
RoB: Risk of Bias
SMD: standardized mean difference
UPDRS III: Unified Parkinson Disease Rating Scale Part III

Edited by S Brini; submitted 15.Oct.2025; peer-reviewed by O Enaworu, CY Su; comments to author 14.Nov.2025; accepted 28.Dec.2025; published 18.Mar.2026.

Please cite as:
Wu J, Zang W, Fang M, Xiao N, Zhang X, Wang S, Zhang Q
Effects of Wearable Devices on Parkinson Disease: Systematic Review and Meta-Analysis of Randomized Controlled Trials Within the International Classification of Functioning, Disability, and Health Framework
J Med Internet Res 2026;28:e85914
URL: https://www.jmir.org/2026/1/e85914
doi:10.2196/85914
PMID:

©Jiarong Wu, Wanli Zang, Mingqing Fang, Ningkun Xiao, Xianzuo Zhang, Su Wang, Qiuxia Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet

XSL•FO
RenderX

Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Determinants and Health Outcomes of Digital Health Literacy in Patients With Cardiovascular Disease: Systematic Review and Meta-Analysis

Eun-Jung Kim[1], RN, MSN; Da-Young Kim[1], RN, MSN; Ha-Jin Kim[2], RN, BSN; Youn-Jung Son[3], RN, CCAPN, PhD

[1]Graduate School of Nursing, Chung-Ang University, Seoul, Republic of Korea
[2]Department of Nursing, Asan Medical Center, Seoul, Republic of Korea
[3]Red Cross College of Nursing, Chung-Ang University, Seoul, Republic of Korea

Corresponding Author:
Youn-Jung Son, RN, CCAPN, PhD
Red Cross College of Nursing
Chung-Ang University
Building 103
84 Heuksukro, Dongjak-gu
Seoul, 06974
Republic of Korea
Phone: 82 028205198
Email: yjson@cau.ac.kr

Abstract

Background: With expansion of technology-enabled care, digital health literacy (DHL) has become integral to effective cardiovascular disease (CVD) management. However, quantitative evidence regarding determinants and health outcomes of DHL in CVD remains limited and heterogeneous, necessitating comprehensive evidence synthesis.

Objective: This study aimed to (1) estimate DHL levels, (2) synthesize DHL-associated factors, and (3) examine DHL-related health outcomes in CVD.

Methods: A systematic review and meta-analysis of DHL in adults with CVD was conducted per PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines. PubMed, Embase, Cochrane CENTRAL, CINAHL, Scopus, Web of Science, and Google Scholar were searched for peer-reviewed studies published between 2006 and January 31, 2026. Quantitative studies enrolling adults with CVD, which reported a measure of DHL were included. Studies focusing exclusively on primary cerebrovascular disease and non–peer-reviewed articles were excluded. Risk of bias (ROB) was assessed using the Appraisal Tool for Cross-Sectional Studies tool, the Newcastle-Ottawa Scale, the Revised Cochrane Risk-of-Bias Tool for Randomized Trials, and the Risk of Bias in Nonrandomized Studies of Interventions. Certainty of evidence was evaluated using the Grading of Recommendations Assessment, Development, and Evaluation approach. Pooled mean eHealth Literacy Scale (eHEALS) scores were synthesized using a random-effects meta-analysis. Heterogeneity was quantified using the I^2 statistic and 95% prediction intervals.

Results: Twenty studies involving 8581 adults with CVD were included. The overall pooled mean eHEALS score was 24.26 (95% CI 21.19-27.32), with substantial heterogeneity (I^2=98.4%; τ^2=15.55; τ=3.94) and a wide 95% prediction interval (14.66-33.85). Lower DHL was consistently associated with older age, lower educational attainment, female sex, limited social support, and less experience with digital technologies. Higher DHL was associated with more favorable health-related outcomes, including health behaviors, better quality of life, and greater use and acceptance of digital health technologies. Subgroup analyses showed no statistically significant differences in DHL by region, disease type, or age group. The certainty of evidence was rated as low to very low, and substantial heterogeneity persisted across analyses.

Conclusions: Our findings underscore DHL as a foundational capability for digitally supported self-management in CVD care and reveal disparities associated with age and socioeconomic factors. By integrating evidence on DHL levels, associated factors, and DHL-related health outcomes in CVD populations, this review provides a more comprehensive, clinically relevant understanding of DHL beyond studies relying on a single instrument (eg, eHEALS) or examining isolated domains. DHL appears to be a

XSL•FO
RenderX

context-dependent competency shaped by broader structural and social determinants. From a clinical and health system perspective, digital health interventions should be accompanied by structured digital inclusion strategies, including routine assessment of DHL and care delivery to patients' digital capacities. Further longitudinal and interventional studies are warranted to clarify the causal pathways linking DHL to health outcomes in adults with CVD and to incorporate provider- and system-level perspectives beyond individual-level assessments.

Trial Registration: PROSPERO International Prospective Register of Systematic Reviews CRD420251068000; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251068000

(J Med Internet Res 2026;28:e89102) doi:10.2196/89102

KEYWORDS

cardiovascular disease; digital health; health literacy; systematic review; meta-analysis

Introduction

Cardiovascular disease (CVD) encompasses disorders of the heart and vasculature, including angina, myocardial infarction, stroke, heart failure, arrhythmia, and valvular heart disease [1]. CVD is typically chronic and requires lifelong management. Effective patient self-management, which includes medication adherence, physical activity, and dietary management, is essential for optimal outcomes [2]. Self-management is an active process in which patients understand their conditions, manage symptoms, adhere to prescribed medications, and modify lifestyle habits. These actions are critical for improving health-related quality of life [3]. Digital health technologies have increasingly been applied to support self-management by enabling real-time monitoring, personalized feedback, and behavioral adjustment [4]. Despite variability in effectiveness across modalities and implementation contexts, the adoption of these technologies continues to expand [5,6]. Digital health technologies facilitate the delivery of personalized health information and enable real-time monitoring and feedback through mobile platforms, wearable devices, and remote patient monitoring systems [7,8]. By leveraging digital infrastructures, these technologies facilitate large-scale dissemination of health information across diverse settings, including medically underserved and resource-limited environments, enhancing their relevance to population and public health contexts [7].

However, substantial differences exist in access to digital technologies, digital capabilities, and trust in digital health systems [7,8]. Older adults and socioeconomically disadvantaged populations are more frequently exposed to environments characterized by limited digital health literacy (DHL), restricted access, and reduced trust, thereby increasing the risk of digital exclusion [8]. In a US study of patients with heart failure, more than half of the participants lacked access to mobile health technologies, and limited access was strongly associated with older age, rural residence, and socioeconomic factors [9]. These findings underscore the presence of digital disparities in both access to and effective use of digital health technologies among patients with CVD [9]. Furthermore, the uptake and utilization of digital health technologies are influenced by social determinants of health, including educational attainment, income, race and ethnicity, and sex. DHL has been identified as a key component of digital health equity [10]. Consequently, effective use of digital health technologies is closely associated with DHL, which may determine whether disparities in access translate into successful digitally supported self-management [11].

With the migration of health information delivery to online and mobile platforms, DHL, an extension of traditional health literacy, has emerged as a central construct [12,13]. DHL refers to the ability to access, understand, evaluate, and apply health information obtained in digital environments to support effective health behaviors and improve or maintain health-related quality of life [14,15]. Higher DHL enables individuals to accurately interpret and use digital health information, promoting more effective self-management [16]. Given the complexity and chronicity of CVD, patients require continuous self-management, for which DHL is a critical resource [17].

Studies assessing DHL in CVD have been conducted across diverse settings and have used heterogeneous instruments, most commonly the eHealth Literacy Scale (eHEALS), the Digital Health Literacy Instrument (DHLI), and the eHealth Literacy Questionnaire (eHLQ). These instruments differ in item content and measurement domains depending on their intended purpose and target population and are used to quantify DHL levels [18-20]. In addition, DHL varies across populations according to age, educational attainment, socioeconomic status, and employment status. Older adults and individuals with lower education or income typically report lower DHL, which is associated with reduced use of digital health information and less effective self-management [16,17].

Supporting effective self-management in CVD requires assessment of individual DHL levels, identification of their determinants, and evaluation of associated health outcomes [21]. Despite the widespread availability of mobile apps, remote monitoring tools, and online resources, individuals with low DHL may be unable to use these resources effectively. This limitation increases exposure to inaccurate information and contributes to distrust or avoidance of digital health resources [22]. Emerging evidence suggests that DHL is associated with health behaviors and outcomes in CVD populations; however, quantitative data on DHL levels, determinants, and associations with clinical and patient-reported outcomes remain limited and heterogeneous [21,23].

This systematic review and meta-analysis aimed to assess DHL levels, synthesize factors associated with DHL, and examine DHL-related health outcomes among patients with CVD.

XSL•FO
RenderX

Methods

Study Design, Protocol, and Registration

This systematic review and meta-analysis synthesized factors associated with DHL and its health outcomes among patients with CVD. The protocol was registered with PROSPERO (CRD420251068000). Reporting followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 expanded checklist, the PRISMA 2020 (checklist provided in Multimedia Appendix 1) for abstracts checklist [24], and PRISMA-S (PRISMA literature search extension) [25].

Two deviations from the registered protocol were made. First, although the PROSPERO record specified inclusion of English-language publications only, no language restrictions were applied during the final search to minimize potential language bias. Second, studies that focused exclusively on stroke or transient ischemic attack were excluded during full-text screening to enhance clinical homogeneity.

Eligibility Criteria

Eligibility criteria were based on the participants, intervention or interest, comparisons, outcomes, and study design (PICOS) framework, as follows: adults aged ≥18 years diagnosed with CVD (participants); studies measuring DHL among patients with CVD (intervention or interest); no comparison group was required, as this review did not focus on comparative studies (comparison); determinants of DHL and its health outcomes (outcomes); and quantitative studies and mixed methods studies containing quantitative data were included (study design). Purely qualitative studies and those reporting only DHL scores without examining associated factors or health outcomes were excluded.

For this review, CVD was operationalized as cardiac disease, including coronary artery disease, myocardial infarction, heart failure, arrhythmias, and valvular disease. Primary cerebrovascular disease, such as stroke and transient ischemic attack, was excluded to minimize confounding of DHL assessment by postevent language or cognitive impairment [26]. To capture DHL comprehensively, related constructs including eHealth literacy, electronic health literacy, and mobile health literacy were also considered [14]. The literature search included studies published from 2006 onward, the year in which the concept of eHealth literacy was first introduced [14,27]. No language restrictions were applied. Exclusion criteria were (1) studies not measuring DHL; (2) studies involving children or adolescents; and (3) non–peer-reviewed articles, unpublished works, research protocols, or conference abstracts (Table 1).

Table 1. Eligibility criteria based on the PICOS[a] framework.

Item	Inclusion criteria	Exclusion criteria
Participants	Studies involving adult patients (≥18 years) with CVD[b]	Studies involving children or adolescents (<18 years) or targeting non-CVD populations
Intervention or interest	Studies measuring DHL[c], including eHealth literacy, electronic health literacy, or mobile health literacy	Studies not measuring DHL or reporting only clinician-reported outcomes
Comparison	Not applicable (no comparator required)	Not applicable
Outcomes	Studies reporting DHL levels, associated factors, or health outcomes related to DHL	Studies not reporting DHL outcomes or related factors
Publication period	Published between January 2006 and January 2026	Published before 2006 or after January 2026
Study designs	Original quantitative studies (cross-sectional, cohort, quasi-experimental, randomized controlled trials, or mixed methods with quantitative outcomes)	Reviews, qualitative-only studies, protocols, conference abstracts, dissertations, editorials, or commentaries

[a]PICOS: participants, intervention or interest, comparison, outcomes, and study design.

[b]CVD: cardiovascular disease.

[c]DHL: digital health literacy.

Information Sources

A systematic literature search on DHL in patients with CVD was conducted in PubMed (via the National Library of Medicine), Embase (via Elsevier), Cochrane Central Register of Controlled Trials (CENTRAL; via EBSCOhost), CINAHL Plus with Full Text (via EBSCOhost), Scopus (via Elsevier), Web of Science (via Clarivate), and Google Scholar. No additional study registries or online resources were searched. Studies published from 2006 through June 30, 2025, were included. To ensure inclusion of the most recent literature, an updated search was conducted on January 31, 2026, covering the period up to that date. Reference lists of all included studies were manually screened for additional relevant publications; however, no further eligible studies were identified. We contacted the corresponding authors of studies with missing or incomplete data to request additional information; however, no additional data were provided.

Search Strategy

A combination of MeSH (Medical Subject Headings) terms, free-text keywords, and Boolean operators was used to ensure comprehensive retrieval. The search strategy was independently developed for this study by synthesizing terms from preliminary literature reviews and MeSH databases, without using pre-established search filters or existing search strings. Search terms were developed based on DHL and CVD. In Google Scholar, where Boolean operators and advanced filters are limited, the main keywords (“digital health literacy or eHealth literacy” AND “cardiovascular disease”) were applied. The

XSL•FO
RenderX

search strategy was developed in consultation with a librarian and reviewed by the research team prior to implementation. Detailed information on the search process is provided in Multimedia Appendix 2.

Study Selection

All search records were imported into EndNote 21, and duplicate records were removed electronically. After removing duplicates, the remaining records were exported to Microsoft Excel for screening. During the initial screening of titles and abstracts, the reviewers manually identified and removed any additional duplicates not detected by the software. Two reviewers (EJK and YJS), both experienced in conducting systematic reviews and meta-analyses, independently assessed the records against the inclusion and exclusion criteria and evaluated full texts for eligibility. Discrepancies were resolved through discussion until consensus was reached, and eligible studies were included in the review.

Data Extraction

Data were extracted on study descriptors (first author, publication year, and country), participant characteristics (sample size, age, and sex), study measures (DHL measurement tools, scale structure and domains, and DHL level), and outcome indicators (factors associated with DHL and related health outcomes). Adjusted estimates for factors associated with DHL and health outcomes were extracted whenever available. When both unadjusted and adjusted results were reported, the final adjusted estimates from multivariable analysis were included. When effect estimates could be derived from reported values, reviewers systematically calculated and extracted them.

Risk of Bias Assessment

Risk of bias (ROB) and methodological quality were independently assessed by two reviewers (EJK and YJS). The Appraisal Tool for Cross-Sectional Studies (AXIS) was used for cross-sectional studies, the Newcastle-Ottawa Scale (NOS) for cohort or case-control studies, the Revised Cochrane Risk-of-Bias Tool for Randomized Trials (RoB 2), and the Risk of Bias in Nonrandomized Studies of Interventions (ROBINS-I) for nonrandomized interventional studies [28-31]. Disagreements were resolved through consensus.

AXIS comprises 20 items covering study design, sampling, analysis, reporting, and discussion, with each item rated as yes, no, or do not know [28]. NOS evaluates selection, comparability, and outcome, awarding stars within each domain; total scores range from 0 to 9, with higher values indicating better quality [29]. RoB 2 appraises 5 domains: the randomization process, deviations from intended interventions, missing outcome data, measurement of outcomes, and selection of the reported result. Each domain is rated as low risk, some concerns, or high risk. Overall risk is classified as low if all domains are low risk, as some concerns if at least one domain raises concerns but none are high risk, and as high if any domain is high risk or multiple domains raise concerns [30]. ROBINS-I assesses 7 domains: confounding, participant selection, intervention classification, deviations from intended interventions, missing data, outcome measurement, and selection of the reported result. Each domain is graded as low, moderate, serious, or critical ROB, with an overall judgement derived accordingly [31].

Data Analysis

When at least two studies reported mean eHEALS scores, the means were meta-analyzed using an inverse-variance–weighted random-effects model [32,33]. To avoid double-counting in these analyses, multiple publications from the same cohort were checked, and country-specific studies with clearly defined samples were prioritized to ensure the independence of observations. For assessing determinants and health outcomes of DHL, effect estimates were extracted where reported. Owing to substantial heterogeneity in outcome definitions and analytical approaches, these findings were synthesized narratively.

Random-effects meta-analyses were conducted using the Hartung-Knapp-Sidik-Jonkman method to obtain more robust CIs. This approach accounts for uncertainty in estimating between-study variance and is particularly appropriate when heterogeneity is high or the number of studies is limited [34]. Heterogeneity was evaluated using Cochran Q and quantified with Higgins I^2; values of 25%, 50%, and 75% were interpreted as low, moderate, and high heterogeneity, respectively [35]. Although I^2 was reported as a descriptive measure of heterogeneity, it was interpreted with caution because it does not reflect the magnitude of variability in true effects across populations or settings.

Therefore, where applicable, 95% prediction intervals were calculated to quantify between-study heterogeneity in a clinically interpretable manner and to facilitate the real-world interpretation of the pooled estimates. Prediction intervals were derived from the random-effects model, incorporating the estimated between-study variance [36].

Meta-regression was prespecified to explore potential sources of heterogeneity; however, it was not conducted owing to the limited number of included studies [37]. Subgroup analyses were performed to examine variability in mean scores according to three predefined criteria: (1) study region (Asia vs Europe), (2) disease type (heart failure only vs mixed CVD), and (3) mean age (<65 years vs ≥65 years). A leave-one-out sensitivity analysis was performed to assess robustness by examining the contribution of each study to the pooled estimate [38]. All analyses were conducted in R (version 4.4.2; R Foundation for Statistical Computing) using the meta and metafor packages [39-41]. Small-study effects were not assessed in meta-analyses including fewer than 10 studies, as funnel plots and related statistical tests, such as Egger regression, have limited power in analyses with a small number of studies [42,43].

Certainty of the Evidence

The certainty of evidence for each key outcome was assessed using the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach. Assessment domains included ROB, inconsistency, indirectness, imprecision, and publication bias. The GRADE approach was applied to associations between DHL and health outcomes. Based on this assessment, the certainty of evidence for each outcome was classified as high, moderate, low, or very low. Two reviewers

XSL•FO
RenderX

(EJK and DYK) independently assessed the certainty of evidence for each outcome, and any disagreements were resolved through consensus [44].

Ethics Approval

The institutional review board of Chung-Ang University approved the study protocol (1041078-20250603-HR-172).

Results

Study Selection

The initial database search, conducted through June 30, 2025, yielded 17,001 records. Screening and eligibility assessment were performed according to prespecified criteria, and the study selection process is summarized in the PRISMA 2020 flow diagram (Figure 1). After exporting records to EndNote 21 and removing duplicates, 13,341 records remained for title and abstract screening. Of these, 1578 records were excluded owing to incompatible study designs (eg, qualitative studies, protocols, and letters), 4943 records were excluded because participants were not adults with CVD, 6615 records were excluded because DHL was not assessed, and 129 records were removed as remaining duplicates. Full texts of 76 studies were assessed for eligibility, and 58 were excluded owing to ineligible study design (eg, conference abstracts and protocols), ineligible population, absence of DHL assessment, or lack of DHL-related factors or health outcomes. Eighteen studies were initially selected from the primary search. To ensure inclusion of the most recent literature, an updated search was conducted in January 2026, yielding 40 additional records. Following the same screening and eligibility process, 38 records were excluded, and 2 additional studies were included. Consequently, a total of 20 studies met the inclusion criteria and were included in this review (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of study selection. This figure illustrates the process of identification, screening, eligibility assessment, and inclusion of studies in the systematic review in accordance with the PRISMA 2020 guidelines. Records were identified through database searching up to June 30, 2025, with an updated search conducted in January 2026. DHL: digital health literacy.

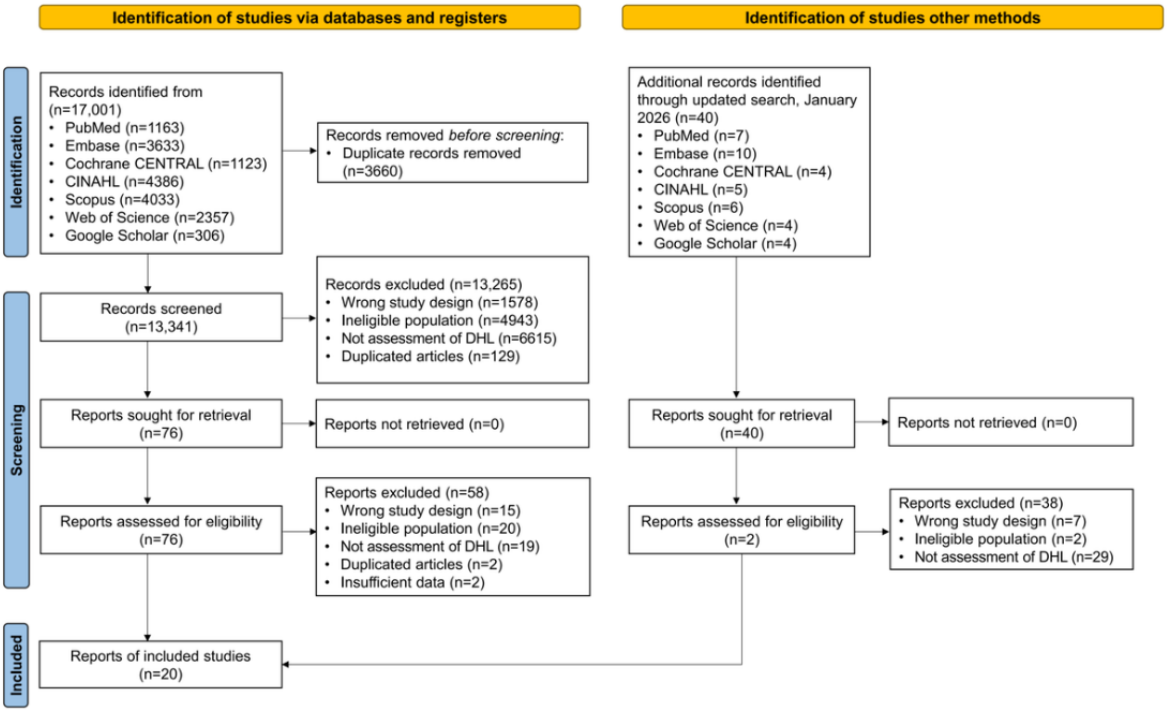

Study Characteristics

The 20 studies included in this review were published from 2018 onward, with 14 published after 2023. Most studies used a cross-sectional design (n=11, 55%), followed by 3 quasi-experimental studies and 2 studies each with randomized controlled trial, longitudinal observational, and prospective observational designs. In 5 experimental and cohort studies [45-49], DHL was assessed at multiple time points. The studies were conducted across several countries, predominantly in Europe and Asia.

Heart failure was the most common condition among the study populations (n=10, 50%). Participants included patients who had undergone coronary artery bypass grafting, those with ischemic heart disease, those who had valve surgery or percutaneous coronary intervention, and those with atrial fibrillation, cardiac arrhythmias with a cardiac implantable electronic device, atherosclerosis, or coronary artery disease. The reported mean age of participants ranged from 57.6 to 79.9 years. Most studies recruited adults aged ≥18 years, although some enrolled specific age groups, such as those aged ≥40 years [49], ≥45 years [50], or ≥60 years [51,52]. The total sample size across all studies was 8581 participants, with proportions of males ranging from 49.5% to 83% and females ranging from 17% to 50.5% (Table 2).

XSL•FO
RenderX

Table 2. Characteristics of included studies (N=20).

Authors (year)	Study design	Country	Health status	Sample size	Age (years)	Sex (female), n (%)
Melholt et al (2018) [45]	Quasi-experimental	Denmark	CABG[a], HF[b], IHD[c], VS[d]	49	• Mean 60.64 (SD 10.75)	9 (18.4)
Chuang et al (2019) [53]	Cross-sectional	Taiwan	HF	141	• Mean 65.2 (SD 11.9)	63 (44.7)
Lin et al (2020) [51]	Longitudinal	Iran	HF	468	• Mean 69.3 (SD 7.3)	230 (49.2)
Rodríguez Parrado et al (2022) [46]	Quasi-experimental	Colombia	HF	28	• Mean 67.3 (SD 12.9)	7 (25.0)
Spindler et al (2022) [47]	Randomized controlled trial	Denmark	HF	137	• Intervention group: mean 61.73 (SD 10.75) • Control group: mean 61.36 (SD 11.46)	32 (23.4)
Yun et al (2022) [54]	Randomized controlled trial	Spain	HF	178	• Median 74	73 (41.0)
Bakhshayesh et al (2023) [55]	Cross-sectional	Iran	HF	200	• ≤50 years: n=45 (22.5%) • >50 years: n=155 (77.5%)	70 (35.0)
Bäuerle et al (2023) [56]	Cross-sectional	Germany	CHF[e], IHD	290	• Mean 57.59 (SD 13.33)	149 (51.4)
Brørs et al (2023) [48]	Prospective	Denmark and Norway	PCI[f]	2924	• Mean 66 (SD 11)	582 (19.9)
Ramstad et al (2023) [57]	Prospective	Norway	PCI	1970	• Mean 66 (SD 11)	427 (21.7)
Rush et al (2023) [58]	Cross-sectional	Canada	AF[g]	195	• Mean 65.36 (SD 10.32)	73 (37.4)
Son et al (2023) [49]	Quasi-experimental	Korea	HF	100	• Mean 58.78 (SD 8.83)	17 (17.0)
Vitolo et al (2023) [59]	Cross-sectional	Italy	Arrhythmia, CIED[h] recipient	300	• Median 75 (IQR 66-84)	109 (36.3)
Mohajeri et al (2024) [60]	Cross-sectional	Germany	Atherosclerosis	310	• Mean 61.72 (SD 11.44)	130 (41.9)
van Schalkwijk et al (2024) [61]	Longitudinal	Netherlands	Arrhythmia, CHD[i], HF	305	• Mean 64.07 (SD 9.98)	112 (36.7)
Astuti et al (2025) [62]	Cross-sectional	Indonesia	HF	119	• Mean 60.18 (SD 13.13)	58 (48.7)
Cuppen et al (2025) [63]	Cross-sectional	Netherlands	HF	61	• Mean 79.9 (SD 9.5)	24 (39.3)
Zhao et al (2025) [52]	Cross-sectional	China	CHD	396	• Mean 69.78 (SD 6.36)	200 (50.5)
Dibek et al (2025) [64]	Cross-sectional	Turkey	HF	250	• ≤60 years: n=76 (30.4%) • >60 years: n=174 (69.6%)	113 (45.2)
Sun et al (2025) [50]	Cross-sectional	China	CHD	594	• 45-59 years: n=203 (34.2%) • 60-74 years: n=264 (44.4%) • ≥75 years: n=127 (21.4%)	235 (42.6)

[a]CABG: coronary artery bypass grafting.

[b]HF: heart failure.

[c]IHD: ischemic heart disease.

[d]VS: valve surgery.

[e]CHF: congestive heart failure.

[f]PCI: percutaneous coronary intervention.

XSL•FO
RenderX

[g]AF: atrial fibrillation.

[h]CIED: cardiac implantable electronic device.

[i]CHD: coronary heart disease.

The most frequently used instrument for assessing DHL was the eHEALS [18] (n=11), followed by the DHLI [19] (n=2) and the eHLQ [20] (n=1). The Digital Health Readiness Questionnaire (DHRQ) [65] was used in one study (n=1). Because newer DHL instruments, such as the eHLQ and DHRQ, incorporate attitudinal components (eg, motivation and attitudes) into the construct of DHL, some studies additionally assessed digital-related attitudes, such as internet anxiety and digital confidence [56,60]. One study used a self-developed instrument [46] (Table 3; Multimedia Appendix 3 [36-48,51,53-56]). In studies with incomplete or unreported essential data [55,56,60], the corresponding authors were contacted to request additional information; however, no responses were received.

XSL•FO
RenderX

Table 3. Digital health literacy measurement tools and levels of the included studies.

Authors (year)	DHL[a] measurement	DHL level
Melholt et al (2018) [45]	eHEALS[b]	• Mean 28.80 (SD 5.96)[c] (baseline)
Chuang et al (2019) [53]	eHEALS	• Mean 26.2 (SD 5.7)
Lin et al (2020) [51]	eHEALS	• Mean 28.16 (SD 5.46)
Rodríguez Parrado et al (2022) [46]	Researcher-developed questionnaire	• Mean 2.33 (SD 1.25)[c] (baseline)
Spindler et al (2022) [47]	eHLQ[d]	• Intervention group: mean 2.98 (SD 3.22) • Control group: mean 2.71 (SD 3.17) (after 6 months)[e]
Yun et al (2022) [54]	Researcher-developed questionnaire	• 72% lower ICT[f] skills (n=128) • 28% middle or higher ICT skills (n=50)
Bakhshayesh et al (2023) [55]	eHEALS	• Mean 18.09 (SD 9.08)
Bäuerle et al (2023) [56]	Internet anxiety, digital confidence, prior experiences with mobile health interventions	• Not reported
Brørs et al (2023) [48]	eHEALS	• Mean 27.27 (SD 6.28) (baseline)
Ramstad et al (2023) [57]	eHEALS	• Mean 25.71 (SD 6.22) (baseline)
Rush et al (2023) [58]	30-item composite tool	• Computer self-efficacy: mean 5.38 (SD 1.27) • Health technology self-efficacy: mean 5.63 (SD 1.01) • Attitude toward health technology: mean 5.46 (SD 0.87)
Son et al (2023) [49]	eHEALS	• Intervention group: mean 24.32 (SD 9.55) • Control group: mean 25.58 (SD 9.79) (baseline)
Vitolo et al (2023) [59]	DHLI[g]	• Mean 48.58 (SD 24.16)
Mohajeri et al (2024) [60]	Internet anxiety, digital confidence, digital overload	• Internet anxiety: mean 1.58 (SD 0.80) • Digital confidence: mean 3.87 (SD 0.98) • Digital overload: mean 1.94 (SD 1.01)
van Schalkwijk et al (2024) [61]	DHLI	• Mean 2.97 (SD 0.60)
Astuti et al (2025) [62]	eHEALS	• Mean 26.35 (SD 7.40)
Cuppen et al (2025) [63]	DHRQ[h]	• Mean 38.4 (SD 17.7)
Zhao et al (2025) [52]	eHEALS	• Mean 18.18 (SD 10.11)
Dibek et al (2025) [64]	eHEALS	• Mean 21.9 (SD 9.7)
Sun et al (2025) [50]	eHEALS	• Median 21 (IQR 13-29)

[a]DHL: digital health literacy.
[b]eHEALS: eHealth Literacy Scale.
[c]Scores calculated by authors.
[d]eHLQ: eHealth Literacy Questionnaire.
[e]Baseline data were not reported in this study.
[f]ICT: information and communications technology.
[g]DHLI: Digital Health Literacy Instrument.
[h]DHRQ: Digital Health Readiness Questionnaire.

XSL•FO
RenderX

Quality Appraisal

The included studies comprised 11 cross-sectional, 2 longitudinal observational, 2 prospective observational, 2 randomized controlled trials, and 3 quasi-experimental studies. ROB and methodological quality were appraised using tools appropriate for each study design, and a summary of these assessments is presented in Table 4.

Table 4. Summary of quality assessment of the included studies[a].

Appraisal tool and study design	Overall quality assessment
AXIS[b]	
Cross-sectional	
Chuang et al (2019) [53]	18/20
Bakhshayesh et al (2023) [55]	17/20
Bäuerle et al (2023) [56]	17/20
Rush et al (2023) [58]	18/20
Vitolo et al (2023) [59]	16/20
Mohajeri et al (2024) [60]	18/20
Astuti et al (2025) [62]	17/20
Cuppen et al (2025) [63]	16/20
Zhao et al (2025) [52]	18/20
Dibek et al (2025) [64]	17/20
Sun et al (2025) [50]	17/20
NOS[c]	
Longitudinal/prospective	
Lin et al (2020) [51]	7/9
Brørs et al (2023) [48]	6/9
Ramstad et al (2023) [57]	5/9
van Schalkwijk et al (2024) [61]	5/9
RoB 2[d]	
Randomized controlled trial	
Spindler et al (2022) [47]	High risk
Yun et al (2022) [54]	Some concerns
ROBINS-I[e]	
Quasi-experimental	
Melholt et al (2018) [45]	Moderate
Rodríguez Parrado et al (2022) [46]	Moderate
Son et al (2023) [49]	Moderate

[a]Methodological quality was assessed using validated appraisal tools appropriate for each study design. Overall quality ratings or scores are reported as provided by each appraisal framework.

[b]AXIS: Appraisal Tool for Cross-Sectional Studies.

[c]NOS: Newcastle-Ottawa Quality Assessment Scale.

[d]RoB 2: Revised Cochrane Risk-of-Bias Tool for Randomized Trials.

[e]ROBINS-I: Risk of Bias in Nonrandomized Studies of Interventions.

The 11 cross-sectional studies were evaluated using the AXIS tool, with quality scores ranging from 16 to 18 out of a maximum of 20 points. However, most studies did not report information on nonresponders or provide response rates. In addition, characteristics of nonresponders or comparisons with responders were not reported (Multimedia Appendix 4 [50,52,53,55,56,58-60,62-64]). The 4 cohort studies (longitudinal, prospective observational studies) were assessed using the NOS, with scores ranging from 5 to 7 out of 9. According to the Agency for Healthcare Research and Quality (AHRQ) thresholds, 2 studies were rated as good quality and 2 as poor quality. The studies rated as poor quality relied on

XSL•FO
RenderX

self-reported outcomes and lacked clear reporting of follow-up rates or attrition (Multimedia Appendix 5 [48,51,57,61]).

The 2 randomized controlled trials were appraised using RoB 2 and were rated as having "some concerns" and "high ROB," respectively. In both studies, the randomization process and blinding procedures were not clearly reported, and insufficient information was available to support a low risk in these domains. Furthermore, neither study provided adequate details regarding missing outcome data (Multimedia Appendix 6 [47,54]). The 3 quasi-experimental studies were evaluated using ROBINS-I and were all judged to have a moderate ROB. Concerns primarily arose in the measurement of the outcomes domain, where outcome assessors may have been aware of participants' intervention status (Multimedia Appendix 7 [45,46,49]).

Factors Associated With DHL

Across the 20 included studies, factors associated with DHL were categorized into 5 domains: demographic (age, sex, education, and residence), clinical (cognitive function, laboratory findings, and disease severity), physical health (functional status and frailty), psychosocial (social support and caregiving needs), and digital technology use (prior experience) (Table 5; Multimedia Appendix 8 [45-49,51-63]). Lower educational attainment (n=6) and older age (n=5) were most frequently associated with lower DHL. Other factors included female sex (n=2), disease severity (n=2), lower or higher social support (n=2), and limited prior experience with digital technologies (n=2).

Table 5. Factors associated with digital health literacy of the included studies.

Category/subcategory	Significant factors[a]	References
Demographic variables		
Age	Older age	[49,50,54,55,64]
Sex	Female	[54,58]
Education	Lower educational attainment	[45,50,54-56,64]
Marital status	Married status	[64]
Economic status	Lower economic status	[64]
Employment status	Unemployment	[64]
Residence	Rural residence	[56]
Clinical variables		
Cognitive function	Lower cognitive function	[54]
Laboratory findings	Lower BMI, DBP[b], Hb[c], NT-proBNP[d]	[54]
Diagnosis duration	>6 years	[64]
Disease severity	Higher NYHA[e] (or HF stage) functional class	[54,64]
Prior cardiac intervention	Absence of PCI or CABG	[50]
Physical health		
Functional status	Greater functional dependence	[54]
Frailty	Higher frailty status	[59]
Smoking status	Smokers	[64]
Psychosocial variables		
Health behaviors	Lower health behaviors	[64]
Social support	Higher social support	[54]
Social support	Lower social support	[52]
Caregiving needs	Greater caregiver needs	[54]
Digital technology use		
Prior experience with digital technology	Low use of digital tools (computer, internet, national health portal)	[45,57]

[a]All listed factors were negatively associated with digital health literacy.

[b]DBP: diastolic blood pressure.

[c]Hb: hemoglobin.

[d]NT-proBNP: N-terminal pro-B-type natriuretic peptide.

[e]NYHA: New York Heart Association classification of heart failure functional status.

XSL•FO
RenderX

Association Between DHL and Health Outcomes

Health outcomes associated with DHL among adults with CVD were categorized as cognitive outcomes (disease knowledge and empowerment), health behaviors (smoking, self-care behaviors, physical activity, and medication adherence), clinical outcomes (cardiac events), health-related outcomes (depression, anxiety, and quality of life), and digital health-related outcomes (use of digital health technology and acceptance of digital health; Table 6).

Across the included studies, higher DHL was most frequently associated with digital health–related outcomes, particularly greater use and higher acceptance of digital health technologies (n=8). Other reported associations included improved self-care behaviors (n=3) and enhanced quality of life (n=2).

Most studies reported their results using regression analysis, while others used correlation analysis, logistic regression, or Cox regression for survival analyses. Effect sizes were reported in various forms depending on study design, including standardized beta coefficient (β), unstandardized coefficient (*B*), odds ratio, correlation coefficient (*r*), and hazard ratio (Multimedia Appendix 8).

XSL•FO
RenderX

Table 6. Association between digital health literacy and health outcomes.

Category, subcategory, and observed health outcomes	References	Certainty[a]
Cognitive outcomes		
Disease knowledge		
Greater HF[b] knowledge	[53]	⊕OOO
Empowerment		
Increased empowerment	[46]	⊕OOO
Health behaviors		
Smoking		
Lower likelihood of smoking	[48]	⊕OOO
Self-care behaviors		
Higher self-care behaviors	[53,55,62]	⊕OOO
Physical activity		
Higher physical activity	[48]	⊕OOO
Medication adherence		
Improved medication adherence	[51]	⊕OOO
Clinical outcomes		
Cardiac events		
Lower risk of cardiac events	[51]	⊕OOO
Symptom status		
Lower symptom burden	[64]	⊕OOO
Health-related outcomes		
Depression		
Reduced depressive symptoms	[48]	⊕OOO
Anxiety		
Reduced anxiety symptoms	[48]	⊕OOO
Quality of life		
Improved quality of life	[51,55]	⊕OOO
Digital health-related outcomes		
Use of digital health technology		
Greater use of the internet, mobile apps, and electronic health applications	[47,57,61]	⊕⊕OO[c] ⊕OOO[d]
Higher technology self-efficacy and motivation; lower technophobia	[52]	⊕OOO
Acceptance of digital health		
Increased acceptance of mobile health; greater willingness to participate in telemonitoring	[56,60,63]	⊕OOO
Higher telehealth satisfaction and more positive attitudes toward health care technology; lower telehealth satisfaction with higher computer self-efficacy	[58]	⊕OOO

[a]The certainty of evidence for each outcome was evaluated using the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) approach: ⊕⊕⊕⊕ indicates high certainty, ⊕⊕⊕O moderate certainty, ⊕⊕OO low certainty, and ⊕OOO very low certainty.

[b]HF: heart failure.

[c]Certainty of evidence based on randomized controlled trial evidence.

[d]Certainty of evidence based on observational studies.

XSL•FO
RenderX

Meta-Analysis Results

Nine studies reported mean eHEALS scores, which were pooled using a random-effects model. The overall pooled mean eHEALS score was 24.26 (95% CI 21.19-27.32; $P<.001$; $I^2=98.4\%$; $\tau^2=15.55$; $\tau=3.94$). The 95% prediction interval ranged from 14.66 to 33.85, indicating substantial variability in true eHEALS scores across populations and settings. Subgroup analyses were conducted by region (n=9), disease type (n=9), and mean age (n=7). Studies conducted in Europe reported slightly higher eHEALS scores than those conducted in Asia. Studies with a mean age <65 years exhibited higher scores than those with a mean age ≥65 years. Studies that included only patients with heart failure demonstrated eHEALS scores nearly identical to those reported in studies of mixed CVD populations. No statistically significant subgroup differences were observed in analyses stratified by region ($P=.50$), disease type ($P=.98$), or age ($P=.38$; Figure 2). Leave-one-out sensitivity analysis indicated that the pooled eHEALS score ranged from 23.70 to 25.02, with a maximum variation of 1.32 points across iterations.

XSL•FO
RenderX

Figure 2. Forest plot of pooled mean eHEALS scores across included studies: (A) overall pooled mean eHEALS score, (B) subgroup meta-analysis by study region (Asia vs Europe), (C) subgroup meta-analysis by disease type (heart failure only vs mixed CVD conditions, including percutaneous coronary intervention or coronary heart disease), and (D) subgroup meta-analysis by mean age group (<65 years vs ≥65 years). This figure summarizes results from 9 studies published between 2018 and 2025, including 5 cross-sectional studies, 2 quasi-experimental studies, 1 longitudinal study, and 1 prospective cohort study. Study populations consisted of adult patients with heart failure or other cardiovascular diseases, recruited across Asia (Taiwan, Iran, South Korea, Indonesia, and China) and Europe (Denmark, Norway, and Turkey). Sample sizes ranged from 49 to 1970 participants per study, with a total of 3693 participants included in the overall meta-analysis [45-47,51,54,56,60,62,63]. eHEALS: eHealth Literacy Scale; HKSJ: Hartung-Knapp-Sidik-Jonkman method; MRAW: mean raw; PCI/CHD: percutaneous coronary intervention/coronary heart disease.

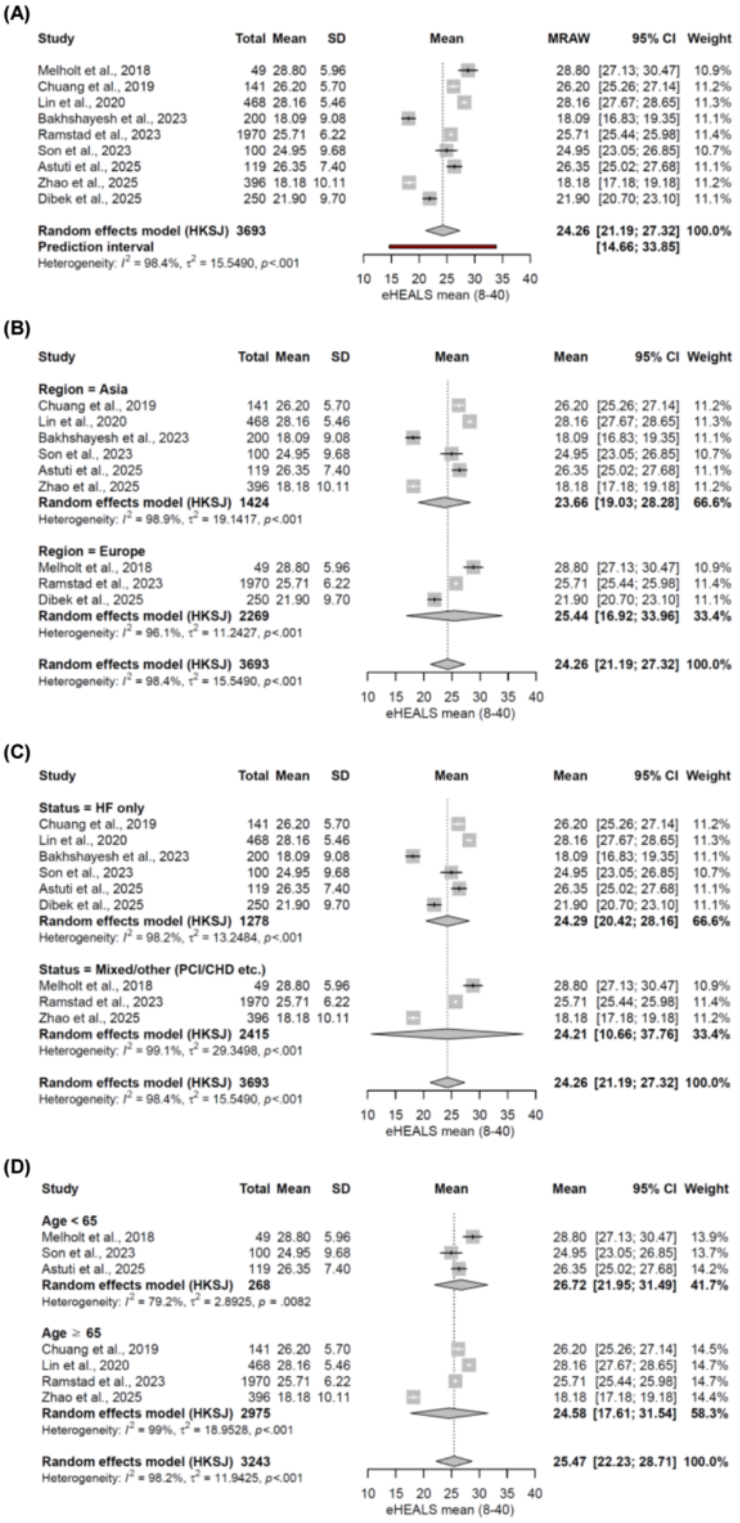

XSL•FO
RenderX

Assessment of Reporting Biases

Assessment of reporting biases was not conducted because fewer than 10 studies were included in the meta-analysis, as described in the Methods section.

Certainty of Evidence

The certainty of evidence for each outcome, as assessed using the GRADE framework, was rated as low or very low (Table 6). The evidence was initially low, reflecting its primary reliance on observational studies. Certainty was further downgraded for imprecision, attributable to the limited number of studies and wide CIs. Detailed ratings and the rationale for downgrades are provided in Multimedia Appendix 9.

Discussion

Principal Findings

This review synthesized evidence on DHL levels, factors associated with DHL, and related health outcomes among adults with CVD, a population for whom sustained self-management is critical. Overall, the findings indicate that DHL levels among adults with CVD are relatively low and characterized by substantial between-study heterogeneity. The wide prediction interval further reflects this heterogeneity, indicating that DHL levels may vary considerably across clinical contexts even when the pooled mean estimate is similar [36]. Sociodemographic characteristics, particularly older age, lower educational attainment, and female sex, were commonly associated with lower DHL. Higher DHL was associated with more favorable health outcomes, including improved self-care behaviors, better quality of life, and greater acceptance of digital health technologies. Collectively, these findings provide an empirical overview of the role of DHL in supporting self-management in patients with CVD.

DHL Levels Across Studies

DHL is increasingly recognized as a key determinant of both public and individual health [66]. Using the commonly applied cutoff of 26 on the eHEALS to indicate limited DHL [67], the pooled eHEALS score identified in this review was below this threshold, suggesting generally low DHL levels among adults with CVD. Compared with findings from a previous systematic review of individuals with chronic diseases, including diabetes mellitus and hypertension [37], DHL levels observed in the present review were comparatively lower. According to the American Heart Association, the prevalence of CVD increases substantially with advancing age, with the highest burden observed among older adults [68]. Accordingly, the relatively low DHL levels identified in this review should be interpreted in the context of the older age distribution of CVD populations. This finding is consistent with a previous meta-analysis by Kanejima et al [26], which reported a high prevalence of limited health literacy among patients with CVD. Effective management of CVD requires adherence to medication regimens, dietary modifications, and timely symptom management and response [12]. These tasks increasingly depend on accessing, interpreting, and applying health information delivered through digital platforms [69]. The cognitive and technological demands associated with CVD management may contribute to challenges in engaging with digital health information and may partially explain the lower observed DHL levels observed in this population [70].

Subgroup analyses were limited to variables consistently reported across studies, including region, disease type, and mean age. No statistically significant subgroup differences were identified, which may reflect limited statistical power due to the small number of included studies [71]. Nevertheless, directional trends warrant consideration. Studies conducted in Europe tended to report higher DHL levels than those conducted in Asia, potentially reflecting differences in digital health infrastructure, access to health technologies, and system-level support [72]. In addition, higher DHL levels observed among participants younger than 65 years suggest that age-related disparities in DHL persist within CVD. The CIs around the pooled estimate reflect uncertainty in the average DHL level across studies and should be interpreted with caution, given the substantial between-study heterogeneity. Moreover, the wide prediction interval indicates considerable variability across settings [36]. This variability suggests that DHL proficiency is not uniformly low among adults with CVD but is likely influenced by contextual and system-level factors [73].

Determinants and Health Outcomes of DHL

The factors identified in this review suggest that DHL among adults with CVD is shaped by both individual characteristics and broader social and digital contexts. Notably, the included studies predominantly focused on individual-level digital experiences, competencies, and perceptions, rather than structural or system-level digital environments. In addition, older age and lower educational attainment were consistently associated with lower DHL in adults with CVD. According to the 2026 American Heart Association Heart Disease and Stroke Statistics, CVD and heart failure disproportionately affect older adults [68]. Furthermore, large-scale cohort studies consistently demonstrate that lower educational attainment is associated with an increased risk of CVD [74]. Collectively, these sociodemographic characteristics provide critical context for interpreting DHL levels in patients with CVD. In this review, additional factors frequently associated with lower DHL included female sex, rural residence, poorer self-rated health, lower social support, and limited prior experience with digital technologies. These findings align with a recent meta-analysis of 17 studies in mixed populations, which identified age, ethnicity, income, employment status, education, and health status as key correlates of eHealth literacy [75].

However, rather than framing these demographic characteristics solely as individual deficits, they should be interpreted through the lens of Digital Determinants of Health [76]. According to the American Heart Association, Digital Determinants of Health—which encompass DHL, digital infrastructure, access, and digital inequity—play a pivotal role in shaping cardiovascular outcomes [77]. Most studies included in this review treated DHL as an individual-level attribute, emphasizing personal skills and behaviors. This approach reflects a broader tendency in the literature to frame DHL as an individual responsibility, with limited consideration of health system structures, provider practices, or service-level contexts that may

XSL•FO
RenderX

influence digital engagement [76]. From this perspective, disparities in DHL may not be fully explained by individual-level characteristics alone, highlighting the potential relevance of structural and provider-level factors [76,77].

Health outcomes associated with DHL clustered into 5 domains: cognitive, health behavior, clinical, health-related, and digital health–related outcomes. Across these domains, higher DHL was most consistently associated with improvements in cognitive understanding and health behaviors, including greater disease knowledge, increased empowerment, and more effective self-care practices such as medication adherence and physical activity. These findings suggest that DHL may facilitate self-management by enhancing information processing and supporting the translation of health information into everyday behaviors [14]. Digital health–related outcomes were most consistently associated with DHL in this review. Higher DHL was associated with greater acceptance and use of digital health technologies, stronger technology self-efficacy, and lower technophobia. These digital indicators appear closely aligned with the functional attributes of DHL, conceptualized as the ability to seek, appraise, and apply health information in digital contexts [14,27]. Therefore, DHL functions as a foundational competency that underpins a range of outcomes related to the use of digital health tools, supporting sustained self-management [22].

In this review, evidence linking DHL to clinical outcomes was comparatively limited. While some included studies reported associations between higher DHL and lower risks of cardiac events [51] or improved symptom status [64], these findings were based on a small number of studies and heterogeneous outcome measures. Further longitudinal research is required to elucidate the pathways linking DHL to hard clinical outcomes, including studies that integrate patient-reported perspectives on providers and system-level digital support.

The overall certainty of these findings is limited by study design characteristics and ROB across the included studies. Most synthesized studies were observational and therefore inherently susceptible to confounding. In addition, quality assessment revealed methodological limitations, particularly incomplete reporting of participant flow, nonresponders, and attrition. Given that DHL is influenced by sociodemographic factors such as age, educational attainment, and socioeconomic status [20], insufficient information on sample selection and follow-up may introduce selection bias and affect the interpretation of DHL estimates.

As digital health environments continue to evolve rapidly, transparent reporting of participant characteristics and study procedures is increasingly critical for contextualizing findings [78]. Improving reporting of sample representativeness and participant flow would enhance the interpretability of future studies and contribute to a more robust evidence base on DHL among adults with CVD. Considering the observed heterogeneity, ROB, and limitations identified in the GRADE framework, the overall certainty of evidence regarding associations between DHL and health outcomes was rated as low. Consequently, these findings should be interpreted as indicative of broad trends rather than a precise estimate applicable to all individual patients.

Implications for Clinical Practice and Health Policy

These findings have key implications for clinical practice. DHL and digital access should be considered central determinants of health equity that influence patients' capacity for effective self-management, rather than merely reflecting individual-level limitations [73]. The emphasis on individual-level factors in the existing literature underscores the need to also consider the health care system and provider-level influences on patients' digital engagement [76]. Accordingly, health care systems and providers may play a critical role in supporting equitable self-care by addressing digital barriers [77].

In clinical settings, assessing patients' digital access and competencies as part of routine assessment—alongside other social and contextual factors—may help identify individuals at risk of digital exclusion [73]. These assessments should not be treated solely as screening tools, but as a foundation for adapting service delivery and provider practices to accommodate varying levels of digital readiness [76]. For patients with low DHL or limited access, the mere provision of digital technologies may be insufficient. System-level strategies—such as supported-use models integrated into routine care and service designs that enhance accessibility—may be required, alongside patient education programs or involvement of family caregivers where appropriate [69,73,77]. Adopting a digital inclusion–oriented approach may help ensure that digital health technologies act as facilitators rather than barriers to care, thereby supporting more effective and sustained self-management among adults with CVD [69,77].

Strengths and Limitations

This review has several strengths. First, to our knowledge, it is the first systematic review to comprehensively synthesize health outcomes associated with DHL in patients with CVD using a multidimensional framework encompassing 5 distinct domains. Second, by incorporating findings from multidimensional instruments such as the DHLI [19], eHLQ [20], and DHRQ [65], this study aligns with the evolving conceptual definition of DHL [14]. This approach extends beyond traditional eHealth literacy and captures both functional skills and attitudinal components (eg, motivation and digital confidence) required in contemporary digital health environments.

Several limitations warrant consideration. First, most included studies were cross-sectional and showed substantial variability in analytic methods and confounder adjustment, limiting comparability. Second, evidence was limited for psychosocial determinants as well as objective clinical endpoints and biological markers. Many biomarkers and clinical events were evaluated in only a small number of studies, often from single-center samples with modest sizes, which constrains generalizability. Third, reporting of nonresponders and participant flow was often incomplete. This limited transparency may compromise assessments of sample representativeness and increase the risk of selection bias, particularly given known sociodemographic gradients in DHL [21]. Finally, substantial between-study heterogeneity was observed across study

XSL•FO
RenderX

populations, settings, measurement instruments, and analytic approaches. Accordingly, pooled estimates and prediction intervals should be interpreted with caution.

Conclusions

To our knowledge, this review is one of the first to integrate evidence on DHL levels, associated factors, and DHL-related health outcomes among adults with CVD. Overall, DHL varied widely and showed consistent gradients by age and socioeconomic disadvantage, with lower DHL generally associated with poorer engagement with digital health resources and less favorable self-management–related outcomes. Although the pooled mean eHEALS score indicates modest DHL on average, the wide prediction interval indicates considerable variability across populations and clinical settings. By synthesizing DHL levels, determinants, and health outcomes in CVD populations, this review offers an integrated evidence base that extends beyond single-instrument assessments (eg, eHEALS) and supports interpretation of DHL in relation to structural and health system contexts.

Our findings highlight the need for routine assessment of DHL and targeted support for patients with limited DHL, as well as the importance of designing and delivering digital tools that are accessible and easy to use in routine CVD care. Future studies using longitudinal and interventional designs are needed to clarify causal pathways linking DHL to self-management and health outcomes, and to examine how provider- and health system–level factors shape and support DHL in routine care.

Acknowledgments

We acknowledge the contributions of individuals who supported this work but did not meet the criteria for authorship. Generative AI tools were not used in the preparation of this manuscript.

Funding

This work was supported by a grant from the National Research Foundation of Korea (NRF), funded by the Korean government (MSIT) (grant RS-2025-00562535). The funder had no role in the study design, data collection, analysis, interpretation, or manuscript preparation.

Data Availability

The data underlying the findings of this study are available from the corresponding authors upon reasonable request.

Authors' Contributions

Conceptualization: EJK, YJS
Data curation: EJK, DYK
Formal analysis: EJK, DYK, HJK
Methodology: YJS
Project administration: YJS
Software: EJK, DYK, HJK
Supervision: YJS
Validation: EJK, DYK, HJK, YJS
Visualization: EJK, DYK, HJK
Writing – original draft: EJK, DYK, HJK, YJS
Writing – review & editing: EJK, DYK, YJS

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA 2020 reporting checklists (PRISMA 2020, PRISMA Abstract, and PRISMA-S).
[DOCX File , 44 KB - jmir_v28i1e89102_app1.docx]

Multimedia Appendix 2
Literature search strategy.
[DOCX File , 30 KB - jmir_v28i1e89102_app2.docx]

Multimedia Appendix 3
Summary of digital health literacy measurement tools included in this review.
[DOCX File , 33 KB - jmir_v28i1e89102_app3.docx]

XSL•FO
RenderX

Multimedia Appendix 4
Quality assessment for cross-sectional studies using the AXIS (Appraisal Tool for Cross-Sectional Studies) tool.
[DOCX File , 30 KB - jmir_v28i1e89102_app4.docx]

Multimedia Appendix 5
Newcastle–Ottawa Quality Assessment of longitudinal observational and prospective observational studies.
[DOCX File , 25 KB - jmir_v28i1e89102_app5.docx]

Multimedia Appendix 6
Risk-of-bias tool for randomized controlled trials.
[DOCX File , 26 KB - jmir_v28i1e89102_app6.docx]

Multimedia Appendix 7
Risk of bias in nonrandomized studies of interventions.
[DOCX File , 28 KB - jmir_v28i1e89102_app7.docx]

Multimedia Appendix 8
Study-level digital health literacy levels, associated factors, and health outcomes.
[DOCX File , 34 KB - jmir_v28i1e89102_app8.docx]

Multimedia Appendix 9
GRADE (Grading of Recommendations Assessment, Development, and Evaluation) assessment.
[DOCX File , 27 KB - jmir_v28i1e89102_app9.docx]

References

1. What is cardiovascular disease? American Heart Association. 2024 Jan 10. URL: https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease [accessed 2026-03-09]
2. Cruz-Ramos NA, Alor-Hernández G, Colombo-Mendoza LO, Sánchez-Cervantes JL, Rodríguez-Mazahua L, Guarneros-Nolasco LR. mHealth apps for self-management of cardiovascular diseases: a scoping review. Healthcare (Basel) 2022;10(2):322 [FREE Full text] [doi: 10.3390/healthcare10020322] [Medline: 35206936]
3. Huang Y, Li S, Lu X, Chen W, Zhang Y. The effect of self-management on patients with chronic diseases: a systematic review and meta-analysis. Healthcare (Basel) 2024;12(21):2151 [FREE Full text] [doi: 10.3390/healthcare12212151] [Medline: 39517362]
4. Yuen E, Winter N, Savira F, Huggins CE, Nguyen L, Cooper P, et al. Digital health literacy and its association with sociodemographic characteristics, health resource use, and health outcomes: rapid review. Interact J Med Res 2024;13:e46888 [FREE Full text] [doi: 10.2196/46888] [Medline: 39059006]
5. Graham-Brown MPM, Smith AC, Greenwood SA. Digital health interventions in chronic kidney disease: levelling the playing field? Clin Kidney J 2023;16(5):763-767 [FREE Full text] [doi: 10.1093/ckj/sfac259] [Medline: 37151422]
6. Endalamaw A, Zewdie A, Wolka E, Assefa Y. A scoping review of digital health technologies in multimorbidity management: mechanisms, outcomes, challenges, and strategies. BMC Health Serv Res 2025;25(1):382 [FREE Full text] [doi: 10.1186/s12913-025-12548-5] [Medline: 40089752]
7. Hu S, Song D, Wan S, Zhang S, Luo C, Li N, et al. Digital health: current applications, challenges, and future directions for enhancing healthcare quality and safety. Front Public Health 2025;13:1646802. [doi: 10.3389/fpubh.2025.1646802] [Medline: 41089861]
8. Livieri G, Mangina E, Protopapadakis ED, Panayiotou AG. The gaps and challenges in digital health technology use as perceived by patients: a scoping review and narrative meta-synthesis. Front Digit Health 2025;7:1474956. [doi: 10.3389/fdgth.2025.1474956] [Medline: 40212901]
9. Diamond JE, Kaltenbach LA, Granger BB, Fonarow GC, Al-Khalidi HR, Albert NM, et al. Access to mobile health interventions among patients hospitalized with heart failure: insights into the digital divide from the CONNECT-HF mHealth substudy. Circ Heart Fail 2024;17(2):e011140. [doi: 10.1161/CIRCHEARTFAILURE.123.011140] [Medline: 38205653]
10. Kim DS, Eltahir AA, Ngo S, Rodriguez F. Bridging the gap: how accounting for social determinants of health can improve digital health equity in cardiovascular medicine. Curr Atheroscler Rep 2024;27(1):9. [doi: 10.1007/s11883-024-01249-9] [Medline: 39576395]
11. Wamala Andersson S, Gonzalez MP. Digital health literacy – a key factor in realizing the value of digital transformation in healthcare. Front Digit Health 2025;7:1461342. [doi: 10.3389/fdgth.2025.1461342] [Medline: 40538571]

XSL•FO
RenderX

12. Magnani JW, Mujahid MS, Aronow HD, Cené CW, Dickson VV, Havranek E, et al. Health literacy and cardiovascular disease: fundamental relevance to primary and secondary prevention: a scientific statement from the American Heart Association. Circulation 2018;138(2):e48-e74 [FREE Full text] [doi: 10.1161/CIR.0000000000000579] [Medline: 29866648]
13. Ndayishimiye C, Lopes H, Middleton J. A systematic scoping review of digital health technologies during COVID-19: a new normal in primary health care delivery. Health Technol (Berl) 2023;13(2):273-284 [FREE Full text] [doi: 10.1007/s12553-023-00725-7] [Medline: 36628261]
14. Ban S, Kim Y, Seomun G. Digital health literacy: a concept analysis. Digit Health 2024;10:20552076241287894 [FREE Full text] [doi: 10.1177/20552076241287894] [Medline: 39381807]
15. Griebel L, Enwald H, Gilstad H, Pohl A, Moreland J, Sedlmayr M. eHealth literacy research—quo vadis? Inform Health Soc Care 2018;43(4):427-442. [doi: 10.1080/17538157.2017.1364247] [Medline: 29045164]
16. Shao YJ, Duan XC, Xu XJ, Guo HY, Zhang ZY, Zhao S, et al. Latent profile and determinants of self-management behaviors among older adult patients with chronic diseases: a cross-sectional study. Front Public Health 2025;13:1506545. [doi: 10.3389/fpubh.2025.1506545] [Medline: 39975786]
17. Ullrich G, Bäuerle A, Vogt H, Mahabadi AA, Paldán K, Messiha D, et al. Digital health literacy and attitudes toward eHealth technologies among patients with cardiovascular disease and their implications for secondary prevention: survey study. JMIR Form Res 2025;9:e63057 [FREE Full text] [doi: 10.2196/63057] [Medline: 40106277]
18. Norman CD, Skinner HA. eHEALS: the eHealth literacy scale. J Med Internet Res 2006;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
19. van der Vaart R, Drossaert C. Development of the digital health literacy instrument: measuring a broad spectrum of Health 1.0 and Health 2.0 skills. J Med Internet Res 2017;19(1):e27 [FREE Full text] [doi: 10.2196/jmir.6709] [Medline: 28119275]
20. Kayser L, Karnoe A, Furstrand D, Batterham R, Christensen KB, Elsworth G, et al. A multidimensional tool based on the ehealth literacy framework: development and initial validity testing of the eHealth literacy questionnaire (eHLQ). J Med Internet Res 2018;20(2):e36 [FREE Full text] [doi: 10.2196/jmir.8371] [Medline: 29434011]
21. Estrela M, Semedo G, Roque F, Ferreira PL, Herdeiro MT. Sociodemographic determinants of digital health literacy: a systematic review and meta-analysis. Int J Med Inform 2023;177:105124 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.105124] [Medline: 37329766]
22. Qi Y, Mohamad E, Azlan AA, Zhang C, Ma Y, Wu A. Digital health solutions for cardiovascular disease prevention: systematic review. J Med Internet Res 2025;27:e64981 [FREE Full text] [doi: 10.2196/64981] [Medline: 39847411]
23. Verweel L, Newman A, Michaelchuk W, Packham T, Goldstein R, Brooks D. The effect of digital interventions on related health literacy and skills for individuals living with chronic diseases: a systematic review and meta-analysis. Int J Med Inform 2023;177:105114. [doi: 10.1016/j.ijmedinf.2023.105114] [Medline: 37329765]
24. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
25. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
26. Kanejima Y, Shimogai T, Kitamura M, Ishihara K, Izawa KP. Impact of health literacy in patients with cardiovascular diseases: a systematic review and meta-analysis. Patient Educ Couns 2022;105(7):1793-1800. [doi: 10.1016/j.pec.2021.11.021] [Medline: 34862114]
27. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
28. Downes MJ, Brennan ML, Williams HC, Dean RS. Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). BMJ Open 2016;6(12):e011458 [FREE Full text] [doi: 10.1136/bmjopen-2016-011458] [Medline: 27932337]
29. Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. The Ottawa Hospital Research Institute. URL: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp [accessed 2026-03-09]
30. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
31. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016;355:i4919 [FREE Full text] [doi: 10.1136/bmj.i4919] [Medline: 27733354]
32. Barili F, Parolari A, Kappetein PA, Freemantle N. Statistical primer: heterogeneity, random- or fixed-effects model analyses? Interact Cardiovasc Thorac Surg 2018;27(3):317-321 [FREE Full text] [doi: 10.1093/icvts/ivy163] [Medline: 29868857]
33. Shim SR, Kim S. Intervention meta-analysis: application and practice using R software. Epidemiol Health 2019;41:e2019008 [FREE Full text] [doi: 10.4178/epih.e2019008] [Medline: 30999738]
34. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014;14:25 [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]

XSL•FO
RenderX

35. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327(7414):557-560 [FREE Full text] [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
36. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023;12(4):101014 [FREE Full text] [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
37. Zaghloul H, Fanous K, Ahmed L, Arabi M, Varghese S, Omar S, et al. Digital health literacy in patients with common chronic diseases: systematic review and meta-analysis. J Med Internet Res 2025;27:e56231 [FREE Full text] [doi: 10.2196/56231] [Medline: 40854231]
38. Liu X, Valdez D, Parker MA, Mai A, Walsh-Buhi ER. Quality of cancer-related information on new media (2014-2023): systematic review and meta-analysis. J Med Internet Res 2025;27:e73185 [FREE Full text] [doi: 10.2196/73185] [Medline: 41061257]
39. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2024. URL: https://www.R-project.org/ [accessed 2026-03-09]
40. Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. Evid Based Ment Health 2019;22(4):153-160 [FREE Full text] [doi: 10.1136/ebmental-2019-300117] [Medline: 31563865]
41. Viechtbauer W. Conducting meta-analyses in R with the meta for package. J Stat Softw 2010;36(3):1-48. [doi: 10.18637/jss.v036.i03]
42. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ. Cochrane Handbook for Systematic Reviews of Interventions, Version 6.5. The Cochrane Collaboration. 2024 Aug. URL: https://www.cochrane.org/handbook [accessed 2026-03-09]
43. Schünemann H, Brożek J, Guyatt G, Oxman A. GRADE Handbook for Grading the Quality of Evidence and the Strength of Recommendations Using the GRADE Approach. Cochrane. 2013 Oct. URL: https://gdt.gradepro.org/app/handbook/handbook.html [accessed 2026-03-09]
44. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336(7650):924-926 [FREE Full text] [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
45. Melholt C, Joensson K, Spindler H, Hansen J, Andreasen JJ, Nielsen G, et al. Cardiac patients' experiences with a telerehabilitation web portal: implications for eHealth literacy. Patient Educ Couns 2018;101(5):854-861. [doi: 10.1016/j.pec.2017.12.017] [Medline: 29305064]
46. Rodríguez Parrado IY, Achury Saldaña DM. Digital health literacy in patients with heart failure in times of pandemic. Comput Inform Nurs 2022;40(11):754-762 [FREE Full text] [doi: 10.1097/CIN.0000000000000883] [Medline: 35234702]
47. Spindler H, Dyrvig AK, Schacksen CS, Anthonimuthu D, Frost L, Gade JD, et al. Increased motivation for and use of digital services in heart failure patients participating in a telerehabilitation program: a randomized controlled trial. Mhealth 2022;8:25 [FREE Full text] [doi: 10.21037/mhealth-21-56] [Medline: 35928510]
48. Brørs G, Dalen H, Allore H, Deaton C, Fridlund B, Norman CD, et al. The association of electronic health literacy with behavioural and psychological coronary artery disease risk factors in patients after percutaneous coronary intervention: a 12-month follow-up study. Eur Heart J Digit Health 2023;4(2):125-135 [FREE Full text] [doi: 10.1093/ehjdh/ztad010] [Medline: 36974270]
49. Son Y, Kim S, Lee W, Won H, Cho JH, Hong J, et al. The effects of a 24-week interactive text message-based mobile health intervention for enhancing self-care behaviours of patients with heart failure: a quasi-experimental study. Nurs Open 2023;10(9):6309-6319 [FREE Full text] [doi: 10.1002/nop2.1878] [Medline: 37313589]
50. Sun S, Guo R, Wang Y, Jiao L, Zhao C, Qiao S, et al. Correlation analysis of electronic health literacy, compliance behavior, and quality of life in middle-aged and older patients with coronary heart disease: a cross-sectional study. Front Public Health 2025;13:1680950. [doi: 10.3389/fpubh.2025.1680950] [Medline: 41446526]
51. Lin C, Ganji M, Griffiths MD, Bravell ME, Broström A, Pakpour AH. Mediated effects of insomnia, psychological distress and medication adherence in the association of eHealth literacy and cardiac events among Iranian older patients with heart failure: a longitudinal study. Eur J Cardiovasc Nurs 2020;19(2):155-164. [doi: 10.1177/1474515119873648] [Medline: 31516036]
52. Zhao J, Hu D, Du H, Wang H, Tu X, Wang A. Social support and technophobia in older patients with coronary heart disease: the mediating roles of eHealth literacy and healthcare technology self-efficacy. PLOS One 2025;20(6):e0325202 [FREE Full text] [doi: 10.1371/journal.pone.0325202] [Medline: 40465587]
53. Chuang HW, Kao CW, Lin WS, Chang YC. Factors affecting self-care maintenance and management in patients with heart failure: testing a path model. J Cardiovasc Nurs 2019;34(4):297-305. [doi: 10.1097/JCN.0000000000000575] [Medline: 31033857]
54. Yun S, Enjuanes C, Calero-Molina E, Hidalgo E, José-Bazán N, Ruiz M, et al. Usefulness of telemedicine-based heart failure monitoring according to 'eHealth literacy' domains: insights from the iCOR randomized controlled trial. Eur J Intern Med 2022;101:56-67. [doi: 10.1016/j.ejim.2022.04.008] [Medline: 35483994]

XSL•FO
RenderX

55. Bakhshayesh M, Hojatansari M, Rabie Siahkali S. Investigating the relationship of electronic health literacy and self-care behaviors with life quality in patients with heart failure. Prevent Care Nurs Midwifery J 2023;13(4):30-40. [doi: 10.61186/pcnm.13.4.30]
56. Bäuerle A, Mallien C, Rassaf T, Jahre L, Rammos C, Skoda E, et al. Determining the acceptance of digital cardiac rehabilitation and its influencing factors among patients affected by cardiac diseases. J Cardiovasc Dev Dis 2023;10(4):174 [FREE Full text] [doi: 10.3390/jcdd10040174] [Medline: 37103053]
57. Ramstad KJ, Brørs G, Pettersen TR, Deaton C, Palm P, Rotevatn S, et al. eHealth technology use and eHealth literacy after percutaneous coronary intervention. Eur J Cardiovasc Nurs 2023;22(5):472-481. [doi: 10.1093/eurjcn/zvac087] [Medline: 36190843]
58. Rush KL, Burton L, Seaton CL, Loewen P, O'Connor BP, Moroz L, et al. Telehealth satisfaction in patients receiving virtual atrial fibrillation care: quantitative exploratory study. JMIR Hum Factors 2023 Sep 14;10:e50232. [doi: 10.2196/50232] [Medline: 37707881]
59. Vitolo M, Ziveri V, Gozzi G, Busi C, Imberti JF, Bonini N, et al. DIGItal health literacy after COVID-19 outbreak among frail and non-frail cardiology patients: the DIGI-COVID study. J Pers Med 2022;13(1):99 [FREE Full text] [doi: 10.3390/jpm13010099] [Medline: 36675760]
60. Mohajeri D, Jahre LM, Bäuerle A, Schieffers T, Messiha D, Rammos C, et al. Acceptance, needs, and demands for nutritional mhealth support in patients with cardiovascular disease. Nutrients 2024;16(23):4155 [FREE Full text] [doi: 10.3390/nu16234155] [Medline: 39683549]
61. van Schalkwijk D, Lodder P, Everaert J, Widdershoven J, Habibović M. Latent profiles of telehealth care satisfaction during the COVID-19 pandemic among patients with cardiac conditions in an outpatient setting. Cardiovasc Digit Health J 2024;5(2):85-95 [FREE Full text] [doi: 10.1016/j.cvdhj.2023.11.022] [Medline: 38765625]
62. Astuti RP, Nurdina G, Herdiman H. Factors influencing self-care among patients with heart failure in Indonesia. Med Surg J Nurs Res 2025;4(1):11-22. [doi: 10.70331/jpkmb.v4i1.47]
63. Cuppen S, van Leunen M, Henken T, Goevaerts M, Scherrenberg M, Falter M, et al. Association of technology-related skills and self-efficacy with willingness to participate in heart failure telemonitoring: cross-sectional observational study. JMIR Form Res 2025;9:e68992 [FREE Full text] [doi: 10.2196/68992] [Medline: 40505140]
64. Dibek EN, Eren MG, Sert H, Kocayigit I. Relationship of health behaviors and symptom status mediated by digital health literacy in persons with heart failure. Nurs Res 2025;74(5):349-356. [doi: 10.1097/NNR.0000000000000828] [Medline: 40307233]
65. Scherrenberg M, Falter M, Kaihara T, Xu L, van Leunen M, Kemps H, et al. Development and internal validation of the digital health readiness questionnaire: prospective single-center survey study. J Med Internet Res 2023;25:e41615 [FREE Full text] [doi: 10.2196/41615] [Medline: 36897627]
66. Barbati C, Maranesi E, Giammarchi C, Lenge M, Bonciani M, Barbi E, et al. Effectiveness of eHealth literacy interventions: a systematic review and meta-analysis of experimental studies. BMC Public Health 2025;25(1):288 [FREE Full text] [doi: 10.1186/s12889-025-21354-x] [Medline: 39849354]
67. Xie L, Zhang S, Xin M, Zhu M, Lu W, Mo PK. Electronic health literacy and health-related outcomes among older adults: a systematic review. Prev Med 2022;157:106997. [doi: 10.1016/j.ypmed.2022.106997] [Medline: 35189203]
68. Palaniappan LP, Allen NB, Almarzooq ZI, Anderson CAM, Arora P, Avery CL, et al. 2026 heart disease and stroke statistics: a report of US and global data from the American Heart Association. Circulation 2026;153(9):e275-e906 [FREE Full text] [doi: 10.1161/CIR.0000000000001412] [Medline: 41562125]
69. Avoke D, Elshafeey A, Weinstein R, Kim CH, Martin SS. Digital health in diabetes and cardiovascular disease. Endocr Res 2024;49(3):124-136. [doi: 10.1080/07435800.2024.2341146] [Medline: 38605594]
70. Shao Y, Hou X, Peng Q, Hu M, Li C. A meta-synthesis of qualitative studies on cardiovascular disease patients' experiences using digital health tools. Front Public Health 2025;13:1709562. [doi: 10.3389/fpubh.2025.1709562] [Medline: 41346736]
71. Zeymer U, Freund A, Hochadel M, Ostadal P, Belohlavek J, Rokyta R, et al. Venoarterial extracorporeal membrane oxygenation in patients with infarct-related cardiogenic shock: an individual patient data meta-analysis of randomised trials. Lancet 2023;402(10410):1338-1346. [doi: 10.1016/S0140-6736(23)01607-0] [Medline: 37643628]
72. Okpechi IG, Muneer S, Ye F, Zaidi D, Ghimire A, Tinwala MM, et al. Global eHealth capacity: secondary analysis of WHO data on eHealth and implications for kidney care delivery in low-resource settings. BMJ Open 2022;12(3):e055658 [FREE Full text] [doi: 10.1136/bmjopen-2021-055658] [Medline: 35321893]
73. Sieck CJ, Sheon A, Ancker JS, Castek J, Callahan B, Siefer A. Digital inclusion as a social determinant of health. NPJ Digit Med 2021;4(1):52. [doi: 10.1038/s41746-021-00413-8] [Medline: 33731887]
74. Tao J, Zhao X, Li B, Sun H, Hu Y, Chen S, et al. Associations of educational attainment and traditional risk factor control with cardiovascular disease. Am J Prev Cardiol 2025;23:101031 [FREE Full text] [doi: 10.1016/j.ajpc.2025.101031] [Medline: 40585338]
75. Hua Z, Yuqing S, Qianwen L, Hong C. Factors influencing eHealth literacy worldwide: systematic review and meta-analysis. J Med Internet Res 2025;27:e50313 [FREE Full text] [doi: 10.2196/50313] [Medline: 40063939]
76. Chidambaram S, Jain B, Jain U, Mwavu R, Baru R, Thomas B, et al. An introduction to digital determinants of health. PLOS Digit Health 2024 Jan;3(1):e0000346 [FREE Full text] [doi: 10.1371/journal.pdig.0000346] [Medline: 38175828]

XSL•FO
RenderX

77. Powell-Wiley TM, Brewer LC, Burke LE, Hernandez R, Landsbaugh Kaar J, Kepper M, et al. Role of technology in promoting heart healthy behavior change to increase equity in optimal cardiovascular health: a scientific statement from the American Heart Association. Circulation 2025;151(18):e972-e985 [FREE Full text] [doi: 10.1161/CIR.0000000000001314] [Medline: 40190270]
78. Perrin Franck C, Babington-Ashaye A, Dietrich D, Bediang G, Veltsos P, Gupta PP, et al. iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res 2023;25:e46694 Erratum in: J Med Internet Res. 2023;25:e49027. doi: 10.2196/49027 [FREE Full text] [doi: 10.2196/46694] [Medline: 37163336]

Abbreviations

AHRQ: Agency for Healthcare Research and Quality
AXIS: Appraisal Tool for Cross-Sectional Studies
CVD: cardiovascular disease
DHL: digital health literacy
DHLI: Digital Health Literacy Instrument
DHRQ: Digital Health Readiness Questionnaire
eHEALS: eHealth Literacy Scale
eHLQ: eHealth Literacy Questionnaire
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
MeSH: Medical Subject Headings
NOS: Newcastle-Ottawa Scale
PICOS: participants, intervention or interest, comparisons, outcomes, and study design
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
ROB: risk of bias
RoB 2: Revised Cochrane Risk-of-Bias Tool for Randomized Trials
ROBINS-I: Risk of Bias in Nonrandomized Studies of Interventions

Edited by S Brini; submitted 08.Dec.2025; peer-reviewed by D Nutbeam, M Harris; comments to author 27.Jan.2026; revised version received 22.Feb.2026; accepted 24.Feb.2026; published 24.Mar.2026.

Please cite as:
Kim EJ, Kim DY, Kim HJ, Son YJ
Determinants and Health Outcomes of Digital Health Literacy in Patients With Cardiovascular Disease: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e89102
URL: https://www.jmir.org/2026/1/e89102
doi:10.2196/89102
PMID:41874540

©Eun-Jung Kim, Da-Young Kim, Ha-Jin Kim, Youn-Jung Son. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Robot-Assisted Therapy for Upper Limb Rehabilitation After Stroke: Umbrella Review

Sijia Liu[1], MD; Xianghong Zhang[2], MSc; Lijun Zhou[2], MBBS; Jiashan Zhang[2], MBBS; Jialin Liu[1], MD; Chengqi He[1], MD

[1]West China Hospital of Sichuan University, Chengdu, China

[2]West China School of Clinical Medicine, Sichuan University, Chengdu, China

Corresponding Author:
Chengqi He, MD
West China Hospital of Sichuan University
No.37 Guoxuexiang, Wuhou District
Chengdu, 610041
China
Phone: 86 13880982177
Fax: 1 388 098 2177
Email: hxkfhcq2015@126.com

Abstract

Background: Stroke is a leading cause of long-term upper limb disability, severely impacting patients' independence and quality of life. Robot-assisted therapy (RAT) has emerged as a promising, high-intensity rehabilitation alternative. However, conclusions from existing systematic reviews on its efficacy are inconsistent and often lack a holistic framework, limiting their use for guiding personalized clinical decisions.

Objective: This study aims to systematically synthesize recent evidence on RAT for upper limb rehabilitation after stroke. Guided by the International Classification of Functioning, Disability and Health framework, it moves beyond singular outcomes to provide a multidimensional evaluation across body function, activity, and participation levels. The review aims to provide stratified guidance for clinical decision-making based on patient- and intervention-specific characteristics, thereby supporting evidence-based practice and informing future research.

Methods: This study included systematic reviews and meta-analyses published from January 1, 2019, to December 26, 2025, comparing RAT with conventional therapy for upper limb rehabilitation after stroke. Overall, 6 databases, including PubMed, Web of Science, and Embase, were searched. Two reviewers (XZ and LZ) independently performed study selection, data extraction, and quality assessment using the AMSTAR 2 tool. The synthesis integrated outcome measures and subgroup analyses derived from the included studies.

Results: This umbrella review included 21 meta-analyses encompassing 535 randomized controlled trials and 27,598 patients across acute, subacute, and chronic stroke stages. According to AMSTAR 2, 17 reviews were high quality, 3 moderate, and 1 critically low. The synthesis demonstrated that RAT was superior in improving upper limb motor function, but no statistically significant advantages were observed in activities of daily living compared to conventional therapy. Subgroup analyses revealed that treatment effects were influenced by stroke stage, upper limb motor impairment level, and robot type.

Conclusions: RAT is an effective intervention for improving upper limb motor function after stroke. However, its benefits are primarily observed at the level of body function, with limited evidence for long-term maintenance. The current evidence is constrained by significant outcome heterogeneity and methodological limitations inherent to umbrella reviews. Future research should validate these findings in broader clinical practice, focus on translating functional gains into sustained improvements in daily activities and participation, and include cost-effectiveness evaluations.

Trial Registration: PROSPERO CRD42024497183; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024497183

(J Med Internet Res 2026;28:e79363) doi:10.2196/79363

KEYWORDS

robot-assisted rehabilitation; stroke; upper limb rehabilitation; umbrella review; robot therapy effectiveness

XSL•FO
RenderX

Introduction

As one of the diseases with the highest disability rate worldwide, stroke severely affects upper limb motor function, significantly impacting patients' ability to perform activities of daily living (ADL) and their quality of life [1]. Upper limb motor impairment is highly prevalent post stroke, affecting nearly 80% of patients in the acute phase, with persistent deficits observed in 50%-60% of survivors at 6 months [2,3]. Consequently, early and effective motor intervention is crucial to improve upper limb function and enhance performance in daily activities. Although traditional physical therapy can effectively improve upper limb function in survivors with stroke, it is resource-intensive, costly, and often relies on potentially limited specialized facilities [4,5].

In recent years, robot-assisted therapy (RAT) has emerged as a major focus of research in upper limb rehabilitation, owing to its advantages, including quantifiability, high reproducibility, and task-oriented exercise [6,7]. Evidence from clinical studies suggests that RAT can enhance motor outcomes and promote neuroplasticity [8]. For instance, robotic-exoskeleton training has been shown to improve motor performance and cortical excitability [9]. Furthermore, task-oriented upper limb training with robotic assistance can modify sensorimotor cortex neuroplasticity and support motor control and learning [8]. Numerous systematic reviews have concluded that RAT can improve upper limb motor function and functional activity in survivors with stroke, irrespective of the specific device used [10-12].

However, the efficacy of RAT remains controversial [13]. For example, one study concluded that robotic rehabilitation did not yield significant advantages over conventional therapy in measures such as Fugl-Meyer Assessment-Upper Extremity (FMA-UE), Action Research Arm Test, and Motor Activity Log scores [14], while other studies have reported positive effects [15,16]. This heterogeneity may stem from factors such as variability in study populations (eg, stroke type) [17], differences in intervention protocols (eg, robot type) [18], and inconsistent assessment tools [19-21]. More critically, existing systematic reviews predominantly focus on single outcome measures (eg, motor function) and lack a comprehensive analysis of multidimensional recovery within the International Classification of Functioning, Disability and Health (ICF) framework [22], which distinguishes between body function, activity, and participation levels. This limitation prevents the provision of stratified guidance for clinical decision-making, which is tailored to patients' specific recovery goals and functional levels.

An umbrella review represents an advanced form of systematic review, distinguished by its capacity to synthesize findings from multiple existing systematic reviews or meta-analyses, thereby offering a more holistic and integrated perspective on a given research topic [23]. This methodology is particularly suited to mapping contradictions, identifying consistent patterns across reviews, and assessing the overall strength of evidence in a fragmented field. Building upon this methodology, the present study aims to systematically integrate the available systematic review evidence concerning RAT for upper limb rehabilitation following stroke. Through comprehensive analysis, we seek to elucidate not only the overall therapeutic effects but also identify key moderating factors that influence outcomes. Therefore, under the guidance of the ICF framework, this umbrella review aims to (1) synthesize and appraise evidence from existing systematic reviews on robot-assisted upper limb rehabilitation after stroke; (2) examine the consistency and strength of evidence across ICF domains; and (3) provide multidimensional, stratified guidance for clinical practice and future research directions.

Methods

Overview

The design, conduct, and reporting of this umbrella review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Multimedia Appendix 1) and its PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) checklist (Multimedia Appendix 2) [24,25]. Additionally, we adhered to the PRIOR (Preferred Reporting Items for Overviews of Reviews) checklist (Multimedia Appendix 3) to ensure comprehensive reporting [26].

Eligibility Criteria

Given the complexity of upper limb motor function and its critical role in daily life, upper limb rehabilitation robots have reached a relatively mature stage of research, development, and application. These systems enable high-precision, high-intensity, and personalized training, which is designed to promote functional recovery, quality of life, and self-care ability. Accordingly, this review specifically investigated the effects of robotic interventions on upper limb recovery post stroke. We excluded reviews that focused only on general stroke outcomes or those evaluating robotic interventions for lower limb rehabilitation. Furthermore, included studies were required to demonstrate clinical relevance through rigorous methodological quality and provide complete datasets to support their conclusions. To account for the rapid technological evolution in this field, we limited inclusion to meta-analyses published within the past 7 years.

Eligible studies involved adult patients with stroke undergoing upper limb rehabilitation and compared interventions incorporating RAT (as a stand-alone or adjunct treatment) against conventional therapy. The outcomes focused on clinical efficacy, encompassing upper limb motor function [27], muscle strength [28], spasticity [29], and ADL [30]. Meta-analyses lacking sufficient data and non-English publications were excluded.

Search Strategy

We performed a systematic literature search on December 26, 2025, using the Sichuan University Library Discovery System. The following databases were searched: PubMed, Web of Science, Embase, the Cochrane Library, IEEE Xplore, and Scopus. The search was restricted to peer-reviewed journal articles published in English between January 1, 2019, and the search date. The complete search syntax is detailed in Table S1

XSL•FO
RenderX

in Multimedia Appendix 4. The search strategy was developed de novo through team discussion to ensure conceptual adequacy. It was not based on any previously published strategy and was not updated after its initial execution. No supplementary search techniques were used, such as citation tracking, manual searches of grey literature, or contacting authors for unpublished data. The study protocol was prospectively registered on PROSPERO (CRD42024497183), and no additional trial registries were consulted. All retrieved records were imported into EndNote 20 for management. Duplicates were removed using the software's automated function, supplemented by manual verification, to compile a unique set of citations for subsequent screening.

Selection Process

Two reviewers (XZ and LZ) independently appraised the methodological quality of each included meta-analysis using the AMSTAR 2 tool [31]. Any disagreements were resolved through discussion with a third reviewer (SL) to reach consensus. The overall quality of each study was rated as high, moderate, low, or critically low.

Data Items

Two reviewers (XZ and LZ) independently performed data extraction using Microsoft Excel 2019. For the qualitative evidence synthesis, key information was extracted from each meta-analysis. This included all outcome comparisons between RAT and conventional therapy, as well as the moderators examined in prespecified subgroup analyses. The extracted moderators comprised intervention intensity, stroke stage (subacute or chronic), robot type (unilateral vs bilateral), training design (exoskeleton vs end-effector), and baseline upper limb motor impairment level.

Assessment of Overlap

To assess the degree of primary study overlap across the included systematic reviews, we constructed a citation matrix in tabular form and used the corrected covered area (CCA) index to quantify the overall extent of overlap, calculated using the following formula: $CCA = (N - r) \times (r \times c) - r$

where represents the total number of included publications (including duplicates), *r* denotes the number of index publications (the count of unique primary studies or the number of rows in the matrix), and is the number of reviews (or the number of columns in the matrix).

The CCA value was interpreted using commonly applied thresholds derived from Pieper et al [32] work: <5% (slight), 5%-10% (moderate), 11%-15% (high), >15% (very high). For meta-analyses with incomplete information, the following handling approach was adopted: (1) Maximized matching: using the available author and year information, cross-referencing and matching were performed against the references in the other 20 meta-analyses. (2) Conservative estimation: for studies that could not be uniquely matched, they were treated as independent studies in the calculation of the overall CCA to avoid underestimating overlap.

Synthesis Methods

Given the methodological limitations of statistically pooling effect sizes from different meta-analyses, this study used a narrative synthesis approach to summarize and evaluate the evidence. The results for each outcome measure were categorized based on the direction of the effect and its statistical significance: if RAT was significantly superior to conventional therapy, it was classified as "Significant;" if there was no significant difference or if conventional therapy was superior, it was classified as "Non-significant." The determination of statistical significance was based on whether the 95% CIs reported in the original meta-analyses included the null value.

In the synthesis and interpretation of the evidence, we fully incorporated considerations of heterogeneity and risk of bias. For heterogeneity, in addition to examining statistical indicators such as I^2, we focused on analyzing its methodological sources, including population characteristics, intervention protocols, and measurement tools. When heterogeneity was high, interpretations were made cautiously, relying on subgroup results. For the risk of bias, we considered both the methodological quality of the included meta-analyses (assessed using AMSTAR 2) and potential biases in the original studies. These factors were integrated to grade and critically discuss the evidence during result interpretation.

Protocol Amendments

During the study screening phase, the publication date range specified in the registration protocol (CRD42024497183) was updated from December 30, 2013-December 30, 2023, to January 1, 2019-December 26, 2025, to obtain the most recent evidence. All other methodological aspects followed the original protocol.

Results

Overview

From an initial pool of 1307 records identified, 21 meta-analyses [10,16,19-22,33-47] were included following deduplication and multiple rounds of screening. The study selection flow is detailed in Figure 1. The 21 meta-analyses collectively cited primary studies 591 times. After deduplication, 310 independent primary studies were obtained. Among these, multiple studies were repeatedly included in different systematic reviews. All studies that were cited 2 or more times (≥2) and their citation frequencies are summarized in Table S2 in Multimedia Appendix 5. The CCA was 4.532%, indicating a slight degree of overlap among the primary studies in this research.

XSL•FO
RenderX

Figure 1. Flowchart of study selection process in the umbrella review of robot-assisted therapy for upper limb rehabilitation poststroke.

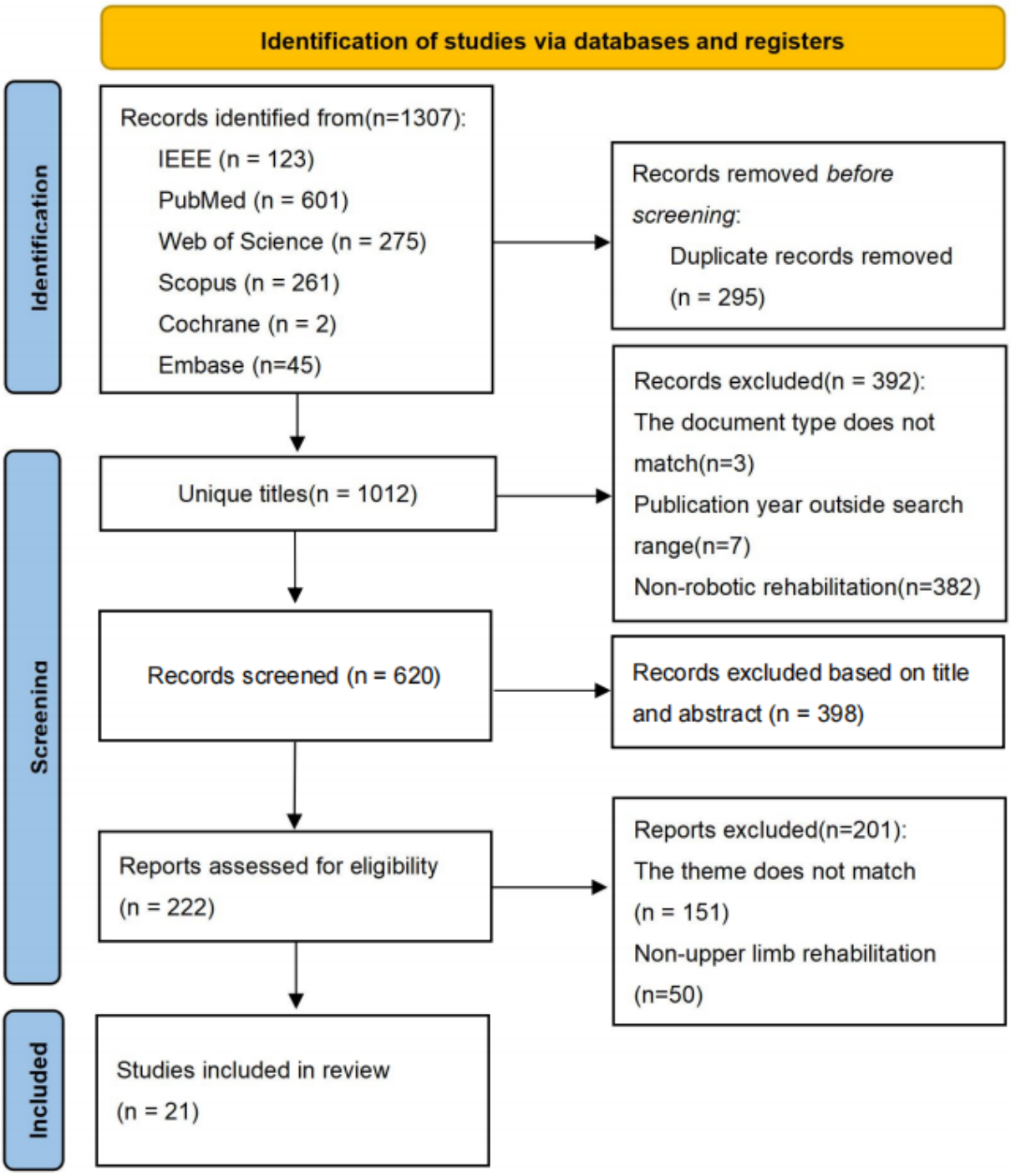

Overview of the Study Included

This review ultimately included 21 meta-analyses [10,16,19-22,33-47], encompassing a total of 535 randomized controlled trials and involving 27,598 patients with upper limb dysfunction after stroke. The study participants included patients in the acute, subacute, and chronic phases of stroke. The types of robots included are end-effector, exoskeleton, soft robotic gloves, etc. Table 1 provides a detailed overview of all the included meta-analyses.

XSL•FO
RenderX

Table 1. Characteristics of the meta-analyses included in the umbrella review on robot-assisted therapy for upper limb rehabilitation post stroke.

Authors	Number of papers	Sample size	Clinical status	Follow-up	Experimental group	Control group	Robotic type	Measurement instrument	Outcome
Tseng et al [33] (2024)	9	295 participants (*EG*[a], *n*=150; *CG*[b], *n*=145)	Survivors with stroke	__[c]	Portable rehabilitation robot. The average duration was 87.78 min. The average number of training sessions per week is 4.56. The average training period was 4.22 weeks	Conventional therapy or no therapy	End-effector-based, exoskeleton-based, and orthosis	FMA[d]	Portable robots prove to be effective (FMA: SMD[e]=0.696, 95%=0.099 to .293, *P*<.05).
Wu et al [34] (2021)	41	1916	Patients with stroke	—	Unilateral RAT[f] or bilateral RAT	Dose-matched conventional rehabilitation	End effector robots, exoskeleton robots	FMA	RAT is an effective intervention for improving upper extremity motor impairment in patients with stroke.
Marotta et al [35] (2021)	26	892	Acute-subacute (<8 weeks) people, chronic people	—	RAT	Conventional therapy	Robotic exoskeleton	FMA	RAT is significantly useful for people with stroke.
Ko et al [36] (2023)	8	309	Patients with post-stroke hemiparesis who had received or were scheduled to receive rehabilitation	—	Rehabilitation programs involving soft robotic gloves or other similar devices	Physical therapy and occupational therapy	Soft robotic gloves or other similar devices	FMA, MBI[g], MAS[h], FIM[i], WMFT[j]	Soft robotic gloves can promote the functional abilities of the upper extremities.
Bazan et al [37] (2022)	9	142, intervention group 73, and control group 69	Adults with an objective diagnosis of unilateral spatial neglect after stroke	—	Robotic limb activation in patients with unilateral spatial neglect after stroke	Conventional rehabilitation for unilateral spatial neglect after stroke	Exoskeleton or end-effector	MVPT-3[k], LBT[l], SCT[m], AT[n], CBS[o]	Limb activation through robotic therapy can improve midline perception. However, there was no impact on tasks assessing visual scanning, functionality, or activities of daily living.

XSL•FO
RenderX

Authors	Number of papers	Sample size	Clinical status	Follow-up	Experimental group	Control group	Robotic type	Measurement instrument	Outcome
Johansen et al [21] (2023)	18	1295, with an age range from 20 to 95 years of age	Persons with stroke	—	Commercialized RAT	Traditional occupational and physiotherapy	End-effector robots, exoskeleton	MAS, Jamar, BI[p]	A statistically significantly higher treatment effect in the robotic-assisted exercise group (*P*=<0.0001) compared to the traditional treatment group, with a total effect size of 0.44 (CI 0.22 to –0.65).
Carrillo et al [19] (2023)	14	1141, the age range of 20 to 85 years, with a median age of 57 years	Participants (age > 18 years) with stroke resulting in functional deficits in their upper extremities.	—	RAT	Conventional therapy	End-effector robots, exoskeleton	FMA	It is unclear whether robot-assisted therapy accelerates upper extremity recovery poststroke when used in conjunction with conventional therapy.
Huo et al [38] (2023)	13	330 (EG, n=175; CG, n=155)	Patients who have been suffering stroke, subacute	—	EMG[q]-based robot	Conventional rehabilitation	EMG-based robot	FMA, MAS, ADL[r]	The outcomes postintervention were significantly improved in the EMG-based robot group.
Chen et al [22] (2020)	35	2241	Patients diagnosed with stroke and having upper limb motor dysfunction	—	RAT. Time per session ranged from 30 minutes to 5 hours. The duration of the intervention ranged from 2 weeks to 12 weeks.	Conventional occupational therapy, physical therapy, task-specific training, ADL training, and constraint-induced movement therapy	Exoskeleton and end-effector	FMA, ARAT[s], WMFT, BBT[t], 9-HPT[u], CAHAI[v], AMAT[w]	RAT was slightly superior in motor impairment recovery.

XSL•FO
RenderX

Authors	Number of papers	Sample size	Clinical status	Follow-up	Experimental group	Control group	Robotic type	Measurement instrument	Outcome
Chien et al [10] (2020)	11	493, the participants were aged 18–65 years	6 months after the onset of stroke	—	The number of sessions for RT ranged from 9 to 40, and each session lasted for 30–120 mins. Participants received RAT for 5 days per week for 2–12 weeks	Usual care	End-effector robots, exoskeleton	FMA, FIM, BI, ARAT, WMFT	RAT produced benefits similar, but not significantly superior, to those from usual care for improving functioning and disability in patients diagnosed with stroke within 6 months.
Iaco et al [39] (2024)	86	4240	Patients with upper limb limitations post-stroke	—	Upper limb-robot therapy. Intervention ranged from 2 weeks to 12 weeks, with a mean upper limb-robot therapy time of 53.55 minutes per session and 4.56	Any other rehabilitative intervention (usual care or specific interventions), placebo, or no treatment	End-effector, exoskeleton	FMA	Small significant improvements in upper limb-muscle synergism, muscle power, motor performance, and basic ADLs.
Moggio et al [40] (2022)	5	149(78 males and 71 females)	Patients with finger-hand motor impairment stroke	—	Exoskeleton and end-effector RAT	Traditional or conventional rehabilitation therapy	End-effector (Amadeo) and exoskeleton	FMA, Quick-DASH[x]	MI[y] showed a significant improvement (P<.05) in the robotic intervention group.
Yang et al [16] (2023)	14	1275	Patients aged 18-80 years with a stroke diagnosis	—	RAT	Any comparative therapy, as well as treatment, as usual or no treatment	End-effector	FMA, MBI, MAS, FIM, WMFT	RAT can significantly enhance the upper limb motor function and activities of daily life in patients with stroke undergoing upper limb rehabilitation.

XSL•FO
RenderX

Authors	Number of papers	Sample size	Clinical status	Follow-up	Experimental group	Control group	Robotic type	Measurement instrument	Outcome
Zhang et al [20] (2022)	46	2533 participants with a mean age ranging from 46.20 to 75.5 years	The stroke patients were over 18 years old	≥ three months	The arm robot used in the intervention group included the Mirror Image Movement Enabler, UL-EXO7, Amadeo Robotic System, InMotion ARM 2.0 Robot, Aremo Spring, Bi-Manu-Track, Myomo e100, Neuro-Rehabilitation Robot, electromyography-driven robot, REJOYCE robot, Pneu-WREX, ReoGo system, and Gloreha robot. Patients received RAT 4 sessions per week for 6 weeks.	Non-robotic therapy	End-effector, exoskeleton, hybrid robot	FMA, MBI, FIM	RAT has significant immediate benefits for motor control and functional activity of the hemiparetic upper limb in patients after stroke.
Zhao et al [41] (2022)	22	758	Participants of either gender over 18 years of age after any duration of stroke	—	Robot-assisted distal training. The training lasted for ≤ 4 weeks with 20 sessions conducted 5 times per week for ≤ 60min each.	Therapist-assisted training or passive range of motion exercise	Exoskeleton robot, end-effector robot, and self-developed devices	FMA, 9-HPT, BBT, MAS, MRC^{z}, MI	The overall effect of robot-assisted distal training on the motor function of the wrists and hands was a significant improvement.
Boardwell et al [42] (2025)	54	2744	Patients with stroke	Ranging from 2 weeks to 6 months	Guide patients to perform high-intensity, repetitive upper limb task training using mechanical-assisted devices.	Conventional rehabilitation	End-effector, Exoskeleton	FIM, BI, ADL, MAL	Robotic rehabilitation had a small, statistically significant positive effect on upper limb capacity compared with conventional rehabilitation (SMD 0.14, 95% CI 0.02-0.26).
Hwang et al [43] (2024)	31	708	Patients with acute stroke	—	RAT as the assistive technology	Conventional treatments	End effectors such as the Armeo Spring, sensor-based devices, InMotion2, ReHapticKnob, and Reo therapy systems.	FMA, ARAT, BBT, WMFT, FIM, MI	Upper-limb robots did not demonstrate significant superiority over conventional treatments in improving the function of upper limbs.

XSL•FO
RenderX

Authors	Number of papers	Sample size	Clinical status	Follow-up	Experimental group	Control group	Robotic type	Measurement instrument	Outcome
Jin et al [44] (2025)	15	574	Patients with stroke	—	The intervention included robot-assisted task-oriented training, without restrictions on the types of robots, training durations, intensities, or frequencies.	Conventional therapies	End-effector, Exoskeleton	FMA-UE[aa], MBI	Robot-assisted task-oriented training significantly enhances the rehabilitation of upper limb function and the recovery of daily living skills in patients with stroke.
Su et al [45] (2024)	18	573	Patients with stroke	—	Upper limb robot rehabilitation training	Conventional rehabilitation	End-effector, Exoskeleton	FMA-UE, WMD[ab], ARAT, MBI, MAS	Upper limb robot–assisted training is superior to conventional training in terms of improving upper limb motor impairment, ability to perform daily living activities, and muscle tone recovery.
Verola et al [46] (2025)	85	3452	Patients with stroke	—	RAT	Conventional physiotherapy	Exoskeleton, End-effector, Hand and Finger Rehabilitation Robots, Bilateral or Mirror-training Robots, Hybrid or Integrated Robotic Systems	FMA, BI, SIS[ac], FIM, BBT, ARAT, WMFT, MAS	RAT produces some significant improvements for the upper limb, but these differences are not clinically relevant when compared to other therapies.

XSL•FO
RenderX

Authors	Number of papers	Sample size	Clinical status	Follow-up	Experimental group	Control group	Robotic type	Measurement instrument	Outcome
Wang et al [47] (2025)	31	1538	Patients with stroke	—	RAT+conventional rehabilitation therapy	Conventional rehabilitation therapy	End-effector robot-assisted therapy, exoskeleton robot-assisted therapy	FIM, FMA-UE, MAS, MBI	RAT combined with routine rehabilitation therapy can effectively improve the upper limb motor function and activities of daily life of patients with stroke.

[a]EG: experimental group.
[b]CG: control group.
[c]NA: not applicable.
[d]FMA: Fugl-Meyer Assessment.
[e]SMD: standardized mean difference.
[f]RAT: robot-assisted therapy.
[g]MBI: Modified Barthel Index.
[h]MAS: Modified Ashworth Scale.
[i]FIM: Functional Independence Measure.
[j]WMFT: Wolf Motor Function Test,
[k]MVPT-3: motor-free visual perception test 3rd.
[l]LBT: line bisection test.
[m]SCT: star cancellation test.
[n]AT: Albert's test.
[o]CBS: Catherine Bergego Scale.
[p]BI: Barthel Index.
[q]EMG: Electromyography.
[r]ADL: activity of daily living.
[s]ARAT: Action Research Arm Test.
[t]BBT: Box and Blocks Test.
[u]9-HPT: Nine Hole Peg Test.
[v]CAHAI: Chedoke Arm and Hand Activity Inventory.
[w]AMAT: Arm Motor Ability Test.
[x]QuickDASH: quick version of disabilities of the arm, shoulder, and hand.
[y]MI: motricity index.
[z]MRC: Medical Research Council Scale.
[aa]FMA-UE: Fugl-Meyer Assessment-upper extremity.
[ab]WMD: weighted mean differences.
[ac]SIS: Stroke Impact Scale.

Methodological Quality of Included Reviews

The methodological quality of the included reviews was assessed using the AMSTAR 2 tool. We synthesized the reported findings from these reviews rather than conducting a new meta-analysis of primary data. According to the AMSTAR 2 appraisal, 17 of the 21 included meta-analyses were of high quality [10,16,20-22,34,35,37,39-47], 3 were of moderate quality [19,33,36], and one was of critically low quality [38] (Table 2). Downgrading was primarily due to factors such as unmet justification for significant statistical heterogeneity [36] or inadequate consideration of potential biases from the primary randomized controlled trials [38]. Despite these limitations, the evidence base is reliable, as the key conclusions are underpinned by a preponderance of medium- to high-quality studies (95.2%), indicating robust primary findings [40].

XSL•FO
RenderX

Table 2. Methodological quality assessment of included meta-analyses using the AMSTAR 2 tool.

Studies	Item 1	Item 2*	Item 3	Item 4*	Item 5	Item 6	Item 7*	Item 8	Item 9*	Item 10	Item 11*	Item 12	Item 13*	Item 14	Item 15*	Item 16	AMSTAR 2 overall rating
Tseng et al [33] (2024)	Yes	Yes	Yes	Partial Yes	No	No	Partial Yes	Partial Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Moderate
Wu et al [34] (2021)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Yes	Partial Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Marotta et al [35] (2021)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Partial Yes	Partial Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Ko et al [36] (2023)	Yes	Partial Yes	Yes	Partial Yes	Yes	Yes	Yes	Partial Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Moderate
Bazan [37] (2022)	Yes	Partial Yes	Yes	Partial Yes	Yes	Yes	Yes	Partial Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Johansen [21] (2023)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Partial Yes	Partial Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Carrillo [19] (2023)	Yes	Partial Yes	Yes	Partial Yes	Yes	Yes	Partial Yes	Partial Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Moderate
Huo [38] (2023)	Yes	Yes	Yes	Partial Yes	Yes	Yes	No	Partial Yes	Yes	No	Yes	Yes	Yes	No	No	Yes	Critically low
Chen et al [22] (2020)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Partial Yes	Partial Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Chien et al [10] (2020)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Partial Yes	Partial Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Iaco et al [39] (2024)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Partial Yes	Partial Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Moggio et al [40] (2022)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Partial Yes	Partial Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Yang et al [16] (2023)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Partial Yes	Partial Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	High
Zhang et al [20] (2022)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Partial Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Zhao et al [41] (2022)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Partial Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High

XSL•FO
RenderX

Studies	Item 1	Item 2*	Item 3	Item 4*	Item 5	Item 6	Item 7*	Item 8	Item 9*	Item 10	Item 11*	Item 12	Item 13*	Item 14	Item 15*	Item 16	AMSTAR 2 overall rating
Boardsworth et al [42] (2025)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Hwang et al [43] (2024)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	High
Jin et al [44] (2025)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Su et al [45] (2024)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Partial Yes	Yes	No	Yes	Yes	High
Verola et al [46] (2025)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Wang et al [45] (2025)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Partial Yes	Yes	No	Yes	Yes	High

*Indicates that the corresponding item is critical.

Primary Outcome: Upper Limb Motor Function

The Fugl-Meyer Assessment (FMA) is the “gold standard” for assessing upper-extremity sensorimotor function, due to its ability to assess aspects such as movement within synergies, mixing synergies, reflexes, wrist, hand, grip, coordination, and speed movements, thus providing a large amount of information that is very useful for understanding the sensorimotor capacity of the affected upper limb after a stroke [48]. The FMA was adopted as the primary outcome measure in 18 out of the 21 included meta-analyses (85.7%) to evaluate the restoration of upper limb motor function.

However, the specific clinical questions explored by various studies using FMA differ significantly, covering multiple dimensions such as the immediacy and sustainability of therapeutic effects [36], responses in populations at different stroke stages [16,21], and impacts on other functional domains [35]. By synthesizing these studies, the consistent use of FMA provides highly comparable and reliable evidence for a comprehensive, multifaceted evaluation of the efficacy of robot-assisted therapy.

Immediate Effects on Upper Limb Function

This study synthesizes evidence on the immediate effects of RAT from multiple meta-analyses. As shown in Table 3, various studies used standardized mean difference (SMD), mean difference (MD), weighted mean differences (WMD), and Hedges gas effect size metrics. Eight independent effect size estimates from different meta-analyses demonstrated statistically significant positive effects of RAT [20,33,34,36,42,44,45,47], supporting its immediate benefits. However, one study reported a non-significant result [43], and variations were observed in the magnitude of effects. The meta-analysis demonstrated moderate heterogeneity (I^2=35%), with variations among the included studies in terms of population characteristics, device types, and treatment protocols. Moreover, the study’s own assessment indicated potential risks of bias in areas such as intervention implementation and completeness of outcome data.

XSL•FO
RenderX

Table 3. Summary of meta-analyses on the immediate effects of robot-assisted therapy compared to conventional therapy on upper limb function in patients with stroke.

Studies	Effect size	Value	95% CI	Statistical significance
Tseng et al [33] (2024)	SMD[a]	0.696	0.09-1.293	Significant
Zhang [20] (2022)	SMD	0.2	0.08-0.32	Significant
Ko et al [36] (2023)	MD[b]	6.52	3.65-9.39	Significant
Wu et al [34] (2021)	Hedges *g*	0.25	0.11-0.38	Significant
Boardsworth et al [42] (2025)	SMD	0.14	0.02-0.26	Significant
Hwang et al [43] (2024)	MD	4.99	–0.07 to 10.05	Non-significant
Jin et al [44] (2025)	SMD	1.01	0.57-1.45	Significant
Su et al [45] (2024)	WMD	5.27	3.36-7.17	Significant
Wang et al [47] (2025)	MD[c]	5.92	3.52-8.32	Significant

[a]SMD: standardized mean difference.

[b]MD: mean difference.

[c]WMD: weighted mean differences.

Long-Term Maintenance of Upper Limb Function

Regarding the long-term maintenance effects of RAT, this review confirms that RAT demonstrates sustained short-term efficacy in improving upper limb function [20,36,42,45], as shown in Table 4. Meanwhile, our comprehensive analysis reveals a significant time-gradient effect: while significant functional improvements were observed immediately after the intervention and during short-term follow-up (≤12 weeks) [36,45], these advantages generally diminished or even disappeared during long-term follow-up (>12 weeks) [20,42]. This time-dependent pattern of efficacy decay indicates notable limitations in current robotic rehabilitation strategies for maintaining long-term therapeutic effects.

Table 4. Effects of robot-assisted therapy on the long-term maintenance of upper limb function in stroke survivors, stratified by follow-up duration.

Studies	Follow-up	Effect size	Value	95% CI	Statistical significance
Ko et al [36] (2023)	≤12 weeks	MD[a]	7.79	5.03-10.55	Significant
Zhang et al [20] (2022)	>12weeks	SMD[b]	–0.07	–0.21 to 0.07	Non-significant
Boardsworth et al [42] (2025)	NA	SMD	0.05	–0.13 to 0.24	Non-significant
Su et al [45] (2024)	＞4weeks	WMD[c]	6.63	3.46-9.8	Significant
Su et al [45] (2024)	≤4 weeks	WMD	4.49	2.11-6.88	Significant

[a]MD: mean difference.

[b]SMD: standardized mean difference.

[c]WMD: weighted mean differences.

Secondary Outcomes

Spasticity

The impact of RAT on spasticity represents another critical dimension for evaluation. The Motor Assessment Scale (MAS) serves as a core, standardized tool for assessing upper limb spasticity following stroke [49]. Although 4 meta-analyses have investigated this issue [16,21,38,41], and multiple studies have reported that robotic training can significantly improve spasticity, as shown in Table 5, their conclusions have been challenged by key research. For instance, some studies contained directional contradictions or reporting flaws, making the results difficult to interpret reasonably [41].

XSL•FO
RenderX

Table 5. Meta-analysis results of the effects of robot-assisted therapy on upper limb spasticity (measured by Modified Ashworth Scale) in patients after stroke.

Studies	Effect size	Value	95% CI	*P* value	Statistical significance
Johansen et al [21] (2023)	MD[a]	–0.24	–1.33 to 0.22	<.00001	Significant
Huo et al [38] (2023)	MD	–0.42	–0.82 to –0.03	.03	Non-significant
Zhao et al [41] (2022)	MD	0.18	–0.32 to –0.04	.01	Significant
Yang et al [16] 2023	SMD[b]	–1.49	–2.85 to –0.12	.05	Non-significant

[a]MD: mean difference.

[b]SMD: standardized mean difference.

Current evidence demonstrates marked heterogeneity (MD –0.42, 95% CI –0.82 to –0.03, I^2=83%, *P*=.03) [38], primarily due to differences among the included studies in terms of population characteristics, intervention protocols, and assessment methods. This heterogeneity makes it difficult to directly compare or synthesize findings across studies. Therefore, the use of the RAT was not more significant than conventional treatment in improving spasticity [16].

Activity of Daily Living

Six meta-analyses evaluated the effect of robot-assisted therapy on the ADL of patients with stroke [20-22,41,42,46]. The evidence regarding its overall immediate effects remains inconsistent, as shown in Table 6. Although some studies reported significant positive effects at the end of the intervention period, heterogeneity among the studies limits the robustness of the conclusions [20,41].

Table 6. Effect sizes for the impact of robot-assisted therapy on the activity of daily living in patients with stroke.

Studies	SMD[a]	95% CI	*P* value	Statistical significance
Johansen et al [21] (2023)	0.11	–0.04 to 0.25	.17	Non-significant
Chen et al [22] (2023)	0.0049	–0.055 to 0.17	.15	Non-significant
Zhang et al [20] (2022)	0.32	0.16-0.47	<.0001	Significant
Zhao et al [41] (2022)	0.7	–0.29 to 1.11	<.001	Significant
Boardsworth et al [42] (2025)	0.04	–0.05 to 0.13	.86	Non-significant
Verola et al [46] 2025	0.29	0.15-0.43	＜.01	Significant

[a]SMD: standardized mean difference.

Long-term follow-up data further indicate that the initial therapeutic benefits may not be sustainable. For instance, a meta-analysis by Zhang et al [20] demonstrated that the effect size diminished and became statistically non-significant during the follow-up period (SMD=0.09, 95% CI –0.06 to 0.23, I^2=38%), suggesting that the improvement in ADL from RAT may diminish over time. Subgroup analysis provided key insights for understanding the heterogeneity in treatment effects, indicating that the degree of ADL improvement is moderated by patients' baseline upper limb motor impairment levels and robotic training modalities. Among these, the level of active engagement during training emerged as a significant influencing factor [20].

Subgroup Analysis Results

Device Types and Patient Upper Limb Motor Impairment Levels

Based on current evidence, the differential efficacy of RAT for patients with mild-to-moderate versus severe upper limb motor impairment demonstrates a complex interaction pattern, as shown in Table 7. Device type may well be the key moderating factor explaining this heterogeneity. Wu et al [34] demonstrated that end-effector robots exhibited clear advantages (Hedges *g*=0.22, 95% CI 0.09-0.36, I^2=35.4%). This finding provides a crucial explanation for their overall subgroup results. However, the study by Su et al [45] found that exoskeletons (WMD 6.90, 95% CI 4.33-9.47) demonstrated a superior effect on upper limb motor impairment compared to end-effector devices (WMD 3.28, 95% CI 0.44-6.12).

XSL•FO
RenderX

Table 7. Subgroup analysis of the efficacy of robot-assisted therapy on upper limb function in patients with stroke, stratified by baseline motor impairment level.

Subgroups and studies	Effect size	Value	95% CI	Statistical significance
Mild-to-moderate				
Wu et al [34] (2021)	Hedges *g*	0.19	–0.01 to 0.4	Non-significant
Zhang et al [20] (2022)	SMD[a]	0.26	0.09-0.42	Significant
Severe				
Wu et al [34] (2021)	Hedges *g*	0.27	0.08-0.46	Significant
Zhang et al [20] (2022)	SMD	0.14	–0.01 to 0.3	Non-significant

[a]SMD: standardized mean difference.

The meta-analysis by Zhang et al [20], which included end-effector, exoskeleton, and hybrid robotic devices, yielded results more supportive of the therapy's benefits for patients with mild-to-moderate upper limb motor impairment. The discrepancy between the findings of these 2 studies strongly suggests a specific matching relationship between robot types and patients' upper limb motor impairment levels: end-effector robots may hold unique value for patients with severe impairment, while more complex devices like hybrid systems may provide additional benefits for those with mild-to-moderate upper limb motor impairment [20].

Stroke Stage

Based on current meta-analytical evidence, the efficacy of RAT for upper limb rehabilitation following stroke demonstrates a distinct phase-dependent characteristic [50]. As shown in Table 8, within the population of patients with chronic stroke, the evidence demonstrates a high degree of consistency. Six studies [16,21,34-36,38] reported statistically significant functional improvements, supporting the effectiveness of this therapy during this phase.

Table 8. Subgroup analysis comparing the efficacy of robot-assisted therapy on upper limb function in patients with subacute stroke.

Subacute subgroup and studies	Effect size	Value	95% CI	Statistical significance
Johansen et al [21] (2023)	SMD[a]	0.5	0.2-0.8	Significant
Huo et al [38] (2023)	SMD	1.18	0.29-2.07	Significant
Marotta et al [35] (2021)	SMD	4.09	1.31-6.87	Significant
Yang et al [16] (2023)	SMD	–0.16	–0.56 to 0.24	Non-significant

[a]SMD: standardized mean difference.

In contrast, the evidence for patients in the subacute phase demonstrates marked heterogeneity. Among four relevant studies [16,21,35,38], three reported significant benefits [21,35,38], while one showed no significant effects [36], indicating that a clear consensus regarding its efficacy at this stage has yet to be established. These divergent outcomes may primarily stem from variations in intervention protocols, types of robotic devices, training dosage, and program design.

Secondly, the ambiguity in study population definitions may introduce bias. The criteria for defining the "subacute phase" across studies could lead to differences in the actual recovery stages and potential of included patients [51]. Finally, methodological heterogeneity cannot be overlooked. Variations in inclusion criteria, statistical methods, and bias risk assessment among different meta-analyses could all affect the final synthesis of results. These factors collectively highlight the need for caution when interpreting the evidence, and future research should aim to clarify these complex relationships through more detailed subgroup analyses and standardized definition criteria.

Discussion

Principal Findings

Based on the systematic synthesis of existing meta-analytical evidence conducted in this umbrella review, we are able to examine the value and position of RAT in poststroke upper limb rehabilitation from a broader perspective. Importantly, our analysis is guided by the ICF framework, which provides a structured approach to evaluating outcomes across different dimensions of recovery. The main findings, clinical implications, and future research directions of this study are outlined as follows.

Efficacy Across ICF Levels: From Body Function to Activity

This review confirms that RAT demonstrates certain efficacy in improving upper limb motor function at the ICF body functions level, as measured by the FMA [36]. However, our synthesis reveals a crucial gradient in treatment effects across ICF levels. First, the therapeutic effects show significant stage-dependency, with the most robust and consistent evidence supporting its effectiveness for patients in the chronic phase

XSL•FO
RenderX

[16,34,35], while its efficacy for subacute patients remains inconsistent due to variations in intervention protocols and patient characteristics [16]. Second, a prominent finding is the limited translation of benefits to the ICF activity level. Although it effectively improves upper limb function and reduces spasticity [34,41], its effects on higher-level functions such as hand dexterity, ADL, and social participation are limited and inconsistent [22,41].

Specifically, when assessing manual dexterity, studies using the Nine-Hole Peg Test demonstrated that the robot-assisted group completed tasks significantly faster than the conventional therapy group [22,38], whereas studies using the Box and Block Test found no significant advantage [41]. More importantly, although RAT performed no worse than conventional therapist-guided training in improving both ADL and enhancing social participation, neither domain showed statistically significant superiority for the robotic approach [22,41]. This dissociation between body functions and activity-level outcomes represents a key challenge in robotic rehabilitation.

Moderators of Treatment Response

Within the ICF framework, we identified several critical moderators that explain the substantial heterogeneity in treatment effects. Intervention intensity is a critical factor influencing therapeutic efficacy. Zhang et al [20] found that when the total treatment duration exceeded 15 hours, the RAT group showed significant improvements in both motor control and functional activity. However, Yang et al [16] reported that excessively long treatment cycles may lead to diminishing returns. Notably, Carrillo et al [19] evaluated the impact of different training intensities on FMA scores and demonstrated that higher-intensity interventions were most beneficial for functional recovery. Furthermore, multiple dimensions of training intensity (session duration, frequency, total cycle) need to be considered synergistically [45,47]. Johansen et al [21] found that significant therapeutic effects were achieved when the total intervention dose reached 1375.33 minutes, while Ko et al [36] subgroup analysis regarding total training duration indicated no significant improvements, regardless of whether the training period exceeded or fell short of 2 weeks. This finding suggests that we need to consider interactions among dose parameters [36].

Device Characteristics and Training Modalities

Technical characteristics significantly moderate therapeutic efficacy. Wu et al [34] found that end-effector robots demonstrated superior effectiveness in improving upper limb motor impairment compared to conventional rehabilitation therapy (Hedges g=0.22; 95% CI 0.09-0.36, I^2=35.4%), potentially attributable to their multijoint coordination training mechanism [52]. Moggio et al [40] further compared device types and revealed that exoskeleton devices showed advantages in enhancing overall hand function and reducing disability, while the end-effector device demonstrated better performance in specific motor control. The selection of training modes is equally crucial for achieving functional gains. Wu et al [34] confirmed that unilateral robot training demonstrated clear efficacy, whereas the advantages of bilateral training and combined unilateral-bilateral training regimens remain unsubstantiated [34,38]. Yang et al [16] emphasized that training modes requiring active patient participation are essential for achieving improvements in ADL.

Integration of Secondary Outcomes

When comprehensively evaluating the effects of RAT on poststroke upper limb function, changes in muscle strength, as one of the core indicators at the ICF body functions level, hold significant clinical reference value. This umbrella review synthesizing relevant meta-analyses found that robotic therapy demonstrates a positive trend in improving muscle strength, though the strength of evidence remains limited. For instance, the analysis by Zhao et al [41] indicated that the upper limb strength in the RAT group was significantly greater than in the control group (SMD=0.42, 95% CI 0.07-0.78, I^2=4%), a finding consistent with the results reported by Johansen et al [21] (SMD=0.43, 95% CI 0.16-0.71, I^2=46%). However, evidence in this area shows inconsistency, as some studies found no significant between-group differences (MD=0.51, 95% CI –0.06 to 1.09, I^2=0%) [41]. This heterogeneity may stem from differences in assessment tools, insufficient specificity of training protocols for strength enhancement, and variations in patients' baseline levels.

Clinical Translation and Health Economic Perspectives

From a health economic perspective, the cost-effectiveness of rehabilitation robotics still warrants careful evaluation. Currently, the price range for such assistive robotic products varies widely, ranging from US $9000 to US $100,000 [33,53]. Robot-assisted rehabilitation requires not only a high initial investment but also involves relatively high system maintenance and operational costs [54,55]. Notably, the ICF framework helps explain the current cost-effectiveness challenges: the clinical benefits are largely confined to the body functions level, with limited impact on the activity level and virtually no high-quality evidence regarding participation-level outcomes. Furthermore, the implementation of robotic rehabilitation is influenced by health care system disparities and cultural contexts [56]. The high costs may pose significant barriers in resource-limited settings, while cultural differences in technology acceptance can affect patient engagement and treatment adherence [57]. Concurrently, the clinical benefits derived from robotic rehabilitation by both patients and health care institutions remain relatively limited, resulting in a cost-effectiveness ratio that has not yet reached an ideal level [58].

Furthermore, patients' attitudes toward RAT present a mixed picture. Multiple studies indicate that participants perceive robotic therapy as more precise and less invasive, and hold higher expectations regarding treatment outcomes [59,60]. Such positive perceptions may enhance clinical acceptance of the technology. However, a considerable proportion of patients express concerns about the risk of device malfunctions, with some even mistakenly believing that robotic therapy has higher error rates than conventional methods [60-62]. This crisis of trust could significantly impact treatment adherence and ultimately compromise therapeutic outcomes [63].

Currently, evidence regarding the ability of rehabilitation robots to reduce overall health care costs remains insufficient, and their

XSL•FO
RenderX

potential advantages in controlling medical expenditures require further validation [64]. Considering multiple factors, including equipment costs, patient acceptance, health care system differences, and cultural contexts, the pathway for clinical integration of robot-assisted rehabilitation still demands careful planning. Future research should not only focus on more rigorous economic evaluations but also address how to enhance patient trust through technical transparency and improved clinician-patient communication, while developing implementation strategies adapted to diverse health care systems and cultural environments, thereby ultimately improving the cost-benefit profile.

Limitations and Future Directions

It should be noted that this umbrella review has several limitations. First, although our synthesis approach accounted for heterogeneity and bias, the inability to reanalyze individual participant data from the primary studies limited a deeper exploration of heterogeneity sources. Second, some included meta-analyses contained methodological flaws or had missing outcome data, factors that could affect the accuracy of the conclusions. Additionally, as a secondary synthesis of existing evidence, this review cannot circumvent potential biases inherent in the evidence base, including possible publication bias (where studies with positive results are more likely to be published) and the fact that some primary studies may have received commercial funding or involved conflicts of interest. Such factors could compromise the objectivity of the evidence. Third, the comprehensive cross-checking and deduplication of the included systematic reviews and meta-analyses, which themselves contain extensive references to indexed primary studies, required immense and highly time-consuming effort [65]. This posed a significant methodological challenge in the implementation of this study.

Building on the ICF framework, we recommend several key directions for future research. Future studies should prioritize the investigation of participation level outcomes to fully understand the technology's impact on patients' lives. Furthermore, research should explicitly target the transfer of gains from body functions to activities and participation levels. Future umbrella reviews could enhance the reliability of their findings by expanding the literature search scope, strengthening the assessment of bias risks in primary studies, and raising the quality thresholds for included studies. In future research, artificial intelligence technologies could be leveraged to further optimize the identification and processing of literature overlap, thereby enhancing both efficiency and accuracy.

These limitations suggest that the conclusions of this review should be interpreted with due caution. However, despite these limitations, our application of the ICF framework provides a robust structure for understanding and advancing the field of robot-assisted upper limb rehabilitation.

Conclusions

This umbrella review demonstrates that RAT serves as an effective intervention for poststroke upper limb rehabilitation, with the most robust and consistent evidence supporting its efficacy in improving motor function, particularly demonstrating clear benefits for patients in the chronic phase. However, its therapeutic advantages are primarily concentrated in short-term functional improvement, while long-term efficacy maintenance remains insufficient, constituting a core bottleneck in current clinical application. The therapeutic effect is significantly modulated by robot type, training mode, and intervention intensity. Future research should focus on developing stratified and individualized treatment protocols based on patient injury characteristics, rehabilitation stage, and device features, and conduct cross–health care system health economic evaluations to advance the clinical translation of precision rehabilitation technologies.

Data Availability

All data generated or analyzed during this study are included in its supplementary information files.

Funding

This study was funded by grants from the National Natural Science Foundation of China (No.82102680) and the Sichuan Provincial Department of Science and Technology (No.2024YFFK0143). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Authors' Contributions

Conceptualization: LS (lead) and ZX (equal)
Data curation: ZX
Formal analysis: ZX (lead), ZL (supporting), and ZJ (supporting)
Funding acquisition: LS
Investigation: LS
Methodology: LS and ZX
Project administration: LS (lead), LJ (equal), and HC (equal)
Supervision: HC (lead) and LJ (equal)
Validation: HC
Visualization: ZX
Writing – original draft: LS (lead) and ZX (equal)

XSL•FO
RenderX

Writing – review and editing:LS (lead) and ZX (equal)

Conflicts of Interest

The authors declare that they have no competing interests.

Multimedia Appendix 1
PRISMA Checklist.
[PDF File (Adobe PDF File), 97 KB - jmir_v28i1e79363_app1.pdf]

Multimedia Appendix 2
PRISMA-S Checklist.
[PDF File (Adobe PDF File), 65 KB - jmir_v28i1e79363_app2.pdf]

Multimedia Appendix 3
PRIOR Checklist.
[PDF File (Adobe PDF File), 260 KB - jmir_v28i1e79363_app3.pdf]

Multimedia Appendix 4
Search processes for different databases.
[PDF File (Adobe PDF File), 73 KB - jmir_v28i1e79363_app4.pdf]

Multimedia Appendix 5
Studies cited two or more times and their citation frequencies.
[PDF File (Adobe PDF File), 96 KB - jmir_v28i1e79363_app5.pdf]

References

1. Kuriakose D, Xiao Z. Pathophysiology and treatment of stroke: present status and future perspectives. Int J Mol Sci 2020;21(20):7609 [FREE Full text] [doi: 10.3390/ijms21207609] [Medline: 33076218]
2. Langhorne P, Coupar F, Pollock A. Motor recovery after stroke: a systematic review. Lancet Neurol 2009;8(8):741-754. [doi: 10.1016/S1474-4422(09)70150-4] [Medline: 19608100]
3. Wen X, Li L, Li X, Zha H, Liu Z, Peng Y, et al. Therapeutic role of additional mirror therapy on the recovery of upper extremity motor function after stroke: a single-blind, randomized controlled trial. Neural Plast 2022;2022:8966920 [FREE Full text] [doi: 10.1155/2022/8966920] [Medline: 36624743]
4. Hu J, Zhao L. Rehab mirror: the application of virtual mirror therapy in post-stroke upper limb rehabilitation. : IEEE; 2024 Presented at: Proceedings of the 17th International Symposium on Computational Intelligence and Design (ISCID); December 14-15, 2024; Hangzhou, China p. 222-225. [doi: 10.1109/iscid63852.2024.00057]
5. Shen YC, Kim AS, Hsia RY. Treatments and patient outcomes following stroke center expansion. JAMA Netw Open 2024;7(11):e2444683 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.44683] [Medline: 39535793]
6. Morone G, Cocchi I, Paolucci S, Iosa M. Robot-assisted therapy for arm recovery for stroke patients: state of the art and clinical implication. Expert Rev Med Devices 2020;17(3):223-233. [doi: 10.1080/17434440.2020.1733408] [Medline: 32107946]
7. Zhang Y, Zhang Y, Zheng B, Chen S, Yu H, Dai L, et al. The effects of combining anodal transcranial direct current stimulation with robot-assisted gait training on lower limb motor function and the motor cortex regulation of stroke patients. J Neuroeng Rehabil 2025;22(1):230 [FREE Full text] [doi: 10.1186/s12984-025-01731-8] [Medline: 41163056]
8. Xie H, Li X, Huang W, Yin J, Luo C, Li Z, et al. Effects of robot-assisted task-oriented upper limb motor training on neuroplasticity in stroke patients with different degrees of motor dysfunction: a neuroimaging motor evaluation index. Front Neurosci 2022;16:957972 [FREE Full text] [doi: 10.3389/fnins.2022.957972] [Medline: 36188465]
9. Singh N, Saini M, Kumar N, Srivastava MVP, Mehndiratta A. Evidence of neuroplasticity with robotic hand exoskeleton for post-stroke rehabilitation: a randomized controlled trial. J Neuroeng Rehabil 2021;18(1):76 [FREE Full text] [doi: 10.1186/s12984-021-00867-7] [Medline: 33957937]
10. Chien WT, Chong YY, Tse MK, Chien CW, Cheng HY. Robot-assisted therapy for upper-limb rehabilitation in subacute stroke patients: a systematic review and meta-analysis. Brain Behav 2020;10(8):e01742 [FREE Full text] [doi: 10.1002/brb3.1742] [Medline: 32592282]
11. Mehrholz J, Pohl M, Platz T, Kugler J, Elsner B. Electromechanical and robot-assisted arm training for improving activities of daily living, arm function, and arm muscle strength after stroke. Cochrane Database Syst Rev 2018;9(9):CD006876 [FREE Full text] [doi: 10.1002/14651858.CD006876.pub5] [Medline: 30175845]

XSL•FO
RenderX

12. Mehrholz J, Pollock A, Pohl M, Kugler J, Elsner B. Systematic review with network meta-analysis of randomized controlled trials of robotic-assisted arm training for improving activities of daily living and upper limb function after stroke. J Neuroeng Rehabil 2020;17(1):83 [FREE Full text] [doi: 10.1186/s12984-020-00715-0] [Medline: 32605587]
13. Park JM, Park HJ, Yoon SY, Kim YW, Shin JI, Lee SC. Effects of robot-assisted therapy for upper limb rehabilitation after stroke: an umbrella review of systematic reviews. Stroke 2025;56(5):1243-1252. [doi: 10.1161/STROKEAHA.124.048183] [Medline: 40115991]
14. Coskunsu DK, Akcay S, Ogul OE, Akyol D, Ozturk N, Zileli F, et al. Effects of robotic rehabilitation on recovery of hand functions in acute stroke: a preliminary randomized controlled study. Acta Neurol Scand 2022;146(5):499-511. [doi: 10.1111/ane.13672] [Medline: 35855628]
15. Bertani R, Melegari C, De Cola MC, Bramanti A, Bramanti P, Calabrò RS. Effects of robot-assisted upper limb rehabilitation in stroke patients: a systematic review with meta-analysis. Neurol Sci 2017;38(9):1561-1569. [doi: 10.1007/s10072-017-2995-5] [Medline: 28540536]
16. Yang X, Shi X, Xue X, Deng Z. Efficacy of robot-assisted training on rehabilitation of upper limb function in patients with stroke: a systematic review and meta-analysis. Arch Phys Med Rehabil 2023;104(9):1498-1513. [doi: 10.1016/j.apmr.2023.02.004] [Medline: 36868494]
17. Qu H, Zeng F, Tang Y, Shi B, Wang Z, Chen X, et al. The clinical effects of brain-computer interface with robot on upper-limb function for post-stroke rehabilitation: a meta-analysis and systematic review. Disabil Rehabil Assist Technol 2024;19(1):30-41. [doi: 10.1080/17483107.2022.2060354] [Medline: 35450498]
18. Carrión-Téllez V, Pastor-Zaplana J, Compañ-Gabucio LM, Peral-Gómez P, García-Aracil NM. Use of robotic devices in upper limb rehabilitation for adult stroke patients: protocol for an umbrella review. Health Sci Rep 2026;9(1):e71684 [FREE Full text] [doi: 10.1002/hsr2.71684] [Medline: 41467237]
19. Carrillo C, Tilley D, Horn K, Gonzalez M, Coffman C, Hilton C, et al. Effectiveness of robotics in stroke rehabilitation to accelerate upper extremity function: systematic review. Occup Ther Int 2023;2023:7991765 [FREE Full text] [doi: 10.1155/2023/7991765] [Medline: 37927581]
20. Zhang L, Jia G, Ma J, Wang S, Cheng L. Short and long-term effects of robot-assisted therapy on upper limb motor function and activity of daily living in patients post-stroke: a meta-analysis of randomized controlled trials. J Neuroeng Rehabil 2022;19(1):76 [FREE Full text] [doi: 10.1186/s12984-022-01058-8] [Medline: 35864524]
21. Johansen T, Sørensen L, Kolskår KK, Strøm V, Wouda MF. Effectiveness of robot-assisted arm exercise on arm and hand function in stroke survivors - a systematic review and meta-analysis. J Rehabil Assist Technol Eng 2023;10:20556683231183639 [FREE Full text] [doi: 10.1177/20556683231183639] [Medline: 37426037]
22. Chen Z, Wang C, Fan W, Gu M, Yasin G, Xiao S, et al. Robot-assisted arm training versus therapist-mediated training after stroke: a systematic review and meta-analysis. J Healthc Eng 2020;2020:8810867 [FREE Full text] [doi: 10.1155/2020/8810867] [Medline: 33194159]
23. Jain R, Meena ML, Rana KB. Risk factors of musculoskeletal symptoms among mobile device users during work from home. Int J Occup Saf Ergon 2022 Dec;28(4):2262-2268. [doi: 10.1080/10803548.2021.1979318] [Medline: 34514964]
24. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
25. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
26. Gates M, Gates A, Pieper D, Fernandes RM, Tricco AC, Moher D, et al. Reporting guideline for overviews of reviews of healthcare interventions: development of the PRIOR statement. BMJ 2022;378:e070849 [FREE Full text] [doi: 10.1136/bmj-2022-070849] [Medline: 35944924]
27. Chen P, Liu T, Tse MMY, Lai CKY, Tsoh J, Ng SSM. The predictive role of hand section of fugl-meyer assessment and motor activity log in action research arm test in people with stroke. Front Neurol 2022;13:926130 [FREE Full text] [doi: 10.3389/fneur.2022.926130] [Medline: 35873769]
28. Naqvi U, Margetis K, Sherman A. Muscle Strength Grading. Treasure Island (FL): StatPearls Publishing URL: https://www.ncbi.nlm.nih.gov/books/NBK436008/ [accessed 2025-11-03]
29. Freire B, Bochehin do Valle M, Lanferdini FJ, Foschi CVS, Abou L, Pietta-Dias C. Cut-off score of the Modified Ashworth Scale corresponding to walking ability and functional mobility in individuals with chronic stroke. Disabil Rehabil 2023;45(5):866-870. [doi: 10.1080/09638288.2022.2037753] [Medline: 35171052]
30. Hong I, Woo HS, Shim S, Li C, Yoonjeong L, Velozo CA. Equating activities of daily living outcome measures: the functional independence measure and the Korean version of modified barthel index. Disabil Rehabil 2018;40(2):217-224. [doi: 10.1080/09638288.2016.1247468] [Medline: 27930889]
31. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017;358:j4008 [FREE Full text] [doi: 10.1136/bmj.j4008] [Medline: 28935701]

XSL•FO
RenderX

32. Pieper D, Antoine S, Mathes T, Neugebauer EAM, Eikermann M. Systematic review finds overlapping reviews were not mentioned in every other overview. J Clin Epidemiol 2014;67(4):368-375. [doi: 10.1016/j.jclinepi.2013.11.007] [Medline: 24581293]
33. Tseng KC, Wang L, Hsieh C, Wong AM. Portable robots for upper-limb rehabilitation after stroke: a systematic review and meta-analysis. Ann Med 2024;56(1):2337735 [FREE Full text] [doi: 10.1080/07853890.2024.2337735] [Medline: 38640459]
34. Wu J, Cheng H, Zhang J, Yang S, Cai S. Robot-assisted therapy for upper extremity motor impairment after stroke: a systematic review and meta-analysis. Phys Ther 2021;101(4):pzab010. [doi: 10.1093/ptj/pzab010] [Medline: 33454787]
35. Marotta N, Demeco A, Moggio L, Ammendolia A. The adjunct of transcranial direct current stimulation to robot-assisted therapy in upper limb post-stroke treatment. J Med Eng Technol 2021;45(6):494-501. [doi: 10.1080/03091902.2021.1922527] [Medline: 34038313]
36. Ko MJ, Chuang YC, Ou-Yang LJ, Cheng YY, Tsai YL, Lee YC. The application of soft robotic gloves in stroke patients: a systematic review and meta-analysis of randomized controlled trials. Brain Sci 2023;13(6):900 [FREE Full text] [doi: 10.3390/brainsci13060900] [Medline: 37371378]
37. Bazan R, Fonseca BHDS, Miranda JMDA, Nunes HRDC, Bazan SGZ, Luvizutto GJ. Effect of robot-assisted training on unilateral spatial neglect after stroke: systematic review and meta-analysis of randomized controlled trials. Neurorehabil Neural Repair 2022;36(8):545-556. [doi: 10.1177/15459683221110894] [Medline: 35880666]
38. Huo Y, Wang X, Zhao W, Hu H, Li L. Effects of EMG-based robot for upper extremity rehabilitation on post-stroke patients: a systematic review and meta-analysis. Front Physiol 2023;14:1172958 [FREE Full text] [doi: 10.3389/fphys.2023.1172958] [Medline: 37256069]
39. De Iaco L, Veerbeek JM, Ket JCF, Kwakkel G. Upper limb robots for recovery of motor arm function in patients with stroke: a systematic review and meta-analysis. Neurology 2024;103(2):e209495. [doi: 10.1212/WNL.0000000000209495] [Medline: 38870442]
40. Moggio L, de Sire A, Marotta N, Demeco A, Ammendolia A. Exoskeleton end-effector robot-assisted therapy for finger-hand motor recovery in stroke survivors: systematic review and meta-analysis. Top Stroke Rehabil 2022;29(8):539-550. [doi: 10.1080/10749357.2021.1967657] [Medline: 34420498]
41. Zhao M, Wang G, Wang A, Cheng LJ, Lau Y. Robot-assisted distal training improves upper limb dexterity and function after stroke: a systematic review and meta-regression. Neurol Sci 2022;43(3):1641-1657. [doi: 10.1007/s10072-022-05913-3] [Medline: 35089447]
42. Boardsworth K, Rashid U, Olsen S, Rodriguez-Ramirez E, Browne W, Alder G, et al. Upper limb robotic rehabilitation following stroke: a systematic review and meta-analysis investigating efficacy and the influence of device features and program parameters. J Neuroeng Rehabil 2025;22(1):164 [FREE Full text] [doi: 10.1186/s12984-025-01662-4] [Medline: 40671078]
43. Hwang S, Min KC, Song CS. Assistive technology on upper extremity function for stroke patients: a systematic review with meta-analysis. J Hand Ther 2024;37(4):507-519. [doi: 10.1016/j.jht.2023.12.014] [Medline: 38796397]
44. Jin C, Chen Y, Ma Y. Effectiveness of robot-assisted task-oriented training intervention for upper limb and daily living skills in stroke patients: a meta-analysis. PLoS One 2025;20(1):e0316633 [FREE Full text] [doi: 10.1371/journal.pone.0316633] [Medline: 39752454]
45. Su T, Wang M, Chen Z, Feng L. Effect of upper robot-assisted training on upper limb motor, daily life activities, and muscular tone in patients with stroke: a systematic review and meta-analysis. Brain Behav 2024;14(11):e70117 [FREE Full text] [doi: 10.1002/brb3.70117] [Medline: 39482838]
46. Verola S, Ugolini A, Pellicciari L, Di Bari M, Paci M. Clinical relevance of the effects of robotic rehabilitation for upper limb recovery after stroke in randomized studies: a systematic review with meta-analysis. Arch Physiother 2025;15:118-130 [FREE Full text] [doi: 10.33393/aop.2025.3209] [Medline: 40416935]
47. Wang H, Wu X, Li Y, Yu S. Efficacy of robot-assisted training on upper limb motor function after stroke: a systematic review and network meta-analysis. Arch Rehabil Res Clin Transl 2025;7(1):100387 [FREE Full text] [doi: 10.1016/j.arrct.2024.100387] [Medline: 40463769]
48. de Blas-Zamorano P, Montagut-Martínez P, Pérez-Cruzado D, Merchan-Baeza J. Fugl-meyer assessment for upper extremity in stroke: a pyschometric systematic review. J Hand Ther 2025;S0894-1130(25):00057-00052 [FREE Full text] [doi: 10.1016/j.jht.2025.04.004] [Medline: 40533328]
49. Ahmad Puzi A, Sidek SN, Mat Rosly H, Daud N, Md Yusof H. Modified Ashworth Scale (MAS) model based on clinical data measurement towards quantitative evaluation of upper limb spasticity. IOP Conf Ser: Mater Sci Eng 2017;260:012024. [doi: 10.1088/1757-899x/260/1/012024]
50. Wade E, Winstein CJ. Virtual reality and robotics for stroke rehabilitation: where do we go from here? Top Stroke Rehabil 2011;18(6):685-700. [doi: 10.1310/tsr1806-685] [Medline: 22436307]
51. Jung HY. Rehabilitation in subacute and chronic stage after stroke. In: Stroke revisited: Diagnosis and Treatment of Ischemic Stroke. Cham: Springer; 2017:351-360.

XSL•FO
RenderX

52. Qiu X, Rahimzamani A, Wang L, Ren B, Mao Q, Durham T, et al. Inferring causal gene regulatory networks from coupled single-cell expression dynamics using scribe. Cell Syst 2020;10(3):265-274.e11 [FREE Full text] [doi: 10.1016/j.cels.2020.02.003] [Medline: 32135093]
53. Qian Z, Bi Z. Recent development of rehabilitation robots. Advances in Mechanical Engineering 2014;7(2):563062. [doi: 10.1155/2014/563062]
54. Zanatta F, Giardini A, Pierobon A, D'Addario M, Steca P. A systematic review on the usability of robotic and virtual reality devices in neuromotor rehabilitation: patients' and healthcare professionals' perspective. BMC Health Serv Res 2022;22(1):523 [FREE Full text] [doi: 10.1186/s12913-022-07821-w] [Medline: 35443710]
55. Lai TJ, Heggie R, Kamaruzaman HF, Bouttell J, Boyd K. Economic evaluations of robotic-assisted surgery: methods, challenges and opportunities. Appl Health Econ Health Policy 2025;23(1):35-49. [doi: 10.1007/s40258-024-00920-1] [Medline: 39333303]
56. Ali A. Ethical and regulatory considerations in rehabilitation robotics: ensuring safe and equitable care. In: Design and Control of Rehabilitation Robots. Studies in Systems, Decision and Control. Cham: Springer; 2025:399-422.
57. Demofonti A, Carpino G, Zollo L, Johnson MJ. Affordable robotics for upper limb stroke rehabilitation in developing countries: a systematic review. IEEE Trans Med Robot Bionics 2021;3(1):11-20. [doi: 10.1109/tmrb.2021.3054462]
58. Tang Y, Dou B. Cost-effectiveness analysis of robotic surgery in healthcare for older individuals: a systematic review based on randomized controlled trials. Front Public Health 2025;13:1614654 [FREE Full text] [doi: 10.3389/fpubh.2025.1614654] [Medline: 40880918]
59. Chang J, Wu C, Hinton Z, Ryan S, Jiranek W, Bolognesi M, et al. Patient perceptions and interest in robotic-assisted total joint arthroplasty. Arthroplast Today 2024;26:101342 [FREE Full text] [doi: 10.1016/j.artd.2024.101342] [Medline: 38481560]
60. Pagani NR, Moverman MA, Puzzitiello RN, Menendez ME, Barnes CL, Kavolus JJ. Online crowdsourcing to explore public perceptions of robotic-assisted orthopedic surgery. J Arthroplasty 2021;36(6):1887-1894.e3. [doi: 10.1016/j.arth.2021.02.027] [Medline: 33741241]
61. Wu Q, Pei H, Ran X, Chen X, Jiang L, Wei A, et al. Qualitative study on the information needs of patients undergoing da vinci robotic surgery. Clin Nurs Res 2023;32(2):433-440. [doi: 10.1177/10547738221103337] [Medline: 35699515]
62. Muaddi H, Zhao X, Leonardelli GJ, de Mestral C, Nathens A, Stukel TA, et al. Fear of innovation: public's perception of robotic surgery. Surg Endosc 2022;36(8):6076-6083. [doi: 10.1007/s00464-022-09011-5] [Medline: 35059838]
63. Jauniaux B, Anand A, Abbas R, Harji DP. From expectations to experiences: a systematic review of patient and public perspectives on robotic surgery. J Robot Surg 2025;19(1):484. [doi: 10.1007/s11701-025-02649-y] [Medline: 40810856]
64. Cano-de-la-Cuerda R, Blázquez-Fernández A, Marcos-Antón S, Sánchez-Herrera-Baeza P, Fernández-González P, Collado-Vázquez S, et al. Economic cost of rehabilitation with robotic and virtual reality systems in people with neurological disorders: a systematic review. J Clin Med 2024;13(6):1531 [FREE Full text] [doi: 10.3390/jcm13061531] [Medline: 38541755]
65. Ballard M, Montgomery P. Risk of bias in overviews of reviews: a scoping review of methodological guidance and four-item checklist. Res Synth Methods 2017;8(1):92-108. [doi: 10.1002/jrsm.1229] [Medline: 28074553]

Abbreviations

ADL: activity of daily living
CCA: corrected covered area
FMA: Fugl-Meyer Assessment
FMA-UE: Fugl-Meyer Assessment-Upper Extremity
ICF: International Classification of Functioning, Disability and Health
MAS: Modified Ashworth Scale
MD: mean difference
PRIOR: Preferred Reporting Items for Overviews of Reviews
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
RAT: robot-assisted therapy
SMD: standardized mean difference
WMD: weighted mean difference

XSL•FO
RenderX

Edited by S Brini; submitted 20.Jun.2025; peer-reviewed by N Döring, H Wang; comments to author 17.Oct.2025; accepted 02.Feb.2026; published 25.Mar.2026.

Please cite as:
Liu S, Zhang X, Zhou L, Zhang J, Liu J, He C
Robot-Assisted Therapy for Upper Limb Rehabilitation After Stroke: Umbrella Review
J Med Internet Res 2026;28:e79363
URL: https://www.jmir.org/2026/1/e79363
doi:10.2196/79363
PMID:41879816

©Sijia Liu, Xianghong Zhang, Lijun Zhou, Jiashan Zhang, Jialin Liu, Chengqi He. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Technology-Based Interventions for Prevention of Type 2 Diabetes Following Gestational Diabetes: Systematic Review and Meta-Analysis

Claire Eades[1], BSc, MSc, PhD; Anh Nguyen-Hoang[1], BSc, MSc, PhD; Louise Hoyle[1], BNurs, MSc, PhD; Dawn Cameron[2], BSN, MSc, PhD; Josie MM Evans[3], MA, MPH, PhD

[1]Health Sciences, Faculty of Health Sciences and Sport, University of Stirling, Stirling, Stirlingshire, United Kingdom
[2]School of Health Sciences and Life Sciences, University of the West of Scotland, Lanarkshire, South Lanarkshire, United Kingdom
[3]Public Health Scotland, Edinburgh, Lothian, United Kingdom

Corresponding Author:
Claire Eades, BSc, MSc, PhD
Health Sciences
Faculty of Health Sciences and Sport
University of Stirling
Pathfoot Building
Stirling, Stirlingshire, FK9 4LA
United Kingdom
Phone: 44 01786466340
Email: c.e.eades@stir.ac.uk

Abstract

Background: Previous gestational diabetes incurs an 8-fold risk of developing type 2 diabetes, but lifestyle change can prevent or delay progression. Technology-based interventions may help overcome challenges women face in making postpartum lifestyle changes.

Objective: This study aimed to assess whether technology-based diabetes prevention interventions improve outcomes related to the onset of type 2 diabetes among women with a previous diagnosis of gestational diabetes.

Methods: Cochrane Central Register of Controlled Trials, CINAHL, Embase, PsycINFO, and Midwives Information and Resource Service were searched to October 2025 using subject headings and free-text terms. Titles and abstracts were independently screened by 2 authors, as were retrieved full-text articles. Studies were eligible if they examined technology-based diabetes prevention interventions delivered between gestational diabetes diagnosis and any time post partum, assessing anthropometric outcomes, glycemic control, health behavior, or psychological outcomes. Risk of bias was assessed by 1 reviewer using the National Institute for Clinical Excellence checklist, and certainty of evidence was assessed by 2 reviewers using the Grading of Recommendations Assessment, Development, and Evaluation. Data were summarized narratively, and results were pooled, where possible, using a random effects model.

Results: This review identified 15 studies, including 1257 participants. Pooled analysis of 7 studies showed significantly greater weight loss among those receiving technology-based interventions (mean difference –1.01, SE 0.35, 95% CI –1.86 to –0.16 kg; *P*=.03). Interventions delivered using technology only showed increased weight loss (mean difference –1.13, 95% CI –3.12 to 0.86 kg) as did those with a longer follow-up (mean difference –1.58, 95% CI –3.93 to 0.76 kg) compared with combined technology and telemedicine approaches (mean difference –0.89, 95% CI –2.51 to 0.73 kg) and studies with shorter follow-up (mean difference –0.7, 95% CI –1.21 to –0.18 kg), but these differences were not significant (mode of delivery: χ^2_1=0.08; *P*=.78; follow-up: χ^2_1=1.06; *P*=.30). Meta-analysis showed no significant differences in BMI (mean difference –0.22, SE 0.1, 95% CI –0.4 to –0.01 kg/m^2; *P*=.27; n=2 studies), fasting glucose (mean difference –0.03, SE 0.16, 95% CI –0.49 to 0.49 mmol/L; *P*=.99; n=4 studies), 2-hour glucose (mean difference 0.12, SE 0.19, 95% CI –0.47 to 0.72 mmol/L; *P*=.56; n=4 studies), hemoglobin A_{1c}(mean difference –0.01%, SE 0.02%, 95% CI –0.24% to 0.23%; *P*=.74; n=2 studies), or homeostasis model assessment of insulin resistance (mean difference 0.07, SE 0.02, 95% CI –0.16 to 0.31; *P*=.16; n=2). Certainty of evidence for all pooled outcomes was very low.

XSL•FO
RenderX

Conclusions: Technology-based interventions may help support women in reducing their risk of type 2 diabetes following gestational diabetes mellitus, but substantial heterogeneity, significant risk of bias, and very low certainty in the evidence mean that the findings should be interpreted cautiously. Trials with larger samples and longer follow-up are required to draw firm conclusions.

Trial Registration: PROSPERO CRD42024324019; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024324019

(J Med Internet Res 2026;28:e78841) doi:10.2196/78841

KEYWORDS

diabetes prevention; gestational diabetes; health behavior change; mHealth; type 2 diabetes

Introduction

Diabetes is a growing global public health concern and a priority for preventive efforts [1]. The number of adults globally with diabetes in 2017 was estimated to be 451 million and is expected to rise to 693 million by 2045, with type 2 diabetes accounting for over 90% of cases [2]. Women who have had gestational diabetes mellitus (GDM) are at an 8-fold increased risk of developing type 2 diabetes [3]. GDM affects around 5% of pregnant women in Europe and is defined as glucose intolerance that is first diagnosed in pregnancy [4]. Although normal glucose regulation usually returns shortly after delivery, rates of type 2 diabetes after a diagnosis of GDM have been reported to be as high as 70% [5], with the greatest risk thought to be in the first 5 years following a pregnancy with GDM [6]. However, women who make positive lifestyle changes can prevent or delay progression to type 2 diabetes [7].

Challenges that women face in making lifestyle changes in the postpartum period include competing demands from work and home life, and a tendency to underestimate their personal risk and the severity of type 2 diabetes [8-10]. Technology-based interventions offer a possible solution to help to overcome some of these challenges by offering behavior change education and support remotely via messages, smartphone apps, or websites. A recent systematic review of technology-driven diabetes prevention interventions found that around two-thirds of included interventions were effective in achieving shorter-term weight loss and one-third were effective in achieving longer-term weight loss, but the review excluded studies focusing on participants with a previous diagnosis of GDM [11].

Although technology-driven interventions have been shown to be effective in supporting the management of GDM [12], to date, no reviews have been identified that focus on interventions primarily delivered using technology for diabetes prevention in women who have previously had GDM. A systematic review and meta-analysis carried out by Halligan et al [13] investigated the effectiveness of interventions with any technology or telemedicine component for women with previous GDM and found clinically relevant but non–statistically significant improvements in weight and BMI, but currently the effectiveness of primarily technology-based interventions is not known. The review by Halligan et al [13] included interventions with any technology component, and so for many of the included articles, the technology-based intervention was an adjunct to a telemedicine or face-to-face intervention. Telemedicine interventions involve real-time communication with a health care provider via telephone or video call, and although they have some practical benefits over a face-to-face appointment, telemedicine still relies on both parties arranging a convenient time for synchronous contact with a health care provider and so cannot offer the flexible access to advice and help that digital interventions can offer. Interventions delivered primarily using technology may offer the most benefits in terms of flexibility for women in the postpartum period and have the potential to be more cost-effective than telemedicine.

The review by Halligan et al [13] did not assess the behavior change techniques (BCTs) present in the interventions reviewed. BCTs are the active ingredients of an intervention that are designed to promote behavior change [14], and a taxonomy has been developed providing a list of standardized labels and definitions of BCTs. Using the shared language and descriptions provided by the BCT taxonomy allows the content of interventions to be unpicked and described in a consistent and comparable way, which may help to provide insight into the effective components of technology-based interventions. The objective of this review is to assess whether interventions focusing on diabetes prevention (not self-management of GDM), that are primarily technology-based, are effective in improving outcomes related to the onset of type 2 diabetes among women with a previous diagnosis of GDM and to identify and describe the BCTs present in interventions.

Methods

Overview

The systematic review and meta-analysis were conducted and are reported according to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses; Multimedia Appendix 1) guidelines [15], PRISMA abstract checklist (Multimedia Appendix 2) and the complementary PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension; Multimedia Appendix 3) for reporting literature searches within systematic reviews [16]. The protocol for the systematic review was registered with the International Prospective Register of Systematic Reviews, and no amendments were made to the protocol [17].

Eligibility Criteria

The inclusion and exclusion criteria used to assess eligibility in the review are presented in Textbox 1.

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Adults aged 18 years or older who have received a previous diagnosis of gestational diabetes.
- Studies that examine the use of a technology-based intervention (eg, text, smartphone app, email, video, and website) in the antenatal or postpartum period with the stated aim of preventing or reducing the risk of type 2 diabetes.
- Interventions using both technology-based and face-to-face delivery, where the technology-based delivery is the primary means of intervention delivery and the technology-based intervention could feasibly be delivered as a stand-alone intervention.
- The outcomes of interest are anthropometric (weight, BMI, percent body fat, and waist or hip circumference), glycemic control (hemoglobin A_{1c}, fasting plasma glucose, insulin resistance, prediabetes incidence/prevalence, and type 2 diabetes incidence), behavioral outcomes (diet, physical activity, or sedentary behavior), and psychological outcomes (beliefs about type 2 diabetes, self-efficacy for health behavior change, and behavioral intentions).
- Experimental and nonexperimental study designs, including observational studies, randomized trials, and nonrandomized trials.
- Studies published in English.

Exclusion criteria

- Studies of women who have received a diagnosis of type 1 or type 2 diabetes.
- Studies describing interventions where the technology-based component was only used to supplement a face-to-face intervention or to deliver real-time support with a human coach.
- Studies focusing on the self-management of gestational diabetes.
- Studies reporting only feasibility and acceptability related outcomes.
- Studies using a qualitative design only.

Information Sources and Search Strategy

The databases MEDLINE, Central Register of Controlled Trials, CINAHL, Embase, PsycINFO, and Midwives Information and Resource Service were searched separately via the EBSCO platform in March 2024 using a combination of database subject headings and free-text words (Multimedia Appendix 4). No dedicated peer review of the search strategy was conducted. The search was rerun in October 2025 to check for papers published since the initial search in March 2024. No restrictions were applied for language, type, or date of publication, and no published search filters were used. No online resources, print sources, or citing references of included papers were searched. The reference lists of included papers were checked to identify any other potentially relevant papers, but experts in the field were not contacted due to the time-consuming nature of this process. Duplicates were removed in RefWorks (ProQuest LLC).

Selection Process

The titles and abstracts of all articles were screened by 1 author (CE), and independent screening was split between 3 other authors, with DC screening one-quarter, LH screening one-quarter, and AN-H screening the other half. The full texts of papers were retrieved for studies that were considered relevant, but also for those that contained insufficient information to allow judgment of relevance. These were checked against the inclusion criteria by CE and independently by AN-H, LH, and DC. Where there were disagreements between reviewers about the inclusion of a paper, a consensus was reached through discussion among all authors.

Data Collection Process and Data Items

A data extraction form was developed for this review and piloted on 3 papers, and then refined. Information was extracted by 1 reviewer (CE) under the following headings: authors, study location, type of intervention, mode of delivery, theory or model base for intervention, aim of the study, study design, inclusion criteria, size of the population, baseline characteristics, type of comparator, tailoring/individual patient–based, duration of intervention and measurement, outcome measures, measurement tools, sample size intervention group, sample size control group, dropout, findings summary, baseline outcomes intervention, baseline outcomes control, follow-up outcomes intervention, and follow-up outcomes control. Data were extracted for all available follow-up time points and outcomes available within each study. We attempted to contact authors where data were missing or not reported in the correct format for meta-analysis. If the data were not available or we received no response, studies were excluded from the meta-analysis. A second author (AN-H) independently checked all outcome data, and any errors were discussed and checked against the published paper by both CE and AN-H. The BCT taxonomy (version 1) [14] was used by 1 reviewer (AN-H) to code the BCTs present in the descriptions of included interventions. A second reviewer (CE) independently coded a random selection of 20% of interventions, and any disagreements were resolved by discussion.

Risk of Bias Assessment

Internal and external validity of included studies was assessed by 1 reviewer (CE) using the National Institute for Clinical Excellence [18] checklist for quantitative intervention studies of any design. This checklist contains 27 items addressing 4 specific domains (population, method of allocation to intervention, outcomes, and analyses) and gives an appraisal of internal and external validity using a “++,” “+,” or “–“ rating scale depending on the extent to which criteria were satisfied,

XSL•FO
RenderX

with a "++" indicating lowest risk of bias and "–" highest risk of bias.

Data Analysis

All data were grouped by outcome for synthesis and summarized narratively. Meta-analysis was planned for any outcome of interest recorded at both baseline and follow-up by 2 or more studies. Consistent with the Cochrane Collaboration guidance, only studies with a randomized controlled trial design were included in the meta-analysis [19]. After extraction, the following outcomes were synthesized using meta-analysis: weight, BMI, fasting plasma glucose, 2-hour glucose, hemoglobin A_{1c} (HbA_{1c}), and homeostasis model assessment of insulin resistance (HOMA-IR). Results were pooled in SPSS (version 28; IBM Corp) using a random effects model to analyze the between-group differences in each outcome at follow-up. Forest plots were used to display individual study results and the syntheses. When there were follow-up data available at multiple time points, data from the last follow-up time point were used in the analysis. Where there were multiple publications reporting on the same sample, the data were only included in the analyses once to avoid double-counting of participants. When SD values were not reported, these were calculated from 95% CIs when possible. All pooled outcomes were continuous and were summarized using unstandardized mean difference with Sidik-Jonkman estimator and Knapp-Hartung SE adjustment as recommended for meta-analyses of a small number of studies [20]. Two-sided *P* values and 95% CIs were calculated for each outcome. Heterogeneity was quantified using between-study variance τ^2. The I^2 statistic is reported but interpreted with caution [21]. Prediction intervals were not calculated as the number of studies in the analyses fell below the recommended minimum number [22]. Analysis of publication bias was not conducted due to the small number of studies eligible for inclusion. Subgroup analyses were planned for the mode of intervention delivery (entirely digital or mixed) and length of follow-up (less than 6 months versus 6 months or more).

Certainty Assessment

Certainty of evidence was evaluated using the Grading of Recommendations Assessment, Development, and Evaluation (GRADE; [23]) criteria for the main outcomes (weight, fasting glucose, 2-hour glucose, HbA_{1c}, or HOMA-IR). GRADE classifies evidence into 4 levels: high, moderate, low, and very low based on assessment of risk of bias, inconsistency, indirectness, imprecision, and publication bias. With evidence rated as high quality, we can have very high certainty that the true effect is close to the estimated effect, whereas in evidence rated as very low, we can have almost no certainty in the estimates reported. Two reviewers (CE and LH) independently rated studies included in the meta-analysis using GRADEpro, and any disagreements were resolved by discussion. The table summarizing the results of the assessment can be found in Multimedia Appendix 5.

Results

Study Selection

Figure 1 shows a PRISMA flow diagram of studies identified by the search. A total of 4704 papers were returned by the search, of which 4582 were excluded at the title and abstract screening stage. We reviewed the full text of 122 papers, and 15 studies [24-38] reported in 17 papers [39,40] were deemed eligible for inclusion in the systematic review. At the full-text screening stage, the following exclusions were made: 26 nondigital interventions, 23 protocols, 16 interventions where the technology-based component was a supplement to face-to-face, 12 interventions for the management or prevention of gestational diabetes, 11 papers that did not report relevant outcomes, 5 duplicates, 4 review papers, 2 papers not published in English, 2 papers that included women with additional cardiometabolic risk factors in the sample, and 1 conference abstract. A total of 4 conference abstracts were reviewed that did not contain enough information to assess eligibility. We searched the 4704 abstracts identified by our review by author name and intervention name, as outlined in the conference publications, to identify any further publications relating to these interventions. For 3 of the 4 conference abstracts, further publications with more information were identified that allowed assessment of eligibility. There were no related papers found for the fourth conference abstract, and no email addresses could be found for the authors through web searches, so the study was excluded from the review [41].

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

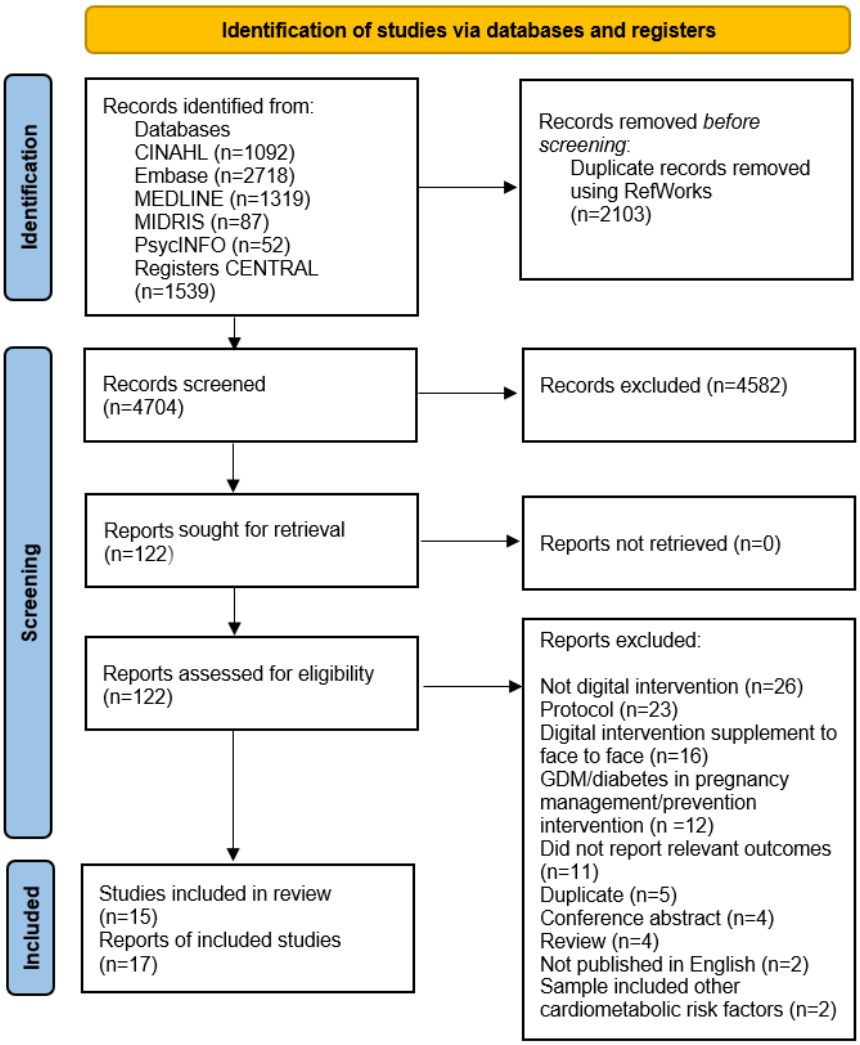

Study Characteristics

The characteristics of the 15 included studies are presented in Table 1. Studies were published between 2012 and 2024, with almost two-thirds published between 2020 and 2024 (n=10, 66%). They were conducted in the United States (n=6) [28,31,32,34,36,37], Australia (n=4) [24,25,35,38], Asia (n=4) [26,28-30], and Europe (n=1) [33]. The mean age of participants among studies that reported age (n=13) [24-26,29-38] ranged from 31.5 to 40.4 years, with most means falling between 31 and 34 years (n=9, 69.2%). Study design was primarily randomized controlled trial (n=11) [24,25,27,29,30,32-36,38], with 2 studies adopting a before-and-after single-arm design [31,37] and 2 a nonrandomized controlled trial design [26,28].

XSL•FO
RenderX

Table 1. Summary of the characteristics of included studies.

Author (year), location	Enrolment setting	Design	Definition GDM[a] used	IG[b]/CG[c] sample size, n	Attrition IG/CG, %	Outcomes assessed	Follow-up duration	Risk of Bias	Age in IG (years), mean (SD)	Age in CG (years), mean (SD)	BMI in IG (years), mean (SD)	BMI in CG (years), mean (SD)
Cheung (2019), Sydney, Australia [24]	1 tertiary care hospital	RCT[d]	FPG[e] ≥ 5.5 mmol/L and/or a 2hPG[f] ≥ 8 mmol/L on 75 g OGTT[g]	40/20	28/35	Weight, PA[h], diet	8.5 months	++/+	34 (4)	34 (4)	NR[i]	NR
Cheung (2024), Sydney, Australia [25]	3 tertiary care hospitals	RCT	FPG ≥ 5.5 mmol/L and/or a 2hPG ≥ 8 mmol/L on 75 g OGTT (at 2 hospitals) or one of the following: FPG ≥ 7.0 mmol/L; 2hPG ≥ 11.1 mmol/L on 75 g OGTT; random plasma glucose ≥ 11.1 mmol/L (at 1 hospital)	88/89	7/12	Weight, PA, diet	6 months	++/+	32.2 (4.6)	32.2 (4.6)	Prepregnancy: 29.4 (6.9)	Prepregnancy: 28.4 (5.9)
Ghaderi (2019), Tehran, Iran [26]	1 treatment center	Quasi-experimental with control group	NR	45/45	2/4	Risk perceptions for T2D[j]	NR	++/+	33.66 (3.93)	32.55 (4.44)	Prepregnancy: 26.44 (1.63)	Prepregnancy: 27.09 (2.13)
Kim (2012), Michigan, United States [27]	University health system (large nonprofit managed care plan with several private practices)	RCT	NR	28/21	7/23	Weight, WC[k], FPG, 2hPG, PA, self-efficacy for diet and PA, risk perceptions type 2 diabetes	3 months	++/++	NR	NR	29.8 (6.8)	29.8 (6.8)

XSL•FO
RenderX

Author (year), location	Enrolment setting	Design	Definition GDM[a] used	IG[b]/CG[c] sample size, n	Attrition IG/CG, %	Outcomes assessed	Follow-up duration	Risk of Bias	Age in IG (years), mean (SD)	Age in CG (years), mean (SD)	BMI in IG (years), mean (SD)	BMI in CG (years), mean (SD)
Kim (2021), South Korea [28]	3 hospitals	Quasi-experimental with control group	NR	64/64	8/3	Weight, body fat, FPG, HbA_{1c}[l], diet, health-promoting lifestyle profile, T2D knowledge	3 months	+/+	Over 35 45.6%	Over 35 50%	NR	NR
Liew (2023), Singapore [29]	NR	RCT	NR	31/30	10/6	Weight, BMI, FPG, 2hPG	2.5 months	++/+	36.5 (3.7)	35.4 (3.7)	24.0 (4.4)	22.7 (4.0)
Lim (2021), Singapore [30]	1 hospital	RCT	FPG ≥ 5.1 mmol/L and/or 1hPG[m] ≥ 10.0 and/or 2hPG ≥ 8.5 mmol/L on 75 g OGTT	101/99	5/6	Weight, WC, diet, FPG, 2hPG, health directed behavior	4 months	++/+	32.6 (4.5)	32.4 (4.2)	Median 26.9 (IQR 23.0–29.5)	Median 25.8 (IQR 22.8–28.6)
Nicholson (2016), United States [31]	2 university-based prenatal clinics and 1 community-based practice	Single-arm pre-post study	NR	16/none	30	Weight, HbA_{1c}, self-efficacy for PA	NR	+/++	31.5 (4.7)	NR	Prepregnancy: 29.4 (11.2)	NR
Nicklas (2014 [25], 2019 [39], 2020 [40]), Boston, United States	1 hospital	RCT	Meets ≥ 2 following 100 g OGTT: FPG ≥ 5.3 mmol/L, 1hPG ≥ 10.0 mmol/L, 2hPG ≥ 8.6 mmol/L, 3hPG[n] ≥ 7.8 mmol/L. Or medical record documented clinician diagnosis	36/39	17/13	Weight, BMI, WC, FPG, 2hPG, HbA_{1c}, HOMA-IR[o], IGT[p] and T2D prevalence, diet, PA	12 months	++/+	33.6 (4.8)	33.3 (5.8)	31.2 (5.8)	31.6 (5.5)

XSL•FO
RenderX

Author (year), location	Enrolment setting	Design	Definition GDM[a] used	IG[b]/CG[c] sample size, n	Attrition IG/CG, %	Outcomes assessed	Follow-up duration	Risk of Bias	Age in IG (years), mean (SD)	Age in CG (years), mean (SD)	BMI in IG (years), mean (SD)	BMI in CG (years), mean (SD)
Potzel (2022), Germany [33]	University medical centre in Munich, Germany; Diabetes center in Dusseldorf; Institute for diabetes research in Tubingen	RCT	FPG ≥ 5.1 mmol/L and/or a 2hPG ≥ 8.5 mmol/L on 75 g OGTT	33/33	18/15	Weight, body fat, IGT prevalence, PA, diet, change in health habits	6 months	++/+	37.0 (4.4)	35.7 (3.1)	Median 26 (IQR 22.6–28.9)	Median 27.3 (IQR 23.9–30.3)
Reutrakul (2022), Chicago, United States [34]	NR	RCT	NR	9/6	0/17	Sleep duration, PA, FPG, 2hPG, area under the curve for glucose, HOMA-IR	1.5 months	+/++	42.0 (2.9)	38.7 (6.0)	32.7 (5.3)	32.6 (4.6)
Rollo (2020), New South Wales, Australia [35]	NR	3-arm RCT	NR	28 (15 low personalization [HP] and 13 low personalization [LP])/14	29/21	Weight, BMI, WC, body fat, HbA_{1c}, diet quality, PA	6 months	++/++	HP[q]: 34 (4.5); LP[r]: 32.8 (3.6)	33.6 (3.8)	HP: 33.9 (3.6), LP: 33.9 (3.6)	31.1 (4.8)
Saxon (2023), Boston and Colorado, United States [36]	NR	RCT	Meets ≥ 2 following 100 g OGTT: FPG ≥ 5.3 mmol/L, 1hPG ≥ 10.0 mmol/L, 2hPG ≥ 8.6 mmol/L, 3hPG ≥ 7.8 mmol/L	91/90	48/49	Diet and PA self-efficacy, readiness to change	24 months	-/-	32.5 (4.9)	32.2 (5.5)	31.4 (5.9)	31.3 (5.8)
Seely (2020), Boston, United States [37]	Single federally qualified community health center	Single-arm pre-post study	NR	21/none	14	Weight, self-efficacy for PA and healthy eating	1.5 months	-/++	Whole sample: 33 (6.9)	NR	NR	NR

XSL•FO
RenderX

Author (year), location	Enrolment setting	Design	Definition GDM[a] used	IG[b]/CG[c] sample size, n	Attrition IG/CG, %	Outcomes assessed	Follow-up duration	Risk of Bias	Age in IG (years), mean (SD)	Age in CG (years), mean (SD)	BMI in IG (years), mean (SD)	BMI in CG (years), mean (SD)
Taylor (2022), New South Wales, Australia [38]	NR	3-arm RCT	NR	48 (HP 25, medium personalization [MP] 23)/28	16/4	Weight, diet quality, PA, self-efficacy weight management	3 months	+/++	HP: 36.0 (5.4); MP[s]: 34.6 (5.3)	37.3 (5.3)	30.7 (4.3)	30.0 (4.8)

[a]GDM: gestational diabetes mellitus.
[b]IG: intervention group.
[c]CG: control group.
[d]RCT: randomized controlled trial.
[e]FPG: fasting plasma glucose.
[f]2hPG: 2-hour plasma glucose.
[g]OGTT: oral glucose tolerance test.
[h]PA: physical activity.
[i]NR: not reported.
[j]T2D: type 2 diabetes.
[k]WC: waist circumference.
[l]HbA1c: hemoglobin A_{1c}.
[m]1hPG: 1-hour plasma glucose.
[n]3hPG: 3-hour plasma glucose.
[o]HOMA-IR: homeostatic model assessment of insulin resistance.
[p]IGT: impaired glucose tolerance.
[q]HP: high personalization.
[r]LP: low personalization.
[s]MP: medium personalization.

Intervention Characteristics

Table 2 summarizes the characteristics of the interventions reported in the 15 included studies. Among the included studies, 2 [36,38] were full trials of interventions piloted in other studies included in the review [32,35]. Most interventions targeted diet and physical activity (n=12) [24,25,27-33,35,37,38]. Among the studies that did not target diet and physical activity, 1 targeted risk perceptions for type 2 diabetes [26], 1 addressed sleep [34], and 1 targeted self-efficacy for diet and physical activity [36]. Of the 12 studies targeting diet and physical activity behaviors, 7 also targeted additional behaviors or factors in the intervention, including breastfeeding [25,32], risk perceptions [27,31], self-efficacy [27], parenting [25], infant health/first aid [25,28], stress [28,29], sleep [29,33], and emotions [33]. A total of 4 studies stated that they were informed by 1 or more theories [25,31,33,37], with social cognitive theory the most frequently used (n=3). Among trials with a control arm (n=11), control participants were provided with information about gestational diabetes or diabetes prevention in 5 studies [24,28,32,33,36], received routine care in 4 studies [25-27,30], received the same app/website/devices as the intervention group but without messaging or coaching 2 studies [29,38], were on a waitlist for the intervention in 1 study [35], and received weekly emails with general health education and weekly brief telephone contact from a coach in 1 study [34].

XSL•FO
RenderX

Table 2. Summary of intervention characteristics.

Author (year)	Intervention name; description	Intervention duration	Time period intervention started	Intervention target	Mode of delivery	Control	Theoretical basis	Any tailoring
Cheung (2019) [24]	Smart Mums with Smart Phones; Lifestyle advice, education and support delivered via text messages and counseling, including feedback from activity monitor	8.5 months	Birth of baby	Diet, PA[a]	Text messaging (3 per week), provision of activity monitor, 2 × 30 min diet counseling sessions (1 remote and 1 face-to-face)	Leaflet	NR[b]	Text messages customized according to activity monitor data, breastfeeding status and ethnicity
Cheung (2024) [25]	Smart Mums with Smart Phones 2; Healthy lifestyle promotion delivered via text messages including feedback from activity monitor	6 months	Birth of baby	Diet, PA, breastfeeding, infant health, parenting	Text messaging (4 per week), activity monitor	Routine care	TRA[c], SCT[d], HBM[e]	Text messages customized to breastfeeding status and activity monitor data
Ghaderi (2019) [26]	NR; Education on risk of type 2 diabetes	NR	NR	Risk perceptions	Smartphone app	Routine care	NR	NR
Kim (2012) [27]	NR; Structured, web-based pedometer program giving personalized goals and education regarding lifestyle modification	3 months	Within 3 years of GDM diagnosis	Diet, PA, risk perceptions, self-efficacy	Website, text messaging, internet forum	Routine care	NR	Customized step count goals
Kim (2021) [28]	NR; Virtual reality program where various exercises could be viewed and performed and diet recorded	3 months	Birth of baby	Diet, PA, stress, infant first aid	Smartphone, virtual reality application and headset	Written educational materials on GDM[f]	NR	Adequacy of diet and calories were presented based upon BMI
Liew (2023) [29]	NR; Self-monitoring intervention using health tracker. Data from tracker reviewed by health coaches and personalized recommendations delivered by coaches	1 month	Within 10 years of GDM diagnosis	Diet, PA, sleep, stress	Smartphone app, continuous glucose monitor, health tracking ring, 2 sessions of chatroom messaging with health coach	Same app & devices as IG[g] but no messaging	NR	Customized recommendations for diet, PA, sleep and stress
Lim (2021) [30]	SPAROW; App offering short videos on lifestyle, diet and activity tracking, personalized goals and live chat function	4 months	Birth of baby	Diet, PA	Smartphone app, live chat with health care professionals (5 to 10 min a day for first week then 2 to 3 times a week)	Routine care	NR	Customized weight and activity goals
Nicholson (2016) [31]	GooDMomS; Web-based lessons on healthy lifestyle, weight tracking and feedback on weight	7.5 months	From diagnosis of GDM	Diet, PA, risk perceptions	Website, 4 live webinars, message board, text messages/emails	NR	SCT	Customized feedback about weight, PA and diet

XSL•FO
RenderX

Author (year)	Intervention name; description	Intervention duration	Time period intervention started	Intervention target	Mode of delivery	Control	Theoretical basis	Any tailoring
Nicklas (2014) [32]; (2019) [39]; (2020) [40]	Balance after Baby; Lifestyle modification adapted from DPP[h] delivered via website and coach. Goals, weight and physical activity recorded on website	12 months	6 weeks post partum	Diet, PA, breastfeeding	Website, email and weekly telephone coaching for 12 weeks then every second week for 12 weeks and monthly thereafter	Leaflet	NR	NR
Potzel (2022) [33]	TRIANGLE; Self-paced DPP-based intervention with individualized challenges, chat coaching and behavior change technique library	6 months	3–18 months following GDM diagnosis	Diet, PA, sleep, emotions	Smartphone app, chat-based coaching	Leaflet	Kok's taxonomy of behavior change methods	Customized and interactive challenges
Retrakul (2022) [34]	Sleep-Extend; Sleep tracker and app, didactic content and telephone coaching aimed at improving sleep	1.5 months	NR	Sleep	Smartphone app, wearable sleep tracker, email, weekly telephone coaching (5 to 20 min each)	Weekly health education emails (eg, vaccination for adults, cancer screening) & weekly brief telephone contact from the coach	NR	During coaching feedback given on sleep data, progress, and barriers explored and goals set
Rollo (2020) [35]	Body Balance Beyond; Website-based content on health lifestyles, goal setting, feedback on diet and activity from coaches, and text messaging support	6 months	Within 2 years of GDM diagnosis	Diet, PA	Website, text messaging, 6 coaching video calls (20–30 min each) with dietician and exercise physiologist	Waitlist control	NR	Coaching sessions reviewed diet and PA, explored barriers to diet and PA, refined goals and personalized strategies for achieving goals
Saxon (2020) [36]	Balance after Baby; See Nicklas 2014 description	12 months	6 weeks post partum	Self-efficacy and readiness to change for PA and diet	Website and weekly telephone, email and text messaging coaching for 12 weeks then every second week for 12 weeks and monthly thereafter	Access to websites with information about diabetes prevention	NR	NR
Seely (2020) [37]	Hola Bebé, Adiós Diabetes; Audio visual modules on healthy eating and physical activity delivered via app	2 months	Within 5 years of GDM diagnosis	Diet, PA	Smartphone app	NR	SCT	NR

XSL•FO
RenderX

Author (year)	Intervention name; description	Intervention duration	Time period intervention started	Intervention target	Mode of delivery	Control	Theoretical basis	Any tailoring
Taylor (2022) [38]	Body Balance Beyond; See Rollo 2020 description	3 months	Within 5 years of GDM diagnosis	Diet, PA	Website, text messaging, 5 coaching video calls (20–30 min each) with a dietician or exercise physiologist	Access to intervention website only	NR	Coaching sessions reviewed diet and PA, explored barriers to diet and PA, set goals and strategies for achieving goals

[a]PA: physical activity.

[b]NR: not reported.

[c]TRA: theory of reasoned action.

[d]SCT: social cognitive theory.

[e]HBM: health belief model.

[f]GDM: gestational diabetes mellitus.

[g]IG: intervention group.

[h]DPP: diabetes prevention program.

The duration of interventions ranged from 1 to 12 months. In most studies, the intervention was carried out in the time period between diagnosis of GDM and 18 months post partum (8/13, 60% of studies reporting when the intervention was delivered) [24,25,28,30-33,36]. A total of 4 studies offered the intervention to women between a maximum of 2 and 5 years following diagnosis of GDM [27,35,37,38], and 1 study offered the intervention up to 10 years following diagnosis [29]. The technology-based modes of delivery used included smartphone apps [26,28-30,33,34,37], SMS text messaging [24,25,27,31,35,36,38], emails [31,32,34,36], websites [27,31,32,35,36,38], online message boards [27,31], virtual reality headset [28], activity/sleep trackers [24,25,34], webinars [31], and video calls [35,38], with most studies used multiple technology-based methods of delivery (n=11). A total of 7 studies used only technology-based modes of intervention delivery [25-30] and 8 used technology-based delivery plus synchronous contact with a health coach/interventionist through live chats [33], telephone calls [32,34,36], video calls [35,38], a combination of face-to-face and remote methods [24], or via webinars [31].

Risk of Bias and Certainty of Evidence

A summary of the quality assessments can be found in Table 1. A total of 9 studies achieved a high-quality rating for internal validity, indicating that they minimized bias across multiple criteria [24-27,29,30,32,33,35]. Additionally, 6 studies received a high-quality rating for external validity, indicating that they minimized risk of bias in the generalizability of the study population across multiple criteria [27,31,34,35,37,38]. The GRADE assessment indicated that the certainty of evidence was very low for all pooled outcomes (weight, fasting plasma glucose, 2-hour plasma glucose, HbA_{1c}, and HOMA-IR; Table 3).

XSL•FO
RenderX

Table 3. Grading of Recommendations Assessment, Development, and Evaluation assessment.

Outcome and certainty assessment							Patients, n		Effect		Certainty	Importance
Studies, n	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Technology-based interventions	Nontechnology-based interventions	Relative (95% CI)	Absolute (95% CI)		
Weight (kg; follow-up: range 10-52 weeks)												
7	Randomized trials	Serious[a,b]	Not serious	Serious[c]	Serious[d]	None	181	159	NR[e]	MD[f] 1.01 SD lower (1.86-0.16 lower)	very low[a,b,c,d]	Critical
BMI (kg/m^2; follow-up: range 10-13 weeks)												
2	Randomized trials	Serious[a,b]	Not serious	Serious[c,g]	Serious[d]	None	47	51	NR	MD 0.22 SD lower (1.48-1.03 higher)	very low[a,b,c,d,g]	Critical
Fasting plasma glucose (follow-up: range 10-52 weeks)												
4	Randomized trials	Serious[a,b]	Not serious	Serious[c]	Serious[d]	None	96	91	NR	MD 0 SD (0.5-0.49 higher)	very low[a,b,c,d]	Critical
2-hour plasma glucose (follow-up: range 10-52 weeks)												
4	Randomized trials	Serious[a,b]	Not serious	Serious[c]	Serious[d]	None	96	91	NR	MD 0.12 SD higher (0.47-0.72 higher)	very low[a,b,c,d]	Critical
Hemoglobin A_{1c} (follow-up: range 26-52 weeks)												
2	Randomized trials	Serious[a,b]	Not serious	Serious[c]	Serious[d]	None	54	53	NR	MD 0.01 SD lower (0.24-0.23 higher)	very low[a,b,c,d]	Critical
HOMA-IR[h] (follow-up: range 6-52 weeks)												
2	Randomized trials	Serious[a,b]	Not serious	Serious[c]	Serious[d]	None	48	44	NR	MD 0.07 SD higher (0.16-0.31 higher)	very low[a,b,c,d]	Critical

[a]The nature of the technological interventions in this review (websites and smartphone apps) makes blinding difficult to achieve. A few studies managed

XSL•FO
RenderX

to blind researchers and outcome assessors, but most studies did not blind participants or researchers or did not report in detail.

[b]Some studies did not adequately describe allocation concealment.

[c]Most studies are small, single region, pilot studies using surrogate/short-term outcomes which limits the directness of the evidence and applicability to the broader population of women with previous gestational diabetes.

[d]All studies have very small sample sizes of <400 and wide CIs.

[e]NR: not reported.

[f]MD: mean difference.

[e]Variation in interventions is present.

[h]HOMA-IR: homeostatic model assessment of insulin resistance.

Anthropometric Outcomes

Weight was reported as an outcome in 12 studies [24,25,27-33,35,37,38], waist circumference in 4 [27,30,32,35], body composition in 3 [28,33,35], and BMI in 2 studies [27,29]. A total of 5 studies could not be included in the meta-analysis for weight: 2 were single-arm studies [31,37], 1 reported the proportion of women who met weight goals [25], 1 did not report SDs, and they could not be calculated [30], and 1 was a nonrandomized controlled trial [28]. Of the 5 studies that could not be pooled, only 1 [28] reported findings that favored the intervention in relation to weight loss.

Pooled analysis showed significantly greater weight loss in the intervention group compared with the control (mean difference –1.01, SE 0.35, 95% CI –1.86 to –0.16 kg; *P*=.03; n=7; GRADE very low; Figure 2 [24,27,29,32,33,35,38]). τ^2=0.59; τ=0.77; and I^2=69% indicated higher heterogeneity. The effect did not remain significant across different modes of delivery and lengths of follow-up in subgroup analyses, although there was a larger reduction in weight noted among studies using technology-based delivery only (mean difference –1.13, 95% CI –3.12 to 0.86 kg for technology-based only, mean difference –0.89, 95% CI –2.51 to 0.73 kg for mixed delivery; Figure 2) and with a longer follow-up duration (mean difference –0.70, 95% CI –1.21 to –0.18 kg for follow-up <6 months; mean difference –1.58, 95% CI –3.93 to 0.76 kg for 6-12 months; Figure 3 [24,27,29,32,33,35,38]) but between-group differences were not significant for mode of delivery (χ^2_1=0.08; *P*=.78) or follow-up duration (χ^2_1=1.06; *P*=.30).

Figure 2. Random effects meta-analysis of the mean difference in weight (kg) comparing the intervention group and the control group according to the mode of delivery (digital only vs mixed) [24,27,29,32,33,35,38].

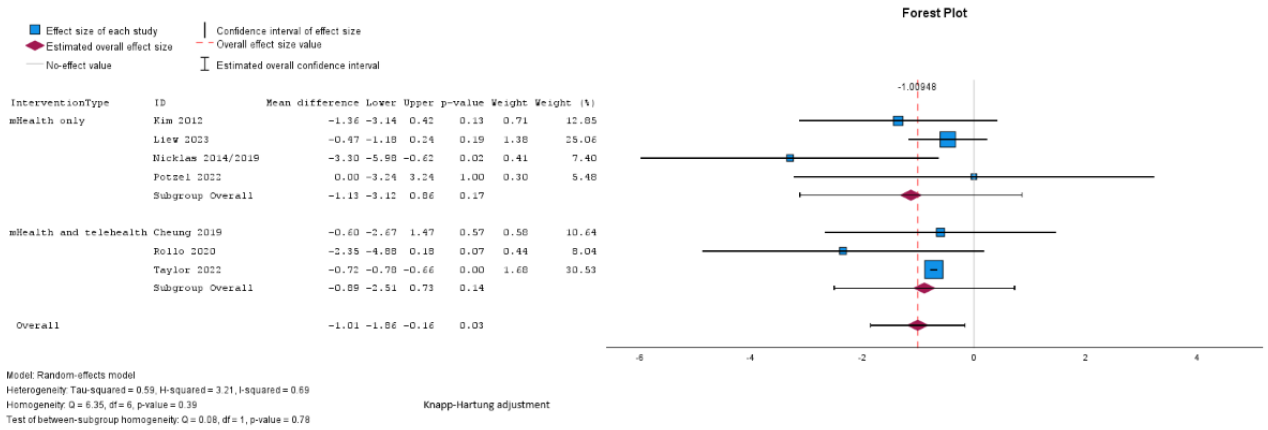

Figure 3. Random effects meta-analysis of the mean difference in weight (kg) comparing the intervention group and the control group according to the length of follow-up [24,27,29,32,33,35,38].

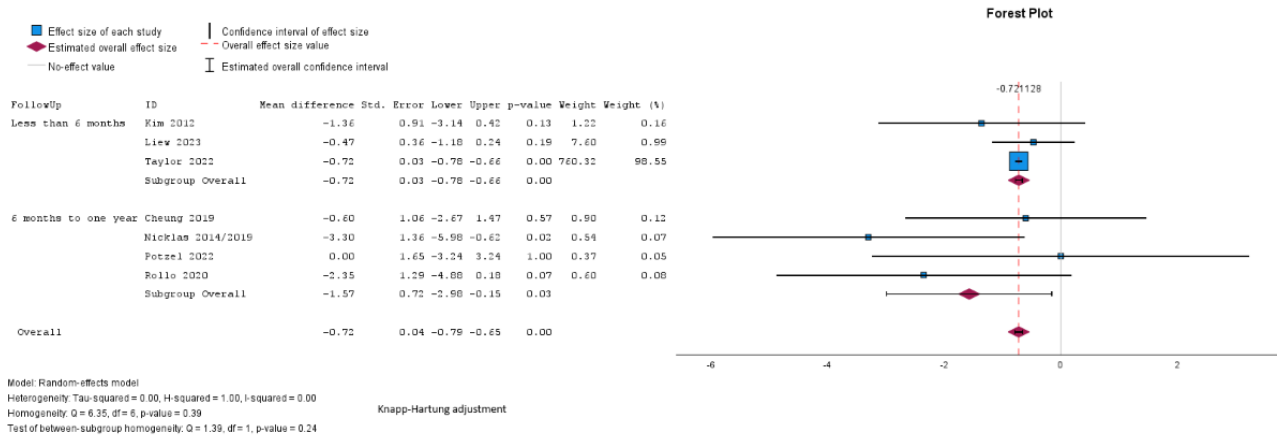

XSL•FO
RenderX

Pooled analysis showed a reduction in BMI in the intervention group compared to the control group, but the difference was not statistically significant (mean difference –0.22, SE 0.1, 95% CI –0.4 to –0.01 kg/m^2; P=.27; n=2 [27,29]; GRADE very low; Figure 4 [27,29]). The I^2=41%; τ=0.1; and τ^2=0.01 indicated lower heterogeneity. Subgroup analyses were not possible for BMI because of the small number of studies assessing these outcomes and the uneven distribution of studies across the different subgroups.

Figure 4. Random effects meta-analysis of the mean difference in BMI comparing the intervention group and the control group [27,29].

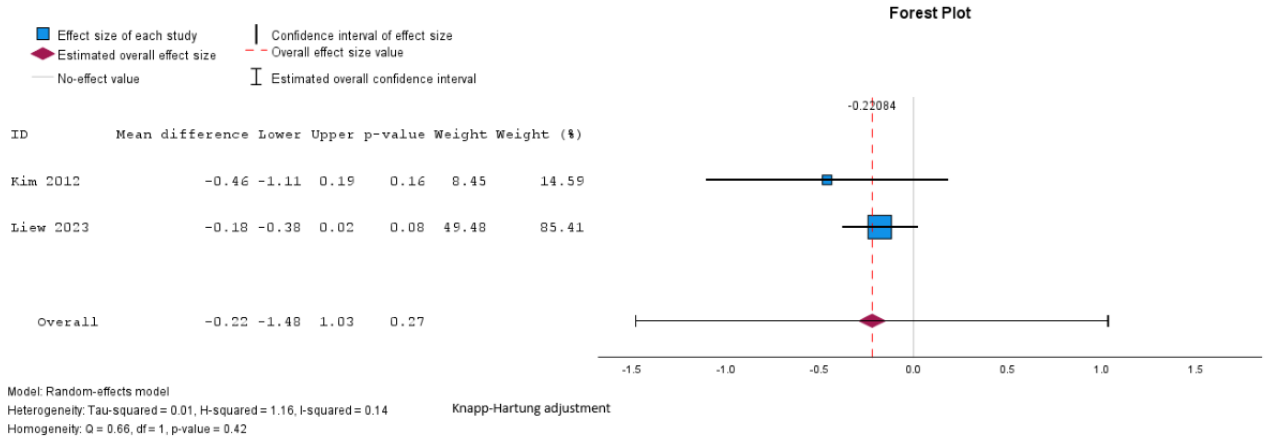

None of the 4 studies assessing waist circumference reported significant differences between the intervention and control groups [27,30,32,35]. Of the 3 studies measuring body composition [28,33,35], only 1 found a significant reduction in body fat percentage in the intervention group compared to the control group [28]. There were too few studies reporting data on waist circumference and body composition at both baseline and follow-up to allow for meta-analysis of these outcomes.

Glycemic Outcomes

Fasting glucose levels were reported in 5 studies [27-29,32,34], 2-hour glucose following an oral glucose tolerance test in 5 [27,29,30,34,39], HbA_{1c} in 4 [28,31,35,39], and HOMA-IR in 2 studies [34,39]. Only 1 study found a significant difference in glycemic outcomes, reporting lower fasting glucose and HbA_{1c} in the intervention group compared to the control group [28]. One study assessed the prevalence of impaired glucose tolerance [33] and another reported both impaired glucose tolerance and type 2 diabetes prevalence [32], with neither finding significant differences in prevalence among women receiving the intervention compared to the control group. Three studies reporting glycemic outcomes could not be pooled: SDs were not reported and could not be calculated for fasting and 2-hour glucose in 1 study [30], 1 study assessing fasting glucose and HbA_{1c} was a nonrandomized controlled trial [28], and another reporting HbA_{1c} was a single-arm design [31]. Subgroup analyses were not possible for glycemic outcomes because of the small number of studies assessing these outcomes and the uneven distribution of studies across the different subgroups of mode of delivery and length of follow-up.

Pooled analysis showed no significant difference between groups in fasting glucose (mean difference –0.03, SE 0.16, 95% CI –0.49 to 0.49 mmol/L; P=.99; n=4; GRADE very low; Figure 5 [27,29,32,34]), 2-hour glucose (mean difference 0.12, SE 0.19, 95% CI –0.47 to 0.72 mmol/L; P=.56; n=4; GRADE very low; Figure 6 [27,29,32,34]), HbA_{1c} (mean difference –0.01%, SE 0.02%, 95% CI –0.24% to 0.23%; P=.74; n=2 studies; GRADE very low; Figure 7 [32,35]), or HOMA-IR (mean difference 0.07, SE 0.02, 95% CI –0.16 to 0.31; P=.16; n=2; GRADE very low; Figure 8 [32,34]). The τ=0.24; τ^2=0.06; and I^2=55% indicate higher heterogeneity among the studies assessing fasting plasma glucose, and 2-hour glucose (I^2=42%; τ=0.58; τ^2=0.34), and lower heterogeneity in the studies assessing HbA_{1c} (I^2=31%; τ=0.02; τ^2=0.0004) and HOMA-IR (I^2=0%; τ=0.001; τ^2=0.000002).

XSL•FO
RenderX

Figure 5. Random effects meta-analysis of the mean difference in fasting glucose comparing the intervention group and the control group [27,29,32,34].

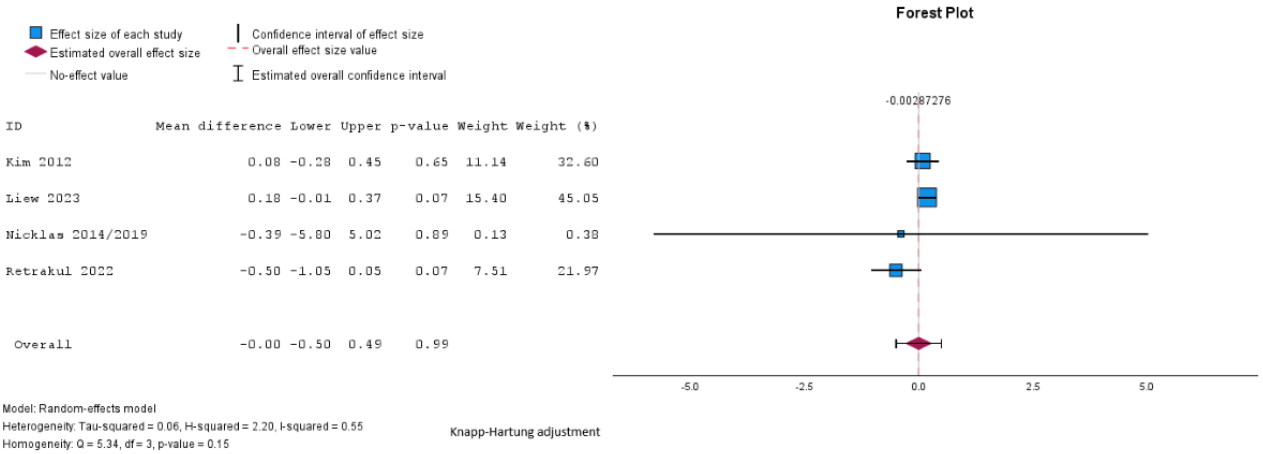

Figure 6. Random effects meta-analysis of the mean difference in 2-hour glucose comparing the intervention group and the control group [27,29,32,34].

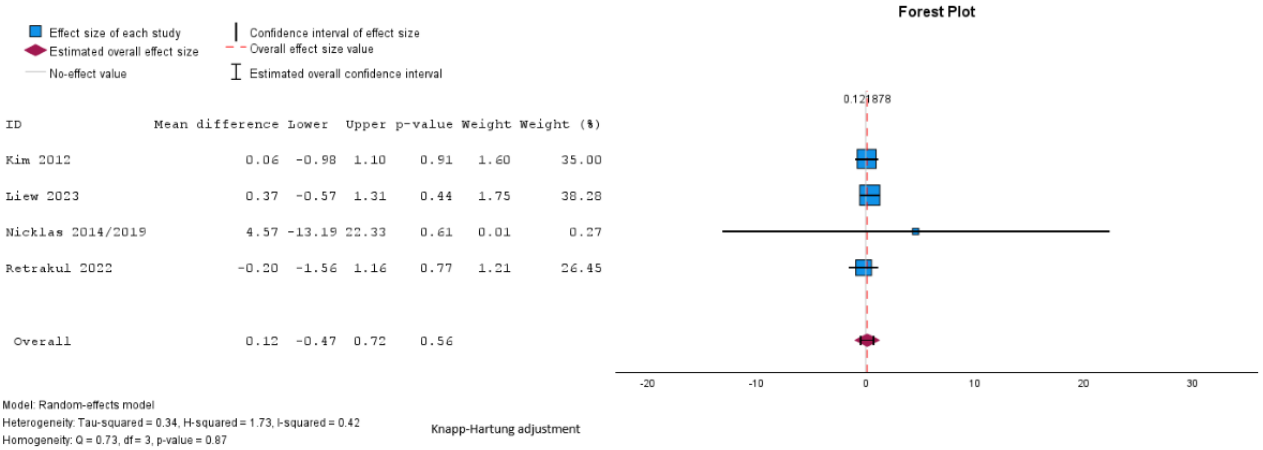

Figure 7. Random effects meta-analysis of the mean difference in hemoglobin A1c comparing the intervention group and the control group [32,35].

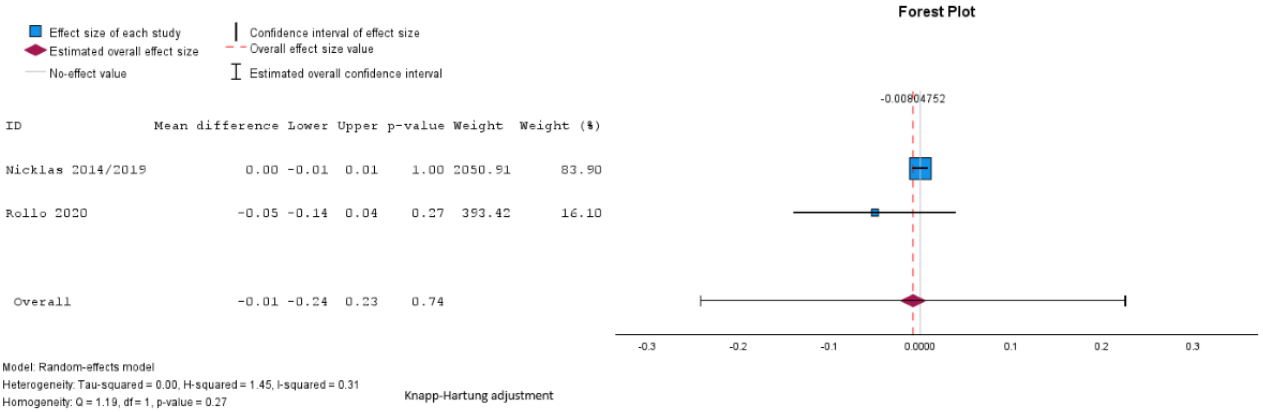

XSL•FO
RenderX

Figure 8. Random effects meta-analysis of the mean difference in homeostasis model assessment of insulin resistance comparing the intervention group and the control group [32,34].

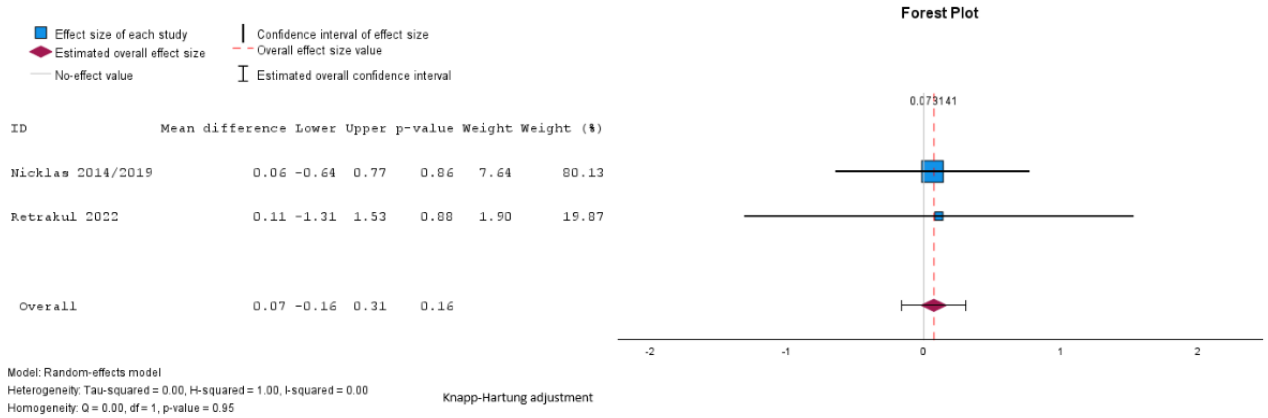

Behavioral Outcomes

A total of 8 studies reported diet outcomes [24,25,28,30,33,35,38,42], and 8 reported physical activity outcomes [24,25,27,32-35,38]. Only 1 study, testing an intervention aimed at improving sleep quality, reported any significant improvements in physical activity among women in the intervention group in comparison to the control group [34]. Two studies reported a significantly reduced calorie intake among those receiving the intervention compared to the control group [30,32], and 1 reported significantly more positive dietary habits among intervention participants [28]. Two studies reported on general measures of health-promoting lifestyles [28,30] and both found that the intervention group had significantly higher scores compared to the control group, indicating more healthy behaviors overall. Similarly, significantly more women in the intervention group of the study by Potzel et al [33] reported making changes to their health-related habits compared to the control group. It was not possible to synthesize behavioral outcomes using meta-analysis due to the variability in how outcomes were measured and reported.

Psychological Outcomes

A total of 4 studies measured self-efficacy for physical activity [27,31,36,37], 3 self-efficacy for diet [27,36,37], and 1 study measured self-efficacy for weight management [38]. Two studies found significant differences in self-efficacy for physical activity [36,37], but in 1, the differences were accounted for by the fact that the control group's self-efficacy worsened, whereas the intervention group remained stable [36]. One study found significant differences in self-efficacy for weight management [38], and 1 found significant differences in self-efficacy for diet [37]. Risk perceptions for type 2 diabetes were reported in 2 studies [26,27], with 1 finding significantly more changes in risk perceptions among women in the intervention group compared to the control group [26]. It was not possible to synthesize psychological outcomes using meta-analysis due to the variability in how these outcomes were measured and reported.

BCTs

A total of 34 different BCTs were coded across the 15 included studies out of the 93 BCTs described in the BCT taxonomy. The number of BCTs in each study ranged from 6 to 17, with a mean of 11.3 (SD 2.82) per study. A 2-tailed independent samples *t* test found no significant difference in the mean number of BCTs in effective interventions (mean 10.5, SD 1.5) and noneffective interventions (mean 12.3, SD 3.7; $t_{7.7}$=−1.19; *P*=.27). Table 4 summarizes the BCTs coded in all studies, and in effective vs noneffective interventions. The most commonly used BCTs were "instruction on how to perform the behavior," which was included in all studies, "credible source" (14/15, 93% of studies), "self-monitoring of behavior" (13/15, 87%), "feedback on behavior" (11/15, 73%), "adding objects to the environment" (10/15, 66%), "prompts and cues" (10/15, 66%), "goal setting (behavior)" (10/15, 66%), "action planning" (10/15, 66%), "problem-solving" (9/15, 60%), and "information about health consequences" (8/15, 53%).

XSL•FO
RenderX

Table 4. Summary of behavior change techniques in effective and noneffective interventions.

Behavior change technique	All studies (n=15), n (%)	Effective any outcome (n=8), n (%)	Not effective for any outcome (n=7), n (%)
Goals and planning			
Goal setting (behavior)	10 (66)	4 (50)	6 (86)
Problem-solving	9 (60)	4 (50)	5 (71)
Goal setting (outcome)	2 (13)	2 (25)	0 (0)
Action planning	10 (66)	5 (63)	5 (71)
Review behavior goal(s)	6 (40)	2 (25)	4 (57)
Discrepancy between current behavior and goal	1 (7)	0 (0)	1 (14)
Review outcome goals	1 (7)	1 (13)	0 (0)
Feedback and monitoring			
Feedback on behavior	11 (73)	5 (63)	6 (86)
Self-monitoring of behavior	13 (87)	7 (88)	6 (86)
Self-monitoring of outcomes of behavior	4 (27)	3 (38)	1 (14)
Feedback on outcomes of behavior	1 (7)	0 (0)	1 (14)
Social support			
Social support (unspecified)	9 (60)	3 (38)	6 (89)
Social support (practical)	2 (13)	2 (25)	0 (0)
Social support (emotional)	1 (7)	1 (13)	0 (0)
Shaping knowledge			
Instruction on how to perform the behavior	15 (100)	8 (100)	7 (100)
Natural consequences			
Information about health consequences	8 (53)	4 (50)	4 (57)
Salience of consequences	3 (20)	3 (38)	0 (0)
Information about social and environmental consequences	4 (27)	2 (25)	2 (29)
Information about emotional consequences	1 (7)	1 (13)	0 (0)
Comparison on behavior			
Demonstration of the behavior	7 (47)	6 (75)	1 (14)
Social comparison	1 (7)	0 (0)	1 (14)
Associations			
Prompts/cues	10 (66)	6 (75)	4 (57)
Repetition and substitution			
Behavioral practice/rehearsal	2 (13)	2 (25)	0 (0)
Habit formation	1 (7)	1 (13)	0 (0)
Graded tasks	2 (13)	0 (0)	2 (29)
Comparison on outcomes			
Credible source	14 (93)	8 (100)	6 (86)
Pros and cons	1 (7)	0 (0)	1 (14)
Reward and threat			
Material reward (behavior)	1 (7)	0 (0)	1 (14)
Nonspecific reward	1 (7)	0 (0)	1 (14)
Social reward	5 (33)	2 (25)	3 (43)
Regulation			

XSL•FO
RenderX

Behavior change technique	All studies (n=15), n (%)	Effective any outcome (n=8), n (%)	Not effective for any outcome (n=7), n (%)
Reduce negative emotions	1 (7)	0 (0)	1 (14)
Antecedents			
Restructuring the physical environment	1 (7)	0 (0)	1 (14)
Adding objects to the environment	10 (66)	7 (88)	3 (43)
Identity			
Framing/reframing	4 (27)	1 (13)	3 (43)

There were 6 BCTs identified as being present in 75% (6/8) or more of effective interventions: “self-monitoring of behavior,” “instruction on how to perform the behavior,” “demonstration of the behavior,” “prompts/cues, credible source,” and “adding objects to the environment.” However, these BCTs were all present in very similar proportions of the studies that were not effective, except for “demonstration of the behavior,” which was found only in 14% (1/7) of noneffective studies compared to 75% (6/8) of effective ones, and “adding objects to the environment,” which was present in 43% (3/7) of noneffective interventions compared to 88% (7/8) of effective interventions.

Differences between effective and noneffective interventions were noted for 3 other BCTs: “goal setting (behavior),” “social support (unspecified),” “salience of consequences.” The “goal setting (behavior)” BCT was used more frequently in noneffective interventions (6/7, 86%) than in effective interventions (4/8, 50%). Effective interventions less frequently used the BCT “social support (unspecified)” (3/8, 38%) compared to noneffective interventions (6/7, 86%), but more frequently used practical (2/8, 25%) or emotional social support (1/8, 13%), which were not used in any noneffective interventions. The “salience of consequences” BCT was used in 38% (3/8) of effective interventions and not used in any noneffective interventions. However, all of the comparisons between effective and noneffective interventions must be interpreted with caution due to the small number of studies in each group.

Discussion

Principal Findings

This systematic review identified 15 studies [24-38] reported in 17 papers [39,40], investigating interventions aiming to prevent type 2 diabetes among women with a history of gestational diabetes using primarily technology-based delivery. Weight loss was the most frequently assessed outcome [24,25,27-33,35,37,38], but other outcomes related to diabetes risk assessed included waist circumference [27,30,32,35], body composition [28,33,35], BMI [27,29], glycemic control [27-32,34,35,39], health behaviors [24,25,27,28,30,32-35,38], self-efficacy for behavior change [27,31,36-38], and risk perceptions of type 2 diabetes [26,27]. Only 1 study assessed type 2 diabetes prevalence and found no significant differences between the intervention and control group [32]. Pooled analysis of 7 studies [24,27,29,32,33,35,38] revealed a significantly greater weight loss among those receiving technology-based interventions. Interventions delivered using technology [27,29,32,33] and with a longer follow-up [25,32,33,35] resulted in increased weight loss compared to those using combined technology and telemedicine approaches [24,35,38] and studies with a shorter follow-up [27,29,38]. Meta-analysis showed no significant differences in BMI [27,29], fasting glucose [27,29,32,34], 2-hour glucose [27,29,32,34], HbA_{1c} [32,35], or HOMA-IR [32,34] between intervention and control groups. There was heterogeneity in the study populations, dropout rates, target, timing, and duration of the intervention, and modes of delivery. Comparable heterogeneity has been reported in other systematic reviews of technology-based or telemedicine interventions for diabetes prevention [11,13].

Comparison With Other Literature

Although most studies included in this review did not individually report significant effects of technology-based interventions on weight loss, pooled analysis revealed a significantly greater weight loss among those receiving technology-based interventions with a mean difference of –1.01 (95% CI –1.86 to –0.16) kg. Although this difference is relatively small, 1 study found that weight loss of 1 kg in 1079 participants of a diabetes prevention intervention reduced the risk of type 2 diabetes by 16% over 3 years [43]. A systematic review of primarily face-to-face lifestyle interventions for women with a history of gestational diabetes reported a mean difference of –1.06 (95% CI −1.68 to −0.44) kg in pooled analysis of 5 trials, suggesting that technology-based interventions may be close to face-to-face interventions in effectiveness for weight loss in this population [44], but with possible benefits in terms of the resources required for delivery. Research has supported the cost-effectiveness of digital health interventions for supporting behavioral change in other populations [45,46], and so it is possible that technology-based interventions could offer a cost-effective means of supporting lifestyle change among women following a diagnosis of GDM. However, none of the studies in this review included any economic analysis, and this should therefore be explored in future trials.

Subgroup analyses carried out in this meta-analysis showed that interventions delivered using technology-based methods only resulted in more weight loss compared to those interventions using a combination of technology-based and telemedicine approaches, but these differences were not statistically significant. Studies with a follow-up of 6 months or more also resulted in increased weight loss than those with a shorter follow-up, which is consistent with a previous systematic review by Goveia et al [47] that included both technology-based and

XSL•FO
RenderX

face-to-face lifestyle interventions for diabetes prevention in women with GDM. Intervention length and follow-up length in studies in this review were largely similar, and so these findings suggest that technology-based interventions for this population should have a duration of 6 months or more to maximize effectiveness. The need for help over a longer period is supported by women's descriptions of the challenges of managing lifestyle change alongside the multiple roles they fulfill in the postpartum period [10].

The findings in relation to glycemic outcomes were less positive than those for weight. Only 1 of the 5 studies assessing glycemic control reported significant differences favoring the intervention group, and pooled analyses reported no significant differences in fasting glucose, 2-hour glucose, HbA_{1c}, or HOMA-IR. The review by Goveia et al [47] also found no significant effect of lifestyle intervention on fasting and plasma glucose but did find a positive effect for HbA_{1c}. Goveia et al [47] found a reduction in the incidence of type 2 diabetes, which was larger in interventions that started within 6 months of women giving birth. Only 1 study in this review assessed type 2 diabetes and found no significant differences in prevalence between the intervention and control group, but the intervention started up to 3 years postdelivery in this study.

It was not possible to pool studies on physical activity and diet in this review, but individually, most studies did not show significant effects of the intervention on either outcome. In contrast, half of the studies assessing self-efficacy for physical activity and diet, or risk perceptions, reported significant differences that favored the intervention group. This suggests that technology-based interventions included in this review do appear to be capable of changing the antecedents of behavior, even if studies do not demonstrate significant changes in lifestyle behavior. However, it is difficult to draw conclusions from individual studies in this review due to the limitations of the studies in relation to sample size, as discussed in the Limitations section below.

There was no significant difference in the number of BCTs in effective studies compared to noneffective studies in this review. This is in contrast to the findings of a systematic review of technology-driven type 2 diabetes prevention interventions in high-risk groups, excluding women with GDM, by Van Rhoon et al [11], which found that effective interventions had more BCTs on average. In our review, there was an average of 11.3 (SD 2.82) BCTs across all interventions compared to an average of 9 in the interventions included in Van Rhoon et al's review [11]. It has been argued that it is important to include fewer, more effective BCTs in technology-based interventions where attrition rates can be high, and too many BCTs might make the intervention more time-consuming and less engaging [48]. Research on BCTs in technology-based interventions aiming to promote physical activity and reduce sedentary behavior in adults suggests that the specific combination of BCTs may also be important, with certain combinations being more effective than others [48]. Given that there was no difference in the number of BCTs between effective and noneffective interventions in this review, and the challenges and competing demands that women report when attempting lifestyle change in the postpartum period [10], future research should arguably focus on what BCTs are likely to be effective in women with GDM, rather than the quantity of BCTs.

There was little difference in the BCTs contained in effective interventions and noneffective interventions in this review except for "demonstration of behavior," "adding objects to the environment," and specific types of "social support" (eg, "emotional or practical") being more common in effective interventions, whereas less effective interventions were more likely to contain "unspecified social support." Furthermore, descriptions of interventions were often lacking in detail, and these differences in social support may represent differences in reporting rather than true differences in intervention approaches. Clearer specification of interventions is needed in trials to build a cumulative evidence base of effective interventions that can be replicated and refined.

The most commonly included BCTs across all studies in this review were: "instruction on how to perform the behavior," "credible source," "self-monitoring of behavior," "feedback on behavior," "adding objects to the environment," "prompts and cues," "goal setting (behavior)," "problem-solving," "action planning," and "information about health consequences." The systematic review of technology-driven type 2 diabetes prevention interventions in high-risk groups, excluding women with GDM, by Van Rhoon et al [11] found that 90% of effective interventions included the BCTs "social support (unspecified)," "goal setting (outcome/behavior)," "feedback on behavior," and "self-monitoring of outcomes" [11]. Of these, "social support (unspecified)," "goal setting (outcomes)," and "self-monitoring (outcomes)" were not among the most used BCTs in this review, and future research should consider the inclusion of these BCTs and assess their transferability from other high-risk groups to women with previous GDM.

Limitations

Limitations of the review include that non-English papers were excluded, gray literature was not identified, and data were not independently extracted but instead checked by a second reviewer. Prediction intervals could not be calculated in the meta-analysis because of the small number of eligible studies. By reporting only CIs and not prediction intervals, we cannot draw conclusions about the possible variation in the intervention effect across different populations and settings [22]. Key limitations of the studies included in the review were the small sample sizes and lack of power calculations [49]. Sample sizes in all included studies were very small, with all studies containing around 100 participants or fewer in each arm. Two-thirds of studies included in the review either did not report carrying out a power calculation or reported being underpowered. Given that meta-analysis showed significant reductions in weight favoring the interventions when most individual studies did not, it is possible that interventions did result in behavioral changes, but studies were not adequately powered to detect these differences [49]. The certainty of the evidence for all pooled outcomes in this review was rated as very low.

Interventions were often poorly described, making it difficult to code and identify potentially effective ingredients, and most

XSL•FO
RenderX

lacked an explicit theoretical basis to explain and understand how they might work. The explicit use of theory in interventions may increase the likelihood that an intervention will be effective by ensuring that the causal determinants of the behavior are understood and addressed in the intervention [50]. There were not enough studies using theory in this review to suggest a particular model or framework that should be used, but instead, we recommend that, to maximize the potential effectiveness of technology-based interventions in this group and facilitate replicability, future research should clearly describe the content of interventions and consider the theoretical basis for the intervention. There was also a variety in the mode of delivery of digital interventions, and future research should explore whether providing participants with additional devices (eg, fitness trackers) offers any significant benefit over them using their own devices (eg, smartphones) that outweighs the costs of these devices.

Conclusions

This review is the first to summarize interventions with primarily technology-based delivery to prevent type 2 diabetes in women following a diagnosis of GDM. Previously published reviews included interventions with any digital component, and in many of the studies reviewed, the digital aspect was an adjunct to in-person or telemedicine delivery, meaning that conclusions could not be clearly drawn about the benefit of primarily digital interventions [13]. This review is also the first to code the BCTs in interventions with the aim of preventing type 2 diabetes in this population, and so can suggest techniques that should be explored in future interventions. The findings of the review suggest that there may be potential for technology-based interventions to support women in reducing their risk of type 2 diabetes following GDM. However, substantial heterogeneity, significant risk of bias, and very low certainty in the evidence according to the GRADE assessment mean that the findings must be interpreted cautiously. Further, high-quality trials with larger samples, longer follow-up, and clearer reporting of interventions are required to allow firm conclusions to be drawn.

Acknowledgments

The authors thank the Faculty of Health Sciences and Sport, University of Stirling, for supporting the publication of this manuscript. No generative artificial intelligence was used at any stage of this meta-analysis or writing of the manuscript.

Data Availability

The datasets generated during this study are available in the additional supporting files for this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA checklist.
[PDF File (Adobe PDF File), 190 KB - jmir_v28i1e78841_app1.pdf]

Multimedia Appendix 2
PRISMA abstract.
[PDF File (Adobe PDF File), 143 KB - jmir_v28i1e78841_app2.pdf]

Multimedia Appendix 3
PRISMA-S checklist.
[PDF File (Adobe PDF File), 137 KB - jmir_v28i1e78841_app3.pdf]

Multimedia Appendix 4
Full search strategy.
[DOCX File , 26 KB - jmir_v28i1e78841_app4.docx]

Multimedia Appendix 5
Data file.
[XLSX File (Microsoft Excel File), 9 KB - jmir_v28i1e78841_app5.xlsx]

References

XSL•FO
RenderX

1. Cho NH, Shaw JE, Karuranga S, Huang Y, da Rocha Fernandes JD, Ohlrogge AW, et al. IDF diabetes atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res Clin Pract 2018;138:271-281. [doi: 10.1016/j.diabres.2018.02.023] [Medline: 29496507]
2. Xu G, Liu B, Sun Y, Du Y, Snetselaar LG, Hu FB, et al. Prevalence of diagnosed type 1 and type 2 diabetes among US adults in 2016 and 2017: population based study. BMJ 2018;362:k1497 [FREE Full text] [doi: 10.1136/bmj.k1497] [Medline: 30181166]
3. Dennison RA, Chen ES, Green ME, Legard C, Kotecha D, Farmer G, et al. The absolute and relative risk of type 2 diabetes after gestational diabetes: a systematic review and meta-analysis of 129 studies. Diabetes Res Clin Pract 2021;171:108625 [FREE Full text] [doi: 10.1016/j.diabres.2020.108625] [Medline: 33333204]
4. Eades CE, Cameron DM, Evans JM. Prevalence of gestational diabetes mellitus in Europe: a meta-analysis. Diabetes Res Clin Pract 2017;129:173-181. [doi: 10.1016/j.diabres.2017.03.030] [Medline: 28531829]
5. Ferrara A. Increasing prevalence of gestational diabetes mellitus: a public health perspective. Diabetes Care 2007;30 Suppl 2:S141-S146. [doi: 10.2337/dc07-s206] [Medline: 17596462]
6. Kim C, Newton KM, Knopp RH. Gestational diabetes and the incidence of type 2 diabetes: a systematic review. Diabetes Care 2002;25(10):1862-1868. [doi: 10.2337/diacare.25.10.1862] [Medline: 12351492]
7. Ratner RE, Christophi CA, Metzger BE, Dabelea D, Bennett PH, Pi-Sunyer X, Diabetes Prevention Program Research Group. Prevention of diabetes in women with a history of gestational diabetes: effects of metformin and lifestyle interventions. J Clin Endocrinol Metab 2008;93(12):4774-4779 [FREE Full text] [doi: 10.1210/jc.2008-0772] [Medline: 18826999]
8. Lie MLS, Hayes L, Lewis-Barned NJ, May C, White M, Bell R. Preventing type 2 diabetes after gestational diabetes: women's experiences and implications for diabetes prevention interventions. Diabet Med 2013;30(8):986-993. [doi: 10.1111/dme.12206] [Medline: 23534548]
9. Eades CE, France EF, Evans JMM. Postnatal experiences, knowledge and perceptions of women with gestational diabetes. Diabet Med 2018;35(4):519-529. [doi: 10.1111/dme.13580] [Medline: 29338094]
10. Parsons J, Ismail K, Amiel S, Forbes A. Perceptions among women with gestational diabetes. Qual Health Res 2014;24(4):575-585. [doi: 10.1177/1049732314524636] [Medline: 24682021]
11. Van Rhoon L, Byrne M, Morrissey E, Murphy J, McSharry J. A systematic review of the behaviour change techniques and digital features in technology-driven type 2 diabetes prevention interventions. Digit Health 2020;6:2055207620914427 [FREE Full text] [doi: 10.1177/2055207620914427] [Medline: 32269830]
12. Leblalta B, Kebaili H, Sim R, Lee SWH. Digital health interventions for gestational diabetes mellitus: a systematic review and meta-analysis of randomised controlled trials. PLOS Digit Health 2022;1(2):e0000015 [FREE Full text] [doi: 10.1371/journal.pdig.0000015] [Medline: 36812531]
13. Halligan J, Whelan ME, Roberts N, Farmer AJ. Reducing weight and BMI following gestational diabetes: a systematic review and meta-analysis of digital and telemedicine interventions. BMJ Open Diabetes Res Care 2021;9(1):e002077 [FREE Full text] [doi: 10.1136/bmjdrc-2020-002077] [Medline: 33853849]
14. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
15. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021;372:n160 [FREE Full text] [doi: 10.1136/bmj.n160] [Medline: 33781993]
16. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
17. PROSPERO. URL: https://www.crd.york.ac.uk/prospero/ [accessed 2026-02-26]
18. National Institute for Health and Care Excellence. Process and methods guides. In: Methods for the Development of NICE Public Health Guidance [Internet]. London: National Institute for Health and Care Excellence (NICE); 2020.
19. Cumpston M, Li T, Page MJ, Chandler J, Welch VA, Higgins JP, et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane handbook for systematic reviews of interventions. Cochrane Database Syst Rev 2019;10(10):ED000142 [FREE Full text] [doi: 10.1002/14651858.ED000142] [Medline: 31643080]
20. IntHout J, Ioannidis JPA, Borm GF. The hartung-knapp-sidik-jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard derSimonian-laird method. BMC Med Res Methodol 2014;14:25 [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
21. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. Res Synth Methods 2017;8(1):5-18. [doi: 10.1002/jrsm.1230] [Medline: 28058794]
22. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023;12(4):101014 [FREE Full text] [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]

XSL•FO
RenderX

23. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;64(4):383-394. [doi: 10.1016/j.jclinepi.2010.04.026] [Medline: 21195583]
24. Cheung NW, Blumenthal C, Smith BJ, Hogan R, Thiagalingam A, Redfern J, et al. A pilot randomised controlled trial of a text messaging intervention with customisation using linked data from wireless wearable activity monitors to improve risk factors following gestational diabetes. Nutrients 2019;11(3):590 [FREE Full text] [doi: 10.3390/nu11030590] [Medline: 30862052]
25. Cheung NW, Simmons D, Marschner S, Thiagalingam A, Pasupathy D, Smith BJ, et al. Randomised controlled trial of a customised text messaging and activity monitor program for lifestyle improvement after gestational diabetes. Nutrients 2024;16(6):820 [FREE Full text] [doi: 10.3390/nu16060820] [Medline: 38542731]
26. Ghaderi M, Farahani MA, Hajiha N, Ghaffari F, Haghani H. The role of smartphone-based education on the risk perception of type 2 diabetes in women with gestational diabetes. Health Technol 2019;9(5):829-837. [doi: 10.1007/s12553-019-00342-3]
27. Kim C, Draska M, Hess ML, Wilson EJ, Richardson CR. A web-based pedometer programme in women with a recent history of gestational diabetes. Diabet Med 2012;29(2):278-283 [FREE Full text] [doi: 10.1111/j.1464-5491.2011.03415.x] [Medline: 21838764]
28. Kim SH, Kim HJ, Shin G. Self-management mobile virtual reality program for women with gestational diabetes. Int J Environ Res Public Health 2021;18(4):1539 [FREE Full text] [doi: 10.3390/ijerph18041539] [Medline: 33562853]
29. Liew SJ, Soon CS, Chooi YC, Tint MT, Eriksson JG. A holistic approach to preventing type 2 diabetes in Asian women with a history of gestational diabetes mellitus: a feasibility study and pilot randomized controlled trial. Front Clin Diabetes Healthc 2023;4:1251411 [FREE Full text] [doi: 10.3389/fcdhc.2023.1251411] [Medline: 37841647]
30. Lim K, Chan SY, Lim SL, Tai BC, Tsai C, Wong SR, et al. A smartphone app to restore optimal weight (SPAROW) in women with recent gestational diabetes mellitus: randomized controlled trial. JMIR Mhealth Uhealth 2021;9(3):e22147 [FREE Full text] [doi: 10.2196/22147] [Medline: 33724204]
31. Nicholson WK, Beckham AJ, Hatley K, Diamond M, Johnson L, Green SL, et al. The gestational diabetes management system (GooDMomS): development, feasibility and lessons learned from a patient-informed, web-based pregnancy and postpartum lifestyle intervention. BMC Pregnancy Childbirth 2016;16(1):277 [FREE Full text] [doi: 10.1186/s12884-016-1064-z] [Medline: 27654119]
32. Nicklas JM, Zera CA, England LJ, Rosner BA, Horton E, Levkoff SE, et al. A web-based lifestyle intervention for women with recent gestational diabetes mellitus: a randomized controlled trial. Obstet Gynecol 2014;124(3):563-570 [FREE Full text] [doi: 10.1097/AOG.0000000000000420] [Medline: 25162257]
33. Potzel AL, Gar C, Banning F, Sacco V, Fritsche A, Fritsche L, et al. A novel smartphone app to change risk behaviors of women after gestational diabetes: a randomized controlled trial. PLoS One 2022;17(4):e0267258 [FREE Full text] [doi: 10.1371/journal.pone.0267258] [Medline: 35476681]
34. Reutrakul S, Martyn-Nemeth P, Quinn L, Rydzon B, Priyadarshini M, Danielson KK, et al. Effects of sleep-extend on glucose metabolism in women with a history of gestational diabetes: a pilot randomized trial. Pilot Feasibility Stud 2022;8(1):119 [FREE Full text] [doi: 10.1186/s40814-022-01076-2] [Medline: 35659776]
35. Rollo ME, Baldwin JN, Hutchesson M, Aguiar EJ, Wynne K, Young A, et al. The feasibility and preliminary efficacy of an eHealth lifestyle program in women with recent gestational diabetes mellitus: a pilot study. Int J Environ Res Public Health 2020;17(19):7115 [FREE Full text] [doi: 10.3390/ijerph17197115] [Medline: 32998401]
36. Saxon CE, Seely EW, Bertin KB, Suresh K, Skurnik G, Roche AT, et al. Self-efficacy and readiness to change among women with recent gestational diabetes engaging in a web-based lifestyle intervention: the balance after baby intervention trial. Am J Lifestyle Med 2026;20(3):385-395. [doi: 10.1177/15598276231155147] [Medline: 41640988]
37. Seely EW, Weitzman PF, Cortes D, Romero Vicente S, Levkoff SE. Development and feasibility of an app to decrease risk factors for type 2 diabetes in Hispanic women with recent gestational diabetes (hola bebé, adiós diabetes): pilot pre-post study. JMIR Form Res 2020;4(12):e19677 [FREE Full text] [doi: 10.2196/19677] [Medline: 33382039]
38. Taylor R, Rollo ME, Baldwin JN, Hutchesson M, Aguiar EJ, Wynne K, et al. Evaluation of a type 2 diabetes risk reduction online program for women with recent gestational diabetes: a randomised trial. Int J Behav Nutr Phys Act 2022;19(1):35 [FREE Full text] [doi: 10.1186/s12966-022-01275-3] [Medline: 35346229]
39. Nicklas JM, Rosner BA, Zera CA, Seely EW. Association between changes in postpartum weight and waist circumference and changes in cardiometabolic risk factors among women with recent gestational diabetes. Prev Chronic Dis 2019;16:E47 [FREE Full text] [doi: 10.5888/pcd16.180308] [Medline: 31002638]
40. Nicklas JM, Skurnik G, Roche T, Schultz C, Suresh K, Seely EW. 191-OR: a web-based lifestyle intervention to reduce postpartum weight retention in women with recent gestational diabetes: the balance after baby intervention trial. Diabetes 2020;69(Supplement_1):191-OR. [doi: 10.2337/db20-191-or]
41. Beeharry D, Swarbrick H, Ahmed H, Nair S, Pond J, Joseph F. Clinical care and other categories posters: education and self-management: P310 multimedia approach to diabetes structured education: the effects of “Keeping Healthy After Gestational Diabetes (GDM)”, a multimedia based education program on patient knowledge, self-efficacy and acceptability. Diabetic Medicine 2017;34(S1):107-130 [FREE Full text] [doi: 10.1111/dme.24_13304]

XSL•FO
RenderX

42. Equality Act 2010. legislation.gov.uk. 2010. URL: https://www.legislation.gov.uk/ukpga/2010/15/contents [accessed 2026-02-26]
43. Hamman RF, Wing RR, Edelstein SL, Lachin JM, Bray GA, Delahanty L, et al. Effect of weight loss with lifestyle intervention on risk of diabetes. Diabetes Care 2006;29(9):2102-2107 [FREE Full text] [doi: 10.2337/dc06-0560] [Medline: 16936160]
44. Gilinsky AS, Kirk AF, Hughes AR, Lindsay RS. Lifestyle interventions for type 2 diabetes prevention in women with prior gestational diabetes: a systematic review and meta-analysis of behavioural, anthropometric and metabolic outcomes. Prev Med Rep 2015;2:448-461 [FREE Full text] [doi: 10.1016/j.pmedr.2015.05.009] [Medline: 26844102]
45. Kyaw TL, Ng N, Theocharaki M, Wennberg P, Sahlen KG. Cost-effectiveness of digital tools for behavior change interventions among people with chronic diseases: systematic review. Interact J Med Res 2023;12:e42396 [FREE Full text] [doi: 10.2196/42396] [Medline: 36795470]
46. Gentili A, Failla G, Melnyk A, Puleo V, Tanna GLD, Ricciardi W, et al. The cost-effectiveness of digital health interventions: a systematic review of the literature. Front Public Health 2022;10:787135 [FREE Full text] [doi: 10.3389/fpubh.2022.787135] [Medline: 36033812]
47. Goveia P, Cañon-Montañez W, Santos DDP, Lopes GW, Ma RCW, Duncan BB, et al. Lifestyle intervention for the prevention of diabetes in women with previous gestational diabetes mellitus: a systematic review and meta-analysis. Front Endocrinol (Lausanne) 2018;9:583 [FREE Full text] [doi: 10.3389/fendo.2018.00583] [Medline: 30344509]
48. Schroé H, Van Dyck D, De Paepe A, Poppe L, Loh WW, Verloigne M, et al. Which behaviour change techniques are effective to promote physical activity and reduce sedentary behaviour in adults: a factorial randomized trial of an e- and m-health intervention. Int J Behav Nutr Phys Act 2020;17(1):127 [FREE Full text] [doi: 10.1186/s12966-020-01001-x] [Medline: 33028335]
49. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. Emerg Med J 2003;20(5):453-458 [FREE Full text] [doi: 10.1136/emj.20.5.453] [Medline: 12954688]
50. Michie S, Prestwich A. Are interventions theory-based? Development of a theory coding scheme. Health Psychol 2010;29(1):1-8 [FREE Full text] [doi: 10.1037/a0016939] [Medline: 20063930]

Abbreviations

BCT: behavior change technique
GDM: gestational diabetes mellitus
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
HbA1c: hemoglobin A1c
HOMA-IR: homeostasis model assessment of insulin resistance
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension

Edited by S Brini; submitted 12.Jun.2025; peer-reviewed by V Mohan, M Lof; comments to author 08.Sep.2025; accepted 13.Feb.2026; published 01.Apr.2026.

Please cite as:
Eades C, Nguyen-Hoang A, Hoyle L, Cameron D, Evans JMM
Technology-Based Interventions for Prevention of Type 2 Diabetes Following Gestational Diabetes: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e78841
URL: https://www.jmir.org/2026/1/e78841
doi:10.2196/78841
PMID:

©Claire Eades, Anh Nguyen-Hoang, Louise Hoyle, Dawn Cameron, Josie MM Evans. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Predictive Value of Machine Learning for Poststroke Mortality Risk: Systematic Review and Meta-Analysis

Yujie Chen[1*]; Zhujing Ou[1*]; Yiting Deng[1]; Anling Luo[1]; Xuzi Li[1]; Yujia Yang[1]; Qin Yang[1]; Xintong Wu[1*]; Muke Zhou[1*]

Department of Neurology, West China Hospital, Sichuan University, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Muke Zhou
Department of Neurology, West China Hospital, Sichuan University
37 Guo Xue Xiang
Chengdu,
China
Phone: 86 18980601686
Email: muke.zhou@scu.edu.cn

Abstract

Background: People with stroke face a high mortality risk, and an accurate prediction model is essential to the guidance of clinical decision-making in this population. Recently, with growing attention paid to machine learning (ML) in stroke care, some researchers have investigated the effectiveness of ML in predicting the mortality risk in stroke. However, systematic evidence is still lacking for its effectiveness.

Objective: This systematic review aims to evaluate the value of ML in predicting the stroke mortality risk. The findings are expected to offer an evidence-based basis for developing and assessing clinical risk prediction tools.

Methods: A search was made in Cochrane Library, PubMed, Embase, and Web of Science up to June 23, 2025, and studies that reported a complete performance of ML in predicting stroke mortality were included. Studies with only risk factors analyzed were excluded. The risk of bias of the included studies was assessed using PROBAST (Prediction model Risk of Bias Assessment Tool). Pooled risk ratios with 95% CIs and prediction intervals (PIs) were derived using the Hartung-Knapp-Sidik-Jonkman method under a random-effects model. Subgroup analyses were also conducted by model type, stroke type, patient source, and treatment background. Moreover, a metaregression was conducted on the C-index for out-of-hospital mortality at different time points to explore the influence of time factors on the model's predictive performance.

Results: Sixty-eight studies were included (23 predicting in-hospital mortality and 45 predicting out-of-hospital mortality), describing the development of 75 prediction models and 43 external validations. The follow-up period was 1 month to 15 years. For predicting in-hospital mortality, the external validation set had a pooled C-index of 0.727 (95% CI 0.677-0.781, 95% PI 0.521-1.000), with sensitivity and specificity of 0.64 (95% CI 0.57-0.70) and 0.74 (95% CI 0.70-0.77), respectively. For predicting out-of-hospital mortality, the pooled C-index was 0.847 (95% CI 0.808-0.887, 95% PI 0.750-0.956) in the external validation set, with sensitivity and specificity of 0.71 (95% CI 0.55-0.82) and 0.76 (95% CI 0.74-0.78), respectively. Comparatively, the overall pooled C-indexes were 0.788 (95% CI 0.766-0.810, 95% PI 0.621-0.999) and 0.812 (95% CI 0.798-0.826, 95% PI 0.693-0.952), respectively. The metaregression revealed a gradual decline in the predictive performance of the overall model and logistic regression model alone, whereas a random forest model maintained sustained performance. Age, National Institutes of Health Stroke Scale score, and stroke-related complications were the most frequently used variables for modeling.

Conclusions: This is the first meta-analysis to demonstrate that ML-based prediction of stroke mortality is feasible. The performance of ML supports its role as an auxiliary tool for identifying high-risk populations, thereby optimizing clinical monitoring and resource allocation. However, due to substantial heterogeneity and a relatively high risk of bias in available studies, caution is warranted in real-world application. The effectiveness of ML may vary across settings, and external validation is recommended before broader implementation.

Trial Registration: PROSPERO CRD420251086321; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251086321

(J Med Internet Res 2026;28:e83821) doi:10.2196/83821

XSL•FO
RenderX

KEYWORDS

stroke; artificial intelligence; mortality; prognosis; logistic regression; prediction model

Introduction

Rationale

Stroke, an acute focal neurological deficit with cerebrovascular causes, is primarily classified into ischemic stroke and hemorrhagic stroke (intracerebral hemorrhage and subarachnoid hemorrhage) according to its pathogenesis. Acute ischemic stroke (AIS) is the most frequent type, accounting for 60%-80% of the total [1]. As a leading cause of mortality and long-term disability globally, stroke greatly reduces patients' quality of life and brings a heavy socioeconomic burden [2,3]. Notably, stroke mortality is projected to rise by 50% from 2020 to 2050 [4]. Therefore, stroke has become a serious threat to human health.

The functional prognosis of some patients with ischemic stroke (especially cases within the time window) can benefit from aggressive treatments (eg, intravenous and mechanical thrombolysis). However, a large number of patients are still at risk of adverse outcomes, which deserves adequate attention. The global stroke mortality is high, with approximately 7.5 million deaths in 2021 [5]. In countries with high population density (eg, China), people with AIS have an in-hospital mortality or discharge against medical advice of about 6% [6], a 3-month mortality of 1.5%-3.2%, and a 1-year mortality of 3.4%-6% [7,8]. The 90-day mortality of AIS shows no significant difference between patients undergoing endovascular thrombectomy and those not undergoing endovascular thrombectomy [9]. Moreover, people with intracerebral hemorrhage have an in-hospital mortality or discharge against medical advice of approximately 21.8%, and a 1-year mortality of 17.9% [10]. Notably, unlike highly lethal acute-phase hemorrhagic stroke, ischemic stroke presents a unique pattern of "low acute-phase mortality and high medium- to long-term risks" [11,12], and its mortality is mostly attributed to complications, recurrence, and long-term dysfunction. As a result, time-dynamic risk prediction tools are urgently needed for patients with stroke. Developing stage-specific prediction protocols for high-risk stroke is of great value in early intervention, medical counseling, and resource optimization.

Although numerous studies have sought to predict poststroke functional outcome [13,14], relevant models apply primarily to those patients with rehabilitation potential. However, critically ill intensive care unit patients are faced with a unique and urgent requirement during clinical decision-making, that is, to accurately assess mortality risk to guide physician-patient communication, treatment strategies, and allocation of health care resources [15]. Nowadays, validated prediction tools are still lacking for the stroke mortality risk, and clinical assessment relies on traditional tools (eg, Acute Physiology and Chronic Health Evaluation [APACHE] II score). However, such a single-mode assessment is static and subjective, ignoring the inherent complexity of conditions [16]. Some studies have also been conducted on specific biomarkers. For example, the triglyceride-glucose index (an important biochemical indicator for insulin resistance) can predict AIS mortality [17], and adiponectin is independently associated with post-AIS 5-year mortality [18]. However, early prediction of poststroke death by these prediction tools is still lacking in clinical practice. In addition, the Delphi method can also be applied to the prediction of mortality risk, but the resulting bias is a serious clinical challenge. Therefore, developing early prediction tools for the mortality risk is urgently needed.

Objectives

As big data and artificial intelligence (AI) emerge, modeling by the growing volume of patient-related data has become a hot spot, which enables individualized outcome prediction [19]. For example, Bonkhoff and Grefkes [20] argued that machine learning (ML) plus clinical, electrophysiological, and neuroimaging data exhibit a high potential for poststroke functional recovery; Akay et al [21] found that AI achieves individualized outcome prediction and assists in treatment decision-making by analyzing multimodal data. The advantage of ML lies in its capability to integrate multisource heterogeneous data from traditional clinical practice, such as medical histories and medical images, and transform them into efficient predictive systems [22]. Unlike traditional methods reliant on rigorous statistical hypotheses, ML presets no data distributions and possesses the capability to process high-dimensional, large-scale data. Therefore, ML demonstrates higher accuracy and efficacy in disease classification and prognosis evaluation [23,24]. Given that patients with stroke often present with complex conditions and multiple complications, and predictions rely heavily on subjective neurological examinations nowadays, it is significant to integrate multimodal clinical data by ML for enhancing the predictive power for stroke mortality [25]. Under this context, some studies have also further investigated ML for predicting short- [26,27] or long-term [28,29] poststroke mortality based on individual data. However, the predictive performance of these models lacks systematic evidence due to the diversity of ML methods and variables, which is a challenge against the subsequent development of simpler clinical prediction tools. No studies have yet quantitatively assessed the effects of these ML models, so the clinical usefulness of the available predictive systems in routine practice remains unclear. Therefore, we conducted this meta-analysis with the following objectives: (1) to measure the predictive performance of ML for stroke mortality across scenarios and summarize the application status of different ML models in stroke mortality prediction, and (2) to evaluate the methodological quality of available studies and determine whether ML can serve as a novel tool for predicting in-hospital and out-of-hospital stroke mortality. The findings are expected to offer an evidence-based basis for facilitating AI development in this field and further developing simple and clinically useful ML-based prediction tools.

XSL•FO
RenderX

Methods

Study Registration

This meta-analysis followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 Guidelines [30] in Multimedia Appendix 1. The protocol was prospectively registered with PROSPERO (CRD420251086321). To incorporate constructive refinements suggested during the peer-review process, several deviations from the original PROSPERO protocol were implemented to enhance the study's rigor. These updates primarily include (1) restructuring the inclusion or exclusion criteria based on the PICOS framework; (2) expanding the analytical scope, including the addition of subgroup and sensitivity (SE) analyses; and (3) refining the statistical methodology by adopting the Hartung-Knapp-Sidik-Jonkman (HKSJ) approach. Accordingly, the protocol has been updated to version 2.0 on the PROSPERO platform to ensure full alignment with the final manuscript.

Eligibility Criteria

The eligibility criteria are as follows (Textbox 1):

Textbox 1. Eligibility criteria for study selection.

Inclusion criteria

- Populations: patients with stroke, including acute ischemic stroke and unspecified stroke types.
- Model construction: studies using artificial intelligence techniques such as machine learning for automated prediction.
- Outcomes: a complete prediction model for mortality risk.
- Study type: case-control, cohort, nested case-control, or case-cohort studies.
- Language: studies reported in English.

Exclusion criteria

- Populations: only patients with hemorrhagic stroke were included.
- Model construction: only risk factors were analyzed, without constructing a mortality prediction model; validation studies solely assessing established scales.
- Outcomes: any of the following evaluation metrics were lacking: receiver operating characteristic curve, C-index, sensitivity, specificity, accuracy, recall, precision, confusion matrix, F_1-score, or calibration curves.
- Study type: conference abstracts published without peer review, meta-analysis/review/guidelines/letter/case report/protocol.
- Publication year: earlier than 2005.
- Language: non-English language original studies.

Information Sources and Search Strategy

Literature search was carried out in adherence to PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) guidelines [31]. Cochrane Library (CENTRAL), PubMed (NCBI), Embase (Embase.com), and Web of Science (Clarivate) were thoroughly searched as of June 21, 2025. In addition, manual searches were conducted on references cited in systematic review reports on the same or similar topics. Developed based on the database requirements, the search strategy combined both medical subject headings ("Stroke," "Machine learning," "Deep learning," and "Mortality") and free-text keywords, as well as Boolean operators. To ensure a comprehensive retrieval, free-text keywords and relevant subordinate terms were identified via PubMed and Embase in addition to primary MeSH or Emtree terms and were further supplemented based on our clinical expertise. Details of the search strategy are available in Multimedia Appendix 2. Duplicate records were removed with EndNote 20 software (Clarivate Analytics). Search parameters were set to include only peer-reviewed papers, with no restrictions on geographical location or publication year, and no published search filters were used. No peer review of the search strategy was conducted.

Selection Process

All identified studies were imported into EndNote 20 software. All identified studies were imported into EndNote 20 software. After duplicate publications were removed, we used EndNote's filtering functionality to exclude records based on publication type (eg, reviews, case reports, letters, protocols, animal studies, and meeting abstracts), language (non-English), and publication date (pre-2005). Subsequently, 2 independent investigators (YJC and ZJO) screened the titles and abstracts of the remaining records based on predefined criteria regarding study population (stroke), intervention (machine learning), and outcomes (mortality). Thereafter, a rigorous full-text review was performed to exclude studies lacking model evaluation parameters or those that did not involve the development of original models. The screen was followed by cross-checking. In case of any disagreements, a third investigator (MKZ) was consulted for resolution.

Data Collection Process and Data Items

A spreadsheet was created to extract data, including title, first author, type of study, year of publication, patient source, dataset source, stroke type, treatment background (whether to undergo acute-phase vascular recanalization), follow-up time, total number of death cases in the training set, overfitting methods, number of cases in the validation set, handling methods for

XSL•FO
RenderX

missing value, variable selection, model types, and modeling variables. During data extraction, studies with missing outcome data were excluded directly from the meta-analysis. Three independent investigators (YJC, YTD, and ALL) implemented the data extraction, followed by cross-checking. In case of any disagreements, a third investigator (XTW) was consulted for resolution.

Risk of Bias Assessment

The included studies were assessed for risk of bias (RoB) using the PROBAST (Prediction model Risk of Bias Assessment Tool) [32] by structured questions across 4 domains: participants (2 questions), predictors (3 questions), outcomes (6 questions), and analysis (9 questions). Each question was answered as "Yes," "Probably yes," "No," "Probably no," or "No information." Each domain was rated as high RoB if 1 or more questions were answered as "No" or "Probably no," as low RoB if all answers were "Yes" or "Probably yes," and as unclear RoB if answered "No information." The overall risk was deemed low only when all domains were rated as low RoB, whereas any high RoB was considered overall high risk. In the absence of high-risk domains, the overall risk was unclear when 1 or more domains had unclear RoB. Two investigators (YJC and YJY) independently performed RoB assessment and cross-checking. Discrepancies were settled by third-party adjudication.

Synthesis Methods

The C-index, a measure for the overall predictive accuracy of ML models, underwent a meta-analysis. If the C-index was reported without its 95% CI or standard error, the SE was estimated using the method described by Debray et al [33]. Similarly, SE and specificity (SP) were synthesized by a bivariate mixed-effects model. Meta-analyses of these metrics typically rely on diagnostic 4-fold tables, but most primary studies included did not report such tables. In this case, we sought to reconstruct the diagnostic 4-fold table by leveraging other available performance metrics, such as SE, SP, precision, accuracy, and the total number of events (deaths). Some studies were based on the same database, but significant heterogeneity was present across these studies in the data extraction period, variable selection, model construction, and parameter adjustment rule. Therefore, all eligible studies were included, and given this potential heterogeneity, a random-effects model was adopted for meta-analysis [34]. Multiple entries from the same study represented distinct prediction models or validation methods, which were treated as independent units of evidence [35]. We performed random-effects meta-analyses using the HKSJ method to estimate pooled effect sizes with 95% CIs. The HKSJ method was chosen as it provides more robust inference, particularly when the number of studies is small. In addition, we calculated 95% prediction intervals (PIs) to present the expected range of true effects in similar studies, across clinical settings, or in future studies, and to better express the extent of heterogeneity [36,37]. The width of PIs was directly influenced by the number of included studies and the extent of heterogeneity [38]. Model performance was evaluated separately in training and validation sets. Furthermore, the performance of models for in-hospital and out-of-hospital mortality was separately synthesized.

To explore potential sources of heterogeneity, subgroup analyses were conducted, stratified by stroke type (AIS vs not specific), patient source (public database vs local hospital), treatment background (reperfusion therapy vs not specific), and model type. Furthermore, to evaluate the generalizability of the models and minimize optimistic bias associated with internal validation [39], we conducted a subgroup analysis in the validation set by validation method (internal validation vs external validation). When reporting the results, we prioritized the performance of the external validation subgroup based on the primary importance of external validation results for assessing clinical generalization [35]. The overall pooled performance (combining internal and external validation) was subsequently reported as a reference during development setting. Given the limited number of external validation studies, subgroup analyses were based on the overall performance. We used a random-effects model to calculate pooled estimates within each subgroup and a fixed-effects (plural) model to compare effect sizes across subgroups. A random-effects metaregression was conducted to investigate the association between the C-index of model performance and the follow-up time after discharge in studies reporting out-of-hospital mortality. The small-study effect in the validation sets of included studies and different models (≥10 studies) in the validation sets was qualitatively and quantitatively assessed with funnel plots and Egger regression tests, respectively. Funnel plot asymmetry was visualized to detect the small-study effect.

SE analyses were performed in 2 stages. In the first stage, to assess the potential impact of patient cohort overlap from using the same public databases (Medical Information Mart for Intensive Care [MIMIC-III/IV], eICU Collaborative Research Database [eICU-CRD], and Sentinel Stroke National Audit Programme [SSNAP]) on the pooled results, only 1 study with a large sample size, high methodological quality (based on PROBAST), and complete analysis (using external validation) was retained per public database. All studies from single-center or independent regional registries were also retained. Using independent datasets filtered in the 2-stage ways, a meta-analysis of the C-index for in-hospital and out-of-hospital mortality prediction models was conducted.

To mitigate potential bias arising from studies reporting multiple models, we established a priority rule for the second stage: selecting only the best-performing model from each study. Specifically, for studies with multiple models, we prioritized the optimal model among those with external validation; if no external validation was performed, the model with the highest internal performance was selected. In this way, we could obtain the evidence of the model with the greatest potential for clinical generalization or optimal performance from each study of independent sets, thereby assessing the robustness of the main findings under this stricter criterion. Additionally, following this criterion (the principle of selecting the study with the largest sample size among overlapping validation sets and subsequently selecting the optimal model within each independent validation set), a separate meta-analysis of the C-index was conducted in the external validation set. It should be noted that the SE

XSL•FO
RenderX

analyses were performed post hoc, and thus they were not prespecified in the preregistered study protocol or initial data analysis plan. Statistical analyses and visualizations were performed using Stata (version 15.0; StataCorp, LLC) and R (version 4.5.1; R Foundation for Statistical Computing), using metafor, forestplot, and ggplot2 packages.

Protocol Deviations and Updates

To incorporate constructive refinements from the peer-review process, several deviations from the original PROSPERO protocol were implemented to enhance the study's rigor. These updates primarily include (1) restructuring the inclusion or exclusion criteria according to the PICOS framework, (2) refining the statistical methodology by adopting the HKSJ method and reporting PIs, and (3) expanding the scope of analysis by adding SE and subgroup analyses (eg, internal vs external validation). Accordingly, the study protocol has been updated to version 2.0 on the PROSPERO platform to ensure full alignment with the final paper.

Results

Study Selection

We initially retrieved 21,177 records, of which 6491 duplicate publications were eliminated. After title and abstract review, 14,544 studies were further excluded. Then the full text of the remainder was examined, of which 38 were excluded due to no modeling, 31 due to no definite outcome metrics, and 5 due to a focus only on validation of previous scores. Finally, 68 studies were included [26-29,40-103] (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.

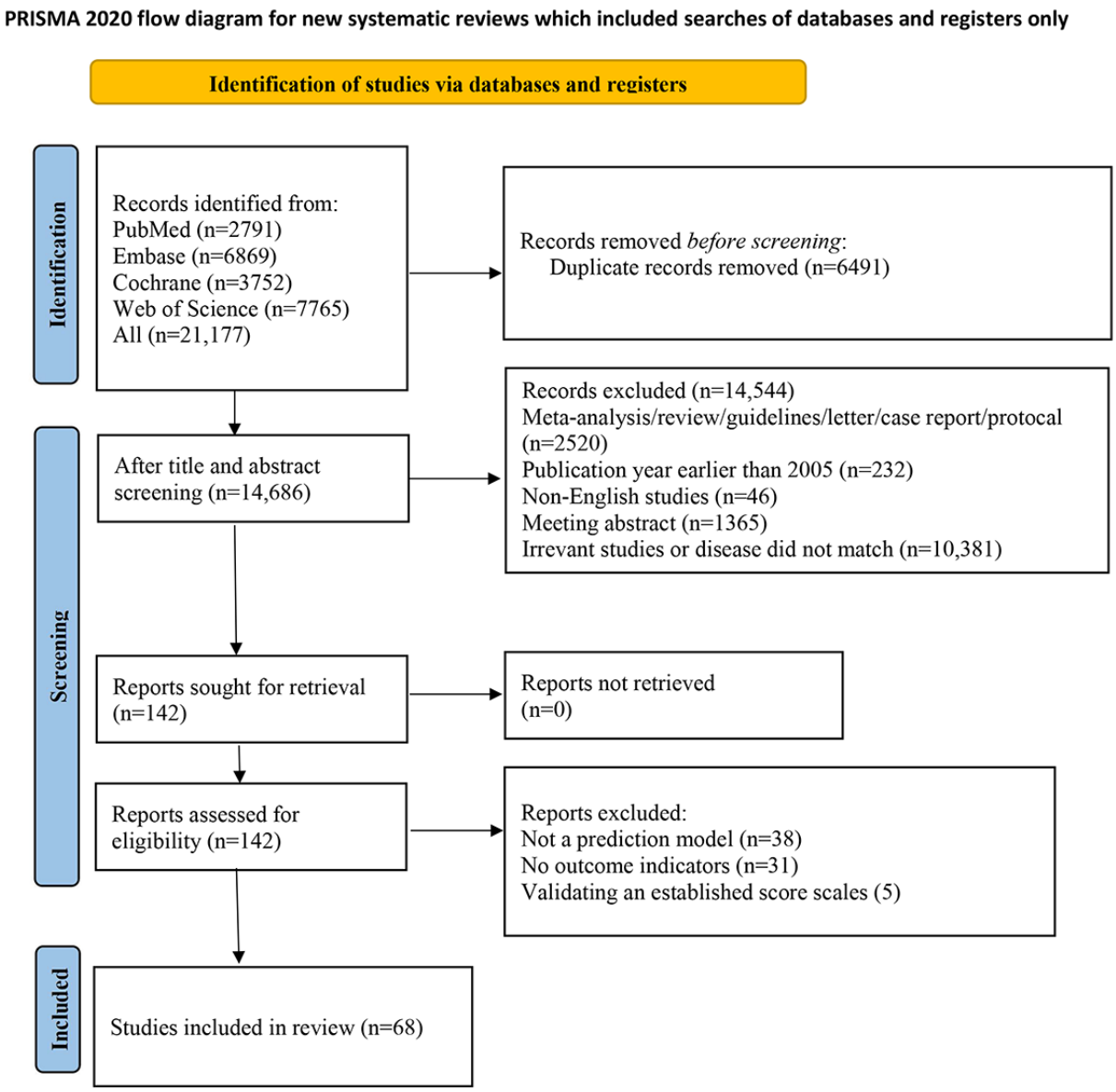

XSL•FO
RenderX

Study Characteristics

We included 68 studies published between 2005 and 2025, mainly from 2007 to 2025. They came from 19 countries and involved 1,759,472 participants, of whom 225,598 (12.8%) died. Twenty-three studies predicted the in-hospital mortality risk, and 45 studies predicted the out-of-hospital mortality risk. Three studies reported both in-hospital and out-of-hospital mortality [41,67,100]. Out-of-hospital mortality was followed up for 1-12 months, as well as 2 years [67], 3 years [46,72], 5 years [82], 10 years [45,75], and 15 years [40] in a few studies. Public databases, including MIMIC, eICU-CRD, the Minimum Basic Data Set, and SSNAP, were used in 28 studies, and there were 26 single-center studies and 14 multicenter studies. Forty-two studies focused on AIS, and the remaining 26 did not specify the stroke type. Seven studies predicted the mortality risk following acute-phase vascular recanalization. Logistic regression (LR) was the most frequently used model (51/68, 75%), followed by extreme gradient boosting (XGBoost) (21/68, 31%), random forest (RF) (20/68, 29%), support vector machine (11/68, 16%), and deep learning (DL) (2/68, 3%). For modeling variables, radiomics features were adopted in only 2 studies, and clinical features were adopted in the remainder. The generation method for the validation set was clearly described in 64 studies. Thirteen studies performed independent external validation, and 11 studies adopted both internal and external validation. For internal validation, random sampling was used in 44 studies, Bootstrap in 9 studies, k-fold cross-validation in 6 studies, holdout validation in 2 studies, and temporal validation in 1 study (Table S1 in Multimedia Appendix 2).

RoB in Studies

For participant selection, PROBAST considers retrospective cohort or case-control studies that do not use public databases with high RoB. Since 8 case-control studies did not use publicly available databases, they were rated as high RoB. All included studies developed reasonable eligibility criteria, so no RoB was introduced in this dimension.

The predictors had the same definition and assessment among participants in all studies, and they were all valid, so all studies were rated as low RoB. The outcome event was death, and the possibility of assessment with knowledge of the outcome could not be ruled out for case-control studies, so we rated 8 studies as high RoB.

The outcome in this meta-analysis was death or nondeath, and it was reasonably categorized and defined. The definition of the outcome eliminated the influence of predictors and was consistent among participants, and the outcome was assessed objectively with or without knowledge of the predictor. Therefore, all studies were rated as low RoB in this dimension.

For statistical analysis, 11 studies were assessed as high RoB because of inappropriate sample sizes; 11 as high RoB because subjects with missing data were directly excluded, without appropriate handling, and 3 as unclear RoB due to no description of the handling method for missing data. The RoB in 13 studies was high since only univariate analysis was conducted. The RoB in 10 studies was high for failing to account for data complexity, including high dimensionality, multicollinearity, or unbalanced data handling. Three studies were rated as high RoB for inappropriate internal validation. All studies were rated as low RoB since they all appropriately handled and incorporated continuous or categorical independent variables and assessed model performance.

To sum up, the overall RoB was low in 33 studies and high in 35 studies (Table S2 in Multimedia Appendix 2 [26-29,40-103] and Multimedia Appendix 3).

Results of Synthesis

In-hospital Mortality

Twenty-five ML models for predicting in-hospital mortality were extracted and included 1,345,411 patients, of whom 159,409 (11.7%) died. The pooled C-index was 0.815 (95% CI 0.783-0.848, 95% PI 0.666-0.997) (Table 1 and Figure 2). The C-index was 0.807 (95% CI 0.770-0.846, 95% PI 0.669-1.000) in studies on AIS, 0.811 (95% CI 0.771-0.853, 95% PI 0.652-1.000) in studies based on public databases, 0.825 (95% CI 0.759-0.897, 95% PI 0.657-1.000) in studies based on hospital databases, and 0.774 (95% CI 0.745-0.803, 95% PI 0.739-0.810) in studies explicitly reporting acute-phase vascular recanalization. LR, RF, and XGBoost models were dominated, and the LR model had a C-index of 0.803 (95% CI 0.742-0.868, 95% PI 0.638-1.000).

In addition, diagnostic metrics could be calculated in 24 studies in the training set, with a pooled SE of 0.81 (95% CI 0.70-0.88) and SP of 0.82 (95% CI 0.76-0.87). The SE and SP of the LR model were 0.71 (95% CI 0.61-0.79) and 0.78 (95% CI 0.69-0.85) (Table 2 and Figures S1 and S2 in Multimedia Appendix 2).

XSL•FO
RenderX

Table 1. Results of meta-analysis of C-index of machine learning for predicting in-hospital mortality in stroke.

Subgroup analysis	Training set						Validation set					
	n	Sample size	C-index (95% CI)	95% PI[a]	Tau	I^2 (%)	n	Sample size	C-index (95% CI)	95% PI	Tau	I^2 (%)
Model												
ANN[b]	1	14,8891	0.651 (0.646-0.656)	N/A[c]	N/A	N/A	8	6837	0.794 (0.696-0.905)	0.544-1.000	0.151	98.40
DT[d]	—[e]	—	—	—	—	—	5	5233	0.720 (0.648-0.800)	0.581-0.892	0.068	73.00
EL[f]	—	—	—	—	—	—	2	845	0.864 (0.373-1.000)	N/A	N/A	N/A
GBM[g]	2	2633	0.842 (0.400-1.000)	N/A	N/A	N/A	2	3813	0.813 (0.248-1.000)	N/A	N/A	N/A
KNN[h]	—	—	—	—	—	—	9	10,555	0.756 (0.707-0.805)	0.590-0.967	0.101	92.80
LASSO[i]	1	77,653	0.879 (0.871-0.886)	N/A	N/A	N/A	1	9141	0.884 (0.863-0.905)	N/A	N/A	N/A
LightGBM[j]	—	—	—	—	—	—	4	2223	0.828 (0.685-1.000)	0.549-1.000	0.114	92.80
LR[k]	8	17,2088	0.803 (0.742-0.868)	0.638-1.000	0.091	99.10	26	419,933	0.802 (0.762-0.844)	0.631-1.000	0.114	99.80
NB[l]	2	1300	0.838 (0.720-0.976)	N/A	N/A	N/A	5	2464	0.621 (0.339-1.139)	0.146-1.000	0.475	95.40
RF[m]	5	15,8453	0.789 (0.727-0.856)	0.647-0.963	0.065	98.30	14	13,721	0.811 (0.773-0.851)	0.682-0.963	0.077	93.20
SVM[n]	1	2031	0.924 (0.909-0.939)	N/A	N/A	N/A	7	7056	0.763 (0.715-0.814)	0.643-0.906	0.064	90.50
XGBoost[o]	5	12,128	0.844 (0.738-0.965)	0.609-1.000	0.107	99.30	16	269,990	0.803 (0.761-0.847)	0.650-0.991	0.096	97.60
Stroke type												
AIS[p]	20	56,6742	0.807 (0.770-0.846)	0.669-1.000	0.099	99.80	65	108,059	0.788 (0.766-0.810)	0.642-0.965	0.101	99.60
Not specific	5	8435	0.848 (0.769-0.936)	0.669-1.000	0.078	97.00	34	643,752	0.784 (0.733-0.840)	0.551-1.000	0.171	97.20
Data resource												
Public database	18	561,389	0.811 (0.771-0.853)	0.652-1.000	0.101	99.80	87	723,433	0.775 (0.753-0.799)	0.613-0.980	0.117	99.50
Local hospital	7	13,788	0.825 (0.759-0.897)	0.657-1.000	0.087	93.50	12	28,378	0.879 (0.840-0.921)	0.749-1.000	0.070	95.70
Treatment background												
EVT[q]/IVT[r]	4	572,769	0.774 (0.745-0.803)	0.739-0.810	0.000	0.00	—	—	—	—	—	—
Not specific	21	2408	0.823 (0.786-0.862)	0.663-1.000	0.101	99.80	—	—	—	—	—	—
Validation method												
Internal validation	—	—	—	—	—	—	67	480,774	0.815 (0.796-0.835)	0.681-0.976	0.089	99.60
External validation	—	—	—	—	—	—	32	271 037	0.727 (0.677-0.781)	0.521-1.000	0.170	96.20
Overall	25	575,177	0.815 (0.783-0.848)	0.666-0.997	0.096	99.70	99	751,811	0.788 (0.766-0.810)	0.621-0.999	0.119	99.50

[a]PI: prediction interval.

XSL•FO
RenderX

[b]ANN: artificial neural network.

[c]N/A: not applicable.

[d]DT: decision tree.

[e]Not available.

[f]EL: ensemble learning.

[g]GBM: gradient boosting machine.

[h]KNN: k-nearest neighbor.

[i]LASSO: least absolute shrinkage and selection operator.

[j]LightGBM: light gradient boosting machine.

[k]LR: logistic regression.

[l]NB: Naive Bayes.

[m]RF: random forest.

[n]SVM: support vector machine.

[o]XGBoost: extreme gradient boosting.

[p]AIS: acute ischemic stroke.

[q]EVT: endovascular thrombectomy.

[r]IVT: intravenous thrombolysis.

Figure 2. Forest plot for meta-analysis of C-index of predictive models for in-hospital mortality in stroke in the training set. AIS: acute ischemic stroke; ANN: artificial neural network; EVT: endovascular thrombectomy; GBM: gradient boosting machine; IVT: intravenous thrombolysis; LASSO: least absolute shrinkage and selection operator; LR: logistic regression; NB: Naive Bayes; PI: prediction interval; RF: random forest; SVM: support vector machine; XGBoost: extreme gradient boosting.

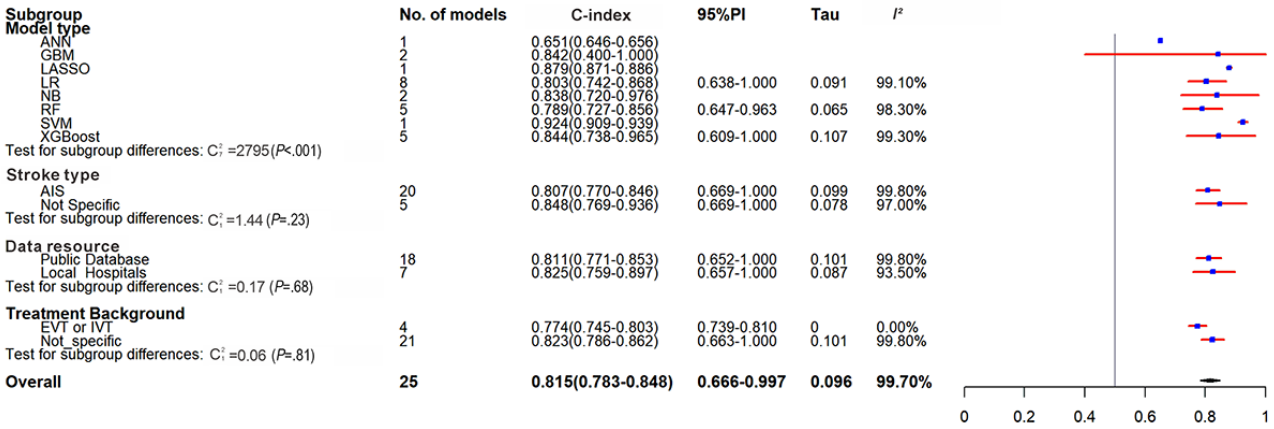

XSL•FO
RenderX

Table 2. Results of meta-analysis of sensitivity and specificity of machine learning for predicting in-hospital mortality in stroke.

Subgroup analysis	Training set				Validation set			
	n	Sample size	SE[a] (95% CI)	SP[b] (95% CI)	n	Sample size	SE (95% CI)	SP (95% CI)
Model								
ANN[c]	2	14,9231	0.92 (0.86-0.98)	0.81 (0.64-0.97)	9	7177	0.74 (0.58-0.85)	0.79 (0.72-0.85)
DT[d]	—[e]	—	—	—	5	5233	0.65 (0.44-0.81)	0.78 (0.52-0.92)
EL[f]	—	—	—	—	2	845	0.88 (0.82-0.94)	0.89 (0.86-0.91)
GBM[g]	2	2633	0.88 (0.72-0.79)	0.79 (0.75-0.83)	2	3813	0.58 (0.33-0.83)	0.78 (0.58-0.98)
KNN[h]	—	—	—	—	9	10,555	0.52 (0.34-0.69)	0.83 (0.70-0.91)
LASSO[i]	1	77,653	0.81	0.81	1	9141	0.79	0.84
LightGBM[j]	—	—	—	—	4	2223	0.73 (0.58-0.83)	0.82 (0.71-0.90)
LR[k]	7	166,373	0.71 (0.61-0.79)	0.78 (0.69-0.85)	19	189,436	0.72 (0.66-0.78)	0.78 (0.70-0.84)
NB[l]	2	1300	0.82 (0.70-0.93)	0.74 (0.69-0.77)	5	2464	0.64 (0.36-0.85)	0.70 (0.61-0.77)
RF[m]	4	152,738	0.66 (0.62-0.70)	0.78 (0.68,0.86)	13	12,292	0.74 (0.63-0.82)	0.81 (0.69-0.89)
SVM[n]	1	2031	0.90	0.85	7	7056	0.61 (0.51-0.70)	0.77 (0.71-0.82)
XGBoost[o]	5	6753	0.78 (0.62-0.88)	0.90 (0.77-0.96)	13	138,201	0.69 (0.62-0.75)	0.73 (0.68-0.78)
Stroke type								
AIS[p]	17	549,597	0.73 (0.66-0.79)	0.82 (0.75-0.88)	58	102,373	0.70 (0.66-0.73)	0.77 (0.74-0.80)
Not specific	7	9115	0.92 (0.67-0.98)	0.83 (0.71-0.90)	31	286,063	0.67 (0.57-0.75)	0.82 (0.74-0.88)
Data resource								
Public database	15	544,244	0.74 (0.67-0.80)	0.82 (0.75-0.88)	76	359,666	0.70 (0.66-0.73)	0.75 (0.72-0.77)
Local hospital	9	14,468	0.89 (0.67-0.97)	0.82 (0.72-0.89)	13	28,770	0.63 (0.43-0.80)	0.93 (0.87-0.97)
Treatment background								
EVT[q]/IVT[r]	4	2408	0.68 (0.64-0.72)	0.76 (0.69-0.82)	N/A[s]	N/A	N/A	N/A
Not specific	20	556,304	0.83 (0.71-0.91)	0.83 (0.77-0.88)	N/A	N/A	N/A	N/A
Validation method								
Internal validation	—	—	—	—	56	111,5921	0.71 (0.66-0.76)	0.81 (0.77-0.85)
External validation	—	—	—	—	33	272,515	0.64 (0.57-0.70)	0.74 (0.70-0.77)
Overall	24	558,712	0.81 (0.70-0.88)	0.82 (0.76-0.87)	89	388,436	0.69 (0.65-0.73)	0.79 (0.76-0.82)

[a]SE: sensitivity.
[b]SP: specificity.
[c]ANN: artificial neural network.
[d]DT: decision tree
[e]Not available.
[f]EL: ensemble learning.
[g]GBM: gradient boosting machine.
[h]KNN: k-nearest neighbor.
[i]LASSO: least absolute shrinkage and selection operator.
[j]LightGBM: light gradient boosting machine.
[k]LR: logistic regression.
[l]NB: Naive Bayes.
[m]RF: random forest.
[n]SVM: support vector machine.
[o]XGBoost: extreme gradient boosting.

XSL•FO
RenderX

[p]AIS: acute ischemic stroke.

[q]EVT: endovascular thrombectomy.

[r]IVT: intravenous thrombolysis.

[s]N/A: not applicable.

Based on externally validated models, the analysis indicated a moderate level of generalizability, with a pooled C-index of 0.727 (95% CI 0.677-0.781, 95% PI 0.521-1.000), with an SE of 0.64 (95% CI 0.57-0.70) and an SP of 0.74 (95% CI 0.70-0.77). These performance estimates were lower than the overall pooled results which yielded a C-index of 0.788 (95% CI 0.766-0.810, 95% PI 0.621-0.999), an overall SE of 0.69 (95% CI 0.65-0.73), and an SP of 0.79 (95% CI 0.76-0.82). The models using hospital databases outperformed those using public databases; the former had a C-index of 0.879 (95% CI 0.840-0.921, 95% PI 0.621-0.999), with SP and SE of 0.93 (95% CI 0.87-0.97) and 0.63 (95% CI 0.43-0.80), and the latter had a C-index of 0.775 (95% CI 0.753-0.799, 95% PI 0.613-0.980), with SP and SE of 0.75 (95% CI 0.72-0.77) and 0.70 (95% CI 0.66-0.73). The studies on AIS had a C-index of 0.788 (95% CI 0.766-0.810, 95% PI 0.642-0.965), with SP and SE of 0.70 (95% CI 0.66-0.73) and 0.77 (95% CI 0.74-0.80), and the LR model had a C-index of 0.802 (95% CI 0.762-0.844, 95% PI 0.631-1.000), with SE and SP of 0.72 (95% CI 0.66-0.78) and 0.78 (95% CI 0.70-0.84), consistent with the overall results. The internal validation subgroup aligned with the overall trend but yielded more optimistic estimates, with a pooled C-index of 0.815 (95% CI 0.796-0.835, 95% PI 0.681-0.976), an SE of 0.71 (95% CI 0.66-0.76), and an SP of 0.81 (95% CI 0.77-0.85) (Tables 1 and 2, Figure 3, and Figures S3 and S4 in Multimedia Appendix 2).

Figure 3. Forest plot for meta-analysis of C-index of predictive models for inhospital mortality in stroke in the validation set. AIS: acute ischemic stroke; ANN: artificial neural network; DT: decision tree; EL: ensemble learning; GBM: gradient boosting machine; KNN: k-nearest neighbor; LASSO: least absolute shrinkage and selection operator; LightGBM: light gradient boosting machine; LR: logistic regression; NA: Not applicable; NA:Not applicable; NB: Naive Bayes; PI: prediction interval; RF: random forest; SVM: support vector machine; XGBoost: extreme gradient boosting.

Subgroup	No. of models	C-index	95%PI	Tau	I^2
Model type					
ANN	8	0.794(0.696-0.905)	0.544-1.000	0.151	98.40%
DT	5	0.720(0.648-0.800)	0.581-0.892	0.068	73.00%
EL	2	0.864(0.373-1.000)	Na	Na	Na
GBM	2	0.813(0.248-1.000)	Na	Na	Na
KNN	9	0.756(0.707-0.805)	0.590-0.967	0.101	92.80%
LASSO	1	0.884(0.863-0.905)	Na	Na	Na
LightGBM	4	0.828(0.685-1.000)	0.549-1.000	0.114	92.80%
LR	26	0.802(0.762-0.844)	0.631-1.000	0.114	99.80%
NB	5	0.621(0.339-1.139)	0.146-1.000	0.475	95.40%
RF	14	0.811(0.773-0.851)	0.682-0.963	0.077	93.20%
SVM	7	0.763(0.715-0.814)	0.643-0.906	0.064	90.50%
XGBoost	16	0.803(0.761-0.847)	0.650-0.991	0.096	97.60%
Test for subgroup differences: C^2_{11}=64 (P<.001)					
Stroke type					
AIS	65	0.788(0.766-0.810)	0.642-0.965	0.101	99.60%
Not Specific	34	0.784(0.733-0.840)	0.551-1.000	0.171	97.20%
Test for subgroup differences: C^2_1=2795(P<.001)					
Data resource					
Public Database	87	0.775(0.753-0.799)	0.613-0.980	0.117	99.50%
Local Hospitals	12	0.879(0.840-0.921)	0.749-1.000	0.070	95.70%
Test for subgroup differences: C^2_1=0.01 (P<.001)					
Validation method					
Internal validation	67	0.815(0.796-0.835)	0.681-0.976	0.089	99.60%
External validation	32	0.727(0.677-0.781)	0.511-1.000	0.170	96.20%
Test for subgroup differences:C^2_1=9.56 (P=.002)					
Overall	**99**	**0.788(0.766-0.810)**	**0.621-0.999**	**0.119**	**99.50%**

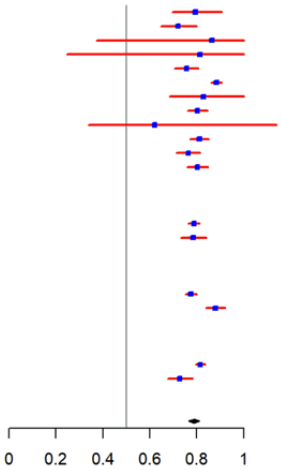

Out-of-Hospital Mortality

Fifty ML models for predicting out-of-hospital mortality were extracted and included 414,061 patients, of whom 66,189 (16%) died. The pooled C-index was 0.823 (95% CI 0.798-0.848, 95% PI 0.672-1.000) (Table 3 and Figure 4), and the diagnostic SE and SP were 0.66 (95% CI 0.60-0.72) and 0.88 (95% CI 0.83-0.91) (Figures S5 and S6 in Multimedia Appendix 2).

XSL•FO
RenderX

Table 3. Results of meta-analysis of C-index of machine learning for predicting out-of-hospital mortality in stroke.

Subgroup analysis	Training set						Validation set					
	n	Sample size	C-index (95% CI)	95% PI[a]	Tau	I^2 (%)	n	Sample size	C-index (95% CI)	95% PI	Tau	I^2 (%)
Model												
ANN[b]	3	4330	0.813 (0.768-0.860)	0.779-0.847	0.001	46.80	4	2615	0.822 (0.653-1.000)	0.498-1.000	0.140	95.3
DT[c]	1	474	0.793 (0.697-0.889)	N/A[d]	N/A	N/A	3	1906	0.69 5 (0.585-0.825)	0.556-0.868	0.025	0.0
EL[e]	—[f]	—	—	—	—	—	3	17,053	0.857 (0.708-1.000)	0.588-1.000	0.075	97.4
GBM[g]	3	3333	0.863 (0.660-1.000)	0.515-1.000	0.103	92.20	5	2076	0.788 (0.717-0.866)	0.652-0.953	0.058	57.6
KNN[h]	2	2859	0.669 (0.0356-1.000)	N/A	N/A	N/A	4	5205	0.737 (0.637-0.852)	0.578-0.939	0.065	80.1
LASSO[i]	—	—	—	—	—	—	—	—	—	—	—	—
LightGBM[j]	1	1388	0.840 (0.790-0.890)	N/A	N/A	N/A	2	2116	0.926 (0.665-1.000)	N/A	N/A	N/A
LR[k]	24	151,284	0.822 (0.795-0.850)	0.698-0.968	0.077	98.10	27	149,308	0.817 (0.793-0.841)	0.708-0.942	0.068	96.8
NB[l]	3	3333	0.832 (0.628-1.000)	0.490-1.000	0.105	86.60	6	6687	0.809 (0.741-0.883)	0.651-1.000	0.076	86.3
RF[m]	5	9502	0.835 (0.699-0.996)	0.549-1.000	0.137	99.50	16	17,729	0.833 (0.801-0.865)	0.724-0.957	0.063	83.1
RSF[n]	—	—	—	—	—	—	2	383	0.789 (0.630-0.986)	N/A	N/A	N/A
SVM[o]	2	2859	0.762 (0.249-1.000)	N/A	N/A	N/A	7	7225	0.789 (0.721-0.863)	0.620-1.000	0.091	92.1
XGBoost[p]	3	7660	0.857 (0.685-1.000)	0.552-1.000	0.088	97.70	10	11,327	0.800 (0.747-0.856)	0.643-0.994	0.091	96.8
DL[q]	1	1546	0.955 (0.925-0.984)	N/A	N/A	N/A	2	1161	0.873 (0.799-0.955)	N/A	N/A	N/A
COX[r]	1	425	0.822 (0.804-0.839)	N/A	N/A	N/A	5	4991	0.792 (0.720-0.871)	0.632-0.993	0.074	92.2
CatBoost[s]	1	1471	0.895 (0.878-0.912)	N/A	N/A	N/A	5	3902	0.818 (0.715-0.937)	0.599-1.000	0.102	94.1
AdaBoost[t]	—	—	—	—	—	—	1	1968	0.850 (0.811-0.889)	N/A	N/A	N/A
Stroke type												
AIS[u]	43	176,770	0.827 (0.801-0.855)	0.678-1.000	0.097	98.90	66	163,990	0.796 (0.778-0.813)	0.677-0.935	0.080	91.4
Not specific	7	13,694	0.796 (0.719-0.882)	0.597-1.000	0.110	99.20	36	71,662	0.841 (0.819-0.862)	0.733-0.964	0.066	97.8
Data resource												
Public database	24	171,811	0.826 (0.798-0.854)	0.696-0.979	0.081	99.50	30	161,344	0.789 (0.763-0.815)	0.661-0.941	0.085	99.4
Local hospital	26	18,653	0.818 (0.776-0.862)	0.633-1.000	0.015	91.70	72	74,308	0.824 (0.808-0.841)	0.712-0.955	0.073	87.1
Treatment background												
EVT[v]/IVT[w]	3	920	0.853 (0.796-0.914)	0.783-0.929	0.008	0.00	7	1338	0.817 (0.760-0.879)	0.681-0.981	0.068	77.5

XSL•FO
RenderX

Subgroup analysis	Training set						Validation set					
	n	Sample size	C-index (95% CI)	95% PI[a]	Tau	I^2 (%)	n	Sample size	C-index (95% CI)	95% PI	Tau	I^2 (%)
Not specific	47	189,544	0.821 (0.795-0.848)	0.666-1.000	0.103	99.10	95	234,314	0.811 (0.797-0.826)	0.691-0.953	0.081	98.4
Validation method												
Internal validation	—	—	—	—	—	—	92	230,828	0.808 (0.793-0.823)	0.687-0.950	0.081	98.5
External validation	—	—	—	—	—	—	10	4824	0.847 (0.808-0.887)	0.750-0.956	0.050	71.3
Overall	50	190,464	0.823 (0.798-0.848)	0.672-1.000	0.099	99.10	102	235,652	0.812 (0.798-0.826)	0.693-0.952	0.080	98.3

[a]PI: prediction interval.
[b]ANN: artificial neural network.
[c]DT: decision tree.
[d]Not applicable.
[e]EL: ensemble learning.
[f]Not available.
[g]GBM: gradient boosting machine.
[h]KNN: k-nearest neighbors.
[i]LASSO: least absolute shrinkage and selection operator.
[j]LightGBM: light gradient boosting machine.
[k]LR: logistic regression.
[l]NB: Naive Bayes.
[m]RF: random forest.
[n]RSF: random survival forest.
[o]SVM: support vector machine.
[p]XGBoost: extreme gradient boosting.
[q]DL: deep learning.
[r]COX: Cox proportional hazards.
[s]CatBoost: categorical boosting.
[t]AdaBoost: adaptive boosting.
[u]AIS: acute ischemic stroke.
[v]EVT: endovascular thrombectomy.
[w]IVT: intravenous thrombolysis.

XSL•FO
RenderX

Figure 4. Forest plot for meta-analysis of C-index of predictive models for out-of-hospital mortality in stroke in the training set. AIS: acute ischemic stroke; ANN: artificial neural network; CatBoost: categorical boosting; COX: Cox proportional hazards; DL: deep learning; DT: decision tree; EVT: endovascular thrombectomy; GBM: gradient boosting machine; IVT: intravenous thrombolysis; KNN: k-nearest neighbor; LightGBM: light gradient boosting machine; LR: logistic regression; NA: Not applicable; NB: Naive Bayes; PI: prediction interval; RF: random forest; SVM: support vector machine; XGBoost: extreme gradient boosting.

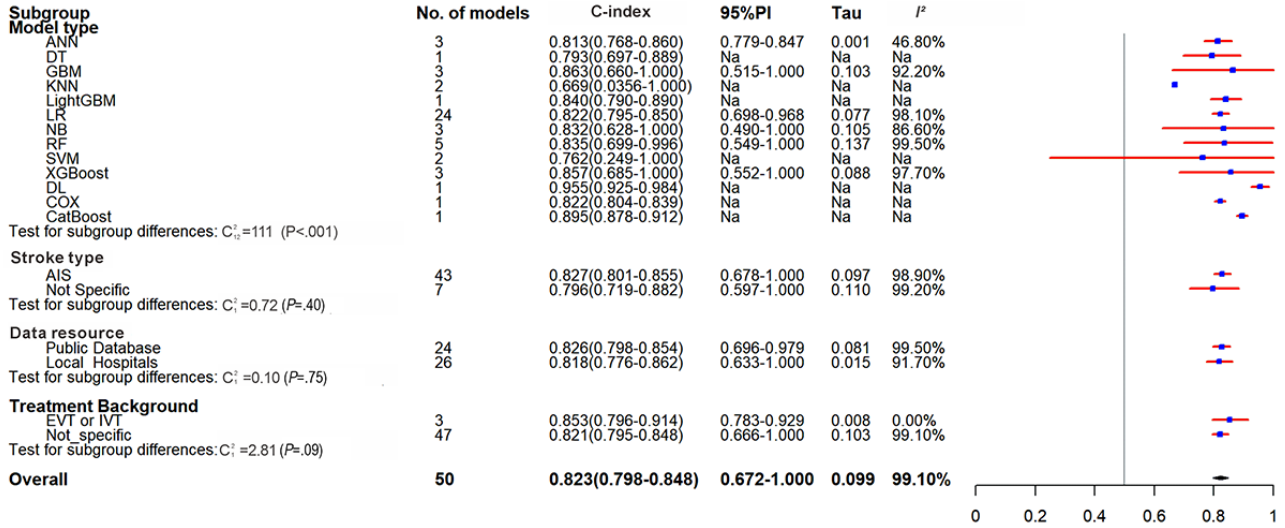

The pooled C-index was 0.827 (95% CI 0.801-0.855, 95% PI 0.678-1.000) in studies on AIS, 0.826 (95% CI 0.798-0.854, 95% PI 0.696-0.979) in studies based on public databases, 0.818 (95% CI 0.776-0.862, 95% PI 0.633-1.000) in studies based on hospital databases, and 0.853 (95% CI 0.796-0.914, 95% PI 0.783-0.929) in studies reporting acute-phase vascular recanalization. LR remained the most used model (24/50, 48%), with a C-index of 0.824 (95% CI 0.798-0.850) (Table 4 and Figure 4).

For out-of-hospital mortality, the pooled performance in external validation yielded a C-index of 0.847 (95% CI 0.808-0.887, 95% PI 0.750-0.956), with an SE of 0.71 (95% CI 0.55-0.82) and an SP of 0.76 (95% CI 0.74-0.78). The overall pooled analysis yielded a C-index of 0.812 (95% CI 0.798-0.826, 95% PI 0.693-0.952), with an SE of 0.68 (95% CI 0.63-0.72) and an SP of 0.82 (95% CI 0.78-0.85). Studies on AIS had a pooled C-index of 0.796 (95% CI 0.778-0.813, 95% PI 0.677-0.935), with SE and SP of 0.65 (95% CI 0.59-0.71) and 0.81 (95% CI 0.75-0.85). Studies based on public databases had a pooled C-index of 0.789 (95% CI 0.763-0.815, 95% PI 0.661-0.941), with SE and SP of 0.69 (95% CI 0.62-0.76) and 0.77 (95% CI 0.74-0.80). Studies based on hospital databases had a pooled C-index of 0.824 (95% CI 0.808-0.841, 95% PI 0.712-0.955), with SE and SP of 0.68 (95% CI 0.62-0.74) and 0.84 (95% CI 0.79-0.88). Studies reporting acute-phase vascular recanalization had a pooled C-index of 0.817 (95% CI 0.760-0.879, 95% PI 0.681-0.981), with SE and SP of 0.77 (95% CI 0.69-0.84) and 0.79 (95% CI 0.72-0.85). Internal validation subgroup yielded a C-index of 0.808 (95% CI 0.793-0.823, 95% PI 0.687-0.950), with SE and SP of 0.67 (95% CI 0.62-0.72) and 0.82 (95% CI 0.78-0.85), respectively. The subgroup results were all consistent with the overall trend. In the validation process for out-of-hospital mortality prediction models, those constructed with data sourced from private hospitals show superior overall performance when contrasted with models that rely on data from public datasets.

LR also demonstrated better performance, with a C-index, SE, and SP of 0.817 (95% CI 0.793-0.841, 95% PI 0.708-0.942), 0.74 (95% CI 0.69-0.79), and 0.79 (95% CI 0.73-0.84), respectively. DL models for out-of-hospital mortality were created in 2 studies, and they exhibited good discriminatory power (C-index: 0.873, 95% CI 0.799-0.955) but unsatisfactory diagnostic performance (SE 0.48, SP 0.76). (Tables 3 and 4, Figure 5, and Figures S7 and S8 in Multimedia Appendix 2).

XSL•FO
RenderX

Table 4. Results of meta-analysis of sensitivity and specificity of machine learning for predicting out-of-hospital mortality in stroke.

Subgroup analysis	Training set				Validation set			
	n	Sample size	SE[a] (95% CI)	SP[b] (95% CI)	n	Sample size	SE (95% CI)	SP (95% CI)
Model								
ANN[c]	3	4330	0.58 (0.55-0.61)	0.80 (0.73-0.87)	5	2860	0.61 (0.38-0.80)	0.85 (0.74-0.92)
DT[d]	—[e]	—	—		4	2217	0.62 (0.24-0.89)	0.72 (0.45-0.89)
EL[f]	—	—	—	—	3	17,053	0.76 (0.71-0.81)	0.80 (0.78-0.81)
GBM[g]	2	2859	0.55 (0.39-0.71)	0.84 (0.75-0.93)	5	2076	0.56 (0.38-0.72)	0.80 (0.52-0.94)
KNN[h]	2	2859	0.45 (0.45-0.46)	0.92 (0.92-0.93)	4	5205	0.53 (0.33-0.72)	0.80 (0.69-0.88)
AdaBoost[ii]	—	—	—	—	1	1968	0.80	0.81
LightGBM[j]	1	1388	0.53	0.83	2	2116	0.77 (0.67-0.86)	0.90 (0.82-0.97)
LR[k]	14	44,677	0.76 (0.69-0.81)	0.80 (0.76-0.83)	21	29,892	0.74 (0.69-0.79)	0.79 (0.73-0.84)
NB[l]	2	2859	0.51 (0.50-0.52)	0.87 (0.82-0.92)	6	6687	0.68 (0.55-0.79)	0.81 (0.67-0.89)
RF[m]	3	3313	0.68 (0.52-0.83)	0.81 (0.68-0.93)	11	13,978	0.64 (0.46-0.79)	0.88 (0.72-0.95)
RSF[n]	—	—	—	—	2	383	0.45 (0.21-0.69)	0.59 (0.42-0.76)
SVM[o]	2	2859	0.49 (0.41-0.57)	0.81 (0.75-0.86)	7	7225	0.67 (0.53-0.78)	0.87 (0.73-0.94)
XGBoost[p]	1	1471	0.33	0.98	8	9179	0.70 (0.48-0.86)	0.84 (0.70-0.91)
DL[q]	2	3821	0.92 (0.83-1.00)	0.99 (0.99-1)	2	1161	0.47 (0.47-0.48)	0.76
COX[r]	1	425	0.62	0.87	3	1101	0.70 (0.61-0.78)	0.74 (0.70-0.78)
CatBoost[s]	1	1471	0.68	0.90	3	2741	0.71 (0.52-0.89)	0.78 (0.77-0.79)
Stroke type								
AIS[t]	29	62,174	0.64 (0.56-0.70)	0.89 (0.84-0.93)	49	33,411	0.65 (0.59-0.71)	0.81 (0.75-0.85)
Not specific	5	10,158	0.78 (0.74-0.80)	0.75 (0.74-0.76)	38	72,431	0.70 (0.64-0.76)	0.83 (0.78-0.87)
Data resource								
Public database	17	55,453	0.64 (0.57-0.71)	0.89 (0.84-0.93)	30	54,088	0.69 (0.62-0.76)	0.77 (0.74-0.80)
Local hospital	17	16,879	0.69 (0.58-0.78)	0.87 (0.77-0.93)	57	51,754	0.68 (0.62-0.74)	0.84 (0.79-0.88)
Treatment background								
EVT[u]/IVT[v]	—	—	—	—	5	1044	0.77 (0.69-0.84)	0.79 (0.72-0.85)
Not specific	—	—	—	—	82	104,798	0.67 (0.62-0.72)	0.82 (0.78-0.85)
Validation method								
Internal validation	—	—	—	—	82	103,116	0.67 (0.62-0.72)	0.82 (0.78-0.85)
External validation	—	—	—	—	5	2726	0.71 (0.55-0.82)	0.76 (0.74-0.78)
Overall	34	72,332	0.66 (0.60-0.72)	0.88 (0.83-0.91)	87	105,842	0.68 (0.63- 0.72)	0.82 (0.78-0.85)

[a]SE: sensitivity.
[b]SP: specificity.
[c]ANN: artificial neural network.
[d]DT: decision tree
[e]Not available.
[f]EL: ensemble learning.
[g]GBM: gradient boosting machine.
[h]KNN: k-nearest neighbor.
[i]AdaBoost: adaptive boosting.

XSL•FO
RenderX

[j]LightGBM: light gradient boosting machine.

[k]LR: logistic regression.

[l]NB: Naive Bayes.

[m]RF: random forest.

[n]RSF: random survival forest.

[o]SVM: support vector machine.

[p]XGBoost: extreme gradient boosting.

[q]DL: deep learning.

[r]COX: Cox proportional hazards.

[s]CatBoost: categorical boosting.

[t]AIS: acute ischemic stroke.

[u]EVT: endovascular thrombectomy.

[v]IVT: intravenous thrombolysis.

Figure 5. Forest plot for meta-analysis of C-index of predictive models for out-of-hospital mortality in stroke in the validation set. AdaBoost: adaptive boosting; AIS: acute ischemic stroke; ANN: artificial neural network;CatBoost: categorical boosting; COX: Cox proportional hazards; DL: deep learning; DT: decision tree; EL: ensemble learning; EVT: endovascular thrombectomy; GBM: gradient boosting machine; IVT: intravenous thrombolysis; KNN: k-nearest neighbor; LightGBM: light gradient boosting machine; LR: logistic regression; NA:Not applicable; NB: Naive Bayes; PI: prediction interval; RF: random forest; RSF: random survival forest; SVM: support vector machine; XGBoost: extreme gradient boosting.

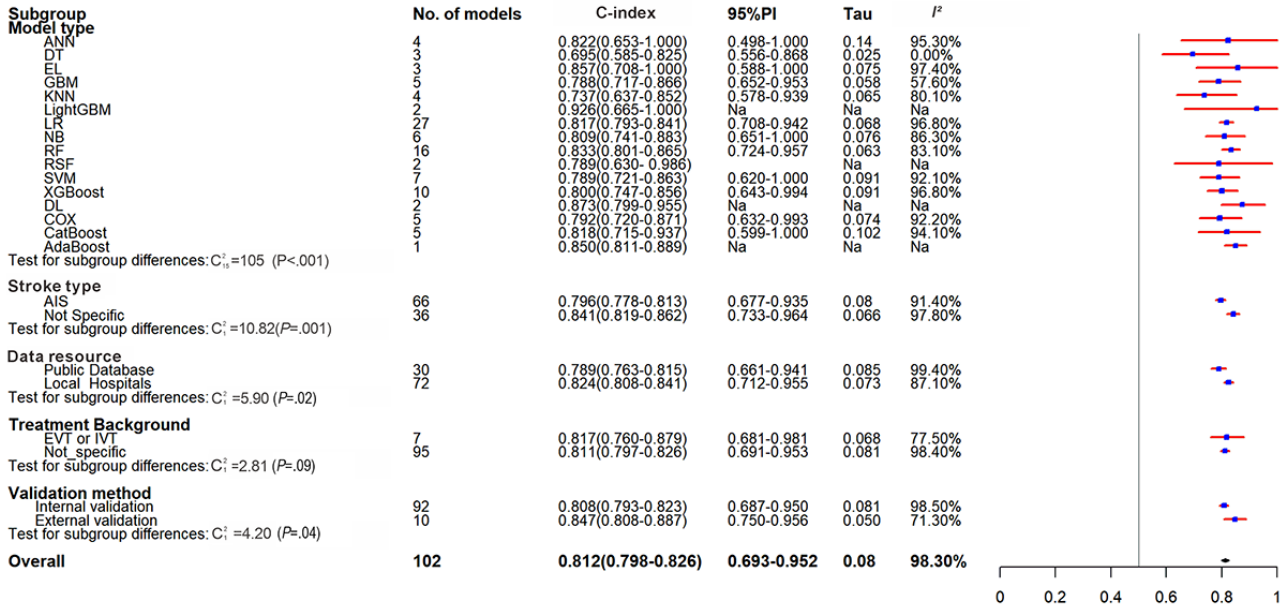

Surprisingly, only a few studies adopted survival analysis modeling (only 1 COX [Cox Proportional Hazards] model in the training set, and only 5 COX models and 2 random survival forest models in the validation set). Furthermore, metaregression revealed a gradual decline in the predictive performance of the LR models and a sustained performance of the RF and XGBoost models (Figure 6A and B and Figures S9-S11 in Multimedia Appendix 2).

XSL•FO
RenderX

Figure 6. Metaregression bubble plot of follow-up time in the predictive models for (A) training set and (B) validation set.

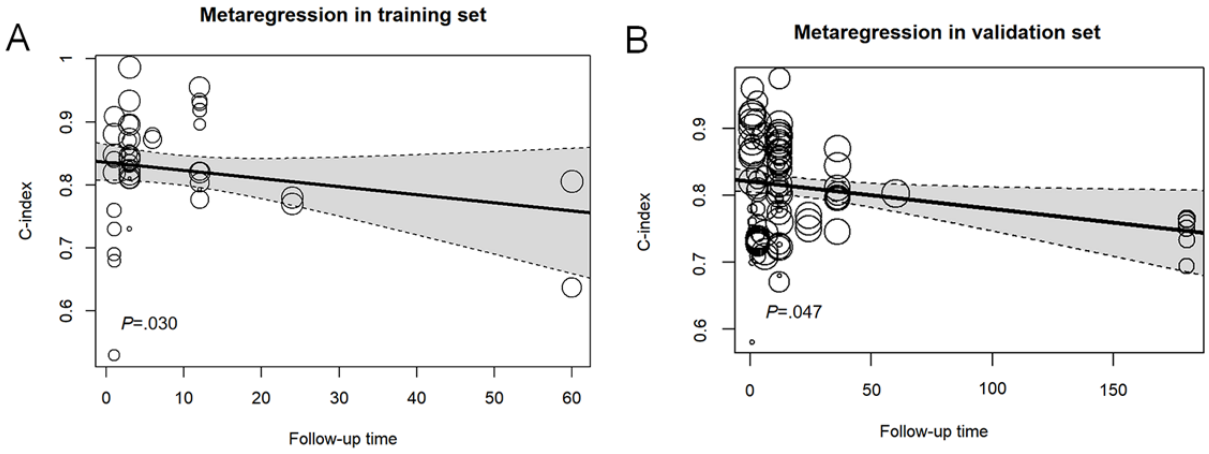

Modeling Variables

Of the 155 clinical variables included, high-frequency variables (>5 occurrences) are shown in Figure 7, with age (n=36), National Institutes of Health Stroke Scale (NIHSS) score (n=16), and hypertension (n=16) as the top 3. These variables were distributed across comorbidities (smoking, previous stroke history, cancer, and in-hospital infections), acute severity of illness (NIHSS, Glasgow Coma Scale, and premorbid modified Rankin Scale [mRS] scores), and admission characteristics (time of admission). Few studies involved biomarkers such as platelet-to-neutrophil ratio (PNR) (n=1) and neutrophil-percentage-to-albumin ratio (NPAR) (n=1).

Figure 7. Machine learning models established by predictors for poststroke mortality. GCS: Glasgow Coma Scale; mRS: modified Rankin Scale; NIHSS: National Institutes of Health Stroke Scale.

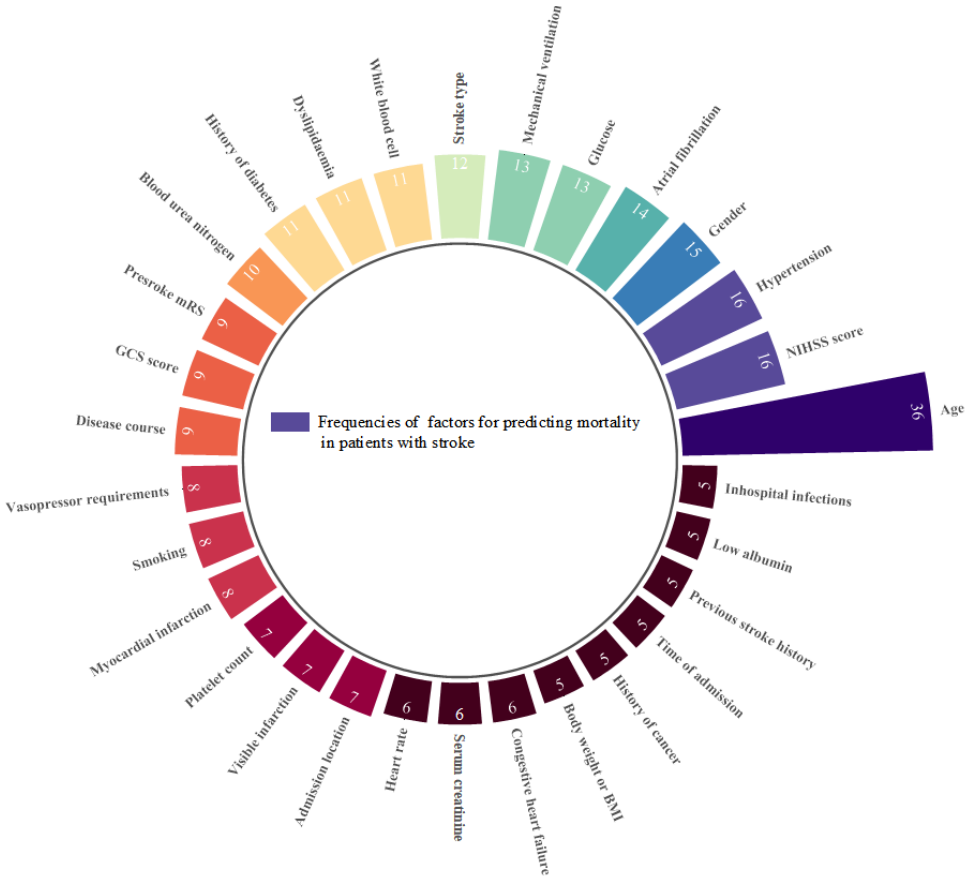

XSL•FO
RenderX

Small-Study Effect Assessment

As revealed by the funnel plot asymmetry (with fewer studies skewing toward lower performance values) and significant results of Egger tests (P<.001), significant small study effects were present in the included studies on out-of-hospital mortality (Figure S13 in Multimedia Appendix 2). Meanwhile, no small study effects were present in the included studies on in-hospital mortality and LR and RF models (Figures S12-S15 in Multimedia Appendix 2).

SE Analysis

SE analyses were performed after controlling for cohort overlap, yielding pooled C-indexes based on selected studies: For in-hospital mortality, Ouyang et al [50] (MIMIC-III), Cummins et al [48] (MIMIC-IV), Huang et al [58] (eICU-CRD), and Wang et al [43] (SSNAP) were selected; for out-of-hospital mortality, Li et al [81] (MIMIC-III) and Zhu et al [96] (MIMIC-IV) were selected as representatives. Thus, the pooled C-indexes for in-hospital and out-of-hospital mortality prediction models were 0.801 (95% CI 0.777-0.828, 95% PI 0.682-0.992) and 0.825 (95% CI 0.811-0.840, 95% PI 0.711-0.957), respectively. Similar to the primary analysis, these results indicated that the main findings of this meta-analysis were robust to cohort overlap. From further single original studies, 16 and 32 optimal models for predicting in-hospital and out-of-hospital mortality were retained, respectively, with pooled C-indexes of 0.852 (95% CI 0.811-0.895, 95% PI 0.699-1.000) and 0.842 (95% CI 0.817-0.868, 95% PI 0.720-0.986), respectively (Figures S16 [27,43,45,48,50,51,58,67,68,70,76,83,99,101-103] and S17 [29, 40, 42, 44, 46, 49, 53-56, 59, 60, 65, 67, 69, 71, 72, 74, 75, 77-82, 86, 89, 91, 93, 95, 96, 100] in Multimedia Appendix 2). Besides, 5 and 6 external validation studies were included for in-hospital and out-of-hospital mortality, respectively. They yielded pooled C-indexes of 0.785 (95% CI 0.698-0.884, 95% PI 0.594-1.000) and 0.853 (95% CI 0.781-0.931, 95% PI 0.708-1.000), respectively (Figures S18 [50,51,53,63,99] and S19 [29,54,56,71,74,79] in Multimedia Appendix 2).

Discussion

Main Findings

To our knowledge, this is the first meta-analysis to evaluate ML as a promising tool for predicting stroke mortality. Different from previous reviews, we conducted subgroup analyses across multiple dimensions, including the model type, patient source, and time. Sixty-eight studies were included in this study, involving numerous ML models for predicting poststroke mortality. Despite methodological concerns of these studies, the most rigorous evidence on generalizability derived from the externally validated subset. The results demonstrated favorable discriminatory power, with pooled C-indices of approximately 0.73 for in-hospital and 0.85 for out-of-hospital mortality, alongside maintaining balanced SE and SP for both in-hospital and out-of-hospital mortality.

The predictive performance of LR models for long-term out-of-hospital mortality might gradually decline, while RF and XGBoost achieved a robust and sustained performance. Subgroup analyses suggested that the models for AIS and the presence or absence of acute-phase vascular recanalization had no great difference from the overall model. Accounting for approximately 50% of the total, LR had slightly inferior performance to other ML methods. Besides, the PI for the C-index for in-hospital or out-of-hospital mortality did not cover the reference line, suggesting that the LR model consistently outperforms random prediction across clinical settings and possesses translational potential. Notably, the PIs were broader in some subgroups due to the limited number of included studies, so further validation is needed in the future. To sum up, ML holds significant value in predicting both in-hospital and out-of-hospital stroke mortality. These findings offer an evidence basis for developing precise, dynamic clinical risk prediction tools in the future.

Comparison With Previous Reviews

Previous reviews have also investigated ML for early judgment of stroke prognosis. Yang et al [104] conducted a meta-analysis of AI models for predicting ischemic stroke outcomes by binary prediction of 90-day mRS, with 7 studies included. They found a pooled area under the curve (AUC) up to 0.872 of the fixed-effects model, confirming the value of AI in predicting a good prognosis of stroke (mRS ≤2). However, they ignored the assessment of mortality (mRS=6). Schwartz et al [105] and Wang et al [106] included 25 and 7 original studies, respectively, and argued that ML achieves better predictive effects on short- and long-term stroke mortality. However, they performed only a qualitative assessment, thus quantitative evidence was lacking. In addition, available studies on ML for predicting stroke mortality do not classify prediction models and consider the influence of time factors on the model accuracy, generating concerns about heterogeneity in population and methodology. Therefore, we expanded the model performance beyond previous systematic reviews by quantitatively synthesizing evidence from external validations to evaluate model generalizability, in addition to assessing the goodness-of-fit in training sets.

Modeling Variables

Modeling variables (clinical features, radiomics features, genetic information, and microbiomics) also constituted a challenge to this study. Due to the critical conditions of severe stroke, some examinations such as magnetic resonance imaging are difficult and consume vast human and material resources. Only 2 studies used radiomics [28,54], which relied primarily on interpretable clinical features. We synthesized interpretable clinical features and confirmed that age, NIHSS score, and hypertension were the most used modeling variables, consistent with previous reviews [21,105]. Meanwhile, some studies also focused on specific laboratory indicators. NPAR combines inflammation with nutrient metabolic status, which is more reflective of the body's acute stress response and overall reserve status [107]. Therefore, NPAR possesses greater predictive power. Chunjuan et al [62] created a nomogram using NPAR and the systematic inflammatory response index for predicting stroke mortality, with an AUC of 0.637. Chung et al [108] proposed that NPAR in combination with simple clinical indicators demonstrates excellent predictive performance for severe illness, with an AUC (0.929) comparable with the Sequential Organ Failure Assessment score, which enables better clinical judgment and

XSL•FO
RenderX

interventions. Geng et al [93] adopted PNR to reflect thrombotic and inflammatory status, in which AUC increases from 0.736 (PNR alone) to 0.888 (PNR incorporating age and NIHSS scores). Therefore, the association of changes in some variables with the mortality risk can be observed in the future.

For intensive care unit patients, simple biomarkers and clinical parameters can be combined with laboratory tests to predict the disease severity. In addition, radiomics can be further introduced to reflect individual neuropathologic features and predict prognosis [109]. Besides, models from individual hospitals generally yielded higher C-indexes. The possible reason is as follows: electronic health records at medical centers often provide more detailed, abundant, and clinically relevant predictor variables than public databases covering broad populations, enabling models to capture more complex prediction patterns [110,111].

Model Selection

Besides modeling variables, the model types (well-interpretable LR, COX, competing risk, and DT) should also be taken into account. These models can quantify the association of variables with risk and thus assist in determining the specific scoring tools, but their accuracy is unsatisfactory. In addition, traditional ML displays high accuracy [112], but its interpretability is challenging. Although ML models interpret "black-box models" using various Shapley additive explanations (SHAP) diagrams at the local and global levels in some studies, the SHAP value under its computing paradigm provides the mean contribution of features in the global scope rather than local logical explanations for specific predictions. Due to this essential attribute, SHAP is difficult to meet the requirements for causality and fidelity in high-reliability scenarios such as clinical diagnosis. Consequently, SHAP values often fail to accurately capture the actual impact of features and may generate misleading feature importance rankings, which restricts their credibility and applicability in high-risk decision-making [113,114]. In the subgroup analysis, LR was dominant in the included studies, with favorable SE and C-index, suggesting that routine interpretable models can early predict stroke mortality. Because of its theoretical simplicity and ease of implementation, LR is widely favored in clinical prediction models [115]. In particular, datasets may be relatively simple and exhibit linear trends in the distribution when data predicting stroke mortality are primarily from clinical history and testing. Therefore, the performance of LR is often comparable with that of complex ML models [112,116]. Meanwhile, LR can quantify the association between various influencing factors and mortality risk and further create a nomogram to better visualize results and thus develop convenient online prediction tools [117]. Thus, despite growing attention to ML, LR remains dominant in data analysis.

Besides predictive accuracy, model interpretability must also be considered in current clinical applications. Generally, highly interpretable models, such as LR, COX, and DT, exhibit limitations in predictive accuracy. In contrast, less interpretable models, such as RF, support vector machine, neural networks, and DL, possess superior accuracy but low interpretability due to their "black-box" property [118,119]. Consequently, less interpretable models face challenges in clinical practice, including limited use by health care providers and difficulty in patient management [120]. Therefore, highly interpretable models should be prioritized. If their accuracy fails to meet clinical requirements, model interpretation techniques (eg, Grad-CAM and LIME) should be introduced to enhance understanding and trust in complex model outputs while maintaining predictive performance [121]. Such interpretable methods are crucial for advancing AI integration and application in clinical settings [122,123]. Besides, DL has high efficiency for image processing and excellent time effect capture ability when processing time series data [124,125], so it has the potential for judging stroke site and time. However, DL is less applied to the early prediction of poststroke mortality, possibly attributed to the difficulty in its interpretability and high requirements for techniques [126]. In the future, DL using multimodal data may be more accurate in prognostic prediction of stroke.

Our meta-analysis revealed a gradual decline in the predictive performance of the overall model and LR model for long-term mortality, consistent with previous reviews [127,128]. The reason may be related to variations in clinical background, alterations in body status, and introduction of time-varying confounders. Notably, RF achieved a robust performance in long-term mortality prediction, with high accuracy, suggesting that RF may be an effective ML model for predicting stroke mortality. Given rehabilitation progresses postdischarge, condition and complications of patients with stroke often present dynamic evolution, so the reliability of AI and ML models relying solely on information at a single time point during hospitalization is restricted in prediction and assessment. Therefore, real-time collection and analysis of patient data are crucial for accurately identifying individual phenotypic changes [129]. Establishing a collaborative real-time data network has become a key foundation for long-term, dynamic clinical decision-making [19]. Moreover, for predicting out-of-hospital mortality, survival analysis modeling was adopted in very few studies. The clinical significance greatly varied between survival analysis modeling and nonsurvival analysis modeling, while the 2 had a minor difference in short-term follow-up; dynamic mortality risk prediction was often a focus in clinical practice, which could be better achieved by survival analysis modeling, with a statistical basis for decision-making [130]. In contrast, nonsurvival analysis modeling was difficult to achieve dynamic mortality risk prediction. Although LR dominates in the analysis currently, dynamic prediction models should be more considered for long-term mortality in the future to develop predictive strategies in real time.

Prospects and Outlook

The potential of ML for predicting stroke mortality was confirmed in this study, but several technical challenges remain. The population is unbalanced data and the dataset balance is a serious challenge [131]. This meta-analysis found a mortality of about 13% from public databases and 11% from hospital databases. However, the impact of data imbalance was considered in few studies. We synthesized SE and SP of models of unbalanced data and found that despite good performance in predicting negative events (ie, survival), these models were less

XSL•FO
RenderX

accurate in predicting positive events (ie, mortality) and were difficult to produce clinical benefits [132]. Therefore, the impact of data imbalance should be taken into account in future modeling.

Another challenge arises from the limited availability and divergent performance patterns of external validation evidence. For in-hospital mortality, the performance in external validation was more conservative than the overall performance and performance in internal validation, probably due to optimistic bias in high-dimensional prognostic settings and simple internal validations [133,134], However, the external validation set remained clinically meaningful and represented an appropriate basis for assessing model generalizability. Conversely, regarding out-of-hospital mortality, the external validation set had more satisfactory C-index than internal validation. This phenomenon was observed in all the 4 included studies [29,54,71,79], which might be related to the fact that patients had more severe conditions, higher mortality, and greater case-mix variation in external validation, thus enhancing the model's discriminatory power [29,135]. However, due to the limited number and potential representativeness concerns [136], the generalizability of this finding requires cautious interpretation.

This meta-analysis revealed significant small study effects in out-of-hospital mortality, with small sample studies tending to report lower C-index values. This may also be attributed to limited standardized data, introducing data noise that attenuated the model's apparent performance [137]. Furthermore, the optimal performance still generally fell below the average of large sample studies. Small sample studies also possessed a narrower case-mix (ie, more homogeneous population), and established evidence suggests that the C-index is naturally lower in homogeneous populations [39]. The lower pooled C-index observed in small sample studies likely reflected clinical heterogeneity or implementation quality, rather than publication bias or outcome reporting bias [138]. Consequently, the inclusion of these studies probably yielded a conservative underestimate of the true performance of robust models developed with sufficient resources. Also, our findings of subgroup analyses should be interpreted with caution. Since algorithm types were not randomly assigned across studies, these differences were observed and might be confounded by other study-level factors rather than the algorithms themselves. Therefore, these results represented associations rather than definitive causal evidence [139].

Furthermore, the translation of these models into clinical practice requires rigorous validation. The vast majority of included studies did not distinguish independent test sets and lacked valid external validation. As a result, the system suitability of the AI-based clinical decision support system was limited in clinical workflows [21]. Therefore, future research should conduct external validation of algorithms by program standardization in larger prospective cohorts and compare multiple existing models by external validation to determine the most useful one [140]. Moreover, the models can be compared with relevant clinicians' judgment, such as providing clinicians (including intensivists and neurologists) with routine clinical information. Additional metrics can also be assessed, such as family acceptance of AI results, resource utilization and planning, and adjustments to subsequent medical plans [141]. Additionally, these algorithms still need improvement to identify mortality in all at-risk populations (low-, medium-, and high-risk populations), and they should be trained by a large number of outlier samples or oversampling techniques, such as random or boundary oversampling or adaptive synthetic sampling. These approaches will facilitate the critical deployment of AI in stroke mortality prediction.

Advantages and Limitations

Advantages

This meta-analysis first quantitatively reviewed the value of ML in predicting in-hospital and out-of-hospital mortality in stroke. The findings are expected to offer an accurate framework for poststroke mortality risk assessment and clinical decision-making.

Limitations

Some limitations are worth noting. First, this meta-analysis exhibited great heterogeneity, which can be attributed to the variations in algorithms and study populations. Few studies reporting acute-phase vascular recanalization were included, and most of the studies did not classify the treatment background in detail, which may affect the overall performance assessment. Also, data mostly came from available cohort studies or conventional registries, and some were single-center studies, introducing some bias and restricting the accuracy. Therefore, prospective multicenter studies with detailed subgroup analyses (including treatment backgrounds) are needed. Second, outcome metrics were missing in a small number of studies, so reasonable estimation was implemented, but the model's predictive accuracy estimates were still affected. Third, internal validation was used in the majority of the studies and external validation in only 20% of the studies, weakening the model's extrapolation capability. Fourth, multiple studies included were from a small number of large public databases, potentially leading to cohort overlap. Although the limited impact of potential cohort overlap on the pooled effect was confirmed by the SE analysis, we recognized that this method could not fully eliminate residual overfitting bias, given that such bias may arise from systematic similarities in patient populations and clinical protocols across these databases [142]. This concealed overlap could still inflate the summary effect estimates and complicate the interpretation of heterogeneity. Therefore, the extrapolation of our conclusions to more heterogeneous clinical settings warrants caution.

Conclusions

ML demonstrates promising performance in stroke mortality prediction, and some ML models are more valuable in predicting long-term mortality. However, more efforts are needed before application in clinical practice. For example, we should focus on population SP, search for more specific biomarkers, introduce more multimodal data, optimize model balance, and increase external validations. Furthermore, ML models should be used to develop prediction tools for stroke mortality risk applicable in clinical settings to advance their translation into clinical practice. In this way, data-based doctor-patient shared decision-making can be facilitated, and more efficient resource

XSL•FO
RenderX

allocation can be achieved in health care institutions, ultimately ameliorating patient outcomes.

Acknowledgments

The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript.

Funding

This study was funded by the Regional Innovation Cooperation Project of the Science and Technology Department of Sichuan Province (2024YFHZ0142), Sichuan Science and Technology Program (no. 2023YFQ0109), and Health Commission of Sichuan Province Medical Science and Technology Program (2023-109 and 24LCYJPT08), as well as the National Key R&D Program of China (no. 2022YFC2503805). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

All data generated or analyzed during this study are included in this published paper and its supplementary information files.

Authors' Contributions

All authors contributed to the study conception and design. YC and ZO participated in writing – original draft preparation, and writing – review and editing. YD contributed to conceptualization. YC and YY contributed to methodology. YC, ZO, YD, AL, XL, YY, and QY conducted the formal analysis and investigation. MZ and XW participated in supervision. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 97 KB - jmir_v28i1e83821_app1.pdf]

Multimedia Appendix 2
Details of the search strategy.
[DOCX File , 3381 KB - jmir_v28i1e83821_app2.docx]

Multimedia Appendix 3
Results of Prediction model Risk of Bias Assessment Tool assessment of risk of bias in included studies.
[PNG File , 57 KB - jmir_v28i1e83821_app3.png]

References

1. Hilkens NA, Casolla B, Leung TW, de Leeuw FE. Stroke. Lancet 2024;403(10446):2820-2836. [doi: 10.1016/S0140-6736(24)00642-1] [Medline: 38759664]
2. Feigin VL, Brainin M, Norrving B, Martins SO, Pandian J, Lindsay P, et al. World stroke organization: Global stroke fact sheet 2025. Int J Stroke 2025;20(2):132-144 [FREE Full text] [doi: 10.1177/17474930241308142] [Medline: 39635884]
3. Pendlebury ST, Rothwell PM. Prevalence, incidence, and factors associated with pre-stroke and post-stroke dementia: a systematic review and meta-analysis. Lancet Neurol 2009;8(11):1006-1018. [doi: 10.1016/S1474-4422(09)70236-4] [Medline: 19782001]
4. Feigin VL, Owolabi MO. Pragmatic solutions to reduce the global burden of stroke: a world stroke organization-lancet neurology commission. Lancet Neurol 2023;22(12):1160-1206. [doi: 10.1016/S1474-4422(23)00277-6] [Medline: 37827183]
5. Global, regional, and national burden of stroke and its risk factors, 1990-2021. A systematic analysis for the global burden of disease study 2021. Lancet Neurol 2024;23(10):973-1003 [FREE Full text] [doi: 10.1016/S1474-4422(24)00369-7] [Medline: 39304265]
6. Gu HQ, Yang X, Wang CJ, Zhao XQ, Wang YL, Liu LP, et al. Clinical characteristics, management, and in-hospital outcomes in patients with stroke or transient ischemic attack in china. JAMA Netw Open 2021;4(8):e2120745 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.20745] [Medline: 34387677]

XSL•FO
RenderX

7. Wang M, Wang CJ, Gu HQ, Meng X, Jiang Y, Yang X, et al. Sex differences in short-term and long-term outcomes among patients with acute ischemic stroke in china. Stroke 2022;53(7):2268-2275 [FREE Full text] [doi: 10.1161/STROKEAHA.121.037121] [Medline: 35130717]
8. Tu WJ, Chao BH, Ma L, Yan F, Cao L, Qiu H, et al. Case-fatality, disability and recurrence rates after first-ever stroke: a study from bigdata observatory platform for stroke of China. Brain Res Bull 2021;175:130-135. [doi: 10.1016/j.brainresbull.2021.07.020] [Medline: 34329730]
9. Goyal M, Menon BK, van Zwam WH, Dippel DWJ, Mitchell PJ, Demchuk AM, HERMES Collaborators. Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. Lancet 2016;387(10029):1723-1731. [doi: 10.1016/S0140-6736(16)00163-X] [Medline: 26898852]
10. Tu WJ, Wang LD. China stroke surveillance report 2021. Mil Med Res 2023;10(1):33 [FREE Full text] [doi: 10.1186/s40779-023-00463-x] [Medline: 37468952]
11. van Asch CJ, Luitse MJ, Rinkel GJ, van der Tweel I, Algra A, Klijn CJ. Incidence, case fatality, and functional outcome of intracerebral haemorrhage over time, according to age, sex, and ethnic origin: a systematic review and meta-analysis. Lancet Neurol 2010;9(2):167-176. [doi: 10.1016/S1474-4422(09)70340-0] [Medline: 20056489]
12. Wu S, Wu B, Liu M, Chen Z, Wang W, Anderson CS, et al. Stroke in China: advances and challenges in epidemiology, prevention, and management. Lancet Neurol 2019;18(4):394-405. [doi: 10.1016/S1474-4422(18)30500-3] [Medline: 30878104]
13. Montellano FA, Ungethüm K, Ramiro L, Nacu A, Hellwig S, Fluri F, et al. Role of blood-based biomarkers in ischemic stroke prognosis. Stroke 2021;52(2):543-551. [doi: 10.1161/strokeaha.120.029232]
14. Kremers F, Venema E, Duvekot M, Yo L, Bokkers R, Lycklama À Nijeholt G, et al. Outcome prediction models for endovascular treatment of ischemic stroke: Systematic review and external validation. Stroke 2022;53(3):825-836 [FREE Full text] [doi: 10.1161/STROKEAHA.120.033445] [Medline: 34732070]
15. Xie P, Hu Y, Li J, Ma Y, Xiao J. Unlocking the potential of real-time ICU mortality prediction: redefining risk assessment with continuous data recovery. NPJ Digit Med 2025;8(1):733 [FREE Full text] [doi: 10.1038/s41746-025-02114-y] [Medline: 41315868]
16. Solomon OM, Gu Y, Wei Y. Risk factors for enteral nutrition feeding intolerance in stroke patients: a scoping review. Nurs Crit Care 2025;30(3):e13183. [doi: 10.1111/nicc.13183] [Medline: 39428862]
17. Yang Y, Huang X, Wang Y, Leng L, Xu J, Feng L, et al. The impact of triglyceride-glucose index on ischemic stroke: a systematic review and meta-analysis. Cardiovasc Diabetol 2023;22(1):2 [FREE Full text] [doi: 10.1186/s12933-022-01732-0] [Medline: 36609319]
18. Efstathiou SP, Tsioulos DI, Tsiakou AG, Gratsias YE, Pefanis AV, Mountokalakis TD. Plasma adiponectin levels and five-year survival after first-ever ischemic stroke. Stroke 2005;36(9):1915-1919. [doi: 10.1161/01.STR.0000177874.29849.f0] [Medline: 16109902]
19. Cecconi M, Greco M, Shickel B, Vincent JL, Bihorac A. Artificial intelligence in acute medicine: a call to action. Crit Care 2024;28(1):258. [doi: 10.1186/s13054-024-05034-7] [Medline: 39075468]
20. Bonkhoff AK, Grefkes C. Precision medicine in stroke: towards personalized outcome predictions using artificial intelligence. Brain 2022;145(2):457-475 [FREE Full text] [doi: 10.1093/brain/awab439] [Medline: 34918041]
21. Akay EMZ, Hilbert A, Carlisle BG, Madai VI, Mutke MA, Frey D. Artificial intelligence for clinical decision support in acute ischemic stroke: a systematic review. Stroke 2023;54(6):1505-1516 [FREE Full text] [doi: 10.1161/STROKEAHA.122.041442] [Medline: 37216446]
22. Zhang J, Wu J, Zhou XS, Shi F, Shen D. Recent advancements in artificial intelligence for breast cancer: image augmentation, segmentation, diagnosis, and prognosis approaches. Semin Cancer Biol 2023;96:11-25. [doi: 10.1016/j.semcancer.2023.09.001] [Medline: 37704183]
23. Borchert RJ, Azevedo T, Badhwar A, Bernal J, Betts M, Bruffaerts R, et al. Artificial intelligence for diagnostic and prognostic neuroimaging in dementia: a systematic review. Alzheimers Dement 2023;19(12):5885-5904. [doi: 10.1002/alz.13412] [Medline: 37563912]
24. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023;23(1):689 [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
25. Radley NM, Soh I, Saad AM, Wickramarachchi M, Dawson A, Hin JNC, et al. Risk of bias assessment of post-stroke mortality machine learning predictive models: systematic review. J Stroke Cerebrovasc Dis 2025;34(6):108291 [FREE Full text] [doi: 10.1016/j.jstrokecerebrovasdis.2025.108291] [Medline: 40089217]
26. Liu W, Ma W, Bai N, Li C, Liu K, Yang J, et al. Identification of key predictors of hospital mortality in critically ill patients with embolic stroke using machine learning. Biosci Rep 2022;42(9):BSR20220995 [FREE Full text] [doi: 10.1042/BSR20220995] [Medline: 35993194]
27. Kurtz P, Peres IT, Soares M, Salluh JIF, Bozza FA. Hospital length of stay and 30-day mortality prediction in stroke: a machine learning analysis of 17,000 ICU admissions in Brazil. Neurocrit Care 2022;37(Suppl 2):313-321. [doi: 10.1007/s12028-022-01486-3] [Medline: 35381967]

XSL•FO
RenderX

28. Kniep HC, Elsayed S, Nawabi J, Broocks G, Meyer L, Bechstein M, et al. Imaging-based outcome prediction in posterior circulation stroke. J Neurol 2022;269(7):3800-3809 [FREE Full text] [doi: 10.1007/s00415-022-11010-4] [Medline: 35257203]
29. Bray BD, Campbell J, Cloud GC, Hoffman A, James M, Tyrrell PJ, et al. Derivation and external validation of a case mix model for the standardized reporting of 30-day stroke mortality rates. Stroke 2014;45(11):3374-3380. [doi: 10.1161/STROKEAHA.114.006451] [Medline: 25293667]
30. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
31. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
32. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019;170(1):W1-W33 [FREE Full text] [doi: 10.7326/M18-1377] [Medline: 30596876]
33. Debray TP, Damen JA, Riley RD, Snell K, Reitsma JB, Hooft L, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. Stat Methods Med Res 2019;28(9):2768-2786 [FREE Full text] [doi: 10.1177/0962280218785504] [Medline: 30032705]
34. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods 2010;1(2):97-111 [FREE Full text] [doi: 10.1002/jrsm.12] [Medline: 26061376]
35. Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ 2017;356:i6460 [FREE Full text] [doi: 10.1136/bmj.i6460] [Medline: 28057641]
36. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]
37. Röver C, Knapp G, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. BMC Med Res Methodol 2015;15:99 [FREE Full text] [doi: 10.1186/s12874-015-0091-1] [Medline: 26573817]
38. Partlett C, Riley RD. Random effects meta-analysis: coverage performance of 95% confidence and prediction intervals following REML estimation. Stat Med 2017;36(2):301-317 [FREE Full text] [doi: 10.1002/sim.7140] [Medline: 27714841]
39. Damen JAA, Moons KGM, van Smeden M, Hooft L. How to conduct a systematic review and meta-analysis of prognostic model studies. Clin Microbiol Infect 2023;29(4):434-440 [FREE Full text] [doi: 10.1016/j.cmi.2022.07.019] [Medline: 35934199]
40. Norouzi S, Ahmadi S, Alinia S, Farzipoor F, Shahsavari A, Hajizadeh E, et al. Machine learning predictive models for survival in patients with brain stroke. Health Promot Perspect 2025;15(1):63-72. [doi: 10.34172/hpp.025.43635] [Medline: 40453683]
41. Zhou Z, Chen B, Mei Z, Chen W, Cao W, Xu E, et al. Machine learning-based scoring model for predicting mortality in ICU-admitted ischemic stroke patients with moderate to severe consciousness disorders. Front Neurol 2025;16:1534961 [FREE Full text] [doi: 10.3389/fneur.2025.1534961] [Medline: 40170899]
42. Xu L, Li C, Zhang J, Guan C, Zhao L, Shen X, et al. Personalized prediction of mortality in patients with acute ischemic stroke using explainable artificial intelligence. Eur J Med Res 2024;29(1):341 [FREE Full text] [doi: 10.1186/s40001-024-01940-2] [Medline: 38902792]
43. Wang W, Rudd AG, Wang Y, Curcin V, Wolfe CD, Peek N, et al. Risk prediction of 30-day mortality after stroke using machine learning: a nationwide registry-based cohort study. BMC Neurol 2022;22(1):195 [FREE Full text] [doi: 10.1186/s12883-022-02722-1] [Medline: 35624434]
44. Wang K, Gu L, Liu W, Xu C, Yin C, Liu H, et al. The predictors of death within 1 year in acute ischemic stroke patients based on machine learning. Front Neurol 2023;14:1092534 [FREE Full text] [doi: 10.3389/fneur.2023.1092534] [Medline: 36908612]
45. Someeh N, Mirfeizi M, Asghari-Jafarabadi M, Alinia S, Farzipoor F, Shamshirgaran SM. Predicting mortality in brain stroke patients using neural networks: outcomes analysis in a longitudinal study. Sci Rep 2023;13(1):18530 [FREE Full text] [doi: 10.1038/s41598-023-45877-8] [Medline: 37898678]
46. Scrutinio D, Ricciardi C, Donisi L, Losavio E, Battista P, Guida P, et al. Machine learning to predict mortality after rehabilitation among patients with severe stroke. Sci Rep 2020;10(1):20127 [FREE Full text] [doi: 10.1038/s41598-020-77243-3] [Medline: 33208913]
47. Petrović I, Broggi S, Killer-Oberpfalzer M, Pfaff JAR, Griessenauer CJ, Milosavljević I, et al. Predictors of in-hospital mortality after thrombectomy in anterior circulation large vessel occlusion: a retrospective, machine learning study. Diagnostics (Basel) 2024;14(14):1531 [FREE Full text] [doi: 10.3390/diagnostics14141531] [Medline: 39061668]
48. Cummins JA, Gerber BS, Fukunaga MI, Henninger N, Kiefe CI, Liu F. In-hospital mortality prediction among intensive care unit patients with acute ischemic stroke: a machine learning approach. Health Data Sci 2025;5:0179 [FREE Full text] [doi: 10.34133/hds.0179] [Medline: 40099281]

XSL•FO
RenderX

49. Park E, Chang HJ, Nam HS. A bayesian network model for predicting post-stroke outcomes with available risk factors. Front Neurol 2018;9:699 [FREE Full text] [doi: 10.3389/fneur.2018.00699] [Medline: 30245663]
50. Ouyang Y, Cheng M, He B, Zhang F, Ouyang W, Zhao J, et al. Interpretable machine learning models for predicting in-hospital death in patients in the intensive care unit with cerebral infarction. Comput Methods Programs Biomed 2023;231:107431. [doi: 10.1016/j.cmpb.2023.107431] [Medline: 36827826]
51. Luo X, Li B, Zhu R, Tai Y, Wang Z, He Q, et al. Development and validation of an interpretable machine learning model for predicting in-hospital mortality for ischemic stroke patients in ICU. Int J Med Inform 2025;198:105874. [doi: 10.1016/j.ijmedinf.2025.105874] [Medline: 40073651]
52. Mai TD, Nguyen DT, Tran CC, Duong HQ, Nguyen HN, Dang DP, et al. Factors associated with 90-day mortality in Vietnamese stroke patients: prospective findings compared with explainable machine learning, multicenter study. PLoS One 2024;19(9):e0310522 [FREE Full text] [doi: 10.1371/journal.pone.0310522] [Medline: 39302916]
53. Lehmann ALCF, Alfieri DF, de Araújo MCM, Trevisani ER, Nagao MR, Pesente FS, et al. Immune-inflammatory, coagulation, adhesion, and imaging biomarkers combined in machine learning models improve the prediction of death 1 year after ischemic stroke. Clin Exp Med 2022;22(1):111-123. [doi: 10.1007/s10238-021-00732-w] [Medline: 34120242]
54. Kim C, Kwon JM, Lee J, Jo H, Gwon D, Jang JH, et al. Deep learning model integrating radiologic and clinical data to predict mortality after ischemic stroke. Heliyon 2024;10(10):e31000 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e31000] [Medline: 38826743]
55. Huang R, Liu J, Wan TK, Siriwanna D, Woo YMP, Vodencarevic A, et al. Stroke mortality prediction based on ensemble learning and the combination of structured and textual data. Comput Biol Med 2023;155:106176. [doi: 10.1016/j.compbiomed.2022.106176] [Medline: 36805232]
56. Huang J, Liu X, Jin W. Clinical decision support systems for 3-month mortality in elderly patients admitted to ICU with ischemic stroke using interpretable machine learning. Digit Health 2024;10:20552076241280126 [FREE Full text] [doi: 10.1177/20552076241280126] [Medline: 39314817]
57. Huang J, Jin W, Duan X, Liu X, Shu T, Fu L, et al. Twenty-eight-day in-hospital mortality prediction for elderly patients with ischemic stroke in the intensive care unit: interpretable machine learning models. Front Public Health 2022;10:1086339 [FREE Full text] [doi: 10.3389/fpubh.2022.1086339] [Medline: 36711330]
58. Huang J, Chen H, Deng J, Liu X, Shu T, Yin C, et al. Interpretable machine learning for predicting 28-day all-cause in-hospital mortality for hypertensive ischemic or hemorrhagic stroke patients in the ICU: a multi-center retrospective cohort study with internal and external cross-validation. Front Neurol 2023;14:1185447 [FREE Full text] [doi: 10.3389/fneur.2023.1185447] [Medline: 37614971]
59. Hoffman H, Wood J, Cote JR, Jalal MS, Otite FO, Masoud HE, et al. Development and internal validation of machine learning models to predict mortality and disability after mechanical thrombectomy for acute anterior circulation large vessel occlusion. World Neurosurg 2024;182:e137-e154. [doi: 10.1016/j.wneu.2023.11.060] [Medline: 38000670]
60. Fernandez-Lozano C, Hervella P, Mato-Abad V, Rodríguez-Yáñez M, Suárez-Garaboa S, López-Dequidt I, et al. Random forest-based prediction of stroke outcome. Sci Rep 2021;11(1):10071 [FREE Full text] [doi: 10.1038/s41598-021-89434-7] [Medline: 33980906]
61. Fang J, Song B, Li L, Tong L, Jiang M, Yan J. RGX ensemble model for advanced prediction of mortality outcomes in stroke patients. BME Front 2024;5:0077 [FREE Full text] [doi: 10.34133/bmef.0077] [Medline: 39600589]
62. Chunjuan Z, Yulong W, Xicheng Z, Xiaodong M. Machine learning consensus clustering for inflammatory subtype analysis in stroke and its impact on mortality risk: a study based on NHANES (1999-2018). Front Neurol 2025;16:1562247 [FREE Full text] [doi: 10.3389/fneur.2025.1562247] [Medline: 40276469]
63. Chen Y, Yang Z, Liu Y, Li Y, Zhong Z, McDowell G, et al. Exploring the prognostic impact of triglyceride-glucose index in critically ill patients with first-ever stroke: insights from traditional methods and machine learning-based mortality prediction. Cardiovasc Diabetol 2024;23(1):443 [FREE Full text] [doi: 10.1186/s12933-024-02538-y] [Medline: 39695656]
64. Chen H, Yang F, Duan Y, Yang L, Li J. A novel higher performance nomogram based on explainable machine learning for predicting mortality risk in stroke patients within 30 days based on clinical features on the first day ICU admission. BMC Med Inform Decis Mak 2024;24(1):161 [FREE Full text] [doi: 10.1186/s12911-024-02547-7] [Medline: 38849903]
65. Che Nawi CMNH, Mohd Hairon S, Wan Yahya WNN, Wan Zaidi WA, Musa KI. Machine learning models for predicting stroke mortality in malaysia: an application and comparative analysis. Cureus 2023;15(12):e50426 [FREE Full text] [doi: 10.7759/cureus.50426] [Medline: 38222138]
66. Abujaber A, Yaseen S, Imam Y, Nashwan A, Akhtar N. Machine learning-based prediction of one-year mortality in ischemic stroke patients. Oxf Open Neurosci 2024;3:kvae011. [doi: 10.1093/oons/kvae011] [Medline: 39569400]
67. Abedi V, Avula V, Razavi SM, Bavishi S, Chaudhary D, Shahjouei S, et al. Predicting short and long-term mortality after acute ischemic stroke using EHR. J Neurol Sci 2021;427:117560 [FREE Full text] [doi: 10.1016/j.jns.2021.117560] [Medline: 34218182]
68. Zhou L, Wu Y, Wang J, Wu H, Tan Y, Chen X, et al. Development of a predictive nomogram for intra-hospital mortality in acute ischemic stroke patients using LASSO regression. Clin Interv Aging 2024;19:1423-1436 [FREE Full text] [doi: 10.2147/CIA.S471885] [Medline: 39139210]

XSL•FO
RenderX

69. Zhang X, Yuan K, Wang H, Gong P, Jiang T, Xie Y, et al. Nomogram to predict mortality of endovascular thrombectomy for ischemic stroke despite successful recanalization. J Am Heart Assoc 2020;9(3):e014899 [FREE Full text] [doi: 10.1161/JAHA.119.014899] [Medline: 31973604]
70. Zhang RX, Zhang WW, Luo YT, Liu GW. An mNUTRIC-based nomogram for predicting the in-hospital death risk in patients with acute stroke. Eur J Clin Nutr 2022;76(10):1464-1469. [doi: 10.1038/s41430-022-01127-0] [Medline: 35411025]
71. Zhang J, Zhuang S, Zhang Y, An L, Li D, Liu Z, et al. Ultrasonic optic nerve sheath diameter as a new predictor for the mortality of patients with large hemispheric infarction. Sci Rep 2025;15(1):460 [FREE Full text] [doi: 10.1038/s41598-024-84720-6] [Medline: 39748001]
72. Zeng M, Smith L, Bird A, Trinh VQ, Bacchi S, Harvey J, et al. Predictions for functional outcome and mortality in acute ischaemic stroke following successful endovascular thrombectomy. BMJ Neurol Open 2024;6(1):e000707 [FREE Full text] [doi: 10.1136/bmjno-2024-000707] [Medline: 38932996]
73. Çelik G, Baykan Ö, Kara Y, Tireli H. Predicting 10-day mortality in patients with strokes using neural networks and multivariate statistical methods. J Stroke Cerebrovasc Dis 2014;23(6):1506-1512. [doi: 10.1016/j.jstrokecerebrovasdis.2013.12.018] [Medline: 24674954]
74. Wen R, Wang M, Bian W, Zhu H, Xiao Y, He Q, et al. Nomogram to predict 6-month mortality in acute ischemic stroke patients treated with endovascular treatment. Front Neurol 2023;14:1330959 [FREE Full text] [doi: 10.3389/fneur.2023.1330959] [Medline: 38249750]
75. Szlachetka WA, Pana TA, Mamas MA, Bettencourt-Silva JH, Metcalf AK, Potter JF, et al. Predicting 10-year stroke mortality: development and validation of a nomogram. Acta Neurol Belg 2022;122(3):685-693 [FREE Full text] [doi: 10.1007/s13760-021-01752-9] [Medline: 34406610]
76. Sun W, Li G, Liu Z, Miao J, Yang Z, Zhou Q, et al. A nomogram for predicting the in-hospital mortality after large hemispheric infarction. BMC Neurol 2019;19(1):347 [FREE Full text] [doi: 10.1186/s12883-019-1571-4] [Medline: 31884967]
77. Solberg OG, Dahl M, Mowinckel P, Stavem K. Derivation and validation of a simple risk score for predicting 1-year mortality in stroke. J Neurol 2007;254(10):1376-1383. [doi: 10.1007/s00415-007-0555-2] [Medline: 17934885]
78. Sha L, Xu T, Ge X, Shi L, Zhang J, Guo H. Predictors of death within 6 months of stroke onset: a model with barthel index, platelet/lymphocyte ratio and serum albumin. Nurs Open 2021;8(3):1380-1392 [FREE Full text] [doi: 10.1002/nop2.754] [Medline: 33378600]
79. Ni H, Zhao L, Ma X, Lv P, Ding Y, Liu Z, et al. Technical risk stratification nomogram model for 90-day mortality prediction in patients with acute basilar artery occlusion undergoing endovascular thrombectomy: a multicenter cohort study. J Am Heart Assoc 2024;13(6):e032107 [FREE Full text] [doi: 10.1161/JAHA.123.032107] [Medline: 38471827]
80. Liljehult J, Christensen T, Christensen KB. Early prediction of one-year mortality in ischemic and haemorrhagic stroke. J Stroke Cerebrovasc Dis 2020;29(4):104667. [doi: 10.1016/j.jstrokecerebrovasdis.2020.104667] [Medline: 32044222]
81. Li XD, Li MM. A novel nomogram to predict mortality in patients with stroke: a survival analysis based on the MIMIC-III clinical database. BMC Med Inform Decis Mak 2022;22(1):92 [FREE Full text] [doi: 10.1186/s12911-022-01836-3] [Medline: 35387672]
82. Kim TJ, Lee JS, Oh MS, Kim JW, Park SH, Yu KH, et al. Stroke measures analysis of pRognostic testing-mortality nomogram predicts long-term mortality after ischemic stroke. Int J Stroke 2025;20(1):75-84. [doi: 10.1177/17474930241278808] [Medline: 39150095]
83. Lee J, Morishima T, Kunisawa S, Sasaki N, Otsubo T, Ikai H, et al. Derivation and validation of in-hospital mortality prediction models in ischaemic stroke patients using administrative data. Cerebrovasc Dis 2013;35(1):73-80. [doi: 10.1159/000346090] [Medline: 23429000]
84. Jin G, Hu W, Zeng L, Ma B, Zhou M. Prediction of long-term mortality in patients with ischemic stroke based on clinical characteristics on the first day of ICU admission: an easy-to-use nomogram. Front Neurol 2023;14:1148185 [FREE Full text] [doi: 10.3389/fneur.2023.1148185] [Medline: 37122313]
85. Jin G, Hu W, Zeng L, Diao M, Chen H, Chen J, et al. Development and verification of a nomogram for predicting short-term mortality in elderly ischemic stroke populations. Sci Rep 2023;13(1):12580 [FREE Full text] [doi: 10.1038/s41598-023-39781-4] [Medline: 37537270]
86. Gao H, Sun X, Li W, Gao Q, Zhang J, Zhang Y, et al. Development and validation of a risk score to predict 30-day mortality in patients with atrial fibrillation-related stroke: GPS-GF score. Neurol Res 2018;40(7):532-540. [doi: 10.1080/01616412.2018.1451431] [Medline: 29544401]
87. Fang L, Zhou M, Mao F, Diao M, Hu W, Jin G. Development and validation of a nomogram for predicting 28-day mortality in patients with ischemic stroke. PLoS One 2024;19(4):e0302227 [FREE Full text] [doi: 10.1371/journal.pone.0302227] [Medline: 38656987]
88. Cappellari M, Zivelonghi C, Turcato G, Forlivesi S, Micheletti N, Tomelleri G, et al. A nomogram to predict the probability of mortality after first-ever acute manifestations of cerebral small vessel disease. J Neurol Sci 2018;385:92-95. [doi: 10.1016/j.jns.2017.12.020] [Medline: 29406922]

XSL•FO
RenderX

89. Schwartz J, Wang Y, Qin L, Schwamm LH, Fonarow GC, Cormier N, et al. Incorporating stroke severity into hospital measures of 30-day mortality after ischemic stroke hospitalization. Stroke 2017;48(11):3101-3107. [doi: 10.1161/STROKEAHA.117.017960] [Medline: 28954922]
90. Reitzle L, Rohmann JL, Kurth T, Audebert HJ, Piccininni M. External validation of risk prediction models for post-stroke mortality in Berlin. BMJ Open 2025;15(6):e089320 [FREE Full text] [doi: 10.1136/bmjopen-2024-089320] [Medline: 40480675]
91. Hwangbo L, Kang YJ, Kwon H, Lee JI, Cho H, Ko J, et al. Stacking ensemble learning model to predict 6-month mortality in ischemic stroke patients. Sci Rep 2022;12(1):17389 [FREE Full text] [doi: 10.1038/s41598-022-22323-9] [Medline: 36253488]
92. Huang Y, Douiri A, Fahey M. A dynamic model for predicting survival up to 1 year after ischemic stroke. J Stroke Cerebrovasc Dis 2020;29(10):105133 [FREE Full text] [doi: 10.1016/j.jstrokecerebrovasdis.2020.105133] [Medline: 32912566]
93. Geng L, Zuo L, Li H, Wang Y, Zhang Q, Ran J, et al. Association of platelet-to-neutrophil ratios with 1-year outcome and mortality in patients with acute ischemic stroke. Neurosci Lett 2023;798:137016. [doi: 10.1016/j.neulet.2022.137016] [Medline: 36529389]
94. Zhang S, Wang J, Pei L, Liu K, Gao Y, Fang H, et al. Interpretability analysis of one-year mortality prediction for stroke patients based on deep neural network. IEEE J Biomed Health Inform 2022;26(4):1903-1910. [doi: 10.1109/JBHI.2021.3123657] [Medline: 34714758]
95. Abujaber AA, Albalkhi I, Imam Y, Nashwan A, Akhtar N, Alkhawaldeh IM. Machine learning-based prognostication of mortality in stroke patients. Heliyon 2024;10(7):e28869 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e28869] [Medline: 38601648]
96. Zhu E, Chen Z, Ai P, Wang J, Zhu M, Xu Z, et al. Analyzing and predicting the risk of death in stroke patients using machine learning. Front Neurol 2023;14:1096153 [FREE Full text] [doi: 10.3389/fneur.2023.1096153] [Medline: 36816575]
97. Chang Y, Zhang L, Li Y, Wang D, Fang Q, Tang X. Derivation and validation of a new visceral adiposity index for predicting short-term mortality of patients with acute ischemic stroke in a chinese population. Brain Sci 2023;13(2):297 [FREE Full text] [doi: 10.3390/brainsci13020297] [Medline: 36831840]
98. Yan X, Xia P, Tong H, Lan C, Wang Q, Zhou Y, et al. Development and validation of a dynamic nomogram for predicting 3-month mortality in acute ischemic stroke patients with atrial fibrillation. Risk Manag Healthc Policy 2024;17:145-158 [FREE Full text] [doi: 10.2147/RMHP.S442353] [Medline: 38250220]
99. Wang W, Otieno JA, Eriksson M, Wolfe CD, Curcin V, Bray BD. Developing and externally validating a machine learning risk prediction model for 30-day mortality after stroke using national stroke registers in the UK and Sweden. BMJ Open 2023;13(11):e069811 [FREE Full text] [doi: 10.1136/bmjopen-2022-069811] [Medline: 37968001]
100. Easton JF, Stephens CR, Angelova M. Risk factors and prediction of very short term versus short/intermediate term post-stroke mortality: a data mining approach. Comput Biol Med 2014;54:199-210. [doi: 10.1016/j.compbiomed.2014.09.003] [Medline: 25303114]
101. Lea-Pereira MC, Amaya-Pascasio L, Martínez-Sánchez P, Rodríguez Salvador MDM, Galván-Espinosa J, Téllez-Ramírez L, et al. Predictive model and mortality risk score during admission for ischaemic stroke with conservative treatment. Int J Environ Res Public Health 2022;19(6):3182 [FREE Full text] [doi: 10.3390/ijerph19063182] [Medline: 35328867]
102. Süt N, Celik Y. Prediction of mortality in stroke patients using multilayer perceptron neural networks. Turk J Med Sci 2011:886-893. [doi: 10.3906/sag-1105-20]
103. Gattringer T, Posekany A, Niederkorn K, Knoflach M, Poltrum B, Mutzenbach S, et al. Predicting early mortality of acute ischemic stroke. Stroke 2019;50(2):349-356. [doi: 10.1161/strokeaha.118.022863]
104. Yang Y, Tang L, Deng Y, Li X, Luo A, Zhang Z, et al. The predictive performance of artificial intelligence on the outcome of stroke: a systematic review and meta-analysis. Front Neurosci 2023;17:1256592 [FREE Full text] [doi: 10.3389/fnins.2023.1256592] [Medline: 37746141]
105. Schwartz L, Anteby R, Klang E, Soffer S. Stroke mortality prediction using machine learning: systematic review. J Neurol Sci 2023;444:120529. [doi: 10.1016/j.jns.2022.120529] [Medline: 36580703]
106. Wang W, Kiik M, Peek N, Curcin V, Marshall IJ, Rudd AG, et al. A systematic review of machine learning models for predicting outcomes of stroke with structured data. PLoS One 2020;15(6):e0234722 [FREE Full text] [doi: 10.1371/journal.pone.0234722] [Medline: 32530947]
107. Yu Y, Zhong Z, Yang W, Yu J, Li J, Guo X, et al. Neutrophil percentage-to-albumin ratio and risk of mortality in patients on peritoneal dialysis. J Inflamm Res 2023;16:6271-6281 [FREE Full text] [doi: 10.2147/JIR.S437256] [Medline: 38146321]
108. Chung J, Ahn J, Ryu JA. Beyond SOFA and APACHE II, novel risk stratification models using readily available biomarkers in critical care. Diagnostics (Basel) 2025;15(9):1122 [FREE Full text] [doi: 10.3390/diagnostics15091122] [Medline: 40361939]
109. Lin J, Frontera JA. Decompressive hemicraniectomy for large hemispheric strokes. Stroke 2021;52(4):1500-1510 [FREE Full text] [doi: 10.1161/STROKEAHA.120.032359] [Medline: 33719518]
110. Maddox TM, Rumsfeld JS, Payne PRO. Questions for artificial intelligence in health care. JAMA 2019;321(1):31-32. [doi: 10.1001/jama.2018.18932] [Medline: 30535130]

XSL•FO
RenderX

111. Lewis AE, Weiskopf N, Abrams ZB, Foraker R, Lai AM, Payne PRO, et al. Electronic health record data quality assessment and tools: a systematic review. J Am Med Inform Assoc 2023;30(10):1730-1740 [FREE Full text] [doi: 10.1093/jamia/ocad120] [Medline: 37390812]
112. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12-22. [doi: 10.1016/j.jclinepi.2019.02.004] [Medline: 30763612]
113. Oka S, Yamazaki T, Takefuji Y. Addressing bias in machine learning feature importance for food quality assessment. Food Chem 2025;494:146171 [FREE Full text] [doi: 10.1016/j.foodchem.2025.146171] [Medline: 40902562]
114. Allgaier J, Mulansky L, Draelos RL, Pryss R. How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare. Artif Intell Med 2023;143:102616 [FREE Full text] [doi: 10.1016/j.artmed.2023.102616] [Medline: 37673561]
115. Zabor EC, Reddy CA, Tendulkar RD, Patil S. Logistic regression in clinical studies. Int J Radiat Oncol Biol Phys 2022;112(2):271-277. [doi: 10.1016/j.ijrobp.2021.08.007] [Medline: 34416341]
116. Yang J, Zeng S, Cui S, Zheng J, Wang H. Predictive modeling of acute respiratory distress syndrome using machine learning: systematic review and meta-analysis. J Med Internet Res 2025;27:e66615 [FREE Full text] [doi: 10.2196/66615] [Medline: 40359510]
117. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform 2008;77(2):81-97 [FREE Full text] [doi: 10.1016/j.ijmedinf.2006.11.006] [Medline: 17188928]
118. La Cava WG, Lee PC, Ajmal I, Ding X, Solanki P, Cohen JB, et al. A flexible symbolic regression method for constructing interpretable clinical prediction models. NPJ Digit Med 2023;6(1):107 [FREE Full text] [doi: 10.1038/s41746-023-00833-8] [Medline: 37277550]
119. Petch J, Di S, Nelson W. Opening the black box: the promise and limitations of explainable machine learning in cardiology. Can J Cardiol 2022;38(2):204-213 [FREE Full text] [doi: 10.1016/j.cjca.2021.09.004] [Medline: 34534619]
120. Bergomi L, Nicora G, Orlowska MA, Podrecca C, Bellazzi R, Fregosi C, et al. Which explanations do clinicians prefer? A comparative evaluation of XAI understandability and actionability in predicting the need for hospitalization. BMC Med Inform Decis Mak 2025;25(1):269 [FREE Full text] [doi: 10.1186/s12911-025-03045-0] [Medline: 40671048]
121. Nazim S, Alam MM, Rizvi SS, Mustapha JC, Hussain SS, Suud MM. Advancing malware imagery classification with explainable deep learning: a state-of-the-art approach using SHAP, LIME and grad-cam. PLoS One 2025;20(5):e0318542 [FREE Full text] [doi: 10.1371/journal.pone.0318542] [Medline: 40435333]
122. Kuo RYL, Harrison C, Curran TA, Jones B, Freethy A, Cussons D, et al. Artificial intelligence in fracture detection: a systematic review and meta-analysis. Radiology 2022;304(1):50-62 [FREE Full text] [doi: 10.1148/radiol.211785] [Medline: 35348381]
123. Khokhar PB, Gravino C, Palomba F. Advances in artificial intelligence for diabetes prediction: insights from a systematic literature review. Artif Intell Med 2025;164:103132 [FREE Full text] [doi: 10.1016/j.artmed.2025.103132] [Medline: 40258308]
124. Ruan X, Fu S, Storlie CB, Mathis KL, Larson DW, Liu H. Real-time risk prediction of colorectal surgery-related post-surgical complications using GRU-D model. J Biomed Inform 2022;135:104202 [FREE Full text] [doi: 10.1016/j.jbi.2022.104202] [Medline: 36162805]
125. Bednarski BP, Singh AD, Zhang W, Jones WM, Naeim A, Ramezani R. Temporal convolutional networks and data rebalancing for clinical length of stay and mortality prediction. Sci Rep 2022;12(1):21247 [FREE Full text] [doi: 10.1038/s41598-022-25472-z] [Medline: 36481828]
126. Shurrab S, Guerra-Manzanares A, Magid A, Piechowski-Jozwiak B, Atashzar SF, Shamout FE. Multimodal machine learning for stroke prognosis and diagnosis: a systematic review. IEEE J Biomed Health Inform 2024;28(11):6958-6973. [doi: 10.1109/JBHI.2024.3448238] [Medline: 39172620]
127. Armoundas AA, Narayan SM, Arnett DK, Spector-Bagdady K, Bennett DA, Celi LA, et al. Use of artificial intelligence in improving outcomes in heart disease: a scientific statement from the American Heart Association. Circulation 2024;149(14):e1028-e1050 [FREE Full text] [doi: 10.1161/CIR.0000000000001201] [Medline: 38415358]
128. Mansournia MA, Etminan M, Danaei G, Kaufman JS, Collins G. Handling time varying confounding in observational research. BMJ 2017;359:j4587. [doi: 10.1136/bmj.j4587] [Medline: 29038130]
129. Braun RG, Edwardson MA, Aldridge CM, Cramer SC, Cruchaga C, Engelter ST, et al. Blood biospecimen recommendations for research on stroke outcomes and recovery. Stroke 2025;56(10):e279-e290. [doi: 10.1161/STROKEAHA.125.050793] [Medline: 40982604]
130. Fisher LD, Lin DY. Time-dependent covariates in the cox proportional-hazards regression model. Annu Rev Public Health 1999;20:145-157. [doi: 10.1146/annurev.publhealth.20.1.145] [Medline: 10352854]
131. Chen X, Wang L, W YYYEDCL. Mitigating data imbalance and noise: a divergence-based approach with enhanced sample selection. Neurocomputing 2024;605:1-10. [doi: 10.1016/j.neucom.2024.128269]
132. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020;368:m689 [FREE Full text] [doi: 10.1136/bmj.m689] [Medline: 32213531]

XSL•FO
RenderX

133. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol 2003;56(5):441-447. [doi: 10.1016/s0895-4356(03)00047-7] [Medline: 12812818]
134. Dubray-Vautrin A, Gravrand V, Marret G, Lamy C, Klijanienko J, Vacher S, et al. Internal validation strategy for high dimensional prognosis model: a simulation study and application to transcriptomic in head and neck tumors. Comput Struct Biotechnol J 2025;27:3792-3802 [FREE Full text] [doi: 10.1016/j.csbj.2025.08.035] [Medline: 40989918]
135. Teale EA, Forster A, Munyombwe T, Young JB. A systematic review of case-mix adjustment models for stroke. Clin Rehabil 2012;26(9):771-786. [doi: 10.1177/0269215511433068] [Medline: 22257504]
136. Peters U, Sherling HR, Chin-Yee B. Hasty generalizations and generics in medical research: a systematic review. PLoS One 2024;19(7):e0306749 [FREE Full text] [doi: 10.1371/journal.pone.0306749] [Medline: 38968284]
137. An D, Lim M, Lee S. Challenges for data quality in the clinical data life cycle: systematic review. J Med Internet Res 2025;27:e60709 [FREE Full text] [doi: 10.2196/60709] [Medline: 40266662]
138. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011;343:d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
139. Borenstein M, Higgins JPT. Meta-analysis and subgroups. Prev Sci 2013;14(2):134-143. [doi: 10.1007/s11121-013-0377-7] [Medline: 23479191]
140. Labarère J, Renaud B, Fine MJ. How to derive and validate clinical prediction models for use in intensive care medicine. Intensive Care Med 2014;40(4):513-527. [doi: 10.1007/s00134-014-3227-6] [Medline: 24570265]
141. Lulin Z, Yiranbon E, Asante Antwi H. Complementarity of clinician judgment and evidence based models in medical decision making: antecedents, prospects, and challenges. Biomed Res Int 2016;2016:1425693 [FREE Full text] [doi: 10.1155/2016/1425693] [Medline: 27642588]
142. Jeong S, Shivakumar M, Jung S, Won H, Nho K, Huang H, et al. Addressing overfitting bias due to sample overlap in polygenic risk scoring. Alzheimers Dement 2025;21(4):e70109. [doi: 10.1002/alz.70109] [Medline: 40189831]

Abbreviations

AI: artificial intelligence
AIS: acute ischemic stroke
AUC: area under the curve
DL: deep learning
eICU-CRD: eICU Collaborative Research Database
HKSJ: Hartung-Knapp-Sidik-Jonkman
LR: logistic regression
MIMIC: Medical Information Mart for Intensive Care
ML: machine learning
mRS: modified Rankin Scale
NIHSS: National Institutes of Health Stroke Scale
NPAR: neutrophil-percentage-to-albumin ratio
PI: prediction interval
PNR: platelet-to-neutrophil ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
PROBAST: Prediction model Risk of Bias Assessment Tool
RF: random forest
RoB: risk of bias
SE: sensitivity
SHAP: Shapley additive explanations
SP: specificity
SSNAP: Sentinel Stroke National Audit Programme
XGBoost: extreme gradient boosting

XSL•FO
RenderX

Edited by S Brini; submitted 09.Sep.2025; peer-reviewed by GI Han, JMI Arockiasamy; comments to author 31.Oct.2025; accepted 19.Feb.2026; published 02.Apr.2026.

Please cite as:
Chen Y, Ou Z, Deng Y, Luo A, Li X, Yang Y, Yang Q, Wu X, Zhou M
Predictive Value of Machine Learning for Poststroke Mortality Risk: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e83821
URL: https://www.jmir.org/2026/1/e83821
doi:10.2196/83821
PMID:

©Yujie Chen, Zhujing Ou, Yiting Deng, Anling Luo, Xuzi Li, Yujia Yang, Qin Yang, Xintong Wu, Muke Zhou. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Artificial Intelligence, Connected Care, and Enabling Digital Health Technologies in Rare Diseases With a Focus on Lysosomal Storage Disorders: Scoping Review

Alberta MC Spreafico[1,2,3], MSc, MBA, PhD; Luca Neri[4,5], MD; Kim Angel[6]; Cinzia Maria Bellettato[7], PhD; Maurizio Scarpa[7], Prof Dr Med

[1]Health Innovation, EVERSANA INTOUCH, Salerno, Italy
[2]Government, Health and Not for Profit Division, SDA Bocconi School of Management, Milano, Italy
[3]Innovation for Global Health Institute, Milan, Italy
[4]WINFOCUS, Milan, Italy
[5]Iuliu Ha ieganu University of Medicine and Pharmacy, Cluj-Napoca, Romania
[6]International MPS Network, Vancouver, BC, Canada
[7]MetabERN, Regional Coordinating Center for Rare Diseases, Udine University Hospital, Udine, Italy

Corresponding Author:
Alberta MC Spreafico, MSc, MBA, PhD
Health Innovation
EVERSANA INTOUCH
Palazzo Innovazione Piazza Abate Conforti
Salerno, 84121
Italy
Phone: 39 3386337713
Email: alberta.spreafico@eversana.com

Abstract

Background: Rare diseases affect more than 300 million people globally, and only about 5% have approved therapies. Lysosomal storage disorders (LSDs) exemplify the diagnostic and long-term care complexity typical of rare diseases, and digital health technologies (DHTs), especially artificial intelligence (AI) and connected care (CC), are emerging tools to support LSD management.

Objective: We aimed to map and synthesize peer-reviewed and gray literature from the past decade on DHTs relevant for LSD care, with a primary analytic focus on AI-enabled and CC solutions and a contextual mapping of other enabling DHTs. Evidence distribution was charted by population, care-journey phase, and outcome domains to identify gaps, methodological limitations, and timely priorities relevant for research, clinical practice implementation, and policies.

Methods: We conducted a scoping review guided by a population, concept, context framework and operationalized through a Population, Intervention, Comparison, and Outcome (PICO)-informed data-charting structure to map study characteristics and reported outcomes, without causal or effectiveness assumptions and without risk-of-bias assessment. We searched PubMed, Google Scholar, and ClinicalTrials.gov for studies published between October 2015 and September 2024, complemented by AI-assisted discovery tools for citation extension. Reproducibility logs (search strings, run dates, filters, and stepwise counts) were maintained. Of 1751 records retrieved, 245 were included. Evidence was charted by LSD population, intervention class (AI, CC, and other enabling DHTs), outcome domains (patient, health care, and societal), and phase of the care journey.

Results: Among 245 included records, 92.2% (226/245) were peer-reviewed, and 7.8% (19/245) were gray literature; no completed and published randomized controlled trials or LSD-specific systematic reviews were identified, with evidence dominated by small, single-center observational studies. Overall, 40 peer-reviewed records reported AI-driven DHTs, 89 reported CC DHTs, and 144 reported other enabling DHTs (some multilabeled). Evidence was concentrated mostly in Gaucher and Fabry diseases. Nearly half of the mapped literature focused on screening and diagnosis, with fewer records addressing treatment intensification, rehabilitation, and end-of-life care. Outcomes were predominantly health care delivery performance measures, with fewer patient and societal outcomes. AI applications mainly supported diagnostic decision support, phenotyping, monitoring, tracking, and risk stratification; CC commonly involved telemedicine, remote monitoring, and patient-engagement platforms; enabling DHTs included interoperable data systems, registries, and digital infrastructures.

XSL•FO
RenderX

Conclusions: The evidence base is appreciable for a niche field and reflects growing interest in AI and CC for LSD care, but heterogeneity and methodological limitations preclude inferences on effectiveness or routine implementation. This evidence map highlights relatively stronger areas and gaps, providing a structured foundation to inform timely expert consensus-building and research prioritization. Key priorities include interoperable data infrastructures and data availability, prospective multicenter evaluations, transparent reporting of algorithms and workflows, and implementation-relevant outcomes to support safe, equitable, and scalable adoption aligned with evolving European Union and global rare-disease priorities.

(J Med Internet Res 2026;28:e73612) doi:10.2196/73612

KEYWORDS

digital health; artificial intelligence; connected care; telemedicine; rare diseases; lysosomal storage disorders; remote monitoring; registries

Introduction

Rationale

Lysosomal storage disorders (LSDs) are a subset of rare diseases (RDs), resulting from inherited deficiencies in lysosomal enzymes or transporters. They affect approximately 1 in 7000 to 8000 newborns and encompass over 70 disorders. They exemplify rare, chronic, and multisystem diseases that require coordinated, longitudinal, multispecialty care [1-4]. The 2021 United Nations resolution on “Addressing the Challenges of Persons Living with a Rare Disease and their Families” [5], together with several policy statements from the European Union (EU) and the World Health Organization, recognizes RDs as a global policy priority and calls for coordinated action and strengthened care networks to advance equitable Universal Health Coverage by 2030 [3].

Across RDs, the diffusion of digital health technologies (DHTs) has accelerated in both research and care delivery, reflecting the need to enhance early and accurate diagnosis, reduce geographic barriers, improve inclusion for dispersed populations, and support longitudinal follow-up [6,7]. DHTs have also been increasingly adopted in RD clinical trials to enable remote assessments and decentralized research models [8]. In parallel, regulators and health systems are integrating digital health into health policies and health technology assessment (HTA) frameworks, with growing emphasis on clinical validation, transparency, and safety. These cross-cutting developments are especially relevant to LSDs, where chronic multisystem trajectories require coordinated multidisciplinary care, timely specialist input, and monitoring across the life course.

However, despite this rapid diffusion and growing policy attention, the landscape of DHT applications in LSD care remains poorly characterized. While the burden and care complexity of RDs, such as LSDs, are well established, the evolving contributions of artificial intelligence (AI) and connected care (CC), as distinct yet complementary classes of DHTs with the potential to enhance LSD management, have not been systematically mapped [3,5,9]. Over the past decade, AI applications (eg, pattern recognition for early diagnosis, predictive analytics, and data-driven decision support systems [DSS]) and CC solutions (eg, televisits, remote monitoring, and digital medical devices [DMD]) have been increasingly reported in LSD contexts. However, evidence is dispersed across heterogeneous study designs, gray literature, and emerging evaluations, which limits informed decision-making for clinicians, patient organizations, and policymakers. Prior reviews survey broader rare-disease contexts or single DHT classes, but none provide an LSD-specific evidence map specifying the studied interventions, populations, care-journey stages, and outcome domains—information fundamental to responsible research, clinical adoption, and policy development.

Objectives

Considering this evolving context, we conducted, to our knowledge, the first scoping review conducted in accordance with PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [10] to describe and map the published and gray literature from the past decade on DHTs relevant to LSD care, with a primary analytic focus on AI-enabled and CC applications and a contextual mapping of other enabling DHTs.

Our aims were to (1) clarify scope and applications across AI, CC, and other enabling DHTs; (2) chart the distribution of evidence across 4 axes (populations, intervention types, care-journey phases, and outcome domains) using an explicit charting framework (Textbox 1 and Figures 1-2); and (3) identify gaps, methodological limitations, and research priorities relevant to clinical implementation and policy development [3,5,9].

This review is explicitly descriptive in nature and does not seek to determine effectiveness or infer causal impact.

XSL•FO
RenderX

Textbox 1. Intervention-based categorization framework to map, cluster, and assess literature for AI-driven, connected care, and other digital health technologies.

Artificial intelligence–driven digital health technologies

- Artificial intelligence (AI) algorithms
  - Machine learning (ML) models, deep learning (DL) and neural networks (NN), natural language processing (NLP), and large language models (LLMs)
- AI-driven digital health technology (DHT) clinical applications
  - Risk stratification, facial recognition, clinical imaging analysis, predictive analytics, decision support, intelligent genomic counselling, virtual agents, and assistants
- AI-driven DHTs research and development applications
  - Drug development, enhanced or simulated clinical trials, disease and phenotypic pattern recognition, and disease epidemiological and progression modeling

Connected care DHTs

- Telemedicine
  - Health care provider (HCP)-patient televisits (including tele-genetics), HCP-HCP teleconsultations, collaborative health care networks
- Digital medical devices
  - Remote patient monitoring (RPM) devices, electronic patient-reported outcomes (ePRO), Mobile Health (mHealth) apps, patient care support programs, digital therapeutics (DTx)
- Data integration systems
  - Electronic health records (EHRs), health information systems (HIS), and national and international registries
- Patient engagement platforms
  - Online patient portals, networks, and peer-to-peer communities

Other DHTs

- Clinical DHTs
  - Precision medicine platforms, digital microfluidics (DMF) platforms, high-performance liquid chromatography (HPLC)–tandem mass spectrometry (MS/MS), bioinformatics tools, enzyme replacement therapy (ERT) optimization tools, virtual reality (VR) & augmented reality (AR), biomedical imaging, decision support systems (DSS)
- Data integration systems enablers
  - Blockchain for health data security
- Research and development DHTs
  - Clinical trial simulation tools, computational modeling
  - Multipurpose omics and molecular analysis DHTs
  - Genomics (including next-generation sequencing), pharmacogenomics, metabolomics, and proteomics applications and platforms.

XSL•FO
RenderX

Figure 1. Outcome-based framework to map, cluster, and assess literature for patient, health care, and societal impact. HCP: health care provider.

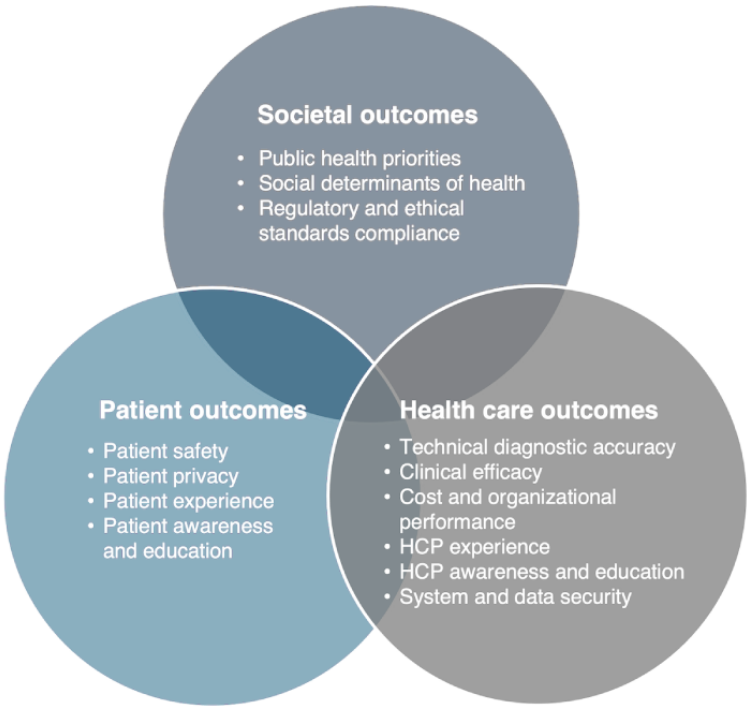

Figure 2. Timed framework to map, cluster, and assess literature for LSD natural history and care journey phases.

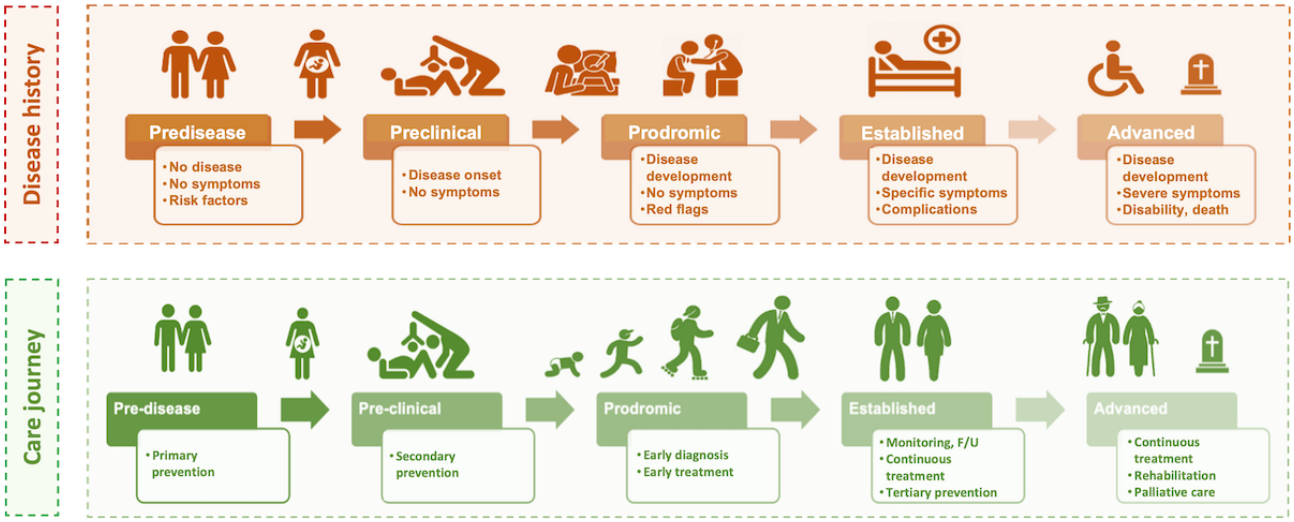

Methods

Overview

The dynamic and rapidly advancing field of DHTs and the variability in the quality of existing evidence render traditional systematic reviews and graded recommendations unsuitable, while making scoping reviews particularly applicable and informative. We used a multidimensional, PICO (population, intervention, comparison, and outcome)-informed, charting framework to classify and map evidence (Textbox 1 and Figures 1-2), while framing the overarching review question using a population–concept–context (PCC) logic. In this study, "PICO-informed" refers to a structured approach to categorize populations, intervention, and outcome domains for mapping purposes, not to a causal comparison framework—consistently with the aim of scoping reviews to map the breadth, characteristics, and gaps of evidence rather than to estimate comparative effectiveness.

The work adheres to the 5-stage scoping review framework advanced by Arksey and O'Malley [11] and follows the PRISMA-ScR guidelines [10,12]. Specifically, reporting is aligned with the 22-item PRISMA-ScR checklist (Multimedia Appendix 1). A prospective protocol was not registered; we acknowledge this as a limitation for transparency and therefore provide full search strings, dates, and flows to support reproducibility.

Identifying the Research Question

Our review question was as follows: In adult and pediatric populations with LSDs (including at-risk/suspected cases), their caregivers, and/or health care professionals (Population), what types of DHTs, particularly AI-enabled and connected care solutions, along with other enabling digital technologies

XSL•FO
RenderX

(Concept), have been reported over the last decade, and how is the evidence distributed across the LSD care-journey phases and across patient, health care delivery, and societal impact domains (Context/Outcomes)?

Comparators were not required for inclusion; when present, they were recorded descriptively. Outcomes were extracted as reported and mapped across predefined domains (Figure 1) to support descriptive synthesis and identification of evidence gaps, rather than causal inference.

Identifying Relevant Studies

We searched PubMed (including MEDLINE) and Google Scholar as core sources, complemented by queries on ClinicalTrials.gov and AI-assisted discovery tools, to broaden recall beyond records explicitly indexed in the core databases. Filters were applied to include studies published in the last decade, regardless of language, and to capture relevant gray literature, including non-peer-reviewed preprint studies, trial registries, theses, reports, expert interviews, book chapters, websites, and institutional or patient association portals. Overall, the multisource search retrieved 1751 records after prescreening. Full database-by-database reproducibility logs (strings, run-dates, filters, and stepwise counts) are provided in Tables S1-S7 in Multimedia Appendix 2.

On September 30, 2024, the baseline PubMed query returned 218 records (full stepwise string and fixed publication date filter reported in #8 in Multimedia Appendix 2). To increase sensitivity while preserving specificity for LSDs, we expanded the DHT concept using 55 keywords, consolidated from >80 DHT intervention terms identified during preliminary mapping, yielding an additional 190 records (total PubMed yield 408; see #5 and #10 in Multimedia Appendix 2). The term “rare disease” was not included, as it captured an excessive number of papers (>1900) with poor specificity for LSDs.

Google Scholar was searched on the same date using the predefined query and year filter (2015-2024), yielding approximately 4070 results (including duplicates and gray literature). To run specific reproducible screening under platform ranking variability, we applied a prespecified stopping rule by exporting and screening only the first 1000 results sorted by relevance (see #24 and #25 in Multimedia Appendix 2).

A similar research query was run on the randomized controlled trial (RCT) registry ClinicalTrials.gov (full string and fixed publication-date filter reported in #32 in Multimedia Appendix 2). We preliminarily identified 14 ongoing or completed, but unpublished, RCTs.

Consensus, SciSpace, and Connected Papers were used as supplementary AI-driven search tools (query expansion and citation chasing) on September 30, 2024, retrieving respectively 50, 163, and 102 records, including duplicates across the outputs (see #34, #35, and #47 in Multimedia Appendix 2). These tools were used only to identify candidate records; they did not autonomously determine eligibility, extract data, or draw conclusions. For reproducibility, we logged for each tool the run date and reference period, inputs (keywords, filters, or seed articles), and the number of candidate records retrieved. The author team reviewed all retrieved outputs and underwent the same deduplication and screening workflow as all other records.

An additional 14 records were finally added from citations, authors’ suggestions, and general web searches.

Study Selection

Records were de-duplicated using Paperpile (non-AI**,** reference manager software) and then double-checked manually. Title and abstract screening was performed by one reviewer (LN) after calibration with a second reviewer (AMCS) to refine eligibility criteria and resolve ambiguities. To mitigate single-screener selection error, AMCS independently screened (1) all records marked as uncertain and (2) a random sample of excluded records; discrepancies were resolved by discussion and, when needed, adjudication by a third reviewer (MS). LN and AMCS assessed full-text eligibility independently, with MS adjudicating disagreements. All researchers reached consensus on the final selection of studies included for data synthesis.

Given the emerging nature of this research field, the eligibility criteria were kept broad to include all study types published over the past 10 years in any language. Gray literature was also taken into consideration if particularly relevant. However, only records that directly addressed our scoping review question, explicitly involving both LSDs and at least one DHT mapped in our taxonomy (Textbox 1), were selected for analysis.

Inclusion criteria were (1) populations with a confirmed LSD diagnosis or at-risk LSD status (including screening, suspected, or undifferentiated LSD cases), and/or their caregivers or health care professionals; (2) digital interventions classified as AI, CC, or other DHTs according to the taxonomy in Textbox 1; and (3) outcomes that could be mapped to at least one of the predefined patient, health care delivery, or societal impact domains.

We excluded records for which (1) full text, or a minimum set of extractable data, could not be retrieved; or records that were (2) not LSD-specific and/or did not involve a DHT as defined in Textbox 1; where cited in Discussion, it served contextual purposes only. Non–peer-reviewed items were retained only when a stable, traceable source was accessible, and core data could be verified.

Data Charting and Synthesis

To convey source heterogeneity, we labeled each included record by peer-reviewed and gray source type, and summarized counts and distribution by evidence tier (Table 1). During synthesis, peer-reviewed items were always summarized separately from gray literature. Source validation and interpretive weighting followed a simple hierarchy: peer-reviewed original studies and reviews formed the primary basis for interpretation; gray sources were used descriptively only to contextualize emerging initiatives and evidence gaps. No clinical or system-level statement in this review rests solely on non–peer-reviewed sources.

XSL•FO
RenderX

Table 1. Literature source types of included records (distribution): corpus split into peer-reviewed and gray literature with subcategories per tier.

Codes and source type		Unique papers, n	% Source tier[a], % (n/N)	% Total papers[b], % (n/N)	Role in synthesis
L.1 Peer-reviewed		226	100 (226/226)	92.2 (226/245)	__[c]
L.1.1	RCTs[d], systematic reviews (SR), meta-analyses	0	0	0	Primary
L.1.2	Narrative, scoping reviews	44	19.5 (44/226)	18 (44/245)	Secondary
L.1.3	Original research (non-RCT)	176	77.9 (176/226)	71.8 (176/245)	Secondary
L.1.4	Case reports, case series	6	2.7 (6/226)	2.4 (6/245)	Context only
L.1.5	Editorials, letters, commentaries, perspectives	0	0	0	Context only
L.2 Gray literature		19	100 (19/19)	7.8 (19/245)	—
L.2.1	Registered, ongoing RCTs	7	36.8 (7/19)	2.9 (7/245)	Context only
L.2.2	Preprints	2	10.5 (2/19)	0.8 (2/245)	Context only
L.2.3	Conference abstracts, proceedings	1	5.3 (1/19)	0.4 (1/245)	Context only
L.2.4	Theses, dissertations	4	21.1 (4/19)	1.6 (4/245)	Context only
L.2.5	Books, book chapters	2	10.5 (2/19)	0.8 (2/245)	Context only
L.2.6	Websites, web posts	1	5.3 (1/19)	0.4 (1/245)	Context only
L.2.7	Reports, white papers, interviews	1	5.3 (1/19)	0.4 (1/245)	Context only
L.2.8	Patents	1	5.3 (7/19)	0.4 (1/245)	Context only
L.1+L.2 All included records		245	—	100 (245/245)	—

[a]% of Total Papers uses 245 as the denominator.

[b]% of Source Tier uses 226 for peer-reviewed (182 research articles + 44 reviews) and 19 for gray literature (7 registered RCTs + 12 miscellaneous sources).

[c]Not available.

[d]RCT: randomized controlled trial.

In line with scoping review guidance, and due to the heterogeneity of study designs, objectives, and sources of evidence, no formal assessment of the risk of bias or methodological quality was performed. Instead, we charted design, sample size, setting, and other quality-relevant features and used these descriptively to temper interpretation in the Results and Discussion sections.

Evidence was mapped and synthesized along the 4 dimensions derived from our multidimensional review question: Population, Intervention class (Textbox 1), Outcome domains (patient, health care delivery, and societal-level outcomes; Figure 1); and Care-journey phase (Figure 2). This population–intervention–outcome–care-phase analytic framework accommodated heterogeneous designs; it was descriptive and label-based, relying on counts and cross-tabulation rather than formal meta-analysis or effect-size estimation.

Population in scope was represented by adults and children diagnosed (or at-risk of being affected) with LSDs, with a particular focus on Gaucher disease (GD), Fabry disease (FD), mucopolysaccharidosis (I, II, IIIA, IIIB, IVA, VI, VII), Pompe disease (PD), Niemann-Pick disease, Krabbe disease, metachromatic leukodystrophy, GM1 and GM2 Gangliosidosis (GM1G and GM2G), Batten disease, and aspartylglucosaminuria, among more than 70 classified entities. Where possible, insights were charted and reported by age group and LSD subtype to reflect differences in clinical presentation and care needs.

We clustered DHT interventions into 3 domains that were applied consistently during study selection, charting, and synthesis: AI-driven DHTs, CC DHTs, and other DHTs (Textbox 1). AI-driven DHTs are adaptive and, to varying degrees, autonomous systems capable of learning, reasoning, and generating outputs from input data (eg, machine learning [ML] and deep learning [DL] models, natural language processing [NLP and large language models [LLMs]). In clinical contexts, these tools typically support early diagnosis and phenotyping, risk prediction, and decision support. CC DHTs include telemedicine; DMDs such as devices monitoring clinical parameters and electronic patient-reported outcomes (ePROs), mobile health (mHealth) apps, and digital therapeutics (DTx); interoperable data-integration systems (electronic health record [EHR], health information systems [HIS], registries, and emerging health-data spaces); and patient-engagement platforms. The category of other DHTs comprises clinical and research and development (R&D) tools that are neither AI-based nor connected-care systems (eg, digital microfluidics, omics and bioinformatics platforms, non-AI imaging tools, and trial-simulation software).

XSL•FO
RenderX

The impact of DHTs was then assessed across outcomes grouped into 3 domains (Figure 1): patient-related outcomes (patient safety, privacy, experience, awareness and education), health care delivery outcomes (technical diagnostic accuracy, clinical efficacy, cost and organizational performance, system and data security, health care provider [HCP] experience, awareness and education), and societal outcomes (public health and social determinants of health, regulatory and ethical compliance). Each included record was coded to one or more (≥1) outcome domains; when a study reported multiple domains, it was multilabeled. As a result, totals in outcome tables may exceed the number of unique records. The Discussion interpreted cross-domain patterns (rather than effectiveness) in line with scoping aims.

We also generalized the natural history of LSDs and summarized how DHTs were positioned across key care-journey phases**:** primary prevention, secondary prevention (screening) and diagnosis, treatment, tertiary prevention and rehabilitation, monitoring and follow-up, and end-of-life care (Figure 2; Multimedia Appendix 3). Each included record was mapped to ≥1 phase depending on where in the journey the intervention (AI, CC, or other DHTs) was evaluated or deployed; multiphase studies were multicoded. The visual representation does not imply causation; it provides placement along the pathway to support descriptive synthesis in the Results and higher-order interpretation in the Discussion.

To interpret plausible causal pathways and specify testable hypotheses for LSD care, we also drew on evidence from adjacent non-LSD clinical domains where AI (eg, diagnostic triage and risk prediction) and CC (eg, televisit, teleconsultation, remote follow-up and monitoring adoption, and adherence) have been more extensively studied. These analogies were cited only as contextual signals (not pooled with LSD-specific data) and used to motivate prospective evaluations (eg, multisite designs, external validation, and standardized outcomes) in LSD settings.

Along the charting process, risks, challenges, and biases associated with DHTs, evidence limitations, and future research needs were identified and reported in the Results and Discussion sections. Key examples are also summarized in Multimedia Appendix 4.

Results

Overview

After the identification of 1751 records, we applied a rigorous selection process adapted to the nature of this multidimensional scoping review and mapping framework (PRISMA-ScR flow in Figure 3). The screening first led to 1225 items after removing duplicates and ineligible entries. The following screening process on titles and abstracts led to the exclusion of 437 records that did not meet relevance criteria coherent with the review question and eligibility criteria. A detailed assessment was then conducted on 788 records. Thorough eligibility assessment ensured each document was relevant to the specific focus, by embracing both LSDs management and DHTs integration, resulting in 245 records being included for final consideration: 182 research articles, 44 reviews, 19 “grey” literature records (including 7 ongoing or unpublished registered RCTs and other sources such as preprints, thesis, book chapters, periodicals, reports, and web posts). Finally, within the peer-reviewed literature, 40 items focused on AI-driven DHTs, 89 on CC DHTs, and 144 on other enabling DHTs, with some reporting more than one category. Nearly all records were published in English, only a few in Italian, German, and Russian.

In addition to LSD-focused records, the searches surfaced many articles on digital health applications in other rare and non-rare conditions. A subset of 492 journal articles and 51 pieces of gray literature was consulted only for contextual illustration. However, these contextual references were not systematically reviewed and did not influence study selection, data extraction, or synthesis.

Data were summarized in a PRISMA-ScR flow diagram (Figure 3), reporting identification, screening, eligibility, and inclusion counts for the corpus. Numbers correspond to records after deduplication and reflect all intervention classes (AI, CC, other DHTs). Reasons for exclusion at full text are summarized in the right-hand box and detailed in the Methods section. Transparency is provided via full search strings, logs, and the PRISMA-ScR checklist (Multimedia Appendix 1).

Expanded study-level evidence charting tables, showing distribution across literature sources, and comparing population characteristics, DHT categories, and care-journey mapping, are provided in Table 1 and Multimedia Appendix 5.

XSL•FO
RenderX

Figure 3. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) flow diagram adapted for a multidimensional scoping review and evidence-mapping framework [9]. AI: artificial intelligence; DHT: digital health technology; LSD: lysosomal storage disorder; RCT: randomized controlled trial.

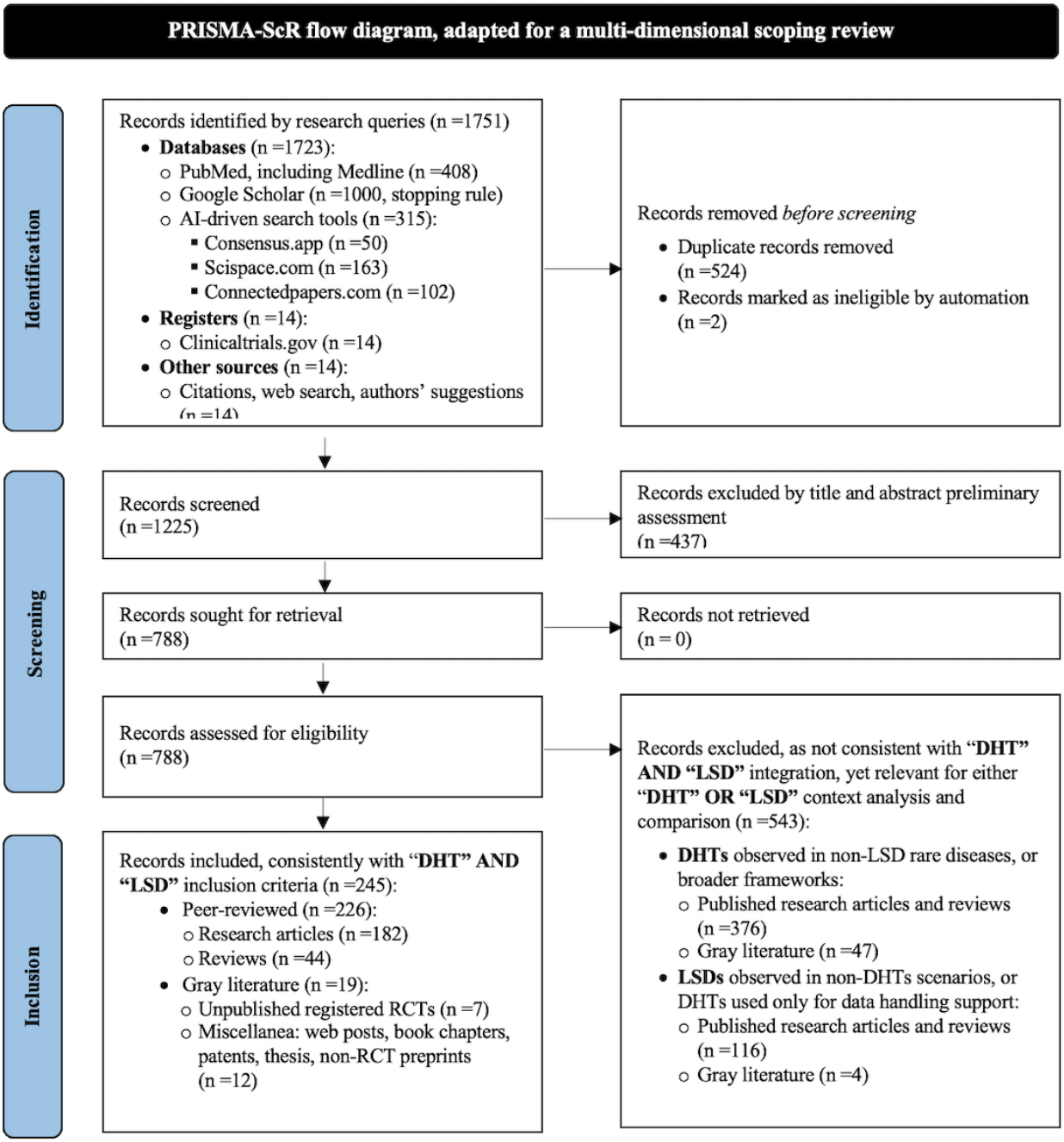

A synthesis at-a-glance of literature source distribution (Table 1) shows that, across the 245 included records, more than 90% (226/245) are peer-reviewed journal items and 7.8% (19/245) are gray literature (registered or ongoing trials, theses, reports, web-based documents, and other non–peer-reviewed sources). No completed RCTs, systematic reviews, or meta-analyses with LSDs-specific digital interventions were identified. Instead, the corpus is dominated by original observational, nonrandomized research and narrative or scoping reviews, with a small tail of case reports and contextual gray sources. Consistent with our scoping aim, peer-reviewed studies form the core basis for interpretation, whereas gray sources are used descriptively to map context and evidence gaps.

Disease-level comparative labeling (Table S1 in Multimedia Appendix 5) shows that evidence is unevenly distributed across LSDs. GD and FD together account for nearly 60% (130/226) of peer-reviewed papers, with additional clusters in mucopolysaccharidosis and PD, and only sparse coverage for other LSDs. To avoid overgeneralization, our synthesis is weighted by this coverage: where an LSD has multiple, convergent peer-reviewed studies, we treat findings as relatively more robust; where counts are very low, we label insights as signal-only and refrain from extrapolating cross-disease claims. This approach preserves disease-specific detail while explicitly acknowledging both sounder evidence and gaps.

XSL•FO
RenderX

Standardized intervention labeling (Table S2 in Multimedia Appendix 5) yielded 380 labels across AI, CC, and other DHTs. Although AI-specific labels represent only about 14% (52/380) of all intervention labels, they appear in approximately 18% (40/226) of peer-reviewed papers, largely in diagnostic and clinical-decision support applications. CC labels account for 32% of all labels (122/380), with DMDs and telemedicine applications most frequently represented. The remaining half of labels fall under other, mostly enabling DHTs (eg, omics platforms, bioinformatics pipelines, and non-AI imaging tools), which provide critical infrastructure. Since multicategory records were multicoded, labeled-paper totals may exceed; we therefore interpret these distributions qualitatively as relative signal strength rather than prevalence estimates.

Outcome-domain labeling (Table S3 in Multimedia Appendix 5) yielded 430 outcome labels across 245 records. Roughly two-thirds of labels (279/430, 64.9%) fall within health care delivery outcomes, led by technical diagnostic accuracy (157/430, 36.5%), with cost and organizational performance plus system and data security together contributing a further 75/430 (17.4%) labels. Patient-level outcomes account for just over one quarter of labels (117/430, 27.2%), dominated by patient experience (45/430, 10.5%) and privacy (39/430, 9.1%), whereas patient safety, awareness, and education remain relatively infrequent (33/430, 7.7% combined). Societal outcomes are least represented, with only 34 of 430 labels (7.9%), 10 of 430 (2.3%) for public health and social determinants of health, and 24 of 430 (5.6%) for regulatory and ethical compliance. Technical diagnostic accuracy alone accounts for more than one-third of all outcome labels and about half of peer-reviewed outcome reports, underscoring a strong focus on accuracy, detection, and classification—for now, more than on clinical efficacy and long-term patient or societal impact. This distribution reflects an evidence landscape characteristic of emerging innovations, in which establishing technical reliability and safety benchmarks precedes the generation of robust data on clinical effectiveness, patient-centered outcomes, and broader societal impact.

Care-journey mapping (Table S4 in Multimedia Appendix 5) shows that nearly half of all phase labels (222/472, 47%) concentrate in secondary prevention (screening) and diagnosis, followed by monitoring and follow-up (116/472, 24.6%) and treatment (95/472, 20.1%), while preconception primary prevention accounts for only 11/472 (2.3%) labels. Rehabilitation, tertiary prevention, and end-of-life care together contribute 28/472 (5.9%) labels, indicating much slimmer digital activity in later or more complex phases of the care journey. Across phases, other DHTs dominate diagnostic and treatment phases (eg, 120/222 labels, 54.1%, in secondary prevention and diagnosis), whereas CC labels are relatively more prominent in monitoring and follow-up (47/116, 40.5%), and AI labels remain numerically modest in downstream phases. These patterns, aligned with the timed framework in Figure 2, highlight a key digital footprint and exploration of applications to enhance early diagnosis and ongoing monitoring, with comparatively fewer digital interventions evaluated in primary prevention and in later phases of the care journey. The subsequent Results subsections unpack these phase-specific signals descriptively, while the Discussion interprets their implications for evaluation and implementation.

Population-Specific Insights by Disease

Coverage Patterns Across LSDs

As shown in Table S1 in Multimedia Appendix 5, disease coverage is highly asymmetric: FD and GD account for >50% of peer-reviewed records, whereas most other LSDs have limited or no DHT literature. We therefore synthesize findings by LSD and avoid cross-disease inference when evidence is sparse.

Although age was not charted as a separate axis, several records describe distinct digital use-cases in pediatric vs adult cohorts. In pediatric care, tools emphasize early detection (eg, neonatal blood spot screening and facial recognition), adherence and progression monitoring, and child- and family-centered education (including virtual reality [VR] and gamified approaches), alongside concerns about limited evidence, access inequities, and privacy [13-20]. In adults, mHealth apps, portals, and telemedicine emphasize self-management and access, while analytics support personalized planning and detection of late-onset manifestations [21-23]. Data scarcity, heterogeneity, and privacy governance recur across age groups [13-23].

Fabry Disease

In FD, AI use-cases cluster around early detection, phenotyping and organ involvement: facial recognition and image analytics (eg, magnetic resonance imaging [MRI], urinary sediment imaging, electrocardiography, echo, and MRI) to detect brain, renal, and cardiac manifestations, and support prognosis, plus EHR-based ML to flag high-risk patients for screening [24-33]. Variant-level prediction of pharmacologic responsiveness from structural and proteomic features is also reported [34-37].

CC and enabling tools support longitudinal management through telemedicine, DMDs, ePROs, registries, and EHRs for symptom tracking, adherence support, and remote follow-up, including app-based support programs integrating wearable or patient-entered data [22,38-43]. Enabling diagnostics include non-AI EHR risk models, digital microfluidics (DMF) + tandem mass spectrometry (MS/MS) newborn screening, and next-generation sequencing (NGS)–based confirmation and precision workflows [14,44-47]. Reported risks include overreliance on digital outputs and discordance with patient experience and clinical status [38,48,49].

Gaucher Disease

In GD, AI is most often applied to imaging-based phenotyping (eg, MRI-derived bone complication prediction; automated spleen-volume estimation) and to EHR and registry analytics supporting risk stratification and screening [8,22,44,50-59]. CC components include televisits, teleconsultations (highlighted during COVID-19) and, in some records, wearables, mHealth, and ePROs for intervisit monitoring of gait, sleep, pain, and stress [6,7,22,60,61]. Enabling diagnostics include genomic-analysis workflows and DMF-based assays supporting mutation detection and newborn-screening pathways [62-64]. Reported limitations include small datasets, bias risk, and limited evidence on long-term impact [56].

XSL•FO
RenderX

Mucopolysaccharidosis

For mucopolysaccharidosis, records describe ML-supported differential diagnosis and prediction of treatment response, complemented by CC (telemedicine and tele-genetics) for remote consultation, adherence monitoring, and education [22,60,65-67]. Enabling tools include genomic pipelines for mutation detection and DMF platforms for diagnostic assays [47,68]. Reported challenges include data integration, regulatory compliance, and evolving practices that require ongoing updating of digital tools [67,69].

Pompe Disease

For PD, AI-enabled tools are reported for EHR-based case finding and progression prediction, and for variant, enzymatic analytics within multiomics research pipelines [35,70-73]. CC applications include telemedicine-supported physiotherapy supervision and cognitive training, plus mHealth trackers and ePROs to monitor neuromuscular function and correlate patient-reported metrics with clinical outcomes, particularly in late-onset PD [60,74-79]. Genomic workflows support mutation identification and personalization [47,80]. Reported implementation barriers include technical and regulatory constraints, costs, fragmented care pathways, and variable user acceptance [74,75,81].

Other LSDs

Beyond FD, GD, mucopolysaccharidosis, and PD, evidence is sparse and typically disease-specific.

In Niemann-Pick disease, AI-driven tools and metabolomic profiling enhance diagnostic accuracy, and DMF platforms streamline diagnostic assays [68,82,83].

Computational tools enhance newborn screening in Krabbe disease by improving specificity and reducing costs, effectively differentiating true-positive from false-positive cases [84].

In MLD, AI algorithms aid in quantifying demyelination load in MRI scans [85,86] and simulate disease mechanisms to predict drug efficacy before clinical trials [87].

Telehealth services improve health care access and satisfaction for patients with Batten disease, enhancing care delivery and potential long-term benefits [20].

MRI scoring and quantitative spectroscopy can be used as surrogates of clinical disease progression in GM1G, as they correlate with ambulation and expressive language gradual impairment [88]. Wearable devices and ePROs may monitor patients with GM2G remotely, assessing health aspects like fatigue, stress, and self-esteem [89].

ML analysis of thalamus MRI images may diagnose aspartylglucosaminuria in patients with neurocognitive symptoms like progressive learning disabilities, with low error rates and promising diagnostic potential [90].

Intervention-Specific Insights by AI-Driven DHTs

Overview

As summarized in Table S2 in Multimedia Appendix 5, intervention subtypes by record, AI-driven DHTs constitute one of the 3 main classes of interventions mapped in this review, with studies clustering predominantly around diagnostic decision support, image and signal analysis, risk prediction, and data-driven patient stratification. The subsections below provide illustrative examples of how AI is being applied along LSD care pathways, highlighting typical use-cases rather than attempting a formal performance ranking.

AI for Risk Prediction and Patient Stratification

ML models identify patients at risk of developing specific diseases, predict the risk of complications, and stratify patients based on disease severity. This is generally achieved by analyzing patient demographics, clinical features, and genetic information [30,56,59,71].

AI for Clinical Diagnostics and Therapeutics

ML models identify potential LSD drug candidates by screening compound libraries and matching data against protein or genomic patient-specific data. They optimize treatment strategies, such as enzyme replacement therapy (ERT), and predict the responsiveness of individual variants to specific treatments [8,28,35,37,52,56]. AI-powered DSS shows use in metabolic conditions where overlapping symptoms often cause diagnostic delays. Intelligent systems may achieve high accuracy (87.4%) in differential diagnosis of LSDs by generating a list of diagnostic hypotheses based on clinical data [16,91,92].

DL algorithms, as a subset of ML that uses artificial neural networks with multiple layers to analyze complex patterns, are widely used for analyzing medical images (eg, brain or cardiac MRI and computed tomography, echocardiography, electrocardiogram, tissue or blood smears, facial features) to detect clinical biomarkers, structural anomalies, and disease-specific LSD patterns. For example, AI applications improve early detection of spleen and liver enlargement [54] or osteonecrosis in GD [51-53], or brain [26,90] and cardiac [28-32,34,93-95] involvement, mostly in FD and PD. AI-driven facial recognition software, used in diagnosing GD and FD, detects distinctive facial features commonly presented in large patient cohorts with early diagnosis potential [24,96,97].

NLP and LLMs extract critical information from unstructured data sources, such as clinical notes, interviews, research papers, insurance claims, and other textual sources. These insights can be integrated into AI-powered symptom assessment or DSS, informing diagnostic and treatment support for LSDs, particularly where clinical expertise is limited or symptoms are challenging to diagnose. NLP enables computers to understand and process human language, allowing them to extract meaningful insights from textual data, identifying patterns and correlations. LLMs have demonstrated remarkable capabilities in understanding and generating human-like text, summarizing patient narratives, extracting key medical concepts from clinical notes or visits, and generating comprehensive reports [33,98,99].

AI for R&D

There are already significant applications of ML and DL models in research on LSDs, including disease modeling, clinical trial simulations, drug discovery, and data analysis. ML tools identify disease-relevant compounds, predict their potential as inhibitors or chaperones, and link them to disease targets, advancing R&D

XSL•FO
RenderX

efforts [37,72,73,100,101]. DL and facial recognition software aid in the investigation and analysis of facial phenotypes in patients with LSDs, contributing to the understanding of these RDs [21,102]. Researchers use AI and simulations to study lysosomal enzymes and their role in storage diseases. These tools may speed up drug development, especially for gene therapies and targeted treatments. AI may optimize preclinical trials and simulate disease progression and treatment response, predict how diseases may evolve, explore new treatments, and simulate genetic mutations and their effects, aiding in establishing standardization and personalization workflows.

Intervention-Specific Insights by CC DHTs

Overview

As summarized in Table S2 in Multimedia Appendix 5, CC solutions are most often used to extend specialist input across distance, support long-term follow-up, and coordinate multidisciplinary care around complex, chronic trajectories. The narrative examples below illustrate how teleconsultation, home-based and hybrid models of care, and registry- or platform-based coordination are being deployed in practice, while acknowledging that formal comparative, cost-effectiveness, and equity analyses remain limited.

Telemedicine

Televisits among providers and patients (HCP-patient) enhance LSD care delivery. Evidence highlights the key role televisits played during the COVID-19 pandemic when in-person consultations were limited [6,15,17,39-41,60,74,103-112].

Platforms supporting virtual visits, genetic counseling (tele-genetics), and real-time monitoring have proven invaluable for maintaining continuity of care, particularly for patients in remote or underserved areas. For instance, telemedicine ensured that patients receiving ERT for conditions like GD and FD continued to receive necessary follow-ups, reducing “missed appointments” rates. Tele-genetics has specifically improved access to genetic testing and counseling, enhancing screening, early diagnosis, and treatment planning [20,43,66,67,75,113].

Tele-consultations among HCPs and collaborative health care networks enable multidisciplinary teams, including geneticists, pediatricians, cardiologists, neurologists, nephrologists, and primary care providers, to discuss complex cases. This collaborative approach has been crucial for LSDs like GD and FD, where care often requires input from various specialists, and nonspecialized professionals may need support from expert centers to enhance early diagnosis or continuous care. European Reference Networks already foster multidisciplinary care for RDs and, specifically, for LSDs through cross-border collaboration and could be enhanced by telemedicine and digital health applications [105,114-116].

Digital Medical Devices (DMDs)

Certified medical device software, such as remote patient monitoring devices, including medical-grade wearables and mHealth apps supported by proof of clinical efficacy, allows for reiterative tracking of real-world inter-visit biomarkers, like heart rate variability, respiratory function, mobility (including gait and fatigue patterns), providing early warnings for disease progression and symptom exacerbations. Examples include monitoring cardiovascular health for FD and activity trackers to help measure physical function as a surrogate for disease progression in GD, FD, PD, and G2MG [7,43,61,76-78,89].

Monitoring of ePROs has also demonstrated efficacy in collecting long-term data from patients with RDs, providing HCPs with detailed insights into daily symptom variations. For instance, patients with FD use ePROs to report neuropathic pain symptoms, enabling clinicians to timely adjust pain management protocols [43,49,61,78,79,89,117-121].

The adoption of digital patient care support programs, often delivered through mHealth applications, correlates with better adherence to management and treatment plans. They empower patients to take an active role in managing their care, leading to improved outcomes. Studies have shown that mHealth apps enhance patient experience, medication adherence, and help track daily symptoms, which are crucial for the long-term care of patients with LSD [7,41,43,61,76-78].

Safe and effective DTx [122] may enhance care through digitally-delivering treatments for common LSD symptoms and comorbidities, in particular, psychological symptoms, such as depression and sleeping disorders [123]. Digitally delivered cognitive behavioral-based interventions are already recommended in recent depression and insomnia treatment guidelines [124,125] and in Germany, as of December 2025, 29 DTx for mental health are already reimbursed and can be prescribed, 3 of which for sleeping disorders [126].

Patient Engagement Tools

Patient portals and online patient networks and communities have fostered greater patient engagement, particularly for RDs like LSDs. In managing long-term conditions like FD, patients who regularly access portals to review their health records and appointments are more likely to adhere to their treatments [8,57,117]. While it is important to ensure correct and up-to-date information, social platforms have created spaces for patients with LSDs to share experiences, discuss treatment side effects, and find emotional support, which in turn improves overall adherence to therapy.

Health Care Data Integration Systems and Disease Registries

Shared EHRs, HIS, and patient registries enhance RDs’ care and research, and enhance multidisciplinary collaboration across care teams and disease networks. In LSD management, EHRs allow geneticists, nephrologists, and other specialists to collaborate on treatment plans, streamlining the care process and improving decision-making. Research highlights that EHRs reduce the likelihood of miscommunication, which is critical for patients requiring long-term monitoring and multiple specialists [44,55,58,70,71].

Integrated international registries, often reported among the sources of published studies, have contributed significantly to understanding disease progression and treatment response in LSDs. These interfaces facilitated data sharing across health care institutions and HCPs, improving research into RDs by

XSL•FO
RenderX

offering large-scale epidemiological data on disease prevalence and long-term outcomes [2,22,38,59,118,127-134].

Intervention-Specific Insights by Other DHTs

Beyond AI and CC, other enabling DHTs underpin much of LSD diagnostics and research and often supply the data and infrastructure that later feed AI or CC applications (Table S2 in Multimedia Appendix 5). Key examples include: high-throughput biochemical diagnostics (DMF and advanced MS/high-performance liquid chromatography-MS/MS workflows), genomics and pharmacogenomics pipelines, bioinformatics and multiomics platforms, digitized imaging and 3D modeling, VR and augmented reality education, selected cognitive and rehabilitation use-cases, non-AI DSS, secure data architectures (including blockchain) for registry- and trial-enabling data sharing, and computational and trial-simulation models for progression and study-design optimization [10,13,16,42,46,47,62,64,68,72,73,80,83-87,92,100,101,135-156].

Outcome-Specific Insights by Domain

Overview

Table S3 in Multimedia Appendix 5 summarizes how outcomes were labeled across the 245 records, grouped into patient, health care delivery, and societal domains. Across all intervention classes, 65% of labels fall within health care delivery outcomes, 27% within patient outcomes, and fewer than 10% within societal outcomes. The following subsections provide a detailed, outcome-based synthesis of the results, grouped across the 3 predefined key domain areas.

Patient Outcome Domain

At the patient level, labeled outcomes mainly address safety, privacy, experience awareness, and education, and are less frequently represented than health care delivery outcomes. AI-enabled predictive analytics, telemedicine, and remote-monitoring DMDs are typically assessed for their ability to anticipate deterioration, flag early warning signs, and reduce avoidable unscheduled visits or admissions in chronic LSD manifestations (eg, cardiomyopathy and skeletal complications) [41,55,61]. Patient experience is usually captured through satisfaction, convenience, perceived access to expertise, usability of telehealth and mHealth tools, while privacy, awareness, and education outcomes appear in studies on consent, information provision, and trust in digital platforms [43,49,60,61,63,78,79,103,108,118,120,157,158].

Observational studies and pilots report that AI-driven risk-prediction tools, telemonitoring platforms, and wearable DMDs are used to identify early signals of deterioration or treatment-related complications, with outcomes reported as fewer emergency contacts, earlier detection of complications, or improved adherence to safety-monitoring protocols [7,41,55,61]. However, safety signals are predominantly process-oriented (alerts generated; actions triggered) rather than hard clinical endpoints, and sample sizes remain small. Reported risks include alert fatigue, overreliance on algorithm outputs, and missed events due to connectivity or data-quality failures [20,41,61].

Privacy outcomes are mostly described through technical and governance safeguards (encryption, access control, and occasionally blockchain-based architectures for registry and genomic data) and through documentation of GDPR-aligned policies, consent capture, and cross-border data-flow governance [63,103,108,152,153,157]. Direct assessment of privacy perceptions is less common, but some reports note that transparency on data use and clear governance arrangements support sustained trust in AI and CC deployments [21,60,69,159,160].

Telemedicine visits, home-based ERT programs, mHealth apps, wearables, VR tools, and online communities are frequently evaluated via patient experience and satisfaction endpoints, with commonly reported benefits including reduced travel burden to expert centers, perceived continuity of specialist input, and more flexible scheduling and follow-up [43,60,61,76-79,118]. Remote monitoring and digital support programs are also linked to perceived gains in self-management and shared decision-making, while digital literacy, connectivity, and device access can limit uptake and drive preference for hybrid models combining in-person care with digital touchpoints [20,60,69,107].

Digital support programs, disease-specific portals, and AI-enhanced educational tools (eg, symptom-assessment apps and interactive e-clinics) deliver tailored information on disease mechanisms, treatment options, and self-management; outcomes typically include self-reported improvements in knowledge, confidence, and engagement with care plans [20,33,99,107,116,117,158]. Online communities and social media groups may add peer-to-peer support, but raise concerns about variable information quality and the need for expert moderation [69,117].

Health Care Delivery Outcome Domain

Health care delivery outcomes account for nearly two-thirds of all outcome labels, dominated by technical diagnostic accuracy, clinical efficacy, system and data security, cost and organizational performance, and health care provider experience and education. AI, CC, and other DHTs are most evaluated on how they improve detection and classification, streamline workflows, support multidisciplinary coordination, and integrate data across settings [46,51-54,68,70,80,97,102,128,134,161].

Security outcomes are reported mainly for registries, EHR-linked algorithms, telemedicine platforms, and blockchain-based prototypes [63,70,103,108,128,152,153]. Evaluations focus on adherence to security standards (encryption, authentication, and audit trails), resilience to cyber-risks, and governance arrangements for access control and data sharing. Some registry and blockchain initiatives explicitly frame security and privacy as prerequisites for cross-border research collaboration and patient trust, but empirical testing of security robustness is rarely detailed beyond high-level descriptions [69,128,152,153].

Several reports and value-based care initiatives describe gains in organizational performance from telemedicine, home-based ERT, digital registries, and automated data flows, including reduced travel and visit burden, fewer missed appointments,

XSL•FO
RenderX

and perceived efficiencies in multidisciplinary coordination [20,23,50,76,116,128,134,161]. Cost-related outcomes are reported as descriptive estimates of resource savings (eg, avoided hospital visits and reduced inpatient days) or as qualitative assessments by clinicians and managers. Up-front investment needs, maintenance costs, and training requirements are also highlighted, particularly in under-resourced settings [23,69,159,161].

HCP experience outcomes concern the usability and perceived value of teleconsultation networks, interoperable HER and registry platforms, and DSS (AI-enabled or rule-based) [16,99,116,139,146,149,150]. Reported benefits include faster access to expert opinions, reduced professional isolation, and more structured approaches to complex diagnostic workups. At the same time, several studies note workflow disruption, documentation burden, and uncertainty around medicolegal responsibility as barriers to sustained adoption, especially when digital tools are introduced without sufficient co-design and training [116,149,162,163].

Digital education platforms, simulation tools, and decision-support interfaces are used to disseminate guidelines, RD curricula, and case-based learning for LSDs and other RDs [116,139,146,149,158]. Outcomes commonly include self-reported changes in knowledge, confidence, or diagnostic accuracy in simulated scenarios. Evidence on long-term behavior change or patient-level impact is limited, but these initiatives are frequently cited as enablers for broader DHT uptake and for strengthening expertise in non-specialist settings.

Societal Outcome Domain

Societal outcomes are least represented and cluster into public-health and social-determinants impacts, and regulatory and ethical compliance, appearing mainly in policy reports, registry initiatives, and mixed-methods studies on access, equity, and governance [1,60,69,153,160]. Public health endpoints are described primarily for telemedicine, networked registries, and outreach programs (eg, geographic reach, referral patterns, and continuity of care during disruptions), with digital divides in connectivity, device access, and digital literacy as recurrent constraints [1,6,60,69,105,109,130,153]. Regulatory and ethical outcomes focus on alignment with evolving frameworks for AI, medical devices, HTA, and data protection (including EU AI Act and national DTx pathways), emphasizing transparency, consent, privacy, and postmarket oversight while noting limited empirical evaluation of regulatory impact [9,21,38,122,126,128,152,153,160,162-164].

Outcome-Specific Insights by Care Journey Phase

Overview

Table S4 in Multimedia Appendix 5 and Figure 2, integrated by Multimedia Appendix 3, summarize how AI, CC and other DHTs applications and relative research are distributed along the LSD care journey; the paragraphs below provide a descriptive narrative of how these technologies are deployed across primary prevention, secondary prevention and diagnosis, treatment, monitoring and follow-up, rehabilitation, tertiary prevention, and end-of-life care, without inferring comparative effectiveness.

Primary Prevention

Primary prevention evidence is sparse and centers on preconception genetic risk identification. Genomic screening pipelines (including NGS, pharmacogenomics, and related digital platforms) and, in some reports, AI-enabled analytics support risk assessment and counseling, including via tele-genetics [14,45-47,62-64,66,68,80,84,165,166]. Ethical considerations include privacy and genetic discrimination risks.

Secondary Prevention and Diagnosis

Across the included studies, AI is used mainly for diagnostic support and phenotyping (eg, facial analysis; imaging analytics across MRI, CT, ultrasound, and smears; EHR-based case finding and severity prediction) [16,24,26,31,33,70,91,92,97-99,102,167]. A smaller set of papers reports NLP and LLM approaches to extract clinically relevant signals from unstructured text to support diagnostic reasoning [33,98,99]. CC solutions complement this pathway through televisits or teleconsultation and EHR-linked data integration, including ePRO and mHealth streams that support longitudinal assessment and diagnostic refinement, particularly for geographically dispersed families [6,20,40,43,60,61,66,76,89].

In secondary prevention and diagnosis, digital tools cluster around high-throughput screening, confirmatory pipelines, and decision support. DMF and related biochemical screening technologies underpin prenatal and neonatal screening and can accelerate early identification of disorders such as FD and GD [14,45]. Genomic workflows (NGS and other omics pipelines) support mutation detection or confirmation, and precision profiling, supplying structured data streams to downstream analytics [46,47,80,168].

Treatment

Treatment-phase applications emphasize personalization and longitudinal adjustment. AI-enabled analytics leverage genomic, proteomic, and metabolomic inputs to support therapy tailoring and to model treatment responsiveness, with reports spanning compound discovery and variant-level prediction [8,28,35,36,72]. CC components (televisits, mHealth, and teleconsultation) support timely initiation and adjustment of therapies such as ERT or substrate reduction, facilitate follow-up and multidisciplinary coordination, and perform continuous-care models integrating ePROs and biometric data for responsive management in conditions such as PD and FD [22,39-41,60,74,76,169].

Monitoring and Follow-Up

For long-term follow-up, CC and monitoring DHTs enhance assessment from episodic visits to inter-visit real-world data. Remote-monitoring DMDs, wearables, ePROs, and digital patient support programs are used to track multidimensional manifestations (eg, cardiac and respiratory function, mobility and gait, sleep, pain, and mood), enabling timely treatment adjustments and adherence support [7,40,43,61,76-78]. Telemedicine and teleconsultation platforms extend specialist oversight and multidisciplinary coordination, while EHR-linked integration (and, in some reports, AI-enabled analytics) consolidates wearable and clinical data into more holistic care

XSL•FO
RenderX

plans, particularly for chronic complications such as cardiomyopathy in FD and skeletal disease in GD [22,29,33,51].

Rehabilitation

Evidence in rehabilitation is limited, but suggests that mHealth apps, patient support programs, tele-rehabilitation platforms, and selected DTx may support at-home therapy and monitoring, sometimes integrating psychological support [61,76-78,118,123,148]. The literature emphasizes the need for clinical validation and implementation integration before these approaches can be scaled [69].

Tertiary Prevention

Tertiary-prevention records mainly describe remote monitoring tools (wearables and apps), and, in some cases, AI-enabled DSS to adjust care plans using real-world markers to prevent complications [71,76-78,89,128]. Blockchain is discussed to strengthen privacy and trust for long-term data sharing; equity, accessibility, and preservation of human contact are recurrent implementation considerations [38,69,103].

End-of-Life Care

End-of-life care evidence is limited, but suggests that telemedicine and remote monitoring can support palliative coordination by facilitating family and team communication and timely symptom-responsive interventions, with emphasis on preserving comfort, dignity, and ethical governance [6,20,43,61,66,109,110,170].

Discussion

Principal Findings

This scoping review shows a rich research landscape on digital health applications for LSDs care, although the overall quality and maturity of the evidence remain limited. Most studied digital interventions cluster upstream along the care journey, focusing on enhancing diagnosis and early monitoring, with comparatively fewer applications in treatment intensification, rehabilitation, tertiary prevention, and end-of-life care. DHTs are used to shorten the road to diagnosis and to stabilize follow-up, and less to support care interventions, where disability, caregiver burden, and health system management tend to peak. In the context of the LSDs' natural history and care journey phases (Figures 2 and 4), this pattern indicates that current evidence is more consistent on shortening diagnostic delays and optimizing surveillance, with a potential positive impact on years lived with disability and years of life lost.

Figure 4. Towards a digitally enhanced care journey for lysosomal storage diseases. HCP: health care provider; LSD: lysosomal storage disorder.

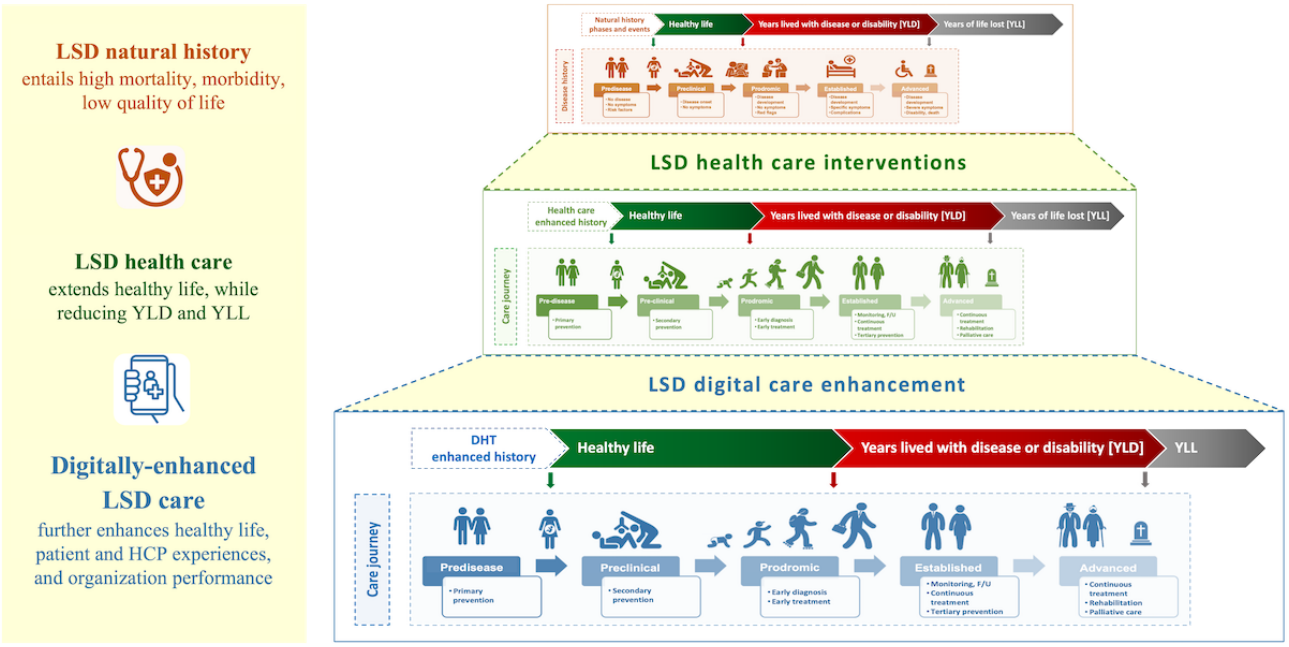

Within this landscape, CC tools anchor real-world delivery (telemedicine, remote patient monitoring, ePRO, and registries), AI is used mainly for diagnostic decision support and risk stratification, and enabling informatics (omics, bioinformatics, non-AI imaging analyses) underpin discovery pipelines. Coverage is uneven across diseases: evidence is concentrated in FD and GD, with only small clusters for other LSDs and little or no activity for many of the 70+ additional entities (Table S2 in Multimedia Appendix 5). To avoid overgeneralization, we therefore synthesized findings stratified by LSD, labeled very low-count areas as signal-only or evidence gaps, and prioritized them in the Future Research section.

Across the evidence base, most outcomes relate to health care delivery performance, particularly technical diagnostic accuracy and workflow coordination. Technical accuracy alone represents over one-third of all outcome labels and nearly half of peer-reviewed reports, indicating that DHT evaluations in LSDs remain centered on proximal process measures (how well tools detect, classify, or transmit information) rather than on longer-term clinical, patient-centered, or societal outcomes. This pattern is typical of an emerging innovation field: early studies prioritize technical reliability and feasibility, while robust assessments of clinical benefit, quality of life, HCP and patient experience, cost-effectiveness, equity, or broader public-health impact are still limited.

XSL•FO
RenderX

These patterns must be methodologically interpreted also within the constraints of the underlying evidence profile. Although title and abstract screening was primarily conducted by one reviewer (AMCS), we used calibration and second-reviewer (LN) audits; residual selection error remains possible. This review was explicitly descriptive and aimed to map what has been studied rather than to infer effectiveness. Across included records, signals for AI (eg, diagnostic support, early detection, and risk stratification) and CC (eg, televisits, teleconsultations, remote follow-up, and monitoring) are emerging, but the predominance of small, single-center, observational designs and the presence of gray literature limit generalizability and constrain inference. Consistent with the scoping-review methodology, we did not conduct a formal risk-of-bias assessment. Instead, we charted study-level features and caveats descriptively and used them to temper interpretation. We accordingly avoided causal language and treated findings as hypothesis-generating and as implementation cues to inform future, more robust evaluations (eg, multisite studies using transparent datasets, predefined comparators, and standardized outcome measures). Our interpretations prioritized peer-reviewed evidence, while policy and strategy documents were used as contextual background only; we did not advance causal or policy claims that were not supported by study design and reporting quality. Additional constraints include language and time-window limits and potential publication and selection biases typical of rapidly evolving digital scenarios.

Taken together, these outcome- and phase-level patterns delineate where evidence is accumulating versus where it is still largely hypothetical. They highlight a need for future evaluations that (1) move beyond accuracy and process metrics toward robust clinical efficacy measures, as well as patient-centered, system-level, and societal endpoints; (2) include later care-journey phases such as rehabilitation, transition to adulthood, and palliative care; and (3) explicitly measure equity, safety, and cost implications. The subsequent Discussion subsections on challenges, counterevidence, and future research interpret these gaps as priorities for staged, theory-informed evaluation rather than as a mandate for immediate large-scale implementation.

Challenges, Risks, and Biases

The evidence base also includes null or inconclusive findings, such as CC applications that did not improve adherence or use due to alert fatigue, workflow misalignment, or connectivity limitations; engagement platforms showing low sustained uptake; and AI models whose performance declined outside the derivation site or failed to outperform simpler baselines in small, heterogeneous cohorts [21,22,38,103,108].

Overall, Figures 2 and 4 suggest a functional split: AI is most often applied in early phases (screening, diagnosis, phenotyping), while CC supports continuity (follow-up, coordination, long-term monitoring). Given the largely descriptive evidence, deployment should be coupled with prospective, multisite, and equity-aware evaluations and transparent reporting [21,38,69,70].

AI-related risks reflect rare-disease data constraints: small single-center cohorts, incomplete or heterogeneous EHR data, missingness, and limited external validation drive overfitting and bias risk [21,38,69,70]. Proposed mitigation strategies include cross-registry pooling under findable, accessible, interoperable, and reusable principles, transfer learning, data augmentation or synthetic data, and routine auditing for bias and performance drift, aligned with emerging governance expectations (eg, EU AI Act) [103,164,171-173]. For CC deployments, the dominant concerns are interoperability with HIS and EHR, and equity (digital literacy, device, and connectivity access), such that CC can either bridge or widen the digital divide depending on implementation design [20-22,38,40,103,108,134,162,163].

Across included records, recurrent barriers to DHT scale-up in LSDs include infrastructure constraints, limited alignment across regulatory, HTA, and reimbursement requirements, and workforce skill gaps [38,128]. Interoperability, provider training, and robust privacy and security governance are repeatedly positioned as prerequisites, alongside patient trust, and clear accountability for data use [14,21,69,102].

Future Research Priorities

Given rapid innovation and evolving governance, multistakeholder consensus processes (eg, consensus conferences informed by PICO-structured evidence syntheses) can help translate emerging evidence into fit-for-purpose recommendations for development, adoption, and evaluation of AI and CC in LSD care and in RDs more broadly [35,36].

Most included evidence is nonrandomized, single-center, with small samples and limited follow-up, constraining causal inference and generalizability [37,71,72,80]. Priorities, therefore, include standardized outcome sets, interoperable data capture, and multisite designs (including pragmatic and, where feasible, randomized evaluations) that can assess longer-term effectiveness, safety, equity, and implementation outcomes [28,29,35,72,79,80]. In rare, geographically dispersed LSD populations, DHTs may need to act both as interventions and as trial enablers (supporting decentralized recruitment, continuous outcome capture, and longitudinal follow-up) [28,79].

Conclusions

This PRISMA-ScR–conformant review provides, to the authors' knowledge, the first comprehensive mapping of DHT applications in LSDs care, spanning disease populations, intervention classes, outcome domains, and care-journey phases. Across 245 records, research activity clustered around a subset of LSDs, particularly FD and GD, and was concentrated in earlier care-journey phases such as screening, diagnosis, and early monitoring, with markedly fewer applications reported in treatment intensification, rehabilitation, transition, or end-of-life care. Care-delivery outcomes (including diagnostic accuracy, workflow coordination, and access to specialist expertise) dominated reporting, whereas patient-centered and societal outcomes remained comparatively scarce. Figure 4 synthesizes how these digitally enhanced activities align with LSDs' natural history and conventional care pathways [24,102].

However, current evidence is insufficient to infer effectiveness or justify routine implementation. Across AI, CC, and other

XSL•FO
RenderX

DHTs, most studies were small, single-center, observational, and exploratory, with only a modest contribution from gray literature. Consistent with the aims of a scoping review, these findings should be interpreted as descriptive and hypothesis-generating. Our mapping is intended to characterize patterns and identify gaps, not to support causal claims or comparative-effectiveness judgments.

Considering these findings, near-term priorities include stronger evidence generation (through RCTs, high-quality real-world evidence studies, and pragmatic, theory-informed implementation trials of AI and CC technologies), using transparent datasets and standardized outcome sets aligned with established evidence standards. Given that the current evidence base is appreciable in size, but heterogeneous and largely observational, expert consensus processes are needed to translate emerging signals into appropriate, practice-relevant guidance and to define minimum requirements for safe and meaningful clinical use. Policy integration and cross-border collaboration are not conclusions drawn from the current data but represent necessary enabling conditions for high-quality data availability in RDs, harmonized governance, interoperability, and future evidence generation. The 4-axis mapping framework and care-journey perspective presented here may help guide the design of these future evaluations and consensus efforts for LSDs and, where appropriate, other RDs.

Acknowledgments

Consistently with *Journal of Medical Internet Research* Publications' guidance on generative artificial intelligence (GenAI), we disclose that GenAI tools, such as Consensus, SciSpace, and Connected Papers, were used to support operational refinement and expansion of the literature search strategy (eg, brainstorming and harmonizing candidate digital health technologies keywords later manually verified and incorporated by the authors) and for language-level editing and quality-control checks (grammar, clarity, translation, terminology harmonization, and internal consistency checks, including verification of internal cross-references to figures, tables, and multimedia appendices). These GenAI tools were not used for study selection decisions, data extraction, data analysis, evidence synthesis, interpretation, or formulation of conclusions. All suggested edits were reviewed, validated, and approved by the author team, who remain fully accountable for the manuscript content. The search prompts used are available in Multimedia Appendix 2 (AI-assisted search reproducibility logs).

Funding

The study was funded by Chiesi Global Rare Diseases. The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' Contributions

AMCS, LN, and MS conceptualized the article. AMCS and LN conducted the detailed literature review and analysis. AMCS provided expertise on digital health innovations and health policies, LN guided the methodology, and MS provided clinical expertise on lysosomal storage disorders and rare diseases. CB complemented the clinical perspective, and KA provided expertise from an international patient association network perspective. AMCS drafted the initial manuscript, which was integrated by LN, KA, and MS. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The 22-item PRISMA-ScR checklist for scoping reviews.
[DOCX File , 2397 KB - jmir_v28i1e73612_app1.docx]

Multimedia Appendix 2
Reproducibility logs of the scoping review search queries.
[DOCX File , 6639 KB - jmir_v28i1e73612_app2.docx]

Multimedia Appendix 3
Examples of AI-driven and Connected Care (CC) Digital Health Technology (DHT) applications to enhance Lysosomal Storage Disease care pathways across all stages.
[DOCX File , 3540 KB - jmir_v28i1e73612_app3.docx]

Multimedia Appendix 4
Examples of challenges for systemic integration of AI-driven and Connected Care (CC) Digital Health Technologies (DHTs).
[DOCX File , 4175 KB - jmir_v28i1e73612_app4.docx]

XSL•FO
RenderX

Multimedia Appendix 5
Comparative and distribution evidence charting tables across LSD population, digital health interventions, patient-healthcare-societal outcomes, and care journey phases.
[DOCX File , 4518 KB - jmir_v28i1e73612_app5.docx]

References

1. Adachi T, El-Hattab AW, Jain R, Nogales Crespo KA, Quirland Lazo CI, Scarpa M, et al. Enhancing equitable access to rare disease diagnosis and treatment around the world: a review of evidence, policies, and challenges. Int J Environ Res Public Health 2023;20(6):4732 [FREE Full text] [doi: 10.3390/ijerph20064732] [Medline: 36981643]
2. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet 2020;28(2):165-173 [FREE Full text] [doi: 10.1038/s41431-019-0508-0] [Medline: 31527858]
3. Rare diseases. European Commission. 2024. URL: https://health.ec.europa.eu/rare-diseases-and-european-reference-networks/rare-diseases_en [accessed 2026-01-01]
4. Tanpaiboon P. Practical management of lysosomal storage disorders (LSDs). Transl Sci Rare Dis 2019;4(3-4):133-157. [doi: 10.3233/trd-190047]
5. Addressing the challenges of persons living with a rare disease and their families : revised draft resolution / Brazil, Central African Republic, Côte d'Ivoire, Cyprus, Ecuador, Equatorial Guinea, France, Italy, Peru, Portugal, Qatar, South Africa, Spain, Ukraine and Vanuatu. United Nations Digital Library. 2021. URL: https://digitallibrary.un.org/record/3947524 [accessed 2026-01-01]
6. Donald A, Cizer H, Finnegan N, Collin-Histed T, Hughes DA, Davies EH. Measuring disease activity and patient experience remotely using wearable technology and a mobile phone app: outcomes from a pilot study in Gaucher disease. Orphanet J Rare Dis 2019;14(1):212 [FREE Full text] [doi: 10.1186/s13023-019-1182-6] [Medline: 31488169]
7. Hurvitz N, Dinur T, Revel-Vilk S, Agus S, Berg M, Zimran A, et al. A feasibility open-labeled clinical trial using a second-generation artificial-intelligence-based therapeutic regimen in patients with Gaucher disease treated with enzyme replacement therapy. J Clin Med 2024;13(11):3325 [FREE Full text] [doi: 10.3390/jcm13113325] [Medline: 38893036]
8. Spreafico AMC, Tarricone R, Stern AD. Comprehensive policies for scaling systemic and equitable integration of digital health technologies. NPJ Digit Med 2025;8(1):404 [FREE Full text] [doi: 10.1038/s41746-025-01820-x] [Medline: 40615530]
9. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). European Union EUR-Lex. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj [accessed 2026-01-01]
10. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473. [doi: 10.7326/m18-0850] [Medline: 30178033]
11. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005;8(1):19-32. [doi: 10.1080/1364557032000119616]
12. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
13. Millington DS, Sista R, Eckhardt A, Rouse J, Bali D, Goldberg R, et al. Digital microfluidics: a future technology in the newborn screening laboratory? Semin Perinatol 2010;34(2):163-169 [FREE Full text] [doi: 10.1053/j.semperi.2009.12.008] [Medline: 20207266]
14. Millington DS, Bali DS. Current state of the art of newborn screening for lysosomal storage disorders. Int J Neonatal Screen 2018;4(3):24 [FREE Full text] [doi: 10.3390/ijns4030024] [Medline: 33072946]
15. Shur N, Atabaki SM, Kisling MS, Tabarani A, Williams C, Fraser JL, et al. Rapid deployment of a telemedicine care model for genetics and metabolism during COVID-19. Am J Med Genet A 2021;185(1):68-72 [FREE Full text] [doi: 10.1002/ajmg.a.61911] [Medline: 33051968]
16. Kobrinskii BA, Blagosklonov NA, Demikova NS, Nikolaeva EA, Kotalevskaya YY, Melikyan LP, et al. A computer system for differential diagnosis of lysosomal storage diseases based on artificial intelligence methods [Article in Russian]. Bull Sib Med 2022;21(2):67-73. [doi: 10.20538/1682-0363-2022-2-67-73]
17. Aydemir S, Ocak S, Saygılı S, Hopurcuoğlu D, Haşlak F, Kıykım E, et al. Telemedicine applications in a tertiary pediatric hospital in Turkey during COVID-19 pandemic. Telemed J E Health 2021;27(10):1180-1187. [doi: 10.1089/tmj.2020.0381] [Medline: 33301373]
18. Bouwman MG, Teunissen QGA, Wijburg FA, Linthorst GE. 'Doctor Google' ending the diagnostic odyssey in lysosomal storage disorders: parents using internet search engines as an efficient diagnostic strategy in rare diseases. Arch Dis Child 2010;95(8):642-644. [doi: 10.1136/adc.2009.171827] [Medline: 20418338]
19. Scherr JF, Albright C, de Los Reyes E. Utilizing telehealth to create a clinical model of care for patients with batten disease and other rare diseases. Ther Adv Rare Dis 2021;2:26330040211038564 [FREE Full text] [doi: 10.1177/26330040211038564] [Medline: 37181116]

XSL•FO
RenderX

20. Uribe-Carretero E, Rey V, Fuentes JM, Tamargo-Gómez I. Lysosomal dysfunction: connecting the dots in the landscape of human diseases. Biology (Basel) 2024;13(1):34 [FREE Full text] [doi: 10.3390/biology13010034] [Medline: 38248465]
21. Elstein D, Giugliani R, Muenzer J, Schenk J, Schwartz IVD, Anagnostopoulou C. Impact of the COVID-19 pandemic on the standard of care for patients with lysosomal storage diseases: a survey of healthcare professionals in the Fabry, Gaucher, and Hunter Outcome Survey registries. Mol Genet Metab Rep 2021;28:100788 [FREE Full text] [doi: 10.1016/j.ymgmr.2021.100788] [Medline: 34367919]
22. Nuovo Monitor 47: Telemedicina e Intelligenza Artificiale a Supporto Dell'assistenza Territoriale. Vol II. Agenas [Web page in Italian]. AGENAS - Agenzia Nazionale per i Servizi Sanitari Regionali. 2022. URL: https://www.agenas.gov.it/monitor-47 [accessed 2026-01-01]
23. Nishith N, Siddiqui SH, R Raja SK, Agrawal N, Phadke S, Sharma S. Utility of morphologic assessment of bone marrow biopsy in diagnosis of lysosomal storage disorders. Indian J Pathol Microbiol 2023;66(1):91-95 [FREE Full text] [doi: 10.4103/ijpm.ijpm_725_21] [Medline: 36656217]
24. D'Angelantonio E, Tocci T, Sansone L, Mencattini A, Russo M, Pallotti A. Fabry disease automatic recognition through image processing. : IEEE; 2022 Presented at: Proceedings of the 2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA); June 22-24, 2022; Messina, Italy p. 1-3. [doi: 10.1109/memea54994.2022.9856583]
25. Montella A, Tranfa M, Scaravilli A, Barkhof F, Brunetti A, Cole J, et al. Assessing brain involvement in Fabry disease with deep learning and the brain-age paradigm. Hum Brain Mapp 2024;45(5):e26599 [FREE Full text] [doi: 10.1002/hbm.26599] [Medline: 38520360]
26. Uryu H, Migita O, Ozawa M, Kamijo C, Aoto S, Okamura K, et al. Automated urinary sediment detection for Fabry disease using deep-learning algorithms. Mol Genet Metab Rep 2022;33:100921 [FREE Full text] [doi: 10.1016/j.ymgmr.2022.100921] [Medline: 36186840]
27. Moura ARS. Detecting important electrocardiogram characteristics for the diagnosis of Fabry disease via statistics and machine learning techniques. ProQuest. 2022. URL: https://search.proquest.com/openview/372a62f0566d6dd4e2056eff49218b4b/1?pq-origsite=gscholar&cbl=2026366&diss=y [accessed 2026-01-01]
28. Jefferies J, Aguiar P, Biondetti G, Warnock D, Kallish S, Nelson M, et al. Estimation of arrhythmia risk in patients with Fabry disease using a machine learning model. J Heart Lung Transplant 2023;42(4):S331. [doi: 10.1016/j.healun.2023.02.765]
29. Vijapurapu R. Multi parametric cardiac magnetic resonance imaging and arrhythmic risk stratification in Fabry disease. University of Birmingham - Institute of Cardiovascular Sciences. 2023. URL: http://etheses.bham.ac.uk/id/eprint/13276/9/Vijapurapu2023PhD_Redacted.pdf [accessed 2026-01-01]
30. Akiki E, Abbasi M, El-Am E, Geske JB, Bos M, Ackerman MJ, et al. Artificial intelligence electrocardiography for the evaluation of cardiac involvement in Fabry disease. J Am Coll Cardiol 2024;83(13):2581. [doi: 10.1016/s0735-1097(24)04571-6]
31. Vijapurapu R, Nordin S, Baig S, Liu B, Rosmini S, Augusto J, et al. Global longitudinal strain, myocardial storage and hypertrophy in Fabry disease. Heart 2019;105(6):470-476. [doi: 10.1136/heartjnl-2018-313699] [Medline: 30282640]
32. Hwang IC, Chun EJ, Kim PK, Kim M, Park J, Choi HM, et al. Artificial intelligence-derived extracellular volume fraction for diagnosis and prognostication in patients with light-chain cardiac amyloidosis. medRxiv 2024 Mar 22. [doi: 10.1101/2024.03.20.24304642]
33. Beck M, Ramaswami U, Hernberg-Ståhl E, Hughes DA, Kampmann C, Mehta AB, et al. Twenty years of the Fabry outcome survey (FOS): insights, achievements, and lessons learned from a global patient registry. Orphanet J Rare Dis 2022;17(1):238 [FREE Full text] [doi: 10.1186/s13023-022-02392-9] [Medline: 35725623]
34. Woodard J, Zheng W, Zhang Y. Protein structural features predict responsiveness to pharmacological chaperone treatment for three lysosomal storage disorders. PLoS Comput Biol 2021;17(9):e1009370 [FREE Full text] [doi: 10.1371/journal.pcbi.1009370] [Medline: 34529671]
35. Tebani A, Mauhin W, Abily-Donval L, Lesueur C, Berger MG, Nadjar Y, et al. A proteomics-based analysis reveals predictive biological patterns in Fabry disease. J Clin Med 2020;9(5):1325 [FREE Full text] [doi: 10.3390/jcm9051325] [Medline: 32370284]
36. Tebani A, Barbey F, Dormond O, Ducatez F, Marret S, Nowak A, et al. Deep next-generation proteomics and network analysis reveal systemic and tissue-specific patterns in Fabry disease. Transl Res 2023;258:47-59. [doi: 10.1016/j.trsl.2023.02.006] [Medline: 36863609]
37. Michalski AA, Lis K, Stankiewicz J, Kloska SM, Sycz A, Dudziński M, et al. Supporting the diagnosis of fabry disease using a natural language processing-based approach. J Clin Med 2023;12(10):3599 [FREE Full text] [doi: 10.3390/jcm12103599] [Medline: 37240705]
38. Laney D, Germain D, Oliveira J, Burlina AP, Cabrera GH, Hong GR, et al. Fabry disease and COVID-19: international expert recommendations for management based on real-world experience. Clin Kidney J 2020;13(6):913-925 [FREE Full text] [doi: 10.1093/ckj/sfaa227] [Medline: 33391734]
39. Riccio E, Pieroni M, Limoneglli G, Pisani A. Impact of COVID-19 pandemic on patients with fabry disease: an Italian experience. Mol Genet Metab 2020;131(1-2):124-125 [FREE Full text] [doi: 10.1016/j.ymgme.2020.07.008] [Medline: 32741663]

XSL•FO
RenderX

40. Nowicki M, Bazan-Socha S, Kłopotowski M, Błażejewska-Hyżorek B, Kusztal M, Pawlaczyk K, et al. Considerations for home-based treatment of fabry disease in Poland during the COVID-19 pandemic and beyond. Int J Environ Res Public Health 2021;18(16):8242 [FREE Full text] [doi: 10.3390/ijerph18168242] [Medline: 34443990]
41. Study based on electronic health records to identify patients at high-risk of Fabry disease (HOPE Fabry). ClinicalTrials.gov. 2024. URL: https://clinicaltrials.gov/study/NCT05671770?cond=Lysosomal%20Storage%20Diseases&intr=electronic&rank=1 [accessed 2026-01-01]
42. D'Amore S, Mckie M, Fahey A, Bleloch D, Grillo G, Hughes M, et al. Fabry app: the value of a portable technology in recording day-to-day patient monitored information in patients with Fabry disease. Orphanet J Rare Dis 2024;19(1):13 [FREE Full text] [doi: 10.1186/s13023-023-02999-6] [Medline: 38212814]
43. Gragnaniello V, Burlina AP, Commone A, Gueraldi D, Puma A, Porcù E, et al. Newborn screening for fabry disease: current status of knowledge. Int J Neonatal Screen 2023;9(2):31 [FREE Full text] [doi: 10.3390/ijns9020031] [Medline: 37367212]
44. Implementation and evaluation of a rare disease algorithm to identify persons at risk of Gaucher disease using data from electronic health records (EHRs) in the United States (project searchlight). ClinicalTrials.gov. 2024. URL: https://clinicaltrials.gov/study/NCT05908656?cond=Lysosomal%20Storage%20Diseases&intr=electronic&rank=2 [accessed 2026-01-01]
45. Zanetti A, D'Avanzo F, Bertoldi L, Zampieri G, Feltrin E, De Pascale F, et al. Setup and validation of a targeted next-generation sequencing approach for the diagnosis of lysosomal storage disorders. J Mol Diagn 2020;22(4):488-502 [FREE Full text] [doi: 10.1016/j.jmoldx.2020.01.010] [Medline: 32036093]
46. Málaga DR, Brusius-Facchin AC, Siebert M, Pasqualim G, Saraiva-Pereira ML, Souza CFD, et al. Sensitivity, advantages, limitations, and clinical utility of targeted next-generation sequencing panels for the diagnosis of selected lysosomal storage disorders. Genet Mol Biol 2019;42(1 suppl 1):197-206 [FREE Full text] [doi: 10.1590/1678-4685-GMB-2018-0092] [Medline: 30985853]
47. Biegstraaten M, Arngrímsson R, Barbey F, Boks L, Cecchi F, Deegan PB, et al. Recommendations for initiation and cessation of enzyme replacement therapy in patients with Fabry disease: The European fabry working group consensus document. Orphanet J Rare Dis 2015;10:36 [FREE Full text] [doi: 10.1186/s13023-015-0253-6] [Medline: 25885911]
48. Berry L, Walter J, Johnson J, Alton J, Powers J, Llòria X, et al. Patient-reported experience with Fabry disease and its management in the real-world setting: Results from a double-blind, cross-sectional survey of 280 respondents. Orphanet J Rare Dis 2024;19(1):153 [FREE Full text] [doi: 10.1186/s13023-024-03090-4] [Medline: 38605390]
49. Kadali S, Naushad SM, Radha Rama Devi A, Bodiga VL. Biochemical, machine learning and molecular approaches for the differential diagnosis of mucopolysaccharidoses. Mol Cell Biochem 2019;458(1-2):27-37. [doi: 10.1007/s11010-019-03527-6] [Medline: 30903511]
50. Sharma GB, Robertson DD, Laney DA, Gambello MJ, Terk M. Machine learning based analytics of micro-MRI trabecular bone microarchitecture and texture in type 1 Gaucher disease. J Biomech 2016;49(9):1961-1968. [doi: 10.1016/j.jbiomech.2016.04.010] [Medline: 27109052]
51. Valero-Tena E, Roca-Espiau M, Verdú-Díaz J, Diaz-Manera J, Andrade-Campos M, Giraldo P. Advantages of digital technology in the assessment of bone marrow involvement in Gaucher's disease. Front Med (Lausanne) 2023;10:1098472 [FREE Full text] [doi: 10.3389/fmed.2023.1098472] [Medline: 37250646]
52. Yu B, Whitmarsh T, Riede P, McDonald S, Kaggie JD, Cox TM, et al. Deep learning-based quantification of osteonecrosis using magnetic resonance images in Gaucher disease. Bone 2024;186:117142 [FREE Full text] [doi: 10.1016/j.bone.2024.117142] [Medline: 38834102]
53. Azuri I, Wattad A, Peri-Hanania K, Kashti T, Rosen R, Caspi Y, et al. A deep-learning approach to spleen volume estimation in patients with Gaucher disease. J Clin Med 2023;12(16):5361 [FREE Full text] [doi: 10.3390/jcm12165361] [Medline: 37629403]
54. Wilson A, Chiorean A, Aguiar M, Sekulic D, Pavlick P, Shah N, et al. Development of a rare disease algorithm to identify persons at risk of Gaucher disease using electronic health records in the United States. Orphanet J Rare Dis 2023;18(1):280 [FREE Full text] [doi: 10.1186/s13023-023-02868-2] [Medline: 37689674]
55. Tenenbaum A, Revel-Vilk S, Gazit S, Roimi M, Gill A, Gilboa D, et al. A machine learning model for early diagnosis of type 1 Gaucher disease using real-life data. J Clin Epidemiol 2024;175:111517. [doi: 10.1016/j.jclinepi.2024.111517] [Medline: 39245415]
56. Revel-Vilk S. Digital health platform customized for patients with Gaucher disease. ClinicalTrials.gov. 2024. URL: https://clinicaltrials.gov/study/NCT04997772?cond=Lysosomal%20Storage%20Diseases&viewType=Card&intr=digital&rank=1 [accessed 2026-01-01]
57. Lobel CJ, Laney DA, Yang J, Jacob D, Rickheim A, Ogg CZ, et al. FDrisk: development of a validated risk assessment tool for fabry disease utilizing electronic health record data. J Rare Dis (Berlin) 2024;3(1):2 [FREE Full text] [doi: 10.1007/s44162-023-00026-7] [Medline: 38187171]
58. Andrade-Campos MM, de Frutos LL, Cebolla JJ, Serrano-Gonzalo I, Medrano-Engay B, Roca-Espiau M, et al. Identification of risk features for complication in Gaucher's disease patients: a machine learning analysis of the Spanish registry of Gaucher disease. Orphanet J Rare Dis 2020;15(1):256 [FREE Full text] [doi: 10.1186/s13023-020-01520-7] [Medline: 32962737]

XSL•FO
RenderX

59. Karabul N, Rose F, Cipra G. Home treatment for lysosomal diseases during COVID-19: German experience. Mol Genet Metab 2021;132(2):S54. [doi: 10.1016/j.ymgme.2020.12.119]
60. Elmonem MA, Belanger-Quintana A, Bordugo A, Boruah R, Cortès-Saladelafont E, Endrakanti M, et al. The impact of COVID-19 pandemic on the diagnosis and management of inborn errors of metabolism: a global perspective. Mol Genet Metab 2020;131(3):285-288 [FREE Full text] [doi: 10.1016/j.ymgme.2020.09.004] [Medline: 33004274]
61. La Cognata V, Guarnaccia M, Morello G, Ruggieri M, Polizzi A, Cavallaro S. Design and validation of a custom NGS panel targeting a set of lysosomal storage diseases candidate for NBS applications. Int J Mol Sci 2021;22(18):10064 [FREE Full text] [doi: 10.3390/ijms221810064] [Medline: 34576242]
62. Ryan E, Jong T, Sidransky E. Newborn screening in Gaucher disease: a bright and complicated future. OBM Genet 2022;06(03):1-21. [doi: 10.21926/obm.genet.2203165]
63. Hirachan R, Horman A, Burke D, Heales S. Evaluation, in a highly specialised enzyme laboratory, of a digital microfluidics platform for rapid assessment of lysosomal enzyme activity in dried blood spots. JIMD Rep 2024;65(2):124-131 [FREE Full text] [doi: 10.1002/jmd2.12413] [Medline: 38444576]
64. D'Souza A, Ryan E, Sidransky E. Facial features of lysosomal storage disorders. Expert Rev Endocrinol Metab 2022;17(6):467-474 [FREE Full text] [doi: 10.1080/17446651.2022.2144229] [Medline: 36384353]
65. Bajaj S, Thawani S, Iyer G, Setty PN, Kore M, Jadhav V. Telegenetics aids the diagnosis of hunter syndrome caused due to a novel IDS variant in rural India, during COVID-19 pandemic. Int J Contemp Pediatr 2022;9(9):851. [doi: 10.18203/2349-3291.ijcp20222126]
66. İnci A, Okur İ, Tümer L, Biberoğlu G, Öktem M, Ezgü F. Clinical and event-based outcomes of patients with mucopolysaccharidosis VI receiving enzyme replacement therapy in Turkey: a case series. Orphanet J Rare Dis 2021;16(1):438 [FREE Full text] [doi: 10.1186/s13023-021-02060-4] [Medline: 34666789]
67. Millington D, Norton S, Singh R, Sista R, Srinivasan V, Pamula V. Digital microfluidics comes of age: High-throughput screening to bedside diagnostic testing for genetic disorders in newborns. Expert Rev Mol Diagn 2018;18(8):701-712 [FREE Full text] [doi: 10.1080/14737159.2018.1495076] [Medline: 30004274]
68. Mistry PK, Kishnani P, Wanner C, Dong D, Bender J, Batista JL, et al. Rare lysosomal disease registries: lessons learned over three decades of real-world evidence. Orphanet J Rare Dis 2022;17(1):362 [FREE Full text] [doi: 10.1186/s13023-022-02517-0] [Medline: 36244992]
69. Lin S, Nateqi J, Weingartner-Ortner R, Gruarin S, Marling H, Pilgram V, et al. An artificial intelligence-based approach for identifying rare disease patients using retrospective electronic health records applied for Pompe disease. Front Neurol 2023;14:1108222 [FREE Full text] [doi: 10.3389/fneur.2023.1108222] [Medline: 37153672]
70. Rustamov J, Rustamov Z, Mohamad MS, Zaki N, Al Tenaiji A, Al Harbi M, et al. An expert rule-based approach for identifying infantile-onset pompe disease patients using retrospective electronic health records. Sci Rep 2024;14(1):21523 [FREE Full text] [doi: 10.1038/s41598-024-72259-5] [Medline: 39277702]
71. Gómez-Cebrián N, Gras-Colomer E, Poveda Andrés JL, Pineda-Lucena A, Puchades-Carrasco L. Omics-based approaches for the characterization of Pompe disease metabolic phenotypes. Biology (Basel) 2023;12(9):1159 [FREE Full text] [doi: 10.3390/biology12091159] [Medline: 37759559]
72. Esmail S, Danter WR. DeepNEU: artificially induced stem cell (aiPSC) and differentiated skeletal muscle cell (aiSkMC) simulations of infantile onset POMPE disease (IOPD) for potential biomarker identification and drug discovery. Front Cell Dev Biol 2019;7:325 [FREE Full text] [doi: 10.3389/fcell.2019.00325] [Medline: 31867331]
73. Brunetti-Pierri N, Fecarotta S, Staiano A, Strisciuglio P, Parenti G. Ensuring continuity of care for children with inherited metabolic diseases at the time of COVID-19: the experience of a metabolic unit in Italy. Genet Med 2020;22(7):1178-1180 [FREE Full text] [doi: 10.1038/s41436-020-0831-4] [Medline: 32409735]
74. Priya K. ClinicalTrials.gov. 2021. URL: https://clinicaltrials.gov/study/NCT02950298?cond=Lysosomal%20Storage%20Diseases&intr=telemedicine&rank=1 [accessed 2026-01-01]
75. Ricci G, Baldanzi S, Seidita F, Proietti C, Carlini F, Peviani S, et al. A mobile app for patients with Pompe disease and its possible clinical applications. Neuromuscul Disord 2018;28(6):471-475. [doi: 10.1016/j.nmd.2018.03.005] [Medline: 29655528]
76. Claramunt M, Idelssonhn S, James M, Corti M, Anton V, Byrne B, et al. P.84 Gait analysis of patients with Pompe disease using a portable system. Neuromuscular Disorders 2022;32:S76. [doi: 10.1016/j.nmd.2022.07.143]
77. Hamed A, Curran C, Gwaltney C, DasMahapatra P. Mobility assessment using wearable technology in patients with late-onset Pompe disease. NPJ Digit Med 2019;2:70 [FREE Full text] [doi: 10.1038/s41746-019-0143-8] [Medline: 31341956]
78. Harfouche M, Kishnani PS, Krusinska E, Gault J, Sitaraman S, Sowinski A, et al. Use of the patient-reported outcomes measurement information system (PROMIS®) to assess late-onset Pompe disease severity. J Patient Rep Outcomes 2020;4(1):83 [FREE Full text] [doi: 10.1186/s41687-020-00245-2] [Medline: 33034771]
79. Sniderman King L, Pan Y, Nallamilli BRR, Hegde M, Jagannathan L, Ramachander V, et al. Pompe disease ascertained through the lantern project, 2018-2021: next-generation sequencing and enzymatic testing to overcome obstacles to diagnosis. Mol Genet Metab 2023;139(1):107565 [FREE Full text] [doi: 10.1016/j.ymgme.2023.107565] [Medline: 37087815]

XSL•FO
RenderX

80. Pruniski B, Lisi E, Ali N. Newborn screening for Pompe disease: impact on families. J Inherit Metab Dis 2018;41(6):1189-1203. [doi: 10.1007/s10545-018-0159-2] [Medline: 29594646]
81. Carmona CJ, German-Morales M, Elizondo D, Ruiz-Rodado V, Grootveld M. Urinary metabolic distinction of Niemann-Pick class 1 disease through the use of subgroup discovery. Metabolites 2023;13(10):1079 [FREE Full text] [doi: 10.3390/metabo13101079] [Medline: 37887404]
82. Moreno-Barea FJ, Franco L, Elizondo D, Grootveld M. Data augmentation techniques to improve metabolomic analysis in Niemann-Pick type C disease. : Springer International Publishing; 2022 Presented at: Proceedings of the Computational Science – ICCS 2022 22nd International Conference; June 21–23, 2022; London, UK p. 78-91. [doi: 10.1007/978-3-031-08757-8_8]
83. Minter Baerg MM, Stoway SD, Hart J, Mott L, Peck DS, Nett SL, et al. Precision newborn screening for lysosomal disorders. Genet Med 2018;20(8):847-854 [FREE Full text] [doi: 10.1038/gim.2017.194] [Medline: 29120458]
84. Tillema J, Derks MG, Pouwels PJW, de Graaf P, van Rappard DF, Barkhof F, et al. Volumetric MRI data correlate to disease severity in metachromatic leukodystrophy. Ann Clin Transl Neurol 2015;2(9):932-940 [FREE Full text] [doi: 10.1002/acn3.232] [Medline: 26401514]
85. Clas P, Groeschel S, Wilke M. A semi-automatic algorithm for determining the demyelination load in metachromatic leukodystrophy. Acad Radiol 2012;19(1):26-34. [doi: 10.1016/j.acra.2011.09.008] [Medline: 22142678]
86. Esmail S, Danter WR. Artificially induced pluripotent stem cell-derived whole-brain organoid for modelling the pathophysiology of metachromatic leukodystrophy and drug repurposing. Biomedicines 2021;9(4):440 [FREE Full text] [doi: 10.3390/biomedicines9040440] [Medline: 33923989]
87. Regier DS, Kwon HJ, Johnston J, Golas G, Yang S, Wiggs E, et al. MRI/MRS as a surrogate marker for clinical progression in GM1 gangliosidosis. Am J Med Genet A 2016;170(3):634-644. [doi: 10.1002/ajmg.a.37468] [Medline: 26646981]
88. Davies EH, Johnston J, Toro C, Tifft CJ. A feasibility study of mHealth and wearable technology in late onset GM2 gangliosidosis (Tay-Sachs and Sandhoff Disease). Orphanet J Rare Dis 2020;15(1):199 [FREE Full text] [doi: 10.1186/s13023-020-01473-x] [Medline: 32746863]
89. Ruohola A, Salli E, Roine T, Tokola A, Laine M, Tikkanen R, et al. Detection of aspartylglucosaminuria patients from magnetic resonance images by a machine-learning-based approach. Brain Sci 2022;12(11):1522 [FREE Full text] [doi: 10.3390/brainsci12111522] [Medline: 36358448]
90. Kobrinskii BA, Blagosklonov NA, Gribova VC, Shalfeeva EA. Expert system for the diagnosis of orphan diseases. : Springer International Publishing; 2023 Presented at: Proceedings of the Sixth International Scientific Conference “Intelligent Information Technologies for Industry” (IITI’22); October 31 – November 6, 2022; Istanbul, Turkey p. 251-260. [doi: 10.1007/978-3-031-19620-1_24]
91. Blagosklonov N, Kobrinskii B. Model of integral evaluation of expert knowledge for the diagnosis of lysosomal storage diseases. CEUR Workshop Proceedings 2020;2648(22):1-20 [FREE Full text]
92. Kamel MA, Abbas MT, Kanaan CN, Awad KA, Baba Ali N, Scalia IG, et al. How artificial intelligence can enhance the diagnosis of cardiac amyloidosis: a review of recent advances and challenges. J Cardiovasc Dev Dis 2024;11(4):118 [FREE Full text] [doi: 10.3390/jcdd11040118] [Medline: 38667736]
93. Martini N, Aimo A, Barison A, Della Latta D, Vergaro G, Aquaro GD, et al. Deep learning to diagnose cardiac amyloidosis from cardiovascular magnetic resonance. J Cardiovasc Magn Reson 2020;22(1):84 [FREE Full text] [doi: 10.1186/s12968-020-00690-4] [Medline: 33287829]
94. Aimo A, Martini N, Barison A. 70 Deep learning to diagnose cardiac amyloidosis from cardiac magnetic resonance findings. Eur Heart J Suppl 2020;22(Supplement_N):N116-N130. [doi: 10.1093/eurheartj/suaa205]
95. Hsieh TC, Lesmann H, Krawitz PM. Facilitating the molecular diagnosis of rare genetic disorders through facial phenotypic scores. Curr Protoc 2023;3(10):e906. [doi: 10.1002/cpz1.906] [Medline: 37812136]
96. Reiter AMV, Pantel JT, Danyel M, Horn D, Ott C, Mensah MA. Validation of 3 computer-aided facial phenotyping tools (DeepGestalt, GestaltMatcher, and D-Score): comparative diagnostic accuracy study. J Med Internet Res 2024;26:e42904 [FREE Full text] [doi: 10.2196/42904] [Medline: 38477981]
97. Liévin V, Hansen JM, Lund A, Elstein D, Matthiesen ME, Elomaa K, et al. FindZebra online search delving into rare disease case reports using natural language processing. PLOS Digit Health 2023;2(6):e0000269 [FREE Full text] [doi: 10.1371/journal.pdig.0000269] [Medline: 37384616]
98. Pankow A, Meißner N, Fouquette L, Türk E, Gilbert S, Tübler T, et al. POS1551-HPR medical expert knowledge meets AI: how expert knowledge can improve symptom assessment apps - a new approach in rare diseases. Ann Rheum Dis 2022;81:1120. [doi: 10.1136/annrheumdis-2022-eular.3734]
99. Gómez J, Artigas L, Valls R, Gervas-Arruga J. An in silico approach to identify early damage biomarker candidates in metachromatic leukodystrophy. Mol Genet Metab Rep 2023;35:100974 [FREE Full text] [doi: 10.1016/j.ymgmr.2023.100974] [Medline: 37275681]
100. Pade LR, Stepler KE, Portero EP, DeLaney K, Nemes P. Biological mass spectrometry enables spatiotemporal 'omics: from tissues to cells to organelles. Mass Spectrom Rev 2024;43(1):106-138 [FREE Full text] [doi: 10.1002/mas.21824] [Medline: 36647247]

XSL•FO
RenderX

101. Pantel JT, Hajjir N, Danyel M, Elsner J, Abad-Perez AT, Hansen P, et al. Efficiency of computer-aided facial phenotyping (DeepGestalt) in individuals with and without a genetic syndrome: Diagnostic accuracy study. J Med Internet Res 2020;22(10):e19263 [FREE Full text] [doi: 10.2196/19263] [Medline: 33090109]
102. Akgun A, Gokcay G, Mungan NO, Sivri HS, Tezer H, Zeybek CA, et al. Expert-opinion-based guidance for the care of children with lysosomal storage diseases during the COVID-19 pandemic: an experience-based Turkey perspective. Front Public Health 2023;11:1092895 [FREE Full text] [doi: 10.3389/fpubh.2023.1092895] [Medline: 36794069]
103. Koracin V, Loeber JG, Mlinaric M, Battelino T, Bonham JR, Groselj U, et al. Global impact of COVID-19 on newborn screening programmes. BMJ Glob Health 2022;7(3):e007780 [FREE Full text] [doi: 10.1136/bmjgh-2021-007780] [Medline: 35236661]
104. Lampe C, Dionisi-Vici C, Bellettato CM, Paneghetti L, van Lingen C, Bond S, et al. The impact of COVID-19 on rare metabolic patients and healthcare providers: results from two MetabERN surveys. Orphanet J Rare Dis 2020;15(1):341 [FREE Full text] [doi: 10.1186/s13023-020-01619-x] [Medline: 33272301]
105. Mlinaric M, Bonham JR, Kožich V, Kölker S, Majek O, Battelino T, et al. Newborn screening in a pandemic-lessons learned. Int J Neonatal Screen 2023;9(2):21 [FREE Full text] [doi: 10.3390/ijns9020021] [Medline: 37092515]
106. Özalp Akın E, Eminoğlu FT, Doğulu N, Koç Yekeduz M, Öncül U, Akpınar F, et al. Unmet needs of children with inherited metabolic disorders in the COVID-19 pandemic. Turk Arch Pediatr 2022;57(3):335-341 [FREE Full text] [doi: 10.5152/TurkArchPediatr.2022.21367] [Medline: 35781238]
107. Ramaswami U, D'Amore S, Finnegan N, Hughes D, Kazemi M, Lysosomal Disorders Team, Royal Free London NHS Foundation Trust. Impact of SARS-CoV-2 (COVID-19) pandemic on patients with lysosomal storage disorders and restoration of services: experience from a specialist centre. Intern Med J 2021;51(10):1580-1593 [FREE Full text] [doi: 10.1111/imj.15473] [Medline: 34487419]
108. Schwartz IVD, Randon DN, Monsores N, Moura de Souza CF, Horovitz DDG, Wilke MVMB, et al. SARS-CoV-2 pandemic in the Brazilian community of rare diseases: a patient reported survey. Am J Med Genet C Semin Med Genet 2021;187(3):301-311 [FREE Full text] [doi: 10.1002/ajmg.c.31883] [Medline: 33474836]
109. Tobór-Świętek E, Sykut-Cegielska J, Bik-Multanowski M, Walczak M, Rokicki D, Kałużny Ł, et al. COVID-19 pandemic and patients with rare inherited metabolic disorders and rare autoinflammatory diseases-organizational challenges from the point of view of healthcare providers. J Clin Med 2021;10(21):4862 [FREE Full text] [doi: 10.3390/jcm10214862] [Medline: 34768381]
110. Zubarioglu T, Hopurcuoglu D, Uygur E, Ahmadzada S, Oge-Enver E, Isat E, et al. The impact of telemedicine for monitoring and treatment of phenylketonuria patients on metabolic outcome during coronavirus disease-19 outbreak. Telemed J E Health 2022;28(2):258-265. [doi: 10.1089/tmj.2020.0569] [Medline: 33719565]
111. Paneghetti L, Bellettato CM, Sechi A, Stepien KM, Scarpa M. One year of COVID-19: infection rates and symptoms in patients with inherited metabolic diseases followed by MetabERN. Orphanet J Rare Dis 2022;17(1):109 [FREE Full text] [doi: 10.1186/s13023-022-02247-3] [Medline: 35246188]
112. Sun A. Lysosomal storage disease overview. Ann Transl Med 2018;6(24):476 [FREE Full text] [doi: 10.21037/atm.2018.11.39] [Medline: 30740407]
113. Tumiene B, Graessner H, Mathijssen IM, Pereira AM, Schaefer F, Scarpa M, et al. European reference networks: challenges and opportunities. J Community Genet 2021;12(2):217-229 [FREE Full text] [doi: 10.1007/s12687-021-00521-8] [Medline: 33733400]
114. Heard JM, Bellettato C, van Lingen C, Scarpa M, MetabERN collaboration group. Research activity and capability in the European reference network metabERN. Orphanet J Rare Dis 2019;14(1):119 [FREE Full text] [doi: 10.1186/s13023-019-1091-8] [Medline: 31142374]
115. Talarico R, Aguilera S, Alexander T, Amoura Z, Andersen J, Arnaud L, et al. The added value of a European reference network on rare and complex connective tissue and musculoskeletal diseases: Insights after the first 5 years of the ERN ReCONNET. Clin Exp Rheumatol 2022;40 Suppl 134(5):3-11 [FREE Full text] [doi: 10.55563/clinexprheumatol/d2qz38] [Medline: 35349419]
116. Rossi A, Hoogeveen IJ, Lubout CMA, de Boer F, Fokkert-Wilts MJ, Rodenburg IL, et al. A generic emergency protocol for patients with inborn errors of metabolism causing fasting intolerance: a retrospective, single-center study and the generation of www.emergencyprotocol.net. J Inherit Metab Dis 2021;44(5):1124-1135 [FREE Full text] [doi: 10.1002/jimd.12386] [Medline: 33844307]
117. Collin-Histed T, Stoodley M, Beusterien K, Elstein D, Jaffe DH, Revel-Vilk S, et al. A global neuronopathic gaucher disease registry (GARDIAN): a patient-led initiative. Orphanet J Rare Dis 2023;18(1):195 [FREE Full text] [doi: 10.1186/s13023-023-02828-w] [Medline: 37480076]
118. Boychuk NA, Mulrooney NS, Kelly NR, Goldenberg AJ, Silver EJ, Wasserstein MP. Parental depression and anxiety associated with newborn bloodspot screening for rare and variable-onset disorders. Int J Neonatal Screen 2022;8(4):59 [FREE Full text] [doi: 10.3390/ijns8040059] [Medline: 36412585]
119. Morand O, Johnson J, Walter J, Atkinson L, Kline G, Frey A, et al. Symptoms and quality of life in patients with Fabry disease: results from an international patient survey. Adv Ther 2019;36(10):2866-2880 [FREE Full text] [doi: 10.1007/s12325-019-01061-x] [Medline: 31435831]

XSL•FO
RenderX

120. Hughes DA, Aguiar P, Deegan PB, Ezgu F, Frustaci A, Lidove O, et al. Early indicators of disease progression in Fabry disease that may indicate the need for disease-specific treatment initiation: findings from the opinion-based PREDICT-FD modified Delphi consensus initiative. BMJ Open 2020;10(10):e035182 [FREE Full text] [doi: 10.1136/bmjopen-2019-035182] [Medline: 33039984]
121. ISO/TR 11147:2023(en), health informatics — personalized digital health — digital therapeutics health software systems. International Organization for Standardization (ISO). 2023. URL: https://www.iso.org/obp/ui/#iso:std:iso:tr:11147:ed-1:v1:en [accessed 2026-01-01]
122. Staretz-Chacham O, Choi JH, Wakabayashi K, Lopez G, Sidransky E. Psychiatric and behavioral manifestations of lysosomal storage disorders. Am J Med Genet B Neuropsychiatr Genet 2010;153B(7):1253-1265. [doi: 10.1002/ajmg.b.31097] [Medline: 20872765]
123. Riemann D, Baglioni C, Bassetti C, Bjorvatn B, Dolenc Groselj L, Ellis JG, et al. European guideline for the diagnosis and treatment of insomnia. J Sleep Res 2017;26(6):675-700. [doi: 10.1111/jsr.12594] [Medline: 28875581]
124. Depression in adults: treatment and management. National Institute for Health and Care Excellence (NICE). 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK583074/ [accessed 2026-01-01]
125. DiGA-Verzeichnis – Digitale Gesundheitsanwendungen (DiGA) [Web page in German]. Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM). 2024. URL: https://diga.bfarm.de/de [accessed 2026-01-01]
126. Muenzer J, Jones SA, Tylki-Szymańska A, Harmatz P, Mendelsohn NJ, Guffon N, et al. Ten years of the hunter outcome survey (HOS): insights, achievements, and lessons learned from a global patient registry. Orphanet J Rare Dis 2017;12(1):82 [FREE Full text] [doi: 10.1186/s13023-017-0635-z] [Medline: 28464912]
127. Klein TL, Bender J, Bolton S, Collin-Histed T, Daher A, De Baere L, et al. A rare partnership: patient community and industry collaboration to shape the impact of real-world evidence on the rare disease ecosystem. Orphanet J Rare Dis 2024;19(1):262 [FREE Full text] [doi: 10.1186/s13023-024-03262-2] [Medline: 38987844]
128. Clarke LA, Giugliani R, Guffon N, Jones SA, Keenan HA, Munoz-Rojas MV, et al. Genotype-phenotype relationships in mucopolysaccharidosis type I (MPS I): insights from the international MPS I registry. Clin Genet 2019;96(4):281-289 [FREE Full text] [doi: 10.1111/cge.13583] [Medline: 31194252]
129. Choudhury MC, Chaube P. Integrating rare disease management in public health programs in India: exploring the potential of national health mission. Orphanet J Rare Dis 2022;17(1):43 [FREE Full text] [doi: 10.1186/s13023-022-02194-z] [Medline: 35144652]
130. Malavera A, Cadilhac DA, Thijs V, Lim JY, Grabsch B, Breen S, et al. Screening for fabry disease in young strokes in the Australian stroke clinical registry (AuSCR). Front Neurol 2020;11:596420 [FREE Full text] [doi: 10.3389/fneur.2020.596420] [Medline: 33324335]
131. Opladen T, Gleich F, Kozich V, Scarpa M, Martinelli D, Schaefer F, et al. U-IMD: the first unified European registry for inherited metabolic diseases. Orphanet J Rare Dis 2021;16(1):95 [FREE Full text] [doi: 10.1186/s13023-021-01726-3] [Medline: 33602304]
132. Wanner C, Ortiz A, Wilcox WR, Hopkin RJ, Johnson J, Ponce E, et al. Global reach of over 20 years of experience in the patient-centered Fabry Registry: advancement of Fabry disease expertise and dissemination of real-world evidence to the Fabry community. Mol Genet Metab 2023;139(3):107603 [FREE Full text] [doi: 10.1016/j.ymgme.2023.107603] [Medline: 37236007]
133. Schoenmakers DH, van den Berg S, Timmers L, Adang LA, Bäumer T, Bosch A, et al. Framework for multistakeholder patient registries in the field of rare diseases: focus on neurogenetic diseases. Neurology 2024;103(6):e209743. [doi: 10.1212/WNL.0000000000209743] [Medline: 39173102]
134. Gucciardi A, Legnini E, Di Gangi IM, Corbetta C, Tomanin R, Scarpa M, et al. A column-switching HPLC-MS/MS method for mucopolysaccharidosis type I analysis in a multiplex assay for the simultaneous newborn screening of six lysosomal storage disorders. Biomed Chromatogr 2014;28(8):1131-1139. [doi: 10.1002/bmc.3133] [Medline: 24449175]
135. La Cognata V, Cavallaro S. Detection of structural variants by NGS: revealing missing alleles in lysosomal storage diseases. Biomedicines 2022;10(8):1836 [FREE Full text] [doi: 10.3390/biomedicines10081836] [Medline: 36009380]
136. Ducatez F, Mauhin W, Boullier A, Pilon C, Pereira T, Aubert R, et al. Parsing fabry disease metabolic plasticity using metabolomics. J Pers Med 2021;11(9):898 [FREE Full text] [doi: 10.3390/jpm11090898] [Medline: 34575675]
137. Gojobori T, Ikeo K, Katayama Y, Kawabata T, Kinjo AR, Kinoshita K, et al. VaProS: a database-integration approach for protein/genome information retrieval. J Struct Funct Genomics 2016;17(4):69-81 [FREE Full text] [doi: 10.1007/s10969-016-9211-3] [Medline: 28012137]
138. Heinken A, El Kouche S, Guéant-Rodriguez RM, Guéant JL. Towards personalized genome-scale modeling of inborn errors of metabolism for systems medicine applications. Metabolism 2024;150:155738 [FREE Full text] [doi: 10.1016/j.metabol.2023.155738] [Medline: 37981189]
139. Pinto E Vairo F, Rojas Málaga D, Kubaski F, Fischinger Moura de Souza C, de Oliveira Poswar F, Baldo G, et al. Precision medicine for lysosomal disorders. Biomolecules 2020;10(8):1110 [FREE Full text] [doi: 10.3390/biom10081110] [Medline: 32722587]
140. Parenti G, Medina DL, Ballabio A. The rapidly evolving view of lysosomal storage diseases. EMBO Mol Med 2021;13(2):e12836 [FREE Full text] [doi: 10.15252/emmm.202012836] [Medline: 33459519]

XSL•FO
RenderX

141. Underhill HR, Golden-Grant K, Garrett LT, Uhrich S, Zielinski BA, Scott CR. Detecting the effects of Fabry disease in the adult human brain with diffusion tensor imaging and fast bound-pool fraction imaging. J Magn Reson Imaging 2015;42(6):1611-1622 [FREE Full text] [doi: 10.1002/jmri.24952] [Medline: 26018987]
142. Sodi A, Lenzetti C, Bacherini D, Finocchio L, Verdina T, Borg I, et al. Quantitative analysis of conjunctival and retinal vessels in Fabry disease. J Ophthalmol 2019;2019:4696429 [FREE Full text] [doi: 10.1155/2019/4696429] [Medline: 31093369]
143. Korlimarla A, Spiridigliozzi GA, Crisp K, Herbert M, Chen S, Malinzak M, et al. Novel approaches to quantify CNS involvement in children with Pompe disease. Neurology 2020;95(6):e718-e732 [FREE Full text] [doi: 10.1212/WNL.0000000000009979] [Medline: 32518148]
144. Kung S, Walters M, Claes P, Goldblatt J, Le Souef P, Baynam G. A dysmorphometric analysis to investigate facial phenotypic signatures as a foundation for non-invasive monitoring of lysosomal storage disorders. JIMD Rep 2013;8:31-39 [FREE Full text] [doi: 10.1007/8904_2012_152] [Medline: 23430517]
145. Mayhew D, Palmer K, Wilson I, Watson S, Stepien KM, Jenkins P, et al. Airway and anaesthetic management of adult patients with mucopolysaccharidoses undergoing cardiac surgery. J Clin Med 2024;13(5):1366 [FREE Full text] [doi: 10.3390/jcm13051366] [Medline: 38592237]
146. Cox-Brinkman J, Vedder A, Hollak C, Richfield L, Mehta A, Orteu K, et al. Three-dimensional face shape in Fabry disease. Eur J Hum Genet 2007;15(5):535-542. [doi: 10.1038/sj.ejhg.5201798] [Medline: 17327876]
147. Ulaşlı AM, Türkmen U, Toktaş H, Solak O. The complementary role of the kinect virtual reality game training in a patient with metachromatic leukodystrophy. PM R 2014;6(6):564-567. [doi: 10.1016/j.pmrj.2013.11.010] [Medline: 24412636]
148. Ronicke S, Hirsch MC, Türk E, Larionov K, Tientcheu D, Wagner AD. Can a decision support system accelerate rare disease diagnosis? Evaluating the potential impact of Ada DX in a retrospective study. Orphanet J Rare Dis 2019;14(1):69 [FREE Full text] [doi: 10.1186/s13023-019-1040-6] [Medline: 30898118]
149. Alves R, Piñol M, Vilaplana J, Teixidó I, Cruz J, Comas J, et al. Computer-assisted initial diagnosis of rare diseases. PeerJ 2016;4:e2211 [FREE Full text] [doi: 10.7717/peerj.2211] [Medline: 27547534]
150. Kobrinskii BA, Demikova NS, Blagosklonov NA. Knowledge engineering in construction of expert systems on hereditary diseases. In: Communications in Computer and Information Science. Cham: Springer International Publishing; 2018:35-45.
151. Piccininni M, Rohmann JL, Logroscino G, Kurth T. Blockchain-based innovations for population-based registries for rare neurodegenerative diseases. Front. Blockchain 2020;3. [doi: 10.3389/fbloc.2020.00020]
152. Vervoort D, Guetter CR, Peters AW. Blockchain, health disparities and global health. BMJ Innov 2021;7(2):506-514. [doi: 10.1136/bmjinnov-2021-000667]
153. Ghasemi M, Rahgozar M, Kavousi K. Complex disease genes identification using a heterogeneous network embedding approach. IEEE/ACM Trans Comput Biol Bioinform 2023;20(2):875-882. [doi: 10.1109/TCBB.2022.3175598] [Medline: 35594221]
154. Thirumal Kumar D, Eldous HG, Mahgoub ZA, George Priya Doss C, Zayed H. Computational modelling approaches as a potential platform to understand the molecular genetics association between Parkinson's and Gaucher diseases. Metab Brain Dis 2018;33(6):1835-1847. [doi: 10.1007/s11011-018-0286-3] [Medline: 29978341]
155. Warnock DG. 7th update on Fabry disease: biomarkers, progression and treatment opportunities in 2022. Nephron 2022;146 Suppl 1:1-51 [FREE Full text] [doi: 10.1159/000524933] [Medline: 35613555]
156. Abdallah S, Sharifa M, I Kh Almadhoun MK, Khawar MM, Shaikh U, Balabel KM, et al. The impact of artificial intelligence on optimizing diagnosis and treatment plans for rare genetic disorders. Cureus 2023;15(10):e46860 [FREE Full text] [doi: 10.7759/cureus.46860] [Medline: 37954711]
157. Al-Jasmi F, Moldovan L, Clarke JT. Hunter disease eClinic: interactive, computer-assisted, problem-based approach to independent learning about a rare genetic disease. BMC Med Educ 2010;10:72 [FREE Full text] [doi: 10.1186/1472-6920-10-72] [Medline: 20973983]
158. Di Brino E, Basile M, Rumi F, Fortunato A, Antonini D, Annicchiarico G, et al. Value for Patient - results of a value based healthcare workshop for the management of lysosomal storage disease in 3 Italian regions [Article in Italian]. GIHTAD 2022 Dec;15(5):1-7 [FREE Full text]
159. Big data and artificial intelligence for achieving universal health coverage: an international consultation on ethics. World Health Organization. URL: https://www.who.int/publications/i/item/WHO-HMM-IER-REK-2018-2 [accessed 2026-01-01]
160. Rao N, Kanago D, Morris M, Narayan V, Varshney K, Gn S, et al. Telegenetics: the experience of an Indian center (Centre for Human Genetics) during the COVID-19 pandemic. J Genet Couns 2021;30(5):1224-1232 [FREE Full text] [doi: 10.1002/jgc4.1517] [Medline: 34596296]
161. Almyranti M, Sutherland E, Ash N, Eiszele S. Artificial intelligence and the health workforce: perspectives from medical associations on AI in health. Organisation for Economic Co-Operation and Development (OECD). URL: https://www.oecd.org/en/publications/artificial-intelligence-and-the-health-workforce_9a31d8af-en.html [accessed 2024-11-20]
162. Tarricone R, Petracca F, Weller HM. "Towards harmonizing assessment and reimbursement of digital medical devices in the EU through mutual learning". NPJ Digit Med 2024;7(1):268 [FREE Full text] [doi: 10.1038/s41746-024-01263-w] [Medline: 39354125]

XSL•FO
RenderX

163. Spreafico A, Coder M, Boccardo V. DTx policy pathways: the evolving scenario in Europe. DTx Policy Pathways. 2023. URL: https://dtxpolicylandscape.org/ [accessed 2026-01-01]
164. Yoldaş Çelik M, Canda E, Yazıcı H, Erdem F, Kalkan Uçar S, Çöker M. Impact of the COVID-19 pandemic on inherited metabolic diseases: evaluation of enzyme replacement treatment adherence with telemedicine. J Pediatr Res Galenos Yayinevi 2022;9(4):391-396. [doi: 10.4274/jpr.galenos.2022.04206]
165. Zierhut HA, MacFarlane IM, Ahmed Z, Davies J. Genetic counselors' experiences and interest in telegenetics and remote counseling. J Genet Couns 2018;27(2):329-338. [doi: 10.1007/s10897-017-0200-x] [Medline: 29362948]
166. Hu H, Tjaden A, Knapp S, Antolin AA, Müller S. A machine learning and live-cell imaging tool kit uncovers small molecules induced phospholipidosis. Cell Chem Biol 2023;30(12):1634-1651.e6 [FREE Full text] [doi: 10.1016/j.chembiol.2023.09.003] [Medline: 37797617]
167. La Cognata V, Cavallaro S. A comprehensive, targeted NGS approach to assessing molecular diagnosis of lysosomal storage diseases. Genes (Basel) 2021;12(11):1750 [FREE Full text] [doi: 10.3390/genes12111750] [Medline: 34828358]
168. Fiumara A, Lanzafame G, Sapuppo A, Arena A, Cirnigliaro L, Barone R. Positive impact of home ERT for mucopolysaccharidoses and pompe disease: the lesson learnt from the COVID-19 pandemic. Healthcare (Basel) 2023;11(8):1176 [FREE Full text] [doi: 10.3390/healthcare11081176] [Medline: 37108010]
169. Finucane AM, O'Donnell H, Lugton J, Gibson-Watt T, Swenson C, Pagliari C. Digital health interventions in palliative care: a systematic meta-review. NPJ Digit Med 2021;4(1):64 [FREE Full text] [doi: 10.1038/s41746-021-00430-7] [Medline: 33824407]
170. Welter D, Juty N, Rocca-Serra P, Xu F, Henderson D, Gu W, et al. FAIR in action - a flexible framework to guide FAIRification. Sci Data 2023;10(1):291 [FREE Full text] [doi: 10.1038/s41597-023-02167-2] [Medline: 37208349]
171. Decherchi S, Pedrini E, Mordenti M, Cavalli A, Sangiorgi L. Opportunities and challenges for machine learning in rare diseases. Front Med (Lausanne) 2021;8:747612 [FREE Full text] [doi: 10.3389/fmed.2021.747612] [Medline: 34676229]
172. AI brings hope for patients with lysosomal storage disease. Today's Clinical Lab. 2023. URL: https://www.clinicallab.com/ai-brings-hope-for-patients-with-lysosomal-storage-disease-27351 [accessed 2026-01-01]
173. The act texts. EU Artificial Intelligence Act. URL: https://artificialintelligenceact.eu/the-act/ [accessed 2026-01-01]

Abbreviations

AI: artificial intelligence
CC: connected care
DHT: digital health technology
DL: deep learning
DMD: digital medical device
DMF: digital microfluidics
DSS: decision support system
DTx: digital therapeutics
EHR: electronic health record
ePRO: electronic patient-reported outcome
ERT: enzyme replacement therapy
EU: European Union
FD: Fabry disease
GD: Gaucher disease
GM1G: GM1 Gangliosidosis
GM2G: GM2 Gangliosidosis
HCP: health care provider
HIS: health information systems
HTA: health technology assessment
LLM: large language model
LSD: lysosomal storage disorder
mHealth: mobile health
ML: machine learning
MRI: magnetic resonance imaging
MS/MS: tandem mass spectrometry
NGS: next-generation sequencing
NLP: natural language processing
PCC: population, concept, context
PD: Pompe disease
PICO: Population, Intervention, Comparison, and Outcome

XSL•FO
RenderX

PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RCT: randomized controlled trial
RD: rare disease
VR: virtual reality

Edited by J Sarvestan; submitted 22.Mar.2025; peer-reviewed by B Wolfe-Piggott, E Ofoka; comments to author 18.Aug.2025; accepted 02.Feb.2026; published 02.Apr.2026.

Please cite as:
Spreafico AMC, Neri L, Angel K, Bellettato CM, Scarpa M
Artificial Intelligence, Connected Care, and Enabling Digital Health Technologies in Rare Diseases With a Focus on Lysosomal Storage Disorders: Scoping Review
J Med Internet Res 2026;28:e73612
URL: https://www.jmir.org/2026/1/e73612
doi:10.2196/73612
PMID:41661224

©Alberta MC Spreafico, Luca Neri, Kim Angel, Cinzia Maria Bellettato, Maurizio Scarpa. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Measuring Technology-Facilitated Sexual Violence and Abuse: Scoping Review of Existing Measures

Sharon Hoi Lam Pak[1], MPhil; Linruo Zhang[1], MSc; Chanchan Wu[2], PhD; Pui Hing Chau[1], PhD; Caroline Bradbury-Jones[3], PhD; Edmond Pui Hang Choi[1], PhD

[1]School of Nursing, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China (Hong Kong)

[2]School of Nursing, Faculty of Health and Social Science, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)

[3]Department of Nursing and Midwifery, University of Birmingham, Birmingham, United Kingdom

Corresponding Author:
Edmond Pui Hang Choi, PhD
School of Nursing
Li Ka Shing Faculty of Medicine
The University of Hong Kong
5/F, HKUMed Academic Building, 3 Sassoon Road, Pokfulam
Hong Kong, 0000
China (Hong Kong)
Phone: 852 3917 6972
Email: h0714919@connect.hku.hk

Abstract

Background: Technology-facilitated sexual violence and abuse (TFSVA) encompasses sexual harms perpetrated through digital means. Rapid developments in this field pose challenges in precisely defining and measuring TFSVA. Consequently, existing measures may fail to capture its multifaceted nature.

Objective: This scoping review aimed to map and appraise current knowledge on measures used to assess TFSVA, specifically to identify and summarize existing measures of TFSVA and its associated constructs, including victimization, perpetration, related attitudes, motivations, and impacts.

Methods: Following the Joanna Briggs Institute framework, we searched 10 English and Chinese databases (PubMed, Embase, CINAHL, Scopus, ProQuest, Web of Science, Chinese National Knowledge Infrastructure, SinoMed, Index to Taiwan Periodical Literature, and Ariti Library) from inception to January 13, 2026. We included peer-reviewed empirical studies focused on developing, adapting, or validating any TFSVA measures across populations and settings. Two independent reviewers screened records and extracted data. Data were analyzed using structural topic modeling (STM) and thematic analysis. Risk-of-bias assessment was not conducted, in accordance with scoping review methodology.

Results: Among the 319 included studies, 218 distinct measures were identified. The analysis revealed a fragmented and conceptually narrow measurement landscape. Most measures focused on specific subdomains, primarily nonconsensual sexting (130/218, 59.6%), with less attention to image-based sexual abuse (30/218, 13.8%), online sexual harassment (27/218, 12.4%), online grooming (13/218, 6.0%), and cyber sexual dating abuse (8/218, 3.7%). Nearly half (108/218, 49.5%) assessed both victimization and perpetration, while 96 out of 218 (44.0%) focused solely on victimization, 12 out of 218 (5.5%) on perpetration, 1 out of 218 (0.5%) on the bystander perspective, and the remaining 1 out of 218 (0.5%) addressed all aspects. The vast majority (199/218, 91.3%) measured behaviors, whereas attitudes, motivations, and impacts were rarely assessed. STM identified 9 key topics, synthesized into 3 core dimensions: consent violation and coercive control; status degradation and demeaning harassment; and contextual and relational framing. Thematic analysis highlighted 5 themes: TFSVA acts, attitudes and beliefs, motivational factors, tactical and contextual dynamics, and survivor impact. Integrating these findings reveals a fundamental conceptual misalignment between existing measures and the full scope of TFSVA.

Conclusions: This first comprehensive synthesis of existing TFSVA measures innovatively integrates STM and thematic analysis. While the absence of a risk-of-bias assessment precludes definitive conclusions regarding measurement robustness, this review moves beyond a descriptive catalog to identify substantive gaps within TFSVA-specific measures, most notably the omission of artificial intelligence–facilitated tactics, intersectional vulnerabilities, and institutional factors. These findings provide

XSL•FO
RenderX

an empirical basis for developing robust, culturally sensitive measures essential for targeted research, effective policy, and responsive social support services in a rapidly evolving digital landscape.

Trial Registration: Open Science Framework 10.17605/OSF.IO/Q5ETW; https://doi.org/10.17605/OSF.IO/Q5ETW

(J Med Internet Res 2026;28:e90068) doi:10.2196/90068

KEYWORDS

image-based sexual abuse; online sexual violence; measures; scoping review; structural topic modeling; technology-facilitated sexual violence

Introduction

Rationale

Technology-facilitated sexual violence and abuse (TFSVA) is a pervasive and urgent societal challenge, representing a convergence of digital innovation and sexual violence. Defined as any sexual violence and abuse perpetrated or amplified through digital technologies—from smartphones and social media platforms to artificial intelligence (AI)–driven applications and GPS tracking—TFSVA encompasses a spectrum of harms, including image-based sexual abuse (IBSA), cyberstalking, and targeted online sexual harassment [1,2]. Specific manifestations range from the nonconsensual sharing of intimate content to the doxing of sensitive personal information, such as sexual orientation or sexually transmitted infection status [3]. This review uses TFSVA as the overarching term for these technology-mediated sexual harms [3].

The rapid proliferation of digital technologies has precipitated a rise in TFSVA worldwide, with studies indicating that 16.6%-20% of adults in the United Kingdom and Australia have reported online sexual harassment in their lifetime, disproportionately affecting women and youth [4-6]. A meta-analysis and systematic review of 25 studies found that 8.8% of individuals have experienced nonconsensual sharing of sexual images or sexually explicit messages (sexting), with many facing threats or unauthorized image capture [7]. TFSVA produces severe consequences, including anxiety, depression, posttraumatic stress disorder, and increased suicidal risk, often stemming from profound privacy violations, stigma, and blaming the individual [1,8]. The pervasive and enduring nature of these digital harms underscores the urgent need for effective identification methods and comprehensive intervention strategies.

A robust and comprehensive measure of TFSVA is critical for understanding its prevalence, dynamics, and impact, as well as for informing prevention, intervention, and policy. However, the landscape of TFSVA is complex and rapidly evolving [6]. This poses a significant challenge in defining and categorizing TFSVA, and existing measures often struggle to capture its multifaceted and context-specific nature [6,7]. One of the primary difficulties in addressing this issue lies in the absence of a universally agreed-upon terminology or definition to describe the myriad ways in which technology is used to perpetrate acts of sexual abuse [6,7]. The absence of shared definitions can lead to underrecognition and underreporting of TFSVA [9], impeding accurate estimation of prevalence and effects [10]. Additionally, the lack of standardized definitions and measures can complicate cross-study comparisons and the aggregation of data [11]. Therefore, it is important to continue gathering data on whether this field of research is reaching a consensus on developing a uniform and comprehensive measure.

To the best of our knowledge, no prior review has systematically mapped existing TFSVA measures. One explanation for this absence is that conducting a systematic review on this topic is challenging due to the lack of precise definitions and standardized research variables in the current evidence literature. This has led to a wide array of measures being used to assess different forms and outcomes of TFSVA. Given these challenges, a scoping review is the most suitable approach for examining emerging evidence before conducting a systematic review [12]. A scoping review is particularly useful for mapping existing knowledge on a topic based on the available literature, thereby providing an overview of the current state of research and identifying gaps in the evidence [13]. It is a transparent and increasingly recognized methodology that can address a broad range of research questions and identify evidence gaps [13]. Although the concept of TFSVA is not yet well defined, to allow for a comprehensive exploration of the topic, TFSVA in this review refers to any form of sexual violence, exploitation, or harassment through the misuse of digital technologies [3].

Objectives

This scoping review aimed to map and appraise current knowledge on measures used to assess TFSVA. The primary question guiding the review was “What measures are currently available to assess TFSVA?” Subquestions addressed measurement domains, psychometric properties, target populations, and contextual applications to ensure comprehensive coverage of the field. The objectives were to (1) identify and summarize existing measures for TFSVA and its associated constructs, including experiencing harm, perpetration, related attitudes, motivations, and impacts; (2) clarify key concepts and definitions of TFSVA in the literature; (3) evaluate the applicability and validity of these measures across different populations and settings; and (4) identify content coverage and gaps in current measurement approaches.

Methods

Protocol and Registration

The scoping review was conducted according to the 5 standard stages of scoping review methodology outlined in the a priori protocol [14], which was registered with the Open Science Framework [15] on March 13, 2024. The review adhered to Joanna Briggs Institute guidance and the PRISMA-ScR

XSL•FO
RenderX

(Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) reporting standards [12,16]. Deviations from the published protocol were explicitly noted in the relevant sections below.

Eligibility Criteria

The screening of studies was guided by the Population-Concept-Context (PCC) framework [17]. The core inclusion criteria were as follows: (1) any population, including survivors, perpetrators, and bystanders, without restriction on sexual orientation, gender, age, or ethnicity; (2) empirical studies focused on the development, cultural adaptation, or validation of a measure for TFSVA; and (3) any geographical, cultural, or online setting. All peer-reviewed journal articles in English or Chinese that reported quantitative, qualitative, or mixed methods empirical data were included. Studies that described technology misuse without a sexual or intimate focus were excluded. For this review, a "measure" was defined as any instrument, tool, or technique used to assess, quantify, or characterize TFSVA or its associated constructs. This included multiitem scales, single-item questions, qualitative questions, and adapted or modified versions of existing measures. No minimum number of items was required for inclusion. Both validated and author-developed measures used in empirical studies were eligible.

Information Sources

The literature search and its reporting followed the PRISMA-S (extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews) guidelines [18]. A systematic search was conducted across 10 databases from database inception to January 13, 2026, including 6 English-language electronic databases (PubMed, Embase, CINAHL, Scopus, ProQuest, and Web of Science) and 4 Chinese databases (Chinese National Knowledge Infrastructure, SinoMed, Index to Taiwan Periodical Literature, and Airiti Library). Study registries were not searched, as the inclusion criteria focused exclusively on published peer-reviewed studies. Registries primarily contain protocols, preprints, and trial records, which were not eligible for inclusion. Online resources and browsing were not conducted, as database searching comprehensively covered the published literature and was supplemented by citation searching of included studies. No other search methods were used beyond the database and citation searches described.

Search

The search strategy combined keywords and controlled vocabulary terms, such as Medical Subject Headings, for the key concepts of the PCC framework. Search filters were not applied to maintain search sensitivity. The search strategy relied on controlled vocabulary and keywords only.

Searches were limited to English- and Chinese-language peer-reviewed empirical studies. To ensure the review reflects the most current literature, the search was executed in 2 phases. The original search, detailed in the published protocol [14], covered records from database inception to December 31, 2024. An updated search was conducted using an identical strategy to capture records published from January 1, 2025, to January 13, 2026. As this was not an update of a prior review, search updates were not applicable. The full electronic search strategy for all databases is provided in Multimedia Appendix 1. To identify additional relevant papers, the reference lists of included studies were hand-searched. We did not contact individual researchers, experts, or organizations for unpublished measures or gray literature, as this scoping review aimed to map the publicly available published literature rather than solicit unpublished measures.

Selection of Sources of Evidence

Following the search, all records were collated and managed using EndNote 20 (Clarivate Plc). No other systematic review software was used for screening or data extraction. After automatic duplicate removal using EndNote, a manual check was performed. Two reviewers independently screened titles, abstracts, and full texts against the inclusion criteria. Any disagreements were resolved through consensus or consultation with a third researcher. Studies excluded at the full-text review stage were documented with specific reasons, aligned with the PCC framework.

Data Items and Data Charting Process

Data were extracted using a comprehensive data charting form detailed in the published protocol [14]. The form captured variables across several domains aligned with the PCC framework: (1) study identification and design; (2) population characteristics (eg, age group, sexual orientation, role as survivor, perpetrator, or bystander); (3) geographical setting; (4) the specific type and definition of TFSVA measured; (5) details of the measure used, including its name, language, and reported psychometric properties; and (6) key outcomes related to TFSVA.

The extraction process began with a pilot phase, during which 2 reviewers independently extracted data from 5 studies to establish consensus and refine the form. The remaining data were extracted by 1 reviewer and confirmed by a second reviewer. Discrepancies were resolved through discussion, and a third reviewer was consulted when necessary.

Synthesis of Results

Overview

Our analysis followed a mixed methods design to integrate quantitative and qualitative insights. This was implemented through a 2-stage process: (1) a descriptive synthesis guided by the PAGER (Patterns, Advances, Gaps, Evidence for practice, and Research Recommendations) Framework to generate higher-level insights [19]; and (2) a content analysis using both structural topic modeling (STM) [20] and reflexive thematic analysis.

Descriptive Synthesis Using the PAGER Framework

The PAGER Framework provided a structured approach to synthesize and report findings [19]. We applied this framework to systematically categorize the descriptive characteristics of the included studies and their measures across 5 dimensions:

- Patterns: identifying dominant trends in the literature.
- Advances: highlighting methodological and conceptual progress.

XSL•FO
RenderX

- Gaps: recording underrepresented populations, contexts, or TFSVA domains.
- Evidence for practice: distilling findings applicable to current research and policy.
- Research recommendations: formulating priorities for future measure development.

Analysis of Measure Content

To analyze the substantive content of the measures, STM and reflexive thematic analysis were conducted on the same set of questionnaire items. This approach aligns with the methodological principle that qualitative and computational text analyses are complementary and analogous processes for identifying patterns in textual data [21]. STM provided a quantitative and reproducible overview of latent conceptual patterns across the full dataset, while reflexive thematic analysis offered qualitative depth, capturing detailed meaning, contextual nuances, and specific question types that purely frequency-based analysis might overlook.

STM is a text-mining technique that identifies latent topics within a text corpus. Each questionnaire item was treated as a separate document. Text preprocessing included lowercasing, removal of punctuation and digits, whitespace normalization, lemmatization, synonym mapping (eg, “photos or images or pics” → picture), and stop-word removal. We fitted STM models with the number of topics (K) ranging from 3 to 10, evaluating model fit using held-out likelihood, residuals, lower bound, semantic coherence, and exclusivity. After identifying K=7, 9, and 10 as offering the best balance of fit, we estimated unsupervised STMs (without covariates) and selected K=9 for final reporting based on balanced fit and conceptual interpretability. All analyses were conducted in R (R Foundation) using the stm (modeling and diagnostics), ggplot2 (visualization), and textstem (lemmatization) packages. This analysis yielded a quantitative map of the dominant conceptual patterns across all measurement items. Interpretive synthesis was performed to translate statistical topics into higher-order conceptual dimensions.

Reflexive thematic analysis was conducted following the 6-phase approach by Clarke and Braun [22]: (1) familiarization with the measurement items; (2) initial code generation; (3) theme search; (4) theme review; (5) theme definition and naming; and (6) report production. Analytical memos documented interpretive decisions and theoretical insights throughout, ensuring thematic robustness and grounding in the data. This independent analysis provided rich, detailed descriptions of question types, contextual nuances, and underlying constructs, complementing and contextualizing the broader statistical patterns identified by STM.

The findings from these 2 analyses were integrated to identify gaps and misalignment within the TFSVA measurement field. Consistent with scoping review methodology, risk-of-bias assessment was omitted [16], although only peer-reviewed articles were included to maintain quality standards.

Ethics Considerations

This scoping review involved a secondary analysis of previously published, publicly available literature. All included studies reported obtaining ethical approval from their respective institutional review boards or equivalent bodies, as documented in the original publications. No additional ethical approval was required for this review, as it did not involve primary data collection from human participants.

Results

Overview of Included Studies and TFSVA Measures

The database searches yielded a total of 17,640 studies. After removing 4046 duplicates, 13,594 references remained for further screening. Specifically, 12,811 records were excluded after title and abstract screening. Of the 783 remaining records for full-text screening, 14 were not retrieved and 450 were excluded for specific reasons. Finally, 319 studies were included in the analysis. From the perspective of publication language, all included studies were in English, and no relevant studies were published in Chinese. Study screening and selection were conducted following the PRISMA flow diagram [23]. The detailed study selection process and specific reasons for exclusion are presented in Figure 1. Also see Multimedia Appendices 2 and 3 for PRISMA-ScR and PRISMA-S checklists, respectively.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of study selection. This diagram shows the number of records identified, screened, assessed for eligibility, and included in the scoping review of technology-facilitated sexual violence and abuse measures, from database searching through final inclusion.

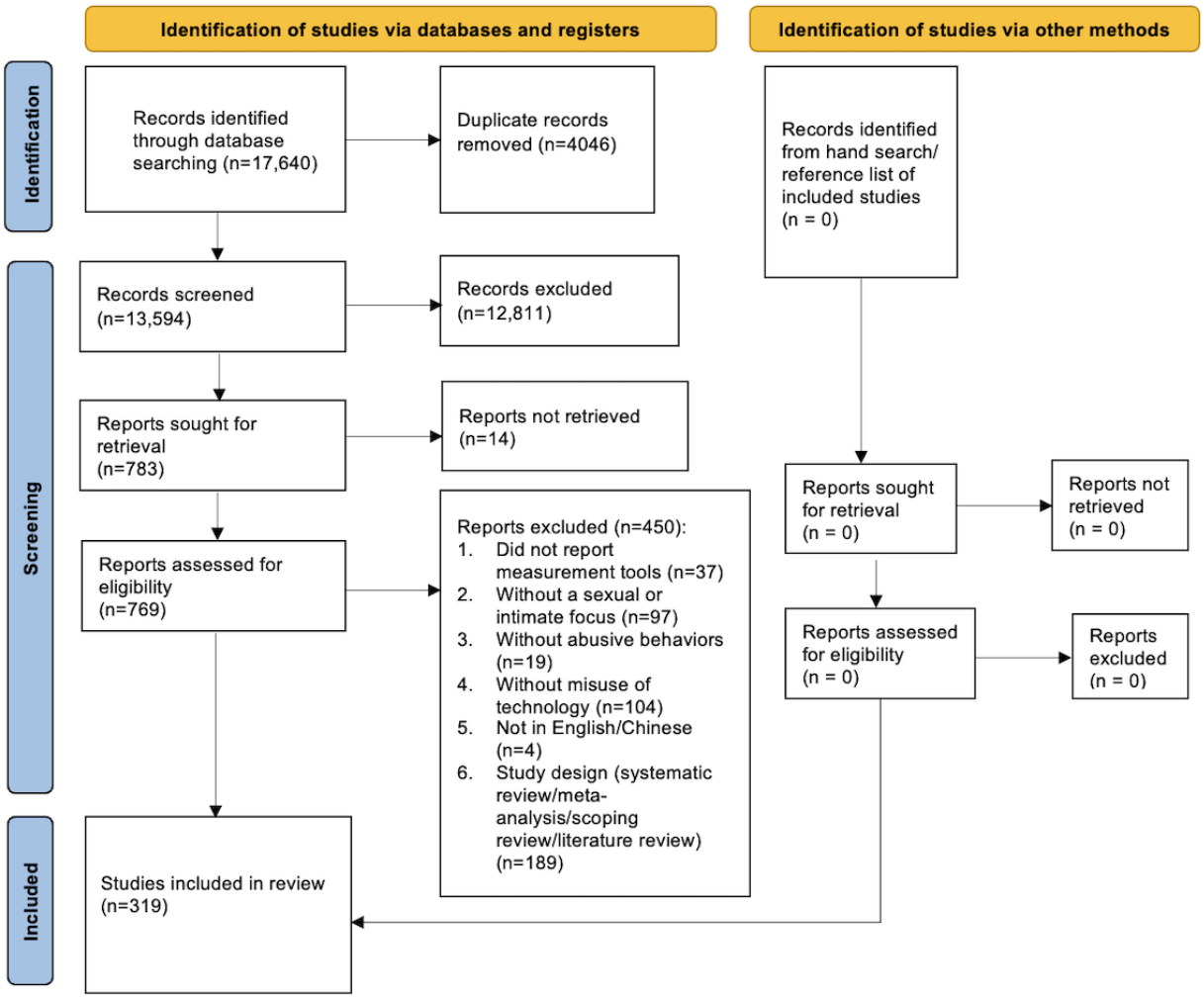

Included studies were published from September 2011 to January 2026 (Multimedia Appendix 4). Study populations originated from diverse geographical regions, including 42 countries. Most studies were conducted in a single country, with the highest proportions from Europe (129/319, 40.4%) and North America (123/319, 38.6%), followed by Asia (40/319, 12.5%), South America (8/319, 2.5%), Africa (5/319, 1.6%), and Oceania (2/319, 0.6%). The remaining 12 (3.8%) studies were multicountry. Most studies targeted adolescents (aged 10-18 years; 130/319, 40.7%) and young adults (aged 18-26 years; 93/319, 29.1%), with 9 (2.8%) studies targeting both adolescents and young adults. Fewer studies focused on adult populations (aged ≥18 years; 68/319, 21.3%). The remaining 19 (6.0%) studies did not specify the age group of the target population. In terms of gender, 294 (92.2%) studies included both female and male populations, while 14 (4.4%) focused only on female populations and 6 (1.9%) focused only on male populations. Five studies (1.6%) included gender minority individuals, such as transgender and gender-diverse populations. Regarding sexual orientation, 314 (98.4%) studies did not specify the sexual orientation of the target population, and only 5 (1.6%) focused on sexual orientation minority populations. The studies used a variety of methodologies, including quantitative surveys (303/319, 95.0%), qualitative interviews (12/319, 3.8%), and mixed methods approaches (4/319, 1.2%).

Among the 319 included studies, 218 distinct measures of TFSVA were identified. Most of these measures were related to nonconsensual sexting and sextortion (130/218, 59.6%). Other areas addressed included IBSA or revenge pornography (30/218, 13.8%), online sexual harassment (27/218, 12.4%), online grooming or solicitation (13/218, 6.0%), and cyber sexual dating abuse (8/218, 3.7%). The remaining 10 (4.6%) measures assessed general TFSVA behaviors. In terms of measurement focus, 96 (44.0%) measures focused on exposure to violence, 12 (5.5%) on perpetration, and 108 (49.5%) measured both aspects. One (0.5%) measure focused on the bystander role, and 1 (0.5%) measured all aspects. In terms of outcomes measured, most studies assessed TFSVA behaviors (199/218, 91.3%), including outcomes related to the prevalence and experiences of TFSVA. Five (2.3%) measures assessed attitudes toward TFSVA, and 4 (1.8%) evaluated motivations for committing TFSVA. The remaining 10 (4.6%) measures assessed more than 1 outcome. Characteristics of the included studies are presented in Table 1 and TFSVA measures are summarized in Table 2.

XSL•FO
RenderX

Table 1. Characteristics of the included studies. This table summarizes the key methodological and contextual details of the included studies (N=319).

Characteristics	Value, n (%)
Study population	
Europe	129 (40.4)
North America	123 (38.6)
Asia	40 (12.5)
South America	8 (2.5)
Africa	5 (1.6)
Oceania	2 (0.6)
Included more than 1 country	12 (3.8)
Age group (years)	
Adolescents (aged 10-18)	130 (40.7)
Young adults (aged 18-26)	93 (29.1)
Adolescents and young adults	9 (2.8)
Adults (aged 18 or above)	68 (21.3)
Any age group	19 (6.0)
Gender	
Both genders	294 (92.2)
Female only	14 (4.4)
Male only	6 (1.9)
Gender minority (ie, transgender and gender-diverse)	5 (1.6)
Sexual orientation	
Not specified	314 (98.4)
Sexual orientation minority (eg, LGBTQ+[a] or men who have sex with men)	5 (1.6)
Study methodology	
Quantitative survey	303 (95.0)
Qualitative interview	12 (3.8)
Mixed methods study	4 (1.2)

[a]LGBTQ+: lesbian, gay, bisexual, transgender, queer (or questioning), and others not explicitly listed.

XSL•FO
RenderX

Table 2. TFSVA[a] measures (n=218).

Characteristics	Value, n (%)
Types of TFSVA	
Nonconsensual sexting	130 (59.6)
Image-based sexual abuse	30 (13.8)
Online sexual harassment	27 (12.4)
Online grooming	13 (6.0)
Cyber sexual dating abuse	8 (3.7)
General TFSVA behaviors	10 (4.6)
Measurement focus	
Exposure to violence	96 (44.0)
Perpetration	12 (5.5)
Exposure to violence and perpetration	108 (49.5)
Bystander	1 (0.5)
All aspects	1 (0.5)
Outcome measures	
Behaviors	199 (91.3)
Attitudes	5 (2.3)
Motivations	4 (1.8)
More than 1 outcome	10 (4.6)

[a]TFSVA: technology-facilitated sexual violence and abuse.

Analysis of publication trends from 2011 to 2025 reveals the evolving focus of TFSVA measures in research. From 2011 to 2023, nonconsensual sexting emerged as the dominant type of TFSVA being measured, with numerous studies focusing on its prevalence, risks, and consequences, particularly among adolescents. However, since 2018, the field has expanded to include more diverse types of TFSVA, such as IBSA, online sexual harassment, and online grooming, reflecting the evolving nature of digital harms. The term TFSVA first appeared prominently in academic literature around 2018, marking a shift toward a more comprehensive understanding of technology-mediated sexual violence, and studies measuring TFSVA have increased significantly since then. By 2024, TFSVA had become the dominant focus of investigation, surpassing the narrower emphasis on nonconsensual sexting, with a broader spectrum of digital sexual violence and its multifaceted impacts being addressed. Data from 2025 indicate a continued increase in studies dedicated to these specific subtypes. The time trend of publications across different types of TFSVA is detailed in Figure 2.

XSL•FO
RenderX

Figure 2. Time trends in publications on different types of technology-facilitated sexual violence and abuse (TFSVA), 2011-2025. This stacked bar chart displays the annual number of publications included in this review, categorized by primary focus.

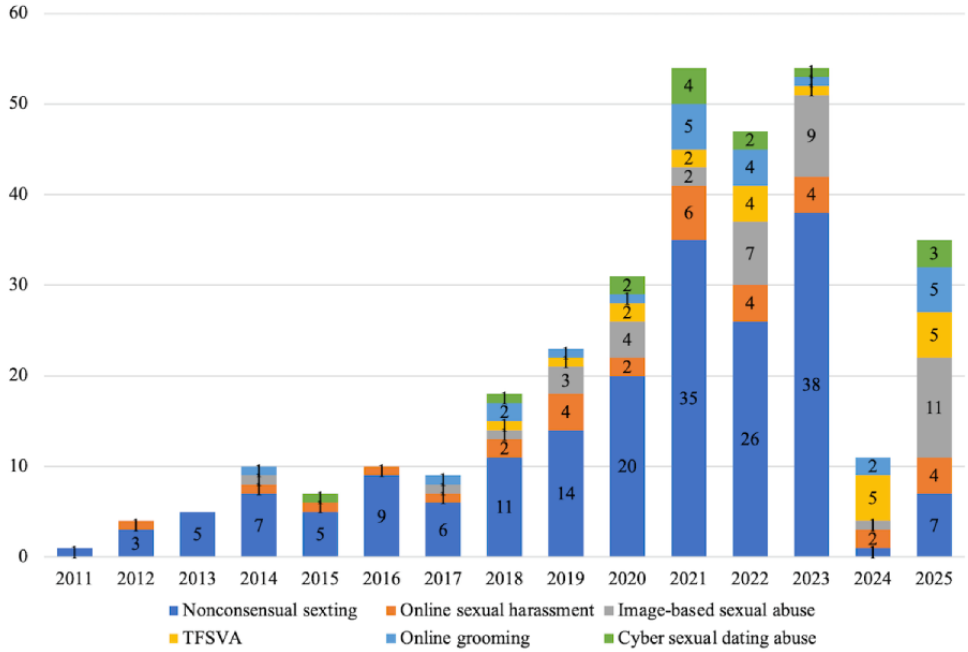

Among all included studies, only 55 reported details on psychometric properties. A total of 138 studies used nonvalidated measures, and 126 studies adapted questionnaires from existing measures. The distribution of validated, nonvalidated, and adapted measures by type of TFSVA across all included studies is presented in Figure 3.

Figure 3. Validation status of measures by type of technology-facilitated sexual violence and abuse (TFSVA). This stacked bar chart displays the distribution of validated, nonvalidated, and adapted measures for each TFSVA type identified in the included studies.

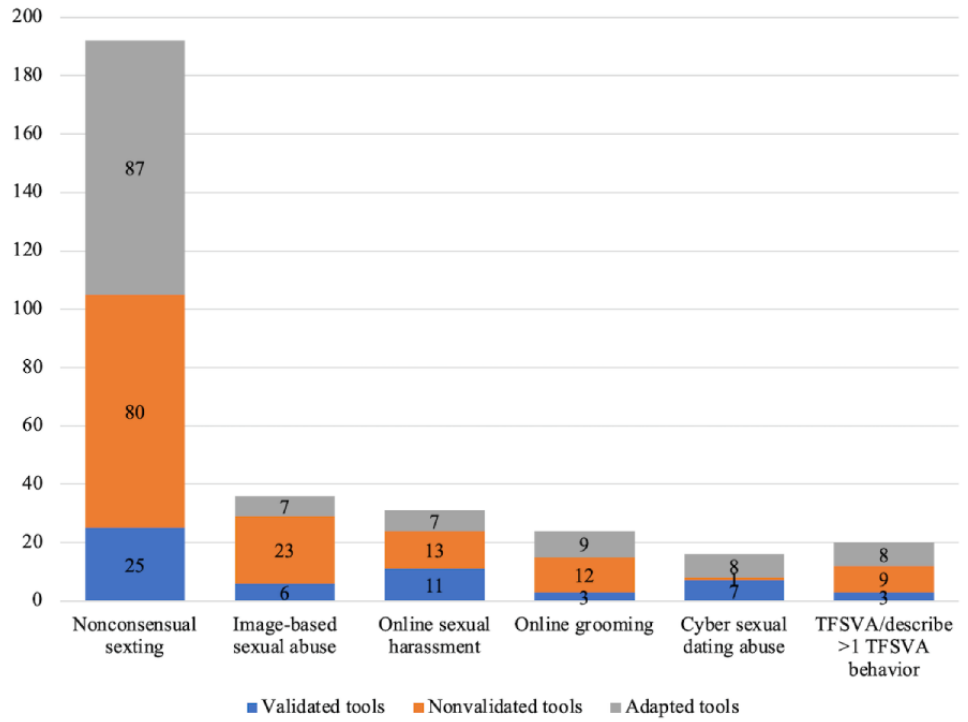

Measures by Types of TFSVA

Nonconsensual Sexting

Nonconsensual sexting, defined as sending or receiving sexually explicit messages or images without one's permission, was the most extensively studied form of TFSVA (130/218, 59.6%). In terms of age group, 56 out of 130 (43.1%) measures targeted adolescents, 39 out of 130 (30%) focused on young adults, and 32 out of 130 (24.6%) focused on adult populations; 2 out of 130 (1.5%) measures targeted both adolescents and young adults, and the remaining 1 out of 130 (0.8%) included all age groups. Three nonconsensual sexting measures specifically

XSL•FO
RenderX

targeted female participants, and none of the measures specified the sexual orientation of the target population.

Nearly all studies related to nonconsensual sexting (n=127) focused on assessing the behavioral domain, with questions primarily addressing the frequency and targets of sexting. Six studies measured attitudes toward nonconsensual sexting, exploring general perceptions and perceived risks associated with the practice. Additionally, 7 studies investigated the motivations behind such behavior, examining reasons for engagement, including sexual purposes, emotional influences, body-image reinforcement, and relationship enhancement.

In terms of psychometric properties, 25 measures were validated, of which 17 reported validity analyses, such as exploratory factor analysis or confirmatory factor analysis. All validated measures reported reliability, including internal consistency (eg, Cronbach α or McDonald ω). Among the validated measures, the Sexting Motivation Scale [24], the Sexting Behavior Scale [25], the Sexting Questionnaire [26], and the Sexting Attitude Scale [27] were the most commonly used. All of these demonstrated strong validity (ie, satisfactory exploratory factor analysis or confirmatory factor analysis) and reliability (ie, Cronbach α >0.8). However, these tools were predominantly validated among adolescents and young adults (18/25, 72%), with limited application to adults or diverse cultural groups.

Online Sexual Harassment

Online sexual harassment refers to any unwelcome or unwanted sexual behavior, advances, or communication that occurs through digital platforms, including social media, messaging apps, email, or online gaming environments [28]. This form of harassment encompasses a broad range of behaviors, such as sexualized cyberbullying, nonconsensual sharing of intimate content, sexually explicit comments, cyberflashing, and sexual coercion facilitated by digital means. Among the included studies (n=218), 27 (12.4%) measured online sexual harassment. In terms of age group, 11 out of 27 (41%) targeted adolescents, 8 out of 27 (30%) targeted young adults, 6 out of 27 (22%) targeted the adult population, and the remaining 2 (7%) included all age groups; 6 out of 27 (22%) measures were distributed to the female population only, and 1 targeted males only. One measure specifically focused on lesbian, gay, bisexual, transgender, queer (or questioning), and others not explicitly listed (LGBTQ+) individuals and examined sexuality-based cyberbullying [29].

Almost all measures included behavioral outcomes (26/27), with 1 study also measuring the motivation outcome. One study explored the attitude outcome from a bystander perspective; 19 out of 27 (70%) measures focused on exposure to violence, 3 (11%) on perpetration, and the remaining 5 (19%) on both aspects. Most questions in these studies inquired about personal experiences of online sexual harassment, such as receiving offensive sexist messages, encountering negative comments about one's gender or sexual orientation, or being subjected to inappropriate sexual remarks about one's body on social media. One measure with a motivation outcome asked about reasons for online sexual harassment.

As many as 11 (41%) measures demonstrated satisfactory validity and reliability. The Online Sexual Victimization Scale was the most commonly used measure [26].

Image-Based Sexual Abuse

IBSA, such as the nonconsensual sharing of intimate images, was assessed in 30 out of 218 (13.8%) studies; 5 out of these 30 (17%) studies targeted adolescents, 13 (43%) targeted young adults, and 6 (20%) targeted adults; 4 (13%) studies included participants across all age groups, and the remaining 2 (7%) did not specify the target population. One study focused on the male population only, while the others did not specify gender or sexual orientation.

Most measures focused on behavioral outcomes (n=25), while 4 studies assessed attitude outcomes and 1 included motivation outcomes. Among the 30 studies, 14 (47%) focused on IBSA exposure, 3 (10%) on perpetration, 12 (40%) included both exposure to violence and perpetration, 1 (3%) focused on a bystander perspective, and 1 (3%) included all perspectives. The measurement questions explored experiences related to sending, receiving, or requesting intimate, suggestive, or provocative images or videos (eg, "I have sent/received/requested intimate/suggestive/provocative images or videos"), reasons and motivations for sharing such content (eg, "I have sent dick pics hoping to receive sexy pictures in return"), and attitudes toward IBSA behaviors (eg, "If a man sends a nude or sexual image to a partner, he can't expect it will remain private").

A total of 6 (20%) measures were validated, demonstrating high validity and reliability. Among these, the Sexual Image-Based Abuse Myth Acceptance Scale was the most widely adapted and utilized [30].

Online Grooming

Online grooming refers to the process by which an individual, typically an adult, establishes an emotional connection with a child or adolescent through digital platforms, such as social media or messaging apps, with the intent of exploiting them for sexual purposes [31]. Of the 218 measures, 13 (6.0%) were designed to examine online grooming. Of these 13, 7 (54%) measures targeted adolescents, 2 (15%) targeted adults, 1 (8%) targeted young adults, and the remaining 3 (23%) focused on both adolescents and young adults. No studies specifically targeted sexual and gender minorities.

All studies focused on behavioral outcomes, with 1 also exploring motivations underlying such behavior. Of the 13 studies, 11 (85%) assessed online grooming exposure, while the remaining 2 (15%) examined both online grooming exposure and perpetration. Most measurement questions centered on the frequency and experiences of online grooming (eg, "In the past 12 months, how often has an adult asked me for pictures or videos of myself with sexual content?", "Someone has made sexual jokes or comments about me").

In these studies, 3 (23%) scales were validated with high reliability and validity, among which the Questionnaire for Online Sexual Solicitation and Interactions With Adults was the most commonly used [32].

XSL•FO
RenderX

Cyber Sexual Dating Abuse

Cyber sexual dating abuse, encompassing behaviors such as monitoring, control, and harassment within romantic relationships, was addressed in 8 out of 218 (3.7%) studies. Of these 8, 5 (63%) focused on adolescents and 3 (38%) on adults. One study specifically targeted sexual and gender minority girls and feminine teens.

All studies focused on behavioral outcomes, with 5 (63%) specifically addressing exposure and 3 (38%) exploring both exposure and perpetration. Some questionnaire content overlapped with items in IBSA and sexual aggression in dating or intimate relationships (eg, “Sending and/or uploading photos, images, and/or videos with intimate or sexual content without permission,” “Pressured my partner to send sexual or naked photos of him or her to me”).

Of these 8 measures, 7 (88%) were validated, demonstrating high reliability and validity. Among these, the Cyber Dating Abuse Questionnaire was the most frequently adapted scale [33].

TFSVA as a Whole

Since 2018, the term TFSVA has gained prominence as a comprehensive descriptor for a wide range of technology-mediated sexual harms [28]. Of the 218 studies, 10 (4.6%) explored TFSVA as a broader category: 2 (20%) targeted adolescents, 7 (70%) targeted adults, and the remaining 1 (10%) had no age restrictions. No studies specifically targeted sexual and gender minorities.

Of the 10 measures, 9 (90%) focused on behavioral outcomes, and 1 (10%) on attitudes toward TFSVA. Among these studies, 6 (60%) examined TFSVA exposure, while 4 (40%) assessed both exposure and perpetration. The measures encompassed various forms of abuse, including IBSA, online sexual harassment, and gender- or sexuality-based harassment, often grouping these behaviors under the umbrella of TFSVA. Questionnaire items focused on experiences of TFSVA (eg, “Nude or seminude image taken without permission,” “Unwanted sexual experience with someone met online,” “Receiving sexuality- or sexual identity-based offensive and/or degrading messages, comments, or other content”).

Of the 10 measures, 3 (30%) were validated, with the 21-item Technology-Facilitated Sexual Violence Victimization Scale being widely adapted and utilized [34].

Detailed Analysis of Validated Measures

To gain detailed insights into the most methodologically robust measures in the field, an in-depth analysis of the 55 validated measures was conducted. This subanalysis focused on their structural characteristics, including TFSVA-specific item length, measurement outcomes, target population, and psychometric properties.

The analysis focused on the TFSVA-specific item content within each validated measure. The number of items dedicated to assessing TFSVA constructs ranged from 1 to 63, with many measures embedding these items as a subscale within a broader questionnaire on general risky online behaviors or intimate partner violence. Validation was heavily concentrated in Western, high-income countries, notably the United States and Spain. Most measures targeted adolescent and young adult populations.

Psychometric reporting among validated measures followed a clear pattern. Internal consistency reliability (Cronbach alpha and McDonald ω) was commonly reported, with typical values exceeding 0.80. However, evidence for temporal stability (test-retest reliability) and construct validity (eg, convergent, discriminant, predictive) was reported less consistently and comprehensively. This indicates that the foundational reliability of TFSVA measures has been established, but the full spectrum of validation evidence is not yet universal. Detailed information on the validated measures and their use is synthesized in Multimedia Appendix 5.

Content of Validated Measures

Content Analysis Using STM

Topic interpretation primarily explains the topic words that describe each topic. STM identifies 9 topics based on high-frequency words. Figure 4 presents the top topics generated from unsupervised STM.

XSL•FO
RenderX

Figure 4. Summary of top topics identified by unsupervised structural topic modeling (STM). This figure presents the 9 most prevalent topics identified using unsupervised STM. For each topic, the bar represents overall prevalence, and the accompanying text lists the highest-probability keywords.

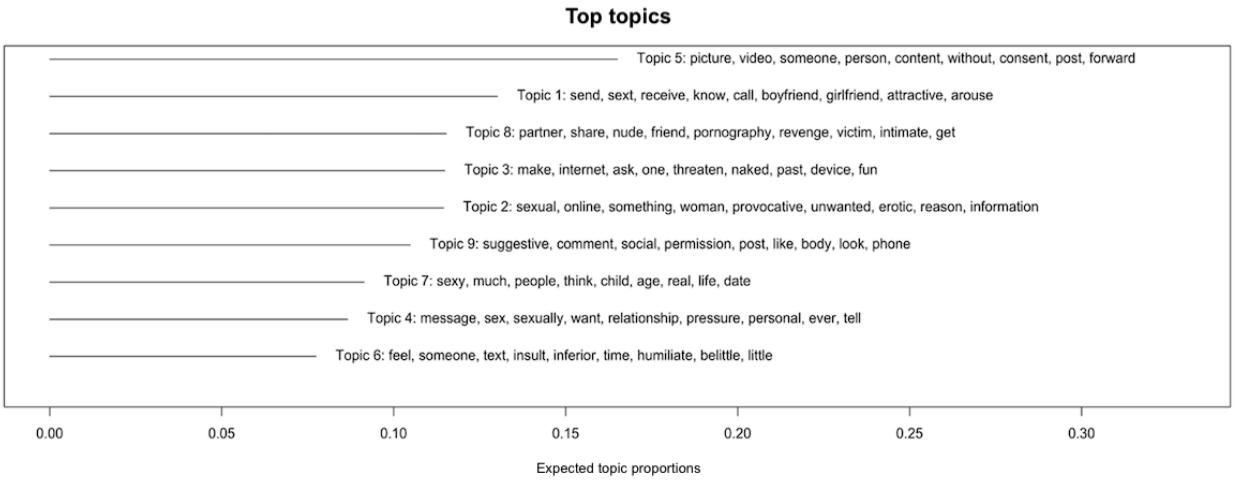

Topics were synthesized into 3 core dimensions representing the content of existing measures of TFSVA: (1) consent violation and coercive control; (2) status degradation and demeaning harassment; and (3) contextual and relational framing. Details of the measurement focus and representative topics for each theme are shown in Table 3.

Table 3. Core measurement dimensions identified from unsupervised structural topic modeling. This table presents the 3 primary dimensions identified through unsupervised structural topic modeling analysis of the included validated measures, listing their measurement focus and representative topics.

Dimensions	Measurement focus	Representative topics
Consent violation and coercive control	Measures acts that override consent, breach trust, and apply pressure or threats to dominate the survivor, representing an abuse of power.	3 (digital threats and control), 4 (sexual pressure online), and 5 (nonconsensual image distribution)
Status degradation and demeaning harassment	Measures acts designed to insult, humiliate, and lower the survivor's social standing, which causes social and psychological harm.	2 (unwanted sexual advances and perceived context), 6 (psychological abuse), and 9 (body shaming)
Contextual and relational framing	Provides the situational setting by defining the specific relationship contexts or vulnerable populations that characterize the abuse.	1 (pressured sexting in relationships), 7 (sexual exploitation of minors), and 8 (intimate partner revenge porn and betrayal)

The first dimension, consent violation and coercive control, encompasses measures of acts that override consent, breach authorization, or apply pressure. It is constituted by topics 3, 4, and 5. For instance, topic 5, the most prevalent topic, captures the nonconsensual distribution of intimate images, anchoring a central TFSVA construct in consent violation. Topics 3 and 4 complement this by measuring coercive tactics, with topic 3 focusing on digital threats and topic 4 capturing direct sexual pressure.

The second dimension, status degradation and demeaning harassment, captures measures of acts intended to insult, humiliate, and lower the survivor's social standing. Represented by topics 2, 6, and 9, it includes unwanted sexual advances that may carry attributions of blame (topic 2), psychological abuse and humiliation (topic 6), and body shaming in social media spaces (topic 9).

The final dimension, contextual and relational framing, defines the specific relational contexts and vulnerable populations that characterize TFSVA. It is represented by topics 1, 7, and 8. Topics 1 and 8 frame abuse within intimate relationships, measuring pressured sexting and revenge porn, respectively. Topic 7 captures online grooming and the sexual exploitation of minors.

Topic prevalence by types of TFSVA is shown in a heatmap (Figure 5). First, topic 5 (nonconsensual image distribution) demonstrates consistently high probability across all TFSVA types, suggesting it is a core element in TFSVA measures. Second, the relatively lower overall prevalence of topic 7 (online grooming and sexual exploitation of minors) may indicate that this form of abuse is underrepresented in general TFSVA instruments or is more often specialized in measures targeting child sexual exploitation. Third, the high specificity of certain topics—such as topic 8's strong association with IBSA and the notable absence of topics 3 (digital threats) and 6 (psychological abuse) from the same type—indicates that IBSA is operationalized primarily around the violation of consent for visual content rather than coercive or psychological tactics.

XSL•FO
RenderX

Figure 5. Topic prevalence by technology-facilitated sexual violence and abuse (TFSVA) type. This heatmap displays the prevalence of each STM topic (columns) across literature focused on specific TFSVA types (rows), highlighting which topics are central or peripheral to TFSVA measures.

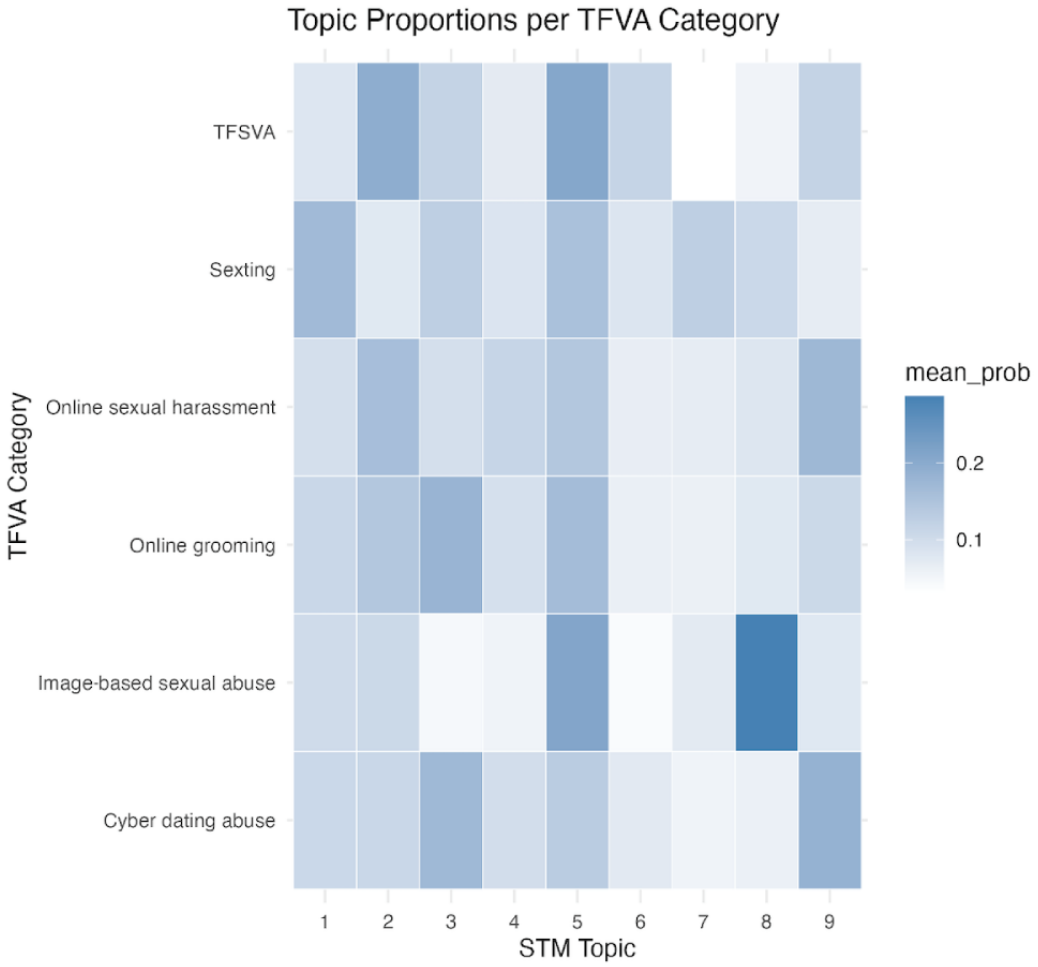

Thematic Analysis

Theme 1: TFSVA Acts and Experiences

Results revealed that the conceptual foundation of TFSVA measures is the direct quantification of concrete acts and exposure experiences. This behavioral core is operationalized through a mirrored structure across instruments: perpetration tools inventory behaviors enacted, while victimization tools inventory exposure to those same behaviors. This is evidenced by the parallelism in items such as a perpetration subscale asking, “In the past year, have you asked someone to send naked pictures of them to you? When received, I’ve forwarded or shared suggestive or sexual text messages of other people” [35,36], and a victimization scale probing, “I’ve received suggestive or sexual text messages” [35].

Beyond this dichotomy, measures capture critical behavioral specifics that define the phenomenon in the modern context. This included the nature of content shared (eg, “suggestive or sexual text messages/videos/images”), the exploitation of specific digital platforms (eg, “social media like Facebook, YouTube, Twitter, Instagram/messaging and communication apps like WhatsApp/online forums and chatrooms/websites and blogs”), and the relational dynamics between parties involved (eg, “friends or peers/online friends/stranger/romantic partner). Furthermore, instruments employed frequency metrics (eg, “number of times in the past 6 months that the participant had sent/received a sexual message”) to quantify experiences [37]. A few questionnaires also included behavioral continuums ranging from online harassment to physical violence (eg, “threatened on a digital device to physically hurt me”) [38].

Theme 2: TFSVA-Related Attitudes and Beliefs

The analysis identified that a distinct subset of instruments moved beyond measuring acts to assess the underlying cognitive frameworks that perpetuate TFSVA. These tools were designed to capture the beliefs that justify, minimize, or shift responsibility for abusive behaviors, and they revealed a critical triangulation of perspectives across different populations. Perpetrator-focused instruments measured aggression-supporting cognitions that legitimize harmful acts, exemplified by items asserting that “Women should be flattered if a partner or ex-partner shows nude pics of her to some close friends” [30]. Conversely, survivor-focused tools often assessed internalized blame, capturing the harmful internalization of these same societal myths through items such as, “Someone who sends a nude or sexual image to their partner, should not be surprised and is responsible if the image ends up online” [30].

XSL•FO
RenderX

Furthermore, instruments gauged societal-level myths and misconceptions that form the bedrock of these attitudes, including normative assertions like “It’s only natural for a guy to brag to his mates by showing them a nude or sexual image of his partner” [30].

Theme 3: Motivational Factors in TFSVA

This theme examined the complex motivational factors underlying TFSVA, capturing the distinct methodological approaches used to ascertain reasons for perpetration. The analysis revealed that motivations were conceptualized along a continuum spanning internally generated, autonomous reasons and externally driven, coerced reasons, typically measured through direct self-report questions posed to perpetrators. Autonomous motivations encompassed internally driven drivers such as relationship enhancement (eg, “sometimes I sent sexts to increase intimacy in my dating relationship”), sexual arousal (eg, “I sent sexts to feel sexually aroused”), and self-validation (eg, “I sent sexts to test whether I am attractive enough/ to verify whether my body is okay”) [24]. In direct contrast, coerced motivations were characterized by external pressure or instrumental gain, encompassing factors such as blackmail (eg, “sent threatening messages on digital devices”) [38], compliance with external pressures (eg, “After you sent the sext, did anyone praise you/did you feel like you fit in better with your friends/peers?”) [39], and the pursuit of tangible benefits such as financial rewards (eg, “offered me money or other things in exchange”) [40]. A critical methodological divergence was also identified, whereby perpetrator motivation was alternatively assessed through the survivor’s interpretation of the event (eg, “perpetrator did this to humiliate me/to feel a sense of dominance and control”) [41].

Theme 4: The Tactical and Contextual Dynamics of TFSVA

Measures of TFSVA demonstrated a sophisticated focus on the specific perpetrator tactics and contextual dynamics that define TFSVA. Instruments consistently captured a core set of direct harassment behaviors, quantifying the nonconsensual sharing of intimate images and the use of persistent online sexual threats through items probing experiences of “unwanted sexual communication” and “IBSA.”

Beyond these direct acts, instruments increasingly measured more nuanced and calculated strategies. This included deception and impersonation, assessed through items asking respondents whether they had been contacted by “fake profiles” or had experienced “false rumors about their sexual behaviors being spread online” [42]. A critical tactical subset involved overt coercion, prominently featured in the operationalization of sextortion. This was measured through items detailing how perpetrators manipulate survivors through threats, such as “Has anyone ever threatened to share your private images unless you sent money or more photos?,” which captures the dual mechanisms of financial coercion and sextortion [40].

Theme 5: Survivor Impact and Harm

Measures of impact revealed a primary focus on the psychological and relational consequences for survivors of TFSVA. Existing instruments predominantly assessed 2 core internalized harms: internalized blame and trauma-related emotional effects. The first focused on items that probe self-directed blame and responsibility, such as survivors endorsing statements such as “senders should expect images to be shared” [41]. The second captured the profound emotional fallout, quantifying experiences of humiliation, violated trust, and social anxiety through items like “victims find it difficult to trust others/often experience feelings of humiliation/being harassed or annoyed.” [41,43].

Beyond internal states, a significant portion of questions addressed broader relational outcomes, often framed by a tension between perceived positive effects (eg, intimacy-building within a consensual context) and the more frequently documented negative consequences of regret, trauma, and the reinforcement of power imbalances (eg, “Did you experience regret immediately after sending the sext?/I sent sexts because I am forced by someone”) [39,44]. Crucially, the nonconsensual nature of the acts was central to the conceptualization of harm. The profound impact was explicitly linked to the experience of violation, whether through the unauthorized sharing of intimate content, the coercion of sextortion, or the fundamental breach of privacy (eg, “disseminating or uploading to the Internet photos or videos of you with erotic or sexual content without your consent”) [45].

Integrated Synthesis

To integrate the findings from STM and thematic analysis, an integrated evidence and gap map was developed and is presented in Figure 6. This conceptual synthesis does not propose a direct correlation but serves as an evaluative framework. It positions the 3 core measurement dimensions identified through STM against the 5 themes derived from thematic analysis. For each dimension, judgment was made regarding its coverage of each thematic construct. The resulting matrix reveals a structural pattern in which robust measurement of behavior differs significantly from deficits in assessing attitudes, motivations, tactical dynamics, and harm-related components.

XSL•FO
RenderX

Figure 6. Synthesis of content coverage from structural topic modeling and thematic analysis. The table compares the 3 core measurement dimensions identified by structural topic modeling with the 5 themes derived from thematic analysis, indicating their coverage in the current evidence as strong, limited, or missing.

Dimensions \ Themes	Acts and experiences	Attitude and beliefs	Motivational factors	Tactical and contextual dynamics	Victim impact and harm
Violation of consent and coercion	●	◎	◎	◎	○
Status degradation and demeaning humiliation	●	○	○	◎	◎
Contextual and relational framing	●	◎	◎	○	○

● Strong/ good coverage ◎ Partial/ limited coverage ○ Gap/ missing coverage

Discussion

Summary of Evidence

This scoping review provides a novel diagnostic analysis of the TFSVA measurement field. Through the integration of STM and thematic analysis of 319 studies encompassing 218 distinct measures, we reveal a fundamental structural misalignment in the field. Current measures are heavily concentrated on cataloging specific behavioral acts while offering only limited or partial coverage of underlying attitudes, motivations, harm-based dimensions, and systemic contexts. This fragmentation helps explain why the field continues to struggle with conceptual and methodological challenges in capturing the multifaceted nature of TFSVA.

Interpretation and Comparison With Existing Literature

Scope and Definition of TFSVA

This review confirms the definitional fragmentation of TFSVA noted in prior literature. Previous studies have observed that TFSVA encompasses a diverse range of behaviors studied in isolation using inconsistent terminology [28]. Henry and Powell's [28] foundational work identified a critical lack of empirical research on TFSVA against adults and noted that only a few existing studies provide reliable data on the nature, scope, and impacts of TFSVA. Nearly a decade later, our synthesis provides empirical confirmation of this fragmentation. The field has produced an abundance of measures, yet they remain focused on specific behavioral subdomains with minimal conceptual integration.

Three key findings from our analysis directly inform the need for an updated definition. First, the narrow behavioral focus of existing measures reveals that current conceptualizations fail to capture the full spectrum of TFSVA. Second, our analysis identified emerging forms of TFSVA, such as AI-generated deepfakes, that are absent from existing measures. Third, our findings show that TFSVA operates within broader structural contexts that shape both the nature of abuse and its consequences.

Grounded in this evidence, we propose an updated definition that also engages with distinctions raised in restorative and transformative justice scholarship [46]. Building on this foundation, we define TFSVA as any form of sexual violence, abuse, coercion, or exploitation that is initiated, facilitated, intensified, or sustained through digital technologies. Such acts may result in psychological, reputational, social, economic, physical, or sexual harm, though the presence or absence of harm is not definitional to the act itself.

This formulation separates the act from its potential consequences. It acknowledges that (1) individuals may experience TFSVA without immediate measurable harm yet still warrant recognition and support; (2) the same act may produce different harms across individuals and contexts; and (3) harm may manifest differently across cultural and geographical settings [46].

The definition encompasses specific acts documented in our review, including IBSA, AI-generated sexual exploitation, cyberstalking leading to physical violence, and algorithmically amplified harassment. It remains adaptable to emerging forms not yet captured in existing measures. Crucially, it recognizes that TFSVA operates within systems of gendered, racialized, and intersectional oppression. These dimensions, which often transcend digital or physical boundaries and jurisdictional borders, are currently absent from existing measures.

Additionally, the definition embeds recognition of systemic oppression and acknowledges that technology is not merely a neutral tool. For instance, algorithms can amplify harassment, platform design may facilitate sexual abuse, and networked technologies enable harm across jurisdictional boundaries [47]. These dimensions follow directly from a definition centered on technological and structural contexts.

XSL•FO
RenderX

Temporal Trends

The increase in TFSVA publications observed in our temporal analysis suggests growing academic and societal recognition. A notable trend is the decline in studies on specific forms, such as sexting and IBSA, from 2011 to 2017. The use of the holistic term "TFSVA" began to rise sharply from 2018, suggesting a conceptual shift in which previously distinct behaviors are being consolidated under a broader umbrella [3]. Additionally, the decline in specific terms may reflect a move away from fragmented terminology toward more comprehensive approaches that capture the spectrum of technology-facilitated sexual harms. Alternatively, it may indicate that earlier research on sexting, IBSA, online sexual harassment, and online grooming reached a saturation point, prompting scholars to explore broader themes or emerging forms of TFSVA. Further investigation is needed to determine whether this trend reflects evolving academic priorities or a response to new technological developments that blur the boundaries between specific subtypes of TFSVA.

Gender, Geographic, and Cultural Considerations

A core finding of this review is the gap in demographic representation within TFSVA measures. Henry and Powell [28] observed that TFSVA may be predominantly gender-, sexuality-, and age-based, with young women being overrepresented as survivors. However, our analysis reveals that the field lacks the measures needed to adequately investigate these patterns.

The gendered nature of TFSVA is well documented, with women and sexual minorities disproportionately affected [48]. Emerging evidence also suggests that men experience TFSVA, albeit in different forms and with varying impacts [49,50]. A recent study assessing the content validity of a TFSVA measure further highlighted significant variations in how individuals of different genders and sexual orientations interpret and perceive TFSVA [51].

Yet, our synthesis reveals 2 critical gaps that prevent the field from advancing this understanding. First, existing measures may lack content sensitivity to diverse identities. Items often use generic language that may not capture experiences specific to sexual and gender minorities. Second, even if measures could capture diverse experiences, they have rarely been validated with diverse populations. Of the 319 studies, 314 (98.4%) did not specify participants' sexual orientation, and only 5 (1.6%) targeted gender minority populations. The implications of these gaps for measure design and validation are addressed in the subsequent discussion.

Moreover, the geographic focus of TFSVA research is heavily skewed. For instance, our results indicate that 260 (81.5%) of the 319 included studies were conducted in Western, high-income countries, while studies from Asian, African, and low-income contexts remain scarce. This imbalance is critical, as cultural norms and contextual factors, such as varying digital literacy, internet access, and legal frameworks, can significantly influence how TFSVA is perceived and experienced [3]. Consequently, existing measures may lack validity in these regions, highlighting the need to develop culturally adapted measures that account for specific social, legal, and technological environments.

Gaps and Recommendations

The thematic analysis of questionnaire content confirmed that existing measures capture behavioral data and certain attitudinal and motivational factors, but lack depth. Our integrated synthesis places these content gaps within a structural framework, revealing substantial limitations in addressing critical dimensions, notably evolving technological threats, intersectional vulnerabilities, and institutional factors.

Emerging technologies present challenges for TFSVA measures. Our analysis showed that emerging forms of TFSVA, such as AI, deepfakes, zoom-bombing or digital intrusion, and abuse via smart devices, were not covered in the measurement content. These gaps render existing questionnaires outdated and create a persistent development lag, as the evolution of abuse tactics outpaces instrument updates [51]. To remain relevant, future questionnaires should be dynamically updated to integrate assessments of emerging tactics, such as GPS stalking and AI-facilitated abuse.

Results from our analysis reveal limitations in capturing intersectional experiences at 2 levels. First, in terms of measure design, the content of existing measures lacks sensitivity to how TFSVA manifests across individuals of different genders and sexual orientations. Items often use generic language (eg, "Have you received unwanted sexual images?") that does not account for specific experiences among sexual and gender minority populations. This limits the validity of these measures for diverse groups. Second, in terms of study application, our descriptive synthesis shows that most included studies did not specify participants' sexual orientation. This indicates that studies rarely conduct and report the intersectional analyses needed to understand disparities.

Existing literature on other sexual violence measures offers concrete guidance for improving inclusivity. For example, the Daily Heterosexist Experiences Questionnaire includes items that specifically assess discrimination and harassment tied to sexual minority identity, such as "People assuming you are heterosexual because you have children," moving beyond generic violence exposure questions [52]. Similarly, measures of intimate partner violence for sexual and gender minority populations adapt items from the Revised Conflict Tactics Scale to capture experiences unique to these populations, including identity-based coercion and outing as a tactic of abuse [53]. These examples demonstrate that inclusive measures require consultation with diverse communities during item development to ensure content validity [52]. A second strategy is the use of language that reflects the lived experiences of specific groups rather than assuming universal applicability [53]. A third is validating measures separately across different populations to establish measurement equivalence [53]. Incorporating these strategies into TFSVA measure development is crucial for capturing nuanced harms, such as gendered and sexuality-based abuse [5]. Future measure development should prioritize inclusive language and item wording, community engagement in the design process, and rigorous validation across diverse gender and sexual orientation groups.

A related and critical gap is the lack of cross-cultural validity. Our findings show that most validated measures were developed

XSL•FO
RenderX

in Western, high-income countries. To move forward, the adaptation of TFSVA measures should follow a rigorous framework to ensure validity across different contexts [54]. This involves (1) ensuring conceptual equivalence through co-creation with local experts and cognitive testing; (2) conducting statistical measurement invariance testing (eg, multigroup confirmatory factor analysis); and (3) establishing criterion validity. This structured process moves beyond simple translation to develop measures that are valid for cross-cultural research and intervention.

Findings from the thematic analysis of measurement content indicate that current TFSVA-specific measures critically underassess structural and institutional factors, such as interactions with legal systems and digital platforms. There is a notable absence of items within TFSVA measures assessing key mediators of survivor outcomes, such as a survivor's trust in, or experience with, platform reporting systems, legal authorities, or support services. It is important to note that this finding is limited to the content of TFSVA-specific measures. Our review did not capture whether included studies administered additional, standalone measures assessing institutional trust, platform responsiveness, or help-seeking behaviors alongside their TFSVA measures. Therefore, although these constructs are absent from TFSVA-specific measures, we cannot draw conclusions regarding their overall availability or use elsewhere in the literature. Nevertheless, the absence of these constructs from TFSVA-specific measures has important implications. A growing body of evidence demonstrates that institutional responses profoundly influence survivor outcomes [55]. Research on situational betrayal theory has shown that platform reporting processes can themselves become sources of secondary trauma when they are hostile, opaque, or ineffective [56]. Institutional responses also shape bystander behaviors and help-seeking pathways [57]. For instance, bystanders' willingness to intervene is influenced by their perceptions of safety, responsibility, and the anticipated consequences of action, which are shaped by broader institutional and social norms [57]. Similarly, survivors' decisions to seek support are mediated by their trust in institutional responses and fear of further harm during or following disclosure [57]. Consequently, when TFSVA measures are used in isolation, they can quantify the abusive act but cannot assess the institutional and structural context in which harm is experienced and mediated. To build an evidence base for effective interventions, future TFSVA measure development should consider integrating items that assess constructs such as institutional trust, platform responsiveness, and multilevel help-seeking behaviors.

Our analysis of TFSVA-specific measures revealed limited attention to psychological and relational impacts, such as long-term trauma, reputational harm, or intimate partner dynamics. Results from STM and thematic analysis indicate that when impacts are included, they tend to frame consequences as isolated emotional states rather than examining how they interact with, or amplify, other mental health challenges [58]. However, this finding is limited to the content of TFSVA measures, as we did not capture whether included studies administered separate validated instruments assessing psychological outcomes alongside their TFSVA-specific measures. Future research should examine how TFSVA measures are used in conjunction with established mental health measures. Additionally, TFSVA measure development should consider integrating items that capture long-term and relational impacts to provide a more comprehensive understanding of survivors', perpetrators', and bystanders' outcomes.

Limitations

This scoping review has several limitations. First, by design, it maps the available evidence but does not appraise study quality or synthesize findings, which precludes definitive conclusions regarding measurement robustness. Second, STM was applied to short questionnaire items. The required text preprocessing may have fragmented the original wording; however, topics were interpreted against the original texts to ensure validity. Third, the inclusion of only English and Chinese publications may have introduced language bias, potentially omitting relevant studies. Fourth, as TFSVA is an evolving field, this review may not capture all emerging forms of the phenomenon. Fifth, our analysis focused specifically on TFSVA measures rather than all measures used within the included studies. Future research could examine how TFSVA measures are used alongside other measures assessing related constructs, such as institutional trust or mental health outcomes.

Conclusions

This scoping review provides the first comprehensive synthesis of existing TFSVA measures through the innovative integration of STM and thematic analysis. By examining 319 studies and 218 distinct measures, we move beyond cataloging to systematically identify the core dimensions reflected in existing measures, such as consent violations, coercive control, and contextual-relational framing. The review also reveals a field that is structurally misaligned with the realities of TFSVA: (1) a lack of measures addressing emerging AI-facilitated tactics; (2) the systematic omission of intersectional factors, such as gender and sexuality; (3) insufficient attention to structural and institutional mediators, such as platform responsiveness; and (4) superficial assessments of psychological and relational impacts, including long-term trauma. Importantly, these findings reflect gaps within TFSVA-specific measures rather than the broader literature, as we did not capture whether studies administered separate measures assessing these constructs alongside their TFSVA measures.

This approach quantifies the field's conceptual fragmentation and pinpoints empirical blind spots. The primary contribution is an evidence-based analysis that redirects measures from narrow, tactic-focused content toward capturing more diverse contextual dimensions. To address the identified gaps, future efforts must prioritize: (1) developing measures with inclusive, community-informed content that reflects diverse experiences across gender, sexuality, and cultural contexts; (2) validating measures across diverse populations to establish measurement equivalence; and (3) integrating items that assess institutional and structural factors shaping TFSVA outcomes. This shift is essential for generating the comprehensive evidence needed to inform responsive policy, platform governance, and social support services in the rapidly evolving digital age.

XSL•FO
RenderX

Data Availability

This study analyzed secondary data sourced from the existing published literature. The complete set of references is provided in Multimedia Appendix 4.

Funding

This study was funded by the General Research Fund, Research Grants Council, Hong Kong (reference: 17617721).

Authors' Contributions

Conceptualization: EPHC (lead), SHLP (equal)
Data curation: SHLP (lead), CW (supporting)
Formal analysis: SHLP (lead), LZ (supporting)
Supervision: EPHC (lead), PHC (equal), CBJ (equal)
Writing – original draft: SHLP
Writing – review & editing: SHLP (lead), EPHC (equal), PHC (equal), CBJ (equal), LZ (supporting), CW (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Keywords and search strategies.
[DOCX File , 33 KB - jmir_v28i1e90068_app1.docx]

Multimedia Appendix 2
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 99 KB - jmir_v28i1e90068_app2.pdf]

Multimedia Appendix 3
PRISMA-S (PRISMA Statement for Reporting Literature Searches in Systematic Reviews) checklist. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.
[PDF File (Adobe PDF File), 83 KB - jmir_v28i1e90068_app3.pdf]

Multimedia Appendix 4
All included references.
[DOCX File , 81 KB - jmir_v28i1e90068_app4.docx]

Multimedia Appendix 5
Summary of validated measures.
[DOCX File , 46 KB - jmir_v28i1e90068_app5.docx]

References

1. Henry N, Flynn A, Powell A. Policing image-based sexual abuse: stakeholder perspectives. Police Practice and Research 2018 Sep 20;19(6):565-581. [doi: 10.1080/15614263.2018.1507892]
2. Buil-Gil D, Kemp S, Kuenzel S, Coventry L, Zakhary S, Tilley D, et al. The digital harms of smart home devices: a systematic literature review. Computers in Human Behavior 2023 Aug;145:107770. [doi: 10.1016/j.chb.2023.107770]
3. Sheikh MMR, Rogers MM. Technology-facilitated sexual violence and abuse in low and middle-income countries: a scoping review. Trauma Violence Abuse 2024 Apr 31;25(2):1614-1629 [FREE Full text] [doi: 10.1177/15248380231191189] [Medline: 37650241]
4. Salerno-Ferraro AC, Erentzen C, Schuller RA. Young women's experiences with technology-facilitated sexual violence from male strangers. J Interpers Violence 2022 Oct 16;37(19-20):NP17860-NP17885 [FREE Full text] [doi: 10.1177/08862605211030018] [Medline: 34271829]
5. Powell A, Scott AJ, Henry N. Digital harassment and abuse: experiences of sexuality and gender minority adults. European Journal of Criminology 2018 Jul 30;17(2):199-223. [doi: 10.1177/1477370818788006]
6. Henry N, Flynn A, Powell A. Technology-facilitated domestic and sexual violence: a review. Violence Against Women 2020 Dec;26(15-16):1828-1854. [doi: 10.1177/1077801219875821] [Medline: 32998673]

XSL•FO
RenderX

7. Patel U, Roesch R. The prevalence of technology-facilitated sexual violence: a meta-analysis and systematic review. Trauma Violence Abuse 2022 Apr 15;23(2):428-443. [doi: 10.1177/1524838020958057] [Medline: 32930064]
8. Champion A, Oswald F, Pedersen CL. Technology-facilitated sexual violence and suicide risk: a serial mediation model investigating bullying, depression, perceived burdensomeness, and thwarted belongingness. The Canadian Journal of Human Sexuality 2021 Apr 01;30(1):125-141. [doi: 10.3138/cjhs.2020-0044]
9. Bates S. Revenge porn and mental health. Feminist Criminology 2016 Jul 24;12(1):22-42. [doi: 10.1177/1557085116654565]
10. Snaychuk LA, O'Neill ML. Technology-facilitated sexual violence: prevalence, risk, and resiliency in undergraduate students. Journal of Aggression, Maltreatment & Trauma 2020 Jan 28;29(8):984-999. [doi: 10.1080/10926771.2019.1710636]
11. Dragiewicz M, Burgess J, Matamoros-Fernández A, Salter M, Suzor NP, Woodlock D, et al. Technology facilitated coercive control: domestic violence and the competing roles of digital media platforms. Feminist Media Studies 2018 Mar 20;18(4):609-625. [doi: 10.1080/14680777.2018.1447341]
12. Peters M. Updated methodological guidance for the conduct of scoping reviews. JBI evidence synthesis 2020;18(10):2119-2126. [doi: 10.11124/jbies-20-00167]
13. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. International Journal of Social Research Methodology 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
14. Pak SHL, Choi EPH, Wu C, Chau PH, Bradbury-Jones C. Measuring technology-facilitated sexual violence and abuse: a scoping review protocol of existing measurements. BMJ Open 2024 Dec 27;14(12):e089296 [FREE Full text] [doi: 10.1136/bmjopen-2024-089296] [Medline: 39732482]
15. Measuring technology-facilitated sexual violence: a scoping review protocol of existing instruments. Open Science Framework. 2024 Mar 13. URL: https://archive.org/details/osf-registrations-q5etw-v1 [accessed 2026-03-28]
16. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, Tunçalp, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473. [doi: 10.7326/m18-0850]
17. Pollock D, Peters MDJ, Khalil H, McInerney P, Alexander L, Tricco AC, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. JBI evidence synthesis 2023;21(3):520-532. [doi: 10.11124/jbies-22-00123]
18. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Systematic reviews 2021;10(1):a-19. [doi: 10.31219/osf.io/sfc38]
19. Bradbury-Jones C, Aveyard H, Herber OR, Isham L, Taylor J, O'Malley L. Scoping reviews: the PAGER framework for improving the quality of reporting. International Journal of Social Research Methodology 2021 Mar 18;25(4):457-470. [doi: 10.1080/13645579.2021.1899596]
20. Beckmeyer JJ, Herbenick D, Eastman-Mueller H. Sexting with romantic partners during college: who does it, who doesn't and who wants to. Sexuality & Culture 2021 May 31;26(1):48-66. [doi: 10.1007/s12119-021-09878-x]
21. Gillies M, Murthy D, Brenton H, Olaniyan R. Theme and topic: how qualitative research and topic modeling can be brought together. arXiv. Published online May 11, 2021 [FREE Full text] [doi: 10.48550/arXiv.2210.00707]
22. Clarke V, Braun V. Thematic analysis. The Journal of Positive Psychology 2016 Dec 09;12(3):297-298. [doi: 10.1080/17439760.2016.1262613]
23. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
24. Bianchi D, Morelli M, Baiocco R, Chirumbolo A. Psychometric properties of the Sexting Motivations Questionnaire for adolescents and young adults. Rassegna di Psicologia 2016;35(3):5-18.
25. Dir AL, Coskunpinar A, Steiner JL, Cyders MA. Understanding differences in sexting behaviors across gender, relationship status, and sexual identity, and the role of expectancies in sexting. Cyberpsychol Behav Soc Netw 2013 Aug;16(8):568-574. [doi: 10.1089/cyber.2012.0545] [Medline: 23675996]
26. Gámez-Guadix M, Almendros C, Borrajo E, Calvete E. Prevalence and association of sexting and online sexual victimization among Spanish adults. Sex Res Soc Policy 2015 Mar 14;12(2):145-154. [doi: 10.1007/s13178-015-0186-9]
27. Weisskirch RS, Delevi R. "Sexting" and adult romantic attachment. Computers in Human Behavior 2011 Sep;27(5):1697-1701. [doi: 10.1016/j.chb.2011.02.008]
28. Henry N, Powell A. Technology-facilitated sexual violence: a literature review of empirical research. Trauma Violence Abuse 2018 Apr 16;19(2):195-208. [doi: 10.1177/1524838016650189] [Medline: 27311818]
29. Muñoz-Fernández N, Del Rey R, Mora-Merchán JA, Elipe P. Sexual diversity bullying and cyberbullying questionnaires: an inclusive approach to measure sexuality-based bullying. J Homosex 2024 Oct 14;71(12):2740-2763. [doi: 10.1080/00918369.2023.2260922] [Medline: 37756388]
30. Powell A, Henry N, Flynn A, Scott AJ. Image-based sexual abuse: the extent, nature, and predictors of perpetration in a community sample of Australian residents. Computers in Human Behavior 2019 Mar;92:393-402. [doi: 10.1016/j.chb.2018.11.009]
31. Whittle H, Hamilton-Giachritsis C, Beech A, Collings G. A review of young people's vulnerabilities to online grooming. Aggression and Violent Behavior 2013 Jan;18(1):135-146. [doi: 10.1016/j.avb.2012.11.008]

XSL•FO
RenderX

32. Gámez-Guadix M, De Santisteban P, Alcazar. The construction and psychometric properties of the questionnaire for online sexual solicitation and interaction of minors with adults. Sex Abuse 2018 Dec 18;30(8):975-991. [doi: 10.1177/1079063217724766] [Medline: 28821214]
33. Borrajo E, Gámez-Guadix M, Pereda N, Calvete E. The development and validation of the cyber dating abuse questionnaire among young couples. Computers in Human Behavior 2015 Jul;48:358-365. [doi: 10.1016/j.chb.2015.01.063]
34. Powell A, Henry N. Technology-facilitated sexual violence victimization: results from an online survey of Australian adults. J Interpers Violence 2019 Sep 03;34(17):3637-3665. [doi: 10.1177/0886260516672055] [Medline: 27697966]
35. del Rey R, Ojeda M, Casas J. Validation of the Sexting Behavior and Motives Questionnaire (SBM-Q). Psicothema 2021 May 01;33(2):287-295. [doi: 10.7334/psicothema2020.207] [Medline: 33879302]
36. Choi HJ, Mori C, Van Ouytsel J, Madigan S, Temple JR. Adolescent sexting involvement over 4 years and associations with sexual activity. J Adolesc Health 2019 Dec;65(6):738-744 [FREE Full text] [doi: 10.1016/j.jadohealth.2019.04.026] [Medline: 31277988]
37. Lucić M, Baćak V, Štulhofer A. The role of peer networks in adolescent pornography use and sexting in Croatia. Journal of Children and Media 2019 Jul 10;14(1):110-127. [doi: 10.1080/17482798.2019.1637356]
38. Brown C, Hegarty K. Development and validation of the TAR Scale: a measure of technology-facilitated abuse in relationships. Computers in Human Behavior Reports 2021 Jan;3:100059. [doi: 10.1016/j.chbr.2021.100059]
39. Cornelius TL, Bell KM, Kistler T, Drouin M. Consensual sexting among college students: the interplay of coercion and intimate partner aggression in perceived consequences of sexting. Int J Environ Res Public Health 2020 Sep 29;17(19):7141 [FREE Full text] [doi: 10.3390/ijerph17197141] [Medline: 33003567]
40. Gámez-Guadix M, De Santisteban P, Wachs S, Wright M. Unraveling cyber sexual abuse of minors: psychometrics properties of the Multidimensional Online Grooming Questionnaire and prevalence by sex and age. Child Abuse Negl 2021 Oct;120:105250. [doi: 10.1016/j.chiabu.2021.105250] [Medline: 34399230]
41. Harper CA, Smith L, Leach J, Daruwala NA, Fido D. Development and validation of the Beliefs About Revenge Pornography Questionnaire. Sex Abuse 2023 Sep 05;35(6):748-783 [FREE Full text] [doi: 10.1177/10790632221082663] [Medline: 35382636]
42. Sánchez V, Muñoz-Fernández N, Vega-Gea E. Peer sexual cybervictimization in adolescents: development and validation of a scale. Int J Clin Health Psychol 2017 May;17(2):171-179 [FREE Full text] [doi: 10.1016/j.ijchp.2017.04.001] [Medline: 30487892]
43. López-Cepero J, Vallejos-Saldarriaga J, Merino-García M. Digital intimate partner violence among Peruvian youths: validation of an instrument and a theoretical proposal. J Interpers Violence 2021 Jun 03;36(11-12):5167-5185. [doi: 10.1177/0886260518803610] [Medline: 30280639]
44. Oswald F, Lopes A, Skoda K, Hesse CL, Pedersen CL. I'll show you mine so you'll show me yours: motivations and personality variables in photographic exhibitionism. J Sex Res 2020 Jul 18;57(5):597-609. [doi: 10.1080/00224499.2019.1639036] [Medline: 31318606]
45. Penado M, Rodicio-García ML, Marcos Cuesta M, Corrás T. Construction and validation of the intimate images diffusion scale among adolescents. Front Psychol 2019 Jun 28;10:1485 [FREE Full text] [doi: 10.3389/fpsyg.2019.01485] [Medline: 31316439]
46. McMahon S, Christensen MC, Todić J. Transformative justice and restorative justice approaches to campus sexual assault: a scoping review. International journal of restorative justice 2024;7(1):12-42. [doi: 10.5553/tijrj.000156]
47. Gomes AB, Sultan A. Problematizing content moderation by social media platforms and its impact on digital harm reduction. Harm Reduct J 2024 Nov 09;21(1):194 [FREE Full text] [doi: 10.1186/s12954-024-01104-9] [Medline: 39516798]
48. Gámez-Guadix M, Sorrel MA, Martínez-Bacaicoa J. Technology-facilitated sexual violence perpetration and victimization among adolescents: a network analysis. Sex Res Soc Policy 2022 Dec 09;20(3):1000-1012. [doi: 10.1007/s13178-022-00775-y]
49. Champion AR, Oswald F, Khera D, Pedersen CL. Examining the gendered impacts of technology-facilitated sexual violence: a mixed methods approach. Arch Sex Behav 2022 Apr 18;51(3):1607-1624. [doi: 10.1007/s10508-021-02226-y] [Medline: 35182285]
50. Choi EPH, Wong JYH, Fong DYT. An emerging risk factor of sexual abuse: the use of smartphone dating applications. Sex Abuse 2018 Jun 02;30(4):343-366. [doi: 10.1177/1079063216672168] [Medline: 27698090]
51. Pak SHL, Wu C, Choi KWY, Choi EPH. Measuring technology-facilitated sexual violence and abuse in the Chinese context: development study and content validity analysis. JMIR Form Res 2024 Nov 19;8:e65199 [FREE Full text] [doi: 10.2196/65199] [Medline: 39561365]
52. Balsam KF, Beadnell B, Molina Y. The Daily Heterosexist Experiences Questionnaire: measuring minority stress among lesbian, gay, bisexual, and transgender adults. Meas Eval Couns Dev 2013 Jan 09;46(1):3-25 [FREE Full text] [doi: 10.1177/0748175612449743] [Medline: 24058262]
53. Dyar C, Messinger AM, Newcomb ME, Byck GR, Dunlap P, Whitton SW. Development and initial validation of three culturally sensitive measures of intimate partner violence for sexual and gender minority populations. J Interpers Violence 2021 Aug 05;36(15-16):NP8824-NP8851 [FREE Full text] [doi: 10.1177/0886260519846856] [Medline: 31057032]

XSL•FO
RenderX

54. Boateng GO, Neilands TB, Frongillo EA, Melgar-Quiñonez HR, Young SL. Best practices for developing and validating scales for health, social, and behavioral research: a primer. Front Public Health 2018 Jun 11;6:149 [FREE Full text] [doi: 10.3389/fpubh.2018.00149] [Medline: 29942800]
55. Sandoval M, de Almeida Vau M, Solaas J, Rodrigues L. Threat of deepfakes to the criminal justice system: a systematic review. Crime Sci 2024 Nov 17;13(1):41-46. [doi: 10.1186/s40163-024-00239-1]
56. Christl M, Pham KT, Rosenthal A, DePrince AP. When institutions harm those who depend on them: a scoping review of institutional betrayal. Trauma Violence Abuse 2024 Oct 22;25(4):2797-2813. [doi: 10.1177/15248380241226627] [Medline: 38258307]
57. Kromann SB, Flynn A. Bystander intervention in image-based sexual abuse: a scoping review. Trauma Violence Abuse 2025 Jun 27:15248380251343191. [doi: 10.1177/15248380251343191] [Medline: 40574536]
58. Oram S. Sexual violence and mental health. Epidemiol Psychiatr Sci 2019 Apr 12;28(6):592-593. [doi: 10.1017/s2045796019000106]

Abbreviations

AI: artificial intelligence
IBSA: image-based sexual abuse
LGBTQ+: lesbian, gay, bisexual, transgender, queer (or questioning), and others not explicitly listed
PAGER: Patterns, Advances, Gaps, Evidence for practice, and Research Recommendations
PCC: Population-Concept-Context
PRISMA-S: PRISMA Statement for Reporting Literature Searches in Systematic Reviews
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
STM: structural topic modeling
TFSVA: technology-facilitated sexual violence and abuse

Edited by S Brini; submitted 21.Dec.2025; peer-reviewed by L Graham, KS Chan; comments to author 11.Jan.2026; revised version received 02.Mar.2026; accepted 19.Mar.2026; published 14.Apr.2026.

Please cite as:
Pak SHL, Zhang L, Wu C, Chau PH, Bradbury-Jones C, Choi EPH
Measuring Technology-Facilitated Sexual Violence and Abuse: Scoping Review of Existing Measures
J Med Internet Res 2026;28:e90068
URL: https://www.jmir.org/2026/1/e90068
doi:10.2196/90068
PMID:

©Sharon Hoi Lam Pak, Linruo Zhang, Chanchan Wu, Pui Hing Chau, Caroline Bradbury-Jones, Edmond Pui Hang Choi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Multimodal Intelligent Monitoring of Parkinson Disease: Scoping Review of Progress and Translational Challenges

Jianhui Tan[1], BD; Xiuzhen Deng[2], BD; Cuilan Wu[1], BD; Molin Yao[1], BD; Jingyi Liao[1], BD; Haoli Zheng[1,3], LLM; Changlin Lian[4], PhD

[1]The 2nd Clinical Medical College, Southern Medical University, Guangzhou, Guangdong, China

[2]School of Nursing, Southern Medical University, Guangzhou, Guangdong, China

[3]Zhujiang Hospital, Southern Medical University, Guangzhou, Guangdong, China

[4]Guangdong Cardiovascular Institute, Guangdong Provincial People's Hospital, Southern Medical University, Guangzhou, Guangdong, China

Corresponding Author:
Changlin Lian, PhD
Guangdong Cardiovascular Institute
Guangdong Provincial People's Hospital
Southern Medical University
No. 106, Zhongshan Road, Yuexiu District
Guangzhou, Guangdong, 510080
China
Phone: 86 1 324 274 0191
Email: lianchanglin@hotmail.com

Abstract

Background: Parkinson disease (PD) is a progressive neurodegenerative disorder with a rapidly growing global prevalence. Current clinical assessments, such as the Unified Parkinson Disease Rating Scale, are limited by subjectivity and episodic application, creating a need for continuous, objective monitoring solutions. While previous reviews have often focused on single technologies, there is a growing trend toward integrating multiple data sources to provide a more holistic view of PD.

Objective: This scoping review synthesizes progress in multimodal intelligent monitoring systems for PD, focusing on the quantification of motor and nonmotor symptoms, algorithm development, and the clinical translation of remote monitoring platforms. Furthermore, we propose a novel heuristic framework (Care-Platform Transformation in PD [CPT-PD]) that provides a forward-looking conceptual design for integrating these technologies into clinical workflows, demonstrating promising potential for future development.

Methods: A targeted literature search was conducted on August 15, 2025, in PubMed, Web of Science, and China National Knowledge Infrastructure for research published between January 1, 2019, and December 31, 2024. The final search was rerun on January 22, 2026, solely to ensure completeness of coverage for this time window; no articles published after December 31, 2024, were included.

Results: Wearable sensors (n=9) demonstrated high concordance with clinical scores in validation studies (eg, 99% for tremor detection), while computer vision (n=6) achieved moderate agreement with clinician ratings in controlled assessments (intraclass correlation coefficient 0.74 for bradykinesia). For nonmotor symptoms, intelligent systems (n=7) demonstrated sleep disturbance detection with up to 92.9% accuracy and autonomic dysfunction monitoring (n=7) via heart rate variability (area under the curve 0.90) and voice analysis (94.55% accuracy). Algorithm studies (n=16) explored single-modality feature extraction and cross-modal fusion, with emerging applications in federated learning. Remote platforms (n=22) improved medication adherence (172/201, 85.6%) and reduced outpatient visits (by 29% in one study). A heuristic CPT-PD framework was proposed to integrate key components of diagnosis, treatment, and management. Collectively, these advancements demonstrate the technical viability and clinical benefits of shifting from episodic, subjective assessments toward a data-driven, continuous, and multimodal approach to PD management.

Conclusions: While current evidence largely reflects multisensor systems rather than deeply integrated multimodal platforms, the field holds promise for advancing toward genuine data fusion that could further improve clinical decision-making. Persistent challenges include fragmented symptom focus, algorithmic heterogeneity, and barriers to adoption among older adults. Future

XSL•FO
RenderX

efforts should build on integrated frameworks such as CPT-PD to develop patient-centered ecosystems, ultimately enabling precision medicine in PD management.

(J Med Internet Res 2026;28:e89057) doi:10.2196/89057

KEYWORDS

Parkinson disease; multimodal intelligent monitoring; wearable sensors; algorithms; remote monitoring platforms

Introduction

Parkinson disease (PD) is a chronic, progressive neurodegenerative disorder that predominantly affects individuals in middle-to-late adulthood. Classic motor symptoms include bradykinesia, muscular rigidity, tremor, and postural instability, which collectively have a pronounced impact on patient quality of life and the ability to perform daily activities [1]. As the global population continues to age, the prevalence of PD is rising sharply—projections suggest that by 2050, PD will affect 25.2 million people worldwide, representing a 112% increase compared with 2021 [2]. This mounting burden not only impairs individual quality of life but also places escalating pressure on health care systems [3].

Current clinical diagnosis of PD relies heavily on the Unified Parkinson Disease Rating Scale (UPDRS), supplemented by limited laboratory investigations such as magnetic resonance imaging, computed tomography, positron emission tomography, and other modalities for assessing brain metabolism or dopamine synthesis [1]. Nevertheless, these diagnostic approaches have major limitations: they are highly subjective, exhibit substantial interrater variability, and existing methods are insufficient for accurate, real-time monitoring of motor fluctuations in home environments. Consequently, clinicians often lack timely insights into changes in patients' condition, which hinders optimal adjustment of medication dosages and treatment plans [4]. An urgent need exists for continuous monitoring tools (wearables and smartphones) that can enhance clinical insight while reducing patient burden.

Recent years have witnessed a rapid expansion of intelligent software technologies tailored for chronic disease management. Advances in wearable sensors, the adoption of sophisticated algorithms, and the development of remote monitoring platforms have opened novel avenues for PD assessment. Among these, remote monitoring platforms—integrating diverse sensing modalities (eg, motion, physiological, speech, vision) with advanced algorithms—have emerged as a focal point in research, aiming to bridge gaps in traditional evaluation methods and support precise, home-based monitoring.

It is important to clarify that, although the term "multimodal" conceptually implies deep algorithmic fusion of heterogeneous data streams, the current technological landscape for PD monitoring largely consists of multisensor systems that collect data from multiple sources but analyze them in parallel or with limited integration. Throughout this review, we use "multimodal intelligent monitoring" as a broad descriptor encompassing this spectrum, and we explicitly indicate when a system achieves genuine multimodal fusion versus when it operates at a multisensor level.

This review focuses on progress in remote monitoring platforms and the management of PD from 2019 to 2024, with an emphasis on 3 primary domains:

- Clinical performance of wearable sensors: assessment of their utility and patient compliance in quantifying PD symptoms, including both motor and nonmotor manifestations.
- Algorithmic approaches: analysis of unimodal and cross-modal feature extraction strategies, evaluating efficiency and persisting challenges.
- Clinical translation of remote monitoring platforms: appraisal of current integration status, clinical outcomes, and translational obstacles.

By systematically reviewing advances and challenges in these areas, this review aims to inform future optimization and successful clinical adoption of closed-loop PD management systems, emphasizing a comprehensive "diagnosis-treatment-management" approach.

Methods

Search Strategy and Study Identification

This review employed a targeted, purposive sampling method for the literature search rather than undertaking a fully exhaustive systematic review. This strategy aimed to identify representative, high-impact research published between 2019 and 2024 that documents essential breakthroughs and persistent hurdles in remote monitoring platforms for PD. The methodological design and reporting conformed to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews; Multimedia Appendix 1) extension guidelines, thereby ensuring transparency and rigor throughout the review process [5]. A comprehensive literature search was initially conducted on August 15, 2025, and then rerun on January 22, 2026.

Key Term Definitions

To ensure conceptual clarity and facilitate understanding for readers from diverse backgrounds (including clinicians), we define the key technical terms used throughout this review in Table 1. Readers are encouraged to refer to this table when encountering these terms in the following sections.

XSL•FO
RenderX

Table 1. Definitions of the term "multimodal intelligent monitoring," "remote monitoring platform," and "(platform) integration."

Key terms	Definition
Multimodal intelligent monitoring	Systems that collect and analyze data from at least two sensing modalities (eg, motion, speech, or vision) for Parkinson disease assessment. Approaches range from parallel data analysis in multisensor setups to advanced joint algorithmic integration, although true multimodal fusion remains challenging. Throughout this review, we use the umbrella term "multimodal intelligent monitoring" to encompass the full spectrum from basic multisensor data collection to advanced algorithmic integration, while explicitly noting the level of fusion achieved by specific systems when evidence permits.
Remote monitoring platform	An integrated software-hardware system enabling continuous, autonomous collection, transmission, and processing of patient data outside clinical settings (primarily at home). It delivers actionable insights to health care providers, patients, or both, to support clinical decision-making, with continuous monitoring as its core function.
(Platform) integration	The seamless interfacing and synergistic operation of diverse data sources, algorithms, user interfaces, and clinical workflows within a remote monitoring platform, creating a cohesive system with greater combined utility than the sum of its parts.
Wearable sensor	A body-worn device (eg, wristwatch, patch, insole) incorporating inertial measurement units or physiological sensors to continuously capture motor and nonmotor signals in free-living environments.
Computer vision	An artificial intelligence field that employs cameras and deep learning algorithms (eg, pose estimation) to quantitatively analyze human movement and behavior from video, enabling contactless assessment of motor symptoms.
Digital biomarker	A physiological or behavioral indicator collected via digital tools (eg, sensors, smartphones), such as gait speed or tremor frequency, used to measure disease presence, severity, or progression.

Search Strategy

We conducted a comprehensive literature search in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) guidelines to ensure transparent and reproducible reporting [6]. We searched 3 electronic databases—PubMed (National Center for Biotechnology Information website), Web of Science Core Collection, and CNKI (China National Knowledge Infrastructure)—from January 1, 2019, to December 31, 2024. The searches were conducted initially on August 15, 2025, and updated on January 22, 2026. It is important to note that the updated search in 2026 was performed solely to verify the completeness of the 2019–2024 period; no articles published after December 31, 2024, were included in this review. We performed the search in each database individually, adapting the search strings accordingly.

To maximize sensitivity and precision, we developed a structured search strategy based on 4 core concept groups:

- Population: PD.
- Sensing modalities: wearable, inertial measurement unit (IMU), computer vision, speech.
- Intelligent methods: machine learning, deep learning, algorithms, multimodal fusion, digital biomarker.
- Platform level: remote monitoring, telemedicine, mobile health (mHealth), digital health.

Synonyms and related terms within each group were combined using the OR operator, and the 4 groups were then combined using AND.

Additionally, the 3-step search strategy recommended by the Joanna Briggs Institute for scoping reviews was employed to maximize sensitivity and comprehensiveness while balancing feasibility [7]. We used phrase searching (with quotation marks) and other precision techniques to ensure accurate retrieval. The full search strategies for all databases, including the exact search strings, filters, and limits used, are provided in Multimedia Appendix 2. For CNKI, the search terms were carefully translated from English, and the search was limited to peer-reviewed journal articles.

We used CiteSpace 6.3.R1 to remove duplicate records identified across the different databases after the initial retrieval. The primary selection of studies was then performed through manual screening of titles, abstracts, and full texts to identify representative and landmark studies that document essential breakthroughs and persistent hurdles in remote monitoring platforms for PD. In addition to database searching, we performed citation searching (snowballing) by manually reviewing the reference lists of all included studies and relevant key reviews to identify additional eligible articles [8].

We did not contact study authors or subject experts to identify unpublished or additional studies. The search strategy was not formally peer reviewed by an information specialist or librarian.

Eligibility Criteria

Study Type

(1) Eligible studies include original research involving human participants (such as randomized controlled trials, cohort studies, case-control studies, cross-sectional studies) or comprehensive reviews (such as systematic reviews and scoping reviews). (2) Studies must be published between January 1, 2019, and December 31, 2024. (3) Only publications in English or Chinese are considered. (4) All included literature must be peer-reviewed, formal publications.

Participants

The study population must consist of patients with clinically diagnosed PD, with no restrictions on age, disease stage, or treatment background.

XSL•FO
RenderX

Core Concept

Studies must closely revolve around the theme of "monitoring technologies for Parkinson disease." Specifically, they should report data or provide in-depth discussion on at least one of the following aspects: (1) *technological performance*, such as the accuracy, sensitivity, specificity, reliability, or feasibility of the monitoring technology; (2) *clinical validation*, such as the effectiveness of the technology in real-world or clinical settings, its correlation with clinical assessment outcomes, or its impact on the clinical decision-making process; and (3) *implementation challenges*, such as barriers, limitations, cost, accessibility, or ethical issues encountered during the practical promotion and application of the technology.

Context

This study does not impose restrictions on the application settings of the technologies (eg, home, community, clinic, hospital) or on geographic or cultural backgrounds, aiming to comprehensively gather relevant evidence from a global perspective.

Exclusion Criteria

(1) Non–peer-reviewed publications, including editorials, commentaries, letters, conference abstracts (unless they provide sufficiently detailed technical data), and preprints; (2) studies limited to animal models, in vitro experiments (cell studies), or primarily focused on neurodegenerative diseases other than PD; (3) publications not specifically centered on PD monitoring or those that mention the topic only incidentally; and (4) duplicate publications or instances in which the full text is unavailable.

Study Selection and Data Extraction

Duplicates were eliminated using CiteSpace 6.3.R1, after which 2 independent reviewers (JHT and XZD) screened article titles and abstracts for relevance. The full texts of potentially eligible studies were then evaluated, with any disagreements resolved by consensus or by involving a third reviewer (CLW). The study selection workflow is outlined in a PRISMA-ScR flow diagram (Figure 1).

Analytical Framework and Synthesis

This review utilized an a priori analytical framework organized into 3 key technological domains: (1) Symptom Assessment (Hardware Systems); (2) Algorithmic Implementation (Software Methods); and (3) Remote Monitoring Platforms (Integrated Systems).

Studies meeting the eligibility criteria were thematically classified into the respective technological domains to facilitate structured synthesis. For each study, relevant data were extracted concerning technological features, performance metrics, and identified limitations.

Limitations of Method

This synthesis does not incorporate risk-of-bias assessments or evidence grading due to its nature as a scoping review rather than a full systematic review. The heterogeneity observed in study metrics, technological modalities, and patient populations precluded the conduct of meta-analyses. Nevertheless, this evidence mapping helps identify methodological gaps for future regulatory-compliant research.

Substantial clinical and methodological diversity was identified among the included studies, precluding direct meta-analysis and limiting the comparability of performance results. As a formal risk-of-bias evaluation was not performed, the effect of study quality on reported outcomes may remain a confounding factor and should be interpreted cautiously.

Results

All search results were exported and managed using EndNote reference management software (Clarivate Plc). The number of records identified from each source and the selection process are reported in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram (Figure 1).

The selection flowchart summarizing the included studies is depicted in Figure 1. A total of 67 studies met the eligibility criteria and were included in this scoping review. Details are provided in Multimedia Appendix 3. Of these, 49 were original research articles and 18 were reviews. Tables 2 and 3 maps the 49 original studies across study design, technology modality, symptom focus, and study maturity, providing a visual gap map of the current primary evidence. Insights from the 18 reviews are integrated throughout the thematic sections below to contextualize the findings and highlight persistent challenges.

XSL•FO
RenderX

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) flow diagram of the study selection process for a scoping review of multimodal intelligent monitoring systems in Parkinson disease (2019-2024). PD: Parkinson disease; RCT: randomized controlled trial.

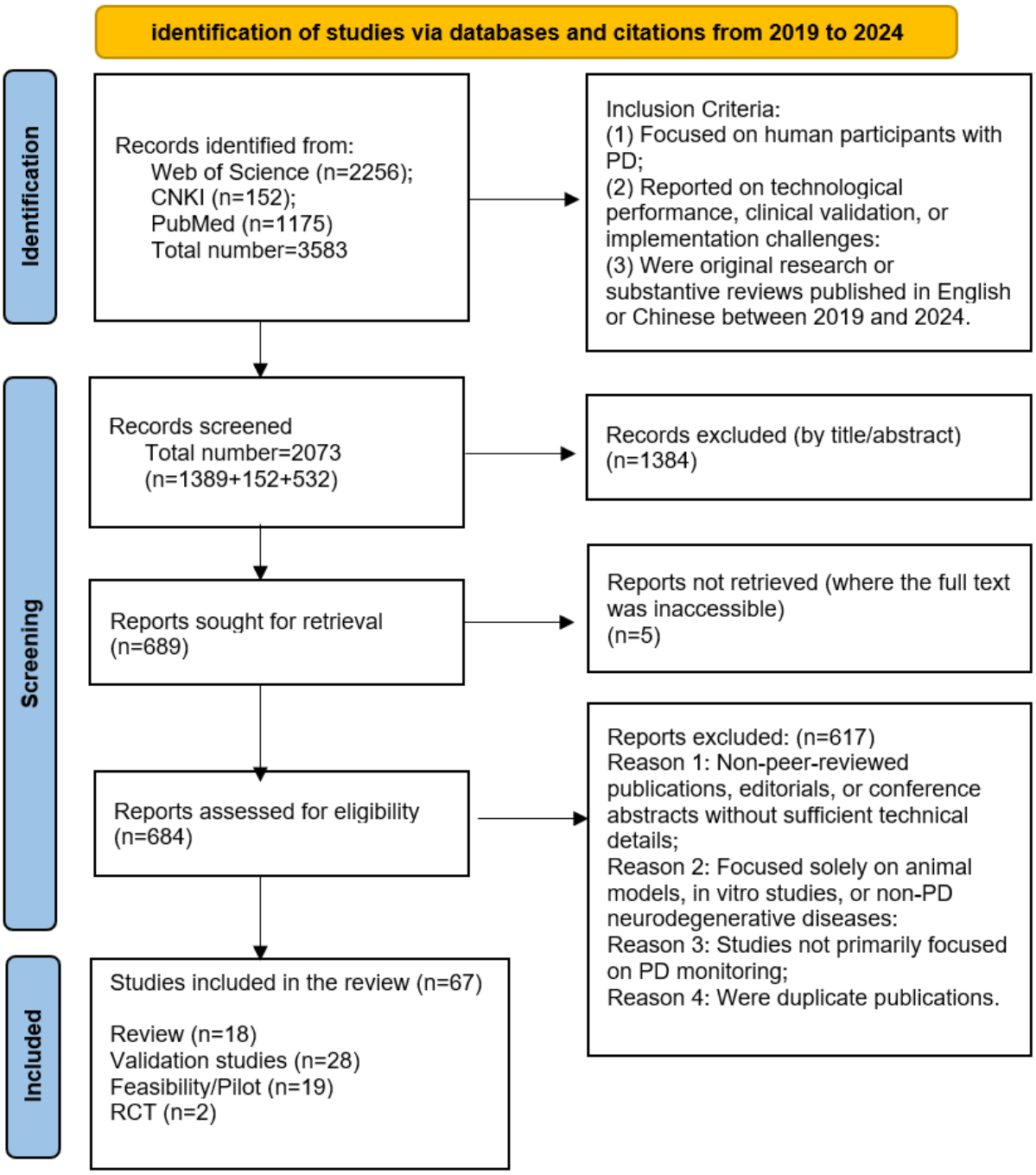

XSL•FO
RenderX

Table 2. Characteristics of included original studies: study type, technological modalities, and symptom domains.

References	Study type				Technology modality					Symptom domain				
	Technical validation	Clinical validation	Feasibility/pilot	RCT[a]	Wearable-Inertial measurement unit	Wearable-physiological	Computer vision	Speech analysis	Multimodal fusion	Motor	Sleep	Autonomic	Neuropsychiatric	Speech
Joshi et al [9]		✓			✓					✓				
Antonini et al [10]		✓			✓					✓				
Cabo-Lopez et al [11]			✓	✓	✓					✓				
Isaacson et al [12]			✓	✓	✓					✓				
Farid et al [13]	✓	✓			✓				✓	✓				
Botros et al [14]	✓		✓		✓				✓	✓				
Morinan et al [15]	✓	✓					✓			✓				
Chien et al [16]	✓	✓					✓			✓				
Yang et al [17]	✓	✓			✓				✓	✓	✓	✓		
Sigcha et al [18]	✓	✓	✓		✓					✓				
San-Segundo et al [19]	✓	✓	✓		✓					✓				
Borzì et al [20]	✓				✓					✓				
Amato et al [21]	✓							✓						✓
Tsakanikas et al [22]	✓				✓					✓				
Ko et al [23]	✓				✓	✓					✓			
Skibińska and Hosek [24]	✓						✓	✓	✓	✓				✓
Moore et al [25]	✓		✓		✓		✓		✓	✓				
Slemenšek et al [26]	✓				✓	✓			✓	✓				
Lipsmeier et al [27]		✓			✓			✓	✓	✓			✓	✓
Shin et al [28]	✓	✓				✓		✓	✓					✓
Shen et al [29]	✓				✓					✓				
Madrid-Navarro et al [30]	✓		✓		✓	✓			✓		✓			

XSL•FO
RenderX

References	Study type				Technology modality					Symptom domain				
	Technical validation	Clinical validation	Feasibility/pilot	RCT[a]	Wearable-Inertial measurement unit	Wearable-physiological	Computer vision	Speech analysis	Multimodal fusion	Motor	Sleep	Autonomic	Neuropsychiatric	Speech
Wu and Cronin-Golomb [31]	✓		✓		✓				✓		✓		✓	
Mirelman et al [32]	✓		✓		✓				✓		✓		✓	
Schalkamp et al [33]		✓			✓	✓			✓		✓	✓	✓	
Gnarra et al [34]	✓		✓		✓	✓			✓	✓	✓			
Oz et al [35]	✓	✓				✓			✓		✓			
van Wamelen et al [36]		✓			✓					✓		✓	✓	
Raschellà et al [37]	✓	✓	✓		✓				✓		✓			
Obayashi et al [38]		✓				✓					✓			
Lavorgna et al [39]			✓	✓	✓	✓			✓	✓	✓		✓	
Suzuki et al [40]	✓	✓			✓	✓						✓		
Almeida et al [41]	✓							✓						✓
Carro´n et al [42]	✓		✓					✓						✓
Polverino et al [43]			✓			✓						✓		
Lim et al [44]	✓	✓					✓	✓	✓	✓				✓
Kyritsis et al [45]	✓	✓			✓					✓				
Silburn et al [46]	✓	✓	✓							✓				
Alves et al [47]		✓	✓							✓			✓	
Thankathuraipandian et al [48]	✓	✓	✓		✓	✓				✓			✓	
Hoffman et al [49]	✓				✓	✓				✓			✓	
Múrias Lopes et al [50]	✓	✓			✓					✓				
Ile an et al [51]	✓				✓					✓				
Wijers et al [52]		✓												
Tarousi et al [53]	✓				✓		✓			✓				

XSL•FO
RenderX

References	Study type				Technology modality					Symptom domain				
	Technical validation	Clinical validation	Feasibility/pilot	RCT[a]	Wearable-Inertial measurement unit	Wearable-physiological	Computer vision	Speech analysis	Multimodal fusion	Motor	Sleep	Autonomic	Neuropsychiatric	Speech
Katie and Nicky [54]	✓				✓				✓	✓				
Gatsios et al [55]			✓		✓					✓				
Fabbri et al [56]			✓						✓					
Song et al [57]	✓							✓						✓

[a]RCT: randomized controlled trial.

XSL•FO
RenderX

Table 3. Characteristics of included original studies: study maturity and sample size.

References	Study maturity					Sample size			
	Lab/controlled	Short-term home	Long-term home	Multicenter validation	Clinically available	Small (<30)	Medium (30-100)	Large (>100)	Multicenter (>200)
Joshi et al [9]				✓	✓		✓		
Antonini et al [10]				✓	✓		✓		
Cabo-Lopez et al [11]		✓			✓		✓		
Isaacson et al [12]		✓			✓	✓			
Farid et al [13]	✓				✓	✓			
Botros et al [14]		✓				✓			
Morinan et al [15]				✓					✓
Chien et al [16]	✓						✓		
Yang et al [17]			✓	✓					
Sigcha et al [18]	✓					✓			
San-Segundo et al [19]	✓	✓				✓			
Borzì et al [20]		✓					✓		
Amato et al [21]	✓						✓		
Tsakanikas et al [22]	✓					✓			
Ko et al [23]		✓					✓		
Skibińska and Hosek [24]	✓							✓	
Moore et al [25]	✓	✓				✓			
Slemenšek et al [26]	✓					✓			
Lipsmeier et al [27]		✓		✓					✓
Shin et al [28]	✓	✓		✓			✓		
Shen et al [29]		✓					✓		
Madrid-Navarro et al [30]	✓	✓					✓		
Wu and Cronin-Golomb [31]		✓				✓			
Mirelman et al [32]		✓				✓			
Schalkamp et al [33]			✓					✓	
Gnarra et al [34]		✓				✓			
Oz et al [35]	✓	✓					✓		
van Wamelen et al [36]		✓						✓	
Raschellà et al [37]	✓	✓				✓			
Obayashi et al [38]		✓						✓	
Lavorgna et al [39]			✓				✓		
Suzuki et al [40]	✓					✓			
Almeida et al [41]	✓						✓		
Carro´n et al [42]	✓	✓					✓		
Polverino et al [43]		✓				✓			
Lim et al [44]	✓							✓	
Kyritsis et al [45]	✓	✓	✓			✓			
Silburn et al [46]	✓				✓	✓			
Alves et al [47]	✓				✓	✓			
Thankathuraipandian et al [48]		✓					✓		

XSL•FO
RenderX

References	Study maturity					Sample size			
	Lab/controlled	Short-term home	Long-term home	Multicenter validation	Clinically available	Small (<30)	Medium (30-100)	Large (>100)	Multicenter (>200)
Hoffman et al [49]		✓				✓			
Múrias Lopes et al [50]	✓				✓			✓	
Ile an et al [51]	✓					✓			
Wijers et al [52]			✓		✓			✓	
Tarousi et al [53]		✓				✓			
Katie and Nicky [54]	✓								✓
Gatsios et al [55]		✓					✓		
Fabbri et al [56]									
Song et al [57]	✓							✓	

The following sections report various performance metrics (eg, accuracy, sensitivity, area under the curve [AUC], intraclass correlation coefficient [ICC], adherence rates) derived from the included studies. It is important to emphasize that these metrics are highly context-dependent. They reflect performance within the specific study designs, participant cohorts, validation protocols, and technological setups from which they originate. Direct numerical comparisons across different studies, technologies, or symptom domains are not appropriate and may be misleading. These figures are presented to illustrate the current state of evidence and technological potential within each domain, not to establish a definitive hierarchy of performance or clinical readiness.

Digital Evaluation of PD Motor Symptoms

Wearable Sensors

Wearable sensors, especially devices utilizing IMUs, have markedly improved the objectivity and continuity of home-based monitoring for PD motor symptoms through multisite deployment and functional integration (Multimedia Appendix 3). Wearable systems employ either distributed placement (eg, wrist, ankle, waist) or targeted designs, enabling concurrent monitoring of core motor features such as tremor, bradykinesia, and gait abnormalities, as well as quantitative tracking of motor fluctuations and dyskinesia [9-12]. Multisensor solutions—such as PDMonitor and Kinesia 360—enhance symptom detection sensitivity by simultaneously monitoring the limbs and trunk, while single-device platforms, such as Personal KinetiGraph (PKG) smartwatches and STAT-ON, prioritize efficient identification of specific symptom domains [9-11,13].

Recent innovations in sensor design extend beyond conventional motion signal detection. For instance, smart insoles such as FeetMe Monitor deliver multidimensional analysis of gait and plantar pressure by merging IMU and pressure sensor data [13]. However, clinical implementation of multisensor technologies is impeded by concerns regarding wearability and a tendency toward lower long-term adherence as the number of sensors increases. Small-scale studies report high adherence (13-16 hours/day) with 3 or fewer sensors [14], although this may not generalize beyond motivated populations. Increased sensor burden is linked to discomfort and practical obstacles (eg, difficulty dressing), elevating dropout rates in less controlled and longer-term deployments. Furthermore, older patients frequently report usability challenges, leading to lower acceptance of complex multisensor systems [58].

Beyond initial device costs, the sustainability of mobile PD monitoring technologies poses an economic challenge. A review by the International Parkinson and Movement Disorder Society Task Force on Technology emphasizes the current lack of robust business models to incentivize health care payer reimbursement, coupled with a dearth of cost-effectiveness analyses [59]. Hidden costs such as maintenance, data subscriptions, and clinician interpretation time compound the economic burden and create significant barriers for resource-limited populations and health care systems. This highlights the urgent need for simplified, affordable monitoring solutions.

The clinical translation of emerging sensor modalities (eg, sweat analyzers, smart insoles) is hampered by data standardization challenges: (1) the absence of cross-manufacturer calibration standards impedes direct comparison of metrics such as gait symmetry or plantar pressure across devices (eg, FeetMe vs pressure mats); (2) temporal alignment difficulties in fusing heterogeneous sensor streams (such as synchronizing IMU and biochemical data) complicate reliable multimodal analysis and commonly require sophisticated hardware or software solutions [58].

These advances improve symptom quantification and treatment planning. Bio-integrated remote monitoring platforms with flexible, stretchable electronics are particularly promising for overcoming wearability constraints. By forming comfortable, skin-conformal interfaces, these next-generation systems minimize motion artifacts and enhance patient acceptance, representing a crucial advance for sustained long-term monitoring [60].

The diverse landscape of wearable technologies, as summarized in Table S1 in Multimedia Appendix 3, demonstrates a common pursuit of high objective performance (eg, accuracy and specificity often >0.85 for PDMonitor). However, direct comparison of these metrics across devices is challenging due

XSL•FO
RenderX

to pronounced methodological heterogeneity. Studies employed different gold standards (eg, UPDRS items, Abnormal Involuntary Movement Scale, expert diary), patient cohorts (varying disease stages), and outcome definitions (eg, thresholds for symptom detection). For instance, although PKG data led to treatment changes in most visits (71/85, 84%), its symptom detection profile (bradykinesia, 15/30, 50%, and 10/30, 33%, dyskinesia) reflects its specific algorithmic focus and validation context, not necessarily an absolute performance ranking. This heterogeneity indicates that high accuracy figures represent proof of efficacy within specific study frameworks, highlighting potential rather than universal clinical readiness. Key translational gaps, consistent across devices, include the need for standardized validation protocols and real-world cost-effectiveness analyses.

Computer Vision Technology

The adoption of computer vision technology combined with algorithms has introduced a powerful approach for remote, quantitative assessment of PD motor symptoms. Applications encompass diverse domains, from analyzing respiratory patterns to evaluating total body movements. Deep learning and pose estimation algorithms (eg, MediaPipe, OpenPose) enable automated measurement of indicators such as finger-tapping amplitude, hand speed, and limb agility via standard cameras or depth sensors. Notable achievements include multicenter validations of OpenPose, which, under controlled assessment protocols, obtained over 80% binary classification accuracy for leg movements and moderate concordance with clinician ratings (ICC 0.74) [15]. ICC over 0.7, while encouraging for research, falls short of the standard required (ie, >0.9) for clinical replacement, limiting these tools to adjunctive roles. Similarly, analyses indicate that computational metrics can be statistically correlated with Movement Disorder Society-UPDRS (MDS-UPDRS) scores (eg, $\rho=0.361$, $P<0.1$) [16], yet this magnitude accounts for only about 13% of clinical variance, suggesting that current computer vision approaches have yet to fully capture the spectrum of symptom severity as evaluated clinically. In parallel, novel approaches such as Massachusetts Institute of Technology's respiratory analysis system leverage long short-term memory models to interpret nocturnal breathing for disease severity prediction, achieving an AUC of 0.89 in their validation cohort, indicating strong discriminative ability in a research context. However, as with all laboratory-derived metrics, its performance in unselected real-world populations remains to be validated [17]. These advances promote objective, contactless evaluation and facilitate home-based monitoring and remote care.

Nonetheless, the deployment of computer vision in uncontrolled home environments faces robustness limitations. The accuracy of skeletal tracking algorithms is sensitive to nonoptimal conditions: subpar lighting, occlusion, and camera angles can diminish performance. Comprehensive reviews confirm that both camera parameters and environmental factors critically affect pose estimation accuracy [61] and that current markerless systems have strict requirements for acquisition environments [62]. Enhancing robustness for home use requires strategies such as data augmentation (simulating varying lighting and backgrounds), expanding training datasets to include multiple scenarios, and employing low-light imaging hardware and algorithms.

Additionally, the use of video-based assessments raises explicit patient privacy concerns. Video data—including facial and bodily motion—constitute biometric identifiers and are protected health information under the Health Insurance Portability and Accountability Act (HIPAA). Compliant systems must employ comprehensive safeguards: technical measures (encryption, anonymization), administrative protocols (business associate agreements and strict access controls), and physical security (secured acquisition and storage devices). For example, deidentifying video data before analysis aligns with HIPAA's Privacy Rule and is foundational to privacy-preserving clinical practice [63].

The current reliance on standardized laboratory tasks, while facilitating initial validation, restricts ecological validity and scalability. A future priority is capturing spontaneous, real-world motor symptoms in unstructured daily settings, as this represents a critical next step for computer vision to achieve continuous and clinically meaningful PD assessment.

A comparative view of computer vision systems (see Table S2 in Multimedia Appendix 3) reveals a trend toward robust performance in controlled settings (eg, AUCs of 0.85-0.91 for the MIT respiratory system). However, a critical appraisal must consider the validation context. High ICC values (eg, 0.74 for KELVIN) and correlations with UPDRS, although statistically significant, account for only a limited portion of clinical variance and originate from protocols using standardized, scripted movements. This indicates a current limitation in ecological validity—the ability to assess spontaneous symptoms in naturalistic environments. The performance decline observed in some applications when moving from controlled to uncontrolled settings further illustrates this "bench-to-bedside" gap. For example, one study reported that voice analysis accuracy decreased from 92% in laboratory conditions to 71% in home environments—a 21-percentage-point decline that highlights the vulnerability of laboratory-trained models to real-world environmental variability (background noise, inconsistent recording conditions). This finding underscores that high accuracy in controlled studies does not guarantee clinical utility in home settings. Therefore, the metrics in Table S2 in Multimedia Appendix 3 primarily serve as benchmarks for technical validation under optimal conditions. The next developmental leap requires demonstrating that these systems can maintain accuracy, robustness to environmental variables (eg, lighting, occlusion), and user-friendliness during continuous, unstructured home monitoring, which remains a significant translational challenge.

Intelligent Recognition of PD Nonmotor Symptoms

Nonmotor symptoms, including sleep disturbances, autonomic dysfunction, speech impairment, mood disorders, and sensory abnormalities, affect the majority of patients with PD and often precede the onset of classical motor symptoms. These manifestations substantially impair quality of life but are frequently underrecognized in routine clinical assessments due to their subjective nature and episodic evaluation. Intelligent identification technologies, therefore, offer particular value by

XSL•FO
RenderX

enabling continuous, objective, and ecologically valid monitoring of nonmotor symptom fluctuations.

In this section, studies are categorized according to sensing modalities, including single-modality and multimodal approaches, without implying uniform levels of data fusion or integration.

The cited accuracy and AUC values, while promising, are predominantly derived from controlled pilot studies or validation cohorts with specific inclusion criteria. Their generalizability to broader, unselected PD populations and home environments requires further confirmation.

Intelligent Recognition of Sleep Disturbances

Sleep disturbances are widely prevalent in PD. Monitoring technologies range from single-motion recorders to multimodal systems. Flexible electrode arrays combining electroencephalogram, electrooculogram, and electromyography signals have achieved strong agreement with the gold-standard laboratory video polysomnography, demonstrating a sensitivity of 76.8% for detecting muscle atonia during rapid eye movement sleep [35]. Wrist-worn actigraphy devices, using triaxial accelerometry, have delivered up to 92.9% screening accuracy for rapid eye movement sleep behavior disorder in laboratory settings; during 2-week home monitoring, models achieved 100% accuracy within specific PD subgroups and 94.4% accuracy among controls with insomnia [37]. However, the generalizability of these findings requires substantiation through larger, multicenter cohorts.

The Kronowise 3.0 device, integrating wrist temperature, acceleration, and environmental light exposure, reveals no significant discrepancies in core sleep parameters when compared with video polysomnography [30]. Multimodal signal technologies enhance both accuracy and convenience, particularly in home environments, thereby supporting early detection and screening of sleep disturbances in PD.

Environmental factors also significantly influence sleep in PD, as verified by interventional studies. Insufficient daylight exposure and excessive nocturnal light have a direct negative impact on sleep efficiency among patients [38]. Concurrently, a decreased frequency of nighttime bed-turning—confirmed by waist-worn accelerometers—correlates with worsening motor symptoms [64]. Integrated smartphone and wearable trackers have identified a bidirectional causal link between poor sleep quality and subsequent daytime anxiety levels [31], and sleep headbands demonstrate that longer supine sleep periods correspond with heightened dyskinesia severity [34]. Collectively, these technologies provide valuable insights for clinical management.

Despite these advancements, monitoring performance often declines when conducted in real-life home environments due to environmental noise, variability in sensor placement, and inconsistent user compliance. This indicates that findings from controlled settings may not readily generalize.

In addition, systems based on wearable devices have limited sensitivity in detecting lighter sleep stages, particularly transitional states such as N1 sleep. Most reported performance indicators are derived from pilot groups with specific inclusion criteria and should therefore be interpreted with caution. Overall, the existing evidence supports the feasibility of intelligent sleep monitoring in the management of PD, but its reliability and generalizability in long-term, free-living environments still require further confirmation.

Intelligent Recognition of Autonomic Dysfunction

Autonomic nervous system dysfunction in PD affects cardiovascular, gastrointestinal, and thermoregulatory domains. Intelligent assessment technologies directly capture physiological signals (Table S4 in Multimedia Appendix 3).

In cardiovascular monitoring, chest straps (eg, POLAR H10) and wristwatches (eg, POLAR V800) facilitate 24-hour heart rate interval recording. Heart rate variability metrics such as SD of normal-to-normal intervals and coefficient of variation of R-R intervals demonstrate high diagnostic utility (AUC 0.90) and can be detected before overt clinical symptoms [40]. Wireless blood pressure monitors offer simultaneous measurement of blood pressure, heart rate, oxygen saturation, and body temperature, autonomously detecting and signaling episodes of orthostatic hypotension with intelligent alerts and support for teleconsultations [43].

In auditory symptom detection, professional and smartphone microphones have demonstrated impressive accuracy rates of 94.55% and 92.94%, respectively, for PD voice analysis [41]. Mobile voice analytics systems also show efficacy in supporting symptom identification in controlled settings [42].

Multimodal intelligence is key for early detection. Combined biometric analysis using smartphone cameras and voice signals from laboratory-controlled samples allows joint assessment of facial expression and voice, aiding early-stage PD diagnosis (AUC 0.85); however, broader clinical validation is needed. Smartwatches can monitor swallowing dysfunction by detecting delayed hand-to-mouth movements [45], while PKG recorders link bradykinesia to constipation, providing quantifiable gastrointestinal dysfunction metrics [36].

Although wearable systems can correlate sleep activity with autonomic function scores, personalized prediction remains challenging [33]. Interpretation of metrics such as heart rate variability is complicated by inter- and intraindividual variability, particularly during longitudinal monitoring. Addressing these challenges requires the establishment of personalized reference baselines and advanced analytical techniques (eg, dynamic baseline models, covariate adjustments). Improving accuracy will require multimodal devices capable of integrating photoplethysmography, electrodermal activity, and movement data.

Most nonmotor symptom monitoring studies remain at the pilot or experimental validation stage, with limited evidence from long-term deployment in routine clinical practice. Importantly, reported performance metrics vary considerably across studies due to differences in sensor types, monitoring duration, sample size, and validation settings; therefore, direct quantitative comparisons between studies are not appropriate. This heterogeneity constrains the generalizability of current findings

XSL•FO
RenderX

and should be carefully considered when interpreting reported performance.

Overall, intelligent identification of nonmotor symptoms in PD remains an evolving field. Although existing studies demonstrate technical feasibility across multiple symptom domains, most systems have been validated in small-scale or short-term settings. Future research should prioritize longitudinal monitoring, multimodal integration, and clinically interpretable outputs to enhance translational value.

Clinical Decision-Making Limitations

Nonmotor monitoring demonstrates technical feasibility but limited clinical utility. Conceptually, sleep data could guide clinical decisions: rapid eye movement sleep behavior disorder detection might prompt melatonin or dopaminergic timing adjustments [35,37], sleep efficiency metrics could inform light exposure therapy [38], and reduced mobility might trigger fall prevention counseling [64]. Similarly, autonomic data could guide cardiovascular risk management [40] or orthostatic hypotension interventions [43].

However, these pathways remain largely theoretical. No study has demonstrated that acting on algorithm-detected sleep or autonomic abnormalities improves patient outcomes, reduces hospitalizations, or enables superior medication optimization compared with standard care. Key barriers include the inability to separate mixed symptom etiologies (eg, sleep fragmentation due to hypokinesia versus depression), the lack of validated thresholds linking specific values to specific actions, and the absence of prospective trials with hard clinical end points. At present, nonmotor monitoring primarily documents physiological patterns without supporting evidence for evidence-based therapeutic decisions.

These clinical limitations reflect deeper structural barriers. Nonmotor monitoring lags behind motor technologies in regulatory clearance and outcome trials [9,10,12]. Four structural barriers help explain this gap.

- Signal: motor symptoms produce direct measurements, whereas nonmotor proxies such as heart rate variability lack specificity.
- Validation: motor assessment relies on the reliable MDS-UPDRS part III, whereas nonmotor gold standards are often expensive or episodic [24].
- Environment: motor patterns are relatively stable, whereas nonmotor measures may degrade in home environments [24,33].
- Actionability: motor symptoms trigger clearer clinical decisions [9,12], whereas nonmotor data often have multiple causes without validated thresholds.

These fundamental barriers require new sensors, personalized baselines, and outcome trials—not just improved algorithms.

Interpreting the Metrics

The accuracy, AUC, and ICC values reported in this section represent technical validation results rather than evidence of clinical efficacy. Three limitations apply: studies used different methods and populations, preventing direct comparison; performance often declines in real-world use [35,41]; and no study has demonstrated that these measures improve patient outcomes. These figures illustrate technological potential but do not support conclusions about clinical readiness or comparative effectiveness.

Multimodal Data Analysis Algorithms

The high-performance metrics reported for algorithmic approaches (eg, accuracy, sensitivity) must be interpreted within their methodological context. These results are typically achieved using curated datasets under constrained evaluation protocols, which may not fully represent the heterogeneity and noise encountered in continuous, real-world monitoring.

Single-Modality Feature Extraction

Single-modality approaches use one type of data or sensor to build models (Table S5 in Multimedia Appendix 3). Convolutional neural networks (CNNs), frequently combined with other algorithms, are prominent for extracting spatial features from movement data in PD [18]. San-Segundo et al [19] reported laboratory-based use of CNNs for tremor percentage estimation, achieving error rates below 5%. While this indicates strong performance under controlled conditions, laboratory-optimized algorithms often experience accuracy declines when deployed in uncontrolled home environments due to signal variability. Additionally, lightweight models tailored for wearable sensors are advancing. Borzì et al [20] combined simple thresholding with CNNs for home-based freezing-of-gait detection, achieving 96% sensitivity. However, sensitivity alone does not capture the full picture; specificity and real-time performance in varied daily contexts are equally critical for clinical utility.

Beyond classic motion sensors, modalities including eye tracking, plantar pressure, and voice analysis (eg, studies by Amato, Chudzik, and Tsakanikas [21,22,65]) are broadening the scope of single-modality PD feature extraction.

Despite their utility, single-modal techniques often yield lower accuracy in real-world, nonlaboratory conditions due to the variability of daily activities and the heterogeneity of PD symptoms. As demonstrated by Sigcha et al [18], consumer smartwatches and deep learning improve early-stage PD monitoring; however, additional validation is essential for advanced-stage disease with more pronounced symptoms. Error propagation remains a concern (eg, rest-period misclassification), requiring more parameters and increasing model complexity. Overall, single-modal approaches have limited predictive validity for the total UPDRS score, as they mainly quantify motor function [66]. For nonmotor symptoms such as sleep disturbances, precision remains to be optimized [23].

Cross-Modality Fusion Strategies

Multimodal intelligent monitoring integrates multiple heterogeneous data modes to achieve a comprehensive, continuous, and accurate assessment of PD. Multimodal integration relies on hierarchical data fusion, and common strategies mainly include feature-level fusion and decision-level fusion (see Table S6 in Multimedia Appendix 3).

XSL•FO
RenderX

In feature-level fusion, the research constructs a unified high-dimensional feature representation by integrating the original or intermediate features from different modes. For example, by simultaneously analyzing the patient's voice signal and facial video features and using machine learning algorithms for modeling, the PD detection accuracy can reach as high as 83% [24]. In addition, to more comprehensively assess fall risk in a free-living environment, gait data and video eye-tracking data are combined. By using the fine-tuned YOLOv8 target detection algorithm to identify objects (such as obstacles and stairs) and potential hazards in the environment, and combining this with an overlap detection algorithm to analyze the intersection of the patient's fixation point, walking path, and hazards, the approach can provide key environmental context information for gait fluctuations and enhance the perception of fall risk [25].

At the same time, techniques such as multigranularity computation, rough set theory, and composite algorithms strengthen robustness. For instance, one study reported 98.9% accuracy for gait activity recognition using a CNN-recurrent neural network ensemble [67]. Although this suggests excellent performance on their specific dataset, such high values often reflect optimization on curated, task-specific data and may not generalize to the heterogeneity of free-living conditions.

In decision-level fusion, models first classify or regress data from each modality independently and then integrate the outputs from individual models. For example, constructing a digital health technology composite score by combining sensor features from multiple active and passive tests can provide a more robust assessment of overall PD severity. In addition, machine learning models have been used to predict MDS-UPDRS total scores or distinguish "on/off" states by fusing multisource features, enabling multidimensional information integration and final decision-making. These methods demonstrate good reliability and validity in early-stage PD populations and provide a digital basis for tracking disease progression [27].

Cross-modal fusion faces many technical challenges, especially feature synchronization. Consistency across data streams is vital to suppress false classifications. Diverse sampling rates and structures necessitate complex temporal alignment. In addition, domain adaptation is indispensable to counteract device and environment biases, and data augmentation strategies help offset small sample limitations (eg, Shin et al's [28] variable window shifting). These methods collectively help maintain stable performance across limited and heterogeneous datasets.

Federated Learning Applications

Federated learning—a distributed model training approach without raw data sharing—offers promise for cross-center collaboration in PD monitoring by enhancing privacy and scalability (see Table S7 in Multimedia Appendix 3). The study emphasizes that clinical data for PD encompass multimodal information on both motor and nonmotor features, such as images and wearable sensor data, providing abundant resources for federated learning applications [68]. The Diversity-Aware Activity Recognition framework, for example, clusters users by gait, activity, and social factors using K-means, enabling personalized model initialization and feature extraction [29]. This approach improves recognition of "cold-start" new users by 5.5%, and diversity-aware models demonstrate strong generalization across multiple activity tasks [29].

Key challenges include significant data distribution differences across users or devices, which affect the model's generalization ability. The introduction of noise in differential privacy mechanisms may reduce data utility; therefore, it is necessary to further optimize the privacy budget and noise mechanisms to balance privacy and utility [29]. The deployment of federated learning frameworks on mobile devices is also limited by computing resources and energy consumption [29]. In addition, medical data are mostly unstructured and lack unified annotation, which affects the quality of model training, and the design and implementation of the entire system must comply with data privacy regulations such as the General Data Protection Regulation and HIPAA [68].

Future directions include the use of federated learning to promote multimodal intelligent monitoring of PD. First, promote multimodal feature fusion by integrating wearable sensors, voice, and image data [69]. Second, develop personalized and adaptive models and use meta-learning, transfer learning, and other technologies to help models better adapt to heterogeneous data [29]. Third, promote lightweight and edge-computing deployment of models to adapt to the resource constraints of mobile devices and enable real-time monitoring and feedback [29]. Fourth, carry out cross-center clinical verification and evaluate the effectiveness and safety of the federated learning system in real clinical scenarios through multiagency collaboration [68]. Continued advancements in multimodal symptom monitoring will be achieved by optimizing feature extraction and fusion, combined with deeper integration of federated learning for secure, efficient data sharing and model training. This convergence will drive systems toward greater intelligence, personalization, and clinical applicability, underpinning precision diagnosis and management of PD.

From Technology to Translation: Challenges and Integrated Strategies for PD Remote Monitoring Platforms

Opportunities

The swift expansion of telemedicine has created ample opportunities for deploying intelligent monitoring technology in PD [43,70,71]. Since the advent of COVID-19, remote neurology consultations have increased by 17%. Notably, 98% of surveyed deep brain stimulation patients express a strong interest in remote management [46], representing a major catalyst for remote monitoring platforms. Accompanying these trends, acceptance rates for wearable and environmental sensors reach 87% and 100%, respectively [14], setting a robust social foundation for clinical translation.

The technical foundations have simultaneously matured: digital monitoring of PD symptoms demonstrates notable progress (PDMonitor tremor detection reports 99% concordance with MDS-UPDRS-III [10]), and cross-modal algorithmic optimization elevates fall risk prediction to 83% accuracy [13]. Together, these improvements unite hardware and software for reliable remote monitoring platforms.

XSL•FO
RenderX

Transformational Results

Remote monitoring platforms now span the entire continuum of PD management. Health care interaction–optimized platforms facilitate secure, efficient communication for individualized care (eg, MoveONParkinson achieves a System Usability Scale score of 79.50, 95% CI 73.70-85.30 [47]; RM App attains 85.5%, 172/201, adherence among deep brain stimulation patients [48]). Symptom-oriented platforms provide real-time physical symptom tracking and detailed assessment (QDG-Care and iHandU focus on tremor quantification [49,50], and PDxOne on gait analysis [51]). Comprehensive platforms, such as My Parkinsoncoach, integrate wearables and smartphones to collect data, deliver feedback, reduce outpatient visits by 29%, and enhance patient empowerment; solutions such as uMotif, Fox Wearable Companion, and PD Dr optimize medication tracking [72]. Nonetheless, integrated remote monitoring platforms encompassing both motor and nonmotor symptoms are still evolving—MoveONParkinson, for example, is limited to conversational interactions, while PDxOne focuses solely on gait. Such segmentation hinders holistic disease management and longitudinal care decisions.

Meanwhile, innovative rehabilitation solutions are broadening the implementation ecosystem for PD. Platforms such as “Shou Pa” incorporate Tai Chi to boost adherence, and initiatives such as ACTIVA, i-PROGNOSIS, and Rhythm Workers use virtual games for intensive motor training [53], illustrating a growing diversity of technological approaches.

The Care-Platform Transformation in PD (CPT-PD) Framework

While these platforms demonstrate technical excellence in their respective domains—including motor assessment, decision support, and treatment optimization—they are primarily focused on single functions or specific points in the care pathway. For example, SMaRT-PD provides a rule-based clinical decision support system that integrates multimodal data (wearable sensors, patient-reported outcomes, and care records) to generate personalized, symptom-specific management recommendations [54]. QDG-Care delivers validated quantitative metrics of motor symptoms through a fully integrated connected platform, with components spanning a dedicated digitography device, patient app, cloud-based algorithm service, provider dashboard, and electronic health record integration [49].

Identifying recurring challenges and themes in existing platforms informs the development of the Care-Platform Transformation in PD (CPT-PD) framework, which synthesizes key ingredients for effective PD monitoring (Figure 2). This is a reference model that draws on retrospective lessons rather than offering prescriptive advice and is intended to support the design of holistic, integrated remote monitoring platforms (Textbox 1).

Figure 2. The Care-Platform Transformation in Parkinson Disease (CPT-PD) framework: a conceptual model integrating key components for the design of effective remote monitoring platforms. PD: Parkinson disease; SCT: social cognitive theory.

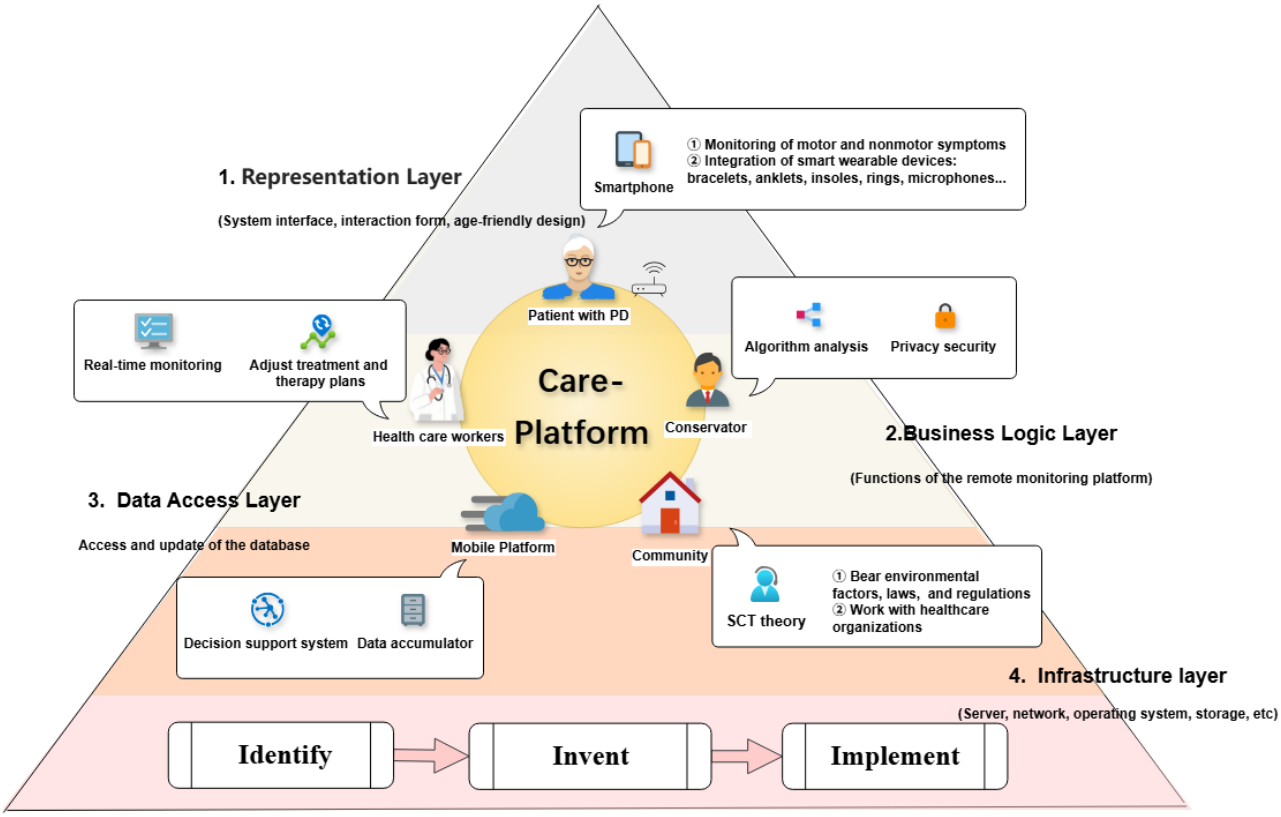

XSL•FO
RenderX

Textbox 1. Framework for the reference model.

1. Longitudinal, clinically driven pathway—identify, invent, implement

Initially, stakeholder needs—including monitoring efficiency, symptom assessment, and doctor-patient interaction—are clarified [46]. Clinical pilot validation and iterative algorithm development follow, culminating in the evaluation of health outcomes and cost-effectiveness for platform deployment.

2. Four-layer horizontal architecture

User-centric design in the presentation layer prioritizes accessibility (eg, large font, intuitive controls) for older patients with PD. The business logic layer emphasizes core symptom monitoring and facilitates clinician-patient exchanges. The data access layer upholds regulatory compliance and multimodal data security, leveraging frameworks such as Fast Healthcare Interoperability Resources and encrypted transmission [55]. The infrastructure layer provides essential network support for system reliability.

3. Collaborative Five-Element Model

Patients with PD submit multimodal data; health care workers use these data for decision-making; administrators ensure platform operation; integrated remote monitoring platforms support system functions; and the community provides supportive networks. Together, these elements interlink into a closed-loop "diagnosis-treatment-management" intelligent management cycle.

The synthesized framework (Textbox 1) offers a structured tool for analyzing platform strengths and weaknesses and a road map for overcoming fragmentation and enhancing clinical integration.

To validate the clinical applicability of this framework, we present the "My Parkinsoncoach" platform, which has successfully reduced outpatient burden in PD. This case example demonstrates how it organically integrates the longitudinal pathway, horizontal architecture, and collaborative ecosystem of the CPT-PD framework [52].

The platform's development clearly follows the longitudinal, clinically driven pathway: it began by identifying the unsustainable burden of frequent outpatient visits; invented an integrated solution comprising periodic electronic questionnaires, visual symptom-domain scoring, clinician-patient communication, and educational modules; and implemented the system through a prospective cohort study that verified a 29% reduction in outpatient visits and a 39% decrease in health care costs. This structured development process is a key reason for the platform's success.

We now explain the platform through the horizontal 4-layer architecture: its presentation layer provides role-specific, user-friendly interfaces for patients and clinicians; the business logic layer automatically converts questionnaire responses into trend scores across 15 symptom domains (eg, motor symptoms, activities of daily living) and implements a clinical response protocol whereby specialized nurses handle initial alerts and neurologists provide backup; the data access layer securely manages these multimodal patient-reported data; and the entire system operates on a stable infrastructure layer (network connectivity).

Within this architecture, the 5 key collaborative elements work in synergy to sustain stable platform operation: patients submit data via the mobile interface; specialized nurses and neurologists interpret the data and intervene based on prompts from the business logic layer; platform operators ensure system stability; the platform itself serves as the integrated functional carrier; and the broader community (eg, family caregivers)—although not highlighted in the original study—represents a natural extension for scaling the model into home-based care.

Applying this case to the CPT-PD framework demonstrates its ability to deconstruct the design logic and core components of a successful platform. At the same time, it reveals areas for framework refinement—notably, the integration of the "community" element remains underdeveloped. This insight points the way toward future optimization of both the "My Parkinsoncoach" platform and the CPT-PD framework itself.

Transformational Bottleneck

Despite progress, only a minority of monitoring devices have transitioned into remote monitoring platforms. This translational bottleneck fundamentally stems from a mismatch between reported performance and clinical readiness. While many studies demonstrate promising results in controlled pilot settings (Technology Readiness Level ~4-5), only a few progress to robust validation in heterogeneous, real-world clinical environments (Technology Readiness Level 7-8), creating a "readiness gap" that hinders broad adoption [73]. This gap is compounded by several persistent bottlenecks:

- Clinical research needs greater incorporation of patient- and community-centered social cognitive theory [47,56], increasing primary care collaboration to prioritize key patient needs.
- Technical fragmentation results in isolated symptom focus and insufficient holistic assessment, and the detachment of development from clinical practice undermines nonmotor symptom monitoring [74].
- Device and algorithm limitations stem from a lack of standardization—the PKG and STAT-ON offer inconsistent data structures, forcing manual curation by clinicians and increasing errors. Artificial intelligence (AI)–sensor integration faces data standardization and regulatory obstacles, as well as extensive validation requirements [75], requiring collaboration among neurologists, AI experts, and regulators.
- Low smartphone adoption among older adult patients necessitates cognitive load–sensitive, age-friendly design with voice user interfaces, simple user interfaces [57], and home assistive device compatibility [71,76-78]. This challenge is particularly evident in studies such as eSanjeevani, which reported low adherence to wearable devices among older patients with PD in India, with health care providers citing "low penetration rates and insufficient

XSL•FO
RenderX

convenience (especially among the elderly)" [78]. Notably, only 32.4% of compiled smart device studies used randomized controlled trials focusing on clinical experience, highlighting a research-to-practice gap [79]. Addressing these issues demands interdisciplinary breakthroughs in algorithms and a paradigm shift toward patient-centered, senior-accessible design.

Discussion

This review provides a critical evaluation of the latest advancements and persisting challenges in multimodal intelligent monitoring for PD. However, as highlighted in our terminology (Table 1), the field currently includes many multisensor systems with limited algorithmic fusion. Around the 3 core dimensions proposed in the "Introduction" section, the main findings of this study are as follows. First, in terms of the clinical performance of wearable sensors, the multisite deployment of IMUs and new sensing technologies has significantly improved the objectivity and continuity of motor symptom monitoring, but wearing comfort, long-term compliance, and data standardization remain the main obstacles to clinical translation. Second, at the level of algorithms and methods, single-mode feature extraction technologies perform well under laboratory conditions, but their accuracy declines in real-world environments. Cross-modal fusion strategies enable more comprehensive disease phenotype analysis by integrating multisource data. Distributed computing frameworks such as federated learning show potential for multicenter collaboration under the premise of ensuring data privacy. Third, in terms of the clinical translation of remote monitoring platforms, existing platforms cover multiple dimensions such as medical interaction optimization, symptom-oriented monitoring, and comprehensive management, but integrated platforms that truly achieve coordinated monitoring of motor and nonmotor symptoms are still in the early stages of development.

Unlike prior reviews focusing on isolated technologies, we adopt an integrative perspective encompassing the quantification of both motor and nonmotor symptoms, the development of analytical algorithms, and the evolution toward remote monitoring platforms. Evidence indicates that platforms integrating wearable sensors, computer vision, and biosignal analytics can substantially enhance objectivity and continuity in symptom tracking (see Tables S2-S5 in Multimedia Appendix 3). Advanced multimodal algorithms and distributed computing demonstrate potential strengths in secure, multicenter data sharing, privacy protection, and model generalizability (see Tables S6 and S7 in Multimedia Appendix 3), in addition to improving medication adherence [72]. The widespread adoption of remote monitoring platforms further validates the clinical value of a closed-loop "diagnosis-treatment-management" approach, optimizing health care resource use and reducing outpatient burden [52,72].

The heterogeneity documented across studies—in technologies, populations, protocols, and outcome measures—while reflecting a vibrant and innovative field, presents a primary challenge for synthesizing evidence and translating devices into practice. This variability precludes simple quantitative comparisons and complicates the assessment of a technology's true clinical readiness. To move beyond this stage, future research should prioritize standardization (eg, validation protocols against common clinical benchmarks), harmonization (of data formats and outcome measures), and pragmatic study designs that evaluate performance in intended-use settings with representative patient populations. Addressing these methodological challenges is as crucial as algorithmic innovation for realizing the potential of multimodal monitoring.

Despite these promising developments, several persistent obstacles impede translation into practice. A major limitation is that most existing solutions remain focused on narrow symptom domains, lacking coordinated quantification across motor and nonmotor features [33]. Moreover, algorithmic adaptability is constrained by heterogeneity in devices and data, underscoring the need for robust multicenter validation [68]. Finally, practical barriers to widespread adoption include platform complexity and low rates of smart device usage among older adult populations [58].

The heterogeneity recorded across studies in technology, population, protocols, and outcome measures—although reflecting the vigorous innovation in this field—also brings systemic challenges for evidence integration and clinical translation. To move beyond this stage, the field must address several key research gaps. Based on the comprehensive evidence in this review, we identified 5 key gaps that, if addressed, would accelerate progress:

First, and most fundamental, there is a lack of longitudinal real-world validation. The vast majority of studies are still limited to short-term, small-sample, laboratory or semicontrolled verification, which fails to capture the progressive nature of PD and lacks longitudinal follow-up data for months or even years in real home environments. As Espay et al [58] stated, this leads to an extremely limited understanding of the long-term compliance with technology, data stability, and its dynamic association with disease progression. Therefore, conducting longitudinal real-world validation has substantial clinical value and cost-effectiveness.

Second, at the system level, current intelligent monitoring generally remains fragmented, focusing on individual symptoms. The lack of integrated platforms quantifying interactions among motor fluctuations, sleep, mood, and autonomic function creates a bottleneck, limiting progression from symptom monitoring to holistic disease management. In view of the high variability of PD symptoms, it has been proposed to develop "multidomain, multisensor, intelligent technologies" to "identify individual disease fingerprints and achieve truly personalized treatment" [58].

Third, from sensor deployment to algorithm performance and clinical reference standards, studies lack unified normative frameworks and exhibit high methodological heterogeneity. The absence of standardized clinical outcome indicators makes it extremely difficult to compare the performance of different technologies or synthesize data across studies and also hinders regulatory approval of digital outcome indicators by agencies such as the Food and Drug Administration [33,58,73,75].

XSL•FO
RenderX

Fourth, although patient-centered design has been repeatedly emphasized, existing intelligent monitoring platforms still show deficiencies in user interface friendliness, cognitive load management, and aging adaptation when serving the older adult population with a high incidence of PD. This has directly led to low technology acceptance and poor long-term compliance in the target group [57,58,78].

At a deeper level, existing research has focused too much on improving prediction accuracy through feature- or decision-level fusion but has largely ignored the pathophysiological logic underlying multimodal data [58]. In fact, true multimodal intelligence should be able to explain the interactions among different modalities rather than simply pursuing higher classification accuracy.

Therefore, addressing these methodological and translational gaps is as critical as algorithmic innovation to realize the potential of multimodal monitoring. To overcome these challenges, future research should prioritize the development of an integrated, precise, and scalable intelligent management ecosystem for continuous PD monitoring. A phased approach is proposed to guide this effort.

In the short term, emphasis should be placed on the coevolution of hardware and algorithms. On the one hand, efforts are needed to develop more wearable, comfortable, and multimodal sensing hardware capable of long-term continuous monitoring, integrating diverse data streams such as sleep patterns, heart rate variability, and speech [13,30,35,40,42,45,56]. On the other hand, at the algorithmic level, capabilities in signal processing and multimodal information fusion must be strengthened, leveraging adaptive fusion mechanisms to enable deeper disease phenotyping [29,64,66,67,80]. Over the medium to long term, the establishment of a secure and compliant multicenter collaborative framework is essential. This will require the implementation of privacy-preserving computational techniques, such as federated learning, to aggregate knowledge from localized data processing securely, alongside embedded privacy-enhancing technologies to meet regulatory and ethical requirements [68,81]. Concurrently, in advancing toward international standards, it is critical to institute dynamic patient-informed consent mechanisms and proactive governance—not merely to fulfill legal obligations but to cultivate a trustworthy, patient-centered, and privacy-aware intelligent platform [68,81].

Importantly, insights drawn from the CPT-PD framework highlight that system design and deployment must thoroughly account for the needs of older users. Consequently, future platforms should integrate AI-driven clinical decision support with aging-adaptive design principles, employing user-friendly devices, stepwise low-burden interactions, and clear, timely feedback to minimize cognitive and operational demands. Such an approach is key to enhancing compliance and achieving truly patient-centric, progressive health management [78,82].

In summary, the innovation of this study lies in its departure from merely commenting on a single technology or symptom dimension. Instead, based on existing evidence, it pioneers a unified framework, CPT-PD, that integrates multimodal perception, edge intelligence, privacy computation, and aging-adaptive design. More importantly, by clearly identifying key research gaps in longitudinal effectiveness, symptom fragmentation, and lack of standardization, this review provides a prioritized road map. In stark contrast to existing research that typically focuses on the performance of specific algorithms or the validation of single devices, this study emphasizes the collaborative quantification of motor and nonmotor symptoms, the coevolution of hardware and algorithms, and the deep integration of technology and humanistic care. Consequently, the key contribution of this study to the field is that it provides a clear road map and theoretical foundation for PD management to transition from a fragmented, technology-centric “tool accumulation” phase to a systematic, patient-centered “intelligent ecosystem.” In practical terms, this means we are poised to overcome the bottlenecks of low technology conversion rates and poor patient acceptance. By constructing a truly accessible, reliable, and user-friendly monitoring system for older adults, we can shift from passive diagnosis and treatment to proactive, continuous home health management. This, in turn, will tangibly improve patients’ quality of life, reduce the overall health care burden, and ultimately advance the goals of precision medicine and scalable, patient-centered disease management.

Acknowledgments

We appreciate the contributions of all authors involved in this study. Their collective efforts in research design, data collection, analysis, and manuscript preparation were crucial to the successful completion of this work. The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT (Generative Artificial Intelligence Delegation Taxonomy; 2025) guidelines, the following task was delegated to GAI tools under full human supervision: The GAI tool used was DeepSeek (Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co., Ltd). Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data Availability

All data generated or analyzed during this study are included in this published article or its multimedia appendices.

XSL•FO
RenderX

Funding

The project was supported by the National Natural Youth Foundation (grant 82301436) and the China Postdoctoral Science Foundation (grant 2023M730597).

Authors' Contributions

Conceptualization: JT
Study design: JT
Supervision: HZ, CL, JL
Writing – original draft: JT, XD, CW, MY
Writing – review & editing: JT, HZ, CL, JL

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed search strategy and process.
[DOCX File , 21 KB - jmir_v28i1e89057_app1.docx]

Multimedia Appendix 2
Detailed search strategy and process.
[DOCX File , 36 KB - jmir_v28i1e89057_app2.docx]

Multimedia Appendix 3
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 203 KB - jmir_v28i1e89057_app3.pdf]

References

1. Tolosa E, Garrido A, Scholz SW, Poewe W. Challenges in the diagnosis of Parkinson's disease. The Lancet Neurology 2021 May;20(5):385-397. [doi: 10.1016/s1474-4422(21)00030-2]
2. Su D, Cui Y, He C, Yin P, Bai R, Zhu J, et al. Projections for prevalence of Parkinson's disease and its driving factors in 195 countries and territories to 2050: modelling study of Global Burden of Disease Study 2021. BMJ 2025 Mar 05;388:e080952 [FREE Full text] [doi: 10.1136/bmj-2024-080952] [Medline: 40044233]
3. WANG Gang, XU Gang, XIE Xinyi, WANG Hualong, HUANG Jingxuan, HUANG Zishuo, QIU Gonghang, DU Yunlan, DAI Ruolian, LU Zhongjiao, CHEN Gang, DONG Qing, ZHOU Hongyu, QIN Yan, LI Juan, WANG Ying, SHANG Huifang, CHEN Xianwen, CHEN Shengdi. China Parkinson's disease report, 2025. Journal of Neurology and Neurorehabilitation 2025;21(2):63-98 (forthcoming) [FREE Full text] [doi: 10.12022/jnnr.2025-0003]
4. Weintraub D, Aarsland D, Chaudhuri KR, Dobkin RD, Leentjens AF, Rodriguez-Violante M, et al. The neuropsychiatry of Parkinson's disease: advances and challenges. The Lancet Neurology 2022 Jan;21(1):89-102. [doi: 10.1016/s1474-4422(21)00330-6]
5. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
6. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
7. Aromataris EL, Porritt K, Pilla B, Jordan Z, editors. JBI Manual for Evidence Synthesis. JBI. 2024. URL: https://doi.org/10.46658/JBIMES-24-01 [accessed 2024-03-26]
8. Choong MK, Galgani F, Dunn AG, Tsafnat G. Automatic evidence retrieval for systematic reviews. J Med Internet Res 2014 Oct 01;16(10):e223 [FREE Full text] [doi: 10.2196/jmir.3369] [Medline: 25274020]
9. Joshi R, Bronstein JM, Keener A, Alcazar J, Yang DD, Joshi M, et al. PKG movement recording system use shows promise in routine clinical care of patients with Parkinson's disease. Front Neurol 2019 Oct 1;10:1027-1011 [FREE Full text] [doi: 10.3389/fneur.2019.01027] [Medline: 31632333]
10. Antonini A, Reichmann H, Gentile G, Garon M, Tedesco C, Frank A, et al. Toward objective monitoring of Parkinson's disease motor symptoms using a wearable device: wearability and performance evaluation of PDMonitor. Front Neurol 2023 May 16;14:1080752-1080720 [FREE Full text] [doi: 10.3389/fneur.2023.1080752] [Medline: 37260606]

XSL•FO
RenderX

11. Cabo-Lopez I, Puy-Nuñez A, Redondo-Rafales N, Teixeira Baltazar S, Calderón-Cruz B. Holter STAT-ON™ against other tools for detecting MF in advanced Parkinson's disease: an observational study. Front Neurol 2023 Aug 18;14:1249385-1249387 [FREE Full text] [doi: 10.3389/fneur.2023.1249385] [Medline: 37662044]
12. Isaacson SH, Boroojerdi B, Waln O, McGraw M, Kreitzman DL, Klos K, et al. Effect of using a wearable device on clinical decision-making and motor symptoms in patients with Parkinson's disease starting transdermal rotigotine patch: a pilot study. Parkinsonism Relat Disord 2019 Jul;64:132-137. [doi: 10.1016/j.parkreldis.2019.01.025] [Medline: 30948242]
13. Farid L, Jacobs D, Do Santos J, Simon O, Gracies J, Hutin E. FeetMe® Monitor-connected insoles are a valid and reliable alternative for the evaluation of gait speed after stroke. Top Stroke Rehabil 2021 Mar 13;28(2):127-134 [FREE Full text] [doi: 10.1080/10749357.2020.1792717] [Medline: 32654627]
14. Botros A, Schütz N, Camenzind M, Urwyler P, Bolliger D, Vanbellingen T, et al. Long-term home-monitoring sensor technology in patients with Parkinson's disease-acceptance and adherence. Sensors (Basel) 2019 Nov 26;19(23):5169 [FREE Full text] [doi: 10.3390/s19235169] [Medline: 31779108]
15. Morinan G, Dushin Y, Sarapata G, Rupprechter S, Peng Y, Girges C, et al. Computer vision quantification of whole-body Parkinsonian bradykinesia using a large multi-site population. NPJ Parkinsons Dis 2023 Jan 27;9(1):10-12 [FREE Full text] [doi: 10.1038/s41531-023-00454-8] [Medline: 36707523]
16. Chien C, Ouyang C, Chiu Y, Lin L, Yang R, Chang Y, et al. An automatic approach of video-based landmark detection and movement analysis for assessing symptoms of bradykinesia in Parkinson’s disease. J Med Biol Eng 2024 Jun 15;44(3):460-470. [doi: 10.1007/s40846-024-00876-6]
17. Yang Y, Yuan Y, Zhang G, Wang H, Chen Y, Liu Y, et al. Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. Nat Med 2022 Oct 22;28(10):2207-2215 [FREE Full text] [doi: 10.1038/s41591-022-01932-x] [Medline: 35995955]
18. Sigcha L, Pavón I, Costa N, Costa S, Gago M, Arezes P, et al. Automatic resting tremor assessment in Parkinson's disease using smartwatches and multitask convolutional neural networks. Sensors (Basel) 2021 Jan 04;21(1):1-29 [FREE Full text] [doi: 10.3390/s21010291] [Medline: 33406692]
19. San-Segundo R, Zhang A, Cebulla A, Panev S, Tabor G, Stebbins K, et al. Parkinson's disease tremor detection in the wild using wearable accelerometers. Sensors (Basel) 2020 Oct 14;20(20):1-23 [FREE Full text] [doi: 10.3390/s20205817] [Medline: 33066691]
20. Borzì L, Sigcha L, Olmo G. Context recognition algorithms for energy-efficient freezing-of-gait detection in Parkinson's disease. Sensors (Basel) 2023 Apr 30;23(9):4426 [FREE Full text] [doi: 10.3390/s23094426] [Medline: 37177629]
21. Amato F, Borzì L, Olmo G, Orozco-Arroyave JR. An algorithm for Parkinson's disease speech classification based on isolated words analysis. Health Inf Sci Syst 2021 Dec 30;9(1):32-15 [FREE Full text] [doi: 10.1007/s13755-021-00162-8] [Medline: 34422258]
22. Tsakanikas V, Ntanis A, Rigas G, Androutsos C, Boucharas D, Tachos N, et al. Evaluating gait impairment in Parkinson's disease from instrumented insole and IMU sensor data. Sensors (Basel) 2023 Apr 12;23(8):3902 [FREE Full text] [doi: 10.3390/s23083902] [Medline: 37112243]
23. Ko Y, Kuo P, Wang C, Chen Y, Chuang P, Li S, et al. Quantification analysis of sleep based on smartwatch sensors for Parkinson's disease. Biosensors (Basel) 2022 Jan 27;12(2):74 [FREE Full text] [doi: 10.3390/bios12020074] [Medline: 35200335]
24. Skibińska J, Hosek J. Computerized analysis of hypomimia and hypokinetic dysarthria for improved diagnosis of Parkinson's disease. Heliyon 2023 Nov;9(11):e21175 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e21175] [Medline: 37908703]
25. Moore J, Celik Y, Stuart S, McMeekin P, Walker R, Hetherington V, et al. Using video technology and AI within Parkinson's disease free-living fall risk assessment. Sensors (Basel) 2024 Jul 29;24(15):1-17 [FREE Full text] [doi: 10.3390/s24154914] [Medline: 39123961]
26. Slemenšek J, Geršak J, Bratina B, van Midden VM, Pirtošek Z, Šafarič R. Wearable online freezing of gait detection and cueing system. Bioengineering (Basel) 2024 Oct 20;11(10):11101048 [FREE Full text] [doi: 10.3390/bioengineering11101048] [Medline: 39451423]
27. Lipsmeier F, Taylor KI, Postuma RB, Volkova-Volkmar E, Kilchenmann T, Mollenhauer B, et al. Reliability and validity of the Roche PD mobile application for remote monitoring of early Parkinson's disease. Sci Rep 2022 Jul 15;12(1):12081 [FREE Full text] [doi: 10.1038/s41598-022-15874-4] [Medline: 35840753]
28. Shin B, Lee SH, Kwon K, Lee YJ, Crispe N, Ahn S, et al. Automatic clinical assessment of swallowing behavior and diagnosis of silent aspiration using wireless multimodal wearable electronics. Adv Sci (Weinh) 2024 Sep;11(34):e2404211. [doi: 10.1002/advs.202404211] [Medline: 38981027]
29. Shen Q, Feng H, Song R, Song D, Xu H. Federated meta-learning with attention for diversity-aware human activity recognition. Sensors (Basel) 2023 Jan 17;23(3):1 [FREE Full text] [doi: 10.3390/s23031083] [Medline: 36772123]
30. Madrid-Navarro CJ, Puertas Cuesta FJ, Escamilla-Sevilla F, Campos M, Ruiz Abellán F, Rol MA, et al. Validation of a device for the ambulatory monitoring of sleep patterns: a pilot study on Parkinson's disease. Front Neurol 2019 Apr 11;10:356-315 [FREE Full text] [doi: 10.3389/fneur.2019.00356] [Medline: 31031690]

XSL•FO
RenderX

31. Wu JQ, Cronin-Golomb A. Temporal associations between sleep and daytime functioning in Parkinson's disease: a smartphone-based ecological momentary assessment. Behav Sleep Med 2020 Jun 14;18(4):560-569. [doi: 10.1080/15402002.2019.1629445] [Medline: 31199163]
32. Mirelman A, Ben Or Frank M, Melamed M, Granovsky L, Nieuwboer A, Rochester L, et al. Detecting sensitive mobility features for Parkinson's disease stages via machine learning. Mov Disord 2021 Sep;36(9):2144-2155. [doi: 10.1002/mds.28631] [Medline: 33955603]
33. Schalkamp A, Harrison NA, Peall KJ, Sandor C. Digital outcome measures from smartwatch data relate to non-motor features of Parkinson's disease. NPJ Parkinsons Dis 2024 May 29;10(1):110 [FREE Full text] [doi: 10.1038/s41531-024-00719-w] [Medline: 38811633]
34. Gnarra O, Calvello C, Schirinzi T, Beozzo F, De Masi C, Spanetta M, et al. Exploring the association linking head position and sleep architecture to motor impairment in Parkinson's disease: an exploratory study. J Pers Med 2023 Nov 10;13(11):1-11 [FREE Full text] [doi: 10.3390/jpm13111591] [Medline: 38003906]
35. Oz S, Dagay A, Katzav S, Wasserman D, Tauman R, Gerston A, et al. Monitoring sleep stages with a soft electrode array: comparison against vPSG and home-based detection of REM sleep without atonia. J Sleep Res 2023 Oct 02;32(5):e13909-e13913. [doi: 10.1111/jsr.13909] [Medline: 37132065]
36. van Wamelen DJ, Hota S, Podlewska A, Leta V, Trivedi D, Rizos A, et al. Non-motor correlates of wrist-worn wearable sensor use in Parkinson's disease: an exploratory analysis. NPJ Parkinsons Dis 2019 Oct 02;5(1):22 [FREE Full text] [doi: 10.1038/s41531-019-0094-4] [Medline: 31602393]
37. Raschellà F, Scafa S, Puiatti A, Martin Moraud E, Ratti P. Actigraphy enables home screening of rapid eye movement behavior disorder in Parkinson's disease. Ann Neurol 2023 Feb 20;93(2):317-329. [doi: 10.1002/ana.26517] [Medline: 36193943]
38. Obayashi K, Saeki K, Tai Y, Yamagami Y, Esaki Y, Yoshikawa T, et al. Daily light exposure profiles and the association with objective sleep quality in patients with Parkinson's disease: the PHASE study. Sleep 2024 Aug 14;47(8):1-10. [doi: 10.1093/sleep/zsae036] [Medline: 38330229]
39. Torrado JC, Husebo BS, Allore HG, Erdal A, Fæø SE, Reithe H, et al. Digital phenotyping by wearable-driven artificial intelligence in older adults and people with Parkinson's disease: protocol of the mixed method, cyclic ActiveAgeing study. PLoS One 2022;17(10):e0275747 [FREE Full text] [doi: 10.1371/journal.pone.0275747] [Medline: 36240173]
40. Suzuki M, Nakamura T, Hirayama M, Ueda M, Hatanaka M, Harada Y, et al. Wearable sensor device-based detection of decreased heart rate variability in Parkinson's disease. J Neural Transm (Vienna) 2022 Oct 14;129(10):1299-1306. [doi: 10.1007/s00702-022-02528-y] [Medline: 35835890]
41. Almeida JS, Rebouças Filho PP, Carneiro T, Wei W, Damaševičius R, Maskeliūnas R, et al. Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques. Pattern Recognition Letters 2019 Jul;125:55-62. [doi: 10.1016/j.patrec.2019.04.005]
42. Carrón J, Campos-Roca Y, Madruga M, Pérez CJ. A mobile-assisted voice condition analysis system for Parkinson's disease: assessment of usability conditions. Biomed Eng Online 2021 Nov 21;20(1):114 [FREE Full text] [doi: 10.1186/s12938-021-00951-y] [Medline: 34802448]
43. Polverino P, Ajčević M, Catalan M, Bertolotti C, Furlanis G, Marsich A, et al. Comprehensive telemedicine solution for remote monitoring of Parkinson's disease patients with orthostatic hypotension during COVID-19 pandemic. Neurol Sci 2022 Jun;43(6):3479-3487 [FREE Full text] [doi: 10.1007/s10072-022-05972-6] [Medline: 35301614]
44. Lim WS, Chiu SI, Peng PL, Jang JSR, Lee SH, Lin CH, et al. A cross-language speech model for detection of Parkinson's disease. J Neural Transm (Vienna) 2025 Apr;132(4):579-590. [doi: 10.1007/s00702-024-02874-z] [Medline: 39739129]
45. Kyritsis K, Fagerberg P, Ioakimidis I, Chaudhuri KR, Reichmann H, Klingelhoefer L, et al. Assessment of real life eating difficulties in Parkinson's disease patients by measuring plate to mouth movement elongation with inertial sensors. Sci Rep 2021 Jan 15;11(1):1632 [FREE Full text] [doi: 10.1038/s41598-020-80394-y] [Medline: 33452324]
46. Silburn P, DeBates S, Tomlinson T, Schwark J, Creek G, Patel H, et al. Rapid development of an integrated remote programming platform for neuromodulation systems through the biodesign process. Sci Rep 2022 Feb 10;12(1):2269 [FREE Full text] [doi: 10.1038/s41598-022-06098-7] [Medline: 35145143]
47. Alves B, Mota PR, Sineiro D, Carmo R, Santos P, Macedo P, et al. MoveONParkinson: developing a personalized motivational solution for Parkinson's disease management. Front Public Health 2024 Aug 19;12:1420171 [FREE Full text] [doi: 10.3389/fpubh.2024.1420171] [Medline: 39224558]
48. Thankathuraipandian S, Greenleaf W, Kyani A, Tomlinson T, Balasingh B, Ross E, et al. Development of a remote therapeutic monitoring platform: applications for movement disorders. Sci Rep 2024 Dec 01;14(1):29837 [FREE Full text] [doi: 10.1038/s41598-024-80567-z] [Medline: 39617767]
49. Hoffman SL, Schmiedmayer P, Gala AS, Wilkins KB, Parisi L, Karjagi S, et al. Comprehensive real time remote monitoring for Parkinson's disease using quantitative DigitoGraphy. NPJ Parkinsons Dis 2024 Jul 27;10(1):137 [FREE Full text] [doi: 10.1038/s41531-024-00751-w] [Medline: 39068150]
50. Lopes EM, Vilas-Boas MDC, Dias D, Rosas MJ, Vaz R, Cunha JPS. iHandU: a novel quantitative wrist rigidity evaluation device for deep brain stimulation surgery. Sensors (Basel) 2020 Jan 07;20(2):331 [FREE Full text] [doi: 10.3390/s20020331] [Medline: 31936023]

XSL•FO
RenderX

51. Ile an RR, Cordo C, Mihăilă L, Fle ar R, Popescu A, Perju-Dumbravă L, et al. Proof of concept in artificial-intelligence-based wearable gait monitoring for Parkinson's disease management optimization. Biosensors (Basel) 2022 Mar 23;12(4):189 [FREE Full text] [doi: 10.3390/bios12040189] [Medline: 35448249]
52. Wijers A, Hochstenbach L, Tissingh G. Telemonitoring via questionnaires reduces outpatient healthcare consumption in Parkinson's disease. Mov Disord Clin Pract 2021 Oct;8(7):1075-1082 [FREE Full text] [doi: 10.1002/mdc3.13280] [Medline: 34631943]
53. Tarousi M, Sarafidis M, Androutsou T, Manta O, Giannakopoulou O, Koutsouri G, et al. Serious games for Parkinson's disease management as implemented in PROPHETIC platform. Health Informatics J 2021;27(2):14604582211011231 [FREE Full text] [doi: 10.1177/14604582211011231] [Medline: 33902340]
54. Katie BK, Nicky S. SMaRT-PD: a clinical decision support system for the management of Parkinson disease. Journal of Neurology, Neurosurgery & Psychiatry 2024;95:a [FREE Full text] [doi: 10.1136/jnnp-2024-abn.151]
55. Gatsios D, Antonini A, Gentile G, Marcante A, Pellicano C, Macchiusi L, et al. Feasibility and utility of mHealth for the remote monitoring of Parkinson disease: ancillary study of the PD_manager randomized controlled trial. JMIR Mhealth Uhealth 2020 Jun 29;8(6):e16414 [FREE Full text] [doi: 10.2196/16414] [Medline: 32442154]
56. Fabbri M, Caldas AC, Ramos JB, Sanchez-Ferro, Antonini A, Růžička E, et al. Moving towards home-based community-centred integrated care in Parkinson's disease. Parkinsonism Relat Disord 2020 Sep;78:21-26. [doi: 10.1016/j.parkreldis.2020.07.001] [Medline: 32674024]
57. Song Y, Yang Y, Cheng P. The investigation of adoption of voice-user interface (VUI) in smart home systems among Chinese older adults. Sensors (Basel) 2022 Feb 18;22(4):1 [FREE Full text] [doi: 10.3390/s22041614] [Medline: 35214513]
58. Espay AJ, Bonato P, Nahab FB, Maetzler W, Dean JM, Klucken J, Movement Disorders Society Task Force on Technology. Technology in Parkinson's disease: challenges and opportunities. Mov Disord 2016 Sep;31(9):1272-1282 [FREE Full text] [doi: 10.1002/mds.26642] [Medline: 27125836]
59. Hobert MA, Maetzler W, Aminian K, Chiari L. Technical and clinical view on ambulatory assessment in Parkinson's disease. Acta Neurol Scand 2014 Sep 01;130(3):139-147. [doi: 10.1111/ane.12248] [Medline: 24689772]
60. Ray TR, Choi J, Bandodkar AJ, Krishnan S, Gutruf P, Tian L, et al. Bio-integrated wearable systems: a comprehensive review. Chem Rev 2019 Apr 24;119(8):5461-5533. [doi: 10.1021/acs.chemrev.8b00573] [Medline: 30689360]
61. Cheng X, Jiao Y, Meiring RM, Sheng B, Zhang Y. Reliability and validity of current computer vision based motion capture systems in gait analysis: a systematic review. Gait Posture 2025 Jul;120:150-160 [FREE Full text] [doi: 10.1016/j.gaitpost.2025.04.016] [Medline: 40250127]
62. Wade L, Needham L, McGuigan P, Bilzon J. Applications and limitations of current markerless motion capture methods for clinical gait biomechanics. PeerJ 2022;10:e12995 [FREE Full text] [doi: 10.7717/peerj.12995] [Medline: 35237469]
63. Moore W, Frye S. Review of HIPAA, part 1: history, protected health information, and privacy and security rules. J Nucl Med Technol 2019 Dec;47(4):269-272 [FREE Full text] [doi: 10.2967/jnmt.119.227819] [Medline: 31182664]
64. Mirelman A, Hillel I, Rochester L, Del Din S, Bloem BR, Avanzino L, et al. Tossing and turning in bed: nocturnal movements in Parkinson's disease. Mov Disord 2020 Jun 20;35(6):959-968. [doi: 10.1002/mds.28006] [Medline: 32080891]
65. Chudzik A, Śledzianowski A, Przybyszewski AW. Machine learning and digital biomarkers can detect early stages of neurodegenerative diseases. Sensors (Basel) 2024 Feb 29;24(5):s24051572 [FREE Full text] [doi: 10.3390/s24051572] [Medline: 38475108]
66. Chandrabhatla AS, Pomeraniec IJ, Ksendzovsky A. Co-evolution of machine learning and digital technologies to improve monitoring of Parkinson's disease motor symptoms. NPJ Digit Med 2022 Mar 18;5(1):32 [FREE Full text] [doi: 10.1038/s41746-022-00568-y] [Medline: 35304579]
67. Slemenšek J, Fister I, Geršak J, Bratina B, van Midden VM, Pirtošek Z, et al. Human gait activity recognition machine learning methods. Sensors (Basel) 2023 Jan 09;23(2):1 [FREE Full text] [doi: 10.3390/s23020745] [Medline: 36679546]
68. Khanna A, Adams J, Antoniades C, Bloem BR, Carroll C, Cedarbaum J, et al. Accelerating Parkinson's disease drug development with federated learning approaches. NPJ Parkinsons Dis 2024 Nov 21;10(1):225 [FREE Full text] [doi: 10.1038/s41531-024-00837-5] [Medline: 39567515]
69. Qadri YA, Ahmad K, Kim SW. Artificial general intelligence for the detection of neurodegenerative disorders. Sensors (Basel) 2024 Oct 16;24(20):1 [FREE Full text] [doi: 10.3390/s24206658] [Medline: 39460138]
70. Del Din S, Kirk C, Yarnall AJ, Rochester L, Hausdorff JM. Body-worn sensors for remote monitoring of Parkinson's disease motor symptoms: vision, state of the art, and challenges ahead. JPD 2021 Jul 16;11(s1):S35-S47. [doi: 10.3233/jpd-202471]
71. Chepke C, Shaughnessy L, Brunton S, Farmer J, Rosenzweig A, Grossberg G, et al. Using telemedicine to assess and manage psychosis among outpatients with neurodegenerative disease. IJGM 2021 Dec;Volume 14:10271-10280. [doi: 10.2147/ijgm.s335739]
72. Tripathi S, Malhotra A, Qazi M, Chou J, Wang F, Barkan S, et al. Clinical review of smartphone applications in Parkinson's disease. Neurologist 2022 Jul 01;27(4):183-193. [doi: 10.1097/NRL.0000000000000413] [Medline: 35051970]
73. Aksenova EI, Gorbatov SY, Pivovarova OA. Determining the level of technological readiness of developments in medicine based on the TRL methodology. Probl Sotsialnoi Gig Zdravookhranenniiai Istor Med 2021 Aug;29(Special Issue):1395-1399. [doi: 10.32687/0869-866X-2021-29-s2-1395-1399] [Medline: 34792895]

XSL•FO
RenderX

74. Espay AJ, Hausdorff JM, Sánchez-Ferro, Klucken J, Merola A, Bonato P, Movement Disorder Society Task Force on Technology. A roadmap for implementation of patient-centered digital outcome measures in Parkinson's disease obtained using mobile health technologies. Mov Disord 2019 May;34(5):657-663. [doi: 10.1002/mds.27671] [Medline: 30901495]
75. Valerio JE, Aguirre Vera GDJ, Fernandez Gomez MP, Zumaeta J, Alvarez-Pinzon AM. AI-driven advances in Parkinson's disease neurosurgery: enhancing patient selection, trial efficiency, and therapeutic outcomes. Brain Sci 2025 May 09;15(5):1 [FREE Full text] [doi: 10.3390/brainsci15050494] [Medline: 40426665]
76. MacPherson MK. Cognitive load affects speech motor performance differently in older and younger adults. J Speech Lang Hear Res 2019 May 21;62(5):1258-1277. [doi: 10.1044/2018_JSLHR-S-17-0222] [Medline: 31051090]
77. Ellis TD, Earhart GM. Digital therapeutics in Parkinson’s disease: practical applications and future potential. JPD 2021 Jul 16;11(s1):S95-S101 [FREE Full text] [doi: 10.3233/jpd-202407]
78. Bagrodia V, Holla VV, Kamble NL, Pal PK, Yadav R. Parkinson's disease and wearable technology: an Indian perspective. Ann Indian Acad Neurol 2022;25(5):817-820 [FREE Full text] [doi: 10.4103/aian.aian_653_22] [Medline: 36560983]
79. Ma B, Yang J, Wong FKY, Wong AKC, Ma T, Meng J, et al. Artificial intelligence in elderly healthcare: a scoping review. Ageing Res Rev 2023 Jan;83:101808. [doi: 10.1016/j.arr.2022.101808] [Medline: 36427766]
80. Barrachina-Fernández M, Maitín AM, Sánchez-Ávila C, Romero JP. Wearable technology to detect motor fluctuations in Parkinson's disease patients: current state and challenges. Sensors (Basel) 2021 Jun 18;21(12):1 [FREE Full text] [doi: 10.3390/s21124188] [Medline: 34207198]
81. Nahai F. General Data Protection Regulation (GDPR) and data breaches: what you should know. Aesthet Surg J 2019 Jan 17;39(2):238-240. [doi: 10.1093/asj/sjy296] [Medline: 30371747]
82. Tiersen F, Batey P, Harrison MJC, Naar L, Serban A, Daniels SJC, et al. Smart home sensing and monitoring in households with dementia: user-centered design approach. JMIR Aging 2021 Aug 11;4(3):e27047 [FREE Full text] [doi: 10.2196/27047] [Medline: 34383672]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
CNKI: China National Knowledge Infrastructure
CNN: convolutional neural network
CPT-PD: Care-Platform Transformation in PD
GAI: generative artificial intelligence
GAIDeT: Generative Artificial Intelligence Delegation Taxonomy
HIPAA: Health Insurance Portability and Accountability Act
ICC: intraclass correlation coefficient
IMU: inertial measurement unit
MDS: Movement Disorder Society
mHealth: mobile health
PD: Parkinson disease
PKG: Personal KinetiGraph
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
UPDRS: Unified Parkinson Disease Rating Scale

Edited by S Brini; submitted 06.Dec.2025; peer-reviewed by MZ Hossain, S Jiang, M Chakit; comments to author 30.Dec.2025; accepted 05.Mar.2026; published 17.Apr.2026.

Please cite as:
Tan J, Deng X, Wu C, Yao M, Liao J, Zheng H, Lian C
Multimodal Intelligent Monitoring of Parkinson Disease: Scoping Review of Progress and Translational Challenges
J Med Internet Res 2026;28:e89057
URL: https://www.jmir.org/2026/1/e89057
doi:10.2196/89057

©Jianhui Tan, Xiuzhen Deng, Cuilan Wu, Molin Yao, Jingyi Liao, Haoli Zheng, Changlin Lian. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Apr.2026. This is an open-access article distributed under the

XSL•FO
RenderX

terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Toward a Common Set of Interface Requirements for Genomic Data Management: Scoping Review

Valeria Resendez[1], PhD; Funda Yıldırım[1], PhD; Eugenio Gaeta[2], PhD; Giuseppe Fico[2], PhD; Simone Borsci[1,3], PhD

[1]University of Twente, Enschede, The Netherlands
[2]Universidad Politécnica de Madrid, Madrid, Spain
[3]Imperial College London, London, United Kingdom

Corresponding Author:
Valeria Resendez, PhD
University of Twente
Capitool 15 (building no. 78), room 347
Enschede, 7521PL
The Netherlands
Phone: 31 0612439689
Email: v.d.c.resendez@utwente.nl

Abstract

Background: Genomic data can advance precision medicine; however, to continue developing more targeted treatments, genomic datasets need to be integrated with health care data and become more disease-focused. This integration, in turn, amplifies existing challenges in health care data management, such as handling large data volumes, adhering to data standards, and protecting sensitive information. Addressing these challenges calls for unified digital ecosystems that combine data collection, standardization, analysis, and governance within a single platform, thereby reducing the technical burden for users. Currently, a clear set of indications about functional and nonfunctional requirements to help designers translate stakeholder needs into actionable design specifications is missing.

Objective: This scoping review aimed to identify the functional and nonfunctional requirements most frequently discussed in the literature from the perspective of end users (eg, clinicians and data analysts) to inform the design of a health and genomic data management platform that supports data sharing and analysis in clinical settings by conducting a PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Review) review.

Methods: We searched for peer-reviewed English studies that focused on platforms for managing genomic data from a user-centered perspective. We considered studies from 2014 to 2024 that were extracted from Scopus, PubMed, Web of Science, and Google Scholar for the scoping review. Insights were extrapolated for a thematic analysis to develop an initial set of requirements. We charted the functional and nonfunctional requirements according to their frequency of occurrence in the literature to provide a structured overview of the most commonly reported requirements.

Results: From 410 initial items, 210 items were preliminarily selected, and 53 items were included in the final analysis. Three primary groups of 26 interface functional requirements emerged: (1) general data management (acquisition, standardization, and sharing), (2) data processing and analysis (preprocessing and analysis pipelines), and (3) data visualization and reporting. Twenty nonfunctional requirements were identified and organized in 4 groups: (1) communication and support, (2) platform technical infrastructure, (3) user experience and user interface characteristics, and (4) security and compliance. We also investigated the issues that need to be resolved to develop an ideal platform.

Conclusions: We identified and mapped the most frequently reported functional and nonfunctional requirements of clinical and data professionals when discussing a health and genomic data management platform. The 3 key functional requirements should be supported by nonfunctional requirements such as secure technical infrastructure and governance mechanisms that enable compliant data processing and sharing. Designers may use these insights and mapping to develop standardized data platforms that promote efficient data exchange between institutions and experts while ensuring regulatory compliance and secure access, as proposed by the European Health Data Space.

(J Med Internet Res 2026;28:e78405) doi:10.2196/78405

XSL•FO
RenderX

KEYWORDS

functional requirements; nonfunctional requirements; genomics; electronic health records; interfaces; data sharing

Introduction

Background

Since the completion of the Human Genome Project in 2003, which marked a turning point in genomic research, advancements in sequencing technologies have increased both the volume and complexity of genomic data [1,2]. For nearly 2 decades, there has been a continuous evolution in genomic data management systems and the analytical methods used in genomic research. Genomic data play a vital role in enhancing patient treatments, from pharmaceutical developments [3] to improving outcomes of organ transplants [4]. Currently, many genomic and health care datasets are isolated or not disease-focused [5]. This certainly limits clinicians' and researchers' ability to advance precision medicine and develop more targeted treatments [6]. However, when health care data are combined with genomic data, the interoperability, computational, and legal challenges increase in complexity due to the quantity and size of data and to the stricter processing requirements. As Alzu'bi et al [1] highlight, platforms designed to manage, share, and analyze genomic data must tackle at least four challenges, which are (1) handle overwhelming volumes of data that often require preprocessing and standardization, especially important for genomic data; (2) simplify complex analysis for highly specialized (large set of) genomic data [1,4]; (3) manage data in line with the legal, social, and ethical issues associated with personal genomic information (eg, preprocessing and analyzing genomic data requires safeguarding patient privacy in line with national and international regulatory frameworks, as well as control to determine who is authorized to access the data [1,4]); and (4) ensure the security and privacy of genomic data through regulatory compliance. This means incorporating regulatory principles such as the General Data Protection Regulation (GDPR) [7], the European Data Act [8], and other European Regulations from the earliest stages of design. Both the platform and its intended use should be developed to maximize data protection and to promote secure data handling practices by design.

The latter challenge (ie, ensuring the security and privacy of the data) aligns with broader European initiatives aimed at creating a framework for health data management and exchange. For instance, the European Health Data Space [9] represents a key pillar of the European health global strategy [10]. The European Health Data Space is the ecosystem of rules, standards, practices, and infrastructures that aims to address the ethical and technical challenges of exchanging interoperable health data in a very diverse and complex environment like the European Union. Interoperability in this context refers to the ability of organizations to work collaboratively toward shared goals by exchanging information and knowledge [11]. The European context is not the only one that deals with the issues of ensuring the privacy and security of systems for exchanging health data. Recently, the US Department of Health and Human Services published an update on the US health care privacy rule [12]. How the EU and US companies, as well as practitioners in both systems, will access and exchange data with reciprocal advantages is the potential point of connection among the regulatory frameworks currently under discussion [13].

Enabling well-regulated ecosystems to exchange data (to a certain extent and in line with regulations) is an essential aspect of supporting responsible research, fostering innovation, informing policymaking, enhancing patient safety, and streamlining regulatory activities [14]. For the specific case of genomics, global initiatives, such as the Global Alliance for Genomics and Health (GA4GH) [15], ELIXIR [16], and the Genomic Data Infrastructure (GDI) [17], are working to tackle the challenges described above by promoting standards for genomic data sharing. For instance, GA4GH focuses on creating a common framework that enables effective and responsible data sharing, driving progress in genomic research and medicine. ELIXIR works to coordinate infrastructural resources, such as databases, software tools, training materials, cloud storage, and supercomputers, facilitating data discovery, expertise exchange, and the establishment of best practices. Meanwhile, GDI strives to provide access to genomic, phenotypic, and clinical data across Europe by establishing a federated, sustainable, and secure infrastructure for data access. Similarly, the European Union Joint Action Towards a European Health Data Space (TEHDAS) program reflects the broader European move toward a federated approach to health data exchange, including genomic information. Under a federated approach, data remains locally stored but can be queried and analyzed across sites [18]. Together, these initiatives illustrate an international recognition that responsible data sharing is fundamental to advancing health and genomic research and realizing the potential of precision medicine [19]. At the same time, the initiatives show a shift toward infrastructures that are interoperable, secure, and designed for collaboration at scale, ensuring that sensitive data can be used to generate scientific and clinical insights while respecting legal and ethical constraints.

A key challenge in improving health and genomic data management, as well as facilitating data exchange in a federated modality, is developing digital ecosystems that can unify data collection, standardization, exchange, and analysis on a single platform. By integrating data access, governance controls, preprocessing tools, analytical workflows, and visualization capabilities into a single interface platform, the main advantage will be providing clinical experts with access to the same datasets. This will also harmonize the analytical power and possibilities of analysis for operators in different countries [20,21]. These types of platforms can also support regulatory compliance and broaden participation by making genomic analysis accessible beyond bioinformaticians [2,22]. Enabling the accessibility of such information to different stakeholders is key to further advancing precision medicine and allowing more personalized treatments [23].

The development of a unifying federated digital ecosystem (ie, unified front-end with distributed data) needs the definition of requirements. Requirements specify the services a system must

XSL•FO
RenderX

provide and the constraints under which it must operate [24]. Describing requirements through a categorization into functional and nonfunctional enables the first high-level translation of needs into specifications that (with further refinements) can be used by designers and developers to build a platform. In this work, we build on previous definitions and define such requirements as follows:

1. Functional: These requirements refer to the platform's ability to provide operations that serve (stated or implied) the needs of the end users [25]. These functional requirements represent what the system can and must do [24,26].
2. Nonfunctional: These requirements are qualities of the platform that support the different functionalities to be used appropriately (eg, accessibility and usability) [27]. These nonfunctional requirements also define the constraints of the software architecture [24,26,28]. Nonfunctional requirements often apply to the system as a whole rather than individual system features or services (eg, security, access control, and availability).

Functional and nonfunctional requirements are entangled aspects; for example, empowering nontechnical users to analyze data easily could involve functional requirements like drag-and-drop interfaces or prebuilt analysis templates, and is connected to nonfunctional requirements such as accessibility and perceived usability [2,29-31].

Previous Work on Genomics Platform Requirements

Several projects have been in the past (and are currently) investigating and developing functional and nonfunctional requirements to enable, for instance, more secure exchanges of health and genomic data. For example, the IntelliOmics project is developing a distributed system that anonymizes data before making it accessible to users [4]. Similarly, the Integrated Microbial Genomes platform offers an interface with privacy functional requirements that allow users to decide whether they want their annotations and data to be public or private [32], as well as the Beacon v2 platform of GDI that provides data discovery functionalities designed to facilitate the search for genomic variants and associated information across distributed datasets without compromising data privacy [33,34].

St. Jude Cloud software (St. Jude Children's Research Hospital) provides a facilitated process of data integration by standardizing data collection and management through different steps of the data analysis process [2]. Additionally, this system uses open-source tools and common data standards to make it easier for researchers to share and combine diverse datasets for in-depth analysis [2]. In a similar way, to simplify data processing, the MetaboAnalyst initiative developed a system with a set of functionalities that allow for easier data preprocessing while providing interactive exploration for users [31]. Another example is the open platform BinaRena, which aims to enhance data visualization by rendering interactive scatter plots that integrate multiple data layers, thereby facilitating the delineation of microbial communities even in extensive datasets [35]. These projects collectively address key challenges in genomic data management. Nevertheless, we are not aware of any structured and systematic investigation that has mapped and listed a set of the most relevant functional and nonfunctional requirements for the potential end users and stakeholders of health and genomic data management.

A systematic and in-depth analysis of the functional and nonfunctional requirements reported in literature can provide a baseline set of requirements that can be used for future investigations with different stakeholders (ie, clinicians, developers, and statisticians). This analysis can also facilitate the development of genomic and health data platforms that are recognized for their levels of perceived quality in usage, such as usability [36] and user experience (UX; ISO 9241-210 [37]). Mapping functional and nonfunctional requirements is the first step in this process. It translates user needs, workflows, and system qualities into design characteristics that specify what the interface must offer. This indicates possible directions for the development of the system. These design characteristics align with the human-centric usability engineering process for medical [38] and software systems [39]. This process can result in the development of effective, useful, interoperable platforms for different stakeholders, thus maximizing the possibility of uncovering insights that can contribute to more targeted diagnoses and treatments, and improved health outcomes.

Aim of This Work

To our knowledge, no previous work has systematically mapped literature in order to identify a set of functional and nonfunctional requirements for genomic data management platforms commonly identified as important by experts. To address this gap, we conducted a scoping literature review following the PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) guidelines [39], which is usually adopted to explore a given field and establish standard practices and requirements, for example, study by Piškur et al [40]. Our research question aimed to identify the key requirements (functional and nonfunctional) considered important by experts and should be incorporated into platforms for genomic and health care data management.

This review provides 2 main contributions. First, it translates scattered insights from the literature into a structured set of functional and nonfunctional requirements, which can inform the development of a unified federated ecosystem. Second, it quantifies these requirements to identify design priorities and areas needing further attention. These contributions are particularly relevant for developers and designers of health care tools seeking to improve genomic research through safer, usable, accessible, and acceptable platforms.

Methods

Overview

We conducted a scoping review of platforms managing genomic data and reported our findings according to the PRISMA-ScR (Multimedia Appendix 1) [41]. The project was registered retrospectively under the open science framework [42].

XSL•FO
RenderX

Information Sources and Search Strategy

We conducted a scoping review to map the relevant requirements of platforms managing genomic data. Data searches were conducted between August 9 and August 14, 2024, in Scopus, PubMed, Web of Science, and Google Scholar, searching for peer-reviewed publications published in the last 10 years. The search strategy was organized around three core concepts, (1) data-related practices, (2) genomics, and (3) user-centered or human-centered approaches. Boolean operators ("AND" and "OR"), truncation, and phrase searching were applied to capture variations in terminology.

The inclusion of user-centered and human-centered terms reflects the goal to focus on the identification of user requirements, instead of focusing on features, technical limitations, and feasibility from a technical perspective. Specifically, we wanted to take as much as possible the perspective of the potential end users, including, for instance, researchers, clinicians, or data managers who interact with genomic data. Understanding user needs is key to capturing the functional requirements (eg, what the platform must be able to do) without presupposing the technical solutions or implementation details. This approach allowed us to identify user-facing functionalities, for example, data visualization or data analysis workflows, without the need to commit to technical details (eg, programming language, database schema, or cloud architecture).

For this reason, the following search terms were used to filter titles, abstracts, and keywords in Scopus, PubMed, and Web of Science: "data management" OR "data processing" OR "data exploration" OR "data discovery" OR "data sharing" OR "data analysis" OR "data integration" AND "genomic" OR "genes" OR "DNA" AND "user-centered" OR "user centered" OR "human-centered" OR "human centered" OR "user experience" OR "user requirements" OR "usability."

In Google Scholar, where filtering by specific fields is not possible, we only applied the same search string to retrieve relevant articles. The aggregated search results were exported into Rayyan, a platform developed for literature reviews, where duplicates were removed. The final selection of the search results can be accessed in the Open Science Framework (OSF) repository for this project [42].

Eligibility Criteria

The authors established inclusion and exclusion criteria through discussions before the review (refer to Multimedia Appendix 2 for summary). We focused on articles describing key functional and nonfunctional requirements for ensuring a high-level UX, as well as those relevant to user-centered design of health and genomic data platforms. We excluded non-English articles, backend-focused and technical papers, inaccessible content, studies not involving human genomic data management or focusing on specific statistical software analyses, and literature reviews lacking platform requirements. We included only items published from August 2014 to 2024.

Selection of Source of Evidence

Following PRISMA-ScR guidelines, 1 reviewer (VR) initially screened the abstracts for duplicates and relevance. Then, 2 reviewers (VR and SB) independently assessed titles and abstracts, with a third reviewer (FY) resolving any disagreements. We extracted article metadata (citation count, DOI, publication year, keywords, source, and authors) and both functional and nonfunctional platform requirements from selected articles.

Data Charting Process

One of the authors (VR) identified potential codes based on the features and aspects reported in the articles as key to the experience of the stakeholders with data platforms. In addition, 2 authors (SB and FY) independently reviewed the initial codes, and disagreements were resolved through iterative discussion until consensus was achieved. The agreed-upon codes were then compared across articles and grouped into broader thematic categories, which formed the basis for the functional and nonfunctional requirements. Requirements describing specific actions, operations, or services the platform must perform (such as data upload, search functions, and workflow execution) were classified as functional. Requirements describing system qualities or constraints (such as security, usability, and scalability) were classified as nonfunctional. The final categorization was reviewed and approved by all authors. The frequency with which requirements appeared across articles was recorded as an indicator of attention in the literature, although this measure should be interpreted cautiously, as it may reflect publication patterns rather than user priorities.

Data Items

The final selection included 53 data items. From each item, we extracted general bibliographic information, such as title, digital object identifier, publication year, authors, and article type, to enable the easy identification of the data items. In addition, following our data charting procedure, we extracted 108 variables capturing both functional requirements (eg, data access, data upload, and data search) and nonfunctional requirements (eg, security, interoperability, scalability, performance, and accessibility). A full list of the selected data items and the complete set of extracted requirements is available in the project's OSF repository [42].

Synthesis of the Results

The full texts of the articles were analyzed to identify all the functionalities of the platform and its characteristics. Through an iterative process of data charting and thematic analysis, these coded characteristics were organized into 2 overarching themes—functional and nonfunctional requirements. These themes reflect how platforms operate and what they do. The functional requirements incorporated themes related to general data management, data processing and analysis, and data visualization and reporting. Nonfunctional requirements capture qualities that enable or constrain these functions. These qualities include the platform's technical infrastructure, user interface (UI) and UX characteristics, security and compliance mechanisms, and communication and support.

XSL•FO
RenderX

Study Risk of Bias Assessment

Given the nature of the scoping review and its aim to map the requirements of genomic data management platforms, a formal risk of bias assessment tool was not applied. We adopted an iterative refinement of these criteria before data extraction to further ensure rigor and consistency. The risk of bias in the process of inclusion or exclusion was assessed through a collaborative approach. Two reviewers (VR and SB) independently screened the titles and abstracts of all retrieved articles to minimize subjective biases during the selection process. Disagreements between the 2 reviewers were resolved by consulting a third reviewer (FY).

Results

Study Selection

The PRISMA-ScR flowchart is presented in Figure 1. A total of 410 studies were identified during the database search (151 records from Scopus, 69 from PubMed, 83 from Web of Science, and 107 from Google Scholar). After removing 200 duplicates, the titles and abstracts of the remaining 210 items were screened. After removing 29 items, a total of 181 items were read in full. About 128 items were also excluded. The remaining 53 items met the criteria, that is, providing information regarding essential functional and nonfunctional requirements of platforms related to health and genomic data management.

Figure 1. Overview of PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) procedure.

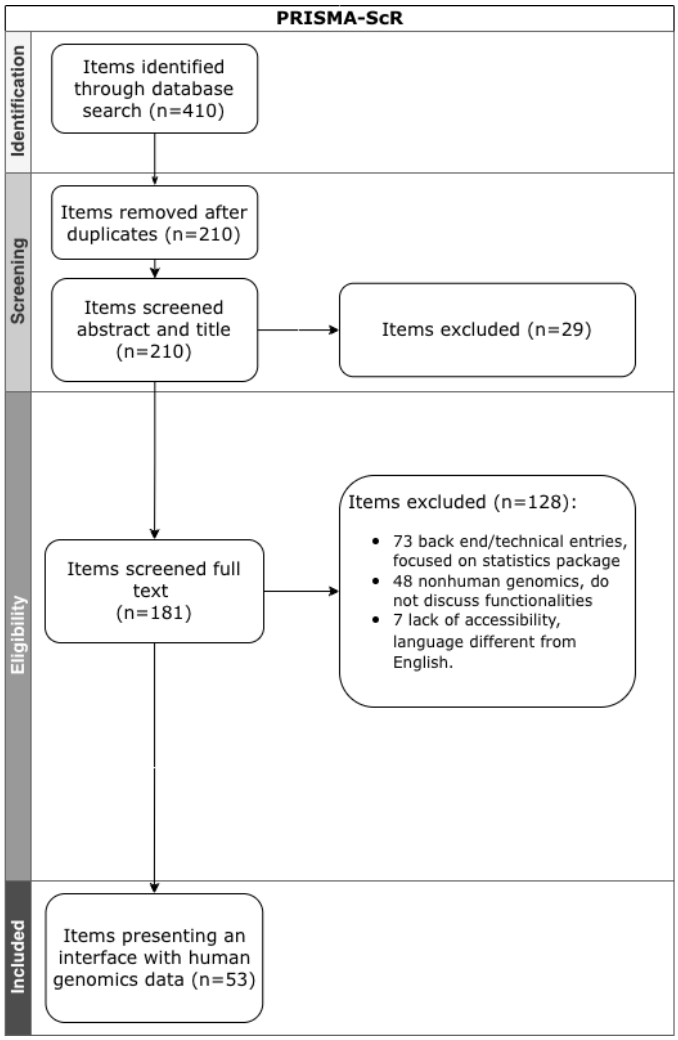

Study Characteristics

The items selected for the review include studies investigating features of genomic data management platforms from 2014 to 2024 (peaking during 2019-2022 with 8-9 articles annually), published in a few key journals, for example, "*Nucleic Acids Research*" (8 articles) [31,43-49], "*BMC Bioinformatics*" (5 articles) [50-54], and "*BMC Genomics*" (2 articles) [32,55]. Geographically, the majority of the articles originate from the United States (21 studies) [2,22,29,32,35,44,52,54-67], followed by Canada (8 studies) [31,47,50,68-72], Germany (5 studies) [45,48,53,73-75], Italy (4 studies) [51,53,76,77], the United Kingdom (3 studies) [49,78,79], Belgium (2 studies) [80,81], China (2 studies) [43,82], and 1 article each from Estonia [46], France [83], Greece [84], India [85], Japan [86], Pakistan [87], Poland [4], and Russia [88]. First authors were predominantly male (n=39) versus female (n=14).

XSL•FO
RenderX

Functional Requirements Groups

Overview

We identified 26 functional requirements organized in 3 main groups and associated subgroups (Figure 2) commonly reported in the literature (refer to Multimedia Appendix 3 for full overview). We have described these groups as follows:

1. General data management: This includes essential functional requirements discussed in 98% (n=52) of the articles to ensure appropriate data acquisition, data standardization, and data sharing.
2. Data processing and analysis: This group (discussed in n=53, 100% of the articles) includes key functional requirements for the stakeholders, such as data preprocessing and analysis pipelines and methods.
3. Visualizing data and generating outputs for reporting: This group was discussed in 90% (n=48) of the articles and included the main options for stakeholders to visualize specific types of data and generate reports on the platform.

Figure 2. Summary of the key functional requirements for the genomic research process.

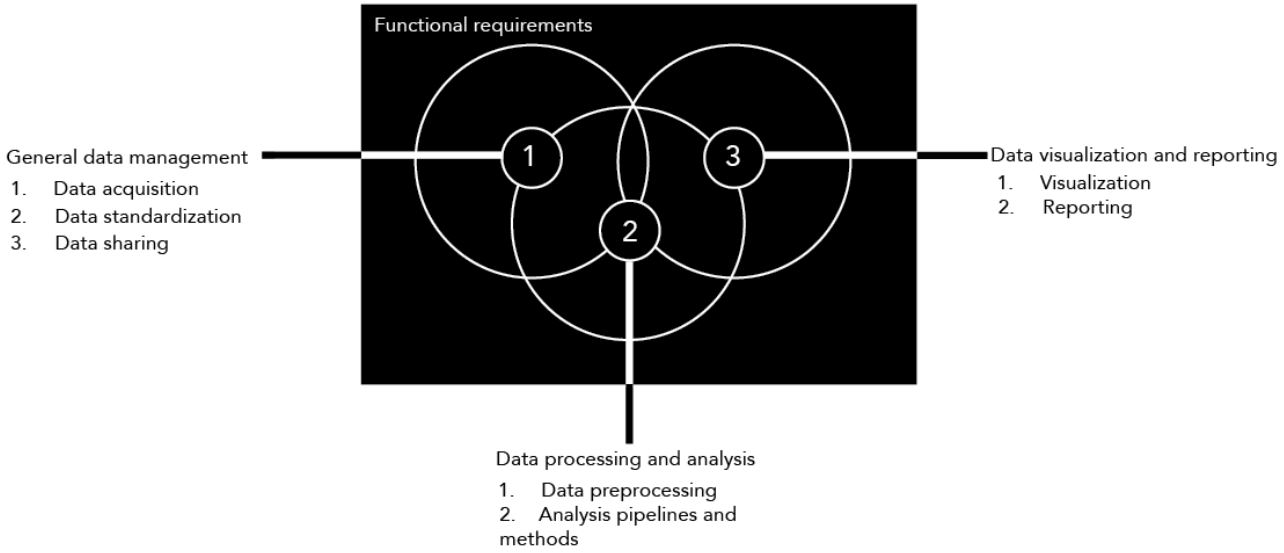

Associated with the functional requirements, we also mapped the nonfunctional requirements and constraints of the system, that is, inherent characteristics of the platform [24]. Specifically, these requirements include (1) communication and support (39 articles) [2,4,22,31,32,35,43-46,48,53,55-58,60-64,66,68,69,71,72,74-79,81-87], (2) UX and UI characteristics (50 articles) [2,4,22,29,31,32,35,43-56,58-65,67-85,87,88], (3) security and compliance (32 articles) [2,4,22,29,31,32,48,52-56,58-60,62-67,70,71,73-75,78,82,84,86-88], and (4) platform technical infrastructure (32 articles) [2,4,22,29,31,46,49,51-54,56-58,60-67,71,72,74,78,80,82-85]. Table 1 presents an overview of the extent to which various articles address the functional and nonfunctional requirements identified in the analysis.

XSL•FO
RenderX

Table 1. Functional and nonfunctional groups of requirements that are mentioned by authors in each selected article.

Study	Main groups of functional requirements			Main group of nonfunctional requirements			Security and compliance
	General data management	Data processing and analysis	Data visualization and reporting	Communication and support	Platform technical infrastructure	UI[a] and UX[b] characteristics	
Demchak et al [56]	✓	✓	✓	✓	✓	✓	✓
Xia et al [68]	✓	✓	✓	✓	__[c]	✓	—
Sante et al [80]	✓	✓	✓	—	✓	✓	—
Xia et al [43]	✓	✓	✓	✓	—	✓	—
Suciu et al [50]	✓	✓	✓	—	—	✓	—
Sauria et al [57]	✓	✓	—	✓	✓	—	—
Calabria et al [51]	✓	✓	✓	—	✓	✓	—
Wolf et al [73]	✓	✓	✓	—	—	✓	✓
Bhuvaneshwar et al [52]	✓	✓	✓	—	✓	✓	✓
Murtagh et al [78]	✓	✓	—	✓	✓	✓	✓
Ding et al [44]	✓	✓	✓	✓	—	✓	—
Chen I et al [32]	✓	✓	✓	✓	—	✓	✓
Albuquerque et al [69]	✓	✓	—	✓	—	✓	—
Davis-Turak et al [29]	✓	✓	✓	—	✓	✓	✓
Takai-Igarashi et al [86]	✓	✓	✓	✓	—	—	✓
Danahey et al [58]	✓	✓	✓	✓	✓	✓	✓
Lau et al [22]	✓	✓	✓	✓	✓	✓	✓
Crawford et al [59]	✓	✓	✓	—	—	—	✓
Das et al [70]	✓	✓	✓	—	—	✓	✓
Warner et al [60]	✓	✓	✓	✓	✓	✓	✓
Mohr et al [74]	✓	✓	✓	✓	✓	✓	✓
Pearce et al [61]	✓	✓	✓	✓	✓	✓	—
Hombach et al [45]	✓	✓	✓	✓	—	✓	—
Raudvere et al [46]	✓	✓	✓	✓	✓	✓	—
Nanni et al [53]	✓	✓	✓	✓	✓	✓	✓
Verbruggen and Menschaert [81]	✓	✓	✓	✓	—	✓	—
Wünsch et al [75]	✓	✓	✓	✓	—	✓	✓
Cappelli et al [76]	✓	✓	✓	✓	✓	✓	—
Kuzmenkov et al [88]	✓	✓	✓	—	—	✓	✓
Kounelis et al [84]	✓	✓	✓	✓	✓	✓	✓
Rodchenkov et al [47]	✓	✓	✓	—	—	✓	—
Bonomi et al [62]	✓	✓	✓	✓	✓	✓	✓
Yousif et al [54]		✓	✓	—	✓	✓	✓
Yukselen et al [55]	✓	✓	✓	✓	—	✓	✓
Holtgrewe et al [48]	✓	✓	✓	✓	—	✓	✓
Canakoglu et al [77]	✓	✓	✓	✓	—	✓	—
Ma et al [63]	✓	✓	✓	✓	✓	✓	✓
Reska et al [4]	✓	✓	✓	✓	✓	✓	✓

XSL•FO
RenderX

Study	Main groups of functional requirements			Main group of nonfunctional requirements			Security and compliance
	General data management	Data processing and analysis	Data visualization and reporting	Communication and support	Platform technical infrastructure	UI[a] and UX[b] characteristics	
Pang et al [31]	✓	✓	✓	✓	✓	✓	✓
Campbell et al [64]	✓	✓	✓	✓	✓	✓	✓
Ochoa et al [49]	✓	✓	✓	—	✓	✓	—
McLeod et al [2]	✓	✓	✓	✓	✓	✓	✓
Ullah et al [87]	✓	✓	✓	✓	—	✓	✓
Raveendran et al [65]	✓	✓	✓	—	✓	✓	✓
Gilbert et al [79]	✓	✓	✓	✓	—	✓	✓
Li et al [82]	✓	✓	✓	✓	✓	✓	✓
Osmond et al [71]	✓	✓	✓	✓	✓	✓	✓
Reiff et al [66]	✓	✓	✓	✓	✓	✓	—
Amer-Yahia et al [83]	✓	✓	✓	✓	✓	✓	—
Pavia et al [35]	✓	✓	✓	✓	—	✓	—
Post et al [67]	✓	✓	—	—	✓	✓	✓
Gill et al [72]	✓	✓	✓	✓	✓	✓	—
Ware et al [85]	✓	✓	—	✓	✓	✓	—

[a]UI: user interface.

[b]UX: user experience.

[c]Not applicable.

General Data Management

General data management encompasses key functionalities of the platform required to enable data analysis and visualization of results [4]. These functionalities can be clustered in three subgroups as described below (Section S1 in Multimedia Appendix 3).

First, data acquisition functionalities refer to the information collected from various sources, including electronic health records (EHRs), laboratory results, open-source databases, and proprietary data from different hospitals [4,83]. These functionalities encompass two main subfunctionalities: (1) the platform's ability to aggregate data from various sources, including databases, graph structures, and textual sources [83]; and (2) the ability of the platform for users to upload their data [66,72,83,85]. The first subfunctionality is noted in 45 articles [2,4,22,29,31,32,35,45-52,54-59,61-64,66-77,79,80,83-88], and it aims to serve users with abundant information to find undiscovered patterns. The second subfunctionality (ie, allowing users to upload their data) is discussed in 30 articles [2,4,22,29,31,32,44-46,48,50-52,55-57,59,63,68-70,73-76,80,84,85,87,88]. Across the articles, the authors describe a broad range of data formats that users can rely on for upload. These include raw genomic data and genomic variants, such as FASTQ, BAM, and VCF [2,22,29,45,61,80], but also data regarding processed genomic and multiomics, such as CNA, SNP, CpG methylation, mRNA expression, proteomics, or metabolomics profiles [4,61,70], and gene sets or annotation tables including GMT files, CSV, TXT, Excel spreadsheets, BED tracks, and JSON query files [67,72,74]. Several platforms also support uploads of additional auxiliary data types, such as images (eg, PNG, JPG, medical CT images, and GIS files) or general documentation files (PDFs, Word, and Excel reports), when these are relevant for metadata management or quality control [4,51,67,74]. Including an option to upload data introduces additional requirements for the platform (eg, data security, standardization of data formats, and robust system performance), as it enables researchers and clinicians to manage and analyze their own information, independent of predefined sources [65]. At the same time, it provides flexibility by allowing researchers to work with original or observational data that aligns with their specific research needs [4], ultimately enhancing the platform's adaptability and relevance to solve research questions. This combination of characteristics (ie, aggregating data from various sources and allowing users to upload their data) can ensure that the platform maintains a robust and up-to-date collection of data sources, fostering a more collaborative environment that benefits broader research initiatives [2].

Second, data standardization functionalities ensure that data from different sources and in different formats is properly aggregated. Aligned to the goal of GA4GH, the standardization of the data involves harmonizing and unifying data to establish compatible and consistent formats. Overall, standardization is recognized as important, with 34 articles highlighting its relevance [2,4,22,29,31,46,49,51,53,58-60,63,65-68,70-80,82,83,85,87,88]. However, approaches to data standardization vary by platform. For instance, 9 items only explicitly highlight the

XSL•FO
RenderX

importance of standardization (eg, studies by Danahey et al [58], Cappelli et al [76], and Ullah et al [87]), 4 items suggested using libraries and Application Programming Interfaces (APIs), such as HTSJDK is an open source Java library, as a tool to standardize [63,66,71,73], and 3 articles propose using pipelines, such as phenotype-genotype harmonization pipelines, for standardizing data (eg, studies by Crawford et al [59], Das et al [70], and Ochoa et al [49]). The standardization process requires constraints to guide different automated approaches; we mapped the various data models and frameworks that emerged in the literature. These models and frameworks serve as nonfunctional requirements, simplifying efforts toward achieving consistency. However, we observed that despite the need for consistency, 10 articles specifically mentioned using customized data models [22,46,52,60,66,74,76,77,83,89]. As explained by Gill et al [72], a reason may be due to the lack of tools that support the implementation of data standards. Other articles refer to established data models that facilitate standardization, including frameworks such as DataSHaPER [77], PATRIC [89], Genomics Data Commons, Genomic Data Model [53], Variant Call Format [75], and EUROCAST [88]. Additionally, global ontologies [83] play a relevant role in providing consistent vocabularies for standardization. Another method involves converting various genes, proteins, and probe identifiers to a common reference, such as the Ensembl gene identifier [90]. Together, these frameworks, ontologies, and identifier systems underpin efforts to achieve data consistency, enable data sharing, and ensure interoperability across studies and platforms.

Third, data sharing functionalities refer to the ability to distribute, access, and exchange data across various platforms, institutions, or user groups, allowing users to access open-source data while also enabling them to apply for access to proprietary datasets [2,79,87]. These were discussed in 30 items and encompass characteristics, such as collaboration tools, that allow users to work together on specific datasets, with shared access, viewing, and modification rights in certain cases. Collaboration characteristics were mentioned in 25 articles [2,4,29,32,45,48,49,52,53,55,64-67,70,71,74,76-79,82-84,88]. These characteristics facilitate coordinated research efforts and joint analysis among researchers, institutions, or community members, often under controlled permissions. Linked to collaboration, customizable access permissions provide tailored control settings that grant distinct levels of access based on users' roles or needs within a system. These may include read-only access, full editing rights, or specific access to data, cases, or experiment groups. The flexibility from customizable access permissions helps organizations manage data security, ensure data integrity, and restrict sensitive information as needed while still enabling collaboration (mentioned in 20 articles [2,22,32,49,51-53,55,58,63,65-67,71,74,77,78,83,84,88]). Additionally, the user application for data access is connected to the process of sharing data. User application to data refers to the structured process where users request permission for specific datasets, especially when data sensitivity or security requires controlled access. This process typically involves filling out forms, agreeing to the terms of use, and, in some cases, obtaining approval from a data access committee. This is addressed in 20 articles [2,32,45,47,48,50,51,56,59,66,69,73,76-79,82-84,86].

Data Processing and Analysis

The second group among the functional requirements of a well-designed platform of health data management concerns the step of data processing and analysis [4]. This step encompasses 2 essential functionalities (ie, data preprocessing and data analysis methods) aimed at ensuring data quality and deriving meaningful insights (Section S2 in Multimedia Appendix 3). We clustered the functionalities as follows:

First, data preprocessing functionalities allow users to establish standards and parameters to assess the quality, consistency, and suitability of data before computational tasks are undertaken. The functionalities included in data processing involve evaluating data completeness, redundancy, and alignment with research requirements to prepare the data for analysis [35,49,72]. This process aims to ensure that the data are findable (ie, data sharing), accessible (ie, data standardization), interoperable (ie, data standardization), reusable (ie, data acquisition), and fair. In a way, preprocessing the data also entails evaluating its quality. The specific aspect of inspecting data quality was mentioned in 17 articles [35,51,52,57,59,63,66-69,72-75,81,82,88]. The assessment of the quality can be done by completeness and redundancy (contamination) assessment [72], establishing automated checks that alert users of missing or improperly formatted data [63], evaluating sequencing data with quality control on specific format (ie, bam) alignments, and validation of results from, for example, ChIP-seq, RNA-seq, and ATAC-seq [75]. When the data does not fulfil research requirements, functionalities for data normalization are suggested in 19 articles [4,29,31,35,43,45,47,57,59,61,63,65,66,68,70,76,77,82,83]. Normalization functionalities aim to make the data consistent by applying various mathematical transformations [66]. Thus, while standardization is typically applied to aggregate data, many studies highlight the importance of verifying data quality and performing normalization before analysis to ensure the data aligns with specific research requirements.

Second, data analysis methods: these functionalities are related to enabling and guiding (pipelines) users to perform different types of analysis (noted in 48 articles [2,4,22,29,31,32,43-50,52-57,59-86,88]. The types of analysis go from sequencing data analysis (such as expressed microRNAs in cancer) and analyzing data from multiple "omics" levels (eg, RNA, protein, and DNA methylation), discussed in 15 articles [2,22,44,49,50,53-55,66,68-70,81,83,85], to statistical analysis that included regressions (discussed in 7 articles [4,31,63,64,70,82,88]). Moreover, 15 articles [4,29,32,44,46,47,49,52,56,62,63,68,71,75,87] mention specialized analysis, like, for instance, pathway analysis; that is, the systematic study of biological pathways, such as metabolic, signaling, or gene regulatory pathways, to understand how specific molecular changes (eg, mutations and gene expression) affect biological processes and diseases. Similarly, network analysis (eg, studies by Pavia et al [35], Raveendran et al [65], Reiff et al [66], and Gill et al [72]) involves examining the relationships and connections between biological entities,

XSL•FO
RenderX

such as genes, proteins, metabolites, or pathways, to explore how their interactions shape biological processes and contribute to diseases. To facilitate end users, automated and customizable pipelines are often suggested as key functionalities, that is, a streamlined end-to-end workflow that automatically performs a series of data processing and analysis steps, allowing a standardized approach for analysis, for example, studies by Ullah et al [87], Amer-Yahia et al [83], Post et al [67], and Gill et al [72]. Moreover, the possibility to use command-line option for advanced analysis is remarked in about 15 articles [2,4,22,29,57,59,62,65,69,70,72,75,82,85,87]. Finally, an important aspect highlighted in the data analysis functionality is the inclusion of features that ensure the reproducibility of findings, checking that research findings are consistent, reliable, and can be independently verified. These characteristics were mentioned in 28 articles [2,4,22,29,31,46,48,52,53,55,62-66,70,72,74-78,80-83,86,87]. Together, these data analysis functionalities enable researchers to perform thorough, flexible, and reproducible analyses, enhancing the effectiveness of their insights.

Data Visualization and Reporting

The third group of functional requirements for a well-designed platform for health data management focuses on exploring insights through data visualization and producing outputs for reporting (Section S3 in Multimedia Appendix 3). Two of the requirements were identified in the literature.

First, visualizations of data simplify and support the interpretation of complex information, enabling users to identify patterns, trends, and relationships within the data [2,49,91]. Visualization of data was mentioned as a key function in 40 articles [2,4,22,29,31,32,35,43,44,46,47,49-56,59-66,70,73-76,79-84,87,88]. Data visualization might involve different types of visualizations like networks [56,82,83], scatter plots [35,50,64], genome browsers [55,65,68], heatmaps [54,57,64], pie charts [61,81,88], histograms [35,61,64], and dashboards [63] to present information clearly, easily, engagingly, and understandable.

Second, reporting functionality facilitates the creation of files and documents that effectively communicate the results of the analysis. This ensures that findings can be shared, published, and used for further research or practical applications [2,35,55,82]. The ability to generate reports that provide an overview of the data in different formats and modalities is a well-discussed feature discussed in 22 articles [2,4,29,35,45,46,48,49,51,55,60,64,68,71,73,75,77,80-83,88]. Related to generating reports is the option to download data in different formats, which appears in 31 articles [2,31,32,35,43,44,46,49-53,55,56,59,61,62,64-66,68,70,72-76,80,82,86,87]: Tab-Separated Values, images, text files, and JavaScript Object Notation files. Generation of reports is also connected to knowledge dissemination (mentioned 9 times in articles, such as studies by Reska et al [4], Yukselen et al [55], and Campbell et al [64]), as it facilitates sharing insights with others. To enhance the efficiency of sharing insights, incorporating citation buttons for easy integration into reports is also considered a useful approach [65,87].

Nonfunctional Requirements

Overview

User interactions with a platform are influenced not just by the direct actions that users can perform, but also by the platform's characteristics (nonfunctional requirements). We identified a total of 20 nonfunctional aspects discussed in the literature, and we clustered such aspects into 4 groups (Multimedia Appendix 4). An overview of the functional and nonfunctional requirements is presented in Figure 3.

Figure 3. Overview of the different functional and nonfunctional requirements mapped from the literature.

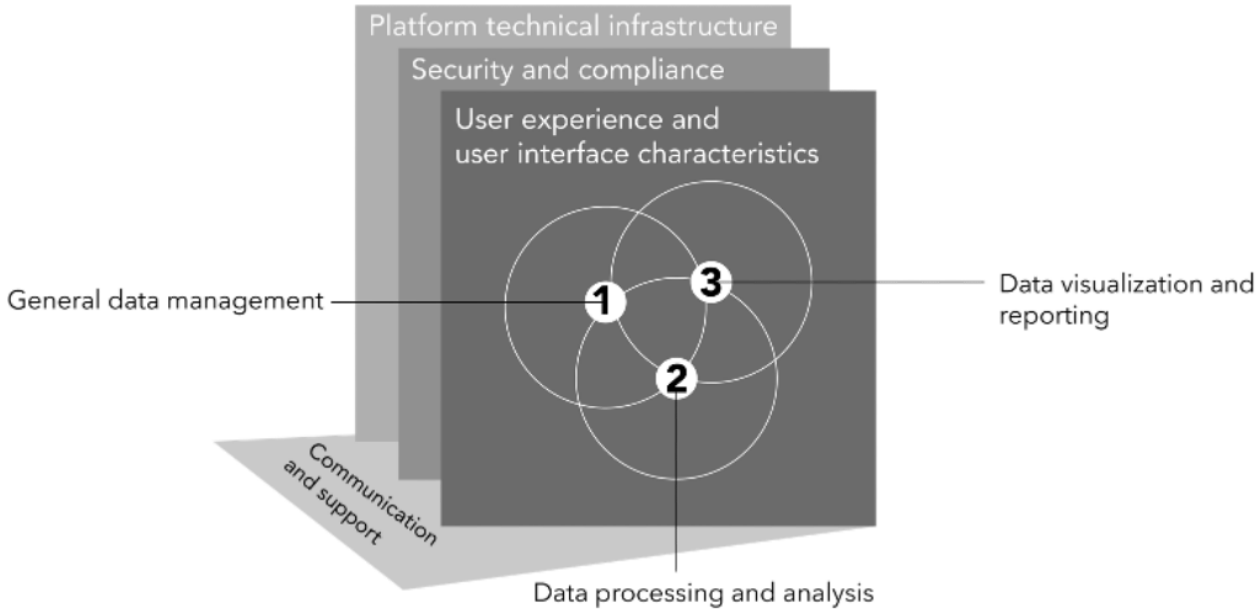

XSL•FO
RenderX

Communication and Support Requirements

These nonfunctional aspects are a unique intersection of functional and nonfunctional aspects, encompassing both user interactions with the platform and the platform's autonomous functionalities. The integration of these aspects enables communication between the user and the system while establishing operational constraints. For example, the function of sending reminders a certain amount of time introduces a timeline constraint, a nonfunctional requirement that aims to facilitate compliance. The communication and support functionality comprises 4 primary characteristics. First is *enabling feedback* from internal and external users [85]. This feature was mentioned in 16 articles [2,22,31,44,58,60,61,64,68,72,75,81,83-86], and it refers to information, comments, or data provided to improve, inform decisions to reinforce the improvement of the system.

A second key aspect is to include *documentation about the platform*, for example, to help users understand, navigate, and use a platform effectively. This includes tutorials, user guides, instructional manuals, and technical documentation. Providing users with thorough documentation has been highlighted in 28 articles [2,22,31,32,35,43-46,48,53,55-57,62-64,66,69,72,74-78,82,83,87], and it serves as a necessary support tool for users to learn about the platform's characteristics, workflows, and best practices.

A third key aspect is to enable users to set up *notifications*. This feature was mentioned 4 times (eg, studies by Gilbert et al [79] and Osmond et al [71]). Notifications in this context refer to alerts or messages that inform users (such as participants, data managers, or researchers) about new or important updates related to data or research. For instance, when new data becomes available, when participants are eligible to participate in a new study, or when there are updates on the research results. Finally, some articles mentioned that *allowing for communication between patients and clinicians* is also relevant. This feature was mentioned in 2 items [50,79].

Platform Technical Infrastructure

For the platform technical infrastructure, we see that a federated approach to data management was mentioned in 8 articles [2,29,53,71,73,74,78,82]. As federated infrastructure enables decentralized data storage and computation, ensuring data privacy by keeping sensitive information local while allowing collaborative analysis across sites, there is a trend over recent years, indicating its growing adoption. Another relevant feature within the architecture was ensuring the scalability and adaptability of the architecture. As noted in 22 articles [4,22,29,31,49,51-54,57,58,62-64,66,67,72,74,80,82,84,85], this feature highlights the need for systems designed to handle future growth, including increased data volume or user demand, to ensure the robustness of applications as needs evolve. Finally, the use of APIs to facilitate the ease of data integration and exchange of information was mentioned in 13 articles [2,22,46,49,51,53,56,60,61,63,65,67,76]. These articles further emphasize the need for scalability and adaptability within the system. The platform architecture category highlights the growing importance of federated data management, scalability, adaptability, and API usage. Federated approaches, as seen in multiple studies, enable decentralized data storage and secure collaborative analysis, reflecting a trend toward broader adoption. The scalability and adaptability of systems, emphasized in numerous articles, ensure that platforms can grow and evolve to meet increasing data and user demands. Additionally, the integration of APIs facilitates seamless data exchange, further supporting the flexibility and expansion of the architecture to accommodate future needs.

UX and UI Characteristics

Several key characteristics were highlighted in the Core UI and UX category to improve platform interactivity and UX. For example, the importance of a usable interface was emphasized in 35 articles [4,29,31,35,43,45-50,52,54,55,58,61,62,65-68,70,71,73,75,76,78-85,87], particularly given that not all users have the technical expertise required to navigate complex programming interfaces for data analysis.

One main feature to make the platforms user-friendly was searching for data, discussed in 36 articles [2,4,22,29,31,32,43-46,49-53,55,58,60,63-67,69-71,73,74,76,77,80,83,84,87,88]. Searching for data refers to the process of locating specific datasets or pieces of information within a more extensive database or data system. This can involve querying databases using various search methods, like entering keywords, applying filters, or selecting specific attributes (eg, accession number, item type, or metadata). Searching allows users to efficiently find relevant data by navigating through records and retrieving data based on specific criteria.

However, other specific characteristics could enhance user interaction with visual data, as highlighted in multiple studies. For example, 10 articles [32,35,43,45,46,60,69,72,83,88] mention the importance of highlighting specific regions, while 11 articles [35,43,50,61,64-66,70,80,84,88] emphasize the usefulness of zooming in on graphs to explore certain genes more in-depth. Additionally, 6 articles [43,46,50,61,84,88] noted the relevance of hover effects. First, highlighting specific regions involves visually distinguishing some aspects of the data, typically using color or other visual cues to make them stand out. This helps users quickly identify and focus on the areas they wish to explore further, such as selecting contigs or data points in a chart or plot. Second, zooming on graphs allows users to dynamically adjust the level of detail displayed, enabling them to magnify specific sections of a graph, map, or plot. Third, hover effects provide users with additional information or functionality when they hover the mouse over a specific element. Together, these characteristics facilitate a more efficient and focused exploration of data, allowing users to analyze trends deeply while maintaining an overview of the broader dataset.

Other relevant characteristics that enhance user interaction with visual data include dropdown menus and drag-and-drop functionality, all of which were discussed across multiple studies. For instance, dropdown menus refer to allowing users to select options or filter data within an interface efficiently, as mentioned in 11 articles [35,45,50,51,58,60,64,66,72,77,84]. This feature is particularly useful for navigating large datasets or switching between different settings without cluttering the UI. Similarly, the drag-and-drop feature (mentioned in 9 articles

XSL•FO
RenderX

[31,35,46,55,56,64-66,84]) allows users to easily upload files or rearrange graphical elements by dragging them across the screen. This feature increases the customization of visual data and enhances the interactivity of the platform, making it more intuitive and user-friendly.

Finally, 2 characteristics addressing the accessibility and device compatibility of the platform were having mobile-friendly interfaces (discussed in 8 articles [31,32,44,51,61,65,79,84]) and supporting multiple languages [52].

Security and Compliance

A total of 4 key characteristics emerged from the literature review in the security and compliance category. The first key feature is the *implementation of privacy-protective measures*. These include methods, such as data deidentification, anonymization, and other data protection measures designed to safeguard sensitive information by obscuring or removing personal identifiers, ensuring compliance with privacy laws and ethical standards. Privacy-protective measures were discussed in 20 articles [2,4,29,32,48,52,54,58,64,65,70,73-75,78,79,82,84,86,88]. The second feature identified is user registration. User registration is when individuals create an account or register on a platform, often by providing essential information such as a name, email address, and other necessary details. Registered users are usually granted access to more secure or restricted platform areas, ensuring that sensitive data is only available to authorized individuals. This feature (ie, user registration) is noted in 13 articles [2,22,31,52,53,56,59,60,67,71,74,78,84], and it allows platforms to guarantee the security of sensitive data and track activity to maintain accountability while offering more personalized services to the users. Closely related to user registration, the third feature is user authentication. This mechanism continuously verifies users through multifactor authentication and real-time access control processes that monitor and regulate their access to systems or data. Authentication mechanisms were highlighted in 10 articles [52,55,56,58,60,67,70,74,84,86]. The fourth feature is consent management. From the patient's perspective, this refers to systems that enable individuals to provide, withdraw, or modify consent for how their data are used over time. Notably, 8 articles [4,58,59,63,71,79,82,86] highlighted the importance of consent systems to empower patients to adapt their preferences to changing needs, ensuring respect for autonomy and ethical data use.

Discussion

Main Findings

There is a consensus among experts that data-driven and evidence-based health care is possible only if we (1) establish standardized mechanisms for data exchange [92,93]; (2) enable all the actors to provide, access, and analyze the data for specific and regulated purposes [94]; (3) enable the digital transformation in a way that is respectful of the patients' ownership of the data.

The tension between the need to access and manipulate data for scientific research while respecting the privacy and safety of the patients is at the center of the discussion, often creating fragmentation. The literature reflects an ongoing debate about the features that should be included in health and genomic management platforms. Many of the proposed functionalities aim to facilitate data analysis, making genomic workflows more accessible to stakeholders beyond bioinformaticians (eg, studies by Gill et al [72], Duong et al [95], Almeida and Oliveira [96], and Melles et al [97]). At the same time, the discussion emphasizes the need to strengthen data ownership, privacy, and security (eg, studies by Arneson et al [86], Bonomi et al [62], and Almeida and Oliveira [96]), alongside the development of technical infrastructure that can enable data access, while maintaining privacy, for example, federated vs centralized data governance [98,99]. Together, these elements highlight the underlying sensitivity of genomic data and its importance for advancing precision medicine.

This work attempted to map, organize, and make sense of such a scattered discussion in literature. Our goal was to provide designers and developers with a set of functional and nonfunctional requirements discussed in relation to data management platforms for health care and genomics data that consider their users in the process. The functional and nonfunctional requirements for genomic data management help distinguish operational functionalities from platform qualities. The functional requirements outline the platform's core features that address the explicit or implicit needs of end users, while the nonfunctional requirements describe the platform's qualities (eg, accessibility and usability) that ensure these functionalities can be used effectively and appropriately. These mapped requirements can help designers and developers understand diverse user needs and anticipate potential trade-offs before requirements elicitation, ultimately enabling more informed and user-centered design decisions.

The nonfunctional requirements, such as the platform's technical infrastructure (eg, federated or centralized), security compliance, and UI characteristics, shape its overall functionality. These requirements must align with industry standards and stakeholder needs to ensure effectiveness. By serving as both a framework and a set of constraints, these requirements define the platform's "possibilities of action" in terms of functioning. Whereas communication and support (eg, manuals, help, and artificial intelligence (AI)–conversational agents) fall under the nonfunctional requirements, as it helps users learn how to use the systems appropriately, it also serves an additional role by compensating for design issues such as missing functionalities, interactive problems, malfunctioning, etc. In this sense, the nonfunctional requirements comprised the infrastructure necessary for integrating the 3 key functional requirements most frequently discussed in the genomic research process.

Overview of the Functional Requirements

The following functional requirements are relevant to ensure that the platform serves the needs of all key stakeholders.

General Data Management

The acquisition, integration, and upload of data from diverse sources is essential. Such platforms must aggregate information from EHRs, laboratory results, open-source databases, and proprietary hospital data to construct robust and comprehensive datasets [4,83]. Some platforms even allow users to upload their

XSL•FO
RenderX

own data files (eg, FASTQ, BAM, and VCF) [2,22,45,53]. The diversity of information enhances the research capacity by including data that reflects various real-world scenarios [2,45,52]. However, there are 2 important functionalities that should be considered before the acquisition of the data. First, the standardization of the data is necessary to ensure compatibility across datasets. Implementing predefined templates, controlled vocabularies, and standardized APIs enhances data standardization and can improve data quality and ensure reliability [2,77]. Although data standardization has been crucial in advancing health and genomic research, there remains a lack of consensus on the standards that should be adopted [49]. This issue is critical to address because nonstandardized data are either unusable or require more time and effort to process [49,78]. In this context, although GA4GH represents one of the largest initiatives in the field, it is mentioned in only a few articles. Thus, despite its significance, it remains underrepresented in the review. This could be due to its ongoing nature, the fact that not all standards are universally applicable, and other evolving factors. Therefore, it is essential for the health and genomic data fields to acknowledge more widely the current initiatives (eg, GA4GH, ELIXIR, and GDI), further agree on a unified approach for standardization, and promote it. In fact, the management of sensitive data, such as genomic data, demands adherence to strict regulations and standards. Meeting these standards further depends on the strength of the technical infrastructure. As such, the operationalization and definition of nonfunctional requirements are key to the success of such functionalities. Second, within the acquisition and standardization of the data, it is important to consider the usability and the time of the data application process for stakeholders to access health data [49,78]. To standardize the overall application procedure for consistency might streamline the process, concurrently it could be useful to provide examples or a wizard on the application form to guide users through the process [49,78].

Data Processing and Analysis

Genomic research platforms must support diverse types of analysis, including pathway, network, and regression analyses, to address the different research needs [72,83]. Within the analysis functionality, having the ability to run predefined, automated, and customizable pipelines (ie, workflows) is also crucial to ensure efficiency and flexibility in the analysis process to address specific research questions [49,87]. Furthermore, the reproducibility of analytical workflows is essential to ensure that research findings are reliable, independently verifiable, and scientifically credible [72]. These aspects are crucial due to the complexity of the data, where intricate, multistep, and specialized tools are needed [29]. As many analysis workflows resemble tangled spaghetti code rather than standardized, reproducible clinical processes [29], health data management platforms must simplify the analysis process, enabling researchers—including nonexperts—to extract meaningful insights more efficiently. In addition, the integration of AI into the analysis platform can identify patterns in large datasets, recommend the most suitable analytical methods, and predict outcomes based on historical data trends. For instance, pathway and network analyses, frequently used for understanding gene interactions and biological pathways, could be optimized by AI algorithms that adaptively select the most fitting analysis based on specific research questions and dataset characteristics. This capability would not only improve research efficiency but also support more accurate and insightful conclusions in genomics and broader health care research.

Visualizing and Reporting Data

This is relevant for making complex data accessible and interpretable. Characteristics such as charts, graphs, and maps help users identify patterns and trends that might be obscured in raw data. Similar to the role of automated and customizable analytical workflows in simplifying data analysis for nonexperts, visualization tools enhance data comprehension and insight extraction [35,56]. In addition, the integration of some nonfunctional requirements, such as highlighting DNA regions, zooming on graphs, and hovering, facilitates detailed exploration of genomic data without overwhelming the user. Allowing for transitions between broad overviews and granular analysis, as well as sensory accessibility, serves to reduce cognitive load and improve data comprehension through dynamic visual cues [100]. When integrated with reporting capabilities, these tools further support the dissemination of analytical findings, enabling more efficient communication of results [82]. In a way, the visualization and reporting characteristics enhance the utility of data management platforms for researchers, facilitating more informed decision-making and collaborative research efforts.

From Functions to Operationalization

The previous functional requirements mapping can serve as a basis for rethinking any current or future health management platform. However, once the functional requirements are translated into real systems, a second layer of requirements emerges. This second layer contains how functional requirements are operationalized, that is, through nonfunctional requirements. An example of this is the general data management functional requirement. This requirement specifies that the platforms should pull information from EHRs, laboratory systems, open databases, and user-uploaded genomic files, such as FASTQ, BAM, and VCF [1-5,29]. In order to operationalize the functionalities of data management, designers must design an interface that can guide users through complex tasks of visual information retrieval and by implementing search tools that might help users navigate, interact with, and understand large datasets (eg, studies by Mohr et al [74], Cappelli et al [76], Rodchenkov et al [47], and Canakoglu et al [77]). This process could rely on an interface that enables users to examine data at a meta level rather than accessing the underlying individual-level information directly, thereby supporting compliance with current regulatory frameworks [7,11,13]. In practice, this could be achieved through a federated approach in which data remain within institution-controlled nodes [99], and the UX and UI provide access only to metadata or aggregated outputs [2,4]. In this sense, nonfunctional requirements, such as UX and UI design, should not be viewed as merely aesthetic or technical add-ons but as part of an iterative design process that optimizes usability while remaining aligned with additional nontechnical requirements, such as the technical infrastructure and security and compliance protocols.

XSL•FO
RenderX

Platforms that aim to support the needed functions for the users, like multiple analytical methods, customizable workflows, reproducible pipelines, and even AI-assisted insights, should also consider the quality in the operationalization of such functionalities, that is, nonfunctional requirements. We are claiming that the success of future platforms for genomic data exchange is not only in the option of data processing and analysis, but also in the ability of the platform to handle high computational demand in a scalable way. To support this aspect, in literature (eg, studies by Ochoa et al [19], Pang et al [31], Demchak et al [56], and Rodchenkov et al [47]), several strategies were identified for managing high-dimensional datasets. For instance, Sauria et al [57] suggest a message passing interface parallelization process to allow for scalability, while Mohr et al [74] highlight scalability through high-performance computing clusters and cloud services integrated with automated workflows, which distribute analyses across interconnected computer resources. These forms of internal computational scalability also underpin federated infrastructures, where data remain at their source (eg, local hospitals) and computation is distributed across institutions [99]. In a federated approach, each site must possess sufficient local compute power or cloud access to process its own data, enabling multisite analysis without the need for data centralization [99].

Although federated data management was mentioned in only a small number of articles (eg, studies by McLeod et al [2], Wolf et al [73], and Mohr et al [74]), it represents a promising approach for managing health and genomic data. Particularly, since decentralizing data storage and processing ensures that sensitive data remains in one place, reducing privacy concerns while enabling collaborative analysis across multiple sites [2,49,78]. Instead of transferring raw data, machine learning models are trained across decentralized locations, sharing only model updates, such as learned parameters. Similarly, clinicians and researchers are not exposed to raw data; instead, the platform provides structured quality indicators, metadata profiles, and model-ready summaries that allow researchers to understand the suitability of a dataset without directly querying its content. On the one hand, this offers clinicians and researchers an environment that removes the need for programming expertise, enabling broader participation from groups that are traditionally excluded from genomic analytics. On the other hand, the ability for institutions to contribute their datasets through secure, institution-controlled nodes, rather than transferring individual-level genomic data externally, ensures alignment with strict data governance requirements and reduces institutional barriers to adoption. This approach lowers operational friction and makes it feasible for hospitals with diverse policies and technical capabilities to engage in multisite federated studies. More broadly, a federated approach helps preserve data privacy, ensuring compliance with European and American regulations [7,11,13] and the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule [12]. It also responds to the challenges posed by data sovereignty laws, which restrict the movement of sensitive data across regional or national borders. However, for these projects to succeed, it is essential to enable cross-border collaboration.

Enabling Collaboration in Federated Infrastructures

For designers and developers aiming to build federated infrastructures, it is crucial to prioritize features that support cross-border collaboration. Such collaboration requires coordination among researchers, legal experts, and technical professionals to develop communities that balance privacy with data accessibility [2,4,32]. Importantly, collaboration in this context hinges not only on technical capability but also on trust and shared expectations. Institutions must have confidence that their data will be handled responsibly and that they will benefit equitably from participation in shared projects, such as developing federated infrastructures. In this regard, although technical infrastructures can facilitate collaboration, the literature also highlights the importance of social and organizational conditions for making such collaboration possible [2,4,32]. For instance, federated infrastructures rely on institutions aligning their data models and colabeling datasets so that shared analyses become feasible [2,29,71]. Conversely, the establishment of a shared infrastructure can itself promote collaboration by providing a common environment, vocabulary, and set of tools through which partners can coordinate their work. This reciprocal relationship, where collaboration is both a prerequisite for and a product of shared infrastructures, extends beyond the definition of functional and nonfunctional requirements. Nevertheless, certain nonfunctional requirements, such as communication features and documentation, can play a fundamental role in the collaboration process. For instance, Holtgrewe et al [48] note that providing tutorials lowers the barrier for new users or institutions to join and contribute to the ecosystem. Similarly, Gill et al [72] highlight that tutorials and protocols help users approach the platform with a shared understanding, while also emphasizing the relevance of responding to user feedback to improve the usability and reciprocity in the collaborations. These examples illustrate how communication and documentation are important to support the collaboration. Yet, despite their importance, we found relatively few examples in literature that explicitly address these requirements. To fill this gap, future genomic data management platforms should incorporate features, such as documentation explaining user roles and responsibilities, real-time notifications that keep participants informed of data access and usage, and feedback mechanisms that allow users to report concerns and suggest improvements. Together, these elements would enhance usability, promote collaboration, and encourage broader adoption.

Limitations

While this scoping review offers a range of requirements for designers and developers seeking to create an integrated digital ecosystem for genomic and health care data, several limitations should be acknowledged.

First, the review may be affected by publication bias. Studies from regions with well-established digital infrastructure are more likely to be published, whereas experiences from regions with limited infrastructure may be underrepresented. This imbalance can lead to an overemphasis on certain challenges, such as inclusive and robust analysis, while other critical issues, such as financial barriers to implementation, receive less

XSL•FO
RenderX

attention. Consequently, the findings may reflect a skewed perspective on what constitutes effective genomic data management, and the applicability of certain solutions may vary depending on local infrastructure and resources.

Second, the search strategy prioritized user-centered design and usability to capture end users' perspectives on their needs as reported in the literature. This focus enabled the mapping of functional requirements, meaning what users want from these systems, although it offers less insight into methods for implementing such requirements. Although technical papers are referenced, an in-depth discussion of implementation approaches falls outside the scope of this study. Future reviews could target literature on nonfunctional requirements, such as high-performance computing optimization and cloud-native genomic workflows, to better capture the architectural and computational foundations that support scalable genomic analysis. Such work would complement the user-centered perspective presented here by illustrating how various workflows and pipelines are managed at the infrastructure level.

Third, this review focused exclusively on English-language articles, which may have excluded relevant work published in other languages. Additionally, while the inclusion of peer-reviewed articles and gray literature ensures scientific rigor, valuable insights from nonacademic sources, such as industry reports or early-stage studies not yet appearing in peer-reviewed journals, may have been overlooked.

To address these limitations, future stages of this research will adopt a Delphi approach to involve key stakeholders in reviewing and achieving consensus on the findings presented here [101]. This process will be more inclusive, engaging international experts from diverse countries alongside industry professionals and patient advocates, thereby incorporating perspectives that this review could not fully capture.

Conclusion

The key functional and nonfunctional requirements identified in this work can potentially form the starting point for defining a core set of functionalities (standards) for genomic data management interfaces worldwide. While there is a clear need to define ways to aggregate and protect important personal health data from inappropriate access and use, there is also a tension with the fact that other multiple parties need or want to access such data for different purposes. The complexity of such a sociotechnical system is further increased by the need to enable and maximize the control of the data owners (the patients) over the use of such information. While we leave the definitions of when it is appropriate to access the data, by whom, and for what purposes to the ethics and legal experts, from the human factors and psychological point of view, it is essential to identify what is acceptable, what is considered trustworthy by the stakeholders, and how to communicate to the patients (the owners of the data) and to all other stakeholders their rights, their duties, and their options for action. We suggest that advancements in genomic data management systems should focus on three primary functional capabilities: (1) general data management, including data acquisition, mechanisms for data sharing, and standardization; (2) the preprocessing and analysis of the data; and (3) along with the visualization and reporting of insights derived from that analysis. These functional elements must be supported by nonfunctional requisites, including secure infrastructure, compliance with legal and ethical standards, and structured governance to facilitate controlled and trustworthy data sharing.

We hope that this work will feed a virtuous circle that will rapidly lead to a human-centered design of health data platforms that are able to address different users and their needs, while also addressing key challenges, such as privacy and technical scalability.

Acknowledgments

We would like to thank the members of the PROTECT-CHILD consortium for their contributions, insights, and collaboration throughout this project.

Funding

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement 101137423.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR checklist.
[DOCX File , 27 KB - jmir_v28i1e78405_app1.docx]

Multimedia Appendix 2
Criteria used to select the articles.
[DOCX File , 15 KB - jmir_v28i1e78405_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Functional requirements.
[DOCX File , 333 KB - jmir_v28i1e78405_app3.docx]

Multimedia Appendix 4
Nonfunctional requirements.
[DOCX File , 17 KB - jmir_v28i1e78405_app4.docx]

References

1. Alzu'bi A, Zhou L, Watzlaf V. Personal genomic information management and personalized medicine: challenges, current solutions, and roles of HIM professionals. Perspect Health Inf Manag 2014;11(Spring):1c. [Medline: 24808804]
2. McLeod C, Gout AM, Zhou X, Thrasher A, Rahbarinia D, Brady SW, et al. Jude cloud: a pediatric cancer genomic data-sharing ecosystem. Cancer Discov 2021;11(5):1082-1099. [doi: 10.1101/2020.08.24.264614]
3. Li Y, Van Den Berg EH, Kurilshikov A, Zhernakova D, Gacesa R, Hu S, Lifelines Cohort Study, et al. Genome-wide studies reveal genetic risk factors for hepatic fat content. Genomics Proteomics Bioinformatics 2024;22(2):qzae031 [FREE Full text] [doi: 10.1093/gpbjnl/qzae031] [Medline: 39142818]
4. Reska D, Czajkowski M, Jurczuk K, Boldak C, Kwedlo W, Bauer W, et al. Integration of solutions and services for multi-omics data analysis towards personalized medicine. Biocybernetics and Biomedical Engineering 2021;41(4):1646-1663. [doi: 10.1016/j.bbe.2021.10.005]
5. Dutta D, Chatterjee N. Expanding scope of genetic studies in the era of biobanks. Hum Mol Genet 2025:ddaf054. [doi: 10.1093/hmg/ddaf054] [Medline: 40312842]
6. Gallagher CS, Ginsburg GS, Musick A. Biobanking with genetics shapes precision medicine and global health. Nat Rev Genet 2025;26(3):191-202. [doi: 10.1038/s41576-024-00794-y] [Medline: 39567741]
7. General Data Protection Regulation (GDPR). 2016/679 2016. European Parliament, European Council. 2016. URL: https://gdpr-info.eu/ [accessed 2023-08-02]
8. Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act). European Union. 2023. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32023R2854 [accessed 2022-08-02]
9. Proposal for a regulation of the European Parliament and of the Council on the European Health Data Space. European Parliament. 2022. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0197 [accessed 2024-07-02]
10. European Commission. Global health. EU Glob Health Strategy. URL: https://health.ec.europa.eu/internationalcooperation/global-health_en [accessed 2025-01-08]
11. Proposal for a regulation of the European parliament and of the council on the European health data space regulation (EU). European Parliament, European Council. 2022. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0197 [accessed 2024-07-02]
12. HIPAA privacy rule to support reproductive health care privacy. Office for Civil Rights (OCR), Office of the Secretary, Department of Health and Human Services. 2024. URL: https://www.govinfo.gov/content/pkg/FR-2024-04-26/pdf/2024-08503.pdf [accessed 2025-01-08]
13. Marcus JS, Martens B, Carugati C, Bucher A, Godlovitch I. The European health data space. SSRN Journal 2022. [doi: 10.2139/ssrn.4300393]
14. Horgan D, Hajduch M, Vrana M, Soderberg J, Hughes N, Omar MI, et al. European health data space—an opportunity now to grasp the future of data-driven healthcare. Healthcare 2022;10(9):1629. [doi: 10.3390/healthcare10091629]
15. Rehm HL, Page AJ, Smith L, Adams JB, Alterovitz G, Babb LJ, et al. GA4GH: International policies and standards for data sharing across genomic research and healthcare. Cell Genom 2021;1(2):100029 [FREE Full text] [doi: 10.1016/j.xgen.2021.100029] [Medline: 35072136]
16. Bacall F, Apaolaza A, Andrabi M, Child C, Goble C, Sand O, et al. Making bioinformatics training events and material more discoverable using teSS, the ELIXIR training portal. Curr Protoc 2023;3(2):e682. [doi: 10.1002/cpz1.682] [Medline: 36809564]
17. Schmitt T, Poirel HA, Cauët E, Delnord M, Van den Bulcke M. Unlocking the genomic landscape: Results of the beyond 1 million genomes (B1MG) pilot in belgium towards genomic data infrastructure (GDI). Health Policy 2024;143:105060. [doi: 10.1016/j.healthpol.2024.105060] [Medline: 38598968]
18. Raab R, Küderle A, Zakreuskaya A, Stern AD, Klucken J, Kaissis G, et al. Federated electronic health records for the European Health Data Space. Lancet Digit Health 2023;5(11):e840-e847 [FREE Full text] [doi: 10.1016/S2589-7500(23)00156-5] [Medline: 37741765]
19. Ochoa D, Hercules A, Carmona M, Suveges D, Baker J, Malangone C, et al. The next-generation open targets platform: reimagined, redesigned, rebuilt. Nucleic Acids Res 2023;51(D1):D1353-D1359 [FREE Full text] [doi: 10.1093/nar/gkac1046] [Medline: 36399499]

XSL•FO
RenderX

20. Bartlett G, Gagnon J. Physicians and knowledge translation of statistics: Mind the gap. CMAJ 2016;188(1):11-12 [FREE Full text] [doi: 10.1503/cmaj.151254] [Medline: 26621998]
21. Bashira MS. Assessment of basic and advanced knowledge in biostatistics and clinical research among health care professionals at King Fahd medical city, Riyadh. J Biometr Biostats 2019;1:5 [FREE Full text]
22. Lau JW, Lehnert E, Sethi A, Malhotra R, Kaushik G, Onder Z, Seven Bridges CGC Team. The cancer genomics cloud: Collaborative, reproducible, and democratized-a new paradigm in large-scale computational research. Cancer Res 2017;77(21):e3-e6 [FREE Full text] [doi: 10.1158/0008-5472.CAN-17-0387] [Medline: 29092927]
23. Lightbody G, Haberland V, Browne F, Taggart L, Zheng H, Parkes E, et al. Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. Brief Bioinform 2019;20(5):1795-1811 [FREE Full text] [doi: 10.1093/bib/bby051] [Medline: 30084865]
24. Sommerville I. Software engineering. Tenth edition. Boston, US: Pearson; 2016.
25. Rawashdeh A, Matalkah B. A new software quality model for evaluating COTS components. J Comp Sci 2006;2(4):373-381. [doi: 10.3844/jcssp.2006.373.381]
26. Glinz M. On non-functional requirements. 2007 Presented at: 15th IEEE Int Requir Eng Conf RE; October 15; Delhi p. 21-26. [doi: 10.1109/re.2007.45]
27. Taraborelli D. Feature binding and object perception. Does object awareness require feature conjunction? European Society for Philosophy and Psychology 2003:1 [FREE Full text]
28. Cleland-Huang J, Settimi R, Zou X, Solc P. The detection and classification of non-functional requirements with application to early aspects. 2006 Presented at: 14th IEEE Int Requir Eng Conf RE06; Sep 11; United States p. 39-48. [doi: 10.1109/re.2006.65]
29. Davis-Turak J, Courtney SM, Hazard ES, Glen WB, da Silveira WA, Wesselman T, et al. Genomics pipelines and data integration: challenges and opportunities in the research setting. Expert Rev Mol Diagn 2017;17(3):225-237 [FREE Full text] [doi: 10.1080/14737159.2017.1282822] [Medline: 28092471]
30. Kho AN, Rasmussen LV, Connolly JJ, Peissig PL, Starren J, Hakonarson H, et al. Practical challenges in integrating genomic data into the electronic health record. Genet Med 2013;15(10):772-778 [FREE Full text] [doi: 10.1038/gim.2013.131] [Medline: 24071798]
31. Pang Z, Chong J, Zhou G, De Lima Morais DA, Chang L, Barrette M, et al. MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. In: Nucleic Acids Res. Oxford, UK: Nucleic Acids Res Oxford University Press; 2021:W388-W396.
32. Chen IMA, Markowitz VM, Palaniappan K, Szeto E, Chu K, Huang J, et al. Supporting community annotation and user collaboration in the integrated microbial genomes (IMG) system. BMC Genomics 2016;17(1):307 [FREE Full text] [doi: 10.1186/s12864-016-2629-y] [Medline: 27118214]
33. Rambla J, Baudis M, Ariosa R, Beck T, Fromont LA, Navarro A, et al. Beacon v2 and beacon networks: A "lingua franca" for federated data discovery in biomedical genomics, and beyond. Hum Mutat 2022;43(6):791-799 [FREE Full text] [doi: 10.1002/humu.24369] [Medline: 35297548]
34. Rueda M, Ariosa R, Moldes M, Rambla J. Beacon v2 reference implementation: a toolkit to enable federated sharing of genomic and phenotypic data. Bioinformatics 2022;38(19):4656-4657. [doi: 10.1093/bioinformatics/btac568] [Medline: 35980167]
35. Pavia M, Chede A, Wu Z, Cadillo-Quiroz H, Zhu Q. BinaRena: a dedicated interactive platform for human-guided exploration and binning of metagenomes. Microbiome 2023;11(1):186 [FREE Full text] [doi: 10.1186/s40168-023-01625-8] [Medline: 37596696]
36. Ergonomics of human-system interaction? part 11: usability: definitions and concepts (ISO 9241-11). International Organization for Standardization. 1998. URL: https://www.iso.org/standard/63500.html [accessed 2026-03-14]
37. Ergonomics of human-system interaction? Part 210: human-centred design for interactive systems (ISO 9241-210). International Organization for Standardization. 2010. URL: https://www.iso.org/standard/77520.html [accessed 2026-03-14]
38. Medical devices? Application of usability engineering to medical devices (ISO 62366). International Organization for Standardization. 2015. URL: https://www.iso.org/standard/72704.html [accessed 2026-03-14]
39. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
40. Piškur B, Beurskens AJ, Jongmans MJ, Ketelaar M, Norton M, Frings CA, et al. Parents' actions, challenges, and needs while enabling participation of children with a physical disability: A scoping review. BMC Pediatr 2012;12(1):177 [FREE Full text] [doi: 10.1186/1471-2431-12-177] [Medline: 23137074]
41. PRISMA. URL: https://www.prisma-statement.org [accessed 2025-01-08]
42. t. OSF. URL: https://osf.io/p5bxr [accessed 2025-11-28]
43. Xia M, Liu CJ, Zhang Q, Guo AY. GEDS: A gene expression display server for mRNAs, miRNAs and proteins. Cells MDPI 2019;8(7):675. [doi: 10.3390/cells8070675] [Medline: 31277321]

XSL•FO
RenderX

44. Ding J, Blencowe M, Nghiem T, Ha SM, Chen YW, Li G, et al. Mergeomics 2.0: a web server for multi-omics data integration to elucidate disease networks and predict therapeutics. Nucleic Acids Res 2021 Jul 02;49(W1):W375-W387 [FREE Full text] [doi: 10.1093/nar/gkab405] [Medline: 34048577]
45. Hombach D, Schuelke M, Knierim E, Ehmke N, Schwarz J, Fischer-Zirnsak B, et al. MutationDistiller: User-driven identification of pathogenic DNA variants. Nucleic Acids Res 2019;47(W1):W114-W120 [FREE Full text] [doi: 10.1093/nar/gkz330] [Medline: 31106342]
46. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. G:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res 2019;47(W1):W191-W198 [FREE Full text] [doi: 10.1093/nar/gkz369] [Medline: 31066453]
47. Rodchenkov I, Babur O, Luna A, Aksoy BA, Wong JV, Fong D, et al. Pathway commons 2019 update: Integration, analysis and exploration of pathway data. Nucleic Acids Res 2020;48(D1):D489-D497 [FREE Full text] [doi: 10.1093/nar/gkz946] [Medline: 31647099]
48. Holtgrewe M, Stolpe O, Nieminen M, Mundlos S, Knaus A, Kornak U, et al. VarFish: Comprehensive DNA variant analysis for diagnostics and research. Nucleic Acids Res 2020;48(W1):W162-W169 [FREE Full text] [doi: 10.1093/nar/gkaa241] [Medline: 32338743]
49. Ochoa D, Hercules A, Carmona M, Suveges D, Gonzalez-Uriarte A, Malangone C, et al. Open Targets Platform: Supporting systematic drug-target identification and prioritisation. Nucleic Acids Res 2021;49(D1):D1302-D1310 [FREE Full text] [doi: 10.1093/nar/gkaa1027] [Medline: 33196847]
50. Suciu RM, Aydin E, Chen BE. GeneDig: A web application for accessing genomic and bioinformatics knowledge. BMC Bioinformatics 2015;16(1):67 [FREE Full text] [doi: 10.1186/s12859-015-0497-0] [Medline: 25888373]
51. Calabria A, Spinozzi G, Benedicenti F, Tenderini E, Montini E. adLIMS: A customized open source software that allows bridging clinical and basic molecular research studies. BMC Bioinformatics 2015;16(S9):S5. [doi: 10.1186/1471-2105-16-s9-s5]
52. Bhuvaneshwar K, Belouali A, Singh V, Johnson RM, Song L, Alaoui A, et al. G-DOC Plus - an integrative bioinformatics platform for precision medicine. BMC Bioinformatics 2016;17(1):193 [FREE Full text] [doi: 10.1186/s12859-016-1010-0] [Medline: 27130330]
53. Nanni L, Pinoli P, Canakoglu A, Ceri S. PyGMQL: Scalable data extraction and analysis for heterogeneous genomic datasets. BMC Bioinformatics 2019;20(1):560 [FREE Full text] [doi: 10.1186/s12859-019-3159-9] [Medline: 31703553]
54. Yousif A, Drou N, Rowe J, Khalfan M, Gunsalus KC. NASQAR: A web-based platform for high-throughput sequencing data analysis and visualization. BMC Bioinformatics 2020;21(1):267 [FREE Full text] [doi: 10.1186/s12859-020-03577-4] [Medline: 32600310]
55. Yukselen O, Turkyilmaz O, Ozturk AR, Garber M, Kucukural A. DolphinNext: A distributed data processing platform for high throughput genomics. BMC Genomics 2020;21(1):310. [doi: 10.1186/s12864-020-6714-x]
56. Demchak B, Hull T, Reich M, Liefeld T, Smoot M, Ideker T, et al. Cytoscape: the network visualization tool for GenomeSpace workflows. F1000Research Faculty of 1000 Ltd 2014;3:151 [FREE Full text] [doi: 10.12688/f1000research.4492.2] [Medline: 25165537]
57. Sauria MEG, Phillips-Cremins JE, Corces VG, Taylor J. HiFive: A tool suite for easy and efficient HiC and 5C data analysis. Genome Biol 2015;16(1):237 [FREE Full text] [doi: 10.1186/s13059-015-0806-y] [Medline: 26498826]
58. Danahey K, Borden BA, Furner B, Yukman P, Hussain S, Saner D, et al. Simplifying the use of pharmacogenomics in clinical practice: Building the genomic prescribing system. J Biomed Inform 2017;75:110-121 [FREE Full text] [doi: 10.1016/j.jbi.2017.09.012] [Medline: 28963061]
59. Crawford KM, Gallego-Fabrega C, Kourkoulis C, Miyares L, Marini S, Flannick J, et al. Cerebrovascular disease knowledge portal. Stroke 2018;49(2):470-475. [doi: 10.1161/strokeaha.117.018922]
60. Warner JL, Prasad I, Bennett M, Arniella M, Beeghly-Fadiel A, Mandl KD, et al. SMART cancer navigator: A framework for implementing ASCO workshop recommendations to enable precision cancer medicine. JCO Precis Oncol 2018(2):1-14 [FREE Full text] [doi: 10.1200/PO.17.00292] [Medline: 30238071]
61. Pearce TM, Nikiforova MN, Roy S. Interactive browser-based genomics data visualization tools for translational and clinical laboratory applications. J Mol Diagn 2019;21(6):985-993 [FREE Full text] [doi: 10.1016/j.jmoldx.2019.06.005] [Medline: 31382034]
62. Bonomi L, Huang Y, Ohno-Machado L. Privacy challenges and research opportunities for genomic data sharing. Nat Genet 2020;52(7):646-654 [FREE Full text] [doi: 10.1038/s41588-020-0651-0] [Medline: 32601475]
63. Ma C, Sridharan M, Al-Sayegh H, Li A, Guo D, Auclair M, et al. London WB. Building a harmonized datamart by integrating cross-institutional systems of clinical, outcome, and genomic data: The pediatric patient informatics platform (PPIP). JCO Clin Cancer Inform Wolters Kluwer 2021;5:202-215 [FREE Full text] [doi: 10.2196/mhealth.3861] [Medline: 25830810]
64. Campbell EM, Boyles A, Shankar A, Kim J, Knyazev S, Cintron R, et al. MicrobeTrace: Retooling molecular epidemiology for rapid public health response. PLoS Comput Biol 2021;17(9):e1009300 [FREE Full text] [doi: 10.1371/journal.pcbi.1009300] [Medline: 34492010]

XSL•FO
RenderX

65. Raveendran K, Freese N, Kintali C, Tiwari S, Bole P, Dias C, et al. BioViz : Web application linking cyVerse cloud resources to genomic visualization in the integrated genome browser. Front Bioinform 2022;2:764619 [FREE Full text] [doi: 10.3389/fbinf.2022.764619] [Medline: 36304269]
66. Reiff SB, Schroeder AJ, Kırlı K, Cosolo A, Bakker C, Mercado L, et al. The 4D nucleome data portal as a resource for searching and visualizing curated nucleomics data. Nat Commun 2022;13(1):2365 [FREE Full text] [doi: 10.1038/s41467-022-29697-4] [Medline: 35501320]
67. Post AR, Ho N, Rasmussen E, Post I, Cho A, Hofer J, et al. Hypermedia-based software architecture enables Test-Driven Development. JAMIA Open 2023;6(4):ooad089 [FREE Full text] [doi: 10.1093/jamiaopen/ooad089] [Medline: 37860604]
68. Xia J, Benner MJ, Hancock REW. NetworkAnalyst. Integrative approaches for protein-protein interaction network analysis and visual exploration. Nucleic Acids Res Oxford University Press 2014;42(W1):W167-W174. [doi: 10.1093/nar/gku443]
69. Albuquerque MA, Grande B, Ritch EJ, Pararajalingam P, Jessa S, Krzywinski M, et al. Enhancing knowledge discovery from cancer genomics data with Galaxy. Gigascience 2017;6(5):1-13 [FREE Full text] [doi: 10.1093/gigascience/gix015] [Medline: 28327945]
70. Das S, Lecours Boucher X, Rogers C, Makowski C, Chouinard-Decorte F, Oros Klein K, et al. Integration of data and phenotypic data within a unified extensible multimodal framework. Front Neuroinform 2018;12:91 [FREE Full text] [doi: 10.3389/fninf.2018.00091] [Medline: 30631270]
71. Osmond M, Hartley T, Dyment DA, Kernohan KD, Brudno M, Buske OJ, et al. Outcome of over 1500 matches through the matchmaker exchange for rare disease gene discovery: The 2-year experience of Care4Rare Canada. Genet Med 2022;24(1):100-108 [FREE Full text] [doi: 10.1016/j.gim.2021.08.014] [Medline: 34906465]
72. Gill IS, Griffiths EJ, Dooley D, Cameron R, Sarah KS, John N, et al. The dataHarmonizer: A tool for faster data harmonization, validation, aggregation and analysis of pathogen genomics contextual information. Microb Genomics Microbiology Society 2023;9(1):000908. [doi: 10.1099/mgen.0.000908]
73. Wolf B, Kuonen P, Dandekar T, Atlan D. DNAseq workflow in a diagnostic context and an example of a user friendly implementation. Biomed Res Int 2015;2015:403497 [FREE Full text] [doi: 10.1155/2015/403497] [Medline: 26137478]
74. Mohr C, Friedrich A, Wojnar D, Kenar E, Polatkan AC, Codrea MC, et al. qPortal: A platform for data-driven biomedical research. PLoS One 2018;13(1):e0191603 [FREE Full text] [doi: 10.1371/journal.pone.0191603] [Medline: 29352322]
75. Wünsch C, Banck H, Müller-Tidow C, Dugas M. AMLVaran: A software approach to implement variant analysis of targeted NGS sequencing data in an oncological care setting. BMC Med Genomics 2020;13(1):17 [FREE Full text] [doi: 10.1186/s12920-020-0668-3] [Medline: 32019565]
76. Cappelli E, Cumbo F, Bernasconi A, Canakoglu A, Ceri S, Masseroli M, et al. OpenGDC: Unifying, modeling, integrating cancer genomic data and clinical metadata. Applied Sciences 2020;10(18):6367. [doi: 10.3390/app10186367]
77. Canakoglu A, Bernasconi A, Colombo A, Masseroli M, Ceri S. GenoSurf: Metadata driven semantic search system for integrated genomic datasets. Database 2019;1:baz132. [doi: 10.1093/database/baz132]
78. Murtagh MJ, Turner A, Minion JT, Fay M, Burton PR. International data sharing in practice: new technologies meet old governance. Biopreserv Biobank 2016;14(3):231-240 [FREE Full text] [doi: 10.1089/bio.2016.0002] [Medline: 27200470]
79. Gilbert RM, Sumodhee D, Pontikos N, Hollyhead C, Patrick A, Scarles S, et al. Collaborative research and development of a novel, patient-centered digital platform (MyEyeSite) for rare inherited retinal disease data: Acceptability and feasibility study. JMIR Form Res 2022;6(1):e21341 [FREE Full text] [doi: 10.2196/21341] [Medline: 35099396]
80. Sante T, Vergult S, Volders PJ, Kloosterman WP, Trooskens G, De Preter K, et al. ViVar: A comprehensive platform for the analysis and visualization of structural genomic variation. PLoS ONE 2014;9(12):e113800. [doi: 10.1371/journal.pone.0113800]
81. Verbruggen S, Menschaert G. mQC: A post-mapping data exploration tool for ribosome profiling. Comput Methods Programs Biomed 2019;181:104806 [FREE Full text] [doi: 10.1016/j.cmpb.2018.10.018] [Medline: 30401579]
82. Li L, An Y, Ma L, Yang M, Yuan P, Liu X, et al. Msuite2: All-in-one DNA methylation data analysis toolkit with enhanced usability and performance. Comput Struct Biotechnol J 2022;20:1271-1276 [FREE Full text] [doi: 10.1016/j.csbj.2022.03.005] [Medline: 35317233]
83. Amer-Yahia S, Koutrika G, Braschler M, Calvanese D, Lanti D, Lücke-Tieke H, et al. INODE: Building an end-to-end data exploration system in practice. ACM SIGMOD Rec 2022;50(4):23-29. [doi: 10.1145/3516431.3516436]
84. Kounelis F, Kanterakis A, Kanavos A, Pandi MT, Kordou Z, Manusama O, et al. Documentation of clinically relevant genomic biomarker allele frequencies in the next-generation FINDbase worldwide database. Hum Mutat 2020;41(6):1112-1122. [doi: 10.1002/humu.24018] [Medline: 32248568]
85. Ware AP, Satyamoorthy K, Paul B. CmirC update 2024: A multi-omics database for clustered miRNAs. Funct Integr Genomics 2024;24(4):133. [doi: 10.1007/s10142-024-01410-2] [Medline: 39085735]
86. Takai-Igarashi T, Kinoshita K, Nagasaki M, Ogishima S, Nakamura N, Nagase S, et al. Security controls in an integrated Biobank to protect privacy in data sharing: Rationale and study design. BMC Med Inform Decis Mak 2017;17(1):100 [FREE Full text] [doi: 10.1186/s12911-017-0494-5] [Medline: 28683736]
87. Ullah S, Ullah F, Rahman W, Karras DA, Ullah A, Ahmad G, et al. The Cancer Research Database (CRDB): Integrated platform to gain statistical insight into the correlation between cancer and COVID-19. JMIR Cancer 2022;8(2):e35020 [FREE Full text] [doi: 10.2196/35020] [Medline: 35430561]

XSL•FO
RenderX

88. Kuzmenkov AY, Trushin IV, Vinogradova AG, Avramenko AA, Sukhorukova MV, Malhotra-Kumar S, et al. AMRmap: An interactive web platform for analysis of antimicrobial resistance surveillance data in Russia. Front Microbiol 2021;12:620002 [FREE Full text] [doi: 10.3389/fmicb.2021.620002] [Medline: 33776956]
89. Sullivan DE, Gabbard Jr. JL, Shukla M, Sobral B. Data integration for dynamic and sustainable systems biology resources: Challenges and lessons learned. Chem Biodivers 2010;7(5):1124-1141 [FREE Full text] [doi: 10.1002/cbdv.200900317] [Medline: 20491070]
90. Martínez-García M, Hernández-Lemus E. Data integration challenges for machine learning in precision medicine. Front Med (Lausanne) 2021;8:784455 [FREE Full text] [doi: 10.3389/fmed.2021.784455] [Medline: 35145977]
91. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012 May;2(5):401-404 [FREE Full text] [doi: 10.1158/2159-8290.CD-12-0095] [Medline: 22588877]
92. Konopik J, Blunck D. Development of an evidence-based conceptual model of the health care sector under digital transformation: Integrative review. J Med Internet Res 2023;25:e41512 [FREE Full text] [doi: 10.2196/41512] [Medline: 37289482]
93. Schmidt AE, Bobek J, Mathis-Edenhofer S, Schwarz T, Bachner F. Cross-border healthcare collaborations in Europe (2007–2017): Moving towards a European Health Union? Health Policy 2022;126(12):1241-1247. [doi: 10.1016/j.healthpol.2022.10.011]
94. Cascini F, Pantovic A, Al-Ajlouni YA, Puleo V, De Maio L, Ricciardi W. Health data sharing attitudes towards primary and secondary use of data: A systematic review. EClinicalMedicine 2024;71:102551 [FREE Full text] [doi: 10.1016/j.eclinm.2024.102551] [Medline: 38533128]
95. Duong BQ, Arwood MJ, Hicks JK, Beitelshees AL, Franchi F, Houder JT, et al. Development of customizable implementation guides to support clinical adoption of pharmacogenomics: Experiences of the Implementing GeNomics In pracTicE (IGNITE) Network. PGPM 2020;Volume 13:217-226. [doi: 10.2147/pgpm.s241599]
96. Almeida JR, Oliveira JL. MONTRA2: A web platform for profiling distributed databases in the health domain. Informatics in Medicine Unlocked 2024;45:101447. [doi: 10.1016/j.imu.2024.101447]
97. Melles M, Albayrak A, Goossens R. Innovating health care: Key characteristics of human-centered design. Int J Qual Health Care 2021;33(Supplement_1):37-44 [FREE Full text] [doi: 10.1093/intqhc/mzaa127] [Medline: 33068104]
98. Bisquert A, Hmimou A, Berral J, Gutierrez-Torre A, Romero O. HealthMesh: An architectural framework for federated healthcare data management. 2024 Presented at: 26th Int Workshop Des Optim Lang Anal Process Big Data; 2024 Mar 25; Paestum, Italy.
99. Rujano MA, Boiten JW, Ohmann C, Canham S, Contrino S, David R, et al. Sharing sensitive data in life sciences: An overview of centralized and federated approaches. Brief Bioinform 2024;25(4):bbae262 [FREE Full text] [doi: 10.1093/bib/bbae262] [Medline: 38836701]
100. Cui Z, Badam SK, Yalçin MA, Elmqvist N. DataSite: Proactive visual data exploration with computation of insight-based recommendations. Information Visualization 2018;18(2):251-267. [doi: 10.1177/1473871618806555]
101. Wilson YA, Smithers‐Sheedy H, Ostojic K, Waight E, Kruer MC, Fahey MC, et al. Common data elements to standardize genomics studies in cerebral palsy. Develop Med Child Neuro 2022;64(12):1470-1476. [doi: 10.1111/dmcn.15245]

Abbreviations

AI: artificial intelligence
API: Application Programming Interface
EHR: electronic health record
GA4GH: Global Alliance for Genomics and Health
GDI: Genomic Data Infrastructure
GDPR: General Data Protection Regulation
HIPAA: Health Insurance Portability and Accountability Act
OSF: Open Science Framework
PRISMA-ScR: Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Review
TEHDAS: Towards a European Health Data Space
UI: user interface
UX: user experience

XSL•FO
RenderX

Edited by A Schwartz; submitted 02.Jun.2025; peer-reviewed by J Du, M Taylor; comments to author 12.Nov.2025; accepted 06.Jan.2026; published 27.Apr.2026.

Please cite as:
Resendez V, Yıldırım F, Gaeta E, Fico G, Borsci S
Toward a Common Set of Interface Requirements for Genomic Data Management: Scoping Review
J Med Internet Res 2026;28:e78405
URL: https://www.jmir.org/2026/1/e78405
doi:10.2196/78405
PMID:42043847

©Valeria Resendez, Funda Yıldırım, Eugenio Gaeta, Giuseppe Fico, Simone Borsci. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Media Exposure and Its Association With Vaccine Attitudes, Intentions, and Hesitancy: Systematic Review

Francesco Leonforte[1*], MD; Vito Nicosia[2*], MD; Paola Comite[2], MD, PhD; Giustino Morlino[3], MD; Antonio Mistretta[2,4], MD, PhD

[1]Department of Integrated Hygiene, Organizational, and Service Activities (Structural Department), Health Management, University Hospital Polyclinic "G. Rodolico—San Marco", Catania, Italy

[2]Department of Medical, Surgical Sciences and Advanced Technologies "G.F. Ingrassia", University of Catania, Catania, CT, Italy

[3]Accreditation and supply network Area, Local Health Authority "Roma 1", Rome, Italy

[4]Scientific Communication Service, National Institute of Public Health (Istituto Superiore di Sanità), Rome, Italy

[*]these authors contributed equally

Corresponding Author:
Vito Nicosia, MD
Department of Medical, Surgical Sciences and Advanced Technologies "G.F. Ingrassia"
University of Catania
Via Santa Sofia
Catania, CT, 87 - 95123
Italy
Phone: 39 3479244709
Email: vitonicosia6@gmail.com

Abstract

Background: Vaccine hesitancy, amplified by the COVID-19 "infodemic," has emerged as a pressing public health challenge. Communication strategies are pivotal for enhancing vaccine literacy, countering misinformation, and sustaining immunization programs.

Objective: This systematic review evaluates the association between communication channels and vaccine hesitancy and adherence, while examining the moderating role of sociodemographic factors.

Methods: A PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)–compliant search was conducted across PubMed, Scopus, and Web of Science, yielding 17,407 records screened according to predefined eligibility criteria (peer-reviewed studies with N>1000 adults assessing communication media targeting vaccine hesitancy and adherence, excluding pediatric, health care–specific, and non-English research). After full-text assessment, studies were appraised using the Modified Medical Education Research Study Quality Instrument for methodological quality and the Joanna Briggs Institute Critical Appraisal Checklist, ROBIN-I (Risk of Bias in Nonrandomized Studies of Interventions), or RoB 2.0 (Risk-of-Bias 2.0 tool for randomized trials) tools for risk of bias.

Results: Thirty-six studies were included (26 cross-sectional, 4 quasi-experimental, 4 randomized controlled trials, 1 cohort, and 1 global analysis). Randomized and nonrandomized experimental studies demonstrated that tailored communication strategies delivered via radio, web platforms, and social media significantly improved vaccine acceptance. Adaptive public health campaigns achieved up to an 8% weekly increase in uptake in Madagascar (relative risk 1.08; 95% CI 1.01-1.15) and a 7.8% higher vaccination rate among Nigerian adults at first follow-up compared with controls. Digital and social media campaigns effectively reduced hesitancy and enhanced trust among hesitant pregnant women in the United States. Sociodemographic factors significantly moderated communication outcomes: a COVID-19 chatbot proved most effective among individuals with lower education and minority backgrounds, while religiosity (b=0.17; 95% CI 0.05-0.30; t810=2.80; P=.005) and cultural congruence (odds ratio 1.89; P<.01) influenced message credibility and engagement, respectively. The persuasive effect of online memes on COVID-19 vaccine intentions was not significantly influenced by gender (P=.83), age (P=.60), or political orientation (P=.44). Age-specific effects were observed, with greater responsiveness to a social media campaign among adults aged 25-34 years and reduced hesitancy among older groups. Multiple cross-sectional studies indicated higher adherence among audiences exposed to traditional media (television, radio, newspapers) and lower trust among social media users. Other studies suggested significant influences of gender, age, socioeconomic status, education level, and political orientation.

XSL•FO
RenderX

Conclusions: By synthesizing fragmented evidence, this review provides a systematic examination of the interplay between multichannel media and vaccine acceptance. It diverges from existing literature by integrating both traditional and digital media perspectives through the lens of sociodemographic moderation. This work offers a critical framework for public health interventions, advocating for rigorous longitudinal research to establish definitive causal links between communication and behavior. Consequently, these findings support a "precision" communication model, enabling the development of culturally congruent strategies tailored to specific recipient profiles to bolster vaccine adherence.

Trial Registration: PROSPERO CRD42025637441; https://tinyurl.com/4r9w83kw

(J Med Internet Res 2026;28:e74280) doi:10.2196/74280

KEYWORDS

vaccine literacy; vaccine hesitancy; vaccine adherence; communication; traditional media; digital media; sociodemographic factors; systematic review

Introduction

Communication in the context of vaccination is a global priority, essential to the success of vaccination campaigns and the safeguarding of public health [1]. Although vaccines are a key tool for preventing infectious diseases, vaccine hesitancy continues to undermine coverage and expose vulnerable populations to risk [2,3]. The COVID-19 pandemic has exacerbated these issues, highlighting the importance of targeted, effective communication strategies to rebuild trust in vaccines and promote acceptance [4].

The World Health Organization (WHO) defines vaccine hesitancy as a delay in acceptance or refusal of vaccination despite availability, driven by distrust, safety concerns, misinformation, and sociocultural influences [5]. By contrast, vaccine literacy is the ability to acquire, understand, and apply vaccine-related information; it is strengthened by transparent, empathetic, and participatory communication and is particularly effective for individuals who are uncertain or misinformed [6].

The COVID-19 "infodemic" has exemplified the massive spread of unverified claims, conspiracy theories, and misleading digital information that spreads faster than scientific evidence [7].

In this context, communication channels play a decisive role. It is possible to distinguish between traditional media (television, radio, and print) and digital/social media (internet-based platforms such as Facebook, Twitter, and Instagram). These channels differ substantially in communication dynamics, speed, audience targeting, and perceived credibility [8]. Traditional media generally operate through a 1-way flow of information, in which professional gatekeepers deliver content to a broad audience. Television and radio are used for public service announcements, informative spots, and educational programs, and are particularly useful in communities with limited internet access [9,10]. Print media, such as newspapers and magazines, offer analyses and contextualization of complex data, providing scientific and technical insights [11,12]. Social and digital media enable 2-way, interactive exchanges and promote user-generated content, offering immediacy and flexibility that are otherwise hindered by production and editorial constraints in traditional outlets [13,14]. These platforms also allow for highly targeted engagement and real-time monitoring of public sentiment, leveraging emotional appeals and establishing a personal connection with the audience [15]. However, they are also more vulnerable to misinformation and echo chambers. Although traditional media are often perceived as more credible and reliable, they can sometimes present negative or inaccurate messages, thereby increasing confusion and uncertainty among the public [16,17].

Despite their differences, the 2 systems are increasingly interdependent. Social platforms frequently amplify the reach of traditional outlets by directing audience attention, while traditional media often incorporate social media trends into their programming. This mutual influence has given rise to a hybrid, integrated, and more diversified media environment [18]. Vaccine communication is a widely debated topic in recent literature [19]. Numerous academic works have detailed the impact of various strategies aimed at enhancing vaccine literacy and counteracting hesitancy [20-23], with particular attention to sociodemographic variables [24,25]. Despite growing evidence, it remains crucial to address the lack of systematic reviews in this field.

This systematic review, therefore, aims to provide a deeper understanding of the role of communication means in promoting vaccine literacy and countering vaccine hesitancy. The specific objectives are as follows:

- Evaluate and compare the correlation between communication means and vaccine attitudes, hesitancy, and adherence.
- Investigate the role of sociodemographic factors.
- Identify gaps in the literature and future directions.

Methods

Overview

This systematic review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The completed PRISMA checklist is provided in Multimedia Appendix 1, ensuring transparency and reproducibility of the results [26]. The screening process was conducted in a double-blind manner in pairs (FL with VN and PC with GM); in cases of conflicting opinions, a third reviewer (AM) made the final decision. The protocol for this systematic review has been officially registered in the international PROSPERO (International Prospective Register of Systematic Reviews) database (CRD42025637441).

XSL•FO
RenderX

Identification of Articles

The search was conducted in 3 main databases (PubMed, Scopus, and Web of Science) using structured queries. The full search strategy is available in Multimedia Appendix 2. The key terms included the following:

- "vaccin*," "communication," "literacy," "hesitancy," and "acceptance"
- References to tools and platforms such as "social media," "chatbot," and "television"

The use of Boolean operators ("AND," "OR," and "NOT") allowed the combination of terms to maximize sensitivity and specificity.

The identified articles were downloaded and imported into the Rayyan software (Qatar Foundation) for screening. The search produced an initial total of 17,407 articles.

Application of Restrictions

Several restrictions were progressively applied to refine the selection (Textbox 1).

Textbox 1. Applied restrictions.

1. Temporal restriction (January 2014-October 2025)

   The selection focused on the last 10 years, including recent developments in technology and vaccination, especially after the COVID-19 pandemic.

2. Language restriction (English)

   Only articles published in English were included to ensure comparability and minimize interpretive ambiguity.

3. Exclusion of irrelevant types

   Reviews, letters, editorials, commentaries, and non–peer-reviewed articles were excluded.

4. Removal of duplicates.

Application of Inclusion/Exclusion Criteria

The inclusion criteria were defined based on 4 main elements according to the PICO (Population, Intervention/Exposure, Comparison, and Outcome) framework:

- Population: Studies on adults (≥18 years) with a sample size larger than 1000.
- Intervention/exposure: Strategies designed to improve vaccine attitudes, increase vaccination adherence, and reduce hesitancy.
- Comparison: Studies comparing different communication strategies or using control groups.
- Outcome: Indicators such as vaccine attitudes, vaccination adherence, or vaccine hesitancy.

Only peer-reviewed studies were considered, including surveys, ecological, case-control, cross-sectional, cohort, prevalence, quasi-experimental, and randomized controlled trials (RCTs).

The exclusion criteria were defined as follows:

- Studies not conducted on humans or conducted exclusively in pediatric populations or among health care professionals.
- Reviews, abstracts, editorials, letters, preprints, and other article types not covered by the inclusion criteria.
- Articles not in English.
- Articles that do not concern vaccines or strategies aimed at improving adherence or communication in the vaccination context.

The review process was carried out based on titles and abstracts.

Data Extraction

For the 99 articles, a rigorous full-text review process was subsequently conducted to confirm full compliance with the inclusion and exclusion criteria and, if confirmed, to proceed with the extraction of relevant information and compilation of a specifically prepared Microsoft Excel spreadsheet. The review and data extraction processes were performed by 2 reviewers (VN and FL) and subsequently checked by others (PC, GM, and AM). The information extracted included the following: title, authors, year of publication, country/context setting, methodology, objectives/purpose of the study, study population and sample size, vaccine, communication method and details, duration, main results, public health implications, and limitations.

Risk of Bias Assessment

The risk of bias was assessed using tools appropriate to the design of individual studies. For cross-sectional studies, the Joanna Briggs Institute Critical Appraisal Checklist for analytical cross-sectional studies was used [27]. It consists of 8 key criteria evaluating aspects such as participant selection, measurement validity, confounding factors, and statistical analysis. For RCTs, risk of bias was evaluated using the RoB 2.0 (Risk-of-Bias tool for randomized trials), developed by Cochrane, to appraise 5 domains: randomization, deviations from intended interventions, missing outcome data, measurement of outcomes, and selection of reported results. Each domain was rated using signaling questions, leading to an overall judgment of "low risk," "some concerns," or "high risk" of bias [28].

In nonrandomized experimental studies, risk of bias was evaluated using the ROBINS-I (Risk of Bias in Nonrandomized Studies of Interventions) tool, which assesses 7 domains: confounding, selection of participants, classification of interventions, deviations from intended interventions, missing data, outcome measurement, and selection of reported results. Judgments were made for each domain to determine an overall risk-of-bias rating: low, moderate, serious, critical, or no information [29].

Quality appraisal was performed by 2 researchers (VN and FL), and any discrepancies were resolved by a third (AM). Three

XSL•FO
RenderX

separate graphical representations of the risk-of-bias assessment were generated using the robvis tool [30].

Quality Assessment

Eligible studies were subjected to a rigorous evaluation of methodological quality using the Modified Medical Education Research Study Quality Instrument (M-MERSQI) tool, a quantitative assessment comprising 12 items across 7 domains, with a total possible score of 100 [31]. Studies were categorized into 3 quality levels: low quality (scores ranging from 0 to 35), medium quality (scores ranging from 36 to 65), and high quality (scores ranging from 66 to 100).

Data Analysis and Synthesis

Despite the presence of quantitative data in several included studies, a meta-analysis was deemed unfeasible due to substantial heterogeneity in study designs, populations, communication exposures, and outcome measures. Therefore, we conducted a narrative and tabular synthesis of the findings from the included studies, organized around key themes and study designs: the association between traditional and social media and vaccine adherence, and the role of sociodemographic factors. For each synthesis, studies were selected based on the type of communication intervention evaluated. Where available, regression coefficients, relative risks, or odds ratios with 95% CIs were tabulated and summarized through narrative synthesis. Outcomes not explicitly designated as primary or secondary for the purposes of this review were excluded from both the results table and the synthesis.

Results

Overview of the Included Studies

At the end of the full-text review process (Figure 1), a total of 36 studies were included, all of which fully met the inclusion and exclusion criteria (see Tables 1-5) [32-67]. Among these, the vast majority were cross-sectional (n=26) [32-57], followed by nonrandomized controlled (quasi-experimental) studies (n=4) [60-63] and RCTs (n=4) [64-67]. One study was classified as a global cross-national analysis (n=1) [59], and another as a cohort study (n=1) [58]. Conversely, 63 studies were excluded for the following reasons: 33 either addressed vaccines without exploring constructs related to hesitancy, adherence, or attitudinal factors, or focused on communication channels and media not specifically evaluated within a vaccination context; 4 conducted content analyses of digital or social media platforms without considering a study population; 3 employed unsuitable study designs; 12 did not meet the sample size criterion; 2 were RCT protocols without reporting results; and 9 had irrelevant outcomes.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the study selection process. The figure shows the number of records identified through database searching, along with the screening and eligibility stages, and the final number of studies included in the review. A total of 7275 records were excluded using database filters, and a further 7116 articles were excluded during title and abstract screening. RCT: randomized controlled trial.

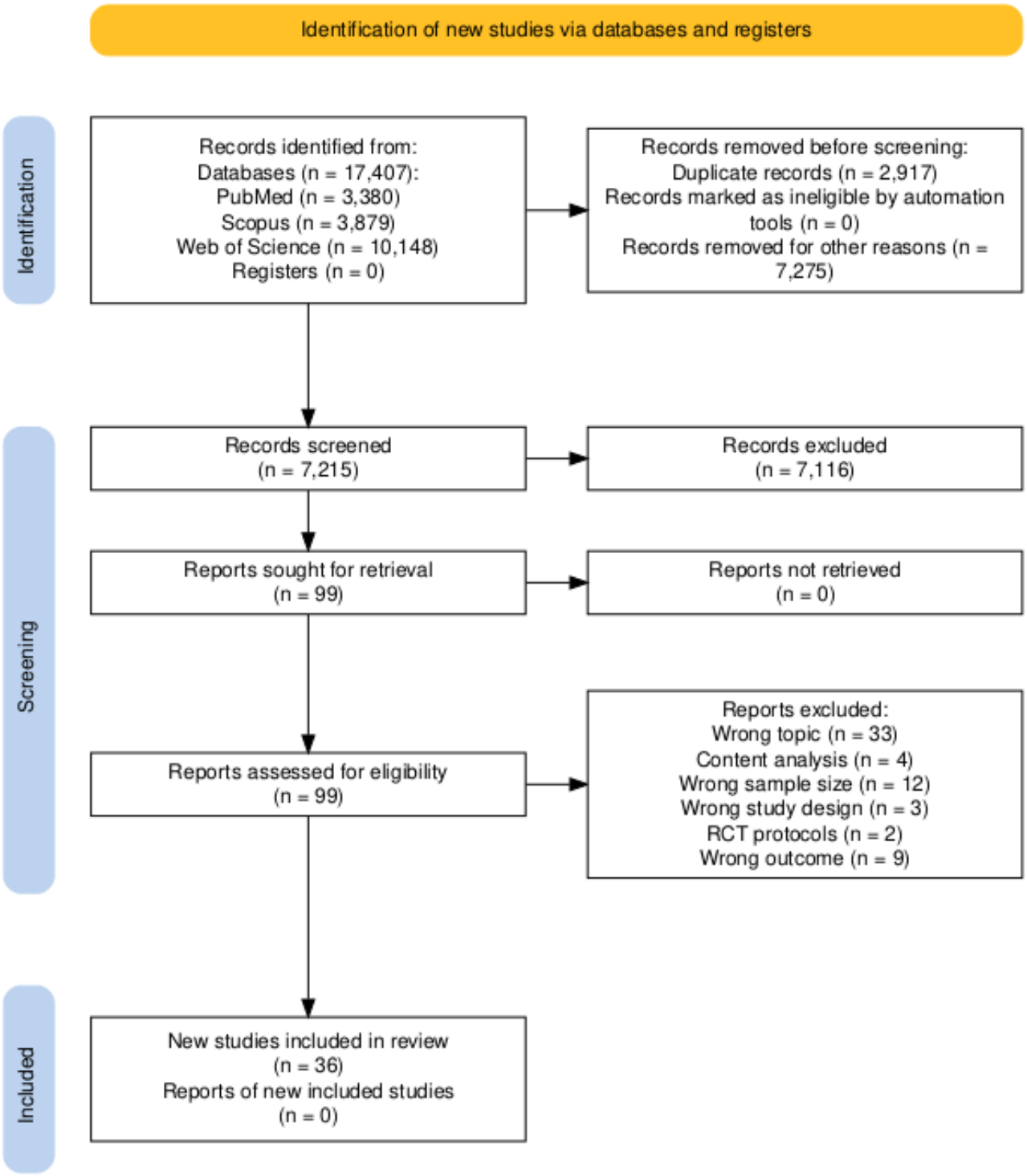

XSL•FO
RenderX

Table 1. Characteristics of cross-sectional studies on communication evidence and vaccination (n=26).

Study	Study design	Sample size	Country of study	Quality score (M-MERSQI[a])	Vaccine discussed	Communication method	Main results
De Giorgio et al [32]	Cross-sectional study	1003 adult participants (median age 40 years) from Croatia, recruited via nonprobability snowball sampling.	Croatia	61	COVID-19	Multiple communication tools, including social network, television/radio, and internet forums	Participants informed via television/radio (odds ratio 2.35, 95% CI 1.71-3.23) or family doctors (odds ratio 2.53, 95% CI 1.78-3.61) were more likely to be vaccinated, whereas reliance on social networks (odds ratio 0.36, 95% CI 0.27-0.49) or blogs (odds ratio 0.34, 95% CI 0.22-0.52) reduced uptake.
Garg et al [33]	Cross-sectional study	Nationally representative sample of US adults from the Health Information National Trends Survey 6 survey; specific N varies per analysis (eg, millions weighted).	United States	67	Human papillomavirus	Health-related videos on social media	Increased exposure to health-related videos improved human papillomavirus vaccine awareness, with the strongest effects among adults aged 18-40 and higher education groups.
Garrett et al [34]	Cross-sectional study	1043 US adults, country-wide quota-based sample, ensuring data quality checks for online participation.	United States	61	COVID-19	Social media	TikTok users reported lower knowledge of COVID-19 guidelines, associated with greater vaccine hesitancy.
Hwang and Shah [35]	Cross-sectional study	4174 US parents of children under 18, including 138 parents with a childhood autism diagnosis, drawn from a nationwide survey.	United States	72	Vaccines in general	Various health information sources (eg, social media, magazines)	Valuing magazines was positively associated with perceived vaccination benefits; valuing television correlated with program maintenance; and social media use correlated negatively with perceived benefits.
Jia et al [36]	Cross-sectional study	1141 US adults; mean age 33.8 (SD 10.8; range 18-79) years; 55% male; 83% White.	United States	62	COVID-19	10 social media posts (from the CoVaxxy dataset, classified by Health Belief Model constructs)	Perceived barriers decreased vaccination intention (b=–0.12, P=.01), while cues to action (b=0.19, P=.002) and self-efficacy messages (b=0.06, P=.04) increased it. "Likes" mediated effects of susceptibility (b=0.03, P=.02) and cues to action (b=0.14, P=.01) on intentions. Celebrities generated more likes (b=0.04, P=.05) and politicians more shares (b=0.05, P=.03).
Jin et al [37]	Cross-sectional study	1946 internet users across Pakistan adults (54.83% men; 44.6% aged between 18 and 30) exposed to COVID-19 vaccine infodemics.	Pakistan	76	Polio	Public service advertisements and digital media platforms	Public service advertisements significantly increased polio vaccine acceptance among Pakistani parents, counteracting misinformation and religious fatalism.
Jung et al [38]	Cross-sectional study	Pooled data from 151,209 women (15+ with child <5 years) from 13 sub-Saharan African countries.	13 countries of Sub-Saharan Africa	75	Childhood vaccines in general	Traditional media (radio, television, and newspaper)	Among 151,209 women in sub-Saharan Africa, daily television/radio consumption was associated with higher childhood vaccination uptake (Bacillus Calmette-Guérin vaccine for tuberculosis; diphtheria, tetanus, and pertussis vaccine; polio; and measles).

XSL•FO
RenderX

Study	Study design	Sample size	Country of study	Quality score (M-MERSQI[a])	Vaccine discussed	Communication method	Main results
Jung [39]	Cross-sectional study	1571 married women (aged between 20 and 40) with underage children from South Korea, China, and Japan (nationally representative online sample).	South Korea, China, and Japan	75	Bacillus Calmette-Guérin vaccine, diphtheria, tetanus, and pertussis vaccine, polio vaccine, and measles vaccine	Television, radio, book, and newspaper	Chinese and South Korean mothers using the internet were less likely to vaccinate children, while traditional media exposure increased uptake.
Khanijahani et al [40]	Cross-sectional study	Pooled data from 220,570 noninstitutionalized US adults over 7 years (2012-2018) from the National Health Interview Survey.	United States	72	Influenza	Internet use for health information and communication	US adults using the internet for health information were more likely to be vaccinated against influenza (odds ratio 1.52, 95% CI 1.45-1.59).
Kim and Jung [41]	Cross-sectional study	1367 Korean adults, nationally representative sample from a mixed method survey (face-to-face and web-based).	South Korea	72	Vaccines in general	General mass media (radio, newspaper, smartphone, and internet)	Higher socioeconomic background in South Korea correlated positively with vaccination adherence.
Kim et al [42]	Cross-sectional study	9584 community-dwelling Medicare beneficiaries (weighted N=50,029,030) from the Winter 2021 Medicare Current Beneficiary Survey.	United States	72	COVID-19	Informal sources of information (social media, internet, and friends/family) and formal source (traditional news, government guidance, and health care providers)	Reliance on informal sources reduced the likelihood of COVID-19 vaccination (odds ratio 0.65, 95% CI 0.56-0.75), testing (odds ratio 0.85, 95% CI 0.74-0.98), and preventive behaviors (odds ratio 0.61, 95% CI 0.50-0.74).
Kobayashi et al [43]	Cross-sectional study	10,192 users of the Corowa-kun chatbot in Japan (median age 55 years, 74% female), participating in an in-chatbot survey.	Japan	61	COVID-19	Corowa-kun chatbot, free chatbot in LINE	Use of the Corowa-kun chatbot revealed higher hesitancy among individuals aged 16-34 (odds ratio 3.7, 95% CI 3.0-4.6) and women (odds ratio 2.4, 95% CI 2.1-2.8).
Lv et al [44]	Cross-sectional study	1673 older adult participants (≥60 years) in Beijing, China, selected via multistage stratified random sampling.	China	72	Influenza	Television, community message boards, and doctors	Older adults exposed to television (odds ratio 1.40, 95% CI 1.12-1.75) and community bulletin boards (odds ratio 1.81, 95% CI 1.45-2.27) were more likely to receive influenza vaccination; coverage was higher in rural versus urban areas (odds ratio 2.57, 95% CI 1.80-3.66).
Melovic et al [45]	Cross-sectional study	1593 parents from Montenegro, Serbia, and Bosnia and Herzegovina responding to online questionnaires.	Montenegro, Serbia, and Bosnia and Herzegovina	68	Vaccines in general	Online media and social marketing	Online media in Montenegro, Serbia, and Bosnia-Herzegovina influenced parental attitudes and correlated with trust in vaccines.

XSL•FO
RenderX

Study	Study design	Sample size	Country of study	Quality score (M-MERSQI[a])	Vaccine discussed	Communication method	Main results
Nah et al [46]	Cross-sectional study	1000 Cameroonians (43.8% female, mean age 20.74 years).	Cameroon	65	COVID-19	Social media use and medical mistrust	Social media use and medical mistrust were positively associated with misinformation beliefs and vaccine hesitancy.
Nah et al [47]	Cross-sectional study	1136 African Americans (53.5% female, age 19-87 years).	United States	65	COVID-19	Social media use and medical mistrust	Same trend observed among African Americans, with social media use and mistrust linked to misinformation and hesitancy.
Piltch-Loeb et al [48]	Cross-sectional study	2650 US individuals (18+) in vaccine priority groups, recruited via mobile survey platform (Pollfish RDE).	United States	72	COVID-19	Traditional media (television, newspapers, and radio) and social media (Facebook, Instagram, YouTube, Twitter, and TikTok)	High trust in television and newspapers increased COVID-19 vaccine acceptance, while reliance on social media or mixed sources decreased acceptance.
Recio-Román et al [49]	Cross-sectional study	27,524 individuals (15+ years) residing in European Union member states, from the EURO-BAROMETRO 91.2 survey (2019).	European Union	79	Vaccines in general	Traditional media (television, radio, newspaper, magazine, social networks, internet sites, etc) and institution (political parties, regional or local public authorities, national government or parliament, European Union, etc)	In 27,524 Europeans, television (β=−.094), radio (β=–.066), and newspapers/magazines (β=−.194) reduced hesitancy; online media weaker (β=−.041); and lack of information increased hesitancy (β=.049).
Sakamoto et al [50]	Cross-sectional study	4062 Japanese university students, recruited via online questionnaire (Learning Management System).	Japan	58	COVID-19	Multiple communication tools. Specifically evaluated a 14-minute-long educational movie to provide information about the COVID-19 vaccine.	Japanese university students reported positive attitudes when informed via television/radio, while negative attitudes were more common among YouTube/Instagram users; the female gender was associated with positivity.
Sunohara et al [51]	Cross-sectional study	3348 controls (20-65 years) from Sapporo, Japan, matched for age and sex to the general population, without COVID-19 diagnosis.	Japan	72	COVID-19	Multiple communication tools (eg, television, newspapers, social media)	Lower odds of incomplete vaccination among television (odds ratio 0.31, 95% CI 0.21-0.44) and newspaper users (odds ratio 0.32, 95% CI 0.20-0.50); higher odds for Facebook (odds ratio 2.36, 95% CI 1.24-4.48) and blogs/bulletins (odds ratio 4.81, 95% CI 2.72-8.48). Older age and male gender were linked to lower acceptance.

XSL•FO
RenderX

Study	Study design	Sample size	Country of study	Quality score (M-MERSQI[a])	Vaccine discussed	Communication method	Main results
Teoh et al [52]	Cross-sectional study	1037 young adults (18-26 years) at the Minnesota State Fair, convenience sample (predominantly White, educated, insured, and vaccinated).	United States	61	Human papillomavirus	Graphic messages: 8 Instagram graphic mockups promoting human papillomavirus vaccination to prevent cancer (4 categories: infographics, disease photos, young adult cancer patient photos, and humorous graphics).	The Instagram human papillomavirus campaign showed modest demographic differences; women rated patient-photo graphics more favorably.
Thompson et al [53]	Cross-sectional study	1192 parents of children/adolescents (9-17 years) from 13 North Texas counties, via web-based surveys.	United States	61	Human papillomavirus	Social media, trust in providers, internet verification skills, and demographics	Fathers had a lower likelihood of human papillomavirus vaccine acceptance than mothers (adjusted odds ratio 0.42, 95% CI 0.27-0.64).
Wada and Smith [54]	Cross-sectional study	3140 Japanese individuals (20-69 years), recruited from an online survey company.	Japan	61	Vaccines in general	Health care providers, national or local government, family, friends, television, newspapers, internet, and books	Women trusted family (odds ratio 1.60), newspapers (odds ratio 1.56), internet (odds ratio 2.19), and books (odds ratio 2.99) more; men trusted friends (odds ratio 1.96).
Wamba et al [55]	Cross-sectional study	3108 adults from France (2001) and South Africa (1107), members of a market research firm's panel.	France and South Africa	75	COVID-19	Traditional media, specialized health sources, family or friends, social media, and web-based sources	Identified 4 opinion profiles; enthusiasts, sceptics, and followers trusted traditional and health sources; conspiracy theorists showed low trust in all channels. Social media influence did not differ by profile.
Yoneoka et al [56]	Cross-sectional study	30,053 participants in a nationwide online survey in Japan (February to March 2021), using quota sampling (matched for age, gender, prefecture population).	Japan	61	COVID-19	Multiple media channels (ie, medical professionals, television, newspapers, The Novel Coronavirus Expert Meeting, and local government)	Uncertain individuals avoided authoritative sources; reliance on expert meetings/local government associated with unwillingness, possibly due to perceived pressure. Under 50s sought authoritative information; over 50s relied more on social media.
Al Zahrani [57]	Cross-sectional study	7159 adults (18+) living in Saudi Arabia.	Saudi Arabia	61	Vaccines in general	Promotion by health care leaders	Men were more responsive than women to health care leaders' influence in promoting vaccination.

[a]M-MERSQI: Modified Medical Education Research Study Quality Instrument.

XSL•FO
RenderX

Table 2. Characteristics of cohort studies on communication evidence and vaccination (n=1).

Study	Study design	Sample size	Country of study	Quality score (M-MERSQI[a])	Vaccine discussed	Communication method	Main results
DeMora et al [58]	Cohort study	1768 Americans from a probability-based longitudinal survey conducted over 10 months.	United States	71	COVID-19 and influenza	Social media	Social media engagement increased vaccination rates, with effects moderated by political orientation (Democrats: pathogen information exposure; Republicans: perceived peer vaccination).

[a]M-MERSQI: Modified Medical Education Research Study Quality Instrument.

Table 3. Characteristics of global cross-national analysis on communication evidence and vaccination (n=1).

Study	Study design	Sample size	Country of study	Quality score (M-MERSQI[a])	Vaccine discussed	Communication method	Main results
Wilson and Wiysonge [59]	Global cross-national analysis	Examination of public attitudes on vaccine safety in 137 countries, vaccination rates in 166 countries, and 258,769 geocoded vaccination-related tweets	Worldwide	64	Vaccines in general	Social media	Significant relationship between the use of social media for organizing offline action and public doubts about vaccine safety. Substantial connection between foreign disinformation campaigns and declining mean vaccination rates over time.

[a]M-MERSQI: Modified Medical Education Research Study Quality Instrument.

XSL•FO
RenderX

Table 4. Characteristics of nonrandomized controlled studies (quasi-experimental) on communication evidence and vaccination (n=4).

Study	Study design	Sample size	Country of study	Quality score (M-MERSQI[a])	Vaccine discussed	Communication method	Main results
Evans et al [60]	Nonrandomized controlled study (quasi-experimental)	1933 Nigerians (adults 18+) from all 37 states, recruited via Facebook, followed for 10 months.	Nigeria	69	COVID-19	Social media (Facebook/Instagram)	Facebook/Instagram campaign in Nigeria increased COVID-19 vaccination rates compared with control participants.
Kim et al [61]	Nonrandomized controlled study (quasi-experimental)	5804 respondents from Tanzania (3442 before campaign and 2362 after campaign), convenience sample.	Tanzania	58	COVID-19	"The One by One: Target COVID-19" social media campaign. High-profile and high-impact influencers and viral content on social media platforms	The Tanzanian social media campaign reduced hesitancy in adults ≥35 years and increased uptake in those aged 25-34, but showed no effect among 18-24-year-olds.
Pavoncello et al [62]	Cross-sectional study	854 adults at phase 1 survey (48.1% male and 51.9% female), 1034 adults at phase 2 survey; total vaccine doses delivered were 24,888.	Madagascar	58	COVID-19	"CoBoGo" campaign (multiple communication tools) using tailored adaptive strategy	Weekly vaccine uptake rose by 8% in phase 2 (relative risk 1.08; 95% CI 1.01-1.15). Vaccine hesitancy was unchanged (Δ=0.02; 95% CI –0.04 to 0.08), knowledge of reliable sources improved (Δ=0.04; 95% CI 0.003-0.08), perception of insufficient information decreased (Δ=–0.11; 95% CI –0.15 to –0.07), and recognition of community health workers increased (Δ=0.07; 95% CI 0.02-0.12).
Schuh et al [63]	Nonrandomized controlled study (quasi-experimental)	13,889 participants in a nationally representative quasi-experimental study in the United States, recruited via RIWI RDIT[b] technology.	United States	63	COVID-19	Informational animated videos, with online experiments testing debunking/prebunking texts addressing messenger RNA vaccine misinformation versus control texts.	COVID-19 educational videos with ethnically congruent protagonists increased the likelihood of full viewing (odds ratio 1.89; *P*<.01).

[a]M-MERSQI: Modified Medical Education Research Study Quality Instrument.

[b]RIWI RDIT: Real-Time Interactive World-Wide Intelligence Random Domain Intercept Technology.

XSL•FO
RenderX

Table 5. Characteristics of RCTs[a] on communication evidence and vaccination (n=4).

Study	Study design	Sample size	Country of study	Quality score (M-MERSQI[b])	Vaccine discussed	Communication method	Main results
Daley et al [64]	Three-arm RCT	1093 pregnant women (≥18 years) with internet access and health insurance, followed until their child was 12-15 months.	United States	67	Vaccines in general	Website with vaccine information and social media components (vaccine social media arm); website with vaccine information only (vaccine information arm)	Both website-based strategies (with/without social media components) significantly improved parental vaccination attitudes and reduced concerns compared with usual care.
Geniole et al [65]	RCT	1584 UK residents (67.6% women; mean age 34.91 years) recruited via Prolific (online labor market).	United Kingdom	61	COVID-19	Internet memes	Exposure to pro-vaccination memes modestly but significantly increased COVID-19 vaccination intention by approximately 3.3 points on a 100-point scale, independent of age, gender, and political orientation. The effect was context-dependent—robust before the announcement of a safe/effective vaccine (~7-point increase) but negligible thereafter—with no corresponding pre/postannouncement change in the control group.
Lee et al [66]	RCT	2045 adult guardians of unvaccinated children and seniors in Thailand, Hong Kong, and Singapore, with 748 participants completing the study after follow-up.	Thailand, Hong Kong, and Singapore	63	COVID-19	COVID-19 chatbots (ChatSure or D24H via Messenger/WhatsApp)	COVID-19 chatbot increased acceptance and trust, particularly among minorities and lower-education groups.
Schmid et al [67]	RCT	Experiment 1: 1382 German adults; experiment 2: 1054 unvaccinated German adults, both recruited via the online panel provider respondi.de using quota sampling.	Germany	60	COVID-19	Online experiments testing debunking/prebunking texts addressing messenger RNA vaccine misinformation versus control texts	Religiosity moderated credibility judgments of debunking messages at the 2-month follow-up, whereas prebunking was not affected by spirituality.

[a]RCT: randomized controlled trial.

[b]M-MERSQI: Modified Medical Education Research Study Quality Instrument.

The publication period spanned from 2015 to 2024. Notably, the highest number of studies were published in 2022 (n=8) [42,43,50,53,56,60,63,66] and 2023 (n=8) [32,37,40,45,48,52,59,67], followed by 2024 (n=6) [35,38,39,41,44,54], 2018 (n=3) [33,34,61], 2021 (n=2) [46,49], 2020 (n=2) [51,55], 2025 (n=2) [64,65], and 2015 (n=2) [36,62]. A single study was published in each of the following years: 2019 [58], 2017 [47], and 2016 [57]. The studies originated from several countries: the United States accounted for the largest number (n=13) [33-36,40,42,47,48,52,53,58,63,64], followed by multicountry studies (n=7) [43,45,50,51,54,56,66] and Japan (n=5) [38,39,49,55,59]. Croatia [32], China [44], Cameroon [46], Germany [67], Nigeria [60], Pakistan [37], United Kingdom [65], South Korea [41], Saudi Arabia [57], Madagascar [62], and Tanzania [61] each contributed 1 study.

The included studies evaluated different vaccines. The COVID-19 vaccine was the most studied (n=19) [32,35,37-41,43-45,48,50,52-54,56,59,66,67], followed by studies on multiple vaccines (n=11) [33,34,36,47,57,58,60-62,64,65]. Vaccines for human

XSL•FO
RenderX

papillomavirus (n=3) [42,46,63], influenza (n=2) [49,51], and polio (n=1) [55] were less represented.

Regarding sample size, 4 studies involved more than 10,000 participants [43,49,56,63], while 3 included more than 100,000 participants [38,40,59]. All remaining studies included more than 1000 participants [32-37,39,41,42,44-48, 50-55,57,58,60-62,64-67].

The overall methodological quality was assessed as "moderate" for 19 studies [32,34,36,43,46,47,50,52-54,56,57,59, 61-63,65-67] and "high" for the remaining 17 [33,35,37-42,44,45,48,49,51,55,58,60,64]. The M-MERSQI score for each study is provided in Tables 1-5.

Across the cross-sectional studies, the risk of bias was generally low for selection/context (D1-D2), case definition (D4), control of confounders (D5-D6), and analytic appropriateness (D8); the most frequent limitations were exposure and outcome measurement (D3 and D7), due to self-report and limited external validation. Overall, 9 studies had a low risk of bias [32,33,35,38,39,44,48,49,55], and 17 had a moderate risk of bias [34,36,37,40-43,45-47,50-54,56,57] (Figure 2; also see [32-57]). The only cohort study achieved good comparability but had an unclear risk for selection and outcome assessment (Figure 3; also see [58]). The nonrandomized experimental studies were judged to have an overall moderate risk of bias. This determination was driven primarily by bias due to confounding (D1), rated as moderate across all studies, and by missing data (D5) and incomplete outcome data (D6), which were rated as moderate in 3 of the 4 studies [60,61,63]. Overall, no study was judged to have a serious risk of bias (Figure 4; see also [60-63]).

XSL•FO
RenderX

Figure 2. Risk of bias across cross-sectional studies (traffic-light plot). Domains assessed were as follows: D1, inclusion criteria; D2, description of participants and setting; D3, exposure measurement; D4, criteria for measuring the condition; D5, identification of confounders; D6, strategies to address confounding; D7, outcome measurement; and D8, statistical analysis. Each row represents an included study (n=26), and the rightmost column shows the overall study-level judgment derived from the domain ratings.

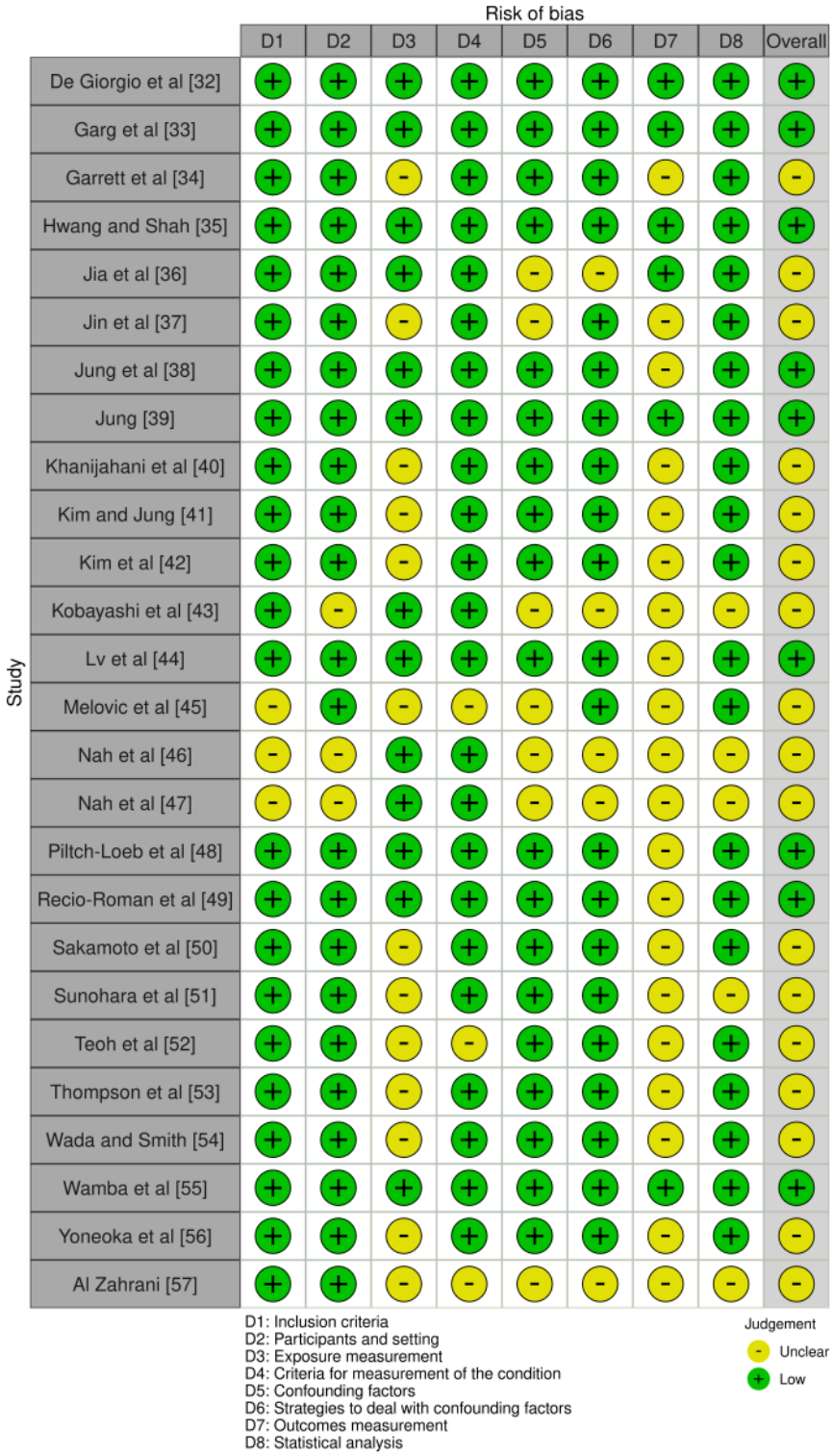

XSL•FO
RenderX

Figure 3. Risk of bias assessment for cohort studies (traffic-light plot). Domains assessed were as follows: D1, selection; D2, comparability; and D3, outcome/exposure. The overall risk-of-bias judgment is shown in the rightmost column.

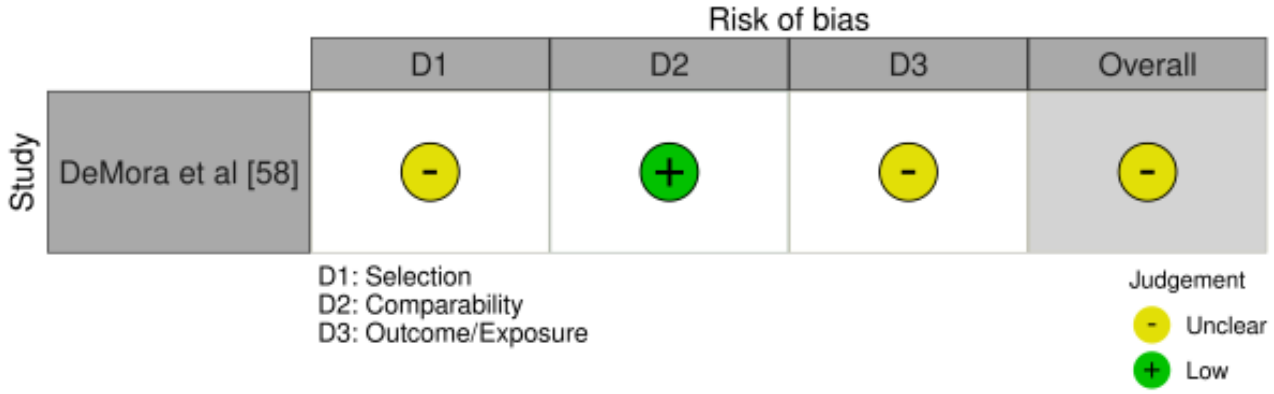

Figure 4. Risk of bias across included nonrandomized studies (traffic-light plot). Domains assessed were as follows: D1, bias due to confounding; D2, bias due to selection of participants; D3, bias in classification of interventions; D4, bias due to deviations from intended interventions; D5, bias due to missing data; D6, bias in measurement of outcomes; and D7, bias in selection of the reported result. The overall risk-of-bias judgment is shown in the rightmost column.

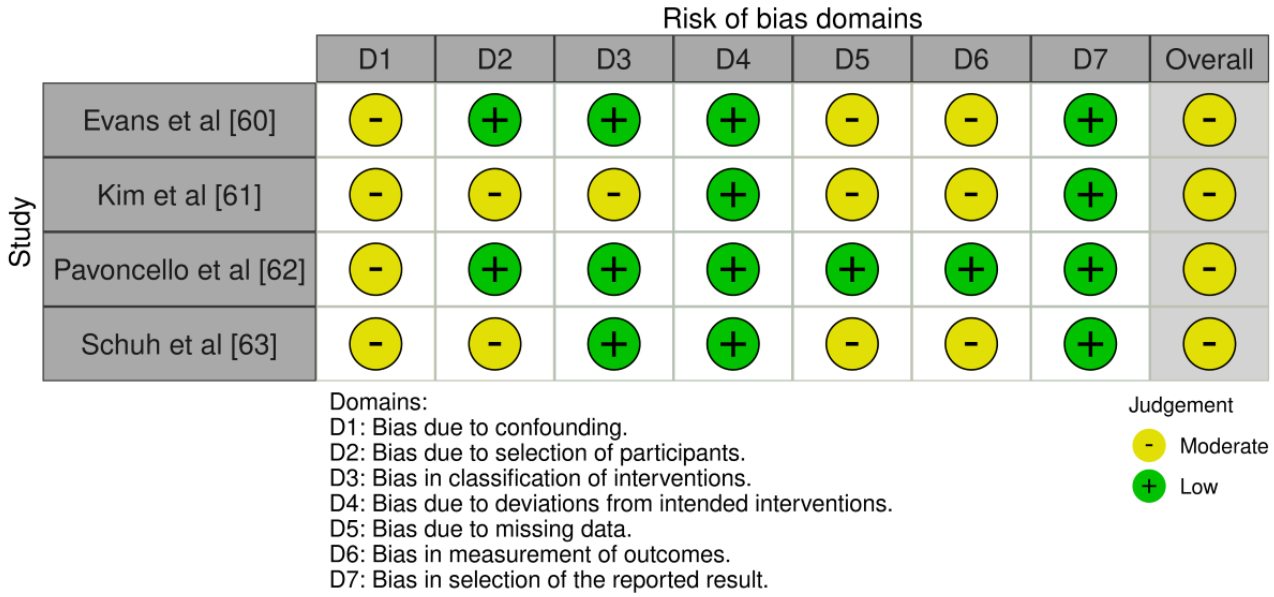

Finally, RCTs showed a low risk of bias for randomization and intervention fidelity, with the main concerns related to outcome measurement and missing data. Selective reporting was occasionally unclear. Overall, most trials had some concerns (Figure 5; see also [64-67]).

XSL•FO
RenderX

Figure 5. Risk of bias in randomized controlled trials (traffic-light plot). Domains assessed were as follows: D1, bias arising from the randomization process; D2, bias due to deviations from intended interventions; D3, bias due to missing outcome data; D4, bias in measurement of the outcome; and D5, bias in selection of the reported result. Each row represents an included RCT (n=4), with the overall risk-of-bias judgment shown in the rightmost column.

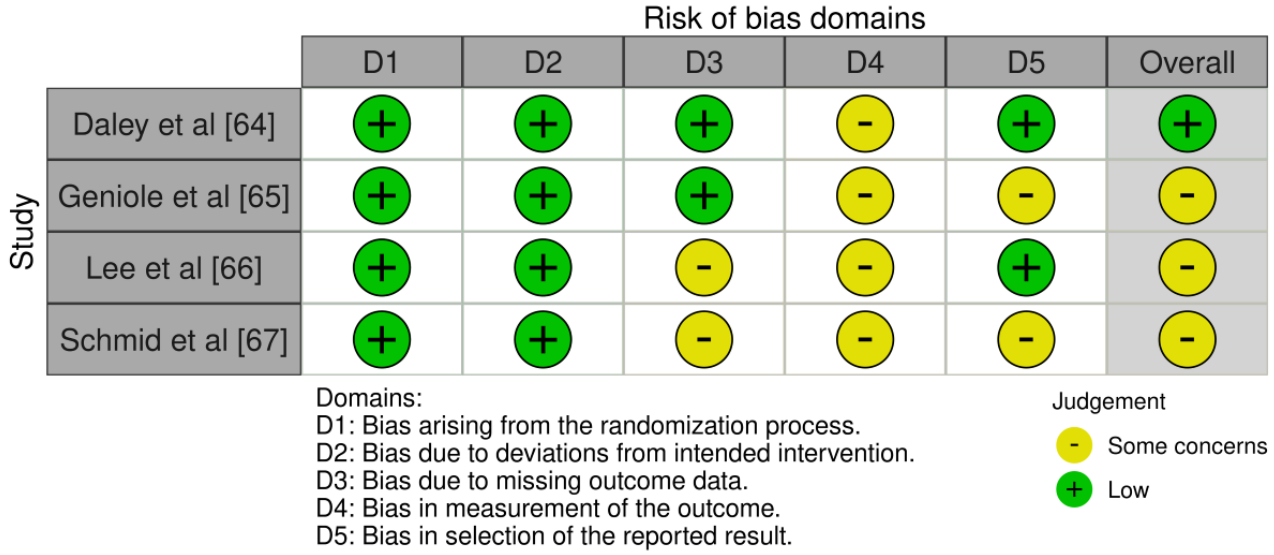

Observational Evidence About Media Exposure and Vaccine Attitudes, Acceptance, and Hesitancy

Sixteen cross-sectional studies [32,34-40,42,44,45,48-51,56] and 1 global cross-national analysis [59] examined the effect of media on vaccination acceptance. Recio-Román et al [49], in a study of 27,524 European citizens, underscored a pronounced positive correlation between traditional media and the mitigation of vaccine hesitancy, compared with new media. Television (β=−.094; *P*<.001; 95% CI −0.132 to −0.056), radio (β=−.066; *P*<.001; 95% CI −0.092 to −0.041), and newspapers/magazines (β=−.194; *P*<.001; 95% CI −0.229 to −0.162) exerted considerable influence. The effect on vaccine adherence was mainly mediated by reductions in hesitancy, except for television, which did not show a significant effect on increasing vaccine uptake, potentially due to the context and sociodemographic characteristics of the recipients. Conversely, online media demonstrated less influence in reducing hesitancy (β=−.041; *P*<.001; 95% CI −0.068 to −0.014) and indirect influence in increasing adherence (β=.021; *P*<.05; 95% CI 0.003-0.038), while other internet sources showed no statistical significance. Lack of information increased vaccine hesitancy (β=.049; *P*<.001; 95% CI 0.008-0.089) and decreased vaccine adherence (β=−.063; *P*<.001; 95% CI −0.090 to 0.037), highlighting the detrimental effects of unreliable communication and information. Similar findings were reported by Sakamoto et al [50], who observed positive attitudes toward COVID-19 vaccination among 4062 Japanese university students, particularly among those relying on television and radio, whereas negative attitudes were more frequent among users of YouTube and Instagram. Likewise, Jung et al [38], in a study involving 151,209 women from sub-Saharan African countries, showed that daily exposure to television and radio increased the likelihood of childhood vaccination against tuberculosis (Bacillus Calmette-Guérin); diphtheria, tetanus, and pertussis; polio; and measles. Similarly, Jung [39] demonstrated that radio use was significantly associated with a higher likelihood of vaccinating children (adjusted odds ratio 1.577, 95% CI 1.040-2.393; *P*<.05) among Japanese, Chinese, and South Korean mothers. Conversely, internet use in China and South Korea was associated with a lower likelihood of vaccinating children.

Hwang and Shah [35] found that the value attributed to magazines by 4174 US parents was positively associated with their perception of vaccination benefits. Additionally, the value attributed to television was positively correlated with the maintenance of vaccination programs. By contrast, social media use was negatively associated with the perception of vaccination benefits. De Giorgio et al [32] highlighted that participants informed by television, radio (odds ratio 2.35, 95% CI 1.71-3.23; *P*<.001), or family doctors (odds ratio 2.53, 95% CI 1.78-3.61; *P*<.001) were more likely to be vaccinated, whereas reliance on social networks (odds ratio 0.36, 95% CI 0.27-0.49; *P*<.001), general internet forums/blogs (odds ratio 0.34, 95% CI 0.22-0.52; *P*<.001), or friends/acquaintances (odds ratio 0.66, 95% CI 0.48-0.91; *P*=.01) was linked to lower uptake. Lv et al [44] found that older adults receiving information through television (odds ratio 1.403, 95% CI 1.124-1.752; *P*<.01) and community bulletin boards (odds ratio 1.812, 95% CI 1.446-2.270; *P*<.01) were more likely to be vaccinated against influenza. Similarly, Kim et al [42] showed that adults using informal sources were less likely to receive the COVID-19 vaccine (odds ratio 0.65, 95% CI 0.56-0.75), undergo COVID-19 testing (odds ratio 0.85, 95% CI 0.74-0.98), or adopt preventive behaviors (odds ratio 0.61, 95% CI 0.50-0.74), whereas reliance on diversified sources increased all 3 outcomes.

XSL•FO
RenderX

The negative role of social media was further highlighted by Garrett et al [34], who found lower knowledge of COVID-19 guidelines among TikTok users (β=–.29; 95% CI –0.58 to –0.004; *P*=.047). Instagram (β=–.40; 95% CI –0.68 to –0.12; *P*=.005) and Twitter (β=–.33; 95% CI –0.58 to –0.08; *P*=.01) use was significantly associated with perceiving the guidelines as overly strict. Additionally, Facebook (β=–.23; 95% CI –0.42 to –0.043; *P*=.02) and TikTok (β=–.25; 95% CI –0.50 to –0.009; *P*=.04) use was linked to lower adherence to the guidelines. Among Cameroonians and African Americans, social media use, along with medical mistrust, was positively associated with beliefs in misinformation about the COVID-19 vaccine and, inferentially, with vaccine hesitancy [46,47]. Wilson and Wiysonge [59], in a global cross-national observational study using survey data from 137 countries (Wellcome Global Monitor), vaccination rates from 166 countries (WHO), and 258,769 geocoded tweets (2018-2019), showed that social media use for organizing action was significantly associated with doubts about vaccine safety. Foreign disinformation campaigns were linked to declining vaccination rates (≈2% yearly decrease per scale point) and a 15% increase in negative vaccine-related tweets.

Piltch-Loeb et al [48] illustrated that local television was the most commonly used source of COVID-19 vaccine information (61.0%). National television (relative risk reduction 1.25; 95% CI 1.02-1.53; *P*<.05), local television (relative risk reduction 1.75; 95% CI 1.40-2.19; *P*<.001), and national newspapers (relative risk reduction 1.81; 95% CI 1.45-2.24; *P*<.001) were associated with higher acceptance, particularly when supported by high levels of trust. Acceptance significantly decreased among individuals relying on social media (relative risk reduction 0.45; 95% CI 0.32-0.64; *P*<.001) or on both social and traditional media (relative risk reduction 0.81; 95% CI 0.66-1.00; *P*=.05). Sunohara et al [51] reported lower odds of incomplete second vaccinations among television (odds ratio 0.31, 95% CI 0.21-0.44) and newspaper users (odds ratio 0.32, 95% CI 0.20-0.50), and higher odds among those relying on books (odds ratio 3.34, 95% CI 1.58-7.06), commercial video sites (odds ratio 2.22, 95% CI 1.44-3.43), Facebook (odds ratio 2.36, 95% CI 1.24-4.48), and personal blogs or electronic bulletin boards (odds ratio 4.81, 95% CI 2.72-8.48).

Likewise, Yoneoka et al [56] noted that 30,053 Japanese individuals who were hesitant about vaccination often avoided authoritative sources such as medical professionals, the government, and television. Jin et al [37] showed that public service advertisements positively influenced offspring polio vaccine acceptance among 2160 Pakistani parents, counteracting misinformation, fake news, and religious fatalism.

However, the positive and responsible use of the internet and social media can be a valuable strategy for disseminating high-quality information and increasing trust in vaccines, as highlighted by Khanijahani et al [40]. Adults using the internet for health-related information were more likely to receive influenza vaccinations (odds ratio 1.52, 95% CI 1.45-1.59) compared with non–health information seekers and noninternet users. Melovic et al [45], evaluating 1593 participants in Montenegro, Serbia, and Bosnia-Herzegovina, demonstrated that online media play a substantial role in shaping parents' attitudes toward child vaccination and correlate with the degree of trust in vaccines. Jia et al [36] tested how Health Belief Model constructs and social media messaging affect COVID-19 vaccine intentions among 1141 unvaccinated US adults. Perceived barriers were negatively (*b*=–0.12; *P*=.01) and cues to action positively (*b*=0.19; *P*=.002) associated with vaccination intention; self-efficacy messages also increased intention (*b*=0.06; *P*=.04). Engagement, particularly "likes," significantly mediated the effect of susceptibility and cues to action on both vaccination intention and persuading others.

Observational Evidence About the Role of Sociodemographic Factors

A total of 15 cross-sectional studies [32,33,37,41,43,44,49-57] and 1 cohort study [58] addressed the role of sociodemographic factors. Recio-Román et al [49] showed that older age groups had lower odds of vaccination than younger ones, while no statistically significant differences were observed after adjusting for occupation. Higher educational attainment, better economic circumstances, and higher social class were associated with greater uptake. Families with 1 child were more likely to vaccinate, whereas no statistically significant differences were found for families with more than 1 child. Political orientation also mattered, with left-leaning individuals showing 8.7% higher odds of vaccination compared with centrists. Garg et al [33] demonstrated greater awareness of the human papillomavirus vaccine among US adults exposed to health-related videos on social media, particularly among younger and more educated adults (18-40 years). De Giorgio et al [32] confirmed that individuals with a bachelor's degree (odds ratio 2.25, 95% CI 1.14-4.46; *P*=.02) or a PhD (odds ratio 1.97, 95% CI 1.01-3.52; *P*=.02) were more likely to receive COVID-19 vaccination compared with those with only a high school education.

Age differences also played a role. Yoneoka et al [56] reported that uncertain individuals aged under 50 years typically sought vaccine information from authoritative sources, including health care professionals, expert meetings, and both local and central government agencies. By contrast, individuals aged over 50 years relied more on social media. However, Kobayashi et al [43] found that young adults (odds ratio 3.7, 95% CI 3.0-4.6), aged 16-34 years, and women (odds ratio 2.4, 95% CI 2.1-2.8) were more likely to develop hesitancy when exposed to chatbot-based COVID-19 vaccine information. Sunohara et al [51] also documented lower acceptance among older individuals and males using social media.

Regarding gender, Teoh et al [52] examined the appeal of different graphic formats for human papillomavirus vaccine promotion among 1037 young adults (aged 18-26 years) in Minnesota. They found significant differences, albeit modest in effect size, related to demographic characteristics such as gender, ethnicity, and level of digital health literacy. Women rated patient photos higher than men (mean difference 0.28; 95% CI 0.14-0.41; *P*=.0001). Participants with high eHealth literacy provided higher ratings for infographics, humorous graphics, and disease photos. Sakamoto et al [50] also highlighted that female gender was associated with a positive attitude toward the vaccine. Wada and Smith [54] illustrated gender differences related to trust in government

XSL•FO
RenderX

recommendations and perceived reliability of various information sources. Specifically, women tended to consider information from family (odds ratio 1.60, 95% CI 1.23-1.99), newspapers (odds ratio 1.56, 95% CI 1.03-2.15), the internet (odds ratio 2.19, 95% CI 1.58-2.73), and books (odds ratio 2.99, 95% CI 2.19-3.40) more reliable, whereas men tended to place more trust in information from friends (odds ratio 1.96, 95% CI 1.24-2.60). Conversely, Al Zahrani [57] highlighted that men were significantly more inclined than women to accept the influence of health care leaders in promoting vaccination. Thompson et al [53] found that fathers had a significantly lower likelihood of accepting human papillomavirus vaccination compared with mothers (adjusted odds ratio 0.42, 95% CI 0.27-0.64; *P*<.001). Regarding contextual differences, Lv et al [44] found that vaccination coverage among the older adult population in Beijing was significantly higher in rural areas (odds ratio 2.57, 95% CI 1.801-3.655; *P*<.01) compared with urban areas.

DeMora et al [58] demonstrated that the relationship between social media use and vaccine uptake was shaped by political orientation, being linked to information about emerging pathogens among Democrats and to the belief that individuals within their social circle are receiving the recommended vaccines among Republicans.

Beliefs and opinions play a critical role in influencing vaccine acceptance. Wamba et al [55] identified 4 opinion profiles: enthusiasts, skeptics, followers, and conspiracy theorists. Enthusiasts, followers, and skeptics generally placed greater trust in traditional media, friends and family, online resources, and specialized health sources, whereas conspiracy theorists demonstrated low trust in all these sources. Social media use showed no significant differences across sociodemographic profiles, suggesting that it is not the primary channel of influence. Furthermore, Jin et al [37] suggested that religious fatalism significantly moderated the association between risk perception and polio vaccine acceptance, showing an inverse effect (β=−.27; *P*=.01). Kim and Jung [41] highlighted that higher socioeconomic status facilitated access to reliable information channels, supporting vaccination adherence.

Experimental Evidence About Media Exposure and Vaccine Attitudes, Acceptance, and Hesitancy

A randomized controlled trial [64] and 2 nonrandomized experimental studies [60,62] provided particularly strong evidence on the relationship between media exposure and vaccine acceptance.

Pavoncello et al [62], in their quasi-experimental study, evaluated the CoBoGo campaign in Madagascar (based on radio spots, community-based initiatives, and endorsement by community health workers), which applied a tailored, adaptive strategy. Phase 2 adaptations produced a significant 8% weekly increase in uptake (relative risk 1.08; 95% CI 1.01-1.15). Vaccine hesitancy remained stable (Δ=0.02; 95% CI –0.04 to 0.08), while sustainability improved through the training of 340 staff, facility upgrades, and integration of standard operating procedures. In an RCT by Daley et al [64], 1093 US pregnant women were randomized to one of the following: a website with vaccine information and social media components; a website with vaccine information only; or usual care. When comparing baseline with time point 1 among vaccine-hesitant parents, both strategies were associated with significant improvements in vaccination attitudes relative to the usual care group. Similarly, when comparing baseline with time point 2, both arms were linked to significant reductions in parental concerns about vaccination.

Similarly, Evans et al [60] showed that Nigerian adults exposed to a public health campaign on Facebook and Instagram exhibited higher vaccination rates. At the first follow-up, the vaccination rate was 31.7% in the treatment states compared with 25.3% in the comparison states, with an adjusted difference of 7.8 percentage points (*P*=.02). By the second follow-up, vaccination rates had increased to 44.1% in the treatment states and 35.0% in the comparison states, with a nonsignificant adjusted difference of 11.0 percentage points (*P*=.07).

Experimental Evidence About the Role of Sociodemographic Factors

Three RCTs [65-67] provided significant evidence on the role of sociodemographic determinants. Schmid et al [67], in a large-scale RCT, demonstrated that at 2-month follow-up, religiosity exerted a strong moderating effect on individuals' credibility judgments of a false myth–debunking message regarding messenger RNA vaccines (*b*=0.17; 95% CI 0.05-0.30; t_{810}=2.80; *P*=.005), whereas no such effect was observed in the short term (*b*=0.01; 95% CI −0.09-0.10; t_{1377}=0.10; *P*=.92). By contrast, the effectiveness of a prebunking message appeared to be unaffected by spirituality (*b*=0.04; 95% CI –0.05 to 0.14; t_{1049}=0.89; *P*=.37). Beyond religiosity, another RCT emphasized the role of education. Lee et al [66], testing the effectiveness of a COVID-19 chatbot among unvaccinated or hesitant individuals, observed the greatest improvements in vaccine acceptance and trust among participants belonging to minority groups and those with lower levels of education. Geniole et al [65] showed that exposure to provaccination memes modestly but significantly increased COVID-19 vaccination intention by about 3.3 points on a 100-point scale, independent of age, gender, and political orientation. The effect was context-dependent—robust before the announcement of a safe and effective vaccine (~7-point increase) but negligible thereafter—with no corresponding pre/postannouncement change in the control group. Regarding age, Kim et al [61] highlighted that a social media campaign aimed at boosting COVID-19 vaccination confidence in Tanzania increased uptake mainly among individuals aged 25-34 years, with limited effects in younger groups. A significant reduction in vaccine hesitancy was observed among individuals aged 35 years and older. Regarding the role of cultural adaptation, Schuh et al [63] showed that ethnic congruence between the protagonist of educational videos about the COVID-19 vaccine and the viewer increased the likelihood of watching the video in full (odds ratio 1.89; *P*<.01).

XSL•FO
RenderX

Discussion

Overview and Rationale

Amid the COVID-19 "infodemic," understanding how information delivery is associated with vaccine literacy and public trust has become a critical public health priority. Therefore, this systematic review analyzes the correlation between communication media and vaccination attitudes, intentions, and hesitancy across different study designs, while specifically examining the moderating influence of sociodemographic determinants.

Main Findings

The review synthesizes evidence from 36 studies, predominantly cross-sectional, showing that tailored communication strategies across both digital and traditional media channels play a significant role in enhancing vaccine acceptance and mitigating hesitancy. While adaptive digital interventions and social media campaigns are particularly effective in reaching specific cohorts, such as hesitant populations and individuals from minority or lower-education backgrounds, traditional media appear to be associated with higher levels of public trust and adherence. Furthermore, the efficacy of these public health messages is strongly moderated by sociodemographic and cultural factors, including religiosity and cultural congruence, which influence message credibility and engagement.

Cross-Sectional Evidence: Patterns and Sociodemographic Moderation

The synthesis of cross-sectional evidence suggests that exposure to traditional media is strongly associated with reduced vaccine hesitancy and higher uptake rates across large-scale studies, likely due to the editorial oversight inherent in these channels [68-70]. Conversely, social media use is frequently observed as a major vehicle for antivaccination narratives and conspiracy theories, exacerbated by the echo chamber effect [71-73]. However, this relationship appears ambivalent [74]: purposeful internet use for health-seeking behavior remains a positive predictor of vaccination [40], suggesting that user intent moderates the medium's effect.

Across sociodemographic lines, high educational attainment and socioeconomic status generally appear to buffer against hesitancy. Complex influences of age and gender are observed. Young adults appear to prefer official and verified information sources, whereas older adults rely more heavily on social media, reflecting the "digital divide" phenomenon [75]. Furthermore, older individuals may have concerns regarding potential interactions between vaccines and preexisting conditions [76]. Regarding gender, women were found to prefer official sources of vaccine information, suggesting a greater reliance on messages emphasizing family or community protection [77]. Nevertheless, previous studies indicate that women are also more likely to delay or refuse vaccination due to fear of side effects and increased exposure to antivaccine content focusing on maternal health and fertility [78,79]. Cultural factors were observed to significantly shape trust in information sources, indicating that messaging must be contextually adapted [70,80]. The validity of these findings is supported by large sample sizes and generally moderate-to-high quality scores, but the cross-sectional design precludes causal inference, limiting conclusions to associations rather than direct effects. It is important to note that 17 out of 26 studies [34,36,37,40-43,45-47,50-54,56,57] have a moderate risk of bias, primarily due to measurement error, as most studies rely on self-reported exposure and vaccination status, potentially introducing recall and social desirability biases.

Longitudinal Cohort Study and Global Analysis

The review goes beyond cross-sectional data by including 1 cohort study [58] and 1 global cross-national analysis [59]. The global study of 137 countries showed that foreign disinformation campaigns are strongly associated with declining vaccination rates, with an estimated 2% annual decrease in coverage for each increase in disinformation, alongside a 15% increase in negative vaccine-related tweets.

The cohort study, based on a 10-month US survey, showed that social media engagement could increase vaccination uptake, but only when moderated by political orientation. Democrats respond more to information about emerging health threats, whereas Republicans are influenced by perceptions of peer vaccination [58,59]. These results are adequately controlled for the main confounding factors (eg, sociodemographic factors), but a residual risk of bias remains due to self-reported outcomes and potential inconsistencies with selection criteria. Together, these studies suggest that social media can promote vaccination through normative influence but remains highly vulnerable to disinformation and external manipulation.

Experimental Evidence: Validating Targeted Interventions

While observational studies suggest a negative correlation between general social media use and vaccine trust, experimental findings confirm that well-structured communication campaigns (based on traditional media such as radio, community-based initiatives, and the active involvement of trusted health care professionals, local institutions, and well-known testimonials) can effectively reverse this trend, depending on design, composition, and content quality [60,62,64]. These results are supported by low outcome measurement bias, as they are based on objective administrative data. This aligns with the example of the Mediterranean diet, which illustrates how well-coordinated communication campaigns can significantly shape health-related behaviors [81,82]. Furthermore, digital tools can bridge equity gaps noted in cross-sectional sociodemographic analyses, as shown by Lee et al [66] regarding minorities and individuals with lower socioeconomic status and education levels, who are more vulnerable to inequities and distrust toward scientific sources [83,84]. Similarly, creative communication strategies based on humor and storytelling can overcome the influence of sociodemographic determinants [65,85], as emotional engagement helps sustain attention and improves message retention [86]. However, some concerns arise regarding these results due to sampling and dropout bias, which are common in online experiments and may introduce systematic differences between participants who complete the study and those who withdraw prematurely.

XSL•FO
RenderX

The evidence regarding cultural moderators remains complex. The influence of religiosity at 2-month follow-up on the effectiveness of debunking messages is consistent with the "Worldview Backfire Effect" model, according to which deeply rooted spiritual and cultural beliefs tend to prevail even when confronted with contradictory information [67,87]. Conversely, prebunking messages—designed to present scientific facts and anticipate misinformation to build resistance against false claims—are not influenced by levels of spirituality, representing valuable tools for public health communication [88]. Likewise, the significant preference for vaccine information content featuring individuals from the same ethnic background [55,63]—consistent with social and ethnic identification theories—needs to be considered in communication strategies [89]. These findings should be interpreted with caution, as they are primarily based on self-reported outcomes.

However, despite the strength of randomized designs in establishing causality, this body of evidence is characterized by substantial heterogeneity in interventions (ranging from internet memes to community radio) and outcome measures (hypothetical intention vs validated medical records). Moreover, restricting inclusion to studies with a sample size greater than 1000 potentially introduces selection bias, leading to the omission of valuable evidence from smaller experimental studies.

Strengths and Limitations

To the best of our knowledge, this review is among the first comprehensive examinations of media-based vaccine communication strategies aimed at enhancing vaccine adherence. It is methodologically robust, drawing on a very large evidence base and integrating both traditional and digital media influences on vaccination. It advances a precision-focused communication model shaped by key sociodemographic factors, and its strict adherence to PRISMA, PROSPERO, and multiple quality-assessment tools ensures a transparent and reliable synthesis.

Nevertheless, several limitations should be acknowledged. First, many studies assess vaccination intentions rather than actual behaviors and often exclude booster doses, which are critical for successful immunization programs. Second, the predominance of cross-sectional designs restricts causal inference. Experimental studies are relatively few, comprising only 8 out of the 36 [60-67] included studies. RCTs generally showed a low risk of bias in randomization and intervention delivery, supporting the evidence that certain communication strategies can positively influence vaccination attitudes, although confounding and self-reported outcomes require cautious interpretation. Third, data collected from specific online platforms introduce selection bias due to participants' distinct sociodemographic and technological profiles. Fourth, publication bias may have occurred, as studies reporting significant communication effects are more likely to be published. To mitigate this risk, strict inclusion criteria and comprehensive quality assessments were applied. Finally, although this review included a broad set of relevant studies, some pertinent works may still have been missed.

Conclusions

In conclusion, by synthesizing large-scale, fragmented evidence on the correlation between media and vaccine acceptance, and highlighting the role of sociodemographic determinants, this review moves beyond isolated analyses, offering a systematic and comprehensive examination of how multichannel media strategies can be leveraged to mitigate vaccine hesitancy and bolster adherence.

This work departs from previous studies in its dual focus on both traditional and digital media through the lens of sociodemographic moderation. By identifying that tailored communication—particularly through ethnic identification, storytelling, and prebunking—can overcome deeply rooted barriers, it provides a critical framework for future public health interventions.

To advance the field, future research priorities must shift toward rigorous longitudinal and experimental designs—particularly in low-income settings—and the standardized evaluation of emerging tools to validate causal links between communication and behavioral change.

The results of this review may provide starting points for orienting communication strategies toward a "precision" model. In this sense, profiling sociodemographic factors may be applied in designing strategies tailored to recipients' characteristics and cultural contexts.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the Guidelines for the Assessment of Intervention Description and Training (GAIDeT) taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing and translation. The GAI tool used was ChatGPT 4.0 (OpenAI). Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Data Availability

A template reporting the extracted data may be available upon motivated request and approval by the research team.

Authors' Contributions

Conceptualization: AM (lead), FL (equal), VN (equal), GM (equal)

XSL•FO
RenderX

Data curation: VN (lead), PC (equal), GM (equal)
Formal analysis: FL (lead), VN (equal)
Investigation: AM (lead), FL (equal), VN (equal), PC (equal)
Methodology: AM (lead), FL (equal), VN (equal)
Project administration: AM
Resources: AM (lead), VN (supporting), PC (supporting)
Supervision: AM
Validation: AM (lead), FL (supporting), VN (supporting), GM (supporting)
Visualization: FL (lead), VN (equal), GM (equal), PC (supporting)
Writing – original draft: VN (lead), FL (equal), AM (equal)
Writing – review & editing: FL (lead), AM (equal), VN (equal), PC (equal), GM (equal)

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[DOCX File , 34 KB - jmir_v28i1e74280_app1.docx]

Multimedia Appendix 2
Full search strategy.
[DOCX File , 16 KB - jmir_v28i1e74280_app2.docx]

References

1. Batteux E, Mills F, Jones LF, Symons C, Weston D. The effectiveness of interventions for increasing COVID-19 vaccine uptake: a systematic review. Vaccines (Basel) 2022 Mar 03;10(3):10030386 [FREE Full text] [doi: 10.3390/vaccines10030386] [Medline: 35335020]
2. Calabrò GE, Vitale F, Rizzo C, Pugliese A, Boccalini S, Bechini A, et al. [The new 15-valent pneumococcal conjugate vaccine for the prevention of infections in pediatric age: a Health Technology Assessment]. J Prev Med Hyg 2023 Mar;64(1 Suppl 1):E1-E160 [FREE Full text] [doi: 10.15167/2421-4248/jpmh2023.64.1s1] [Medline: 37655211]
3. Amodio E, D'Anna A, Verso MG, Leonforte F, Genovese D, Vitale F. Rotavirus vaccination as a public health strategy to reduce the burden of hospitalization: the field experience of Italy (2008-2018). J Med Virol 2023 Aug;95(8):e29000. [doi: 10.1002/jmv.29000] [Medline: 37515481]
4. Dubé E, Vivion M, MacDonald NE. Vaccine hesitancy, vaccine refusal and the anti-vaccine movement: influence, impact and implications. Expert Rev Vaccines 2015 Jan;14(1):99-117. [doi: 10.1586/14760584.2015.964212] [Medline: 25373435]
5. Dubé E, Laberge C, Guay M, Bramadat P, Roy R, Bettinger J. Vaccine hesitancy: an overview. Hum Vaccin Immunother 2013 Aug;9(8):1763-1773 [FREE Full text] [doi: 10.4161/hv.24657] [Medline: 23584253]
6. Lorini C, Santomauro F, Donzellini M, Capecchi L, Bechini A, Boccalini S, et al. Health literacy and vaccination: a systematic review. Hum Vaccin Immunother 2018 Feb 01;14(2):478-488 [FREE Full text] [doi: 10.1080/21645515.2017.1392423] [Medline: 29048987]
7. Borges do Nascimento IJ, Pizarro AB, Almeida JM, Azzopardi-Muscat N, Gonçalves MA, Björklund M, et al. Infodemics and health misinformation: a systematic review of reviews. Bull World Health Organ 2022 Oct 01;100(9):544-561 [FREE Full text] [doi: 10.2471/BLT.21.287654] [Medline: 36062247]
8. Xu J. Does the medium matter? A meta-analysis on using social media vs. traditional media in crisis communication. Public Relations Review 2020 Nov;46(4):101947. [doi: 10.1016/j.pubrev.2020.101947]
9. Dubé E, Gagnon D, Vivion M. Optimizing communication material to address vaccine hesitancy. Can Commun Dis Rep 2020 Feb 06;46(2-3):48-52 [FREE Full text] [doi: 10.14745/ccdr.v46i23a05] [Medline: 32167088]
10. Jarrett C, Wilson R, O'Leary M, Eckersberger E, Larson HJ, SAGE Working Group on Vaccine Hesitancy. Strategies for addressing vaccine hesitancy - a systematic review. Vaccine 2015 Aug 14;33(34):4180-4190 [FREE Full text] [doi: 10.1016/j.vaccine.2015.04.040] [Medline: 25896377]
11. Giguère A, Légaré F, Grimshaw J, Turcotte S, Fiander M, Grudniewicz A, et al. Printed educational materials: effects on professional practice and healthcare outcomes. Cochrane Database Syst Rev 2012 Oct 17;10:CD004398. [doi: 10.1002/14651858.CD004398.pub3] [Medline: 23076904]
12. Goldstein S, MacDonald NE, Guirguis S, SAGE Working Group on Vaccine Hesitancy. Health communication and vaccine hesitancy. Vaccine 2015 Aug 14;33(34):4212-4214 [FREE Full text] [doi: 10.1016/j.vaccine.2015.04.042] [Medline: 25896382]
13. Steffens MS, Dunn AG, Leask J, Wiley KE. Using social media for vaccination promotion: practices and challenges. Digit Health 2020 Nov 03;6:2055207620970785 [FREE Full text] [doi: 10.1177/2055207620970785] [Medline: 35173976]

XSL•FO
RenderX

14. Limaye RJ, Opel DJ, Dempsey A, Ellingson M, Spina C, Omer SB, et al. Communicating with vaccine-hesitant parents: a narrative review. Acad Pediatr 2021 May;21(4S):S24-S29. [doi: 10.1016/j.acap.2021.01.018] [Medline: 33958087]
15. Chou WS, Budenz A. Considering emotion in COVID-19 vaccine communication: addressing vaccine hesitancy and fostering vaccine confidence. Health Commun 2020 Dec;35(14):1718-1722. [doi: 10.1080/10410236.2020.1838096] [Medline: 33124475]
16. Agergaard TE, Smith ME, Gammelby AK, Tørring ML, Nielsen KH. Complexity and controversy in media coverage of Human papillomavirus (HPV) vaccination: a qualitative content analysis of news coverage in Denmark 2008–2018. Front Commun 2023 Mar 2;8:e1032460. [doi: 10.3389/fcomm.2023.1032460]
17. Catalan-Matamoros D, Peñafiel-Saiz C. How is communication of vaccines in traditional media: a systematic review. Perspect Public Health 2019 Jan;139(1):34-43. [doi: 10.1177/1757913918780142] [Medline: 29877122]
18. Larson HJ, Lin L, Goble R. Vaccines and the social amplification of risk. Risk Anal 2022 Jul;42(7):1409-1422 [FREE Full text] [doi: 10.1111/risa.13942] [Medline: 35568963]
19. Omer SB, Amin AB, Limaye RJ. Communicating about vaccines in a fact-resistant world. JAMA Pediatr 2017 Oct 01;171(10):929. [doi: 10.1001/jamapediatrics.2017.2219]
20. Olson O, Berry C, Kumar N. Addressing parental vaccine hesitancy towards childhood vaccines in the United States: a systematic literature review of communication interventions and strategies. Vaccines (Basel) 2020 Oct 08;8(4):590 [FREE Full text] [doi: 10.3390/vaccines8040590] [Medline: 33049956]
21. Kahan DM. Social science. A risky science communication environment for vaccines. Science 2013 Oct 04;342(6154):53-54. [doi: 10.1126/science.1245724] [Medline: 24092722]
22. Whitehead HS, French CE, Caldwell DM, Letley L, Mounier-Jack S. A systematic review of communication interventions for countering vaccine misinformation. Vaccine 2023 Jan 27;41(5):1018-1034 [FREE Full text] [doi: 10.1016/j.vaccine.2022.12.059] [Medline: 36628653]
23. White P, Alberti H, Rowlands G, Tang E, Gagnon D, Dubé. Vaccine hesitancy educational interventions for medical students: a systematic narrative review in Western countries. Hum Vaccin Immunother 2024 Dec 31;20(1):2397875 [FREE Full text] [doi: 10.1080/21645515.2024.2397875] [Medline: 39323010]
24. Hudson A, Montelpare WJ. Predictors of vaccine hesitancy: implications for COVID-19 public health messaging. Int J Environ Res Public Health 2021 Jul 29;18(15):ijerph18158054 [FREE Full text] [doi: 10.3390/ijerph18158054] [Medline: 34360345]
25. Truong J, Bakshi S, Wasim A, Ahmad M, Majid U. What factors promote vaccine hesitancy or acceptance during pandemics? A systematic review and thematic analysis. Health Promot Int 2022 Feb 17;37(1):daab105. [doi: 10.1093/heapro/daab105] [Medline: 34244738]
26. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
27. Moola S, Munn Z, Tufanaru C, Aromataris E, Sears K, Sfetcu R, et al. Chapter 7: Systematic reviews of etiology risk. In: Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis. Adelaide, SA, Australia: The Joanna Briggs Institute; Mar 01, 2024:252-313.
28. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
29. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016 Oct 12;355:i4919 [FREE Full text] [Medline: 27733354]
30. McGuinness LA, Higgins JPT. Risk-of-bias VISualization (robvis): an R package and Shiny web app for visualizing risk-of-bias assessments. Res Synth Methods 2020 Apr 26:55-61 [FREE Full text] [doi: 10.1002/jrsm.1411] [Medline: 32336025]
31. Al Asmri M, Haque MS, Parle J. A Modified Medical Education Research Study Quality Instrument (MMERSQI) developed by Delphi consensus. BMC Med Educ 2023 Jan 25;23(1):63 [FREE Full text] [doi: 10.1186/s12909-023-04033-6] [Medline: 36698117]
32. De Giorgio A, Kuvačić G, Maleš D, Vecchio I, Tornali C, Ishac W, et al. Willingness to receive COVID-19 booster vaccine: associations between green-pass, social media information, anti-vax beliefs, and emotional balance. Vaccines (Basel) 2022 Mar 21;10(3):10030481 [FREE Full text] [doi: 10.3390/vaccines10030481] [Medline: 35335113]
33. Garg A, Nyitray AG, Roberts JR, Shungu N, Ruggiero KJ, Chandler J, et al. Consumption of health-related videos and human papillomavirus awareness: cross-sectional analyses of a US national survey and YouTube from the urban-rural context. J Med Internet Res 2024 Jan 15;26:e49749 [FREE Full text] [doi: 10.2196/49749] [Medline: 38224476]
34. Garrett C, Qiao S, Li X. The role of social media in knowledge, perceptions, and self-reported adherence toward COVID-19 prevention guidelines: cross-sectional study. JMIR Infodemiology 2024 Feb 16;4:e44395 [FREE Full text] [doi: 10.2196/44395] [Medline: 38194493]
35. Hwang J, Shah DV. Health information sources, perceived vaccination benefits, and maintenance of childhood vaccination schedules. Health Commun 2019 Oct;34(11):1279-1288. [doi: 10.1080/10410236.2018.1481707] [Medline: 29869894]

XSL•FO
RenderX

36. Jia X, Ahn S, Morgan SE. The role of social media messages and content creators in shaping COVID-19 vaccination intentions. Front Digit Health 2025;7:1448884 [FREE Full text] [doi: 10.3389/fdgth.2025.1448884] [Medline: 40196399]
37. Jin Q, Raza SH, Yousaf M, Munawar R, Shah AA, Hassan S, et al. Ingraining polio vaccine acceptance through public service advertisements in the digital era: the moderating role of misinformation, disinformation, fake news, and religious fatalism. Vaccines (Basel) 2022 Oct 17;10(10):10101733 [FREE Full text] [doi: 10.3390/vaccines10101733] [Medline: 36298598]
38. Jung M, Lin L, Viswanath K. Effect of media use on mothers' vaccination of their children in sub-Saharan Africa. Vaccine 2015 May 21;33(22):2551-2557. [doi: 10.1016/j.vaccine.2015.04.021] [Medline: 25896379]
39. Jung M. The effect of maternal decisional authority on children's vaccination in East Asia. PLoS One 2018 Jul 12;13(7):e0200333 [FREE Full text] [doi: 10.1371/journal.pone.0200333] [Medline: 30001397]
40. Khanijahani A, Calhoun B, Kiel J. Internet use habits and influenza vaccine uptake among US adults: results from seven years (2012-2018) of the National Health Interview Survey. Public Health 2021 Jun;195:76-82. [doi: 10.1016/j.puhe.2021.04.007] [Medline: 34062275]
41. Kim J, Jung M. Associations between media use and health information-seeking behavior on vaccinations in South Korea. BMC Public Health 2017 Sep 11;17(1):700 [FREE Full text] [doi: 10.1186/s12889-017-4721-x] [Medline: 28893212]
42. Kim Y, Kim J, Li Y. Association of sources of COVID-19 information with vaccine uptake, preventive behaviors, and perceived severity of COVID-19. J Am Med Dir Assoc 2023 Aug;24(8):1143-1150 [FREE Full text] [doi: 10.1016/j.jamda.2023.05.024] [Medline: 37400059]
43. Kobayashi T, Nishina Y, Tomoi H, Harada K, Tanaka K, Matsumoto E, et al. Corowa-kun: a messenger app chatbot delivers COVID-19 vaccine information, Japan 2021. Vaccine 2022 Jul 30;40(32):4654-4662 [FREE Full text] [doi: 10.1016/j.vaccine.2022.06.007] [Medline: 35750541]
44. Lv M, Fang R, Wu J, Pang X, Deng Y, Lei T, et al. The free vaccination policy of influenza in Beijing, China: the vaccine coverage and its associated factors. Vaccine 2016 Apr 19;34(18):2135-2140. [doi: 10.1016/j.vaccine.2016.02.032] [Medline: 26917011]
45. Melovic B, Jaksic Stojanovic A, Vulic TB, Dudic B, Benova E. The impact of online media on parents' attitudes toward vaccination of children-social marketing and public health. Int J Environ Res Public Health 2020 Aug 11;17(16):1-27 [FREE Full text] [doi: 10.3390/ijerph17165816] [Medline: 32796740]
46. Nah S, Williamson LD, Kahlor LA, Atkinson L, Ntang-Beb J, Upshaw SJ. COVID-19 vaccine hesitancy in Cameroon: the role of medical mistrust and social media use. J Health Commun 2023 Sep 02;28(9):619-632. [doi: 10.1080/10810730.2023.2250287] [Medline: 37622325]
47. Nah S, Williamson L, Kahlor L, Atkinson L, Upshaw S, Ntang-Beb JL. The roles of social media use and medical mistrust in Black Americans' COVID-19 vaccine hesitancy: the RISP model perspective. Health Commun 2024 Aug;39(9):1833-1846. [doi: 10.1080/10410236.2023.2244169] [Medline: 37551159]
48. Piltch-Loeb R, Savoia E, Goldberg B, Hughes B, Verhey T, Kayyem J, et al. Examining the effect of information channel on COVID-19 vaccine acceptance. PLoS One 2021;16(5):e0251095 [FREE Full text] [doi: 10.1371/journal.pone.0251095] [Medline: 33979370]
49. Recio-Román A, Recio-Menéndez M, Román-González MV. Influence of media information sources on vaccine uptake: the full and inconsistent mediating role of vaccine hesitancy. Computation 2023 Oct 23;11(10):208. [doi: 10.3390/computation11100208]
50. Sakamoto M, Ishizuka R, Ozawa C, Fukuda Y. Health information and COVID-19 vaccination: beliefs and attitudes among Japanese university students. PLoS ONE 2022 Nov 9;17(11):e0277435. [doi: 10.1371/journal.pone.0277435]
51. Sunohara S, Asakura TR, Kimura T, Saijo M, Tamakoshi A. Traditional and social media usage associated with COVID-19 vaccine uptake in Sapporo, Japan. Asia Pac J Public Health 2024 Mar 30;36(4):358-365. [doi: 10.1177/10105395241240952]
52. Teoh D, Shaikh R, Schnaith A, Lou E, McRee A, Nagler RH, et al. Evaluation of graphic messages to promote human papillomavirus vaccination among young adults: a statewide cross-sectional survey. Prev Med Rep 2019 Mar;13:256-261 [FREE Full text] [doi: 10.1016/j.pmedr.2019.01.002] [Medline: 30723659]
53. Thompson EL, Preston SM, Francis JKR, Rodriguez SA, Pruitt SL, Blackwell J, et al. Social media perceptions and internet verification skills associated with human papillomavirus vaccine decision-making among parents of children and adolescents: cross-sectional survey. JMIR Pediatr Parent 2022 Sep 14;5(3):e38297 [FREE Full text] [doi: 10.2196/38297] [Medline: 36103216]
54. Wada K, Smith DR. Mistrust surrounding vaccination recommendations by the Japanese government: results from a national survey of working-age individuals. BMC Public Health 2015 Apr 26;15:426 [FREE Full text] [doi: 10.1186/s12889-015-1772-8] [Medline: 25928236]
55. Wamba S, Guthrie C, Queiroz M, Twinomurinzi H. Digital technologies and COVID-19 vaccine acceptance: evidence from France and South Africa. Journal of Global Information Management 2023;31(1):333611. [doi: 10.4018/JGIM.333611]
56. Yoneoka D, Eguchi A, Nomura S, Kawashima T, Tanoue Y, Murakami M, et al. Identification of optimum combinations of media channels for approaching COVID-19 vaccine unsure and unwilling groups in Japan. Lancet Reg Health West Pac 2022 Jan;18:100330 [FREE Full text] [doi: 10.1016/j.lanwpc.2021.100330] [Medline: 34927110]

XSL•FO
RenderX

57. Al Zahrani E. The role of healthcare leaders in promoting vaccine acceptance in Saudi Arabia. JHL 2024 Jul;Volume 16:279-286. [doi: 10.2147/jhl.s470522]
58. Morris ME, Aguilera A. Mobile, social, and wearable computing and the evolution of psychological practice. Prof Psychol Res Pr 2012 Dec;43(6):622-626 [FREE Full text] [doi: 10.1037/a0029041] [Medline: 25587207]
59. Wilson SL, Wiysonge C. Social media and vaccine hesitancy. BMJ Glob Health 2020 Oct;5(10):004206. [doi: 10.1136/bmjgh-2020-004206] [Medline: 33097547]
60. Evans WD, Bingenheimer JB, Long M, Ndiaye K, Donati D, Rao NM, et al. Outcomes of a social media campaign to promote COVID-19 vaccination in Nigeria. PLoS One 2023;18(9):e0290757 [FREE Full text] [doi: 10.1371/journal.pone.0290757] [Medline: 37713381]
61. Kim S, Lilani A, Redemptus C, Campana K, Tozan Y. A pre-post evaluation study of a social media-based COVID-19 communication campaign to improve attitudes and behaviors toward COVID-19 vaccination in Tanzania. PLoS One 2024;19(5):e0300206 [FREE Full text] [doi: 10.1371/journal.pone.0300206] [Medline: 38709712]
62. Pavoncello V, Kislaya I, Andrianarimanana DK, Marchese V, Rakotomalala R, Rasamoelina T, et al. Optimizing vaccine uptake in sub-Saharan Africa: a collaborative COVID-19 vaccination campaign in Madagascar using an adaptive approach. Implement Sci 2025 Jan 09;20(1):2 [FREE Full text] [doi: 10.1186/s13012-024-01412-5] [Medline: 39789551]
63. Schuh HB, Rimal RN, Breiman RF, Orton PZ, Dudley MZ, Kao L, et al. Evaluation of online videos to engage viewers and support decision-making for COVID-19 vaccination: how narratives and race/ethnicity enhance viewer experiences. Front Public Health 2023;11:1192676 [FREE Full text] [doi: 10.3389/fpubh.2023.1192676] [Medline: 37670826]
64. Daley MF, Narwaney KJ, Shoup JA, Wagner NM, Glanz JM. Addressing parents' vaccine concerns: a randomized trial of a social media intervention. Am J Prev Med 2018 Jul;55(1):44-54. [doi: 10.1016/j.amepre.2018.04.010] [Medline: 29773490]
65. Geniole SN, Bird BM, Witzel A, McEvoy JT, Proietti V. Preliminary evidence that brief exposure to vaccination-related internet memes may influence intentions to vaccinate against COVID-19. Comput Human Behav 2022 Jun;131:107218 [FREE Full text] [doi: 10.1016/j.chb.2022.107218] [Medline: 35125639]
66. Lee KY, Dabak SV, Kong VH, Park M, Kwok SLL, Silzle M, et al. Effectiveness of chatbots on COVID vaccine confidence and acceptance in Thailand, Hong Kong, and Singapore. NPJ Digit Med 2023 May 25;6(1):96 [FREE Full text] [doi: 10.1038/s41746-023-00843-6] [Medline: 37231110]
67. Schmid P, Betsch C. Benefits and pitfalls of debunking interventions to counter mRNA vaccination misinformation during the COVID-19 pandemic. Sci Commun 2022 Oct;44(5):531-558 [FREE Full text] [doi: 10.1177/10755470221129608] [Medline: 38603361]
68. Ashfield S, Donelle L, Tryphonopoulos P, Dubé È, Smith M. Digital health literacy, vaccine information sources, and vaccine acceptance among parents in Ontario: quantitative findings from a mixed methods study. PLOS Glob Public Health 2024;4(5):e0003154 [FREE Full text] [doi: 10.1371/journal.pgph.0003154] [Medline: 38758822]
69. Jejaw M, Demissie KA, Teshale G, Geberu DM, Tiruneh MG, Hagos A, et al. Estimated effect of media use on mothers' vaccination of their children in Sub-Saharan Africa: a quasi-experimental propensity score matching analysis using DHS data. BMC Public Health 2025 May 31;25(1):2018. [doi: 10.1186/s12889-025-23258-2] [Medline: 40450220]
70. Grygarová D, Kožený J, Tišanská L, Havlík M, Horáček J. Trust in official information as a key predictor of COVID-19 vaccine acceptance: evidence from a Czech longitudinal survey study. BMC Public Health 2025 Feb 25;25(1):770. [doi: 10.1186/s12889-025-21988-x] [Medline: 40001023]
71. Cascini F, Pantovic A, Al-Ajlouni YA, Failla G, Puleo V, Melnyk A, et al. Social media and attitudes towards a COVID-19 vaccination: a systematic review of the literature. EClinicalMedicine 2022 Jun;48:101454 [FREE Full text] [doi: 10.1016/j.eclinm.2022.101454] [Medline: 35611343]
72. Burki T. Vaccine misinformation and social media. The Lancet Digital Health 2019 Oct;1(6):e258-e259. [doi: 10.1016/S2589-7500(19)30136-0]
73. Puri N, Coomes EA, Haghbayan H, Gunaratne K. Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases. Hum Vaccin Immunother 2020 Jul 21:1-8. [doi: 10.1080/21645515.2020.1780846] [Medline: 32693678]
74. Skafle I, Nordahl-Hansen A, Quintana DS, Wynn R, Gabarron E. Misinformation about COVID-19 vaccines on social media: rapid review. J Med Internet Res 2022 Aug 04;24(8):e37367 [FREE Full text] [doi: 10.2196/37367] [Medline: 35816685]
75. van Dijk J, Hacker K. The digital divide as a complex and dynamic phenomenon. The Information Society 2003 Sep;19(4):315-326. [doi: 10.1080/01972240309487]
76. Piticchio T, Savarino F, Volpe S, Prinzi A, Costanzo G, Gamarra E, et al. Inflammatory profile assessment in a highly selected athyreotic population undergoing controlled and standardized hypothyroidism. Biomedicines 2024 Jan 22;12(1):N.A-N.A [FREE Full text] [doi: 10.3390/biomedicines12010239] [Medline: 38275410]
77. Hawkes S, Buse K. Gender and global health: evidence, policy, and inconvenient truths. Lancet 2013 May 18;381(9879):1783-1787 [FREE Full text] [doi: 10.1016/S0140-6736(13)60253-6] [Medline: 23683645]
78. Zintel S, Flock C, Arbogast AL, Forster A, von Wagner C, Sieverding M. Gender differences in the intention to get vaccinated against COVID-19: a systematic review and meta-analysis. Z Gesundh Wiss 2022 Jan 07:1-25 [FREE Full text] [doi: 10.1007/s10389-021-01677-w] [Medline: 35018277]

XSL•FO
RenderX

79. Toshkov D. Explaining the gender gap in COVID-19 vaccination attitudes. Eur J Public Health 2023 Jun 01;33(3):490-495. [doi: 10.1093/eurpub/ckad052] [Medline: 37178211]
80. Grant MD, Markowitz DM, Sherman DK, Flores A, Dickert S, Eom K, et al. Ideological diversity of media consumption predicts COVID-19 vaccination. Sci Rep 2024 Nov 22;14(1):28948 [FREE Full text] [doi: 10.1038/s41598-024-77408-4] [Medline: 39578487]
81. D'Innocenzo S, Biagi C, Lanari M. Obesity and the Mediterranean diet: a review of evidence of the role and sustainability of the Mediterranean diet. Nutrients 2019 Jun 09;11(6):1306 [FREE Full text] [doi: 10.3390/nu11061306] [Medline: 31181836]
82. Oteri V, Galeano F, Panebianco S, Piticchio T, Le Moli R, Frittitta L, et al. Influence of Mediterranean diet on sexual function in people with metabolic syndrome: a narrative review. Nutrients 2024 Oct 06;16(19):3397. [doi: 10.3390/nu16193397] [Medline: 39408364]
83. Friis K, Lasgaard M, Rowlands G, Osborne RH, Maindal HT. Health literacy mediates the relationship between educational attainment and health behavior: a Danish population-based study. J Health Commun 2016;21(sup2):54-60. [doi: 10.1080/10810730.2016.1201175] [Medline: 27668691]
84. Fiorilla C, Contarino F, Sorrentino M, Bella F, Leonforte F, Nicosia V, et al. Socioeconomic deprivation and human papillomavirus vaccination inequities: evidence from Southern Italy. Vaccine 2026 Jan 01;69:127986. [doi: 10.1016/j.vaccine.2025.127986] [Medline: 41223688]
85. Gaysynsky A, Heley K, Chou WS. An overview of innovative approaches to support timely and agile health communication research and practice. Int J Environ Res Public Health 2022 Nov 16;19(22):N.A-N.A [FREE Full text] [doi: 10.3390/ijerph192215073] [Medline: 36429796]
86. Hakim H, Driedger SM, Gagnon D, Chevrier J, Roch G, Dubé E, et al. Digital gamification tools to enhance vaccine uptake: scoping review. JMIR Serious Games 2024 Feb 29;12:e47257 [FREE Full text] [doi: 10.2196/47257] [Medline: 38421688]
87. Swire-Thompson B, DeGutis J, Lazer D. Searching for the backfire effect: measurement and design considerations. J Appl Res Mem Cogn 2020 Sep;9(3):286-299 [FREE Full text] [doi: 10.1016/j.jarmac.2020.06.006] [Medline: 32905023]
88. Vivion M, Anassour Laouan Sidi E, Betsch C, Dionne M, Dubé E, Driedger SM, et al. Prebunking messaging to inoculate against COVID-19 vaccine misinformation: an effective strategy for public health. Journal of Communication in Healthcare 2022 Mar 04;15(3):232-242. [doi: 10.1080/17538068.2022.2044606]
89. Andsager JL, Bemker V, Choi H, Torwel V. Perceived similarity of exemplar traits and behavior. Communication Research 2016 Jun 30;33(1):3-18. [doi: 10.1177/0093650205283099]

Abbreviations

M-MERSQI: Modified Medical Education Research Study Quality Instrument
PICO: Population, Intervention/Exposure, Comparison, and Outcome
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO: International Prospective Register of Systematic Reviews
RCT: randomized controlled trial
RoB: Risk-of-Bias 2.0 tool for randomized trials
ROBINS: Risk of Bias in Non-randomized Studies of Interventions
WHO: World Health Organization

Edited by S Brini; submitted 21.Mar.2025; peer-reviewed by A Law, T Piticchio, E Vitale; comments to author 02.Sep.2025; accepted 09.Jan.2026; published 28.Apr.2026.

Please cite as:
Leonforte F, Nicosia V, Comite P, Morlino G, Mistretta A
Media Exposure and Its Association With Vaccine Attitudes, Intentions, and Hesitancy: Systematic Review
J Med Internet Res 2026;28:e74280
URL: https://www.jmir.org/2026/1/e74280
doi:10.2196/74280

©Francesco Leonforte, Vito Nicosia, Paola Comite, Giustino Morlino, Antonio Mistretta. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Current Practices and a Novel Operational Framework for Planning Research on Digital Health Promotion Interventions From Development to Implementation: Scoping Review

Claire Collin[1], MSc, PhD; Corinne Alberti[1,2], MD, PhD; Philippe Martin[1,2,3], MPH, PhD; Graham Moore[4,5], MSc, PhD; Clara Eyraud[1], MPH; Enora Le Roux[1,2,6], MSc, PhD

[1]Université Paris Cité, Inserm, ECEVE, Paris, France

[2]AP-HP, Hôpital Universitaire Robert Debré, Clinical Epidemiology Unit, CIC 1426, Paris, France

[3]UR14 - Sexual and Reproductive Health and Rights, Institut National d'Etudes Démographiques (INED), Aubervilliers, France

[4]Centre for Development, Evaluation, Complexity and Implementation in Public Health Improvement (DECIPHer), School of Social Sciences, Cardiff University, Cardiff, United Kingdom

[5]Wolfson Centre for Young People's Mental Health, Cardiff University, Cardiff, United Kingdom

[6]University Centre for Adolescent and Young Adult Health, Fondation Santé des Etudiants de France, Paris, France

Corresponding Author:
Claire Collin, MSc, PhD
Université Paris Cité, Inserm, ECEVE
10 avenue de Verdun
Paris, 75010
France
Phone: 33 643968895
Email: ccollin.contact@gmail.com

Abstract

Background: The UK Medical Research Council's Guidance on Developing and Evaluating Complex Interventions (MRC GDECI) outlines a 4-phase framework for structuring research programs on interventions: development, feasibility, evaluation, and implementation. However, it provides limited practical direction on how researchers should select which phases to conduct or determine when and whether to progress between phases. This gap is particularly challenging in the context of digital health interventions (DHIs), given their fast-paced and rapidly evolving nature.

Objective: This scoping review examined the research phases conducted, how researchers progressed through them, and the intervention characteristics associated with overall program structure and duration in DHI research, to inform the design of future research programs.

Methods: We searched PubMed, Embase, CINAHL, PsycINFO, and ClinicalTrials.gov to identify complex DHIs promoting health among adolescents and young adults, implemented between 2017 and 2026, for which at least 2 phases of the MRC GDECI were reported, including the evaluation phase. For each eligible intervention, all related protocols, preprints, and published articles were retrieved to reconstruct the full research program. For each program, we analyzed the presence of each research phase, its organization (ie, phase arrangements), and the mechanisms guiding progression between phases (ie, progression mechanisms). Phase-specific and overall program durations were recorded.

Results: A total of 31 research programs, covering 31 interventions and reported in 130 articles, were included. Development, feasibility, evaluation, and implementation phases were reported in 26, 23, 31, and 7 research programs, respectively. Three types of phase arrangements were identified: sequential, iterative, and overlapping. Progression mechanisms between phases included automatic progression, conditional progression based on researchers' appraisal of findings without prespecified criteria, and progression based on predefined quantitative criteria. Six main research program structures were observed, combining phase arrangements and progression mechanisms. Iterative arrangements were most common, observed in 22 research programs, followed by overlapping (n=10) and strictly sequential structures (n=7). Most progressions relied on researchers' appraisal of findings without prespecified criteria. Justifications for phase iteration, omission, or progression decisions were rarely reported. The median program duration was 5.8 (IQR 3.8-6.6) years (n=13). Based on these findings, a novel 4-step operational framework and visualization tools were developed to guide the design and planning of DHIs, highlighting key considerations for each step, as well as the strengths, limitations, and risks associated with each phase arrangement and progression mechanism.

XSL•FO
RenderX

Conclusions: This scoping review is the first to systematically examine phase arrangements and progression mechanisms in DHI research programs. Beyond descriptive reporting, it provides a conceptualization of research program structures and offers a flexible operational framework to support the concrete implementation of the MRC GDECI. Greater explicitness in decisions about program structure may enhance methodological rigor, reduce research waste, and improve the integrity and reproducibility of interventions.

Trial Registration: PROSPERO CRD42023401979; https://tinyurl.com/mvc265y3

(J Med Internet Res 2026;28:e82611) doi:10.2196/82611

KEYWORDS

complex interventions; digital health; health promotion; methodological framework; research program design

Introduction

Complex interventions constitute a significant part of health and social care and require robust methodological approaches to produce reliable evidence for decision-making. To address this, the UK Medical Research Council (MRC) developed the Guidance on Developing and Evaluating Complex Interventions (GDECI), a foundational framework providing theoretical and methodological recommendations across 4 research phases: development, feasibility, evaluation, and implementation [1]. While the MRC GDECI acknowledges that all phases may not be necessary, may not follow a strictly sequential order, and may be revisited when uncertainties persist, it offers limited practical recommendations on which phases to conduct or how to decide when and whether to progress between phases.

This absence of practical methodological direction, while allowing flexible application across diverse interventions, creates challenges for researchers in a context where little research has been conducted on the operationalization of the MRC GDECI in relation to these 2 aspects: phase selection and progression mechanisms between phases.

With regard to phase selection, the boundaries defining what each phase entails have been reported to be insufficiently delineated for the development [2], feasibility [3], and implementation [4] phases, respectively, resulting in blurred distinctions and reduced transparency in reporting [5]. Development phases, in particular, have been shown to be underreported or mixed within feasibility studies, limiting insight into how development processes shape subsequent phases and outcomes [2,5]. Similarly, the proliferation of terminology used to describe feasibility studies has hindered clear differentiation between feasibility, development, and evaluation phases [6]. Persistent ambiguity also remains regarding whether modifications made during feasibility work constitute a return to development or represent adaptations that can appropriately occur within the feasibility phase itself [3]. More broadly, insufficient justification for omitting, merging, or underreporting phases has been highlighted to deprive the scientific community of valuable learning opportunities (both in terms of good practice and failed attempts), compromising transparency, evidence synthesis, and field advancement [7,8], and potentially contributing to research waste [5].

With regard to progression mechanisms between phases, the literature is more developed in some areas, particularly in relation to progression from feasibility to evaluation. Indeed, studies have focused predominantly on the definition and use of predefined quantitative progression criteria to inform decisions about whether to proceed to a full-scale randomized controlled trial (RCT) [9-11]. In contrast, progression decisions in other instances, such as from development to feasibility, from evaluation to implementation, or within feasibility studies that do not rely on quantitative criteria, have received comparatively little attention. The absence of clearly articulated progression mechanisms has been associated with an increased risk of interventions with weaknesses prematurely advancing to later phases, ultimately failing to demonstrate meaningful effects [12-14]. Conversely, prolonged feasibility testing before a full evaluation may impede the timely delivery of evidence to end users, in a context where public health research is expected to produce findings rapidly and efficiently [15,16]. Taken together, the limited documentation of progression mechanisms, particularly beyond feasibility to evaluation, underscores the need for a comprehensive examination of how progression decisions are currently made and reported across intervention research programs.

Among complex interventions, digital health interventions (DHIs) represent a rapidly growing category, typically comprising multiple components addressing simultaneous aims (eg, education, behavior change, or monitoring, and peer engagement) [17,18]. DHIs are increasingly deployed among adolescents and young adults (AYAs), who frequently engage with digital platforms for health information [19]. In this context, DHIs have emerged as promising, innovative, and relatively low-cost approaches to promote healthy behaviors [20,21].

However, the inherently perishable nature of DHIs introduces unique challenges when applying the MRC GDECI. Completing all research phases raises viability concerns, as interventions may become outdated by the time the full research cycle is completed [22]. This is particularly concerning given that traditional phased approaches from intervention development to implementation often span several years, with few interventions ultimately adopted in routine practice [23]. For DHIs targeting AYAs, whose digital preferences [24] and behaviors [25] evolve rapidly, this temporal mismatch between research timelines and technological relevance poses substantial difficulties.

Given the absence of practical guidance on operationalizing the MRC GDECI for different types of complex interventions, notably DHIs, investigation of current research practices is needed to inform complementary guidance. This study,

XSL•FO
RenderX

therefore, aimed to examine how researchers conduct phases and progress between them in digital health research. Taking DHIs promoting health among AYAs as a study framework, our specific objectives were to (1) identify the phases conducted in DHI research programs, (2) identify the mechanisms used to progress from one phase to another, (3) quantify the time required to complete programs from development to implementation, and (4) explore the relationship between intervention characteristics and program structure and duration. The ultimate goal of the study is to provide practical guidance on designing and conducting research programs that build on MRC recommendations by being tailored to DHIs.

Methods

Study Design

This study is part of a broader project based on a scoping review examining the methodology used to develop and evaluate DHIs for AYAs. The review followed PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [26] (see Multimedia Appendix 1 for the completed report and abstract checklists). Full methods are described in a prospectively registered PROSPERO (International Prospective Register of Systematic Reviews) protocol (CRD42023401979) and the primary article, which explored the selection, analysis, and interpretation of multiple outcomes in determining intervention success [27]. No deviations from the registered protocol occurred. However, given the breadth and richness of the extracted data, the analyses were structured into 2 complementary articles, each addressing distinct objectives. This study investigates research program design, focusing on phase arrangements, progression mechanisms, and program duration.

Eligibility Criteria

Eligible interventions were complex DHIs, defined as interventions comprising at least 2 distinct components (eg, videos, quizzes, and informational content) and engaging users throughout at least 2 different modalities, ranging from passive information delivery to interactive engagement with peers and professionals [1,18]. Interventions were required to be delivered exclusively through digital technologies (eg, computer programs, games, mobile apps, social media platforms, and text or voice messages) and to target AYAs, defined as populations with a mean or median participant age between 10 and 24 years, in accordance with the World Health Organization [28]. Eligible interventions had to address health promotion and primary prevention objectives, be implemented between 2017 and 2026, and be published in English or French. The 2017 start date was chosen to ensure the technological relevance of included DHIs in a rapidly evolving digital landscape. This date also coincides with the publication of several seminal papers in late 2016 that established key conceptual and methodological foundations for DHI development and evaluation, with potential downstream influence on intervention design, conduct, and reporting [17,20,22,29-32].

In addition, interventions required preprints or published articles that covered at least 2 of the 4 MRC GDECI phases, whether or not the authors explicitly referred to a phased approach. Phases could be identified through explicit labeling or inferred from commonly used terminology corresponding to the development, feasibility, evaluation, and implementation phases. At least 2 phases were necessary to allow examination of the progression mechanisms between phases. One of these phases had to be the evaluation phase, as the primary article from this review focused on outcome selection and the accuracy of conclusions in that phase. This eligibility criterion was applied to ensure that all 3 study objectives (ie, phase selection, progression mechanisms, and research program duration) could be addressed.

Information Sources

A total of 4 databases (PubMed, Embase, PsycINFO, and CINAHL) and 1 study registry (ClinicalTrials.gov) were individually searched on February 9, 2023, with updates on May 2, 2023, and January 2, 2024.

For each identified intervention, all associated protocols, published articles, and preprints were retrieved using three complementary approaches: (1) citation tracking using the “Cited By” tool in the PubMed database, (2) reference list screening of included articles, and (3) targeted searches using the intervention name or acronym as a search term across databases, to ensure comprehensive identification of relevant reports.

To capture any additional relevant records related to the included interventions published since January 2, 2024, additional searches were conducted on April 15, 2025, and January 7, 2026. Articles unavailable via open access, university subscriptions, or interlibrary loan were requested directly from the corresponding authors by email, and all eligible articles were successfully retrieved.

Search Strategy

The search strategy was iteratively developed by the research team in collaboration with the university library and comprised 5 groups of relevant keywords: “eHealth,” “intervention research,” “evaluation,” “health promotion and prevention,” and “adolescents and young adults.” This strategy was applied across all selected databases and registries. The completed PRISMA-S (Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension) checklist [33] and detailed search strategies for each database are provided in Multimedia Appendix 2.

Data Extraction

Overview

Title and abstract screening, full-text screening, and data extraction were performed independently by 2 researchers (CC and CE), with discrepancies resolved by a third researcher (ELX). All screening steps were conducted using Covidence systematic review software (Veritas Health Innovation). Data extraction was conducted using EpiData (version 3.1; EpiData Association).

For each research program, data were systematically extracted across 4 domains (see Multimedia Appendix 3 for the data charting form).

XSL•FO
RenderX

Phase Arrangements

Using MRC phase definitions [1] and commonly used synonyms in the DHI literature [34], we documented which phases were conducted (development, feasibility, evaluation, and implementation). In addition, we coded whether authors explicitly referenced the MRC GDECI and, if so, in relation to which phase. We also documented whether authors used the term "phase" in a manner consistent with MRC definitions to describe the studies conducted within their research programs, even when the MRC framework was not explicitly cited. Multimedia Appendix 4 details the labels used in the original articles and describes how these were mapped to the corresponding MRC phases.

For each program, we recorded the chronological order of phases as reported by the authors, as well as the presence of phase iteration and phase overlap. Phases were classified as "iterated" when (1) the authors explicitly reported the use of an iterative process during intervention development and (2) multiple pilot studies, efficacy or effectiveness trials, or implementation studies of the same intervention were conducted within the feasibility, evaluation, or implementation phases, respectively. Phases were classified as "overlapping" when (1) for the development and feasibility phases, both were conducted concurrently within a single "formative research" study and reported as a single research output, or when development and piloting activities were interwoven (eg, an initial development stage followed by a pilot study and subsequent redevelopment) and (2) for the evaluation and implementation phases, the authors explicitly reported the use of a hybrid trial design (either Type I or Type II) which by definition simultaneously evaluates the clinical intervention and the implementation strategy [35], or when implementation-related data were collected during the evaluation phase using the same participant sample.

Progression Mechanisms

We documented the decision-making processes reported for advancing from one phase to the next, based on authors' descriptions of methods in the Methods sections and their interpretation of findings in the Discussion sections of the included articles. The following four types of progression situations were identified: (1) authors do not discuss phase-specific results in relation to announced subsequent phases, or progression occurs despite unfavorable or inconclusive findings; (2) authors explicitly discuss phase results and justified progression to the next phase on the basis of positive findings and completion of the current phase, but without predefined progression criteria, relying primarily on a qualitative appraisal of results; (3) authors report the use of prespecified quantitative thresholds, often operationalized through a traffic-light system, to inform progression decisions; and (4) authors do not describe at all the next planned research phase.

Duration

Phase-specific and total research program durations were extracted from published articles when reported. Phase-specific duration was defined as the elapsed time between the start and end of each research phase. When phase durations were not reported in the articles, information from trial registries was used as a proxy for phase duration, including ClinicalTrials.gov ("Study Start" and "Study Completion"), International Standard Randomized Controlled Trial Number (ISRCTN; "Date of first enrollment" and "Study completion"), and Australian New Zealand Clinical Trials Registry (ANZCTR; "Date of first participant enrollment" and "Date of last data collection"). This approach was primarily applied to pilot RCTs corresponding to the feasibility phase and to RCTs or quasi-experimental studies corresponding to the evaluation phase, for which trial registrations are typically requested. Total research program duration was defined as the elapsed time from the start of the first phase to the completion of the final phase. Accordingly, total program duration could be calculated for interventions for which both a start date for the first phase and a completion date for the final phase were available, even when phase-specific durations within the research program were not reported.

Intervention Characteristics

Several intervention characteristics were recorded including whether the intervention was newly developed or adapted from an existing one (new vs adapted), content was personalized based on participants' characteristics or standardized (personalized vs standardized), content evolved during delivery or remained fixed (dynamic vs fixed), multiple digital technologies or a single one were used (multiple vs single technology), development was guided by behavioral theory (theory-driven vs empirically-driven), and a participatory approach was used (participatory vs nonparticipatory). We also documented the evaluation study design (RCT vs quasi-experimental design; as these were the only designs used to evaluate eligible interventions) and funding characteristics (amount in US dollars and source: public, private, or mixed).

Consistent with the specific objectives of this scoping review and PRISMA-ScR guidance, no formal critical appraisal of included sources was conducted.

Data Analysis

Drawing on extracted data, 1 researcher (CC) developed initial typologies of phase arrangements and progression mechanisms, along with operational definitions for each arrangement and mechanism. These typologies were discussed and collaboratively refined with 2 researchers (ELR and CA) until agreement was reached. To assess the reliability of the typologies, 8 (26%) of the 31 research program structures, defined as the specific combination of phase arrangements and progression mechanisms, were independently coded by 2 researchers (CC and CE). Interrater reliability was evaluated using agreement percentages and Cohen κ. The results indicated almost perfect agreement for phase identification, substantial to almost perfect agreement for phase arrangement identification, and substantial to perfect agreement for the identification of progression mechanisms [36]. Detailed interrater agreement percentages and Cohen κ values for each typology category are reported in Multimedia Appendix 5. Following this double coding, the final typologies and corresponding definitions (presented in the Results section) were refined and subsequently applied by CC to derive the program structures of the remaining interventions. Similar structures were grouped and validated by all coauthors.

XSL•FO
RenderX

Phase-specific and total program durations were calculated as medians with first and third quartiles (IQR), expressed in months and years, respectively.

To examine whether researchers tailored programs to intervention-specific characteristics, we explored relationships between intervention characteristics and both program structure and duration through descriptive analyses using counts and percentages only. Further details are provided in Multimedia Appendix 6.

Results

Study Selection

The initial searches (February 9, 2023; May 2, 2023; and January 2, 2024) yielded a total of 6691 records, from which 31 interventions, described across 119 articles, met the inclusion criteria. Updated searches restricted to already identified interventions (April 15, 2025, and January 7, 2026) identified 7 and 4 additional articles, respectively. This resulted in a final sample of 31 interventions, with their research programs documented across 130 articles (Figure 1). A median of 4 (IQR 3-6) articles per intervention was included.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the literature search. WHO: World Health Organization. AYA: Adolescent and Young Adult.

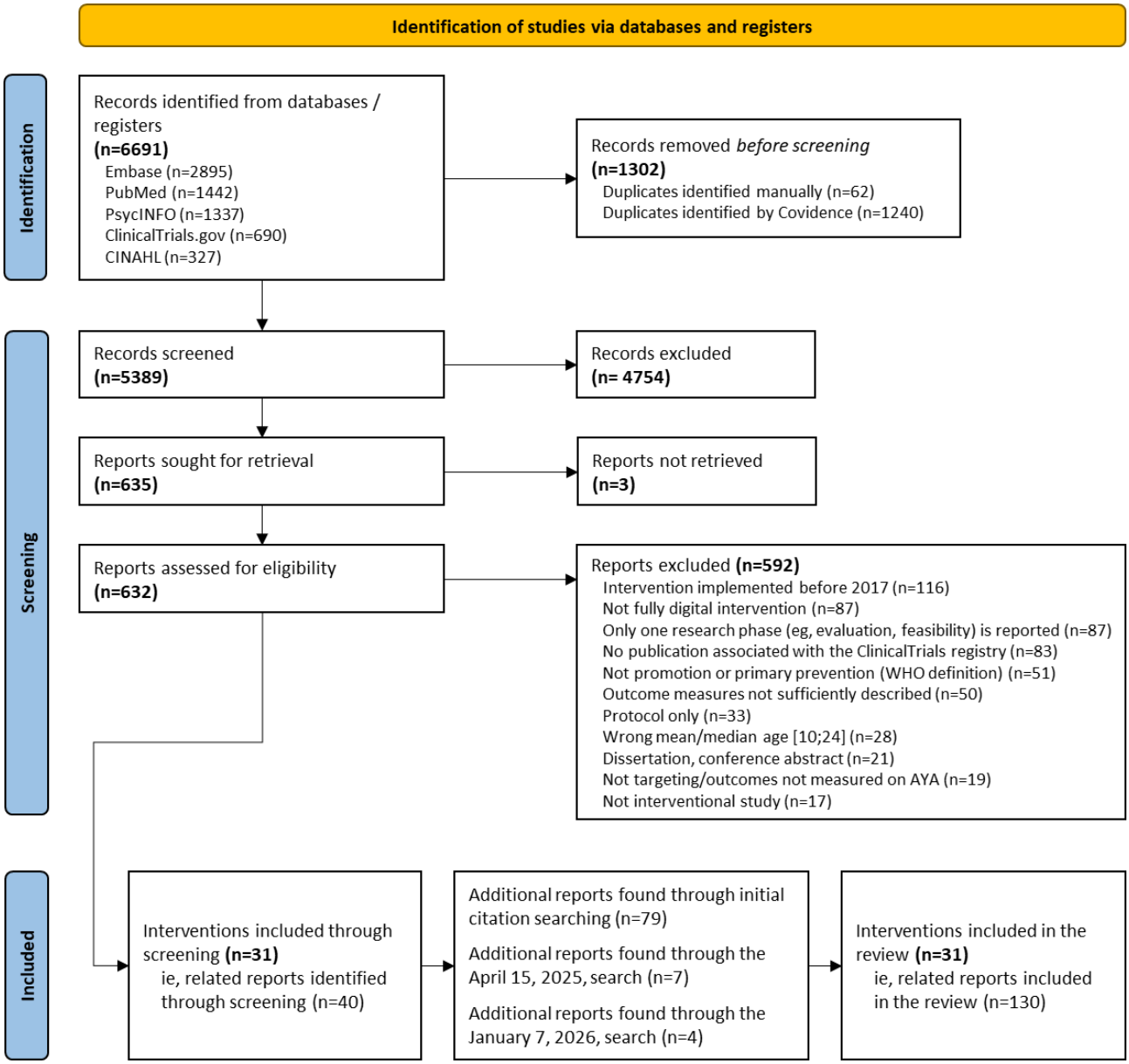

Research Program Characteristics

Of the 31 research programs, 26 included a development phase, 23 a feasibility phase, 31 an evaluation phase (including efficacy or effectiveness evaluation: n=31; process evaluation: n=26; and economic evaluation: n=4), and 7 an implementation phase. One intervention reported a transferability phase, defined as a phase assessing the extent to which positive outcomes observed in a successful intervention, previously evaluated (and implemented) in a primary context, can be replicated in a different target context [37].

XSL•FO
RenderX

In 5 research programs, the MRC GDECI was explicitly referenced, in relation to the development (n=2) and the evaluation (n=3) phases. Four reports highlighted the guidance's usefulness for intervention development and process evaluation, whereas 1 questioned the applicability of the MRC's relatively lengthy phased approach to the rapid evolution of digital interventions. A total of 17/31 of the included interventions used the term "phase" in a manner consistent with the MRC GDECI to describe the stages of their research programs.

Research programs were predominantly structured as 3-phase programs, with some following 2- or 4-phase structures (Table 1).

Table 1. Description of the research program phases across 31 digital health interventions targeting adolescents and young adults.

Phases	All interventions, n (%)
2-phase structure	10 (32)
Development and evaluation	6 (19)
Feasibility and evaluation	4 (13)
3-phase structure	16 (52)
Development, feasibility, and evaluation	13 (42)
Development, evaluation, and implementation	1 (3)
Feasibility, evaluation, and implementation	2 (7)
4-phase structure	5 (16)
Development, feasibility, evaluation, and implementation	4 (13)
Development, feasibility, evaluation, and transferability	1 (3)

Phase Arrangements

The following three types of phase arrangements were observed: sequential, iterative, and overlapping (Table 2).

All interventions used sequential arrangements, with only 7/31 being strictly sequential. The remaining 24/31 combined sequential arrangements with iteration (14/24), overlapping (2/24), or both iteration and overlapping (8/24).

Among the 22 interventions that incorporated iterative arrangements, iteration most commonly occurred during the development (n=17) phase, followed by feasibility (n=5), evaluation (n=1), and implementation (n=1). In 13/22 interventions, the original authors explicitly noted phase iteration as a methodological strength, although the rationale for and frequency of iterations were seldom explained.

Table 2. Typology of the phase arrangements used in the digital health intervention research programs.

Phase arrangement	Description	Notation	Visualization
Sequential	Progression is strictly linear from one phase to the next.	>	
Iterative	One or more phases of the program are repeated at least 2 times.	(i)	
Overlapping	At least 2 phases of the program are conducted simultaneously.	[...+...]	

Among the 10 interventions with overlapping arrangements, the most common pattern was "evaluation + implementation" overlap (n=6), followed by "development + feasibility" overlap (n=3), and multiple overlaps ("development + feasibility" and "evaluation + implementation"; n=1).

Progression Mechanisms

Across the 31 interventions, the following three mechanisms facilitated progression between research phases: automatic progression, conditional progression based on researchers' appraisal of findings without prespecified criteria, and conditional progression based on predefined quantitative criteria with thresholds (Table 3).

XSL•FO
RenderX

Table 3. Typology of progression mechanisms used to move between research phases in digital health research programs.

Progression mechanism[a]	Description	Notation	Visualization
Automatic progression	The next phase begins automatically upon completion of the previous phase, regardless of findings.	*(auto)*	
Conditional progression based on researchers' appraisal of findings without prespecified criteria	Progression is informed by researchers' appraisal and interpretation of available findings, without reliance on prespecified or clearly reported criteria. Decisions may involve adapting the intervention or methods before progressing.	*(appr)*	
Conditional progression based on predefined quantitative criteria	Progression is informed by preestablished, quantitative criteria (ie, progression criteria). A traffic-light system (or green-amber-red) determines whether to proceed, repeat the phase, or terminate the intervention based on established thresholds.	*(traffic)*	

[a]Progression mechanisms that were not reported are represented with the following notation: (NR), and the following visualization: NR. Worked examples: automatic progression: in Wright et al [38], progression from feasibility to evaluation was classified as automatic, as the pilot randomized controlled trial (RCT) of the MIDY intervention showed no intervention effects, yet the research team proceeded directly to a full-scale RCT. Conditional progression based on researchers' appraisal without prespecified criteria: in Reiter et al [39], progression from feasibility to evaluation for the Outsmart HPV intervention was based on the authors' narrative appraisal of positive pilot RCT findings (eg, preliminary efficacy and favorable assessment of delivery methods), in the absence of predefined progression criteria or thresholds. Conditional progression based on predefined quantitative criteria: in Garbett et al [40], progression from feasibility to evaluation for the Warna-Warni Waktu intervention was guided by four prespecified quantitative criteria (retention, adherence, data quality, and harm), using a traffic-light system (green=proceed; amber=modify; red=reconsider or terminate), with progression decisions based on observed performance against these thresholds.

A total of 48 phase-to-phase progressions were documented across all included interventions. Among these, the most common mechanism was conditional progression based on researchers' appraisal of findings without prespecified criteria, accounting for 36/48 mechanisms. Automatic progression was documented in 7 cases: development to feasibility (n=2), development to evaluation (n=1), feasibility to feasibility (n=1), feasibility to evaluation (n=2), and evaluation to implementation (n=1). Only 1 intervention explicitly mentioned the use of predefined progression criteria (traffic-light system) to proceed from feasibility to evaluation. Four mechanisms could not be classified due to insufficient reporting.

Research Program Structures

By combining phase arrangements and progression mechanisms, the following six groups of program structures were identified and are presented in Table 4: (1) entirely sequential structure, with or without all phases (n=6); (2) entirely sequential structure, with conditional progression based on predefined quantitative criteria (n=1); (3) sequential structure with iterative development (n=9), iterative feasibility (n=3), iterative evaluation (n=1) or iteration of 2 different phases ("development and feasibility") (n=1); (4) overlapping development and feasibility phases, followed by a sequential structure (n=3); (5) sequential early phases (development and feasibility) followed by overlapping evaluation and implementation (n=6); and (6) structure with multiple overlaps ("development + feasibility" and "evaluation + implementation"; n=1). Multimedia Appendix 7 provides one worked example for each of the 6 research structures, detailing the phase arrangements and progression mechanisms as reported by the authors and classified according to the developed typologies. Multimedia Appendix 8 presents the overall program structure and phase duration for each included intervention.

XSL•FO
RenderX

Table 4. Main research program structures identified in 31 digital health interventions.

Structure	Description	Visual description	Notation
Structure 1	Sequential structure only, with automatic progression or progression based on researchers' appraisal without prespecified criteria (n=6)		D *(auto)* > F *(appr)* > E *(auto)* > I
Structure 2	Sequential structure only, with progression based on predefined quantitative criteria (n=1)		D *(appr)* > F *(traffic)* > E
Structure 3	Iterative development (n=9), iterative feasibility (n=3), iterative evaluation (n=1) or 2-phase iteration (n=1) within a sequential structure, with automatic progression or progression based on researchers' appraisal without prespecified criteria (n=14)		D(i) *(appr)* > F *(appr)* > E
Structure 4	Overlapping development and feasibility phases (with or without iteration) within a sequential structure, with automatic progression or progression based on researchers' appraisal without prespecified criteria (n=3)		[D(i)+F] *(appr)* > E
Structure 5	Overlapping evaluation and implementation phases (with or without iteration) within a sequential structure, with automatic progression or progression based on researchers' appraisal without prespecified criteria (n=6)[a]		D(i) *(appr)* > F *(auto)* > [E+I]
Structure 6	Multiple overlaps within a sequential structure, with progression based on researchers' appraisal without prespecified criteria (n=1)		[D(i)+F] *(appr)* > [E+I]

[a]Count (n=6) includes 1 intervention in which a transferability phase replaced the implementation phase. Phases: D: development (n=26); F: feasibility (n=23); E: evaluation (n=31); I: implementation (n=7). Phase arrangements are as follows: >: sequential; (i): iterative; [+]: overlapping. Progression mechanisms: *(auto)*: automatic progression; *(appr)*: conditional progression based on researchers' appraisal of findings without prespecified criteria; *(traffic)*: conditional progression based on predefined quantitative criteria. Main structures may be further divided into substructures based on the progression mechanisms used to move from one phase to the next. Multimedia Appendix 7 provides a worked example for each main research structure.

Research Program Duration

The complete research program duration was available for 13 interventions. The median total duration of those programs was 5.8 (IQR 3.8-6.6) years (n=13).

Phase-specific durations were available for varying numbers of interventions. Phase-specific median durations were 18 (IQR 9-31) months for development (n=11); 5 (IQR 1-9) months for feasibility (n=16); 15 (IQR 8-27) months for evaluation (n=30); and 48 (IQR 48-48) months for implementation (n=1).

Cumulative median time for research phases only, excluding gaps between phases, was 3.5 (IQR 2.2-4.4) years (n=13).

Relationships Between Intervention Characteristics and Program Structure and Duration

Intervention characteristics, along with descriptive statistics and tests of hypotheses between intervention characteristics and program structure and duration, are presented in Multimedia Appendix 9. Exploratory analyses were limited by missing data on funding (15/31, 48%) and duration (13/31, 42%), and small subgroup sizes.

Regarding program structure, newly developed interventions compared to adapted ones (15/19, 79% vs 5/12, 42%) and interventions with dynamic content compared to those with fixed content (4/5, 80% vs 16/26, 62%) showed greater iteration in the early phases. Studies with higher research funding were more complete 3- or 4-phase research programs. Specifically,

XSL•FO
RenderX

all 3 interventions with funding over US $5 million used either 3-phase (n=1) or 4-phase (n=2) programs.

Regarding program duration, these well-funded programs tended to have longer evaluation phases (lasting over 24 months).

No patterns were observed between other intervention characteristics and program structure or duration.

Synthesizing Results: A Novel Operational Framework for Research Program Design

Building on the findings related to research program characteristics, structure, and duration, and informed by the MRC GDECI, a novel 4-step operational framework was developed to support the planning, reporting, and critical appraisal of DHI research programs (Figure 2).

Figure 2. Operational framework for designing and planning digital health promotion interventions in terms of phase arrangements and progression mechanisms.

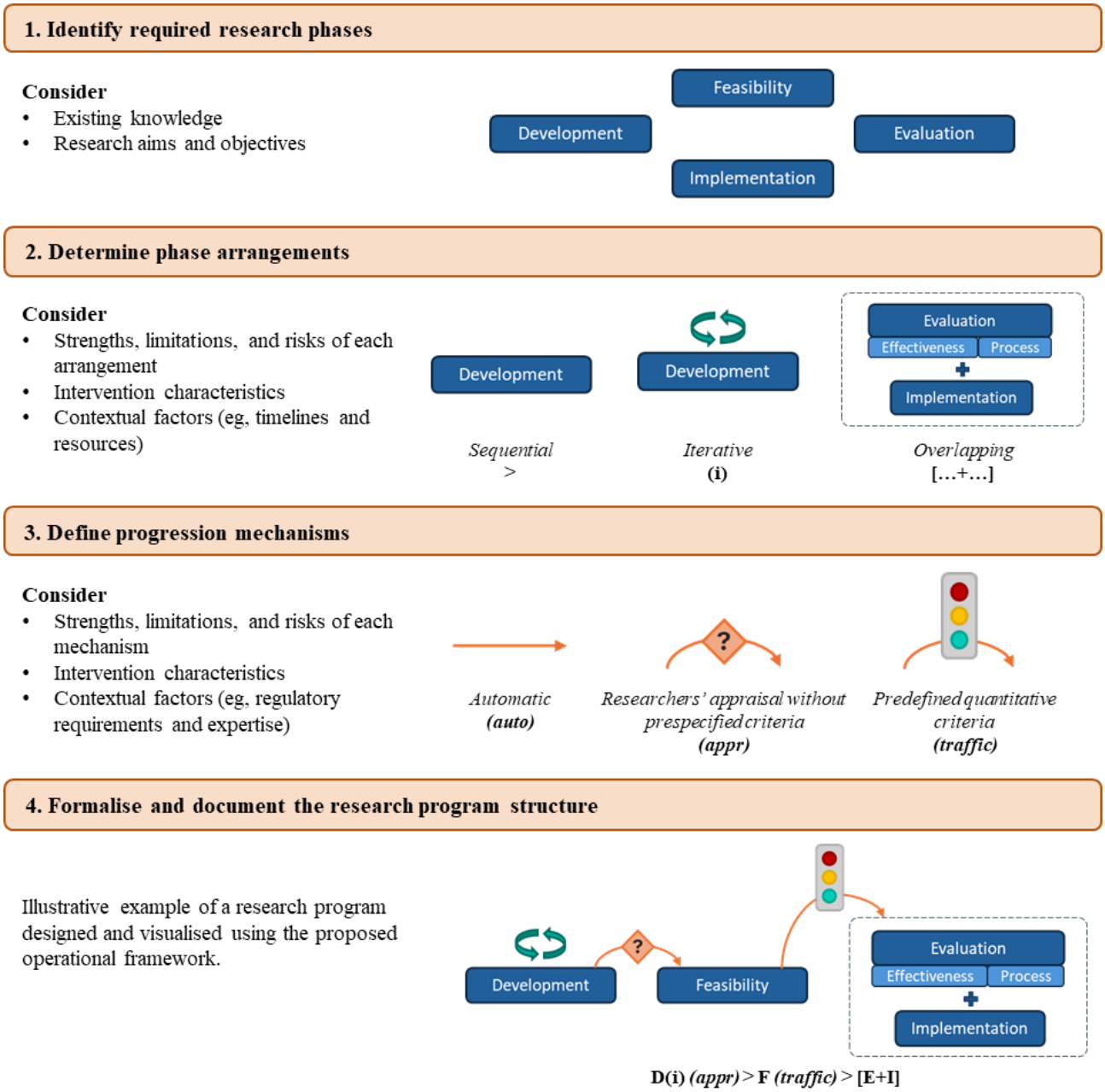

The framework promotes structured decision-making across 4 sequential steps. Step 1 involves explicit identification of the research phases to be included (ie, development, feasibility, evaluation, and implementation). Step 2 requires specification of how these phases will be arranged (ie, sequentially, iteratively, or in overlapping arrangements). Step 3 concerns the progression mechanisms between phases (ie, automatic progression, progression based on researchers' appraisal without prespecified criteria, or progression based on predefined quantitative criteria). Finally, Step 4 entails formalization of the overall research program structure using the proposed visual tools and associated notation.

At each step, the framework highlights key considerations for decision-making: research aims guide phase selection, while intervention characteristics (eg, degree of novelty, behavioral complexity, and technological uncertainty) and contextual

XSL•FO
RenderX

constraints inform choices regarding phase arrangements and progression mechanisms.

For example, novel interventions incorporating untested technologies or complex behaviors may benefit from iterative development with predefined quantitative criteria to ensure adequate refinement before costly evaluation. Conversely, interventions with established effectiveness may use overlapping phases to accelerate implementation. Regarding progression mechanisms, conditional progression based on researchers' appraisal of findings without prespecified criteria enables nuanced decision-making but requires transparent documentation of decision criteria and rationale. In contrast, predefined progression criteria maximize transparency and reproducibility but demand careful selection of meaningful metrics and thoughtful interpretation.

To complement the framework, and drawing on both the existing literature and the patterns identified in this review, Table 5 synthesizes the key strengths, potential risks, and methodological limitations of each phase arrangement and progression mechanism [5,13,41-44].

XSL•FO
RenderX

Table 5. Strengths, risks, and limitations of the phase arrangements and progression mechanisms identified in digital health interventions research programs.

	Strengths	Risks and limitations
Phase arrangements		
Sequential	• Provides a clear structure with distinct phase boundaries • Ensures systematic progression through research stages (or justification as to why one is omitted) • Facilitates reporting, outcome selection, and interpretation, and scientific valorization of each phase • Supports targeted funding applications and allocation per phase	• May increase total program duration (compared to overlapping arrangements) • Limits flexibility to respond to emerging uncertainties (risk of overadherence to preestablished plans) • May delay the implementation of promising interventions
Iterative	• Enables unresolved uncertainties to be addressed before progression • Leads to more refined and context-sensitive interventions • Supports real-time learning and adaptation throughout the program	• Extends the duration of individual phases • Increases research costs • Requires transparent reporting of rationale and changes at each iteration • More complex to design, coordinate, and manage • Requires criteria to determine when further iteration is unnecessary or unproductive
Overlapping	• Enables simultaneous work on multiple phases, potentially reducing total program duration • Enables cross-phase interpretation of outcomes while discussing intervention success (notably evaluation and implementation outcomes)	• Blurs methodological boundaries between phases • Increases complexity in research planning, coordination, and outcome interpretation • More resource-intensive (in terms of funding, staffing, expertise, and logistics)
Progression mechanisms		
Automatic progression	• Ensures timely progression between phases, reducing delays • Simplifies program planning and management	• May allow progression of underdeveloped or unfeasible interventions • Risk of propagating early flaws in the intervention into subsequent phases • Involves minimal to no decision-making regarding progression
Conditional progression based on researchers' appraisal of findings without prespecified criteria	• Enables holistic assessment drawing on multiple data sources • Allows integration of both quantitative and qualitative findings • Permits context-sensitive decision-making informed by intervention-specific knowledge	• Prone to reporting (selective reporting) and confirmation biases • Lacks standardization and transparency in decision-making • Increases planning complexity • Limits reproducibility and comparability across studies
Conditional progression based on predefined quantitative criteria (progression criteria)[a]	• Ensures transparent and standardized decision-making • Provides objective thresholds to determine success • Reduces the risk of reporting bias • Enhances study transparency, replicability, and reproducibility	• Decision-making precedes experimentation, risking overlooking unexpected findings • Risk of setting arbitrary or inappropriate thresholds • Potential disconnect between metrics and meaningful intervention success • Reduces flexibility to adapt to emerging insights

[a]With respect to the use of predefined quantitative criteria, Mellor et al [13] propose a four-step framework (ie, design, conduct, analysis, and reporting) to guide researchers in the application of such progression mechanisms.

Discussion

Summary of Findings

This study examined how DHI research programs align with the MRC GDECI, focusing on how research phases are arranged and how progression between phases occurs. Our analysis of 31 programs revealed no single dominant structure, but several structures characterized by 3 types of phase arrangements (sequential, iterative, and overlapping) and 3 progression mechanisms (automatic, appraisal-based, and predefined criteria). Iteration, particularly of early phases (development and feasibility), and overlaps were common. Most progressions relied on researchers' appraisal or occurred automatically, with only 1 program using predefined quantitative criteria. Justifications for phase iteration or omission, or the choice of progression mechanisms, were rarely documented, corroborating existing literature [9]. Median program duration was 5.8 (IQR 3.8-6.6) years (n=13), while the median cumulative duration of

XSL•FO
RenderX

individual phases was 3.5 (IQR 2.2-4.4) years (n=13). Finally, a novel, 4-step operational framework, along with the strengths, potential risks, and methodological limitations of each phase arrangement and progression mechanism, is proposed to support researchers in making informed decisions in designing their DHI research programs.

Interpretation and Implications of Findings

Phase Arrangements

Strictly sequential arrangements, arguably aligning most closely with a traditional MRC GDECI interpretation, were relatively uncommon. This may reflect differential guidance application in response to intervention-specific requirements.

Iterative arrangements were frequent, likely reflecting the dynamic nature of DHIs [20]. These interventions often require multiple rounds of user testing, feedback incorporation, and technological refinement [17], consistent with agile and user-centered methodologies increasingly promoted in digital research [45-47]. Iterative approaches may be particularly relevant for DHIs targeting AYAs, to sustain engagement among digitally savvy participants with elevated expectations for sophisticated technologies as found in gaming and social media [48,49]. However, repeated early-phase iterations may prolong research timelines and increase costs [50], necessitating a balance between enhancing relevance and mitigating the risk of obsolescence in rapidly evolving digital environments [22].

The use of overlapping phases, while not explicitly endorsed in the MRC GDECI, reflects emerging interests in hybrid research designs intended to accelerate translation from evaluation to implementation [35]. For DHIs, such arrangements may provide simultaneous insights into multiple research phases while addressing rapid technological evolution [41]. However, hybrid research designs require advanced methodological expertise to prevent evaluation and implementation of underdeveloped interventions [51], potentially increasing research waste. The rise of overlapping arrangements also suggests that the strict phase boundaries outlined in the MRC GDECI may not fully reflect the realities of DHI research. This observation supports the need for more flexible, context-sensitive adaptations of the guidance, as has already been advocated in other fields, including palliative care and rehabilitation research [52-54].

Progression Mechanisms

The important reliance on conditional progression based on researchers' appraisal of findings without prespecified criteria and automatic progression may stem from several factors. First, current funding structures often support evaluation-focused programs [55], potentially incentivizing progression regardless of early-phase outcomes. Second, predefined progression criteria, originally formalized within clinical research, have probably not yet been adopted in DHI research; as Perski and Short [56] have, for instance, shown for DHI acceptability assessment, for which thresholds are rarely specified, and, when present, are typically not grounded in clear theoretical or empirical justification. Third, the complexity of DHIs may preclude reliance on simple, predefined quantitative thresholds to determine phase completeness [17].

Nevertheless, the limited use of predefined progression criteria is noteworthy given their endorsement in methodological literature for enhancing research integrity and transparency, while limiting biases [9,13]. This gap is especially problematic given the risk of inadequate documentation or underreporting of progression decisions when explicit criteria are not used [11]. However, the use of such criteria warrants critical attention: many lack an objective scientific basis, and their overmechanistic use may prove problematic [10].

Program Duration

The median program duration of 5.8 (IQR 3.8-6.6) years is substantially shorter than the often-cited 17-year development to implementation timeline for complex health care innovations [23], potentially reflecting inherently faster DHI development driven by competitive pressures and the relative ease of digital deployment compared to face-to-face interventions [22]. However, several alternative explanations may account for this observed difference. These include the limited number of implementation studies in our sample, the substantial amount of missing duration data, and our reliance on trial registry dates to approximate phase-specific duration. The latter approach may have underestimated phase duration, particularly when the "date of last data collection" was the only available proxy for phase completion. In practice, the activities required to formally conclude a phase, such as data analysis, interpretation of findings, decision-making regarding progression to subsequent phases, and publication, often extend for several additional months and, in some cases, years beyond the end of data collection [57].

While shorter timelines appear advantageous for DHIs, the gap between total program duration and cumulative phase duration (5.8 vs 3.5 years, respectively) still reveals substantial interphase delays, likely due to funding acquisition, ethical approvals, or regulatory processes [58]. Such delays are particularly concerning for DHIs given the risk of rapid technological obsolescence [22].

Implications and Applicability of the Operational Framework

The developed framework, alongside the identified strengths, potential risks, and limitations of phase arrangements and progression mechanisms, represents one of the main contributions of this review. It is designed to support researchers in making transparent, reflexive, and methodologically informed decisions when structuring research programs and defining progression between phases. Importantly, it is not intended as a prescriptive or universal model. Rather, it provides a set of "building blocks," namely, phase arrangements and progression mechanisms, that can be selected, combined, and adapted according to intervention characteristics, contextual realities, and practical constraints, provided that such selections are explicitly justified and transparently reported.

While based on the MRC GDECI and developed from DHI-specific findings, the developed framework has potential for applicability beyond the guidance and DHIs to prevention and health promotion interventions more broadly. While our results suggested that not all authors used the MRC GDECI,

XSL•FO
RenderX

phased approaches to structuring research programs were explicitly reported in more than half of the interventions and could be inferred in all included DHIs. Providing operational guidance for more transparent arrangements and progressions between research phases may therefore benefit a wide range of researchers, including those who do not explicitly reference the MRC GDECI.

The framework is also likely applicable to non-DHI interventions, although the resulting research structures may differ, as DHIs and non-DHIs have distinct characteristics. DHIs may require shorter timelines and benefit from iterative early phases to produce engaging, user-ready tools. In contrast, more stable health promotion interventions, such as school-based programs, may support longer, sequential programs with a greater emphasis on implementation. These differences highlight the potential for adapting the framework to diverse intervention types while retaining its core principles for phased research program structuring.

Strengths and Limitations

This study is the first to systematically examine DHI research program structures against established methodological guidance, offering valuable insights into current practices. The proposed framework turns these findings into a practical, visual tool to assist with future research planning and funding decisions.

Limitations must be acknowledged. Exploratory analyses of the relationships between intervention characteristics and program structure and duration were constrained by missing data on funding and duration, and by small sample sizes within subgroups of intervention characteristics and phase arrangements. Although the absence of clear patterns in our analyses could be interpreted as a lack of relationships between the selected intervention characteristics and program structure or duration, it is equally plausible that limited sample sizes hindered the detection of such relationships. As a result, future research with larger datasets is needed to assess these relationships more robustly and to support the development of additional propositions and recommendations to better guide research program design. Furthermore, the relatively small number of implementation studies identified in the review raises questions regarding the chosen eligibility criteria and search strategies. With respect to eligibility criteria, requiring included interventions to report at least 2 research phases and an evaluation component may have led to the underrepresentation of single-phase studies, nonevaluated pilots, and pilots that were discontinued early due to negative feasibility findings or lack of further funding. Although this criterion was necessary to document progression mechanisms, it privileged research programs structured around an evaluation phase. Nevertheless, this focus remains highly relevant, as the effectiveness of interventions is assessed during the evaluation phase and constitutes a critical stage in intervention development and testing. With respect to search strategies, the focus on academic databases may have led to the overlooking of some implementation studies published outside peer-reviewed journals or described using terminology not captured by the search strategy. This may have limited our ability to provide a more comprehensive account of the conduct, arrangement, and progression toward real-world implementation phases within research programs. However, this pattern may also reflect broader structural challenges in conducting and publishing implementation research, including funding mechanisms that prioritize development and evaluation over implementation and scale-up.

Conclusion

To our knowledge, this scoping review is the first to systematically examine how DHI research programs are structured with respect to phase arrangements and progression mechanisms, and to explore how intervention characteristics are associated with program structure and duration. By synthesizing current practices and the decision-making processes underpinning program design, this review advances the field by moving beyond descriptive reporting to offer a conceptualization of research program structures that provides practical direction for the concrete operationalization of the MRC GDECI framework.

The resulting operational framework offers a flexible, nonprescriptive tool that supports transparent, context-sensitive decision-making in research planning, making an original contribution to the field. These findings carry practical implications for researchers, funders, journal editors, and the broader scientific community. Greater explicitness in defining phase arrangements and progression mechanisms may enhance methodological rigor, reduce research waste, and improve the integrity and reproducibility of DHIs, as well as of prevention and health promotion interventions more broadly.

Future research should further test, refine, and validate this framework across diverse intervention types and investigate factors associated with decision-making for phase arrangements and progression mechanisms.

Acknowledgments

The authors thank Morgane Michel for her valuable and insightful comments on this paper.

Funding

The authors have not declared a specific grant for this research from any funding agency in the public, commercial, or not-for-profit sectors. CC is supported by a doctoral contract funded by the École des Hautes Études en Santé Publique.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are available in the Zenodo repository [59].

Authors' Contributions

CC, ELR, and CA conceptualized and designed the study. CC wrote the study protocol. CC, ELR, and PM created the search strategy. CC and CE screened abstracts and full texts and extracted data. ELR resolved disagreements. CC conducted the formal analysis, with contributions from CE, ELR, CA, and PM for the visualization of results. CC, ELR, CA, and GM interpreted the findings. CC drafted the original manuscript, which was reviewed and edited by CA, PM, GM, CE, and ELR. ELR and CA supervised the research. All authors read and approved the final manuscript. CC and ELR are the guarantors for the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Completed PRISMA-ScR 2019 checklist for report and abstract.
[DOCX File , 54 KB - jmir_v28i1e82611_app1.docx]

Multimedia Appendix 2
PRISMA-S checklist and search strategies by database.
[DOCX File , 50 KB - jmir_v28i1e82611_app2.docx]

Multimedia Appendix 3
Data charting form.
[DOCX File , 46 KB - jmir_v28i1e82611_app3.docx]

Multimedia Appendix 4
Terminology used to name research phases.
[DOCX File , 41 KB - jmir_v28i1e82611_app4.docx]

Multimedia Appendix 5
Interrater agreement on classifying phase conduct, phase arrangement, and progression mechanisms for 8 digital health interventions.
[DOCX File , 26 KB - jmir_v28i1e82611_app5.docx]

Multimedia Appendix 6
Hypotheses linking intervention characteristics and research program structure and duration.
[DOCX File , 39 KB - jmir_v28i1e82611_app6.docx]

Multimedia Appendix 7
Worked examples of digital health research program structures (A-F).
[DOCX File , 139 KB - jmir_v28i1e82611_app7.docx]

Multimedia Appendix 8
Research program structure, duration, and corresponding references for each intervention included in the study.
[DOCX File , 527 KB - jmir_v28i1e82611_app8.docx]

Multimedia Appendix 9
Description of intervention characteristics and hypothesis testing.
[DOCX File , 54 KB - jmir_v28i1e82611_app9.docx]

References

1. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021;374:n2061 [FREE Full text] [doi: 10.1136/bmj.n2061] [Medline: 34593508]

XSL•FO
RenderX

2. Duncan E, O'Cathain A, Rousseau N, Croot L, Sworn K, Turner KM, et al. Guidance for reporting intervention development studies in health research (GUIDED): an evidence-based consensus study. BMJ Open 2020;10(4):e033516 [FREE Full text] [doi: 10.1136/bmjopen-2019-033516] [Medline: 32273313]
3. Hallingberg B, Turley R, Segrott J, Wight D, Craig P, Moore L, et al. Exploratory studies to decide whether and how to proceed with full-scale evaluations of public health interventions: a systematic review of guidance. Pilot Feasibility Stud 2018;4:104 [FREE Full text] [doi: 10.1186/s40814-018-0290-8] [Medline: 29854417]
4. Pfadenhauer LM, Gerhardus A, Mozygemba K, Lysdahl KB, Booth A, Hofmann B. Making sense of complexity in context and implementation: the context and implementation of complex interventions (CICI) framework. Implementation Sci 2017;12:21 [FREE Full text] [doi: 10.1186/s13012-017-0552-5]
5. O'Cathain A, Croot L, Duncan E, Rousseau N, Sworn K, Turner KM, et al. Guidance on how to develop complex interventions to improve health and healthcare. BMJ Open 2019;9(8):e029954 [FREE Full text] [doi: 10.1136/bmjopen-2019-029954] [Medline: 31420394]
6. Eldridge SM, Lancaster GA, Campbell MJ, Thabane L, Hopewell S, Coleman CL, et al. Defining feasibility and pilot studies in preparation for randomised controlled trials: development of a conceptual framework. PLoS One 2016;11(3):e0150205 [FREE Full text] [doi: 10.1371/journal.pone.0150205] [Medline: 26978655]
7. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet 2009;374(9683):86-89 [FREE Full text] [doi: 10.1016/s0140-6736(09)60329-9]
8. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, et al. Reducing waste from incomplete or unusable reports of biomedical research. Lancet 2014;383(9913):267-276 [FREE Full text] [doi: 10.1016/s0140-6736(13)62228-x]
9. Avery KNL, Williamson PR, Gamble C, O'Connell Francischetto E, Metcalfe C, Davidson P, et al. Informing efficient randomised controlled trials: exploration of challenges in developing progression criteria for internal pilot studies. BMJ Open 2017;7(2):e013537 [FREE Full text] [doi: 10.1136/bmjopen-2016-013537] [Medline: 28213598]
10. Mellor K, Eddy S, Peckham N, Bond CM, Campbell MJ, Lancaster GA, et al. Progression from external pilot to definitive randomised controlled trial: a methodological review of progression criteria reporting. BMJ Open 2021;11(6):e048178 [FREE Full text] [doi: 10.1136/bmjopen-2020-048178] [Medline: 34183348]
11. Mellor K, Dutton SJ, Hopewell S, Albury C. How are progression decisions made following external randomised pilot trials? A qualitative interview study and framework analysis. Trials 2022;23(1):132 [FREE Full text] [doi: 10.1186/s13063-022-06063-9] [Medline: 35144670]
12. Beets MW, von Klinggraeff L, Weaver RG, Armstrong B, Burkart S. Small studies, big decisions: the role of pilot/feasibility studies in incremental science and premature scale-up of behavioral interventions. Pilot Feasibility Stud 2021;7(1):173 [FREE Full text] [doi: 10.1186/s40814-021-00909-w] [Medline: 34507624]
13. Mellor K, Albury C, Dutton SJ, Eldridge S, Hopewell S. Recommendations for progression criteria during external randomised pilot trial design, conduct, analysis and reporting. Pilot Feasibility Stud 2023;9(1):1-10 [FREE Full text] [doi: 10.1186/s40814-023-01291-5] [Medline: 37061720]
14. McGauran N, Wieseler B, Kreis J, Schüler YB, Kölsch H, Kaiser T. Reporting bias in medical research - a narrative review. Trials 2010;11:37 [FREE Full text] [doi: 10.1186/1745-6215-11-37] [Medline: 20388211]
15. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. Framework for the development and evaluation of complex interventions: gap analysis, workshop and consultation-informed update. Health Technol Assess 2021;25(57):1-132 [FREE Full text] [doi: 10.3310/hta25570] [Medline: 34590577]
16. Lipworth W, Kerridge I, Stewart C, Silva D, Upshur R. The fragility of scientific rigour and integrity in "Sped up Science": research misconduct, bias, and hype and in the COVID-19 pandemic. J Bioeth Inq 2023;20(4):607-616 [FREE Full text] [doi: 10.1007/s11673-023-10289-w] [Medline: 38064166]
17. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016;51(5):843-851 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
18. West PR, Michie PS. A Guide to Development and Evaluation of Digital Behaviour Change Interventions in Healthcare. Sutton, United Kingdom: Silverback Publishing; 2016.
19. Raeside R, Jia SS, Redfern J, Partridge SR. Navigating the online world of lifestyle health information: qualitative study with adolescents. JMIR Pediatr Parent 2022;5(1):e35165 [FREE Full text] [doi: 10.2196/35165] [Medline: 35147506]
20. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. J Med Internet Res 2017;19(6):e232 [FREE Full text] [doi: 10.2196/jmir.7126] [Medline: 28663162]
21. Stanic T, Avsar TS, Gomes M. Economic evaluations of digital health interventions for children and adolescents: systematic review. J Med Internet Res 2023;25:e45958 [FREE Full text] [doi: 10.2196/45958] [Medline: 37921844]
22. Patrick K, Hekler EB, Estrin D, Mohr DC, Riper H, Crane D, et al. The pace of technologic change: implications for digital health behavior intervention research. Am J Prev Med 2016;51(5):816-824 [FREE Full text] [doi: 10.1016/j.amepre.2016.05.001] [Medline: 27745681]
23. Grant J, Green L, Mason B. Basic research and health: a reassessment of the scientific basis for the support of biomedical science. Res Eval 2003;12(3):217-224 [FREE Full text] [doi: 10.3152/147154403781776618]

XSL•FO
RenderX

24. Moreno MA, Binger K, Zhao Q, Eickhoff J, Minich M, Uhls YT. Digital technology and media use by adolescents: latent class analysis. JMIR Pediatr Parent 2022:e35540 [FREE Full text] [doi: 10.2196/35540]
25. Baird S, Choonara S, Azzopardi PS, Banati P, Bessant J, Biermann O. A call to action: the second Lancet commission on adolescent health and wellbeing. Lancet 2025:1945-2022 [FREE Full text] [doi: 10.1016/s0140-6736(22)01529-x]
26. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169:467-473 [FREE Full text] [doi: 10.7326/m18-0850]
27. Collin C, Eyraud C, Martin P, Michel M, Le Roux E, Alberti C. A scoping review of outcome selection and accuracy of conclusions in complex digital health interventions for young people (2017-2023): methodological proposals for population health intervention research. BMC Med 2025:400 [FREE Full text] [doi: 10.1186/s12916-025-04245-1]
28. Adolescent and young adult health. World Health Organization. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/adolescents-health-risks-and-solutions [accessed 2024-11-20]
29. Yardley L, Choudhury T, Patrick K, Michie S. Current issues and future directions for research into digital behavior change interventions. Am J Prev Med 2016;51(5):814-815 [FREE Full text] [doi: 10.1016/j.amepre.2016.07.019] [Medline: 27745680]
30. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016;51(5):833-842 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
31. McNamee P, Murray E, Kelly MP, Bojke L, Chilcott J, Fischer A. Designing and undertaking a health economics study of digital health interventions. Am J Prev Med 2016;51:852-860 [FREE Full text] [doi: 10.1016/j.amepre.2016.05.007]
32. Hekler EB, Michie S, Pavel M, Rivera DE, Collins LM, Jimison HB, et al. Advancing models and theories for digital behavior change interventions. Am J Prev Med 2016;51(5):825-832 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.013] [Medline: 27745682]
33. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10:39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z]
34. Enam A, Torres-Bonilla J, Eriksson H. Evidence-based evaluation of eHealth interventions: systematic literature review. J Med Internet Res 2018;20(11):e10971 [FREE Full text] [doi: 10.2196/10971] [Medline: 30470678]
35. Curran GM, Bauer M, Mittman B, Pyne JM, Stetler C. Effectiveness-implementation hybrid designs. Med Care 2012;50:217-226 [FREE Full text] [doi: 10.1097/mlr.0b013e3182408812]
36. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174 [FREE Full text] [doi: 10.2307/2529310]
37. Cambon L, Minary L, Ridde V, Alla F. Transferability of interventions in health education: a review. BMC Public Health 2012;12:497 [FREE Full text] [doi: 10.1186/1471-2458-12-497]
38. Wright C, Dietze PM, Agius PA, Kuntsche E, Livingston M, Black OC, et al. Mobile phone-based ecological momentary intervention to reduce young adults' alcohol use in the event: a three-armed randomized controlled trial. JMIR Mhealth Uhealth 2018;6(7):e149 [FREE Full text] [doi: 10.2196/mhealth.9324] [Medline: 30030211]
39. Reiter PL, Katz ML, Bauermeister JA, Shoben AB, Paskett ED, McRee AL. Increasing human papillomavirus vaccination among young gay and bisexual men: a randomized pilot trial of the Outsmart HPV intervention. LGBT Health 2018;5(5):325-329 [FREE Full text] [doi: 10.1089/lgbt.2018.0059] [Medline: 29979642]
40. Garbett KM, Craddock N, Haywood S, Nasution K, White P, Saraswati LA, Girl Effect, Percolate Galactic, et al. A novel, scalable social media-based intervention ("Warna-Warni Waktu") to reduce body dissatisfaction among young Indonesian women: protocol for a parallel randomized controlled trial. JMIR Res Protoc 2022;11(1):e33596 [FREE Full text] [doi: 10.2196/33596] [Medline: 35089154]
41. Matson TE, Hermes EDA, Lyon AR, Quanbeck A, Schueller SM, Wilson SM. A framework for designing hybrid effectiveness-implementation trials for digital health interventions. Ann Epidemiol 2025;104:35-47 [FREE Full text] [doi: 10.1016/j.annepidem.2025.02.007]
42. Jacobs MA, Graham AL. Iterative development and evaluation methods of mHealth behavior change interventions. Curr Opin Psychol 2016;9:33-37 [FREE Full text] [doi: 10.1016/j.copsyc.2015.09.001]
43. Mohr DC, Lyon AR, Lattie EG, Reddy M, Schueller SM. Accelerating digital mental health research from early design and creation to successful implementation and sustainment. J Med Internet Res 2017;19(5):e153 [FREE Full text] [doi: 10.2196/jmir.7725] [Medline: 28490417]
44. Rauwerdink A, Spinazze P, Gijsbers H, Molendijk J, Zwolsman S, Schijven MP. Approaches to evaluating digital health technologies: scoping review. J Med Internet Res 2024:e50251 [FREE Full text] [doi: 10.2196/50251]
45. Blandford A, Gibbs J, Newhouse N, Perski O, Singh A, Murray E. Seven lessons for interdisciplinary research on interactive digital health interventions. Digit Health 2018;4:2055207618770325 [FREE Full text] [doi: 10.1177/2055207618770325]
46. Kokol P, Blažun Vošner H, Kokol M, Završnik J. Role of agile in digital public health transformation. Front Public Health 2022;10:899874 [FREE Full text] [doi: 10.3389/fpubh.2022.899874] [Medline: 35646754]
47. Kilfoy A, Hsu TCC, Stockton-Powdrell C, Whelan P, Chu C, Jibb L. An umbrella review on how digital health intervention co-design is conducted and described. NPJ Digit Med 2024:1-13 [FREE Full text] [doi: 10.1038/s41746-024-01385-1]

XSL•FO
RenderX

48. Raeside R. Advancing adolescent health promotion in the digital era. Health Promot Int 2025;40(2):daae172 [FREE Full text] [doi: 10.1093/heapro/daae172] [Medline: 40037909]
49. Li Z, Lu F, Wu J, Bao R, Rao Y, Yang Y, et al. Usability and effectiveness of eHealth and mHealth interventions that support self-management and health care transition in adolescents and young adults with chronic disease: systematic review. J Med Internet Res 2024;26:e56556 [FREE Full text] [doi: 10.2196/56556] [Medline: 39589770]
50. Wynn DC, Eckert CM. Perspectives on iteration in design and development. Res Eng Design 2017;28:153-184 [FREE Full text] [doi: 10.1007/s00163-016-0226-3]
51. Ullman AJ, Beidas RS, Bonafide CP. Methodological progress note: hybrid effectiveness-implementation clinical trials. J Hosp Med 2022;17(11):912-916 [FREE Full text] [doi: 10.1002/jhm.12936] [Medline: 35934981]
52. Evans C, Higginson I, Manthrope J. MRC Guidance on Developing and Evaluating Complex Interventions: Application to Research on Palliative and End of Life Care. London: NIHR School of Social Care Research, LSE; 2013.
53. Levack WM, Gross DP, Martin RA, Every-Palmer S, Kiekens C, Cordani C, et al. Designing studies and reviews to produce informative, trustworthy evidence about complex interventions in rehabilitation: a narrative review and commentary. Eur J Phys Rehabil Med 2024;60(6):1088-1099 [FREE Full text] [doi: 10.23736/s1973-9087.24.08459-4]
54. Connell L, Ackerley S, Rycroft-Malone J. Applying the updated MRC framework for developing and evaluating complex interventions with integrated implementation conceptual knowledge: an example using NeuroRehabilitation OnLine. Front Health Serv 2025 [FREE Full text] [doi: 10.3389/frhs.2025.1562627]
55. Turner KM, Rousseau N, Croot L, Duncan E, Yardley L, O'Cathain A, et al. Understanding successful development of complex health and healthcare interventions and its drivers from the perspective of developers and wider stakeholders: an international qualitative interview study. BMJ Open 2019;9(5):e028756 [FREE Full text] [doi: 10.1136/bmjopen-2018-028756] [Medline: 31152042]
56. Perski O, Short C. Acceptability of digital health interventions: embracing the complexity. Transl Behav Med 2021;11:1473-1480 [FREE Full text] [doi: 10.1093/tbm/ibab048]
57. Showell MG, Cole S, Clarke MJ, DeVito NJ, Farquhar C, Jordan V. Time to publication for results of clinical trials. Cochrane Database Syst Rev 2024:1 [FREE Full text] [doi: 10.1002/14651858.mr000011.pub3]
58. Blümle A, Haag T, Balmford J, Rücker G, Schumacher M, Binder N. A multi-state model analysis of the time from ethical approval to publication of clinical research studies. PLoS One 2020;15(3):e0230797 [FREE Full text] [doi: 10.1371/journal.pone.0230797] [Medline: 32218603]
59. Dataset_Scientific Article_Planning Research on Digital Health Promotion Interventions_2026January19. Zenodo. URL: https://zenodo.org/records/18302609 [accessed 2026-04-21]

Abbreviations

AYA: adolescent and young adult
DHI: digital health intervention
GDECI: Guidance on Developing and Evaluating Complex Interventions
ISRCTN: International Standard Randomized Controlled Trial Number
MRC: UK Medical Research Council
PRISMA-S: Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
PROSPERO: International Prospective Register of Systematic Reviews
RCT: randomized controlled trial

Edited by S Brini; submitted 19.Aug.2025; peer-reviewed by M Lange, Y Kong, A Shivanna, B Breznen; comments to author 28.Nov.2025; revised version received 25.Feb.2026; accepted 05.Mar.2026; published 06.May.2026.

Please cite as:
Collin C, Alberti C, Martin P, Moore G, Eyraud C, Le Roux E
Current Practices and a Novel Operational Framework for Planning Research on Digital Health Promotion Interventions From Development to Implementation: Scoping Review
J Med Internet Res 2026;28:e82611
URL: https://www.jmir.org/2026/1/e82611
doi:10.2196/82611
PMID:

XSL•FO
RenderX

©Claire Collin, Corinne Alberti, Philippe Martin, Graham Moore, Clara Eyraud, Enora Le Roux. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Automated Approaches of Text Simplification of Patient Education Materials: Scoping Review

Cornelia Krenn[1], BSc, MSc, Dr.scient.med.; Christine Loder[1], Mag., MPH; Natalie Berger[1], BSc, MSc; Klaus Jeitler[1,2], Dr med; Thomas Semlitsch[1], Mag.rer.nat.; Andrea Siebenhofer[1,3], Dr med, MBA; Denise Wilfling[1,4], BSc, MSc, Dr.rer.hum.biol.

[1]Institute of General Practice and Evidence-based Health Services Research, Medical University of Graz, Graz, Austria

[2]Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria

[3]Institute of General Practice, Goethe University Frankfurt am Main, Frankfurt, Germany

[4]Institute of Nursing and Health Sciences, Medical University of Graz, Graz, Austria

Corresponding Author:
Cornelia Krenn, BSc, MSc, Dr.scient.med.
Institute of General Practice and Evidence-based Health Services Research, Medical University of Graz
Neue Stiftingtalstraße 6
Graz, 8010
Austria
Phone: 43 316385 ext 73567
Email: cornelia.krenn@medunigraz.at

Abstract

Background: Patient education materials (PEMs) often exceed the American Medical Association's (AMA) recommended sixth-grade reading grade level (RGL). While artificial intelligence (AI) offers potential for automated text simplification, concerns persist regarding linguistic quality, content fidelity, and the understandability of simplified PEMs by laypeople.

Objective: This scoping review maps existing evidence on automated language processing technologies for simplifying PEMs for laypeople.

Methods: Following the Joanna Briggs Institute (JBI) methodology and the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guideline, 5 bibliographic databases (Ovid MEDLINE, Embase, CINAHL, PsycInfo, and IEEE Xplore) were systematically searched from 2019 to May 2025, supplemented by reference screening and gray literature searches. Eligible sources were peer-reviewed empirical studies published in English that examined large language models (LLMs), AI-supported writing assistants, AI-based conversational agents, or AI-supported tools designed for automatic text simplification of PEMs. Targeted outcomes included linguistic quality (ie, linguistic comprehensibility, linguistic correctness) and content fidelity (ie, factual accuracy, factual completeness) of simplified PEMs. Excluded sources comprised rule-based systems, manual text simplification, non-laypeople target groups, and technology-focused performance metrics. Results were synthesized via thematic analysis across the domains of targeted outcomes. In accordance with JBI methodology, a risk-of-bias assessment was not performed.

Results: A total of 31 eligible studies met the inclusion criteria, examining various LLMs, including OpenAI's GPT series, Gemini, Bard, Claude, Copilot, and Llama. Specifically, GPT-4.0 achieved the most consistent improvements in standardized readability metrics (eg, the Flesch-Kincaid Grade Level [FKGL]). However, achieving predefined target RGLs remained challenging across all LLMs, particularly at lower RGLs. Findings on content fidelity were inconsistent: despite high content similarity scores, content accuracy was often compromised.

Conclusions: This is the first scoping review to comprehensively synthesize evidence on automated technologies for text simplification in PEMs. The review identified 2 critical validation gaps. First, no study examined the linguistic correctness (eg, grammar and typographical errors) of automatically simplified PEMs. Second, and most notably, the understandability of the simplified PEMs was assessed exclusively by experts, with no empirical evaluation involving laypeople. Although LLMs effectively reduce text complexity as measured by objective readability metrics, reliance on these formulas represents a critical limitation, as they serve merely as structural proxies. Improvements in readability do not guarantee the maintenance of content accuracy or laypeople's understandability. Current evidence is further limited by the lack of systematic prompt quality evaluation and the predominant focus on English-language PEMs in US contexts, restricting generalizability. This review provides a foundation for

XSL•FO
RenderX

future research by highlighting the need for validated evaluation frameworks that encompass layperson testing and content verification. For clinical practice, LLMs should currently serve as assistive tools, with mandatory expert review remaining essential to verify content fidelity before disseminating LLM-simplified PEMs to laypeople.

(J Med Internet Res 2026;28:e88365) doi:10.2196/88365

KEYWORDS

artificial intelligence; large language models; patient education materials; automatic text simplification; linguistic quality; content fidelity

Introduction

Rationale

Effective health communication is recognized as a public health priority [1]. Health communication aims to improve health by ensuring effective understanding and application of health information. Central to this is health literacy, defined as the degree to which individuals have the capacity to obtain and understand health information needed to make appropriate health decisions [2]. Patients with higher health literacy are more likely to engage in health-promoting behaviors, utilize health care services, and effectively manage chronic diseases [3]. However, the use of complex medical language poses a significant barrier to patient understanding. Health care professionals are encouraged to use plain language tailored to patients' comprehension levels, which can be challenging, especially in time-sensitive clinical care settings [4,5]. Patient education materials (PEMs) play a central role in supporting patient-physician interactions by providing clear and accessible information on health conditions, treatments, and health promotion [6]. Personalized PEMs have been shown to improve patient care through shared decision-making, enhanced patient satisfaction, and better physical and psychosocial well-being [7]. To maximize accessibility of PEMs, leading organizations such as the National Institutes of Health and the American Medical Association (AMA) recommend writing PEMs at or below a sixth-grade reading level (RGL) [8,9]. However, numerous studies have demonstrated that a significant portion of existing PEMs fail to meet this benchmark, often being written at a level too complex for many patients to understand. An analysis of PEMs showed average readability scores ranging from 8th to 15th RGL across various medical fields [10-16]. This readability concern has not improved between 2001 and 2022. These findings underscore that simplified versions of PEMs are necessary, and health care professionals are encouraged to provide easy-to-read PEMs to patients [17,18].

Artificial intelligence (AI) offers promising opportunities to enhance effective health communication. In particular, large language models (LLMs) have emerged as transformative tools in natural language processing (NLP), enabling diverse applications such as answering patient questions; summarizing, translating, or simplifying medical texts; supporting clinical paperwork; and providing individualized medical guidance [19-21]. Importantly, LLMs have the potential to enhance the accessibility of medical knowledge by making complex medical language more comprehensible to laypeople, thereby enabling patients to better understand their health conditions [22,23]. Text simplification as an NLP task has advanced significantly in recent years, especially driven by developments in LLMs since 2019. Whereas earlier technologies relied primarily on rule-based systems or machine learning models, LLMs have enabled a paradigm shift in NLP, making these capabilities more widely accessible to health care professionals and health researchers. In addition, LLM capabilities have rapidly evolved, demonstrating increasingly sophisticated language understanding and generation abilities [24,25].

Despite promising results, important challenges remain. These include the risk of factual errors (hallucinations), critical omissions of information, and the unintended loss of meaning between original and simplified text versions. Furthermore—and most critically—the difficulty of verifying understandability persists, as strong performance on standard quality metrics does not guarantee that simplified texts are actually understandable to laypeople [26]. Moreover, the rapid evolution of AI language processing technologies makes it challenging for researchers and health care providers to maintain an up-to-date overview of available tools and supporting evidence.

Objectives

This scoping review aimed to identify and map existing evidence on the use of automated language processing technologies for the simplification of PEMs into layperson-friendly language. For this review, layperson-friendly language is defined as text characterized by an accessible reading level, simple sentence structures, and explanation or avoidance of medical jargon. This scoping review is guided by the following research questions:

- What automated language processing technologies are currently used to simplify PEMs into layperson-friendly language?
- How are the linguistic quality and content fidelity of these simplified texts evaluated?

Methods

Review Principles and Protocol

This scoping review followed the methodological framework of the Joanna Briggs Institute (JBI) [27] and is reported according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [28] (Multimedia Appendix 1) and the PRISMA-S (PRISMA Statement for Reporting Literature Searches in Systematic Reviews) [29] (Multimedia Appendix 2) guidelines. The review methodology was registered in the Open Science Framework [30].

XSL•FO
RenderX

Deviation From the Protocol

The preregistered protocol specified the inclusion of studies on automatic text simplification of any complex medical or health-related text. As Figure 1 shows, full-text screening revealed a large and highly heterogeneous body of literature, encompassing different document types such as radiology reports, electronic health records, discharge letters, PEMs, informed consent forms, and scientific papers. This heterogeneity presents challenges for meaningful synthesis and comparison across fundamentally different health information materials with varying purposes, audiences, and complexity levels.

Additionally, this study was framed as part of a larger project, the A+CHIS project [31], which aims to develop a system that provides users with a selection of diverse health documents and preliminary PEMs tailored to their individual information needs and cognitive prerequisites. To present a more focused and coherent synthesis of evidence that would yield insights for patient-facing health communication systems such as A+CHIS, a post hoc decision was made to refine the scope to studies concerning PEM simplification specifically. This narrowed focus aligns with the primary goal of improving health information accessibility for laypeople and ensures methodological consistency in evaluating simplification approaches.

This post hoc decision was made after the initial full-text screening but before full-text data extraction. All other methods remained unchanged from the preregistered protocol.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for study selection. PEM: patient education material.

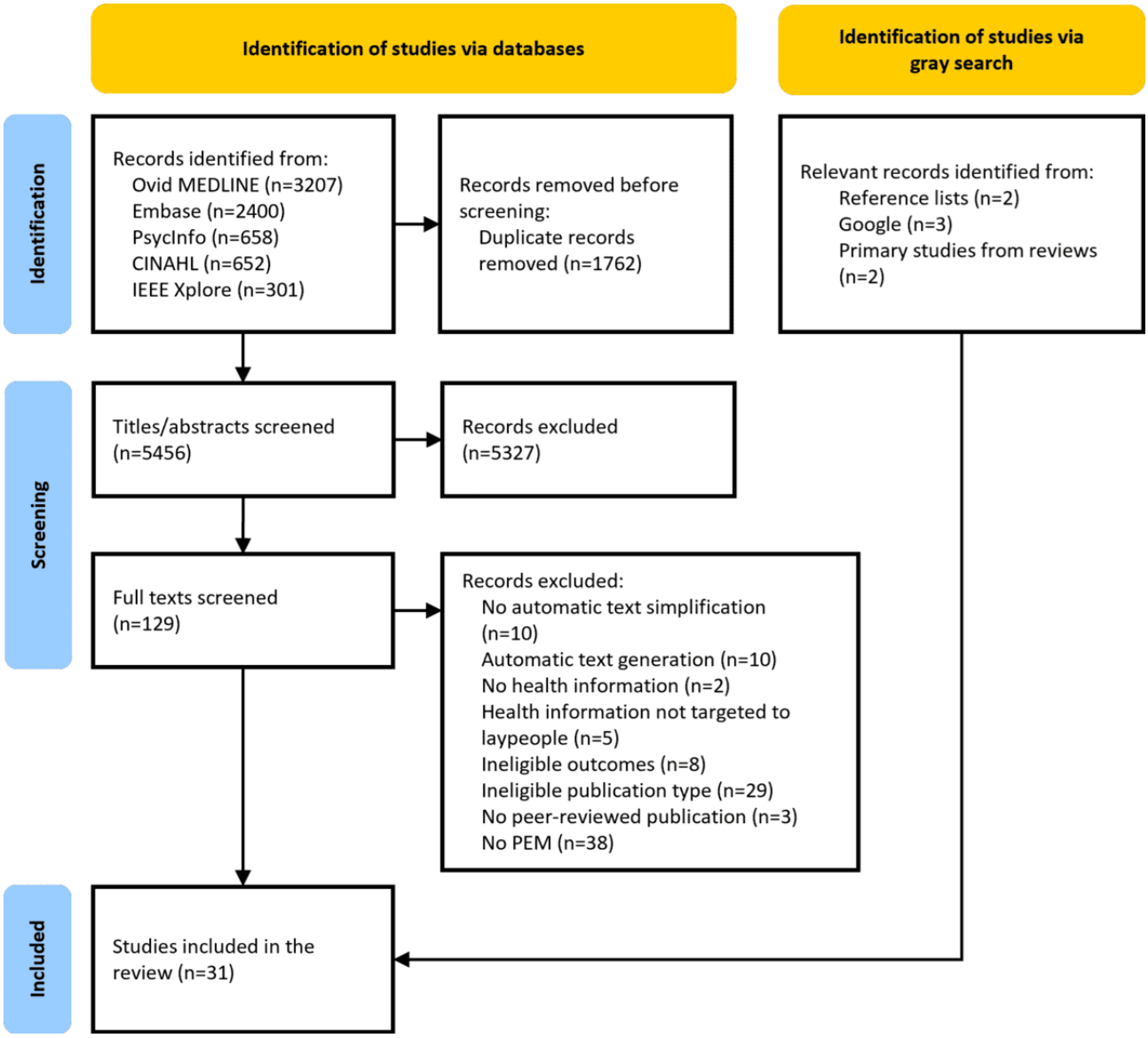

Screening and Identifying Relevant Studies: Information Sources and Search

A comprehensive literature search was conducted in May 2025 across 5 bibliographic databases: MEDLINE, Embase, CINAHL, PsycINFO, and IEEE Xplore. The MEDLINE and Embase strategies were run simultaneously as a multifile search in Ovid, and results were deduplicated using the Ovid deduplication tool. The search strategy was based on database-specific controlled vocabulary and free-text terms. To identify relevant search

XSL•FO
RenderX

terms, the MeSH Browser, a keyword analysis tool [32], and a synonym identification tool [33] were used. The search was limited to the period from 2019 to 2025, as this period marks significant advances in automated language processing technologies, particularly with the emergence of LLMs that have substantially enhanced automatic text simplification capabilities. No search filter was applied. The search block for "automated language processing technologies" was based on 2 existing search strategies [34,35]. Detailed search strategies for the databases are provided in Multimedia Appendix 3. The MEDLINE search strategy was peer reviewed by another research team member (KJ).

Reference lists of included studies were manually screened to identify additional studies. We also searched gray literature in July 2025 using the search string "patient education material" AND ("large language model" OR "ChatGPT") AND "simplifying" in Google. The first 5 pages of each search result were screened. We did not contact the authors of included studies for clarification, as the available data were sufficient to conduct the scoping review. No other methods were used to locate additional studies. Eligible studies were imported into EndNote 21.5 (Clarivate Plc) to identify and remove duplicates.

Eligibility Criteria

Framework Applied

The Population-Concept-Context (PCC) framework was used to define the eligibility criteria and guide the search strategy [27].

- Population:
  The target population was defined as laypeople requiring PEMs, including patients and their families, the general public, individuals with limited health literacy, and health care professionals without specific medical expertise.
- Concept:
  The central concept was the use of automated language processing technologies for the simplification of PEMs.
- Context:
  The context was limited to PEM simplification, defined as any text (full text or excerpt) designed to inform laypeople about medical or health-related topics. Nontext formats, such as video, audio materials, or infographics, were not considered in this review.

Building upon the PCC framework, the following inclusion and exclusion criteria were applied for study selection:

Inclusion Criteria

- Studies that assessed automated language processing technologies, specifically LLMs such as AI-supported writing assistants (eg, DeepL Write) and AI-based conversational agents (eg, ChatGPT, Gemini), or other AI-supported tools designed for automatic text simplification.
- Studies that assessed text quality indicators of simplified PEMs in at least one of two domains: (1) linguistic quality, encompassing linguistic comprehensibility (eg, readability, text complexity, word choice, structural clarity, medical jargon) and linguistic correctness (eg, grammatical accuracy, typographical errors); and (2) content fidelity, encompassing factual correctness (eg, hallucinations, exaggerations, falsifications, understatements, misinterpretations) and factual completeness (eg, content retention).
- Studies that measured outcomes either objectively (eg, FKGL) or subjectively (eg, expert ratings, user feedback).
- Peer-reviewed empirical study designs (quantitative, qualitative, and case studies) published in English, including primary studies identified through relevant reviews.

Exclusion Criteria

- Studies that evaluated nonlearning or rule-based systems, manual text simplification or human postediting of AI output, automated language translation without simplification intent, or text analysis software tools.
- Studies that evaluated the effectiveness of automated simplified PEMs targeted at health care professionals.
- Studies that evaluated only technology-oriented performance metrics (eg, Bilingual Evaluation Understudy [BLEU] or Recall-Oriented Understudy for Gisting Evaluation [ROUGE] scores).
- Non–peer-reviewed publications (eg, preprints, editorials, commentaries, letters to the editor, opinions), studies lacking a complete methodological description, and publications in languages other than English.

Selection of Sources of Evidence

Following JBI guidelines [27], 2 authors (CK, DW, CL, or NB) independently screened all titles and abstracts against the inclusion criteria. A pilot stage involving 25 titles and abstracts was conducted to ensure consistent application of the criteria, with conflicts resolved through discussion. Full texts of potentially eligible studies were subsequently assessed by 2 independent reviewers (CK, DW, CL, or NB) using the same eligibility criteria. Disagreements were resolved through discussion or by involving a third reviewer (CK, DW, CL, or NB).

Data Charting

A standardized data extraction form was developed and piloted on 3 eligible studies to assess its clarity and completeness. Following the pilot test and discussion within the research team, 1 additional item (the language of simplified PEMs) was added to the final form.

Data Items

The following data from the included studies were extracted: bibliographic details (authors, year of publication, and country), technology details (name and version of LLM and prompts used), source text (medical field of PEMs, number of analyzed materials, and language), outcomes (text quality indicators and measurement methods), and key findings (main results related to linguistic quality and content fidelity). As recommended by the JBI methodology for scoping reviews [27], 1 reviewer (CK) extracted the data, which was verified by a second reviewer (DW).

XSL•FO
RenderX

Critical Appraisal

In line with JBI guidelines [27], critical appraisal of eligible studies was not required.

Collating, Summarizing, and Reporting the Results

Following the JBI scoping review methodology [27], we conducted a descriptive synthesis of the extracted data. We did not perform an analytical synthesis of outcomes but instead mapped and summarized findings descriptively. The extracted data were collated iteratively and organized using a framework-based approach, categorizing findings into the 2 predefined outcome domains: linguistic quality (encompassing linguistic comprehensibility and linguistic correctness) and content fidelity (encompassing factual correctness and factual completeness). Within each domain, descriptive qualitative content analysis was conducted by the first author (CK) to summarize the types of automated language processing technologies evaluated, the measurement methods applied, and the reported direction of effect (eg, improvements in readability scores or identified factual inaccuracies). Results were presented descriptively through narrative summaries, tables, and figures to address the review questions. Frequency counts were used where appropriate to quantify the occurrence of specific technologies or outcome measures. All descriptive analyses were conducted using Microsoft Excel.

Results

Selection of Sources of Evidence

The systematic database search identified 7218 references. Following deduplication, 5456 titles and abstracts were screened, and 129 full texts were subsequently assessed for eligibility. This process resulted in the inclusion of 24 [36-59] studies specifically addressing PEMs. In addition, 2 studies [60,61] identified from relevant reviews were included. Two further studies [62,63] were identified through reference list screening of the included studies, and a supplementary Google search yielded 3 additional relevant studies [64-66]. In total, this scoping review included 31 studies focusing on PEMs. Figure 1 shows the study selection process in a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

Studies excluded at the full-text screening stage and their primary reasons for exclusion are provided in Multimedia Appendix 4.

Characteristics of Sources of Evidence

The main characteristics of all included studies are presented in Table 1. All included studies investigated LLMs exclusively, evaluating 8 different models. OpenAI's GPT series emerged as the most extensively studied models: GPT-4.0 (n=10) [36-43,62,64], GPT-3.5 (n=8) [44-49,60,61], and GPT-3.0 (n=1) [50]. Other evaluated LLMs included Google's Gemini (n=4) [51,52,65,66] and Bard (n=5) [53-56,63], Anthropic's Claude (n=2) [57,65], Microsoft's Copilot (n=1) [66], and Meta's Llama (n=1) [57], with 12 studies conducting comparative analyses across multiple models [51-59,63,65,66]. As illustrated in Multimedia Appendix 5, all included studies were published very recently, with 11 in 2025 [36-40,44,51,57,64-66], 19 in 2024 [41-43,45-48,50,52-56,58-63], and 1 in 2023 [49]. Although the literature search covered the period from 2019 onward, no relevant studies were identified before 2023, highlighting the emerging nature of this field.

XSL•FO
RenderX

Table 1. Summary of main characteristics of the included studies.

Study	Country	Study design	LLM[a]; number of PEMs[b] analyzed (language of PEMs)	Medical field in PEMs	Readability scale	Target RGL[c,d]
Spina et al [36]	United States	Not reported	GPT-4.0; 9 (English)	Glaucoma	FKGL[e] and FRE[f]	5
Reaver et al [37]	United States	Cross-sectional study	GPT-4.0; 57 (English) and 56 (Spanish)	Orthopedic surgery	FRE, Fry, LIX[g], RIX[h], SMOG[i], GPM[j] Fry, FHRI[k], and SOL[l]	6
Picton et al [38]	United States	Not reported	GPT-4.0; 340[m] (English)	Neurology and neurological surgery	FKGL and FRE	5
Li et al [44]	United States	Cross-sectional comparative study	GPT-3.5; 50 (English)	Cataract surgery	FKGL, SMOG, GFI[n], and CLI[o]	5
Dihan et al [51]	United States	Cross-sectional comparative study	GPT-3.5, GPT-4.0, and Gemini Advanced; 20 (English)	Dry eye disease	FKGL and SMOG	6
Chandra et al [39]	United States	Not reported	GPT-4.0; 30 (English)	Orthopedic surgery	SMOG	6-8
Busigo Torres et al [40]	United States	Not reported	GPT-4.0; 77 (Spanish)	Orthopedic	FHRI and SOL	5
Andalib et al [57]	United States	Not reported	GPT-3.5, GPT-4.0, Claude 2, and Llama 2; 48 (English)	Orthopedic	FKGL and FRE	5
Will et al [65]	United States	Cross-sectional study	GPT-4.0, Gemini 1.5-flash, and Claude 3.5 Sonnet; 60 (English)	Heart disease, cancer, and stroke	FKGL, FRE, SMOG, and GFI	5
Naghdi et al [66]	The Netherlands	Comparative observational study	GPT-3.5, GPT-4.0, Copilot, and Gemini; 30 (English)	Reproductive genetics	FKGL, FRE, SMOG, GFI, CLI, and LWF[p]	6-8
Singh et al [64]	United States	Not reported	GPT-4.0; 25 (English)	Neurosurgery	FKGL, FRE, SMOG, CLI, and ARI[q]	8
Zaki et al [41]	United States	Not reported	GPT-4.0; 73 (English)	Interventional radiology procedures	FRE, GFI, and ARI	5
Vallurupalli et al [45]	United States	Not reported	GPT-3.5; 18[r] (English)	Hand surgery	Combined calculator: FKGL, FRE, SMOG, ARI, GFI, LWF, and CLI	6-8
Shehab et al [42]	United States	Cross-sectional study	GPT-4.0; 124 (English)	Cleft lip and palate	FKGL and FRE	6
Patel et al [43]	United States	Proof - of - concept study	GPT-4.0; 71 (English)	Otolaryngology	FKGL, FRE, SMOG, and GFI	6
Oliva et al [46]	United States	Cross-sectional study	GPT-3.5; 109 (English)	Otolaryngology	FKGL and FRE	5
Kianian et al [53]	United States	Not reported	GPT-4.0 and Bard; 9 (English)	Uveitis	FKGL	6
Rasika et al [47]	United States	Not reported	GPT-3.5; 15 (English and Spanish)	Ophthalmology	FKGL, FRE, GFI, FHRI, Crawford Nivel-de-Grado, Gutiérrez, Szigriszt-Pazos/INFLESZ[s], and Legibilidad-µ	Nr
Gupta et al [52]	United States	Not reported	GPT-4.0 and Gemini; 7 (English)	Radiology	FKGL, FRE, SMOG, and GFI	6
Garcia Valencia et al [58]	United States and Thailand	Not reported	GPT-3.5 and GPT-4.0; 27[t] (English)	Living kidney donation	FKGL	8
Fanning et al [59]	United States	Not reported	GPT-3.5 and GPT-4.0; 75 (English)	Plastic surgery	FKGL, FRE, Fry, SMOG, GFI, and Raygor Estimate	6

XSL•FO
RenderX

Study	Country	Study design	LLM[a]; number of PEMs[b] analyzed (language of PEMs)	Medical field in PEMs	Readability scale	Target RGL[c,d]
Dihan et al [54]	United States	Cross-sectional comparative study	GPT-3.5, GPT-4.0, and Bard; 20 (English)	Idiopathic intracranial hypertension	FKGL and SMOG	6
Dihan et al [55]	United States	Cross-sectional comparative study	GPT-3.5, GPT-4.0, and Bard; 20 (English)	Childhood glaucoma	FKGL and SMOG	6
Dihan et al [56]	United States	Cross-sectional comparative study	GPT-3.5, GPT-4.0, and Bard; 20 (English)	Pediatric cataract	FKGL and SMOG	6
Baldwin [50]	United Kingdom	Not reported	GPT-3.0; 50 (English)	Burns first aid	FKGL, FRE, SMOG, GFI, and CLI	6
Ayre et al [48]	Australia	Observational study	GPT-3.5; 26[u] (English)	Not restricted[v]	SMOG	8
Manasyan et al [60]	United States	Not reported	GPT-3.5; 34 (English)	Alveolar bone grafting	FKGL, FRE, and GFI	5
Vallurupalli et al [61]	United States	Not reported	GPT-3.5; 18 (English)	Craniofacial procedures	Combined calculator: FKGL, FRE, SMOG, ARI, GFI, LWF, and CLI	8
Abreu et al [62]	United States	Cross-sectional study	GPT-4.0; 34 (English)	Cancer	FKGL, SMOG, Fry, and GFI	6
Rouhi et al [63]	United States	Pilot study	GPT-3.5 and Bard; 21 (English)	Aortic stenosis	FKGL, FRE, SMOG, and GFI	5
Kirchner et al [49]	United States	Proof-of-concept study	GPT-3.5; 20 (English)	Orthopedic	FKGL and FRE	5

[a]LLM: large language model.
[b]PEM: patient education material.
[c]RGL: reading grade level.
[d]According to the US grade school-level system.
[e]FKGL: Flesch-Kincaid Grade Level.
[f]FRE: Flesch Reading Ease.
[g]LIX: Läsbarhetsindex Index.
[h]RIX: Rate Index.
[i]SMOG: Simple Measure of Gobbledygook.
[j]GPM Fry: Gilliam Peña Mountain Fry Graph.
[k]FHRI: Fernandez-Huerta Readability Index.
[l]SOL: Spanish Orthographic Length.
[m]274 PEMs about neurology, 66 PEMs about neurological surgery.
[n]GFI: Gunning Fog Index.
[o]CLI: Coleman-Liau Index.
[p]LWF: Linsear Write Formula.
[q]ARI: Automated Readability Index.
[r]Excerpts of PEMs.
[s]INFLESZ: Índice Flesch-Szigriszt.
[t]Frequently asked questions.
[u]Extracts were at least 300 words.
[v]Online health information published by recognized national and international health information provider websites.

The geographical distribution showed a strong US dominance (n=27) [36-47,49,51-57,59-65], with additional contributions from the United Kingdom [50], Australia [48], the Netherlands [66], and a US-Thailand collaboration [58]. This geographic focus is reflected in the languages studied, with 30 studies examining English PEMs [36-39,41-66], 2 evaluating both English and Spanish PEMs [37,47], and 1 focusing solely on Spanish PEMs [40].

Sample sizes varied considerably, ranging from 7 to 340 PEMs per study, including full texts, text excerpts [45,48], and frequently asked question sections [58].

Figure 2 illustrates the distribution of medical fields addressed in the PEMs in the included studies. Surgery was the most frequently investigated field (n=14), encompassing orthopedics, neurosurgery, plastic surgery, oral and maxillofacial surgery, craniofacial surgery, and otolaryngology. Ophthalmology was the second most common field (n=7), covering topics such as

XSL•FO
RenderX

glaucoma, cataract, uveitis, and dry eye disease. Additional fields included internal medicine (n=4; heart disease, aortic stenosis, cancer), neurology (n=3; general neurology, stroke, idiopathic intracranial hypertension), and diagnostic radiology (n=2). Single studies addressed emergency medicine (burns first aid) [50], obstetrics and gynecology (reproductive genetics) [66], and urology (renal transplantation) [58]. No studies were identified in the following medical fields: allergy and immunology, anesthesiology, dermatology, dentistry, family medicine, medical genetics, nuclear medicine, pathology, pediatrics, physical medicine and rehabilitation, preventive medicine, psychiatry, or radiation oncology.

To evaluate the effect of prompt engineering, 1 study [59] compared the outputs of GPT-3.5 and GPT-4.0 using 2 different prompts. The first was a simple prompt with a general instruction to simplify the text while maintaining its structure. The second, more detailed prompt provided explicit constraints, referencing specific readability scales and including examples to guide the model's output [59]. All prompts used in the included studies are presented in Multimedia Appendix 6.

A comprehensive overview of the reported outcomes across all included studies is provided in Multimedia Appendix 7.

Figure 2. Distribution of medical fields represented in patient education materials across the included studies.

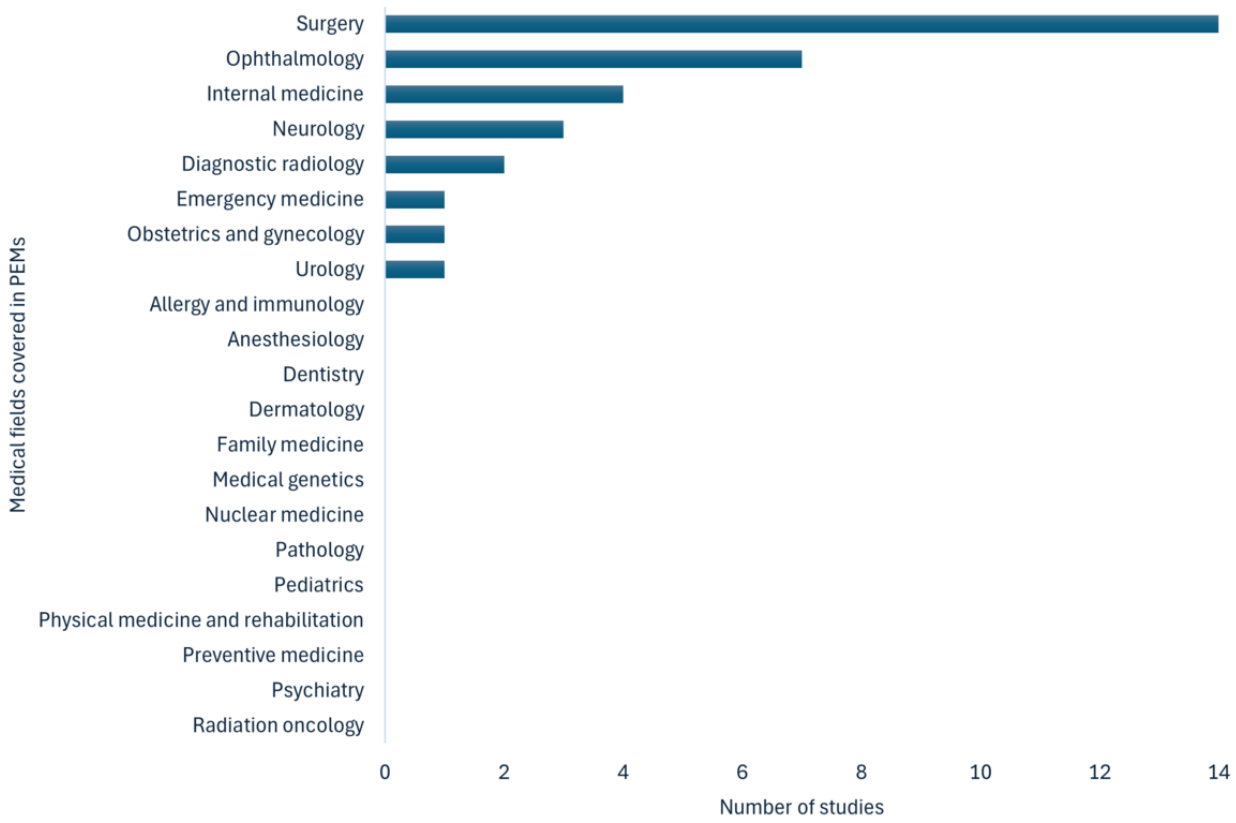

Linguistic Quality

Assessment of Linguistic Comprehensibility Versus Absence of Linguistic Correctness Evaluation

All 31 included studies assessed the linguistic quality of simplified PEMs using at least one indicator of linguistic comprehensibility. Notably, none of the studies evaluated linguistic correctness, such as grammatical accuracy or typographical errors.

Linguistic Comprehensibility

Readability

All included studies assessed linguistic comprehensibility using at least one readability scale. The most frequently used scales for English-language PEMs were the FKGL (n=23 studies) [38,42-44,46,47,49-60,62-66], Flesch Reading Ease (FRE; n=18) [36-38,41-43,46,47,49,50,52,57,59,60,63-66], and Simple Measure of Gobbledygook (SMOG; n=17) [37,39,43,44,48,50-52,54-56,59,62-66]. Additional metrics included the Gunning Fog Index (GFI; n=12) [41,43,44,47,50,52,59,60,62,63,65,66], Coleman-Liau Index (CLI; n=4) [44,50,64,66], Automated Readability Index (ARI; n=3) [41,52,64], Fry score (n=3) [37,59,62], and single-use scales such as the Raygor Estimate, Rate Index (RIX) [59], Läsbarhetsindex (LIX) [37], and Linsear Write Formula (LWF) [66]. Two studies used a composite readability scoring system based on the average of 7 readability indices (FRE, FKGL, SMOG, GFI, CLI, ARI, and LWF) [45,61]. Several studies employed multiple readability scales for comparative analysis. Table 2 provides an overview of the readability scales used to assess English-language PEMs in the included studies, along with their underlying scoring components used to measure readability.

XSL•FO
RenderX

As Figure 3 shows, readability improvement varied by LLM and readability scale in English-language PEMs. GPT-4.0, the most frequently evaluated model, demonstrated the most consistent performance, with a proportion of analyses showing readability improvement falling into the “91%-100%” category across all metrics, except CLI and ARI (67%). By contrast, GPT-3.5 showed more variable results: a 100% success rate on SMOG but lower rates of improvement for FKGL (77%), FRE (75%), and GFI (71%). Other LLMs (GPT-3.0, Gemini, Bard, Claude, Llama 2, and Copilot) also achieved high proportions of analyses showing significant improvements; however, evidence remains limited to single studies for most models.

For Spanish-language PEMs, 3 studies [37,40,47] employed language-specific readability scales such as the Fernández-Huerta Readability Index, GPM Fry Graph, LIX, RIX, Spanish Orthographic Length (SOL), Crawford Nivel-de-Grado, Gutiérrez, Szigriszt-Pazos/Índice Flesch-Szigriszt (INFLESZ), and Legibilidad-µ, achieving improvements in 9 out of 12 (75%) analyses.

Table 2. Readability scales used in included studies and corresponding scoring components [67,68].

Scale	Components to score readability
Flesch-Kincaid Grade Level	• Average number of syllables per word • Average number of words per sentence
Flesch Reading Ease	• Average number of syllables • Average number of words per sentence • Average number of sentences
Simple Measure of Gobbledygook	• Average number of words with ≥3 syllables • Average number of sentences
Gunning Fog Index	• Number of sentences • Number of words • Number of words with ≥3 syllables
Coleman-Liau Index	• Average number of letters per 100 words • Average number of sentences per 100 words
Automated Readability Index	• Average number of characters per word (eg, any letters, numbers, symbols) • Average number of words per sentence
Fry	• Average number of sentences • Syllables per 100 words
Raygor Estimate	• Average number of sentences • Long words (≥6 characters) per 100 words
Rate Index	• Number of long words • Number of sentences
Läsbarhetsindex Index	• Average number of words per sentence • Percentage of words with >6 letters
Linsear Write Readability	• Number of easy words (words with ≤2 syllables) • Number of hard words (words with ≥3 syllables)

XSL•FO
RenderX

Figure 3. Map analysis of significant readability improvements in English-language patient education materials across all studies, achieved using various large language models and assessed with different readability scales. ARI: Automated Readability Index; CLI: Coleman-Liau Index; FKGL: Flesch-Kincaid Grade Level; FRE: Flesch Reading Ease; GFI: Gunning Fog Index; LIX: Läsbarhetsindex Index; LWF: Linsear Write Formula; RIX: Rate Index; SMOG: Simple Measure of Gobbledygook.

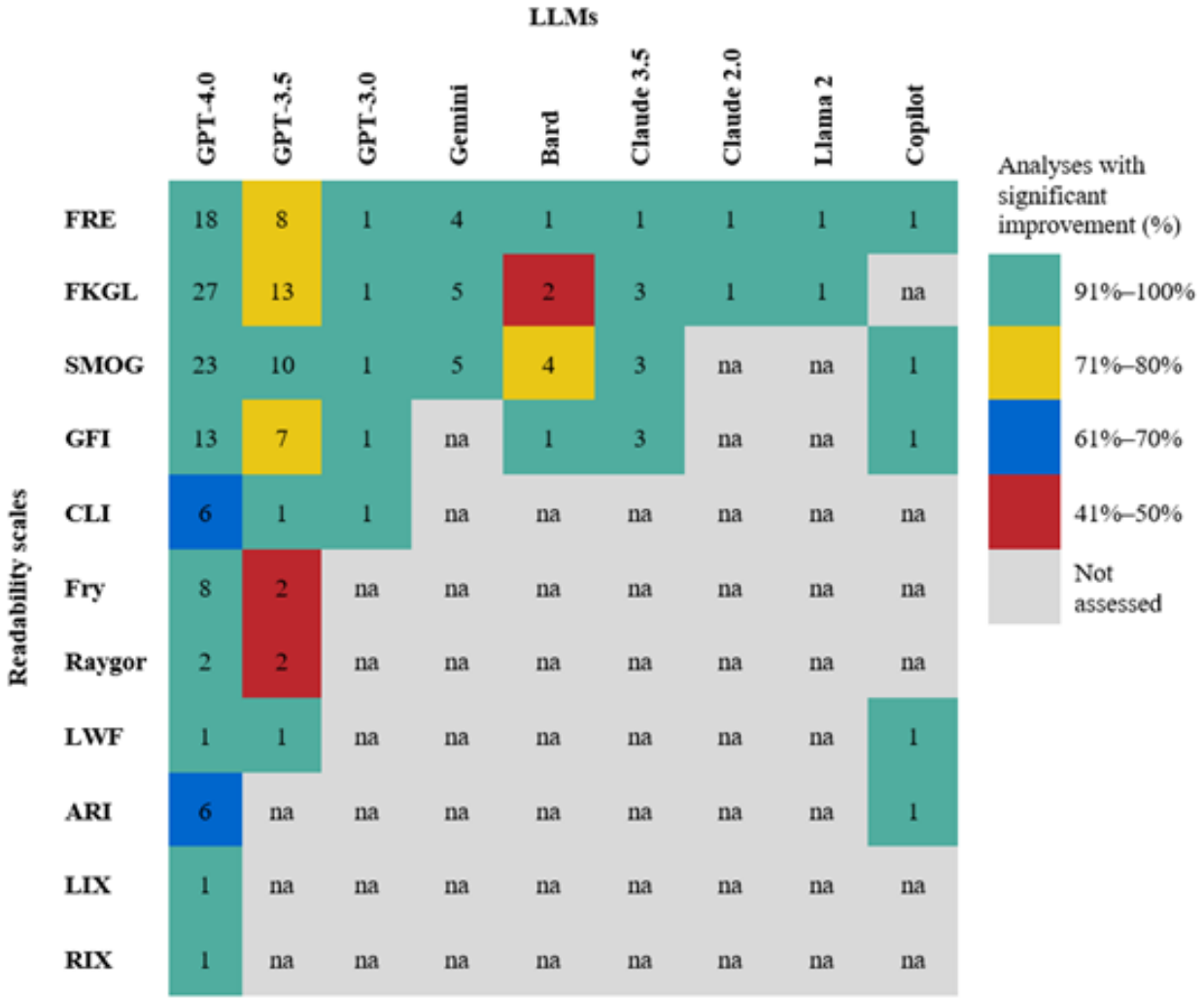

Reading Grade Levels

In addition to objective measurements of readability, nearly all included studies (n=30) investigated whether LLMs could achieve predefined target reading grade levels (RGLs) specified in the prompts, typically ranging from fifth to eighth grade. The detailed original, target, and achieved RGLs after prompting LLMs to rewrite PEMs to a specific RGL across all included studies can be found in Figure 4 (also see [38-40,42-46,48-52,54-59,61-63,66-68]).

As Figure 4 shows, the studies demonstrated variable success rates for GPT models. At the fifth-grade target level, GPT-3.5 and GPT-4.0 demonstrated their lowest performance, achieving success in 4 out of 10 (40%) and 2 out of 7 (29%) attempts, respectively. Performance improved notably at the sixth-grade level, where GPT-3.5 reached the target in 3 out of 6 (50%) cases, whereas GPT-4.0 showed success in 8 out of 12 (67%) cases. At the eighth-grade level, an inverse pattern emerged: GPT-3.5 achieved a success rate of over 65% (2/3, 67%), whereas GPT-4.0 succeeded in only <35% of attempts (2/6, 33%).

Among other LLMs, Gemini demonstrated the strongest performance, with an 83% success rate, followed by Claude models at 75%. Conversely, Bard exhibited notably poor performance, achieving target RGLs in only 20% of attempts. Evidence for GPT-3.0 and Llama 2 is limited to single studies, precluding meaningful comparison.

Three studies explored flexible target ranges (sixth to eighth grade) rather than fixed targets, with only 2 of 6 analyses achieving the sixth-grade level recommended by the AMA [39,45,66].

Analysis of Spanish-language PEMs achieving specific RGLs showed mixed results: 1 study successfully reduced readability from sixth to fifth grade using GPT-4.0 [40], whereas another achieved only a minimal reduction (ninth to eighth grade) despite targeting the sixth grade [37]. The third study did not report a specific target RGL [52].

XSL•FO
RenderX

Figure 4. Comparison of original and achieved RGL after prompting various LLMs to simplify PEMs to a predefined target RGL. LLMs: large language models, PEM: patient education material, RGL: reading grade level.

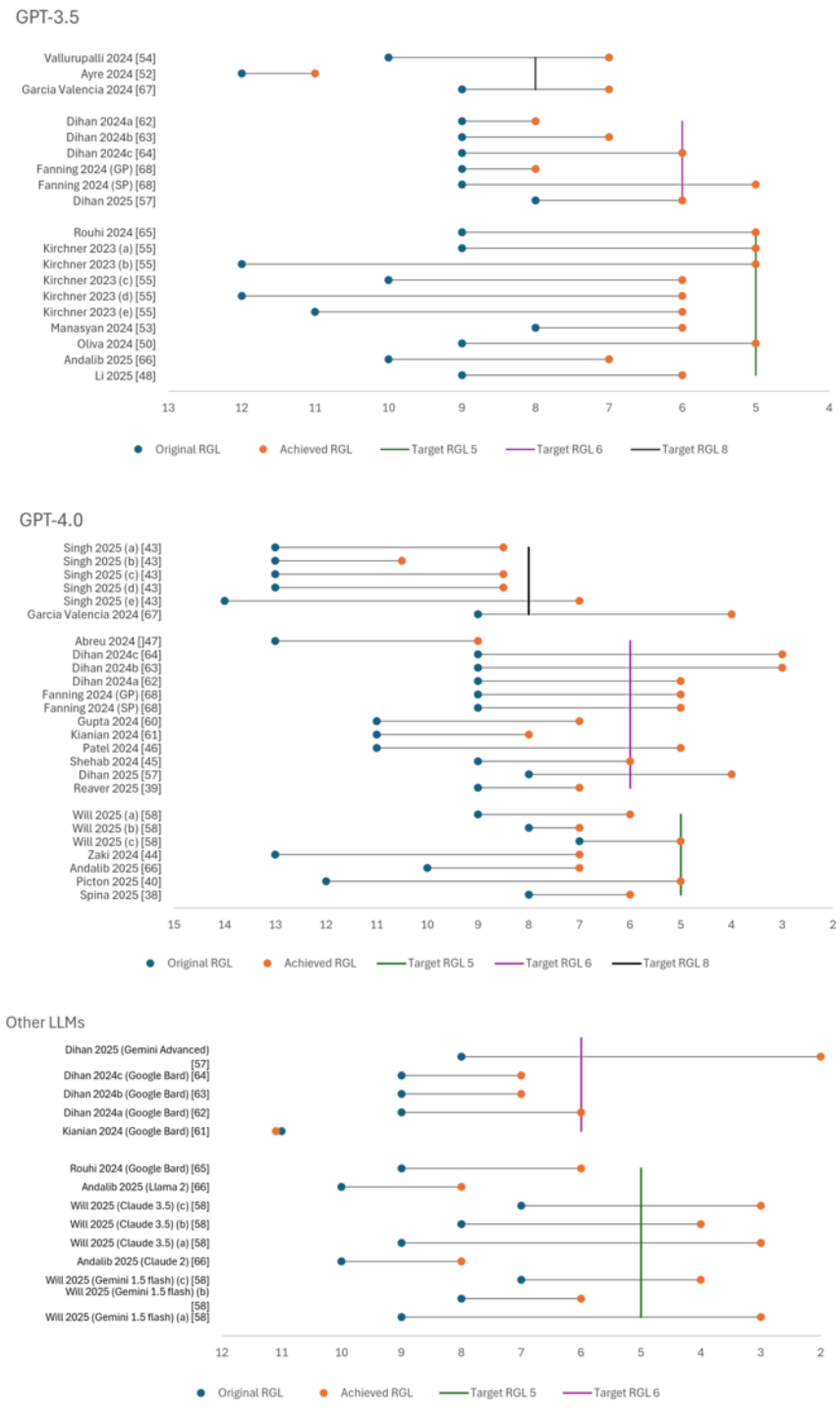

Other Linguistic Comprehensibility Indicators

The ability to reduce *text length* was assessed through word count in 14 studies [37,38,40,42,48,51,52,54-57,60,65,66] and sentence count in 5 studies [51,54,55,57,66]. GPT-3.5 reduced word count in all 8 analyses, and GPT-4.0 achieved reductions in nearly 80% of cases (15/19, 79%). Both models maintained readability improvements alongside text reduction in most instances: GPT-3.5 in 75% and GPT-4.0 in 89%. Other LLMs, such as Bard, Claude (2.0 and 3.5), Llama 2, and Copilot, similarly achieved word count reductions while maintaining readability. Sentence count decreased in 13 out of 17 (76%) analyses, with no notable differences between LLMs.

Lexical complexity was assessed in 8 studies [37,51,54-57,60,62]. All tested LLMs (GPT-3.5, GPT-4.0, Bard, Gemini, and Llama 2) reduced syllable-based metrics (polysyllabic words, syllables per word, and syllables per

XSL•FO
RenderX

sentence), except Claude 2 in 1 analysis [57]. GPT-4.0 also reduced the number of words exceeding 6 characters [62].

Syntactic complexity was assessed in 3 studies [37,62,66]. Multiple LLMs (GPT-3.5, GPT-4.0, Gemini, and Copilot) successfully reduced long sentences (>20-22 words) and passive voice usage.

Vocabulary complexity was assessed in 2 studies. LLMs effectively simplified vocabulary, reducing medical jargon, acronyms, uncommon words, and complex words [48,66].

Spanish-language studies also confirmed reductions in word count, syllable count, and long sentence frequency [37,40].

Content Fidelity

Identified Studies

A total of 20 studies examined whether the simplified texts remained factually correct and complete [36-38,40,43,46,48,49,51,52,54,55,57,58,61-66].

Factual Correctness

Factual correctness was assessed through content similarity in 6 studies [36-38,57,58,62] and content accuracy in 16 studies [37,38,40,43,46,49,51,54,55,58,61-66].

Content similarity was evaluated through automated and human assessment methods. Automated analyses using latent semantic analysis with cosine similarity values consistently demonstrated high semantic preservation. GPT-4.0 showed the strongest evidence base, producing "near identical" outputs in all 4 analyses [36,37,57,62], with 3 additional studies reporting "high similarity" [37,38,58]. GPT-3.5, Claude 2, and Llama 2 each achieved "near identical" similarity in single analyses [57]. Expert evaluations, conducted by either research team members or specialized health care professionals, further supported these findings. Human raters consistently confirmed that simplified texts did not contain extraneous information across all tested LLMs [57]. One study reporting quantitative expert assessment of GPT-4.0 yielded a mean similarity score of 0.72 on a 0-1 scale (where 1 represents identical content) [38].

Content accuracy was generally high but varied across studies. Automated F_1-scores, which reflect precision and recall (ie, whether a statement is true or false and whether it is present or absent), ranged from 72% to 92% for GPT-4.0 across different studies [37,38,62]. Expert assessments were inconsistent. For GPT-4.0, 8 studies reported no factual errors [43,51,53-55,58,64,65], 1 reported 90%-100% accuracy across 3 raters [52], and 2 studies using 5-point Likert scales reported accuracy between 3.55 and 4.1 (5=completely accurate) [38,66]. One notable outlier reported only 52% factual correctness [40]. For GPT-3.5, several studies reported no factual inaccuracies [46,49,51,54,55,61,63], and where a 5-point Likert scale was used, it scored 3.8 [66]. Other LLMs showed similar variability: Gemini received expert accuracy or suitability ratings of 48%-71% (with up to 14% judged inaccurate) [52], whereas other studies reported error-free outputs [51,65] or 10% inaccuracy rates [65]. Claude showed 5% inaccuracy in some analyses; Copilot received a mean score of 3.51; and Bard achieved 88.9% accuracy and suitability in 1 study [53] while being reported as error-free in other studies [54,55,63].

Factual Completeness

Factual completeness was assessed through content retention [43,46,48,49,52,66] and content omission [66] in fewer studies.

Content retention varied considerably across the evaluated LLMs. GPT-4.0 demonstrated the most robust performance when compared with GPT-3.5, Gemini, and Copilot [66]. Furthermore, GPT-4.0 maintained sufficient detail with some redundancies (eg, chronic cough was mentioned as a symptom of bronchitis 4 times) [43], and 95% of its outputs retained at least 75% of information [52]. Conversely, Gemini demonstrated reduced retention, with only 68% of outputs achieving the same retention threshold [52]. GPT-3.5 showed inconsistent retention, with studies reporting sufficient detail preservation [49], 80% average content retention [48], or truncation issues in 9% of outputs [46].

Content omission was evaluated in 1 study [66]. GPT-4.0 achieved the highest ratings for preserving essential information while simplifying, followed by Copilot. Conversely, Gemini received the lowest scores [66].

Overview of Text Quality Indicator Assessment Methods

As highlighted in Figure 5, linguistic comprehensibility was predominantly assessed using automated methods, as the manual assessment of objective indicators such as readability, sentence length, or word count is inefficient. By contrast, verifying factual correctness and completeness, as well as evaluating the understandability of simplified PEMs, inherently requires human judgment and cannot be fully automated. While content fidelity of simplified PEMs in the included studies was assessed by experts, this occurred less frequently overall. Crucially, the analysis reveals 2 major validation gaps: first, no study assessed the linguistic correctness of the simplified PEMs. Second, and most notably, no study involved laypeople in assessing the actual understandability of the simplified PEMs.

XSL•FO
RenderX

Figure 5. Evidence gap map of assessment methods used across studies for outcome domains (linguistic quality, content fidelity, and understandability) of simplified patient education materials, stratified by assessment type (laypeople, experts, and automated methods). AM: automated method.

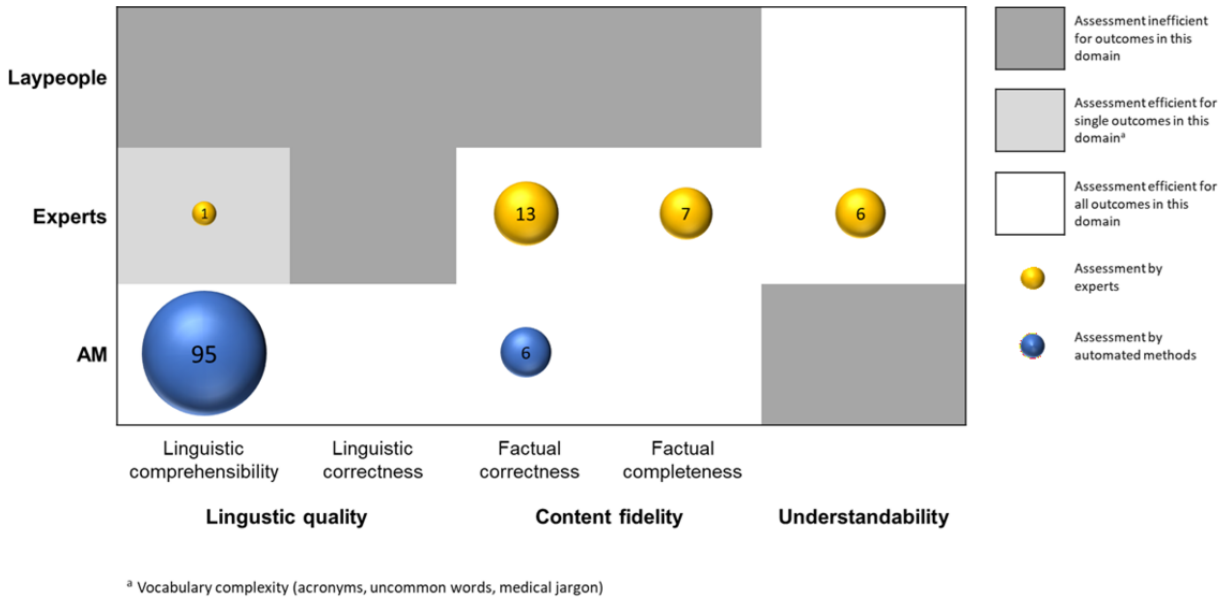

Relationship Between Readability, Content Accuracy, and Understandability of Simplified PEMs

A total of 6 studies evaluated the understandability and clarity of simplified PEMs through human expert assessments, including research team members or health care professionals, using validated tools such as the Patient Education Materials Assessment Tool [51,54,55,65] and subjective assessments of whether the PEM was understandable for an average patient or successfully improved accessibility without losing key information [52,58]. However, only 4 studies directly examined the relationship between readability improvement, content accuracy, and overall expert-rated understandability of simplified PEMs [54,55,58,65] (Table 3).

Table 3. Relationship between readability, content accuracy, and understandability.

Large language model	Readability improvement[a], n/N (%)	Content accuracy maintenance[b], n/N (%)	Understandability success[c], n/N (%)	Medical topic	Studies
GPT-4.0	9/9 (100)	6/6 (100)	3/6 (50)	✓ [d]Childhood glaucoma, cancer, and living kidney donation ✗ [e]Idiopathic intracranial hypertension, heart disease, and stroke	[54,55,58,65]
GPT-3.5	3/4 (75)	2/2 (100)	0/2 (0)	✗ Idiopathic intracranial hypertension and childhood glaucoma	[54,55]
Bard	4/4 (100)	2/2 (100)	0/2 (0)	✗ Idiopathic intracranial hypertension and childhood glaucoma	[54,55]
Gemini 1.5	4/4 (100)	2/3 (67)	1/3 (33)	✓ Cancer ✗ Heart disease and stroke	[65]
Claude 3.5	4/4 (100)	1/3 (33)	1/3 (33)	✓ Heart disease ✗ Cancer and stroke	[65]

[a]Number of analyses with significant improvement in at least one readability metric (Flesch-Kincaid Grade Level, Flesch Reading Ease, Simple Measure of Gobbledygook, or Gunning Fog Index).

[b]Number of analyses where no factual errors were detected.

[c]Judgment by human experts

[d]Understandable.

[e]Not understandable.

The studies confirmed that GPT-3.5, GPT-4.0, Bard, Gemini 1.5, and Claude 3.5 consistently improved readability scores across standard metrics, except for GPT-3.5 using FKGL in 1 analysis [54]. Content accuracy was generally high but varied by LLM and medical field. GPT-3.5, GPT-4.0, and Bard achieved 100% content accuracy across multiple conditions. However, Gemini 1.5 demonstrated 10% inaccuracy for heart disease content while maintaining 100% accuracy for cancer

XSL•FO
RenderX

and stroke. Claude 3.5 showed the opposite pattern: 100% accuracy for heart disease but 5% inaccuracy for cancer and stroke. Understandability proved most challenging. GPT-4.0 emerged as the most capable, successfully generating readable, accurate, and understandable materials for childhood glaucoma [55] and kidney donation [62]; however, it failed for idiopathic intracranial hypertension [45]. Other models demonstrated greater limitations: GPT-3.5 and Bard generated readable and accurate outputs that experts deemed not understandable across multiple evaluations. Disease-specific performance varied substantially: GPT-4.0 generated understandable PEMs for cancer but not for cardiovascular conditions, whereas Claude 3.5 achieved understandability only for heart disease content [65].

Discussion

Principal Findings

This scoping review synthesized evidence on automated language processing technologies currently used to simplify PEMs into layperson-friendly language. A total of 31 studies met the inclusion criteria and examined LLMs for automatic text simplification. Notably, despite a comprehensive literature search, no studies specifically examined AI-supported writing assistants (eg, DeepL Write) or other AI-supported tools designed for automated text simplification. This scoping review has 3 key findings.

First, LLMs consistently improve readability metrics, with GPT-4.0 demonstrating the most reliable performance in comparison with other LLMs. However, achieving predefined RGLs remains challenging across all LLMs, particularly for GPT models at the lowest target level (fifth grade). Notably, GPT-4.0 does not consistently outperform GPT-3.5; rather, performance varies by target RGL. This indicates that newer models are not automatically better at achieving specific RGLs than older ones. Although Gemini and Claude show higher success rates, limited analyses preclude definitive conclusions. At the current stage of technological development, LLMs struggle to consistently achieve the recommended RGL for optimal lay comprehension, particularly the sixth-grade level recommended by the AMA. However, based on the available data, it remains unclear whether the failure to consistently achieve target RGLs stems from inherent model limitations or from heterogeneity in prompting strategies across the included studies.

Second, content fidelity assessments showed considerable variability. Although content similarity scores were generally high, content accuracy ranged from 48% to 100% across studies. This level of inaccuracy is concerning in medical contexts, where even minor errors can compromise patient safety.

Third, and most critically, this scoping review reveals a fundamental validation gap: no study has evaluated linguistic correctness (eg, grammar or typographical errors), and no study has assessed whether laypeople actually understand the simplified PEMs. The absence of reporting on linguistic correctness indicators suggests that authors implicitly assume that LLM-simplified texts are grammatically correct. However, this assumption can be problematic in medical contexts, where ambiguous phrasing or grammatical errors can alter the meaning of health instructions and potentially cause harm to patients. For instance, even minor errors in negation or word order can result in dangerous misinterpretations. Therefore, the failure to systematically assess and report linguistic correctness represents both a methodological oversight and a patient safety concern. Furthermore, the few studies that assessed overall understandability relied exclusively on expert assessments. However, experts such as health care professionals cannot reliably predict what laypeople understand or do not understand. This validation gap reveals a critical conceptual conflation: readability, as measured by formulas, captures surface-level text characteristics such as sentence length, word length, and syllable counts. However, it does not measure whether a reader actually understands the content. By contrast, understandability refers to the extent to which readers can extract meaning from text, apply the information to their own situation, and make informed decisions based on what they read. Consequently, improved readability metric scores do not guarantee that medical content is also factually correct, contextually appropriate, or genuinely understandable to the target group. This distinction is particularly significant in medical contexts, where patients must not only read but also correctly interpret and act upon health information. The complete absence of patient-centered outcome assessment represents the most significant finding of this scoping review.

Comparison With Previous Work

To the best of our knowledge, this is the first scoping review focusing specifically on automatic text simplification methods for PEMs. This distinguishes our work from previous research on medical AI, which has predominantly focused on text generation [69-78] rather than text simplification or the simplification of clinical records [79-82], rather than PEMs.

Nguyen et al [83] conducted a systematic review and meta-analysis of online PEMs related to cleft lip and palate. In line with our findings, they identified the critical absence of understandability testing, as only 1 study directly assessed patient understanding of simplified texts. In contrast to their disease-specific approach, our review spans a broader range of medical topics, evaluates linguistic quality and content fidelity, and employs a broader search strategy.

In a broader scoping review, Aydin et al [84] examined LLM applications across multiple domains of patient care, including education, engagement, workload reduction, patient-centered health customization, and communication. However, their search was limited to PubMed and was conducted in June 2024. They similarly reported readability improvements in studies addressing automatic text simplification. Their conclusion that LLMs can create accessible materials, help interpret complex information, and enhance patient-provider communication—while also noting that accuracy, readability issues, and ethical concerns require further development—aligns with the findings of our review.

The complete absence of direct testing of understandability with laypeople in all 31 studies is significant, reflecting a systemic methodological limitation in this field: the uncritical acceptance

XSL•FO
RenderX

of readability formulas as valid measures of text simplification. These findings are further supported by studies indicating that standard readability formulas have important limitations. They primarily count variables such as sentence length, word length, or polysyllabic words, but do not assess whether the text aligns with the actual comprehension, context, expertise, semantic understanding, and textual coherence of the target audience. Common readability formulas cannot adequately assess the actual understandability of medical texts [85]. Furthermore, readability formulas may judge short technical terms and abbreviations as simple, despite being far less understandable than longer, everyday descriptions [86,87]. For example, the term "HbA_{1c}" is very short but often unknown, whereas the term "long-term blood sugar" is longer but more self-explanatory. Additionally, in-text explanations of terms are often needed for comprehension; however, they can lengthen and complicate sentences [88].

Research Implications

Overview

The findings of this review highlight several research gaps, constituting an urgent call to action for future studies. The validation gap identified in this review—namely, the complete absence of patient-centered comprehension testing and linguistic correctness assessment—should be addressed as a priority in future research. The following research gaps should guide the field forward:

Patient-Centered Evaluation

This represents the most significant research gap. Future studies must shift from purely algorithmic readability evaluation to patient-centered understandability testing [37,46,47,55-57,60,62,66]. Future studies should involve laypeople with varying health literacy levels, age, education, and cultural backgrounds in understandability assessments using validated methods, rather than relying solely on expert judgment or readability formulas [44,48,58]. For instance, Ondov et al [89] suggested 2 methods of comprehension assessment:

- Multiple-choice questions: After reading the original or simplified text, users answer constructed items that require understanding of the text to respond correctly. These questions are reliable when developed and validated properly, but are labor-intensive and require domain and assessment expertise.
- Cloze tests: These tests involve deleting words in a passage and asking users to supply the missing terms from context. They correlate well with other comprehension measures and are largely automatable. Common variants include basic cloze (masking important content words) and multiple-choice cloze (using distractors for each blank).

Prompt Engineering

Prompt engineering requires systematic investigation. In this scoping review, only 1 study [59] systematically investigated prompt design. Fanning et al [59] compared general and specific prompts using GPT-3.5 and GPT-4.0 and demonstrated that prompt design can markedly influence output quality. The study shows that well-designed, specific prompts enabled GPT-3.5 to achieve readability improvements comparable to GPT-4.0, suggesting that prompt optimization may be more influential than model generation for text simplification tasks. Their specific prompts included explicit references to readability measures, detailed instructions for writing style (eg, shortening sentences and simplifying words), specification of target readability levels, and explicit constraints requiring that rewritten texts maintain the original meaning and information. While these findings are based on a single study, they offer a possible explanation for the inconsistent results observed across the included studies. The wide variability in content accuracy (48%-100%) and frequent failure to achieve specific RGLs could be attributed to differences in prompt design rather than solely to inherent model limitations. However, further research is needed to confirm this. Future studies should establish best practices for prompt engineering [36,37,39,41,42,45,47,51,53,59,61,65]. This involves systematically comparing different prompting strategies. For instance, moving beyond basic zero-shot prompts to incorporate advanced techniques such as few-shot prompting (providing models with examples of desired simplified outputs) [37,65] or chain-of-thought prompting (guiding the model through a step-by-step reasoning process). Chain-of-thought prompting could instruct the LLM to first identify complex medical terms, then suggest layperson equivalents, then shorten sentences, and finally check for factual accuracy [90]. Similarly, exploring the impact of stylistic constraints, detailed context-specific instructions, and the use of word-substitution lists is essential [37]. Further, future studies should focus on iterative prompt design with feedback loops for continuous refinement based on expert and patient evaluations, as well as evaluating prompt stability across LLM generations [47,61]. The reliability of prompts should also be assessed through test-retest consistency, ensuring that identical prompts yield consistent, high-quality outputs over time [91]. Finally, a key long-term goal should be the development of a validated, open-access prompt library with demonstrated efficacy for medical text simplification, enabling standardization across research and clinical applications. While existing technical prompt collections in the medical domain primarily address reasoning and question-answering tasks (eg, Microsoft MedPrompt), they were neither designed nor clinically validated for medical text simplification [92].

Technological Advancements

The rapid progress in LLM technology requires continuous evaluation. Future studies should systematically compare emerging LLMs, particularly beyond the GPT series, to identify models with unique strengths (eg, long-context handling with Claude 2) [44,47,48,59,65]. Furthermore, cost-effectiveness comparisons between commercial and open-source LLMs are also needed [47,55,63].

Geographic and Linguistic Diversity

The included studies predominantly focused on English-language PEMs within the US health care context, with only 3 studies [37,40,47] evaluating Spanish texts. Consequently, the generalizability of these findings to other languages and cultural health care systems is limited. Languages with complex morphology (eg, German or Slavic languages)

XSL•FO
RenderX

or distinct syntactic structures may present different challenges for automated simplification than English. Furthermore, cultural differences in medical communication styles may affect how "layperson-friendly" is defined and achieved in other regions. Therefore, future studies should investigate simplification performance across diverse linguistic and cultural contexts [48,49,62,65,66].

Medical Field Scope

The included studies predominantly focused on surgical specialties and ophthalmology, while numerous medical fields remain unexplored with regard to LLM-based text simplification of PEMs. No studies addressed, for example, dentistry, dermatology, psychiatry, or preventive medicine. This is notable, as these fields frequently rely on PEMs containing complex terminology. Dentistry, for example, involves jargon-heavy procedures in areas such as orthodontics that may be difficult for patients to understand [93]. Future research should investigate LLM-based text simplification across a broader range of medical fields, especially rare diseases and less prevalent health conditions, where LLMs with limited training data may be more susceptible to errors [48,65].

Linguistic Correctness

None of the included studies assessed text quality indicators related to linguistic correctness. Thus, future evaluations should incorporate metrics for grammar and typographical errors, as these factors are fundamental to establishing credibility and trust, independent of their direct effect on comprehension [94,95].

Practice Implications

Based on the findings of this scoping review, this study has several recommendations for clinical practice. First, LLMs should be used as assistive rather than autonomous tools for simplifying PEMs used in clinical practice [51]. This implication is supported by a recent study showing that LLMs can approximate targeted RGLs when simplifying health information materials about psychiatry, but their outputs are inconsistent, with significant variability in reading levels and deviations from the intended content, making them unsuitable for standalone deployment in health care settings [96]. Second, all AI-simplified PEMs must undergo mandatory expert review by qualified health care professionals to verify content fidelity and clinical appropriateness before dissemination to laypeople [38-41,45-49,52,59-61,66]. Third, health care organizations using AI-assisted text simplification tools should establish clear protocols defining quality checkpoints and approval workflows. Continuous monitoring of performance can track both readability improvements and error rates across different medical domains [59].

Although LLMs hold considerable promise for enhancing patient health literacy through simplified text versions, their use in clinical practice requires careful attention to inherent limitations, including the risk of reinforcing biases and stereotypes embedded in training data. The use of AI in health care communication requires an ongoing commitment to the accuracy of generated content [50].

Limitations

Although this scoping review followed rigorous and systematic methods, some limitations must be acknowledged. First, rule-based technologies were excluded from this review. This decision was made to focus on the contemporary and rapidly evolving landscape dominated by generative AI, which offers greater scalability and adaptability for text simplification. Second, although different prompting strategies were used across the included studies, the quality of these prompts was not evaluated. In accordance with scoping review methodology, which aims to map the extent and nature of available evidence rather than critically appraise methodological quality, a systematic assessment of prompt design was not conducted. Consequently, it was not possible to determine whether the observed failure to meet specific target RGLs reflects model limitations or suboptimal prompting strategies. However, such an analysis could have provided deeper insights into optimal prompt engineering for PEM simplification. Lastly, the database search was conducted in May 2025, and the gray literature search in July 2025. Given the rapid pace of AI development, particularly regarding LLMs, new relevant studies may have emerged since, and some findings may already become outdated shortly after publication. Finally, the included studies were heavily skewed toward English-language PEMs and US-based contexts, limiting the generalizability of the findings, as text simplification approaches may perform differently across languages with varying grammatical complexity (eg, morphology, syntax). Moreover, LLMs are predominantly trained on English data [97], which may lead to lower performance in other languages. Additionally, the readability metrics used (eg, US RGLs) are specific to the US educational system and may not directly translate to plain-language guidelines in other cultural health care systems.

Conclusions

To our knowledge, this is the first scoping review to comprehensively synthesize evidence on automated language processing technologies for PEM simplification, systematically mapping linguistic, quality, and content fidelity outcomes.

The findings of our scoping review indicate that although LLMs can improve the readability of PEMs, they predominantly fail to achieve the recommended sixth-grade RGL. However, the most significant finding is the identification of a critical validation gap: no study has assessed whether laypeople actually understood the simplified PEMs, nor has any study evaluated linguistic correctness. Coupled with variable content accuracy and reliance solely on readability formulas that poorly predict actual understanding by laypeople, these findings have important implications for clinical practice. Currently, LLMs should serve as assistive tools rather than autonomous solutions. All AI-simplified materials must undergo mandatory expert review to verify content fidelity before dissemination to laypeople. Future research must urgently shift from purely algorithmic evaluation to patient-centered validation to directly assess laypeople's comprehension.

XSL•FO
RenderX

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the Generative AI Disclosure and Transparency (GAIDeT) taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing, translation, and reformatting. The GAI tool used was DeepL Write. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This study was funded by the Austrian Science Fund (FWF; grant FG 11-B), as part of the project "Human-Centered Interactive Adaptive Visual Approaches in High-Quality Health Information." The funder had no involvement in the study design, data collection, analysis, interpretation, or writing of the manuscript.

Data Availability

All extracted data relevant to the research aims are included in the manuscript and its multimedia appendices.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[DOCX File , 84 KB - jmir_v28i1e88365_app1.docx]

Multimedia Appendix 2
PRISMA-S (PRISMA Statement for Reporting Literature Searches in Systematic Reviews) checklist.
[DOCX File , 17 KB - jmir_v28i1e88365_app2.docx]

Multimedia Appendix 3
Search strategies for bibliographic databases.
[DOCX File , 21 KB - jmir_v28i1e88365_app3.docx]

Multimedia Appendix 4
Excluded studies.
[DOCX File , 38 KB - jmir_v28i1e88365_app4.docx]

Multimedia Appendix 5
Trend analysis of publication dates for studies evaluating large language models for the text simplification of patient education materials (2019-2025).
[PNG File , 84 KB - jmir_v28i1e88365_app5.png]

Multimedia Appendix 6
Overview of prompts used for automated text simplification in the included studies.
[DOCX File , 62 KB - jmir_v28i1e88365_app6.docx]

Multimedia Appendix 7
Overview of reported outcomes in included studies.
[DOCX File , 66 KB - jmir_v28i1e88365_app7.docx]

References

1. Niederdeppe J, Boyd AD, King AJ, Rimal RN. Strategies for effective public health communication in a complex information environment. Annu Rev Public Health 2025 Apr;46(1):411-431 [FREE Full text] [doi: 10.1146/annurev-publhealth-071723-120721] [Medline: 39656948]
2. Berkman ND, Davis TC, McCormack L. Health literacy: what is it? J Health Commun 2010;15 Suppl 2:9-19. [doi: 10.1080/10810730.2010.499985] [Medline: 20845189]

XSL•FO
RenderX

3. Sudirman. The role of health literacy in improving health outcomes: challenges, interventions, and policies. J Health Lit Qual Res 2022 Mar 31;2(1):15-30. [doi: 10.61194/jhlqr.v2i1.530]
4. Wahrenbrock T, Landry K, Amin DP, Rizvanolli L, Chandrasekaran R, Mycyk MB, et al. Medical jargon is often misunderstood by emergency department patients. Am J Emerg Med 2025 Oct;96:25-29. [doi: 10.1016/j.ajem.2025.06.012] [Medline: 40513550]
5. Kumar PM, Nagate RR, Alqahtani SM, Sarma GSN, Supraja S. Role of jargon in the patient-doctor communication in the dental healthcare sector-a systematic review and meta-analysis. J Educ Health Promot 2023;12:198 [FREE Full text] [doi: 10.4103/jehp.jehp_1442_22] [Medline: 37545998]
6. Delaney FT, Doinn T, Broderick JM, Stanley E. Readability of patient education materials related to radiation safety: what are the implications for patient-centred radiology care? Insights Imaging 2021 Oct 21;12(1):148 [FREE Full text] [doi: 10.1186/s13244-021-01094-3] [Medline: 34674063]
7. Bhattad PB, Pacifico L. Empowering patients: promoting patient education and health literacy. Cureus 2022 Jul;14(7):e27336 [FREE Full text] [doi: 10.7759/cureus.27336] [Medline: 36043002]
8. Deb B. Personal health literacy. Agency for Healthcare Research and Quality (AHRQ). 2023 Aug 30. URL: https://psnet.ahrq.gov/primer/personal-health-literacy [accessed 2025-11-18]
9. Weis B. Health Literacy: A Manual for Clinicians. Chicago, IL: American Medical Association Foundation and American Medical Association; 2003.
10. Ghanem D, Covarrubias O, Harris AB, Shafiq B. Readability of the Orthopaedic Trauma Association patient education tool. J Orthop Trauma 2023 Aug 01;37(8):e307-e311. [doi: 10.1097/BOT.0000000000002593] [Medline: 36862992]
11. Cheng BT, Kim AB, Tanna AP. Readability of online patient education materials for glaucoma. J Glaucoma 2022 Jun 01;31(6):438-442. [doi: 10.1097/IJG.0000000000002012] [Medline: 35283441]
12. Kamath D, Mcintyre S, Peskin E, Stratman S, Agarwal N, Kamath PD, et al. Readability of online patient education materials for interventional pain procedures. Cureus 2020 Sep 27;12(9):e10684 [FREE Full text] [doi: 10.7759/cureus.10684] [Medline: 33133850]
13. Armache M, Assi S, Wu R, Jain A, Lu J, Gordon L, et al. Readability of patient education materials in head and neck cancer: a systematic review. JAMA Otolaryngol Head Neck Surg 2024 Aug 01;150(8):713-724. [doi: 10.1001/jamaoto.2024.1569] [Medline: 38900443]
14. Khan S, Moon J, Martin CA, Bowden E, Chen J, Tsui E, et al. Readability and suitability of online uveitis patient education materials. Ocul Immunol Inflamm 2024 Sep;32(7):1175-1179. [doi: 10.1080/09273948.2023.2203759] [Medline: 37145033]
15. Hartnett DA, Philips AP, Daniels AH, Blankenhorn BD. Readability of online foot and ankle surgery patient education materials. Foot Ankle Spec 2025 Feb;18(1):9-18. [doi: 10.1177/19386400221116463] [Medline: 35934974]
16. Okuhara T, Furukawa E, Okada H, Yokota R, Kiuchi T. Readability of written information for patients across 30 years: a systematic review of systematic reviews. Patient Educ Couns 2025 Jun;135:108656 [FREE Full text] [doi: 10.1016/j.pec.2025.108656] [Medline: 40068244]
17. Caverly TJ, Hayward RA. Dealing with the lack of time for detailed shared decision-making in primary care: everyday shared decision-making. J Gen Intern Med 2020 Oct;35(10):3045-3049 [FREE Full text] [doi: 10.1007/s11606-020-06043-2] [Medline: 32779137]
18. Hemsley B, Balandin S, Worrall L. Nursing the patient with complex communication needs: time as a barrier and a facilitator to successful communication in hospital. J Adv Nurs 2012 Jan;68(1):116-126. [doi: 10.1111/j.1365-2648.2011.05722.x] [Medline: 21831131]
19. Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, et al. A comprehensive overview of large language models. ACM Trans Intell Syst Technol 2025 Aug 18;16(5):1-72. [doi: 10.48550/arXiv.2307.06435]
20. Busch F, Hoffmann L, Rueger C, van Dijk EH, Kader R, Ortiz-Prado E, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med (Lond) 2025 Jan 21;5(1):26 [FREE Full text] [doi: 10.1038/s43856-024-00717-2] [Medline: 39838160]
21. Miao J, Thongprayoon C, Suppadungsuk S, Garcia Valencia OA, Qureshi F, Cheungpasitporn W. Innovating personalized nephrology care: exploring the potential utilization of ChatGPT. J Pers Med 2023 Dec 04;13(12):1-21 [FREE Full text] [doi: 10.3390/jpm13121681] [Medline: 38138908]
22. Eid K, Eid A, Wang D, Raiker RS, Chen S, Nguyen J. Optimizing ophthalmology patient education via chatbot-generated materials: readability analysis of AI-generated patient education materials and The American Society of Ophthalmic Plastic and Reconstructive Surgery patient brochures. Ophthalmic Plast Reconstr Surg 2024;40(2):212-216. [doi: 10.1097/IOP.0000000000002549] [Medline: 37972974]
23. Alessandri-Bonetti M, Liu HY, Palmesano M, Nguyen VT, Egro FM. Online patient education in body contouring: a comparison between Google and ChatGPT. J Plast Reconstr Aesthet Surg 2023 Dec;87:390-402. [doi: 10.1016/j.bjps.2023.10.091] [Medline: 37939643]
24. Espinosa-Zaragoza I, Abreu-Salas J, Lloret E, Pozo P, Palomar M. A review of research-based automatic text simplification tools. 2023 Sep Presented at: The 14th International Conference on Recent Advances in Natural Language Processing; September 4-6, 2023; Shoumen, Bulgaria. [doi: 10.26615/978-954-452-092-2_036]

XSL•FO
RenderX

25. Zhou H, Liu F, Gu B, Zou X, Huang J, Wu J. A survey of large language models in medicine: progress, application, and challenge. arXiv Preprint posted online November 9, 2023 [FREE Full text] [doi: 10.48550/arXiv.2311.05112]
26. Kirk MD, Pires SM, Black RE, Caipo M, Crump JA, Devleesschauwer B, et al. Correction: World Health Organization estimates of the global and regional disease burden of 22 foodborne bacterial, protozoal, and viral diseases, 2010: a data synthesis. PLoS Med 2015 Dec;12(12):e1001940 [FREE Full text] [doi: 10.1371/journal.pmed.1001940] [Medline: 26701262]
27. Aromataris EL, Porritt K, Pilla B, Jordan Z. JBI Manual for Evidence Synthesis - 2024 Edition. JBI. URL: https://synthesismanual.jbi.global [accessed 2025-09-08]
28. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, Tunçalp, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473. [doi: 10.7326/m18-0850]
29. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
30. Krenn C, Wilfling D, Loder C, Jeitler K, Siebenhofer-Kroitzsch A, Semlitsch T. Automated approaches to simplifying complex medical texts for laypersons: a scoping review protocol. Open Science Framework (OSF). URL: https://osf.io/wrb38/ [accessed 2025-11-18]
31. Schreck T. Nutzerzentrierte adaptive ansätze zur gesundheitsinformation (human-centered adaptive approaches for health information). Austrian Science Fund (FWF). Vienna URL: https://www.fwf.ac.at/forschungsradar/10.55776/FG11 [accessed 2025-10-14]
32. Koster J. PubReMiner. PubMed. 2014. URL: https://hgserver2.amc.nl/cgi-bin/miner/miner2.cgi [accessed 2025-05-14]
33. Better synonyms. TopAI. URL: https://topai.tools/t/better-synonyms [accessed 2025-05-14]
34. Zhang C, Liu S, Zhou X, Zhou S, Tian Y, Wang S, et al. Examining the role of large language models in orthopedics: systematic review. J Med Internet Res 2024 Nov 15;26:e59607 [FREE Full text] [doi: 10.2196/59607] [Medline: 39546795]
35. Waldock WJ, Zhang J, Guni A, Nabeel A, Darzi A, Ashrafian H. The accuracy and capability of artificial intelligence solutions in health care examinations and certificates: systematic review and meta-analysis. J Med Internet Res 2024 Nov 05;26:e56532 [FREE Full text] [doi: 10.2196/56532] [Medline: 39499913]
36. Spina A, Fereydouni P, Tang J, Andalib S, Picton B, Fox A. Tailoring glaucoma education using large language models: addressing health disparities in patient comprehension. Medicine (Baltimore) 2025 Jan 10;104(2):e41059 [FREE Full text] [doi: 10.1097/MD.0000000000041059] [Medline: 39792725]
37. Reaver C, Pereira D, Carrillo E, Marcos C, Goldfarb C. Evaluating the performance of artificial intelligence for improving readability of online English- and Spanish-language orthopaedic patient educational material: challenges in bridging the digital divide. The Journal of bone and joint surgery 2025;107(8):e36. [doi: 10.2106/jbjs.24.01078]
38. Picton B, Andalib S, Spina A, Camp B, Solomon SS, Liang J, et al. Assessing AI simplification of medical texts: readability and content fidelity. Int J Med Inform 2025 Mar;195:105743 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105743] [Medline: 39667051]
39. Chandra K, Ghilzai U, Lawand J, Ghali A, Fiedler B, Ahmed AS. Improving readability of shoulder and elbow surgery online patient education material with Chat GPT (Chat Generative Pretrained Transformer) 4. J Shoulder Elbow Surg 2025 Nov;34(11):e1119-e1124. [doi: 10.1016/j.jse.2025.02.025] [Medline: 40118438]
40. Busigó Torres R, Restrepo M, Stern BZ, Yahuaca BI, Buerba RA, García IA, et al. Artificial intelligence shows limited success in improving readability levels of Spanish-language orthopaedic patient education materials. Clin Orthop Relat Res 2025 Feb 11;483(7):1185-1192. [doi: 10.1097/CORR.0000000000003413] [Medline: 39937452]
41. Zaki HA, Mai M, Abdel-Megid H, Liew SQR, Kidanemariam S, Omar AS, et al. Using ChatGPT to improve readability of interventional radiology procedure descriptions. Cardiovasc Intervent Radiol 2024 Aug;47(8):1134-1141. [doi: 10.1007/s00270-024-03803-z] [Medline: 38981939]
42. Shehab AA, Shedd KE, Alamah W, Mardini S, Bite U, Gibreel W. Bridging gaps in health literacy for cleft lip and palate: the role of artificial intelligence and interactive educational materials. Cleft Palate Craniofac J 2026 Jan;63(1):65-71. [doi: 10.1177/10556656241289653] [Medline: 39380385]
43. Patel EA, Fleischer L, Filip P, Eggerstedt M, Hutz M, Michaelides E, et al. The use of artificial intelligence to improve readability of otolaryngology patient education materials. Otolaryngol Head Neck Surg 2024 Aug;171(2):603-608. [doi: 10.1002/ohn.816] [Medline: 38751109]
44. Li G, Lin MX, Cui D, Mathews PM, Akpek EK. Enhancing online cataract surgery patient education materials through artificial intelligence. Can J Ophthalmol 2025 Apr;60(2):e197-e202. [doi: 10.1016/j.jcjo.2024.07.018] [Medline: 39163992]
45. Vallurupalli M, Shah ND, Vyas RM. Optimizing readability of patient-facing hand surgery education materials using Chat Generative Pretrained Transformer (ChatGPT) 3.5. J Hand Surg Am 2024 Oct;49(10):986-991. [doi: 10.1016/j.jhsa.2024.05.007] [Medline: 38970600]
46. Oliva AD, Pasick LJ, Hoffer ME, Rosow DE. Improving readability and comprehension levels of otolaryngology patient education materials using ChatGPT. Am J Otolaryngol 2024;45(6):104502. [doi: 10.1016/j.amjoto.2024.104502] [Medline: 39197330]

XSL•FO
RenderX

47. Sudharshan R, Shen A, Gupta S, Zhang-Nunes S. Assessing the utility of ChatGPT in simplifying text complexity of patient educational materials. Cureus 2024 Mar;16(3):e55304 [FREE Full text] [doi: 10.7759/cureus.55304] [Medline: 38559518]
48. Ayre J, Mac O, McCaffery K, McKay BR, Liu M, Shi Y, et al. New frontiers in health literacy: using ChatGPT to simplify health information for people in the community. J Gen Intern Med 2024 Mar 08;39(4):573-577 [FREE Full text] [doi: 10.1007/s11606-023-08469-w] [Medline: 37940756]
49. Kirchner GJ, Kim RY, Weddle JB, Bible JE. Can artificial intelligence improve the readability of patient education materials? Clin Orthop Relat Res 2023 Nov 01;481(11):2260-2267. [doi: 10.1097/CORR.0000000000002668] [Medline: 37116006]
50. Baldwin AJ. An artificial intelligence language model improves readability of burns first aid information. Burns 2024 Jun;50(5):1122-1127. [doi: 10.1016/j.burns.2024.03.005] [Medline: 38492982]
51. Dihan QA, Brown AD, Chauhan MZ, Alzein AF, Abdelnaem SE, Kelso SD, et al. Leveraging large language models to improve patient education on dry eye disease. Eye (Lond) 2025 Apr;39(6):1115-1122. [doi: 10.1038/s41433-024-03476-5] [Medline: 39681711]
52. Gupta M, Gupta P, Ho C, Wood J, Guleria S, Virostko J. Can generative AI improve the readability of patient education materials at a radiology practice? Clin Radiol 2024 Nov;79(11):e1366-e1371. [doi: 10.1016/j.crad.2024.08.019] [Medline: 39266371]
53. Kianian R, Sun D, Crowell EL, Tsui E. The use of large language models to generate education materials about uveitis. Ophthalmol Retina 2024 Feb;8(2):195-201 [FREE Full text] [doi: 10.1016/j.oret.2023.09.008] [Medline: 37716431]
54. Dihan QA, Brown AD, Zaldivar AT, Chauhan MZ, Eleiwa TK, Hassan AK, et al. Advancing patient education in idiopathic intracranial hypertension: the promise of large language models. Neurol Clin Pract 2025 Feb;15(1):e200366. [doi: 10.1212/CPJ.0000000000200366] [Medline: 39399571]
55. Dihan Q, Chauhan MZ, Eleiwa TK, Hassan AK, Sallam AB, Khouri AS, et al. Using large language models to generate educational materials on childhood glaucoma. Am J Ophthalmol 2024 Sep;265:28-38. [doi: 10.1016/j.ajo.2024.04.004] [Medline: 38614196]
56. Dihan Q, Chauhan MZ, Eleiwa TK, Brown AD, Hassan AK, Khodeiry MM, et al. Large language models: a new frontier in paediatric cataract patient education. Br J Ophthalmol 2024 Sep 20;108(10):1470-1476. [doi: 10.1136/bjo-2024-325252] [Medline: 39174290]
57. Andalib S, Solomon S, Picton B, Spina A, Scolaro J, Nelson A. Source characteristics influence AI-enabled orthopaedic text simplification: recommendations for the future. JB JS Open Access 2025;10(1):1-14 [FREE Full text] [doi: 10.2106/JBJS.OA.24.00007] [Medline: 39781102]
58. Garcia Valencia OA, Thongprayoon C, Miao J, Suppadungsuk S, Krisanapan P, Craici IM, et al. Empowering inclusivity: improving readability of living kidney donation information with ChatGPT. Front Digit Health 2024 Apr 10;6:1366967 [FREE Full text] [doi: 10.3389/fdgth.2024.1366967] [Medline: 38659656]
59. Fanning JE, Escobar-Domingo MJ, Foppiani J, Lee D, Miller AS, Janis JE, et al. Improving readability and automating content analysis of plastic surgery webpages with ChatGPT. J Surg Res 2024 Jul;299:103-111. [doi: 10.1016/j.jss.2024.04.006] [Medline: 38749313]
60. Manasyan A, Lasky S, Jolibois M, Moshal T, Roohani I, Munabi N, et al. Expanding accessibility in cleft care: the role of artificial intelligence in improving literacy of alveolar bone grafting information. Cleft Palate Craniofac J 2025 Nov;62(11):1873-1880. [doi: 10.1177/10556656241281453] [Medline: 39246230]
61. Vallurupalli M, Shah ND, Vyas RM. Validation of ChatGPT 3.5 as a tool to optimize readability of patient-facing craniofacial education materials. Plast Reconstr Surg Glob Open 2024 Feb;12(2):e5575 [FREE Full text] [doi: 10.1097/GOX.0000000000005575] [Medline: 38313589]
62. Abreu AA, Murimwa GZ, Farah E, Stewart JW, Zhang L, Rodriguez J, et al. Enhancing readability of online patient-facing content: the role of AI chatbots in improving cancer information accessibility. J Natl Compr Canc Netw 2024 May 15;22(2 D):e237334. [doi: 10.6004/jnccn.2023.7334] [Medline: 38749478]
63. Rouhi AD, Ghanem YK, Yolchieva L, Saleh Z, Joshi H, Moccia MC, et al. Can artificial intelligence improve the readability of patient education materials on aortic stenosis? A pilot study. Cardiol Ther 2024 Mar;13(1):137-147 [FREE Full text] [doi: 10.1007/s40119-023-00347-0] [Medline: 38194058]
64. Singh A, Gupta N, Chien DL, Singh R, Sachdeva A, Danasekaran K, et al. ChatGPT-4 in neurosurgery: improving patient education materials. Neurosurgery 2026 Jan 01;98(1):147-160. [doi: 10.1227/neu.0000000000003606] [Medline: 40704789]
65. Will J, Gupta M, Zaretsky J, Dowlath A, Testa P, Feldman J. Enhancing the readability of online patient education materials using large language models: cross-sectional study. J Med Internet Res 2025 Jun 04;27:e69955 [FREE Full text] [doi: 10.2196/69955] [Medline: 40465378]
66. Naghdi M, Cao P, Essers R, Heijligers M, Paulussen ADC, van der Lugt A, et al. Artificial intelligence-simplified information to advance reproductive genetic literacy and health equity. Hum Reprod 2025 Sep 01;40(9):1681-1688. [doi: 10.1093/humrep/deaf135] [Medline: 40692125]
67. Scott B. Readability scoring system plus. Readability Formulas. URL: https://readabilityformulas.com/readability-scoring-system.php [accessed 2026-02-03]

XSL•FO
RenderX

68. Hansberry DR, Agarwal N, Gonzales SF, Baker SR. Are we effectively informing patients? A quantitative analysis of on-line patient education resources from the American Society of Neuroradiology. AJNR Am J Neuroradiol 2014 Jul;35(7):1270-1275 [FREE Full text] [doi: 10.3174/ajnr.A3854] [Medline: 24763420]
69. Nasra M, Jaffri R, Pavlin-Premrl D, Kok HK, Khabaza A, Barras C, et al. Can artificial intelligence improve patient educational material readability? A systematic review and narrative synthesis. Intern Med J 2025 Jan 25;55(1):20-34. [doi: 10.1111/imj.16607] [Medline: 39720869]
70. Zhao F, He H, Liang J, Cen J, Wang Y, Lin H, et al. Benchmarking the performance of large language models in uveitis: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, Google Gemini, and Anthropic Claude3. Eye (Lond) 2025 Apr;39(6):1132-1137. [doi: 10.1038/s41433-024-03545-9] [Medline: 39690303]
71. Zhang Q, Wu Z, Song J, Luo S, Chai Z. Comprehensiveness of large language models in patient queries on gingival and endodontic health. Int Dent J 2025 Feb;75(1):151-157 [FREE Full text] [doi: 10.1016/j.identj.2024.06.022] [Medline: 39147663]
72. Zhan Y, Chen X, Ye F, Wu Z, Usman M, Yuan Z, et al. Evaluating AI chatbot responses to postkidney transplant inquiries. Transplant Proc 2025 Mar;57(2):394-405. [doi: 10.1016/j.transproceed.2024.12.028] [Medline: 39814625]
73. Zeljkovic I, Novak M, Jordan A, Lisicic A, Nemeth-Blažić T, Pavlovic N, et al. Evaluating ChatGPT-4's correctness in patient-focused informing and awareness for atrial fibrillation. Heart Rhythm O2 2025 Jan;6(1):58-63 [FREE Full text] [doi: 10.1016/j.hroo.2024.10.005] [Medline: 40224268]
74. Yoo M, Jang CW. Presentation suitability and readability of ChatGPT's medical responses to patient questions about on knee osteoarthritis. Health Informatics J 2025 Jan 19;31(1):14604582251315587 [FREE Full text] [doi: 10.1177/14604582251315587] [Medline: 39828887]
75. Yıldız HA, Söğütdelen E. AI chatbots as sources of STD information: a study on reliability and readability. J Med Syst 2025 Apr 03;49(1):43. [doi: 10.1007/s10916-025-02178-z] [Medline: 40178771]
76. Yılmaz IBE, Doğan L. Talking technology: exploring chatbots as a tool for cataract patient education. Clin Exp Optom 2025 Jan 09;108(1):56-64. [doi: 10.1080/08164622.2023.2298812] [Medline: 38194585]
77. Xu Q, Wang J, Chen X, Wang J, Li H, Wang Z, et al. Assessing the efficacy of ChatGPT prompting strategies in enhancing thyroid cancer patient education: a prospective study. J Med Syst 2025 Jan 17;49(1):11. [doi: 10.1007/s10916-024-02129-0] [Medline: 39820814]
78. Schwartzman JD, Shaath MK, Kerr MS, Green CC, Haidukewych GJ. ChatGPT is an unreliable source of peer-reviewed information for common total knee and hip arthroplasty patient questions. Adv Orthop 2025 Jan 06;2025(1):5534704 [FREE Full text] [doi: 10.1155/aort/5534704] [Medline: 39817149]
79. Yang X, Xiao Y, Liu D, Shi H, Deng H, Huang J, et al. Enhancing physician-patient communication in oncology using GPT-4 through simplified radiology reports: multicenter quantitative study. J Med Internet Res 2025 Apr 17;27:e63786 [FREE Full text] [doi: 10.2196/63786] [Medline: 40245397]
80. Eisinger F, Holderried F, Mahling M, Stegemann-Philipps C, Herrmann-Werner A, Nazarenus E, et al. What's going on with me and how can i better manage my health? The potential of GPT-4 to transform discharge letters into patient-centered letters to enhance patient safety: prospective, exploratory study. J Med Internet Res 2025 Jan 21;27:e67143 [FREE Full text] [doi: 10.2196/67143] [Medline: 39836954]
81. Can E, Uller W, Vogt K, Doppler MC, Busch F, Bayerl N, et al. Large language models for simplified interventional radiology reports: a comparative analysis. Acad Radiol 2025 Feb;32(2):888-898 [FREE Full text] [doi: 10.1016/j.acra.2024.09.041] [Medline: 39353826]
82. Li HH, Moon JT, Kumar S, Ricci J, Sim N, Bercu ZL, et al. Evaluation of multilingual simplifications of IR procedural reports using GPT-4. J Vasc Interv Radiol 2025 Apr;36(4):696-703.e1. [doi: 10.1016/j.jvir.2025.01.002] [Medline: 39793700]
83. Nguyen AT, Li RA, Gosain AK, Galiano RD. Readability of online patient education materials for cleft care: a systematic review and meta-analysis. Cleft Palate Craniofac J 2026 May;63(5):1223-1232. [doi: 10.1177/10556656251327803] [Medline: 40095963]
84. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne) 2024;11:1477898 [FREE Full text] [doi: 10.3389/fmed.2024.1477898] [Medline: 39534227]
85. Zheng J, Yu H. Readability formulas and user perceptions of electronic health records difficulty: a corpus study. J Med Internet Res 2017 Mar 02;19(3):e59 [FREE Full text] [doi: 10.2196/jmir.6962] [Medline: 28254738]
86. Kauchak D, Leroy G. Moving beyond readability metrics for health-related text simplification. IT Prof 2016;18(3):45-51 [FREE Full text] [doi: 10.1109/MITP.2016.50] [Medline: 27698611]
87. Leroy G, Kauchak D. The effect of word familiarity on actual and perceived text difficulty. J Am Med Inform Assoc 2014 Feb;21(e1):e169-e172 [FREE Full text] [doi: 10.1136/amiajnl-2013-002172] [Medline: 24100710]
88. Kandula S, Curtis D, Zeng-Treitler Q. A semantic and syntactic text simplification tool for health content. AMIA Annu Symp Proc 2010 Nov 13;2010:366-370 [FREE Full text] [Medline: 21347002]
89. Ondov B, Attal K, Demner-Fushman D. A survey of automated methods for biomedical text simplification. J Am Med Inform Assoc 2022 Oct 07;29(11):1976-1988 [FREE Full text] [doi: 10.1093/jamia/ocac149] [Medline: 36083212]

XSL•FO
RenderX

90. Zhu D, Xiong Y, Zhang J, Xie X, Xia C. Understanding before reasoning: enhancing chain-of-thought with iterative summarization pre-prompting. arXiv Preprint posted online January 8, 2025 [FREE Full text] [doi: 10.48550/arXiv.2501.04341]
91. Liu J, Liu F, Wang C, Liu S. Prompt engineering in clinical practice: tutorial for clinicians. J Med Internet Res 2025 Sep 15;27:e72644-e72644 [FREE Full text] [doi: 10.2196/72644] [Medline: 40955776]
92. Horvitz E. The power of prompting Microsoft. Microsoft Corporation. URL: https://www.microsoft.com/en-us/research/blog/the-power-of-prompting/ [accessed 2026-02-27]
93. Benning A, Madadian MA, Seehra J, Fan K. Patients understanding of terminology commonly used during combined orthodontic-orthognathic treatment. Surgeon 2021 Oct;19(5):e193-e198. [doi: 10.1016/j.surge.2020.09.012] [Medline: 33423926]
94. Liu J, Song S, Zhang Y. Linguistic features and consumer credibility judgment of online health information. ResearchGate. 2021. URL: https://www.researchgate.net/publication/350511487_Linguistic_features_and_consumer_credibility_judgment_of_online_health_information [accessed 2026-04-24]
95. Witchel HJ, Thompson GA, Jones CI, Westling CEI, Romero J, Nicotra A, et al. Spelling errors and shouting capitalization lead to additive penalties to trustworthiness of online health information: randomized experiment with laypersons. J Med Internet Res 2020 Jun 10;22(6):e15171 [FREE Full text] [doi: 10.2196/15171] [Medline: 32519676]
96. Aich A, Liu T, Giorgi S, Isman K, Bobojonova R, Ungar L, et al. Language models in digital psychiatry: challenges with simplification of healthcare materials. NPP Digit Psychiatry Neurosci 2025 May 22;3(1):10. [doi: 10.1038/s44277-025-00029-w] [Medline: 41361013]
97. Qin L, Chen Q, Zhou Y, Chen Z, Li Y, Liao L, et al. A survey of multilingual large language models. Patterns (N Y) 2025 Jan 10;6(1):101118 [FREE Full text] [doi: 10.1016/j.patter.2024.101118] [Medline: 39896256]

Abbreviations

AI: artificial intelligence
AMA: American Medical Association
ARI: Automated Readability Index
BLEU: Bilingual Evaluation Understudy
CLI: Coleman-Liau Index
FKGL: Flesch-Kincaid Grade Level
FRE: Flesch Reading Ease
GFI: Gunning Fog Index
INFLESZ: Índice Flesch-Szigriszt
JBI: Joanna Briggs Institute
LIX: Läsbarhetsindex Index
LLM: large language model
LWF: Linsear Write Formula
NLP: natural language processing
PCC: Population-Concept-Context
PEM: patient education material
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: PRISMA Statement for Reporting Literature Searches in Systematic Reviews
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RGL: reading grade level
RIX: Rate Index
ROUGE: Recall-Oriented Understudy for Gisting Evaluation
SMOG: Simple Measure of Gobbledygook
SOL: Spanish Orthographic Length

XSL•FO
RenderX

Edited by S Brini; submitted 24.Nov.2025; peer-reviewed by K Chen, Z Liu; comments to author 22.Dec.2025; revised version received 27.Feb.2026; accepted 02.Mar.2026; published 07.May.2026.

Please cite as:
Krenn C, Loder C, Berger N, Jeitler K, Semlitsch T, Siebenhofer A, Wilfling D
Automated Approaches of Text Simplification of Patient Education Materials: Scoping Review
J Med Internet Res 2026;28:e88365
URL: https://www.jmir.org/2026/1/e88365
doi:10.2196/88365
PMID:

©Cornelia Krenn, Christine Loder, Natalie Berger, Klaus Jeitler, Thomas Semlitsch, Andrea Siebenhofer, Denise Wilfling. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Reporting of Telehealth Implementation in Cystic Fibrosis: Scoping Review Using a Novel Theory-Based Evaluation Lens

Tamara Vagg[1,2,3], BA, MSc, PhD; Robyn Doherty[4], BA, MRes, PhD; Sarath C Ranganathan[5,6,7], MBChB, MRCP, FRCPH, FRACP, PhD, FThorSoc, ATSF, FAHMS; Barry J Plant[1,2,3], MB BCH BAO, B.Med.Sc, MRCPI, MFSEM, ECFMG, MD, PCCM, FRCPI

[1]Cork Centre for Cystic Fibrosis, Cork University Hospital, University College Cork, Cork, Ireland
[2]The HRB funded Clinical Research Facility, University College Cork, Cork, Ireland
[3]Department of Medicine, University College Cork, Cork, Ireland
[4]Cystic Fibrosis Registry of Ireland, University College Dublin, Cystic Fibrosis Registry of Ireland, Dublin, Ireland
[5]Respiratory and Sleep Medicine Department, Royal Children's Hospital, Melbourne, Australia
[6]Respiratory Diseases Research, Murdoch Children's Research Institute, Melbourne, Australia
[7]Department of Paediatrics, The University of Melbourne, Melbourne, Australia

Corresponding Author:
Tamara Vagg, BA, MSc, PhD
Cork Centre for Cystic Fibrosis
Cork University Hospital
University College Cork
Wilton
Cork, T12 DC4A
Ireland
Phone: 353 0214920018
Email: tamara.vagg@ucc.ie

Abstract

Background: Many inductive reviews exploring telehealth and its application in health care have identified missing or inconsistently reported implementation data, calling for a standardized approach to telehealth research.

Objective: Using cystic fibrosis (CF) as a case exemplar, this study evaluated the adherence of telehealth research to standardized reporting frameworks through a theory-based evaluation lens to assess implementation reporting quality and identify knowledge gaps and strengths across the literature.

Methods: We conducted an updated systematic review of the PubMed, Scopus, and Web of Science databases using a novel deductive approach to identify relevant scientific papers available in English and focusing on the delivery of telehealth interventions to CF populations as part of or alongside routine CF care. Two relevant reporting checklists were identified in the Equator Network database (Guidelines and Checklist for the Reporting on Digital Health Implementations [iCHECK-DH] and Template for Intervention Description and Replication for Telehealth [TIDiER-telehealth]) to extract data from the papers. Each checklist category was described as being “fully reported” (score=2), “partially reported” (score=1), and “did not report” (score=0) for each paper. An overall score was calculated for adherence to the checklists.

Results: In total, 98 studies published between 2006 and May 2025 were included in this review, with the majority appearing during the COVID-19 pandemic (2021-2022). Most studies were conducted in a single country, predominantly the United States, Australia, and the United Kingdom, and were published in medical journals. Telehealth was variably described, with video call–based models in combination with remote monitoring being most common. The median score was 22/40 (range 11-29, 55.0% adherent) for iCHECK and 15/24 (range 6-23, 62.5% adherent) for TIDiER, demonstrating moderate overall reporting quality. For iCHECK, ≥50% of studies fully reported 6/20 categories, partially reported 9/20 categories, and did not report 3/20 categories. For TIDiER, ≥50% of studies fully reported 4/12 categories, partially reported 6/12 categories, and did not report 1/12 categories, indicating persistent gaps in intervention description despite improved partial reporting.

Conclusions: Key areas, such as justification for telehealth, target populations, and outcomes, are well documented, offering valuable insights into the rationale for and outcomes associated with telehealth. However, implementation processes remain underreported, partly due to the more recent adoption of frameworks like iCHECK and TIDiER. The clinical implications of the

XSL•FO
RenderX

current evidence limit the implementation of telehealth in terms of the ability to assess feasibility and readiness for adoption; understand financial implications and plan sustainably; ensure patient safety, data protection, and equity; interpret outcome data in context; and share, replicate, or scale evidence-based models of care. Strengthening the commitment to standardized telehealth reporting will ultimately support clinical decision-making and improve the effective and equitable integration of telehealth into care.

(J Med Internet Res 2026;28:e86194) doi:10.2196/86194

KEYWORDS

Guidelines and Checklist for the Reporting on Digital Health Implementations; iCHECK-DH; Template for Intervention Description and Replication for Telehealth; TIDiER-telehealth; scoping review; telehealth; cystic fibrosis; implementation

Introduction

Background

Telehealth has existed for several decades but experienced a large increase in adoption during the COVID-19 pandemic [1]. Over time, telehealth's definition has changed to match current technology capabilities and infrastructure contexts. As a result, there is no universally accepted definition of telehealth, and the term is often used interchangeably with other terms, such as telemedicine, eHealth, and virtual monitoring [2,3]. For this review, telehealth refers to interventions/care delivered synchronously by a health professional to a patient over the phone or via video [4]. Much research in this field often aims to synthesize and produce evidence to support the adoption of telehealth in different health contexts, including chronic disease care and management across different disease/condition populations [5-7]. However, similar to the difficulty with terminology and definitions, heterogeneity exists in research approaches, telehealth strategies, data, and reporting, ultimately leading to clinical challenges with implementation and decision-making due to missing or lacking data [8]. An example of this can be seen in studies reviewing telehealth in amyotrophic lateral sclerosis (ALS) care. In 2019, Helleman et al [9] conducted a systematic review on the use of telehealth in ALS care. The authors concluded that telehealth is positively received by patients and carers; however, care teams report mixed feelings for telehealth in ALS care. This is due to implementation barriers surrounding legislation and finances and the need for more data on barriers/facilitators and cost-effectiveness in this space, demonstrating the effects of missing pertinent data [9].

In addition to the identified lack of data exploring the cost-effectiveness of telehealth, in 2022, Houser et al [10] identified a lack of data and research on telehealth practices to manage documentation of telehealth patient visits (eg, the clinical, administrative, and technical information that must be entered into a patient health record). The authors collected survey data from clinical administrators and managers of different physical and mental health facilities in the United States and found that there are inequities in patient access and knowledge of technology, telehealth guidelines are being frequently updated, and telehealth procedures are also being frequently updated, causing difficulty for both patients and care teams. More notably, there are barriers related to associated documentation for reimbursement from health insurers, leading to refusal of telehealth visits and other associated legal issues, again presenting an example of how a lack of data on documentation practices affects telehealth implementation and delivery [10].

These issues with associated telehealth documentation were also found by research conducted in 2017 to explore patient information leaflets on telehealth. Kayyali et al [11] identified information gaps in these leaflets, such as technical assurance, technical support, cost, privacy, safety, and patient choice, related to information about technology reliability, what to do if the technology failed, whom to contact with technical issues, the costs associated with a telehealth visit, and how privacy and confidentiality will be maintained. These information gaps were also believed to contribute negatively to patient experience and the adoption of telehealth services, again demonstrating the need for data to support meaningful and informed patient information leaflet creation as part of telehealth implementation [11].

As patients expressed concern with privacy and safety [11], so too did researchers and care teams. In 2022, Houser et al [12] conducted a complementary systematic review exploring data and privacy concerns in US telehealth research. The authors found that these concerns are caused by three types of key factors: environmental, technological, and operational (eg, broadband infrastructure disparities, the integration of telehealth platforms into electronic health records [EHRs], and staff training). The authors concluded that broad approaches considering multiple domains (technology, digital literacy, accessibility, privacy and safety concerns) are needed to develop best-practice guidelines for the implementation of telehealth into services, presenting an example of pertinent telehealth data that need to be included in telehealth research to support wider synthesis for the development of guidelines [12].

These disparities have continued to be noted across health care services in more recent years. In 2024, Amagai et al [13] retrospectively analyzed EHR data from a large hospital network in the United States to understand telehealth disparities across specialties. The authors noted that telehealth reduces the number of no-shows (did not attend) overall, telehealth use and effectiveness vary depending on specialty, and there are disparities among demographic groups (including race, ethnicity, health insurance type, and access to technology). Moreover, telehealth alone does not eliminate health inequities, and pre-existing structural barriers can persist in telehealth. The authors highlighted that policy makers need to tailor policies to reach different patient cohorts, address the varying specialty needs, and improve equitable telehealth adoption [13],

XSL•FO
RenderX

presenting an example of how a "one size fits all" approach to telehealth does not work and more data are needed to understand and combat health inequities.

In 2025, the aforementioned barriers and considerations were continuously noted. For example, Mahdavi et al [14] reported that successful telehealth implementation is strongly influenced by infrastructure, technical support, and training for patients. Oudbier et al [15] built upon this by demonstrating that implementation is not a singular process and relies on ongoing management and review to adapt to the changing needs of patients and care services. In addition, two systematic reviews [16,17] concluded that telehealth is a beneficial component to care delivery, although technology gaps, regulatory uncertainties, and equity issues are ongoing, and that a collaborative approach is needed to develop sustainable and equitable telehealth frameworks that detail comprehensive training programs and investment in technology. Another 2025 systematic review [1] focusing on telehealth during the COVID-19 pandemic identified that evaluations lacked adequate methods and cohort coverage. These telehealth systems, due to the necessity of the pandemic, were rapidly implemented without sufficient data on effectiveness and implementation pathways. Consequently, Mumporeze et al [1] called for future studies to provide more comprehensive data to describe these pathways. These examples demonstrate how research and telehealth inductive reviews vary in what data are found to be important; however, all studies show that data that support frameworks and pathways are essential for sustaining telehealth.

The aforementioned telehealth studies demonstrate that there are important gaps in the evidence base, especially missing or inconsistently reported data, and that certain data play a critical role in guiding long-term adoption, informing policy development, and shaping implementation strategies that support equitable access and workflow integration. To address these gaps and inconsistencies, emerging standardized reporting methods now provide a means to consistently capture essential research and implementation data, ensuring reproducibility and strengthening the quality of evidence needed to support telehealth scale-up and policy formation.

Rationale and Scoping Review Aims

From an implementation science perspective, closing the gap between what we know and what we do becomes difficult when the evidence base is insufficient. In the aforementioned systematic reviews, the authors deduced missing or important telehealth data from the identified literature in a bottom-up approach. However, there is need for a deductive and structured top-down approach that can be replicated across multiple telehealth care contexts to provide a means of identifying both the strengths and the weaknesses of telehealth reporting in a standardized way. In short, previous systematic reviews have worked from the ground up to piece together what is missing and important, and there is need for novel structured top-down approaches to support reproducibility in identifying what is being reported well and what is missing.

This scoping review sought to use such a structured approach to provide a more complete picture of the current state of evidence and identify areas for improvement in the reporting of telehealth implementation projects. To do so, we examined the extent of telehealth implementation for cystic fibrosis (CF) reported up to 2025 using a novel method that leverages standardized reporting frameworks and a theory-based evaluation (TBE) perspective. The findings of this review will offer a framework that can be applied to other chronic diseases.

To achieve the overall study aim, we:

- Conducted a systematic scoping review of available research on telehealth in CF care
- Examined the reporting of telehealth implementation in the identified studies against a standardized reporting frameworks
- Identified the knowledge gaps and strengths of the existing literature

Implementation of Telehealth in Cystic Fibrosis

CF is a specialized multisystem chronic genetic condition, and its complexity necessitates support from a specialized multidisciplinary care team spanning medicine, nursing, dietetics, physiotherapy, pharmacy, psychology, and social care. Care teams must also regularly engage with additional specialist services (eg, endocrinology, nephrology, cardiology, dermatology, gastroenterology, radiology and imaging, and mental health) to address the diverse complications that can affect multiple organs and give rise to comorbidities, such as liver disease, diabetes, and pancreatic insufficiency [18]. The availability of highly effective modulator therapies has further shifted the clinical landscape, extending life expectancy for many patients with CF and further incorporating other areas of care, such as pregnancy, aging, menopause, and malignancies [19]. As a result, CF care teams now also collaborate more frequently with broader specialties, such as fertility and maternity care, and oncology [20,21].

Given the multifaceted nature of the condition and the need for specialized multidisciplinary teams, telehealth plays an important role in CF care by enabling timely access to specialized support, improving disease management for patients living far from specialist centers, and supporting continuity of care through virtual consultations and remote monitoring [22]. Remote monitoring and telehealth have been explored in CF care since the 1980s [23], offering a long history of engagement with digital health approaches. As such, CF provides an especially relevant model for telehealth implementation.

CF has a clearly defined patient population, which mirrors many chronic conditions that rely on specialist services. This defined cohort brings its own considerations for how telehealth can facilitate disease management and support individuals who are geographically dispersed or distant from specialist centers. CF care also involves long-term follow-up across well-established clinical outcome measures, such as lung function, microbiology, and surveillance for complications [24,25]. CF also carries a substantial burden of treatment; patients with CF typically require three to four reviews annually with specialist centers and must manage a high volume of regular prescribed medications, many of which demand ongoing monitoring [26].

Finally, CF benefits from established international models of care [27] that guide practice in a consistent and uniform manner

XSL•FO
RenderX

across geographic locations, creating a stable framework within which telehealth innovations can be tested, implemented, and evaluated. In Australia, telehealth has become so integrated into standard care that it is written into the national standards of care [28].

As across other disease areas, the implementation and integration of telehealth in CF care was especially amplified during the COVID-19 pandemic, with many practice-based examples of its use during that time being published [8]. Since then, although a return to face-to-face care has prevailed [29], an ongoing interest in the potential of telehealth implementation in CF care has remained, with some surmising that future face-to-face clinics may not always be needed [30], paving the way for greater integration of telehealth and other digital health solutions into standard CF care delivery. Furthermore, two large CF networks (the Cystic Fibrosis Learning Network [CFLN] in the United States and the European CF Society Telehealth for CF Working Group [THCF]) are actively exploring safe and equitable implementation into CF care delivery, also presenting an opportunity for wider contextual understanding [4,31].

Although interest in telehealth for CF care continues, the evidence base remains limited, and the absence of implementation standards restricts its broader adoption [32]. Many promising examples exist [8], often as case studies, but systematic reviews show most research focuses on feasibility rather than outcomes [33,34]. Inconsistencies in study design, limited multicenter trials, and no standardized reporting make it difficult to establish best practices for integrating telehealth into CF care and clinical decision-making [8].

Methods

Defining a Theoretical and Reporting Framework

We followed a TBE approach for this study. TBE is a method of judging how well programs, such as telehealth, are working, focusing on how they are implemented and the results they achieve [35]. TBE is based on two key ideas [36]:

- Program theory: This explains the assumptions about how a program is supposed to work to achieve its intended goals, or outcomes.
- Implementation theory: This looks at the activities involved in running a program, or how a group carries it out.

TBE also considers the role of context and how different settings or conditions can help or hinder the program's success. It examines the mechanisms (processes or structures) that influence whether a program achieves its goals or has unintended effects (see Figure 1) [37]. By combining these ideas, TBE supports evaluations that look at both outcomes (what was achieved) and implementation (how it was done) in context. This approach helps build evidence about what works, for whom, and in what situations. In the case of telehealth for conditions such as CF, TBE is therefore a valuable tool to fill gaps in the evidence.

Figure 1. Relationship between implementation theory, program theory, and context.

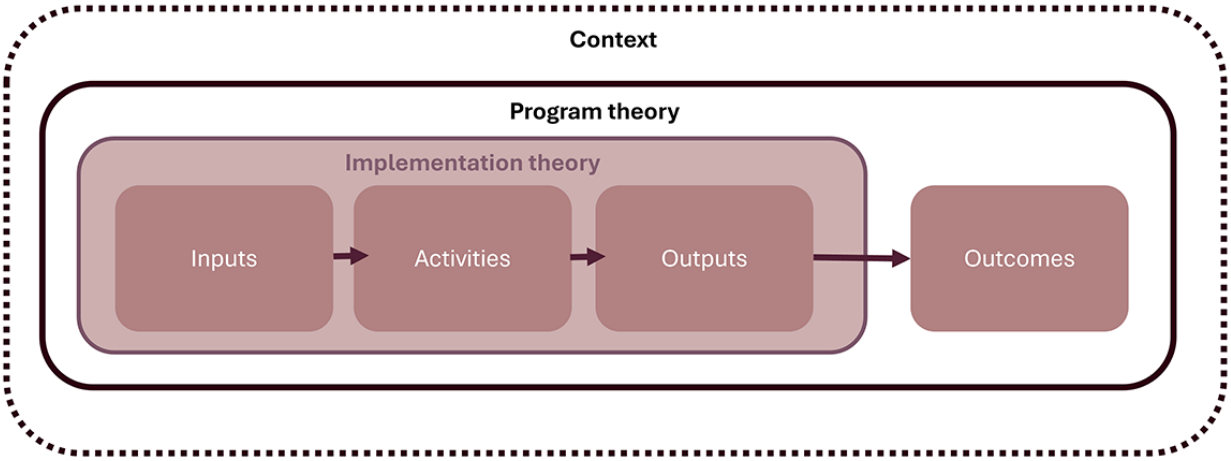

With TBE as a guiding principle, two reporting frameworks were selected from the Equator Network database [38] that individually deal with different aspects of implementation theory, program theory, and the context related to telehealth interventions: Guidelines and Checklist for the Reporting on Digital Health Implementations (iCHECK-DH) and Template for Intervention Description and Replication for Telehealth (TIDiER-telehealth). Collectively, the chosen reporting frameworks identify core components of both implementation theory and program theory.

iCHECK-DH [39] (referred to as iCHECK in this study) was published in 2023 and is a digital health implementation reporting tool. Although this framework does not deal specifically with telehealth interventions, it does capture the factors potentially affecting the implementation of telehealth interventions (see the *iCHECK Checklist* section in Multimedia Appendix 1).

TIDiER-telehealth [40] (referred to as TIDiER in this study) published in 2022, is a framework focused on reporting telehealth interventions in outcome-driven studies. It emphasizes detailing the intervention and its outcomes (i.e., program theory) but does not primarily address implementation reporting (see the *TIDieR-Telehealth* section in Multimedia Appendix 1).

Each framework details a standardized reporting checklist. For this review, we reverse-engineered the aforementioned checklists

XSL•FO
RenderX

to extract data from the selected studies to describe what information is being reported within this space and to identify what knowledge gaps exist.

Protocol

In 2023, we undertook a systematic review exploring CF telehealth studies [8] up to May 2021. This scoping review protocol is an extension of the previous systematic review and is not registered. The original systematic review methodology was repeated to gather more recent studies (May 2021-May 2025). This scoping review was reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) checklist (Multimedia Appendix 2) [41].

Eligibility Criteria

The inclusion criteria for the systematic refinement of studies included only scientific papers available in English. Conference abstracts, reviews, and gray literature were excluded from this review. Papers meeting the following criteria were included for review:

- Population: telehealth services delivered to CF populations or reviewed by CF care teams
- Concept: the implementation of any telehealth intervention adhering to the telehealth definition outlined by the THCF [4]:

  Telehealth provides safe and accessible care/education through remote consultations between a patient/caregiver and healthcare professional using communication technology (e.g., videoconferencing, smartphone). It can be further enhanced using remote patient monitoring technologies coupled with safe and meaningful digital data transfer and storage.
- Context: telehealth interventions delivered alongside and/or as part of routine CF care

Information Sources and Search

The search methods used were reported following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Reporting Literature Searches (PRISMA-S) checklist (Multimedia Appendix 3) [42]. The databases selected for this review included PubMed, Scopus, and Web of Science as they offer complementary strengths. Although all three databases index similar disciplines and research fields, PubMed was chosen as it predominantly indexes clinical, biomedical, and medical content. Scopus was chosen as it builds upon this to also include technology, business, economics, and finance journals. Finally, the Web of Science was chosen as it indexes many computer science, engineering, and technology journals.

The search string included was “Cystic Fibrosis” and one of the following terms, as described in the original systematic review: “Virtual Care,” “Virtual Health,” “Remote Monitoring,” “TeleMonitoring,” “Digital Health,” “Home Monitoring,” “eHealth,” “telehealth,” “telemedicine”, “wearable devices” [8].

Here is an example of the search string used for the Web of Science database:

“TS=(Cystic Fibrosis AND Virtual Health) OR TS=(Cystic Fibrosis AND Remote Monitoring) OR TS=(Cystic Fibrosis AND TeleMonitoring) OR TS=(Cystic Fibrosis AND Digital Health) OR TS=(Cystic Fibrosis AND home Monitoring) OR TS=(Cystic Fibrosis AND ehealth) OR TS=(Cystic Fibrosis AND telehealth) OR TS=(Cystic Fibrosis AND telemedicine) OR TS=(Cystic Fibrosis AND wearable devices) OR TS=(Cystic Fibrosis AND Virtual Care)”

For a full list of search strings for each database and how these strings were developed, see the *Search Strategy* section in Multimedia Appendix 1.

Other search methods, such as simultaneous multidatabase searching, study registry search, purposeful online browsing, citation searching, or contact with researchers, were not used in this search strategy.

Selection of Sources of Evidence

Once all sources of evidence were identified, duplicates were removed using formulas in Microsoft Excel. Titles and abstracts were reviewed to ensure papers met the inclusion criteria, as outlined before. Two reviewers (authors TV and RD) blind-reviewed the papers during each stage and resolved any conflicts through consensus discussions at each of the title, abstract, and full-text review stages. Consensus discussions centered primarily on whether the interventions included met the definition of telehealth, as defined before.

Data Charting and Data Items

The data selected for analysis included (1) study descriptive and (2) checklist data. Study descriptive data included the paper title, author list, year of publication, and journal. For data extraction, all 20 items from iCHECK [39] and 12 items from TIDiER [40] were used (see the data extraction template in Multimedia Appendix 4). The studies were scored against each category available in each checklist: fully reported on a category, score=2; partially reported on a category, score=1; and did not report, score=0. All extracted data were stored in Excel. The same scoring system was used for all category items, and no weighting was provided to any specific category. Although some categories may be seen as more relevant for practical implementation (eg, safety and data governance) and others are not mandatory (as per the checklists), we sought to identify reporting strengths and weaknesses only. Hence, an unweighted category–scoring approach was adopted.

See Table 1 for an example of scoring TIDiER category 8 (When and How Much: “Describe the number of times the intervention was delivered and over what period of time including the number of sessions, their schedule, and their duration, intensity or dose.”)

XSL•FO
RenderX

Table 1. Example of "fully reported," "partially reported," and "did not report" scoring for one checklist category.

Scoring label (score)	Study	Data example
Fully reported (2)	Graziano et al [43]	"The intervention consisted of four individual telehealth sessions, each lasting 30-40 minutes, delivered via online video platforms. The sessions were conducted during the COVID-19 lockdown in Italy, starting in March 2020. Participants received all four planned sessions. Feasibility and satisfaction were assessed one week after the last session."
Partially reported (1)	Verkleij et al [44]	"The intervention consists of eight 45-minute sessions. The initial intake, fourth, and last sessions are completed in real time with a therapist, while the remaining five sessions are self-guided. The paper does not specify the exact frequency or total duration over which these sessions are delivered."
Did not report (0)	Hasan et al [45]	"Not mentioned (the paper does not provide specific details on the number of sessions, their schedule, or duration for the telehealth intervention)"

Data extraction and scoring were performed using Elicit AI Research Assistant 2024 [46]. Elicit is an artificial intelligence (AI) tool that uses large language models (LLMs) to extract data from scientific studies and is reported as being 80%-90% accurate [47]. To extract data, the AI tool is fed prompts that include a label and a description of which data are to be extracted. For this review, we tused he category names and definitions, as described within the checklists, to inform the AI tool. The category names within the checklist were used as data extraction "labels," and the category definitions from the checklist were used as "label descriptions" (Multimedia Appendix 5).

Validation of AI Data Extraction

To evaluate the accuracy of the AI tool, two reviewers (TV and RD) also manually extracted and scored data from 18 studies (Table S1 in Multimedia Appendix 6). We calculated the difference between the manual and AI scoring for each paper and each checklist category. No disagreement in scoring was equal to a difference of 0, a minor disagreement was equal to a difference of 1, and major disagreement was equal to a difference of 2. The frequency of "no disagreements," "mnor disagreements," and "major disagreements" was calculated (Figure S1 and Figure S2 in Multimedia Appendix 6). We found that the title category had the highest number of major disagreements (5 for iCHECK and 12 for TIDiER), which was found to be related to various digital health terms available to describe telehealth (eg, telemedicine, telecare, and telemonitoring). However, for all other checklist categories, and per manuscript comparisons, the number of major disagreements was low (≤3), with the majority of values being 0 (Figure S1 and Figure S2 in Multimedia Appendix 6). Therefore, Elicit AI was found acceptable for this review.

Critical Appraisal of Individual Sources of Evidence

Due to the heterogeneity of the studies and variability in associated study data, a quality appraisal tool was not implemented for this review. Instead, we focused on the quantity of information reported according to the iCHECK and TIDiER checklists.

Synthesis of Results

The maximum total score that a paper can receive with iCHECK is 40 (ie, 20 categories with a maximum score of 2 per category) and with TIDiER is 24 (ie, 12 categories with a maximum score of 2 per category). This maximum score represents a paper that fully reports on all checklist categories, scoring 2 in each category. A paper such as this would be described as 100% adherent to the reporting checklist.

All papers and their checklist scores were stored in Excel. Descriptive statistics were performed on each of the manuscript scores and checklist categories.

Ethical Considerations

This study did not involve human participants.

Results

Selection of Sources of Evidence

From the original systematic review up to May 2021 (N=39), 13 (33%) conference abstracts were removed for this review. In the second phase of the systematic review (May 2021-May 2025), 873 new studies were identified. From this pool, 322 (36.9%) duplicates were removed. The remaining 551 (63.1%) studies underwent systematic refinement, and 419 (76.0%) studies were excluded at title screening, 50 (9.1%) at abstract screening, and 10 (1.8%) at introduction screening, leaving 72 (13.1%) studies that met the inclusion criteria. The 26 studies from the original systematic review and these 72 studies from this scoping review constituted a total of 98 studies from the beginning until May 2025 (see Figure 2). A full list of the final 98 studies [30,43-45,48-141] can be found in Table S2 in Multimedia Appendix 6.

XSL•FO
RenderX

Figure 2. PRISMA flow of the paper selection process using a combination of two reviews. The first considered papers from the beginning to May 2021 (n=26), and the updated review took place from May 2021 to May 2025. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

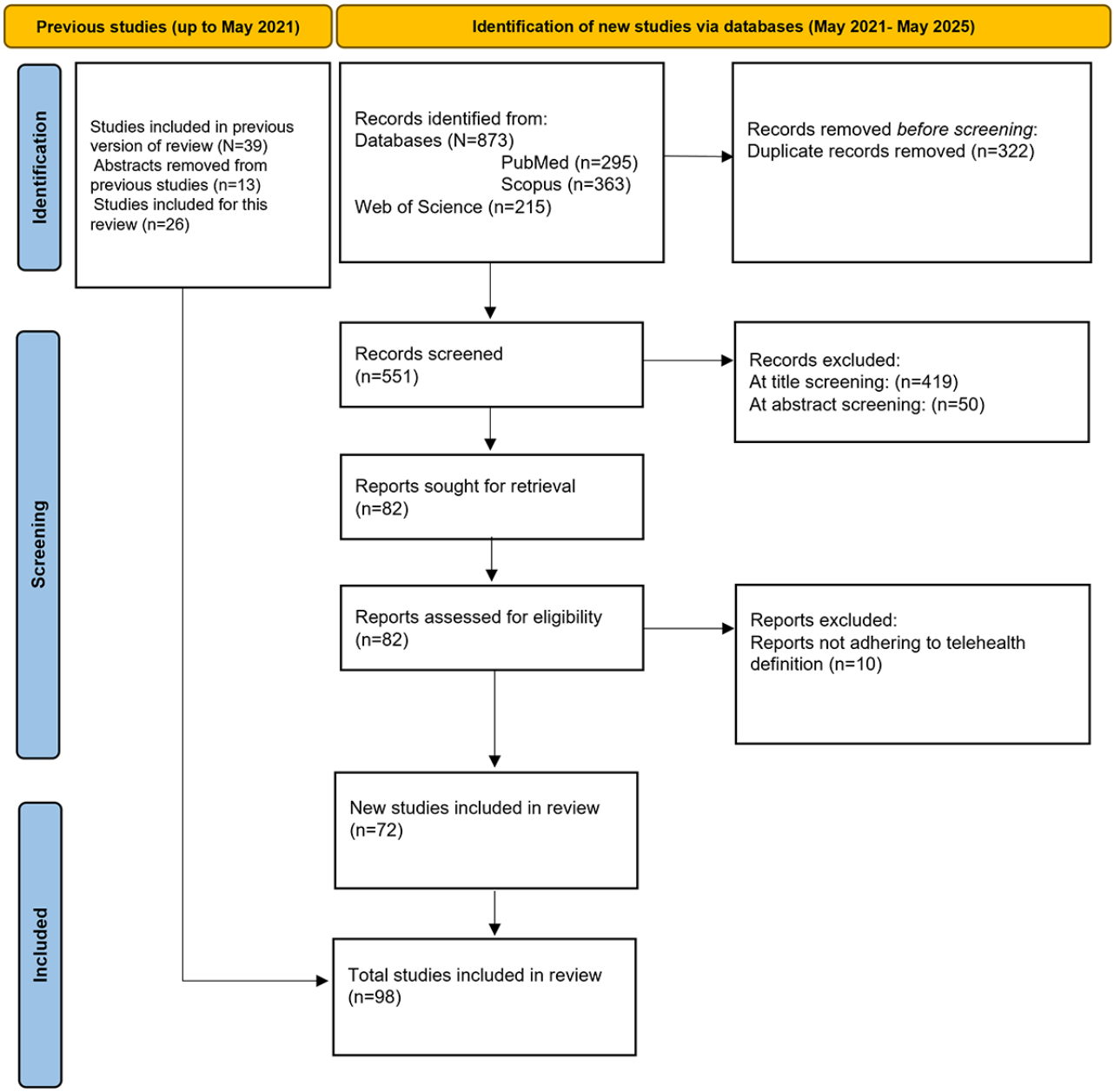

Characteristics of Sources of Evidence

A summary of the 98 studies (study design, population, country of origin, telehealth format, and key findings) can be seen in Table S3 in Multimedia Appendix 6. Of the 98 studies, 90 (91.8%) were conducted in a single country. The United States (n=42, 46.7%), Australia (n=11, 12.2%), and the United Kingdom (n=8, 8.9%) were the three most common countries. In addition, 6 (6.1%) studies were conducted in two or more countries, and 2 (2.0%) studies did not explicitly describe the country where they were conducted, and this information had to be inferred from author affiliations and ethics committee details. The earliest study was published in 2006 (n=1, 1.0%), with the highest number of studies being published during the pandemic (2021: n=28, 28.6%; 2022: n=16, 16.3%; Figure S3 in Multimedia Appendix 6). Furthermore, 79 (80.6%) studies were published in medical journals and the remaining 19 (19.4%) in digital health journals (Figure S4 in Multimedia Appendix 6). Of the 98 studies, 9 (9.2%) used the term “intervention” within the title and 3 (3.1%) used the term “implementation” (Figure S5 in Multimedia Appendix 6). We found 17 digital health terms used within the titles to describe the telehealth service: “telehealth” (n=41, 41.8%), “telemedicine” (n=14, 14.3%), and “home spirometry” (n=8, 8.2%) were the most common digital health terms used within the titles (Figure S6 in Multimedia Appendix 6). In addition, 80 (81.6%) studies used one digital health term, 10 (10.2%) studies used two terms, and 8 (8.2%) studies used no digital health terms. An example of a title with no digital health terms was “Understanding the Acceptability of the Changing Model of Care in Cystic Fibrosis” [136]. The studies used various telehealth modalities in various combinations (eg, video calls

XSL•FO
RenderX

in combination with remote monitoring). The most common telehealth modalities included video calls only (n=22, 22.4%); video calls, phone calls, and remote monitoring (n=20, 20.4%); video calls and remote monitoring (n=16, 16.3%); and phone calls and remote monitoring (n=11, 11.2%).

The population focus across the studies varied from adults with CF, adolescents with CF, and children with CF to their parents/caregivers, partners, and CF care teams, with many of the studies exploring a combination of these populations (eg, adults with CF and parents; adults with CF and CF care teams; or children with CF, adolescents with CF, and parents). Multiple study designs had been used across the 98 papers. The most common study types included cross-sectional studies (n=17, 17.3%), qualitative studies (n=13, 13.3%), randomized control trials (n=9, 9.8%), and observational studies (n=8, 8.2%).

Synthesis of Results: Standardized Reporting Scores

Overall Scores

Table S4 (iCHECK) and Table S5 (TIDiER) in Multimedia Appendix 6 provide a full list of category scores for each of the 98 studies (score=0 is marked in red to assist in visualizing the gaps). The median score calculated for all studies was 22/40 (range 11-29, 55.0% adherent) for iCHECK and 15/24 (range 6-23, 62.5% adherent) for TIDiER.

Category Scores for iCHECK

For iCHECK, across the 20 checklist categories, 11 (55%) scored ≥50% for adherence. Of the remaining 9 (45%) categories, 2 (22.2%) scored <10% (as shown in Table 1). In terms of reporting completeness, we calculated percentage scores across “fully reported,” “partially reported,” and “did not report.” For “fully reported,”, 6 (30%) categories scored ≥50%, while 4 (20%) scored of 0. For “partially reported,” 9 (45%) categories scored ≥50%, with 1 (5%) scoring <10%. Finally, for “did not report,” 3 (15%) categories scored ≥50%, with 3 (15%) other categories scoring 0 (Table 2; also see Table S6 and Figure S7A in Multimedia Appendix 6).

XSL•FO
RenderX

Table 2. Summary of scores for the iCHECK[a] checklist.

Categories	Adherence level (%)[b]	Overall adherence to each category (total score), n (%)	Percentage of papers (number) that:		
			Reported completely, n (%)	Partially reported, n (%)	Did not report, n (%)
Title					
Title	Moderate	147 (75.0)	58 (59.1)	31 (31.6)	9 (9.2)
Abstract					
Abstract	Very good	180 (91.8)	85 (86.7)	10 (10.2)	3 (3.0)
Introduction					
Context	Moderate	103 (52.5	5 (5.1)	93 (94.9)	0
Problem statement	Very good	179 (91.3	81 (82.6)	17 (17.3)	0
Similar interventions	Moderate	98 (50.0)	11 (11.2)	76 (77.5)	11 (11.2)
Methods					
Aims and objectives	Moderate	134 (68.4)	46 (46.9)	42 (42.8)	10 (10.2)
Blueprint summary	Moderate	125 (63.7)	46 (46.9)	37 (37.7)	17 (17.3)
Technical design	Low	62 (31.6)	2 (2.0)	58 (59.1)	38 (38.7
Target	Very good	187 (95.4	89 (90.8)	9 (9.2)	0
Data	Low	58 (29.6)	0	58 (59.2)	40 (40.8)
Interoperability	Very low	15 (7.6	0	15 (15.3)	83 (84.7)
Participating entities	Low	76 (38.7)	1 (1.0)	74 (75.5)	23 (23.4)
Budget planning	Very low	12 (6.1)	0	12 (12.2)	86 (87.8)
Sustainability	Low	44 (22.4)	0	44 (44.9)	54 (55.1)
Results					
Coverage	Moderate	87 (44.3)	10 (10.2)	67 (68.3)	21 (21.4)
Outcomes	Very good	162 (82.7)	75 (76.5)	12 (12.2)	11 (11.2)
Lessons learned	Moderate	96 (48.9)	1 (1.0)	94 (95.9)	3 (3.0)
Unintended consequences[c]	Moderate	100 (51.0)	20 (20.4)	60 (61.2)	18 (18.3)
Discussion					
Conclusion	Very good	165 (84.1	79 (80.6)	7 (7.1)	12 (12.2)
General					
General[c]	Moderate	118 (60.2)	32 (32.7)	54 (55.1)	12 (12.2)

[a]iCHECK: Guidelines and Checklist for the Reporting on Digital Health Implementations.

[b]Adherence level (by % range): very low, 0%-20%; low, 20%-40%; moderate, 40%-60%; good, 60%-80%; very good, 80%-99%.

[c]These categories are not mandatory within the original checklist.

Category Scores for TIDiER

For TIDiER, across the 12 checklist categories, 8 (66.7%) scored ≥50%. Of the remaining 4 (33.3%) categories, none scored <10% (as shown in Table 1). In terms of reporting completeness, 4 (33.3%) categories scored ≥50%, while 1 (2.8%) scored <10% for "fully reported." For "partially reported", 6 (50.0%) categories scored of ≥50%, with 2 (16.7%) scoring <10%. Finally, for "did not report," 1 (2.8%) category scored ≥50%, with 5 (41.7%) scoring <10% (Table 3; also see Table S6 and Figure S7B in Multimedia Appendix 6).

XSL•FO
RenderX

Table 3. Summary of scores for the TIDiER[a] checklist.

Categories	Adherence level (%)[b]	Overall adherence to each category (total score), n (%)	Percentage of papers (number) that:		
			Reported completely, n (%)	Partially reported, n (%)	Did not report, n (%)
Brief name	Low	70 (35.7)	35 (35.7)	0	63 (64.2)
Why	Very good	176 (89.8)	81 (82.7)	14 (14.3)	3 (3.1)
What					
What (materials)	Moderate	113 (57.7)	19 (19.4)	75 (76.5)	4 (4.1)
What (procedures)	Very good	174 (88.8)	84 (85.7)	6 (6.1)	8 (8.2)
Who provided	Low	66 (33.7)	5 (5.1)	56 (57.1)	37 (37.8)
How	Very good	181 (92.3)	83 (84.7)	15 (15.3)	0
Where	Good	157 (80.1)	60 (61.2)	37 (37.8)	1 (1.0)
When and how much	Moderate	109 (55.6)	23 (23.5)	63 (64.2)	12 (12.2)
Tailoring[c]	Moderate	103 (52.5)	27 (27.5)	49 (50.0)	22 (22.4)
Modifications[c]	Low	75 (38.3)	23 (23.5)	29 (29.6)	46 (46.9)
How well					
How (planned)[c]	Moderate	92 (46.9)	16 (16.3)	60 (61.2)	22 (22.4)
How (actual)[c]	Moderate	102 (52.0)	16 (16.3)	54 (55.1)	10 (10.2)

[a]TIDiER-telehealth: Template for Intervention Description and Replication for Telehealth.

[b]Adherence level (by % range): very low, 0%-20%; low, 20%-40%; moderate, 40%-60%; good, 60%-80%; very good, 80%-99%.

[c]These categories are not mandatory within the original checklist.

Discussion

Principal Findings

This scoping review examined how telehealth interventions for CF are reported using a TBE perspective with two established frameworks (TIDieR and iCHECK). To the best of our knowledge, this is the first review to use both TIDiER and iCHECK reporting standards in a structured deductive approach to assess the completeness of information reported in the telehealth literature. We found that although studies have consistently described why telehealth is used, whom it targets, and what outcomes are achieved, far less details are provided regarding how telehealth interventions are implemented in practice. Reporting related to technical design, data handling, interoperability, budget planning, sustainability, and operational modifications is notably limited. These gaps restrict our understanding of implementation processes and hinder the ability to interpret or replicate models of telehealth care across different settings. These findings describe deficiencies within telehealth reporting, not deficiencies within the intervention itself or how telehealth was implemented within each study's context. These reporting deficiencies highlight that reproducing these results in other contexts will be challenging.

Interpretation of Findings in Relation to the Existing Literature

The use of TBE to underpin the choice of reporting frameworks used in this review highlights gaps in understanding the implementation of telehealth interventions. By integrating program theory, implementation theory, and context, TBE guided us to address the extent of outcome reporting (what was achieved), implementation (how it was done), and contextual factors.

When comparing the two checklists, studies showed slightly greater adherence to the TIDiER (intervention) checklist than to the iCHECK (implementation) checklist. This suggests that available data more often frame telehealth as an isolated intervention rather than as a fully implemented care service pathway. Across both checklists, researchers consistently explain why telehealth is being used, identify target populations, and report study outcomes effectively. This provides valuable insight into the impact of telehealth within specific contexts. However, the most underreported areas relate to the practical aspects of implementation, indicating that evidence-based guidance on how to operationalize telehealth remains limited or unclear.

Technical design, data, interoperability, participating entities, budget planning, sustainability, brief name, who provided, and modifications were some of the lowest scoring categories across both checklists. Interoperability, budget planning, brief name, and modifications received the highest incidences of "did not report". Except "brief name," each of the other categories has a potential impact on implementing telehealth into clinical practice. One of the most discussed issues in the wider telehealth literature is that of budget planning [9]. Often, it is believed that telehealth can provide some reductions in cost to the health care system. However, increasing numbers of publications have

XSL•FO
RenderX

identified that there may not be any cost-saving benefits [48] and that costs may increase during initial implementation and later during routine service delivery [142]. This is further complicated by how telehealth is intended to be used within care delivery, for example, whether as a service that primarily operates on telehealth (with face-to-face meetings being used for exceptional circumstances) or as the more commonly described hybrid models [30]. Hybrid models often use telehealth in tandem with an existing service and can be additive to traditional services [142] and/or can cause overuse of care, which can further increase costs [143]. Understanding the context in which the telehealth service has been implemented, its intended use within the service, and the budget planning to support that would allow other services to interpret the transferability of the research into their own context without potential unforeseen financial pressure.

Sustainability (financial, economic, environmental, etc) was underreported and remains an area for continued discussion due to knowledge and evidence gaps. Within telehealth, environmental impact data can be both positive and negative [144-146]. For example, there is potential for reduced greenhouse emissions from less travel (patients and care teams), lower hospital equipment use, and less waste from sanitization procedures [144-146]. However, generating and operating digital systems increases energy use, and poorly managed recyclable wearable and portable devices create more waste [147]. A more prominent challenge is sustaining telehealth services over time [148,149]. Following the COVID-19 surge, many services reported a decline in use [150]. Research is ongoing to understand this trend, sometimes described as “telehealth fatigue” [148,151]. Addressing it requires detailed best practices and delivery strategies [151]. Yet, without sustainability data in publications, identifying effective, context-specific strategies remains difficult, potentially contributing to further decline and impacting areas such as budget planning.

Telehealth is inherently multidisciplinary, blending health care with computing and engineering. Papers are often authored by health care professionals and published in medical journals, which may explain why technical categories (eg, technical design, data, interoperability, modifications) are underreported. Although many arguments support expanding these areas, two worth highlighting are data security/protection and technology literacy. As technology becomes more embedded in health care, implementing proper safety protocols for managing health data is essential. Poorly managed telehealth can pose serious privacy risks [12], reducing patient and organizational trust, leading to decreased use [152], and affecting service sustainability. Technology literacy (described as the skills, attitudes, and knowledge to use technology confidently) is another barrier to telehealth adoption and influences health needs [153]. A lack of digital health literacy (technology and health literacy combined) can worsen existing health inequalities and increase the burden on care teams [154-156]. To improve digital health literacy and data security awareness, better reporting of these categories in telehealth studies is needed.

Finally, the COVID-19 pandemic caused an increase in the adoption, use, and distribution of scientific publications describing telehealth [157,158]. Now, the focus is moving from rapid adoption to consideration of implementation barriers and telehealth sustainability challenges [16,159]. Considering this, it is important that future research in this space continue to be published in adherence to a standardized checklist. This will not only provide more robust and consistent evidence for the use of telehealth but also data to allow care services to interpret and transfer this knowledge into their own context in a relevant, safe, and meaningful way.

Clinical Implications of the Current Evidence Base

Telehealth interventions in the selected studies were described at a clinical level (whom they target, why they are used, and what outcomes they achieve), but the practical realities of implementing them in existing care pathways remain insufficiently detailed or negated entirely. This can lead to several implications for clinical practice. First, without information about workflows, staff roles, technical infrastructure, and required modifications to existing processes, health care teams cannot reliably anticipate the operational impact of adopting a telehealth model. This increases the risk for failed or unsustainable implementations [148,149].

Second, the underreporting of budget planning and sustainability prevents clinicians and service managers from evaluating ongoing resource needs, cost pressures, or unintended cost escalations. This gap is clinically important because budget constraints directly influence workforce allocation, service availability, and long-term viability [9,142,143].

Third, limited reporting on data management, interoperability, and digital literacy considerations impedes care teams’ ability to assess whether the telehealth approach protects patient privacy, accommodates patients with low digital health literacy or access to technology and Wi-Fi (avoiding digital exclusion as part of the digital divide), or exacerbates health inequalities already caused by socioeconomic biases and geographic disparities [12,154-156].

Fourth, without contextual implementation data, clinical leaders cannot distinguish whether an observed improvement (eg, reduced admissions, improved adherence) is due to the telehealth model itself or due to local adaptations, resource differences, or specific contextual enablers [16,160].

Finally, insufficient details on modifications, technical design, and operational processes undermine the transferability and scalability of successful telehealth approaches. This slows down evidence uptake and perpetuates fragmented service models [12,161].

Limitations

First, there are no universally accepted terms and definitions within the digital health space [2,3]. Furthermore, due to the growing interest in said space, more terms are emerging. As a result, the search strategy in this review (10 terms to describe telehealth as per the outlined definition) can be expanded to include other terms with the potential to identify other studies in the future. Similarly, other databases can also be incorporated to widen the search. This review was limited to publications written in English. We also noted that some of the publications in this review predate the selected checklists and also may have

XSL•FO
RenderX

been unknown to the authors of more recent publications. This delay in knowledge transfer and research dissemination is common and somewhat expected, as multiple studies have shown that awareness and uptake of new guidelines often lag for years despite active dissemination efforts [162]. Hence, we did not aim to describe how adherent any individual study is to these checklists and instead focused on knowledge gaps across the literature. It should also be noted that some of the categories in the checklists are "not mandatory" (see Tables 2 and 3) and therefore not essential; however, for this review, these were included.

In conceptualizing and structuring this review, the reporting frameworks identified were individually found to be lacking important aspects of TBE and telehealth. The decision to use two frameworks overcame this issue to a certain extent by attempting to collectively present a more comprehensive picture of reporting of the implementation of telehealth interventions.

Furthermore, many of the included studies did not describe their research as either an intervention outcome study or an implementation study, and as such, it was necessary to score all studies against both checklists to remove potential bias.

This translational challenge of intervention versus implementation in the digital health field is becoming increasingly recognized and explored [163]. Projects such as the European Digital Health Technology Assessment framework (EDiHTA) [164] are specifically developing evaluation methodologies, which may, in time, improve the evaluation of digital health tools and the reporting seen in digital health studies.

The use of both manual and AI scoring of studies highlights the challenges in implementing these kinds of frameworks objectively. For example, in manually scoring the papers against each of the frameworks, we noted the need to clarify category definitions to ensure a consistent scoring mechanism was used. Likewise, the AI tool struggled with more simple categories, such as identifying whether the paper titles reported were linked to telehealth. This was mainly seen within the "title" and "brief name" categories, where AI adopted a lenient scoring method identifying papers using various terms to describe telehealth (telemedicine, remote monitoring, virtual consultation, etc). However, manual scoring found that the terms were too broad and varying and many did not use the term "telehealth" or "implementation" specifically. Currently, many of the AI tools designed to support research are approximately >80% accurate [47,165], and the discrepancies seen in this study are to be expected. Future research in this area could potentially improve this accuracy by dedicating additional time and expertise to AI prompt engineering.

Conclusion

As the volume of telehealth research continues to grow, it is increasingly important to evaluate not just its efficacy but also the robustness of its implementation [166-169]. The TBE approach followed in this study provides a novel deductive and structured means to assess this evidence base using the selected checklists. Collective adherence to established reporting frameworks remains inconsistent, which may limit reproducibility and transferability of findings. Encouragingly, certain checklists categories are well documented, offering valuable insights into telehealth's potential. However, the underreporting of implementation processes persists, in part due to the relative novelty of frameworks such as TIDiER and iCHECK. These findings reflect a deficiency in reporting telehealth implementation research and the potential translation of this knowledge into other contexts. These reporting deficiencies do not reflect a deficiency in the telehealth intervention or the identified authors' ability to implement telehealth. However, strengthening the commitment to comprehensive, framework-aligned reporting is essential to improve the effective and equitable integration of telehealth into routine care.

Although CF is positioned as a case exemplar, the implications of this work extend beyond a single condition. Many complex chronic diseases, such as chronic obstructive pulmonary disorder (COPD), heart failure, diabetes, and rheumatologic conditions, share similar challenges relating to fluctuating illness trajectories, multidisciplinary care needs, and variability in patient digital readiness. As such, the findings of this study are also applicable widely to other chronic conditions. The potential impact of the identified underreported sections will carry a similar impact in other telehealth care cases and need to be considered by all when reporting on telehealth implementations so as to strengthen telehealth reporting consistency across diverse clinical contexts. Improving transparency around implementation processes, contextual factors, and adaptations not only enhances the comparability of telehealth studies but also supports the more reliable translation of evidence into real-world practice for a broad range of chronic conditions. Finally, the novel method presented in this review is transferable and can be used to evaluate telehealth reporting in other health care settings and specialties.

Acknowledgments

We would like to acknowledge Cystic Fibrosis Ireland for their support in this research. Cystic Fibrosis Ireland played no role in the study design; collection, analysis, or interpretation of data; or the writing of the paper.

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT Taxonomy 2025, the following tasks were delegated to GAI tools under full human supervision: data collection, data curation and organization, and data analysis. The GAI tool used was Elicit Plus 2025-2026. Responsibility for the final paper lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. We also used Elicit AI to extract data from the selected studies based on predefined checklists. Elicit AI extracted data under specified structured answers to allow for scoring.

XSL•FO
RenderX

Funding

TV and BJP received research funding from Cystic Fibrosis Ireland from 2023 to 2024, without which this work would not have come to fruition.

Data Availability

Data are presented in the main paper, and data are also provided in Multimedia Appendices available for download, along with the published paper.

Authors' Contributions

Conceptualization: TV, RD, SCR, and BJP
Methodology: TV and RD
Software: TV and RD
Data curation: TV and RD
Writing—original draft preparation: TV, RD, SCR, and BJP
Visualization: : TV, RD, SCR, and BJP
Writing—reviewing and editing: : TV, RD, SCR, and BJP

Conflicts of Interest

TV and BJP received a Vertex CF-RIA award for exploring digital health literacy within cystic fibrosis telehealth. RD and SCR have no conflicts of interest to declare.

Multimedia Appendix 1
Guidelines and Checklist for the Reporting on Digital Health Implementations (iCHECK-DH) checklist, Template for Intervention Description and Replication for Telehealth (TIDieR-telehealth) checklist, and search strategy for the review.
[DOCX File , 52 KB - jmir_v28i1e86194_app1.docx]

Multimedia Appendix 2
Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) checklist.
[PDF File (Adobe PDF File), 173 KB - jmir_v28i1e86194_app2.pdf]

Multimedia Appendix 3
Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Reporting Literature Searches (PRISMA-S) checklist.
[PDF File (Adobe PDF File), 137 KB - jmir_v28i1e86194_app3.pdf]

Multimedia Appendix 4
Data extraction template.
[XLSX File (Microsoft Excel File), 15 KB - jmir_v28i1e86194_app4.xlsx]

Multimedia Appendix 5
Elicit artificial intelligence (AI) prompts.
[DOCX File , 28 KB - jmir_v28i1e86194_app5.docx]

Multimedia Appendix 6
Manual reviewer data extraction compared with Elicit artificial intelligence (AI) extraction; summary of studies for the review; summary of the final 98 studies, including study design, population, research objective, country of origin, telehealth format, and key findings; scores for all 98 papers for the Guidelines and Checklist for the Reporting on Digital Health Implementations (iCHECK) and Template for Intervention Description and Replication for Telehealth (TIDiER) checklists; summary of scores for iCHECK and TIDiER checklists; iCHECK and TIDiER frequencies of 3 disagreement types recorded per paper and per checklist category for 18 studies; publication frequency based on years; discipline focus of journals (medical and digital health); number of papers that self-described as implementation or intervention in the title; digital health terms used in paper titles; and overall category scores (%) of iCHECK and TIDiER.
[DOCX File , 530 KB - jmir_v28i1e86194_app6.docx]

References

XSL•FO
RenderX

1. Mumporeze C. Evaluating telehealth: a systematic review of its effectiveness and challenges post Covid-19. Public Health Inst J 2025 Jul 14;2025(9) [FREE Full text]
2. Rahmah L, Purnomo AF, Alizadeh M, Askari SS, Lutfiana NC, Mundakir M, et al. Are we all singing from the same song sheet? Standardizing terminology used in inter-professional telehealth education and practice: a mixed method study. BMC Med Educ 2025 May 05;25(1):649 [FREE Full text] [doi: 10.1186/s12909-025-07207-6] [Medline: 40325451]
3. Maheu M, Whitten P, Allen A. E-Health, Telehealth, and Telemedicine: A Guide to Startup and Success. San Francisco, CA: Jossey-Bass (Wiley imprint); Feb 2002.
4. Telehealth in cystic fibrosis. ECFS Telehealth for Cystic Fibrosis Working Group. 2023. URL: https://www.ecfs.eu/telehealth [accessed 2025-01-10]
5. Fernandes G, Figueiredo T, Costa E, Coelho L, Loyens D. Impact of telehealth on health outcomes and quality of life in the older adults population: a systematic review. Front Digit Health 2025 Dec 18;7:1708960 [FREE Full text] [doi: 10.3389/fdgth.2025.1708960] [Medline: 41487302]
6. Parker S, Prince A, Thomas L, Song H, Milosevic D, Harris MF, IMPACT Study Group. Electronic, mobile and telehealth tools for vulnerable patients with chronic disease: a systematic review and realist synthesis. BMJ Open 2018 Aug 29;8(8):e019192 [FREE Full text] [doi: 10.1136/bmjopen-2017-019192] [Medline: 30158214]
7. Bitar H, Alismail S. The role of eHealth, telehealth, and telemedicine for chronic disease patients during COVID-19 pandemic: a rapid systematic review. Digit Health 2021;7:20552076211009396 [FREE Full text] [doi: 10.1177/20552076211009396] [Medline: 33959378]
8. Vagg T, Shanthikumar S, Ibrahim H, O'Regan P, Chapman WW, Kirwan L, et al. Telehealth in Cystic Fibrosis. A systematic review incorporating a novel scoring system and expert weighting to identify a 'top 10 manuscripts' to inform future best practices implementation. J Cyst Fibros 2023 Jul;22(4):598-606 [FREE Full text] [doi: 10.1016/j.jcf.2023.05.012] [Medline: 37230808]
9. Helleman J, Kruitwagen ET, van den Berg LH, Visser-Meily JMA, Beelen A. The current use of telehealth in ALS care and the barriers to and facilitators of implementation: a systematic review. Amyotroph Lateral Scler Frontotemporal Degener 2020 May 26;21(3-4):167-182 [FREE Full text] [doi: 10.1080/21678421.2019.1706581] [Medline: 31878794]
10. Houser SH, Flite CA, Foster SL, Hunt TJ, Morey A, Palmer MN, et al. Patient clinical documentation in telehealth environment: are we collecting appropriate and sufficient information for best practice? Mhealth 2022 Jan;8:6 [FREE Full text] [doi: 10.21037/mhealth-21-30] [Medline: 35178437]
11. Kayyali R, Hesso I, Ejiko E, Nabhani Gebara S. A qualitative study of telehealth patient information leaflets (TILs): are we giving patients enough information? BMC Health Serv Res 2017 May 19;17(1):362 [FREE Full text] [doi: 10.1186/s12913-017-2257-5] [Medline: 28526026]
12. Houser S, Flite C, Foster S. Privacy and security risk factors related to telehealth services - a systematic review. Perspect Health Inf Manag 2023;20(1):1f [FREE Full text] [Medline: 37215337]
13. Amagai S, Vonesh E, Adams J, Luo Y. Closing the gap: addressing telehealth disparities across specialties in the sustained pandemic era. NPJ Digit Med 2024 Aug 21;7(1):217 [FREE Full text] [doi: 10.1038/s41746-024-01201-w] [Medline: 39164391]
14. Mahdavi S, Fekri M, Mohammadi-Sarab S, Mehmandoost M, Zarei E. The use of telemedicine in family medicine: a scoping review. BMC Health Serv Res 2025 Mar 13;25(1):376 [FREE Full text] [doi: 10.1186/s12913-025-12449-7] [Medline: 40082924]
15. Oudbier SJ, Souget-Ruff SP, Chen BSJ, Ziesemer KA, Meij HJ, Smets EMA. Implementation barriers and facilitators of remote monitoring, remote consultation and digital care platforms through the eyes of healthcare professionals: a review of reviews. BMJ Open 2024 Jun 10;14(6):e075833 [FREE Full text] [doi: 10.1136/bmjopen-2023-075833] [Medline: 38858155]
16. Rabbani MG, Alam A, Prybutok VR. Digital health transformation through telemedicine (2020–2025): barriers, facilitators, and clinical outcomes—a systematic review and meta-analysis. Encyclopedia 2025 Dec 04;5(4):206 [FREE Full text] [doi: 10.3390/encyclopedia5040206]
17. Nourani A, Hosseini SM, Rassoulian M, Joudivand L, Samimi T, Rahimi B. A systematic review of methods used for COVID-19 telehealth systems evaluation: lessons from past experiences for future use. BMC Health Serv Res 2025 Aug 16;25(1):1089 [FREE Full text] [doi: 10.1186/s12913-025-13315-2] [Medline: 40819047]
18. Ronan NJ, Elborn JS, Plant BJ. Current and emerging comorbidities in cystic fibrosis. Presse Med 2017 Jun;46(6 Pt 2):e125-e138 [FREE Full text] [doi: 10.1016/j.lpm.2017.05.011] [Medline: 28554721]
19. Appelt D, Fuchs T, Steinkamp G, Ellemunter H. Malignancies in patients with cystic fibrosis: a case series. J Med Case Rep 2022 Jan 19;16(1):27 [FREE Full text] [doi: 10.1186/s13256-021-03234-1] [Medline: 35042562]
20. Parisi GF, Papale M, Pecora G, Rotolo N, Manti S, Russo G, et al. Cystic fibrosis and cancer: unraveling the complex role of CFTR gene in cancer susceptibility. Cancers (Basel) 2023 Aug 24;15(17):4244 [FREE Full text] [doi: 10.3390/cancers15174244] [Medline: 37686519]
21. McGlynn J, DeCelie-Germana JK, Kier C, Langfelder-Schwind E. Reproductive counseling and care in cystic fibrosis: a multidisciplinary approach for a new therapeutic era. Life (Basel) 2023 Jul 12;13(7):1545 [FREE Full text] [doi: 10.3390/life13071545] [Medline: 37511919]

XSL•FO
RenderX

22. Vagg T, Shanthikumar S, Morrissy D, Chapman W, Plant B, Ranganathan S. Telehealth and virtual health monitoring in cystic fibrosis. Curr Opin Pulm Med 2021 Nov 01;27(6):544-553 [FREE Full text] [doi: 10.1097/MCP.0000000000000821] [Medline: 34431789]
23. Finkelstein SM, Budd JR, Warwick WJ, Kujawa SJ, Wielinski CL, Ewing LB. Feasibility and compliance studies of a home measurement monitoring program for cystic fibrosis. J Chronic Dis 1986 Jan;39(3):195-205 [FREE Full text] [doi: 10.1016/0021-9681(86)90024-x] [Medline: 3949943]
24. Acosta N, Heirali A, Somayaji R, Surette MG, Workentine ML, Sibley CD, et al. Sputum microbiota is predictive of long-term clinical outcomes in young adults with cystic fibrosis. Thorax 2018 Nov;73(11):1016-1025 [FREE Full text] [doi: 10.1136/thoraxjnl-2018-211510] [Medline: 30135091]
25. Mingora CM, Flume PA. Pulmonary complications in cystic fibrosis: past, present, and future: adult cystic fibrosis series. Chest 2021 Oct;160(4):1232-1240 [FREE Full text] [doi: 10.1016/j.chest.2021.06.017] [Medline: 34147501]
26. Berlinski A, Chambers MJ, Willis L, Homa K, Com G. Redesigning care to meet national recommendation of four or more yearly clinic visits in patients with cystic fibrosis. BMJ Qual Saf 2014 Apr;23 Suppl 1:i42-i49 [FREE Full text] [doi: 10.1136/bmjqs-2013-002345] [Medline: 24608550]
27. De Wachter E, De Boeck K, Sermet-Gaudelus I, Simmonds N, Munck A, Naehrlich L, ECFS Diagnostic Network Working Group. ECFS standards of care on CFTR-related disorders: towards a comprehensive program for affected individuals. J Cyst Fibros 2024 May;23(3):388-397 [FREE Full text] [doi: 10.1016/j.jcf.2024.01.012] [Medline: 38388234]
28. Roseby C, Ellis K, Edwards R, Douglas T, Hunt S, Schultz A. Cystic fibrosis standards of care, Austrailia. Cystic Fibrosis Australia. 2025. URL: https://www.cysticfibrosis.org.au/standards-of-care/ [accessed 2026-03-26]
29. Gifford AH, Ong T, Dowd C, Van Citters AD, Scalia P, Sabadosa KA, et al. Evaluating barriers to and promoters of telehealth during the COVID-19 pandemic at U.S. cystic fibrosis programs. J Cyst Fibros 2021 Dec;20 Suppl 3:9-13 [FREE Full text] [doi: 10.1016/j.jcf.2021.08.034] [Medline: 34930545]
30. Dixon E, Dick K, Ollosson S, Jones D, Mattock H, Bentley S, et al. Telemedicine and cystic fibrosis: do we still need face-to-face clinics? Paediatr Respir Rev 2022 Jun;42:23-28 [FREE Full text] [doi: 10.1016/j.prrv.2021.05.002] [Medline: 34215541]
31. Van Citters AD, Buus-Frank ME, King JR, Seid M, Holthoff MM, Amin RS, et al. The cystic fibrosis learning network: a mixed methods evaluation of program goals, attributes, and impact. Learn Health Syst 2023 Jul 20;7(3):e10356 [FREE Full text] [doi: 10.1002/lrh2.10356] [Medline: 37731865]
32. Ketchell RI. Telemedicine is the way forward for the management of cystic fibrosis - the case in favour. Paediatr Respir Rev 2018 Mar;26:19-21 [FREE Full text] [doi: 10.1016/j.prrv.2017.03.004] [Medline: 28416301]
33. Desimone ME, Sherwood J, Soltman SC, Moran A. Telemedicine in cystic fibrosis. J Clin Transl Endocrinol 2021 Dec;26:100270 [FREE Full text] [doi: 10.1016/j.jcte.2021.100270] [Medline: 34765457]
34. Cox NS, Alison JA, Rasekaba T, Holland AE. Telehealth in cystic fibrosis: a systematic review. J Telemed Telecare 2012 Mar;18(2):72-78 [FREE Full text] [doi: 10.1258/jtt.2011.110705] [Medline: 22198961]
35. Funnell SC, Rogers PJ. Purposeful Program Theory: Effective Use of Theories of Change and Logic Models. Hoboken, NJ: John Wiley & Sons; Mar 15, 2011.
36. Palumbo DJ, Oliverio A. Implementation theory and the theory-driven approach to validity. Eval Program Planning 1989 Jan;12(4):337-344 [FREE Full text] [doi: 10.1016/0149-7189(89)90050-5]
37. Pawson R, Tilley N. Realistic Evaluation. London, UK: SAGE Publications; Apr 1997.
38. Enhancing the quality and transparency of health research. Equator Network. URL: https://www.equator-network.org [accessed 2025-01-10]
39. Perrin Franck C, Babington-Ashaye A, Dietrich D, Bediang G, Veltsos P, Gupta PP, et al. iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res 2023 May 10;25:e46694 [FREE Full text] [doi: 10.2196/46694] [Medline: 37163336]
40. Rhon DI, Fritz JM, Kerns RD, McGeary DD, Coleman BC, Farrokhi S, et al. TIDieR-telehealth: precision in reporting of telehealth interventions used in clinical trials - unique considerations for the Template for the Intervention Description and Replication (TIDieR) checklist. BMC Med Res Methodol 2022 Jun 02;22(1):161 [FREE Full text] [doi: 10.1186/s12874-022-01640-7] [Medline: 35655144]
41. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, Tunçalp, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/m18-0850]
42. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc 2021 Apr 01;109(2):174-200 [FREE Full text] [doi: 10.5195/jmla.2021.962] [Medline: 34285662]
43. Graziano S, Boldrini F, Righelli D, Milo F, Lucidi V, Quittner A, et al. Psychological interventions during COVID pandemic: telehealth for individuals with cystic fibrosis and caregivers. Pediatr Pulmonol 2021 Jul 27;56(7):1976-1984 [FREE Full text] [doi: 10.1002/ppul.25413] [Medline: 33905614]

XSL•FO
RenderX

44. Verkleij M, Georgiopoulos AM, Friedman D. Development and evaluation of an internet-based cognitive behavioral therapy intervention for anxiety and depression in adults with cystic fibrosis (eHealth CF-CBT): an international collaboration. Internet Interv 2021 Apr;24:100372 [FREE Full text] [doi: 10.1016/j.invent.2021.100372] [Medline: 33816126]
45. Hasan S, Cecilia Lansang M, Salman Khan M, Dasenbrook E. Managing cystic fibrosis related diabetes via telehealth during COVID-19 pandemic. J Clin Transl Endocrinol 2021 Mar;23:100253 [FREE Full text] [doi: 10.1016/j.jcte.2021.100253] [Medline: 33723506]
46. Elicit. Elicit: the AI research assistant. Elicit. 2023. URL: https://support.elicit.com/en/articles/14757952-citing-elicit-search-methodology-and-using-elicit-s-content-in-your-own-work [accessed 2026-05-15]
47. Information and advice from the Elicit team, Elicit's limitations. Elicit Help Cente. URL: https://support.elicit.com/en/articles/549569 [accessed 2025-01-10]
48. Franz N, Rapp H, Hansen RN, Gold LS, Goss CH, Lechtzin N, et al. Health care costs related to home spirometry in the eICE randomized trial. J Cyst Fibros 2022 Jan;21(1):61-69 [FREE Full text] [doi: 10.1016/j.jcf.2021.02.014] [Medline: 33715993]
49. Jaclyn D, Andrew NS, Ryan P, Julianna B, Christopher S, Nauman C, et al. Patient and family perceptions of telehealth as part of the cystic fibrosis care model during COVID-19. J Cyst Fibros 2021 May;20(3):e23-e28 [FREE Full text] [doi: 10.1016/j.jcf.2021.03.009] [Medline: 33775604]
50. Benz C, Middleton A, Elliott A, Harvey A. Physiotherapy via telehealth for acute respiratory exacerbations in paediatric cystic fibrosis. J Telemed Telecare 2021 Mar 16;29(7):552-560 [FREE Full text] [doi: 10.1177/1357633x21998205]
51. Hatziagorou E, Toulia I, Avramidou V, Kampouras A, Tsara V, Tsanakas J. Change in CF care during COVID-19 pandemic: single-center experience in a middle-income setting. Pediatr Pulmonol 2021 Sep;56(9):3065-3067 [FREE Full text] [doi: 10.1002/ppul.25560] [Medline: 34265187]
52. Corcoran J, Marley Campbell C, Ladores S. Transitioning to telehealth during the coronavirus disease 2019 pandemic: perspectives from partners of women with cystic fibrosis and healthcare providers. Chronic Illn 2023 Mar;19(1):95-101 [FREE Full text] [doi: 10.1177/17423953211060257] [Medline: 34812062]
53. Van Citters AD, Dieni O, Scalia P, Dowd C, Sabadosa KA, Fliege JD, et al. Barriers and facilitators to implementing telehealth services during the COVID-19 pandemic: a qualitative analysis of interviews with cystic fibrosis care team members. J Cyst Fibros 2021 Dec;20 Suppl 3:23-28 [FREE Full text] [doi: 10.1016/j.jcf.2021.09.004] [Medline: 34930537]
54. Ong T, Van Citters AD, Dowd C, Fullmer J, List R, Pai SA, et al. Remote monitoring in telehealth care delivery across the U.S. cystic fibrosis care network. J Cyst Fibros 2021 Dec;20 Suppl 3:57-63 [FREE Full text] [doi: 10.1016/j.jcf.2021.08.035] [Medline: 34930544]
55. Beşer Ö, Karaboğa EPU, Hepkaya E, Kılınç Sakallı AA, Dönmez Türkmen A, Dilek TD, et al. The role of telehealth services in children with cystic fibrosis during coronavirus disease 2019 outbreak. Telemed J E Health 2022 Jun;28(6):832-837 [FREE Full text] [doi: 10.1089/tmj.2021.0228] [Medline: 34619057]
56. Hendra K, Neemuchwala F, Chan M, Ly NP, Gibb ER. Patient and provider experience with cystic fibrosis telemedicine clinic. Front Pediatr 2021;9:784692 [FREE Full text] [doi: 10.3389/fped.2021.784692] [Medline: 34900879]
57. Solomon GM, Bailey J, Lawlor J, Scalia P, Sawicki GS, Dowd C, et al. Patient and family experience of telehealth care delivery as part of the CF chronic care model early in the COVID-19 pandemic. J Cyst Fibros 2021 Dec;20 Suppl 3:41-46 [FREE Full text] [doi: 10.1016/j.jcf.2021.09.005] [Medline: 34930542]
58. Dowd C, Van Citters AD, Dieni O, Willis A, Powell L, Sabadosa KA. Design and methods for understanding the state of cystic fibrosis care amid the COVID-19 pandemic. J Cyst Fibros 2021 Dec;20 Suppl 3:3-8 [FREE Full text] [doi: 10.1016/j.jcf.2021.08.028] [Medline: 34930539]
59. Ahmed R, Greenfield M, Morley CP, Desimone M. Satisfaction and concerns with telemedicine endocrine care of patients with cystic fibrosis. Telemed Rep 2022;3(1):93-100 [FREE Full text] [doi: 10.1089/tmr.2021.0053] [Medline: 35720441]
60. Kenis-Coskun Ö, Aksoy AN, Kumaş EN, Yılmaz A, Güven E, Ayaz HH, et al. The effect of telerehabilitation on quality of life, anxiety, and depression in children with cystic fibrosis and caregivers: a single-blind randomized trial. Pediatr Pulmonol 2022 May;57(5):1262-1271 [FREE Full text] [doi: 10.1002/ppul.25860] [Medline: 35174670]
61. Albon D, Thomas L, Hoberg L, Stamper S, Somerville L, Varghese P, CF Learning Network telehealth iLab working group. Cystic fibrosis learning network telehealth innovation lab during the COVID-19 pandemic: a success QI story for interdisciplinary care and agenda setting. BMJ Open Qual 2022 May;11(2):e001844 [FREE Full text] [doi: 10.1136/bmjoq-2022-001844] [Medline: 35589277]
62. Rodkjær L, Jeppesen M, Schougaard L. Management of cystic fibrosis during COVID-19: patient reported outcomes based remote follow-up among CF patients in Denmark - a feasibility study. J Cyst Fibros 2022 Mar;21(2):e106-e112 [FREE Full text] [doi: 10.1016/j.jcf.2021.10.010] [Medline: 34785157]
63. Costa RLF, Costa RF, Gonçalves CP, Cohen RWF, Santana NN. Telemedicine of patients with cystic fibrosis during the COVID-19 pandemic. Rev Paul Pediatr 2022;40:e2021118 [FREE Full text] [doi: 10.1590/1984-0462/2022/40/2021118in] [Medline: 35544907]
64. Nash EF, Choyce J, Carrolan V, Justice E, Shaw KL, Sitch A, et al. A prospective randomised controlled mixed-methods pilot study of home monitoring in adults with cystic fibrosis. Ther Adv Respir Dis 2022;16:17534666211070133 [FREE Full text] [doi: 10.1177/17534666211070133] [Medline: 35274585]

XSL•FO
RenderX

65. Bathgate CJ, Kilbourn KM, Murphy NH, Wamboldt FS, Holm KE. Pilot RCT of a telehealth intervention to reduce symptoms of depression and anxiety in adults with cystic fibrosis. J Cyst Fibros 2022 Mar;21(2):332-338 [FREE Full text] [doi: 10.1016/j.jcf.2021.07.012] [Medline: 34366282]
66. Franciosi AN, Wilcox PG, Quon BS. Cystic fibrosis respiratory microbiology monitoring during a global pandemic: lessons learned from a shift to telehealth. Ann Am Thorac Soc 2022 Mar;19(3):498-500 [FREE Full text] [doi: 10.1513/AnnalsATS.202101-087RL] [Medline: 34406918]
67. Fadaizadeh L, Hassanzad M, Valinejadi A, Taheri SMJ, Farnia P, Hassanzad N, et al. Evaluation of mobile phone-based tele-monitoring of cystic fibrosis patients during the COVID-19 pandemic: a 3-year experience in Iran. Biomed Biotechnol Res J 2022;6(2):261-265 [FREE Full text] [doi: 10.4103/bbrj.bbrj_60_22]
68. Somerville LAL, List RP, Compton MH, Bruschwein HM, Jennings D, Jones MK, et al. Real-world outcomes in cystic fibrosis telemedicine clinical care in a time of a global pandemic. Chest 2022 May;161(5):1167-1179 [FREE Full text] [doi: 10.1016/j.chest.2021.11.035] [Medline: 34896356]
69. Richardson CH, Orr NJ, Ollosson SL, Irving SJ, Balfour-Lynn IM, Carr SB. Initiating home spirometry for children during the COVID-19 pandemic - a practical guide. Paediatr Respir Rev 2022 Jun;42:43-48 [FREE Full text] [doi: 10.1016/j.prrv.2021.02.001] [Medline: 33773928]
70. Verkleij M, Georgiopoulos AM, Barendrecht H, Friedman D. Pilot of a therapist-guided digital mental health intervention (eHealth CF-CBT) for adults with cystic fibrosis. Pediatr Pulmonol 2023 Jul;58(7):2094-2103 [FREE Full text] [doi: 10.1002/ppul.26438] [Medline: 37144856]
71. Watanabe AH, Willis C, Ragsdale R, Biskupiak J, Moore K, Brixner D, et al. Patient perspectives on the use of digital technology to help manage cystic fibrosis. Pulm Med 2023;2023:5082499 [FREE Full text] [doi: 10.1155/2023/5082499] [Medline: 36727045]
72. Shanthikumar S, Ruseckaite R, Corda J, Mulrennan S, Ranganathan S, Douglas T. Telehealth use in Australian cystic fibrosis centers: clinician experiences. Pediatr Pulmonol 2023 Oct;58(10):2906-2915 [FREE Full text] [doi: 10.1002/ppul.26612] [Medline: 37477510]
73. Beaufils F, Enaud R, Gallode F, Boucher G, Macey J, Berger P, et al. Adherence, reliability, and variability of home spirometry telemonitoring in cystic fibrosis. Front Pediatr 2023;11:1111088 [FREE Full text] [doi: 10.3389/fped.2023.1111088] [Medline: 36911035]
74. Collaco JM, Albon D, Ostrenga JS, Flume P, Schechter MS, Cromwell EA. Factors associated with receiving CF care and use of telehealth in 2020 among persons with cystic fibrosis in the United States. J Cyst Fibros 2023 May;22(3):456-463 [FREE Full text] [doi: 10.1016/j.jcf.2022.12.004] [Medline: 36526554]
75. Corinna Morlacchi L, Privitera E, Rossetti V, Santambrogio M, Bellofiore A, Rosso L, et al. Telemonitoring: an opportunity in cystic fibrosis lung transplant recipients. Heliyon 2023 Oct;9(10):e19931 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e19931] [Medline: 37818011]
76. Edmondson C, Westrupp N, Short C, Seddon P, Olden C, Wallis C, et al. Unsupervised home spirometry is not equivalent to supervised clinic spirometry in children and young people with cystic fibrosis: results from the CLIMB-CF study. Pediatr Pulmonol 2023 Oct;58(10):2871-2880 [FREE Full text] [doi: 10.1002/ppul.26602] [Medline: 37503909]
77. Stalker HJ, Jonasson AR, Hopfer SM, Collins MS. Improvement in cystic fibrosis newborn screening program outcomes with genetic counseling via telemedicine. Pediatr Pulmonol 2023 Dec;58(12):3478-3486 [FREE Full text] [doi: 10.1002/ppul.26678] [Medline: 37712603]
78. Layton AM, Irwin AM, Mihalik EC, Fleisch E, Keating CL, DiMango EA, et al. Telerehabilitation using fitness application in patients with severe cystic fibrosis awaiting lung transplant: a pilot study. Int J Telemed Appl 2021;2021:6641853 [FREE Full text] [doi: 10.1155/2021/6641853] [Medline: 33727918]
79. Morsa M, Perrin A, David V, Rault G, Le Roux E, Alberti C, et al. Experiences among patients with cystic fibrosis in the mucoexocet study of using connected devices for the management of pulmonary exacerbations: grounded theory qualitative research. JMIR Form Res 2024 Jan 23;8:e38064 [FREE Full text] [doi: 10.2196/38064] [Medline: 38261372]
80. Poulsen M, Holland AE, Button B, Jones AW. Preferences and perspectives regarding telehealth exercise interventions for adults with cystic fibrosis: a qualitative study. Pediatr Pulmonol 2024 May;59(5):1217-1226 [FREE Full text] [doi: 10.1002/ppul.26889] [Medline: 38289142]
81. Rimbaldo K, Frayman KB, Shanthikumar S. The impact of telehealth based care on paediatric cystic fibrosis outcomes. J Cyst Fibros 2023 Jul;22(4):706-709 [FREE Full text] [doi: 10.1016/j.jcf.2023.03.003] [Medline: 36890065]
82. Bell JM, Dwyer TJ, Cunich M, Dentice RL, Hutchings O, Jo HE, et al. Impact of cystic fibrosis multidisciplinary virtual clinics on patient experience, time commitments and costs. Intern Med J 2024 May;54(5):809-816 [FREE Full text] [doi: 10.1111/imj.16258] [Medline: 37886890]
83. Piazza-Waggoner C, Ferguson KS, Daines C, Acton JD, Powers SW. Case study: providing evidence-based behavioral and nutrition treatment to a toddler with cystic fibrosis and multiple food allergies via telehealth. Pediatr Pulmonol 2006 Oct;41(10):1001-1004 [FREE Full text] [doi: 10.1002/ppul.20470] [Medline: 16871593]
84. Wilkinson OM, Duncan-Skingle F, Pryor JA, Hodson ME. A feasibility study of home telemedicine for patients with cystic fibrosis awaiting transplantation. J Telemed Telecare 2008;14(4):182-185 [FREE Full text] [doi: 10.1258/jtt.2008.070107] [Medline: 18534951]

XSL•FO
RenderX

85. Murgia F, Corona B, Bianciardi F, Romano P, Tagliente I, Bella S. The application of telemedicine in the follow-up of lung transplantation in a patient with cystic fibrosis. Clin Ter 2014;165(5):e382-e383 [FREE Full text] [doi: 10.7417/T.2014.1769] [Medline: 25366959]
86. Womack C, Farsin R, Farsad M, Chaudary N. Emerging alternatives to conventional clinic visits in the era of COVID-19: adoption of telehealth at VCU adult cystic fibrosis center. Int J Gen Med 2020 Nov;13:1175-1186 [FREE Full text] [doi: 10.2147/ijgm.s274193] [Medline: 33235489]
87. Ronan P, Mian A, Carr SB, Madge SL, Lorenc A, Robinson N. Learning to breathe with Tai Chi online - qualitative data from a randomized controlled feasibility study of patients with cystic fibrosis. Eur J Integr Med 2020 Dec;40:101229 [FREE Full text] [doi: 10.1016/j.eujim.2020.101229] [Medline: 33106755]
88. Chen JJ, Cooper DM, Haddad F, Sladkey A, Nussbaum E, Radom-Aizik S. Tele-exercise as a promising tool to promote exercise in children with cystic fibrosis. Front Public Health 2018;6:269 [FREE Full text] [doi: 10.3389/fpubh.2018.00269] [Medline: 30324099]
89. Wood J, Mulrennan S, Hill K, Cecins N, Morey S, Jenkins S. Telehealth clinics increase access to care for adults with cystic fibrosis living in rural and remote Western Australia. J Telemed Telecare 2016 Jul 20;23(7):673-679 [FREE Full text] [doi: 10.1177/1357633x16660646]
90. Tomlinson O, Shelley J, Trott J, Bowhay B, Chauhan R, Sheldon C. The feasibility of online video calling to engage patients with cystic fibrosis in exercise training. J Telemed Telecare 2019 Feb 25;26(6):356-364 [FREE Full text] [doi: 10.1177/1357633x19828630]
91. Shakkottai A, Kaciroti N, Kasmikha L, Nasr SZ. Impact of home spirometry on medication adherence among adolescents with cystic fibrosis. Pediatr Pulmonol 2018 Apr;53(4):431-436 [FREE Full text] [doi: 10.1002/ppul.23950] [Medline: 29457700]
92. Gur M, Nir V, Teleshov A, Bar-Yoseph R, Manor E, Diab G, et al. The use of telehealth (text messaging and video communications) in patients with cystic fibrosis: a pilot study. J Telemed Telecare 2016 May 13;23(4):489-493 [FREE Full text] [doi: 10.1177/1357633x16649532]
93. Sturt J, Dliwayo TR, Forjaz V, Hamilton K, Bryce C, Fraser J, et al. Eliciting the impact of digital consulting for young people living with long-term conditions (LYNC study): cognitive interviews to assess the face and content validity of two patient-reported outcome measures. J Med Internet Res 2018 Oct 11;20(10):e268 [FREE Full text] [doi: 10.2196/jmir.9786] [Medline: 30309832]
94. Chrysochoou EA, Hatziagorou E, Kirvassilis F, Tsanakas J. Telephone monitoring and home visits significantly improved the quality of life, treatment adherence and lung function in children with cystic fibrosis. Acta Paediatr 2017 Nov;106(11):1882 [FREE Full text] [doi: 10.1111/apa.13996] [Medline: 28727186]
95. Warda N, Rotolo SM. Virtual medication tours with a pharmacist as part of a cystic fibrosis telehealth visit. J Am Pharm Assoc 2021;61(5):e119-e125 [FREE Full text] [doi: 10.1016/j.japh.2021.04.005] [Medline: 33931355]
96. De Biase RV, Cristiani L, Paglia C, Alghisi F, Giordani B, Lucidi V, et al. Clinical and microbiological monitoring of cystic fibrosis patients, three years of follow-up via tele-medicine: an empirical research. Clin Ter 2020;171(5):e381-e384 [FREE Full text] [doi: 10.7417/CT.2020.2244] [Medline: 32901778]
97. Perkins RC, Davis J, NeSmith A, Bailey J, Powers MR, Chaudary N, et al. Favorable clinician acceptability of telehealth as part of the cystic fibrosis care model during the COVID-19 pandemic. Ann Am Thorac Soc 2021 Sep;18(9):1588-1592 [FREE Full text] [doi: 10.1513/AnnalsATS.202012-1484RL] [Medline: 33636092]
98. Compton M, Soper M, Reilly B, Gettle L, List R, Bailey M, et al. A feasibility study of urgent implementation of cystic fibrosis multidisciplinary telemedicine clinic in the face of COVID-19 pandemic: single-center experience telemedicine. Telemed J E Health 2020 Aug;26(8):978-984 [FREE Full text] [doi: 10.1089/tmj.2020.0091] [Medline: 32357084]
99. Arden MA, Hutchings M, Whelan P, Drabble SJ, Beever D, Bradley JM, et al. Development of an intervention to increase adherence to nebuliser treatment in adults with cystic fibrosis: CFHealthHub. Pilot Feasibility Stud 2021 Jan 04;7(1):1 [FREE Full text] [doi: 10.1186/s40814-020-00739-2] [Medline: 33390191]
100. Nobili RM, Gambazza S, Spada MS, Tutino AL, Bulfamante AM, Mariani A, et al. Remote support by multidisciplinary teams: a crucial means to cope with the psychological impact of the SARS-COV-2 pandemic on patients with cystic fibrosis and inflammatory bowel disease in Lombardia. Int J Clin Pract 2021 Jul;75(7):e14220 [FREE Full text] [doi: 10.1111/ijcp.14220] [Medline: 33848388]
101. Albon D, Van Citters AD, Ong T, Dieni O, Dowd C, Willis A, et al. Telehealth use in cystic fibrosis during COVID-19: association with race, ethnicity, and socioeconomic factors. J Cyst Fibros 2021 Dec;20 Suppl 3:49-54 [FREE Full text] [doi: 10.1016/j.jcf.2021.09.006] [Medline: 34930543]
102. Wickerson L, Helm D, Gottesman C, Rozenberg D, Singer LG, Keshavjee S, et al. Telerehabilitation for lung transplant candidates and recipients during the COVID-19 pandemic: program evaluation. JMIR Mhealth Uhealth 2021 Jun 17;9(6):e28708 [FREE Full text] [doi: 10.2196/28708] [Medline: 34048354]
103. Kayser MZ, Valtin C, Greer M, Karow B, Fuge J, Gottlieb J. Video consultation during the COVID-19 pandemic: a single center's experience with lung transplant recipients. Telemed J E Health 2021 Jul;27(7):807-815 [FREE Full text] [doi: 10.1089/tmj.2020.0170] [Medline: 33054671]

XSL•FO
RenderX

104. Rad EJ, Mirza AA, Chhatwani L, Purington N, Mohabir PK. Cystic fibrosis telemedicine in the era of COVID-19. JAMIA Open 2022 Apr;5(1):ooac005 [FREE Full text] [doi: 10.1093/jamiaopen/ooac005] [Medline: 35224457]
105. Doumit M, Ledwos R, Plush L, Chuang S, Gray M, Jaffe A, et al. Telehealth application of an ultrasonic home spirometer. Arch Dis Child 2022 Aug;107(8):752-754 [FREE Full text] [doi: 10.1136/archdischild-2021-322536] [Medline: 35277380]
106. Fettes E, Riley M, Brotherston S, Doughty C, Griffiths B, Laverty A, et al. "You're on mute!" Does pediatric CF home spirometry require physiologist supervision? Pediatr Pulmonol 2022 Jan;57(1):278-284 [FREE Full text] [doi: 10.1002/ppul.25708] [Medline: 34581507]
107. Ozsezen B, Emiralioglu N, Tural D, Sunman B, Buyuksahin H, Yalcin E, et al. Telephone surveillance during the COVID-19 pandemic: is it a helpful diagnostic tool for detecting acute pulmonary exacerbations in children with chronic lung disease? J Telemed Telecare 2020 Nov 12;28(9):694-702 [FREE Full text] [doi: 10.1177/1357633x20972008]
108. Cox NS, Eldridge B, Rawlings S, Dreger J, Corda J, Hauser J, Youth Activity Unlimited – A Strategic Research Centre of the UK Cystic Fibrosis Trust. Web-based physical activity promotion in young people with CF: a randomised controlled trial. Thorax 2023 Jan;78(1):16-23 [FREE Full text] [doi: 10.1136/thorax-2022-218702] [Medline: 36180067]
109. Bouteleux B, Beaufils F, Fayon M, Bui S. Home-spirometry exacerbation profiles in children with cystic fibrosis. Pediatr Pulmonol 2024 Mar;59(3):552-561 [FREE Full text] [doi: 10.1002/ppul.26781] [Medline: 38014613]
110. Yoon E, Hur S, Curtis LM, Benavente JY, Wolf MS, Serper M. Patient factors associated with telehealth quality and experience among adults with chronic conditions. JAMIA Open 2024 Jul;7(2):ooae026 [FREE Full text] [doi: 10.1093/jamiaopen/ooae026] [Medline: 38596698]
111. Morrison L, McCrea G, Palmer S. Online activity - a beaming good initiative! Delivering alternative exercise opportunities for people with cystic fibrosis. Physiother Theory Pract 2024 Jul;40(7):1609-1615 [FREE Full text] [doi: 10.1080/09593985.2023.2182654] [Medline: 36809231]
112. Federici A, De Marchis M, Alghisi F, Fiocchi AG, Bella S. Telemonitoring for cystic fibrosis patients of Bambino Gesù Children's Hospital during COVID-19. Clin Ter 2022;173(5):440-442 [FREE Full text] [doi: 10.7417/CT.2022.2460] [Medline: 36155730]
113. Kodjebacheva GD, Tang C, Groesbeck F, Walker L, Woodworth J, Schindler-Ruwisch J. Telehealth use in pediatric care during the COVID-19 pandemic: a qualitative study on the perspectives of caregivers. Children (Basel) 2023 Feb 06;10(2):311 [FREE Full text] [doi: 10.3390/children10020311] [Medline: 36832439]
114. O'Hayer CV, O'Loughlin CM, Nurse CN, Smith PJ, Stephen MJ. ACT with CF: a telehealth and in-person feasibility study to address anxiety and depressive symptoms among people with cystic fibrosis. J Cyst Fibros 2021 Jan;20(1):133-139 [FREE Full text] [doi: 10.1016/j.jcf.2020.11.013] [Medline: 33279467]
115. Oates GR, Mims C, Geurs R, Bergquist R, Hager A, Guimbellot JS, et al. Mobile health platform for self-management of pediatric cystic fibrosis: impact on patient-centered care outcomes. J Cyst Fibros 2023 Sep;22(5):823-829 [FREE Full text] [doi: 10.1016/j.jcf.2023.04.009] [Medline: 37085386]
116. List R, Compton M, Soper M, Bruschwein H, Gettle L, Bailey M, et al. Preserving multidisciplinary care model and patient safety during reopening of ambulatory cystic fibrosis clinic for nonurgent care: a hybrid telehealth model. Telemed J E Health 2021 Feb;27(2):193-199 [FREE Full text] [doi: 10.1089/tmj.2020.0247] [Medline: 32924854]
117. Faiçal AVB, Souza EL, Terse-Ramos R. Family perception of a telehealth program for people with cystic fibrosis during the COVID-19 pandemic in northeastern Brazil. J Bras Pneumol 2023 Nov 17;49(5):e20230242 [FREE Full text] [doi: 10.36416/1806-3756/e20230242] [Medline: 37991072]
118. Enochs C, Filbrun AG, Iwanicki C, Moraniec H, Lehrmann J, Stiffler J, et al. Development of an interdisciplinary telehealth care model in a pediatric cystic fibrosis center. Telemed Rep 2021;2(1):224-232 [FREE Full text] [doi: 10.1089/tmr.2021.0021] [Medline: 35720757]
119. Shanthikumar S, Moore E, Corda J, Reardon N, Louey S, Frayman K, et al. Patient and family perspectives regarding the use of telehealth for cystic fibrosis care. Pediatr Pulmonol 2021 May;56(5):811-813 [FREE Full text] [doi: 10.1002/ppul.25262] [Medline: 33433061]
120. Compton M, List R, Starheim E, Somerville L, Williamson L, Murray R, et al. Home spirometry utilisation in telemedicine clinic for cystic fibrosis care during COVID-19 pandemic: a quality improvement process. BMJ Open Qual 2021 Aug;10(3):e001529 [FREE Full text] [doi: 10.1136/bmjoq-2021-001529] [Medline: 34433580]
121. Horrell LN, Hayes S, Herbert LB, MacTurk K, Lawhon L, Valle CG, et al. Telemedicine use and health-related concerns of patients with chronic conditions during COVID-19: survey of members of online health communities. J Med Internet Res 2021 Feb 18;23(2):e23795 [FREE Full text] [doi: 10.2196/23795] [Medline: 33539307]
122. Thee S, Stahl M, Fischer R, Sutharsan S, Ballmann M, Müller A, et al. A multi-centre, randomized, controlled trial on coaching and telemonitoring in patients with cystic fibrosis: conneCT CF. BMC Pulm Med 2021 Apr 21;21(1):131 [FREE Full text] [doi: 10.1186/s12890-021-01500-y] [Medline: 33882893]
123. Hillen B, Simon P, Schlotter S, Nitsche O, Bähner V, Poplawska K, et al. Feasibility and implementation of a personalized, web-based exercise intervention for people with cystic fibrosis for 1 year. BMC Sports Sci Med Rehabil 2021 Aug 19;13(1):95 [FREE Full text] [doi: 10.1186/s13102-021-00323-y] [Medline: 34412703]

XSL•FO
RenderX

124. Woods BM, Bray LA, Campbell SB, Li P, Kazmerski TM, Hovater C, et al. Implementation and evaluation of a fertility preservation telehealth counseling intervention for males with cystic fibrosis. J Cyst Fibros 2024 Jul;23(4):658-663 [FREE Full text] [doi: 10.1016/j.jcf.2024.06.011] [Medline: 38942720]
125. Kilic K, Vardar-Yagli N, Nayir-Buyuksahin H, Guzelkas I, Dogru D, Saglam M, et al. Exercise intolerance, oxidative stress, and irisin in pediatric cystic fibrosis: can telehealth-based exercise training make a difference? Heart Lung 2024;68:145-153 [FREE Full text] [doi: 10.1016/j.hrtlng.2024.06.016] [Medline: 38981171]
126. Bathgate CJ, Smith ED, Murphy NH, Quittner AL, Riekert KA, Goralski JL, et al. Coping and learning to Manage Stress with cystic fibrosis (CALM): a multisite telehealth randomized controlled trial to reduce depression and anxiety symptoms in adults with cystic fibrosis. J Cyst Fibros 2025 Mar;24(2):310-318 [FREE Full text] [doi: 10.1016/j.jcf.2024.11.002] [Medline: 39581781]
127. Langfelder-Schwind E, Basile M, Moyal-Smith R, Polo J, McGinniss MA, Petersen J, et al. Increasing access to genetic counselors with disease-specific expertise: development of a centralized cystic fibrosis genetic counseling telehealth model. Pediatr Pulmonol 2025 Dec;60(12):e71416 [FREE Full text] [doi: 10.1002/ppul.71416] [Medline: 41451894]
128. Ozyemisci Taskiran O, Albayrak H, Kog C, Atli E, Gonullu E, Yantac AE, et al. The effect of a 12-week tele-exercise using immersive virtual reality on functional capacity in adolescents with cystic fibrosis: a randomized controlled, single (assessor) - blind study. Respir Med 2025 Nov;248:108362 [FREE Full text] [doi: 10.1016/j.rmed.2025.108362] [Medline: 40983137]
129. Davis J, Gibson SL, Perkins RC, Greenberg J, Simoneau T, Sawicki GS. Perceptions of telehealth in pediatric pulmonary clinic post-pandemic. Pediatr Pulmonol 2025 Aug;60(8):e71237 [FREE Full text] [doi: 10.1002/ppul.71237] [Medline: 40827727]
130. Medbo J, Imberg H, Hansen C, Krantz C, de Monestrol I, Svedberg M. Telemedicine and home spirometry in cystic fibrosis: a prospective multicenter study. Pediatr Pulmonol 2024 Nov;59(11):2967-2975 [FREE Full text] [doi: 10.1002/ppul.27166] [Medline: 38963304]
131. Davis J, Perkins R, Bailey J, Chaudary N, Garcia B, Froh D, et al. Acceptability of telehealth post-pandemic among clinicians across the United States caring for people with cystic fibrosis. Pediatr Pulmonol 2025 Feb;60(2):e70000 [FREE Full text] [doi: 10.1002/ppul.70000] [Medline: 39960328]
132. Schechter MS, Molzhon A, Everhart RS, Kang L, Weiskittle R, Castleberry B, et al. Impact of UPLIFT, a group telehealth intervention, on symptoms of depression and anxiety in adults with CF. J Cyst Fibros 2025 Mar;24(2):319-325 [FREE Full text] [doi: 10.1016/j.jcf.2024.11.008] [Medline: 39668003]
133. Cox NS, Dal Corso S, Burge AT, Bondarenko J, Perryman J, Holland AE. Remote assessment of exercise capacity in adults with chronic respiratory disease: safety, reliability and acceptability. Chron Respir Dis 2025;22:14799731251318033 [FREE Full text] [doi: 10.1177/14799731251318033] [Medline: 39881539]
134. Vendrusculo FM, da Costa GA, Bagatini MA, Lemes BMHM, Faria CA, de Oliveira LC, et al. Feasibility of performing the 3-minute step test with remote supervision in children and adolescents with cystic fibrosis: a comparative study. Pediatr Investig 2024 Jun;8(2):83-90 [FREE Full text] [doi: 10.1002/ped4.12436] [Medline: 38910852]
135. O'Hayer CV, Smith PJ, Drescher CF, Bruschwein H, Nurse CN, Kushner HM, et al. ACT with CF: a randomized trial of acceptance and commitment therapy vs supportive psychotherapy for adults with cystic fibrosis. Gen Hosp Psychiatry 2024;91:212-222 [FREE Full text] [doi: 10.1016/j.genhosppsych.2024.11.011] [Medline: 39626332]
136. Doumit M, Pacey V, Jaffe A, Gray K. Understanding the acceptability of the changing model of care in cystic fibrosis. Respir Med 2024;234:107847 [FREE Full text] [doi: 10.1016/j.rmed.2024.107847] [Medline: 39490937]
137. Albon D, Ong T, Horton B, Brighton D, Shen S, List R, et al. Cystic fibrosis learning network telehealth innovation lab during the COVID-19 pandemic: impact on access to care, outcomes, and a new CF care model. Pediatr Pulmonol 2025 May;60(5):e71102 [FREE Full text] [doi: 10.1002/ppul.71102] [Medline: 40325945]
138. Reis LMS, Pinheiro AAC, Silva Júnior MA, Gonçalves CP, Santana NN. Impact of telehealth during the COVID-19 pandemic on clinical and nutritional conditions of adolescents with cystic fibrosis. J Bras Pneumol 2024 May 27;50(2):e20230397 [FREE Full text] [doi: 10.36416/1806-3756/e20230397] [Medline: 38808828]
139. De Marchis M, Rivolta M, De Dominicis S, Ciarnella M, Leone P, Milo F, et al. WEB-REHAB Program for people with cystic fibrosis during COVID-19 pandemic: a feasibility study. Minerva Respir Med 2025 Mar;64(1):1-8 [FREE Full text] [doi: 10.23736/s2784-8477.24.02139-9]
140. Douglas TA, Pourghaderi AR, Ahern S, Corda J, Earnest A, Mulrennan S, et al. The impact of telehealth on clinical outcomes in adults and children with cystic fibrosis in Australia. J Cyst Fibros 2026 Mar;25(2):369-374 [FREE Full text] [doi: 10.1016/j.jcf.2025.12.017] [Medline: 41469307]
141. Qin Y, Hamana K, Gale N. Remote exercise services for people with cystic fibrosis: experiences and perceptions from people with cystic fibrosis and members of cystic fibrosis multidisciplinary teams. Disabil Rehabil 2025 Jun;47(13):3353-3366 [FREE Full text] [doi: 10.1080/09638288.2024.2420832] [Medline: 39539020]
142. Snoswell CL, Taylor ML, Caffery LJ. Why telehealth does not always save money for the health system. JHOM 2021 Jul 15;35(6):763-775 [FREE Full text] [doi: 10.1108/jhom-04-2020-0159]
143. Mahtta D, Daher M, Lee MT, Sayani S, Shishehbor M, Virani SS. Promise and perils of telehealth in the current era. Curr Cardiol Rep 2021 Jul 16;23(9):115 [FREE Full text] [doi: 10.1007/s11886-021-01544-w] [Medline: 34269884]

XSL•FO
RenderX

144. Smith MW, O'Shea AMJ, Wray CM. Health care and climate change-telemedicine's role in environmental stewardship. JAMA Netw Open 2023 Jan 03;6(1):e2253794 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.53794] [Medline: 36719685]
145. Umpierre RN, Mattiello R, Schmitz CAA, Falceto de Barros E, da Silva RS, Gonçalves MR, et al. Greening healthcare and slashing carbon emissions through telemedicine: a cross-sectional study from over 50 thousand remote consults at a leading tertiary hospital. Front Digit Health 2025 May 30;7:1497770 [FREE Full text] [doi: 10.3389/fdgth.2025.1497770] [Medline: 40520214]
146. Robeznieks A. Telehealth virtual visits can be a tool to fight climate change. AMA. URL: https://www.ama-assn.org/public-health/environmental-health/telehealth-virtual-visits-can-be-tool-fight-climate-change [accessed 2026-03-26]
147. Ige TO. A systematic review of the impact of telehealth utilization in the United States of America's healthcare system on the environment. J Biol Agric Healthcare 2024 Jan;14(1):55-65 [FREE Full text] [doi: 10.7176/JBAH/14-1-06]
148. Hudgins A, Garber K, Gustin T. Synchronous telehealth fatigue among health care providers. J Nurse Pract 2024 Sep;20(8):105111 [FREE Full text] [doi: 10.1016/j.nurpra.2024.105111]
149. Uscher-Pines L, Fischer SH. Key findings from RAND health care research on telehealth policy. RAND. URL: https://www.rand.org/pubs/research_briefs/RBA1402-1-v4.html [accessed 2026-04-29]
150. Shilane D, Lu TH. Declining trends in telehealth utilization in the ongoing COVID-19 pandemic. J Telemed Telecare 2023 Sep 28;31(4):563-571 [FREE Full text] [doi: 10.1177/1357633x231202284] [Medline: 37769292]
151. Huang J, Yeung AM, Eiland LA, Huang ES, Raymond JK, Klonoff DC. Telehealth fatigue: is it real? What should be done? J Diabetes Sci Technol 2024 Jan;18(1):196-200 [FREE Full text] [doi: 10.1177/19322968221127253] [Medline: 36205155]
152. Hall JL, McGraw D. For telehealth to succeed, privacy and security risks must be identified and addressed. Health Aff (Millwood) 2014 Mar;33(2):216-221 [FREE Full text] [doi: 10.1377/hlthaff.2013.0997] [Medline: 24493763]
153. Triana AJ, Gusdorf RE, Shah KP, Horst SN. Technology literacy as a barrier to telehealth during COVID-19. Telemed J E Health 2020 Sep;26(9):1118-1119 [FREE Full text] [doi: 10.1089/tmj.2020.0155] [Medline: 32429770]
154. Vagg T, Twohig S, Deasy KF, Ibrahim HIS, Plant BJ. A narrative review of digital health literacy within cystic fibrosis telehealth: are we considering it? mHealth 2025 Apr;11:20 [FREE Full text] [doi: 10.21037/mhealth-24-66] [Medline: 40248749]
155. Digital health literacy key to overcoming barriers for health workers, WHO study says. World Health Organization (WHO). 2023 Sep 18. URL: https://www.who.int/europe/news/item/18-09-2023-digital-health-literacy-key-to-overcoming-barriers-for-health-workers--who-study-says [accessed 2026-04-29]
156. Arias López MDP, Ong BA, Borrat Frigola X, Fernández AL, Hicklent RS, Obeles AJT, et al. Digital literacy as a new determinant of health: a scoping review. PLOS Digit Health 2023 Oct 12;2(10):e0000279 [FREE Full text] [doi: 10.1371/journal.pdig.0000279] [Medline: 37824584]
157. Feng H, Kurata K, Cao J, Itsuki K, Niwa M, Aoyama A, et al. Telemedicine research trends in 2001-2022 and research cooperation between China and other countries before and after the COVID-19 pandemic: bibliometric analysis. Interact J Med Res 2024 Aug 30;13:e40801 [FREE Full text] [doi: 10.2196/40801] [Medline: 39213010]
158. Maugeri A, Barchitta M, Basile G, Agodi A. Public and research interest in telemedicine from 2017 to 2022: infodemiology study of Google trends data and bibliometric analysis of scientific literature. J Med Internet Res 2024 May 16;26:e50088 [FREE Full text] [doi: 10.2196/50088] [Medline: 38753427]
159. Jerjes W, Harding D. Telemedicine in the post-COVID era: balancing accessibility, equity, and sustainability in primary healthcare. Front Digit Health 2024 Aug 21;6:1432871 [FREE Full text] [doi: 10.3389/fdgth.2024.1432871] [Medline: 39233772]
160. Barry R, Green E, Robson K, Nott M. Factors critical for the successful delivery of telehealth to rural populations: a descriptive qualitative study. BMC Health Serv Res 2024 Aug 07;24(1):908 [FREE Full text] [doi: 10.1186/s12913-024-11233-3] [Medline: 39113042]
161. Key considerations in operationalizing a telehealth program. Telehealth. URL: https://www.providence.org/-/media/project/psjh/providence/socal/files/telehealth/telehealthwhitepaper.pdf [accessed 2026-03-26]
162. Fahim C, Prashad A, Silveira K, Chandraraj A, Thombs BD, Tonelli M, et al. Dissemination and implementation of clinical practice guidelines: a longitudinal, mixed-methods evaluation of the Canadian Task Force on Preventive Health Care's knowledge translation efforts. CMAJ Open 2023 Aug 08;11(4):E684-E695 [FREE Full text] [doi: 10.9778/cmajo.20220121] [Medline: 37553226]
163. Husain L, Kitchens K, Raja S, Memon M. Bridging implementation gaps in digital health: a translational research imperative for equitable healthcare innovation. Clin Transl Sci 2025 Oct;18(10):e70375 [FREE Full text] [doi: 10.1111/cts.70375] [Medline: 41044865]
164. The first European Digital Health Technology Assessment framework co-created by all stakeholders in the European Health Ecosystem. EDiHTA. 2025. URL: https://edihta-project.eu/ [accessed 2026-02-16]
165. Hao Q, Xu F, Li Y, Evans J. Artificial intelligence tools expand scientists' impact but contract science's focus. Nature 2026 Jan 14;649(8099):1237-1243 [FREE Full text] [doi: 10.1038/s41586-025-09922-y] [Medline: 41535462]

XSL•FO
RenderX

166. Hamilton AB, Brown A, Loeb T, Chin D, Grills C, Cooley-Strickland M, et al. Enhancing patient and organizational readiness for cardiovascular risk reduction among Black and Latinx patients living with HIV: study protocol. Prog Cardiovasc Dis 2020 Mar;63(2):101-108 [FREE Full text] [doi: 10.1016/j.pcad.2020.02.014] [Medline: 32109483]
167. Chambers DA, Emmons KM. Navigating the field of implementation science towards maturity: challenges and opportunities. Implement Sci 2024 Mar 13;19(1):26 [FREE Full text] [doi: 10.1186/s13012-024-01352-0] [Medline: 38481286]
168. Fisher G, Smith CL, Pagano L, Spanos S, Zurynski Y, Braithwaite J. Leveraging implementation science to solve the big problems: a scoping review of health system preparations for the effects of pandemics and climate change. Lancet Planet Health 2025 Apr;9(4):e326-e336 [FREE Full text] [doi: 10.1016/s2542-5196(25)00056-7]
169. Wensing M, Grol R. Knowledge translation in health: how implementation science could contribute more. BMC Med 2019 May 07;17(1):88 [FREE Full text] [doi: 10.1186/s12916-019-1322-9] [Medline: 31064388]

Abbreviations

AI: artificial intelligence
ALS: amyotrophic lateral sclerosis
CF: cystic fibrosis
EHR: electronic health record
GAI: generative artificial intelligence
iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Reporting Literature Searches
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
TBE: theory-based evaluation
THCF: Telehealth for CF Working Group
TIDiER-telehealth: Template for Intervention Description and Replication for Telehealth

Edited by S Brini; submitted 20.Oct.2025; peer-reviewed by S Ajayi, M Chakit; comments to author 29.Dec.2025; accepted 10.Apr.2026; published 22.May.2026.

Please cite as:
Vagg T, Doherty R, Ranganathan SC, Plant BJ
Reporting of Telehealth Implementation in Cystic Fibrosis: Scoping Review Using a Novel Theory-Based Evaluation Lens
J Med Internet Res 2026;28:e86194
URL: https://www.jmir.org/2026/1/e86194
doi:10.2196/86194
PMID:

©Tamara Vagg, Robyn Doherty, Sarath C Ranganathan, Barry J Plant. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Blockchain-Enabled Self-Sovereign Identity Applications in Health Care: Scoping Review

Abha Pokharel[1], MSc; Surya Kathayat[1], PhD

Norwegian University of Science and Technology, Trondheim, Trøndelag, Norway

Corresponding Author:
Abha Pokharel, MSc
Norwegian University of Science and Technology
Gløshaugen, IT-bygget sydfløy, Sem Sælands vei 7
Trondheim, Trøndelag, 7034
Norway
Phone: 47 48619879
Email: abha.pokharel@ntnu.no

Abstract

Background: Self-sovereign identity (SSI) provides a decentralized approach to digital identity management, enabling individuals to control their personal data without reliance on centralized authorities. Blockchain technology offers a tamper-resistant and distributed infrastructure that can support secure and verifiable identity systems. In health care, where identity fragmentation, privacy risks, and interoperability challenges persist, blockchain-enabled SSI (BC-SSI) has been proposed as a potential solution. However, existing research remains heterogeneous, with varying levels of technical maturity and limited evidence of real-world deployment.

Objective: This study conducts a scoping review to systematically map BC-SSI applications in health care and to analyze their application domains, development stages, study aims, targeted challenges, and technological infrastructures. In addition, this study aims to identify structural gaps in current research and assess the readiness of BC-SSI systems for clinical deployment.

Methods: This review followed the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) methodology. A comprehensive literature search conducted between September 2024 and August 2025 identified 37 peer-reviewed studies that met predefined inclusion criteria. Data were extracted and synthesized using descriptive and thematic analyses across application areas, system maturity, technological components, and reported challenges.

Results: The findings indicate that BC-SSI research in health care remains at an early stage of maturity, with most studies proposing conceptual models or prototype implementations and limited real-world validation. Applications predominantly focus on identity verification, credential management, and privacy-preserving data exchange across domains such as electronic health records, mobile health, and access control systems. Commonly used technologies include decentralized identifiers, verifiable credentials, smart contracts, and privacy-enhancing mechanisms such as zero-knowledge proofs and selective disclosure. Despite rapid technical development, persistent challenges include interoperability limitations, governance gaps, usability concerns, and insufficient integration with health care infrastructures. Notably, a structural gap was identified between technological capability and system-level readiness for clinical deployment.

Conclusions: BC-SSI technologies demonstrate potential for enabling secure, interoperable, and patient-centric identity management in health care. However, current research is predominantly technology-driven and lacks sufficient system-level validation. This study highlights the need for integrated architectural approaches, governance frameworks, and real-world evaluation to bridge the gap between conceptual innovation and clinical implementation. Advancing BC-SSI toward health care adoption will require coordinated progress across technical, organizational, and regulatory dimensions.

(J Med Internet Res 2026;28:e89574) doi:10.2196/89574

KEYWORDS

blockchain; decentralization; health care; identity management; privacy; security; self-sovereign identity

XSL•FO
RenderX

Introduction

Overview

Blockchain-enabled self-sovereign identity (SSI) applications in health care are attracting greater research attention as health data becomes increasingly digital and the need for secure, patient-focused identity management grows. Early blockchain projects in health care primarily focused on maintaining record immutability and enabling secure data exchange. More recently, researchers have explored the use of SSI to give patients greater control over their digital identities and health information [1-3]. The move from centralized to decentralized identity systems is driven by ongoing problems in current health care systems, such as limited interoperability, privacy risks, and fragmented access control [4-6]. These issues are especially evident in situations such as referrals, emergency care, and telehealth, where rapid and reliable identity verification is needed for effective clinical coordination [7].

The need for stronger digital identity systems is clear from the rise in health care data breaches. Growing regulatory demands, such as those under the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA), also underscore the need for better solutions [1,8,9]. In this context, BC-SSI systems ensure this by enabling decentralized identity checks, secure credential management, and private data sharing among health care providers. Still, there are trade-offs between decentralization, scalability, governance, and system performance. These challenges are even more complex in regulated health care settings [10,11].

Even though more research is being conducted on BC-SSI in health care, the studies remain scattered. Most studies focus on concepts or early prototypes [1,12,13]. Few studies bring together architectural designs, technology choices, platforms, and governance to assess whether these systems are ready for real-world health care use. However, these studies have rarely examined how BC-SSI fits with clinical systems, regulations, and daily operations. As a result, there is an incomplete understanding of how practical these systems are in real-world health care settings. To fill this gap, this study makes 4 main contributions. First, it reviews peer-reviewed literature to map out BC-SSI applications in health care. Second, it looks at their technology, development stages, and use. Third, it brings together common architectural features. Fourth, it creates a layered architectural model and assesses the readiness of current BC-SSI systems for real-world health care and clinical use.

Background

SSI and Blockchain

SSI is a decentralized identity model that enables individuals to create, manage, and selectively disclose digital credentials without reliance on centralized identity providers [7,13]. In SSI systems, identity attributes are not stored within institutional databases but are instead held by individuals as verifiable credentials (VCs), which can be presented when required. These systems operate within a trust framework involving issuers (eg, health care institutions), holders (eg, patients), and verifiers (eg, service providers), supported by governance mechanisms such as credential standards, trust registries, and revocation processes. In health care environments, traditional identity systems remain largely centralized and fragmented, requiring repeated identity verification across multiple organizations. This results in inefficiencies, data duplication, and increased risks of identity mismatch and unauthorized access, increasing vulnerability to data breaches. These limitations have driven interest in decentralized identity models such as SSI for health care applications.

Blockchain technology provides the underlying infrastructure that supports SSI through immutable and verifiable record keeping [10]. Within BC-SSI architectures, decentralized identifiers (DIDs) represent user-controlled identities, while VCs enable cryptographic validation of identity attributes without direct reliance on issuing authorities. Blockchain is typically used to anchor DID registries, credential schemas, and revocation mechanisms, while sensitive health data remains off chain to preserve privacy. This combination enables secure and tamper-evident identity verification across distributed health care systems. To operationalize these capabilities, BC-SSI architectures incorporate a set of supporting mechanisms. These include selective disclosure for controlled data sharing, zero-knowledge proofs (ZKPs) for privacy-preserving verification, and smart contracts for automated enforcement of access control and audit policies. Together, these mechanisms enable decentralized identity management while addressing key requirements of trust, security, and accountability in health care contexts.

In studies of health care identity systems, security and privacy are closely related yet distinct concerns. Security focuses on ensuring confidentiality, integrity, and availability of identity and health data through mechanisms such as authentication and access control. Privacy emphasizes minimizing unnecessary disclosure of sensitive information and enabling individuals to control how their data are shared [11,12]. These concerns are particularly critical in health care due to the sensitivity of medical data and the complexity of cross-organizational data exchange [1].

SSI Frameworks and Blockchain Network Types

SSI frameworks define the architectural components required for credential issuance, storage, verification, and trust management. Implementations commonly use platforms such as Hyperledger Indy and Aries to support decentralized identity registries and secure agent-based communication [7,12]. SSI architectures may operate on different blockchain network configurations depending on governance and trust requirements. Public (permissionless) networks such as Bitcoin and Ethereum enable unrestricted participation and transparent validation. Consortium networks restrict participation to authorized entities and are often deployed in multi-institutional collaborations. Private networks are centrally governed and typically implemented within a single organization. These models reflect varying trade-offs between decentralization, scalability, and governance control relevant to identity system design.

XSL•FO
RenderX

Smart Contracts

Smart contracts are programmable components deployed on blockchain platforms that automate identity-related operations. In health care contexts, they can enforce access control policies, manage consent conditions, and record auditable transaction logs without manual intervention [13]. This supports transparent and accountable identity management across distributed health care systems.

Despite increasing research activity in BC-SSI, existing studies remain heterogeneous in terms of application focus, architectural design, and implementation maturity. Many proposed solutions are limited to conceptual models or prototype implementations, with minimal evaluation in real-world health care environments. Furthermore, there is limited consolidated analysis of how BC-SSI mechanisms address health care-specific challenges such as interoperability, governance, and clinical workflow integration.

To address this gap, this study conducts a scoping review to systematically map BC-SSI applications in health care, focusing on application domains, technological components, development stages, and the challenges these systems aim to address.

Methods

Study Design

We used the scoping review methodology to map existing proposals that integrate BC-SSI into health care. A scoping review is particularly appropriate because BC-SSI research is still an emerging area, conceptually diverse, and lacking consolidated empirical evidence. The review followed the framework of Arksey and O'Malley [14], as enhanced by Levac et al [15], and adhered to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) reporting guidelines [16]. The completed PRISMA-ScR checklist is provided in Multimedia Appendix 1. The methodology includes five stages: (1) defining research questions (RQs), (2) identifying relevant studies, (3) selecting eligible publications, (4) charting extracted data, and (5) synthesizing and reporting findings. The following 5 RQs guided the review:

- RQ1: What BC-SSI application areas and data types are addressed in health care?
- RQ2: What development stages (proposal, architecture, and prototype) are used in BC-SSI studies?
- RQ3: What are the main aims of BC-SSI applications in health care?
- RQ4: What challenges do BC-SSI technologies aim to address?
- RQ5: What BC-SSI technologies, frameworks, and infrastructures are used?

Search Strategy and Study Selection

A comprehensive search was conducted across 6 databases, along with their distribution as illustrated in Table 1. Google Scholar was included to capture interdisciplinary publications that may not be indexed in domain-specific databases. The final search update was performed on August 15, 2025. Search strings combined free-text terms related to "self-sovereign identity," "blockchain," and "healthcare" using Boolean operators ("OR" within concept groups and "AND" between concept groups). Synonym groups (eg, health, eHealth, medical, hospital, and clinical) were expanded accordingly. Backward and forward snowballing were conducted to identify additional relevant studies [17]. Full search strings are provided in Multimedia Appendix 2. A total of 345 records were initially retrieved. The distribution of retrieved records by database is presented in Table 1.

Table 1. Captured works of literature by source (N=345).

Source	Captured works, n
Scopus	95
Web of Science	46
Google Scholar	133
Embase	19
Medline	13
IEEE Xplore	39

Eligibility Criteria and Study Selection

Peer-reviewed studies published between 2015 and 2025 were considered eligible. Inclusion criteria included BC-SSI proposals addressing health domain problems (conceptual model, framework, proposal, prototype, experiment, implementation, and pilot), technical descriptions (eg, architecture, prototype, and framework), English-language peer-reviewed primary studies, and studies published between 2015 and 2025. Exclusion criteria included studies unrelated to health care, blockchain-only solutions without SSI components, nonprimary studies (eg, reviews and editorials) or abstract-only publications, and non-English publications.

All references were imported into Zotero (Corporation for Digital Scholarship) for deduplication and screened using Rayyan [18]. Two reviewers independently screened titles, abstracts, and full texts according to predefined criteria. Disagreements were resolved through discussion and consensus to ensure consistent interpretation.

After duplicate removal and eligibility screening, 33 studies were retained. A total of 4 additional studies were identified

XSL•FO
RenderX

through snowballing, resulting in 37 studies included in the final synthesis.

Data Extraction and Charting

Data were charted using a predefined extraction matrix consistent with PRISMA-ScR guidance. The first author performed data extraction, and the second author reviewed all entries. Discrepancies were resolved through discussion, and all final decisions were agreed jointly. The following data were extracted: general information, including author(s) ID and study; publication type, including journal, conference, proceedings, or book chapter; application area (eg, health records, mobile health [mHealth], wearable/embedded systems, health care services); data types handled (eg, medical data, credentials, vaccination records, and location data); and development stage, including proposal, architectural design, and prototype/experimental implementation.

Extracted variables were mapped directly to the RQs: application areas and data types informed RQ1; development-stage classification addressed RQ2; thematic coding of study aims addressed RQ3; reported challenges informed RQ4; and BC-SSI–enabling technologies, blockchain platforms supporting SSI, blockchain types, smart contracts, and storage models were analyzed under RQ5.

Following Petersen et al [19], classification categories for application areas, data types, and development stages were iteratively developed and refined during the charting process. Categories were merged or adjusted as necessary to ensure conceptual consistency and reduce overlap across studies.

Data Synthesis and Reporting

After screening and data extraction, findings were synthesized descriptively and thematically, aligned with the 5 RQs (RQ1-RQ5). Descriptive statistics were used to summarize distributions across application areas, data types, development stages, and technological components. Thematic categorization was applied to study aims and reported challenges to identify recurring patterns across BC-SSI implementations. The complete list of included studies and their extracted characteristics is presented in the Results section.

Results

Overview of Included Studies

A total of 37 studies met the inclusion criteria and were included in the final synthesis. Figure 1 presents the PRISMA-ScR flow diagram summarizing the study selection process. Table 2 presents the key characteristics of the included studies, including publication type, application area, data types addressed, and development stage.

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of study selection.

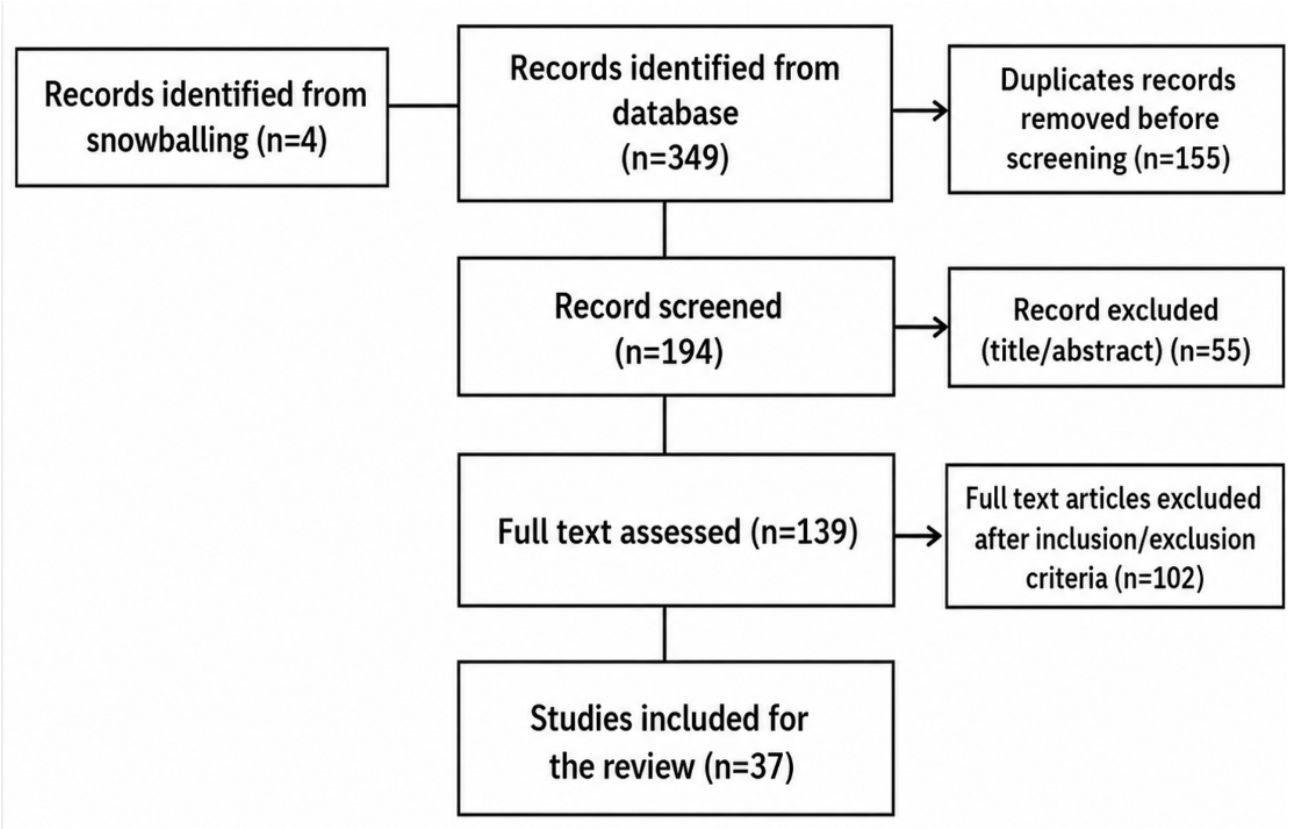

XSL•FO
RenderX

Table 2. Included research papers in the scoping review and their key characteristics.

ID	Study (type)	Application area	Data types	Development stage
1	Nasrin [20] (conference)	Health records	Vaccination and credential	Proposal
2	Hak et al [21] (conference)	Embedded and wearable systems	Sensitive and credential	Architecture
3	Fotopoulos et al [22] (conference)	Embedded and wearable systems	Credential	Proposal
4	Zou et al [23] (journal)	Embedded and wearable systems	Credential	Prototype/experimental
5	Bai et al [24] (journal)	Embedded and wearable systems	Medical and credential	Prototype/experimental
6	Bandara et al [25] (conference)	mHealth[a]	Location and sensitive	Prototype/experimental
7	Kormiltsyn et al [26] (conference)	Health services	Medical	Architecture
8	Harrell et al [12] (journal)	mHealth	Sensitive	Architecture
9	George and Chacko [27] (journal)	mHealth	Sensitive and credential	Prototype/experimental
10	Sahi et al [28] (conference)	Health records	Credential and sensitive	Proposal
11	Kim et al [29] (conference)	mHealth	Credential	Architecture
12	Abid et al [30] (journal)	mHealth	Sensitive	Prototype/experimental
13	Popa et al [31] (journal)	Embedded and wearable systems	Sensitive and credential	Prototype/experimental
14	Saha et al [32] (conference)	Health records	Credential	Proposal
15	Rafid et al [33] (conference)	mHealth	Credential and medical	Proposal
16	Saragih et al [4] (conference)	Health records	Sensitive and credential	Proposal
17	Santos et al [9] (journal)	Health records	Medical and credential	Prototype/experimental
18	Song et al [34] (journal)	Embedded and wearable systems	Sensitive and credential	Prototype/experimental
19	Keil et al [35] (conference)	Embedded and wearable systems	Credential	Architecture
20	Freytsis et al [36] (journal)	Access control	Credential	Architecture
21	Zhuang et al [37] (journal)	mHealth	Sensitive and credential	Architecture
22	Al Muharif and Ahmed [38] (conference)	Health records	Medical and credential	Prototype/experimental
23	Pujari et al [39] (conference)	Health records	Credential and sensitive	Architecture
24	Tcholakian et al [8] (journal)	Health records	Medical	Architecture
25	Sami et al [6] (conference)	Health records	Medical	Proposal
26	Wang et al [40] (journal)	Health services	Credential	Prototype/experimental
27	Kang et al [41] (journal)	Health services	Location and credential	Architecture
28	Pujari et al [42] (journal)	Clinical research	Credential	Prototype/experimental
29	Manoj et al [43] (journal)	Health services	Credential	Prototype/experimental
30	Abubakar et al [44] (conference)	Health records	Credential	Prototype/experimental
31	Mchale et al [45] (conference)	Health records	Vaccination	Prototype/experimental
32	Guerar et al [46] (journal)	Health records	Credential and medical	Proposal
33	de Oliveira et al [47] (conference)	mHealth	Credential and medical	Prototype/experimental
34	Thirasak et al [48] (journal)	Health records	Medical	Prototype/experimental
35	Makina et al [1] (conference)	Health records	Medical	Architecture
36	Boi et al [49] (journal)	Health records	Credential and medical	Prototype/experimental
37	Saidi et al [50] (journal)	Health records	Medical	Prototype/experimental

[a]mHealth: mobile health.

Descriptive Characteristics of Included Studies

Across the 37 included studies, 49% (18/37) were conference publications and 51% (19/37) were journal articles. Google Scholar contributed the largest share of the initially retrieved records (133/345, 39%), whereas Embase contributed the smallest share (19/345, 6%). Among the final set of included publications, Scopus accounted for 65% (24/37) and IEEE

XSL•FO
RenderX

Xplore for 19% (7/37). Publication activity peaked in 2022 for both conference papers (6/18, 33%) and journal papers (5/19, 26%), reflecting growing interest in BC-SSI solutions in health care.

RQ1: BC-SSI Application Areas and Data Types

The review identified 6 major BC-SSI application areas in health care. Health records constituted the largest category (16/37, 43%), encompassing applications related to personal health data management, medical record exchange, and credential verification. mHealth applications represented the second-largest category (8/37, 22%), reflecting the growing use of decentralized identity mechanisms in mHealth platforms and patient-facing applications.

Embedded and wearable systems accounted for 19% (7/37) of the studies and included applications involving the Internet of Medical Things and smart medical devices. These systems often require secure device authentication and identity verification in distributed health care environments. The remaining application areas included health care services (4/37, 11%), clinical studies (1/37, 3%), and facility access control (1/37, 3%).

The studies also addressed several types of health care–related data. Credential data were the most frequent type (27/37, 73%), highlighting the central role of identity verification and authorization in BC-SSI architectures. Medical and other sensitive personal data were commonly associated with health record management and wearable device applications. Location data appeared primarily in mHealth and contact-tracing scenarios, in which decentralized identity mechanisms were used to support privacy-preserving monitoring and verification.

RQ2: Development Stages of BC-SSI Applications

The reviewed studies demonstrate varying levels of development maturity in BC-SSI health care applications. Across 37 studies, 22% (8/37) presented conceptual proposals, 30% (11/37) provided architectural designs, and 49% (18/37) provided prototype or experimental implementations. Although more than one-half of the studies reported prototype-level implementations, most evaluations were limited to simulated environments or small-scale user testing, indicating early-stage technological validation. Only a small number reported real-world deployment or evaluation.

RQ3: Aims of BC-SSI Studies

Thematic analysis of the reviewed studies indicates that BC-SSI research is primarily driven by 3 major objectives. Privacy preservation was the most frequently reported aim, accounting for approximately 38% (14/37) of the studies, followed by secure data management (7/37, 19%) and patient-centric identity control (7/37, 19%). The distribution of these aims is summarized in Table 3.

XSL•FO
RenderX

Table 3. Mapping of included studies to research objectives (research question 3 [RQ3]) and corresponding health care challenges (research question 4 [RQ4]).

ID	Study aim (RQ3)	Challenges (RQ4)
1	Vaccination verification and authentication	I[a], DI[b], DP[c], S[d]
2	Anonymous cross-domain authentication	I, P[e], E[f], DI, DP (IoMT[g])
3	Device authentication and revocation (IoMT)	DI, I, SC[h], E, S
4	Privacy-aware authentication and data exchange	I, DI, S, SC, E
5	Decentralized access control for health data	AC[i], SC, E, P, A[j]
6	Privacy-aware contact tracing	P, SC, I, T[k], DP
7	Secure authentication for medical data exchange	DM[l], AC, S
8	Patient-centric data management and sharing	IDM[m], I, AC, A, DP
9	Credential issuance and access control	AC, I, A, P, S, DP
10	Consent-based access control	IDM, CM[n], I, E, P, S
11	Identity tracking and transparency	IDM, T, DP, A
12	Privacy-preserving certificate verification	DI, DP, A, P
13	Consent-based identity sharing	IDM, I, A, P, S
14	Decentralized credential verification	IDM, P, S, E, DP, A
15	Identity management (rare diseases)	IDM, S, E
16	Cross-institution authentication	DM, S, P, I
17	Identity verification for service access	DM, S, I, P
18	Identity management for users and devices	IDM, A, P, S, DI
19	Attribute-based credential verification	DM, A, P, S, I
20	Anti-profiling access control	A, AC, P
21	Identity provisioning for newborns	P, S
22	Patient-controlled identity and data	P, S, I, A, E
23	Patient-centric credential preservation	DM, P, A
24	Secure EHR[o] exchange	DI, P, S, A
25	Fine-grained access control	AC, P, S, E
26	Credential validation and access control	IDM, P, S, A
27	Unlinkable contact tracing	P, A, E
28	Identity verification for research access	AC, IDM, I, DP, P
29	Secure identity validation	IDM, DP, S, P, A
30	Identity management and authentication	S, P, CM
31	Fraud-resistant certificate verification	S, T, DP, A, DM
32	Identity recovery mechanisms	IDM, S, A
33	Patient-centric data preservation	IDM, P, I, A
34	Decentralized data access control	P, S, AC, I
35	Transparent access control	P, S, T
36	Scalable EHR access control	S, E, SC
37	Privacy-aware access control	AC, P, S, A

[a]I: interoperability.

[b]DI: data integrity.

[c]DP: data provenance.

XSL•FO
RenderX

[d]S: security.
[e]P: privacy.
[f]E: efficiency.
[g]IoMT: Internet of Medical Things.
[h]SC: scalability.
[i]AC: access control.
[j]A: autonomy.
[k]T: transparency.
[l]DM: data management.
[m]IDM: identity management.
[n]CM: consent management.
[o]EHR: electronic health record.

Privacy-related aims focus on enabling controlled disclosure of identity and health-related attributes during verification and data exchange processes. Secure data management aims emphasize strengthening authentication, credential validation, and integrity assurance within health care systems. Patient-centric identity control aims are centered on enabling individuals to manage their digital identities, credentials, and consent in interactions with health care providers.

Less frequently, studies addressed objectives such as secure electronic health record (EHR) exchange, prevention of identity traceability, rare-disease identity management, and research identity verification. Overall, these findings indicate that BC-SSI research is primarily oriented toward enhancing privacy and identity assurance, with comparatively less emphasis on system-level integration and interoperability.

To explicitly link the study aim (RQ3) to the underlying health care challenges (RQ4), Table 3 provides a structured mapping of each included study to its primary aims and the corresponding challenges it addressed.

RQ4: Privacy and Security Challenges Addressed by BC-SSI

Privacy (27/37, 73%) and security (26/37, 70%) were the most frequently addressed challenges across the reviewed studies. These reflect persistent limitations in traditional health care identity and data management systems.

Privacy-related challenges primarily involve overdisclosure of sensitive patient information, limited user control over data sharing, and the risk of identity linkability across health care providers. For example, vaccination and health certification systems [30,45] address the need for attribute-level verification without exposing full identity records through selective disclosure and the use of VCs. Similarly, decentralized access control frameworks [39,40,50] emphasize patient-controlled identity and consent management, while contact tracing and IoMT scenarios [25,41] focus on reducing identity traceability using pseudonymous identifiers and privacy-preserving verification mechanisms.

Security-related challenges center on weak authentication, unauthorized data access, and risks of credential forgery and tampering in distributed health care environments. Cross-domain authentication schemes [21-23] address identity verification weaknesses across health care systems, whereas blockchain-based access control and audit mechanisms [48,50] enhance accountability and authorization. In addition, vaccination systems and fraud-resilient health care platforms [30,45,47] mitigate credential forgery through VCs and immutable blockchain registries.

Overall, the findings indicate that BC-SSI research is primarily driven by confidentiality, integrity, and identity assurance requirements, whereas challenges related to scalability and operational performance remain comparatively less explored.

RQ5: BC-SSI Technologies and Infrastructures

DIDs, VCs, and Digital Wallets

Across the reviewed literature, DIDs serve as identity anchors for patients, providers, and devices, enabling portable, cryptographically verifiable claims across institutional boundaries, as illustrated in Figure 2. VCs encode health care attributes, including vaccination status, professional qualifications, and access permissions. Selective disclosure mechanisms were widely used to support privacy-preserving verification.

XSL•FO
RenderX

Figure 2. Blockchain-Enabled Self-Sovereign Identity (BC-SSI) technologies applied in health care. DID: decentralized identifier; VC: verifiable credential.

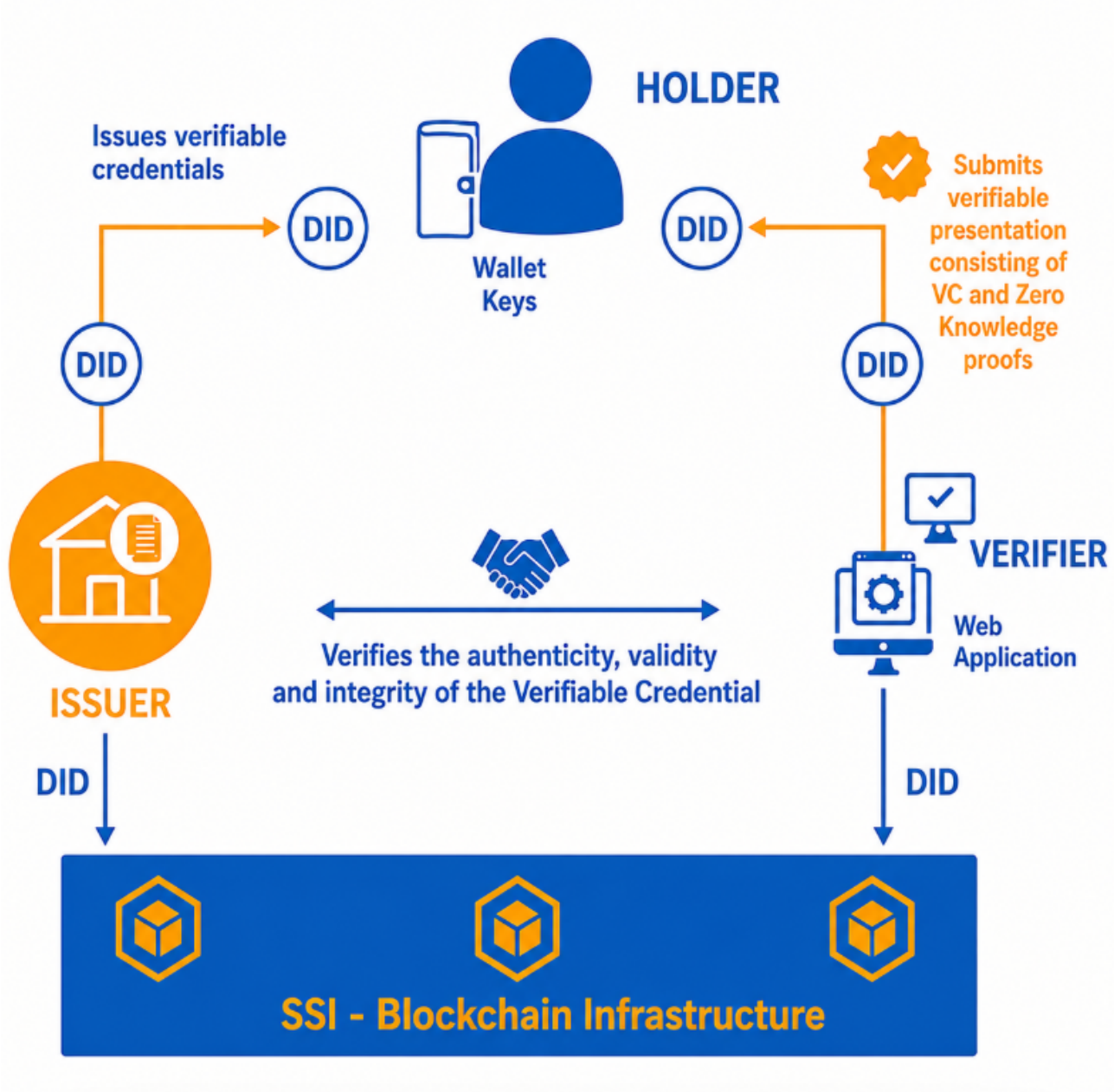

Digital wallets were used to manage keys, credentials, and DID-based connections. Both custodial (server-managed) and noncustodial (client-side) models were observed, reflecting trade-offs between usability and security control.

Blockchain Platforms Supporting SSI

As illustrated in Table 4 , Hyperledger Indy, often combined with Hyperledger Aries, was the most frequently reported SSI-supporting platform (10/37, 27%). Hyperledger Fabric (8/37, 22%) was commonly selected for permissioned consortium settings. Ethereum-based solutions (6/37, 16%) leveraged smart-contract programmability but raised privacy and cost concerns. Notably, 32% (12/37) of studies did not explicitly specify the BC platform, limiting reproducibility and comparative assessment.

XSL•FO
RenderX

Table 4. Blockchain platforms and characteristics of blockchain-enabled self-sovereign identity (BC-SSI) implementations.

BC platform	Studies	Key characteristics
Hyperledger Indy (often with Hyperledger Aries)	[3], [5], [8], [9], [21], [25], [28], [29], [30]	Self-sovereign identity–oriented identity management, decentralized identifiers, verifiable credentials, and permissioned identity ecosystems
Hyperledger Fabric	[1], [49], [10], [11], [13], [15], [29]	Permissioned consortium blockchain, modular architecture, enterprise access control, and privacy-preserving transactions
Hyperledger Besu	[16]	Ethereum-compatible permissioned blockchain supporting enterprise deployment
Ethereum blockchain	[7], [8], [10], [19], [31]	Smart contract programmability, decentralized application support, and public blockchain infrastructure
Not specified (N/A)	[14], [17], [18], [20], [22], [23], [50], [26], [27], [2], [32]	Platform details not explicitly reported

Blockchain Types, Smart Contracts, and Storage Models

Private or permissioned blockchains were used in approximately 51% (19/37) of the reviewed studies. Public blockchain and consortium models were reported less frequently, whereas 43% (16/37) of studies did not explicitly specify the blockchain type. Smart contracts were implemented in 43% (16/37) of studies, primarily for consent management, credential issuance, revocation, and access control.

To reduce on-chain exposure of sensitive health data, 16% (6/37) of studies adopted a hybrid storage architecture. In these approaches, blockchain was used to store identifiers, hashes, or audit logs, whereas medical records were maintained off chain using distributed storage systems such as Interplanetary File System or institutional databases. Privacy-enhancing cryptographic techniques were reported in 35% (13/37) of studies, including ZKPs and selective-disclosure mechanisms that support attribute verification without revealing full data sets.

Comparative Evaluation of BC-SSI Platforms

Table 5 summarizes the comparative characteristics of BC-SSI platforms identified in the reviewed studies. Hyperledger Indy was most frequently used in identity-centric implementations and provides native support for DIDs and VCs. Hyperledger Fabric was commonly selected in permissioned consortium settings but does not natively implement World Wide Web Consortium DID/VC standards. Ethereum-based solutions provide programmability for smart contracts and public-chain deployment models. Hyperledger Besu supports both permissioned and public configurations and is Ethereum compatible. uPort, built on Ethereum, provides decentralized identity functionality through wallet-based implementations. A notable proportion of studies did not explicitly specify the underlying blockchain platform or SSI framework.

Table 5. Comparative evaluation of blockchain-enabled self-sovereign identity (BC-SSI) platforms used in the literature.

Framework	DID[a]	VC[b]	Chain	Smart contract	Health care suitability	Key limitations
Hyperledger Indy	Strong	Strong	Permissioned	Limited	High (privacy-first and identity-centric)	No general-purpose smart contracts
Hyperledger Fabric	Indirect	Limited	Permissioned	Yes	High (enterprise-grade and consortium use)	No native DID/VC support
Ethereum	Strong	Strong	Public	Yes	Medium (high programmability)	Gas fees and public ledger visibility
Hyperledger Besu	Strong	Strong	Permissioned or public	Yes	Medium-high	Deployment and performance complexity
uPort	Strong	Strong	Public Ethereum	Yes	Medium	Dependent on Ethereum scalability
Not specified or custom models	Varies	Varies	Mixed	Varies	Not assessable	Low reproducibility and unclear stack

[a]DID: decentralized identifier.

[b]VC: verifiable credential.

Taken together, the technologies, infrastructures, and governance elements identified across the reviewed studies suggest that BC-SSI health care systems consist of several interacting technical and operational components. These recurring elements are synthesized into a conceptual layered architecture for BC-SSI deployment in health care contexts.

Discussion

Reflection on Principal Findings

This scoping review identified 37 peer-reviewed studies investigating BC-SSI applications in health care, revealing a rapidly growing yet structurally fragmented research landscape and indicating uneven development across technical and

XSL•FO
RenderX

operational dimensions. Although publication activity has increased substantially since 2020, most contributions remain conceptual models, architectural proposals, or prototype implementations evaluated in controlled environments. Large-scale deployments, cross-institutional pilots, and longitudinal evaluations remain rare. This demonstrates that current BC-SSI research remains largely disconnected from real-world health care implementation. This review addresses the identified gap by providing a system-level synthesis of BC-SSI architectures and their readiness for health care deployment.

Across the reviewed literature, BC-SSI systems primarily focus on identity verification, credential management, consent-based access control, and privacy-preserving data exchange. These objectives correspond to longstanding challenges in health care identity management, including fragmented identity infrastructures and centralized trust models. However, a clear translational gap emerges between theoretical benefits and practical deployment, indicating that the proposed solutions have not yet been validated in operational health care settings. While many studies emphasize improved privacy, autonomy, and secure credential exchange, empirical evidence demonstrating integration with health care workflows, regulatory compliance, and cross-organizational interoperability remains limited, indicating insufficient evidence for real-world applicability and system-level integration.

Importantly, technical feasibility does not necessarily translate into operational readiness. Although several studies demonstrate working credential issuance and verification workflows, most systems have not evaluated within real clinical environments or integrated with EHR infrastructures. As a result, the readiness of BC-SSI systems for operational health care deployment remains uncertain, highlighting a lack of clinical validation and feasibility.

Security and Privacy Mechanism in Operational Context

A recurring theme in the reviewed studies is the use of cryptographic mechanisms, including DIDs, VCs, selective disclosure, ZKPs, and smart contracts, to enhance privacy and security. However, these mechanisms are often discussed in the abstract without sufficiently examining their operational implications for health care systems, limiting understanding of their behavior in real-world health care settings. In a real health care setting, selective disclosure enables patients to demonstrate specific attributes, such as vaccination status and professional authorization, without revealing full identity records. ZKPs can support verification while minimizing unnecessary data exposure, potentially reducing the risk of large-scale data breaches. Smart contracts may automate consent enforcement and audit logging, theoretically enhancing traceability and accountability. Yet, these technical advantages introduce new risk vectors, demonstrating that enhanced privacy mechanisms may also increase system complexity and operational risk. Wallet compromise, key loss, credential revocation complexity, cross-institutional correlation attacks, and emergency override requirements represent unresolved operational challenges. Few studies systematically evaluate how these mechanisms perform under health care-specific constraints such as high patient turnover, emergency access demands, and strict compliance frameworks. Notably, to our knowledge, no studies have systematically evaluated these risks in emergency or high-pressure clinical scenarios, which represents a critical gap in the literature.

While the BC-SSI architecture may mitigate risks associated with centralized identity repositories, it also redistributes responsibility between users and institutions, requiring careful governance and usability design. This trade-off demonstrates that security mechanisms must be evaluated not only for technical robustness but also for their impact on patient safety and clinical workflow reliability.

Platform Selection and Architectural Trade-Offs

The comparative evaluation of the blockchain platforms reveals significant architectural divergences. Hyperledger Indy, often combined with Aries, dominates identity-centric implementations because of its native support for DIDs and VCs. Permissioned platforms such as Hyperledger Fabric and Besu offer enterprise-grade governance and controlled participation models, aligning more naturally with regulated health care environments. Public Ethereum-based solutions provide programmability and a mature smart contract ecosystem but raise concerns about transaction transparency, scalability, and regulatory suitability. These choices are not merely technical preferences but reflect underlying governance assumptions, trust models, and regulatory constraints.

In addition, nearly one-third of the studies failed to clearly specify their blockchain infrastructure, thereby limiting reproducibility and technical assessment. This lack of architectural transparency constitutes a methodological weakness in the current literature and limits the ability to assess the suitability of proposed solutions for real-world health care deployment. This demonstrates that platform selection is a critical architectural decision that directly shapes privacy guarantees, interoperability potential, and regulatory alignment.

Stakeholder Implications and Systemic Constraints

The potential of BC-SSI cannot be evaluated solely through a technical lens. Its feasibility depends on stakeholder alignment across patients, providers, regulators, and system developers.

For patients, SSI promises granular control over identity attributes and consent. However, user-facing challenges, including wallet management, credential recovery, cognitive load, and emergency override procedures, are rarely evaluated through formal usability studies. For example, in acute care scenarios, delayed identity recovery or inaccessible credentials could directly affect continuity of care. The limited exploration of recovery frameworks and emergency access models represents a critical gap between conceptual autonomy and practical safety.

For health care providers, integration with established clinical workflows remains largely unaddressed. Limited studies demonstrated integration with existing EHR or hospital information systems, or comprehensive alignment with standards such as Health Level Seven (HL7) and Fast Healthcare

XSL•FO
RenderX

Interoperability Resources (FHIR). Without such integration, BC-SSI remains peripheral to routine care delivery, limiting its practical impact on health care operations.

For regulators and policymakers, the distinction between technical privacy preservation and demonstrable legal compliance is crucial. Although the BC-SSI architecture conceptually aligns with GDPR and HIPAA principles, few studies have evaluated audit mechanisms, credential revocation governance, issuer accreditation models, or the legal enforceability of VCs. Governance design, particularly schema ownership, trust registries, and cross-jurisdictional interoperability, remains underdeveloped. This demonstrates that governance is a primary barrier to adoption rather than a secondary design consideration.

Taken together, these findings indicate that BC-SSI maturity is constrained less by cryptographic capabilities than by governance, interoperability, and human-centered implementation challenges.

Conceptual BC-SSI Health Care Architecture

This section presents a novel contribution of this review by moving beyond descriptive analysis toward an integrated architectural perspective. The synthesis of the reviewed studies indicates that BC-SSI implementations in health care can be understood as a multilayered architecture integrating identity infrastructure, blockchain platforms, privacy-preserving mechanisms, health care system interoperability, and governance structures. While many studies emphasize decentralized identity technologies and cryptographic mechanisms, these components are often addressed in isolation rather than as integrated systems.

Based on the recurring elements identified across the literature, Figure 3 illustrates a conceptual BC-SSI health care architecture that synthesizes these patterns into a unified system-level perspective.

Figure 3. Conceptual multilayered blockchain-enabled self-sovereign identity (BC-SSI) health care architecture. FHIR: Fast Healthcare Interoperability Resources; GDPR: General Data Protection Regulation; HIPAA: Health Insurance Portability and Accountability Act; HL7: Health Level Seven; IoT: internet of things.

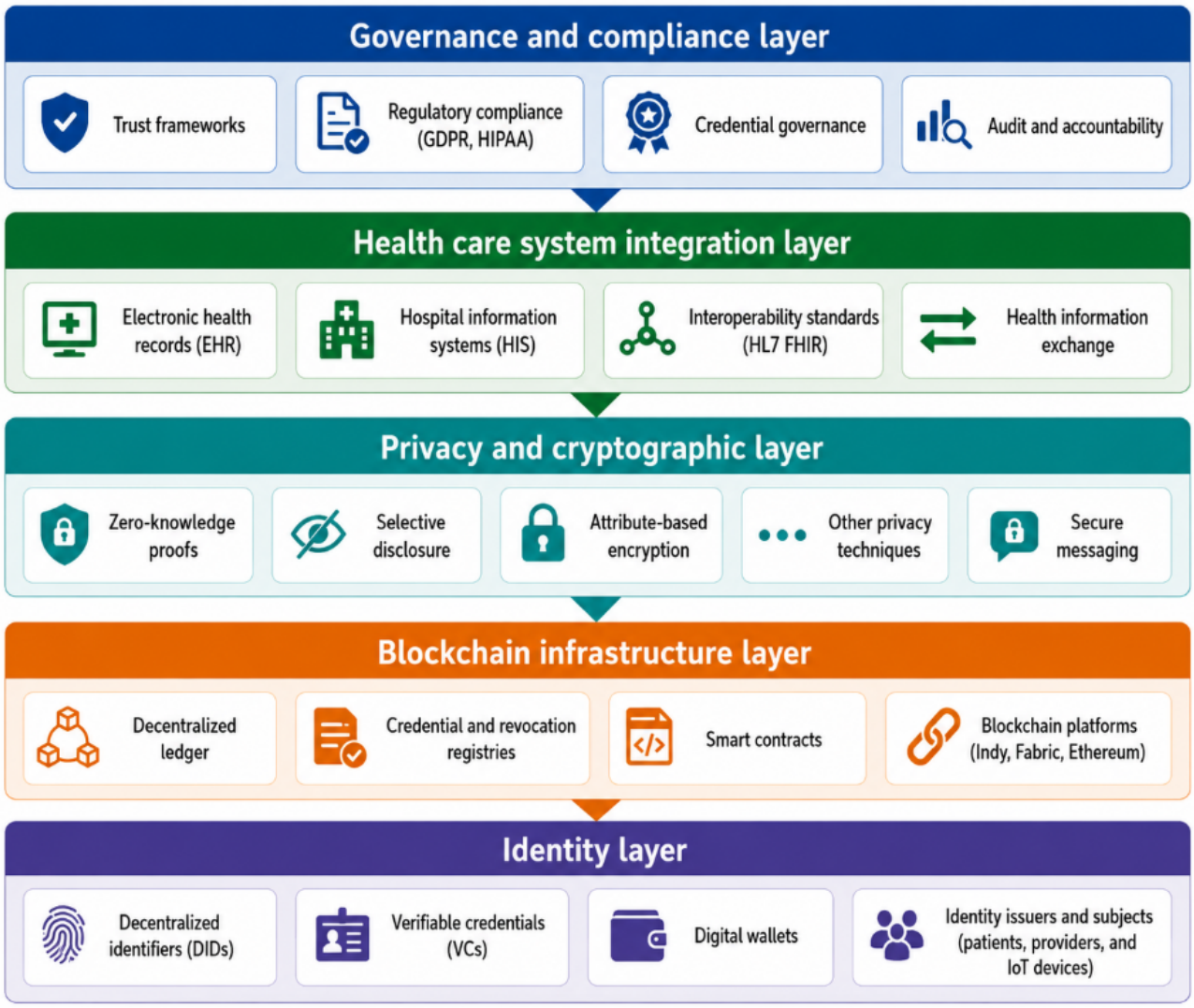

At the foundational level, the identity layer includes DIDs, VCs, and digital wallets that enable patients, providers, and health care devices to manage decentralized digital identities. These identities are supported by the blockchain infrastructure layer, which maintains credential schemas, revocation registries, and

XSL•FO
RenderX

verification records using distributed-ledger platforms such as Hyperledger Indy, Fabric, and Ethereum-based systems.

A privacy and cryptographic layer enables secure verification through mechanisms such as selective disclosure, ZKPs, and cryptographic signatures, allowing health care attributes to be validated without exposing unnecessary personal data. These capabilities interact with the health care system integration layer, in which BC-SSI identity verification interfaces with EHR, hospital information systems, and health care interoperability standards such as HL7 and FHIR.

Finally, a governance and compliance layer provides the regulatory and organizational framework necessary for health care adoption. This layer encompasses credential governance policies, issuer accreditation, trust registries, and compliance with regulatory frameworks such as GDPR and HIPAA.

This layered perspective highlights that current research lacks a holistic architectural approach, limiting its ability to support end-to-end health care identity systems.

From Prototype Innovation to Clinical Readiness

When interpreted through the conceptual architecture illustrated in Figure 3, the findings reveal a structural imbalance in current BC-SSI health care research, with a disproportionate focus on technical components rather than system-level integration. Most reviewed studies concentrate on the lower layers of the architecture, particularly the identity and blockchain infrastructure layers, in which DIDs, VCs, and distributed ledger technologies have been widely implemented and experimentally validated.

In contrast, the upper layers of architecture remain comparatively underdeveloped. Integration with clinical infrastructures, including EHR systems and interoperability standards such as HL7 and FHIR, is rarely demonstrated in existing prototypes. Similarly, governance mechanisms, including credential trust registries, issuer accreditation models, revocation governance, and regulatory alignment with frameworks such as GDPR and HIPAA, are frequently discussed conceptually but seldom implemented or evaluated in operational environments. This demonstrates that current research does not adequately address the requirements for real-world health care deployment.

These observations suggest that the maturity of BC-SSI in health care is constrained less by cryptographic capability than by challenges related to governance design, interoperability alignment, and human-centered implementation. While technical innovation in decentralized identity and credential architectures is advancing rapidly, real-world health care deployment remains limited due to insufficient evaluation of cross-institution workflows, credential lifecycle management, emergency access mechanisms, and usability of patient-facing identity tools.

Rather than evaluating BC-SSI systems solely in terms of functional capabilities, this review demonstrates that clinical readiness depends on system-level integration, governance alignment, and real-world validation. This indicates that the primary barrier to BC-SSI adoption is not technological feasibility but the absence of coordinated validation across technical, organizational, and regulatory dimensions.

Synthesis Across Research Questions

Taken together, the findings of this review provide a coherent response to the RQs. RQ1 and RQ2 are addressed by demonstrating that BC-SSI applications are concentrated in identity-centric use cases and remain largely at the conceptual or prototype stage, with limited progression toward real-world deployment. RQ3 is addressed by identifying the dominant study aims, particularly privacy preservation, secure data exchange, and patient-centric identity control. RQ4 is addressed by revealing that the challenges targeted by these studies, primarily privacy, security, and interoperability, are approached predominantly through technical mechanisms, with limited consideration of operational, governance, and integration constraints.

The discussion further extends these findings by demonstrating that the key limitation of current BC-SSI research is not a lack of technical capability but the absence of system-level validation, governance alignment, and integration with health care infrastructures. This synthesis reinforces the central gap identified in this study and highlights the conditions necessary for advancing BC-SSI toward clinical readiness.

Limitations

This scoping review has several limitations. First, it included only peer-reviewed English-language publications from 2015 to 2025. Relevant gray literature, industrial white papers, pilot deployments, or non-English studies may therefore not be represented. Given the rapidly evolving nature of blockchain and SSI ecosystems, some practical implementations may exist outside academic reporting.

Second, consistent with scoping review methodology, this study did not perform a formal quality appraisal or risk-of-bias assessment of included studies. The objective was to map the breadth and characteristics of BC-SSI research rather than evaluate intervention effectiveness. As a result, findings should be interpreted as a structured synthesis of existing proposals and implementations rather than a quantitative assessment of performance or security robustness.

Third, heterogeneity in terminology across BC-SSI research may have influenced retrieval. Although comprehensive Boolean search strategies and snowballing were used, emerging terminology or alternative phrasing may have led to the exclusion of the relevant studies.

Despite these limitations, adherence to the Arksey and O'Malley framework and PRISMA-ScR guidelines enhances transparency and reproducibility, providing a systematic overview of current BC-SSI health care research. The conceptual architecture presented in this study should therefore be interpreted as a synthesis derived from the reviewed literature rather than as an empirically validated system design.

Implications and Future Research Directions

The findings of this review have important implications for both research and practice. First, it demonstrates that advancing

XSL•FO
RenderX

BC-SSI in health care requires a shift from technology-centric development toward system-level design and validation. While current studies successfully demonstrate cryptographic feasibility, their limited integration with health care workflows, governance frameworks, and regulatory requirements constrains their practical applicability. This suggests that future BC-SSI research must adopt a sociotechnical perspective that considers not only technical performance but also usability, interoperability, and organizational alignment.

Second, the proposed conceptual architecture highlights the need for coordinated development across multiple layers, including identity infrastructure, blockchain platforms, privacy-preserving mechanisms, health care system integration, and governance frameworks. This layered perspective provides a structured foundation for designing and evaluating BC-SSI systems in health care and may serve as a reference model for future implementations.

Future research should prioritize multi-institutional pilot deployments that evaluate BC-SSI systems under realistic health care conditions. This includes assessing interoperability across health care providers, integration with EHR systems, and alignment with established standards such as HL7 and FHIR. Longitudinal evaluations are particularly important for assessing system scalability, performance, and usability over time.

Governance development represents a critical research frontier. Formal trust frameworks, credential schema governance models, revocation mechanisms, and issuer accreditation processes must be systematically designed and validated. In addition, regulatory mapping studies are required to explicitly align decentralized identity operations with frameworks such as GDPR and HIPAA, moving beyond conceptual compliance toward demonstrable accountability.

Human-centered evaluation is equally essential. Future studies should examine usability challenges related to identity wallets, credential recovery, and emergency access mechanisms to ensure that patient autonomy does not compromise safety or continuity of care. Security assessments should also consider real-world threat scenarios, including wallet compromise, key loss, correlation attacks, and cross-institutional vulnerabilities.

Collectively, these directions highlight that the future of BC-SSI in health care depends not only on advancing cryptographic techniques but also on achieving integration, governance maturity, and clinical validation.

Conclusion

This scoping review systematically analyzed 37 peer-reviewed studies investigating BC-SSI applications in health care. The findings indicate that, while research activity has increased significantly in recent years, the field remains characterized by conceptual and prototype-driven development, with limited evidence of real-world deployment or clinical validation. The review demonstrates that BC-SSI research is primarily focused on identity-centric use cases, including credential verification, authentication, and privacy-preserving data exchange. These applications directly address longstanding challenges in health care identity management, particularly those related to privacy, security, and fragmented trust models. However, a clear gap persists between technical capability and operational readiness.

By synthesizing findings across studies, this review shows that the primary limitation of current BC-SSI research is not technological feasibility but the lack of integration with health care systems, governance frameworks, and regulatory requirements. This study therefore advances current literature by bridging the gap between fragmented technical proposals and system-level understanding required for health care implementation. This gap is further highlighted through the proposed conceptual multilayered architecture, which reveals a structural imbalance between well-developed technical components and underdeveloped system-level elements such as interoperability, governance, and clinical integration.

In addressing the RQs, this study provides a comprehensive mapping of BC-SSI applications (RQ1), identifies their developmental maturity (RQ2), analyzes their primary aims (RQ3), and examines the challenges they seek to address (RQ4). Beyond descriptive synthesis, this review contributes an analytical perspective by reframing BC-SSI maturity in terms of translational readiness, emphasizing the importance of system-level validation and real-world applicability.

BC-SSI technologies hold significant potential to enhance patient-centric identity management and secure data exchange in health care. However, their successful adoption will depend on coordinated progress across technical, organizational, and regulatory domains. Future research must therefore move beyond isolated technical innovation toward integrated, governance-aware, and clinically validated identity systems. Only through such efforts can BC-SSI transition from conceptual promise to a reliable foundation for health care identity infrastructures.

Funding

This study received no external funding.

Data Availability

All data analyzed in this study were derived from published peer-reviewed literature. The extracted data supporting the findings of this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

AP conceptualized the study, designed the methodology, conducted the literature search and screening, performed data extraction and analysis, created the visualizations, and drafted the original manuscript. SK contributed to the study design, critically reviewed and edited the manuscript, and provided supervision and methodological guidance. Both authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR checklist.
[PDF File (Adobe PDF File), 248 KB - jmir_v28i1e89574_app1.pdf]

Multimedia Appendix 2
Full search strings.
[DOCX File , 14 KB - jmir_v28i1e89574_app2.docx]

References

1. Makina H, Letaifa AB, Rachedi A. A blockchain and self-sovereign identity-based platform for electronic health records management under GDPR. : IEEE; 2024 Presented at: Proceedings of the 2024 20th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob); October 21-23, 2024; Paris p. 118-123. [doi: 10.1109/wimob61911.2024.10770361]
2. Honório da Silva ML, Velasco G, Pimentel VN, Martins M, Ribeiro GSP, Carvalho ST. Blockchain and self-sovereign identity: a healthcare use case. 2025 Presented at: Proceedings of the Workshop em Blockchain: Teoria, Tecnologias e Aplicações (WBlockchain); May 19-22, 2025; Natal, Rio Grande do Norte p. 154-167. [doi: 10.5753/wblockchain.2025.9506]
3. Yue X, Wang H, Jin D, Li M, Jiang W. Healthcare data gateways: found healthcare intelligence on blockchain with novel privacy risk control. J Med Syst 2016;40(10):218. [doi: 10.1007/s10916-016-0574-6] [Medline: 27565509]
4. Saragih TK, Tanuwijaya E, Wang G. The use of blockchain for digital identity management in healthcare. : IEEE; 2022 Presented at: Proceedings of the 2022 10th International Conference on Cyber and IT Service Management (CITSM); September 20-21, 2022; Yogyakarta p. 1-6. [doi: 10.1109/citsm56380.2022.9935935]
5. Al Mamun A, Azam S, Gritti C. Blockchain-based electronic health records management: a comprehensive review and future research direction. IEEE Access 2022;10:5768-5789. [doi: 10.1109/access.2022.3141079]
6. Sami KT, Toorani M. Blockchain-based access control for electronic health records. In: Abie H, Gkioulos V, Katsikas S, Pirbhulal S, editors. Secure and Resilient Digital Transformation of Healthcare. Cham: Springer; 2024:1884.
7. Siqueira A, da Conceição AF, Rocha V. Blockchains and self-sovereign identities applied to healthcare solutions: a systematic review. ArXiv Preprint posted online on April 26, 2021. [doi: 10.48550/arXiv.2104.12298]
8. Tcholakian M, Gorna K, Laurent M, Kaffel Ben Ayed H, Naghmouchi M. Self-sovereign identity for consented and content-based access to medical records using blockchain. Security and Communication Networks 2023;2023:6025789. [doi: 10.1155/2023/6025789]
9. de R dos Santos Y, de Oliveira NR, Barbosa GNN, Reis LHA, Mendes ACR, de Oliveira MT, et al. Decentralized security in blockchain-based digital health systems: self-sovereign identity, access control, and auditing with smart contracts. Cluster Comput 2025;28(15):940. [doi: 10.1007/s10586-025-05669-3]
10. Martínez AL, Naghmouchi M, Laurent M. Empower healthcare through a self-sovereign identity infrastructure for secure electronic health data access. ArXiv Preprint posted online on January 21, 2025. [doi: 10.48550/arXiv.2501.12229]
11. Torongo AA, Toorani M. Blockchain-based decentralized identity management for healthcare systems. ArXiv Preprint posted online on July 31, 2023. [doi: 10.48550/arXiv.2307.16239]
12. Harrell DT, Usman M, Hanson L, Abdul-Moheeth M, Desai I, Shriram J. Technical design and development of a self-sovereign identity management platform for patient-centric health care using blockchain technology. Blockchain Healthc Today 2022;5 [FREE Full text] [doi: 10.30953/bhty.v5.196] [Medline: 36779027]
13. Kumar D. Blockchain technology for healthcare and self-sovereign identity. In: Blockchain Technology for Cyber Defense, Cybersecurity, and Countermeasures. Boca Raton: CRC Press; 2025:69-93.
14. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005;8(1):19-32. [doi: 10.1080/1364557032000119616]
15. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
16. Tricco AC, Zarin W, Ghassemi M, Nincic V, Lillie E, Page MJ. Same family, different species: methodological conduct and quality varies according to purpose for five types of knowledge synthesis. J Clin Epidemiol 2018;96:133-142. [doi: 10.1016/j.jclinepi.2017.10.014] [Medline: 29103958]

XSL•FO
RenderX

17. Levy Y, Ellis TJ. A systems approach to conduct an effective literature review in support of information systems research. InformingSciJ 2006;9:181-212. [doi: 10.28945/479]
18. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
19. Petersen K, Vakkalanka S, Kuzniarz L. Guidelines for conducting systematic mapping studies in software engineering: an update. Inf Softw Technol 2015;64:1-18. [doi: 10.1016/j.infsof.2015.03.007]
20. Nasrin S. Securing vaccination data using self-sovereign identity, hyperledger fabric and zero trust model. : IEEE; 2023 Presented at: 2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD); September 21-23, 2023; Dhaka p. 290-294. [doi: 10.1109/icict4sd59951.2023.10303620]
21. Hak L, Fugkeaw S, Sato H. LSAC: a lightweight, scalable, and anonymous cross-domain authentication scheme in IoMT. IEEE: IEEE; 2024 Presented at: 2024 IEEE Intelligent Mobile Computing (MobileCloud); April 22-24, 2024; Tokyo p. 8-15. [doi: 10.1109/mobilecloud62079.2024.00009]
22. Fotopoulos F, Malamas V, Dasaklis TK, Kotzanikolaou P, Douligeris C. A blockchain-enabled architecture for IoMT device authentication. : IEEE; 2020 Presented at: 2020 IEEE Eurasia Conference on IoT, Communication and Engineering (ECICE); October 23-25, 2020; Yunlin p. 89-92. [doi: 10.1109/ecice50847.2020.9301913]
23. Zou S, Cao Q, Huangqi C, Huang A, Li Y, Wang C, et al. A physician's privacy-preserving authentication and key agreement protocol based on decentralized identity for medical data sharing in IoMT. IEEE Internet of Things J 2024;11(17):29174-29189. [doi: 10.1109/jiot.2024.3406561]
24. Bai P, Kumar S, Aggarwal G, Mahmud M, Kaiwartya O, Lloret J. Self-sovereignty identity management model for smart healthcare system. Sensors (Basel) 2022;22(13):4714 [FREE Full text] [doi: 10.3390/s22134714] [Medline: 35808211]
25. Bandara E, Liang X, Foytik P, Shetty S. Connect: blockchain and self-sovereign identity empowered contact tracing platform. In: Wireless Mobile Communication and Healthcare. Cham: Springer; 2021:208-223.
26. Kormiltsyn A, Norta A, Nisar S, Dwivedi V. Preventing data-security breaches and patient-safety risks in cross-blockchain e-healthcare systems. In: Chbeir R, Benslimane D, Zervakis M, Manolopoulos Y, Nguyen NT, Tekli J, editors. Management of Digital Ecosystems. Cham: Springer; 2024.
27. George M, Chacko AM. Health passport: a blockchain-based PHR-integrated self-sovereign identity system. Front Blockchain 2023;6:1075083. [doi: 10.3389/fbloc.2023.1075083]
28. Sahi N, Liang A, Van DW, Oikonomou K, Zhang P. Self-sovereign identity in semi-permissioned blockchain networks leveraging ethereum and hyperledger fabric. : IEEE; 2023 Presented at: Proceedings of the 2023 IEEE International Conference on Digital Health (ICDH); July 2-8, 2023; Chicago p. 315-321. [doi: 10.1109/icdh60066.2023.00053]
29. Kim J, Choi M, Lee C, Woo JWK, Hong J. Service applicable blockchain-based self-sovereign identity management system. : IEEE; 2023 Presented at: Proceedings of the 2023 IEEE International Conference on Blockchain and Cryptocurrency (ICBC); May 15-18, 2023; Dubai p. 1-5. [doi: 10.1109/icbc56567.2023.10174875]
30. Abid A, Cheikhrouhou S, Kallel S, Jmaiel M. NovidChain: blockchain-based privacy-preserving platform for COVID-19 test/vaccine certificates. Softw Pract Exp 2022;52(4):841-867 [FREE Full text] [doi: 10.1002/spe.2983] [Medline: 34226768]
31. Popa M, Stoklossa SM, Mazumdar S. ChainDiscipline - towards a blockchain-IoT-based self-sovereign identity management framework. IEEE Trans Serv Comput 2023;16(5):3238-3251. [doi: 10.1109/tsc.2023.3279871]
32. Saha S, Nova SN, Iqbal MI. Healthcare professionals credential verification model using blockchain-based self-sovereign identity. In: Proceedings of the Fourth International Conference on Trends in Computational and Cognitive Engineering. Singapore: Springer; 2022:381-392.
33. Rafid FA, Benissan E, Murr L, Ermakov I, Oikonomou K, Zhang P. A decentralized identity system for accelerating medical communications within rare disease communities. : IEEE; 2022 Presented at: Proceedings of the 2022 IEEE International Conference on Blockchain, Smart Healthcare and Emerging Technologies (SmartBlock4Health); November 3-5, 2022; New York p. 1-8. [doi: 10.1109/smartblock4health56071.2022.10034646]
34. Song X, Xu G, Huang Y, Dong J. DID-HVC-based web3 healthcare data security and privacy protection scheme. Future Gener Comput Syst 2024;158:267-276. [doi: 10.1016/j.future.2024.04.015]
35. Keil KR, Bochnia R, Gudymenko I, Köpsell S, Anke J. Gaining back the control over identity attributes: access management systems based on self-sovereign identity. 2024 Presented at: Proceedings of Open Identity Summit; May 6-7, 2024; Aachen p. 61-72. [doi: 10.18420/OID2024_05]
36. Freytsis M, Barclay I, Radha SK, Czajka A, Siwo GH, Taylor I, et al. Development of a mobile, self-sovereign identity approach for facility birth registration in Kenya. Front Blockchain 2021;4:631341. [doi: 10.3389/fbloc.2021.631341]
37. Zhuang Y, Shyu CR, Hong S, Li P, Zhang L. Self-sovereign identity empowered non-fungible patient tokenization for health information exchange using blockchain technology. Comput Biol Med 2023;157:106778 [FREE Full text] [doi: 10.1016/j.compbiomed.2023.106778] [Medline: 36934533]
38. Al Muharif MHA, Ahmed S. Towards a pervasive paradigm: Enabling patients privacy through decentralized identity and data control in e-health. : IEEE; 2024 Presented at: Proceedings of the 2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT). IEEE; April 25-26, 2024; Gharuan p. 750-754. [doi: 10.1109/incacct61598.2024.10551158]

XSL•FO
RenderX

39. Pujari C, Muniyal B, Chandrakala CB, Rajarajan M. A user-centric self-sovereign identity-based authentication framework for decentralized data management in healthcare settings. : IEEE; 2023 Presented at: International Conference on Recent Advances in Information Technology for Sustainable Development (ICRAIS); December 6-7, 2023; Bangalore p. 89-94. [doi: 10.1109/icrais59684.2023.10367127]
40. Wang D, Chen X, Zhang L, Fang Y, Huang C. A blockchain-based human-to-infrastructure contact tracing approach for COVID-19. IEEE Internet of Things J 2022;9(14):12836-12847. [doi: 10.1109/jiot.2021.3138971]
41. Kang Y, Park YB. Secure access control realization based on self-sovereign identity for cloud CDM. Applied Sciences 2022;12(19):9833. [doi: 10.3390/app12199833]
42. Pujari C, Muniyal B, Chandrakala CB, Rao A, Sadiname V, Rajarajan M. Identity resilience in the digital health ecosystem: a key recovery-enabled framework. Comput Biol Med 2023;167:107702 [FREE Full text] [doi: 10.1016/j.compbiomed.2023.107702] [Medline: 37976822]
43. Manoj T, Makkithaya K, VG N. A blockchain based decentralized identifiers for entity authentication in electronic health records. Cogent Engineering 2022;9(1). [doi: 10.1080/23311916.2022.2035134]
44. Abubakar M, McCarron P, Jaroucheh Z, Al DA, Buchanan B. Blockchain-based platform for secure sharing and validation of vaccination certificates. 2021 Presented at: 14th International Conference on Security of Information and Networks (SIN); December 15-17, 2021; Piscataway p. 1-8. [doi: 10.1109/sin54109.2021.9699221]
45. Mchale S, Murr L, Zhang P. Using decentralized identifiers and interPlanetary file system to create a recoverable rare disease patient identity framework. : Association for Computing Machinery; 2023 Presented at: 2023 7th International Conference on Medical and Health Informatics (ICMHI); May 12-14, 2023; New York p. 142-149. [doi: 10.1145/3608298.3608325]
46. Guerar M, Migliardi M, Russo E, Khadraoui D, Merlo A. SSI-MedRx: a fraud-resilient healthcare system based on blockchain and SSI. Blockchain Res Appl 2025;6(1):100242. [doi: 10.1016/j.bcra.2024.100242]
47. de Oliveira NR, dos Santos YDR, Barbosa GNN, Reis LHA, Mendes ACR, Tuler de Oliveira M. Distributed data security in digital health: Self-sovereign identity, access control, and blockchain-based log records. In: IEEE; 2024 Presented at: 6th International Conference on Blockchain Computing and Applications (BCCA); November 27-29, 2024; Abu Dhabi p. 558-565. [doi: 10.1109/bcca62388.2024.10844453]
48. Thirasak K, Chainarong D, Chuaphanngam T, Fugkeaw S. SSX-EHRs: secure and scalable cross-domain EHRs sharing with blockchain sharding and dynamic proxy re-encryption. EURASIP J on Info Security 2025;2025(1):15. [doi: 10.1186/s13635-025-00200-y]
49. Boi B, Cirillo F, De Santis M, Esposito C. Soulbound tokens: enabler for privacy-aware and decentralized authentication mechanism in medical data storage. Blockchain Healthc Today 2024;7. [doi: 10.30953/bhty.v7.334] [Medline: 39649415]
50. Saidi H, Labraoui N, Ari AAA, Maglaras LA, Emati JHM. DSMAC: privacy-aware decentralized self-management of data access control based on blockchain for health data. IEEE Access 2022;10:101011-101028. [doi: 10.1109/access.2022.3207803]

Abbreviations

BC-SSI: blockchain-enabled SSI
DID: decentralized identifier
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
GDPR: General Data Protection Regulation
HIPAA: Health Insurance Portability and Accountability Act
HL7: Health Level Seven
mHealth: mobile health
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
RQ: research question
SSI: self-sovereign identity
VC: verifiable credential
ZKP: zero-knowledge proof

XSL•FO
RenderX

Edited by Y Zhuang; submitted 16.Dec.2025; peer-reviewed by M Krishnapatnam, Z Zandesh, A Akli; comments to author 28.Feb.2026; revised version received 01.Apr.2026; accepted 24.Apr.2026; published 26.May.2026.

Please cite as:
Pokharel A, Kathayat S
Blockchain-Enabled Self-Sovereign Identity Applications in Health Care: Scoping Review
J Med Internet Res 2026;28:e89574
URL: https://www.jmir.org/2026/1/e89574
doi:10.2196/89574
PMID:

©Abha Pokharel, Surya Kathayat. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effectiveness of Digital Health Interventions to Improve Self-Care in Patients With Chronic Diseases: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Jessica Longhini[1], PhD; Daniel Pedrotti[2], MSN; Federica Foladori[2], MSN; Melania Stedile[2], MSN; Francesca Stefani[2], MSN; Michela Dal Ben[2], MSN; Alessandro Froner[2], MSN; Marta Proietti Pesci[2], MSN; Stefano Toccoli[2], MSN; Anna Brugnolli[2,3], Prof Dr, MSN

[1]Laboratory of Studies and Evidence Based Nursing, Universita degli Studi di Padova Dipartimento di Scienze Cardio-Toraco-Vascolari e Sanita pubblica, University of Padua, Padova, Veneto, Italy

[2]Trento Campus of Health Sciences, Azienda Sanitaria Universitaria Integrata del Trentino, Trento, Italy

[3]Interdepartmental Medical Science Centre, University of Trento, Trento, Italy

Corresponding Author:
Jessica Longhini, PhD
Laboratory of Studies and Evidence Based Nursing
Universita degli Studi di Padova Dipartimento di Scienze Cardio-Toraco-Vascolari e Sanita pubblica
University of Padua
Via Leonardo Loredan, 18
Padova, Veneto, 35131
Italy
Phone: 39 3405322936
Email: jessica.longhini@unipd.it

Abstract

Background: Chronic diseases account for most global morbidity and mortality, increasing the need for effective long-term self-care support. Digital health interventions, such as mobile apps, telemonitoring, and connected devices, are increasingly used to promote self-care; yet, their overall effectiveness across chronic conditions remains unclear.

Objective: This systematic review and meta-analysis evaluated whether digital health interventions improve self-care in adults with chronic diseases.

Methods: We searched PubMed, CINAHL, Scopus, and PsycINFO for randomized controlled trials (RCTs; January 1, 2013, to December 31, 2025) that assessed digital health interventions targeting self-care outcomes, as measured with validated instruments, in patients with chronic conditions. Standardized mean differences (SMDs) were pooled using random-effects models, while results not suitable for meta-analysis were synthesized narratively. Risk of bias was assessed with the Cochrane Risk of Bias 2.0 tool for RCTs and certainty of evidence with Grading of Recommendations Assessment, Development and Evaluation.

Results: A total of 55 RCTs involving 5889 participants were included. Most interventions were multicomponent and mainly based on mobile or web-based applications, telemonitoring, connected devices, and text-messaging support. In diabetes, pooled analyses showed little to no clear improvement across self-care domains measured with the Summary of Diabetes Self-Care Activities, including general diet (3 studies), specific diet (3 studies), exercise (5 studies), foot care (5 studies), and glucose monitoring (4 studies), with low to very low certainty of evidence. In heart failure, digital interventions probably improved self-care monitoring measured with the Self-Care of Heart Failure Index (5 studies, 364 participants; SMD=0.49, 95% CI 0.13-0.85; low certainty), whereas effects on self-care maintenance (5 studies) and on self-care measured with the European Heart Failure Self-Care Behaviour Scale (3 studies) were not clearly demonstrated. In other chronic conditions, narrative synthesis suggested possible benefits in some cardiovascular conditions, chronic hepatitis B, epilepsy, and hypertension, while no significant effects were found in chronic obstructive pulmonary disease and multimorbidity, and mixed findings emerged in Parkinson disease. Across 17 studies, medication adherence showed little to no overall improvement (SMD=0.06, 95% CI –0.31 to 0.42, 95% prediction interval –0.98 to 1.09; very low certainty), indicating that future studies could plausibly show either benefit or no effect. Overall, heterogeneity was substantial, and most evidence was of low or very low certainty.

XSL•FO
RenderX

Conclusions: This review is innovative in providing an up-to-date, cross-condition synthesis focused specifically on self-care as a multidimensional outcome, rather than on clinical end points alone or single diseases. The findings suggest that digital health interventions may be more effective for supporting self-care monitoring than for promoting broader behavioral maintenance or medication adherence. Evidence is limited by methodological heterogeneity, small sample sizes, short follow-up periods, and varied outcome measures. Larger designed trials using standardized self-care metrics and equity-focused approaches are needed to clarify effectiveness and guide implementation.

(J Med Internet Res 2026;28:e88708) doi:10.2196/88708

KEYWORDS

chronic disease; digital; e-health; self-care; self-management; technology

Introduction

Chronic diseases remain the leading cause of morbidity and mortality worldwide in adults, accounting for over 40 million deaths annually from noncommunicable diseases and imposing an escalating burden on health systems and societies [1]. As global populations age, the prevalence of chronic conditions, such as diabetes, heart failure, chronic obstructive pulmonary disease (COPD), cancer, and chronic kidney disease continues to rise, generating escalating demand for long-term management and self-management support [1-3].

In this scenario, self-care in the adult population has emerged as a cornerstone of chronic disease management to ensure sustainable health systems against the shortage of health professionals and limitations in service accessibility [4]. For the purpose of this review, chronic diseases are defined as long-term conditions requiring ongoing management and self-regulation, such as cardiovascular diseases, diabetes, and chronic respiratory conditions, which are characterized by a persistent need for continuous self-care behaviors. According to the World Health Organization (WHO) [4], self-care refers to “the ability of individuals, families, and communities to promote health, prevent disease, maintain health, and cope with illness and disability, with or without the support of a health care provider.” Building on this, Riegel’s middle-range theory of self-care of chronic illness conceptualizes self-care as a multidimensional process encompassing maintenance (health-promoting behaviors, such as diet, physical activity, smoking, alcohol consumption, mental health, medication adherence, and engagement with health care services), monitoring (recognition and interpretation of symptoms, tracking changes in physical or psychological status, and awareness of early warning signs), and management (decision-making and action in response to symptoms, including medication adjustment, seeking professional support, implementing coping strategies, and modifying daily activities) [5,6]. However, despite this well-established theoretical framework, the measurement of self-care remains highly heterogeneous across studies, with different instruments capturing distinct dimensions of the construct, thus limiting comparability and synthesis of evidence. In particular, medication adherence is among the most extensively investigated self-care behaviors across chronic conditions, given the challenges associated with promoting this competence and its substantial impact on hard clinical outcomes. Indeed, evidence consistently demonstrates that adequate self-care in chronic conditions, including medication adherence, is associated with improved quality of life, reduced hospitalizations, and decreased mortality [7,8]. Nevertheless, previous studies and reviews have reported inconsistent effects of interventions aimed at improving self-care, likely due to variability in intervention components, outcome measures, and methodological quality, highlighting the need for more rigorous and comprehensive syntheses.

Over the past decade, particularly during and after the COVID-19 pandemic, digital health interventions have played a pivotal role in supporting adult patients with chronic conditions in self-care and ensuring continuity of chronic disease management when in-person services were disrupted [9,10]. Digital health interventions encompass telemedicine, mobile health (mHealth), remote monitoring, big data, artificial intelligence, and other technology-driven tools to improve health outcomes [10]. However, the available synthesized evidence remains only partially informative regarding their effectiveness in improving self-care. Most systematic reviews have predominantly focused on clinical and service-oriented outcomes, such as hospitalizations, mortality, symptom severity, and physiological parameters [11-14], while rarely identifying self-care as a primary end point. Even when self-care related outcomes are considered, they are often operationalized through proxy constructs such as self-efficacy or focused on a single behavior, such as physical activity or diet adherence [14,15].

When focusing on self-care, available systematic reviews are largely disease-specific. In heart failure, telemonitoring interventions may improve self-care behaviors, although findings remain inconsistent [16]. In dementia, digital technologies have shown potential to support autonomy and self-management, but with heterogeneous and low-certainty evidence [12,17]. Digital interventions have also shown promise in atopic dermatitis and cancer care, although findings remain context-specific [18,19]. More recent trials and reviews continue to expand the field, showing benefits of digital self-care interventions across conditions such as cardiovascular disease, diabetes, Parkinson disease, asthma, multimorbidity, and COPD [14,15,20,21].

However, considerable variability persists in both outcomes and intervention design. Existing evidence is often limited to specific modalities, such as mobile apps and telemonitoring, thereby overlooking the rapid evolution and increasing diversity of newer digital technologies highlighted in recent literature [15]. Moreover, most studies remain focused on single diseases, even though individuals with chronic conditions commonly experience multimorbidity rather than isolated conditions. Two reviews were conducted on self-care and disease management

XSL•FO
RenderX

in chronic disease. One was limited to studies published up to 2020, focused on eHealth, and included only 3 chronic conditions, thereby not reflecting the rapid evolution and diversification of digital health technologies [22]. A recent scoping review mapped the landscape of digital health interventions for chronic disease management. However, by design, it provided a narrative and nonquantitative synthesis, limiting the ability to draw definitive conclusions on effectiveness [15].

Taken together, the literature remains insufficient due to fragmentation across diseases, heterogeneous conceptualization and measurement of self-care, and a lack of an evaluation of self-care as a core outcome. Therefore, there is still no systematic review that specifically synthesizes the effectiveness of digital health interventions for self-care across chronic conditions, while incorporating a wide spectrum of contemporary digital technologies. Addressing this gap is crucial because self-care is a key determinant of long-term outcomes, adherence, and the sustainability of health systems. Clarifying whether, for whom, and under what conditions digital interventions strengthen self-care will inform clinical decision-making, policymakers, and scalable implementation.

The aim of this systematic review and meta-analysis was therefore to evaluate the effectiveness of global digital health interventions in improving self-care among adults with chronic diseases.

Methods

Design

This systematic review and meta-analysis were conducted according to the Cochrane Handbook for Systematic Reviews of Interventions [23] and reported following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [24] (Multimedia Appendix 1) and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) [25] guidelines. The protocol was registered in PROSPERO (International Prospective Register of Systematic Reviews; CRD42023479314). No deviations from the predefined protocol occurred.

Eligibility Criteria

We included studies based on the following criteria: (1) randomized controlled trials (RCTs); (2) adults aged ≥18 years with chronic diseases, defined here as long-lasting conditions requiring ongoing management over time, typically characterized by slow progression and the need for sustained self-care activities; (3) participants recruited in any health care or community setting; (4) “digital health interventions for person,” according to the WHO framework, defined as “the capabilities of digital technology that can be implemented to achieve objectives that are targeted toward persons,” including members of the public who are potential or current users of health services and caregivers [10]. Eligible interventions included mHealth and eHealth approaches (eg, telemonitoring, video consultations, mobile apps, and electronic health records), software-based solution, and emerging technologies (eg, artificial intelligence, big data, and genomics) compared with any type of control condition; (5) self-care abilities, including treatment adherence, measured using validated quantitative questionnaire or instruments; and (6) studies published in English. Studies focusing on patients with cancer, mental illness, pregnant women, or in prisons were excluded.

Literature Search and Study Selection

We searched 4 electronic databases, PubMed (via PubMed), CINAHL (EBSCOhost), Scopus (Elsevier), and PsycINFO (EBSCOhost). We included studies published between January 1, 2013, and December 31, 2025, following the release of 2 key international and European policy documents on digital health in 2012 [26,27]. The search strings are reported in Multimedia Appendix 2. These were developed iteratively based on preliminary scoping searches and refined in consultation with an expert librarian. Additionally, the reference lists of included trials, gray literature, and registries were screened for any additional eligible studies.

All retrieved records were imported into Rayyan (Rayyan Systems Inc), an AI-assisted tool for systematic reviews, which was used to remove duplicates and manage the screening process. Titles and abstracts were independently screened by 2 reviewers (JL and FF). Full texts of potentially eligible studies were assessed independently by 2 reviewers. In this second step, the level of agreement was Cohen κ=0.83 (95% CI 0.71-0.93), demonstrating good interrater reliability. Disagreements were resolved by discussion or consultation with a third reviewer (DP).

Data Extraction

Two reviewers (JL and FF) independently extracted the following data using a standardized Excel form, which was piloted in 2 studies, after the screening was concluded. The data were authors, year, objectives, design, country, intervention, comparator, setting, participant characteristics, and outcomes (quantitative and narrative data), including follow-up duration and measurement instruments. No disagreements were detected. We contacted 6 corresponding authors of the included studies to request clarification or additional information when data were missing.

Quality Assessment

The methodological quality of included studies was assessed using the Cochrane Risk of Bias 2.0 tool for RCTs [28]. Discrepancies were resolved by consensus or by consulting a third reviewer. The quality of evidence for each outcome of interest was independently assessed by 2 reviewers using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system with the GRADEpro software [29]. GRADE is an internationally recognized and widely used framework that provides a transparent, reproducible, and systematic method for rating the certainty of evidence at the outcome level. It considers 5 key domains, risk of bias, inconsistency, indirectness, imprecision, and publication bias, to determine the confidence that the estimated effect is close to the true effect. By evaluating and combining these domains, the overall quality of evidence for each outcome is rated as high, moderate, low, or very low. Language to report confidence in

XSL•FO
RenderX

results according to the certainty of evidence was adherent to the Cochrane Handbook [23].

Data Synthesis and Analysis

The primary outcome was self-care measured with a validated questionnaire. Where sufficient data were available, we performed meta-analyses on comparable outcomes, grouped by outcomes and questionnaires, and homogeneity in digital interventions and conditions. We used standardized mean differences (SMDs, Hedges adjusted *g*) with 95% CIs because different versions of the same questionnaires were used across studies to measure the same outcome. We interpreted an SMD of 0.2, 0.5, and 0.8 as a small, medium, and large effect size, respectively. A random-effects model was used to account for clinical and methodological heterogeneity. Heterogeneity was assessed visually (forest plots) and statistically (I^2 statistic), with thresholds interpreted as follows: 0%-14% negligible, 15%-29% low, 30%-60% moderate, 50%-90% substantial, and 75%-100% considerable heterogeneity [23]. We planned to conduct subgroup analyses if any significant heterogeneity emerged in the pooled results and sensitivity analyses if high-risk-of-bias studies were included in the meta-analysis. Small-study effects were investigated through the visual inspection of the funnel plots and the Egger test when at least 10 studies were available. Prediction intervals were calculated with at least 5 studies and no asymmetry in the funnel plot [23]. No double-counting of participants occurred in the meta-analyses. Each study contributed only independent data, and the same participants were not included more than once within any single analysis. Analysis was performed with RevMan (version 5; Cochrane) and R software (R Foundation for Statistical Computing) with the package *metafor* [30].

When meta-analysis was not feasible (eg, insufficient data and heterogeneous outcome reporting), results were synthesized narratively.

Results

Characteristics of Included Studies

We included 55 RCTs [31-85] (Figure 1), of which almost half (n=26, 47.3%) were published in the past 5 years. The sample sizes ranged from 27 [79] to 1571 [39] (Multimedia Appendix 3 [31-85]). Most studies were conducted in the United States (n=11) [33,48-50,58,63,65,69,79,82,84] and in China (n=10) [43,45,47,61,64,70-72,76,77], followed by Iran [44,53,68], and Australia (n=3) [42,59,75] (Multimedia Appendix 3).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

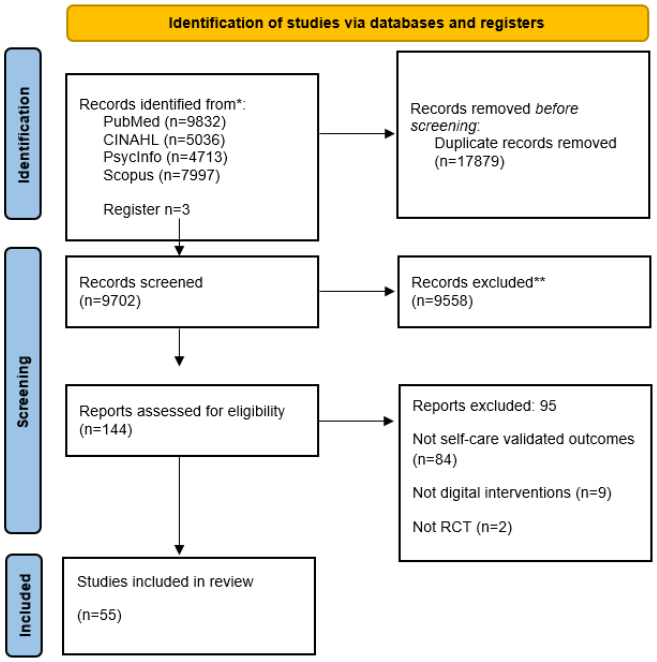

The most common chronic conditions were heart failure (n=13) [37,39,42,44,46,50,55,62,72-75,79], type 1 or type 2 diabetes (n=10) [31,33,35,41,45,47,49,53,59,61], and acute cardiovascular diseases, including myocardial infarction or stroke (n=10) [32,34,36,38,51,56,57,64,65,67], representing the most frequently targeted chronic conditions across the included studies.

Among the included studies, 5 involved participants with multiple chronic diseases [48,58,68,77,78], whereas atrial fibrillation was the primary focus in 4 others [40,52,76,82]. Other less frequent conditions included asthma (n=3) [43,60,84], COPD (n=2) [66,71], hypertension (n=2) [63,85], chronic hepatitis B [54], HIV infection [69], and epilepsy [70].

All studies obtained ethical approval and respected the privacy regulations.

Characteristics of Interventions

Characteristics of interventions and control care are reported in Table 1 and in Multimedia Appendix 3. Across the 55 included studies, digital health interventions for persons fell within the WHO [10] functional categories of “Targeted communication

XSL•FO
RenderX

to persons," "Personal health tracking," and "Person-based reporting."

Mobile or web-based applications were the most frequently used technology (n=39, 70.9%), followed by telemonitoring systems (n=24, 43.6%) and connected medical devices such as blood pressure monitors, glucometers, weight scales, electrocardiogram devices, or activity trackers (n=24, 43.6%). Text-message programs, including SMS text messages, WeChat-based messages, and email/SMS reminders, were used in 18 (32.7%) studies, whereas tablet-based apps were reported in 6 (10.9%) studies.

Overall, interventions were predominantly multicomponent, with the most common configuration combining mobile or web-based applications, telemonitoring, and connected devices (n=12, 21.8%). Stand-alone mobile or web-based applications were used in 10 (18.2%) studies, whereas mobile/web-based application plus text-message programs were reported in 8 (14.5%) studies. Telemonitoring combined with connected devices but without apps was used in 6 (10.9%) studies, and text-message–only interventions were also used in 6 (10.9%) studies.

Interventions can also be characterized according to their main behavior change technique categories [86]. The most common category was prompts/cues, operationalized through reminder messages, push notifications, interactive voice response messages, or automated prompts, which were present in 54 (98.2%) studies. Self-monitoring of behavior or outcomes was also frequent and was usually implemented through apps or connected devices, allowing patients to enter or transmit symptoms, medication use, or physiological parameters; active patient interaction was reported in 44 (80%) studies. Feedback on behavior/biofeedback, often delivered automatically or after clinician review of transmitted data, was commonly paired with alerts for symptoms or high-risk values, which were reported in 33 (60%) studies. A further recurring behavior change technique category was instruction on how to perform the behavior together with information about health consequences, usually delivered through educational modules, videos, app-based content, or structured tele-education, and identifiable in 48 (87.3%) studies. Goal setting, action planning, and behavioral reinforcement were observed in 24 (43.6%) studies, primarily in multicomponent coaching or structured self-management programs. Social support features, such as peer interaction, online discussion boards, caregiver involvement, or shared monitoring functions (eg, "medfriend"), were less common and were identified in 10 (18.2%) studies.

These components were delivered at different levels. Phone calls were incorporated in 21 (38.2%) studies, mainly for follow-up, motivational counseling, or technical support, and face-to-face sessions were included in 12 (21.8%) studies, usually for initial training or reinforcement. All interventions included patient-level components (n=55, 100% studies), such as self-monitoring, reminders, education, and behavioral support. In addition, provider-level involvement, including monitoring of transmitted data, follow-up contacts, or clinical feedback delivered by nurses, pharmacists, physicians, or multidisciplinary teams, was identified in 43 studies (78.2%). In several cases, health care professionals also contributed to treatment adjustment or clinical decision-making based on remotely collected data.

XSL•FO
RenderX

Table 1. Characteristics of the interventions.

Study	Mobile app	Tablet	Other devices (eg, BP[a] and weight)	Text message	Telemonitoring	Reminder/motivational messages	Phone calls	Alerts for symptoms/high risk	Face-to-face sessions	Patient interaction (eg, input parameters)	Conditions
Hoban 2013 [50]	No	No	Yes	No	Yes	No	Yes	Yes	Yes	No	Heart failure
Kirwan 2013 [59]	Yes	No	No	Yes	No	Yes	No	No	No	Yes	Diabetes
Arora 2013 [33]	No	No	No	Yes	No	Yes	No	No	No	No	Diabetes
Heisler 2014 [49]	No	Yes	No	No	No	Yes	Yes	No	Yes	No	Hypertension
Boyne 2014 [37]	No	No	Yes	No	Yes	Yes	No	Yes	No	No	Heart failure
Jahangard-Rafsanjani 2015 [53]	No	No	Yes (glucometer)	No	No	Yes	Yes	No	Yes	Yes	Diabetes
Vuorinen 2014 [73]	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Heart failure
Park 2014 [65]	No	No	No	Yes	No	Yes	No	No	No	No	Coronary heart disease
Hägglund 2015 [46]	No	Yes	Yes (weight)	No	Yes	Yes	No	Yes	No	Yes	Heart failure
Pfaeffli Dale 2015 [67]	Yes (web + mobile)	No	Yes (pedometer)	Yes	No	Yes	No	No	No	Yes	Coronary heart disease
Kamal 2015 [57]	No	No	No	Yes	No	Yes	No	No	No	No	Stroke
Jeon 2016 [54]	Yes	No	No	No	No	Yes	No	Yes	No	Yes	Hepatitis B
Akhu-Zaheya 2016 [32]	No	No	No	Yes	No	Yes	No	No	No	No	Cardiovascular diseases
Kim 2016 [58]	Yes	No	Yes (BP monitor)	No	Yes	Yes	No	Yes	No	Yes	Hypertension
Koufopoulos 2016 [60]	Yes (web+mobile)	No	No	No	No	Yes (peer motivational)	No	No	No	Yes	Asthma
Baron 2017 [35]	Yes	No	Yes (BP, glucose)	No	Yes	Yes	Yes	Yes	No	Yes	Diabetes
Melin 2018 [62]	No	Yes	Yes (scale)	No	Yes	Yes	No	Yes	No	No	Heart failure
Desteghe 2018 [40]	No	No	Yes (MEMS[b])	No	Yes	Yes	Yes	Yes	No	No	Atrial fibrillation
Kamal 2018 [56]	No	No	No	Yes	No	Yes	Yes (IVR[c])	No	No	Yes	Stroke/MI[d]
Morawski 2018 [63]	Yes	No	No	No	No	Yes	No	No	No	Yes	Hypertension
Schnall 2018 [69]	Yes (web app)	Yes	No	No	No	Yes	No	No	No	Yes	HIV
Agarwal 2019 [31]	Yes	No	No	No	No	Yes	No	No	No	Yes	Hypertension

XSL•FO
RenderX

Study	Mobile app	Tablet	Other devices (eg, BP[a] and weight)	Text message	Telemonitoring	Reminder/motivational messages	Phone calls	Alerts for symptoms/high risk	Face-to-face sessions	Patient interaction (eg, input parameters)	Conditions
Sun 2019 [72]	Yes	No	No	Yes (WeChat)	Yes	Yes	Yes	Yes	Yes	Yes	Heart failure
Park 2020 [66]	Yes	No	Yes (pedometer)	Yes	No	Yes	Yes	Yes	Yes	Yes	COPD[e]
Stamenova 2020 [71]	Yes	No	Yes (BP and weight)	No	Yes	Yes	No	Yes	No	Yes	COPD
Ding 2020 [42]	No	No	Yes (scale)	No	Yes	Yes	Yes	Yes	No	Yes	Heart failure
Si 2020 [70]	Yes	No	No	No	Yes	Yes	No	Yes	No	Yes	Epilepsy
Wonggom 2020 [75]	Yes (tablet)	Yes	No	No	No	Yes	No	No	Yes	Yes	Heart failure
Dincer 2020 [41]	Yes	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Diabetes
Hong 2021 [51]	No	No	Yes (BP monitor)	No	Yes	Yes	Yes	Yes	Yes	No	Coronary heart disease
Bruggmann 2021 [38]	Yes	No	Yes (BP, HR[f], and weight)	No	Yes	Yes	Yes	Yes	No	Yes	HF[g], COPD, diabetes
Jiang 2021 [55]	Yes	No	Yes (BP and weight)	No	Yes	Yes	Yes	Yes	Yes	Yes	Heart failure
Hsieh 2021 [52]	Yes (web)	Yes	No	No	Yes	Yes	No	Yes	No	Yes	Atrial fibrillation
Ni 2022 [64]	Yes (WeChat)	No	Yes (BP monitor)	Yes	No	Yes	No	Yes	No	Yes	Coronary heart disease
Ware 2022 [74]	Yes	No	Yes (BP, glucose, and weight)	No	Yes	Yes	No	Yes	No	Yes	Hypertension
Han 2023 [47]	Yes	No	Yes (glucometer)	No	Yes	Yes	Yes	Yes	No	Yes	Type 2 diabetes
Poorcheraghi 2023 [68]	Yes	No	No	No	No	Yes	Yes	Yes	No	Yes	Polypharmacy
Deckwart 2023 [39]	No	No	Yes (ECG[h], BP, weight, and SpO_2)	No	Yes	Yes	Yes	Yes	No	Yes	Heart failure
Guo 2023 [45]	Yes	No	Yes (glucose sensor)	No	Yes	Yes	Yes	Yes	Yes	Yes	Diabetes
Bernal-Jiménez 2024 [36]	Yes	No	No	No	No	Yes	No	No	No	Yes	Coronary heart disease
FarzanehRad 2024 [44]	No	No	No	Yes	No	Yes	Yes	Yes	No	No	Heart failure
Hartch 2024 [48]	Yes	No	No	No	No	Yes	Yes	Yes	No	Yes	Multiple chronic diseases
Babu 2024 [34]	Yes	No	Yes (BP and glucose)	No	Yes	Yes	No	Yes	No	Yes	Stroke

XSL•FO
RenderX

Study	Mobile app	Tablet	Other devices (eg, BP[a] and weight)	Text message	Telemonitoring	Reminder/motivational messages	Phone calls	Alerts for symptoms/high risk	Face-to-face sessions	Patient interaction (eg, input parameters)	Conditions
Ye 2024 [77]	Yes (WeChat)	No	No	Yes	No	Yes	No	No	No	Yes	Diabetes + Hypertension
Xu 2024 [76]	Yes	No	No	No	Yes	Yes	No	Yes	No	Yes	Atrial fibrillation
Erdoğan 2024 [43]	Yes (web-based)	No	No	Yes (email/SMS)	No	Yes	No	No	No	Yes	Asthma
Lee 2024 [61]	Yes	No	No	No	No	Yes	No	Yes	No	Yes	Type 2 diabetes
Hwang 2025 [78]	Yes	No	Yes (BP monitor)	No	Yes	Yes	No	Yes	No	Yes	Hypertension
Keskin 2024 [85]	Yes	No	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Hypertension
Kitsiou 2025 [79]	Yes	No	Yes (BP monitor)	No	Yes	Yes	No	Yes	No	Yes	Hypertension
Lee 2025 [80]	Yes	No	No	Yes	No	Yes	No	No	No	Yes	Type 2 diabetes
Lippke 2025 [81]	Yes	No	No	Yes	No	Yes	No	No	No	Yes	Cardiovascular disease
Magnani 2025 [82]	Yes	No	No	Yes	No	Yes	No	No	No	Yes	Atrial fibrillation
Meyer 2025 [83]	Yes	No	No	No	No	Yes	No	No	No	Yes	Cardiovascular disease
Silberman 2025 [84]	Yes	No	No	Yes	No	Yes	No	No	No	Yes	Cardiovascular disease

[a]BP: blood pressure.

[b]MEMS: Medication Event Monitoring System.

[c]IVR: interactive voice response.

[d]MI: myocardial infarction.

[e]COPD: chronic obstructive pulmonary disease.

[f]HR: heart rate.

[g]HF: heart failure.

[h]ECG: electrocardiogram.

Risk of Bias

Of the 55 RCTs, 34 (61.8%) were rated as having some concerns, 20 (36.4%) as high risk, and 1 (1.8%) as low risk (Multimedia Appendix 4 [31-85]).

Bias arising from the randomization process: most studies adequately described the random sequence generation and allocation procedures. A total of 40 (72.7%) trials were rated as low risk, while 15 (27.3%) were judged as having some concerns.

Bias due to deviations from intended interventions: in this domain, 35 (63.6%) studies were assessed as low risk, 18 (32.7%) as some concerns, and 2 (3.6%) as high risk.

Bias due to missing outcome data: most trials reported adequate follow-up and outcome data. A total of 37 (67.3%) studies were rated as low risk, 16 (29.1%) as some concerns, and 2 (3.6%) as high risk.

Bias in the measurement of the outcome: the majority of studies were judged as having some concerns in this domain (n=53, 96.4% studies), while 2 (3.6%) studies were rated as high risk.

Bias in selection of the reported result: most studies were judged as low risk (n=39, 70.9% studies), whereas 16 (29.1%) were rated as having some concerns.

XSL•FO
RenderX

Outcome Measures

Overview

The results are categorized by self-care and medication adherence. Within both categories, we reported the results by meta-analysis, and narrative results emerged from studies not pooled in the meta-analysis due to insufficient data or high heterogeneity in the studies' characteristics.

Self-Care Outcomes

Self-Care in Diabetes

Summary of Diabetes Self-Care Activities

The results are reported according to the most commonly investigated domains, which include general diet, specific diet, exercise, foot care, and glucose monitoring, as specified in the original instrument.

Foot care: the pooled estimate from 5 RCTs in patients with diabetes mixing different types of digital interventions [31,33,35,45,77] showed that digital interventions may result in little to no improvement in foot care compared with usual care (SMD=0.44, 95% CI –0.60 to 1.47; I^2=95%; Figure 2 [31,33,35,45,77]), but the evidence is very uncertain. According to the GRADE approach, the certainty of the evidence was rated as very low, downgraded for risk of bias, inconsistency, and imprecision (Table 2).

XSL•FO
RenderX

Figure 2. Meta-analysis for the Summary of Diabetes Self-Care Activities (SDSCA) outcomes. HKSJ: Hartung-Knapp-Sidik-Jonkman; REML: restricted maximum likelihood.

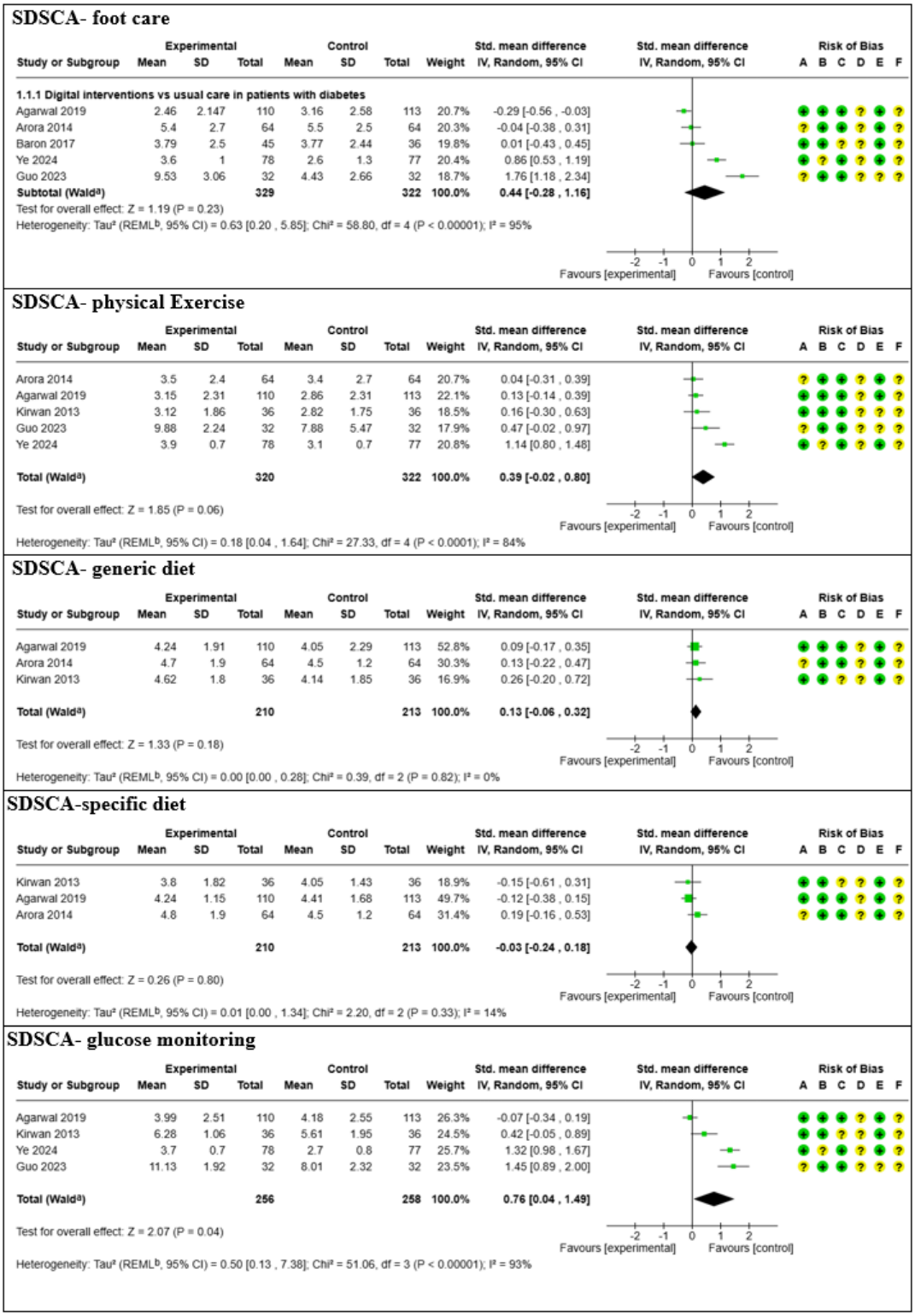

XSL•FO
RenderX

Table 2. Grading of Recommendations Assessment, Development and Evaluation certainty of evidence.

	Certainty assessment							Patients, n		Effect	Certainty
	Studies, n	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	DHIs[a]	Usual care	SMD[b] (95% CI)	
SDSCA[c]-foot care	5	RCTs	Serious[d]	Serious[e]	Not serious	Serious[f]	None	329	322	0.44 SD higher (–0.60 to 1.47)	Very low[d,e,f]
SDSCA-physical exercise	5	RCTs	Serious[d]	Serious[e]	Not serious	Serious[f]	None	320	322	0.39 SD higher (–0.18 to 0.96)	Very low[d,e,f]
SDSCA-generic diet	3	RCTs	Serious[d]	Not serious	Not serious	Serious[f]	None	210	213	0.13 SD higher (–0.06 to 0.32)	Low[d,f]
SDSCA-specific diet	3	RCTs	Serious[d]	Not serious	Not serious	Serious	None	210	213	0.03 SD lower (–0.47 to 0.42)	Low[d]
SDSCA-glucose monitoring	4	RCTs	Serious[d]	Very serious[e]	Not serious	Very serious[f]	None	256	258	0.76 SD higher (–0.40 to 1.93)	Very low,[d,e,f]
European Heart Failure Self-Care Behaviour Scale	3	RCTs	Serious[d]	Very serious[e]	Not serious	Very serious[f]	None	1040	1007	–0.18 SD lower (–0.74 to 0.39)	Very low[d,f,g]
Self-Care of Heart Failure Index-maintenance	5	RCTs	Serious[d]	Very serious[e]	Not serious	Serious[f]	None	208	156	0.48 SD higher (–0.10 to 1.05)	Very low[d,f]
Self-Care of Heart Failure Index-monitoring	5	RCTs	Serious[d]	Serious[e]	Not serious	Not serious	None	208	156	0.49 SD higher (0.13 to 0.85)	Low[d,e]
Medication adherence	17	RCTs	Serious[d]	Serious[e]	Not serious	Serious[f]	None	1372	1383	0.06 SD higher (–0.31 to 0.42)	Very low[d,f]

[a]DHI: digital health intervention.

[b]SMD: standardized mean difference.

[c]SDSCA: Summary of Diabetes Self-Care Activities.

[d]Some concerns were reported in different domains across studies.

[e]Inconsistency was partially explained by study location.

[f]The 95% CI includes no meaningful benefit and no effect or arm.

[g]No reasons were identified to explain the statistical heterogeneity.

In subgroup analysis (Multimedia Appendix 5 [31,33,35,45,59,77]), when considering studies conducted in high-income countries (United States, United Kingdom, and Canada), statistical heterogeneity markedly decreased (I^2=14%) and the pooled effect continued to show no meaningful improvement in self-care (SMD=–0.15, 95% CI –0.57 to 0.27). Studies conducted in China showed no meaningful improvement in self-care, and the statistical heterogeneity remained high (SMD=1.27, 95% CI –4.42 to 6.97; I^2=86%). The sensitivity analyses yielded results consistent with the primary analysis.

Physical exercise: the pooled estimate from 5 RCTs mixing different types of digital interventions [31,33,45,59,77] showed that digital interventions may result in a small improvement in physical exercise self-care compared with usual or augmented usual care (SMD=0.39, 95% CI –0.18 to 0.96; I^2=84%; Figure 2). According to the GRADE approach, the certainty of the evidence was rated as low, downgraded for risk of bias and imprecision (Table 2).

In subgroup analysis (Multimedia Appendix 5), studies conducted in high-income countries (United States, United Kingdom, and Canada) showed no meaningful improvement in physical exercise self-care (SMD=0.11, 95% CI –0.03 to 0.25; I^2=0%), as well as studies conducted in China (SMD=0.83, 95%

XSL•FO
RenderX

CI –3.38 to 5.04; I^2=79%). The sensitivity analyses yielded results consistent with the primary analysis.

Generic diet: the pooled estimate of 3 studies mixing different types of digital interventions [31,33,59] showed that digital interventions may result in little to no improvement in general dietary self-care compared with usual or augmented usual care (SMD=0.13, 95% CI –0.06 to 0.32; I^2=0%; Figure 2). According to the GRADE approach, the certainty of the evidence was rated as low, downgraded for risk of bias and imprecision (Table 2).

Specific diet: the pooled estimate of 3 studies mixing different types of digital interventions [31,33,59] showed that digital interventions may result in little to no difference in specific dietary self-care compared with usual or augmented usual care (SMD=–0.03, 95% CI –0.47 to 0.42; I^2=14%; Figure 2). According to the GRADE approach, the certainty of the evidence was rated as low, downgraded for risk of bias and imprecision (Table 2).

Glucose monitoring: the pooled estimate from 4 RCTs mixing different types of digital interventions [31,45,59,77] showed that digital interventions may result in little to no difference in glucose monitoring self-care compared with usual care (SMD=0.76, 95% CI –0.40 to 1.93; I^2=93%; Figure 2), but the evidence is very uncertain. According to the GRADE approach, the certainty of the evidence was rated as very low, downgraded for risk of bias, inconsistency, and imprecision (Table 2).

In subgroup analysis (Multimedia Appendix 5), studies conducted in high-income countries (United States and United Kingdom) showed no meaningful improvement in glucose monitoring self-care (SMD=0.13, 95% CI –2.98 to 3.25; I^2=70%), whereas studies conducted in China showed a large and statistically significant effect favoring digital interventions (SMD=1.36, 95% CI 0.66 to 2.06; I^2=0%). The difference between subgroups was statistically significant (P<.001). The sensitivity analyses yielded results consistent with the primary analysis.

Narrative Results

Across studies not included in the meta-analysis (Multimedia Appendix 6 [41,45,53,61,67]), digital education and monitoring tools were generally associated with improvements in adherence to dietary recommendations, physical activity, and glucose monitoring. Evidence from mHealth and tele-education programs suggests a positive effect on dietary behaviors and overall self-care, as reflected by improvements in Summary of Diabetes Self-Care Activities (SDSCA) scores across several studies [45,61,77]. Similarly, pharmacist-supported digital education was associated with improvements in broader self-care domains, including diet, glucose monitoring, and foot care [53]. Interventions specifically targeting foot care through mobile apps also demonstrated potential benefits for self-care behaviors [41].

Self-Care in Heart Failure

European Heart Failure Self-Care Behaviour Scale

The pooled estimate from 3 studies mixing different types of digital interventions [37,39,73] showed that digital interventions may result in little to no difference in self-care compared with usual or augmented usual care, but the evidence is very uncertain (SMD=–0.18, 95% CI –1.39 to 1.03; I^2=96%; Figure 3 [37,39,73]). According to the GRADE approach, the certainty of the evidence was rated as very low, downgraded for risk of bias, inconsistency, and imprecision (Table 2).

XSL•FO
RenderX

Figure 3. Meta-analysis for the Self-care in heart failure outcomes. HKSJ: Hartung-Knapp-Sidik-Jonkman; REML: restricted maximum likelihood.

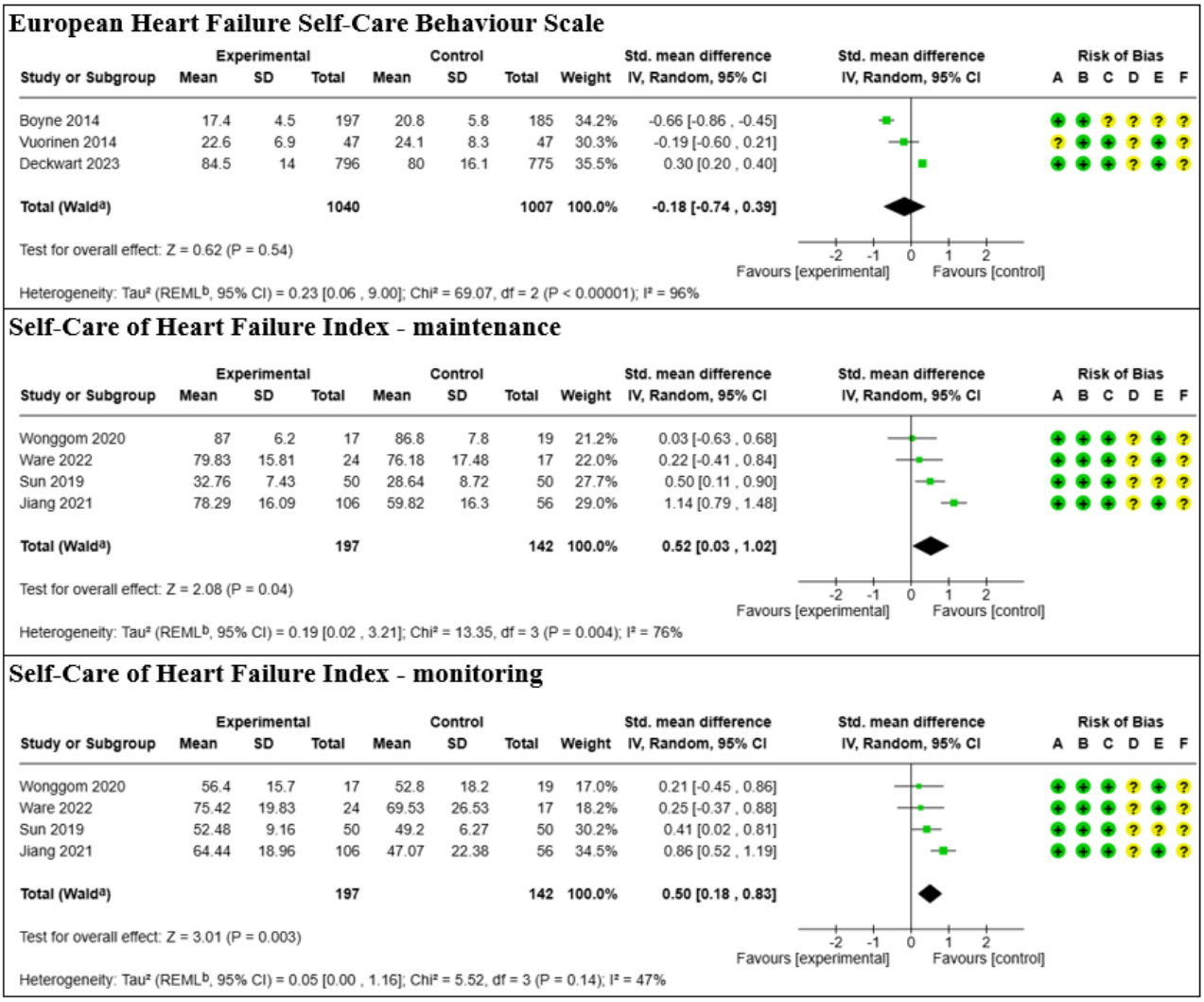

Self-Care of Heart Failure Index

Maintenance: the overall pooled analysis from 5 studies mixing different types of digital interventions based on an application [55,72,74,75,79] that the experimental intervention may result in little to no improvement in the outcome when compared with the control (SMD=0.48, 95% CI –0.10 to 1.05; *P*=.08; Figure 3) with moderate heterogeneity (I^2=70%; *P*=.005). According to the GRADE approach, the certainty of the evidence was rated as very low due to inconsistency, imprecision, and risk of bias (Table 2).

Subgroup analyses (Multimedia Appendix 5 [55,72,74,75,79]) were conducted according to study location. Among studies performed in China [55,72] (n=262), the SMD showed no significant effect (SMD=0.83, 95% CI –3.19 to 4.85; *P*=.009), with high heterogeneity (I^2=82%), as well as studies conducted in Canada-Australia [74,75,79] (n=102; SMD=0.14, 95% CI –0.12 to 0.40; *P*=.15) but with no heterogeneity (I^2=0%). The sensitivity analyses yielded results consistent with the primary analysis.

Monitoring: the overall pooled analysis from 5 studies mixing different type of digital interventions based on an application [55,72,74,75,79] showed that the experimental intervention probably results in a moderate improvement in self-care monitoring when compared with the control (SMD=0.49, 95% CI 0.13-0.85; *P*=.02; Figure 3) with moderate heterogeneity (I^2=38%; *P*=.21). According to the GRADE approach, the certainty of the evidence was rated as low, downgraded due to inconsistency and risk of bias (Table 2).

Subgroup analyses (Multimedia Appendix 5) were conducted according to study location. Among studies conducted in China [55,72], the SMD showed no statistically significant effect (SMD=0.65, 95% CI –2.14 to 3.45; *P*=.21) with substantial heterogeneity (I^2=64%; *P*=.10). In contrast, in the studies conducted in Canada-Australia [74,75], the SMD favored the experimental intervention, showing a moderate and statistically significant effect (SMD=0.25, 95% CI 0.12-0.38; *P*=.01) and no heterogeneity (I^2=0%; *P*=.98). The test for subgroup differences was not statistically significant (*P*=.08). The sensitivity analyses yielded results consistent with the primary analysis.

Narrative Results

Contrasting results emerged from studies not included in the meta-analysis (Multimedia Appendix 6 [37,46,50,62,79]).

XSL•FO
RenderX

Interventions combining daily monitoring with clinician oversight, often involving nurse-led review and feedback, were generally associated with better self-care performance as measured by the European Heart Failure Self-Care Behaviour Scale (EHFScB)-9 and related instruments [46,62]. Positive changes in specific behavioral domains, such as physical activity, were also observed in some interventions, including telemonitoring combined with home-based support, although effects were not consistent across all follow-up time points [50].

Similarly, programs integrating interactive educational components, automated alerts, and patient-provider communication reported improvements in compliance-related outcomes [37]. Conversely, a recent mHealth intervention integrating remote monitoring, education, and motivational messaging did not demonstrate meaningful differences in Self-Care of Heart Failure Index (SCHFI) self-care management compared with usual care [79].

Self-Care in Other Chronic Conditions

Other studies investigated digital interventions in different conditions (Multimedia Appendix 6 [51,54,70,78,80,81]).

Overall, digital interventions incorporating education, monitoring, and communication components showed heterogeneous effects on self-care outcomes depending on the condition and measurement tool. In cardiovascular conditions, interventions combining telemonitoring with educational support and follow-up were associated with improvements in self-care as measured by the Partners in Health Scale [51]. Similarly, app-based self-management interventions incorporating educational modules, symptom tracking, and clinician communication in other conditions, including chronic hepatitis B, epilepsy, and hypertension, were linked to positive changes in self-care [54,70,81]. In contrast, studies conducted in patients with COPD generally reported no meaningful differences in self-care outcomes when using mobile apps or telemonitoring systems, including those integrating remote monitoring and clinician support [66,71].

Recent studies using digital health coaching programs and mobile apps reported mixed findings across domains of the Self-Care of Chronic Illness Inventory. While some interventions suggested improvements in specific components, such as self-care maintenance and monitoring in Parkinson disease [80], others did not show statistically significant differences across self-care domains in patients with multiple chronic conditions [78].

Medication Adherence

Overview

The overall pooled analysis of 17 studies showed that digital interventions may result in little to no improvement in medication adherence compared with usual or augmented usual care (SMD=0.06, 95% CI –0.31 to 0.42; I^2=89%; Figure 4 [32-34,43,44,49,56-58,60,63-65,67]), but the evidence is very uncertain. The 95% prediction interval (–0.98 to 1.09) indicated that future studies could plausibly show either benefit or no effect. According to the GRADE approach, the certainty of the evidence was rated as very low, downgraded for risk of bias, inconsistency, and imprecision (Table 2).

Figure 4. Meta-analysis for the Medication adherence outcome. HKSJ: Hartung-Knapp-Sidik-Jonkman; DL: DerSimonian-Laird.

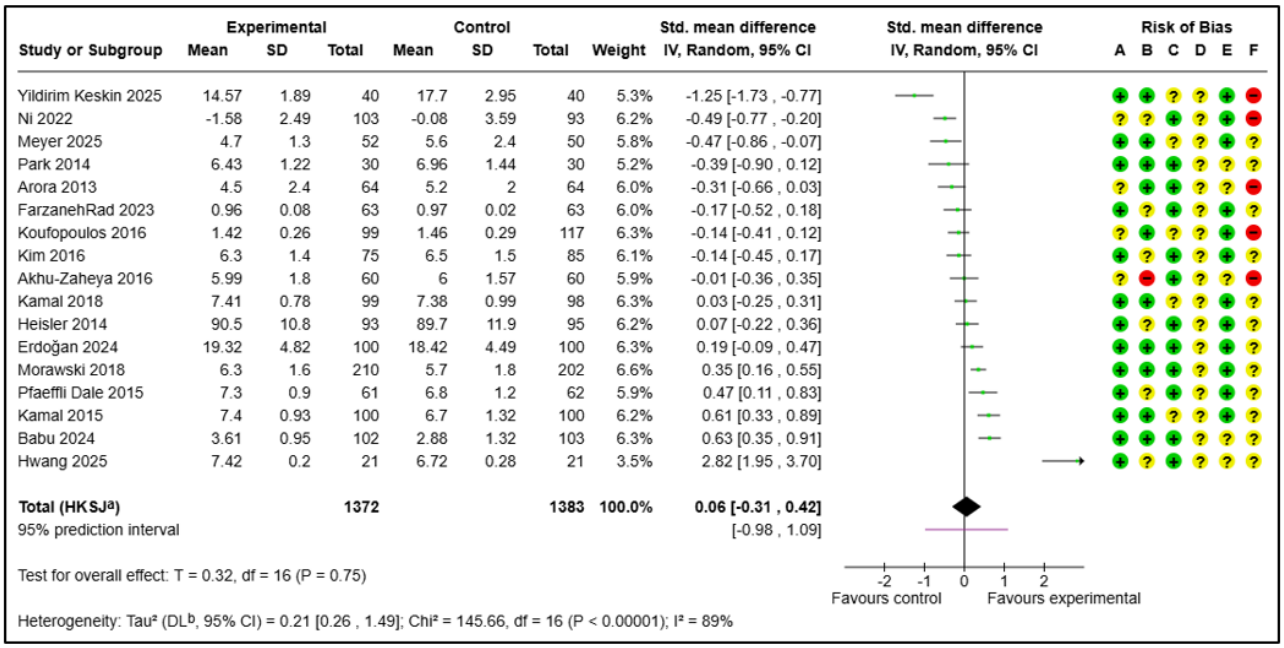

Subgroup analyses did not meaningfully alter these findings. Grouping studies by type of intervention (text messaging/reminders vs multicomponent interventions) did not reduce heterogeneity and confirmed the absence of a significant effect. Grouping by country income level slightly increased the heterogeneity among studies in both the high-income and low-medium income country groups and still showed no significant improvement in adherence. The sensitivity analyses yielded results consistent with the primary analysis.

Narrative Results

Several studies not included in the meta-analysis provide further insights (Multimedia Appendix 6). Across studies using the

XSL•FO
RenderX

Morisky Medication Adherence Scale (MMAS-4 and MMAS-8), interventions combining remote monitoring, educational content, and personalized feedback are generally associated with higher adherence levels across multiple conditions, including atrial fibrillation, multimorbidity, and diabetes [40,53,68]. mHealth apps supporting medication management and patient-provider communication also suggested beneficial effects on adherence in atrial fibrillation [76], as did digitally supported behavioral interventions in coronary heart disease [36]. Studies using alternative adherence measures, including the Medication Adherence Rating Scale (MARS) and the Center for Adherence Support Evaluation Adherence Index, also reported positive changes in adherence with interventions based on web platforms, interactive applications, and tailored digital feedback in patients with atrial fibrillation and HIV [52,69]. In contrast, a smartphone-based intervention incorporating a conversational agent and remote monitoring in patients with atrial fibrillation did not show differences in self-reported nonadherence over follow-up [82]. Findings from studies using the Adherence to Refills and Medication Scale (ARMS) were heterogeneous across conditions. Mobile app–based interventions without intensive support showed no differences in adherence in patients with cardiovascular disease [38], whereas interventions incorporating reminders, feedback, and engagement features reported improved adherence in patients with chronic conditions requiring long-term pharmacotherapy [48]. More recent interventions integrating wearable monitoring, personalized feedback, and behavioral interventions suggested potential benefits for adherence in patients with asthma [84].

Discussion

Principal Findings

This systematic review and meta-analysis evaluated the effectiveness of digital health interventions in improving self-care and medication adherence among adults with chronic diseases. A total of 47 RCTs were included, encompassing a wide range of chronic conditions, geographic settings, and digital modalities. Overall, the evidence suggests that digital interventions can support patient self-care and adherence behaviors; however, the effects were heterogeneous and not consistently statistically significant.

For diabetes, all SDSCA domains (foot care, diet, physical exercise, and glucose monitoring) showed small, nonsignificant effects with low to very low certainty, indicating that confidence in these estimates is limited. Notably, subgroup analyses in glucose monitoring revealed significant effects in studies conducted in China compared to those conducted in the United States and the United Kingdom, suggesting that contextual factors, such as program intensity, cultural adaptation, and professional involvement, may influence outcomes. Across studies not included in the meta-analysis, digital interventions, particularly those integrating education, monitoring, and pharmacist or clinician support, tended to show improvements in multiple self-care domains (eg, diet, glucose monitoring, and foot care). However, these studies are likely affected by methodological limitations highlighted in the risk of bias assessment. Our meta-analysis findings are inconsistent with the existing body of evidence, although the results from the narrative synthesis are consistent with those reported by Liu et al [87], who found significant improvements in hemoglobin A_{1c}, blood pressure, and diabetes self-management activities, especially in short-term interventions (<6 months) and younger populations, and with Shrivastava et al [88], who observed general improvement trends but limited statistical significance due to small samples and unclear risk of bias.

For heart failure, pooled analyses suggested small to moderate improvements in self-care monitoring when measured with the SCHFI instrument, particularly in studies conducted in Canada, the United States, and Australia, compared to those in China, while little to no effects were found in maintenance, nor on self-care when measured with the EHFScB scale. However, certainty of evidence ranged from very low to low, indicating limited confidence in these estimates and that future evidence is likely to change the direction or magnitude of the observed effects. Our narrative synthesis indicates heterogeneous and inconsistent findings; interventions incorporating clinician oversight and interactive components were associated with improvements in specific self-care behaviors, although these findings should be interpreted cautiously, given the generally high or unclear risk of bias across several domains. The absence of an effect on self-care maintenance likely reflects the current capacity of available technologies to support the detection and monitoring of clinical parameters, while remaining insufficient to effectively influence self-care maintenance, defined as behavioral change in key habits [6], such as physical activity, diet, and medication adherence. Overall, our results highlight the uncertainty of the current evidence reporting mixed findings across studies, suggesting that effectiveness remains uncertain, although interactive, feedback-driven models may offer potential benefits [89,90].

This is also supported by the pooled estimate from 17 RCTs on medication adherence across different conditions, which showed no significant overall effect, with very low certainty, indicating that the true effect is likely to be substantially different from the observed estimate and that confidence in this finding is very limited. This was supported by the wide prediction interval, which indicates substantial variability in effects across different settings, suggesting that future studies may plausibly show benefit, no effect, or even harm. No further evidence emerged from subgroup analyses by intervention type and country. This is supported by the narrative synthesis that showed contrasting results, even though findings are predominantly derived from studies at high or moderate risk of bias, requiring attention in interpreting the results. This contrasts with Lanke et al [21] and Kim et al [20], who found significant improvements, especially when apps included interactive functions, enhanced reminders, or data sharing. Shrivastava et al [88] similarly observed improvement trends but noted that few studies reached significance, largely due to methodological limitations. In stroke, Zeng et al [91] reported significant adherence gains from mHealth apps and messaging interventions.

Differences between our findings and prior meta-analyses may reflect broader inclusion criteria, greater heterogeneity in populations and interventions, and the inclusion of multiple

XSL•FO
RenderX

digital modalities (apps, telemonitoring, SMS text messages, and web portals), which may have diluted specific adherence effects. Many included trials were small, short-term, and used heterogeneous adherence measures (MMAS, MARS, ARMS, and Medication Event Monitoring System), which reduced the precision and certainty of the evidence.

In other chronic conditions, narrative findings suggest that multicomponent digital interventions may improve self-care in some diseases, particularly cardiovascular conditions, chronic hepatitis B, epilepsy, and hypertension, where improvements were reported in domains such as symptom monitoring, treatment adherence, and overall self-management. In contrast, no significant effects were observed in COPD, while mixed results emerged for Parkinson disease, where improvements were limited to specific domains such as monitoring and maintenance but not management, and for patients with multimorbidity, among whom no significant changes were reported across self-care domains. However, given that the majority of studies were assessed as having a high or moderate risk of bias, these results should be interpreted with caution.

Overall, the inconsistency in our findings may be explained by broader inclusion criteria, greater heterogeneity in populations and outcomes, and diverse intervention modalities. Our review encompassed multiple chronic conditions and technologies (apps, telemonitoring, SMS text messages, and web portals), which may dilute specific effects on medication adherence. Additionally, many included studies were small, of short duration, or used heterogeneous measurement tools (MMAS, MARS, ARMS, and Medication Event Monitoring System), leading to lower precision and downgraded certainty. Furthermore, despite efforts to ensure methodological consistency, self-care remains a complex and multidimensional construct to measure [92], and its measurement varies substantially across studies. The use of different validated instruments, each capturing distinct dimensions of self-care, may limit comparability and contribute to heterogeneity in the findings.

In terms of clinical implications, this review both corroborates, contrasts, and extends prior literature. It confirms that digital health interventions may hold potential for improving self-care, especially in heart failure and monitoring, while also indicating that their effects are uneven and context-dependent [14,15,20,21]. Specifically, due to inconsistent self-care maintenance results across diabetes, heart failure, and other conditions, it remains necessary to support the implementation of digital health interventions with human components to ensure long-term engagement of the person, such as in-person visits and motivational interviewing [93,94]. Other challenges should also be considered in implementing digital health interventions. Equity and digital health literacy remain critical determinants of effectiveness that are not thoroughly addressed in the included studies. Evidence from Turnbull et al [95] showed that the impact of digital self-care interventions is not uniform across populations. Their review showed that web-based programs can benefit disadvantaged or minority groups (eg, ethnic minorities and individuals with lower socioeconomic status), although digital access, health literacy, and usability continue to pose substantial barriers. Ge et al [96] highlight that digital exclusion among older adults is a multidimensional phenomenon that includes resource, skill, and motivational constraints limiting engagement with digital health interventions. Sociodemographic, functional, and psychological factors, such as limited confidence and technology-related anxiety, also reduce participation [97,98] but were rarely analyzed as moderators in the studies included in this review. Therefore, this systematic review highlights the importance of integrating digital interventions into routine care through structured implementation strategies. These include ensuring education for both patients and professionals, tailoring interventions to individual levels of digital literacy, and embedding digital tools within existing clinical pathways to enhance continuity rather than create parallel workflows. Moreover, co-design approaches and ongoing user engagement should be prioritized to improve usability, acceptability, and adherence [10,93]. Health systems adopting digital self-care solutions should also consider equity-oriented implementation frameworks to prevent the widening of existing disparities and to ensure that innovations benefit all patient populations.

As regards research implications, several priorities emerge. First, future trials should adopt standardized outcome measures to enhance comparability and enable robust meta-analyses. The proliferation of heterogeneous instruments (eg, multiple versions of SDSCA, MARS, or MMAS) hampers synthesis and reduces external validity. Second, longer-term trials are needed to assess the sustainability of intervention effects, as most included studies had follow-up durations shorter than 12 months. Third, equity-focused research is essential. Few RCTs stratified outcomes by socioeconomic status, age, or eHealth literacy, despite evidence that these factors critically determine effectiveness. Fourth, cost-effectiveness analyses are urgently required, given the resource implications of scaling digital health interventions within health systems. Fifth, future studies should investigate advanced technologies, including artificial intelligence, wearable sensors, and adaptive platforms, to determine whether they can deliver more consistent benefits. Finally, there is a need to conduct studies on self-care in chronic diseases grounded in robust theoretical frameworks, such as the middle-range theory [5,6]. Recent versions of measurement instruments incorporate self-care management as a third component, alongside maintenance and monitoring, which is crucial for developing individuals' ability to manage their health autonomously.

A limitation of this review is the inclusion of self-care outcomes only when measured with a validated instrument to ensure consistency across measures. However, many self-care behaviors can also be evaluated through alternative methods, such as step counts for physical activity or pill counts and prescription records for medication adherence. Future research could therefore expand on these findings by adopting a broader approach to measuring self-care behaviors. In addition, the search was conducted on 4 databases and was limited to articles published in English, which may have led to the exclusion of relevant studies and introduced language and publication bias.

Conclusion

In conclusion, current evidence suggests that digital health interventions may provide some benefit in self-care monitoring

XSL•FO
RenderX

in heart failure, while showing no clear or consistent effects on self-care maintenance or when assessed with alternative instruments. In diabetes, effects across all self-care domains (diet, physical activity, foot care, and glucose monitoring) were small and nonsignificant. In other chronic conditions, results varied. Positive changes were observed in some cardiovascular conditions, chronic hepatitis B, epilepsy, and hypertension, whereas no significant effects were found in COPD and multimorbidity, while mixed results emerged in Parkinson disease. Similarly, no clear overall effect was observed for medication adherence across chronic conditions. Overall, the certainty of the evidence is predominantly low to very low, limiting the ability to draw firm conclusions. Results should therefore be interpreted as suggestive but not definitive, emphasizing the need for larger, methodologically robust trials with standardized outcomes and longer follow-up to clarify the true impact of digital health on self-care and adherence. This review synthesizes evidence across multiple chronic conditions, highlighting that the effects of digital health interventions on self-care remain variable and context-dependent. From a practical perspective, integrating digital tools with human support and adapting interventions to patient characteristics may help enhance their effectiveness in routine care.

Funding

The authors declare that no financial support was received for this work.

Data Availability

All data generated or analyzed during this study are included in this published article.

Authors' Contributions

Conceptualization: JL, DP, AB
Methodology: JL, DP, AB
Formal analysis: JL, FF, MS, FS, MDB, AF, MPP, ST
Data curation: FF, MS, FS, MDB, AF, MPP, ST
Writing – review & editing: DP, FF, MS, FS, MDB, AF, MPP, ST, AB

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA 2020 expanded checklist.
[DOC File , 351 KB - jmir_v28i1e88708_app1.doc]

Multimedia Appendix 2
Search strings.
[DOCX File , 18 KB - jmir_v28i1e88708_app2.docx]

Multimedia Appendix 3
Characteristics of the studies included.
[DOCX File , 61 KB - jmir_v28i1e88708_app3.docx]

Multimedia Appendix 4
Risk of bias.
[DOCX File , 436 KB - jmir_v28i1e88708_app4.docx]

Multimedia Appendix 5
Subgroup meta-analysis.
[DOCX File , 565 KB - jmir_v28i1e88708_app5.docx]

Multimedia Appendix 6
Studies not included in the meta-analysis.
[DOCX File , 20 KB - jmir_v28i1e88708_app6.docx]

References

XSL•FO
RenderX

1. World Health Organization. Noncommunicable Diseases Progress Monitor 2025. Geneva: World Health Organization; 2025.
2. United Nations. World Social Report 2023: Leaving No One Behind In An Ageing World. New York: Department of Economic and Social Affairs Publications; 2023.
3. World Health Organization. Integrated Care for Older People (ICOPE): Guidance for Person-Centred Assessment and Pathways in Primary Care. Geneva: World Health Organization; 2024.
4. World Health Organization. WHO Guideline on Self-Care Interventions for Health and Well-Being. Geneva: World Health Organization; 2021.
5. Riegel B, Dickson VV, Faulkner KM. The situation-specific theory of heart failure self-care: revised and updated. J Cardiovasc Nurs 2016;31(3):226-235. [doi: 10.1097/JCN.0000000000000244] [Medline: 25774844]
6. Riegel B, Jaarsma T, Strömberg A. A middle-range theory of self-care of chronic illness. ANS Adv Nurs Sci 2012;35(3):194-204. [doi: 10.1097/ANS.0b013e318261b1ba] [Medline: 22739426]
7. Riegel B, Westland H, Iovino P, Barelds I, Bruins Slot J, Stawnychy MA, et al. Characteristics of self-care interventions for patients with a chronic condition: a scoping review. Int J Nurs Stud 2021;116:103713 [FREE Full text] [doi: 10.1016/j.ijnurstu.2020.103713] [Medline: 32768137]
8. Lee CS, Westland H, Faulkner KM, Iovino P, Thompson JH, Sexton J, et al. The effectiveness of self-care interventions in chronic illness: a meta-analysis of randomized controlled trials. Int J Nurs Stud 2022;134:104322 [FREE Full text] [doi: 10.1016/j.ijnurstu.2022.104322] [Medline: 35907275]
9. Moynihan R, Sanders S, Michaleff ZA, Scott AM, Clark J, To EJ, et al. Impact of COVID-19 pandemic on utilisation of healthcare services: a systematic review. BMJ Open 2021;11(3):e045343 [FREE Full text] [doi: 10.1136/bmjopen-2020-045343] [Medline: 33727273]
10. World Health Organization. Classification of Digital Interventions, Services and Applications in Health. Geneva: World Health Organization; 2023.
11. Ambrosi E, Mezzalira E, Canzan F, Leardini C, Vita G, Marini G, et al. Effectiveness of digital health interventions for chronic conditions management in European primary care settings: systematic review and meta-analysis. Int J Med Inform 2025;196:105820. [doi: 10.1016/j.ijmedinf.2025.105820] [Medline: 39922064]
12. Cornelius G, Hodgson W, Maguire R, Egan K. Wearable technology, smart home systems, and mobile apps for the self-management of patient outcomes in dementia care: systematic review. J Med Internet Res 2025;27:e65385 [FREE Full text] [doi: 10.2196/65385] [Medline: 40839866]
13. Park Y, Kim EJ, Park S, Lee M. Digital health intervention effect on older adults with chronic diseases living alone: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2025;27:e63168. [doi: 10.2196/63168] [Medline: 40163849]
14. Zhuang M, Hassan II, W Ahmad WMA, Abdul Kadir A, Liu X, Li F, et al. Effectiveness of digital health interventions for chronic obstructive pulmonary disease: systematic review and meta-analysis. J Med Internet Res 2025;27:e76323 [FREE Full text] [doi: 10.2196/76323] [Medline: 40418567]
15. Al Mahmud A, Joachim S, Jayaraman PP, Learmonth C, Tyagi S, Forkan ARM, et al. Digital health interventions to support chronic disease management: systematic scoping review. JMIR Mhealth Uhealth 2026;14:e63742 [FREE Full text] [doi: 10.2196/63742] [Medline: 41533959]
16. Nick JM, Roberts LR, Petersen AB. Effectiveness of telemonitoring on self-care behaviors among community-dwelling adults with heart failure: a quantitative systematic review. JBI Evid Synth 2021;19(10):2659-2694. [doi: 10.11124/JBIES-20-00329] [Medline: 33896907]
17. Neal D, van den Berg F, Planting C, Ettema T, Dijkstra K, Finnema E, et al. Can use of digital technologies by people with dementia improve self-management and social participation? A systematic review of effect studies. J Clin Med 2021;10(4):604 [FREE Full text] [doi: 10.3390/jcm10040604] [Medline: 33562749]
18. Cherrez-Ojeda I, Robles-Velasco K, Osorio MF, Ormaza Vera A, Sarfraz Z, Sarfraz A, et al. A systematic review and meta-analysis of mobile health applications and telemonitoring in atopic dermatitis self-management. Dermatol Ther (Heidelb) 2024;14(7):1787-1798. [doi: 10.1007/s13555-024-01213-0] [Medline: 38909171]
19. Lim DSC, Kwok B, Williams P, Koczwara B. The impact of digital technology on self-management in cancer: systematic review. JMIR Cancer 2023;9:e45145. [doi: 10.2196/45145] [Medline: 37991831]
20. Kim SK, Park SY, Hwang HR, Moon SH, Park JW. Effectiveness of mobile health intervention in medication adherence: a systematic review and meta-analysis. J Med Syst 2025;49(1):13. [doi: 10.1007/s10916-024-02135-2] [Medline: 39821698]
21. Lanke V, Trimm K, Habib B, Tamblyn R. Evaluating the effectiveness of mobile apps on medication adherence for chronic conditions: systematic review and meta-analysis. J Med Internet Res 2025;27:e60822. [doi: 10.2196/60822] [Medline: 40743450]
22. Renzi E, Baccolini V, Migliara G, De Vito C, Gasperini G, Cianciulli A, et al. The impact of eHealth interventions on the improvement of self-care in chronic patients: an overview of systematic reviews. Life (Basel) 2022;12(8):1253 [FREE Full text] [doi: 10.3390/life12081253] [Medline: 36013432]
23. Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M. Cochrane Handbook for Systematic Reviews of Interventions version 6.5. USA: Cochrane; 2024.

XSL•FO
RenderX

24. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
25. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
26. World Health Organization and International Telecommunication Union. National eHealth Strategy Toolkit. Geneva: Switzerland: World Health Organization; 2012.
27. eHealth Action Plan 2012-2020. European Commission. 2012. URL: https://health.ec.europa.eu/publications/ehealth-action-plan-2012-2020_en [accessed 2026-05-27]
28. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
29. Schünemann H, Brożek J, Guyatt G, Oxman A. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. 2013. URL: https://gdt.gradepro.org/app/handbook/handbook.html [accessed 2026-05-27]
30. R Core Team. R: a language and environment for statistical computing. In: R Foundation for Statistical Computing. Vienna, Austria: R Core Team; 2022.
31. Agarwal P, Mukerji G, Desveaux L, Ivers NM, Bhattacharyya O, Hensel JM, et al. Mobile app for improved self-management of type 2 diabetes: multicenter pragmatic randomized controlled trial. JMIR Mhealth Uhealth 2019;7(1):e10321 [FREE Full text] [doi: 10.2196/10321] [Medline: 30632972]
32. Akhu-Zaheya LM, Shiyab WY. The effect of short message system (SMS) reminder on adherence to a healthy diet, medication, and cessation of smoking among adult patients with cardiovascular diseases. Int J Med Inform 2017;98:65-75. [doi: 10.1016/j.ijmedinf.2016.12.003] [Medline: 28034414]
33. Arora S, Peters AL, Burner E, Lam CN, Menchine M. Trial to examine text message-based mHealth in emergency department patients with diabetes (TExT-MED): a randomized controlled trial. Ann Emerg Med 2014;63(6):745-54.e6. [doi: 10.1016/j.annemergmed.2013.10.012] [Medline: 24225332]
34. Babu V, Sylaja P, Soman B, Varma RP, Ms M, Gl G, et al. A randomized controlled trial of medication adherence and management of risk factors for secondary prevention of stroke (MaMoRS) using a smartphone-based application. Int J Stroke 2024;19(6):654-664. [doi: 10.1177/17474930241245612] [Medline: 38533606]
35. Baron JS, Hirani SP, Newman SP. Investigating the behavioural effects of a mobile-phone based home telehealth intervention in people with insulin-requiring diabetes: results of a randomized controlled trial with patient interviews. J Telemed Telecare 2017;23(5):503-512. [doi: 10.1177/1357633X16655911] [Medline: 27377790]
36. Bernal-Jiménez M, Calle G, Gutiérrez Barrios A, Gheorghe LL, Cruz-Cobo C, Trujillo-Garrido N, et al. Effectiveness of an interactive mHealth app (EVITE) in improving lifestyle after a coronary event: randomized controlled trial. JMIR Mhealth Uhealth 2024;12:e48756 [FREE Full text] [doi: 10.2196/48756] [Medline: 38648103]
37. Boyne JJJ, Vrijhoef HJM, Spreeuwenberg M, De Weerd G, Kragten J, Gorgels APM, TEHAF investigators. Effects of tailored telemonitoring on heart failure patients' knowledge, self-care, self-efficacy and adherence: a randomized controlled trial. Eur J Cardiovasc Nurs 2014;13(3):243-252. [doi: 10.1177/1474515113487464] [Medline: 23630403]
38. Bruggmann C, Adjedj J, Sardy S, Muller O, Voirol P, Sadeghipour F. Effects of the interactive web-based video "Mon Coeur, Mon BASIC" on drug adherence of patients with myocardial infarction: randomized controlled trial. J Med Internet Res 2021;23(8):e21938 [FREE Full text] [doi: 10.2196/21938] [Medline: 34459744]
39. Deckwart O, Koehler K, Lezius S, Prescher S, Koehler F, Winkler S. Effects of remote patient management on self-care behaviour in heart failure patients: results from the randomized TIM-HF2 trial. Eur J Cardiovasc Nurs 2023;22(8):786-794. [doi: 10.1093/eurjcn/zvad019] [Medline: 36752782]
40. Desteghe L, Vijgen J, Koopman P, Dilling-Boer D, Schurmans J, Dendale P, et al. Telemonitoring-based feedback improves adherence to non-vitamin K antagonist oral anticoagulants intake in patients with atrial fibrillation. Eur Heart J 2018;39(16):1394-1403. [doi: 10.1093/eurheartj/ehx762] [Medline: 29300888]
41. Dincer B, Bahçecik N. The effect of a mobile application on the foot care of individuals with type 2 diabetes: a randomised controlled study. Health Educ J 2020;80(4):425-437. [doi: 10.1177/0017896920981617]
42. Ding H, Jayasena R, Chen SH, Maiorana A, Dowling A, Layland J, et al. The effects of telemonitoring on patient compliance with self-management recommendations and outcomes of the innovative telemonitoring enhanced care program for chronic heart failure: randomized controlled trial. J Med Internet Res 2020;22(7):e17559 [FREE Full text] [doi: 10.2196/17559] [Medline: 32673222]
43. Erdoğan EG, Örsal Ö. The effect of web-designed education on medication adherence, asthma control and fatigue in patients with asthma: a randomized controlled trial. Int J Nurs Pract 2024;30(6):e13288. [doi: 10.1111/ijn.13288] [Medline: 39031297]
44. FarzanehRad A, Allahbakhshian A, Gholizadeh L, Khalili AF, Hasankhani H. Randomized comparison of the effects of tailored text messaging versus pillbox organizers on medication adherence of heart failure patients. BMC Cardiovasc Disord 2024;24(1):244 [FREE Full text] [doi: 10.1186/s12872-024-03884-1] [Medline: 38724943]
45. Guo M, Meng F, Guo Q, Bai T, Hong Y, Song F, et al. Effectiveness of mHealth management with an implantable glucose sensor and a mobile application among Chinese adults with type 2 diabetes. J Telemed Telecare 2023;29(8):632-640. [doi: 10.1177/1357633X211020261] [Medline: 34152238]

XSL•FO
RenderX

46. Hägglund E, Lyngå P, Frie F, Ullman B, Persson H, Melin M, et al. Patient-centred home-based management of heart failure. Findings from a randomised clinical trial evaluating a tablet computer for self-care, quality of life and effects on knowledge. Scand Cardiovasc J 2015;49(4):193-199 [FREE Full text] [doi: 10.3109/14017431.2015.1035319] [Medline: 25968968]
47. Han CY, Zhang J, Ye XM, Lu JP, Jin HY, Xu WW, et al. Telemedicine-assisted structured self-monitoring of blood glucose in management of T2DM results of a randomized clinical trial. BMC Med Inform Decis Mak 2023;23(1):182 [FREE Full text] [doi: 10.1186/s12911-023-02283-4] [Medline: 37710226]
48. Hartch CE, Dietrich MS, Lancaster BJ, Stolldorf DP, Mulvaney SA. Effects of a medication adherence app among medically underserved adults with chronic illness: a randomized controlled trial. J Behav Med 2024;47(3):389-404 [FREE Full text] [doi: 10.1007/s10865-023-00446-2] [Medline: 38127174]
49. Heisler M, Choi H, Palmisano G, Mase R, Richardson C, Fagerlin A, et al. Comparison of community health worker-led diabetes medication decision-making support for low-income Latino and African American adults with diabetes using e-health tools versus print materials: a randomized, controlled trial. Ann Intern Med 2014;161(10 Suppl):S13-S22 [FREE Full text] [doi: 10.7326/M13-3012] [Medline: 25402398]
50. Hoban MB, Fedor M, Reeder S, Chernick M. The effect of telemonitoring at home on quality of life and self-care behaviors of patients with heart failure. Home Healthc Nurse 2013;31(7):368-377. [doi: 10.1097/NHH.0b013e318291fd56] [Medline: 23632274]
51. Hong PC, Chen KJ, Chang YC, Cheng SM, Chiang HH. Effectiveness of theory-based health information technology interventions on coronary artery disease self-management behavior: a clinical randomized waitlist-controlled trial. J Nurs Scholarsh 2021;53(4):418-427 [FREE Full text] [doi: 10.1111/jnu.12661] [Medline: 33844425]
52. Hsieh HL, Kao CW, Cheng SM, Chang YC. A web-based integrated management program for improving medication adherence and quality of life, and reducing readmission in patients with atrial fibrillation: randomized controlled trial. J Med Internet Res 2021;23(9):e30107 [FREE Full text] [doi: 10.2196/30107] [Medline: 34550084]
53. Jahangard-Rafsanjani Z, Sarayani A, Nosrati M, Saadat N, Rashidian A, Hadjibabaie M, et al. Effect of a community pharmacist-delivered diabetes support program for patients receiving specialty medical care: a randomized controlled trial. Diabetes Educ 2015;41(1):127-135. [doi: 10.1177/0145721714559132] [Medline: 25420946]
54. Jeon JH. Evaluation of a smartphone application for self-care performance of patients with chronic hepatitis B: a randomized controlled trial. Appl Nurs Res 2016;32:182-189. [doi: 10.1016/j.apnr.2016.07.011] [Medline: 27969026]
55. Jiang Y, Koh KWL, Ramachandran HJ, Nguyen HD, Lim DS, Tay YK, et al. The effectiveness of a nurse-led home-based heart failure self-management programme (the HOM-HEMP) for patients with chronic heart failure: a three-arm stratified randomized controlled trial. Int J Nurs Stud 2021;122:104026. [doi: 10.1016/j.ijnurstu.2021.104026] [Medline: 34271265]
56. Kamal AK, Khalid W, Muqeet A, Jamil A, Farhat K, Gillani SRA, et al. Making prescriptions "talk" to stroke and heart attack survivors to improve adherence: results of a randomized clinical trial (The Talking Rx Study). PLoS One 2018;13(12):e0197671 [FREE Full text] [doi: 10.1371/journal.pone.0197671] [Medline: 30571697]
57. Kamal AK, Shaikh Q, Pasha O, Azam I, Islam M, Memon AA, et al. A randomized controlled behavioral intervention trial to improve medication adherence in adult stroke patients with prescription tailored Short Messaging Service (SMS)-SMS4Stroke study. BMC Neurol 2015;15:212. [doi: 10.1186/s12883-015-0471-5] [Medline: 26486857]
58. Kim JY, Wineinger NE, Steinhubl SR. The influence of wireless self-monitoring program on the relationship between patient activation and health behaviors, medication adherence, and blood pressure levels in hypertensive patients: a substudy of a randomized controlled trial. J Med Internet Res 2016;18(6):e116 [FREE Full text] [doi: 10.2196/jmir.5429] [Medline: 27334418]
59. Kirwan M, Vandelanotte C, Fenning A, Duncan MJ. Diabetes self-management smartphone application for adults with type 1 diabetes: randomized controlled trial. J Med Internet Res 2013;15(11):e235 [FREE Full text] [doi: 10.2196/jmir.2588] [Medline: 24225149]
60. Koufopoulos JT, Conner MT, Gardner PH, Kellar I. A web-based and mobile health social support intervention to promote adherence to inhaled asthma medications: randomized controlled trial. J Med Internet Res 2016;18(6):e122 [FREE Full text] [doi: 10.2196/jmir.4963] [Medline: 27298211]
61. Lee H, Park G, Lee D, Khang AR, Lee MJ. Long-term effects of an automated personalized self-care program for patients with type 2 diabetes. Nurs Health Sci 2024;26(4):e70008. [doi: 10.1111/nhs.70008] [Medline: 39648515]
62. Melin M, Hägglund E, Ullman B, Persson H, Hagerman I. Effects of a tablet computer on self-care, quality of life, and knowledge: a randomized clinical trial. J Cardiovasc Nurs 2018;33(4):336-343. [doi: 10.1097/JCN.0000000000000462] [Medline: 29369123]
63. Morawski K, Ghazinouri R, Krumme A, Lauffenburger JC, Lu Z, Durfee E, et al. Association of a smartphone application with medication adherence and blood pressure control: the MediSAFE-BP randomized clinical trial. JAMA Intern Med 2018;178(6):802-809 [FREE Full text] [doi: 10.1001/jamainternmed.2018.0447] [Medline: 29710289]
64. Ni Z, Wu B, Yang Q, Yan LL, Liu C, Shaw RJ. An mHealth intervention to improve medication adherence and health outcomes among patients with coronary heart disease: randomized controlled trial. J Med Internet Res 2022;24(3):e27202 [FREE Full text] [doi: 10.2196/27202] [Medline: 35262490]

XSL•FO
RenderX

65. Park LG, Howie-Esquivel J, Chung ML, Dracup K. A text messaging intervention to promote medication adherence for patients with coronary heart disease: a randomized controlled trial. Patient Educ Couns 2014;94(2):261-268. [doi: 10.1016/j.pec.2013.10.027] [Medline: 24321403]
66. Park SK, Bang CH, Lee SH. Evaluating the effect of a smartphone app-based self-management program for people with COPD: a randomized controlled trial. Appl Nurs Res 2020;52:151231. [doi: 10.1016/j.apnr.2020.151231] [Medline: 31955942]
67. Pfaeffli Dale L, Whittaker R, Jiang Y, Stewart R, Rolleston A, Maddison R. Text message and internet support for coronary heart disease self-management: results from the text4heart randomized controlled trial. J Med Internet Res 2015;17(10):e237 [FREE Full text] [doi: 10.2196/jmir.4944] [Medline: 26490012]
68. Poorcheraghi H, Negarandeh R, Pashaeypoor S, Jorian J. Effect of using a mobile drug management application on medication adherence and hospital readmission among elderly patients with polypharmacy: a randomized controlled trial. BMC Health Serv Res 2023;23(1):1192 [FREE Full text] [doi: 10.1186/s12913-023-10177-4] [Medline: 37919712]
69. Schnall R, Cho H, Mangone A, Pichon A, Jia H. Mobile health technology for improving symptom management in low income persons living with HIV. AIDS Behav 2018;22(10):3373-3383 [FREE Full text] [doi: 10.1007/s10461-017-2014-0] [Medline: 29299790]
70. Si Y, Xiao X, Xia C, Guo J, Hao Q, Mo Q, et al. Optimising epilepsy management with a smartphone application: a randomised controlled trial. Med J Aust 2020;212(6):258-262. [doi: 10.5694/mja2.50520] [Medline: 32092160]
71. Stamenova V, Liang K, Yang R, Engel K, van Lieshout F, Lalingo E, et al. Technology-enabled self-management of chronic obstructive pulmonary disease with or without asynchronous remote monitoring: randomized controlled trial. J Med Internet Res 2020;22(7):e18598 [FREE Full text] [doi: 10.2196/18598] [Medline: 32729843]
72. Sun J, Zhang ZW, Ma YX, Liu W, Wang CY. Application of self-care based on full-course individualized health education in patients with chronic heart failure and its influencing factors. World J Clin Cases 2019;7(16):2165-2175 [FREE Full text] [doi: 10.12998/wjcc.v7.i16.2165] [Medline: 31531312]
73. Vuorinen AL, Leppänen J, Kaijanranta H, Kulju M, Heliö T, van Gils M, et al. Use of home telemonitoring to support multidisciplinary care of heart failure patients in Finland: randomized controlled trial. J Med Internet Res 2014;16(12):e282 [FREE Full text] [doi: 10.2196/jmir.3651] [Medline: 25498992]
74. Ware P, Shah A, Ross HJ, Logan AG, Segal P, Cafazzo JA, et al. Challenges of telemonitoring programs for complex chronic conditions: randomized controlled trial with an embedded qualitative study. J Med Internet Res 2022;24(1):e31754 [FREE Full text] [doi: 10.2196/31754] [Medline: 35080502]
75. Wonggom P, Nolan P, Clark RA, Barry T, Burdeniuk C, Nesbitt K, et al. Effectiveness of an avatar educational application for improving heart failure patients' knowledge and self-care behaviors: a pragmatic randomized controlled trial. J Adv Nurs 2020;76(9):2401-2415. [doi: 10.1111/jan.14414] [Medline: 32395836]
76. Xu W, Huang X, Lin Q, Wu T, Guan C, Lv M, et al. Application of Alfalfa app in the management of oral anticoagulation in patients with atrial fibrillation: a multicenter randomized controlled trial. BMC Med Inform Decis Mak 2024;24(1):294 [FREE Full text] [doi: 10.1186/s12911-024-02701-1] [Medline: 39385171]
77. Ye H, Lin L, Zhong D, Chen P, He X, Luo Z, et al. The impact of telehealth education on self-management in patients with coexisting type 2 diabetes mellitus and hypertension: a 26-week randomized controlled trial. J Endocrinol Invest 2024;47(9):2361-2369. [doi: 10.1007/s40618-024-02310-9] [Medline: 38351401]
78. Hwang M, Lee S, Park GE, Park Y. Effectiveness of a digital health coaching self-management program for older adults living alone with multiple chronic conditions: a randomized controlled trial. Geriatr Nurs 2025;65:103509 [FREE Full text] [doi: 10.1016/j.gerinurse.2025.103509] [Medline: 40633238]
79. Kitsiou S, Gerber BS, Buchholz SW, Kansal MM, Sun J, Pressler SJ. Patient-centered mHealth intervention to improve self-care in patients with chronic heart failure: phase 1 randomized controlled trial. J Med Internet Res 2025;27:e55586 [FREE Full text] [doi: 10.2196/55586] [Medline: 39813671]
80. Lee J, Yoo S, Kim Y, Kim E, Park H, Sohn YH, et al. Effect of the Yon PD app on the management of self-care in people with Parkinson disease: randomized controlled trial. J Med Internet Res 2025;27:e62822 [FREE Full text] [doi: 10.2196/62822] [Medline: 40173440]
81. Lippke S, Korte L, Kumar VA, Fach A, Ratz T. Health behavior and disease self-management indicators in patients with cardiovascular diseases using a health app: findings from an RCT. AIMS Public Health 2025;12(1):233-258 [FREE Full text] [doi: 10.3934/publichealth.2025015] [Medline: 40248407]
82. Magnani JW, Lalama CM, Abebe KZ, Ferry D, Rollman BL, Lancet MQ, et al. A mobile relational agent to enhance atrial fibrillation self-care: primary and secondary outcomes of a randomized controlled trial. Am Heart J 2025;290:115-128. [doi: 10.1016/j.ahj.2025.06.009] [Medline: 40545207]
83. Meyer B, Riepenhausen A, Betz LT, Jauch-Chara K, Reshetnik A. Internet-based digital intervention to support the self-management of hypertension compared to usual care: results of the HALCYON randomized controlled trial. BMC Cardiovasc Disord 2025;25(1):256 [FREE Full text] [doi: 10.1186/s12872-025-04698-5] [Medline: 40186109]
84. Silberman J, Sarlati S, Harris B, Lenyoun H, Kaur M, Wagner BG, et al. A digital asthma self-management program for adults: a randomized clinical trial. JAMA Netw Open 2025;8(7):e2521438 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.21438] [Medline: 40674052]

XSL•FO
RenderX

85. Yildirim Keskin A, Özpancar Şolpan N, Değirmenci H. The effect of mobile application follow-up on treatment compliance and self-care management in patients with hypertension: randomized controlled trial. Public Health Nurs 2025;42(1):275-285. [doi: 10.1111/phn.13476] [Medline: 39492655]
86. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
87. Liu F, Li J, Li X, Yang Z, Wang W, Zhao L, et al. Efficacy of telemedicine intervention in the self-management of patients with type 2 diabetes: a systematic review and meta-analysis. Front Public Health 2024;12:1405770 [FREE Full text] [doi: 10.3389/fpubh.2024.1405770] [Medline: 38835608]
88. Shrivastava TP, Goswami S, Gupta R, Goyal RK. Mobile app interventions to improve medication adherence among type 2 diabetes mellitus patients: a systematic review of clinical trials. J Diabetes Sci Technol 2023;17(2):458-466 [FREE Full text] [doi: 10.1177/19322968211060060] [Medline: 34861793]
89. Lu H, Xu S, Yang J, Zhou Y, Gu Z. Effect of telemedicine on self-care in patients with heart failure: a meta-analysis of randomized controlled trials. Eur J Cardiovasc Nurs 2026:zvaf212. [doi: 10.1093/eurjcn/zvaf212] [Medline: 41544652]
90. Ni YX, Liu XH, He L, Wen Y, You GY. Mobile application-based interventions for people with heart failure: a systematic review and meta-analysis. J Nurs Manag 2024;2024:6859795. [doi: 10.1155/2024/6859795] [Medline: 40224789]
91. Zeng Z, Wu T, Lv M, Qian J, Chen M, Fang Z, et al. Impact of mobile health and telehealth technology on medication adherence of stroke patients: a systematic review and meta-analysis of randomized controlled trials. Int J Clin Pharm 2022;44(1):4-14. [doi: 10.1007/s11096-021-01351-x] [Medline: 34800254]
92. Riegel B, De Maria M, Barbaranelli C, Luciani M, Ausili D, Dickson VV, et al. Measuring self-care: a description of the family of disease-specific and generic instruments based on the theory of self-care of chronic illness. J Cardiovasc Nurs 2024;40(2):103-113. [doi: 10.1097/jcn.0000000000001146]
93. Eiselt AK, Kirkendall S, Xiong E, Langner D, Goldfarb M. Achieving clinically meaningful outcomes in digital health: a six-step, cyclical precision engagement framework (ENGAGE). Front Digit Health 2025;7:1713334 [FREE Full text] [doi: 10.3389/fdgth.2025.1713334] [Medline: 41608160]
94. Longhini J, Canzan F, Zambiasi P, Toccoli S, Gios L, Del Greco M, et al. A nurse-led model of care with 2 telemonitoring to manage patients with heart failure in primary health care: a mixed-method feasibility study [response to letter]. Patient Prefer Adherence 2023;17:3305-3306 [FREE Full text] [doi: 10.2147/PPA.S451932] [Medline: 38106369]
95. Turnbull S, Cabral C, Hay A, Lucas PJ. Health equity in the effectiveness of web-based health interventions for the self-care of people with chronic health conditions: systematic review. J Med Internet Res 2020;22(6):e17849 [FREE Full text] [doi: 10.2196/17849] [Medline: 32459632]
96. Ge H, Li J, Hu H, Feng T, Wu X. Digital exclusion in older adults: a scoping review. Int J Nurs Stud 2025;168:105082. [doi: 10.1016/j.ijnurstu.2025.105082] [Medline: 40279791]
97. Hepburn J, Williams L, McCann L. Barriers to and facilitators of digital health technology adoption among older adults with chronic diseases: updated systematic review. JMIR Aging 2025;8:e80000 [FREE Full text] [doi: 10.2196/80000] [Medline: 40934502]
98. Birati Y, Tzemah-Shahar R. Barriers to digital health adoption in older adults: scoping review informed by innovation resistance theory. J Med Internet Res 2026;28:e75591 [FREE Full text] [doi: 10.2196/75591] [Medline: 41627850]

Abbreviations

ARMS: Adherence to Refills and Medication Scale
COPD: chronic obstructive pulmonary disease
EHFScB: European Heart Failure Self-Care Behaviour Scale
GRADE: Grading of Recommendations Assessment, Development and Evaluation
MARS: Medication Adherence Rating Scale
mHealth: mobile health
MMAS: Morisky Medication Adherence Scale
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
PROSPERO: International Prospective Register of Systematic Reviews
RCT: randomized controlled trial
SCHFI: Self-Care of Heart Failure Index
SDSCA: Summary of Diabetes Self-Care Activities
SMD: standardized mean difference
WHO: World Health Organization

XSL•FO
RenderX

Edited by S Brini; submitted 02.Dec.2025; peer-reviewed by S Strandberg, C Escoffery; comments to author 27.Feb.2026; accepted 19.Apr.2026; published 09.Jun.2026.

Please cite as:
Longhini J, Pedrotti D, Foladori F, Stedile M, Stefani F, Dal Ben M, Froner A, Proietti Pesci M, Toccoli S, Brugnolli A
Effectiveness of Digital Health Interventions to Improve Self-Care in Patients With Chronic Diseases: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e88708
URL: https://www.jmir.org/2026/1/e88708
doi:10.2196/88708
PMID:

©Jessica Longhini, Daniel Pedrotti, Federica Foladori, Melania Stedile, Francesca Stefani, Michela Dal Ben, Alessandro Froner, Marta Proietti Pesci, Stefano Toccoli, Anna Brugnolli. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Navigating Diabetes Management in the Digital Era: Scoping Review of Online Health Information-Seeking Behavior

Pin-Heng Tiao[1], PharmD; Hsun-Yu Chan[2], PhD; Yen-Ming Huang[1,3,4], BPharm, MS, PhD

[1]Graduate Institute of Clinical Pharmacy, College of Medicine, National Taiwan University, Taipei City, Taiwan

[2]Department of Industrial Education, National Taiwan Normal University, Taipei City, Taiwan

[3]School of Pharmacy, College of Medicine, National Taiwan University, Taipei City, Taiwan

[4]Department of Pharmacy, National Taiwan University Hospital, Taipei City, Taiwan

Corresponding Author:
Yen-Ming Huang, BPharm, MS, PhD
Graduate Institute of Clinical Pharmacy
College of Medicine
National Taiwan University
Number 33, Linsen South Road, Zhongzheng District,
Taipei City, 100025
Taiwan
Phone: 886 233668784
Fax: 886 233668784
Email: yenming927@ntu.edu.tw

Abstract

Background: Online health information seeking (OHIS) has become a central component of chronic disease management within an increasingly interactive, algorithm-mediated digital ecosystem. For individuals with diabetes, ongoing self-management demands create sustained needs for accessible, actionable health information. Although prior reviews have described general information-seeking behaviors, few have integrated technological evolution, multilevel determinants, and equity considerations specific to diabetes.

Objective: This scoping review maps patterns of OHIS among individuals with diabetes, identifies the types of information sought, synthesizes the multilevel determinants of OHIS, and explores temporal shifts across major phases of digital health development.

Methods: This scoping review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) and Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension (PRISMA-S) reporting guidelines and was guided by the Sample, Phenomenon of Interest, Design, Evaluation, Research type framework. Five electronic databases (PubMed, Scopus, Web of Science, CINAHL, and Embase) were systematically searched for English-language empirical studies from inception to May 4, 2026. Eligible studies included empirical research investigating OHIS behaviors among individuals with type 1 diabetes, type 2 diabetes, or gestational diabetes. Data were extracted using a standardized charting form and synthesized descriptively. Determinants were organized according to the Social Ecological Model, and qualitative findings were analyzed using content analysis. Studies were stratified into 3 periods reflecting shifts in digital infrastructure: early web environments (2002-2010), expansion of social media and mobile technologies (2011-2018), and integrated digital and artificial intelligence (AI)–enabled ecosystems (2019-2025).

Results: Eighty-one studies from 32 countries met the inclusion criteria. The use of digital sources diversified over time. Early studies emphasized search engines and institutional websites, whereas later studies increasingly reported engagement with social media platforms and online communities. Mobile health apps and generative AI chatbots appeared in recent publications, although evidence on AI use remained limited. The most frequently sought content included self-management and lifestyle guidance, general diabetes knowledge, and treatment-related information. Determinants of OHIS operated across multiple levels. At the individual level, younger age, greater educational attainment, higher income, and better eHealth literacy were associated with increased engagement, while psychological factors such as perceived knowledge gaps and a desire for autonomy motivated searching. Interpersonal influences included peer support and clinician communication. Organizational and environmental factors encompassed health care access, digital infrastructure, information quality, and platform characteristics. Persistent disparities were observed among older adults and socioeconomically disadvantaged groups.

XSL•FO
RenderX

Conclusions: This review synthesizes OHIS among individuals with diabetes through the lenses of technological evolution, multilevel determinants, and digital health equity. Unlike previous reviews that focused on specific platforms or general information-seeking behaviors, it maps the transition from web-based resources to social media and emerging AI-enabled ecosystems. This temporally informed synthesis advances understanding of digital engagement in diabetes self-management, identifies key evidence gaps, and informs clinical, organizational, and policy strategies to promote equitable access to trustworthy online health information.

(J Med Internet Res 2026;28:e82081) doi:10.2196/82081

KEYWORDS

behavior; diabetes; digital; health; information; online; seeking

Introduction

Background

The digital transformation of health care has fundamentally reshaped how individuals access, interpret, and apply health information. The internet has shifted from a supplementary resource to a primary channel for health engagement, with online health information seeking (OHIS) becoming a normative component of health behavior in many regions of the world [1,2]. According to Eurostat [3], approximately 60% of individuals aged 16-74 years in European countries searched for health-related information online in 2025, reflecting a 15% increase over the past decade. Similar or higher rates have been reported in parts of Asia [4]. Reliance on digital health information increased further during the COVID-19 pandemic, when in-person access to care was constrained [5,6].

However, OHIS is no longer confined to static websites accessed through search engines. The digital health ecosystem has undergone rapid structural evolution. Early forms of OHIS centered on retrieving information from institutional websites and web-based forums [7]. The subsequent expansion of social media platforms introduced participatory environments characterized by peer exchange, user-generated content, and the networked dissemination of health narratives [8,9]. More recently, generative artificial intelligence (AI) chatbots have emerged as conversational intermediaries capable of producing tailored responses to health queries in real time [10-12]. These shifts represent not merely technological upgrades but fundamental transformations in the epistemic architecture of health information. Health information environments have evolved from retrieval-based systems to interactive, algorithmically mediated knowledge ecosystems. These environments are also becoming increasingly opaque in how information is generated, prioritized, and presented.

Within this evolving digital infrastructure, OHIS functions as both an opportunity and a risk. On the one hand, digital platforms may enhance patient activation, support shared decision-making, and facilitate chronic disease self-management [13]. On the other hand, they expose individuals to information overload, variable credibility, commercial influence, and algorithmic filtering that shapes information visibility [14,15]. The rise of generative AI further complicates this landscape by introducing systems capable of producing fluent but potentially unverifiable health guidance [16]. As such, OHIS should be understood as a cognitively demanding and socially situated process embedded within digitally structured environments rather than merely as a behavioral act [17].

Diabetes provides a relevant context for exploring OHIS. The International Diabetes Federation estimates that diabetes will affect 853 million adults worldwide by 2050, and global diabetes-related expenditures are projected to approach US $1 trillion [18]. Beyond its scale, diabetes is characterized by sustained and complex self-management demands. Effective management requires continuous engagement in medication use, blood glucose monitoring, dietary regulation, and physical activity planning [19]. These activities are not episodic but ongoing, requiring patients to repeatedly search for, interpret, evaluate, and apply health information in daily life [20]. As diabetes self-management occurs largely outside clinical settings [21], individuals frequently turn to digital environments to bridge informational gaps between consultations [22]. More than 60% of individuals with diabetes in the United States engage in OHIS [23]. OHIS, therefore, represents a central component of contemporary diabetes self-care. Within patient-centered care models that emphasize shared decision-making and patient activation [24], digital health information has become intertwined with routine disease management.

Yet, digital engagement in health contexts remains unevenly distributed across populations. The digital divide involves more than internet access alone. Differences in digital competencies, health literacy, technological resources, and socioeconomic conditions also play important roles [25]. These structural differences influence how individuals access, evaluate, and use online health information. For populations already at elevated risk of diabetes-related complications, including older adults and individuals with lower income or educational attainment, limited digital capacity may intensify existing social and health inequities [26]. Digital health environments, therefore, represent sociotechnical spaces that can expand patient empowerment while also reproducing disparities in knowledge access, participation, and health outcomes [27].

Despite expanding research on health information seeking, existing reviews have not comprehensively addressed OHIS among individuals with diabetes within this rapidly evolving digital context. Prior syntheses have examined general information-seeking behaviors [28], aggregated multiple chronic conditions without isolating diabetes-specific patterns [29], or focused narrowly on online communities [28]. These approaches provide important descriptive insights but do not integrate multilevel determinants, technological evolution, and equity

XSL•FO
RenderX

considerations within a unified framework. Moreover, most predate the widespread integration of social media platforms and generative AI systems into routine health information practices.

Objectives

Given the rapid transformation of digital infrastructure and the central role of self-management in diabetes care, an updated, theoretically informed synthesis is needed. This scoping review addressed this gap by mapping the evidence on OHIS behaviors among individuals with diabetes. It focused on the sources of information used, the types of information sought, and the multilevel factors influencing these behaviors. These patterns were examined across evolving digital health environments. In addition, a secondary descriptive analysis was conducted to explore temporal shifts in OHIS patterns across technological eras. By situating OHIS within the broader digital health ecosystem, this review also sought to advance conceptual understanding of how individuals with diabetes engage with increasingly algorithm-mediated information environments. The findings may inform strategies that promote equitable and evidence-informed digital health engagement.

Methods

Overview

Given the heterogeneity of study designs, populations, and outcome measures in this field, a scoping review methodology was adopted to systematically map the existing evidence, clarify key concepts, and identify knowledge gaps. The review was conducted following the framework proposed by Arksey and O'Malley [30] and further informed by methodological refinements from the Joanna Briggs Institute guidance for scoping reviews [31]. The research questions were structured using the Sample, Phenomenon of Interest, Design, Evaluation, Research type (SPIDER) framework (Multimedia Appendix 1) [32]. Reporting followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) [33] and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension (PRISMA-S) [34] guidelines to enhance transparency and reproducibility. The completed PRISMA-ScR and PRISMA-S checklists are provided in Multimedia Appendices 2 and 3, respectively.

Protocol and Registration

The review protocol was not registered before commencement.

Eligibility Criteria

Detailed inclusion and exclusion criteria are presented in Textbox 1. We included empirical studies published in English that described OHIS behaviors among individuals with diabetes. OHIS was defined as the active effort to obtain health-related information from digital sources and platforms to address health information needs [35]. Studies were excluded if they did not directly assess OHIS behaviors, focused solely on health literacy or eHealth literacy without a clear information-seeking component, were nonempirical, or lacked full-text availability.

Textbox 1. Eligibility criteria for the scoping review.

1. Inclusion criteria

- Individuals with diabetes of any age, including type 1 diabetes, type 2 diabetes, or gestational diabetes
- Studies with mixed clinical samples only if diabetes-specific data are reported separately or can be extracted
- Studies describing online health information-seeking behaviors directly
- Empirical primary studies (quantitative, qualitative, or mixed methods) with full-text available
- Articles published in English
- Studies published from database inception to May 2026

2. Exclusion criteria

- Studies did not directly assess online health information-seeking behaviors (eg, content or text analysis of user-generated posts, or trend analysis of online data)
- Studies addressing health/eHealth literacy only without exploring seeking behaviors
- Studies were not full-text articles (eg, abstracts, posters, letters)
- Studies were nonempirical in nature (eg, reviews, protocols, commentaries, editorials)

Information Sources and Search Strategy

A comprehensive literature search was conducted across 5 electronic databases: PubMed, Scopus, Web of Science, CINAHL, and Embase. Each database was searched independently through its web-based interface. Retrieved records were imported into EndNote (Clarivate Plc) for reference management and duplicate removal before screening. Citation tracking was additionally performed through backward reference searching of included studies using Google Scholar (Google LLC). No additional sources, such as trial registries, conference proceedings, organizational websites, or gray literature databases, were searched. Study authors were not contacted, and no supplementary search methods were used beyond those described.

The initial search was conducted in April 2025 and updated on May 4, 2026, to capture newly published studies. The search strategy was developed using the SPIDER framework and

XSL•FO
RenderX

centered on 3 core concepts: diabetes, health information-seeking behavior, and online or digital contexts. Search terms were adapted from previous reviews [13,28] and refined through team discussion. Boolean operators were applied to combine keywords, and searches were performed across database-specific fields (eg, title, abstract, keyword, and topic fields). In PubMed, MeSH (Medical Subject Headings) were additionally used. No limits or filters were imposed. The full reproducible search strategies and the number of records retrieved from each database are provided in Multimedia Appendix 4.

Selection of Sources of Evidence

After duplicate removal, titles and abstracts were screened by 1 reviewer (PHT) according to the predefined eligibility criteria (Textbox 1). Screening decisions were independently verified by a second reviewer (YMH). Discrepancies were resolved through discussion, with a third reviewer (HYC) consulted as needed. Full-text articles deemed potentially eligible were subsequently retrieved and assessed by the review team to confirm final inclusion.

Data Charting Process and Data Items

Data charting was conducted by 1 reviewer (PHT) and cross-verified by other team members to ensure consistency and accuracy. A predefined extraction form, informed by prior literature [13,28], was used to guide data collection. Extracted variables included publication year, author, country, study design, sample characteristics, type of diabetes, disease duration, reported online platforms, types of information sought, and determinants of OHIS. Data were organized using Microsoft Excel.

Synthesis of Results

Descriptive mapping was performed for online platforms and types of information sought; frequencies and percentages were calculated to summarize study characteristics. To identify determinants of OHIS, a mixed deductive-inductive analytic approach was adopted [36]. A deductive content analysis was conducted using the Social Ecological Model as a guiding framework to categorize determinants at the individual, interpersonal, organizational, and digital-environment levels [37]. Codes were iteratively refined through team discussion. To enhance interpretive depth, qualitative findings related to patient motivations, emotional responses, perceived dilemmas, and coping strategies were further synthesized. Themes were generated inductively and organized into higher-order domains that captured recurring psychological and behavioral processes underlying OHIS.

To explore the impact of technological evolution, studies were stratified into 3 periods reflecting major phases of digital health development: early web-based environments (2002-2010), the expansion of social media and mobile technologies (2011-2018), and the integrated digital and AI-enabled ecosystem (2019-2025). Study characteristics, platforms, and determinants were compared descriptively across periods to identify temporal shifts in OHIS patterns.

Critical Appraisal of Individual Sources of Evidence

Consistent with the purpose of a scoping review, this study aimed to map the breadth and characteristics of the available evidence rather than evaluate intervention effectiveness [30]. Therefore, no formal quality appraisal or risk of bias assessment was conducted.

Results

Overview of the Included Studies

The initial search yielded 4589 records, comprising 1524 (33.21%) from Scopus, 924 (20.14%) from Embase, 871 (18.98%) from CINAHL, 691 (15.06%) from Web of Science, and 579 (12.62%) from PubMed. After the removal of 1933 (42.12%) duplicates, 2656 (57.88%) records underwent title and abstract screening. Of these, 161 (3.51%) articles were assessed for full-text eligibility, and 76 (1.66%) studies met the inclusion criteria [23,38-112]. Citation tracking identified an additional 5 eligible studies [113-117], resulting in a final sample of 81 included studies for synthesis [23,38-117] (Figure 1).

XSL•FO
RenderX

Figure 1. Flow diagram of the literature search process.

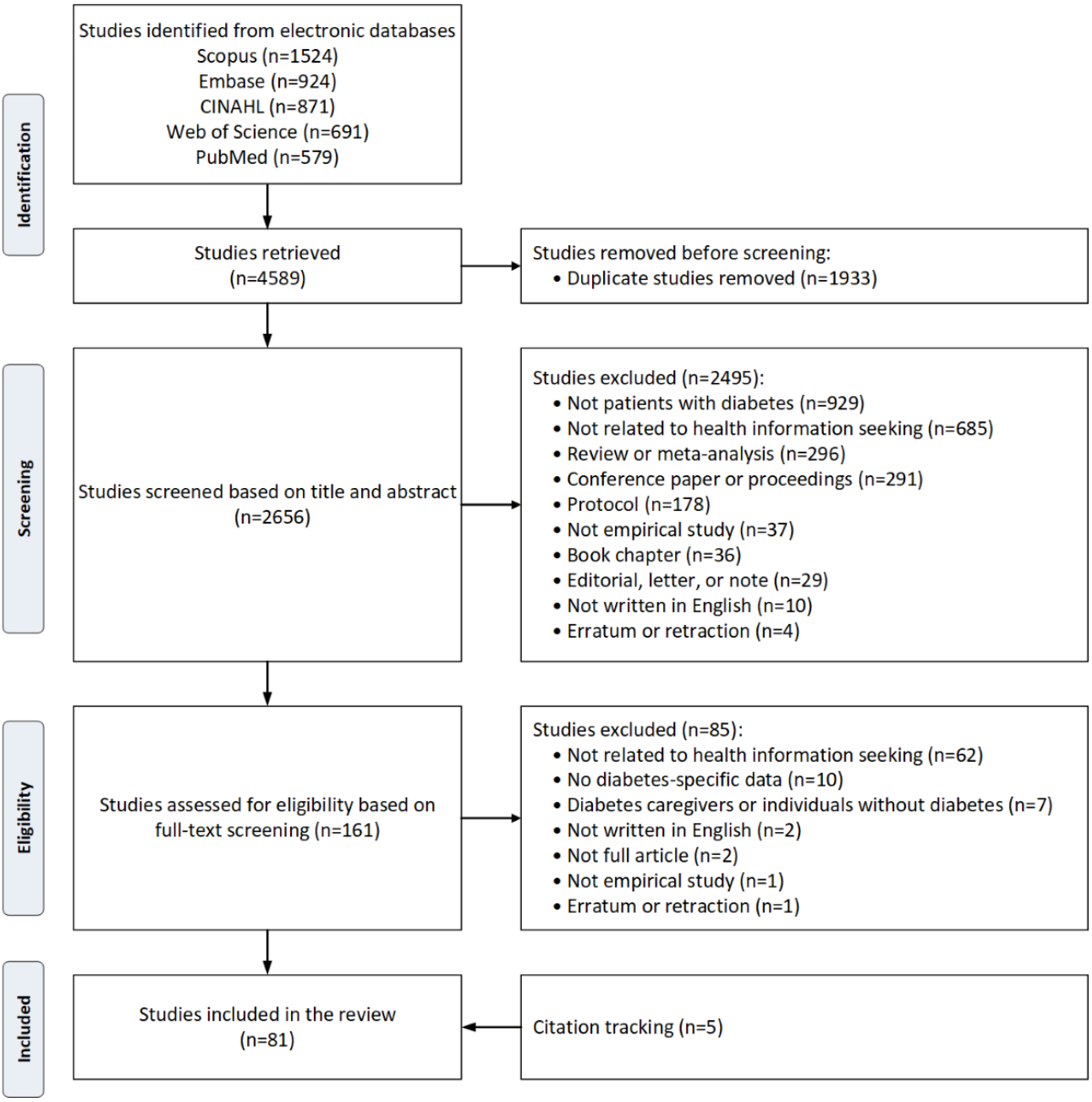

The included studies were conducted across 32 countries spanning 5 continents. The largest number originated from the United States (n=21, 26%), followed by the United Kingdom (n=7, 9%) and Australia (n=5, 6%; Figure 2). Studies were published between 2002 and 2025, with a marked increase in publication frequency over time. A total of 12 (15%) studies [38-49] were published before 2010, 29 (36%) [50-76,113,114] between 2011 and 2020, and 40 (49%) [23,77-112,115-117] between 2021 and 2025, indicating growing scholarly attention to digital health engagement in diabetes. Quantitative cross-sectional surveys were the most common study design (n=42, 52%), followed by qualitative studies (n=30, 37%) and mixed methods studies (n=9, 11%). Qualitative approaches became more prevalent after 2011, whereas mixed methods designs appeared with increasing frequency but remained relatively uncommon across publication periods (Figure 3).

XSL•FO
RenderX

Figure 2. Geographic distribution of studies included in the review.

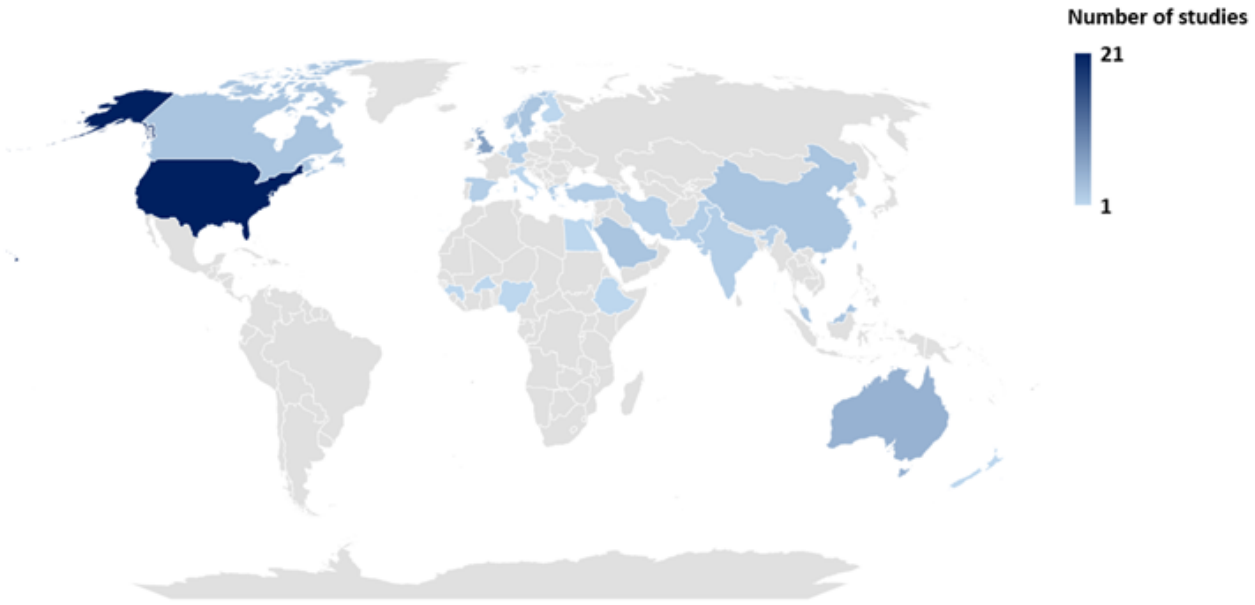

Figure 3. Publication period and study methodology of included studies.

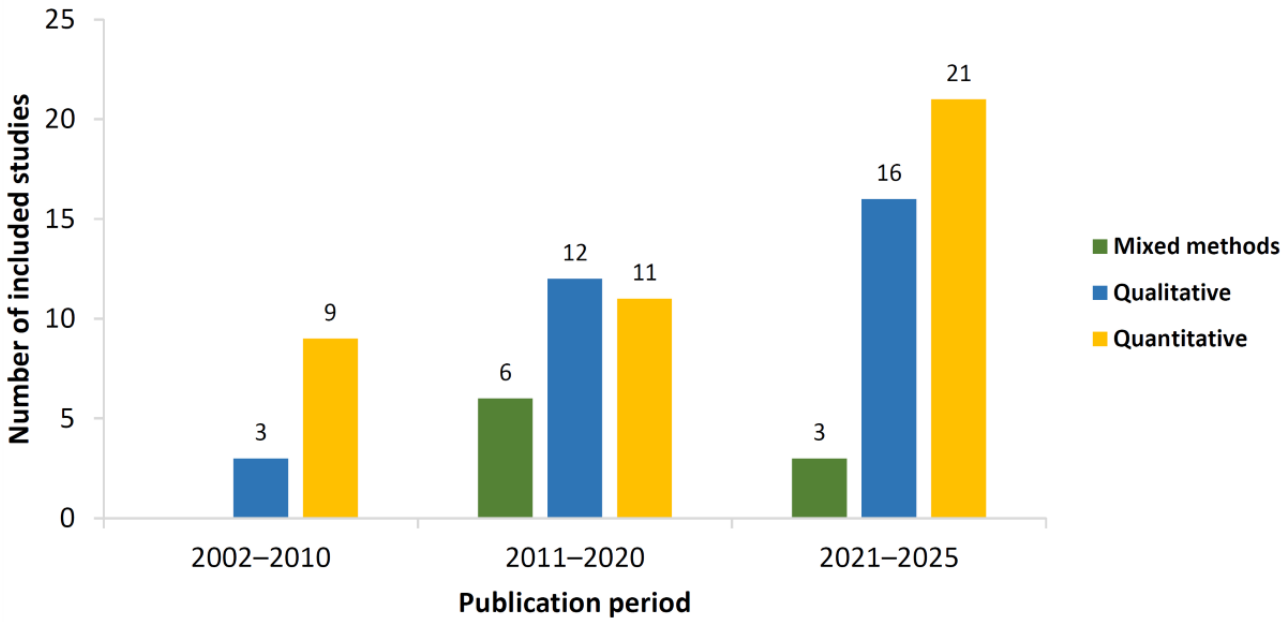

Sample sizes ranged from 8 to 7999 participants. Among the 62 (77%) studies that reported diabetes type, 31 (50%) focused on type 2 diabetes, 8 (13%) on type 1 diabetes, 17 (27%) included both type 1 and type 2 diabetes, and 6 (10%) targeted gestational diabetes. Disease duration was reported in 38 (61%) studies. Detailed study characteristics are presented in Table 1.

XSL•FO
RenderX

Table 1. Characteristics of included studies, including study design, population, and research focus.

Study	Country	Sample size, n	Age (years)	Types of diabetes (duration of diabetes diagnosis)	Study design (data collection method)
Giménez-Pérez et al [38]	Spain	244	Mean 34.3 (SD 12.9)	Type 1 diabetes (mean 11.5 years; SD 9.0)	Quantitative (survey)
Millard and Fintak [39]	United States	1387	Adults (>18)	Not defined	Quantitative (survey)
Eriksson-Backa [40]	Finland	17	Mean 27	Not defined	Qualitative (interview)
Wagner et al [41]	United States	263	>21	Not defined	Quantitative (survey)
Jackson et al [42]	United States	457	Mean 57 (SD 11.0)	Type 2 diabetes	Quantitative (survey)
Nordfeldt et al [43]	Sweden	90	Median 14 (range 5-20)	Type 1 diabetes (≥1.5 years)	Quantitative (survey)
Robertson et al [44]	United Kingdom	70	Range 16-79	Type 1 diabetes and type 2 diabetes (53% were ≥5 years)	Quantitative (survey)
Carlson et al [45]	United States	110	49% aged ≥60	Not defined	Quantitative (survey)
Stevenson et al [46]	United Kingdom	13	≥30	Not defined	Qualitative (focus group)
Cho et al [47]	United States	201	Mean 58.9 (SD 10.4)	Type 2 diabetes	Quantitative (survey)
Longo et al [48]	United States	46	Mean 61 (range 48-77)	Type 1 diabetes and type 2 diabetes (56.5% were ≥5 years)	Qualitative (focus group)
Plotnikoff et al [49]	Canada	244	Mean 60.93 (SD 11.23)	Type 2 diabetes (mean 8.98 years)	Quantitative (survey)
Chisolm et al [50]	United States	223	Range 13-18	Not defined	Quantitative (survey)
Mayberry et al [51]	United States	75	Mean 56.9 (SD 8.8)	Type 2 diabetes	Mixed methods (survey and focus group)
Shaw and Johnson [52]	United States	57	≥21	Type 2 diabetes (mean 7 years)	Quantitative (survey)
Hyman et al [113]	Canada	184	Mean 51.5	Type 2 diabetes	Quantitative (survey)
Janeice Morgan and Trauth [53]	United States	30	Adults aged >18	Type 1 diabetes and type 2 diabetes (≥1 year)	Qualitative (interview)
Meyfroidt et al [54]	Belgium	21	Mean 60 (range 41-85)	Type 2 diabetes	Qualitative (focus group)
Nordfeldt et al [55]	Sweden	24	Range 10-17	Type 1 diabetes	Qualitative (focus group)
Connolly and Crosby [56]	United States	25	Mean 54 (SD 11.6)	Not defined	Qualitative (focus group)
Wiley et al [57]	Australia	150	Range 18-35	Type 1 diabetes	Multimethod (survey and focus group)
Garfield et al [58]	United States	1714	Adults (>18)	Type 1 diabetes and type 2 diabetes	Quantitative (survey)
Jamal et al [59]	Saudi Arabia	344	Mean 53.5 (SD 13.8, range 16-84)	T2D (54.6% were ≥10 years)	Quantitative (survey)
Kalantzi et al [60]	Greece	203	adults (>18); with 57.6% aged ≥60	Type 1 diabetes and type 2 diabetes (48.8% were ≥10 years)	Quantitative (survey)
Lui et al [61]	Australia	3652	Mean 63	Type 2 diabetes (median: 6 years)	Quantitative (survey)
Morgan et al [62]	United States	30	Range 22-72	Type 1 diabetes and type 2 diabetes	Qualitative (interview)
Brady et al [63]	United Kingdom	21	Mean 50	Type 1 diabetes and type 2 diabetes	Qualitative (interview)
Fergie et al [64]	United Kingdom	20	Mean 25.7 (SD 3.7, range 18-30)	Not defined (most were ≥10 years)	Qualitative (interview)
O'Kane et al [65]	United Kingdom	32	Not reported	Type 1 diabetes and type 2 diabetes	Qualitative (interview)

XSL•FO
RenderX

Study	Country	Sample size, n	Age (years)	Types of diabetes (duration of diabetes diagnosis)	Study design (data collection method)
Weymann et al [66]	Germany	10 (interview); 178 (questionnaire)	Mean 62 (SD 10.8, range 36-86)	Type 2 diabetes (mean 11.8 years, SD 10.1)	Mixed methods (survey and interview)
Aponte and Nokes [67]	United States	20	Mean 74 (SD 5.6)	Type 2 diabetes (mean 16.7 years; SD 6.8)	Mixed methods (survey and focus group)
Brady et al [68]	United Kingdom	21	Mean 52	Type 1 diabetes and type 2 diabetes	Qualitative (interview)
Mathiesen et al [69]	Denmark	12	Mean 61 (range 38-70)	Type 2 diabetes (mean 12 years)	Qualitative (interview)
St. Jean 2017 [114]	United States	32	Mean 53.4 (SD 10.6)	Type 2 diabetes (mean 86.7 months; SD 116.0)	Mixed methods (survey and interview)
Vitger et al [70]	Denmark	22	Mean 60	Type 2 diabetes (mean 9 years)	Quantitative (survey)
Leelavathi [71]	Malaysia	380	Mean 60.7 (SD 8.1)	Type 2 diabetes (mean 11.3 years; SD 7.7)	Quantitative (survey)
Martis et al [72]	New Zealand	60	Mean 33 (SD 4.5)	Gestational diabetes	Qualitative(interview)
Dayyani et al [73]	Denmark	11	Range 24-42	Gestational diabetes	Qualitative (interview)
Maneze et al [74]	Australia	18	Mean 69.6 (SD 9.6)	Type 2 diabetes (mean 14.3 years; SD 9.8)	Mixed methods (qualitative phase: interview)
Lee et al [75]	Malaysia and Singapore	673	Adults (>18); with 85% aged ≥41	Type 2 diabetes	Quantitative (survey)
Zhang et al [76]	Singapore	60	73% aged ≥51	Type 1 diabetes and type 2 diabetes	Quantitative (survey)
Abdel Nasser et al [77]	Saudi Arabia	2228	44.6% aged between 45 and 55	Type 2 diabetes (60.5% were ≤5 years)	Quantitative (survey)
Alvarez-Perez et al [78]	Spain and Italy	28	Range 14-75	Type 1 diabetes and type 2 diabetes	Qualitative (focus group)
Dehnavi et al [79]	Iran	8	37.5% aged between 30 and 39	Not defined	Qualitative (interview)
Edwards et al [80]	United Kingdom	10	Not reported	Gestational diabetes	Qualitative (interview)
Guo et al [81]	Taiwan	249	Mean 44.6 (SD 11.0, range 20-65)	Type 2 diabetes (mean 6.1 years, SD 5.6)	Quantitative (survey)
Hughes et al [82]	United States	95	Mean 26.8 (SD 7.2)	Type 1 diabetes (12.3 years, SD 9.2)	Qualitative (survey)
Kostagiolas et al [84]	Greece	106	56.6% were >60	Type 2 diabetes (62.2% were ≥10 years)	Quantitative (survey)
Mansour [85]	Egypt	311	Adults (>18); with 69.5% aged between 46 and 60	Type 2 diabetes (66.5% were ≥10 years)	Quantitative (survey)
Mengiste et al [115]	Ethiopia	423 (survey); 14 (interview)	57.7% aged between 18 and 40	Not defined	Multimethod (survey and interview)
Nazir and Soroya [86]	Pakistan	100	Not reported	Not defined	Quantitative (survey)
Siti Zuhaida et al [83]	Malaysia	174	Adults (>18); with 55.2% aged between 45 and 64	Type 1 diabetes and type 2 diabetes	Quantitative (survey)
Soroya et al [116]	Pakistan	100	43% aged between 41 and 50	Not defined (41% were ≥10 years)	Mixed methods (survey and interview)
Sjöström et al [87]	Sweden	10 (interview); 58 (questionnaire)	Range 41-82	Type 2 diabetes (≤5 years)	Multimethod (survey and interview)
Zhao et al [88]	China	1563	Mean 65.9 (SD 9.7)	Not defined (57.7% were ≥5 years)	Quantitative (survey)

XSL•FO
RenderX

Study	Country	Sample size, n	Age (years)	Types of diabetes (duration of diabetes diagnosis)	Study design (data collection method)
Al Nozha and Elshatarat [89]	Saudi Arabia	211	Mean 42.2 (SD 17.6)	Type 1 diabetes and type 2 diabetes (mean 11.3 years; SD 7.9)	Quantitative (survey)
Broekhuis et al [90]	Scotland	12	Mean 54.8 (SD 8.58)	Type 2 diabetes	Qualitative (online diary)
Eke et al [23]	United States	2903	Adults (>18); with 50.4% aged ≥65	Not defined	Quantitative (survey)
Freeman et al [91]	Australia	37	Mean 54.1 (SD19.8)	Not defined	Mixed methods (survey and interview)
Langford et al [92]	United States	608	Not reported	Not defined	Quantitative (survey)
Rajanala et al [93]	United States	54	Mean 19.9 (SD 1.9)	Type 1 diabetes (68.5% were ≥5 years)	Quantitative (survey)
Subramaniam et al [94]	Singapore	436	Not reported	Type 1 diabetes and type 2 diabetes	Quantitative (survey)
Terkeş et al [95]	Turkey	250	Mean 58.5 (SD 1.03)	Type 2 diabetes (70.8% were ≥5 years)	Quantitative (survey)
Wang et al [96]	United States	7999	>25	Not defined	Quantitative (survey)
Yao et al [117]	China	22	Mean 57.3 (SD 10.8)	Type 2 diabetes (mean 7.3 years)	Qualitative (focus group)
Costa and Camargo-Plazas [97]	Canada	30	Adults (>18); with 60% aged ≥65	Type 1 diabetes and type 2 diabetes (60% were ≥5 years)	Qualitative (interview)
Andersen et al [98]	Denmark	12	Mean 57.1 (SD 7.2)	Type 2 diabetes (mean 6.6 years, SD 4.6)	Qualitative (interview)
Arueyingho et al [99]	Nigeria	110	Mean 54.5 (SD 6.8)	Type 2 diabetes (90% were ≤5 years)	Quantitative (survey)
Dsouza et al [100]	India	10	Mean 40	Type 2 diabetes (80% were ≤10 years)	Qualitative (interview)
Meer et al [101]	Kuwait	22	Adults (>18)	Type 2 diabetes	Qualitative (interview)
Naef et al [102]	Germany	20	Mean 16	Type 1 diabetes (mean 7 years)	Qualitative (interview)
Ouedraogo et al [103]	Guinea and Burkina Faso	92	48.9% aged between 20 and 50	Not defined	Quantitative (survey)
Peimani et al [104]	Iran	1143	Mean 58.8	Type 2 diabetes (mean 10.2 years)	Quantitative (survey)
Roesler et al [105]	Australia	815	Mean 36.5 (SD 4.7)	Gestational diabetes	Qualitative (survey)
Ji and Chi [106]	China	380	Mean 50.7 (SD 12.8)	Type 1 diabetes and type 2 diabetes (56.6% were ≥5 years)	Quantitative (survey)
Xu et al [107]	China	11 (interview); 235 (questionnaire)	Mean 31	Gestational diabetes	Mixed methods (survey and interview)
Birati et al [108]	Israel	24	58% aged between 30 and 39	Gestational diabetes	Qualitative (interview)
Jeon et al [109]	Korea	24	54.2% aged between 20 and 39	Type 1 diabetes and type 2 diabetes (62.5% were ≤5 years)	Qualitative (interview)
Kilinç İşleyen and Özdemir [110]	Turkey	241	Mean 59.8 (SD 7.7)	Type 2 diabetes (mean 11.6 years, SD 8.0)	Quantitative (survey)
Maxwell et al [111]	United States	26	Mean 22.6 (SD 2)	Type 1 diabetes (mean 12.6 years; SD 5.9)	Qualitative (interview)
Sadanandam et al [112]	India	150	44% aged between 44 and 54	Not defined (60.0% were ≥5 years)	Quantitative (survey)

XSL•FO
RenderX

Online Information Sources and Types of Information Sought

Across the 81 included studies [23,38-117], individuals with diabetes engaged with diverse digital platforms for health information seeking (Table 2). The most frequently reported category was general internet use (n=46, 57%), often described as "using the internet" or "looking up diabetes online" without specifying particular platforms. Social media platforms and online forums (eg, Facebook, Twitter, WeChat) were also commonly reported (n=38, 47%). Health-related websites, such as WebMD and Mayo Clinic, were identified in 16 (20%) studies, followed by search engines, primarily Google, in 15 (19%) studies. Less frequently reported sources were diabetes-specific websites (n=9, 11%), government or public health websites (n=7, 9%), and mobile health apps (n=6, 7%). Generative AI chatbots were reported in 1 (1%) recent study, suggesting the early but emerging integration of AI-based tools into OHIS practices.

Table 2. Temporal trends in online health information sources reported across 81 included studies[a].

Type of source	Time			Number of studies	Reported references
	2002-2010, n (%)	2011-2018, n (%)	2019-2025, n (%)		
General internet	8/12 (67)	13/29 (45)	25/40 (63)	46	[23,41,42,44-49,52,54,55,58,60,61,66,67,69-71,73-75,78,79,81,84,85,93-97,101-104,106,107,109,110,113-117]
Social media and online forums (eg, Facebook, WeChat)	2/12 (17)	9/29 (31)	27/40 (68)	38	[39,40,52,55,57,59,63-65,68,72,76,77,80,82,83,85,86,88-94,98,99,101,102,104,105,107,108,110-112,116,117]
Health-related websites (eg, WebMD, Mayo Clinic)	5/12 (42)	7/29 (24)	4/40 (10)	16	[38-40,44,48,50,51,53,56,58,62,65,76,86,90,91]
Search engine (eg, Google, Bing)	1/12 (8)	4/29 (14)	10/40 (25)	15	[43,59,62,64,72,77,80,86-91,97,100]
Diabetes-specific website (eg, American Diabetes Association, Children with Diabetes)	2/12 (17)	4/29 (14)	3/40 (8)	9	[40,48,50,51,57,59,80,90,112]
Government and public health websites (eg, NICE[b] guideline, Public Health Agency of Sweden)	0/0 (0)	3/29 (10)	4/40 (10)	7	[50,64,68,76,87,88,90]
Mobile health apps	0/0 (0)	0/0 (0)	6/40 (15)	6	[88,90,99,107-109]
Generative artificial intelligence chatbots	0/0 (0)	0/0 (0)	1/40 (3)	1	[109]
Other sources (news websites, email, and web-based courses)	2/12 (17)	2/29 (7)	4/40 (10)	8	[40,41,53,65,87,90,99,104]

[a]Values represent the percentage of studies within each publication period that reported each source category (2002-2010, n=12; 2011-2018, n=29; and 2019-2025, n=40). Categories are not mutually exclusive, and individual studies could report multiple sources.

[b]NICE: National Institute for Health and Care Excellence.

Temporal stratification revealed shifts in digital engagement patterns. Earlier studies (2002-2010) predominantly described the use of search engines and static health websites. From 2011 onward, social media platforms and online communities became increasingly prominent. In the most recent period (2019-2025), interactive and mobile platforms were more frequently reported, and AI-based tools appeared in the literature, albeit to a limited extent. The proportion of studies reporting social media and online community use increased from 17% (2/12) during the early web era (2002-2010) to 68% (27/40) during the integrated digital and AI era (2019-2025), whereas reports of health-related website use declined from 42% (5/12) to 10% (4/40) over the same period (Table 2). Mobile health apps and generative AI tools were reported exclusively in the most recent era. These findings suggest a transition from static information retrieval to more participatory, interactive, and technology-enabled digital environments.

Of the included studies, 49 (60%) [42,43,47-53,56-60,62-68,70-73,75,76,78,80-83,87,89,90,92,93,95,97-101,104,105,107,108,111,112] reported on the types of online health information sought (Table 3). Information about self-management and lifestyle was the most frequently reported category (n=28, 57%), including diet, physical activity, glucose monitoring, and stress management. General diabetes knowledge was reported in 21 (43%) studies, and treatment-related information (eg, treatment options and medication side effects) appeared in 18 (37%) studies. Other commonly reported categories included symptoms or complications (n=13, 27%), general health topics (n=12, 24%), and lived experiences and peer support (n=11, 22%). Technology-related information (eg, glucose meters, insulin pumps) was reported in 3 (6%) studies. Less frequently discussed topics included psychosocial concerns (eg, depression), family impact, health insurance, and COVID-19–related information.

XSL•FO
RenderX

Table 3. Categories of online health information types sought across the included studies[a].

Category of information type	Description	Number of studies	Reported references
Self-management and lifestyle	Diet, nutrition, exercise, and stress management	28	[43,48-52,57,59,64,66,71,72,76,78,80,82,83,89,90,99-101,104,105,107,108,111,112]
General diabetes	Information about disease knowledge, causes, and diagnosis	21	[43,47,52,58-60,66,67,70,71,73,76,81,83,90,93,95,97,98,104,112]
Treatment-related information	Medications, side effects, and alternative therapies	18	[43,51-53,59,62,64,66,67,71,76,78,82,83,90,95,97,101]
Symptoms and complications	Symptoms of diabetes and consequences of poor metabolic control	13	[43,59,66,71,76,83,89,90,95,97,101,108,112]
General health topics	Information about general health and wellness	12	[42,43,50,56,59,67,75,81,83,89,90,92]
Lived experiences and peer support	Lived experiences, second opinion, and personal advice from online peers	11	[52,63,65,66,68,76,78,80,83,107,111]
Diabetes device/technology	Blood glucose meter, insulin pump, and medical equipment	3	[43,78,82]
Other specific topics	Impact of diabetes on family life, sensitive health issues, COVID-19 information, and health care resources	4	[50,70,76,87]

[a]Of the 81 included studies, 49 reported extractable data for this outcome. Categories are not mutually exclusive, and individual studies could report multiple information types.

Across diabetes populations, self-management and lifestyle information, general diabetes knowledge, and treatment-related information were most commonly reported in studies involving type 2 diabetes or mixed diabetes populations. Technology-related information was identified exclusively in studies that included individuals with type 1 diabetes, whereas studies of gestational diabetes primarily reported information seeking related to self-management and lifestyle, as well as lived experiences and peer support (Figure 4). Overall, the content of information seeking remained relatively stable across technological eras, with self-management information consistently representing the most frequently reported category. However, more recent studies increasingly described experiential knowledge exchange, emotional support seeking, and peer-to-peer information sharing within social media environments. This pattern suggests that contemporary OHIS extends beyond information retrieval and increasingly incorporates social interaction, collective knowledge construction, and emotional support.

Figure 4. Types of online health information sought across diabetes populations. GD: gestational diabetes; T1D: type 1 diabetes; T2D: type 2 diabetes.

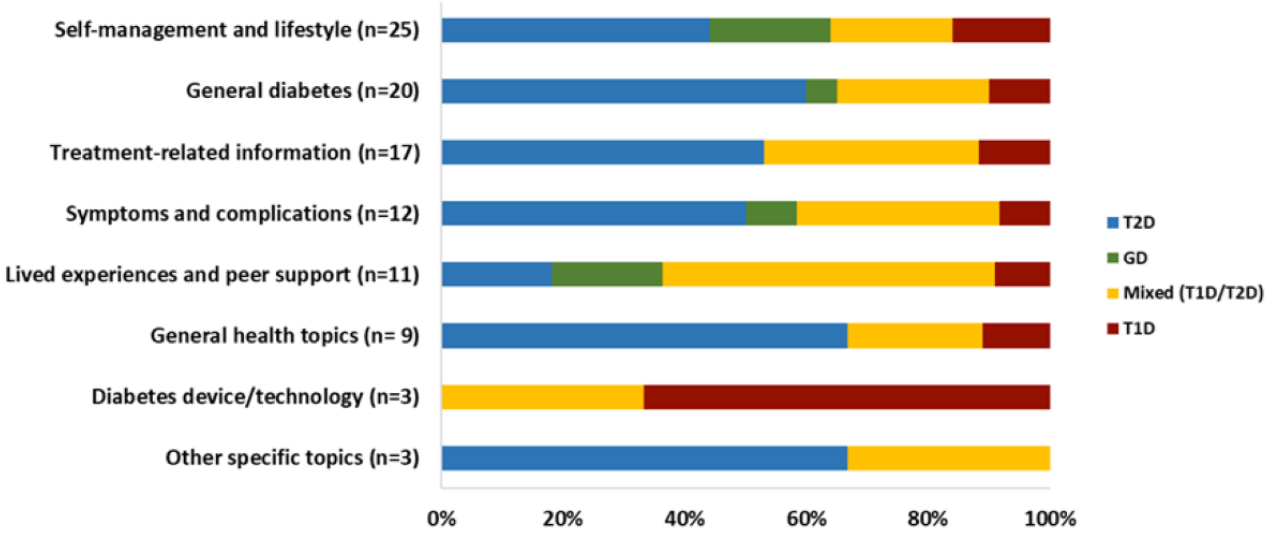

Multilevel Factors Influencing OHIS

Determinants of OHIS

Determinants of OHIS were reported in 54 of the 81 (67%) studies [23,38,40,42-44,46,48,50,51,53-57,59-67,69,71-76,78-80,82,84,87,89-92,96-98,100,102,104,107,108,111,112,114,116,117] and were organized according to the Social Ecological Model levels of individual, interpersonal, organizational, and environmental factors (Table 4).

XSL•FO
RenderX

Table 4. Factors influencing online health information seeking among individuals with diabetes, organized by the Social Ecological Model[a].

Influencing factors	Evidence basis	Direction	Reported references
Individual factors			
Sociodemographic			
Increasing age	QUANT[b] and QUAL[c]	–[d]	[23,40,42,44,59-61,71,75,92,96]
Greater education level	QUANT and QUAL	+[e]	[23,38,42,44,54,59-61,63,71,75,76,96,97,104,112]
Higher household income	QUANT and QUAL	+	[23,60,61,71,96,100,112]
Female sex	QUANT and QUAL	+	[40,96,104,112]
Racial minorities	QUANT	–	[96]
Employment	QUANT	+	[71,112]
Health status			
Shorter duration of diabetes	QUANT	+	[43,60,61]
Poor glycemic control	QUANT	+/–[f]	[60,61]
Type 1 diabetes (vs type 2 diabetes)	QUANT	+	[60,89]
Obesity	QUANT	+	[96]
Reporting neuropathy	QUANT	+	[61]
Reporting severe hypoglycemia	QUANT	+	[38]
Better self-reported health	QUANT	–	[76]
Receiving lifestyle modification only	QUANT	+	[71]
Having a family history of diabetes	QUANT	+	[71]
Newly diagnosed diabetes	QUAL	+	[53]
Psychological			
Perceived insufficient knowledge	QUAL	+	[73,90,91,97,107]
Negative emotions (eg, anxiety, distress, fear)	QUANT and QUAL	+/–	[72,76,80,87,90,98,108]
Desire for greater control	QUAL	+	[53,62,91]
Preference for physical interaction	QUAL	–	[66,91]
Lack of confidence in using digital devices	QUAL	–	[91]
Highly activated patients	QUANT	+	[61]
Higher perceived threat of complications	QUANT	–	[76]
Higher efficacy beliefs in diabetes management	QUANT	–	[76]
Knowledge and skills			
Lower health literacy	QUANT and QUAL	–	[48,50,51,73,98]
Lower eHealth literacy	QUANT and QUAL	–	[56,63,72,78,87,91,104]
Inexperience in internet use	QUAL	–	[53]
Limited digital literacy	QUANT	–	[84,100,116]
Limited exposure to diabetes education	QUANT	–	[59]
Interpersonal factors			
Social support			
Online peer support	QUAL	+	[57,63-65,78,80,82,108]
Family encouragement and assistance	QUAL	+	[51,56,67,80]

XSL•FO
RenderX

Influencing factors	Evidence basis	Direction	Reported references
Sense of community	QUAL	+	[80,82,111]
Family as an information substitute	QUAL	–	[53,62]
Relationships and interactions with health care professionals			
Communication barriers	QUAL	+	[57,66,91,102,107]
Inadequate information provision	QUAL	+	[46,65,66,80,91,97]
Higher trust	QUAL	–	[53,62,65,78,107]
Recommendation and guidance to online sources	QUAL	+	[46,73,91]
Organizational factors			
Health care delivery and organizational structure			
Disorganized care	QUANT	+	[61]
Information overload	QUAL	+	[102]
Health care access and use			
More frequent clinical visits with health care professionals	QUANT	+	[23,43]
Limited health care access	QUAL	+	[53,114]
Environmental factors			
Accessibility			
Limited access to the internet and technology	QUANT and QUAL	–	[38,56,66,90,91,114,116]
Convenience and ease of access	QUAL	+	[80,87,91]
Source characteristics			
Information overload	QUANT and QUAL	–	[48,53,66,67,78,79,84,87]
Conflicting information	QUAL	–	[98,107]
Misinformation	QUAL	–	[63,64,79,82]
Low trust in online sources	QUAL	–	[53,55,57,62,66,91,107,117]
High trust in peer sources	QUAL	+	[63,80]
Perceived usefulness of internet	QUANT and QUAL	+	[46,50,67]
High complexity of information (eg, medical jargon and technical language)	QUAL	–	[48,66,67,74]
Excessive advertising	QUAL	–	[107]
Inappropriate content depth	QUAL	–	[107]
Infrequent updates	QUAL	–	[107]
Language availability of source	QUAL	+	[72,73]
Accessible content presentation and layout	QUAL	+	[55,64,78,111]

[a]Of the 81 included studies, 54 reported extractable data relevant to this outcome.

[b]QUANT: quantitative findings.

[c]QUAL: qualitative findings.

[d]The "–" sign indicates a negative effect (indicating a negative association or hindrance of online health information seeking).

[e]The "+" sign indicates a positive effect (indicating a positive association or facilitation of online health information seeking).

[f]The "+/–" indicates mixed findings (indicating both positive and negative associations, or factors described as both facilitating and hindering online health information seeking across studies).

XSL•FO
RenderX

Individual-Level Factors

Individual-level influences were reported in 41 of the 54 (76%) studies [23,38,40,42-44,48,50,51,53,54,56,59-63,66,69,71-73,75,76,78,80,84,87,89-92,96-98,100,104,107,108,112,116]. Sociodemographic characteristics were consistently associated with OHIS patterns. Higher educational attainment and income were linked to greater OHIS engagement, whereas increasing age was generally associated with lower levels of online searching. Female sex was associated with higher OHIS frequency in several studies [40,96,104], whereas 1 study reported the opposite trend [112]. One study reported lower engagement among racial minority groups [96]. Health status variables showed more heterogeneous associations. Individuals with type 1 diabetes and those with shorter disease duration were more likely to seek information online [43,60,61,89]. Glycemic control demonstrated mixed findings across studies [60,61]. Being newly diagnosed was qualitatively described as a strong motivator for initiating OHIS [53].

Psychological factors played a prominent role. Patients frequently reported initiating OHIS in response to perceived knowledge gaps, abnormal test results, or new symptoms [73,90,91,97,107]. Seeking online information was often framed as a strategy to enhance autonomy and reduce reliance on health care professionals (HCPs) for minor concerns [53,62,91]. Emotional responses operated bidirectionally. Anxiety and perceived threat could prompt searching [76,80,90], whereas fear, confusion, or distress triggered by online content could discourage continued engagement [72,87,98,108]. A preference for in-person consultation reduced reliance on digital sources [66,91].

Knowledge and skills were consistently implicated. Lower eHealth literacy and limited digital competence were associated with reduced OHIS engagement and difficulty evaluating information quality [56,72,78,91]. Conversely, individuals with higher eHealth literacy were more likely to cross-verify information across sources [87]. Limited prior exposure to diabetes education and inexperience with internet use were also reported as barriers [53,59].

Interpersonal-Level Factors

Interpersonal influences were described in 22 of the 54 (41%) studies [46,51,53,56,57,62-67,69,73,78,80,82,91,97,102,107,108,111]. Online peer support communities were frequently characterized as facilitators of OHIS by providing experiential knowledge, emotional reassurance, and a sense of shared identity [57,63-65,78,80,82,108]. These communities were often perceived as safe spaces for discussing concerns not easily raised in clinical settings [80,82,111]. In addition, family involvement showed mixed effects. Encouragement and assistance from family members promoted OHIS [51,56,67,80], whereas reliance on family members perceived as knowledgeable sometimes reduced independent information seeking [53,62].

Interactions with HCPs also affected OHIS. Communication barriers during consultations [57,66,91,102,107] or insufficient explanations [46,65,66,80,91,97] frequently prompted patients to seek supplementary online information. Trust in HCPs was inversely associated with independent OHIS in some studies [53,62,65,69,78,107]. Conversely, explicit recommendations from clinicians to consult reliable online resources encouraged structured engagement [46,73,91].

Organizational-Level Factors

Organizational influences were reported in 6 of the 54 (11%) studies [23,43,53,61,102,114]. Perceived inadequacies in care coordination, such as fragmented services and overwhelming or poorly structured information from health institutions, were associated with increased OHIS [61,102]. Health care utilization showed mixed associations with OHIS. Frequent use of health care services was positively associated with OHIS in 2 studies [23,43], whereas limited access to care led to compensatory reliance on digital information in another [53,114].

Environmental-Level Factors

Environmental-level influences were identified in 31 of the 54 (57%) studies [38,46,48,50,53,55-57,62-64,66,67,69,72-74,78-80,82,84,87,90,91,98,107,111,114,116,117]. Structural access to information technology was a foundational determinant. Limited internet connectivity or lack of appropriate devices constrained OHIS [38,56,66,90,91,114,116], whereas smartphone availability and ease of access facilitated engagement [80,87,91]. In addition, source-related characteristics were prominent determinants. Concerns regarding credibility [53,55,57,62,66,91,107,117], misinformation [63,64,79,82], conflicting content [69,98,107], and information overload [48,53,66,67,78,79,84,87] were widely reported. Content complexity and medical jargon were described as barriers [66,67,74], particularly for individuals with lower health literacy. Facilitators included culturally appropriate content [64], user-friendly layouts and visual aids [55,78,111], and information available in preferred languages [72,73]. Of note, some studies reported greater trust in peer-generated content than in clinician-provided information [63,80], reflecting shifting epistemic dynamics within digital health ecosystems.

Evidence Gap Map

To enhance data visualization and summarize the distribution of evidence, an evidence gap map was developed based on diabetes type and levels of influencing factors (Figure 5). Overall, the evidence was concentrated at the individual and environmental levels, with the highest number of studies focusing on type 2 diabetes. By contrast, organizational-level factors were sparsely represented across all diabetes groups.

XSL•FO
RenderX

Figure 5. Evidence gap map of factors influencing online health information seeking across diabetes types.

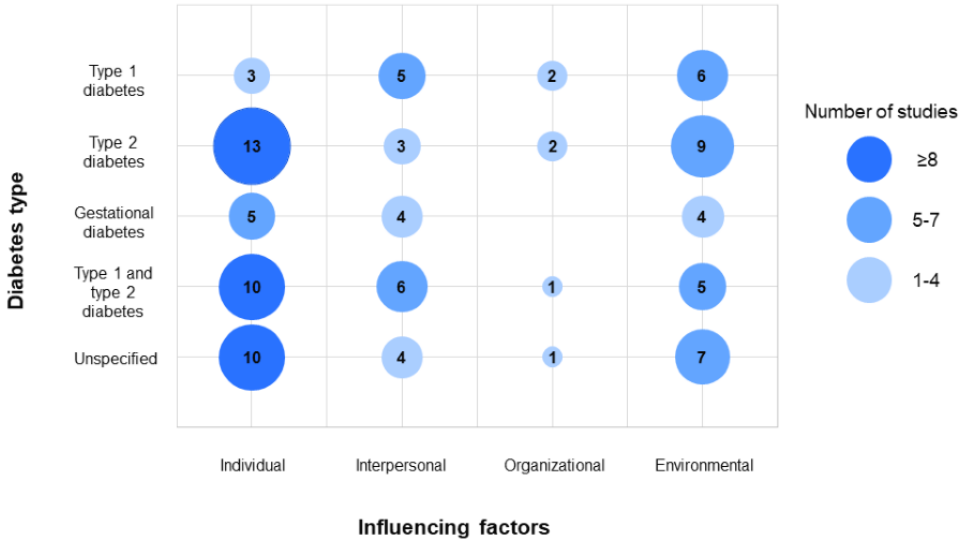

Discussion

Principal Findings

This scoping review provides a comprehensive synthesis of OHIS behaviors among individuals with diabetes within a rapidly evolving digital ecosystem. Among 81 studies, we identified substantial diversification of digital information sources, a temporal shift toward interactive and socially mediated platforms, and a complex constellation of multilevel determinants shaping OHIS engagement. These findings suggest that OHIS has evolved beyond a supplementary activity and now represents an integral component of contemporary diabetes self-management. Compared with earlier reviews [28,29], this synthesis extends prior work by integrating technological evolution, multilevel determinants, and equity considerations within a unified framework. By positioning OHIS within the broader digital health environment, including search engines, social media platforms, and emerging generative AI systems, this review advances a more contextualized understanding of how individuals with diabetes navigate increasingly algorithm-mediated information environments.

Transformation of the Digital Information Ecosystem

Our findings demonstrate a clear temporal shift in information sources. Earlier studies primarily reported reliance on search engines and institutional websites, whereas more recent publications emphasize social media platforms and online communities. These interactive environments offer experiential knowledge, emotional validation, and peer-generated advice that may complement formal medical guidance [118-120]. Engagement in peer support spaces has been associated with improved glycemic outcomes and greater self-management competence, highlighting their potential clinical relevance [121,122].

At the same time, these platforms introduce epistemic vulnerabilities. Information quality is uneven, commercial influences are pervasive, and algorithmic amplification may privilege engaging content over evidence-based information [123]. Rapidly evolving platforms, such as TikTok, exemplify environments in which users may find it difficult to evaluate credibility [124]. The dual role of social media as both a source of support and a channel for misinformation highlights the need for digital stewardship within clinical practice. HCPs may extend their role beyond traditional consultations by curating credible digital resources and supporting digitally informed peer leaders within patient communities [125].

Emergence of AI-Driven Conversational Tools

The integration of large language model–based chatbots represents a further transformation in OHIS [10,126]. Although only 1 included study directly reported the use of generative AI among individuals with diabetes [109], emerging literature suggests increasing experimentation with conversational AI tools for health queries [127]. Unlike traditional search engines, AI chatbots generate synthesized responses in natural language and may reduce cognitive burden while tailoring information to varying literacy levels [128,129]. These tools may also provide affective support through conversational interaction [130].

However, concerns remain regarding accuracy, transparency, and contextual appropriateness [131,132]. The absence of patient-specific data limits clinical reliability, and algorithmic opacity complicates trust calibration [133]. Moreover, individuals with chronic conditions have expressed skepticism toward AI-mediated health advice [134]. Future research should therefore explore how patients integrate AI-generated responses into decision-making processes, how trust is established or eroded, and how these tools interact with clinician guidance.

XSL•FO
RenderX

As conversational AI becomes embedded in routine information practices, its influence on diabetes self-management requires systematic investigation.

Persistent Demand for Practical and Actionable Content

Patients' health information needs appeared remarkably consistent despite rapid changes in digital technologies. Across technological eras, self-management, general diabetes knowledge, and treatment-related information remained the most frequently sought topics [20]. This pattern suggests that OHIS serves as an ongoing complement to diabetes care and self-management beyond clinical encounters. Stratified analyses further indicated that information needs varied across diabetes populations. Technology-related content was more prominent among individuals with type 1 diabetes, whereas individuals with gestational diabetes more frequently sought information related to self-management and peer support (Table 4). These findings suggest that a one-size-fits-all approach to digital health information may be insufficient and highlight the need for resources tailored to the circumstances and priorities of different diabetes populations.

Despite the expanding availability of online health information, barriers related to readability, usability, and information overload remain common [135]. Individuals with limited health literacy may face additional challenges when navigating complex medical terminology and poorly designed websites [136]. Improving digital health information, therefore, requires not only increasing access but also enhancing its relevance and usability. Co-development of online resources involving patients and HCPs, together with the use of structured evaluation tools such as DISCERN (a validated instrument for evaluating the quality of consumer health information on treatment choices) and PEMAT (Patient Education Materials Assessment Tool), may help improve the quality, clarity, and relevance of online content [137]. Regular assessment of the readability and usability of web-based materials by health care organizations may further support equitable access to digital health information and help reduce disparities in engagement and benefit.

Digital Divide and Sociodemographic Stratification

Consistent with broader digital health literature, engagement in OHIS followed clear sociodemographic gradients [35,138]. Older age, lower educational attainment, limited income, and a diagnosis of type 2 diabetes were associated with reduced online engagement. These patterns likely reflect structural digital divides encompassing technological access, digital skills, and confidence [25]. As type 2 diabetes disproportionately affects older adults and socially disadvantaged populations [139], sociodigital inequities may amplify existing health and metabolic vulnerabilities [13].

Although many sociodemographic factors are not modifiable, recognition of their influence enables the implementation of targeted and proportionate support strategies [62,140]. Individuals who are younger or recently diagnosed may benefit from curated digital resource pathways and structured orientation to credible online platforms. For patients who experience barriers to digital engagement, alternative communication modalities, including printed educational materials and structured in-person education, remain essential complements to digital approaches [141]. Addressing digital inequities should therefore be viewed not only as a technological issue but also as a core component of equitable and patient-centered diabetes care.

Psychological Processes and Distributed Health Literacy

Beyond demographic predictors, psychological dynamics also strongly shape OHIS. Perceived knowledge gaps, abnormal clinical results, and desires for greater autonomy were commonly associated with increased online information seeking [142,143], whereas confusion, fear, and exposure to conflicting information sometimes contributed to information avoidance [144]. These findings are consistent with theories of uncertainty management and information avoidance, suggesting that engagement with online health information involves marked emotional and cognitive processes [144]. Online health information seeking may therefore serve not only as information acquisition but also as a strategy for emotional regulation and sense-making. Patients frequently used online communities to seek reassurance, validate personal experiences, and obtain experiential knowledge, particularly when experiencing uncertainty or diabetes-related distress.

Social networks also appeared to play a prominent mediating role in OHIS. Family members and peers often assist with interpreting and contextualizing online health information, which reflects the concept of distributed health literacy within social systems [145]. While such support may enhance understanding and facilitate diabetes self-management, it may also introduce potential pathways for misinformation if mediators lack adequate literacy [146]. As diabetes management commonly occurs within family and community contexts, the interpretation and application of online information may depend not only on individual skills but also on the knowledge and beliefs of broader social networks. These findings suggest that strategies to improve digital engagement and eHealth literacy may benefit from extending beyond individual patients to include family members, caregivers, and peer communities [70].

Clinical Implications and System-Level Considerations

Patient-provider communication emerged as an influential determinant of OHIS behaviors. Patients frequently turned to online sources when consultations left informational or emotional needs unmet [2]. Greater trust in HCPs was associated with reduced independent searching, yet explicit clinician endorsement of credible digital resources seemed to encourage more structured and informed engagement [147]. These findings suggest that OHIS should be incorporated into clinical communication rather than treated as a competing information source [148]. Encouraging patients to discuss online findings during consultations may help facilitate collaborative decision-making and support chronic disease management [149]. Such integration also depends on HCPs possessing adequate eHealth literacy and familiarity with digital resources. Variability in clinicians' digital competencies further highlights

XSL•FO
RenderX

the need for workforce development and additional research in this area [150].

These findings also carry broader implications for health care systems and digital health policy. Health care organizations may need to expand access to evidence-based and user-friendly digital resources while supporting HCPs in digital communication and evaluation of online information quality. As AI-generated and algorithmically curated content becomes increasingly embedded within digital health environments, concerns related to transparency, governance, and digital health equity may become more pronounced. Integrating digital health guidance into routine care may therefore help promote safer, more informed, and more equitable patient engagement within evolving digital ecosystems.

Limitations

Several limitations should be acknowledged. The review was restricted to studies published in English and did not include gray literature. This may have resulted in the omission of relevant evidence from non-English–speaking contexts and introduced potential publication bias. Considerable heterogeneity in study design, populations, and reporting of socioeconomic variables limited detailed subgroup analysis. Although disparities in age, income, and education were noted, inconsistent reporting prevented a standardized synthesis of barriers across low-income or older groups. Most included studies were conducted in Western and high-income settings, which may limit the transferability of findings to low- and middle-income regions where digital infrastructure and health systems differ. In addition, the rapid evolution of digital technologies, including AI-mediated platforms, means that emerging developments may not yet be fully captured. Although this review identified a large and growing body of literature, the rapid expansion of the field may challenge the timeliness of synthesis, given the evolving nature of digital platforms. Future research should include broader geographic representation and more consistent measurement of social determinants to strengthen contextual understanding of OHIS.

Conclusions

This scoping review provides the most comprehensive synthesis to date of OHIS among individuals with diabetes in a rapidly evolving digital health ecosystem. Unlike previous reviews that focused primarily on specific platforms, online communities, or general information-seeking patterns, this review integrates technological evolution, multilevel determinants, and digital health equity within a single conceptual framework. The findings demonstrate that OHIS has become a dynamic, socially embedded, and increasingly algorithm-mediated component of diabetes self-management.

By identifying temporal shifts from static web resources to interactive social media platforms and emerging AI-enabled tools, this review contributes to a broader understanding of how digital ecosystems shape patient engagement, decision-making, and self-management. The evidence also highlights persistent disparities across age, education, income, and digital competencies, alongside notable gaps in organizational-level research. These findings provide actionable insights for HCPs, health care systems, and policy makers seeking to support safe, equitable, and evidence-informed digital engagement. Integrating digital health guidance into routine clinical care, strengthening eHealth literacy, and implementing transparent digital health governance strategies may help ensure that future technological advances improve diabetes outcomes without exacerbating existing health inequities.

Acknowledgments

The authors used ChatGPT-5.5 (OpenAI) for proofreading and language editing during manuscript preparation. All generated content was reviewed, verified, and edited by the authors, who assume full responsibility for the final manuscript. No generative artificial intelligence tool was involved in study design, data collection, data analysis, interpretation of findings, or authorship. Preliminary findings from this study were presented at the 47th Annual Meeting and Scientific Sessions of the Society of Behavioral Medicine (held in Chicago, Illinois, on April 22, 2026).

Data Availability

All data generated or analyzed during this study are included in this published article and its multimedia appendices.

Funding

This research was supported by the Academic Research-Career Development Project (Laurel Research Project NTU-115L7736) of National Taiwan University, funded by the Ministry of Education, Taiwan. The funding body had no role in the study design, data collection, statistical analysis, interpretation of findings, or manuscript preparation.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Use of the SPIDER Framework (Sample, Phenomenon of Interest, Design, Evaluation, Research Type) to guide the research question.
[DOCX File , 15 KB - jmir_v28i1e82081_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist.
[DOCX File , 27 KB - jmir_v28i1e82081_app2.docx]

Multimedia Appendix 3
PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) checklist.
[DOCX File , 20 KB - jmir_v28i1e82081_app3.docx]

Multimedia Appendix 4
Search strategy and procedures used to identify studies in PubMed, Scopus, Web of Science, and CINAHL.
[DOCX File , 18 KB - jmir_v28i1e82081_app4.docx]

References

1. Bundorf MK, Wagner TH, Singer SJ, Baker LC. Who searches the internet for health information? Health Serv Res 2006 Jun;41(3 Pt 1):819-836 [FREE Full text] [doi: 10.1111/j.1475-6773.2006.00510.x] [Medline: 16704514]
2. Liu D, Yang S, Cheng CY, Cai L, Su J. Online health information seeking, eHealth literacy, and health behaviors among Chinese internet users: cross-sectional survey study. J Med Internet Res 2024 Oct 18;26:e54135 [FREE Full text] [doi: 10.2196/54135] [Medline: 39423374]
3. Individuals using the internet for seeking health-related information. Eurostat. 2026 Feb 25. URL: https://ec.europa.eu/eurostat/databrowser/view/tin00101/default/table?lang=en [accessed 2026-04-30]
4. Jia X, Pang Y, Liu LS. Online health information seeking behavior: a systematic review. Healthcare (Basel) 2021 Dec 16;9(12):1740 [FREE Full text] [doi: 10.3390/healthcare9121740] [Medline: 34946466]
5. Di Novi C, Kovacic M, Orso CE. Online health information seeking behavior, healthcare access, and health status during exceptional times. J Econ Behav Organ 2024 Apr;220:675-690. [doi: 10.1016/j.jebo.2024.02.032] [Medline: 38628501]
6. Neely S, Eldredge C, Sanders R. Health information seeking behaviors on social media during the COVID-19 pandemic among American social networking site users: survey study. J Med Internet Res 2021 Jun 11;23(6):e29802 [FREE Full text] [doi: 10.2196/29802] [Medline: 34043526]
7. Lee K, Hoti K, Hughes JD, Emmerton LM. Consumer use of "Dr Google": a survey on health information-seeking behaviors and navigational needs. J Med Internet Res 2015 Dec 29;17(12):e288 [FREE Full text] [doi: 10.2196/jmir.4345] [Medline: 26715363]
8. Elnaggar A, Ta Park V, Lee SJ, Bender M, Siegmund LA, Park LG. Patients' use of social media for diabetes self-care: systematic review. J Med Internet Res 2020 Apr 24;22(4):e14209. [doi: 10.2196/14209] [Medline: 32329745]
9. Tesch Z, Prónay S, Buzás N. Can the group effect dominate the influence of the child on the parent's decision to care for type 1 diabetes? J Pediatr Nurs 2024 May;76:e19-e26. [doi: 10.1016/j.pedn.2024.01.014] [Medline: 38281892]
10. Yun HS, Bickmore T. Online health information-seeking in the era of large language models: cross-sectional web-based survey study. J Med Internet Res 2025 Mar 31;27:e68560 [FREE Full text] [doi: 10.2196/68560] [Medline: 40163112]
11. Ayo-Ajibola O, Davis RJ, Lin ME, Riddell J, Kravitz RL. Characterizing the adoption and experiences of users of artificial intelligence-generated health information in the United States: cross-sectional questionnaire study. J Med Internet Res 2024 Aug 14;26:e55138 [FREE Full text] [doi: 10.2196/55138] [Medline: 39141910]
12. Mendel T, Singh N, Mann DM, Wiesenfeld B, Nov O. Laypeople's use of and attitudes toward large language models and search engines for health queries: survey study. J Med Internet Res 2025 Feb 13;27:e64290 [FREE Full text] [doi: 10.2196/64290] [Medline: 39946180]
13. Zhao YC, Zhao M, Song S. Online health information seeking behaviors among older adults: systematic scoping review. J Med Internet Res 2022 Feb 16;24(2):e34790 [FREE Full text] [doi: 10.2196/34790] [Medline: 35171099]
14. Rodrigues F, Newell R, Rathnaiah Babu G, Chatterjee T, Sandhu NK, Gupta L. The social media iInfodemic of health-related misinformation and technical solutions. Health Policy Technol 2024 Jan 30;13(2):100846. [doi: 10.1016/j.hlpt.2024.100846]
15. Purnat TD, Wilhelm E, White BK, Okan O, Rosario R, Scales D. Health promotion in the algorithmic age: recognizing the information environment as a determinant of health. Health Promot Int 2025 Sep 3;40(5):daaf166. [doi: 10.1093/heapro/daaf166] [Medline: 40990145]
16. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
17. Rosenlund M, Turja T, Jylhä V, Saranto K, Kuusisto H. Experiences of decision-making in healthcare and online health information-seeking among older adults and people with long-term disease: online survey study. J Patient Exp 2026 Jan 20;13:23743735251415086. [doi: 10.1177/23743735251415086] [Medline: 41573246]
18. Duncan BB, Magliano DJ, Boyko EJ. IDF Diabetes Atlas 11th edition 2025: global prevalence and projections for 2050. Nephrol Dial Transplant 2025 Dec 23;41(1):7-9. [doi: 10.1093/ndt/gfaf177] [Medline: 40874767]
19. Davis J, Fischl AH, Beck J, Browning L, Carter A, Condon JE, et al. 2022 national standards for diabetes self-management education and support. Diabetes Spectr 2022;35(2):137-149 [FREE Full text] [doi: 10.2337/ds22-ps02] [Medline: 35668879]

XSL•FO
RenderX

20. Biernatzki L, Kuske S, Genz J, Ritschel M, Stephan A, Bächle C, et al. Information needs in people with diabetes mellitus: a systematic review. Syst Rev 2018 Feb 14;7:27 [FREE Full text] [doi: 10.1186/s13643-018-0690-0] [Medline: 29444711]
21. Hilliard ME, Sparling KM, Hitchcock J, Oser TK, Hood KK. The emerging diabetes online community. Curr Diabetes Rev 2015;11(4):261-272 [FREE Full text] [doi: 10.2174/1573399811666150421123448] [Medline: 25901500]
22. Ahmad F, Joshi SH. Self-care practices and their role in the control of diabetes: a narrative review. Cureus 2023 Jul 5;15(7):e41409 [FREE Full text] [doi: 10.7759/cureus.41409] [Medline: 37546053]
23. Eke R, Jannat KT, Yang XT, Parton JM. Disparity and factors associated with internet health information seeking among US adults living with diabetes mellitus: cross-sectional study. J Med Internet Res 2022 May 30;24(5):e32723 [FREE Full text] [doi: 10.2196/32723] [Medline: 35635741]
24. Abid MH, Kumah A, Newera A, Hafez P. Patient-centered health care: from patient experience to human experience. Glob J Qual Saf Healthc 2024 Jul 10;7(4):144-148. [doi: 10.36401/jqsh-24-x2] [Medline: 39534234]
25. Lythreatis S, Singh SK, El-Kassar AN. The digital divide: a review and future research agenda. Technol Forecast Soc Change 2022 Feb;175:121359. [doi: 10.1016/j.techfore.2021.121359]
26. Badr J, Motulsky A, Denis JL. Digital health technologies and inequalities: a scoping review of potential impacts and policy recommendations. Health Policy 2024 Aug;146:105122 [FREE Full text] [doi: 10.1016/j.healthpol.2024.105122] [Medline: 38986333]
27. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119. [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
28. Kuske S, Schiereck T, Grobosch S, Paduch A, Droste S, Halbach S, et al. Diabetes-related information-seeking behaviour: a systematic review. Syst Rev 2017 Oct 24;6(1):212. [doi: 10.1186/s13643-017-0602-8] [Medline: 29065919]
29. Tesfa GA, Demeke AD, Zewold M, Ngusie HS. Health information-seeking behavior among people living with the two common chronic diseases in low and middle-income countries (LMICs). A systematic review and meta-analysis. Digit Health 2024 Nov 26;10:20552076241302241. [doi: 10.1177/20552076241302241] [Medline: 39600387]
30. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
31. Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020 Oct;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
32. Cooke A, Smith D, Booth A. Beyond PICO: the SPIDER tool for qualitative evidence synthesis. Qual Health Res 2012 Jul 24;22(10):1435-1443. [doi: 10.1177/1049732312452938] [Medline: 22829486]
33. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Sep 4;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
34. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
35. Zimmerman MS, Shaw G. Health information seeking behaviour: a concept analysis. Health Info Libr J 2020 Feb 12;37(3):173-191. [doi: 10.1111/hir.12287] [Medline: 32052549]
36. Pollock D, Peters MDJ, Khalil H, McInerney P, Alexander L, Tricco AC, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. JBI Evid Synth 2023 Mar 1;21(3):520-532. [doi: 10.11124/JBIES-22-00123] [Medline: 36081365]
37. Kilanowski JF. Breadth of the socio-ecological model. J Agromedicine 2017 Jul 25;22(4):295-297. [doi: 10.1080/1059924x.2017.1358971] [Medline: 28742433]
38. Giménez-Pérez G, Gallach M, Acera E, Prieto A, Carro O, Ortega E, et al. Evaluation of accessibility and use of new communication technologies in patients with type 1 diabetes mellitus. J Med Internet Res 2002 Dec;4(3):E16 [FREE Full text] [doi: 10.2196/jmir.4.3.e16] [Medline: 12554547]
39. Millard RW, Fintak PA. Use of the internet by patients with chronic illness. Disease Management and Health Outcomes 2002 Mar;10(3):187-194. [doi: 10.2165/00115677-200210030-00006]
40. Eriksson-Backa K. Who uses the web as a health information source? Health Informatics J 2003 Jun;9(2):93-101. [doi: 10.1177/1460458203009002004]
41. Wagner TH, Baker LC, Bundorf MK, Singer S. Use of the internet for health information by the chronically ill. Prev Chronic Dis 2004 Oct;1(4):A13 [FREE Full text] [Medline: 15670445]
42. Jackson CL, Batts-Turner ML, Falb MD, Yeh HC, Brancati FL, Gary TL. Computer and internet use among urban African Americans with type 2 diabetes. J Urban Health 2005 Dec;82(4):575-583 [FREE Full text] [doi: 10.1093/jurban/jti126] [Medline: 16221917]
43. Nordfeldt S, Johansson C, Carlsson E, Hammersjö JA. Use of the internet to search for information in type 1 diabetes children and adolescents: a cross-sectional study. Technol Health Care 2005 Jan 1;13(1):67-74. [doi: 10.3233/thc-2005-13104]
44. Robertson JL, Akhtar S, Petrie J, Brown F, Jones G, Perry C, et al. How do people with diabetes access information? Pract Diab Int 2005 Jul 19;22(6):207-210 [FREE Full text] [doi: 10.1002/pdi.818]

XSL•FO
RenderX

45. Carlson BA, Neal D, Magwood G, Jenkins C, King MG, Hossler CL. A community-based participatory health information needs assessment to help eliminate diabetes information disparities. Health Promot Pract 2006 Jul;7(3_suppl):213S-222S. [doi: 10.1177/1524839906288694] [Medline: 16760247]
46. Stevenson FA, Kerr C, Murray E, Nazareth I. Information from the internet and the doctor-patient relationship: the patient perspective - a qualitative study. BMC Fam Pract 2007 Aug;8:47 [FREE Full text] [doi: 10.1186/1471-2296-8-47] [Medline: 17705836]
47. Cho AH, Arar NH, Edelman DE, Hartwell PH, Oddone EZ, Yancy WS. Do diabetic veterans use the internet? Self-reported usage, skills, and interest in using My HealtheVet Web portal. Telemed J E Health 2010 Jun;16(5):595-602. [doi: 10.1089/tmj.2009.0164] [Medline: 20575727]
48. Longo DR, Schubert SL, Wright BA, LeMaster J, Williams CD, Clore JN. Health information seeking, receipt, and use in diabetes self-management. Ann Fam Med 2010 Jul 1;8(4):334-340 [FREE Full text] [doi: 10.1370/afm.1115] [Medline: 20644188]
49. Plotnikoff RC, Johnson ST, Karunamuni N, Boule NG. Physical activity related information sources predict physical activity behaviors in adults with type 2 diabetes. J Health Commun 2010 Dec 16;15(8):846-858. [doi: 10.1080/10810730.2010.522224] [Medline: 21170787]
50. Chisolm DJ, Hardin DS, McCoy KS, Johnson LD, McAlearney AS, Gardner W. Health literacy and willingness to use online health information by teens with asthma and diabetes. Telemed J E Health 2011 Nov;17(9):676-682 [FREE Full text] [doi: 10.1089/tmj.2011.0037] [Medline: 21943161]
51. Mayberry LS, Kripalani S, Rothman RL, Osborn CY. Bridging the digital divide in diabetes: family support and implications for health literacy. Diabetes Technol Ther 2011 Oct;13(10):1005-1012. [doi: 10.1089/dia.2011.0055] [Medline: 21718098]
52. Shaw RJ, Johnson CM. Health information seeking and social media use on the internet among people with diabetes. Online J Public Health Inform 2011;3(1):E61000 [FREE Full text] [doi: 10.5210/ojphi.v3i1.3561] [Medline: 23569602]
53. Janeice Morgan A, Trauth EM. Socio-economic influences on health information searching in the USA: the case of diabetes. Inf Technol People 2013 Nov 11;26(4):324-346. [doi: 10.1108/itp-09-2012-0098]
54. Meyfroidt S, Aeyels D, Van Audenhove C, Verlinde C, Peers J, Panella M, et al. How do patients with uncontrolled diabetes in the Brussels-Capital Region seek and use information sources for their diet? Prim Health Care Res Dev 2013 Jul;14(3):229-239. [doi: 10.1017/S1463423612000205] [Medline: 22717540]
55. Nordfeldt S, Angarne-Lindberg T, Nordwall M, Ekberg J, Berterö C. As facts and chats go online, what is important for adolescents with type 1 diabetes? PLoS One 2013 Jun 21;8(6):e67659 [FREE Full text] [doi: 10.1371/journal.pone.0067659] [Medline: 23805322]
56. Connolly KK, Crosby ME. Examining e-Health literacy and the digital divide in an underserved population in Hawai'i. Hawaii J Med Public Health 2014 Feb;73(2):44-48 [FREE Full text] [Medline: 24567867]
57. Wiley J, Westbrook M, Long J, Greenfield JR, Day RO, Braithwaite J. Diabetes education: the experiences of young adults with type 1 diabetes. Diabetes Ther 2014 Jun;5(1):299-321 [FREE Full text] [doi: 10.1007/s13300-014-0056-0] [Medline: 24519150]
58. Garfield SS, Xenakis JJ, Bastian A, McBride M. Experiences of people with diabetes by payer type: an analysis of the roper diabetes data set. Diabetes Ther 2015 Jun;6(2):113-125. [doi: 10.1007/s13300-015-0109-z] [Medline: 25997415]
59. Jamal A, Khan SA, AlHumud A, Al-Duhyyim A, Alrashed M, Bin Shabr F, et al. Association of online health information-seeking behavior and self-care activities among type 2 diabetic patients in Saudi Arabia. J Med Internet Res 2015 Aug 12;17(8):e196 [FREE Full text] [doi: 10.2196/jmir.4312] [Medline: 26268425]
60. Kalantzi S, Kostagiolas P, Kechagias G, Niakas D, Makrilakis K. Information seeking behavior of patients with diabetes mellitus: a cross-sectional study in an outpatient clinic of a university-affiliated hospital in Athens, Greece. BMC Res Notes 2015 Feb 20;8:48 [FREE Full text] [doi: 10.1186/s13104-015-1005-3] [Medline: 25889724]
61. Lui CW, Col JR, Donald M, Dower J, Boyle FM. Health and social correlates of Internet use for diabetes information: findings from Australia's Living with Diabetes Study. Aust J Prim Health 2015 Sep;21(3):327-333. [doi: 10.1071/PY14021] [Medline: 25007788]
62. Morgan AJ, Ferguson YM, Trauth EM. Consumer demographics and internet based health information search in the United States: the intersectionality of gender, race, and class. Int J E-Health Med Commun 2015 Jan 1;6(1):58-72. [doi: 10.4018/ijehmc.2015010105]
63. Brady E, Segar J, Sanders C. “You get to know the people and whether they’re talking sense or not”: negotiating trust on health-related forums. Soc Sci Med 2016 Aug;162:151-157. [doi: 10.1016/j.socscimed.2016.06.029] [Medline: 27359321]
64. Fergie G, Hilton S, Hunt K. Young adults' experiences of seeking online information about diabetes and mental health in the age of social media. Health Expect 2016 Nov 23;19(6):1324-1335 [FREE Full text] [doi: 10.1111/hex.12430] [Medline: 26647109]
65. O'Kane AA, Park SY, Mentis H, Blandford A, Chen Y. Turning to peers: integrating understanding of the self, the condition, and others' experiences in making sense of complex chronic conditions. Comput Support Coop Work 2016 Dec;25(6):477-501 [FREE Full text] [doi: 10.1007/s10606-016-9260-y] [Medline: 32355411]
66. Weymann N, Härter M, Dirmaier J. Information and decision support needs in patients with type 2 diabetes. Health Informatics J 2016 Mar;22(1):46-59 [FREE Full text] [doi: 10.1177/1460458214534090] [Medline: 24916569]

XSL•FO
RenderX

67. Aponte J, Nokes KM. Electronic health literacy of older Hispanics with diabetes. Health Promot Int 2017 Jun 1;32(3):482-489. [doi: 10.1093/heapro/dav112] [Medline: 26681770]
68. Brady E, Segar J, Sanders C. Accessing support and empowerment online: the experiences of individuals with diabetes. Health Expect 2017 Sep 15;20(5):1088-1095 [FREE Full text] [doi: 10.1111/hex.12552] [Medline: 28718928]
69. Mathiesen AS, Thomsen T, Jensen T, Schiøtz C, Langberg H, Egerod I. The influence of diabetes distress on digital interventions for diabetes management in vulnerable people with type 2 diabetes: a qualitative study of patient perspectives. J Clin Transl Endocrinol 2017 Sep;9:41-47 [FREE Full text] [doi: 10.1016/j.jcte.2017.07.002] [Medline: 29067269]
70. Vitger T, Langberg H, Grabowski D. Mutual involvement in families with type 2 diabetes through web-based health care solutions: quantitative survey study of family preferences, challenges, and potentials. JMIR Diabetes 2017 Sep 27;2(2):e23 [FREE Full text] [doi: 10.2196/diabetes.7456] [Medline: 30291067]
71. Leelavathi M, Teh Rohaila J, Thenmolle A, Mohamad Salehuddin MS, Barathan M, Natasha PJ, et al. Online health information seeking among people with diabetes mellitus and its association with self-management. Med & Health 2018 Dec;13(2):72-82. [doi: 10.17576/MH.2018.1302.07]
72. Martis R, Brown J, McAra-Couper J, Crowther CA. Enablers and barriers for women with gestational diabetes mellitus to achieve optimal glycaemic control – a qualitative study using the theoretical domains framework. BMC Pregnancy Childbirth 2018 Apr 11;18(1):91. [doi: 10.1186/s12884-018-1710-8] [Medline: 29642898]
73. Dayyani I, Terkildsen Maindal H, Rowlands G, Lou S. A qualitative study about the experiences of ethnic minority pregnant women with gestational diabetes. Scand J Caring Sci 2019 Sep 11;33(3):621-631. [doi: 10.1111/scs.12655] [Medline: 30653703]
74. Maneze D, Weaver R, Kovai V, Salamonson Y, Astorga C, Yogendran D, et al. “Some say no, some say yes”: receiving inconsistent or insufficient information from healthcare professionals and consequences for diabetes self-management: a qualitative study in patients with type 2 diabetes. Diabetes Res Clin Pract 2019 Oct;156:107830. [doi: 10.1016/j.diabres.2019.107830] [Medline: 31465812]
75. Lee JW, Theng Y, Lee SW. Health information seeking behaviour using mobile devices among people with diabetes: a comparison between Middle and high income country. Digit Health 2020 Sep 11;6:2055207620956457. [doi: 10.1177/2055207620956457]
76. Zhang X, Foo S, Majid S, Chang YK, Dumaual HTJ, Suri VR. Self-care and health-information-seeking behaviours of diabetic patients in Singapore. Health Commun 2019 Jul 15;35(8):994-1003. [doi: 10.1080/10410236.2019.1606134]
77. Abdel Nasser A, Alzahrani RM, Ghandoura AN, Sultan I. Use of electronic health (eHealth) among Saudi type 2 diabetic patients and its association with their diabetic self-management: a cross-sectional study. Cureus 2021 Mar 14;13(3):e13882 [FREE Full text] [doi: 10.7759/cureus.13882] [Medline: 33868846]
78. Alvarez-Perez Y, Perestelo-Perez L, Rivero-Santana A, Wagner AM, Torres-Castaño A, Toledo-Chávarri A, IC-Health Project Consortium. Cocreation of massive open online courses to improve digital health literacy in diabetes: pilot mixed methods study. JMIR Diabetes 2021 Dec 13;6(4):e30603 [FREE Full text] [doi: 10.2196/30603] [Medline: 34898453]
79. Dehnavi Z, Ayatollahi H, Hemmat M, Abbasi R. Patients' and physicians' perspectives about using health information technology in diabetes management in Iran: a qualitative study. Perspect Health Inf Manag 2021;18(4):1i [FREE Full text] [Medline: 34975358]
80. Edwards KJ, Bradwell HL, Jones RB, Andrade J, Shawe JA. How do women with a history of gestational diabetes mellitus use mHealth during and after pregnancy? Qualitative exploration of women's views and experiences. Midwifery 2021 Jul;98:102995. [doi: 10.1016/j.midw.2021.102995] [Medline: 33784541]
81. Guo SHM, Hsing HC, Lin JL, Lee CC. Relationships between mobile eHealth literacy, diabetes self-care, and glycemic outcomes in Taiwanese patients with type 2 diabetes: cross-sectional study. JMIR Mhealth Uhealth 2021 Feb 05;9(2):e18404 [FREE Full text] [doi: 10.2196/18404] [Medline: 33544088]
82. Hughes A, Heydarian N, Gerardo D, Solis I, Morera O. Seeking health information and social support in the diabetes online community. Front Clin Diabetes Healthc 2021 Sep 28;2:708405. [doi: 10.3389/fcdhc.2021.708405] [Medline: 36994327]
83. Siti Zuhaida H, Chung HC, Mohd Said F, Tumingan K, Sahar Shah N, Abu Bakar SH. The best online tools based on media preference reflected by health information received on social media amongst diabetic patients in Hospital Canselor Tuanku Muhriz, Kuala Lumpur, Malaysia. Malays J Med Sci 2021 Jun 30;28(3):118-128 [FREE Full text] [doi: 10.21315/mjms2021.28.3.11] [Medline: 34285650]
84. Kostagiolas P, Tsiligros P, Theodorou P, Tentolouris N, Niakas D. A cross-sectional survey interconnecting health information seeking behavior with clinical data of type 2 diabetes mellitus patients. Library Hi Tech 2021 Jun 21;39(2):448-461. [doi: 10.1108/lht-02-2020-0030]
85. Mansour E. An investigation into the information-seeking behavior of Egyptian adult patients with type 2 diabetes mellitus (T2DM). J Hosp Librariansh 2021 Oct 20;21(4):367-390. [doi: 10.1080/15323269.2021.1982260]
86. Nazir M, Soroya SH. Health informatics: use of internet for health information seeking by Pakistani chronic patients. J Libr Adm 2021 Jan 04;61(1):134-146. [doi: 10.1080/01930826.2020.1845552]
87. Sjöström A, Hajdarevic S, Hörnsten Å, Öberg U, Isaksson U. Experiences of online COVID-19 information acquisition among persons with type 2 diabetes and varying eHealth literacy. Int J Environ Res Public Health 2021 Dec 15;18(24):13240. [doi: 10.3390/ijerph182413240] [Medline: 34948852]

XSL•FO
RenderX

88. Zhao N, Luo X, Zhang H, Dai R, Pan W, Hambly BD, et al. Cross-sectional study on health literacy and internet accessibility among patients with DM in Gansu, China. Front Public Health 2021 Oct 13;9:692089 [FREE Full text] [doi: 10.3389/fpubh.2021.692089] [Medline: 34722434]
89. Al Nozha OM, Elshatarat RA. The association between online health-related information seeking behavior and self-care activities among patients with diabetes mellitus in Saudi Arabia. Univers J Public Health 2022 Aug;10(4):304-315. [doi: 10.13189/ujph.2022.100404]
90. Broekhuis M, van Velsen L, De Franco D, Pease A, Hermens H. Contextual health information behavior in the daily lives of people with type 2 diabetes: a diary study in Scotland. Health Commun 2022 Mar;37(3):272-284. [doi: 10.1080/10410236.2020.1837426] [Medline: 33135473]
91. Freeman T, Fisher M, Foley K, Boyd MA, Ward PR, McMichael G, et al. Barriers to digital health services among people living in areas of socioeconomic disadvantage: research from hospital diabetes and antenatal clinics. Health Prom J Austr 2022 Jul;33(3):751-757. [doi: 10.1002/hpja.540] [Medline: 34510601]
92. Langford AT, Orellana KT, Buderer N. Use of YouTube to watch health-related videos and participation in online support groups among US adults with heart disease, diabetes, and hypertension. Digit Health 2022 Aug 24;8:20552076221118822. [doi: 10.1177/20552076221118822] [Medline: 36046636]
93. Rajanala S, Wilson JK, Mitchell PD, Garvey KC, Fishman LN. Contrasting social media use between young adults with inflammatory bowel disease and type 1 diabetes: cross-sectional study. JMIR Pediatr Parent 2022 Apr 25;5(2):e34466 [FREE Full text] [doi: 10.2196/34466] [Medline: 35139484]
94. Subramaniam M, Lau JH, Asharani PV, Abdin E, Roystonn K, Wang P, et al. Sources of information on diabetes and its demographic correlates: a nationwide survey among Singapore residents. Health Promot Int 2022 Oct 26;37(5):daac107 [FREE Full text] [doi: 10.1093/heapro/daac107] [Medline: 36287520]
95. Terkeş N, Çelik F, Bektaş H. Determination of the technology use status and attitudes towards technology of patients with type 2 diabetes. Cyprus J Med Sci 2022 Sep 22;7(4):477-483. [doi: 10.4274/cjms.2021.2021-82]
96. Wang SY, Yeh HC, Stein AA, Miller ER. Use of health information technology by adults with diabetes in the United States: cross-sectional analysis of National Health Interview Survey Data (2016-2018). JMIR Diabetes 2022 Jan 12;7(1):e27220 [FREE Full text] [doi: 10.2196/27220] [Medline: 35019844]
97. Costa IG, Camargo-Plazas P. The processes of engagement in information-seeking behavior for individuals with diabetes who developed diabetic foot ulcer: a constructivist grounded theory study. Digit Health 2023 Jun 19;9:20552076231177155. [doi: 10.1177/20552076231177155] [Medline: 37361440]
98. Andersen AMJ, Jervelund SS, Maindal HT, Hempler NF. Acquisition, application, and distribution of health literacy from culturally sensitive type 2 diabetes education among Arabic-Speaking migrants in Denmark: a longitudinal qualitative analysis. Scand J Caring Sci 2024 May 22;38(2):523-535. [doi: 10.1111/scs.13228] [Medline: 38031875]
99. Arueyingho O, Aprioku JS, Marshall P, O'Kane AA. Insights Into sociodemographic influences on type 2 diabetes care and opportunities for digital health promotion in Port Harcourt, Nigeria: quantitative study. JMIR Diabetes 2024 Aug 21;9:e56756 [FREE Full text] [doi: 10.2196/56756] [Medline: 39167439]
100. Dsouza SM, Venne J, Shetty S, Brand H. Identification of challenges and leveraging mHealth technology, with need-based solutions to empower self-management in type 2 diabetes: a qualitative study. Diabetol Metab Syndr 2024 Jul 30;16(1):182. [doi: 10.1186/s13098-024-01414-9] [Medline: 39080764]
101. Meer Z, Al-Ozairi E, Fernandes G, Ranganathan S, Patel J. Information behaviours of people with type 2 diabetes in Kuwait: a grounded theory study. BMC Prim Care 2024 Sep 4;25:326. [doi: 10.1186/s12875-024-02577-0] [Medline: 39232660]
102. Naef AN, Fischbock N, Tezcan-Güntekin H, Amelung VE. Adolescents with type 1 diabetes' perspectives on digital health interventions to enhance health literacy: a qualitative study. Front Public Health 2024 Feb 13;12:1340196. [doi: 10.3389/fpubh.2024.1340196] [Medline: 38414892]
103. Ouedraogo I, Some BMJ, Benedikter R, Diallo G. Health and eHealth literacy of patients with diabetes in low-income countries: perspective from Guinea and Burkina Faso. JMIR Diabetes 2024 Dec 03;9:e55677 [FREE Full text] [doi: 10.2196/55677] [Medline: 39626226]
104. Peimani M, Stewart AL, Ghodssi-Ghassemabadi R, Nasli-Esfahani E, Ostovar A. The moderating role of e-health literacy and patient-physician communication in the relationship between online diabetes information-seeking behavior and self-care practices among individuals with type 2 diabetes. BMC Prim Care 2024 Dec 30;25:442. [doi: 10.1186/s12875-024-02695-9] [Medline: 39736551]
105. Roesler A, Butten K, Taylor P, Morrison M, Varnfield M, Holmes‐Truscott E. The experiences of individuals who have had gestational diabetes: a qualitative exploration. Diabet Med 2024 Oct 15;41(11):e15374. [doi: 10.1111/dme.15374] [Medline: 38853396]
106. Ji X, Chi J. Exploring the relationship between eHealth literacy and diabetes knowledge, self-efficacy, and self-care behaviors in Chinese diabetic patients: a cross-sectional study. J Nurs Res 2024 Dec;32(6):e359. [doi: 10.1097/jnr.0000000000000642] [Medline: 39593226]
107. Xu J, Chen Y, Zhao J, Wang J, Chen J, Pan X, et al. Current status of electronic health literacy among pregnant women with gestational diabetes mellitus and their perceptions of online health information: a mixed-methods study. BMC Pregnancy Childbirth 2024 May 28;24:392. [doi: 10.1186/s12884-024-06594-w] [Medline: 38807050]

XSL•FO
RenderX

108. Birati Y, Yefet E, Shavit LY, Spitzer S. Bridging the gap: understanding preferences and perceived benefits of an mHealth app for GDM self-management. Health Policy Technol 2025 Jun;14(3):101011. [doi: 10.1016/j.hlpt.2025.101011]
109. Jeon S, Lee S, Kim EH, Eun J, Lee K, Lim H, et al. Generative AI chatbot for diabetes management: formative 2-part qualitative study using DTalksBot involving patients and clinicians. JMIR Form Res 2025 Nov 12;9:e72553 [FREE Full text] [doi: 10.2196/72553] [Medline: 41223424]
110. Kilinç Işleyen E, Özdemir İ. The relationship between e-Health literacy and health promotion self-care behaviors and predictors of e-Health literacy in individuals with type 2 diabetes mellitus: a cross-sectional study. Turkiye Klinikleri J Nurs Sci 2025 Feb 13;17(1):98-107. [doi: 10.5336/nurses.2024-104274]
111. Maxwell T, Branka L, Asher N, Commissariat P, Laffel L. Young adults with type 1 diabetes and their perspectives on diabetes-related social media: qualitative study. JMIR Diabetes 2025 Jun 25;10:e69243 [FREE Full text] [doi: 10.2196/69243] [Medline: 40561554]
112. Sadanandam P, Sreenivas N, Simon S, Sreenivas G, Bodduna P. Digital health information seeking behaviour among diabetics: a cross sectional study. In J Life Sci Biotechnol Pharma Res 2025 Sep 24;14(9):1871. [doi: 10.69605/ijlbpr_14.9.2025.311]
113. Hyman I, Patychuk D, Zaidi Q, Kljujic D, Shakya YB, Rummens JA, et al. Self-management, health service use and information seeking for diabetes care among recent immigrants in Toronto. Chronic Dis Inj Can 2012 Dec;33(1):12-18. [doi: 10.24095/hpcdp.33.1.02] [Medline: 23294917]
114. St. Jean B. Factors motivating, demotivating, or impeding information seeking and use by people with type 2 diabetes: a call to work toward preventing, identifying, and addressing incognizance. J Assoc Inf Sci Technol 2017 Jan 20;68(2):309-320. [doi: 10.1002/asi.23652]
115. Mengiste M, Ahmed MH, Bogale A, Yilma T. Information-seeking behavior and its associated factors among patients with diabetes in a resource-limited country: a cross-sectional study. Diabetes Metab Syndr Obes 2021 May 17;14:2155-2166 [FREE Full text] [doi: 10.2147/DMSO.S289905] [Medline: 34040402]
116. Soroya SH, Ilyas A, Ameen K. Understanding information behavior of diabetic patients: a case of the diabetic's institute Pakistan. Libr Manag 2021 Dec 11;42(1-2):1-21. [doi: 10.1108/lm-03-2020-0041]
117. Yao M, Zhang DY, Fan JT, Lin K, Haroon S, Jackson D, et al. The experiences of people with type 2 diabetes in communicating with general practitioners in China – a primary care focus group study. BMC Prim Care 2022 Feb 3;23:24. [doi: 10.1186/s12875-022-01632-y] [Medline: 35172752]
118. Oser TK, Oser SM, Parascando JA, Hessler-Jones D, Sciamanna CN, Sparling K, et al. Social media in the diabetes community: a novel way to assess psychosocial needs in people with diabetes and their caregivers. Curr Diab Rep 2020 Feb 20;20(3):10. [doi: 10.1007/s11892-020-1294-3] [Medline: 32080765]
119. Rueger J, Dolfsma W, Aalbers R. Perception of peer advice in online health communities: access to lay expertise. Soc Sci Med 2021 May;277:113117. [doi: 10.1016/j.socscimed.2020.113117] [Medline: 33865092]
120. Ardissone A, Leonowicz-Bukała I, Struck-Peregończyk M. “Can anyone tell me…”. Online health communities in diabetes self-management in Poland and Italy. Health Commun 2025 Mar;40(3):492-499. [doi: 10.1080/10410236.2024.2348842] [Medline: 38687112]
121. Litchman ML, Edelman LS, Donaldson GW. Effect of diabetes online community engagement on health indicators: cross-sectional study. JMIR Diabetes 2018 Apr 24;3(2):e8 [FREE Full text] [doi: 10.2196/diabetes.8603] [Medline: 30291079]
122. Buzás N, Horváth MD, Tesch Z, Hallgató E. How online peer support affects management efficacy and mitigates difficulties of parents caring for children with type 1 diabetes. Prim Care Diabetes 2023 Dec;17(6):607-611. [doi: 10.1016/j.pcd.2023.09.003] [Medline: 37730520]
123. Vasistha S, Kanchibhatla A, Blanchette JE, Rieke J, Hughes AS. The sugar-coated truth: the quality of diabetes health information on TikTok. Clin Diabetes 2025;43(1):53-58. [doi: 10.2337/cd24-0042] [Medline: 39829690]
124. Kong W, Song S, Zhao YC, Zhu Q, Sha L. TikTok as a health information source: assessment of the quality of information in diabetes-related videos. J Med Internet Res 2021 Sep 01;23(9):e30409 [FREE Full text] [doi: 10.2196/30409] [Medline: 34468327]
125. Silverman JA, Chugh A, Hollier JM, Martin N, Raghu VK, Rosas‐Blum E, et al. Using social media for patient care, research, and professional development: a North American Society of Pediatric Gastroenterology, Hepatology, and Nutrition position paper. J Pediatr Gastroenterol Nutr 2024 Feb 19;78(2):414-427. [doi: 10.1002/jpn3.12051] [Medline: 38299267]
126. Wardle C, Urbani S, Wang E. Evolving health information-seeking behavior in the context of Google AI overviews, ChatGPT, and Alexa: interview study using the think-aloud protocol. J Med Internet Res 2025 Oct 07;27:e79961 [FREE Full text] [doi: 10.2196/79961] [Medline: 41055948]
127. Al Shboul MKI, Alwreikat A, Alotaibi FA. Investigating the use of ChatGPT as a novel method for seeking health information: a qualitative approach. Science & Technology Libraries 2023 Sep 08;43(3):225-234. [doi: 10.1080/0194262x.2023.2250835]
128. Ayre J, Mac O, McCaffery K, McKay BR, Liu M, Shi Y, et al. New frontiers in health literacy: using ChatGPT to simplify health information for people in the community. J Gen Intern Med 2024 Mar;39(4):573-577. [doi: 10.1007/s11606-023-08469-w] [Medline: 37940756]

XSL•FO
RenderX

129. Esmaeilzadeh P, Maddah M, Mirzaei T. Using AI chatbots (e.g., CHATGPT) in seeking health-related information online: the case of a common ailment. Comput Hum Behav 2025 Mar;3:100127. [doi: 10.1016/j.chbah.2025.100127]
130. Hwang M, Zheng Y, Cho Y, Jiang Y. AI applications for chronic condition self-management: scoping review. J Med Internet Res 2025 Apr 08;27:e59632 [FREE Full text] [doi: 10.2196/59632] [Medline: 40198108]
131. Yau JYS, Saadat S, Hsu E, Murphy LSL, Roh JS, Suchard J, et al. Accuracy of prospective assessments of 4 large language model chatbot responses to patient questions about emergency care: experimental comparative study. J Med Internet Res 2024 Nov 04;26:e60291 [FREE Full text] [doi: 10.2196/60291] [Medline: 39496149]
132. Kuroiwa T, Sarcon A, Ibara T, Yamada E, Yamamoto A, Tsukamoto K, et al. The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. J Med Internet Res 2023 Sep 15;25:e47621 [FREE Full text] [doi: 10.2196/47621] [Medline: 37713254]
133. Deiana G, Dettori M, Arghittu A, Azara A, Gabutti G, Castiglia P. Artificial intelligence and public health: evaluating ChatGPT responses to vaccination myths and misconceptions. Vaccines (Basel) 2023 Jul 7;11(7):1217 [FREE Full text] [doi: 10.3390/vaccines11071217] [Medline: 37515033]
134. Wu PF, Summers C, Panesar A, Kaura A, Zhang L. AI hesitancy and acceptability-perceptions of AI chatbots for chronic health management and long COVID support: survey study. JMIR Hum Factors 2024 Jul 23;11:e51086 [FREE Full text] [doi: 10.2196/51086] [Medline: 39045815]
135. Chan HY, Wang CC, Jeng W, Huang YM. Strengthening scientific credibility in the face of misinformation and disinformation: viable solutions. J Control Release 2023 Aug;360:163-168. [doi: 10.1016/j.jconrel.2023.05.036] [Medline: 37301268]
136. Fitzpatrick PJ. Improving health literacy using the power of digital communications to achieve better health outcomes for patients and practitioners. Front Digit Health 2023 Nov 17;5:1264780. [doi: 10.3389/fdgth.2023.1264780] [Medline: 38046643]
137. Rees G, Nowell L, Risling T. Shaping the future of digital health education in Canada: prioritizing competencies for health care professionals using the quintuple aim. JMIR Med Educ 2025 Sep 08;11:e75904 [FREE Full text] [doi: 10.2196/75904] [Medline: 40920576]
138. Mirzaei A, Aslani P, Luca EJ, Schneider CR. Predictors of health information-seeking behavior: systematic literature review and network analysis. J Med Internet Res 2021 Jul 2;23(7):e21680 [FREE Full text] [doi: 10.2196/21680] [Medline: 33979776]
139. Gwira JA, Fryar CD, Gu Q. Prevalence of total, diagnosed, and undiagnosed diabetes in adults: United States, August 2021-August 2023. NCHS Data Brief 2024 Nov(516):CS354814. [doi: 10.15620/cdc/165794] [Medline: 40085919]
140. Nouri S, Khoong EC, Lyles CR, Karliner L. Addressing equity in telemedicine for chronic disease management during the Covid-19 pandemic. NEJM Catal Innov Care Deliv 2020 May 4;1(3):1-13. [doi: 10.1056/CAT.20.0123]
141. Huang YM, Pecanac KE, Shiyanbola OO. “Why am I not taking medications?” Barriers and facilitators of diabetes medication adherence across different health literacy levels. Qual Health Res 2020 Dec;30(14):2331-2342. [doi: 10.1177/1049732320945296] [Medline: 32723206]
142. Lee ST, Lin J. A self-determination perspective on online health information seeking: the internet vs. face-to-face office visits with physicians. J Health Commun 2016 Jun;21(6):714-722. [doi: 10.1080/10810730.2016.1157651]
143. Link E, Baumann E, Klimmt C. Explaining online information seeking behaviors in people with different health statuses: German representative cross-sectional survey. J Med Internet Res 2021 Dec 10;23(12):e25963 [FREE Full text] [doi: 10.2196/25963] [Medline: 34890348]
144. Soroya SH, Farooq A, Mahmood K, Isoaho J, Zara SE. From information seeking to information avoidance: understanding the health information behavior during a global health crisis. Inf Process Manag 2021 Mar;58(2):102440. [doi: 10.1016/j.ipm.2020.102440] [Medline: 33281273]
145. Abreu L, Nunes JA, Taylor P, Silva S. Distributed health literacy among people living with type 2 diabetes in Portugal: defining levels of awareness and support. Health Soc Care Community 2018 Jan;26(1):90-101. [doi: 10.1111/hsc.12465] [Medline: 28643399]
146. Muscat DM, Gessler D, Ayre J, Norgaard O, Heuck IR, Haar S, et al. Seeking a deeper understanding of ‘distributed health literacy’: a systematic review. Health Expect 2022 May 20;25(3):856-868. [doi: 10.1111/hex.13450] [Medline: 35178823]
147. Mainous AG, Liu-Galvin R, Durden B, Yin L, Saguil AA. Patient trust in the health system, internet information searching and the patient-provider relationship. Front Med (Lausanne) 2025 Nov 20;12:1665927. [doi: 10.3389/fmed.2025.1665927]
148. Luo A, Qin L, Yuan Y, Yang Z, Liu F, Huang P, et al. The effect of online health information seeking on physician-patient relationships: systematic review. J Med Internet Res 2022 Feb 10;24(2):e23354 [FREE Full text] [doi: 10.2196/23354] [Medline: 35142620]
149. Lu Q, Schulz PJ. Physician perspectives on internet-informed patients: systematic review. J Med Internet Res 2024 Jun 06;26:e47620 [FREE Full text] [doi: 10.2196/47620] [Medline: 38842920]
150. Kirchberg J, Fritzmann J, Weitz J, Bork U. eHealth literacy of German physicians in the pre-COVID-19 era: questionnaire study. JMIR Mhealth Uhealth 2020 Oct 16;8(10):e20099 [FREE Full text] [doi: 10.2196/20099] [Medline: 33064102]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
HCP: health care professional
MeSH: Medical Subject Headings
OHIS: online health information seeking
PEMAT: Patient Education Materials Assessment Tool
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
SPIDER: Sample, Phenomenon of Interest, Design, Evaluation, Research type

Edited by S Brini; submitted 15.Aug.2025; peer-reviewed by N Buzas; comments to author 04.Feb.2026; revised version received 02.Jun.2026; accepted 03.Jun.2026; published 16.Jun.2026.

Please cite as:
Tiao PH, Chan HY, Huang YM
Navigating Diabetes Management in the Digital Era: Scoping Review of Online Health Information-Seeking Behavior
J Med Internet Res 2026;28:e82081
URL: https://www.jmir.org/2026/1/e82081
doi:10.2196/82081
PMID:

©Pin-Heng Tiao, Hsun-Yu Chan, Yen-Ming Huang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Warmth of Medicine: The Irreplaceability of Doctors in the Era of Artificial Intelligence

Yu-ying Yang, MD, PhD; Jian-min Liu, MD, PhD

Department of Endocrine and Metabolic Diseases, Rui-jin Hospital, Shanghai Jiao Tong University School of Medicine; Shanghai Institute of Endocrine and Metabolic Diseases; National Clinical Research Center for Metabolic Diseases (Shanghai), 197 Ruijin Er Road, Shanghai, China

Corresponding Author:
Jian-min Liu, MD, PhD
Department of Endocrine and Metabolic Diseases, Rui-jin Hospital, Shanghai Jiao Tong University School of Medicine; Shanghai Institute of Endocrine and Metabolic Diseases; National Clinical Research Center for Metabolic Diseases (Shanghai), 197 Ruijin Er Road, Shanghai, China

Abstract

This article argues that despite the remarkable advances of artificial intelligence (AI) in medicine —including demonstrated capabilities in image recognition, diagnosis, treatment planning, and even empathic communication in controlled settings—the core of medical practice remains irreducibly human. We identify three domains in which AI cannot replace doctors: the holistic, sensory art of clinical observation and intuition; the longitudinal, trust-based doctor-patient relationship built on genuine emotional connection; and the capacity to embrace clinical uncertainty, exercise moral responsibility, and make courageous decisions in the absence of algorithmic guidance. The intended audience includes clinical doctors, medical students, medical educators, and health policy makers navigating the integration of AI into practice. We conclude that preserving "AI-free clinical time" in medical training and safeguarding the humanistic dimensions of care are essential, and technology is to complement rather than diminish the healing arts.

(J Med Internet Res 2026;28:e97412) doi:10.2196/97412

KEYWORDS

artificial intelligence; doctors; patients; medicine; medical humanities

Introduction

With the rapid advancement of artificial intelligence (AI) technologies, including large language models (LLMs) and multimodal AI, the field of medicine is undergoing unprecedented transformation. AI has made great progress in medical image analysis, nonimage data mining, novel learning paradigms, and human-AI collaborative workflows, and has demonstrated remarkable capabilities in areas such as image recognition, disease prediction, diagnosis, summarizing guidelines, and treatment planning [1,2]. However, no matter how technology progresses, the essence of medicine remains an art that integrates science and humanity, with its core being: "confidence, confidentiality, competence, contract, community responsibility, and commitment" as proposed by David Morrell at 1994 British Medical Association Summit, and then rephrased to "caring, integrity, competence, confidentiality, responsibility, and advocacy" by the conference steering group [3].

It is true that recent clinical studies, including meta-analyses, systematic reviews, and trials, prove AI chatbots and LLM-based AI systems can match or even surpass clinicians in history-taking, following diagnostic pathways, and could deliver more consistent and superior empathetic communication [4-8]. However, these studies were mainly nonclinical, conducted in public online forums and simulated clinical settings, using text-based interactions and standardized patient actors rather than real clinical practice [4-8]. In actual clinical scenarios, the performance of AI in aspects of long-term patient relationships, nonverbal communication rather than synchronous text chat, embodied presence, and professional accountability, especially under clinical conditions of uncertainty and ambiguity, is still not confirmed and is even hard to test. Thus, the notion that AI could replace the central role of doctors is premature and could reflect a "myopic view" [9].

In this paper, we identify three domains in which AI cannot replace doctors: (1) the art of clinical observation and intuition; (2) the trust-based doctor-patient relationship built on genuine emotional connection; and (3) the capacity to embrace clinical uncertainty, exercise moral responsibility, and make courageous decisions in the absence of algorithmic guidance. It is widely accepted that the ultimate goal of medicine is not merely to cure diseases but to care for life—doctors treat the diseases, but more importantly, the patients, the individuals with emotions and dignity. In an era dominated by AI, it is even more crucial for doctors to embody and demonstrate the warmth, wisdom, and value of being human.

XSL•FO
RenderX

The Art of Medicine: The Significance of Comprehensive History-Taking and Physical Examination

AI can rapidly analyze vast amounts of data, but it cannot fully replace face-to-face communication between doctors and patients. Detailed history-taking and physical examination form the foundation of diagnosis. This process relies not only on objective information but also on the doctors' decades-honed expertise, experience, intuition, and humanistic care. A patient's smile or a relaxed conversation can indicate improvement in their condition, while beads of sweat the size of soybeans on the patient's forehead, cold hands, a trembling voice, and a palpable firm mass allow the physician to gauge the severity and even the malignancy of the disorder. Such comprehensive clinical observation and thinking is an "art" that is difficult for AI to replicate.

We acknowledge that in controlled research settings—particularly text-based evaluations such as Objective Structured Clinical Examination (OSCE)-style assessments—AI models have demonstrated comparable and even better performance than clinicians in history-taking and diagnostic reasoning [4,6]. These findings are valuable and suggest that AI can meaningfully assist in structured clinical encounters. However, the sensory and relational dimensions of bedside medicine—observation of subtle physical signs, tactile information from palpation, and clinical reasoning and judgment abilities formed over years of practice—remain domains in which algorithmic processing cannot yet substitute for embodied human presence.

The Cornerstone of Doctor-Patient Trust: The Power of Emotional Connection

Doctor-patient trust is "a real human relationship based on love, caring, and sharing" [3]. This trust is built not only on technical competence but also on the empathy and sense of responsibility demonstrated by doctors. When facing illness, patients' fear and anxiety often extend far beyond the pathology itself. They need love [10]. A comforting word from the doctor, an explanation with patience, or even a reassuring glance can become the motivation for patients to persevere.

AI may be able to provide the "optimal treatment plan," but it cannot understand a mother's anguish over her child's illness or comprehend the desperation of a cancer patient facing death. When a doctor hears a patient with advanced cancer weakly murmur "Why am I still alive?" after a bout of drowsiness, what rises in the doctor's heart is a complex and heavy mix of emotions—empathy for the patient's suffering, helplessness at the limitations of medicine, and the struggle with professional ethics. Only through empathy can a doctor help patients and their families understand the disease, accept reality, and participate in decision-making.

This "trust relationship" and empathy are vital components of treatment and the core of medical humanism, which goes far beyond the mechanical delivery of medical information—it requires a keen awareness of the patient's emotional shifts. When explaining an unfavorable prognosis to patients and their relatives, a doctor should sit facing them, and even hold the patient's hand, talk *with* rather than talk *to* them. The essence of medicine lies in two-way communication, not one-way information delivery. When a patient hears a diagnosis, their eyes may redden, their voice tremble—subtle body language signaling fear, grief, or helplessness. A doctor can instantly adjust their tone, slow their speech, and say, "I understand this is difficult for you and your family, but we'll work through it together." This immediate genuine emotional response is based on the doctor's own professional characteristics and the empathy developed through the diagnostic and treatment process with patients.

It is noteworthy that current studies challenged the assumption that AI cannot demonstrate empathy. It was found that chatbot responses to patient questions were preferred over physician responses in 78.6% of evaluations and were rated "empathetic or very empathetic" at a rate of 45.1% versus 4.6% for physicians [4], a meta-analysis also demonstrated a mean difference of 0.87 favoring AI over human health care professionals in empathic communication [8]. However, these studies were largely based on single-encounter, text-based interactions evaluated by third-party raters or standardized patients—differs in important ways from the empathy required in ongoing therapeutic relationships. In addition, although AI might also provide some emotionally supportive notes to patients, sometimes its responses are modulated by an inferred assessment of the user's underlying emotional needs [11]. The longitudinal trust that develops through repeated face-to-face encounters, the shared emotional experience across a disease trajectory, the non-verbal communication of presence during moments of crisis, and the moral commitment to accompany a patient through suffering—these dimensions of the doctor-patient relationship have not yet been systematically evaluated in AI-comparison studies. It is not clear whether patients would maintain the same preference for AI communication when facing a serious diagnosis in person, when making complex trade-off decisions over months of treatment, or at the end of life.

What is more, it cannot be ignored that patients also observe doctors in daily clinical practice to establish his own trust to doctors. Through direct observation and intuition, patients can discern from doctors' words and demeanor whether the doctors are professional, reliable and steadfast in duty, or arrogant, pretentious and reckless; whether the doctors are ready to take on responsibilities or shift blame when difficulties occur. Such information that patients gain from observing doctors cannot be acquired through AI at present, yet it exerts a vital influence on patients' decision-making and is closely associated with their physical health and even life safety.

The Boundaries of Medicine: The "Uncertainty" Beyond AI's Reach

The value of AI in medicine cannot be judged only by medical efficiency but must also consider the patient's perspective, including benefits and risks [12]. Interindividual differences—

XSL•FO
RenderX

lifestyle habits, social and family background, psychological factors, education level, economic status, and trust in physicians—all influence treatment adherence and outcomes, whereas AI's "standardized" recommendations may overlook these "non-technical factors."

Senior doctors learn from both successes and failures. The most valuable aspect of clinical practice lies precisely in those "lessons from failure." Although extensive quantities of clinical data with negative outcomes are reported in medical corpora and would be used to train AI models, the medical AI systems still face well-documented limitations, such as representational bias [13] and generalizability failure [14]. For example, when using commercial prediction algorithms to manage the health of populations, they exhibited significant racial bias, possibly because the algorithm predicts health care costs rather than illness. This finding suggests the importance of the choice of the label on which the algorithm is trained [13]. Another study revealed that most existing COVID-19 prognostic models show decent discriminative ability but carry high or unclear risk of bias. Their reported performance is overoptimistic and not generalizable to target populations. Such bias arises from model overfitting, improper evaluation (eg, neglect of calibration), and inappropriate management of missing data; thus, "this oversupply of insufficiently validated models is not useful for clinical practice" [14]. Health and disease are complex and holistic; the selection of labels to measure health and diseases is important, while also confusing in constructing a prediction algorithm [13]. Lack of comprehensive understanding of the complexities and uncertainties inherent in medical practice might render their recommendations harboring critical blind spots.

Another concern is that with the widespread application of AI in clinical settings, the new generation of doctors may become increasingly—or even excessively—reliant on algorithmic recommendations, gradually losing their own ability to think independently [2], they may hesitate to explore innovative treatment approaches, and lack the courage to take responsibility for medical decisions.

Furthermore, there are elements in medical practice that AI cannot replace: when technological means are exhausted, a physician's clinical judgment and perseverance often become decisive factors. When lab or imaging test results are ambiguous, a doctor's experience and intuition may be life-saving. In moments of public health crises, when confronting unknown lethal diseases without AI support, doctors still charge forward into life-or-death rescues without hesitation. This embodies the sacred and irreplaceable nature of the medical profession.

Summary: The Symbiosis of Technology and Humanity

AI is a powerful assistant to doctors, but never a replacement. Technological advancement should not come at the cost of medicine's humanistic care. While technology can optimize workflows and reduce errors, the soul of medicine will always lie in person-to-person compassion. In the era of AI, we must emphasize the core values of doctors more than ever: clinical wisdom, humanistic empathy, responsibility, and courage to make decisions.

The future of medicine is not about machines replacing humans, but about human-machine collaboration—where "AI empowers doctors," combining AI's precision with human doctors' clinical insight to serve patients' holistic health needs better.

Therefore, in training the next generation of doctors, we must preserve sufficient "AI-free clinical time." For patients, we should understand that no matter how technology evolves, what they need is not just precise data analysis but also a doctor who listens to their suffering, honors life, offers warmth, and lights their way through darkness—even at life's final moments. We must safeguard those irreplaceable moments of care—the reassuring smile, the empathetic "I understand your worry," the intuitive yet life-saving decision. This is the timeless essence of medicine and the fundamental reason human physicians can never be replaced.

Technological progress should never come at the expense of human connection. Science determines how fast we can move forward in medicine; it is humanism that defines how far we can go. Only the warmth of doctors can transform healthcare into a true healing art—one that never loses medicine's most precious qualities: its humanity and soul.

Acknowledgments

No generative AI was used in the writing of this manuscript.

Funding

No financial support was received for this work.

Conflicts of Interest

None declared.

References

1. Zheng T. Comparative analysis of AI tools for disseminating ADA 2025 Diabetes Care Standards: implications for cardiovascular physicians. J Diabetes 2025 Mar;17(3):e70072. [doi: 10.1111/1753-0407.70072] [Medline: 40051065]

XSL•FO
RenderX

2. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
3. Smith R. Medicine's core values. BMJ 1994 Nov 12;309(6964):1247-1248. [doi: 10.1136/bmj.309.6964.1247] [Medline: 7888834]
4. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
5. Fahrner LJ, Chen E, Topol E, Rajpurkar P. The generative era of medical AI. Cell 2025 Jul 10;188(14):3648-3660. [doi: 10.1016/j.cell.2025.05.018] [Medline: 40645169]
6. Tu T, Schaekermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence. Nature 2025 Jun;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
7. Sorin V, Brin D, Barash Y, et al. Large language models and empathy: systematic review. J Med Internet Res 2024 Dec 11;26:e52597. [doi: 10.2196/52597] [Medline: 39661968]
8. Howcroft A, Bennett-Weston A, Khan A, Griffiths J, Gay S, Howick J. AI chatbots versus human healthcare professionals: a systematic review and meta-analysis of empathy in patient care. Br Med Bull 2025 Sep 22;156(1):ldaf017. [doi: 10.1093/bmb/ldaf017] [Medline: 41115171]
9. Altamimi I, Altamimi A, Alhumimidi AS, Altamimi A, Temsah MH. Artificial intelligence (AI) chatbots in medicine: a supplement, not a substitute. Cureus 2023 Jun;15(6):e40922. [doi: 10.7759/cureus.40922] [Medline: 37496532]
10. Kanne T. Love and medicine. JAMA 2024 Jan 16;331(3):195. [doi: 10.1001/jama.2023.26366] [Medline: 38153707]
11. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. In: Drazen JM, Kohane IS, Leong TY, editors. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
12. How to support the transition to AI-powered healthcare. Nat Med 2024 Mar;30(3):609-610. [doi: 10.1038/s41591-024-02897-9] [Medline: 38504014]
13. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
14. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020 Apr 7;369:m1328. [doi: 10.1136/bmj.m1328] [Medline: 32265220]

Abbreviations

AI: artificial intelligence
LLM: large language model
OSCE: Objective Structured Clinical Examination

Edited by I Steenstra; submitted 06.Apr.2026; peer-reviewed by J Harvey Turner, PF Chen; revised version received 30.May.2026; accepted 04.Jun.2026; published 17.Jun.2026.

Please cite as:
Yang YY, Liu JM
The Warmth of Medicine: The Irreplaceability of Doctors in the Era of Artificial Intelligence
J Med Internet Res 2026;28:e97412
URL: https://www.jmir.org/2026/1/e97412
doi:10.2196/97412

© Yu-ying Yang, Jian-min Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Solidarity or Segregation? ChatGPT Health and US Health Care Disparities

Andrew J Barnhart[1], PhD; Giuseppe Comerci[1], MA; Barbara Prainsack[2], Prof Dr, Dr Phil; Matthias Braun[1], Prof Dr, Dr theol

[1]Department of Social Ethics, University of Bonn, Am Hofgarten 8, Bonn, North Rhine-Westphalia, Germany

[2]Department of Political Science, University of Vienna, Vienna, Austria

Corresponding Author:
Andrew J Barnhart, PhD
Department of Social Ethics, University of Bonn, Am Hofgarten 8, Bonn, North Rhine-Westphalia, Germany

Abstract

ChatGPT Health, an artificial intelligence feature launched by OpenAI in January 2026, integrates personal medical records and consumer health data into a large language model–based chatbot. It positions itself as a digital health tool to help users navigate a fragmented US health care system marked by inaccessible insurance, medical deserts, and workforce shortages. We argue that, rather than closing these gaps, ChatGPT Health risks widening what we call the solidarity gap in health care. By appearing to address unmet needs, it may obscure and entrench the structural conditions that produce health care disparities. We identify specific risks to patient safety, such as the legitimization of dangerous self-rationing and self-medication behaviors, inaccurate triage of clinical emergencies, erosion of effective health communication, and the reinforcement of confirmation and anchoring bias. These risks are particularly acute for individuals facing structural barriers to care. To prevent ChatGPT Health and similar large language model–based health tools from deepening inequities, we propose solidarity-based policy interventions such as upstream reforms that rebuild the institutional foundations of access and health equity and downstream safeguards that embed artificial intelligence tools within accountable, equitable health care infrastructures as complements to rather than substitutes for human care.

(J Med Internet Res 2026;28:e94972) doi:10.2196/94972

KEYWORDS

large language models; digital health policy; solidarity; artificial intelligence; AI; health care inequity

Introduction

Overburdened health care services and multiple therapeutic gaps leave many health needs unmet. Many of those with unmet needs are now turning to artificial intelligence (AI) to fill in these therapeutic gaps themselves, which in turn may unburden health care services [1]. OpenAI, a US-based company developing advanced commercial and consumer AI, published a report in January 2026 that illustrated how people used ChatGPT for health care–related purposes [2]. According to the report, more than 40 million users worldwide consult ChatGPT daily with health-related questions. In the United States alone, nearly 2 million queries per week concern health insurance, whereas in rural communities, approximately 600,000 weekly messages relate to health care. Seven in 10 health care conversations occur outside clinical hours, showing a persistent demand for information and support [2]. Another study of over 500,000 deidentified Microsoft Copilot conversations in January 2026 found that nearly 1 in 5 conversations involved personal symptom assessments or discussions regarding conditions [3]. These conversations rose in frequency in the evening and nighttime hours, times when in-person health care services are less available [3].

For ChatGPT or other large language model (LLM) users who have access to health care, services and data are often fragmented: a laboratory test PDF in one portal, visitation notes in another, symptoms and health data tracked through an app, and no clear way to connect it all for a bigger picture. ChatGPT Health, launched in January 2026, is a new feature of OpenAI's chatbot that attempts to solve the first 2 challenges identified in the report: supporting users in navigating health care systems and helping mitigate access gaps [4]. OpenAI frames their Health feature as a guide through the uncertainty of the US health care system [4]. However, embedding this capability within ChatGPT is a double-edged sword. While it may be of help for some, in other contexts, the Health feature may fall short of closing the intended care gaps and could inadvertently amplify hazardous attitudes and behaviors among users.

When viewed through a governance lens, ChatGPT Health reframes what support looks like. The feature does not change the affordability of health insurance, availability of appointments, or in-person clinical workforce capacity. However, it does change how a person interprets and engages

XSL•FO
RenderX

with their health data and health-related information more broadly. ChatGPT Health is designed to unite personal medical information (eg, laboratory results, medical histories, PDFs, and notes) with consumer health data from apps such as Apple Health, Function Health, and MyFitnessPal [4]. By aggregating these sources, ChatGPT Health can generate responses grounded in a user's own data, promising an overall wellness snapshot, simplifying test results, or proposing a personalized training plan. It presents itself as a tool that reduces administrative complexity for the individual. However, at the same time, it raises questions about the responsibility of health data interpretation, whether this role will create conflicts with medical expertise and clinical institutions, and whether it shifts responsibility toward the individual [5,6].

Centralizing sensitive personal data within ChatGPT Health comes with privacy and security risks. To mitigate this, ChatGPT Health operates as a distinct module with dedicated memory, keeping medical conversations separate from general ChatGPT chats. Sensitive information does not flow into nonhealth interactions, although ChatGPT Health may use relevant context from prior conversations to improve coherence [4]. For instance, when explaining recurrent high blood pressure, it could consider an earlier discussion about job loss as a potential stressor. Given the sensitivity of medical data, ChatGPT Health adds privacy and security measures beyond those of standard ChatGPT, including purpose-built encryption and optional multifactor authentication [4]. Users retain control and can disconnect apps at any time or delete ChatGPT Health's memory.

At the time of writing, OpenAI has begun providing access to ChatGPT Health to a limited group of testers to refine the experience [4]. Partly due to strict data protection rules in Europe, the service is currently offered only outside the European Economic Area, Switzerland, and the United Kingdom. Correspondingly, despite OpenAI's stated commitment to making ChatGPT Health available to all ChatGPT users, the system appears to be optimized primarily for users in the United States. To aggregate medical data, OpenAI relies on b.well, a US-based platform for health data management [4]. As a result, the integration of medical data—intended as a core component of the ChatGPT Health feature—with consumer-facing applications currently seems to be available only in the United States.

Importantly, ChatGPT Health is not the only LLM-based tool entering the health care space. Comparable systems—such as Claude for Healthcare (Anthropic), Perplexity Health (Perplexity), and Copilot Health (Microsoft)—have recently been announced with similar functionalities [5-7]. While this paper focuses on ChatGPT Health, the dynamics and challenges discussed are likely to extend to LLM-based health tools more broadly.

ChatGPT Health is expected to have an immediate and major impact on the US health care landscape. Analyzing its deployment in this context can help anticipate broader implications and potential challenges that may emerge as similar systems are introduced in other countries, making the United States a particularly instructive case study.

Our core argument is that ChatGPT Health, along with similar LLM health modes, risks widening the solidarity gaps in US health care. While they may appear to address unmet needs, such chatbots can obscure and entrench the structural disparities that led to those unmet needs in the first place. These chatbot health interventions risk dangerous self-rationing and self-medication behaviors, inaccurate triage, erosion of effective health communication, and the reinforcement of confirmation bias. We propose that these gaps cannot be closed with digital tools. Instead, reducing these gaps requires solidarity-based policy interventions at 2 levels: upstream reforms that strengthen the institutional foundations of access and affordability and downstream safeguards governing the use of AI tools such as ChatGPT Health within the health care system.

Inaccessible Health Insurance and Health Care in the United States

OpenAI draws attention to 2 systemic issues in US health care [4]. First, health care is hard to navigate: unclear, convoluted information makes decisions (such as choosing an insurance plan) confusing and frustrating. ChatGPT already receives nearly 2 million health insurance messages each week [2]. Access and affordability are worsening as federal subsidies lapse. At the end of 2025, enhanced premium tax credits and cost-sharing reductions enacted by the American Rescue Plan and extended by the Inflation Reduction Act expired, threatening the affordability of Health Insurance Marketplace coverage for people ineligible for Medicaid and without affordable employer plans. Congress is debating whether to reinstate, extend, or make these subsidies permanent [7], although news outlets report that talks have diminished, with no agreed-upon action [8]. Without the subsidies, premiums are projected to more than double and, in some cases, quintuple, pricing out millions of people [7]. The Urban Institute estimates 4.8 million additional uninsured people in 2026 if subsidies are not extended, leaving many Americans more vulnerable in their health and care [7].

The second issue underscored by OpenAI is so-called medical deserts [2]. These are areas where residents lack adequate access to essential services such as physicians, pharmacies, hospitals, and specialists [9]. Contributing factors include work conditions, lifestyle, health care workforce migration, and sociodemographics [9]. The crisis is acute. Since 2010, more than 150 rural hospitals have closed or ended inpatient care, and over 400 face potential closure due to financial instability [10]. Less profitable services such as emergency care, obstetrics, and behavioral health are often cut first [10]. Uneven workforce distribution is also linked to worse outcomes. Many factors contribute to the uneven distribution of health care personnel across the territory: working conditions, workload, salary, and career prospects are generally better in urban areas, resulting in a lower presence of health care personnel in rural areas [7]. As a consequence, patients farther from facilities have lower survival rates, longer hospital stays, and higher rates of missed follow-ups [9]. When structural deficiencies and disparities such as these exist, it seems only natural that those who need health care would gravitate toward a more readily available alternative. And so, we now turn our attention to chatbots and digital health.

XSL•FO
RenderX

Limitations and Plausible Risks of ChatGPT Health Interventions

OpenAI presents ChatGPT Health as a practical tool to help address these shortcomings [2]. With only a smartphone and an internet connection, it is widely accessible and could temporarily mitigate some structural gaps. At the same time, the "cure" that ChatGPT Health promises also poses considerable risks. In what follows, we distinguish between (1) documented structural inequities in health care, (2) known and empirically established limitations of LLMs, and (3) anticipated risks that follow plausibly from points 1 and 2 but have not yet been fully empirically established.

The first concern relates to people with poor or no access to health care services. We anticipate that, here, ChatGPT Health could create a perverse effect. While appearing to mitigate the problem, it may make matters worse by giving false reassurance through persuasive, humanlike, empathetic responses that a chatbot is a suitable stand-in for advice from an experienced and adequately trained professional. For instance, it is well documented that adults with obesity have struggled to access glucagonlike peptide-1 receptor agonists due to limited insurance coverage, leading some to ration use to cut costs [11]. In a plausible scenario enabled by ChatGPT Health, patients who cannot afford medications promptly or who lack access to primary care or pharmacies might turn to ChatGPT Health for advice on questions such as "What happens if I miss a daily dose?" or "What symptoms could occur if I delay my medication by eight hours?" If the answers suggest low immediate risk, patients may feel reassured and choose to skip or delay doses. In this way, ChatGPT Health could provide reliable information that inadvertently legitimizes self-rationing. This practice is risky and potentially fatal for conditions that require continuous treatment, such as diabetes [11].

Another likely consequence of overreliance on ChatGPT Health is self-medication, treating symptoms without clinician supervision. A US survey found that nonprescription antibiotic use is common: 43.6% (246/564) of respondents reported prior use without a prescription [12]. Reported drivers included lack of transportation (echoing medical deserts), lack of insurance, and the cost of physician visits [12]. These documented drivers suggest that ChatGPT Health could become an additional pathway to self-medication. Although OpenAI states that ChatGPT Health is not intended to replace clinicians [2], vulnerable users without immediate access to care may use it to justify self-medication. Indirect prompts such as "What should I take for pain from a suspected fracture?" or "Are antibiotics helpful for burning with urination?" could elicit consistent, seemingly actionable guidance.

Self-medication compounds risk because queries about a drug's suitability often entail self-diagnosis. The likelihood of misdiagnosis is high, and inappropriate medication can cause serious harm. In addition, empirical evidence already indicates concrete triage failures. A study published in *Nature Medicine* found that the ChatGPT Health system's accuracy followed an inverted U-shaped pattern, meaning that it performed worst at clinical extremes, where errors matter most [13]. Most alarmingly, over half (51.6%) of true emergencies were undertriaged, with the system recommending that patients experiencing conditions such as diabetic ketoacidosis or impending respiratory failure seek care within 24 to 48 hours rather than going immediately to the emergency department [13].

Another risk, especially for people without access to formal care, is the erosion of effective health communication. ChatGPT is a sophisticated language model that generates humanlike text but lacks genuine emotional intelligence and cannot fully grasp the implications of its responses as it lacks contextual knowledge [14]. In a digital-only format, it misses nonverbal cues such as facial expressions and body language, making authentic bidirectional feedback difficult and prone to distortion [14]. More broadly, there is a documented gap between theories of effective health communication and their application in the capabilities of chatbots [15]. Effective health communication relies on social constructs; community dynamics; and identifiable, relatable figures that help individuals envision healthier alternatives, all of which are not fully present in chatbot communications [15]. Poor health communication is particularly risky for individuals in vulnerable psychological states, including those with suicidal ideation, because chatbots cannot truly understand, empathize, or fully provide emotional support [14].

ChatGPT Health may also reinforce confirmation or anchoring bias, particular risks that build from established LLM limitations. Research on LLM-mediated information seeking has established that generative AI's personalized, conversational style can strengthen existing beliefs, obscure medical consensus, and spread misinformation [13,16]. We anticipate that these dynamics are likely to be amplified within ChatGPT Health. Users often frame questions to fit their preconceptions, and the model tailors answers to those frames rather than nudging toward neutral, evidence-based perspectives. This may normalize behaviors such as self-medication or self-triaging [13,16]. Opaque customization may create echo chambers, and the model's authoritative tone encourages uncritical acceptance. Compounding the problem, ChatGPT has produced contradictory answers to the same query depending on phrasing, yet both may appear equally convincing to users with limited health literacy [16].

While the limitations and risks briefly outlined above are anticipated to be particularly acute for individuals facing structural gaps to care, they are not exclusive to underserved populations. Risks of confirmation and anchoring biases are present regardless of a person's insurance status or access to a clinician. Inaccurate triage affects any user who consults a chatbot before deciding whether to seek emergency care. However, what distinguishes structurally underserved populations is the clear absence of fundamental safety nets such as trustful relationships with clinicians and a functioning and well-funded health care infrastructure that could otherwise catch and correct errors [17,18]. Without such safety nets, individuals become more vulnerable and, thus, would (at least plausibly) reach for the closest and most easily accessible health resource available—such as ChatGPT Health. In doing so, they can receive immediate benefits such as "health advice," but they

XSL•FO
RenderX

are also taking a risk of greater impacts of harms [19]. Taken together, this leads to a widening of what we call solidarity gaps in health care, as shown in Figure 1.

Figure 1. A process flow diagram depicting how ChatGPT interventions can lead to widening solidarity gaps.

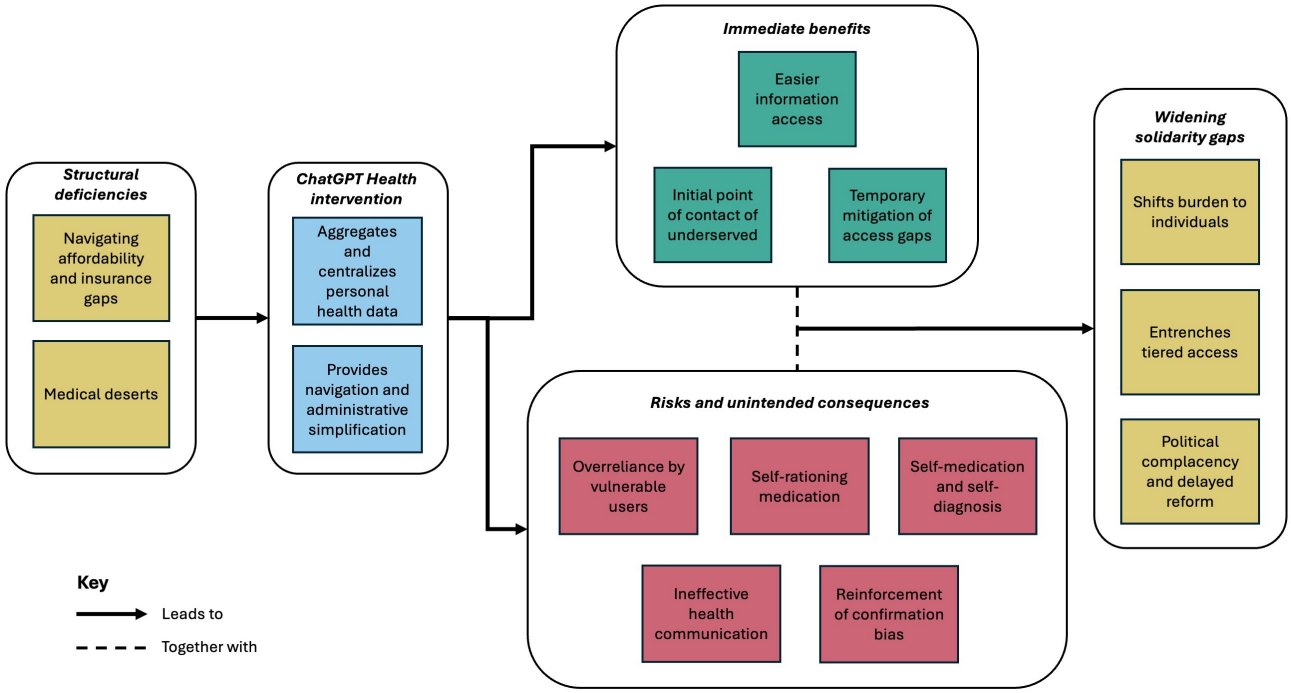

Closing, Not Exacerbating, the Solidarity Gap in Health Care

Solidarity in health care is a practice-based commitment to carry costs and share benefits with others, grounded in the recognition of relevant similarities between people—in this case, every person's vulnerability to disease or injuries [20,21]. Health care systems in Europe and in some other parts of the world [22] have historically been organized around principles of solidarity, whereas the US health care system has largely not. This has produced a situation in which collective support is unevenly realized through governance and access. Tools such as ChatGPT Health risk exacerbating this gap by creating the appearance that unmet needs in health care are being addressed without altering the institutional arrangements that produce unequal access in the first place. Where people must choose between long waiting times, long-distance travel, substantial out-of-pocket payments, or consulting a chatbot, the solidarity gap increases [20]. We define a solidarity gap here as a disparity between the ideal of collective, shared responsibility for health care costs and benefits and the reality of how unevenly collective support is distributed. Furthermore, ChatGPT Health is likely to shift even more burdens onto patients and their families (an anticipated risk that builds on documented patterns of cost shifting in digital health more broadly). It may entrench tiered access where some receive clinician-led care augmented by AI whereas others rely only on a chatbot and foster political complacency that delays investment in clinics, workforce, and coverage. This does not mean that there is no role for ChatGPT and other LLM-based tools in health care. If used and regulated within a well-funded, well-trained, and well-rested workforce within an equitable health care system, such tools can complement human intelligence and experience.

How do we get there? Avoiding a widening of the solidarity gap requires policy that puts solidarity at the center of health system design. Solutions operate both upstream and downstream of tools such as ChatGPT Health. Upstream interventions target the institutional tier of solidarity and are inseparable from justice: they counter structural arrangements that systematically produce inequitable access and outcomes [23]. Addressing medical deserts and insurance gaps requires rebuilding the institutional conditions of accessibility and affordability that make last-resort reliance on chatbots unnecessary. A solidarity-oriented upstream agenda establishes universal, equitable access as a baseline; treats health inequities as products of structural injustice; and expands democratic participation in health system governance so that benefits and burdens are shared fairly. There are proposed democratic participation mechanisms in the context of AI and health policy that could be adapted for the governance of consumer health AI. One such mechanism includes citizens' juries, a form of deliberative democracy that can obtain informed public judgment on health AI [24] or health policy generally [25].

Moreover, solidarity-based approaches would call for an interconnected distribution of resources that guarantees access to health care also for people living in rural and other underserved areas. For example, consistent rotation of mobile clinics, namely, medical units that provide medical care via vehicles, can be implemented. Services offered by mobile clinics generally focus on primary care and can span screenings, vaccinations, dental checks, and mental health [26]. A further key move could be the introduction of educational requirements in medical programs that include mandatory training in rural areas. In this way, in addition to helping address the shortage of professionals, young students could enrich their training by dealing with unfamiliar contexts [27]. From a more general

XSL•FO
RenderX

standpoint, local governments should undertake investments to establish medical infrastructures and facilities and level out discrepancies in the system, guaranteeing democratic and fair access to health care.

Downstream, ChatGPT Health should be embedded in public and clinical infrastructures as a complement to, not a substitute for, human experience, expertise, and care. Evidence indicates that LLMs are most effective in the hands of subject matter experts who can leverage time savings while identifying errors [28]. Therefore, deployment should prioritize augmentation of clinicians and care teams rather than shifting the burden to patients. Ideally, this would directly address the erosion of health communication identified above. Clinicians can read nonverbal cues, apply emotional intelligence, and contextualize responses and, thus, can more readily communicate accurate health information to a patient than a chatbot alone. However, a caveat is that clinicians must not themselves fall prey to the generative pretrained transformer bias risks or cognitive offloading.

Integration should enable warm handoffs to human services, countering the risks of self-rationing and self-medication behaviors described above. However, it is important to note that these service handoffs do not, on their own, address the material reasons for why an individual might initially undertake such courses of action. Safety and equity audits should specifically target triage failures, and unless or until an established threshold of accurate triage is met, the use of ChatGPT Health should be avoided for such purposes entirely. Transparency about model capabilities and limits helps users recognize when chatbot guidance is insufficient, which may mitigate the risks of uncritical acceptance of outputs, and performance evaluations should center on reductions in unmet needs and improved access rather than engagement metrics.

Data governance must likewise be solidarity based. Individual control alone cannot correct power asymmetries between data processors and data subjects or address group-level harms and benefits [26]. Governance should also strengthen collective oversight so that the benefits and costs of data practices are shared fairly and assess data uses by whether they create public value without imposing grave risks [29]. This form of data governance is especially salient for the corresponding LLM companies as they attempt to centralize personal health data while also shifting the interpretative responsibility onto the individual end user. It is salient insofar as it attempts to rebalance such power asymmetries, offers collective data ownership and control, and prioritizes public value. Tools such as PLUTO (Public Value Assessment Tool) serve as practical mechanisms for these value assessments by categorizing data use based on implications for public value while considering power dynamics and sustainability [30]. High–public value, low-risk uses should receive public support; low–public value, high-risk uses should be prohibited and subject to strong enforcement. Because harm can still occur and commercial actors can profit from data use, policy should provide accessible remedies and require that a portion of commercial gains be returned to the public domain [29]. Accessible remedies can take institutional forms such as that of harm mitigation bodies; such bodies can provide support to those who were plausibly harmed by data use and monitor potential harms from big data practices [31]. The return of commercial gains can take the form of data access levies on commercial owners of consumer health AI. The generated revenue could then be redirected to rebuild and maintain traditional health infrastructures and institutions or else put to better use for the people and communities that enabled the success of consumer health AI [29]. In Table 1, we summarize these possible solidarity-based policy interventions and what they target.

Table . Upstream and downstream solidarity-based policy interventions and what they target and/or prevent.

Upstream or downstream	Solidarity gap addressed	Solidarity-based policy intervention	What it targets or is meant to prevent
Upstream	Institutional access	Rebuild institutional conditions of accessibility and affordability	Prevent chatbot use from becoming a de facto substitute for real coverage, appointments, pharmacies, and in-person capacity
Upstream	Coverage equity	Establish universal, equitable access as a baseline	Prevent tiered access where some receive clinician-led care vs those who rely on chatbots
Upstream	Democratic participation	Expand democratic participation in health system governance	Prevent political complacency and delays in investments in clinics, workforces, and coverage
Downstream	Accountability	Only embed chatbots within equitable systems with accountable infrastructures	Prevent the shifting of responsibility for interpretation and care onto (vulnerable) individuals
Downstream	Measurement	Evaluate performance through reductions in unmet needs and improved access	Prevent incentives that prioritize chatbot use over institutional access
Downstream	Data governance	Data solidarity governance: strengthen collective oversight and assess data uses based on public value vs grave risks	Correct power asymmetries between data processors and data subjects; address group-level harms and benefits

XSL•FO
RenderX

Conclusions

The appearance of a health-dedicated function in ChatGPT and similar LLMs deserves careful scrutiny. The promise of filling health care gaps risks colliding with health care systems such as those in the United States and potentially widening those gaps. The potential overreliance on ChatGPT Health by individuals experiencing unequal access to health care services calls for solutions rooted in solidarity. However, a solidaristic health system does not emerge from technology alone, but technology can help only if institutions do their part. Policymakers have an obligation to rebuild and strengthen the foundations of access and affordability and embed AI tools only within accountable, equitable infrastructures. If done correctly, ChatGPT Health can complement rather than replace human care and improve navigation without widening gaps. The measure of success cannot be based on LLM engagement. Instead, the measure should be fewer unmet needs, safer care, and easier access to basic health care infrastructure.

What is unique about the concept of a solidarity gap is that it shows how digital health technologies can mask and entrench structural disparities by appearing to meet needs without altering the conditions that produce them. The concept, in a way, forces a shift in the analysis from a tool in isolation to its interaction with the institutional and social contexts in which it is deployed. Moreover, there are clear upstream and downstream policy reforms that can be put in place and that are based on a principle of solidarity. If convenience and the ideal of short-term cost savings replace solidarity, AI, no matter the form, will become a veneer over deepening inequities. Solidarity is the governing principle that will become the bridge to better care.

Acknowledgments

The authors declare that artificial intelligence was not used in the preparation of this article.

Funding

This work was supported by the Open Access Publication Fund of the University of Bonn and was cofunded by the European Union (European Research Council; SIMTWIN project; 101076822). However, the views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Data Availability

Data sharing is not applicable to this paper as no datasets were generated or analyzed during this study.

Authors' Contributions

Conceptualization: AJB, GC, BP, MB
Funding acquisition: MB
Investigation: AJB, GC, BP, MB
Methodology: AJB, GC, BP, MB
Supervision: MB
Writing—original draft: AJB, GC, BP, MB
Writing—review and editing: AJB, GC, BP, MB

Conflicts of Interest

None declared.

References

1. Opel N, Breakspear M. Transforming mental health research and care through artificial intelligence. Science 2026 Jan 15;391(6782):249-258. [doi: 10.1126/science.adz9193] [Medline: 41538442]
2. AI as a healthcare ally: how Americans are navigating the system with ChatGPT. : OpenAI; 2026 URL: https://cdn.openai.com/pdf/2cb29276-68cd-4ec6-a5f4-c01c5e7a36e9/OpenAI-AI-as-a-Healthcare-Ally-Jan-2026.pdf [accessed 2026-01-16]
3. Costa-Gomes B, Tolmachev P, Taysom E, et al. Public use of a generalist LLM chatbot for health queries. Nat Health 2026 Apr 16. [doi: 10.1038/s44360-026-00117-x]
4. Introducing ChatGPT health. OpenAI. 2026. URL: https://openai.com/index/introducing-chatgpt-health/ [accessed 2026-01-13]
5. Bleher H, Braun M. Diffused responsibility: attributions of responsibility in the use of AI-driven clinical decision support systems. AI Ethics 2022;2(4):747-761. [doi: 10.1007/s43681-022-00135-x] [Medline: 35098247]
6. Kiener M. AI and responsibility: no gap, but abundance. J Appl Philos 2025 Feb;42(1):357-374. [doi: 10.1111/japp.12765]

XSL•FO
RenderX

7. Buettgens M, Simpson M, Levitis J, Hernandez-Lepe F, Banthin J. 4.8 million people will lose coverage in 2026 if enhanced premium tax credits expire. : Urban Institute; 2025 URL: https://www.urban.org/research/publication/48-million-people-will-lose-coverage-2026-if-enhanced-premium-tax-credits [accessed 2026-06-07]
8. Choi J. Senate talks to revive ACA tax credits appear to be fizzling out. The Hill. 2026. URL: https://thehill.com/policy/healthcare/5723559-moreno-collins-proposal-fizzles/ [accessed 2026-03-06]
9. Flinterman LE, González-González AI, Seils L, et al. Characteristics of medical deserts and approaches to mitigate their health workforce issues: a scoping review of empirical studies in Western countries. Int J Health Policy Manag 2023;12:7454. [doi: 10.34172/ijhpm.2023.7454] [Medline: 38618823]
10. Howren MB, Hansen JR. The One Big Beautiful Bill Act-implications for rural health care. JAMA 2025 Oct 7;334(13):1143-1144. [doi: 10.1001/jama.2025.13518] [Medline: 40900579]
11. Chen AS, Borden CG, Canavan ME, Ross JS, Oladele CR, Lipska KJ. Cost-related prescription drug rationing by adults with obesity. JAMA Netw Open 2024 Nov 4;7(11):e2433000. [doi: 10.1001/jamanetworkopen.2024.33000] [Medline: 39499520]
12. Grigoryan L, Paasche-Orlow MK, Alquicira O, et al. Antibiotic use without a prescription: a multisite survey of patient, health system, and encounter characteristics. Clin Infect Dis 2023 Aug 22;77(4):510-517. [doi: 10.1093/cid/ciad241] [Medline: 37094252]
13. Ramaswamy A, Tyagi A, Hugo H, et al. ChatGPT Health performance in a structured test of triage recommendations. Nat Med 2026 May;32(5):1671-1675. [doi: 10.1038/s41591-026-04297-7] [Medline: 41731097]
14. Barnhart AJ, Barnhart JE, Dierickx K. Why ChatGPT means communication ethics problems for bioethics. Am J Bioeth 2023 Oct;23(10):80-82. [doi: 10.1080/15265161.2023.2250278] [Medline: 37812099]
15. Weingott S, Parkinson J. The application of artificial intelligence in health communication development: a scoping review. Health Mark Q 2025;42(1):67-109. [doi: 10.1080/07359683.2024.2422206] [Medline: 39556410]
16. Lopez-Lopez E, Abels CM, Holford D, Herzog SM, Lewandowsky S. Generative artificial intelligence-mediated confirmation bias in health information seeking. Ann N Y Acad Sci 2025 Aug;1550(1):23-36. [doi: 10.1111/nyas.15413] [Medline: 40716036]
17. Bourgois P, Holmes SM, Sue K, Quesada J. Structural vulnerability: operationalizing the concept to address health disparities in clinical care. Acad Med 2017 Mar;92(3):299-307. [doi: 10.1097/ACM.0000000000001294] [Medline: 27415443]
18. Horan TA, Botts NE, Burkhard RJ. A multidimensional view of personal health systems for underserved populations. J Med Internet Res 2010 Aug 4;12(3):e32. [doi: 10.2196/jmir.1355] [Medline: 20685644]
19. Hille EM. Biases as risk factors to patient security? A taxonomy of algorithmic harm in health. Am J Bioeth 2026 Feb;26(2):114-116. [doi: 10.1080/15265161.2025.2608618] [Medline: 41678704]
20. Braun M. How predictive medicine leads to solidarity gaps in health. NPJ Digit Med 2025 Feb 18;8(1):111. [doi: 10.1038/s41746-025-01497-2] [Medline: 39966662]
21. Prainsack B. The “we” in the “me”: solidarity and health care in the era of personalized medicine. Sci Technol Hum Values 2018;43(1):21-44. [doi: 10.1177/0162243917736139]
22. Guy M. Solidarity as a political determinant of health: insights from EU competition policy. J Health Polit Policy Law 2024 Oct 1;49(5):783-803. [doi: 10.1215/03616878-11257016] [Medline: 38567773]
23. Gould CC. Solidarity and the problem of structural injustice in healthcare. Bioethics 2018 Nov;32(9):541-552. [doi: 10.1111/bioe.12474] [Medline: 30044895]
24. van der Veer SN, Riste L, Cheraghi-Sohi S, et al. Trading off accuracy and explainability in AI decision-making: findings from 2 citizens’ juries. J Am Med Inform Assoc 2021 Sep 18;28(10):2128-2138. [doi: 10.1093/jamia/ocab127] [Medline: 34333646]
25. Street J, Duszynski K, Krawczyk S, Braunack-Mayer A. The use of citizens’ juries in health policy decision-making: a systematic review. Soc Sci Med 2014 May;109:1-9. [doi: 10.1016/j.socscimed.2014.03.005] [Medline: 24657639]
26. Leibowitz A, Livaditis L, Daftary G, Pelton-Cairns L, Regis C, Taveras E. Using mobile clinics to deliver care to difficult-to-reach populations: a COVID-19 practice we should keep. Prev Med Rep 2021 Dec;24:101551. [doi: 10.1016/j.pmedr.2021.101551] [Medline: 34522575]
27. Singleton R. The benefits of rural training: producing the expert generalists of the future. Future Healthc J 2020 Feb;7(1):50-52. [doi: 10.7861/fhj.2019-0071] [Medline: 32104766]
28. Noy S, Zhang W. Experimental evidence on the productivity effects of generative artificial intelligence. Science 2023 Jul 14;381(6654):187-192. [doi: 10.1126/science.adh2586] [Medline: 37440646]
29. Prainsack B, El-Sayed S, Forgó N, Szoszkiewicz Ł, Baumer P. Data solidarity: a blueprint for governing health futures. Lancet Digit Health 2022 Nov;4(11):e773-e774. [doi: 10.1016/S2589-7500(22)00189-3] [Medline: 36307191]
30. El-Sayed S, Kickbusch I, Prainsack B. Data solidarity: operationalising public value through a digital tool. Glob Public Health 2025 Dec;20(1):2450403. [doi: 10.1080/17441692.2025.2450403] [Medline: 39789994]
31. McMahon A, Buyx A, Prainsack B. Big data governance needs more collective responsibility: the role of harm mitigation in the governance of data use in medicine and beyond. Med Law Rev 2020 Feb 1;28(1):155-182. [doi: 10.1093/medlaw/fwz016] [Medline: 31377815]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
LLM: large language model
PLUTO: Public Value Assessment Tool

Edited by Z Su; submitted 09.Mar.2026; peer-reviewed by C Chow, J Shapiro; revised version received 28.Apr.2026; accepted 29.Apr.2026; published 22.Jun.2026.

Please cite as:
Barnhart AJ, Comerci G, Prainsack B, Braun M
Solidarity or Segregation? ChatGPT Health and US Health Care Disparities
J Med Internet Res 2026;28:e94972
URL: https://www.jmir.org/2026/1/e94972
doi:10.2196/94972

© Andrew J Barnhart, Giuseppe Comerci, Barbara Prainsack, Matthias Braun. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Enhancing Physician Resilience to Generative AI: Multilevel Framework for Shared Authority, Verification, and Skill Preservation

Hongxia Pan[1], MMed; Jialin Liu[2,3], MD; Siru Liu[4], PhD

[1]Rehabilitation Medicine Center and Institute of Rehabilitation Medicine, West China Hospital of Sichuan University, Chengdu, Sichuan, China
[2]Department of Medical Informatics, West China Hospital of Sichuan University, Guoxue Xiang Street, Chengdu, Sichuan, China
[3]Department of Otolaryngology-Head and Neck Surgery, West China Hospital of Sichuan University, Chengdu, Sichuan, China
[4]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

Corresponding Author:
Jialin Liu, MD
Department of Medical Informatics, West China Hospital of Sichuan University, Guoxue Xiang Street, Chengdu, Sichuan, China

Abstract

As generative artificial intelligence (AI), particularly large language model–based tools, is increasingly integrated into diagnosis, triage, decision support, and treatment planning, it offers potential gains in efficiency and information access. However, real-world deployment also introduces important risks, including hallucinations, miscalibrated confidence, automation bias, and increased verification burden on physicians. This burden may divert attention from independent clinical reasoning, contribute to deskilling, and increase vulnerability when models fail silently or perform poorly in unfamiliar clinical contexts. Existing AI governance frameworks emphasize data quality, transparency, accountability, and ethical deployment, but pay less attention to physician-facing resilience, defined in this paper as the capacity to sustain independent and safe clinical judgment when collaborating with generative AI. In this viewpoint, we propose a multilevel governance framework organized around 3 coordinated domains: cognitive workload shaping, clinical authority governance and allocation, and organizational safety governance and accountability. Together, these domains aim to reduce verification burden, preserve physician decisional authority, and align institutional oversight with safe and context-sensitive AI use. The framework includes mechanisms such as risk-sensitive verification triggers, bounded delegation, structured interprofessional review, and organizational monitoring to support safe clinical integration while minimizing avoidable workflow disruption. At the same time, implementation may be limited by workflow friction, alert fatigue, variable institutional resources, and the need for ongoing monitoring and recalibration to ensure that safeguards remain clinically useful rather than burdensome. Accordingly, this paper outlines a structured governance framework to guide safer integration of generative AI into clinical care and inform future evaluation across specialties, workflows, and institutional settings.

(J Med Internet Res 2026;28:e88058) doi:10.2196/88058

KEYWORDS

cognitive load; authority; accountability; artificial intelligence; AI; generative artificial intelligence; generative AI; clinical safety; physician

Introduction

Large language model–based generative artificial intelligence (AI) is rapidly extending from documentation support to core clinical activities, including diagnosis, clinical decision support (CDS), triage, treatment recommendation, and preventive care planning [1-6]. Early evaluations suggest potential gains in timeliness and efficiency; however, deployment in real-world clinical settings introduces sociotechnical risks with direct implications for safety and professional competence [7,8]. Generative AI can produce fluent but incorrect outputs (hallucinations); express confidence that is not supported by the underlying evidence; and perform inconsistently under distribution shift, prompt artifacts, or algorithmic bias, potentially leading to clinically unjustified differences across demographic or socially marginalized groups [9]. In this paper, “distribution shift” refers to reduced performance when the AI encounters patients, settings, or data patterns that differ from its training data, and “prompt artifacts” refers to misleading changes in output caused by superficial wording or formatting differences rather than clinically meaningful distinctions. These risks impose a substantial verification burden: physicians must divert time and attention from intrinsic clinical reasoning to source checking, uncertainty appraisal, and contextual fit assessment, thereby increasing extraneous cognitive load [10-12]. In addition, assigning complex integrative tasks to AI may contribute to deskilling as independent diagnostic synthesis and risk communication decline, leaving physicians more vulnerable during model failures or out-of-distribution conditions [13,14]. Coupled with automation bias and overtrust,

XSL•FO
RenderX

these pressures create structural vulnerabilities that cannot be mitigated solely through traditional safety training or single-discipline governance [15].

We use "resilience" in the sense established in resilience engineering: the capacity of a system or its actors to sustain safe and effective functioning under variable, degraded, or unanticipated conditions [16,17]. This usage differs from narrower psychological accounts that emphasize individual emotional coping or recovery from adversity. Applied to the physician-AI interface, physician resilience refers to the sustained capacity for independent clinical judgment when collaborating with systems whose outputs may be uncertain, miscalibrated, or context inappropriate. Although AI governance research has extensively examined data quality, transparency, and ethical principles, less attention has been paid to the combined cognitive, professional, and organizational conditions needed to preserve this resilience in physician-AI collaboration.

Sustaining this resilience depends on preserving physician autonomy. Autonomy in this case does not mean isolated clinical discretion or resistance to AI assistance. Rather, it refers to the physician's professionally accountable capacity to interrogate, contextualize, accept, modify, or override AI-generated suggestions based on clinical evidence, patient values, uncertainty, and situational constraints [18,19]. The relationship between resilience and autonomy is reciprocal: resilience requires preserved autonomy so that physicians can adapt safely when AI outputs are incomplete, biased, or miscalibrated; autonomy, in turn, requires resilient cognitive, professional, and organizational conditions that prevent clinical judgment from being displaced by automation bias, excessive verification burden, or institutional pressure to defer to algorithmic outputs [7,10,11,15]. Therefore, the governance challenge is not only whether generative AI outputs are accurate but also whether clinical environments preserve physicians' capacity to remain accountable decision-makers within a model of shared human-AI authority.

To address this gap, we propose a multilevel governance framework organized around 3 coordinated domains (Figure 1): cognitive workload shaping, clinical authority governance and allocation, and organizational safety governance and accountability. This resilience-centered framing explains why governance must address not only model performance but also the cognitive, professional, and institutional conditions under which physicians can sustain safe judgment and accountable authority in AI-supported care. Together, these domains aim to bound and rebalance verification work, preserve physician decisional authority while enabling safe delegation, and strengthen institutional accountability for shared risk. Figure 1 depicts the relationships among these domains and their associated governance mechanisms. To clarify the 2 implementation levels of the framework, "institutional recommendations" refers to governance design, resource allocation, workflow integration, monitoring, and oversight structures. In contrast, "physician-facing recommendations" refers to verification practices, uncertainty appraisal, patient communication, sign-off, and preservation of clinical judgment. These levels cut across all 3 domains rather than mapping onto separate domains because safe physician action often depends on institutional design, resources, and oversight.

Figure 1. Three-pillar governance model for enhancing physician resilience to generative artificial intelligence (AI). ASoP: automation scope of practice; PSI: protected safety investment.

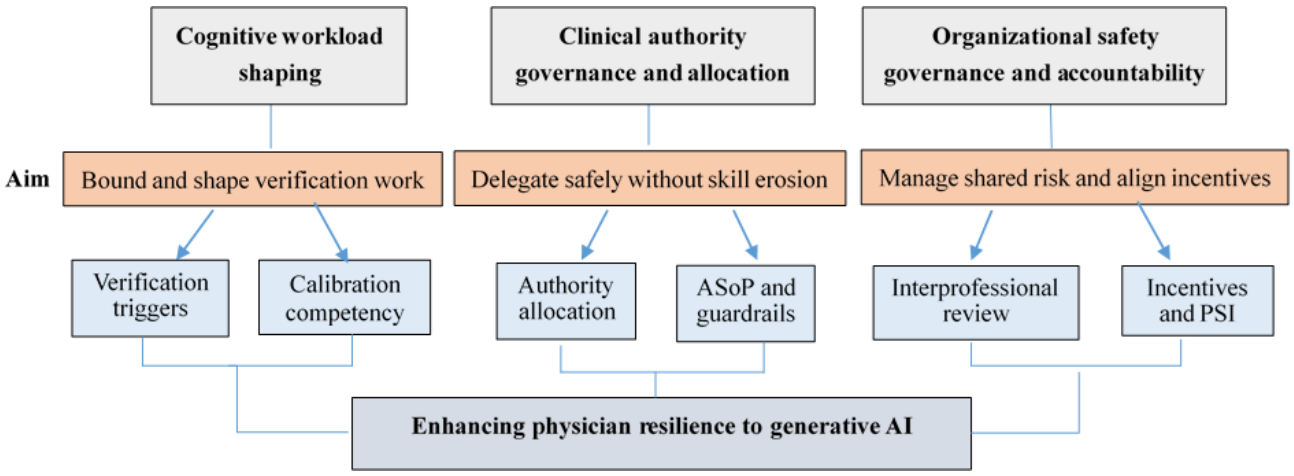

This viewpoint presents a normative, concept-driven framework developed through interdisciplinary, literature-informed synthesis rather than formal empirical data collection or expert consensus procedures. We drew on 5 research areas: resilience engineering, clinical reasoning and metacognition, CDS stewardship, AI governance and regulation, and emerging evidence on automation bias and AI-induced deskilling. From these, we identified recurrent risks at the physician-AI interface and organized corresponding governance responses across the 3 domains. The framework is intended to guide implementation and evaluation of physician-facing safeguards for generative AI in clinical care and extend rather than replicate governance approaches developed for traditional rule-based CDS given the distinct properties of generative AI outputs.

Cognitive Workload Shaping

Overview

In a busy outpatient clinic, a physician uses an electronic health record–integrated generative AI tool to summarize a patient's history and suggest possible explanations for subacute dyspnea. The summary is fluent but omits a recent medication change

XSL•FO
RenderX

and presents several low-probability diagnoses as a coherent leading narrative. The resulting risks are 2-fold: factual incompleteness or distortion and cognitive bias induced by a plausible but misleading narrative frame. The physician may use the AI-framed explanation as an anchor, divert attention to verifying low-probability claims, or overlook clinically important information omitted from the summary. Therefore, the central governance challenge is to reduce verification burden while preserving independent clinical reasoning when AI outputs are coherent but epistemically unreliable.

Verification Triggers and Proportionate Safeguards

Institutions should classify verification triggers using a structured matrix based on 4 dimensions: (1) clinical harm severity if the AI-supported recommendation is wrong; (2) irreversibility and time criticality of the downstream action; (3) uncertainty or conflict intensity, including model uncertainty, weak provenance, or disagreement between AI output and clinician assessment; and (4) equity sensitivity, referring to decisions in which algorithmic bias may plausibly distort recommendations [20,21]. Because algorithmic bias is difficult for individual physicians to detect case by case, equity sensitivity should be operationalized primarily as a system-level trigger supplemented by clinician judgment [20,22]. Institutions should predefine equity-sensitive contexts using local validation results, known subgroup performance gaps, and high-impact domains that affect access, triage, diagnosis, treatment intensity, or follow-up [23,24]. The trigger may activate automatically when an AI-supported recommendation involves these contexts, when patient characteristics fall outside the model's validated population, or when local audits identify subgroup-specific performance concerns [20,25]. Demographic variables used for this purpose should support bias surveillance rather than justify differential clinical treatment. Clinicians should also be able to invoke the trigger when they identify contextual vulnerability or discordance between the AI output and the patient's clinical or social context.

Lower-tier triggers involve reversible, low-uncertainty tasks and require brief verification, whereas higher-tier triggers should prompt independent reasoning and explicit attestation. Escalation or second review should be reserved for cases involving high severity, irreversibility, time criticality, substantial uncertainty, conflict with clinician judgment, or equity-sensitive contexts. Thresholds may vary by service line, but classification logic should remain explicit, auditable, and periodically recalibrated. Attestation interfaces are designed to discourage verbatim or formulaic rationales, with free-text fields structured to elicit clinical specificity rather than generic confirmation. Mandatory pauses and other high-interruption mechanisms are reserved for high-risk decisions, whereas lower-risk situations rely on passive or minimally disruptive cues. Routine safety governance includes monitoring of override rates, bypass patterns, and clinician-reported burden to ensure that safeguards remain salient rather than becoming background noise [25].

Uncertainty and Trust Calibration Competency

Institutions should ensure that physician training shifts from basic AI literacy to verification-oriented competencies and calibrated trust, defined as the ability to adjust reliance on AI according to clinical context, risk level, and uncertainty signals [26,27]. Core competencies include interpreting model uncertainty and calibration indicators as reliability signals rather than proxies for ground truth; calibration indicators refer to signals about whether the model's expressed confidence or uncertainty corresponds to observed correctness in the relevant clinical context [28]. Protocolized plausibility checks against established physiology, guideline constraints, and patient-specific factors are needed, with documented attestation when a check is triggered. Training should also cover prompt hygiene, defined as structuring prompts to reduce ambiguity and prevent misleading context from shaping the model's response, and contextual integrity, defined as maintaining appropriate boundaries around what clinical information should or should not be included in an AI interaction [29]. These practices require institutional templates and interface guardrails rather than individual technical improvisation. Physicians should also appraise the provenance and recency of model-cited evidence to determine whether AI-synthesized rationales should be accepted, discounted, or escalated [30]. Together, these competencies support calibrated trust and help contain verification burden in AI-supported care.

Clinical Authority Governance and Allocation

Overview

In an oncology service, a generative AI system drafts a chemotherapy dosing recommendation for a patient with declining renal function and evolving laboratory abnormalities. The output is plausible but may not reflect local protocols, recent dose adjustments, or temporal trends in laboratory values. Therefore, the central governance challenge is not only whether the recommendation is clinically correct but also who is authorized to decide, verify, and sign off on a high-risk intervention.

Decisional Responsibility and Role Allocation

In AI-supported clinical care, safe practice depends on clearly distinguishing the AI system's informational support role from the physician's decisional responsibility for clinical interpretation, patient communication, and sign-off [22,24]. AI outputs do not carry independent clinical authority and cannot, by themselves, commit a patient to action. Physicians retain responsibility for interpreting outputs; accepting, modifying, or rejecting recommendations; completing required disclosure and consent processes; and signing off on clinically consequential decisions [24,31]. Accountability should be distributed in proportion to operational control: physicians remain responsible for case-specific clinical judgment; institutions remain responsible for local validation, workflow integration, and performance monitoring; and developers remain responsible for model design, governance update, and management of known limitations [22,32]. This allocation preserves physician judgment while avoiding the inappropriate concentration of liability on physicians for system-level AI failures. In value-sensitive or preference-dependent decisions, decisional authority ultimately rests with the patient, whereas

XSL•FO
RenderX

physicians remain responsible for transparent communication and informed, preference-sensitive decision-making.

At the institutional level, a designated AI governance committee or equivalent multidisciplinary body should approve AI deployment within specific workflows, define local validation requirements, and oversee monitoring and escalation pathways. Such a body should include representation from clinical leadership, informatics, quality and safety, legal or compliance, and relevant specialty expertise. Patient-facing transparency, including disclosure of AI involvement and relevant uncertainty, is a core component of safe human oversight and should be tailored to risk level and use case [22,33]. To make disclosure feasible in time-limited encounters, institutions can provide standardized plain-language phrasing, electronic health record–integrated documentation prompts, and risk-stratified consent workflows [31]. For low-risk uses such as AI-assisted summarization or documentation support, brief disclosure that AI was used and reviewed by the clinician may be sufficient. Clinically consequential recommendations, including AI-supported diagnostic interpretation, medication adjustment, triage prioritization, or treatment planning, warrant communication of 4 core points: (1) AI supported the analysis, (2) the physician reviewed and contextualized the output, (3) uncertainty or alternative interpretations remain, and (4) the patient may ask questions or decline AI involvement where legally and ethically applicable. For example, phrasing such as "An AI-supported tool helped check part of the information, but I reviewed it in light of your clinical situation" can be integrated into consent, shared decision-making, or postvisit documentation workflows. When AI outputs conflict with clinical guidelines, documented patient goals, or physician judgment, a verification trigger should activate predefined escalation pathways, with the resulting review, attestation, and rationale recorded in the audit log [20,26,30].

Automation Scope of Practice and Safe Delegation

To operationalize safe delegation, institutions should apply automation scope of practice (ASoP) through a stepwise decision process that links task classification, verification triggers, and authority allocation [23,24,32]. First, the workflow is divided into discrete task units, such as summarization, differential diagnosis generation, medication review, triage prioritization, or treatment recommendation. Second, each task is assigned an initial ASoP tier based on clinical risk, verifiability, time criticality, and dependence on contextual physician judgment. Third, verification triggers, including harm severity, irreversibility and time criticality, uncertainty or conflict intensity, and equity sensitivity, are applied to determine whether the task requires a higher level of review. Fourth, authority allocation rules determine who may act on the output, who must supervise or attest, and when escalation or second review is required.

Institutions can apply a 4-tier ASoP rubric (Table 1). Tier 0 includes nondelegable, physician-led tasks involving high-stakes diagnostic or therapeutic judgment, weak verifiability, time-critical consequences, or potentially irreversible harm. Assistive tasks such as draft documentation, summarization, or structured information retrieval fall under tier 1, where AI output is low risk, readily verifiable, and not directly action enabling. Tier 2 applies to supervised analytic support, including preliminary differentials, medication conflict flags, or care plan drafting, for which physician review, contextualization, and explicit acceptance are required before clinical action. At tier 3, automation may proceed within predefined institutional rules, audit trails, and escalation criteria for conditionally delegated workflow actions such as low-risk protocol-based triage support, follow-up scheduling, or routine monitoring reminders [23,32].

XSL•FO
RenderX

Table . Illustrative automation scope of practice (ASoP) tier assignments for representative clinical tasks[a].

Example task	ASoP tier	Permitted AI[b] role	Verification and authority requirements
Drafting or formatting a clinical note	Tier 1: assistive	Drafting or structuring text	Clinician reviews for accuracy and omissions before signing; escalates if hallucinated or incorrect content is detected.
Summarizing prior laboratory results, imaging, or medication history	Tier 1: assistive	Extracting and summarizing defined source information	Clinician verifies source accuracy, temporal relevance, and omissions; escalates if data are incomplete, outdated, or outside validated sources.
Generating preliminary differential diagnoses	Tier 2: supervised analytic	Suggesting candidates but not determining diagnosis	Independent clinician reasoning and explicit acceptance, modification, or rejection; escalate for high uncertainty, conflict, or high-risk presentations.
Flagging medication interactions or renal dose concerns	Tier 2: supervised analytic	Flagging concerns or drafting options	Clinician, with pharmacist input where relevant, verifies history, laboratory trends, and contraindications; escalates high-harm or protocol-conflicting cases for second review.
Protocol-based follow-up scheduling or monitoring reminders	Tier 3: conditionally delegated	Recommending or initiating predefined workflow steps	Allowed only within institutional rules, audit trails, and retrospective monitoring; escalate for status change, missing data, or out-of-protocol cases.
Low-risk protocol-based triage support such as stable outpatient referrals	Tier 3: conditionally delegated	Supporting prioritization within bounded protocols	Bounded by predefined rules; higher-risk presentations escalate to clinician confirmation; monitor for red flags or atypical features.
Final diagnosis, chemotherapy dosing, treatment selection, or withholding urgent care	Tier 0: nondelegable	Providing information or options only	Physician-led judgment with explicit attestation; second review for high-stakes or contested cases; no autonomous execution pathway permitted.

[a]The categories are examples for local adaptation rather than fixed universal rules.

[b]AI: artificial intelligence.

Tier assignment should be dynamic rather than fixed. Tasks move toward more restrictive tiers when the patient is clinically unstable, input data are incomplete or unreliable, the output falls outside the model's validated use context, the recommendation affects vulnerable or underserved groups, or the consequences of error are difficult to reverse. These assessments determine what may be automated, what must remain under physician control, and what safeguards are required [22,23].

Organizational Safety Governance and Accountability

Overview

In a large health system, a generative AI tool drafts responses to patient portal messages about new symptoms, possible medication adverse effects, and requests for treatment advice. Some drafts are helpful, whereas others are overly reassuring, fail to identify escalation cues, or conflict with local triage protocols. Although individual physicians may detect some errors during review, safe use depends on organizational monitoring thresholds, audit mechanisms, update oversight, and accountability structures. Therefore, the central governance challenge is not only error detection at the point of care but also ensuring that institutional oversight sustains safe physician judgment beyond individual vigilance alone.

Targeted Interprofessional Safety Review

For prespecified high-risk scenarios such as chemotherapy dose adjustment, physicians should retain final decisional authority while convening focused interprofessional review with relevant colleagues, including pharmacists, nurses, and specialists [24,34,35]. The purpose is not to duplicate individual verification but to support team-based assessment of feasibility, safety, and cross-disciplinary workflow fit [34-36]. Activation should follow verification triggers with team-level implications, and review intensity and timing should be adapted to clinical urgency [24,32]. In time-critical situations, review may involve rapid, role-explicit confirmation. When immediate action is

XSL•FO
RenderX

required before review can occur, documented retrospective review should follow. Workflow interfaces should capture role-specific input and physician attestation in the audit log to support traceability and accountability [32,36].

Physician-Facing Incentives and Protected Safety Investment

Organizational policy should align incentives with safety rather than throughput alone [23,37,38]. Institutions should establish protected safety investment to support nonautomatable functions that strengthen physician decision-making, including time for complex case consultation, targeted interprofessional review training, and selected skill retention activities for high-risk scenarios [38]. AI-related gains can be documented through auditable operational measures such as clinician time returned, reduced after-hours documentation, or avoided downstream workflow failures rather than assumed from adoption alone. Where direct monetization is constrained by billing context, these measures can inform local business cases, and safety investment may need justification as prospective quality infrastructure. Performance monitoring should emphasize auditable safety process measures such as verification trigger adherence, appropriate override or decline rates, attested sign-off, and time to decision in triggered cases [37].

Skill Retention and Learning-Oriented Error Review

Skill retention activities should be risk stratified and targeted to high-risk or high-delegation workflows. Where feasible, these activities can be integrated into protected simulation time, case-based continuing medical education, credentialing-linked refreshers, or quality improvement and morbidity and mortality review processes rather than assuming that clinicians have unscheduled downtime [13,38]. Institutions should maintain a local learning-oriented, nonpunitive error taxonomy. This taxonomy would classify AI-related safety events by source and mechanism, including interface design problems, human factor contributors, data or workflow mismatches, and model output failures. The resulting signals can guide focused review, targeted training, and interface improvement rather than individual blame [23,39]. To clarify how the 3 framework domains translate into institution-level and physician-facing responsibilities, Table 2 summarizes the major recommendations, primary responsible parties, and practical implications for implementation.

XSL•FO
RenderX

Table . Framework recommendations by domain and responsible party.

Framework domain and major recommendation	Primary responsible party	Practical implication
Cognitive workload shaping		
Use risk-sensitive verification triggers and proportionate safeguards	Institution led; physician applied	Institutions define thresholds, interface cues, escalation rules, and monitoring processes; physicians apply triggered checks, document rationales, and escalate when needed.
Develop uncertainty and trust calibration competency	Shared institutional and physician responsibility	Institutions provide training, templates, and interface guardrails; physicians interpret uncertainty, assess provenance, perform plausibility checks, and calibrate reliance on AI[a] outputs.
Clinical authority governance and allocation		
Clarify decisional responsibility and role allocation	Shared institutional and physician responsibility	Institutions define disclosure, consent, accountability, and escalation pathways; physicians interpret AI outputs, communicate AI-supported reasoning and uncertainty, and retain sign-off responsibility for consequential decisions.
Define an automation scope of practice for safe delegation	Institution led; physician applied	Institutions classify tasks by risk, verifiability, competence, and supervision level; physicians use AI only within approved delegation boundaries and apply required verification.
Organizational safety governance and accountability		
Activate targeted interprofessional safety review for high-risk scenarios	Shared institutional and physician responsibility	Institutions define team review triggers and documentation requirements; physicians retain final authority while seeking pharmacist, nursing, specialty, or other input when triggered.
Align incentives with protected safety investment and auditable performance monitoring	Institution led	Institutions protect time and resources for complex case consultation, targeted training, and safety monitoring while avoiding throughput-only measures of AI value.
Support skill retention and learning-oriented error review	Shared institutional and physician responsibility	Institutions use risk-stratified simulation, continuing education, credentialing-linked refreshers, quality improvement processes, and a nonpunitive error taxonomy; physicians participate in targeted reassessment and remediation when needed.

[a]AI: artificial intelligence.

Discussion

Relation to Existing Knowledge

The proposed framework is informed by and extends several established research areas. The verification trigger mechanism draws on cognitive load theory and meta-cognitive training in medical education [12,40]. The authority allocation construct builds on resilience and error management frameworks, whereas the organizational safety pillar is informed by World Health Organization guidance, the European Union AI Act, and emerging institutional initiatives such as the Trustworthy and Responsible AI Network [26,32,41,42]. These frameworks establish essential system-level expectations for transparency, accountability, human oversight, and institutional readiness but do not specify the cognitive and workflow conditions under which physicians can sustain independent judgment while collaborating with generative AI [26,32]. Physician autonomy literature further informs the authority governance component by clarifying why physicians must retain accountable authority to interpret, modify, or override AI-supported recommendations [18,19]. Emerging evidence on automation bias and AI-related deskilling informs the verification, delegation, and skill retention components by identifying mechanisms that may erode independent clinical judgment [7,13-15]. This framework complements these foundations by focusing more explicitly on physician-facing safeguards, including verification triggers, bounded delegation, calibrated interpretation of uncertainty, and skill retention monitoring.

Distinct Governance Requirements for Generative AI

This framework extends rather than replicates governance approaches used for traditional rule-based CDS. Conventional alert-based CDS is typically narrower in scope and anchored to prespecified logic; accordingly, governance has often focused on alert thresholds, override appropriateness, and workflow burden [24,25]. In contrast, generative AI systems produce context-sensitive natural language outputs that may appear coherent and persuasive while still being false, inaccurate, biased, or incomplete [28]. Their behavior may also vary with prompt design, the quantity and order of information provided, and model updates [29]. Therefore, governance for physician

XSL•FO
RenderX

resilience in AI-supported clinical care should extend beyond traditional alert management to include bounded verification triggers, calibrated interpretation of uncertainty, protection of unaided reasoning, and explicit boundaries for AI use within clinician-led decisions [32].

Limitations and Implementation Considerations

As a normative, conceptual viewpoint, this framework has several limitations. First, it does not yet provide prospective evidence regarding workflow efficiency, clinician acceptance, patient outcomes, or unintended consequences; feasibility and net benefit require pilot-testing and context-specific evaluation [43,44]. Second, safeguards such as verification triggers, diagnostic time-outs, attestation requirements, and targeted interprofessional review may introduce workflow friction, increased administrative burden, alert fatigue, or delays in time-sensitive settings if applied too broadly [25]. Emergency override pathways may also be overused, underused, or applied inconsistently in practice, requiring explicit criteria and review processes to support timely care without undermining accountability. Third, implementation capacity is likely to vary substantially across institutions, particularly in resource-constrained settings lacking sufficient staffing, digital infrastructure, monitoring capacity, or protected implementation time [23,44]. Fourth, sustaining a living ASoP table and longitudinal skill retention activities requires version control, periodic review, and coordination across multiple functions; institutions will need proportionate update cycles, shared templates, and risk-prioritized review processes to keep these mechanisms sustainable.

Accordingly, the framework should be understood as risk stratified and locally adaptable rather than uniformly applicable. In high-acuity contexts, immediate override with deferred review may be necessary to preserve timely care. Thresholds for verification, delegation, and authority sharing will require ongoing recalibration as models, interfaces, and clinical workflows evolve. Future work should identify which components are most feasible, which provide the greatest safety benefit, and how implementation burden can be minimized while preserving physician judgment. For institutions considering adoption, implementation should proceed incrementally, beginning with governance structures and bounded pilot workflows followed by prospective monitoring, recalibration, and staged expansion, with skill retention monitoring and incentive alignment incorporated from the outset rather than after scale-up [13,43,44].

Conclusions

Physician resilience in AI-supported clinical care depends on more than improved prompting or isolated human oversight requirements. This viewpoint advances a governance approach to generative AI integration that emphasizes bounded verification, preserved physician judgment, and organizational accountability as conditions for safe clinical use. Future work should prospectively evaluate the feasibility, clinical effects, and implementation burden of these mechanisms across specialties, workflows, and institutional settings. Such evaluation should include bounded pilot studies, mixed methods assessments of physician workload and autonomy, safety event monitoring, equity-focused subgroup analyses, and longitudinal assessments of skill retention.

Acknowledgments

After completing the manuscript, the authors used ChatGPT (GPT-5.5; OpenAI) to assist with language refinement. The authors reviewed and edited the output and take full responsibility for the final content of the manuscript.

Funding

The authors declared no financial support was received for this work.

Data Availability

Data sharing is not applicable to this paper as no datasets were generated or analyzed during this study.

Authors' Contributions

JL and SL contributed to conceptualization. JL, HP, and SL contributed to literature review. HP, JL, and SL contributed to writing—original draft. All authors contributed to writing—review and editing.

Conflicts of Interest

None declared.

References

1. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open 2024 Oct 1;7(10):e2440969. [doi: 10.1001/jamanetworkopen.2024.40969] [Medline: 39466245]
2. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun 28;25:e48568. [doi: 10.2196/48568] [Medline: 37379067]

XSL•FO
RenderX

3. Gaber F, Shaik M, Allega F, et al. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. NPJ Digit Med 2025 May 9;8(1):263. [doi: 10.1038/s41746-025-01684-1] [Medline: 40346344]
4. Liu S, Huang SS, McCoy AB, Wright AP, Horst S, Wright A. Optimizing order sets with a large language model-powered multiagent system. JAMA Netw Open 2025 Sep 2;8(9):e2533277. [doi: 10.1001/jamanetworkopen.2025.33277] [Medline: 40986301]
5. Zeng J, Zou X, Li S, et al. Assessing the role of the Generative Pretrained Transformer (GPT) in Alzheimer's disease management: comparative study of neurologist- and artificial intelligence-generated responses. J Med Internet Res 2024 Oct 31;26:e51095. [doi: 10.2196/51095] [Medline: 39481104]
6. Teo ZL, Thirunavukarasu AJ, Elangovan K, et al. Generative artificial intelligence in medicine. Nat Med 2025 Oct;31(10):3270-3282. [doi: 10.1038/s41591-025-03983-2] [Medline: 41053447]
7. Khera R, Simon MA, Ross JS. Automation bias and assistive AI: risk of harm from AI-driven clinical decision support. JAMA 2023 Dec 19;330(23):2255-2257. [doi: 10.1001/jama.2023.22557] [Medline: 38112824]
8. Topaz M, Peltonen LM, Zhang Z. Beyond human ears: navigating the uncharted risks of AI scribes in clinical practice. NPJ Digit Med 2025 Sep 24;8(1):569. [doi: 10.1038/s41746-025-01895-6] [Medline: 40993221]
9. Kim J, Podlasek A, Shidara K, Liu F, Alaa A, Bernardo D. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. Sci Rep 2025 Nov 11;15(1):39426. [doi: 10.1038/s41598-025-22940-0]
10. Kosior K, Wall T, Ferrero S. The role of metacognition in teaching clinical reasoning: theory to practice. Educ Health Prof 2019;2(2):108-114. [doi: 10.4103/EHP.EHP_14_19]
11. Sokol K, Fackler J, Vogt JE. Artificial intelligence should genuinely support clinical reasoning and decision making to bridge the translational gap. NPJ Digit Med 2025 Jun 10;8(1):345. [doi: 10.1038/s41746-025-01725-9] [Medline: 40494886]
12. Ghanbari S, Haghani F, Barekatain M, Jamali A. A systematized review of cognitive load theory in health sciences education and a perspective from cognitive neuroscience. J Educ Health Promot 2020;9:176. [doi: 10.4103/jehp.jehp_643_19] [Medline: 32953905]
13. Natali C, Marconi L, Dias Duran LD, Cabitza F. AI-induced deskilling in medicine: a mixed-method review and research agenda for healthcare and beyond. Artif Intell Rev 2025;58:356. [doi: 10.1007/s10462-025-11352-1]
14. Budzyń K, Romańczyk M, Kitala D, et al. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. Lancet Gastroenterol Hepatol 2025 Oct;10(10):896-903. [doi: 10.1016/S2468-1253(25)00133-5] [Medline: 40816301]
15. Abdelwanis M, Alarafati HK, Tammam MM, Simsekler MC. Exploring the risks of automation bias in healthcare artificial intelligence applications: a Bowtie analysis. J Saf Sci Resil 2024 Dec;5(4):460-469. [doi: 10.1016/j.jnlssr.2024.06.001]
16. Wiig S, Aase K, Billett S, et al. Defining the boundaries and operational concepts of resilience in the resilience in healthcare research program. BMC Health Serv Res 2020 Apr 19;20(1):330. [doi: 10.1186/s12913-020-05224-3] [Medline: 32306981]
17. Hollnagel E, Woods DD, Leveson N, editors. Resilience Engineering: Concepts and Precepts: CRC Press; 2006. [doi: 10.1201/9781315605685]
18. Funer F, Wiesing U. Physician's autonomy in the face of AI support: walking the ethical tightrope. Front Med (Lausanne) 2024;11:1324963. [doi: 10.3389/fmed.2024.1324963] [Medline: 38606162]
19. Grosser J, Düvel J, Hasemann L, Schneider E, Greiner W. Studying the potential effects of artificial intelligence on physician autonomy: scoping review. JMIR AI 2025 Mar 13;4:e59295. [doi: 10.2196/59295] [Medline: 40080059]
20. Chen RJ, Wang JJ, Williamson DF, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat Biomed Eng 2023 Jun;7(6):719-742. [doi: 10.1038/s41551-023-01056-8] [Medline: 37380750]
21. Nagendran M, Festor P, Komorowski M, Gordon AC, Faisal AA. Quantifying the impact of AI recommendations with explanations on prescription decision making. NPJ Digit Med 2023 Nov 7;6(1):206. [doi: 10.1038/s41746-023-00955-z] [Medline: 37935953]
22. Augmented intelligence in medicine. American Medical Association. 2026. URL: https://www.ama-assn.org/practice-management/digital-health/augmented-intelligence-medicine [accessed 2026-03-20]
23. Regulatory considerations on artificial intelligence for health. World Health Organization. 2023. URL: https://www.who.int/publications/i/item/9789240078871 [accessed 2026-01-20]
24. Clinical decision support software: guidance for industry and Food and Drug Administration staff. U.S. Food & Drug Administration. 2026. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software [accessed 2026-02-22]
25. Chaparro JD, Beus JM, Dziorny AC, et al. Clinical decision support stewardship: best practices and techniques to monitor and improve interruptive alerts. Appl Clin Inform 2022 May;13(3):560-568. [doi: 10.1055/s-0042-1748856] [Medline: 35613913]
26. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. World Health Organization. 2025. URL: https://www.who.int/publications/i/item/9789240084759 [accessed 2026-01-16]
27. Sakamoto T, Harada Y, Shimizu T. Facilitating trust calibration in artificial intelligence-driven diagnostic decision support systems for determining physicians' diagnostic accuracy: quasi-experimental study. JMIR Form Res 2024 Nov 27;8:e58666. [doi: 10.2196/58666] [Medline: 39602469]

XSL•FO
RenderX

28. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med 2024 Sep;30(9):2613-2622. [doi: 10.1038/s41591-024-03097-1] [Medline: 38965432]
29. Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. JMIR Med Inform 2024 Apr 8;12:e55318. [doi: 10.2196/55318] [Medline: 38587879]
30. Wu K, Wu E, Wei K, et al. An automated framework for assessing how well LLMs cite relevant medical references. Nat Commun 2025;16(1):3615. [doi: 10.1038/s41467-025-58551-6]
31. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240029200 [accessed 2026-01-20]
32. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (text with EEA relevance). European Union. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng [accessed 2026-01-16]
33. WHO calls for safe and ethical AI for health. World Health Organization. 2023. URL: https://www.who.int/news/item/16-05-2023-who-calls-for-safe-and-ethical-ai-for-health [accessed 2026-01-18]
34. Siegel RD, LeFebvre KB, Temin S, et al. Antineoplastic therapy administration safety standards for adult and pediatric oncology: ASCO-ONS standards. JCO Oncol Pract 2024 Oct;20(10):1314-1330. [doi: 10.1200/OP.24.00216] [Medline: 38776491]
35. Weingart SN, Li JW, Zhu J, et al. US cancer center implementation of ASCO/Oncology Nursing Society chemotherapy administration safety standards. J Oncol Pract 2012 Jan;8(1):7-12. [doi: 10.1200/JOP.2011.000379] [Medline: 22548004]
36. Martin DB, Kaemingk D, Frieze D, Hendrie P, Payne TH. Safe implementation of computerized provider order entry for adult oncology. Appl Clin Inform 2015;6(4):638-649. [doi: 10.4338/ACI-2015-03-RA-0027] [Medline: 26767061]
37. Augmented intelligence in health care. American Medical Association. URL: https://www.ama-assn.org/system/files/2019-01/augmented-intelligence-policy-report.pdf [accessed 2026-01-20]
38. Meeting the moment: addressing barriers and facilitating clinical adoption of artificial intelligence in medical diagnosis. National Academy of Medicine. 2022. URL: https://nam.edu/perspectives/meeting-the-moment-addressing-barriers-and-facilitating-clinical-adoption-of-artificial-intelligence-in-medical-diagnosis/ [accessed 2026-01-15]
39. Hospital survey on patient safety culture. Agency for Healthcare Research and Quality. 2019. URL: https://www.ahrq.gov/sops/surveys/hospital/index.html [accessed 2026-01-20]
40. Chew KS, Durning SJ, van Merriënboer JJ. Teaching metacognition in clinical decision-making using a novel mnemonic checklist: an exploratory study. Singapore Med J 2016 Dec;57(12):694-700. [doi: 10.11622/smedj.2016015] [Medline: 26778635]
41. Anderson JE, Ross AJ, Back J, et al. Implementing resilience engineering for healthcare quality improvement using the CARE model: a feasibility study protocol. Pilot Feasibility Stud 2016;2:61. [doi: 10.1186/s40814-016-0103-x] [Medline: 27965876]
42. Embí PJ, Rhew DC, Peterson ED, Pencina MJ. Launching the Trustworthy and Responsible AI Network (TRAIN): a consortium to facilitate safe and effective AI adoption. JAMA 2025 May 6;333(17):1481-1482. [doi: 10.1001/jama.2025.1331] [Medline: 39928518]
43. Reddy S. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. Implement Sci 2024 Mar 15;19(1):27. [doi: 10.1186/s13012-024-01357-9] [Medline: 38491544]
44. Scaling innovations in public health systems: guidance and toolkit. World Health Organization. 2026. URL: https://www.who.int/publications/i/item/9789240120761 [accessed 2026-02-20]

Abbreviations

AI: artificial intelligence
ASoP: automation scope of practice
CDS: clinical decision support

XSL•FO
RenderX

Edited by A Coristine; submitted 22.Nov.2025; peer-reviewed by J Grosser, KH Lin, R Poss-Doering; revised version received 11.May.2026; accepted 01.Jun.2026; published 24.Jun.2026.

Please cite as:
Pan H, Liu J, Liu S
Enhancing Physician Resilience to Generative AI: Multilevel Framework for Shared Authority, Verification, and Skill Preservation
J Med Internet Res 2026;28:e88058
URL: https://www.jmir.org/2026/1/e88058
doi:10.2196/88058

© Hongxia Pan, Jialin Liu, Siru Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Personal Health Large Language Models and the Negotiation of Medical Authority in Clinical Care: Opportunities, Risks, and Governance

Wenyi Xie[1], MSN; Jialin Liu[2,3], MD; Siru Liu[4], PhD

[1]Medical Intensive Care Unit, West China Hospital of Sichuan University, Chengdu, Sichuan, China
[2]Information Center, West China Hospital of Sichuan University, No. 37 Guo Xue Xiang Street, Chengdu, Sichuan, China
[3]Department of Otolaryngology-Head and Neck Surgery, West China Hospital of Sichuan University, Chengdu, Sichuan, China
[4]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

Corresponding Author:
Jialin Liu, MD
Information Center, West China Hospital of Sichuan University, No. 37 Guo Xue Xiang Street, Chengdu, Sichuan, China

Abstract

Personal health large language models (PH-LLMs) are patient-facing conversational systems that synthesize user-entered information, patient-generated health data, wearable data, and selected personal health records—where users choose to connect them—into personalized, longitudinal, action-oriented health narratives. Unlike generic health chatbots that mainly provide one-off responses to isolated questions, PH-LLMs may generate continuing interpretations, priorities, and candidate next steps that patients bring into clinical encounters. In contrast to electronic health record–tethered clinical artificial intelligence (AI), they often originate outside institutional oversight and may be selected, used, or trusted by patients before professional review. This viewpoint examines how PH-LLMs may reshape the negotiation of medical authority by contributing to a shift from the traditional dyadic clinician-patient relationship toward a triadic model of negotiated authority, in which clinicians may increasingly need to mediate among clinical evidence, patient values, and algorithmic narratives. PH-LLMs may support patient participation by organizing symptoms, contextualizing home-monitoring and wearable data, improving health literacy, assisting chronic disease self-management, and preparing patients for more collaborative visits. Patient-facing AI narratives may also introduce distinct risks. At the individual level, these include inaccurate or incomplete responses arising from imprecise queries or missing context, misinterpretation of otherwise accurate information in the absence of clinical context, and contextually biased or poorly matched advice across demographic, cultural, linguistic, disability-related, or socioeconomic contexts. At the system level, they include authority conflict when AI recommendations diverge from clinical judgment, fragmentation of clinical truth, privacy and data-governance concerns, diffusion of accountability when harm results from advice produced outside clinical governance, and inequitable access to premium tools and continuous monitoring devices. To address these challenges, we propose a 3-layer clinical governance framework for patient-brought PH-LLM narratives. The first layer, evidence and provenance, makes AI-generated narratives epistemically legible by clarifying platform identity, data sources, temporal anchoring, uncertainty, and privacy-relevant data-use and retention conditions. The second layer, clinical arbitration and workflow integration, uses risk-stratified intake, proportionate documentation, escalation triggers, and equity-preserving workflows to embed PH-LLM outputs into routine care. The third layer, competence and accountability, defines the communication competencies, AI literacy supports, institutional responsibilities, vendor accountability, and risk-proportionate verification duties needed for triadic care. This framework is a conceptual and governance-oriented proposal rather than a validated clinical protocol. Future empirical work should evaluate its feasibility, documentation burden, equity effects, clinical safety impact, and acceptability among patients, clinicians, and health systems. Governed through these interdependent layers, PH-LLMs may serve as supporting infrastructure for safer, person-centered longitudinal care.

(J Med Internet Res 2026;28:e91727) doi:10.2196/91727

KEYWORDS

personal health large language models; clinician-patient relationship; artificial intelligence; medical authority; clinical governance; shared decision-making; patient-generated health data; digital health; accountability

Introduction

In the conventional biomedical model, diagnostic reasoning and therapeutic recommendations have been formulated primarily within the clinical encounter, sustained by professional epistemic authority and institutional accountability [1]. This arrangement is increasingly challenged by the emergence of personal health large language models (PH-LLMs). In this viewpoint, PH-LLMs

XSL•FO
RenderX

refer to patient-facing or consumer-facing large language model (LLM)-based conversational systems that synthesize user-entered information, patient-generated health data, wearable data, and selected user-connected personal health records into personalized, longitudinal, action-oriented health narratives [2-4]. This patient-initiated access to personal health records is distinct from bidirectional electronic health record (EHR) integration governed by health systems and clinical workflows. For clarity throughout this viewpoint, by epistemic authority we refer to the capacity of a knowledge source, including clinicians, clinical records, or PH-LLM outputs, to shape how users interpret symptoms, risks, and care options before or between clinical encounters. Compared with generic health chatbots, which mainly provide one-off responses to isolated questions, PH-LLMs may sustain continuing interpretations, priorities, and candidate next steps that patients carry into clinical encounters. They also differ from EHR-tethered clinical artificial intelligence (AI): whereas EHR-tethered AI generally operates within health-system governance and clinician workflows, PH-LLMs are typically initiated, selected, or used by patients before professional review [5]. We adopt this functional definition based on user-facing behavior and deployment context rather than on any specific model architecture, because PH-LLMs evolve rapidly and span both proprietary and open-source implementations.

Recent commercial developments illustrate this emerging category. Examples include OpenAI's ChatGPT Health, which began rolling out in January 2026 as a dedicated health and wellness experience, and Fitbit's Gemini-powered personal health coach, which entered public preview initially for eligible US-based Fitbit Premium Android users in October 2025 [6-10]. These products are referenced solely as time-bound illustrations of how consumer-facing PH-LLMs are entering everyday use, not as evidence of clinical decision-making capability, and their features, availability, and access conditions remain subject to change. Unlike static metric dashboards, PH-LLMs may synthesize wearable data, user-entered information, patient-generated health data, and selected personal health records—when users choose to connect them—into plausible explanations and candidate next steps before professional review or clinical contextualization [11]. However, such synthesis should not be equated with reliable longitudinal medical reasoning. Evidence from clinical AI deployment, longitudinal EHR reasoning, and LLM assurance studies suggests that current systems may remain vulnerable to data shift, temporal-reasoning limitations, and hallucination when interpreting noisy, incomplete, device-dependent, or temporally distributed health data [12-14].

PH-LLM functions should, therefore, be understood along a risk continuum. Low-risk functions include general health education, lifestyle coaching, sleep or fitness support, and visit preparation focused on question framing or symptom organization. Higher-risk functions include chronic disease management, medication-related advice, diagnostic interpretation, escalation guidance, or recommendations that conflict with clinician judgment. In practice, a single user session may span multiple points along this continuum, and the same functional category may carry different levels of risk depending on the patient population, clinical context, and the specificity of the AI-generated recommendation. As patients use these systems across this continuum, they may enter clinical encounters with AI-generated interpretations, priorities, and candidate next steps that prestructure clinical sensemaking. This may contribute to a shift from a dyadic model of clinician-patient interaction toward a triadic model of negotiated authority, in which clinicians may increasingly need to mediate among clinical evidence, patient values, and algorithmic narratives (Figure 1). While this potential shift may support chronic disease self-management and shared decision-making, it also introduces risk pathways, including epistemic conflict, fragmentation of clinical truth, privacy and data-governance concerns, and diffusion of accountability [2,15].

The purpose of this viewpoint is to analyze how PH-LLMs may reshape the negotiation of medical authority and propose a clinical governance framework that preserves safety, accountability, and trust while enabling beneficial use. This viewpoint focuses specifically on patient-facing PH-LLMs whose personalized outputs may enter clinical encounters and influence shared understanding, patient expectations, or clinical decisions; it does not address all health-related chatbots or all forms of clinical AI. We examine the potential implications of this emerging shift for clinical practice, characterize key opportunities and risks, identify recurrent failure modes, and outline governance strategies across 3 domains: (1) evidence and provenance, (2) clinical arbitration and workflow integration, and (3) competence and accountability.

This viewpoint is conceptual and governance-oriented. To avoid overstating the evidence base, we distinguish time-bound examples of consumer-facing deployment [6-10], emerging peer-reviewed evidence on PH-LLM synthesis of wearable and patient-generated data [11], evidence on capability limitations from clinical AI deployment, longitudinal EHR reasoning, and LLM assurance studies [12-14], and conceptual claims about how PH-LLMs may reshape patient interpretation, clinician-patient interaction, and accountability [2,16-20]. Projections of future clinical practice and governance outcomes are conditional on governance design, institutional adoption, and policy uptake rather than predictions of inevitable change. The proposed shift toward a triadic clinician-patient-PH-LLM relationship is therefore presented as an emerging governance concern rather than as an empirically established transformation. The observable markers in Table 1 are intended to support future empirical evaluation of these propositions, including their magnitude, distribution across populations, equity implications, and clinical safety effects.

XSL•FO
RenderX

Figure 1. Transition from a dyadic clinician-patient relationship to a triadic clinician-patient-PH-LLM relationship. (A) The traditional dyadic model, in which clinical reasoning, risk interpretation, and treatment recommendations are negotiated primarily between the clinician and the patient. (B) The emerging triadic model, in which patients may use PH-LLMs outside institutional oversight, such as between visits or before encounters, and bring AI-generated summaries, interpretations, or recommendations into the visit. Solid lines indicate direct interactions between the patient and clinician, and between the patient and PH-LLM. The dotted blue line between the clinician and PH-LLM represents an indirect, patient-mediated relationship rather than direct interoperability, formal EHR integration, or clinician control. Clinicians therefore encounter PH-LLMs mainly through patient-brought outputs and must mediate among clinical evidence, patient values, and algorithmic narratives that may diverge from clinical judgment. This figure also illustrates that PH-LLMs may redistribute epistemic authority across 3 actors and thereby create a need for point-of-care governance mechanisms. AI: artificial intelligence; PH-LLM: personal health large language model.

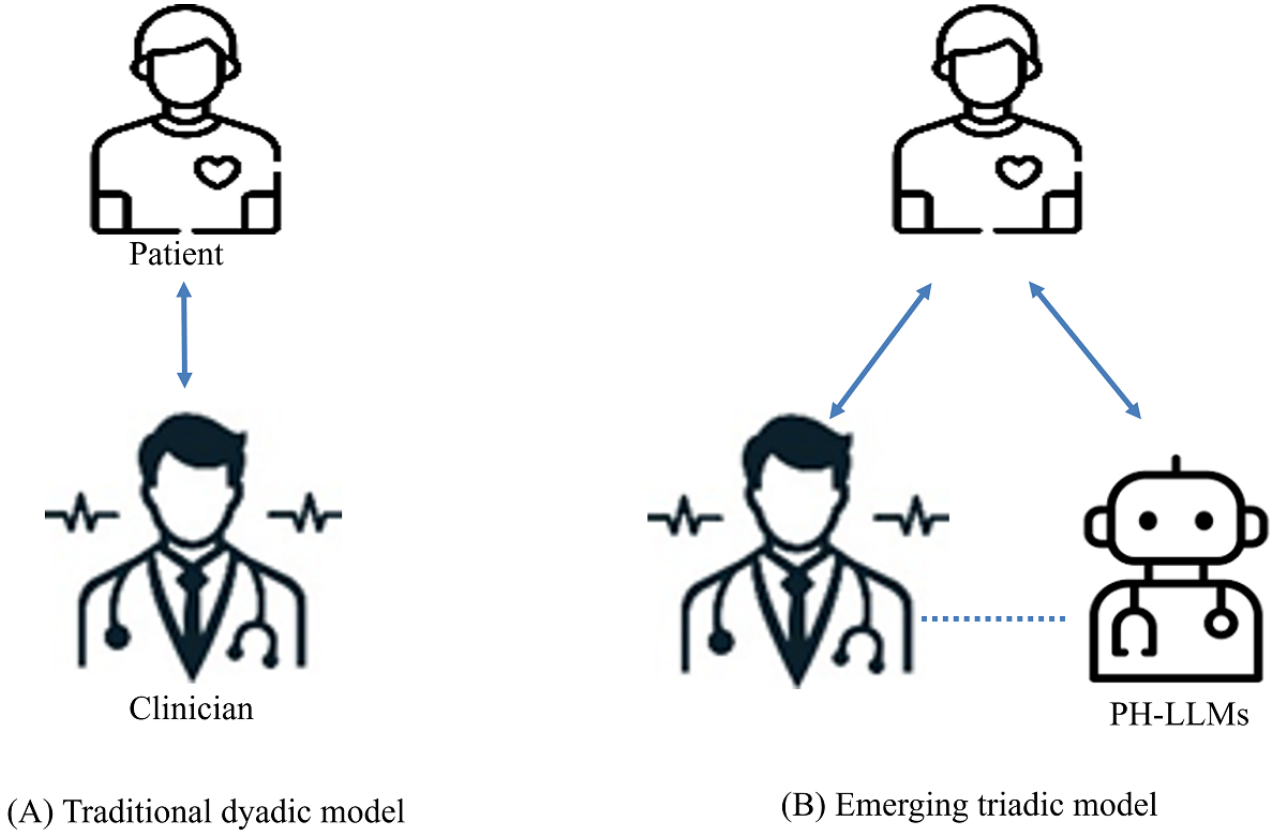

XSL•FO
RenderX

Table . Sociotechnical concepts and candidate observable markers for the empirical evaluation of the integration of personal health large language models (PH-LLMs) in clinical encounters.

Sociotechnical concept	Candidate observable markers[a]
Triadic encounter and negotiated authority	Encounters in which patients introduce AI[b]-generated content, visit time allocated to AI-mediated discussion, and shared decisions in which AI advice is explicitly considered
Authority conflict	Documented AI-clinician disagreement; plan modification, follow-up, or escalation after disagreement; and patient-reported adherence changes after disagreement
Fragmentation of clinical truth	Multiple AI platforms or wearable systems referenced in one visit, conflicting AI-generated interpretations, and clinician time spent reconciling inconsistent narratives
Diffusion of accountability	Complaints or grievances referencing AI advice; AI-related adverse-event reports; and completeness of model, platform, timestamp, and data-source information available for review
Attention gradient and equity	Differences in clinician engagement time, documentation, follow-up, or escalation between patients with and without AI summaries; and differences stratified by language, socioeconomic status, age, disability, rurality, or digital literacy
Documentation burden	Documentation time per AI-related event, length of AI-related documentation, and clinician-reported burden of AI verification and arbitration

[a]Markers are illustrative and intended to support future empirical evaluation; they are not validated quality measures, performance indicators, clinical protocols, or liability-assignment tools. Their validity, feasibility, equity implications, and risk-stratification sensitivity require further study.

[b]AI: artificial intelligence.

Opportunities: Patient Empowerment and the Negotiation of Medical Authority

The potential redistribution of epistemic authority associated with PH-LLM use is not inherently destabilizing. When appropriately governed, these systems may enhance patient participation, improve continuity of care, and support shared decision-making.

Supporting Patient Participation and Health Literacy

PH-LLMs may strengthen patients' capacity to participate actively in clinical encounters by supporting personalized sensemaking and preference-sensitive deliberation, although direct evidence for these effects remains limited. By helping users organize symptoms, contextualize home-monitoring data such as weight, blood pressure, glucose, sleep, and activity, and identify questions or concerns for discussion, these systems may enable patients to arrive better prepared for shared decision-making [11]. These functions should augment patients' participation in clinical reasoning, not substitute for clinician judgment or validated clinical decision support. PH-LLMs may also support health literacy by translating medical terminology, explaining general health concepts at an appropriate reading level, and helping patients formulate questions about test results, care options, and value-sensitive trade-offs [21,22]. Such support may be valuable for individuals with limited baseline health knowledge or low confidence navigating clinical information. At the same time, without safeguards for accuracy, comprehensibility, cultural appropriateness, and bias, PH-LLMs may amplify misunderstanding or misinformation, particularly among underserved populations, despite appearing empathetic and authoritative [2,23].

Between-Visit Support and Care Delivery Efficiency

At the care-delivery level, LLM-enabled between-visit support may strengthen continuity in chronic disease self-management and extend informational or behavioral support between encounters [24]. Potential applications include medication adherence reminders, structured symptom tracking, and prompts that help patients recognize when to seek professional input [25]. When integrated into well-designed care pathways, such support may contribute to more organized between-visit communication and improve the triage of urgent portal messages [26].

Collaborative Mediation and the Reframing of Medical Authority

At the level of clinical encounters, PH-LLMs may support more individualized care discussions by organizing patient-reported outcomes, wearable trends, and prior clinical information, where available, into structured summaries that inform agenda setting and shared decision-making [11,27]. Although these systems are sometimes framed as enabling precision medicine, that characterization should be applied cautiously. Current evidence suggests that these systems are better suited to clinical prioritization and patient preparation than to individualized treatment adjustment, which requires rigorous clinician oversight [2,18,28,29]. In this context, clinician authority may be reframed less as unilateral information control and more as collaborative mediation among clinical evidence, patient goals, and AI-generated narratives [30]. This may create a triadic interaction in which the PH-LLM functions as a shared interpretive object that can structure questions, clarify preferences, and support communication [31].

XSL•FO
RenderX

Risks: Challenges to Medical Order and Medical Authority

Pre-Encounter Sensemaking as a Source of Risk

Personal health LLMs introduce risks that extend beyond generic health misinformation [15,32]. Because they can generate coherent, personalized, and action-oriented narratives whose surface fluency may exceed their reasoning reliability, they may shape patients' interpretations of symptoms, risks, and treatment trade-offs before the clinical encounter [16]. This concern is heightened by current limitations in longitudinal medical reasoning, temporal data integration, distribution shift, and hallucination under realistic data conditions [12-14]. Although upstream sensemaking may improve patient preparedness, plausible but nonvalidated outputs may also lead to unsafe actions, distorted expectations, or erosion of therapeutic alignment [17,18].

Patient-Facing Risks at the Individual Level

Three patient-facing risks may arise at the individual encounter level, even when the model is not overtly hallucinating. First, PH-LLM outputs are sensitive to how patients formulate questions. Users without medical training may describe symptoms ambiguously or omit key context such as red-flag features, comorbidities, or medication use. The system may then generate confident but inaccurate responses that patients carry into clinical encounters as if they reflect validated clinical reasoning [23]. Second, even factually accurate outputs may be misread as personalized recommendations rather than general information requiring individualized clinical judgment. A correct explanation of glycosylated hemoglobin (HbA_{1c}), for example, may be misapplied if a patient interprets a specific value without considering comorbidities, hypoglycemia risk, or treatment goals [19,22]. Third, population-level accuracy does not guarantee individual-level appropriateness. PH-LLM outputs may reflect demographic, linguistic, or socioeconomic biases, particularly for populations underrepresented in training data, so correct general information may be misapplied when a patient's circumstances differ from the model's implicit reference case [2,23]. These risks warrant health-literacy adaptation, fairness monitoring, and clinician review of higher-risk advice.

Equity and Access Risks

Equity concerns arise before any model output is generated. Access to paid or premium PH-LLM services, wearable devices, reliable internet connectivity, compatible smartphones, and the digital literacy needed to critically interpret AI-generated narratives is unevenly distributed across populations [33,34]. The triadic model may therefore initially benefit those already better positioned to monitor, organize, and advocate for their health. This imbalance could create an attention gradient: patients who arrive with detailed AI-generated summaries may receive more clinician attention and individualized discussion, while those without such tools may have fewer opportunities to shape the encounter. PH-LLMs should therefore not be assumed to be equity-enhancing simply because they support personalization; their equity effects may vary according to access, usability, affordability, digital literacy, and the fairness of generated outputs [2,18,33].

Between-Visit Risks: False Alarms and False Reassurance

Between-visit PH-LLM support may produce opposing failure modes. The same functionality that supports continuity may increase utilization through false alarms or delay care through false reassurance when symptom interpretation, triage advice, or escalation guidance is inaccurate or poorly calibrated to clinical urgency. Access and efficiency gains from PH-LLM–enabled between-visit support should therefore be treated as conditional and context-dependent rather than as inherent effects of deployment [35,36].

Authority Conflict: When Algorithmic and Clinical Recommendations Diverge

A central risk arises when PH-LLM–generated recommendations contradict clinical guidance. Whereas traditional search engines return ranked links that require users to synthesize and interpret information, PH-LLMs can generate contextually tailored, assertive, and seemingly well-referenced recommendations that patients may interpret as equivalent to, or even superior to, clinician judgment [19,22]. As a hypothetical illustration, a patient with type 2 diabetes may receive an LLM-generated suggestion to intensify basal insulin based on continuous glucose monitor data, while the endocrinologist, aware of recent hypoglycemic episodes not captured in the data stream, advises against intensification. In such a scenario, 2 apparently credible sources offer contradictory guidance without a transparent basis for adjudication. Such conflicts may lead some patients to privilege algorithmic outputs over professional judgment, potentially contributing to unsafe actions, reduced trust, or erosion of the therapeutic alliance [19,20]. Clinicians may also experience professional strain when their expertise is challenged by outputs whose assumptions, data sources, and evidentiary limits are difficult to inspect. These tensions are compounded by technical failure modes, including domain mismatch, where wellness-trained models pathologize normal variation; temporal anchoring errors, where models rely on outdated clinical markers; and hallucination, where fabricated assertions are presented with unwarranted confidence [12-14]. Even when clinicians recognize these limitations, patients may lack the expertise to determine when algorithmic advice has exceeded its evidentiary warrant.

Fragmentation of Clinical Truth: Competing Narratives Across Platforms

Whereas authority conflict concerns tension between algorithmic and clinical guidance, fragmentation arises when multiple PH-LLMs generate divergent interpretations from partial, heterogeneous, or noninteroperable data sources [33,34]. Consumer-facing systems often operate within proprietary data silos, limiting reconciliation across consumer apps, wearable devices, and clinical records. Patients may therefore receive internally coherent but mutually inconsistent health narratives. Such fragmentation may impair the formation of a coherent personal health narrative and contribute to decision paralysis or motivated reasoning, in which patients favor the most

XSL•FO
RenderX

reassuring platform over the most clinically appropriate interpretation [37-39]. When platform design is optimized for user engagement, conversational coaching may include behavioral nudges that are not transparently separable from clinical reasoning, raising concerns about persuasive design and commercial conflict [40,41]. Clinicians may therefore be drawn into epistemic triage: identifying clinically valid elements, correcting algorithmically anchored misunderstandings, and reconciling competing narratives without eroding therapeutic trust.

Diffusion of Accountability: Who Is Responsible When Harm Occurs?

Diffusion of accountability represents a distinct failure mode that concerns governance rather than information quality alone [2,42]. Conventional clinical care rests on defined structures of responsibility: clinicians are accountable for recommendations within their scope of practice, health systems assume organizational duties of care, and regulators set enforceable standards [43,44]. PH-LLMs complicate this architecture by introducing an external source of personalized, clinically suggestive advice that generates diagnostic-style narratives and action-oriented recommendations without comparable mechanisms for auditability, oversight, or duty assignment [45,46]. When patients act on PH-LLM–mediated advice and harm occurs, responsibility may fragment across patients, clinicians, health systems, developers, platforms, and data intermediaries, particularly when data inputs, model assumptions, update history, and downstream data uses remain difficult to inspect [42,47].

This accountability gap may be reinforced when consumer platforms frame outputs as “informational” in their terms of service while presenting them in a clinical tone and with high confidence [48,49]. Clinicians may be asked to respond to advice they did not generate and cannot meaningfully audit, while health systems may lack visibility into the guidance patients receive between visits. Patients, encountering personalized and continuous recommendations, may reasonably treat these outputs as medically authoritative [18,50]. The result is a mismatch between clinical influence and accountable oversight: patients have limited recourse, clinicians face uncertain professional obligations, and health systems remain exposed to risks from decisions shaped outside institutional governance [42,45,47].

Data-governance opacity further compounds this problem. PH-LLMs may rely on sensitive patient-generated health data, wearable streams, and conversational histories, yet consumer-grade practices may not meet health care–grade expectations for security, use limitation, transparency, and breach response [51]. In a distributed ecosystem of device manufacturers, app developers, model providers, cloud services, and data intermediaries, harms arising from unsafe recommendations, data breaches, or opaque secondary uses may be difficult to attribute and remediate [52].

Privacy and Data Governance

Privacy and data governance warrant dedicated attention in consumer-facing PH-LLMs because users may disclose sensitive health information through open-ended, longitudinal dialog. Unlike static health websites, PH-LLMs may accumulate multisession narratives involving symptoms, medications, mental health concerns, family history, wearable-device trends, and crisis-related disclosures. The extent of such accumulation depends on system design, user settings, and data-retention practices. Privacy risks, therefore, extend beyond unauthorized disclosure to include secondary use, long-term retention, cross-session profiling, inferred sensitive attributes, and blurred boundaries among safety monitoring, product improvement, commercial optimization, and research use [51-53].

Addressing these risks requires context-sensitive governance. General health questions may be managed through baseline consent and adjustable retention settings, whereas mental health disclosures, crisis-related content, and longitudinal behavioral narratives require stronger safeguards, including data minimization, tiered consent, deletion controls, and restrictions on secondary use, advertising, or model training without appropriate authorization [49,51]. Crisis-related content requires an immediate safety response without consent-related delay, followed by access-restricted retention justified by safety or safeguarding needs [2,18,49,51].

Fairness monitoring should be incorporated without making the routine collection of protected-attribute data the default [33,34,51]. Where such data are used for disparity evaluation, they should be separated from interaction content, analyzed in aggregate, and protected through access controls and independent review where feasible [2,33,49,51]. Youth-facing deployments may require additional safeguards for age-appropriate defaults, confidential help-seeking, and crisis-response pathways consistent with applicable child and adolescent data-protection requirements [54,55].

Summary of System-Level Risks

Taken together, the system-level risks discussed above are mutually reinforcing: authority conflict may worsen fragmentation, fragmentation may obscure responsibility for downstream harms, and unclear privacy practices may undermine clinical appraisal and accountability. When data sources, retention practices, or platform responsibilities are insufficiently disclosed, clinicians and health systems may struggle to assess patient-facing AI narratives and attribute responsibility when harms occur. Privacy governance should therefore be treated as a core component of PH-LLM clinical governance rather than as a peripheral compliance issue.

Clinical Governance: Managing the Shift and Its Implications

The governance challenge posed by PH-LLMs is not one of comprehensive exclusion but rather of structured clinical integration. Given their increasing availability in consumer-facing health and wellness contexts [50], the comprehensive prohibition of patient-facing AI tools may be neither feasible nor desirable in many settings. Clinical governance should, therefore, focus on making AI-generated narratives visible, assessable, negotiable, and reversible within the professional encounter.

XSL•FO
RenderX

Limitations of Existing Frameworks

Current regulatory and ethical frameworks for AI in health, including those addressing risk classification, safety assurance, and ethics governance, provide important but incomplete foundations [2,48,56,57]. These frameworks primarily emphasize upstream safety assurance, developer responsibilities, and postdeployment monitoring. They offer limited guidance for situations in which AI-generated narratives originating outside institutional boundaries intersect with clinician judgment, workflow constraints, and shared decision-making. In particular, they do not adequately address: (1) how clinicians should evaluate the credibility of AI outputs generated outside institutional oversight; (2) how conflicts between algorithmic and clinical recommendations should be mediated in practice; and (3) how accountability should be documented when AI advice shapes, but does not determine, clinical decisions. These gaps call for a point-of-care governance model that complements product-level certification with encounter-level governance of patient-brought AI advice.

A 3-Layer Clinical Governance Framework for PH-LLMs

Framework Development

To address these gaps, we propose a 3-layer clinical governance framework comprising interdependent layers: (1) evidence and provenance, (2) clinical arbitration and workflow integration, and (3) competence and accountability (Figure 2). The framework was developed through a viewpoint synthesis appropriate to this article type, rather than a systematic review. The synthesis drew on existing AI governance literature [2,18,56-58], digital health and patient-generated health data literature [33,34,51], conceptual frameworks on clinician-patient-AI interaction [30,31], and published literature and time-bound examples of current and emerging consumer-facing PH-LLM use. These inputs were synthesized to identify recurring governance problems, organize them into a 3-layer framework, and propose candidate observable markers for future empirical evaluation (Table 1). The framework is intended to guide the structured integration of PH-LLM narratives into professional practice while safeguarding clinical safety, institutional accountability, and the therapeutic alliance.

Figure 2. Three-layer clinical governance framework for patient-brought PH-LLM narratives. The framework comprises 3 interdependent layers. Layer 1, evidence and provenance, makes AI-generated narratives legible by identifying the model or platform, data sources, timestamps, uncertainty signals, and privacy-relevant data-use conditions. Layer 2, clinical arbitration and workflow integration, provides a risk-stratified process for determining whether AI advice can be addressed through routine counseling, clinician review, same-day escalation, or specialist input. Layer 3, competence and accountability, clarifies communication skills, documentation expectations, verification duties, and responsibility boundaries among patients, clinicians, institutions, developers, and platforms. Together, these layers are intended to make patient-facing AI narratives visible, assessable, negotiable, privacy-aware, proportionately documented, and accountable within routine clinical workflows. AI: artificial intelligence; PH-LLM: personal health large language model.

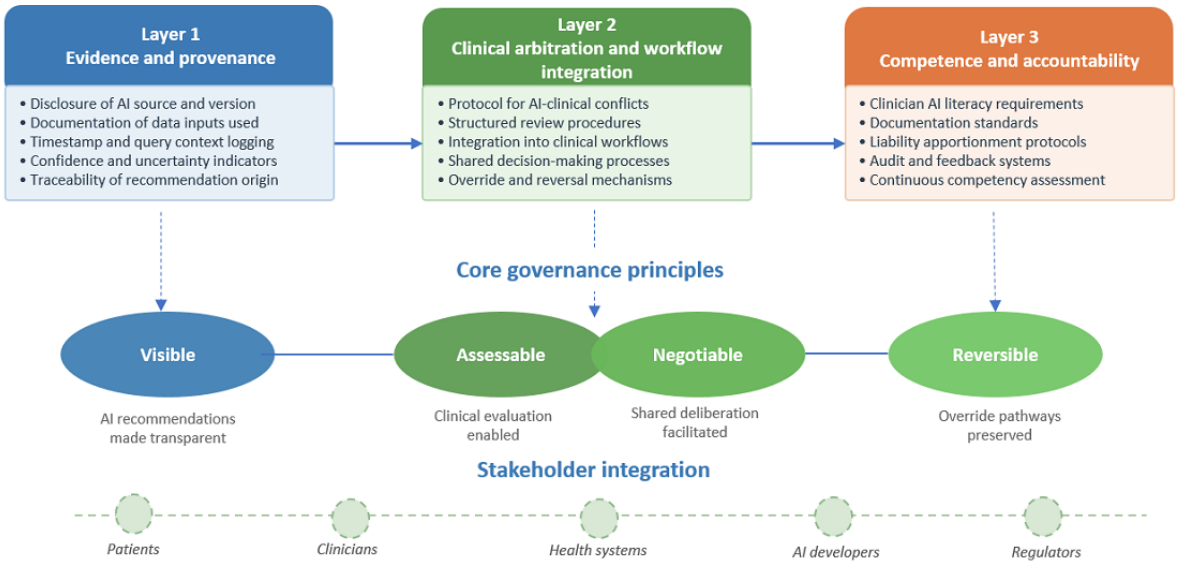

Layer 1: Evidence and Provenance

A foundational requirement for clinically consequential AI-generated recommendations entering clinical discussion is epistemic legibility: clinicians cannot meaningfully evaluate or negotiate AI advice unless its evidentiary basis is sufficiently visible for clinical appraisal [56,57]. This requirement should be understood as a governance standard for PH-LLMs seeking institutional integration or trusted-use designation, not as a transparency obligation that frontline clinicians can impose on proprietary platforms during routine encounters. For clinically consequential AI outputs, developers and platforms should disclose: (1) model identity and version; (2) data sources, such as wearable data, self-reported symptoms, user-connected personal health records, or user-entered information; (3) temporal anchoring, including timestamps and data windows for key variables; and (4) uncertainty signals distinguishing measured data from inferred or extrapolated conclusions [59]. For example, a patient with hypertension may present a PH-LLM summary concluding that “blood pressure control has worsened” and recommending treatment intensification. A layer 1 provenance check would establish whether this conclusion

XSL•FO
RenderX

was based on validated home measurements, wearable estimates, manual entries, or a trend inferred from incomplete data, together with the relevant dates and measurement window. Readings that are recent, consistently elevated, and concordant with the clinical record may inform discussion as patient-provided context; an unclear data source would require confirmation through standard clinical assessment before being treated as clinically actionable evidence.

When provenance is unavailable, clinicians should treat the output as patient-reported contextual information rather than independently verifiable clinical evidence. Clinicians should check clinically consequential claims against the medical record, current clinical findings, guideline-based care, and professional judgment. At the institutional level, health systems may require provenance disclosure, evidence links, audit logs, and data-source labeling as conditions for procurement, formal EHR integration, preferred app lists, or partnerships with commercial PH-LLM vendors [57,58]. Regulators, professional societies, payers, and accreditation bodies may be better positioned to establish enforceable transparency requirements for consumer-facing systems that generate clinically influential recommendations. In practice, full provenance disclosure may remain limited for many consumer-facing PH-LLMs in the near term. Layer 1 should therefore be understood as an aspirational governance standard directed primarily at developers, institutions, and regulators, while layer 2 provides the operational mechanism for handling AI-generated advice at the point of care when provenance is incomplete or unavailable.

Layer 2: Clinical Arbitration and Workflow Integration

Governance must extend beyond transparency to specify how AI-generated advice is handled within clinical workflows. We propose a risk-stratified intake process in which AI-generated advice is categorized as follows: (1) lifestyle and health-promotion guidance, generally low risk; (2) disease-management, medication-related, diagnostic-interpretation, or monitoring advice, moderate to high risk; or (3) safety-critical alerts requiring immediate escalation [17,30,31,48]. To remain feasible, this process should not require extensive documentation for every AI-related discussion. Low-risk advice may be addressed through ordinary clinical counseling and documented, if needed, within the standard encounter note. More structured arbitration should be reserved for cases in which AI-generated advice materially conflicts with clinical judgment, guideline-based care, medication decisions, diagnostic interpretation, or escalation thresholds. Safety-critical alerts, including indications of acute risk such as suicidal ideation, cardiac symptoms, or severe medication interactions, should trigger immediate clinical escalation through established institutional pathways rather than routine arbitration.

Workflow integration should also protect equity in clinical attention because patient-generated health data, wearable devices, and commercial digital tools are unevenly available across populations [33,34]. Triage should remain grounded in clinical acuity, patient vulnerability, and care needs rather than the presence or sophistication of an AI-generated summary. Health systems may reduce inequity through standardized previsit symptom forms, nonproprietary patient-reported outcome tools, interpreter-accessible digital supports, and staff-assisted intake, helping prevent PH-LLM use from becoming an informal gateway to more attentive care.

When AI-generated advice falls into moderate-risk or high-risk categories, documentation should be embedded into existing workflows rather than added as a separate administrative layer. Practical mechanisms may include previsit intake questions, patient-uploaded screenshots with timestamps and data windows where appropriate, EHR smart phrases, check-box templates, and team-based triage before clinician review [34,48,51]. For moderate-risk or high-risk advice, including safety-critical items, minimum documentation should capture the clinically relevant AI-generated claim or recommendation, the reason for accepting, modifying, or rejecting it, any patient preference that materially affected the decision, and the final shared plan. This approach supports accountable reasoning without converting AI-related disagreement into an excessive administrative burden. To make this layer actionable, Table 2 provides an illustrative PH-LLM intake template organized around 4 stages: patient-provided information, staff triage, minimum documentation, and escalation triggers. The template is risk-calibrated and intended for local adaptation rather than as a validated protocol.

XSL•FO
RenderX

Table . Illustrative personal health large language models (PH-LLM) intake template for receiving, triaging, documenting, and escalating patient-brought artificial intelligence (AI) advice in routine encounters.[a]

Operational stage	Recommended elements	Example operationalization
Patient-provided information	Screenshot or text of the AI output; platform or model name if visible; date and time; relevant data window, such as 7 or 30-day wearable trends; patient's reason for raising the advice	Portal upload, check-in photograph, or brief intake note, such as, "I asked [PH-LLM] on [date] about [concern]."
Staff triage	Apply the layer 2 risk-tier classification to incoming AI advice: low risk, moderate or high risk, or safety-critical	EHR[b] smart phrase or intake prompt with risk-tier checkbox; moderate-risk or high-risk and safety-critical items routed for clinician review
Minimum documentation	For moderate-risk, high-risk, and safety-critical advice: AI-generated claim, clinician rationale for accepting, modifying, or rejecting it, patient preference if material, and final shared plan[c]	Embedded EHR smart phrase or check-box template, rather than a separate administrative form
Escalation triggers	Same-day review, specialist consultation, or safety pathway when AI advice may affect medication, diagnostic interpretation, or escalation decisions; conflicts with recommended treatment; poses substantial foreseeable harm; or shows recurrent unsupported reliance	"AI-Flag" routed to clinician inbox; same-day phone or telehealth contact for safety-critical items; specialist consultation or safety event report for clinically consequential or recurrent conflicts

[a]This template is illustrative and risk-calibrated; fields, thresholds, and pathways should be locally adapted and empirically evaluated.

[b]EHR: electronic health record.

[c]Low-risk advice does not require structured documentation and may be addressed within the standard counseling note when clinically relevant.

This layer should also include reversibility mechanisms. Reversibility does not mean overriding competent patient autonomy or prohibiting PH-LLM use; rather, it means revisiting, contextualizing, or limiting the clinical influence of AI advice that is unsupported, outdated, unsafe, or inconsistent with the patient's broader clinical context. Low-risk lifestyle guidance may be managed through shared decision-making, risk communication, monitoring, and follow-up, whereas clinically consequential or safety-critical conflicts may require secondary review, specialist consultation, or institutional escalation [17,20,30]. For external consumer-facing systems outside institutional control, the absence of provenance limits the evidentiary weight of the output, and clinically consequential recommendations should not be relied on without independent verification. This approach reframes conflict not merely as an error to be corrected but as a managed negotiation within accountable care, while avoiding a paternalistic model in which clinician authority automatically displaces AI-informed patient preferences.

Layer 3: Competence and Accountability

Effective governance requires new competencies and clarified responsibility boundaries. For clinicians, this includes AI-related clinical communication competence: the ability to redirect AI-driven disagreement away from comparative intelligence, such as "AI versus clinician," toward evidence quality, risk assessment, patient values, and verifiable outcomes [53]. Clinicians should also recognize automation bias in patients and use concise, nonconfrontational strategies to restore epistemic balance without undermining patient agency [17,19,20]. Patients, in turn, require basic AI literacy. Health systems should provide standardized guidance, analogous to medication education, on red-flag symptoms, data currency, and documentation practices, including screenshots with timestamps and relevant data windows [17,20,53]. At the institutional level, health systems should establish documentation standards, escalation pathways, audit mechanisms, and adverse-event review processes for AI-influenced clinical decisions [42,46,48].

Accountability should be allocated according to control and role. Clinicians should not be held responsible for the internal logic, model design, data-processing errors, or algorithmic failures of PH-LLMs they did not develop, deploy, control, or audit [42,45,47]. However, when AI-generated advice is presented during care and materially influences diagnosis, treatment, escalation, or shared decision-making, clinicians may retain a professional duty to evaluate clinically consequential claims through reasonable, risk-proportionate verification [30,43,47,48]. Developers and platforms remain responsible for product design, provenance, privacy practices, and algorithmic behavior; institutions are responsible for workflow standards, escalation pathways, documentation expectations, and audit mechanisms; and clinicians are responsible for responding reasonably to AI-influenced claims that become visible and clinically relevant [42,45-47]. For example, a patient with chronic kidney disease may report a PH-LLM recommendation for frequent over-the-counter nonsteroidal anti-inflammatory drug use. Under layer 3, the clinician is not responsible for the model's reasoning or design, but has a risk-proportionate responsibility to evaluate a claim that is clinically visible and consequential. This would include verifying renal function and medications, explaining the risk, documenting the decision to reject or modify the advice, and offering a safer alternative. Recurrent or foreseeably harmful recommendations could be escalated through medication-safety review, a digital health governance committee, or vendor-reporting channels.

XSL•FO
RenderX

This distinction helps clarify the liability gray area. Documented verification should not be understood as automatically transferring liability for opaque external systems to clinicians; rather, it creates an auditable record showing that clinically consequential AI advice was accepted, modified, or rejected on the basis of clinical evidence, patient context, and professional judgment [43,47,48]. Conversely, ignoring a clearly safety-relevant AI-generated claim or relying on it without appropriate verification may increase professional or institutional risk under existing negligence standards [43,44,47]. Clearer institutional policies and regulatory guidance are therefore needed to distinguish product-level accountability from clinician-level duties of reasonable verification in triadic clinician-patient-PH-LLM interactions [31,42,44-48].

Toward Empirical Evaluation: Observable Markers

Key concepts in this viewpoint, including epistemic authority, negotiated authority, authority conflict, fragmentation of clinical truth, and diffusion of accountability, describe potential sociotechnical implications associated with PH-LLM use in clinical encounters [30,42,45,53]. To support future empirical examination, Table 1 maps these concepts to candidate observable markers. These markers are illustrative rather than validated performance metrics, quality indicators, clinical protocols, or liability-assignment tools; their validity, feasibility, equity implications, and sensitivity to clinical risk stratification require further study [34,51]. They may be operationalized through structured clinical documentation, EHR smart phrases, encounter recording or time-and-motion studies where ethically approved and consented, patient-reported measures, equity-stratified visit analytics, and patient safety event reporting systems [34,46,48]. Their use should be proportionate to clinical risk and should avoid adding excessive documentation burden or creating incentives to privilege AI-generated summaries over clinically similar concerns raised without AI assistance.

Limitations

This viewpoint has several limitations. First, it is conceptual and governance-oriented; the proposed 3-layer framework, intake template (Table 2), and observable markers (Table 1) are illustrative, have not been empirically validated, and should not be used as performance measures, quality indicators, or clinical protocols without further validation. Second, as a viewpoint study, the literature synthesis is narrative and conceptually motivated rather than systematic; readers seeking exhaustive evidence appraisal on specific PH-LLM capabilities or harms should consult systematic reviews, benchmark studies, and future empirical evaluations as the evidence base matures. Third, the framework is informed primarily by US clinical, regulatory, and liability contexts, and adaptation may be required for other health systems, financing and payment models, professional norms, and legal frameworks. Fourth, the framework focuses primarily on patient-facing conversational PH-LLM narratives brought into clinical encounters and may require adaptation for fully multimodal systems, EHR-integrated tools, or institutionally deployed clinical decision-support systems. Fifth, the framework addresses downstream clinical governance of PH-LLM advice in encounters rather than upstream technical infrastructure of PH-LLM deployment. Service-level and infrastructure factors, such as latency, uptime, update frequency, runtime reconfiguration, and quality-of-service management, may affect deployed PH-LLM reliability, but detailed infrastructure governance is outside the scope of this framework [60,61]. Finally, the consumer-facing product examples cited, such as ChatGPT Health and Fitbit's Gemini-powered health coach, are time-bound illustrations of an evolving ecosystem; their features, availability, and access conditions remain subject to change. Future work should prospectively evaluate the feasibility, documentation burden, equity effects, clinical safety impact, and acceptability of the proposed governance components across diverse care settings.

Conclusions

Ungoverned PH-LLMs may destabilize how clinical authority is negotiated, fragment therapeutic relationships, and diffuse accountability when patient-facing AI narratives enter care. When governed through the proposed 3-layer framework, these narratives may instead be incorporated as supporting infrastructure for safer, person-centered, longitudinal care. The task is not to exclude AI from clinical relationships, but to render AI-generated narratives clinically visible, negotiable, verifiable, and accountable within structured clinical governance.

Acknowledgments

After drafting the manuscript, the authors used ChatGPT (OpenAI) for artificial intelligence-assisted language editing to improve clarity and style. All suggestions were critically evaluated and further substantive revisions were made by the authors, who take full responsibility for the final content's integrity and accuracy.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Data Availability

Data sharing is not applicable to this paper as no new datasets were generated or analyzed during this study.

XSL•FO
RenderX

Authors' Contributions

JL and SL conceived and designed the study. WX, SL, and JL conducted the literature review and contributed to the conceptual synthesis. JL, WX, and SL contributed to writing the original draft. All authors critically revised the manuscript, approved the final version, and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

References

1. Engel GL. The need for a new medical model: a challenge for biomedicine. Science 1977 Apr 8;196(4286):129-136. [doi: 10.1126/science.847460] [Medline: 847460]
2. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. : World Health Organization; 2024 URL: https://iris.who.int/server/api/core/bitstreams/e9e62c65-6045-481e-bd04-20e206bc5039/content [accessed 2026-01-10]
3. Merrill MA, Paruchuri A, Rezaei N, et al. Transforming wearable data into personal health insights using large language model agents. Nat Commun 2026 Jan 12;17(1):1143. [doi: 10.1038/s41467-025-67922-y] [Medline: 41526380]
4. Tilton AK, Caplan BE, Cole BJ. Generative AI in consumer health: leveraging large language models for health literacy and clinical safety with a digital health framework. Front Digit Health 2025;7:1616488. [doi: 10.3389/fdgth.2025.1616488] [Medline: 40933812]
5. Ayo-Ajibola O, Davis RJ, Lin ME, Riddell J, Kravitz RL. Characterizing the adoption and experiences of users of artificial intelligence-generated health information in the United States: cross-sectional questionnaire study. J Med Internet Res 2024 Aug 14;26:e55138. [doi: 10.2196/55138] [Medline: 39141910]
6. Introducing ChatGPT Health. OpenAI. URL: https://openai.com/index/introducing-chatgpt-health [accessed 2026-01-08]
7. What is ChatGPT Health? OpenAI Help Center. URL: https://help.openai.com/en/articles/20001036-what-is-chatgpt-health [accessed 2026-01-10]
8. Google Health Enterprise. URL: https://fitbit.google/enterprise/blog/introducing-the-new-personal-health-coach-powered-by-gemini/ [accessed 2026-01-08]
9. OpenAI launches ChatGPT Health to connect medical records, wellness apps. Reuters. URL: https://www.reuters.com/business/healthcare-pharmaceuticals/openai-launches-chatgpt-health-connect-medical-records-wellness-apps-2026-01-07/ [accessed 2026-01-10]
10. Fitbit's personal health coach in public preview is here. Google. URL: https://blog.google/products-and-platforms/devices/fitbit/personal-health-coach-public-preview/ [accessed 2026-01-08]
11. Khasentino J, Belyaeva A, Liu X, et al. A personal health large language model for sleep and fitness coaching. Nat Med 2025 Oct;31(10):3394-3403. [doi: 10.1038/s41591-025-03888-0] [Medline: 40813712]
12. Subasri V, Krishnan A, Kore A, et al. Detecting and remediating harmful data shifts for the responsible deployment of clinical AI models. JAMA Netw Open 2025 Jun 2;8(6):e2513685. [doi: 10.1001/jamanetworkopen.2025.13685] [Medline: 40465297]
13. Cui H, Unell A, Chen B, et al. TIMER: temporal instruction modeling and evaluation for longitudinal clinical records. NPJ Digit Med 2025 Sep 26;8(1):577. [doi: 10.1038/s41746-025-01965-9] [Medline: 41006898]
14. Omar M, Sorin V, Collins JD, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Commun Med (Lond) 2025 Aug 2;5(1):330. [doi: 10.1038/s43856-025-01021-3] [Medline: 40753316]
15. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023 Aug 11;25:e48009. [doi: 10.2196/48009] [Medline: 37566454]
16. Beale SK, Cohen N, Secheli B, McIntire D, Kho KA. Comparing physician and artificial intelligence chatbot responses to posthysterectomy questions posted to a public social media forum. AJOG Glob Rep 2025 Aug;5(3):100553. [doi: 10.1016/j.xagr.2025.100553] [Medline: 40917303]
17. Kostick-Quenet KM, Gerke S. AI in the hands of imperfect users. NPJ Digit Med 2022 Dec 28;5(1):197. [doi: 10.1038/s41746-022-00737-z] [Medline: 36577851]
18. Ethics and governance of artificial intelligence for health: WHO guidance. : World Health Organization; 2021 URL: https://iris.who.int/server/api/core/bitstreams/f780d926-4ae3-42ce-a6d6-e898a5562621/content [accessed 2026-01-08]
19. Shekar S, Pataranutaporn P, Sarabu C, Cecchi GA, Maes P. People overtrust AI-generated medical advice despite low accuracy. NEJM AI 2025;2(6). [doi: 10.1056/AIoa2300015]
20. Moëll B, Sand Aronsson F. Harm reduction strategies for thoughtful use of large language models in the medical domain: perspectives for patients and clinicians. J Med Internet Res 2025 Jul 25;27:e75849. [doi: 10.2196/75849] [Medline: 40712151]
21. Will J, Gupta M, Zaretsky J, Dowlath A, Testa P, Feldman J. Enhancing the readability of online patient education materials using large language models: cross-sectional study. J Med Internet Res 2025 Jun 4;27:e69955. [doi: 10.2196/69955] [Medline: 40465378]

XSL•FO
RenderX

22. Zeng J, Zou X, Li S, et al. Assessing the role of the generative pretrained transformer (GPT) in Alzheimer's disease management: comparative study of neurologist- and artificial intelligence-generated responses. J Med Internet Res 2024 Oct 31;26:e51095. [doi: 10.2196/51095] [Medline: 39481104]
23. Rosen KL, Sui M, Heydari K, Enichen EJ, Kvedar JC. The perils of politeness: how large language models may amplify medical misinformation. NPJ Digit Med 2025 Nov 6;8(1):644. [doi: 10.1038/s41746-025-02135-7] [Medline: 41198821]
24. Serugunda HM, Jianquan O, Kasujja Namatovu H, et al. Using large language models for chronic disease management tasks: scoping review. JMIR Med Inform 2025 Sep 29;13:e66905. [doi: 10.2196/66905] [Medline: 41021927]
25. Hwang M, Zheng Y, Cho Y, Jiang Y. AI applications for chronic condition self-management: scoping review. J Med Internet Res 2025 Apr 8;27:e59632. [doi: 10.2196/59632] [Medline: 40198108]
26. Liu S, Wright AP, McCoy AB, Huang SS, Steitz B, Wright A. Detecting emergencies in patient portal messages using large language models and knowledge graph-based retrieval-augmented generation. J Am Med Inform Assoc 2025 Jun 1;32(6):1032-1039. [doi: 10.1093/jamia/ocaf059] [Medline: 40220286]
27. Schoonbeek RC, Workum JD, Schuit SCE, et al. Quality and efficiency of integrating customised large language model-generated summaries versus physician-written summaries: a validation study. BMJ Open 2025 Sep 4;15(9):e099301. [doi: 10.1136/bmjopen-2025-099301] [Medline: 40908007]
28. Armitage RC. Implications of large language models for clinical practice: ethical analysis through the principlism framework. J Eval Clin Pract 2025 Feb;31(1):e14250. [doi: 10.1111/jep.14250] [Medline: 39618089]
29. Artsi Y, Sorin V, Glicksberg BS, et al. Challenges of implementing LLMs in clinical practice: perspectives. J Clin Med 2025 Sep 1;14(17):6169. [doi: 10.3390/jcm14176169] [Medline: 40943929]
30. As'ad M, Faran N, Joharji H. AI-supported shared decision-making (AI-SDM): conceptual framework. JMIR AI 2025 Aug 7;4:e75866. [doi: 10.2196/75866] [Medline: 40773762]
31. Campos HDO, Wolfe D, Luan H, Sim I. Generative AI as third agent: large language models and the transformation of the clinician-patient relationship. J Particip Med 2025 Aug 11;17:e68146. [doi: 10.2196/68146] [Medline: 40826831]
32. Chen Y, Esmaeilzadeh P. Generative AI in medical practice: in-depth exploration of privacy and security challenges. J Med Internet Res 2024 Mar 8;26:e53008. [doi: 10.2196/53008] [Medline: 38457208]
33. Canali S, Schiaffonati V, Aliverti A. Challenges and recommendations for wearable devices in digital health: data quality, interoperability, health equity, fairness. PLoS Digit Health 2022 Oct;1(10):e0000104. [doi: 10.1371/journal.pdig.0000104] [Medline: 36812619]
34. Brückner S, Sadare O, Fesl S, Scheibe M, Lang C, Gilbert S. Attitudes of healthcare professionals and researchers toward wearable and app derived patient generated health data. NPJ Digit Med 2025 Mar 30;8(1):186. [doi: 10.1038/s41746-025-01568-4] [Medline: 40159538]
35. Kopka M, von Kalckreuth N, Feufel MA. Accuracy of online symptom assessment applications, large language models, and laypeople for self-triage decisions. NPJ Digit Med 2025 Mar 25;8(1):178. [doi: 10.1038/s41746-025-01566-6] [Medline: 40133390]
36. Wang C, Wang F, Li S, et al. Patient triage and guidance in emergency departments using large language models: multimetric study. J Med Internet Res 2025 May 15;27:e71613. [doi: 10.2196/71613] [Medline: 40374171]
37. Iles IA, Gillman AS, O'Connor LE, Ferrer RA, Klein WMP. Understanding responses to different types of conflicting information about cancer prevention. Soc Sci Med 2022 Oct;311:115292. [doi: 10.1016/j.socscimed.2022.115292] [Medline: 36063595]
38. Nagler RH, Vogel RI, Gollust SE, Yzer MC, Rothman AJ. Effects of prior exposure to conflicting health information on responses to subsequent unrelated health messages: results from a population-based longitudinal experiment. Ann Behav Med 2022 May 18;56(5):498-511. [doi: 10.1093/abm/kaab069] [Medline: 34398961]
39. Schreidah CM, Fahmy LM, Lapolla BA, Geskin LJ. Decision paralysis: recognition and patient-centered discourse. Dermatol Ther (Heidelb) 2023 May;13(5):1211-1217. [doi: 10.1007/s13555-023-00921-3] [Medline: 37046111]
40. Hölbling L, Maier S, Feuerriegel S. A meta-analysis of the persuasive power of large language models. Sci Rep 2025 Dec 12;15(1):43818. [doi: 10.1038/s41598-025-30783-y] [Medline: 41387793]
41. Valentine L, Hinton JDX, Bajaj K, et al. A meta-analysis of persuasive design, engagement, and efficacy in 92 RCTs of mental health apps. NPJ Digit Med 2025 Apr 29;8(1):229. [doi: 10.1038/s41746-025-01567-5] [Medline: 40301581]
42. Lang BH, Nyholm S, Blumenthal-Barby J. Responsibility gaps and black box healthcare AI: shared responsibilization as a solution. Digit Soc 2023 Dec;2(3):52. [doi: 10.1007/s44206-023-00073-z] [Medline: 38596344]
43. Cestonaro C, Delicati A, Marcante B, Caenazzo L, Tozzo P. Defining medical liability when artificial intelligence is applied on diagnostic algorithms: a systematic review. Front Med (Lausanne) 2023;10:1305756. [doi: 10.3389/fmed.2023.1305756] [Medline: 38089864]
44. Aaron DG, Robertson CT, King LP, Sage WM. A new legal standard for medical malpractice. JAMA 2025 Feb 26;333(13):1161-1165. [doi: 10.1001/jama.2025.0097] [Medline: 40009364]
45. Santoni de Sio F, Mecacci G. Four responsibility gaps with artificial intelligence: why they matter and how to address them. Philos Technol 2021;34(4):1057-1084. [doi: 10.1007/s13347-021-00450-x]
46. Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: a narrative review. Heliyon 2024 Feb 29;10(4):e26297. [doi: 10.1016/j.heliyon.2024.e26297] [Medline: 38384518]

XSL•FO
RenderX

47. Price WN II, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. JAMA 2019 Nov 12;322(18):1765-1766. [doi: 10.1001/jama.2019.15064] [Medline: 31584609]
48. Clinical decision support software: guidance for industry and food and drug administration staff. : U.S. Food and Drug Administration (FDA); 2026 URL: https://www.fda.gov/media/109618/download [accessed 2026-02-08]
49. Gerke S, Rezaeikhonakdar D. Privacy aspects of direct-to-consumer artificial intelligence/machine learning health apps. Intell Based Med 2022;6:100061. [doi: 10.1016/j.ibmed.2022.100061]
50. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors 2023 May 17;10:e47564. [doi: 10.2196/47564] [Medline: 37195756]
51. Khatiwada P, Yang B, Lin JC, Blobel B. Patient-generated health data (PGHD): understanding, requirements, challenges, and existing techniques for data security and privacy. J Pers Med 2024 Mar 3;14(3):282. [doi: 10.3390/jpm14030282] [Medline: 38541024]
52. Vallor S, Vierkant T. Find the gap: AI, responsible agency and vulnerability. Minds Mach (Dordr) 2024;34(3):20. [doi: 10.1007/s11023-024-09674-0] [Medline: 38855350]
53. Cao W, Zhang Q, Liu J, Liu S. From agents to governance: essential AI skills for clinicians in the large language model era. J Med Internet Res 2026 Jan 14;28:e86550. [doi: 10.2196/86550] [Medline: 41534067]
54. The Children's Online Privacy Protection Act (COPPA): what parents should know. Federal Trade Commission. URL: https://www.ftc.gov/news-events/topics/protecting-consumer-privacy-security/kids-privacy-coppa [accessed 2026-04-06]
55. Age appropriate design: a code of practice for online services. Information Commissioner's Office (ICO). URL: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/childrens-information/childrens-code-guidance-and-resources/age-appropriate-design-a-code-of-practice-for-online-services/ [accessed 2026-04-06]
56. Artificial Intelligence Risk Management Framework (AI RMF 1.0). : U.S. Department of Commerce; 2023 URL: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf [accessed 2026-06-10]
57. Regulatory considerations on artificial intelligence for health. : World Health Organization; 2023 URL: https://iris.who.int/server/api/core/bitstreams/ad62580f-540f-4e36-b957-e7f2946ae1fb/content [accessed 2026-01-08]
58. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: guidance for industry and food and drug administration staff. : U.S. Food and Drug Administration (FDA); 2024 URL: https://www.fda.gov/media/166704/download [accessed 2026-01-08]
59. Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. Presented at: FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency; Jan 29-31, 2019. [doi: 10.1145/3287560.3287596]
60. Imani H, Peng J, Mohseni P, Amirany A, El-Ghazawi T. QoS-efficient serving of multiple mixture-of-expert llms using partial runtime reconfiguration. Presented at: Proceedings of the 42nd International Conference on Machine Learning; Jul 13-19, 2025 URL: https://openreview.net/pdf?id=g45SHBmZLz [accessed 2026-06-10]
61. Imani H, Amirany A, El-Ghazawi T. Mixture of experts with mixture of precisions for tuning quality of service. Presented at: 2024 IEEE International Conference on Rebooting Computing (ICRC); Dec 16-17, 2024. [doi: 10.1109/ICRC64395.2024.10937027]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
HbA_{1c}: glycosylated hemoglobin
LLM: large language model
PH-LLM: personal health large language model

Edited by A Coristine; submitted 19.Jan.2026; peer-reviewed by A Amirany, Z Li; revised version received 31.May.2026; accepted 31.May.2026; published 25.Jun.2026.

Please cite as:
Xie W, Liu J, Liu S
Personal Health Large Language Models and the Negotiation of Medical Authority in Clinical Care: Opportunities, Risks, and Governance
J Med Internet Res 2026;28:e91727
URL: https://www.jmir.org/2026/1/e91727
doi:10.2196/91727

© Wenyi Xie, Jialin Liu, Siru Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License

XSL•FO
RenderX

(https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Lessons Learned From Implementing a Culturally Tailored Virtual Support Program for Asian American Breast Cancer Survivors With Pain and Depression

Seulgi Ryu[1*], MSN; Dongmi Kim[1*], PhD; Wonshik Chee[2*], PhD; Eun-Ok Im[1*], MPH, PhD

[1]School of Nursing, The University of Texas at Austin, 1710 Red River St, Austin, TX, United States

[2]College of Education, The University of Texas at Austin, Austin, TX, United States

[*]all authors contributed equally

Corresponding Author:
Seulgi Ryu, MSN
School of Nursing, The University of Texas at Austin, 1710 Red River St, Austin, TX, United States

Abstract

Breast cancer is the most diagnosed cancer among women in the United States, with a greater increase in incidence among Asian American women than among others. Despite having resided in the United States for decades, Asian American breast cancer survivors face unique cultural challenges, including stigma and reluctance to disclose their illness. We present a Viewpoint to discuss practical issues encountered in implementing a culturally tailored, technology-based program aimed at reducing pain and depressive symptoms in Asian American breast cancer survivors. This Viewpoint is a simple ancillary content analysis of meeting minutes and research diaries from a parent study of cancer pain management; here, we identify themes related to practical issues in conducting a culturally tailored, technology-based intervention among Asian American breast cancer survivors with pain and depressive symptoms. Key issues identified include (1) reluctance to disclose, (2) variability in engagement based on disease status, (3) the need for personalized support, and (4) intraethnic differences related to various factors. Suggestions for future research include (1) tailoring communication to participants' preferences while building trust over time, (2) applying flexible care strategies, (3) assessing individual needs early and adapting materials based on feedback, (4) matching participants with interventionists according to language and level of acculturation, and (5) respecting differences in cultural identities between ethnic subgroups. Addressing these challenges can improve the effectiveness of technology-based interventions for racial or ethnic minority groups.

Trial Registration: ClinicalTrials.gov NCT06085313; https://clinicaltrials.gov/study/NCT06085313

(J Med Internet Res 2026;28:e81187) doi:10.2196/81187

KEYWORDS

breast cancer; Asian American; technology-based intervention; pain; depressive symptoms

Introduction

Breast cancer remains the most commonly diagnosed cancer among women in the United States, with approximately 1 in 8 women (13%) expected to be diagnosed with breast cancer in their lifetime [1]. From 2012 to 2021, the incidence of breast cancer increased by 1% annually, with a notably steeper rise of 2.6% among Asian American women than among White, Black, and American Indian/Alaska Native women (1%) and among Hispanic women (1.6%) [1]. This increase among Asian American women coincided with the rapid growth of the Asian American population, which is projected to double between 2016 and 2060, primarily due to migration [2].

Among Asian American subgroups, breast cancer is most prevalent among Chinese (32%), Japanese (35%), and Korean (32%) American women [2]. Higher breast cancer rates are found among Asian immigrants than among US-born Asian people [3], indicating the role of immigration-related factors in health outcomes.

Despite long-term residence in the United States, Asian Americans often retain strong cultural values that shape their health care experiences. Many breast cancer survivors within Asian American communities report stigma, emotional suppression, and reluctance to disclose illness, owing to cultural expectations. Cultural factors and expectations have a significant role in gender-specific cancers. This is especially clear in many Asian cultures, where discussions relating to reproductive cancers, such as breast cancer, are often considered taboo or sensitive [4]. Furthermore, in the collective, family-centered nature of decision-making in Asian societies, women may delay or avoid care due to expectations of modesty, composure, and harmony, and may place greater emphasis on family responsibilities rather than on seeking care for themselves. This cultural factor also reflects discomfort in seeking medical

XSL•FO
RenderX

attention and help because of reluctance to reveal their cancer diagnosis [5,6]. Chinese Americans report feelings of social rejection [7], Korean Americans describe fear of disclosure [8,9], and Japanese Americans often avoid seeking help altogether because of beliefs about enduring hardship alone [10]. The burden of fulfilling multiple caregiving roles as mothers, wives, and daughters intensifies this emotional toll [9,11].

Cultural stigma surrounding illness often leads to concealing the side effects of treatments, such as alopecia [8], or even refusing chemotherapy to avoid looking ill [12]. These behaviors contribute to lower levels of emotional support and reduced self-care among Asian American breast cancer survivors [11]. Language barriers and unfamiliarity with the US health care system further compound such challenges [13,14]. In navigating care, recent immigrant Asian Americans often report a lack of accessible, culturally appropriate information [11,15]. Cultural stigma around cancer pain also hinders open communication with providers, limiting access to appropriate treatment [16]. A systematic review has shown that many Asian survivors normalize pain and illness, leading to the underutilization of pain management and worsening of symptoms [17].

To address these challenges, culturally and linguistically tailored interventions are increasingly emphasized [9,18]. Individuals' cultural norms significantly influence how they perceive and respond to pain, reinforcing the importance of culturally tailored strategies to address pain and provide emotional support [19]. However, despite the importance of pain and depressive symptoms in cancer care [8,11], few studies have addressed these issues within culturally tailored programs for Asian American breast cancer survivors [12]. In response, the research team has developed a culturally tailored, technology-based web application intervention for Chinese, Korean, and Japanese American breast cancer survivors. This program focuses on reducing pain and depressive symptoms. However, multiple challenges have emerged during implementation, and the operational framework is presented in Figure 1. This Viewpoint contributes to exploring issues encountered in implementing a culturally tailored, technology-based intervention for Asian American women with breast cancer and to providing recommendations that can inform future research and support research teams in planning, implementing, and evaluating technology-based intervention studies among racial or ethnic minorities.

Figure 1. Operational framework: from implementation challenges to culturally tailored strategies.

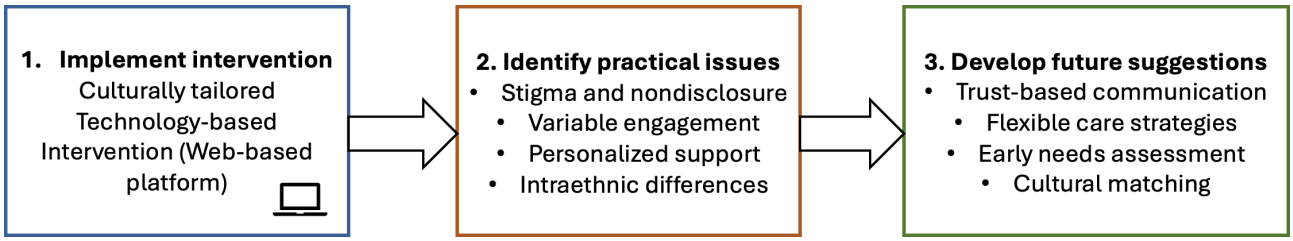

The Parent Study

Our ongoing parent study is a randomized controlled trial designed to improve pain and depressive symptoms among Chinese, Japanese, and Korean American breast cancer survivors in the United States. The 3 ethnic groups were selected because they have the highest prevalence rates of breast cancer among Asian women in the United States [2]. Participants were recruited nationwide, and enrolled participants were randomly assigned to either an intervention or a control group. The intervention group received a web-based coaching and support program with a bilingual interventionist of the same racial/ethnic background through weekly chat sessions, focusing on both pain and depressive symptom management for 12 weeks. The control group received the same web-based one-to-one format, with a primary focus on pain management. Importantly, the assigned interventionists identified themselves by their professional roles (eg, registered nurse or nurse practitioner) and provided support as trained members of the research team. Furthermore, all participants received a Fitbit (Google LLC) to encourage physical activity and symptom management.

Outcomes included cancer pain management and pain experiences, including self-reported pain; digital biomarkers (eg, heart rate); accompanying symptoms, such as depressive symptoms; and quality of life. Assessments were conducted at 3 points (pretest, 1 month posttest, and 3 months posttest) during the 3-month intervention period. The study was grounded in Bandura's social cognitive theory [20] and approved by the institutional review board of the University of Texas at Austin (parent study title: "Cancer Pain Management: A Technology-Based Intervention for Asian American Breast Cancer Survivors"; STUDY00004807).

The technology-based activities consist of three components culturally tailored for Asian American breast cancer survivors: (1) a study website providing online forums for group support where participants share their opinions and experiences based on the selected topics from the educational modules provided, and individual one-to-one coaching with an assigned interventionist; (2) online education modules available in Chinese (simplified and traditional), English, Japanese, and Korean based on topics related to breast cancer care and treatments; and (3) links to verified online sources, including the Centers for Disease Control and Prevention website. Online forums and chatting sessions were administered by bilingual interventionists, and educational modules included both general and culture-specific content for respective subethnic groups. The online resources include information from general and subethnic-specific health organizations and institutes. Anonymity was strictly protected throughout all online forums and chat sessions. All coaching and support interactions were conducted one-on-one to ensure participants had no access to other participants' personal information. Forum participation

XSL•FO
RenderX

was limited to participants of the same racial/ethnic background and was organized into small groups of approximately 10 participants, thereby further supporting privacy and a sense of safety. Additionally, all participants used self-selected nicknames, and no identifying information was visible to other participants.

Analysis

Our research team analyzed meeting minutes and research diaries from the parent study to identify themes related to practical issues in conducting a culturally tailored, technology-based intervention among Asian American breast cancer survivors with pain and depressive symptoms. The research memos were based on data from 51 participants (n=12, 23.5% Japanese, n=17, 33.3% Chinese, and n=22, 43.1% Korean) with an average age of 50.51 (SD 9.24; range 32 - 72) years. Of the participants, 12 (23.5%) communicated in English and 39 (76.5%) communicated in their own language. The average number of years spent in the United States among the 46 participants not born in the United States was 19.89 years. Only 1 (2%) participant, from the Chinese group, dropped out during the 1-year data collection period. Overall, 98% (50/51) of enrolled participants completed the intervention. Adherence to the individual coaching and support sessions was high, and intervention sessions were completed as scheduled. The research team members who conducted interventions with participants were registered nurses or nurse practitioners in the United States with clinical experience for patients with cancer for a range of years. Most of them held either a master's degree in nursing or a doctoral degree, or were doctoral students in nursing. This analysis was performed with a simple content analysis [21]. Using Microsoft Teams, the team held weekly discussions to address challenges and maintained records of online communications, such as meeting minutes. The research team communicated weekly for 1 year (15 - 16 memos per week × 4 weeks per month × 12 months=720 memos).

Research memos from weekly meeting minutes included issues raised during the intervention phases. The data included topics such as passive or active involvement; participants' treatment and side effects, including pain or depressive symptoms; how interventionists assisted them or emotional support provided; technical issues in navigating the website and Fitbit; and participants' reflections. Two authors with experience in qualitative coding independently reviewed and coded the data line by line. Differences were discussed, and the codes were refined to reflect their intended meanings. The themes were finalized through multiple group meetings with the research interventionists. Ultimately, categories representing practical issues in implementing coaching and support sessions were established on the basis of themes extracted from the codes. The interventionists' research memos, which describe their perspectives and share matched themes, are presented in Table 1.

XSL•FO
RenderX

Table . Issues and interventionist perspective quotes.

Issues	Interventionist perspectives (paraphrased)
Reluctance to disclose	• The participant initially showed a passive attitude, responding to questions with only "yes" or "no." In accordance with the intervention protocol, I attempted to build rapport by asking questions drawn from the baseline survey. [J[a]] • At the beginning of the sessions, the participant preferred to be asked essential questions (eg, pain or depressive symptom management, physical activity) rather than sharing details about daily life or other concerns. [K[b]] • This participant did not proactively share their emotions or any pain or depressive symptoms, so I initiated the conversation by asking how their day had been and what their mood was like that week. After 3 - 5 sessions, the participant became more active and open. [C[c]]
Variability in engagement based on disease status	• This participant has been a cancer survivor for over 10 years, so she had relatively fewer depressive symptoms, pain, or discomfort. I provided information on a healthy diet and exercise to maintain her condition. [K] • I encountered a participant who showed great appreciation for the chat sessions for providing useful information to improve side effects and sleep quality. [C and J]
Need for personalized support	• We discussed the side effects of hormone therapy due to her chronic fatigue and aches, and we will discuss bone density next week, as the participant is concerned about osteopenia. [K] • The participant underwent extreme mood fluctuations due to premenstrual symptoms after hormone therapy, so I provided support for her emotional health. [C and K]
Intraethnic differences related to various factors	• I have noticed that English-speaking Chinese are more fitted into the US mainstream culture and understand the medical terms and how health care systems work better than non-English-speaking Chinese with limited English proficiency. On the other hand, I found that some Chinese-speaking participants had difficulty accessing medical care, so as an interventionist, I provided guidance on navigating the US health care system. [C] • This English-speaking participant focused more on themselves rather than the other non-English-speaking participants who mostly centered on family care [C and K] • To reduce the pain, this participant used sauna, cupping therapy, acupuncture, or patches that she shipped from Korea. [K]

[a]J: Japanese interventionist.

[b]K: Korean interventionist.

[c]C: Chinese interventionist.

Observations

Reluctance to Disclose

Three key challenges emerged during the implementation of culturally tailored individual coaching and support sessions delivered by culturally matched interventionists. First, several participants demonstrated a passive communication style during their interactions with interventionists. Often reluctant to express emotions, they frequently responded to questions with minimal engagement, offering brief replies such as "yes" or "no." This behavior aligns with findings in previous literature, which indicate that Asian Americans tend to express emotions passively and to view not bothering others as a culturally virtuous action [11,18].

Second, some participants reported discomfort with the small talk that interventionists adopted to establish rapport [22]. These participants preferred to be asked only essential questions, thereby minimizing their participation in chat sessions. These communication patterns presented challenges for the interventionists, particularly in navigating conversations after addressing key concerns identified in the baseline survey.

Third, interventionists occasionally encountered participants who were highly proactive in managing their own well-being but reluctant to disclose specific details about their difficulties and especially cautious about sharing personal hardships. This is a common phenomenon among Chinese, Korean, and Japanese American participants, and building rapport with participants in this study required an average of 2 to 3 or 4 to 5 out of 12 sessions. Indeed, Asian American breast cancer survivors are often not open to sharing their distress, even with

XSL•FO
RenderX

their physicians [9,18]. Consequently, despite the anonymity and non-face-to-face nature of coaching sessions that we expected to enhance openness, the research team faced significant barriers to establishing effective interactions with participants.

Variability in Engagement Based on Disease Status

Participants' responses to the culturally tailored intervention varied substantially by disease status. Individuals who were not currently experiencing significant pain or depressive symptoms, were not undergoing active treatment, had received a noninvasive breast cancer diagnosis, or had low scores on the Patient Health Questionnaire-9 (which was administered prior to the chatting sessions) [23] generally demonstrated lower levels of engagement. Many of these participants reported fewer unmet needs related to their cancer experiences and thus participated more passively during the chat sessions. Likewise, participants who had completed surgery or primary treatment more than 10 years ago often reported minimal lingering pain or symptoms [24] and exhibited low involvement with the interventionists.

In contrast, participants who were newly diagnosed or actively undergoing treatment were significantly more engaged. They frequently sought timely information, emotional support, and strategies to manage symptoms associated with their diagnosis [18]. Similarly, participants who reported ongoing bodily pain or unresolved physical discomfort were also highly proactive during the chat sessions, often steering the conversations to explore effective pain management strategies. Others with poor sleep quality or heightened anxiety related to fear of cancer recurrence demonstrated strong engagement, initiating conversations and maintaining a high level of interaction during the chat sessions [25].

Necessity for Personalized Support

Our research protocol requires participants to complete weekly mandatory chatting sessions with culturally matched interventionists for 12 weeks. Following these sessions, participants are encouraged to engage in self-directed learning using educational modules and online resources provided by the research team. However, considerable variability was observed in participants' engagement with these materials, which appeared to be influenced by individual differences in baseline knowledge, personal needs, and interest levels.

In response, the research team has continually revised and updated the educational content to better accommodate participants' evolving and diverse needs. These personalized efforts are also integrated into the chatting sessions, allowing interventionists to deliver targeted, relevant information. For instance, one participant undergoing tamoxifen treatment reported experiencing severe premenstrual symptoms, which included anxiety, depressive mood, and extreme mood swings [26]. In this case, the interventionist provided tailored information addressing the definition of premenstrual symptoms, diagnostic criteria, and self-assessment methods. This individualized support not only enabled the participant to objectively recognize her symptoms but also empowered her to communicate more effectively with her health care provider to pursue appropriate treatment. Another example involved a participant undergoing active hormone therapy who was experiencing poor sleep quality. That participant focused her chatting sessions on discussing treatment-related side effects and tracking her sleep patterns using a Fitbit device provided by the research team.

Intraethnic Differences

Two key challenges were identified in implementing our culturally tailored interventions: differences in engagement between English-speaking and non-English-speaking participants, and differences in acculturation levels. Intraethnic differences in engagement were prominent between English-speaking and non-English-speaking Chinese, Korean, and Japanese participants. English-speaking participants were generally more acculturated and exhibited less stigma regarding the disclosure of symptoms and concerns related to their cancer diagnosis [27]. They tended to be more approachable and proactive in their interactions with interventionists and frequently took the lead in conversations. Many openly shared personal information and actively sought both medical and emotional support, which may reflect cultural norms shaped by their greater integration into mainstream American society [27]. This pattern was also observed during the research team's community outreach, in which English-speaking Chinese Americans were more willing to discuss the study than Chinese-speaking Chinese Americans. The interventionists observed that English-speaking participants demonstrated a stronger grasp of medical terminology and a clearer understanding of how the US health care system operates. They also tended to prioritize their individual well-being [28] over traditional familial roles, such as mother, wife, or daughter [11], and actively sought ways to manage their symptoms rather than silently endure them [10].

In contrast, non-English-speaking participants often encountered challenges in accessing medical care [11], in part because of limited English proficiency. These individuals required additional guidance from interventionists to navigate available health care resources [12]. Among the Korean-speaking participants, for example, role conflicts emerged frequently. These individuals, balancing the multiple responsibilities of being mothers, wives, daughters-in-law, and breast cancer survivors, reported significant internal struggles [12]. Many endured their pain quietly and were hesitant to share their difficulties with family members [11], reflecting culturally embedded values such as self-sacrifice and emotional stoicism [12].

Differences in participant engagement across acculturation levels were also observed. Participants who had immigrated more than 20 years ago or were born in the United States were generally more open to sharing personal experiences, demonstrated greater confidence in expressing their symptoms, and actively sought health-related information [29]. Some of those born in the United States adopted a hybrid approach to symptom management. For example, Chinese American participants frequently combined Western medical practices, such as the use of painkillers, with traditional remedies, such as acupuncture and herbal medicine [30]. Similarly, some

XSL•FO
RenderX

Korean American participants reported using culturally familiar complementary methods, such as sauna therapy, to alleviate bodily pain, improve circulation in their swollen arm due to breast cancer surgery, and enhance overall comfort [31]. These practices, rooted in the Korean wellness tradition, were often used alongside conventional medical treatment, reflecting a culturally embedded approach to symptom management. This tendency may be attributed to the continued influence of cultural values transmitted through family, even among English-speaking, American-born individuals [32].

Conversely, recent immigrants were typically less expressive about their emotional and physical challenges and tended to rely more heavily on traditional remedies to manage their pain or depressive symptoms. They also experienced greater difficulties in adjusting to the health care system and identifying culturally appropriate services in the United States. These individuals required more intensive guidance and culturally sensitive support from interventionists, including assistance in identifying culturally relevant support groups within their communities [11,13].

We also found differences in participant engagement across other background characteristics. Although participants may belong to the same broad ethnic category, differences in political identity, language, and social norms were observed between Taiwanese and mainland Chinese participants. Some Taiwanese participants expressed discomfort with being labeled as "Chinese" and rejected this identification, emphasizing the importance of acknowledging their distinct cultural and national identities. This finding indicated the need for a more nuanced approach in implementing the study.

Future Directions

These viewpoints have identified four practical issues in implementing a culturally tailored, technology-based program among Asian American breast cancer survivors: (1) reluctance to disclose, (2) variability in engagement based on disease status, (3) the necessity for personalized support, and (4) intraethnic differences in engagement related to multiple factors. These issues reflect the program's culturally tailored, non-face-to-face, technology-based design. These viewpoints call for the following recommendations for such research with racial or ethnic minorities in the future, especially for research with Asian American breast cancer survivors (Table 2).

Table . Practical issues and related suggestions for future research.

Practical issues	Suggestions
Reluctance to disclose	• Strive to accommodate participants' communication preferences • Understand that building trust with Asian American breast cancer survivors may require sustained effort over time • Take a more proactive stance in providing emotional support and offering relevant information
Variability in engagement based on disease status	• Adopt and implement flexible communication and care strategies
Necessity for personalized support	• Perform a thorough initial assessment to gain a comprehensive understanding of participants • Regularly update and diversify materials based on feedback from participants
Intraethnic differences	• Match participants with interventionists who have similar linguistic backgrounds and levels of acculturation • Address cultural identity sensitivities within each ethnic subgroup

First, interventionists should be trained to effectively support participants who may not engage interactively during chat sessions. Communication strategies can be tailored to participants' preferences, for example, by reducing the duration of chatting sessions and fostering a supportive environment that encourages openness at participants' own pace [33]. Beginning each session with simple, casual questions, such as "How are you feeling today?" or "How has your day been so far?" can help participants enter the conversation and reduce the potential pressure or discomfort associated with more direct inquiries.

For participants who demonstrate a passive communication style, interventionists can adopt a more proactive stance [33]. This may involve providing emotional support and gently introducing relevant information about treatment options, symptom management, and potential side effects [18]. It is crucial to recognize that trust-building with Asian American breast cancer survivors often requires sustained effort over time [34]. Therefore, interventionists should anticipate that it may take at least 2 to 3 sessions, and possibly up to 4 or 5, before participants are sufficiently open to share personal experiences or express emotional concerns in a web-based research study, according to the study's findings. Research teams should also emphasize the advantages of anonymous, non-face-to-face interventions, especially their flexibility regarding time, location, and perceived safety from stigma.

Although Asian American breast cancer survivors are often reluctant to express emotions openly because of cultural norms of emotional restraint [7], interventionists can still foster trust and establish meaningful therapeutic relationships by attending to participants' unmet emotional needs [18]. Prior research has shown that while many Asian American breast cancer survivors internalize feelings of depression or anxiety, they also maintain

XSL•FO
RenderX

a strong desire for emotional support [9,18]. Interventionists can play a vital role in enabling participants to develop effective coping strategies by gently encouraging them to share their feelings or experiences [9].

Second, given the variability in engagement based on disease status, interventionists need to adopt and apply flexible communication and care strategies. For participants who are newly diagnosed or undergoing active treatment, it is essential to prioritize discussions of treatment-related side effects, emotional distress, and associated concerns, and to provide in-depth, individualized coaching and coping strategies [35,36]. Interventionists can support these individuals through timely guidance offered in supportive, open, interactive chat sessions that directly address their current problems and lived experiences.

In contrast, breast cancer survivors who are asymptomatic or no longer undergoing active treatment may require a different approach. These individuals may not always vocalize their physical pain or psychological distress [37]. They may also be unaware of potential long-term side effects or may continue to experience residual symptoms, such as pain [12,24]. Among long-term breast cancer survivors, fatigue is common, so interventionists should be particularly attentive to identifying and addressing such symptoms [38,39]. For these participants, interventionists may focus more on general well-being, self-management, and psychosocial adjustment to survivorship [40]. Interventionists should also remain attentive to subtle cues because even those who report no current physical or emotional symptoms may still have unspoken concerns, such as fear of recurrence, uncertainty, body change, or identity loss [40,41]. Fostering a supportive, participant-centered environment during chat sessions is essential for addressing both explicit and implicit needs.

Third, we recommend enhancing personalization by conducting a thorough initial assessment to ascertain each participant's baseline knowledge, personal needs, and interests [42]. Our research team has provided tailored support to address participants' specific concerns, particularly those related to treatment side effects. This has included providing detailed information on alleviating symptoms, such as bodily aches or pain resulting from hormone therapy [7]; addressing medication-related side effects, such as osteoporosis [43]; guiding lifestyle modifications; providing reliable sources for symptom relief; and offering dietary recommendations and psychosocial support [35]. Regularly updating and diversifying educational content in response to participants' feedback and emerging needs can help ensure that learning remains engaging and addresses individual concerns [44]. A robust feedback system is also essential for gathering insights into the perceived effectiveness of educational materials and for identifying a broad range of participants' needs [44]. This feedback can inform continuous improvements, making the intervention more supportive and meaningful. By implementing these suggestions, researchers can create a more engaging, supportive intervention environment, ultimately improving health outcomes and participants' satisfaction.

Fourth, to accommodate the diverse needs arising from variations in language proficiency and levels of acculturation, participants should be matched with interventionists who share similar cultural and linguistic backgrounds [14]. This will facilitate greater mutual understanding and more effective communication and create a more supportive environment [13]. To account for participants' cultural backgrounds and English-language proficiency, our research team matched first-generation participants with interventionists who had substantial knowledge of participants' native cultural norms and lived experiences, or who had immigrated relatively recently. In contrast, second-generation participants who were more comfortable with speaking English were paired with interventionists who demonstrated a stronger understanding of US culture, had prior experience with American health care settings, and were more familiar with navigating the US health care system [45]. Overall, interventionists with more years of relevant experience were prioritized during matching to enhance the quality of interaction and cultural understanding. To ensure the quality of the matching, all interventionists, regardless of their length of stay in the United States, received standardized training and education prior to the intervention. The materials covered included US health care systems; manuals for the intervention and the forum; guidance for weekly sessions; and evidence-based educational materials to inform participants about treatments, cancer care, and culturally tailored information on pain and depressive symptom management. Furthermore, throughout the regular meetings, we shared our concerns and were consistently supervised by the principal investigators. As the study progressed, our matching strategy appeared to be effective, contributing to a more inclusive and culturally and linguistically appropriate environment for participants. Therefore, we recommend continual monitoring and evaluation of the outcomes of such detailed matching. Tracking indicators such as engagement levels, satisfaction, and perceived support can help researchers refine the matching strategy to further enhance the intervention's overall effectiveness.

Furthermore, given ethnic subgroups' sensitivity regarding cultural identity, participants should be matched with interventionists who share not only the same ethnicity but also similar cultural and linguistic backgrounds, even within broader ethnic categories. To address these concerns, our research team implemented strategies to ensure culturally matched and individualized coaching and support sessions. For example, Taiwanese participants were matched with interventionists of Taiwanese background; Chinese participants who had immigrated from mainland China or were US-born individuals with parents from mainland China, and who used simplified Chinese, were paired with interventionists who shared similar cultural experiences and language use [46]. Even during recruitment, the team developed both Traditional Chinese and Simplified Chinese flyers, following recommendations from Taiwanese interventionists, to appropriately reflect the linguistic and cultural differences between the Taiwanese and Chinese communities. This fostered a respectful, inclusive environment, ultimately enhancing the intervention's effectiveness.

In the era of proliferating artificial intelligence (AI)–assisted tools for cancer care, AI-powered chatbots have benefited

XSL•FO
RenderX

patients with breast cancer by being available 24/7, given the unpredictability of the disease and associated anxiety, and by offering interactive, personalized assistance with easy access [47]. To accommodate the suggestions of this research, AI chatbots could be integrated with human interventionists as a powerful tool for meeting individual needs or for helping individuals become more engaged, especially when stigma, emotional suppression, or reluctance to disclose breast cancer due to cultural norms may hinder help-seeking. Furthermore, they could provide tailored information to address individual needs, offer more dynamic, interactive environments for breast cancer survivors, and potentially support emotional well-being [47,48]. However, concerns have been raised about AI chatbots' ability to understand human emotions [49]. Human touch and interpersonal interactions could therefore complement AI chatbots, helping breast cancer survivors feel genuinely cared for. Furthermore, future studies need to investigate incorporating features for language support, including cultural nuances and even dialects into AI-powered chatbots to provide culturally tailored supportive care for Asian American breast cancer survivors.

Conclusions

On the basis of an ongoing study with Asian American breast cancer survivors, this Viewpoint discusses practical challenges in implementing a culturally tailored, technology-based web application program among Asian American breast cancer survivors with pain and depressive symptoms, and we provide several recommendations for such research in the future. We offer these recommendations to inform the planning and implementation of future culturally tailored, technology-based interventions among racial or ethnic minority groups, specifically Asian American breast cancer survivors.

Acknowledgments

The authors appreciate the efforts of the research team, including the dozen research assistants involved in the study.

Funding

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health (grant R33CA280979). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

Data sharing is not applicable to this paper, as no datasets were generated or analyzed during this study.

Authors' Contributions

Conceptualization: SR, DK, WC, EOI
Formal analysis: SR, DK, WC, EOI
Funding acquisition: WC, EOI
Investigation: SR, DK, WC, EOI
Methodology: SR, DK, WC, EOI
Resources: SR, DK, WC, EOI
Supervision: WC, EOI
Writing—original draft: SR
Writing—review and editing: SR, DK, WC, EOI

Conflicts of Interest

None declared.

References

1. Breast cancer facts & figures 2024–2025. American Cancer Society. 2024. URL: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/2024/breast-cancer-facts-and-figures-2024.pdf [accessed 2025-05-20]
2. Cancer facts & figures for Asian American, native Hawaiian, & other Pacific Islander people 2024–2026. American Cancer Society. 2024. URL: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/aanhpi-cancer-facts-and-figures/aanhpi-cff.pdf [accessed 2025-05-11]
3. Morey BN, Gee GC, von Ehrenstein OS, et al. Higher breast cancer risk among immigrant Asian American women than among US-born Asian American women. Prev Chronic Dis 2019 Feb 14;16:E20. [doi: 10.5888/pcd16.180221] [Medline: 30767860]
4. Mohd Tohit NF, Haque M. Forbidden conversations: a comprehensive exploration of taboos in sexual and reproductive health. Cureus 2024 Aug;16(8):e66723. [doi: 10.7759/cureus.66723] [Medline: 39139803]

XSL•FO
RenderX

5. Holroyd E, Twinn S, Adab P. Socio-cultural influences on Chinese women's attendance for cervical screening. J Adv Nurs 2004 Apr;46(1):42-52. [doi: 10.1111/j.1365-2648.2003.02964.x] [Medline: 15030441]
6. Chesla CA, Kwan CM, Chun KM, Stryker L. Gender differences in factors related to diabetes management in Chinese American immigrants. West J Nurs Res 2014 Oct;36(9):1074-1090. [doi: 10.1177/0193945914522718] [Medline: 24558055]
7. Wang L, Geng X, Ji L, Lu G, Lu Q. Treatment decision-making, family influences, and cultural influences of Chinese breast cancer survivors: a qualitative study using an expressive writing method. Support Care Cancer 2020 Jul;28(7):3259-3266. [doi: 10.1007/s00520-019-05161-w] [Medline: 31735999]
8. Kim IR, Cho J, Choi EK, et al. Perception, attitudes, preparedness and experience of chemotherapy-induced alopecia among breast cancer patients: a qualitative study. Asian Pac J Cancer Prev 2012;13(4):1383-1388. [doi: 10.7314/apjcp.2012.13.4.1383] [Medline: 22799336]
9. Lim JW, Gonzalez P, Wang-Letzkus MF, Baik O, Ashing-Giwa KT, Lim JW. Health behavior changes following breast cancer treatment: a qualitative comparison among Chinese American, Korean American, and Mexican American survivors. J Health Care Poor Underserved 2013 May;24(2):599-618. [doi: 10.1353/hpu.2013.0094] [Medline: 23728031]
10. Peregrina HN, Maglalang DD, Hwang J, Yoo GJ. A qualitative exploration of the continuum of help-seeking among Asian American breast cancer survivors. Soc Work Health Care 2023;62(10):345-358. [doi: 10.1080/00981389.2023.2244012] [Medline: 37532433]
11. Warmoth K, Cheung B, You J, Yeung NC, Lu Q. Exploring the social needs and challenges of Chinese American immigrant breast cancer survivors: a qualitative study using an expressive writing approach. Int J Behav Med 2017 Dec;24(6):827-835. [doi: 10.1007/s12529-017-9661-4] [Medline: 28585073]
12. Lee S, Chen L, Ma GX, Fang CY, Oh Y, Scully L. Challenges and needs of Chinese and Korean American breast cancer survivors: in-depth interviews. N Am J Med Sci (Boston) 2013 Feb 2;6(1):1-8. [doi: 10.7156/najms.2013.0601001] [Medline: 24019995]
13. Yoon H, Chatters L, Kao TS, Saint-Arnault D, Northouse L. Factors affecting quality of life for Korean American cancer survivors: an integrative review. Oncol Nurs Forum 2016 May 1;43(3):E132-E142. [doi: 10.1188/16.ONF.E132-E142] [Medline: 27105204]
14. Choi KS, Jun M, Anderson G. Experiences of Korean-American women with high risk hereditary breast cancer. Asian Oncol Nurs 2012;12(2):175. [doi: 10.5388/aon.2012.12.2.175]
15. Hwang DA, Lee A, Song JM, Han HR. Recruitment and retention strategies among racial and ethnic minorities in web-based intervention trials: retrospective qualitative analysis. J Med Internet Res 2021 Jul 12;23(7):e23959. [doi: 10.2196/23959] [Medline: 34255658]
16. Andres EB, Poco L, Balasubramanian I, et al. Responsive cancer care in Asia: stigma and pain must be acknowledged and addressed. Support Care Cancer 2025 Jan 30;33(2):128. [doi: 10.1007/s00520-025-09173-7] [Medline: 39883219]
17. Kwok W, Bhuvanakrishna T. The relationship between ethnicity and the pain experience of cancer patients: a systematic review. Indian J Palliat Care 2014 Sep;20(3):194-200. [doi: 10.4103/0973-1075.138391] [Medline: 25191006]
18. Kim JH, Ahmed K, Ahn SH, et al. Healthcare professionals' views of supportive care needs for Chinese, Korean, and Vietnamese Americans with metastatic cancer. Asian Am J Psychol 2024 Sep;15(3):233-245. [doi: 10.1037/aap0000329] [Medline: 39430037]
19. Miller ET, Abu-Alhaija DM. Cultural influences on pain perception and management. Pain Manag Nurs 2019 Jun;20(3):183-184. [doi: 10.1016/j.pmn.2019.04.006] [Medline: 31230670]
20. Bandura A. Social Foundations of Thought and Action: A Social Cognitive Theory: Prentice-Hall; 1986.
21. Weber RP. Basic Content Analysis, 2nd edition: Sage Publications; 1990.
22. Cui X. Small talk: a missing skill in the Chinese communicative repertoire. Aust Rev Appl Linguist 2015;38(1):3-23. [doi: 10.1075/aral.38.1.01cui]
23. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
24. Stein KD, Syrjala KL, Andrykowski MA. Physical and psychological long-term and late effects of cancer. Cancer 2008 Jun 1;112(11 Suppl):2577-2592. [doi: 10.1002/cncr.23448] [Medline: 18428205]
25. Breast cancer treatment timeline. Living Beyond Breast Cancer. 2022. URL: https://www.lbbc.org/about-breast-cancer/treatments/breast-cancer-treatment-timeline [accessed 2025-06-01]
26. Carmassi C, Cordone A, Dell'Oste V, et al. Prescribing tamoxifen in patients with mood disorders: a systematic review of potential antimanic versus depressive effects. J Clin Psychopharmacol 2021;41(4):450-460. [doi: 10.1097/JCP.0000000000001412] [Medline: 34166298]
27. Wright AA, Stieglitz H, Kupersztoch YM, et al. United States acculturation and cancer patients' end-of-life care. PLoS ONE 2013;8(3):e58663. [doi: 10.1371/journal.pone.0058663] [Medline: 23536809]
28. Yoo GJ, Aviv C, Levine EG, Ewing C, Au A. Emotion work: disclosing cancer. Support Care Cancer 2010 Feb;18(2):205-215. [doi: 10.1007/s00520-009-0646-y] [Medline: 19434430]
29. Wang JH, Adams I, Huang E, et al. Physical distress and cancer care experiences among Chinese-American and non-Hispanic White breast cancer survivors. Gynecol Oncol 2012 Mar;124(3):383-388. [doi: 10.1016/j.ygyno.2011.11.029] [Medline: 22115854]

XSL•FO
RenderX

30. Nielsen A, Dusek JA, Taylor-Swanson L, Tick H. Acupuncture therapy as an evidence-based nonpharmacologic strategy for comprehensive acute pain care: the Academic Consortium Pain Task Force White Paper update. Pain Med 2022 Aug 31;23(9):1582-1612. [doi: 10.1093/pm/pnac056] [Medline: 35380733]
31. Hannuksela ML, Ellahham S. Benefits and risks of sauna bathing. Am J Med 2001 Feb 1;110(2):118-126. [doi: 10.1016/s0002-9343(00)00671-9] [Medline: 11165553]
32. Zhang D. Language maintenance and language shift among Chinese immigrant parents and their second-generation children in the U.S. Biling Res J 2010 Apr 29;33(1):42-60. [doi: 10.1080/15235881003733258]
33. van Bruinessen IR, van Weel-Baumgarten EM, Gouw H, Zijlstra JM, Albada A, van Dulmen S. Barriers and facilitators to effective communication experienced by patients with malignant lymphoma at all stages after diagnosis. Psychooncology 2013 Dec;22(12):2807-2814. [doi: 10.1002/pon.3352] [Medline: 23897828]
34. Tam Ashing K, Padilla G, Tejero J, Kagawa-Singer M. Understanding the breast cancer experience of Asian American women. Psychooncology 2003;12(1):38-58. [doi: 10.1002/pon.632] [Medline: 12548647]
35. Moore HC. Breast cancer survivorship. Semin Oncol 2020 Aug;47(4):222-228. [doi: 10.1053/j.seminoncol.2020.05.004] [Medline: 32513419]
36. Monteiro-Guerra F, Signorelli GR, Rivera-Romero O, Dorronzoro-Zubiete E, Caulfield B. Breast cancer survivors' perspectives on motivational and personalization strategies in mobile app-based physical activity coaching interventions: qualitative study. JMIR Mhealth Uhealth 2020 Sep 21;8(9):e18867. [doi: 10.2196/18867] [Medline: 32955446]
37. Hsu T, Ennis M, Hood N, Graham M, Goodwin PJ. Quality of life in long-term breast cancer survivors. J Clin Oncol 2013 Oct 1;31(28):3540-3548. [doi: 10.1200/JCO.2012.48.1903] [Medline: 23980087]
38. Götze H, Friedrich M, Taubenheim S, Dietz A, Lordick F, Mehnert A. Depression and anxiety in long-term survivors 5 and 10 years after cancer diagnosis. Support Care Cancer 2020 Jan;28(1):211-220. [doi: 10.1007/s00520-019-04805-1] [Medline: 31001695]
39. Kaur N, Gupta A, Sharma AK, Jain A. Survivorship issues as determinants of quality of life after breast cancer treatment: report from a limited resource setting. Breast 2018 Oct;41:120-126. [doi: 10.1016/j.breast.2018.07.003] [Medline: 30029054]
40. R K, L S, P B, S G, R LP. Psychosocial experiences of breast cancer survivors: a meta-review. J Cancer Surviv 2024 Feb;18(1):84-123. [doi: 10.1007/s11764-023-01336-x] [Medline: 36854799]
41. Melhem SJ, Nabhani-Gebara S, Kayyali R. Latency of breast cancer stigma during survivorship and its influencing factors: a qualitative study. Front Oncol 2023;13:1075298. [doi: 10.3389/fonc.2023.1075298] [Medline: 36998442]
42. Cao W, Cho H. Culture in cancer survivorship interventions for Asian Americans: a systematic review and critical analyses. Asian Am J Psychol 2021 Mar;12(1):65-75. [doi: 10.1037/aap0000225] [Medline: 33959222]
43. Teissonnière M, Point M, Biver E, et al. Bone effects of anti-cancer treatments in 2024. Calcif Tissue Int 2025 Mar 27;116(1):54. [doi: 10.1007/s00223-025-01362-0] [Medline: 40146323]
44. de Jong K, Conijn JM, Gallagher RA, Reshetnikova AS, Heij M, Lutz MC. Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: a multilevel meta-analysis. Clin Psychol Rev 2021 Apr;85:102002. [doi: 10.1016/j.cpr.2021.102002] [Medline: 33721605]
45. Calvo R, Hawkins SS. Disparities in quality of healthcare of children from immigrant families in the US. Matern Child Health J 2015 Oct;19(10):2223-2232. [doi: 10.1007/s10995-015-1740-z] [Medline: 25987471]
46. Farsimadan F, Khan A, Draghi-Lorenz R. On ethnic matching: a review of the research and considerations for practice, training and policy. In: Lago C, editor. The Handbook of Transcultural Counselling & Psychotherapy: McGraw Hill; 2011:65-80.
47. Shaban MM, Osman YM, Mohamed NA, Shaban MM. Empowering breast cancer clients through AI chatbots: transforming knowledge and attitudes for enhanced nursing care. BMC Nurs 2025 Jul 29;24(1):994. [doi: 10.1186/s12912-025-03585-w] [Medline: 40731340]
48. Goumas G, Dardavesis TI, Syrigos K, Syrigos N, Simou E. Chatbots in cancer applications, advantages and disadvantages: all that glitters is not gold. J Pers Med 2024 Aug 19;14(8):877. [doi: 10.3390/jpm14080877] [Medline: 39202068]
49. Lawson McLean A, Hristidis V. Evidence-based analysis of AI chatbots in oncology patient education: implications for trust, perceived realness, and misinformation management. J Cancer Educ 2025 Aug;40(4):482-489. [doi: 10.1007/s13187-025-02592-4] [Medline: 39964607]

Abbreviations

AI: artificial intelligence

XSL•FO
RenderX

Edited by A Mavragani; submitted 23.Jul.2025; peer-reviewed by F Lui, J Giese-Davis; revised version received 30.Apr.2026; accepted 06.May.2026; published 03.Jun.2026.

Please cite as:
Ryu S, Kim D, Chee W, Im EO
Lessons Learned From Implementing a Culturally Tailored Virtual Support Program for Asian American Breast Cancer Survivors With Pain and Depression
J Med Internet Res 2026;28:e81187
URL: https://www.jmir.org/2026/1/e81187
doi:10.2196/81187

© Seulgi Ryu, Dongmi Kim, Wonshik Chee, Eun-Ok Im. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 3.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Unique Digital Images as Incentives in Clinical Trials: A Digital Shift Toward Meaningful Participation

Xavier Tadeo[1,2], PhD; Gyula Seres[1,2], PhD; Peter Wang[1,2,3], PhD; Yoann Sapanel[1,4], PhD; Alexandria Remus[1,2,3,5,6], PhD; Jasmine L Eyal[1]; Jason Kai-Wei Lee[1,2,5,6,7], PhD, FACSM; Simon Chesterman[8,9], DPhil, LLB; Julian Savulescu[10,11,12], PhD; John Eu-Li Wong[13,14], MBBS; Christopher L Asplund[1,2,3,15], PhD; R Brian Stone[16], MA; Benjamin C K Tee[2,17,18,19], PhD; Reza Shokri[20], PhD; Marlena Raczkowska[1,2], PhD; Li Ming Chong[1,2,3], PhD; Siong Peng Kwek[1,2]; Agata Blasiak[1,2,3], PhD; Dean Ho[1,2,3,4,21,22], PhD

[1]Institute for Digital Medicine (WisDM), Yong Loo Lin School of Medicine, National University of Singapore, 28 Medical Drive, #05-COR, Singapore, SG.01, Singapore
[10]Murdoch Children's Research Institute, Royal Children's Hospital Flemington Road, Parkville, Victoria, Australia
[11]Centre for Biomedical Ethics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[12]Oxford Uehiro Centre for Practical Ethics, Faculty of Philosophy, University of Oxford, Oxford, United Kingdom
[13]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[14]Department of Haematology-Oncology, National University Cancer Institute, National University Hospital Singapore, Singapore
[15]Department of Social Sciences, Yale-NUS College, Singapore
[16]Division of Industrial Design, College of Design and Engineering, National University of Singapore, Singapore
[17]Department of Materials Science and Engineering (MSE), National University of Singapore, Singapore
[18]Institute for Health Innovation & Technology (iHealthtech), National University of Singapore, Singapore
[19]Department of Electrical and Computer Engineering (ECE), National University of Singapore, Singapore
[2]The N.1 Institute for Health (N.1), National University of Singapore, Singapore, Singapore
[20]Department of Computer Science, School of Computing, National University of Singapore, Singapore
[21]Department of Pharmacology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[22]The Bia-Echo Asia Centre for Reproductive Longevity and Equality, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[3]Department of Biomedical Engineering, College of Design and Engineering, National University of Singapore, Singapore
[4]Singapore's Health District @ Queenstown, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[5]Heat Resilience & Performance Centre, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[6]Human Potential Translational Research Programme, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[7]Department of Physiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[8]Faculty of Law, National University of Singapore, Singapore
[9]NUS AI Institute, National University of Singapore, Singapore

Corresponding Author:
Xavier Tadeo, PhD
Institute for Digital Medicine (WisDM), Yong Loo Lin School of Medicine, National University of Singapore, 28 Medical Drive, #05-COR, Singapore, SG.01, Singapore

Abstract

Incentivization in clinical trial participation can be challenging, with many studies failing to meet recruitment or retention goals despite traditional compensation strategies. Digital health evolves, and with it, new approaches can emerge to engage participants meaningfully. We propose unique digital images as a novel, symbolic incentive for clinical trials. Digital images combine qualities such as personalization, ownership, and digital visibility, which may drive engagement more effectively than monetary rewards alone. In our illustrative study, participants complete artificial intelligence–personalized digital therapeutics training using CURATE.DTx, generating individualized learning trajectories. These are transformed into digital artworks and minted as nonfungible tokens given as a reward upon trial completion. This concept integrates gamification, personalization, and blockchain technology to support both intrinsic and extrinsic motivation. We explore the implications for decentralized health care, long-term behavior change, and participant recognition in the context of preventive medicine and longevity science. Our aim is to encourage research into the use of digital incentives to transform the trial participant experience and promote sustained engagement in health interventions.

(J Med Internet Res 2026;28:e88022) doi:10.2196/88022

XSL•FO
RenderX

KEYWORDS

digital health; clinical trial engagement; nonfungible tokens; NFT; behavior change; participant incentives; decentralized health care; artificial intelligence; AI

Introduction

Persistent challenges in recruitment and retention remain the greatest bottlenecks in conducting clinical trials [1]. Poor recruitment is the main source of discontinuation of trials, representing 9.9% of interruptions [2]. Monetary compensation and access to new therapies are often insufficient to entice individuals to participate in studies. An empirical analysis found that 19% of trials are either terminated for unsuccessful accrual or completed with less than 85% expected enrollment [3]. Extended timelines, increased costs, underpowered analyses, and trial failure are the main consequences. Collected data may be wasted, and public trust may be undermined as participants contribute their time without meaningful outcomes [4]. Treatments are delayed, and funding opportunities might be lost [5]. This raises the issue of whether traditional incentives are sufficient to sustain participant engagement. Communication techniques, while prevalent, do not show robust improvements. A Cochrane review of 68 trials summarized that timely reminders are a useful addition with limited gains, but even customized communication is insufficient to yield improvements over monetary and nonmonetary participation incentives [6]. These limitations suggest that recruitment and retention are fundamentally behavioral challenges, motivating exploration of alternative incentive structures beyond traditional monetary compensation.

We propose a digital age alternative that merges behavioral science and personalization: digital images as personalized trial incentives. A unique, ownable artifact is a durable extrinsic cue that can be internalized into identity (eg, “I am a longevity pioneer”), a psychological process known as self-signaling [7]. Experimental evidence shows that forgoing monetary rewards to earn a symbolic credential boosts pride and sustained prosocial behavior by reinforcing diagnostic self-inference of valued traits [8]. In the context of clinical trials, a singular digital artwork cocreated from personal performance data serves as an identity anchor, transforming transient participation into a lasting narrative of agency and contribution. Moving toward digital and personalized incentives is timely. A nearly universal adoption of electronic health records in hospitals in high-income countries means that new digital incentive models do not require stand-alone platforms. More importantly, the adoption of patient-facing infrastructure mainly through mobile apps is gaining traction [9,10].

Discussion of value, ownership, and participant agency within decentralized health care systems is warranted. To our knowledge, no precedent exists for using artwork as an incentive in clinical trials. This approach leverages the intrinsic value of art to motivate participation and can find application in everyday digital health use. Gamification in the form of digital rewards can improve adherence to health and wellness programs, changing behaviors that are central to healthy aging and longevity.

At its core, this viewpoint piece aims to explore the feasibility and rationale of using personalized digital artwork as a novel form of incentivization in clinical trials. We examine the conceptual foundation, behavioral implications, and potential applications. Our goal is to initiate a broader conversation with digital health researchers, clinicians working at the intersection of health care and technology, developers, and designers of digital health tools on how digital incentives may reshape participant engagement in the era of digital and decentralized health care. We outline the design of a forthcoming feasibility study (Avatar.DTx) as an illustrative example of trialing digital images as participation incentives in clinical research. Accordingly, no empirical results from this trial are reported in this manuscript.

Conventional Incentives in Clinical Trials

Monetary payment for participating in research has been documented for more than 100 years. In the United States, monetary and nonmonetary compensation to trial participants became mainstream in the 1920s and 1930s, including meals and transportation [11]. Today, most organizations pay for participation in a wide range of human subject research types, but there is still no consensus regarding the adequacy of it. Different commentators on trial compensation define it as wrong, coercive, fair, or necessary [12-14]. Additionally, there is minimal guidance for researchers and institutional review boards (IRBs) to decide whether and how much to pay participants. A long history of research participant compensation has not yielded consensus on the effectiveness, ethics, or design of monetary incentives in clinical trials [15].

Investigators may offer monetary payments to research participants as an incentive to take part, as fair compensation for their contribution, and/or to offset related financial burden. However, there is limited evidence on the necessity of monetary incentives to boost recruitment. Healthy individuals are attracted by money but also for other reasons, such as altruism, curiosity, and attention from physicians [16-18]. Patients affected by the researched condition are often interested in the interventional treatment that could benefit their health, although little research has been conducted on the influence of money in their decision to participate [19]. Money does not ensure more recruitment diversity either given the distrust of the research establishment by some ethnic groups [20]. The American Academy of Pediatrics recommended giving gifts to participating children instead of money following an appreciation model [21]. As we can see, the form and amount of payment can be adjusted depending on the type of research and participants, their contributions and vulnerabilities, local guidelines, and cultural norms. This choice matters: it shapes both recruitment volume and demographic composition. In US studies, the share offering any compensation varies widely (approximately 35% - 63% across trial phases and 10% - 84% across intervention types), underscoring heterogeneity by design and risk profile [22].

XSL•FO
RenderX

According to Bentley and Thacker [23], monetary payments appear to achieve their intended effect of increasing participants' willingness to enroll in research. However, payments could compromise the integrity of study results. First, monetary compensation is associated with increased motivation to participate in research independently of perceived risk, with higher payments further amplifying motivation. Second, higher monetary payments may create incentives for participants to withhold or conceal information regarding restricted activities. The authors conclude that further research is needed to inform sound policies on compensating research participants [23]. Similarly, Treweek et al [6] evaluated strategies to improve recruitment to and retention in randomized controlled trials—studies in which participants are randomly assigned to an intervention or control group to assess causal effects. They concluded that the evidence base remains limited, heterogeneous, and highly context dependent. Recruitment and retention challenges are multifactorial behavioral phenomena influenced by perceived burden, trust, logistical barriers, and expectations of benefit rather than purely by financial considerations. Monetary payments may improve follow-up rates, particularly for low-burden tasks such as survey completion; however, their impact on initial enrollment is inconsistent, and optimal structures, amounts, and delivery mechanisms remain poorly defined. The reviews also highlight the scarcity of rigorous evaluations of innovative engagement strategies, noting that most tested interventions predate contemporary digital and decentralized trial models [4].

Fixed payments may inadequately reflect the value and opportunity cost of participation that varies widely between individuals. Evidence from behavioral economics shows diminishing marginal returns beyond modest thresholds [24]. Monetary incentives often trigger an undesirable "crowding-out" effect where extrinsic rewards dilute intrinsic motivation, especially in altruistically framed activities such as research participation [25,26]. They rarely leave a lasting impression on the participant.

Clinical research increasingly shifts toward remote participation, digital therapeutics—software-driven, evidence-based interventions designed to prevent, manage, or treat medical conditions—and app-based data collection. Consequently, traditional retention tools may be insufficient to sustain long-term engagement. This gap underscores the need for novel, context-appropriate incentive mechanisms specifically designed for modern digital health trials.

Digital Art Incentives

Digital art paradigms such as scarcity, provenance, authorship, and symbolic value shape the meaning attributed to cultural objects, often independently of their direct monetary value [27]. In the context of clinical trials, digital art can function as an incentive, offering recognition, identity, and engagement. Moreover, art has always been relational in varying degrees, a factor of sociability and dialogue, adding another nontransactional layer to this incentive type [28]. Digital art has entered mainstream culture, with the art ecosystem recognizing its value. Among global art forms, digital art saw the biggest uplift in participation and spending in 2024 to 2025. Art collectors are growing comfortable with hybrid forms of exchange [29].

From the perspective of digital health trials, digital art as an incentive presents the advantage of high compatibility with digital infrastructures. High internet penetration and health care data digitalization can enable digital art incentives to be delivered, tracked, and experienced with minimal friction.

Tokenization technologies such as nonfungible tokens (NFTs) represent a convenient implementation platform for digital art–based incentives. Blockchains use fungible tokens that are not only currency but also utility mechanisms that align incentives among participants. Blockchain computation has an economic cost, and the token is the "fuel" that enables operations, tying economic value to network use. Users pay in the native token to execute code, and validators are compensated in the same token for processing transactions. Tokens are incentives for all stakeholders and allow the decentralized network to be sustained [30,31]. NFTs share this principle but, by virtue of being noninterchangeable (unique), can encode properties such as participation, ownership, or rights. NFTs are tools for structuring participation and verifying contributions in distributed systems [32,33].

In a digital health trial context, where incentive alignment is critical, NFTs enable incentive design beyond conventional monetary compensation. They are programmable and can encode participation, provenance, and conditional rights within a clinical research protocol. Blockchain digitally certifies ownership of unique digital assets, and NFT ownership paths are public. This allows behavioral patterns to be inferred (eg, a trial participant's relationship with their tokenized incentive, including engagement, perceived value, and longitudinal interaction).

Since their inception, NFT market activity has concentrated on a few assets and collections, which indicates that engagement and loyalty are bound to certain art ecosystems [34]. In the context of digital health trial incentives, this pattern reveals the significance of aligning incentives with participants' identities and preferences: sustained engagement is more likely to appear from personalized and meaningful incentives.

Monetary rewards may crowd out intrinsic motivation by framing participation as a transactional exchange: payment in exchange for participation. In contrast, a digital artwork, when derived from an individual's own performance or contribution, is expected to carry an intrinsic valuation. This is strengthened by its signaling value. The primary mechanism lies in the act of earning the artifact through task completion: its uniqueness and direct linkage to study participation confer meaning tied to achievement. This characteristic is lacking in other discussed conventional forms of participation incentives. The design allows participants to derive value from their actions ("I completed this training; therefore I am committed"), consistent with self-signaling theory [7]. Public visibility, such as the ability to share the artwork, may reinforce the identity signal, although it is not required for the mechanism to operate [8].

From our perspective, incentive art used in a clinical trial belongs to a distinct category from speculative digital art. Its

XSL•FO
RenderX

economic and behavioral properties are shaped by purpose and context rather than pecuniary motives. Tokenized or not, digital incentives are different from cash payments, with advantages and risks worth considering in trial incentive design.

Conventional, Digital, and Tokenized Incentive Mechanisms

Traditional, nontokenized, and tokenized incentives differ in technical implementation and behavioral properties (Table 1). Monetary payments are immediate, bounded in value, and aimed to offset the inconveniences of participation [35]. Their impact on recruitment and retention remains uncertain, and they carry a potential risk of coercion. Nontokenized digital incentives (eg, digital art, certificates, and badges) are nonmonetary and nontransferable and find value through recognition, identity signaling, and engagement [36]. Tokenized digital incentives are unique, programmable, and able to encode contribution, identity, and provenance. Their value is derived mainly from symbolic meaning and personalization [37].

Table . Comparison of conventional, digital, and tokenized incentive mechanisms in clinical trials.

Feature	Conventional cash incentives	Nontokenized digital incentives	Tokenized digital incentives
Fungible	Yes	Yes	No
Secondary market	No	No	Yes
Identity signaling	No	Possible	Possible
Long-term value	No	Contextual	Potential
Primary incentive logic	Compensation	Recognition	Recognition
Ethical risk	Coercion	Exclusion	Financialization

Nontokenized incentives leverage digital infrastructures without introducing market dynamics. They may have long-term value only in a personal, symbolic context. They introduce an ethical risk among participants with differing levels of digital literacy or cultural expectations, who may perceive symbolic or digital rewards as insufficient recognition of their contribution [38]. Tokenized incentives enable traceability and attribution with a secondary market. This feature introduces the risk of financialization, a shift in participant motivation from engagement to profits. However, incentive design (nontransferability and personalization) can mitigate such behavior.

The motivational spectrum that conventional, nontokenized, and tokenized digital incentives occupy is different. Money is immediate and simple but not meaningful. Nontokenized digital incentives emphasize symbolic value, recognition, and contextual meaning without direct economic relevance. Tokenized digital incentives have behavioral, social, and symbolic components, and they may require governance to avoid market-driven distortions. Due to their different incentive mechanisms, they can be complementary rather than substitutionary. Their effectiveness and ethical profile depend on their design and contextual deployment.

Digital Images as Incentives: Methodological Recommendations

Trials should include both quantitative and qualitative measures to assess the feasibility and impact of using digital art as a participation incentive in clinical research. Suggested outcomes include completion rates, participant retention, artwork retention or trading behavior, and qualitative feedback on acceptability, perceived value, and motivational impact collected through structured interviews or surveys. It is critical to document participants' prior familiarity with digital assets to help account for selection effects and interpret responses. When implementing incentives as NFTs, trials should also report technical implementation details, including wallet setup, data deidentification procedures, and blockchain transaction protocols. Potential biases such as digital literacy disparities, speculative motivations, and unequal access must be anticipated and addressed in the design. In addition, studies should document governance and life cycle considerations related to digital asset ownership, such as procedures for handling loss of access, transferability constraints, or ownership disputes as these factors may influence participant experience and long-term acceptability.

To reduce the risk of speculative participation, all materials—especially consent forms—should clearly state that the artwork is intended as a nonfinancial participation incentive with no guaranteed market value. This helps manage expectations and preserve intrinsic motivation. Unequal access can be minimized by avoiding the requirement of prior familiarity with digital art and/or blockchain technologies and using accessible platforms with minimal transaction costs. Clear communication about privacy safeguards, technical support availability, and poststudy access to the digital artwork is also important to ensure equitable participation and sustained trust. Together, these strategies ensure that participation reflects interest in the intervention rather than in external incentives alone.

An Illustrative Example: The Avatar.DTx Study

Our team has conceived a methodological study to assess the feasibility and acceptability of using images as rewards in a digital therapeutics trial for cognitive training. This pilot trial was approved by the National University of Singapore IRB (NUS-IRB-2023-183) and has been registered on ClinicalTrials.gov (NCT06647030). Recruitment has not started.

XSL•FO
RenderX

This planned trial is an example of how our proposal can be tested (Multimedia Appendix 1).

Trial participants will use CURATE.DTx, a tablet-ready, multitasking serious game successfully investigated in previous clinical trials (NCT04848935 and NCT06240897) [39,40]. CURATE.DTx leverages a small-data (ie, using individual longitudinal data rather than population-scale data), AI-derived platform (CURATE.AI) to dynamically personalize cognitive training by adjusting game intensity. Performance data from each participant are used to generate a unique learning trajectory profile. Once the profile is generated, it allows for real-time adaptation of training inputs to optimize individual learning outcomes. As a result, CURATE.DTx offers a promising method for enhancing cognitive training and potentially preventing age-related cognitive decline. Its interface is a modified version of the Multi-Attribute Task Battery (MATB), the Online MATB [41-43]. This MATB version is an update of the MATB flight deck simulator originally developed by the National Aeronautics and Space Administration and the US Air Force, which was previously leveraged by our research group [44-47].

Figure 1 shows an overview of our study. A total of 15 healthy participants will engage with CURATE.DTx as part of a structured cognitive training program. The platform will capture personalized performance data across a 10-week training period. Those data will then be transformed into a unique, generative artwork by a digital artist. Upon completion of the study, the artwork will be available as a high-resolution image file and, additionally, minted as an NFT and transferred to the participant's digital wallet. The images in this study will be integrated into a gamified, meaningful process. If participants do not complete the full training regimen, they will be compensated modestly per session in local currency, but they will not receive anything else. This structure is expected to reinforce engagement without coercion.

Figure 1. Overview of NFT minting and transfer in the planned Avatar.DTx study. Participants complete CURATE.DTx training, after which deidentified performance data are used to generate personalized digital artwork that is minted as an NFT and transferred to the participant's digital wallet. ETH: Ether; NFT: nonfungible token; DTx: digital therapeutics.

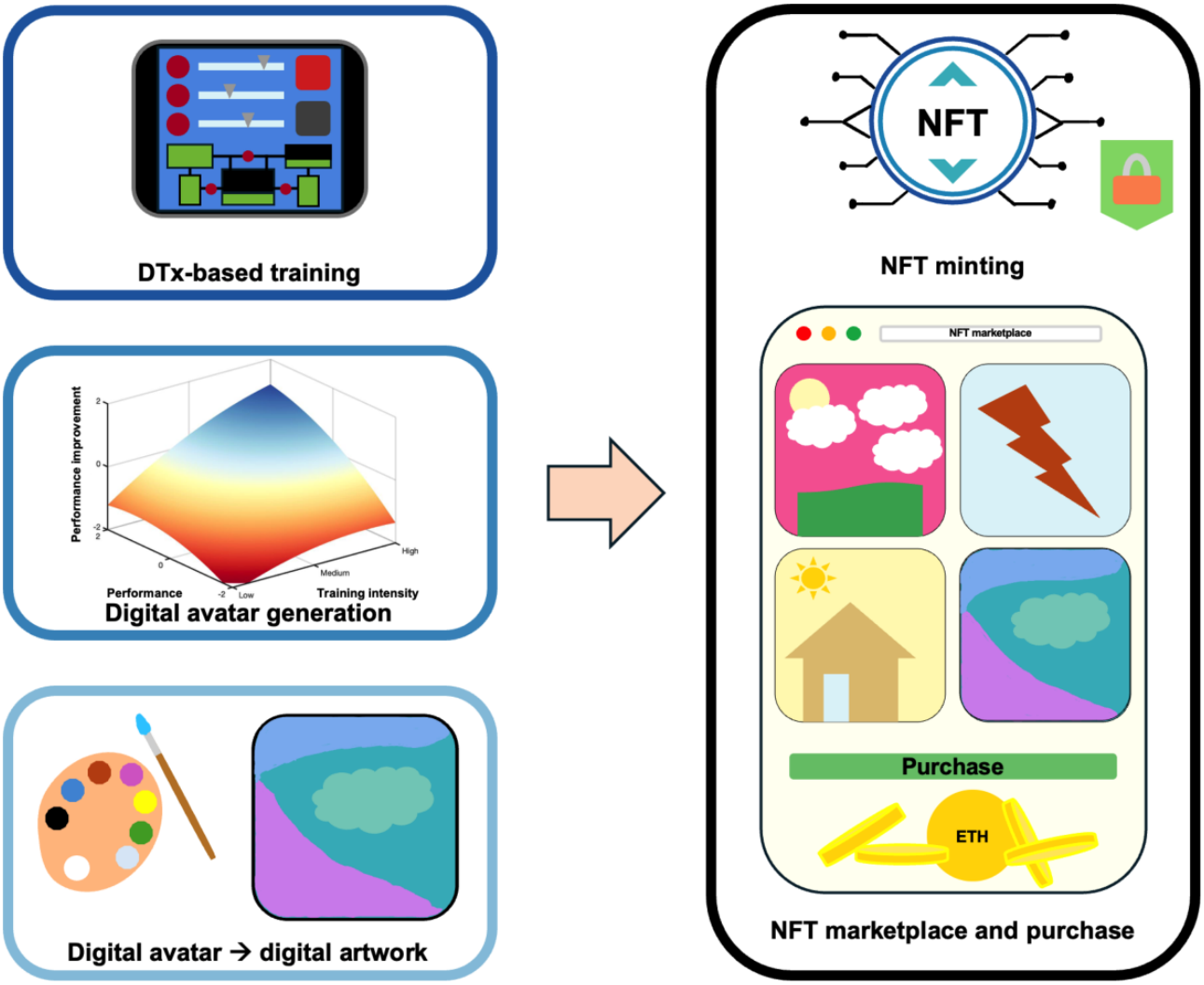

A relevant feature of these images is that they will be cocreated by a human artist and a generative artificial intelligence (AI) model. Although technically they are both unique digital art formats, human-made art is perceived as more valuable [36,48]. This way, participants also receive an otherwise paid service for free, further driving up the perceived value of an artwork.

The digital artist will use models such as generative adversarial networks and variational autoencoders. The algorithm will

XSL•FO
RenderX

introduce noise to ensure variability and prevent identical outputs, and the artist will shape the creation by selecting inputs, adjusting parameters, and polishing results. Noise will operate as a source of controlled randomness, enabling each run to yield distinct variations. For example, when transforming a user's profile into a digital painting, the artist may guide choices of color schemes, background motifs, or line styles while the algorithm determines their precise configuration (Figure 2).

Figure 2. Stylistic variants of personalized digital artworks illustrating outputs that could be generated under a human-guided artificial intelligence (AI) applied to abstracted performance features, including (A) abstract, (B) painterly, (C) generative lines (avatarlike), (D) minimal, (E) biometric inspired, and (F) symbolic milestone representations. In the planned Avatar.DTx trial, a human digital artist will intervene in the artwork creation process by guiding the AI (selecting inputs and adjusting aesthetic parameters); the images shown in this figure are fully AI generated and included solely for illustrative purposes. All images are derived from synthetic data and do not contain real participant information or clinical measurements.

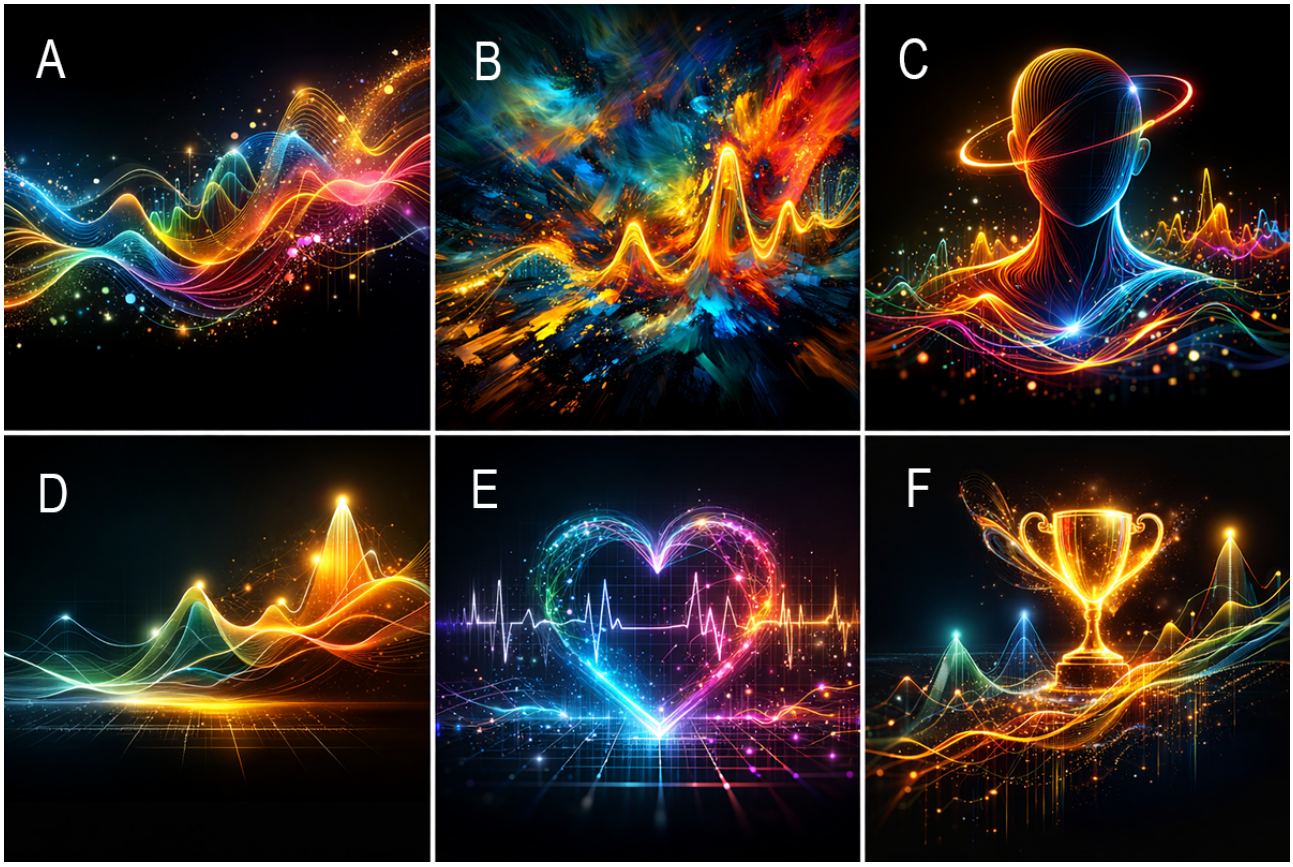

Our main goal is to assess the feasibility and acceptability of digital incentives as a means of encouraging participation. We will conduct in-depth interviews to understand perceptions of human art through digital means, AI-generated art, NFTs, ownership, value, and privacy. Technical and ethical safeguards will be put in place, including data anonymization and transparent consent.

Broader Implications

The potential effects of using digital image–based incentives go beyond recruitment and retention in clinical research, with implications for mechanisms that sustain adherence to health care interventions. At present, the role of AI and data sciences in the emerging fields of healthy aging and human longevity is being widely explored [49-52]. Longevity science and preventive health interventions increasingly rely on behavior change over long periods [53]. Maintaining long-term engagement in physical training, dietary interventions, and cognitive training activities is a major challenge [37]. Devising lasting commitment strategies becomes essential. Digital images, especially when linked to personalized feedback (eg, AI and small-data analytics generation of individualized health profiles), offer a new class of motivational tools for decentralized and digital health. Additionally, the integration of data-driven self-optimization with digital art incentives can foster a digital community where users share milestones and motivate each other [54].

Beyond their incentive value, the practical relevance of digital images in clinical contexts depends on whether they can be implemented within existing health care infrastructures without adding operational or technical burden. AI-based systems are increasingly used to generate or analyze clinically relevant visual outputs across medical domains, including applications in dental imaging and aesthetic planning [55,56]. Digital images can be incorporated into existing clinical systems without changing the core electronic health record. Rather than embedding new software directly, a separate incentive support tool can connect to it using widely adopted health data standards (eg, Fast Healthcare Interoperability Resources) [57,58]. In practical terms, this tool would mitigate privacy concerns by accessing only the limited information necessary to inform incentive selection. For example, appointment history leads to the generation of a recommendation about the type of incentive, when it should be offered, and how it should be communicated

XSL•FO
RenderX

(eg, via patient portal message or SMS text messaging). Recommendations can be delivered within the clinician's existing workflow so that they appear as part of routine care processes, such as after a missed appointment or during discharge planning. Digital images as incentives or tokens of appreciation can be seamlessly integrated in these systems.

This approach allows incentive personalization to operate as an extension of current systems. Safeguards must include maintaining records of recommendations, allowing clinicians to override automated suggestions when appropriate, limiting data use to what is strictly necessary, and monitoring for unintended disparities in how incentives are distributed across patient groups [59].

Digital artwork may also hold promise for underrepresented populations seeking greater personalization or for individuals drawn to new forms of digital participation [60].

It is worth noting that digital incentives are not necessarily exclusive; they can complement rather than replace traditional approaches to participant motivation such as financial compensation or altruism. Artistic forms of compensation add aesthetic and symbolic value. A culture of participant-centered trial conduct is emerging. Within that progression, rewarding participation with personalized digital art represents a coherent step.

Evidence shows that the value of artworks is driven by 2 factors: the intrinsic enjoyment dividend and additional trade value [61]. While we intend to present digital images as primarily nonspeculative incentives, enabling the trade of images in the form of NFTs may increase perceived value for some participants without being a necessary or primary feature of the incentive design. This can be facilitated by platforms such as OpenSea or Rarible, which make the process accessible to less technology-savvy participants.

Ethical and Practical Considerations

Using digital artworks as incentives in clinical research introduces new ethical and methodological questions. When using NFTs, ethical concerns about data visibility, even in abstract artistic forms, require serious consideration. Privacy-preserving algorithms, transparent consent procedures, and fallback contingencies should be implemented. During the consent process, participants should be informed that, if they do not complete the full training regimen, they will receive modest per-session compensation in local currency without eligibility for further incentives. This approach is designed to support engagement while avoiding undue inducement. No personal or clinical data should be stored on-chain. Only minimal metadata—such as a token ID and a link to the digital artwork hosted off-chain—should be recorded on the blockchain [62]. All sensitive participant data should be stored securely off-chain on encrypted institutional servers in accordance with local data governance policies.

To minimize the cognitive burden associated with blockchain interactions, researchers should manage e-wallets on behalf of participants and facilitate onboarding into digital art platforms. To account for the immutability of blockchain records, NFTs should only be minted and transferred after participants complete the study or intervention. If a participant withdraws prior to study completion, no token should be created or assigned. While there are many blockchains available, Ethereum is a sensible choice for NFT minting due to its mature infrastructure and wide support for ERC-721 standards [52]. Since Ethereum's transition to a proof-of-stake consensus mechanism in September 2022, validators secure the network by staking 32 ETH, eliminating the need for energy-intensive mining [63]. This upgrade has reduced Ethereum's energy use by over 99%, improving sustainability [64] and demonstrating that it is possible to choose a blockchain with relatively low environmental impact.

Intrinsic value is inherently random and not directly observable. It varies from person to person depending on individual taste, familiarity with digital art, and perceived emotional or aesthetic resonance [65]. Because of this subjectivity, predicting or standardizing the perceived worth of a digital artwork is difficult. Nonetheless, if the potential value of the incentive is disclosed in advance, it may introduce a selection effect among participants, attracting individuals motivated primarily by financial interest [66]. Such selection effects could bias recruitment and compromise the generalizability of findings. Furthermore, the perceived transactional nature of the incentive could risk undermining intrinsic motivation for behavior change [67]. To prevent a speculative motivation bias whereby participation incentives are dominated by a perception that an NFT is a tradable asset with a certain upside potential, the study protocol must clearly address this by avoiding investment framing, avoiding recruitment channels associated with trading, and controlling for the participants' motivation to join in an exit questionnaire.

The prospect of receiving a tradable asset, even if symbolic, may also exert undue influence among participants unfamiliar with the underlying technology [32,68]. This intersects with concerns of equity as individuals with lower digital literacy may face barriers to understanding, accessing, or benefiting from NFT-based incentives. Additionally, NFTs exist within markets characterized by speculative value and volatility, potentially leading to misconceptions about future worth or resale opportunities [69]. These risks are not inherent to digital incentives themselves and can be mitigated through incentive design and governance, including nontransferability, delayed minting until study completion, and IRB oversight to ensure proportionality and transparency.

Questions regarding data governance, intellectual property (IP), and platform dependence remain unresolved, especially in the context of health care research [70]. Careful data minimization; explicit IP disclosure; and the use of open, portable technical standards can substantially mitigate these risks.

Due to the absence of centralized regulation and standardized licensing practices, a commonly cited issue with NFTs is that buyers often obtain only a token reference (eg, a link to a file) rather than rights to the underlying IP, in this case, the artwork, potentially eroding public trust [71]. In clinical trials, IRB oversight may help mitigate these concerns by ensuring clear and ethical disclosure of compensation and related IP rights. In

XSL•FO
RenderX

this setup, the artwork serves to incentivize participants in clinical trials, and IRB oversight ensures proper disclosure of any IP-related compensation.

Recruitment has not started.

Conclusions

We encourage digital health and clinical research communities to explore incentive models that are more than transactional. Digital art and other forms of meaningful recognition can reformulate participant incentives for contributing to science. In clinical trials, where recruitment, retention, and sustained participant engagement directly determine scientific validity, incentive design remains a persistent and unresolved challenge.

This exploration is particularly timely. Longevity science and preventive health interventions hinge on adhering to diet; exercise; cognitive training; and, when necessary, medication. However, maintaining engagement over time is still a challenge. Digital incentives that combine personalization, perceived value, and public visibility provide a new tool to support both clinical research and long-term interventions.

Health care is shifting toward proactive and personalized models, and the mechanisms for motivating and rewarding individuals must evolve alongside it. Beyond the immediate benefits of recruitment or adherence, digital incentives raise broader questions about recognition and reciprocity in research. In the future, individuals may expect forms of engagement that acknowledge their role as active contributors rather than passive participants. When trial participants receive a unique artifact of their involvement, they are acknowledged. Health care is becoming more digital, allowing for new forms of engagement and reward. Digital images may be one such form.

Acknowledgments

The authors would like to thank Dr Smrithi Vijayakumar, Ryan Tan, and Jaron Cellist for their collaboration, informative insights, and suggestions. The authors declare the use of generative artificial intelligence (GenAI) in the writing process. According to the Generative Artificial Intelligence Delegation Taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing. The GenAI tool used was GPT-5 (OpenAI). Additionally, ChatGPT was used to generate illustrative examples of personalized digital artworks. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

The authors declared no financial support was received for this work.

Authors' Contributions

Conceptualization: DH, XT
Investigation: JLE, SC, JS
Methodology: AR, SPK, MR
Supervision: DH, JL, CA, AB
Visualization: PW, AR, XT
Writing—original draft: XT, GS, AR, YS, LMC, PW
Writing—review and editing: XT, GS, JKWL, RBS, BCKT, RS, JEL, CLA, AB
All authors approved the final version of the manuscript and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.
DH is a co-corresponding author and can be reached via email at biedh@nus.edu.sg.

Conflicts of Interest

DH is Cofounder and shareholder of KYAN Technologies. DH is an advisor to Elyx Life Pte Ltd, Prelude Health, and Foulplay Performance Nutrition. DH, AB, XT, PW, YS, CLA, AR, and MR are coinventors of pending patents pertaining to artificial intelligence–enabled therapy and data-driven behavior change. JS is a Bioethics Committee consultant for Bayer. Applicable conflicts of interest will be managed by an appropriate disclosure at all stages of study conduct: patient recruitment, data analysis, and results publication. There are no commercial relationships with nonfungible token platforms or artists. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Avatar.DTx methodological details.
[DOCX File, 22 KB - jmir_v28i1e88022_app1.docx]

References

1. Chaudhari N, Ravi R, Gogtay NJ, Thatte UM. Recruitment and retention of the participants in clinical trials: challenges and solutions. Perspect Clin Res 2020;11(2):64-69. [doi: 10.4103/picr.PICR_206_19] [Medline: 32670830]

XSL•FO
RenderX

2. Kasenda B, von Elm E, You J, et al. Prevalence, characteristics, and publication of discontinued randomized trials. JAMA 2014 Mar 12;311(10):1045-1051. [doi: 10.1001/jama.2014.1361] [Medline: 24618966]
3. Carlisle B, Kimmelman J, Ramsay T, MacKinnon N. Unsuccessful trial accrual and human subjects protections: an empirical analysis of recently closed trials. Clin Trials 2015 Feb;12(1):77-83. [doi: 10.1177/1740774514558307] [Medline: 25475878]
4. Halpern SD, Karlawish JH, Berlin JA. The continuing unethical conduct of underpowered clinical trials. JAMA 2002 Jul 17;288(3):358-362. [doi: 10.1001/jama.288.3.358] [Medline: 12117401]
5. Kitterman DR, Cheng SK, Dilts DM, Orwoll ES. The prevalence and economic impact of low-enrolling clinical studies at an academic medical center. Acad Med 2011 Nov;86(11):1360-1366. [doi: 10.1097/ACM.0b013e3182306440] [Medline: 21952064]
6. Treweek S, Pitkethly M, Cook J, et al. Strategies to improve recruitment to randomised trials. Cochrane Database Syst Rev 2018 Feb 22;2(2):MR000013. [doi: 10.1002/14651858.MR000013.pub6] [Medline: 29468635]
7. Bodner R, Prelec D. Self-signaling and diagnostic utility in everyday decision making. Psychol Econ Decis 2003;1(105) [FREE Full text]
8. Kirgios EL, Chang EH, Levine EE, Milkman KL, Kessler JB. Forgoing earned incentives to signal pure motives. Proc Natl Acad Sci U S A 2020 Jul 21;117(29):16891-16897. [doi: 10.1073/pnas.2000065117] [Medline: 32631987]
9. Adler-Milstein J, Holmgren AJ, Kralovec P, Worzala C, Searcy T, Patel V. Electronic health record adoption in US hospitals: the emergence of a digital “advanced use” divide. J Am Med Inform Assoc 2017 Nov 1;24(6):1142-1148. [doi: 10.1093/jamia/ocx080] [Medline: 29016973]
10. Andreadis K, Buderer N, Langford AT. Patients’ understanding of health information in online medical records and patient portals: analysis of the 2022 Health Information National Trends Survey. J Med Internet Res 2025 May 30;27:e62696. [doi: 10.2196/62696] [Medline: 40446288]
11. Lederer SE. Subjected to Science: Human Experimentation in America Before the Second World War: JHU Press; 1997.
12. McNeill P. Paying people to participate in research: why not? A response to Wilkinson and Moore. Bioethics 1997 Oct;11(5):390-396. [doi: 10.1111/1467-8519.00079]
13. Payment and reimbursement to research subjects. U.S. Food & Drug Administration. 2018. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/payment-and-reimbursement-research-subjects [accessed 2026-05-15]
14. Dunn LB, Gordon NE. Improving informed consent and enhancing recruitment for research by understanding economic behavior. JAMA 2005 Feb 2;293(5):609-612. [doi: 10.1001/jama.293.5.609] [Medline: 15687316]
15. Grady C. Payment of clinical research subjects. J Clin Invest 2005 Jul;115(7):1681-1687. [doi: 10.1172/JCI25694] [Medline: 16007244]
16. Bigorra J, Baños JE. Weight of financial reward in the decision by medical students and experienced healthy volunteers to participate in clinical trials. Eur J Clin Pharmacol 1990;38(5):443-446. [doi: 10.1007/BF02336681] [Medline: 2379528]
17. Farré M, Lamas X, Camí J. Sensation seeking amongst healthy volunteers participating in phase I clinical trials. Br J Clin Pharmacol 1995 Apr;39(4):405-409. [doi: 10.1111/j.1365-2125.1995.tb04469.x] [Medline: 7640147]
18. Cunny KA, Miller HW. Participation in clinical drug studies: motivations and barriers. Clin Ther 1994;16(2):273-282. [Medline: 8062322]
19. Halpern SD, Karlawish JH, Casarett D, Berlin JA, Asch DA. Empirical assessment of whether moderate payments are undue or unjust inducements for participation in clinical trials. Arch Intern Med 2004 Apr 12;164(7):801-803. [doi: 10.1001/archinte.164.7.801] [Medline: 15078651]
20. Corbie-Smith G, Thomas SB, St George DM. Distrust, race, and research. Arch Intern Med 2002 Nov 25;162(21):2458-2463. [doi: 10.1001/archinte.162.21.2458] [Medline: 12437405]
21. Shaddy RE, Denne SC, Committee on Drugs and Committee on Pediatric Research. Clinical report--guidelines for the ethical conduct of studies to evaluate drugs in pediatric populations. Pediatrics 2010 Apr;125(4):850-860. [doi: 10.1542/peds.2010-0082] [Medline: 20351010]
22. Kolbe A. Use of participant compensation in U.S. clinical research studies. Assistant Secretary for Planning and Evaluation. 2025. URL: https://aspe.hhs.gov/reports/compensation-clinical-research [accessed 2026-04-09]
23. Bentley JP, Thacker PG. The influence of risk and monetary payment on the research participation decision making process. J Med Ethics 2004 Jun;30(3):293-298. [doi: 10.1136/jme.2002.001594] [Medline: 15173366]
24. Halpern SD, Small DS, Troxel AB, et al. Effect of default options in advance directives on hospital-free days and care choices among seriously ill patients: a randomized clinical trial. JAMA Netw Open 2020 Mar 2;3(3):e201742. [doi: 10.1001/jamanetworkopen.2020.1742] [Medline: 32227179]
25. Newman GE, Jeremy Shen Y. The counterintuitive effects of thank-you gifts on charitable giving. J Econ Psychol 2012 Oct;33(5):973-983. [doi: 10.1016/j.joep.2012.05.002]
26. Gneezy U, Rustichini A. Pay enough or don’t pay at all. Q J Econ 1998;115(3):791-810. [doi: 10.1162/003355300554917]
27. Paul C. Digital Art: Thames & Hudson; 2008.
28. Bourriaud N. Relational Aesthetics: Les Presses Du Réel; 2002.
29. McAndrew C. The Art Basel and UBS Art Market Report. : The Art Market URL: https://www.artbasel.com/about/initiatives/the-art-market [accessed 2026-05-15]
30. Ethereum whitepaper. Ethereum. 2014. URL: https://ethereum.org/en/whitepaper/ [accessed 2026-05-15]

XSL•FO
RenderX

31. Cong LW, He Z. Blockchain disruption and smart contracts. National Bureau of Economic Research, NBER Working Paper 2018 [FREE Full text]
32. Entriken W, Shirley D, Evans J, Sachs N. ERC-721: non-fungible token standard. Ethereum Improvement Proposals. 2018. URL: https://eips.ethereum.org/EIPS/eip-721 [accessed 2026-05-15]
33. Catalini C, Gans JS. Some simple economics of the blockchain. SSRN. Preprint posted online on Nov 27, 2016. [doi: 10.2139/ssrn.2874598]
34. Nadini M, Alessandretti L, Di Giacinto F, Martino M, Aiello LM, Baronchelli A. Mapping the NFT revolution: market trends, trade networks, and visual features. Sci Rep 2021 Oct 22;11(1):20902. [doi: 10.1038/s41598-021-00053-8] [Medline: 34686678]
35. Abdelazeem B, Abbas KS, Amin MA, et al. The effectiveness of incentives for research participation: a systematic review and meta-analysis of randomized controlled trials. PLoS One 2022;17(4):e0267534. [doi: 10.1371/journal.pone.0267534] [Medline: 35452488]
36. Fortuna P, Modliński A, McNeill M. Creators matter. Perception and pricing of art made by human, cyborgs and humanoid robots. Empir Stud Arts 2022;41(2):331-351. [doi: 10.1177/02762374221143717]
37. Cappelen AW, Charness G, Ekström M, Gneezy U, Tungodden B. Exercise improves academic performance. J Polit Econ 2026;134(1):397-434. [doi: 10.1086/738251]
38. Nouri SS, Adler-Milstein J, Thao C, et al. Patient characteristics associated with objective measures of digital health tool use in the United States: a literature review. J Am Med Inform Assoc 2020 May 1;27(5):834-841. [doi: 10.1093/jamia/ocaa024] [Medline: 32364238]
39. Remus A, Tadeo X, Kai GN, et al. CURATE.AI COR-Tx platform as a digital therapy and digital diagnostic for cognitive function in patients with brain tumour postradiotherapy treatment: protocol for a prospective mixed-methods feasibility clinical trial. BMJ Open 2023 Oct 24;13(10):e077219. [doi: 10.1136/bmjopen-2023-077219] [Medline: 37879700]
40. Tan SB, Tan J, Raczkowska MN, et al. Digital game-based interventions for cognitive training in healthy adults and adults with cognitive impairment: protocol for a two-part systematic review and meta-analysis. BMJ Open 2023 May 4;13(5):e071059. [doi: 10.1136/bmjopen-2022-071059] [Medline: 37142320]
41. Kee T, Weiyan C, Blasiak A, et al. Harnessing CURATE.AI as a digital therapeutics platform by identifying N‐of‐1 learning trajectory profiles. Adv Therap 2019 Sep;2(9):1900023. [doi: 10.1002/adtp.201900023]
42. Pantuck AJ, Lee DK, Kee T, et al. Modulating BET bromodomain inhibitor ZEN‐3694 and enzalutamide combination dosing in a metastatic prostate cancer patient using CURATE.AI, an artificial intelligence platform. Adv Ther 2018;1(6). [doi: 10.1002/adtp.201800104]
43. Zarrinpar A, Lee DK, Silva A, et al. Individualizing liver transplant immunosuppression using a phenotypic personalized medicine platform. Sci Transl Med 2016 Apr 6;8(333):333ra49. [doi: 10.1126/scitranslmed.aac5954] [Medline: 27053773]
44. Kwek SP, Leong QY, Lee VV, et al. Exploring the general acceptability and user experience of a digital therapeutic for cognitive training in a Singaporean older adult population: qualitative study. JMIR Form Res 2025;9:e63568. [doi: 10.2196/63568]
45. Comstock JR, Arnegard RJ. The multi-attribute task battery for human operator workload and strategic behavior research. NTRS - NASA. 1992. URL: https://ntrs.nasa.gov/citations/19920007912 [accessed 2026-05-15]
46. Miller JW. U.S. air force-developed adaptation of the multi-attribute task battery for the assessment of human operator workload and strategic behavior. National Technical Reports Library. URL: https://ntrl.ntis.gov/NTRL/dashboard/searchResults/titleDetail/ADA537547.xhtml [accessed 2026-05-15]
47. Miller WD, Schmidt KD, Estepp JR, Bowers MA, Davis I. An updated version of the U.S. Air Force Multi-Attribute Task Battery (AF-MATB). Defense Technical Information Center. 2014. URL: https://apps.dtic.mil/sti/citations/ADA611870 [accessed 2026-05-15]
48. Bellaiche L, Shahi R, Turpin MH, et al. Humans versus AI: whether and why we prefer human-created compared to AI-created artwork. Cogn Res Princ Implic 2023 Jul 4;8(1):42. [doi: 10.1186/s41235-023-00499-6] [Medline: 37401999]
49. Wang P, Tadeo X, Chew HS, et al. N-of-1 health optimization: digital monitoring of biomarker dynamics to gamify adherence to metabolic switching. PNAS Nexus 2024;3(6):pgae214. [doi: 10.1093/pnasnexus/pgae214] [Medline: 38881838]
50. Moskalev A, Stambler I, Zhavoronkov A, editors. Artificial Intelligence for Healthy Longevity: Springer; 2023. [doi: 10.1007/978-3-031-35176-1]
51. Marino N, Putignano G, Cappilli S, et al. Towards AI-driven longevity research: an overview. Front Aging 2023;4:1057204. [doi: 10.3389/fragi.2023.1057204]
52. Michie S, Johnston M, Carey R. Behavior change technique. In: Gellman MD, editor. Encyclopedia of Behavioral Medicine: Springer; 2020:206-213. [doi: 10.1007/978-3-030-39903-0_1661]
53. Milkman K. How to Change: The Science of Getting from Where You Are to Where You Want to Be: Random House; 2021.
54. Maher CA, Lewis LK, Ferrar K, Marshall S, De Bourdeaudhuij I, Vandelanotte C. Are health behavior change interventions that use online social networks effective? A systematic review. J Med Internet Res 2014 Feb 14;16(2):e40. [doi: 10.2196/jmir.2952] [Medline: 24550083]

XSL•FO
RenderX

55. Chen K, Qiu L, Xie X, Bai Y. Dynasmile: video-based smile analysis software in orthodontics. SoftwareX 2025 Feb;29:102004. [doi: 10.1016/j.softx.2024.102004]
56. Khaled Ali GA, Abdelbadei KA, Essam M, et al. Smile.AI: a deep learning system for digital smile design. In: 2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS): IEEE; 2023. [doi: 10.1109/ICICIS58388.2023.10391194]
57. Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. The Fast Health Interoperability Resources (FHIR) standard: systematic literature review of implementations, applications, challenges and opportunities. JMIR Med Inform 2021 Jul 30;9(7):e21929. [doi: 10.2196/21929] [Medline: 34328424]
58. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016 Sep;23(5):899-908. [doi: 10.1093/jamia/ocv189] [Medline: 26911829]
59. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17. [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
60. Groom LL, Schoenthaler AM, Mann DM, Brody AA. Construction of the digital health equity-focused implementation research conceptual model - bridging the divide between equity-focused digital health and implementation research. PLOS Digit Health 2024 May;3(5):e0000509. [doi: 10.1371/journal.pdig.0000509] [Medline: 38776354]
61. Towse R. A Textbook of Cultural Economics, 2nd edition: Cambridge University Press; 2019.
62. Kuo TT, Kim HE, Ohno-Machado L. Blockchain distributed ledger technologies for biomedical and health care applications. J Am Med Inform Assoc 2017 Nov 1;24(6):1211-1220. [doi: 10.1093/jamia/ocx068] [Medline: 29016974]
63. Proof-of-stake (PoS). Ethereum. URL: https://ethereum.org/developers/docs/consensus-mechanisms/pos/ [accessed 2026-05-15]
64. Ethereum Energy Consumption Index. Digiconomist. URL: https://digiconomist.net/ethereum-energy-consumption [accessed 2026-05-15]
65. Hutter M, Throsby D. Beyond Price: Value in Culture, Economics, and the Arts: Cambridge University Press; 2008.
66. Gneezy U, Meier S, Rey-Biel P. When and why incentives (don't) work to modify behavior. J Econ Perspect 2011 Nov;25(4):191-210. [doi: 10.2307/41337236]
67. Promberger M, Marteau TM. When do financial incentives reduce intrinsic motivation? Comparing behaviors studied in psychological and economic literatures. Health Psychol 2013 Sep;32(9):950-957. [doi: 10.1037/a0032727] [Medline: 24001245]
68. Halpern SD, Chowdhury M, Bayes B, et al. Effectiveness and ethics of incentives for research participation: 2 randomized clinical trials. JAMA Intern Med 2021 Nov 1;181(11):1479-1488. [doi: 10.1001/jamainternmed.2021.5450] [Medline: 34542553]
69. Dowling M. Is non-fungible token pricing driven by cryptocurrencies? Finance Res Lett 2022 Jan;44:102097. [doi: 10.1016/j.frl.2021.102097]
70. van Haaften-Schick L, Whitaker A. From the artist's contract to the blockchain ledger: new forms of artists' funding using NFTs, fractional equity, and resale royalties. SSRN Electron J 2021. [doi: 10.2139/ssrn.3842210]
71. Xiao B. Copyright law and non-fungible tokens: experience from China. Int J Law Inf Technol 2023;30(4):444-471. [doi: 10.1093/ijlit/eaad007]

Abbreviations

AI: artificial intelligence
IP: intellectual property
IRB: institutional review board
MATB: Multi-Attribute Task Battery
NFT: nonfungible token

Edited by A Schwartz; submitted 18.Nov.2025; peer-reviewed by I Tosun, K Chen, Y Fan; revised version received 17.Apr.2026; accepted 20.Apr.2026; published 08.Jun.2026.

Please cite as:
Tadeo X, Seres G, Wang P, Sapanel Y, Remus A, Eyal JL, Lee JKW, Chesterman S, Savulescu J, Wong JEL, Asplund CL, Stone RB, Tee BCK, Shokri R, Raczkowska M, Chong LM, Kwek SP, Blasiak A, Ho D
Unique Digital Images as Incentives in Clinical Trials: A Digital Shift Toward Meaningful Participation
J Med Internet Res 2026;28:e88022
URL: https://www.jmir.org/2026/1/e88022
doi:10.2196/88022

XSL•FO
RenderX

© Xavier Tadeo, Gyula Seres, Peter Wang, Yoann Sapanel, Alexandria Remus, Jasmine L Eyal, Jason Kai-Wei Lee, Simon Chesterman, Julian Savulescu, John Eu-Li Wong, Christopher L Asplund, R Brian Stone, Benjamin C K Tee, Reza Shokri, Marlena Raczkowska, Li Ming Chong, Siong Peng Kwek, Agata Blasiak, Dean Ho. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Using Ultra-Abridged Individual Difference Scales for Personalization in Digital Mental Health to Improve Uptake, Engagement, and Experiences: Three-Tiered Decision Framework for Scale Shortening

Siu Kit Yeung, MPhil; Alan C Y Tong, PhD; Han Zhao, MEd; Winnie W S Mak, PhD

Department of Psychology, Chinese University of Hong Kong, Rm 333, Chen Kou Bun Building, Hong Kong, China (Hong Kong)

Corresponding Author:
Winnie W S Mak, PhD
Department of Psychology, Chinese University of Hong Kong, Rm 333, Chen Kou Bun Building, Hong Kong, China (Hong Kong)

Abstract

Given the diversity of human characteristics and experiences, personalization in nudges, messages, choice presentations, interventions, and overall product design has been increasingly adopted in digital health to promote engagement. Past studies on moderators and personalization in digital health and mental health services generally focused on demographic and symptom variables, with generally inconsistent findings or null findings. Cognitive, motivational, and decisional psychological attributes are largely overlooked. Psychology often uses long self-report scales to measure various psychological attributes. Although they are useful in tapping into individuals' psychological profiles, when applied in real-life, everyday settings to assess individual differences, people are most likely unwilling to complete them. With the pressing need to personalize digital health platforms to enhance uptake, retention, and engagement, ultrashort versions of these psychological scales may be considered to allow assessment of multiple attributes at the same time. Scale shortening can be achieved through regression analyses of each item, factor analyses, item response theory, ant colony optimization, and machine learning methods, with each method having advantages, disadvantages, and conditions required to make it suitable. To illustrate, we provided examples of regression analyses of each item and factor analyses, with potential implications for personalizing narrative versus research-based messages in digital mental health contexts. We present a 3-tiered decision framework for scale shortening method selection depending on goals and possible constraints, with guidelines on validation methods for ultrashort scales. Moving forward, more validation studies and field studies in digital health platforms are needed to evaluate the ecological validity, reliability, and generalizability of these methods, bearing in mind the limitations and conditions where such shortening methods may not work well. Researchers may compare the effectiveness and limitations of personalization using ultrashort scales with other commonly adopted personalization methods (eg, based on longer scales, behavioral data, and large language models). Ethical concerns need to be considered and mitigated carefully, respecting diverse preferences, informed choices, and the privacy of service users. Our viewpoint piece is primarily intended for digital mental health researchers and practitioners, but may also be informative for the fields of digital health and medicine as well as personalization (eg, personalized health care, personalized nudging, and message matching) more broadly, given the common goal of boosting uptake and engagement as well as improving service users' experiences.

(J Med Internet Res 2026;28:e80662) doi:10.2196/80662

KEYWORDS

personalization; scale shortening; digital mental health; individual differences; uptake; engagement

Scale Length as a Barrier to Personalization in Digital Mental Health

Human beings are heterogeneous, with diverse personalities, social-cognitive characteristics, decision-making tendencies, motivational orientations, perceptions, and experiences, etc [1]. Given that one size does not fit all, personalizing messages, choice presentations, nudges, user interface designs, and interventions are generally more effective in encouraging uptake of healthy behaviors, promoting engagement and health outcomes [1-12]. Such generally positive effects are supported or partially supported by numerous reviews, including meta-analyses in various contexts, including digital health environments [2,8,10-13]. Evidence of personalization advantage in digital mental health contexts is more mixed, with some studies showing the advantage of personalized services over nonpersonalized services [14,15]. However, there are studies demonstrating positive effects of personalized digital mental health services overall but not detecting differences between interventions with and without personalization components (see the recent reviews by Dandil and Kingston

XSL•FO
RenderX

[14] and Schaeuffele et al [15]). Pinpointing the causes of such discrepancies is premature, since null findings do not imply the absence of advantage and digital mental health studies comparing personalized conditions with nonpersonalized conditions are limited [14-16].

Past studies regarding personalization or moderation of mental health services have generally focused on demographic and disorder-focused variables (ie, depression symptom severity, anxiety severity, and clinical diagnosis [17,18]). However, the effects of personalizing based on these variables are inconsistent [17,18], and some service users are concerned about privacy issues with sharing such potentially sensitive mental health information [19,20]. Personalization based on motivation, decision tendencies, and cognitive styles may be promising as they have demonstrated positive effects in the broader health behavior and tailoring or matching literature (see reviews by Joyal-Desmarais et al [2] and Nikoloudakis et al [13]). These psychological attributes capture the variability of humans in ways that are inherently humanistic rather than medical. Many of these variables are closely associated with health decisions [9] and thus provide digital systems with valuable information to respond to the whole person in more nuanced and effective ways. With the availability of numerous validated individual-level psychological scales, measuring them for personalization of digital health services is potentially feasible [21,22]. One major challenge is the length of these scales, especially when added together. To tap into multiple psychological characteristics, users may need to complete dozens or even hundreds of items. Yet, most users are not willing to complete such long questionnaires [23]. Relatedly, many users of digital mental health platforms may be experiencing mental health challenges, which may affect their cognitive load and motivation [24]. Even though users may look for personalized care, completing long questionnaires may require substantial cognitive and motivational resources that add a burden to these already-distressed individuals, possibly leading to abandonment of the platform or service altogether [24,25].

Ultrabrief Scales in Digital Mental Health

To tackle this practical obstacle, we propose shortening psychological scales significantly to an ultrabrief version (ie, 2 - 4 items [13,14,16]), with empirical steps to retain acceptable reliability and validity, so that personalization can be made feasible without overburdening users [15]. In this viewpoint piece, we aim to (1) introduce and recommend various scale shortening methods for personalization purposes in digital health with examples (refer to “Item Selection and Scale Shortening” as well as “Selecting Items for Personalization” sections), (2) discuss validity, reliability, and ethical considerations (refer to “Validity and Reliability Considerations” as well as “Ethical Considerations” sections), and (3) propose a decision framework that takes account into purposes, execution considerations, and constraints while maintaining validity and reliability as well as examining real-world applicability and effects (refer to “Recommendations on Practical Framework” section). We hope this piece will drive more research studies on ultrashort scales for personalization, facilitating applicability and improving user experiences in personalization. We consider digital mental health researchers and practitioners as the primary audience of this piece, as personalization is a popular engagement-boosting method in digital mental health spaces [16]. Given widespread implementations of personalization [4], this piece is also potentially relevant for digital health researchers and practitioners in other non–mental health domains as well as researchers and practitioners who are interested in personalization of health-related services, recommendations, messages, choice presentations, and nudges.

For decades, the use of very short scales to measure health-related constructs has been widely adopted in public health contexts. While there are many studies on scale shortening in the broader psychology and health-medical literature, work validating or applying brief scales for personalization in digital mental health contexts remains scarce. One exemplar is the shortening of Patient Health Questionnaire-9 (PHQ-9) [26] and Generalized Anxiety Disorder-7 (GAD-7) [27] into Patient Health Questionnaire-4 (PHQ-4) [28] that incorporated 2 items from each [29,30]. Hlynsson et al [30] found that Patient Health Questionnaire-2 (PHQ-2) and Generalized Anxiety Disorder-2 (GAD-2) are effective at identifying people with significant symptoms (with satisfactory discriminative validity) throughout different stages, including pretreatment, midtreatment, and posttreatment phases. When compared to their full versions, the short forms demonstrated good reliability and validity, especially at posttreatment for GAD-2. The PHQ-4 has now been used in digital platforms and mobile apps as a screening measure and assists in content, service, or platform recommendations or personalizations [31,32]. Given that the anxiety and depressive symptoms are highly prevalent in the general population, the utility of such short scales can enable quick screening. If non–symptom-focused psychological measurements can also be abridged with sufficient reliability and validity, they can offer a vast array of individual attributes for personalization.

Item Selection and Scale Shortening

Currently, factor analyses (FAs), item response theory (IRT), ant colony optimization (ACO), and machine learning (ML)–based methods [33-35] are established ways to shorten scales. FA is commonly adopted for scale shortening and may involve selecting items based on higher factor loadings [36,37] and tends to require smaller sample sizes than other methods. The method using IRT selects items based on more detailed information, including measurement precision, item information, and item discrimination, and can maintain precise estimation of the score of the original scale, but requires a larger sample size than FAs [38-40]. ACO selects the superior solution among multiple possible solutions but performs better with 500 or more participants in general [41]. ML methods, notably genetic algorithms, also select the short form with better prediction performances after considering many combinations, but also need larger sample sizes, ideally thousands of participants [33]. Each method has its pros, cons, and practical challenges, which are summarized in Table 1. Researchers may choose the appropriate statistical methods based on the sample size, number

XSL•FO
RenderX

of factors, and technical knowledge or experience with relevant methods.

XSL•FO
RenderX

Table . Explanations, advantages, disadvantages or barriers, and suggested software or programs, packages, and/or modules of different scale shortening methods[a].

Methods	Explanation	Advantages	Disadvantages or barriers	Suggested software or programs, packages, and libraries
Factor analyses	• Selecting items with the highest factor loadings, as shown in EFAs[b] [36,37].	• Relatively feasible and easy for most researchers to implement while not requiring very large sample sizes, with many programs and packages available. • Similar goodness-of-fit indices to more complex methods such as ACO[c] when there is one factor [41].	• When there are 2 factors and there is minor model misspecification, poor model-data fit can occur, and FAs[d] perform worse than ACO [41].	• SPSS is a program that requires payment or an institutional subscription, with both point-and--click and syntax functions. The basic SPSS program can run EFA [42], whereas running CFA[e] requires the Amos add-on [43]. • Jamovi [44] is an open-source program that allows point-and-click and syntax, with EFA and CFA functions. • R (an open-source program) Package *lavaan* provides more options (such as estimation methods and multi-group CFA) for conducting EFAs and CFAs, with more information reported (such as about goodness of fit) [45]. • R Package *psych* can help conduct EFAs and internal consistency tests, with more options available [46].

XSL•FO
RenderX

Methods	Explanation	Advantages	Disadvantages or barriers	Suggested software or programs, packages, and libraries
Item response theory	• Selection of items based on detailed information, including measurement precision, item information, and item discrimination of each item [38].	• Provides more detailed information for each item, but not simply a single reliability score for the test in FAs [38]. • Ensures sufficient measurement precision across the whole range of the trait [38].	• More mathematically challenging, and many researchers may not have the relevant fundamentals for implementing IRT[f]. • Requires a large sample size, generally 500 participants and ideally larger, which may not be feasible for some researchers and in some contexts given resource constraints [47]	• IRTPRO is a standalone program designed specifically for various IRT tests, for both unidimensional and multidimensional models with both dichotomous and polytomous data [48]. • MPlus is a program that requires entering syntax, with the advantage of integrating IRT with other data analytical methods such as factor analytical methods, mixture modeling and multilevel mixed modeling [49]. • R Package *mirt* was signed for multidimensional IRT, with both confirmatory and exploratory modeling [50]. • R Package *TAM* is particularly suitable for running IRT tests for large-scale datasets, including multidimensional, multilevel, and multigroup modeling [51,52]. • R Package *ltm* is designed for unidimensional IRT but not multidimensional IRT, and may be more suitable for less complex models [53,54].
Ant colony optimization	• Involves metaheuristic algorithms to select a better solution among multiple possible solutions, with the goal of maximizing model fit [41,55]	• Can better select items with higher data-model fit by considering multiple variations and selecting the superior solution [55]. • Outperforms FAs in goodness of fit when there are 2 factors and model misspecification is present [41].	• Technically more challenging to implement, especially for researchers without computing fundamentals. • Very limited options available that are specifically designed for ACO in short-form development (R Package *ShortForm*). • Generally requires a large sample of at least 500 participants, with the sample split into training and validation samples [40]. Such a large sample size may be less implementable in some contexts.	• R Package *ShortForm* can facilitate short form creation with the goal of maximizing validity with criteria decided by the researcher [35]. • HeuristicLab is a standalone program with ACO functions, requiring users to create fitness or validity criteria for item selection [56].
Machine learning–based methods				

XSL•FO
RenderX

Methods	Explanation	Advantages	Disadvantages or barriers	Suggested software or programs, packages, and libraries
	• Selection of items aiming to predict participants' sum score or mean score in the original scale, with both training dataset and cross-validation dataset [33]. • Considers many possible combinations of short forms in selecting the short form with better prediction performances, through methods such as genetic algorithm [33].	• Can better maintain predictive validity [57]. • Can predict participants' scores in the original scale through a small number of items accurately [33].	• To achieve high prediction performance, large sample sizes, ideally thousands of participants, are needed [33]. Such sample sizes may not be feasible for some studies. • Lack of consistency in items selected in bootstrapped samples [33]. • Requires higher proficiency and expertise in machine learning and programming, which is not common in the field of psychology.	• R Package *GA* offers general-purpose functions for genetic algorithms, but requires more procedures (such as creating functions) by users to shorten the scales and may be more suitable for users with higher expertise in R [58]. • Python DEAP (Distributed Evolutionary Algorithms in Python) is a library in Python that consists of evolutionary algorithms functions, including genetic algorithms, which can be used for optimizing items to be selected [59]. • HeuristicLab allows customization of genetic algorithms, which can be used for scale shortening [56].
Item-level moderated regression analyses	• Conducting separate regression analyses with interaction between each item of the scale and the condition, possibly with alpha correction. • Item selection may be based on 2 - 4 items with the largest interaction regression coefficients.	• Technically simple to implement. • Helps select items with stronger moderation effects, which may be more relevant and efficacious for personalization in a specific context.	• The selected items based on (larger) regression coefficients may not capture the entire construct well. • The items selected may be sample dependent and may lack generalizability to other samples.	• SPSS allows various forms of regression analyses [42], with the Process Macro add-on for simple slope analyses [60]. • Jamovi consists of functions for both logistic and linear regression as well as simple slope analyses [44]. • R glm and lm functions (base R) can be adopted for logistic regression and linear regression, respectively.

[a]The above table provides information on explanations, pros and cons, as well as suggested programs, packages, or libraries of different scale shortening methods. This table is not meant to be a comprehensive review of various scale shortening methods but to encourage scale shortening for the purposes of personalization in digital health contexts. The table only provides a very brief summary of various methods. For more details regarding various scale shortening methods, see [33,34,41].

[b]EFA: exploratory factor analysis.

[c]ACO: ant colony optimization.

[d]FA: factor analysis.

[e]CFA: confirmatory factor analysis.

[f]IRT: item response theory.

Selecting Items for Personalization

Researchers may conduct moderation analyses with the abridged questionnaires to examine whether the individual attributes moderate the effects of conditions (eg, messages and therapies) on outcomes such as uptake, engagement, decision, satisfaction, or well-being. If moderation and simple slope effects with the short form are shown to be practically meaningful with cross-over interaction effects (ie, condition A results in better outcomes than condition B for people higher in a psychological attribute whereas condition B leads to better outcomes than condition A for people lower in a psychological attribute, see the example below [61]), further work can examine whether the short form can be applied in a digital platform for

XSL•FO
RenderX

personalization. Apart from selecting items based on primary data, researchers are also encouraged to share data openly on repositories such as the Open Science Framework, so that others can conduct secondary analyses to shorten the scales for implementation and/or future studies, perhaps through combining samples from multiple shared datasets, with methods that require larger sample sizes, such as IRT and ML methods [33,38-40].

One overlooked but particularly helpful method for selecting items relevant for personalization is to perform moderation analyses on an item level (that has shown evidence for moderation on a scale level) and select the items (preferably 2 to 4 items) with the strongest moderation effects. This method ensures the items selected work well in personalization in a particular context, while other scale shortening methods may not always achieve this goal. In the following, we will explain with an example of narrative versus research messaging, with moderation of thinking and decision styles on the breathing exercise decision.

In a study of 166 young adults, we compared the effects of narrative versus research evidence messages in encouraging stress management, including playing a breathing exercise video and investigated individual difference moderators while conducting exploratory analyses to shorten scales [61]. Individual difference moderators included Experiential Thinking Style (ETS)—indicating tendencies to focus on past experiences, emotions, and narratives in making judgments and decisions [62], and decision style—categorized as rational decision style (deliberate and systematic tendencies in making decisions) and intuitive decision style (emphasis on feelings and fast judgments in making decisions) [63]. Participants were randomized to the narrative message that involves storytelling of a person implementing stress management methods to cope with distress, or the research evidence message that includes statistical information from research studies (such as systematic reviews) regarding the impact of stress management methods [61]. The logistic regressions with standardized premessage stress management behavior as the covariate showed that ETS [62] and decision style [63] (which were standardized) moderated the effects of narrative versus research evidence messages on breathing exercise practice decision [61].

For people with higher ETS and intuitive decision style, narrative messages resulted in higher likelihood to play the breathing exercise video, perhaps because of the match in information presented (story) and judgment or decision style. In contrast, whereas for people with lower ETS, tentative evidence indicated that research evidence messages may outperform narrative messages [61]. The effect sizes for these differences are medium to large, but with *P* values for the simple slope analyses of .03 and .06 at –1 SD level of ETS and intuitive decision style, respectively [61]. As this is an experimental study with young adults, such evidence is considered preliminary and field studies with more diverse samples and perhaps cross-validations with larger sample sizes will be needed to test the reliability of the findings and to prevent overfitting [33].

Supplementary and exploratory regression analyses found that Item 2 (“I often go by my instincts when deciding on a course of action”) and Item 5 (“I tend to use my heart as a guide for my actions”) in the ETS scale [62] have the largest interaction effects, both with *P* values below adjusted alphas (accounting for multiple exploratory tests). These were assessed using standardized regression coefficients in log-odds metrics and corresponding odds ratios (ORs; refer to Table S1 in Multimedia Appendix 1). Combining these 2 items for further analysis also produced significant results (bs indicate unstandardized regression coefficients of the interaction terms), *b*=1.16, 95% CI 0.46-1.86; OR 3.20, 95% CI 1.59-3.43; *P*=.001. These comparable coefficients suggest that the shortened Experiential Thinking Style captured the same moderation pattern as the full scale (*b*=0.91, 95% CI 0.23-1.60; OR 2.49, 95% CI 1.26-4.93; *P*=.009).

Apart from conventional regression analyses for testing moderation effects, different analytical methods, including multiverse analyses, robustness checks, sensitivity analyses, and ML methods as well as replications may be implemented to compare moderation and personalization effects and/or assess the validity and reliability of findings [33]. If multiple methods show support for sizable moderations of certain terms, such items are more likely to be robust and meaningful moderators.

Results indicated evidence of internal reliability. The internal consistency of the full version was Cronbach α=0.67 with the average interitem Pearson correlation as 0.17, and the 2-item short version yielded a Spearman-Brown coefficient of 0.63 with the interitem correlation as 0.46. There was a strong positive correlation between the two versions, *r*(164)=0.84, 95% CI 0.79-0.88; *P*<.001. The short ETS scale also showed positive correlations with the Intuitive Decision Style subscale, *r*(164)=0.74, 95% CI 0.66-0.80; *P*<.001, and negative correlations with the Rational Decision Style subscale, *r*(164)=–0.29, 95% CI –0.42 to –0.14; *P*<.001, providing evidence of convergent validity. Correlations among all scales are presented in Table S5 in Multimedia Appendix 1.

We also conducted similar regression analyses and exploratory factor analyses (EFAs) with Hamilton et al [63] scales, and such results with suggested short forms are reported in Tables S2 and S3 in Multimedia Appendix 1. To summarize, we found support for moderation with both the full version (OR 2.70, 95% CI 1.34-5.45; *P*=.005) and the short form (OR 2.82, 95% CI 1.42-5.61; *P*=.003) of Hamilton et al [63] Intuitive Decision Style. Results indicate evidence of acceptable internal consistency of both long form (α=0.88, average interitem correlation=0.54) and short form (Spearman-Brown coefficient=0.88, interitem correlation=0.78). The 2 selected items (Item 6 and Item 8, refer to Multimedia Appendix 1 for the items) have the largest regression coefficients with *P* values below adjusted alphas (Table S3 in Multimedia Appendix 1) and the highest factor loadings based on EFA (Table S6 in Multimedia Appendix 1).

To complement the above regression analyses, we also conducted EFA with both Maximum Likelihood-Varimax and Minimum Residuals-Oblimin, and the results with 2 methods are highly similar. The EFA with 172 individuals in this study revealed a 3-factor solution, with these 2 items (2 and 5) being the highest factor-loading items on 2 factors, respectively (refer

XSL•FO
RenderX

to Table S4 in Multimedia Appendix 1 for the values). Although Item 8 ("I generally don't depend on my feelings to help me make decisions," reversed, [62]) represents a significant item for the third factor in the EFA, it did not significantly moderate the impact of messages on practice decisions in the item-level regression and is therefore not recommended. Given the goal of item selection is for personalization, an item being significant for a factor does not guarantee its inclusion if the item is not a meaningful moderator. Given these insights, future studies that aim to personalize narrative versus statistical or research evidence messages based on ETS may consider including Item 2 and Item 5 [62] for simplicity to test whether moderation effects exist. Such short forms are intended for testing potential matching and moderation effects in the context of narrative versus statistical or research evidence messages only, but not for other purposes.

Validity and Reliability Considerations

While ultrashort scales may improve engagement and reduce dropout, they inevitably raise questions about validity and reliability [64,65], which we partially examined in the above examples and will conduct further studies. The key is not to assume equivalence with longer scales but to define acceptable benchmarks and trade-offs given the measurement purpose. For reliability, conventional benchmarks such as Cronbach α will no longer be appropriate given a short scale will almost inevitably yield a lower value of Cronbach α, not necessarily because it is unreliable, but because the formula inherently penalizes brevity [66]. Consider the formula of Cronbach α: $\alpha = N \times r / 1 + (N-1) \times r$, where N is the total number of items and r is the mean interitem correlation. For example, a 3-item scale with interitem correlations of 0.50 yields a Cronbach α value of 0.75, while a 10-item scale with the same correlations would reach a Cronbach α value of 0.91. It is not the reliability in itself but the number of items that biases the value downward. Moreover, Cronbach α assumes all items have essentially equal factor loadings (tau-equivalence). This assumption rarely holds in psychological constructs, especially after trimming [67], and violating it causes Cronbach α to misrepresent true reliability [68]. When items differ in their factor loadings, as they often do in ultrashort scales, Cronbach α becomes a particularly inaccurate metric. For 2-item measures, while statistics such as the Spearman-Brown coefficient are recommended as a mathematically superior alternative to Cronbach α [69], these metrics still fundamentally rely on classical test theory assumptions that can become unstable with only 2 data points. Therefore, although they could serve as a reference, traditional reliability tests of this kind are not particularly telling for ultrashort scales.

Instead, we recommend using interitem correlation as the reliability indicator. Briggs and Cheek [70] recommend that the mean of a set of interitem correlations ideally lies between 0.20 and 0.40, whereas greater than 0.50 suggests potential item redundancy. Likewise, Clark and Watson [71,72] recommended that the average interitem correlation and the average interitem correlation for each item fall within the range of 0.15 - 0.50. Together, we suggest somewhere around 0.30 (±0.15-0.20) to be acceptable, but these criteria are arbitrary and not universally applicable. Reliability should not be treated as a fixed and sole criterion but interpreted alongside construct coverage and test-retest stability [34,73].

For validity, reducing scales by large narrows conceptual bandwidth, capturing only the most central indicators of a latent construct [74]. To determine whether an ultrashort scale retains construct validity, if feasible, researchers can evaluate both structural equivalence and conceptual correspondence. For structural equivalence, researchers may fit both the full and shortened scales within the same dataset in confirmatory factor analysis if there are 3 or more items, while examining their factor loadings, configural structure, and fit indices [34,41]. If the short scale reproduces the same factor structure with acceptable fit, structural validity is supported [34,41].

For conceptual correspondence, researchers can correlate scores from the short form with the full-scale scores and external theoretically related variables. Convergent correlation over 0.80 or above between the short form and long form may be considered acceptable [75], but there is no well-established acceptable cutoff. When item removal leads to factor distortion or loss of theoretically important variance, further reduction is unjustified [74]. However, confirmatory factor analysis can possibly be done for short forms with 3 items or more, whereas for models with 2 items, it would be just-identified, resulting in a perfect fit. Therefore, in 2-item scenarios, we recommend cross-checking the variance explained with at least 2 datasets and comparing the explanatory variance loss from the shortened scale compared to the full scale (also see the paper by Smith [74]).

When Using Ultrashort Scales for Personalization Is Problematic

Ultrashort scales should be avoided when measurement precision, reliability, or validity would be compromised [34,65]. They may not be suitable for multifaceted constructs such as mindfulness, where multiple subdomains (cognitive, affective, and behavioral) jointly define the construct [76]. For such constructs, subscales may be more suitable. Constructs that fluctuate rapidly, such as momentary affect and state mindfulness, or other variables that require repeated measurements also cast doubt. Researchers can use experience sampling at different time points using short brief items, preserving representativeness while minimizing fatigue [72].

Ultrashort scales may also be problematic when standardized comparability (eg, a cutoff) is required, as brevity disrupts communication and comparability across research and practice settings [77]. When full instruments are replaced by abbreviated forms, sometimes results may not be well cross-referenced across studies, clinical services, or policy benchmarks, limiting shared understanding and comparability [77]. In such contexts, maintaining fully validated scales ensures consistency, interpretability, and interoperability across systems.

Finally, ultrashort scales risk bias when measurement invariance across demographic or cultural groups is untested, or when the construct itself remains theoretically underdeveloped. In these cases, scale shortening is premature because there is no stable

XSL•FO
RenderX

conceptual foundation for deciding which items are essential for content coverage or certain intended purposes such as diagnosis and screening [74]. This occurs when a construct's definition, dimensionality, or boundaries are still evolving, unclear, or inconsistently operationalized across studies [72]. For example, emerging concepts such as digital alliances are still lacking established theoretical hierarchies. Reducing items without a clear framework risks eliminating important indicators. In such contexts, the priority should be theoretical clarification and factor validation rather than item trimming [72].

Recommendations on Practical Framework

When scale shortening is considered appropriate without substantial problems, researchers developing ultrashort psychological scales for digital platforms can align their reduction methodology with the specific utility of the data. We propose a 3-tiered framework for method selection (Figure 1). For example, a researcher may prioritize both predictive accuracy and identification of items that interact better with conditions. In such situations, a combination of FA or ML approaches plus item-level moderation analyses may be adopted.

Figure 1. Decision framework. The purposes are not mutually exclusive, meaning there can be multiple purposes in scale shortening. The diagram is created with Gemini (version 3 Pro; Google DeepMind) and Nano Banana Pro (Google DeepMind).

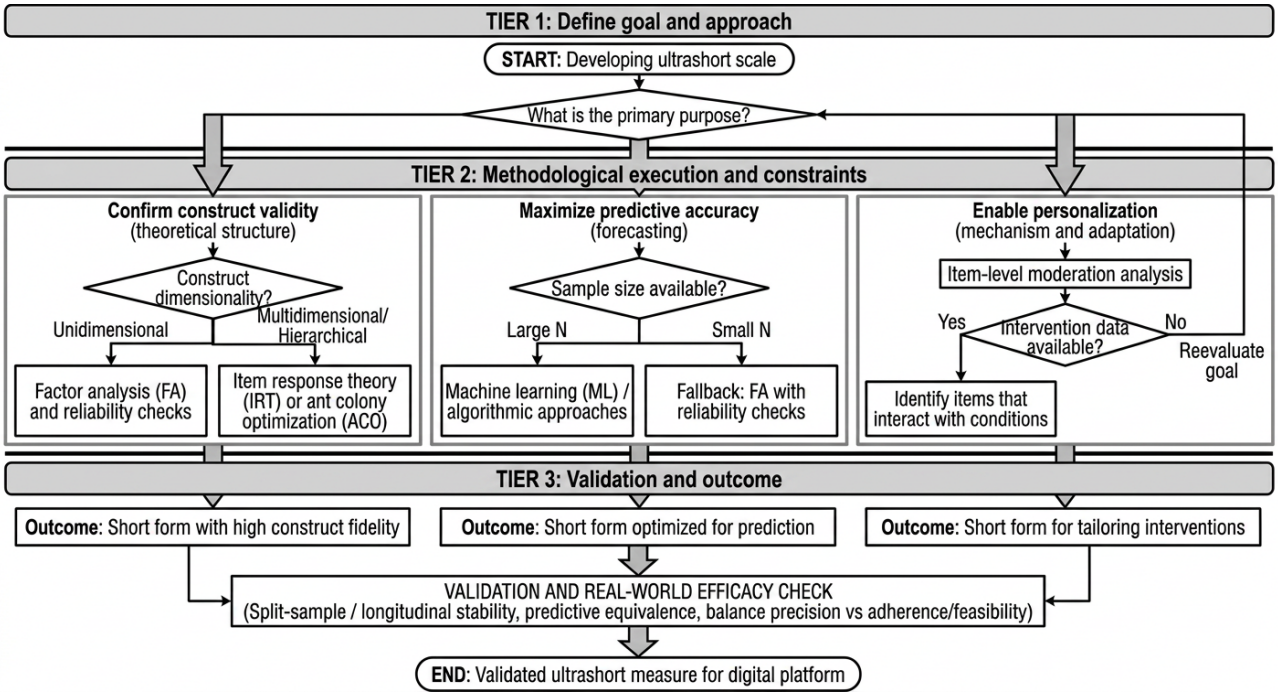

Tier 1: Define Goal

When the goal is to replicate the structure of a full-length assessment (eg, replacing standard intake forms), traditional FA or IRT is recommended [74,78]. These methods prioritize internal consistency and the maintenance of the latent trait.

If the objective is forecasting outcomes, such as behavioral patterns or engagement probabilities, algorithmic approaches such as ML or ACO are generally superior. These methods prioritize variance explanation over theoretical purity [79,80].

When the objective is to personalize, accuracy alone is insufficient. A model may predict engagement with 90% accuracy without revealing how to influence it. While conducting the above scale shortening statistical methods is worthwhile, item-level moderation analyses can be considered. This approach identifies items that interact with intervention conditions to predict differential outcomes. For example, specific items assessing "experiential thinking style" may moderate the effects of narrative versus research-based content. Unlike standard reduction, which discards "noisy" items, personalization retains items that function as distinct mechanisms for adaptation.

The inclusion of a reevaluation loop is critical to safeguard against methodological misalignment. It enforces the sharp distinction between prognostic utility (predicting who is likely to engage or improve) and prescriptive utility (determining which specific intervention will cause that improvement). True personalization relies on detecting heterogeneity of treatment effects, a statistical interaction that mathematically requires variation in intervention conditions [81]. Without distinct data on intervention types (eg, narrative vs didactic), it is premature to isolate the "mechanism of action" required to claim a scale enables personalization. This loop compels researchers to acknowledge this constraint, preventing unsupported claims about adaptability and ensuring they pivot back to methods such as ML or FA that align with the limitations of a static dataset.

Tier 2: Consider Methodological Constraints

Once the goal is defined, technical constraints dictate the specific analytical execution. We should take into account the

XSL•FO
RenderX

factor structure and the sample size available before choosing a suitable method.

Factor Structure

The dimensionality of the construct must be revisited prior to item selection. For unidimensional constructs, strong factor loadings and low residual correlations in FA suggest that a few high-performing items can adequately represent the latent trait [82]. However, multidimensional or hierarchical constructs require IRT or metaheuristic methods (eg, ACO) to ensure the short form retains measurement precision across all necessary subdomains without losing content validity.

Sample Size Practicality

Method selection is often constrained by data availability. FA can yield stable results with smaller cohorts. In contrast, IRT typically requires n>500 for reliable parameter estimation [83], and ML-based approaches ideally require several thousand participants to minimize overfitting during cross-validation [33]. In data-scarce environments, we recommend generating preliminary short forms via FA and reliability checks to be refined later as sample sizes grow.

Tier 3: Validation and Real-World Performance

Regardless of the derivation method, rigorous validation is needed. Split-sample or longitudinal validation needs to demonstrate structural stability, acceptable reliability (interitem correlation 0.15-0.50; [72]), and predictive equivalence. Triangulating methods can be considered. For instance, researchers may retain items that possess both high factor loadings and strong IRT information to enhance robustness [84]. Replications with multiple different samples are also recommended.

The predictive performance of ultrashort scales can be compared with the full scale and evaluated in the context of ecological feasibility. While long scales theoretically offer higher measurement precision, their predictive validity often degrades in real-world digital settings due to fatigue and attrition [85]. Well-chosen items often retain distinctively high power. For instance, the PHQ-4 explains 80% - 90% of the variance in depressive and anxiety outcomes relative to the full PHQ-9 and GAD-7, with minimal loss in classification accuracy [28] while providing the benefits mentioned throughout this paper.

Ethical Considerations

Apart from methodological considerations, researchers and practitioners should consider ethics carefully. Personalization based on a minimal number of items may be preferred by some users but disliked by others, whereas some service users may prefer other methods, such as artificial intelligence chatbot-based personalization but some service users may dislike artificial intelligence–based methods for mental health services [86]. Digital platforms should not impose such methods without the informed consent of users, as this may be perceived as a violation of privacy (see the papers by Matz et al [4] and Teeny et al [8] for related discussions). Transparent, clear, and explicit explanations of personalizations are needed to respect the diverse needs, preferences, and concerns of service users [20]. Explanations regarding both procedures, specifics of tailoring (ie, what and how data will be used for tailoring, for what purposes, such as recommendations, messaging, and user interface designs), advantages, potential benefits, disadvantages, and potential risks with different personalization methods should be provided in the consent forms, facilitating informed and autonomous choices [19]. Given the diversity of characteristics, experiences with different personalization methods, preferences, needs, and concerns, supporting autonomy and agency in choices regarding various personalization methods or not being personalized is essential for ethical and respectful practices [20]. This means that service users can choose between multiple personalization methods, possibly combinations of multiple methods (if available), or choose not to be personalized.

In addition, while there can be multiple personalization methods and multiple scales included for personalization purposes, it is essential to respect service users' choices in the intended purposes they want to or do not want to be personalized for. This ensures that the data collected are limited and relevant for specific purposes preferred by service users (data minimization), as emphasized in the EU General Data Protection Regulation Article 5(1)(c) [87]. The purpose of personalization (such as for mental health self-care activity recommendation and for reminder or persuasive messages encouraging self-care) of a scale can be specified, allowing service users to make informed decisions on whether they fill in such items.

Service users should also have the right to switch between different personalization methods or switch from being personalized to not being personalized (or vice versa), for example, if they feel uncomfortable or have privacy concerns regarding being personalized in a certain way [19]. Service users may also choose to delete personal data provided or modify the personal records [19], such as the responses to the ultrashort scales, as attributes of some people can change over time. Autonomy in personalization or nonpersonalization choices is critical, given implications for enhancing engagement and psychological well-being, which are key priorities in digital mental health services [20].

Moreover, another potential ethical concern is that algorithmic approaches can risk embedding or amplifying systematic biases present in training data. When personalization is based purely on data-driven correlations, demographic features (eg, age, gender, ethnicity, and education) may inadvertently drive recommendations, leading to discriminatory or inequitable tailoring [88]. For instance, a digital mental health algorithm might offer different content intensities or therapeutic suggestions to subgroups not because of psychological need but due to biased data patterns. A related concern is monophily, the phenomenon where individuals' attributes can be inferred not only from their own data but also from their network connections or interaction patterns [89]. ML methods trained on these interactions may cause further information divide.

Limitations and Future Directions

The examples of moderation effects we discussed in this paper are based on a single simple but not multiple samples, and we cannot rule out the possibility of false positives and the

XSL•FO
RenderX

possibility of such items having larger regression coefficients due to chance, even though we applied alpha corrections for multiple testing in item-level moderation analyses [90]. In addition, test-retest reliability of the very short forms has not been confirmed. Further studies with additional participant samples, multiple statistical methods, and test-retest reliability tests, perhaps in a digital mental health platform, are needed to examine the validity and reliability as well as replicability, robustness, and generalizability of such (possible) moderation effects before applying broadly in services [33,34]. We also encourage researchers to conduct studies with multiple samples, possibly with multiple statistical methods, before widespread implementation of personalization based on selected items [33,41,91].

While personalization based on 2 to 3 items may seem practically feasible and beneficial on the surface, more studies are needed to test the ecological validity, robustness, replicability, reliability, and generalizability of such methods. Field studies in digital health platforms can be conducted with a personalized group and a nonpersonalized group [92]. In field studies within digital health platforms, perhaps the personalized group of participants may fill in 2 to 3 items that have been shown to be stronger moderators for personalization purposes (eg, participants are assigned to one of the messages based on individual characteristics, short form). Such a personalized (experimental) group can then be compared to the nonpersonalized group (eg, in which participants are assigned to one of the messages randomly). Both objective (eg, number of modules registered, number of modules completed, and duration of usage) and self-report measures (eg, satisfaction and mental health) can be included in digital platforms [9] to assess the efficacy of personalization based on a minimal number of items. We recognize the current empirical base of personalization based on validated ultrashort scales is limited, especially in digital health contexts. The goal of this viewpoint piece is to highlight the gap and to call for future research, but not to systematically synthesize the very scarce literature. We look forward to systematic reviews on this topic when enough studies are conducted in the digital mental health domain.

In addition, future studies may compare the efficacy and effectiveness of personalization based on 2 to 3 items, personalization based on the original scale, personalization based on behavioral data (eg, type of courses registered) tracked in platforms, and personalization based on large language model chatbots. Another issue is that users are probably more likely to complete the minimal number of items than if the number of items is more than 10 [23], but how much more likely? Future studies are needed to empirically test this. Another worthwhile direction is to investigate users' preferences, attitudes, and concerns regarding being personalized based on a minimal number of items, when compared to being personalized based on a more lengthy scale, tracked behavioral data or being personalized based on large language model chatbots (see the paper by Pieritz et al [20]). It is also possible that people may prefer being personalized based on cognitive, motivational, and decisional characteristics compared to mental health symptoms or diagnosis, as mental health information may be more personally sensitive [19,20]. Future studies can test this speculation. These methods are not mutually exclusive and individual differences data may be captured by a combination of these methods, and personalization choices can be based on a combination or integration of these methods, if considered as acceptable and desirable by individuals.

Conclusion

To summarize, we propose that personalization based on a minimal number of items of psychological attributes, perhaps through scale shortening and/or through selecting the items that show stronger moderation effects, can be worthwhile. This is also likely more practical and can facilitate personalization for more service users than traditional psychological questionnaires that consist of a larger number of items [23], as more service users may be more willing to fill in a small number of items. Researchers and practitioners may choose the suitable methods (eg, FAs, ML methods, IRT, ACO, and item-level regression analyses) considering strengths, limitations, and feasibility of different methods, depending on factors such as sample size, researchers' familiarity with methods, and number of factors (also see the papers by Gonzalez [33], Koğar [34,41], and Raborn et al [91]). Our presented 3-tier decision framework may facilitate such choices and implementations. If feasible, the results of 2 or more methods can be considered in selecting the items. While our work focuses on digital mental health, such methods may also be relevant for other digital health domains and we encourage more research on scale shortening in various digital health domains.

While these ideas may seem promising, we call for more studies attempting to investigate diverse user experiences, different user attitudes, efficacy, effectiveness, ecological validity, generalizability, replicability, challenges, disadvantages, and ethical concerns of personalization based on a very small number of items before widespread implementations. Finally, we encourage researchers and practitioners to consider and analyze carefully and considerately regarding the length of scales adopted for personalization purposes in digital mental health contexts. We can aim to increase the likelihood of participants' filling in personalization-related measures with the potential of further enhancing engagement and reducing participants' reluctance and discomfort with filling in long questionnaires, while maintaining sufficient validity and reliability with shortened scales and respecting participants' informed choices of personalization methods, personalization purposes, or nonpersonalization (including opting out or removal of data).

Acknowledgments

We thank 3 peer reviewers and the editor for their very helpful suggestions, especially regarding psychometric and ethical considerations.

XSL•FO
RenderX

Disclosure of Delegation to generative artificial intelligence (GenAI; adapted based on GAIDeT Declaration Generator by Suchikova et al [93]).

The authors declare the use of GenAI in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: idea generation, literature search and systematization, visualization, proofreading and editing, and reformatting. The GenAI tools used were ChatGPT 5.1, Gemini 3, and Poe Assistant. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

The declaration is submitted by ACYT.

To be specific, OpenAI's ChatGPT (version 5) was used to search for relevant papers, organize and stimulate ideas. Poe Assistant was adopted for converting references from APA style to Vancouver style, as well as proofreading and English language editing. Gemini (version 3 Pro) with Nano Banana Pro was used to create the decision diagram.

Funding

The senior author (WWSM) received the General Research Fund (GRF) from the Research Grants Council (RGC) of Hong Kong SAR China on 2 studies regarding personalization of messages in boosting engagements in digital mental health (Ref No: 14606623), and the first (SKY) and third authors (HZ) are key members of that project. The funder has no involvement in any aspect of this manuscript.

Data Availability

The dataset and codes with outputs for the EFA and regression analyses discussed above [61] are available in the Open Science Framework repository [94].

Conflicts of Interest

The team has worked on numerous digital mental health projects. WWSM and ACYT are key leaders in Jockey Club TourHeart+, a digital mental health platform which incorporates personalization.

Multimedia Appendix 1
Supplementary tables of data analyses.
[DOCX File, 84 KB - jmir_v28i1e80662_app1.docx]

References

1. Banerjee S, Galizzi MM. People are different! and so should be behavioural interventions. In: Samson A, editor. The Behavioral Economics Guide 2024: Behavioral Science Solutions Ltd; 2024. URL: https://eprints.lse.ac.uk/125441/1/BanerjeeGalizzifinalpublishedversion.pdf [accessed 2026-05-08]
2. Joyal-Desmarais K, Scharmer AK, Madzelan MK, See JV, Rothman AJ, Snyder M. Appealing to motivation to change attitudes, intentions, and behavior: a systematic review and meta-analysis of 702 experimental tests of the effects of motivational message matching on persuasion. Psychol Bull 2022;148(7-8):465-517. [doi: 10.1037/bul0000377]
3. Matz SC, Kosinski M, Nave G, Stillwell DJ. Psychological targeting as an effective approach to digital mass persuasion. Proc Natl Acad Sci U S A 2017 Nov 28;114(48):12714-12719. [doi: 10.1073/pnas.1710966114] [Medline: 29133409]
4. Matz SC, Beck ED, Atherton OE, et al. Personality science in the digital age: the promises and challenges of psychological targeting for personalized behavior-change interventions at scale. Perspect Psychol Sci 2024 Nov;19(6):1031-1056. [doi: 10.1177/17456916231191774] [Medline: 37642145]
5. Mills S. Personalized nudging. Behav Public Policy 2022 Jan;6(1):150-159. [doi: 10.1017/bpp.2020.7]
6. Peer E, Egelman S, Harbach M, Malkin N, Mathur A, Frik A. Nudge me right: personalizing online security nudges to people's decision-making styles. Comput Human Behav 2020 Aug;109:106347. [doi: 10.1016/j.chb.2020.106347]
7. Peer E, Mills S. From one, many: how can nudges be personalized? Behav Sci Policy 2026 Apr;12(1):49-59. [doi: 10.1177/23794607251403327]
8. Teeny JD, Siev JJ, Briñol P, Petty RE. A review and conceptual framework for understanding personalized matching effects in persuasion. J Consum Psychol 2021 Apr;31(2):382-414. [doi: 10.1002/jcpy.1198]
9. Yeung SK, Mak WWS, Cheung GMH, Li CW, Yu HC, Zheng Y. A systematic review of nudging in the mental health contexts – progress, findings, and ways forward. Collabra Psychol . [doi: 10.13140/RG.2.2.25096.33288]
10. Krebs P, Prochaska JO, Rossi JS. A meta-analysis of computer-tailored interventions for health behavior change. Prev Med 2010;51(3-4):214-221. [doi: 10.1016/j.ypmed.2010.06.004] [Medline: 20558196]
11. Nye A, Delgadillo J, Barkham M. Efficacy of personalized psychological interventions: a systematic review and meta-analysis. J Consult Clin Psychol 2023 Jul;91(7):389-397. [doi: 10.1037/ccp0000820] [Medline: 37166831]

XSL•FO
RenderX

12. Tong HL, Quiroz JC, Kocaballi AB, et al. Personalized mobile technologies for lifestyle behavior change: a systematic review, meta-analysis, and meta-regression. Prev Med 2021 Jul;148:106532. [doi: 10.1016/j.ypmed.2021.106532] [Medline: 33774008]
13. Nikoloudakis IA, Crutzen R, Rebar AL, et al. Can you elaborate on that? Addressing participants' need for cognition in computer-tailored health behavior interventions. Health Psychol Rev 2018 Dec;12(4):437-452. [doi: 10.1080/17437199.2018.1525571] [Medline: 30221580]
14. Dandil Y, Kingston J. Personalized mobile apps for mental health and well-being in the general population: a systematic review. J Technol Hum Serv 2025:1-34. [doi: 10.1080/15228835.2025.2491347]
15. Schaeuffele C, Zagorscak P, Langerwisch V, Wilke J, Medvedeva Y, Knaevelsrud C. A systematic review on personalization of treatment components in IBIs for mental disorders. Internet Interv 2025 Sep;41(100840):100840. [doi: 10.1016/j.invent.2025.100840] [Medline: 40584576]
16. Matthews P, Rhodes-Maquaire C. Personalisation and recommendation for mental health apps: a scoping review. Behav Inf Technol 2025 Jun 15;44(10):2389-2404. [doi: 10.1080/0144929X.2024.2356630]
17. Norris LA, Kendall PC. Moderators of outcome for youth anxiety treatments: current findings and future directions. J Clin Child Adolesc Psychol 2021;50(4):450-463. [doi: 10.1080/15374416.2020.1833337] [Medline: 33140992]
18. Hornstein S, Zantvoort K, Lueken U, Funk B, Hilbert K. Personalization strategies in digital mental health interventions: a systematic review and conceptual framework for depressive symptoms. Front Digit Health 2023;5:1170002. [doi: 10.3389/fdgth.2023.1170002] [Medline: 37283721]
19. Valentine L, D'Alfonso S, Lederman R. Recommender systems for mental health apps: advantages and ethical challenges. AI Soc 2023 Aug;38(4):1627-1638. [doi: 10.1007/s00146-021-01322-w]
20. Pieritz S, Khwaja M, Faisal AA, Matic A. Personalised recommendations in mental health apps: the impact of autonomy and data sharing. Presented at: CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; May 8-13, 2021; Yokohama, Japan p. 1-12. [doi: 10.1145/3411764.3445523]
21. Alslaity A, Chan G, Orji R. A panoramic view of personalization based on individual differences in persuasive and behavior change interventions. Front Artif Intell 2023;6:1125191. [doi: 10.3389/frai.2023.1125191] [Medline: 37841233]
22. Kaptein M, Markopoulos P, de Ruyter B, Aarts E. Personalizing persuasive technologies: explicit and implicit personalization using persuasion profiles. Int J Hum Comput Stud 2015 May;77:38-51. [doi: 10.1016/j.ijhcs.2015.01.004]
23. Rolstad S, Adler J, Rydén A. Response burden and questionnaire length: is shorter better? A review and meta-analysis. Value Health 2011 Dec;14(8):1101-1108. [doi: 10.1016/j.jval.2011.06.003] [Medline: 22152180]
24. Zhang R, Ringland KE, Paan M, Mohr DC, Reddy M. Designing for emotional well-being: integrating persuasion and customization into mental health technologies. Presented at: CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; May 8-13, 2021; Yokohama, Japan p. 1-13. [doi: 10.1145/3411764.3445771]
25. Sundar SS, Marathe SS. Personalization versus customization: the importance of agency, privacy, and power usage. Hum Commun Res 2010;36(3):298-322. [doi: 10.1111/j.1468-2958.2010.01377.x]
26. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
27. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
28. Kroenke K, Spitzer RL, Williams JBW, Löwe B. An ultra-brief screening scale for anxiety and depression: the PHQ-4. Psychosomatics 2009;50(6):613-621. [doi: 10.1176/appi.psy.50.6.613] [Medline: 19996233]
29. Hlynsson JI, Carlbring P. Diagnostic accuracy and clinical utility of the PHQ-2 and GAD-2: a comparison with long-format measures for depression and anxiety. Front Psychol 2024;15:1259997. [doi: 10.3389/fpsyg.2024.1259997] [Medline: 38800678]
30. Hlynsson JI, Lindner P, Sjöberg A, Carlbring P. Let's talk about digital mental health assessment: utilizing PHQ-2 and GAD-2 for in-treatment monitoring and BBQ and QWB for outcome evaluation in internet interventions. Presented at: The 13th Swedish Congress on Internet Interventions; May 20-21, 2024; Stockholm, Sweden p. 14 URL: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1876267&dswid=6622 [accessed 2026-05-08]
31. Cheng AL, Abraham J, Hartz SM, Laber EB, Miller JP. Evaluating and optimizing just-in-time adaptive interventions in a digital mental health intervention (Wysa for chronic pain) for middle-aged and older adults With chronic pain: protocol for a series of randomized trials. JMIR Res Protoc 2025 Sep 17;14(1):e77532. [doi: 10.2196/77532] [Medline: 40961490]
32. Velmovitsky P, Keown-Stoneman C, J Pfisterer K, et al. Development of a recommendation engine to university student mental health support aligned with stepped care: longitudinal cohort study. J Med Internet Res 2025 Sep 17;27:e72669. [doi: 10.2196/72669] [Medline: 40767642]
33. Gonzalez O. Psychometric and machine learning approaches to reduce the length of scales. Multivariate Behav Res 2021;56(6):903-919. [doi: 10.1080/00273171.2020.1781585] [Medline: 32749158]
34. Koğar H. Development of a short form: methods, examinations, and recommendations. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi 2020;11(3):302-310. [doi: 10.21031/epod.739548]
35. Raborn AW, Leite WL. ShortForm: an R package to select scale short forms with the ant colony optimization algorithm. Appl Psychol Meas 2018 Sep;42(6):516-517. [doi: 10.1177/0146621617752993] [Medline: 30787491]

XSL•FO
RenderX

36. Hull L, Mandy W, Belcher H, Petrides KV. Validation of the Camouflaging Autistic Traits Questionnaire Short Form (CATQ-SF). Compr Psychiatry 2024 Nov;135:152525. [doi: 10.1016/j.comppsych.2024.152525] [Medline: 39191097]
37. Rogers SK, Whitted WM, Rand KL, Cheavens JS. Hope scale-short form: validation of a shortened measure of hope. Int J Appl Posit Psychol 2024 Apr;9(1):411-434. [doi: 10.1007/s41042-023-00132-3]
38. Chio FHN, Lai MHC, Mak WWS. Development of the nonattachment scale-short form (NAS-SF) using item response theory. Mindfulness (N Y) 2018 Aug;9(4):1299-1308. [doi: 10.1007/s12671-017-0874-z]
39. Bjorner JB, Petersen MA, Groenvold M, et al. Use of item response theory to develop a shortened version of the EORTC QLQ-C30 emotional functioning scale. Qual Life Res 2004 Dec;13(10):1683-1697. [doi: 10.1007/s11136-004-7866-x] [Medline: 15651539]
40. Olaru G, Danner D. Developing cross-cultural short scales using ant colony optimization. Assessment 2021 Jan;28(1):199-210. [doi: 10.1177/1073191120918026] [Medline: 32418476]
41. Koğar H. Which scale short form development method is better? A Comparison of ACO, TS, and SCOFA. Int J Assess Tool Educ 2022;9(3):583-592. [doi: 10.21449/ijate.946231]
42. Watkins MW. A Step-by-Step Guide to Exploratory Factor Analysis with SPSS: Routledge; 2021.
43. Arbuckle JL. IBM SPSS Amos 19 User's Guide: Amos Development Corporation; 2010:635 URL: http://amosdevelopment.com/download/amos.pdf [accessed 2026-05-08]
44. Şahi˙n M, Aybek E. Jamovi: an easy to use statistical software for the social scientists. Int J Assess Tool Educ 2019;6(4):670-692. [doi: 10.21449/ijate.661803]
45. Rosseel Y. lavaan: an R package for structural equation modeling. J Stat Soft 2012;48(2). [doi: 10.18637/jss.v048.i02]
46. Revelle W. How to: use the psych package for factor analysis and data reduction. : Department of Psychology, Northwestern University; 2016 URL: https://cran.uni-muenster.de/web/packages/psychTools/vignettes/factor.pdf [accessed 2026-05-08]
47. Zanon C, Hutz CS, Yoo HH, Hambleton RK. An application of item response theory to psychological test development. Psicol Refl Crít 2016 Dec;29(1). [doi: 10.1186/s41155-016-0040-x]
48. Paek I, Han KT. IRTPRO 2.1 for windows (item response theory for patient-reported outcomes). Appl Psychol Meas 2013 May;37(3):242-252. [doi: 10.1177/0146621612468223]
49. Muthén B, Mplus ML. Handbook of Item Response Theory: Chapman and Hall/CRC; 2017:507-518.
50. Chalmers RP. Mirt: a multidimensional item response theory package for the R environment. J Stat Softw 2012;48:1-29. [doi: 10.18637/jss.v048.i06]
51. Robitzsch A, Kiefer T, Wu M, et al. Package "TAM": test analysis modules. CRAN. 2020. URL: 10.32614/CRAN.package.TAM [accessed 2026-05-28]
52. Baghaei P, Robitzsch A. A tutorial on item response modeling with multiple groups using TAM. Educ Methods Psychom 2025;3(1):1-14. [doi: 10.61186/emp.2025.1]
53. Rizopoulos D. ltm: an R package for latent variable modeling and item response theory analyses. J Stat Soft 2007;17(5). [doi: 10.18637/jss.v017.i05]
54. Rizopoulos D, Rizopoulos MD. Package 'ltm': Latent Trait Models under IRT. CRAN. 2022. URL: https://cran.r-project.org/web/packages/ltm/ltm.pdf [accessed 2026-05-28]
55. Koğar E Yılmaz, Koğar H. A validation study of the Self-Compassion Scale-Short Form (SCS-SF) with ant colony optimization in a Turkish sample. Psihologija 2023;56(1):89-114. [doi: 10.2298/PSI211127015K]
56. Wagner S, Affenzeller M. HeuristicLab: a generic and extensible optimization environment. 2005 Presented at: Adaptive and Natural Computing Algorithms: Proceedings of the International Conference; Mar 21-23, 2005; Coimbra, Portugal p. 538-541. [doi: 10.1007/3-211-27389-1_130]
57. Jeong EM, Kim JK. Less is more: machine learning-based shortened sleep questionnaires for efficient clinical practice. Sleep Med Res 2024 Dec;15(4):216-217. [doi: 10.17241/smr.2024.02523]
58. Scrucca L, Sahdra B. GAabbreviate: abbreviating items measures using genetic algorithms. rdrr.io. 2015. URL: https://rdrr.io/cran/GAabbreviate/ [accessed 2026-05-28]
59. Fortin FA, Rainville FM, Gardner MA, Parizeau M, Gagné C. DEAP: evolutionary algorithms made easy. J Mach Learn Res 2012 Jul 1;13:2171-2175 [FREE Full text]
60. Hayes AF. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach, 3rd edition: The Guilford Press; 2022.
61. Zhao H, Yeung SK, Mak WWS. Narrative versus research-based messages in encouraging stress management: decision style and experiential thinking style as moderators. ResearchGate. Preprint posted online on Nov 2, 2025 URL: https://www.researchgate.net/publication/397179578_Narrative_versus_Research-based_Messages_in_Encouraging_Stress_Management_-_Decision_Style_and_Experiential_Thinking_Style_as_Moderators [accessed 2026-05-08]
62. Norris P, Epstein S. An experiential thinking style: its facets and relations with objective and subjective criterion measures. J Pers 2011 Oct;79(5):1043-1079. [doi: 10.1111/j.1467-6494.2011.00718.x] [Medline: 21241307]
63. Hamilton K, Shih SI, Mohammed S. The development and validation of the rational and intuitive decision styles scale. J Pers Assess 2016;98(5):523-535. [doi: 10.1080/00223891.2015.1132426] [Medline: 26967981]

XSL•FO
RenderX

64. Ziegler M, Kemper CJ, Kruyen P. Short scales–five misunderstandings and ways to overcome them. J Individ Differ 2014 Jan 1;35(4):185-189. [doi: 10.1027/1614-0001/a000148]
65. Kruyen PM, Emons WHM, Sijtsma K. On the shortcomings of shortened tests: a literature review. Int J Test 2013 Jul;13(3):223-248. [doi: 10.1080/15305058.2012.703734]
66. Cortina JM. What is coefficient alpha? An examination of theory and applications. J Appl Psychol 1993;78(1):98-104. [doi: 10.1037//0021-9010.78.1.98]
67. Revelle W, Zinbarg RE. Coefficients alpha, beta, omega, and the glb: comments on Sijtsma. Psychometrika 2009 Mar;74(1):145-154. [doi: 10.1007/s11336-008-9102-z]
68. McNeish D. Thanks coefficient alpha, we'll take it from here. Psychol Methods 2018 Sep;23(3):412-433. [doi: 10.1037/met0000144] [Medline: 28557467]
69. Eisinga R, Grotenhuis MT, Pelzer B. The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? Int J Public Health 2013 Aug;58(4):637-642. [doi: 10.1007/s00038-012-0416-3] [Medline: 23089674]
70. Briggs SR, Cheek JM. The role of factor analysis in the development and evaluation of personality scales. J Pers 1986 Mar;54(1):106-148. [doi: 10.1111/j.1467-6494.1986.tb00391.x]
71. Clark LA, Watson D. Constructing validity: basic issues in objective scale development. Psychol Assess 1995;7(3):309-319. [doi: 10.1037/1040-3590.7.3.309]
72. Clark LA, Watson D. Constructing validity: new developments in creating objective measuring instruments. Psychol Assess 2019 Dec;31(12):1412-1427. [doi: 10.1037/pas0000626] [Medline: 30896212]
73. McCrae RR, Kurtz JE, Yamagata S, Terracciano A. Internal consistency, retest reliability, and their implications for personality scale validity. Pers Soc Psychol Rev 2011 Feb;15(1):28-50. [doi: 10.1177/1088868310366253] [Medline: 20435807]
74. Smith GT, McCarthy DM, Anderson KG. On the sins of short-form development. Psychol Assess 2000;12(1):102-111. [doi: 10.1037/1040-3590.12.1.102]
75. Neumann E, Rohmann E, Sattel H. The 10-item short form of the German Experiences in Close Relationships scale (ECR-G-10)—model fit, reliability, and validity. Behav Sci (Basel) 2023 Nov 16;13(11):935. [doi: 10.3390/bs13110935] [Medline: 37998682]
76. Rammstedt B, Beierlein C. Can't we make it any shorter? J Individ Differ 2014 Jan 1;35(4):212-220. [doi: 10.1027/1614-0001/a000141]
77. Lindenmayer JP. Are shorter versions of the Positive and Negative Syndrome Scale (PANSS) doable? A critical review. Innov Clin Neurosci 2017 Dec 1;14(11-12):73-76. [Medline: 29410940]
78. Reise SP. Invited paper: the rediscovery of bifactor measurement models. Multivariate Behav Res 2012 Sep 1;47(5):667-696. [doi: 10.1080/00273171.2012.715555] [Medline: 24049214]
79. Leite WL, Huang IC, Marcoulides GA. Item selection for the development of short forms of scales using an ant colony optimization algorithm. Multivariate Behav Res 2008;43(3):411-431. [doi: 10.1080/00273170802285743] [Medline: 26741203]
80. Schroeders U, Wilhelm O, Olaru G. Meta-heuristics in short scale construction: ant colony optimization and genetic algorithm. PLoS ONE 2016;11(11):e0167110. [doi: 10.1371/journal.pone.0167110] [Medline: 27893845]
81. Kravitz RL, Duan N, Braslow J. Evidence - based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q 2004;82(4):661-687. [doi: 10.1111/j.0887-378X.2004.00327.x] [Medline: 15595946]
82. Costello AB, Osborne JW. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. Pract Assess Res Eval 2005;10(7):1-9. [doi: 10.7275/jyj1-4868]
83. Yen WM, Fitzpatrick AR. Item response theory. In: Educational Measurement 2006, Vol. 4:111-153.
84. Vésteinsdóttir V, Reips UD, Joinson A, Thorsdottir F. An item level evaluation of the Marlowe-Crowne Social Desirability Scale using item response theory on Icelandic internet panel data and cognitive interviews. Pers Individ Dif 2017 Mar;107:164-173. [doi: 10.1016/j.paid.2016.11.023]
85. Zhang Y, Rong J, Wang X, et al. Smartphone app–based survey deployment patterns and longitudinal response rate: randomized controlled trial. J Med Internet Res 2025;27:e73972. [doi: 10.2196/73972]
86. Schepman A, Rodway P. Initial validation of the general attitudes towards artificial intelligence scale. Comput Hum Behav Rep 2020;1:100014. [doi: 10.1016/j.chbr.2020.100014] [Medline: 34235291]
87. Biega AJ, Potash P, Daumé H, Diaz F, Finck M. Operationalizing the legal principle of data minimization for personalization. 2020 Jul 25 Presented at: SIGIR '20; Jul 25-30, 2020; Virtual Event China p. 399-408. [doi: 10.1145/3397271.3401034]
88. Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: mapping the debate. Big Data Soc 2016 Dec;3(2):2053951716679679. [doi: 10.1177/2053951716679679]
89. Altenburger KM, Ugander J. Monophily in social networks introduces similarity among friends-of-friends. Nat Hum Behav 2018 Apr;2(4):284-290. [doi: 10.1038/s41562-018-0321-8] [Medline: 30936539]
90. García-Pérez MA. Use and misuse of corrections for multiple testing. Methods Psychol 2023 Nov;8:100120. [doi: 10.1016/j.metip.2023.100120]
91. Raborn AW, Leite WL, Marcoulides KM. A comparison of metaheuristic optimization algorithms for scale short-form development. Educ Psychol Meas 2020 Oct;80(5):910-931. [doi: 10.1177/0013164420906600] [Medline: 32855564]

XSL•FO
RenderX

92. Rouvere J, Griffith Fillipo IR, Romanelli M, et al. Personalization strategies for increasing engagement with digital mental health resources: sequential multiple assignment randomized trial. JMIR Ment Health 2025 Nov 4;12:e73188. [doi: 10.2196/73188] [Medline: 41187311]
93. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (generative AI delegation taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res 2026 Apr;33(3):2544331. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]
94. Personalization based on a minimal number of items opinion piece. OSF. 2025. URL: osf.io/rkw7h/overview [accessed 2026-05-08]

Abbreviations

ACO: ant colony optimization
ETS: Experiential Thinking Style
FA: factor analysis
GAD-2: Generalized Anxiety Disorder-2
GAD-7: Generalized Anxiety Disorder-7
IRT: item response theory
ML: machine learning
PHQ-2: Patient Health Questionnaire-2
PHQ-4: Patient Health Questionnaire-4
PHQ-9: Patient Health Questionnaire-9

Edited by S Brini; submitted 14.Jul.2025; peer-reviewed by A Alabi, RQ Wang, S Anukam; revised version received 09.Apr.2026; accepted 10.Apr.2026; published 08.Jun.2026.

Please cite as:
Yeung SK, Tong ACY, Zhao H, Mak WWS
Using Ultra-Abridged Individual Difference Scales for Personalization in Digital Mental Health to Improve Uptake, Engagement, and Experiences: Three-Tiered Decision Framework for Scale Shortening
J Med Internet Res 2026;28:e80662
URL: https://www.jmir.org/2026/1/e80662
doi:10.2196/80662

© Siu Kit Yeung, Alan C Y Tong, Han Zhao, Winnie W S Mak. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Ethical Governance of Large Language Models in Health Care: Trust, Responsibility, and Equity in Routine Use

Xiongwen Yang[1,2*], MD, PhD; Chuan Xu[1,2*], MD

[1]Department of Thoracic Surgery, Guizhou Provincial People's Hospital, 83 Zhongshan East Road, Guiyang, Guizhou, China
[2]NHC Key Laboratory of Pulmonary Immunological Diseases, Guizhou Provincial People's Hospital, Guiyang, Guizhou, China
[*]all authors contributed equally

Corresponding Author:
Xiongwen Yang, MD, PhD
Department of Thoracic Surgery, Guizhou Provincial People's Hospital, 83 Zhongshan East Road, Guiyang, Guizhou, China

Abstract

Large language models (LLMs) are becoming increasingly embedded in routine health care communication, raising ethical challenges that extend beyond model performance alone. This Viewpoint argues that ethical risks in LLM-enabled health care emerge through patterns of reliance, institutional embedding, and governance during real-world use. Using "adoption-phase ethics" as an analytic lens, this paper examines 3 interrelated dimensions of ethical risk. First, trust in LLM-enabled health care is shaped not only by technical accuracy, but also by institutional and relational conditions surrounding its use. Second, responsibility may become distributed and ambiguous when LLM-mediated information influences clinical communication and decision-making. Third, equity concerns arise from unequal capacities to interpret, contest, and benefit from LLM-generated information. We argue that ethical governance of LLMs in health care requires continuous, system-level oversight that extends beyond model evaluation alone, including clear accountability structures, role-sensitive implementation, and equity-oriented governance practices. By reframing ethical analysis around routine integration rather than technical performance alone, this Viewpoint aims to support more responsible and sustainable use of LLMs in health care.

(J Med Internet Res 2026;28:e93470) doi:10.2196/93470

KEYWORDS

large language models; medical ethics; ethical governance; health care communication; technology adoption; equity and trust; responsibility

Introduction

Large language models (LLMs) have rapidly evolved from experimental research tools to increasingly visible participants in health care information ecosystems. Early investigations have largely focused on whether these systems can perform clinically relevant tasks—such as summarizing medical documentation, clarifying technical terminology, or supporting elements of decision-making—with acceptable levels of accuracy and efficiency [1-3]. Such performance-oriented research has been essential in establishing technical feasibility and translational promise.

Recent advances in medical artificial intelligence (AI) have also explored techniques such as multicenter knowledge transfer, multimodal learning, and domain adaptation to improve model robustness and generalizability [4-7]. While these approaches primarily address technical performance, they further highlight the importance of considering how such systems are deployed, interpreted, and governed in real-world contexts. As LLMs begin to enter routine clinical and informational contexts, however, additional ethical questions emerge. These questions cannot be addressed solely by whether outputs meet predefined benchmarks under controlled evaluation. Rather, they concern how ethical responsibilities, patterns of reliance, and expectations of accountability may shift when LLMs are integrated into everyday care practices [8-11].

In this Viewpoint, we use the term "adoption-phase ethics" as an analytic lens for examining how ethical risks emerge during the sustained integration of LLMs into routine health care practice. Rather than focusing primarily on predeployment evaluation or technical adoption alone, this perspective emphasizes how patterns of reliance, institutional embedding, and governance arrangements shape ethical outcomes during real-world use.

During experimental evaluation, ethical attention often centers on technical properties such as performance, validation, bias mitigation, and explainability [12]. These concerns remain crucial. Improvements in model accuracy and robustness can meaningfully reduce certain risks. Routine integration raises ethical questions that extend beyond technical performance alone. As LLM-mediated information becomes embedded within clinical communication and patient interpretation practices, ethical concerns increasingly involve how trust is formed, how responsibility is distributed, and how individuals interpret

XSL•FO
RenderX

uncertainty within institutional contexts [10,11,13,14]. These dynamics are particularly significant in health care, where patient vulnerability, professional accountability, and asymmetries of expertise shape reliance on AI-mediated information [11,13,15,16]. Governance conditions, including institutional accountability and uncertainty communication, therefore become central to ethical integration during routine use.

These considerations invite a reframing of AI ethics in medicine. At its core, this Viewpoint argues that the central ethical challenge of LLM integration in health care lies not simply in model performance, but in how patterns of reliance become institutionalized during routine use. Much prior work has emphasized predeployment evaluation, such as performance validation, bias assessment, and transparency requirements. While these remain foundational, ethical challenges in practice do not replace model-level concerns but interact with them in complex ways. Model-level properties, including accuracy, bias, and uncertainty, remain ethically significant. However, ethical risks during routine use are shaped not only by these technical characteristics, but also by how LLMs are embedded within institutional workflows, communication practices, and governance structures. Ethical analysis must therefore extend beyond technical evaluation toward questions of accountability, reliance, and equitable capacity to benefit during sustained use [17,18].

In this Viewpoint, adoption is understood not merely as technical deployment, but as the stabilization of LLM use within everyday clinical and informational practices. Drawing on normative ethical analysis informed by existing empirical discussions, we examine 3 interrelated domains—trust, responsibility, and equity—to clarify how routine LLM use may (1) reshape institutional and interpersonal trust, (2) create responsibility gaps, and (3) amplify disparities in interpretive and practical capacity.

This Viewpoint contributes to existing discussions by reframing ethical analysis from a predominantly model-centric perspective toward one that is governance-oriented and grounded in routine use. It conceptualizes “adoption-phase ethics” as an analytic lens for understanding how patterns of reliance, institutional embedding, and differential capacity shape ethical risk during sustained LLM integration.

This perspective differs from many existing AI ethics frameworks, which primarily emphasize general principles such as fairness, transparency, and accountability at the model or system design level [19-21]. While existing AI ethics frameworks provide important normative guidance, they often emphasize ethical principles as properties to be ensured prior to deployment. In contrast, this perspective focuses on how ethical risks evolve through patterns of reliance, institutional embedding, and governance during routine use [19-22].

This Viewpoint focuses on LLM applications in communication support and non-autonomous decision contexts, rather than fully autonomous diagnostic systems. It aims to examine how ethical risks emerge during routine integration and to propose a governance-oriented framework grounded in trust, responsibility, and equity.

This analysis is structured around 3 key messages. First, trust in LLM-enabled health care is shaped not only by model performance, but also by institutional endorsement and governance structures. Second, responsibility becomes distributed across clinicians, institutions, and system designers, creating potential ambiguity in accountability. Third, equity concerns arise from differences in users’ capacity to interpret and benefit from AI-mediated information.

This Viewpoint is intended for clinicians, health system leaders, and policymakers engaged in the governance and implementation of LLM-enabled health care systems.

Conceptual Approach

As shown in Table S1 in Multimedia Appendix 1, recent empirical studies consistently highlight that ethical challenges in LLM use extend beyond model performance and are shaped by patterns of reliance, responsibility allocation, and governance conditions in real-world settings.

Conceptually, this analysis is situated at the intersection of AI ethics, health care technology adoption, and sociotechnical systems theory. Drawing on existing empirical and conceptual scholarship, this Viewpoint develops a governance-oriented analytical lens for examining how ethical risks emerge during the routine integration of LLMs into health care. Rather than focusing solely on model performance or adoption outcomes, the analysis emphasizes how patterns of reliance, responsibility allocation, and differential capacity to benefit are shaped through sustained real-world use.

The analysis is structured around 3 interrelated domains: trust, responsibility, and equity. These frequently arise in discussions of health care technology integration. Trust refers to institutional and relational expectations surrounding the reliability and governance of LLM-mediated information. Responsibility concerns how accountability and decision authority are distributed among clinicians, institutions, and system designers. Equity addresses differences in access, interpretive capacity, and practical benefit across user groups. Together, these domains provide a framework for examining how ethical risks become embedded within sociotechnical systems during routine LLM use.

The analysis focuses on LLM applications in communication support and nonautonomous decision contexts within health care. Figure 1 illustrates the interaction between model-level properties and system-level governance conditions, emphasizing that ethical risks emerge through their ongoing interplay during routine use rather than through a strictly linear transition between stages.

XSL•FO
RenderX

Figure 1. Interaction-based framework of ethical risk in large language model–enabled health care.

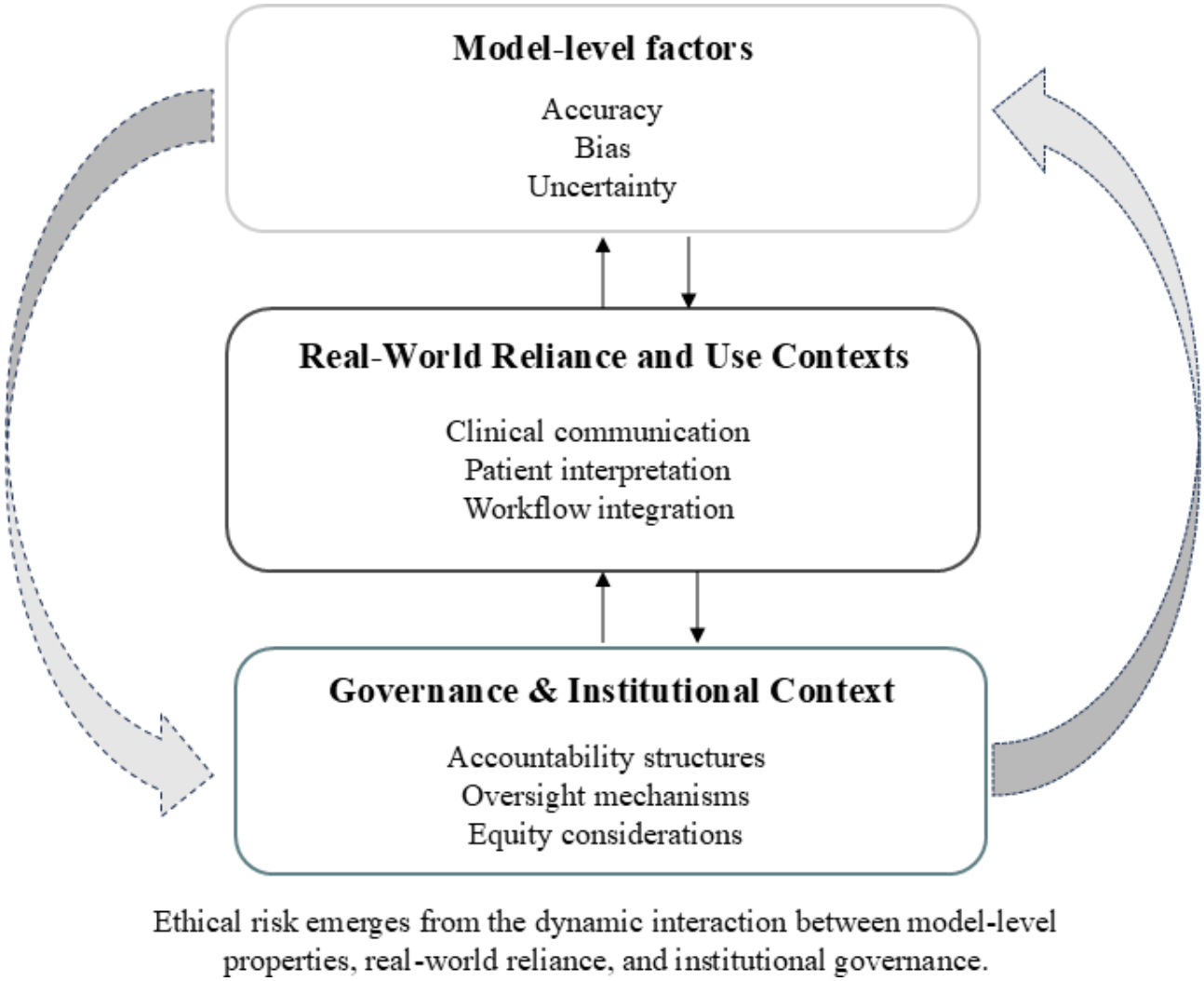

This figure presents a layered and interaction-based framework for understanding ethical risk during routine LLM integration in health care. Model-level properties (eg, accuracy, bias, and uncertainty), patterns of reliance in clinical practice (eg, communication, interpretation, and workflow integration), and governance conditions (eg, accountability structures, oversight mechanisms, and equity considerations) are conceptualized as interacting dimensions within a sociotechnical system.

Rather than representing a linear progression from model evaluation to implementation, the framework emphasizes that ethical risks emerge dynamically through the institutionalization of reliance during real-world use.

Ethical Dynamics of Routine LLM Adoption in Health Care: Trust as an Institutional and Relational Phenomenon

As LLMs move from experimental evaluation into routine clinical and informational contexts, questions of trust take on heightened ethical significance [2,23]. While technical performance remains relevant, trust during sustained use cannot be understood solely as confidence in algorithmic accuracy [16,24]. Instead, it is shaped by expectations regarding transparency, accountability, and institutional protection.

In practice, clinicians often evaluate LLM-supported communication not only on perceived output quality but also on whether use is embedded within governed workflows that clarify oversight and professional responsibility. For example, when LLMs generate patient-facing explanations of diagnostic results, clinicians may need to assess not only factual accuracy but also whether uncertainty is appropriately communicated and aligned with clinical judgment. For patients and caregivers, trust may depend on whether LLM-mediated information is intelligible, appropriately contextualized, and clearly distinguished from professional clinical judgment [9,25,26].

Taken together, these dynamics suggest that trust in adoption-phase LLM use is not merely an individual attitude toward technology. Rather, it emerges as a relational and institutional phenomenon, structured by governance signals that communicate who stands behind LLM-mediated information and how harms would be addressed if they occur. Ethical analysis must therefore consider not only whether models are accurate, but also how institutional endorsement and accountability shape patterns of reliance [10,27,28].

XSL•FO
RenderX

Responsibility and Accountability in Mediated-Information Environments

Routine integration of LLMs into health care communication also raises questions about responsibility allocation. Clinicians remain legally and ethically accountable for patient care, yet LLM-generated summaries, explanations, or drafts may mediate access to information in ways that obscure provenance, limitations, or uncertainty [26,28,29]. This issue is particularly pronounced in health care, where professional responsibility is formally codified, and where reliance on mediated information may have direct implications for patient safety and clinical accountability.

At the same time, patients and caregivers encountering AI-mediated explanations may struggle to distinguish between professional clinical guidance and automated inference. For instance, a patient receiving an LLM-generated explanation of treatment options may interpret the information as authoritative medical advice, even when it lacks formal clinical validation or contextualization. When boundaries are unclear, interpretive burdens can shift onto users who lack equivalent expertise or institutional recourse. This configuration may create the possibility of responsibility gaps—not necessarily because of technical error, but because of structural ambiguity in how accountability is distributed across clinicians, institutions, and system designers [24,27,29].

Importantly, these concerns are not exhausted by improvements in technical performance. Even highly accurate systems may introduce new ambiguities regarding authorship, endorsement, and liability when embedded into routine workflows [28]. Ethical evaluation must therefore address how responsibility is specified, communicated, and governed in contexts where AI mediates clinical information exchange.

Equity, Capability, and Differential Vulnerability

A further ethical dimension of routine LLM adoption concerns uneven capacity to interpret and benefit from AI-mediated health information. Engagement with digital tools is shaped by differences in digital literacy, educational background, language proficiency, and access to institutional support. These factors influence not only initial uptake but also users' ability to contextualize outputs, recognize uncertainty, and seek clarification when needed [10,30,31].

In health care settings, these disparities may have heightened ethical significance, as misinterpretation of information can directly affect patient understanding, decision-making, and health outcomes, particularly among vulnerable populations. In certain contexts, LLMs may function as capability amplifiers: they can enhance access and understanding for users equipped to navigate them effectively, while offering more limited or potentially confusing benefits to those with fewer resources. Users with greater digital literacy may be better positioned to contextualize or verify LLM-generated information, whereas others may rely on it more passively, increasing the risk of misunderstanding or misplaced trust. Without deliberate governance attention, patterns of reliance may therefore reproduce or intensify existing inequities [31,32].

Ethical vulnerability is also asymmetrically distributed across user roles. Patients and caregivers may face epistemic vulnerability when interpreting LLM-mediated information without sufficient authority or recourse [28]. Clinicians, by contrast, may face risks of overreliance, cognitive offloading, or erosion of professional judgment under time pressure. These vulnerabilities differ in kind and degree, underscoring that ethical risks associated with routine adoption are not evenly shared [24,26].

Rethinking AI Ethics Frameworks in Light of Adoption

The dynamics described above invite reconsideration of how AI ethics frameworks are applied to LLM integration in health care. Much prior work in medical AI ethics has emphasized principles such as transparency, explainability, fairness, and bias mitigation—often in the context of predeployment evaluation [7,33,34]. These principles remain foundational. However, routine adoption surfaces ethical challenges that extend beyond model-level properties [17,18,25].

Transparency, for example, is frequently framed as explainability of internal model logic. Yet in practice, users may be more concerned with institutional endorsement, accountability, and recourse mechanisms than with technical architecture. Similarly, fairness is often operationalized through statistical metrics, while disparities in digital literacy, infrastructure, and implementation context may shape who can meaningfully benefit from LLM-mediated information [16,24].

Another limitation of prevailing frameworks lies in their implicit individualization of responsibility. Ethical guidance often presumes that informed users can appropriately interpret and contest AI-generated information. In routine care environments characterized by time pressure and asymmetries in expertise, such assumptions may be unrealistic. Responsibility for ethical use cannot plausibly rest on individual vigilance alone; it must be institutionally structured and explicitly governed [26,29].

Finally, ethical evaluation is sometimes treated as a static, predeployment exercise. Adoption, by contrast, is dynamic. Patterns of reliance evolve, user expectations shift, and organizational norms adapt over time. Ethical oversight must therefore be conceived as an ongoing governance process responsive to emerging vulnerabilities rather than as a one-time compliance checklist [29].

Governance-Oriented Considerations for Routine Integration

If ethical risks become more pronounced during sustained adoption, governance responses must likewise shift in emphasis [18,27]. The patterns summarized in Table S1 in Multimedia Appendix 1 further demonstrate that governance challenges arise not only from model limitations, but from the ways in which LLMs are embedded within clinical workflows and relied upon in practice.

XSL•FO
RenderX

First, trust in LLM-supported communication should be institutionally anchored. At an operational level, governance mechanisms should be embedded within clinical workflows through structured human-in-the-loop review, audit trails documenting LLM-generated outputs and clinician modifications, and clear escalation pathways for uncertain, contested, or potentially harmful outputs. Health care organizations should also clarify how LLM outputs are validated, monitored, and integrated into communication and decision-making processes.

Second, governance should be role-sensitive. Clinicians and patients engage with LLMs under different conditions of authority, expertise, and exposure to harm. Ethical design should therefore support calibrated reliance and auditability for clinicians while ensuring intelligibility, uncertainty signaling, and accessible recourse for patients and caregivers. Examples of accessible recourse include clearly defined complaint or feedback pathways, mechanisms for correcting or retracting erroneous AI-generated information, and the availability of human support to review and clarify contested outputs. Such mechanisms should be clearly assigned to responsible actors (eg, clinical teams and institutional oversight committees) and embedded in patient-facing communication channels to ensure that users can request clarification, correction, or human review in a timely manner.

Third, equity requires proactive attention. Literacy-sensitive interfaces, adaptive explanatory strategies, and targeted support for underserved populations should be regarded as central components of ethically robust adoption. Without such measures, LLM integration risks amplifying disparities under the banner of informational democratization.

Finally, oversight should be continuous. Because adoption is an evolving process, health care systems must establish mechanisms for monitoring, feedback, and iterative revision. An ethics of adoption thus centers on institutional responsibility, inclusivity, and sustained governance rather than on technical safeguards alone. This may involve periodic auditing of system use, monitoring patterns of reliance and error, and incorporating user feedback into governance updates and system redesign.

Conclusion

LLMs are becoming increasingly embedded within everyday health care communication, reshaping how medical information is accessed, interpreted, and relied upon. In this Viewpoint, we argue that ethical challenges in LLM-enabled health care are not driven by model performance alone, but emerge through patterns of reliance, institutional embedding, and governance during routine use.

This analysis reframes ethical evaluation from a predominantly model-centric perspective toward a governance-oriented perspective grounded in real-world integration. By conceptualizing “adoption-phase ethics” as an analytic lens, this paper has highlighted how trust, responsibility, and equity interact as dynamic dimensions of ethical risk during sustained LLM use.

Several limitations should be acknowledged. This analysis is conceptual and does not provide empirical validation of the proposed framework. In addition, while we draw on emerging empirical studies to inform the discussion, the synthesis is not exhaustive and may not capture all implementation contexts.

Future research should examine how governance mechanisms and patterns of reliance shape ethical outcomes in clinical practice, as well as how context-specific governance strategies can be developed and evaluated across different health-care settings.

Ultimately, ethical integration of LLMs in health care will depend not only on technical performance, but also on how institutional governance structures shape reliance, accountability, and equitable capacity to benefit users during routine use.

Acknowledgments

The authors used a generative artificial intelligence tool (ChatGPT; OpenAI) to assist with language refinement, grammar correction, and consistency of the manuscript. The tool was not used for study design, conceptual development, data collection, data analysis, result interpretation, or reference generation. All scientific content, analyses, results, and conclusions were conceived, verified, and approved by the authors. Records of the artificial intelligence–assisted editing interactions have been retained by the authors and can be made available upon request, in accordance with JMIR Publications’ editorial policy on the use of generative artificial intelligence.

Funding

This work was supported by the Talent Fund of Guizhou Provincial People’s Hospital (grant 2024-16), the Science and Technology Fund Project of the Guizhou Provincial Health Commission (grant gzwkj2026-295) and the Basic Research Program of the Guizhou Provincial Department of Science and Technology (Qian Ke He Ji Chu; grant QKH-JC-MS [2025] 492). The funding sources had no role in the study design, analysis, interpretation of data, writing of the manuscript, or the decision to submit the manuscript for publication.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed.

XSL•FO
RenderX

Authors' Contributions

XY: conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft

CX: conceptualization, methodology, writing—review and editing, visualization supervision, project administration, funding acquisition

Conflicts of Interest

None declared.

Multimedia Appendix 1

Empirical studies informing trust, responsibility, and equity considerations in routine LLM integration in health care.
[XLSX File, 18 KB - jmir_v28i1e93470_app1.xlsx]

References

1. Menezes MCS, Hoffmann AF, Tan ALM, et al. The potential of generative pre-trained transformer 4 (GPT-4) to analyse medical notes in three different languages: a retrospective model-evaluation study. Lancet Digit Health 2025 Jan;7(1):e35-e43. [doi: 10.1016/S2589-7500(24)00246-2] [Medline: 39722251]
2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
3. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature New Biol 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
4. Wang C, Qian P, Wang Z, et al. Multicenter knowledge transfer calibration with rapid zeroth-order TSK fuzzy system for small sample epileptic EEG signals. IEEE Trans Fuzzy Syst 2024;32(11):6224-6236. [doi: 10.1109/TFUZZ.2024.3443878]
5. Wang C, Hu W, Wang J, Qian P, Wang S, Ortale R. Consistency and complementarity jointly regularized subspace support vector data description for multimodal data. Int J Intell Syst 2024 Jan;2024(1):1989706. [doi: 10.1155/2024/1989706]
6. Qiu C, Qian P, Wang C, et al. Joint pre-encoding representation and structure embedding for efficient and low-resource knowledge graph completion. 2024 Presented at: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. [doi: 10.18653/v1/2024.emnlp-main.851]
7. Fareed M, Fatima M, Uddin J, Ahmed A, Sattar MA. A systematic review of ethical considerations of large language models in healthcare and medicine. Front Digit Health 2025;7:1653631. [doi: 10.3389/fdgth.2025.1653631] [Medline: 41019285]
8. Yang X, Xiao Y, Liu D, et al. Enhancing physician-patient communication in oncology using GPT-4 through simplified radiology reports: multicenter quantitative study. J Med Internet Res 2025 Apr 17;27:e63786. [doi: 10.2196/63786] [Medline: 40245397]
9. Yang X, Xiao Y, Liu D, et al. Factors influencing adoption of large language models in health care: multicenter cross-sectional mixed methods observational study. J Med Internet Res 2025 Dec 11;27:e84918. [doi: 10.2196/84918] [Medline: 41380031]
10. Chen J, Liu Y, Liu P, Zhao Y, Zuo Y, Duan H. Adoption of large language model AI tools in everyday tasks: multisite cross-sectional qualitative study of Chinese hospital administrators. J Med Internet Res 2025 Apr 1;27:e70789. [doi: 10.2196/70789] [Medline: 40116330]
11. Dennstädt F, Schmerder M, Riggenbach E, et al. Comparative evaluation of a medical large language model in answering real-world radiation oncology questions: multicenter observational study. J Med Internet Res 2025 Sep 23;27:e69752. [doi: 10.2196/69752] [Medline: 40986858]
12. Hanna MG, Pantanowitz L, Jackson B, et al. Ethical and bias considerations in artificial intelligence/machine learning. Mod Pathol 2025 Mar;38(3):100686. [doi: 10.1016/j.modpat.2024.100686] [Medline: 39694331]
13. Shin HS, Williams H, Braykov N, Jahan A, Meller J, Orenstein EW. The influence of artificial intelligence scribes on clinician experience and efficiency among pediatric subspecialists: a rapid, randomized quality improvement trial. Appl Clin Inform 2025 Aug;16(4):1041-1052. [doi: 10.1055/a-2657-8087] [Medline: 40675605]
14. Yang X, Xiao Y, Liu D, et al. Enhancing doctor-patient communication using large language models for pathology report interpretation. BMC Med Inform Decis Mak 2025 Jan 23;25(1):36. [doi: 10.1186/s12911-024-02838-z] [Medline: 39849504]
15. Dhagarra D, Goswami M, Kumar G. Impact of trust and privacy concerns on technology acceptance in healthcare: an Indian perspective. Int J Med Inform 2020 Sep;141:104164. [doi: 10.1016/j.ijmedinf.2020.104164] [Medline: 32593847]
16. Kaur D, Uslu S, Rittichier KJ, Durresi A. Trustworthy artificial intelligence: a review. ACM Comput Surv 2023 Feb 28;55(2):1-38. [doi: 10.1145/3491209]
17. Singhal A, Neveditsin N, Tanveer H, Mago V. Toward fairness, accountability, transparency, and ethics in AI for social media and health care: scoping review. JMIR Med Inform 2024 Apr 3;12:e50048. [doi: 10.2196/50048] [Medline: 38568737]
18. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 2020 Mar 20;368:l6927. [doi: 10.1136/bmj.l6927] [Medline: 32198138]

XSL•FO
RenderX

19. Lekadir K, Frangi AF, Porras AR, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ 2025 Feb 5;388:e081554. [doi: 10.1136/bmj-2024-081554] [Medline: 39909534]
20. Weiner EB, Dankwa-Mullan I, Nelson WA, Hassanpour S. Ethical challenges and evolving strategies in the integration of artificial intelligence into clinical practice. PLOS Digit Health 2025 Apr;4(4):e0000810. [doi: 10.1371/journal.pdig.0000810] [Medline: 40198594]
21. Chan A, Rahimi-Ardabilli H, Rogers WA, Coiera E. The real-world impact of artificial intelligence ethics frameworks across a decade in healthcare: a scoping review. J Am Med Inform Assoc 2025 Nov 1;32(11):1767-1777. [doi: 10.1093/jamia/ocaf167] [Medline: 41093301]
22. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
23. Wong KKL, Han Y, Cai Y, Ouyang W, Du H, Liu C. From trust in automation to trust in AI in healthcare: a 30-year longitudinal review and an interdisciplinary framework. Bioengineering (Basel) 2025 Oct 1;12(10):1070. [doi: 10.3390/bioengineering12101070] [Medline: 41155069]
24. Visser R, Peters TM, Scharlau I, Hammer B. Trust, distrust, and appropriate reliance in (X)AI: a conceptual clarification of user trust and survey of its empirical evaluation. Cogn Syst Res 2025 Jun;91:101357. [doi: 10.1016/j.cogsys.2025.101357]
25. Ebnali Harari R, Altaweel A, Ahram T, Keehner M, Shokoohi H. A randomized controlled trial on evaluating clinician-supervised generative AI for decision support. Int J Med Inform 2025 Mar;195:105701. [doi: 10.1016/j.ijmedinf.2024.105701] [Medline: 39631268]
26. Ryan K, Kasun M, Roberts LW, Kim JP. Information, collaboration, regulation: physician and AI researcher views on ethical considerations in clinical AI integration. Big Data Soc 2025;12(2). [doi: 10.1177/20539517251343853] [Medline: 41355848]
27. Allen LN, Lin J, Segal BM, Ndlovu K, Bilardi D, Pettigrew LM. Artificial intelligence in primary care: frameworks, challenges, and guardrails. Lancet Prim Care 2026 Mar;2(3):100079. [doi: 10.1016/j.lanprc.2025.100079]
28. Nouis SC, Uren V, Jariwala S. Evaluating accountability, transparency, and bias in AI-assisted healthcare decision- making: a qualitative study of healthcare professionals' perspectives in the UK. BMC Med Ethics 2025 Jul 8;26(1):89. [doi: 10.1186/s12910-025-01243-z] [Medline: 40629303]
29. Torkamaan H, Steinert S, Pera MS, et al. Challenges and future directions for integration of large language models into socio-technical systems. Behav Inf Technol :1-20. [doi: 10.1080/0144929X.2024.2431068]
30. Urbina JT, Vu PD, Nguyen MV. Disability ethics and education in the age of artificial intelligence: identifying ability bias in ChatGPT and Gemini. Arch Phys Med Rehabil 2025 Jan;106(1):14-19. [doi: 10.1016/j.apmr.2024.08.014] [Medline: 39216786]
31. Ji Y, Ma W, Sivarajkumar S, et al. Mitigating the risk of health inequity exacerbated by large language models. NPJ Digit Med 2025 May 4;8(1):246. [doi: 10.1038/s41746-025-01576-4] [Medline: 40319154]
32. Ahmed H. Large language models for clinical trials in the Global South: opportunities and ethical challenges. AI Ethics 2026 Feb;6(1):76. [doi: 10.1007/s43681-025-00943-x]
33. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. World Health Organization. 2024. URL: https://www.who.int/publications/i/item/9789240084759 [accessed 2026-05-28]
34. Ethics and governance of artificial intelligence for health. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240029200 [accessed 2026-05-26]

Abbreviations

AI: artificial intelligence
LLM: large language model

Edited by I Steenstra; submitted 13.Feb.2026; peer-reviewed by C Wang, T Mori; revised version received 16.May.2026; accepted 20.May.2026; published 10.Jun.2026.

Please cite as:
Yang X, Xu C
Ethical Governance of Large Language Models in Health Care: Trust, Responsibility, and Equity in Routine Use
J Med Internet Res 2026;28:e93470
URL: https://www.jmir.org/2026/1/e93470
doi:10.2196/93470

XSL•FO
RenderX

© Xiongwen Yang, Chuan Xu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Nontraditional Data in Pandemic Preparedness and Response: Identifying and Addressing First- and Last-Mile Challenges

Mattia Mazzoli[1*], PhD; Irma Varela-Lasheras[2*], PhD; Sónia Namorado[2,3], PhD; Constantino Pereira Caetano[2], PhD; Andreia Leite[2,3], PhD; Lisa Hermans[4], PhD; Niel Hens[4,5], PhD; Polen Türkmen[1,6], MA; Kyriaki Kalimeri[1], PhD; Leo Ferres[1,7], PhD; Ciro Cattuto[1], PhD; Daniela Paolotti[1], PhD; Stefaan Verhulst[1,6,8,9], PhD

[1]ISI Foundation, Via Della Rocca 20, Turin, Italy

[2]Department of Epidemiology, National Institute of Health Doctor Ricardo Jorge, Lisbon, Portugal

[3]NOVA National School of Public Health, Public Health Research Centre, Comprehensive Health Research Center, Associated Laboratory in Translation and Innovation Towards Global Health, Clinical Academic Center, NOVA University Lisbon, Lisbon, Portugal

[4]Interuniversity Institute of Biostatistics and Statistical Bioinformatics, Data Science Institute, Hasselt University, Hasselt, Belgium

[5]Centre for Health Economics Research and Modelling Infectious Diseases, Vaccine and Infectious Disease Institute, University of Antwerp, Antwerp, Belgium

[6]The Data Tank, Brussels, Belgium

[7]Data Science Institute, Universidad del Desarrollo, Santiago de Chile, Chile

[8]The GovLab, New York University, New York, NY, USA

[9]Interuniversity MicroElectronics Center, Studies in Media, Innovation and Technology, Vrije Universiteit Brussels, Brussels, Belgium

[*]these authors contributed equally

Corresponding Author:
Mattia Mazzoli, PhD
ISI Foundation, Via Della Rocca 20, Turin, Italy

Abstract

The COVID-19 pandemic served as an important test case of complementing traditional public health data with nontraditional data, such as mobility traces, social media activity, and wearable data, to inform real-time decision-making. Drawing on an expert workshop and a targeted survey of epidemic modelers in Europe, this study assesses the promise and the persistent limitations of such data in pandemic preparedness and response. We distinguish between “first-mile” challenges (obstacles to accessing and harmonizing data) and “last-mile” challenges (difficulties in translating insights into actionable policy interventions). The expert workshop, convened in March 2024 in Brussels, brought together 50 participants, including public health professionals, data scientists, policymakers, and industry leaders, to reflect on lessons learned and define strategies for better integration of nontraditional data into epidemic modeling and policymaking. The accompanying survey, gathering experiences from 29 modelers, offers empirical evidence of the barriers faced by modelers during the COVID-19 pandemic and highlights areas where key data were unavailable or underused. The experiences collected through the survey and workshop resulted in ten key actions and three overarching recommendations for public entities, data providers, and stakeholders. Our findings reveal ongoing issues with data access, quality, and interoperability, as well as institutional and cognitive barriers to evidence-based decision-making. Approximately 66% of all datasets had at least one access problem, with data sharing reluctance for nontraditional sources being double that of traditional data (30% vs 15%). Only 10% of respondents reported that they could use all the data they needed. These limitations included issues related to timeliness and granularity of data, as well as issues with linkage, comparability, and biases. To overcome these hurdles, we propose a set of enabling mechanisms, including data inventories, standardization protocols, simulation exercises, data stewardship roles, and data collaboratives. For first-mile challenges, solutions focus on technical and legal frameworks for data access. For last-mile challenges, we recommend fusion centers, decision accelerator laboratories, and networks of scientific ambassadors to bridge the gap between analysis and action. We argue that realizing the full value of nontraditional data requires a sustained investment in institutional readiness, cross-sectoral collaboration, and a shift toward a culture of data solidarity. Grounded in the lessons of the COVID-19 pandemic, the study can be used to design a roadmap for using nontraditional data to confront a broader array of public health emergencies, from climate shocks to humanitarian crises.

(J Med Internet Res 2026;28:e85540) doi:10.2196/85540

KEYWORDS

nontraditional data; pandemic preparedness; pandemic response; data science; epidemic modeling

XSL•FO
RenderX

Introduction

Background

Throughout the COVID-19 pandemic, decision-makers around the world looked for timely and quality information to inform their response efforts. Yet, much of the data traditionally used for public health (eg, public health–based surveillance data, health care–based data, and clinical trials) were not available fast enough or at the scale or the coverage needed for a crisis that extended beyond national borders and required a multisector, timely response. As a result, many in the scientific community turned to new, nontraditional data sources and cross-sectoral collaborations to fill these gaps and accelerate knowledge generation to support decision-making. We acknowledge the importance of mathematical, statistical, and artificial intelligence tools to interpret data and provide support for evidence-based policymaking. However, here, we explicitly focus on the two ends of the process, that is, the availability and retrieval of critical data and their power to inform public health decisions.

Nontraditional data (NTD) is often defined as "repurposed data that can be digitally captured (eg, mobile phone and financial data), mediated (eg, social media and digital traces), or observed (eg, satellite imagery)" [1-3], often using new and privately held technology [3]. As such, it is an umbrella term for information not originally collected for public health but repurposed for that purpose (Table 1). As such, feeds are often generated continuously and at a population scale, they offer more granular and more real-time data, helping public health responders, governments, and civil society organizations to respond more effectively.

Table . Properties of traditional and nontraditional data (NTD).

	Traditional data	NTD
Primary intended use	Population health	Business optimization, transport planning, environmental control, algorithm optimization, social behavior analysis, and flu-like illness monitoring
Collection method	Analog and digital	Digital
Coverage type	Groups targeted through recruitment, health care seeking, and notifiable diseases	Groups defined by digital service use and technology adoption - often scalable to the population level
Structure	Structured	Often unstructured
Most common biases	Reporting delays, underreporting, unequal health care–seeking behavior, geo coverage, testing capacity, representativity, unequal health access, heterogeneous reporting standards, and noise	Costly, infrequent sampling, recall bias, self-selection bias, representativity, digital divide, data ownership, low granularity, weak link to contact patterns, unstructured, demographic bias, privacy and interoperability, infrastructure sustainability, incomplete metadata, and heterogeneous quality
Example	Cohort studies, clinical trials, traditional contact tracing, consultation rates, and notifiable diseases	XDR[a] mobility data, participatory surveillance, social media data, and digital contact tracing

[a]XDR: extended detail record.

The potential of NTD to inform the response to public health emergencies has already long been recognized. For example, during the 2014 to 2016 West Africa Ebola outbreak, anonymized mobile phone call records were used to map population movements and shape travel restriction policies in Sierra Leone [4]. Similarly, Google search queries have been used to predict and offer early warnings on diseases, including Zika [5]. The use of NTD has been an important aspect of many of these efforts. Its use was further highlighted during the COVID-19 crisis, illustrating how such a rapid and granular data source can be used to inform the response to a public health crisis.

Typical NTD sources include aggregated and anonymized mobile location traces, information on retail purchases, social media posts, and data sourced from wearable devices, for instance, on heart rate or temperature [3]. NTDs have been classified into 4 main types: health, social mixing, economic, and sentiment data (Figure 1). Applications to pandemic response are exemplified below.

XSL•FO
RenderX

Figure 1. Nontraditional data types and sources. Schematic representation of the 4 types of nontraditional data and a sample of the main respective sources. Readers may refer to the text for a more extensive description of data sources and their use.

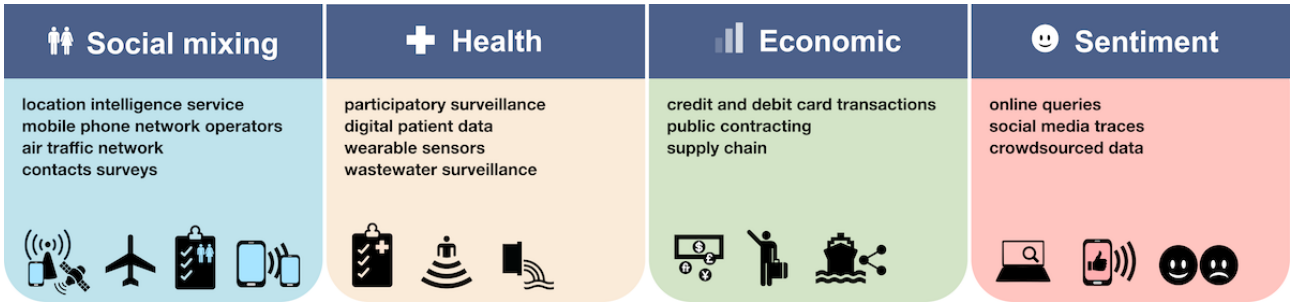

Health

At the core of pandemic response stands the monitoring of population health. At the beginning of the pandemic, researchers quickly repurposed preexisting crowdsourcing platforms to monitor COVID-19 circulation and related behaviors [6,7]. The participatory surveillance [8] platform InfluenzaNet [9] has collected flu-like symptoms and health-related behaviors from volunteers across Europe since 2003 to analyze country-level incidence trends on a weekly basis [10]. While volunteers report symptoms to their respective country's unique platform, all national platforms within the InfluenzaNet umbrella adopt standardized approaches to reporting, facilitating cross-country analysis. InfluenzaNet expanded its metrics infrastructure in 2020 to include COVID-19 data. Similar efforts are found in the United States with OutbreaksNearMe [11], in South Africa with CoughWatch [12], and in Australia and New Zealand with Flutracking [13]. In many cases, symptom reports are collected through mobile apps, as in the case of a mobile app [14] launched by King's College London and ZOE, a science and nutrition company. These data helped public health officials in the United Kingdom to identify early indicators of illness in the emergent stages of the pandemic.

Smart thermometers and other wearable sensors were a valuable data source for analyzing trends in fever in the population. Wearable sensor data [15] collected by Kinsa's network of smart thermometers [16] have been used to analyze fever patterns and develop outbreak forecasting maps across several countries. Thermometer data were already used in 2012 to forecast influenza outbreaks and were repurposed for COVID-19 during the onset of the pandemic. A wide range of stakeholders used these data, including pharmacies, schools, and other decision-makers. In Germany, wearable technology data collected by Thryve, in partnership with the Robert Koch Institute, created a data donation platform [17] for improved fever detection research in the spring of 2020. Among health-related repurposed data are news media records. While being focused on informing the general public of ongoing health threats, global parsers and visualizers, such as HealthMap, help collect and monitor news of disease outbreaks worldwide [18]. Using trusted social and news media, recent efforts have built open-source global datasets and mappers for spatiotemporal health threats surveillance [19,20]. At finer scales, wastewater-based epidemiology, which is based on sampling of sewage waters, proved to be a powerful tool during the pandemic to achieve a more granular and unbiased method to monitor disease circulation [21,22]. Wastewater-based epidemiology can detect viral traces of SARS-CoV-2 among other biological and environmental contaminants and reflect transmission dynamics in untested communities [23], allowing for earlier detection of outbreaks. In Hong Kong, wastewater surveillance efforts found the first evidence of the spread of the Delta variant and informed public health interventions [24].

Social Mixing

On a geographic scale, mobility patterns encode the mixing of residents of different locations, allowing modelers to estimate the spatial spread of the disease and suggest spatially targeted interventions. At the population level, contact data encode interactions at risk of contagion, allowing modelers to study disease transmission across demographic groups and suggest setting-specific interventions. Mobility records were published regularly after the early days of the pandemic through *Data for Good* [25] programs at diverse spatial scales and temporal resolutions. Google, for instance, regularly published human mobility data through its Community Mobility Reports [26]. These data regarded worldwide regions and included trends around commonly visited location types (eg, retail, workplace, and residential areas) that, when combined with census data, helped inform epidemic modeling to support decision-makers. Social media companies' *Data for Good* programs, such as Meta [27] and Cuebiq [28], assisted epidemic modelers in providing critical insights for public health measures and population studies. Telecommunications companies [29] played an important role in tackling the pandemic in their respective countries, helping modelers and public health practitioners monitor and understand the heterogeneous population response to restrictions, as in the case of Telefonica Chile [30] and Orange France [31]. These data helped decision-makers assess the effectiveness of their policies, monitor population compliance, and tailor the public health responses. Beyond mobility patterns, social mixing data were typically collected from traditional contact surveys and diaries, which did not scale well at the national level and did not offer longitudinal or real-time follow-up of human behavior. Following prepandemic efforts such as POLYMOD [32], online contact surveys reemerged in the pandemic's early days thanks to new crowdsourcing tools at the European level, such as CoMix [33], an initiative funded by the European Commission (EC) that quickly gathered public mixing and sentiment data during the COVID-19 pandemic. The weekly survey started in March 2020 in Belgium, the Netherlands, and the United Kingdom before expanding to 17 additional countries across Europe. Marketing firm Ipsos MORI recruited participants via social media advertisements and email

XSL•FO
RenderX

campaigns to crowdsource reported social contacts, assisting with national policy evaluation.

Economic

The pandemic shifted consumers' needs and habits, impacting local businesses' revenues and opening new market opportunities. Economic NTD helped policymakers monitor the impact of the pandemic on local businesses and the population's spending habits. Data on electronic transactions [34] provide information on the time, location, and purpose of people's purchases, proxying their economic wealth. Electronic transaction data from credit cards were used to measure the reduction in revenue [35] of economic activities during the COVID-19 pandemic.

Supply chain data allow monitoring the status of the production and distribution of goods, providing logistics organizations with important insights to optimize the goods pipeline. Using maritime traffic data collected via a global network of Automatic Identification System receivers, the Marine Traffic Research Lab studied the impact of the pandemic on the shipping industry [36]. Open contracting records refer to records of public contracts from administrations. They are used in many countries to increase transparency and combat corruption in public spending. In Ecuador, the National Public Procurement Service launched a public-facing dashboard of emergency procurement contracts during the COVID-19 pandemic [37].

Sentiment

While aiming at controlling disease circulation, interventions had an impact on population mental distress and were partially received by citizens, giving rise to misinformation exposure and loss of adherence to policies over time. Sentiment NTD allowed researchers and organizations to monitor people's perceptions, mental health, and exposure to misinformation. Online search queries and social media data were valuable tools to understand the impact of online behavior on the pandemic progression and monitor the population's perception of restrictions. Google search trends and tweets were used by German researchers to create a future-oriented early alert system [38] that could be used to build phenomenological and mechanistic models to forecast cases and hospitalization trends. Tweets before and during the pandemic were analyzed by researchers in Italy to analyze the potential health threat driven by online misinformation [39] related to the pandemic. The same data were used by researchers in the United States to integrate public health data with social media data, developing an early alert system for outbreak detection [40]. However, sentiment data do not stop at online discourse on social media; the worldwide COVID-19 trends and impact survey [41] conducted by Meta collected important information, among others, on the population's mental health. Over the course of the pandemic, respondents transmitted self-reported feelings of anxiety and depression, for the first time collecting important insights on the population's mental distress and risk perception during a pandemic. While CoMix [33] collected both self-reported behavioral data and sentiment from participants, other efforts aimed to incorporate a wider variety of data streams related to people's perception of the pandemic. For example, 6 European public health agencies collaborated during the pandemic to create PANDEM-Source [42], an IT surveillance tool that integrates both traditional case counts and nontraditional social media reports for pandemic surveillance. Results feed into PANDEM-2 [43], a real-time data dashboard for decision-makers and health experts.

NTD can help fill critical gaps in traditional data during public health emergencies by increasing the speed and volume of data collection. When combined with traditional data, NTD can provide a more comprehensive understanding of the problem and support rapid decision-making. However, alongside the promise, COVID-19 also revealed challenges involved in the use of both traditional data and NTD. The early months of the COVID-19 pandemic were characterized by ad hoc collaborations, which revolved around the use of NTD sources: mobility dashboards built overnight [44], wastewater pilot studies [24], and hastily negotiated data sharing agreements [45].

A noteworthy example of ad hoc collaboration between public bodies and the private sector was the EC's collaboration with European Mobile Network Operators (MNOs) [46]. Within just a few months, 17 MNOs across 22 EU states and Norway provided access to mobile phone data with the aim of understanding the geographical spread of the disease and assessing pandemic interventions. In addition to European-level initiatives, many subnational regions, such as the Valencian region of Spain [47], combined data from mobile operators with data from Facebook, Google, and government ministries to evaluate the impact of mobility restrictions on COVID-19 spread. Challenges that emerged from these pioneering initiatives included data harmonization and privacy guardrails.

Even after the pandemic, we can find virtuous examples of public-private collaborations allowing NTD collection aimed at making data publicly accessible for research. One institutional example is represented by the initiative of the Transport Ministry of Spain [48], regularly making publicly available origin-destination matrices [49] of mobility flows at high temporal and spatial resolution in Spain since early 2020. This ongoing initiative is a collaboration between the Ministry, Orange Spain, and Nommon and provides a curated and poststratified, ready-to-use dataset of human mobility stratified by multiple sociodemographics. On a more bottom-up side, the community of researchers is also responding to the growing lack of NTD accessibility. This is an example of the R2M2P2 Consortium: Readying Regional Mobility data for Modeling Pandemic Preparedness [50], led by the Erasmus Medical Center in collaboration with Dutch universities and the national public health agency RIVM. The aim of this effort is to identify useful mobility data sources and set up a pipeline providing usable and accessible mobility data to inform transmission models to respond to key public health questions on the spread of respiratory infections in the Dutch population.

These innovations helped highlight how diverse data streams, when responsibly harnessed, can quicken detection times and potentially reduce mortality and morbidity. These experiences served as an important test case for the integration of these new sets of data into epidemic modeling and decision-making for pandemic response. However, they also exposed deep structural

XSL•FO
RenderX

weaknesses, such as fragmented governance, limited interoperability, and a chronic skills gap. Questions remain around how these data can best be used in a pandemic crisis, namely, regarding their limitations and the investment needed to replicate efforts from COVID-19 in future public health emergencies. A better understanding of these challenges may provide insights and lessons for future efforts to combat public health emergencies.

Traditional data exhibit known biases (eg, as reporting delay, geographical coverage, and health care accessibility disparities among others) that modelers typically take into account using Bayesian inference, sample poststratification, and artificial intelligence–based synthetic data generation techniques. However, NTDs are not exempt from biases and have their own limitations, most commonly being sample bias resulting from technology adoption divide, data generation noise due to self-reporting bias and unequal representativity among population groups, unequal spatiotemporal coverage, and poor granularity. In Table 1, we summarize the most common biases for both data types. Modelers adopt similar techniques to those used for traditional data to account for NTD biases, but effective approaches are case specific and data specific. Beyond the focus of the survey and workshop on the challenges of accessing and integrating NTD into evidence-based policy, recent studies have mapped and addressed thoroughly known biases of NTD [2,51-53].

This study is based on the structured discussions held during an expert workshop on the use of NTD to counteract pandemics conducted in March 2024 in Brussels, along with the results of a survey on data readiness and availability disseminated among the European modeling community conducted the year before. The workshop convened public health professionals, data scientists, policymakers, industry leaders, and other stakeholders to reflect on lessons learned and define strategies for better integration of NTD into epidemic modeling and policymaking. The accompanying survey offers empirical evidence of the barriers faced by modelers during the COVID-19 pandemic and highlights areas where key data were unavailable or underused. Together, these 2 sources form the basis for 10 key lessons learned and proposed actions about the role of NTD in pandemic preparedness and response, which we outline below.

On the basis of the experts' opinions and the conducted survey, we draw a set of recommendations that establish a framework for the use of NTD in ongoing and future public health and other crises. From the survey responses, 2 main categories of obstacles emerged: "first-mile" and "last-mile" challenges. First-mile challenges refer to the difficulties of rapidly identifying and accessing relevant data when a crisis begins; these include legal constraints, unclear stewardship, and difficulty in the application of the FAIR [54] principles, that is, lack of interoperability and, critically, unavailability of data. Last-mile challenges are more nuanced and often less recognized but are nonetheless critical. They concern the difficulty of translating complex data into actionable insights that decision-makers can understand, trust, and apply. Together, these challenges represent a formidable and persistent obstacle to the use of NTD in public health emergencies. To move beyond improvisation during a crisis, we argue for a more proactive, institutionalized, and scalable approach to using NTD streams before, during, and after public health emergencies.

Survey on Data Availability and Data Needs

In 2024, we conducted a survey [55] to better understand data availability, use, and unmet needs during the COVID-19 pandemic, primarily in the European modeling community. Specifically, the survey aimed to gather information about which types of data respondents drew on during the COVID-19, obstacles and challenges they faced, and whether and how these data were used to shape policy.

The survey was divided into 3 sections. The first section aimed to collect the research questions and modeling approaches used by researchers participating in the survey. The second section aimed to identify (1) what data were available and were used during the pandemic, (2) how data were accessed and what problems were encountered in accessing the data, and (3) what quality problems were encountered when using the data. Finally, the third section aimed to understand what data were needed but not used by these research groups and whether these unmet needs were the result of data unavailability or the result of data quality and access issues. Throughout the survey, the data were categorized into traditional epidemiological data (eg, tests, hospitalizations, deaths, and vaccination) and nontraditional epidemiological data, that is, data types not collected with epidemiological purposes and/or relatively new data types (eg, mobility, contacts, and wastewater data), so that in addition to each individual data type, we can have a broader view on traditional versus NTD. Overall, the survey was composed of 11 questions, of which 10 were closed-ended questions, with the last open question regarding the purpose of the data use.

Using a purposive sampling approach, we circulated the survey through the ECDC Modeling Hubs mailing list, a community populated by many public health emergency responders during the COVID-19 pandemic and the ESCAPE consortium. The aim was to target as many epidemic modelers as possible in Europe. In total, 29 experts (individuals and groups) from academia, government agencies, and the private sector responded to the survey.

Pandemic Response as a Hybrid Enterprise

In this section, we present the results obtained from our surveyed sample of the modeling community in Europe. All reported statistics are purely descriptive and reflect the responses of our sample; as such, we do not infer statistics or generalize to broader populations.

The survey responses suggest that NTD sources have moved from fringe to mainstream, indicating that pandemic modeling and response among participant modelers have in many ways become a hybrid enterprise, in which NTD often complement and augment more traditional (if somewhat slower) sources. For instance, 90% of enquired modelers reported using at least one nontraditional stream alongside more conventional sources.

XSL•FO
RenderX

Among nontraditional data types, mobility was dominant (72% of respondents), alongside other nontraditional sources such as socioeconomic and demographic, contact, and wastewater data (Figure 2). An equally striking aspect of the survey finding was the direct relevance of such data to inform the response to the pandemic. Seventy-nine percent of the respondents stated that they used data to inform public health policies, with the largest number of these (83%) stating that they used mathematical models to translate data into insights used to inform policymaking, followed by statistical models (48% of respondents). More specifically (and in addition to recent work conducted by the GovLab on identifying use cases for the use of NTD [56]), respondents reported their need for NTD to achieve the following goals:

- monitoring adaptive behavior to interventions across sociodemographic groups
- assessing the age-specific impact of interventions on disease transmission
- understanding the role of socioeconomic factors in disease spread
- better informing and validating models assumptions
- better calibrating complex models, particularly behavioral components
- improving short-term predictions at regional and local scales
- assessing the risk of health care system saturation
- better tracking the spatial invasion patterns of disease
- identifying the most infected areas and high-risk population groups with finer higher granularity
- tracing transmission chains and assessing estimating the distributions of secondary cases
- broadening and sharpening the scope of epidemiological studies

Moreover, NTDs were needed to assess the risk of health care system saturation, better track the spatial invasion of disease, assess most infected areas and population groups with higher granularity, trace transmission chains, and assess distribution of secondary cases.

Figure 2. Comparison of the use (x-axis) and the need (y-axis) of the different types of data reported by respondents. Vertical and horizontal gray lines represent the mean number of answers for all data types used and needed, respectively. The upper-right quadrant represents data highly used and needed; the upper-left quadrant represents data highly needed but not highly used; the bottom-right quadrant represents data highly used but not highly needed; and the bottom-left quadrant represents data neither highly used nor highly needed. Reported values are the result of closed-ended multiple-choice questions. ICU: intensive care unit; ILI: influenza-like illness.

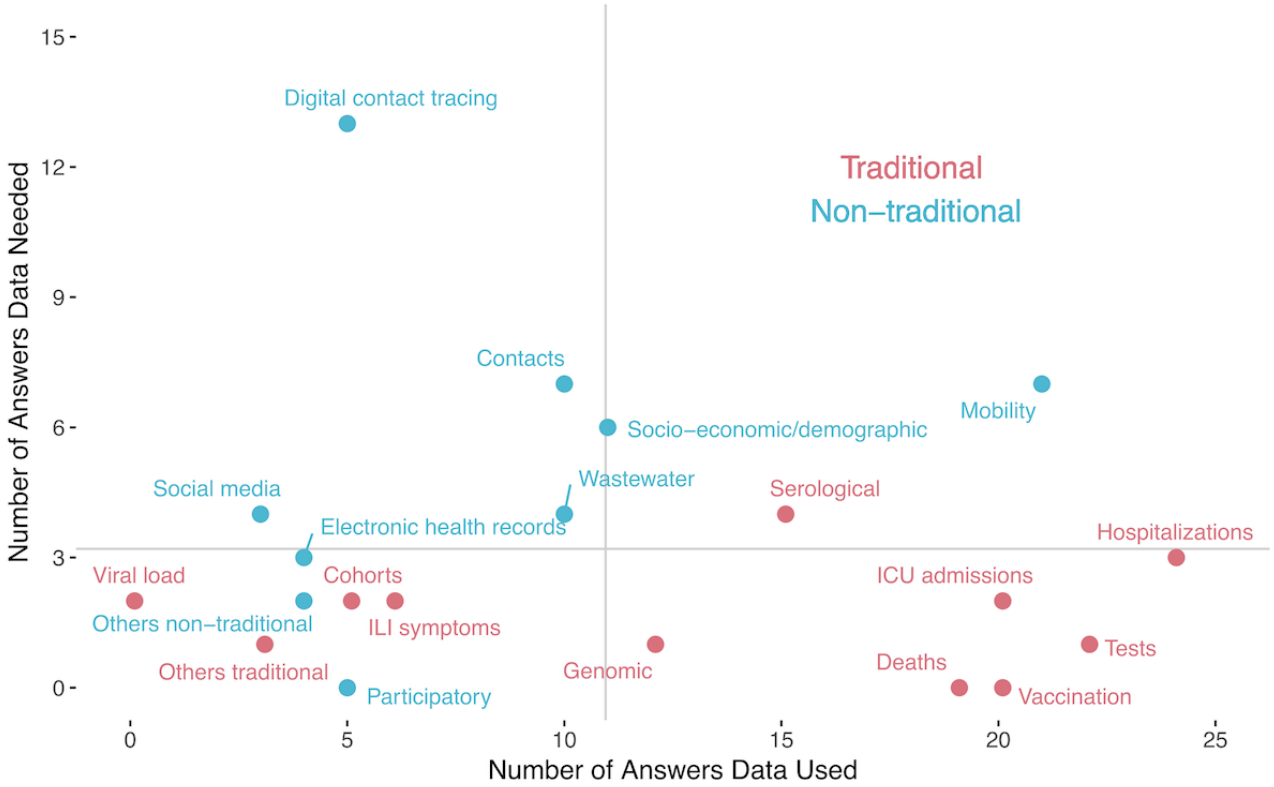

Workshop on First- and Last-Mile Challenges for Data Needs

On March 22, 2024, we held a workshop at the University Foundation in Brussels to better understand the role (including limitations) of NTD. Attendance at the Workshop was upon invitation. We assembled a high-level expert panel reflecting the workshop's interdisciplinary goals, bringing together approximately 50 participants from diverse sectors:

- Private sector representatives from industries, such as telecom companies, which typically own and provide relevant NTD
- Academic sector leading European scientists and project partners who played pivotal roles during the pandemic response

XSL•FO
RenderX

- Public health representative officials from national public health institutes (eg, Portugal and Switzerland) and European institutions (eg, HERA)

Attendees were selected based on their expertise and background. Figure 3 shows the provenance of the invited participants.

Figure 3. Workshop participating organizations. Pie chart representing the sector of the 50 participants in the workshop. See the Acknowledgments section for the full list of participants and affiliations.

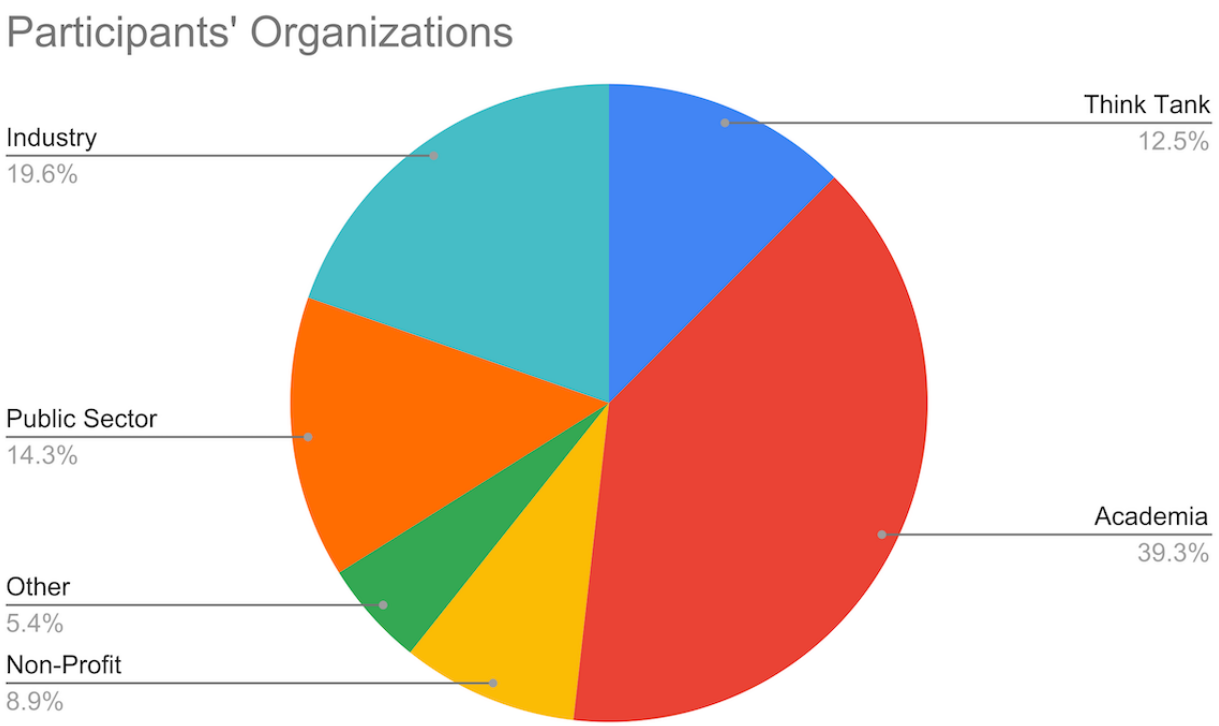

The workshop's specific objectives included the following:

- Advancing the dialogue on leveraging NTD for pandemic preparedness
- Drawing insights from experts and practitioners in relevant fields
- Building meaningful collaborations to enhance pandemic response efforts
- Co-creating high-level recommendations and a roadmap for European decision-makers

Discussions and activities were structured around 2 main themes: "Data for Preparedness" and "Data for Pandemic Response." Participants were divided into working groups to explore these topics through 3 key sessions.

1. First-mile challenges: addressing readiness to access NTD in case of a pandemic. Questions included were as follows: Are we prepared to access and reuse NTD for preparedness and response? What challenges remain from the previous pandemic?
2. Prototyping first-mile solutions: This session focused on refining and defining challenges, gathering quick feedback, identifying areas with the greatest impact or feasibility, and proposing concrete actions and recommendations.
3. Last-mile challenges: Exploring how to ensure data insights are translated into actionable decision-making. Discussions centered on identifying key decision-makers, facilitating partnerships between data, scientific, and policy communities, and fostering effective communication.

We adopted this established framework [57-61] as an organizing principle for the workshop. In pandemic response, the "first mile" refers to the ability to quickly obtain both traditional data and NTD and make them available for reuse. The "last mile" focuses on converting that data-driven knowledge into actionable insights and ensuring it is effectively applied by policymakers and public health authorities.

The final plenary session united all participants to review and discuss the solutions proposed during the previous sessions. Discussions in all sessions were reported by in-person note takers.

Difficulties and Challenges Reported by Participants to the Workshop and Respondents to the Survey

Overview

The pandemic revealed some significant challenges that remain unsolved. Both the convened experts and the survey aimed to better understand the nature of these challenges and how they might be overcome to do better in future health emergencies.

One of the overarching findings of the survey is that the use of data [62] to inform the response to COVID-19 and other public

XSL•FO
RenderX

health emergencies can be broken down into 2 main challenges: access, in which data are identified, sourced, and processed; and translation, in which data are used for policymaking and meaningful interventions. We call these first-mile and last-mile bottlenecks (see Figure 4 for a summary of the two). We used the same division to shape the structure of the workshop sessions, as this is a relevant and useful framework to analyze the bottlenecks of data pipelines in pandemic response. First-mile bottlenecks often get the lion's share of attention; however, last-mile shortcomings are equally significant in the way they impede the conversion of insights into timely action. In other words, access without translation is potentially as limiting as a simple lack of access.

Figure 4. Proposed actions addressing first- and last-mile challenges. Schematic figure illustrating the 10 proposed actions summarized within the first- and last-mile framework.

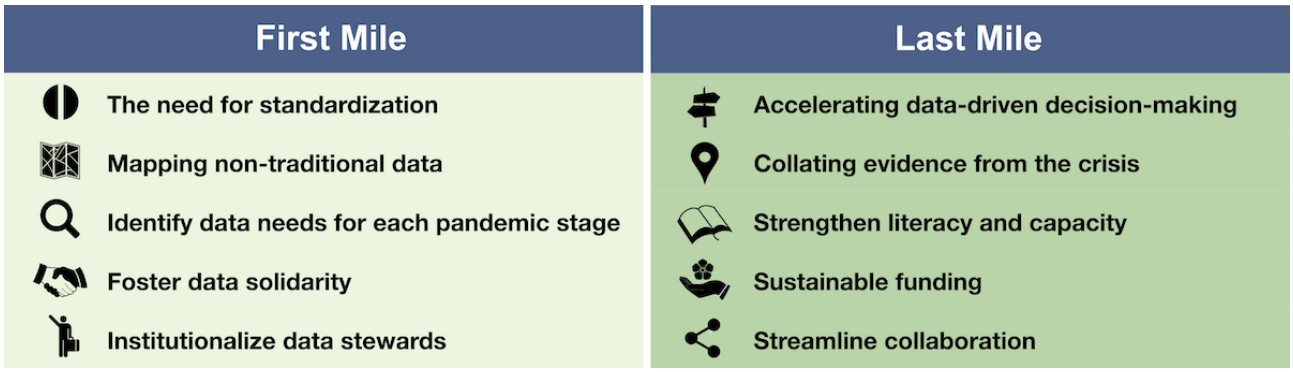

"First-Mile" challenges

First-mile hurdles can arise at any stage of a public health crisis. They refer to difficulties responders face in identifying and accessing relevant data streams [63]. Obstacles can include opaque ownership and licensing, a lack of existing data sharing agreements, missing metadata, and quality issues. Technical fragmentation—including nonstandard formats and incompatible APIs—is another impediment. Two key obstacles highlighted by workshop participants and survey respondents included a lack of standardization and a shortage of incentives for sharing, leading to what the experts labeled reduced "data solidarity" [64].

The survey quantifies some of these pain points. According to experts, approximately 66% of all datasets used by respondents—traditional and nontraditional—had at least one access problem (S1 Fig in Multimedia Appendix 1). The proportion of issues linked to data sharing reluctance in the context of NTD was double that of traditional data (30% vs 15%; see S1 Fig in Multimedia Appendix 1). Furthermore, only 55% of nontraditional streams were publicly available or obtainable via government agreements, compared to more than 83% for traditional data (S2 Fig in Multimedia Appendix 1). In addition, data quality problems were also prevalent. Despite being rare, preparation and processing issues existed across both traditional data and NTD (12% vs 10%; S3 Fig in Multimedia Appendix 1). Only 4% of the datasets that participants reported (both traditional and nontraditional) had no quality problems; perhaps somewhat surprisingly, machine readability [65] issues were far more prevalent with traditional data compared to NTD (11% vs 2%). These quality issues may stem from the fact that individuals and organizations are less familiar with NTD requirements than traditional data. As most *Data for Good* programs were ceased after the pandemic, a further obstacle is found in data impermanence, which hinders data reusability and makes data pipelines more fragile in the face of new threats.

"Last-Mile" Challenges

Last-mile issues surface at the opposite end of the data-to-policy pipeline. They hamper the process of turning analysis into actionable intelligence and meaningful interventions. Once again, there are multiple drivers of these limitations, including gaps in data and technical literacy among officials, weak institutional bridges between scientists and policymakers, and slow feedback loops. Mismatches between the types of data available and requirements for policymaking can also play a role. As we have elsewhere argued [66], taking a question-based approach to evidence-based interventions can help align data with the most pressing and relevant policy issues.

At a broad level, the survey suggests a degree of success on this front. As noted earlier, 79% of respondents stated that their analyses fed directly into public health policy. Nonetheless, a further examination of the survey results suggests a number of obstacles that prevented data from actually being usable (S4 Fig in Multimedia Appendix 1). Besides data availability, these include limitations in the timeliness and granularity of data, issues with linkage [67], comparability, and biases. Only 10% of respondents reported that they could use all the data they needed. This was particularly relevant for NTD (eg, contact tracing [68], mobility, and contact data) and reflects a broader problem with data availability, which remains a significant barrier for data usage in policy making (S4 Fig in Multimedia Appendix 1).

Ten Lessons Learned and 10 Proposed Actions

In this section, we outline key proposed actions in response to 10 lessons learned that emerged from the workshop and survey addressing first- and last-mile challenges.

The Need for Standardization

Participants at the workshop stressed how the experience of reusing NTD during the COVID-19 pandemic revealed

XSL•FO
RenderX

widespread inconsistencies in data formats, metadata, and quality metrics. These issues made it difficult to integrate and analyze data across sources and impeded timely insights. The expert workshop emphasized the urgent need to establish common standards not only for the data itself but also for metadata structures, reproducibility, and quality indicators. In particular, the lack of standardized approaches [69,70] to processing origin-destination matrices [71] in mobility data emerged as a significant obstacle for modeling [72] and understanding epidemics [73]. One suggestion was the establishment of a multisectoral task force to create a common set of questions and methodological standards, enabling smoother collaboration before the next crisis.

One virtuous example is the effort carried out through the MNOdata4OS and Multi-MNO project [74]. This initiative, developed by the European Statistical System [75], aims to create a consistent methodology for processing MNO data for official statistics throughout Europe. The project also includes defined quality standards and an open-source reference implementation of the outlined methodology. Notably, this ongoing work seeks to transition MNO data usage from Experimental Statistics [76] to official statistics by addressing challenges of data privacy, creating consistent legislation for long-term data access partnerships, and implementing standardized methodologies.

Accelerating Evidence-Based Decision-Making

Advanced analytics mean little if they do not feed into policy; as we have seen, the last-mile challenges remain significant. The workshop, therefore, endorsed three “bridges” that could help advance evidence-based policy: (1) fusion centers [77], to merge and make actionable data streams; (2) decision accelerator laboratories [78], to prototype interventions and iterate on policy; and (3) a network of scientific ambassadors trained to translate data insights into clear-language policy and more generally to bridge science and practice. Together, these mechanisms can help move from insight to action more quickly. The group also recommended building a cross-border European community of epidemic modelers to expand access to evidence and ensure that no region is left behind in future emergencies.

Mapping NTD

Discussion at the workshop highlighted how during the COVID-19 pandemic, as with prior and subsequent public health emergencies, many stakeholders were often not aware of NTD. Participants stressed this as a critical barrier and emphasized the need for a “data map,” that is, a comprehensive data inventory describing what datasets exist and rating them for accessibility, relevance, and quality. They also underlined the need for regular audits and harmonized metadata templates so that such an inventory would be searchable and valuable across a range of use cases.

The EU Health Data Space (EHDS) [79], currently collecting health data only, should begin to absorb such demand by building common standards, practices, and governance for health data reuse in research, innovation, and policymaking, integrating NTDs that are essential for fast public health response to health crises. Adopted in January 2025, the EU Health Data Space proposes HealthData@EU [80], an ecosystem of national platforms that establishes standards for data quality and use and facilitates the annual publication of data catalogs to find and use common data types across borders.

Collating Evidence From the Crisis

Participants at the workshop stressed that while hundreds of initiatives during the COVID-19 pandemic used mobility, payments, or social media signals [3], lessons learned remain scattered and difficult to generalize.

The workshop highlighted the need for more research on how NTD is used, during which phases, and with what impact. Such a repository would include not only project descriptions but also tags, pipeline schemes, and metadata that allow sorting by disease stage, data type, and outcome at the level of intervention effectiveness based on these data. In this way, future public health professionals and researchers could quickly identify what worked, avoid redundancy, and inform fast and effective interventions while avoiding previous mistakes.

Identify Data Needs for Each Pandemic Stage

We learned how identifying and aligning data needs with evolving goals at different stages of a pandemic is essential. Workshop participants highlighted the value of tabletop simulation exercises as a practical tool to explore evolving data requirements. The panel of experts suggested tabletop simulations to provide simulated environments [42] to explore hypothetical pandemic scenarios, revealing dynamic data requirements at various stages of public health emergencies. In doing so, they help policymakers and other stakeholders assess preparedness and response strategies, as well as identify critical data needs and data streams ahead of time.

Strengthen Literacy and Capacity

We have noted a number of last-mile hurdles that limit policymakers’ capacity to transform NTD into actionable insights. In addition to those already listed, the challenges also include a significant skills gap among public health officials and other stakeholders. To address this, the workshop expert panel called for investments in both analytical training (eg, interpreting mobility or social media data) and “question literacy” [66]. Question literacy is an essential but underappreciated skill. It aligns data supply with data needs and can help prioritize scarce public resources and policymaker bandwidth (eg, by minimizing time spent searching for or analyzing unhelpful or unnecessary data).

Foster Data Solidarity

A recurring theme of the workshop was the need to foster a culture of “data solidarity,” that is, the idea that data holders, especially in the private sector, should commit to sharing data responsibly and proactively during public health emergencies. This requires not only ethical and legal frameworks but also the cultivation of mutual trust and shared purpose among stakeholders and across sectors. Practical manifestations of data solidarity could include prenegotiated commitment frameworks, increased transparency around how data will be used, and the development of social licenses [81] through public engagement. In addition, data collaboratives [82], that is, expert groups based

XSL•FO
RenderX

on public-private collaborations exchanging data to help solve public problems, and similar bodies have proven effective in promoting trust and data sharing. Across all these mechanisms, it is essential to establish incentives for sharing; these could include naming and framing schemes and access to aggregated insights for strong sharers.

Institutionalize Data Stewards

The expert panel stressed that even when the incentives and structures for data solidarity exist, coordinating actual data sharing is a complex task, especially in the midst of a public health emergency. The workshop suggested the creation of dedicated "data stewards" [83] to help institutionalize and facilitate this process. Data stewards are individuals or groups responsible for building partnerships, conducting audits and risk assessments, facilitating internal coordination, and communicating externally with stakeholders. Their role has proven especially effective in nurturing and sustaining data collaboratives [82]. By making this a formal, resourced role, organizations can ensure they are better prepared to access and share data when a crisis occurs.

Sustainable Funding

Long-term, sustainable funding and investment are critical to contrasting data impermanence, promoting data sharing and effective use of NTD. One of the clearest lessons from the COVID-19 crisis is that ad hoc or one-off funding mechanisms are inadequate and that long-term capacity requires sustained investment in infrastructure, human capital, and technology. Toward this end, the expert group recommended establishing a dedicated data fund, that is, public-private partnerships supporting data availability for evidence-based policies. The fund should be governed transparently, include mechanisms for rapid-response grants, and build in accountability through independent oversight.

Streamline Collaboration Through Advanced Technologies

The expert group highlighted how policy, norms, and society are vital parts of any framework to promote the use of NTD. However, it should not be overlooked that advanced technologies themselves can play a transformative role, especially in encouraging more data sharing. Participants suggested considering technologies such as privacy-enhancing tools [84], Edge Computing [85], Data Sandboxes [86], Data Pods [87], and synthetic data [88], including social contacts [89] and mobility [71], beyond electronic health records. Many of these underlie, the emerging concept of "data spaces"—structured, rule-based environments in which data holders retain control over how their data are shared and used [90]. By investing in these technologies, governments and public health institutions can significantly strengthen their ability to collaborate while still maintaining public trust and protecting privacy and other rights.

Conclusions

Overview

The 10 lessons outlined earlier, as well as the survey results, point to a set of broader priorities for policymaking. In this final section, we distill the experience of COVID-19 into 3 overarching recommendations that can inform not only future responses to public emergencies but also steps that governments and other stakeholders need to take today to prepare for the next pandemic. Taken together, these recommendations emphasize the need for sustained investment, institutional change, and a more collaborative, integrated approach to the reuse of NTD in public health emergencies.

Institutionalize Readiness

Across all themes, a central theme emerges: success during the next crisis will not be determined by improvisation or timely responses, but by the foundations we lay today. Being better prepared means building durable technical, organizational, and human capacity. Governments and other stakeholders must work today to formalize roles such as data stewards, establish data sharing agreements, and invest in tools (eg, Fusion Centers [77], Decision Accelerator Labs [78], and simulation exercises) that can be activated quickly when they are needed. Readiness should be seen as a continuous state, supported by ongoing trust building and community involvement. Some specific actions include (1) establishing and institutionalizing data stewards, empowered with formal mandates; (2) creating and funding fusion centers and decision accelerator laboratories to prototype, evaluate, and scale interventions based on nontraditional (and traditional) data; and (3) conducting regular tabletop simulation exercises involving all stakeholders to enhance readiness and identify weak points or challenges.

Build Data Preparedness

Institutional readiness must be accompanied by data preparedness. Once again, many of the most pressing challenges during the COVID-19 pandemic stemmed from the simple fact that public health representatives were not ready: data were inaccessible and fragmented, and their nature and location were poorly understood. Building data preparedness requires developing standardized approaches to how data are structured, stored, and analyzed—especially for NTDs such as mobility traces or social media signals. Importantly, there is a need to map what data exists and where it resides. Specific actions include (1) developing data inventories and maps to identify nontraditional datasets, assess their accessibility and quality, and ensure metadata are harmonized across sources; (2) launching a multisectoral task force, including epidemic modelers, public health experts, and social scientists to set and maintain data standards for behavioral data integration into models, including protocols for metadata; and (3) create an evidence repository to collate and tag successful (and unsuccessful) uses of NTD during the COVID-19 pandemic to increase data literacy.

Create a Trusted and Collaborative Ecology

Data do not exist in a vacuum. Its value is realized through relationships among sectors, stakeholders, and, importantly,

XSL•FO
RenderX

between those who hold data and those who need it. Trust is the foundation of these relationships and must be deliberately cultivated. This includes establishing clear ethical guidelines for data use but also developing institutional intermediaries such as data stewards and data collaboratives that can manage partnerships and facilitate reuse. An observatory can play a particularly critical role, helping to monitor how NTDs are used in real time, documenting best practices, and ensuring that insights are shared across borders and sectors. Building this ecology will require a cultural shift, from data hoarding to data solidarity.

Specific actions include (1) developing mutual commitment frameworks between private data holders and public entities that may use data to clarify terms of emergency data sharing ahead of crises; (2) creating an observatory for NTD use that tracks deployment, use, and disseminates real-time insights and lessons learned; and (3) encouraging and scaling data collaboratives with clear governance rules and mechanisms for shared infrastructure and trust building across sectors and stakeholders.

Together, these 3 recommendations offer a framework, not a set of independent guidelines or policies. None of these can succeed in isolation, but if pursued together, they can create a virtuous cycle of preparedness, insight, and action.

Crucially, these steps are not only relevant to future pandemics. They can also inform how societies respond to a range of public crises, from climate shocks to humanitarian emergencies. Effective responses to each of these increasingly rely on prepared, tested, informed, and inclusive decision-making.

The COVID-19 pandemic was an indisputable calamity. Nonetheless, it was also a moment of extraordinary innovation, experimentation, and, most importantly, learning. If we absorb its lessons and begin now to build the foundations we lacked, then we may yet emerge from this crisis more resilient and better prepared next time. We may not be able to prevent the next crisis—but we can certainly improve our response and, in so doing, potentially mitigate the scope for another calamity.

Limitations

This work provides exploratory insights from a purposively sampled community of European epidemic modelers rather than generalizable findings from representative research. The survey was distributed through the ECDC Modeling Hub mailing list and the ESCAPE project networks, reaching modeling teams actively engaged in COVID-19 response efforts. Of the contacts reached, 29 modeling teams responded. All survey results were purely descriptive and reflect only the experiences and perspectives of this specific sample—no inferential statistics or population-level generalizations should be drawn from our findings.

Our recruitment strategy inherently created selection bias toward modelers already familiar with and interested in NTD. Survey participants were recruited through networks focused on epidemic modeling, and workshop participants self-selected to attend an event specifically about NTD integration. This likely resulted in confirmation bias, where reported experiences and needs reflect a community already convinced of the value of NTD sources. Our findings should therefore be interpreted as perspectives from early adopters and advocates of NTD, not as representative of the broader modeling community. The barriers and needs identified may be even more pronounced among modelers not yet engaged with these data sources.

The workshop discussions were captured through structured note-taking by designated members of the organizing team rather than through formal transcription or qualitative coding methods. While this approach is standard for policy-oriented workshops, it introduces subjectivity in how themes were identified and synthesized into our 10 lessons learned. These lessons represent expert recommendations informed by survey data and workshop discussions, not findings that emerged purely from systematic data analysis. All ESCAPE consortium members (a substantial portion of workshop participants and co-authors of this manuscript) reviewed and validated the synthesis, but this does not eliminate the interpretive nature of the process.

Our findings reflect primarily European perspectives and the specific context of the COVID-19 pandemic response. Experiences from other regions may reveal different lessons and suggest alternative recommendations. Similarly, the challenges and opportunities associated with NTD may differ for other types of health emergencies (eg, climate-related health events, humanitarian crises, and endemic disease surveillance).

Nontraditional data sources are not exempt from biases and have their own limitations, including technology adoption bias, noise from self-reporting, unequal representativity among population groups, unequal spatiotemporal coverage, and data ownership and privacy issues. Traditional data exhibit known biases (eg, reporting delays, geographical coverage disparities, and health care accessibility differences) that modelers typically address through Bayesian inference, poststratification, and synthetic data generation. Similar techniques can be applied to NTD, although effective approaches are case specific and data specific.

Despite these limitations, our findings and recommendations are formulated to be applicable beyond the specific context of COVID-19 modeling in Europe. The data types discussed and actions proposed are not disease-specific or geographically constrained, although their implementation may require adaptation to local institutional contexts. This work should be viewed as a starting point for broader discussions on improving data-to-policy translation in public health emergencies, not as definitive guidance derived from representative research.

Acknowledgments

Workshop participants, namely:

XSL•FO
RenderX

André Peralta Santos Health Directorate Portugal, Acting Deputy Director, General for Health

Arnon Vandenbergh, King Baudouin Foundation, Data Scientist

Arnout Desmet, Blauwe Cluster Sustainability Advisor

Brecht Ingelbeen, Institute of Tropical Medicine, Antwerp Researcher

Casey Weston, LinkedIn - Data for Impact, Program Senior Manager

Christophe Fraser, University of Oxford - Nuffield, Dept. of Medicine, Professor of Epidemiology

Ciro Cattuto, ISI Foundation, Scientific Director

Constantino Caetano, Portuguese National Health Institute Doutor Ricardo Jorge, Ph.D. Student

Daniela Paolotti, ISI Foundation, Research Leader

David Dab, Microsoft, National Technology Officer

Denis Xavier Renaud, Orange Head of Orange Flux Vision R&D

Emma Hodcroft, Swiss Tropical and Public Health Institute, Assistant Prof. and leader of the Epidemiology and Virus Evolution group

Esteban Moro, Northeastern University, Professor in Physics

Francesco Parino, INSERM, Post-Doc Researcher

Helena Coning, Mastercard, Chief Privacy & Data Responsibility Officer

Hien Vu, CEPS, Associate Researcher

Ilaria del Seppia, European Commission - HERA.03, Policy Officer

Ingmar Weber, Saarland University Alexander von Humboldt, Professor for AI and Chair for Societal Computing

Irma Varela-Lasheras, Portuguese National Health Institute Doutor Ricardo Jorge, Post-Doc Researcher

Jasmin Menten, The Data Tank, Executive Assistant

Javier de la Cueva, IE University, Lawyer, Researcher and Lecturer

Jean-Michel Contet, Orange

Jill Falman, Meta, Data for Good Program Manager

Jonny Shipp, The Data Tank, Consultant

Jose Javier Ramasco, Spanish National Research Council, Researcher

Julia Ebert, Vodafone Institute, Senior Research Manager

Kaylin Bolt, Public Health Seattle King County, Social Research Scientist

Kyriaki Kalimeri, ISI Foundation, Researcher

Laura McGorman, Meta Data for Good, Director

Laura Sandor, The Data Tank, Programme Manager

Leo Ferres, Universidad del Desarrollo, Associate Professor of Computer and Data Science

Linus Bengtsson, Flowminder, Chair of the Board

Lisa Hermans, Hasselt University, Research Manager

Luca Ferretti, University of Oxford - Pandemic Sciences Institute, Career Development Fellow

Luca Pappalardo, Italian National Research Council, Senior Researcher

Lukas Adomavicius, Centre for Information Policy Leadership (CIPL), Privacy Analyst

Manuel Garcia Herranz, UNICEF, Chief Scientist

Manuel Ribeiro, University of Lisbon, Assistant Researcher

Maria Loureiro, European Commission - HERA.02, Intelligence Analyst - Health Threats

XSL•FO
RenderX

Mattia Mazzoli, ISI Foundation, PostDoc Researcher

Mohammad Reza, Rahmanian Haghighi University of Antwerp, PhD Candidate

Moiz Shaikh, The Data Tank, Programme Manager

Natascha Gerlach, Centre for Information Policy Leadership (CIPL), Director of Privacy & Data Policy

Niel Hens, Hasselt University, Director & University of Antwerp, Scientific chair holder

Olena Snidalova, Vodafone Institute, Senior Research Manager

Paul Theyskens, MyData Global, Board Member

Paula Patricio, NOVA School of Science and Technology, Associate Professor

Paula Vasconcelos, Support Unit of National Health Authority and the Emergency Management in Public Health, Public Health Emergencies Coordination Support Operations Centre - Portugal

Paulina Behluli, The Data Tank, Project Associate

Pietro Coletti, Hasselt University/JRC, PostDoc Researcher

Polen Türkmen, The Data Tank, Project Lead

Richard Benjamins, OdiseIA, CEO & Co-Founder

Rik Gosselink, KU Leuven, Ethics Advisor

Slim Turki, Open Data & Data Ecosystems, Senior Researcher

Sofie Bekaert, King Baudouin Foundation, Senior Program Manager

Sonia Namorado, Portuguese National Health Institute Doutor Ricardo Jorge, Researcher

Stefaan Verhulst, The Data Tank, Co-Founder and Principal Scientific Advisor

Stefan Thurner, Complexity Science Hub Vienna, President

Takahiro Yabe, New York University, Assistant Professor

Funding

This project was supported by the ESCAPE project (101095619), funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. MM, PT, LF, KK, CC, DP, and SV acknowledge support from the Lagrange Project of the ISI Foundation, funded by Fondazione CRT. LF acknowledges support from the Fondo de Investigación y Desarrollo en Salud, Fonis, Project SA24I0124.

Authors' Contributions

NH, CC, DP, and SV acquired the funding and conceptualized the work. IV-L, SN, CPC, AL, and DP created the survey. MM, PT, KK, LF, CC, DP, and SV organized the workshop. IV-L analyzed the data and visualized the results. MM, IV-L, and SV wrote the original draft. All authors contributed to reviewing and editing the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary information with insights on the survey on data readiness and availability.
[PDF File, 123 KB - jmir_v28i1e85540_app1.pdf]

References

1. Albanna B, Heeks R, Pawelke A, Boy J, Handl J, Gluecker A. Data-powered positive deviance: combining traditional and non-traditional data to identify and characterise development-related outperformers. Development Engineering 2022;7:100090. [doi: 10.1016/j.deveng.2021.100090]
2. Bolt K, Gil-González D, Oliver N. Unconventional data, unprecedented insights: leveraging non-traditional data during a pandemic. Front Public Health 2024;12:1350743. [doi: 10.3389/fpubh.2024.1350743] [Medline: 38566798]

XSL•FO
RenderX

3. Chafetz H, Zahuranec A, Marcucci S, Davletov B, Verhulst S. The #Data4COVID19 review: assessing the use of non-traditional data during a pandemic crisis. SSRN Journal 2022. [doi: 10.2139/ssrn.4273229]
4. Peak CM, Wesolowski A, Zu Erbach-Schoenberg E, et al. Population mobility reductions associated with travel restrictions during the Ebola epidemic in Sierra Leone: use of mobile phone data. Int J Epidemiol 2018 Oct 1;47(5):1562-1570. [doi: 10.1093/ije/dyy095] [Medline: 29947788]
5. Teng Y, Bi D, Xie G, et al. Dynamic forecasting of Zika epidemics using Google Trends. PLOS ONE 2017;12(1):e0165085. [doi: 10.1371/journal.pone.0165085]
6. Cheng Z, Ruktanonchai NW, Wesolowski A, et al. Social, mobility and contact networks in shaping health behaviours and infectious disease dynamics: a scoping review. Infect Dis Poverty 2025 Dec 3;14(1):123. [doi: 10.1186/s40249-025-01378-6] [Medline: 41331691]
7. Wang C, Han L, Stein G, et al. Crowdsourcing in health and medical research: a systematic review. Infect Dis Poverty 2020 Jan 20;9(1):8. [doi: 10.1186/s40249-020-0622-9] [Medline: 31959234]
8. Wójcik OP, Brownstein JS, Chunara R, Johansson MA. Public health for the people: participatory infectious disease surveillance in the digital age. Emerg Themes Epidemiol 2014;11(1):7. [doi: 10.1186/1742-7622-11-7] [Medline: 24991229]
9. Syndromic surveillance web platform collecting real time data. InfluenzaNet. URL: https://influenzanet.info/ [accessed 2024-02-12]
10. Koppeschaar CE, Colizza V, Guerrisi C, et al. Influenzanet: citizens among 10 countries collaborating to monitor influenza in Europe. JMIR Public Health Surveill 2017 Sep 19;3(3):e66. [doi: 10.2196/publichealth.7429] [Medline: 28928112]
11. A community of 7,954,345 people tracking local COVID-19 and flu outbreaks. Outbreaks Near Me. URL: https://outbreaksnearme.org/us/en-US [accessed 2025-09-15]
12. South Africa's participatory surveillance pilot platform for influenza-like illness. CoughWatchSA. URL: https://coughwatchsa.org.za/ [accessed 2025-09-15]
13. Tracking respiratory illness across Australia and New Zealand. FluTracking. URL: https://info.flutracking.net/ [accessed 2025-09-15]
14. ZOE COVID study app: how king's researchers slowed the spread of COVID-19. King's College London. 2025. URL: https://www.kcl.ac.uk/news/spotlight/zoe-covid-study-app-kings-researchers-slowed-the-spread-covid-19 [accessed 2025-09-15]
15. Wilks J, Davies P. How a smart thermometer saw COVID-19 coming before the experts. Euronews. URL: https://www.euronews.com/health/2021/03/30/how-a-smart-thermometer-saw-covid-19-coming-before-the-experts [accessed 2025-09-15]
16. Early warning system for contagious illness. Kinsa. URL: https://home.kinsahealth.com/ [accessed 2025-09-15]
17. How Thryve built a data donation tool to fight COVID-19 for the German Government. Thryve. URL: https://www.thryve.health/news/how-thryve-built-a-data-donation-tool-to-fight-covid-19-for-the-german-government [accessed 2025-09-15]
18. Flu & ebola map | Virus & contagious disease surveillance. HealthMap. URL: http://healthmap.org [accessed 2025-12-13]
19. Disease events. BEACON. URL: https://beaconbio.org/en/ [accessed 2025-12-13]
20. Badker R, Kipperman N, Ash B, et al. Constructing a global human epidemic database using open-source digital biosurveillance. Sci Data 2025 Feb 26;12(1):344. [doi: 10.1038/s41597-025-04663-z] [Medline: 40011511]
21. Singer AC, Thompson JR, Filho CRM, et al. A world of wastewater-based epidemiology. Nat Water 2023;1(5):408-415. [doi: 10.1038/s44221-023-00083-8]
22. Keshaviah A, Karmali RN, Vohra D, Hu XC, Diamond MB. The role of wastewater data in pandemic management. : Mathematica; 2022 URL: https://www.rockefellerfoundation.org/wp-content/uploads/2022/04/The-Role-of-Wastewater-Data-in-Pandemic-Management-Survey-Research-Brief-Final.pdf [accessed 2026-03-05]
23. Larsen DA, Wigginton KR. Tracking COVID-19 with wastewater. Nat Biotechnol 2020 Oct;38(10):1151-1153. [doi: 10.1038/s41587-020-0690-1] [Medline: 32958959]
24. Deng Y, Xu X, Zheng X, et al. Use of sewage surveillance for COVID-19 to guide public health response: a case study in Hong Kong. Science of The Total Environment 2022 May;821:153250. [doi: 10.1016/j.scitotenv.2022.153250]
25. Buckee C, Balsari S, Schroeder A. Making data for good better. PLOS Digit Health 2022 Jan;1(1):e0000010. [doi: 10.1371/journal.pdig.0000010] [Medline: 36812507]
26. COVID-19 community mobility reports. Google. URL: https://www.google.com/covid19/mobility/ [accessed 2023-12-04]
27. Mazzoli M, Valdano E, Colizza V. Projecting the COVID-19 epidemic risk in France for the summer 2021. J Travel Med 2021 Oct 11;28(7):taab129. [doi: 10.1093/jtm/taab129] [Medline: 34414436]
28. Pepe E, Bajardi P, Gauvin L, et al. COVID-19 outbreak response, a dataset to assess mobility changes in Italy following national lockdown. Sci Data 2020 Jul 8;7(1):230. [doi: 10.1038/s41597-020-00575-2] [Medline: 32641758]
29. Oliver N, Lepri B, Sterly H, et al. Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. Sci Adv 2020 Jun;6(23):eabc0764. [doi: 10.1126/sciadv.abc0764] [Medline: 32548274]
30. Gozzi N, Tizzoni M, Chinazzi M, Ferres L, Vespignani A, Perra N. Estimating the effect of social inequalities on the mitigation of COVID-19 across communities in Santiago de Chile. Nat Commun 2021 Apr 23;12(1):2429. [doi: 10.1038/s41467-021-22601-6] [Medline: 33893279]

XSL•FO
RenderX

31. Pullano G, Valdano E, Scarpa N, Rubrichi S, Colizza V. Evaluating the effect of demographic factors, socioeconomic factors, and risk aversion on mobility during the COVID-19 epidemic in France under lockdown: a population-based study. Lancet Digit Health 2020 Dec;2(12):e638-e649. [doi: 10.1016/S2589-7500(20)30243-0] [Medline: 33163951]
32. Mossong J, Hens N, Jit M, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLOS Med 2008 Mar 25;5(3):e74. [doi: 10.1371/journal.pmed.0050074] [Medline: 18366252]
33. The CoMix study. University of Hasselt. URL: https://www.uhasselt.be/en/aparte-sites-partner-en/epipose/the-comix-study [accessed 2025-09-15]
34. Aladangady A, Aron-Dine S, Dunn W. From transaction data to economic statistics: constructing real-time, high-frequency, geographic measures of consumer spending. In: Big Data for Twenty-First-Century Economic Statistics: University of Chicago Press; 2021:115-145 URL: https://www.nber.org/books-and-chapters/big-data-twenty-first-century-economic-statistics/transaction-data-economic-statistics-constructing-real-time-high-frequency-geographic-measures [accessed 2025-09-15] [doi: 10.7208/chicago/9780226801391.003.0005]
35. Dunn A, Hood K, Driessen A. Measuring the effects of the COVID-19 pandemic on consumer spending using card transaction data. : U.S. Bureau of Economic Analysis; 2020 URL: https://www.bea.gov/sites/default/files/papers/BEA-WP2020-5_0.pdf [accessed 2026-03-05]
36. Millefiori LM, Braca P, Zissis D, et al. COVID-19 impact on global maritime mobility. Sci Rep 2021 Sep 10;11(1):18039. [doi: 10.1038/s41598-021-97461-7] [Medline: 34508144]
37. How COVID-19 and collective intelligence transformed procurement risks into opportunities for transparency in ecuador. Open Contracting Partnership. 2020 Dec 20. URL: https://www.open-contracting.org/2020/12/20/how-covid-19-and-collective-intelligence-transformed-procurement-risks-into-opportunities-for-transparency-in-ecuador/ [accessed 2025-09-15]
38. Wang D, Lentzen M, Botz J, et al. Development of an early alert model for pandemic situations in Germany. Sci Rep 2023;13(1):20780. [doi: 10.1038/s41598-023-48096-3]
39. Gallotti R, Valle F, Castaldo N, Sacco P, De Domenico M. Assessing the risks of “infodemics” in response to COVID-19 epidemics. Nat Hum Behav 2020 Dec;4(12):1285-1293. [doi: 10.1038/s41562-020-00994-6] [Medline: 33122812]
40. Kogan NE, Clemente L, Liautaud P, et al. An early warning approach to monitor COVID-19 activity with multiple digital traces in near real time. Sci Adv 2021 Mar;7(10):eabd6989. [doi: 10.1126/sciadv.abd6989] [Medline: 33674304]
41. Salomon JA, Reinhart A, Bilinski A, et al. The US COVID-19 Trends and Impact Survey: continuous real-time measurement of COVID-19 symptoms, risks, protective behaviors, testing, and vaccination. Proc Natl Acad Sci U S A 2021 Dec 21;118(51):e2111454118. [doi: 10.1073/pnas.2111454118] [Medline: 34903656]
42. Orchard F, Clain C, Madie W, et al. PANDEM-Source, a tool to collect or generate surveillance indicators for pandemic management: a use case with COVID-19 data. Front Public Health 2024;12:1295117. [doi: 10.3389/fpubh.2024.1295117] [Medline: 38572005]
43. Major emergency exercise brings Irish and EU health experts together. PANDEM-2. URL: https://pandem-2.eu/ [accessed 2025-09-15]
44. Corona prediction: telekom supports RKI. Deutsche Telekom blog. 2020 Mar 18. URL: https://www.telekom.com/en/company/details/corona-prediction-telekom-supports-rki-597114 [accessed 2025-09-15]
45. Moorthy V, Henao Restrepo AM, Preziosi MP, Swaminathan S. Data sharing for novel coronavirus (COVID-19). Bull World Health Organ 2020 Mar 1;98(3):150. [doi: 10.2471/BLT.20.251561] [Medline: 32132744]
46. Vespe M, Iacus SM, Santamaria C, Sermi F, Spyratos S. On the use of data from multiple mobile network operators in Europe to fight COVID-19. Data & Policy 2021;3:e8. [doi: 10.1017/dap.2021.9]
47. Oliver N. Data science for social good: the valencian example during the COVID-19 pandemic. EsadeEcPol Center for Economic Policy. 2022. URL: https://www.esade.edu/ecpol/en/publications/data-science-for-social-good-the-valencian-example-during-the-covid-19-pandemic/ [accessed 2025-09-15]
48. Open data movilidad. Ministerio de Transportes y Movilidad Sostenible. URL: https://www.transportes.gob.es/ministerio/proyectos-singulares/estudios-de-movilidad-con-big-data/opendata-movilidad [accessed 2025-09-15]
49. Barbosa H, Barthelemy M, Ghoshal G, et al. Human mobility: models and applications. Phys Rep 2018 Mar;734:1-74. [doi: 10.1016/j.physrep.2018.01.001]
50. R2M2P2 consortium: readying regional mobility data for modelling pandemic preparedness. ZonMw Projects. URL: https://projecten.zonmw.nl/nl/project/r2m2p2-consortium-readying-regional-mobility-data-modelling-pandemic-preparedness [accessed 2025-09-15]
51. Liu C, Holme P, Lehmann S, Yang W, Lu X. Nonrepresentativeness of human mobility data and its impact on modeling dynamics of the COVID-19 pandemic: systematic evaluation. JMIR Form Res 2024 Jun 28;8(1):e55013. [doi: 10.2196/55013] [Medline: 38941609]
52. Lu X, Feng J, Lai S, et al. Human mobility in epidemic modeling. Phys Rep 2026 Feb;1157:1-45. [doi: 10.1016/j.physrep.2025.10.010]
53. Brownstein JS, Chu S, Marathe A, et al. Combining participatory influenza surveillance with modeling and forecasting: three alternative approaches. JMIR Public Health Surveill 2017;3(4):e83. [doi: 10.2196/publichealth.7344]

XSL•FO
RenderX

54. Wilkinson MD, Dumontier M, Aalbersberg I, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3(1):160018. [doi: 10.1038/sdata.2016.18]
55. Mazzoli M, Varela-Lasheras I, Namorado S, et al. Survey on Data Availability and Data Needs for Modelers during the COVID-19 Pandemic 2025.
56. Social data for health: advancing health and wellbeing research across disciplines. Wellcome. 2025 Nov 6. URL: https://wellcome.org/insights/reports/social-data-for-health [accessed 2025-12-31]
57. Chen K, Hellerstein JM, Parikh TS. Data in the first mile. Presented at: CIDR 2011, Fifth Biennial Conference on Innovative Data Systems Research; Jan 9-12, 2011; Asilomar, CA, USA p. 203-206.
58. The last mile problem. Darkhorse Visualization blog. URL: https://darkhorsevisualization.com/blog/the-last-mile-problem [accessed 2025-12-31]
59. Bridging the last mile in data analytics. Harvard Business Review. 2017. URL: https://hbr.org/sponsored/2017/11/bridging-the-last-mile-in-data-analytics [accessed 2025-12-31]
60. Li H, Wei D, Moffitt K, Vasarhelyi MA. Addressing the last mile problem in open government data: using AIS technologies to enhance governmental financial reporting. SSRN. Journal of Emerging Technologies in Accounting. Preprint posted online on Mar 12, 2023. [doi: 10.2139/ssrn.4385883]
61. Outpacing pandemics: solving the first and last mile challenges of data-driven policy making. Medium blog. 2024 Feb 6. URL: https://medium.com/data-policy/outpacing-pandemics-solving-the-first-and-last-mile-challenges-of-data-driven-policy-making-c7f8ef6662a4 [accessed 2024-02-12]
62. Schwalbe N, Wahl B, Song J, Lehtimaki S. Data sharing and global public health: defining what we mean by data. Front Digit Health 2020;2:612339. [doi: 10.3389/fdgth.2020.612339] [Medline: 34713073]
63. Jiang JX, Cram P, Qi K, Bai G. Challenges and dynamics of public health reporting and data exchange during COVID-19: insights from US hospitals. Health Aff Sch 2024 Jan;2(1):qxad080. [doi: 10.1093/haschl/qxad080] [Medline: 38756405]
64. Prainsack B, El-Sayed S, Forgó N, Szoszkiewicz Ł, Baumer P. Data solidarity: a blueprint for governing health futures. Lancet Digit Health 2022 Nov;4(11):e773-e774. [doi: 10.1016/S2589-7500(22)00189-3] [Medline: 36307191]
65. Amadi D, Kiwuwa-Muyingo S, Bhattacharjee T, et al. Making metadata machine-readable as the first step to providing findable, accessible, interoperable, and reusable population health data: framework development and implementation study. Online J Public Health Inform 2024;16(1):e56237. [doi: 10.2196/56237]
66. Verhulst S. Inquiry as infrastructure: defining good questions in the age of data and AI. SSRN. Preprint posted online on 2025. [doi: 10.2139/ssrn.5229339]
67. Eisinger-Mathason TSK, Leshin J, Lahoti V, Fridsma DB, Mucaj V, Kho AN. Data linkage multiplies research insights across diverse healthcare sectors. Commun Med (Lond) 2025 Mar 4;5(1):58. [doi: 10.1038/s43856-025-00769-y] [Medline: 40038513]
68. Irwin N, Aisyah DN, Rahman FM, Manikam L. Digital contact tracing technology in the COVID-19 pandemic: a systematic review. Health Technol 2024 Nov;14(6):1229-1239. [doi: 10.1007/s12553-024-00857-4]
69. Yabe T, Luca M, Tsubouchi K, Lepri B, Gonzalez MC, Moro E. Enhancing human mobility research with open and standardized datasets. Nat Comput Sci 2024 Jul;4(7):469-472. [doi: 10.1038/s43588-024-00650-3] [Medline: 38956383]
70. Kostandova N, Schluth C, Arambepola R, et al. A systematic review of using population-level human mobility data to understand SARS-CoV-2 transmission. Nat Commun 2024 Dec 3;15(1):10504. [doi: 10.1038/s41467-024-54895-7] [Medline: 39627231]
71. Jiang R, Wang Z, Cai Z. Countrywide origin-destination matrix prediction and its application for COVID-19. In: Dong Y, Kourtellis N, Hammer B, editors. Machine Learning and Knowledge Discovery in Databases Applied Data Science Track: Springer International Publishing; 2021:319-334. [doi: 10.1007/978-3-030-86514-6_20]
72. Gallotti R, Maniscalco D, Barthelemy M, De Domenico M. Distorted insights from human mobility data. Commun Phys 2024;7(1):421. [doi: 10.1038/s42005-024-01909-x]
73. Kostandova N, Prosperi C, Mutembo S, et al. Comparing and integrating human mobility data sources for measles transmission modeling in Zambia. PLOS Glob Public Health 2025;5(5):e0003906. [doi: 10.1371/journal.pgph.0003906] [Medline: 40392858]
74. Mobile Network Operator data for Official Statistics (MNOdata4OS). Eurostat CROS. URL: https://cros.ec.europa.eu/MNOdata4OS [accessed 2025-09-15]
75. What is the european statistical system? eurostat. URL: https://ec.europa.eu/eurostat/web/european-statistical-system [accessed 2025-09-15]
76. Experimental statistics - overview. eurostat. URL: https://ec.europa.eu/eurostat/web/experimental-statistics [accessed 2025-09-15]
77. Fusion center. Wikipedia. 2025 Jul 25. URL: https://en.wikipedia.org/w/index.php?title=Fusion_center&oldid=1302479378 [accessed 2025-09-15]
78. Verhulst SG. From data to decision intelligence: the potential of decision accelerator labs. Medium. URL: https://medium.com/data-stewards-network/from-data-to-decision-intelligence-the-potential-of-decision-accelerator-labs-33f4060734a8 [accessed 2025-09-15]

XSL•FO
RenderX

79. European health data space (EHDS) | updates. European Health Data Space. URL: https://www.european-health-data-space.com/ [accessed 2025-09-15]
80. HealthData@EU central platform. European Union. URL: https://acceptance.data.health.europa.eu/healthdata-central-platform?locale=en [accessed 2025-09-15]
81. Verhulst S, Zahuranec AJ, Zable A, Addo P. Reimagining data governance for AI: operationalizing a social license for data reuse. SSRN Journal. 2025. [doi: 10.2139/ssrn.5255677]
82. Data collaboratives canvas. Observatory of Public Sector Innovation. 2018. URL: https://oecd-opsi.org/toolkits/data-collaboratives-canvas/ [accessed 2025-09-15]
83. Verhulst SG. Wanted: Data Stewards - (Re-)defining the roles and responsibilities of Data Stewards for an age of data collaboration. Medium. 2020. URL: https://medium.com/data-policy/wanted-data-stewards-re-defining-the-roles-and-responsibilities-of-data-stewards-for-an-age-of-75ed59f65e0a [accessed 2026-03-05]
84. Privacy enhancing technologies. Organisation for Economic Co-operation and Development. URL: https://www.oecd.org/en/topics/sub-issues/privacy-enhancing-technologies.html [accessed 2025-09-15]
85. Rancea A, Anghel I, Cioara T. Edge computing in healthcare: innovations, opportunities, and challenges. Future Internet 2024;16(9):329. [doi: 10.3390/fi16090329]
86. Kalkar U, Saxena S, Verhulst V. Data sandboxes: managing the open data spectrum. : Open Data Policy Lab; 2023 URL: https://incubator.opendatapolicylab.org/files/DataSandboxesPrimer.pdf [accessed 2025-09-15]
87. Crum ED, Taelman R, Buelens B, Ertaylan G, Verborgh R. Personalized medicine through personal data pods. 2024 Presented at: SWAT4HCLS 2024: Semantic Web Applications and Tools for Health Care and Life Sciences 2024; Feb 20-26, 2024; Leiden University Medical Center, Leiden University, and Naturalis, The Netherlands p. 167-168 URL: http://hdl.handle.net/1854/LU-01JKQQH3PMCX6P8B6NY88MD6W9 [accessed 2026-03-05]
88. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. npj Digit Med 2023;6(1):1-8. [doi: 10.1038/s41746-023-00927-3]
89. Mistry D, Litvinova M, Pastore y Piontti A, et al. Inferring high-resolution human mixing patterns for disease modeling. Nat Commun 2021;12(1):1. [doi: 10.1038/s41467-020-20544-y]
90. Bacco M, Kocian A, Chessa S, Crivello A, Barsocchi P. What are data spaces? Systematic survey and future outlook. Data Brief 2024 Dec;57:110969. [doi: 10.1016/j.dib.2024.110969] [Medline: 39957735]

Abbreviations

EC: European Commission
EHDS: EU Health Data Space
MNO: mobile network operator
NTD: nontraditional data

Edited by A Mavragani; submitted 09.Oct.2025; peer-reviewed by N Madhav, S Mitra; accepted 03.Mar.2026; published 29.Apr.2026.

Please cite as:
Mazzoli M, Varela-Lasheras I, Namorado S, Pereira Caetano C, Leite A, Hermans L, Hens N, Türkmen P, Kalimeri K, Ferres L, Cattuto C, Paolotti D, Verhulst S
Nontraditional Data in Pandemic Preparedness and Response: Identifying and Addressing First- and Last-Mile Challenges
J Med Internet Res 2026;28:e85540
URL: https://www.jmir.org/2026/1/e85540
doi:10.2196/85540

© Mattia Mazzoli, Irma Varela-Lasheras, Sónia Namorado, Constantino Pereira Caetano, Andreia Leite, Lisa Hermans, Niel Hens, Polen Türkmen, Kyriaki Kalimeri, Leo Ferres, Ciro Cattuto, Daniela Paolotti, Stefaan Verhulst. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Rheumatic Digital Twin: Proposed Machine Learning–Based Multimodal Framework to Inform Clinical Decision-Making

Daniyal Selani[1,2*], MSc; Rachel Knevel[2*], MD, PhD; Marcel Reinders[1,3], PhD; Erik B van den Akker[1,3], PhD

[1]Pattern Recognition and Bioinformatics, Faculty Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands

[2]Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands

[3]Department of Biomedical Data Sciences, Leiden University Medical Center, Einthovenweg 20, Leiden, The Netherlands

[*]these authors contributed equally

Corresponding Author:
Erik B van den Akker, PhD
Pattern Recognition and Bioinformatics, Faculty Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands

Abstract

Rheumatic diseases are chronic, immune-mediated conditions characterized by significant heterogeneity in presentation and disease course. However, current clinical approaches often rely on snapshot-based assessments that fail to capture the complex longitudinal evolution of these conditions. To address these limitations and support the implementation of precision medicine, we present the design for the Rheumatic Digital Twin, a novel, modular conceptual framework intended to integrate heterogeneous multimodal data, ranging from electronic health records and clinical notes to imaging and omics, into a dynamic, computational representation of the patient journey. Our theoretical architecture addresses challenges related to data silos and variable availability of data modalities through a multistage approach that envisions the use of domain-specific foundation models to independently process distinct data modalities. To effectively model the temporal progression inherent in chronic diseases, the proposed design utilizes Transformer architectures, leveraging self-attention mechanisms to treat patient events, such as lab results or medication changes, as sequential data tokens. We describe how these unimodal representations would subsequently be fused via joint embedding techniques to construct a shared, multimodal representational space. Envisioned to function analogously to a recommender system, the Rheumatic Digital Twin framework is modeled to map patients into a latent space where proximity reflects clinical and biological similarity. By identifying "nearest neighbors," historical patients with comparable trajectories, the system aims to enable in silico cohorting, theoretically allowing clinicians to forecast key clinical events, predict treatment responses, and identify likely disease courses based on the outcomes of similar peers.

(J Med Internet Res 2026;28:e86763) doi:10.2196/86763

KEYWORDS

digital twin; artificial intelligence; AI; rheumatology; electronic health records; machine learning; multimodal data

Introduction

Rheumatic diseases (RDs) are chronic, immune-mediated conditions marked by persistent inflammation affecting joints, connective tissues, and internal organs. RDs are progressive, often disabling, and notoriously heterogeneous in both presentation and disease course. Early diagnosis and individualized treatment are essential to improving outcomes and preventing irreversible damage, yet current clinical approaches remain ill-equipped to handle the complexity of these diseases [1]. Traditional RD classification and diagnostic frameworks rely on clinical assessment, patient-reported symptoms, imaging, and blood-based biomarkers. While valuable, these sources often offer limited resolution, capturing only fragments of a patient's disease profile. Crucially, they fail to reflect the underlying molecular mechanisms and individualized risk factors that drive disease heterogeneity. Collectively, the current classification leads to patient groupings that appear clinically similar but may differ substantially in their biology, prognosis, and treatment response.

To address these limitations, researchers have turned to multimodal data integration, combining clinical variables with genetic, transcriptomic, proteomic, and imaging data [2]. These efforts have shown promise; for example, transcriptomic profiling has been used to define novel inflammatory endotypes in rheumatoid arthritis (RA) [3], and combined clinical-genetic models have improved flare prediction [4]. Yet, most approaches remain constrained by siloed data analysis, one-time snapshots, and limited longitudinal context. The full potential of multimodal data in rheumatology has yet to be realized.

We argue that a more comprehensive, dynamic, and patient-specific model is needed; one that evolves with the patient and enables interpretation within the broader disease

XSL•FO
RenderX

context. This is where the concept of the medical digital twin (MDT) becomes particularly compelling. A digital twin (DT) is a continuously updating computational representation of an individual patient that utilizes data-driven models to integrate diverse data streams over time, ranging from electronic health records (EHR) and omics profiles to imaging, wearables, and lifestyle data [5]. MDTs have already demonstrated promise in oncology [6], cardiology, and critical care [7], enabling simulation of disease progression, real-time risk prediction, and adaptive therapy planning. The application of MDTs within rheumatology, however, is still in its infancy [8]. Existing MDT approaches often fall short when it comes to effectively integrating high-dimensional, multimodal data into a unified representation of the patient [9]. This integration is critical; the utility of an MDT depends not just on data accumulation but on mapping these data modalities into a shared space in which features of a patient can be aligned and compared meaningfully with those of previously admitted patients.

In response to these challenges, this paper presents the conceptual design and architectural specifications for the Rheumatic Digital Twin (RDT). The primary novelty of our work lies in the proposal of an architecture specifically engineered to address the unique complexities of RDs, which are characterized by an exceptional degree of multimodal heterogeneity and long-term, fluctuating disease courses. Unlike existing models that often rely on static data snapshots, the proposed RDT architecture introduces a modular, Transformer-based fusion framework designed to capture the dynamic, temporal evolution of the patient's journey.

The contribution of the proposed RDT framework is threefold. First, it enables the integration of diverse, high-dimensional data into a unified representational space. Second, the modular design allows for independent handling of each modality, making the system resilient to missing data and scalable for clinical environments with varying resources. Finally, by positioning individual patient journeys within a broader population context, the RDT facilitates "in silico cohorting," allowing clinicians to forecast outcomes and treatment responses by analogy with similar historical cases. This application-specific approach provides a necessary foundation for moving beyond traditional RD classification toward a truly personalized, data-driven model of care.

While the RDT framework is applicable across the spectrum of rheumatology, it is particularly positioned to benefit complex, heterogeneous conditions such as RA, systemic lupus erythematosus, and systemic sclerosis. In these diseases, clinical presentation often diverges significantly from underlying molecular pathology; patients may share similar symptoms but follow vastly different trajectories regarding joint erosion, renal involvement, or interstitial lung disease. By integrating multimodal streams ranging from ultrasound synovitis scores to transcriptomic profiles, the RDT aims to better resolve these diverse endotypes than traditional clinical classification criteria.

Recommender Systems as DTs: An Intuitive Definition

There is currently no broad consensus on the definition of DTs in general, let alone for applications in the medical domain. In their most ambitious implementations, DTs aim to exhaustively model all aspects of a system [10]. However, demands on the available data are often limiting, thus creating implementation boundaries. In the medical context, these limitations involve the variety, the variable availability, and the variable sampling frequencies of data types measured for patients.

We define MDTs as an information system or a collection of systems capable of integrating patient data aggregated during the patient journey to generate patient representations that are clinically actionable and predictive of outcomes, useful for supporting clinical decision-making tasks. As our MDT is rooted in the longitudinal sequence of events, it constructs a DT of the patient's journey, treating the evolving clinical narrative as the defining feature of the patient's representation. This is particularly important in our multimodal approach, where each event can encompass data from various modalities (eg, lab results, imaging, clinical notes).

It is important to acknowledge that the pursuit of a full-scale MDT exists alongside a prevailing clinical preference for simpler, established predictive models. In many practical settings, clinicians may prioritize targeted omics studies or validated "snapshot" biomarkers because they offer high interpretability and lower resource requirements compared to the inherent complexity of a DT. These simpler models are often easier to implement within existing time-sensitive workflows and provide clear, actionable evidence for specific clinical questions. However, we argue that while these traditional approaches are relevant, they are often ill-equipped to handle the deep heterogeneity and chronic, evolving nature of RDs. Relying solely on static, unidimensional analysis risks missing the latent disease states and emergent phenotypes that only a dynamic, multimodal representation can reveal. The RDT is not intended to replace these tools; in fact, it ingests exactly these mutually complementary data points to comprehensively capture the whole patient journey, thereby aiming to provide a level of precision that single-instance models cannot achieve.

Our proposed MDT seeks to embed patient data in a space where patient similarities and differences are amplified. Similar patients are mapped close together, whereas dissimilar patients are mapped further apart. This concept is analogous to how recommender systems for services such as Amazon and Netflix work [11]. Even though these are surface-level similarities, the task of recommending movies or shopping items based on previous customers with very similar preferences and interests provides an intuitive way to understand our DT proposal. For instance, in our MDT, "shop items" or "movies" become "treatment options," and "customer preferences and interests" are replaced by all recorded interactions between the patient and the health care system, that is, the multimodal patient data. Ochoa et al [12] have already shown how a recommendation system–like approach can be utilized for treatment decision support.

XSL•FO
RenderX

Recommender systems for commercial services typically define a space in which users with similar interests are closely aligned, effectively mapping similar users into the same neighborhoods or clusters using embedding methods or clustering algorithms. The underlying assumption is that users who exhibit similar behavior, recorded via interactions with the website, will have similar interests and, hence, would want to buy or watch similar items. Interestingly, these systems also track user behavior over time to account for changes in taste or interest. This temporal aspect is equally crucial for applications in the medical domain, where patient journeys unfold over time, and incorporating time-resolved multimodal data allows us to capture the evolving nature of the patient's condition in high detail [5].

Drawing parallels, our hypothesis is that patients with closely resembling phenotypic, genetic, demographic, and clinical traits may have similar disease states and could, therefore, respond to similar treatments. That is, patients who have some underlying similarities based on the complex interaction of factors might develop similar or identical disease states. This is particularly relevant when considering multimodal data, as the "interaction of factors" can involve complex relationships among genetic predispositions, lifestyle choices, and various clinical measurements. This is useful because we can use existing data from previous patients to stratify new patients into clinically relevant clusters and trajectories. Placing individual patients within such multimodal representational spaces that encapsulate the full heterogeneity of the broader RD population unlocks powerful possibilities for personalized medicine and translational insight. Patients positioned near each other in this space share not just superficial clinical features but deeper molecular and phenotypic similarities. This contextualization allows clinicians and researchers to infer likely disease trajectories, therapeutic responses, and even comorbidity risks by analogy with similar patients, essentially enabling in silico cohorting.

For instance, if a newly diagnosed patient maps closely to a cluster of individuals who responded favorably to a specific biologic therapy, this evidence can inform treatment selection, even in the absence of extensive prior clinical history. Beyond treatment prediction, this context also allows for early identification of at-risk patients, continuous monitoring for deviations from expected patterns, and hypothesis generation for underexplored subgroups. Moreover, the evolving structure of the patient space itself can reveal latent disease states, emergent phenotypes, or signatures of response that are not evident through unidimensional analysis. In this way, DTs not only support individual patient care but also serve as a foundation for continuous learning and discovery within the rheumatology domain.

A Modular Multimodal MDT

Given the practical constraints of data availability and the varying nature of health care data, we propose a flexible architecture that can adapt to diverse tasks and data types. In clinical settings, data modalities can vary significantly in their ease of collection, granularity, and frequency. For example, questionnaire data can be easily collected for many patients, while genomic data is often restricted by high costs, technical expertise, and privacy concerns. The architecture of our proposed MDT (Figure 1) is designed to be modular, allowing for the integration of different types of data depending on the specific needs of the task at hand and the resources available. Instead of developing a single model that tries to process and integrate all available data types simultaneously, we propose the independent handling of each data modality (Figure 1, panel 1), which can then be combined into a joint embedding (Figure 1, panel 2). This joint embedding, once generated, can be integrated into a task-specific pipeline that allows the system to perform specific clinical tasks, such as clustering patients based on phenotypic similarities (Figure 1, panel 3) or predicting treatment responses (Figure 1, panel 4).

XSL•FO
RenderX

Figure 1. Rheumatic Digital Twin (RDT). Architecture of the multimodal RDT for precision patient stratification and outcome forecasting. (1) Diverse clinical and biomedical data sources, including physical assessments, billing codes (in the *International Classification of Diseases, 10th Revision* [*ICD-10*] standard), blood tests, and, optionally, genetic profiles and patient questionnaires, undergo domain-specific transformations. (2) These reduced representations are integrated through joint embedding to construct a patient-patient network that captures interpatient similarity across modalities. (3) The resulting representations enable unsupervised clustering to uncover disease subtypes. By analyzing these representations longitudinally, the model captures the evolving "patient journey" through distinct states, building on the concept of "cluster journeys" [13,14] to define treatment response profiles and relapse trajectories over time. (4) New patients are mapped onto this latent space to forecast outcomes by identifying their nearest neighbors among patients with known trajectories. GPs: general practitioners; ML: machine learning.

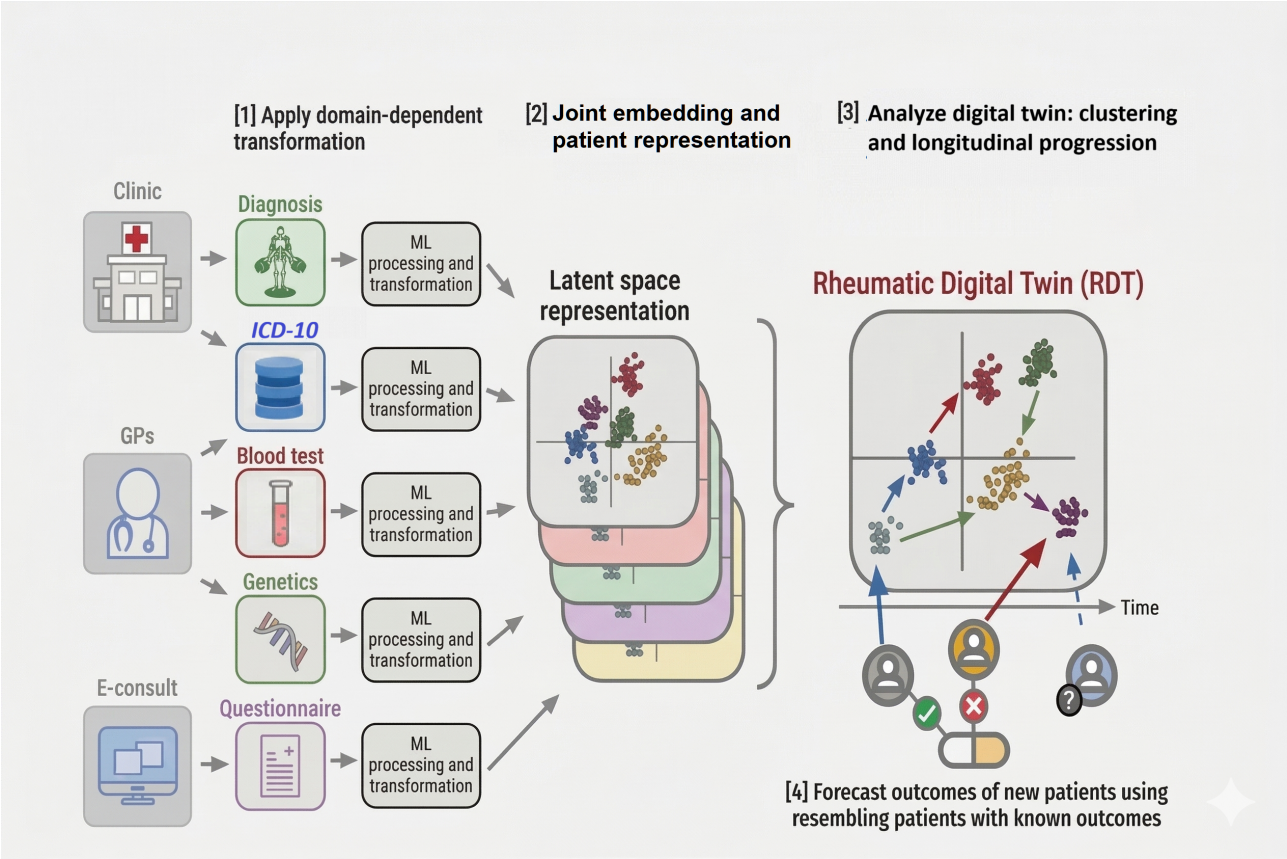

For the embedding of each modality (eg, clinical data, genetic data, imaging, etc), we leverage the concept of foundation models, that is, neural networks trained in an unsupervised manner on large modality-specific datasets to extract meaningful embeddings [15]. These models are trained independently for each modality to preserve their domain-specific characteristics. To combine the individual embeddings into a unified representation, we draw on recent advances in multimodal learning, such as unsupervised contrastive learning [16] and supervised joint embedding models [17] (Figure 2).

XSL•FO
RenderX

Figure 2. Schematic overview of the multistage workflow for building task-specific multimodal digital twins. Modality-specific data undergoes self-supervised pretraining to yield unimodal foundation models. These models are aligned through a joint embedding training phase, followed by task-specific fine-tuning, culminating in a specialized multimodal model tailored to the target application.

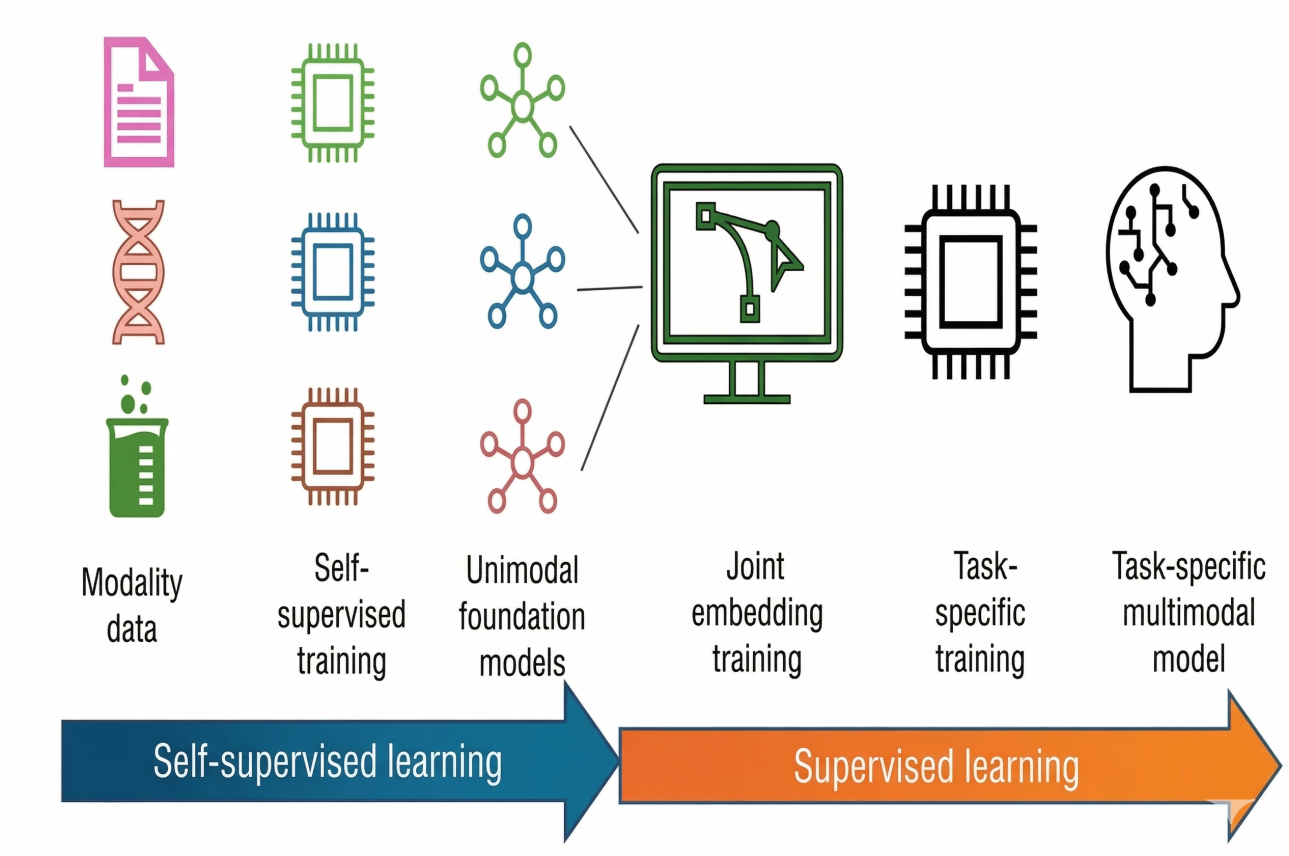

Once a patient's data is embedded into the shared space, it can evolve over time. As new events in a patient's journey, such as a change in disease activity score or the addition of a new medication, are recorded, the patient's embedding can be updated. The integration process does not require the entire model to be retrained. Instead, new events can be incorporated by updating the patient's representation in the embedding space and running the pipeline again. This creates a dynamic 2-way flow of information: decisions informed by the MDT affect the patient's journey, while new events in the patient's journey further inform the MDT. The resulting patient embeddings can then be used as inputs for additional models designed to predict disease trajectories, treatment outcomes, or potential complications.

The modular design enhances the system's flexibility and scalability, as it allows for the addition of new data modalities or tasks without requiring significant reconfiguration of the entire system. For example, the system can initially be developed using readily available clinical data and later be expanded to include genomic or imaging data as these modalities become more accessible. This also enables continuous refinement of the MDT system, which can evolve alongside advancements in medical technology and the availability of new data sources. As the system grows and more modalities are incorporated, the resulting MDT becomes an increasingly robust tool for providing personalized clinical recommendations.

Modeling Single Modalities

In chronic diseases, patient care unfolds over extended periods, during which patients' characteristics, disease states, symptoms, and treatment responses evolve. This dynamic nature of the patient journey necessitates a method of data modeling that can capture the progression of the disease over time and the interaction between past events and future outcomes. Unlike static datasets, longitudinal patient data must reflect not only the sequence of events but also the way each event influences subsequent events, incorporating both medical history and external factors. This ongoing temporal progression means that patients cannot be modeled as isolated snapshots; instead, their health status should be seen as part of an evolving narrative, a "patient journey" that spans multiple visits, treatments, and changes in condition.

To effectively model this continuous and evolving process, we propose using Transformers, a deep learning architecture particularly well suited for sequential data [18]. Transformers are not only capable of modeling complex sequential data but can also learn intricate dependencies across long temporal sequences, making them ideal for capturing the unfolding nature of patient data. Transformers are the backbone of state-of-the-art models in natural language processing because of their ability to handle long-range dependencies, contextual relationships, and the interdependencies between elements within a sequence. In the context of medical data, these "elements" are patient events, such as appointments, treatments, lab results, and disease

XSL•FO
RenderX

activity scores. Each of these events can be seen as part of a larger, context-dependent journey, and Transformers excel at modeling these relationships.

At the core of Transformers lies the self-attention mechanism, which is essential for processing sequences with complex interdependencies. The self-attention mechanism works by computing a weighted sum of all elements in the sequence, where the weights reflect the relevance of each element (or event) to others within the sequence. This is crucial when modeling patient journeys, as a single medical event, such as a change in lab results or a prescribed medication, can significantly impact the patient's subsequent health status. The self-attention mechanism enables the model to prioritize the most relevant events in a patient's history when making predictions or recommendations, allowing it to focus on the parts of the journey that are most predictive of future outcomes. For example, a previous diagnosis of a comorbidity may strongly influence treatment decisions for a current condition. Transformers can automatically assign a higher weight to such events when making predictions about treatment efficacy or disease progression. The specific definition of an event depends on the modality being modeled. For example, if we are modeling clinical visits, each event might correspond to a single visit, with all associated information, such as diagnoses, treatments, and prescriptions. If we are modeling disease activity scores, each event could represent a new score recorded during a visit or a specific time point in the disease progression.

By treating these patient events as tokens in a sequence, we can apply the same Transformer architecture that has proven successful in natural language processing tasks like machine translation or sentiment analysis. The key to this approach is the self-supervised training method known as masked language modeling. In masked language modeling, portions of the input data are randomly masked, and the model is tasked with predicting the missing information based on the surrounding context. This pretraining strategy encourages the model to learn a deep understanding of the relationships between different events in the patient journey, helping it generate meaningful representations of the patient's health history. For instance, Zhang et al [7] demonstrated the use of Transformers in modeling patient journeys represented by International Classification of Diseases, 10th Revision (*ICD-10)* billing codes, where the sequence of medical visits and associated codes was treated as events. This approach enabled the Transformer to develop clinically relevant representations of patients, which could then be fine-tuned for tasks such as predicting future disease progression or treatment responses. Similarly, by adopting a Transformer-based model for patient data, we can ensure that each patient's journey is captured in a rich, contextualized manner, accounting for the temporal relationships between events.

To effectively process the inherent complexity and irregular sampling of EHRs, we employ a specialized tokenization strategy that converts heterogeneous medical data into a standardized sequence for the Transformer. Since rheumatic patient data is often irregularly sampled, we address temporal sparsity by explicitly modeling the time elapsed between clinical encounters. This is achieved by incorporating "gap event" tokens or time-delta embeddings into the sequence, which allows the model to differentiate between events occurring in rapid succession and those separated by months or years. This ensures that the Transformer incorporates the duration between observations as a primary feature, preserving the temporal context of the "evolving narrative" of the patient's journey.

While alternative architectures, such as temporal convolutional networks [19] or recurrent neural network (RNN) variants like gated recurrent unit with decay [20], are commonly used for sequential data, we prioritize Transformers due to their superior ability to handle long-range dependencies and complex contextual relationships [21]. Unlike strictly recurrent models, which may suffer from vanishing gradients over the multi-year timelines typical of RD [22,23], the self-attention mechanism allows the RDT to prioritize specific historical events, such as a distant comorbidity, that remain highly predictive of current treatment outcomes.

The conversion of specific data types into tokens is modality-dependent. Structured clinical data, such as *ICD-10* billing codes and specific diagnoses, are treated as discrete tokens similar to words in a vocabulary. For continuous variables, such as irregularly sampled disease activity scores or laboratory values, we either discretize the values into bins or use linear layers to project the raw numerical data into a high-dimensional embedding space. Unstructured data, such as clinical notes from e-consults or physical assessments, are processed via hierarchical tokenization. In this approach, a specialized model like Clinical BERT (Bidirectional Encoder Representations from Transformers) generates a single vector for an entire note. To address the token limits of traditional architectures, we leverage current state-of-the-art methods such as Longformers [24] and sparse attention models [25]. These advancements, driven by improved hardware capabilities, support greatly increased context windows, effectively mitigating truncation issues and ensuring that even lengthy clinical narratives are captured before entering the central Transformer as a single "event token." By standardizing these diverse inputs into a unified token stream, the Transformer can leverage its self-attention mechanism to identify dependencies between disparate events, regardless of their original format (Figure 3).

XSL•FO
RenderX

Figure 3. Transformer integration within the Rheumatic Digital Twin framework. The schematic illustrates the model hierarchy, where the Transformer acts as the central temporal core. It processes heterogeneous patient data as sequential tokens (left) to generate contextualized embeddings (center), which are then used by downstream predictive heads for forecasting and treatment simulation (right). EHR: electronic health record.

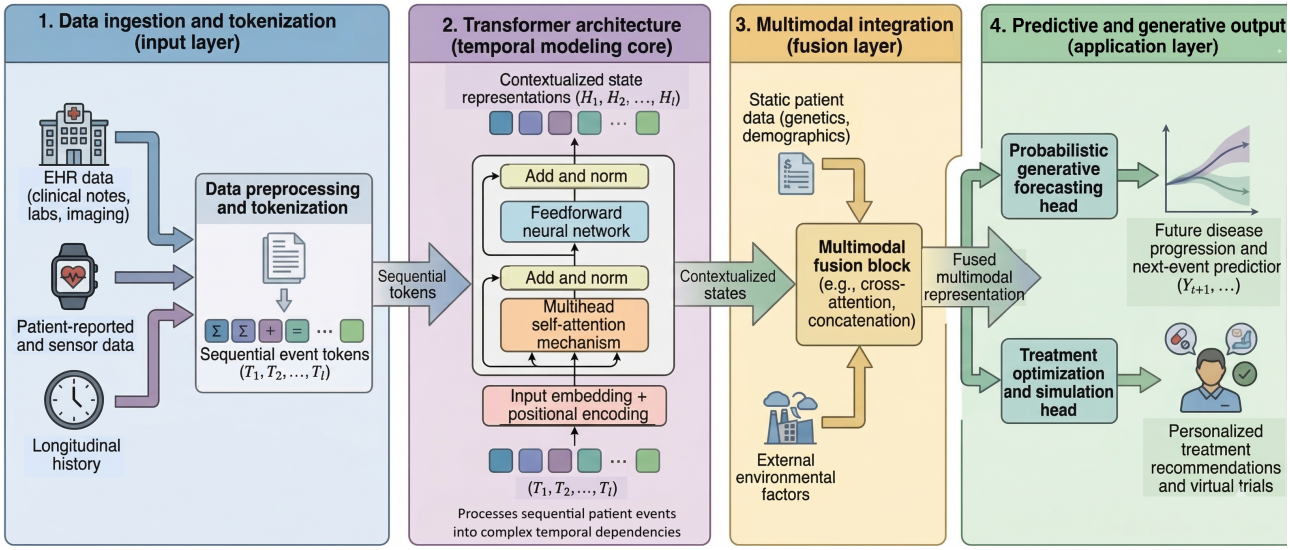

Joint Embeddings

To integrate these diverse perspectives captured by different data modalities, our framework utilizes an intermediate fusion strategy. Rather than aligning data at the raw input level, which would be highly sensitive to noise and formatting differences, each modality is independently processed by its respective foundation model to generate domain-specific embeddings. These embeddings are then projected into a shared latent space via a multimodal Transformer.

A significant challenge in this phase is the inherent "complete missingness" of high-dimensional data, such as genetic or imaging profiles, which may not be available for every patient. To maintain alignment despite these gaps, we employ a self-attention masking mechanism. In this approach, missing modality embeddings are represented as masked or zero-padded tokens before being fed into the joint Transformer. The self-attention mechanism is specifically configured to ignore these masked tokens, ensuring that the learned cross-modal relationships are derived solely from available data. This allows the model to remain resilient; it can still compute a robust patient representation by leveraging the correlations among whatever subset of modalities is present. By treating each modality as an independent but relatable "token" in the patient's profile, the architecture avoids the rigid requirement of complete data matrices common in traditional multivariate analysis.

To ensure that the joint embedding space preserves meaningful cross-modal relationships rather than merely concatenating data, we propose 2 primary self-supervised learning objectives. First, we use contrastive learning, which is designed to minimize the distance between embeddings of semantically related instances, such as a specific diagnostic imaging feature and its corresponding clinical description, while distinguishing them from dissimilar instances [26,27]. Second, we incorporate a cross-modal masked modeling strategy, wherein portions of one modality (eg, specific lab results) are hidden, requiring the model to reconstruct the missing information using context from available modalities (eg, longitudinal clinical notes) [28,29]. By training the model to predict across different data types, the system is designed to capture the underlying biological and clinical correlations between disparate sources [30,31]. This alignment can enable the RDT to map patients with similar underlying disease mechanisms to the same neighborhood, maintaining robustness even when data profiles are incomplete or derived from differing modalities [17]. These shared embeddings can then be used for tasks such as cross-modal retrieval, where a query image could return the most relevant medical records, or for transfer learning to apply knowledge gained from one data modality to another, less well-studied modality [32,33].

Clinical Decision Making

Once a patient's data has been embedded into the multimodal MDT space, the core question becomes: "How can this representation be used to inform clinical decisions?" While embeddings offer a powerful means of integrating diverse sources of information into a unified framework, they are not the end goal. They must be translated into actionable insights. Depending on the clinical use case and the nature of the available data, there are several principled ways in which predictions can be made from this space.

One intuitive and transparent approach is neighborhood-based reasoning (Figure 4). This method compares a patient's embedding to others in the latent space, assuming proximity indicates clinical similarity. By identifying nearest neighbors, the system can analyze prior treatments and outcomes. If a therapy was successful among close neighbors, it might be recommended for the current patient. This local inference, like case-based reasoning, offers a familiar paradigm for clinicians.

XSL•FO
RenderX

Figure 4. Conceptual illustration of the patient embedding process. (A) High-dimensional input: raw EHR data are characterized as high-dimensional, sparse, and heterogeneous, comprising diverse inputs such as unstructured notes, laboratory values, demographics, and longitudinal visit timelines. (B) Latent embedding space: the embedding function transforms these inputs into a low-dimensional, dense mathematical representation (latent space). In this space, distance correlates with clinical similarity: patients with shared chronic conditions (eg, arthritis) cluster closely together (high similarity), while remaining distinct from patients with unrelated phenotypes, such as acute fractures or healthy controls. This illustrates that "embedding" refers to a learned feature space, not anatomical joints. EHR: electronic health record.

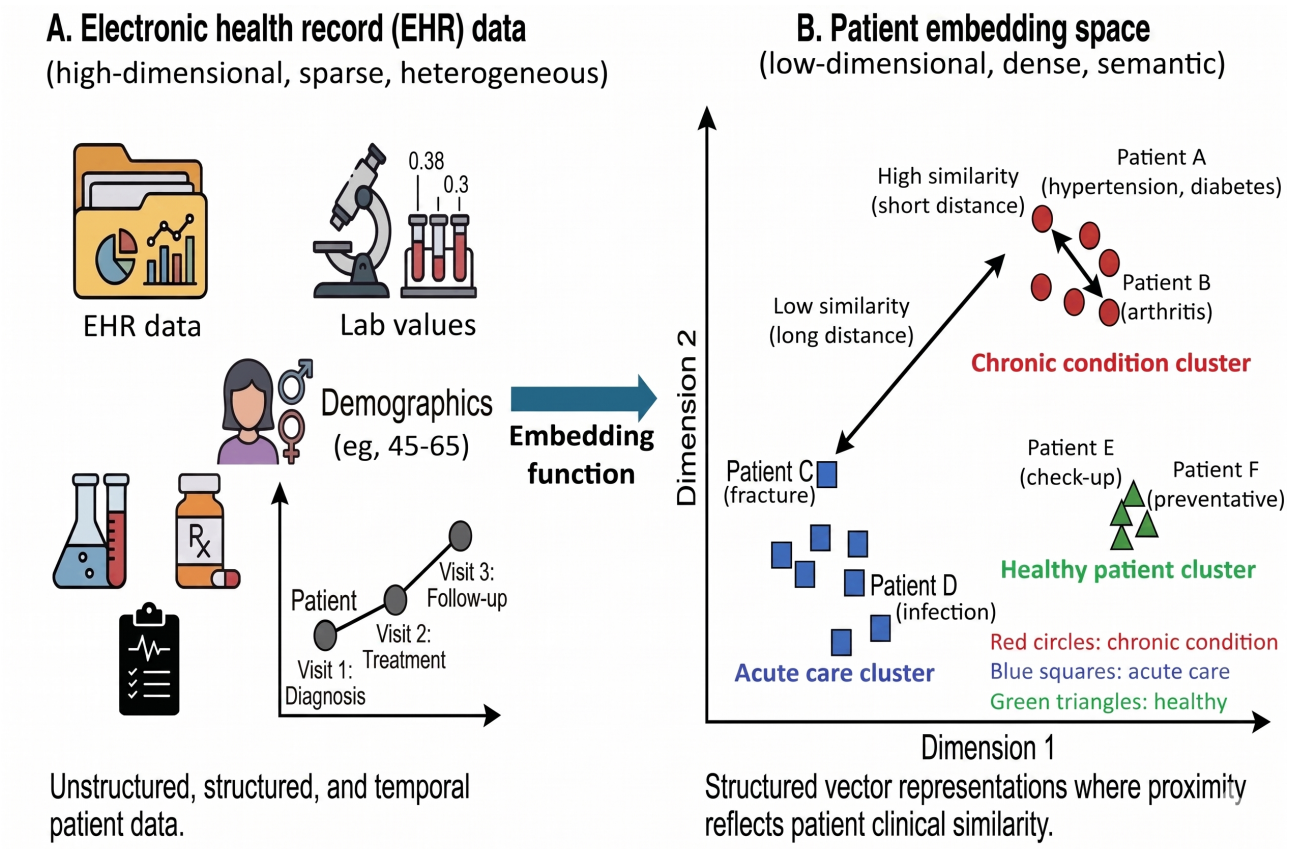

The utility of neighborhood-based reasoning depends on how "similarity" is defined and measured within the multimodal representational space [34]. In our framework, clinical similarity is quantified using mathematical distance metrics, such as cosine similarity or Euclidean distance, applied to the learned latent embeddings. Proximity in this shared space indicates that 2 patients share deeply aligned multimodal profiles, even if their superficial characteristics differ.

To make this actionable, these distances can be mapped to specific clinical metrics: for instance, a newly diagnosed patient might be found to reside in a "neighborhood" where 80% of historical cases responded favorably to tumor necrosis factor inhibitors or shared a high likelihood of flares within 6 months. By identifying these nearest neighbors, clinicians can transition from abstract data points to concrete comparative outcomes, such as identifying shared phenotypic trajectories or similar risks for specific comorbidities. This transforms the latent space from a purely computational construct into a clinically actionable map that informs treatment selection based on the proven successes and failures of "look-alike" patients.

Alternatively, a predictive modeling approach can be employed by training downstream classifiers or regressors directly on the embedding space. Here, embeddings serve as input features for models predicting specific clinical targets, such as disease progression, treatment efficacy, or adverse event risk. This method leverages the compact, information-rich nature of embeddings, which are often better suited for learning than high-dimensional raw data. For example, a simple feedforward neural network, logistic regression, or gradient-boosted trees could predict treatment response likelihood, enabling scalable and fast clinical decision support.

The MDT system can also be used for more specific, yet critical, tasks. It enables forecasting key clinical events, such as predicting the likelihood or timing of significant medical interventions, specialist visits, or treatment changes at future points (eg, 150, 270, 360 d) in the patient's care pathway (Figure 5A). The MDT system also facilitates early disease diagnosis prediction, allowing for the estimation of a specific diagnosis's probability within a future window based on a patient's historical data, which is crucial for timely intervention (Figure 5B,C).

XSL•FO
RenderX

Figure 5. Strategies for clinical event forecasting within the Rheumatic Digital Twin. (A) Fixed-window prediction: illustration of a static forecasting task where patient data gathered during a fixed initial window (eg, the first 150 d) is used to predict specific downstream events, such as a future rheumatic (REU) appointment or a disease-modifying antirheumatic drug (DMARD) prescription at set time points (270 or 360 d). (B and C) Iterative dynamic prediction: illustration of a sliding-window approach where predictions are updated continuously as the patient journey evolves. (B) In the early phase, data available up to point Ma are used to predict a diagnosis within the next n days; the probability is low as the event lies outside the prediction window. (C) As the observation window extends to Mb, the specific diagnosis event falls within the prediction horizon (Mb + n days), resulting in a higher probability forecast. Dx: diagnosis; Ma: earlier observation time point; Mb: later observation time point.

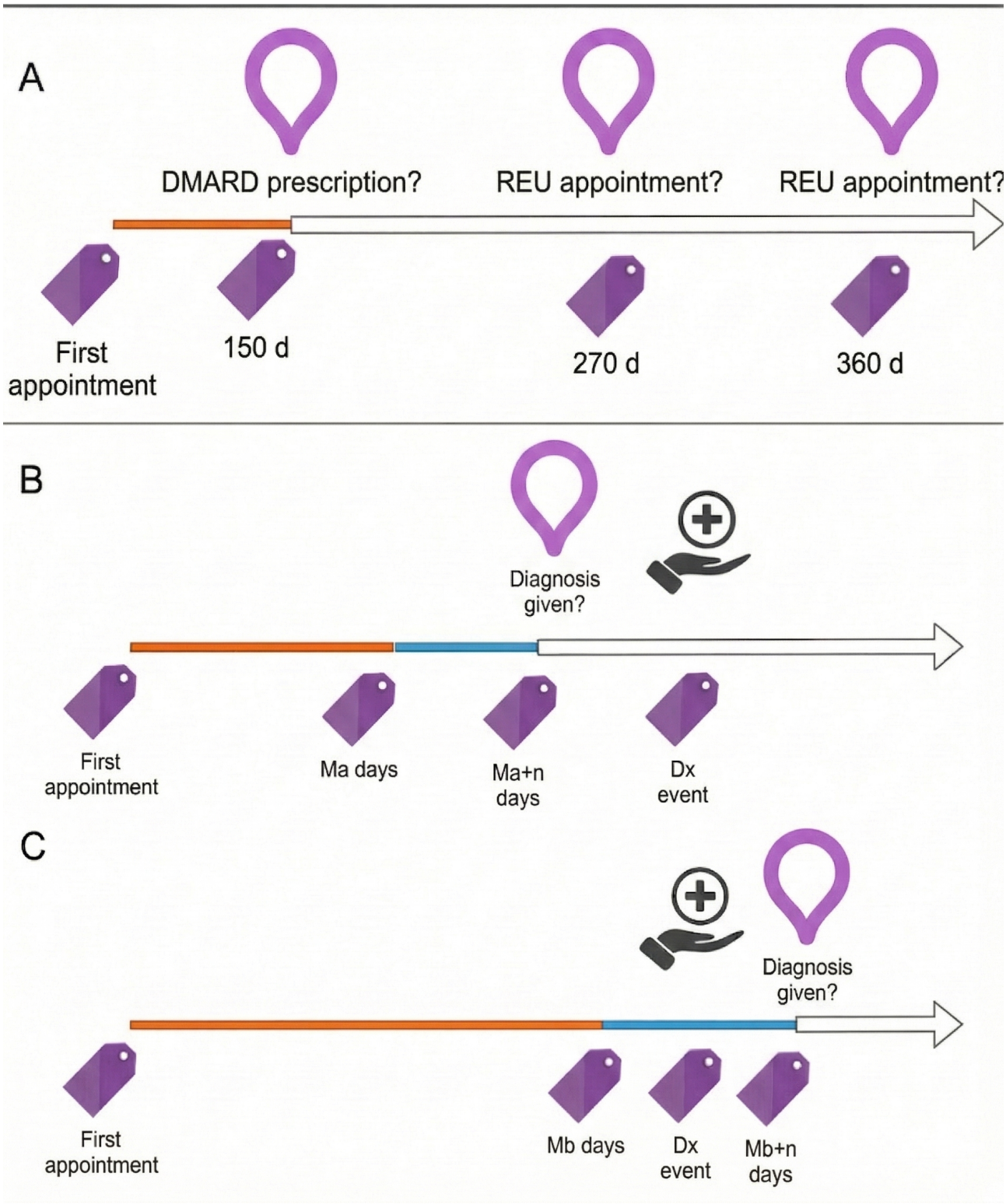

In practice, these approaches are not mutually exclusive. A robust DT system could combine them, using neighbor-based reasoning for transparency, predictive models for speed, generative models for scenario planning, and active learning for uncertainty management. Together, they provide a rich toolkit for turning multimodal embeddings into personalized, data-driven clinical actions.

XSL•FO
RenderX

Case Vignette: Dynamic Stratification in RA

To illustrate the practical utility of the RDT, consider 2 hypothetical patients, "Patient A" and "Patient B," both presenting with early RA. At baseline, both patients appear clinically identical: they are of similar age, have high disease activity (Disease Activity Score-28 >5.1), and test positive for anti–cyclic citrullinated peptide serology. Under current standard-of-care guidelines, both are viewed as "high-risk" candidates and are prescribed methotrexate as a first-line therapy. However, the RDT would resolve the underlying heterogeneity by ingesting their full multimodal profiles, utilizing *ICD-10* billing codes, longitudinal lab values, and unstructured doctor notes from previous consultations.

While their structured billing codes and diagnosis labels are identical, the joint embedding layer detects subtle, latent differences. Patient A's embedding is heavily influenced by specific phrases in the doctor's notes regarding extra-articular manifestations and a pattern of escalating inflammatory lab values, whereas Patient B's embedding aligns with standard seropositive profiles. Consequently, the RDT projects them into different "neighborhoods" of the latent space despite their superficial similarity. The model maps Patient B to a cluster of historical "methotrexate responders," validating the standard course of action where the patient proceeds to achieve remission on methotrexate alone. Conversely, the model maps Patient A to a "refractory" cluster, where nearest neighbors historically failed methotrexate and suffered early joint erosion. Instead of waiting 6 months to confirm treatment failure via a physical flare, the RDT triggers an early alert for Patient A, highlighting the divergence from Patient B's trajectory and supporting a clinical decision to switch disease-modifying antirheumatic drugs (eg, to a targeted biologic) earlier in the care pathway. This comparative analysis demonstrates how the RDT resolves heterogeneity that is invisible to snapshot-based assessments, ensuring Patient A receives aggressive intervention, while Patient B avoids unnecessary escalation.

Discussion

Key Contributions

To deliver on the promise of precision medicine, we require tools that can identify meaningful disease subgroups and provide treatment guidance with minimal trial and error. The necessary multimodal patient data, ranging from clinical notes and wet lab diagnostics to imaging and genomic profiling, often exists but is siloed and underused. The modular MDT architecture presented here offers a flexible and scalable approach for harmonizing, modeling, and learning from this diverse data landscape.

The central contribution of this work is the proposal of a modular, multimodal DT architecture that is designed to be robust and clinically relevant. By dividing the learning task into modality-specific foundation models, each trained on a particular data type, the architecture allows for independent development and validation. This modularity enables model optimization for each data type, supports gradual system expansion as new data streams become available, and facilitates cross-institutional collaboration. A major strength is the architecture's ability to operate with missing data modalities. Embeddings can be created from available modalities, allowing predictions and patient similarity calculations to be made even when data is incomplete. The embedding neighborhood around a patient enables case-based reasoning, providing a human-interpretable link between new and historical cases and offering immediate clinical insight. The architecture also supports incremental learning and dynamic model updates. As more patient data become available, embeddings can be recomputed without retraining the entire system. This adaptability is crucial for real-world deployment. Additionally, it improves interpretability by allowing clinicians to trace recommendations back to similar historical cases. Finally, by structuring predictions around multimodal embeddings, the architecture is compatible with recent advances in large language models. These large language models could further enrich the system by supporting natural language queries, summarizing patient records, or generating clinical hypotheses.

Comparative analysis with existing frameworks: to contextualize the novelty of the RDT, it is necessary to contrast it with existing machine learning architectures commonly used in medical informatics. While standard approaches such as unimodal time-series models (eg, RNNs/long short-term memories) or early fusion static models have proven useful for specific tasks, they face significant limitations when applied to the heterogeneous and fragmented nature of RD data. As summarized in Table 1, the RDT distinguishes itself through its intermediate fusion strategy and Transformer-based backbone. Unlike early fusion models that require complete data vectors (often necessitating artificial imputation) or late fusion ensembles that fail to capture feature interactions, the RDT leverages joint embeddings to maintain high resilience against missing data while preserving cross-modal learning.

XSL•FO
RenderX

Table . Comparative analysis of the proposed Rheumatic Digital Twin framework vs traditional machine learning architectures.

Feature	Unimodal time-series (RNN[a]/LSTM[b])	Early fusion static models	Late fusion ensembles	Proposed RDT[c] (intermediate fusion)
Data integration	Single modality only (eg, labs)	Concatenates raw inputs at start	Averages predictions at end	Fuses learned embeddings in latent space
Temporal handling	Good, but struggles with irregular gaps	None (static snapshots)	Limited (varies by submodel)	High (Transformers+Time-delta tokens)
Missing data	Low resilience (gaps break sequence)	Low resilience (needs full vector)	High resilience	High resilience (via attention masking)
Cross-modal learning	None	High (but brittle)	Low (no feature interaction)	High (via joint embedding alignment)
Scalability	Low (hard to add new data types)	Low (must retrain entire model)	High	High (modular components)

[a]RNN: recurrent neural network.

[b]LSTM: long short-term memory.

[c]RDT: Rheumatic Digital Twin.

Key Challenges

Implementing this architecture in practice involves overcoming several fundamental challenges. First, multimodal data integration is difficult because health care data is heterogeneous in structure, format, and semantics. Harmonizing data across institutions also introduces variation and potential bias. A related challenge is embedding coherence. Because each modality is embedded independently, there is a risk of losing crucial cross-modal interactions. Future work may need to move toward joint embedding models or employ alignment techniques to preserve these dependencies.

It is important to note that while we discuss federated learning (FL) as a promising architecture for privacy preservation, it represents only one potential pathway for implementation among other centralized or hybrid approaches. In scenarios where FL is adopted, our framework assumes that paired samples (eg, imaging and reports) for a given patient are colocated on the same node. This assumption facilitates local contrastive learning by ensuring simultaneous access to paired modalities during the optimization of local objectives. However, this setup shifts the technical challenge toward global alignment: while local nodes can align their own modalities effectively, ensuring that the learned latent spaces remain consistent across heterogeneous nodes (non–independent and identically distributed data) without centralized data sharing requires robust aggregation strategies to prevent model divergence.

Validation is another substantial hurdle. A DT system that integrates multiple types of predictions and use cases defies easy benchmarking. Furthermore, the MDT's influence on patient care can create feedback loops that shift data distribution, requiring continuous evaluation and drift detection [35]. Building gold-standard validation datasets that reflect real clinical variation remains an open challenge. Finally, methodological challenges are inherent to real-world health care data. Noise, missingness, and systemic bias are pervasive and can compromise model quality [36]. Correcting these issues requires statistical techniques and continual clinician involvement [37].

While the reliance on foundation models offers significant predictive power, implementing such an architecture in practice involves overcoming substantial computational and practical hurdles. Developing true foundation models typically requires massive data volumes and high-performance computing resources, which can be prohibitively expensive for individual academic medical centers.

Our modular architecture specifically mitigates these challenges by separating the learning task into independent, modality-specific streams. Rather than attempting to train a single monolithic model on all data types simultaneously, a task that would require immense memory and synchronized data access, our framework allows for the distributed, parallel training of smaller unimodal models. This approach enables institutions to optimize and validate a text-based model (eg, for clinical notes) or an image-based model independently, significantly reducing the immediate hardware requirements for any single training phase. By leveraging pretrained weights from existing domain-specific models and focusing local resources on fine-tuning and joint embedding, the RDT becomes a computationally feasible solution for real-world clinical environments.

Key Considerations

Even with technical challenges addressed, several broader considerations remain. Two of the most important considerations are trust and interpretability. Clinicians must understand not just what the system recommends, but why, and they must retain final decision-making authority. Scalability and infrastructure are also critical. The system must integrate with hospital information systems, meet stringent privacy and security requirements, and be usable in time-sensitive clinical workflows [38]. This requires careful design of interfaces and data governance mechanisms. Ethical and regulatory concerns must also be considered [39]. Systems must be designed with safeguards that ensure transparency, fairness, and accountability. Risks, such as reinforcing existing disparities, must be explicitly mitigated. Lastly, there is the question of generalizability. An MDT that performs well in one hospital may not generalize to others without recalibration.

XSL•FO
RenderX

To address the challenge of poor generalizability across institutions, we propose a strategy that leverages the modularity of the RDT through tiered transfer learning. In this framework, modality-specific foundation models are first pretrained on large, multi-institutional datasets to capture broad, generalizable representations of medical data. Because these models are domain-specific rather than site-specific, they can be developed through collaborative efforts, such as, but not limited to, FL and local fine-tuning, which allow for the extraction of global knowledge while maintaining strict data privacy.

It is crucial to distinguish this resource-intensive model creation phase from the subsequent model adoption phase. While the development of a new foundation model requires high-performance computing provided by a research consortium, individual clinical centers can then adopt these globally pretrained foundation models and perform local fine-tuning on the joint embedding layer and task-specific pipelines. This ensures that the computational burden of training from scratch is borne by the collective federation, while local deployment remains computationally feasible for standard hospital infrastructure.

Individual clinical centers can then adopt these global foundation models and perform local fine-tuning on the joint embedding layer and task-specific pipelines. This localized step allows the system to adapt to institution-specific coding nuances, local demographics, and varying clinical practices without requiring the immense resources needed to build a model from scratch. However, this approach is not without risks; we must remain vigilant against "catastrophic forgetting," where the model loses its generalizable knowledge during local adaptation, and "domain overfitting," where the DT becomes so attuned to local biases that it loses its predictive utility for outlier patients. Continuous evaluation and drift detection mechanisms are, therefore, essential to ensure that the RDT remains robust as it moves between different health care environments.

Implementation Roadmap: To bridge the gap between proposal and practice, we outline a realistic implementation roadmap centered on 3 pillars of feasibility.

1. Incremental Integration: Institutions can initiate the RDT using common data sources, such as EHR clinical notes and structured laboratory results. More complex and high-cost modalities, such as genomic profiling or advanced imaging, can be integrated into the existing joint embedding layer later without requiring a total system overhaul.
2. Leveraging Foundation Models: The use of pretrained foundation models significantly reduces the local computational burden. By adopting models already trained on massive multimodal datasets (eg, using open-source models like ClinicalBERT for text), clinical centers only need to invest in the relatively low-cost phase of local fine-tuning and joint alignment. In cases where domain-specific models do not yet exist, institutions may participate in FL consortia to distribute the high cost of training a new model; however, once established, the local operational requirement remains limited to efficient fine-tuning rather than the "monolithic" training of entire systems from scratch.
3. Distributed Parallel Training: The modular architecture allows for the parallel processing of different data streams. This distributed approach enables specialized departments (eg, radiology for imaging, clinical chemistry for routine lab values) to maintain and validate their specific foundation models independently, making the overall system more manageable within the existing organizational silos of academic medical centers.

Legal Liability: MDTs raise fundamental questions about responsibility in clinical decision-making. Because MDTs currently function as decision-support tools rather than autonomous agents, existing regulatory frameworks (eg, European Union Medical Device Regulation; Food and Drug Administration Clinical Decision Support guidance) operate under a human-in-the-loop principle: clinicians remain responsible for interpreting MDT outputs within the broader clinical context. At the same time, responsibility is shared with developers, who must ensure model validity, documentation, and ongoing performance, and with health care institutions that govern data quality and safe deployment. MDTs, however, increase the causal complexity of recommendations because predictions arise from a continually updated model instance. When suboptimal outcomes occur, attributing fault across clinician judgment, data provenance, and model design becomes challenging. This underscores the need for rigorous audit trails, version control, and traceability mechanisms and may ultimately require new legal frameworks capable of addressing distributed or tiered liability in hybrid human-artificial intelligence decisions.

Ownership of the MDT: Continually evolving MDTs also challenge traditional notions of data and model ownership. The underlying personal health data remain the legal rights of the patient, whereas the model architecture, algorithms, and proprietary components are typically owned by the developer. Health care institutions, in turn, control the clinical instantiation of the MDT and bear responsibility for its operational use. Because an MDT becomes a patient-specific and dynamically updated derivative of all 3 layers, clear ownership boundaries are difficult to define. A hybrid ownership model is therefore likely to be required, one that distinguishes among rights over data, rights over the computational model, and rights over the instantiated DT used in clinical care. Future governance frameworks will need to delineate these rights explicitly, address portability between institutions, and ensure accountability for the evolving digital representation of the patient.

Validation Strategies

Given that the RDT is presented here as a prospective framework, establishing a robust validation pipeline is critical for its eventual translation into clinical practice. We propose a multitiered validation strategy designed to evaluate both the technical integrity of the model and its practical utility for rheumatologists.

Technical Evaluation of Embeddings: The quality and information density of the learned latent space will be primarily assessed through self-supervised reconstruction tasks. By measuring how accurately the foundation models can reconstruct original input data, such as specific lab values or clinical text

XSL•FO
RenderX

spans, from their latent representations, we can evaluate whether the embeddings effectively capture the essential features of the patient journey. Additionally, the alignment among modalities can be tested using cross-modal retrieval metrics, determining whether a clinical query can accurately retrieve the corresponding imaging or genomic profile in the shared space.

Assessment of Clinical Utility: To confirm clinical usefulness, the model must be validated against objective "gold standards" in rheumatology. We propose benchmarking the RDT's predictive capabilities on high-stakes tasks, such as forecasting the timing of disease flares, predicting treatment switches (eg, failure of a specific disease-modifying antirheumatic drug), and estimating future disease activity scores (DAS-28).

To rigorously evaluate performance on these tasks, we propose a comprehensive benchmarking strategy. First, to ensure clinical relevance, the model will be validated against established clinical gold standards, such as the European Alliance of Associations for Rheumatology recommendations and human expert consensus. While we acknowledge that human consensus can be subject to interrater variability, it remains the necessary standard for validating utility in fundamental decision-making tasks.

Complementing this, we will employ technical comparisons to isolate architectural benefits. To demonstrate the added value of integrating longitudinal, multimodal data over standard clinical practice, predictions will be compared against traditional "snapshot" baselines (eg, logistic regression based on the most recent clinical assessment). Finally, to isolate the specific performance gains attributable to the proposed Transformer architecture, we will compare the RDT against established longitudinal methods, such as RNNs or long short-term memories or gradient-boosted trees (eg, eXtreme Gradient Boosting) trained on aggregated historical features.

Interpretability as Validation: Finally, we propose "neighborhood-based reasoning" as a qualitative validation method. By examining a patient's nearest neighbors in the embedding space, clinicians can verify that mathematical proximity correlates with established medical knowledge. For instance, validation would involve confirming that "neighboring" patients in the DT space historically shared similar inflammatory endotypes or biological responses to specific therapies. This step ensures the model identifies meaningful medical patterns rather than spurious data correlations, providing a transparent and verifiable basis for clinical decision-making.

However, it is important to acknowledge that, while the retrieval of historical neighbors offers clinical transparency, the similarity metric itself is derived from high-dimensional, nonlinear embeddings. Therefore, validation must confirm that mathematical proximity in this latent space consistently maps to clinically interpretable features, thereby establishing a verifiable, if not fully transparent, basis for decision-making.

Concluding Remarks

In conclusion, the health care system is undergoing rapid and pervasive digitization, resulting in an unprecedented availability of data across all stages of care. This data ubiquity creates new opportunities for purely data-driven approaches, such as the hereby-proposed RDT system, to support critical clinical tasks, including disease subtyping and therapy referral. As ongoing technological advances continue to expand our capacities for data acquisition, exchange, storage, and computation, we believe that such DT-based frameworks will play an increasingly central role in enabling more precise, personalized, and effective care, particularly for intrinsically heterogeneous clinical application fields such as rheumatologic care.

Acknowledgments

The authors sincerely thank the reviewers for their constructive feedback process. During the preparation of this work, the author(s) used ChatGPT (GPT-5.1) and Google Gemini (Gemini 2.5) to assist with writing in the form of grammar correction, as well as for text sharpening and upscaling in figures. After using these tools or services, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

This project has received funding from the Horizon Europe programme under grant agreement number 10108071 (Stratification of Patients using advanced Integrative modeling of Data for the Rheumatic Diseases [SPIDERR]). We thank all SPIDERR consortium members for their useful discussions on MDT implementation within rheumatology.

Data Availability

No original datasets were generated or analyzed during the current study. Data sharing is not applicable to this study, as no new data were created or analyzed within this theoretical framework.

Authors' Contributions

Writing – original draft: DS, RK, MR, EBA
Writing – review & editing: DS, RK, MR, EBA

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Katsoulakis E, Wang Q, Wu H, et al. Digital twins for health: a scoping review. NPJ Digit Med 2024 Mar 22;7(1):77. [doi: 10.1038/s41746-024-01073-0] [Medline: 38519626]
2. Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review. Brief Bioinformatics 2022 Mar 10;23(2):bbab569. [doi: 10.1093/bib/bbab569]
3. Lewis MJ, Barnes MR, Blighe K, et al. Molecular portraits of early rheumatoid arthritis identify clinical and treatment response phenotypes. Cell Rep 2019 Aug 27;28(9):2455-2470.e5. [doi: 10.1016/j.celrep.2019.07.091] [Medline: 31461658]
4. Knevel R, le Cessie S, Terao CC, et al. Using genetics to prioritize diagnoses for rheumatology outpatients with inflammatory arthritis. Sci Transl Med 2020 May 27;12(545):eaay1548. [doi: 10.1126/scitranslmed.aay1548] [Medline: 32461333]
5. He Y, Huang F, Jiang X, et al. Foundation model for advancing healthcare: challenges, opportunities and future directions. IEEE Rev Biomed Eng 2025;18:172-191. [doi: 10.1109/RBME.2024.3496744] [Medline: 39531565]
6. Karaman I, Sebin B. From data-driven cities to data-driven tumors: dynamic digital twins for adaptive oncology. Front Artif Intell 2025;8:1624877. [doi: 10.3389/frai.2025.1624877] [Medline: 40785836]
7. Halpern GA, Nemet M, Gowda DM, Kilickaya O, Lal A. Advances and utility of digital twins in critical care and acute care medicine: a narrative review. J Yeungnam Med Sci 2025;42:9. [doi: 10.12701/jyms.2024.01053] [Medline: 39587767]
8. Sequí-Sabater JM, Benavent D. Artificial intelligence in rheumatology research: what is it good for? RMD Open 2025 Jan 8;11(1):e004309. [doi: 10.1136/rmdopen-2024-004309] [Medline: 39778924]
9. Ringeval M, Etindele Sosso FA, Cousineau M, Paré G. Advancing health care with digital twins: meta-review of applications and implementation challenges. J Med Internet Res 2025 Feb 19;27:e69544. [doi: 10.2196/69544] [Medline: 39969978]
10. National Academies of Sciences, Engineering, and Medicine; National Academy of Engineering; Division on Earth and Life Studies; Division on Engineering and Physical Sciences; Board on Atmospheric Sciences and Climate; Board on Life Sciences; Computer Science and Telecommunications Board; Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences and Analytics; Committee on Foundational Research Gaps and Future Directions for Digital Twins. Foundational Research Gaps and Future Directions for Digital Twins: National Academies Press; 2024. [doi: 10.17226/26894]
11. Zhang S, Yao L, Sun A, Tay Y. Deep learning based recommender system: a survey and new perspectives. ACM Comput Surv 2019;52(1):1-38. [doi: 10.1145/3285029]
12. Ochoa JGD, Csiszár O, Schimper T. Medical recommender systems based on continuous-valued logic and multi-criteria decision operators, using interpretable neural networks. BMC Med Inform Decis Mak 2021 Jun 11;21(1):186. [doi: 10.1186/s12911-021-01553-3] [Medline: 34112161]
13. Maarseveen TD, Maurits MP, Coletto LA, et al. Location and amount of joint involvement differentiates rheumatoid arthritis into different clinical subsets. NPJ Digit Med 2025 Oct 23;8(1):623. [doi: 10.1038/s41746-025-01997-1] [Medline: 41131344]
14. Steinz N, Maarseveen TD, van den Akker EB, et al. Time-independent disease state identification defines distinct trajectories determined by localised vs systemic inflammation in patients with early rheumatoid arthritis. Ann Rheum Dis 2025 Aug;84(8):1301-1312. [doi: 10.1016/j.ard.2025.04.011] [Medline: 40348636]
15. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. arXiv. Preprint posted online on Aug 16, 2021. [doi: 10.48550/arXiv.2108.07258]
16. Wang WC, Ahn E, Feng D, Kim J. A review of predictive and contrastive self-supervised learning for medical images. Mach Intell Res 2023 Aug;20(4):483-513. [doi: 10.1007/s11633-022-1406-4]
17. Caruso CM, Soda P, Guarrasi V. MARIA: a multimodal transformer model for incomplete healthcare data. Comput Biol Med 2025 Sep;196(Pt C):110843. [doi: 10.1016/j.compbiomed.2025.110843] [Medline: 40784080]
18. Zhou HY, Yu Y, Wang C, et al. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. Nat Biomed Eng 2023;7(6):743-755. [doi: 10.1038/s41551-023-01045-x]
19. Bai S, Kolter JZ, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv. Preprint posted online on Mar 4, 2018. [doi: 10.48550/arXiv.1803.01271]
20. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Sci Rep 2018 Apr 17;8(1):6085. [doi: 10.1038/s41598-018-24271-9] [Medline: 29666385]
21. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Presented at: Advances in Neural Information Processing Systems 30 (NIPS 2017); Dec 4-9, 2017 URL: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html [accessed 2026-04-04]
22. Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw 1994;5(2):157-166. [doi: 10.1109/72.279181] [Medline: 18267787]
23. Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks. Presented at: ICML'13: Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28; Jun 16-21, 2013; Atlanta GA, USA p. 1310-1318 URL: https://dl.acm.org/doi/10.5555/3042817.3043083 [accessed 2026-04-04]

XSL•FO
RenderX

24. Beltagy I, Peters ME, Cohan A. Longformer: the long-document transformer. arXiv. Preprint posted online on Apr 10, 2020 URL: https://arxiv.org/abs/2004.05150 [accessed 2026-04-04]
25. Zaheer M, Guruganesh G, Dubey A, et al. Big bird: transformers for longer sequences. Presented at: NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems; Dec 6-12, 2020; Vancouver BC, Canada p. 17283-17297 URL: https://dl.acm.org/doi/abs/10.5555/3495724.3497174 [accessed 2026-04-04]
26. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. Presented at: ICML'20: Proceedings of the 37th International Conference on Machine Learning; Jul 13-18, 2020; Virtual (online) p. 1597-1607 URL: https://dl.acm.org/doi/10.5555/3524938.3525087 [accessed 2026-04-04]
27. Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision. Presented at: Proceedings of the 38th International Conference on Machine Learning; Jul 18-24, 2021; Virtual Event p. 8748-8763 URL: https://proceedings.mlr.press/v139/radford21a.html [accessed 2026-04-04]
28. Zhang Y, Jiang H, Miura Y, Manning CD, Langlotz CP. Contrastive learning of medical visual representations from paired images and text. Presented at: Machine Learning for Healthcare Conference; Aug 5-6, 2022; Durham, NC, USA p. 2-25 URL: https://proceedings.mlr.press/v182/zhang22a/zhang22a.pdf [accessed 2026-04-04]
29. He K, Chen X, Xie S, Li Y, Dollar P, Girshick R. Masked autoencoders are scalable vision learners. Presented at: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 18-24, 2022; New Orleans, LA, USA p. 16000-16009. [doi: 10.1109/CVPR52688.2022.01553]
30. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature 2023 Apr;616(7956):259-265. [doi: 10.1038/s41586-023-05881-4] [Medline: 37045921]
31. Baltrusaitis T, Ahuja C, Morency LP. Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell 2019 Feb;41(2):423-443. [doi: 10.1109/TPAMI.2018.2798607] [Medline: 29994351]
32. Caruana R. Multitask learning. Mach Learn 1997 Jul;28(1):41-75. [doi: 10.1023/A:1007379606734]
33. Wang T, Li F, Zhu L, Li J, Zhang Z, Shen HT. Cross-Modal Retrieval: A Systematic Review of Methods and Future Directions. Proc IEEE 2024;112(11):1716-1754. [doi: 10.1109/JPROC.2024.3525147]
34. Sharafoddini A, Dubin JA, Lee J. Patient similarity in prediction models based on health data: a scoping review. JMIR Med Inform 2017 Mar 3;5(1):e7. [doi: 10.2196/medinform.6730] [Medline: 28258046]
35. Sahiner B, Chen W, Samala RK, Petrick N. Data drift in medical machine learning: implications and potential remedies. Br J Radiol 2023 Oct;96(1150):20220878. [doi: 10.1259/bjr.20220878] [Medline: 36971405]
36. Roh Y, Heo G, Whang SE. A survey on data collection for machine learning: a big data - AI integration perspective. IEEE Trans Knowl Data Eng 2021;33(4):1328-1347. [doi: 10.1109/TKDE.2019.2946162]
37. Austin PC, White IR, Lee DS, van Buuren S. Missing data in clinical research: a tutorial on multiple imputation. Can J Cardiol 2021 Sep;37(9):1322-1331. [doi: 10.1016/j.cjca.2020.11.010] [Medline: 33276049]
38. Oladele S. Challenges and solutions for interoperability in health informatics: a focus on data privacy. Health Inform Rev 2024 [FREE Full text]
39. Howe EG, Elenberg F. Ethical challenges posed by big data. Innov Clin Neurosci 2020;17(10-12):24-30. [Medline: 33898098]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
DT: digital twin
EHR: electronic health record
FDA: Food and Drug Administration
FL: federated learning
***ICD-10*:** International Classification of Diseases, 10th Revision
MDT: medical digital twin
RA: rheumatoid arthritis
RD: rheumatic disease
RDT: Rheumatic Digital Twin
RNN: recurrent neural network
SPIDERR: Stratification of Patients using advanced Integrative modeling of Data for the Rheumatic Diseases

XSL•FO
RenderX

Edited by A Coristine; submitted 29.Oct.2025; peer-reviewed by KV Sangaraju, T Hügle; revised version received 11.Feb.2026; accepted 14.Feb.2026; published 04.May.2026.

Please cite as:
Selani D, Knevel R, Reinders M, van den Akker EB
Rheumatic Digital Twin: Proposed Machine Learning–Based Multimodal Framework to Inform Clinical Decision-Making
J Med Internet Res 2026;28:e86763
URL: https://www.jmir.org/2026/1/e86763
doi:10.2196/86763

© Daniyal Selani, Rachel Knevel, Marcel Reinders, Erik B van den Akker. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 4.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI in Clinical Decision Support Systems: Promising Applications and Strategies for Managing Data Challenges

Jennifer E Daly[1], PhD; Dursun Delen[2,3], PhD; Zheng Han[4], PhD; River Smith[5], PhD; Jacqueline Honerlaw[6], RN; Kelly Cho[6,7], PhD; Bridget Bennett[8], MD; Jennifer Sippel[8], PhD

[1]Office of the Chief of Staff, Eastern Oklahoma VA Health Care System, 1101 Honor Heights Drive, Muskogee, United States

[2]Department of Management Science and Information Systems, Spears School of Business, Center for Health Systems Innovation, Oklahoma State University, Stillwater, OK, United States

[3]Faculty of Engineering and Natural Sciences, Department of Industrial Engineering, Istinye University, Sariyer, Istanbul, Turkey

[4]Department of Biomedical Engineering, University of Central Oklahoma, Edmond, OK, United States

[5]Office of the Associate Chief of Staff for Research, Eastern Oklahoma VA Health Care System, Muskogee, OK, United States

[6]Division of Population Health and Data Science, Massachusetts Veterans Epidemiology Research and Information Collaborative, VA Boston Healthcare System, Boston, MA, United States

[7]Division of Aging, Mass General Brigham, Harvard University, Cambridge, Boston, MA, United States

[8]Spinal Cord Injuries and Disorders National Program Office, United States Department of Veterans Affairs, Washington, DC, United States

Corresponding Author:
Jennifer E Daly, PhD
Office of the Chief of Staff, Eastern Oklahoma VA Health Care System, 1101 Honor Heights Drive, Muskogee, United States

Abstract

The translation of big data analytics and artificial intelligence (AI) into clinical decision support systems (CDSSs) has advanced from proof of concept to real-world clinical practice. AI-informed CDSSs show measurable improvements in diagnostic accuracy, risk stratification, resource use, and patient outcomes compared to traditional models, offering the potential to assist clinicians in managing symptom complexity and uncertainty in health care delivery. Despite this potential, access to large amounts of high-quality and granular data remains one of the most significant bottlenecks to AI-enabled CDSSs. We argue that as health care systems increasingly adopt data-driven decision support, addressing the challenges of data accessibility and protection is essential to realizing the full potential of AI in clinical medicine. We use selected case examples of AI-informed CDSSs in oncology, organ transplantation, diabetic retinopathy, epilepsy, spinal cord injury, rare disease diagnosis, and emergency medicine to illustrate opportunities and challenges related to AI's potential to improve patient outcomes. We discuss public and semipublic, medical institutional and commercial, and government and national data sources that are currently available for the development of CDSSs and highlight the practical and ethical constraints associated with these data. We consider alternative data resources and ways in which health care systems can strengthen data ecosystems to increase AI-driven CDSS efficacy and implementation to improve patient outcomes.

(J Med Internet Res 2026;28:e71532) doi:10.2196/71532

KEYWORDS

artificial intelligence; AI; delivery of health care; decision-making; machine learning; precision medicine; big data; population health; decision support systems; clinical medical informatics

Introduction

Artificial intelligence (AI) and big data analytics have emerged as powerful tools for medical decision support, offering the potential to augment clinicians' diagnostic accuracy, improve prognostic estimation, and support more timely and personalized clinical decisions. AI-based clinical decision support systems (CDSSs) leverage large volumes of clinical, imaging, patient-generated, and administrative data to assist clinicians in managing symptom complexity and uncertainty in their delivery of health care.

Early CDSSs emphasized rule-based models intended to reduce errors and improve adherence to evidence-based care [1]. As electronic health records (EHRs) became more widely available, machine learning (ML) models were found to match or exceed clinician performance in domains such as medical imaging interpretation, early detection of clinical deterioration, and prediction of disease progression [2,3]. Natural language processing techniques further extended decision support by extracting clinically meaningful information from unstructured EHR data, enabling more comprehensive risk stratification and care planning [4]. Systematic reviews suggest that these tools

XSL•FO
RenderX

may reduce cognitive burden and variability in decision-making while improving efficiency [5,6].

Recent research documents a maturation of AI-enabled CDSSs with a focus on real-world implementation. Studies highlight advances in explainable AI, clinician-AI interactions, and evaluation frameworks that move beyond algorithmic accuracy to assess clinical usefulness and workflow integration [5,6]. These studies identify key determinants of clinician use and trust, establishing factors such as transparency, usability, workflow integration, and training as central to adoption [7,8]. A meta-analysis grounded in the unified theory of acceptance and use of technology revealed that performance expectancy, effort expectancy, social influence, and facilitating conditions significantly influence health care practitioners' intentions to use AI-enabled CDSSs [8]. Concurrently, research emphasizes the importance of explainable AI methodologies to support interpretability and user confidence, as well as human-centered evaluation frameworks that align technical designs with real-world clinical needs [6,9,10]. Additional work explores persistent problems and barriers associated with implementation, including technical fit, perceived risk, and task alignment, suggesting that successful CDSS integration hinges not only on algorithmic performance but also on organizational, human, and contextual factors [7,11-13]. These studies collectively underscore that while AI-driven CDSSs have strong potential to enhance clinical decision-making, their effective deployment requires attention to human factors, trust, usability, and structured implementation strategies.

The literature also identifies substantial challenges that limit the widespread and equitable adoption of AI in medical decision support. In addition to concerns about interpretability, bias, and workflow integration, access to high-quality data remains a critical barrier. Many health care systems face fragmented data infrastructures, restrictive data governance policies, and privacy regulations that limit data sharing across institutions [14]. Research has shown that AI models trained on narrow or nonrepresentative datasets may perform poorly when deployed in diverse clinical settings, exacerbating disparities in care [15]. Furthermore, smaller or resource-limited organizations often lack the technical capacity or data volume required to develop or adopt robust AI systems.

This paper aims to clarify both the promise of AI-driven medical decision support and the structural data access challenges that must be addressed to realize its full clinical value. We argue that as health care systems increasingly adopt data-driven decision support, addressing the challenges of data accessibility and protection is essential to realizing the full potential of AI in clinical medicine. We review case studies and discuss synthetic data and federated networks as potential solutions for working with sensitive data. We offer a set of core principles to guide the development of hybrid data ecosystems for AI-enabled CDSSs that can make high-quality data easy to discover, use, and trust. Figure 1 provides a visual overview of the promise and challenges of AI-enabled CDSSs and potential solutions, which are discussed in greater detail in this paper.

Figure 1. Visual overview of artificial intelligence (AI)–informed clinical decision support for selected case examples and potential solutions to overcome data access challenges.

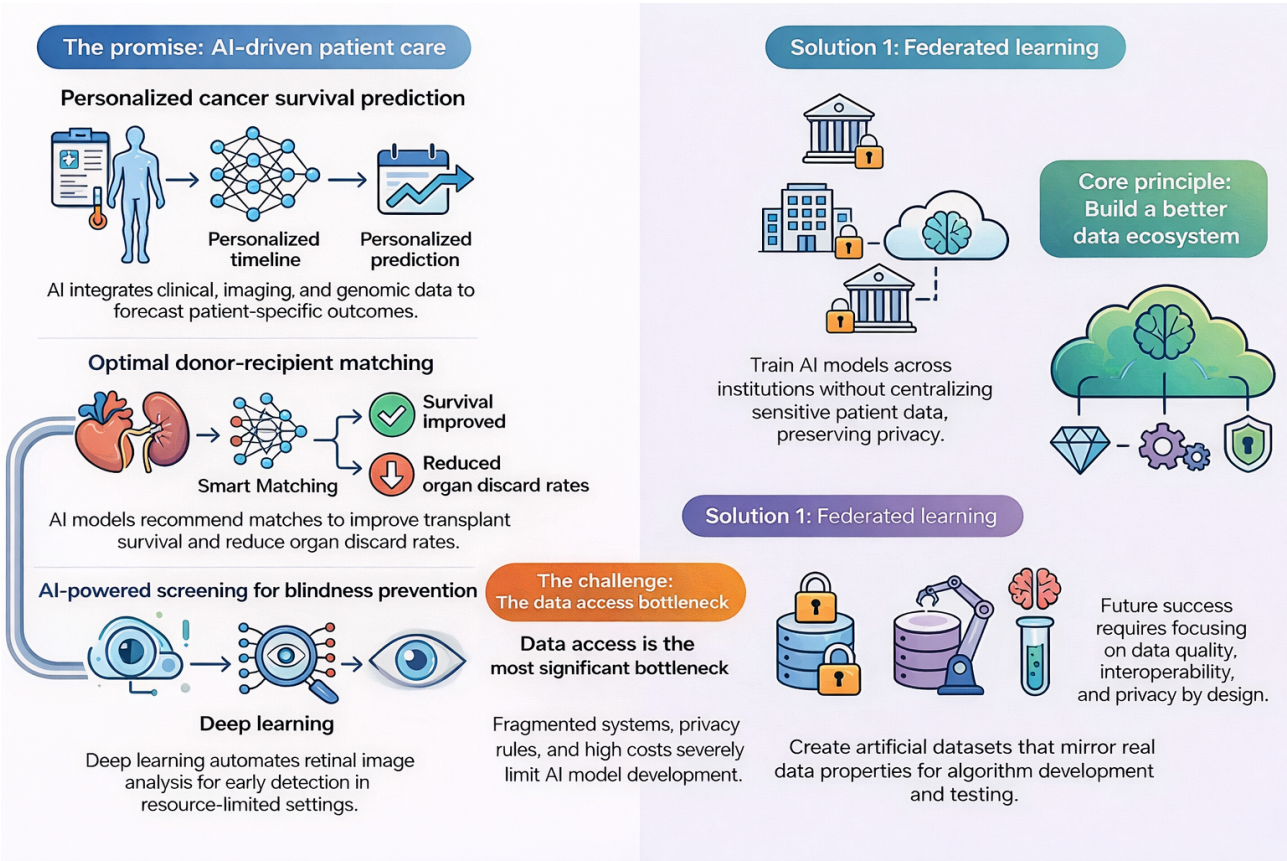

XSL•FO
RenderX

Data Access and Challenges of AI-Enabled CDSSs

Selected Applications

The translation of big data analytics and AI in CDSSs has moved from proof of concept to real-world clinical impact. Data-driven CDSSs demonstrate measurable improvements in diagnostic accuracy, risk stratification, resource utilization, and patient outcomes across multiple areas, including oncology, organ transplantation, chronic disease management, emergency medicine, rare disease diagnosis, and spinal cord injury. These case studies illustrate the opportunities and challenges in deploying AI-enabled CDSSs in clinical medicine.

Oncology: Personalized Cancer Characterization and Survival Prediction

Cancer care has been among the earliest and most fertile domains for AI-driven CDSSs due to the availability of high-dimensional data, including clinical variables, imaging, genomic profiles, and longitudinal outcomes. Early work demonstrated that ML models could outperform traditional statistical approaches in classifying cancer subtypes and predicting survival, laying the foundation for modern precision oncology [16].

Recent studies have advanced this paradigm by integrating heterogeneous data sources to enable patient-specific survival prediction and explainable risk stratification [17]. Deep learning and ensemble models have been used to uncover latent patterns across tumor biology, treatment response, and demographic factors, enabling individualized prognostic assessments rather than population-level averages [18]. Importantly, contemporary research emphasizes model interpretability, using explainable AI techniques to identify which clinical or molecular factors drive predictions, thereby increasing clinician trust and facilitating shared decision-making [19-21].

AI-driven CDSSs support oncologists' treatment selection, risk-adjusted follow-up scheduling, and identification of high-risk patients who may benefit from aggressive or experimental therapies. However, challenges remain, including data imbalance across cancer subtypes, model generalizability across institutions, and ethical concerns surrounding algorithmic bias when predictions influence life-critical decisions.

Organ Transplantation: AI-Enabled Optimal Donor-Recipient Matching

Organ transplantation presents a high-stakes optimization problem where mismatches can lead to graft failure, patient morbidity, and wasted organs. Traditional allocation systems rely on rule-based scoring methods that may not fully capture complex, nonlinear interactions among the donor, the recipient, and contextual factors [22].

Recent AI-driven CDSS approaches apply ML, optimization algorithms, and predictive analytics to maximize graft utilization while improving posttransplant survival [20]. These systems analyze large registries of donor and recipient data, learning patterns associated with graft longevity, rejection risk, and long-term outcomes. By simulating multiple allocation scenarios, AI models can recommend matches that balance equity, urgency, and expected benefit more effectively than static scoring systems [21].

Notably, these approaches highlight the role of big data in addressing systemic inefficiencies—reducing organ discard rates and improving fairness in allocation. Nevertheless, their adoption faces regulatory, ethical, and transparency challenges, particularly when algorithmic recommendations conflict with established allocation policies or clinician judgment.

Diabetic Retinopathy: AI-Based Screening in Resource-Limited Settings

Diabetic retinopathy (DR) is a leading cause of preventable blindness, yet early detection remains limited in rural and underserved regions due to shortages of ophthalmologists [23]. AI-enabled big data analytics has emerged as a transformative solution, leveraging deep learning models trained on large retinal image datasets to automatically detect DR with accuracy that is comparable to that of expert clinicians [24]. These CDSS applications enable scalable, low-cost screening at the point of care, allowing for early referral and intervention before irreversible vision loss occurs [25]. Their integration into primary care and mobile screening units demonstrates how big data and AI can bridge health care access gaps and improve population-level outcomes.

However, despite the promise of DR CDSSs, challenges persist regarding dataset representativeness, robustness across imaging devices, and clinical workflow integration. Additionally, regulatory approval and medico-legal accountability remain critical considerations when automated systems are used for diagnostic decision-making.

Traumatic Brain Injury and Posttraumatic Epilepsy

Traumatic brain injury (TBI) is a major global health concern, affecting millions annually and often leading to severe neurological sequelae, including posttraumatic epilepsy (PTE) [26]. Predicting the risk of PTE among patients with TBI is crucial for effective prevention. ML models that consolidate longitudinal patient encounter data, disease history, and laboratory records have emerged as a promising alternative to radiological or pathological data for risk assessment [27-29]. However, traditional ML models face limitations, including data sparsity due to their reliance on Euclidean data structure [30], requiring imputation that can introduce bias [16,17], and challenges in representing temporal data despite the rich sequence-based information embedded in patient encounters [31,32].

As a solution, Ramamurthy et al [33] leveraged a heterogeneous graph attention network (HeteroGAT) on large-scale data from the Cerner Real-World Data. By encoding hospital visits as patient-to-diagnosis edges within the graph structure, this HeteroGAT model eliminates the need for imputation, efficiently compresses data, and retains temporal information. The trained model was shown to outperform traditional models in predicting PTE risk among patients with TBI.

XSL•FO
RenderX

Spinal Cord Injury: Improving Rehabilitation Planning and Prevention

A growing body of evidence shows that AI and ML models improve the prediction of clinically meaningful outcomes after spinal cord injury or disorder (SCI) [34-37]. These models forecast Spinal Cord Independence Measure scores [34], gait recovery [36], and changes in American Spinal Injury Association Impairment Scale grade [37] more accurately than traditional methods, supporting improved risk stratification and individualized rehabilitation planning. Multimodal approaches integrating clinical notes, imaging, electrophysiology, biomarkers, and social determinant data have enabled explainable, clinician-facing tools that enhance both predictive performance and interpretability. Advances in neuromodulation and device-assisted rehabilitation further suggest complementary pathways in which ML-guided patient selection and dosing may increase real-world benefit [38-48]. Additional applications include algorithms to predict postoperative health-related outcomes and guide surgical decisions [38,39], AI-trained robotics to improve upper extremity function through noninvasive electrical stimulation in tetraplegia [41], and molecular discovery efforts targeting neural regeneration by modulating inflammatory processes at the site of SCI [42].

Despite benefits and advances, most SCI studies to date are retrospective, single center, or registry based, restricting external generalizability and increasing the risk of overfitting. Public SCI datasets are relatively small and heterogeneous, and although there is a greater amount of Veterans Health Administration (VHA) data, they can be fragmented across clinical systems, complicating efforts to build robust, interoperable models. Moreover, few studies extend beyond prediction to demonstrate whether ML-guided care improves long-term outcomes such as functional recovery, quality of life, or reduced health care use.

Rare Disease Characterization: Accelerating Diagnosis and Treatment Pathways

Rare diseases such as systemic lupus erythematosus and Crohn disease often present with heterogeneous symptoms that evolve over time, which can lead to delayed or missed diagnoses. AI-driven CDSSs have shown promise in characterizing disease subtypes, identifying early diagnostic signatures, and predicting disease progression using longitudinal EHR data [49].

By mining structured and unstructured clinical data, these systems uncover subtle patterns that may elude traditional diagnostic approaches. Early and accurate diagnosis enables timely treatment initiation, improved symptom control, and reduced long-term complications—outcomes particularly critical in diseases in which diagnostic delays can span years [50].

The primary challenge with rare diseases lies in limited sample sizes and fragmented data, which complicate model training and validation. Federated learning and multi-institutional data sharing, which are discussed in later sections, are emerging as potential solutions to overcome these barriers while preserving patient privacy.

Emergency Medicine: Reducing Emergency Department Bounce Backs Using Unstructured Data

Emergency department (ED) bounce backs—unplanned return visits shortly after discharge—are costly, burdensome, and often indicative of missed diagnoses or inadequate discharge planning. Recent CDSS research leverages natural language processing and ML to analyze unstructured clinician notes, capturing clinical nuance missing from coded data [51].

These systems identify patients at high risk of bounce back by modeling symptom descriptions, clinician assessments, and contextual factors documented in free text. By flagging high-risk cases in real time, CDSSs support more informed discharge decisions, targeted follow-up, and resource allocation, ultimately improving patient safety and reducing unnecessary ED use [52].

Key challenges with ED models include variability in clinical documentation, model explainability, and real-time integration into fast-paced ED workflows. Nonetheless, this case exemplifies the untapped value of unstructured EHR data in advancing AI-driven clinical decision support.

Summary of AI-Enabled CDSS Applications

Collectively, these case studies demonstrate that big data and AI-enabled CDSSs can meaningfully improve diagnosis, prevention, risk stratification, and outcomes across diverse clinical domains. However, their success depends not only on algorithmic sophistication but also on data quality, interpretability, ethical governance, and seamless clinical integration. As health care systems increasingly adopt data-driven decision support, addressing these challenges is essential to realizing the full potential of AI in clinical medicine.

Big Data Resources for AI-Enabled CDSSs

Overview

Despite the demonstrated value of AI and advanced analytics for improving health care decision-making and patient outcomes, access to large amounts of high-quality and granular clinical data remains a major bottleneck. Health care data are inherently sensitive, fragmented across institutions, governed by complex regulatory frameworks, and costly to curate and integrate. Although a small number of public and semipublic datasets have enabled important advances in AI-enabled CDSSs, most clinically rich datasets remain difficult to access for scientific and translational research.

Public and Semipublic Health Care Datasets: Opportunities and Limitations

Several large-scale datasets have played a foundational role in advancing AI-driven CDSSs by providing researchers with access to real-world patient data at scale.

The Surveillance, Epidemiology, and End Results Program provides longitudinal cancer incidence, treatment, and survival data across 22 US geographic regions, representing approximately 48% of cancer cases nationwide and supporting population-level oncology research [53].

XSL•FO
RenderX

The United Network for Organ Sharing maintains a comprehensive transplant registry containing data on all US organ donation and transplant events since 1987, enabling modeling of donor-recipient matching and graft survival.

The Centers for Medicare and Medicaid Services (CMS) databases offer large administrative claims datasets for studying health care costs, quality, and outcomes. CMS uses a tiered access model that includes publicly available synthetic public use files for exploratory analyses [54] and fee-based access to patient-level data through the Research Data Assistance Center, requiring data use agreements, institutional review board (IRB) approval, and additional safeguards for limited or identifiable datasets [55,56]. This multitiered model demonstrates a pragmatic approach to data democratization that lowers barriers for exploratory research while maintaining rigorous governance for sensitive analyses.

Medical Institution–Based Data Warehouses and Commercial Real-World Data Platforms

To address the limitations of public datasets, researchers have turned to medical institution–managed and commercial real-world data (RWD) repositories, which aggregate richer EHR data across multiple health care systems. These platforms offer greater clinical granularity—including clinical notes, laboratory results, medications, imaging metadata, and care pathway details—enabling more nuanced AI-driven CDSS development and validation. However, their data sharing models, access mechanisms, and computational infrastructures differ markedly from those of government-sponsored repositories, presenting both opportunities and challenges for researchers.

Oracle Real-World Data (ORWD), formerly Cerner Real-World Data, is one of the largest commercial aggregators of real-world evidence, curating deidentified EHR data from over 100 million patients across diverse health care settings in the United States and internationally [57]. The platform integrates claims data, laboratory results, clinical documentation, and medication records from hospital systems, ambulatory clinics, and specialty care centers. ORWD is distinguished by its data sharing and analytics infrastructure designed for pharmaceutical research, regulatory submissions, and comparative effectiveness studies. Access is governed by subscription-based licensing and data use agreements that define permitted research questions, data elements, cohorts, and publication rights; access is often limited to organizations that contribute data to Oracle Health. ORWD provides a secure, cloud-based analytical environment hosted on Amazon Web Services where approved researchers access harmonized datasets without direct download to maintain privacy. The platform supports SQL-based tools (eg, Spark) and R and Python integrations, enabling large-scale model development and validation using distributed computing resources.

The Epic Systems Cosmos platform has emerged as a powerful force in RWD aggregation, leveraging Epic's market position as the leading EHR vendor in the United States [58]. As of 2024, Cosmos contains deidentified records from over 300 million patients across more than 1000 health care organizations. Similar to ORWD, Epic's data sharing structure operates through a collaborative research network model in which participating health systems agree to contribute data in exchange for access. Notably, Cosmos has enabled large-scale studies on COVID-19 outcomes [59], opioid prescribing patterns [60], and other health issues. Cosmos represents another pragmatic example of how EHR vendors such as Oracle Health can leverage their market position to create valuable research infrastructure while balancing commercial interests with public health benefit.

Collectively, ORWD and Epic Cosmos illustrate a commercial data sharing paradigm that contrasts sharply with government-sponsored open access models. These platforms offer exceptional clinical granularity, scale, and near–real-time data availability, enabling research that would be infeasible with traditional registry or claims-based datasets. Their cloud-based analytical environments and integrated development tools lower technical barriers for approved researchers, facilitating rapid cohort identification and model development. However, subscription-based or partnership-driven access models introduce substantial financial and institutional barriers, limiting participation to well-resourced organizations and raising equity concerns. While these platforms provide access to structured data and, in some cases, unstructured clinical notes to support AI-driven CDSS development, access remains constrained by membership requirements, high costs, contractual restrictions, institutional review processes, and governance constraints. Data harmonization across sites also remains a significant challenge, limiting model portability and reproducibility.

Government and National Health System Data Repositories

National and government-sponsored initiatives also offer uniquely rich datasets for model development and clinical research.

Veterans Health Affairs has a robust and integrated data infrastructure that supports clinical care, operational decision-making, and research across the VHA. Central to this infrastructure is the US Department of Veterans Affairs Informatics and Computing Infrastructure (VINCI), a resource of the Veterans Affairs (VA) Office of Research and Development that provides researchers with nationwide access to VA patient data [61]. VINCI partners with the Corporate Data Warehouse (CDW), which aggregates health care–relevant clinical and operational data from across the VHA into a relational database to facilitate enterprise-level reporting, performance measurement, management decision-making, and research. In addition to hosting all CDW data, VINCI provides unique longitudinal views of patient care and supports advanced analytics through the Observational Medical Outcomes Partnership Common Data Model. The Observational Medical Outcomes Partnership Common Data Model is technology neutral and optimized for large-scale computational analysis, enabling evaluation of associations between interventions and outcomes across datasets containing millions of patients and billions of clinical observations. To support data reuse and reproducibility, standardized phenotypes with documented curation methods and decision rules are maintained in the VA's Centralized Interactive Phenomics Resource knowledge base

XSL•FO
RenderX

[62,63]. Figure 2 illustrates the current architecture of VHA data.

Figure 2. Veterans Health Administration (VHA) data warehousing architecture. AAC: Austin Automation Center; ADR: Administrative Data Repository; CMS: Centers for Medicare and Medicaid Services; CSC: Clinical Studies Center; DoD: Department of Defense; DSS: decision support system; HDR: Health Data Repository; HSR&D: Health Services Research and Development; OPP: Office of Planning and Programs; OQP: Office of Quality and Performance; PCS: Office of Patient Care Services; VA: US Department of Veterans Affairs; VISION: Veterans Integrated Service Network; VistA: Veterans Health Information Systems and Technology Architecture.

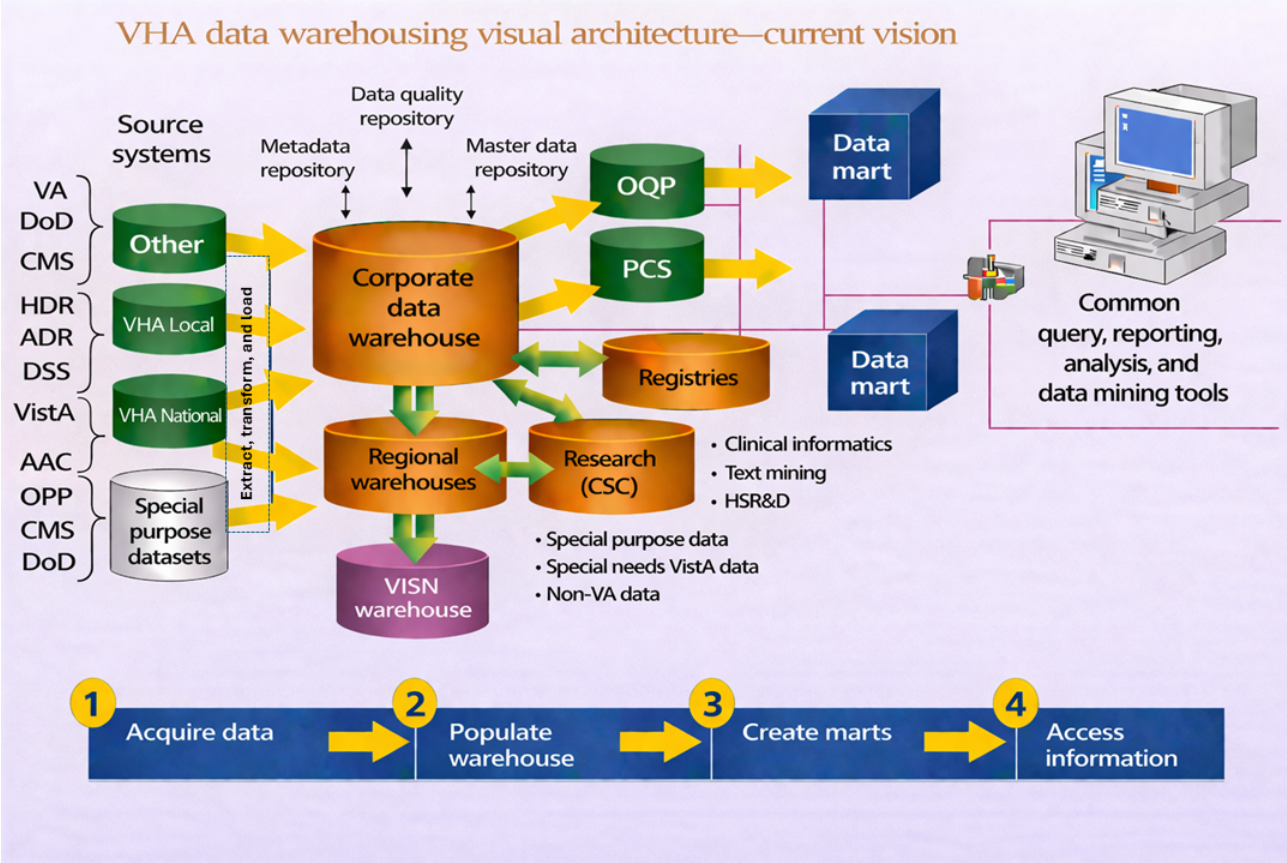

Complementing the VINCI and CDW are specialized VHA registries that enable in-depth study of rare, complex, and veteran-specific conditions. These registries include detailed longitudinal clinical assessments and EHR data for veterans receiving specific treatments or follow-up care, such as the VA Central Cancer Registry, the Health Outcomes Military Exposures registry—which encompasses multiple exposure-related surveillance programs—and the VA Spinal Cord Injuries and Disorders Registry [64-66]. Patients are tracked over time and remain in registries even after death, supporting epidemiologic research and monitoring of disease incidence, treatment patterns, care efficiency, outcomes, and costs. Access to VA data is restricted to VA employees and non-VA researchers who obtain without-compensation employee status and complete a formal approval process, including submission of a data management plan, IRB approval, and research protocols through VINCI.

The strength of VHA data lies in the size of patient cohorts; their depth and granularity; and their longitudinal nature, which allows for tracking changes over time, establishing causality and temporal dynamics, controlling for individual differences, increasing statistical power, understanding developmental processes, and reducing bias. However, although VA data are notably rich and more complete than many civilian datasets, they are limited by data quality, such as inaccuracies, missing data, and inconsistencies, as well as the complexity of integrating data from various sources with differing formats and standards (eg, the US Department of Defense, CMS, individual VHA facilities, and overall national VHA data). Additionally, ensuring data privacy and security, managing the vast volume of data, addressing potential biases and representativeness issues, and requiring advanced technical expertise and computational power can present significant hurdles to effective big data analytics and AI implementation.

The UK Biobank offers one of the most comprehensive population-scale datasets worldwide, integrating EHRs, imaging, genomics, lifestyle data, and long-term follow-up [67]. Launched in 2006, it enrolled over 500,000 participants aged 40 to 69 years across the United Kingdom, collecting extensive health records, genomic data, multimodal imaging (including brain, cardiac, and abdominal magnetic resonance imaging), lifestyle questionnaires, physical measurements, and longitudinal outcomes via linkage to national registries. Beyond its scale, the UK Biobank is distinguished by its transparent open access model designed to maximize global research utility while maintaining participant trust. Approved researchers worldwide can apply through a streamlined online process, with applications reviewed solely based on scientific merit and feasibility rather than institutional affiliation or commercial interests [68]. Researchers execute a material transfer agreement,

XSL•FO
RenderX

commit to peer-reviewed publication, return derived data and algorithms, and submit annual progress reports. Data access is subject only to nominal cost-recovery fees, and research outputs must be publicly available, fostering transparency and reproducibility. The application-to-access timeline typically spans weeks, accelerating the research life cycle. As of 2024, the UK Biobank has supported over 9000 peer-reviewed publications across diverse disease domains, including genome-wide association studies, cardiovascular risk prediction models, and AI-based retinal image analyses for systemic disease detection [68-71]. This model demonstrates that large-scale data sharing, coupled with robust yet nonburdensome governance and privacy protections, can accelerate discovery without compromising ethical stewardship [72].

The National Institutes of Health All of Us research program exemplifies a shift toward inclusive, participant-centered precision medicine in the United States, emphasizing innovative data access infrastructure [73]. Launched in 2018 to enroll 1 million diverse participants, All of Us addresses the underrepresentation of diverse populations in large-scale health studies. The program collects EHRs, genomic data, biospecimens, wearable sensor data, and self-reported health information, creating a rich, longitudinal dataset reflecting US diversity. A key feature is its cloud-based, open science data sharing model through the Researcher Workbench—a secure, scalable platform democratizing data and computational tool access. The Workbench operates on a tiered access system: a public tier with deidentified aggregate data; a registered tier with deidentified, aggregated data for researchers upon completing ethics training; and a controlled tier with individual-level genomic and detailed data for researchers with additional credentials and IRB approval. The platform reduces privacy risks by providing analysis tools without requiring data download, enabling research using preconfigured Jupyter notebooks and R and Python environments [73].

All of Us prioritizes participant engagement and data return, providing participants with access to their genetic results and health summaries, thereby fostering transparency and trust [74]. This bidirectional data flow respects participant autonomy while maximizing public benefit. Recent studies using All of Us data have advanced genetic diversity understanding in pharmacogenomics, identified novel disease associations, and validated AI-driven clinical models across diverse demographics [74,75]. By emphasizing diversity, participant empowerment, accessible infrastructure, and open science, the All of Us research program advances precision medicine and models an ethically grounded, technologically sophisticated approach to big data governance. Initiatives such as this one, alongside the UK Biobank and CMS's tiered model, demonstrate that rigorous, transparent data sharing and collaboration can transform global health care delivery and outcomes. These national and government-sponsored data resources demonstrate the power of centralized, well-governed data infrastructures, but access is often limited by citizenship, institutional affiliation, or strict use agreements, which reduces global accessibility and, ultimately, affects their potential for use in the development and continuous improvement of AI-enabled CDSSs.

Emerging Alternatives and Privacy-Preserving Solutions for Practical and Ethical Challenges

Overview

Using EHR data and AI to predict health outcomes presents practical challenges, particularly given patients' vulnerabilities and the need for equitable access to high-quality care. Medical data are often fragmented across systems, incomplete, or inconsistently documented, limiting the accuracy and generalizability of predictive models. Patients may also have unique or context-specific health profiles, and models trained on historical data may fail to reflect new treatments, evolving standards of care, or shifting disease patterns. Additional limitations include "truth in data" issues—such as the use of diagnostic codes to rule out rather than document conditions—and collider bias, which can distort model outputs [76,77]. Furthermore, the opaque or "black box" nature of many ML models can undermine clinician trust and impede adoption in high-stakes clinical settings where understanding the rationale behind predictions is essential.

Ethical concerns further complicate the use of AI in clinical decision-making, particularly related to bias, fairness, and privacy. Training datasets that underrepresent certain populations, such as women or racial and ethnic minority groups, risk perpetuating inequities in care. Privacy risks are heightened because medical data are highly sensitive and vulnerable to breaches, and overreliance on algorithmic predictions may unintentionally diminish clinician judgment and patient-centered care. Addressing these challenges requires deliberate safeguards, transparency, and accountability, as well as continued involvement of clinicians and patients in system design and use. Balancing the benefits of AI with the protection of patient rights, dignity, and equity is essential to the responsible integration of AI and ML into precision medicine. Given these ethical and legal constraints surrounding patient data, several alternative approaches are gaining traction.

Synthetic Data Generation

Advances in generative AI now enable the creation of high-fidelity synthetic health care datasets that preserve statistical properties and clinical relationships while minimizing reidentification risk. When carefully validated, synthetic data can support algorithm development, benchmarking, and educational use. However, challenges remain in ensuring clinical realism, avoiding hidden biases, and validating downstream model performance.

The main advantage of synthetic data is that they provide a dynamic, data-driven simulation of real-world systems, which can enable prediction, optimization, and decision-making without violating privacy or disrupting patient care. However, because synthesized data rely on collected and modeled data to mirror a real system, they are only as reliable as the quality, level of detail, and timeliness of the source data. Incomplete, biased, or outdated inputs can lead to misleading outputs, reinforcing poor decisions rather than improving them. There is also a risk of overfitting, where the model learns patterns that

XSL•FO
RenderX

fit past data too closely, making it less able to handle new or unexpected situations. Additionally, maintaining a synthetic dataset is resource intensive, requiring constant synchronization, governance, and security oversight—making them vulnerable to drift, privacy risks, and escalating costs. Synthesized patient data also raise privacy concerns, as building detailed digital counterparts often requires access to sensitive or personal data that can be exposed or misused if not properly protected. Finally, synthetic data struggle with generalization: insights derived from one context may not transfer well to different environments, contributing to misleading conclusions when applied outside the conditions under which the synthesized dataset was trained or calibrated. All these concerns must be considered before such data can inform policy or treatment decisions.

Federated Data Networks

Federated learning enables collaborative AI development across health care institutions without centralizing sensitive patient data. Instead of aggregating data in one place, a federated learning algorithm sends the model to each site, where institutions train it locally and share only parameter updates—not raw records that need to be constrained within institutional firewalls. A central server aggregates these updates (eg, via federated averaging) to iteratively refine a global model. This process continues until convergence, producing a model that learns from multi-institutional data while maintaining privacy and data sovereignty. This approach is particularly promising for rare diseases, multicenter studies, and regulatory-compliant AI development, although it introduces technical complexity and coordination overhead.

However, federated learning networks are limited by diverse organizational and technical landscapes, infrastructure variances, and differences in secure processing environments, as well as the extent of data harmonization, fragmentation, heterogeneity across data sites, and generalizability. Addressing these challenges would put tremendous strain on the existing IT infrastructure. Nonetheless, federated data provide a way for researchers to access extensive patient data across various locations to design algorithms with high accuracy, broad applicability, and minimal resource utilization.

Proposed Framework for Synthesis: Balancing Access, Privacy, and Innovation

AI holds immense potential to transform health care by providing data-driven decision support, optimizing workflow, and improving patient outcomes. However, the tension between data accessibility and data protection defines the current landscape and constrains the development of AI-driven clinical health care models. While public datasets and national repositories provide essential foundations, they are insufficient on their own to support the next generation of data-driven CDSSs. Scalable progress requires a hybrid ecosystem that combines curated RWD, synthetic data augmentation, and federated architectures that are supported by clear governance frameworks and incentives for responsible data sharing. Figure 3 illustrates a detailed outline for a framework to overcome data bottlenecks and improve patient and system outcomes.

XSL•FO
RenderX

Figure 3. Proposed framework for overcoming data bottlenecks to improve patient and system outcomes. AI: artificial intelligence; CDSS: clinical decision support system; CMS: Centers for Medicare and Medicaid Services; CT: computed tomography; DL: deep learning; EHR: electronic health record; ML: machine learning; MRI: magnetic resonance imaging; NLP: natural language processing; SEER: Surveillance, Epidemiology, and End Results; UNOS: United Network for Organ Sharing; VHA: Veterans Health Administration; XAI: explainable AI.

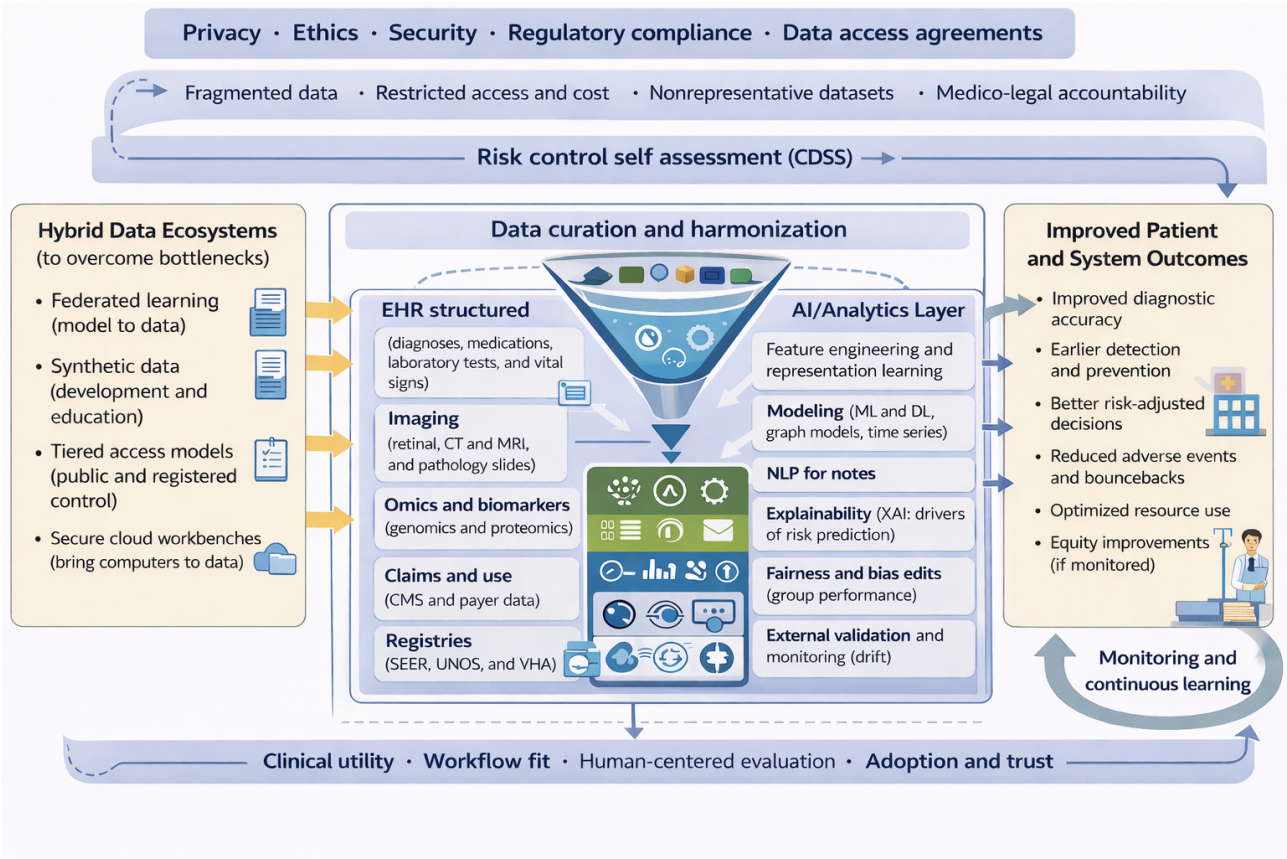

Building hybrid data ecosystems for AI-enabled CDSSs requires a reliable, secure, and well-governed foundation that makes high-quality data easy to discover, use, and trust. Core principles include the following:

- Quality over quantity—establish strong data quality standards, validation, and continuous monitoring. For AI systems, this includes validating ground truth labels and detecting and correcting systematic errors in data collection.
- Interoperability and reusability—use open table formats and common vocabularies so data can move across tools and teams. Adopt standard instruments and frameworks for clinical data with consistent preprocessing and design schemas.
- Privacy and ethics by design—include consent, deidentification, and access controls into pipelines rather than as afterthoughts or preuse requirements. Monitor performance across demographic groups to continuously assess bias and fairness.
- Observability and lineage—trace where data came from, how they were transformed, and how they impact models to increase transparency and explainability, making processes understandable to clinicians and affected individuals.
- Self-service with guardrails—empower users to safely discover, understand, and request data via catalogs and data products. Establish data stewards with clear roles and responsibilities, definitions, quality rules, and catalog curation. Use data protection impact assessments to audit compliance, security, and privacy.

Ultimately, overcoming data availability challenges is not merely a technical problem—it is an organizational, regulatory, and ethical one. Addressing these barriers will be essential to translating AI innovation into sustained improvements in patient outcomes.

Ethical Considerations

This study did not involve human participants or identifiable data, so ethics approval was not required. Studies completed by other researchers and reviewed in this paper were conducted in accordance with the ethical standards of the institutional review boards of affiliated institutions. All data used in these studies were deidentified and compliant with the 1996 HIPAA (Health Insurance Portability and Accountability Act) regulations to ensure patient privacy and confidentiality. The ethical considerations related to secondary use of existing health care data for AI research were rigorously upheld in all studies conducted by the authors of this paper.

XSL•FO
RenderX

Acknowledgments

The generative artificial intelligence tool ChatGPT by OpenAI was used to generate figures, improve linguistic clarity, and decrease the word count of the manuscript. The views expressed in this paper are those of the authors and do not necessarily represent the views of the US Department of Veterans Affairs or other affiliated institutions. The authors would like to acknowledge the contributions of Dr Moncarm Fouche, chief of staff, Eastern Oklahoma VA Health Care System, and Evan Erwin, executive director, Veterans Research and Education Foundation, for their support and assistance with this project.

Funding

Funding for this study was provided by the Veterans Research and Education Foundation.

Authors' Contributions

JED, DD, and ZH contributed to the conceptualization of this viewpoint. JED, DD, and ZH drafted the revised manuscript, and all authors participated in the review and editing of the final version.

Conflicts of Interest

None declared.

References

1. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. JAMA 2018 Dec 4;320(21):2199-2200. [doi: 10.1001/jama.2018.17163] [Medline: 30398550]
2. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Feb 2;542(7639):115-118. [doi: 10.1038/nature21056] [Medline: 28117445]
3. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1:18. [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
4. Eguia H, Sánchez-Bocanegra CL, Vinciarelli F, Alvarez-Lopez F, Saigí-Rubió F. Clinical decision support and natural language processing in medicine: systematic literature review. J Med Internet Res 2024 Sep 30;26:e55315. [doi: 10.2196/55315] [Medline: 39348889]
5. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019 Oct 29;17(1):195. [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
6. Salimparsa M, Sedig K, Lizotte DJ, Abdullah SS, Chalabianloo N, Muanda FT. Explainable AI for clinical decision support systems: literature review, key gaps, and research synthesis. Informatics 2025;12(4):119. [doi: 10.3390/informatics12040119]
7. Tun HM, Rahman HA, Naing L, Malik OA. Trust in artificial intelligence-based clinical decision support systems among health care workers: systematic review. J Med Internet Res 2025 Jul 29;27:e69678. [doi: 10.2196/69678] [Medline: 40772775]
8. Dingel J, Kleine AK, Cecil J, Sigl AL, Lermer E, Gaube S. Predictors of health care practitioners' intention to use ai-enabled clinical decision support systems: meta-analysis based on the unified theory of acceptance and use of technology. J Med Internet Res 2024 Aug 5;26:e57224. [doi: 10.2196/57224] [Medline: 39102675]
9. Giebel GD, Raszke P, Nowak H, et al. Improving AI-based clinical decision support systems and their integration into care from the perspective of experts: interview study among different stakeholders. JMIR Med Inform 2025 Jul 7;13:e69688. [doi: 10.2196/69688] [Medline: 40623684]
10. Gambetti A, Han Q, Shen H, Soares C. A survey on human-centered evaluation of explainable AI methods in clinical decision support systems. arXiv. Preprint posted online on Feb 14, 2025. [doi: 10.48550/arXiv.2502.09849]
11. Bayor AA, Li J, Yang IA, Varnfield M. Designing Clinical Decision Support Systems (CDSS)-a user-centered lens of the design characteristics, challenges, and implications: systematic review. J Med Internet Res 2025 Jun 20;27:e63733. [doi: 10.2196/63733] [Medline: 40540451]
12. Zheng R, Jiang X, Shen L, et al. Investigating clinicians' intentions and influencing factors for using an intelligence-enabled diagnostic clinical decision support system in health care systems: cross-sectional survey. J Med Internet Res 2025 Apr 7;27:e62732. [doi: 10.2196/62732] [Medline: 40194276]
13. Labkoff S, Oladimeji B, Kannry J, et al. Toward a responsible future: recommendations for AI-enabled clinical decision support. J Am Med Inform Assoc 2024 Nov 1;31(11):2730-2739. [doi: 10.1093/jamia/ocae209] [Medline: 39325508]
14. Sáez C, Ferri P, García-Gómez JM. Resilient artificial intelligence in health: synthesis and research agenda toward next-generation trustworthy clinical decision support. J Med Internet Res 2024 Jun 28;26:e50295. [doi: 10.2196/50295] [Medline: 38941134]
15. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
16. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 2005 Jun;34(2):113-127. [doi: 10.1016/j.artmed.2004.07.002] [Medline: 15894176]

XSL•FO
RenderX

17. Dag A, Asilkalkan A, Aydas OT, Caglar M, Simsek S, Delen D. A parsimonious tree augmented naive bayes model for exploring colorectal cancer survival factors and their conditional interrelations. Inf Syst Front 2025 Jun;27:1209-1225. [doi: 10.1007/s10796-024-10517-7]
18. Nath G, Coursey A, Ekong J, et al. Determining the temporal factors of survival associated with brain and nervous system cancer patients: a hybrid machine learning methodology. Int J Healthc Manag 2024;17(3):453-467. [doi: 10.1080/20479700.2023.2196101]
19. Dag AZ, Akcam Z, Kibis E, Simsek S, Delen D. A probabilistic data analytics methodology based on Bayesian belief network for predicting and understanding breast cancer survival. Knowl Based Syst 2022 Feb;242:108407. [doi: 10.1016/j.knosys.2022.108407]
20. Topuz K, Abdulrashid I, Davazdahemami B, Coussement K, Delen D. Developing an interpretable decision support framework for artifact design in organ transplantation. Inf Syst Front 2025 Dec;27:2573-2597. [doi: 10.1007/s10796-025-10652-9]
21. Amini M, Bagheri A, Delen D. An explanatory analytics model for identifying factors indicative of long- versus short-term survival after lung transplantation. Decis Anal J 2022 Jun;3:100058. [doi: 10.1016/j.dajour.2022.100058]
22. Oztekin A, Al-Ebbini L, Sevkli Z, Delen D. A decision analytic approach to predicting quality of life for lung transplant recipients: a hybrid genetic algorithms-based methodology. Eur J Oper Res 2018;266(2):639-651. [doi: 10.1016/j.ejor.2017.09.034]
23. Piri S, Delen D, Liu T, Zolbanin HM. A data analytics approach to building a clinical decision support system for diabetic retinopathy: developing and deploying a model ensemble. Decis Support Syst 2017;101:12-27. [doi: 10.1016/j.dss.2017.05.012]
24. Wang R, Miao Z, Liu T, et al. Derivation and validation of essential predictors and risk index for early detection of diabetic retinopathy using electronic health records. J Clin Med 2021 Apr 2;10(7):1473. [doi: 10.3390/jcm10071473] [Medline: 33918304]
25. Wang R, Miao Z, Liu T, et al. Derivation and validation of essential predictors and risk index for early detection of diabetic retinopathy using electronic health records. J Clin Med 2021 Apr 2;10(7):1473. [doi: 10.3390/jcm10071473]
26. Pease M, Gonzalez-Martinez J, Puccio A, et al. Risk factors and incidence of epilepsy after severe traumatic brain injury. Ann Neurol 2022 Oct;92(4):663-669. [doi: 10.1002/ana.26443] [Medline: 35713346]
27. An S, Malhotra K, Dilley C, et al. Predicting drug-resistant epilepsy - a machine learning approach based on administrative claims data. Epilepsy Behav 2018 Dec;89:118-125. [doi: 10.1016/j.yebeh.2018.10.013] [Medline: 30412924]
28. Lhatoo SD, Bernasconi N, Blumcke I, et al. Big data in epilepsy: clinical and research considerations. Report from the Epilepsy Big Data Task Force of the International League Against Epilepsy. Epilepsia 2020 Sep;61(9):1869-1883. [doi: 10.1111/epi.16633] [Medline: 32767763]
29. Fernandes M, Cardall A, Jing J, et al. Identification of patients with epilepsy using automated electronic health records phenotyping. Epilepsia 2023 Jun;64(6):1472-1481. [doi: 10.1111/epi.17589] [Medline: 36934317]
30. Ahmed N. Graph representation of patient's data in EHR for outcome prediction [Master's thesis]. : Otto-von-Guericke University Magdeburg; 2023 URL: https://wwwiti.cs.uni-magdeburg.de/iti_db/publikationen/ps/auto/MastersThesis:ahmed:2023.pdf [accessed 2026-03-24]
31. Li J, Wang M, Steinbach MS, Kumar V, Simon GJ. Don't do imputation: dealing with informative missing values in EHR data analysis. In: 2018 IEEE International Conference on Big Knowledge (ICBK): IEEE Computer Society; 2018:415-422. [doi: 10.1109/ICBK.2018.00062]
32. Haneuse S, Arterburn D, Daniels MJ. Assessing missing data assumptions in EHR-based studies: a complex and underappreciated task. JAMA Netw Open 2021 Feb 1;4(2):e210184. [doi: 10.1001/jamanetworkopen.2021.0184] [Medline: 33635321]
33. Ramamurthy P, Han Z, Delen D, et al. Graph attention network with comorbidity connectivity embedding for post-traumatic epilepsy risk prediction using sparse time-series electronic health records. Int J Med Inform 2026 Mar 15;208:106239. [doi: 10.1016/j.ijmedinf.2025.106239] [Medline: 41478164]
34. Kato C, Uemura O, Sato Y, Tsuji T. Functional outcome prediction after spinal cord injury using ensemble machine learning. Arch Phys Med Rehabil 2024 Jan;105(1):95-100. [doi: 10.1016/j.apmr.2023.08.011] [Medline: 37714506]
35. Yoo HJ, Koo B, Yong CW, Lee KS. Prediction of gait recovery using machine learning algorithms in patients with spinal cord injury. Medicine (Baltimore) 2024 Jun 7;103(23):e38286. [doi: 10.1097/MD.0000000000038286] [Medline: 38847729]
36. Maki S, Furuya T, Inoue T, et al. Machine learning web application for predicting functional outcomes in patients with traumatic spinal cord injury following inpatient rehabilitation. J Neurotrauma 2024 May;41(9-10):1089-1100. [doi: 10.1089/neu.2022.0383] [Medline: 37917112]
37. Håkansson S, Tuci M, Bolliger M, Curt A, Jutzeler CR, Brüningk SC. Data-driven prediction of spinal cord injury recovery: an exploration of current status and future perspectives. Exp Neurol 2024 Oct;380:114913. [doi: 10.1016/j.expneurol.2024.114913] [Medline: 39097073]
38. Tao G, Yang S, Xu J, Wang L, Yang B. Global research trends and hotspots of artificial intelligence research in spinal cord neural injury and restoration-a bibliometrics and visualization analysis. Front Neurol 2024;15:1361235. [doi: 10.3389/fneur.2024.1361235] [Medline: 38628700]

XSL•FO
RenderX

39. Khan O, Badhiwala JH, Grasso G, Fehlings MG. Use of machine learning and artificial intelligence to drive personalized medicine approaches for spine care. World Neurosurg 2020 Aug;140:512-518. [doi: 10.1016/j.wneu.2020.04.022] [Medline: 32797983]
40. Ghahramani Z. Probabilistic machine learning and artificial intelligence. Nature 2015 May 28;521(7553):452-459. [doi: 10.1038/nature14541] [Medline: 26017444]
41. Ajiboye AB, Willett FR, Young DR, et al. Restoration of reaching and grasping movements through brain-controlled muscle stimulation in a person with tetraplegia: a proof-of-concept demonstration. Lancet 2017 May 6;389(10081):1821-1830. [doi: 10.1016/S0140-6736(17)30601-3] [Medline: 28363483]
42. Kosuri S, Borca CH, Mugnier H, et al. Machine-assisted discovery of chondroitinase ABC complexes toward sustained neural regeneration. Adv Healthc Mater 2022 May;11(10):e2102101. [doi: 10.1002/adhm.202102101] [Medline: 35112508]
43. Luther SL, Thomason SS, Sabharwal S, et al. Machine learning to develop a predictive model of pressure injury in persons with spinal cord injury. Spinal Cord 2023 Sep;61(9):513-520. [doi: 10.1038/s41393-023-00924-z] [Medline: 37598263]
44. Spungen AM, Dematt EJ, Biswas K, et al. Exoskeletal-assisted walking in veterans with paralysis: a randomized clinical trial. JAMA Netw Open 2024 Sep 3;7(9):e2431501. [doi: 10.1001/jamanetworkopen.2024.31501] [Medline: 39230903]
45. Bogie KM, Roggenkamp SK, Zeng N, et al. Development of predictive informatics tool using electronic health records to inform personalized evidence-based pressure injury management for veterans with spinal cord injury. Mil Med 2021 Jan 25;186(Suppl 1):651-658. [doi: 10.1093/milmed/usaa469] [Medline: 33499541]
46. Burkhart L, Siddiqui S, Smith A, et al. Development of a decision support tool to prevent community acquired pressure injuries for use in the spinal cord injury clinic using the delphi technique. Spinal Cord 2023 Dec;61(12):667-683. [doi: 10.1038/s41393-023-00937-8] [Medline: 37828368]
47. Habibi MA, Naseri Alavi SA, Soltani Farsani A, Mousavi Nasab MM, Tajabadi Z, Kobets AJ. Predicting the outcome and survival of patients with spinal cord injury using machine learning algorithms: a systematic review. World Neurosurg 2024 Aug;188:150-160. [doi: 10.1016/j.wneu.2024.05.103] [Medline: 38796146]
48. Kapoor D, Xu C. Spinal cord injury AIS predictions using machine learning. eNeuro 2023 Jan;10(1):ENEURO.0149-22.2022. [doi: 10.1523/ENEURO.0149-22.2022] [Medline: 36543536]
49. Reddy BK, Delen D. Predicting hospital readmission for lupus patients: an RNN-LSTM-based deep-learning methodology. Comput Biol Med 2018 Oct 1;101:199-209. [doi: 10.1016/j.compbiomed.2018.08.029] [Medline: 30195164]
50. Reddy BK, Delen D, Agrawal RK. Predicting and explaining inflammation in Crohn's disease patients using predictive analytics methods and electronic medical record data. Health Informatics J 2019 Dec;25(4):1201-1218. [doi: 10.1177/1460458217751015] [Medline: 29320910]
51. Davazdahemami B, Peng P, Delen D. A deep learning approach for predicting early bounce-backs to the emergency departments. Healthc Anal 2022 Nov;2:100018. [doi: 10.1016/j.health.2022.100018]
52. Peng P, Davazdahemami B, Delen D, Shapiro J, Manini AF. 288 Predicting potential bouncebacks to the emergency department: a machine learning approach. Ann Emerg Med 2019 Oct;74(4):S114. [doi: 10.1016/j.annemergmed.2019.08.246]
53. Park HS, Lloyd S, Decker RH, Wilson LD, Yu JB. Overview of the Surveillance, Epidemiology, and End Results database: evolution, data variables, and quality assurance. Curr Probl Cancer 2012;36(4):183-190. [doi: 10.1016/j.currproblcancer.2012.03.007] [Medline: 22481006]
54. Health insurance exchange public use files (exchange pufs). URL: https://www.cms.gov/marketplace/resources/data/public-use-files [accessed 2026-04-02]
55. Ayanian JZ, Buntin MB. Pausing the proposed CMS data access policy: first, do no harm. JAMA Health Forum 2024 May 3;5(5):e241328. [doi: 10.1001/jamahealthforum.2024.1328] [Medline: 38819799]
56. CMS Limited data set (LDS) files. URL: https://www.cms.gov/data-research/cms-data/limited-data-set-lds-files [accessed 2026-04-02]
57. Ehwerhemuepha L, Carlson K, Moog R, et al. Cerner Real-World Data (CRWD) - a de-identified multicenter electronic health records database. Data Brief 2022 Mar;42:108120. [doi: 10.1016/j.dib.2022.108120] [Medline: 35434225]
58. Epic cosmos. URL: https://cosmos.epic.com/ [accessed 2026-04-02]
59. Son M, Gallagher K, Lo JY, et al. Coronavirus disease 2019 (COVID-19) pandemic and pregnancy outcomes in a U.S. population. Obstet Gynecol 2021;138(4):542-551. [doi: 10.1097/AOG.0000000000004547] [Medline: 34433180]
60. Zindani S, Rashid Z, Woldesenbet S, Khalil M, Pawlik TM. Investigating racial/ethnic disparities in opioid use among patients undergoing gastrointestinal cancer surgery. Am J Surg 2026 Mar;253:116799. [doi: 10.1016/j.amjsurg.2025.116799] [Medline: 41500095]
61. VINCI data services: an overview. U.S. Department of Veterans Affairs. URL: https://www.hsrd.research.va.gov/for_researchers/cyber_seminars/archives/video_archive.cfm?SessionID=7769 [accessed 2026-04-02]
62. Honerlaw J, Ho YL, Fontin F, et al. Framework of the Centralized Interactive Phenomics Resource (CIPHER) standard for electronic health data-based phenomics knowledgebase. J Am Med Inform Assoc 2023 Apr 19;30(5):958-964. [doi: 10.1093/jamia/ocad030] [Medline: 36882092]
63. Honerlaw J, Ho YL, Fontin F, et al. Centralized Interactive Phenomics Resource: an integrated online phenomics knowledgebase for health data users. J Am Med Inform Assoc 2024 Apr 19;31(5):1126-1134. [doi: 10.1093/jamia/ocae042] [Medline: 38481028]

XSL•FO
RenderX

64. Sippel JL, Daly JE, Poggensee L, et al. Modernization of a large spinal cord injuries and disorders registry: the veterans administration experience. Arch Rehabil Res Clin Transl 2022 Oct;4(4):100237. [doi: 10.1016/j.arrct.2022.100237] [Medline: 36545529]
65. Sippel JL, Willenberg R, Evans CT, et al. Evolution of the Veterans Health Administration Spinal Cord Injuries and Disorders (SCI/D) Registry (VHA SCIDR): characterization from 1994 to 2022. J Spinal Cord Med 2026 Mar;49(2):281-291. [doi: 10.1080/10790268.2024.2434305] [Medline: 39661743]
66. Health Outcomes Military Exposures (HOME). U.S. Department of Veterans Affairs. URL: https://www.publichealth.va.gov/PUBLICHEALTH/index.asp [accessed 2026-03-25]
67. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 2015 Mar;12(3):e1001779. [doi: 10.1371/journal.pmed.1001779] [Medline: 25826379]
68. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 2018 Oct;562(7726):203-209. [doi: 10.1038/s41586-018-0579-z] [Medline: 30305743]
69. Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK Biobank. Nat Genet 2018 Nov;50(11):1593-1599. [doi: 10.1038/s41588-018-0248-z] [Medline: 30349118]
70. Littlejohns TJ, Sudlow C, Allen NE, Collins R. UK Biobank: opportunities for cardiovascular research. Eur Heart J 2019 Apr 7;40(14):1158-1166. [doi: 10.1093/eurheartj/ehx254] [Medline: 28531320]
71. Raisi-Estabragh Z, Petersen SE. Cardiovascular research highlights from the UK Biobank: opportunities and challenges. Cardiovasc Res 2020 Jan 1;116(1):e12-e15. [doi: 10.1093/cvr/cvz294] [Medline: 31778147]
72. Tseng RM, Rim TH, Shantsila E, et al. Validation of a deep-learning-based retinal biomarker (Reti-CVD) in the prediction of cardiovascular disease: data from UK Biobank. BMC Med 2023 Jan 24;21(1):28. [doi: 10.1186/s12916-022-02684-8] [Medline: 36691041]
73. All of Us Research Hub with Data Tiers. All of Us Research Hub. URL: https://www.researchallofus.org/data-tools/data-access/ [accessed 2026-04-13]
74. Rim TH, Lee CJ, Tham YC, et al. Deep-learning-based cardiovascular risk stratification using coronary artery calcium scores predicted from retinal photographs. Lancet Digit Health 2021 May;3(5):e306-e316. [doi: 10.1016/S2589-7500(21)00043-1] [Medline: 33890578]
75. All of Us Research Program Investigators, Denny JC, Rutter JL, et al. The “All of Us” research program. N Engl J Med 2019 Aug 15;381(7):668-676. [doi: 10.1056/NEJMsr1809937] [Medline: 31412182]
76. Ramirez AH, Sulieman L, Schlueter DJ, et al. The All of Us research program: data quality, utility, and diversity. Patterns (N Y) 2022 Aug 12;3(8):100570. [doi: 10.1016/j.patter.2022.100570] [Medline: 36033590]
77. Olaker VR, Fry S, Terebuh P, et al. With big data comes big responsibility: Strategies for utilizing aggregated, standardized, de-identified electronic health record data for research. Clin Transl Sci 2025 Jan;18(1):e70093. [doi: 10.1111/cts.70093] [Medline: 39740190]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
CDW: Corporate Data Warehouse
CMS: Centers for Medicare and Medicaid Services
DR: diabetic retinopathy
ED: emergency department
EHR: electronic health record
HIPAA: Health Insurance Portability and Accountability Act
IRB: institutional review board
ML: machine learning
ORWD: Oracle Real-World Data
PTE: posttraumatic epilepsy
RWD: real-world data
SCI: spinal cord injury or disorder
TBI : traumatic brain injury
VA: United States Department of Veterans Affairs
VHA: Veterans Health Administration
VINCI : United States Department of Veterans Affairs Informatics and Computing Infrastructure

XSL•FO
RenderX

Edited by J Sarvestan; submitted 20.Jan.2025; peer-reviewed by A Pour-Rashidi, F Skelton, J Patrianakos; revised version received 05.Feb.2026; accepted 05.Feb.2026; published 04.May.2026.

Please cite as:
Daly JE, Delen D, Han Z, Smith R, Honerlaw J, Cho K, Bennett B, Sippel J
AI in Clinical Decision Support Systems: Promising Applications and Strategies for Managing Data Challenges
J Med Internet Res 2026;28:e71532
URL: https://www.jmir.org/2026/1/e71532
doi:10.2196/71532

© Jennifer E Daly, Dursun Delen, Zheng Han, River Smith, Jacqueline Honerlaw, Kelly Cho, Bridget Bennett, Jennifer Sippel. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 4.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

From Pilot Trap to Institutional Capacity: A Governance Framework for Sustainable Clinical AI Implementation in Health Systems

Jin Tian[1], MME, MHS; Zengren Zhao[1], MD; Longmei Tang[2,3], PhD; Yongzhao Song[4], MSc; Yuchang Li[1], MPH; Nan Jiang[2], DPH

[1]Hospital Management Innovation Center, The First Hospital of Hebei Medical University, 89 Donggang Street, Shijiazhuang, Hebei, China
[2]Department of Social Medicine and Health Services Management, Hebei Medical University, Shijiazhuang, Hebei, China
[3]Hebei Province Key Laboratory of Environmental and Human Health, Shijiazhuang, Hebei, China
[4]Department of Public Health, The First Hospital of Hebei Medical University, Shijiazhuang, Hebei, China

Corresponding Author:
Jin Tian, MME, MHS
Hospital Management Innovation Center, The First Hospital of Hebei Medical University, 89 Donggang Street, Shijiazhuang, Hebei, China

Abstract

Clinical artificial intelligence (AI) applications frequently fail to transition from short-term pilot projects into sustained components of routine clinical care, a phenomenon referred to in this viewpoint as the pilot trap. This persistent gap reflects not only technical or regulatory limitations but also insufficient governance capacity within health care organizations. We argue that such capacity is not fully established before deployment; rather, it develops through implementation as real-world operational tensions clarify organizational ownership, accountability boundaries, and coordination mechanisms. Drawing on an 18-month implementation of a provincial clinical AI platform in China, we develop a 6-module governance framework encompassing institutional carrier formation, infrastructure governance, regulatory and ethical governance, interdisciplinary coordination, translational scaling, and lifecycle evaluation and oversight. These modules represent functional governance conditions observed during implementation rather than a prescriptive institutional architecture to be installed prior to deployment. We further introduce the concept of functional transferability and position the framework as an upstream complement to existing international governance standards, which typically specify what governance should achieve, but often assume that the organizational capacity to implement it already exists. Advancing clinical AI beyond demonstration, therefore, depends less on model performance alone than on the ability of health systems to develop and sustain the institutional capacity required for routine clinical use.

(J Med Internet Res 2026;28:e92680) doi:10.2196/92680

KEYWORDS

clinical artificial intelligence; AI governance; health systems; institutional infrastructure; implementation science; governance framework

Introduction

Substantial investment in clinical artificial intelligence (AI) has generated a rapidly expanding portfolio of tools proposed for health care delivery, including applications for diagnostic support, clinical decision-making, and operational coordination [1,2]. Despite these advances, most clinical AI systems remain confined to short-lived pilots or proof-of-concept demonstrations rather than becoming durable components of routine care across health systems [3-5]. This persistent gap between technological progress and sustained clinical integration suggests that the central constraint lies not only in model performance but also in the institutional arrangements required to authorize, operate, and sustain AI as accountable clinical capacity. Understanding these institutional conditions therefore becomes central to explaining why technically promising AI systems fail to persist in routine clinical practice.

This pattern is frequently described as a “pilot trap,” in which technically promising systems fail to transition into stable operational use [4-7]. Under such conditions, ownership structures remain fragmented, accountability boundaries are unclear, and deployment depends on temporary resources or individual champions rather than durable institutional arrangements. What appears as a challenge of technological adoption therefore reflects a deeper problem of institutional capacity formation, not merely the absence of governance structures, but the absence of the implementation conditions through which such structures are generated and stabilized within routine clinical practice.

China provides an analytically informative setting for examining these dynamics. Clinical AI has been incorporated into national strategies for health care modernization and embedded within evolving regulatory frameworks governing software as a medical device [8,9]. Within this policy environment, provincial

XSL•FO
RenderX

platforms and tertiary hospitals, where administrative legitimacy, clinical scale, and technical capacity converge, represent a structurally significant configuration for AI integration. These conditions make governance formation observable as a determinant of whether AI transitions from episodic deployment to institutional infrastructure. Yet, despite this policy alignment, the conditions enabling governance formation, including stable organizational ownership, standardized infrastructure, and sustained operational accountability, remain inconsistently realized.

Existing scholarship on clinical AI has largely concentrated on model validation, predictive performance, and application-level effectiveness [10-12]. Governance has increasingly been recognized as relevant to implementation outcomes, yet available frameworks remain limited in operational specificity [13-16]. High-level governance principles often lack clearly defined organizational carriers through which governance becomes actionable [17], while implementation-oriented approaches tend to emphasize evaluation and oversight without addressing upstream conditions of ownership, financing, and sustained operational responsibility [6,18]. Critical questions therefore remain underspecified, including who governs clinical AI, under what institutional mandate, and how accountability is distributed across the system lifecycle. Without addressing these questions, the institutional conditions required for sustained clinical AI integration remain poorly understood.

In this viewpoint, we draw on an 18-month real-world implementation of a provincial clinical AI platform in China to develop a governance framework for the institutionalization of clinical AI. By analyzing contemporaneous implementation materials, including governance documents, regulatory filings, and operational records, we identify 6 interdependent governance modules that repeatedly emerged as necessary conditions for sustained deployment. These modules describe the organizational and governance capacities required for clinical AI to transition from episodic pilot projects to routinized clinical infrastructure. We further introduce the concept of functional transferability to distinguish governance functions that remain stable across contexts from the institutional forms through which they may be realized. The framework is intended both to explain why many clinical AI systems remain pilot-bound and to help identify governance conditions that support durable AI integration across health systems.

Conceptual Framing

The conceptual framing of this study is grounded in 3 interrelated constructs: institutionalization, infrastructure, and governance. Institutionalization refers to the stabilization of AI-enabled practices as routinized clinical processes that persist beyond pilot deployments and project-based implementation. Infrastructure denotes the technical and operational substrate required for reliable and repeatable system use, encompassing data pipelines, interoperability systems, computational capacity, and access control. Governance refers to the allocation of decision rights and accountability arrangements that authorize system use, assign responsibility, and oversee system updates across the deployment lifecycle [19,20].

These constructs are analytically distinct yet operationally sequential. Infrastructure constitutes the technical substrate that enables system operation. Governance organizes the human and institutional arrangements through which that substrate is authorized, monitored, and maintained. Institutionalization denotes the observable outcome that emerges when governed systems persist in routine use over time. Governance and institutionalization are related but not identical. Governance refers to the arrangements through which accountability and oversight are structured, whereas institutionalization denotes the condition that becomes observable when such arrangements stabilize sufficiently to sustain routine practice beyond project-specific resources or individual champions. Governance may exist without institutionalization having yet been achieved, whereas institutionalization presupposes the presence of governance arrangements capable of sustaining routine practice.

In this study, these constructs are not treated as fully established preconditions prior to implementation. Governance capacity in particular is conceptualized as emergent rather than antecedent. It develops through the operational demands generated during implementation rather than being fully specified in advance. Implementation processes therefore play a constitutive role in shaping governance arrangements, through which accountability structures, decision rights, and operational responsibilities gradually become stabilized within clinical environments [21]. Institutionalization emerges when these governance arrangements persist and sustain routine clinical practice over time. The sequential relationship among these 3 constructs is illustrated in Figure 1.

XSL•FO
RenderX

Figure 1. Infrastructure provides the technical substrate that enables system operation, including data, computing, and interoperability capacities. Governance organizes coordination and accountability across the lifecycle of deployment by structuring authorization, oversight, and operational responsibility. Institutionalization represents the sustained integration of artificial intelligence–enabled practices into routine clinical workflows as governed systems persist over time. The figure illustrates the sequential relationship through which infrastructure enables governance and governance supports institutionalization.

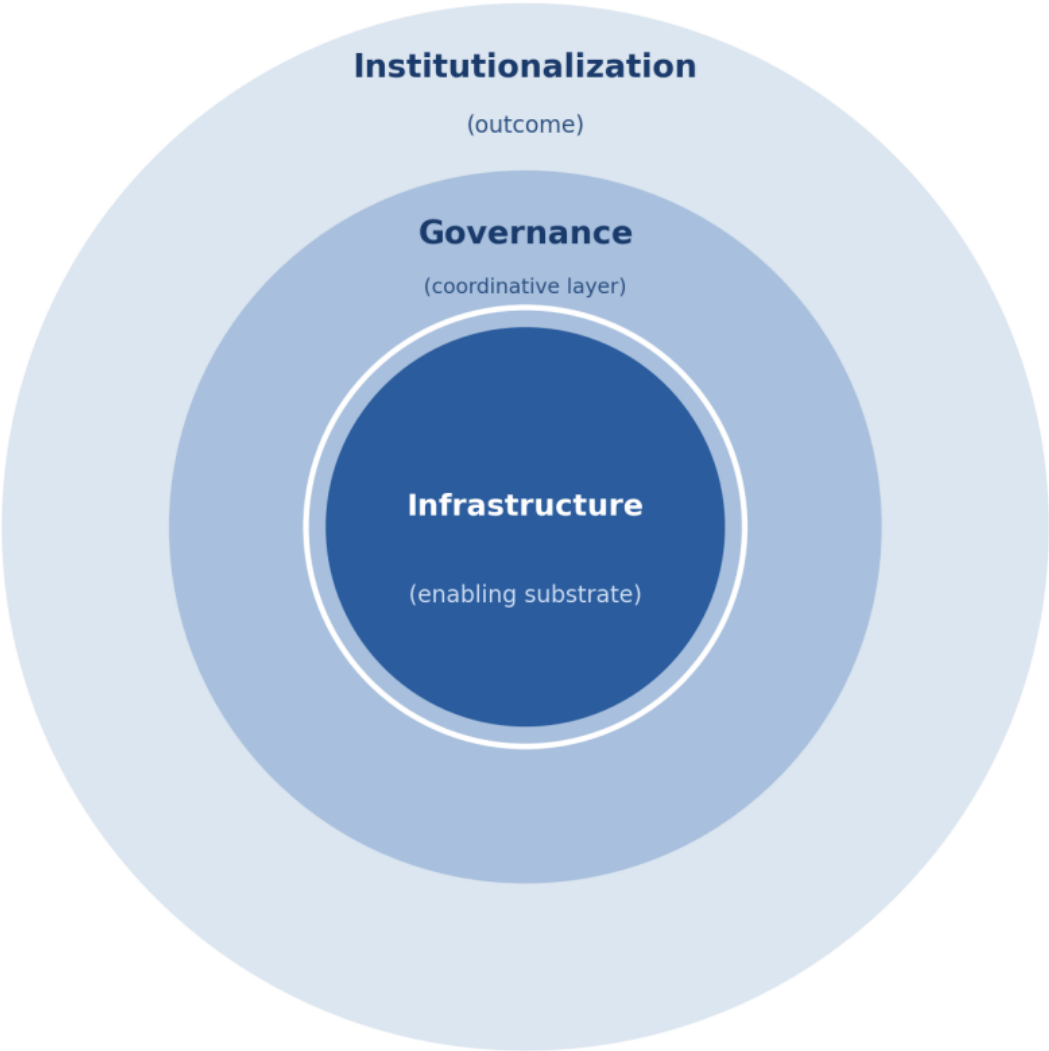

Case Context

To examine how the conceptual relationships outlined above operate in practice, this study analyzes the implementation of a clinical AI platform at a large tertiary academic hospital in Hebei Province, China. Over an 18-month period, the institution deployed a set of AI-enabled systems aimed at supporting both clinical decision-making and operational coordination across departments and affiliated institutions. Implementation was organized around 3 clinical pathways, encompassing intelligent preconsultation triage, oncology multidisciplinary team (MDT) decision support, and therapeutic drug monitoring (TDM). Each pathway represents a domain in which AI systems intersect directly with routine clinical workflows and institutional decision structures and therefore provides an observable setting for examining governance formation during deployment.

The hospital functions as a regional medical center and maintains substantial digital infrastructure, including integrated clinical information systems and data management capabilities. As a large tertiary hospital operating within China's national regulatory framework for AI-enabled medical technologies, it represents a setting where clinical scale, digital infrastructure, and regulatory oversight converge. This institutional configuration creates an analytically valuable context for observing how governance arrangements evolve as AI systems transition from experimental deployment to sustained integration within routine clinical practice.

XSL•FO
RenderX

From Implementation Materials to Governance Framework: Analytical Approach

This study adopts a qualitative, case-based process-tracing approach to examine how governance structures emerged during an 18-month clinical AI platform implementation at a large tertiary academic hospital in Hebei Province, China. The analytical strategy follows qualitative case study logic commonly used to examine institutional processes, organizational coordination, and decision trajectories within complex health system settings.

Empirical materials were derived from institutional artifacts generated through routine platform development and deployment. These materials comprised 17 institutional artifacts, including 13 text-based documents and 4 visual implementation materials, encompassing governance frameworks, regulatory filings, infrastructure certification records, internal operational monitoring records, operational deployment artifacts, monthly work reports, meeting materials, organizational documents, budget requests, workflow diagrams, and platform architecture and feature maps. These materials documented platform development milestones, coordination events, governance decisions, and operational activity across the review period from July 2, 2024, to January 24, 2026. Internal operational monitoring records additionally provided aggregated operational indicators documenting system use during platform deployment. These indicators were used descriptively to characterize continuity of deployment, scope of workflow embedding, and organizational consolidation rather than as formal effectiveness endpoints.

Materials were organized chronologically to identify key decision points and coordination tensions, defined as documented conflicts, delays, or breakdowns requiring institutional responses. Governance modules were derived inductively from recurrent coordination tensions and the institutional mechanisms developed to resolve them. Data analysis followed an inductive thematic analysis approach in which coordination tensions identified across implementation artifacts were iteratively grouped into higher-level governance functions that subsequently informed the 6-module framework. A tension was considered governance-relevant when it recurred across multiple implementation episodes and was corroborated by more than one category of institutional artifact. Initial coding and module development were conducted by two authors and subsequently reviewed by team members with distinct clinical, administrative, and policy perspectives to strengthen interpretive robustness and reduce the risk of post hoc rationalization. Members of the research team participated in documenting institutional processes during the implementation period, enabling access to contemporaneous institutional materials while maintaining analytical separation from AI model development, technical validation, and system performance evaluation activities. Coding was conducted manually by the authors without the use of qualitative data analysis software.

This retrospective, case-based study received formal approval from the Clinical Research Ethics Committee of The First Hospital of Hebei Medical University (approval number: 2026; research review number: 032; approval date: February 5, 2026). The analysis reported in this manuscript was based on institutional documents and aggregated operational materials and did not involve the use of identifiable individual patient-level data.

Governance Framework for Clinical AI Institutionalization

Development of the Governance Framework

Analysis of implementation materials, conducted through the process-tracing approach described above, identified recurrent operational tensions arising during the deployment of clinical AI systems. These tensions were primarily associated with infrastructure integration, regulatory alignment, and the embedding of AI outputs within routine clinical workflows. Recurrent solutions to these implementation tensions were grouped into 6 interdependent modules. These modules represent governance conditions observed in practice rather than a prescriptive institutional design.

The 6 modules collectively constitute a governance cycle through which clinical AI systems transition from experimental deployment to stabilized institutional use. The first two modules, institutional carrier formation and infrastructure governance, represent upstream enabling conditions that establish organizational ownership and sustain technical continuity prior to clinical activation. Regulatory and ethical governance defines authorization boundaries and distributes accountability across the deployment lifecycle. The remaining modules, interdisciplinary operational coordination, translational scaling, and lifecycle evaluation and oversight, support active deployment by enabling cross-departmental embedding, contextual expansion, and adaptive governance as clinical and regulatory conditions evolve. Although analytically presented as 6 modules, the framework operates as a recursive governance cycle in which implementation generates feedback that informs subsequent governance adjustments and institutional learning. The governance framework is illustrated in Figure 2.

XSL•FO
RenderX

Figure 2. Six governance modules (M1-M6) represent organizational, technical, regulatory, and clinical functions required for sustained AI deployment. These modules collectively constitute governance capacity that enables the transition of clinical AI systems from pilot initiatives to institutionalized clinical practice. The framework operates as an adaptive learning cycle in which lifecycle oversight informs continuous governance formation. AI: artificial intelligence.

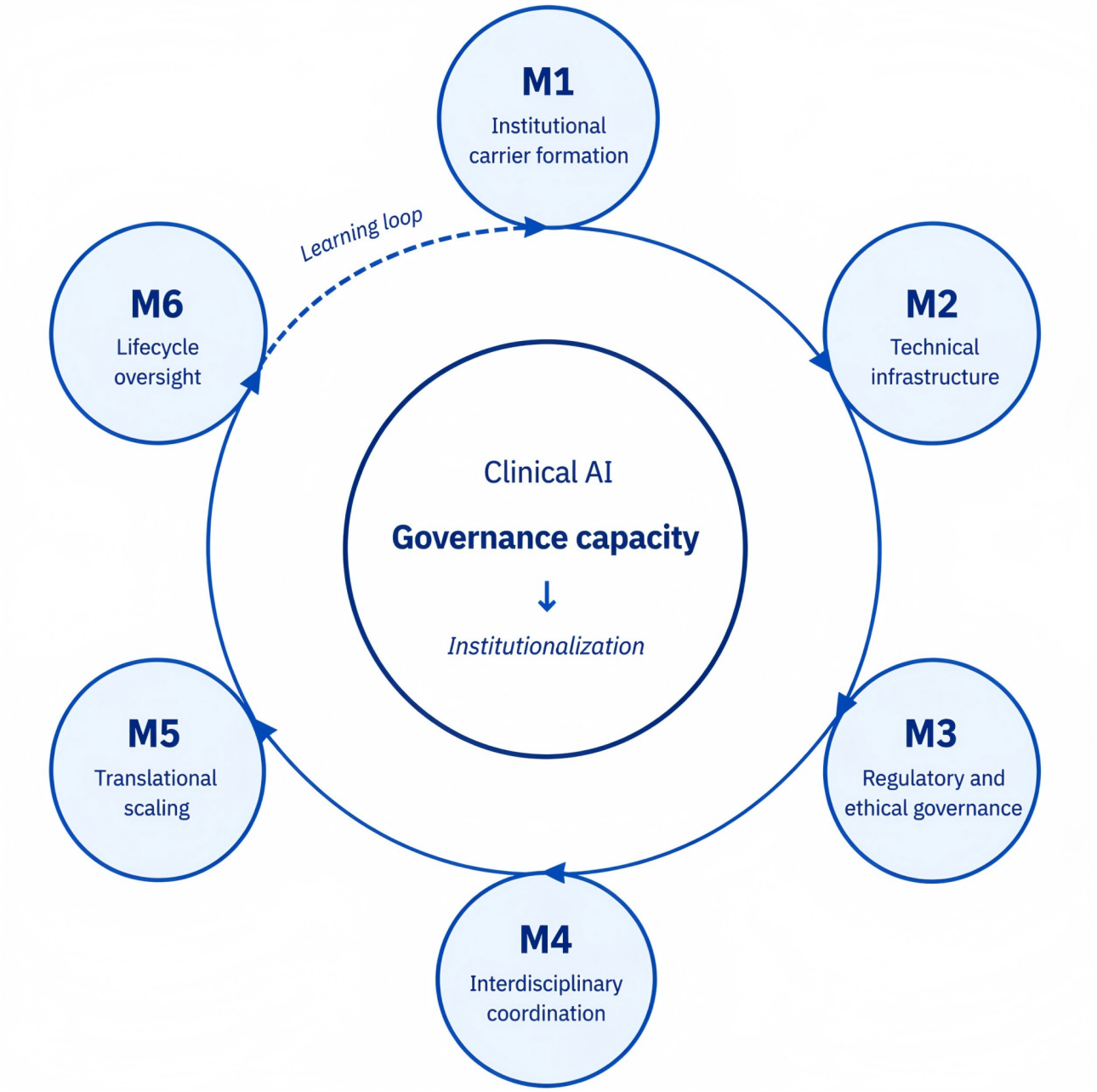

Overview of the Governance Framework

The framework comprises 6 interdependent modules representing the organizational, technical, regulatory, and clinical functions required to stabilize AI systems as routinized components of clinical practice. Each module reflects a recurring governance tension identified during implementation and is supported by empirical process evidence drawn from the Hebei provincial clinical AI platform. Table 1 summarizes the modules, their corresponding implementation artifacts, and the transferable governance requirements derived from them.

XSL•FO
RenderX

Table . Governance framework modules, empirical implementation artifacts, and transferable governance requirements.

Governance module	Empirical evidence from implementation	Transferable governance requirement
M1: Institutional carrier formation	Establishment of a dedicated medical AI[a] laboratory through formal administrative orders; defined organizational structure integrating algorithm development, clinical research, and application translation; designation as a provincial engineering research center.	Clinical AI deployment requires a designated institutional carrier with formal mandate, decision authority, and organizational independence from temporary project teams.
M2: Infrastructure governance	Integration of high-performance computing infrastructure with hospital information systems; EMR[b] functional maturity certification; interoperability standardization assessments; documented institutional data governance policies.	Data pipelines, computational capacity, and interoperability mechanisms must be governed as durable institutional infrastructure with defined access, security, and update controls.
M3: Regulatory and ethical governance	Algorithm filing under the national regulatory framework administered by the Cyberspace Administration of China; documentation defining the system's purpose, use boundaries, and institutional accountability.	Regulatory authorization and accountability structures should be formally specified prior to clinical activation of AI systems.
M4: Interdisciplinary operational coordination	Three-division laboratory governance structure; integration of AI outputs into multidisciplinary tumor board workflows; documented role allocation and decision rights across clinical and technical teams; temporary suspension of a pathway pending clarification of update authority.	Operational deployment requires clearly defined roles, decision rights, and accountability structures linking algorithm development, clinical decision processes, and institutional oversight.
M5: Translational scaling	Expansion of the AI-supported preconsultation pathway across multiple outpatient departments; cross-department workflow integration protocols; establishment of a provincial clinical AI research institute supporting broader deployment.	Scaling requires governance pathways that translate localized operational success into institution-wide organizational capability.
M6: Lifecycle evaluation and oversight	Documentation of system use boundaries; role-based pathway supervision; model version update logs; institutional review of operational feedback during implementation.	Lifecycle governance requires monitoring mechanisms tied to institutional authority to manage system updates, operational feedback, and evolving clinical conditions.

[a]AI: artificial intelligence.

[b]EMR: electronic medical record.

Across the observation period, the platform was progressively integrated into routine clinical workflows spanning multiple service lines. Risk assessment tools were incorporated into outpatient triage processes to support early identification of high-risk patients, while decision-support functions were embedded within multidisciplinary case discussions to facilitate structured cross-departmental evaluation. AI-assisted documentation tools, including generative medical record support and natural-language clinical assistants, were concurrently introduced to support routine documentation and information retrieval. The platform also enabled structured capture of research indicators and integration of clinical data management with institutional quality-monitoring functions. To provide high-level proof-of-concept evidence that the platform progressed beyond bounded pilot deployment, selected operational indicators are summarized in Table 2.

XSL•FO
RenderX

Table . Reported operational indicators and measurement definitions[a].

Indicator	Definition and measurement basis	Observed value
Cumulative patient interactions (preconsultation pathway)	Total number of completed patient-initiated interactions recorded in system logs across all 5 enrolled outpatient departments from the first departmental onboarding to the end of the observation period	>24,000
Recent average daily use	Mean number of completed interactions per calendar day calculated over the postrollout operational phase, following phased onboarding of all 5 departments	~110 encounters/day
Continuous operational duration (preconsultation pathway)	Number of consecutive months the pathway remained operationally active following departmental rollout	>12 months
Number of clinical pathways in sustained deployment	Count of artificial intelligence–enabled clinical pathways that remained in active institutional use across the focal observation period	3

[a]The operational indicators reported in this table were derived from aggregated institutional operational monitoring records generated during routine platform deployment. They are presented to improve interpretive transparency rather than as indicators of clinical effectiveness. The recent daily use figure refers to the mature postrollout phase following phased departmental onboarding and should not be interpreted as a rate that applied uniformly across the entire cumulative review period. No identifiable individual patient-level data were used.

Empirical Context of Platform Operation

During the observation period, the clinical AI platform was progressively integrated into routine clinical workflows across multiple service lines. Operational deployment was organized around 3 clinical pathways, each representing a domain in which AI-assisted functions were embedded within routine clinical decision processes and institutional coordination mechanisms.

The intelligent preconsultation pathway deployed a conversational interface based on a hospital-developed and locally deployed large language model–enabled system, rather than a commercial vendor platform, to elicit structured symptom histories from patients prior to clinical encounters. The pathway generated summarized risk assessments that were automatically transmitted to the electronic medical record (EMR) system, thereby providing preliminary clinical information to support outpatient triage and physician assessment at the point of care.

The oncology multidisciplinary decision-support pathway embedded AI-assisted treatment recommendations within tumor board workflows. It was supported by a hospital-developed oncology decision-support model based on a transformer architecture and adapted for oncology-specific text generation and clinical recommendation tasks. Model-generated outputs were reviewed alongside imaging findings, pathology reports, and patient-specific clinical data under defined specialist decision authority, with final treatment determinations remaining within established clinical accountability structures.

The TDM pathway applied an in-house machine learning pipeline that integrated principal component analysis for feature optimization and a transformer-based deep learning regression model to generate plasma drug concentration estimates for duloxetine and support individualized dosing decisions. These estimates were derived from patient-specific clinical parameters and incorporated into routine medication-management workflows to inform dosing adjustments.

In addition to these clinical pathways, AI-assisted documentation tools were deployed to support routine information management, including generative medical record functions and natural-language clinical assistants facilitating clinical documentation and information retrieval. The platform also enabled structured capture of research indicators and integration of clinical data management with institutional quality-monitoring and downstream analytical processes. Sustained operation across these domains created an empirical context in which AI-enabled systems engaged directly with institutional governance arrangements and cross-departmental coordination structures. This provided the empirical basis for examining how governance mechanisms emerged and evolved during real-world implementation.

Three forms of implementation evidence were observable during the review period, together indicating that the platform had moved beyond bounded pilot deployment. Sustained operation was evident across 3 clinical pathways, with the preconsultation pathway remaining in routine use for more than 12 months and recording more than 24,000 completed patient interactions across 5 outpatient departments. At the same time, AI outputs were incorporated into existing clinical workflows rather than remaining external demonstrations, including automatic transmission of preconsultation summaries to the EMR system, integration of oncology decision support into MDT deliberation, and incorporation of duloxetine concentration prediction into medication-management processes. Continuity of deployment was also accompanied by formal governance consolidation: regulatory authorization was documented in the algorithm filing record, role-based operational oversight was specified in pathway supervision protocols, update procedures were recorded in version-control logs, and organizational ownership was formalized through the medical AI laboratory mandate. These observations collectively provide high-level implementation evidence that the platform functioned as an institutionalized

XSL•FO
RenderX

operational capability rather than remaining a time-limited pilot initiative.

Institutional Carrier Formation

A foundational governance challenge in clinical AI deployment concerns the absence of a stable organizational locus of ownership. Under such conditions, the institutional authority required to coordinate development, deployment, and lifecycle oversight remains fragmented, preventing AI capabilities from transitioning beyond pilot implementation into durable institutional use.

When AI initiatives operate as project-based activities, accountability for cross-departmental data access, clinical validation, and regulatory engagement becomes diffusely distributed across organizational units and external collaborators. In the absence of a designated institutional carrier, technically functional systems frequently fail to persist beyond initial pilot phases because no organizational entity possesses the authority required to align technical development with clinical governance and institutional decision structures.

Governance consolidation in the studied setting occurred through the establishment of a dedicated medical AI laboratory serving as an institutional carrier with a formal mandate and defined decision rights. The laboratory integrated algorithm development, clinical research, and application translation within a unified organizational structure, thereby concentrating responsibility for AI system oversight within a single accountable entity. This carrier was established prior to broader clinical integration and subsequently provided the administrative foundation upon which additional governance arrangements were constructed, including designation as a provincial engineering research center.

Institutional arrangements may vary across health systems in organizational form and administrative configuration. A consistent governance principle nevertheless emerges, whereby sustained clinical AI integration requires a durably designated organizational entity responsible for coordinating development, deployment, and lifecycle oversight. Such carriers may take the form of dedicated laboratories, formally chartered governance committees, or federated institutional arrangements, but must remain structurally distinct from temporary project teams and independent of short-term funding cycles.

Infrastructure Governance

A second governance challenge arises from the treatment of data and computational resources as project-bound technical assets rather than as governed institutional infrastructure. When infrastructure provision remains embedded within temporary project arrangements, the standardization, interoperability, and operational continuity required for routine clinical deployment cannot be reliably sustained. Under such conditions, AI systems remain dependent on configurations that are structurally incompatible with long-term institutional operation.

Fragmented information systems, inconsistent data standards, and reliance on temporary computing environments constrain stable model operation and impede integration into clinical workflows. These limitations typically arise not from data scarcity but from the absence of institutional governance mechanisms capable of ensuring infrastructure continuity, access control, and system-wide coordination across heterogeneous hospital environments. Without formal governance of underlying infrastructure, technically validated AI systems may nevertheless fail to operate reliably at the point of clinical deployment, creating conditions in which deployment continuity depends on informal negotiation rather than defined accountability structures.

Infrastructure governance within the studied platform was progressively incorporated into institutional oversight arrangements. High-performance computing capacity was established and integrated with existing clinical information systems, while external certifications, including EMR functional maturity assessment and interoperability standardization evaluation, confirmed that infrastructure conditions satisfied requirements for clinical deployment. Data standardization across legacy systems further required iterative coordination among clinical, technical, and administrative units, as nonstandardized terminologies embedded within departmental workflows impeded the establishment of unified data pipelines.

Infrastructure architectures differ across institutions in scale, technical configuration, and organizational management. A consistent governance principle nevertheless emerges, whereby data pipelines, computational resources, and interoperability systems must be governed as durable institutional infrastructure, maintained under defined accountability arrangements and independent of the project-bound funding cycles that characterize early-stage AI deployment.

Regulatory and Ethical Governance

A further governance challenge concerns the specification of regulatory authorization and accountability prior to clinical deployment. When regulatory alignment is deferred until after deployment decisions are made, accountability arrangements remain incompletely defined during the period of greatest institutional uncertainty, and the boundaries of legitimate system use are established retrospectively rather than by design.

In the absence of prior regulatory specification, system purpose, intended use boundaries, and designated institutional responsibility may remain insufficiently defined at the point of clinical activation. Effective governance therefore requires that regulatory authorization and accountability structures be established prior to deployment, particularly in regulatory environments that continue to evolve alongside emerging clinical AI capabilities.

Within the Hebei provincial platform, regulatory authorization was obtained prior to broader clinical activation of the oncology decision-support pathway through national algorithm filing procedures administered under the Cyberspace Administration of China framework. The filing process clarified system purpose, defined intended use boundaries, and formally designated institutional responsibility for ongoing deployment. Regulatory requirements therefore informed operational design decisions at the predeployment stage rather than constraining systems already embedded in clinical practice, and oversight functions were incorporated within existing institutional governance

XSL•FO
RenderX

arrangements rather than constituted as a separate compliance structure.

Regulatory frameworks vary substantially across jurisdictions in scope, procedural requirements, and institutional applicability. A consistent governance principle nevertheless emerges, whereby accountability specification and the definition of authorized use boundaries should precede clinical activation. When regulatory alignment occurs only after deployment, governance uncertainty persists during implementation and institutional oversight across the system lifecycle becomes structurally constrained.

Interdisciplinary Operational Coordination

Effective clinical deployment of AI systems depends on structured coordination across clinical, technical, and administrative domains. When such coordination is not institutionally specified, technically functional systems may operate without agreed procedures for clinical use, and accountability for AI-generated outputs may remain diffuse or unassigned.

Ambiguity frequently emerges at the interface between algorithm development and clinical decision-making. Technical personnel may not possess the authority to modify systems in response to operational feedback, while clinicians may lack formally defined procedures for incorporating AI outputs into diagnostic or therapeutic workflows. Without governance arrangements that connect these domains, AI systems may remain technically operational but institutionally unstable.

Operational coordination in the studied implementation was institutionalized through explicit role allocation across a 3-division laboratory structure responsible for algorithm development, clinical research, and application translation. In the oncology multidisciplinary decision-support pathway, AI outputs were embedded within established team workflows under defined decision rights. Model-generated outputs were reviewed during structured multidisciplinary meetings alongside imaging findings, pathology reports, and patient-specific clinical data, while final treatment decisions remained within established clinical accountability structures. Early deployment further revealed that responsibility boundaries required clarification, and one pathway experienced temporary suspension pending resolution of ambiguity regarding authority over model update decisions.

Institutional coordination mechanisms differ across health systems in organizational design and governance structure. A consistent governance principle nevertheless emerges, whereby sustained clinical deployment requires clearly defined roles, decision rights, and accountability structures linking algorithm development, clinical use, and institutional oversight.

Translational Scaling

Durable institutional integration of clinical AI systems requires governance mechanisms capable of extending operational capacity beyond localized implementations. Without such mechanisms, systems demonstrating reliable performance within circumscribed clinical contexts may remain confined to isolated deployments, preventing broader organizational adoption and durable integration at scale.

Localized operational success does not automatically generate the institutional authority required for expansion across departments or clinical service lines. Effective scaling therefore depends on governance arrangements that connect deployment experience with institutional decision structures, enabling operational practices developed in one context to be translated into formally recognized organizational capabilities.

In the studied implementation, scaling occurred through progressive integration of AI-supported pathways into institutional governance structures rather than replication of isolated applications. The preconsultation risk assessment pathway was extended across multiple outpatient departments through standardized workflow integration, enabling its routine use across distinct clinical service lines. Additional AI-supported pathways were subsequently introduced in coordination with clinical and administrative units, each accompanied by governance documentation specifying operational roles, workflow procedures, and system use parameters. Over time, these expansions transformed initially localized tools into institutionally coordinated clinical services embedded within routine care processes. Institutional recognition further reinforced this process through research center designation and the establishment of a provincial clinical AI research institute.

Scaling mechanisms vary across institutional contexts in governance structure and policy environment. A consistent governance principle nevertheless emerges, whereby durable expansion of clinical AI capacity requires governance pathways linking localized deployment with institution-wide authority structures, allowing operational experience to translate into formally recognized institutional capability.

Lifecycle Evaluation and Oversight

Long-term clinical deployment of AI systems requires governance mechanisms capable of monitoring performance, managing system updates, and adapting deployment conditions over time. Without such mechanisms, operational systems may gradually diverge from their intended use conditions as clinical workflows change, data distributions shift, and regulatory expectations develop over time.

In the absence of structured lifecycle oversight, responsibility for monitoring system performance may become diffuse across organizational units. Version control may be inconsistently documented, and operational feedback generated by clinical users may not systematically reach those with authority to modify deployed systems. This configuration introduces the risk of operational drift as AI systems expand beyond initial pilot environments into routine clinical practice, underscoring the necessity of governance mechanisms capable of maintaining accountability across the full system lifecycle.

Within the observed implementation, lifecycle oversight practices were incorporated into existing institutional governance arrangements rather than established as a fully separate governance protocol. Monitoring activities included documentation of system use boundaries, clarification of responsible units for pathway supervision, and iterative

XSL•FO
RenderX

adjustments in response to operational feedback. System updates were conducted within established institutional decision structures and aligned with regulatory requirements. Standardized version-control documentation remained under iterative development during the observation period as institutional governance procedures matured.

Lifecycle governance arrangements vary across institutional contexts in their degree of formalization and organizational structure. Across these variations, a consistent governance principle applies, whereby sustained clinical deployment requires monitoring mechanisms linked to defined institutional authority, ensuring that system updates, clinical workflow adjustments, and evolving regulatory expectations remain subject to accountable oversight throughout the operational lifecycle.

Discussion

From Implementation to Governance Formation

Persistent pilot dependence in clinical AI is commonly attributed to governance deficits, including the absence of accountability structures, ownership arrangements, or regulatory alignment that would enable AI systems to move from experimental deployment into routine clinical use [22,23]. While this interpretation identifies important institutional constraints, it does not fully explain why governance arrangements so often fail to emerge even when pilot implementations are technically successful. The present analysis points to a more precise formulation. In this view, the pilot trap reflects not only the absence of governance but also insufficient implementation depth for governance structures to take shape [4,24]. Systems remain project-bound not because governance was unavailable in principle, but because deployment ended before the operational tensions capable of generating governance had fully materialized. What distinguishes sustained integration from persistent pilot dependence is therefore not the prior existence of a governance framework but the capacity of implementation processes to generate governance through iterative responses to recurrent coordination demands.

Reconceptualizing the pilot trap in institutional terms carries direct implications for how it is addressed. Existing approaches often position governance as a precondition for deployment by emphasizing accountability frameworks, compliance structures, and ownership arrangements before clinical integration proceeds [25,26]. Evidence from this case indicates a different dynamic. Governance capacity did not precede implementation; it emerged through it. Data-standardization conflicts, responsibility-boundary disputes, and scaling misalignments created the institutional conditions under which governance structures became necessary and were progressively consolidated [27]. Deployment therefore proceeded not from a complete governance architecture but from a minimal institutional foundation that accumulated governance capacity as implementation deepened. In contrast to frameworks that emphasize technical integration requirements or clinician adoption barriers as the primary constraints on clinical AI sustainability, the present analysis foregrounds the governance formation process itself. Attention is directed toward the mechanisms through which organizational ownership, accountability boundaries, and coordination structures become stabilized as durable institutional capacity. Across the implementation trajectory, 6 governance modules repeatedly emerged as functional conditions necessary for sustained deployment rather than as elements of a predefined governance design. Considered together, these modules constitute governance capacity as a system-level capability through which sociotechnical resources are consolidated into routinized clinical infrastructure. The operational indicators summarized in Table 2 provide empirical grounding for this interpretation and suggest that governance formation was accompanied by sustained operational embedding rather than remaining a purely organizational or aspirational construct.

These indicators are not reported as performance benchmarks but as governance-relevant evidence. Sustained use across more than 12 months of continuous operation, encompassing more than 24,000 patient interactions within a phased multidepartment deployment, indicates that governance arrangements were sufficient to support accountable clinical operation over time rather than confining the platform to a bounded pilot episode. Operational continuity across the preconsultation pathway is consistent with the stabilization of M1 and M2, without which system continuity would be more likely to depend on informal coordination than on defined accountability structures. Regulatory authorization obtained prior to clinical activation of the oncology decision-support pathway is consistent with the functioning of M3, as documented in the algorithm filing record and related compliance materials. The maintenance of defined clinical decision authority throughout the oncology MDT pathway is consistent with the operationalization of M4. The phased expansion of the preconsultation pathway across multiple outpatient departments is indicative of M5, as documented in deployment materials and departmental rollout records. Iterative system updates conducted within institutional governance structures during the observation period are indicative of M6 functioning as an active governance mechanism rather than a nominal policy commitment. Although governance functions vary in their observability through use metrics, each of the 6 modules is linked to at least one category of traceable institutional artifact, including operational records, regulatory filings, deployment materials, and organizational documents. This supports the descriptive reproducibility of the framework as an analytical lens, whereas its broader transferability across heterogeneous health system contexts is discussed in the following section. Collectively, these indicators suggest that the governance modules described in this framework were associated with observable operational continuity and that the transition from pilot deployment to institutionalized clinical practice was sustained across multiple service lines and governance domains.

From Implementation Tensions to Governance Mechanisms

Implementation of the platform was accompanied by several operational frictions that required institutional responses, through which governance structures gradually emerged [24,28]. Rather than resulting from deliberate institutional design, governance arrangements developed through iterative responses

XSL•FO
RenderX

to practical operational challenges. The analysis identifies 3 generative mechanisms that shaped this process: authorization, responsibility allocation, and coordination.

Authorization translated regulatory expectations into operationally defined deployment boundaries and clarified the legitimate scope and conditions under which clinical AI systems could be used [25,29]. Responsibility allocation embedded AI-related tasks within existing organizational structures by assigning accountable ownership across institutional units and professional roles [30]. Coordination across clinical, technical, and administrative domains was gradually stabilized as decision rights, communication channels, and operational interfaces became more clearly defined [31].

These mechanisms emerged through concrete operational friction rather than prior institutional design. During early deployment of the TDM pathway, inconsistencies in laboratory data formatting and record structures across hospital information systems temporarily limited automated data ingestion and required manual reconciliation. Resolving this tension required sustained coordination among clinical departments, the hospital information center, and the AI development team, ultimately producing standardized data interfaces and validation procedures, which were later incorporated into the governance framework.

A second coordination tension arose during deployment of the oncology MDT decision-support pathway. Clinician uncertainty regarding the appropriate role of AI-generated recommendations within established multidisciplinary decision-making processes led to a temporary operational suspension. Following administrative review, responsibility boundaries and documentation requirements were clarified, and the governance framework formalized AI outputs as advisory inputs under defined clinician oversight with role-based monitoring arrangements.

These episodes illustrate that institutionalization does not occur through seamless technological integration but through the iterative resolution of organizational tensions [27,32]. Governance neither simply precedes implementation nor follows it as a downstream regulatory layer. Instead, governance is produced through the implementation process itself. Institutionalization emerges when governance arrangements stabilize in response to recurrent operational demands, transforming iterative problem-solving into durable institutional capacity embedded within routine organizational practice.

From Governance First to Governance Fit

Effective clinical AI deployment requires governance fit rather than governance expansion. Governance-first approaches are sometimes criticized for introducing procedural rigidity that slows technological innovation. This analysis addresses that concern by distinguishing governance functions from the institutional forms through which they are implemented. This distinction also clarifies why the framework is analytically reproducible across settings even when organizational forms differ.

Certain governance functions constitute nonnegotiable conditions for safe and accountable integration. These include clearly defined authorization boundaries, accountable ownership, and mechanisms for lifecycle oversight. The institutional arrangements through which such functions are enacted, however, may vary substantially in scale, procedural intensity, and degree of centralization across health system contexts. Accordingly, reproducibility in this context lies not in replicating identical organizational structures but in assessing whether comparable governance functions can be identified through traceable institutional evidence.

Governance fit refers to the calibrated alignment between governance intensity and system characteristics such as clinical risk, update frequency, and operational complexity [25,33]. Under this perspective, governance architectures should not impose uniform procedural requirements across all deployment environments. Instead, effective governance preserves adaptive capacity by calibrating oversight intensity according to the potential consequences and reversibility of system changes.

Several mechanisms illustrate how such calibration can be operationalized in practice. Risk-tiered update thresholds allow routine model refinements to proceed under simplified authorization procedures while reserving stricter review processes for high-impact system modifications. Predefined authorization pathways for low-risk adjustments further reduce procedural bottlenecks. Staged deployment strategies similarly enable systems to evolve within monitored operational environments before wider clinical integration [34].

Both extremes generate systemic challenges. Insufficient governance creates fragmentation and reinforces persistent pilot dependence, whereas excessive governance may introduce bureaucratic inertia that constrains technological adaptation [35,36]. Health systems can therefore balance accountability with technological agility by implementing governance arrangements that differentiate routine system evolution from high-risk modification processes while maintaining clear institutional responsibility for oversight. The broader question of how these governance functions travel across institutional environments is addressed in the following section on functional transferability.

From Clinician-Centric to Participatory Governance Maturation

The absence of structural patient and community involvement in the present case invites a more precise analytical observation than simply noting a gap. Existing literature on patient participation in clinical AI often conflates two distinct roles. Patients may participate as design informants who help shape decisions about system purpose, or as governance stakeholders who contribute to oversight of how deployed systems operate within clinical environments [37,38]. Most existing guidance addresses the former by emphasizing co-design principles and participatory development methodologies. By contrast, patient involvement in operational governance has received substantially less analytical attention, and the institutional conditions required to make such participation structurally meaningful remain insufficiently specified.

Participatory governance capacity is not a function of inclusion per se but of the institutional anchoring through which

XSL•FO
RenderX

participatory inputs acquire an accountable pathway to governance decisions. Meaningful patient involvement in AI governance requires an organizational carrier capable of receiving, processing, and acting on participatory inputs within established responsibility structures. Without such a carrier, participation risks becoming consultative rather than consequential, producing inputs that lack a defined institutional pathway to governance decisions [39]. The sequencing observed here, therefore, reflects a structural constraint rather than a deliberate exclusion. Participatory mechanisms were introduced following initial stabilization of the accountability architecture, a sequencing that reflected the specific institutional conditions of this implementation context rather than a principled position on the optimal timing of patient involvement. Contemporary co-design frameworks and implementation science literature frequently advocate for patient involvement from the earliest stages of governance development, and parallel co-design of governance and participation structures represents a well-documented and viable alternative in other health system contexts.

As institutionalization advances and governance architecture stabilizes, participatory capacity becomes both feasible and functionally necessary. Within the proposed framework, patient and community involvement is most appropriately integrated into lifecycle governance mechanisms. Defined feedback channels, transparent communication of system use boundaries, and structured review of operational incidents allow oversight to extend beyond professional actors while preserving clear lines of institutional accountability [38,40]. This integration addresses not only the legitimacy requirements identified in the literature but also the operational question of how participatory inputs are processed within functioning governance structures. Design-focused participation frameworks rarely specify these institutional pathways, leaving the governance dimension of participation largely unresolved [37,41].

From Structural Replication to Functional Transferability

Although the framework emphasizes governance capacity rather than technological sophistication, several baseline conditions support its applicability across health system contexts [42]. These conditions represent minimal enabling environments for sustained clinical AI operation rather than structural requirements of the framework itself. Institutional authority to assign responsibility for AI oversight must be present. Basic digital infrastructure capable of supporting clinical data capture, storage, and computational processing is also required, together with regulatory or administrative mechanisms that permit AI tools to be activated within clinical workflows. Interdisciplinary coordination capacity must additionally connect clinical practice, technical development, and organizational management to sustain cross-functional collaboration. Where such conditions are only partially present, functional equivalents such as shared infrastructure platforms, external technical partnerships, or regional governance arrangements may provide alternative implementation pathways [43].

Framework transferability therefore lies in functional reproduction rather than structural replication [22]. In this paper, functional transferability is used as an analytic formulation to distinguish transferable governance functions from the setting-specific organizational forms through which they are instantiated. Institutionalization does not require identical organizational architectures but depends on the presence of governance functions capable of stabilizing clinical AI as durable institutional capacity. Six functional requirements underpin this process, including durable ownership arrangements, governed infrastructure, defined authorization boundaries, cross-functional coordination mechanisms, structured scaling pathways, and lifecycle oversight. Under this interpretation, governance transferability depends on whether equivalent institutional functions can be realized through different organizational arrangements across settings, rather than on replication of a single structural form. In this sense, reproducibility should be understood descriptively rather than structurally. What can be reproduced across settings is not an identical organizational form but an analytic approach that examines whether comparable governance functions are evidenced through traceable institutional artifacts. Table 3 illustrates how these requirements may be instantiated across heterogeneous health system contexts. It presents both the institutional arrangements observed in the study setting and illustrative analogs that may fulfill equivalent governance roles in more decentralized health systems. The purpose of this mapping is therefore to operationalize functional equivalence across contexts, rather than to prescribe identical organizational structures.

XSL•FO
RenderX

Table . Functional governance requirements and cross-system analogs.

Governance module	Functional requirement	Example instantiation in study context	Possible analog in decentralized health systems
M1: Institutional carrier formation	Accountable ownership formally assigned beyond project teams with defined decision authority across deployment, maintenance, and update cycles.	Provincial engineering research center and hospital AI[a] laboratory.	Health system AI governance office or clinical AI steering committee.
M2: Infrastructure governance	Data and computational resources governed as institutional infrastructure ensuring continuity of access and interoperability with clinical systems.	Institutionally governed data platforms and high-performance computing environment.	Cloud-based infrastructure operating under institutional data governance agreements.
M3: Regulatory and ethical governance	Predeployment authorization establishing legal legitimacy and defined accountability boundaries for clinical AI use.	National algorithm filing and institutional compliance documentation.	Software-as-a-medical-device approval pathways and institutional compliance review processes.
M4: Interdisciplinary coordination	Stable cross-functional coordination with defined roles and decision rights linking clinical, technical, and administrative actors.	Three-division laboratory integrating clinical, technical, and translational roles.	Cross-functional clinical AI committees linking clinicians, engineers, and administrators.
M5: Translational scaling	Governance mechanisms linking localized deployment with higher-level institutional authority to support controlled expansion.	Regional medical alliances extending AI services across affiliated hospitals.	Multisite health system networks or collaborative implementation consortia.
M6: Lifecycle evaluation and oversight	Sustained monitoring and controlled system evolution through performance review, version control, and governance adjustment mechanisms.	Role-based monitoring and documented change-control procedures.	Model oversight committees and postdeployment monitoring programs.

[a]AI: artificial intelligence.

Economic feasibility represents an important consideration for broader adoption of governance-oriented implementation models [43]. The framework does not require the independent construction of large-scale computational infrastructure. The operative condition is continuity of access to governed data and computational resources rather than infrastructural ownership. In practice, this requirement may be fulfilled through shared regional platforms, cloud-based infrastructures operating under institutional data governance agreements, or staged implementation strategies focusing on high-value clinical pathways. These configurations allow institutions across resource levels to establish governance capacity without replicating the infrastructural scale characteristic of tertiary referral centers. Resource pooling, coordinated implementation, and shared accountability structures may therefore support deployment across diverse institutional configurations, including regional health networks, public-private partnerships, and externally supported platforms.

This functional interpretation situates the framework within the broader landscape of international AI governance initiatives while clarifying its distinct contribution. Existing frameworks, including ISO/IEC 42001 [44], World Health Organization guidance on health AI governance [19], and regulatory pathways developed by agencies such as the US Food and Drug Administration [45], share a common structural assumption that the organizational capacity to implement governance already exists within deploying institutions. These frameworks specify what governance should achieve, including accountability documentation, risk management structures, and compliance requirements. However, they presuppose the existence of an institutional carrier capable of activating and sustaining these arrangements [25,28]. The framework developed here addresses the prior question of how such capacity comes into existence. It identifies the conditions under which organizational ownership, coordination authority, and oversight mechanisms emerge through implementation rather than being fully established in advance [22,24]. These two levels should therefore be seen as complementary rather than competing. Normative standards define the accountability architecture within which clinical AI should operate, while the present framework explains the institutional formation process through which that architecture becomes operationally viable at the point of deployment. The framework is therefore transferable not because institutions can or should replicate the Hebei case in structural detail, but because comparable governance functions may be identified, examined, and adapted across heterogeneous implementation settings.

From Viewpoint to Research Agenda

The framework presented in this study derives from a single institutional case observed during an early stage of governance consolidation, which necessarily constrains generalizability [46]. Embedded observation enabled detailed insight into implementation dynamics but also introduced interpretive proximity. Triangulation across traceable institutional artifacts and iterative review by coauthors with distinct clinical, administrative, and governance roles were used to mitigate this

XSL•FO
RenderX

limitation. Lifecycle governance mechanisms were still evolving during the observation period, and participatory governance remained only partially developed. The scope of claims the present analysis can support is accordingly bounded by these conditions. Although the framework may be analytically transferable across settings, its empirical adequacy and practical consequences remain to be tested through comparative research.

Beyond the constraints of this single case, a more fundamental limitation lies in the current state of the field. Clinical AI governance research has generated a substantial body of normative frameworks, implementation principles, and policy guidance, yet the relationship between governance arrangements and clinical outcomes remains largely unexamined [47]. Governance frameworks are frequently justified by the assumption that they improve care quality, patient safety, and operational reliability. However, empirical investigation of this assumption remains conspicuously limited. Governance studies rarely measure clinical outcomes, and outcome studies rarely examine governance structures [47]. The result is a structural circularity in which the value of governance is asserted but seldom demonstrated empirically.

Breaking this circularity requires a specific methodological intervention, namely longitudinal comparative studies that track governance formation and clinical outcomes simultaneously, treating governance artifacts, including carrier establishment, authorization records, coordination protocols, and oversight documentation, as traceable intermediate variables in a causal pathway from implementation to clinical impact [48]. Investigations of this kind would need to span multiple institutional contexts to distinguish governance effects from institutional confounders and would require follow-up periods sufficient to capture the maturation of lifecycle governance mechanisms beyond early implementation phases. Economic analyses characterizing the cost structures of alternative governance configurations constitute an equally pressing methodological demand [49], particularly for institutions operating with resource constraints that preclude replication of tertiary-center infrastructure. Most fundamentally, the field requires an agreed methodology for attributing clinical and operational outcomes to governance arrangements rather than to technical model performance alone [14,47], a distinction that existing evaluation frameworks are currently ill-equipped to draw.

Conclusions

Clinical AI adoption is typically framed as a problem of technology transfer, understood as the movement of capable systems from development into clinical practice. The present analysis points toward a more fundamental reconceptualization. What health systems are required to construct is not a pathway for technology to enter institutions but the institutional capacity through which technology becomes governable, accountable, and sustained. Institutional capacity of this kind is not established in advance of implementation; it is generated through it. The operational tensions that arise during deployment are not obstacles to governance formation. They are its generative conditions. The 6-module framework developed in this study identifies the functional arrangements through which this formation process stabilizes into sustained institutional capacity. The framework thus reorients institutional strategy away from governance prespecification and toward the cultivation of implementation conditions capable of generating governance over time. Institutions that recognize governance as a product of implementation rather than its prerequisite are better positioned to support the iterative process through which clinical AI transitions from experimental artifact to organizational infrastructure. Without such capacity, technically mature systems remain perpetually promising pilots awaiting conditions that never arrive.

Acknowledgments

We acknowledge the clinical departments, information technology teams, and administrative units involved in the development and routine operation of the provincial clinical artificial intelligence platform. Their contributions to system integration, workflow embedding, and governance documentation provided essential implementation context for this viewpoint. During manuscript preparation, generative artificial intelligence tools were used for language editing and clarity improvements. All substantive scientific content, interpretations, and final wording were reviewed and approved by the authors, who take full responsibility for the manuscript.

Funding

No external funding was received for this study.

Data Availability

The datasets generated and analyzed during this study consist of institutional governance documents, regulatory filings, infrastructure certification records, internal operational monitoring records, and other operational implementation artifacts. These materials are not publicly available due to institutional governance policies and regulatory restrictions. Requests for access to deidentified or aggregated information may be directed to the corresponding author and will be considered subject to institutional review and approval by the relevant authorities.

Authors' Contributions

Conceptualization: JT, ZZ

XSL•FO
RenderX

Formal analysis: JT, YL
Investigation: YS, NJ
Methodology: JT, YL
Supervision: ZZ
Writing – original draft: JT
Writing – review & editing: JT, ZZ, LT, YS, YL, NJ

Conflicts of Interest

None declared.

References

1. Obermeyer Z, Emanuel EJ. Predicting the future – big data, machine learning, and clinical medicine. N Engl J Med 2016 Sep 29;375(13):1216-1219. [doi: 10.1056/NEJMp1606181] [Medline: 27682033]
2. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019 Jan;25(1):30-36. [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]
3. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019 Oct 29;17(1):195. [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
4. Ahmed MI, Spooner B, Isherwood J, Lane M, Orrock E, Dennison A. A systematic review of the barriers to the implementation of artificial intelligence in healthcare. Cureus 2023 Oct;15(10):e46454. [doi: 10.7759/cureus.46454] [Medline: 37927664]
5. Maddox TM, Rumsfeld JS, Payne PRO. Questions for artificial intelligence in health care. JAMA 2019 Jan 1;321(1):31-32. [doi: 10.1001/jama.2018.18932] [Medline: 30535130]
6. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
7. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019 Sep;25(9):1337-1340. [doi: 10.1038/s41591-019-0548-6] [Medline: 31427808]
8. Hassan M, Kushniruk A, Borycki EM. Barriers to and facilitators of artificial intelligence adoption in health care: scoping review. JMIR Hum Factors 2024 Aug 29;11:e48633. [doi: 10.2196/48633] [Medline: 39207831]
9. Wang F, Casalino LP, Khullar D. Deep learning in medicine—promise, progress, and challenges. JAMA Intern Med 2019 Mar 1;179(3):293-294. [doi: 10.1001/jamainternmed.2018.7117] [Medline: 30556825]
10. Wang Z. Artificial intelligence in Chinese healthcare: a review of applications and future prospects. Biomed Eng Lett 2025 Nov;15(6):1065-1072. [doi: 10.1007/s13534-025-00515-2] [Medline: 41280149]
11. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
12. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019 Oct;1(6):e271-e297. [doi: 10.1016/S2589-7500(19)30123-2] [Medline: 33323251]
13. Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: mapping the debate. Big Data Soc 2016 Dec;3(2):2053951716679679. [doi: 10.1177/2053951716679679]
14. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. N Engl J Med 2018 Mar 15;378(11):981-983. [doi: 10.1056/NEJMp1714229] [Medline: 29539284]
15. Hogg HDJ, Al-Zubaidy M, Technology Enhanced Macular Services Study Reference Group, et al. Stakeholder perspectives of clinical artificial intelligence implementation: systematic review of qualitative evidence. J Med Internet Res 2023 Jan 10;25:e39742. [doi: 10.2196/39742] [Medline: 36626192]
16. Choudhury A, Asan O. Role of artificial intelligence in patient safety outcomes: systematic literature review. JMIR Med Inform 2020 Jul 24;8(7):e18599. [doi: 10.2196/18599] [Medline: 32706688]
17. Morley J, Floridi L, Kinsey L, Elhalal A. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci Eng Ethics 2020 Aug;26(4):2141-2168. [doi: 10.1007/s11948-019-00165-5] [Medline: 31828533]
18. Shaw J, Rudzicz F, Jamieson T, Goldfarb A. Artificial intelligence and the implementation challenge. J Med Internet Res 2019 Jul 10;21(7):e13659. [doi: 10.2196/13659] [Medline: 31293245]
19. Ethics and governance of artificial intelligence for health. World Health Organization. 2021 Jun 28. URL: https://www.who.int/publications/i/item/9789240029200 [accessed 2026-04-22]
20. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. J Am Med Inform Assoc 2020 Mar 1;27(3):491-497. [doi: 10.1093/jamia/ocz192] [Medline: 31682262]
21. Coiera E. The last mile: where artificial intelligence meets reality. J Med Internet Res 2019 Nov 8;21(11):e16323. [doi: 10.2196/16323] [Medline: 31702559]

XSL•FO
RenderX

22. Panch T, Mattie H, Celi LA. The “inconvenient truth” about AI in healthcare. NPJ Digit Med 2019;2:77. [doi: 10.1038/s41746-019-0155-4] [Medline: 31453372]
23. Greenhalgh T, Abimbola S. The NASSS framework – a synthesis of multiple theories of technology implementation. Stud Health Technol Inform 2019 Jul 30;263:193-204. [doi: 10.3233/SHTI190123] [Medline: 31411163]
24. Nilsen P, Seing I, Ericsson C, Birken SA, Schildmeijer K. Characteristics of successful changes in health care organizations: an interview study with physicians, registered nurses and assistant nurses. BMC Health Serv Res 2020 Feb 27;20(1):147. [doi: 10.1186/s12913-020-4999-8] [Medline: 32106847]
25. Nilsen P, Svedberg P, Neher M, et al. A framework to guide implementation of AI in health care: protocol for a cocreation research project. JMIR Res Protoc 2023 Nov 8;12:e50216. [doi: 10.2196/50216] [Medline: 37938896]
26. Bates DW, Auerbach A, Schulam P, Wright A, Saria S. Reporting and implementing interventions involving machine learning and artificial intelligence. Ann Intern Med 2020 Jun 2;172(11 Suppl):S137-S144. [doi: 10.7326/M19-0872] [Medline: 32479180]
27. Ajami S, Bagheri-Tadi T. Barriers for adopting electronic health records (EHRs) by physicians. Acta Inform Med 2013;21(2):129-134. [doi: 10.5455/aim.2013.21.129-134] [Medline: 24058254]
28. Liao F, Adelaine S, Afshar M, Patterson BW. Governance of clinical AI applications to facilitate safe and equitable deployment in a large health system: key elements and early successes. Front Digit Health 2022;4:931439. [doi: 10.3389/fdgth.2022.931439] [Medline: 36093386]
29. Liu Y, Yu W, Dillon T. Regulatory responses and approval status of artificial intelligence medical devices with a focus on China. NPJ Digit Med 2024 Sep 18;7(1):255. [doi: 10.1038/s41746-024-01254-x] [Medline: 39294318]
30. Guiding principles for registration review of artificial intelligence medical devices. Center for Medical Device Evaluation, National Medical Products Administration. 2022 Mar 10. URL: https://www.cncsdr.org/ggtz/ggzz/202203/t20220311_304188.html [accessed 2026-05-05]
31. Wang F, Beecy A. Implementing AI models in clinical workflows: a roadmap. BMJ Evid Based Med 2025 Sep 22;30(5):285-287. [doi: 10.1136/bmjebm-2023-112727] [Medline: 38914450]
32. Nair M, Svedberg P, Larsson I, Nygren JM. A comprehensive overview of barriers and strategies for AI implementation in healthcare: mixed-method design. PLoS ONE 2024;19(8):e0305949. [doi: 10.1371/journal.pone.0305949] [Medline: 39121051]
33. Greenhalgh T, Papoutsi C. Studying complexity in health services research: desperately seeking an overdue paradigm shift. BMC Med 2018 Jun 20;16(1):95. [doi: 10.1186/s12916-018-1089-4] [Medline: 29921272]
34. Sendak MP, Ratliff W, Sarro D, et al. Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. JMIR Med Inform 2020 Jul 15;8(7):e15182. [doi: 10.2196/15182] [Medline: 32673244]
35. McKee M, Wouters OJ. The challenges of regulating artificial intelligence in healthcare; Comment on “Clinical decision support and new regulatory frameworks for medical devices: are we ready for it? – A viewpoint paper”. Int J Health Policy Manag 2023;12:7261. [doi: 10.34172/ijhpm.2022.7261] [Medline: 36243948]
36. Gama F, Tyskbo D, Nygren J, Barlow J, Reed J, Svedberg P. Implementation frameworks for artificial intelligence translation into health care practice: scoping review. J Med Internet Res 2022 Jan 27;24(1):e32215. [doi: 10.2196/32215] [Medline: 35084349]
37. Busuioc M. Accountable artificial intelligence: holding algorithms to account. Public Adm Rev 2021;81(5):825-836. [doi: 10.1111/puar.13293] [Medline: 34690372]
38. Trocin C, Mikalef P, Papamitsiou Z, Conboy K. Responsible AI for digital health: a synthesis and a research agenda. Inf Syst Front 2023 Dec;25(6):2139-2157. [doi: 10.1007/s10796-021-10146-4]
39. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. PLoS Med 2018 Nov;15(11):e1002689. [doi: 10.1371/journal.pmed.1002689] [Medline: 30399149]
40. Mittelstadt B. Principles alone cannot guarantee ethical AI. Nat Mach Intell 2019;1(11):501-507. [doi: 10.1038/s42256-019-0114-4]
41. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell 2019;1(9):389-399. [doi: 10.1038/s42256-019-0088-2]
42. Owoyemi A, Owoyemi J, Osiyemi A, Boyd A. Artificial intelligence for healthcare in Africa. Front Digit Health 2020;2:6. [doi: 10.3389/fdgth.2020.00006] [Medline: 34713019]
43. Alami H, Lehoux P, Auclair Y, et al. Artificial intelligence and health technology assessment: anticipating a new level of complexity. J Med Internet Res 2020 Jul 7;22(7):e17707. [doi: 10.2196/17707] [Medline: 32406850]
44. Benraouane SA. AI Management System Certification According to the ISO/IEC 42001 Standard, 1st edition: Productivity Press; 2024. [doi: 10.4324/9781003463979]
45. Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. US Food and Drug Administration. 2021 Jan. URL: https://www.fda.gov/media/145022/download [accessed 2026-05-05]
46. Yin RK. Case Study Research: Design and Methods, 5th edition: SAGE; 2014.
47. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]

XSL•FO
RenderX

48. Bates DW, Levine D, Syrowatka A, et al. The potential of artificial intelligence to improve patient safety: a scoping review. NPJ Digit Med 2021 Mar 19;4(1):54. [doi: 10.1038/s41746-021-00423-6] [Medline: 33742085]
49. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019 Apr 4;380(14):1347-1358. [doi: 10.1056/NEJMra1814259] [Medline: 30943338]

Abbreviations

AI: artificial intelligence
EMR: electronic medical record
MDT: multidisciplinary team
TDM: therapeutic drug monitoring

Edited by A Coristine; submitted 02.Feb.2026; peer-reviewed by L Fletscher, V Palama; revised version received 19.Mar.2026; accepted 06.Apr.2026; published 07.May.2026.

Please cite as:
Tian J, Zhao Z, Tang L, Song Y, Li Y, Jiang N
From Pilot Trap to Institutional Capacity: A Governance Framework for Sustainable Clinical AI Implementation in Health Systems
J Med Internet Res 2026;28:e92680
URL: https://www.jmir.org/2026/1/e92680
doi:10.2196/92680

© Jin Tian, Zengren Zhao, Longmei Tang, Yongzhao Song, Yuchang Li, Nan Jiang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Governing Patient-Facing AI-Generated Video in Digital Health: A Risk-and-Ethics Matrix for Deployment, Monitoring, and Change Control

Yongzheng Hu[1,2], MD; Wei Jiang[1,2], MD

[1]Department of Nephrology, Affiliated Hospital of Qingdao University, 16 Jiangsu Road, Shinan District, Qingdao, Shandong, China

[2]Department of Nephrology, Qingdao University, Qingdao, Shandong, China

Corresponding Author:
Wei Jiang, MD
Department of Nephrology, Affiliated Hospital of Qingdao University, 16 Jiangsu Road, Shinan District, Qingdao, Shandong, China

Abstract

In this Viewpoint, we argue that patient-facing high-fidelity artificial intelligence (AI)–generated video requires governance that is operational, life cycle based, and embedded in existing institutional review pathways rather than limited to predeployment checks alone. Patient-facing high-fidelity AI-generated video—synthetic or substantially AI-mediated video that presents realistic human likeness, voices, or clinical communication cues—is rapidly entering patient education and clinical communication. We propose a risk-and-ethics matrix that combines residual clinical risk (likelihood × severity after mitigations) with an ethical alignment score that operationalizes autonomy, beneficence, nonmaleficence, and justice to yield actionable dispositions (encourage, permit with oversight, restrict or redesign, or prohibit). The framework links each disposition to dossier-based review, minimum controls, and postdeployment monitoring triggers—focused on measurable outcomes (eg, comprehension, content-attributable follow-up burden, incidents and complaints, and equity gaps) as well as provenance and change control—to support auditable, revisitable decisions over the system life cycle.

(J Med Internet Res 2026;28:e91940) doi:10.2196/91940

KEYWORDS

artificial intelligence; AI; digital health; deepfakes; risk management; postdeployment monitoring

Introduction

High-fidelity artificial intelligence (AI)–generated video—from text-to-video patient explainers to deepfake-style clinician avatars—is entering digital health via patient portals, telehealth workflows, and social platforms [1,2]. In this Viewpoint, we use this term to refer to synthetic or substantially AI-mediated video that presents realistic human likeness, voice, or other clinically salient communication cues in ways that may influence patient trust, comprehension, or decisions. We use “operational governance” to mean the institutional processes through which such systems are reviewed, approved, monitored, and re-evaluated over time. “Real-world deployment” refers to routine use outside controlled testing environments, including use through patient portals, telehealth workflows, apps, and social media channels where content may be redistributed or consumed without direct clinician mediation. “Iterative system change” refers to postdeployment modifications to models, prompts, templates, scripts, rendering pipelines, distribution channels, or disclosure and provenance controls that may materially alter system behavior. Early published examples suggest potential value for patient education and communication, including usability-tested patient digital twins for critical care education, avatar-based educational interventions associated with improved parental knowledge and care skills in hydrocephalus, and pilot or specialty use cases in radiology and postoperative patient communication [3-6]. However, governance often lags behind routine deployment, where content is redistributed across channels and iteratively updated (model, prompt, or template changes). These deployments are judged by outcomes beyond technical accuracy. Real-world clinical performance should be assessed using measurable end points that capture patient and system impact, including comprehension (eg, teach back checks), content-attributable follow-up burden, incident and complaint rates, and equity gaps across language and health literacy groups. This gap motivates a workflow-integrated approach that links upfront review to postdeployment monitoring, incident response, and change control across the life cycle [7,8].

The same properties that make AI-generated video attractive in digital health—realism, personalization, and rapid iteration—also create failure modes that are difficult to detect and manage once content is deployed across heterogeneous channels [9,10]. In routine settings, videos may be reposted or clipped outside institutional portals, updated as models and prompt templates change, and consumed without clinician

XSL•FO
RenderX

context—conditions that allow for small errors to propagate into clinically consequential misinformation, unnecessary follow-up burden, or delayed care [11-13]. Identity cues embedded in video (eg, clinician likeness, institutional branding, or emotionally resonant avatars) can amplify perceived authority and trust, increasing the impact of misstatements, undisclosed synthetic identity, and privacy misuse. Before introducing residual clinical risk, it is important to distinguish it from clinical risk more broadly. In this paper, "clinical risk" refers to the possibility that patient-facing AI-generated video contributes to clinically relevant harm, including misinformation that changes care, delayed help seeking, unnecessary follow-up burden, privacy or identity misuse, psychological distress, or inequitable performance across patient groups. Risk becomes residual when these foreseeable failure modes are reassessed after proposed safeguards—such as clinician script review, constrained generation, authenticated distribution, disclosure, provenance controls, and escalation pathways—have been specified. Therefore, the matrix classifies the clinically relevant risk that remains after mitigation rather than the unmitigated theoretical hazard.

We propose a workflow-integrated governance approach: the risk-and-ethics matrix. It links residual clinical risk—defined here as the remaining likelihood and severity of clinically relevant harm after mitigation—to a principlism-based ethical alignment score (EAS) to support deployment decisions for patient-facing AI-generated video. Plotting these dimensions yields actionable dispositions (encourage, permit with oversight, restrict or redesign, and prohibit) and links each to minimum controls—dossier-based documentation, disclosure requirements, and human oversight where appropriate—as well as predefined postdeployment monitoring metrics and rereview triggers. We use representative scenarios to show how health systems can translate ethical commitments and probabilistic harms into auditable, revisitable decisions across the life cycle, particularly as content is updated and redistributed beyond its original workflow. We then summarize key risk mechanisms, present the scoring rubric and workflow, and map representative use cases of monitoring and change control actions.

The framework is intended for institutional decision-makers rather than for platform-wide moderation. In this paper, the relevant videos are those created, commissioned, adapted, or sponsored for patient-facing use. Typical producers include health systems; clinicians; patient education teams; digital health vendors; and researchers working in care delivery, education, or protocolized specialist settings. The primary governing bodies are local institutional actors such as institutional review boards (IRBs), digital health or clinical governance committees, patient education and communications leaders, and safety and IT oversight teams. Their role is to decide whether a proposed use case should be approved, under what minimum controls, and with what monitoring and rereview conditions. Existing laws, policies, and AI governance frameworks remain essential, but they often operate at a higher level of abstraction and do not specify how institutions should translate transparency, consent, safety, equity, provenance, and change control expectations into case-level deployment decisions for patient-facing AI-generated video use. What the matrix adds is not a replacement for law or formal regulation but an operational layer for institutional decision-making. It converts broad requirements such as transparency, human oversight, safety, equity, provenance, and accountability into case-level classifications, minimum controls, and life cycle management actions for specific patient-facing AI-generated video use cases [14].

The aim of this Viewpoint is to argue that governance of patient-facing AI-generated video should connect residual risk assessment and ethical alignment to concrete institutional decisions, life cycle monitoring, and change control. To support this argument, we outline the main risk mechanisms; present the risk-and-ethics matrix and its workflow; and then discuss implementation, validation, and adaptation across institutional and regulatory contexts.

Risk Mechanisms in Routine Digital Health Deployment of AI-Generated Video

When AI-generated video is deployed routinely across patient portals, telehealth workflows, and social platforms, failure modes emerge that are not well captured by predeployment validation and, therefore, require postdeployment monitoring, incident response, and change control. First, misinformation and content or performance drift (eg, model updates, guideline changes, prompt template changes, and channel shifts) pose direct hazards to patient decision-making [15]. Hyperrealistic "clinician" avatars can convey inaccurate advice with a credibility premium that textual chatbots rarely command. Subtle script hallucinations and the lack of standardized clinical review workflows in many deployments amplify the chance that viewers will act on falsehoods before clinicians can intervene [16,17]. These risks are heightened in asynchronous, public-facing channels where corrections lag behind dissemination and platform ranking may preferentially surface engaging content.

Second, identity misuse and privacy infringements are uniquely salient when the video is the medium because identity cues drive trust calibration and downstream adherence. Unauthorized cloning of a clinician or patient's likeness undermines autonomy, consent, and informational self-determination. Even ostensibly therapeutic recreations such as those involving deceased relatives raise unresolved questions about posthumous privacy and family interests [18,19]. Because mere visual plausibility confers trust, impersonation can catalyze fraud and degrade the informational environment far beyond the index case.

Third, psychological impact is bidirectional and context dependent. Video immersive qualities can deepen engagement, although they may also retraumatize, induce overattachment to synthetic figures, or blur boundaries between memory and simulation in grief and trauma work [20,21]. Minimizing these harms requires careful screening, clear framing, and predefined discontinuation criteria with escalation pathways to human care—not only technical guardrails.

Operationalizing these concerns should prioritize minimal, measurable guardrails. For comprehension and misinformation, institutions should track user-reported confusion, unplanned

XSL•FO
RenderX

follow-up contacts attributable to the content, and brief comprehension checks (eg, teach back–style questions) in representative patient groups [22]. For identity and autonomy, all patient-facing deployments should meet a baseline disclosure standard, including clear labeling of synthetic content, explicit affirmation that no real clinician is speaking, and an accessible opt out. Psychological risk warrants prescreening, short validated distress scales, and predefined stop rules. Equity should be audited through stratified analyses of comprehension, incident or complaint rates, and follow-up burden (eg, by language, age, and health literacy). These metrics make benefits and harms visible enough to guide iterative redesign and trigger rereviews when thresholds are crossed.

Fourth, authenticity and institutional trust are collective goods at stake. As synthetic media saturate telehealth, patients may begin to doubt legitimate communications ("Is this my doctor or an AI?") [23,24]. The resulting frictions—hesitation to follow instructions and demand for redundant confirmation—impose hidden costs on clinicians and organizations. Therefore, provenance signals and disclosure norms matter not as mere formalities but as trust-preserving infrastructure.

Finally, justice and equity considerations cut across all preceding risks. Benefits may accrue first to well-resourced settings that can build multilingual, culturally attuned avatars, whereas harms—deception, confusion, and exploitation—disproportionately fall on groups with lower health literacy or access to verification tools [25,26]. Thus, equity-oriented design, performance disaggregation, and complaint path accessibility are ethical requirements, not optional enhancements.

Operational Governance Framework: The Risk-And-Ethics Matrix

We present an operational governance framework—the risk-and-ethics matrix—that supports deployment decisions for patient-facing AI-generated video by linking residual clinical risk to a principlism-based EAS and to predefined monitoring and rereview actions. Existing laws, policies, and AI governance frameworks establish essential high-level expectations. They do not usually specify how institutions should adjudicate a concrete patient-facing AI-generated video use case, what minimum controls should accompany approval, or when iterative changes should trigger rereview. Purely technical risk scoring tends to underweight autonomy and justice [27,28], whereas principle-first approaches can ignore how likely and severe harms are in actual practice [28-31]. Our integration preserves the strengths of both approaches and translates them into decisions that IRBs, hospital digital health and patient education governance groups, communications leaders, and safety and IT oversight committees can defend, document, and audit.

Here, we distinguish inherent or unmitigated clinical hazard from residual clinical risk, which is the basis for governance classification. On the risk axis, we adapt a hospital-grade matrix consistent with common clinical risk management concepts in which risk reflects the combination of probability and severity. For each use case, we score (1) the likelihood that a specified harm scenario will occur on a 4-level ordinal scale (rare, unlikely, possible, and likely) and (2) the severity of plausible consequences on a 4-level scale (negligible, minor, major, and catastrophic). Cross-tabulation yields composite tiers of low, moderate, high, and extreme residual risk. Assessment proceeds by enumerating the failure modes specific to synthetic video—misinformation that could change care, identity or privacy breaches, psychologically triggering content, or equity harms [19,32,33]—and then rating each mode and assigning the overall tier according to the highest credible residual risk after proposed mitigations. Mitigations—such as clinician review of scripts, constrained generation, authenticated distribution, and provenance or disclosure controls—are recorded in the dossier with versioning and change logs so that residual (not theoretical) risk is the basis of classification over time.

On the ethics side, we operationalize the 4 principles—autonomy, beneficence, nonmaleficence, and justice—into an EAS ranging from 0 to 8. Each principle receives a score of 0 when violated, a score of 1 when partially upheld, and a score of 2 when clearly upheld, guided by concrete criteria that map abstract duties to observable practices. Autonomy considers the transparency of AI use, the accuracy of identity representation, voluntariness, and the adequacy of consent in a video medium [34]; beneficence requires a credible, evidence-informed benefit that is proportionate to the foreseeable burdens [35]; nonmaleficence emphasizes minimizing physical, psychological, informational, and reputational harms and guarding against foreseeable misuse [36]; and justice attends to equitable access and performance across groups, bias mitigation, nonexploitation of vulnerable populations, and preservation of public trust [37]. We band the EAS scores as high (7-8), medium (4-6), or low (0-3); where evidence is limited, conservative scoring and explicit uncertainty statements are needed. Table 1 shows the scales and rubric.

XSL•FO
RenderX

Table . Scales and rubric for the risk-and-ethics matrix.

Component and level or principle	Definition or criterion
Likelihood	
Rare	Rare under routine conditions; requires multiple safeguards to fail
Unlikely	Single lapse or unusual context
Possible	Common precursor conditions present
Likely	Reproducible under routine conditions
Severity	
Negligible	No decision impact; self-correcting (eg, brief uncertainty without behavior change)
Minor	Transient confusion; extra contact (eg, 1 follow-up call or portal message for clarification)
Major	Clinically consequential misinformation or marked psychological harm (eg, delayed care, inappropriate self-management, or significant distress requiring clinician intervention)
Catastrophic	Severe harm or system-level misinformation (eg, serious injury, widespread harmful misinformation, or crisis-level psychological destabilization)
EAS[a]—autonomy	
0	No or unclear disclosure; misleading identity; no opt out
1	Disclosure present but incomplete or hard to understand
2	Clear disclosure; accurate identity; voluntary, informed consent
EAS—beneficence	
0	No credible benefit
1	Plausible benefit; limited evidence
2	Evidence-informed benefit; proportional to burdens
EAS—nonmaleficence	
0	Foreseeable significant harms
1	Harms possible with partial mitigation
2	Robust mitigation (human in the loop, constrained generation, or crisis plan)
EAS—justice	
0	Exacerbates inequity or exploitation
1	Neutral or unclear
2	Equitable access; bias mitigation; accessible complaint path
EAS banding	
High	7-8
Medium	4-6
Low	0-3

[a]EAS: ethical alignment score.

Residual clinical risk is scored on a 4-level likelihood scale (rare, unlikely, possible, and likely) and a 4-level severity scale (negligible, minor, major, and catastrophic) interpreted after proposed mitigations. Ethical alignment is scored using the EAS (0-8), operationalizing autonomy, beneficence, nonmaleficence, and justice on a scale from 0 to 2 per principle and then banded as high (7-8), medium (4-6), or low (0-3). Ratings should reflect residual (not theoretical) risk, with conservative defaults and explicitly recorded uncertainty when evidence is sparse, and should be versioned for rereview after material changes.

Plotting the residual risk tier against EAS bands yields 4 deployment dispositions with explicit entry rules (Figure 1): encourage, permit with oversight, restrict or redesign, and prohibit. “Encourage” applies when residual risk is low and ethical alignment is high, supporting routine deployment with disclosure and periodic quality assurance. “Permit with

XSL•FO
RenderX

oversight" covers moderate risk with at least medium ethical alignment (or low risk with medium ethical alignment) and requires human-in-the-loop review where appropriate, audit trails, incident reporting, time-limited approvals, and a postdeployment monitoring plan with predefined triggers for rereview. "Restrict or redesign" is appropriate when residual risk is high or when ethical alignment is low in the absence of intrinsic deception (ie, the use case is not fundamentally based on impersonation, undisclosed synthetic clinicians, or manipulative identity cues); here, scope narrowing, stronger transparency and safety guardrails, and protocolized pilots are prerequisites for reconsideration. "Prohibit" is reserved for extreme risk or for low ethical alignment tied to intrinsic deception or manipulation that exploits vulnerabilities; in such cases, deployment is disallowed and takedown, reporting, and other platform- or legal-level remedies may be warranted. These thresholds aim less for numerical precision than for defensible consistency across cases and clarity about what concrete changes would move an application toward a safer, more ethically aligned quadrant.

Figure 1. Risk-and-ethics matrix for patient-facing generative artificial intelligence video in health care. Residual clinical risk (likelihood × severity after mitigations) is plotted against ethical alignment (ethical alignment score; EAS), yielding 4 deployment dispositions: encourage, permit with oversight, restrict or redesign, and prohibit. The horizontal axis represents the residual risk tier (low, moderate, high, or extreme), and the vertical axis represents EAS band (high, medium, and low), where EAS operationalizes autonomy, beneficence, nonmaleficence, and justice on a scale from 0 to 8. Entry rules prioritize residual (not theoretical) risk and link each disposition to minimum controls and rereview triggers.

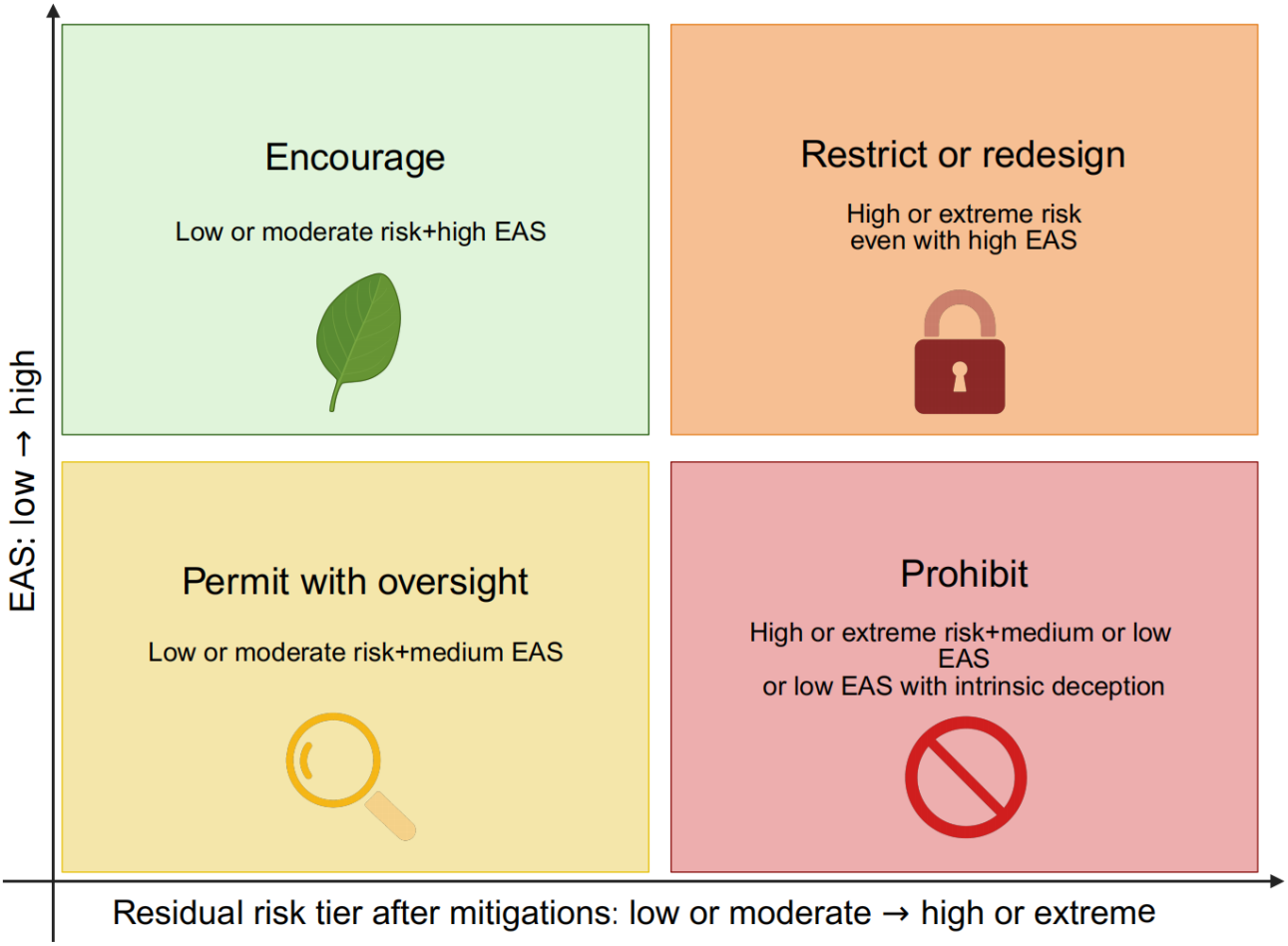

To support consistency, the rubric anchors abstract principles to concrete artifacts (eg, disclosure language, escalation pathways, evidence of benefit, and provenance controls) so that different panels can converge even when data are sparse [38,39]. Interrater reliability is promoted through independent prescoring, structured reconciliation, and written rationales for deviations from precedent. Because both risk and ethics are provisional in fast-moving sociotechnical contexts [40], institutions should version scores with date-stamped assumptions and require rereview after model updates, distribution channel changes, or sentinel events. Therefore, classification is not a verdict but a living record of judgment under stated conditions.

Finally, we specify a lightweight workflow that fits existing governance rather than creating parallel structures. Because review burden should be proportional to risk and novelty, we do not assume a single fixed evaluation time for all use cases. Low-risk, template-based, clinician-vetted educational videos that closely follow prior approved formats may undergo an expedited review focused on dossier updates, disclosure, and any material changes, whereas novel, higher-risk, psychologically sensitive, identity-based, or publicly disseminated use cases warrant fuller panel deliberation and documentation. Proponents submit a use case dossier describing purpose and audience, generation pipeline, distribution channel, mitigation plan, anticipated failure modes, and versioning or change logs. A triad panel—a clinician or health educator,

XSL•FO
RenderX

bioethicist, and safety or IT lead—scores risk and EAS independently, reconciles differences, and documents residual disagreements; panels may co-opt patient advocacy or health literacy expertise for patient-facing deployments when needed. Decisions link directly to the chosen disposition and to a postdeployment monitoring plan with predefined indicators (eg, misinformation incidents, user-reported confusion, complaint rates, follow-up burden, and equity gaps) and time-bound rereview triggers. Operational steps are summarized in Figure 2; a printable evaluator's checklist and monitoring triggers can be found in Multimedia Appendix 1, and a structured use case dossier template can be found in Multimedia Appendix 2. For institutional use, matrix-guided evaluation should be embedded into release governance such that publication or distribution through official channels requires completed dossier documentation, named sign-off, and versioned approval records. Materials that bypass review or breach minimum controls should trigger withholding of institutional dissemination; pause or takedown; incident review; and, where applicable, corrective action under local policy. In this sense, the key incentive structure is not speed-based reward. It is the linkage of authorization, traceability, and consequences to the right to deploy patient-facing AI-generated video. Where feasible, dossiers should trace content provenance from prompt to rendered asset to distribution and specify abuse-resistant defaults (eg, prohibited impersonation classes and persistent labeling), with escalation procedures when out-of-distribution use or unequal performance emerges. Table 2 shows governance categories and minimum controls.

XSL•FO
RenderX

Figure 2. Workflow from use case dossier to classification and postdeployment monitoring. Proponents submit a versioned use case dossier; a triad panel (clinician or health educator, bioethicist, and safety or IT lead) independently prescores residual risk and ethical alignment score (EAS), reconciles differences, and records a written rationale. The resulting disposition maps to minimum controls (Table 2) and a monitoring plan with predefined indicators and rereview triggers (eg, incidents, user-reported confusion, equity gaps, and version changes). Operational checklist and monitoring triggers are provided in Multimedia Appendix 1; the dossier template is provided in Multimedia Appendix 2.

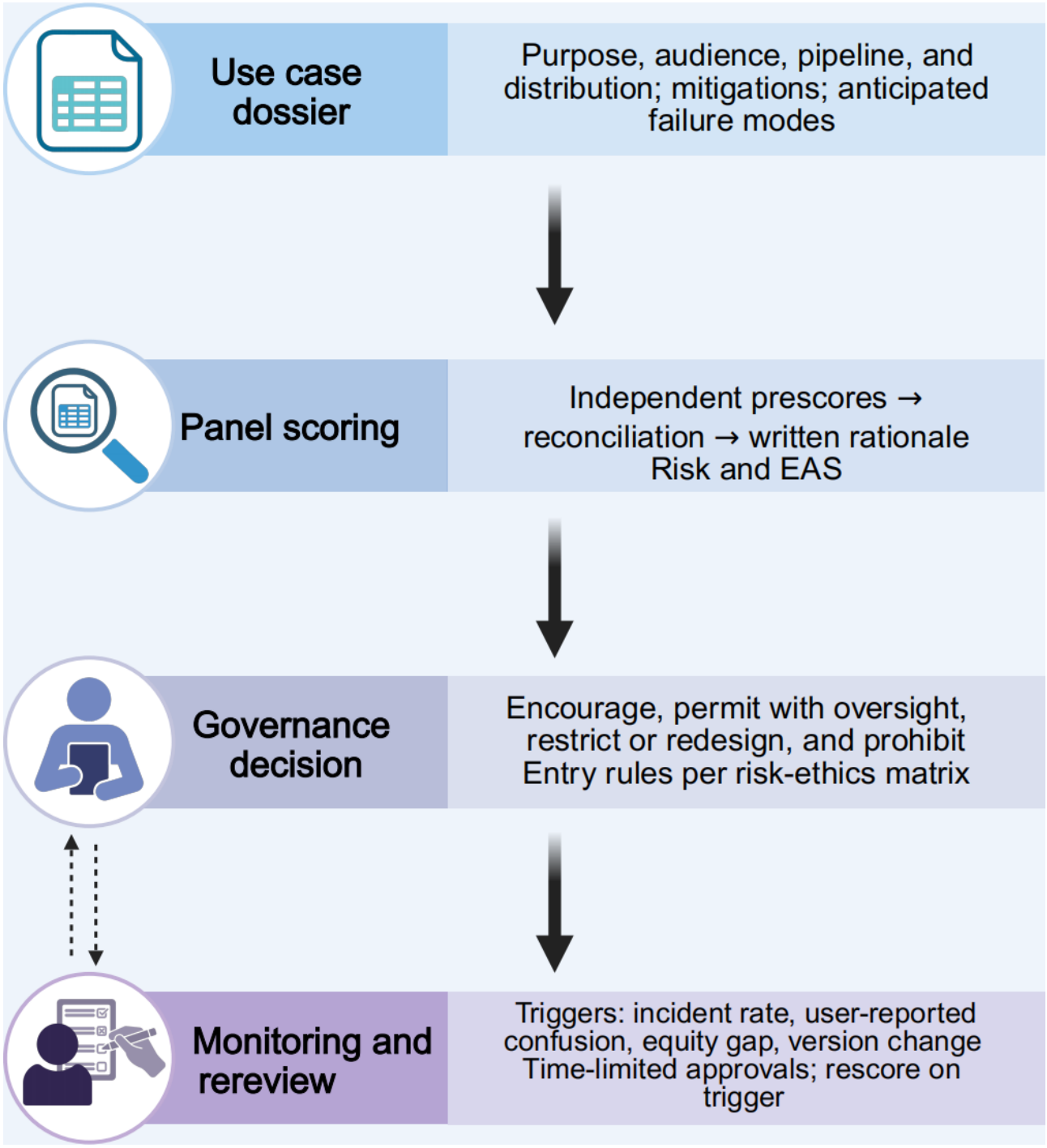

XSL•FO
RenderX

Table . Deployment dispositions, entry rules, and minimum controls.

Category	Residual risk tier × EAS[a] band	Minimum controls[b]	Illustrative examples[c]
Encourage	Low risk and high EAS	Disclosure and routine QA[d]	Transparent, clinician-vetted patient education avatar
Permit with oversight	Moderate risk and ≥medium EAS or low risk and medium EAS	Human in the loop, audit trail, incident reporting, and time-limited approval	Protocolized "deepfake therapy" in specialist or IRB[e] settings
Restrict or redesign	High risk or low EAS (no intrinsic deception)	Scope narrowing, stronger transparency and safety, and piloting then rescoring	Free-text explainer without clear disclosure
Prohibit	Extreme risk or low EAS with intrinsic deception or manipulation	Takedown and platform- or legal-level remedies	Impersonated physician endorsements

[a]EAS: ethical alignment score.

[b]Minimum controls specify documentation, human oversight, incident reporting, time-limited approval, and rereview triggers as applicable.

[c]Examples illustrate typical placements.

[d]QA: quality assurance.

[e]IRB: institutional review board.

Use Case Mapping: From Dossier Evidence to Monitoring Plans

The following use cases illustrate a repeatable mapping from dossier evidence to residual risk or EAS scoring, a deployment disposition, and a minimal monitoring plan with explicit triggers for rereview.

Case 1: Patient Education Avatar (Transparent, Vetted, and Multilingual)

A hospital deploys short, tailored videos explaining procedures and postoperative care through a clearly disclosed AI avatar delivered via an authenticated patient portal. Scripts are clinician vetted, linguistically and culturally adapted, and version controlled. The portal supports replay, adjustable playback speed, and an easy pathway to request human follow-up or report confusion. Residual risk is low: under constrained scripts plus clinical review and secure distribution, misinformation is rare or unlikely and typically minor (eg, transient confusion or extra contact), and privacy exposure is limited because the content is generic. Ethical alignment would likely be high (autonomy=2, beneficence=2, nonmaleficence=2, and justice=2): disclosure and an opt out support autonomy, measurable gains in comprehension and reduced avoidable follow-up burden support beneficence, constrained generation and review reduce foreseeable harms, and multilingual access advances justice. The resulting disposition is to encourage, with routine quality assurance and a time-bounded review cadence plus rereview triggers for guideline changes, prompt template updates, channel changes, or stratified performance gaps across language and health literacy groups [41,42].

Case 2: "Deepfake Therapy" for Grief (Therapist Led, Protocolized, and Time Limited)

Under IRB-approved protocols, a psychotherapist offers a time-bounded intervention in which a synthetic likeness of a deceased relative delivers scripted messages to facilitate goodbye rituals. A research or specialist setting matters because it enables structured screening, standardized outcome capture, adverse event reporting, and enforceable stop rules. Residual risk is moderate to high: even with careful preparation, clinically meaningful psychological harms may occur (severity level: major), and the likelihood may be possible or likely in vulnerable subgroups. Ethical alignment would likely be medium (autonomy=1, beneficence=2, nonmaleficence=1, and justice=1): beneficence may be credible for selected patients; autonomy depends on robust consent that frames the video as a simulation and checks understanding; nonmaleficence relies on screening, therapist presence, predefined discontinuation criteria, and escalation pathways; and justice requires equitable access criteria and avoidance of coercive commercialization. The resulting disposition is to permit with oversight only in research or specialist settings, with predefined outcome measures (including follow-up windows), incident logging, and immediate cessation upon adverse reactions; approvals should be time limited, with rereview triggers tied to protocol deviations, adverse events, or material changes to the model or pipeline [5,43,44].

Case 3: Impersonated Physician Endorsements (No Consent and Public Dissemination)

A synthetic clone of a prominent clinician appears on social media platforms to promote unverified health products or claims. Residual risk is extreme: harm is likely because the video exploits identity-based trust and can divert patients from evidence-based care; severity level is major to catastrophic if it prompts medication changes, delays appropriate care, or amplifies misinformation at scale. Ethical alignment would score low across all 4 principles: deception negates autonomy; benefits are not patient centered; harms are foreseeable and unmitigated; and the practice exploits vulnerable audiences, undermining justice and public trust. The resulting disposition is to prohibit. Rapid takedown and reporting workflows, provenance checks, public clarification through verified institutional channels, and legal remedies are warranted; organizations should precommit to a zero-tolerance policy for unauthorized likeness use and log incidents to strengthen postmarket monitoring and prevention [44-46].

XSL•FO
RenderX

Discussion

Implications for Digital Health Implementation

Our risk-and-ethics matrix translates life cycle AI governance expectations into a practical format for clinical decision-makers by pairing probabilistic risk appraisal with principled ethics in a way that is explicit, documentable, and revisitable. The distinctive governance value of the risk-and-ethics matrix is that it bridges the gap between high-level regulatory expectations and local operational decisions. Rather than offering another abstract set of principles, it enables institutions to classify a specific use case; document the rationale for approval or restriction; assign minimum controls; and connect deployment to monitoring, incident response, and change control over time. This orientation is consistent with major governance frameworks that emphasize postdeployment monitoring, mechanisms for capturing user input, incident response, and change management as core components of responsible AI use in real-world settings [47]. We localize these expectations to patient-facing generative AI video by grounding "risk tiering" in concrete clinical harm scenarios (eg, misinformation that changes care, identity misuse, psychological triggering content, and equity harms) and by converting principlism—autonomy, beneficence, nonmaleficence, and justice—into action-guiding criteria through an EAS. Together, the 2 axes yield implementable dispositions—encourage, permit with oversight, restrict or redesign, and prohibit—whose entry rules can be recorded, audited, and defended as part of evaluating real-world clinical performance.

Beyond classification, the matrix provides a shared structure for interdisciplinary deliberation and practical redesign. The framework also functions as a design instrument: because dossiers trace why a proposal lands in a given disposition, developers are directed toward concrete modifications—clear disclosure and comprehension-checked consent to strengthen autonomy; scope limitation, constrained generation, and human-in-the-loop review to reduce residual risk; and stratified monitoring to strengthen justice. In parallel, professional guidance on generative AI in medicine underscores the importance of preserving human oversight and aligning deployments with clinical workflows rather than displacing them—an emphasis that is especially salient for persuasive patient-facing media. Conversely, the matrix clarifies "red-line" cases grounded in intrinsic deception (eg, impersonated clinician endorsements), supporting rapid takedown, institutional clarification through verified channels, and incident logging to strengthen future prevention.

National-level safeguards become especially important in cases in which institutional incentives favor speed, visibility, or monetization over careful review. In China, the relevant governance architecture is emerging but remains distributed across health sector, platform, and generative AI rules. Existing measures already provide building blocks, including synthetic content labeling and traceability, filing and disclosure for certain public-facing AI services, and health sector expectations for account registration and monitoring. The next step is to connect these elements through sector-specific requirements for disclosure, provenance, verified identity, monitoring, rapid correction or takedown, and enforceable accountability. Such national-level guardrails would not replace local review tools such as the risk-and-ethics matrix; rather, they would create the incentive environment in which institutional review is more likely to be performed seriously and consistently.

Feasibility in smaller or resource-limited settings will depend on tiered implementation rather than assuming the full model from the outset. The core minimum is not a large committee but a documented review pathway with clearly assigned accountability, use case documentation, explicit disclosure checks, and a mechanism for escalation when risk exceeds local expertise. For familiar low-risk educational videos, institutions with limited resources may use a simplified pathway involving a clinically accountable reviewer plus a second reviewer with operational or technical oversight supported by a standardized checklist and basic postdeployment signals such as complaints, follow-up contacts, and disclosure failures. The fuller triad panel model—with dedicated bioethics input, richer analytics, formal incident reporting, and equity stratification—should be viewed as an expanded configuration for higher-risk or more mature settings. Where dedicated bioethics expertise is unavailable, regional ethics consortia, shared review pools, tele-ethics consultation, or referral pathways to larger centers may provide a practical alternative, especially for first-in-class, psychologically sensitive, identity-based, or publicly disseminated use cases.

Limitations and Validation Agenda

This approach has limitations. Early deployments will often rely on expert judgment because empirical evidence on the frequency and magnitude of novel harms remains sparse, and both residual risk estimates and EAS components can vary with local context [48-51]. To temper subjectivity, we emphasize independent prescoring, structured reconciliation, and written rationales anchored to observable artifacts (eg, disclosure language, evidence of benefit, escalation pathways, and provenance controls) [52]. Because the EAS is intended as an operational rubric rather than a purely intuitive checklist, institutions should prospectively calibrate and evaluate its reliability before routine use. A practical approach would be to begin with a set of anchor case vignettes spanning low-, medium-, and high-alignment scenarios; require independent prescoring by panel members; conduct structured reconciliation with written reasons for disagreement; and repeat calibration periodically using shared cases across panels or sites [53]. For the ratings of 0 to 2 assigned to each principle, agreement could be summarized using percentage agreement and weighted κ, whereas the reliability of the summed EAS from 0 to 8 could be examined using an intraclass correlation coefficient. Institutions could additionally track agreement on EAS banding and on the final deployment disposition because these outputs are directly tied to governance decisions. Content validity could be strengthened through multidisciplinary expert review of whether the rubric adequately captures observable manifestations of autonomy, beneficence, nonmaleficence, and justice, with iterative refinement through pilot-testing or Delphi-style consensus procedures [54]. Construct validity could then be explored by testing whether the EAS discriminates

XSL•FO
RenderX

between use cases that are expected a priori to differ in ethical alignment (eg, transparent clinician-vetted education avatars vs impersonated clinician endorsements) [55]. Importantly, classification should be treated as a living record—versioned with date-stamped assumptions—so that uncertainty becomes auditable and revisable rather than implicit [56,57]. The same use case may plausibly yield different profiles across clinical domains (eg, perioperative education vs mental health) or across resource settings; documenting contextual assumptions and applying prespecified domain modifiers can improve consistency without suppressing legitimate local variation. Finally, risk severity overlaps conceptually with nonmaleficence, and beneficence often embeds risk-benefit trade-offs [58,59]. Therefore, treating the axes as orthogonal is a usability heuristic—intended to promote clarity and reproducibility—rather than a claim of theoretical independence; cross-referencing during deliberation should be expected.

Measurement, Monitoring, and Rereview Triggers

These caveats point to a concrete real-world evaluation agenda. Retrospective incident reviews and prospective pilots can stress test thresholds and calibrate rubrics using patient-centered and workflow-relevant end points (eg, comprehension or teach back performance, unplanned follow-up contacts attributable to content, complaint and incident rates, and stratified equity gaps). Textbox 1 summarizes a core monitoring set, operational data sources, and example rereview triggers that can be embedded into routine digital health workflows. Governance-relevant measurement should be paired with life cycle mechanisms for capturing user input and adjudicating overrides and with explicit incident response and recovery pathways—elements foregrounded in the National Institute of Standards and Technology's risk management guidance for deployed AI systems [14]. Harmonization with provenance and disclosure standards can further improve auditability and reduce identity-related misuse, enabling versioned reassessments as models, prompts, distribution channels, and guardrails evolve. For systems that will undergo iterative change, "change control" should be planned rather than improvised; regulatory thinking around predetermined change control plans provides a useful template for specifying anticipated modifications and the evidence required to validate them over time. Comparative ethical analysis may also be valuable in edge cases where principlism and alternative lenses diverge; documenting such divergences can refine decision rules while preserving usability [60,61].

Textbox 1. Core monitoring metrics, operational data sources, and rereview triggers for patient-facing artificial intelligence–generated video.

Core metrics (minimum set)

- Comprehension: brief teach back–style checks or short postview questions and user-reported confusion
- Content-attributable follow-up burden: messages, calls, or telehealth follow-ups attributable to the video (eg, tagged reason codes or postview "contact clinician" clicks)
- Incidents and complaints: safety reports and formal complaints linked to the content (misinformation, identity misuse, or privacy concerns)
- Equity gaps: stratified differences in comprehension, follow-up burden, and incidents or complaints (eg, by language and health literacy proxies)
- Provenance and trust: visibility of synthetic content disclosure, verification friction (eg, "Is this my doctor?" queries), and confirmed impersonation attempts

Data sources (digital health operations)

- Patient portal and telehealth analytics (views, completion, and click-through to contact or obtain support)
- Secure messaging and call center logs (tags and reason codes and clinician note templates for attribution)
- Incident reporting and complaint systems (patient safety and privacy or security tickets)
- Patient feedback channels (postview survey and "report confusion/request human help" buttons)
- If distributed externally, verified channel monitoring (eg, takedown requests and platform reports)

Rereview triggers (examples; adapt locally)

- Material change: model, prompt or template, script or guideline, channel, or language rollout updates
- Signal excursion: sustained rise in follow-up burden or confusion reports vs baseline
- Safety event: any major incident or clustered minor incidents attributable to the content
- Equity flag: new or widening stratified gaps in end points
- Provenance or identity event: confirmed impersonation, unauthorized likeness use, or disclosure failure

In early deployments, trigger thresholds should be treated as provisional rather than fixed regulatory cutoff points. A pragmatic starting approach is to establish a local baseline during an initial pilot or first complete rollout cycle and then update thresholds iteratively as more observations accumulate. During this early phase, institutions may use structured expert consensus or Delphi-style calibration among early adopters to define provisional trigger ranges, with wider tolerance bands and explicit uncertainty notes until local rates stabilize. Thresholds should ideally be interpreted against rolling local baselines rather than in isolation and should be re-estimated after material workflow, model, channel, or language changes. For equity

XSL•FO
RenderX

analyses in particular, subgroup differences should not be overinterpreted when denominators are sparse; institutions should prespecify minimum subgroup counts before treating observed gaps as decision relevant and, where counts remain small, use descriptive flagging and continued data collection rather than strong statistical conclusions.

Here, "rereview" denotes the procedural reassessment triggered by monitoring signals, incidents, or material changes. "Reclassification" denotes a substantive change in risk tier, EAS band, or deployment disposition that may result from that reassessment. Because ethical priorities and risk tolerance vary across cultures and health systems, the framework is designed to be portable yet tunable [62,63]. To preserve comparability while allowing for local adaptation, the framework distinguishes core elements that should remain stable across sites from parameters that may be tuned to local context (Table 3).

Table . Core standardized elements and locally tunable parameters of the risk-and-ethics matrix.

Domain	Core elements (preserve across sites)[a]	Tunable parameters (adapt locally)[b]
Ethical architecture	Four principles; 0-2 EAS[c] scoring logic; high, medium, and low EAS banding	Local case anchors, examples, and training materials
Residual risk assessment	Residual risk logic after mitigation; 4-level likelihood and severity structure	Default risk modifiers for high-vulnerability domains or populations
Deployment decisions	Four disposition categories (encourage, permit with oversight, restrict or redesign, and prohibit)	Local approving body, escalation route, and implementation authority
Documentation and oversight	Use case dossier, rationale, versioning, accountability, and human oversight	Dossier format; full triad panel vs simplified or shared review pathway
Disclosure and provenance	Minimum synthetic content disclosure, identity accuracy, and provenance expectations	Disclosure wording, language level, format, and content credential implementation
Monitoring and rereview	Monitoring, incident capture, equity review, and rereview trigger logic	Indicator thresholds, baseline methods, observation windows, subgroup definitions, and language coverage minimums

[a]Core elements are intended to preserve conceptual and operational comparability across sites.

[b]Tunable parameters may be adapted to local workflow capacity, legal context, language needs, and risk tolerance provided that deviations are specified prospectively and documented consistently within the adopting institution.

[c]EAS: ethical alignment score.

Crucially, classification must remain revisitable. Approvals in "permit with oversight" should be time limited and coupled with predefined rereview triggers (eg, model or prompt updates, distribution channel changes, sentinel incidents, or emerging inequities). Successful mitigation and accumulating evidence may move a use case toward "encourage," whereas incidents or drift may push it toward "restrict or redesign" or "prohibit." Embedding this cadence operationalizes continuous risk management and aligns oversight with the broader shift toward postdeployment monitoring systems for AI in real-world settings.

The risk-and-ethics matrix is intended to complement, not replace, formal regulatory review. Its role is to translate broad regulatory expectations into case-level institutional governance for patient-facing AI-generated video use. In cross-jurisdictional terms, the framework's disclosure, identity, and provenance elements align with transparency-oriented requirements; its human-in-the-loop review, escalation pathways, and authority to pause or withdraw deployments align with expectations around human oversight; its dossier, versioning, and documented rationale align with technical documentation and record-keeping expectations; and its monitoring indicators, incident triggers, and rereview cadence align with postmarket monitoring and iterative change control requirements. This means that, where a use case is already subject to sector-specific regulation—such as the European Union AI Act's risk-based obligations for certain AI systems or medical device review pathways that incorporate predetermined change control planning—the matrix is not a substitute for those legal processes. Rather, it provides a local operational layer that helps institutions implement, document, and monitor responsible use under real-world conditions.

Conclusions and Next Steps for Implementation

AI-generated video is becoming a routine modality for patient-facing communication, making life cycle governance essential to safe real-world deployment [1,64,65]. This Viewpoint presents an operational risk-and-ethics matrix that links residual clinical risk and ethical alignment to auditable, revisitable deployment decisions.

Future work should prioritize 3 deliverables. First, real-world evaluation should calibrate thresholds with patient-centered and workflow-relevant end points (eg, comprehension and teach back, content-attributable follow-up burden, incidents and complaints, and stratified equity gaps) and embed monitoring, incident response, and change management with time-limited approvals and rereview triggers. Second, provenance should be strengthened through clear disclosure and interoperable content credentials to reduce identity misuse and support verification. Third, change control should be planned rather than improvised: institutions should prespecify anticipated model, prompt, or pipeline updates and the evidence required to maintain assurance over time, drawing on the Food and Drug Administration's predetermined change control plan approach for AI-enabled systems.

XSL•FO
RenderX

Acknowledgments

The authors used ChatGPT (OpenAI) for limited language translation and language editing during manuscript preparation. The authors reviewed and revised all such output and take full responsibility for the final manuscript.

Funding

This research was supported by grants from the National Natural Science Foundation of China (82370724), the Qingdao Key Health Discipline Development Fund, and the Qingdao Key Clinical Specialty Elite Discipline project.

Authors' Contributions

WJ and YH conceived the presented idea and developed the theoretical framework for the risk-and-ethics evaluation scaffold for patient-facing generative artificial intelligence video. YH conducted the literature synthesis, drafted the manuscript, and designed the figures. WJ supervised the project, secured funding, and provided critical revision of the manuscript for important intellectual content. Both authors discussed the concepts and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Operational tool pack (evaluator's checklist and monitoring triggers). This pack operationalizes the risk-ethics matrix for routine governance. Part A provides a 10- to 12-item evaluator's checklist. Part B lists monitoring indicators with example triggers. Crossing a trigger prompts rereview and reclassification per Table 2 and Figure 2.

[DOC File, 314 KB - jmir_v28i1e91940_app1.doc]

Multimedia Appendix 2

Use case dossier template (versioned submission form). The dossier captures purpose and audience, generation pipeline, distribution channel, content and privacy, failure modes with mitigations, ethical alignment score (EAS) evidence (0-2 per principle), residual risk tier, governance ask, monitoring plan and thresholds, sign-offs, and change log. It aligns with Table 1 (scales and EAS rubric), Figure 1 (quadrant mapping), Table 2 (entry rules and minimum controls), and Figure 2 (workflow). Panels perform independent prescores, reconciliation, and written rationales; approvals in yellow zones are time limited, with predefined triggers for rereview.

[DOC File, 236 KB - jmir_v28i1e91940_app2.doc]

References

1. Netland T, von Dzengelevski O, Tesch K, Kwasnitschka D. Comparing human-made and AI-generated teaching videos: an experimental study on learning effects. Comput Educ 2025 Jan;224:105164. [doi: 10.1016/j.compedu.2024.105164]
2. Lin J, Gu Y, Du G, et al. 2D/3D image morphing technology from traditional to modern: a survey. Inf Fusion 2025 May;117:102913. [doi: 10.1016/j.inffus.2024.102913]
3. Rovati L, Gary PJ, Cubro E, et al. Development and usability testing of a patient digital twin for critical care education: a mixed methods study. Front Med (Lausanne) 2024;10:1336897. [doi: 10.3389/fmed.2023.1336897] [Medline: 38274456]
4. Islam MZ, Wang G. Avatars in the educational metaverse. Vis Comput Ind Biomed Art 2025 Jun 10;8(1):15. [doi: 10.1186/s42492-025-00196-9] [Medline: 40493326]
5. Hoek S, Metselaar S, Ploem C, Bak M. Promising for patients or deeply disturbing? The ethical and legal aspects of deepfake therapy. J Med Ethics 2025 Jul 7;51(7):481-486. [doi: 10.1136/jme-2024-109985] [Medline: 38981659]
6. Düzgün MV, İşler A, Kazan MS. Effect of avatar-based education program in hydrocephalus on ventriculoperitoneal shunt complications and parents' knowledge and care skills: multicenter randomized controlled trials. Pediatr Neurol 2025 Aug;169:131-139. [doi: 10.1016/j.pediatrneurol.2025.05.019] [Medline: 40505427]
7. Queiroz ABL, Sartori LRM, Lima GDS, Moraes RR, Lima GS. Editorial policies for use and acknowledgment of artificial intelligence in dental journals. J Dent 2025 Oct;161:105923. [doi: 10.1016/j.jdent.2025.105923] [Medline: 40545230]
8. Wekenborg MK, Gilbert S, Kather JN. Examining human-AI interaction in real-world healthcare beyond the laboratory. NPJ Digit Med 2025 Mar 19;8(1):169. [doi: 10.1038/s41746-025-01559-5] [Medline: 40108434]
9. Eutamene HB, Hamidouche W, Keita M, Taleb-Ahmed A, Hadid A. Integrating perceptual quality analysis and caption-based features for robust deepfake video detection. Comput Electr Eng 2025 Dec;128:110699. [doi: 10.1016/j.compeleceng.2025.110699]
10. Benezeth Y, Krishnamoorthy D, Botina Monsalve DJ, Nakamura K, Gomez R, Mitéran J. Video-based heart rate estimation from challenging scenarios using synthetic video generation. Biomed Signal Process Control 2024 Oct;96:106598. [doi: 10.1016/j.bspc.2024.106598]

XSL•FO
RenderX

11. Zahedi FM, Zhao H, Sanvanson P, Walia N, Jain H, Shaker R. My real avatar has a doctor appointment in the wepital: a system for persistent, efficient, and ubiquitous medical care. Inf Manag 2022 Dec;59(8):103706. [doi: 10.1016/j.im.2022.103706]
12. Sestino A, D'Angelo A. My doctor is an avatar! The effect of anthropomorphism and emotional receptivity on individuals' intention to use digital-based healthcare services. Technol Forecast Soc Change 2023 Jun;191:122505. [doi: 10.1016/j.techfore.2023.122505]
13. Decety J, Li J. The value of empathy in medical practice: a neurobehavioral perspective. Soc Sci Humanit Open 2025;12:101956. [doi: 10.1016/j.ssaho.2025.101956]
14. Tabassi E. Artificial intelligence risk management framework (AI RMF 1.0). : National Institute of Standards and Technology; 2023 URL: https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10 [accessed 2026-04-27]
15. Masayoshi K, Katada Y, Ozawa N, Ibuki M, Negishi K, Kurihara T. Deep learning segmentation of non-perfusion area from color fundus images and AI-generated fluorescein angiography. Sci Rep 2024 May 11;14(1):10801. [doi: 10.1038/s41598-024-61561-x] [Medline: 38734727]
16. Piot MA, Attoe C, Billon G, Cross S, Rethans JJ, Falissard B. Simulation training in psychiatry for medical education: a review. Front Psychiatry 2021;12:658967. [doi: 10.3389/fpsyt.2021.658967] [Medline: 34093275]
17. Zhang J, Ding Y, Zhang H, et al. An experimental study on embodiment forms and interaction modes in affective robots for anxiety relief and emotional connection. Int J Hum Comput Stud 2025 Sep;203:103584. [doi: 10.1016/j.ijhcs.2025.103584]
18. Soto-Sanfiel MT, Wu Q. How audiences make sense of deepfake resurrections: a multilevel analysis of realism, ethics, and cultural meaning. Comput Hum Behav 2026 Jan;174:108822. [doi: 10.1016/j.chb.2025.108822]
19. Ma'arif A, Maghfiroh H, et al. Social, legal, and ethical implications of AI-generated deepfake pornography on digital platforms: a systematic literature review. Soc Sci Humanit Open 2025;12:101882. [doi: 10.1016/j.ssaho.2025.101882]
20. Wagner R, Pardi G, Müller J, Brucker B, Schwarzer S, Gerjets P. Listening to scientists in immersive videos: how levels of immersion and points of view influence learning experiences. Comput Educ 2025;234:1-19. [doi: 10.1016/j.compedu.2025.105326]
21. Bujić M, Salminen M, Hamari J. Effects of immersive media on emotion and memory: an experiment comparing article, 360-video, and virtual reality. Int J Hum Comput Stud 2023 Nov;179:103118. [doi: 10.1016/j.ijhcs.2023.103118]
22. Fridman I, Bylund CL, Elston Lafata J. Trust of social media content and risk of making misinformed decisions: survey of people affected by cancer and their caregivers. PEC Innov 2024;5:100332. [doi: 10.1016/j.pecinn.2024.100332] [Medline: 39323933]
23. Diwanji VS. Should your brand hire virtual influencers? How realism and gender presentation shape trust and purchase intentions. J Retail Consum Serv 2026 Jan;88:104491. [doi: 10.1016/j.jretconser.2025.104491]
24. Letafati M, Otoum S. On the privacy and security for e-health services in the metaverse: an overview. Ad Hoc Netw 2023 Nov;150:103262. [doi: 10.1016/j.adhoc.2023.103262]
25. Pirhofer J, Bükki J, Vaismoradi M, Glarcher M, Paal P. A qualitative exploration of cultural safety in nursing from the perspectives of Advanced Practice Nurses: meaning, barriers, and prospects. BMC Nurs 2022 Jul 4;21(1):178. [doi: 10.1186/s12912-022-00960-9] [Medline: 35787799]
26. Sterponi L, Fatigante M, Zucchermaglio C, Alby F. Companions in immigrant oncology visits: uncovering social dynamics through the lens of Goffman's footing and Conversation Analysis. SSM Qual Res Health 2024 Jun;5:100432. [doi: 10.1016/j.ssmqr.2024.100432]
27. Krijger J. What about justice and power imbalances? A relational approach to ethical risk assessments for AI. DISO 2024;3:56. [doi: 10.1007/s44206-024-00139-6]
28. Ploug T, Jørgensen RF, Motzfeldt HM, Ploug N, Holm S. The need for patient rights in AI-driven healthcare - risk-based regulation is not enough. J R Soc Med 2025 Aug;118(8):248-252. [doi: 10.1177/01410768251344707] [Medline: 40562393]
29. Baard P, Sandin P. Principlism and citizen science: the possibilities and limitations of principlism for guiding responsible citizen science conduct. Res Ethics 2022;18(4):304-318. [doi: 10.1177/17470161221116558]
30. Clouser KD, Gert B. A critique of principlism. J Med Philos 1990 Apr;15(2):219-236. [doi: 10.1093/jmp/15.2.219] [Medline: 2351895]
31. Mittelstadt B. Principles alone cannot guarantee ethical AI. Nat Mach Intell 2019;1(11):501-507. [doi: 10.1038/s42256-019-0114-4]
32. Whittaker L, Kietzmann J, Letheren K, Mulcahy R, Russell-Bennett R. Brace yourself! Why managers should adopt a synthetic media incident response playbook in an age of falsity and synthetic media. Bus Horiz 2023;66(2):277-290. [doi: 10.1016/j.bushor.2022.07.004]
33. Mylrea M. The generative AI weapon of mass destruction: evolving disinformation threats, vulnerabilities, and mitigation frameworks. In: Lawless W, Mittu R, Sofge D, Fouad H, editors. Interdependent Human-Machine Teams: The Path to Autonomy: Academic Press; 2024:315-347. [doi: 10.1016/B978-0-443-29246-0.00007-9]
34. Omrani N, Rivieccio G, Fiore U, Schiavone F, Agreda SG. To trust or not to trust? An assessment of trust in AI-based systems: concerns, ethics and contexts. Technol Forecast Soc Change 2022 Aug;181(2):121763. [doi: 10.1016/j.techfore.2022.121763]

XSL•FO
RenderX

35. Liao SM, Haykel I, Cheung K, Matalon T. Navigating the complexities of AI and digital governance: the 5W1H framework. J Responsible Technol 2025 Sep;23:100127. [doi: 10.1016/j.jrt.2025.100127]
36. Braga CM, Serrano MA, Fernández-Medina E. Towards a methodology for ethical artificial intelligence system development: a necessary trustworthiness taxonomy. Expert Syst Appl 2025 Aug;286:128034. [doi: 10.1016/j.eswa.2025.128034]
37. Pettersson M, Hedström M, Höglund AT. The ethics of DNR-decisions in oncology and hematology care: a qualitative study. BMC Med Ethics 2020 Jul 31;21(1):66. [doi: 10.1186/s12910-020-00508-z] [Medline: 32736556]
38. Lanerolle G, Roberts ES, Haroon A, Shetty A. Introduction. In: Quality Assurance Management: A Comprehensive Overview of Real-World Applications for High Risk Specialties: Elsevier Science; 2024:1-21.
39. Li S, Wang Z, Shang Y, et al. Developing federated time-to-event scores using heterogeneous real-world survival data. Comput Biol Med 2025 Oct;197(Pt B):111084. [doi: 10.1016/j.compbiomed.2025.111084] [Medline: 40976210]
40. Patriarca R, Falegnami A, Costantino F, Di Gravio G, De Nicola A, Villani ML. WAx: an integrated conceptual framework for the analysis of cyber-socio-technical systems. Saf Sci 2021 Apr;136:105142. [doi: 10.1016/j.ssci.2020.105142]
41. Badawy MK, Khamwan K, Carrion D. A pilot study of generative AI video for patient communication in radiology and nuclear medicine. Health Technol 2025 Mar;15:395-404. [doi: 10.1007/s12553-025-00945-z]
42. Adeboye W, Tayal V, Odubanjo E, et al. Artificial intelligence in the delivery of patient care: avatar-generated videos for patient education post breast surgery. Eur J Surg Oncol 2024 May;50:108076. [doi: 10.1016/j.ejso.2024.108076]
43. Emotional documentary explores new compassionate possibilities of VR. Unreal Engine. 2020. URL: https://www.unrealengine.com/developer-interviews/emotional-documentary-explores-new-compassionate-possibilities-of-vr?lang=fr [accessed 2025-10-17]
44. Pizzoli SF, Monzani D, Vergani L, Sanchini V, Mazzocco K. From virtual to real healing: a critical overview of the therapeutic use of virtual reality to cope with mourning. Curr Psychol 2023;42(11):8697-8704. [doi: 10.1007/s12144-021-02158-9] [Medline: 34429574]
45. Liu T, Wang P, Pan D, Liu R. Credibility of AI generated and human video doctors and the relationship to social media use. Front Public Health 2025;13:1559378. [doi: 10.3389/fpubh.2025.1559378]
46. Stokel-Walker C. Deepfakes and doctors: how people are being fooled by social media scams. BMJ 2024 Jul 17;386:q1319. [doi: 10.1136/bmj.q1319] [Medline: 39019557]
47. Nordling L. Scientists are falling victim to deepfake AI video scams - here's how to fight back. Nature New Biol 2024 Aug 7. [doi: 10.1038/d41586-024-02521-3] [Medline: 39112581]
48. Oehmen J, Locatelli G, Wied M, Willumsen P. Risk, uncertainty, ignorance and myopia: Their managerial implications for B2B firms. Industrial Marketing Management 2020 Jul;88:330-338. [doi: 10.1016/j.indmarman.2020.05.018]
49. Singh S, Dhumane A. Unmasking digital deceptions: an integrative review of deepfake detection, multimedia forensics, and cybersecurity challenges. MethodsX 2025;15:103632. [doi: 10.1016/j.mex.2025.103632] [Medline: 41080432]
50. McAndrew T, Reich NG. An expert judgment model to predict early stages of the COVID-19 pandemic in the United States. PLoS Comput Biol 2022 Sep;18(9):e1010485. [doi: 10.1371/journal.pcbi.1010485] [Medline: 36149916]
51. Awodi NJ, Liu YK, Ayodeji A, Adibeli JO. Expert judgement-based risk factor identification and analysis for an effective nuclear decommissioning risk assessment modeling. Prog Nucl Energy 2021 Jun;136:103733. [doi: 10.1016/j.pnucene.2021.103733]
52. Maltby KM, Howes E, Lincoln S, et al. Marine climate change risks to biodiversity and society in the ROPME Sea Area. Clim Risk Manag 2022;35(2121):100411. [doi: 10.1016/j.crm.2022.100411]
53. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor Quant Methods Psychol 2012;8(1):23-34. [doi: 10.20982/tqmp.08.1.p023] [Medline: 22833776]
54. Polit DF, Beck CT. The content validity index: are you sure you know what's being reported? Critique and recommendations. Res Nurs Health 2006 Oct;29(5):489-497. [doi: 10.1002/nur.20147] [Medline: 16977646]
55. Boateng GO, Neilands TB, Frongillo EA, Melgar-Quiñonez HR, Young SL. Best practices for developing and validating scales for health, social, and behavioral research: a primer. Front Public Health 2018;6:149. [doi: 10.3389/fpubh.2018.00149] [Medline: 29942800]
56. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). NPJ Digit Med 2024 Jul 8;7:183. [doi: 10.1038/s41746-024-01157-x] [Medline: 38977771]
57. Lekens AL, Drageset S, Hansen BS. Knowing how, arguing why: nurse anaesthetists' experiences of nursing when caring for the surgical patient. BMC Nurs 2025 Feb 7;24(1):144. [doi: 10.1186/s12912-025-02752-3] [Medline: 39920699]
58. Jansen SN, Kamphorst BA, Mulder BC, et al. Ethics of early detection of disease risk factors: a scoping review. BMC Med Ethics 2024 Mar 5;25(1):25. [doi: 10.1186/s12910-024-01012-4] [Medline: 38443930]
59. Pathni RK. Beyond algorithms: ethical implications of AI in healthcare. Med J Armed Forces India 2025;81(6):630-636. [doi: 10.1016/j.mjafi.2024.10.014] [Medline: 41268011]
60. Rauprich O. Principlism. In: Chadwick R, editor. Encyclopedia of Applied Ethics: Academic Press; 2012:590-598.
61. Bello P, Bridewell W. Self-control on the path toward artificial moral agency. Cogn Syst Res 2025 Jan;89:101316. [doi: 10.1016/j.cogsys.2024.101316]

XSL•FO
RenderX

62. Lysaght T, Chan HY, Scheibner J, Toh HJ, Richards B. An ethical code for collecting, using and transferring sensitive health data: outcomes of a modified Policy Delphi process in Singapore. BMC Med Ethics 2023 Oct 4;24(1):78. [doi: 10.1186/s12910-023-00952-7] [Medline: 37794387]
63. Guenduez AA, Walker N, Demircioglu MA. Digital ethics: global trends and divergent paths. Gov Inf Q 2025 Sep;42(3):102050. [doi: 10.1016/j.giq.2025.102050]
64. Vehi J, Mujahid O, Beneyto A, Contreras I. Generative artificial intelligence in diabetes healthcare. iScience 2025 Aug 15;28(8):113051. [doi: 10.1016/j.isci.2025.113051] [Medline: 40703444]
65. Crowe B, Shah S, Teng D, et al. Recommendations for clinicians, technologists, and healthcare organizations on the use of generative artificial intelligence in medicine: a position statement from the Society of General Internal Medicine. J Gen Intern Med 2025 Feb;40(3):694-702. [doi: 10.1007/s11606-024-09102-0] [Medline: 39531100]

Abbreviations

AI: artificial intelligence
EAS: ethical alignment score
IRB: institutional review board

Edited by A Mavragani; submitted 22.Jan.2026; peer-reviewed by A Shamsi, Z Su; revised version received 14.Apr.2026; accepted 15.Apr.2026; published 08.May.2026.

Please cite as:
Hu Y, Jiang W
Governing Patient-Facing AI-Generated Video in Digital Health: A Risk-and-Ethics Matrix for Deployment, Monitoring, and Change Control
J Med Internet Res 2026;28:e91940
URL: https://www.jmir.org/2026/1/e91940
doi:10.2196/91940

© Yongzheng Hu, Wei Jiang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Actionable Innovation Day Approach: Participatory Model for Advancing Critical Care Innovation

Brett N Hryciw[1]; Cecilia Tran[1]; Rashi Ramchandani[1,2]; Cameron Love[1]; Christine Caron[1]; Aimee Sarti[1]; Annelise Miller[3]; Suzanne Madore[1]; Michael Chasse[4]; Andy Pan[5,6,7]; Simon Didcote[8]; Scott Millington[1,9]; Heather Galley[1,9]; Kwadwo Kyeremanteng[9,10,11]; Andrew Seely[10,12,13]

[1]Department of Medicine, University of Ottawa, Ottawa, ON, Canada
[10]Clinical Epidemiology Program, Ottawa Hospital Research Institute, 451 Smyth Road, Ottawa, ON, Canada
[11]Division of Palliative Medicine, Ottawa Hospital, Ottawa, Canada
[12]Department of Thoracic Surgery, University of Ottawa, Ottawa, ON, Canada
[13]School of Epidemiology & Public Health, University of Ottawa, Ottawa, ON, Canada
[2]Institute of Health Policy, Evaluation and Management, University of Toronto, Toronto, ON, Canada
[3]Department of Family Medicine, University of Ottawa, Ottawa, ON, Canada
[4]Department of Intensive Care, Université de Montréal, Montreal, Canada
[5]Department of Emergency Medicine, The Ottawa Hospital, Ottawa, ON, Canada
[6]Division of Critical Care Medicine, Montfort Hospital, Ottawa, ON, Canada
[7]Ornge (Transport Medicine), Mississauga, ON, Canada
[8]OBS Medical (United Kingdom), Gloucestershire, United Kingdom
[9]Department of Critical Care, Ottawa Hospital, Ottawa, ON, Canada

Corresponding Author:
Andrew Seely
Clinical Epidemiology Program, Ottawa Hospital Research Institute, 451 Smyth Road, Ottawa, ON, Canada

Abstract

Health care innovation is essential for improving patient outcomes, enhancing system efficiency, and preparing for future challenges; however, meaningful progress is often hindered by entrenched barriers such as resistance to change, fragmented interdisciplinary collaboration, and constrained financial and human resources. These persistent obstacles make it difficult for health care systems to translate creative ideas into sustainable, real-world improvements, underscoring the need for structured approaches that support collaboration and reduce implementation friction. To address these challenges, we developed the Actionable Innovation Day (AID) approach, a structured, participatory model designed to generate consensus-based, low-cost recommendations that are feasible for system improvement. The first regional AID event in Eastern Ontario gathered 57 multidisciplinary participants, including clinicians, administrators, patient partners, and industry leaders, for a full-day series of presentations, facilitated discussions, and targeted breakout sessions focused on critical care. Through guided deliberation and collaborative analysis, participants synthesized diverse perspectives into a prioritized set of improvement opportunities. The process yielded 28 actionable recommendations across 4 domains: health care innovation, regionalized care, critical care practices, and the use of artificial intelligence. A postevent survey (86% response rate) showed strong agreement, with 23 recommendations rated above 4 on a 5-point scale. The highest-ranked proposals emphasized the value of strengthening research-industry-clinical partnerships, integrating families more intentionally into intensive care unit rehabilitation and recovery processes, and implementing centralized regional coordination to optimize critical care capacity. Together, these findings illustrate not only the feasibility of the AID model but also the AID model's ability to surface strategic, context-appropriate solutions that resonate across stakeholder groups. The AID process offers a scalable and adaptable template for advancing health care innovation through collaborative, real-world problem-solving. While this initial event focused on critical care, the underlying principles of structured engagement, iterative consensus building, and interdisciplinary co-design are broadly applicable to many sectors of health care. We encourage institutions, regional networks, and health system leaders to adopt and tailor the AID framework to their own local priorities, recognizing that inclusive innovation processes can accelerate system improvement even in resource-limited settings. Ultimately, the AID approach serves as both a methodology and a call to action: by empowering teams to collectively identify, refine, and champion actionable ideas, health care organizations can build the capacity and culture necessary to drive meaningful and sustained innovation across diverse clinical and operational domains.

(J Med Internet Res 2026;28:e73614) doi:10.2196/73614

XSL•FO
RenderX

KEYWORDS

innovation; critical care; interdisciplinary collaboration; patient-centered care; artificial intelligence; AI

The Current Landscape of Health Care Innovation

Innovation is foundational to the advancement of any discipline, particularly health care. It encompasses technological advancements, process improvements, and cultural and organizational shifts, all of which enhance the quality, efficiency, and accessibility of health care services [1]. As health care systems face challenges related to resource limitations and increased demand, leading to burnout among health care providers, innovative thinking becomes essential to ensure sustainability and progress [2,3].

In health care, innovation acts as a catalyst for improving patient outcomes, optimizing workflows, and enhancing patient experiences [4]. The acute care system in Canada has acknowledged the need to address overcrowded emergency departments, prolonged wait times, and health care provider burnout [3,5,6]. These issues underscore the need for practical innovation to improve both patient care and operational efficiency. However, health care innovation often faces significant barriers, including organizational resistance to change, the complexities of evaluating multiple stakeholder opinions for idea validation, and resource limitations for implementation [7].

The Actionable Innovation Day Approach as a Catalyst for Innovation

While innovation workshops and ideation events are increasingly used in health care to identify areas for health care innovation, they often lack structured follow-up, fail to produce actionable outputs, or exclude critical stakeholder groups. The Actionable Innovation Day (AID) approach was designed to address these gaps by combining multidisciplinary participation, a mandate for concrete recommendations, and consensus-based validation, all within a scalable, low-cost framework. The AID framework provides a structured model to bring together diverse health care stakeholders to propose actionable changes within a local health authority. The AID approach was conceptually adapted from the Agile Innovation Framework (AIF), which emphasizes iterative, stakeholder-driven development cycles. By embedding this philosophy into a single-day participatory format, the AID approach aimed to translate these principles into a practical model for health care innovation.

The regional AID event aimed to establish a shared understanding of health care priorities and develop practical solutions that can be scaled across regions to drive health care transformation. This paper describes the AID approach through its first implementation, which focused on acute and critical care within a regional health system. Critical care was chosen as the focus due to ongoing challenges related to coordination, resource constraints, and patient outcomes, making it a compelling context in which to evaluate the AID model in practice.

The first regional AID event was implemented with a focus on acute and critical care, engaging stakeholders across the Champlain Local Health Integration Network to address challenges specific to intensive care unit (ICU) and emergency care delivery. Stakeholders invited included physicians, nurses, respiratory therapists, and administrators from hospitals, as well as patient partners, government representatives, and industry participants, thereby allowing for the representation of broad multidisciplinary perspectives. The event was held both in person and virtually to mitigate geographical barriers to participation. It focused on coordinating expedient patient transfers when necessary to provide timely access to appropriate levels of care while minimizing unwarranted risks and expenses of unnecessary patient transfers. Table 1 describes the agenda for the regional AID framework.

Table . The regional Actionable Innovation Day event agenda.

Time	Session
7 AM-8:15 AM	Coffee and registration
8:15 AM-8:30 AM	Welcome address
8:30 AM-10:00 AM	Speaker presentations: Future Innovation
10 AM-10:30 AM	Networking
10:30 AM-noon	Speaker presentations: Present Innovation
Noon-1 PM	Networking and lunch
1 PM-2:15 PM	Small breakout group discussion: Innovation in Critical Care Regionalization
2:15 PM-2:30 PM	Networking and coffee break
2:30 PM-3:45 PM	Summary of small group discussions
3:45 PM-4:45 PM	Summary and rating of recommendations
4:45 PM-5:00 PM	Closing remarks

XSL•FO
RenderX

The planning committee, composed of representatives from the Ottawa Hospital, the Ottawa Hospital Research Institute, and affiliated academic and industry partners, designed the AID event to foster multidisciplinary collaboration around regional critical care innovation. Speakers were identified through a consultative process that sought representation from key stakeholder groups, including physicians, nurses, respiratory therapists, allied health professionals, hospital administrators, patient partners, and industry leaders. Each invited speaker was asked to deliver a concise presentation, highlighting a current challenge within their domain and concluding with 1 to 3 actionable recommendations for system improvement. The day was structured into two 90-minute presentation blocks, each followed by a brief discussion period to clarify ideas and prepare for the afternoon breakout sessions focused on translating these recommendations into practical solutions. All recommendations were systematically documented by designated recorders (CT and RR) during the AID. Following the morning presentations, participants joined one of three 90-minute small group breakout sessions in the afternoon to explore recommendations to address the key theme of "How to optimize regional critical care delivery?" Participants were randomly assigned to breakout groups, each consisting of 15 people. Moderators facilitated these sessions, helping attendees identify actionable recommendations and items related to the key theme. Recommendations from the breakout sessions were then presented by a small group speaker for a further whole-group discussion and organized into consensus summary recommendations.

At the end of the day, all recommendations were summarized. A follow-up survey using Mentimeter for immediate audience polling allowed attendees to anonymously rate their level of agreement with the suggested recommendations using a 5-point Likert scale. Items with high agreement were prioritized for future implementation.

The AID Approach in Action: Attendance and Impact

Participant Demographics

In total, 110 stakeholders were identified and invited to attend the AID event. The attendance rate was 52%, as 57 participants, including invited speakers, attended the regional AID event in person. An additional 10 participants attended the event virtually through Microsoft Teams. Of the 67 people with whom the survey was shared, responses were received from 49 (73%) individuals. Respondents ranged in age from 23 to 65 years, with a median age of 47 (IQR 37 - 55) years. The participants included 6 (12.2%) physicians, 2 (4.1%) registered nurses, 1 (2.0%) respiratory therapist, 17 (34.7%) industry partners, 4 (8.2%) allied health care providers, 7 (14.3%) research staff, 5 (10.2%) administrative staff, 5 (10.2%) leadership staff, and 2 (4.1%) patient partners. Gender distribution was 42.9% (n=21) male and 57.1% (n=28) female, and 81.6% (n=40) identified as White. Experience levels ranged from individuals in training to those with over 25 years of experience (Table 2).

XSL•FO
RenderX

Table . Summary of demographic characteristics of regional Actionable Innovation Day survey respondents (n=49).

Demographic variables	Survey respondents, n (%)
Gender	
Male	21 (42.9)
Female	28 (57.1)
Age group (y)	
<30	4 (8.2)
30 - 39	11 (22.4)
40 - 49	15 (30.6)
50 - 59	13 (26.5)
>60	4 (8.2)
No response	2 (4.1)
Years of experience	
In training	2 (4.1)
0 - 5	4 (8.2)
5 - 10	4 (8.2)
11 - 15	6 (12.2)
16 - 20	9 (18.4)
21 - 25	13 (26.5)
>25	11 (22.4)
Ethnicity	
White	40 (81.6)
South Asian	3 (6.1)
East or Southeast Asian	4 (8.2)
Black	1 (2.0)
Other	1 (2.0)
Occupations	
Physician	6 (12.2)
Registered nurse	2 (4.1)
Respiratory therapist	1 (2.0)
Allied health care	4 (8.2)
Industry partners	17 (34.7)
Research staff	7 (14.3)
Administration staff	5 (10.2)
Leadership staff	5 (10.2)
Patient partners	2 (4.1)

Applying the AID Approach Into Critical Care Innovation

Overview

The event featured 16 presentations from various professional backgrounds, which covered future innovations in health care, patient-centered care, critical care advancements, nursing and respiratory therapy improvements, and innovations in urban and resource-limited settings. Topics also included the integration of AI, predictive monitoring, effective interhospital transfers, social work innovations, and enhanced partnerships between academia, industry, and health care. Presenters focused on using existing resources to enact meaningful change within the Champlain Local Health Integration Network, and their recommendations were recorded (Table 3).

XSL•FO
RenderX

Table . Summary of key points and actionable recommendations from presentations of health care professionals of innovation in health care.

Speaker	Title	Recommendation summary
Cameron Love (Hospital CEO)	Future of innovation at TOH[a]	1. Communicate and instill a sense of pride and importance of discovery within your teams. 2. Create strong program structure and process that incorporates discovery into everything we do and talk about. 3. Identify 3 innovation priorities annually to move forward with.
Kwadwo Kyeremanteng (Chair, Department of Critical Care)	Future of innovation in critical care	1. Pilot program for wearable monitors in hospitals. 2. Pilot of adopting AI[b]-based early warning system.
Christine Caron (patient representative)	Patient-centered innovation	1. Alleviate delirium and PTSD[c] in critical care patients through: • ICU[d] diary (communicates discharge management with diagrams about steps or risks). • sounds (headphone and sunlamp): headphone Halo from patients to be protected from sounds in ICU (spa music for physiotherapy, brown noise, and pink noise). • physical therapy: teach families OT[e] or PT[f] exercises to perform during their ICU stay to mitigate physical impairments or disabilities (eg, ICU cycle and yoga bands).
Annelise Miller (ER MD)	Innovation in resource-limited community or rural ED[g]	1. Understand the context (who/what/where/why/how) with your community partners. • Clearly define providers, materials, time, geographic distance, and physical space. 2. Foster innovation (support local teams or local management, ensure it facilitates timely action, and builds interinstitutional relationship). 3. Innovation should occur alongside advocacy.
Suzanne Madore (Hospital COO and VP)	Innovation in TOH nursing	1. Foster a culture of innovation with support from leadership. 2. Embed innovation into strategy and comprehend the essence of innovation. 3. Recognize and reward innovation.
Lauralynn McIntyre (ICU Department Research Lead)	Research innovation	1. Monitor and aim to increase the number of patients enrolled in critical care studies at TOH. 2. Promotion of critical care research at TOH. 3. Training to develop future innovation researchers. a. Early-career investigator support for innovation researchers.
Matt Bromwich (surgeon, CEO, Hospital Innovation Lead)	Innovation in scientific academic industry partnerships	1. Innovation should be subtractive (one solution replaces 9 other tasks). 2. Have a team around you (to speak the language and fill in the gaps).

XSL•FO
RenderX

Speaker	Title	Recommendation summary
Pierre Cardinal (intensivist, educator)	Innovation in documentation and education	1. After innovation work and findings on free forums such as Google Firebase Community.
Frank Fiorenza (respiratory therapist, CEO)	Innovation in respiratory therapy	1. Review mechanical ventilation policy and add a minimum of 1 new solution annually. 2. Identify if ventilation protocols are lung protective 3. Review critical patient transport protocols and add a minimum of 1 solution annually. 4. Do you have an idea for a new device or way to improve patient care—bring it forward. 5. Have you been presented with an innovation recently? Is it viable at your facility? Take on a minimum of 1 new project annually—be the champion.
Michael Chasse (intensivist scientist)	Where is AI in critical care	1. AI tools should learn to use data "as is": Tools should use unstandardized data inputs to improve their adaptability and effectiveness in real-world settings. 2. Health care practitioners should prepare for the inevitable partial replacement by AI in specific tasks, embracing the shift as part of the ongoing evolution of medical practice. 3. Anticipate and decide on nonreplaceable tasks: identify and prioritize the tasks that humans should retain, even if AI could perform them more efficiently and/or better.
Andrew Seely (surgeon, intensivist, scientist, CEO)	Monitoring-based predictive decision support	1. Implement extubation advisor at TOH in a multidisciplinary quality improvement project focused on reducing extubation failure. • Continue to seek to get more value from monitoring, by the implementation and evaluation of predictive clinical decision support tools.
Scott Millington (intensivist, VCC[h] Lead)	Virtual critical care	1. Hub-and-spoke model for other specialties that deals with any acute patients. 2. Use VCC workflow and technology to take responsibility for region.
Heather Galley (Critical Care SW[i])	Innovation in social work	1. Admission checklist → automatic social work referral. 2. Review Critical Care Family Assistance Program criteria and implement updates.
Alexandre Tran (surgeon, intensivist, scientist)	Predicting and preventing ward deterioration	1. Improve data capture and analysis to personalize the prediction of critical events. 2. Capitalize on the human element to maximize technology effectiveness.
Andy Pan (intensivist, regional transport lead)	Improving interhospital critical care transfers	1. Involve the transport physician early on and determine the best means of transfer. 2. Pretransfer optimization and minimize transition points.
Andrew Kramer (data scientist) and Simon Didcote (developer, CEO)	Continuous predictive monitoring	

XSL•FO
RenderX

Speaker	Title	Recommendation summary
		1. Implement continuous noninvasive monitoring outside ICU combined with suitable regulated analytics such as VSI[j]—especially for high-risk patients. 2. Consider investigating use of regulated EWS[k] such as ViSIG in ICU to improve outcomes. 3. Use a device-agnostic platform to tie analytics together to avoid vendor lock-in.

[a]TOH: The Ottawa Hospital.
[b]AI: artificial intelligence.
[c]PTSD: posttraumatic stress disorder.
[d]ICU: intensive care unit.
[e]OT: occupational therapy.
[f]PT: physical therapy.
[g]ED: emergency department.
[h]VCC: virtual critical care.
[i]SW: social work.
[j]VSI: Visensia Safety Index.
[k]EWS: early warning system.

Small group discussions then provided a multidisciplinary platform for further exploring these as well as additional solutions to improve the efficiency and effectiveness of regionalized acute care. In total, 4 themes emerged: (1) coordination and communication, (2) resource management, (3) education, and (4) family and patient engagement (Figure 1). Coordination and communication focused on establishing centralized critical care coordination, real-time data dashboards, and repatriation protocols for optimizing patient transfers. Resource management emphasized ICU utilization, severity, predictive monitoring pilots, and ensuring adequate community resources. Education was highlighted as vital for fostering collaboration among stakeholders, including community members, academic institutions, and health care providers. Family and patient engagement included strategies such as promoting journaling and home-based rehabilitation to enhance patient recovery. All recommendations from the presentations and small group sessions were then synthesized into a comprehensive list of recommendations generated from the event.

Figure 1. Summary of actionable recommendations to optimize regionalization of critical care for a local health integration network in Ontario, Canada, generated from small group discussions organized by key themes. ICU: intensive care unit; VCC: virtual critical care.

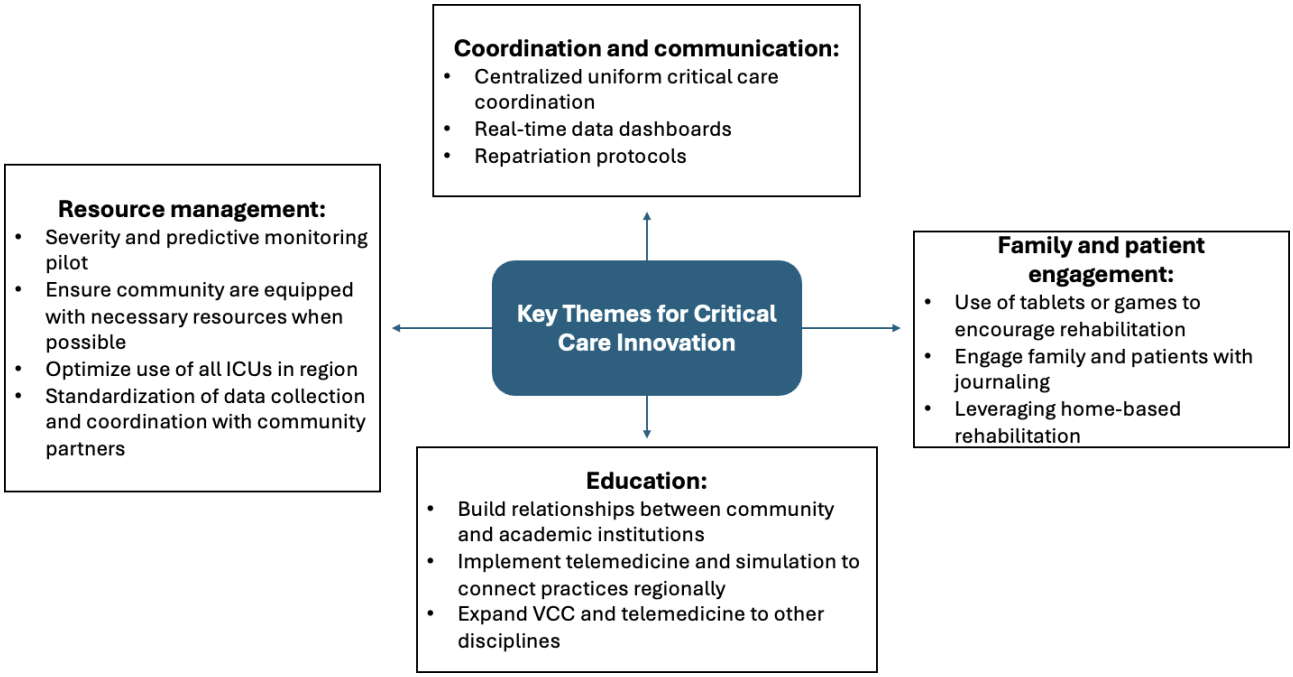

XSL•FO
RenderX

Actionable Recommendations

A total of 16 speakers provided a total of 42 actionable recommendations during individual presentations on the morning of the AID event, followed by discussion and questions after each. In addition, a set of afternoon parallel small group sessions came up with recommendations to address improved regional coordination of critical care. Subsequently, a total of 28 actionable recommendations emerged from a synthesis of the individual presenters and discussion by participants during the day (Multimedia Appendix 1), as prepared by 2 independent observers (CT and RR) present and recording all events of the day. They summarized the recommendations and discussion into four key domains: (1) innovation in health care, (2) innovation in regionalized care, (3) critical care practices, and (4) use of AI in health care. No preexisting framework was used for classification. A follow-up email survey sent to all participants showed broad support for these recommendations, with agreement ratings ranging from 3.66 to 4.50 on a 5-point Likert scale (Multimedia Appendix 1). Of the 28 recommendations, 23 were ranked greater than 4 and 13 were ranked greater than 4.25. These highly ranked recommendations are included in Table 4.

Table . Summary of highly ranked actionable recommendations for critical care innovation.

Key domains and recommendation[a]	Rating[b]
Innovation in health care	
Innovation must be fostered through interinstitutional relationships as well as partnerships between research, industry, and health care	4.50
Innovation in regionalized care	
We need to train and mentor early-career investigators to develop future innovation researchers	4.31
We need to promote critical care research in the region	4.26
We need to review policies, protocols, and medical directives to identify 1 to 3 innovation priorities annually to move forward with	4.25
Critical care practices	
Involve families in patient physical therapy in ICU[c] and in home rehabilitation	4.46
Implement centralized regional critical care coordination to optimize specialized care and bed occupancy	4.44
Implement lung-safe ventilation and follow evidence-based guideline recommendations	4.43
Implement a pilot program for wearable monitors in the region	4.42
Protect patients from sounds in the ICU with the use of headphones or Halo	4.41
Use of AI[d] in health care	
To avoid vendor lock-in, device-agnostic platforms must be used to tie analytics together	4.32
We must continue to seek to get more value from monitoring, by implementation and evaluation of predictive clinical decision support tools	4.30
We must improve data capture and availability to personalize critical event prediction	4.30
We must identify and prioritize the tasks that humans should retain; we need to capitalize on the human element to maximize effectiveness	4.27

[a]Recommendations with an overall Likert rating of ≥4.25 were included in the table.

[b]Rating was evaluated on a 5-point Likert scale.

[c]ICU: intensive care unit.

[d]AI: artificial intelligence.

In the domain of "innovation in health care," participants emphasized fostering partnerships between research, industry, and health care institutions as a key driver of innovation (rating: 4.50). Embedding innovation into organizational strategies received strong support, with a focus on recognizing local innovation efforts and sharing practical tools with related areas to enhance broader applicability. For "innovation in regionalized care," mentoring early-career investigators and promoting regional critical care research were prioritized, with participants stressing the importance of developing future researchers (ratings: 4.31 and 4.26, respectively). Another key recommendation was to review policies, protocols, and medical

XSL•FO
RenderX

directives annually to identify and move forward with 1 to 3 innovation priorities (rating: 4.25). In "critical care practices," involving families in patient physical therapy and home rehabilitation was highlighted as an important strategy for enhancing patient outcomes (rating: 4.46). Establishing centralized regional critical care coordination to optimize care and bed occupancy also received strong support (rating: 4.44). Additional recommendations included implementing lung-safe ventilation protocols, initiating a pilot program for wearable monitors, and reducing noise exposure in the ICU by using headphones (ratings: 4.43, 4.42, and 4.41, respectively). In the "use of AI in health care" domain, participants highlighted the importance of using device-agnostic platforms to prevent vendor lock-in and promote cohesive data analytics (rating: 4.32). Participants also supported improving data capture and using predictive clinical decision support tools to extract more value from monitoring (rating: 4.30 and 4.30, respectively). Emphasizing the human element by identifying and retaining tasks best suited to humans rather than AI was also considered critical (rating: 4.27).

Redefining Innovation in Health Care

Health care innovation is frequently reduced to a radical process that yields disruptive, revolutionary outcomes. In contrast, incremental innovation can represent disciplined, user-centered improvements that address practical challenges [8]. The optimal approach to a sustainable, innovative culture may therefore be achieved incrementally through regular, collaborative assessment of real problems faced by health care providers, patients, and institutions [9]. The regional AID event embodies this definition of innovation, focusing on low-cost, actionable improvements designed to enhance patient outcomes and health care system efficiency. Unlike typical innovation workshops that emphasize idea generation alone, the AID event formalizes the process by requiring structured output, involving multilevel stakeholders, and using postevent consensus polling. These features position the AID framework as a replicable model for driving practical innovation across health care settings, although the inaugural event focused on critical care delivery.

Furthermore, the AID framework's participatory model promotes easily implemented ideas that are subsequently positioned for rapid iteration and validation based on real-world feedback. By centering on the immediate needs of stakeholders, the AID event moves away from the often imagined "lightbulb moment" of innovation toward a collaborative, methodical approach that balances creativity with practical implementation. Nonetheless, it is recognized that subsequent implementation of AID is a significantly greater barrier.

Building on the AIF

The regional AID event shares many parallels with and builds on the principles of Agile Innovation, which emphasizes iterative cycles of planning, execution, and adaptation [10]. Within the AIF, the fifth step includes an innovation forum—a structured design event that brings various stakeholders to brainstorm potential innovations. The AID model represents an enhanced and expanded version of this forum. In the outlined AID approach, stakeholder engagement was the first critical component. By involving individuals from multiple disciplines, including those who control resources and those directly affected by care, the AID approach ensured that the identified problems had practical and systemic significance. Our process begins with presentations from local experts, which provide the necessary context and highlight the nuances of specific challenges within their domains. These expert-driven insights were crucial in framing the problems effectively and in proposing potential solutions tailored to real needs. The stakeholders conclude their presentations with actionable recommendations informed by the literature, professional experience, and/or expert opinion. This structured knowledge-sharing phase frames the problem effectively and seeds the ideation process. Following the presentations, we facilitated group breakout sessions that enabled stakeholders to refine the presented solutions and participate in a creative, judgment-free ideation session. These sessions provided a multidisciplinary platform for the exploration of innovative approaches and the assessment of the feasibility of each idea. The collective prioritization of actionable recommendations allowed the AID contributors to establish a foundation for subsequent development and validation phases of the AIF.

Beyond its direct impact on critical care, the AID approach has broader relevance, given the pressing challenges of an aging population and rising health care expenditures. With the projected increase in chronic diseases and long-term care needs, health care systems require structured, multidisciplinary approaches that can rapidly identify and implement cost-effective, sustainable solutions. The iterative nature of the AID framework allows for practical, system-level interventions that can be adapted to various domains, from geriatrics to community-based preventive health programs.

Overcoming Conventional Barriers to Innovation

Systemic barriers often impede health care innovation [11]. The AID approach deliberately addressed these barriers by incorporating specific strategies designed to facilitate a conducive environment to implement change. One significant barrier to health care innovation is the misalignment between administrative priorities and the adaptive, rapid changes required for practical innovation [11]. Policymakers and administrators may be reluctant to embrace such changes due to concerns about potential risks, costs, or disruptions to established processes. The AID event actively involved administrative leaders throughout the event to align innovation priorities with organizational goals, enabling smoother adoption and increasing the likelihood of long-term integration. Furthermore, the intention to promote incremental innovation, as opposed to radical invention, ensures that the implementation process can be more readily navigated. A second substantial barrier involves the limited availability of financial and infrastructure resources [7]. High-cost innovations can be particularly challenging to implement in resource-constrained environments such as Canadian health care. The AID framework addressed this challenge by emphasizing low-cost, high-impact solutions that required minimal resource investment. This approach not only made the proposed innovations more feasible for a wide range of institutions but also reduced barriers related to resource allocation, thereby promoting more equitable innovation. Third, the diffusion of innovation is another critical aspect of

XSL•FO
RenderX

sustainable health care change [12,13]. The AID approach promoted the diffusion of recommendations in several ways. The recommendations are naturally distributed throughout the health authority by involving innovators and early adopters across most disciplines and from sites across the region. Similarly, because of the collaborative approach to developing the recommendations, those involved are more likely to appreciate the relevance and alignment of the proposed solutions to the specific context of their local community. Finally, by inviting administrators, we hope to promote a cultural shift that promotes innovation from those in leadership roles. Leaders who actively participate in and advocate for innovative initiatives set a positive example, foster a supportive culture, and help embed innovation as an organizational value [14]. Barriers such as resistance to change often hinder the adoption of new practices. This resistance can be exacerbated by fears that new innovations will add to the workload of already overburdened health care teams. The AID framework sought to counter these concerns by focusing on innovations designed to streamline workflows and reduce unnecessary burdens. It is anticipated that approaches such as the AID framework could help foster a more positive perception of innovation, thereby encouraging broader acceptance and adoption among health care teams.

From the AID Approach to Implementation: Next Steps in the AIF

The AIF provides a pathway for advancing the actionable recommendations from the regional AID event beyond early development to institutional or regional integration. The AID approach corresponds to the innovation forum described in the fifth stage of the AIF, serving as an initial platform for generating and refining innovative ideas. However, to move these ideas toward real-world implementation, the subsequent AIF stages, which are focused on continuous iteration, refinement, and early implementation, must be undertaken. These stages facilitate “development sprints” that iterate rapidly on solutions, incorporating ongoing feedback to prepare innovations for broader evaluation and deployment [10]. For these next steps, ongoing communication and collaboration among AID event participants would be a cornerstone of success. This will ensure the development of practical, context-specific solutions that can move forward effectively. The AID model can also be conducted at regular intervals to maintain momentum and evaluate progress. Future iterations of the AID event can facilitate follow-up discussions on implementation progress, allowing for both synchronous and asynchronous advancement of innovations.

Limitations

One of the challenges during a communal collegial effort to distill consensus recommendations is the need to summarize information during a discussion process into a summary list of recommendations. This is variably facilitated by a chair and therefore lacks standardization. This would be a limitation of this work and the method in general. In addition, a synthesis of recommendations may also be subject to typically unconscious bias. Nonetheless, there are recognized best practice means to amalgamate and synthesize information, which would benefit this type of single-day collaborative educational event designed to generate consensus recommendations. The Delphi method is a validated, iterative consensus-building technique in which experts anonymously rate and refine a series of statements across multiple rounds. When used after an innovation conference, it provides a systematic means to transform raw, diverse input into consolidated, prioritized, and consensus-informed recommendations, reducing bias while ensuring broad stakeholder alignment.

Moving Forward: The Role of the AID Framework in Health Care Innovation

The AID process represents a possible organizational format for a meeting to engage health care leaders on the topic and process of innovation, supporting the generation of consensus recommendations and hopefully serving as a catalyst for innovation that improves care. The design is aimed to create a participatory model that engages multidisciplinary and multilevel stakeholders. By aligning with the AIF, a structured, practical approach focused on low-cost, scalable solutions to advancing critical care is provided, with the aim of developing innovations that are both impactful and sustainable. However, we appreciate that the next steps are to follow through on the recommendations, that is, turning recommendations into action, which is the admittedly more challenging next step of implementation. The AID process offers a practical framework for initiating, developing, and refining innovation within health care. Although this approach was demonstrated in critical care, we encourage health care leaders and stakeholders to adapt the AID process to their own specialties and local contexts by engaging relevant partners within their institutions and regions. Ultimately, the AID event approach can be flexibly applied and scaled to support innovation across diverse health care settings.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Disclaimer

The authors used the generative artificial intelligence tool ChatGPT by OpenAI to improve grammar and clarity in the Results and Discussion sections of the paper. All the speakers mentioned in the agenda have provided consent to disclose their names for publication.

XSL•FO
RenderX

Authors' Contributions

BNH contributed to the conceptualization, investigation, methodology, validation, writing of the original draft, and reviewing and editing. CT contributed to the data curation, formal analysis, investigation, visualization, writing of the original draft, and reviewing and editing. RR contributed to the data curation, formal analysis, investigation, writing of the original draft, and reviewing and editing. CL, CC, A Sarti, AM, S Madore, MC, AP, SD, S Millington, and HG contributed to conceptualization, methodology, project administration, and resources. A Seely performed the roles of conceptualization, methodology, project administration, resources, supervision, writing, and reviewing and editing.

Conflicts of Interest

A Seely is the founder and CEO of Therapeutic Monitoring Systems Inc, a company that develops variability-derived predictive point-of-care clinical decision support software tools. A Seely manages this active conflict of interest with agreements with the Ottawa Health Research Institute and disclosures on grants, presentations, and publications. All other authors declare no other relevant conflicts of interest.

Multimedia Appendix 1
Attendee agreement ratings for recommendations related to (1) innovation in health care, (2) innovation in regionalized care, (3) critical care practices, and (4) use of artificial intelligence in health care. Agreement was rated on a Likert scale (0=no agreement and 5=strong agreement).
[PNG File, 672 KB - jmir_v28i1e73614_app1.png]

References

1. Kosiol J, Silvester T, Cooper H, Alford S, Fraser L. Revolutionising health and social care: innovative solutions for a brighter tomorrow - a systematic review of the literature. BMC Health Serv Res 2024 Jul 12;24(1):809. [doi: 10.1186/s12913-024-11099-5] [Medline: 38997711]
2. Duong D. Primary care is facing a capacity crisis — can pandemic lessons help chart a path forward? CMAJ 2022 Nov 7;194(43). [doi: 10.1503/cmaj.1096023]
3. Singh J, Poon DEO, Alvarez E, et al. Burnout among public health workers in Canada: a cross-sectional study. BMC Public Health 2024 Jan 2;24(1):48. [doi: 10.1186/s12889-023-17572-w] [Medline: 38166742]
4. Kremsdorf R. Using innovation technology to enhance patient care delivery. Nurs Outlook 2003;51(3):S16-S20. [doi: 10.1016/s0029-6554(03)00099-x] [Medline: 12830124]
5. Drummond AJ. No room at the inn: overcrowding in Ontario's emergency departments. CJEM 2002 Mar;4(2):91-97. [doi: 10.1017/s1481803500006187] [Medline: 17612426]
6. Guttmann A, Schull MJ, Vermeulen MJ, Stukel TA. Association between waiting times and short term mortality and hospital admission after departure from emergency department: population based cohort study from Ontario, Canada. BMJ 2011 Jun 1;342:d2983. [doi: 10.1136/bmj.d2983] [Medline: 21632665]
7. Parmar J, Sacrey LA, Anderson S, et al. Facilitators, barriers and considerations for the implementation of healthcare innovation: a qualitative rapid systematic review. Health Soc Care Community 2022 May;30(3):856-868. [doi: 10.1111/hsc.13578] [Medline: 34558143]
8. Rubin GD, Abramson RG. Creating value through incremental innovation: managing culture, structure, and process. Radiology 2018 Aug;288(2):330-340. [doi: 10.1148/radiol.2018171239] [Medline: 29969069]
9. Segarra-Oña M, Peiró-Signes Á, Verma R. Fostering innovation through stakeholders' engagement at the healthcare industry: tapping the right key. Health Policy 2020 Aug;124(8):895-901. [doi: 10.1016/j.healthpol.2020.05.013] [Medline: 32522366]
10. Holden RJ, Boustani MA, Azar J. Agile Innovation to transform healthcare: innovating in complex adaptive systems is an everyday process, not a light bulb event. BMJ Innov 2021 Apr;7(2):499-505. [doi: 10.1136/bmjinnov-2020-000574]
11. MacNeil M, Koch M, Kuspinar A, Juzwishin D, Lehoux P, Stolee P. Enabling health technology innovation in Canada: barriers and facilitators in policy and regulatory processes. Health Policy 2019 Feb;123(2):203-214. [doi: 10.1016/j.healthpol.2018.09.018] [Medline: 30352755]
12. Berwick DM. Disseminating innovations in health care. JAMA 2003 Apr 16;289(15):1969-1975. [doi: 10.1001/jama.289.15.1969] [Medline: 12697800]
13. Dearing JW, Cox JG. Diffusion of innovations theory, principles, and practice. Health Aff (Millwood) 2018 Feb;37(2):183-190. [doi: 10.1377/hlthaff.2017.1104] [Medline: 29401011]
14. van den Hoed MW, Backhaus R, de Vries E, Hamers JPH, Daniëls R. Factors contributing to innovation readiness in health care organizations: a scoping review. BMC Health Serv Res 2022;22(1):997. [doi: 10.1186/s12913-022-08185-x]

Abbreviations

AID: Actionable Innovation Day
AIF: Agile Innovation Framework

XSL•FO
RenderX

ICU: intensive care unit

Edited by A Mavragani, N Cahill; submitted 08.Mar.2025; peer-reviewed by YK Lee, Z Dumont; revised version received 19.Dec.2025; accepted 31.Dec.2025; published 08.May.2026.

Please cite as:
Hryciw BN, Tran C, Ramchandani R, Love C, Caron C, Sarti A, Miller A, Madore S, Chasse M, Pan A, Didcote S, Millington S, Galley H, Kyeremanteng K, Seely A
The Actionable Innovation Day Approach: Participatory Model for Advancing Critical Care Innovation
J Med Internet Res 2026;28:e73614
URL: https://www.jmir.org/2026/1/e73614
doi:10.2196/73614

© Brett N Hryciw, Cecilia Tran, Cameron Love, Christine Caron, Aimee Sarti, Annelise Miller, Suzanne Madore, Michael Chasse, Andy Pan, Simon Didcote, Scott Millington, Heather Galley, Kwadwo Kyeremanteng, Andrew Seely. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Digital Exposome: A Life Course Framework for Health in the Digital Age

Pascal Petit[1], PhD; Nicolas Vuillerme[1,2], PhD

[1]Univ. Grenoble Aples, CNRS, Grenoble INP, LIG, SANGRIA, Grenoble, France
[2]Institut Universitaire de France, Paris, France

Corresponding Author:
Pascal Petit, PhD
Univ. Grenoble Aples, CNRS, Grenoble INP, LIG, SANGRIA, Grenoble, France

Abstract

Digital technologies are reshaping human behavior, health care delivery, and population health; however, their cumulative effects across the lifespan remain underexplored. This viewpoint argues that exposures arising from interactions with digital technologies should be formally integrated into exposome science as a distinct, measurable component of the human environment. Our aims are to (1) redefine the digital component of the exposome (the digital exposome) within the broader exposome framework, (2) examine its life course implications for health and equity, and (3) outline a research and policy agenda to enable its systematic measurement and integration into clinical and public health practice. Digital technology–related exposures can confer benefits such as enhanced health monitoring, personalized interventions, improved access to care, and the promotion of healthy behaviors. However, they may also introduce potential risks, including mental health challenges, cognitive and circadian disruptions, sedentary lifestyles, exposure to misinformation, and widening inequities among vulnerable populations. Despite their ubiquity, digital technology–related exposures remain poorly integrated into clinical medicine, epidemiology, or public and global health policies. Drawing on interdisciplinary evidence from exposure science, epidemiology, and digital phenotyping research, we propose a refined conceptual definition of the digital exposome grounded in the classical exposome domains. We propose redefining the digital exposome as the full spectrum of exposures resulting from interactions or proximities with digital technologies and their combined influence on health across the lifespan. This framework conceptualizes digital technology–related exposures as a dynamic set of environmental influences operating through sociotechnical, behavioral, and biological pathways over the life course. To operationalize this framework, we discuss practical approaches using validated behavioral instruments, objective device use logs, ecological momentary assessments, smartphone-based digital phenotyping, and wearable sensing technologies. Systematic measurement, large-scale longitudinal studies, and harmonized exposure metrics are needed to characterize the cumulative health impacts of digital environments more accurately. Emerging tools such as digital markers or biomarkers and digital phenotypes offer promising opportunities to link real-world technology use with physiological and biological outcomes, thereby supporting precision medicine and population health strategies. Ethical governance, privacy safeguards, and equity considerations must be embedded from the start, drawing on emerging exposomethics frameworks. Recognizing the digital exposome as a modifiable determinant of health offers a foundation for evidence-based guidance, prevention strategies, and policy interventions suited to increasingly digital societies. By integrating digital technology–related exposures into exposome science, clinical practice, and public health research, this viewpoint seeks to foster interdisciplinary dialogue, guide future empirical work, and support the development of safer and more equitable digital environments across the lifespan.

(J Med Internet Res 2026;28:e90153) doi:10.2196/90153

KEYWORDS

digital health; life course health; digital exposome; digital determinants of health; health equity; digital public health; digital technologies; mobile phone; digital phenotype; wearables

A Constantly Evolving Digital World

Societies worldwide are undergoing a profound digital transformation, as information and communication technologies increasingly shape everyday life and influence health and well-being across the lifespan [1-3]. From early childhood to older age, digital technologies (DTs) influence education, health care delivery, communication, and social connection, fundamentally reshaping how people live, learn, and interact [1,4]. Individuals now navigate environments characterized by rapid connectivity, omnipresent media platforms, and instantaneous access to health-related information [1-3].

DTs have become powerful enablers of health and well-being, supporting the broader vision of health for all [5]. Artificial intelligence (AI)-driven tools, wearable devices, mobile health, and eHealth technologies (eg, telemedicine platforms) facilitate exposure and health monitoring, improve access to care, and

XSL•FO
RenderX

support personalized interventions [4-8]. In digital psychiatry, for instance, passive data streams and AI-based tools are increasingly used to infer mental health states and support personalized care [8,9]. Educational technologies can also foster learning and cognitive development [4].

However, the influence of DTs on health is inherently dual. While they offer substantial benefits, they also introduce potential risks, particularly for vulnerable populations such as children, older adults, and individuals facing barriers related to digital literacy, infrastructure, or socioeconomic conditions [3,4,10,11]. For example, some forms of digital media use, such as exchanging SMS text messages with close friends or viewing positive or nature-related images, have been associated with improvements in well-being [12]. In contrast, prolonged screen time or exposure to artificial light at night has been linked to adverse health outcomes [12].

The emerging concept of the digital determinants of health highlights how technological factors influence disparities in access, affordability, and health outcomes [13,14]. These determinants interact dynamically with social determinants of health (SDOH), often exacerbating inequalities and influencing quality of care and accessibility [13,14]. Recognizing these interactions is essential for achieving equitable health progress in increasingly digital societies. The dual nature of digital influence underscores the need for conceptual frameworks that address both the benefits and the emerging risks of digital transformation.

In this context, the exposome concept [15,16] provides a promising framework for understanding the cumulative impact of digital environments on health. This viewpoint develops a conceptual and operational framework for integrating DT-related exposures into the exposome paradigm. Our primary aims are to (1) redefine the digital exposome within the broader exposome framework, (2) articulate its implications for health and equity across the life course, and (3) outline a research and policy agenda to support its measurement and integration into clinical and public health practice.

The intended audience includes digital health researchers, exposome scientists, clinicians, public health professionals, policymakers, sociologists, and ethics scholars. By bridging these disciplines, this viewpoint seeks to translate the concept of the digital exposome into actionable research agendas, measurement strategies, and policy-relevant interventions. It aims to stimulate interdisciplinary dialogue; guide future empirical research (eg, digital phenotyping and digital marker or biomarker development); and inform decision-making in clinical, public health, digital health, and policy contexts.

A targeted narrative literature review informed this viewpoint by identifying relevant publications from databases such as PubMed, Web of Science, Dimensions, and Google Scholar, using search terms related to the exposome, DT, and health.

Mapping Digital-Related Exposures

The exposome concept, representing the totality of environmental influences from conception to death [15,16], offers a powerful paradigm for characterizing the full spectrum of DT-related exposures across the life course. However, despite their ubiquity in modern societies, DT-related exposures remain largely absent from current exposome frameworks [1,2]. Only a limited number of studies have begun to explore their integration into exposomic research [17,18].

Recent investigations illustrate both the feasibility and potential value of incorporating DT-related exposures into exposome science. For instance, a data-driven exposome-wide association study (ExWAS) introduced digital exposomic risk scores that quantify cumulative exposure to digital risks such as general usage, cyberbullying, and problematic or addictive behavior in adolescents [17]. These scores explained a substantial proportion of the variance in self-reported mental health and enhanced understanding of how DT-related exposures contribute to mental health inequities [17]. Similarly, machine learning analyses of UK Biobank data examined dementia risk in relation to 128 exposome factors, including DT-related exposures, and found an inverse association between dementia and the duration of mobile phone use [18]. Although causal inferences remain uncertain, these findings highlight the potential relevance of DT-related exposures within broader environmental health research.

These pioneering studies highlight the necessity, feasibility, and potential utility of integrating DT-related exposures into exposomic research at both the population and individual levels. As individuals spend increasing portions of their lives online or connected digitally, these exposures must be recognized as meaningful components of the exposome [1]. Consequently, the question is no longer whether DT-related exposures belong within exposome science, but rather how they can be systematically measured, modeled, and integrated. Expanding the exposome framework to encompass DT-related exposures may help bridge critical knowledge gaps and support the development of patient-centered, evidence-based interventions in an increasingly digitalized world.

Redefining the Digital Exposome

Previous efforts have defined the digital exposome as “the whole set of tools and platforms (including contents) that an individual uses and the activities and processes that an individual engages with as part of his digital life” [1]. While foundational, this definition remains too narrow to capture the full complexity of lifelong DT-related exposures.

We propose redefining the digital exposome as the complete and cumulative collection of direct and indirect exposures deriving from interactions or proximities with DTs and their combined influence on health across the lifespan. This definition emphasizes exposure rather than mere usage, aligning with the broader exposome principle that environmental factors operate cumulatively and dynamically throughout life.

This redefinition can be positioned within the classical 3-domain structure of the exposome (Figure 1) [15]. Figure 1 synthesizes established exposome theory with emerging scholarship on DT use, digital footprints and tracing [1], digital markers or biomarkers [19,20], and digital phenotyping [2,21,22], resulting in an integrative view of the digital exposome. Each exposome

XSL•FO
RenderX

domain is organized into thematic categories. This conceptual representation is informed in interdisciplinary evidence spanning public health, neuroscience, behavioral science, and digital health studies.

Figure 1. Conceptual framework of the digital component of the exposome. The classical 3-domain exposome structure [15] was adapted for digital technology–related exposures. Bidirectional arrows represent dynamic lifespan interactions, with heightened vulnerability of individuals during developmental windows (eg, childhood and adolescence). FOMO: fear of missing out; mHealth: mobile health.

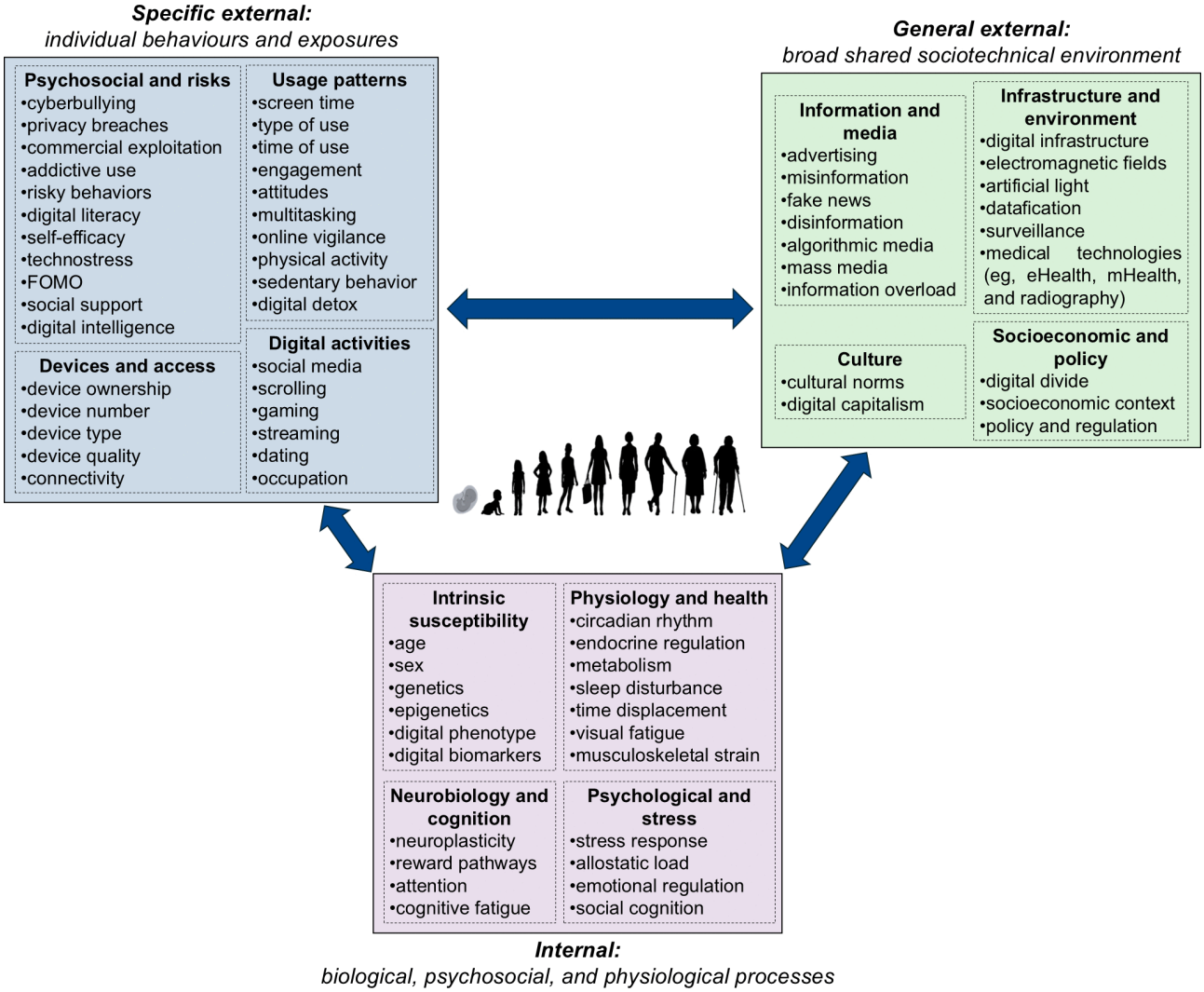

The upper part of Figure 1 depicts the 2 external domains of the digital exposome. The general external digital exposome (upper right) refers to population-level, contextual DT-related exposures that operate broadly across societies. These include structural and environmental influences such as misinformation, information overload, blue light exposure, electromagnetic fields, digital advertising ecosystem, digital infrastructure, and algorithmic content curation.

In contrast, the specific external digital exposome (upper left) encompasses individualized and behavioral DT-related exposures shaped by personal usage patterns and access conditions. Examples include screen time, digital literacy, multitasking, cyberbullying experiences, and occupational digital practices. Both external domains involve direct and indirect DT-related exposures. Direct exposures stem from immediate physical or behavioral interactions with DT (eg, screen time, sedentary behavior, and blue light exposure), whereas indirect exposures arise through mediated sociotechnical processes such as algorithmic filtering, addictive digital design, misinformation, or enforced digital disengagement strategies.

The lower part of Figure 1 represents the internal digital exposome domain, which captures biological, physiological, neurocognitive, and psychosocial responses to external digital stimuli. This domain includes processes such as neuroplasticity, stress activation, cognitive fatigue, sleep disturbance, appetite changes, and circadian disruption or metabolic alterations.

The conceptualization assumes dynamic and bidirectional interactions between internal susceptibilities and external DT-related exposures throughout the life course. The 3 domains of the digital exposome interact dynamically across the lifespan and are particularly influential during sensitive developmental stages such as childhood and adolescence, when exposures can have lasting and disproportionate effects [12,23]. For example, adolescence represents a sensitive period in which many young people spend more time on digital platforms (eg, smartphones, tablets, and computers) than with peers or family members [12,24].

XSL•FO
RenderX

Health Impacts Across the Lifespan

DTs exert complex, bidirectional effects across the lifespan [25]. The digital exposome influences multiple organ systems through physiological and behavioral pathways, including cardiovascular strain from sedentary lifestyles, metabolic dysregulation, and endocrine and immune alterations [2,10,17]. The adverse consequences of DT-related exposures can be broadly categorized into direct and indirect effects. Direct effects arise from the content or functional characteristics of digital media (eg, multitasking, push notifications, and blue light exposure), while indirect effects result from displacement of other essential activities, including sleep, physical exercise, and schoolwork [12].

Mental health and cognitive development represent the most extensively studied domains, particularly since the COVID-19 pandemic [12]. DT use has been associated with depression, anxiety, low self-esteem, addiction, and suicidal ideation [2,10,17], although causal relationships remain difficult to establish [24]. For example, among adolescents, very high daily time spent on digital media (≥6 hours/day), excluding schoolwork, correlates with lower self-esteem, depressive symptoms, and poorer sleep quality [12]. In addition, generative AI technologies may expose children to unsafe or inappropriate content, further compounding risk [11].

Beyond mental health, DTs contribute to a range of other disorders, including ocular conditions [4,10], noise-induced hearing loss [4], musculoskeletal disorders and pain syndromes [4,10,25,26], and obesity [10]. Blue light exposure during daytime may enhance attention, but nighttime exposure disrupts circadian rhythms and is associated with sleep disturbances and other health risks [27]. Prolonged screen use similarly impairs sleep quality, disrupts circadian regulation, and reduces next-day alertness [10,28]. Infodemics during COVID-19 illustrate population-level consequences of digital misinformation that negatively affects disease outbreak management [29,30]. Excessive reliance on DT may also contribute to cognitive decline, a phenomenon often described as digital dementia [3,31]. The digitalization of workplaces has further reshaped occupational settings, improving efficiency in some contexts while simultaneously heightening psychosocial stress and ergonomic risks [25].

Importantly, these effects are unevenly distributed, as illustrated by the digital determinants of health framework [13,14]. Populations with limited digital literacy, inadequate access, or lack of algorithmic transparency face greater exposure risks and fewer benefits, perpetuating the digital divide and reinforcing health inequities [5].

Barriers to Integration

Integrating the digital exposome into health science presents methodological, practical, and ethical challenges [2,24,32-34]. DT-related exposures are ubiquitous, rapidly evolving, and deeply interwoven with daily life, making them difficult to measure and mitigate.

Current assessments rely predominantly on self-reported measures or indicators such as screen time or digital media use, which correlate only modestly with objective device logs and suffer from low validity [12,24,33-36]. These measures fail to capture content quality, engagement type, multitasking, contextual use (eg, educational vs recreational), or biological correlates [33,34,36]. Ecological momentary assessment offers real-time behavioral data in naturalistic settings through ambulatory techniques such as diaries and physiological recordings [37,38]. By prompting multiple daily self-reports, ecological momentary assessment minimizes recall bias but struggles with longitudinal participant engagement [38].

Screen exposure quantification remains inconsistent. Self-reporting is subjective [39], eye-tracking glasses are invasive (particularly for children) and visually selective [39], and built-in smartphone apps lack cross-device compatibility and user verification [39,40]. A recent study proposed combining egocentric cameras with object detection to track multiscreen exposure patterns [39].

Passive smartphone and wearable sensing holds promise [16,39] but raises complex ethical challenges akin to broader exposome research. Continuous, granular data streams (location tracking, tappigraphy, and physiological signals) are often reidentifiable, creating issues of privacy, informed consent, data sovereignty, equity, governance, and actionability [1,32,41,42]. The General Data Protection Regulation–aligned solutions (eg, data minimization, on-device preprocessing, and deidentification) are essential but leave open questions about longitudinal analysis, secure storage, responsible sharing, and equitable benefit-sharing with disproportionately burdened populations [2,42]. Current normative frameworks remain insufficient for exposome data governance [41].

Traditional cross-sectional designs provide static snapshots, missing within-person variation over time and context [38,43]. Inconsistent DT-related exposure definitions impede study comparability [1,10,17], while SDOH confounding complicates causal inference [13,14]. SDOH are rarely discrete or easily measured, making it difficult to establish the well-defined exposures required for valid causal analysis [44,45]. While some SDOH components (eg, specific socioeconomic status pathways) may be more actionable than commonly recognized in epidemiology [46], causal inference from observational data still requires causal consistency (ie, specifying plausible interventions to shift exposure states) [44]. Decisions about modeling distal SDOH as confounders versus causal variables carry important methodological implications that require explicit justification in exposome research [46]. Robust modeling, confounder management, and interpretable pipelines remain essential to avoid spurious associations [45]. The expotype concept, defined as the vector of exposures accumulated over time, further underscores that clinical integration requires causal validation [46].

Future Directions

Addressing these challenges requires multidisciplinary collaboration (eg, clinicians, epidemiologists, neuroscientists, psychologists, sociologists, computer scientists, patients, and

XSL•FO
RenderX

digital health experts). Priority actions include developing harmonized metrics for DT-related exposures, embedding digital variables into ExWAS [16], and linking digital data with health outcomes [2,21,22].

Strengthening digital literacy through education and clinical training can empower health care professionals to recognize and manage digital risks in practice [4]. Clinical assessments should incorporate patterns of digital usage, behavioral addiction screening, and circadian disruption indicators, supporting personalized interventions such as guided digital detoxification or AI-supported cognitive therapies [8,9].

Large-scale longitudinal studies are urgently needed to clarify how DT-related exposures shape health trajectories across the life course, examining cumulative burden, timing, intensity, and context [23,33,43]. Key questions include the following: At which life stages are DT-related exposures most harmful or beneficial? How do lifelong digital habits interact with aging, chronic disease, and physical and cognitive decline? Which environmental factors influence the adoption, frequency, intensity, and type (eg, problematic) of DT use?

Evidence-based public health guidelines must evolve to reflect the complexity of modern digital environments. Current screen time recommendations remain limited, often overlooking important distinctions between passive use and active, purposeful engagement [10,47,48]. Updated guidance should consider environmental factors, developmental stages, and emerging evidence on both the benefits and risks of DT-related exposure [10,48].

Operationalizing the digital exposome requires objective metrics and validated instruments such as the Media and Technology Usage and Attitudes Scale [49], Cyberbullying Questionnaire [50], problematic internet use scales [51], and eHealth literacy tools [52]. Standardized metrics should include device logs (eg, unlocks, app categories, and late-night use) and contextual self-reports (eg, purpose and weekday or weekend patterns). Behavioral indicators (eg, app switches, concurrent windows, and notification responsiveness) require harmonization across studies. Objective monitoring approaches are becoming increasingly feasible, including in-vivo measures of online behavior (eg, platform-specific time-use data and coded analysis of social media interactions) [43] and smartphone apps that can passively log screen-on duration [36,53], track fine-grained touchscreen interactions ("tappigraphy") [40], or merge logging with experience sampling [54]. For example, among older adults, analyses of call detail records have been used to explore the relationship between social asymmetry and depressive symptoms, demonstrating how digital trace data may reveal psychosocial vulnerability [55]. Wearable and mobile sensors enable highly individualized, high-resolution exposure and physiological monitoring [16]. Embedding digital variables into ExWAS could facilitate systematic linkage of DT-related exposures with biological and clinical outcomes.

Occupational medicine should quantify "digital task load" (eg, hours of intensive email use, number of online meetings, after-hours connectivity, screen breaks, or multitasking patterns) and relate it to adverse health outcomes such as psychosocial hazards (eg, burnout) and musculoskeletal issues.

Digital markers or biomarkers, defined as objective and quantifiable physiological or behavioral measures derived from digital devices, offer new opportunities to link continuous real-world data with health outcomes [19,20]. Complementing this approach, the digital phenotype refers to the measurable footprint of an individual's digital behavior, encompassing patterns of interaction, communication, and technology use [1,2,19,21,22]. Smartphone-based digital phenotyping provides scalable, longitudinal assessment that links environmental context and behavioral data with chronic disease trajectories [22], allowing high-resolution profiling of mobility, behavior, and environment [22]. Together, these tools can bridge digital behavior and biological processes, enabling early identification of risk and resilience. Integrating digital markers or biomarkers and phenotypes into exposome research will enhance our ability to monitor, predict, and personalize health trajectories, advancing precision and personalized medicine in the digital age [21,22].

A Call to Action

The digital revolution is irreversible; however, its potential health consequences remain within our control. The digital exposome represents a defining determinant of contemporary health, encompassing both risks and opportunities. Its systematic integration into exposome science, clinical medicine, digital health, and public health is essential to ensure that digital environments promote rather than compromise well-being. This systematic integration is not merely conceptual but practically achievable through validated instruments [49,50], objective digital logging [53,54], longitudinal phenotyping [22], and harmonized DT-related exposure metrics.

The coming decade must turn awareness into action. Coordinated collaboration among researchers, clinicians, educators, policymakers, and funders can transform the digital exposome from a conceptual frontier into a foundation for healthier societies. By systematically recognizing and addressing DT-related exposures as integral components of the human environment, we can shape a future in which technological progress and public health advance together.

Funding

This work was partially supported by funding from the French government, managed by the National Research Agency (ANR) under the France 2030 program (MIAI Cluster) (grant ANR-23-IACL-0006) and under the "Investissements d'avenir" program (grants ANR-10-AIRT-0005 and ANR-15-IDEX-0002). The funding sources had no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the paper for publication. The content is solely the responsibility of the authors and does not necessarily represent the official views of any of the funders or collaborators.

XSL•FO
RenderX

Data Availability

Data sharing is not applicable to this paper as no datasets were generated or analyzed during this study.

Authors' Contributions

Conceptualization: PP
Funding acquisition: NV (lead), PP (supporting)
Investigation: PP
Methodology: PP
Project administration: PP
Supervision: PP
Visualization: PP
Writing – original draft: PP
Writing – review & editing: PP (lead), NV (supporting)

Conflicts of Interest

None declared.

References

1. Lopez-Campos G, Merolli M, Martin-Sanchez F. Biomedical informatics and the digital component of the exposome. Stud Health Technol Inform 2017;245:496-500. [Medline: 29295144]
2. Martin-Sanchez F, Bellazzi R, Casella V, Dixon W, Lopez-Campos G, Peek N. Progress in characterizing the human exposome: a key step for precision medicine. Yearb Med Inform 2020 Aug;29(1):115-120. [doi: 10.1055/s-0040-1701975] [Medline: 32303099]
3. Small GW, Lee J, Kaufman A, et al. Brain health consequences of digital technology use. Dialogues Clin Neurosci 2020 Jun;22(2):179-187. [doi: 10.31887/DCNS.2020.22.2/gsmall] [Medline: 32699518]
4. Cullen J, Muntz A, Marsh S, et al. Impacts of digital technologies on child and adolescent health: recommendations for safer screen use in educational settings. N Z Med J 2024 Jul 5;137(1598):9-13. [doi: 10.26635/6965.6565] [Medline: 38963926]
5. Neter E, Western MJ, Cooper R, Silva AM, König LM. Towards bridging the digital divide: training healthcare professionals for digitally inclusive healthcare systems. Glob Health Res Policy 2025 Jul 30;10(1):31. [doi: 10.1186/s41256-025-00433-x] [Medline: 40739280]
6. Gaona GA, Kassamali AA, Isakadze N, Martin SS. Harnessing wearables and digital technologies to decode the cardiovascular exposome. Methodist Debakey Cardiovasc J 2024;20(5):59-70. [doi: 10.14797/mdcvj.1459] [Medline: 39525376]
7. Johnson T, Kanjo E, Woodward K. DigitalExposome: quantifying impact of urban environment on wellbeing using sensor fusion and deep learning. Comput Urban Sci 2023;3(1):14. [doi: 10.1007/s43762-023-00088-9] [Medline: 36970599]
8. Torous J, Bucci S, Bell IH, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021 Oct;20(3):318-335. [doi: 10.1002/wps.20883] [Medline: 34505369]
9. Burr C, Morley J, Taddeo M, Floridi L. Digital psychiatry: risks and opportunities for public health and wellbeing. IEEE Trans Technol Soc 2020;1(1):21-33. [doi: 10.1109/TTS.2020.2977059]
10. Khanani MI, Khan MR, Farooqi MF, Fazal J, Aabideen Z, Alkuwaiti NS. Digital media use and screen time exposure among youths: a lifestyle-based public health concern. Cureus 2025 Jul;17(7):e88373. [doi: 10.7759/cureus.88373] [Medline: 40837898]
11. Towards digital safety by design for children: OECD digital economy papers. Organization for Economic Co-operation and Development. 2024. URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/06/towards-digital-safety-by-design-for-children_f1c86498/c167b650-en.pdf [accessed 2025-11-10]
12. Frielingsdorf H, Fomichov V, Rystedt I, Lindstrand S, Korhonen L, Henriksson H. Associations of time spent on different types of digital media with self-rated general and mental health in Swedish adolescents. Sci Rep 2025 Jan 6;15(1):993. [doi: 10.1038/s41598-024-83951-x] [Medline: 39762280]
13. Chidambaram S, Jain B, Jain U, et al. An introduction to digital determinants of health. PLOS Digit Health 2024 Jan;3(1):e0000346. [doi: 10.1371/journal.pdig.0000346] [Medline: 38175828]
14. Petretto DR, Carrogu GP, Gaviano L, et al. Digital determinants of health as a way to address multilevel complex causal model in the promotion of digital health equity and the prevention of digital health inequities: a scoping review. J Public Health Res 2024 Jan;13(1):22799036231220352 Retracted inJ Public Health Res. Dec 23, 2025;14(4):22799036251411945. [doi: 10.1177/22799036251411945] [Medline: 41467230]
15. Chung MK, House JS, Akhtari FS, et al. Decoding the exposome: data science methodologies and implications in exposome-wide association studies (ExWASs). Exposome 2024;4(1):osae001. [doi: 10.1093/exposome/osae001] [Medline: 38344436]

XSL•FO
RenderX

16. Sarigiannis D, Anesti O, Papaioannou N, Karakoltzidis A, Karakitsios S. Computational standards and tools for exposome-wide association studies linking the human exposome with health outcomes. Environ Int 2026 Feb;208:110117. [doi: 10.1016/j.envint.2026.110117] [Medline: 41691951]
17. Pagliaccio D, Tran KT, Visoki E, DiDomenico GE, Auerbach RP, Barzilay R. Probing the digital exposome: associations of social media use patterns with youth mental health. NPP Digit Psychiatry Neurosci 2024;2:5. [doi: 10.1038/s44277-024-00006-9] [Medline: 39464493]
18. Camacho M, Atehortúa A, Wilkinson T, Gkontra P, Lekadir K. Low-cost predictive models of dementia risk using machine learning and exposome predictors. Health Technol 2025 Mar;15:355-365. [doi: 10.1007/s12553-024-00937-5]
19. Powell D. Walk, talk, think, see and feel: harnessing the power of digital biomarkers in healthcare. NPJ Digit Med 2024 Feb 24;7(1):45. [doi: 10.1038/s41746-024-01023-w] [Medline: 38396034]
20. Macias Alonso AK, Hirt J, Woelfle T, Janiaud P, Hemkens LG. Definitions of digital biomarkers: a systematic mapping of the biomedical literature. BMJ Health Care Inform 2024 Apr 8;31(1):e100914. [doi: 10.1136/bmjhci-2023-100914] [Medline: 38589213]
21. Jain SH, Powers BW, Hawkins JB, Brownstein JS. The digital phenotype. Nat Biotechnol 2015 May;33(5):462-463. [doi: 10.1038/nbt.3223] [Medline: 25965751]
22. Yi L, Hart JE, Straczkiewicz M, et al. Measuring environmental and behavioral drivers of chronic diseases using smartphone-based digital phenotyping: intensive longitudinal observational mHealth substudy embedded in 2 prospective cohorts of adults. JMIR Public Health Surveill 2024 Oct 11;10:e55170. [doi: 10.2196/55170] [Medline: 39392682]
23. Korte M. The impact of the digital revolution on human brain and behavior: where do we stand? Dialogues Clin Neurosci 2020 Jun;22(2):101-111. [doi: 10.31887/DCNS.2020.22.2/mkorte] [Medline: 32699510]
24. Hale L, Guan S. Screen time and sleep among school-aged children and adolescents: a systematic literature review. Sleep Med Rev 2015 Jun;21:50-58. [doi: 10.1016/j.smrv.2014.07.007] [Medline: 25193149]
25. Demissie B, Bayih ET, Demmelash AA. A systematic review of work-related musculoskeletal disorders and risk factors among computer users. Heliyon 2024;10(3):e25075. [doi: 10.1016/j.heliyon.2024.e25075] [Medline: 38318034]
26. Al-Hadidi F, Bsisu I, AlRyalat SA, et al. Association between mobile phone use and neck pain in university students: a cross-sectional study using numeric rating scale for evaluation of neck pain. PLoS One 2019;14(5):e0217231. [doi: 10.1371/journal.pone.0217231] [Medline: 31107910]
27. Haghani M, Abbasi S, Abdoli L, et al. Blue light and digital screens revisited: a new look at blue light from the vision quality, circadian rhythm and cognitive functions perspective. J Biomed Phys Eng 2024 Jun;14(3):213-228. [doi: 10.31661/jbpe.v0i0.2106-1355] [Medline: 39027713]
28. Chang AM, Aeschbach D, Duffy JF, Czeisler CA. Evening use of light-emitting eReaders negatively affects sleep, circadian timing, and next-morning alertness. Proc Natl Acad Sci U S A 2015 Jan 27;112(4):1232-1237. [doi: 10.1073/pnas.1418490112] [Medline: 25535358]
29. Cinelli M, Quattrociocchi W, Galeazzi A, et al. The COVID-19 social media infodemic. Sci Rep 2020 Oct 6;10(1):16598. [doi: 10.1038/s41598-020-73510-5] [Medline: 33024152]
30. The Lancet Infectious Diseases. The COVID-19 infodemic. Lancet Infect Dis 2020 Aug;20(8):875. [doi: 10.1016/S1473-3099(20)30565-X] [Medline: 32687807]
31. Ali Z, Janarthanan J, Mohan P. Understanding digital dementia and cognitive impact in the current era of the internet: a review. Cureus 2024 Sep;16(9):e70029. [doi: 10.7759/cureus.70029] [Medline: 39449887]
32. Calabrò RS. Context is medicine: integrating the exposome into neurorehabilitation. Brain Sci 2025 Nov 7;15(11):1198. [doi: 10.3390/brainsci15111198] [Medline: 41300205]
33. Dienlin T, Johannes N. The impact of digital technology use on adolescent well-being. Dialogues Clin Neurosci 2020 Jun;22(2):135-142. [doi: 10.31887/DCNS.2020.22.2/tdienlin] [Medline: 32699513]
34. Scharkow M. The accuracy of self-reported internet use—a validation study using client log data. Commun Methods Meas 2016;10(1):13-27. [doi: 10.1080/19312458.2015.1118446]
35. Plackett R, Blyth A, Schartau P. The impact of social media use interventions on mental well-being: systematic review. J Med Internet Res 2023 Aug 11;25:e44922. [doi: 10.2196/44922] [Medline: 37565693]
36. Júdice PB, Sousa-Sá E, Palmeira AL. Discrepancies between self-reported and objectively measured smartphone screen time: before and during lockdown. J Prev 2023 Jun;44(3):291-307. [doi: 10.1007/s10935-023-00724-4]
37. Götzl C, Hiller S, Rauschenberg C, et al. Artificial intelligence-informed mobile mental health apps for young people: a mixed-methods approach on users' and stakeholders' perspectives. Child Adolesc Psychiatry Ment Health 2022 Nov 17;16(1):86. [doi: 10.1186/s13034-022-00522-6] [Medline: 36397097]
38. Prochnow T, Wang WL, Wang S, et al. Understanding longitudinal ecological momentary assessment completion: results from 12 months of burst sampling in the TIME study. JMIR Mhealth Uhealth 2025 Oct 22;13:e67117. [doi: 10.2196/67117] [Medline: 41124691]
39. Li X, Holiday S, Cribbet M, et al. Non-invasive screen exposure time assessment using wearable sensor and object detection. Annu Int Conf IEEE Eng Med Biol Soc 2022 Jul;2022:4917-4920. [doi: 10.1109/EMBC48229.2022.9871903] [Medline: 36086530]

XSL•FO
RenderX

40. Massar SA, Chua XY, Soon CS, et al. Trait-like nocturnal sleep behavior identified by combining wearable, phone-use, and self-report data. NPJ Digit Med 2021 Jun 2;4(1):90. [doi: 10.1038/s41746-021-00466-9] [Medline: 34079043]
41. Safarlou CW, Jongsma KR, Vermeulen R, Bredenoord AL. The ethical aspects of exposome research: a systematic review. Exposome 2023 Apr 12;3(1):osad004. [doi: 10.1093/exposome/osad004] [Medline: 37745046]
42. Sillé FC, Belkadi M, Koehler K, et al. Charting exposomethics: a roadmap for the ethical foundations of the human exposome project. Hum Genomics 2026 Jan 17;20(1):34. [doi: 10.1186/s40246-025-00906-7] [Medline: 41547936]
43. Moroney E, Lee SS, Ebbert AM, Luthar SS. Digital media use frequency, online behaviors, and risk for negative mental health in adolescents in high-achieving schools. Clin Child Psychol Psychiatry 2023 Jan;28(1):237-254. [doi: 10.1177/13591045221108834] [Medline: 35770791]
44. Kaufman JS. Causal inference challenges in the relationship between social determinants and cardiovascular outcomes. Can J Cardiol 2024 Jun;40(6):976-988. [doi: 10.1016/j.cjca.2024.02.005] [Medline: 38365089]
45. Ponzano M, Rotem RS, Bellavia A. Complex methods for complex data: key considerations for interpretable and actionable results in exposome research. Eur J Epidemiol 2025 Dec;40(12):1399-1403. [doi: 10.1007/s10654-025-01281-2] [Medline: 40768123]
46. Vineis P. Prevention, precision prevention and precision medicine. Exposome 2025;5(1):osaf011. [doi: 10.1093/exposome/osaf011] [Medline: 41209440]
47. McArthur BA, Volkova V, Tomopoulos S, Madigan S. Global prevalence of meeting screen time guidelines among children 5 years and younger: a systematic review and meta-analysis. JAMA Pediatr 2022 Apr 1;176(4):373-383. [doi: 10.1001/jamapediatrics.2021.6386] [Medline: 35157028]
48. Tadpatrikar A, Sharma MK, Murthy P. Digital dilemmas and existing recommendations for healthy screen time use for children and adolescents. Asian J Psychiatr 2024 Sep;99:104154. [doi: 10.1016/j.ajp.2024.104154] [Medline: 39047352]
49. Rosen LD, Whaling K, Carrier LM, Cheever NA, Rokkum J. The Media and Technology Usage and Attitudes Scale: an empirical investigation. Comput Human Behav 2013 Nov 1;29(6):2501-2511. [doi: 10.1016/j.chb.2013.06.006] [Medline: 25722534]
50. Gámez-Guadix M, Villa-George F, Calvete E. Psychometric properties of the Cyberbullying Questionnaire (CBQ) among Mexican adolescents. Violence Vict 2014;29(2):232-247. [doi: 10.1891/0886-6708.vv-d-12-00163r1] [Medline: 24834745]
51. Kuss DJ, Lopez-Fernandez O. Internet addiction and problematic internet use: a systematic review of clinical research. World J Psychiatry 2016 Mar 22;6(1):143-176. [doi: 10.5498/wjp.v6.i1.143] [Medline: 27014605]
52. Monkman H, Kushniruk AW, Barnett J, Borycki EM, Greiner LE, Sheets D. Are health literacy and eHealth literacy the same or different? Stud Health Technol Inform 2017;245:178-182. [Medline: 29295077]
53. Christensen MA, Bettencourt L, Kaye L, et al. Direct measurements of smartphone screen-time: relationships with demographics and sleep. PLoS One 2016;11(11):e0165331. [doi: 10.1371/journal.pone.0165331] [Medline: 27829040]
54. Johannes N, Meier A, Reinecke L, et al. The relationship between online vigilance and affective well-being in everyday life: combining smartphone logging with experience sampling. Media Psychol 2021;24(5):581-605. [doi: 10.1080/15213269.2020.1768122]
55. Aubourg T, Demongeot J, Renard F, Provost H, Vuillerme N. Association between social asymmetry and depression in older adults: a phone call detail records analysis. Sci Rep 2019 Sep 18;9(1):13524. [doi: 10.1038/s41598-019-49723-8] [Medline: 31534178]

Abbreviations

AI: artificial intelligence
DT: digital technology
ExWAS: exposome-wide association study
SDOH: social determinants of health

Edited by A Mavragani; submitted 22.Dec.2025; peer-reviewed by A Corlateanu, S Kuoppamaki; revised version received 05.Mar.2026; accepted 30.Mar.2026; published 08.May.2026.

Please cite as:
Petit P, Vuillerme N
The Digital Exposome: A Life Course Framework for Health in the Digital Age
J Med Internet Res 2026;28:e90153
URL: https://www.jmir.org/2026/1/e90153
doi:10.2196/90153

XSL•FO
RenderX

© Pascal Petit, Nicolas Vuillerme. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Co-Lifecycle Governance for Learning Medical AI: A Hybrid Convergence Framework for Adaptive Regulatory Oversight

Jae Hyun Lee[1], JD, DDS; Boram Choi[1], DDS, PhD; Kwunho Jeong[1], BS; Sang Won Suh[2], MD, PhD; Ju Han Kim[3], MD, PhD; Dae-Soon Son[4,5,6], PhD

[1]Global Research Center, JNPMEDI, Seoul, Republic of Korea

[2]Department of Physiology, College of Medicine, Hallym University, Chuncheon, Gangwon-do, Republic of Korea

[3]Division of Biomedical Informatics, Seoul National University Biomedical Informatics (SNUBI), College of Medicine, Seoul National University, Seoul, Republic of Korea

[4]Major in Bio-Healthcare Convergence, College of Natural Sciences, Hallym University, 1 Hallymdaehak-gil, Chuncheon, Gangwon-do, Republic of Korea

[5]Division of Big Data and Artificial Intelligence, Institute of New Frontier Research, College of Medicine, Hallym University, Chuncheon, Gangwon-do, Republic of Korea

[6]Hallym AI-BioHealth R&BD Center, Research Institute of Medical-Bio Convergence, Hallym University, Chuncheon, Gangwon-do, Republic of Korea

Corresponding Author:
Dae-Soon Son, PhD
Major in Bio-Healthcare Convergence, College of Natural Sciences, Hallym University, 1 Hallymdaehak-gil, Chuncheon, Gangwon-do, Republic of Korea

Abstract

Artificial intelligence (AI) in health care is increasingly defined not by static algorithms but by adaptive intelligence—systems that evolve over time through interactions with data, clinicians, and clinical environments. This adaptive capacity creates a structural mismatch with regulatory frameworks built for technologies whose behavior remains static. As AI models drift, recalibrate, or degrade in real-world contexts, they dissolve the linear boundaries between design, deployment, and clinical interpretation. These temporal, epistemic, and organizational frictions expose responsibility gaps that cannot be resolved through incremental modifications to legacy oversight structures. Regulators across major jurisdictions are beginning to respond to these challenges, though with differing orientations. The United States advances mechanisms for predictable adaptation, including Predetermined Change Control Plans, real-world evidence frameworks, and life cycle–oriented quality management reforms. The European Union emphasizes precautionary, rights-based governance through the European Union Artificial Intelligence Act (AI Act) and modernized liability rules. South Korea, operating within a hyperconnected digital health ecosystem, has introduced the Digital Medical Products Act (DMPA), one of the world's first comprehensive statutory frameworks for learning medical AI. Despite philosophical differences, these regulatory trajectories converge on a shared insight: learning AI systems cannot be governed by static rules or episodic evaluation. This viewpoint proposes Co-Lifecycle Governance as a conceptual framework to synchronize regulatory oversight with adaptive intelligence. Rather than treating oversight as a discrete event, Co-Lifecycle Governance frames regulation as a continuous, synchronized process grounded in 4 pillars: continuous validation, agile change management, proactive performance surveillance, and distributed accountability. Each pillar functions as a structural antidote to the responsibility frictions that arise when AI systems evolve faster than expectations surrounding them. Together, these pillars provide a governance grammar capable of supporting safe, iterative model improvement while maintaining system-level trust. Drawing from the strengths of US predictability, European Union accountability, and Korean scalability, this paper outlines a hybrid convergence pathway that synthesizes predictability, accountability, and operational feasibility. Learning AI will not wait for governance to catch up; oversight must evolve in lockstep with adaptive intelligence. Co-Lifecycle Governance offers a foundation for regulatory systems that not only regulate learning AI but also learn with it—at the speed at which adaptive intelligence actually changes.

(J Med Internet Res 2026;28:e90654) doi:10.2196/90654

KEYWORDS

medical artificial intelligence; medical AI; adaptive artificial intelligence; adaptive AI; learning health systems; Co-Lifecycle Governance; regulatory science; United States Food and Drug Administration Predetermined Change Control Plan; European Union AI Act; Digital Medical Products Act; hybrid governance

XSL•FO
RenderX

Introduction

Medical artificial intelligence (AI) is no longer a fixed device; it is a learning system whose behavior shifts as its data, context, and interactions change. Traditional medical technologies move through segmented and predictable phases, whereas learning AI operates through continuous feedback loops. Evidence from explainability research, clinical deployment challenges, unintended consequences, big data fragility, ethical risks, calibration drift, adversarial vulnerabilities, and explainability limitations demonstrates that learning systems do not maintain a single fixed "performance state" [1-8].

Throughout this viewpoint, we use "learning medical AI" as an umbrella term for AI-enabled medical software whose behavior can change after deployment through planned updates, recalibration, or context-dependent drift. This does not imply autonomous real-time online learning at the bedside, which is neither assumed nor advocated in this framework. In many medical domains, definitive outcome labels are delayed ("ground truth lag"), which makes life cycle governance dependent on staged reassessment rather than continuous real-time outcome learning. Accordingly, throughout this viewpoint, we use "continuous" strictly in a life cycle sense—ongoing, repeated, and trigger-based—rather than as uninterrupted real-time outcome learning.

This paper begins from a consequential insight: the life cycle of a medical product and the learning cycle of an AI system are structurally distinct yet increasingly intertwined. Embedding AI into clinical environments reshapes its behavior in ways that static regulatory snapshots cannot adequately represent. As a result, foundational governance questions arise: Who is responsible when models drift? How should updates be reviewed? Which evidence should guide continuous performance assessment?

Regulators worldwide are experimenting with different answers. The United States is developing structured pathways for predictable adaptation, such as Predetermined Change Control Plans (PCCPs) [9-11]. The European Union (EU) emphasizes precautionary accountability through the AI Act and accompanying liability reforms [12,13]. Korea has enacted the DMPA, establishing a statutory foundation for adaptive oversight within a highly digitized health care ecosystem [14].

Taken together, these regulatory responses point to a shared structural realization: learning medical AI cannot be effectively governed through static rules or episodic review. While each jurisdiction approaches the problem from a distinct regulatory philosophy—adaptive flexibility in the United States, precautionary accountability in the EU, and infrastructural hybridization in Korea—all 3 are converging on the need for oversight systems that remain responsive to ongoing model evolution. These systems share the same problem statement, but they operationalize it through different regulatory logics. This convergence, coupled with persistent divergence in implementation, creates the conditions for a governance framework that can integrate adaptability, accountability, and operational feasibility within a single life cycle–aware structure.

This viewpoint proposes Co-Lifecycle Governance, a framework that integrates regulatory, organizational, and technical responsibilities to synchronize oversight with adaptive intelligence. Rather than replicating any single jurisdictional model, Co-Lifecycle Governance articulates a life cycle–aware structure capable of operationalizing the core requirements embedded across PCCP-based update governance, risk-based AI regulation, and statutory digital medical product oversight.

Regulatory DNA of Learning AI: Divergent Philosophies and Converging Pressures

Regulatory systems designed for fixed-performance technologies were never built to follow AI that continues to change itself after approval. What distinguishes learning AI is not only its evolving performance but also the mechanisms through which that evolution emerges—data drift, contextual adaptation, workflow coupling, and distributed system interactions. These dynamics force regulators to articulate what kinds of change should be permitted, anticipated, monitored, or constrained. The result has been the crystallization of 3 regulatory archetypes, each representing a deeper institutional philosophy about uncertainty, risk, and technological evolution. Although they often appear as regional policy variations, they are better understood as expressions of distinct regulatory "genotypes": different ways of encoding how a governance system adapts to the presence of learning systems.

In this context, "regulatory DNA"—previously articulated in comparative form between the United States and the EU [15]—does not refer to a normative ideal or a shared philosophical blueprint. Rather, it describes the structural response patterns that different regulatory systems have activated in reaction to the same underlying pressure: the inability of fixed, static regulatory models to govern learning systems that continue to change after deployment. While the United States, the EU, and Korea exhibit divergent regulatory forms, they are responding to a common constraint—the need to govern AI across its life cycle rather than at a single point in time.

United States: Predictable Adaptation Through Structured Flexibility

The United States has taken the clearest steps toward treating model evolution not as a postmarket anomaly but as a governable life cycle property. The United States Food and Drug Administration (FDA)'s introduction of PCCPs [9] represents a foundational shift in regulatory logic: manufacturers may prespecify which aspects of a model will change, how they will change, and what validation evidence will be required before updates are deployed. This preauthorization mechanism acknowledges that learning systems cannot be frozen at the moment of approval. Instead, it reframes updates as predictable events within a predefined envelope of acceptable change.

This philosophy has roots in the FDA's long-standing Total Product Life Cycle approach, which emphasizes that the safety of software-based products is shaped not only by design but also by postmarket performance and the manufacturer's quality processes. The modernization of the Quality Management

XSL•FO
RenderX

System Regulation [11] reinforces this orientation by embedding continuous monitoring, process discipline, and organizational accountability into regulatory expectations. Together with the International Medical Device Regulators Forum's foundational software as a medical device (SaMD) frameworks (definitions, risk categorization, and clinical evaluation) [16-18], these instruments operationalize a life cycle–thinking mindset in which regulators oversee systems in motion, not artifacts at rest.

Importantly, the US governance approach also relies on real-world evidence (RWE) to monitor performance and validate changes [10]. RWE offers regulators a dynamic evidentiary substrate for detecting drift, inequities, or clinically significant shifts that may not appear during premarket testing. This integration of real-world learning into regulatory oversight represents a pragmatic recognition: learning AI is safer when regulators explicitly permit controlled evolution rather than forcing manufacturers to choose between compliance and improvement.

Still, the US approach carries trade-offs. Its emphasis on developer responsibility and on organizational quality systems places substantial trust in manufacturers to monitor themselves. Moreover, while PCCPs offer clarity about planned updates, they provide less formal structure for unplanned evolution arising from distributional drift or environmental coupling—precisely the kinds of changes most likely to generate risk. Nevertheless, as an archetype, the US model contributes a vital regulatory virtue: predictable adaptation, which aligns closely with the Co-Lifecycle Governance pillar of agile change management.

EU: Precautionary Governance Anchored in Accountability

The EU approaches learning AI through a fundamentally different regulatory lens. Where the United States structures change, the EU seeks to control risk and ensure rights. The AI Act [12] is built on a precautionary logic: high-risk AI systems—especially those used in medical contexts—must undergo extensive ex ante conformity assessment, maintain comprehensive technical documentation, demonstrate risk mitigation measures, and preserve human oversight throughout their life cycle. The premise is not that AI will necessarily evolve safely, but that institutions must be able to reconstruct and audit system behavior at every stage.

This approach reflects a long-standing European regulatory tradition: documentation as governance. By requiring detailed data provenance records, model design dossiers, transparency reports, and continuous postmarket monitoring plans, the AI Act constructs an auditable trail that can reveal how a model changes, how those changes were validated, and how they may contribute to harm. In this sense, the EU does not primarily regulate learning as a dynamic engineering process; rather, it regulates the evidence environment needed to hold systems accountable.

The updated Product Liability Directive (PLD) [13] further reinforces this architecture by expanding liability to software-driven harms and update-related failures. Under the new PLD, manufacturers may face strict liability if insufficient documentation prevents courts from determining how a failure occurred. This makes transparency not only a compliance requirement but also a legal survival mechanism. If a model's learning trajectory is not well documented, its developer becomes more—not less—exposed.

Yet the EU approach has its own trade-offs. Its strong ex ante compliance burden can slow down adaptation, creating tension for AI systems that rely on frequent updates to remain safe or clinically relevant. The AI Act permits modifications, but the pathways for high-risk systems remain comparatively rigid, requiring additional documentation and potential reassessment. At scale, this may create a mismatch between the speed of learning and the velocity of regulatory processes.

Still, the EU model contributes a distinct regulatory strength that no other jurisdiction provides as strongly: rights-based accountability. It articulates a governance stance in which AI evolution is allowed only when it remains transparent, contestable, and traceable. In Co-Lifecycle Governance terms, the EU strengthens the pillar of distributed accountability, ensuring that oversight does not simply trust manufacturers but constrains them through enforceable obligations.

Korea: Hybridization in a Hyperconnected Ecosystem

South Korea offers a regulatory trajectory shaped not only by policy choice but also by structural conditions that make fixed, static oversight particularly difficult to sustain. As one of the world's most digitally connected health care environments—with near-universal electronic health record adoption, integrated claims databases, nationwide broadband infrastructure, and high digital literacy—Korea operates in a setting where AI-enabled medical products are likely to evolve rapidly once deployed. In such an environment, regulatory frameworks that rely solely on premarket authorization or episodic review are inherently mismatched to real-world use. The DMPA [14,19] emerges from this context as one of the first statutory frameworks explicitly designed to govern digital medical technologies, including adaptive AI, across their operational life cycle.

In this sense, the DMPA does not simply borrow from US or European regulatory traditions; it reflects a convergence driven by necessity. Korea's regulatory environment faced simultaneous pressures: the need to accommodate iterative software updates without stalling clinical innovation, and the need to ensure traceability, documentation, and enforceability in a highly data-intensive health system. Elements commonly associated with US governance—such as acceptance of planned postmarket modification and structured update pathways—are reflected in the Ministry of Food and Drug Safety (MFDS) guidance emphasizing proportionality and predefined validation of changes [20]. At the same time, the DMPA incorporates features characteristic of European regulation, including explicit documentation requirements and auditable risk management structures, which are essential in a system where large-scale data integration amplifies both benefit and harm.

XSL•FO
RenderX

Korea's broader Digital Health Innovation Strategy [21] further reinforces this hybridization by embedding AI governance within national health care modernization. Institutional coherence across the MFDS, the Health Insurance Review and Assessment Service, the National Institute of Medical Device Safety Information, and clinical providers creates conditions under which postmarket oversight can be coordinated at scale. Importantly, this infrastructural strength should not be interpreted as implying continuous raw-data access for private developers to retrain models. Rather, Korea's centralized claims and safety-monitoring systems primarily enable mediated, privacy-preserving surveillance and audit mechanisms. Under such an approach, population-level data can support regulator- or institution-led monitoring, signal detection, and periodic revalidation without necessitating unrestricted data transfer to manufacturers. In this sense, the infrastructure strengthens the learning cycle of oversight and recalibration. It does not imply autonomous, developer-driven continuous model updating. At the same time, centralized oversight architectures introduce their own governance trade-offs, including heightened privacy sensitivity, potential centralization risks, and institutional dependency. These constraints should be recognized as part of the design conditions for scalable life cycle governance, not as automatic advantages. The boundary between medical AI and broader health care AI is particularly salient in Korea's fast-moving digital ecosystem. Tools that begin as wellness, administrative, or population-health software (eg, lifestyle coaching, symptom checkers, appointment triage, or operational risk stratification) may become "medical" AI once their intended use shifts toward diagnosis, treatment selection, or clinical risk management. For learning systems, this boundary can evolve over time as workflow integration deepens, creating "category drift," which challenges static regulatory categorization and reinforces the need for life cycle–aware governance.

A practical Korean example of boundary management can be seen in the MFDS's 2020 approval of Samsung Electronics' blood pressure measurement mobile app as a SaMD [22]. Rather than treating the broader mobile platform as a medical device in its entirety, the MFDS attached device oversight to the specific software function and its intended use: once the app claimed noninvasive measurement and display of blood pressure and pulse, it became subject to medical device performance standards and approval [22]. Although this case is not itself a postdeployment learning event, it demonstrates the operative Korean rule that regulatory status follows function and intended use rather than the digital platform label alone. For adaptive AI, the same logic implies that category drift can occur when iterative updates or deeper workflow integration transform a health-management tool into clinically actionable measurement, triage, or treatment-support software.

These structural conditions directly shape how the 4 pillars of Co-Lifecycle Governance are operationalized in Korea. Surveillance and agile change management are most visibly enabled by Korea's infrastructure, but the remaining pillars are equally implicated. Continuous validation becomes practicable because performance can be reassessed using real-world population data rather than isolated device-level follow-up studies. Distributed accountability is reinforced by the ability to trace responsibility across developers, deploying institutions, and regulators through unified data and documentation pathways. Rather than emphasizing a single pillar, Korea's contribution lies in demonstrating how continuous validation, change management, surveillance, and accountability can operate simultaneously within one statutory and infrastructural ecosystem.

The DMPA is still evolving, and implementation challenges remain. Yet Korea's approach illustrates a form of operational feasibility that neither US flexibility nor European accountability alone can guarantee. Where the US model excels at structuring change and the EU model excels at enforcing responsibility, Korea shows how life cycle–aware governance can be instantiated at scale. In this respect, Korea's experience is not merely nationally specific but offers a transferable reference for health systems seeking to align regulatory oversight with the realities of learning AI in data-rich environments.

Convergence

Despite philosophical divergence, the regulatory genotypes expressed in the US, EU, and Korean frameworks converge on a shared structural recognition: learning AI cannot be governed through fixed rules designed for static technologies. Whether through adaptation envelopes, rights-based documentation regimes, or infrastructural hybridization, each jurisdiction is articulating mechanisms for continuous oversight, proportional risk management, and distributed responsibility. Beyond national regulation, international bodies have similarly emphasized the need for continuous governance, responsible data stewardship, and life cycle–aware oversight in digital health and AI systems [23,24]. These 3 regulatory archetypes are summarized in Table 1.

XSL•FO
RenderX

Table . Comparative regulatory DNA of the United States, European Union, and Korea.

	United States	European Union	Korea
Philosophy	Predictable adaptation	Precautionary accountability	Hybrid operationalization
Key regulatory instruments	PCCP[a], RWE[b], and QMSR[c]	AI Act[d] and PLD[e]	DMPA[f]
Oversight mode	Preauthorized change	High-risk controls	Mixed adaptive pathways
Liability	Tort[g] + regulatory	Strict liability	Hybrid evolving
Strength	Predictability	Accountability	Scalability

[a]PCCP: Predetermined Change Control Plan.
[b]RWE: real-world evidence.
[c]QMSR: Quality Management System Regulation.
[d]AI Act: European Union Artificial Intelligence Act.
[e]PLD: Product Liability Directive.
[f]DMPA: Digital Medical Products Act.
[g]Tort refers to civil liability for wrongful acts or omissions, including negligence or product-related harm, outside contractual obligations.

The Erosion of Linear Responsibility: Structural Frictions in Learning AI

As AI systems evolve during real-world use, responsibility across developers, clinicians, institutions, and regulators becomes unstable. Drift, opacity, and adversarial vulnerabilities erode the linear chain linking design to deployment [1,4-8]. Responsibility gaps arise from three structural frictions:

1. A temporal friction, in which performance changes after approval in ways not reflected in initial validation [6]
2. An epistemic friction, in which failures become invisible to clinicians due to opacity and design constraints [1,7,8]
3. An organizational friction, driven by distributed control across developers, vendors, institutions, and regulators [25-28]

These frictions produce responsibility gaps and "accountability overload," in which responsibility is simultaneously diluted across multiple actors and disproportionately concentrated on those least able to influence system behavior. Learning AI thus destabilizes legacy responsibility doctrines that rely on static role allocation and linear chains of control. Addressing this breakdown requires a governance structure capable of tracking responsibility across time, reallocating obligations as systems evolve, and maintaining traceability between technological change and regulatory accountability. The resulting life cycle responsibility matrix is summarized in Table 2.

Table . Responsibility matrix across the life cycle of learning artificial intelligence systems.

	Developer	Institution	Clinician	Regulator
Design	Architecture	—[a]	—	Standards
Deployment	Documentation	Integration quality	Correct use	Authorization
Real-world learning	Patch cycles	Monitoring	Interpretation	Surveillance
Failure	Root cause	Incident detection	Reporting	Enforcement
Update	Validation	Rollout oversight	Communication	PCCP[b] and DMPA[c] rules

[a]No primary responsibility is assigned to that actor at the corresponding life cycle stage.
[b]PCCP: Predetermined Change Control Plan.
[c]DMPA: Digital Medical Products Act.

These structural frictions point to a deeper limitation of existing governance approaches. Learning medical AI systems do not merely introduce new risks; they alter the temporal and organizational conditions under which responsibility is exercised. Addressing these challenges therefore requires a governance framework that treats technological change as a core design feature rather than an exception to be managed after deployment. The central task is not only to allocate responsibility when failures occur but also to synchronize regulatory oversight with the ongoing evolution of learning systems.

Co-Lifecycle Governance: 4 Pillars for Learning Systems

Co-Lifecycle Governance operationalizes the coevolutionary relationship between learning medical AI and regulatory oversight through 4 interdependent pillars. Rather than functioning as isolated controls, these pillars work together to ensure that adaptation remains both governable and accountable across the product life cycle, as illustrated in Figure 1.

XSL•FO
RenderX

Figure 1. Co-Lifecycle Governance framework for learning medical artificial intelligence (AI). This figure illustrates the parallel and interconnected evolution of medical AI systems and regulatory oversight across the product life cycle.

The first pillar, continuous validation, addresses temporal instability introduced by learning systems. Because performance and calibration may shift as data distributions and clinical workflows evolve, validation cannot remain confined to premarket evaluation. However, "continuous" validation is not synonymous with real-time outcome learning. In settings with delayed end points (eg, oncology or chronic disease), validation must operate across time horizons: (1) short-horizon monitoring of input drift, uncertainty, and process indicators; (2) intermediate validation using proxy outcomes, chart-review audits, or surrogate markers; and (3) longer-horizon confirmation once definitive outcomes mature. This staged structure allows governance to respond to drift early while preserving scientific rigor when ground truth is lagged [2,4,6]. Continuous validation reframes regulatory assurance as an ongoing life cycle obligation, not a one-time premarket certification event. Such life cycle–based oversight inevitably entails financial, technical, and organizational burdens, which must be proportionately allocated across developers, institutions, and regulators. In practice, these burdens may be addressed through structured cost-sharing mechanisms. For example, multi-institutional data-sharing consortia can distribute validation costs across participating sites, while subscription-based service models may incorporate ongoing surveillance and validation as part of software maintenance. In some settings, public infrastructure may support baseline monitoring functions, reducing the burden on individual institutions.

The second pillar, agile change management, structures how updates are proposed, validated, authorized, communicated, and rolled out. Mechanisms such as PCCPs and risk-proportionate pathways under the DMPA exemplify how model evolution can be made transparent, reviewable, and predictable [9,14]. Importantly, agility must be paired with version governance to avoid inequities across care environments. Co-Lifecycle Governance therefore treats version transparency (who is running which model, with what validated performance), minimum performance floors, and update dissemination policies as ethical and operational requirements rather than mere engineering choices.

The third pillar, proactive performance surveillance, responds to the epistemic opacity of learning AI by providing early warning signals. Surveillance is primarily passive and signal-oriented: it continuously tracks aggregate and subgroup performance proxies (eg, drift metrics, calibration signals, alert rates, and incident reports) to identify anomalies that warrant investigation [1,7,8]. By design, surveillance does not "reprove"

XSL•FO
RenderX

effectiveness; it detects deviations and triggers targeted validation when thresholds are crossed. This operational distinction clarifies why surveillance and validation can share data infrastructures yet remain functionally nonredundant.

Finally, distributed accountability aligns obligations with actual control across developers, deploying institutions, clinicians, and regulators. Here, "distributed" refers to operational responsibility (who monitors, investigates, communicates, and executes mitigations), not the dilution of legal liability. In strict liability regimes, financial liability may remain concentrated on manufacturers, but operational accountability must still be traceable across the sociotechnical system. Co-Lifecycle Governance therefore emphasizes traceability: a chain of custody for model versions, validation evidence, deployment context, and corrective actions, which supports both clinical trust and legal enforceability [25-28]. This distinction preserves compatibility with strict liability regimes, where enforceability depends on clear forensic attribution. This structure also remains compatible with the learned intermediary doctrine, under which clinicians retain final decision-making authority in patient care.

For this structure to remain clinically viable, explainability must function as a deployment and update authorization requirement rather than an optional design feature. Initial deployment and any material recalibration or postmarket update must be conditioned on evidence that the system provides interpretable outputs, uncertainty signals, and traceable decision pathways sufficient for clinicians to understand the basis, confidence, and limits of model recommendations within the relevant workflow. Continuous validation must reassess these explainability properties after substantial model or interface changes, and performance surveillance should treat systematic clinician confusion, abnormal override patterns, or unexplained subgroup performance shifts as safety signals. Where these conditions cannot be met, the system should not be deployed or should be restricted to narrower use contexts, because clinician authority would otherwise be reduced to nominal endorsement rather than meaningful mediation. In this sense, explainability is not merely desirable; it is a governance gate that preserves both clinical trust and the practical viability of the learned intermediary doctrine.

Taken together, these 4 pillars provide coordinated answers to the structural frictions identified earlier: continuous validation mitigates temporal drift, performance surveillance addresses epistemic opacity, and distributed accountability—working in concert with agile change management—resolves organizational fragmentation. In combination, they form a life cycle–aware governance architecture capable of synchronizing technological evolution with regulatory expectations.

A robust life cycle governance framework must also specify an exit strategy. When surveillance and validation indicate persistent, nonmitigable risk, governance should support staged controls: conditional suspension (use restricted to defined contexts), rollback to a prior validated version, and decommissioning or recall when risks cannot be acceptably reduced. We position this "safe exit" protocol as a necessary complement to update pathways, ensuring that life cycle governance is not biased toward improvement alone.

However, executing rollback or decommissioning in clinical environments is not frictionless. Technical dependencies between AI systems and electronic health record infrastructures may complicate rapid reversion to prior versions. Abrupt changes in system behavior may also disrupt clinical workflows and introduce new safety risks, particularly if clinicians have adapted to the updated system. These constraints highlight the need for preplanned and institutionally integrated exit strategies. Table 3 maps the 4 pillars of Co-Lifecycle Governance to representative regulatory anchors in the United States, EU, and Korea.

Table . Mapping the 4 pillars of Co-Lifecycle Governance to regulatory anchors.

Pillar	United States	European Union	Korea
Continuous validation	RWE[a] and QMSR[b]	AI Act[c] documentation	DMPA[d] performance tracking
Agile change management	PCCP[e]	Change subject to conformity assessment and recertification triggers	Adaptive update pathways
Surveillance	RWE	Mandatory monitoring	National-scale infrastructure
Accountability	Shared	Strict liability	Hybrid

[a]RWE: real-world evidence.

[b]QMSR: Quality Management System Regulation.

[c]AI Act: European Union Artificial Intelligence Act.

[d]DMPA: Digital Medical Products Act.

[e]PCCP: Predetermined Change Control Plan.

Because continuous validation and performance surveillance both rely on postmarket data streams, Table 4 clarifies their distinct operational roles.

XSL•FO
RenderX

Table . Operational comparison between continuous validation and performance surveillance.

Dimension	Continuous validation (pillar 1)	Surveillance (pillar 3)
Primary purpose	Confirm or update performance claims	Detect early warning signals
Mode	Active, hypothesis-driven testing	Passive, continuous monitoring
Evidence standard	Higher (reference standards and audits)	Preliminary (proxies and alerts)
Typical data	Curated labeled sets, chart review, surrogate outcomes, and matured end points	Drift metrics, calibration signals, alert rates, incident reports, and subgroup signals
Trigger	Scheduled revalidation or triggered by surveillance	Always on and triggers validation when thresholds are crossed
Output	Revalidation report, updated performance claims, and recalibration decision	Alerts/flags, monitoring reports, and escalation tickets
Decision action	Approve recalibration or update or safe exit	Escalate to validation and change control

Illustrative Clinical Scenario

To illustrate how the proposed framework operates in practice, consider a sepsis prediction algorithm deployed in an intensive care unit.

Over time, clinicians begin to observe fewer alerts for high-risk patients, potentially increasing the risk of delayed intervention in deteriorating patients. Concurrently, performance surveillance systems detect a decline in calibration across specific patient subgroups (pillar 3), triggering an institutional alert.

This signal initiates targeted validation (pillar 1), in which recent patient data and chart review audits are used to assess whether the observed drift reflects true performance degradation. Upon confirmation, the developer implements a recalibration update under a prespecified change management plan (pillar 2), with updated performance metrics communicated to clinicians prior to deployment.

Throughout this process, version tracking and audit logs ensure traceability of decisions and responsibility allocation (pillar 4). If recalibration fails to restore acceptable performance, the system may enter a staged exit pathway, including conditional suspension or rollback.

Hybrid Convergence: Synthesizing Predictability, Accountability, and Scale

As previously analyzed in comparative context [15], the United States contributes predictable adaptation through PCCPs, RWE, and life cycle–oriented quality systems [9-11,16-18]. The EU contributes enforceable accountability through risk-based governance and liability reform under the AI Act and PLD [12,13]. Korea contributes scalable operationalization through statutory digital medical product frameworks and nationally coordinated digital health infrastructure [14,19-21].

Rather than proposing a single harmonized regulatory procedure, the hybrid convergence should be understood as a governance grammar—a set of life cycle concepts and operational primitives that can be instantiated within different legal traditions. In practice, manufacturers operate under one primary jurisdiction at a time, while global developers may translate shared governance artifacts (eg, version traceability records, validation documentation, and monitoring plans) into jurisdiction-specific compliance packages. Figure 2 illustrates this convergence pathway: distinct regulatory logics feed into a shared life cycle–oriented governance architecture without requiring legal harmonization.

This framing addresses the apparent friction between rapid iteration and precautionary compliance. US-style preauthorized change pathways enable structured model updates, whereas the EU high-risk regime emphasizes conformity assessment, documentation, and recertification triggers. Co-Lifecycle Governance does not claim to eliminate these compliance burdens. Instead, it renders them governable by (1) classifying changes according to risk and regulatory impact, (2) prespecifying validation evidence proportional to each change category, and (3) maintaining audit-ready documentation that preserves traceability across updates. In this sense, agile change management can operate within precautionary systems as a discipline for anticipating and controlling recertification thresholds rather than as a mechanism for unconstrained speed.

The focus on the United States, EU, and Korea reflects their representation of 3 archetypal regulatory logics—predictable adaptation, precautionary accountability, and infrastructural scalability—rather than an exhaustive global survey.

XSL•FO
RenderX

Figure 2. Convergence of archetypal regulatory logics into a hybrid life cycle governance architecture. This schematic figure shows how the United States (predictability), the European Union (EU; accountability), and Korea (scalability) converge toward a shared hybrid life cycle governance architecture.

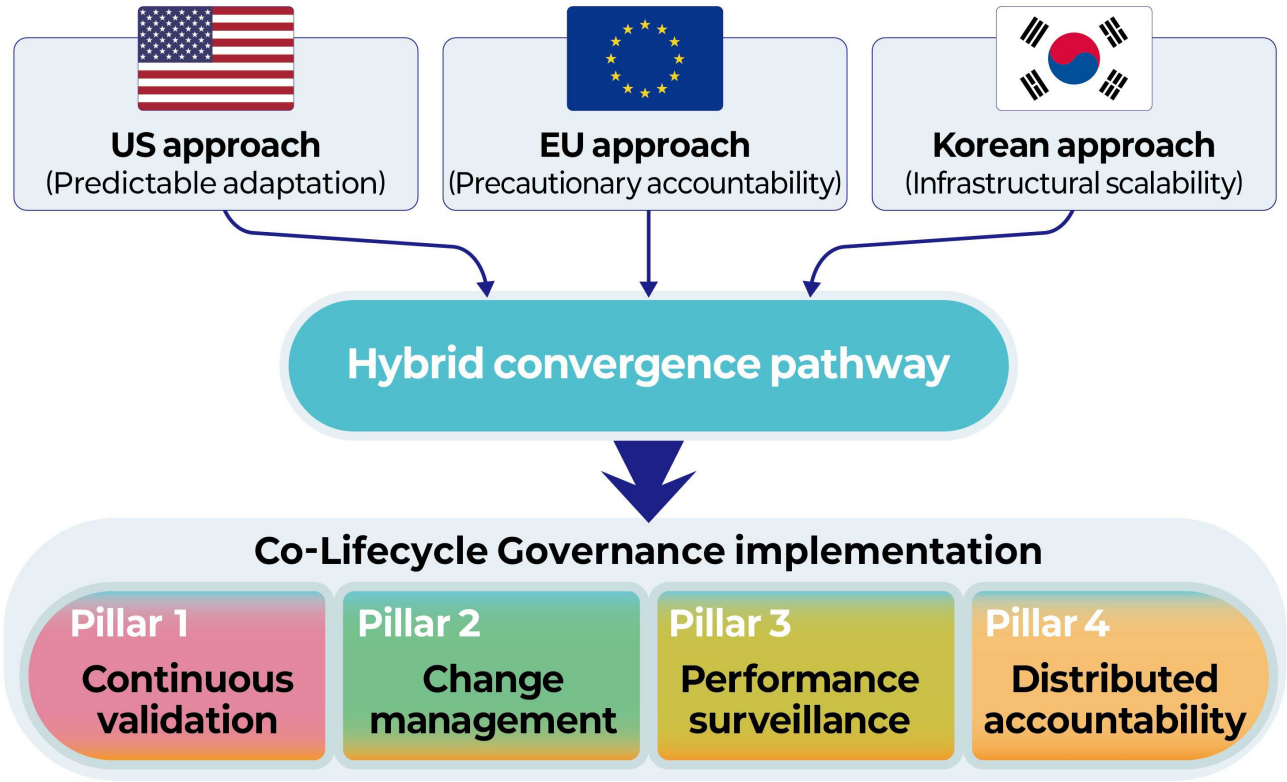

Other jurisdictions, including Japan's postapproval change management approaches, Singapore's life cycle guidance for software medical devices, and China's classification principles for AI-based medical software, demonstrate parallel movement toward life cycle–aware governance. China has also issued classification and technical review guidance for AI-based medical software, emphasizing risk-based categorization and performance evaluation requirements tied to intended use and clinical function. Japan has explored controlled postapproval update pathways, while Singapore emphasizes life cycle monitoring and quality system–based regulatory flexibility. These examples reinforce the broader claim of structural convergence while preserving regional divergence in implementation.

Finally, scalability is not synonymous with simplicity. National infrastructures that enable coordinated surveillance also introduce governance prerequisites and risks, including robust data governance, privacy safeguards, institutional capacity for monitoring, and clear legal pathways for corrective action. Co-Lifecycle Governance treats these as design constraints to be explicitly managed rather than as automatic advantages. Therefore, the hybrid convergence represents not a unified regime but a convergent architecture through which diverse regulatory systems can align life cycle oversight with adaptive intelligence.

In the context of unplanned postmarket drift, regulatory approaches also diverge. In the United States, such events are primarily addressed through manufacturer-led monitoring within quality management systems, with less formalized pathways for unplanned changes outside PCCP-defined updates. In contrast, the Korean framework enables more centralized signal detection through national-scale infrastructure, supporting coordinated revalidation and corrective action. This distinction reflects a broader difference between distributed monitoring and infrastructure-enabled oversight.

Conclusions

Learning AI exposes foundational limits of legacy governance structures. Evidence of drift, opacity, and adversarial vulnerability demonstrates why static oversight cannot ensure safety in dynamic environments [1,4-8]. Co-Lifecycle Governance reframes oversight as a synchronized, adaptive process grounded in 4 structural pillars [2-8,25-28]. These pillars provide a shared language for distributing responsibility and coordinating obligations across the life cycle.

Internationally, predictable adaptation (United States) [9-11], enforceable accountability (EU) [12,13], and operational scalability (Korea) [14,19-21] represent complementary strengths. No single philosophy suffices in isolation. A durable hybrid governance pathway must synthesize these approaches while maintaining public trust.

The pace of adaptive AI development increasingly challenges the capacity of static governance models to respond. Co-Lifecycle Governance offers a foundation for regulatory systems capable of learning and adapting with the technologies they oversee.

XSL•FO
RenderX

Acknowledgments

The authors used generative AI tools (Google's Gemini 3 Flash and OpenAI's ChatGPT [GPT-4–class model]) solely for literature discovery, organizing background notes, and linguistic editing. All content and final interpretations remain the sole responsibility of the authors.

Funding

This research was supported by the Regional Innovation System & Education (RISE) Glocal University 30 Project program through the Gangwon RISE Center, funded by the Ministry of Education (MOE) and the Gangwon State (GS), Republic of Korea (2025-RISE-10-009).

Conflicts of Interest

JHL is the chief of staff at JNPMEDI. KJ is the founder and chief executive officer of JNPMEDI and holds a financial interest in the company. JNPMEDI provides clinical trial data management and related services. The authors emphasize that this viewpoint does not evaluate, endorse, or promote any specific product or service. All other authors declare no other conflicts of interest.

References

1. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. In: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: Association for Computing Machinery; 2016. [doi: 10.1145/2939672.2939778]
2. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019 Oct 29;17(1):195. [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
3. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA 2017 Aug 8;318(6):517-518. [doi: 10.1001/jama.2017.7797] [Medline: 28727867]
4. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018 Apr 3;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
5. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. PLoS Med 2018 Nov;15(11):e1002689. [doi: 10.1371/journal.pmed.1002689] [Medline: 30399149]
6. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc 2017 Nov 1;24(6):1052-1061. [doi: 10.1093/jamia/ocx030] [Medline: 28379439]
7. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science 2019 Mar 22;363(6433):1287-1289. [doi: 10.1126/science.aaw4399] [Medline: 30898923]
8. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1):310. [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
9. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions. U.S. Food and Drug Administration. 2025. URL: https://www.fda.gov/media/166704/download [accessed 2026-05-01]
10. Use of real-world evidence to support regulatory decision-making for medical devices. U.S. Food and Drug Administration. 2025. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-real-world-evidence-support-regulatory-decision-making-medical-devices [accessed 2026-02-15]
11. Quality Management System Regulation (QMSR). U.S. Food and Drug Administration. URL: https://www.fda.gov/medical-devices/postmarket-requirements-devices/quality-management-system-regulation-qmsr [accessed 2026-05-01]
12. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending regulations (EC) no 300/2008, (EU) no 167/2013, (EU) no 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (text with EEA relevance). European Union. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng [accessed 2026-02-15]
13. Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC (text with EEA relevance). European Union. 2024. URL: https://eur-lex.europa.eu/eli/dir/2024/2853/oj/eng [accessed 2026-05-01]
14. Digital Medical Products Act. Ministry of Food and Drug Safety, Republic of Korea. URL: https://www.law.go.kr/LSW/lsInfoP.do?lsiSeq=259299#0000 [accessed 2026-05-01]
15. Lee JH, Choi B, Jeong K, et al. Diverging regulatory DNA in adaptive medical AI: US agility and EU accountability in lifecycle governance. Front Med (Lausanne) 2026;13:1758708. [doi: 10.3389/fmed.2026.1758708]
16. Software as a Medical Device (SaMD): key definitions. International Medical Device Regulators Forum. 2013. URL: https://www.imdrf.org/documents/software-medical-device-samd-key-definitions [accessed 2026-02-15]

XSL•FO
RenderX

17. Software as a Medical Device: possible framework for risk categorization and corresponding considerations. International Medical Device Regulators Forum. 2014. URL: https://www.imdrf.org/documents/software-medical-device-possible-framework-risk-categorization-and-corresponding-considerations [accessed 2026-02-15]
18. Software as a Medical Device (SaMD): clinical evaluation. International Medical Device Regulators Forum. 2017. URL: https://www.imdrf.org/documents/software-medical-device-samd-clinical-evaluation [accessed 2026-02-15]
19. Business guidelines for the implementation of Digital Medical Products Regulations. Ministry of Food and Drug Safety, Republic of Korea. 2025. URL: https://www.mfds.go.kr/brd/m_1060/view.do?seq=15629 [accessed 2026-02-15]
20. Guidelines for the approval and review of digital medical devices applying artificial intelligence technology. Ministry of Food and Drug Safety, Republic of Korea. 2025. URL: https://www.mfds.go.kr/brd/m_1060/view.do?seq=15657 [accessed 2026-02-16]
21. Ministry of Health and Welfare (Republic of Korea). Healthcare 4.0 Era for Healthy Korea. Ministry of Health and Welfare. 2023. URL: https://www.mohw.go.kr/board.es?act=view&bid=0032&list_no=375996&mid=a20401000000 [accessed 2026-02-15]
22. The world's first blood pressure measurement mobile app approved as medical device. Ministry of Food and Drug Safety, Republic of Korea. 2020. URL: https://www.mfds.go.kr/eng/brd/m_61/view.do?seq=9 [accessed 2026-04-22]
23. Ethics and governance of artificial intelligence for health. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240029200 [accessed 2026-02-15]
24. Health data governance for the digital age. Organisation for Economic Co-operation and Development. 2022. URL: https://www.oecd.org/en/publications/health-data-governance-for-the-digital-age_68b60796-en.html [accessed 2026-02-15]
25. Price WN 2nd, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. JAMA 2019 Nov 12;322(18):1765-1766. [doi: 10.1001/jama.2019.15064] [Medline: 31584609]
26. Gerke S, Minssen T, Cohen IG. Ethical and legal challenges of artificial intelligence–driven health care. In: Bohr A, Memarzadeh K, editors. Artificial Intelligence in Healthcare: Academic Press; 2020:295-336. [doi: 10.1016/B978-0-12-818438-7.00012-5]
27. Leslie D. Understanding artificial intelligence ethics and safety: a guide for the responsible design and implementation of AI systems in the public sector. SSRN. Preprint posted online on Aug 18, 2020 URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3403301 [accessed 2026-05-01]
28. Molnár-Gábor F. Implementing the human right to science in the context of health-related data processing. J Law Biosci 2024;11(1):lsae004. [doi: 10.1093/jlb/lsae004] [Medline: 38495856]

Abbreviations

AI: artificial intelligence
AI Act: European Union Artificial Intelligence Act
DMPA: Digital Medical Products Act
EU: European Union
FDA: United States Food and Drug Administration
MFDS: Ministry of Food and Drug Safety
PCCP: Predetermined Change Control Plan
PLD: Product Liability Directive
RWE: real-world evidence
SaMD: software as a medical device

Edited by A Coristine; submitted 31.Dec.2025; peer-reviewed by R Omachonu, T Adachi; revised version received 23.Apr.2026; accepted 23.Apr.2026; published 19.May.2026.

Please cite as:
Lee JH, Choi B, Jeong K, Suh SW, Kim JH, Son DS
Co-Lifecycle Governance for Learning Medical AI: A Hybrid Convergence Framework for Adaptive Regulatory Oversight
J Med Internet Res 2026;28:e90654
URL: https://www.jmir.org/2026/1/e90654
doi:10.2196/90654

© Jae Hyun Lee, Boram Choi, Kwunho Jeong, Sang Won Suh, Ju Han Kim, Dae-Soon Son. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet

XSL•FO
RenderX

Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Frailty-Based Remote Monitoring in Older Adults With Heart Failure: Conceptual Framework for Adaptive Digital Health Strategies

Rémi Esser[1], MD; Olivier Maurou[1], MD; Marine Larbaneix[1], MD; Alejandro Mondragon[1], MD; Marlène Esteban[1], MD; Christine Farges[1], MD; Nicolas Pages[2], MD; Sophie Nisse Durgeat[2], PharmD; Marc Harboun[1], MD

[1]Hôpital La Porte Verte, 6 avenue du Maréchal Franchet d'Esperey, Versailles, Île-de-France, France

[2]NP Medical, Bordeaux, France

Corresponding Author:
Rémi Esser, MD
Hôpital La Porte Verte, 6 avenue du Maréchal Franchet d'Esperey, Versailles, Île-de-France, France

Abstract

Remote monitoring is increasingly used in heart failure care, but most programs rely on uniform models that insufficiently reflect the heterogeneity of older adults, particularly with respect to frailty, cognitive impairment, functional dependency, and caregiver availability. This viewpoint argues that frailty should be considered a central determinant of remote monitoring design in older adults with heart failure, rather than a secondary modifier of conventional digital health pathways. Drawing on evidence from heart failure telemonitoring, geriatric medicine, and real-world cardiogeriatric experience, we propose a frailty-adaptive framework structured around four clinical trajectories: robust, prefrail, frail, and palliative. Each trajectory is associated with distinct monitoring objectives, workflow adaptations, and response pathways. Robust patients may benefit primarily from optimization, self-management support, and guideline-directed therapy titration; prefrail patients from early detection of deterioration and functional decline; frail patients from proxy-supported reporting, nurse-led triage, and rapid-access cardiogeriatric reassessment; and palliative patients from simplified symptom-guided monitoring focused on comfort, hospitalization avoidance, and caregiver support. This framework reframes remote monitoring as a stratified clinical process rather than a purely technological intervention. By aligning digital strategies with frailty status, functional capacity, and care goals, frailty-adaptive remote monitoring may improve clinical relevance, promote digital health equity, and support more sustainable models of care for older adults with heart failure.

(J Med Internet Res 2026;28:e88681) doi:10.2196/88681

KEYWORDS

digital health strategy; remote monitoring; frailty; aging population; heart failure; telemedicine workflow

Introduction

The Misalignment Between Digital Health and Aging Realities

Remote monitoring has become an integral component of heart failure (HF) management, aiming to reduce rehospitalizations and improve chronic disease control. Early telemonitoring trials such as Tele-HF (Telemonitoring to Improve Heart Failure Outcomes), TIM-HF (Telemedical Interventional Monitoring in Heart Failure), and TIM-HF2 (Telemedical Interventional Management in Heart Failure 2) predominantly enrolled relatively younger and less frail populations, limiting the generalizability of their findings to very old adults with multimorbidity and geriatric syndromes. More recent large-scale real-world programs, such as TELESAT-HF, have demonstrated heterogeneous but clinically meaningful results, strongly influenced by program structure, patient selection, and response workflows [1-3].

Much of the evidence base and many telemonitoring workflows were initially developed and validated in populations that were younger and less frail. In parallel, several studies have shown that a substantial proportion of older adults experience "telemedicine unreadiness," driven by cognitive impairment, sensory limitations, and reduced digital literacy [4-6]. Together, these observations suggest that existing telemonitoring models may not be fully aligned with the characteristics and needs of very old and frail populations, in whom HF is closely intertwined with multimorbidity, functional decline, and dependency.

Frailty is a well-established clinical construct reflecting reduced physiological reserve and increased vulnerability to stressors. Seminal work by Fried et al [7] and Rockwood et al [8] defined frailty as a measurable and clinically meaningful condition, later expanded by Clegg et al [9] as a major determinant of health outcomes at the population level. In HF, frailty has been consistently associated with higher risks of hospitalization, disability, and mortality, often outperforming traditional cardiac

XSL•FO
RenderX

markers in prognostic accuracy among older adults [10-13]. At the same time, the global burden of HF continues to rise with population aging [14]. However, frailty remains largely absent from the design of most telemonitoring strategies, including patient selection, alert thresholds, and care pathways.

Although several studies have shown that older adults can successfully engage in telemonitoring when appropriately supported [4-6], adherence, comprehension, and responsiveness vary substantially according to frailty level.

This viewpoint argues that frailty should be considered a primary determinant of remote monitoring design in older adults with HF, rather than a secondary modifier. We propose a structured frailty-adaptive framework based on 4 major clinical trajectories—robust, prefrail, frail, and palliative—each associated with distinct monitoring objectives, clinical responses, and organizational requirements.

By reframing remote monitoring as a stratified, patient-centered strategy rather than a purely technological intervention, this approach aims to improve clinical relevance, optimize resource allocation, and promote digital health equity. Integrating frailty into telemonitoring design is essential to ensure that digital interventions remain accessible, effective, and safe across the full spectrum of aging and vulnerability.

Conceptual Basis of the Framework

Conceptual Basis and Sources

This viewpoint is based on a narrative synthesis of the literature combined with real-world clinical experience from a cardiogeriatric telemonitoring program. To our knowledge, existing telemonitoring approaches do not explicitly integrate frailty into digital health planning for older adults with HF nor link segmentation to practical determinants such as proxy reporting, nurse-led triage, or rapid-access pathways.

Frailty can be assessed using validated tools such as the Fried phenotype or the Clinical Frailty Scale. The 4 profiles proposed in this framework correspond to clinically meaningful frailty trajectories rather than rigid categories, allowing flexibility based on clinical judgment and validated instruments.

Frailty Segmentation Approach

In this framework, frailty is used as an integrative construct reflecting global vulnerability, rather than being limited to digital capability or functional independence.

Frailty should not be interpreted solely as a proxy for digital literacy or autonomy. While some frail individuals may retain the ability to engage with digital tools, functional independence, cognitive status, and caregiver availability represent complementary dimensions that influence telemonitoring feasibility.

Previous studies have demonstrated that HF trajectories in older adults are highly heterogeneous and strongly influenced by frailty and functional status rather than left ventricular ejection fraction alone [9-13,15-17].

Within this perspective, remote monitoring is conceptualized not as a purely technological intervention but as a stratified care strategy, in which monitoring objectives, intensity, and response pathways are adapted to frailty level.

Illustrative Real-World Telemonitoring Program

To support the proposed framework, we describe an illustrative cardiogeriatric telemonitoring program reflecting real-world clinical practice.

This large real-world cardiogeriatric telemonitoring program involved approximately 568 older adults with HF over a median monitoring duration of 18 months within a care pathway that integrates telemonitoring, nurse-led triage, and cardiogeriatric day-hospital reassessment. These figures are provided to contextualize the clinical experience underlying this viewpoint and should not be interpreted as original outcome data.

Patient follow-up relied on a structured symptom-based questionnaire completed twice weekly, focusing on dyspnea, peripheral congestion, weight evolution, fatigue, and other indicators of HF decompensation. When patients were unable to complete the questionnaire themselves—frequent among frail or cognitively impaired individuals—data were entered by family caregivers, home-care nurses, or residential-care staff.

Responses were processed through an automated algorithm generating 3 levels of alerts (green, orange, red) based on predefined clinical thresholds. Alerts were transmitted to a specialized cardiogeriatric nursing team, who performed structured telephone assessments to confirm symptoms and evaluate clinical status.

When indicated, interventions included adjustment of oral diuretic therapy according to predefined protocols or rapid referral to a cardiogeriatric day-hospital pathway.

This workflow illustrates a model in which remote monitoring is supported by algorithm-based alert systems, nurse-led triage, and rapid access to in-person reassessment. The program described here is provided as an illustrative organizational model, rather than a device-specific intervention. The underlying principles are consistent with telemonitoring pathways implemented in France within the ETAPES (Expérimentations de Télémédecine pour l'Amélioration des Parcours en Santé) program and are broadly applicable across different digital health platforms.

The program described here is provided as an illustrative organizational model, rather than a device-specific intervention. The underlying principles are consistent with telemonitoring pathways implemented in France within the ETAPES program and are broadly applicable across different digital-health platforms.

Adaptation to Functional and Digital Limitations

An essential component of this framework is the adaptation of telemonitoring to differences in digital literacy, sensory capacity, and functional autonomy among older adults.

In real-world practice, a substantial proportion of very old and frail patients do not perform digital reporting independently, making caregiver- or professional-supported reporting a key component of coordinated telemonitoring workflows. This shifts remote monitoring from a purely digital interaction to a

XSL•FO
RenderX

coordinated clinical process, where caregivers and nursing teams play a key operational role.

This shifts remote monitoring from a purely digital interaction to a coordinated clinical process, where caregivers and nursing teams play a key operational role.

Positioning telemonitoring as a flexible, caregiver-supported intervention ensures that frailty, cognitive impairment, or limited digital skills do not constitute exclusion criteria. Instead, these characteristics inform program design and organization, enabling telemonitoring to remain feasible, safe, and equitable across the full spectrum of frailty.

The Frailty-Adaptive Remote Monitoring Framework

Overview

The proposed framework defines 4 distinct frailty-based trajectories—robust, prefrail, frail, and palliative—each associated with specific monitoring objectives, workflow adaptations, and clinical responses.

Figure 1 provides a schematic overview of the frailty-adaptive telemonitoring framework, illustrating how patient stratification translates into differentiated monitoring strategies and care pathways.

Figure 1. Frailty-adaptive telemonitoring framework for older adults with heart failure. This figure illustrates how remote monitoring strategies are adapted according to frailty status. After frailty assessment using validated tools (eg, Fried phenotype or Clinical Frailty Scale), patients are stratified into 4 trajectories: robust, prefrail, frail, and palliative. Each profile is associated with specific monitoring objectives, workflow adaptations, and response pathways, ranging from self-monitoring and optimization to nurse-led triage, rapid-access care, and symptom-guided support. Overall, the framework highlights the need for a stratified, patient-centered digital health approach.

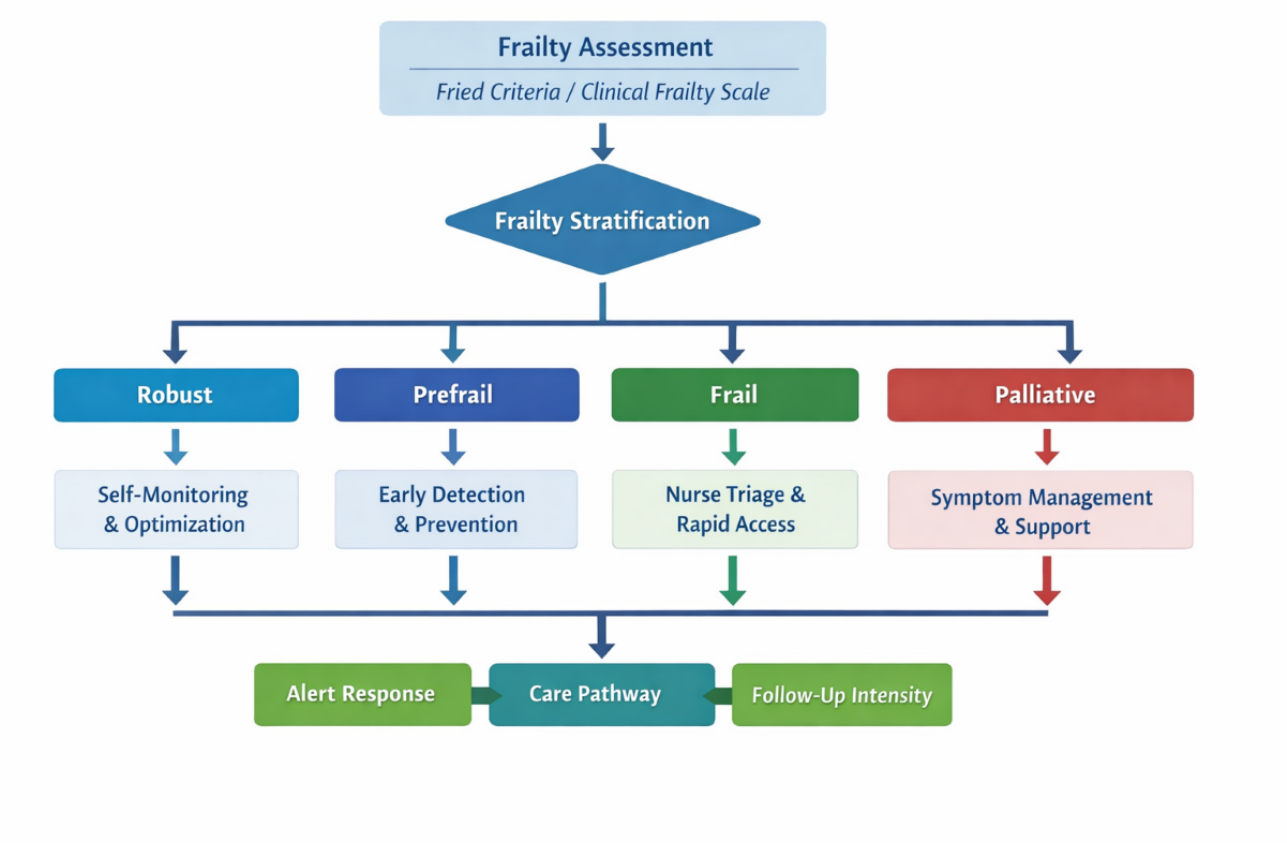

Table 1 provides a structured overview of how remote-monitoring strategies can be aligned with frailty status to optimize clinical relevance, feasibility, and resource allocation.

XSL•FO
RenderX

Table . Frailty-based profiles and their adapted remote-monitoring strategies.

Frailty profile	Clinical characteristics	Primary monitoring objectives	Monitoring intensity & workflow adaptations	Typical interventions triggered
Robust	Preserved mobility, good cognition, low comorbidity burden	Optimization, empowerment, GDMT[a] titration	Standard symptom questionnaire; patient self-reporting; minimal caregiver involvement	Diuretic adjustment, lifestyle advice, GDMT optimization
Prefrail	Early vulnerability, mild decline in gait, fatigue, early cognitive signs	Detection of early deterioration and reduction of risk of functional decline and hospitalization	Slightly increased monitoring frequency; simplified alerts; closer coordination with primary care and nurses	Early titration, nutritional support, mobility interventions, closer follow-up
Frail	Multimorbidity, polypharmacy, functional dependency, cognitive impairment	Crisis avoidance, stabilization, early identification of decompensation	Proxy reporting common (caregiver/nurse); structured triage; rapid-access day-hospital pathway; nurse-led adjustments	IV[b] diuretics, iron repletion, geriatric assessment, medication adjustment
Palliative	Severe frailty, refractory symptoms, limited life expectancy, goal: comfort	Symptom-guided care, hospitalization avoidance, support at home	Highly simplified monitoring; caregiver-driven reporting; alerts focused on comfort/symptom distress rather than disease control	Symptom relief, comfort titration, avoiding ED[c] visits, supporting home-based end-of-life care

[a]GDMT: guideline-directed medical therapy.

[b]IV: intravenous.

[c]ED: emergency department.

Robust Older Adults: Optimization and Prevention

Robust older adults, characterized by preserved mobility, intact cognition, and low comorbidity burden, are the population most closely represented in early telemonitoring trials. In this group, remote monitoring primarily supports optimization of HF management, including early detection of congestion, reinforcement of adherence, lifestyle optimization, and guideline-directed medical therapy titration.

Evidence from trials such as Tele-HF, TIM-HF, TIM-HF2, and MORE-CARE, as well as meta-analyses, suggests that benefits in this population depend on program responsiveness and integration into multidisciplinary HF care [2,3,18,19].

Patients typically perform self-reporting independently, with high adherence and stable engagement. In this context, telemonitoring functions as a preventive strategy aimed at maintaining clinical and functional stability.

Prefrail Patients: Early Detection and Functional Preservation

Prefrail individuals represent a transitional stage characterized by early vulnerability, including mild sarcopenia, reduced gait speed, fatigue, or early cognitive changes.

In this trajectory, the primary role of remote monitoring is the early detection of subtle clinical or functional deterioration. Changes in symptom patterns—such as increased fatigue, reduced activity tolerance, progressive dyspnea for similar exertion, or reduced ability to complete questionnaires independently—may signal early decompensation or loss of functional reserve. Increased caregiver involvement or delayed responses to alerts may also reflect declining autonomy.

Although direct evidence that telemonitoring prevents transition to frailty is lacking, early detection of clinical deterioration and timely intervention may contribute to preserving functional status and reducing hospitalization risk. This supports a model of slightly intensified monitoring, simplified alert systems, and closer coordination with primary care, nursing teams, and caregivers.

Frail Patients: Stabilization and Crisis Avoidance

In frail older adults, HF is embedded within multimorbidity, polypharmacy, cognitive impairment, and functional dependency. These patients experience frequent exacerbations, high hospitalization rates, and reduced ability to engage independently in self-monitoring. Frailty has been consistently identified as a major predictor of hospitalization and mortality in populations with HF [11-13,17,20].

In this group, the effectiveness of telemonitoring depends less on patient digital capacity than on organizational structure. Proxy-assisted reporting by caregivers or health care professionals is common and enables continued participation despite functional limitations [4-6,21,22].

Remote monitoring therefore operates as a stabilization tool within an integrated care pathway. Key components include nurse-led triage, structured clinical verification of alerts, and rapid access to in-person reassessment. Typical interventions include intravenous diuretics, iron repletion, medication adjustment, and comprehensive geriatric assessment. Integrated cardiogeriatric models have demonstrated improved symptom control and reduced hospital utilization in this population [12,22,23].

The primary objective in this trajectory is crisis avoidance through early intervention and rapid care escalation rather than patient-driven self-management.

XSL•FO
RenderX

Palliative Patients: Symptom Control and Home-Based Care

Patients in the palliative trajectory typically present with advanced HF, severe frailty, refractory symptoms, and limited life expectancy. Many express a preference to avoid hospitalization and remain at home. However, in practice, a large proportion still die in hospital, often following acute decompensation events [24,25].

When adapted to patient goals, remote monitoring may support symptom-guided care, early identification of distress, and avoidance of unnecessary emergency department visits. Monitoring strategies in this group are simplified and often rely on caregiver-driven reporting, with alert systems focused on symptom burden rather than disease optimization.

Evidence from palliative telehealth and home-based HF care suggests that such approaches can improve symptom control, reduce acute care use, and support home-based end-of-life care [26-28]. In this trajectory, remote monitoring shifts from disease management to a supportive role aimed at maintaining comfort and enabling care aligned with patient preferences.

These four trajectories illustrate how remote monitoring can be operationalized as a stratified, frailty-responsive care strategy rather than a uniform digital intervention.

Clinical and Health System Implications

This viewpoint proposes a frailty-adaptive framework for remote monitoring in older adults with HF, structured around 4 clinical trajectories with differentiated monitoring objectives, workflows, and clinical responses. This framework provides an operational rather than purely conceptual approach by linking frailty profiles to concrete monitoring pathways, clinical responses, and organizational requirements. It challenges the prevailing model of uniform telemonitoring by positioning frailty as a primary determinant of digital health design rather than a secondary characteristic.

Addressing Common Misconceptions and Barriers

A common misconception is that frail or palliative patients lack the capacity to engage with digital health interventions. In practice, remote monitoring in these populations rarely relies solely on patient-driven input. Data collection and alert responses can be supported by caregivers, home nurses, or health care professionals, and monitoring systems can be simplified or structured around nurse-led workflows. Multiple studies have shown that older adults—including those with frailty or mild cognitive impairment—can successfully participate in telemonitoring when programs are appropriately adapted [4-6,20-22].

Another concern relates to the potential burden on health care teams. However, evidence suggests that structured telemonitoring programs, particularly those integrating early detection and rapid intervention pathways, can reduce hospitalizations and emergency department visits, thereby decreasing overall health care use in high-risk HF populations [2,3,19,23,24,29].

Ethical concerns are also raised regarding the use of telemonitoring in palliative patients. However, studies in palliative telehealth demonstrate that remote monitoring can improve symptom control, reduce distressing hospital transfers, and support alignment with patient preferences, particularly when centered on comfort, shared decision-making, and caregiver support [26-28].

These considerations indicate that telemonitoring is not inherently unsuitable for frail or palliative populations; rather, its effectiveness depends on the extent to which programs are adapted to their specific needs and goals.

Implications for Clinical Practice and Health Systems

Integrating frailty into the design of remote monitoring has direct implications for clinical programs and health systems. From a digital health perspective, this approach shifts telemonitoring from a uniform technological deployment to a stratified system design integrating patient-level heterogeneity. Telemonitoring strategies should be structured according to frailty profiles, with differentiated objectives, alert thresholds, follow-up intensity, and care pathways.

In practice, this requires tiered monitoring models, nurse-led triage, rapid-access pathways, and coordination with caregivers and community-based providers. At the system level, policy frameworks should support the integration of geriatric expertise into HF care, promote multidisciplinary coordination, and incentivize organizational models that link telemonitoring with rapid clinical response [22,23,30,31]. Digital health solutions should allow flexible configuration, including customizable alert thresholds, caregiver interfaces, and simplified reporting processes adapted to varying levels of functional and cognitive capacity.

Digital Equity as a Core Dimension

Digital equity represents a central challenge in the implementation of remote monitoring in aging populations. Older adults frequently face structural barriers, including limited digital literacy, sensory impairments, cognitive decline, and variable access to caregivers. Without adaptation, these factors may lead to the exclusion of the most vulnerable patients from digital health interventions.

A frailty-adaptive approach addresses these challenges by embedding equity into program design. Key elements include simplified interfaces, proxy-supported data entry, educational support tailored to functional capacity, and monitoring workflows that do not rely exclusively on patient autonomy.

Positioning Within Existing Literature

The effectiveness of telemonitoring in HF has been highly variable across studies, largely due to differences in program design, patient selection, and response workflows. By explicitly

XSL•FO
RenderX

incorporating frailty into these dimensions, the proposed framework provides a potential explanation for part of this heterogeneity.

Variability in outcomes may reflect misalignment between monitoring strategies and patient vulnerability rather than limited efficacy. Frailty-based segmentation may therefore represent a key lever to improve both clinical impact and implementation success. This framework may also provide a basis for future risk stratification models and the integration of frailty into digital health algorithms.

Frailty should be considered not only a prognostic marker but a structuring principle for digital care pathways.

Limitations

This viewpoint presents a conceptual framework rather than results from an interventional or comparative study. Although grounded in published evidence and real-world clinical experience, the proposed model has not been empirically validated as a unified strategy.

Some elements are derived from a French telemonitoring program implemented within the national ETAPES framework; however, the principles described—including frailty segmentation, proxy-supported reporting, and nurse-led triage—are intended to be generalizable across different health care systems and digital infrastructures.

Future research should prospectively evaluate this framework, including stratified outcomes according to frailty level and its impact on hospitalization, quality of life, and health care utilization.

Toward Frailty-Adaptive Digital HF Care

Remote monitoring should not remain a uniform technological intervention applied identically across a heterogeneous aging population. Its value depends on its ability to adapt both to frailty status and to the dynamic progression of HF. Robust patients may benefit primarily through optimization and empowerment; prefrail patients through early detection and timely intervention; frail patients through crisis avoidance and structured rapid-access care; and palliative patients through comfort-focused support, stabilization, and the preservation of home-based dignity.

A frailty-adaptive remote monitoring strategy is also essential for digital health equity, ensuring that older adults with the greatest vulnerabilities are not excluded from digital transformation but are meaningfully included within it.

A frailty-based conceptual model represents a necessary evolution in digital HF care and provides a framework for more meaningful and sustainable telemonitoring programs. One-size-fits-all telemonitoring is fundamentally misaligned with the biology and care trajectories of older adults.

Acknowledgments

The authors thank the nursing and allied health teams of the Cardiogeriatrics Department at Hôpital La Porte Verte for their daily commitment and valuable contributions to the implementation of this innovative model of care.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-forprofit sectors. The work was carried out as part of the employment of the authors at Hôpital La Porte Verte (Versailles, France). The funder had no role in the writing of the manuscript, its editing, approval, or the decision to submit for publication.

Conflicts of Interest

None declared.

References

1. Chaudhry SI, Mattera JA, Curtis JP, et al. Telemonitoring in patients with heart failure. N Engl J Med 2010 Dec 9;363(24):2301-2309. [doi: 10.1056/NEJMoa1010029] [Medline: 21080835]
2. Koehler F, Koehler K, Deckwart O, et al. Efficacy of telemedical interventional management in patients with heart failure (TIM-HF2): a randomised, controlled, parallel-group, unmasked trial. Lancet 2018 Sep 22;392(10152):1047-1057. [doi: 10.1016/S0140-6736(18)31880-4] [Medline: 30153985]
3. Girerd N, Barbet V, Seronde MF, et al. Association of a remote monitoring programme with all-cause mortality and hospitalizations in patients with heart failure: national-scale, real-world evidence from a 3-year propensity score analysis of the TELESAT-HF study. Eur J Heart Fail 2025 Sep;27(9):1658-1669. [doi: 10.1002/ejhf.3563] [Medline: 39807086]
4. Cajita MI, Gleason KT, Han HR. A systematic review of mHealth-based heart failure interventions. J Cardiovasc Nurs 2016;31(3):E10-E22. [doi: 10.1097/JCN.0000000000000305] [Medline: 26544175]
5. Kruse C, Fohn J, Wilson N, Nunez Patlan E, Zipp S, Mileski M. Utilization barriers and medical outcomes commensurate with the use of telehealth among older adults: systematic review. JMIR Med Inform 2020 Aug 12;8(8):e20359. [doi: 10.2196/20359] [Medline: 32784177]

XSL•FO
RenderX

6. Lam K, Lu AD, Shi Y, Covinsky KE. Assessing telemedicine unreadiness among older adults in the United States during the COVID-19 pandemic. JAMA Intern Med 2020 Oct 1;180(10):1389-1391. [doi: 10.1001/jamainternmed.2020.2671] [Medline: 32744593]
7. Fried LP, Tangen CM, Walston J, et al. Frailty in older adults: evidence for a phenotype. J Gerontol A Biol Sci Med Sci 2001 Mar;56(3):M146-M156. [doi: 10.1093/gerona/56.3.m146] [Medline: 11253156]
8. Rockwood K, Mitnitski A. Frailty in relation to the accumulation of deficits. J Gerontol A Biol Sci Med Sci 2007 Jul;62(7):722-727. [doi: 10.1093/gerona/62.7.722] [Medline: 17634318]
9. Clegg A, Young J, Iliffe S, Rikkert MO, Rockwood K. Frailty in elderly people. Lancet 2013 Mar 2;381(9868):752-762. [doi: 10.1016/S0140-6736(12)62167-9] [Medline: 23395245]
10. Afilalo J, Alexander KP, Mack MJ, et al. Frailty assessment in the cardiovascular care of older adults. J Am Coll Cardiol 2014 Mar 4;63(8):747-762. [doi: 10.1016/j.jacc.2013.09.070] [Medline: 24291279]
11. McNallan SM, Chamberlain AM, Gerber Y, et al. Measuring frailty in heart failure: a community perspective. Am Heart J 2013 Oct;166(4):768-774. [doi: 10.1016/j.ahj.2013.07.008] [Medline: 24093859]
12. Vidán MT, Blaya-Novakova V, Sánchez E, Ortiz J, Serra-Rexach JA, Bueno H. Prevalence and prognostic impact of frailty and its components in non-dependent elderly patients with heart failure. Eur J Heart Fail 2016 Jul;18(7):869-875. [doi: 10.1002/ejhf.518] [Medline: 27072307]
13. Denfeld QE, Winters-Stone K, Mudd JO, Gelow JM, Kurdi S, Lee CS. The prevalence of frailty in heart failure: a systematic review and meta-analysis. Int J Cardiol 2017 Jun 1;236:283-289. [doi: 10.1016/j.ijcard.2017.01.153] [Medline: 28215466]
14. Savarese G, Lund LH. Global public health burden of heart failure. Card Fail Rev 2017 Apr;3(1):7-11. [doi: 10.15420/cfr.2016:25:2] [Medline: 28785469]
15. Rodríguez-Pascual C, Paredes-Galán E, Ferrero-Martínez AI, et al. The frailty syndrome is associated with adverse health outcomes in very old patients with stable heart failure: a prospective study in six Spanish hospitals. Int J Cardiol 2017 Jun 1;236:296-303. [doi: 10.1016/j.ijcard.2017.02.016] [Medline: 28215465]
16. Uchmanowicz I, Gobbens RJJ. The relationship between frailty, anxiety and depression, and health-related quality of life in elderly patients with heart failure. Clin Interv Aging 2015;10:1595-1600. [doi: 10.2147/CIA.S90077] [Medline: 26491276]
17. Pandey A, Kitzman D, Whellan DJ, et al. Frailty among older decompensated heart failure patients: prevalence, association with patient-centered outcomes, and efficient detection methods. JACC Heart Fail 2019 Dec;7(12):1079-1088. [doi: 10.1016/j.jchf.2019.10.003] [Medline: 31779931]
18. Boriani G, Da Costa A, Quesada A, et al. Effects of remote monitoring on clinical outcomes and use of healthcare resources in heart failure patients with biventricular defibrillators: results of the MORE-CARE multicentre randomized controlled trial. Eur J Heart Fail 2017 Mar;19(3):416-425. [doi: 10.1002/ejhf.626] [Medline: 27568392]
19. Inglis SC, Clark RA, Dierckx R, Prieto-Merino D, Cleland JGF. Structured telephone support or non-invasive telemonitoring for patients with heart failure. Cochrane Database Syst Rev 2015 Oct 31;2015(10):CD007228. [doi: 10.1002/14651858.CD007228.pub3] [Medline: 26517969]
20. Afilalo J, Karunananthan S, Eisenberg MJ, Alexander KP, Bergman H. Role of frailty in patients with cardiovascular disease. Am J Cardiol 2009 Jun 1;103(11):1616-1621. [doi: 10.1016/j.amjcard.2009.01.375] [Medline: 19463525]
21. Eberly LA, Kallan MJ, Julien HM, et al. Patient characteristics associated with telemedicine access for primary and specialty ambulatory care during the COVID-19 pandemic. JAMA Netw Open 2020 Dec 1;3(12):e2031640. [doi: 10.1001/jamanetworkopen.2020.31640] [Medline: 33372974]
22. Kvedar JC, Fogel AL, Elenko E, Zohar D. Digital medicine's march on chronic disease. Nat Biotechnol 2016 Mar;34(3):239-246. [doi: 10.1038/nbt.3495] [Medline: 26963544]
23. Whellan DJ, Goodlin SJ, Dickinson MG, et al. End-of-life care in patients with heart failure. J Card Fail 2014 Feb;20(2):121-134. [doi: 10.1016/j.cardfail.2013.12.003] [Medline: 24556532]
24. Warraich HJ, Yancy CW. Failure is not an option: optimizing care for heart failure patients at the end of life. JACC Heart Fail 2019 Jul;7(7):558-560. [doi: 10.1016/j.jchf.2019.02.006] [Medline: 31078472]
25. Gadoud A, Jenkins SMM, Hogg KJ. Palliative care for people with heart failure: summary of current evidence and future direction. Palliat Med 2013 Oct;27(9):822-828. [doi: 10.1177/0269216313494960] [Medline: 23838375]
26. Bakitas MA, Dionne-Odom JN, Ejem DB, et al. Effect of an early palliative care telehealth intervention vs usual care on patients with heart failure: the ENABLE CHF-PC randomized clinical trial. JAMA Intern Med 2020 Sep 1;180(9):1203-1213. [doi: 10.1001/jamainternmed.2020.2861] [Medline: 32730613]
27. Calton B, Shibley WP, Cohen E, et al. Patient and caregiver experience with outpatient palliative care telemedicine visits. Palliat Med Rep 2020;1(1):339-346. [doi: 10.1089/pmr.2020.0075] [Medline: 34223495]
28. Gelfand SL, Lakin JR, Sciacca KR, et al. Specialty-aligned palliative care: responding to the needs of a tertiary care health system. J Pain Symptom Manage 2022 Dec;64(6):e341-e346. [doi: 10.1016/j.jpainsymman.2022.08.013] [Medline: 36031081]
29. Koehler F, Winkler S, Schieber M, et al. Telemedicine in heart failure: pre-specified and exploratory subgroup analyses from the TIM-HF trial. Int J Cardiol 2012 Nov 29;161(3):143-150. [doi: 10.1016/j.ijcard.2011.09.007] [Medline: 21982700]

XSL•FO
RenderX

30. Naylor MD, Brooten DA, Campbell RL, Maislin G, McCauley KM, Schwartz JS. Transitional care of older adults hospitalized with heart failure: a randomized, controlled trial. J Am Geriatr Soc 2004 May;52(5):675-684. [doi: 10.1111/j.1532-5415.2004.52202.x] [Medline: 15086645]
31. Boyd CM, Landefeld CS, Counsell SR, et al. Recovery of activities of daily living in older adults after hospitalization for acute medical illness. J Am Geriatr Soc 2008 Dec;56(12):2171-2179. [doi: 10.1111/j.1532-5415.2008.02023.x] [Medline: 19093915]

Abbreviations

ETAPES: Expérimentations de Télémédecine pour l'Amélioration des Parcours en Santé
HF: heart failure
Tele-HF: Telemonitoring to Improve Heart Failure Outcomes
TIM-HF: Telemedical Interventional Monitoring in Heart Failure

Edited by J Piera-Jiménez; submitted 29.Nov.2025; peer-reviewed by A Spaulding, J Edwards; revised version received 22.Mar.2026; accepted 09.Apr.2026; published 20.May.2026.

Please cite as:
Esser R, Maurou O, Larbaneix M, Mondragon A, Esteban M, Farges C, Pages N, Nisse Durgeat S, Harboun M
Frailty-Based Remote Monitoring in Older Adults With Heart Failure: Conceptual Framework for Adaptive Digital Health Strategies
J Med Internet Res 2026;28:e88681
URL: https://www.jmir.org/2026/1/e88681
doi:10.2196/88681

© Rémi Esser, Olivier Maurou, Marine Larbaneix, Alejandro Mondragon, Marlène Esteban, Christine Farges, Nicolas Pages, Sophie Nisse Durgeat, Marc Harboun. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Use of Speculative Fiction in Future-Focused Health Care Research

Sjors Groeneveld[1,2], MSc; Jobke Wentzel[3], PhD; Melissa Laurens[1,4], PhD; Lisette van Gemert-Pijnen[5], PhD; Rudolf M Verdaasdonk[2], PhD; Harmieke van Os-Medendorp[6,7], PhD; Marjolein den Ouden[1,4], PhD

[1]Research Group Technology, Health & Care, School of Social Work, Saxion University of Applied Sciences, Postbox 70.000, Enschede, Overijssel, The Netherlands

[2]TechMed Center, Health Technology Implementation, University of Twente, Enschede, The Netherlands

[3]Research Group ICT Innovation in Care, Windesheim University of Applied Sciences, Zwolle, The Netherlands

[4]Research Group Care and Technology, Regional Community College of Twente, Hengelo, The Netherlands

[5]Centre for eHealth and Wellbeing Research, Section of Psychology, Health and Technology, University of Twente, Enschede, The Netherlands

[6]Faculty Health, Sports, and Social Work, Inholland University of Applied Sciences, Amsterdam, The Netherlands

[7]Spaarne Gasthuis Academy, Haarlem, The Netherlands

Corresponding Author:
Sjors Groeneveld, MSc
Research Group Technology, Health & Care, School of Social Work, Saxion University of Applied Sciences, Postbox 70.000, Enschede, Overijssel, The Netherlands

Abstract

Health care is undergoing rapid change due to digitalization, artificial intelligence, data-driven decision-making, and shifting patient needs. These developments raise complex ethical, social, and organizational questions that cannot always be addressed by conventional research methods alone. There is a growing need for tools that help stakeholders imagine alternative futures to surface underlying values. Futures studies and speculative fiction respond to this need by presenting “what if” scenarios that make uncertain futures concrete and discussable, enabling dialogue among diverse stakeholders such as health care professionals, researchers, policy makers, and decision makers. This paper examines how speculative fiction can be used as a structured yet imaginative tool in health care research. It positions speculative fiction as a tool within futures studies and participatory research approaches, outlining its conceptual grounding and clarifying its role that stimulates interpretation and reflection within methods such as focus groups, workshops, and surveys. In this way, speculative fiction complements established approaches such as human-centered design, contextual inquiry, and value specification by addressing what does not yet exist and making abstract future issues tangible. The paper presents a case study, The Digital Data Divide, in which speculative fiction was used to explore the use of personal data in health care. Two contrasting short films were developed to stimulate dialogue and invite participants to reflect on associated values. Insights from this case are structured into a 6-step practical reporting guide: determining whether speculative fiction fits the research aim, choosing an appropriate form of speculative fiction, creating or selecting speculative scenarios, engaging participants, analyzing responses, and sharing and disseminating results. Across these steps, the paper discusses methodological and ethical considerations, including alignment between scenarios and study aims, balancing utopian and dystopian elements, questions of plausibility and interpretability, and the need for researcher reflexivity. Overall, this paper contributes to the growing discussion on future-oriented tools in health care research by showing how speculative fiction can help address complex and uncertain challenges in ways that are accessible to a wide range of stakeholders and that support dialogue across perspectives, provided that its use is ethically transparent, methodologically explicit, and carefully reported. The paper concludes with a call to other researchers to also experiment with speculative fiction and to share their experiences with the health care research community to learn and advance its use.

(J Med Internet Res 2026;28:e89699) doi:10.2196/89699

KEYWORDS

health care research; speculative fiction; Futures Studies; research tools; narratives; health technology; digital health

Introduction

Health care is changing fast because of new technologies, social developments, and changing patient needs [1]. Digitalization, artificial intelligence (AI), data-driven decision-making, and new models of care delivery are reshaping how health care is organized and how professionals work [2]. At the same time, aging populations, rising multimorbidity, workforce shortages, and increasing expectations for personalized care add further pressure to the system. These developments generate complex

XSL•FO
RenderX

questions that are not purely technical but also ethical, social, and organizational. For example: How should personal data be used responsibly? What forms of professional judgment will matter most in data-driven health care? How can care remain person-centered when technologies become more dominant? And how should changing roles and responsibilities be negotiated between professionals, patients, and digital systems? Many of these questions cannot be fully addressed through conventional research alone and benefit from approaches that present alternative futures, challenge underlying assumptions, and help define ethical boundaries by engaging stakeholders in collective sense-making around uncertain futures [3-5]. This is particularly important for ethical or value-based issues, where the aim is not to predict the most *probable* future but to explore and discuss (multiple) *possible* futures in a structured, open, and reflective manner [5,6]. Futures studies and speculative fiction can help do this.

The aim of this paper is to provide inspiration and guidance on how speculative fiction can be used in health care research. Specifically, this Viewpoint (1) positions speculative fiction as a structured probe within futures-oriented health research and (2) offers a practical, stepwise guide illustrated through the case of The Digital Data Divide. By doing so, we clarify both the conceptual rationale and the practical application of speculative fiction for health care researchers seeking to engage stakeholders in reflection on possible futures.

Speculative Fiction: A Tool for Exploring Possible Futures

> *It is 2055. Chronic diseases like diabetes and heart failure are managed entirely at home. AI monitors vitals in real time, robots assist with daily care, and virtual nurses check in via hologram. Hospitals focus only on emergencies, while neighborhood health hubs handle the rest.*

For some, this future promises efficiency and independence; for others, it raises urgent concerns: what happens to human connection in care? Who is accountable when algorithms fail? How do we ensure that no one gets left behind? This scenario isn't a prediction; it is speculative fiction, a broad term for stories, images, or experiences that deliberately depart from consensus reality to explore alternative futures and their consequences [7]. Unlike traditional forecasting that projects the future based on the most *probable* developments and quantitative trends, speculative fiction aims to explore *possible* futures and its consequences [8]. This distinction resonates with established futures studies frameworks, such as Voros' Futures Cone [9]. Speculative fiction uses imagination to reflect on questions, from ethics to equity, and to spark conversations about the world we want...or want to avoid [10].

We recognize that speculative fiction overlaps with adjacent traditions such as speculative and critical design, design fiction, and experiential futures [11-14]. While these approaches differ in disciplinary roots and primary outputs, they share a commitment to engaging audiences through encounters with imagined futures. In this paper, we use speculative fiction as an umbrella term for narrative, visual, experiential, and material forms when they function primarily as stories about a possible future world. Our focus is therefore less on design as production and more on fiction as a narrative device that enables reflection, collective sense-making, and dialogue within health care research.

Speculative fiction is widely used beyond academic research. It thrives in culture and society. From dystopian television series such as "Star Trek," "Black Mirror," or "The Handmaid's Tale" to computer games and visionary works of science fiction literature, it captivates audiences by exploring alternative realities and posing provocative "what if" questions. These stories challenge our assumptions about technology, ethics, and society, often serving as a mirror to reflect on contemporary anxieties or aspirations. There is a body of educational and sociological scholarship that suggests that fictional narratives, such as science fiction [15,16], can function as powerful sense-making tools. By inviting readers and viewers to temporarily inhabit alternative perspectives and social worlds, these stories render abstract or complex societal dynamics more tangible and emotionally resonant. Their capacity to create cognitive and moral distance, while simultaneously fostering identification and critical reflection, may help explain why speculative fiction is particularly well suited to exploring possible futures of technology and care.

For decades, Futures Studies have systematically explored possible, probable, and (un)desirable futures to inform policy, strategy, and societal development [17,18]. In this paper, we position speculative fiction as a tool within this established academic field of Futures Studies. Speculative fiction as a tool is not a data collection method: it acts as a probe that stimulates interpretation and reflection within methods of data collection such as focus groups, workshops, facilitated dialogues, or surveys. Speculative fiction thereby complements existing approaches such as participatory methods, contextual inquiry, value specification, and design thinking. While these approaches share the goal of uncovering values, interests, and perspectives [19,20], they typically focus on current realities or probable futures. Speculative fiction, however, excels in addressing what does not yet exist. For example, contextual inquiry seeks to understand current needs and behaviors, but speculative fiction pushes this exploration into hypothetical futures. Similarly, value specification and service road mapping aim to clarify priorities and map pathways, but speculative fiction introduces unfamiliar or disruptive contexts, forcing stakeholders to confront uncertainties and ethical dilemmas that otherwise might be overlooked. Even within design thinking's iterative cycles, speculative approaches act as a "what if" sandbox [12], enabling teams to stress-test ideas against possible future conditions. In this way, it complements these approaches by making the intangible tangible.

By deploying speculative fiction in health care research, we align with the field's core goal of fostering democratic participation and cross-sectoral dialogue [3], while cultivating futures literacy: the competence to reflect on diverse attitudes toward the future and generate alternative scenarios to navigate complexity and uncertainty [17,21].

XSL•FO
RenderX

The Role of Speculative Fiction in Health Care

As discussed earlier, speculative fiction can complement established research methods by addressing what does not yet exist. An example from health care is the CeHReS Roadmap, a health care–specific, design-oriented approach that emphasizes contextual values and stakeholder needs early in the development of eHealth technologies [22]. While the CeHRes Roadmap excels in mapping current priorities and behaviors, it focuses on existing realities or imaginable situations. Speculative fiction adds a critical layer by projecting these values and needs into hypothetical futures. For example, while the CeHRes Roadmap might uncover a care team's current frustrations with digital tools, speculative fiction can explore how those frustrations might evolve in a future where AI-driven diagnostics or robotic care assistants fundamentally reshape their roles. By creating speculative scenarios, it enables stakeholders, including those who rarely engage with future-oriented questions in their daily work, such as frontline caregivers or administrative staff, to confront complex dilemmas, uncover hidden assumptions, and co-design adaptive solutions. Speculative fiction's strength lies in its ability to democratize futures thinking, surface unspoken concerns, and bridge abstract trends and action. Speculative fiction takes many forms, each offering unique ways to engage with possible futures and their implications. For example, initiatives such as the "Expo Future of work in health care and wellbeing," developed by a Dutch center of expertise for long-term care, used an interactive, traveling exhibition to engage health care professionals in exploring future scenarios for the care sector [23]. The expo presents 4 speculative visions of care work in 2040, combining visual storytelling, interactive installations, and discussion prompts to start dialogue about technological, organizational, and societal changes. Although inspiring, such projects are still an exception rather than the rule.

While its potential is vast, a recent scoping literature review [3] shows that Futures Studies primarily focus on macro-level themes (eg, national strategies, pandemic preparedness, and workforce dynamics). These studies highlight *what* needs to be explored but often overlook *how* tools such as speculative fiction can be deployed within organizations to mobilize internal stakeholders in cocreating their futures. Building on the established value of speculative fiction, the following section of this paper shifts from theory to application, offering practical insights learned from a case study.

Practical Insights Learned in Using Speculative Fiction

Overview

This section focuses on the practical aspects of using speculative fiction in research. This is illustrated through the introduction of a case project, The Digital Data Divide, from which practical insights are drawn and structured into six steps: (1) determining whether speculative fiction aligns with the research aim, (2) choosing the form of speculative fiction, (3) creating or selecting speculative fiction scenarios, (4) engaging participants during data collection, (5) analyzing responses, and (6) sharing results. These steps are subsequently synthesized into a 6-step reporting guide to facilitate clarity, transparency, and transferability in future applications.

Case: The Digital Data Divide

The growing ability to collect and analyze personal data is transforming many sectors, including health care [24]. Data from wearable devices, smart home sensors, and digital health applications increasingly inform clinical decisions and public health strategies [25]. While these developments hold promise for more personalized and efficient care, they also raise complex questions about privacy, autonomy, trust, and the ownership of data [26]. These issues are not only technical but also social and ethical, requiring spaces for dialogue and reflection across different parts of society.

To create such a space and engage a broad audience in discussions about data-driven futures, the project The Digital Data Divide is developed [27] (see Figure 1 for the project poster). This project is designed as an interactive speculative experience that can be used in both public festival contexts and more structured educational or research settings, using speculative fiction to provoke reflection and dialogue. Participants begin the experience as they are invited to make a symbolic choice between a red and a blue pill, each representing a different future scenario. Based on their choice, participants view 1 of 2 short speculative films that follow an ordinary person (Maarten or Maartje) throughout a day in a data-saturated society. In both films, personal data play a central role in everyday life. However, in the "red" scenario, data systems subtly work against the person, shaping opportunities, reinforcing inequalities, enabling surveillance, and limiting autonomy and agency over personal data. In the "blue" scenario, data infrastructures are depicted as supportive, seemingly enhancing well-being, safeguarding privacy, and enabling choice and control. Themes such as privacy, autonomy, surveillance, commercial data use, and social (in)equality recur in both films but are framed in contrasting ways. Differences in aesthetic elements such as color palette and music further reinforce the emotional tone of each world.

XSL•FO
RenderX

Figure 1. The Digital Data Divide poster.

After viewing the films, participants reunite in a shared space and use dialogue cards to discuss what they had seen; reflect on the ethical, personal, and societal implications of these imagined futures; and make underlying values, assumptions, and emotional responses explicit. The contrasting scenarios surfaced participants' underlying values and assumptions about technology and data, prompting reflection not only on desirable futures but also on the societal conditions required to achieve them.

The project is used across a wide variety of contexts, including public festivals, educational institutions, research projects, and professional events. Figure 2 illustrates its appearance in a festival context. Over the course of time, The Digital Data Divide reached a diverse audience, from health care professionals and students to members of the public. By situating the conversation in both informal public settings and more structured educational or research contexts, the project aims to lower the barrier for participation, while also enabling deeper reflection. Across settings, it supports participants in exploring and articulating views beyond their usual professional or social boundaries.

XSL•FO
RenderX

Figure 2. The Digital Data Divide at a festival setting.

Across its implementations, The Digital Data Divide has been used in 2 distinct but related contexts. In public festival settings, the project functioned as an exploratory live experience to test whether speculative fiction could stimulate dialogue in informal environments. The project reached members of the general public from diverse backgrounds. Participants provided feedback through brief flyers containing visual analogue scales (eg, perceived awareness of the impact of data and evaluation of the experience) and open comment fields. Responses were collected on-site and qualitatively clustered to identify recurring reflections and suggestions for improvement.

In a later, more structured qualitative focus group study with international health care students (n=40), the same films were used as prompts for moderated discussion. These sessions were recorded and analyzed using an inductive thematic approach. While the festival implementation primarily served exploratory development purposes, the focus group setting allowed for deeper analysis of how participants reasoned about imagined data-driven futures.

Across both applications, several recurring tensions became visible. Participants valued the use of contrasting extremes, noting that the utopian or dystopian framing helped clarify their own positions. The films heightened awareness of the societal impact of personal data, though responses ranged from concern to skepticism about plausibility. Some participants experienced the format as highly thought-provoking, while others expressed a desire for clearer facilitation or more practical information about current realities. These insights informed further refinement of the dialogue format. In particular, we became more attentive to the delicate balance between realism and imaginative projection: scenarios need to feel credible enough to enable identification, while remaining sufficiently speculative to open space for critical reflection. This realization shaped our recommendation to carefully calibrate plausibility and contrast when using speculative fiction as a structured tool for future-oriented dialogue.

Step 1: Determine Whether Speculative Fiction Fits the Research Aim

The first step in using speculative fiction is to establish whether the research question benefits from imaginative exploration. Speculative approaches work particularly well for topics marked by uncertainty, ethical complexity, or potential changes in (professional) practice, especially when questions involve ambiguity, value tensions, and several plausible outcomes [28,29]. In health care, this could include (the use of) emerging technologies, evolving models of patient care, or shifts in professional roles, responsibilities, and ethical boundaries. Rather than predicting the most likely future, speculative fiction invites participants to imagine a possible future and reflect on its desirability, risk, and implications, resulting in rich insights into the underlying values and preferences that shape how people relate to future technologies. Speculative fiction is less suitable for highly defined operational questions or established interventions. A useful guiding question is: Does this topic involve futures for which no empirical data yet exist, but where societal and ethical reflection is necessary?

In The Digital Data Divide project, the research focus centered on the increasing use of personal data in health care and its ethical and societal consequences. Other possible topics could be the future role of nurses in AI-supported care, scenarios for end-of-life decision-making, or future models of empathy and access to care. Establishing a focus at this conceptual level ensures that speculative fiction is not used as a storytelling exercise alone but as a structured method for exploring emerging health futures.

Step 2: Choosing the Form of Speculative Fiction

After identifying the research focus, the next step is to select an appropriate form. Speculative work can take many shapes including written narratives, audiovisual media, interactive formats, and design fiction artifacts [30]. Although these forms vary widely, we propose grouping them into 4 overarching

XSL•FO
RenderX

categories including anticipated strengths and limitations to support practical comparison (Table 1). Each form supports a different kind of audience, mode of participation, and depth of discussion, and requires different resources. Through written formats (ie, short stories, diary fragments, and poetry), it delves into deep narrative worlds, using speculative storytelling to question assumptions, reimagine systems, and propose bold alternatives [31]. Written narratives often work well for individual reflection and deeper introspection, while audiovisual media such as short films or animations can evoke strong emotional responses and intellectual engagement that provide a powerful starting point for shared dialogue [32]. Interactive or performative formats, for example, installations or live role-playing through participatory theater, are particularly effective for larger or public audiences where collective participation and experiential learning are desired. Design fiction artifacts concretize future scenarios by introducing tangible objects or representations, enabling participants to engage in concrete discussions about aspects such as usability and everyday practice.

When deciding which form to use, we suggest considering four criteria: (1) the audience (individual, small groups, or larger groups), (2) the desired mode of participation (reflection, group dialogue, or public engagement), (3) the practical feasibility and resources available, and (4) the desired depth of discussion or analysis.

Table . Examples of possible forms of speculative fiction.

Form of speculative fiction	Most suitable for	Strengths	Considerations or limitations
Written narratives (short stories, diary fragments, and letters from the future)	Individual reflection; sensitive or value-laden topics	Supports introspection and slower, thoughtful engagement	Requires reading time and capacities; less suitable for large or public audiences
Audiovisual media (short films, animations, and illustrated scenarios)	Group dialogue, emotional engagement, and mixed audiences	High accessibility, strong emotional impact, and effective for shared reflection	Resource-intensive, requires production skills, and risk of oversimplification
Interactive formats (installations, role-playing, experiential futures, and theater play)	Large groups, public events, and festivals	Immersive and memorable; encourages active participation	Logistical complexity, requires facilitation, and harder to standardize
Design fiction artifacts (prototypes, speculative objects, interfaces, and fictional brochures)	Stakeholder workshops, design reviews, policy dialogue, and education	Makes future technologies feel concrete, helps discuss assumptions about usability, and everyday practice	Risk of being interpreted as “the” solution, requires careful framing, and production effort varies depending on fidelity

In The Digital Data Divide project, original film-based scenarios were used to create an emotionally engaging experience suitable for both informal public festivals and more educational or research environments. The choice for film supported accessibility, narrative depth, and emotional relatedness. Two contrasting short films (1 utopian and 1 dystopian) were made with the goal to start a dialogue on relevant values from multiple perspectives. This contrast helped in the discussion to reflect on what is important for you as a person in a *possible* future, instead of finding consensus as participants in the most *probable* future.

Step 3: Creating or Selecting Speculative Fiction Scenarios

Once the form is chosen, the next step is deciding whether to develop new speculative material or adapt existing fictional work. Creating original speculative narratives offers flexibility and ensures alignment with the research aims, but it requires more time, resources, and often interdisciplinary collaboration. When taking this route, the competency of the creators becomes crucial: scenario development benefits from a combination of domain knowledge (such as health care, ethics, laws, or regulations), creative skills (such as storytelling and design), and methodological awareness to ensure that the material is relevant, balanced, and responsibly framed. Original speculative scenarios can be created individually or in cocreation with a wide range of stakeholders, including not only domain experts and end users but also designers, ethicists, creatives, and other voices who can meaningfully shape how a future is imagined. Individual writing allows for very specific controlled narratives, while group cocreation supports diversity of perspectives and greater creative range. Creative workshops, often including structured ideation methods such as brain writing, futures wheel, SCAMPER, or the Experiential Futures Ladder [14,33], can help translate abstract concerns into concrete situations, characters, and moral dilemmas.

Using existing stories or cultural references can reduce development time, but it also requires careful assessment of the material’s quality and suitability, a concern that aligns with methodological discussions in vignette studies [34]. Researchers should examine whether the fiction aligns with the research focus, offers sufficient nuance for meaningful reflection, and presents a balanced mix of positive and negative elements. It is also important to consider whether the material avoids unnecessary bias and does not unintentionally steer participants toward a particular viewpoint. Although concepts such as validity and reliability take on a different form when creating or selecting speculative fiction scenarios, focusing on plausibility and interpretability rather than factual accuracy, transparency about narrative choices, and ethical intent remains crucial to ensure that (adapted) material is used in a responsible and balanced way. In this regard, involving intended participants or end users during scenario development or selection can help ensure that scenarios feel realistic, balanced, and meaningful for those who will later respond to it.

In The Digital Data Divide project, speculative scenarios were developed through cocreation in a multidisciplinary session in

XSL•FO
RenderX

which researchers, health care professionals, ethicists, students, designers, and film professionals explored emerging questions around personal data in health care. This informal yet structured session generated early narrative ideas, ethical dilemmas, and possible character storylines. A scriptwriter then translated these insights into draft scripts. Filmmakers, actors, and a composer were involved to refine tone, pacing, emotional resonance, and visual language. Importantly, the development process moved iteratively from concept to scenario to finished film. This shows that speculative scenarios are not static elements but coconstructed narratives formed by a wide range of involved people, creative expertise, and production realities.

Step 4: Engaging Participants During Data Collection

Engaging participants is essential for turning speculative scenarios into meaningful insights rather than leaving them as stand-alone stories [12,14,29]. To ensure that speculative scenarios contribute to the research aim, participants need structured support to move from simply reading, viewing, or participating in a fictional future to considering what it might mean for their own practice, values, or professional identity. Clear prompts, guiding questions, or dialogue tools help make this transition explicit. Speculative fiction is most effective when participants are encouraged not only to react to a story but also to position themselves within the imagined situation and articulate which futures they consider desirable, problematic, or uncertain. In this way, the speculative element becomes a tool for research-relevant reflection and meaning-making. Ethical considerations are essential as speculative scenarios can influence how participants think or feel about an issue, especially when they include strong dystopian or utopian elements. Clear informed consent [35], openness about the purpose of the scenario, and attention to potential emotional impact help ensure that speculative materials are used responsibly.

The Digital Data Divide project was deliberately designed to be accessible and engaging, suitable for both diverse public audiences and more structured research or educational environments. In both applications, participants first viewed contrasting speculative films individually, after which structured reflection was facilitated in a shared setting. Dialogue cards encouraged participants to express emotional responses, consider ethical tensions, and explore how data-driven futures might influence their professional roles or personal autonomy. This process deepened the conversation and linked the speculative narrative to the ongoing construction of professional identity and practice.

In the festival setting, feedback was collected using brief hard copy flyers with visual analogue scales and open comment fields. This pragmatic method was intentionally selected to suit an informal festival context, where lengthy questionnaires or structured interviews would have created participation barriers. In the qualitative focus group study, the same films were used as prompts for moderated discussion, and conversations were recorded and analyzed qualitatively. Together, these applications illustrate how speculative scenarios can function flexibly as a tool across different contexts.

Given that speculative scenarios may evoke sensitive reflections, ethical considerations were addressed in both applications of The Digital Data Divide, although in different ways reflecting their respective contexts. As said, the festival setting functioned as an exploratory live experience rather than a formal research study. No identifiable personal data were collected, and no audio or video recordings were made. Participants were informed in advance (both verbally and through the flyer) about the nature of the experience, including the presentation of contrasting future scenarios and the voluntary character of participation. Feedback was provided using the brief hard copy flyers, and the anonymous responses were stored on a protected research drive. Prior to the festival experience, the approach was discussed with ethics and privacy advisors to consider potential personal (data) risks. For the focus group study (where the films were used as one component within a broader qualitative research design), formal ethical approval was obtained through the relevant institutional review procedures at participating universities. Participants received an information letter and signed informed consent prior to participation. They were explicitly informed about the recording of the sessions, their right to withdraw at any time without consequences, and the handling of their data. Recordings were encrypted, transcribed without identifying information, and stored on a secure, password-protected research drive accessible only to the research team.

Step 5: Analyzing Responses

The fifth step concerns the analysis of participant responses. As noted earlier, speculative fiction is not a data collection method on its own; it is a tool that functions as a probe that triggers interpretation and reflection within methods of data collection such as focus groups, workshops, facilitated dialogues, or surveys. The data analysis of discussions based on speculative fiction may draw on approaches such as thematic analysis [36], narrative analysis [37], or mixed methods designs [38]. The focus is typically on meaning-making: how participants articulate values, ethical concerns, and desirable or undesirable futures. Rather than searching for “correct” answers, analysis aims to surface patterns in reasoning, identity formation, and expectations about emerging technologies. In mixed methods approaches, these qualitative insights can be combined with quantitative measures (such as surveys or rating scales) to provide a more comprehensive understanding of how participants make sense of the fictional futures.

The exploratory festival implementation of The Digital Data Divide necessarily introduced limitations. Measures of awareness were simplified, context factors (eg, festival atmosphere and time of day) that could influence responses, and the open nature of the setting could introduce information bias. These constraints raise important questions about validity and reliability in open, participatory contexts. Nonetheless, the method demonstrated that speculative fiction could generate meaningful insights outside traditional research environments, particularly when the goal is exploratory sense-making rather than precise causal inference. In the focus group study where the films from The Digital Data Divide project were used with students, the resulting qualitative data were analyzed using an inductive thematic approach, through which themes were

XSL•FO
RenderX

constructed from the students' reflections. While the analysis followed a conventional qualitative procedure, the goal remained exploratory: to understand how participants reasoned about imagined futures, rather than explanatory to treat their responses as "data" about current reality.

Step 6: Sharing Results

A final consideration is how to share insights generated through speculative fiction. Unlike some traditional research, speculative approaches do not only inform scholarly debate but also aim to circulate ideas, provoke reflection, and shape broader conversations about the future in society. Dissemination therefore extends beyond academic publications to include public exhibitions, community dialogue sessions, educational activities, and creative media outputs: formats that in some cases resemble those used in practice-based research. Returning insights to society is particularly relevant when imagined futures may influence real ones: presenting speculative scenarios and the reflections they generate can contribute to collective sense-making and democratic discussion about emerging technologies. In this context, collective sense-making refers to the shared interpretive process by which groups negotiate meanings, articulate underlying values, and coconstruct understandings of plausible futures. Drawing on literature on sense-making, the collective aspect emphasizes how participants collaboratively make sense of imagined futures rather than merely expressing isolated opinions [39]. By encountering contrasting scenarios and discussing them together, participants do not only reflect individually but also shape shared interpretations and concerns, supporting broader insights about the impact of technology.

In The Digital Data Divide project, this principle guided the choice to present findings not only in academic settings but also through continued public engagement at festivals, health care organizations, and educational programs. By sharing results in the same open spaces where the dialogue began, the project aimed to extend reflection, spark new conversations, and reinforce a feedback loop in which speculative visions and societal responses mutually shape one another.

Binding It Together

Together, the aforementioned insights, structured into 6 steps, show an example of how speculative fiction can be used as a structured but creative research tool in health care. By choosing a future-focused question, selecting the right speculative format, creating or selecting scenarios, involving participants, analyzing their reflections, and sharing results with society, speculative fiction goes beyond storytelling and becomes a tool for future-focused research. It builds ethical awareness, critical thinking, and open discussion in areas where evidence is limited and the future is uncertain. Unlike traditional scenario approaches, which often aim to map trends or outline plausible developments, speculative fiction creates experiential and value-rich situations that help participants feel and interpret possible futures. To enhance transparency and reusability, Table 2 provides a compact reporting guide aligned with the 6 steps. Researchers using speculative fiction may consider reporting on the following elements. This checklist is not intended as a rigid protocol but as a guide to support transparent and context-sensitive reporting when speculative fiction is used as a research tool.

Table . Six-step reporting guide for using speculative fiction in health care research.

Step	What to report	Key considerations
1. Research aim	Why speculative fiction was chosen and its foreseen contribution	Nature of uncertainty, ethical/value tensions, and intended contribution (exploratory, dialogical, and analytical)
2. Form of speculative fiction	Chosen format and rationale	Intended audience, level of participation, resources, and depth of engagement
3. Scenario development	How scenarios were created or selected	Cocreation process, balance of perspectives, plausibility checks, mitigation of bias, and involvement of stakeholders
4. Participant engagement	Facilitation approach and data collection	Prompts or dialogue tools used, setting of data collection, informed consent procedure, emotional safeguards, ethical issues, and handling the data
5. Analysis approach	How responses were interpreted	Qualitative or quantitative methods, analytic lens (eg, thematic and narrative), and reflexivity of researchers
6. Sharing and dissemination	How insights were circulated	Academic outputs, public engagement, feedback loops, and contribution to collective sense-making

Discussion

Principal Findings

This paper demonstrates how speculative fiction can support the exploration of uncertain or sensitive future issues in health care, aligning with the increasing need for future-oriented approaches to address the sector's wicked problems [22,40]. Many intended participants, such as patients, informal caregivers, and health care professionals, may find it difficult to envision future scenarios or anticipate their consequences. Because these groups are often most affected by change,

XSL•FO
RenderX

methods are needed that help unlock their perspectives and latent knowledge and enable creative, reflective contributions [41]. Speculative fiction bridges this gap by offering concrete, imaginative future scenarios that people can react to, rather than asking them to invent those futures themselves. This creates a more inclusive starting point, giving participants (regardless of education, experience, or role) an equal seat at the table. By making hypothetical futures tangible, speculative fiction supports those who struggle to express abstract ideas or anticipate change, enabling richer discussions about values, concerns, and desirable directions for health care [42]. Speculative fiction thereby complements widely used participatory and design-based methods such as Human-Centered Design, which rely on participants' articulating needs and imagining alternatives [43,44].

In addition, much participatory research relies on very verbalized methods: participants need to be able to comprehend the language the researcher uses and need to be able to express their own thoughts, desires, values, and needs in words. Not all participants can do this with ease, leading to possible biases [40]. Researchers have sought to overcome such biases by using visual stimuli to provoke thoughts and encourage idea-sharing among their participants using visual methods such as photo elicitation [45], which demonstrate how images can help participants express experiences and perspectives that are difficult to verbalize through conventional interviews alone. Speculative fiction can complement frameworks such as Design Thinking [46] by introducing future scenarios when imagining or verbalizing such situations often lies beyond the capacity of participants. In this sense, speculative fiction differs from other tools by presenting a possible future for participants to respond to, rather than asking them to construct it themselves. Speculative fiction enhances inclusivity by providing multiple modes of sense-making (verbal, visual, narrative, and experiential). This allows participants to engage with future scenarios without relying solely on abstract reasoning or verbal articulation. This enables involvement and opinion-forming among groups who may be difficult to engage through traditional research methods. Speculative fiction thus creates a shared entry point for exploring future directions across system levels. At the micro level, it helps individuals reflect on their values, identity, and responses to future situations. At the meso level, it enables teams and organizations to build shared visions and explore strategic choices with internal stakeholders. At the macro level, it can inform policy debates and spark wider societal dialogue [47]. In this way, speculative fiction supports meaningful conversations about the future across all levels of the health care system.

In their scoping review, Meskó and colleagues [3] conclude that a more systematic or guideline-based approach is needed for the practice of future methods. We wish to elaborate on this call, based on our experiences, with speculative fiction. *First*, in line with good and ethical research practice, transparency about the researcher's goal and chosen approach is key [35]. Speculative fiction is a powerful tool: it can trigger strong emotional responses, shape opinions, and foreground certain futures while obscuring others. Similar debates in bioethics highlight that speculative work can serve different purposes, and that standards for impact and good practice should be adjusted accordingly [48]. Researchers' own values and assumptions inevitably influence the scenarios they create or select, as well as how participants respond to them. Much like in qualitative research, this calls for continuous reflection on the own subjective role of researchers [49]. Involving multiple viewpoints, iterating between data and interpretation, and using triangulation can help safeguard balance and transparency. We argue that research papers that use speculative fiction should explicitly incorporate these ethical considerations, as this will help develop responsible, transparent, and effective ways of applying speculative fiction in health care research. Similar calls for greater methodological transparency have been made in related domains, including participatory eHealth design and human-computer interaction research [22,40,41]. *Second*, speculative fiction can function as a meaningful research tool only when its design is aligned with the study's aims. Researchers need to ensure that the speculative fiction scenarios support the type of insights they want to generate. Depending on the goal of the study, for example, describing, comparing, and measuring effects, such techniques should be chosen deliberately during study preparation. *Third*, speculative fiction scenarios should be created or selected carefully, avoiding narratives that are unnecessarily extreme or disconnected from reality. A key consideration is the involvement of diverse voices in scenario development. Engaging stakeholders early helps ensure that scenarios feel grounded, balanced, and meaningful. Strong or exaggerated scenarios can be valuable tools for stimulating reflection but only when researchers are transparent about why such provocation is used. Speculative scenarios can influence or steer opinions, especially when they lean toward highly utopian or dystopian futures. Researchers must remain alert to this potential "dark side."

Limitations and Future Work

While this paper outlines insights learned from applying speculative fiction in health care research, these insights are primarily grounded in our own project experience. As such, several of the recommendations, such as the proposed grouping of speculative forms and the anticipated strengths and limitations presented in Table 1, reflect informed judgment based on practice rather than systematically gathered evidence. We therefore regard this paper as a starting point for a broader methodological conversation rather than a definitive framework. Additional perspectives from other research contexts, projects, and disciplines are essential to refine, challenge, and expand the ideas presented here. Continued sharing of practical experiences will strengthen the knowledge base and support more robust guidance for using speculative fiction in health care research.

Additionally, future research could compare how participants respond to different types of speculative scenarios. Such comparisons may help explain how speculative fiction influences reasoning, value expression, and sense-making. Although this was outside the scope of this paper, these studies would be valuable for assessing the method's rigor and its ability to generate reliable insights about long-term attitudes, decision-making, or collaboration in health care.

XSL•FO
RenderX

Finally, cross-domain comparisons could deepen understanding of how speculative fiction works in health care compared with fields such as media studies, marketing, or public policy. Such studies may clarify how varying the level of utopian or dystopian elements shapes engagement and interpretation.

Conclusions

This paper shows how speculative fiction can function as a structured yet imaginative tool for exploring uncertain futures in health care. By positioning speculative fiction within futures studies and participatory design traditions, and by outlining insights from a case study, the paper demonstrates how speculative scenarios can support reflection, surface values, and stimulate dialogue across diverse groups. Whether researchers choose to adapt existing scenarios or create new ones, speculative fiction can be tailored to different research aims, settings, and levels of analysis.

To advance this emerging field, a more systematic and transparent approach is needed. Developing guidelines would help ensure ethical clarity, balanced scenario design, and transparent reporting of researcher choices. Such guidance would not *replace* existing frameworks. Rather, it would *complement* them by offering a structured way to engage with uncertain futures and by strengthening the methodological grounding of speculative fiction in health care research. We therefore propose a 6-step reporting guide for using speculative fiction in health care research. With this paper, we contribute to the growing debate on future-oriented approaches in health care research and aim to strengthen the foundation for the thoughtful use of speculative fiction. We encourage other researchers to experiment with speculative fiction and to share their experiences with the health care research community to learn and advance its use.

Acknowledgments

The authors would like to thank all involved people during the creation and use of the case project The Digital Data Divide.

Funding

This work was supported by a project (023.017.064) of the research program, Doctoral Grant for Teachers, which is (partly) financed by NWO (the Dutch Research Council).

Conflicts of Interest

None declared.

References

1. Sinha R. The role and impact of new technologies on healthcare systems. Discov Health Syst 2024;3(1):96. [doi: 10.1007/s44250-024-00163-w]
2. Groeneveld SWM, van Os-Medendorp H, van Gemert-Pijnen J, et al. Essential competencies of nurses working with AI-driven lifestyle monitoring in long-term care: a modified Delphi study. Nurse Educ Today 2025 Jun;149:106659. [doi: 10.1016/j.nedt.2025.106659] [Medline: 40056483]
3. Meskó B, Kristóf T, Dhunnoo P, Árvai N, Katonai G. Exploring the need for medical futures studies: insights from a scoping review of health care foresight. J Med Internet Res 2024 Oct 9;26:e57148. [doi: 10.2196/57148] [Medline: 39383528]
4. Vaughn LM, Jacquez F. Participatory research methods—choice points in the research process. J Participatory Res Methods 2020;1(1). [doi: 10.35844/001c.13244]
5. Breuer J, Maier MJ, Bansen A, et al. Design-based methods for exploring ethical questions in the field of neurotechnologies. J Responsible Innov 2024 Dec 31;11(1):2408814. [doi: 10.1080/23299460.2024.2408814]
6. Mesko B, Kristóf T, Dhunnoo P, Árvai N, Katonai G. A practical guide to using futures methods in health care: approaches, applications, and case studies. J Med Internet Res 2025;27:e82820. [doi: 10.2196/82820]
7. Oziewicz M. Speculative fiction. In: Lynch DS, editor. Oxford Research Encyclopedia of Literature: Oxford Academic; 2017. [doi: 10.1093/9780197851470.001.0001]
8. Börjeson L, Höjer M, Dreborg KH, Ekvall T, Finnveden G. Scenario types and techniques: towards a user's guide. Futures 2006 Sep;38(7):723-739. [doi: 10.1016/j.futures.2005.12.002]
9. Voros J. A generic foresight process framework. Foresight (Colch) 2003 Jun 1;5(3):10-21. [doi: 10.1108/14636680310698379]
10. Sutherland S. Speculative texts and what's next. Explorations Media Ecol 2024 Mar 1;23(1):47-51. [doi: 10.1386/eme_00192_7]
11. Johannessen LK, Keitsch MM, Pettersen IN. Speculative and critical design—features, methods, and practices. Presented at: Proceedings of the design society: international conference on engineering design; Aug 5-8, 2019; Delft the Netherlands p. 1623-1632. [doi: 10.1017/dsi.2019.168]
12. Dunne A, Raby F. Speculative Everything–Design, Fiction, and Social Dreaming: MIT Press; 2013.
13. Bleecker J. Design fiction: a short essay on design, science, fact, and fiction. In: Machine Learning and the City: Applications in Architecture and Urban Design: Wiley Online Library; 2022:561-578. [doi: 10.1002/9781119815075]

XSL•FO
RenderX

14. Candy S, Dunagan J. Designing an experiential scenario: The People Who Vanished. Futures 2017 Feb;86:136-153. [doi: 10.1016/j.futures.2016.05.006]
15. Hughes JJ, Lantos JD. Medical ethics through the Star Trek lens. Lit Med 2001;20(1):26-38. [doi: 10.1353/lm.2001.0004] [Medline: 11411031]
16. Ford A, Pretorius L. Balancing the needs of the many against the needs of the few: aliens, holograms and discussions of medical ethics. In: Teaching Medicine and Medical Ethics Using Popular Culture: Springer; 2017:133-147. [doi: 10.1007/978-3-319-65451-5_8]
17. Kristóf T. Development tendencies and turning points of futures studies. Eur J Futures Res 2024;12(1):9. [doi: 10.1186/s40309-024-00231-7]
18. Kristóf T, Nováky E. The Story of Futures Studies: an Interdisciplinary Field Rooted in Social Sciences. Soc Sci (Basel) 2023;12(3):192. [doi: 10.3390/socsci12030192]
19. Holtzblatt K, Beyer H. Principles of contextual inquiry. In: Contextual Design: Design for Life: Morgan Kaufmann; 2017:43-80.
20. Auernhammer J, Roth B. The origin and evolution of Stanford University's design thinking: from product design to design thinking in innovation management. J of Product Innov Manag 2021 Nov;38(6):623-644. [doi: 10.1111/jpim.12594]
21. Vidergor HE. Teaching futures thinking literacy and futures studies in schools. Futures 2023 Feb;146:103083. [doi: 10.1016/j.futures.2022.103083]
22. Kip H, Beerlage-de Jong N, van Gemert-Pijnen LJEWC, Kelders SM. The CeHRes Roadmap 2.0: update of a holistic framework for development, implementation, and evaluation of eHealth technologies. J Med Internet Res 2025 Jan 13;27:e59601. [doi: 10.2196/59601] [Medline: 39805104]
23. Design thinking in de zorg [Web page in Dutch]. Vilans. 2023. URL: https://www.vilans.nl/wat-doen-we/projecten/design-thinking-in-de-zorg [accessed 2025-12-19]
24. Carmichael L, Hall W, Boniface M. Personal data store ecosystems in health and social care. Front Public Health 2024;12:1348044. [doi: 10.3389/fpubh.2024.1348044] [Medline: 38384893]
25. Morita PP, Sahu KS, Oetomo A. Health monitoring using smart home technologies: scoping review. JMIR Mhealth Uhealth 2023 Apr 13;11:e37347. [doi: 10.2196/37347] [Medline: 37052984]
26. Groeneveld S, Bin Noon G, den Ouden MEM, et al. The cooperation between nurses and a new digital colleague "AI-driven lifestyle monitoring" in long-term care for older adults: viewpoint. JMIR Nurs 2024 May 23;7:e56474. [doi: 10.2196/56474] [Medline: 38781012]
27. Groeneveld SWM, Os-Medendorp H, Gemert-Pijnen J, Verdaasdonk RM, Ouden MEM. The digital data divide: using speculative fiction to facilitate public dialogue on the increased use of personal data. 2025 Presented at: Proceedings of the 5th International Scientific Conference Society Technology Solutions; Oct 15-17, 2025; Valmiera, Latvia p. 1-8.
28. Meskus M, Tikka E. Speculative approaches in social science and design research: methodological implications of working in 'the gap' of uncertainty. Qual Res 2024 Apr;24(2):209-228. [doi: 10.1177/14687941221129808]
29. Schuijer JW, Broerse JEW, Kupper F. Citizen science fiction: the potential of situated speculative prototyping for public engagement on emerging technologies. Nanoethics 2021 Apr;15(1):1-18. [doi: 10.1007/s11569-020-00382-4]
30. Blythe M. Research through design fiction: narrative in real and imaginary abstracts. 2014 Presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; Apr 26 to May 1, 2014; Toronto, Ontario, Canada p. 703-712. [doi: 10.1145/2556288.2557098]
31. LeMaster L, Johnson A. Speculative fiction, criticality, and futurity: an introduction. Commun Crit Cult Stud 2021 Jul 3;18(3):280-282. [doi: 10.1080/14791420.2021.1953699]
32. Olson W, Pierce J. Exploring the use of speculative concept films for co-speculation around data ethics. Presented at: CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems; Apr 26 to May 1, 2025; Yokohama Japan p. 1-21. [doi: 10.1145/3706598.3713245]
33. van 't Veer J, Wouters E, Veeger M, van der Lugt R. Ontwerpen voor Zorg en Welzijn [Book in Dutch]: Coutinho; 2020.
34. Sheringham J, Kuhn I, Burt J. The use of experimental vignette studies to identify drivers of variations in the delivery of health care: a scoping review. BMC Med Res Methodol 2021 Apr 22;21(1):81. [doi: 10.1186/s12874-021-01247-4] [Medline: 33888077]
35. Vereniging H. Nederlandse gedragscode wetenschappelijke integriteit [Report in Dutch]. : DANS Data Station Physical and Technical Sciences; 2018. [doi: 10.17026/DANS-2CJ-NVWU]
36. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
37. Riessman CK. Narrative Methods for the Human Sciences: Sage; 2008.
38. Creswell JW, Inoue M. A process for conducting mixed methods data analysis. J Gen Fam Med 2025 Jan;26(1):4-11. [doi: 10.1002/jgf2.736] [Medline: 39776872]
39. Urquhart C, Cheuk B, Lam L, Snowden D. Sense‐making, sensemaking and sense making—a systematic review and meta‐synthesis of literature in information science and education: An Annual Review of Information Science and Technology (ARIST) paper. Asso for Info Science & Tech 2025 Jan;76(1):3-97. [doi: 10.1002/asi.24866]

XSL•FO
RenderX

40. van Velsen L, Ludden G, Grünloh C. The limitations of user-and human-centered design in an eHealth context and how to move beyond them. J Med Internet Res 2022 Oct 5;24(10):e37341. [doi: 10.2196/37341] [Medline: 36197718]
41. Vandekerckhove P, de Mul M, Bramer WM, de Bont AA. Generative participatory design methodology to develop electronic health interventions: systematic literature review. J Med Internet Res 2020 Apr 27;22(4):e13780. [doi: 10.2196/13780] [Medline: 32338617]
42. Tsekleves E, Yong MH, Lee CAL, Giga S, Hwang JS, Lau SL. Rethinking how healthcare is conceptualised and delivered through speculative design in the UK and Malaysia: a comparative study. Design J 2019 Apr 1;22(sup1):429-444. [doi: 10.1080/14606925.2019.1595430]
43. Qi X, Yu J. Participatory design in human-computer interaction: cases, characteristics, and lessons. 2025 Apr 26 Presented at: CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems; Apr 26 to May 1, 2025; Yokohama, Japan p. 1-26 URL: https://dl.acm.org/doi/proceedings/10.1145/3706598 [doi: 10.1145/3706598.3713436]
44. Moore G, Wilding H, Gray K, Castle D. Participatory methods to engage health service users in the development of electronic health resources: systematic review. J Particip Med 2019 Feb 22;11(1):e11474. [doi: 10.2196/11474] [Medline: 33055069]
45. Kyololo OM, Stevens BJ, Songok J. Photo-elicitation technique: utility and challenges in clinical research. Int J Qual Methods 2023 Oct;22:16094069231165714. [doi: 10.1177/16094069231165714]
46. Brown T. Design thinking. Harv Bus Rev 2008 Jun;86(6):84-92. [Medline: 18605031]
47. Barrish P. Speculative fiction and the political economy of healthcare: Chang-Rae Lee's On Such a Full Sea. J Med Humanit 2019 Sep;40(3):297-313. [doi: 10.1007/s10912-018-9514-4] [Medline: 29860646]
48. Johnson T, Romanis EC. The relationship between speculation and translation in Bioethics: methods and methodologies. Monash Bioeth Rev 2023 Dec;41(S1):1-19. [doi: 10.1007/s40592-023-00181-z]
49. Berger R. Now I see it, now I don't: researcher's position and reflexivity in qualitative research. Qual Res 2015 Apr;15(2):219-234. [doi: 10.1177/1468794112468475]

Abbreviations

AI: artificial intelligence

Edited by B Mesko; submitted 19.Dec.2025; peer-reviewed by E Kendal, L Pretorius, X Ling; revised version received 24.Feb.2026; accepted 09.Mar.2026; published 22.May.2026.

Please cite as:
Groeneveld S, Wentzel J, Laurens M, Gemert-Pijnen LV, Verdaasdonk RM, Os-Medendorp HV, Ouden MD
The Use of Speculative Fiction in Future-Focused Health Care Research
J Med Internet Res 2026;28:e89699
URL: https://www.jmir.org/2026/1/e89699
doi:10.2196/89699

© Sjors Groeneveld, Jobke Wentzel, Melissa Laurens, Lisette van Gemert-Pijnen, Rudolf M Verdaasdonk, Harmieke van Os-Medendorp, Marjolein den Ouden. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Rerouting Eye Care: How AI and Telemedicine Are Reshaping Ophthalmology Patient Journeys in the United Kingdom and Germany

Broder Poschkamp[1,2], MSc, MD; Ariel Yuhan Ong[2,3,4,5], MD; Justin Engelmann[2,3], PhD; Michael Grün[2,6], MD; Andreas Stahl[1], MD; Pearse Andrew Keane[2,3,5], MD; David Adrian Merle[2,3,5,7], MD, MSc

[1]Department of Ophthalmology, Universitätsmedizin Greifswald, Ferdinand Sauerbruch Street, Greifswald, Mecklenburg-Vorpommern, Germany
[2]Institute of Ophthalmology, University College London, London, England, United Kingdom
[3]Moorfields Eye Hospital NHS Foundation Trust, London, England, United Kingdom
[4]Oxford Eye Hospital, Oxford University Hospitals NHS Trust, Oxford, England, United Kingdom
[5]NIHR Moorfields Biomedical Research Centre, London, England, United Kingdom
[6]Department of Ophthalmology, St. Franziskus Hospital, Münster, North Rhine-Westphalia, Germany
[7]Center of Ophthalmology, University Hospital Tuebingen, Tuebingen, Germany

Corresponding Author:
Broder Poschkamp, MSc, MD
Department of Ophthalmology, Universitätsmedizin Greifswald, Ferdinand Sauerbruch Street, Greifswald, Mecklenburg-Vorpommern, Germany

Abstract

Artificial intelligence (AI) as a medical device is now progressively entering routine ophthalmic care, yet its impact is still mostly evaluated in terms of diagnostic performance rather than how it reshapes patient care pathways. This viewpoint argues that careful pathway design is crucial to implementing AI in ophthalmology so that it translates into real, practical benefits for patients and services. We propose a framework that classifies AI- and telemedicine-enabled eye care pathways by the initial point of contact and the role of AI: from direct human assessment to grader- or telemedicine-based use to AI-first contact and fully autonomous AI gatekeeping. We apply this framework to 2 contrasting health systems, the United Kingdom and Germany, focusing on common retinal diseases and cardiovascular risk assessment from retinal images. In the United Kingdom, AI is being introduced mainly as a modular upgrade to standardized programs layered onto a gatekeeper model centered on community optometrists and general practitioners. In Germany, direct access to office-based ophthalmologists, opportunistic screening, and commercial retail offerings are producing more fragmented clinician- and market-driven AI adoption. Our comparison shows that the same AI technologies can generate very different patient journeys across health systems. Therefore, their value depends not only on diagnostic performance but also on intentional pathway design, including clear escalation rules and structured information transfer. This includes appropriate task shifting from ophthalmologists to trained staff or new roles, such as image grading, tele-ophthalmology triage, or pathway management. Without such design, AI risks duplicate testing, incomplete or poorly coordinated referrals, and further fragmentation of care.

(J Med Internet Res 2026;28:e93170) doi:10.2196/93170

KEYWORDS

patient pathways; artificial intelligence as a medical device; tele-ophthalmology; ophthalmic artificial intelligence; ophthalmic AI; artificial intelligence; AI screening

Introduction

Artificial intelligence (AI) has emerged as a powerful tool for ophthalmic screening since around 2018, following the first pivotal trials of autonomous diabetic retinopathy (DR) systems [1]. It is now available in multiple regulatory jurisdictions across the European Union, the United Kingdom, China, India, and other regions [2,3]. Current regulatory approvals for artificial intelligence as a medical device (AIaMD) encompass the detection and monitoring of DR, glaucoma, and age-related macular degeneration (AMD), as well as emerging oculomics tasks such as biological age estimation and cardiovascular risk prediction [2]. However, despite strong diagnostic performance in controlled studies (typically retrospective), real-world uptake remains slow and heterogeneous, with substantial variation in how AI is integrated—if at all—into regional and national eye care pathways and health systems.

At the same time, the demand for eye care is rising: ophthalmology is now the busiest outpatient specialty in the United Kingdom, contributing the largest number of outpatient

XSL•FO
RenderX

visits and with demand projected to rise by more than 40% over the next 20 years [4]. In Germany, over the past 15 years, demand for ophthalmic care has risen by approximately 15% to 34%, whereas effective care capacity has remained almost unchanged [5].

The technical backbone of most ophthalmic AI screening programs is digital fundus photography, which permits the detection of multiple retinal and optic nerve diseases from a single, noninvasive imaging modality [6]. The wider availability of affordable cameras and stable internet connection has enabled tele-ophthalmology models in which images captured in primary care, community clinics, or retail settings are analyzed remotely by AIaMD and/or human graders. These models support digital-first pathways that triage which patients truly require in-person specialist care [7].

In this viewpoint and conceptual analysis, we examine how the interplay between currently available AIaMD and telemedicine may reshape patient journeys for DR, AMD, glaucoma, and cardiovascular risk prediction in the United Kingdom and Germany. We selected these 2 high-income health systems because they have similar burdens of these conditions but fundamentally different access structures. The UK health care system is comparatively more centralized and nationally coordinated, whereas the German system is more fragmented. Both countries are early adopters of AIaMD and tele-ophthalmology, yet they differ substantially in their care structures. This contrast makes them well suited to illustrate how the same technologies can generate divergent patient journeys. We first introduce a pragmatic framework to classify ophthalmic care pathways by care setting and level of AI autonomy. We then apply this framework to current and potential pathways in both health systems. Finally, we outline clinical, organizational, and policy implications for designing future AI- and telemedicine-enabled eye care pathways that streamline patient access to eye care while maintaining safety and equity. This viewpoint is intended for clinicians, service designers, policymakers, and other stakeholders involved in implementing AI-enabled eye care pathways.

Conceptual Framework

In this viewpoint, “screening” denotes first-line examinations in asymptomatic at-risk individuals to detect previously undiagnosed eye disease. “Surveillance” refers to repeated examinations to monitor patients with known or suspected disease or high-risk groups over time to trigger treatment or retreatment [8]. “Tele-ophthalmology” describes remote eye care using digital imaging and/or video. Within this domain, we distinguish between remote image grading, which is limited to diagnostic assessment, and telemedicine (or teleconsultation), which includes treatment capabilities such as issuing prescriptions or initiating therapy. Current AIaMD can be applied to autonomous diagnostic tasks such as image grading and to the surveillance of existing disease over time.

In this paper, a care pathway refers to the overall sequence from first presentation to diagnosis, referral, treatment, and follow-up, whereas “triage” denotes one step within that pathway that determines prioritization or onward referral. For our comparison of UK and German patient journeys, we introduce the term “retail-based screening” to describe eye assessments performed in supermarkets, optometry shops, or pharmacies outside traditional medical practices.

We distinguish 4 generic AI-enabled screening pathways (Figure 1). Whether AI only assists a grader or independently decides who is referred for further treatment determines whether a pathway is classified as semiautomated or fully autonomous. Pathway A represents conventional care, where the patient is examined directly by an ophthalmologist; AI may be used by the physician in the background, but it is not visible in the direct patient contact. In pathway B, images or cases are first assessed by a grader or through a telemedicine contact, which then decides whether referral to an ophthalmologist is required. In pathway C, an AI provides an initial assessment, but cases are still reviewed by a grader or via telemedicine before patients reach the treating ophthalmologist (semiautomated model: every case is seen by a human, in contrast to fully automated pathways where only a subset of cases is forwarded). In pathway D, AI makes the primary screening decision without a grader and decides which cases are referred for in-person treatment.

XSL•FO
RenderX

Figure 1. Four generic artificial intelligence (AI)–supported care pathways are distinguished. Pathway A is a physician-based pathway in which the patient is seen in person by an ophthalmologist; AI may be used in the background, but the patient does not have direct AI contact. In pathway B, the first contact is digital, for example, via image grading by a grader or a telemedicine consultation with a physician. Pathway C is a confirmatory model in which AI provides the initial assessment, followed by review by a grader or telemedicine physician before any in-person consultation. Pathway D is a model in which an AI system first analyzes all patients and, in a fully automated manner, decides which patients need to be seen by a physician. In pathways C and D, red arrows indicate whether a human review step is present (semiautomated) or absent (fully automated), with absence indicating a higher level of autonomy.

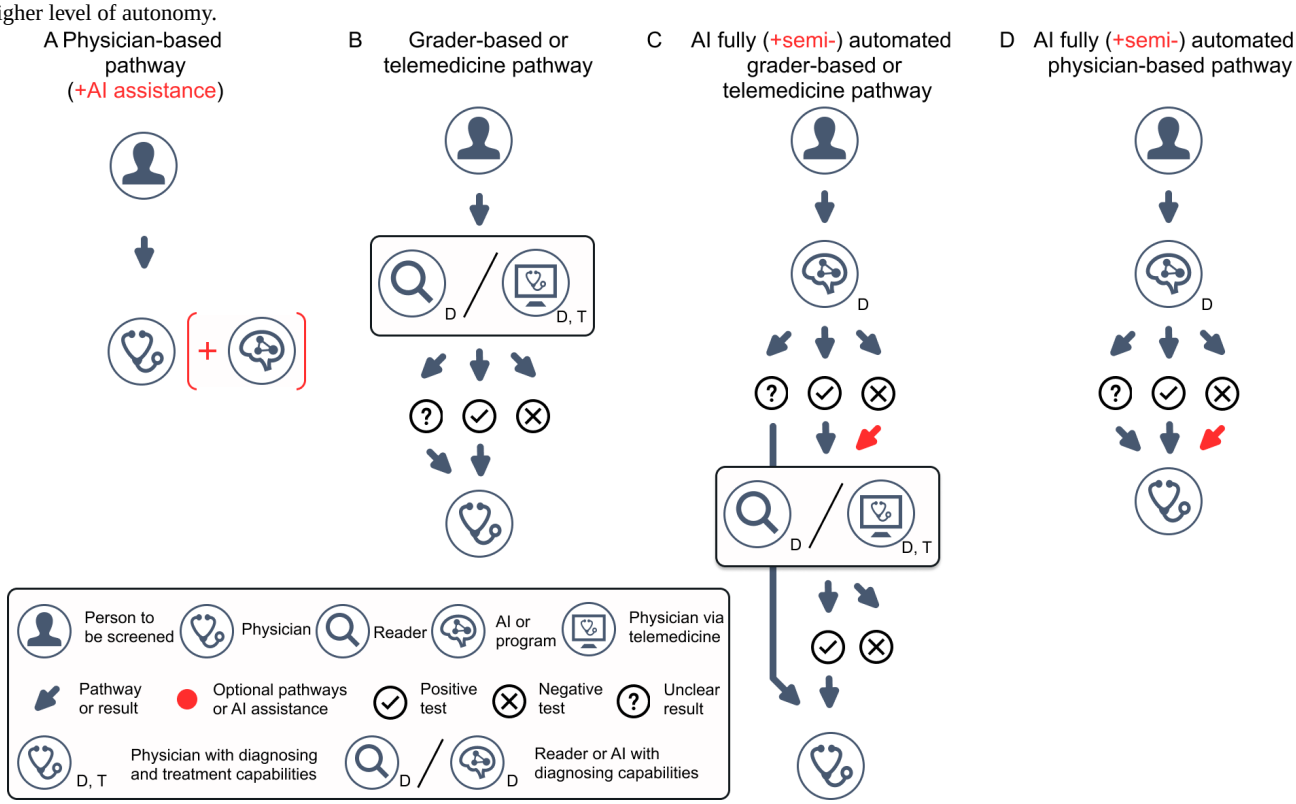

Patient Journeys in the United Kingdom

The United Kingdom's National Health Service (NHS) is a tax-funded national health system that provides universal coverage and generally free care at the point of use, with community eye care delivered largely through general ophthalmic services contracts and hospital eye services funded directly from NHS budgets. The NHS operates on a strict "gatekeeper" model, where access to specialist ophthalmology care is controlled by primary care providers rather than permitting direct specialist access. UK patients generally cannot consult a hospital ophthalmologist without a formal referral except in cases of acute emergency. While direct access to ophthalmologists is possible via the private practice route, this is not available in the public sector.

This structural constraint has necessitated specific detection and referral pathways to manage the flow of patients into secondary care. In this model, the community optometrist plays a significantly expanded clinical role compared to those in many other European nations, acting as the primary triage filter for the health system [9]. As the vast majority of these optometrists operate within commercial retail environments ("high-street optometry practices"), the UK model represents a public-private interface where retail technology adoption directly influences medical pathways. Patients presenting with visual symptoms or those attending for routine refraction generally visit a community optometrist first. If pathology is suspected, the optometrist issues a referral to the hospital eye service. In the context of our framework, this standard UK model represents conventional care (pathway A) but with a unique community-based prescreening layer performed by optometrists rather than physicians (Figure 2A).

XSL•FO
RenderX

Figure 2. Patient journeys in the United Kingdom. (A) The United Kingdom follows a gatekeeper model, meaning that patients usually cannot access hospital eye clinics directly. Patients with eye problems enter via pharmacy minor ailment schemes (often with referral to community optometry), general practitioners (GPs), or emergency departments (EDs), which can refer onward to the hospital eye clinic. Patients with diabetes may also be referred via the diabetic eye screening program (DESP). (B) Artificial intelligence (AI) as a medical device–enabled pathways are disease and service specific. In diabetic retinopathy, image assessment in the DESP (by a grader and/or AI) can trigger referral to hospital, bypassing the GP or optometry. In suspected neovascular age-related macular degeneration (AMD), referrals can originate from optometry, GPs, or the DESP and are channeled via an urgent fast track (approximately 2 weeks) to an AMD service. The glaucoma pathway may include refinement by specialist optometrists before hospital escalation, and retail screening can also direct patients into National Health Service care (eg, via the GP). Red arrows indicate urgent referral routes.

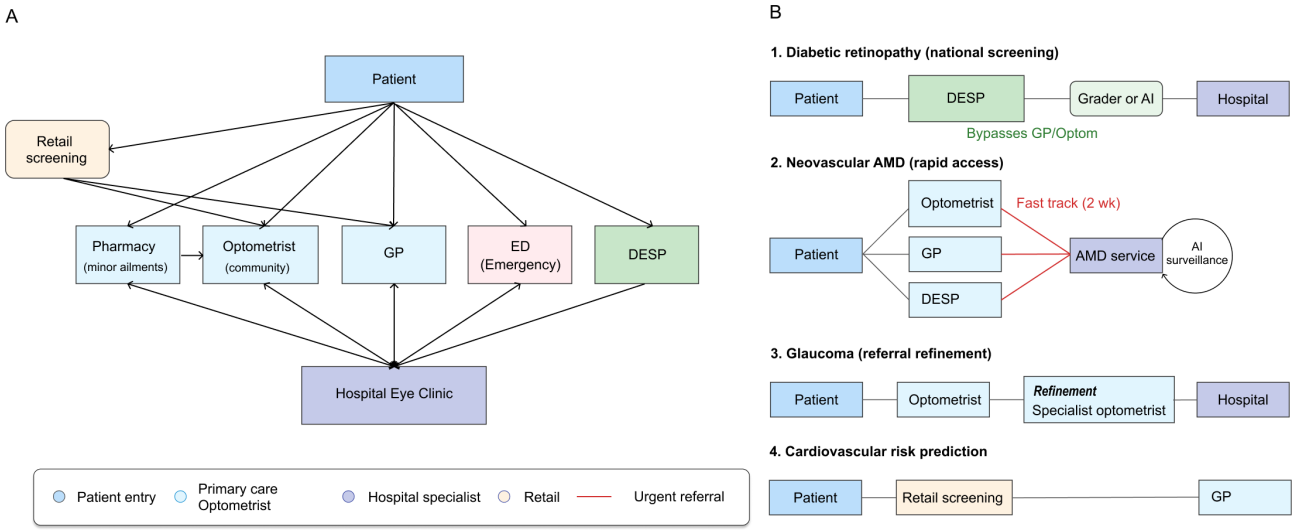

For DR, the patient journey shifts significantly from this opportunistic model to a systematic, active screening approach. The United Kingdom supports one of the world's most mature tele-ophthalmology programs, the NHS diabetic eye screening program (DESP), with each of the 4 nations (England, Scotland, Wales, and Northern Ireland) operating its own DESP with slightly different modes of operation [10]. All eligible individuals with diabetes are invited to dedicated community clinics or mobile units for regular screening. In this workflow, digital fundus images are captured by technicians and transmitted to a centralized grading center, where they are reviewed by trained human graders following a strict national protocol. This represents a telemedicine or human grading (pathway B) model. The English DESP is currently validating fully autonomous grader-based filtering models (pathway C), in which AI systems act as a filter to exclude "no disease" images (Figure 2B) [11], whereas the Scottish DESP has already deployed such an automated first-line grading system (eg, iGradingM) since 2011 (pathway C) [12].

In contrast, the journey for patients with neovascular age-related macular degeneration (nAMD) is usually symptom driven and time critical. Patients typically present to a primary care optometrist with acute symptoms such as distortion or central vision loss. To minimize delay, optometrists use "fast-track" electronic referral pathways to hospital eye services to achieve the national target of a specialist assessment within 2 weeks of referral [13]. These referrals can come from general practitioners (GPs) or from the DESP as well [14]. This represents a streamlined conventional care approach (pathway A); unlike the DR pathway, there is no intermediate telemedical grading step for diagnosis as the clinical suspicion raised by the optometrist triggers an immediate face-to-face specialist assessment. Recent data from a randomized clinical trial showed that embedding a tele-ophthalmology step into this urgent macula referral pathway can substantially reduce false-positive urgent referrals [15]. In parallel, an AI-enabled decision tool for treatment monitoring in nAMD is currently being prospectively evaluated within NHS retina clinics explicitly focusing on clinical safety and pathway integration (Figure 2B) [16].

Glaucoma detection in the United Kingdom remains largely opportunistic, occurring primarily during routine sight tests at optical shops, many of which are NHS funded for eligible risk groups such as individuals with a first-degree relative with glaucoma [17]. Historically, this approach has generated high false-positive referral rates due to the difficulty of diagnosing glaucoma based on intraocular pressure or optic disc appearance alone. To mitigate this, many regions commission glaucoma referral refinement services [18]. In this workflow, a patient flagged by a standard optometrist is not referred directly to the hospital but is instead re-examined by a specialist community optometrist with advanced qualifications. Only if this specialist confirms the suspicion is the patient referred to secondary care. This constitutes a sophisticated variation of conventional care (pathway A), introducing a community-based human filter to triage cases (Figure 2B) [19].

The United Kingdom is beginning to introduce oculomics into the eye care pathway to identify systemic health risks, although these applications remain exploratory, with limited evidence to date [20]. In this emerging model, AI algorithms analyze retinal fundus images captured during routine retail eye examinations to predict cardiovascular risk or risk factors [21,22]. Early deployments have been reported mainly in private optometry practices and private clinics, where cardiovascular risk scores are offered as an add-on to standard eye examinations on a self-pay basis outside formal NHS commissioning [23]. In the context of our framework, this corresponds to a fully autonomous digital-first pathway (pathway D), in which an AI system generates a systemic risk estimate that could, in principle,

XSL•FO
RenderX

prompt direct referral to a GP for cardiovascular assessment and management as necessary. However, these services remain at an exploratory stage, and uptake in routine optometric practice is limited, with outstanding questions regarding reimbursement models, workflow integration, governance, and professional responsibility (Figure 2B) [20].

Patient Journeys in Germany

In Germany, health care is mainly financed through statutory health insurance (GKV) with income-related contributions and near-universal coverage, with outpatient eye care provided largely by office-based ophthalmologists reimbursed by sickness funds and hospital eye services funded via statutory insurance and diagnosis-related group–based hospital payments [24]. In contrast to the United Kingdom's gatekeeper model, the German health care system is defined by a direct-access specialist structure. Patients seeking eye care possess the freedom to consult an office-based ophthalmologist directly [25]. However, the provision framework is more complex than a purely physician-based model. While eye care services are formally provided by ophthalmologists (and, to a minor extent, by GPs), opticians play a nonnegligible role in the initial contact. Although legally classified within the handicraft sector rather than as health care professionals, German opticians are entitled to perform screening tests, including tonometry, perimetry, and examinations of the anterior and posterior segment provided they refer any suspected abnormalities to a physician for formal diagnosis [26]. While the specialized role of the optometrist is gaining importance, it is not yet as widely available or as clearly defined in scope as in the United Kingdom. Consequently, despite this informal prescreening layer, the final human filter in Germany remains the ophthalmologist. In the context of our framework, the standard German model corresponds almost exclusively to conventional care (pathway A), where the final screening decision, diagnosis, and management are performed face-to-face by a physician (Figure 3A) [27].

Figure 3. Patient journeys in Germany. (A) In Germany, patients can directly consult an ophthalmologist. Entry may also occur via a general practitioner (GP) or via opticians or retail screening, which can direct patients to ophthalmology. Hospital eye clinics are typically involved for surgery and second opinions rather than first contact. (B) Artificial intelligence (AI) as a medical device–enabled pathways are disease and service specific. Diabetic retinopathy is detected during opportunistic examinations by ophthalmologists, with escalation to hospital when required. In neovascular age-related macular degeneration (AMD), care can remain decentralized, with possible AI-supported surveillance. Glaucoma includes additional entry routes (eg, opticians and self-paid or individual health services [IGeL]) before specialist assessment, with hospital involvement mainly for surgical care or second opinions. For cardiovascular risk prediction from retinal images, retail screening can direct patients onward to a GP or ophthalmology. *Surgeries or second opinion.

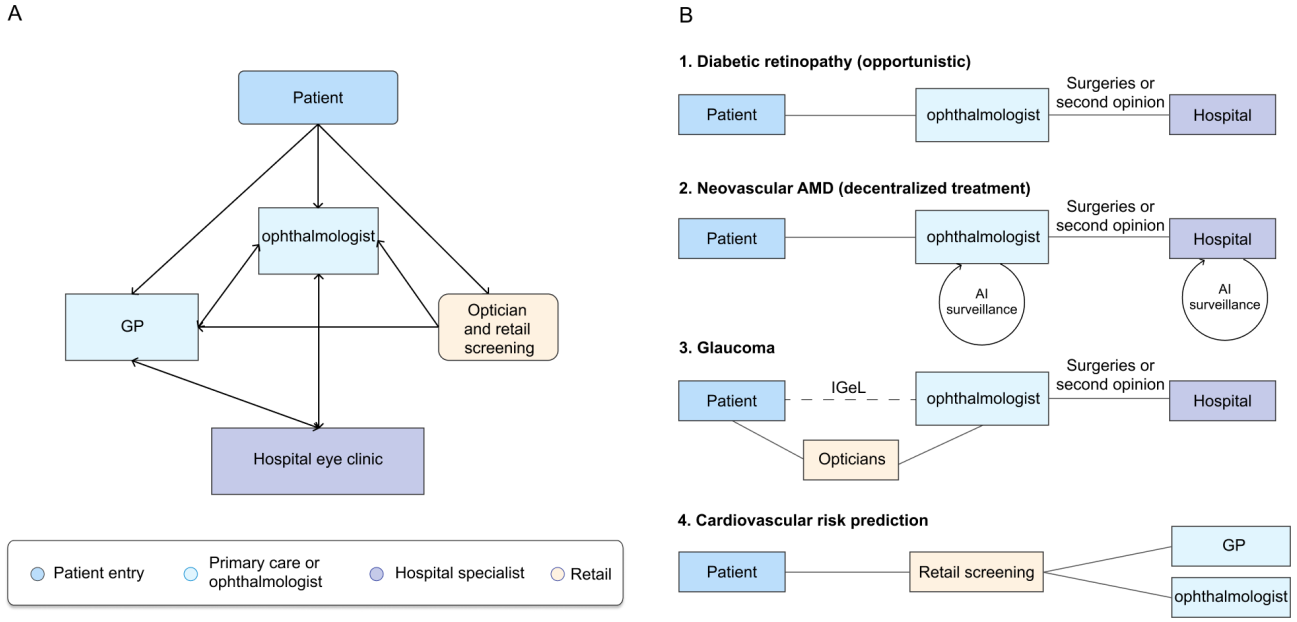

For DR, Germany relies on an opportunistic screening model. While national guidelines mandate annual or biannual eye examinations for patients with diabetes with and without comorbidities (eg, arterial hypertension or diabetes known for >10 years), respectively, the responsibility lies with the patient and their diabetologist to ensure that these visits occur [28]. Screening is performed exclusively by office-based ophthalmologists using slit lamp biomicroscopy with or without fundus photography (pathway A). Consequently, telemedicine or AI integration is almost nonexistent in routine care and remains largely limited to specific pilot projects [29]. More recently, retail-based screening models have introduced an additional semiautomated (all cases are reviewed), grader-based pathway (pathway C) [30]. For DR, hospital eye care is only necessary for surgery or complex cases or to provide a second opinion (Figure 3B).

The journey for nAMD in Germany is similarly shaped by office-based specialists. Patients noticing visual changes typically self-refer directly to their local ophthalmologist. Many German office-based ophthalmologists are equipped to administer intravitreal injections themselves or work in close networks with specialized injection centers [27]. Ophthalmologists in both office-based practices and hospital clinics can equip their imaging devices with AI tools that assist with AMD biomarker segmentation and classification [31].

XSL•FO
RenderX

However, these are pilot implementations, and they represent a physician-based pathway with AI support that does not meaningfully change the patient journey (Figure 3B).

Glaucoma detection follows an opportunistic and often commercially driven logic. As GKV does not routinely cover asymptomatic glaucoma screening, intraocular pressure measurements and optic disc assessments are frequently offered as individual health services, which patients must pay for out of pocket [24]. This creates a socioeconomic filter rather than a clinical one. When a patient accepts this screening, the diagnosis is confirmed directly by the ophthalmologist. Unlike the United Kingdom's referral refinement schemes, which use specialist optometrists to filter false positives, the German system absorbs this diagnostic load directly within the primary specialist layer (pathway A; Figure 3B).

The use of oculomics for cardiovascular disease risk assessment is only beginning to enter the German market and, at present, is confined to exploratory offerings in the retail optical sector [32]. As GKV does not yet provide reimbursement for retinal cardiovascular screening, this innovation currently follows a commercial pathway corresponding to pathway D. Similar to the United Kingdom, patients pay out of pocket at an optician for a retinal scan that can provide both an AI analysis of DR, glaucoma, and AMD and an AI-based cardiovascular risk assessment [33]. While such models illustrate how high-technology oculomics screening could be brought into community settings, they currently operate without formal integration into GKV-funded care pathways or clearly defined referral arrangements to GPs, cardiologists, or ophthalmologists (Figure 3B).

Digital-First Pathways

Digital-first models, particularly those emerging in the retail sector, create new entry points into eye care largely operated by nonmedical staff. While these services can improve access, they must not function as diagnostic dead ends without clear pathways to treatment or clinical decision-making. For DR, programs in which screening results are directly linked to a scheduled ophthalmology appointment have been shown to achieve substantially higher attendance rates (up to 5-fold higher) compared with models that only ask patients to make an appointment with an eye care specialist [34].

Critically, patients with acute "red flag" symptoms, such as sudden vision loss, new flashes and floaters, or severe ocular pain, should bypass routine screening workflows and be directed immediately to urgent specialist assessment rather than being delayed in retail or digital screening pathways [35]. This also implies that screening services themselves should not routinely route patients into the emergency department. In the United Kingdom, tele-ophthalmology services that triage urgent referrals have demonstrated the ability to reduce the burden on hospital eye departments by safely triaging cases to reduce hospital referrals from the community [15].

AI tools used in digital-first pathways should explicitly state their intended use, for example, whether they are designed for fully autonomous screening, semiautomated decision support, or purely as a grading aid. This intended use influences acceptable thresholds for sensitivity and specificity, legal responsibility, and how results are communicated to patients [36]. In addition, those implementing and governing AI-enabled services must define the intended downstream action of AI outputs. In the emerging field of oculomics, where retinal images are used to estimate systemic risk such as cardiovascular disease, it is particularly important to clarify and define the management pathway, for example, whether a high-risk result should trigger referral to a GP, cardiologist, or ophthalmologist rather than simply labeling patients as "high risk" without a clear pathway for follow-up.

Task Allocation and Information Transfer

For digital-first models to improve care rather than fragment it, examinations that trigger onward referral must provide sufficient structured information for the next clinician in the chain. AI reports should include key clinical variables (eg, relevant systemic risk factors, image quality, suspected diagnosis and severity, urgency, and recommended follow-up interval) rather than only a binary "refer/no refer" flag [14,28]. This is analogous to structured referral forms used in German diabetes eye care pathways, which record systemic risk factors such as diabetes type and duration, hemoglobin A_{1c} level, blood pressure, vascular complications, and an overall risk stratification, or to structured AMD referral templates in UK guidance. These reports should also be designed for seamless integration into existing electronic health record systems to ensure continuity and accessibility of information. Explainability may also be useful in this regard. Clinicians receiving or triaging the referral may wish to access the original image and the AI outputs to build trust in the AI system's decisions.

Comparison Between the United Kingdom and Germany

Although the United Kingdom and Germany are both early adopters of ophthalmic AIaMD and tele-ophthalmology, the organizational context in which these technologies are deployed differs substantially. In the United Kingdom, digital and tele-ophthalmology services are increasingly layered onto a gatekeeper-based system in which community optometrists and GPs act as primary entry points into eye care and control access to hospital eye services [37]. These differences influence where AI is inserted, who retains oversight, and whether implementation occurs as part of coordinated pathways or as isolated diagnostic add-ons.

In Germany, in contrast, many patients can access office-based ophthalmologists directly [24], and there is currently no significant nationwide reimbursement framework for AI-based medical devices in ophthalmology. In addition, most AI use is confined to pilot projects or local physician-based decision support tools that do not fundamentally alter the patient journey [38]. Given Germany's direct access to specialist care, physician-based AI (pathway A) may help save clinical time, but without coordinated pathway design, its use remains fragmented and largely dependent on individual

XSL•FO
RenderX

ophthalmologists' willingness and capacity to integrate these tools [39]. Table 1 summarizes the key structural contrasts between the 2 systems and their implications for AI-enabled ophthalmology pathways.

Table . Artificial intelligence (AI)–enabled ophthalmology pathways in the United Kingdom and Germany.

Dimension	United Kingdom	Germany
Health system organization	More centralized; nationally coordinated	More fragmented; regionally variable
Entry into eye care	Gatekeeper based	Direct specialist access
Main first contact	Community optometrist, GP[a], or organized screening	Office-based ophthalmologist
Existing tele-ophthalmology infrastructure	Stronger, especially in DR[b] screening	Limited; mainly pilot based
Current AI deployment pattern	More at the pathway level and programmatic	More local; physician or market driven
Position of human oversight	Often before specialist referral	Often at specialist level, who is the first contact
Likely AI implementation by use case	DR: AI-supported screening or grader prefiltering within organized programs; AMD[c]: referral refinement or treatment monitoring support; glaucoma: community-based referral refinement and risk stratification; oculomics: retail- or optometry-based cardiovascular risk estimation with unclear downstream integration	DR: opportunistic practice-based screening support or retail prescreening; AMD: physician-facing imaging decision support in specialist care; glaucoma: self-pay screening or in-practice risk support; oculomics: retail or self-pay cardiovascular risk assessment with unclear downstream integration
Main risk	Complex escalation and accountability across tiers	Fragmentation and lack of pathway integration; high number of false positives
Main policy priority	Safe scaling within existing referral systems	Building coordinated pathways around otherwise isolated AI tools

[a]GP: general practitioner.

[b]DR: diabetic retinopathy.

[c]AMD: age-related macular degeneration.

At the same time, in both the United Kingdom and Germany, emerging retail-based digital-first services are beginning to create retail-initiated pathways (pathways C and D), typically as self-pay services that have huge potential to reduce ophthalmologist workload. In both countries, many imaging tests (eg, color fundus photography and optical coherence tomography) are reimbursed only for specific indications and otherwise paid out of pocket. If retail screening is not integrated into shared care pathways and data flows, these risks duplicate testing and double charging.

Conclusions

Our comparison shows that the same AI technologies can lead to very different care pathways across health care systems. Therefore, to deploy AI at scale and translate it into meaningful benefit for patients and services, the next steps are not only better algorithms but also intentional pathway design: defining where AI sits in the journey, who receives its outputs, how follow-up is organized, how use is reimbursed, and how responsibility and oversight are shared. In practical terms, this includes establishing processes for low-quality or ungradable images, integrating AI outputs into routine records, and specifying clear escalation routes and follow-up intervals for urgent or uncertain findings. Experience from other countries and health system contexts, including India and broader implementation efforts across Europe, Australia, and America, suggests that ophthalmic AI can improve access and support screening but only when workflow design, referral pathways, and local health system context are addressed alongside algorithm performance [2,3,7]. Implementation may also require new roles in grading, tele-ophthalmology triage, referral refinement, and pathway management, together with the safe and feasible transfer of selected routine tasks from highly specialized ophthalmologists to appropriately trained staff. Future work should focus on co-designing and prospectively evaluating AI-enabled pathways with patients, clinicians, and payers in each system so that digital-first models improve access and efficiency without compromising safety or professional accountability.

Acknowledgments

During manuscript preparation, GPT-5.4 (OpenAI) and Microsoft Word's built-in spelling and grammar tools were used solely for language polishing, grammar correction, and improvement of phrasing. These tools were not used to generate scientific content, interpret findings, analyze data, or identify references. All output was critically reviewed, edited, and verified by the authors, who take full responsibility for the final manuscript.

XSL•FO
RenderX

Funding

This work received no dedicated project funding. BP is supported by the Gerhard Domagk Nachwuchsförderprogramm (Clinician Scientist Program Rural Age scholarship; project 493623784). AYO is supported by a National Institute for Health and Care Research Moorfields Eye Charity doctoral fellowship (NIHR303691). DAM is supported by the Euretina Retinal Medicine Clinical Research Award. MG is supported by the Dr. Werner Jackstädt Foundation. The funders had no involvement in the design, conduct, analysis, interpretation, or writing of this manuscript.

Data Availability

Data sharing is not applicable to this paper as no datasets were generated or analyzed during this study.

Conflicts of Interest

JE provides consulting services to industry clients via Machine Intelligence Solutions Ltd., including Skleo Health GmbH. These activities are unrelated to this work. PAK is a cofounder of Cascader Ltd; has acted as a consultant for Retina Consultants of America, Roche, Boehringer-Ingelheim, and Bitfount; and is an equity owner in Big Picture Medical. He has received speaker fees from Zeiss, Thea, Apellis, and Roche. He has received travel support from Bayer and Roche. He has attended advisory boards for Topcon Healthcare, Bayer, Boehringer-Ingelheim, and Roche. All other authors declare no other conflicts of interest.

References

1. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med 2018;1:39. [doi: 10.1038/s41746-018-0040-6] [Medline: 31304320]
2. Ong AY, Taribagil P, Sevgi M, et al. A scoping review of artificial intelligence as a medical device for ophthalmic image analysis in Europe, Australia and America. NPJ Digit Med 2025 May 29;8(1):323. [doi: 10.1038/s41746-025-01726-8] [Medline: 40442400]
3. Ramasamy K, Mishra C, Kannan NB, Namperumalsamy P, Sen S. Telemedicine in diabetic retinopathy screening in India. Indian J Ophthalmol 2021 Nov;69(11):2977-2986. [doi: 10.4103/ijo.IJO_1442_21] [Medline: 34708732]
4. Das A, Smith D, Mathew RG. Predictors of ophthalmology career success (POCS) study. BMJ Open Ophthalmol 2021;6(1):e000735. [doi: 10.1136/bmjophth-2021-000735] [Medline: 34322599]
5. Schuster AK, Wolfram C, Pfeiffer N, Finger RP. Ophthalmology 2019-where do we stand?: An analysis of the treatment situation in Germany [Article in German]. Ophthalmologe 2019;116(9):829-837. [doi: 10.1007/s00347-019-0894-2] [Medline: 31053964]
6. Li T, Bo W, Hu C, et al. Applications of deep learning in fundus images: a review. Med Image Anal 2021 Apr;69:101971. [doi: 10.1016/j.media.2021.101971] [Medline: 33524824]
7. Korot E, Wood E, Weiner A, Sim DA, Trese M. A renaissance of teleophthalmology through artificial intelligence. Eye 2019 Jun;33:861-863. [doi: 10.1038/s41433-018-0324-8]
8. Wilson JM, Jungner G. Principles and practice of screening. World Health Organization. 1968. URL: https://iris.who.int/server/api/core/bitstreams/762f0d2f-4225-46df-b060-c39b37b9d76d/content [accessed 2026-05-15]
9. Baker H, Ratnarajan G, Harper RA, Edgar DF, Lawrenson JG. Effectiveness of UK optometric enhanced eye care services: a realist review of the literature. Ophthalmic Physiol Opt 2016 Sep;36(5):545-557. [doi: 10.1111/opo.12312] [Medline: 27580754]
10. Scanlon PH. The English National Screening Programme for diabetic retinopathy 2003-2016. Acta Diabetol 2017 Jun;54(6):515-525. [doi: 10.1007/s00592-017-0974-1] [Medline: 28224275]
11. Macdonald T, Zhelev Z, Liu X, et al. Generating evidence to support the role of AI in diabetic eye screening: considerations from the UK National Screening Committee. Lancet Digit Health 2025 May;7(5):100840. [doi: 10.1016/j.landig.2024.12.004] [Medline: 40185647]
12. Philip S, Fleming AD, Goatman KA, et al. The efficacy of automated “disease/no disease” grading for diabetic retinopathy in a systematic screening programme. Br J Ophthalmol 2007 Nov;91(11):1512-1517. [doi: 10.1136/bjo.2007.119453] [Medline: 17504851]
13. Age related macular degeneration services: recommendations. : The Royal College of Ophthalmologists; 2024 URL: https://www.rcophth.ac.uk/wp-content/uploads/2021/08/AMD-Services-Commissioning-Guidance-Recommendtions.pdf [accessed 2026-05-15]
14. Chandra S, McKibbin M, Mahmood S, et al. The Royal College of Ophthalmologists Commissioning guidelines on age macular degeneration: executive summary. Eye (Lond) 2022 Nov;36(11):2078-2083. [doi: 10.1038/s41433-022-02095-2] [Medline: 35624304]
15. Sharma A, Hussain R, Learoyd AE, et al. Teleophthalmology versus standard of care for community optometry referrals of retinal disease (HERMES): a cluster randomised controlled trial with linked prospective diagnostic accuracy assessment of artificial intelligence support. Lancet Prim Care 2025 Oct;1(4):100041. [doi: 10.1016/j.lanprc.2025.100041]

XSL•FO
RenderX

16. Hogg HD, Brittain K, Teare D, et al. Safety and efficacy of an artificial intelligence-enabled decision tool for treatment decisions in neovascular age-related macular degeneration and an exploration of clinical pathway integration and implementation: protocol for a multi-methods validation study. BMJ Open 2023 Feb 1;13:e069443. [doi: 10.1136/bmjopen-2022-069443]
17. Free NHS eye tests and optical vouchers. National Health Service. URL: https://www.nhs.uk/nhs-services/opticians/free-nhs-eye-tests-and-optical-vouchers/ [accessed 2026-01-12]
18. Harper RA, Gunn PJ, Spry PG, Fenerty CH, Lawrenson JG. Care pathways for glaucoma detection and monitoring in the UK. Eye (Lond) 2020 Jan;34(1):89-102. [doi: 10.1038/s41433-019-0667-9] [Medline: 31700149]
19. Carmichael J, Abdi S, Balaskas K, Costanza E, Blandford A. The effectiveness of interventions for optometric referrals into the hospital eye service: a review. Ophthalmic Physiol Opt 2023 Nov;43(6):1510-1523. [doi: 10.1111/opo.13219] [Medline: 37632154]
20. Zhu Z, Wang Y, Qi Z, et al. Oculomics: current concepts and evidence. Prog Retin Eye Res 2025 May;106:101350. [doi: 10.1016/j.preteyeres.2025.101350] [Medline: 40049544]
21. Arnould L, Meriaudeau F, Guenancia C, et al. Using artificial intelligence to analyse the retinal vascular network: the future of cardiovascular risk assessment based on oculomics? A narrative review. Ophthalmol Ther 2023 Apr;12(2):657-674. [doi: 10.1007/s40123-022-00641-5] [Medline: 36562928]
22. Vaghefi E, An S, Corbett R, Squirrell D. Association of retinal image-based, deep learning cardiac BioAge with telomere length and cardiovascular biomarkers. Optom Vis Sci 2024 Jul 1;101(7):464-469. [doi: 10.1097/OPX.0000000000002158] [Medline: 38935034]
23. The future of heart health—now available through your eyes. John Rose Eyecare. 2025. URL: https://johnroseeyecare.co.uk/2025/05/31/the-future-of-heart-health-now-available-through-your-eyes/ [accessed 2026-01-12]
24. Busse R, Blümel M, Knieps F, Bärnighausen T. Statutory health insurance in Germany: a health system shaped by 135 years of solidarity, self-governance, and competition. Lancet 2017 Aug 26;390(10097):882-897. [doi: 10.1016/S0140-6736(17)31280-1] [Medline: 28684025]
25. Blümel M, Spranger A, Achstetter K, Maresso A, Busse R. Germany: health system review. Health Syst Transit 2020 Dec;22(6):1-272 [FREE Full text] [Medline: 34232120]
26. Seo JM, Kim HS, Sickenberger W, Kim SH. A study of the educational system for opticians and optometrists in Germany I. J Korean Ophthalmic Opt Soc 2019 Sep;24(3):205-214. [doi: 10.14479/jkoos.2019.24.3.205]
27. Weißbuch zur situation der ophthalmologischen versorgung in Deutschland (2023): ein leitfaden für die zukunft der ophthalmologie in Deutschland. Deutsche Ophthalmologische Gesellschaft. 2023. URL: https://dog.org/fuer-aerzte/weissbuecher-der-dog/dog-weissbuch-situation-der-ophthalmologischen-versorgung-2023 [accessed 2026-05-15]
28. S3-Leitlinie Nationale VersorgungsLeitlinie (NVL) typ-2-diabetes. AWMF. URL: https://register.awmf.org/de/leitlinien/detail/nvl-001 [accessed 2025-12-02]
29. Poschkamp B, Kantz L, Augstein P, et al. Customizing AI-based screening with real-world data: practical insights from diabetic retinopathy. Acta Ophthalmol 2026 May;104(3):323-332. [doi: 10.1111/aos.17591] [Medline: 40932047]
30. Skleo. URL: https://skleo.de/en/home/ [accessed 2026-05-15]
31. Gutfleisch M, Heimes-Bussmann B, Aydin S, et al. A post-hoc analysis of intravitreal aflibercept-treated nAMD patients from ARIES & ALTAIR: predicting treatment intervals and frequency for aflibercept treat-and-extend therapy regimen using machine learning. Graefes Arch Clin Exp Ophthalmol 2025 Jul;263(7):1885-1897. [doi: 10.1007/s00417-025-06812-x] [Medline: 40210713]
32. Netzhautanalyse mit KI-Auswertung bei Geertz Optik. Geertz Optik. URL: https://www.geertz-optik.de/service/netzhautanalyse/ [accessed 2026-01-13]
33. Hutton D. Toku obtains CE and UKCA marks for AI cardiovascular risk assessments through the eye. Ophthalmology Times Europe. 2024. URL: https://europe.ophthalmologytimes.com/view/toku-obtains-ce-and-ukca-marks-for-ai-cardiovascular-risk-assessments-through-the-eye [accessed 2025-12-05]
34. Sim DA, Mitry D, Alexander P, et al. The evolution of teleophthalmology programs in the United Kingdom: beyond diabetic retinopathy screening. J Diabetes Sci Technol 2016 Feb 1;10(2):308-317. [doi: 10.1177/1932296816629983] [Medline: 26830492]
35. Wilson H, Bhogal-Bhamra GK, Dhawahir-Scala F, Tromans C, Harper RA. A review of UK primary care pathways for acute ophthalmic conditions. Eye (Lond) 2025 Jan;39(1):45-56. [doi: 10.1038/s41433-024-03440-3] [Medline: 39496740]
36. Abràmoff MD, Tobey D, Char DS. Lessons learned about autonomous AI: finding a safe, efficacious, and ethical path through the development process. Am J Ophthalmol 2020 Jun;214:134-142. [doi: 10.1016/j.ajo.2020.02.022] [Medline: 32171769]
37. Anderson M, Pitchforth E, Edwards N, Alderwick H, McGuire A, Mossialos E. United Kingdom: health system review. Health Syst Transit 2022 May;24(1):1-194. [Medline: 35579557]
38. Beuse A, Grohmann C, Schadwinkel HM, Skevas C, Spitzer MS. Artificial intelligence for the detection of diabetic retinopathy. Klin Monbl Augenheilkd 2025 Sep;242(9):885-893. [doi: 10.1055/a-2545-1192] [Medline: 40456265]

XSL•FO
RenderX

39. Farah L, Davaze-Schneider J, Martin T, Nguyen P, Borget I, Martelli N. Are current clinical studies on artificial intelligence-based medical devices comprehensive enough to support a full health technology assessment? A systematic review. Artif Intell Med 2023 Jun;140:102547. [doi: 10.1016/j.artmed.2023.102547] [Medline: 37210155]

Abbreviations

AI: artificial intelligence
AIaMD: artificial intelligence as a medical device
AMD: age-related macular degeneration
DESP: diabetic eye screening program
DR: diabetic retinopathy
GKV: German statutory health insurance
GP: general practitioner
nAMD: neovascular age-related macular degeneration
NHS: National Health Service

Edited by A Mavragani; submitted 10.Feb.2026; peer-reviewed by D Gunasekeran, Z Su; revised version received 21.Apr.2026; accepted 30.Apr.2026; published 27.May.2026.

Please cite as:
Poschkamp B, Ong AY, Engelmann J, Grün M, Stahl A, Keane PA, Merle DA
Rerouting Eye Care: How AI and Telemedicine Are Reshaping Ophthalmology Patient Journeys in the United Kingdom and Germany
J Med Internet Res 2026;28:e93170
URL: https://www.jmir.org/2026/1/e93170
doi:10.2196/93170

© Broder Poschkamp, Ariel Yuhan Ong, Justin Engelmann, Michael Grün, Andreas Stahl, Pearse Andrew Keane, David Adrian Merle. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Extrinsic Trust as a Contractual Framework for Accountable AI in Health Care: Viewpoint

Anthony Kelly[1,2], PhD

[1]Department of Electronic and Computer Engineering, University of Limerick, Castletroy, Limerick, Ireland
[2]Health Research Institute, University of Limerick, Limerick, Ireland

Corresponding Author:
Anthony Kelly, PhD
Department of Electronic and Computer Engineering, University of Limerick, Castletroy, Limerick, Ireland

Abstract

Artificial intelligence (AI) promises efficiency and equity in health care. However, adoption remains fragmented due to weak foundations of trust. This Viewpoint highlights the gap between intrinsic trust, based on interpretability, and extrinsic trust, based on functional validation. We propose a contractual framework between AI systems and users defined by 3 promises: reliability, scope and equity, and shift and uncertainty. Illustrated through a vignette, we show how health systems can operationalize these promises through structured evidence and governance, translating trustworthy AI into accountable clinical deployment.

(J Med Internet Res 2026;28:e83903) doi:10.2196/83903

KEYWORDS

artificial intelligence; AI; explainable artificial intelligence; explainable AI; XAI; trust; machine learning; ML; mental health; decision support system; intrinsic trust

Introduction

Artificial intelligence (AI) has the potential to transform clinical care, with applications spanning triage, diagnosis, risk prediction, and resource allocation [1]. However, despite this technical promise, real-world adoption remains fragmented. Trust is a central barrier to adoption as clinicians and decision-makers remain hesitant to rely on AI in safety-critical environments, where errors carry significant consequences for patient outcomes, liability, and public confidence [2].

Adoption requires structured, credible evidence. A recent scoping review identified impediments including limited external validation data, transparency and equity concerns, workflow integration difficulties, and unclear accountability and highlighted facilitators such as robust validation, governance, and postdeployment monitoring [3].

Conceptually, the gap between AI engagement and adoption corresponds to a distinction between intrinsic and extrinsic trust. Intrinsic trust is the sense that clinicians gain when AI outputs are interpretable and aligned with clinical reasoning; extrinsic trust (or functionality trust) [4] rests on empirical evidence that the model performs reliably across diverse populations and under real-world conditions [5]. Intrinsic trust is necessary to encourage engagement and build confidence in the system, but extrinsic trust ultimately influences deployment decisions [5,6].

Establishing extrinsic trust is fundamentally a data and infrastructure challenge. We argue that health systems require structured approaches to generate, verify, and act on validation evidence under real-world conditions. Accordingly, we introduce a contractual framing of extrinsic trust supported by a minimum evidence package and a clinical vignette. Rather than proposing a new reporting artifact, the framework complements existing tools such as model cards, algorithmic audits, and assurance cases by treating trust as a small set of explicit, testable promises. Each promise is linked to defined evidence requirements, verification roles, and breach conditions that trigger governance action across development, deployment, and postdeployment monitoring.

Intrinsic and Extrinsic Trust in Clinical AI

Clinicians engage with AI systems in stages, beginning with how outputs align with their professional reasoning [7]. This early stage of intrinsic trust develops when predictions are interpretable and presented in ways that resonate with clinical heuristics [8-10].

Intrinsic trust provides a foundation for engagement, but decisions about deployment depend on further evidence. Clinicians require independent validation that the model performs reliably across different patient groups and that it can indicate when deference to human judgment is warranted in the face of predictive uncertainty [11-15]. Interpretable AI models are central to the sensemaking involved in intrinsic trust, but decisions about adoption require evidence of performance under real-world conditions [16]. To express this issue more succinctly, explainability fosters engagement, but validation governs deployment [5,12,14]. Intrinsic trust lowers the barrier to initial use, while extrinsic trust provides the assurance needed

XSL•FO
RenderX

for safe integration into workflows and health system decision-making.

Extrinsic trust validates that the model can generalize from the training examples in the dataset to previously unseen examples. As such, it must be either invariant to shift in the statistical distribution of the unseen data or convey increasing uncertainty commensurate with the shift. Since supervised machine learning models are known to be sensitive to out-of-distribution data shift, appropriate validation involves verifying that the model's sensitivity to out-of-distribution data shift is reliably conveyed through an appropriate uncertainty metric. Shifts within the distribution may occur in subgroups of the data, for example, in demographic or clinical subtypes. Appropriate validation for such in-distribution shifts involves verifying performance invariance. AI model calibration is concerned with how well a model's predicted probabilities align with actual outcomes and is therefore the foundation of reliability [17]. Calibration provides a foundation for trust by enabling users to gauge whether the AI "knows when it knows." Without external validation, especially under real-world conditions where data distributions may shift, clinicians may withhold trust or misplace it, leading to underuse or overreliance [12,14,18].

Consequently, the following issues need to be addressed through validation:

- The model should be reliable for in-distribution unseen data examples, demonstrating that predicted probabilities align with actual outcomes.
- Reliability should be maintained for in-distribution subgroups.
- Uncertainty metrics should convey higher uncertainty for out-of-distribution data, demonstrating that the model "knows when it knows."

Together, these questions define a pragmatic minimum for extrinsic trust rather than a complete taxonomy. They reflect 3 recurrent failure modes in clinical AI deployment: miscalibrated confidence, uneven subgroup performance, and silent degradation under distributional shift. Reliability and uncertainty are complementary. Reliability characterizes confidence alignment within scope, while uncertainty governs safe behavior as that alignment weakens. Scope and equity are treated jointly because subgroup shifts represent clinically salient in-distribution variation that population-level metrics can obscure.

Although these requirements are most relevant for discriminative AI models that predict classifications, there is emerging evidence that generative AI text models (large language models) can be queried about their reliability and uncertainty [19], perhaps allowing for future application of this framework.

Core Data Requirements for Extrinsic Trust

Similarly to trust in people, trust in AI has been framed as a contract [11,20,21]. Following this notion of contractual trust, users can rely on an AI system when explicit promises are stated and kept. Therefore, users must know what the AI is being trusted with and possess a means of evaluating whether the contract is adhered to [11]. This adherence depends on the context, and the means of determining adherence are dependent on the nature of the contract [8,11]. Drawing on these validation questions, we proposed 3 evidence clauses that should be adhered to for safe deployment (Figure 1) and detail the required metrics in Textbox 1.

XSL•FO
RenderX

Figure 1. Concentric framework of extrinsic trust. The 3 contractual promises of trust—reliability, scope and equity, and shift and uncertainty—are represented as radial sectors. Each promise operates across 3 levels: model-level evidence (inner ring), user-level trust and workflow fit (middle ring), and system-level governance and accountability (outer ring). At the core lies extrinsic trust, integrating these dimensions into the basis for responsible clinical artificial intelligence adoption. ECE: expected calibration error; SCE: static calibration error.

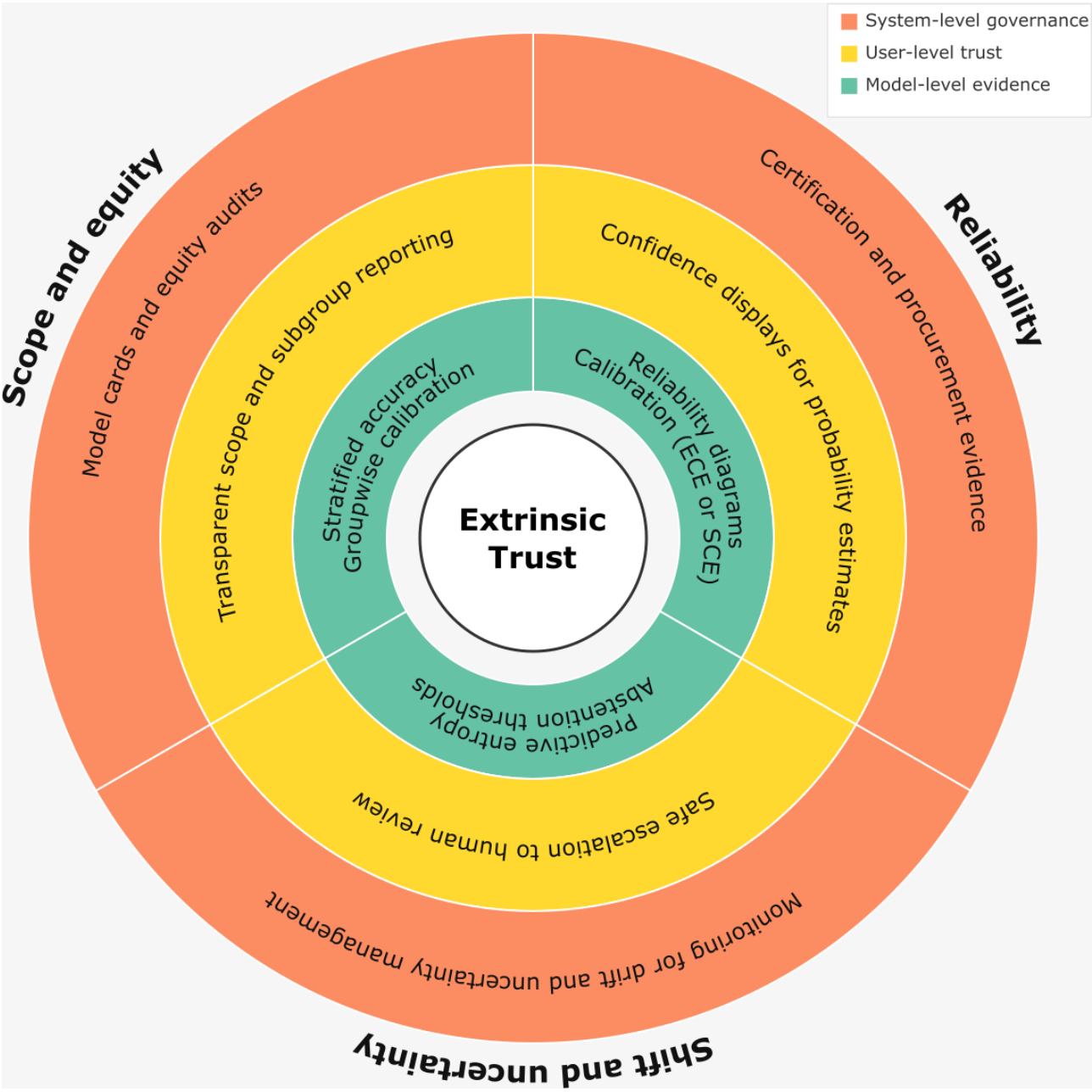

XSL•FO
RenderX

Textbox 1. Metrics for operationalizing the 3 promises of extrinsic trust in artificial intelligence systems.

Reliability promise

- Population-level performance metrics (eg, sensitivity, area under the receiver operating characteristic curve, and F_1-score)
- Expected calibration error or static calibration error: scalar measures of how closely predicted probabilities align with observed outcomes
- Reliability diagrams: visual summaries that clinicians can interpret to judge whether model confidence reflects real-world accuracy

Scope and equity promise

- Stratified performance metrics (eg, sensitivity, area under the receiver operating characteristic curve, and F_1-score): reported across clinically and demographically relevant subgroups to detect disparities
- Groupwise calibration plots: showing whether confidence is systematically over or underestimated for particular populations

Shift and uncertainty promise

- Negative log-likelihood: evaluates the quality of probabilistic predictions under noise or perturbation
- Predictive entropy or abstention rate: label-agnostic measures of uncertainty, useful for detecting out-of-distribution cases and supporting safe deferral to human review

The term "contractual" refers to a professional and governance compact rather than a legal indemnity framework and does not define liability. It makes explicit the performance promises that an AI system claims to uphold, the evidence required to support them, and the governance actions that follow if they are breached, thereby informing procurement and oversight.

In this compact, developers or vendors make explicit performance claims, health systems commission and govern verification, and clinicians rely on outputs under defined workflow conditions. A promise is meaningful only if keeping or breaching it is observable through agreed artifacts and escalation rules.

Recent work on trust in clinical AI emphasizes that trust is relational and bidirectional, emerging from interactions among models, clinicians, workflows, and institutional safeguards rather than from model properties alone. From this perspective, AI systems depend on human inputs, data quality, and organizational practices to remain safe and effective over time [22]. The contractual promises proposed in this paper are intended to complement this view by specifying the minimum evidence and governance commitments required to sustain trust within such sociotechnical systems, particularly through algorithmic audits and postdeployment monitoring.

Reliability Promise

The first requirement is a reliability promise: an AI system should provide probability estimates that clinicians can trust in the intended use context. High-confidence predictions should typically be correct, and uncertainty should be visible when the model is unsure.

Reliability depends on calibration: a well-calibrated model produces probability estimates that align with observed outcomes, giving clinicians a clear sense of when the system's confidence can be relied upon [23,24]. This alignment is crucial in safety-critical settings, where misplaced confidence can lead to harmful decisions. Tools such as reliability diagrams and scalar metrics such as the expected calibration error or its multiclass analogue, the static calibration error (SCE), provide concise ways of demonstrating this alignment by showing a good fit of the reliability curve to the diagonal and low values for the metrics (perfectly calibrated models have an SCE or expected calibration error of 0) [17,25,26]. Monitoring changes in these metrics or the reliability curve fit is useful as acceptable threshold levels are context dependent. Postcalibration scaling can further improve poorly calibrated models [26,27].

For health systems, requiring calibration evidence ensures that AI models offer dependable guidance that supports safe decision-making [12].

Scope and Equity Promise

The second requirement is a scope and equity promise: AI systems should clearly state who and what they are designed for and then provide evidence that they perform consistently across relevant clinical and demographic subgroups. Declaring scope makes explicit the intended populations, settings, and workflow assumptions, while equity validation ensures that no patient group is systematically disadvantaged.

Validation of the scope and equity promise involves stratified evaluation. Standard performance measures such as accuracy, sensitivity, area under the receiver operating characteristic curve, or F_1-score [28] should be reported not just at the population level but separately for subgroups defined by age, sex, ethnicity, or clinical subtype [29]. Calibration should also be examined within each stratum to detect systematic over- or underestimation. These analyses highlight where models are consistent and where subgroup-specific recalibration or data augmentation may be required. It should be acknowledged that small sample sizes may result from intersectional subgroups that may limit the ability to perform subgroup analysis in some cases.

From a regulatory and governance perspective, AI model cards offer a practical tool for this promise. They function as structured "labels," detailing intended use cases, target populations, known limitations, subgroup performance, and fairness considerations [30]. This transparency allows clinicians,

XSL•FO
RenderX

procurement teams, and regulators to assess whether systems are safe and appropriate for deployment.

Equity is fundamental because health populations are heterogeneous. A model that performs well overall may still underperform in underrepresented groups, amplifying disparities in care [12-14]. Mandating scope declarations and subgroup validation enables health systems to recognize both strengths and limits, supporting targeted safeguards where needed.

Shift and Uncertainty Promise

The third requirement is a shift and uncertainty promise: AI systems must provide safeguards when confronted with inputs that differ from the data on which they were trained. In evolving clinical environments, populations, referral patterns, and data quality change over time. Without mechanisms to detect and manage such shifts, models risk silent failure in precisely the contexts where reliability is most critical [12,31].

This promise requires evidence that the system can signal uncertainty, degrade gracefully, and defer to human review when appropriate. Metrics such as negative log-likelihood or predictive entropy capture probabilistic quality and uncertainty in unfamiliar cases [26,32,33]. In practice, confidence should decline, and error signals should increase in a predictable manner when inputs deviate or contain errors.

By enabling abstention or escalation under uncertainty, models help protect patients from misleading outputs, preserve equity across heterogeneous populations, and reduce risks of overreliance [34]. Embedding uncertainty clauses into validation frameworks ensures that decision-makers have the information to guarantee that AI remains trustworthy under real-world conditions.

Minimum Evidence Package for Extrinsic Trust

To formalize the contractual framework, we outline a minimum evidence package (Table 1) that specifies what constitutes proof of each promise, who is responsible for generating and verifying that proof, and what breach conditions require mitigation or suspension of deployment. In many cases, there is no accepted threshold for many of the metrics that constitute proof, although statistical threshold limits may be calculated on held-out calibration sets. For example, a material breach may be defined as a statistically significant upward shift in calibration error relative to the validated deployment baseline as determined by locally agreed monitoring procedures. Therefore, the minimum evidence package provides the foundation for a contract between the AI system and its users: a shared understanding that the model's behavior is recorded and that deviations can trigger governance actions to mitigate emerging risks before they manifest in clinical practice.

XSL•FO
RenderX

Table . Minimum evidence package for extrinsic trust. Baselines refer to the model's validated performance at deployment or the last approved update, with material deviation determined relative to locally defined governance thresholds.

Promise	Artifacts	Producer or verifier	Breach condition and action
All	• Model card documenting intended population, features, and workflow assumptions	• Producer: vendor or developer • Verifier: health system governance and clinical safety officer • Trigger: model commissioning and model update	• Breach: model card not satisfactory • Action: remedial documentation or clarification required before deployment or update
Reliability	• Population-level performance metrics (eg, sensitivity, AUC[a], and F_1-score) • Calibration metrics: ECE[b] or SCE[c] • Reliability diagrams	• Producer: vendor or developer • Verifier: health system governance and clinical safety officer • Trigger: annually; model update or change in patient profiles	• Breach: performance or calibration metrics deviate materially from the validated deployment baseline • Action: (1) recalibration or retraining, (2) investigation of the root cause, or (3) possible suspension
Scope and equity	• Stratified performance metrics (eg, subgroup: sensitivity, AUC, and F_1-score) • Groupwise calibration plots	• Producer: vendor or developer • Verifier: health system governance and clinical safety officer • Trigger: annually; model update or change in patient profiles	• Breach: subgroup performance or calibration deviates materially from the validated deployment baseline • Action: (1) subgroup analysis, (2) recalibration, (3) model card update to exclude subgroups, or (4) possible suspension
Shift and uncertainty	• Predictive entropy or abstention rate • NLL[d]	• Producer: vendor or developer • Verifier: health system governance and clinical safety officer • Trigger: annually; model update or change in patient profiles	• Breach: uncertainty metrics (eg, entropy or NLL) deviate materially from the validated deployment baseline • Action: (1) retraining of model with new data, (2) reverification of model, (3) model card update, or (4) possible suspension

[a]AUC: area under the receiver operating characteristic curve.

[b]ECE: expected calibration error.

[c]SCE: static calibration error.

[d]NLL: negative log-likelihood.

For clinicians and health system decision-makers, the contractual framing also provides a practical cognitive benefit. By organizing technical, regulatory, and statistical requirements into a small number of clear promises, it offers a concise mental model for what trustworthy AI deployment entails, helping cut through the complexity of standards and guidance.

The 3 promises align closely with emerging international regulatory requirements for the analysis and management of AI risks. For example, the European Union AI Act [35] requires transparency (Article 13), accuracy and robustness (Article 15), and risk monitoring (Articles 9 and 72). Although this applies primarily to AI medical devices [36], it nevertheless also implies best practice for lower-risk AI. Table 2 maps each promise to obligations in the European Union AI Act [35], Food and Drug Administration predetermined change control plan [37], and International Organization for Standardization and International Electrotechnical Commission 42001 standard [38], demonstrating that the proposed contractual framework provides a practical implementation framework consistent with regulatory trajectories in high-risk clinical AI.

XSL•FO
RenderX

Table . Mapping the 3 promises to emerging regulatory requirements.

Promise	EU[a] AI[b] Act (Regulation [EU] 2024/1689)	FDA[c] PCCP[d]	ISO[e] and IEC[f] 42001:2023 standard (AI management system)
Reliability promise	• Article 15: high-risk AI must achieve appropriate accuracy, robustness, and cybersecurity, and AI providers must declare accuracy metrics and levels and test or validate them • Article 9: requires an ongoing risk management system • Article 11+annex IV: technical documentation must include information such as performance characteristics and metrics and testing or validation evidence (as specified in annex IV)	The PCCP comprises (1) description of modifications, (2) modification protocol, and (3) impact assessment. The modification protocol should specify verification and validation activities, including performance evaluation methods, performance metrics, statistical tests, and predefined acceptance criteria, and describe how the manufacturer will determine that a modification is acceptable before implementation.	Clause 8.1 (operational planning and control) requires planned and controlled AI operations. Clause 9.1 (monitoring, measurement, analysis, and evaluation) requires organizations to determine what needs to be monitored and measured and how and when to evaluate AI system performance. Clause 10 (improvement) requires correction and continuous improvement when performance deviates.
Scope and equity promise	• Article 10: training, validation, and testing data must be sufficiently relevant and representative for the intended purpose, and AI providers must examine and mitigate bias risks • Article 13: transparency and instructions enabling appropriate use, including information needed for deployers to interpret outputs and use the system properly	The evidence and testing data supporting the PCCP should be representative of intended use populations (eg, demographic factors where relevant). The PCCP and its modifications must keep the device within the authorized intended use and indications; a change to what was described or authorized (eg, new population not covered) may require a new marketing submission.	Clauses 4.1 - 4.3 require definition of organizational context, interested parties, and the scope of the AI management system. Clause 6.1 (actions to address risks and opportunities) requires identification and treatment of AI risks, including risks to individuals and groups. Clause 6.1.4 (AI system impact assessment) requires assessment of impacts on individuals and groups affected by AI systems.
Shift and uncertainty promise	• Article 72: AI providers must operate a postmarket monitoring system to collect, document, and analyze performance and risks during real-world use • Article 15: robustness expectations (accuracy, robustness, and cybersecurity) support managing performance degradation in foreseeable conditions	The PCCP's modification protocol should define how the manufacturer will detect the need for change and evaluate modifications; examples include identifying triggers, such as when drift in data is observed. The protocol should specify how performance will be assessed after the change (metrics, tests, or acceptance criteria) and state that modifications should not be implemented if acceptance criteria are not met.	Clause 9.1 requires ongoing monitoring and evaluation of AI systems. Clause 9.2 (internal audit) requires periodic audits to detect nonconformity. Clause 10.2 (nonconformity and corrective action) requires organizations to respond to deviations, investigate causes, and implement corrective actions supporting structured responses to model drift and performance degradation.

[a]EU: European Union.
[b]AI: artificial intelligence.
[c]FDA: Food and Drug Administration.
[d]PCCP: predetermined change control plan.
[e]ISO: International Organization for Standardization.
[f]IEC: International Electrotechnical Commission.

Case Study Vignette: Operationalizing the 3 Promises Using the Probabilistic Integrated Scoring Model

To illustrate how extrinsic trust can be operationalized, we turn to the Probabilistic Integrated Scoring Model (PrISM) [39], a prototype AI classification system developed for initial treatment screening in routine mental health care. PrISM is a transparent, interpretable neural network that integrates the structured logic of psychometric instruments with probabilistic estimation.

The PrISM study involved a secondary analysis of questionnaire and demographic data from 1068 patients treated in Internetpsykiatrien, Denmark's national internet-based mental health service. Further details related to this vignette can be found in Multimedia Appendix 1.

Regarding the reliability promise, reliability diagrams show how closely the model's predicted probabilities align with observed outcomes. Figure 2 shows that PrISM confidence estimates remain near the ideal diagonal, with an SCE of 4.10%. For a clinician reviewing a referral, this provides assurance that, when the model appears confident, it is generally correct,

XSL•FO
RenderX

supporting decisions about when to trust AI guidance and when to seek further input.

Figure 2. Illustrative example of a reliability diagram for a predictive multiclass model (Probabilistic Integrated Scoring Model). The static calibration error (SCE) of 4.10% and close adherence to the diagonal indicate that the predicted probabilities closely match observed outcomes.

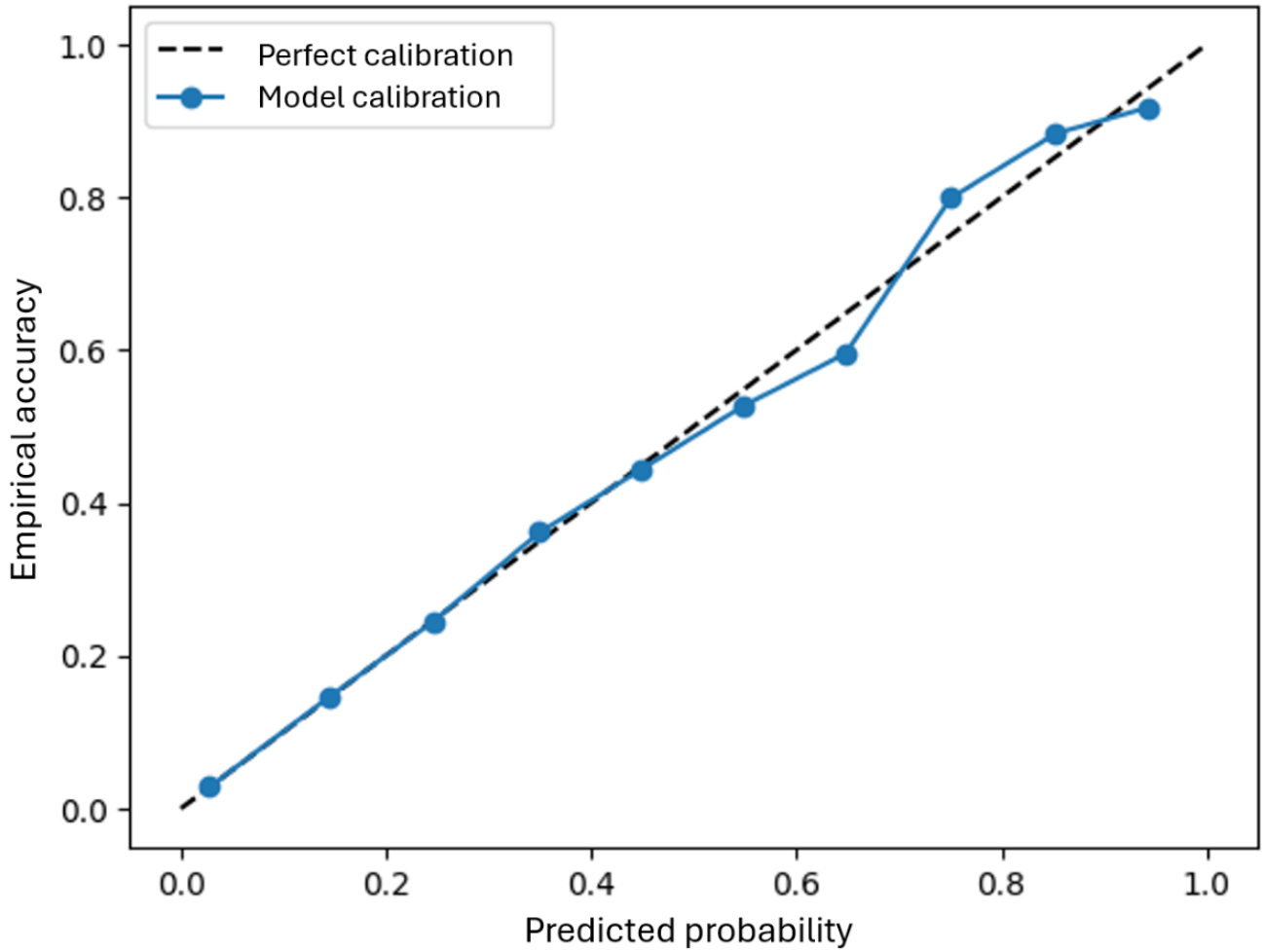

Regarding the scope and equity promise, PrISM is explicitly designed and framed as a tool for screening referrals in a digital clinic for mild to moderate depression and anxiety, clearly stating who and what it is designed for. Validation includes stratified calibration and discrimination (eg, by sex or presenting problem), with deployment contingent on subgroup performance remaining within a predefined deviation from the population baseline. This analysis highlights where performance is consistent and where disparities might emerge, enabling clinicians and service managers to anticipate limitations and plan mitigations rather than encountering them unexpectedly in practice.

Regarding the shift and uncertainty promise, in deployment, AI systems encounter cases that differ from their training data. PrISM addresses this by signaling uncertainty. When data inputs contain errors or deviate from their familiar distribution, the uncertainty metric (negative log-likelihood) rises, and the system may abstain from issuing a recommendation, routing the case for human review (Figure 3). For example, cases exceeding the 95th percentile of validation set uncertainty may trigger abstention and clinician review, with abstention rates monitored longitudinally as a safety signal. For clinicians, this safeguard functions as a clear indicator that the model recognizes its limits, ensuring that safe escalation is possible.

XSL•FO
RenderX

Figure 3. Illustrative example of deterioration of uncertainty (Probabilistic Integrated Scoring Model). As data noise increases (x-axis), the negative log-likelihood (NLL) uncertainty metric increases, whereas the accuracy performance metric stays relatively constant. This illustrates that the model's performance may be robust to data errors, but the model signals its increasing lack of confidence appropriately.

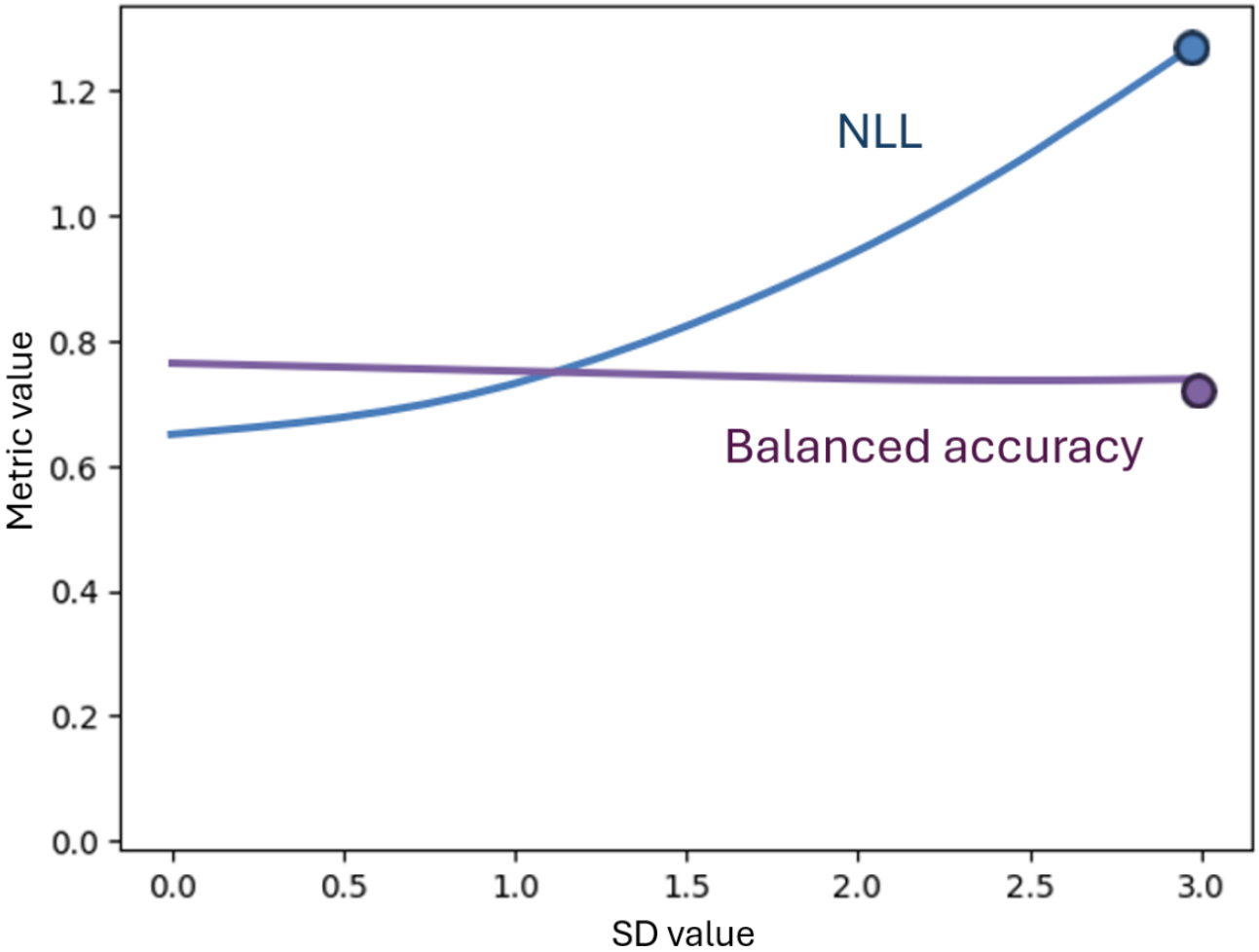

This vignette demonstrates that extrinsic trust can be embedded through design choices that make reliability, scope, and uncertainty measurable. The PrISM vignette shows how trust promises can be translated into clinical evidence, allowing health systems to evaluate, monitor, and govern AI as accountable decision support tools.

Health System Implications and Future Directions

Embedding extrinsic trust requires health systems to translate the 3 contractual promises into routine governance and clinical practice. In practical terms, this means requiring calibration, subgroup performance, and uncertainty evidence at procurement; defining verification roles within clinical safety structures; and specifying escalation actions when promises are breached.

Evidence of trust must be accessible and actionable at the point of care. Simple signals such as confidence or abstention indicators can reduce cognitive burden for frontline clinicians, while dashboards can surface detailed calibration, equity, and drift metrics for governance teams. When systems defer under uncertainty, clear escalation pathways are needed to avoid automation bias or clinician deskilling and preserve human oversight.

Looking ahead, this contractual framing aligns with regulatory shifts toward continuous assurance rather than one-time approval. By treating reliability, scope and equity, and shift and uncertainty as enforceable commitments, health systems gain a lightweight but robust mechanism to govern clinical AI across its life cycle as real-world conditions evolve.

Conclusions

Extrinsic trust is the decisive factor that determines whether clinical AI remains experimental or becomes a sustainable component of care. While intrinsic trust, supported by interpretability, enables clinician engagement, adoption in safety-critical settings depends on structured evidence of performance under real-world conditions.

By framing extrinsic trust as a contractual relationship defined by 3 promises of reliability, scope and equity, and shift and uncertainty, this Viewpoint provides a practical lens for evaluating, governing, and sustaining clinical AI systems. The PrISM vignette demonstrates that these promises can be operationalized through calibration, subgroup validation, and uncertainty-aware escalation. Together, this approach shifts trust from an abstract aspiration to a testable, accountable basis for responsible clinical AI deployment.

XSL•FO
RenderX

Acknowledgments

The author would like to express gratitude to the researchers and staff of Internetpsykiatrien, Denmark's national internet-based mental health service, whose conversations and interactions helped inspire this viewpoint. Generative artificial intelligence was used to improve clarity and language in some author-written paragraphs.

Funding

The author is funded by Innovation Fund Denmark.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to ethical and legal constraints surrounding patient data but can be made available from the corresponding author on reasonable request.

Authors' Contributions

AK conceived the viewpoint, developed the models and results, and wrote the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details of the Probabilistic Integrated Scoring Model vignette.
[DOCX File, 40 KB - jmir_v28i1e83903_app1.DOCX]

References

1. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
2. Afroogh S, Akbari A, Malone E, Kargar M, Alambeigi H. Trust in AI: progress, challenges, and future directions. Humanit Soc Sci Commun 2024;11(1):1568. [doi: 10.1057/s41599-024-04044-8]
3. Hassan M, Kushniruk A, Borycki E. Barriers to and facilitators of artificial intelligence adoption in health care: scoping review. JMIR Hum Factors 2024 Aug 29;11:e48633. [doi: 10.2196/48633] [Medline: 39207831]
4. Choung H, David P, Ross A. Trust and ethics in AI. AI Soc 2023 Apr;38(2):733-745. [doi: 10.1007/s00146-022-01473-4]
5. Shin D. The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI. Int J Hum Comput Stud 2021 Feb;146:102551. [doi: 10.1016/j.ijhcs.2020.102551]
6. Choung H, David P, Ross A. Trust in AI and its role in the acceptance of AI technologies. Int J Hum Comput Interact 2023 May 28;39(9):1727-1739. [doi: 10.1080/10447318.2022.2050543]
7. Lipton ZC. The mythos of model interpretability. Commun ACM 2018 Sep 26;61(10):36-43. [doi: 10.1145/3233231]
8. Kelly A, Bhardwaj N, Holmberg Sainte-Marie TT, et al. Investigating how clinicians form trust in an AI-based mental health model: qualitative case study. JMIR Hum Factors 2025 Dec 19;12:e79658. [doi: 10.2196/79658] [Medline: 41417472]
9. Ali S, Abuhmed T, El-Sappagh S, et al. Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. Inf Fusion 2023 Nov;99:101805. [doi: 10.1016/j.inffus.2023.101805]
10. Joyce DW, Kormilitzin A, Smith KA, Cipriani A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. NPJ Digit Med 2023 Jan 18;6(1):6. [doi: 10.1038/s41746-023-00751-9] [Medline: 36653524]
11. Jacovi A, Marasović A, Miller T, Goldberg Y. Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI. Presented at: FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; Mar 3-10, 2021. [doi: 10.1145/3442188.3445923]
12. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digit Med 2021 Jan 5;4(1):4. [doi: 10.1038/s41746-020-00367-3] [Medline: 33402680]
13. Malinin A, Gales M. Predictive uncertainty estimation via prior networks. Presented at: NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems; Dec 3-8, 2018. [doi: 10.5555/3327757.3327808]
14. Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. Nat Mach Intell 2019;1(1):20-23. [doi: 10.1038/s42256-018-0004-1]
15. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci Rep 2016 May 17;6:26094. [doi: 10.1038/srep26094] [Medline: 27185194]
16. Gawlikowski J, Tassi CRN, Ali M, et al. A survey of uncertainty in deep neural networks. Artif Intell Rev 2023 Oct;56(S1):1513-1589. [doi: 10.1007/s10462-023-10562-9]

XSL•FO
RenderX

17. Riley RD, Archer L, Snell KIE, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ 2024 Jan 15;384:e074820. [doi: 10.1136/bmj-2023-074820] [Medline: 38224968]
18. Zhu F, Zhang XY, Cheng Z, Liu CL. Revisiting confidence estimation: towards reliable failure prediction. IEEE Trans Pattern Anal Mach Intell 2024 May;46(5):3370-3387. [doi: 10.1109/TPAMI.2023.3342285] [Medline: 38090830]
19. Wei J, Karina N, Chung HW, et al. Measuring short-form factuality in large language models. arXiv. Preprint posted online on Nov 7, 2024. [doi: 10.48550/arXiv.2411.04368]
20. Hawley K. Trust, distrust and commitment. Nous 2014 Mar;48(1):1-20. [doi: 10.1111/nous.12000]
21. Tallant J, Donati D. Trust: from the philosophical to the commercial. Philos Manag 2020 Mar;19(1):3-19. [doi: 10.1007/s40926-019-00107-y]
22. Sagona M, Dai T, Macis M, Darden M. Trust in AI-assisted health systems and AI's trust in humans. NPJ Health Syst 2025;2(1):10. [doi: 10.1038/s44401-025-00016-5]
23. Nixon J, Dusenberry MW, Jerfel G, Zhang L, Tran D. Measuring calibration in deep learning. Presented at: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019); Jun 15-20, 2019 URL: https://openaccess.thecvf.com/content_CVPRW_2019/html/Uncertainty_and_Robustness_in_Deep_Visual_Learning/Nixon_Measuring_Calibration_in_Deep_Learning_CVPRW_2019_paper.html [accessed 2026-02-26]
24. Ovadia Y, Fertig E, Ren J, et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Presented at: NIPS'19: 33rd International Conference on Neural Information Processing Systems; Dec 8-14, 2019. [doi: 10.5555/3454287.3455541]
25. Naeini MP, Cooper GF, Hauskrecht M. Obtaining well calibrated probabilities using Bayesian binning. Proc AAAI Conf Artif Intell 2015 Jan;2015:2901-2907 [FREE Full text] [Medline: 25927013]
26. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. Presented at: 34th International Conference on Machine Learning (ICML 2017); Aug 6-11, 2017 URL: https://proceedings.mlr.press/v70/guo17a.html [accessed 2026-02-26]
27. Laves MH, Ihler S, Kortmann KP, Ortmaier T. Well-calibrated model uncertainty with temperature scaling for dropout variational inference. Presented at: 4th Workshop on Bayesian Deep Learning (NeurIPS 2019); Dec 13-14, 2019 URL: https://www.bayesiandeeplearning.org/2019/papers/77.pdf [accessed 2026-02-26]
28. Hicks SA, Strümke I, Thambawita V, et al. On evaluation metrics for medical applications of artificial intelligence. Sci Rep 2022 Apr 8;12(1):5979. [doi: 10.1038/s41598-022-09954-8] [Medline: 35395867]
29. James G, Witten D, Hastie T, Tibshirani R, Taylor J. An Introduction to Statistical Learning: With Applications in Python: Springer; 2023. URL: https://link.springer.com/book/10.1007/978-3-031-38747-0 [doi: 10.1007/978-3-031-38747-0]
30. Gilbert S, Adler R, Holoyad T, Weicken E. Could transparent model cards with layered accessible information drive trust and safety in health AI? NPJ Digit Med 2025 Feb 25;8(1):124. [doi: 10.1038/s41746-025-01482-9] [Medline: 40000736]
31. Ku WL, Min H. Evaluating machine learning stability in predicting depression and anxiety amidst subjective response errors. Healthcare (Basel) 2024 Mar 10;12(6):625. [doi: 10.3390/healthcare12060625] [Medline: 38540589]
32. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); Dec 4-9, 2017. [doi: 10.5555/3295222.3295387]
33. Rasmussen CE, Quiñonero-Candela J, Sinz F, Bousquet O, Schölkopf B. Evaluating predictive uncertainty challenge. In: Quiñonero-Candela J, Dagan I, Magnini B, d'Alché-Buc F, editors. Machine Learning Challenges Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment: Springer; 2006:1-27. [doi: 10.1007/11736790_1]
34. He W, Jiang Z, Xiao T, Xu Z, Li Y. A survey on uncertainty quantification methods for deep learning. ACM Comput Surv 2023;37(4):111 [FREE Full text]
35. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). European Union. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng [accessed 2026-02-27]
36. van Kolfschooten H, van Oirschot J. The EU Artificial Intelligence Act (2024): implications for healthcare. Health Policy 2024 Nov;149:105152. [doi: 10.1016/j.healthpol.2024.105152] [Medline: 39244818]
37. Predetermined change control plans for medical devices. U.S. Food and Drug Administration. 2024. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/predetermined-change-control-plans-medical-devices [accessed 2026-02-27]
38. ISO/IEC 42001:2023 - information technology — artificial intelligence — management system. International Organization for Standardization. 2023. URL: https://www.iso.org/standard/42001 [accessed 2026-02-27]
39. Kelly A, Jensen EK, Grua EM, Mathiasen K, Van de Ven P. An interpretable model with probabilistic integrated scoring for mental health treatment prediction: design study. JMIR Med Inform 2025 Mar 26;13:e64617. [doi: 10.2196/64617] [Medline: 40138679]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
PrISM: Probabilistic Integrated Scoring Model
SCE: static calibration error

Edited by A Coristine; submitted 10.Sep.2025; peer-reviewed by S Mohanadas, T Dai; revised version received 02.Feb.2026; accepted 04.Feb.2026; published 05.Mar.2026.

Please cite as:
Kelly A
Extrinsic Trust as a Contractual Framework for Accountable AI in Health Care: Viewpoint
J Med Internet Res 2026;28:e83903
URL: https://www.jmir.org/2026/1/e83903
doi:10.2196/83903

© Anthony Kelly. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Twins for Just-in-Time Adaptive Interventions (JITAIs): Framework for Optimizing and Continually Improving JITAIs

Asim H Gazi[1], PhD; Daiqi Gao[1], PhD; Susobhan Ghosh[1], BS; Ziping Xu[1], PhD; Anna L Trella[1], BS; Predrag Klasnja[2], PhD; Susan A Murphy[1], PhD

[1]School of Engineering and Applied Sciences, Harvard University, 29 Oxford Street, Cambridge, MA, United States

[2]School of Information, University of Michigan–Ann Arbor, Ann Arbor, MI, United States

Corresponding Author:
Asim H Gazi, PhD
School of Engineering and Applied Sciences, Harvard University, 29 Oxford Street, Cambridge, MA, United States

Abstract

In the context of digital health, just-in-time adaptive interventions (JITAIs) are nascent precision medicine systems that can extend personalized health care support to everyday life. A challenge in designing JITAIs is that personalized support often involves sophisticated decision-making algorithms. These decision-making algorithms can require numerous nontrivial design decisions that must be made between successive JITAI deployments (eg, hyperparameter selection for an artificial intelligence algorithm). Making design decisions between deployments—rather than during deployment—ensures intervention fidelity and enhances the ability to replicate results. Yet, each deployment can be costly, precluding the use of A/B testing for every design decision. How should design decisions be made strategically between JITAI deployments? This paper introduces "digital twins for just-in-time adaptive interventions (JITAI-Twins)" to address this question. JITAI-Twins are "digital twins of a subpopulation" (term used in the 2023 National Academies workshop proceedings on digital twins). JITAI-Twins are used to virtually simulate the potential outcomes of a JITAI's design decisions for an upcoming deployment. Based on simulation results, design decisions are made for the deployed JITAI. To continually improve the JITAI, data collected during deployment are used to update the JITAI-Twin—and this bidirectional feedback between deployments and simulation environments continues. JITAI-Twins are thus "fit-for-purpose" (term used in the National Academies 2024 consensus report on digital twins) instantiations of the digital twin concept. In this paper, we elucidate the specifics and design process of JITAI-Twins, with examples of prior use in clinical settings. JITAI-Twins highlight continuity over the course of a JITAI's optimization and continual improvement, emphasizing the need for bidirectional feedback between versions of a simulation environment and a JITAI's deployments.

(J Med Internet Res 2026;28:e72830) doi:10.2196/72830

KEYWORDS

digital twin; just-in-time adaptive intervention; mobile health; digital health; reinforcement learning; artificial intelligence; machine learning; precision medicine; simulation; deployment; JITAI; mHealth; AI; RL

Introduction

Motivation

Just-in-time adaptive interventions (JITAIs) are promising intervention systems that have the potential to impact mobile health (mHealth) [1,2]. By leveraging wireless technologies such as wearable sensors to infer when an individual is at risk for negative change and/or receptive to positive change [3], JITAIs can proactively provide support in everyday life that is adapted to an individual's health state or environmental context via "decision rules" [4]. Decision rules specify how intervention delivery should be adapted depending on "tailoring variables" such as the weather outside or an individual's heart rate.

While adaptive intervention delivery can improve outcomes [5], it is nontrivial to design effective decision rules, and their design has increasingly relied on sophisticated artificial intelligence (AI) and control engineering methods. These methods optimize decision rules to better match the target individual or sample of individuals, rather than individuals from whom previous data were collected [6,7]. To use AI and control engineering methods, nontrivial design decisions must often be made, including which tailoring variables to use, which AI or control engineering method to use, or what hyperparameters should be selected for an AI algorithm [8].

Two competing forces make it challenging to make such design decisions for a JITAI. First, deployments can be costly and are critical to a JITAI's continued development, especially in clinical settings [9,10]. It is imperative that a JITAI performs as optimally as possible when deployed. Second, JITAIs must remain autonomous during each deployment [7]. It is unscalable to require manual modifications to a JITAI's decision-making algorithms during deployment, and unplanned modifications can compromise intervention fidelity and replicability [11,12].

XSL•FO
RenderX

The desire for a JITAI to perform optimally must be balanced with the need for a JITAI to operate autonomously—motivating the development of methods to evaluate and make informed design decisions before a JITAI is deployed.

Digital Twins for Just-in-Time Adaptive Interventions

In this paper, we introduce the framework "digital twin for just-in-time adaptive intervention (JITAI-Twin)" and elucidate how the JITAI-Twin framework can be used to strategically make design decisions prior to a JITAI's deployments (Figure 1). As discussed in a recent Consensus Report from the National Academies [13], digital twins are characterized by two components: (1) the "modeling and simulation to create a virtual representation of a physical counterpart" and (2) a "bidirectional interaction between the virtual and the physical." This bidirectional interaction refers to the use of the virtual twin to inform decisions made for the physical twin, and the use of data from the physical twin to inform the virtual twin. JITAI-Twins include both virtual representation and bidirectional feedback components. JITAI-Twins take data from a JITAI's target subpopulation (physical twin) and construct a set of virtual approximations to the subpopulation's behavior (virtual twin). These virtual approximations are computational models of the subpopulation's behavior when the JITAI is deployed (also known as "simulation environments"). Computational models are fit using prior deployment data and then used to simulate the subpopulation's behavior. Simulations help evaluate multiple sets of potential design decisions (ie, candidate decision-making algorithms), to evaluate which candidate decision-making algorithms are likely to perform optimally. Based on simulation results, the JITAI design team can decide which decision-making algorithm to deploy on the JITAI for a sample of the subpopulation. This deployment produces additional data from the JITAI's target subpopulation, and these data are again used to update and fit the computational models to simulate upcoming deployments and evaluate design decisions. This bidirectional feedback cycle continues for the lifespan of the JITAI-Twin.

Figure 1. Digital twin for just-in-time adaptive intervention framework. Just-in-time adaptive interventions are designed for a subpopulation of individuals—not just a single individual, as highlighted in green to emphasize a key difference from most digital twins in health. This subpopulation serves as the physical twin. When a just-in-time adaptive intervention is deployed, data are collected from (a sample of) the physical twin. These data are used to construct simulation models that serve as the virtual twin. These models are used to simulate the performance of candidate decision-making algorithms for an upcoming deployment. The best-performing algorithm is then deployed within the just-in-time adaptive intervention. Data collected during deployment are then used to update the simulation models—and the modeling, simulation, decision-making, and data collection cycle continues—with bidirectional feedback taking place between just-in-time adaptive intervention deployments, as highlighted in green. Note the existence of data-deployment mismatches highlighted in pink, where data-deployment mismatches could involve differences in the prior data's sample from the upcoming deployment's sample or an inclusion of new variables. JITAI-Twin: digital twin for just-in-time adaptive intervention.

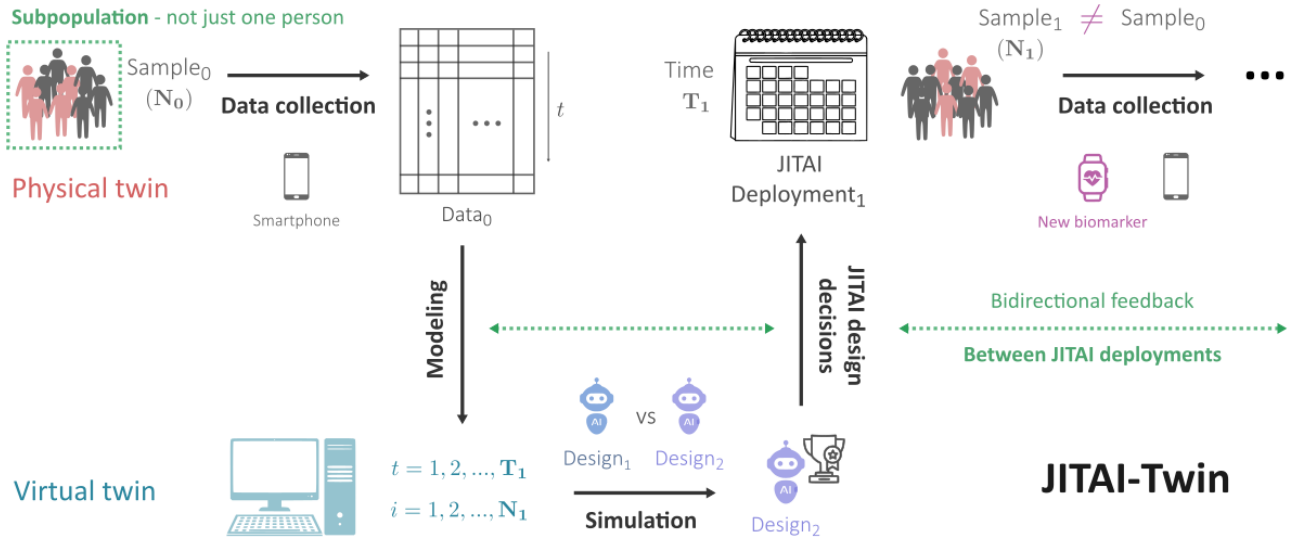

JITAI-Twins are "fit for [the] purpose" of modeling a deployment subpopulation's behavior over the course of a JITAI's continual improvement [13-15], and this presents unique challenges that differ from challenges discussed in previous literature on digital twins for health [16-23] (refer to Section S1 of Multimedia Appendix 1 for a comprehensive overview of related work). A specific deployment's sample rarely shares participants with a previous or follow-up deployment's samples (Figure 1). Thus, models of the subpopulation are not fit or updated to replicate the behavior of specific individuals. Instead, the models should approximate the behavior and heterogeneity of the entire subpopulation, while only sampling from that subpopulation. In contrast, most digital twin frameworks in health focus on modeling a single individual to make model-informed decisions (eg, model predictive control for an artificial pancreas' insulin delivery), where the person-specific model is updated with data specifically from the target individual [20]. Exceptions to this include "digital twins of a subpopulation" for rare disease treatment, discussed in the context of in silico clinical trials to better understand the potency of a drug [24]. Although only briefly mentioned in workshop proceedings of the National Academies without specific guidance on design or implementation examples [24], digital twins of a subpopulation for rare disease treatment are the most relevant example of digital twins. In fact, JITAI-Twins are another instantiation of the digital twin of a subpopulation concept. Other challenges that commonly arise for JITAI-Twins include "data-deployment mismatches." Data-deployment

XSL•FO
RenderX

mismatches are mismatches between the data available to construct the virtual twin and the upcoming real-world deployment characteristics that need to be simulated. Bidirectional feedback for JITAI-Twins takes place at the timescale of deployments, rather than any fixed unit of time (Figure 1). Due to the varying time lengths of deployments and the months or years that could pass between subsequent JITAI deployments, mismatches between previously collected data and future deployments can arise as society, science, and digital health technology evolve.

This Paper's Contributions

Compared to existing literature, the JITAI-Twin framework is the first instantiation of the digital twin of a subpopulation concept with associated guidance on design and implementation examples. Digital twins of a subpopulation are important to making informed design decisions and maintaining intervention fidelity when deploying a JITAI for a group of individuals. Relative to the literature on simulation testbed design for reinforcement learning (RL) algorithms for JITAIs [7], the JITAI-Twin framework highlights the need for bidirectional feedback. Feedback from virtual to physical involves design decisions that are informed via JITAI-Twin simulations. Feedback from physical to virtual helps improve a JITAI-Twin and ensure it matches the ever-changing deployment subpopulation for a JITAI. JITAIs are not designed and deployed as one-time efforts. JITAIs require continual improvement (eg, redesign and deployment) as scientific progress is made, technology evolves, and human behavior changes. Similarly, modeling, simulation, decision-making, deployment, and data collection should be viewed as an ongoing cycle (Figure 1).

This paper is organized as follows. We first provide a motivating example for JITAI-Twins based on a JITAI in the physical activity space that has required continual improvement over multiple deployments [5,25,26]. We then provide guidance on JITAI-Twin design and elucidate the step-by-step process of designing and updating JITAI-Twins. We also enumerate potential solutions to data-deployment mismatches that often complicate a JITAI-Twin's design and updates. Penultimately, we discuss the iterative evaluation process of JITAI-Twins under the digital twin framework of verification, validation, and uncertainty quantification (VVUQ). Finally, we conclude with guidance on selecting a JITAI's decision-making algorithm based on JITAI-Twin simulation results and discussion. We thereby make the following contributions:

1. Formalize the JITAI-Twin framework to instantiate digital twins of a subpopulation for the optimization and continual improvement of JITAIs.
2. Provide guiding principles for the design and updates of a JITAI-Twin.
3. Elucidate the design and iterative evaluation of JITAI-Twins, with examples of imputation and sensitivity analyses to address data-deployment mismatches.

Motivating Example of a JITAI's Continual Improvement: HeartSteps

HeartSteps is a JITAI that has been designed, optimized, and continually improved over the past decade to promote physical activity for sedentary adults [5,27]. In this section, we outline the development journey of HeartSteps, highlighting specific aspects of HeartSteps' development that benefited from the use of a JITAI-Twin. We also highlight the challenges faced in designing the JITAI-Twin due to real-world considerations with HeartSteps.

The initial version of HeartSteps ("HeartSteps V1") was aimed at increasing physical activity levels for sedentary adults by encouraging bouts of physical activity during daily life [5]. HeartSteps V1 consisted of a Jawbone Up wrist-worn activity tracker and a smartphone app that provided self-monitoring and "push" intervention components. The push interventions included 2 types of contextually tailored activity suggestions—intended to help individuals disrupt their sedentary behavior and take short (~10 minutes) walks throughout the day, respectively—and planning prompts, intended to help participants plan a longer activity bout for the following day.

To quantify the effects of the activity suggestions on subsequent physical activity and optimize future development of a JITAI to intelligently deliver activity suggestions targeting walking bouts (ie, walking suggestions), a 6-week microrandomized trial (MRT) was conducted. MRTs are designed to quantify how potential intervention components affect near-term outcomes of interest (eg, step counts), how these effects are moderated by context (eg, weather outside), and how these effects change over time [14]. The data from this initial MRT were then used to optimize the provision of both types of activity suggestions in the revised version of HeartSteps ("HeartSteps V2").

To optimize the provision of walking suggestions for HeartSteps V2, the team developed an RL algorithm that personalized the provision of walking suggestions. This personalization was aimed at maximizing the impact of walking suggestions on the intended proximal outcome: step count in the subsequent 30 minutes [26]. To develop an appropriate RL algorithm and select its associated hyperparameters, the team designed simulation environments that we frame here as the first iteration of a JITAI-Twin for HeartSteps. The JITAI-Twin helped evaluate candidate RL algorithms and strategically select the appropriate RL algorithm to deploy [26].

A challenge in designing the JITAI-Twin for HeartSteps was the existence of data-deployment mismatches. First, HeartSteps V2 focused on a different population—individuals with type 1 hypertension living in a different geographical location (Seattle for V2 vs Michigan for V1)—and involved changes to the data collection methodology compared to HeartSteps V2. Second, the Jawbone sensor was no longer available, so HeartSteps V2 used Fitbit Versa smartwatches to track and collect physical activity data. Third, HeartSteps V2 included daily and weekly collection of ecological momentary assessment data that were different from the data collected when HeartSteps V1 was

XSL•FO
RenderX

deployed, in part due to the team's growing understanding of the psychosocial variables that mediate and moderate physical activity patterns. For example, an app engagement variable was included in HeartSteps V2 to help the RL algorithm personalize the provision of walking suggestions, but HeartSteps V1 did not include measurements of app engagement. After addressing these data-deployment mismatches (discussed in the section "Designing and Updating JITAI-Twins"), the JITAI-Twin simulation results informed the RL algorithm that was then deployed and evaluated for HeartSteps V2 [26].

To address limitations of the prior RL algorithm and incorporate the latest findings from the science of behavior change literature, the team is now working on an improved RL algorithm for deployment (this latest iteration of the JITAI is termed "HeartSteps V3" in this paper) [25,28]. The RL algorithm for HeartSteps V2 is optimized for increased step counts [26], but recent evidence demonstrates that long-term commitment to behavior change, such as the maintenance of physical activity, is more closely determined by affective attitudes toward the desired behavior, rather than short-term bouts of the desired behavior [29-31]. To incorporate this insight into HeartSteps V3, the upcoming deployment will use affective attitudes toward physical activity as outcome variables, and the walking suggestion component was redesigned to pair suggestions for walking with affectively positive stimuli to both increase receptivity to the suggestions and impact affective attitudes. The JITAI-Twin for HeartSteps is being updated accordingly based on the data collected during the deployment of HeartSteps V2. The JITAI-Twin is again being used to evaluate and select between candidate RL algorithms for HeartSteps V3 [25], and is again faced with data-deployment mismatches, as we describe in the section "Designing and Updating JITAI-Twins."

Guiding Principles to Follow for JITAI-Twin Design

When designing and updating a JITAI-Twin, several modeling and simulation decisions must be made. To inform these decisions, we propose the following guiding principles (Table 1) based on lessons learned from prior JITAI-Twin design and related work [7,26,32,33].

Table . Guiding principles for JITAI-Twin[a] design.

Guiding principles	Brief description
Documentation	Document the purpose and plans for the design of a JITAI-Twin, adjustments made during design and updates, and simulation outcomes prior to each deployment
Team science	Include domain scientist(s), data scientist(s), and a deployment lead to ensure the necessary interdisciplinary expertise exists on the JITAI-Twin design team
Model generalizability	Use data modeling strategies that maximize generalizability from prior data to upcoming deployments
The devil is in the details	Judiciously investigate each aspect of a JITAI-Twin's design, especially aspects that are critical to simulation performance
Refrain from data removal	To the extent possible, refrain from removing data that is representative of real-world deployments (eg, disengaged participants) when designing and updating the virtual twin

[a]JITAI-Twin: digital twin for just-in-time adaptive intervention.

Documentation

A JITAI-Twin should be fit for purpose [13], and this purpose should be clearly documented. The design and update plans should be detailed, assumptions made in modeling the physical twin should be documented, and the objectives for each of the simulations conducted should be prespecified. During the JITAI-Twin design, adjustments to the plan will be necessary—these design decisions should be documented for computational reproducibility and continuity [34]. In general, a well-designed and documented JITAI-Twin has the potential to outlive its initial designers and provide a quality return on investment. Since JITAI-Twin simulations are conducted offline between JITAI deployments, computational cost is less likely to be a limiting factor compared to the time and effort in designing and maintaining a JITAI-Twin (ie, labor costs). Thus, like any computational resource, poor documentation can lead to wasted time and effort during updates to a JITAI-Twin, which can incur significantly more costs compared to spending time initially and at each update step on documentation.

Team Science

Just as an interdisciplinary team is recommended when designing and deploying a JITAI, an interdisciplinary team is recommended when designing and leveraging a JITAI-Twin for decision-making [35]. The JITAI-Twin design team should, at a minimum, consist of the deployment lead, a domain scientist, and a computational scientist. The deployment lead is the researcher in charge of executing the JITAI deployment. The deployment lead's insight is necessary to faithfully model practical aspects of a deployment and the target subpopulation (eg, expected disengagement from the JITAI). The domain scientist is the expert in physiology, psychology, etc, relevant to the health application (eg, suicidologist for a suicide prevention study). This domain scientist's insight is necessary to design simulation models that reasonably approximate psychophysical changes and outcomes. Finally, the data scientist

XSL•FO
RenderX

(eg, computer scientist, engineer, or statistician) is needed to mathematically abstract psychophysical characteristics into virtual constructs (eg, simulation models). Ideally, this data scientist has interdisciplinary training to improve cross-disciplinary communication and abstraction.

Model Generalizability

"All models are wrong," but "useful" models have explanatory power, capturing relationships that generalize from past data to future deployments [36]. We highlight 2 interpretations of explanatory power relevant to JITAI-Twin design [37,38]. First, explanatory power reflects a model's ability to quantitatively capture known qualitative relationships, a characteristic often attributed to mechanistic models. This is especially important in the context of counterfactual outcomes, where a simulation model will need to generate plausible outcomes for alternative intervention options that were not selected in prior data. A recommended approach is to assess the limitations of prior data and ensure aspects of the model informed by domain knowledge account for blind spots in the data. For example, if prior data do not account for a portion of the JITAI's target subpopulation, construct variants of the computational models that account for potential differences in behavior informed by domain knowledge [39]. Second, explanatory power refers to a model's ability to generate data that align with real-world observations, a strength of data-driven machine learning models. JITAI-Twin models should balance the strengths of mechanistic and data-driven modeling to enhance generalizability [40]. Generalizability refers to how well computational models can explain or reconstruct data that were not used to train the models and has been extensively discussed in the literature [41,42]. A wealth of literature exists on the challenges of model selection for generalizability and approaches to consider [43,44]. In general, approaches to enhance generalizability ensure that a model's complexity is commensurate with data availability. This is related to the bias-variance trade-off that has been shown empirically to be of value to model generalizability [45]. The bias-variance trade-off generally indicates that generalization error will follow a U-like curve, where large errors with too few parameters indicate too much bias and too little variance (ie, too little flexibility to capture the underlying patterns), and large errors with too many parameters indicate too little bias and too much variance (ie, too much flexibility causing the model to overfit). Recent work on the surprising performance of overparameterized machine learning models adds a new layer to this guidance, showing that the "double descent phenomenon" can be observed with overparameterized models (eg, deep neural networks) [46]. The complexity of model selection makes team science even more critical, as data scientists may have helpful intuition from experience that contradicts general guidance for a specific application setting.

Note that a JITAI-Twin's simulation models should not only capture the underlying relationships of interest in the data (eg, relationships between intervention options and outcomes of interest), but also the unstructured variability within the data (ie, the "noise" characteristics). When candidate decision-making algorithms are evaluated, simulations will need to be conducted with stochastically generated noise to evaluate how well decision-making algorithms perform in settings that are not accounted for by the learned aspects of the models. Generalizable models are critical because a JITAI-Twin's simulations inform real-world decisions for settings that will likely differ from prior data.

The Devil Is in the Details

"Inattention to detail is the hallmark of mediocrity" for JITAI-Twin design [47]. JITAIs and real-world deployments are sophisticated, and because JITAI-Twins are used to make design decisions for real-world JITAI deployments, details cannot be overlooked or the risk of failure can become significant (eg, the Mars Orbiter that crashed partly due to a mismatch between metric and imperial units [48]). JITAI-Twin design must match the level of sophistication by attending to details such as the assumptions and limitations associated with a modeling approach (eg, assuming linearity vs modeling nonlinearities that may require more data). Of course, "perfect is the enemy of good," so a JITAI-Twin design team must judiciously worry about aspects of a JITAI-Twin's design that are critical to modeling the physical twin accurately (ie, "worrying selectively") [36]. Team science helps guide the team in worrying selectively. Protocol details that are critical to model accurately can be highlighted by deployment leads. Domain scientists can prioritize which psychophysical relationships are most important to model. Data scientists can clearly articulate the limitations and the assumptions of modeling approaches they use. Due to the level of detail required in designing a JITAI-Twin, its design should be viewed as a research project itself. Adequate time must be budgeted for collaborative prioritization, modeling, simulation, evaluation, and iteration. Iteration will help data scientists learn what aspects of the target subpopulation and deployment are most important to model accurately for informative simulations. Iteration will also help deployment leads and domain scientists learn what correspondence is reasonable between the physical twin and virtual twin and maintain that standard.

Refrain From Data Removal

Data removal should be avoided to the extent possible when designing a JITAI-Twin. JITAI-Twins approximate the true behavior of a target subpopulation using a JITAI. A fundamental truth of JITAI deployments is the harsh reality of user disengagement, missing data, etc. When disengaged users or users with insufficient data are removed or ignored, the JITAI-Twin becomes biased toward the subsample of prior participants or users that remained engaged [49,50]. However, future JITAI deployments will not only include engaged users, but will also include users who disengage. For a JITAI that seeks to personalize intervention delivery, disengaged users may in fact be the most important to consider [32]. Removing incomplete data would overlook users who may benefit the most from a well-designed decision-making algorithm. That said, if data are incorrect or are unrepresentative of the physical twin due to measurement or recording discrepancies, these data may need to be removed. Hence, the suggestion is not to avoid data removal entirely. Rather, the prudent JITAI-Twin designer should judiciously refrain from removing data representative of upcoming deployments.

XSL•FO
RenderX

Designing and Updating JITAI-Twins

Designing and updating a JITAI-Twin involves 5 overarching steps, as visually summarized in Figure 2:

Figure 2. Overarching steps to designing (and updating) digital twins for just-in-time adaptive intervention. The 5 steps are illustrated from left to right and detailed in the main text. Verification and validation are iterative and continue until the base simulation testbed meets the desired performance. The construction of testbed variants is a form of uncertainty quantification that, along with other uncertainty quantification strategies, ensures just-in-time adaptive intervention design decisions are made in an uncertainty-informed manner. JITAI: just-in-time adaptive intervention.

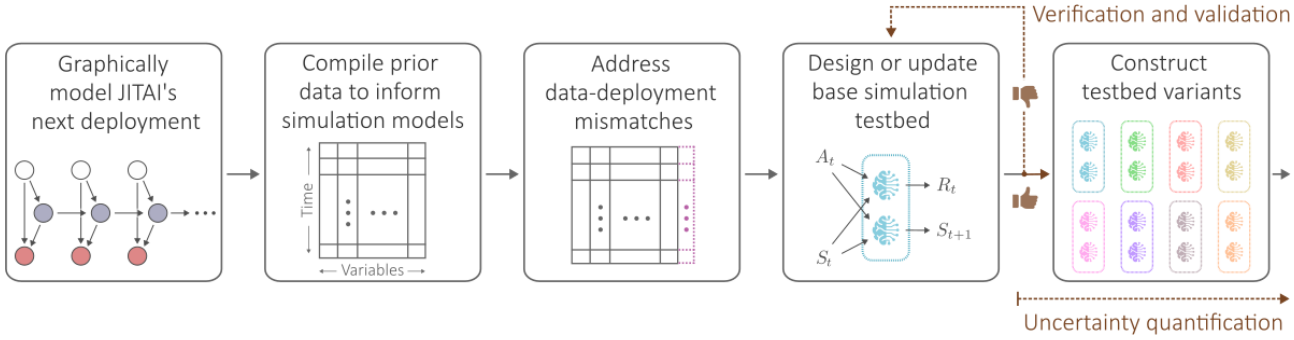

1. The first step is to graphically model the upcoming deployment's characteristics (eg, how often intervention can be delivered), intervention targets (eg, intervention will try to enhance positive affect toward physical activity), and target subpopulation (eg, sedentary adults). A recommended approach is to construct a causal directed acyclic graph (DAG) for the upcoming deployment to qualitatively represent expected user behavior in response to interventions, moderated by context variables [51,52].
2. The second step is to compile prior data that mirror the data to be collected for the upcoming deployment to the greatest extent possible. These data might be compiled from a previously conducted MRT, as was the case for HeartSteps V1 data [5,26]. Note that data from an MRT are not required, as will be discussed in the context of data-deployment mismatches.
3. The third step is to address data-deployment mismatches. For example, if the upcoming deployment involves an intervention component that was not present in prior deployments, the effect size for the intervention component must be imputed. The imputed effect size captures what domain scientists anticipate regarding the strength of the intervention component's effects on outcomes, relative to non–intervention-related changes in outcomes (ie, the "signal-to-noise" ratio). The imputed effect size enables simulations of hypothetical responses to delivery of the intervention [53].
4. The fourth step is to construct a base "simulation testbed" informed by the graphical model, using the data compiled and imputation decisions made. A simulation testbed is defined as a collection of computational models used to simulate the performance of a JITAI's decision-making algorithms. The base simulation testbed will include a set of simulation models (eg, outcome and state transition models) and parameters to vary. Some parameters will correspond to values imputed or specified to address data-deployment mismatches. Since JITAI-Twins are digital twins of a subpopulation and not of a single individual, the simulation models must capture heterogeneity between individuals in the subpopulation. The base testbed will need to be iteratively evaluated and redesigned until satisfactory evaluation results, as illustrated in Figure 2 and discussed in the next section, "Iterative Evaluation: VVUQ."
5. The fifth step is to create variants of the base simulation testbed. These variants are used to evaluate how robust candidate algorithms are to potentially incorrect modeling assumptions. For example, if an intervention's effect size had to be imputed (in step 3) or the intervention's effects may differ from prior data, one can form testbed variants by varying the intervention's effect size to understand how candidate algorithms perform across potential scenarios. This is a form of uncertainty quantification (Figure 2), as discussed in the next section, "Iterative Evaluation: VVUQ." The variants and the base simulation testbed together form the virtual twin of the JITAI-Twin.

To make explanations more concrete, the remainder of this paper will focus on JITAI-Twins that simulate the performance of candidate RL algorithms (eg, the HeartSteps JITAI-Twin). The steps we describe remain applicable to simulating the performance of other decision-making algorithms for JITAIs, such as control engineering methods, decision trees, expert systems, etc [54]. RL algorithms are AI algorithms that iteratively optimize the selection of actions based on previously observed rewards and states [55]. In the context of JITAIs, RL algorithms are decision-making algorithms that can select between intervention options (actions) to optimize proximal outcomes (rewards) based on an individual's tailoring variables (states). Example actions could include whether a physical activity suggestion is delivered or which suggestion is delivered. An example reward variable could include an individual's step count as a surrogate for moderate-to-vigorous physical activity. Example state variables could include an individual's environmental context, such as the weather or information about the individual, such as their current location. State-of-the-art JITAIs use RL algorithms to personalize intervention delivery, progressively determining which intervention option is likely to optimize outcomes based on the inferred state of an individual [26]. For a primer on RL and simulation modeling, refer to Section S2 of Multimedia Appendix 1.

XSL•FO
RenderX

Graphically Model the Subpopulation and JITAI Deployment Characteristics

A graphical model helps ground subsequent JITAI-Twin design and update steps. With a graphical model, the research team can identify variables of interest and compile data accordingly for training models. The team can also identify which variables or relationships are important for which no data constitutes data impoverishment. When designing the base simulation testbed, the computational models used can then be structured according to the relationships specified in the graphical model. The graphical model can also help inform which variants to construct to investigate uncertainties in the graphical model.

For HeartSteps V3, the deployment characteristics and subpopulation behavior are currently being modeled by the causal DAG shown in Figure 3, adapted from Gao et al [25]. Note that this DAG is still under refinement but has undergone multiple iterations to collaboratively develop. A DAG visually describes causal relationships between the specified variables using directed arrows that are acyclic [51]. Importantly, arrows between variables that are omitted from the DAG reflect effects that are believed to be weak or nonexistent. Constructing a graphical model of this form requires the team to first identify the variables of importance to the upcoming trial. At a minimum, this consists of context variables (C in Figure 3), intervention options (A in Figure 3), and outcome variables (R and E in Figure 3). Identifying these variables translates to specifying states, actions, and rewards for the RL algorithm simulation testbed. The conceptual model should also specify mediators (M and N in Figure 3) that may affect the relationships between context, interventions, and outcomes. Since JITAI deployments involve decisions that are made over multiple time points (ie, sequential decision-making), it is important that the conceptual model captures time dependencies. For example, Figure 3 also illustrates dependencies across consecutive weeks and across consecutive days within a week. Assumed independence in time is also captured with the absence of arrows. Although independence between variables may not perfectly mirror the physical twin, it is important to selectively worry about the relationships most consequential to the upcoming deployment. A recommended approach to do this is to rank order the relationships between variables and decide on a cutoff point below which all relationships will be assumed inconsequential. Variants of the simulation testbed can be used to test the robustness of candidate decision-making algorithms to relationships that fall slightly below the cutoff point [25], as discussed in the later section “Construct Variants of the Base Simulation Testbed.”

Note that the variables included in the graphical model map to the subpopulation and deployment characteristics and not to the JITAI directly. This distinction is critical between JITAI-Twin models and a JITAI’s models used for decision-making. The JITAI may not measure or have access to all the variables that are important for simulations. The variables important for simulations are the variables needed to satisfactorily approximate how the target subpopulation will respond to the JITAI for the upcoming deployment. The graphical model needs to encompass all variables that are crucial for the JITAI-Twin’s simulation models.

XSL•FO
RenderX

Figure 3. Example causal directed acyclic graph used for the HeartSteps digital twin for just-in-time adaptive intervention. The directed acyclic graph shown is being iteratively refined to graphically model HeartSteps V3. Each arrow indicates a causal relationship from the variable at the start of the arrow to the variable at the end of the arrow. The subscript w represents week (ie, w-1 refers to previous week), so the subscript w,1 refers to the first decision time within the week, while subscript w,K refers to the Kth decision time within the week. The context, C, represents an individual's recent step count, GPS location, etc The proximal outcome, M, is the step count after an intervention is or is not delivered. Intervention, A, represents the decision to send a physical activity suggestion or not. Proximal outcome, N, represents whether a user clicks on an intervention message or opens the app. Intermediate outcome, E, represents a higher-level "app engagement" variable. The outcome variable, R, represents affective attitude toward physical activity.

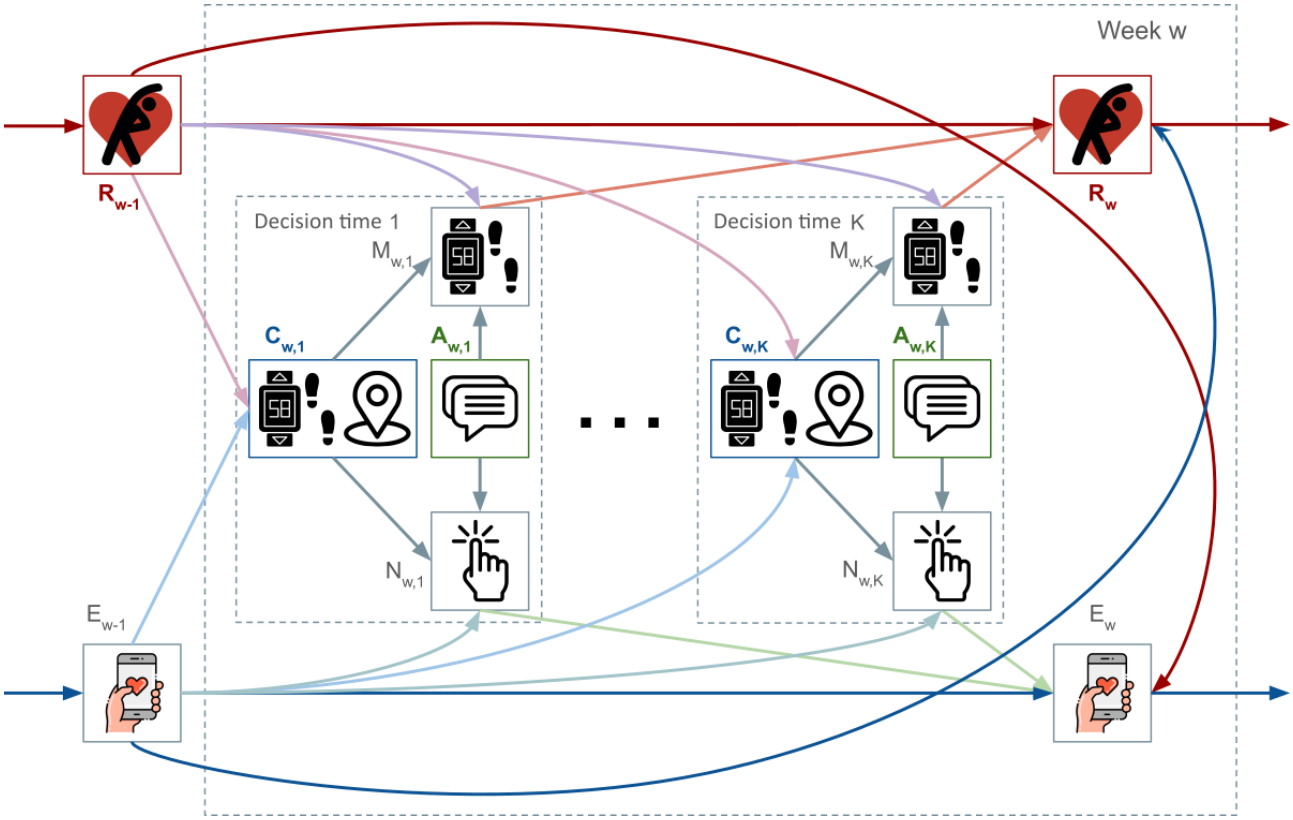

Compile Prior Data to Inform Simulation Models

Prior data are necessary to train and otherwise inform the computational models used for a JITAI-Twin. If prior deployment data are not available (ie, the JITAI-Twin is being designed for the first time, rather than being updated), the study team should identify prior data most relevant to the upcoming deployment's simulations (similar variables, timescales of measurement, etc). The graphical model discussed in the previous section ("Graphically Model the Subpopulation and JITAI Deployment Characteristics") can help guide data compilation. Relevant prior data should help inform relationships between variables in the graphical model.

For HeartSteps, the initial JITAI-Twin was designed using data from the HeartSteps MRT (ie, HeartSteps V1). MRT data are ideal for JITAI-Twin creation because MRT data contain dynamic measurements of outcomes and context relevant to the intervention components being investigated—and the provision of interventions is microrandomized. The microrandomization of intervention provision provides data for use in quantifying the relationships between intervention options and outcomes, and how the relationships are moderated by context variables. The outcomes and context variables map to rewards and state variables that a simulation testbed can generate in response to an RL algorithm's simulated actions (ie, intervention provision).

When compiling prior data, it is important to avoid data leakage [56]. The study team should structure prior data in a way that mirrors how data will be used to inform algorithms for deployment. For example, some RL algorithms, such as Thompson Sampling algorithms used by HeartSteps, enable the use of informative Bayesian priors or other hyperparameters [57]. If the plan is to use a single dataset for both simulation and to inform the Bayesian priors, then the data should be split into disjoint datasets for modeling and Bayesian prior construction or hyperparameter selection, respectively [58].

Address Data-Deployment Mismatches

The data compiled (see previous section, "Compile Prior Data to Inform Simulation Models") are likely missing variables of interest from the graphical model (see previous section, "Graphically Model the Subpopulation and JITAI Deployment Characteristics") or are deficient in some way when it comes to informing relationships of interest for the simulation models. We term these deficiencies "data-deployment mismatches." Data-deployment mismatches are a reality of JITAI-Twin design and updates. During the continual improvement process, research teams often want to incorporate the latest innovations in sensing and intervention technology [59]. The science of behavior change, psychophysiology, and other relevant domain sciences is also progressing rapidly [28,60-62]. This progress and rapidly evolving technology often add additional variables

XSL•FO
RenderX

or dependencies to consider for the upcoming deployment that are absent from previous deployment data. This absence of variables in prior data necessary to simulate upcoming deployment decisions is termed "data impoverishment." Other forms of data-deployment mismatches include domain shift and sampling bias. Deployments may take place years apart, where characteristics of the subpopulation or their behavior may change. The time of the year or the specific sample investigated for each deployment may differ from previous deployments.

To characterize the data impoverishment faced for JITAI-Twin's design or updates, identify the missing variables and/or relationships in the graphical model that cannot be fit using the data compiled. This data impoverishment can be overcome through judicious imputation decisions and the variation of unknown model parameters. These variations are a form of sensitivity analysis, where the uncertainty in the model parameters is propagated through to the simulation results (eg, if parameter x falls below value a, then decision-making algorithm 1 performs better than algorithm 2). Team science is critical to plausibly overcoming data impoverishment. Domain scientists and trial leads help ensure imputation and sensitivity analyses are conducted realistically.

Identifying domain shift and sampling bias requires careful comparison of the prior data, the overall subpopulation, and the upcoming deployment's characteristics. For many deployments, it is not possible to fully anticipate the upcoming sample's characteristics. However, it is important to recognize any oversampling or undersampling of a particular participant group in prior data. Sampling bias may also occur with respect to time. Prior data collected in the winter for a physical activity study will differ from what is expected if deploying a physical activity JITAI in the summer. When sampling bias is less obvious to the eye, methods to detect sampling bias include statistical tests to compare sample demographics and other characteristics to mean values or proportions known to be representative of the subpopulation. Domain shifts can still take place even if sampling was done in a representative manner. Over time, the characteristics of the subpopulation can change (eg, expectations of technology can change rapidly, as with the recent proliferation of generative AI tools), and how technology or a JITAI is deployed can also change. The detection of significant domain shifts (ie, "distribution shifts" or "dataset shifts") is an active area of research, with methods that involve combining dimensionality reduction (of features such as baseline survey responses) and statistical testing performing the best in practice [63]. To overcome these data-deployment mismatches, domain knowledge can be used to adjust parameters of the simulation models according to expected differences between the domain or sample present in prior data and the upcoming deployment. Parameters of the simulation model that capture variability in the subpopulation or data collection environment can also be enlarged. Performing simulations with enlarged variances can help account for sources of variability in the upcoming deployment that may not have been observed in prior data due to sampling bias or domain shift.

For HeartSteps V3, outcome variables are impoverished [25]. Progress in the science of behavior change indicates that optimizing for short-term bouts of a desired activity is less indicative of sustained behavior change compared to optimizing affective attitudes toward the desired activity [28], which applies to the maintenance of physical activity [64-67]. Affective attitudes toward physical activity were not collected in prior data. To address this data impoverishment, affective attitudes toward physical activity are now being generated when updating the JITAI-Twin to simulate the upcoming deployment. Specifically, a linear state-space model is fit using system identification to infer latent (ie, not measured) changes in affective attitudes toward physical activity from observed (ie, measured) bouts of unprompted and prompted physical activity [25]. The relationship between affective attitudes toward physical activity and bouts of unprompted physical activity was captured in the DAG (Figure 3). For more examples of how data impoverishment has been addressed in the past for state and action variables, see Section S3 of Multimedia Appendix 1.

Design the Base Simulation Testbed

Designing the base simulation testbed is an iterative process. Computational models are designed to generate (1) outcomes and (2) state transitions each time a candidate decision-making algorithm takes an action. These models must account for heterogeneity between individuals in the subpopulation. The computational models are then iteratively evaluated to ensure plausible outputs are generated. A detailed discussion of iterative JITAI-Twin evaluation is provided in the next section, "Iterative Evaluation: VVUQ." Note that the model fitting and evaluation procedure for simulation models for JITAI-Twins differs from the procedures used for prediction and statistical inference models typically discussed in the machine learning and statistics literature. A simulation model for a JITAI-Twin is used to evaluate candidate decision-making algorithms, so the simulation model should not only be able to capture the underlying relationships between decisions, context variables, and outcomes of interest—but the simulation model should also be able to model noise characteristics properly. Simulations will need to be conducted with stochastically generated noise to evaluate how well candidate decision-making algorithms perform in settings that are not accounted for by the learned aspects of the models.

Outcome Modeling

Outcome generation requires a model that maps the state of the simulation environment and the action an RL algorithm takes to some set of outcomes. These outcomes can then be used to form rewards for an RL algorithm. Note the subtle distinction between outcomes and rewards. Rewards are used by an RL algorithm for learning and can differ from one RL algorithm to the next. Outcomes are generated by the JITAI-Twin regardless of which RL algorithm is used. The RL algorithms may use those outcomes to form their own reward signals.

Outcomes can be modeled probabilistically, $y(s, a)=P(Y_t=y|S_t^w=s, A_t=a)$. The superscript W emphasizes that this is the "world state" (ie, the all-encompassing state of the simulation environment, rather than the limited-in-knowledge state of specific RL algorithms to be evaluated). See the RL primer in the supplement for further explanation (Section S2 of Multimedia Appendix 1). The

XSL•FO
RenderX

outcome model will need to be trained on prior data, after addressing data-deployment mismatches (section "Address Data-Deployment Mismatches"). The model used for $y(s, a)$ can be of several forms [68]. See the work on HeartSteps V2 and HeartSteps V3 for examples [25,26]. The guiding principles presented in this paper and the availability of data should inform the model class and the model fitting process used [69,70].

State Transition Modeling

A state transition model is of the form W(s′, s, a)=P(St+1w=s′|Stw=s, At=a). Again, the superscript W emphasizes that this is the "world state." See Section S2 of Multimedia Appendix 1 for further explanation. This model quantifies the probability of reaching state S′, given that the current state is S and the RL algorithm for the JITAI takes action a. The simulation environment's state transition model dictates how the simulation environment's state changes over time in response to actions taken by the simulated JITAI, regardless of which simulated JITAI is used. Our ongoing work for HeartSteps V3 learns a state transition model to continually improve the JITAI-Twin for HeartSteps [25]. Note that by concatenating the next state, St+1w and the outcome, Y_t, into a single vector, a single probabilistic model conditioned on S_t and A_t can be used to produce a vector of next state and outcome variables. This approach is used by Gao et al [25] and captured by the DAG in Figure 3, where next states are not differentiated from outcomes.

A simpler approach taken for HeartSteps V2 was to view state generation from a model-based data augmentation perspective [26]. Models are used only for the outcome function and to address state impoverishments. The prior data's state sequences are then used as is to generate state transitions. This preserves the dependence across time in state variables but makes the simplifying assumption that the actions of the candidate decision-making algorithms have minimal effects on the sequences of states observed in prior data (eg, states such as weather and location). This decision is based on the principle that digital therapeutics tend to have little immediate effect on autonomous state dynamics relative to residual noise (ie, implicit assumption of a simpler RL algorithm called a "contextual bandit").

Modeling Heterogeneity Between Individuals in a Subpopulation

A JITAI-Twin's outcome and state transition models must account for heterogeneity between individuals [71]. Several approaches exist for modeling heterogeneity between individuals in a subpopulation. One approach that has been used for the HeartSteps JITAI-Twin is the construction of several user-specific models to represent a virtual cohort of users [25,26]. By fitting a user-specific model to each user present in prior data, the variability across model parameters can be used to quantify the heterogeneity across individuals. The user-specific modeling approach is useful when variability lacks a clear distributional structure. However, the downside of the approach is that it assumes prior data are representative of the subpopulation. An alternative approach is to assume that the heterogeneity across individuals in the subpopulation exhibits some structure (eg, normal distribution) or can be stratified in some way (eg, categorizing users by some phenotype). Stratified modeling or mixed-effects modeling (eg, with random effects that are normally distributed) can be used to form a single model or a smaller subset of models to simulate the responses of a virtual cohort of users. An inherent trade-off that exists is that imposing structure may lead to underfitting, but assuming no structure risks overfitting to prior data. More flexible data generation approaches (eg, via artificial neural networks) that use proxy features to account for user identities or characteristics could be used to generate a virtual cohort [72]. The downside of more flexible data generation methods is that they typically require extensive data to produce realistic outputs.

Additional mHealth Considerations

A few additional aspects of mHealth interventions are important to model. Engagement with an intervention tends to decay over time [5,73]. This can translate to reduced effect sizes of the intervention (captured by the outcome model) over the course of a simulated trial, as well as participant dropout. As long as data on participants who dropped out were not removed from the prior trial's data, the prior trial's data can be used to inform the expected dropout rate over the course of a simulated trial. User burden in response to the intervention should also be considered. If an intervention is delivered too often, habituation may reduce the effect size of the intervention over time (captured by the state transition model's dynamics). Engagement may also diminish over time.

Construct Variants of the Base Simulation Testbed

Variants of the JITAI-Twin's base simulation testbed can be used to evaluate the robustness of candidate algorithms to potentially incorrect assumptions made during graphical modeling or decisions made to overcome data-deployment mismatches. The use of testbed variants is critical because the base simulation testbed on its own will fail to replicate characteristics of the upcoming deployment. However, the base testbed along with a well-designed set of variants can be used to cast a broader net over the possible scenarios the JITAI may encounter in the upcoming deployment.

The variants constructed should all satisfy "matching conditions." Matching conditions are defined here as marginal statistics or other outputs of the simulation testbeds that must be met to ensure scientifically plausible simulations. For example, if the study team decides that the degree of between-subject variability observed must be greater than some value, then plausible variants must produce between-subject variability above that value. Calibration can be used to ensure matching conditions are satisfied.

For HeartSteps V3, a simplified DAG is used to inform the RL algorithm (simplified version of the DAG shown in Figure 3; see Gao et al for details) [25]. Certain variants of the simulation testbed intentionally cause mismatches between the DAG used to inform the RL algorithm and the DAG used to generate the JITAI-Twin simulation testbed's rewards. The RL algorithm selected for deployment should be robust to these mismatches. Imputed effect size trajectories were also varied to construct additional variants. These variants are used to test the

XSL•FO
RenderX

performance of candidate RL algorithms for the entire range of plausible effect sizes (a matching condition) that could be observed during the upcoming deployment.

Iterative Evaluation: VVUQ

The evaluation process for JITAI-Twins is iterative and aligns with the digital twin evaluation framework of VVUQ [13,74], as shown in Figure 4. Verification ensures that the virtual constructs (eg, simulation models) created computationally (eg, using floating point approximations) mirror the desired mathematical descriptions. Validation ensures that the digital twin adequately represents the physical twin for the purposes of its design. Uncertainty quantification encompasses sensitivity analyses, the quantification of confidence in simulation results, and the quantification of uncertainty in data collected when designing and updating a JITAI-Twin. Sensitivity analyses determine how simulation outcomes are affected by varying learned parameters and/or modeled variables. Confidence measures are produced by performing numerous stochastic simulations and by using statistical methods to quantify confidence and credible intervals (ie, uncertainty quantification) for simulation outcomes. Data uncertainty can arise due to corrupted data, among other sources of uncertainty. Although difficult, characterizing sources of uncertainty (eg, heterogeneity in the subpopulation or stochasticity in simulation) and quantifying uncertainty accordingly are critical tasks.

Figure 4. Verification, validation, and uncertainty quantification for digital twins for just-in-time adaptive interventions. Verification, validation, and uncertainty quantification is an iterative process. For each iteration of the digital twin for just-in-time adaptive intervention, multiple repetitions of verification and validation are often necessary before satisfactory simulation performance is achieved. Uncertainties must then be quantified in simulations to inform decision-making for a just-in-time adaptive intervention's deployment. Data from a just-in-time adaptive intervention's deployment will contain sources of error that also need to be quantified when updating the digital twin for just-in-time adaptive intervention's computational models. The verification, validation, and uncertainty quantification cycle continues. JITAI: just-in-time adaptive intervention; JITAI-Twin: digital twin for just-in-time adaptive intervention; VVUQ: Verification, validation, and uncertainty quantification.

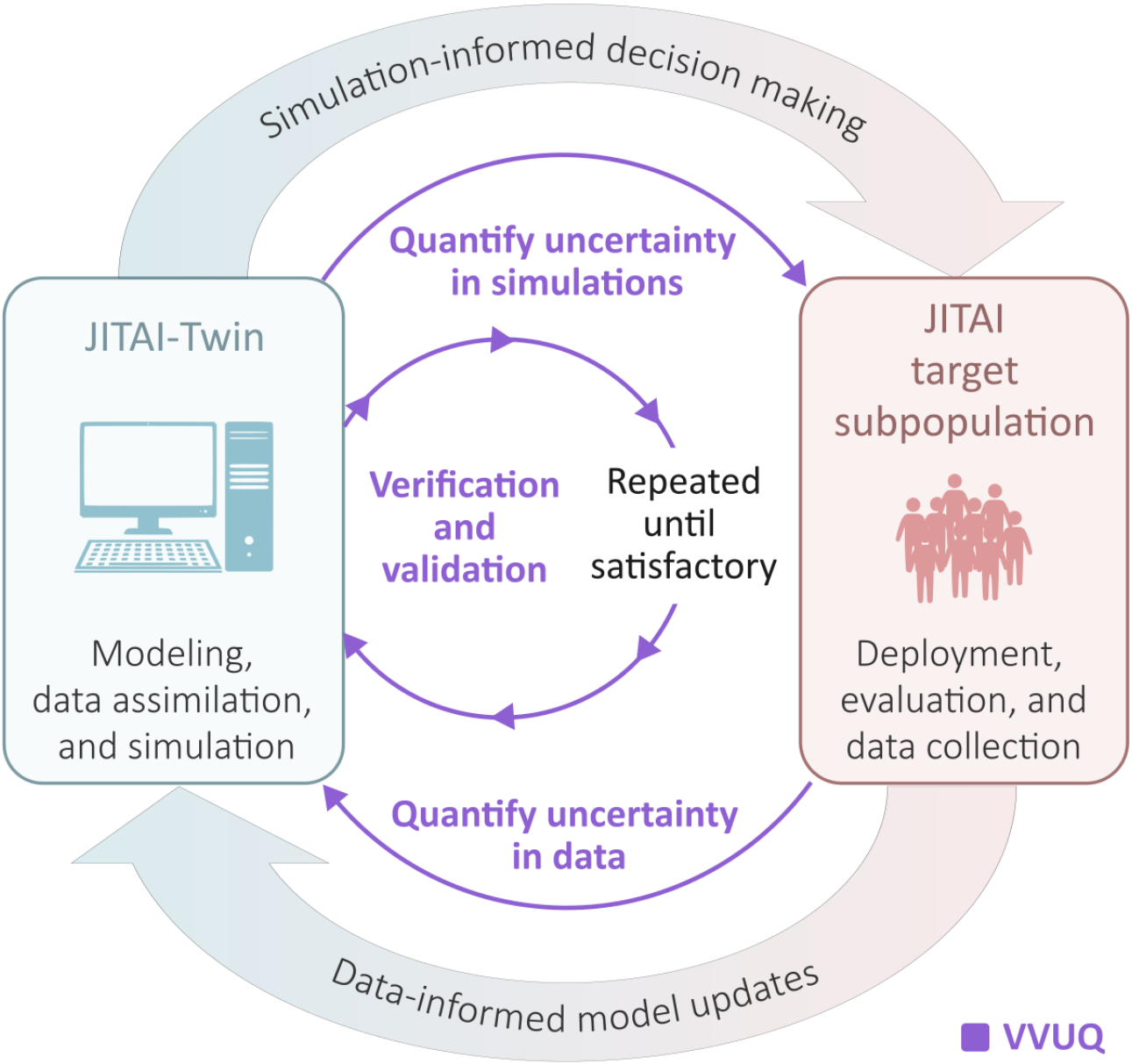

XSL•FO
RenderX

Verification

Verification is necessary for any computational model or piece of code. Standard practices to ensure proper code execution, negligible floating-point approximation errors, and appropriate data generation should all be first-pass evaluation criteria. One verification technique relevant to generative modeling is replication analysis. One can ensure replicable simulations by keeping track of the seeds used for random number generation and rerunning for verification. Other ways to improve verification include publicly sharing the code, procedural information, and data (if possible). Other researchers and practitioners can then participate in verification and improvements to the procedures and code.

Validation

Data from the upcoming deployment are necessary to truly validate a JITAI-Twin's simulations. However, once data are collected for the upcoming deployment, it is too late to update the JITAI-Twin according to validation results. Thus, alternative strategies for predeployment validation are necessary to ascertain JITAI-Twin quality. We separate JITAI-Twin validation strategies into 2 categories: mechanistic and empirical.

Mechanistic

Mechanistic validations are used to determine whether the generated simulation data are reasonable in the context of domain science and deployment experience. Specific examples include simulated engagement with a JITAI or participant dropout. Decaying engagement with an intervention and participant dropout rates should be modeled in a manner that aligns with deployment expectations. Other examples include simulated dynamics for state variables, outcomes, etc A concrete example from HeartSteps is to evaluate whether generated step count outcomes should decrease over time when the simulated virtual participant experiences an increase in positive affect toward physical activity (should generally not be the case in this example).

Empirical

Empirical validation requires detailed, data-driven evaluation of the JITAI-Twin's generative models and their outputs (see the supplementary material of Gao et al for an example from HeartSteps) [25]. Empirical evaluation strategies can be separated into 3 categories. The first category involves goodness-of-fit assessments on prior data used to fit the simulation models. Goodness-of-fit measures describe how well the trained model can generate outputs (in the absence of noise or residual components) similar to the prior data outputs when given prior data inputs (eg, coefficient of determination for regression). Coefficients of the model can also be assessed to ensure that the weights are of the appropriate scale, sign, and magnitude according to domain science.

The second category of empirical validation involves out-of-sample prediction (eg, using held-out data). For models fit to several participants' data simultaneously, a subset of participants' data can be held out to evaluate agreement between simulated and real data when the models are used to predict the held-out outcomes and state transitions given the held-out states and actions. When simulations are conducted without noise, generated data should align with the held-out data, on average (ie, the expected values, in a statistical sense, should align). When simulations are conducted with noise (as is done over multiple iterations when evaluating decision-making algorithms), the out-of-sample data should align with the distribution of data stochastically generated by the models. Importantly, the aim of out-of-sample prediction is not to replicate the held-out data during data generation. The aim of out-of-sample prediction is to instead evaluate how well marginal statistics and other diagnostic metrics agree with held-out data. In cases where a set of participant-specific models are trained to act as virtual participant models, participant-specific data can be held out and used for validation.

The third category of empirical validation diagnoses whether the models generate data that satisfy diagnostic checks. This is related to model calibration [75,76]. These diagnostic checks evaluate whether statistics computed across multiple simulated trials (eg, effect sizes and variability across virtual participants [7,25]) agree with the statistics computed from physical trials. It is recommended to inspect several statistics when validating a JITAI-Twin's generated data.

Uncertainty Quantification

Uncertainty quantification ensures that decisions informed by a JITAI-Twin are made with an understanding of the uncertainty in JITAI-Twin modeling and simulation results [77]. In this context, uncertainty quantification refers to computing measures of variability and alternative outcomes to accompany a JITAI-Twin's simulation results, such as 95% CIs for a candidate decision-making algorithm's simulated performance or simulation performances using a different testbed variant. There are several sources of uncertainty that, to the best of the JITAI-Twin design team's capacity, need to be considered, quantified, and propagated to quantify uncertainty for a JITAI-Twin's simulation results. We outline several important sources of uncertainty for JITAI-Twins in this subsection.

One source of uncertainty is stochasticity in simulations. Stochasticity is critical because simulations often involve several probabilistic outcomes that may differ from one simulation to the next (eg, the outcome of a coin flip). Stochasticity manifests either through probabilistic modeling used for the outcome and state transition models or probabilistic decision-making via a candidate decision-making algorithm. To account for this stochasticity, several seeds should be used for random number generation to investigate the variability in results across multiple simulation runs. CIs can be constructed accordingly, and algorithm evaluation decisions can then be made with a notion of statistical significance [32]. Note that for implementation purposes, the simulations corresponding to each random seed can be parallelized to potentially save time.

An additional source of uncertainty is the uncertainty in trained parameters and imputed variables. Parameters of a simulation model that are trained using prior data may not be accurate in representing upcoming deployment characteristics due to limitations in the quality and quantity of prior data. Variables that are imputed using domain knowledge to address data impoverishment may behave differently from what is observed on average in domain sciences. Sensitivity analyses (eg, via

XSL•FO
RenderX

variants of a simulation testbed) can be a useful way to investigate how simulation results vary as a function of changes in imputed variables or changes in learned parameters [78].

A final source of uncertainty we discuss here is uncertainty in the data compiled from prior deployments and studies, as shown in Figure 4. Accounting for uncertainty in collected data is important because the data compiled will significantly influence what the JITAI-Twin's simulation models capture during training. Uncertainty in compiled data can exist due to inference and measurement errors. Inference errors exist when a computational model is used to map measured data to variables of interest, such as context or outcome variables (eg, an algorithm that computes step counts will produce estimates with error). Measurement errors exist due to the inherent uncertainty in measurements, such as surveys of psychological constructs or sensor-based estimates of quantities that can be corrupted by external sources of noise. Distinguishing uncertainty in measurements from uncertainty in model predictions due to limited knowledge has been discussed extensively in the statistics and machine learning literature [79]. These sources of uncertainty can be simulated and propagated through to the simulation results by introducing noise terms with variances that can be scaled in magnitude to help quantify the sensitivity of simulation results to these errors [80].

VVUQ Is Continual

The VVUQ process is continual for a JITAI-Twin. Once a JITAI-Twin is used to inform decision-making in practice, a JITAI-Twin should be reevaluated following the completed deployment. This reevaluation will help inform the JITAI-Twin prior to its use in decision-making for the subsequent deployment—and the cycle continues, as depicted in Figure 4.

Evaluating Candidate Algorithms According to JITAI-Twin Simulation Results

Once a JITAI-Twin has been evaluated satisfactorily, the JITAI-Twin's simulation results can be used to decide which decision-making algorithm to deploy as part of the upcoming JITAI. Note that these simulations should be conducted with noise variables that match the characteristics of the noise observed in previously collected data (eg, informed by the simulation models' residuals when fit to prior data). This helps evaluate how well candidate decision-making algorithms will perform in real-world settings that are not completely accounted for by the learned aspects of the simulation models. For each candidate decision-making algorithm, the JITAI-Twin will produce a set of simulation results. These results will correspond to multiple simulated deployments for each variant of the base simulation testbed, as illustrated in Figure 5 (for a real example of simulation results, see Trella et al [7]). The goal at this stage is to identify the specific algorithm to be used for deployment based on the results shown across all simulations.

XSL•FO
RenderX

Figure 5. Illustrative example of a digital twin for just-in-time adaptive intervention's simulation results. For each simulation testbed variant of a digital twin for just-in-time adaptive intervention, candidate digital twin for just-in-time adaptive intervention decision-making algorithms will be evaluated according to metrics that the team deems useful for deciding which algorithm to deploy onboard the real digital twin for just-in-time adaptive intervention. Performance metrics may not always agree with respect to which algorithm to select. Trade-offs must often be made based on the performance priorities for the digital twin for just-in-time adaptive intervention to be deployed. JITAI: digital twin for just-in-time adaptive intervention; JITAI-Twin: digital twin for just-in-time adaptive intervention.

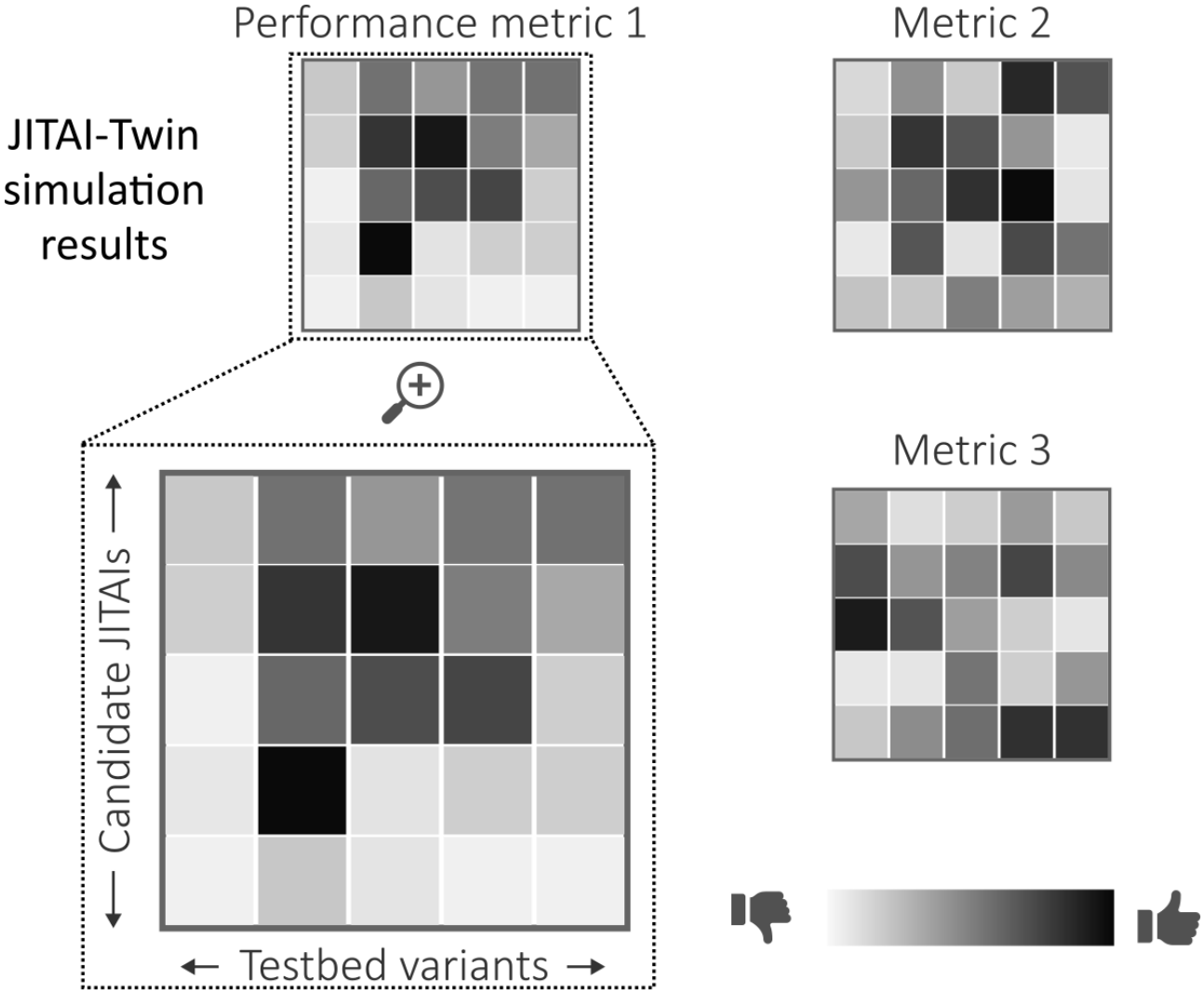

The metrics used to evaluate candidate algorithms and decide which to deploy will depend on the research team's priorities. If the primary goal of an intervention is to elicit a positive impact on average across the entire deployment sample, then a measure of central tendency across all virtual participants is likely sufficient. An example measure would be the mean across all virtual participants of the average cumulative reward across all simulated deployments (ie, 1N∑n=1N1S∑s=1S∑t=1TRt, s, n, where s indexes random seed, *n* indexes virtual participant, and *t* indexes time within each simulated deployment). However, if one goal of an intervention is to ensure that individuals with disorders that are more treatment resistant experience some benefit, then a lower percentile might be taken into consideration (eg, 25th percentile of 1S∑s=1S∑t=1TRt,s,n). Other metrics could be used to investigate considerations related to time. For example, if an intervention trial may be terminated early if outcomes are not met [81], the algorithms may be evaluated for an initial subset of time (ie, instead of 1N∑n=1N1S∑s=1S∑t=1TRt,s,n, one could use 1N∑n=1N1S∑s=1S∑t=1T`Rt,s,n for T′<T). If early participant disengagement is a primary concern, several T` values can be used to compare the speed of learning across the candidate RL algorithms. Table 2 summarizes the 3 example metrics described. Note that if multiple metrics are used to compare algorithms, design trade-offs will be necessary, as illustrated in Figure 5.

XSL•FO
RenderX

Table . Example metrics to evaluate candidate algorithms[a].

Example metric	Use case	Formula
Mean cumulative reward	Choose the algorithm that elicits the greatest positive impact on average across the entire deployment sample	1N∑n=1N1S∑s=1S∑t=1TRt,s,n
Lower percentile of the cumulative reward (eg, 25th percentile)	Choose the algorithm that improves outcomes the most for those that are least benefited (eg, treatment resistant)	Lower percentile of 1S∑s=1S∑t=1TRt,s,n
Mean, early terminated reward	Choose the algorithm that most improves the mean positive impact within a shorter time period than the full deployment	1N∑n=1N1S∑s=1S∑t=1T`Rt,s,n for T′<T

[a]$s \in \{1,2,...,S\}$ indexes random seed, $n \in \{1,2,...,N\}$ indexes virtual participant, and $t \in \{1,2,...,T\}$ indexes time within each simulated deployment.

For HeartSteps V2, the average total reward was used as the primary metric [26]. For HeartSteps V3, the average and 95% CIs for the cumulative reward are also reported [25]. CIs provide an indication of variability across virtual participants.

Discussion

Implications for Research and Practice

The JITAI-Twin framework we elucidate and the guiding principles we provide are informed by prior work in designing RL algorithms for JITAIs. Aside from the HeartSteps JITAI described in Guiding Principles to Follow for JITAI-Twin Design [25,26], 2 RL-empowered JITAIs that have been deployed in clinical trials were used to inform this paper. For the “Oralytics” JITAI designed to improve oral health in disadvantaged populations [82], the first iteration of a JITAI-Twin was designed and used to inform design decisions (hyperparameters of the RL algorithm’s reward [53]) for the JITAI’s clinical trial that started in fall 2023 and concluded in summer 2024 [83]. A unique challenge faced when designing the Oralytics JITAI—that was simulated as part of the JITAI-Twin—was delayed access to oral self-care behavior data from participants due to the use of a commercial smartbrush. Issues with data accessibility are common with commercial products and should be simulated by a JITAI-Twin. For the “MiWaves” JITAI designed to reduce cannabis use in emerging adults, the first iteration of a JITAI-Twin was used to select the reBandit algorithm over baseline approaches [32]. MiWaves was then deployed in a preregistered feasibility and acceptability study which was completed in fall 2024 [84]. A challenge faced with MiWaves was that the reward for the RL algorithm was a measure of engagement with the JITAI, but engagement with a JITAI is latent and difficult to quantify. A surrogate reward was thus defined and used for JITAI-Twin simulations to evaluate candidate algorithms. In general, the JITAI-Twin framework can be extended to many other domains, so long as the guiding principles of team science and attention to detail are followed to address domain-specific challenges.

The more costly it is to evaluate design decisions in the real world, the more valuable a JITAI-Twin is for optimizing and continually improving a JITAI. A/B testing is a tried-and-true approach to comparing design choices in practice [85], and is the basis for MRTs and other optimization experiments [14]. A significant downside to A/B testing and optimization experiments is that factorial designs can quickly become untenable when the cost of collecting data is nonnegligible and several nonobvious design decisions need to be made prior to the deployment of a JITAI. Especially for JITAIs that leverage RL and control engineering approaches, several design decisions must often be made just for model selection [8,43]. On the other hand, some design decisions may be difficult to model using prior data and simulate via a JITAI-Twin. An example would be whether a new intervention component that differs substantially from previous components should be added to a JITAI (eg, noninvasive vagus nerve stimulation in tandem with mindfulness interventions for stress management [86]). Optimization experiments are best reserved for these difficult-to-simulate design decisions, especially if the experimental cost of evaluating decisions is non-negligible. Thus, the recommended path forward is to use JITAI-Twins and optimization experiments in tandem to optimize and continually improve JITAIs.

Compared to person-specific digital twins, there are practical advantages to using digital twins of a subpopulation when optimizing and continually improving a JITAI. A person-specific digital twin becomes obsolete as soon as the specific person is no longer being considered for upcoming deployments of a JITAI. Digital twins of a subpopulation remain useful so long as the JITAI is being continually improved for the target subpopulation. Practically speaking, digital twins of a subpopulation also reduce the need to collect large amounts of data from specific individuals. Although person-specific digital twins enhance personalization and circumvent the need to model subpopulation heterogeneity, prior work shows that months of exploratory data collection can be required to satisfactorily construct person-specific models for a JITAI [87]. In behavioral settings where adherence remains an issue and disengagement often spells the end of an individual’s participation, collecting large amounts of data from a single individual may be a challenge. Future work should consider balancing the 2 approaches to better balance learning within a deployment for improved personalization (person-specific) and learning between deployments to continually improve a JITAI system (subpopulation-level).

Other opportunities for future research on JITAI-Twins include the use of large language models to improve the abstraction of domain science (often described using natural language) into computational constructs, the use of large language models to

XSL•FO
RenderX

simulate participant interactions with a JITAI [88], and the use of foundation models to leverage large sums of previously collected data that may be openly available but rife with data-deployment mismatches. See Section S4 of the Multimedia Appendix 1 for further discussion of future work.

Limitations

Our discussion of JITAI-Twins is limited in a few respects. The underlying concepts behind JITAI-Twins may find broader applicability beyond JITAI deployments. Other adaptive digital interventions may also benefit from digital twins. However, this paper is restricted to JITAI applications to ensure guidance is informed by practical experience. The example JITAI-Twins in this paper only include those that the coauthors of this paper have designed. Although others have designed JITAIs empowered by RL algorithms [89,90], the use of JITAI-Twins is not yet standardized and the publication of simulation environments and results is not yet commonplace, limiting our ability to include other simulations and JITAIs as examples. One of the goals of this paper is to bring JITAI-Twins to the forefront of discussion to overcome this gap. Future work is also necessary to better elucidate practical considerations, such as the computational and labor costs associated with designing and maintaining a JITAI-Twin and potential comparisons with the costs of conducting optimization experiments for the same design decisions. Other practical considerations include scalability. Although the framework as described is not dependent on the amount of data available or the scale of a JITAI's deployment, future work is necessary to evaluate the generalizability of the JITAI-Twin framework to larger-scale deployments. Guidance is needed on adapting and improving RL algorithms based on JITAI-Twin simulation results. Although the previous section of this paper provides initial guidance on RL algorithm selection, we leave further discussion to future work, as the scope of this paper focuses on JITAI-Twins. Future work is also needed to empirically validate the benefit of each of the proposed guiding principles and design steps for JITAI-Twins to potentially improve or provide additional evidence for aspects of the framework.

Conclusions

In this paper, we present the JITAI-Twin framework as an instantiation of "digital twins of a subpopulation" for JITAI optimization and continual improvement [24]. We detail how JITAI-Twin simulation results can be used to make difficult design decisions, such as which decision-making algorithm to deploy as part of a JITAI. This serves as information flow from the virtual twin to the physical twin. We also show how data collected from a JITAI's deployment can be used to update the JITAI-Twin prior to the subsequent deployment, demonstrating information flow from the physical twin to the virtual twin. JITAI-Twins are thus "fit for [the] purpose" of simulating the behavior of a JITAI deployment's subpopulation [13]. The impact that JITAI-Twins can have would be greatly enhanced through collaborative efforts to share computational resources for designing and evaluating JITAI-Twins (eg, codebases); curate data that could be used to initialize, train, or update JITAI-Twins; and establish benchmarks to help practitioners better understand when a JITAI-Twin is ready to inform decisions or to help researchers evaluate new methods for JITAI-Twin modeling or simulation. Toolboxes that enhance the accessibility of JITAI-Twin design (eg, as a possible addition to the JustIn platform [91]) would also facilitate more widespread adoption of JITAI-Twin methods to improve the optimization and continual improvement of JITAIs.

Acknowledgments

In preparing this article, generative artificial intelligence was used for tasks that fall within classification 1 of the taxonomy published by the International Association of Scientific, Technical & Medical Publishers: "refinement, correction, editing, or formatting the manuscript to improve the clarity of language."

Funding

The work of AHG was supported by Schmidt Science Fellows, in partnership with the Rhodes Trust. This research was funded by National Institutes of Health grants K99EB037411, P50DA054039, P41EB028242, R01HL125440, 2R01HL125440, U01CA229445, R01LM013107, UH3DE028723, and P30AG073107-03 GY3 Pilots.

Data Availability

This paper is accompanied by Multimedia Appendix 1. Section S1 provides a broader overview of related literature, Section S2 provides a primer on RL algorithms and simulation models in the context of JITAIs, Section S3 provides additional examples of how data impoverishment was addressed in prior work, and Section S4 describes open questions and future directions in JITAI-Twin design. Example simulation testbeds used as single deployment iterations of JITAI-Twins (ie, these simulation testbeds were used to inform a single deployment – future work is necessary to continually adapt these testbeds and form true JITAI-Twins) are available online [92,93].

Authors' Contributions

Conceptualization: AHG, SAM
Writing – original draft: AHG
Visualization: AHG, DG

XSL•FO
RenderX

Funding Acquisition: AHG, PK, SAM
Writing – review & editing: DG, SG, ZX, AT, PK, SAM
Supervision: SAM

Conflicts of Interest

SAM holds concurrent appointments at Harvard University and as an Amazon Scholar. This paper describes work performed at Harvard University and is not associated with Amazon.

Multimedia Appendix 1
Includes a more extensive review of related literature, a primer on reinforcement learning algorithms and simulation models, additional examples of data impoverishment faced during digital twin for just-in-time adaptive intervention design and updates, and additional discussion on open questions and future directions in digital twin for just-in-time adaptive intervention design.
[DOCX File, 310 KB - jmir_v28i1e72830_app1.docx]

References

1. Spruijt-Metz D, Nilsen W. Dynamic models of behavior for just-in-time adaptive interventions. IEEE Pervasive Comput 2014;13(3):13-17. [doi: 10.1109/MPRV.2014.46]
2. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462. [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
3. Nahum-Shani I, Hekler EB, Spruijt-Metz D. Building health behavior models to guide the development of just-in-time adaptive interventions: a pragmatic framework. Health Psychol 2015 Dec;34S:1209-1219. [doi: 10.1037/hea0000306] [Medline: 26651462]
4. Spruijt-Metz D, Wen CKF, O'Reilly G, et al. Innovations in the use of interactive technology to support weight management. Curr Obes Rep 2015 Dec;4(4):510-519. [doi: 10.1007/s13679-015-0183-6] [Medline: 26364308]
5. Klasnja P, Smith S, Seewald NJ, et al. Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of HeartSteps. Ann Behav Med 2019 May 3;53(6):573-582. [doi: 10.1093/abm/kay067] [Medline: 30192907]
6. Hekler EB, Rivera DE, Martin CA, et al. Tutorial for using control systems engineering to optimize adaptive mobile health interventions. J Med Internet Res 2018 Jun 28;20(6):e214. [doi: 10.2196/jmir.8622] [Medline: 29954725]
7. Trella AL, Zhang KW, Nahum-Shani I, Shetty V, Doshi-Velez F, Murphy SA. Designing reinforcement learning algorithms for digital interventions: pre-implementation guidelines. Algorithms 2022 Aug;15(8):255. [doi: 10.3390/a15080255] [Medline: 36713810]
8. Probst P, Boulesteix AL, Bischl B. Tunability: importance of hyperparameters of machine learning algorithms. J Mach Learn Res 2019;18:1-32. [doi: 10.5555/3322706.3361994]
9. Russell LB, Norton LA, Pagnotti D, et al. Using clinical trial data to estimate the costs of behavioral interventions for potential adopters: a guide for trialists. Med Decis Making 2021 Jan;41(1):9-20. [doi: 10.1177/0272989X20973160] [Medline: 33218296]
10. Monashefsky A, Alon D, Baranowski T, et al. How much did it cost to develop and implement an eHealth intervention for a minority children population that overlapped with the COVID-19 pandemic? Contemp Clin Trials 2023 Feb;125:107044. [doi: 10.1016/j.cct.2022.107044] [Medline: 36473682]
11. Mohr DC, Schueller SM, Riley WT, et al. Trials of intervention principles: evaluation methods for evolving behavioral intervention technologies. J Med Internet Res 2015 Jul 8;17(7):e166. [doi: 10.2196/jmir.4391] [Medline: 26155878]
12. An M, Dusing SC, Harbourne RT, Sheridan SM, START-Play Consortium. What really works in intervention? Using fidelity measures to support optimal outcomes. Phys Ther 2020 May 18;100(5):757-765. [doi: 10.1093/ptj/pzaa006] [Medline: 31944249]
13. National Academies of Sciences Engineering and Medicine. Foundational Research Gaps and Future Directions for Digital Twins: The National Academies Press; 2024:1-202. [doi: 10.17226/26894]
14. Qian T, Walton AE, Collins LM, et al. The microrandomized trial for developing digital interventions: experimental design and data analysis considerations. Psychol Methods 2022 Oct;27(5):874-894. [doi: 10.1037/met0000283] [Medline: 35025583]
15. Collins LM, Murphy SA, Strecher V. The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): new methods for more potent eHealth interventions. Am J Prev Med 2007 May;32(5 Suppl):S112-S118. [doi: 10.1016/j.amepre.2007.01.022] [Medline: 17466815]
16. Laubenbacher R, Mehrad B, Shmulevich I, Trayanova N. Digital twins in medicine. Nat Comput Sci 2024 Mar;4(3):184-191. [doi: 10.1038/s43588-024-00607-6] [Medline: 38532133]
17. Kamel Boulos MN, Zhang P. Digital twins: from personalised medicine to precision public health. J Pers Med 2021 Jul 29;11(8):745. [doi: 10.3390/jpm11080745] [Medline: 34442389]

XSL•FO
RenderX

18. Björnsson B, Borrebaeck C, Elander N, et al. Digital twins to personalize medicine. Genome Med 2020 Dec;12(1):1-4. [doi: 10.1186/s13073-019-0701-3]
19. Laubenbacher R, Sluka JP, Glazier JA. Using digital twins in viral infection. Science 2021 Mar 12;371(6534):1105-1106. [doi: 10.1126/science.abf3370] [Medline: 33707255]
20. Katsoulakis E, Wang Q, Wu H, et al. Digital twins for health: a scoping review. NPJ Digit Med 2024 Mar 22;7(1):77. [doi: 10.1038/s41746-024-01073-0] [Medline: 38519626]
21. Zhang K, Zhou HY, Baptista-Hon DT, et al. Concepts and applications of digital twins in healthcare and medicine. Patterns (N Y) 2024 Aug 9;5(8):101028. [doi: 10.1016/j.patter.2024.101028] [Medline: 39233690]
22. Elkefi S, Asan O. Digital twins for managing health care systems: rapid literature review. J Med Internet Res 2022 Aug 16;24(8):e37641. [doi: 10.2196/37641] [Medline: 35972776]
23. Armeni P, Polat I, De Rossi LM, Diaferia L, Meregalli S, Gatti A. Digital twins in healthcare: is it the beginning of a new era of evidence-based medicine? A critical review. J Pers Med 2022 Jul 30;12(8):8. [doi: 10.3390/jpm12081255] [Medline: 36013204]
24. National Academies of Sciences Engineering and Medicine. Opportunities and challenges for digital twins in biomedical research—proceedings of a workshop in brief 2023. : The National Academied Press; 2023. [doi: 10.17226/26922]
25. Gao D, Lai HY, Klasnja P, Murphy SA. Harnessing causality in reinforcement learning with bagged decision times. . Preprint posted online on Oct, 2024 URL: https://arxiv.org/abs/2410.14659v1 [accessed 2024-11-04]
26. Liao P, Greenewald K, Klasnja P, Murphy S. Personalized HeartSteps: a reinforcement learning algorithm for optimizing physical activity. Proc ACM Interact Mob Wearable Ubiquitous Technol 2020 Mar;4(1):1. [doi: 10.1145/3381007] [Medline: 34527853]
27. Spruijt-Metz D, Marlin BM, Pavel M, et al. Advancing behavioral intervention and theory development for mobile health: the HeartSteps II protocol. Int J Environ Res Public Health 2022 Feb 17;19(4):2267. [doi: 10.3390/ijerph19042267] [Medline: 35206455]
28. Science of Behavior Change (SOBC). National Institutes of Health. URL: https://commonfund.nih.gov/behaviorchange [accessed 2024-11-04]
29. Simonton KL. Testing a model of personal attributes and emotions regarding physical activity and sedentary behaviour. Int J Sport Exerc Psychol 2021 Sep 3;19(5):848-865. [doi: 10.1080/1612197X.2020.1739112]
30. Mullen SP, Olson EA, Phillips SM, et al. Measuring enjoyment of physical activity in older adults: invariance of the physical activity enjoyment scale (paces) across groups and time. Int J Behav Nutr Phys Act 2011 Sep 27;8(1):103. [doi: 10.1186/1479-5868-8-103] [Medline: 21951520]
31. Mohiyeddini C, Pauli R, Bauer S. The role of emotion in bridging the intention–behaviour gap: the case of sports participation. Psychol Sport Exerc 2009 Feb;10(2):226-234. [doi: 10.1016/j.psychsport.2008.08.005]
32. Ghosh S, Guo Y, Hung PY, et al. ReBandit: random effects based online RL algorithm for reducing cannabis use. . Preprint posted online on Feb, 2024 URL: http://arxiv.org/abs/2402.17739 [accessed 2024-08-13]
33. Holford NHG, Hale M, Ko HC, et al. Simulation in drug development: good practices. : CDDS; 1999 URL: https://holford.fmhs.auckland.ac.nz/docs/simulation-in-drug-development-good-practices.pdf [accessed 2025-02-23]
34. Ghosh S, Guo Y, Hung PY, et al. Reproducible workflow for online AI in digital health. Preprint posted online on Sep, 2025 URL: https://www.arxiv.org/pdf/2509.13499 [accessed 2025-09-29]
35. Hall KL, Vogel AL, Huang GC, et al. The science of team science: a review of the empirical evidence and research gaps on collaboration in science. Am Psychol 2018;73(4):532-548. [doi: 10.1037/amp0000319] [Medline: 29792466]
36. Box GEP. Science and statistics. J Am Stat Assoc 1976 Dec;71(356):791-799. [doi: 10.1080/01621459.1976.10480949]
37. Baker RE, Peña JM, Jayamohan J, Jérusalem A. Mechanistic models versus machine learning, a fight worth fighting for the biological community? Biol Lett 2018 May;14(5):5. [doi: 10.1098/rsbl.2017.0660] [Medline: 29769297]
38. Biecek P, Burzykowski T. Explore, explain, and examine predictive models. In: Explanatory Model Analysis: Chapman and Hall/CRC; 2021. [doi: 10.1201/9780429027192]
39. Shivaswamy P, Chandrashekar A. Bias-variance decomposition for ranking. 2021 Mar 8 Presented at: WSDM 2021 - Proceedings of the 14th ACM International Conference on Web Search and Data Mining; Virtual Event Israel p. 472-480. [doi: 10.1145/3437963.3441772]
40. Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. Nat Rev Phys 2021;3(6):422-440. [doi: 10.1038/s42254-021-00314-5]
41. Wagg D, Burr C, Shepherd J, Xuereb Conti Z, Enzer M, Niederer S. The philosophical foundations of digital twinning. . Preprint posted online on Jan, 2024. [doi: 10.31224/3500]
42. Geron A. Hands-on Machine Learning with Scikit-Learn, Keras & Tensorflow, 2nd edition: O'Reilly Media; 2019.
43. Forster MR. Key concepts in model selection: performance and generalizability. J Math Psychol 2000 Mar;44(1):205-231. [doi: 10.1006/jmps.1999.1284] [Medline: 10733865]
44. Kernbach JM, Staartjes VE. Foundations of machine learning-based clinical prediction modeling: Part II—generalization and overfitting. Acta Neurochir 2022:15-21. [doi: 10.1007/978-3-030-85292-4_3]

XSL•FO
RenderX

45. Kohavi R, Wolpert D. Bias plus variance decomposition for zero-one loss functions. 1996 Presented at: Proceedings of the Thirteenth International Conference on Machine Learning; Jul 3-6, 1996 URL: https://dl.acm.org/doi/10.5555/3091696.3091730 [accessed 2026-02-18]
46. Dar Y, Muthukumar V, Baraniuk RG. A farewell to the bias-variance tradeoff? An overview of the theory of overparameterized machine learning. . Preprint posted online on Sep, 2021 URL: https://arxiv.org/pdf/2109.02355 [accessed 2026-02-09]
47. Benk J, Beyersdorf F. Commentary: Inattention to detail is the hallmark of mediocrity. J Thorac Cardiovasc Surg 2022 Dec;164(6):1888-1889. [doi: 10.1016/j.jtcvs.2021.09.026] [Medline: 34602255]
48. Why the mars probe went off course - IEEE spectrum. IEEE Spectr. URL: https://spectrum.ieee.org/why-the-mars-probe-went-off-course [accessed 2025-09-29]
49. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O. A survey on missing data in machine learning. J Big Data 2021;8(1):140. [doi: 10.1186/s40537-021-00516-9] [Medline: 34722113]
50. Singh J, Sato M, Ohkuma T. On missingness features in machine learning models for critical care: observational study. JMIR Med Inform 2021 Dec 8;9(12):e25022. [doi: 10.2196/25022] [Medline: 34889756]
51. Digitale JC, Martin JN, Glymour MM. Tutorial on directed acyclic graphs. J Clin Epidemiol 2022 Feb;142:264-267. [doi: 10.1016/j.jclinepi.2021.08.001] [Medline: 34371103]
52. Lipsky AM, Greenland S. Causal directed acyclic graphs. JAMA 2022 Mar 15;327(11):1083-1084. [doi: 10.1001/jama.2022.1816] [Medline: 35226050]
53. Trella AL, Zhang KW, Nahum-Shani I, Shetty V, Doshi-Velez F, Murphy SA. Reward design for an online reinforcement learning algorithm supporting oral self-care. Proc Innov Appl Artif Intell Conf 2023 Jun;37(13):15724-15730. [doi: 10.1609/aaai.v37i13.26866]
54. Turban E. Review of expert systems technology. IEEE Trans Eng Manage 1988 May;35(2):71-81. [doi: 10.1109/17.6007]
55. Sutton R, Barto A. Reinforcement Learning: An Introduction: MIT Press; 2018. URL: http://incompleteideas.net/book/the-book-2nd.html [accessed 2026-02-18]
56. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (N Y) 2023 Sep 8;4(9):100804. [doi: 10.1016/j.patter.2023.100804] [Medline: 37720327]
57. Daniel JR, Benjamin V, Abbas K, Ian O, Zheng W. A tutorial on Thompson sampling. Found Trends Mach Learn 2018 Jul 12;11(1):1-99. [doi: 10.1561/2200000070]
58. van de Schoot R, Depaoli S, King R, et al. Bayesian statistics and modelling. Nat Rev Methods Primers 2021;1(1):1-26. [doi: 10.1038/s43586-020-00001-2]
59. Cuff A. The evolution of digital health and its continuing challenges. BMC Digit Health 2023;1(1):3. [doi: 10.1186/s44247-022-00004-x] [Medline: 38014367]
60. Sumner JA, Carey RN, Michie S, Johnston M, Edmondson D, Davidson KW. Using rigorous methods to advance behaviour change science. Nat Hum Behav 2018 Nov;2(11):797-799. [doi: 10.1038/s41562-018-0471-8] [Medline: 30931398]
61. Berntson GG, Cacioppo JT, Sarter M. Psychophysiology. Biol Psychiatry 2002 Jul:123-138. [doi: 10.1002/0470854871]
62. Bach DR, Castegnetti G, Korn CW, Gerster S, Melinscak F, Moser T. Psychophysiological modeling: current state and future directions. Psychophysiology 2018 Nov;55(11):e13214. [doi: 10.1111/psyp.13209] [Medline: 30175471]
63. Rabanser S, Unnemann S. Failing loudly: an empirical study of methods for detecting dataset shift. Presented at: Proceedings of the 33rd International Conference on Neural Information Processing Systems; Dec 8-14, 2019. [doi: 10.5555/3454287.3454412]
64. Van Cappellen P, Rice EL, Catalino LI, Fredrickson BL. Positive affective processes underlie positive health behaviour change. Psychol Health 2018 Jan;33(1):77-97. [doi: 10.1080/08870446.2017.1320798] [Medline: 28498722]
65. Rhodes RE, Kates A. Can the affective response to exercise predict future motives and physical activity behavior? A systematic review of published evidence. Ann Behav Med 2015 Oct;49(5):715-731. [doi: 10.1007/s12160-015-9704-5] [Medline: 25921307]
66. Brand R, Ekkekakis P. Affective–reflective theory of physical inactivity and exercise. Ger J Exerc Sport Res 2018 Mar;48(1):48-58. [doi: 10.1007/s12662-017-0477-9]
67. Williams DM, Dunsiger S, Ciccolo JT, Lewis BA, Albrecht AE, Marcus BH. Acute affective response to a moderate-intensity exercise stimulus predicts physical activity participation 6 and 12 months later. Psychol Sport Exerc 2008 May;9(3):231-245. [doi: 10.1016/j.psychsport.2007.04.002] [Medline: 18496608]
68. Sankaran K, Holmes SP. Generative models: an interdisciplinary perspective. Annu Rev Stat Appl 2023 Mar 10;10(1):325-352 [FREE Full text] [doi: 10.1146/annurev-statistics-033121-110134]
69. Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma. Neural Comput 1992 Jan;4(1):1-58. [doi: 10.1162/neco.1992.4.1.1]
70. Roelof R, Fridovich-Keil S, Miller J, et al. A meta-analysis of overfitting in machine learning. 2019 Presented at: Proceedings of the 33rd International Conference on Neural Information Processing Systems; Dec 8-14, 2019. [doi: 10.5555/3454287.3455110]
71. Bryan CJ, Tipton E, Yeager DS. Behavioural science is unlikely to change the world without a heterogeneity revolution. Nat Hum Behav 2021 Aug;5(8):980-989. [doi: 10.1038/s41562-021-01143-3] [Medline: 34294901]

XSL•FO
RenderX

72. Nikbakht M, Gazi AH, Zia J, et al. Synthetic seismocardiogram generation using a transformer-based neural network. J Am Med Inform Assoc 2023 Jun 20;30(7):1266-1273. [doi: 10.1093/jamia/ocad067] [Medline: 37053380]
73. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11. [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
74. Sel K, Hawkins-Daarud A, Chaudhuri A, et al. Survey and perspective on verification, validation, and uncertainty quantification of digital twins for precision medicine. NPJ Digit Med 2025 Jan 17;8(1):40. [doi: 10.1038/s41746-025-01447-y] [Medline: 39825103]
75. Bingham D, Butler T, Estep D. Inverse problems for physics-based process models. Annu Rev Stat Appl 2024 Apr 22;11(1):461-482. [doi: 10.1146/annurev-statistics-031017-100108]
76. Song H, Song M, Liu X. Online autonomous calibration of digital twins using machine learning with application to nuclear power plants. Appl Energy 2022 Nov;326:119995. [doi: 10.1016/j.apenergy.2022.119995]
77. Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. Nat Mach Intell 2019;1(1):20-23. [doi: 10.1038/s42256-018-0004-1]
78. Saltelli A, Aleksankina K, Becker W, et al. Why so many published sensitivity analyses are false: a systematic review of sensitivity analysis practices. Environ Model Softw 2019 Apr;114:29-39. [doi: 10.1016/j.envsoft.2019.01.012]
79. Gal Y, Koumoutsakos P, Lanusse F, Louppe G, Papadimitriou C. Bayesian uncertainty quantification for machine-learned models in physics. Nat Rev Phys 2022;4(9):573-577. [doi: 10.1038/s42254-022-00498-4]
80. Lin DJ, Gazi AH, Kimball J, Nikbakht M, Inan OT. Real-time seismocardiogram feature extraction using adaptive Gaussian mixture models. IEEE J Biomed Health Inform 2023 Aug;27(8):3889-3899. [doi: 10.1109/JBHI.2023.3273989] [Medline: 37155395]
81. Ellenberg SS, Shaw PA. Early termination of clinical trials for futility - considerations for a data and safety monitoring board. NEJM Evid 2022 Jul;1(7):EVIDctw2100020. [doi: 10.1056/EVIDctw2100020] [Medline: 38319261]
82. Trella AL, Zhang KW, Carpenter SM, et al. Oralytics reinforcement learning algorithm. . Preprint posted online on Jun, 2024 URL: https://arxiv.org/abs/2406.13127v1 [accessed 2024-08-13]
83. Trella AL, Zhang KW, Jajal H, et al. A deployed online reinforcement learning algorithm in an oral health clinical trial. . Preprint posted online on Sep, 2024 URL: https://arxiv.org/abs/2409.02069v1 [accessed 2024-09-13]
84. Ghosh S, Guo Y, Hung PY, et al. MiWaves reinforcement learning algorithm. Aug. Preprint posted online on 2024 URL: https://arxiv.org/abs/2408.15076v1 [accessed 2024-09-11]
85. Siroker D, Koomen P. A/B Testing: The Most Powerful Way to Turn Clicks into Customers: John Wiley & Sons; 2015. URL: https://books.google.com/books/about/A_B_Testing.html?id=9xUWCgAAQBAJ [accessed 2025-09-29]
86. Gazi AH, An S, Sanchez-Perez JA, et al. Modeling latent dynamics of the autonomic nervous system in response to trauma recall and non-invasive vagus nerve stimulation. IEEE Trans Biomed Eng 2025;73(1):220-233. [doi: 10.1109/TBME.2025.3580051]
87. Park J, Kim M, El Mistiri M, et al. Advancing understanding of just-in-time states for supporting physical activity (Project JustWalk JITAI): protocol for a system ID study of just-in-time adaptive interventions. JMIR Res Protoc 2023 Sep 26;12:e52161. [doi: 10.2196/52161] [Medline: 37751237]
88. Karine K, Marlin BM. Enhancing adaptive behavioral interventions with LLM inference from participant-described states. Jul. Preprint posted online on 2025 URL: https://arxiv.org/pdf/2507.03871 [accessed 2025-09-30]
89. Figueroa CA, Aguilera A, Chakraborty B, et al. Adaptive learning algorithms to optimize mobile applications for behavioral health: guidelines for design decisions. J Am Med Inform Assoc 2021 Jun 12;28(6):1225-1234. [doi: 10.1093/jamia/ocab001] [Medline: 33657217]
90. Gönül S, Namlı T, Coşar A, Toroslu İ. A reinforcement learning based algorithm for personalization of digital, just-in-time, adaptive interventions. Artif Intell Med 2021 May;115:102062. [doi: 10.1016/j.artmed.2021.102062] [Medline: 34001322]
91. Introducing JustIn. d3center. URL: https://d3c.isr.umich.edu/justin/ [accessed 2025-09-29]
92. ReBandit. GitHub. URL: https://github.com/susobhang70/reBandit [accessed 2026-02-09]
93. StatisticalReinforcementLearningLab. GitHub. URL: https://github.com/StatisticalReinforcementLearningLab/oralytics_algorithm_design [accessed 2026-02-09]

Abbreviations

AI: artificial intelligence
DAG: directed acyclic graph
JITAI: just-in-time adaptive intervention
JITAI-Twin: digital twin for just-in-time adaptive intervention
mHealth: mobile health
MRT: microrandomized trial
RL: reinforcement learning
VVUQ: verification, validation, and uncertainty quantification

XSL•FO
RenderX

Edited by J Sarvestan; submitted 09.May.2025; peer-reviewed by KH Lin, VG Prabhu; revised version received 19.Dec.2025; accepted 20.Dec.2025; published 06.Mar.2026.

Please cite as:
Gazi AH, Gao D, Ghosh S, Xu Z, Trella AL, Klasnja P, Murphy SA
Digital Twins for Just-in-Time Adaptive Interventions (JITAIs): Framework for Optimizing and Continually Improving JITAIs
J Med Internet Res 2026;28:e72830
URL: https://www.jmir.org/2026/1/e72830
doi:10.2196/72830

© Asim H Gazi, Daiqi Gao, Susobhan Ghosh, Ziping Xu, Anna L Trella, Predrag Klasnja, Susan A Murphy. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI Agents and Epidemic Intelligence on Respiratory Infectious Diseases: Toward a Conceptual Framework Integrating Decision Support

Liuyang Yang[1,2,3*], PhD; Liyu Shan[1*], MD; Xiaolin Cao[4*], PhD; Jinzhao Cui[2], MPH; Michael Tong[5], PhD; Yan Niu[6], PhD; Ting Zhang[2], PhD

[1]The Third Affiliated Hospital of Kunming Medical University, Yunnan Cancer Hospital, Peking University Cancer Hospital Yunnan, Kunming, Yunnan, China

[2]School of Population Medicine and Public Health, Chinese Academy of Medical Sciences & Peking Union Medical College, No. 9 Dongdan Santiao, Beijing, Beijing, China

[3]School of Data Science, Fudan University, Shanghai, Shanghai, China

[4]Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, China

[5]National Centre for Epidemiology and Population Health, The Australian National University, Canberra, Australian Capital Territory 2601, Australia

[6]Public Health Emergency Center, Chinese Center for Disease Control and Prevention, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Ting Zhang, PhD
School of Population Medicine and Public Health, Chinese Academy of Medical Sciences & Peking Union Medical College, No. 9 Dongdan Santiao, Beijing, Beijing, China

Abstract

Traditional epidemic intelligence relies heavily on human epidemiologists for data interpretation and reporting, which makes it resource intensive, slow to respond, and vulnerable to variability in professional expertise. To overcome these limitations, we propose an expanded conceptual epidemic intelligence quadripartite framework that extends the classical trinity of (1) surveillance, (2) risk evaluation, and (3) early warning with a fourth pillar, (4) decision support and intervention optimization through AI agents. Acting as 24/7 digital epidemiologists, multiagent systems can integrate heterogeneous signals from multisource surveillance systems, conduct contextual risk evaluation and adaptive forecasting, generate tailored early warnings, and provide actionable recommendations for targeted control—closing the loop between detection and response. Embedding interpretability and mandatory human-in-the-loop oversight enhances trust and accountability. Nonetheless, real-world deployment requires addressing context-specific challenges of data quality, interoperability, robustness, governance, circular reporting, and equity. If designed with transparency, inclusiveness, and resilience, AI agents have the potential to transform epidemic intelligence into a continuously adaptive and globally connected system.

(J Med Internet Res 2026;28:e86936) doi:10.2196/86936

KEYWORDS

artificial intelligence agents; respiratory infectious diseases; surveillance; risk evaluation; early warning; decision support

Introduction

Respiratory diseases, such as seasonal influenza, respiratory syncytial virus (RSV), and emerging zoonotic respiratory pathogens, remain among the most significant causes of global morbidity and mortality. Seasonal influenza infects up to 1 billion people annually, causing an estimated 290,000 to 650,000 deaths each year [1], while RSV is responsible for approximately 33 million cases and 100,000 child deaths worldwide each year [2]. The rapid transmissibility, unpredictable evolution, and potential for interspecies spillover events make these pathogens a persistent global health threat. Recent outbreaks, such as the COVID-19 pandemic, underscore the urgent need for advanced epidemic intelligence. In particular, capturing the critical "golden window"— a short but decisive period during which timely detection and intervention can accelerate responses such as vaccine development and deployment—is essential to limit transmission, prevent missed opportunities for effective containment, and reduce the broader societal and economic impacts of respiratory infectious diseases.

Traditional respiratory disease surveillance systems, relying on clinician reporting, laboratory confirmation, and manual data collation, are often constrained by delayed signal detection and fragmented workflows. While event-based surveillance platforms, such as ProMED-mail and HealthMap, have integrated nontraditional signals, the transition from detection

XSL•FO
RenderX

to actionable response remains suboptimal. Emerging artificial intelligence (AI)–driven frameworks, such as BlueDot [1] and the World Health Organization (WHO) Mosaic surveillance framework [2,3], demonstrate that fusing multisource signals with real-time analytics can substantially reduce analytical latency. However, technical acceleration alone cannot bridge the gap to intervention; the "golden window" is frequently missed due to governance bottlenecks, including bureaucratic friction and varying political will [4]. Consequently, there is an evolving need for orchestrated multiagent systems (MAS) that move beyond passive dashboards to provide active, evidence-based decision support, helping to navigate both data complexity and institutional inertia [5].

The "epidemic intelligence trinity," a framework recently consolidated by the authors, comprises three functional stages: (1) surveillance and detection, (2) risk evaluation, and (3) early warning [6]. Surveillance aggregates heterogeneous outbreak indicators, risk evaluation contextualizes them to estimate potential spread and impact, and early warning translates these insights into alerts and intervention triggers. However, in current practice, these stages often remain linear and siloed. Information flows from surveillance to warning but often lacks the feedback loop required to optimize interventions immediately.

This Viewpoint presents a conceptual framework that advances beyond existing models by introducing "decision support" as a fourth, explicit pillar and using AI agents to drive the workflow. Unlike the WHO Mosaic framework, which primarily focuses on the diversity of data sources ("tiles"), or the traditional trinity, which emphasizes situational awareness, our quadripartite framework prioritizes autonomy and action. It proposes moving from passive dashboards to active MAS that not only display data but autonomously reason over it to suggest specific interventions, closing the loop between intelligence and response. Throughout this paper, we illustrate this potential through a recurring case scenario of a hypothetical zoonotic spillover to demonstrate tangible impacts.

AI agents—autonomous computational entities capable of perceiving their environment, reasoning, and taking actions—offer a pathway to unify this process. Moving beyond static rule-based algorithms, modern AI agents use large language models (LLMs) to perform chain-of-thought reasoning and tool orchestration. This viewpoint examines how AI agents can enhance respiratory epidemic intelligence within this expanded framework (Figure 1).

Figure 1. Artificial intelligence (AI) agents–driven integrated closed-loop framework for respiratory infectious disease surveillance, risk evaluation, early warning, and decision support. The framework illustrates an autonomous cycle for respiratory infectious disease management consisting of four interactive stages: (1) surveillance and detection—autonomous collection, integration, and preprocessing of multisource data; (2) risk evaluation—automated prediction and estimation of epidemic risk using machine learning and deep learning; (3) early warning—generation of timely alerts based on risk analysis; and (4) decision support—provision of targeted prevention and control recommendations through a decision database. At the core, the AI agent serves as the "brain," performing reasoning and orchestration to manage data flow and workflow across these stages. The framework supports a hybrid interaction model: it primarily operates in a proactive mode by pushing integrated "surveillance-to-action" reports to decision-makers, while simultaneously supporting reactive queries from expert supervisors through an iterative question and answer process.

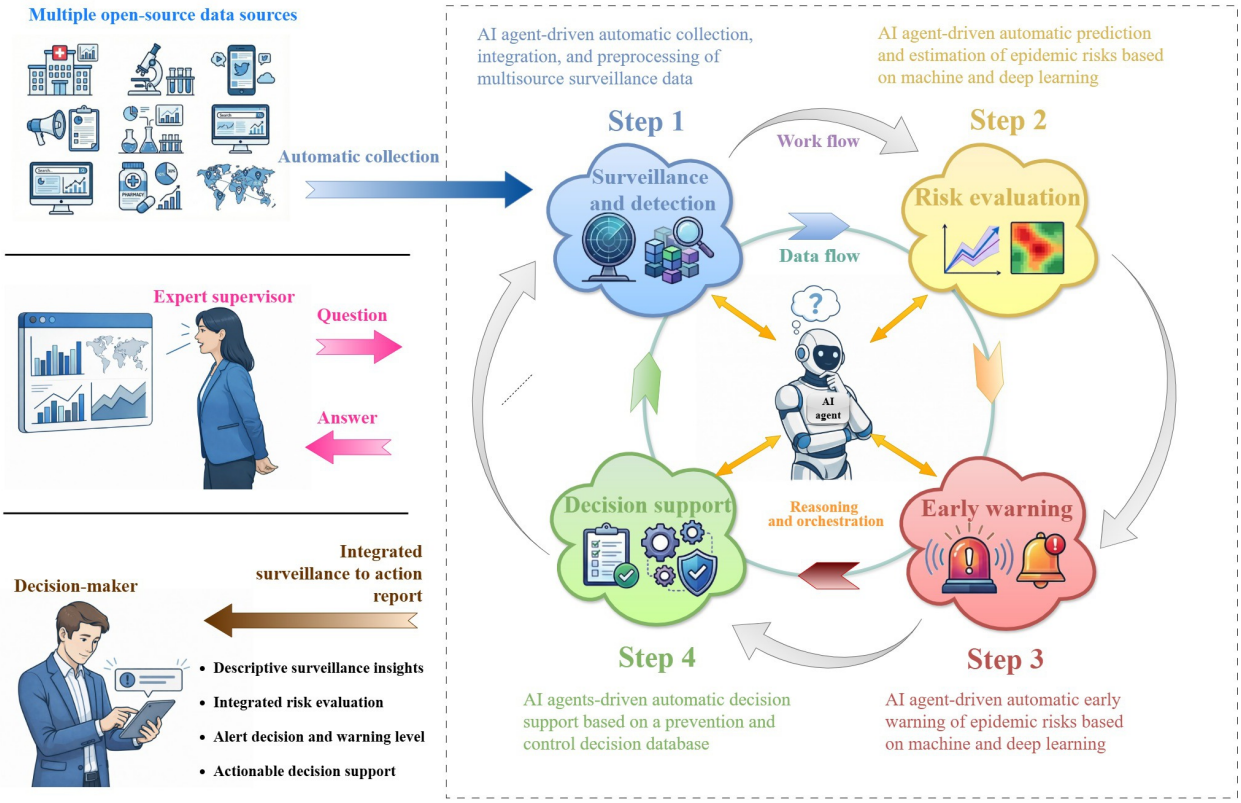

XSL•FO
RenderX

Defining AI Agents in Epidemiology

To understand the potential of this framework, it is necessary to define what constitutes an "AI agent" in this context. Unlike traditional deep learning models that passively process input to produce a prediction (eg, forecasting case counts), an AI agent is an autonomous system capable of three core functions.

Second, reasoning—using LLMs as a cognitive engine, the agent decomposes complex epidemiological problems into steps (chain-of-thought), interprets context, and plans tasks. For example, it can decide that a spike in "fever" search terms requires validation against pharmacy sales data.

Third, action—the agent executes tasks using external tools (application programming interface; API), such as running a transmission simulation model, drafting an alert email to health authorities, or querying a clinical guideline database.

In a MAS, specialized agents collaborate, handing off tasks to one another. This shift allows epidemic intelligence to move from a human-dependent "pull" system to an AI-driven "push" system, where the software proactively identifies threats and proposes solutions.

Surveillance and Detection

Surveillance and detection constitute the first pillar of epidemic intelligence, aiming to identify early signals of outbreaks or abnormal disease activities. Traditional respiratory surveillance predominantly relies on single-source data, such as hospital reports of influenza-like illness (ILI), laboratory confirmation of pathogens, or clinician-reported unusual cases. While effective for confirmed cases, this passive and delayed approach may fail to detect emerging threats, particularly novel or geographically dispersed events.

In contrast, multisource "mosaic" surveillance integrates heterogeneous data streams to create a more sensitive and timely view of epidemic activity. The WHO has advocated a mosaic framework for respiratory virus surveillance, emphasizing that combining multiple "tiles" of data produces a more complete outbreak picture [2].

Given the complexity of diverse data modalities, an MAS approach is superior to a single monolithic agent. Specialized agents can autonomously ingest data from formal channels, such as EHRs, laboratory test results, pharmacy sales, and sentinel clinic reports, and from informal sources, including news media, ProMED-mail bulletins, social media trends, search index [7], participatory symptom reports, animal health bulletins, and environmental or climate datasets. Unconventional indicators, such as increased sales of cough or cold medication, spikes in fever-related online searches, and elevated SARS-CoV-2 RNA levels in wastewater, have been shown to precede clinical surges by days or weeks [8]. AI agents can continuously integrate and analyze such signals, functioning as tireless "digital epidemiologists."

Traditional ILI surveillance has relied heavily on passive reporting, where clinicians manually record and submit case counts after aggregating data from their observations. This approach is inherently subjective, prone to underreporting, and often delayed. Increasingly, health systems are shifting to automated extraction of ILI indicators from EHRs, leveraging structured fields such as chief concerns, physical examination findings, and diagnostic codes. These systems can automatically flag and count encounters matching ILI case definitions in near real time.

Natural language processing further enhances surveillance by enabling AI agents to mine structured and unstructured data for outbreak clues. In addition to EHRs, agents can process informal text sources such as local news and social media [9], extracting disease-relevant facts, classifying signals by syndrome, and distinguishing credible reports from rumors. For instance, an agent might detect an increase in local language posts mentioning "pneumonia," cross-reference this with wastewater testing data showing a coronavirus RNA spike, and confirm concurrent increases in web searches for "fever" and "cough." The system can produce composite alerts with higher confidence and shorter detection lag than manual reporting or single-source monitoring by fusing such diverse signals. A critical advantage of AI agents lies in their ability to filter out noise while retaining sensitivity to subtle signals. In contrast, traditional systems often raise alert thresholds to avoid false alarms, which may result in missed events.

Moreover, AI agents can flag low-intensity anomalies and internally validate them against other data streams or reference knowledge bases. For instance, an agent detecting a sudden rise in emergency visits for respiratory illness could immediately check for similar patterns in nearby regions or seek plausible explanations, such as environmental events. The system would escalate an alert to human experts only when multiple indicators converge, thereby reducing false positives while preserving early warning capacity.

By automating and democratizing multisource data collection and analysis, AI agents extend surveillance to regions with limited public health infrastructure. This capability shortens detection lead times and reduces "blind spots" in low-resource settings, contributing to more equitable and timely global epidemic intelligence.

Case Scenario

An AI agent designed for respiratory threats might continuously scan unstructured data. In a hypothetical scenario of a novel zoonotic spillover, the "surveillance agent" detects a statistically significant cluster of social media posts mentioning "severe cough" and "poultry die-off" in a specific province. Simultaneously, it cross-references this with real-time wastewater data showing an unidentified viral RNA spike in the same locality. Unlike a passive dashboard, the agent autonomously links these disparate signals and flags them for immediate processing. The practical superiority of these orchestrated systems in reducing analytical latency is further evidenced by recent empirical studies, which show significant improvements in lead time and sensitivity compared to traditional methods (Table 1 provides a detailed comparison).

XSL•FO
RenderX

Table . Comparison of traditional surveillance systems vs artificial intelligence (AI)–driven systems.

Study	Conceptual approach and data integration	Performance advantage (timeliness and accuracy)	Role in decision support and public health utility
Du et al [5]	LLM[a]-driven multimodal fusion: uses LLMs to synthesize heterogeneous data streams (news, reports, and clinical data) dynamically rather than static statistical modeling.	Real-time integration: outperforms baseline models in complex dynamics by processing unstructured data in real time.	Operational readiness and resource mobilization: provides rapid situational awareness to justify early resource allocation before clinical confirmation is available.
Idahor et al [10]	Paradigm shift (reactive to proactive): a theoretical framework contrasting rigid, siloed traditional reporting with fluid, big data–driven AI ecosystems.	Lag reduction: traditional systems lag by weeks; AI enables near–real-time monitoring by bypassing bureaucratic reporting delays.	Strategic planning and gap analysis: supports high-level policy shifts from reactive containment to proactive mitigation; identifies "blind spots" in existing infrastructure.
Xie et al [9]	Digital signal detection (social sensing): leverages LLMs to interpret social media sentiment and symptom reporting (Twitter [subsequently rebranded X, X Corp]) as a proxy for early epidemic signals.	Early warning (approximately 1 week lead): achieves approximately 7.63 days lead time over official epidemiological data; strong temporal correlation.	Preclinical alerting and risk communication: triggers public health messaging and targeted testing before hospitals see a surge, allowing for "flattening the curve" interventions.
Zhang et al [11]	Hybrid surveillance (environmental+clinical): integrates wastewater viral load monitoring with traditional clinical case reporting to form a dual verification system.	Leading indicator (0-7 days): wastewater signals precede clinical reports; combined systems show higher robustness than single-source methods.	Noninvasive community risk assessment: enables decision-makers to assess community transmission levels without mass individual testing; guides localized lockdowns or interventions.
Zhang et al [12]	Algorithmic anomaly detection: compares machine learning (supervised or unsupervised) against static statistical thresholds (ADTM)[b] for outbreak detection.	Sensitivity enhancement: warnings issued approximately 5 days earlier. Unsupervised ML[c] achieved 100% sensitivity vs 71% for traditional thresholds.	Automated triage and sensitivity tuning: reduces "missed alarms" for critical threats; allows officials to adjust sensitivity thresholds based on available response capacity.

[a]LLM: large language model.

[b]ADTM: algorithmic detection and threshold monitoring.

[c]ML: machine learning.

Risk Evaluation

Risk evaluation is the second core function of epidemic intelligence in respiratory infectious disease, focusing on interpreting detected signals to determine an outbreak's likelihood, severity, and potential consequences. While detection identifies unusual patterns, risk evaluation contextualizes them, distinguishing events requiring urgent intervention from those of limited concern [3].

Traditional risk assessments rely on expert review of epidemiological, clinical, and contextual information. However, manual processes can be slow, particularly when integrating heterogeneous datasets such as pathogen genomic sequences, case demographics, population immunity, mobility data, and health care system capacity. AI agents can accelerate this process by automating the collation, integration, and preliminary analysis of such datasets in near real time.

Modern AI agents can apply predictive models to estimate outbreak growth rates, reproduction numbers, and geographical spread, incorporating uncertainties from multiple data sources [13]. They can also use machine learning to identify populations or regions at higher risk based on vaccination coverage, comorbidities, demographic profiles, and environmental factors. For example, an agent might assess whether a novel influenza strain detected in wastewater coincides with low immunity levels in nearby populations, indicating a heightened potential for community spread.

Importantly, risk evaluation by AI agents is not limited to epidemiological modeling. AI agents can incorporate operational factors, such as testing capacity, supply chain status for antivirals, or intensive care unit (ICU) bed availability, to provide decision-makers with a comprehensive risk profile. However, it is important to acknowledge that obtaining real-time operational data is challenging due to the fragmented nature of health information systems and inventory databases. These assessments can be updated dynamically as new information becomes available, ensuring that public health authorities receive timely, data-driven insights to guide intervention priorities.

Case Scenario

On receiving the signal from the surveillance agent, the "risk evaluation agent" automatically initiates a probabilistic assessment. It retrieves current population immunity data for influenza subtypes in the affected region, factors in local hospital capacity (ICU bed availability), and runs a transmission simulation. It calculates that the effective reproduction number (Rt) could exceed 2.5 within 7 days if left unchecked, classifying the event as "high risk."

XSL•FO
RenderX

Early Warning

Early warning aims to alert relevant stakeholders quickly enough for interventions to prevent widespread transmission [14,15]. Traditional systems often delay alerts until events are verified through official channels, which can reduce timeliness. AI agents can bridge this gap by issuing graded alerts based on probability thresholds, while continuing to refine assessments as new data become available.

An AI agent's early warning process typically follows three steps: (1) aggregating multisource indicators into a composite risk score; (2) comparing the score against predefined alert thresholds; and (3) communicating the alert in structured, actionable formats. By using anomaly detection and probabilistic forecasting models, agents can provide both short-term outbreak forecasts and medium-term epidemic scenarios [16].

Unlike static rule-based systems, AI agents can adapt their alerting logic over time by incorporating feedback from both human experts and real-world outcomes, thereby creating a continuously improving feedback loop. In practice, this adaptation relies on human-in-the-loop review, where epidemiologists confirm or dismiss alerts, as well as retrospective validation once ground truth data become available. Alerts can be stratified by urgency; for example, an "advisory" for low-probability but high-impact threats and an "action alert" when confidence is high and immediate measures are recommended.

Dissemination channels can also be diversified: AI agents can automatically generate reports for public health dashboards, push notifications for field epidemiologists, and plain language summaries for community stakeholders. This ensures that early warnings reach the right audience in time to act, a critical factor in respiratory pathogen control where delays of even a few days can lead to exponential case growth.

Case Scenario

The "early warning agent" takes the "high risk" classification and drafts a stratified alert. It generates a technical "action alert" for the local Centers for Disease Control, detailing the viral anomalies and projected trajectory. Simultaneously, it prepares a plain language "advisory" draft for community health workers, emphasizing vigilance for patients who have had contact with poultry. These drafts are held for human verification but are prepared instantly, saving hours of manual drafting.

Decision Support

Decision support extends beyond warning, guiding the choice and timing of interventions to minimize health and socioeconomic impacts. Respiratory epidemic control includes recommendations for targeted vaccination, nonpharmaceutical interventions, antiviral distribution, and communication strategies.

AI agents can synthesize epidemiological forecasts, resource availability, and intervention effectiveness data to recommend optimal response strategies [4]. For example, given an emerging RSV outbreak in a region with limited ICU capacity, an agent might recommend targeted nonpharmaceutical interventions for high-risk populations, rapid deployment of prophylactic treatments, and public messaging to reduce exposure risk.

Given the severe safety risks associated with LLM hallucination (fabrication) in medical contexts, decision support modules must implement strict guardrails. Recommendations must be grounded in verified clinical protocols and subjected to retrieval-augmented generation techniques to ensure factual accuracy. Decision support also benefits from scenario simulation. AI agents can run counterfactual models to estimate their potential impact by comparing projected outcomes with and without specific interventions. These simulations can incorporate real-time updates, allowing response plans to adapt dynamically as the situation evolves.

Transparency is essential: decision recommendations must be explainable, clearly indicating underlying assumptions, data sources, and model limitations. This ensures that human decision-makers can assess the validity and acceptability of AI-driven guidance before it is implemented. In this role, AI agents act not as autonomous decision-makers but as analytical partners—augmenting human expertise with rapid, data-rich, and evidence-based insights that can shorten reaction times and improve intervention precision.

Case Scenario

The "decision support agent" queries a database of evidence-based interventions. On the basis of the projected Rt and local resource constraints, it proposes a ranked list of interventions: (1) immediate targeted closure of live poultry markets in the identified district, (2) reallocation of antiviral stockpiles from the neighboring province, and (3) activation of a fever-screening protocol at local transport hubs. These recommendations are presented to human decision-makers with citations to relevant clinical protocols.

Challenges and Practical Implementation Guidance

While promising, real-world deployment demands addressing specific challenges through concrete practical measures.

Data Quality and Interoperability

AI agents depend on high-quality, diverse inputs. Informal sources may be noisy or biased, while data gaps in low-income regions risk reinforcing inequities, necessitating careful curation and cross-verification to prevent the system from optimizing primarily for data-rich regions. A critical risk is "circular reporting," where agents amplify their own signals; therefore, rigorous source provenance tracking is essential to distinguish between organic signals and AI-generated content. To address data fragmentation, such as inconsistent formats in wastewater surveillance reports or a lack of standardized electronic medical record coding, implementation must adhere to open standards, such as HL7 Fast Healthcare Interoperability Resources and WHO protocols, to ensure seamless interoperability [12].

XSL•FO
RenderX

Furthermore, to overcome data sovereignty barriers, privacy-preserving technologies such as federated learning can enable global model training without transferring sensitive local data.

Model Robustness and Validation

Complex models risk false positives or false negatives, potentially triggering inappropriate responses or missing critical threats. Rigorous validation is necessary, measuring lead time, sensitivity, specificity, and forecast accuracy against benchmarks. However, optimizing these metrics requires a delicate balance; an overly sensitive system risks generating "alert fatigue," where a high volume of low-confidence warnings desensitizes human analysts. If operators are inundated with false alarms, they may eventually ignore genuine threats, rendering the surveillance system ineffective. Validating on historical data using internet-connected LLMs carries a risk of "look-ahead bias," as these models may have "seen" the outcomes during training. Therefore, "prospective shadow testing" is recommended as a more rigorous alternative, where the agent runs in parallel to existing systems without releasing active alerts, allowing for safe evaluation of performance. Transparent, adjustable decision thresholds and iterative human-AI feedback loops can enhance reliability. Initial deployment should involve close human oversight, with gradual autonomy as confidence in model performance grows.

Explainability and Trust

High-stakes decisions require explainable outputs. Explainable AI approaches can improve transparency by highlighting relevant data signals and causal rules. LLM agents can generate narrative rationales, but these must reflect actual reasoning to avoid conveying misleading confidence. Trust in AI agents will grow as they demonstrate value under expert supervision; however, high-profile errors could set progress back.

Political and Bureaucratic Friction

Even the most precise AI-driven warnings may encounter friction as they move upward through institutional hierarchies. Decisions such as introducing lockdowns or prioritizing vaccine distribution are not solely forecasting problems; they are high-stakes governance and optimization problems influenced by economic concerns, political incentives, and competing narratives. AI agents must therefore be viewed not as a panacea for delay but as tools to provide clearer evidence-based support under radical uncertainty, potentially reducing, although not eliminating, the friction within political decision processes.

Ethical and Legal Considerations

Privacy, consent, and accountability are critical concerns. AI agents must respect anonymity and avoid individual identification without consent. Legal frameworks for automated public health decisions remain underdeveloped, raising questions of liability in the event of errors. Human-in-the-loop governance must be framed not merely as a method for enhancing trust but as a mandatory legal and safety requirement before any "decision support" outputs are acted upon. Equity demands ensuring benefits for low-resource settings, and WHO-led global AI intelligence services could help provide access for all countries [17]. Furthermore, the economic sustainability of 24/7 digital epidemiologists must be considered, as continuous inference on large models is computationally expensive [18].

Cybersecurity and Misuse

Adversarial attacks, such as false data injection, could disrupt or manipulate surveillance outputs. Systems must authenticate data and resist manipulation. Risks of misuse, including political exploitation for restrictive measures, necessitate algorithmic transparency, independent audits, and adherence to human rights norms.

Workforce, Infrastructure, and Human-AI Interaction

Effective integration requires both skilled personnel and adequate infrastructure. Human epidemiologists should work alongside AI agents (digital epidemiologists) to validate outputs and enrich datasets. However, we must strictly mitigate the risks of "automation bias" and overreliance. If public health officials passively accept algorithmic outputs without scrutiny, it can lead to dangerous "misprioritization," where resources are diverted based on high-confidence but erroneous model predictions. Furthermore, long-term dependence on automated agents risks a "loss of expertise," where the workforce's ability to independently interpret raw epidemiological signals and nuances degrades. Therefore, workflows must be designed to keep humans cognitively engaged—using AI as a "triage tool" rather than a replacement—and regular "manual mode" training simulations should be mandated to maintain core analytical competencies. Low-resource settings may rely on cloud-based services to host AI capabilities. Regular evaluations, including simulated outbreak drills, are necessary to test readiness and refine performance.

Strategic Road Map

To transition this framework from concept to reality, we propose four priority actions: (1) standardization—develop global data dictionaries and API standards for respiratory surveillance to facilitate cross-border AI agent collaboration; (2) validation—establish standard "epidemic benchmark datasets" to rigorously test AI agent performance against historical ground truths; (3) pilot programs—deploy shadow-mode AI agents in sentinel surveillance sites to evaluate real-world utility and workflow integration; and (4) governance frameworks—draft international guidelines defining liability and accountability for AI-supported public health decisions.

Conclusions

The convergence of advanced AI and the urgent need for improved epidemic intelligence offers a transformative opportunity. By extending the traditional trinity into a proactive loop with decision support, AI agents embody the "universal

XSL•FO
RenderX

epidemic intelligence model" [7]. Through autonomous workflows, ranging from detecting weak signals in wastewater to recommending targeted market closures, agents can contextualize respiratory outbreaks days or weeks sooner than traditional methods. Realizing this vision demands investment, transparency, and equity, alongside the practical safeguards and road map outlined earlier. If implemented responsibly, these "force multipliers" can network globally to create a distributed immune system, unifying fragmented data into actionable knowledge—a capability essential in an interconnected world where pathogens ignore borders. The integration of AI agents into epidemic intelligence represents a transformative opportunity to enhance global health security. By addressing the outlined challenges and implementing the proposed road map, we can create a continuously adaptive and globally connected system. We call upon policymakers, researchers, and public health professionals to collaborate in realizing this vision, ensuring a more resilient future against respiratory infectious diseases.

Acknowledgments

The authors thank colleagues and collaborators who provided valuable discussions and conceptual input during the development of this manuscript. The authors also acknowledge the support of institutional research platforms that facilitated literature retrieval and manuscript preparation. No additional individuals or organizations contributed in a manner that meets the criteria for authorship. The authors declare that no generative artificial intelligence tools were used in the writing, data analysis, figure preparation, or editing of this manuscript. All content was produced, reviewed, and approved solely by the authors, who take full responsibility for the accuracy, originality, and integrity of the work.

Funding

This research was funded by the National Key Research and Development Program of China (grant 2023YFC2308701) and Yunnan Fundamental Research Projects (grant 202501AU070024).

Authors' Contributions

LY: conceptualization, framework design, writing—original draft, funding acquisition
LS: conceptualization, visualization, writing—review and editing
XC: conceptualization, literature review, writing—review and editing
JC: literature review, writing—review and editing
MT: conceptualization, critical revision of the manuscript
YN: writing—review and editing
TZ: supervision, project administration, funding acquisition, critical revision of the manuscript
All authors: intellectual development of the manuscript and final approval

Conflicts of Interest

None declared.

References

1. Transforming public health surveillance. BlueDot. URL: https://bluedot.global/transforming-public-health-surveillance/ [accessed 2026-02-22]
2. Mott JA, Bergeri I, Lewis HC, Mounts AW, Briand SC. Facing the future of respiratory virus surveillance: "the mosaic surveillance framework". Influenza Other Respir Viruses 2023 Mar 21;17(3):e13122. [doi: 10.1111/irv.13122] [Medline: 36970570]
3. Morgan OW, Abdelmalik P, Perez-Gutierrez E, et al. How better pandemic and epidemic intelligence will prepare the world for future threats. Nat Med 2022 Aug;28(8):1526-1528. [doi: 10.1038/s41591-022-01900-5] [Medline: 35764683]
4. Koren O, Weidmann NB. Infectious disease outbreaks drive political mistrust. Proc Natl Acad Sci U S A 2025 Jul 22;122(29):e2506093122. [doi: 10.1073/pnas.2506093122] [Medline: 40674429]
5. Du H, Zhao Y, Zhao J, et al. Advancing real-time infectious disease forecasting using large language models. Nat Comput Sci 2025 Jun;5(6):467-480. [doi: 10.1038/s43588-025-00798-6] [Medline: 40481184]
6. Zhang T, Feng L, Lai S, Peng Z, Lan Y, Yang W. Epidemic intelligence trinity: detection, risk assessment, and early warning. Chin Med J (Engl) 2024 Jun 5;137(11):1313-1315. [doi: 10.1097/CM9.0000000000002856] [Medline: 37758650]
7. Yang L, Zhang T, Han X, et al. Influenza epidemic trend surveillance and prediction based on search engine data: deep learning model study. J Med Internet Res 2023 Oct 17;25:e45085. [doi: 10.2196/45085] [Medline: 37847532]
8. Wu F, Zhang J, Xiao A, et al. SARS-CoV-2 titers in wastewater are higher than expected from clinically confirmed cases. mSystems 2020 Jul 21;5(4):e00614-20. [doi: 10.1128/mSystems.00614-20] [Medline: 32694130]

XSL•FO
RenderX

9. Xie J, Zhang Z, Zeng S, et al. Leveraging large language models for infectious disease surveillance-using a web service for monitoring COVID-19 patterns from self-reporting tweets: content analysis. J Med Internet Res 2025 Feb 20;27:e63190. [doi: 10.2196/63190] [Medline: 39977859]
10. Idahor CO, Esomu EJO, Ogbonna N, et al. Infectious disease surveillance in the era of big data and AI: opportunities and pitfalls. Cureus 2025 Oct;17(10):e93929. [doi: 10.7759/cureus.93929] [Medline: 41200593]
11. Zhang X, Lei Z, Wang Q, et al. Comparative performance of wastewater, clinical, and digital surveillance indicators for COVID-19 monitoring in routine practice: retrospective observational study. J Med Internet Res 2025 Nov 6;27:e70232. [doi: 10.2196/70232]
12. Zhang T, Yang L, Fan Z, et al. Comparison between threshold method and artificial intelligence approaches for early warning of respiratory infectious diseases - Weifang City, Shandong Province, China, 2020-2023. China CDC Wkly 2024 Jun 28;6(26):635-641. [doi: 10.46234/ccdcw2024.119] [Medline: 38966311]
13. Reich NG, Brooks LC, Fox SJ, et al. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. Proc Natl Acad Sci U S A 2019 Feb 19;116(8):3146-3154. [doi: 10.1073/pnas.1812594116] [Medline: 30647115]
14. Kostkova P. A roadmap to integrated digital public health surveillance: the vision and the challenges. Presented at: Proceedings of the 22nd International Conference on World Wide Web; May 13-17, 2017 URL: https://dl.acm.org/doi/10.1145/2487788.2488024 [accessed 2026-02-22]
15. Allam Z. The rise of machine intelligence in the COVID-19 pandemic and its impact on health policy. In: Allam Z, editor. Surveying the Covid-19 Pandemic and Its Implications: Urban Health, Data Technology and Political Economy: Elsevier; 2020:89-96. [doi: 10.1016/B978-0-12-824313-8.00006-1]
16. Smolinski MS, Crawley AW, Olsen JM, Jayaraman T, Libel M. Participatory disease surveillance: engaging communities directly in reporting, monitoring, and responding to health threats. JMIR Public Health Surveill 2017 Oct 11;3(4):e62. [doi: 10.2196/publichealth.7540] [Medline: 29021131]
17. Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models: World Health Organization; 2025.
18. Nagarajan R, Kondo M, Salas F, et al. Economics and equity of large language models: health care perspective. J Med Internet Res 2024 Nov 14;26:e64226. [doi: 10.2196/64226] [Medline: 39541580]

Abbreviations

AI: artificial intelligence
API: application programming interface
EHR: electronic health record
ICU: intensive care unit
ILI: influenza-like illness
LLM: large language model
RSV: respiratory syncytial virus
WHO: World Health Organization

Edited by A Coristine; submitted 01.Nov.2025; peer-reviewed by K Adegoke, M Chakit, Z Liu; revised version received 27.Jan.2026; accepted 29.Jan.2026; published 09.Mar.2026.

Please cite as:
Yang L, Shan L, Cao X, Cui J, Tong M, Niu Y, Zhang T
AI Agents and Epidemic Intelligence on Respiratory Infectious Diseases: Toward a Conceptual Framework Integrating Decision Support
J Med Internet Res 2026;28:e86936
URL: https://www.jmir.org/2026/1/e86936
doi:10.2196/86936

© Liuyang Yang, Liyu Shan, Xiaolin Cao, Jinzhao Cui, Michael Tong, Yan Niu, Ting Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Deep Research Agents: Major Breakthrough or Incremental Progress for Medical AI?

Matthew Yu Heng Wong[1], BA, MB, BChir; Ariel Yuhan Ong[2,3], MBChB, FRCOphth; David A Merle[2,3], MD, FEBO; Pearse A Keane[2,3], MD, FRCOphth

[1]School of Clinical Medicine, University of Cambridge, Addenbrooke's Hospital, Hills Rd, Cambridge, United Kingdom
[2]NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust, London, United Kingdom
[3]Institute of Ophthalmology, University College London, London, United Kingdom

Corresponding Author:
Matthew Yu Heng Wong, BA, MB, BChir
School of Clinical Medicine, University of Cambridge, Addenbrooke's Hospital, Hills Rd, Cambridge, United Kingdom

Abstract

Deep research agents are autonomous large language model–based systems capable of iterative web search, retrieval, and synthesis. They are increasingly positioned as the next major leap in medical artificial intelligence. In this viewpoint, we argue that while these agents mark progress in information access and workflow automation, they represent an incremental evolution rather than a paradigm shift. We review current applications of deep research agents in biomedical scenarios, including literature review generation, clinical evidence synthesis, guideline comparison, and patient education. Across these early use cases, the tools demonstrate the ability to rapidly gather and structure up-to-date information, often producing outputs that appear comprehensive and well-referenced. However, these strengths coexist with unresolved and clinically significant limitations. Citation fidelity remains inconsistent across models, with subtle misinterpretations or unreliable references still common. Their retrieval processes and evidence-ranking mechanisms remain opaque, raising concerns about reproducibility and hidden biases. Moreover, overreliance on artificial intelligence–generated syntheses risks eroding clinicians' critical appraisal skills and may introduce automation bias at a time when medicine increasingly requires deeper scrutiny of information sources. Safety constraints are also less predictable within multistep research pipelines, increasing the risk of harmful or inappropriate outputs. Finally, current evidence is largely limited to proof-of-concept evaluations, with little evidence from real-life clinical deployment. We contend that deep research agents should be embraced as assistive research tools rather than pseudoexperts. Their value lies in accelerating information gathering, not replacing rigorous human judgment. Realizing their potential will require transparent retrieval architectures, robust benchmarking, and explicit educational integration to preserve clinicians' evaluative reasoning. Used judiciously, these systems could enrich medical research and practice; used uncritically, they risk amplifying errors at scale.
We contend that deep research agents should be embraced as assistive research tools rather than pseudoexperts. Their value lies in accelerating information gathering, not replacing rigorous human judgment. Realizing their potential will require transparent retrieval architectures, robust benchmarking, and explicit educational integration to preserve clinicians' evaluative reasoning. Used judiciously, these systems could enrich medical research and practice; used uncritically, they risk amplifying errors at scale.

(J Med Internet Res 2026;28:e88195) doi:10.2196/88195

KEYWORDS

artificial intelligence; AI; medical research; scientific writing; large language models; LLMs

Introduction

Large language models (LLMs) have now become extensively tested and used for a range of medical tasks and capabilities [1,2]. However, current LLMs are limited by issues such as hallucinations, opaque reasoning processes, and reliance on prior static training data. Retrieval-augmented generation (RAG) aims to mitigate some of these limitations by allowing LLMs to draw from external information sources (eg, specialized medical textbooks and journals) [3]. In late 2024, another advancement in the development of LLMs was introduced in the form of "deep research agents"—autonomous research assistants that combine LLMs with real-time internet search and citation capabilities [4]. These agents leverage the capabilities of RAG by replacing this external data source with up-to-date, live internet access [5].

Beyond simple responses to standard chat-based prompting, these deep research agents can autonomously perform web searches, retrieve, cross-reference, and cite literature, positioning themselves as potential replacements for a myriad of human-led tasks. OpenAI's Deep Research (launched in February 2025) was designed to rapidly access and analyze vast datasets of literature, synthesize their findings, and then generate tailored written outputs [6]. These tools promise to bridge the gap

XSL•FO
RenderX

between an ever-growing medical knowledge base and the limited capacity clinicians have to digest it. With minimal user input, they can autonomously generate comprehensive, well-structured, and referenced reviews of prespecified topics. In this viewpoint, we use "deep research agents" (lowercase) to denote the general class of agentic, multistep, research-oriented LLM systems. The capitalized term "Deep Research" is reserved exclusively for specific commercial implementations (eg, OpenAI's Deep Research) and is used only where product-level distinctions are necessary.

Deep research agents are a relatively recent development, and their application in biomedical contexts remains only partially investigated. In this study, we review how deep research agents have been explored in early use cases in medicine, ranging from literature reviews and medical education to clinical queries, highlighting their anticipated potential. We also identify areas where these agents are limited and require further study and careful validation as usage and adoption grow. Overall, we argue that while these agents mark progress in information access and workflow automation, they represent an incremental evolution in medical artificial intelligence (AI) rather than a paradigm shift.

Deep Research Agents Versus Conventional LLM Chatbots for Medicine

Deep research agents represent a new evolution of conventional chatbot-style LLMs, which integrate RAG workflows and more complex attention mechanisms. Unlike standard LLM chatbots, which typically rely on the model's internal weights and a single input context window, deep research systems typically embed a multistep reasoning pipeline: they first query external knowledge sources, retrieve relevant documents, and then feed those passages into the generator, which uses cross-attention to integrate information across multiple sources before producing a response (Figure 1). Importantly, unlike single-shot retrieval in traditional RAG pipelines, web exploration and source retrieval are iterative and agent-driven, enabling deeper coverage of sparse or scattered evidence [7].

Figure 1. Deep research agents autonomously guide the entire workflow: starting from user input, optionally clarifying intent and planning the search, iteratively using tools for offline data retrieval (databases, vector stores), online search (APIs, browsers), and higher-level functions such as analytics or multimodal content creation (text, image), before assembling the findings into a structured report [8]. API: application programming interface; LLM: large language model.

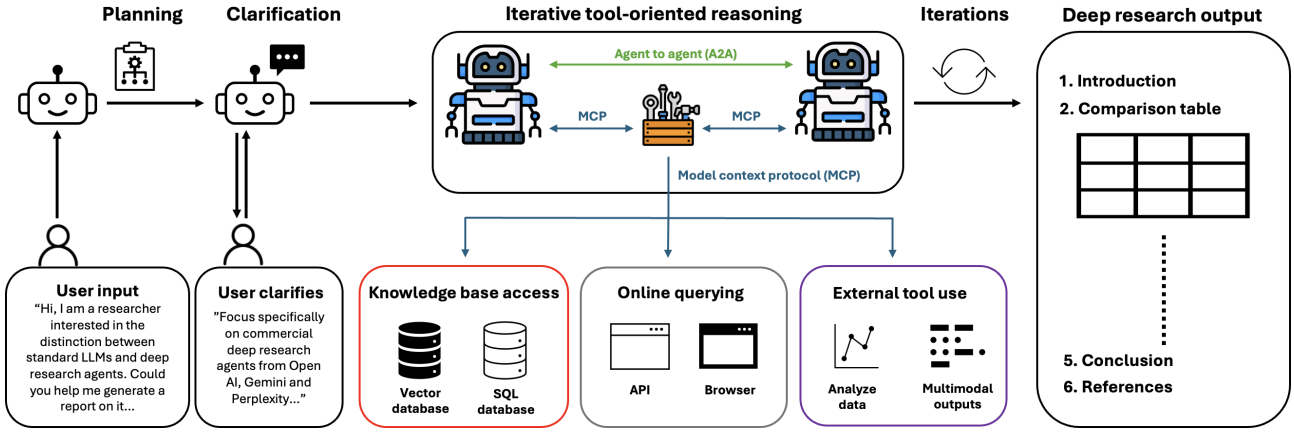

Deep research agents extend RAG by combining retrieval with real-time tool use and iterative reasoning, allowing the system not only to fetch relevant documents but also to plan, refine, and synthesize information as part of a single workflow. Dynamically engaging with external tools and managing multistage research tasks in real time ostensibly allows these agents to provide greater autonomy, context-awareness, and accuracy in their outputs (Table 1) [8].

XSL•FO
RenderX

Table . Architectural and functional comparison of deep research agents vs standard retrieval-augmented generation chatbots.

Dimension	Deep research agents (eg, OpenAI Deep Research, Gemini Deep Research, Perplexity Pro)	Standard RAG[a] or web-enabled Chatbots (single-shot retrieval + generation; optional browsing)
Core behavior	Multistep autonomous pipeline: *plan* → *search* → *retrieve* → *read* → *synthesize* → *cite*. Persistent reasoning and external grounding.	Nonagentic workflow: answer generation supported by one-shot retrieval or ad hoc browsing; limited planning and iteration.
External information integration	Integrated web and file retrieval; dynamic multisource querying with internal memory.	Single-pass retrieval (*top-k* docs); weaker systematic coverage and cross-source reconciliation.
Processing time or run length	Long-running sessions (≈5 - 30 min typical) processing many documents and verifying citations.	Short latency (seconds to a few minutes) typically fewer tool calls and shallower document review.
Citation handling or attribution	Structured outputs with traceable references and in-line links; still requires human verification for claim-source alignment.	Can provide citations when configured or prompted. May show weaker claim-source alignment.
RAG	Full multihop RAG with iterative query refinement, retrieval, cross-checking, and synthesis.	Single-hop RAG; limited multihop behavior unless manually orchestrated.
Control and reasoning architecture	Agentic controller layer atop the LLM[b] for planning, tool selection, and multisource integration.	Fixed LLM + retriever. Reasoning and evidence integration occur implicitly within a single LLM pass.
Output style	Structured research reports with citations and sectioning (eg, "Findings," "Limitations"), can be tailored with prompt engineering.	Conversational prose or short summaries without specific prompting.
Limitations or risks	Improved grounding but still susceptible to retrieval errors, misweighting of evidence, opaque rankings of sources, and hallucinations.	Similar risks (retrieval noise, cherry-picked snippets, hallucinations) and weaker systematic coverage. Higher risk of hallucinated references.
Potential use cases	Systematic reviews and evidence synthesis, where provenance and depth matter, and slow generation is acceptable.	Point queries, quick updates, short summaries with citations. For tasks where speed matters.

[a]RAG: retrieval-augmented generation.

[b]LLM: large language model.

Compared with their early 2025 iterations, current deep research agents have shifted from text-only, single-model research bots into multimodal, tool-integrated, multimodel research systems. This evolution includes enhanced perceptual capabilities, allowing agents to interpret not only text but also visual elements such as figures, images, and complex documents, alongside expanded access to public sources (eg, PDFs from the web) through external connectors. In parallel, successive model upgrades have incorporated advanced self-reflection mechanisms, increases in scalability, and longer context windows. For example, newer iterations of Gemini's Deep Research employ "Flash Thinking" to support more explicit planning and iterative self-critique during synthesis [9].

From a clinical perspective, this architecture could one day transform how evidence is synthesized and applied in medicine and health care. By autonomously retrieving and cross-referencing peer-reviewed studies, clinical guidelines, and trial reports, deep research agents could assist in rapid reviews**,** clinical decision support, and precision medicine literature surveillance (Table 2). For instance, a clinician could enter a query to "Identify emerging 2024 - 2025 biomarkers predictive of immunotherapy response across cancer types" and receive a sourced, citation-backed synthesis reflecting the latest evidence—something standard LLM chatbots, limited to static parametric memory, cannot reliably provide. This output strongly hints at the future of AI-assisted knowledge retrieval. Moreover, the use of cross-attention between retrieved documents allows the model to contextualize and reconcile conflicting studies**,** potentially supporting meta-analytic or guideline-development workflows [8]. We briefly tested this workflow and evaluated the resulting retrieval and output of the model (Figures 2 and 3).

XSL•FO
RenderX

Table . Illustrative deep research queries showcasing the breadth of use cases for clinicians and academics.

Category	Example query	Potential benefits
Clinical updates	"What emerging biomarkers beyond programmed death-ligand 1 predict response to immunotherapy, based on recent clinical studies?"	Provides clinicians with current knowledge without manually searching journals and reviewing multiple articles.
Guideline comparison	"How do European Society for Medical Oncology and American Society of Clinical Oncology guidelines address immunotherapy response in lung cancer?"	Rapid evidence synthesis across guideline bodies.
Basic science research	"What biological mechanisms underlie emerging biomarkers predictive of immunotherapy response?"	Elucidates underlying molecular mechanisms and pathways.
Translational research	"Which emerging biomarkers are sufficiently validated and assay-ready to inform immunotherapy patient selection?"	Highlights cutting-edge translational trends.
Patient communication	"Explain in simple terms to a patient how their biomarkers may help predict whether immunotherapy will work."	Generates layperson-friendly patient information.
Academic writing	"Draft an introduction for a review summarizing emerging biomarkers predictive of immunotherapy response."	Saves time on first-draft writing.
Whimsical or exploratory	"What will clinicians recognize as the moment when immunotherapy response first became meaningfully predictable?"	Engages creativity and unlocks novel perspectives.

Figure 2. An example of how a deep research agent (OpenAI's Deep Research) engages in an iterative dialog with the user to clarify intent before conducting literature searches. By confirming details such as cancer type and biomarker scope, the agent supports a more focused and structured synthesis—a workflow that extends beyond typical single-turn large language model interactions.

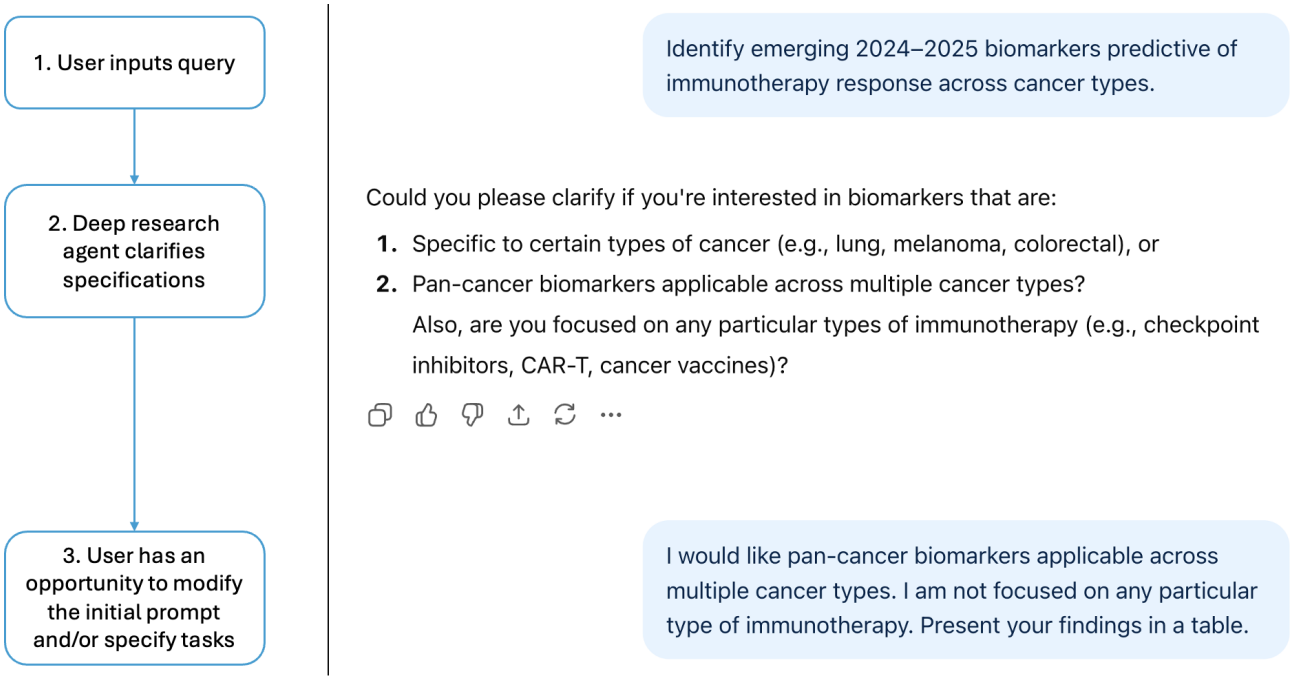

XSL•FO
RenderX

Figure 3. Example of a structured evidence table generated by a deep research agent in response to the earlier query, "identify emerging 2024 - 2025 biomarkers predictive of immunotherapy response across cancer types." The agent autonomously retrieved, synthesized, and formatted recent literature into a reference-supported summary in less than 8 minutes. In this instance, all Deep Research–generated content and citations were independently checked against the original publications and were found to accurately reflect the source literature, with no hallucinated references. This figure, therefore, illustrates a successful deep research workflow. However, polished outputs of this kind should not be assumed to be uniform across tasks or systems. The superscript annotations highlight illustrative areas where known risks may arise: †potential inconsistencies in citation fidelity (eg, incorrect or semantically mismatched source attribution) and ‡possible overstatement or imprecise reporting of quantitative effects. These annotations are hypothetical and do not imply errors in the example shown but are included to emphasize the need for verification despite apparent surface reliability. Citations in this figure are shown in Multimedia Appendix 1. AUC: area under the curve; CTRAM: class I–restricted T-cell–associated molecule; ICI: immune checkpoint inhibitor; MSI: microsatellite instability; OS: overall survival; PD-L1: programmed death-ligand 1; PFS: progression-free survival; TMB: tumor mutational burden.

Biomarker	Type	Immune Mechanism/Pathway	Cancer Types Studied	Key Findings	Source
CRTAM expression	Transcriptomic (T cell/NK cell marker)	Activation/differentiation of cytotoxic and regulatory T cells (CRTAM is a transmembrane protein on T/NK cells)	Pan-cancer (diverse solid tumors; highest expression in melanoma and small intestine)	High tumor CRTAM transcript levels were linked to significantly longer PFS and OS on ICI therapy. In a 514-patient pan-cancer cohort, "high" CRTAM expression (top ≥25%) predicted better response to checkpoint blockade (e.g. HR≈0.60 for PFS, HR≈0.39 for OS)‡ even after adjusting for PD-L1, MSI, and TMB[A]†. This association was not seen in patients who did not receive ICI, indicating CRTAM's predictive value.	[A]
Third Dimensional Stromal (TDS) score	Transcriptomic (stromal gene signature)	Stromal-rich tumor ecosystem (distinct from immune-infiltrated "hot" tumors)	Pan-cancer (multiple solid tumors including lung, urinary, skin, kidney, etc.)	A novel 66-gene "TDS" signature (enriched in stromal cells) was identified. **High TDS** predicted *poor* ICI outcomes: meta-analysis showed OS HR≈1.56 and PFS HR≈1.65 (p<0.0001)‡ for high-TDS tumors[A]†. High-TDS tumors lost the usual benefit of ICI versus chemotherapy, underscoring TDS as a negative predictive biomarker.	[A]
Immunotherapy-Responsive Ecotype Signature (IRE.Sig)	Transcriptomic (single-cell-derived)	Pro-immunogenic tumor "ecosystem" (cell-state composition)	Pan-cancer (validated across multiple cancer types and ICI-treated cohorts)	IRE.Sig is a gene expression signature reflecting a favorable immune "ecotype." In analysis of 34 scRNA-seq datasets and bulk pan-cancer cohorts (11,001 patients, 32 cancer types), IRE.Sig robustly predicted response to ICI. Machine-learning models using IRE.Sig achieved AUC ~0.72 in validation cohorts[B]†, indicating it generalizes across cancers.	[B]

Importantly, although multihop retrieval and iterative self-correction introduce greater autonomy and coverage, the underlying architecture of these agents remains conceptually continuous with established RAG techniques. Their principal innovation lies in the orchestration layer that coordinates planning, retrieval, and tool use, rather than in a fundamentally new reasoning paradigm (eg, no symbolic reasoning or formal logic introduced). These agents may therefore represent a stepwise functional improvement in autonomy but are only an incremental evolution in architecture.

As such, deep research agents inherit many of the same fundamental limitations as prior RAG-based systems. Current implementations remain limited by their dependence on open-access data, lack of transparent ranking mechanisms, and inability to independently verify information. Consequently, these systems continue to operate as unpredictable black boxes, generating probabilistic responses with limited reasoning transparency. As these constraints are progressively addressed, deep research tools may increasingly contribute to the early stages of systematic reviews and evidence-based clinical support.

Deep Research Agents for Literature Reviews and Knowledge Synthesis

One potential use of deep research agents is drafting medical literature reviews or summaries of evidence. Ong et al [5] evaluated several such tools (Gemini 1.5 Pro Deep Research, Perplexity's Deep Research, and OpenAI's Deep Research) by tasking them with writing a literature review on "oculomics" (ocular biomarkers of systemic health). They found that the OpenAI agent produced the most compelling draft—a logically structured, readable, and informative narrative that nonexpert readers could hardly distinguish from a human-written review. In another study, Jesudason et al [10] reported that OpenAI's Deep Research could generate a multipage referenced text on complex medical topics in minutes, including controversial clinical questions such as antithrombotic therapy for carotid dissection. Meanwhile, Wang et al's [11] DeepEvidence allows deep research to be performed across various biomedical knowledge graphs, achieving substantial performance gains in evidence synthesis across drug discovery and clinical trial development. Such comprehensive reports, complete with citations, suggest deep research agents could save researchers substantial time on initial literature reviews.

Another study benchmarked multiple agents for systematic review title and abstract screening in radiology. They found traditional LLMs showed variable sensitivity but high specificity in classifying abstracts, while deep research agents (OpenAI and Google's Gemini) could autonomously retrieve literature with high precision but moderate recall [12]. Another recent evaluation by Loke et al [13] benchmarked deep research agents (OpenAI, Perplexity, and Gemini) and autonomous agentic systems like Manus AI against a published systematic review and meta-analysis on myopia risk. Although this work shares

XSL•FO
RenderX

partial authorship with the present manuscript, it suggests systems such as OpenAI and Perplexity Deep Research can identify studies meeting inclusion criteria and perform generally accurate qualitative data extraction, indicating utility for certain stages of the review process.

However, these studies also reveal important caveats. Citation accuracy arises as a common concern. Ong et al [5] reported multiple instances of hallucinations or erroneous citations, including references to nonexistent articles and incorrect attribution of fictitious articles to real researchers. In independent tests, Keplinger et al [14] tasked 5 deep research agent models with generating a dermatology literature review. They found OpenAI's Deep Research performed best in reference accuracy—approximately 95% of its citations were identifiable, with approximately 70% completely correct. In contrast, Google Gemini 2.5 Pro Deep Research and Perplexity AI Deep Research frequently produced fabricated citations: roughly 47% to 50% of references from those models had fake authors or titles. Another study similarly found OpenAI's Deep Research and xAI's Grok 3 achieved citation precision exceeding 90%, surpassing Gemini and Perplexity but still clearly lagging behind human reviewers [15]. Together, these findings show that while deep research agents can rapidly synthesize literature, citation reliability remains inconsistent across models and still falls short of human standards.

Citation–claim alignment is another notable problem across the board. Keplinger et al [14] found that over 50% of OpenAI's Deep Research's cited statements contained at least one subtle inaccuracy or misrepresentation of the source. The AI often misinterpreted study results, cited out of context, or hallucinated details not in the cited papers. These citation errors are difficult to detect without thorough cross-checking by an expert, raising concerns that an AI-generated review could unwittingly propagate misinformation. Ong et al [5] similarly observed factual errors in the AI-generated oculomics review that appeared innocuous to nonexperts—for example, claims about nonexistent trials or incorrect statements about regulatory approvals. Meanwhile, Loke et al [13] found that none of the evaluated deep research agents could conduct a genuine meta-analysis or generate accurate, publication-quality figures and tables. For instance, frequent omissions and hallucinations were observed when agents were asked to construct summary tables of included studies.

Besides accuracy, bias represents a key concern. Deep research agents often rely on RAG-modeled retrieval to select what they consider the "best" or most relevant evidence, using vector similarity search, ranking algorithms, and other techniques. However, the opacity of what the model deems to be "optimal" or "most relevant" presents a real challenge to the generation of verifiable, reproducible knowledge in clinical and translational research [16]. Given that English-language publications from Western nations dominate the internet, deep research tools may be more likely to draw from these, which risks reinforcing global inequalities in biomedical research. Without conscious efforts to diversify their input sources, these agents risk perpetuating existing publication biases and overlooking unconventional or minority perspectives in research [17].

In summary, deep research agents can dramatically speed up information gathering or evidence summarizing in medicine, producing text that is well organized and richly referenced. However, the consensus among all studies is that human oversight remains crucial for now.

Deep Research Agents for Patient Education and Clinical Q&A

Another emerging medical application of deep research agents is in patient education and answering clinical questions. By integrating up-to-date information from the web, these agents can potentially serve as advanced, real-time medical chatbots. Gültekin et al [18] compared OpenAI's Deep Research with DeepSeek R1 (a contemporary model with a similar "deep think" web-search capability) for answering common patient questions about anterior cruciate ligament surgery. Notably, OpenAI's Deep Research delivered more comprehensive answers, covering questions in greater depth (completeness score 4.0 vs 3.2 for DeepSeek). DeepSeek's answers, however, were clearer and more accessible to laypeople, written at a significantly lower reading level (8.9 vs 14.2). The authors noted that excessive detail may overwhelm patients, while oversimplification risks omitting critical information. This suggests integrating the strengths of both models (the thoroughness of ChatGPT and the plain-language clarity of DeepSeek) could be ideal for patient education materials. Nevertheless, readability need not be an intrinsic model limitation: deep research systems can ultimately be prompted to adjust language complexity, allowing clinicians to tailor explanations to patient literacy levels.

A similar multimodel evaluation by Yetkin et al [19] queried 11 AI chatbots—including GPT-4 Deep Research, Perplexity's Deep Research mode, and others—with common patient questions about sarcoidosis. The models that integrated RAG (ChatGPT-4.0 Deep Research, Perplexity, and xAI's Grok Deep Search) achieved the highest scores on validated metrics such as DISCERN (measuring the quality of health publications) and Patient Education Materials Assessment Tool (PEMAT; patient education and understandability). These deep research–enabled models significantly outperformed standard chatbots that lacked web access. However, the improved comprehensiveness came again at the expense of readability. The top-performing models delivered information at a very high reading level—Flesch-Kincaid grade more than 16 (college-level difficulty). This reiterates the importance of thoughtful prompting and clinician oversight to align outputs with patient health literacy.

Taken together, these studies show that deep research agents can dramatically accelerate the generation of high-quality patient education content, but their output may not yet be patient-accessible without clinical mediation. In practice, a clinician could use deep research agents to quickly gather accurate medical information for a patient, then refine the output into more digestible terms. Moreover, patients have become increasingly AI literate and can adopt simple prompting strategies (eg, "rewrite this explanation for an eighth-grade reading level" or "explain this using analogies and avoid medical jargon") to request clearer, lay-friendly reformulations of

XSL•FO
RenderX

AI-generated content. However, while prompting can improve readability, it cannot guarantee accuracy. Patients need to be made aware of the potential pitfalls of AI-generated medical information, including the importance of corroborating outputs against trusted sources (eg, National Health Service patient information resources). As time progresses, deep research agents could gradually shift the clinician's role from information provider to information steward: verifying accuracy, simplifying complexity, and ensuring alignment with a patient's health literacy level and emotional needs.

Emerging Medical-Specific Deep Research Agents and Benchmarks

As deep research agents have matured, evaluation has shifted from general retrieval toward complex, multistep reasoning in medical contexts. Medically tailored benchmarks have been introduced alongside the development of domain-specific deep research agents.

For example, MedBrowseComp was introduced in 2025 as the first benchmark for autonomous, multihop biomedical research [20]. It featured over 1000 physician-curated questions that simulate realistic clinical research tasks. Importantly, MedBrowseComp evaluated agents not solely on output accuracy but on their ability to locate evidence across multiple authoritative sources and reconcile fragmented biomedical information. The benchmark revealed that leading proprietary deep research agents, including Open AI's o3, Gemini 2.5 Pro, and Perplexity, could answer only around a quarter of these multistep questions correctly.

Meanwhile, domain-specialized agents such as IQVIA's Med-R1 (an 8B parameter medical reasoning model with an agentic framework) and the open-source MedResearcher-R1 (32B) emerged. These incorporated dense medical knowledge and custom retrieval tools, enabling them to match and even outperform much larger general models on certain biomedical tasks. On MedBrowseComp, MedResearcher-R1 (32B) achieved a new state-of-the-art 27.5% accuracy, surpassing OpenAI's o3 Deep Research (25.5%) and Google's Gemini 2.5 Pro (~25%) [21]. Likewise, IQVIA's Med-R1 Deep Research Agent achieved comparable results to OpenAI's o3 and Google's Gemini 2.5 Pro despite using only an 8B parameter model [22]. These medical-specific systems demonstrate more advanced multihop clinical reasoning and synthesis across specialized databases (eg, PubMed, clinical trial registries, and Food and Drug Administration data). By leveraging domain-specific retrieval engines and expert knowledge, they may be able to synthesize complex chains of biomedical facts that generalist agents may miss. In addition, these specialized deep research agents attain their results with relatively compact model sizes (8B for IQVIA or 32B parameters for MedResearcher-R1). This efficiency suggests promising potential for deployment in computationally constrained health care settings, where smaller, domain-focused AI systems may be integrated.

Current limitations underscore that adapting these techniques to medical contexts is still in its early stages. Tasks requiring beyond 4 to 5 reasoning hops remain largely unsolved, as even state-of-the-art agents saw a sharp drop in accuracy on 5-hop questions. Their performance profile on these benchmarks reinforces that such systems should not yet be considered reliable clinical assistants, but rather exploratory tools requiring substantial human oversight. Moreover, the scope of the MedBrowseComp benchmark is itself constrained: it uses automated answer judging (without full clinical expert review) and focuses on hematology and oncology datasets alone, omitting certain subdomains such as advanced genomics or multimodal cases.

Benefits and Limitations of Current Deep Research Agents

Current studies illustrate an emerging picture of the potential roles of deep research agents in biomedical settings. In summary, these agents are being explored as advanced research assistants and information mediators: to draft literature reviews, summarize evidence on controversies, assist in writing academic papers, educate patients with up-to-date information, and even support clinical decision queries. The common thread is their ability to autonomously comb through and synthesize vast text corpora to produce a structured, referenced answer in a fraction of the time a human might take.

This has already led some to herald a new era of AI-augmented medical research. Editorials have described "AI's deep research revolution" and the "next evolutionary leap" [23], noting that current iterations of LLMs can speed up the summarization of literature and even facilitate comparing study results—such that "for text-based research, such as systematic reviews or meta-analyzes, AI may replace them efficiently" [15]. Scientific publishing may be entering a "post-LLM" era, where AI agents equipped with deep reasoning and retrieval capabilities are reshaping how literature is synthesized and reviewed.

On the other hand, virtually all medical evaluations of these tools so far have included strong warnings (Figure 4). A recurring concern is the integrity of citations and content. While the latest models have largely solved the problem of blatantly fabricated references that plagued earlier LLMs, subtler issues persist—slight errors in bibliographic details or correct references attached to unsupported claims. These are not always obvious, especially to nonexperts or when an AI output looks polished. If clinicians or researchers take AI-generated text at face value, there is a risk of false information seeping into publications or clinical guidance.

XSL•FO
RenderX

Figure 4. Limitations for deep research agents include system-level reliability and safety, human and societal risks, and the lack of real-world clinical validation. AI: artificial intelligence.

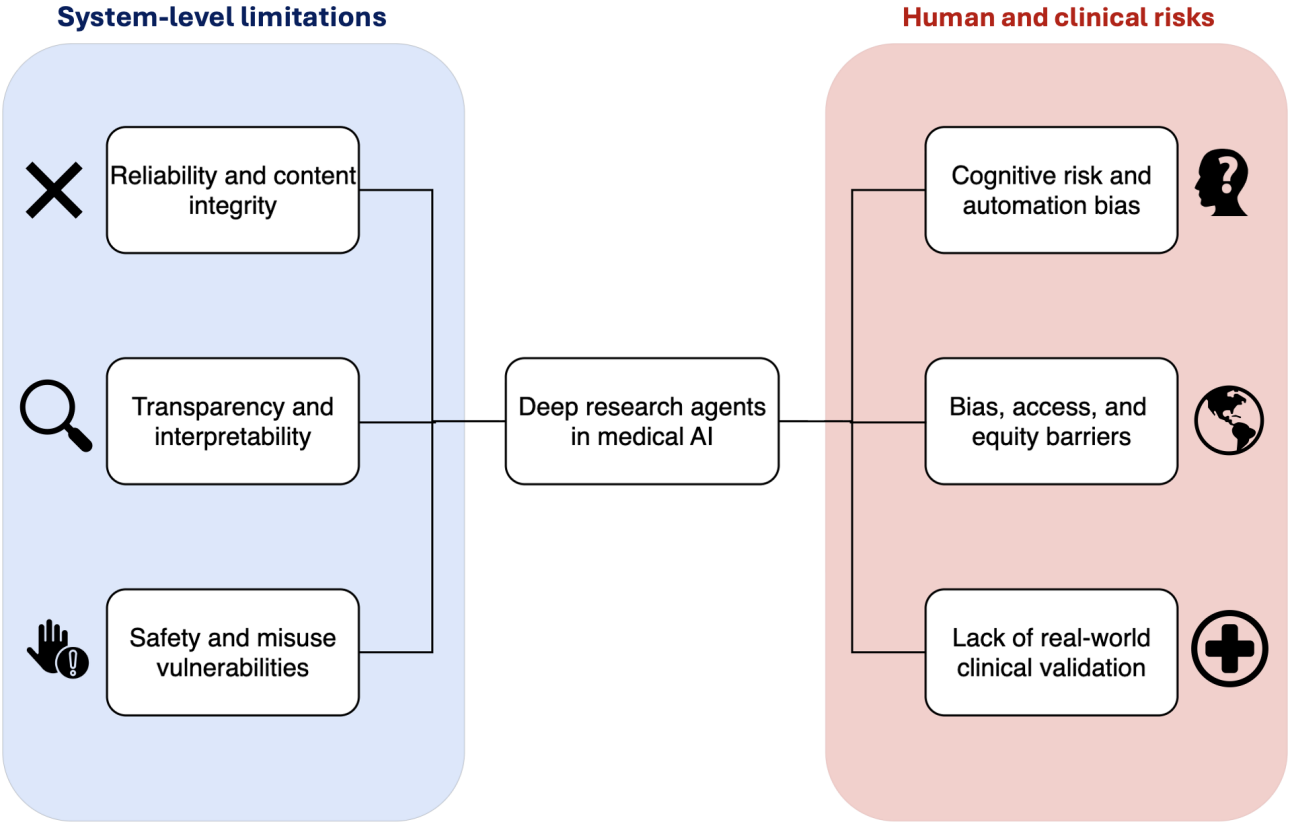

Another concern is the potential for overreliance. Jesudason et al [10] urge that medical trainees should avoid leaning too heavily on AI-prepared syntheses, as they may lose the ability to critically appraise evidence themselves. Cognitive off-loading, where learners delegate lower-level mental tasks to external systems to free up working memory, can support higher-order reasoning and enhance learning if used judiciously [24]. By contrast, automation bias—an overreliance on automated systems for important tasks—arises when users defer to system outputs even when independent judgment is required. In medical training, critical appraisal of studies, evidence weighting, data interpretation, and clinical reasoning represent non-negotiable competencies that should not be off-loaded. Excessive off-loading of these tasks risks fostering not only automation bias (overreliance on automated systems) but also deskilling (loss of previously acquired competencies), never-skilling (failure to develop essential expertise), and mis-skilling (reinforcement of incorrect reasoning due to AI errors or bias) [25]. Over time, excessive reliance on predigested syntheses may erode clinicians' capacity for deep critical thinking, analogous to early concerns about the erosion of manual skills with surgical robotics. In the same vein, Keplinger et al [14] expressed concerns that outsourcing the initial draft of a review to AI could distort authors' own mental map of the field and undermine their engagement with the literature, ultimately harming scientific creativity.

Another key but often overlooked aspect is safety. Recent work has raised broader concerns that increasing autonomy and multistep planning in AI systems introduces qualitatively new safety challenges, particularly around oversight and loss of control [26]. In addition, it has been suggested that deep research agents may circumvent traditional safety filters under certain conditions. Unlike standalone LLMs, which are engineered to refuse unsafe prompts, deep research agents' multistep reasoning pipelines may potentially weaken those refusal mechanisms. Chen et al [27] note that deep research agents may process a broader range of query types than standard chat interfaces because the research pipeline is treated as an information-gathering task rather than a direct response to a prompt. This could inadvertently allow malicious or high-risk instructions—such as instructions related to biological weapons development, hacking, or privacy breaches—to be pursued under the guise of research. This does not mean safety filters are obsolete or permanently bypassed in research-oriented systems. Rather, they highlight a current tension between autonomous information-gathering workflows and content moderation mechanisms that were originally designed for short "chatbot" conversational use. As vendors continue to iterate on agentic architectures and develop guardrails specific to these agents, such safety vulnerabilities may be mitigated.

Importantly, ethical and practical limitations exist. Deep research agents are constrained to open-access information sources, as they cannot directly retrieve paywalled journal content, which could bias them toward certain sources. Their outputs also reflect the data they find—if the online content is skewed or outdated, the AI's output will inevitably be biased as well. Moreover, access to deep research agents themselves is likely to be uneven. Owing to their computational intensity, such systems are expected to be deployed through subscription-based or premium access models. This introduces an economic barrier that may limit adoption in resource-constrained settings, potentially reinforcing disparities in global medical research. As a result,

XSL•FO
RenderX

agentic research workflows may not only reflect prevailing publication biases but also exacerbate inequities in who is able to perform large-scale, synthesis-driven research. Over time, wider licensing or open-source alternatives may mitigate these effects, but economic access remains an important consideration in the near term. Moreover, the development of smaller, lightweight models with deep research capabilities (eg, IQVIA's Med-R1 8B model) may enable more computationally affordable evidence synthesis in resource-limited settings.

Transparency is another issue: it is not always clear how AI algorithms prioritize which sources to use or what reasoning led to a given conclusion. This opacity can make it hard for users to fully trust the results without verification, especially in medicine, where decisions require provenance and auditability. This is in contrast to evidence retrieval platforms such as OpenEvidence, which demonstrate a more transparent paradigm. OpenEvidence pulls directly from curated biomedical databases (eg, PubMed/Medline, clinical trial registries) and exposes every citation used, allowing users to trace each claim back to the original paper or paragraph [28,29]. In the future, deep research systems could adopt a similar architecture—combining autonomous reasoning with constrained, citation-verifiable retrieval pipelines.

Finally, there remains a notable lack of studies that rigorously evaluate the real-world capabilities of deep research agents within medicine. The existing literature largely consists of perspective pieces and proof-of-concept studies, with few empirical studies providing systematic benchmarking of these agents against established evidence synthesis standards or real-world clinical deployment. This gap underscores the need for methodologically robust research to determine whether deep research agents can reliably replicate or augment real-world medical workflows [30]. Recent efforts have begun to recognize this benchmarking gap. Du et al [31] introduced *DeepResearchBench*, the first comprehensive, multidomain benchmark specifically designed for deep research agents. Other initiatives, such as Rigorous Bench, further highlight how current evaluation schemes for LLMs fail to capture the multistep reasoning, citation integrity, and evidence synthesis capabilities characteristic of deep research systems [32]. Together, these studies emphasize that robust, clinically tailored benchmarks are urgently needed to measure not only the generative fluency of these agents but also their fidelity to source material, transparency of reasoning, and reproducibility of evidence synthesis. Moving forward, such benchmarks should also incorporate clinically relevant outcomes, including time saved, workload reduction, and decision support quality.

A Practical Verification and Transparency Framework for Clinical Use

To support the judicious use of deep research agents in clinical and research settings, we propose a structured evidence verification and transparency framework operating at both the system level and the end-user level.

At the system level, developers and vendors should adopt standards for logging retrieval queries, source rankings, and cross-attention or contribution weights linking generated statements to retrieved passages. While these signals do not provide full interpretability of model reasoning, they could enable partial auditing of multistep workflows by exposing which documents were retrieved and which sources most strongly influenced each synthesized claim [33]. Importantly, these logs should be made requestable and interpretable for clinicians and researchers, recognizing that domain experts (rather than system developers) are best positioned to interpret biomedical relevance and evidence quality. As these signals are not currently exposed by most proprietary systems, they should be regarded as forward-looking policy and design recommendations for model developers rather than tools currently available for clinical end-users.

At the user level, clinicians can follow a three-stage workflow when assessing agent outputs: (1) checking that citations originate from authoritative clinical repositories such as PubMed, Food and Drug Administration, guideline bodies, etc; (2) conducting selective audits of the most clinically actionable statements against original papers or guidelines, ensuring these claims are backed by multiple independent sources; and (3) assessing semantic coherence and task alignment to confirm the output remains focused on the specified clinical scope and does not exhibit internal contradictions or overconfidence in weak evidence. By combining architectural transparency with verification at the clinician level, this framework may structure responsible evaluation and improve confidence in the adoption of deep research agents for biomedical evidence synthesis.

Ultimately, trade-offs between the speed advantages of deep research agents and the time required for rigorous manual verification are likely, particularly in high-stakes clinical use. In practice, the net time savings of these systems may arise less from end-to-end automation than from accelerating early-stage information gathering and narrowing the search space for subsequent human verification.

Conclusion

The evolution of general-purpose LLMs to more specialized research-oriented agents, including deep research agents (eg, Open AI's Deep Research, Google's Gemini Deep Research, and Perplexity's Deep Research), marks significant progress in how medical knowledge can be accessed and synthesized. However, while they may represent a stepwise functional improvement in workflow automation, they remain an incremental extension of existing reasoning architectures and inherit many of their limitations. Current evaluations highlight the need for careful, evidence-based integration into medical practice. Their outputs continue to require human scrutiny to ensure fidelity, as they can miscite, misrepresent, or overgeneralize evidence in ways that are not immediately obvious. In medicine, these tools should complement, but not replace, rigorous human-driven research and review processes.

From a practical perspective, the idea of a "deep research" revolution may already be here, and if used judiciously, it can empower health care professionals—speeding up literature analysis and freeing up human experts to focus on critical thinking, creative insights, and patient-centered

XSL•FO
RenderX

decision-making. We believe the most sustainable path forward lies in collaboration with these agents, grounded in transparency and clinical accountability.

Acknowledgments

MYHW is supported by a pre-employment grant from the Ministry of Health Holdings, Singapore. AYO is supported by a National Institute for Health Research (NIHR)—Moorfields Eye Charity (MEC) Doctoral Fellowship (NIHR303691). DAM is supported by a EURETINA Retinal Medicine Clinical Research Grant. PAK is supported by a UK Research & Innovation Future Leaders Fellowship (MR/T019050/1) and The Rubin Foundation Charitable Trust. The views expressed in this publication are those of the authors and not necessarily those of the National Health Service, the NIHR, or the Department of Health and Social Care.

Funding

The authors declared no financial support was received for this work.

Disclaimer

ChatGPT 4o and o1 (OpenAI) were used to conduct the experiments discussed in this article. Generative artificial intelligence was not used in the writing of this Viewpoint.

Data Availability

The data that support the findings of this study are included in this published article and its supplementary information file.

Authors' Contributions

Conceptualization: MYHW, PAK
Methodology: MYHW
Data curation: MYHW, AYO, DAM
Supervision: PAK
Visualization: MYHW
Writing – original draft: MYHW
Writing – review & editing: MYHW, AYO, DAM, PAK

Conflicts of Interest

PAK is a cofounder of Cascader Ltd and has acted as a consultant for Retina Consultants of America, Roche, Boehringer Ingelheim, and Bitfount and is an equity owner in Big Picture Medical. He has received speaker fees from Zeiss, Thea, Apellis, and Roche. He has received travel support from Bayer and Roche. He has attended advisory boards for Topcon, Bayer, Boehringer Ingelheim, and Roche. The remaining authors do not have any conflicts of interest to declare.

Multimedia Appendix 1
Raw chat logs with Open AI's Deep Research.
[DOCX File, 483 KB - jmir_v28i1e88195_app1.docx]

References

1. Yu E, Chu X, Zhang W, et al. Large language models in medicine: applications, challenges, and future directions. Int J Med Sci 2025;22(11):2792-2801. [doi: 10.7150/ijms.111780] [Medline: 40520893]
2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
3. Brown A, Roman M, Devereux B. A systematic literature review of retrieval-augmented generation: techniques, metrics, and challenges. Big Data Cogn Comput 2025;9(12):320. [doi: 10.3390/bdcc9120320]
4. Jones N. OpenAI's "Deep Research" tool: is it useful for scientists? Nature New Biol 2025 Feb 6. [doi: 10.1038/d41586-025-00377-9] [Medline: 39915598]
5. Ong AY, Merle DA, Wagner SK, Keane PA. Exploring the dilemma of AI use in medical research and knowledge synthesis: a perspective on deep research tools. J Med Internet Res 2025 Jul 15;27:e75666. [doi: 10.2196/75666] [Medline: 40663724]
6. Introducing deep research. OpenAI. URL: https://openai.com/index/introducing-deep-research/ [accessed 2026-01-17]
7. Zhang W, Li X, Zhang Y, et al. Deep research: a survey of autonomous research agents. arXiv. Preprint posted online on Aug 18, 2025. [doi: 10.48550/arXiv.2508.12752]
8. Huang Y, Chen Y, Zhang H, et al. Deep research agents: a systematic examination and roadmap. arXiv. Preprint posted online on Jun 22, 2025. [doi: 10.48550/arXiv.2506.18096]
9. Gemini deep research. Gemini. URL: https://gemini.google/overview/deep-research/ [accessed 2026-01-17]

XSL•FO
RenderX

10. Jesudason D, Gao C, Seth I, Chan WO, Bacchi S. OpenAI's 'Deep Research' for the generation of comprehensive referenced medical text: uses and cautions. ANZ J Surg 2025 Nov;95(11):2250-2252. [doi: 10.1111/ans.70161] [Medline: 40325519]
11. Wang Z, Chen Z, Yang Z, et al. DeepEvidence: empowering biomedical discovery with deep knowledge graph research. arXiv. Preprint posted online on Dec 23, 2025. [doi: 10.48550/arXiv.2601.11560]
12. Dogra S, Arabshahi S, Wei J, et al. Evaluating large language models for radiology systematic review title and abstract screening. Acad Radiol 2025 Dec;32(12):7023-7034. [doi: 10.1016/j.acra.2025.08.014] [Medline: 40849232]
13. Loke WT, Srinivasan S, Yang GD, et al. Can “Deep Research” agents and general AI agentic systems autonomously perform systematic review and meta-analysis? Eye (Lond) 2026 Jan;40(1):147-149. [doi: 10.1038/s41433-025-04138-w] [Medline: 41315732]
14. Keplinger LE, Frashure LK, Duran SA, Hu G. Assessment of Deep Research for dermatology literature reviews: deep concern over the hype. J Eur Acad Dermatol Venereol 2025 Sep 4:40904191. [doi: 10.1111/jdv.70035] [Medline: 40904191]
15. Thurzo A, Varga I. Revisiting the role of review articles in the age of AI-agents: integrating AI-reasoning and AI-synthesis reshaping the future of scientific publishing. Bratisl Med J 2025 Apr;126(4):381-393. [doi: 10.1007/s44411-025-00106-8]
16. Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: a survey. arXiv. Preprint posted online on Dec 18, 2024. [doi: 10.48550/arXiv.2312.10997]
17. Yang R, Wong MYH, Li H, et al. Retrieval-augmented generation in medicine: a scoping review of technical implementations, clinical applications, and ethical considerations. arXiv. Preprint posted online on Nov 8, 2025. [doi: 10.48550/arXiv.2511.05901]
18. Gültekin O, Inoue J, Yilmaz B, et al. Evaluating DeepResearch and DeepThink in anterior cruciate ligament surgery patient education: ChatGPT-4o excels in comprehensiveness, DeepSeek R1 leads in clarity and readability of orthopaedic information. Knee Surg Sports Traumatol Arthrosc 2025 Aug;33(8):3025-3031. [doi: 10.1002/ksa.12711] [Medline: 40450565]
19. Yetkin NA, Baran B, Rabahoğlu B, Tutar N, Gülmez İ. Evaluating the reliability and quality of sarcoidosis-related information provided by AI chatbots. Healthcare (Basel) 2025 Jun 5;13(11):1344. [doi: 10.3390/healthcare13111344] [Medline: 40508957]
20. Chen S, Moreira P, Xiao Y, et al. MedBrowseComp: benchmarking medical deep research and computer use. 2025 Presented at: 39th Conference on Neural Information Processing Systems (NeurIPS 2025); Dec 6-7, 2025 URL: https://openreview.net/pdf?id=pSDvq2ZgPk [accessed 2026-03-11]
21. Yu A, Yao L, Liu J, et al. MedResearcher-R1: expert-level medical deep researcher via a knowledge-informed trajectory synthesis framework. . Preprint posted online on Aug 20, 2025. [doi: 10.48550/arXiv.2508.14880]
22. Introducing IQVIA med-R1 deep research agent. IQVIA. URL: https://www.iqvia.com/blogs/2025/12/introducing-iqvia-med-r1-deep-research-agent [accessed 2026-01-18]
23. Wang YM, Chen TJ. AI’s deep research revolution: transforming biomedical literature analysis. J Chin Med Assoc 2025;88(6):415-416. [doi: 10.1097/JCMA.0000000000001231] [Medline: 40128176]
24. Chirayath G, Premamalini K, Joseph J. Cognitive offloading or cognitive overload? How AI alters the mental architecture of coping. Front Psychol 2025;16:1699320. [doi: 10.3389/fpsyg.2025.1699320] [Medline: 41356052]
25. Abdulnour REE, Gin B, Boscardin CK. Educational strategies for clinical supervision of artificial intelligence use. N Engl J Med 2025 Aug 21;393(8):786-797. [doi: 10.1056/NEJMra2503232] [Medline: 40834302]
26. Bengio Y, Cohen M, Fornasiere D, et al. Superintelligent agents pose catastrophic risks: can scientist AI offer a safer path? Superintell Robot Saf Align 2025;2(5). [doi: 10.70777/si.v2i5.15569]
27. Chen S, Li Z, Han Z, et al. Deep research brings deeper harm. arXiv. Preprint posted online on Oct 13, 2025. [doi: 10.48550/arXiv.2510.11851]
28. Hurt RT, Stephenson CR, Gilman EA, et al. The use of an artificial intelligence platform OpenEvidence to augment clinical decision-making for primary care physicians. J Prim Care Community Health 2025;16:21501319251332215. [doi: 10.1177/21501319251332215] [Medline: 40238861]
29. Jagarapu J, Babata K, Chamarthi S, Hoyt R. The accuracy and repeatability of OpenEvidence on complex medical subspecialty scenarios: a pilot study. medRxiv. Preprint posted online on Dec 4, 2025. [doi: 10.64898/2025.11.29.25341091]
30. Moons P, Dou B, Desmedt CP. Google’s AI co-scientist and OpenAI’s deep research: new partners in health research? Eur J Cardiovasc Nurs 2025 Jul 21;24(5):800-807. [doi: 10.1093/eurjcn/zvaf066] [Medline: 40678974]
31. Du M, Xu B, Zhu C, Wang X, Mao Z. DeepResearch bench: a comprehensive benchmark for deep research agents. arXiv. Preprint posted online on Jun 13, 2025. [doi: 10.48550/arXiv.2506.11763]
32. Yao Y, Wang Y, Zhang Y, et al. A rigorous benchmark with multidimensional evaluation for deep research agents: from answers to reports. arXiv. Preprint posted online on Oct 2, 2025. [doi: 10.48550/arXiv.2510.02190]
33. Zhu Y. Applications of attention mechanisms in explainable machine learning. Acad J Comput Inf Sci 2025;8(10). [doi: 10.25236/AJCIS.2025.081008]

Abbreviations

AI: artificial intelligence
LLM: large language model

XSL•FO
RenderX

PEMAT: Patient Education Materials Assessment Tool
RAG: retrieval-augmented generation

Edited by A Coristine; submitted 21.Nov.2025; peer-reviewed by KM Sajjadul Islam, K Chen, S Saini; revised version received 25.Feb.2026; accepted 25.Feb.2026; published 26.Mar.2026.

Please cite as:
Wong MYH, Ong AY, Merle DA, Keane PA
Deep Research Agents: Major Breakthrough or Incremental Progress for Medical AI?
J Med Internet Res 2026;28:e88195
URL: https://www.jmir.org/2026/1/e88195
doi:10.2196/88195

© Matthew Yu Heng Wong, Ariel Yuhan Ong, David A Merle, Pearse A Keane. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

From Knowledge Graphs to Digital Twins: Perspectives on Modeling Patient Outcomes for Health Care Quality Assessment

Anna-Katharina Nitschke[1*], MSc; Juan G Diaz Ochoa[2*], Dr rer nat; Simone Neumaier[3], Dr med; Markus Knott[4], MSc

[1]Department of Physics, Heidelberg University, Heidelberg, Germany
[2]PerMediQ, Pelargusstraße 2, Stuttgart, Germany
[3]Institut für klinische und experimentelle Transfusionsmedizin, Tübingen, Germany
[4]Klinikum Stuttgart, Stuttgart Cancer Center, Stuttgart, Germany
[*]these authors contributed equally

Corresponding Author:
Juan G Diaz Ochoa, Dr rer nat
PerMediQ, Pelargusstraße 2, Stuttgart, Germany

Abstract

Medical applications of mathematical modeling, including machine learning models, knowledge graphs, and health digital twins, primarily involve the prediction of patient outcomes. This expert perspective examines how mathematical modeling can contribute to health care quality management. Definitions of procedures, patient outcomes, and quality metrics are provided with a quantitative focus. The emphasis is subsequently placed on 3 categories of patient-centered quality of care, namely, patient safety, procedure accuracy, and procedure efficacy, for which a conceptual and mathematical description is provided. Different levels of modeling tasks essential for managing patient-centered quality of care are identified. This article facilitates a deeper understanding of the topic by assigning relevant publications to these 3 quality categories. Focus is placed on the applicability of graph-based methods, including knowledge graphs and health digital twins, to improve quality management in health care. We have presented a clinical scenario and provided information on methodological limitations, future research directions, and practical implications.

(J Med Internet Res 2026;28:e81946) doi:10.2196/81946

KEYWORDS

quality management; medicine; patient outcomes; machine learning; knowledge graphs; digital twins; human digital shadows

Introduction

In recent years, considerable efforts have been made to improve patient outcomes by standardizing medical care across health care systems for patients [1]. This has been achieved in particular through the creation and application of evidence-based medical guidelines [2]. Here, standardization (ie, the design of medical procedures via guidelines aimed at the systematic integration of scientific evidence) plays a relevant role. This is achieved by selecting the most appropriate health decisions required during the patient's journey. Standardizing patient care is expected to improve cost and workflow efficiency, as well as resource allocation [3].

Many forms of medical interventions (eg, the selection of a therapy) can be seen as part of well-defined processes (ie, a series of actions taken to achieve an outcome) [4]. This concept intersects with the industry, where processes refer to a set of interconnected tasks that convert input into specific outputs. Therefore, modern medicine can be seen as a set of planned processes recommended by guidelines. This overlap has led to industrialization in medicine, with both positive (evidence-based treatments) and negative aspects (eg, dehumanization of medicine and focus on economic aspects) [3].

Furthermore, process-oriented health care focuses on process optimization and economic optimization instead of patient care (for a review of the industrialization of health care and its criticisms, see the chapter by Da Silva [5]). On the other hand, a patient-centered and process-oriented perspective can lead to positive aspects, such as breaking silos in clinical data, optimizing outcomes (ie, patient recovery after therapy), minimizing medical errors and expenses for diagnosis and treatment, and maximizing clinical outcomes [6]. In such cases, the balance between standardization and customization is a paramount problem [3].

Records containing medical interventions generated when a patient receives medical treatment (eg, electronic health records [EHRs]) are essentially defined as processes with different steps, including diagnosis, medical intervention (such as medication, surgery, etc), and tracking of a patient's condition to prevent further health problems [7]. Thus, it is assumed that a patient's process can be modeled in a similar way to the dynamics of a productive industrial process. Through the logical integration of different steps, both the well-being of the patient and the economic performance of the institution (eg, hospital) can be optimized.

XSL•FO
RenderX

Mathematical modeling can significantly improve the way medical processes are implemented and meet quality standards. However, medical data are often not used beyond direct patient care. The reasons for this are manifold and include a lack of digitalization, inadequate structured health data, stringent data protection laws, insufficient technical infrastructure, and limited personnel or financial capacities. Health care providers often use medical jargon and classification schemes that are country-specific and even center-specific (an exception is the use of the *International Classification of Diseases* [*ICD*] for disease classification), and information contained in the form of unstructured data is often not machine-readable.

In addition, health care practitioners (HCPs) are still not sufficiently trained in machine learning (ML) technologies. Moreover, IT specialists rarely understand clinical processes in health care facilities. In these dynamic times, it is crucial to bring different disciplines together if ML applications are to provide real benefits in everyday clinical practice [8]. Assuming that notions of causality can be identified in medicine and used to identify mechanisms and derive predictive models [9,10], ML applications can provide information about patient outcomes, as well as unexpected or undesired events, and support medical decision-making.

This article presents a novel perspective on modeling patient outcomes to improve health care quality assessment. The applicability of mathematical models (focusing on graph-based methods and health digital twins [HDTs]) is considered within a quantitative concept of health care quality management, considering 3 categories of patient-centered quality of care (PCQC), namely, patient safety, procedure accuracy, and procedure efficacy, thereby providing a consistent taxonomy. To date, this relevant topic has been explored only to a limited extent. This work aims to contribute to improving health and well-being (according to the UN sustainability goals, SDG3) [11]. We discuss the potential, limitations, and practical implications of the application of these technologies for modeling PCQC.

Definition of Patient Outcomes for Quantitative Quality Assessment

Overview

In medicine, it is generally difficult to define patient outcomes quantitatively (related to measurable information: exact metrics and observables captured by sensors) and qualitatively (related to nonmeasurable descriptions: categories) because patients are heterogeneous and complex organisms, and an overall assessment of their health status therefore remains a challenge.

Our goal is to identify these variables of interest and place them in a broader understanding of quality assessment, paving the way for mathematical modeling in this field. To this end, we aim to introduce a description of patient outcomes as a fundamental concept for evaluating quality indicators (QIs). These indicators are relevant for decision-making because they provide an estimate for which option holds the greatest desirability or value [12], that is, improvement of the quality of care (eg, improvement of patient well-being, avoidance of patient death, optimal use of resources, etc).

Patient Outcomes

From a medical perspective, a patient outcome is defined in terms of what is meaningful and valuable to the individual patient [13]. However, this definition is very general and difficult to relate to a quantifiable metric. There is still a conceptual problem with the general definition of results in the medical context, such as patient-relevant outcomes (morbidity, mortality, and quality of life) and surrogate or biomarker outcomes [14]. These outcomes can differ significantly depending on the underlying disease. Liu et al [15] summarized the definition of patient outcomes related to clinical practice "as any change [within the patient's health status] that results from health care [for which each profession] has developed outcome measures that focus on the standards, activities, and impact of its discipline."

Thus, patient outcomes are manifold and range from PCQC [14] to institutional performance [16] and to patient-reported outcomes [17], which can be assessed from different data sources (for a comprehensive overview, see information provided by Busse et al [18]).

In this perspective article, we focus only on the description of PCQC (Table 1), for which a comprehensive review was published by Kersting et al [14]. We aim to provide a reference for the general concept, which is why we generally refer to medical procedures as follows: A *medical procedure* is the overarching phrasing for any type of observation, measurement, diagnosis, or treatment performed on the patient.

However, not all patient outcomes can be measured objectively, and only a limited number can be translated into clear, quantifiable metrics. In other words, a mathematical model can reflect only a fraction of the clinical reality.

XSL•FO
RenderX

Table . Subdivision of patient outcomes into 3 areas (patient-centered quality of care, institutional performance, and patient-reported outcomes).

Variable[a]	Patient-centered quality of care	Institutional performance	Patient-reported outcomes
Definition	Patient-centered QIs[b] in health care [19] measured or captured by HCPs[c] [14].	Process execution and resource consumption on an institutional statistical evaluation level [16].	Patient's subjective evaluation of health experience. Report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else [17].
Exemplary categories	Medical evaluation of: • Patient safety • Procedure accuracy • Procedure efficacy	Financial evaluation of: • Patient waiting time • Economic efficiency of medical procedures • Productivity	Questionnaires evaluating: • Symptoms • Quality of life • Health care experience

[a]For each area, a definition and a further subdivision into categories are given to enable a quantitative description of the different areas.

[b]QI: quality indicator.

[c]HCP: health care practitioner.

In the context of clinical trials, great efforts have been made to establish quantifiable patient outcomes. Clinical trials use predefined endpoints to measure patient outcomes. Typical endpoints are patient survival at a given time, incidence of clinical events (such as stroke), clinical performance measures, or patient-reported outcomes [20].

Considering the impact that EHRs have on the quality of care, we provide a definition of patient outcomes on the basis of changes registered in the health record and defined as events [21] (assuming that the data relevant for the definition of patient outcomes are machine-readable). We refer to *patient outcomes* as events registered in the health records. Such events can include changes in a patient's health state or changes in medical procedures.

These outcomes enable health care providers and organizations to measure the impact of their services on patient well-being, identify key areas for improvement, and enhance the overall quality of care.

Medical Guidelines

Medical guidelines are systematically developed statements that reflect the current state of medical knowledge. Medical experts prepare these guidelines in cooperation with professional associations. These evidence-based recommendations are based on clinical studies published in scientific literature and/or other existing evidence informed by expert experience and consensus.

The extracted information is then assessed and published by organizations, such as the World Health Organization, for international guidance or by country-specific professional societies (in Germany, the *Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e. V.* [22]). Guidelines are not a static corpus and are continuously adapted and further developed on the basis of current clinical studies and scientific findings. They intend to support the decision-making of HCPs in order to enable appropriate care for certain health problems. Medical guidelines consist of complex workflows in mostly narrative form and can also contain multiple, equivalent treatment alternatives for the same disease unit. In general, comparing performed procedures with guidelines is challenging, especially considering that guideline recommendations should not be understood as rigid limits. In medical guidelines, a distinction is made between several grades of recommendations ("can" or "should") on the basis of available evidence (see, for example, the AWMF register for prostate carcinoma [in German] [23]). Additionally, medical guidelines are country-specific and may differ across national states and regions.

Information on QIs

To assess in part how guidelines are applied within an institution and to improve the quality of care, several QIs have been developed by different institutions. These QIs serve as internal quality management tools for medical institutions, enable benchmarking with other institutions, and aim to improve the quality of care for patients [24]. In Germany and Switzerland, routine billing data from hospitals are used for this purpose, if a health care facility decides to participate voluntarily in such initiatives [25]. In Germany, QIs for cancer treatment are defined annually by the German Cancer Society (*Deutsche Krebsgesellschaft* [DKG]). Hospitals that meet the criteria defined by the DKG are eligible to receive DKG certification. The outcomes of cancer patients treated in these DKG-certified centers are superior in terms of patient survival [26]. However, these QIs are assessed retrospectively on an institutional basis and cover aspects of the application of medical guidelines.

Quality Metrics for PCQC

Mathematical modeling is used to represent a real-world phenomenon by capturing and analyzing the most significant relationships between different observables through formalized relationships between mathematical parameters. For the quantitative description of patient outcomes, they need to be described by numerically or categorically measurable parameters. These parameters capture various relevant aspects of the overall performance of patient-centered care quality. These numerical and categorical parameters are referred to as quality metrics (QMs) because they enable an evaluation of patient care. There is a hierarchical ordering that assigns each QI to several QMs, which might be defined in a disease- or context-specific way, to place them into the perspective of

XSL•FO
RenderX

mathematical modeling. We refer to *QM* as a quantitative measure of QIs that relates to adherence to medical guidelines (procedural accuracy); QM also refers to patient-centered care (patient safety and effectiveness of the procedure).

Information on PCQC

Overview

We present an overview of the broad concept of PCQC. According to the classification provided in Table 1, we define in detail the 3 basic categories used for the further discussion of health care quality assessment: patient safety, procedure accuracy, and procedure efficacy.

For PCQC, there is a strong reference to medical guidelines and appropriate clinical practice. Owing to the increasing complexity of medical knowledge reflected within medical guidelines, it is becoming increasingly difficult for health care professionals to keep track of changes. The application of mathematical modeling is intended to help health care professionals comply with the guidelines (ie, a sequence of recommended procedures and support for medical decision-making). Medical practitioners select their decisions on the basis of the probability of initiating a change in the patient's health state (assuming that physicians are either using heuristics or making rational decisions on the basis of available information [27]). Use cases for the application of mathematical modeling for PCQC are presented in Multimedia Appendix 1.

Mathematical modeling can be included for predicting QMs to enhance health care quality. In this context, the patient's journey can be understood as a complex chain of procedures that can be logically interconnected. Therefore, it is imperative to look at the whole patient journey as a process in which each procedure is just a point in time (t_{i-1}, t_{i+1}). As illustrated in Figure 1, the general way in which health care quality is assessed can be described as follows: collecting data from a patient (element 1), which a clinician uses to make a medical decision regarding the patient's further procedure (element 2) and whose impact is measured after the procedure is carried out by looking at the patient outcome (element 3), which allows assessment of the quality of the health care provided (element 4). One can distinguish between two methods of implementation: (1) use the prediction in a prospective way (dashed arrow; meaning that before the medical decision is made, the clinician receives feedback from the model and can adjust the decision, leading to potentially different outcomes and improved health care quality) and (2) use a retrospective inclusion (dotted arrow), allowing an evaluation of the observed patient outcome and an assessment of the quality of care.

Another distinguishing feature of mathematical modeling is whether the model works for individual patients or for patient cohorts. Hence, for modeling the QM, we can consider the following representation levels:

1. Patient-cohort predictions for quality management of patient populations: It involves statistical modeling of the patient population for controlling domestic and global medical practices. At this level, mathematical models describe an "average and black box virtual patient" that tests or corroborates published and domestic guidelines.
2. Patient-centered predictions: It involves modeling of specific targets and the inclusion of personal information in gray (partial knowledge of model details) or white-box (full knowledge of model details) models to model "virtual individual patients" and evaluate published or domestic guidelines at an individual level. This implies that models should reflect and predict individual deviations from the expected results of the main patient cohort (see, for example, information provided by Peng et al [28]).

The patient journey can be represented as a chain of observed patient outcomes (for example, the main disease and corresponding codiseases) and possible medical decisions that lead again to distinct patient outcomes over time that are patient journey dependent. Medical guidelines specify possible medical decisions following patient outcomes. The quality of the health care received during the patient journey can be assessed by comparing different elements within this decision-tree structure. Three categories of PCQC can be distinguished: patient safety, procedure accuracy, and procedure efficacy. In the following sections, the 3 different categories of modeled patient outcomes are defined by describing their context to the patient journey in a qualitative and quantitative way, as well as identifying suitable QMs.

XSL•FO
RenderX

Figure 1. Visualization of key parameters for health care quality management and mathematical model integration throughout the patient's journey over time. The predicted quality metric of the mathematical model can be implemented in 2 ways: prospective (dashed line) or retrospective (dotted line). It includes the following elements: (1) data collection from the patient, (2) medical decision made by a clinician, (3) patient outcome after the procedure, (4) quality assessment, (5) quality metric prediction from the mathematical model, and (6) data storage. This can analogously be performed at the population level (patient-cohort predictions) instead of at the patient level (patient-centered predictions).

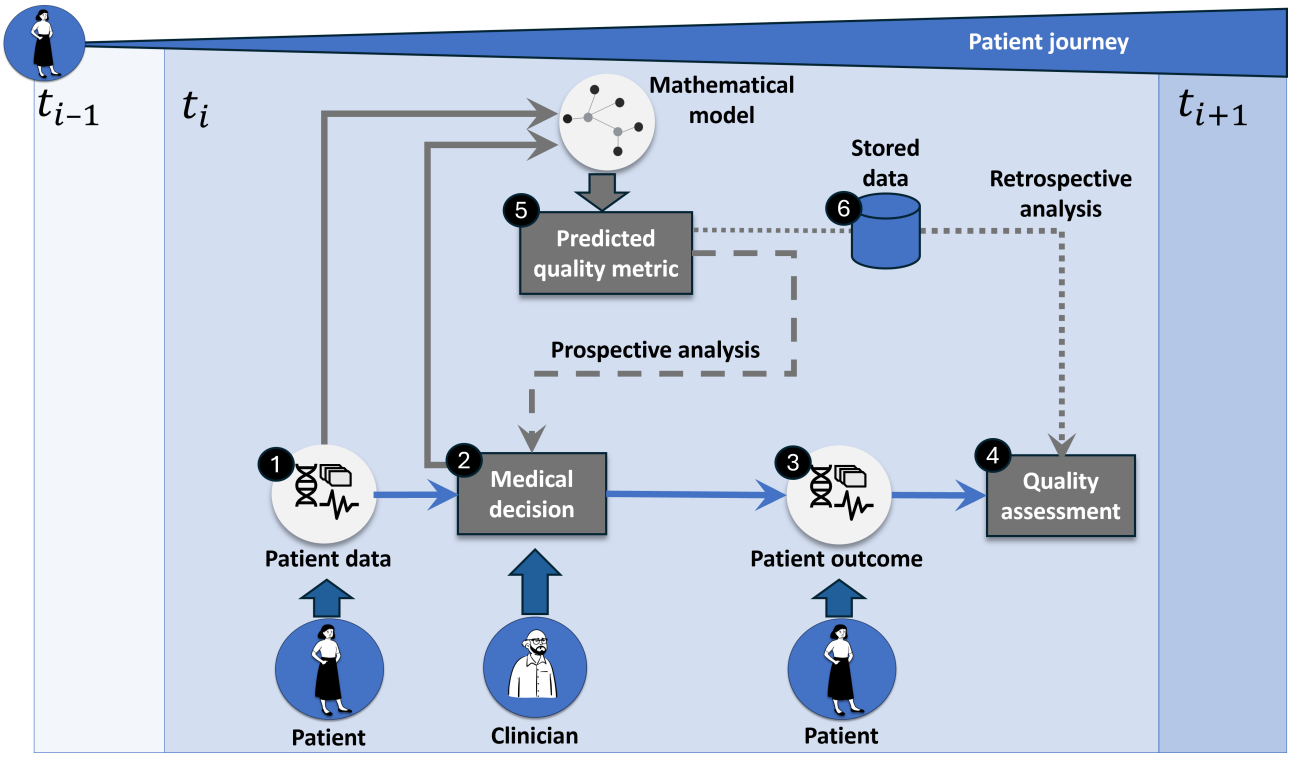

Patient Safety

Providing an exact definition of patient safety is challenging, as this term remains under debate in the literature [29]. Patient safety (related to the taxonomy introduced in Table 1) can be measured by patient safety indicators, which refer mostly to the avoidance of accidental and unwanted events in health care. Patient safety is partially measured within health care facilities via predefined inpatient QIs. Patient safety indicators and inpatient QIs were developed by the Agency for Healthcare Research and Quality (AHRQ [30]), and together, they provide a comprehensive picture of QIs for patient safety. These QIs are mostly determined from the billing data of the hospitals and the medical documentation data contained therein (*ICD* codes for diagnoses and International Classification of Health Interventions [ICHI] codes [31,32] for medical procedures, although each country can have its own classification schema). Possible examples of QIs that reflect patient safety can be extracted from the AHRQ [30], such as complication-related indicators, procedure-related errors, medical-related errors, and hospital system indicators.

There are a variety of possible patient outcomes, some of which are desirable and some of which are adverse (an adverse outcome is evaluated depending on the specific clinical context). Assuming that physicians are rational decision makers, the medical decision is an evaluation of the risk of previous and future outcomes (according to the classic *Theory of Rational Decision Making*; see, for example, information provided previously [33]). A mathematical model for patient safety can provide support through checking (retrospective) or predicting (prospective) how likely the outcome of a medical decision will be an element of the set of adverse outcomes.

QMs can be related to the respective QIs for the quantitative representation of patient safety. Examples can be defined as follows: (1) Mortality rate and failure to rescue, which have emerged as important QMs [34]; (2) Emergency readmission to the hospital after a procedure among outpatients, which can happen, for instance, if prophylaxis has been forgotten or if safety guidelines have not been properly followed; (3) Number of reoperations or complications shortly after the operation, which might indicate a need for investigation within the facility; and (4) Admission to intensive care units (ICUs) and length of stay (LOS) in the ICU among inpatients (different from the intermediate medical care unit).

Table 2 depicts the relationships of the example QMs to the respective QIs, the patient status, the data type, and the QM type. The "Combination event with a time period" column indicates that the observed metric is provided in conjunction with a time parameter.

XSL•FO
RenderX

Table . QIs[a] and QMs[b], data type, correlation involving patient status, and combination event with a time period for patient safety.

Patient safety - QM	QI	QM kind	Data type	Correlation involving patient status	Combination event with a time period
Failure to rescue	Patient safety	Disease specific	Binary	ICHI[c] with mortality	Yes (eg, within 30 days)
Hospital readmission	Hospital readmission	Context specific	Binary	—[d]	Yes (eg, within 30 days)
Number of reoperations or complications	Patient safety	Context specific	Numeric	—	Yes (eg, within 90 days or 1 year)
Admission to ICU[e] and ICU-LOS[f]	Patient safety	Context specific	Binary	ICHI with critical care	Yes (eg, within 48 - 72 hours)

[a]QI: quality indicator.

[b]QM: quality metric.

[c]ICHI: International Classification of Health Interventions.

[d]Not applicable.

[e]ICU: intensive care unit.

[f]ICU-LOS: length of stay in the intensive care unit.

Procedure Accuracy

The purpose of medical guidelines is to reduce errors and disparities in medical care while supporting best practices and responsibilities in medicine on the basis of the current state of scientific knowledge. Following established medical guidelines increases the probability of achieving better clinical outcomes [35-37]. Therefore, an appropriate measure of PCQC (related to the taxonomy introduced in Table 1) is the adherence of HCPs to medical guidelines, including the correct and accurate application of a medical procedure, which we refer to as procedure accuracy. To date, there is no structured or automated assessment of whether recommended medical guidelines have been followed by HCPs for an individual patient.

For the inclusion of mathematical modeling for health care quality assessment, it is necessary to identify metrics that can be defined precisely (if possible, with binary or clear multilevel indicators). In most cases, these metrics cannot be defined universally and require disease-specific development. Thus, with respect to procedure accuracy, mathematical modeling can provide useful insights as follows:

- Anomaly detection and mathematical models can help discover common patterns in the relationship between diagnoses and medical procedures [38] or in the retrospective analysis of clinical narratives via natural language processing [39]. Such pattern recognition can be used to discover for which patient the pattern deviates from the guidelines and if such anomalies are an indication of a patient-centered procedure [40].
- Local guidelines (which are not merely deviations from national standards but adaptive tools shaped by real-world experience) can be created by hospitals or health care institutions. These include local adjustments to general guidelines when differences are identified, and the efficacy of alternative methods has been demonstrated [41].
- Patient-centered interpretation guidelines can be created, and the integration of additional data, such as physiological or genetic data, into models can provide recommendations for action that consider specific patient variations. Hence, models can enable the implementation of patient-centered quality management.

Clear results include checking whether the recommendations of a guideline for performing a standardized procedure or a set of procedures have been followed in an individual patient. This information can be retrieved from EHRs, laboratory information management systems, or even reimbursement data. In addition to compliance with a guideline, medical errors can also be reduced. Examples of QMs for procedure accuracy are as follows (Table 3):

- Detection of patient outliers: The average amount of hospitalization required by a medical procedure (LOS in hospital compared with the average) and the appearance of adverse events (number of infections, etc, compared with the average) can be assessed.
- Expected medical procedure: On the basis of the *ICD* diagnosis, specific ICHI procedures can be predicted, and deviations in the ICHI code from the prediction can be detected.
- Revision operation: The number of necessary additional surgical interventions after surgery (eg, due to complications of the initial surgical intervention) can be assessed.

While these metrics are easy to extract at the site level, they risk missing relevant contextual factors. It is desirable to integrate additional qualitative or clinical data points for a more holistic assessment. However, in clinical practice, simpler modeling techniques are used.

XSL•FO
RenderX

Table . QIs[a] and QMs[b], data type, correlation involving patient status, and combination event with a time period for procedure accuracy.

Procedure accuracy - QM	QI	QM kind	Data type[c]	Correlation involving patient status	Combination event with a time period
Detection of patient outliers	Accuracy of procedure application	Context specific	Binary	ICHI[d] (patient) with ICHI (population)	No
Number of deviations of procedures with respect to guidelines (Δ[e] ICHI)	Accuracy in the application of medical guidelines	Context specific	Multilabel	ICHI (patient) with ICHI (population)	No
Revision of operations	Patient-specific management	Disease specific	Binary	__[f]	Yes (eg, within 90 days or 1 year)

[a]QI: quality indicator.

[b]QM: quality metric.

[c]Binary data correspond to 2 possible values (yes/no or 0/1).

[d]ICHI: International Classification of Health Interventions.

[e]The symbol Δ refers to a difference with respect to a reference value.

[f]Not applicable.

Procedure Efficacy

A general definition of efficacy has been provided by Lynch [42] ("efficacy is the capacity [of a medical intervention] to produce an effect"). For quality assessment, procedure efficacy (related to the taxonomy introduced in Table 1) refers to the optimal selection of procedures aiming at achieving the best possible patient outcome.

Retrospective models could compare the medical decisions chosen by HCPs with other possible choices, which may lead to better outcomes. Hence, procedure efficacy can also be considered for monitoring a decision (made under imperfect information) by observing the patient's response or for making a prognosis and a subsequent decision under imperfect information. Owing to its nature, procedure efficacy requires the implementation of mechanisms and causal relationships where target/trial emulation is needed. This type of modeling contrasts with conventional predictive risk modeling tasks (which are applied, for instance, to evaluate patient safety and procedure accuracy).

The following are relevant QMs for procedure efficacy (Table 4):

- Biomarker levels: Predictions and projections of biomarker levels and their relationships with accepted levels to assess and predict a patient's condition (acceptable [ie, considering parameters below a threshold] or deteriorating [ie, parameters above a threshold]) can lead to specific medical decisions. For example, a stable blood glucose range over time indicates the efficacy of diabetes treatment.
- Tumor-free survival: Time (in months or years) from complete tumor resection until tumor recurrence.
- Symptom risk score: This score quantifies, for example, the risk of an improvement or deterioration in a specific medical condition.
- Overall survival: Indicates the time (in months or years) from diagnosis until death from any cause.
- Pathologic complete remission: The proportion of patients showing no residual viable tumor cells on histopathological assessment after the completion of tumor therapy.

XSL•FO
RenderX

Table . QIs[a] and QMs[b], data type, correlation involving patient status, and combination event with a time period for procedure efficacy.

Procedure efficacy - QM	QI	QM kind	Data type[c]	Correlation involving patient status	Combination event with a time period
Biomarker levels	Surrogate for disease control	Disease specific	Time series	Biomarker with ICHI[d]	—[e]
Tumor-free survival	Tumor control	Disease/context specific	Binary	Biomarker/time	Yes (eg, after 1, 3, or 5 years)
Recurrence-free survival	Tumor control	Disease/context specific	Binary	Biomarker	Yes (eg, 12, 24, or 36 months)
Symptom risk scores	Satisfactory patient condition	Disease specific	Binary	—	Yes (eg, directly or within 6 - 12 weeks after treatment)
Overall survival	Tumor control	Context specific	Binary	Biomarker	Yes (eg, after 1, 3, or 5 years)
Pathologic complete remission	Tumor control	Disease specific	Binary	Biomarker	Yes (eg, directly or within 6 - 12 weeks after treatment)

[a]QI: quality indicator.

[b]QM: quality metric.

[c]Binary data correspond to 2 possible values (yes/no or 0/1).

[d]ICHI: International Classification of Health Interventions.

[e]Not applicable.

In legal or patient-communication contexts, terms such as “survivability” are typically used, which refer to an inherent capacity or probability of survival but are not considered valid clinical endpoints. In this study, we refer to well-defined outcome measures such as tumor-free survival.

Mathematical Modeling of PCQC

Overview

A mathematical representation of the patient’s condition to predict the outcome is essential for a precise evaluation of quality management. In summary, patient safety (Table 2) and procedure efficacy (Table 4) are related to clinical endpoints, whereas procedure accuracy (Table 3) is closely related to precision in the implementation of clinical guidelines and in reference to medical evidence. To avoid conceptual conflation in QMs, procedure accuracy is always defined with respect to documents (scientific literature or guidelines) containing medical or scientific evidence. Importantly, some events, such as patient outliers (Table 3), are estimated inside a time horizon and are mapped into binary states. Despite their convenience for mathematical modeling, there are relevant trade-offs, such as information loss or threshold choices, which can be problematic for the final validity of the models.

In the context of the 3 quality management categories, mathematical models need to perform distinct modeling tasks. First, the information coming from the patient needs to be stored and ordered in a data storage system (graph-based structures, such as knowledge graphs [KGs], can be used to encode patient data and information in a systematic manner). On the basis of the available data, single mathematical models (we focus on ML models) are trained to accomplish tasks that can either be classification tasks of the patient state or prediction tasks of the evolution of a parameter describing the patient response. The model predictions can be either prospective or retrospective.

We focus on KGs to integrate information from single events within the patient journey. By analyzing patterns from these KGs extracted via methods, such as graph neural networks (GNNs), the PCQC can be evaluated. For some QMs, it is necessary to consider the medical process in its specific context within the patient’s journey, which can be understood as a chain of events. This is an important aspect when assessing procedure efficacy, as it strongly depends on the previous and subsequent procedures. We present the concept of HDTs as a model integration framework, enabling a patient path representation.

The growing interest and relevance of mathematical modeling of patient outcomes are indicated in Table 5, which includes the number of published articles from the last 7 years (see Multimedia Appendix 2 for the methodology used to extract these data). However, the number of publications including digital twins (DTs) or GNNs is only a fraction of the total number of publications addressing the application of ML to predict patient outcomes (for a comprehensive review, see the article by Kline et al [43]).

XSL•FO
RenderX

Table . Publications in the last 7 years.

Variable[a]	Year, n						
	2019	2020	2021	2022	2023	2024	2025
Machine learning	899	1517	2289	2662	3192	4701	7146
Graph neural network	13	7	10	16	20	39	55
Digital twin	2	5	7	7	20	63	161

[a]Queries considering "patient outcomes" and "machine learning," "graph neural network," or "digital twin" (see Multimedia Appendix 2 for the method used to construct the table; the content was actualized on January 5, 2026).

Graph-Based Methods for PCQC

With EHRs, extended research integrating different patient data for health care quality assessment is now possible. Regarding this development, Si et al [44] published a comprehensive survey about the use of current advances in patient representation based on data stored in EHRs.

In recent years, the application of linked knowledge in graphs has become relevant, in part because it allows holistic individualized representation, either by considering individual characteristics in an entire population [38] or by providing a patient-centered representation of the disease and the corresponding outcomes. Data representation as a graph is a method to leverage transparency in model construction. Graph models are also a way to map unstructured data, such as clinical narratives, into machine-readable data via large language models (LLMs). Additionally, graph models, including GNNs, have been considered to represent a way to close the gap between ML and symbolic reasoning [45], introducing the desired interpretability for ML methods.

The sample health care KG in Figure 2 can serve to represent the concept of graph models in health care: the health status of a patient can be modeled as the interlinking of various factors, such as the disease's attributes (including type, symptoms, and anatomy), the patient's genetics (including pathways and biological processes), the kind of drug compounds, and the characterization of side effects. Hence, graph models can function as holistic representations of a patient's health status through the integration of different data sources into a single model.

XSL•FO
RenderX

Figure 2. A simple health care knowledge graph inspired by the article of Abu-Salih et al [46]. The abstract graph shows how a disease is connected to its attributes (such as type, symptoms, and anatomy) and to compounds (and attributes such as side effects and pharmacological classes), and how it is associated with genes (and attributes such as pathways and biological processes). Interactions with the disease, which we refer to as medical procedures (such as diagnosis, medication, and measurement), can be added to the graph and can lead to the development of the disease over time (t_i–t_{i+1}).

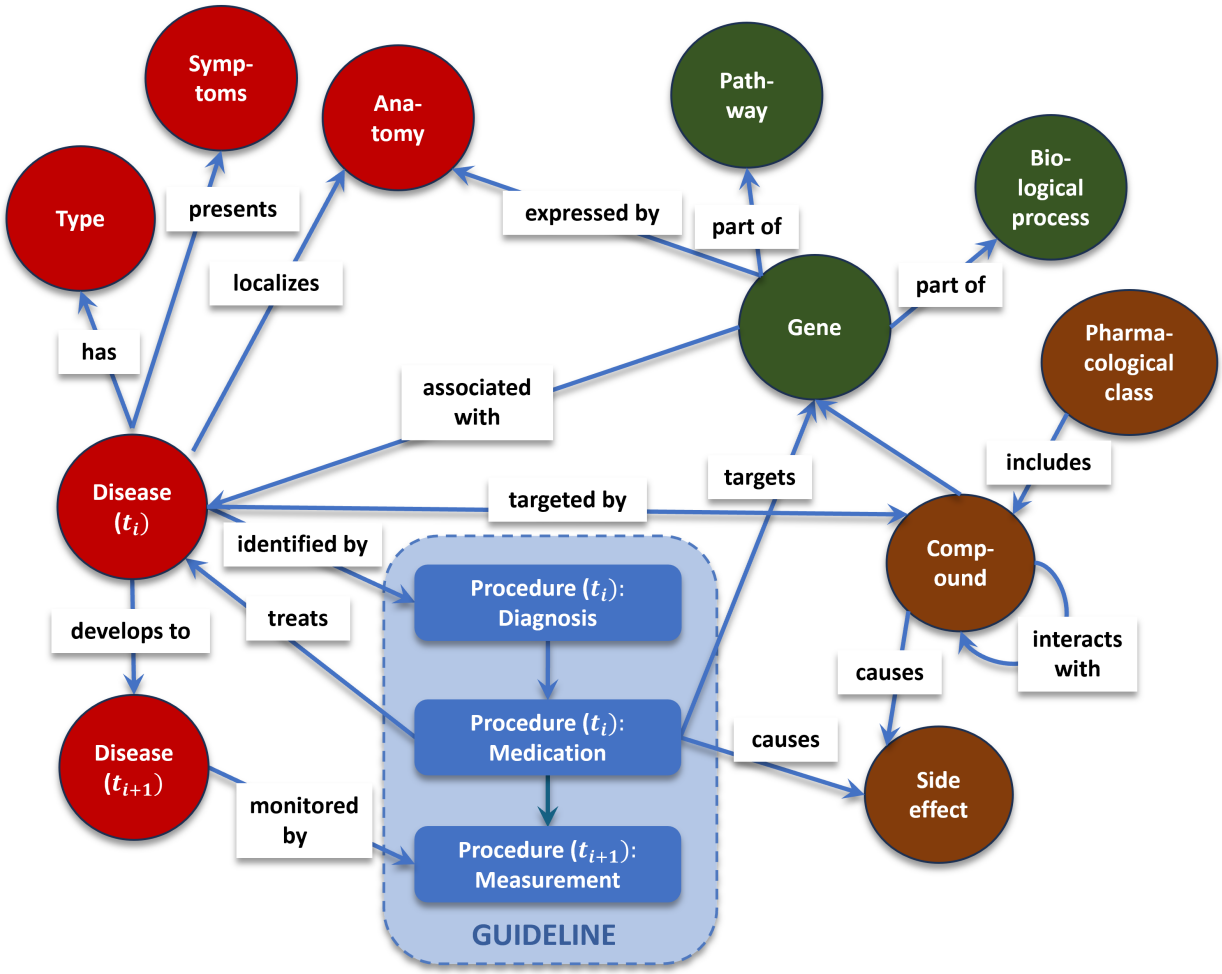

In graph models, representation learning can be applied to convert multivariate time series and static features into nodes of the graph [47], for example, when multiple steps, such as various procedures in different time steps or disease stages, are converted into nodes in the entire graph.

Hence, starting from the disease stage at time point t_i, procedures, such as diagnosis, can be performed, which introduces a new node into the KG. As clinical work is ideally performed within the frame of the guidelines, suitable medication will be proposed and applied. Any action taken through the performance of a medical procedure can lead to an evolution of the patient's disease state, which, at a time point, can be monitored by the next procedure, again inducing new nodes in the KG. Relating medical procedures to specific patient characteristics (such as the genetic profile of the patient, side effects, etc) enables a holistic view, allowing guidelines to be tailored to the individual patient and to differ from institutional guidelines. Graphs are not limited to representing a single event and can represent a whole process (ie, the patient's journey). This is useful for DT implementation, which will be discussed in the following section.

In addition to patient-centered representations, graph models for disease-centered representations can also be constructed, as shown in Figure 3. These can be created from electronic patient records because hospital records contain rich information about the interrelationship of diagnoses (diagnoses and corresponding codiagnoses), the interrelationship of medical procedures (one medical intervention leads to another), and the interrelationship of diagnoses and procedures (due to the causal relationship between the diagnoses and the associated procedures). If a malignant neoplasm of the lungs (C34) is diagnosed, it is already known from earlier observations that it is often accompanied by chronic obstructive pulmonary disease (J44) and secondary malignant neoplasms of the brain (C79.3). Against the background of these relationships, 4 possible methods can be identified: computed tomography, bronchoscopy, magnetic resonance imaging, and positron emission tomography.

In Multimedia Appendix 3, we present an overview of the literature concerning the application of graph models for patient safety, procedure accuracy, and procedure efficacy.

XSL•FO
RenderX

Figure 3. Exemplary knowledge graph related to disease diagnoses and medical procedures. In this example, we present the relationship involving a malignant neoplasm of the lung (C34). We also present typical codiagnoses of this disease on the left of the image (J44 and C79.3). On the right of the image, we present corresponding medical procedures that integrate the main diagnosis and codiagnoses (such as computed tomography, bronchoscopy, magnetic resonance imaging, and positron emission tomography). This example has been extracted from guidelines for lung cancer in oncology [48]. *ICD*: *International Classification of Diseases*; OPS: Operationen- und Prozedurenschlüssel.

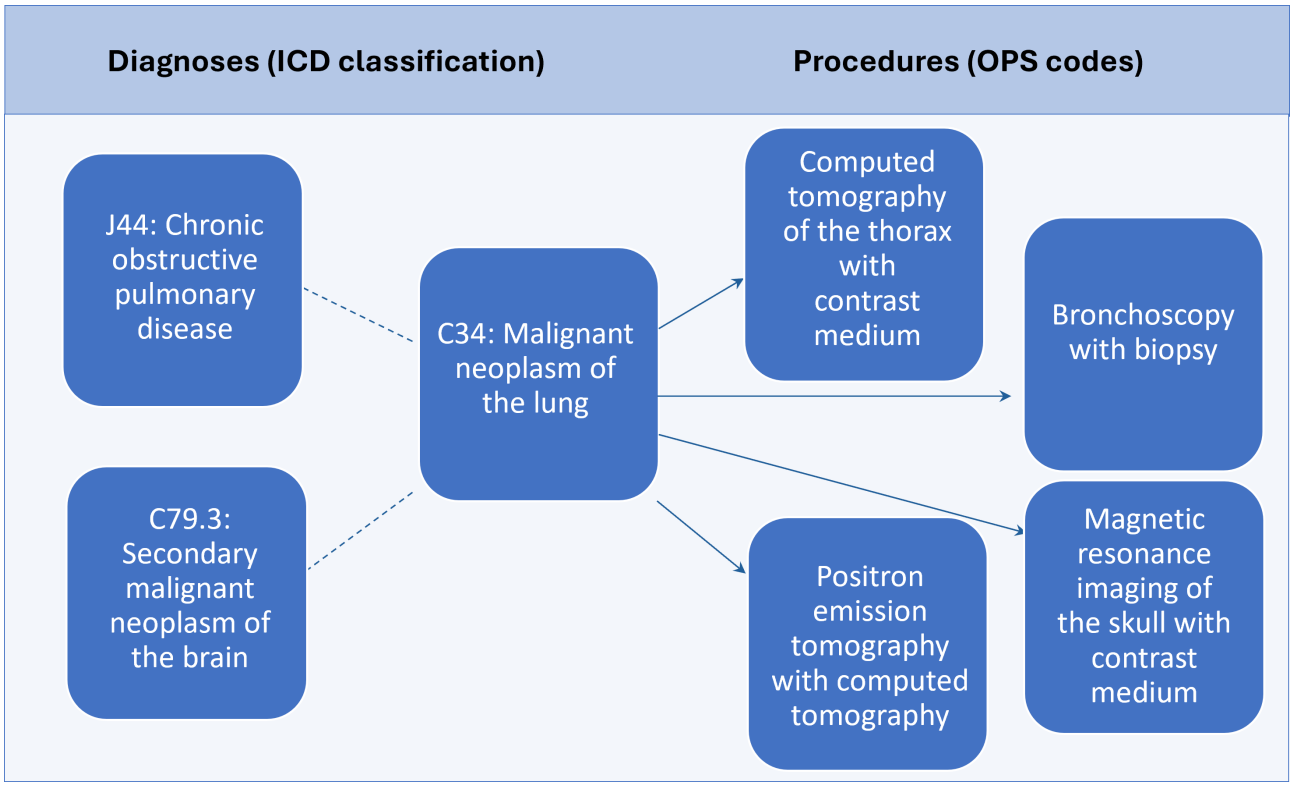

Limitations of Graph-Based Methods for PCQC

As the graph grows in size, querying and updating can become increasingly resource-intensive, potentially impacting the model's performance. In addition, algorithms are often based on the message-passing paradigm, which consists of node-by-node iteration (where each node aggregates information from neighboring nodes), which has significant limitations for graphs consisting of highly connected nodes [49]. For this reason, graphs with explicit heterogeneous edge distributions pose a problem and are difficult to evaluate because of the excessive compression of exponentially growing information in fixed-size vectors [50]. Recently, methods based on a different approach than the node-centered approach, which is based on differential geometry, have shown promising results for overcoming graph bottlenecks [51].

KGs often have incomplete data. If the graph does not have all the relevant entities and relationships, the model may lack important information. Incorrect or outdated information can lead to erroneous insights and conclusions.

HDTs and Health Digital Shadows for Modeling Procedure Efficacy

To model procedure efficacy, it is necessary to analyze the system's (ie, patient's) response to an external perturbation (eg, administration of a drug). It is crucial to gain deep insights into a patient's condition to assess the efficacy of procedures. Here, we want to present how these modeling tasks can be accomplished by HDTs [52] and by health digital shadows (HDSs), specifically when using graph-based methods.

For modeling procedure efficacy, the targets are not only binary values indicating patient outcomes or medical procedures but also biomarkers recorded as time series. For example, it is currently not possible to predict in advance whether a patient will face an allergic reaction or intolerance to medication. However, the ability to predict such events is relevant, as these reactions can be life-threatening and lead to treatment delays or morbidity.

Another problem is that physicians lack the time to review all available medical records of a patient. In the case of complex diseases, such as cancer, with various genetic mutations that could change during the procedure, prior or regular modeling is needed. Procedure adaptations or medication adjustments can be made in cases of side effects or low procedure efficacy.

For example, genetic polymorphisms in drug metabolism are already known for some substances, which must be determined before the substance is administered (eg, dihydropyrimidine dehydrogenase testing before 5-fluorouracil administration). Possible drug interactions must also be considered if several diseases are present. Many patients take several medications, which can sometimes lead to unforeseeable interactions. For patients, these interactions are increasingly difficult to oversee and evaluate, and individual factors also play a role in tolerability. The combination of different target values, along with different time spans, makes the problem definition much more complex than for patient safety and procedure accuracy.

XSL•FO
RenderX

Hence, an algorithmic architecture capable of considering a wide range of patient information and potential interactions would help model patient responses to medical procedures, predict patient outcomes, and test different counterfactual scenarios (ie, estimating a patient's outcome on the basis of a treatment regime that is distinct from the one that was actually followed and the observed outcome) [53,54]. Therefore, we model the integration of different procedures within a process. Technically speaking, this task represents the need for the integration of various models along a process (ie, patient's journey).

For this modeling task, the concept of a DT, which originated from engineering sciences [55], can be adapted into a clinical context. The aim of a DT is to generate a digital representation of an existing object within a process or time evolution in a digital space. Its concept is designed to combine time-evolving information from real objects, such as data generated from the Internet of Things, and mathematical modeling in the digital space [56]. Therefore, this digital representation allows the adaptation or optimization of processes in a digital space before transformations or modifications of real processes are made. For example, by integrating different and disparate information sources, it is possible to create different models not only for industrial production but also for the solution of critical and strategic problems, such as pollution, by generating universal avatars [57]. In summary, the design of a DT has the following main characteristics [56]:

1. Enablement of bidirectional communication between the model and real assets.
2. Integration of different information sources, such as genetic information, comorbidities, medication history, medical history, and family history.
3. Integration of different models, such as mechanistic models (eg, physiologically based pharmacokinetic modeling [PBPK] models [58]) or inductive models (statistical models such as structured nested mean models or ML models), at one time point or for one procedure [53].
4. Coupling of individual procedure models within a process, for instance, in a patient journey.
5. Simulation of procedures over time and process.
6. Model improvement through learning from the digital cohort (sum of all DTs).
7. Compliance with context-specific requirements arising from the field of application (eg, clinical health service).

In the industry, objects can be fully digitalized (with a bidirectional flow between the object and the digital object), which is limited only by the current evidence and understanding of the underlying complexity of the physical or biochemical processes or materials that should be mimicked. However, in the medical context, patients are not only too biologically complex but also cannot be objectivized in a straightforward way [59]. For this reason, we prefer to refer to HDSs, even when the literature currently mainly uses the term HDTs [60].

Furthermore, the connection between KGs and HDSs represents a logical step in the way different and disparate information sources are connected to create descriptive and predictive models. Relevant patient information as well as medical guidelines (in the form of KGs) can be integrated into a full graph containing patient data (eg, genetic or socioeconomic data) and additional relevant information (eg, drug structure and drug interactions).

In its role as an information and model integrating method, an HDS is well-suited for the following [61]:

- Generic representations: The predicted value of a quantity of interest is within the expected range of the global population. The representations are on an abstract and qualitative scale.
- Population-specific representations: The predicted value of a quantity of interest is evaluated with respect to a patient cohort. Such representations are more precise, enabling quantitative predictions.
- Specific representations: The predicted value of a quantity of interest has high accuracy in relation to the expected value for a single patient. These models are used for personalized predictions.

The HDS concept is interesting for prospective analysis, that is, in the simulation and prediction of a patient's health status and response to procedures. For prognostic purposes, it includes the ability to integrate all available patient information. Doctors treating patients can verify in advance the planned treatment regimen, such as medication or radiotherapy, on the patient's digital representation.

The HDS model runs through all the interactions between the procedure and the patient. In the case of pharmacological treatment, medication interactions are simulated (eg, by integrating ML models for drug-drug interactions [62]). Then, an individual recommendation for the substance and the individual dosage and medication can be obtained by applying appropriate mathematical models (ie, ML). Additionally, if the treatment response is not sufficient, the HDS could propose additional procedures or further treatment strategies. Current databases for potential drug targets could also be considered.

Figure 4 shows how all these steps are included in a closed information loop. Bidirectional communication between the real world and the model is enabled. Therefore, the general model obtains patient-specific information at time point t_i (element 1). This enables the human HDS to make a prognosis (element 2), which can prospectively be used by the clinician to adapt the medical decision (element 3). Alternatively, model prediction can be used retrospectively to evaluate medical decisions by examining patient outcomes at time points t_{i+1} (element 4). As time evolves, the patient's journey produces increasing patient-specific information, leading to better possible adaptation of the general model to a personalized human HDS. As the HDS cohort grows, model readjustments for the general model are performed after certain time periods (element 5).

XSL•FO
RenderX

Figure 4. Representation of a health digital shadow (HDS) as an information system. The HDS integrates patient data into a generally derived mathematical model (1) and allows model prognosis (2). Thus, the model supports the work of medical personnel as additional information for medical decision-making (3), which leads to a patient outcome at the next time point. In this way, bidirectional communication between the HDS and the real world is enabled. The model is also able to simulate the patient's journey, although the model's prognosis can be used to directly estimate the patient's outcome (4). The information gained by the individual human HDS can be used to readjust the general model constructed by the digital cohort (5).

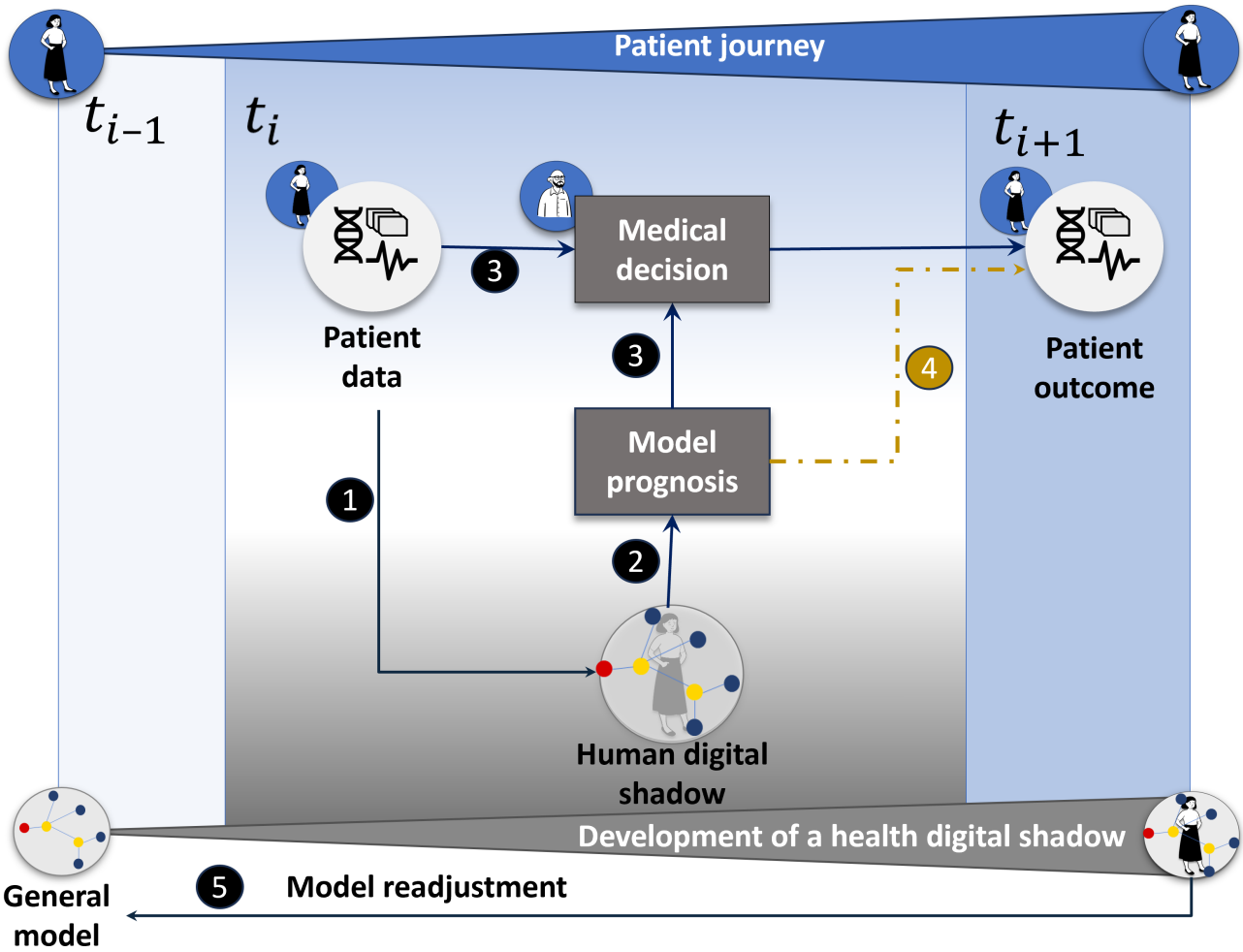

Thus, HDSs could increase the safety of drug therapy and its efficacy by providing personalized recommendations based on all available information [63]. For example, a similar concept, which is based on a whole-body representation via PBPK models and the integration of multiagent systems representing hepatocyte metabolism, can be used for pharmacology, particularly for the representation of the avatar's response to a medication [64,65]. This example demonstrates that through the analysis of the avatar's response, it is possible to plan procedures and even consider individual patient characteristics (eg, personalized metabolite levels that depend on the patient's genetics) that lead to personalized patient care.

This example implies that procedures can be tested on the DHS to forecast procedure success; therefore, the recommendation level of a procedure can be assessed before being applied to a real patient. Such an application is equivalent to the so-called dynamic HDT, where forecasts based on time series are performed [65].

Consequently, HDSs can be useful for modeling procedure efficacy, as in the following examples [66]:

- Human HDSs for inpatients:
  - We analyze which combination of medical procedures is most appropriate for patients depending on their socioeconomic context, environment, genetics, etc.
  - We analyze whether medical decisions will have an effective and desired outcome.
- Human HDSs for outpatients:
  - Help physicians recognize dangerous changes in a patient's health state through the analysis of patient data that can be delivered in real time (wearables or machines, for instance, in nephrology).

HDSs are increasingly being used for the estimation of the quality of patient care [55]. To this end, the prediction of patient outcomes is relevant, for example, in a personalized patient assessment to decide whether a procedure with high risk (death, side effects, patient burden, etc) is needed. This prevents the deterioration of a patient's quality of life. Figure 5 shows an exemplary representation of a human HDS that integrates different data sources (such as omics, EHR, and documentation), which are measured in different time periods along the patient journey, into different models to predict patient outcomes.

XSL•FO
RenderX

Figure 5. Schematic representation of a health digital shadow (HDS). An HDS is not a single model but a collection of different models to perform different predictions depending on the system's parameters of interest. Such predictions can include the analysis and prediction of time series at different time points of the patient's journey. In this image, we present a scheme of a patient's journey: after making the first medical decision on the basis of clinical guidelines (1), different information sources are integrated to predict quality metrics in different time periods depending on the patient's evolution (2). For example, the pharmacological response can be estimated via a whole-body model (eg, the physiologically based pharmacokinetic model [67]). Patient outcomes can be subsequently estimated via, for example, cluster analysis (3) [68]. EHR: electronic health record; HDT: health digital twin.

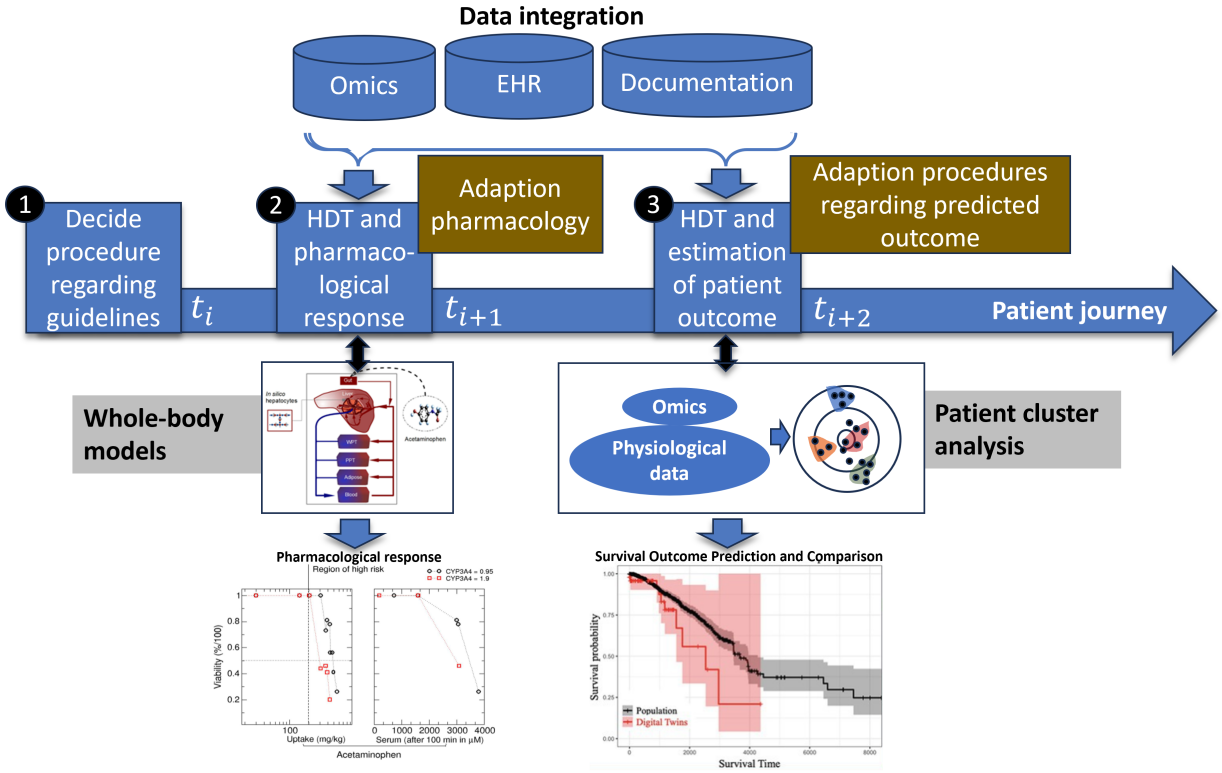

As shown in Figure 5, after choosing the best possible medical procedure according to clinical guidelines, different information sources are integrated to predict QMs in different time periods depending on the patient's evolution. For example, a whole-body (PBPK) model can be used to integrate patient-specific data within the HDT to predict a patient's pharmacological response. Using pharmacological response prediction and patient data, another model can estimate patient outcomes (eg, through cluster analysis) for survival time prediction.

Chang et al presented a remarkable example of possible DHS realization through the creation of a pipeline for predicting patient outcomes, particularly survival heterogeneity, for patients with breast cancer [68,69]. In this implementation, 3 different data sources were integrated, where low-dimensional embeddings of clinical and molecular features enriched by an external annotation database were the main data sources. In this algorithm, a dimensional reduction method, Uniform Manifold Approximation and Projection, is implemented. This is a general-purpose manifold learning and dimension reduction algorithm to reduce the dimension of the data space. A dynamic HDS requires a pipeline that is accordingly adapted (ie, new data and additional models must be integrated and appropriately adapted according to other outcomes) [68].

Nitschke et al [68] presented a general and unspecialized design that is applicable for HDSs and is predictive, modular, evolving, informed, interpretable, and explainable. This design combines KGs and ensemble learning to represent the patient's entire clinical journey and assist clinicians in their decision-making. Broad clinical application was shown by presenting 2 explicit cases: prostate cancer biopsy and glioma treatment decisions [68].

Limitations of HDSs for Modeling Procedure Efficacy

HDSs often oversimplify real systems. These approximations can lead to discrepancies between the model and the actual system. The accuracy of the HDS depends on the quality and granularity of the data used. Since more than one model is coupled to generate a full HDS, different models trained on different data and granularities can lead to unbalanced representations. Finally, poor or sparse data can lead to unreliable models [70]. Notably, all these problems are exacerbated by issues related to the aggregation and integration of different types of data that occur in health care, along with the strict privacy policies that must be followed [71], a problem that may be alleviated with the introduction of EHRs [72].

Creating and maintaining an HDS can be computationally intensive, requiring significant resources for simulation and real-time data processing. Integrating an HDS with other

XSL•FO
RenderX

systems, sensors, and data sources can be complex and may require sophisticated interfaces and protocols. Ensuring that the HDS reflects real-time changes in the physical system can be challenging, especially in systems with high dynamics [73].

An interesting option is the relation and validation of HDSs using in vitro methods, such as organ-on-a-chip. This option opens the possibility to compensate for potential defects in model predictions and identify their potential problems by validating and comparing results with advanced in vitro patient representations [74]. This, however, is still a challenge in the field of oncology and requires more research to obtain a correct representation of patient outcomes.

Example of the Implementation of KGs and HDSs

The clinical scenario in Textbox 1 illustrates the integration of KGs and HDSs and demonstrates how our proposed framework evaluates all 3 QM categories while simultaneously assessing treatment safety in patients with comorbidities, ensuring guideline accuracy while allowing for personalized adaptation, and optimizing efficacy through predictive modeling of multiple therapeutic pathways for patients with non–small cell lung cancer. This case highlights how KG/HDS systems add value beyond static guidelines by dynamically integrating genomic data, resistance patterns, and real-world evidence to support complex clinical decision-making and establish prospective monitoring protocols.

Textbox 1. Clinical scenario.

A 64-year-old never-smoker with advanced lung adenocarcinoma undergoes comprehensive genomic profiling. Data: electronic health record showing diagnosis code C34.1 with brain metastases present; next-generation sequencing results showing EGFR L858R primary, T790M resistance mutation, and concurrent TP53; PD-L1 tumor proportion score showing a result of 2%; and baseline liver function test showing an alanine aminotransferase level of 45 U/L. Knowledge graph identifies treatment hierarchy according to National Comprehensive Cancer Network guidelines and maps resistance patterns to subsequent therapy options. Health digital shadow predicts the intracranial response rate with osimertinib (efficacy quality metric [QM]=0.70), hepatotoxicity probability (safety QM=0.18), and time to next progression (median 11.2 months). Accuracy assessment flags missing baseline brain magnetic resonance imaging according to the guidelines (accuracy QM=0.85). The system recommends osimertinib 80 mg daily with a central nervous system–specific imaging schedule, hepatic monitoring every 2 weeks initially, and liquid biopsy at 3 months for emerging resistance detection. The model also preidentifies next-line options based on likely resistance mechanisms, enabling proactive treatment planning.

Issues and Future Research Directions

Our analysis revealed that patient safety, the accuracy of the procedure, and the effectiveness of the procedure are paramount. Using the present taxonomy, we can define the types of challenges and the type of ML implementation to be performed. However, this taxonomy is far from perfect. For example, distinguishing between a diagnostic procedure and a therapeutic procedure may require a more specialized and precise taxonomy. In addition, the metrics required to evaluate patient outcomes and estimate the QI may not always be perfectly calibrated [75], which can limit the applicability of such methods. In general, patient outcomes are still highly debated [76]. Despite this, we believe that it is important to keep any classification and taxonomy as simple as possible to facilitate the implementation of mathematical models.

We concluded that holistic quality management analysis (especially regarding procedure efficacy) requires the integration of several data sources. This is because there is no single decision (in decision support) or event (assignment of a diagnosis), but whole processes impact the patient's health status. Quality management and assessment can be very challenging for model implementation since it requires the integration of a constant information flow, which can continuously change within the process. As a possible solution to address these issues, we have introduced the concept of the human HDS.

Guidelines serve as guidance to define and implement processes necessary for the management of patients' health status. They are the result of clinical studies and are intended to help doctors make decisions. However, they contain complex workflows that should be evaluated individually. While a deviation from a guideline may be a potential problem in disease management for one patient because a step in the workflow has not been taken into account, which is a reduction in the quality of the hospital (ie, a decrease in procedural accuracy), for other patients, this deviation can be seen as a necessary adjustment needed for more personalized health care (ie, increasing the effectiveness of the procedure). In this context, especially in the evaluation of personalized practices, ML methods can influence the way guidelines are applied. Furthermore, we are cautious about full model–assisted personalization as well. Some critics have reported that data integration and target modeling have the potential to move away from real personalized patient centers and more human medicine [77]. In this case, individual modeling, which is embedded in a full modeling lifecycle, as shown in Multimedia Appendix 4, is just a way to quantitatively evaluate guidelines and their potential effects on the expected quality management of the disease at the patient level.

In this framework, graph representation is appealing, given that graph models are essentially an integration of data and information. Furthermore, when time is considered, it is possible to generate a digital representation of the patient (ie, an HDS). Furthermore, this concept has considerable potential for both inpatients and outpatients. This assumes that data generated by portable devices can track the patient's condition at home. Furthermore, this concept can be coupled with digitally stored patient journeys (ie, the digital cohort) [78].

Currently, an HDS essentially involves the integration of different models within a process. We believe that in the future, the application of KGs will contribute to the rapid development of HDSs. For example, recent work has suggested that the combination of KGs and LNNs could be the key to the implementation of productive avatars in medicine [79].

Novel decentralized data storage concepts may help train and define accurate models with sufficient data volumes. For

XSL•FO
RenderX

example, by integrating different and disparate information sources and real-world data built on the FAIR (findability, accessibility, interoperability, and reusability) principle [80], some of which are interconnected through the semantic web (according to the Worldwide Web 3.0 standards, for example, semantic web, such as Solid [81]), it is possible to create models not only for industrial production but also for the solution of critical and strategic problems by generating universal avatars [57]. Several recent initiatives are developing such ideas to implement DTs on the basis of semantic web principles (see, for instance, the EDITH-CSA project [61]). In such initiatives, it is remarkable to find not only the definition and exploration of models but also the definition of data standards required for data interoperability, most of which are based on Fast Healthcare Interoperability Resource (FHIR) standards.

HDSs face additional challenges. In the event of poor data safety standards, criminals could gain access to medical records and HDSs for blackmail or ransom demands. Companies, such as health insurance or disability insurance, might also demand the patient's HDS prognosis before enrollment. Patients could be denied insurance, insurance rates could be increased, or existing insurance conditions could be adapted regularly on the basis of HDSs. This could ultimately lead to the selection of people who will not receive insurance at all or who will receive only reduced coverage. Institutions and employers could also demand an HDS model before hiring people in order to select healthy people.

Another relevant issue is data contamination by ML models. For example, if a machine catches early signs of sepsis and doctors treat it, this creates a "contaminated association" in the data. Furthermore, an intervention triggered by one model can quietly disrupt another, even if they are focused on entirely different outcomes [82]. This can be particularly critical for HDTs. A potential solution is to preserve clean data from datasets where predictive models are not used.

The level of mandatory treatment recommendations based on HDSs must be discussed and assessed. Patients' wishes are crucial. What happens if the physician or patient deviates from an HDS's recommendations? In this case, the insurance company may not cover the alternative treatments since the HDS does not predict them to be most effective. This question is related to the overall reliability of HDSs. In retrospect, the accuracy of such recommendations can be assessed (eg, by checking procedure efficacy and tolerability).

In summary, maintaining a healthy balance between the sense of euphoria and the concern for validity in the implementation of mathematical models is imperative. Graph models require clean labels, accurate graph structures, and sufficient initial node features; however, real-world graph data often have noise and sparse labels, whereas different datasets have distinct feature constructions [83]. Furthermore, DHSs often oversimplify the systems they represent, are challenged by in vivo conditions, and are difficult to maintain.

Recently, foundation large x models (LxMs) trained on large databases (eg, LLMs) have accelerated the implementation of most mathematical models while reducing development costs. However, these models are expensive and difficult to generalize (even LLMs have limited applicability). In such cases, a balance between LxMs (for instance, LLMs) and tailored models (for instance, BERT [Bidirectional Encoder Representations from Transformers]) can be the most suitable, cost-effective, and safe solution to all these challenges.

Implications of the Study for Practice

The following steps and checks should help minimize some of the limitations of KGs applied to PCQC:

- Define a realistic system size that can be represented by KGs and maintain a critical view of the overscaling and size of KGs.
- Define aggregation functions to derive partial graph folding and compensate for incomplete node categories [51].
- Keep the graph updated with new knowledge and ensure consistency across different sources [84].
- Define appropriate data structures, such as Resource Description Frameworks, to make KGs interoperable.

Similarly, the following checklist can guide the correct implementation of HDSs applied to PCQC:

- Ensure data harmonization and minimization/access control of different data sources and digital records.
- Implement human-in-the-loop overriding for the correct interpretation of HDSs and avoidance of problems such as alert fatigue.
- Use drift detection/recertification and calibration standards (eg, using organ-on-a-chip) for safety in QM implementation on HDSs.
- Implement staged rollouts (eg, silent trials) before HDS activation.

Finally, the following are the key findings and conclusions derived throughout this article:

- Patient outcomes are not abstract quality terms since they can be formalized and linked to measurable clinical processes.
- Medical guidelines are essential but often underutilized in quality monitoring.
- Real-world clinical decisions are complex and require personalization.
- KGs and GNNs help bridge fragmented clinical data.
- HDSs offer a vision for prospective, patient-specific quality evaluation.
- Quality management should shift from static checklists to dynamic, data-driven guidance.

Conclusion

The integration of KGs and DTs into health care holds transformative potential for improving clinical outcomes and the quality of care. However, this optimistic view is feasible only when the technologies are implemented in an ethical way, avoiding blind customer overconfidence in the delivered results. Furthermore, implementation requires accounting for the perspectives of stakeholders (as well as their own cultural environment and background): What is considered a desirable outcome from a patient-centered perspective (eg, symptom relief

XSL•FO
RenderX

and preserved quality of life) may diverge from institutional or systemic goals such as shorter hospital stay or cost efficiency.

As such, mathematical models for health care quality must clearly define whether the primary focus lies in maximizing individual patient well-being, the operational performance of the health care system, or economic sustainability. These dimensions are not mutually exclusive, but the prioritization among them fundamentally shapes the modeling framework and interpretation of safety-related events. From this perspective, we also provide concrete checklists and recommendations for the correct deployment of models, for example, for the generation of KGs or the deployment of HDSs.

The combination of patient outcomes, procedures, and guidelines is not only interesting from a technical point of view but also extremely relevant for improving the quality of hospitals and health centers. The integration of guidelines implies the application of ML in different ways, from alerting and predicting critical conditions to the application of similar methods used in fraud detection to recognize potential deviations from guidelines.

The topic of quality management in medicine and the implementation of more detailed mathematical definitions are expected to become as relevant as diagnosis and image recognition are today, and we expect an increasing interest in further ML applications for quality management in the next few years.

Acknowledgments

AKN would like to thank Matthias Weidemüller for insightful and inspiring discussions, as well as for the invaluable support provided throughout this work. Additionally, AKN would like to thank Fabian Egersdörfer and Carlos Brandl for proofreading the text and providing recommendations. JGDO would like to thank Faizan E Mustafa for inspiring discussions and providing support in the research about graph knowledge, and Elena Ramirez for contributing to the assessment of the relationship of medical decisions in the framework of the theory of rational decision-making. Moreover, JGDO would like to thank Felix Weil for providing constant support and MaryLena Bleile for pointing out the relationship between health digital shadows and testing different counterfactual scenarios. We are grateful to the editor and reviewers who provided comments and helped us improve this article. No identifiable patient-level data were accessed or analyzed in this conceptual viewpoint.

All the ideas, texts, and figures were generated by the authors. No artificial intelligence tool was used to generate substantive content, and only language polishing was supported by the use of such a tool. According to the STM taxonomy [85], the artificial intelligence tool used for the preparation of this manuscript belongs to category 1: language refinement was carried out via Rubriq to detect spelling and typographical errors.

Funding

This project was supported by the Ministry for Economics, Labor and Tourism from Baden-Württemberg, Germany via grant agreement number BW1_1456 (AI4MedCode).This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2181/1 - 390900948 (the Heidelberg STRUCTURES Excellence Cluster). This project was also supported by broad funding to the Klinikum Stuttgart, Stuttgart Cancer Center – Tumorzentrum Eva Mayr-Stihl, from the Eva Mayr-Stihl Foundation, Waiblingen, Germany.

Authors' Contributions

The initial idea was proposed and drafted by AKN and JGDO. JGDO contributed to the overview of the knowledge graphs. SN, AKN, and JGDO contributed to the digital twin overview. SN and MK contributed to the medical background, examples, and critical evaluation of the various methods described in this article. All the authors contributed to the final version of this article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Insight into the different aspects of the assessment of patient-centered quality of care through possible mathematical modeling applications.
[DOCX File, 21 KB - jmir_v28i1e81946_app1.docx]

Multimedia Appendix 2
Analysis of the available literature on PubMed.
[DOCX File, 17 KB - jmir_v28i1e81946_app2.docx]

Multimedia Appendix 3
Literature overview of the application of graph models for patient safety, procedure accuracy, and procedure efficacy.

XSL•FO
RenderX

[DOCX File, 31 KB - jmir_v28i1e81946_app3.docx]

Multimedia Appendix 4
Lifecycle of the data and models in relation to clinical guidelines.
[DOCX File, 530 KB - jmir_v28i1e81946_app4.docx]

References

1. McLachlan S, Kyrimi E, Dube K, Hitman G, Simmonds J, Fenton N. Towards standardisation of evidence-based clinical care process specifications. Health Informatics J 2020 Dec;26(4):2512-2537. [doi: 10.1177/1460458220906069] [Medline: 32186428]
2. Timmermans S, Mauck A. The promises and pitfalls of evidence-based medicine. Health Aff (Millwood) 2005;24(1):18-28. [doi: 10.1377/hlthaff.24.1.18] [Medline: 15647212]
3. Sinsky CA, Bavafa H, Roberts RG, Beasley JW. Standardization vs customization: finding the right balance. Ann Fam Med 2021;19(2):171-177. [doi: 10.1370/afm.2654] [Medline: 33685879]
4. Alexopoulos K, Chryssolouris G. Process. In: Chatti S, Laperrière L, Reinhart G, Tolio T, editors. CIRP Encyclopedia of Production Engineering: Springer; 2019:1349-1352. [doi: 10.1007/978-3-662-53120-4_6567]
5. Da Silva N. The industrialization of healthcare and its critiques. In: Diaz-Bone R, Larquier G, editors. Handbook of Economics and Sociology of Conventions: Springer; 2023:1-25. [doi: 10.1007/978-3-030-52130-1_45-2]
6. Dodwad SS. Quality management in healthcare. Indian J Public Health 2013;57(3):138-143. [doi: 10.4103/0019-557X.119814] [Medline: 24125927]
7. Rojas E, Munoz-Gama J, Sepúlveda M, Capurro D. Process mining in healthcare: a literature review. J Biomed Inform 2016 Jun;61:224-236. [doi: 10.1016/j.jbi.2016.04.007] [Medline: 27109932]
8. May M. Eight ways machine learning is assisting medicine. Nat Med 2021 Jan;27(1):2-3. [doi: 10.1038/s41591-020-01197-2] [Medline: 33442003]
9. Bach JF. Causality in medicine. C R Biol 2019;342(3-4):55-57. [doi: 10.1016/j.crvi.2019.03.001] [Medline: 30981720]
10. Diaz Ochoa JG. A unified method for assessing the observability of dynamic complex systems. Comput Biol Med 2023 Jun;160:107012. [doi: 10.1016/j.compbiomed.2023.107012] [Medline: 37187137]
11. Ensure healthy lives and promote well-being for all at all ages. United Nations. URL: https://sdgs.un.org/goals/goal3 [accessed 2025-08-01]
12. Steele K, Stefánsson HO. Decision theory. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy (Winter 2020 Edition): Metaphysics Research Lab, Stanford University; 2020. URL: https://plato.stanford.edu/archives/win2020/entries/decision-theory/ [accessed 2024-05-12]
13. Epstein RM, Street RL. The values and value of patient-centered care. Ann Fam Med 2011;9(2):100-103. [doi: 10.1370/afm.1239] [Medline: 21403134]
14. Kersting C, Kneer M, Barzel A. Patient-relevant outcomes: what are we talking about? A scoping review to improve conceptual clarity. BMC Health Serv Res 2020 Jun 29;20(1):596. [doi: 10.1186/s12913-020-05442-9] [Medline: 32600321]
15. Liu Y, Avant KC, Aungsuroch Y, Zhang XY, Jiang P. Patient outcomes in the field of nursing: a concept analysis. Int J Nurs Sci 2014 Mar;1(1):69-74. [doi: 10.1016/j.ijnss.2014.02.006]
16. Hadian SA, Rezayatmand R, Shaarbafchizadeh N, Ketabi S, Pourghaderi AR. Hospital performance evaluation indicators: a scoping review. BMC Health Serv Res 2024 May 1;24(1):561. [doi: 10.1186/s12913-024-10940-1] [Medline: 38693562]
17. Wang XS, Gottumukkala V. Patient-reported outcomes: is this the missing link in patient-centered perioperative care? Best Pract Res Clin Anaesthesiol 2021 Dec;35(4):565-573. [doi: 10.1016/j.bpa.2020.10.006] [Medline: 34801218]
18. Busse R, Klazinga N, Panteli D. In: Quentin W, editor. Improving Healthcare Quality in Europe: Characteristics, Effectiveness and Implementation of Different Strategies: European Observatory on Health Systems and Policies; 2019.
19. Dovetail Editorial Team. Patient outcomes: overview, measures & ways to improve. Dovetail. 2023. URL: https://dovetail.com/patient-experience/what-are-patient-outcomes/ [accessed 2025-08-06]
20. McLeod C, Norman R, Litton E, Saville BR, Webb S, Snelling TL. Choosing primary endpoints for clinical trials of health care interventions. Contemp Clin Trials Commun 2019 Dec;16:100486. [doi: 10.1016/j.conctc.2019.100486] [Medline: 31799474]
21. Yanamadala S, Morrison D, Curtin C, McDonald K, Hernandez-Boussard T. Electronic health records and quality of care. Medicine (Baltimore) 2016;95(19):e3332. [doi: 10.1097/MD.0000000000003332] [Medline: 27175631]
22. Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e V [Website in German]. URL: https://www.awmf.org/ [accessed 2025-08-06]
23. S3-Leitlinie Prostatakarzinom [Article in German]. AWMF online. URL: https://register.awmf.org/de/leitlinien/detail/043-022OL [accessed 2025-08-06]
24. Langer T, Follmann M. Das Leitlinienprogramm Onkologie (OL): Nukleus einer evidenzbasierten, patientenorientieren, interdisziplinären Onkologie? [Article in German]. Z Evid Fortbild Qual Gesundhwes 2015;109(6):437-444. [doi: 10.1016/j.zefq.2015.09.007] [Medline: 26474648]

XSL•FO
RenderX

25. Schick J. IQM Positionspapier Ergebnisqualität [Article in German]. Initiative Qualitätsmedizin e V. 2025. URL: https://www.initiative-qualitaetsmedizin.de/positionspapier [accessed 2025-08-06]
26. Schmitt J, Klinkhammer-Schalke M, Bierbaum V, et al. Initial cancer treatment in certified versus non-certified hospitals. Dtsch Arztebl Int 2023 Sep 29;120(39):647-654. [doi: 10.3238/arztebl.m2023.0169] [Medline: 37583089]
27. Gigerenzer G, Gray JAM. Bessere Ärzte, Bessere Patienten, Bessere Medizin Aufbruch in Ein Transparentes Gesundheitswesen: Mit Einem Vorwort von Günther Jonitz Strüngmann [Book in German]: Medizinisch Wissenschaftliche Verlagsgesellschaft; 2013.
28. Peng J, Jury EC, Dönnes P, Ciurtin C. Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: applications and challenges. Front Pharmacol 2021;12:720694. [doi: 10.3389/fphar.2021.720694] [Medline: 34658859]
29. Emanuel L, Berwick D, Conway J, et al. What exactly is patient safety? J Med Regul 2009 Mar;95(1):13-24. [doi: 10.30770/2572-1852-95.1.13]
30. Agency for Healthcare Research and Quality. URL: https://www.ahrq.gov/ [accessed 2025-08-06]
31. Nimptsch U, Mansky T, Busse R. G-IQI – German Inpatient Quality Indicators Version 54: Universitätsverlag der TU Berlin; 2022.
32. Zaiss A, Dauben HP. ICHI-International Classification of Health Interventions: a balancing act between the demands of statistics and reimbursement [Article in German]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2018 Jul;61(7):778-786. [doi: 10.1007/s00103-018-2747-6] [Medline: 29808283]
33. Decision-making as reasoning. In: Cognitive Economics: Springer; 2008:49-67. [doi: 10.1007/978-3-540-71347-0_4]
34. Lafonte M, Cai J, Lissauer ME. Failure to rescue in the surgical patient: a review. Curr Opin Crit Care 2019 Dec;25(6):706-711. [doi: 10.1097/MCC.0000000000000667] [Medline: 31567517]
35. Greenes RA. Clinical Decision Support: The Road to Broad Adoption, 2nd edition: Academic Press; 2014.
36. Barth JH, Misra S, Aakre KM, et al. Why are clinical practice guidelines not followed? Clin Chem Lab Med 2016 Jul 1;54(7):1133-1139. [doi: 10.1515/cclm-2015-0871] [Medline: 26650076]
37. Börm P. Leitlinienbasierter Clinical Decision Support – Anforderungen an evidenzbasierte Entscheidungsunterstützungssysteme [Article in German]. OP-Journal 2021 Apr;37(1):28-35. [doi: 10.1055/a-1284-3193]
38. Diaz Ochoa JG, Mustafa FE. Graph neural network modelling as a potentially effective method for predicting and analyzing procedures based on patients' diagnoses. Artif Intell Med 2022 Sep;131:102359. [doi: 10.1016/j.artmed.2022.102359] [Medline: 36100347]
39. Becker M, Kasper S, Böckmann B, Jöckel KH, Virchow I. Natural language processing of German clinical colorectal cancer notes for guideline-based treatment evaluation. Int J Med Inform 2019 Jul;127:141-146. [doi: 10.1016/j.ijmedinf.2019.04.022] [Medline: 31128826]
40. Alharbe N, Rakrouki MA, Aljohani A. A healthcare quality assessment model based on outlier detection algorithm. Processes (Basel) 2022 Jun;10(6):1199. [doi: 10.3390/pr10061199]
41. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
42. Lynch SS. Evaluating drug efficacy and safety. MSD Manuals. URL: https://www.msdmanuals.com/professional/clinical-pharmacology/concepts-in-pharmacotherapy/drug-efficacy-and-safety [accessed 2024-08-08]
43. Kline A, Wang H, Li Y, et al. Multimodal machine learning in precision health: a scoping review. NPJ Digit Med 2022 Nov 7;5(1):171. [doi: 10.1038/s41746-022-00712-8] [Medline: 36344814]
44. Si Y, Du J, Li Z, et al. Deep representation learning of patient data from electronic health records (EHR): a systematic review. J Biomed Inform 2021 Mar;115:103671. [doi: 10.1016/j.jbi.2020.103671] [Medline: 33387683]
45. Liu M, Zhang F, Huang P, Niu S, Ma F, Zhang J. Learning the satisfiability of pseudo-boolean problem with graph neural networks. In: Simonis H, editor. Principles and Practice of Constraint Programming CP 2020 Lecture Notes in Computer Science, Vol 12333: Springer; 2020:885-898. [doi: 10.1007/978-3-030-58475-7_51]
46. Abu-Salih B, Al-Qurishi M, Alweshah M, Al-Smadi M, Alfayez R, Saadeh H. Healthcare knowledge graph construction: a systematic review of the state-of-the-art, open issues, and opportunities. J Big Data 2023;10(1):81. [doi: 10.1186/s40537-023-00774-9] [Medline: 37274445]
47. Zheng Z, Guo C, Chen J, Li J. Graph neural network-based representation learning for medical time series. In: Iliadis L, Papaleonidas A, Angelov P, Jayne C, editors. Artificial Neural Networks and Machine Learning – ICANN 2023: Springer Nature; 2023:194-205. [doi: 10.1007/978-3-031-44223-0_16]
48. S3-Leitlinie Prävention, Diagnostik, Therapie und Nachsorge des Lungenkarzinoms [Report in German]. : Leitlinienprogramm Onkologie; 2024 URL: https://www.leitlinienprogramm-onkologie.de/fileadmin/user_upload/Downloads/Leitlinien/Lungenkarzinom/Version_3/LL_Lungenkarzinom_Langversion_3.0.pdf [accessed 2025-08-06]
49. Sato R. A survey on the expressive power of graph neural networks. arXiv. Preprint posted online on Mar 9, 2020. [doi: 10.48550/arXiv.2003.04078]
50. Alon U, Yahav E. On the bottleneck of graph neural networks and its practical implications. Open Review. 2021. URL: https://openreview.net/forum?id=i80OPhOCVH2 [accessed 2022-09-07]

XSL•FO
RenderX

51. Topping J, Di Giovanni F, Chamberlain BP, Dong X, Bronstein MM. Understanding over-squashing and bottlenecks on graphs via curvature. Open Review. 2022. URL: https://openreview.net/forum?id=7UmjRGzp-A [accessed 2025-12-01]
52. Kulkarni C, Quraishi A, Raparthi M, et al. Hybrid disease prediction approach leveraging digital twin and metaverse technologies for health consumer. BMC Med Inform Decis Mak 2024 Apr 5;24(1):92. [doi: 10.1186/s12911-024-02495-2] [Medline: 38575951]
53. Chakraborty B, Moodie EEM. Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine: Springer; 2013. [doi: 10.1007/978-1-4614-7428-9]
54. Chakraborty B, Moodie EEM. Estimation of optimal DTRs by directly modeling regimes. In: Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine: Springer; 2013:79-100. [doi: 10.1007/978-1-4614-7428-9_5]
55. Vallée A. Digital twin for healthcare systems. Front Digit Health 2023;5:1253050. [doi: 10.3389/fdgth.2023.1253050] [Medline: 37744683]
56. Richter J, Lange F, Scheper T, Solle D, Beutel S. Digitale Zwillinge in der Bioprozesstechnik – Chancen und Möglichkeiten [Article in German]. Chemie Ingenieur Technik 2023 Apr;95(4):498-510. [doi: 10.1002/cite.202200166]
57. Akroyd J, Mosbach S, Bhave A, Kraft M. Universal digital twin - a dynamic knowledge graph. Data-Centric Engineering 2021;2:e14. [doi: 10.1017/dce.2021.10]
58. Peters SA. Physiologically‐Based Pharmacokinetic (PBPK) Modeling and Simulations: Principles, Methods, and Applications in the Pharmaceutical Industry: John Wiley & Sons, Inc; 2012. [doi: 10.1002/9781118140291]
59. John A, Alhajj R, Rokne J. A systematic review of AI as a digital twin for prostate cancer care. Comput Methods Programs Biomed 2025 Aug;268:108804. [doi: 10.1016/j.cmpb.2025.108804] [Medline: 40347618]
60. Katsoulakis E, Wang Q, Wu H, et al. Digital twins for health: a scoping review. NPJ Digit Med 2024 Mar 22;7(1):77. [doi: 10.1038/s41746-024-01073-0] [Medline: 38519626]
61. Building the European virtual human twin. : Edith; 2023 URL: https://www.edith-csa.eu/wp-content/uploads/2023/04/EDITH_D3.1_vision_roadmap_outline_final.pdf [accessed 2025-08-06]
62. Mei S, Zhang K. A machine learning framework for predicting drug-drug interactions. Sci Rep 2021 Sep 2;11(1):17619. [doi: 10.1038/s41598-021-97193-8] [Medline: 34475500]
63. Diasio RB, Offer SM. Testing for dihydropyrimidine dehydrogenase deficiency to individualize 5-fluorouracil therapy. Cancers (Basel) 2022 Jun 30;14(13):3207. [doi: 10.3390/cancers14133207] [Medline: 35804978]
64. Diaz Ochoa JG, Bucher J, Niklas J, Mauch K. Multi-scale modeling for individualized spatiotemporal prediction of drug effects. Toxicol Lett 2013 Aug;221:S83-S84. [doi: 10.1016/j.toxlet.2013.05.093]
65. Mulder ST, Omidvari AH, Rueten-Budde AJ, et al. Dynamic digital twin: diagnosis, treatment, prediction, and prevention of disease during the life course. J Med Internet Res 2022 Sep 14;24(9):e35675. [doi: 10.2196/35675] [Medline: 36103220]
66. Sun T, He X, Li Z. Digital twin in healthcare: recent updates and challenges. Digit Health 2023;9:20552076221149651. [doi: 10.1177/20552076221149651] [Medline: 36636729]
67. Diaz Ochoa JG, Bucher J, Péry ARR, Zaldivar Comenges JM, Niklas J, Mauch K. A multi-scale modeling framework for individualized, spatiotemporal prediction of drug effects and toxicological risk. Front Pharmacol 2012;3:204. [doi: 10.3389/fphar.2012.00204] [Medline: 23346056]
68. Nitschke AK, Brandl C, Egersdörfer F, Görtz M, Hohenfellner M, Weidemüller M. Design for a digital twin in clinical patient care. NPJ Health Syst 2025;3(1):11. [doi: 10.1038/s44401-025-00060-1]
69. Chang HC, Gitau AM, Kothapalli S, Welch DR, Sardiu ME, McCoy MD. Understanding the need for digital twins' data in patient advocacy and forecasting oncology. Front Artif Intell 2023;6:1260361. [doi: 10.3389/frai.2023.1260361] [Medline: 38028666]
70. Margaria T, Ryan S. Data and data management in the context of digital twins. In: Crespi N, Drobot AT, Minerva R, editors. The Digital Twin: Springer International Publishing; 2023:253-278. [doi: 10.1007/978-3-031-21343-4_10]
71. Awrahman BJ, Aziz Fatah C, Hamaamin MY. A review of the role and challenges of big data in healthcare informatics and analytics. Comput Intell Neurosci 2022;2022:5317760. [doi: 10.1155/2022/5317760] [Medline: 36210978]
72. Venkatesh KP, Raza MM, Kvedar JC. Health digital twins as tools for precision medicine: considerations for computation, implementation, and regulation. NPJ Digit Med 2022 Sep 22;5(1):150. [doi: 10.1038/s41746-022-00694-7] [Medline: 36138125]
73. Tao F, Zhang H, Zhang C. Advancements and challenges of digital twins in industry. Nat Comput Sci 2024 Mar;4(3):169-177. [doi: 10.1038/s43588-024-00603-w] [Medline: 38532139]
74. Zhai X, Chen M, Esfahani SS, et al. Heterogeneous system-on-chip-based Lattice-Boltzmann visual simulation system. IEEE Systems Journal 2020;14(2):1592-1601. [doi: 10.1109/JSYST.2019.2952459]
75. Endo H, Uchino S, Hashimoto S, et al. Development and validation of the predictive risk of death model for adult patients admitted to intensive care units in Japan: an approach to improve the accuracy of healthcare quality measures. J Intensive Care 2021 Feb 15;9(1):18. [doi: 10.1186/s40560-021-00533-z] [Medline: 33588956]
76. Reiss J, Ankeny RA. Philosophy of medicine. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy (Spring 2022 Edition): Metaphysics Research Lab, Stanford University; 2022. URL: https://plato.stanford.edu/archives/spr2022/entries/medicine/ [accessed 2023-08-02]

XSL•FO
RenderX

77. Abettan C, Welie JVM. The impact of twenty-first century personalized medicine versus twenty-first century medicine's impact on personalization. Philos Ethics Humanit Med 2020 Nov 18;15(1):11. [doi: 10.1186/s13010-020-00095-2] [Medline: 33203434]
78. Böhler D, Friebe M. (Digital) patient journey and empowerment: digital twin. In: Friebe M, editor. Novel Innovation Design for the Future of Health: Springer; 2022:169-178. [doi: 10.1007/978-3-031-08191-0_17]
79. Nye L. Digital twins for patient care via knowledge graphs and closed-form continuous-time liquid neural networks. arXiv. Preprint posted online on Jul 8, 2023. [doi: 10.48550/arXiv.2307.04772]
80. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3:160018. [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
81. About Solid. Solid. URL: https://solidproject.org/about [accessed 2025-08-06]
82. Vaid A. Medical AI can transform medicine - but only if we carefully track the data it touches. Nature New Biol 2025 Jun;642(8069):864-866. [doi: 10.1038/d41586-025-01946-8] [Medline: 40550881]
83. Huang C, Ren X, Tang J, Yin D, Chawla N. Large language models for graphs: progresses and directions. Presented at: ACM Web Conference 2024; May 13-17, 2024. [doi: 10.1145/3589335.3641251]
84. Lu Z, Liu Y, Wen G, Zhou B, Zhang W, Zhang J. Noise-resistant graph neural networks with manifold consistency and label consistency. Expert Syst Appl 2024 Jul;245:123120. [doi: 10.1016/j.eswa.2023.123120]
85. Do you allow the use of ChatGPT or other generative language models and how should this be reported? JMIR Publications. 2025. URL: https://support.jmir.org/hc/en-us/articles/13387268671771-Do-you-allow-the-use-of-ChatGPT-or-other-generative-language-models-and-how-should-this-be-reported [accessed 2026-01-05]

Abbreviations

AHRQ: Agency for Healthcare Research and Quality
BERT: Bidirectional Encoder Representations from Transformers
DKG: Deutsche Krebsgesellschaft (German Cancer Society)
DT: digital twin
EHR: electronic health record
GNN: graph neural network
HCP: health care practitioner
HDS: health digital shadow
HDT: health digital twin
ICD: *International Classification of Diseases*
ICHI: International Classification of Health Interventions
ICU: intensive care unit
KG: knowledge graph
LLM: large language model
LOS: length of stay
LxM: large x model
ML: machine learning
PBPK: physiologically based pharmacokinetic modeling
PCQC: patient-centered quality of care
QI: quality indicator
QM: quality metric

Edited by J Sarvestan; submitted 06.Aug.2025; peer-reviewed by A Shivanna, S Narayan, UK Chilakalapalli; revised version received 15.Jan.2026; accepted 16.Jan.2026; published 31.Mar.2026.

Please cite as:
Nitschke AK, Diaz Ochoa JG, Neumaier S, Knott M
From Knowledge Graphs to Digital Twins: Perspectives on Modeling Patient Outcomes for Health Care Quality Assessment
J Med Internet Res 2026;28:e81946
URL: https://www.jmir.org/2026/1/e81946
doi:10.2196/81946

© Anna-Katharina Nitschke, Juan G Diaz Ochoa, Simone Neumaier, Markus Knott. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.Mar.2026. This is an open-access article distributed under the terms of the Creative

XSL•FO
RenderX

Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Integrating Generative AI Into Patient-Centered Clinical Decision Support: Viewpoint on Research and Practice Considerations

Prashila Dullabh[1], MD; Courtney Zott[1], MPH; Nicole Gauthreaux[1], MPH; Caroline Peterson[1], MPH; Abigail Aronoff[1], MPH; Kistein Monkhouse[2], MPA; Dean F Sittig[3], PhD

[1]Health Sciences Department, NORC at the University of Chicago, 1828 L Street NW, 9th Floor, Washington, DC, United States

[2]Patient Orator Inc, New York, NY, United States

[3]Informatics Review LLC, Lake Oswego, OR, United States

Corresponding Author:
Prashila Dullabh, MD
Health Sciences Department, NORC at the University of Chicago, 1828 L Street NW, 9th Floor, Washington, DC, United States

Abstract

There is growing interest in understanding how generative artificial intelligence (GenAI) can support patients and caregivers in making informed health care decisions, known as patient-centered clinical decision support (PC CDS). In this viewpoint, we present example applications for GenAI-supported PC CDS for patients, caregivers, clinicians, and patient-clinician interactions and examine the opportunities, challenges, and potential solutions associated with these applications. We conducted a targeted document review of our work in the Agency for Healthcare Research and Quality's Clinical Decision Support Innovation Collaborative focusing on GenAI-enabled PC CDS, supplemented by snowball sampling and targeted searches to identify additional applications. Findings were refined and validated through solicited feedback from a 20-member multidisciplinary expert panel. Through our work, we highlight six critical needs that must be addressed to fully realize GenAI's potential in PC CDS: (1) engage and ensure representation of patients and caregivers in design and development; (2) build the science of effective PC CDS implementation to support patient engagement; (3) develop risk-based policies for when to use GenAI; (4) establish independent testing and vetting criteria; (5) periodically reassess to identify and address algorithmic drift and verify performance; and (6) establish policies to promote transparency and patient consent in the use of GenAI. Understanding the applications and their potential implications for health care quality is essential to further the beneficial, ethical, and safe development of GenAI-supported PC CDS.

(J Med Internet Res 2026;28:e81628) doi:10.2196/81628

KEYWORDS

generative artificial intelligence; clinical decision support systems; patient-centered care; patient involvement; artificial intelligence; AI; quality of health care

Introduction

The clinical landscape is experiencing rapid growth in artificial intelligence (AI), especially generative AI (GenAI) systems that learn the underlying structure of their training data to create new content based on those learned patterns, such as new text, images, and data prompts [1]. This technology has sparked interest in its potential to integrate new data and analytics into patient-centered clinical decision support (PC CDS). PC CDS encompasses decision-making tools that significantly incorporate patient-centered factors across 4 key dimensions: knowledge (evidence based on comparative effectiveness research and research findings on patient-centered outcomes), informed by data that are directly generated from patients (ie, patient-generated health data [PGHD], patient-reported outcomes [PROs], and/or preferences), delivery (engaging patients and caregivers across settings through patient portals, apps, and other digital tools), and use (facilitating shared decision-making [SDM], a process where patients, caregivers, and care teams share and discuss health information and patients' values and preferences to reach mutually acceptable health-related decisions) [2]. Studies have suggested that GenAI has the potential to impact health care delivery, facilitate more personalized care, influence patient outcomes, and reshape the clinician-patient and clinician-caregiver relationship [3-5]. At the same time, researchers have identified barriers to realizing GenAI's potential, including limited GenAI implementation frameworks that consider the full sociotechnical environment [6], variability in requirements for reporting and validating GenAI [7], and mixed perceptions about the utility and safety of the technology among clinicians and patients [8]. Additionally, while many articles have described GenAI's short-term impact on clinicians and health care systems, such as relieving clinician workload and improving operational efficiency, to our knowledge, relatively few focus on the decision-making tools that support patient-centered care or on tools where the patient is the user [9-11]. Because these tools

XSL•FO
RenderX

directly engage patients in understanding health information, managing their conditions, and participating in care decisions, they raise distinct expectations around health literacy, such as supporting patients' ability to obtain, process, and understand health information [12]; self-management [13], such as strengthening patients' problem-solving, decision-making, and action-planning skills; and SDM, such as enhancing patients' knowledge, awareness, and ability to cope and engage in collaborative decisions with their care teams [1,14]. There is a need to understand the various ways and intricacies of leveraging GenAI to engage patients and caregivers in their care and how these tools can enhance patient-clinician communications and support the delivery of evidence-based care.

This viewpoint summarizes the novel applications of GenAI in PC CDS through illustrative use cases that can provide a better understanding of its potential among patients, caregivers, clinicians, and PC CDS developers. On the basis of these illustrative use cases, we identified common themes related to the benefits and challenges of GenAI-supported PC CDS tools on health care quality and delivery, with the purpose of acknowledging their advantages while mitigating their potential risks. Drawing from the challenges, we discuss ways to advance trustworthy GenAI-based PC CDS to chart a path forward for research and practice.

Targeted Review Approach

We used 3 methods to identify the use cases.

First, we conducted an in-depth review of 4 reports from the Agency for Healthcare Research and Quality's Clinical Decision Support Innovation Collaborative that addressed AI applications, including a landscape assessment on using AI to scale PC CDS [15], a report on patient and caregiver perspectives on GenAI in PC CDS [16], and 2 assessments of AI-supported PC CDS tools [17,18].

Second, we used a snowball sampling approach to identify additional relevant literature from the reference lists and conducted targeted searches in the PubMed and Google Scholar databases. In total, we identified 53 peer-reviewed sources. From these sources, we abstracted the type of AI technology used, the specific use cases, and the intended end users of these systems.

Third, we solicited feedback from a 20-member Steering Committee including patients, clinicians, informaticians, researchers, electronic health record (EHR) developers, payers, and policymakers. After the research team (PD, CZ, NG, and AA) completed the targeted review analysis and developed a draft use case exhibit, the lead author presented the preliminary findings to the Steering Committee and invited feedback on additional GenAI-supported PC CDS use cases to consider, the categorization of use cases, the validation of the identified benefits and considerations, and recommendations regarding critical needs for the effective use of GenAI-supported PC CDS. Meeting minutes documented key points of agreement, which were subsequently incorporated by the research team.

Use Cases for GenAI-Supported PC CDS

Figure 1 depicts use cases for GenAI-supported PC CDS that operate across categories based on their primary users and functions. Before reaching end users, GenAI can streamline the technical development of PC CDS for developers and implementers by optimizing knowledge artifacts and reducing development time and cost [15,19]. For end users, patient- and caregiver-focused tools emphasize health and well-being, chronic condition self-management, and educational resources. Tools supporting patient-clinician interactions facilitate communication through conversational agents, symptom monitoring systems, and treatment personalization for patients, which are discussed through SDM with their clinician. Clinician-focused applications assist with information management, particularly for analyzing large volumes of PGHD, while also providing diagnostic support and risk prediction. The potential impact and clinical significance of these use cases will vary across specialties, care settings, and patient populations, and the figure is not intended to imply a hierarchy of importance among them.

XSL•FO
RenderX

Figure 1. Use cases of generative artificial intelligence in patient-centered clinical decision support (PC CDS) technology. PGHD: patient-generated health data.

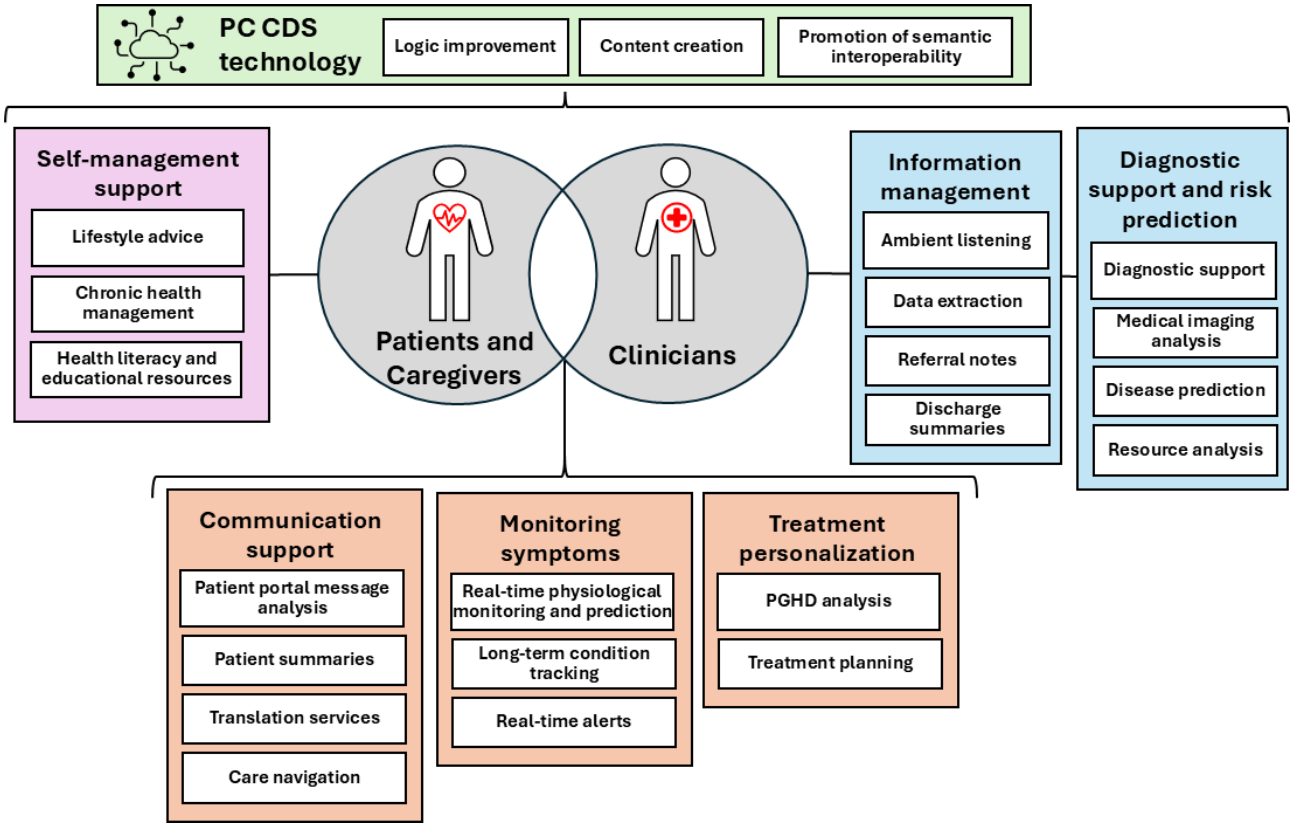

Salient Benefits and Considerations on the Use of GenAI in PC CDS

Overview

The use cases for GenAI-supported PC CDS illuminate several crosscutting themes that highlight the benefits and challenges of implementing these systems in health care settings. The themes do not apply uniformly across all use cases but rather vary in relevance depending on factors such as the clinical context, the decisions being made, and the patient-clinician interactions involved. Some themes also naturally intersect. For example, challenges such as hallucinations and misinformation cut across multiple applications and mediate a range of GenAI's benefits. In the following discussion, the examples specify the applicable use cases and raise challenges where they are highly relevant, illustrating how these themes can manifest across different applications.

Personalization and Precision

GenAI can support the delivery of personalized care that is responsive to a patient's unique health profile. GenAI is equipped to rapidly process extensive amounts of patient-centered health information from a range of sources, such as patients' EHRs, genomic data, medical devices, wearable technologies, and questionnaires assessing patients' preferences and goals [20]. It can then leverage complex algorithms to generate personalized treatment advice and recommendations, facilitating SDM [21-25]. For example, an AI-powered remote monitoring system for diabetes can collect and analyze patients' blood glucose levels, exercise levels, and eating habits to generate personalized meal plans and exercise routines for patients [26]. One study developed an AI-assisted Internet of Things wearable smartwatch prototype for older people to proactively detect and manage frailty through the collection of their physical activity data [27].

However, one of the key limitations of using GenAI to provide individualized health recommendations is its inconsistent reasoning capabilities and clinical judgment, particularly when it must synthesize information across long conversations, complex clinical scenarios, or high-volume data sources such as wearables. A systematic review found several issues with large language models (LLMs) such as fragile reasoning performance, diagnostic inaccuracies, and overly cautious clinical judgments across a range of use cases including treatment personalization, monitoring systems, communication support, and diagnostic support and risk prediction [28]. Others have raised concerns about personalized AI's potential to limit exposure to critical information, fostering "echo chambers" that reinforce users' specific beliefs [29].

Patient Engagement and Empowerment

Emerging use cases for GenAI center on its potential to empower patients through self-management and communication support. GenAI can make health information more understandable and equip patients with the information needed to take a more active role in their care. For example, health systems can use GenAI-powered chatbots or digital health assistants as patient-facing self-management resources and/or as communication support tools for clinicians to provide health information, answer questions, help assess symptoms, share

XSL•FO
RenderX

medication reminders, and facilitate appointment scheduling [30,31].

GenAI has the potential to support SDM between patients and clinicians by providing data or educational resources [32]. One study explored perspectives on the use of GenAI in SDM for knee replacement surgery, finding that patients viewed GenAI as another source of information that could enhance their understanding of their risk profile, empowering them to make decisions based on their values and preferences [10]. Despite its potential, patients with less experience using technology or navigating health information from electronic sources (ie, limited digital health literacy) may find using these tools challenging, leading to a sense of disempowerment [16]. With GenAI in particular, emerging evidence suggests that patients may mistake the confident tone of model outputs for factual accuracy, increasing the likelihood that they follow guidance that is incomplete or inaccurate [33-35].

Quality of Care and Safety

GenAI shows promise in improving patient safety and quality of care. It can facilitate real-time physiological monitoring and prediction of patients' conditions and send recommendations to patients or alerts to clinicians to support care coordination [36]. An ongoing study of a GenAI-supported PC CDS monitoring system for patients with asthma uses voice biomarker technology to recognize variations in the patient's recorded voice and sends immediate alerts to the patient for treatment intervention. The digital tool calculates a Respiratory Symptoms Risk Score and allows for remote care coordination, as necessary [37]. GenAI can also enhance diagnostic accuracy and risk prediction based on patient health histories [38], unstructured clinical notes [39,40], PGHD [41], or medical imaging analysis [42]. In some cases, this can inform allocation of resources and treatments based on individual- and system-level data. Mercy Healthcare System integrated the Chen Chemotherapy Model into their EHR to predict hospitalization risk from chemotherapy side effects in adult patients without leukemia. The texting platform prompts patients to report and rate their symptoms, then sends this PRO data to clinicians for review [43]. This allows clinicians to proactively manage symptoms before hospitalization becomes necessary. Furthermore, GenAI can summarize disparate information for clinicians to allow them more time for direct patient interactions, which can improve care quality [44]. With ambient listening technology, GenAI-supported tools convert recorded patient-clinician conversations into structured visit summary notes. These tools reduce the time clinicians spend on documentation during visits, enabling them to focus more attention on direct patient communication and interaction [45].

At the same time, overreliance on GenAI can pose risks to patient safety. One study found that GenAI models exhibited larger cognitive biases for medical decisions when compared with practicing clinicians, illustrating how LLMs could lead to diagnostic or treatment errors in complex medical cases [46]. Another study found that more than half of a sample of primary care physicians reviewing AI-generated patient portal messages containing errors did not identify and remedy all the errors and 35% to 45% of physicians submitted erroneous messages entirely unedited, indicating the need for additional guardrails beyond physician oversight [47]. Additionally, liability questions arise when GenAI is involved in decision-making, especially for "black box" algorithms where the internal decision-making process is hidden, meaning users can only see the input data they provide and the resulting output without understanding the reasoning process for medical predictions or decisions [48]. This makes it more difficult to determine who is accountable if there are adverse outcomes or whether the health system should take responsibility if they do not properly implement, maintain, or train users on the GenAI-supported system [5,49].

Access to Care

GenAI-supported PC CDS has the potential to improve access to medical care. It can support translation service use cases by using speech recognition and generating speech output in any language, thus overcoming language barriers between patients and their clinicians [50]. It can also be used for self-management support use cases in settings where there is a shortage of health care professionals, such as in rural areas or when more immediate access to clinical expertise is needed (regardless of geographic setting) to assist patients with nonurgent medical conditions. For example, Spänig et al [51] demonstrated that an autonomous medical AI interface designed to identify patients at risk for type 2 diabetes mellitus could function as an initial point of guidance in settings with limited clinician availability by offering preliminary risk assessment and directing patients toward appropriate follow-up care.

Despite this potential, GenAI systems may produce uneven health outcomes if not carefully implemented. Models trained on limited or skewed datasets could lead to less accurate self-management recommendations or physiological monitoring predictions for different patient populations [52]. Biased GenAI outputs can result from a range of common data quality issues in training models, such as limited demographic information [53,54], inaccurate or missing data [55], and unrepresentative data that do not include adequate information on specific patient subpopulations [56,57]. While bias can be mitigated by preprocessing data used to train AI models with retrieval-augmented generation (coupling models with an external and reliable database or knowledge base) [58], selecting models that prioritize transparency, and postprocessing the model output to correct for bias, these approaches require developer assessments of what constitutes fairness [59].

Beyond its potential for bias and error, a wide range of factors limit GenAI's integration into patients' lives, which can influence access to care when a GenAI-supported PC CDS tool serves as the first point of guidance in patient-facing use cases (eg, chronic health management and treatment planning). There may be usability challenges if digital and health literacy levels, community-specific factors, and local context are not considered during the design of GenAI-supported tools [60,61]. Additionally, chief medical information officers and other executives are essential since GenAI-supported tools require technological and financial resources to deploy and train care teams on their effective use. This creates the potential to exacerbate the digital divide between patients in health care organizations with robust health information technology support

XSL•FO
RenderX

systems versus patients in lower-resourced health care organizations, leaving some patients without access to GenAI-supported self-management, communication, or monitoring tools that can help them recognize when care is needed and/or remain engaged between visits.

Development of Clinical Decision Support Artifacts and Scaling

In addition to the clinician and patient and caregiver use cases, GenAI has the potential to improve the technology driving PC CDS. It can write code, map variables, create value sets [15], and generate realistic synthetic patient data that reduce the time needed for training and validation cycles [62]. GenAI can also suggest improvements to PC CDS logic [19] by analyzing alert overrides that contribute to clinician fatigue and personalizing alert criteria [63]. Furthermore, GenAI could solve a critical challenge in health care interoperability: the use of multiple incompatible information systems to record health data [64]. By using LLMs to quickly transform heterogeneous data into structured, standardized formats, it can facilitate the sharing of patient information across health systems. Despite this potential, developing interoperable GenAI systems is complex due to the quality of available patient data, the lack of implementation frameworks and standards for GenAI applications, and the cost of computational resources [65].

Areas to Advance GenAI-Supported PC CDS

Overview

The successful integration of GenAI in PC CDS across the use cases requires health system leadership, researchers, developers, informaticians, and policymakers to address at least the 6 areas discussed in subsequent sections to ensure these tools deliver meaningful, safe, and unbiased benefits. As with the benefits and challenges, these areas of opportunity are interconnected in nature and vary in relevance based on the clinical context, the decisions being made, and the patient-clinician interactions involved in the use case. Several of these areas align with the National Academy of Medicine's AI Code of Conduct framework for the development and application of responsible AI in health and medicine [66]. In addition, since PC CDS does not occur in isolation, its success depends on effective public health strategies that promote community health and well-being.

Engaging and Ensuring Representation of Patients and/or Caregivers in Design and Development

To capitalize on GenAI's ability to support the PC CDS use cases illustrated in this manuscript, engaging patients and caregivers during the design and development process is essential [67]. As reflected in the National Academy of Medicine's AI Code of Conduct, by involving end users in usability testing and other design activities—particularly those with varying language preferences, cultural norms, and digital and health literacy—developers can ensure GenAI-supported tools are tailored to account for the unique preferences, contexts, communication styles, and needs of individual patients [68,69]. Additionally, patient input can help ensure personalized treatment recommendations provided by GenAI are actionable for different populations [70] and accommodate the complexity of patients with multimorbidities. The degree to which patients and caregivers influence the design of GenAI-supported PC CDS tools will vary by use case, with greater influence typically possible in patient-facing applications and more limited involvement in clinician-facing functions such as diagnostic support or risk prediction.

Building the Science of Effective PC CDS Implementation to Support Patient Engagement

Advances in methods to promote the adoption of evidence-based practices are needed to address the barriers to patient engagement posed by patient-facing GenAI-supported PC CDS in particular. Research is needed on the multilevel factors that (1) influence the use of communication support tools such as chatbots and digital health assistants, such as inaccurately processing patient inputs or lack of personal connection [70]; (2) support their seamless integration into clinical workflows and patient lifeflows so that the information provided by the PC CDS does not undermine clinician communication; and (3) support GenAI's ability to facilitate SDM through visualizations that contain the appropriate level of detail [71,72]. A randomized clinical trial for a GenAI-enabled clinical decision aid used some of the aforementioned strategies through the incorporation of PRO measures, patient education, and tailoring to patient preferences, resulting in improvements in patients' decisional quality, SDM, and functional outcomes [73]. Although the research on effective PC CDS implementation is nascent, this evidence supports the importance of further exploring multilevel factors influencing patient engagement to improve uptake of GenAI-supported PC CDS.

Developing Risk-Based Policies for Deciding When GenAI Use Is Appropriate and What Level of Clinician Involvement Is Required

GenAI's ability to integrate into such a wide range of PC CDS use cases highlights the importance of developing clear policies for its use. Risk-based guardrails should define when GenAI can operate autonomously in supporting patients' knowledge and data capture (eg, by translating clinician-authored recommendations into plain language, generating visit summaries, or helping patients record symptoms and goals for minor issues) and where clinician review and sign-off are required before outputs are used to guide care (eg, for patients with multimorbidities who may face conflicting guidelines) [74]. Human-in-the-loop represents a promising approach for risk-based guardrails. Although its current use has largely focused on clinician-facing use cases such as verifying the accuracy and reliability of AI-generated discharge summaries or personalized recommendations, it also has potential to support patients with low digital health literacy in patient-facing use cases such as self-management support, communication support, and monitoring systems, where information is delivered digitally and patients may require additional support in interpretation. As seen in SDM research, patients with low digital health literacy benefit from clinician facilitation to interpret and apply complex health information, suggesting similar human-in-the-loop support may be beneficial for patient-facing GenAI-supported PC CDS tools [75]. While organizations bear

XSL•FO
RenderX

primary responsibility for ensuring the safe and appropriate use of the GenAI tools they deploy, patients' own GenAI literacy can also play a role in promoting appropriate use. Critical health AI literacy, the ability to critically evaluate GenAI outputs, recognizing how health systems and structural factors may shape them, and using this awareness to make informed decisions, is an emerging concept that can complement organizational safeguards by helping patients engage more safely and effectively with GenAI PC CDS [76]. Overall, well-defined policies will support postdeployment monitoring and help mitigate risks such as overreliance on GenAI, unclear liability, and inappropriate delegation of clinical judgment to technology.

Establishing Independent Testing and Vetting Criteria to Ensure Safety and Accuracy

To ensure the quality and safety of GenAI-supported PC CDS, developers must rigorously evaluate and independently test processes before deployment [74]. Unlike static decision support systems, GenAI-supported tools require initial validation not only for accuracy but also for fairness, interpretability, and robustness against biases [46]. Suggested approaches include conducting centralized testing through third parties, as well as requiring local oversight by adapting models such as the Clinical Laboratory Improvement Amendments for AI [77]. Establishing standardized oversight for GenAI helps build trust in its recommendations and ensure its deployment improves, not compromises, care quality.

Periodically Reassessing to Ensure No Algorithmic Drift and Verify Performance

GenAI algorithms and the data they operate on evolve, necessitating reassessment by the organization implementing GenAI-supported PC CDS to identify and address algorithmic drift, where shifts in data patterns lead to degraded performance or unintended outputs [74]. Because GenAI-supported PC CDS is integrated into real clinical workflows, monitoring approaches and metrics need to remain feasible and usable by clinical and organizational leaders, not just data scientists and researchers [78]. Several early prototypes, such as recent work developing human-centered AI monitoring systems for health care, demonstrate the utility of evaluating models along 4 core dimensions: performance, process, outcomes, and fairness [78]. For GenAI-supported PC CDS, outcomes and fairness can be assessed using similar principles, that is, evaluating whether recommendations lead to desired clinical and safety outcomes overall and across patient subgroups. Performance and process evaluation are more complex for GenAI because outputs are nondeterministic and language-based, and validated measures and methods are still emerging. Monitoring must therefore address behaviors such as hallucinations [79], reasoning quality [80], sycophantic behavior [81], and input robustness or feature drift [82]. While some elements require human expert review, automated and semiautomated tools are emerging. Examples include LLM-as-judge for evaluating correctness and guideline adherence or assessing fidelity to expected answers or text-based summaries of clinical information [83]; and statistical monitors that track token-level patterns or vocabulary shifts indicating changes in model behavior [82]. Organizations should establish clear triggers for reassessment—such as major guideline updates, new evidence affecting care recommendations, demographic changes in the patient population, the passage of time, or early signals of performance deterioration—and establish a reassessment frequency based on the clinical risk and operational impact of the GenAI-supported application [84,85].

Establishing Policies to Promote Transparency and Patient Consent in the Use of GenAI

Patients need clear explanations of when and how GenAI is used in PC CDS tools that influence their care, along with the technology's limitations [86]. The fair, appropriate, valid, effective, and safe principles [87] offer a blueprint for providing this transparency through specific, standardized technical and performance information. Furthermore, GenAI-supported PC CDS tools may require access to sensitive health information (eg, ambient listening tools and monitoring systems) with which patients may have varying levels of familiarity and comfort. Patients should have the right to opt out of the use of GenAI in their care and have full knowledge of who will access their data and how it will be used if they choose to participate. The development of patient-centered transparency and consent guidelines represents a critical area for advancing GenAI-supported PC CDS, including establishing which information is most relevant for patients to understand, determining how it should be effectively communicated, and considering when dynamic consent models may be warranted given the adaptive nature of GenAI systems [88].

Conclusions

The integration of GenAI into PC CDS holds promise for transforming health care delivery. GenAI can enhance the personalization and precision of care; empower patients through improved engagement and SDM; and support clinicians, patients, and caregivers in making more informed decisions. However, the successful implementation of GenAI in PC CDS requires consideration of several critical factors. These include addressing potential biases in GenAI algorithms, ensuring the transparency and explainability of GenAI-driven recommendations, and fostering trust among patients and clinicians. Additionally, it is essential to engage patients and caregivers in the design and development of GenAI-supported tools, develop clear policies for GenAI use, and establish rigorous testing and validation processes to ensure safety and accuracy. By addressing these challenges, the full potential of GenAI can begin to be harnessed to improve patient outcomes and advance the field of PC CDS.

Acknowledgments

The authors would like to acknowledge Edwin Lomotan, MD, and James Swiger, MBE, for their critical review and feedback on the manuscript. In addition, the authors acknowledge the following members of the Clinical Decision Support Innovation

XSL•FO
RenderX

Collaborative (CDSiC) Steering Committee and CDSiC Innovation Center Planning Committee: Dr Joel Andress; Dr James Cimino; Deborah Collyar; Angela Dobes; Dr Jordan Everson; Dr Sonja Fulmer; Dr Robert Greenes; Dr Tonya Hongsermeier; Dr Kensaku Kawamoto; Dr Gilad Kuperman; Dr Brian Levy; Dr Dave Little; Dr David Lobach; Dr J Marc Overhage; Tiffany Peterson; Dr Gerasimos Petratos; Dr Jonathan Teich; Dr Lipika Samal; Wesley Sargent; Patrick Schoen; Dr Richard Schreiber; Dr Scott Weingarten; Michael Wittie; and Dr Haipeng (Mark) Zhang.

Funding

This work is based on research conducted by NORC at the University of Chicago under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, Maryland (contract number 75Q80120D00018/75Q80121F32003). The views expressed in this paper are those of the authors and do not necessarily represent the official positions of the AHRQ or the US Department of Health and Human Services.

Data Availability

No datasets were generated or analyzed during this study.

Authors' Contributions

PD conceived of the presented idea and supervised the project. PD, CZ, NG, CP, and AA wrote the manuscript with input from all authors. DFS and KM provided critical feedback and helped shape the figure and manuscript.

Conflicts of Interest

None declared.

References

1. Feuerriegel S, Hartmann J, Janiesch C, Zschech P. Generative AI. Bus Inf Syst Eng 2024 Feb;66:111-126. [doi: 10.1007/s12599-023-00834-7]
2. Dullabh P, Sandberg SF, Heaney-Huls K, et al. Challenges and opportunities for advancing patient-centered clinical decision support: findings from a horizon scan. J Am Med Inform Assoc 2022 Jun 14;29(7):1233-1243. [doi: 10.1093/jamia/ocac059] [Medline: 35534996]
3. Ouanes K, Farhah N. Effectiveness of artificial intelligence (AI) in clinical decision support systems and care delivery. J Med Syst 2024 Aug 12;48(1):74. [doi: 10.1007/s10916-024-02098-4] [Medline: 39133332]
4. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
5. Braun M, Hummel P, Beck S, Dabrock P. Primer on an ethics of AI-based decision support systems in the clinic. J Med Ethics 2020 Apr 3;47(12):e3. [doi: 10.1136/medethics-2019-105860] [Medline: 32245804]
6. McCradden MD, Joshi S, Anderson JA, London AJ. A normative framework for artificial intelligence as a sociotechnical system in healthcare. Patterns (N Y) 2023 Nov 10;4(11):100864. [doi: 10.1016/j.patter.2023.100864] [Medline: 38035190]
7. Liaw ST, Liyanage H, Kuziemsky C, et al. Ethical use of electronic health record data and artificial intelligence: recommendations of the Primary Care Informatics Working Group of the International Medical Informatics Association. Yearb Med Inform 2020 Aug;29(1):51-57. [doi: 10.1055/s-0040-1701980] [Medline: 32303098]
8. Anjara SG, Janik A, Dunford-Stenger A, et al. Examining explainable clinical decision support systems with think aloud protocols. PLoS One 2023 Sep 14;18(9):e0291443. [doi: 10.1371/journal.pone.0291443] [Medline: 37708135]
9. Funer F, Schneider D, Heyen NB, et al. Impacts of clinical decision support systems on the relationship, communication, and shared decision-making between health care professionals and patients: multistakeholder interview study. J Med Internet Res 2024 Aug 23;26:e55717. [doi: 10.2196/55717] [Medline: 39178023]
10. Gould DJ, Dowsey MM, Glanville-Hearst M, et al. Patients' views on AI for risk prediction in shared decision-making for knee replacement surgery: qualitative interview study. J Med Internet Res 2023 Sep 18;25:e43632. [doi: 10.2196/43632] [Medline: 37721797]
11. Bjerring JC, Busch J. Artificial intelligence and patient-centered decision-making. Philos Technol 2021 Jun;34:349-371. [doi: 10.1007/s13347-019-00391-6]
12. Sørensen K, Van den Broucke S, Fullam J, et al. Health literacy and public health: a systematic review and integration of definitions and models. BMC Public Health 2012 Jan 25;12:80. [doi: 10.1186/1471-2458-12-80] [Medline: 22276600]
13. Lorig KR, Holman H. Self-management education: history, definition, outcomes, and mechanisms. Ann Behav Med 2003 Aug;26(1):1-7. [doi: 10.1207/S15324796ABM2601_01] [Medline: 12867348]
14. Elwyn G, Durand MA, Song J, et al. A three-talk model for shared decision making: multistage consultation process. BMJ 2017 Nov 6;359:j4891. [doi: 10.1136/bmj.j4891] [Medline: 29109079]
15. Kawamoto K, Ryan S, Heaney-Huls K, et al. Implementation, Adoption, and Scaling Workgroup: Landscape Assessment on the Use of Artificial Intelligence to Scale PC CDS: Agency for Healthcare Research and Quality; 2024.

XSL•FO
RenderX

16. Desai P, Dobes A, Shah A, et al. Trust and Patient-Centeredness Workgroup: Patient and Caregiver Perspectives on Generative Artificial Intelligence in Patient-Centered Clinical Decision Support [Internet]: Agency for Healthcare Research and Quality (US); 2024.
17. Zott C, Sittig DF, Gauthreaux N, et al. PAIGE Chatbot For Patient-Clinician Communication: Usability and Utility Assessment: Agency for Healthcare Research and Quality (US); 2024.
18. Gauthreaux N, Zott C, Boxwala A, Dullabh PM, Sittig DF. Quartz App to Support Medication Adherence: Usability and Feasibility Assessment [Internet]: Agency for Healthcare Research and Quality (US); 2024.
19. Liu S, Wright AP, Patterson BL, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 2023 Jun 20;30(7):1237-1245. [doi: 10.1093/jamia/ocad072] [Medline: 37087108]
20. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. In: Artificial Intelligence in Healthcare: Academic Press; 2020:25-60. [doi: 10.1016/B978-0-12-818438-7.00002-2]
21. Love-Koh J, Peel A, Rejon-Parrilla JC, et al. The future of precision medicine: potential impacts for health technology assessment. Pharmacoeconomics 2018 Dec;36(12):1439-1451. [doi: 10.1007/s40273-018-0686-6] [Medline: 30003435]
22. Horiuchi D, Tatekawa H, Shimono T, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. Neuroradiology 2024 Jan;66(1):73-79. [doi: 10.1007/s00234-023-03252-4] [Medline: 37994939]
23. Fathima M, Moulana M. Revolutionizing breast cancer care: AI-enhanced diagnosis and patient history. Comput Methods Biomech Biomed Engin 2025 Apr;28(5):642-654. [doi: 10.1080/10255842.2023.2300681] [Medline: 38178694]
24. Ramkumar PN, Haeberle HS, Ramanathan D, et al. Remote patient monitoring using mobile health for total knee arthroplasty: validation of a wearable and machine learning-based surveillance platform. J Arthroplasty 2019 Oct;34(10):2253-2259. [doi: 10.1016/j.arth.2019.05.021] [Medline: 31128890]
25. Fozoonmayeh D, Le HV, Wittfoth E, et al. A scalable smartwatch-based medication intake detection system using distributed machine learning. J Med Syst 2020 Feb 28;44(4):76. [doi: 10.1007/s10916-019-1518-8] [Medline: 32112271]
26. Khalifa M, Albadawy M, Iqbal U. Advancing clinical decision support: the role of artificial intelligence across six domains. Comput Methods Programs Biomed Update 2024;5:100142 Retracted inComput Methods Programs Biomed Update. 2025;8:100204. [doi: 10.1016/j.cmpbup.2025.100204]
27. Ciubotaru BI, Sasu GV, Goga N, et al. Prototype results of an Internet of Things system using wearables and artificial intelligence for the detection of frailty in elderly people. Appl Sci 2023;13(15):8702. [doi: 10.3390/app13158702]
28. Souza GD, Melo G, Schneider D. Tailoring treatment in the age of AI: a systematic review of large language models in personalized healthcare. Informatics 2025;12(4):113. [doi: 10.3390/informatics12040113]
29. Kostick-Quenet KM. A caution against customized AI in healthcare. NPJ Digit Med 2025 Jan 7;8(1):13. [doi: 10.1038/s41746-024-01415-y] [Medline: 39774331]
30. Thorat V, Rao P, Joshi N, Talreja P, Shetty AR. Role of artificial intelligence (AI) in patient education and communication in dentistry. Cureus 2024 May 7;16(5):e59799. [doi: 10.7759/cureus.59799] [Medline: 38846249]
31. Altamimi I, Altamimi A, Alhumimidi AS, Altamimi A, Temsah MH. Artificial intelligence (AI) chatbots in medicine: a supplement, not a substitute. Cureus 2023 Jun 25;15(6):e40922. [doi: 10.7759/cureus.40922] [Medline: 37496532]
32. Abbasgholizadeh Rahimi S, Cwintal M, Huang Y, et al. Application of artificial intelligence in shared decision making: scoping review. JMIR Med Inform 2022 Aug 9;10(8):e36199. [doi: 10.2196/36199] [Medline: 35943793]
33. Hart R. Chatbots can trigger a mental health crisis. What to know about "AI psychosis". TIME. 2025 Aug 5. URL: https://time.com/7307589/ai-psychosis-chatgpt-mental-health/ [accessed 2025-12-10]
34. Dober C. Using generative AI for therapy might feel like a lifeline—but there's danger in seeking certainty in a chatbot. The Guardian. 2025 Aug 3. URL: https://www.theguardian.com/commentisfree/2025/aug/03/generative-ai-chatbot-therapy-dangers-risks [accessed 2025-12-10]
35. Draelos RL, Afreen S, Blasko B, et al. Large language models provide unsafe answers to patient-posed medical questions. arXiv. Preprint posted online on Jul 25, 2025. [doi: 10.48550/arXiv.2507.18905]
36. A novel patient-facing mobile platform to collect and implement patient-reported outcomes and voice biomarkers in underserved adult patients with asthma. Agency for Healthcare Research and Quality. URL: https://digital.ahrq.gov/ahrq-funded-projects/novel-patient-facing-mobile-platform-collect-and-implement-patient-reported [accessed 2025-02-11]
37. ASTHMAXcel voice mobile application to improve chronic disease management and patient outcomes.: Agency for Healthcare Research and Quality URL: https://web.archive.org/web/20251110170833/https://digital.ahrq.gov/program-overview/research-stories/asthmaxcel-voice-mobile-application-improve-chronic-disease [accessed 2025-12-03]
38. Feng T. Applications of artificial intelligence to diagnosis of neurodegenerative diseases. Stud Health Technol Inform 2023 Nov 23;308:648-655. [doi: 10.3233/SHTI230896] [Medline: 38007795]
39. Jee J, Fong C, Pichotta K, et al. Automated real-world data integration improves cancer outcome prediction. Nature 2024 Dec;636(8043):728-736. [doi: 10.1038/s41586-024-08167-5] [Medline: 39506116]
40. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. NPJ Digit Med 2022 Dec 26;5(1):194. [doi: 10.1038/s41746-022-00742-2] [Medline: 36572766]

XSL•FO
RenderX

41. Ye J, Woods D, Jordan N, Starren J. The role of artificial intelligence for the application of integrating electronic health records and patient-generated data in clinical decision support. AMIA Jt Summits Transl Sci Proc 2024 May 31;2024:459-467. [Medline: 38827061]
42. Armato SG, Drukker K, Hadjiiski L. AI in medical imaging grand challenges: translation from competition to research benefit and patient care. Br J Radiol 2023 Oct;96(1150):20221152. [doi: 10.1259/bjr.20221152] [Medline: 37698542]
43. DeFreitas M. Can AI help monitor chemotherapy side effects? HealthLeaders. 2024 Jan 8. URL: https://www.healthleadersmedia.com/technology/can-ai-help-monitor-chemotherapy-side-effects [accessed 2025-12-09]
44. Maleki Varnosfaderani S, Forouzanfar M. The role of AI in hospitals and clinics: transforming healthcare in the 21st century. Bioengineering (Basel) 2024 Mar 29;11(4):337. [doi: 10.3390/bioengineering11040337] [Medline: 38671759]
45. Eastwood B. Ambient listening in healthcare: dictation, documentation and AI. HealthTech. 2024 Aug 7. URL: https://healthtechmagazine.net/article/2024/08/ambient-listening-inhealthcare-perfcon [accessed 2025-04-08]
46. Wang J, Redelmeier DA. Cognitive biases and artificial intelligence. NEJM AI 2024 Nov 27;1(12). [doi: 10.1056/AIcs2400639]
47. Biro JM, Handley JL, Malcolm McCurry J, et al. Opportunities and risks of artificial intelligence in patient portal messaging in primary care. NPJ Digit Med 2025 Apr 24;8(1):222. [doi: 10.1038/s41746-025-01586-2] [Medline: 40275104]
48. Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J Med Ethics 2021 Mar 18:medethics-2020-106820. [doi: 10.1136/medethics-2020-106820] [Medline: 33737318]
49. Maliha G, Gerke S, Cohen IG, Parikh RB. Artificial intelligence and liability in medicine: balancing safety and innovation. Milbank Q 2021 Sep;99(3):629-647. [doi: 10.1111/1468-0009.12504] [Medline: 33822422]
50. Kim H, Jin HM, Jung YB, You SC. Patient-friendly discharge summaries in Korea based on ChatGPT: software development and validation. J Korean Med Sci 2024 Apr 29;39(16):e148. [doi: 10.3346/jkms.2024.39.e148] [Medline: 38685890]
51. Spänig S, Emberger-Klein A, Sowa JP, Canbay A, Menrad K, Heider D. The virtual doctor: an interactive clinical-decision-support system based on deep learning for non-invasive prediction of diabetes. Artif Intell Med 2019 Sep;100:101706. [doi: 10.1016/j.artmed.2019.101706] [Medline: 31607340]
52. Bullerman B. Joint commission on accreditation of healthcare organizations. United States of America. Implicit Bias in Healthcare: A Quick Safety Review. 2026 Mar 23. URL: https://coilink.org/20.500.12592/gbpfpg [accessed 2024-05-01]
53. Akay EM, Hilbert A, Carlisle BG, Madai VI, Mutke MA, Frey D. Artificial intelligence for clinical decision support in acute ischemic stroke: a systematic review. Stroke 2023 Jun;54(6):1505-1516. [doi: 10.1161/STROKEAHA.122.041442] [Medline: 37216446]
54. Jeong HK, Park C, Henao R, Kheterpal M. Deep learning in dermatology: a systematic review of current approaches, outcomes, and limitations. JID Innov 2022;3(1):100150. [doi: 10.1016/j.xjidi.2022.100150] [Medline: 36655135]
55. Balla Y, Tirunagari S, Windridge D. Pediatrics in artificial intelligence era: a systematic review on challenges, opportunities, and explainability. Indian Pediatr 2023 Jul 15;60(7):561-569. [Medline: 37424120]
56. Clement J, Maldonado AQ. Augmenting the transplant team with artificial intelligence: toward meaningful AI use in solid organ transplant. Front Immunol 2021;12:694222. [doi: 10.3389/fimmu.2021.694222] [Medline: 34177958]
57. Buchlak QD, Esmaili N, Leveque JC, et al. Machine learning applications to clinical decision support in neurosurgery: an artificial intelligence augmented systematic review. Neurosurg Rev 2020 Oct;43(5):1235-1253. [doi: 10.1007/s10143-019-01163-8] [Medline: 31422572]
58. Shi W, Zhuang Y, Zhu Y, Iwinski H, Wattenbarger M, Wang MD. Retrieval-augmented large language models for adolescent idiopathic scoliosis patients in shared decision-making. Presented at: ACM-BCB 2023: 14th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics; Sep 3-6, 2023. [doi: 10.1145/3584371.3612956]
59. Ferrara E. Fairness and bias in artificial intelligence: a brief survey of sources, impacts, and mitigation strategies. Sci 2024;6(1):3. [doi: 10.3390/sci6010003]
60. Aung YYM, Wong DCS, Ting DSW. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. Br Med Bull 2021 Sep 10;139(1):4-15. [doi: 10.1093/bmb/ldab016] [Medline: 34405854]
61. Nair M, Svedberg P, Larsson I, Nygren JM. A comprehensive overview of barriers and strategies for AI implementation in healthcare: mixed-method design. PLoS One 2024;19(8):e0305949. [doi: 10.1371/journal.pone.0305949] [Medline: 39121051]
62. Reddy S. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. Implement Sci 2024 Mar 15;19(1):27. [doi: 10.1186/s13012-024-01357-9] [Medline: 38491544]
63. Liu S, McCoy AB, Peterson JF, et al. Leveraging explainable artificial intelligence to optimize clinical decision support. J Am Med Inform Assoc 2024 Apr 3;31(4):968-974. [doi: 10.1093/jamia/ocae019] [Medline: 38383050]
64. Iroju O, Soriyan A, Gambo I, Olaleke J. Interoperability in healthcare: benefits, challenges and resolutions. Int J Innov Appl Stud 2013;3(1):262-270 [FREE Full text]
65. Yim D, Khuntia J, Parameswaran V, Meyers A. Preliminary evidence of the use of generative AI in health care clinical services: systematic narrative review. JMIR Med Inform 2024 Mar 20;12:e52073. [doi: 10.2196/52073] [Medline: 38506918]
66. An Artificial Intelligence Code of Conduct for Health and Medicine: Essential Guidance for Aligned Action: National Academy of Medicine; 2025.

XSL•FO
RenderX

67. Dullabh P, Dungan R, Raj M, et al. Trust & Patient-Centeredness Workgroup: Methods for Involving End-Users in PC CDS Co-Design [Internet]: Agency for Healthcare Research and Quality (US); 2023.
68. Desai PJ, Zott C, Gauthreaux N, et al. Trust & Patient-Centeredness Workgroup: An Introductory Handbook for Patient Engagement Throughout the Patient-Centered Clinical Decision Support Lifecycle [Internet]: Agency for Healthcare Research and Quality (US); 2023.
69. Pham Q, Gamble A, Hearn J, Cafazzo JA. The need for ethnoracial equity in artificial intelligence for diabetes management: review and recommendations. J Med Internet Res 2021 Feb 10;23(2):e22320. [doi: 10.2196/22320] [Medline: 33565982]
70. Nadarzynski T, Knights N, Husbands D, et al. Achieving health equity through conversational AI: a roadmap for design and implementation of inclusive chatbots in healthcare. PLOS Digit Health 2024 May 2;3(5):e0000492. [doi: 10.1371/journal.pdig.0000492] [Medline: 38696359]
71. Milne-Ives M, de Cock C, Lim E, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res 2020 Oct 22;22(10):e20346. [doi: 10.2196/20346] [Medline: 33090118]
72. Robinson R, Liday C, Lee S, et al. Artificial intelligence in health care-understanding patient information needs and designing comprehensible transparency: qualitative study. JMIR AI 2023;2:e46487. [doi: 10.2196/46487] [Medline: 38333424]
73. Jayakumar P, Moore MG, Furlough KA, et al. Comparison of an artificial intelligence-enabled patient decision aid vs educational material on decision quality, shared decision-making, patient experience, and functional outcomes in adults with knee osteoarthritis: a randomized clinical trial. JAMA Netw Open 2021 Feb 1;4(2):e2037107. [doi: 10.1001/jamanetworkopen.2020.37107] [Medline: 33599773]
74. Labkoff S, Oladimeji B, Kannry J, et al. Toward a responsible future: recommendations for AI-enabled clinical decision support. J Am Med Inform Assoc 2024 Nov 1;31(11):2730-2739. [doi: 10.1093/jamia/ocae209] [Medline: 39325508]
75. Gartner B, Leysen D, Mcgowan H, et al. Digitally supported shared decision-making for exercise prescription in the secondary prevention of cardiovascular disease. Eur J Prev Cardiol 2025 May 19;32(Supplement_1). [doi: 10.1093/eurjpc/zwaf236.362]
76. Campos H, Salmi L. Critical AI health literacy as liberation technology: a new skill for patient empowerment. NAM Perspect 2025;12. [doi: 10.31478/202512a]
77. Jackson BR, Sendak MP, Solomonides A, Balu S, Sittig DF. Regulation of artificial intelligence in healthcare: Clinical Laboratory Improvement Amendments (CLIA) as a model. J Am Med Inform Assoc 2025 Feb 1;32(2):404-407. [doi: 10.1093/jamia/ocae296] [Medline: 39657218]
78. Salwei ME, Davis SE, Reale C, et al. Human-centered design of an artificial intelligence monitoring system: the Vanderbilt Algorithmovigilance Monitoring and Operations System. JAMIA Open 2025 Oct;8(5):ooaf136. [doi: 10.1093/jamiaopen/ooaf136] [Medline: 41180894]
79. Asgari E, Montaña-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. NPJ Digit Med 2025 May 13;8(1):274. [doi: 10.1038/s41746-025-01670-7] [Medline: 40360677]
80. Qiu P, Wu C, Liu S, et al. Quantifying the reasoning abilities of LLMs on clinical cases. Nat Commun 2025 Nov 6;16(1):9799. [doi: 10.1038/s41467-025-64769-1] [Medline: 41198657]
81. Chen S, Gao M, Sasse K, et al. When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. NPJ Digit Med 2025 Oct 17;8(1):605. [doi: 10.1038/s41746-025-02008-z] [Medline: 41107408]
82. Kellogg KC, Ye B, Hu Y, Savova GK, Wallace B, Bitterman DS. Large language models require a new form of oversight: capability-based monitoring. arXiv. Preprint posted online on Nov 5, 2025. [doi: 10.48550/arXiv.2511.03106]
83. Croxford E, Gao Y, First E, et al. Evaluating clinical AI summaries with large language models as judges. NPJ Digit Med 2025 Nov 5;8(1):640. [doi: 10.1038/s41746-025-02005-2] [Medline: 41193667]
84. Davis SE, Walsh CG, Matheny ME. Open questions and research gaps for monitoring and updating AI-enabled tools in clinical settings. Front Digit Health 2022;4:958284. [doi: 10.3389/fdgth.2022.958284] [Medline: 36120717]
85. Davis SE, Embí PJ, Matheny ME. Sustainable deployment of clinical prediction tools-a 360° approach to model maintenance. J Am Med Inform Assoc 2024 Apr 19;31(5):1195-1198. [doi: 10.1093/jamia/ocae036] [Medline: 38422379]
86. Sittig DF, Singh H. Recommendations to ensure safety of AI in real-world clinical care. JAMA 2025 Feb 11;333(6):457-458. [doi: 10.1001/jama.2024.24598] [Medline: 39602298]
87. Health data, technology, and interoperability: certification program updates, algorithm transparency, and information sharing. Office of the National Coordinator for Health Information Technology (ONC), Department of Health and Human Services (HHS). 2023. URL: https://healthit.gov/wp-content/uploads/2023/12/hti-1-final-rule.pdf [accessed 2025-03-17]
88. Dullabh P, Dhopeshwarkar R, Leaphart D, et al. Trustworthy Artificial Intelligence (TAI) for Patient-Centered Outcomes Research (PCOR): Office of the Assistant Secretary for Planning and Evaluation (ASPE); 2023.

Abbreviations

AI: artificial intelligence
EHR: electronic health record
GenAI: generative artificial intelligence
LLM: large language model

XSL•FO
RenderX

CDS: patient-centered clinical decision support
PGHD: patient-generated health data
PRO: patient-reported outcome
SDM: shared decision-making

Edited by A Mavragani; submitted 01.Aug.2025; peer-reviewed by D Hynes, H Lv, K Kier, N Newton; revised version received 03.Feb.2026; accepted 27.Feb.2026; published 01.Apr.2026.

Please cite as:
Dullabh P, Zott C, Gauthreaux N, Peterson C, Aronoff A, Monkhouse K, Sittig DF
Integrating Generative AI Into Patient-Centered Clinical Decision Support: Viewpoint on Research and Practice Considerations
J Med Internet Res 2026;28:e81628
URL: https://www.jmir.org/2026/1/e81628
doi:10.2196/81628

© Prashila Dullabh, Courtney Zott, Nicole Gauthreaux, Caroline Peterson, Abigail Aronoff, Kistein Monkhouse, Dean F Sittig. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 1.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Beyond Missingness: Systematizing Methods for Comprehensive Data Fitness Assessment in Clinical Research

Hanieh Razzaghi[1], MPH, PhD; Kaleigh Wieand[1], BSc; Kimberley L Dickinson[1], MS; Michael G Kahn[2], MD, PhD; Jason Roy[3], PhD; Clair Blacketer[4], MPH, PMP; Dimitri A Christakis[5], MD, MPH; Christopher B Forrest[1], MD, PhD; Jane Greenberg[6], MS, PhD; Harold P Lehmann[7], MD, PhD; Keith A Marsolo[8], PhD; Jennifer Sciolla[9], CTRS, CCLS, MS; Mark G Weiner[10], MD; Nicole G Weiskopf[11], PhD; L Charles Bailey[1], MD, PhD

[1]Applied Clinical Research Center, Children's Hospital of Philadelphia, 3401 Civic Center Blvd, Philadelphia, PA, United States
[10]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, United States
[11]Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR, United States
[2]Analytics Resource Center, University of Colorado Anschutz Medical Campus, Aurora, CO, United States
[3]Department of Biostatistics and Epidemiology, Rutgers School of Public Health, Piscataway, NJ, United States
[4]Epidemiology Analytics, Janssen Research and Development, Titusville, NJ, United States
[5]Department of Pediatrics, Seattle Children's Hospital, Seattle, WA, United States
[6]Department of Information Science, Drexel University, Philadelphia, PA, United States
[7]Department of Medicine, Johns Hopkins School of Medicine, Baltimore, MD, United States
[8]Department of Population Health Sciences, Duke University School of Medicine, Durham, NC, United States
[9]Biomedical Research Informatics Center, Nemours/Alfred I. DuPont Hospital for Children, Wilmington, DE, United States

Corresponding Author:
Hanieh Razzaghi, MPH, PhD
Applied Clinical Research Center, Children's Hospital of Philadelphia, 3401 Civic Center Blvd, Philadelphia, PA, United States

Abstract

Secondary use of clinical data offers unprecedented opportunities to rapidly conduct large-scale research and improve patient care. However, incomplete understanding of data quality requirements for a study often causes significant delays in executing analyses and validating results. Current practice has largely followed 2 paths. First, multi-institutional networks have developed general data quality programs, but these are typically tied to unique network characteristics and do not address study-specific requirements well. Second, models have been proposed to formalize the requirements for data fitness analyses without extending to the methods needed to meet these requirements. More recently, tools have been developed to conduct cohort-centric screening, focusing on generally applicable structural checks such as missingness or facial implausibility. These provide a first level of information but incompletely capture the fitness requirements of an analysis. In turn, investigators conduct per-study exploratory analyses, but these efforts are typically ad hoc and partially reported, which can hinder reproducible science and delay advances in patient care. Analogously to advances over the past decade in data modeling and reproducible analytics, there is a need for a more systematic, capable approach to study-specific data quality assessment (SSDQA). We discuss such a model, which guides improved SSDQA design and implementation, including metadata for consistent annotation and reporting of data quality assessment results. The model integrates theoretical principles of data quality testing with pragmatic considerations of application to clinical data, providing a consistent approach to specifying data quality assessment checks. Additionally, it proposes to regularize check application through a standard set of options. The SSDQA model builds on current practice, providing a path toward more complete, sound, and reproducible assessments. These characteristics foster multidisciplinary collaboration to identify data quality issues that, in turn, inform decisions about study design and provide important context that has a bearing on adoption of results.

(J Med Internet Res 2026;28:e76398) doi:10.2196/76398

KEYWORDS

data quality; electronic health records; research readiness; automation; data fitness

Introduction

Network Data Quality Assessment

Secondary use of clinical data offers unprecedented opportunities to conduct large-scale research and derive insights into patient care. This is particularly true when using data from multi-institutional networks such as PEDSnet [1], PCORnet [2], the National COVID Cohort Collaborative [3], All of Us [4], or others [5-9]. However, clinical data are complex, and problems related to data quality [10-16] have significantly

XSL•FO
RenderX

hindered research across many networks. General data quality assessment (DQA) programs have been developed for identifying and resolving issues in real-world data [17-32], including in multi-institutional networks, but efforts have often been siloed by unique network characteristics, differences in data models, or lack of automation and reproducibility [33-35]. This tight coupling of DQA programs with existing infrastructure limits reuse in the wider community.

Data quality frameworks have traditionally focused on data model conformance, completeness, and plausibility components [27] rather than targeting specific clinical content, which can miss significant threats to analytic validity. This is not simply a lack of effective implementation. Rather, it reflects that general DQA at the network level addresses separate analytic problems from study-specific DQA (SSDQA), although the 2 domains are related. For example, a network data quality metric may evaluate clinical procedures as a single group to ensure appropriate mappings and compute procedure records per patient. However, if a particular study relies on the presence of specific biopsy procedures, this can only provide a starting point. A gap in data extraction that maps these procedures to a valid but less specific code or misses these mappings entirely—whether overall or for the study group specifically—would not be discovered during broader network data quality checks. Further testing would be needed to assess whether the needed information could be recovered by changing extraction of procedure data or the presence of the biopsy must be inferred from other data. As a result, most network-focused data quality programs lack the flexibility needed in study-specific contexts. Furthermore, most current DQA research focuses primarily on harmonization of terms to *describe* data quality rather than developing standardized models to *produce* standardized assessments.

Addressing Study-Specific Data Quality

While less widely considered, there have been several attempts to address the need for domain-specific or SSDQA. Efforts focused on regulatory science targeting the US Food and Drug Administration and European Medicines Agency have produced planning guides for evaluating fitness in real-world data, such as SPIFD (Structured Process to Identify Fit-for-Purpose Data) and STaRT-RWE (Structured Template for Planning and Reporting on the Implementation of Real-World Evidence Studies), which have considerably advanced the field [36-38]. Similar work derived from model-based software design has addressed analogous steps [39]. These frameworks formalize study design and data requirements, underlining what is needed to produce reliable results for regulatory decision-making. However, providing specific methods or software to standardize and systematically report on findings was beyond the scope of these efforts. Furthermore, these and similar frameworks recognize the importance of well-delineated reporting on data quality but focus on reporting as part of a bespoke study design rather than building reusable informatics frameworks or tools.

Separately, several tools have taken pragmatic approaches to interrogating study datasets. Some address specific topics, such as missingness or “never events” in DQe-c [40] or differences in case mix in the ENACT Data Quality Explorer [35]. Another effort, dataquieR [41,42], focuses on highly flexible testing of a priori constraints; the resulting tools have a wide potential range but concretely address a smaller set of data quality needs and provide limited options to probe semantic issues that may arise in the data. A third direction, exemplified by the Observational Health Data Sciences and Informatics CohortDiagnostics package [43], has been to transfer tests that are effective at the network level for application to study cohorts to produce more specific reporting.

While each of these efforts has been a step forward, evaluations of the current state of SSDQA for clinical data have consistently identified the need for more standard, applicable, automatable approaches [33,34]. A recent systematic review of DQAs and tools concluded that most SSDQAs are developed on a project-by-project basis and that this approach is not practical because of time and resource constraints [33]. Finally, prior efforts have focused on solving a specific problem rather than developing an underlying model to serve a wide range of data quality requirements in a systematic and reproducible way.

Systematizing SSDQA

The challenge, then, is that SSDQA *content* must be specific to the semantic needs of each analysis, whereas SSDQA *methods* need to be standardized and reproducible. The field has advanced significantly in the past decade, but current practice does not fully attain either of these goals. There remains a need to bridge gaps between broad network DQA, high-level proposals for SSDQA needs, and current ad hoc practice of SSDQA. In this paper, we discuss how this challenge can be met, building from existing science including our own previous work [44] and expert consensus to articulate a model for a systematic approach to SSDQA and reporting. Specifically, we focus on the critical space between formalizing the high levels of SSDQA [36-39] and the specifics of implementing tests [35,40,41], aiming to articulate methods that satisfy the former’s needs and systematize approaches to the latter. We address (1) well-founded, analysis-aware data quality testing and (2) the construction of practical, reusable data quality metrics that can be widely adopted to drive consistent and concrete data quality check development. Advancing in these areas is important to clinical informaticians as methodologists and stewards of data resources; to clinical researchers, who are reliant on large datasets to produce valid results; and to patients and clinicians, who must be informed consumers of research results.

An Expanded Model for Assessing Data Fitness

Articulating Requirements for Clinical Data Fitness

We first articulate a set of requirements for advancing the practice of SSDQA. We include both high-level strategies, or goals that an effective model must reflect, as well as specific design principles that motivate the process of check construction and application, which are shown in Textbox 1.

XSL•FO
RenderX

Textbox 1. Design principles for developing study-specific data quality assessment.

- Data quality assessment should correspond to the different stages of a research project, ranging from cohort selection to assessment of minor covariates.
- The principles of a framework should be standardized and reusable across multiple study-specific contexts, and software packages should streamline implementation.
- Data quality tests need descriptive terms that are specific and pair methods with output.
- Data quality software packages should be configurable without the need for extensive coding.
- Methods used in data quality checks should be interpretable, with a range of standard options to evaluate results.
- Data quality checks should be pragmatic and informative, without requiring deep knowledge of underlying theory for use.
- Temporality should be an important dimension for data quality checks and not be treated as a separate check itself.

We then synthesize these considerations into a concrete model (Figure 1) for constructing fitness-oriented data quality checks. The model provides consistent guidance for standardizing check types that evaluate data fitness across a range of study-specific requirements. It delineates the processes of check development and execution through two interdependent processes labeled on the left-hand side of Figure 1: (1) SSDQA development and tools, which focuses on check development and back-end automation; and (2) user process, which highlights the interaction between the user and data quality checks. We elaborate on these processes in the 2 sections below.

Figure 1. Model for systematic construction of study-specific data quality assessment (SSDQA) checks showing the conceptual tracks of check development and user process and the stages of initial specification followed by application.

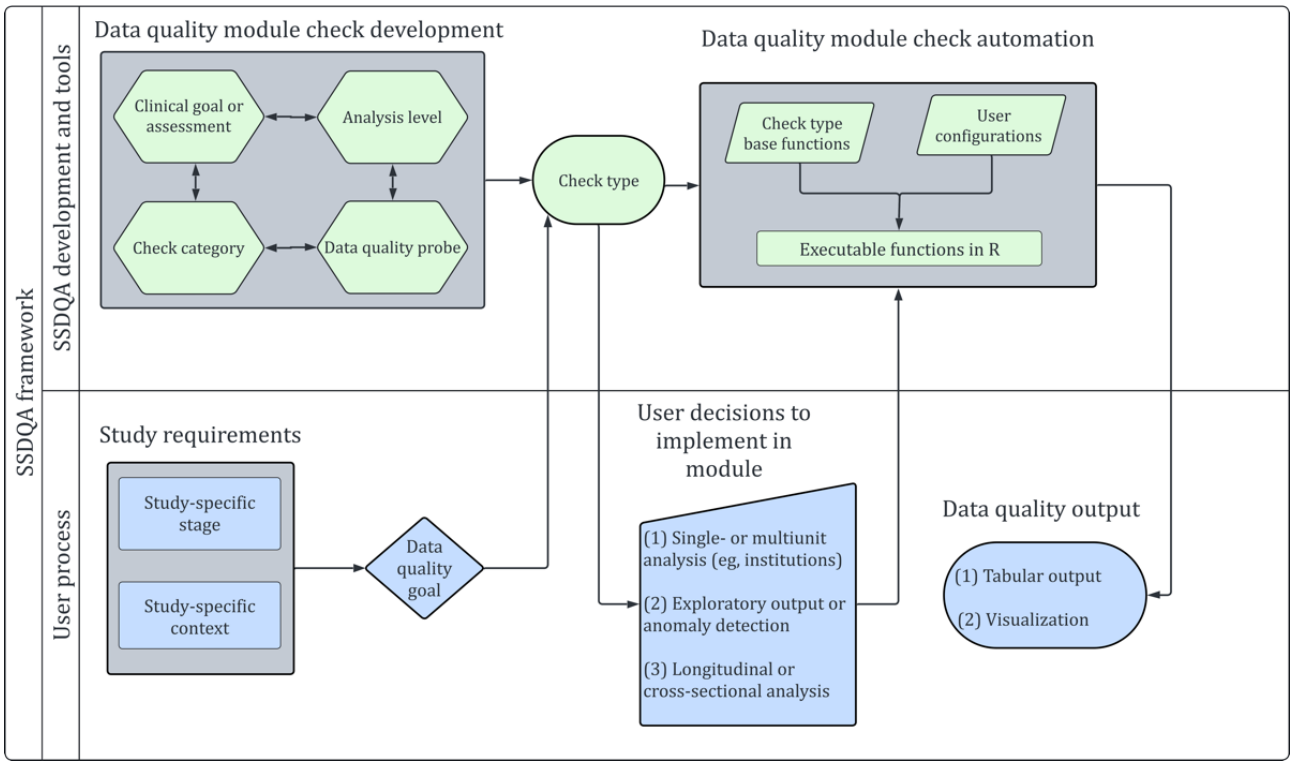

Adding Rigor to SSDQA Development

To promote better coverage of SSDQA needs as well as improved understanding of what the measures produce, the development of a data quality check type should incorporate 4 separate but integrated components. The first is the clinical goal or assessment, which describes the clinical truth that the user needs to evaluate; this aligns in use with purposive frameworks such as SPIFD [37]. For example, an investigator might posit that patients in a study cohort should have confirmatory clinical data (eg, end-stage renal disease and evidence of dialysis) to establish basic face validity. While check types are broader than specific study goals, what goals a check type could address is an important consideration. The second component is check category, which expresses the evaluated aspect of data quality; here we encompass current work focused on widely adopted harmonized terms [45] such as completeness, plausibility, or conformance. The third component is a data quality probe, which, in contrast to the clinical goal, designates a data-centric

XSL•FO
RenderX

purpose for a data quality check, such as assessing eligibility criteria or mapping errors in the data. Finally, analysis level describes the application of a check across levels of aggregation, including person-level, event-level, or visit-level analysis. The integration of these 4 components forms a check type, which is discoverable through these attributes; a set of metadata terms reflecting current practice is provided in Textbox 2. Use of these and similar terms allows check types to be findable for reuse in different studies and promotes interoperability and reusability in the comparison of results across studies.

Textbox 2. Metadata terms associated with attributes of study-specific data quality assessment check types.

Check category

- Concordance
- Conformance
- Completeness
- Consistency
- Information representation
- Plausibility

Data quality probe

- Misclassification
- Eligibility criteria
- Missing expected data
- Anomalous information density
- Temporal inconsistency
- Anomalous value from internal distribution
- Data representation errors
- External benchmarking
- Selection error or bias

Clinical goal or assessment

- Clinical consistency
- Clinical follow-up
- Clinical complexity
- Confirmatory clinical data
- Expected clinical event representation
- Targeted patient population
- Utilization patterns
- Valid diagnostic criteria

Analysis level

- Person level
- Visit level
- Event level

Once a check type has been specified, application to the data could be mediated by a common code base instantiating the desired analyses, spanning the set of user configurations (see below) required to execute a specific data quality check of that type. This approach, which we term “data quality modules,” facilitates consistent application of the check type through a set of base functions incorporating user choices via parameters without requiring reimplementation or code editing and the consequent risk of divergent computation across use cases [46].

Standardizing Application of DQA Checks

Application of study-specific data quality testing must by definition be adapted to each study’s requirements. Data quality checks are deployed to inform decision-making about study design and method choices, assess the validity of the dataset, and identify potential biases or stratifications present in the data. Data quality requirements may vary based on the stage of the study or the background of the user. For example, data quality

XSL•FO
RenderX

for a clinical researcher planning cohort selection will require a different set of evaluation measures from those for the validation of primary outcomes, and testing at the design stage will differ from a statistician investigating an anomalous result during analysis. This is denoted as "study-specific stage" in Figure 1.

In parallel, "study-specific context" compels users to formally define the data constructs (eg, cohort, variables, and concept sets) around which to identify potential gaps in data quality or constraints in study design that require testing. This step drives the creation of the "study-specific" requirements at the core of SSDQA application. On the basis of these criteria, users go on to select the appropriate check types to be applied to the desired data elements.

Three final decision points determine how a particular check is applied and interpreted. First, applications may focus on single- or multi-unit analysis. Most often, units are participating institutions, and multiunit output facilitates comparisons among institutions, whereas single-unit output permits more in-depth analysis within a particular institution. In other circumstances, it may be more correct to combine all institutions as a single unit of analysis or stratify by other characteristics, such as clinical specialty. Second, a particular data quality analysis may focus on manual exploration or be used to formally detect outliers. The former produces descriptive visualizations, which are particularly useful for assessing patterns relying on topical expertise. In contrast, the latter computes quantitative metrics to identify potential outliers, which is useful when assessing complex patterns or for automation. Third, users may wish to analyze their data longitudinally to understand changes over time or cross-sectionally covering a specified period to examine average effect.

Augmenting the Scope of SSDQA

This model for SSDQA allows us to consider current practice centered on analytic requirements rather than software output and identify areas for expansion. We begin with a set of 18 data quality check types (Table 1) that cover foundational requirements for SSDQA; while not exhaustive, these check types span a large fraction of use cases encountered in clinical studies. Although many check types can be applied at multiple phases of a study, we group them here based on likely associations. For example, during the "cohort identification" phase of a study, research teams might deploy the "attrition step" or "sensitivity to selection criteria" check types. The primary difference between "cohort fitness" and "dataset fitness" is that the former evaluates the fitness of a specific cohort against study requirements, whereas the latter tests for the logical cohesion of a dataset. Therefore, anomalies found in the former require investigating study-specific criteria, whereas those found in the latter identify broader logical inconsistencies in a dataset. "Data conformance" check types evaluate a dataset for conformance to a required set of standards, such as ensuring that units for a particular laboratory test conform to the standard or accepted reporting for that test. "Variable testing" focuses on study variables, that is, a set of operational clinical definitions represented in an analysis. Nearly all check types can be reasonably configured for single-site vs multi-site analysis, exploratory output vs anomaly detection, and longitudinal vs cross-sectional analysis. As a result, a single check type may yield 8 different types of outputs.

XSL•FO
RenderX

Table . Catalog of check types derived from the study-specific data quality assessment model.

Check type	Check description	Metadata terms
Cohort identification		
Attrition step	Assess heterogeneity in cohort selection through attrition steps	"Plausibility," "eligibility criteria," "missing expected data," "selection error or bias," "target patient population," and "person level"
Sensitivity to selection criteria	Evaluates the impact of alternate cohort definitions by comparing demographics, follow-up time, utilization patterns, outcomes, and other key variables	"Consistency," "eligibility criteria," "selection error or bias detection," "clinical consistency," "target patient population," and "person level"
Cohort fitness		
Patient facts	Assesses the availability of clinical events such as drug utilization or laboratory events per year of follow-up	"Completeness," "missing expected data," "anomalous information density," "clinical follow-up," "utilization patterns," and "visit level"
Patient event sequencing	Evaluates the plausibility of dates in relation to clinical events (eg, chronic kidney disease diagnosis precedes order for dialysis)	"Consistency," "plausibility," "misclassification," "expected clinical event representation," "confirmatory clinical data," "utilization patterns," and "person level"
Patient record consistency	Tests whether the clinical data in a patient's record are consistent and confirmatory (eg, patients with leukemia should be receiving chemotherapy)	"Consistency," "plausibility," "missing expected data," "clinical consistency," "confirmatory clinical data," "utilization patterns," "valid diagnostic criteria," and "person level"
Expected facts present	Evaluates whether a cohort has specific types of clinical data, such as BMI, office visit procedures, or vital signs	"Completeness," "missing expected data," "valid diagnostic criteria," "expected clinical event representation," and "visit level"
Clinical data values and ranges	Assesses whether clinical data outcomes or values align with patient cohort characteristics	"Consistency," "misclassification," "confirmatory clinical data," "valid diagnostic criteria," and "person level"
Dataset fitness		
Clinical events and specialty agreement	Evaluates the concordance of specific events (eg, diagnosis of type 1 diabetes) and specialist clinicians or care sites (eg, endocrinologist or endocrinology clinic) at the visit level	"Concordance," "data representation errors," "confirmatory clinical data," "utilization patterns," and "visit level"
Visit clinical data agreement	Determines whether expected clinical events occur within the same clinical encounter (eg, ventilator support in the ICU[a] or initial antihypertensive prescription and blood pressure reading)	"Concordance," "missing expected data," "clinical consistency," "expected clinical event representation," and "visit level"
Date sequencing	Detects outliers and anomalous values such as dates too far in the past or future and whether dates assigned to clinical events occur in a reasonable order or proximity	"Plausibility," "temporality inconsistency," "clinical consistency," "utilization patterns," and "event level"
Data conformance		
Clinical metadata	Assesses whether clinical facts are accompanied by appropriate metadata (eg, prescription drugs with dosing information and laboratory values with assigned specimens)	"Conformance," "missing expected data," "data representation errors," "expected clinical event representation," and "event level"
Unit and value alignment	Determines whether drug prescriptions and administrations or laboratory results contain units that are conformant to expected standards	"Conformance," "misclassification," "clinical consistency," and "event level"
Duplicate record check	Identifies where there are duplicate rows or values in a given dataset	"Conformance," "anomalous information density," "utilization patterns," and "event level"
Variable testing		
Expected variables present	Checks for the presence of expected variables and presents a variety of distributions of these variables in the dataset	"Completeness," "missing expected data," "external benchmarking," "expected clinical event representation," "valid diagnostic criteria," and "person level"

XSL•FO
RenderX

Check type	Check description	Metadata terms
Quantitative variable distributions	Evaluates quantitative distributions of clinical values (eg, laboratory values or BMI) or patient characteristics (eg, number of visits per patient or follow-up time)	"Plausibility," "anomalous value from internal distribution," "valid diagnostic criteria," "clinical consistency," and "event level"
Concept set testing		
Concept set distribution	Examines distributions of concepts (eg, codes) that represent a particular variable (eg, wheezing or uncomplicated asthma are the most common concepts represented in primary care cohort, but severe asthma may be the most common in patients with prolonged respiratory specialty care)	"Information representation," "data representation errors," "expected clinical event representation," and "event level"
Source and concept vocabularies	Provides source-to-concept mappings and their distributions to identify potential problems related to information loss	"Information representation," "data representation errors," "clinical consistency," and "event level"
Unmapped concepts	Shows the potential impact of unmapped concepts within a dataset	"Information representation," "misclassification," "missing expected data," "expected clinical event representation," and "event level"

[a]ICU: intensive care unit.

Using this taxonomy, we see that current DQA tools operate largely within a few types. Tests for missingness or data density, a focus of many packages [26,27,40,41,43], fall within the "expected facts present" type or, in some cases, assess "clinical metadata" [27]. Tests for never-valid implausibility [26,27,41,43] most often assess date sequencing or, in some cases, patient record consistency. Conformance-oriented packages [17,27,43] may add to these tests matching the "unmapped concepts" and less often the "unit and value alignment" or "duplicate record" check types. The ENACT Data Quality Explorer focuses on diagnosis code distribution [35], a form of the "expected facts present" type, as an opportunity to infer other data quality gaps.

Viewed through this lens, it becomes clear that there is good coverage of structural and study-autonomous aspects of data quality that can disrupt analyses, but there are few resources for systematic interrogation of the clinical semantics that are equally critical to study validity. It is in these areas that we believe that the calls for increased standardization in data quality testing are most apt [33,34].

Applying Well-Formed DQA to Clinical Research

Identifying Requirements During Study Design

To illustrate how the approach that we discuss can improve clinical studies, we examine a specific need in clinical research design: assessing whether persons and data available for a study correctly identify the at-risk population, that is, form a clinically reasonable basis to ask the research question. There are several implicit aspects of this question, including the following: (1) Does the cohort accurately reflect the medical state of interest? (2) Are there undiscovered selection biases or stratifications in the cohort? (3) Are the data elements needed to find patients present? For this discussion, we will focus on the specific question of whether the cohort behaves as one would expect for eligible patients; that is, does it have facial and possibly construct validity? The study development path we propose encapsulates these ideas in a single check type that evaluates whether the *variables* (ie, clinical concepts) that a study's design expects to find are present in a study dataset (check type: expected variables present), applied during the cohort identification and fitness phases of the study. We note that this is different from the more commonly addressed question of whether *data elements* (ie, labels that ideally capture specific facts for the study) are present; this would follow the "expected facts present" type. The model distinguishes between the 2 because they are semantically different, notwithstanding the fact that they are often confused in practice. Several other check types will be germane to the study's process of fitness testing. In turn, this should not only inform study design but also produce knowledge that allows others to compare this to other studies and design newer studies more efficiently. However, for the sake of this example, we will limit discussion to 1 check type. We present this as an illustration of how the framework might be used rather than as a systematic evaluation of its effect.

Having identified a relevant check type, we describe the user process phase in a hypothetical study of children with sickle cell disease (SCD). This is an uncommon condition in the general population and, hence, unlikely to drive findings in network-wide data quality testing. However, it is a major cause of both health burden and health care utilization in affected children. For this example, we set 4 expectations: evidence of transcranial Doppler procedure, a standard practice to assess risk of stroke; selected SCD-related laboratory results; pain diagnoses; and hydroxyurea prescription.

Single-Unit Analysis

The single-unit "expected variables present" visualization (Figure 2) allows users to explore data quality as a single unit rather than a comparison of entities to evaluate the dataset in aggregate. Figures 2A and 2B highlight the difference between exploratory visualizations and formal anomaly detection. In this example, distributions of variables indicate that more patients in this cohort have an associated hemoglobin test result (approximately 90%) than other SCD-related tests (approximately 73%). However, there is also a high degree of overlap between patients with hemoglobin and other tests

XSL•FO
RenderX

(Figure 2B), and the clinical expectation is that all patients followed for SCD will have results from both sets of tests. This would imply that data quality problems are more common for capturing or representing SCD-specific rather than general test results, as the former are required for a smaller set of patients. Analogously, the use of transcranial Doppler ultrasound, indicated for patients with the more physiologically severe SCD-SS and SCD-Sβ^0 variants of SCD, correlates fairly well with patients who are taking hydroxyurea, indicated for the same subset of patients, suggesting that patient data are more likely to be complete in this subset of patients requiring more follow-up and clinical observation. The low proportion of patients prescribed hydroxyurea overall may in itself point to a data quality problem, as the drug is indicated for all patients in this subcohort. Possible explanations are that the cohort criteria are too broad (SCD-SS and SCD-Sβ^0 represent 65%-75% of all patients with SCD) and the drug has only been formally recommended in pediatrics since 2017.

Figure 2. Single-unit results from an "expected values present" study-specific data quality assessment check type: (A) exploratory visualization showing the proportion of the cohort having each of several clinical characteristics that would be expected for the study design, (B) exploratory visualization showing co-occurrence of variable pairs (the Jaccard index is able to account for incomplete capture of each variable, and self-comparisons are omitted), (C) longitudinal trends in completeness of several expected data elements, and (D) anomaly detection using control charts applied to longitudinal trend data.

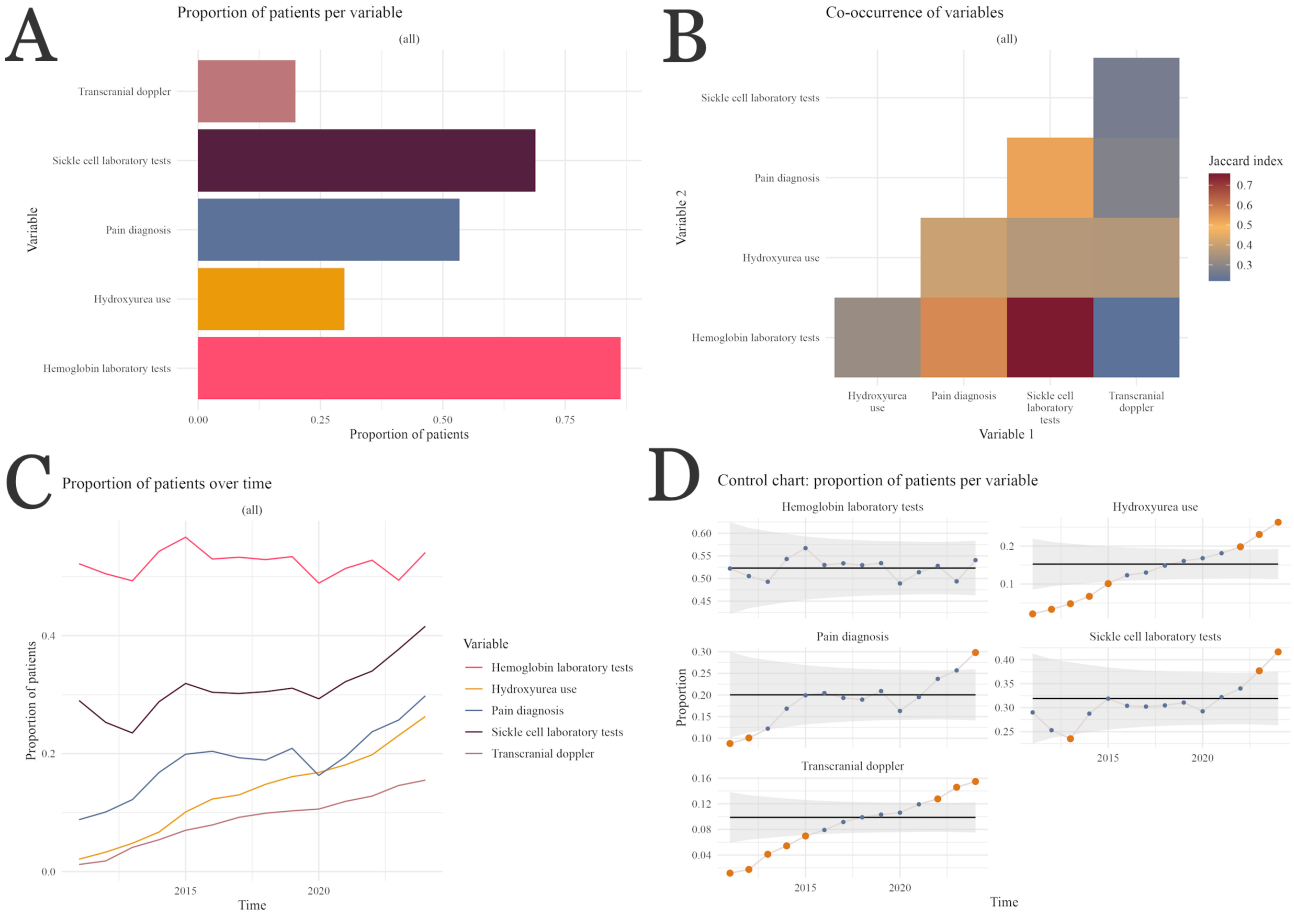

Figures 2C and 2D show temporal trends for the selected variables, which may elucidate whether the low observed proportions for hydroxyurea use are driven exclusively by approval date. Figure 2C is a time-series analysis where all variables are plotted as a proportion of eligible patients over time. It presents relatively stable rates for hemoglobin testing and pain diagnoses, steady increases over time in both transcranial Doppler and hydroxyurea use that reflect updated clinical patient guidelines, and relatively sharp increases in 2013 for laboratory testing variables that may suggest timing of electronic health record implementation rather than differences in care or utilization patterns. Figure 2D shows control charts, with anomalies indicated by orange circles. These graphs reflect similar trends to those exhibited in the exploratory graphs; the hydroxyurea output in particular indicates that the 2017 approval influenced prescribing patterns, although multi-institutional data may clarify whether trends and proportions are heterogeneous. Overall health care utilization still remains lower than expected at approximately 70%.

Multi-Unit Analysis

Figure 3 again shows both the exploratory and anomaly detection visualizations for the same variables in a multiunit (site contributor) analysis. The exploratory visualization shows a heat map with proportions of patients who have evidence of the indicated variable in the dataset (Figure 3A). The display showing tests for anomalous values shows a dot plot with the same color scale but with a star indicating statistical anomalies and the size of the dot indicating the mean proportion across all institutions (Figure 3B). The size and color of the hemoglobin results indicate that all sites capture this variable well, ranging from 78% (site A) to 95% (site J). Site A shows

XSL•FO
RenderX

lower-than-expected capture of non–diagnosis-related clinical data, whereas site D captures laboratory tests well for patients but, conversely, captures pain diagnoses, hydroxyurea use, and transcranial Doppler poorly compared with other sites. Notably, site K captures all use well except for transcranial Doppler, likely marking a problem with extracting procedure data from clinical source systems. These graphs illustrate the most frequent anomalies for transcranial Doppler, SCD-specific laboratory tests, and hydroxyurea utilization variables. Hydroxyurea in particular illustrates how heterogeneous data capture or clinical practice might appear, with sites B and D showing low representation (approximately 11%) and sites H and K showing high representation (42% and 52%, respectively).

Figure 3. Multiunit results from applying an "expected values present" study-specific data quality assessment check type: (A) exploratory visualization showing the proportion of the cohort having each of several clinical characteristics that would be expected for the study design, now stratified by institution; (B) analogous visualization identifying sites with statistically anomalous proportions for each variable; (C) smoothed (locally estimated scatterplot smoothing [LOESS] with 95% CI) longitudinal trends in the proportion of patients receiving hydroxyurea; and (D) use of the Euclidean distance to summarize the differences in longitudinal trends shown in the prior graph to facilitate identification of outliers.

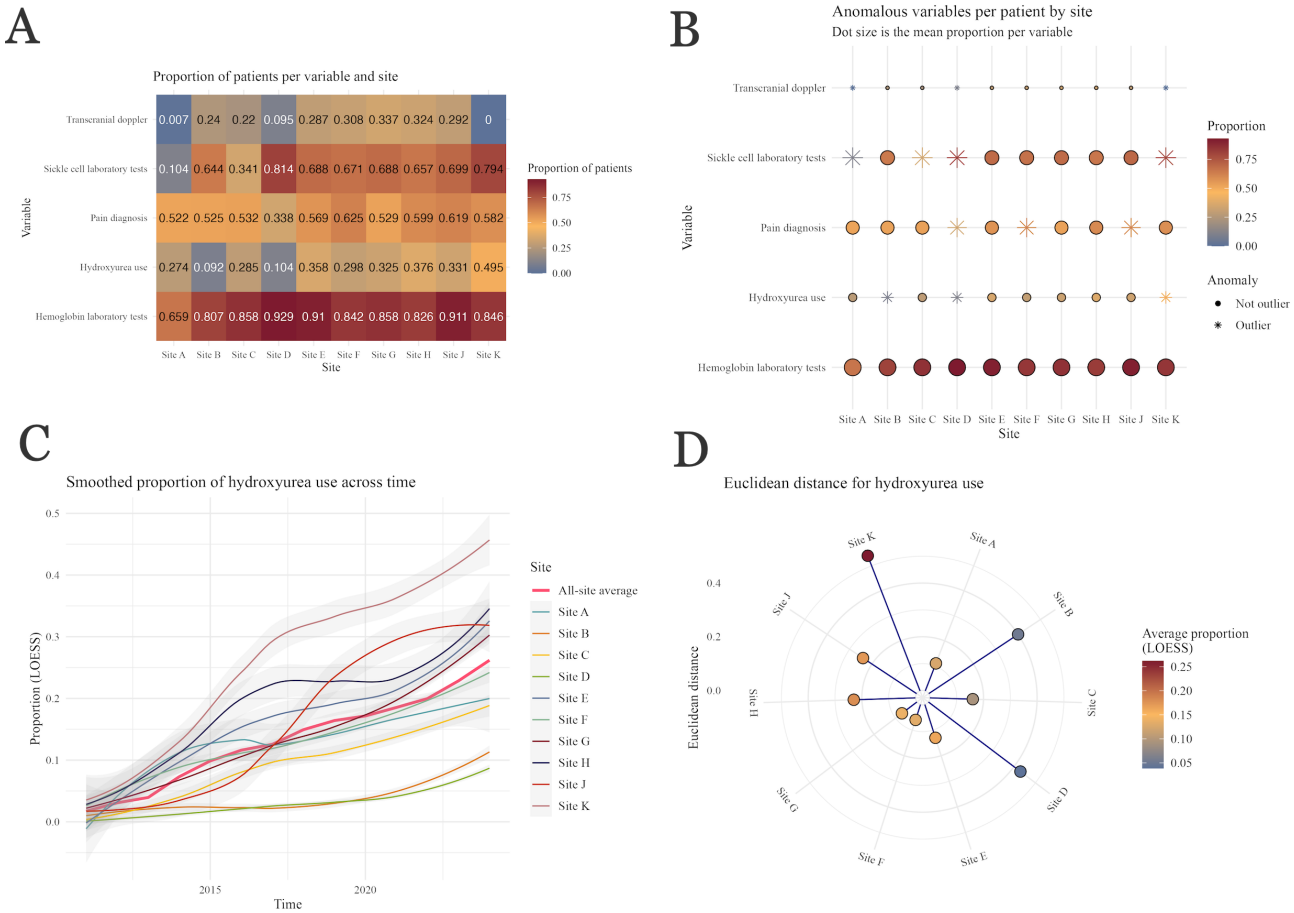

Finally, we illustrate trends across time, facilitating deeper exploration of the heterogeneity shown in the overall results. Taking hydroxyurea use as an example, Figure 3C reinforces the single-unit observation that the Food and Drug Administration drug approval in 2017 increased use across time. Site J appears the most impacted, with a sharp uptick in 2017 and stable use afterward, in contrast to other sites' steady rise. The longitudinal data also reinforce the observation that sites B and D have a low proportion of patients prescribed hydroxyurea, and while slightly increasing in more recent years, the proportions remain nearly constant. Site B is shown with high-quality data in other domains, as would be observed with drug-specific data quality problems. In contrast, site D shows anomalous data across domains. Figure 3D displays the overall distance between each site and the all-site median across time. This visualization reveals that sites K, B, and D trend markedly differently from the all-site median, with site K trending higher and, as expected, sites B and D trending lower.

Taken together, this single check type (expected variables present) can produce diagnostic outputs that inform decision-making about several potential data quality problems related to study-specific requirements. Some of these findings would doubtless be identified by a study team performing ad hoc exploratory analyses but likely not all, and certainly with limited potential for cross-study reuse of either tests or results. Importantly, we describe check types designed to augment rather than replace the type of information available from network-wide DQA, allowing for interrogation of semantics critical to proposed analyses and not just data structure.

Summary

Structuring the ways in which we approach SSDQA contributes to better and more systematic research in several ways. First, checks can be driven by a theoretically sound framework that integrates clinical and technical considerations while retaining

XSL•FO
RenderX

the opportunity to configure how it is applied to meet the requirements of the user. Second, the adoption of a common framework enables reproducibility and standardization of SSDQA checks for dissemination in a wider community in keeping with principles of findability, accessibility, interoperability, and reusability [47,48]. Third, the adoption of a shared and well-grounded framework further facilitates transparency about data beyond simple descriptions when sharing of datasets is not possible due to patient privacy considerations. Consumers of research results are better able to understand data characteristics that may have a bearing on the applicability of the results to their circumstances. Similarly, other researchers can account for differences in data source to validate or augment previous work. For example, complex differences in use of the health system across cohorts may bias study outcomes, which is only crudely reported in current publications.

The problem is complex, and efforts will necessarily be incremental for many reasons. First, the universe of potential SSDQA testing is quite large, and while the theoretical model we discuss is expansible, a finite set of check types cannot cover all possible requirements. However, SSDQA for any single study will exercise only a subset of those possibilities; the check types we highlight cover most high-frequency cases. Second, effective SSDQA requires an interdisciplinary team comprising technical, methodological, and clinical domain expertise; systematizing DQA methods does not replace the need for this expertise. Nonetheless, for such teams, a repeatable approach is particularly useful both for study-wide screening for risks and when a DQA issue is suspected. For the former, check types for cohort identification and variable exploration may inform study design, variable definitions, and analytic requirements. For the latter, data anomalies can be investigated during a study in a more predictable way that will not require designing novel checks when time is of the essence. Third, a greater diversity of check types brings with it complexity, requiring familiarization to make the most effective use of available resources. While adoption of well-founded methods is advantageous in the longer term, a culture shift will be needed in areas where practice is currently ad hoc. Finally, our discussion currently focuses on structured data. Given the rapid increase in the use of free-text or imaging data, translation of these ideas will be needed. Nevertheless, the check type components apply across data domains even if the subsequent development of tools differs.

Effective SSDQA is a critical but complex part of conducting valid research using clinical data. As with other steps of the research process, SSDQA benefits from evolving methodological practices. We provide an opportunity to reflect on ways to improve SSDQA effectiveness and reporting. Additional work will undoubtedly be needed as clinical research methods and data sources continue to advance.

Acknowledgments

The authors are grateful for the contributions of data scientists and researchers who shared their experience assessing study-specific data quality. They are also grateful to the patients, families, and clinicians in networks such as PEDSnet who make large-scale learning about ways to improve health possible.

Funding

This work was funded by Patient-Centered Outcomes Research Institute award ME-2020C3-21199. The funder played no role in the design of the study; method development; collection, analysis, or interpretation of the data; writing of the manuscript; or decision to submit it for publication. The views expressed in this paper do not represent those of the Patient-Centered Outcomes Research Institute or participating institutions.

Conflicts of Interest

JS is a current member of the Ronald McDonald House Charities of Greater Delaware, as well as a consultant for the Association of Child Life Professionals. C Blacketer is an employee of Johnson & Johnson, holds stock and stock options, and is involved in the development of the Observational Health Data Sciences and Informatics Data Quality Dashboard and CohortDiagnostics tools. KAM is involved in the development of the PCORnet Data Curation package. These conflicts did not interfere with the work performed in this project. No authors have any financial interest in the software tools developed during this work, nor have they been involved in the development of other software evaluated in this study. All other authors declare no other conflicts of interest.

References

1. Forrest CB, Margolis PA, Bailey LC, et al. PEDSnet: a national pediatric learning health system. J Am Med Inform Assoc 2014;21(4):602-606. [doi: 10.1136/amiajnl-2014-002743] [Medline: 24821737]
2. Forrest CB, McTigue KM, Hernandez AF, et al. PCORnet® 2020: current state, accomplishments, and future directions. J Clin Epidemiol 2021 Jan;129:60-67. [doi: 10.1016/j.jclinepi.2020.09.036] [Medline: 33002635]
3. Haendel MA, Chute CG, Bennett TD, et al. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. J Am Med Inform Assoc 2021 Mar 1;28(3):427-443. [doi: 10.1093/jamia/ocaa196] [Medline: 32805036]

XSL•FO
RenderX

4. All of Us Research Program Investigators, Denny JC, Rutter JL, et al. The “All of Us” research program. N Engl J Med 2019 Aug 15;381(7):668-676. [doi: 10.1056/NEJMsr1809937] [Medline: 31412182]
5. Tarabichi Y, Frees A, Honeywell S, et al. The Cosmos collaborative: a vendor-facilitated electronic health record data aggregation platform. ACI Open 2021 Jan;5(1):e36-e46. [doi: 10.1055/s-0041-1731004] [Medline: 35071993]
6. Drozda JP Jr, Graham J, Muhlestein JB, et al. Multi-institutional distributed data networks for real-world evidence about medical devices: building unique device identifiers into longitudinal data (BUILD). JAMIA Open 2022;5(2):ooac035. [doi: 10.1093/jamiaopen/ooac035] [Medline: 35663113]
7. Pedrera-Jimenez M, Garcia-Barrio N, Hernandez-Ibarburu G, et al. Building an i2b2-based population repository for COVID-19 research. Stud Health Technol Inform 2022 May 25;294:287-291. [doi: 10.3233/SHTI220460] [Medline: 35612078]
8. Ohno-Machado L, Agha Z, Bell DS, et al. pSCANNER: patient-centered Scalable National Network for Effectiveness Research. J Am Med Inform Assoc 2014;21(4):621-626. [doi: 10.1136/amiajnl-2014-002751] [Medline: 24780722]
9. Desai RJ, Marsolo K, Smith J, et al. The FDA Sentinel Real World Evidence Data Enterprise (RWE-DE). Pharmacoepidemiol Drug Saf 2024 Oct;33(10):e70028. [doi: 10.1002/pds.70028] [Medline: 39385712]
10. Joshua Lin K, Jin Y, Gagne J, et al. Longitudinal data discontinuity in electronic health records and consequences for medication effectiveness studies. Clin Pharmacol Ther 2022 Jan;111(1):243-251. [doi: 10.1002/cpt.2400] [Medline: 34424534]
11. Koffman L, Levis AW, Arterburn D, et al. Investigating bias from missing data in an electronic health records-based study of weight loss after bariatric surgery. Obes Surg 2021 May;31(5):2125-2135. [doi: 10.1007/s11695-021-05226-y] [Medline: 33462670]
12. Boyd AD, Gonzalez-Guarda R, Lawrence K, et al. Potential bias and lack of generalizability in electronic health record data: reflections on health equity from the National Institutes of Health Pragmatic Trials Collaboratory. J Am Med Inform Assoc 2023 Aug 18;30(9):1561-1566. [doi: 10.1093/jamia/ocad115] [Medline: 37364017]
13. Hubbard RA, Lett E, Ho GY, Chubak J. Characterizing bias due to differential exposure ascertainment in electronic health record data. Health Serv Outcomes Res Methodol 2021 Sep;21(3):309-323. [doi: 10.1007/s10742-020-00235-3] [Medline: 34366704]
14. Harton J, Mitra N, Hubbard RA. Informative presence bias in analyses of electronic health records-derived data: a cautionary note. J Am Med Inform Assoc 2022 Jun 14;29(7):1191-1199. [doi: 10.1093/jamia/ocac050] [Medline: 35438796]
15. Haneuse S, Arterburn D, Daniels MJ. Assessing missing data assumptions in EHR-based studies: a complex and underappreciated task. JAMA Netw Open 2021 Feb 1;4(2):e210184. [doi: 10.1001/jamanetworkopen.2021.0184] [Medline: 33635321]
16. Weiskopf NG, Dorr DA, Jackson C, Lehmann HP, Thompson CA. Healthcare utilization is a collider: an introduction to collider bias in EHR data reuse. J Am Med Inform Assoc 2023 Apr 19;30(5):971-977. [doi: 10.1093/jamia/ocad013] [Medline: 36752649]
17. Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. J Am Med Inform Assoc 2021 Sep 18;28(10):2251-2257. [doi: 10.1093/jamia/ocab132] [Medline: 34313749]
18. Castellanos EH, Wittmershaus BK, Chandwani S. Raising the bar for real-world data in oncology: approaches to quality across multiple dimensions. JCO Clin Cancer Inform 2024 Jan;8:e2300046. [doi: 10.1200/CCI.23.00046] [Medline: 38241599]
19. Feder SL. Data quality in electronic health records research: quality domains and assessment methods. West J Nurs Res 2018 May;40(5):753-766. [doi: 10.1177/0193945916689084] [Medline: 28322657]
20. Huser V, Li X, Zhang Z, et al. Extending Achilles heel data quality tool with new rules informed by multi-site data quality comparison. Stud Health Technol Inform 2019 Aug 21;264:1488-1489. [doi: 10.3233/SHTI190498] [Medline: 31438195]
21. Johnson SG, Speedie S, Simon G, Kumar V, Westra BL. Quantifying the effect of data quality on the validity of an eMeasure. Appl Clin Inform 2017 Oct;8(4):1012-1021. [doi: 10.4338/ACI-2017-03-RA-0042] [Medline: 29241241]
22. Kapsner LA, Kampf MO, Seuchter SA, et al. Moving towards an EHR data quality framework: the MIRACUM approach. Stud Health Technol Inform 2019 Sep 3;267:247-253. [doi: 10.3233/SHTI190834] [Medline: 31483279]
23. Khare R, Utidjian LH, Razzaghi H, et al. Design and refinement of a data quality assessment workflow for a large pediatric research network. EGEMS (Wash DC) 2019 Aug 1;7(1):36. [doi: 10.5334/egems.294] [Medline: 31531382]
24. Lee K, Weiskopf N, Pathak J. A framework for data quality assessment in clinical research datasets. AMIA Annu Symp Proc 2017;2017:1080-1089. [Medline: 29854176]
25. Mohamed Y, Song X, McMahon TM, et al. Tailoring rule-based data quality assessment to the Patient-Centered Outcomes Research Network (PCORnet) Common Data Model (CDM). AMIA Annu Symp Proc 2022;2022:775-784. [Medline: 37128433]
26. Pfaff ER, Girvin AT, Gabriel DL, et al. Synergies between centralized and federated approaches to data quality: a report from the national COVID cohort collaborative. J Am Med Inform Assoc 2022 Mar 15;29(4):609-618. [doi: 10.1093/jamia/ocab217] [Medline: 34590684]
27. Qualls LG, Phillips TA, Hammill BG, et al. Evaluating foundational data quality in the National Patient-Centered Clinical Research Network (PCORnet®). EGEMS (Wash DC) 2018 Apr 13;6(1):3. [doi: 10.5334/egems.199] [Medline: 29881761]

XSL•FO
RenderX

28. Sengupta S, Bachman D, Laws R, et al. Data quality assessment and multi-organizational reporting: tools to enhance network knowledge. EGEMS (Wash DC) 2019 Mar 29;7(1):8. [doi: 10.5334/egems.280] [Medline: 30972357]
29. Sidky H, Young JC, Girvin AT, et al. Data quality considerations for evaluating COVID-19 treatments using real world data: learnings from the National COVID Cohort Collaborative (N3C). BMC Med Res Methodol 2023 Feb 17;23(1):46. [doi: 10.1186/s12874-023-01839-2] [Medline: 36800930]
30. Taggart J, Liaw ST, Yu H. Structured data quality reports to improve EHR data quality. Int J Med Inform 2015 Dec;84(12):1094-1098. [doi: 10.1016/j.ijmedinf.2015.09.008] [Medline: 26480872]
31. Weiskopf NG, Bakken S, Hripcsak G, Weng C. A data quality assessment guideline for electronic health record data reuse. EGEMS (Wash DC) 2017 Sep 4;5(1):14. [doi: 10.5334/egems.218] [Medline: 29881734]
32. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 1;20(1):144-151. [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
33. Lewis AE, Weiskopf N, Abrams ZB, et al. Electronic health record data quality assessment and tools: a systematic review. J Am Med Inform Assoc 2023 Sep 25;30(10):1730-1740. [doi: 10.1093/jamia/ocad120] [Medline: 37390812]
34. Ozonze O, Scott PJ, Hopgood AA. Automating electronic health record data quality assessment. J Med Syst 2023 Feb 13;47(1):23. [doi: 10.1007/s10916-022-01892-2] [Medline: 36781551]
35. Wang TD, Henderson DW, Weber GM, et al. Understanding data differences across the ENACT federated research network. medRxiv. Preprint posted online on Jan 17, 2025. [doi: 10.1101/2025.01.17.25320686] [Medline: 39867368]
36. Gatto NM, Reynolds RF, Campbell UB. A structured preapproval and postapproval comparative study design framework to generate valid and transparent real-world evidence for regulatory decisions. Clin Pharmacol Ther 2019 Jul;106(1):103-115. [doi: 10.1002/cpt.1480] [Medline: 31025311]
37. Gatto NM, Vititoe SE, Rubinstein E, Reynolds RF, Campbell UB. A structured process to identify fit-for-purpose study design and data to generate valid and transparent real-world evidence for regulatory uses. Clin Pharmacol Ther 2023 Jun;113(6):1235-1239. [doi: 10.1002/cpt.2883] [Medline: 36871138]
38. Wang SV, Pinheiro S, Hua W, et al. STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies. BMJ 2021 Jan 12;372:m4856. [doi: 10.1136/bmj.m4856] [Medline: 33436424]
39. Diaz-Garelli JF, Bernstam EV, Lee M, Hwang KO, Rahbar MH, Johnson TR. DataGauge: a practical process for systematically designing and implementing quality assessments of repurposed clinical data. EGEMS (Wash DC) 2019 Jul 25;7(1):32. [doi: 10.5334/egems.286] [Medline: 31367649]
40. Bergquist T, Hassan M. DQe-c-v2. GitHub. URL: https://github.com/data2health/DQe-c-v2/ [accessed 2026-04-03]
41. Kapsner LA, Mang JM, Mate S, et al. Linking a consortium-wide data quality assessment tool with the MIRACUM metadata repository. Appl Clin Inform 2021 Aug;12(4):826-835. [doi: 10.1055/s-0041-1733847] [Medline: 34433217]
42. Struckmann S, Mariño J, Kasbohm E, Salogni E, Schmidt CO. dataquieR 2: an updated R package for FAIR data quality assessments in observational studies and electronic health record data. J Open Source Softw 2024;9(98):6581. [doi: 10.21105/joss.06581]
43. Rao GA, Shoaibi A, Makadia R, et al. CohortDiagnostics: phenotype evaluation across a network of observational data sources using population-level characterization. PLoS One 2025;20(1):e0310634. [doi: 10.1371/journal.pone.0310634] [Medline: 39820599]
44. Razzaghi H, Greenberg J, Bailey LC. Developing a systematic approach to assessing data quality in secondary use of clinical data based on intended use. Learn Health Syst 2021;6(1):e10264. [doi: 10.1002/lrh2.10264] [Medline: 35036548]
45. Kahn MG, Callahan TJ, Barnard J, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC) 2016;4(1):1244. [doi: 10.13063/2327-9214.1244] [Medline: 27713905]
46. Razzaghi H, Wieand K, Dickinson K, Bailey C. Data quality modules. PEDSpace. URL: https://pedsnet.org/metadata/communities/57fd85d3-0b50-4239-8f05-8db5e0e65a6a/browse/author [accessed 2026-03-27]
47. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3:160018. [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
48. Haux C, Knaup P. Using FAIR metadata for secondary use of administrative claims data. Stud Health Technol Inform 2019 Aug 21;264:1472-1473. [doi: 10.3233/SHTI190490] [Medline: 31438187]

Abbreviations

DQA: data quality assessment
SCD: sickle cell disease
SPIFD: Structured Process to Identify Fit-for-Purpose Data
SSDQA: study-specific data quality assessment
STaRT-RWE: Structured Template for Planning and Reporting on the Implementation of Real-World Evidence Studies

XSL•FO
RenderX

Edited by A Mavragani; submitted 23.Apr.2025; peer-reviewed by A Almeida, C Hebert, 민 임; revised version received 13.Jan.2026; accepted 28.Jan.2026; published 14.Apr.2026.

Please cite as:
Razzaghi H, Wieand K, Dickinson KL, Kahn MG, Roy J, Blacketer C, Christakis DA, Forrest CB, Greenberg J, Lehmann HP, Marsolo KA, Sciolla J, Weiner MG, Weiskopf NG, Bailey LC
Beyond Missingness: Systematizing Methods for Comprehensive Data Fitness Assessment in Clinical Research
J Med Internet Res 2026;28:e76398
URL: https://www.jmir.org/2026/1/e76398
doi:10.2196/76398

© Hanieh Razzaghi, Kaleigh Wieand, Kimberley L Dickinson, Michael G Kahn, Jason Roy, Clair Blacketer, Dmitri A Christakis, Christopher B Forrest, Jane Greenberg, Harold P Lehmann, Keith A Marsolo, Jennifer Sciolla, Mark G Weiner, Nicole G Weiskopf, L Charles Bailey. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Scalable Agile Framework for Execution in AI for Medical AI Ethics Policy Design in Small- and Medium-Sized Enterprises

Ion Nemteanu[1], BS, MS; Adir Mancebo Jr[2], BA, MA, PhD; Leslie Joe[2], MSDS; Ryan Lopez[2], BS; Patricia Lopez[2], ME, MASC; Warren Woodrich Pettine[3,4], MD

[1]Nemsee LLC, San Diego, CA, United States

[2]Data Science Alliance, San Diego, CA, United States

[3]Department of Psychiatry, University of Utah, 501 Chipeta Way, Salt Lake City, UT, United States

[4]Mountain Biometrics, Inc., Salt Lake City, UT, United States

Corresponding Author:
Warren Woodrich Pettine, MD
Department of Psychiatry, University of Utah, 501 Chipeta Way, Salt Lake City, UT, United States

Abstract

Artificial intelligence (AI) is transforming patient care, but it also raises ethical questions, such as bias and transparency. While a range of well-established frameworks exist to guide responsible AI practice, most were designed for academic or regulatory settings and can be hard to operationalize within fast-moving, resource-limited small and medium-sized enterprises (SMEs). We report on the collaborative design of the SAFE-AI (Scalable Agile Framework for Execution in AI), an approach that embeds ethical safeguards, including fairness, transparency, responsibility metrics, and continuous monitoring, directly into standard Agile development cycles. In keeping with established Agile principles, SAFE-AI provides "just enough structure" to integrate ethical oversight into existing workflows without prescribing extensive new governance layers. Similar to other Agile frameworks, such as Scrum, which is described as a "lightweight framework" designed to help teams solve complex problems through iterative learning and minimal process overhead, SAFE-AI aims to remain practical for organizations that may not have dedicated ethics or compliance staff. Rather than simplifying technical methods, SAFE-AI simplifies when and how ethical review is triggered and documented, making responsible AI practices feasible even in environments with limited ethics, governance, or compliance resources. SAFE-AI assumes the presence of qualified data scientists and engineers, and it does not replace the need for statistical or technical expertise but instead provides a lightweight structure for coordinating and documenting work that those experts already perform. We followed a design-science, practice-oriented approach over 20 weeks. After a discovery workshop, a cross-functional team was assembled that included SME employees, ethics researchers, and academic partners. The SME's role was limited to informing design constraints and feasibility considerations during the co-design phase. No operational pilot or production deployment was conducted as part of this study. To reduce the risk of internal design bias and improve generalizability, we also consulted external stakeholders through structured feedback sessions, including clinicians, health care domain experts, and regulatory specialists. Their feedback was incorporated into each prototype-feedback cycle, ensuring that priorities reflected not only the SME's immediate context but also broader clinical and regulatory perspectives. The co-design process produced a 4-phase SAFE-AI life cycle: discovery, assessment, development, and monitoring. SAFE-AI's phase-specific checklists meld acceptance, fairness, and transparency metrics into each Agile sprint. A novel scenario-based probability analogy mapping method was added to translate model risk and uncertainty into plain-language narratives for nontechnical stakeholders, forming the framework's core "responsibility metrics" layer. SAFE-AI is presented as a proposed framework showing that meaningful ethical safeguards can be embedded easily within common workflows used by SMEs that already use basic Agile or iterative development practices. Its checklist-driven phases and automatic review triggers provide a defensible way to track fairness, transparency, and responsibility throughout the model lifecycle.

(J Med Internet Res 2026;28:e80028) doi:10.2196/80028

KEYWORDS

digital health; eHealth; artificial intelligence; machine learning; medical ethics; small business; agile methodology; software development life cycle; risk assessment; fairness metrics; transparency; trustworthy AI

XSL•FO
RenderX

Introduction

Background

Artificial intelligence (AI) systems deployed in health care, whether directly in clinical decision-making or through supporting technologies, carry significant ethical responsibilities. Their outputs can influence diagnoses, treatment options, operational workflows, and ultimately patient well-being. For organizations developing or integrating these tools, including small and medium-sized enterprises (SMEs) that support patient care, ethical considerations, such as fairness, transparency, and accountability, are essential from the outset.

While significant academic work has explored the ethics of deploying AI in medical decision support systems, these insights are often inaccessible to SMEs [1-6]. The complexity of AI ethics, including concepts like fairness, transparency, and accountability, presents a significant barrier, as does the prohibitive cost of dedicated ethics teams. Furthermore, many existing toolkits lack the practical guidance or real-world case studies necessary for SMEs [7,8]. Additionally, embedding ethics into AI development can be transformative for SMEs. Investments in ethical governance contribute to the value of intangible assets, such as intellectual property, loss aversion, brand reputation, and organizational goodwill. In health care markets, where regulatory trust and patient safety are paramount, demonstrable ethical alignment can also serve as a differentiator that attracts ethically aware customers, clinical partners, and investors. As a result, operationalizing AI ethics is not merely a compliance strategy but a value-creating asset for emerging technology companies. Proactive approaches tend to emphasize ways in which investments drive important intangibles that are reflected in a differentiated reputation and robust corporate culture. Mature firms link these intangible outcomes directly to investments in AI ethics and governance, using criteria, such as recognition of industry leadership and support of environmental, social, and governance efforts, as well as the long-term ability to manage risks [9].

In April 2024, HHS OCR (Department of Health and Human Services - Office for Civil Rights) clarified that decision-support tools, including AI, must not produce discriminatory outcomes under §1557 of the Affordable Care Act [10]. Outside software as a medical device (SaMD) pathways, comprehensive federal requirements remain unsettled. The US Food and Drug Administration (FDA) oversight and emerging guidance coexist with evolving proposals [11-13]. In this environment, organizations must anticipate stricter expectations while operating with practical, defensible governance. SAFE-AI (Scalable Agile Framework for Execution in AI) offers a pragmatic path for SMEs to meet today's expectations and prepare for tomorrow's.

To facilitate SMEs in addressing this challenge, we introduce a simplified framework in alignment with established Agile principles for responsibly integrating AI in medical settings. SAFE-AI aims to provide "just enough structure" to guide ethical oversight without imposing extensive new governance requirements, consistent with industry definitions of lightweight frameworks, such as Scrum, which emphasizes minimal process overhead and iterative adaptation [14,15,16]. Using this framework will help ensure SMEs develop medical AI solutions that meet the standard of large hospital systems and prioritize patient care.

Emphasizing AI responsibility, SAFE-AI addresses fairness, transparency, and ethical deployment in a manner proportional to organizational maturity and available expertise. A well-defined process facilitates collaboration between health care practitioners and data scientists, ensuring that ethical considerations, fairness, and equity are systematically assessed and documented. This approach ensures that each phase of AI development aligns with these critical values, leading to more effective and responsible AI tools for health care.

Study Design and Project Context

This work followed a design-science, practice-oriented methodology to develop SAFE-AI, a scalable internal governance framework intended for broad adoption by SMEs building medical-AI products. The project was a collaboration among Medical Timeseries Networks (MTN; an SME), the Data Science Alliance (DSA), Nemsee LLC, and the University of Utah. MTN served as the primary design partner, while DSA, Nemsee, and academic partners contributed ethics, regulatory, and methodological expertise. MTN's role in this study was limited to articulating operational constraints, workflow realities, and feasibility considerations during framework development; no operational deployment or live pilot was conducted as part of this study.

The purpose of the consortium was to embed ethics, regulatory, and methodological expertise into the SAFE-AI framework itself, rather than to provide ongoing expert participation in day-to-day development workflows.

Informal Evidence Gathering

Between December 2024 and April 2025, the design team conducted an ad hoc search of peer-reviewed literature, regulatory guidance (eg, FDA, SaMD, Good Machine-Learning Practice [17], HHS §1557 rule [15], and EU (European Union; AI Act drafts) [18], and industry white papers. Searches were performed opportunistically in PubMed, Scopus, IEEE Xplore, and gray-literature repositories as questions arose during development. All relevant sources were captured in a shared Zotero library and tagged by topic. Key findings were reviewed in weekly design committee meetings.

Framework Outputs

The co-design process yielded:

1. A 4-phase life-cycle diagram (discovery, assessment, development, and monitoring)
2. Stage-specific checklists covering acceptance, fairness, and transparency metrics
3. Practical implementation guidance (eg, role assignments and artifact templates) for each phase.

XSL•FO
RenderX

Consensus and Documentation

All design decisions were made through real-time discussion until unanimous verbal agreement was reached. No formal voting or Delphi procedures were used. Version history, meeting minutes, and rationale for changes were maintained in a locked Google Drive folder that serves as the project's audit trail.

Future Work: Planned Pilot Implementation and Evaluation

The following pilot implementation and evaluation activities are planned for future work and were not conducted as part of the present design study. SAFE-AI will be applied to 2 upcoming MTN initiatives, including a workflow-orchestration product and a data-engineering pipeline. An evaluation plan was prespecified to capture:

1. Process metrics: time (hours) required to complete each SAFE-AI phase.
2. Governance metrics: count and category of fairness metrics defined per project.
3. Usability: postimplementation survey of participating staff (5-item Likert scale; 1=strongly disagree, 5=strongly agree).

Descriptive statistics will summarize these data after the first pilot quarter.

Ethical Considerations

The project involved policy development only; no human participants or identifiable health data were used. Institutional Review Board (IRB) oversight was therefore not required.

SAFE-AI Ethical Evaluation Framework

We present the SAFE-AI, which is a structured process designed to embed ethical evaluation into the technical development life cycle of AI products (summary in Figure 1, full process in Figure 2). A concise checklist distinguishing required and recommended steps is provided in Checklist 1 to support implementation consistency. SAFE-AI enters the workflow after business requirements or system objectives have been defined and prioritized, and tasks are ready for execution by the development team. It complements Agile and Scrum methodologies by introducing structured ethical oversight at critical stages without disrupting project momentum.

Figure 1. SAFE-AI (Scalable Agile Framework for Execution in AI) summary workflow, highlighting each core phase and feedback loop from prioritization to deployment and monitoring. AI: artificial intelligence.

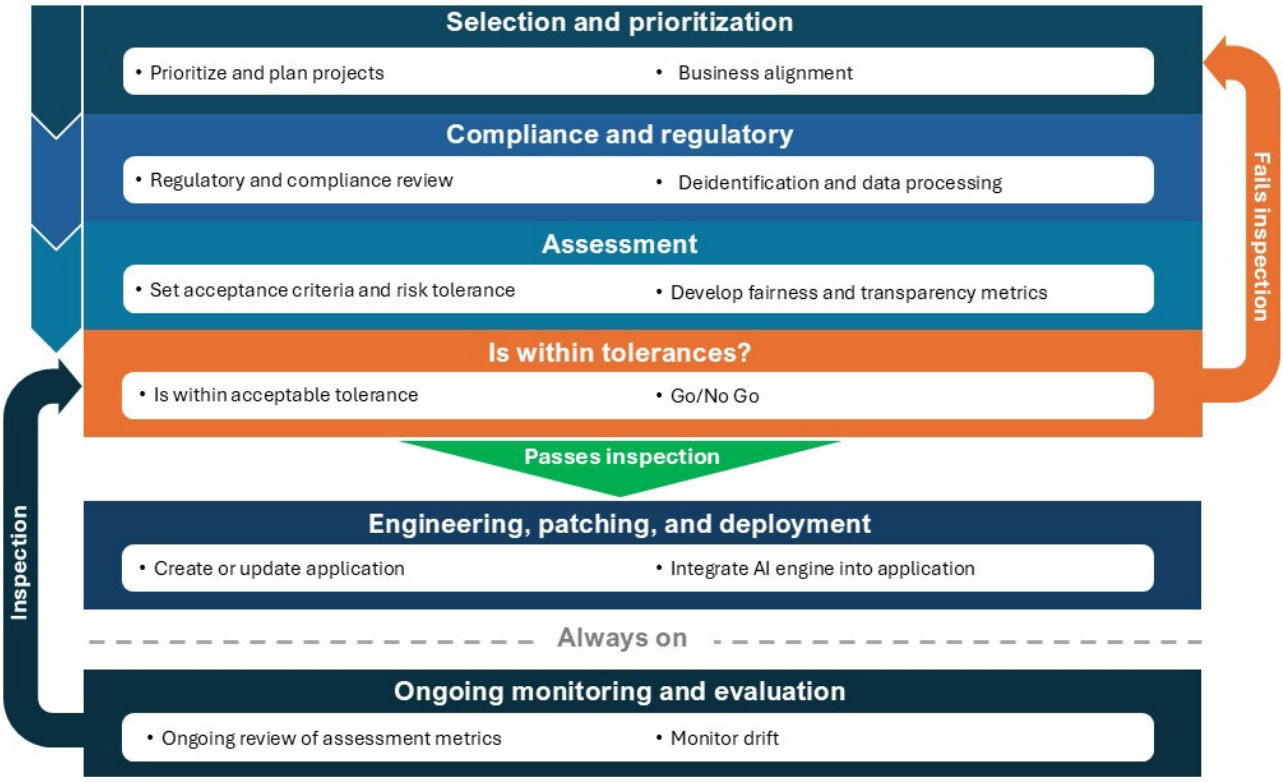

XSL•FO
RenderX

Figure 2. Detailed ethical evaluation process SAFE-AI (Scalable Agile Framework for Execution in AI) - conducted continuously throughout product development as illustrated above. The process begins with a discussion of priorities and alignment and proceeds through model building and inference implementation, emphasizing appropriate tolerance levels at each iteration. AI: artificial intelligence; SDLC: software development life cycle.

XSL•FO
RenderX

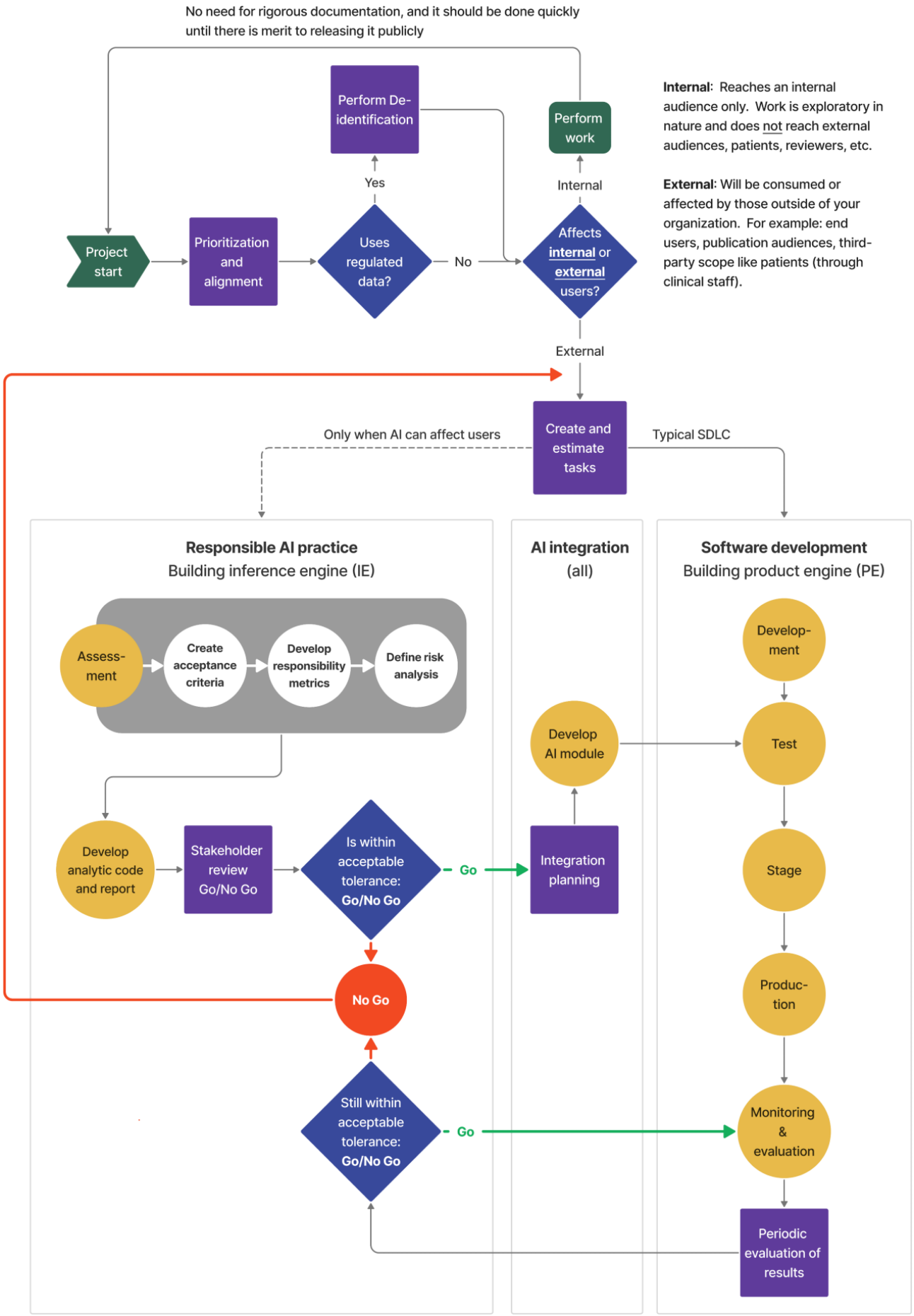

XSL•FO
RenderX

The SAFE-AI framework is not a one-time gate. It is designed to be re-entered at any point in the development lifecycle whenever changes occur that could materially affect the behavior, fairness, or transparency of the AI system. This includes retraining with new data, hyperparameter adjustments, updates to feature selection, or integrating external datasets.

Decisions to reactivate safe-ai should be made jointly by the product owner and technical lead, in consultation with the cross-functional team during natural Agile "sprint review" mechanisms when all the stakeholders are already present. This ensures that ethical checkpoints are triggered not only by technical changes (eg, retraining, parameter tuning) but also by product or stakeholder concerns (eg, shifts in clinical use cases). Embedding this responsibility into sprint reviews and release planning creates a consistent governance mechanism without requiring a separate oversight body.

Practitioners are encouraged to revisit SAFE-AI whenever there are any modifications to their models, regardless of perceived simplicity. Even seemingly minor technical tweaks may cause unintended shifts in model performance or equity. Further recommendations are outlined in the section titled "Quick Note on Model Updates and Considerations," where we suggest re-evaluating the model against all the previously established metrics (refer to Part 2 [Assessment]) to ensure continued ethical integrity.

In this paper, the reference to a "data science team" is used as a generalized term to encompass data scientists and the other skilled roles and internal stakeholders that contribute to the project, including product and program managers, data engineers, analysts, statisticians, software engineers, and development operations personnel. However, when referring specifically to data scientists, we mean those roles that are directly responsible for researching and building the AI algorithms for use in the business and in their products.

SAFE-AI: Part 1 (Discovery)

Determine if the Project Will Use Regulated Data

First, we will determine if the planned work will use or access regulated data. Organizations must comply with all relevant regional, state, federal, and local laws before any access and manipulation of data begins. Protected health information and personally identifiable information are among the many regulated data types. Existing acts, such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) set specific compliance requirements.

Best practices include anonymizing or aggregating data to protect individual privacy. Additionally, informed consent is required for legally effective participation in research and must be obtained in accordance with the principle of respect for persons [19]. This principle ensures that individuals are given the opportunity to choose whether they wish to participate in research studies.

A comprehensive regulatory review should be conducted before any AI development begins. If specific handling techniques require documentation, they should be securely stored and referenced throughout development. When designing systems that impact external audiences, regulatory requirements should be listed as acceptance criteria and revisited at every development stage. If the same data are used repeatedly, the regulatory review can cover repeated use and should have a documented process (if necessary) to ensure compliance for future access and proper use. This will ensure that resources do not have to review the data every time it is accessed.

Determine if Scope of Work Impacts Internal or External Users

iDetermine External Effect

During typical prioritization meetings, it should be discussed whether the work will affect external stakeholders like patients, clinicians, or other health care providers. This is an important distinction since there are many tasks that should be done quickly and match the pace of the organization and do not require the extra burden of documentation. Realistically, not every task requires burdensome processes and documentation. This framework aims to only target the tasks that have a direct effect on the health care outcomes as a direct result of the product being developed. The first step in this process is a quick discussion during the planning meeting that determines the upcoming scope of work for the development team and clarifies if outputs can affect external stakeholders. Clarifications are detailed below.

Internal Stakeholders

Project work that includes exploratory research, feasibility studies, or internal management tools that do not directly impact patient outcomes can proceed without additional scrutiny. The process involves moving forward with the task as normal, without the need for additional scrutiny. Upon completion, the process is restarted from the beginning.

External Stakeholders

Work that affects those outside of your organization. For example, patients, clinicians, or other health care providers; end users, and publication audiences. In this case, the outcome of these tasks can or will have a direct impact on an external audience. Internal projects may also carry external implications. For instance, human resource initiatives that use protected employee data involve internal stakeholders, but mishandling that data can create risks that extend beyond the organization. The goal is to recognize when the project scope introduces potential external risk, even if it originates internally. In this case, the process involves moving on to the next step.

Create and Estimate Tasks

This phase begins with breaking down high-level project goals into smaller, manageable user stories or tasks, as seen in Agile Software Development Processes. These tasks often span data ingestion, cleaning, feature engineering, model development, evaluation, and integration. During planning, the data science team estimates each task's duration based on the scope of work for iterative delivery while leaving room for learning, model iteration, and integration into production workflows.

To ensure that ethical and social considerations are not deferred until the end of development, or left out entirely, this framework incorporates structured ethical reasoning into the task creation

XSL•FO
RenderX

process from the start. The foundation for this integration draws from the RESOLVEDD process (Review, Estimate, Solutions, Outcomes, Likely impact, Values, Evaluate, Decide, and Defend), a structured ethical decision-making framework originally introduced by Pfeiffer and Forsberg [20,21]. RESOLVEDD guides practitioners through steps, such as reviewing the facts, estimating the ethical conflict, surveying alternatives, opening options to stakeholders, listing risks, valuing impacts, evaluating options, deciding, and defending the decision. While originally applied to workplace ethics, the logic of RESOLVEDD has broad applicability in responsible AI contexts.

Building on this foundation, Vakkuri et al. [22] introduced ECCOLA, a card-based toolkit tailored to Agile AI and autonomous systems development. ECCOLA does not follow an acronym, but its sprint-by-sprint format maps well to the RESOLVEDD process. It consists of a deck of 21 cards aligned to 8 core AI ethics themes (eg, transparency, accountability, and bias), each designed to provoke discussion, surface risks, and frame decisions during active development. The cards can be used in planning meetings to ensure that ethical issues are not only acknowledged but also actively scoped into the project as discrete tasks.

SAFE-AI adapts this card logic but moves one step further: rather than selecting from a predefined deck, teams create their own custom cards directly from the objectives and risks identified in the project. Each card captures the problem, the related ethical concern, and the concrete analysis tasks that will address it. These cards are then written into the backlog as user stories or research spikes, estimated during sprint planning, and assigned clear ownership. By embedding them directly into the backlog, ethical concerns are tracked with the same visibility and accountability as technical work. For example, a card generated from the objective “Minimize Gender Bias” might be translated into a user story with acceptance criteria tied to distributional checks, fairness metrics, and subgroup error analyses.

This approach converts abstract ethical principles into concrete, time-boxed actions that can be tracked and revisited throughout the sprint cycle. It ensures that responsibility is not just a post hoc evaluation, but a proactive part of scoping and estimating work. By embedding the card principle in this way, SAFE-AI enables collaborative assignment of each card and its tasks, supports ongoing discussion between stakeholders, and provides a shared method for everyone to follow. A simple template for these cards anchors the ethical problem with other “User Stories,” with the analysis tasks listed below (Table 1).

Table . SAFE-AI[c] ethical card template for planning and backlog integration (eg, Minimize Gender Bias[a,b]).

What to do	Output
Training data balance. Quantify overall and conditional distributions by gender (and salient covariates)	• Distribution tables and joint density plots
Split integrity. Verify gender balance across training, validation, and testing (or rolling production batches).	• Per-split distribution summaries
Performance by subgroup: Report declared metrics by gender (eg, AU-ROC[d], F_1-score, PPV[e], or NPV[f] calibration) and compute gaps and confidence intervals.	• Per-subgroup metric table • Gap estimates • Calibration plots
Downstream effect at acceptance testing: Compare consequential outcomes by gender (eg, referral or redirect rates) under final acceptance criteria.	• Outcome distributions with statistical tests and effect sizes

[a]Do all genders have similar performance and minimized bias?

[b]Problem: The model may yield systematically different performance and error profiles across genders, a protected characteristic in health care decision support.

[c]SAFE-AI: Scalable Agile Framework for Execution in AI.

[d]AU: area under the receiver operating characteristic.

[e]PPV: positive predictive value.

[f]NPV: negative predictive value.

These tasks are customized per project and sprint. They should be time-boxed and assigned within the sprint backlog, ensuring fairness concerns are scoped into the same process as technical work.

Integrating Into Agile Software Development Life Cycle

In practice, these SAFE-AI cards should be treated in the same way as user stories or research spikes within Agile planning. During sprint planning, teams can pull the relevant cards into the backlog and frame them as discrete work items with acceptance criteria, effort estimates, and clear ownership. This ensures that ethical questions are not abstract side discussions but concrete backlog items tracked alongside feature development. For example, a card highlighting fairness concerns may become a “bias audit story,” or a transparency card may generate a “model interpretability spike.” By embedding these cards directly into sprint discussions and backlog management, responsibility metrics are scoped, prioritized, and delivered with the same visibility as any other technical deliverable without introducing new governance layers or slowing iterative delivery.

Importantly, integrating SAFE-AI into Agile requires only minimal additional ceremony. Teams typically spend about 50 minutes during preplanning to surface ethical risks and create the initial cards and approximately 10 minutes during

XSL•FO
RenderX

sprint-planning ceremonies to refine and convert those cards into user stories or spikes (Figure 3). These brief additions occur within meetings that already exist, meaning the procedural overhead is minimal. The more substantial time investment (generally up to 6 hours within the sprint) is not overhead created by SAFE-AI itself but rather the time spent actually performing the ethical tasks, such as fairness checks, interpretability analyses, drift reviews, or documentation improvements. In many teams, these activities either occurred informally, inconsistently, or were never undertaken at all. SAFE-AI simply makes this work visible, trackable, and accountable by assigning it a discrete story with an owner and acceptance criteria.

Figure 3. SAFE-AI (Scalable Agile Framework for Execution in AI) card workflow integrated into the Agile software development life cycle, including an estimate of the additional time required to incorporate SAFE-AI activities into standard Agile ceremonies. In preplanning, project objectives are defined, and ethical risks are identified, leading to the creation of custom cards that specify the problem, analysis tasks, and outputs. During sprint planning, these cards are converted into backlog items, such as user stories or research spikes, estimated, assigned, and given acceptance criteria. In the sprint phase, the ethical tasks are executed and validated alongside technical stories, ensuring responsibility metrics are tracked within the same Agile workflow with about 1-7 hours of additional time per sprint depending on the needs.

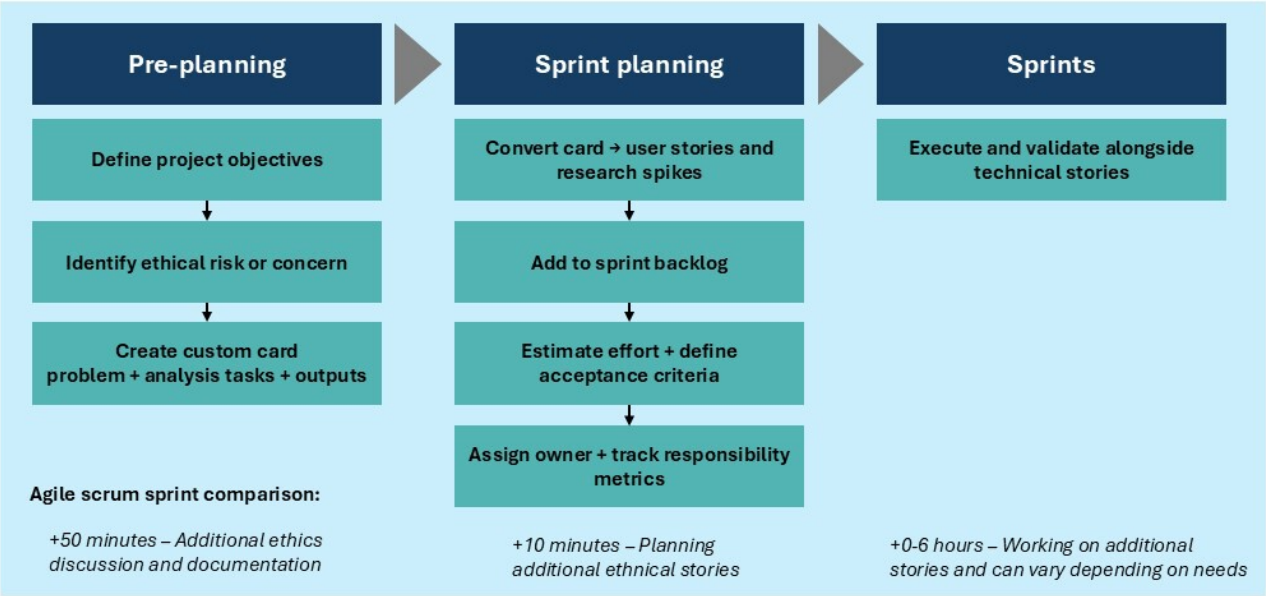

SAFE-AI cards progress from the identification of ethical risks in preplanning to conversion into backlog items during sprint planning and finally to execution and validation alongside technical stories (Figure 3). This demonstrates that SAFE-AI adds only a small amount of process time while ensuring that ethical responsibilities are now explicitly planned, resourced, and delivered with the same rigor as technical features.

SAFE-AI: Part 2 (Assessment)

Overview

This section occurs during the inference engine (IE) phase and outlines key steps the data science team, both technical and product stakeholders, should complete before progressing to development and integration. These steps ensure that AI IEs operate within defined boundaries of uncertainty and acceptable risk, aligning with ethical standards and strategic business goals.

It is important that the following steps be incorporated alongside traditional software testing as a routine part of development practice, with consistent attention to their execution. While these steps may initially appear to add time or complexity—requiring deeper evaluation, more test cycles, and added stakeholder input—when planned appropriately, these activities do not disrupt schedules. They are instead embedded naturally as part of the process and become part of the schedule.

The SAFE-AI framework is designed to be iterative and collaborative, ensuring that ethical assessments are embedded incrementally and transparently throughout the development lifecycle. Crucially, tasks are not imposed in isolation, as they are co-developed and agreed upon by all stakeholders, including data scientists, engineers, product leads, and domain experts, which keeps timelines aligned and priorities clear.

Ethical testing is treated with the same rigor and visibility as functional testing, making it easier to forecast and manage. While often overlooked due to perceived resource constraints or competitive pressure, these steps are essential to building systems that are robust, compliant, and trustworthy. The framework's goal is to define the minimum necessary ethical safeguards required to ensure responsible AI while preserving agility and long-term strategic value.

There are 3 tasks that are especially relevant during the assessment phase, each of which is detailed below

Creating Acceptance Criteria

Acceptance criteria establish the conditions a system must meet before release. Borrowed from traditional software development life cycle (SDLC), they provide clear, testable expectations. In AI systems, this still begins with baseline functionality "Does the model work as intended?" For example, a clinical diagnostic tool might require ≥95% accuracy overall before being considered ready for integration into workflows. Such criteria

XSL•FO
RenderX

help establish shared expectations and support regulatory alignment before AI models are integrated into clinical workflows. Atlassian describes acceptance criteria as “clear, concise, and testable statements that focus on providing positive customer results” [23,24].

In AI systems, especially those used in high-stakes domains like health care, defining acceptance criteria is not just a technical step—it is a normative one. The choice of what counts as “acceptable” performance can have real-world implications for safety, equity, and trust. It is therefore essential that these baseline criteria be co-developed with key stakeholders, including clinicians, patients, domain experts, and end users. AI engineers may assume they understand what matters most to users, but this assumption often overlooks critical perspectives—particularly around risk tolerance, fairness, and real-world utility. Engaging stakeholders in defining success conditions helps ensure that the system reflects the values and priorities of those it is intended to serve.

Acceptance criteria are only the first step. Once minimum functionality is defined, additional layers of responsibility metrics are needed to answer two further questions: “Does it work fairly, and is it understandable to those making directions on behalf of the company and those it serves?” Fairness and transparency extend acceptance criteria into the ethical domain, ensuring that the system not only functions but also functions responsibly.

Develop Responsibility Metrics

The Importance of Responsibility Metrics

Once baseline acceptance criteria confirm that a model works, the next step is to evaluate how it will work in practice. Responsibility metrics extend beyond functionality to address whether the system is fair, trustworthy, and aligned with stakeholder values. These metrics help organizations detect bias, communicate uncertainty, and ensure that model outputs can be understood and acted upon responsibly. Together, they provide the ethical guardrails that acceptance criteria alone cannot supply.

Fairness Metrics

Bias and equitable performance across groups within AI systems will be defined and measured. As BSA (Business Software Alliance [25]) states, fairness is subjective and requires selecting the most appropriate metrics for the system being developed. These metrics should be documented and reviewed collaboratively by subject matter experts and data scientists [25]. In practice, fairness metrics are used to detect and mitigate disparate impacts, ensuring that one group or class (eg, based on race, gender, age, or geography) is not disproportionately harmed or excluded compared to others. This may involve evaluating performance across subgroups, measuring disparities in false positives or negatives, or assessing equal access to beneficial outcomes. The goal is not only to identify statistical imbalance but also to actively choose metrics that reflect the ethical priorities and risk tolerances of the system’s intended users and affected populations [26-30]

Transparency Metrics

Metrics will be defined that focus on ensuring that AI model predictions are interpretable, explainable, and actionable by end users. One technique used in this framework is a scenario-based probability analogy mapping method (SPAMM)—a structured approach to communicating AI predictions by framing them as outcome narratives grounded in real-world context and relevant to the business goals. This method, which we refer to as the SPAMM, enhances transparency by mapping probabilities of outcomes related to fairness metrics and other acceptance and risk criteria to familiar analogies that illustrate the likelihood of different scenarios, the conditions under which errors occur, and their potential consequences. Rather than focusing solely on model internals, SPAMM presents predictions in narrative form - conveying not only what the model predicts, but why, how alternative outcomes compare, and which inputs most influence the result (Figure 4). This approach builds stakeholder trust by making uncertainty and risk both visible and relatable.

Figure 4. Example of scenario-based probability analogy mapping method compared to technical metrics. AI: artificial intelligence; SHAP: Shapley Additive Explanations; SPAMM: scenario-based probability analogy mapping method.

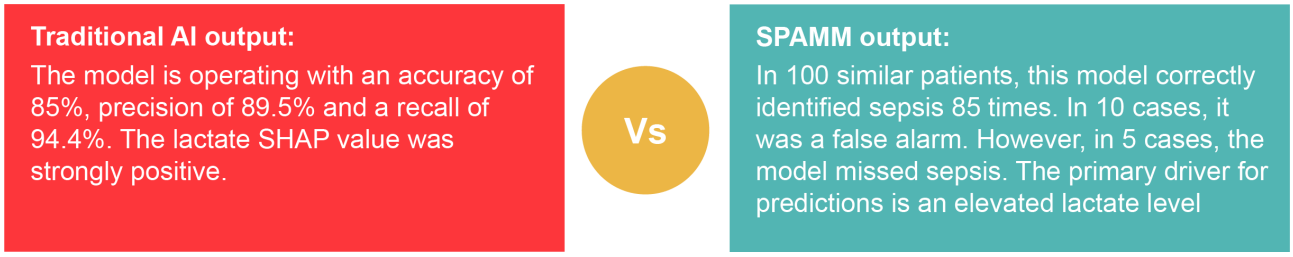

This approach is particularly important in high-stakes domains like health care, where decision-makers, such as clinicians, patients, or administrators must weigh the risks of false positives and false negatives. By highlighting the probability of misclassification and mapping those scenarios to familiar contexts, stakeholders can assess how often and under what circumstances the model may be wrong - and what the real-world consequences of those errors might be. For example, if a model is correct 85% of the time but consistently fails on a particular patient subgroup, this insight can guide mitigation strategies, prompt workflow adjustments, or justify additional human oversight. These dimensions of model behavior—error distribution, uncertainty, and impact—are often absent in standard AI evaluation pipelines.

Traditional interpretability tools like SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations), partial dependence plots (PDP), and accumulated local effects (ALE) are valuable for understanding model

XSL•FO
RenderX

behavior. SHAP, which is based on cooperative game theory, assigns each feature an importance value representing its contribution to a particular prediction—similar to calculating how much each player contributed to a team's win. LIME works by perturbing the input data near a given instance and fitting a simple model to explain the local decision boundary. PDP and ALE methods help assess how the average model output changes when a single feature is varied, which supports global interpretability, especially when exploring relationships between inputs and outcomes.

These methods help data scientists and auditors understand what the model is doing, but they may not fully address the kinds of interpretive needs that arise when communicating overall impact with clinical or business stakeholders [31-35]. Technical performance metrics that evaluate true and positive rates, such as root mean square error, precision, and/or recall, can be difficult to translate to nontechnical audiences. As Ribeiro et al [36] noted, trust in machine learning (ML) models depends on the user's ability to anticipate and understand model behavior in practical terms— not just accuracy, but reliability and risk.

As a result, we apply SPAMM, which adds a critical layer to the transparency toolkit by reframing predictive outputs in terms that mirror how people naturally reason under uncertainty. At its core, the method walks users through a model's output by anchoring it to the specific input conditions that led to that prediction and by showing how often those conditions have historically produced similar outcomes.

Presenting outputs in this context helps stakeholders understand not just what the model predicted, but why it arrived at that decision, how likely alternative outcomes were, and how much confidence they can place in it. While this can become complex in models with high-dimensional input spaces or large numbers of interacting variables, the point is not to replicate the entire model logic, but to provide a real-world testing lens that frames predictions through the lens of practical "what-if" scenarios. This helps surface critical edge cases, increase interpretability, and build trust, especially in settings where human oversight and risk-based decisions are essential [37-39].

Crucially, this framing also enables teams to assign tangible costs to errors by explicitly asking "What can happen if this prediction is wrong?" and "What is the risk impact to the organization or patient?" For example, a false negative in a sepsis prediction model may delay critical treatment, leading to clinical harm, regulatory exposure, and reputational damage. A false positive in an insurance fraud detector may result in a wrongful denial of benefits, eroding user trust and increasing legal liability. By mapping these risks to predicted outcomes and associated variables, organizations can better understand which errors matter most, estimate their financial or operational consequences, and prioritize safeguards accordingly. Integrating these cost-of-error considerations into the transparency process transforms interpretability from a purely technical exercise into a tool for risk mitigation and strategic alignment [40,41].

Ultimately, this approach supports a more robust, stakeholder-centered view of transparency. It enables users to understand reliability, anticipate error, and make informed decisions based on modeled outcomes—an essential capability for ethical and accountable AI deployment in health care and other sensitive domains (Table 2).

Table . Illustrative acceptance criteria and responsibility metrics across common project types, including enterprise automation, consumer-facing artificial intelligence, and clinical decision support.

Project type	Acceptance criteria	Responsibility metrics
Enterprise automation	• API[a] response time ≤200 ms • Workflow accuracy ≥98% across standard inputs	• Fairness: Ensure no subgroup of users is excluded from automation • Transparency: Apply SPAMM[b] to communicate error risks, automation actions, and results, and maintain audit logs
Consumer-facing AI	• Recommendation relevance ≥35% via click-through, ≥85 satisfaction • Subgroup accuracy within ±5%, ≥10% Interaction with the next action	• Fairness: Balanced recommendation exposure across demographic groups • Transparency: Apply SPAMM to explain recommendations, actions, and results, and expose key features
Clinical decision support	• Diagnostic accuracy ≥95% overall • Subgroup disparity ≤5% across age, race, and gender	• Fairness: Subgroup error analysis, including false positives and negatives • Transparency: Apply SPAMM to frame predictions and risks in a clinical context

[a]API (application programming interface):

[b]SPAMM (scenario-based probability analogy mapping method):

This framework invites further development in areas such as visualization, user-centered metrics, and storytelling techniques for AI transparency. While this paper does not formalize the full methodology, it highlights the importance of designing explanation strategies that reflect how real-world users evaluate uncertainty, reliability, and relevance—not just technical performance.

In addition to defining fairness and transparency requirements, teams should specify the statistical methods that will be used later in monitoring to verify whether these metrics hold in

XSL•FO
RenderX

production. For example, subgroup outcome disparities can be tracked using chi-square or Fisher exact tests, distribution shifts can be detected with Kolmogorov–Smirnov or Jensen–Shannon divergence, and overall population stability can be assessed with Population Stability Index (PSI). By embedding these statistical guardrails directly into the acceptance criteria, SAFE-AI ensures that monitoring processes in Part 4 are both preplanned and directly linked to the responsibility metrics established during development.

Define Risk Analysis

Risk analysis should be conducted collaboratively, including consultation with relevant stakeholders like product teams, data scientists, and legal experts. There are many possible approaches, and the following are a few examples of structured risk analysis processes that already exist in the field of AI and health care.

1. The NIST (National Institute of Standards and Technology) AI Risk Management Framework provides structured guidelines for evaluating risks in AI systems, emphasizing trustworthiness and resilience through iterative assessments [42,43].
2. McKinsey's "Derisking AI by Design" approach highlights the necessity of integrating risk management throughout the AI lifecycle, focusing on proactive monitoring and governance [44].
3. Stanford's AI health care risk assessment framework outlines liability risks in medical AI applications, suggesting adaptive monitoring strategies to address unforeseen ethical and operational challenges [45].
4. The US Food and Drug Administration's (FDA) Total Product Life Cycle (TPLC) framework ensures continuous oversight from AI model development through postmarket surveillance. The FDA writes how generative artificial intelligence (GenAI) presents potential risks that may require varying levels of risk controls for different applications, as is true of other technologies. For example, hallucinations, particularly those that may appear to be authentic outputs to users, may present a significant challenge in certain health care applications where highly accurate, truthful information is critical. They state how current evaluation approaches, such as those used to evaluate computer-assisted triage, detection, and diagnostic devices, may still be applicable for GenAI-enabled devices, albeit with additional supporting evidence. However, it may be challenging to determine the evidence that may be needed for certain GenAI implementations [46,47].

All of the above approaches include hazard identification, risk assessment, and mitigation strategies tailored to AI's dynamic nature in health care. By incorporating structured risk evaluation, teams can proactively address potential harms and regulatory challenges. Key considerations include:

1. Misuse potential: Could the product be used incorrectly, leading to harm?
2. Error costs: What are the consequences of false predictions?
3. Societal implications: Are there broader ethical concerns?

Legal experts should be engaged to ensure compliance with evolving regulatory landscapes and provide specific tests and metrics that can be used to create risk boundaries for AI models that can help pass regulatory reviews as well as build trust with the public.

Quick Note on Model Updates and Considerations

Once an IE has been approved and integrated into a production environment, it is common practice to revisit the model for improvements through techniques, such as retraining with new data, hyperparameter tuning, or feature modifications. While these changes may be perceived as incremental or routine, they carry the potential to significantly alter the algorithm's behavior, calibration, and fairness properties. Even minor updates can reintroduce latent biases, shift subgroup performance, or affect the interpretability and reliability of outputs.

From an ethical standpoint, any modification to the IE, including changes to a single model or updates to the system of models, should be treated as a new model iteration when following the SAFE-AI framework, regardless of whether the foundational architecture remains unchanged. Although it is not necessary to recreate the entire ethical evaluation pipeline, it is considered best practice to reapply the original assessment metrics established during initial development (outlined in Part 2 [Assessment]). This allows teams to validate that the updated model continues to meet ethical expectations for fairness, transparency, and performance consistency.

In high-stakes domains, such as health care, the tension between deployment speed and responsible oversight is a persistent challenge. Rapid iteration is often essential for competitiveness; however, this urgency must be weighed against the risk of unintended harm. Ethical oversight should not be perceived as a barrier but rather as a safeguard that ensures continuity of ethical alignment across the model's lifecycle.

Therefore, organizations are strongly encouraged to adopt a policy of proactive reassessment following any material change to an AI system. By incorporating these evaluations into their standard operating procedures, teams can preserve ethical integrity, mitigate risks, and maintain regulatory and stakeholder confidence.

SAFE-AI: Part 3 (Development and Integration)

Transition to Development

After defining metrics and performing the necessary reviews, the next step is to move from validated concepts into functional systems. This phase encompasses 3 key stages already defined in Part 2: the IE, the product engine (PE), and AI integration.

1. IE: As described earlier, the IE represents the culmination of algorithm development, transforming exploratory research into validated, production-ready code. Activities include hypothesis testing, feature engineering, model training, and validation, until the algorithm meets defined SAFE-AI thresholds for accuracy, risk, and subgroup reliability.
2. PE: In parallel, product engineers build the functional application layer. Here, validated models are embedded into the product architecture so they can import data,

XSL•FO
RenderX

process it through the IE, and output results for downstream systems and users. Inputs, outputs, and thresholds are explicitly defined and jointly agreed upon. This work ensures the product is both technically robust and usable for stakeholders.

3. AI integration: Integration is the convergence point where the IE and PE are combined into a single, deployable system. At this stage, focus shifts from feasibility to operational viability, ensuring compatibility with data pipelines, embedding monitoring systems, and preparing the organization for deployment (training, regulatory compliance, user support, and related functions).

Together, these stages operationalize the SAFE-AI framework by moving from validated research into accountable, production-ready systems that can be deployed and monitored in real-world settings.

Develop Analytic Code and Report

As part of the IE phase, data scientists transform exploratory research into validated inference code. This includes feature engineering, data exploration, model training and tuning, and refining the algorithmic process until it meets the established acceptance criteria and responsibility metrics. The key output of this phase is a report, which serves as the deliverable for review in the assessment and go or no-go decision.

This report functions as a consensus-building document: it (1) demonstrates how the model performs against the minimum functional standards defined in the acceptance criteria; (2) presents the results of fairness and transparency evaluations under the responsibility metrics, including subgroup performance, parity checks, error distributions, and interpretability narratives, such as SPAMM; (3) assesses overall risks and costs with implications for both users and the organization; (4) establishes a baseline of monitoring metrics and visualizations for long-term oversight; and (5) ultimately facilitates an informed go or no-go decision among stakeholders.

Stakeholder Review Go or No-Go

After completing the above, the decision-making stakeholders should come together as part of an existing review or showcase discussion to review the report, analyze the results, and come to a consensus on whether or not the work is “good enough” to proceed to deployment. If the group determines that further refinement is needed, then the team updates the relevant metrics, documents the discussion, and begins another iteration of the assessment process within the same Agile cycle. Importantly, this review is not a new governance requirement being added to Agile. Most teams already make informal go or no-go decisions in sprint reviews, demos, or ad hoc leadership discussions. SAFE-AI simply formalizes and documents these joint decisions, so the rationale is transparent and traceable. It is important to include organizational leadership with the authority to make the executive decision in this regard, including people with ownership of balancing the risks of the algorithm against the needs of the organization.

AI developers and AI deployers should maintain a governance framework that is backed by sufficient executive oversight. In addition to developing and approving the substance of the governance framework’s policies, senior management should play an active role in overseeing the company’s AI product development lifecycle. For high-risk systems that may negatively impact people in consequential ways, company leadership should be accountable for making “go or no-go” decisions [25]. SAFE-AI does not add a separate approval gate, but rather aligns the accountability expected in high-risk AI with the existing cadence of Agile decision points, making explicit what is often handled implicitly.

Integration Planning

Now that the inferential process has been completed and all stakeholders are aligned on the algorithm, the data scientist can work with the software engineers to lock down the deployment constraints. In some cases, this can be done before the “go or no-go” decision if inputs and outputs will not change significantly. It is important to have these completed before proceeding, although in practice, many teams already discuss these topics informally earlier in the development cycle. SAFE-AI does not introduce new tasks here, but simply formalizes conversations that often happen ad hoc, providing clearer documentation and coordination. It is good practice to perform integration planning as early as possible. However, performing this too early can create conflicts when the inputs and outputs change during the research phases. It’s best to perform this planning phase when there is good certainty that the inputs and outputs are well defined and not likely to change. By treating these steps as structured, documented alignment points rather than informal hallway conversations, teams avoid misunderstandings and reduce rework without adding meaningful overhead. For example:

1. Input or output contract: The data engineering process will need to know what data to create to feed into the algorithm and how often, while also asking for the algorithm outputs in return. This contract can be version controlled to say, “The system will always give you this data in this format, and your IE will always give certain data in a certain format in return.” This makes updating the algorithm in the future easier, which will be discussed later.
2. Update process: This should be a plan for which version of the production will integrate the first model. It is typical that an algorithm’s code may change more often than the production version and should have a process to deploy updates to the algorithm quickly without waiting for a full product release.

Develop an AI Module

Now it is time to integrate the models as a new feature within the home application’s development environment that is now aligned with a planned production release. At this point, the work follows typical SDLC processes where the algorithms are engineered into the overall product architecture and involves performing all the other typical processes involved with a new feature release, such as marketing, educating stakeholders, etc

Integration

Once the code is ready, developers move it into production, following the standard SDLC. It is essential to recognize that ML models and the broader application often have different

XSL•FO
RenderX

deployment schedules, requiring a well-orchestrated Continuous Integration and Continuous Delivery process as illustrated in Textbox 1.

Textbox 1. Defining distinct task categories that are relevant to the SAFE-AI (Scalable Agile Framework for Execution in AI) process.

Developing product engine (PE): Traditional software development where the product architecture is built. These tasks are deterministic and repeatable, such as constructing databases, application programming interfaces, or user interfaces. Outcomes are binary (eg, schema exists or not; field matches or does not), and successful runs always yield identical results. The PE provides the stable scaffolding into which artificial intelligence (AI) components can later be embedded.

Developing an inference engine (IE): Unlike deterministic product code, IEs rely on models that incorporate randomness, sampling, or continuous learning, meaning the same code can yield different outputs across runs. This makes validation more complex, where predictions must be evaluated across multiple runs, data splits, and subgroups to ensure they fall within acceptable thresholds rather than simple pass or fail testing. As such, running the same code twice can produce different results. IE development spans from exploratory research (hypotheses, tuning cycles, and modeling notebooks) to validated inference code ready for integration.

Organizations often manage two parallel codebases during this phase: (1) research code is maintained by data scientists and (2) validated prediction engine code is used within the product engine. Only when the team confirms that the engine meets uncertainty and risk criteria does it become eligible for integration.

AI Integration (AII): The convergence point where prediction engine code is embedded into the product codebase. AI transforms research into an operational component that accepts input, runs the algorithm, and returns predictions through the product interface.

By the time the code reaches this phase, the model has been fully validated and is no longer being evaluated for feasibility—just functionality. The code is versioned independently and integrated in coordination with product releases to ensure seamless deployment.

To illustrate these differences, Figure 5 demonstrates how these phases operate in practice.

As shown, the IE and PE evolve on separate but parallel development tracks, each with its own branching and deployment cadence. For example, Live IE version 1.1 may remain in production while a newer Dev IE v1.2 is being validated, just as Dev PE v2.2 advances independently of earlier releases. The AII phase serves as the coordination point where validated AI models are embedded into product-ready systems, and inputs and outputs are aligned and locked. This moment (marked by the star icon in Figure 5) represents the culmination of both tracks, where a stable IE is paired with its corresponding product version for release. Once deployed, the system enters a new live state (eg, Live PE v2.2 with Live IE v1.2), ensuring that only approved, tested combinations reach end users. This disciplined integration workflow supports transparency, traceability, and accountability across the AI development lifecycle.

It is important to note that changes to the IE don't always involve modifications to input or output formats. The structure of the data pipeline may remain consistent even as the underlying model evolves through algorithm updates, parameter tuning, or other enhancements. The purpose of the diagram is simply to illustrate that, at some point, the IE and PE must converge during the AI Integration phase. However, since IE updates can occur independently of PE deployments, it's critical that any such changes, regardless of how minimal they may appear, still undergo the SAFE-AI evaluation process. This ensures that even behind-the-scenes model improvements are held to the same standards of validation, risk assessment, and accountability.

Figure 5. Example of product engine and inference engine version control with parallel development cycles. AI: artificial intelligence; PE: product engine.

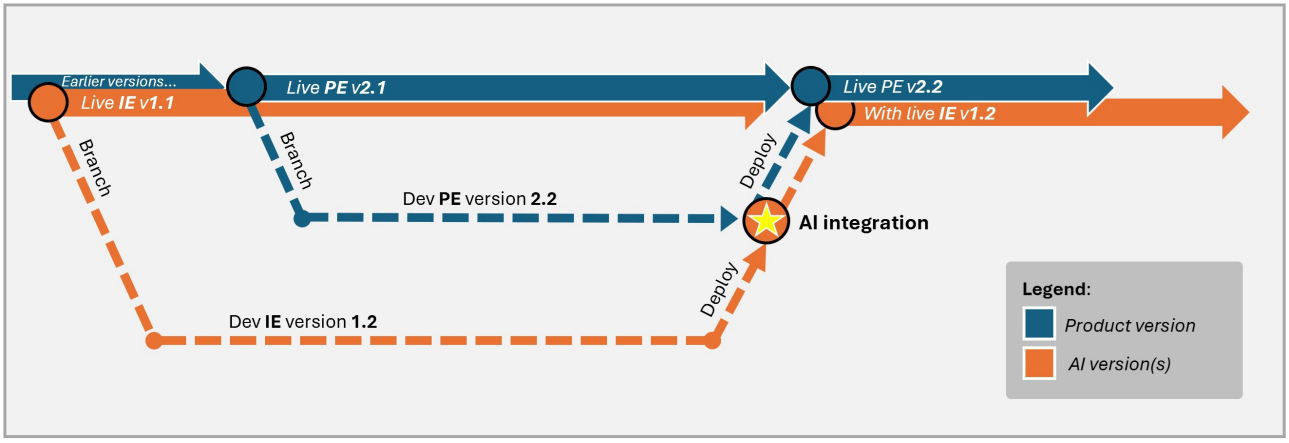

For example, ML models may need to be retrained and updated more frequently than full product releases. This raises important considerations, such as how to introduce new parameters or model updates without disrupting the existing environment.

Ensuring a smooth deployment process requires clear coordination between the data science and software development teams. These discussions should focus on versioning strategies, model refresh cycles, and seamless integration workflows to maintain system stability while enabling continuous improvements (Textbox 2).

XSL•FO
RenderX

Textbox 2. Example illustration of SPAMM (Scenario-based Probability Analogy Mapping Method) in practice

To illustrate SPAMM, the project is assumed to be a system that helps to triage emergency room visits for nonurgent patients only (not to include resuscitation or emergent patients) based on incoming patient data and demographics to try and reduce patient wait times and complications caused by long waits and overcrowding. The goal is to translate your inference engine's preferred error metrics, such as recall, accuracy, root mean square error, and other technical metrics, into the real actions and events that will happen as a result of the predictions.

This is a demonstration example only. It illustrates how the SPAMM method can be applied, but it is not a real clinical analysis.

Step 1 (complete assessment steps): Before building scenarios, all relevant metrics and goals found in the acceptance criteria, responsibility metrics, and risk analysis steps in SAFE-AI (Scalable Agile Framework for Execution in AI) should be defined and documented. That will ensure the organization has a shared baseline of "what matters," how to measure it, and the cost of getting it wrong.

Step 2 (translate acceptance criteria into scenarios): For a relevant metric, these should be associated with the simulated outcomes from the inference engine that is translated into how those predictions relate to the real-world actions and consequences of the model.

Overall Scenarios

- The Emergency Department receives about 152 "nonurgent" patients a day, with 4 out of 5 of those patients experiencing no change in how they would normally be treated.
- If all of the prescribed actions are implemented:
- About 15% (23/152) of these patients would end up waiting longer than they normally would, since the system would reprioritize other cases ahead of them.
- About 5% (8/152) of these patients would be classified as higher urgency than they normally would, meaning they would be escalated for faster treatment than under standard triage.
- Of the 15% (23/152) of patients who would experience longer wait times, only about 1 in 1000 would experience an adverse effect due to the delay (approximately 9 cases per year).
- For the remaining ~23 patients per day who waited longer but did not experience harm, their delayed care effectively freed capacity for more urgent patients. This reallocation reduced other ER-related adverse events by an estimated 5% (around 25 cases per year).
- **Demographic Scenarios** • Age: For every 100 older patients, about 8 are predicted incorrectly. These patients have a 50% higher chance of harm and accounted for nearly all of the adverse events caused by longer wait times. • Race: For every 100 Black patients, about 6 are predicted incorrectly, with about a 5% higher chance of harm. • Gender: There are no meaningful differences in error rates by gender.

Conclusion

- Most patients (121/152, 80%) are triaged the same way they would be under the current system.
- Roughly 8 patients per day are upgraded, meaning they receive care faster than before.
- About 24 patients per day are downgraded, which creates some additional risk, but only around 1 in 1000 of these cases leads to actual harm (about 9 cases per year).
- By comparison, the reallocation of resources reduces other ER-related harms by an estimated 25 cases per year, suggesting an overall safety benefit.
- The most significant concern is that the small number of adverse events disproportionately affects older patients, and Black patients show slightly higher error rates.
- These fairness issues, together with the rare but high-impact missed urgent cases, represent the biggest risks to the hospital's reputation and liability. Smaller risks, such as false alarms or mild misclassifications, carry lower per-case costs but can still influence efficiency.

Overall Demonstration Conclusion: This example illustrates how a triage support system could reduce wait times and improve patient flow while also exposing the disproportionate costs of rare but severe errors and subgroup disparities. From a business perspective, the primary risks are tied to patient harm, liability exposure, and reputational impact. The exercise is not a real analysis but a demonstration of how SPAMM can be applied to create scenario-based narratives that connect acceptance criteria, responsibility metrics, and risk analysis into a single story for decision-makers.

SAFE-AI: Part IV- Monitoring and Evaluation

System monitoring and evaluation at regular intervals can help ensure the prevention of potential harm creation or perpetuation. While initial model results may meet responsibility metrics during development, changes in data patterns or underlying system characteristics could cause deviations in production performance.

According to Data Science Alliance [48] and Guan et al [49], 5 important types of concerns exist when conducting monitoring:

1. Concept drift: Also known as model drift, this occurs when the patterns or relationships estimated by the model shift from their state at deployment. These shifts can manifest instantaneously, gradually, cyclically, or temporarily, potentially compromising the model's effectiveness. This can occur abruptly (eg, new treatments altering outcome rates) or gradually ("AI aging"), requiring continuous recalibration [49].
2. Data drift: This phenomenon emerges when the distribution of production data significantly differs from the training data. While the negative impact of data drift on model

XSL•FO
RenderX

performance typically develops slowly, its effects compound over time, leading to increasingly unreliable results [48].

3. Data integrity: Production environments may introduce unexpected variables, incompatible data types, or inconsistent measurements that differ from the model's training environment. These integrity issues can severely impact the model's accuracy and reliability, requiring immediate attention and correction [48].
4. Bias drift: Even a model developed with careful consideration for fairness and bias mitigation can become unfair in production due to bias drift. As societies evolve, existing biases may diminish while new ones emerge, requiring continuous monitoring and adjustment to maintain ethical performance [48].
5. Calibration drift: Highlighted by Guan et al [49] as a distinct concern in health care: a model may maintain discrimination but lose calibration, producing systematically over- or underestimated risk scores that undermine trust and safety [49].

The categories of drift outlined by Data Science Alliance [48] can be directly linked to statistical methods described by Guan et al [49], which provide concrete means of distinguishing meaningful changes from random variation. By pairing each drift type with corresponding identification strategies, it becomes possible to operationalize conceptual risks into measurable safeguards. Table 3 shows how SAFE-AI connects Data Science Alliance's [48] framework of monitoring concerns with Guan et al [49] recommended statistical approaches.

Table . Statistical tests and drift identification strategies (adapted from Guan et al [49]).

Drift type and formula or test	When to use
Concept drift	
Cumulative Sum and Shewhart or SPC[a] control charts	• For monitoring performance metrics, such as AUC[b] or calibration slope, over time to detect gradual "AI[c] aging" or abrupt performance changes that are effective for real-time detection of sustained deviations.
Data drift	
Maximum mean discrepancy	• For high-dimensional feature spaces (eg, latent embeddings from imaging or EHR[d] data) • It detects subtle covariate shifts.
Wasserstein Distance (Earth Mover's Distance)	• For comparing distributions of latent features, like in medical imaging or pathology data • It is sensitive to spatial domain shifts
Kolmogorov–Smirnov test	• For univariate continuous variables (eg, lab values and vital signs) • It outputs a test statistic and a *P* value to assess whether the distributions differ significantly.
Jensen–Shannon divergence	• For comparing feature or prediction distributions across temporal windows, be robust to temporal shifts.
Bias drift	
Chi-square test	• For categorical variables (eg, sex, race, and hospital unit) • Applied in metadata to detect subgroup distribution shifts
Fisher exact test	• For small sample sizes where chi-square assumptions fail (eg, minority subgroup outcome analysis).
Calibration drift	
Calibration slope or intercept and dynamic calibration curves	• For tracking whether predicted probabilities remain aligned with observed outcome frequencies.
Data integrity	
Rule-based validation and schema checks	• For detecting anomalies, such as new, unexpected, or missing variable encodings and inconsistent measurement units

[a]SPC: statistical process control.
[b]AUC: area under the curve.
[c]AI: artificial intelligence.
[d]EHR: electronic health record.

XSL•FO
RenderX

To distinguish these forms of drift from random variation, Guan et al [49] emphasize the importance of embedding formal statistical testing within monitoring pipelines. Data drift can be identified using statistical distance measures, such as Maximum mean discrepancy or Wasserstein distance for high-dimensional features, Kolmogorov-Smirnov tests for continuous variables, and Jensen-Shannon divergence for monitoring temporal distributional shifts. Subgroup outcome disparities should be routinely assessed with categorical association tests, such as chi-square or Fisher exact test, ensuring fairness is maintained across demographic groups. Calibration drift requires continuous evaluation of calibration slope and intercept, or dynamic calibration curves, to confirm that predicted probabilities remain aligned with observed outcomes. For concept drift and temporal performance degradation, including the gradual loss of discrimination, sequential statistical process control methods, such as cumulative summation and Shewhart control charts, are effective for detecting "AI aging" in deployed systems (Table 3). In this way, the SAFE-AI framework operationalizes Guan et al [49] recommendations by mapping conceptual drift categories to concrete identification strategies relevant for health care contexts.

Operationalizing these methods requires automated pipelines that compute drift, calibration, and fairness metrics at scheduled intervals, with alerts triggered when thresholds are exceeded. However, detection alone is insufficient [49,50]. The need for multidisciplinary review is stressed, where data scientists, clinicians, and ethics experts interpret signals and determine whether retraining, recalibration, or workflow changes are warranted. Equally important is the structured documentation of drift events, calibration measures, and corrective actions, which provides transparency, supports regulatory alignment, and builds trust. Embedding such statistical guardrails within governance processes reduces false alarms and enables timely interventions before performance degradation compromises patient care.

Below are the recommended steps to implement an effective monitoring structure:

1. Automate monitoring and alerting: Deploy automated pipelines that calculate drift and fairness metrics at scheduled intervals. Integrate with alert systems (eg, dashboards, email, and Slack notifications) to flag deviations as soon as they occur. For example, trigger an alert if PSI >0.25 or subgroup accuracy difference >10%.
2. Scheduled reviews: Implement periodic review sessions (eg, monthly or quarterly) involving stakeholders (data scientists, clinicians, and ethics experts) to analyze monitoring reports and determine if intervention, such as model retraining or recalibration, is needed.
3. Document changes: Maintain a structured log (eg, an SQL table or Pandas DataFrame) capturing timestamps, versions, drift statistics, fairness scores, and actions taken.

Table 4 illustrates how SAFE-AI's automated monitoring and documentation process operates in practice. The table represents a simplified monthly log of model performance and fairness metrics tracked through an automated pipeline. In this example, Model 1.2.1 initially performs within acceptable bounds, showing low data drift (PSI <0.10) and stable fairness scores.

Table . Example of monitoring statistics table.

Date	Model_ID	Data_Drift (PSI[a])	Accuracy	F_1-score	Fairness_Parity	Bias_Flag	Notes
January 3, 2025	1.2.1	0.07	0.86	0.91	0.94	Pass	Stable
January 4, 2025	1.2.1	0.08	0.81	0.92	0.93	Pass	Stable
January 5, 2025	1.2.1	0.11	0.79	0.89	0.90	Warning	Stable
January 6, 2025	1.2.1	0.15	0.74	0.85	0.88	Alert	Notified developer
January 7, 2025	1.2.1	0.19	0.68	0.80	0.84	Fail	Failed
January 8, 2025	1.3.0	0.05	0.88	0.92	0.95	Pass	Updated

[a]PSI (population stability index):

By May 2025, drift begins to rise and accuracy drops slightly, triggering a warning flag. In June, the system automatically issues an alert notification to developers as drift and subgroup disparity increase. By July, both performance and fairness fall below defined SAFE-AI thresholds, producing a fail status that prompts model retraining.

A month later, version 1.3.0 was deployed. The updated model restores accuracy, stability, and fairness parity to acceptable levels, confirming successful remediation.

This example demonstrates how SAFE-AI's automated monitoring, alerting, and documentation pipeline detects early signs of degradation, escalates issues for review, and records corrective actions, creating a transparent, auditable trail of model health over time.

1. User feedback integration: Establish channels for end-user (clinician and patient) feedback regarding both model performance and explainability. Use structured surveys and focus groups to gather qualitative data that can inform further improvements.
2. Iterative improvement: Based on audit findings and feedback, implement retraining or fine-tuning cycles that reincorporate recent data. Validate updated models against the established responsibility metrics before redeployment.

XSL•FO
RenderX

Effective system monitoring and evaluation represent crucial components of responsible AI deployment. Potential issues can be identified and addressed early on with the right tools in place. This proactive approach, combined with structured feedback mechanisms and continuous improvement processes, helps ensure that AI systems remain both technically sound and ethically aligned with their intended purposes.

Much of the monitoring and evaluation effort falls within the "0 - 6 hours" of additional sprint time associated with SAFE-AI integration (Figure 3). Importantly, this represents the structured version of work that most teams already perform informally, such as checking whether performance has shifted, manually reviewing error cases, or responding to unexpected model behavior in production, rather than a new category of tasks introduced by SAFE-AI. This framework formalizes existing monitoring responsibilities, provides statistical methods to replace ad hoc practices, and ensures that time spent diagnosing drift, fairness issues, or degraded calibration is tracked as intentional sprint work with owners and acceptance criteria. In this way, the monitoring step reflects not additional governance layers, but a clearer, more accountable articulation of work that must occur to sustain responsible AI performance over time.

Discussion

Balancing Ethical Oversight With Business Objectives

The SAFE-AI framework distinguishes itself by aligning ethical rigor with business priorities and development speed. Whereas existing frameworks, such as ECCOLA [51] or academic or regulatory models, are often perceived as prescriptive or cumbersome, SAFE-AI integrates oversight into Agile workflows without slowing delivery. Ethical review is framed not as external compliance but as a strategic safeguard against regulatory, reputational, and financial risks.

SAFE-AI frames ethical assessment not as an overhead cost, but as a strategic risk mitigation process that helps ensure model integrity, maintain user trust, and prevent potential regulatory, reputational, or financial harm. SAFE-AI is purposefully designed to complement the natural rhythms of modern software and AI product development by integrating ethical review seamlessly into existing Agile and Scrum workflows.

One of the primary distinctions of SAFE-AI is its emphasis on efficiency without compromising ethical standards. Rather than exhaustive checklists or governance layers, SAFE-AI establishes the minimum necessary safeguards to preserve responsibility while remaining agile. This proportional approach makes it especially practical for SMEs with limited resources.

A key aspect of this balance is recognizing that the time SAFE-AI introduces into the development cycle, such as the additional 6 hours of work during a sprint, is not a totally new burden but a structured articulation of tasks that development teams typically perform informally. Most teams already spend time investigating unexpected model behavior, resolving data inconsistencies, responding to stakeholder concerns, or informally reviewing fairness or performance anomalies. SAFE-AI does not create totally new categories of work, but it simply moves these ad hoc activities into the sprint backlog, where they can be estimated, tracked, owned, and completed with the same discipline as technical tasks to ensure they are completed. By making previously invisible labor explicit rather than adding new governance layers, SAFE-AI preserves development speed while ensuring that ethical safeguards are delivered consistently and transparently.

A central innovation is the SPAMM. Unlike SHAP, LIME, or other interpretability tools that emphasize technical explanations, SPAMM translates model behavior into real-world narratives and cost-based analogies. This enables nontechnical stakeholders to grasp the implications of model errors, facilitating informed trade-offs between ethical risk and business priorities.

SAFE-AI also emphasizes iterative oversight. AI systems evolve through retraining, tuning, and updates. SAFE-AI requires previously defined acceptance and responsibility metrics to be revisited after each substantive change. This ensures updates do not reintroduce bias or compromise fairness while avoiding unnecessary rework.

Finally, SAFE-AI frames ethical responsibility as a shared duty across cross-functional teams, embedding accountability within everyday workflows rather than delegating it to specialized ethics committees.

The Critical Role of Risk Interpretability and Metric Development

A core principle of ethical AI development in health care is ensuring that the inherent risks of automated decision-making are both interpretable and measurable. Modeling risk interpretability is not a secondary concern but a foundational necessity for responsible AI deployment. Without clear and agreed-upon mechanisms for understanding how an algorithm's predictions translate into real-world outcomes, health care providers and affected individuals are left to navigate an opaque and potentially harmful system.

The introduction of the SAFE-AI framework addresses this challenge directly by developing probability analogy methods to bridge the gap between complex model behavior and the practical, human-driven ethical decision-making processes within health care settings. Unlike traditional explainability tools, which often focus exclusively on mathematical metrics or technical explanations, while emphasizing narrative-driven, probabilistic scenarios that map model behavior to understandable analogies. This approach acknowledges the reality that nontechnical stakeholders, including clinicians, patients, and regulatory bodies, must ultimately be able to evaluate and engage with the outputs of AI systems in order to trust them.

The significance of this method is especially relevant during the assessment phase of SAFE-AI. Setting responsibility metrics is not merely a technical exercise but an ethical obligation. Without these metrics, it is impossible to contextualize and evaluate how well the algorithm is performing against real-world risks. The development of responsibility metrics, including fairness and transparency metrics, must occur in tandem with risk interpretation strategies to ensure that metrics are not abstract or arbitrary but rather aligned with the actual impact on the user population.

XSL•FO
RenderX

Additionally, one of the most overlooked sources of ethical risk in AI systems is introduced not during the initial development but during subsequent algorithmic updates, such as model retraining or hyperparameter tuning. Even seemingly minor updates can alter the balance of an algorithm in subtle ways that inadvertently introduce ethical imbalances or exacerbate existing disparities. As described in earlier sections, SAFE-AI mandates that any update to an existing inference engine—whether through additional training data, tuning, or rebalancing—should be treated as a new project cycle if there is any risk that new training data can create predictions that fall out of bounds. Although previously established assessment metrics can and should be reused, they must be explicitly re-evaluated to verify that any modifications have not unintentionally undermined fairness, transparency, or safety.

This approach balances the need for iterative improvement with the ethical imperative to avoid harm. Too often, pressure to deploy quickly or incrementally improve system performance results in bypassing critical ethical assessments; however, many algorithms use random samples as part of their retraining process, and a quick iteration can inadvertently create vastly different outcomes. SAFE-AI stresses that, even in fast-paced environments, evaluating updated models against established metrics is essential. The cost of skipping this step is not merely technical debt but ethical liability, with real consequences for patient safety and equity.

Finally, it is important to emphasize that the development of these metrics and risk interpretability mechanisms cannot be left solely to technical teams. The most effective way to safeguard against ethical harm is through multidisciplinary collaboration. Risk interpretation is inherently contextual, requiring input from health care professionals, regulatory experts, ethicists, and product stakeholders who understand the social and operational realities in which the AI system will function. This collaborative approach ensures that metrics are not only technically valid but also ethically relevant and meaningful to the people most impacted by the model's behavior.

By prioritizing transparency, risk interpretability, and ongoing evaluation using the approaches outlined in the "Develop Responsibility Metrics" section, the framework supports the development of AI systems that are not only effective but also fair, accountable, and ethically sound.

Validating the Framework in Real-World Health Care AI Settings

To assess SAFE-AI's real-world relevance and impact, future work should focus on comprehensive field-testing across diverse health care AI applications. The empirical analysis would involve implementing the SAFE-AI framework with SMEs developing various medical AI systems, including diagnostic support tools, patient outcome assessment, and wound trajectory prediction. Using a pre-post intervention design, these studies would evaluate the framework's impact on development practices, fairness outcomes, and organizational dynamics in real-world health care AI deployments.

Within health care settings, the field studies would specifically examine how the SAFE-AI framework performs across various clinical applications, where ethical considerations regarding patient care and outcomes are particularly critical. The research would investigate how effectively SAFE-AI addresses issues, such as training on biased datasets, transparency requirements, preventing provider overreliance on AI systems across different clinical contexts, and how to ensure its seamless integration into existing processes. By studying implementation across diverse medical specialties, it is possible to assess SAFE-AI's adaptability to varying clinical workflows, data types, and ethical priorities.

The proposed implementation process would include a baseline assessment to capture preframework metrics, structured implementation of the responsibility protocols, and postimplementation evaluation at multiple intervals. This systematic approach would allow for the identification of potential adoption challenges, including resource constraints, workflow disruptions, and organizational resistance. Field testing would evaluate various implementation strategies, such as gradual integration, iterative refinement, and leadership engagement models, to determine which approaches most effectively support framework adoption while maintaining rigorous ethical standards.

Building on these future validation efforts, the framework presented here offers a practical approach to embedding ethical considerations throughout health care AI development while maintaining efficient workflows. By transforming abstract principles into concrete processes with measurable outcomes, this framework helps bridge the gap between technical implementation and ethical governance. As health care AI systems become increasingly integrated into clinical decision-making, such structured approaches to responsible development will be essential for building trustworthy systems that genuinely advance patient care without introducing new forms of bias or harm.

Comparison With Existing Frameworks

Table 5 situates SAFE-AI alongside ECCOLA and CARE-AI (Collaborative Assessment for Responsible and Ethical AI Implementation), two of the most prominent ethical AI frameworks. While ECCOLA provides broad governance prompts and CARE-AI emphasizes health care assurance cases, both approaches are resource-intensive and often external to fast-paced product development. In contrast, SAFE-AI was co-designed with practitioners to embed ethical safeguards directly into Agile or Scrum sprints, collaboratively customize responsibility metrics to the project at hand, and lower adoption barriers for SMEs. While the development of SAFE-AI benefited from an expert consortium, this involvement intentionally front-loaded ethical and regulatory complexity into the framework design so that downstream SME users can apply its artifacts without requiring comparable ethics or academic partnerships.

XSL•FO
RenderX

Table . Example of SPAMM[a] compared to technical metrics.

Criteria	ECCOLA	CARE-AI[b]	SAFE-AI[c]
Origin or context	Academic or consortium-led, broad AI[d] governance guidelines	Health care–oriented ethical assurance	SME[e]-practical, co-designed with practitioners
Workflow integration	Provides ethical prompts or checklists, but external to development sprints	Process guidance tied to assurance cases; heavy documentation	Fully embedded into Agile or Scrum sprints with sprint-based checkpoints
Fairness coverage	High-level bias prompts, no quantitative thresholds	Addresses equity in health care safety cases	Defines fairness metrics with subgroup thresholds and SPAMM narratives
Transparency coverage	Suggests explainability but lacks concrete methods	Assurance-case documentation for regulators	Combines methods like SHAP[f] or LIME[g] with SPAMM for all domains (scenario-based narrative analogies)
Monitoring	Minimal; assumes postdeployment audits	Focus on compliance reporting and auditability	Continuous monitoring with statistical drift detection triggers
Adoption burden	High (lengthy prompts or checklists and steep learning curve)	High (documentation intensive; requires assurance specialists)	Low to moderate (checklists and metrics integrated into existing development tools)
Scalability to SMEs	Limited; designed for larger institutions	Limited; high compliance load	Explicitly designed for SMEs; pragmatic, “minimum necessary” safeguards

[a]SPAMM: scenario-based probability analogy mapping method.

[b]CARE-AI: (Collaborative Assessment for Responsible and Ethical AI Implementation).

[c]SAFE-AI: Scalable Agile Framework for Execution in AI.

[d]AI: artificial intelligence.

[e]SME: small and medium-sized enterprise.

[f]SHAP: Shapley Additive Explanations.

[g]LIME: Local Interpretable Model-Agnostic Explanations.

A key differentiator is SAFE-AI’s explicit use of fairness metrics and SPAMM narratives to operationalize equity and transparency across domains, whereas ECCOLA and CARE-AI stop at higher-level guidance. Similarly, SAFE-AI integrates continuous monitoring with statistical drift detection, allowing ongoing oversight without the compliance burden of full assurance-case documentation.

Taken together, these differences highlight SAFE-AI’s pragmatic contribution, as it delivers a “minimum necessary” governance structure that is rigorous enough to support fairness, transparency, and monitoring, yet lean enough for SMEs to adopt without specialized compliance staff.

Conclusion

The SAFE-AI ethical development framework proposed in this paper offers a practical, business-aligned approach for integrating responsible AI practices into the everyday workflows of product and data science teams. By mirroring established Agile and Scrum methodologies, SAFE-AI avoids treating ethics as a parallel or external process, instead embedding ethical evaluation directly into the structure of software and AI development. This alignment ensures that ethical considerations are addressed without compromising delivery timelines or development agility—a critical factor for SMEs operating under resource constraints.

Central to SAFE-AI is the principle that AI development introduces uncertainty and risk that cannot be adequately managed through traditional software engineering approaches alone. Accordingly, SAFE-AI emphasizes the early and ongoing use of measurable ethical metrics—acceptance criteria, fairness metrics, and transparency metrics—to evaluate model behavior. These metrics serve as practical tools for managing the probabilistic nature of AI systems, enabling teams to assess whether their models meet not only performance expectations but also ethical and social obligations.

Importantly, the SAFE-AI framework extends beyond initial deployment by incorporating a monitoring and evaluation phase that supports continuous ethical oversight. It recognizes that AI systems are dynamic and must be re-evaluated whenever retraining, tuning, or environmental shifts occur. This ensures that ethical integrity is not treated as a static checkpoint but rather as an evolving requirement that must be actively maintained throughout the product lifecycle.

A key innovation introduced by SAFE-AI is the scenario-based probability analogy mapping process as an interpretability technique designed to enhance trust among nontechnical stakeholders. Translating probabilistic model behavior into familiar, narrative-based analogies with associated risks and costs helps explain how and why AI systems reach certain predictions—especially in high-stakes health care settings where

XSL•FO
RenderX

transparency is critical. Unlike traditional explainability tools that focus on model internals, SAFE-AI centers on real-world interpretability, offering an accessible way to communicate model behavior, associated risks, and uncertainty in decision-making. As such, it not only supports transparency but also serves as a bridge between ethical modeling and user-centered design.

Taken together, the contributions of SAFE-AI establish a scalable pathway for responsible AI adoption in health care and beyond where modeling excellence, software best practices, and ethical integrity are not at odds but rather deeply interwoven into a cohesive and actionable development process.

Acknowledgments

The authors gratefully acknowledge Professor David Danks, University of California, San Diego, for insightful comments that improved the clarity and rigor of the manuscript. The author prepared portions of this manuscript with the assistance of ChatGPT (GPT-5, OpenAI, 2025) to support phrasing refinement. The author reviewed, verified, and revised all AI-assisted text for accuracy and originality, ensuring that all factual content, references, and interpretations are the author's own. No generative AI was used for data analysis, study design, or interpretation of results.

Funding

This study was partially supported by the National Institute on Aging's Artificial Intelligence and Technology Collaboratories for Aging Research Grant P30AG073104 (WWP), the US Army xTech AI Grand Challenge (WWP), the Huntsman Mental Health Foundation (WWP); as well as Mountain Biometrics, Inc. (WWP), Nemsee, LLC (IN), and the Data Science Alliance (AM, LJ, RL, and PL). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the US Army or the US Government.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during this study.

Authors' Contributions

Conceptualization: IN, AM, LJ, RL, WWP
Investigation: IN, AM, LJ, RL, WWP
Methodology: IN, AM, LJ, RL, WWP
Project administration: IN, WWP, PL
Supervision: IN, WWP, PL
Validation: IN, AM, RL, WWP
Visualization: IN, LJ, WWP
Writing – original draft: IN
Writing – review & editing: IN, AM, WWP

Conflicts of Interest

WWP is a shareholder in Mountain Biometrics Inc., and IN is a shareholder in Nemsee, LLC. All other authors declare no competing interests.

Checklist 1
SAFE-AI (Scalable Agile Framework for Execution in AI) phase summary and implementation checklist.
[XLSX File, 10 KB - jmir_v28i1e80028_app1.xlsx]

References

1. What is Scrum? Scrum.org. URL: https://www.scrum.org/resources/what-scrum-module [accessed 2025-12-02]
2. The 2020 Scrum Guide. Scrum Guides. URL: https://scrumguides.org/scrum-guide.html [accessed 2025-12-02]
3. Ali SJ, Christin A, Smart A, Katila R. Walking the walk of AI ethics: organizational challenges and the individualization of risk among ethics entrepreneurs. Presented at: FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency; Jun 12-15, 2023; Chicago IL USA p. 217-226 URL: https://dl.acm.org/doi/proceedings/10.1145/3593013 [doi: 10.1145/3593013.3593990]
4. Crockett K, Colyer E, Gerber L, Latham A. Building trustworthy AI solutions: a case for practical solutions for small businesses. IEEE Trans Artif Intell 2023 Aug;4(4):778-791. [doi: 10.1109/TAI.2021.3137091]
5. Hajianhosseinabadi H, Lindau N. Startup interaction with AI - A multiple case study on startups' ethical navigation when using Generative AI tools.: University of Gothenburg; 2024. URL: https://gupea.ub.gu.se/handle/2077/82378 [accessed 2025-02-26]

XSL•FO
RenderX

6. Kostick-Quenet KM, Cohen IG, Gerke S, et al. Mitigating racial bias in machine learning. J Law Med Ethics 2022;50(1):92-100. [doi: 10.1017/jme.2022.13] [Medline: 35243993]
7. Ning Y, Liu X, Collins GS, et al. An ethics assessment tool for artificial intelligence implementation in healthcare: CARE-AI. Nat Med 2024 Nov;30(11):3038-3039. [doi: 10.1038/s41591-024-03310-1]
8. Berman G, Goyal N, Madaio M. A scoping study of evaluation practices for responsible AI tools: steps towards effectiveness evaluations. Presented at: CHI '24: CHI Conference on Human Factors in Computing Systems; May 11-6, 2024; Honolulu, HI p. 1-24 URL: https://dl.acm.org/doi/proceedings/10.1145/3613904 [doi: 10.1145/3613904.3642398]
9. Winecoff AA, Watkins EA. Artificial concepts of artificial intelligence: institutional compliance and resistance in AI startups. AIES '22: AAAI/ACM Conference on AI, Ethics, and Society :788-799. [doi: 10.1145/3514094.3534138]
10. Fjeld J, Achten N, Hilligoss H, Nagy A, Srikumar M. Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI. SSRN Journal 2020. [doi: 10.2139/ssrn.3518482]
11. Clinical decision support software frequently asked questions (FAQs). US Food and Drug Administration. 2025. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/clinical-decision-support-software-frequently-asked-questions-faqs [accessed 2025-01-23]
12. Clinical decision support software guidance for industry and food and drug administration staff. US Food and Drug Administration. 2022. URL: https://www.fda.gov/media/109618/download [accessed 2025-01-23]
13. Rep. arrington JC [R-T-19]. H.R.1 - 119th congress (2025-2026): one big beautiful bill act. Congress.gov. 2025. URL: https://www.congress.gov/bill/119th-congress/house-bill/1/text [accessed 2025-06-30]
14. Domin H, Rossi F, Goehring B, Ganapini M, Berente N, Bevilacqua M. On the ROI of AI ethics and governance investments: from loss aversion to value generation. California Management Review. 2024 Jul 29. URL: https://cmr.berkeley.edu/2024/07/on-the-roi-of-ai-ethics-and-governance-investments-from-loss-aversion-to-value-generation/ [accessed 2025-09-19]
15. Nondiscrimination in health programs and activities. Centers for Medicare & Medicaid Services. 2024. URL: https://www.federalregister.gov/documents/2024/05/06/2024-08711/nondiscrimination-in-health-programs-and-activities [accessed 2025-01-23]
16. Bevilacqua M, Berente N, Domin H, Goehring B, Rossi F. The return on investment in AI ethics: a holistic framework. arXiv. Preprint posted online on Nov 7, 2023. [doi: 10.48550/arXiv.2309.13057]
17. Good machine learning practice for medical device development: guiding principles. US Food and Drug Administration.: FDA; 2023. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles [accessed 2025-01-23]
18. Artificial intelligence and machine learning in software as a medical device. US Food and Drug Administration. 2025. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device [accessed 2025-01-23]
19. Informed consent faqs. US Department of Health and Human Services. URL: https://www.hhs.gov/ohrp/regulations-and-policy/guidance/faq/informed-consent/index.html [accessed 2025-02-05]
20. Pfeiffer RS, Forsberg RP. Ethics on the Job: Cases and Strategies, 4th edition: Wadsworth; 2013.
21. Pfeiffer RS. Teaching ethical decision-making: the case study method and the RESOLVEDD strategy. Teach Philos 1992 May 1;15(2):175-184. [doi: 10.5840/teachphil199215221]
22. Vakkuri V, Kemell KK, Jantunen M, Halme E, Abrahamsson P. ECCOLA — A method for implementing ethically aligned AI systems. J Syst Softw 2021 Dec;182:111067. [doi: 10.1016/j.jss.2021.111067]
23. Acceptance criteria: examples and best practices. Atlassian. URL: https://www.atlassian.com/work-management/project-management/acceptance-criteria [accessed 2025-01-23]
24. Everything you need to know plus examples. Scrum Alliance. URL: https://resources.scrumalliance.org/Article/need-know-acceptance-criteria [accessed 2025-01-23]
25. Confronting bias: BSA's framework to build trust in AI. : BSA; 2021 URL: https://www.bsa.org/files/reports/2021bsaaibias.pdf
26. Ratwani RM, Sutton K, Galarraga JE. Addressing AI algorithmic bias in health care. JAMA 2024 Oct 1;332(13):1051-1052. [doi: 10.1001/jama.2024.13486] [Medline: 39230911]
27. Assessment. Fairlearn. URL: https://fairlearn.org/v0.12/user_guide/assessment/index.html [accessed 2025-01-23]
28. Babic B, Gerke S, Evgeniou T, Cohen IG. Beware explanations from AI in health care. Science 2021 Jul 16;373(6552):284-286. [doi: 10.1126/science.abg1834]
29. Abujaber AA, Nashwan AJ. Ethical framework for artificial intelligence in healthcare research: a path to integrity. World J Methodol 2024 Sep 20;14(3):94071. [doi: 10.5662/wjm.v14.i3.94071] [Medline: 39310239]
30. Weerts H, Dudík M, Edgar R, Jalali A, Lutz R, Madaio M. Fairlearn: assessing and improving fairness of AI systems. arXiv. Preprint posted online on Mar 29, 2023. [doi: 10.48550/arXiv.2303.16626]
31. Molnar C. Local Model-Agnostic Methods. In: Interpret Mach Learn 2018. URL: https://christophm.github.io/interpretable-ml-book/agnostic.html [accessed 2025-02-05]
32. Watson D. Conceptual challenges for interpretable machine learning. SSRN Journal 2020 Jan:65. [doi: 10.2139/ssrn.3668444]

XSL•FO
RenderX

33. Kim S. Explainable AI (XAI) methods part 3 - accumulated local effects (ALE). Towards Data Science. 2022. URL: https://towardsdatascience.com/explainable-ai-xai-methods-part-3-accumulated-local-effects-ale-cf6ba3387fde/ [accessed 2026-01-25]
34. Mutegeki H, Nahabwe A, Nakatumba-Nabende J, Marvin G. Interpretable machine learning-based triage for decision support in emergency care. 2023 Presented at: 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI); Apr 11-13, 2023; Tirunelveli, India p. 983-990. [doi: 10.1109/ICOEI56765.2023.10125918]
35. The AI/ML wars: “explain” or test black box models? error stat philos. Error Statistics. 2022. URL: https://errorstatistics.com/2022/03/23/the-ai-ml-wars-explain-or-test-black-box-models/ [accessed 2025-02-05]
36. Ribeiro MT, Singh S, Guestrin C. Model-agnostic interpretability of machine learning. arXiv. Preprint posted online on Jun 16, 2016. [doi: 10.48550/arXiv.1606.05386]
37. Aamodt A, Plaza E. Case-based reasoning: foundational issues, methodological variations, and system approaches. Eur J Artif Intell 1994;7(1):39-59. [doi: 10.3233/AIC-1994-7104]
38. Cai CJ, Jongejan J, Holbrook J. The effects of example-based explanations in a machine learning interface. 2019 Mar 17 Presented at: IUI ’19; Mar 17-20, 2019; Marina del Ray California p. 258-262 URL: https://dl.acm.org/doi/proceedings/10.1145/3301275 [doi: 10.1145/3301275.3302289]
39. IONOS Editorial Team. What is case-based reasoning? IONOS. 2025. URL: https://www.ionos.com/digitalguide/websites/web-development/case-based-reasoning/ [accessed 2025-02-05]
40. Costa M. Social return on investment. In: Idowu SO, Capaldi N, Zu L, Gupta AD, editors. Encyclopedia of Corporate Social Responsibility: Springer; 2013:2238-2244. [doi: 10.1007/978-3-642-28036-8_734]
41. Millar R, Hall K. Social Return on Investment (SROI) and Performance Measurement: the opportunities and barriers for social enterprises in health and social care. Public Manag Rev 2013 Sep;15(6):923-941. [doi: 10.1080/14719037.2012.698857]
42. NIST AI RMF Playbook. NIST AIRC. URL: https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook [accessed 2025-01-23]
43. Blueprint for trustworthy AI implementation guidance and assurance for healthcare version 10. : Coalition for Health AI; 2023 URL: https://assets.ctfassets.net/7s4afyr9pmov/4AXIWGIlcrjWDaW2ueTaRS/f98e5cb2528187635895cce6ba5ec309/Blueprint_for_Trustworthy_AI.pdf
44. Baquero JA, Burkhardt R, Govindarajan A, Wallace T. Derisking AI: risk management in AI development. McKinsey. 2020. URL: https://www.mckinsey.com/capabilities/quantumblack/our-insights/derisking-ai-by-design-how-to-build-risk-management-into-ai-development?utm_source=chatgpt.com [accessed 2025-02-14]
45. Mello MM, Guha N. Understanding liability risk from healthcare AI. : Stanford University: Human-Centered Artificial Intelligence; 2024 URL: https://hai-production.s3.amazonaws.com/files/2024-02/Liability-Risk-Healthcare-AI.pdf
46. Total product life cycle for medical devices. US Food and Drug Administration. 2023. URL: https://www.fda.gov/about-fda/cdrh-transparency/total-product-life-cycle-medical-devices [accessed 2025-02-14]
47. Reddy GP, Pavan Kumar YV, Prakash KP. Hallucinations in large language models (LLMs). 2024 Apr 25 Presented at: 2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream); Apr 25-25, 2024; Vilnius, Lithuania p. 1-6. [doi: 10.1109/eStream61684.2024.10542617]
48. The framework for responsible data science practices. : Data Science Alliance; 2024 URL: https://drive.google.com/file/d/17cdW49APhyv7qEy0EyQV6vv7qkw_EcIQ/view?usp=embed_facebook
49. Guan H, Bates D, Zhou L. Keeping medical AI healthy and trustworthy: a review of detection and correction methods for system degradation. arXiv. Preprint posted online on Dec 2, 2025. [doi: 10.48550/arXiv.2506.17442]
50. Saenz AD, McCoy T, Mantha AB, et al. Establishing responsible use of AI guidelines: a comprehensive case study for healthcare institutions. npj Digit Med 2024 Nov 30;7(1):1-6. [doi: 10.1038/s41746-024-01300-8]
51. Vakkuri V, Kemell KK, Abrahamsson P. Ethically aligned design: an empirical evaluation of the RESOLVEDD-strategy in software and systems development context. 2019 Presented at: 2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA); Aug 28-30, 2019; Kallithea-Chalkidiki, Greece p. 46-50. [doi: 10.1109/SEAA.2019.00015]

Abbreviations

AI: artificial intelligence
ALE: accumulated local effects
BSA: Business Software Alliance
CCPA: California Consumer Privacy Act
DSA: Data Science Alliance
EU: European Union
FDA: US Food and Drug Administration
GDPR: General Data Protection Regulation
GenAI: generative artificial intelligence
HHS OCR: Department of Health and Human Services - Office for Civil Rights
HIPAA: Health Insurance Portability and Accountability Act

XSL•FO
RenderX

IE: inference engine
IRB: Institutional Review Board
LIME: Local Interpretable Model-Agnostic Explanations
ML: machine learning
MTN: Medical Timeseries Networks
NIST: National Institute of Standards and Technology
PDP: partial dependence plots
PE: product engine
PSI: Population Stability Index
SAFE-AI: Scalable Agile Framework for Execution in AI
SaMD: software as a medical device
SDLC: software development life cycle
SHAP: Shapley Additive Explanations
SME: small and medium-sized enterprise
SPAMM: Scenario-based Probability Analogy Mapping Method
TPLC: Total Product Life Cycle

Edited by A Coristine; submitted 03.Jul.2025; peer-reviewed by A Munoz-Zavala, VV Sangaraju, X Liang; accepted 18.Dec.2025; published 25.Feb.2026.

Please cite as:
Nemteanu I, Mancebo Jr A, Joe L, Lopez R, Lopez P, Pettine WW
Scalable Agile Framework for Execution in AI for Medical AI Ethics Policy Design in Small- and Medium-Sized Enterprises
J Med Internet Res 2026;28:e80028
URL: https://www.jmir.org/2026/1/e80028
doi:10.2196/80028

© Ion Nemteanu, Adir Mancebo Jr, Leslie Joe, Ryan Lopez, Patricia Lopez, Warren Woodrich Pettine. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

From Agents to Governance: Essential AI Skills for Clinicians in the Large Language Model Era

Weiping Cao[1,2], MMed; Qing Zhang[1*], MD; Jialin Liu[3,4*], MD; Siru Liu[5], PhD

[1]Department of Cardiology, West China Hospital, Sichuan University, Chengdu, Sichuan, China
[2]Department of Cardiology, The People's Hospital of Leshan, Leshan, Sichuan, China
[3]Information Center, West China Hospital, Sichuan University, Chengdu, China
[4]Department of Medical Informatics, West China Medical School, Sichuan University, Chengdu, China
[5]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States
[*]these authors contributed equally

Corresponding Author:
Jialin Liu, MD
Information Center
West China Hospital
Sichuan University
No.37 Guo Xue Xiang Street
Chengdu, 610041
China
Phone: 86 28 85422416
Fax: 86 28 85422607
Email: dljl8@163.com

Abstract

Large language models are rapidly transitioning from pilot schemes to routine clinical practice. This creates an urgent need for clinicians to develop the necessary skills to strike the right balance between seizing opportunities and taking accountability. We propose a 3-tier competency framework to support clinicians' evolution from cautious users to responsible stewards of artificial intelligence (AI). Tier 1 (foundational skills) defines the minimum competencies for safe use, including prompt engineering, human–AI agent interaction, security and privacy awareness, and the clinician-patient interface (transparency and consent). Tier 2 (intermediate skills) emphasizes evaluative expertise, including bias detection and mitigation, interpretation of explainability outputs, and the effective clinical integration of AI-generated workflows. Tier 3 (advanced skills) establishes leadership capabilities, mandating competencies in ethical governance (delineating accountability and liability boundaries), regulatory strategy, and model life cycle management—specifically, the ability to govern algorithmic adaptation and change protocols. Integrating this framework into continuing medical education programs and role-specific job descriptions could enhance clinicians' ability to use AI safely and responsibly. This could standardize deployment and support safer clinical practice, with the potential to improve patient outcomes.

(J Med Internet Res 2026;28:e86550) doi:10.2196/86550

KEYWORDS

large language model; clinician; agent; competency; artificial intelligence; education; continuing medical education

Introduction

The convergence of generative artificial intelligence (AI) and health care is catalyzing a paradigm shift in clinical practice, with significant implications for the future of medicine [1-3]. Large language models (LLMs), exemplified by recent advances, such as GPT-4 and Gemini, demonstrate a transformative capacity to process multimodal data and generate context-aware responses, increasingly positioning them as integral components in frontline clinical decision support [4,5].

Although LLMs have the potential to improve clinical effectiveness, ensuring that their application optimizes patient safety, ethical alignment, and long-term benefits remains a substantial challenge [2,5]. This complexity is compounded by the intersection of regulatory requirements and ethical obligations. Evolving legal frameworks, such as the European Union (EU) AI Act and the US Food and Drug Administration

XSL•FO
RenderX

(FDA) guidance, explicitly mandate human oversight for high-risk AI systems [6,7]. Simultaneously, global ethical standards from the World Health Organization (WHO) and the American Medical Association emphasize the necessity of physician leadership and accountability [2,8,9]. However, a gap remains in translating these high-level mandates into actionable clinical skills. Without the active leadership and input of clinicians, these technologies risk imposing unintended burdens and may fail to achieve their full potential [1].

The imperative for advanced AI governance arises from a fundamental shift from passive information retrieval to autonomous task execution. While conventional LLM paradigms rely on user-initiated prompt response exchanges, clinicians query the model and verify its text outputs, and the system does not autonomously call external tools. By contrast, emerging agentic workflows introduce a perceive-plan-act (and often reflect) loop [10]. In this mode, the system interprets high-level clinical intents (eg, hypertension management); decomposes them into subtasks; and autonomously executes actions via application programming interfaces, such as accessing electronic health record (EHR) data or calculating risk scores [11,12]. This transition reframes supervision; clinicians must move beyond prompt engineering to govern how autonomy is delegated, how actions are constrained, and how escalation pathways are formalized.

To address these regulatory, ethical, and technical demands, we propose a foundational, tiered AI competency framework for clinicians. The framework is structured around progressive tiers: tier 1 (foundational skills), tier 2 (intermediate skills), and tier 3 (advanced skills). We describe the core competencies at each tier, outline the framework's limitations, and propose priority directions for validation to sustain its relevance amid an evolving regulatory landscape.

AI Competency Framework

LLM-enabled care necessitates a transition in the roles of clinicians (physicians, nurses, pharmacists, and allied health professionals)—from interpreting predictive outputs to supervising agentic workflows. Drawing on previous research, a narrative synthesis of evolving digital health competencies, and an analysis of the technical capabilities of LLMs [13-17], we propose a 3-tier, governance-aligned framework that articulates core LLM competencies. As illustrated in Figure 1, the framework progresses from foundational safe use (tier 1) to evaluative proficiency (tier 2) and ultimately to governance and leadership (tier 3).

Figure 1. Essential artificial intelligence (AI) skills for clinicians in the large language model era.

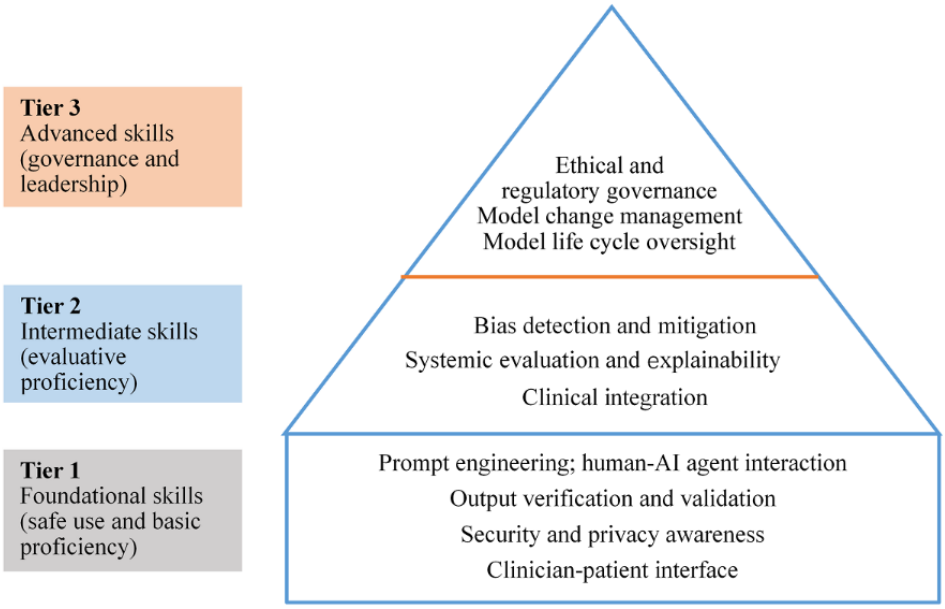

Tier 1: Foundational Skills (Safe Use and Basic Proficiency)

These entry-level competencies prioritize basic interaction with LLMs, enabling clinicians to leverage AI for routine tasks without compromising clinical autonomy. The key elements are described subsequently. First, prompt engineering (task specification for clinician-initiated and hybrid workflows) is used to craft precise, context-aware instructions—with explicit roles, required inputs, constraints, task steps, and output formats (including citation and traceability requirements)—to elicit task-appropriate outputs (eg, structured outlines for differential diagnosis). This competency primarily supports clinician-initiated chat and hybrid workflows, as fully agentic perceive-plan-act execution is typically governed by system-level prompts and policies rather than user-generated prompts. When paired with verification and source grounding, this approach may reduce hallucinations and improve relevance and completeness [18]. Second, human-AI agent interaction (agent supervision) ensures that agents operate within bounded autonomy with explicit roles, goals, and guardrails. Clinicians must maintain awareness of least privilege tool permissions and system constraints (eg, data minimization, time and step limits, and sandboxed execution) [19], with clear termination and escalation criteria. Clinicians also monitor and validate the perceive-plan-act-reflect loop using provenance and citation requirements; protected health information redaction; and audit

XSL•FO
RenderX

logging of prompts, tool calls, and human overrides [20-22]. When confidence or calibration thresholds are not met (eg, coverage targets and abstention or deferral rules), clinicians intervene, interrupt the agent, or revert to manual workflows and document the event for review. Third, output verification and validation involve clinicians critically evaluating individual LLM outputs for accuracy, relevance, and internal consistency. Generated content (eg, summaries, diagnostic considerations, and treatment plans) is cross-referenced against the EHR, structured data, and established clinical evidence to detect hallucinations, omissions, or misstatements. In culturally diverse settings, clinicians must assess outputs for cultural safety and linguistic accuracy. This involves verifying that translated instructions and culturally specific dietary or lifestyle advice are appropriate for the patient's context. Clinicians should also check for Anglocentric bias that could conflict with local norms or the patient's language proficiency. This human-in-the-loop verification is essential for ensuring patient safety in individual clinical encounters [23,24]. The fourth element is security and privacy awareness. To comply with regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation, clinicians must adhere to foundational safeguards centered on data minimization and appropriate tool use. For nonintegrated or open access LLM interfaces, this includes avoiding entry of protected health information and direct identifiers and, when clinically necessary to discuss a case, removing or generalizing nonessential identifiers before input [2,5,25,26]. In contrast, for authorized, integrated enterprise agents operating within a secure EHR environment, manual deidentification is often neither feasible nor necessary; instead, clinicians verify minimum necessary access, confirm the agent is scoped to the correct patient context, and ensure permissions are aligned with the specific clinical task through role-based access control and least privilege settings. Rather than conducting technical audits themselves, clinicians prevent inadvertent privacy breaches by distinguishing approved tools from unapproved ones and escalating permission or access-scope concerns through institutional channels. These baseline competencies are prerequisites for safe AI use in routine clinical workflows [5]. Fifth, clinician-patient interface (transparent communication and shared decision-making) involves incorporating AI-assisted content into the clinical encounter without undermining patient trust or the therapeutic alliance. Clinicians should disclose when AI is used (eg, AI-scribed summaries, patient-portal messages, and patient education materials) to uphold patient autonomy and informed consent [26].

Tier 2: Intermediate Skills (Evaluative Proficiency)

Building on foundational knowledge, these competencies center on critically assessing and integrating LLMs into clinical reasoning workflows while addressing bias and uncertainty in generative AI. First, bias detection and mitigation require clinicians to interpret algorithmic bias audit reports and uncertainty quantification outputs (eg, confidence intervals, prediction intervals, or conformal prediction sets when provided) to assess reliability across patient subgroups. Clinicians initiate and validate remediation actions—such as recommending prompt or workflow adjustments and defining escalation and deferral rules—in coordination with technical teams, ensuring adherence to prespecified fairness metrics and minimum subgroup performance thresholds [27,28]. For example, in tumor grading, clinicians review reported subgroup performance using minimum sample-size thresholds, calibration and coverage, abstention rates, and uncertainty displays (including confidence sets). They assess model rationale and interpret between-group performance gaps. Second, systemic evaluation and explainability involve moving beyond checking individual outputs to assessing the broader reliability, calibration, and failure modes of the AI system. Clinicians should be able to interpret model performance metrics (eg, sensitivity, specificity, error and hallucination rates, and performance in specific subpopulations) and evaluate available explainability outputs (eg, feature importance scores, reason codes, or saliency maps where available) to understand why a model reached a conclusion [29]. This evaluation must include equity audits that assess model performance across distinct subgroups (eg, race, ethnicity, and language) [28]. For instance, a model may demonstrate high overall accuracy but fail disproportionately for languages spoken by minority groups or specific dialects. Clinicians leading the evaluation must identify such disparities and determine if the model is safe for deployment in diverse populations. These skills enable clinicians to judge systemic trustworthiness and identify appropriate clinical use cases and target populations for which the model is calibrated, effective, and equitable [30]. Third, clinical integration requires clinicians to use domain knowledge to refine model outputs (eg, align treatment suggestions with evidence-based guidelines) while monitoring for potential deskilling. Clinicians maintain human-AI collaboration and specify deferral and escalation rules (eg, abstention thresholds and human-review triggers) and document these events for auditability [31,32].

Tier 3: Advanced Skills (Governance and Leadership)

Unlike the foundational skills in tiers 1 and 2, this tier represents a specialized track for clinician-leaders, clinical informaticists, and physician builders assuming governance roles. These competencies focus on the strategic oversight and architectural direction of AI systems. The main competencies involved are described subsequently. First, ethical and regulatory governance involves overseeing the development of institutional policies for LLM use to ensure alignment with ethical principles, data protection laws (eg, General Data Protection Regulation and HIPAA), and international guidance [26]. This requires establishing governance infrastructure—such as AI steering committees and ethics review boards—to specify authorized use cases, roles and responsibilities, liability frameworks, and compliance protocols. Crucially, policies must explicitly delineate accountability boundaries among supervising clinicians, health care institutions, and AI developers and vendors, particularly for autonomous or semiautonomous agentic workflows. In this capacity, clinician leaders do not personally conduct technical audits; instead, they serve as the strategic link between medical staff and technical bodies, ensuring that institutional processes reflect clinical realities and patient safety risks. Second, model change management requires supervision of domain adaptation (eg, task- or specialty-specific tuning) within a multidisciplinary process. In this capacity, clinicians

XSL•FO
RenderX

bridge clinical needs and technical implementation, upholding standards for validity, equity, and safety. This supervision necessitates predefined evaluation plans, comprehensive documentation (eg, model cards), and rigorous external validation (including multicenter, temporal, and geographic shift tests) before production deployment. Leaders must specify minimum performance thresholds and mandate shadow deployment phases to validate safety before full patient exposure [33,34]. Third, model life cycle oversight entails governing AI systems across their full life cycle—from validation through postmarket monitoring, updating, and decommissioning. This includes orchestrating institutional processes for drift detection, performance re-evaluation, and version control [35,36] (Textbox 1). Leaders must navigate complex regulatory mechanisms for iterative improvement, such as the predetermined change control plans (PCCPs) by the FDA [6] and the postmarket surveillance requirements of the EU AI Act [7]. They collaborate with informatics, regulatory, and quality teams to ensure that updates, retraining, or expanded indications are clinically justified, transparently communicated, and supported by robust evidence and incident review protocols [35,37].

Textbox 1. Clinical vignette—governance in action.

Scenario: executing a manual rollback protocol

- A clinical informatics director oversees a deployed discharge-summary agent. During routine postmarket surveillance, the monitoring dashboard signals that the model's summarization accuracy has dropped below the prevalidated threshold of 95% (a metric specified in the Food and Drug Administration–accepted predetermined change control plan). Attributing the decline to data drift caused by a recent update in the hospital's note-template format, the director initiates a rollback protocol—leveraging either institutional version control or a vendor-mediated pathway specified in the service-level agreement. The deployment is rolled back to the previous stable version (version 2.0) while the technical team remediates and revalidates the updated model (version 2.1).
- In deployments where direct rollback is technically unsupported (eg, some software as a service–based integrations), the protocol mandates pausing or disabling the agent and reverting to manual workflows until remediation is complete.

Alignment and Differentiation From Existing Frameworks

This framework is broadly aligned with the American Medical Association's guidance on augmented intelligence, prioritizing physician leadership, transparency, and patient benefit [38,39]. Furthermore, it adheres to competency-based digital education frameworks from the WHO and the Association of American Medical Colleges, both of which prioritize observable behaviors and measurable learning outcomes [40-42]. It also builds on recent competency proposals in AI and digital health that foreground digital health literacy, awareness of data bias, and the ethical use of assistive tools [10,43]. As summarized in Table 1, our contribution lies in extending these earlier frameworks to the agentic LLM era. First, we explicitly differentiate between predictive and informational paradigms and agentic workflows. Accordingly, we move from clinicians interpreting decision support outputs to supervising and governing active, tool-using agents. Second, we introduce model life cycle literacy as an explicit competency domain, encompassing familiarity with mechanisms for ongoing monitoring, updating, and regulatory adaptation. Within this broader, jurisdiction-agnostic concept, PCCPs in the FDA context are presented as one concrete example, alongside emerging requirements under frameworks, such as the EU AI Act. To our knowledge, previous frameworks have not explicitly integrated agent supervision and life cycle–oriented governance into a tiered, clinician-facing competency model.

Table 1. Comparison of the agent to the governance framework and existing digital health competency frameworks.

Feature and dimension	Agent to governance framework	Existing frameworks (eg, World Health Organization, American Medical Association, and Association of American Medical Colleges)
Technological scope	Agentic and autonomous: agentic workflows (perceive-plan-act loops) and tool-using large language models that execute multistep tasks	Predictive and informational: clinical decision support, diagnostic classifiers, and standard information retrieval systems
Clinician's role	Supervisor and governor: human-on-the-loop oversight for task delegation, monitoring agent behavior, and managing bounded autonomy	Interpreter and decision-maker: human-in-the-loop integration, focusing on the critical appraisal of risk scores and diagnostic suggestions
Verification skills	Output verification and logic checking: detection of hallucinations in generative text and verification of agentic tool calls (eg, application programming interface actions)	Statistical and evidence-based appraisal: evaluation of model performance metrics (eg, sensitivity and specificity), data quality, and automation bias
Regulatory and life cycle	Life cycle management: specific literacy in predetermined change control plans, algorithmic drift detection, and postmarket surveillance (eg, European Union Artificial Intelligence Act and Food and Drug Administration)	Foundational ethics and compliance: adherence to core bioethical principles (beneficence and equity), privacy standards (Health Insurance Portability and Accountability Act and General Data Protection Regulation), and informed consent
Target audience and structure	Tiered differentiation: distinguishes between frontline users (tiers 1 and 2) and a specialized leadership track (tier 3) for governance	Universal digital literacy: baseline digital health competencies applicable to the broad health care workforce to ensure safe general use

XSL•FO
RenderX

Operationalizing Competencies for Education and Assessment

Translation of this framework into continuing medical education (CME) curricula requires the specification of observable, assessable behaviors aligned with competency-based medical education principles. Given that the clinical workforce encompasses diverse roles—including physicians, nurses, and allied health professionals—implementation and assessment should be tailored to role-specific scope of practice and role-based EHR access controls. For example, behavioral indicators involving least privilege enforcement or the rejection of agent actions may be operationalized differently depending on the individual's credentialed permissions and administrative privileges. To ensure implementation feasibility and mitigate workforce burden, only tier 1 competencies are intended for the general clinical workforce, whereas tiers 2 and 3 are reserved for smaller groups of superusers and clinician leaders in formal governance roles. To avoid adding entirely new courses, these competencies are designed to be integrated into existing curricula (eg, evidence-based medicine, clinical reasoning, and quality and safety) and CME activities. Institutions are responsible for resourcing and coordinating these training activities, ensuring that individual clinicians are not expected to acquire advanced competencies (tiers 2 and 3) without appropriate organizational support and protected time.

Table 2 links each tier to behavioral indicators written as active, measurable learning outcomes. Indicators span tier 1 (eg, identifying hallucinations) to tier 3 (eg, initiating life cycle protocols) and should be tailored to role-specific responsibilities and decision rights. These indicators provide curriculum developers with a concrete scaffolding to design simulation-based, workplace-based, and microlearning assessments that verify skill acquisition in clinical practice. Ultimately, these anchors facilitate the incorporation of this framework into CME curricula and clinical job descriptions, thereby promoting institutional transparency, accountability, and regulatory alignment [40,43,44].

Table 2. Sample behavioral indicators for continuing medical education assessment and clinical application.

Core competency	Behavioral indicator (observable action)
Tier 1: foundational (frontline user)	
Prompt engineering	Formulates a context-aware prompt that includes explicit role definitions (eg, act as a cardiologist), constraints, and required output formats, without disclosing PHI[a]
Human-AI[b] agent interaction	Identifies and intercepts inappropriate agent requests (eg, social history for refills) and enforces denial or escalation protocols based on least privilege guardrails and predefined termination and handoff criteria
Output verification and validation	Detects and corrects a hallucinated reference or dosage in a large language model–generated draft by cross-referencing with the patient's structured laboratory data and trusted guidelines
Security and privacy awareness	Uses minimum necessary data; deidentifies data for nonintegrated tools; for electronic health record agents, verifies patient context and least privilege access, and escalates PHI or policy risks
Clinician-patient interface	Informs patients when AI is used, explains AI-derived insights in patient-appropriate language (including uncertainties and limitations), and documents consent or refusal when clinically indicated.
Tier 2: intermediate (superuser or champion)	
Bias detection and mitigation	Interprets stratified subgroup performance and uncertainty reports, flags clinically meaningful disparities, triggers mitigation (eg, threshold adjustments or human-review rules), and verifies improvement via updated audit reports
Systemic evaluation	Evaluates a confusion matrix for a diagnostic AI tool to determine if the false-negative rate is acceptable for a specific screening population
Explainability	Interprets available explainability outputs (eg, feature importance, reason codes, or saliency maps where available) to detect spurious cues and document potential failure modes
Clinical integration	Defines where AI outputs enter the workflow; assigns roles, documentation, and escalation steps; and maintains clinician accountability when AI recommendations conflict
Tier 3: advanced (governance leader)	
Ethical and regulatory governance	Drafts and implements an institutional policy that establishes escalation pathways and explicitly delineates accountability and liability boundaries among the supervising clinician, the institution, and the AI developer for autonomous agentic workflows
Model change management	Initiates and justifies model change requests (eg, recalibration, retraining, or expanded indication), defining the clinical rationale, validation plan, and monitoring criteria consistent with the predetermined change control plan
Model life cycle oversight	Oversees monitoring of model performance and drift (eg, calibration, error rates, and data shifts); ensures execution of incident protocols and predefined controls (eg, roll back and human review)

[a]PHI: protected health information.

[b]AI: artificial intelligence.

XSL•FO
RenderX

Limitations and Future Work

We propose a governance-aligned competency framework designed to guide clinicians in the safe and effective use of LLMs in clinical practice. However, several limitations should be acknowledged. First, external validity may differ by specialty, care setting (inpatient vs ambulatory), health-system maturity, and EHR integration capacity. Critically, the institutional infrastructure required for tier 3—specifically, the establishment of AI steering committees—may currently be feasible only in resource-rich academic medical centers. Mandating such governance structures in resource-constrained community hospitals may be impractical. This feasibility gap extends to global health contexts; the framework requires adaptation in low- and middle-income settings where informatics infrastructure, governance capacity, and regulatory regimes differ substantially. Moreover, the objective structured clinical examination blueprint [44] and key performance indicators remain surrogate end points. By themselves, these measures do not guarantee improvements in patient-centered outcomes (eg, adverse events and readmissions). Second, a prospective, multicenter external evaluation is still necessary. Although we specify fairness analyses and minimum subgroup sample size and performance thresholds, real-world coverage across languages, cultural contexts, pediatrics, geriatrics, and rare-disease pathways is likely incomplete. Third, the regulatory environment remains dynamic as harmonized standards under the EU AI Act and FDA change control frameworks (eg, PCCPs) continue to evolve. Accordingly, operationalized procedures and performance thresholds should be periodically reassessed—particularly following material model updates—to sustain regulatory compliance and clinical relevance. Fourth, the automation paradox (skill decay) warrants attention. While AI agents improve efficiency, they may precipitate clinician deskilling over time. Safe fallback protocols (eg, reverting to manual workflows during system failure) are feasible only if clinicians maintain underlying diagnostic and procedural competence. To mitigate this risk, organizations should integrate automation-off scenarios into simulation training and downtime contingency plans to ensure clinicians remain capable of detecting errors and safely resuming control.

Given the rapid evolution of clinical AI, this framework requires ongoing refinement across 5 strategic areas. First, validation studies use expert consensus (eg, Delphi methods) and multicenter educational trials to link tiered competencies to observable clinical behaviors, such as error interception, safe deferral, and workflow efficiency. Second, assessment science strengthens psychometric measurement by refining objective structured clinical examination stations, bias and calibration checklists, and reliability targets (eg, generalizability coefficient [G] ≥0.70) [45,46]. Third, capacity building establishes faculty development programs and reusable educational resources—including deidentified sandboxes, annotated audit log exemplars, and HIPAA-aligned exercise sets—to support cross-specialty implementation. Fourth, advanced equity and uncertainty quantification cultivates practical competence in algorithmic fairness and uncertainty management through routine subgroup audits, coverage targets and reporting (using conformal prediction where appropriate), and bedside remediation playbooks. Fifth, simulation and institutional integration evaluate the feasibility and effectiveness of embedding automation-on and automation-off scenarios and rollback procedures within existing simulation programs and downtime contingency planning (eg, multidisciplinary team discussions). Outcomes include competency attainment, error interception during failures, auditability, and pathways for formal recognition within CME credit structures and institutional job descriptions.

Conclusions

The progressive competency model integrates technical proficiency with ethical governance to provide clinicians with essential AI skills for the LLM era. By embedding these competencies into CME standards and job descriptions—using clear, observable behaviors—institutions can standardize safe and accountable AI use. Preparing for an AI-augmented future requires integrating governance-focused skills into medical training and professional development. This approach positions clinicians as responsible stewards of AI, ensuring adoption remains sustainable and centered on patient care.

Acknowledgments

Following the initial drafting of the manuscript, the authors used ChatGPT (OpenAI) for artificial intelligence–assisted language editing to enhance clarity and style. All suggested edits were subject to critical review, and the authors subsequently made additional substantive revisions. The authors assume complete responsibility for the integrity and accuracy of the final content.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Data Availability

Data sharing is not applicable to this paper as no new datasets were generated or analyzed during this study.

XSL•FO
RenderX

Authors' Contributions

JL, QZ, and SL conceived and designed the study. WC and JL led data curation, with literature analysis by WC, JL, QZ, and SL. JL, WC, QZ, and SL contributed to writing the original draft. All authors critically revised the manuscript, approved the final version, and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

References

1. Angus DC, Khera R, Lieu T, Liu V, Ahmad FS, Anderson B, JAMA Summit on AI. AI, health, and health care today and tomorrow: the JAMA Summit Report on Artificial Intelligence. JAMA 2025 Nov 11;334(18):1650-1664. [doi: 10.1001/jama.2025.18490] [Medline: 41082366]
2. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. World Health Organization. 2025 Mar 25. URL: https://www.who.int/publications/i/item/9789240084759 [accessed 2025-09-10]
3. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun 28;25:e48568 [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
4. Liu S, Huang SS, McCoy AB, Wright AP, Horst S, Wright A. Optimizing order sets with a large language model-powered multiagent system. JAMA Netw Open 2025 Sep 02;8(9):e2533277 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.33277] [Medline: 40986301]
5. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, FUTURE-AI Consortium. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ 2025 Feb 05;388:e081554 [FREE Full text] [doi: 10.1136/bmj-2024-081554] [Medline: 39909534]
6. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions. United States Food and Drug Administration. 2025 Aug. URL: https://tinyurl.com/5dr3xx6j [accessed 2025-09-12]
7. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). European Union. 2024 Jun 13. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng [accessed 2025-09-10]
8. Augmented intelligence in health care. American Medical Association. 2018. URL: https://www.ama-assn.org/system/files/2019-01/augmented-intelligence-policy-report.pdf [accessed 2025-09-16]
9. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. 2021 Jun 28. URL: https://www.who.int/publications/i/item/9789240029200 [accessed 2025-09-10]
10. Xi Z, Chen W, Guo X, He W, Ding Y, Hong B, et al. The rise and potential of large language model based agents: a survey. Sci China Inf Sci 2025 Jan 17;68(2):121101. [doi: 10.1007/s11432-024-4222-0]
11. Jiang Y, Black KC, Geng G, Park D, Zou J, Ng AY, et al. MedAgentBench: a virtual EHR environment to benchmark medical LLM agents. NEJM AI 2025 Aug 28;2(9):1-9. [doi: 10.1056/aidbp2500144]
12. Skalidis I, Maurizi N, Salihu A, Fournier S, Cook S, Iglesias JF, et al. Artificial intelligence and advanced digital health for hypertension: evolving tools for precision cardiovascular care. Medicina (Kaunas) 2025 Sep 04;61(9):1597 [FREE Full text] [doi: 10.3390/medicina61091597] [Medline: 41010987]
13. Liu J, Liu F, Fang J, Liu S. The application of Chat Generative Pre-trained Transformer in nursing education. Nurs Outlook 2023 Nov;71(6):102064. [doi: 10.1016/j.outlook.2023.102064] [Medline: 37879261]
14. Schuitmaker L, Drogt J, Benders M, Jongsma K. Physicians' required competencies in AI-assisted clinical settings: a systematic review. Br Med Bull 2025 Jan 16;153(1):ldae025. [doi: 10.1093/bmb/ldae025] [Medline: 39821209]
15. Scott IA, Shaw T, Slade C, Wan TT, Barmanray R, Coorey C, et al. Proposing core competencies for physicians in using artificial intelligence tools in clinical practice. Intern Med J 2025 Aug;55(8):1403-1409. [doi: 10.1111/imj.70112] [Medline: 40576330]
16. Goodman KE, Rodman AM, Morgan DJ. Preparing physicians for the clinical algorithm era. N Engl J Med 2023 Aug 10;389(6):483-487. [doi: 10.1056/NEJMp2304839] [Medline: 37548320]
17. Russell RG, Lovett Novak L, Patel M, Garvey KV, Craig KJ, Jackson GP, et al. Competencies for the use of artificial intelligence-based tools by health care professionals. Acad Med 2023 Mar 01;98(3):348-356. [doi: 10.1097/ACM.0000000000004963] [Medline: 36731054]
18. Liu J, Liu F, Wang C, Liu S. Prompt engineering in clinical practice: tutorial for clinicians. J Med Internet Res 2025 Sep 15;27:e72644 [FREE Full text] [doi: 10.2196/72644] [Medline: 40955776]
19. Naveen K, Sajja WS, Nerella A. Building secure AI agents for autonomous data access in compliance/regulatory-critical environments. Comput Fraud Secur 2024;2024(9):363-373 [FREE Full text] [doi: 10.52710/cfs.746]
20. Huang K. Agentic AI: Theories and Practices. Cham, Switzerland: Springer; 2025.

XSL•FO
RenderX

21. Acharya DB, Kuppan K, Divya B. Agentic AI: autonomous intelligence for complex goals—a comprehensive survey. IEEE Access 2025;13:18912-18936. [doi: 10.1109/access.2025.3532853]
22. Borkowski AA, Ben-Ari A. Multiagent AI systems in health care: envisioning next-generation intelligence. Fed Pract 2025 May;42(5):188-194. [doi: 10.12788/fp.0589] [Medline: 40831649]
23. Lu JG, Song LL, Zhang LD. Cultural tendencies in generative AI. Nat Hum Behav 2025 Nov 20;9(11):2360-2369. [doi: 10.1038/s41562-025-02242-1] [Medline: 40542181]
24. Grazhdanski G, Vasilev V, Vassileva S, Taskov D, Antova I, Koychev I, et al. SynthMedic: utilizing large language models for synthetic discharge summary generation, correction and validation. J Biomed Inform 2025 Oct;170:104906 [FREE Full text] [doi: 10.1016/j.jbi.2025.104906] [Medline: 40962129]
25. Jonnagaddala J, Wong ZS. Privacy preserving strategies for electronic health records in the era of large language models. NPJ Digit Med 2025 Jan 16;8(1):1-3 [FREE Full text] [doi: 10.1038/s41746-025-01429-0] [Medline: 39820020]
26. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023 Aug 11;25:e48009 [FREE Full text] [doi: 10.2196/48009] [Medline: 37566454]
27. Ganta T, Kia A, Parchure P, Wang MH, Besculides M, Mazumdar M, et al. Fairness in predicting cancer mortality across racial subgroups. JAMA Netw Open 2024 Jul 01;7(7):e2421290 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.21290] [Medline: 38985468]
28. Omar M, Sorin V, Agbareia R, Apakama DU, Soroush A, Sakhuja A, et al. Evaluating and addressing demographic disparities in medical large language models: a systematic review. Int J Equity Health 2025 Feb 26;24(1):57 [FREE Full text] [doi: 10.1186/s12939-025-02419-0] [Medline: 40011901]
29. Mesinovic M, Watkinson P, Zhu T. Explainability in the age of large language models for healthcare. Commun Eng 2025 Jul 17;4(1):128 [FREE Full text] [doi: 10.1038/s44172-025-00453-y] [Medline: 40676176]
30. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature 2025 Jun 09;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
31. Berzin TM, Topol EJ. Preserving clinical skills in the age of AI assistance. The Lancet 2025 Oct;406(10513):1719. [doi: 10.1016/s0140-6736(25)02075-6]
32. Kim SH, Wihl J, Schramm S, Berberich C, Rosenkranz E, Schmitzer L, et al. Human-AI collaboration in large language model-assisted brain MRI differential diagnosis: a usability study. Eur Radiol 2025 Sep 07;35(9):5252-5263. [doi: 10.1007/s00330-025-11484-6] [Medline: 40055233]
33. Dorfner FJ, Dada A, Busch F, Makowski MR, Han T, Truhn D, et al. Evaluating the effectiveness of biomedical fine-tuning for large language models on clinical tasks. J Am Med Inform Assoc 2025 Jun 01;32(6):1015-1024. [doi: 10.1093/jamia/ocaf045] [Medline: 40190132]
34. Makhni S, Rico J, Cerrato P, Hill B, Overgaard S, Wu J, et al. A comprehensive approach to responsible AI development and deployment. Mayo Clin Proc Digit Health 2025 Dec;3(4):100294 [FREE Full text] [doi: 10.1016/j.mcpdig.2025.100294] [Medline: 41282583]
35. M S AR, C R N, B R S, Lahza H, Lahza HF. A survey on detecting healthcare concept drift in AI/ML models from a finance perspective. Front Artif Intell 2023 Apr 17;5:955314 [FREE Full text] [doi: 10.3389/frai.2022.955314] [Medline: 37139355]
36. Kore A, Abbasi Bavil E, Subasri V, Abdalla M, Fine B, Dolatabadi E, et al. Empirical data drift detection experiments on real-world medical imaging data. Nat Commun 2024 Feb 29;15(1):1887 [FREE Full text] [doi: 10.1038/s41467-024-46142-w] [Medline: 38424096]
37. Moskalenko V, Kharchenko V. Resilience-aware MLOps for AI-based medical diagnostic system. Front Public Health 2024 Mar 27;12:1342937 [FREE Full text] [doi: 10.3389/fpubh.2024.1342937] [Medline: 38601490]
38. Augmented intelligence development, deployment, and use in health care. American Medical Association. 2024 Nov. URL: https://www.ama-assn.org/system/files/ama-ai-principles.pdf [accessed 2025-09-16]
39. Augmented intelligence in medicine. American Medical Association. URL: https://www.ama-assn.org/practice-management/digital-health/augmented-intelligence-medicine [accessed 2025-09-10]
40. Digital education for building health workforce capacity. World Health Organization. 2020 Apr 14. URL: https://www.who.int/publications/i/item/9789240000476 [accessed 2025-09-10]
41. Global competency and outcomes framework for the essential public health functions. World Health Organization. 2024 Jun 10. URL: https://www.who.int/publications/i/item/9789240091214 [accessed 2025-09-18]
42. Artificial intelligence competencies for medical educators. Association of American Medical Colleges. URL: https://tinyurl.com/44t88c4c [accessed 2025-10-21]
43. Gazquez-Garcia J, Sánchez-Bocanegra CL, Sevillano JL. AI in the health sector: systematic review of key skills for future health professionals. JMIR Med Educ 2025 Feb 05;11:e58161 [FREE Full text] [doi: 10.2196/58161] [Medline: 39912237]
44. Milan FB, Grochowalski JH. A resource efficient and reliable standard setting method for OSCEs: borderline regression method using standardized patients as sole raters in clinical case encounters with medical students. Med Teach 2022 Aug;44(8):878-885. [doi: 10.1080/0142159X.2022.2041586] [Medline: 35234562]
45. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. Med Teach 2012;34(11):960-992. [doi: 10.3109/0142159X.2012.703791] [Medline: 23140303]

XSL•FO
RenderX

46. Park SY, Lee SH, Kim MJ, Ji KH, Ryu JH. Acceptability of the 8-case objective structured clinical examination of medical students in Korea using generalizability theory: a reliability study. J Educ Eval Health Prof 2022 Sep 08;19:26 [FREE Full text] [doi: 10.3352/jeehp.2022.19.26] [Medline: 36071557]

Abbreviations

AI: artificial intelligence
CME: continuing medical education
EHR: electronic health record
EU: European Union
FDA: Food and Drug Administration
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
PCCP: predetermined change control plan
WHO: World Health Organization

Edited by A Coristine; submitted 26.Oct.2025; peer-reviewed by S Palmieri, A Algumaei; comments to author 24.Nov.2025; revised version received 23.Dec.2025; accepted 23.Dec.2025; published 14.Jan.2026.

Please cite as:
Cao W, Zhang Q, Liu J, Liu S
From Agents to Governance: Essential AI Skills for Clinicians in the Large Language Model Era
J Med Internet Res 2026;28:e86550
URL: https://www.jmir.org/2026/1/e86550
doi:10.2196/86550
PMID:

©Weiping Cao, Qing Zhang, Jialin Liu, Siru Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Collaborative and Cooperative Hospital "In-House" Medical Device Development and Implementation in the AI Age: The European Responsible AI Development (EURAID) Framework Compatible With European Values

Anett Schönfelder[1], MBA; Maria Eberlein-Gonska[2], Prof Dr; Manfred Hülsken-Giesler[3], Prof Dr; Florian Jovy-Klein[4], MSc; Jakob Nikolas Kather[1,5,6,7], Prof Dr; Elisabeth Kohoutek[8], RA; Thomas Lennefer[9], Dr rer nat, MSc; Elisabeth Liebert[10], MA; Myriam Lipprandt[11], Prof Dr; Rebecca Mathias[1], MSc; Hannah Sophie Muti[1,12], MD, Dr med; Julius Obergassel[13], MD, MHBA; Thomas Reibel[4], MSc; Ulrike Rösler[14], PhD; Moritz Schneider[15], MSc; Larissa Schlicht[16], MSc; Hannes Schlieter[17], Prof Dr; Malte L Schmieding[18], MD, MSc; Nils Schweingruber[13], MD; Martin Sedlmayr[19], Prof Dr; Reinhard Strametz[20], Prof Dr; Barbara Susec[21], BA; Magdalena Katharina Wekenborg[1], Dr rer nat; Eva Weicken[22], MD; Katharina Weitz[22], Dr rer nat; Anke Diehl[10*], Dr med, MD; Stephen Gilbert[1,17*], Prof Dr

[1]Else Kroener Fresenius Center for Digital Health, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Saxony, Germany

[2]University Hospital Carl Gustav Carus, Dresden, Germany

[3]Department of Nursing Science, Institute of Health Research and Education, University of Osnabrück, Osnabrück, Germany

[4]Institute for Technology and Innovation Management (TIM), RWTH Aachen University, Aachen, Germany

[5]Department of Medicine I, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Saxony, Germany

[6]Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany

[7]Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, United Kingdom

[8]Luther Rechtsanwaltsgesellschaft mbH, Cologne, Germany

[9]Department for Prevention, AOK Federal Association, Berlin, Germany

[10]Department for Digital Transformation, University Medicine Essen, Essen, Germany

[11]Institute of Medical Informatics, University Hospital RWTH Aachen, Aachen, North Rhine-Westphalia, Germany

[12]Department for Visceral, Thoracic and Vascular Surgery, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany

[13]University Medical Center Hamburg - Eppendorf, Hamburg, Germany

[14]Federal Institute for Occupational Safety and Health, Berlin, Germany

[15]Institute for Occupational Safety and Health of the German Social Accident Insurance, Sankt Augustin, Germany

[16]Faculty of Humanities and Social Sciences, Karlsruhe Institute of Technology, Karlsruhe, Germany

[17]Faculty of Business and Economics, TUD Dresden University of Technology, Dresden, Germany

[18]German Federal Ministry of Health, Berlin, Germany

[19]Institute for Medical Informatics and Biometry, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany

[20]Wiesbaden Institute for Healthcare Economics and Patient Safety (WiHelP), Wiesbaden, Germany

[21]ver.di, Berlin, Germany

[22]Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, Berlin, Germany

[*]these authors contributed equally

Corresponding Author:
Anett Schönfelder, MBA
Else Kroener Fresenius Center for Digital Health
Faculty of Medicine and University Hospital Carl Gustav Carus
TUD Dresden University of Technology
Fetscherstr. 74
Dresden, Saxony, 01307
Germany

XSL•FO
RenderX

Phone: 49 35145815840
Email: anett.schoenfelder@ukdd.de

Abstract

The last years have seen an acceleration in the development and uptake of artificial intelligence (AI) systems by "early adopter" hospitals, caught between the pressures to "perform" and "transform" in a struggling health care system. This transformation has raised concerns among health care providers as their voices and location-specific workflows have often been overlooked, resulting in technologies that fail to integrate meaningfully into routine care and worsen rather than improve care processes. How can positive AI implementation be carried out in health care, aligned with European values? Based on a perspective that spans all stakeholders, we have created EURAID (European Responsible AI Development), a practical, human-centric framework for AI development and implementation based on agreed goals and values. We illustrate this approach through the co-development of a narrow-purpose "in-house" AI system, designed to help bridge the AI implementation gap in real-world clinical settings. This example is then expanded to address the broader challenges associated with complex, multiagent AI systems. By portraying all key stakeholders across the AI development life cycle and highlighting their roles and contributions within the process, real use cases, and methods for achieving iterative consensus, we offer a unique practical approach for safe and fast progress in hospital digital transformation in the AI age.

(J Med Internet Res 2026;28:e80754) doi:10.2196/80754

KEYWORDS

AI Act; digital transformation; in-house medical device development; agentic AI; AI life cycle; artificial intelligence

The Transformation of Future Medicine Through Artificial Intelligence Technologies

Will the slogans already heard in health care system strikes, such as "Trust Nurses, Not AI" and "AI has got to go!"[1,2], become more common? They reflect growing concerns about the evolving role of health care professionals (HCPs) in a changing health system, which persist despite reports that 20% of National Health Service (NHS) doctors are already using artificial intelligence (AI) daily [3]. Although the importance of digital transformation to enhance the efficiency of care delivery and to provide better models of care suited to modern age [4-6] is well recognized within care systems [7-11], it often cannot be comprehensively addressed, as health care systems worldwide find themselves caught between the need to both "perform" and "transform" in a system facing "firefighting" ongoing challenges [12-17]. The application of AI technologies has the potential to address some of those aspects (Table 1), as it can speed digital transformation and can (at least if applied well and if the associated potential barriers and uncertainties are jointly recognized and resolved) make health care more accessible, effective, and economically sustainable [18]. Examples of the positive impact of good AI implementation are (1) enhancement of clinical practice, particularly in areas such as diagnosis and personalized medicine [11,19,20]; (2) workflow improvements, by supporting administrative tasks such as transcription, patient communication, and patient-related recordkeeping [21,22]; and (3) increased operational efficiency, through the optimization of routine processes, enabling HCPs to work in a more patient-centered way [23], and potentially contribute to cost reductions [24,25]. With the recent introduction of "agentic AI" [26-29] and autonomous AI-enabled systems [30,31], far more systematic complexity can be handled by AI [32].

XSL•FO
RenderX

Table 1. Problems artificial intelligence (AI)–enabled transformation can address, approaches, challenges, and possible unintended consequences.

Current health system problems	Possible digital and AI[a] solutions	Implementation challenges and risks
Administrative workload unrelated to direct patient care [33,34], inefficient workflows, and fragmented communication burden on HCPs[b].	Automation of administrative and routine tasks, and AI-driven workflow optimization, allowing people to focus on patients.	• Different perspectives on which tasks to automate. • Increase in workload in some cases. • Risk of overreliance on AI outcomes with insufficient human oversight. • Automation of the current way of providing care without restructuring and rethinking processes. • Concerns about job security, the transformation of job roles, and medical malpractice.
Stress, duplication (eg, medical history) [35], and discontinuous care resulting from disconnected devices, limited interoperability, and manual coordination.	Adjusting the hospital's IT environment as an AI-sustained platform, characterized by high interoperability in itself and with other providers supporting seamless patient journeys.	• Deficient data quality, data silos, inadequate computational resources, a shortage of specialized expertise, and poor or nonexistent infrastructure between providers. • Concerns about safety and regulation.
Poor information flow and HCP training deficit.	AI-supported knowledge management to build confidence in usage.	• Shortage of HCPs limits time for training. • Various adoption readiness levels among HCPs. • Concerns about trust in technologies.

[a]AI: artificial intelligence.

[b]HCP: health care professional.

However, AI is not a panacea, and initial evaluations of real-world performance in clinical settings are mixed [36-38]. One reason is that AI implementation projects have often underestimated the importance of individual AI medical devices operating as interconnected clinical and technological infrastructures rather than being a collection of isolated, standalone algorithms. AI in health care over the next years needs to be seen as interacting, interdependent, and flexible applications [39], involving both broad- and narrow-purpose tools and models that closely interact with and reshape human workflows, while simultaneously, human workflows, adaptations, and experience reshape the use of AI, particular to the local setting and local approach to health care delivery.

Integration of Interactive AI Systems in Clinical Workflows Requires HCPs at the Core, Not as Observers

This future model needs HCPs at its core, not only as users interacting with AI systems, but as active participants in their co-design, procurement, implementation, monitoring, and evaluation. This idea is rooted in organizational and implementation theories, such as the "socio-technical systems theory" [40], that emphasizes the importance of a holistic perspective to jointly bridge human and technological capabilities, particularly in the context of autonomous technologies [41,42], and the "normalization process theory" [43], which acknowledges users' cognitive participation and collective action as key determinants in implementing, embedding, and integrating complex and new interventions (eg, AI systems) in daily practice [44,45]. "Human-centered AI" can take a cross-theoretical perspective by viewing AI systems not as stand-alone technologies, but as integral components of a broader sociotechnical system. Two perspectives are relevant: humans being able to understand AI and AI being able to understand humans [46]. For example, explainable AI (XAI) methods should not only address the technical transparency of machine learning models but also focus on human understanding [47]. On the other hand, AI systems need to take into account the needs, requirements, and mental models of humans [48] and the context of clinical decisions [49] to create explanations that are supportive in the clinical setting.

Yet, despite the substantial body of research on theoretical foundations, the translation of the underlying principles into everyday implementation of AI systems and clinical reality is lagging behind [50-54], often key aspects are neglected, and many implementation projects fail [55]. Problems often begin during the development of AI systems, which are frequently designed and tested in settings that are far removed from the everyday realities of clinical practice [56], and with HCPs and location-specific workflows often overlooked. The consequences of systems designed without sustained input from HCP and patients [57] are visible as they fail to demonstrate their suitability and worsen rather than improve processes, leading to the perception that the introduction of digital technologies into health care adds to the burden [57,58] (Figure 1), although general relief through well-implemented work aids would be very welcomed. That misalignment has been associated with increased stress among HCPs [59] (including "technostress" [60,61]), disconnected patient care [62,63], and has even resulted in other unintended negative consequences, such as HCPs resisting the use of the technologies [15], using technologies in unanticipated ways [64], or developing workarounds that may endanger patient care [65]. Insufficient digital health literacy and training among HCPs amplify these effects, leaving HCPs unprepared for the demands of interacting with intelligent systems [66]. Other consequences appearing in real-world implementation are model uncertainty [67], "AI hallucinations" or clinically harmful recommendations, bias [14], and context misalignment [68], which risk fragmented care and diminish patients' trust in technology-assisted decisions.

XSL•FO
RenderX

Figure 1. The introduction of artificial intelligence (AI) into clinical workflows is changing everyday clinical care and could, at least theoretically, enhance satisfaction, empower, upskill, and provide a better work environment and better interactions for health care professionals (HCPs) and patients; however, the reality is often much less positive. The upper circle is showing the current situation of health care delivery, which is characterized by a low level of digitalization and an ever-increasing amount of nonpatient-related activities, causing moderate satisfaction and happiness among both HCPs and patients. Care delivery transformation through AI can bring positive effects as shown in the green circle on the right (such as delivering better, more efficient and even more patient-centered care through optimized processes and well-balanced support systems) or, as is frequently the case, negative effects (red circle on the left), causing frustration, disconnection and stress of HCPs and patients because of interoperability issues with AI implementations that were never properly designed with the user needs in mind.

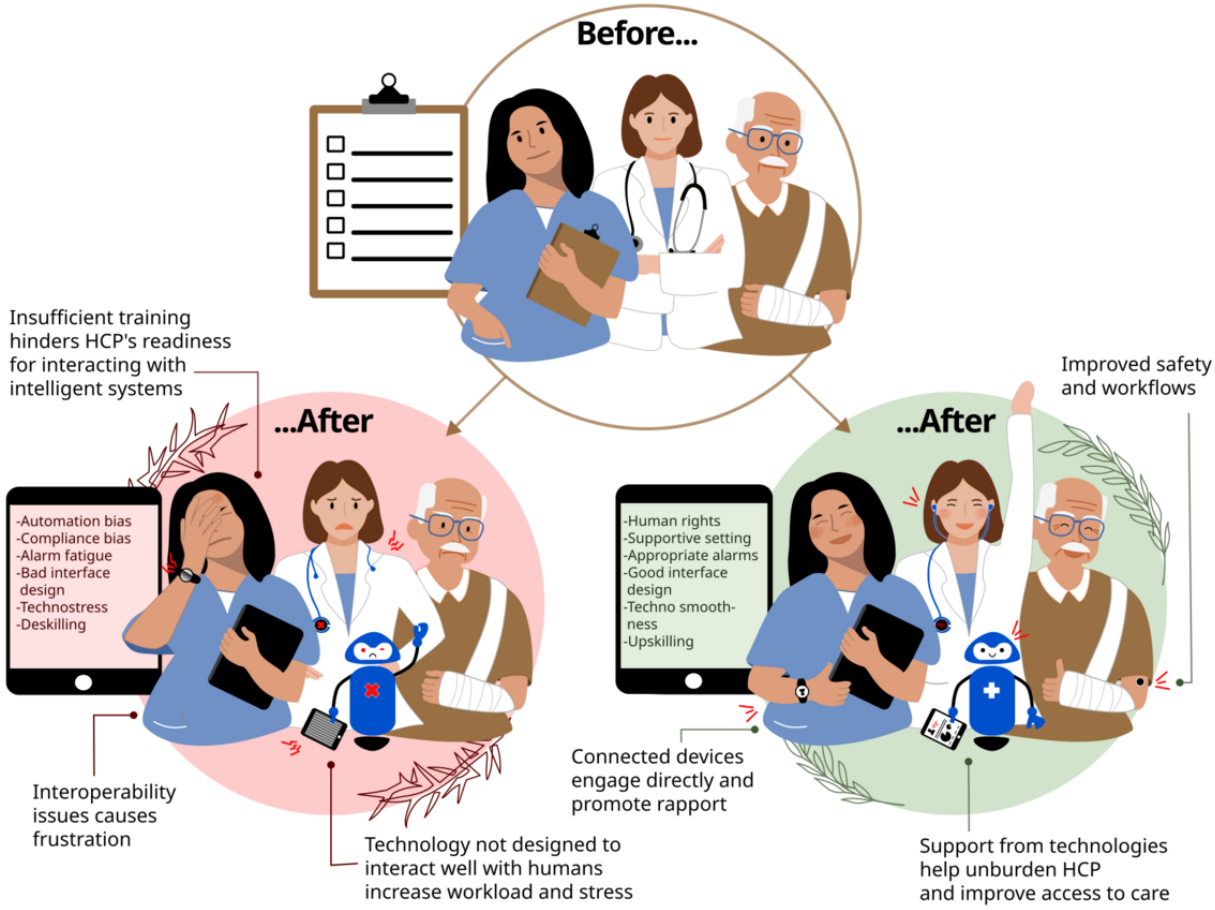

Improving Adoption by Co-Development Across the AI Life Cycle

Overview

The real-world challenges discussed underscore that successful AI development and implementation are less a technical task than a comprehensive change management process [57] that needs active participation, transparent governance, continuous feedback, and development beyond technical metrics, including systematic real-world evaluation of human-AI interaction, and a focus on non-technical design criteria such as usability, workflow fit, trust, and acceptance.

To bridge this gap, we propose EURAID (European Responsible AI Development), a practical framework of human-centric AI development and implementation in hospitals, which is cooperative and collaborative and based on shared goals in accordance with European values according to Article 2 of the Treaty on European Union (TEU; ie, human dignity [69,70], freedom [69,71], democracy [72], equality [69], rule of law [73], and human rights [69,74]) and European laws (Table 2).

XSL•FO
RenderX

Table 2. Regulations in the European Union and its member states that guide AI use in health care (nonexhaustive).

Regulation or law	Scope	Approach
Medical Device Regulation (MDR; 2017/745)	Governs medical devices (including digital systems) used for diagnostic or therapeutic purposes.	• The medical devices' intended purpose defines the associated performance claims, which must be substantiated through clinical evaluation. • GSPRs[a] must be met, including structured risk management (ISO[b] 14971:2019), a certified QMS[c] (ISO 13485:2016), usability engineering (IEC[d] 62366-1:2015+A1:2020), and a planned and documented development process (IEC 62304:2006+A1:2015), depending on the respective product category.
Artificial Intelligence Act (AI Act; 2024/1689)	Governs the development, market entry, and use of AI[e] systems.	• Classifies high-risk AI systems (including AI-enabled medical devices, class IIa+) and GPAI[f] (that can perform a wide range of tasks, not limited to one clear intended purpose) and LLM[g] models, depending on both the function performed and the systems' intended purpose. • Additional transparency obligations apply for certain systems such as emotion recognition, biometric categorization, and interactive or generative AI.
EU Occupational Safety and Health Directive (89/391/EEC 1989) and national laws	Ensures workers' health and safety.	• Systematic risk assessments and preventive measures. • Worker consultation and participation.
Professional regulations (eg, Federal Medical Code for doctors) and labor laws (eg, German Works Constitution Acts)	Defines autonomy and participation rights of HCPs[h].	• Protection of professional independence in decision-making. • Co-determination rights of employee representatives, for example, when adopting systems that monitor behavior and/or performance.

[a]GSPR: general safety and performance requirements.

[b]ISO: International Organization for Standardization.

[c]QMS: quality management system.

[d]IEC: International Electrotechnical Commission.

[e]AI: artificial intelligence.

[f]GPAI: general-purpose artificial intelligence.

[g]LLM: large language model.

[h]HCP: health care professional.

In detail, we describe the appropriate stakeholder circle, the approaches needed for implementing new and highly integrated, localized, and adaptive AI models, and optimal techniques for building consent. While this paper emphasizes that AI systems are increasingly evolving into system-level tools with broad intended purposes, it is nevertheless valuable to explore the development of a narrow-purpose, limited-functionality tool as a simple entry point in the consideration of AI system implementation. This example serves as a foundation for discussing the broader challenges associated with a broad intended purpose and multiagent AI systems. We describe the co-development of an "in-house" AI system [75] that is developed within a health institution to address specific needs [76,77], rather than the implementation of an externally developed "off the shelf" AI system, as this allows more aspects of the collaborative process to be described.

This pragmatic approach was developed in part through in-depth individual consultation and 4 flexible multistakeholder workshops, which are described in more detail in Table 3. By bringing together all the relevant players in the health care ecosystem, we were able to set agreed goals and processes for the development, integration, use, and oversight of health AI. These insights from the workshops informed aspects of the development of the overall framework presented in this viewpoint, alongside the perspective of the authors.

XSL•FO
RenderX

Table 3. Methodological design of the stakeholder workshops. Since workshops are platforms to jointly identify and explore complex domains, and help to gain relevant insights beyond the individual stakeholders' scope of knowledge [78], they offer a valuable basis for a framework that has consensus-building at its core.

Aspect	Approach
Stakeholder definition	• An individual or group who is affected by or can influence the digital transformation in hospitals, particularly with a focus on AIa-enabled systems.
Identification of stakeholders	• Stakeholders were identified using the 7Ps framework [79], which serves as a guide for engaging diverse and relevant interest groups. We modified the categories and definitions of the 7Ps according to our context: • Patients and the public: As this is not a traditional patient-focused study, but rather a practical, expert-driven implementation guide for human-centric digital transformation in a hospital setting, stakeholders were viewed both as domain experts and as potential patients. Additionally, we had feedback from two different international patient representative organizations. • Providers: Individuals who provide care to patients and offer relevant insights from their clinical work were included. The selected clinicians represent various medical fields, including psychology, and are balanced in their seniority and professional position. • Purchasers: Since digital transformation must be financed individually by each hospital, we included stakeholders responsible for the high-level management of digital transformation in large hospitals who manage strategic decisions about cost underwriting based on a specific internal budget. • Payers: In Germany, digital hospital transformation is supported through federal programs. Therefore, we involved stakeholders working at the Federal Ministry of Health and stakeholders who are actively translating those programs into clinical practice. Additionally, we included employees of insurance companies, as insurers play, in general, a critical role in creating patient-centric digital ecosystems and in incentivizing digital health solutions. • Policy makers: Policy makers and supporters of digital transformation in hospitals were included, particularly those who support a human-centric approach while ensuring the rights of HCPs[b] and patients are in place, spanning stakeholders from labor unions to occupational health and safety experts, as well as relevant legal and ethical perspectives. • Product makers: As EURAID[c] highlights the need for a well-balanced stakeholder group developing and implementing AI in health care, the stakeholders representing the "in-house" manufacturers are in their profession AI system developers, psychologists and human-centered AI development professionals, as well as experts in medical device regulation, quality and clinical risk management, medical informatics, and in occupational health and safety at work. • Principal investigators: The researchers included were from a background of clinical AI, medical device regulation, nursing science, medical informatics, digital health, patient safety, psychology, and ethics.
Stakeholder engagement	• Objectives: The goal of stakeholder engagement was to achieve a common agreement on the theme by balancing the differences of individual viewpoints (eg, between calls for greater space for innovation or rather tighter regulation), and developing a framework that all stakeholders agree with. • Methods: Stakeholders were engaged through participatory workshops (three dealt with relevant aspects EURAID should focus on and were initiated by the German Federal Institute for Occupational Safety and Health (BAuA), in 2024 and 2025, with 25, 24, and 17 participants respectively; and one dealt with aspects of HCP integration and current health system problems AI-enabled transformation might solve (Table 1) and was organized by the Else Kröner Fresenius Center for Digital Health in February 2025, with 5 participants). The participating stakeholders spanned the whole 7P categories. Based on this data and a critical review of the literature exploring existing frameworks and gaps, AS and SG developed the concept for the paper and wrote the first draft of EURAID. The stakeholders reviewed the paper, validated its content, and provided further expert insights during a 4-month iterative consensus process.

[a]AI: artificial intelligence.

[b]HCP: health care professional.

[c]EURAID: European Responsible AI Development.

Step 1: Comprehensive and Inclusive Stakeholder Involvement to Build Consensus and Ensure Goal-Oriented Development and Implementation

The selection and active participation of stakeholders and the building of consensus are critical to the success of the AI system development and implementation. The stakeholders involved should be balanced in disciplinarity (clinical, technical, and administrative [80]) and operational responsibilities (professional positions, employee representatives, etc) as well as in age and gender. In Table 4, we highlight the key stakeholders involved, and in particular their role in the implementation process. Each stakeholder is selected for their contribution, ranging from strategic aspects (management board) to safety perspectives (employee representatives, quality management, clinical experts, and users), and data-driven issues (AI system developer, data scientists, and IT and regulatory specialists). In principle, stakeholders in their profession are not mutually exclusive; instead, one could fulfill several roles simultaneously.

XSL•FO
RenderX

Table 4. Key stakeholders and their roles in shaping and guiding AI development and implementation in health care. Each stakeholder is selected for their contribution to the process and expertise.

Stakeholder	Important areas of stakeholder involvement and key aspects they can address
Management Board	The management board sets an overall vision and strategy, leading change management [57,81], and providing investment [82] in staff, hardware, and supporting infrastructure [17]. They foster an institutional culture that tolerates experimentation (and failure) [80], serve as the institution's most credible communicator (ensuring transparency around risks and benefits), and manage external relationships by forging alliances with industry innovators, researchers, professional associations, and policymakers.
Employee Representatives	The foremost priority of employee representatives is to defend and improve working conditions, including occupational safety, workload management, and job security. Although large-scale staff redundancies are unlikely consequences of the near-term implementation of AI[a] in hospital health care systems, which are operating against a backdrop of large staff shortage [83,84], anxiety about automation and transformation of job roles is real [85]. Employee representatives ensure that AI is implemented in a way that eases staff workload and safeguards their well-being and autonomy. In the mid- to long-term, they also negotiate fair compensation policies [86] and career development frameworks that reflect changing roles and skills in an increasingly digital workplace.
AI-System Owner and team[b]	The AI system owner holds primary accountability for the system's performance, safety, and operational impact. They lead the project and ensure alignment with strategic goals and regulatory compliance, while understanding the users' "pain points" both from a clinical and organizational perspective. Their responsibilities include bridging the communication gap between technical and nontechnical language, balancing different perspectives, and developing educational approaches [66] to increase user adoption.
Clinical Experts	Clinical experts identify clinical relevance and utility, which are interpreted and transcribed into a specific scope (intended purpose that specifies clinical indication and initial target group). They provide crucial input to clinical validation and safety, ensuring the AI system integrates effectively into workflows, as well as initiate, oversee, and conduct clinical trial–based AI studies.
AI System Developer	To design and develop machine learning algorithms tailored to specific needs, the AI system developer must integrate and harmonize data from different sources [15]. They also validate the AI model and detect and mitigate model bias to ensure the systems are fair, scalable, adaptable, and verifiable in real-world environments [87].
Users (HCP[c] or patient)	Users with varying levels of digital literacy [57,88] provide real-world, iterative feedback on the system's usability, workflow integration, and perceived value. They often become multipliers for AI adoption, and by their active participation in co-designing educational materials [66], they support evolving digital competence among peers.
Data Scientist	The data scientist safeguards the quality of the data foundation on which the AI system depends during preparation, collection, and checking of the data, for example, by keeping data collection protocols and detecting data imbalance, bias, or outliers across age, sex, gender, race, or ethnicity to prevent disparities and underperformance before they arise [89].
IT Specialists	This role provides the essential technical infrastructure and ensures secure, seamless integration with existing systems, like EHR[d] platforms or laboratory systems, requiring technical, syntactic, semantic, and organizational interoperability [15,90]. Beyond integration, they build and maintain structures for data security, access control, and real-time support, and establish data backup and disaster-recovery systems.
Regulatory Specialists	Regulatory Specialists provide expertise in medical device and AI law, data protection, and human rights. They ensure regulation standards (like the MDR[e] and the AI Regulation) are met throughout the product lifecycle, which is essential to mitigate legal risks and prevent potential breaches.
Notified Body	The role of the Notified Body is to assess whether medical devices meet European legislation, like MDR. This includes determining the correct classification, evaluating legal compliance, and reviewing technical documentation [75,91]. The Notified Body only has a direct role where a CE[f]-mark is sought for medium or high-risk AI systems.
Quality Management	Quality management ensures continuous patient safety by monitoring and measuring performance, outcomes, and the integrity of clinical workflows [87]. They establish comprehensive risk management systems (eg, handling device failures or malfunctions) [87] and drive standardization. This role also promotes safe system use by co-designing educational programs [66] for both HCP and patients.

[a]AI: artificial intelligence.

[b]Role of the stakeholders whose input is coordinated through the AI-System Owner.

[c]HCP: health care professional.

XSL•FO
RenderX

[d]EHR: electronic health record.

[e]MDR: Medical Device Regulation (2017/745).

[f]CE: Conformité Européenne.

An interactive environment, with all critical stakeholder groups adequately represented, enables and encourages the integration of stakeholder insights and experiential learnings, while promoting careful consideration of how AI systems are best built to be suited to clinical workflows, as well as where existing workflows may need to be modified to adapt to the AI system. This does not mean that every stakeholder group is involved in every decision and has an equal say in the progress of digitalization. Creating this impression could lead to disillusionment and eroded trust in digitalization, and would probably slow down the whole process. Each stakeholder group is involved in some part of the process, with their precise stages of involvement and roles depending on their potential contribution to the process, and it is essential that each stakeholder is aware of the degree of their involvement.

A crucial success factor alongside the development and implementation is the role of the "product owner," who takes the coordinating lead. As in-house development in health care institutions often does not have a commercial development focus, we use the term "AI-System Owner" to denote the "product owner." Although the title may vary by organization, this role usually combines both the entire lifecycle product ownership responsibilities and the domain expertise in health care and AI. The absence of a single person taking responsibility for the development and performance of the system will generally result in a range of negative consequences, such as poor stakeholder communication, a lack of clear vision, scope, and prioritization, and other issues, as real-world examples [92] have shown. We therefore highlight the AI-System Owner as a central stakeholder leading a team of other stakeholders (Figure 2).

Figure 2. The (ongoing) product development in a dynamic team led by the AI-System Owner. The AI-System Owner fulfils a crucial role as he is leading a core team of relevant stakeholders during the process of development and implementation. In a hospital setting, team members will often fulfill several roles simultaneously. AI: artificial intelligence; HCP: health care professional.

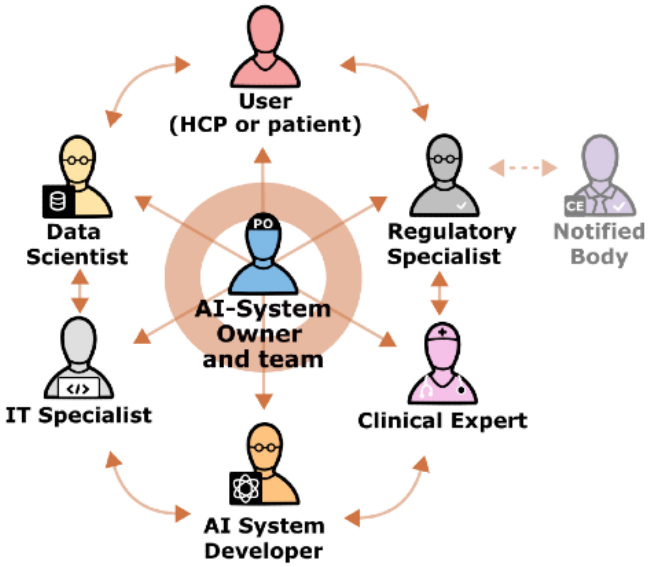

Step 2: Agreement on the Overall Goals and "Device" Purpose

The collaborative and effective implementation of an AI system into clinical workflows starts with a collective agreement on the goals of the implementation, for example, using methods such as SMART (specific, measurable, attainable, relevant, and time-bound), particularly the specific user (generally an HCP or patient) whose needs the system is intended to address. These identified needs are then interpreted and transcribed into a specific scope of the device, known as the "Intended Purpose," which specifies the clinical indication, how the system addresses this clinical indication, and the (initial) target group needs.

Although the regulations for AI-system design and implementation do not formally require the direct involvement of any other health care system actors than the "user" of the AI and its "deployer" (in a broad sense), we argue that the sustainable and beneficial implementation of AI systems needs early and proportional agreement on goals and input from all stakeholders. This includes discussion between the management board, employee representatives, quality management, and the AI-System Owner and their team (Figure 3). Later product development steps require feedback between the AI-System Owner team (including clinical experts and the users of the system), and selected stakeholders (as shown in Figure 3), with management "checkpoints" periodically to ensure that the development of the AI-system is following the initially agreed plan for the AI system. Given the complexity of multistakeholder involvement, it is useful to have a set of rules for working together at the beginning, and to repeatedly build consent along the AI development life cycle, for which we highlight techniques in Figure 3.

XSL•FO
RenderX

Figure 3. Stakeholder interaction and consent-building across the AI life cycle. The figure describes the co-development of an "in-house" AI system, ie, one that is developed within a health care organization to address specific needs. During the development phases, which build on each other and can be repeated, different groups of stakeholders interact to improve the AI system by providing feedback and optimizing the system's adaptation to health care professional workflows. Building consent with a range of different stakeholders with varying levels of experience and backgrounds is not easy. We therefore highlight techniques for building consent at each stage of development to ensure an efficient and safe process that is in line with European values and regulations. AI: artificial intelligence; MDR: Medical Device Regulation; ML: machine learning; PCCP: predetermined change control plan; RWP: real-world performance; SaMD: software as a medical device.

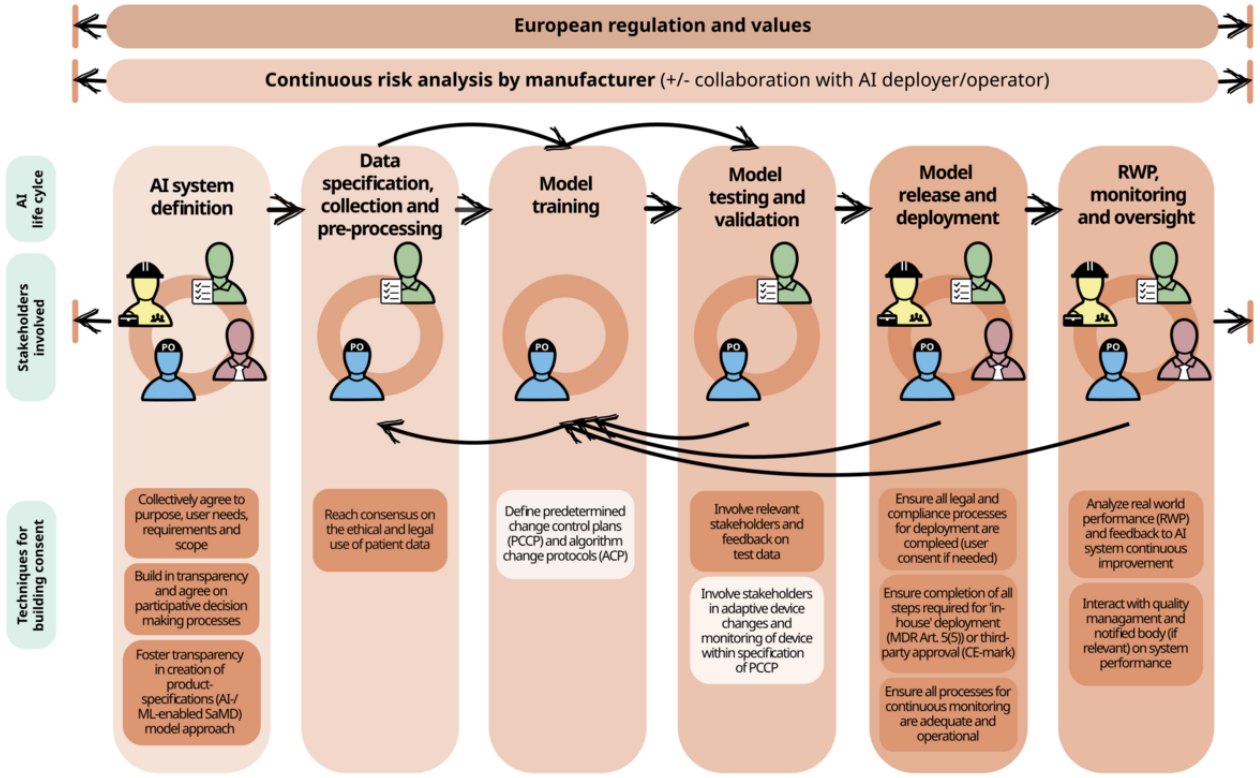

Step 3: AI System Development "In-House"

While generally medical devices must undergo a conformity assessment procedure and must be marked with a CE (Conformité Européenne)-mark before being used, the European Union (EU) exempts certain devices from this general obligation and allows individual health institutions to develop and use "in-house" medical AI systems involved in the diagnosis or therapy of disease without the obligation to conduct a conformity assessment procedure, as long as safety standards and those for quality management are in place. Based on Article 5(5) of the EU Medical Device Regulation (MDR; 2017/745) [75], this exemption applies only for in-house use on a nonindustrial scale and if the needs of the targeted patient groups cannot be met through available and equivalent devices on the market [75,93]. Also covered is the in-house combination or modification of existing systems or devices [93,94]. For example, in Table 5, we have outlined 3 practical examples of AI systems, which have been developed in-house in a German hospital setting, each of them with a unique intended purpose, clinical indication, and target group. We highlight for each the technical approach used as well as the stakeholders included during development and potential prospective trial designs.

XSL•FO
RenderX

Table 5. Practical examples of AI applications developed in-house and their stakeholder integration. The AI use cases presented originate from the SmartHospital.NRW [95] research project, funded by the Federal State of North Rhine-Westphalia, Germany. The project is limited to research and development activities; therefore, the use cases are confined to the development stage. Clinical testing and product commercialization are explicitly beyond the project's scope.

Use case	Automated discharge summary	AI[a]-powered voice assistant for bedside patient support	AI-supported prevention of adverse events
Intended purpose	Automates and optimizes the creation of discharge letters within hospital workflows to reduce clinician workload and improve communication regarding patient care.	Enables patients at the bedside to interact via natural speech, facilitating access to medication schedules, personal calendars, diary management, and support to overcome language barriers through oral translation and simplified language.	Focuses on early and reliable detection of nursing-relevant risks by enhancing existing risk models based on structured nursing assessments and integrating LLMs[b] to analyze clinical progress notes and identify patient-specific risk factors.
Clinical indication	Addresses the challenge of time-intensive medical documentation, particularly discharge summaries following inpatient stays.	Designed for patients requiring accessible communication support, especially those experiencing language barriers, vision impairments, or limited mobility, while promoting autonomy without providing direct medical advice.	Designed to support systematic, early identification of nursing-related risks, including falls, pressure ulcers, and malnutrition, augmenting safety and enabling individualized care planning.
Target group	Primarily, hospital physicians with indirect patient benefits, such as improved continuity of care and efficient information transfer to general practitioners.	Hospitalized patients who require assistance in accessing information and communicating effectively.	Nursing staff responsible for patient care and hospitalized patients are actively involved in care processes.
Technical approach	Uses generative AI language models interfaced with hospital information systems to autonomously extract structured clinical data and generate contextually relevant text suggestions for documentation.	Uses on-premises LLMs within dedicated patient devices; enables localized processing of voice input streams independent of hospital system integration, thereby preserving data sovereignty.	Integrates structured clinical data, unstructured data derived from speech-to-text conversion of nursing assessments, and patient-reported outcomes to facilitate comprehensive risk detection.
Stakeholders included during development	Management Board, AI System Developer, AI-System Owner, IT Specialists, Clinical Experts, and Users.	Management Board, AI System Developer, IT Specialists, Clinical Experts, and Users.	Management Board, AI System Developer, AI-System Owner, IT Specialists, Clinical Experts, and Users.
Experience of development	Developed iteratively as a prototype, validated with real clinical data, while ensuring compliance with regulatory, privacy, and interoperability standards.	Followed an iterative development approach with thorough curation of informational content; faced technical challenges such as limited server access before full deployment of open-source models.	Development prioritized screening instruments to assess signs and symptoms of nursing care, optimization of AI risk detection models, and ensuring data privacy using pseudonymization and anonymization techniques.
Potential prospective trial designs	Cluster-randomized controlled trial at the ward level, comparing standard discharge processes versus AI-assisted summaries. Primary endpoints: clinician documentation time and report quality (as judged by independent review).	Patient-level crossover trial with and without AI voice assistant support. Main outcomes: patient autonomy, effectiveness of information access, and user satisfaction, controlling for intrapatient variability.	Pragmatic controlled trial in clinical wards comparing standard care with and without AI-based risk detection algorithms. Outcomes: incidence of adverse events (falls, pressure ulcers, and malnutrition), timeliness of risk identification, and changes in clinical workflow.

[a]AI: artificial intelligence.

[b]LLM: large language model.

In contrast to commercial deployments, in-house systems offer a distinctive opportunity for embedding participatory ethics, iterative design cycles, and real-world validation and feedback loops directly into the lifecycle of medical AI. This allows the creation of a highly customized solution that fits in location-specific clinical workflows and staff practices, which can be extended to multiple systems within the same platform and institution [76]. Moreover, a key advantage is the use of the hospital's own data; however, this requires a well-developed data infrastructure and processes for obtaining patient consent. Considerations include interoperability and data preparation, such as labeling (although label-free approaches are becoming more common), structuring, and collection (requirements also under the AI Regulation), in order to know which data can be used for a specific solution.

Step 4: AI-System Testing, Validation, and Clinical Evaluation

Health care AI demands rigorous, multidimensional evaluation that must encompass not only technical performance, but also clinical integration, and verify safety, usability, ethical robustness, and regulatory compliance.

XSL•FO
RenderX

Independent assessment of device performance can be generated through statistically sound test plans, which generate information separate from the training data set [96]. Since validation in real-world settings is still a bottleneck [97], prospective, noninterventional silent trials [98,99] (where AI is tested within the clinical pathway in real time without affecting patients) can enhance transparency and facilitate informed deployment decisions. For large language models (LLMs) and, in particular, adaptive AI models that evolve over time, continuous validation frameworks are needed [100]. Recent studies have highlighted, substantial challenges to the reliability and safety of LLMs in health care persist, including hallucinations [101], metacognitive deficiencies [102], vulnerability to bias [103] and data-poisoning [104], and problems in integration in existing workflows [105], making single evaluation dimensions insufficient. Therefore, multidimensional methods could help to operationalize feasibility, score diagnostic accuracy or unsafe recommendations, and detect bias and usability issues. Examples are "QUEST" [106] to score outputs, or agentic-based simulations such as "CRAFT-MD" [107] for clinical workflow evaluation. Alignment with international AI standards (eg, ISO/IEC [International Organization for Standardization/ International Electrotechnical Commission] 42001:2023 [108], FG-AI4H [Focus Group on AI for Health] clinical evaluation framework [109]) further strengthens interoperability and safety.

Beyond objective data and algorithm quality, subjective feedback from users is essential [57,110]. Evaluations should capture how AI systems integrate into existing workflows and routines, their ease of use, and their perceived performance and interface design. Researchers highlighted several approaches for evaluation, such as through integrated feedback systems [110,111] or through organizational internalization by creating an "AI-QI"-unit responsible for quality improvement and assurance [87], interacting as a "glue" between different entities.

Evaluation should follow a risk-tiered approach that links the level of regulatory and ethical scrutiny to the severity of the health decision involved (Figure 4). For instance, AI systems used for administrative optimization or appointment scheduling may require a lower level of risk mitigation, while those supporting diagnostic or therapeutic decisions demand significantly higher safeguards. This tiering can draw on the EU AI Act's risk classes and MDR risk classifications, and should be developed in consensus with relevant stakeholders, including clinical risk management and regulatory specialists.

XSL•FO
RenderX

Figure 4. Risk-based tiering of safeguards. With a proportional approach to regulatory and ethical safeguards aligned with the severity of the health decisions affected by an AI system, this provides a useful link between risk classification (eg, under Medical Device Regulation or the EU AI Act) and the required level of human oversight, transparency, and stakeholder involvement. AI: artificial intelligence; EHR: electronic health record; EU: European Union; LLM: large language model.

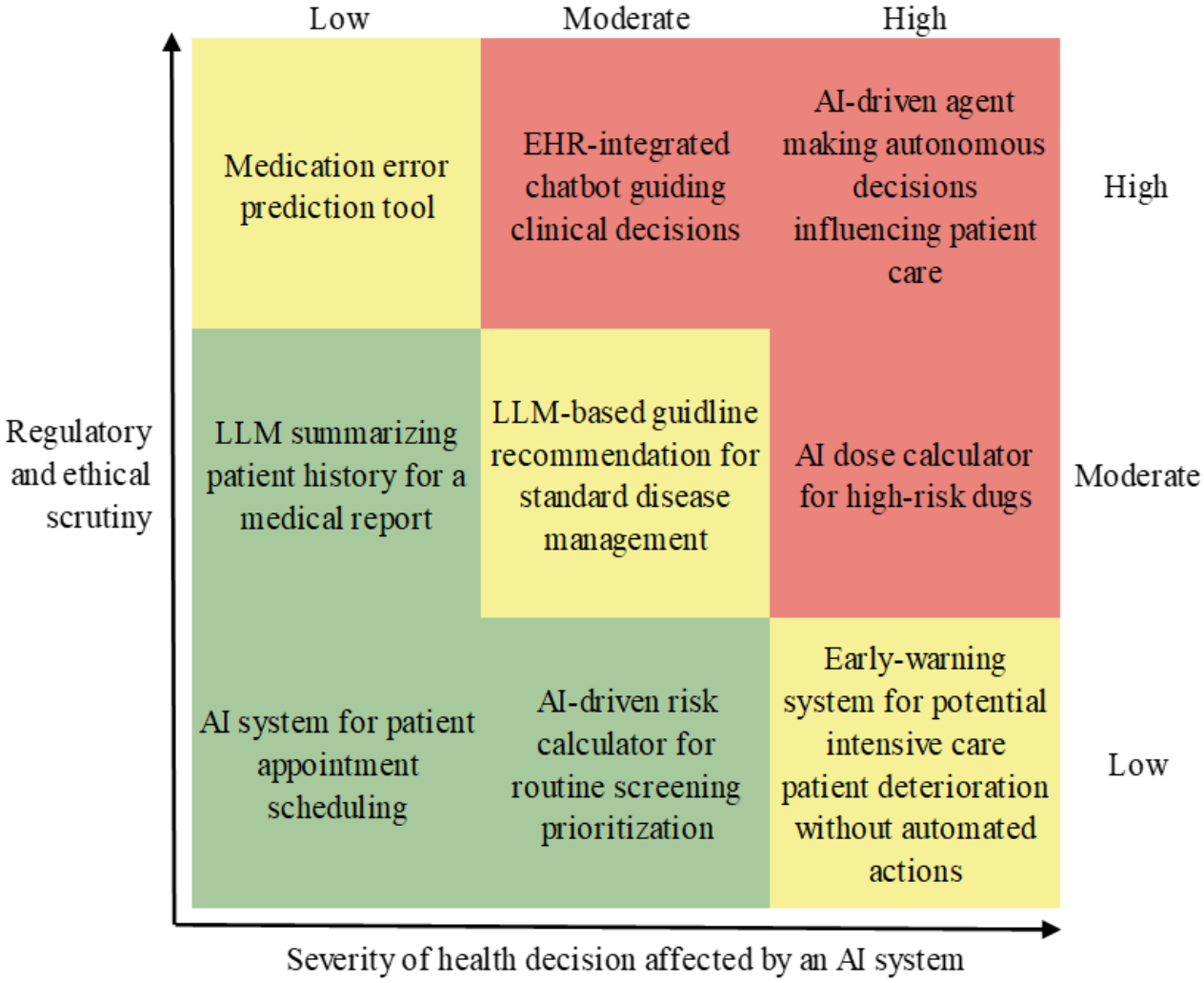

To ensure that the AI system is compatible with European values, ethics-based auditing frameworks like capAI, grounded in the EU AI Act, can guide risk identification in each phase of the AI lifecycle from an ethical point of view [112]. The integration of tools like the self-assessment list for trustworthy AI (ALTAI) [113], developed by the EU High-Level Expert Group on AI, into ethics-based auditing of AI systems can further support responsible usage of AI and foster user trust. Yet, ethical guidelines are just that: guidelines. They rarely or incompletely answer concrete ethical questions regarding the use of an AI system in a specific situation, such as the question of specific moral responsibility if mistakes of AI systems lead to patient harm. This is a highly discussed topic in ethics [114] and becomes even more severe in the context of black-box problems, eventually leading to moral responsibility gaps [115]. Other still unsolved ethical questions occur, for example, regarding data ownership in the context of the principle of beneficence (ie, promoting others' benefit and preventing harm [116,117]) and informed consent [118] or anthropomorphization of AI [119]. Therefore, embedding ethical points of view into the whole life cycle of AI is necessary [120].

Step 5: Development and Deployment of Training Approaches

The successful adoption of AI by hospital employees correlates with continuous development and training [88]. Although training is also a requirement of the EU AI Act [121], it is of note that only 24% of the health care institutions provide AI training programs and workshops [122]. This underscores a gap in education and certification, leaving clinicians without the necessary tools to harness the full potential of AI. However, there are various ways to support confidence in AI technologies among HCPs. For example, (1) by investing in comprehensive training programs that help to gain necessary skills [88] while also extending existing programs with AI literacy, or (2) by developing and provisioning resources and mechanisms to build and strengthen connections among peers and innovators to share their AI-related knowledge and experiences [80]. And more importantly, AI training should be a fixed part of any professional education and competency assessment, as well as included in further training (eg, through integration into Continual Medical Education programs) [123] to build confidence in its use among the next generation of HCP and achieve a symbiotic relationship between humans and AI [124].

XSL•FO
RenderX

In order to build AI literacy among HCP in a safe and controlled environment, training methods such as simulation-based modules [125,126] (ie, practice in realistic settings [125,127]), case-based exercises [128], and interactive workshops [129] can help to explore tools repeatedly without risking patient safety while facilitating experimental learning. Another method of providing HCPs with hands-on experience using AI tools in a controlled environment is to conduct a pilot phase, during which AI is tested by selected clinical users in a narrow area of practice, or shadow deployment, in which AI operates in shadow mode alongside clinicians in real time and is guided by predefined safety and workflow indicators [130]. This will also influence trust and adoption among users and foster psychological safety, since evidence from human-computer interaction research indicates that a positive attitude toward AI is not only a function of system transparency or explainability, but also depends on users' self-efficacy, previous experience, and the perceived fairness and predictability of the system [131]. With regard to content, it is important to define responsibilities within the company regarding who will take ownership of training the users in basic competencies of AI literacy. The AI-System Owner and his or her team would be the best fit, as they combine the entire use case-relevant expertise through different perspectives, ranging from clinical experts to system developers.

Training should foster understanding of AI systems and facilitate interaction and use of AI systems, and is relevant not just for direct users, but for all HCPs who will work alongside care systems influenced by AI (Figure 5) [15,66]. Key competencies are a basic understanding of when and how to use AI, knowledge about the use of the systems' elements, the ability to make informed decisions based on a risk-benefit analysis, the awareness of legal and ethical considerations, and, to adapt to new tools and applications [123,132]. Components of health care AI training that are generic do not need to be developed de novo by the health institution. However, specific training directly related to the AI-system to be deployed will generally be required, and it is often necessary to provide ongoing training which takes account of the learning curve of the HCP in the use of the AI, emergent problems such as automation bias [133] and deskilling [134], and changes and further development of the AI-systems.

Figure 5. The learning curve of the health care professional (HCP) in the use of artificial intelligence (AI) systems in health care. After training in the basic AI principles and their use, as well as health care–specific guidelines for AI integration, on the first day of the system's clinical use, the HCP should be trained in the operational knowledge of the system being deployed. The HCP will then develop their skills through experience in their use.

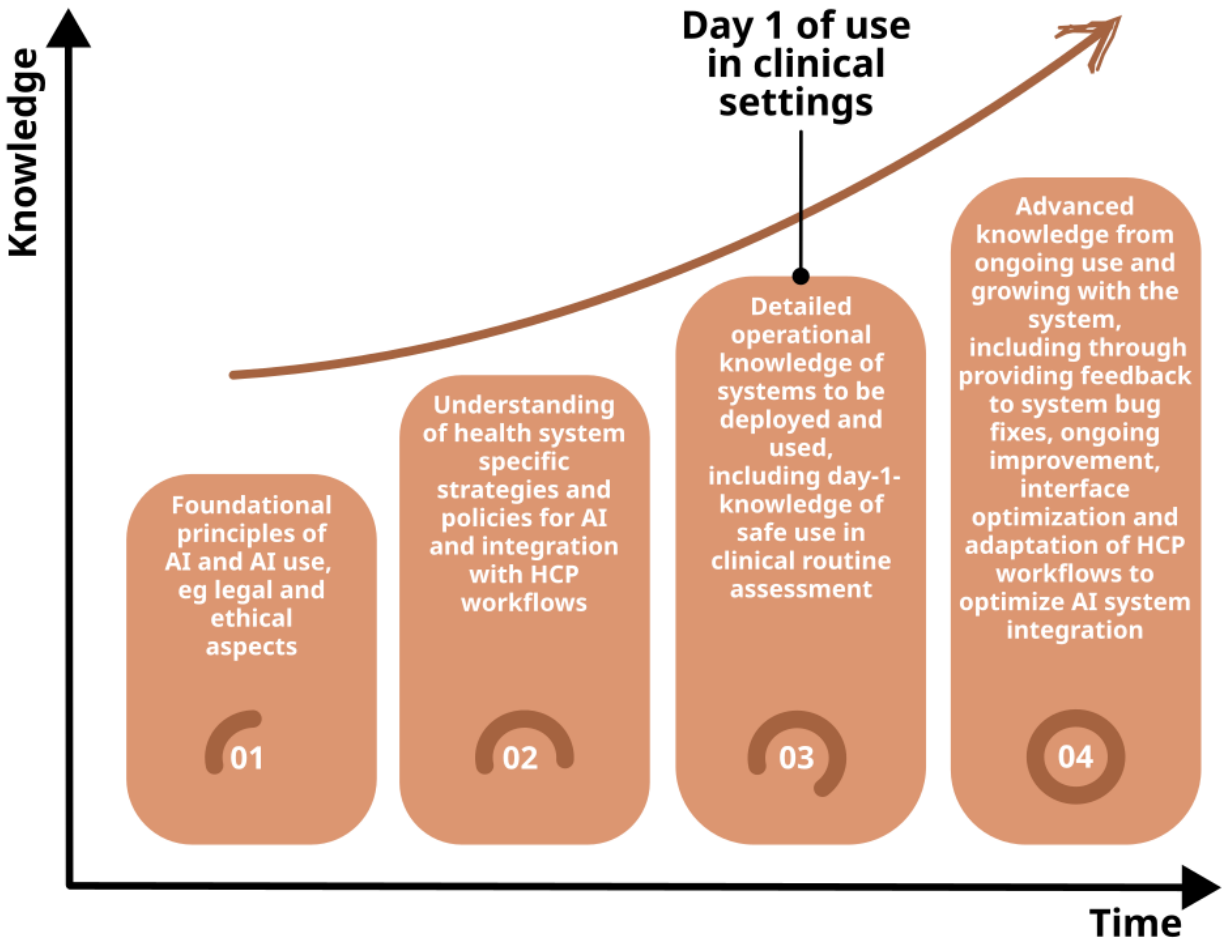

XSL•FO
RenderX

Step 6: AI-System Deployment, Real-World Performance Monitoring, and Later Decommissioning

After model creation and testing, the goal is to place the system in real-world clinical settings to improve patient care and outcomes [135] according to the previously defined overall goals and device purpose. This needs transparency, and compliance with legal and ethical processes (user consent), as well as the completion of all steps required for the exemption to conduct a conformity assessment under in-house deployment ("MDR Article 5(5)") or third-party approval (CE-mark). Therefore, looping in all stakeholders is needed to collaboratively address associated challenges. A key role is played by the management board and AI-System Owner to provide a clear external and internal communication that signals the prioritization of human well-being during the whole process, and users as multipliers to promote trust for broad widespread acceptance and use.

Involving all stakeholders also applies to monitoring and oversight of real-world performance, as it needs constant feedback from different perspectives to improve system performance and data-related processes. The goal of monitoring is to raise an alarm when unintended or special cases occur [87], which emphasizes the importance of finding solutions through collaboration and collective intelligence. The "AI-QI" unit described above could consolidate and strengthen the established stakeholder structure within the company long term. In addition, algorithmic audits can serve as a framework for continuously monitoring AI systems and understanding errors, how and why these adverse events occurred, while anticipating their potential consequences [136]. Real-world performance monitoring must adequately account for model drift (degradation of AI system performance over time) due to changes in external factors such as patient populations, data collection, or medical practice [137].

Running a "legacy system" usually means facing layers of technical debt, which slows down development and complicates maintenance, and leads to several risks, such as the technology becoming less reliable and decreasing in performance, or exposing systems to vulnerabilities such as cyberattacks. However, decommissioning can be an option to abstract and secure data in a newer system [138]. This process needs to be carried out by IT and regulatory specialists, as well as data scientists and quality management, in consultation with users, the management board, and employee representatives, and notified bodies where required.

Special Considerations for Adaptive, "Agentic" and "off-the-Shelf" AI Systems

Some recent AI approaches are developed so that they learn and adapt from data and feedback from the real world, allowing them to change continuously without explicit interventions from the developer [139,140]. Ensuring such systems are safe, effective, and of high quality while being flexible requires a more interactive and participatory approach than traditional systems that follow static and predefined rules. This is especially true when self-learning systems are combined with agentic AI systems that are able to handle multilevel tasks, coordinate tools, centralize human communication, and basically act as health care teammates [26-29]. Autonomous AI systems and LLM-enabled clinical decision systems have already been approved in Europe [30,141,142]. As the approval and use increase, and as these systems continuously encounter new settings and tasks, it is essential to define clear boundaries, controlled environments with clinician oversight [27], ongoing auditing [26], and adequate training capacities for HCPs [27]. As broad models may be applied across multiple hospital departments and clinical contexts (eg, simultaneously used in an emergency department and psychiatry clinic) with dynamic or variable workflow integration, transparent communication, and iterative feedback across stakeholders (as presented in this paper) are also critical to ensure adaptability and to address the more complex ethical, legal, and social implications.

For off-the-shelf AI systems provided by external companies, the interaction between stakeholders should be focused on integration, compliance, and validation to meet operational and regulatory needs. These systems may limit the level of innovation achievable (no bottom-up activism from internal users and developers to continually contribute improvements and features that better meet unique requirements) and may lead to trust issues due to less transparency in the handling of data and underlying algorithms [14], requiring proactive communication and change management. Responsibilities for monitoring and model updating, especially with proprietary algorithms, become more complex and need to be clarified between external collaborators and internal stakeholders [87]. Platforms for delivering off-the-shelf AI systems now allow the co-hosting of in-house developed AI models, alongside the CE-marked models, enabling both approaches to coexist, and making clear the need and possibilities for the co-design, embedding, and co-implementation of commercial and in-house approaches [143].

Discussion

Studies show a persistent gap between research and clinical implementation [144,145], with medical AI adoption still very slow [144,146] and limited to a few use cases [147]. Reasons include the difficulty of aligning diverse stakeholder perspectives within complex health care systems, the rigidity of regulatory frameworks, and the limited consideration of design approaches of work and organizational psychology [148]. As a result, achieving both technological effectiveness, in the sense of medical accuracy and system performance, and user acceptance among HCP and patients is often perceived as conflicting goals.

A balance is therefore needed between ensuring safety and enabling innovation [149]. EURAID finds this "sweet spot," accelerating digital transformation in a human-centric way. Unlike existing frameworks, which focus narrowly on user perspectives [80,150,151], isolated implementation aspects [150,152-155] (such as evaluation, safety, or ethics), serve as decision support tool for choosing the most fitting available AI solution [156], or have a limited clinical scope [157-160], EURAID explicitly maps all key stakeholders across the AI development life cycle, clarifies their roles and key aspects they

XSL•FO
RenderX

can address (Table 4) in co-creating, guiding, and governing "in-house" AI development and deployment. It also details stakeholder roles in real use cases, and methods for achieving iterative consensus at each development stage across disciplines that reflect shared goals in alignment with European values, and strengthening the understanding of training methods, content, and key competencies.

However, EURAID has some limitations. The resources or specialized staff needed for iterative development and testing are more limited in smaller hospitals, necessitating concentrating multiple roles on fewer people, which can lead to a shortage of expertise, but, on the other hand, may also speed up processes. Although our approach can likely better address creative problem-solving, traditional, rigid, and hierarchical structures common in health care may hinder stakeholder selection based on their contributions and expertise rather than their positions and level of seniority. Although "in-house" AI devices may not require CE marking, they are not exempt from regulation and have legal liability implications. Health institutions must comply with a number of obligations that may discourage them from doing it at all, which slows down both innovation and digitalization. A practical solution is to designate key staff for legal or ethical liaison roles or establish a multidisciplinary AI advisory board and data governance council within the institution to ensure compliance and continuity.

Conclusions

EURAID is a pragmatic, solution-oriented framework, compatible with European values and regulations, and ensures that barriers to "in-house" AI development and implementation in hospitals are acknowledged early and resolved through collaborative problem-solving. The underlying principle is that the likely future of medicine, driven by integrated, localized, and adaptive AI technologies, will need all critical stakeholders (which we portray individually in this paper) adequately represented, and their various perspectives embedded in the co-design, procurement, implementation, and oversight of AI systems, ensuring that digital transformation in health care truly benefits the people who will use them every day. Additionally, as AI systems used vary by type and clinical setting, we propose a risk-tiered approach that provides a useful link between risk classification and the required level of human oversight, transparency, and stakeholder involvement.

To translate EURAID into action, hospitals should begin by conducting internal readiness assessments, establishing cross-functional AI governance structures, and defining clear, role-specific responsibilities for ethical, legal, technical, and clinical oversight. Regulators and professional bodies should, in parallel, create structures that connect local innovation with next-generation European legislation, for governance that is as intelligent as the technology built.

Acknowledgments

We acknowledge the use of the ChatGPT language model (GPT-3.5, GPT-4, and GPT-5; OpenAI) for assisting in refining some text of this paper. Responsibility for the final manuscript lies entirely with the authors. The graphical elements in this paper were designed using Inkscape.

Funding

This work was supported by the European Commission under the Horizon Europe Program, as part of the project ASSESS-DHT (101137347) via funding to SG and RM. The views and opinions expressed herein are, however, the authors' responsibility only, and do not necessarily reflect those of the European Union, the United Kingdom, the European Health and Digital Executive Agency (HaDEA), UK Research and Innovation (UKRI), or the National Institute for Health and Care Excellence (NICE); the European Union, United Kingdom, and granting authorities cannot be held responsible for the views, opinions, and information contained herein.

Authors' Contributions

AS and SG developed the concept of the study. AS and SG wrote the first draft of the paper. AS, MEG, MHG, FJK, JNK, EK, TL, EL, ML, RM, HSM, JO, TR, UR, M Schneider, LS, HS, MLS, NS, M Sedlmayr, RS, BS, MKW, EW, KW, AD, and SG contributed to the writing, interpretation of the content, and editing of the paper, revising it critically for important intellectual content. AS, MEG, MHG, FJK, JNK, EK, TL, EL, ML, RM, HSM, JO, TR, UR, M Schneider, LS, HS, MLS, NS, M Sedlmayr, RS, BS, MKW, EW, KW, AD, and SG had final approval of the completed version. AS, MEG, MHG, FJK, JNK, EK, TL, EL, ML, RM, HSM, JO, TR, UR, M Schneider, LS, HS, MLS, NS, M Sedlmayr, RS, BS, MKW, EW, KW, AD, and SG take accountability for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

The authors expertise ranges from medical device regulation (AS, SG, RM, and ML), to high-level management of digital transformation in big hospitals (AD), and includes experts in quality and clinical risk management (RS, MEG, and SG), medical informatics (ML, M Schneider, SG, MLS, and M Sedlmayr) and occupational health and safety at work (UR, LS, and TL), as well as relevant insights from clinical experts and HCP (EW, JNK, HSM, NS, JO, and AD), AI system developers (JNK, NS, and JO) and expertise in psychology and human-centered AI development (MKW, KW, and TL). In addition, we included relevant legal (EK), ethical (EL), and federal policy (MLS) perspectives, health and social accident insurance companies (TL and M Schneider), labor unions (BS), and from academia (MHG, HS, TR, FJK, AS, RM, HSM, SG, JNK, and MKW).

XSL•FO
RenderX

Conflicts of Interest

SG declares a nonfinancial interest as an Advisory Group member of the EY-coordinated "Study on Regulatory Governance and Innovation in the field of Medical Devices" conducted on behalf of the Directorate-General for Health and Food Safety (SANTE) of the European Commission. He declares the following competing financial interests: SG has or has had consulting relationships with Una Health GmbH, Lindus Health Ltd, Flo Ltd, ICURA ApS, Rock Health Inc, Thymia Ltd, FORUM Institut für Management GmbH, High-Tech Gründerfonds Management GmbH, Directorate-General for Research and Innovation of the European Commission, and Ada Health GmbH, and holds share options in Ada Health GmbH. JNK declares consulting services for Bioptimus, France; Panakeia, UK; AstraZeneca, UK; and MultiplexDx, Slovakia. Furthermore, he holds shares in StratifAI, Germany, Synagen, Germany, and Ignition Lab, Germany; has received an institutional research grant from GSK; and has received honoraria from AstraZeneca, Bayer, Daiichi Sankyo, Eisai, Janssen, Merck, MSD, BMS, Roche, Pfizer, and Fresenius. JO has received travel grants from Abbott and research grants from German Heart Foundation (DSHF), German Center for Cardiovascular Research (DZHK), the University of Hamburg (UHH), and the German Federal Ministry of Education and Research (BMBF), and is co-founder and former managing director of IDM GmbH. MLS reports no conflicts of interest. The opinions expressed in this article are his own and do not necessarily reflect the views held by the German Federal Ministry of Health. None declared by the other authors.

References

1. Blum K. California nurses protest 'untested' AI as it proliferates in health care. Association of Health Care Journalists. URL: https://healthjournalism.org/blog/2024/08/california-nurses-protest-untested-ai-as-it-proliferates-in-health-care/ [accessed 2024-08-09]
2. Bruce G. Nurses protest AI at Kaiser Permanente. Becker's Health IT. URL: https://www.beckershospitalreview.com/healthcare-information-technology/nurses-protest-ai-at-kaiser-permanente/ [accessed 2024-04-22]
3. Blease CR, Locher C, Gaab J, Hägglund M, Mandl KD. Generative artificial intelligence in primary care: an online survey of UK general practitioners. BMJ Health Care Inform 2024;31(1):e101102 [FREE Full text] [doi: 10.1136/bmjhci-2024-101102] [Medline: 39288998]
4. Fernandopulle R. We must stop trying to deliver 21st-century care with a 19th-century delivery model. MedGenMed 2005;7(2):50 [FREE Full text] [Medline: 16369428]
5. Kennedy PJ. Our health system is built on an antiquated model of care. The Hill. 2020 Aug 25. URL: https://thehill.com/opinion/healthcare/513615-our-health-system-is-built-on-an-antiquated-model-of-care/ [accessed 2025-04-03]
6. Mele M. Antiquated methods put patients at risk. Beckers's Clinical Leadership. URL: https://www.beckershospitalreview.com/quality/antiquated-methods-put-patients-at-risk/ [accessed 2019-03-14]
7. Mauro M, Noto G, Prenestini A, Sarto F. Digital transformation in healthcare: assessing the role of digital technologies for managerial support processes. Technol Forecast Soc Change 2024;209:123781. [doi: 10.1016/j.techfore.2024.123781]
8. Marques ICP, Ferreira JJM. Digital transformation in the area of health: systematic review of 45 years of evolution. Health Technol 2019;10(3):575-586. [doi: 10.1007/s12553-019-00402-8]
9. Barbieri C, Neri L, Stuard S, Mari F, Martín-Guerrero JD. From electronic health records to clinical management systems: how the digital transformation can support healthcare services. Clin Kidney J 2023;16(11):1878-1884 [FREE Full text] [doi: 10.1093/ckj/sfad168] [Medline: 37915897]
10. Mulukuntla S, Pamulaparthyvenkata S. Digital transformation in healthcare: assessing the impact on patient care and safety. Int J Med Health Sci 2020;6(3) [FREE Full text]
11. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023;23(1):689 [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
12. Otero-García L, Mateos JT, Esperato A, Llubes-Arrià L, Regulez-Campo V, Muntaner C, et al. Austerity measures and underfunding of the Spanish health system during the COVID-19 pandemic-perception of healthcare staff in Spain. Int J Environ Res Public Health 2023;20(3):2594 [FREE Full text] [doi: 10.3390/ijerph20032594] [Medline: 36767958]
13. MOSCIARO M, KAIKA M, ENGELEN E. Financializing healthcare and infrastructures of social reproduction: How to bankrupt a hospital and be unprepared for a pandemic. J Soc Policy 2022;53(2):261-279. [doi: 10.1017/s004727942200023x]
14. Dennstädt F, Hastings J, Putora PM, Schmerder M, Cihoric N. Implementing large language models in healthcare while balancing control, collaboration, costs and security. NPJ Digit Med 2025;8(1):143 [FREE Full text] [doi: 10.1038/s41746-025-01476-7] [Medline: 40050366]
15. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, Martinez EZ, Zucoloto ML, Østengaard L, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023;6(1):161 [FREE Full text] [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
16. Rane N, Choudhary S, Rane J. Acceptance of artificial intelligence: key factors, challenges, and implementation strategies. SSRN Electron J 2024:19. [doi: 10.2139/ssrn.4842167]
17. Karpathakis K, Morley J, Floridi L. A justifiable investment in AI for healthcare: aligning ambition with reality. SSRN Electron J 2024. [doi: 10.2139/ssrn.4795198]

XSL•FO
RenderX

18. Artificial intelligence in healthcare. European Commission. URL: https://health.ec.europa.eu/ehealth-digital-health-and-care/artificial-intelligence-healthcare_en [accessed 2025-12-20]
19. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, et al. Towards accurate differential diagnosis with large language models. Nature 2025;642(8067):451-457 [FREE Full text] [doi: 10.1038/s41586-025-08869-4] [Medline: 40205049]
20. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature 2025;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
21. Anderson BJ, Zia ul Haq M, Zhu Y, Hornback A, Cowan AD, Mott M, et al. Development and evaluation of a model to manage patient portal messages. NEJM AI 2025;2(3) [FREE Full text] [doi: 10.1056/aioa2400354]
22. Hassan H, Zipursky AR, Rabbani N, You JG, Tse G, Orenstein E, et al. Clinical implementation of artificial intelligence scribes in health care: a systematic review. Appl Clin Inform 2025;16(4):1121-1135 [FREE Full text] [doi: 10.1055/a-2597-2017] [Medline: 40306686]
23. Olson KD, Meeker D, Troup M, Barker TD, Nguyen VH, Manders JB, et al. Use of ambient AI scribes to reduce administrative burden and professional burnout. JAMA Netw Open 2025;8(10):e2534976 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.34976] [Medline: 41037268]
24. Chatzikou M, Latsou D, Apostolidis G, Billis A, Charisis V, Rigas ES, et al. Economic evaluation of artificially intelligent (AI) diagnostic systems: Cost consequence analysis of clinician-friendly interpretable computer-aided diagnosis (ICADX) tested in cardiology, obstetrics, and gastroenterology, from the HosmartAI horizon 2020 project. Healthcare (Basel) 2025;13(14):1661 [FREE Full text] [doi: 10.3390/healthcare13141661] [Medline: 40724686]
25. El Arab RA, Al Moosa OA. Systematic review of cost effectiveness and budget impact of artificial intelligence in healthcare. NPJ Digit Med 2025;8(1):548 [FREE Full text] [doi: 10.1038/s41746-025-01722-y] [Medline: 40858882]
26. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature 2023;616(7956):259-265. [doi: 10.1038/s41586-023-05881-4] [Medline: 37045921]
27. Zou J, Topol EJ. The rise of agentic AI teammates in medicine. The Lancet 2025;405(10477):457. [doi: 10.1016/s0140-6736(25)00202-8]
28. Moritz M, Topol E, Rajpurkar P. Coordinated AI agents for advancing healthcare. Nat Biomed Eng 2025;9(4):432-438. [doi: 10.1038/s41551-025-01363-2] [Medline: 40169759]
29. Qiu J, Lam K, Li G, Acharya A, Wong TY, Darzi A, et al. LLM-based agentic systems in medicine and healthcare. Nat Mach Intell 2024;6(12):1418-1420. [doi: 10.1038/s42256-024-00944-1]
30. DERM makes medical history as world's first autonomous skin cancer detection system is approved for clinical decisions in Europe. Skin Analytics. URL: https://skin-analytics.com/news/regulatory-certification/derm-class-iii-ce-mark/ [accessed 2025-12-20]
31. Gilbert S, Dai T, Mathias R. Consternation as congress proposal for autonomous prescribing AI coincides with the haphazard cuts at the FDA. NPJ Digit Med 2025;8(1):165 [FREE Full text] [doi: 10.1038/s41746-025-01540-2] [Medline: 40102664]
32. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J 2021;8(2):e188-e194 [FREE Full text] [doi: 10.7861/fhj.2021-0095] [Medline: 34286183]
33. Myny D, Van Goubergen D, Gobert M, Vanderwee K, Van Hecke A, Defloor T. Non-direct patient care factors influencing nursing workload: a review of the literature. J Adv Nurs 2011;67(10):2109-2129. [doi: 10.1111/j.1365-2648.2011.05689.x] [Medline: 21722164]
34. Woolhandler S, Himmelstein DU. Administrative work consumes one-sixth of U.S. physicians' working hours and lowers their career satisfaction. Int J Health Serv 2014;44(4):635-642. [doi: 10.2190/hs.44.4.a]
35. Steinkamp J, Kantrowitz JJ, Airan-Javia S. Prevalence and sources of duplicate information in the electronic medical record. JAMA Netw Open 2022;5(9):e2233348 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.33348] [Medline: 36156143]
36. Fritz P, Kleinhans A, Raoufi R, Sediqi A, Schmid N, Schricker S, et al. Evaluation of medical decision support systems (DDX generators) using real medical cases of varying complexity and origin. BMC Med Inform Decis Mak 2022;22(1):254 [FREE Full text] [doi: 10.1186/s12911-022-01988-2] [Medline: 36153527]
37. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023;330(1):78-80 [FREE Full text] [doi: 10.1001/jama.2023.8288] [Medline: 37318797]
38. Ng JJW, Wang E, Zhou X, Zhou KX, Goh CXL, Sim GZN, et al. Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: a systematic review. BMC Med Inform Decis Mak 2025;25(1):236 [FREE Full text] [doi: 10.1186/s12911-025-03061-0] [Medline: 40598136]
39. Mathias R, McCulloch P, Chalkidou A, Gilbert S. Digital health technologies need regulation and reimbursement that enable flexible interactions and groupings. NPJ Digit Med 2024;7(1):148 [FREE Full text] [doi: 10.1038/s41746-024-01147-z] [Medline: 38890404]
40. Appelbaum SH. Socio - technical systems theory: an intervention strategy for organizational development. Management Decision 1997;35(6):452-463. [doi: 10.1108/00251749710173823]
41. Behymer KJ, Flach JM. From autonomous systems to sociotechnical systems: designing effective collaborations. She Ji J Des Econ Innov 2016;2(2):105-114 [FREE Full text] [doi: 10.1016/j.sheji.2016.09.001]

XSL•FO
RenderX

42. Kudina O, Van de Poel I. A sociotechnical system perspective on AI. Minds Mach 2024;34(3):21. [doi: 10.1007/s11023-024-09680-2]
43. May C, Finch T. Implementing, embedding, and integrating practices: an outline of normalization process theory. Sociology 2009;43(3):535-554. [doi: 10.1177/0038038509103208]
44. Finch TL, Rapley T, Girling M, Mair FS, Murray E, Treweek S, et al. Improving the normalization of complex interventions: measure development based on normalization process theory (NoMAD): study protocol. Implement Sci 2013;8:43 [FREE Full text] [doi: 10.1186/1748-5908-8-43] [Medline: 23578304]
45. Murray E, Treweek S, Pope C, MacFarlane A, Ballini L, Dowrick C, et al. Normalisation process theory: a framework for developing, evaluating and implementing complex interventions. BMC Med 2010;8:63 [FREE Full text] [doi: 10.1186/1741-7015-8-63] [Medline: 20961442]
46. Riedl MO. Human‐centered artificial intelligence and machine learning. Hum Behav & Emerg Tech 2019;1(1):33-36. [doi: 10.1002/hbe2.117]
47. Dawoud K, Samek W, Eisert P, Lapuschkin S, Bosse S. Human-centered evaluation of XAI methods. : IEEE; 2023 Presented at: Proceedings of the 2023 IEEE International Conference on Data Mining Workshops (ICDMW); Dec 4, 2023; Shanghai, China p. 912-921. [doi: 10.1109/icdmw60847.2023.00122]
48. Holzinger A, Kargl M, Kipperer B, Regitnig P, Plass M, Muller H. Personas for artificial intelligence (AI) an open source toolbox. IEEE Access 2022;10:23732-23747. [doi: 10.1109/access.2022.3154776]
49. Combi C, Amico B, Bellazzi R, Holzinger A, Moore JH, Zitnik M, et al. A manifesto on explainability for artificial intelligence in medicine. Artif Intell Med 2022;133:102423 [FREE Full text] [doi: 10.1016/j.artmed.2022.102423] [Medline: 36328669]
50. Woolf SH. The meaning of translational research and why it matters. JAMA 2008;299(2):211-213. [doi: 10.1001/jama.2007.26] [Medline: 18182604]
51. Westerlund A, Sundberg L, Nilsen P. Implementation of implementation science knowledge: the research-practice gap paradox. Worldviews Evid Based Nurs 2019;16(5):332-334 [FREE Full text] [doi: 10.1111/wvn.12403] [Medline: 31603625]
52. Sanderson C, Douglas D, Lu Q, Schleiger E, Whittle J, Lacey J, et al. AI ethics principles in practice: perspectives of designers and developers. IEEE Trans Technol Soc 2023;4(2):171-187. [doi: 10.1109/tts.2023.3257303]
53. Tidjon LN, Khomh F. The different faces of AI ethics across the world: a principle-to-practice gap analysis. IEEE Trans Artif Intell 2023;4(4):820-839. [doi: 10.1109/tai.2022.3225132]
54. Lukkien DRM, Nap HH, Buimer HP, Peine A, Boon WPC, Ket JCF, et al. Toward responsible artificial intelligence in long-term care: a scoping review on practical approaches. Gerontologist 2023;63(1):155-168 [FREE Full text] [doi: 10.1093/geront/gnab180] [Medline: 34871399]
55. Oludapo S, Carroll N, Helfert M. Why do so many digital transformations fail? A bibliometric analysis and future research agenda. J Bus Res 2024;174:114528. [doi: 10.1016/j.jbusres.2024.114528]
56. Wekenborg MK, Gilbert S, Kather JN. Examining human-AI interaction in real-world healthcare beyond the laboratory. NPJ Digit Med 2025;8(1):169 [FREE Full text] [doi: 10.1038/s41746-025-01559-5] [Medline: 40108434]
57. Safi S, Thiessen T, Schmailzl KJ. Acceptance and resistance of new digital technologies in medicine: qualitative study. JMIR Res Protoc 2018;7(12):e11072 [FREE Full text] [doi: 10.2196/11072] [Medline: 30514693]
58. Sujan M, Baber C, Salomon P, Pool R, Chozos N, Aceves-González C. Human factors ergonomics in healthcare AI. Chartered Institute of Ergonomics & Human Factors 2021:45. [doi: 10.13140/RG.2.2.22455.85924]
59. Wosny M, Strasser LM, Hastings J. Experience of health care professionals using digital tools in the hospital: qualitative systematic review. JMIR Hum Factors 2023;10:e50357 [FREE Full text] [doi: 10.2196/50357] [Medline: 37847535]
60. Wekenborg MK, Förster K, Schweden F, Weidemann R, Bechtolsheim FV, Kirschbaum C, et al. Differences in physicians' ratings of work stressors and resources associated with digital transformation: cross-sectional study. J Med Internet Res 2024;26:e49581 [FREE Full text] [doi: 10.2196/49581] [Medline: 38885014]
61. Brod C. Technostress: The Human Cost of the Computer Revolution. Boston, MA: Addison-Wesley; 1984.
62. Alkureishi MA, Choo ZY, Rahman A, Ho K, Benning-Shorb J, Lenti G, et al. Digitally disconnected: qualitative study of patient perspectives on the digital divide and potential solutions. JMIR Hum Factors 2021;8(4):e33364 [FREE Full text] [doi: 10.2196/33364] [Medline: 34705664]
63. Tabche C, Raheem M, Alolaqi A, Rawaf S. Effect of electronic health records on doctor-patient relationship in Arabian gulf countries: a systematic review. Front Digit Health 2023;5:1252227 [FREE Full text] [doi: 10.3389/fdgth.2023.1252227] [Medline: 37877127]
64. Zheng K, Abraham J, Novak LL, Reynolds TL, Gettinger A. A survey of the literature on unintended consequences associated with health information technology: 2014–2015. Yearb Med Inform 2018;25(01):13-29. [doi: 10.15265/iy-2016-036]
65. Holden RJ, Rivera-Rodriguez AJ, Faye H, Scanlon MC, Karsh B. Automation and adaptation: Nurses' problem-solving behavior following the implementation of bar coded medication administration technology. Cogn Technol Work 2013;15(3):283-296 [FREE Full text] [doi: 10.1007/s10111-012-0229-4] [Medline: 24443642]
66. Antecedents of constructive human-AI collaboration: an exploration of human actors' key competencies. In: IFIP Advances in Information and Communication Technology. Cham, Switzerland: Springer International Publishing; 2021:113-124.

XSL•FO
RenderX

67. Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach Learn 2021;110(3):457-506 [FREE Full text] [doi: 10.1007/s10994-021-05946-3]
68. Dung L. Current cases of AI misalignment and their implications for future risks. Synthese 2023;202(5):138 [FREE Full text] [doi: 10.1007/s11229-023-04367-0]
69. Charter of fundamental rights of the European union (2000/C 364/01). European Parliament, the Council and the Commission of the European Union. URL: https://www.europarl.europa.eu/charter/pdf/text_en.pdf [accessed 2025-12-24]
70. Human dignity in the European Union (EU). Values@VET. 2025. URL: https://valuesatvet.si/files/2025/06/Human-dignity-in-the-European-Union.pdf [accessed 2025-12-20]
71. Freedom in the European Union (EU). Values@VET. URL: https://valuesatvet.si/files/2025/06/Freedom-in-the-European-Union.pdf [accessed 2025-12-20]
72. EU mechanism on democracy, the rule of law and fundamental rights: European Parliament resolution of 25 October 2016 with recommendations to the commission on the establishment of an EU mechanism on democracy, the rule of law and fundamental rights (2015/2254(INL)) (2018/C 215/25). European Parliament. 2018. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX%3A52016IP0409&utm [accessed 2025-12-20]
73. Klamert M, Kochenov D. Article 2 TEU. In: The EU Treaties and the Charter of Fundamental Rights: A Commentary. New York: Oxford Academic; 2019:22-30.
74. European Convention on Human Rights, as amended by protocols nos. 11, 14 and 15 supplemented by protocols nos. 1, 4, 6, 7, 12, 13 and 16. European Court of Human Rights. URL: https://www.echr.coe.int/documents/d/echr/convention_ENG [accessed 2025-12-20]
75. Regulation (EU) 2017/745 of 5 April 2017 on medical devices, amending directive 2001/83/EC, regulation (EC) No 178/2002 and regulation (EC) No 1223/2009 and repealing council directives 90/385/EEC and 93/42/EEC. European Parliament and Council of the European Union. URL: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng [accessed 2025-03-25]
76. Gilbert S, Mathias R, Schönfelder A, Wekenborg M, Steinigen-Fuchs J, Dillenseger A, et al. A roadmap for safe, regulation-compliant Living Labs for AI and digital health development. Sci Adv 2025;11(20):eadv7719 [FREE Full text] [doi: 10.1126/sciadv.adv7719] [Medline: 40367163]
77. Calderaro J, Morement H, Penault-Llorca F, Gilbert S, Kather JN. The case for homebrew AI in diagnostic pathology. J Pathol 2025;266(4-5):390-394 [FREE Full text] [doi: 10.1002/path.6438] [Medline: 40613320]
78. Ørngreen R, Levinsen KT. Workshops as a research methodology. Electron J E-Learn 2017;15(1):70-81 [FREE Full text]
79. Concannon TW, Meissner P, Grunbaum JA, McElwee N, Guise J, Santa J, et al. A new taxonomy for stakeholder engagement in patient-centered outcomes research. J Gen Intern Med 2012;27(8):985-991 [FREE Full text] [doi: 10.1007/s11606-012-2037-1] [Medline: 22528615]
80. Understanding healthcare workers confidence in artificial intelligence (AI) (Part 1). NHS Artificial Intelligence (AI) Lab, Health Education England (HEE). 2022. URL: https://digital-transformation.hee.nhs.uk/building-a-digital-workforce/dart-ed/horizon-scanning/understanding-healthcare-workers-confidence-in-ai [accessed 2025-12-20]
81. Hess T, Matt C, Benlian A, Wiesböck F. Options for formulating a digital transformation strategy. In: Strategic Information Management: Theory and Practice. Oxfordshire, UK: Routledge; 2020:494.
82. Kejriwal M. AI in practice and implementation: issues and costs. In: Artificial Intelligence for Industries of the Future. Cham, Switzerland: Springer International Publishing; 2023:25-45.
83. Džakula A, Relić D. Health workforce shortage - doing the right things or doing things right? Croat Med J 2022;63(2):107-109 [FREE Full text] [doi: 10.3325/cmj.2022.63.107] [Medline: 35505643]
84. Global strategy on human resources for health: workforce 2030. World Health Organization. 2016. URL: https://iris.who.int/handle/10665/250368 [accessed 2025-02-13]
85. Rony MKK, Parvin MR, Wahiduzzaman M, Debnath M, Bala SD, Kayesh I. "I Wonder if my years of training and expertise will be devalued by Machines": Concerns about the replacement of medical professionals by artificial intelligence. SAGE Open Nurs 2024;10:23779608241245220 [FREE Full text] [doi: 10.1177/23779608241245220] [Medline: 38596508]
86. Kochan TA. Artificial intelligence and the future of work: a proactive strategy. AI Mag 2021;42(1):16-24. [doi: 10.1002/j.2371-9621.2021.tb00006.x]
87. Feng J, Phillips RV, Malenica I, Bishara A, Hubbard AE, Celi LA, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. NPJ Digit Med 2022;5(1):66 [FREE Full text] [doi: 10.1038/s41746-022-00611-y] [Medline: 35641814]
88. Kumawat E, Datta A, Prentice C, Leung R. Artificial intelligence through the lens of hospitality employees: a systematic review. Int J Hosp Manag 2025;124:103986. [doi: 10.1016/j.ijhm.2024.103986]
89. de Hond AAH, Leeuwenberg AM, Hooft L, Kant IMJ, Nijman SWJ, van Os HJA, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med 2022;5(1):2 [FREE Full text] [doi: 10.1038/s41746-021-00549-7] [Medline: 35013569]
90. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. NPJ Digit Med 2019;2(1):79 [FREE Full text] [doi: 10.1038/s41746-019-0158-1] [Medline: 31453374]

XSL•FO
RenderX

91. Regulation (EU) 2017/746 of 5 April 2017 on in vitro diagnostic medical devices and repealing directive 98/79/EC and commission decision 2010/227/EU. European Parliament and Council of the European Union. URL: https://eur-lex.europa.eu/eli/reg/2017/746/oj/eng [accessed 2025-12-20]
92. Lohr S. What ever happened to IBM's Watson? The New York Times. 2016. URL: https://www.nytimes.com/2021/07/16/technology/what-happened-ibm-watson.html [accessed 2025-12-20]
93. Guidance on the health institution exemption under Article 5(5) of Regulation (EU) 2017/745 and Regulation (EU) 2017/746 (MDCG 2023-1). Medical Device Coordination Group (MDCG). 2023. URL: https://dskb.dk/wp-content/uploads/2021/09/In-house-guidance_stakeholders.pdf [accessed 2025-12-20]
94. Boyle G, Melvin T, Verdaasdonk RM, Van Boxtel RA, Reilly RB. Hospitals as medical device manufacturers: keeping to the medical device regulation (MDR) in the EU. BMJ Innov 2024;10(3):74-80. [doi: 10.1136/bmjinnov-2023-001150]
95. Mit Künstlicher Intelligenz das Krankenhaus von morgen gestalten. SmartHospital.NRW. URL: https://smarthospital.nrw/ [accessed 2025-12-20]
96. Good machine learning practice for medical device development: guiding principles. Medicines and Healthcare products Regulatory Agency (MHRA). URL: https://www.gov.uk/government/publications/good-machine-learning-practice-for-medical-device-development-guiding-principles/good-machine-learning-practice-for-medical-device-development-guiding-principles#guiding-principles [accessed 2021-10-27]
97. Arun S, Grosheva M, Kosenko M, Robertus JL, Blyuss O, Gabe R, et al. Systematic scoping review of external validation studies of AI pathology models for lung cancer diagnosis. NPJ Precis Oncol 2025;9(1):166 [FREE Full text] [doi: 10.1038/s41698-025-00940-7] [Medline: 40483288]
98. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019;25(9):1337-1340. [doi: 10.1038/s41591-019-0548-6] [Medline: 31427808]
99. McCradden MD, London AJ, Gichoya JW, Sendak M, Erdman L, Stedman I, et al. CANAIRI: the collaboration for translational artificial intelligence trials in healthcare. Nat Med 2025;31(1):9-11. [doi: 10.1038/s41591-024-03364-1] [Medline: 39762426]
100. Hellmeier F, Brosien K, Eickhoff C, Meyer A. Beyond one-time validation: a framework for adaptive validation of prognostic and diagnostic AI-based medical devices. ArXiv Preprint posted online on September 7, 2024. [doi: 10.48550/ARXIV.2409.04794]
101. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature 2024;630(8017):625-630 [FREE Full text] [doi: 10.1038/s41586-024-07421-0] [Medline: 38898292]
102. Griot M, Hemptinne C, Vanderdonckt J, Yuksel D. Large language models lack essential metacognition for reliable medical reasoning. Nat Commun 2025;16(1):642 [FREE Full text] [doi: 10.1038/s41467-024-55628-6] [Medline: 39809759]
103. Omar M, Soffer S, Agbareia R, Bragazzi NL, Apakama DU, Horowitz CR, et al. Sociodemographic biases in medical decision making by large language models. Nat Med 2025;31(6):1873-1881. [doi: 10.1038/s41591-025-03626-6] [Medline: 40195448]
104. Alber DA, Yang Z, Alyakin A, Yang E, Rai S, Valliani AA, et al. Medical large language models are vulnerable to data-poisoning attacks. Nat Med 2025;31(2):618-626. [doi: 10.1038/s41591-024-03445-1] [Medline: 39779928]
105. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med 2024;30(9):2613-2622. [doi: 10.1038/s41591-024-03097-1] [Medline: 38965432]
106. Tam TYC, Sivarajkumar S, Kapoor S, Stolyar AV, Polanska K, McCarthy KR, et al. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med 2024;7(1):258 [FREE Full text] [doi: 10.1038/s41746-024-01258-7] [Medline: 39333376]
107. Mehandru N, Miao BY, Almaraz ER, Sushil M, Butte AJ, Alaa A. Evaluating large language models as agents in the clinic. NPJ Digit Med 2024;7(1):84 [FREE Full text] [doi: 10.1038/s41746-024-01083-y] [Medline: 38570554]
108. ISO/IEC 42001:2023 Information technology - artificial intelligence - management system. International Organization for Standardization. 2023. URL: https://www.iso.org/standard/81230.html#lifecycle [accessed 2025-12-20]
109. FG-AI4H DEL7.4 - Clinical evaluation of AI for health. International Telecommunication Union. 2023. URL: https://www.itu.int/pub/T-FG-AI4H-2023-3 [accessed 2025-12-20]
110. Welzel C, Cotte F, Wekenborg M, Vasey B, McCulloch P, Gilbert S. Holistic human-serving digitization of health care needs integrated automated system-level assessment tools. J Med Internet Res 2023;25:e50158 [FREE Full text] [doi: 10.2196/50158] [Medline: 38117545]
111. Mathias R, Vasey B, Chalkidou A, Riedemann L, Melvin T, Gilbert S. Safe AI-enabled digital health technologies need built-in open feedback. Nat Med 2025;31(2):370-375. [doi: 10.1038/s41591-024-03397-6] [Medline: 39905271]
112. Floridi L, Holweg M, Taddeo M, Amaya Silva J, Mökander J, Wen Y. capAI - A procedure for conducting conformity assessment of AI systems in line with the EU artificial intelligence act. SSRN Electron J 2022:91. [doi: 10.2139/ssrn.4064091]
113. Directorate General for Communications Networks, Content and Technology. The assessment list for trustworthy artificial intelligence (ALTAI) for self assessment. European Commission. 2020. URL: https://data.europa.eu/doi/10.2759/002360 [accessed 2025-06-07]

XSL•FO
RenderX

114. Coeckelbergh M. Artificial intelligence, responsibility attribution, and a relational justification of explainability. Sci Eng Ethics 2020;26(4):2051-2068 [FREE Full text] [doi: 10.1007/s11948-019-00146-8] [Medline: 31650511]
115. Santoni de Sio F, Mecacci G. Four responsibility gaps with artificial intelligence: why they matter and how to address them. Philos Technol 2021;34(4):1057-1084 [FREE Full text] [doi: 10.1007/s13347-021-00450-x]
116. Beauchamp T. The principle of beneficence in applied ethics. The Stanford Encyclopedia of Philosophy. 2019. URL: https://plato.stanford.edu/archives/spr2019/entries/principle-beneficence/ [accessed 2025-12-20]
117. Varkey B. Principles of clinical ethics and their application to practice. Med Princ Pract 2021;30(1):17-28 [FREE Full text] [doi: 10.1159/000509119] [Medline: 32498071]
118. Porsdam Mann S, Savulescu J, Sahakian BJ. Facilitating the ethical use of health data for the benefit of society: electronic health records, consent and the duty of easy rescue. Philos Trans A Math Phys Eng Sci 2016;374(2083):20160130 [FREE Full text] [doi: 10.1098/rsta.2016.0130] [Medline: 28336803]
119. Placani A. Anthropomorphism in AI: hype and fallacy. AI Ethics 2024;4(3):691-698 [FREE Full text] [doi: 10.1007/s43681-024-00419-4]
120. McLennan S, Fiske A, Tigard D, Müller R, Haddadin S, Buyx A. Embedded ethics: a proposal for integrating ethics into the development of medical AI. BMC Med Ethics 2022 26;23(1):6 [FREE Full text] [doi: 10.1186/s12910-022-00746-3] [Medline: 35081955]
121. Regulation (EU) 2024/1689 of the European Parliament and of the council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). European Parliament and Council of the European Union. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng [accessed 2025-12-20]
122. Early successes, untapped potential, lingering questions: AI adoption in healthcare report 2024. Healthcare Information and Management Systems Society (HIMSS), Medscape. 2024. URL: https://cdn.sanity.io/files/sqo8bpt9/production/68216fa5d161adebceb50b7add5b496138a78cdb.pdf [accessed 2025-12-20]
123. Schubert T, Oosterlinck T, Stevens RD, Maxwell PH, van der Schaar M. AI education for clinicians. EClinicalMedicine 2025;79:102968 [FREE Full text] [doi: 10.1016/j.eclinm.2024.102968] [Medline: 39720600]
124. Zirar A, Ali SI, Islam N. Worker and workplace artificial intelligence (AI) coexistence: emerging themes and research agenda. Technovation 2023;124:102747. [doi: 10.1016/j.technovation.2023.102747]
125. Elendu C, Amaechi DC, Okatta AU, Amaechi EC, Elendu TC, Ezeh CP, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore) 2024;103(27):e38813 [FREE Full text] [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
126. So HY, Chen PP, Wong GKC, Chan TTN. Simulation in medical education. J R Coll Physicians Edinb 2019;49(1):52-57. [doi: 10.4997/jrcpe.2019.112]
127. Datta R, Upadhyay K, Jaideep C. Simulation and its role in medical education. Med J Armed Forces India 2012;68(2):167-172 [FREE Full text] [doi: 10.1016/S0377-1237(12)60040-9] [Medline: 24623932]
128. Thistlethwaite JE, Davies D, Ekeocha S, Kidd JM, MacDougall C, Matthews P, et al. The effectiveness of case-based learning in health professional education. A BEME systematic review: BEME Guide No. 23. Medical Teacher 2012;34(6):e421-e444. [doi: 10.3109/0142159x.2012.680939]
129. Mukurunge E, Reid M, Fichardt A, Nel M. Interactive workshops as a learning and teaching method for primary healthcare nurses. Health SA 2021;26:1643 [FREE Full text] [doi: 10.4102/hsag.v26i0.1643] [Medline: 34956654]
130. Daye D, Wiggins WF, Lungren MP, Alkasab T, Kottler N, Allen B, et al. Implementation of clinical artificial intelligence in radiology: who decides and how? Radiology 2022;305(3):555-563 [FREE Full text] [doi: 10.1148/radiol.212151] [Medline: 35916673]
131. Hoff KA, Bashir M. Trust in automation: integrating empirical evidence on factors that influence trust. Hum Factors 2015;57(3):407-434. [doi: 10.1177/0018720814547570] [Medline: 25875432]
132. Jiang T, Sun Z, Fu S, Lv Y. Human-AI interaction research agenda: a user-centered perspective. Data Inf Manag 2024;8(4):100078. [doi: 10.1016/j.dim.2024.100078]
133. Vered M, Livni T, Howe PDL, Miller T, Sonenberg L. The effects of explanations on automation bias. Artificial Intelligence 2023;322:103952. [doi: 10.1016/j.artint.2023.103952]
134. Choudhury A, Chaudhry Z. Large language models and user trust: consequence of self-referential learning loop and the deskilling of health care professionals. J Med Internet Res 2024;26:e56764 [FREE Full text] [doi: 10.2196/56764] [Medline: 38662419]
135. Ng MY, Kapur S, Blizinsky KD, Hernandez-Boussard T. The AI life cycle: a holistic approach to creating ethical AI for health decisions. Nat Med 2022;28(11):2247-2249 [FREE Full text] [doi: 10.1038/s41591-022-01993-y] [Medline: 36163298]
136. Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. Lancet Digit Health 2022;4(5):e384-e397. [doi: 10.1016/s2589-7500(22)00003-6]
137. Faust L, Wilson P, Asai S, Fu S, Liu H, Ruan X, et al. Considerations for quality control monitoring of machine learning models in clinical practice. JMIR Med Inform 2024;12:e50437 [FREE Full text] [doi: 10.2196/50437] [Medline: 38941140]

XSL•FO
RenderX

138. Planning for managing legacy systems and decommissioning digital healthcare technologies. NHS AI and Digital Regulations Service for Health and Social Care. URL: https://www.digitalregulations.innovation.nhs.uk/regulations-and-guidance-for-adopters/all-adopters-guidance/planning-for-managing-legacy-systems-and-decommissioning-digital-healthcare-technologies/ [accessed 2023-11-13]
139. Sharma A, Nayancy, Verma R. The Confluence of Cryptography, Blockchain and Artificial Intelligence. Florida, USA: CRC Press; 2025.
140. MHRA, Brunel University. Project Report: Research into Methodology for Determining Significant Change in the Way That an Adaptive AI Algorithm Medical Device Is Working and How Such Change Should Be Regulated. https://www.gov.uk. URL: https://www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device/software-and-artificial-intelligence-ai-as-a-medical-device [accessed 2025-03-29]
141. The most trusted AI in mental healthcare: scale behavioral health with clinical AI. Limbic. 2025. URL: https://www.limbic.ai/ [accessed 2025-12-20]
142. We provide validated information for healthcare professionals. Prof. Valmed - Validated Medical Information GmbH. URL: https://profvalmed.com/ [accessed 2025-12-20]
143. Frequently asked questions. deepc GmbH. URL: https://www.deepc.ai/learn/faq [accessed 2025-12-20]
144. Study on the deployment of AI in healthcare: final report. Publications Office of the European Union. 2025. URL: https://data.europa.eu/doi/10.2875/2169577 [accessed 2025-08-07]
145. Eskofier BM, Klucken J. Predictive models for health deterioration: understanding disease pathways for personalized medicine. Annu Rev Biomed Eng 2023;25(1):131-156 [FREE Full text] [doi: 10.1146/annurev-bioeng-110220-030247] [Medline: 36854259]
146. Goldfarb A, Taska B, Teodoridis F. Artificial intelligence in health care? Evidence from online job postings. AEA Pap Proc 2020;110:400-404 [FREE Full text] [doi: 10.1257/pandp.20201006]
147. Wu K, Wu E, Theodorou B, Liang W, Mack C, Glass L, et al. Characterizing the clinical adoption of medical AI devices through U.S. insurance claims. NEJM AI 2024;1(1). [doi: 10.1056/aioa2300030]
148. Ulfert AS, Le Blanc P, González-Romá V, Grote G, Langer M. Are we ahead of the trend or just following? The role of work and organizational psychology in shaping emerging technologies at work. Eur J Work Organ Psychol 2024;33(2):120-129. [doi: 10.1080/1359432x.2024.2324934]
149. Gilbert S, Anderson S, Daumer M, Li P, Melvin T, Williams R. Learning from experience and finding the right balance in the governance of artificial intelligence and digital health technologies. J Med Internet Res 2023;25:e43682 [FREE Full text] [doi: 10.2196/43682] [Medline: 37058329]
150. Ganesan S, Somasiri N. Navigating the integration of machine learning in healthcare: challenges, strategies, and ethical considerations. J Comput Cogn Eng 2024. [doi: 10.47852/bonviewJCCE42023600]
151. Developing healthcare workers' confidence in artificial intelligence (AI) (Part 2). NHS Artificial Intelligence (AI) Lab, Health Education England (HEE). 2023. URL: https://digital-transformation.hee.nhs.uk/building-a-digital-workforce/dart-ed/horizon-scanning/developing-healthcare-workers-confidence-in-ai [accessed 2025-12-20]
152. Reddy S, Rogers W, Makinen VP, Coiera E, Brown P, Wenzel M, et al. Evaluation framework to guide implementation of AI systems into healthcare settings. BMJ Health Care Inform 2021;28(1):e100444 [FREE Full text] [doi: 10.1136/bmjhci-2021-100444] [Medline: 34642177]
153. Moreno-Sánchez PA, Ser JD, Gils MV, Hernesniemi J. A design framework for operationalizing trustworthy artificial intelligence in healthcare: requirements, tradeoffs and challenges for its clinical adoption. Information Fusion 2025;127:103812. [doi: 10.2139/ssrn.5249603]
154. Nair M, Nygren J, Nilsen P, Gama F, Neher M, Larsson I, et al. Critical activities for successful implementation and adoption of AI in healthcare: towards a process framework for healthcare organizations. Front Digit Health 2025;7:1550459 [FREE Full text] [doi: 10.3389/fdgth.2025.1550459] [Medline: 40453810]
155. Nilsen P, Svedberg P, Neher M, Nair M, Larsson I, Petersson L, et al. A framework to guide implementation of AI in health care: protocol for a cocreation research project. JMIR Res Protoc 2023;12:e50216 [FREE Full text] [doi: 10.2196/50216] [Medline: 37938896]
156. Dagan N, Devons-Sberro S, Paz Z, Zoller L, Sommer A, Shaham G, et al. Evaluation of AI solutions in health care organizations — The OPTICA tool. NEJM AI 2024;1(9). [doi: 10.1056/aics2300269]
157. Mittermaier M, Raza M, Kvedar JC. Collaborative strategies for deploying AI-based physician decision support systems: challenges and deployment approaches. NPJ Digit Med 2023;6(1):137 [FREE Full text] [doi: 10.1038/s41746-023-00889-6] [Medline: 37543707]
158. Davahli MR, Karwowski W, Fiok K, Wan T, Parsaei HR. Controlling safety of artificial intelligence-based systems in healthcare. Symmetry 2021;13(1):102 [FREE Full text] [doi: 10.3390/sym13010102]
159. Labkoff S, Oladimeji B, Kannry J, Solomonides A, Leftwich R, Koski E, et al. Toward a responsible future: recommendations for AI-enabled clinical decision support. J Am Med Inform Assoc 2024;31(11):2730-2739. [doi: 10.1093/jamia/ocae209] [Medline: 39325508]
160. Lekadira K, Osuala R, Gallin C. FUTURE-AI: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. ArXiv Preprint posted online on September 20, 2021. [doi: 10.48550/arXiv.2109.09658]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
CE: Conformité Européenne
EU: European Union
EURAID: European Responsible AI Development
FG-AI4H: Focus Group on AI for Health
HCP: health care professional
ISO/IEC: International Organization for Standardization/ International Electrotechnical Commission
LLM: large language model
MDR: Medical Device Regulation
NHS: National Health Service
SMART: specific, measurable, attainable, relevant, and time-bound
TEU: Treaty on European Union
XAI: explainable AI

Edited by J Sarvestan; submitted 16.Jul.2025; peer-reviewed by I Schlömer, KH Lin; comments to author 22.Aug.2025; revised version received 05.Nov.2025; accepted 06.Nov.2025; published 29.Jan.2026.

Please cite as:
Schönfelder A, Eberlein-Gonska M, Hülsken-Giesler M, Jovy-Klein F, Kather JN, Kohoutek E, Lennefer T, Liebert E, Lipprandt M, Mathias R, Muti HS, Obergassel J, Reibel T, Rösler U, Schneider M, Schlicht L, Schlieter H, Schmieding ML, Schweingruber N, Sedlmayr M, Strametz R, Susec B, Wekenborg MK, Weicken E, Weitz K, Diehl A, Gilbert S
Collaborative and Cooperative Hospital "In-House" Medical Device Development and Implementation in the AI Age: The European Responsible AI Development (EURAID) Framework Compatible With European Values
J Med Internet Res 2026;28:e80754
URL: https://www.jmir.org/2026/1/e80754
doi:10.2196/80754
PMID:

©Anett Schönfelder, Maria Eberlein-Gonska, Manfred Hülsken-Giesler, Florian Jovy-Klein, Jakob Nikolas Kather, Elisabeth Kohoutek, Thomas Lennefer, Elisabeth Liebert, Myriam Lipprandt, Rebecca Mathias, Hannah Sophie Muti, Julius Obergassel, Thomas Reibel, Ulrike Rösler, Moritz Schneider, Larissa Schlicht, Hannes Schlieter, Malte L Schmieding, Nils Schweingruber, Martin Sedlmayr, Reinhard Strametz, Barbara Susec, Magdalena Katharina Wekenborg, Eva Weicken, Katharina Weitz, Anke Diehl, Stephen Gilbert. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

AI Triage in Primary Care: Building Safer and More Equitable Real-World Evidence

Aymn Alamoudi[1,2], MSc; Evangelos Kontopantelis[3,4], PhD; Salwa S Zghebi[1], PhD; Benjamin Brown[1], PhD

[1]Division of Population Health, Health Services Research and Primary Care, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, England, United Kingdom

[2]Department of Public Health, School of Nursing and Health Sciences, Jazan University, Jazan, Jazan Region, Saudi Arabia

[3]Division of Informatics, Imaging and Data Sciences, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, England, United Kingdom

[4]Division of Family Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

Corresponding Author:
Aymn Alamoudi, MSc
Division of Population Health, Health Services Research and Primary Care, School of Health Sciences, Faculty of Biology, Medicine and Health
University of Manchester
Williamson Building, 5th Floor
Oxford Road
Manchester, England, M13 9PL
United Kingdom
Phone: 44 161 306 6000
Email: aymn.alamoudi@postgrad.manchester.ac.uk

Abstract

Artificial intelligence triage in general practice is developing rapidly within the primary care digital transformation, promising efficiency gains and safety standardization in overwhelmed primary care systems. However, current evidence is drawn from retrospective validations, emergency settings, or vignettes, with scant evaluation of real-world outcomes and almost no equity-stratified safety data, despite known disparities across age, ethnicity, language, and deprivation. From a sociotechnical standpoint, which considers the fit between people, tasks, technology, and organizational context, risks arise not only from algorithmic bias and undertriage but also from human factors, workflow misalignment, governance gaps, and inadequate postdeployment monitoring. We argue that ensuring artificial intelligence triage is safe and equitable requires real-world evaluations in primary care settings, equity-focused performance reporting using theoretically informed frameworks, and rigorous postmarket surveillance. Without these, deployment may widen existing health inequalities rather than moderate them.

(J Med Internet Res 2026;28:e88396) doi:10.2196/88396

KEYWORDS

artificial intelligence; AI; triage; primary care; general practice; patient safety; health equity; real-world evidence; clinical decision support

Introduction

Globally, primary care faces sustained growth in demand, increased patient complexity, and a workforce whose full-time equivalent growth has not matched demand, resulting in persistent access pressures and delays [1,2]. The COVID-19 pandemic accelerated the adoption of remote and digital access, including online consultations and reinforced strategic commitments to “digital front door” models within health systems, such as the National Health Service [3-5]. Online consultation submissions in England rose from approximately 2.7 million in October 2023 to a peak of 8.3 million in October 2025, as seen in Figure 1, highlighting the rapid and sustained growth of digital entry points into general practice (GP) [6]. Evidence also suggests that digital access is not equity neutral. A systematic review of inequalities in remote GP consultations found differential use by sociodemographic characteristics, with internet-based consultations more frequently used by younger, more affluent, and more educated groups, and noted that the impact of these inequalities on clinical outcomes remains uncertain [7].

XSL•FO
RenderX

Figure 1. Growth in monthly online consultation submissions in England. Data source: National Health Service England release [6]. Information from NHS England, licenced under the current version of the Open Government Licence.

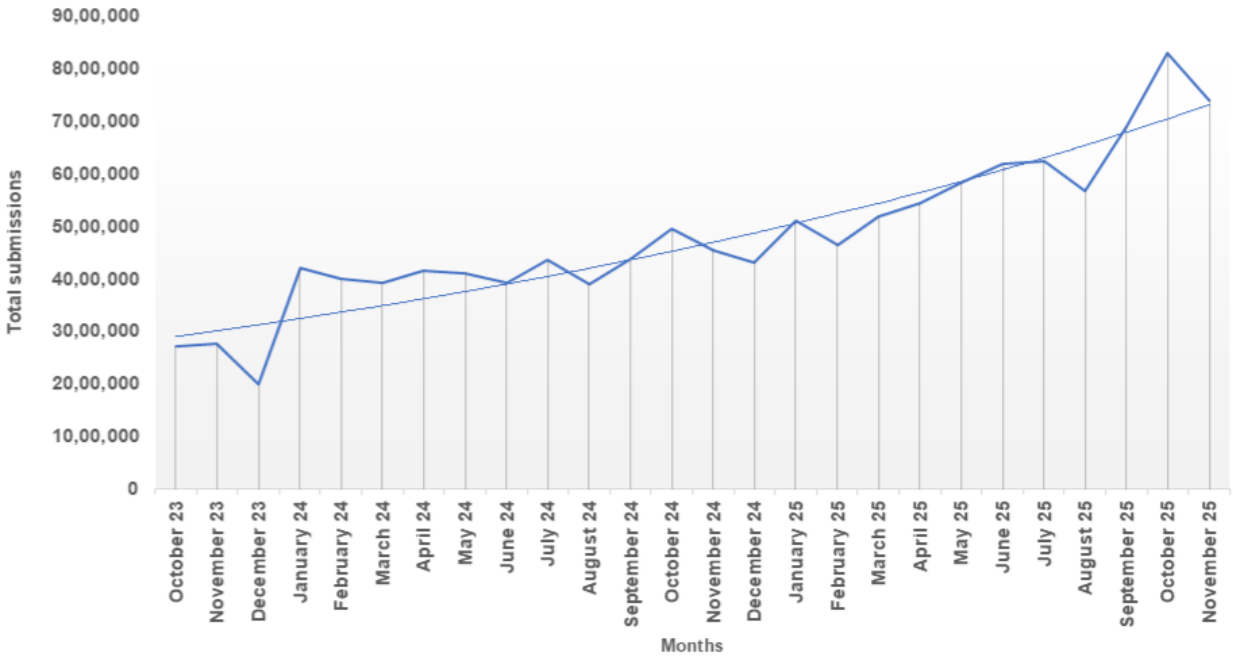

In this context, artificial intelligence (AI)–enabled triage combines structured questions, red-flag pathways, and machine learning (ML) risk stratification with electronic health record (EHR) integration and clinician oversight to route patients more efficiently and potentially improve safety [8-10]. In this viewpoint, we distinguish between 3 related but conceptually distinct system types: symptom checkers, clinical decision support systems (CDSSs), and AI triage. Symptom checkers are patient-facing digital tools that provide health advice or triage recommendations directly to users, often without clinician oversight, and have been widely evaluated in consumer and emergency contexts [8]. CDSSs are clinician-facing tools embedded within clinical workflows or EHRs that support decision-making through alerts, risk scores, or guideline-based recommendations [9,10]. AI-enabled triage refers to digital systems that collect patient-reported information and generate urgency or routing recommendations (eg, self-care, routine review, urgent GP assessment, or emergency referral), with or without clinician oversight [11]. These systems may be embedded within online consultation platforms, patient-facing symptom checkers, or CDSS. Importantly, not all online consultation systems are AI enabled and not all AI-enabled triage systems function as stand-alone symptom checkers.

This viewpoint advances 3 linked arguments. First, we argue that triage in primary care is a safety-critical and equity-sensitive function, such that errors or delays can produce serious harm and unequal outcomes. Second, we show that the current evidence base for AI-supported triage is dominated by emergency department (ED), vignette, and retrospective studies, with little real-world or equity-stratified evaluation in GP. Third, we argue that AI triage operates as a sociotechnical system shaped by human behavior, workflows, and governance, meaning that algorithmic accuracy alone cannot guarantee safety or fairness. This viewpoint aims to outline a practical agenda for evaluating and governing AI-enabled triage in GP that integrates real-world safety outcomes, equity-stratified performance reporting, and sociotechnical implementation and surveillance. The intended audience includes GP clinicians and practice leaders, digital health and AI developers, evaluators and implementation scientists, and policymakers and regulators responsible for deployment and monitoring. Our contribution is to consolidate a practical, real-world evaluation and governance agenda for AI triage in GP that integrates sociotechnical safety (workflow and human factors), equity-stratified performance reporting (including an example of fairness), and postdeployment surveillance.

Triage in Primary Care Is Safety Critical and Equity Sensitive

In health care, triage refers to the systematic process of assessing patient urgency and risk to determine the appropriate level, timing, and pathway of care [12,13]. In primary care, triage does not establish a diagnosis but prioritizes patients for self-care, routine review, urgent GP assessment, or emergency referral based on presenting symptoms, clinical risk, and service capacity [14,15]. This function is safety critical because misclassification can lead to delayed diagnosis, inappropriate self-management, or unnecessary escalation [8-10]. AI-enabled triage promises standardization and auditability but introduces novel patient-safety risks, such as automation bias, algorithmic mistriage, and digital exclusion, particularly for socially disadvantaged groups [8,16,17]. Moreover, model performance (eg, sensitivity, specificity, calibration, and error rates) may vary by age, ethnicity, language, or limited digital access, unless these dimensions are intentionally tested and monitored [18-22].

XSL•FO
RenderX

The absence of equity evaluation may pose a significant risk. If models are calibrated primarily on majority language, younger, or White-majority cohorts, AI triage may systematically de-escalate or deprioritize patients whose symptom descriptions diverge due to cultural or linguistic factors. Coupled with cognitive and automation biases in human users, the most vulnerable groups risk unsafe disposition, such as self-care advice when urgent assessment is indicated.

The "equity blind spot" in AI triage is not merely a technical glitch; it reflects broader systemic oversight. To operationalize safe, equitable AI, embedding framework-informed stratification, for example, PROGRESS-Plus, fairness metrics, and multidimensional performance reporting into every stage of model development, validation, and deployment is needed. Without this, AI triage may reinforce and even amplify existing health care disparities if deployed without adequate safeguards.

What the Evidence Shows and What It Misses

What Current Studies Show

Controlled evaluations of AI triage report high technical performance, with area under the receiver operating characteristic (singular) curve values typically between 0.82 and 0.94 and sensitivities often exceeding 0.75. However, these studies are predominantly retrospective, vignette based, or conducted in EDs and hospital settings. They rarely reflect routine workflows in GP [23-26].

Recently, Abualruz et al [12] reviewed 22 studies on AI-supported triage, most of which were carried out in emergency, acute, or hospital-based settings. Only a small subset examined outpatient or primary-care use. As a result, the current literature tells us little about how AI triage performs in everyday GP or how it affects patient safety in real workflows. Table 1 presents the distribution of published AI-supported triage studies by clinical setting.

Table 1. Distribution of published artificial intelligence–supported triage studies by clinical setting (N=22).

Clinical setting	Study type	Studies, n (%)
Emergency department or hospital	Real patient data	19 (86)
Primary care	Clinical vignettes or qualitative studies	3 (14)
Primary care	Real patient data	0 (0)

Most studies have been conducted in emergency or hospital settings using real patient data (19/22, 86%). In total, 14% (3/22) of the studies relied on clinical vignettes or qualitative interviews, and none (0/22, 0%) evaluated AI-supported triage using real-world patient data in routine GP.

Why This Evidence Is Not Sufficient for Safe and Equitable GP Deployment

Equity reporting is also sparse. Few studies disaggregate performance by age, ethnicity, language, or socioeconomic status [27-30]. Intersectional analyses, for example, age×ethnicity or ethnicity×deprivation, are almost absent. Subgroup calibration, false-negative rates, and false-positive rates (FPRs) are rarely reported.

Study design further limits interpretability. Vignette-based and retrospective analyses do not capture real-world pressures, such as workload variation, free-text symptom input, clinician overrides, or case-mix drift [31]. Prospective designs, such as controlled interrupted time series or cluster-randomized trials, are almost never used in GP settings [1,3]. Without these designs, safety effects cannot be attributed reliably to AI deployment.

Postdeployment monitoring is also underdeveloped (Table 2). Few studies report ongoing calibration checks, subgroup performance dashboards, or systematic incident reporting aligned to the World Health Organization (WHO) International Classification for Patient Safety [32,33]. As a result, health systems lack visibility into how AI triage safety changes over time.

Table 2. What the evidence on artificial intelligence triage shows and what it misses.

Domain	What studies typically show	What is usually missing
Accuracy	High area under the receiver operating characteristic curve (≈0.82-0.94) and reasonable sensitivity in retrospective and vignette-based studies	Performance under real general practitioner workload, with free-text input, comorbidity, and clinician override
Setting	Predominantly emergency departments, acute care, or simulated cases	Routine general practice, community clinics, and longitudinal follow-up
Safety outcomes	Agreement with clinicians or reference standards	Delayed diagnosis, avoidable emergency use, or patient harm
Equity	Rare or absent subgroup reporting	Performance by age, ethnicity, language, deprivation, or intersectional groups
Monitoring	One-off validation at model development	Postdeployment drift, subgroup miscalibration, and incident tracking

XSL•FO
RenderX

Consistent with this, an ED-focused scoping review found limited demographic breakdowns and no multidimensional analyses, leaving equity implications unclear [27]. Within UK GP settings, stratified data on undertriage or performance by deprivation or ethnicity are rare [16,29], and subgroup calibration metrics or true-positive rate (TPR) and FPR reporting are notably absent. A recent international review found variable triage accuracy, poor calibration reporting, and limited deployment-level evaluation, reinforcing that the evidence gap is global [34].

Why AI Triage Is a Sociotechnical System

Overview

AI-enabled triage systems are not isolated algorithms; they operate within complex care delivery systems where human factors, workflows, and trust dynamics profoundly shape safety. A sociotechnical system perspective, exemplified by the systems engineering initiative for patient safety framework, analyzes how people, tasks, tools, and organizational structures interact to influence patient safety.

By contrast, implementation frameworks focus on organizational readiness, technology adoption, and sustainability [35,36]. Together, these complementary approaches emphasize that deploying an accurate ML model alone does not guarantee safe outcomes; safety depends on both workflow integration and organizational adoption.

Human factors and trust are central. Health care professionals and patients must interpret AI-generated recommendations within their cognitive, ethical, and emotional contexts. A recent qualitative work on AI-based triage in Swedish primary care underlines how trust emerges from lived experience, transparency, and perceived reliability. Both patients' and professionals' trust is contingent on real-world usability and clear decision roles and not just model accuracy [31].

Similar issues are emerging in teledentistry and dental triage, where AI-enabled chatbots and triage systems are used to prioritize pain, infection, trauma, or urgent referral pathways [37]. Early work includes prototype "intelligent dental triage systems" and evaluations of AI chatbots for dental queries, but the same core risks apply: safety-critical undertriage, unequal performance for patients with language barriers or limited digital access, and workflow integration challenges in busy dental practices [38]. AI tools in dental assessment and smile analysis, such as Dynasmile, a video-based AI smile analysis platform in orthodontics, illustrate the expanding role of AI beyond workflow triage into diagnostic and aesthetic decision support in oral care [39].

This cross-domain comparison reinforces our central claim: AI triage should be evaluated as a sociotechnical intervention with equity-stratified safety reporting and postdeployment monitoring, regardless of clinical specialty.

Explainable AI to Support Calibrated Trust and Reduce Automation Bias

In safety-critical triage, explanations should aim to support calibrated trust not persuasion. Evidence from clinical decision-support research suggests that well-designed explanations can improve clinician understanding and trust calibration, while poorly designed explanations can increase overreliance and automation bias [40,41].

In practice, explainable AI for GP triage should be workflow integrated and low burden, including the following: (1) a short list of the main drivers for escalation (eg, red-flag symptoms and abnormal risk profile), (2) uncertainty indicators or confidence bands, (3) an "override required" prompt for high-risk edge cases, and (4) safety-netting text that is consistent with the triage rationale. Explanation stability is also important; near-identical inputs should not produce inconsistent rationales, as this undermines trust and may increase unsafe deference [40,42].

These requirements for explanation design reinforce the importance of sociotechnical fit described subsequently.

What Real-World Deployments Show

AI triage does not operate as a stand-alone algorithm. In GP, it is embedded in symptom checkers, online consultations, patient apps, and EHR-based decision-support tools. These systems influence how patients describe symptoms, how clinicians prioritize work, and how care is delivered.

Evidence from multiple countries shows potential efficiency gains. In Iceland, an ML triage model for respiratory symptoms improved previsit risk stratification in community clinics [43]. Similarly, Brazil's primary care referral triage system demonstrated improved appropriateness of specialist referrals [44]. Studies from Sweden and Italy reported improved workflow transparency but persistent concerns about trust, usability, and clinician acceptance [31,33].

However, most deployments remain early stage, small scale, or limited to specific pathways. Systematic reviews of symptom checkers across Europe, the United States, Spain, Canada, and Asia report wide variability in triage accuracy and frequent mismatches between algorithmic and clinician assessments, particularly for complex multimorbidity and non–native-language users [8,14,34,45].

Why Workflow, Trust, and Integration Matter

Qualitative evidence highlights that sociotechnical fit is critical. Steerling et al [31] found that both health care professionals and patients require alignment with clinical judgment, transparency, and oversight before trusting AI-based triage. Similarly, Siira et al [46] identified 3 interacting barriers in Swedish primary care: (1) professional skepticism or resistance and trust, (2) organizational readiness and digital maturity, and (3) technical limitations and poor EHR integration. Successful sites mitigated these by hands-on leadership and staff training, "superuser" networks, and iterative codevelopment with vendors. Even where efficiency gains were perceived, unresolved integration gaps and complex case-mix sustained workload and safety concerns.

Evidence from the UK primary care e-visits (14 practices; 16 staff, 24 patients; 2020-2021) identified 7 concrete AI use cases—workflow routing, directing, prioritization, postsubmission adaptive questioning, writing assistance,

XSL•FO
RenderX

self-help information, and autobooking—and found acceptability hinged on clinical oversight, timely responses, and ongoing evaluation. Perceived upsides were workload relief and faster help, while risks were depersonalization and mistriage, if poorly implemented [47].

What This Means for Safety

These findings show that AI triage safety depends on how tools interact with people, workflows, and organizational routines. This aligns with sociotechnical safety theory, particularly the systems engineering initiative for patient safety framework, which emphasizes the fit between tasks, technology, and organizational context [32,48,49].

AI also offers enhanced structured data capture, natural language processing (NLP)–enabled symptom interpretation, EHR-integrated safety netting, and auditable decision trails [50]. NLP-enabled symptom interpretation can, for example, recognize "feeling pressure in chest" as equivalent to angina, supporting safer triage for patients who do not use standard terminology [26,45]. However, these benefits are offset by risks. Algorithmic mistriage, automation bias, and poor integration can delay escalation or overload urgent pathways [8,13].

Without governance mechanisms, such as version control, audit logs, safety dashboards, and periodic revalidation, performance may degrade over time as populations, language, and risk profiles change [32]. Therefore, sociotechnical alignment is not optional; it appears essential for safe AI-enabled triage.

The Equity Blind Spot

Although AI-enabled triage systems are promoted as equitable tools for managing primary care demand, the current evidence reveals a persistent equity blind spot, driven by underrepresentation, limited fairness measurement, and neglect of equity-focused monitoring [51].

Equity in digital health demands more than equal treatment; it requires fair opportunity to achieve safe, good outcomes, especially when baseline disadvantages exist. Equity reporting should use structured tools. The PROGRESS-Plus framework, developed by the Cochrane-Campbell equity methods group, was designed to systematically identify and report on equity-relevant factors in health research [17]. It extends the original PROGRESS (place of residence, race and ethnicity, occupation, gender, religion, education, socioeconomic status, and social capital) acronym with "plus" dimensions, including age, disability, and language [17]. The framework was created to help researchers illuminate disparities that might otherwise be masked in aggregated data and has since been widely applied in clinical trials, systematic reviews, and digital health evaluations.

Alternative frameworks, such as the health equity impact assessment tool, as seen in Textbox 1, are used prospectively to anticipate equity impacts before interventions are deployed [52]. The SIITHIA (Strengthening the Integration of Intersectionality Theory in Health Inequality Analysis) checklist provides structured criteria for identifying inequities in digital health [53]. More recently, the digital health equity framework extends this approach to digital interventions and multidimensional analyses [54]. In practice, these can be operationalized by setting thresholds (eg, sensitivity gaps ≤5% between groups), with breaches triggering model review and corrective action. Complementing this, algorithmic fairness metrics, such as equal opportunity (equal TPRs), equalized odds (matching both TPR and FPR), and calibration integrity, are critical for measuring subgroup performance and detecting systematic bias. Without these frameworks, inequities may go undetected beneath aggregated performance.

Textbox 1. Frameworks for evaluating safety and equity in artificial intelligence triage.

Safety and sociotechnical performance

- Systems engineering initiative for patient safety describes how people, tasks, tools, and organizational context interact to shape patient safety [48].
- World Health Organization International Classification for Patient Safety provides a standardized taxonomy for reporting and classifying safety incidents [32,49,55].

Implementation and adoption

- Consolidated Framework for Implementation Research assesses organizational readiness and barriers to and facilitators of implementation.
- Nonadoption, abandonment, scale-up, spread, and sustainability [35,36] framework evaluates the complexity and long-term viability of digital health technologies.
- Reach, effectiveness, adoption, implementation, and maintenance framework evaluates long-term reach, effectiveness, adoption, and maintenance [56].
- Human, organization, and technology-fit framework examines alignment between human, organizational, and technical factors [57].

Equity and fairness

- PROGRESS-Plus identifies social stratifiers, such as age, ethnicity, language, deprivation, and disability [17].
- Health Equity Impact Assessment tool evaluates potential equity impacts before deployment [52].
- The SIITHIA (Strengthening the Integration of Intersectionality Theory in Health Inequality Analysis) checklist and the digital health equity framework support intersectional and digital-specific equity analysis [53,54].

XSL•FO
RenderX

The frameworks presented in Textbox 1 allow AI triage to be evaluated across safety, implementation, and equity dimensions. Combining them enables a more comprehensive assessment than any single lens can provide.

Textbox 2 provides an illustrative example of how to report equal opportunity (true positive rate; sensitivity) across intersectional PROGRESS-Plus strata.

Textbox 2. Illustrative example: reporting equal opportunity across intersectional strata.

- Equal opportunity requires similar true positive rates (TPRs; sensitivity) across groups for individuals who truly need urgent care.
- Step 1 includes defining the safety-critical outcome (Y=1), for example, "urgent same-day clinical assessment required" based on a reference standard (eg, clinician adjudication, emergency department attendance within 24 to 48 hours, or diagnosis of a time-critical condition).
- Step 2 involves computing TPR within each subgroup.
- TPR (sensitivity) = true positive (TP) / (TP + false negative [FN])
- Step 3 involves reporting TPR across PROGRESS-Plus strata and intersectional strata, for example, age group×ethnicity or ethnicity×deprivation quintile. An example of reporting is as follows:
- White, least deprived: TP=180; FN=20 → TPR=0.90
- White, most deprived: TP=150; FN=30 → TPR=0.83
- Minority ethnicity, least deprived: TP=70; FN=20 → TPR=0.78
- Minority ethnicity, most deprived: TP=55; FN=25 → TPR=0.69
- Step 3a involves quantifying uncertainty in subgroup estimates. For each subgroup, TPRs should be reported with measures of uncertainty (eg, 95% CIs or standard errors), particularly where subgroup sample sizes are small. This enables assessment of whether observed differences are robust or compatible with random variation.
- Step 4 summarizes the disparity. This approach makes equity risks visible that would be hidden in overall performance metrics.
- Absolute gap (maximum TPR – minimum TPR) = 0.90 – 0.69 = 0.21
- Flag threshold: investigate if the gap is greater than 0.05 or if any subgroup TPR is less than 0.80
- These figures are illustrative. In practice, equity assessments should consider statistical uncertainty (eg, CI overlap and subgroup sample size) alongside point estimates. Empirical data on equity-stratified artificial intelligence triage performance in primary care remain limited.

Several dimensions of bias have been documented or are anticipated in an AI-driven triage system, as mentioned subsequently.

The first dimension is age. Older adults are frequently underrepresented in development datasets, increasing the risk of atypical presentations or those with limited digital literacy [58,59].

The second dimension is language and ethnicity. NLP models are extremely sensitive to linguistic variation, dialects, multilingual input, or limited English proficiency. However, model evaluations rarely account for these, threatening safe triage in diverse populations [60].

Furthermore, broader AI research (outside primary care) shows that racial and ethnic biases in algorithmic systems persist [61]. For instance, algorithms that relied on health care cost as a proxy for illness systematically undertriaged Black patients due to unequal access-driven cost differences [60]. Similarly, AI in imaging frequently underdiagnoses emergent pathology in marginalized groups. Black women have shown significantly higher underdiagnosis rates in medical imaging models.

Telephone or digital triage evaluations suggest that low-income individuals, ethnic minority groups, and displaced patients experience worse outcomes, though quantitative data remain sparse [62].

Single-axis analysis (age or ethnicity) is insufficient. Intersecting vulnerabilities (eg, older adults from minority ethnic backgrounds with language barriers) can compound risk and increase mistriage. However, only a limited number of studies disaggregate safety performance by intersectional subgroups, leaving some of the most disadvantaged populations effectively invisible in assessments. Such analyses often lack power; therefore, findings should be treated as exploratory unless supported by large, multisite cohorts.

What Must Change: a Research and Policy Agenda

In this viewpoint paper, we argue that AI triage stands at a crossroad; it has the potential to improve safety and access in primary care but requires real-world evaluation, equity-focused monitoring, and sociotechnical governance. Figure 2 summarizes the proposed real-world evaluation and governance loop for AI triage in GP. Digital entry points feed into AI-supported triage and clinician workflow, producing safety and service-use outcomes that inform monitoring and governance (including equity dashboards, drift detection, and incident reporting). Governance outputs drive model and workflow updates, enabling continuous improvement.

XSL•FO
RenderX

Figure 2. Conceptual loop for evaluating and governing artificial intelligence (AI)–enabled triage in general practice as a sociotechnical intervention.

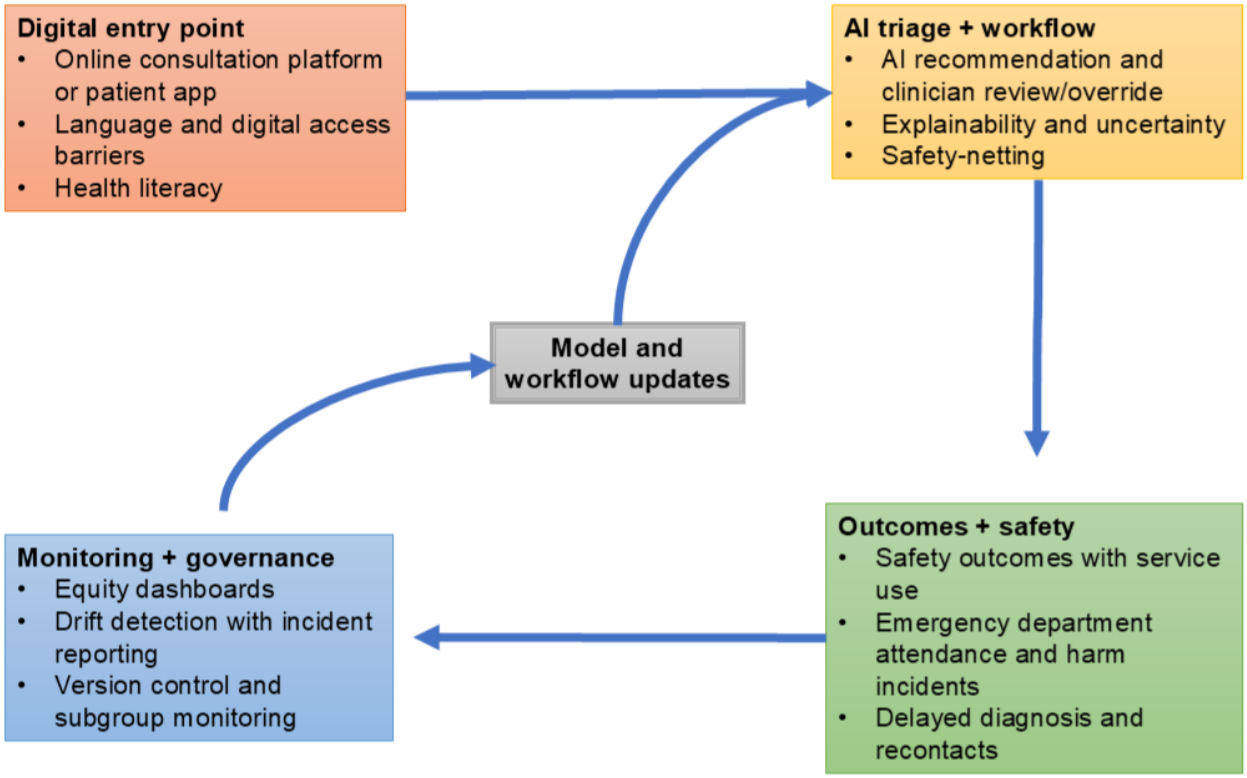

To shift AI triage from a hypothetical promise to an equitable, safe reality in GP and primary care settings, we propose 5 interrelated priorities, as mentioned subsequently.

The first priority is real-world evaluations in primary care (prospective or retrospective). Current evidence is dominated by vignette experiments or ED contexts. Prospective, real-world evaluations—such as controlled interrupted time series or (cluster) randomized controlled trials (RCTs)—that assess patient safety outcomes (eg, delayed diagnoses and avoidable emergency use), workflow effects, and override behaviors in GP practice are urgently needed. Where randomization is infeasible, controlled interrupted time series with matched controls and prespecified safety outcomes can provide strong quasi-experimental evidence.

The second priority is equity-stratified performance reporting. AI triage systems should be evaluated through the equity lenses, such as PROGRESS-Plus and other fairness metrics. This means disaggregating performance (TPR, FPR, and calibration) by age, ethnicity, language, deprivation, and their multidimensional aspects to identify and mitigate differential risks. Without such reporting, disparities will remain hidden.

The third priority is causal evaluation designs. Observational signals are helpful, but causality demands rigorous designs. Interrupted time series around AI deployment and RCTs, where feasible (to attribute safety effects more definitively to AI interventions), and propensity score methods, instrumental variables, or target trial emulation, where RCTs are infeasible, should be conducted.

The fourth priority is postmarket surveillance and governance infrastructure. Effective governance is not optional. Organizations should adopt frameworks such as people, process, technology, and operations to ensure structured oversight across personnel, process, technology, and operations. AI triage requires continuous monitoring, with statistically principled detection of drift, subgroup miscalibration, and emerging hazards [63].

The fifth priority is human-AI collaboration and implementation research. Research should shift from algorithm-centric evaluation to sociotechnical integration. Studies should examine how clinicians interpret, override, or trust AI suggestions; how AI supports (rather than disrupts) workflow; and how organizational culture shapes safe AI adoption. Mixed methods research combining qualitative insights with quantitative safety metrics will be critical.

Prospective evaluations in GP are challenging but feasible. Cluster-randomized trials and interrupted time series require careful handling of contamination between clinicians, cointerventions during rollout, seasonal and demand shocks, and variation in practice digital maturity. Outcome measurement also depends on data linkage (eg, EHR, urgent care, ED attendance, and diagnostic follow-up), and governance processes can slow implementation. Although such designs remain underused for AI-enabled triage, they are well-established in evaluating complex service interventions in primary care and are methodologically appropriate for this context [64,65].

Conclusions

AI triage offers potential for improving primary care efficiency, safety, and consistency, but current evidence leaves critical gaps. Without intersectional, real-world safety evaluations,

XSL•FO
RenderX

implementation is not just uncertain; it may be ethically risky and may inadvertently magnify existing health inequities. Over the coming years, this field must commit to responsible, equity-focused, system-aware evidence generation. That means embedding AI evaluation within the disordered realities of practice, governance mechanisms that ensure fairness and transparency, and human-AI systems that augment care rather than add workload. Operationally, this requires three commitments: (1) prospective, real-world evaluations in GP; (2) equity-stratified performance reporting guided by frameworks, such as PROGRESS-Plus; and (3) rigorous postmarket surveillance with drift and subgroup monitoring and WHO International Classification for Patient Safety–aligned incident reporting. Without these, deployment risks amplify inequities rather than reducing them. With these commitments, AI triage can better deliver on its potential of safer, more equitable primary care.

Acknowledgments

Generative artificial intelligence (ChatGPT, version 5.2; OpenAI) was used to support language editing for clarity and readability. All content was reviewed and revised by the authors, who take full responsibility for the final manuscript.

Funding

This work received no specific funding. EK is partly funded by the National Institute for Health and Care Research HealthTech Research Centre in Emergency and Acute Care (grant NIHR205301) and the Manchester British Heart Foundation Centre for Research Excellence (grant RE/24/130017).

Conflicts of Interest

None declared.

References

1. Sidaway-Lee K, Pereira Gray SD, Khan N, Abraham L, Evans P. GP continuity: the keystone of general practice. InnovAiT 2024 Apr 24;17(7):313-320. [doi: 10.1177/17557380241246742]
2. Hobbs FD, Bankhead C, Mukhtar T, Stevens S, Perera-Salazar R, Holt T, National Institute for Health Research School for Primary Care Research. Clinical workload in UK primary care: a retrospective analysis of 100 million consultations in England, 2007-14. Lancet 2016 Jun 04;387(10035):2323-2330 [FREE Full text] [doi: 10.1016/S0140-6736(16)00620-6] [Medline: 27059888]
3. The NHS long term plan. NHS England. 2019 Jan. URL: https://www.longtermplan.nhs.uk/ [accessed 2025-05-25]
4. Greenhalgh T, Wherton J, Shaw S, Morrison C. Video consultations for COVID-19. BMJ 2020 Mar 12;368:m998. [doi: 10.1136/bmj.m998] [Medline: 32165352]
5. Reidy C, Papoutsi C, Kc S, Gudgin B, Laverty AA, Greaves F, et al. Qualitative evaluation of the implementation and national roll-out of the NHS App in England. BMC Med 2025 Jan 21;23(1):20 [FREE Full text] [doi: 10.1186/s12916-024-03842-w] [Medline: 39838384]
6. Submissions via online consultation systems in general practice. NHS England. URL: https://digital.nhs.uk/data-and-information/publications/statistical/submissions-via-online-consultation-systems-in-general-practice [accessed 2026-01-24]
7. Parker RF, Figures EL, Paddison CA, Matheson JI, Blane DN, Ford JA. Inequalities in general practice remote consultations: a systematic review. BJGP Open 2021 Jun;5(3):040 [FREE Full text] [doi: 10.3399/BJGPO.2021.0040] [Medline: 33712502]
8. Chambers D, Cantrell AJ, Johnson M, Preston L, Baxter SK, Booth A, et al. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. BMJ Open 2019 Aug 01;9(8):e027743 [FREE Full text] [doi: 10.1136/bmjopen-2018-027743] [Medline: 31375610]
9. Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J Am Med Inform Assoc 2003;10(6):523-530 [FREE Full text] [doi: 10.1197/jamia.M1370] [Medline: 12925543]
10. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 2005 Apr 02;330(7494):765 [FREE Full text] [doi: 10.1136/bmj.38398.500764.8F] [Medline: 15767266]
11. Technologies T. Technologies to address wait times in the emergency department. Health Technologies, National Library of Medicine 2025 Jul:EN0058 [FREE Full text] [Medline: 40857378]
12. Abualruz H, Yasin I, Abu Sabra MA, Abunab HY, Azayzeh R, Zubidi Y, et al. The role of artificial intelligence in enhancing triage decisions in healthcare settings: a systematic review. Appl Nurs Res 2025 Dec;86:152024. [doi: 10.1016/j.apnr.2025.152024] [Medline: 41330654]
13. Peta D, Day A, Lugari WS, Gorman V, Ahayalimudin N, Pajo VM. Triage: a global perspective. J Emerg Nurs 2023 Nov;49(6):814-825. [doi: 10.1016/j.jen.2023.08.004] [Medline: 37925222]

XSL•FO
RenderX

14. Tahernejad A, Sahebi A, Abadi AS, Safari M. Application of artificial intelligence in triage in emergencies and disasters: a systematic review. BMC Public Health 2024 Nov 18;24(1):3203 [FREE Full text] [doi: 10.1186/s12889-024-20447-3] [Medline: 39558305]
15. Singh H, Sittig DF. Advancing the science of measurement of diagnostic errors in healthcare: the Safer Dx framework. BMJ Qual Saf 2015 Feb;24(2):103-110 [FREE Full text] [doi: 10.1136/bmjqs-2014-003675] [Medline: 25589094]
16. Smart C, Newman C, Hartill L, Bunce S, McCormick J. Workload effects of online consultation implementation from a job-characteristics model perspective: a qualitative study. BJGP Open 2022 Nov 21;7(1):BJGPO.2022.0024. [doi: 10.3399/bjgpo.2022.0024]
17. O'Neill J, Tabish H, Welch V, Petticrew M, Pottie K, Clarke M, et al. Applying an equity lens to interventions: using PROGRESS ensures consideration of socially stratifying factors to illuminate inequities in health. J Clin Epidemiol 2014 Jan;67(1):56-64. [doi: 10.1016/j.jclinepi.2013.08.005] [Medline: 24189091]
18. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019 Jan 7;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
19. Barocas S, Hardt M, Narayanan A. Fairness And Machine Learning: Limitations And Opportunities. Cambridge, MA: The MIT Press; 2023.
20. Akter S, Dwivedi YK, Sajib S, Biswas K, Bandara RJ, Michael K. Algorithmic bias in machine learning-based marketing models. J Bus Res 2022 May;144:201-216. [doi: 10.1016/j.jbusres.2022.01.083]
21. Jain A, Brooks JR, Alford CC, Chang CS, Mueller NM, Umscheid CA, et al. Awareness of racial and ethnic bias and potential solutions to address bias with use of health care algorithms. JAMA Health Forum 2023 Jun 02;4(6):e231197 [FREE Full text] [doi: 10.1001/jamahealthforum.2023.1197] [Medline: 37266959]
22. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci U S A 2020 Jun 09;117(23):12592-12594 [FREE Full text] [doi: 10.1073/pnas.1919012117] [Medline: 32457147]
23. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019 Oct;1(6):e271-e297 [FREE Full text] [doi: 10.1016/S2589-7500(19)30123-2] [Medline: 33323251]
24. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020 Mar 25;368:m689 [FREE Full text] [doi: 10.1136/bmj.m689] [Medline: 32213531]
25. Delshad S, Dontaraju VS, Chengat V. Artificial intelligence-based application provides accurate medical triage advice when compared to consensus decisions of healthcare providers. Cureus 2021 Aug;13(8):e16956 [FREE Full text] [doi: 10.7759/cureus.16956] [Medline: 34405077]
26. Ivanov O, Wolf L, Brecher D, Lewis E, Masek K, Montgomery K, et al. Improving ED emergency severity index acuity assignment using machine learning and clinical natural language processing. J Emerg Nurs 2021 Mar;47(2):265-78.e7 [FREE Full text] [doi: 10.1016/j.jen.2020.11.001] [Medline: 33358394]
27. Tyler S, Olis M, Aust N, Patel L, Simon L, Triantafyllidis C, et al. Use of artificial intelligence in triage in hospital emergency departments: a scoping review. Cureus 2024 May;16(5):e59906 [FREE Full text] [doi: 10.7759/cureus.59906] [Medline: 38854295]
28. Bernardi FA, Alves D, Crepaldi N, Yamada DB, Lima VC, Rijo R. Data quality in health research: integrative literature review. J Med Internet Res 2023 Oct 31;25:e41446 [FREE Full text] [doi: 10.2196/41446] [Medline: 37906223]
29. Wise J. Electronic consultations offer few benefits for GP practices, says study. BMJ 2017 Nov 06:j5141. [doi: 10.1136/bmj.j5141]
30. Karran EL, Cashin AG, Barker T, Boyd MA, Chiarotto A, Dewidar O, et al. Using PROGRESS-plus to identify current approaches to the collection and reporting of equity-relevant data: a scoping review. J Clin Epidemiol 2023 Nov;163:70-78 [FREE Full text] [doi: 10.1016/j.jclinepi.2023.09.017] [Medline: 37802205]
31. Steerling E, Svedberg P, Nilsen P, Siira E, Nygren J. Influences on trust in the use of AI-based triage-an interview study with primary healthcare professionals and patients in Sweden. Front Digit Health 2025 May 20;7:1565080 [FREE Full text] [doi: 10.3389/fdgth.2025.1565080] [Medline: 40463579]
32. Runciman W, Hibbert P, Thomson R, Van Der Schaaf T, Sherman H, Lewalle P. Towards an international classification for patient safety: key concepts and terms. Int J Qual Health Care 2009 Feb;21(1):18-26 [FREE Full text] [doi: 10.1093/intqhc/mzn057] [Medline: 19147597]
33. Mahlknecht A, Engl A, Piccoliori G, Wiedermann CJ. Supporting primary care through symptom checking artificial intelligence: a study of patient and physician attitudes in Italian general practice. BMC Prim Care 2023 Sep 04;24(1):174 [FREE Full text] [doi: 10.1186/s12875-023-02143-0] [Medline: 37661285]
34. Riboli-Sasco E, El-Osta A, Alaa A, Webber I, Karki M, El Asmar ML, et al. Triage and diagnostic accuracy of online symptom checkers: systematic review. J Med Internet Res 2023 Jun 02;25:e43803 [FREE Full text] [doi: 10.2196/43803] [Medline: 37266983]

XSL•FO
RenderX

35. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 07;4:50 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
36. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
37. Tuzlalı M, Baki N, Aral K, Aral CA, Bahçe E. Evaluating the performance of AI chatbots in responding to dental implant FAQs: a comparative study. BMC Oral Health 2025 Oct 08;25(1):1548 [FREE Full text] [doi: 10.1186/s12903-025-06863-w] [Medline: 41063105]
38. Kaushik R, Rapaka R. A patient-centered perspectives and future directions in AI-powered teledentistry. Discoveries (Craiova) 2024;12(4):e199. [doi: 10.15190/d.2024.18] [Medline: 40109877]
39. Chen K, Qiu L, Xie X, Bai Y. Dynasmile: Video-based smile analysis software in orthodontics. SoftwareX 2025 Feb;29:102004. [doi: 10.1016/j.softx.2024.102004]
40. Abbas Q, Jeong W, Lee SW. Explainable AI in clinical decision support systems: a meta-analysis of methods, applications, and usability challenges. Healthcare (Basel) 2025 Aug 29;13(17):2154 [FREE Full text] [doi: 10.3390/healthcare13172154] [Medline: 40941506]
41. Abdelwanis M, Alarafati HK, Tammam MM, Simsekler MC. Exploring the risks of automation bias in healthcare artificial intelligence applications: a Bowtie analysis. J Saf Sci Resil 2024 Dec;5(4):460-469. [doi: 10.1016/j.jnlssr.2024.06.001]
42. Salimparsa M, Sedig K, Lizotte DJ, Abdullah SS, Chalabianloo N, Muanda FT. Explainable AI for clinical decision support systems: literature review, key gaps, and research synthesis. Informatics 2025 Oct 28;12(4):119. [doi: 10.3390/informatics12040119]
43. Ellertsson S, Hlynsson HD, Loftsson H, Sigur Sson EL. Triaging patients with artificial intelligence for respiratory symptoms in primary care to improve patient outcomes: a retrospective diagnostic accuracy study. Ann Fam Med 2023;21(3):240-248 [FREE Full text] [doi: 10.1370/afm.2970] [Medline: 37217331]
44. Vergara PO, de Oliveira JC, Mattiello R, Montelongo A, Roman R, Katz N, et al. Accuracy of artificial intelligence for gatekeeping in referrals to specialized care. JAMA Netw Open 2025 Jun 02;8(6):e2513285 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.13285] [Medline: 40459894]
45. Wallace W, Chan C, Chidambaram S, Hanna L, Acharya A, Daniels E, et al. Evaluating the diagnostic and triage performance of digital and online symptom checkers for the presentation of myocardial infarction: a retrospective cross-sectional study. PLOS Digit Health 2024 Aug;3(8):e0000558. [doi: 10.1371/journal.pdig.0000558] [Medline: 39102377]
46. Siira E, Tyskbo D, Nygren J. Healthcare leaders' experiences of implementing artificial intelligence for medical history-taking and triage in Swedish primary care: an interview study. BMC Prim Care 2024 Jul 24;25(1):268. [doi: 10.1186/s12875-024-02516-z] [Medline: 39048973]
47. Moschogianis S, Darley S, Coulson T, Peek N, Cheraghi-Sohi S, Brown B. Seven opportunities for artificial intelligence in primary care electronic visits: qualitative study of staff and patient views. Ann Fam Med 2025 May 27;23(3):214-222 [FREE Full text] [doi: 10.1370/afm.240292] [Medline: 40425478]
48. Holden RJ, Carayon P, Gurses AP, Hoonakker P, Hundt AS, Ozok AA, et al. SEIPS 2.0: a human factors framework for studying and improving the work of healthcare professionals and patients. Ergonomics 2013;56(11):1669-1686 [FREE Full text] [doi: 10.1080/00140139.2013.838643] [Medline: 24088063]
49. Patient safety: making health care safer. World Health Organization (WHO). 2017. URL: https://www.who.int/publications/i/item/WHO-HIS-SDS-2017.11 [accessed 2025-07-14]
50. Lobach D, Sanders GD, Bright TJ, Wong A, Dhurjati R, Bristow E, et al. Enabling health care decisionmaking through clinical decision support and knowledge management. Evid Rep Technol Assess (Full Rep) 2012 Apr(203):1-784. [Medline: 23126650]
51. Darley S, Coulson T, Peek N, Moschogianis S, van der Veer SN, Wong DC, et al. Understanding how the design and implementation of online consultations affect primary care quality: systematic review of evidence with recommendations for designers, providers, and researchers. J Med Internet Res 2022 Oct 24;24(10):e37436 [FREE Full text] [doi: 10.2196/37436] [Medline: 36279172]
52. Olyaeemanesh A, Takian A, Mostafavi H, Mobinizadeh M, Bakhtiari A, Yaftian F, et al. Health Equity Impact Assessment (HEIA) reporting tool: developing a checklist for policymakers. Int J Equity Health 2023 Nov 18;22(1):241 [FREE Full text] [doi: 10.1186/s12939-023-02031-0] [Medline: 37980523]
53. Government of Canada. How to integrate intersectionality theory in quantitative health equity analysis? A rapid review and checklist of promising practices. Public Health Agency of Canada. Ottawa, ON URL: https://www.canada.ca/en/public-health/services/publications/science-research-data/how-integrate-intersectionality-theory-quantitative-health-equity-analysis.html [accessed 2026-01-24]
54. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119 [FREE Full text] [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
55. Carayon P. Sociotechnical systems approach to healthcare quality and patient safety. Work 2012;41 Suppl 1(0 1):3850-3854 [FREE Full text] [doi: 10.3233/WOR-2012-0091-3850] [Medline: 22317309]

XSL•FO
RenderX

56. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
57. Yusof MM, Kuljis J, Papazafeiropoulou A, Stergioulas LK. An evaluation framework for health information systems: human, organization and technology-fit factors (HOT-fit). Int J Med Inform 2008 Jun;77(6):386-398. [doi: 10.1016/j.ijmedinf.2007.08.011] [Medline: 17964851]
58. Chu CH, Donato-Woodger S, Khan SS, Nyrup R, Leslie K, Lyn A, et al. Age-related bias and artificial intelligence: a scoping review. Humanit Soc Sci Commun 2023 Aug 17;10(1):510. [doi: 10.1057/s41599-023-01999-y]
59. Shiwani T, Relton S, Evans R, Kale A, Heaven A, Clegg A, Ageing Data Research Collaborative (Geridata) AI group, et al. New Horizons in artificial intelligence in the healthcare of older people. Age Ageing 2023 Dec 01;52(12):afad219 [FREE Full text] [doi: 10.1093/ageing/afad219] [Medline: 38124256]
60. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453 [FREE Full text] [doi: 10.1126/science.aax2342] [Medline: 31649194]
61. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019 Apr 04;380(14):1347-1358. [doi: 10.1056/nejmra1814259]
62. Williams C, Shang D. Telehealth usage among low-income racial and ethnic minority populations during the COVID-19 pandemic: retrospective observational study. J Med Internet Res 2023 May 12;25:e43604 [FREE Full text] [doi: 10.2196/43604] [Medline: 37171848]
63. Maleki Varnosfaderani S, Forouzanfar M. The role of AI in hospitals and clinics: transforming healthcare in the 21st century. Bioengineering (Basel) 2024 Mar 29;11(4):337 [FREE Full text] [doi: 10.3390/bioengineering11040337] [Medline: 38671759]
64. Bernal JL, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol 2017 Feb 01;46(1):348-355 [FREE Full text] [doi: 10.1093/ije/dyw098] [Medline: 27283160]
65. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M, Medical Research Council Guidance. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ 2008 Sep 29;337:a1655 [FREE Full text] [doi: 10.1136/bmj.a1655] [Medline: 18824488]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
ED: emergency department
EHR: electronic health record
FPR: false-positive rate
GP: general practice
ML: machine learning
NLP: natural language processing
RCT: randomized controlled trial
SIITHIA: Strengthening the Integration of Intersectionality Theory in Health Inequality Analysis
TPR: true-positive rate
WHO: World Health Organization

Edited by A Mavragani; submitted 24.Nov.2025; peer-reviewed by KMS Islam, M Kokash, J Grosser, K Chen; comments to author 07.Jan.2026; accepted 28.Jan.2026; published 04.Mar.2026.

Please cite as:
Alamoudi A, Kontopantelis E, Zghebi SS, Brown B
AI Triage in Primary Care: Building Safer and More Equitable Real-World Evidence
J Med Internet Res 2026;28:e88396
URL: https://www.jmir.org/2026/1/e88396
doi:10.2196/88396
PMID:

©Aymn Alamoudi, Evangelos Kontopantelis, Salwa S Zghebi, Benjamin Brown. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN

XSL•FO
RenderX

1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Digital Health as an Enabler of Personalized Medicine in Ghana: Gaps, Opportunities, and Future Directions

Godsway Sackey[1,2], MSc; Babajide Owoyele[1,2], MSc; Frank E Baiden[3], Prof Dr; Stefan Konigorski[1,2,4], Dr rer nat

[1]Hasso Plattner Institute for Digital Engineering, Potsdam, Brandenburg, Germany
[2]Digital Engineering Faculty, University of Potsdam, Potsdam, 14482, Germany
[3]Fred N Binka School of Public Health, University of Health and Allied Sciences, Ho, Ghana
[4]Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, NY, United States

Corresponding Author:
Stefan Konigorski, Dr rer nat
Hasso Plattner Institute for Digital Engineering
Prof.-Dr.-Helmert-Str. 2-3
Potsdam, Brandenburg, 14482
Germany
Phone: 49 3315509 ext 4858
Email: stefan.konigorski@hpi.de

Abstract

Digital health solutions and personalized medicine are increasingly promoted as pathways to improve health care delivery in low-resource settings, including Ghana. Drawing on insights from our examination of the published literature and our engagement with digital health research in this context, we present a scholarly viewpoint on how digital health has been positioned in relation to personalized medicine in Ghana, where progress has been uneven and largely oriented toward population-level interventions. We observe that most digital health initiatives in Ghana focus on mobile health apps and health information systems that support service delivery and access, with limited translation toward truly personalized models of care. Although personalized medicine is frequently discussed as a future goal, it remains weakly operationalized in practice, and approaches such as N-of-1 trials—often cited as exemplars of individualized care—are notably absent from the existing literature. Importantly, the limited uptake of personalized approaches does not reflect a lack of relevance in Ghana, but rather the constraints of population-level digital health strategies that, while essential, have shown limited capacity to address individual heterogeneity in treatment responses, adherence, and long-term outcomes. We argue that this absence reflects structural, methodological, and policy-related challenges. At the same time, emerging digital health infrastructure, policy interest, and research capacity present opportunities to reposition digital health as an enabler of personalized medicine. Helping many single individuals through scalable digital personalized approaches may be a valuable innovative approach to public health. This viewpoint articulates key gaps, contextual constraints, and future directions, with the aim of informing researchers, policymakers, and implementers seeking to advance personalized, data-driven care in Ghana and comparable settings.

(J Med Internet Res 2026;28:e70696) doi:10.2196/70696

KEYWORDS

personalized medicine; digital health; health interventions; Ghana; N-of-1 trials

Reframing Digital Health and Personalized Medicine in Ghana

Digital health has become a central pillar of health system strengthening efforts in many low- and middle-income countries, including Ghana, where mobile health apps, electronic health records, and digital decision-support tools are increasingly integrated into service delivery [1-3]. These initiatives are commonly framed around improving access, efficiency, and continuity of care, particularly in settings constrained by limited infrastructure and health workforce capacity [4-6]. However, considerably less attention has been given to how digital health might meaningfully enable personalized models of care that respond to individual patient needs rather than population averages [6-9].

Personalized medicine, broadly understood as the tailoring of prevention, diagnosis, and treatment strategies to individual characteristics, has gained prominence in high-income settings but remains conceptually and operationally underdeveloped in many low-resource contexts. In Ghana, discussions of

XSL•FO
RenderX

personalized medicine are often aspirational, with limited clarity on how existing digital health systems could support individualized decision-making in routine practice [3,10]. As a result, digital health and personalized medicine are frequently discussed in parallel rather than as mutually reinforcing domains [11-13].

Our examination of the published literature on digital health in Ghana suggests that most existing initiatives prioritize generalized service delivery improvements, such as appointment reminders, health education, surveillance, and data management [3,14,15]. While these efforts are important, they rarely extend toward designs that support individual-level experimentation, adaptive treatment strategies, or sustained personalization over time. Notably, approaches such as N-of-1 trials—which are single-person crossover trials that are often highlighted in the personalized medicine discourse as a rigorous method for individual-level evaluation—are absent from the Ghanaian digital health literature. We contend that this absence reflects structural, methodological, and policy-related barriers rather than a lack of relevance or potential [5,10,16].

This viewpoint argues that reframing digital health as a potential enabler of personalized medicine in Ghana requires greater conceptual precision, contextual sensitivity, and deliberate alignment between technological design, research methods, and health system priorities. Rather than treating the current evidence base as a deficit, we use it as a foundation to articulate key gaps, clarify misconceptions, and outline future directions for research, policy, and implementation.

What the Existing Evidence Suggests About Digital Health and Personalized Medicine in Ghana

Our perspective is informed by a protocol-guided examination of the published literature on digital health, personalized medicine, and N-of-1 trials in Ghana. This examination followed a preregistered protocol to ensure transparency and methodological rigor, with the protocol publicly available through the Open Science Framework [17]. We systematically searched multiple bibliographic databases, including PubMed, Scopus, IEEE Xplore, the Cochrane Library, and Google Scholar, covering the period from 2000 to 2024. Retrieved records were screened and charted using predefined criteria to map intervention types, thematic focus, and the extent to which individualized approaches were operationalized in practice. This examination revealed several consistent patterns relevant to the advancement of individualized care.

First, the existing body of work is relatively small—in total, 40 relevant studies were identified across the entire search period—and unevenly distributed, with most studies concentrated on digital health interventions designed to improve access, efficiency, and service delivery at the population level. Mobile health apps and electronic health record systems dominate this landscape, reflecting national and institutional priorities around digitization and health system strengthening [6,7,18]. More specifically, mobile health interventions constituted approximately 60% of identified studies, followed by electronic health record and broader eHealth system implementations (approximately one quarter of studies). Explicitly personalized medicine applications accounted for less than 10% of publications, and no empirical applications of N-of-1 trials were identified. This distribution illustrates the strong orientation toward population-level digital service delivery, with comparatively limited operationalization of individualized care models.

Second, personalized medicine remains weakly operationalized in practice within Ghana and possibly also other low-resource settings, despite being frequently invoked in conceptual discussions. Only a small subset of the literature engages explicitly with personalization beyond stratified or protocol-driven approaches, and even fewer studies articulate how individual-level data could be used to iteratively adapt care over time. This pattern suggests that personalization is often treated as a future aspiration rather than a design principle embedded in current digital health initiatives [11,12].

Third, we did not identify any empirical applications of N-of-1 trials within the Ghanaian digital health literature, despite the theoretical alignment between digital health and individualized experimentation and despite frequent conceptual discussion of personalized medicine in the literature. N-of-1 trials are widely described in the personalized medicine literature as a rigorous approach to evaluating treatment effects at the individual level, particularly for chronic and heterogeneous conditions [19-21]. We contend that their absence in the Ghanaian context reflects structural, methodological, and regulatory barriers rather than a lack of relevance or potential.

Finally, the Ghana-based studies within the reviewed literature reflect both national investment in digital health and the way research agendas are shaped by funding priorities, policy emphasis, and data availability. While insights from comparable low-resource settings are informative, the Ghanaian case illustrates how contextual factors—such as governance structures, digital infrastructure, and workforce capacity—mediate the translation of digital health tools into more personalized models of care [22-24]. For instance, assessments of Ghana's eHealth workforce reveal that limited training in health informatics restricts the ability of frontline workers to utilize digital tools for complex, adaptive decision-making beyond basic data entry. Similarly, the fragmentation of electronic health record systems often prevents the interoperability needed to generate the longitudinal patient profiles essential for personalized care [3,5]. Taken together, these observations provide a foundation for rethinking how digital health initiatives might evolve beyond generalized interventions toward approaches that better support individual-level decision-making.

Although this analysis is grounded in evidence from Ghana, several of the patterns identified are likely to extend to other low- and middle-income settings pursuing digital health-enabled health system strengthening. Across studies in Ghana and comparable low- and middle-income countries, digital health interventions are frequently designed as fixed, protocol-driven tools aimed at standardized service delivery, often embedded within predefined reporting and administrative workflows

XSL•FO
RenderX

[6,7,25]. Although appropriate for scaling essential services, such designs rarely incorporate mechanisms for systematic within-person learning or adaptive treatment refinement over time. While these approaches have delivered important gains, they have also revealed shared limitations in addressing individual heterogeneity in disease trajectories, treatment response, and long-term care needs, particularly as health systems face a growing burden of chronic and non-communicable conditions [26,27]. For example, mobile health interventions for noncommunicable diseases often rely on standardized SMS text messaging protocols that do not dynamically adapt to a patient's specific adherence barriers or changing health status [26]. Similarly, clinical decision support systems used in maternal care are frequently optimized for strict guideline adherence, often missing the unique, evolving risk profiles of individual patients [28].

The Ghanaian case thus offers broader insight into how digital health investments, when primarily oriented toward population-level objectives, may inadvertently constrain the development of more individualized models of care. Challenges related to workforce capacity, data governance, and the use of routinely collected digital data for adaptive decision-making recur across many resource-constrained settings. At the same time, Ghana's relatively mature digital health ecosystem illustrates how existing infrastructure could be leveraged to support more personalized approaches without requiring fundamentally new systems. These insights may therefore inform efforts in other settings seeking to move beyond generalized digital health interventions toward more personalized, data-driven care.

Implications for Research, Policy, and Practice

The patterns observed in the existing literature have important implications for how digital health and personalized medicine are conceptualized and pursued in Ghana. First, there is a need to move beyond viewing digital health solely as a mechanism for improving access and efficiency toward recognizing its potential to support adaptive, patient-centered care. This shift requires greater attention to how data generated through routine digital health interventions can be leveraged to inform individual-level decision-making over time, rather than serving primarily administrative or surveillance functions [25,29].

Beyond methodological considerations, the existing literature points to a set of recurring system-level constraints that limit the translation of digital health into more personalized models of care in Ghana. Across studies, digital health interventions are frequently designed as fixed, protocol-driven tools aimed at standardized service delivery, with limited flexibility for individual-level adaptation over time [14,28,30]. Data generated through these systems are often used primarily for reporting, surveillance, or administrative purposes, rather than as inputs for ongoing clinical decision-making at the individual level. This design orientation constrains the ability of digital platforms to support learning about what works for whom, even when longitudinal data are technically available.

In addition, workforce capacity, data governance arrangements, and regulatory uncertainty emerge as cross-cutting barriers. Health workers are rarely trained to interpret or act on individual-level digital data beyond predefined workflows, and guidance on secondary data use, experimentation, and adaptive care remains limited. These constraints are not unique to Ghana, but they are amplified in resource-constrained settings where incentives favor scale and coverage over experimentation and iteration. As a result, even well-established digital health systems may inadvertently reinforce standardized care pathways rather than enabling personalization.

Importantly, these barriers help explain the absence of approaches such as N-of-1 trials in the Ghanaian literature. Their lack of uptake reflects structural and institutional conditions rather than conceptual misalignment or lack of potential. In particular, the rigid architecture of current electronic health record systems is primarily designed for aggregate reporting to national databases, making it difficult to integrate the granular, high-frequency patient feedback loops that N-of-1 trials require [2,3]. Addressing these constraints—through targeted capacity building, clearer governance frameworks, and deliberate design choices—may therefore be as important as introducing new technologies when seeking to advance personalized, data-driven care.

From a research perspective, advancing personalized medicine in Ghana will require methodological innovation that is sensitive to local constraints. While randomized controlled trials remain important, alternative designs such as pragmatic trials, adaptive studies, and N-of-1 trials may offer more feasible and contextually appropriate pathways for evaluating individualized interventions in resource-constrained settings [19,20]. Building capacity in these methods, alongside strengthening data governance and ethical oversight, will be critical to ensuring rigor and trust. Calls for innovative evidence-generation approaches in digital health emphasize the need for pragmatic and adaptive evaluation models that reflect real-world implementation constraints rather than relying exclusively on traditional randomized controlled trials [6]. In resource-constrained settings, digitally supported N-of-1 designs using open-source platforms such as the StudyU platform [21] may represent one such approach, offering methodological rigor while remaining scalable and embedded within routine service delivery. Aligning digital infrastructure with adaptive research designs could therefore bridge the current gap between population-scale digital health implementation and genuinely individualized care.

Global analyses of digital health ethics and governance similarly emphasize that regulatory frameworks often prioritize data protection, reporting compliance and system interoperability rather than enabling responsible secondary use of routinely collected data for adaptive care or individual-level experimentation [29,31]. Without clearer guidance on permissible experimental designs within routine care settings, digital platforms risk remaining administratively efficient but clinically static. In Ghana, ongoing digital health and health information system reforms provide an opportunity to embed personalization principles into national strategies rather than treating them as add-on innovations [6,31].

XSL•FO
RenderX

Finally, implementers and health system leaders must grapple with the practical realities of integrating personalized approaches into routine care. This includes aligning digital tools with clinical workflows, supporting health workers through training and decision support, and ensuring that personalization efforts do not exacerbate existing inequities. Without deliberate attention to these issues, digital health risks reinforcing standardized, one-size-fits-all models of care, even as it expands technological capacity [24].

Concluding Perspective

A recurring critique of personalized approaches in low-resource settings, including Ghana, is that they appear misaligned with urgent population-level public health needs. This concern is understandable, given the central role of population-based strategies in expanding coverage, improving equity, and strengthening health systems. However, Ghana's experience with digital health over the past decade suggests that population-level approaches alone—while necessary—have not been sufficient to address persistent challenges related to individual variability in disease trajectories, treatment response, adherence, and long-term care, particularly for chronic and noncommunicable conditions. Many digital health interventions remain designed around standardized protocols and average effects, which can obscure unmet needs at the individual level even when population-level gains are achieved.

In this context, personalized approaches should not be viewed as competing with public health priorities but as a complementary evolution enabled by digital health. By leveraging routinely collected individual-level data, scalable digital personalized approaches offer a pathway to improve effectiveness without sacrificing reach or equity. Helping many single individuals through such scalable personalized strategies may therefore represent a pragmatic extension of public health practice in Ghana rather than a departure from it.

Digital health has already transformed aspects of health care delivery in Ghana [3,32], yet its potential to support personalized medicine remains largely unrealized. By reframing existing digital health investments as platforms for individualized, adaptive care—and by addressing the methodological, policy, and system-level barriers that currently constrain personalization—there is an opportunity to move beyond incremental gains toward more patient-centered models of care. This viewpoint calls for a deliberate shift in how digital health is conceptualized, evaluated, and implemented, positioning personalized medicine not as a distant aspiration but as an achievable goal within Ghana's evolving health system.

In this setting, N-of-1 trials provide a practical illustration of how personalized approaches could be implemented through digital health without undermining population-level priorities. Although such designs have not yet been applied in Ghana, their absence reflects structural, methodological, and system-level constraints rather than limited relevance. Examining whether and how digitally supported N-of-1 trials could be adapted to Ghana's health system represents a logical next step for research and implementation.

Data Availability

The data underlying this study consist of bibliographic records identified through systematic database searches covering the period from January 2000 to April 2024. The full search strategy is provided in Multimedia Appendix 1. Extracted study characteristics and the cleaned integrated dataset generated during the literature examination are available from the corresponding author upon request. No individual-level participant data were generated or analyzed in this study.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Authors' Contributions

GS, BO, and SK conceptualized the viewpoint and defined its overall scope and aims. GS led the examination and synthesis of the existing literature and drafted the initial manuscript. BO, FEB, and SK provided scholarly oversight and critical input throughout the development of the argument. All authors contributed to refining the perspectives presented, critically reviewed the manuscript, and approved the final version for submission.

GS is co-corresponding author. He can be reached at Godsway.Sackey@student.hpi.uni-potsdam.de.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy and protocol details.
[PDF File (Adobe PDF File), 73 KB - jmir_v28i1e70696_app1.pdf]

References

XSL•FO
RenderX

1. Amoakoh HB, Klipstein-Grobusch K, Amoakoh-Coleman M, Agyepong IA, Kayode GA, Sarpong C, et al. The effect of a clinical decision-making mHealth support system on maternal and neonatal mortality and morbidity in Ghana: study protocol for a cluster randomized controlled trial. Trials 2017 Apr 04;18(1):157 [FREE Full text] [doi: 10.1186/s13063-017-1897-4] [Medline: 28372580]
2. Lee S, Lee YJ, Kim S, Choi W, Jeong Y, Rhim NJ, et al. Perceptions on data quality, use, and management following the adoption of tablet-based electronic health records: results from a pre–post survey with district health officers in Ghana. JMDH 2022 Jul 12;15:1457-1468 [FREE Full text] [doi: 10.2147/jmdh.s368704]
3. Achampong EK. Implementation of electronic health record system in Ghana: a review. TOPHJ 2022 Oct 10;15:e187494452208181 [FREE Full text] [doi: 10.2174/18749445-v15-e2208181]
4. Long LA, Pariyo G, Kallander K. Digital technologies for health workforce development in low- and middle-income countries: a scoping review. Glob Health Sci Pract 2018 Oct 10;6(Supplement 1):S41-S48. [doi: 10.9745/ghsp-d-18-00167]
5. Ogoe HA, Asamani JA, Hochheiser H, Douglas GP. Assessing Ghana's eHealth workforce: implications for planning and training. Hum Resour Health 2018 Nov 27;16(1):65 [FREE Full text] [doi: 10.1186/s12960-018-0330-8] [Medline: 30482223]
6. Labrique AB, Wadhwani C, Williams KA, Lamptey P, Hesp C, Luk R, et al. Best practices in scaling digital health in low and middle income countries. Global Health 2018 Nov 03;14(1):103 [FREE Full text] [doi: 10.1186/s12992-018-0424-z] [Medline: 30390686]
7. Karamagi HC, Muneene D, Droti B, Jepchumba V, Okeibunor JC, Nabyonga J, et al. eHealth or e-Chaos: The use of digital health interventions for health systems strengthening in Sub-Saharan Africa over the last 10 years: a scoping review. J Glob Health 2022 Dec 03;12:04090 [FREE Full text] [doi: 10.7189/jogh.12.04090] [Medline: 36462201]
8. Fatehi F, Samadbeik M, Kazemi A. What is digital health? Review of definitions. Stud Health Technol Inform 2020 Nov 23;275:67-71. [doi: 10.3233/shti200696]
9. Holeman I, Cookson T, Pagliari C. Digital technology for health sector governance in low-and middle-income countries. J Glob Health 2016 Dec;6(2):020408. [doi: 10.7189/jogh.06.020408]
10. Kesse-Tachi A, Asmah AE, Agbozo E. Factors influencing adoption of eHealth technologies in Ghana. Digit Health 2019;5:2055207619871425 [FREE Full text] [doi: 10.1177/2055207619871425] [Medline: 31523448]
11. Mathur S, Sutton J. Personalized medicine could transform healthcare. Biomed Rep 2017 Jul;7(1):3-5 [FREE Full text] [doi: 10.3892/br.2017.922] [Medline: 28685051]
12. Shabaruddin FH, Fleeman ND, Payne K. Economic evaluations of personalized medicine: existing challenges and current developments. Pharmgenomics Pers Med 2015 Jun 24;8:115-126. [Medline: 26309416]
13. Mitropoulos K, Cooper DN, Mitropoulou C, Agathos S, Reichardt JKV, Al-Maskari F, et al. Genomic medicine without borders: which strategies should developing countries employ to invest in precision medicine? A new "fast-second winner" strategy. OMICS 2017 Nov;21(11):647-657 [FREE Full text]
14. Mohammed A, Acheampong PR, Otupiri E, Osei FA, Larson-Reindorf R, Owusu-Dabo E. Mobile phone short message service (SMS) as a malaria control tool: a quasi-experimental study. BMC Public Health 2019 Aug 29;19(1):1193 [FREE Full text] [doi: 10.1186/s12889-019-7336-6] [Medline: 31464623]
15. Rokicki S, Fink G. Assessing the reach and effectiveness of mHealth: evidence from a reproductive health program for adolescent girls in Ghana. BMC Public Health 2017 Dec 20;17:969 [FREE Full text] [doi: 10.1186/s12889-017-4939-7] [Medline: 29262823]
16. Barkman C, Weinehall L. Policymakers and mHealth: roles and expectations, with observations from Ethiopia, Ghana and Sweden. Glob Health Action 2017 Jun;10(Suppl3):1337356. [Medline: 28838303]
17. Exploring digital health solutions: personalised medicine and n-of-1 trials in Ghana: a scoping review. Open Science Framework. URL: https://osf.io/n2krw/overview [accessed 2026-04-08]
18. Abul-Husn NS, Kenny EE. Personalized medicine and the power of electronic health records. Cell 2019 Mar 21;177(1):58-69 [FREE Full text] [doi: 10.1016/j.cell.2019.02.039] [Medline: 30901549]
19. Lillie E, Patay B, Diamant J, Issell B, Topol EJ, Schork NJ. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? Per Med 2011 Mar;8(2):161-173 [FREE Full text] [doi: 10.2217/pme.11.7] [Medline: 21695041]
20. Alemayehu C, Nikles J, Mitchell G. N-of-1 trials in the clinical care of patients in developing countries: a systematic review. Trials 2018 Apr 23;19(1):246 [FREE Full text] [doi: 10.1186/s13063-018-2596-5] [Medline: 29685163]
21. Konigorski S, Wernicke S, Slosarek T, Zenner A, Strelow N, Ruether DF, et al. StudyU: a platform for designing and conducting innovative digital N-of-1 trials. J Med Internet Res 2022 Jul 05;24(7):e35884 [FREE Full text] [doi: 10.2196/35884] [Medline: 35787512]
22. Dzando G, Akpeke H, Kumah A, Agada E, Lartey AA, Nortu J, et al. Telemedicine in Ghana: insight into the past and present, a narrative review of literature amidst the Coronavirus pandemic. J Public Health Afr 2022 May 24;13(1):2024 [FREE Full text] [doi: 10.4081/jphia.2022.2024] [Medline: 35720800]
23. Koduah A, Anim Boadi J, Azeez JNK, Adu Asare B, Yevutsey S, Gyansa-Lutterodt M, et al. Institutionalizing health technology assessment in ghana: enablers, constraints, and lessons. Health Syst Reform 2023 Dec 31;9(3):2314519 [FREE Full text] [doi: 10.1080/23288604.2024.2314519] [Medline: 38715200]
24. Saner H. Digital health implementation: how to overcome the barriers? Eur J Prev Cardiol 2019 Jul;26(11):1164-1165. [doi: 10.1177/2047487319848222] [Medline: 31046444]

XSL•FO
RenderX

25. Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions-A call for innovative evidence generation approaches. NPJ Digit Med 2020;3:110 [FREE Full text] [doi: 10.1038/s41746-020-00314-2] [Medline: 32904379]
26. Opoku D, Busse R, Quentin W. Achieving sustainability and scale-up of mobile health noncommunicable disease interventions in sub-Saharan Africa: Views of policy makers in Ghana. JMIR Mhealth Uhealth 2019 May 03;7(5):e11497 [FREE Full text] [doi: 10.2196/11497] [Medline: 31066706]
27. Williams K, Markwardt S, Kearney SM, Karp JF, Kraemer KL, Park MJ, et al. Addressing implementation challenges to digital care delivery for adults with multiple chronic conditions: stakeholder feedback in a randomized controlled trial. JMIR Mhealth Uhealth 2021 Feb 01;9(2):e23498 [FREE Full text] [doi: 10.2196/23498] [Medline: 33522981]
28. Amoakoh HB, Klipstein-Grobusch K, Grobbee DE, Amoakoh-Coleman M, Oduro-Mensah E, Sarpong C, et al. Using mobile health to support clinical decision-making to improve maternal and neonatal health outcomes in Ghana: insights of frontline health worker information needs. JMIR Mhealth Uhealth 2019 May 24;7(5):e12879 [FREE Full text] [doi: 10.2196/12879] [Medline: 31127719]
29. Vayena E, Haeusermann T, Adjekum A, Blasimme A. Digital health: meeting the ethical and policy challenges. Swiss Med Wkly 2018 Jan 16;148:w14571 [FREE Full text] [doi: 10.4414/smw.2018.14571] [Medline: 29376547]
30. Amoakoh HB, Klipstein-Grobusch K, Agyepong IA, Amoakoh-Coleman M, Kayode GA, Reitsma JB, et al. Can an mhealth clinical decision-making support system improve adherence to neonatal healthcare protocols in a low-resource setting? BMC Pediatr 2020 Nov 27;20(1):534 [FREE Full text] [doi: 10.1186/s12887-020-02378-1] [Medline: 33243172]
31. Kostkova P. Grand challenges in digital health. Front Public Health 2015 May 5;3:134 [FREE Full text] [doi: 10.3389/fpubh.2015.00134] [Medline: 26000272]
32. Rooney L, Rimpiläinen S, Morrison C, Nielsen SL. Review of emerging trends in digital health and care. Digital Health and Care Institute. Glasgow: University of Strathclyde 2018 Nov 20:1 [FREE Full text] [doi: 10.17868/67860]

Edited by A Schwartz; submitted 30.Dec.2024; peer-reviewed by KL Mauco, S Mitra; comments to author 03.Jul.2025; revised version received 26.Feb.2026; accepted 10.Mar.2026; published 16.Apr.2026.

Please cite as:
Sackey G, Owoyele B, Baiden FE, Konigorski S
Digital Health as an Enabler of Personalized Medicine in Ghana: Gaps, Opportunities, and Future Directions
J Med Internet Res 2026;28:e70696
URL: https://www.jmir.org/2026/1/e70696
doi:10.2196/70696

©Godsway Sackey, Babajide Owoyele, Frank E Baiden, Stefan Konigorski. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Rethinking Trust in Synthetic Health Data: Lessons From 7 European Research Initiatives

Jens Declerck[1,2], MSc; Dipak Kalra[1], Prof Dr; Antti Airola[3]; Ahmed Youssef Ali Amer[4]; Christos Chatzichristos[5]; Maria del Mar Mañu Pereira[6]; Bruno M. de Brito Robalo[7]; Francesco Ghini[8]; Alberto Gutierrez-Torre[9]; Sem Hoogteijling[10]; Susanne Hultsch[11]; Jan Ramon[12]; Sara Reidel[13,14]; Francesco Regazzoni[15,16]; Luís Silva[17]; Inês Silveira[17]; Tsekeridou Sofia[18], PhD; Christophe Maes[1,2]

[1]The European Institute for Innovation through Health Data, Oosterzele, Belgium
[2]Unit of Medical Informatics and Statistics, Department of Public Health and Primary Care, The University of Ghent, Ghent, Belgium
[3]Department of Computing, University of Turku, Turku, Finland
[4]Innovative Medicine, Johnson & Johnson, Beerse, Belgium
[5]Department of Electrical Engineering, KU Leuven, Leuven, Belgium
[6]University Hospital Vall d'Hebron, Barcelona, Spain
[7]Department of Obstetrics and Gynaecology, Erasmus MC, University Medical Center, Rotterdam, The Netherlands
[8]Laboratory of Translational Research, Reggio Emilia, Italy
[9]Barcelona Supercomputing Center, Barcelona, Catalonia, Spain
[10]Department of Neurology and Neurosurgery, University Medical Center Utrecht Brain Center, Utrecht, The Netherlands
[11]VEIL.AI, Helsinki, Finland
[12]Institut national de recherche en sciences et technologies du numérique, Lille, France
[13]Vall d'Hebron Research Institute, University Hospital Vall d'Hebron, Barcelona, Spain
[14]BarcelonaTech (UPC), Universitat Politècnica de Catalunya, Barcelona, Spain
[15]University of Amsterdam, Amsterdam, The Netherlands
[16]Università della Svizzera italiana, Lugano, Ticino, Switzerland
[17]Laboratory for Instrumentation, Biomedical Engineering and Radiation Physics, NOVA School of Science and Technology, Caparica, Portugal
[18]Netcompany BELUX, Athens, Greece

Corresponding Author:
Jens Declerck, MSc
The European Institute for Innovation through Health Data
Merebaaistraat 10
Oosterzele, 9860
Belgium
Phone: 32 0474538199
Email: jens.declerck@i-hd.eu

Abstract

Synthetic data generation (SDG) structured health data is increasingly promoted as a solution to longstanding barriers in health data access. It is offering the promise of privacy-preserving data reuse for research, innovation, and policy. Despite rapid technical advances, the adoption of synthetic health data in real-world settings remains limited. Shaped by challenges around data quality, representativeness, infrastructure readiness, trust, and legal uncertainty, this viewpoint draws on experiences from 7 European research initiatives within the HealthData4EU cluster to reflect on how SDG is being operationalized in practice. It synthesizes cross-project insights to highlight recurring methodological and governance tensions and to examine their implications for trust and responsible use. The analysis argues that trustworthy SDG cannot be achieved through technical optimization alone but requires alignment between evaluation practices, upstream data stewardship, regulatory clarity, and sustained stakeholder engagement. Addressing these conditions is essential for moving synthetic data from experimental pilots toward a credible and sustainable component of European health research ecosystems.

(J Med Internet Res 2026;28:e83369) doi:10.2196/83369

XSL•FO
RenderX

KEYWORDS

health data quality; synthetic data generation; health data; data quality dimensions; privacy; project

Introduction

As health care systems are largely digitized and become increasingly data-driven, access to large-scale, high-quality data is essential for clinical research [1-3], digital health innovation [4], and evidence-based policymaking [5]. However, accessing patient-level health data remains a persistent challenge [6]. Privacy regulations [7], data quality issues [8,9], and data fragmentation [10] continue to limit data access and sharing, especially in cross-organizational studies [9,11]. These barriers not only slow innovation but also risk underrepresenting certain diseases and populations, such as rare conditions, in secondary data use. As a result, many promising analytical and clinical applications struggle to move beyond pilots.

Synthetic data generation (SDG) has emerged as a promising response to the barriers limiting access and reuse of patient-level health data [12]. By producing synthetic datasets that resemble the statistical and structural properties of real health data, without exposing personal identities, SDG can drive innovation while maintaining privacy [7]. It enables safer collaboration across different types of health data (eg, electronic health records, genomics, medical imaging) and offers value in areas such as artificial intelligence (AI) development [7,13], rare disease research [14], and model testing [13,15]. The main aim of SDG is to retain the essential characteristics of original datasets while retaining privacy (in most use cases). These can be grouped into quantitative [16] (eg, statistical metrics, analytical patterns, signal-based features), qualitative [16] (eg, expert-driven assessments), and domain-specific attributes [17] that depend on the data type and specific use case. Conceptual work has also emphasized the importance of incorporating domain knowledge into SDG processes to improve realism and relevance for clinical applications [18]. Despite this growing interest, the adoption of synthetic health data in real-world research and clinical environments remains limited [12,19]. Concerns persist regarding data quality, representativeness, interpretability, and legal status [20,21], as well as the uncertainty about how synthetic data exceed what current methods and data infrastructures can deliver [13]. This gap between promise and practice has contributed to skepticism among multiple stakeholders, for whom trust in data sources is essential [22]. This lack of trust is partly explained by the focus of SDG literature on technical methods and evaluation metrics [7,22,23] while giving limited attention to the organizational, legal, and sociotechnical conditions that shape real-world adoption [7,22].

Several EU-funded initiatives are currently testing SDG in real-world health research settings, all facing a shared question: how can synthetic health data be made trustworthy, useful, and safe? Despite this activity, there has been limited cross-initiative reflection on how these challenges are being addressed in practice.

The aim of this viewpoint is to examine why trust in synthetic health data remains fragile despite growing technical maturity and to reflect on the nontechnical conditions required for its responsible adoption. This article is intended for health data researchers, clinicians, data stewards, infrastructure developers, regulators, and policymakers engaged in secondary use of health data across Europe. This article provides a viewpoint informed by experiences from 7 European research initiatives within the HealthData4Eu cluster that work with structured synthetic health data. It synthesizes cross-project insights to highlight recurring challenges related to data quality, interoperability, regulatory alignment, and stakeholder acceptance and reflects on their implications for trust, governance, and responsible use of synthetic data in European health research.

Basis of This Viewpoint

Scope and Context of the Included Projects

This viewpoint is informed by 7 EU-funded projects that constitute the HealthData4EU cluster, a Horizon Europe initiative focused on advancing SDG and secure data use in health research across Europe. Collectively, these projects address complementary aspects of the synthetic data landscape, including data generation methods, federated and privacy-preserving architectures, secure data sharing, and AI-enabled clinical applications. All operate within a shared European policy and funding framework while engaging with diverse clinical and institutional contexts relevant to the European Health Data Space.

The included initiatives represent the full set of synthetic data projects participating in the HealthData4EU cluster at the time of writing. While they operate within a shared European policy and funding framework, they span a wide range of clinical domains, data modalities (eg, tabular, imaging, longitudinal, multimodal data), and technical approaches. This diversity within a coordinated structure provides a useful basis for reflecting on how SDG is being operationalized across health care settings and where common challenges and tensions emerge. An overview of the participating initiatives with their clinical domains is provided in Table 1.

XSL•FO
RenderX

Table 1. Overview of projects, clinical domains, and example use cases involving synthetic data.

Project	Clinical domain(s)	Example use cases
AISym4MED	Neurological and chronic diseases (ie, diffuse large B-cell lymphoma, lung cancer, multiple myeloma, Alzheimer disease, type 2 diabetes, breast cancer)	• Synthetic data for multimodal applications • Support for AI[a] model testing
FLUTE	Oncology (prostate cancer)	• Risk stratification • Federated AI validation for prostate cancer diagnosis
PHASE IV AI	Oncology and neurology (ie, lung cancer, prostate cancer, ischemic stroke)	• Early lung cancer detection • Synthetic magnetic resonance imaging for stroke and model testing
PHEMS	Pediatric diseases (ie, congenital cardiac conditions, sepsis, hemophilia)	• Sepsis prediction in pediatric intensive care units • Hemophilia care optimization
SECURED	Multiple domains (i.e.ie, mammography, histopathology, chest radiography, cardiotocography)	• Privacy-preserving SDG[b] • Federated AI tool development
SYNTHEMA	Rare hematological diseases (ie, sickle-cell disease, acute myeloid leukemia)	• Synthetic datasets for rare disease research • Federated validation
SYNTHIA	Multimodal and personalized medicine	• Benchmarking and validation of SDG methods across diverse clinical scenarios

[a]AI: artificial intelligence.

[b]SDG: synthetic data generation.

The insights synthesized in this paper draw on cross-project engagement conducted during the active lifetime of the initiatives. These insights were informed by cluster-level workshops, project presentations, and recurring exchanges among project teams and by review of the project documentation during the active implementation phase. Importantly, these interactions took place while projects were still ongoing, allowing challenges and emerging practices to be discussed as they arose rather than retrospectively. The exchanges involved multidisciplinary contributors, including technical developers, clinicians, data stewards, infrastructure providers, and legal and ethical experts.

While these insights are not derived form a formal empirical study, they reflect sustained, longitudinal engagement with projects that are actively developing and validating the real-world use of synthetic health data within regulated clinical and research environments. To improve transparency and address requests for additional project-level context, we have incorporated a table in Multimedia Appendix 1 providing a descriptive overview of the included projects.

Lessons From 7 SDG Projects

A Shared Landscape of Methodological Tensions

Across projects, SDG is not pursued as a single technical task but as a complex sociotechnical intervention embedded in heterogeneous health systems [22]. In practice, SDG sits at the intersection of machine learning, clinical data infrastructures, regulatory governance, and institutional trust.

Despite differences in disease focus, data modalities, and maturity, projects consistently encounter several recurring challenges that shape how SDG is operationalized in practice. Across the initiatives examined in this paper, 3 themes emerged repeatedly: the absence of shared consensus on how synthetic data quality should be defined and evaluated, the limitations of existing data infrastructures to support reliable SDG workflows, and uncertainty surrounding the regulatory and governance status of synthetic data. These challenges reflect the structural characteristics of health data ecosystems, where technical innovation must coexist with legal safeguards, clinical accountability, and fragmented infrastructures. Understanding SDG through these challenges helps explain why synthetic data adoption remains difficult to scale beyond pilot environments.

Data Quality Without Consensus

A central tension concerns how “data quality” in synthetic health data is defined and demonstrated. In practice, data quality is assessed through overlapping but fragmented lenses: statistical fidelity, analytical or clinical utility, and privacy protection. These dimensions are widely acknowledged across initiatives, yet there is no shared understanding of how they should be balanced, prioritized, or interpreted in different contexts [24].

Existing tools and metrics often assume that real-world source data are themselves unbiased, complete, and representative [3,6,8,9]. This assumption is rarely valid in health systems shaped by structural inequities, fragmented data capture, and variable clinical practices across institutions and regions. When real data reflect structural inequities or incomplete capture, synthetic data derived from them may inherit the same distortions [20].

As a result, SDG risks reinforcing existing biases rather than mitigating them. While synthetic data are frequently promoted as a mechanism for improving representativeness or enabling rare disease research, representational properties are frequently assessed primarily at the level of the resulting dataset [6]. In practice, however, representativeness is shaped not only by the

XSL•FO
RenderX

quality of the synthetic generation process but also by the characteristics of the underlying source data and the transformations applied during the data lifecycle (eg, data extraction, harmonization, mapping to common data models, other extraction transfer load steps) [25]. These can introduce or amplify biases by selectively filtering variables, standardizing heterogeneous data sources, or excluding smaller subpopulations. These upstream influences may remain difficult to detect when evaluation focuses primarily on statistical similarity. As a result, population coverage and the potential impact of data preparation processes on representation may receive less attention in current evaluation practices.

In this sense, data quality cannot be reduced to statistical resemblance alone. We argue that there is a need for a more explicit data quality assessment at source and across the data lifecycle. In that sense, results need to be interpreted in relation to the populations, decisions, and governance contexts in which synthetic data are intended to be used.

Infrastructure as a Bottleneck to Trust

Federated and decentralized SDG architectures are frequently presented as solutions to privacy and governance constraints [13]. In practice, infrastructure maturity at data sources emerges as a critical limiting factor. Heterogeneous hospital IT systems, inconsistent data models, and uneven data governance practices complicate federated training and validation. These challenges affect not only technical integration but also the reliability of data inputs and the feasibility of data quality assessments.

Interoperability challenges persist even when projects adopt common data models such as the Observational Medical Outcomes Partnership Common Data Model. Differences in local coding practices, laboratory workflows, and semantic interpretation (eg, variation in LOINC [Logical Observation Identifiers Names and Codes] mappings for similar laboratory measurements) require substantial upfront coordination. These nuances complicate federated validation and cross-site comparability, reinforcing that standardization alone does not eliminate heterogeneity in real-world health data. Within the HealthData4EU cluster, several projects reported that aligning local coding practices and laboratory mappings across institutions required substantial preprocessing effort before federated SDG or validation could be performed.

These infrastructural constraints shape what forms of SDG are feasible and who can participate, reinforcing a gap between technical potential and institutional readiness.

Trust, Transparency, and Regulatory Uncertainty

Trust is Not Guaranteed by Compliance

Most initiatives emphasize compliance with data protection and ethical frameworks. While compliance is necessary, it is not sufficient to generate trust. Trust emerges through transparency, interpretability, and meaningful stakeholder engagement across the SDG lifecycle [22].

In practice, these dimensions are unevenly operationalized. Some initiatives invest in dataset documentation, model cards, and participatory validation involving clinicians or data stewards. Others limit engagement to formal approval processes or internal review. This variability affects adoption, particularly in clinical and regulatory contexts where accountability and explainability are essential.

In the European context, this includes emerging obligations under the EU Artificial Intelligence Act, specific for high-risk medical AI systems, where the provenance and validation of training data, including synthetic data, may become subject to increased scrutiny.

The Unresolved Legal Status of Synthetic Data

Regulatory ambiguity remains one of the most significant barriers to uptake. Synthetic data are often assumed to fall outside data protection regulation, yet institutions frequently adopt conservative interpretations, particularly of rare diseases or small cohorts [26]. The absence of harmonized guidance leaves data controllers and ethics boards navigating uncertainty case by case.

Clarification is particularly needed in relation to General Data Protection Regulation (GDPR) anonymization thresholds, the interaction with the EU Artificial Intelligence Act, and national interpretations by data protection authorities and ethics committees.

This uncertainty raises a fundamental but often unspoken question: what is the value of synthetic data if it cannot be confidently reused beyond its original context? If synthetic datasets remain confined to local pilots due to legal or institutional caution, their potential to support cross-border research, reproducibility, and capacity building is limited. Addressing reusability therefore becomes central to the trust debate, linking legal clarity, documentation, and transparent validation of the broader promise of synthetic data as a shared research asset rather than a one-off technical item.

Emerging Responses and Their Limits

Encouragingly, initiatives are beginning to respond through federated validation pipelines, benchmarking libraries, interoperability standards, and public-facing platforms. However, these responses remain fragmented, and no shared consensus has yet emerged on what constitutes trustworthy SDG across contexts.

Reflections and Recommendations

The experiences discussed in this article suggest that SDG in health research has reached a point where technical feasibility is no longer the primary bottleneck. Instead, progress increasingly depends on how synthetic data are evaluated, governed, communicated, and trusted in practice.

First, there is a need to reframe how success in SDG is defined and assessed. Much of the current emphasis remains on demonstrating technical performance, such as statistical similarity or model accuracy. While these measures are important, they do not on their own indicate whether synthetic data are appropriate for research or operational contexts. In practice, synthetic datasets may serve different purposes (eg, exploratory analysis, educational use, or hypothesis testing). These uses imply different expectations regarding the level of fidelity, reliability, and validation required. Clarifying the

XSL•FO
RenderX

intended context of use can therefore help guide how synthetic datasets are evaluated and interpreted. This does not imply that all potential analytical applications must be evaluated in advance, but rather that evaluation strategies should be transparent about the contexts for which synthetic datasets are considered suitable. At the same time, general evaluation metrics remain essential to characterize the overall fidelity and privacy properties of synthetic datasets. In many cases, findings derived from exploratory analysis using synthetic data may subsequently require validation using the original data sources where appropriate access mechanisms exist.

Second, greater attention should be given to data stewardship and governance upstream of SDG. Synthetic data reflects the characteristics and limitations of the infrastructures from which it is generated. Without robust practices for data quality management, interoperability, and documentation at source, SDG risks perpetuating existing shortcomings rather than addressing them. Aligning SDG initiatives with broader efforts to strengthen health data infrastructures would therefore increase both the reliability of synthetic outputs and institutional confidence in their reuse.

Third, addressing the legal uncertainty surrounding synthetic health data is essential for wider adoption. In the absence of shared guidance, institutions often adopt cautious positions that limit reuse, particularly in sensitive or cross-border contexts. Developing common interpretations, best practices, or certification pathways that clarify when synthetic data can be considered sufficiently anonymized would provide a more stable foundation for decision-making. Such approaches would not eliminate the need for contextual oversight, but they would reduce fragmentation and support more consistent application of safeguards.

Finally, trust building must be recognized as an active and ongoing process. Trust does not arise automatically from compliance with legal or ethical requirements nor from technical sophistication alone. It depends on transparency, interpretability, and the ability of stakeholders to understand how synthetic data were generated, evaluated, and intended to be used. Embedding mechanisms such as clear documentation, accessible validation summaries, and engagement with clinicians, data stewards, and other end users can support accountability and enable more nuanced judgments about risk and value.

Overall, these reflections point to a broader shift in focus: from demonstrating that synthetic data can be generated to establishing the conditions under which it can be used responsibly, communicated transparently, and governed consistently. Advancing SDG along this path will require coordination across technical, institutional, and regulatory domains. If achieved, synthetic data can become not only a powerful analytical tool but also a credible and trusted component of health research ecosystems.

Acknowledgments

The authors would like to thank all 7 HealthData4EU projects for their valuable input, collaboration, and contributions to this paper. Their openness and insights were essential to shaping the cross-case analysis. The authors are also grateful to Iraia Nuñez and Anna Lorenzini for their dedicated coordination and communications support. While not involved in the technical content of this study, their efforts in organizing exchanges and facilitating collaboration across the projects played a vital role in enabling this synthesis. No generative artificial intelligence tools were used in the generation of this paper.

Funding

This paper was funded by the Innovative Health Initiative through the SYNTHIA project. The study draws on insights from 7 research projects that form part of the HealthData4EU cluster. Projects within the cluster and their funding sources are as follows. SYNTHEMA, PHASE IV AI, SECURED, FLUTE, and AISym4MED are funded by the Horizon Europe program. PHEMS is funded by Horizon Europe and UK Research and Innovation. SYNTHIA is funded through the Innovative Health Initiative.

Conflicts of Interest

SH is employed by VEIL.AI, a company developing data anonymization technologies. The work presented in this manuscript was conducted as part of the PHEMS consortium and was not influenced by the employer, and SH declares no direct financial or commercial interests related to this publication. The other authors have no conflicts of interest to declare.

Multimedia Appendix 1
Overview of all HealthData4EU projects, their objectives, data modalities, use cases, and methodological challenges.
[DOCX File , 43 KB - jmir_v28i1e83369_app1.docx]

References

1. Eden R, Burton-Jones A, Scott I, Staib A, Sullivan C. Effects of eHealth on hospital practice: synthesis of the current literature. Aust Health Rev 2018 Sep;42(5):568-578. [doi: 10.1071/AH17255] [Medline: 29986809]
2. Zheng K, Abraham J, Novak LL, Reynolds TL, Gettinger A. A survey of the literature on unintended consequences associated with health information technology: 2014–2015. Yearb Med Inform 2018 Mar 06;25(01):13-29. [doi: 10.15265/iy-2016-036]

XSL•FO
RenderX

3. Declerck J, Lee J, Sen A, Palmeri A, Oostenbrink R, Giannuzzi V, et al. The potential to leverage real-world data for pediatric clinical trials: a proof-of-concept study. J Med Internet Res 2025 May 30;27:e72573 [FREE Full text] [doi: 10.2196/72573] [Medline: 40446289]
4. Wang Z, Penning M, Zozus M. Analysis of anesthesia screens for rule-based data quality assessment opportunities. Stud Health Technol Inform 2019;257:473-478 [FREE Full text] [Medline: 30741242]
5. Wiebe N, Xu Y, Shaheen AA, Eastwood C, Boussat B, Quan H. Indicators of missing electronic medical record (EMR) discharge summaries: a retrospective study on Canadian data. Int J Popul Data Sci 2020 Dec 11;5(1):1352 [FREE Full text] [doi: 10.23889/ijpds.v5i3.1352] [Medline: 34007880]
6. Declerck J, Kalra D, Vander Stichele R, Coorevits P. Frameworks, dimensions, definitions of aspects, and assessment methods for the appraisal of quality of health data for secondary use: comprehensive overview of reviews. JMIR Med Inform 2024 Mar 06;12:e51560 [FREE Full text] [doi: 10.2196/51560] [Medline: 38446534]
7. Shabani M, Marelli L. Re-identifiability of genomic data and the GDPR: assessing the re-identifiability of genomic data in light of the EU General Data Protection Regulation. EMBO Rep 2019 Jun;20(6):e48316 [FREE Full text] [doi: 10.15252/embr.201948316] [Medline: 31126909]
8. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC) 2016;4(1):1244 [FREE Full text] [doi: 10.13063/2327-9214.1244] [Medline: 27713905]
9. Declerck J, Vandenberk B, Deschepper M, Colpaert K, Cool L, Goemaere J, et al. Building a foundation for high-quality health data: multihospital case study in Belgium. JMIR Med Inform 2024 Dec 20;12:e60244 [FREE Full text] [doi: 10.2196/60244] [Medline: 39727158]
10. Wei W, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, Chai HS, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. J Am Med Inform Assoc 2012;19(2):219-224 [FREE Full text] [doi: 10.1136/amiajnl-2011-000597] [Medline: 22249968]
11. Oja M, Tamm S, Mooses K, Pajusalu M, Talvik H, Ott A, et al. Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model: lessons learned. JAMIA Open 2023 Dec;6(4):ooad100 [FREE Full text] [doi: 10.1093/jamiaopen/ooad100] [Medline: 38058679]
12. Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: a narrative review. PLOS Digit Health 2023 Jan;2(1):e0000082 [FREE Full text] [doi: 10.1371/journal.pdig.0000082] [Medline: 36812604]
13. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. NPJ Digit Med 2023 Oct 09;6(1):186 [FREE Full text] [doi: 10.1038/s41746-023-00927-3] [Medline: 37813960]
14. Al-Dhamari I, Abu Attieh H, Prasser F. Synthetic datasets for open software development in rare disease research. Orphanet J Rare Dis 2024 Jul 15;19(1):265 [FREE Full text] [doi: 10.1186/s13023-024-03254-2] [Medline: 39010138]
15. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. BMC Med Res Methodol 2020 May 07;20(1):108 [FREE Full text] [doi: 10.1186/s12874-020-00977-1] [Medline: 32381039]
16. Park N, Mohammadi M, Gorde K, Jajodia S, Park H, Kim Y. Data synthesis based on generative adversarial networks. Proc VLDB Endow 2018 Jun;11(10):1071-1083. [doi: 10.14778/3231751.3231757]
17. Esteban C, Hyland S, Rätsch G. Real-valued (medical) time series generation with recurrent conditional gans. ArXiv Preprint posted online on December 4, 2017. [doi: 10.48550/arXiv.1706.02633]
18. Hashemi A, Soliman A, Lundström J, Etminani K. Domain knowledge-driven generation of synthetic healthcare data. In: Volume 302: Caring is Sharing – Exploiting the Value in Data for Health and Innovation. Amsterdam, Netherlands: IOS Press eBooks; May 22, 2023:352-353.
19. Ogwel B, Mzazi VH, Awuor AO, Otieno G, Ogolla S, Nyawanda BO, et al. A quarter-century of synthetic data in healthcare: Unveiling trends with structural topic modeling. Digit Health 2025;11:20552076251404530 [FREE Full text] [doi: 10.1177/20552076251404530] [Medline: 41346942]
20. Alaa A, Van BB, Saveliev E, Van DSM. editors. arXiv 2022 Jul 13:1-17. [doi: 10.48550/arXiv.2102.08921]
21. Stenger M, Leppich R, Foster I, Kounev S, Bauer A. Evaluation is key: a survey on evaluation measures for synthetic time series. J Big Data 2024 May 07;11(1):66. [doi: 10.1186/s40537-024-00924-7]
22. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
23. Jordon J, Szpruch L, Houssiau F, Bottarelli M, Cherubin G, Maple C. Synthetic data -- what, why and how? ArXiv Preprint posted online on May 6, 2022. [doi: 10.48550/arXiv.2205.03257]
24. Kaabachi B, Despraz J, Meurers T, Otte K, Halilovic M, Kulynych B, et al. A scoping review of privacy and utility metrics in medical synthetic data. NPJ Digit Med 2025 Jan 27;8(1):60 [FREE Full text] [doi: 10.1038/s41746-024-01359-3] [Medline: 39870798]
25. Liaw S, Guo JGN, Ansari S, Jonnagaddala J, Godinho MA, Borelli AJ, et al. Quality assessment of real-world data repositories across the data life cycle: a literature review. J Am Med Inform Assoc 2021 Jul 14;28(7):1591-1599 [FREE Full text] [doi: 10.1093/jamia/ocaa340] [Medline: 33496785]

XSL•FO
RenderX

26. Pilgram L, Ko H, Tung A, El Emam K. Protecting patient privacy in tabular synthetic health data: a regulatory perspective. NPJ Digit Med 2025 Nov 28;8(1):732 [FREE Full text] [doi: 10.1038/s41746-025-02112-0] [Medline: 41315669]

Abbreviations

AI: artificial intelligence
GDPR: General Data Protection Regulation
LOINC: Logical Observation Identifiers Names and Codes
SDG: synthetic data generation

Edited by A Mavragani; submitted 01.Sep.2025; peer-reviewed by D Liu, M Deschepper, L Busatta, N Jackson, A Hashemi; comments to author 29.Oct.2025; revised version received 04.Mar.2026; accepted 24.Mar.2026; published 29.Apr.2026.

Please cite as:
Declerck J, Kalra D, Airola A, Amer AYA, Chatzichristos C, del Mar Mañu Pereira M, M. de Brito Robalo B, Ghini F, Gutierrez-Torre A, Hoogteijling S, Hultsch S, Ramon J, Reidel S, Regazzoni F, Silva L, Silveira I, Sofia T, Maes C
Rethinking Trust in Synthetic Health Data: Lessons From 7 European Research Initiatives
J Med Internet Res 2026;28:e83369
URL: https://www.jmir.org/2026/1/e83369
doi:10.2196/83369
PMID:

©Jens Declerck, Dipak Kalra, Antti Airola, Ahmed Youssef Ali Amer, Christos Chatzichristos, Maria del Mar Mañu Pereira, Bruno M. de Brito Robalo, Francesco Ghini, Alberto Gutierrez-Torre, Sem Hoogteijling, Susanne Hultsch, Jan Ramon, Sara Reidel, Francesco Regazzoni, Luís Silva, Inês Silveira, Tsekeridou Sofia, Christophe Maes. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Clinical AI is Not (Yet) Trustworthy-But It Could Be

Ali Saad[1*], PhD; Sofia B Dias[2*], PhD; Ghada Alhussein[3*], PhD; David Lyreskog[4], PhD; Ioannis Gerasimou[5], MSc; Beatriz Alves[6], MSc; Maarten de Vos[7], Prof Dr; Ioannis Drivas[8], PhD; John Zaras[9], MSc; Andreas Stergioulas[10], MSc; Iskanter Bensenousi[1], MSc; Leontios Hadjileontiadis[3,5*], Prof Dr[‡]; Christos Chatzichristos[7*], PhD; Stelios Hadjidimitriou[5*], PhD

[1]AINIGMA Technologies, Leuven, Belgium
[2]Faculdade de Motricidade Humana, Universidade de Lisboa, Centro Interdisciplinar de Estudo da Performance Humana, Lisbon, Portugal
[3]Department of Biomedical Engineering and Biotechnology, College of Medicine and Health Sciences, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates
[4]NEUROSEC, Department of Psychiatry, University of Oxford, Oxford, United Kingdom
[5]Department of Electrical and Computer Engineering, School of Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece
[6]Faculdade de Motricidade Humana, University of Lisbon, Lisbon, Portugal
[7]STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics, Department of Electrical Engineering, KU Leuven, Leuven, Belgium
[8]Diadikasia Business Consulting Symvouloi Epicheiriseon AE, Athens, Greece
[9]Squaredev, Brussels, Belgium
[10]Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece
[‡]AI-PROGNOSIS consortium
[*]these authors contributed equally

Corresponding Author:
Leontios Hadjileontiadis, Prof Dr
Department of Electrical and Computer Engineering
School of Engineering
Aristotle University of Thessaloniki
Univ. Campus, D Building, 6th floor, Of#26
Thessaloniki, 54124
Greece
Phone: 30 6977269398
Email: leontios@auth.gr

Abstract

The growing emphasis on trustworthy artificial intelligence (AI) in health care reflects a shift away from models optimized for predictive performance toward governable and auditable systems that can be adopted and sustained in clinical practice. Nonetheless, many clinical AI applications continue to privilege technical performance while underaddressing ethical, regulatory, and societal considerations, leading to concerns around robustness, transparency, and clinical adoption. To address this, governance frameworks such as the Assessment List for Trustworthy Artificial Intelligence (ALTAI) have been proposed to operationalize trust-related requirements across the AI lifecycle. However, evidence on the practical use of these frameworks remains limited. In this Viewpoint, we describe the application of ALTAI as a procedural governance framework within the Horizon Europe AI-PROGNOSIS project, which aims to support Parkinson disease diagnosis and care through predictive models and digital biomarkers derived from everyday devices. The seven ALTAI requirements (ie, human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, nondiscrimination, and fairness; societal and environmental well-being; and accountability) were mapped to key stages of the AI lifecycle within the project, including design and specification, data preparation, model development and validation, user interface and user experience deployment, external prospective validation, and overarching management and workflow. To examine how these requirements were perceived in practice, we conducted a structured internal survey among AI developers and data scientists involved in the AI-PROGNOSIS project (n=10). Participants rated the relevance of the 17 ALTAI subdomains using a three-point prioritization scale. Technical accuracy, data governance, and privacy were consistently rated as highly relevant, whereas societal impact received the lowest prioritization. This pattern reflects a documented tension in AI development, where technical teams tend to deprioritize broader societal concerns under delivery and performance constraints. Nonetheless, this work should be interpreted as a context-specific case study rather than a

XSL•FO
RenderX

validation of ALTAI. The small sample size and project-specific setting limit generalizability, and these findings should not be considered as representative of broader clinical AI development. Overall, by making prioritization gaps explicit and embedding multidisciplinary review across lifecycle checkpoints, this case study illustrates how structured governance frameworks can surface implementation tensions and support accountable AI development. While these approaches do not resolve all of the aforementioned challenges, they provide practical guidance for integrating trust-related considerations into clinical AI projects.

(J Med Internet Res 2026;28:e85433) doi:10.2196/85433

KEYWORDS

trustworthy artificial intelligence (AI); clinical AI; ALTAI Framework; assessment list for trustworthy artificial intelligence; AI-PROGNOSIS European research initiative; lifecycle safeguards; ethical AI integration

Introduction

Artificial intelligence (AI) has seen accelerating development and adoption across health care domains, driven by advances in machine learning (ML), data availability, and computational power [1]. While this technological momentum has led to significant gains in diagnostic accuracy, prognostic modeling, and treatment optimization, the integration of AI systems into routine health care remains cautious and uneven. This reticence, while partially attributable to regulatory inertia and data access limitations, is fundamentally rooted in a deeper concern: the perceived trustworthiness of AI-driven systems [2,3]. This is, in fact, a critical challenge, as performance alone is insufficient without ensuring reliability, ethical alignment, and public trust [4], especially as health care AI systems transition from proof of concept to clinical deployment.

Trust in clinical AI transcends conventional performance metrics. It is not reducible to algorithmic accuracy or validation statistics alone, but rather represents a composite property that encompasses transparency, interpretability, accountability, and alignment with both clinical values and ethical principles [5,6]. Recent works have highlighted how the implementation of AI in health care is increasingly shaped not only by technical feasibility, but by the complex interplay of governance, institutional norms, and frontline practices [7,8]. Trustworthiness in AI is increasingly recognized as a multidimensional construct that encompasses not only compliance with technical benchmarks, but also alignment with ethical principles and regulatory standards [9]. Such an alignment must extend beyond end-product validation to permeate the entire lifecycle of system development. Procedural approaches, those that embed trust-oriented safeguards from design to deployment, are essential to achieving this [10]. From the Viewpoint of end users, including clinicians, patients, and institutions, trust is not simply earned by retrospective audit or certification; instead, it is cultivated over time, shaped by system behavior, user experience (UX), and organizational context [11].

Despite the proliferation of frameworks and normative guidance, ranging from high-level ethical principles to emerging regulatory instruments, a persistent implementation gap remains [11,12]. Existing instruments frequently emphasize outcomes rather than mechanisms; they evaluate trust post hoc, rather than embedding it procedurally throughout the AI lifecycle. This disconnect highlights the absence of a pragmatic scaffolding to guide trustworthy AI design and deployment within high-stakes environments, such as health care.

This Viewpoint advances a procedural approach to trustworthiness in clinical AI, drawing upon the Assessment List for Trustworthy Artificial Intelligence (ALTAI) developed by the High-Level Expert Group on AI [13], as a representative tool for operationalizing ethical and regulatory principles. Unlike purely aspirational codes, the ALTAI framework offers a practical, step-by-step checklist that can be directly integrated into project workflows. It delivers concrete guidance throughout the entire AI lifecycle, including design specification, data governance, model development, evaluation, and deployment, and includes defined metrics and procedures to promote compliance and transparency [13]. To contextualize this approach, we examine the AI-PROGNOSIS project [14], a European initiative focused on the development of predictive models for Parkinson disease (PD), as a case study. AI-PROGNOSIS was selected as the case study for applying the ALTAI framework because it concentrates on several recurrent challenges in clinical AI. The project aims to generate individualized PD risk scores, PD progression forecasts, and medication response estimates from multimodal health data. This requires analyzing data from heterogeneous data sources and longitudinal modelling, aiming, for instance, to provide support for therapeutic decision-making. These characteristics make AI-PROGNOSIS a suitable and demanding test case for examining how ALTAI can guide trustworthiness, transparency, and risk mitigation in complex clinical AI systems beyond this specific consortium, illustrating how a procedural framework can shape real-world system design.

By anchoring trustworthiness in procedural steps rather than retrospective assessments, this Viewpoint contributes to a growing body of literature calling for actionable strategies to embed ethical and regulatory principles into real-world AI systems [7,8]. In the sections that follow, we articulate the conceptual rationale for procedural trust, outline the ALTAI framework, illustrate its instantiation across the AI development continuum via the AI-PROGNOSIS case study, and reflect on the practical, ethical, and sociotechnical tensions encountered during implementation in health care innovation.

From Principles to Procedure: Why Trust in AI Needs a Blueprint

The proliferation of ethical guidelines for AI has revealed a growing consensus: trust is essential for the responsible deployment of AI in health care. Yet, despite the emergence of high-level principles, such as fairness, transparency, and accountability, there remains a persistent gap between normative aspirations and practical implementation [15,16]. This

XSL•FO
RenderX

disconnect has prompted calls for procedural frameworks that can translate abstract values into actionable design and governance strategies [17]. This approach has several benefits over traditional principle-based frameworks of trust, which typically rely on specific targets and retrospective evaluations.

Nevertheless, trust in AI is neither a monolithic concept nor an intrinsic property of the system; rather, it is a system-level outcome shaped by dynamic interactions among users, institutions, and broader sociotechnical environments [18]. It emerges from the interplay of interrelated technical and ethical dimensions, such as explainability, robustness, and fairness, none of which are sufficient in isolation. Cultivating trust, therefore, entails iterative, lifecycle-spanning processes that embed normative safeguards into the design, development, deployment, and governance of AI systems [19,20].

In alignment with the ALTAI framework [13] and International Organization for Standardization/International Electrotechnical Commission Technical Specification (ISO/IEC TS) 5723:2022 [21], we identify eight core dimensions, that is, robustness, generalization, explainability, accountability, transparency, reproducibility, fairness, and privacy, as foundational to Trustworthy AI in health care. These dimensions reflect a synthesis of normative principles and practical requirements for clinical-grade AI systems. In particular, robustness refers to a system's capacity to perform reliably under uncertainty, such as noisy inputs, adversarial perturbations, or incomplete records, without significant loss of function [9]. However, robustness must coexist with usability and interpretability, particularly in clinical environments. Closely related is generalization, the model's ability to extrapolate to unseen data, which remains a fundamental challenge given the risk of underfitting or overfitting, especially in small or biased datasets [19,22,23].

Explainability is a context-sensitive construct that varies across stakeholders, clinicians, patients, regulators, and developers [20]. It may be achieved through post hoc methods (eg, Shapley additive explanations and local interpretable model-agnostic explanations) or interpretable models, each with trade-offs in fidelity and scalability [24]. Accountability demands clear traceability of decisions to responsible entities [25], while transparency, its enabling counterpart, requires open disclosure of model purpose, data provenance, and performance characteristics [26]. Achieving transparency often involves managing organizational or proprietary constraints. Reproducibility, a pillar of scientific integrity, remains elusive in ML due to nondeterminism in training processes and hardware dependencies [27]. Fairness, arguably the most socially charged dimension, seeks to mitigate bias introduced during data collection, model design, and deployment [9,28]. Technical responses span pre-, in-, and postprocessing interventions and must be informed by socioethical theories of discrimination and equity [22,28,29]. Finally, privacy safeguards not only identifiers but also individual autonomy over data use [12]. While techniques like differential privacy, de-identification, and data minimization offer protection, they may constrain model expressiveness or transparency [30]. These trade-offs underscore the necessity of procedural frameworks that treat trust not as a checklist but as a dynamic property shaped by ongoing design, validation, and governance.

While often treated as discrete targets, the aforementioned dimensions are deeply interwoven and must be addressed through integrated, context-sensitive design strategies across the AI lifecycle [23]. Attempts to enhance one dimension, such as increasing transparency, can inadvertently compromise another, such as protecting proprietary data or patient privacy. This interplay reinforces the need for procedural frameworks that consider trustworthiness as a system-wide property, developed iteratively and contextually throughout the AI lifecycle.

It is useful to clarify that the eight ISO/IEC TS 5723:2022 dimensions are introduced as foundational properties that define what constitutes a trustworthy clinical AI system, whereas the seven ALTAI requirements provide a procedural structure for how these properties can be operationalized across the AI lifecycle. Several of the ISO dimensions (eg, robustness, transparency, fairness, privacy, and accountability) map directly onto ALTAI's requirements, while others (eg, generalization and reproducibility) are implicitly addressed through ALTAI's emphasis on technical robustness, documentation, and oversight. This alignment positions ALTAI not as an alternative to the ISO dimensions, but as a practical mechanism for embedding them into development workflows, as discussed next.

ALTAI as a Procedural Anchor

While core dimensions of Trustworthy AI, such as robustness, fairness, and transparency, provide conceptual structure, their realization in clinical settings requires systematic, context-sensitive implementation. From the Asilomar AI Principles [31] and the Montreal Declaration [32], to institutional efforts like AI4People [6] and the Organisation for Economic Co-operation and Development guidelines [33], much of this work has emphasized normative commitments, fairness, accountability, transparency, and safety. National strategies, including those from China [34], the United Kingdom [35], the United States [36], and the European Union [37], have begun translating these principles into policy and regulation. In the health care domain, oversight by bodies such as the US Food and Drug Administration [36] and the National Institute of Standards and Technology [38] adds additional complexity, particularly for high-risk systems. In particular, the US Food and Drug Administration guidance on AI and ML-enabled software as a medical device covers good ML practices, including data management, validation, and postmarket monitoring [36]. Similarly, the National Institute of Standards and Technology AI Risk Management Framework provides voluntary lifecycle risk management, emphasizing governance, mapping, measurement, and management of risks such as bias, robustness, and explainability in health care settings [38].

Among these efforts, the European Union's (EU's) ALTAI [39], developed by the high-level expert group on artificial intelligence in 2020 [33], stands out as a concrete procedural framework for embedding trust across the AI development lifecycle. Unlike principle-driven charters, ALTAI codifies seven actionable requirements: human agency and oversight, technical robustness and safety, privacy and data governance,

XSL•FO
RenderX

transparency, diversity and fairness, societal well-being, and accountability. These dimensions are designed not as abstract endpoints, but as iterative checkpoints, aligning design practices with ethical and regulatory imperatives at each stage of AI development. Its web-based tool supports structured self-assessment and generates visual diagnostics, such as radar plots, summarizing strengths and deficiencies, thus enabling continuous monitoring, recommended next steps, and targeted refinement.

Albeit scoring details remain opaque, ALTAI represents one of the most widely adopted instruments for proceduralizing trust in AI workflows. To assess its translational value in applied clinical research, we adapted ALTAI within the AI-PROGNOSIS project. This adaptation aligned technical design efforts with structured trust requirements, providing a foundation for identifying risk points, embedding ethical safeguards, and supporting internal reflection among development teams. In the following section, we present this case study as an applied instantiation of ALTAI, illustrating how trust-oriented governance can be realized through procedural integration.

Operationalizing Procedural Trust: The AI-PROGNOSIS Case Study

To evaluate the ALTAI procedural trust framework in clinical AI, we conducted a structured assessment within the AI-PROGNOSIS project, which aims to generate individualized PD risk scores, PD progression forecasts, and medication response estimates, using ML techniques applied to multimodal health data. AI-PROGNOSIS was selected as the case study since it embodies several characteristics that typify the broader challenges of clinical AI. The project integrates multimodal health data, including clinical records, wearable-derived digital biomarkers, and longitudinal assessments, and supports high-stakes decisions related to PD. These features create a complex technical and ethical landscape in which issues of robustness, fairness, transparency, and data governance are particularly salient. As such, AI-PROGNOSIS provides a demanding and representative context for stress-testing how a procedural framework like ALTAI can be operationalized across the AI lifecycle.

A targeted survey (Table S1 in Multimedia Appendix 1) was administered to evaluate the relevance of ALTAI's seven requirements in this context. The respondents included ten AI developers and data scientists across five partner institutions (two academic centers and three SMEs) of the AI-PROGNOSIS consortium. The group included eight male and two female participants, with a mean age of 40.2 (SD 8.9) years, and an average of 13.3 (SD 11) years of professional experience. All participants were directly involved in technical design or data pipeline development for the AI-PROGNOSIS platform.

The survey asked participants to rate the perceived importance of 19 ALTAI subgroups (Figure S1 in Multimedia Appendix 1), such as fallback plans, explainability, fairness, and environmental impact, using a three-point scale (low=1, medium=2, and high=3). When a subgroup was considered by a participant to be unrelated to the project, it was assigned a value of zero. The average importance rating was then calculated using only the non-zero ratings, averaged across all subgroups and participants. To increase the resolution of the resulting average importance score, it was rescaled from 1-3 to the range of 1-10 by multiplying its value by a factor of 10/3.

The main aim was to identify which components were considered most critical for a health care-focused AI system in development. In fact, the survey was designed as an internal prioritization exercise rather than a statistical evaluation of ALTAI. The 10 respondents were members of the AI-PROGNOSIS technical work packages and were selected because they were directly responsible for model development, data engineering, or pipeline integration within the project. In this context, roles such as "AI developer" and "data scientist" overlap substantially, and the survey did not aim to distinguish between them. The sample size reflects the bounded size of the consortium's technical team, and no claims of statistical significance or generalizability are made. For this reason, descriptive characteristics are reported without inferential interpretation, and the survey is explicitly framed as capturing the perspectives of this specific project team rather than representing broader trends in clinical AI. Its purpose was to surface internal prioritization patterns across ALTAI subdomains to inform project-specific design decisions, not to validate the ALTAI framework itself. From this perspective, potential for selection bias can be acknowledged, as all participants were selected precisely because they were directly involved in AI-PROGNOSIS development, and their perspectives therefore reflect the priorities and constraints of this specific consortium rather than those of the broader clinical AI community.

Figure 1 illustrates the average importance ratings provided by AI experts across the seven ALTAI requirements. More specifically, subgroups related to accuracy, privacy, reliability, fallback planning, human agency and autonomy, explainability, communication, and auditability received the highest range of average importance score (8.4-7.1), respectively. Bias mitigation, stakeholder participation, data governance, human oversight, and resilience to attack and security received an average importance score within the range of 6.8-6.0, respectively. The third range of 5.8-5.0 included accessibility and universal design, risk management, general safety, and traceability, respectively. In contrast, requirements addressing the impact on work and skills, along with the impact on society at large or democracy, were rated as less critical with average importance scores of 4.8 and 4.7, respectively. Finally, the environmental well-being was considered as not related to the project and resulted in zero. In addition, Figure S1 in Multimedia Appendix 1 offers a more granular view by displaying the average importance scores for each ALTAI subgroup. Table S2 in Multimedia Appendix 1 provides the full set of ALTAI checklist responses alongside the corresponding system-generated recommendations. Moreover, to support integration across the AI-PROGNOSIS development cycle, ALTAI's seven requirements were mapped to six core lifecycle stages, namely: design and specification, data preparation, model development and validation, interface design and deployment, external validation, and cross-cutting governance and workflow (Figure 2). This mapping, also presented in Table S1 in Multimedia Appendix 1, aligns with the high-level expert group

XSL•FO
RenderX

on artificial intelligence's guidance on adapting trustworthiness frameworks to specific system contexts. It enabled the identification of stage-specific trust touchpoints and informed the implementation of safeguards, such as adversarial testing, explainability benchmarking, and privacy-preserving data handling. The integration of these lifecycle stages (Figure 2) is further explored below.

Figure 1. Average importance scores per requirement for ALTAI subgroups. Ratings were derived from a structured survey conducted with AI experts in the AI-PROGNOSIS project, reflecting the perceived relevance of each ALTAI requirement to clinical AI development. ALTAI: Assessment List for Trustworthy Artificial Intelligence, AI: Artificial Intelligence.

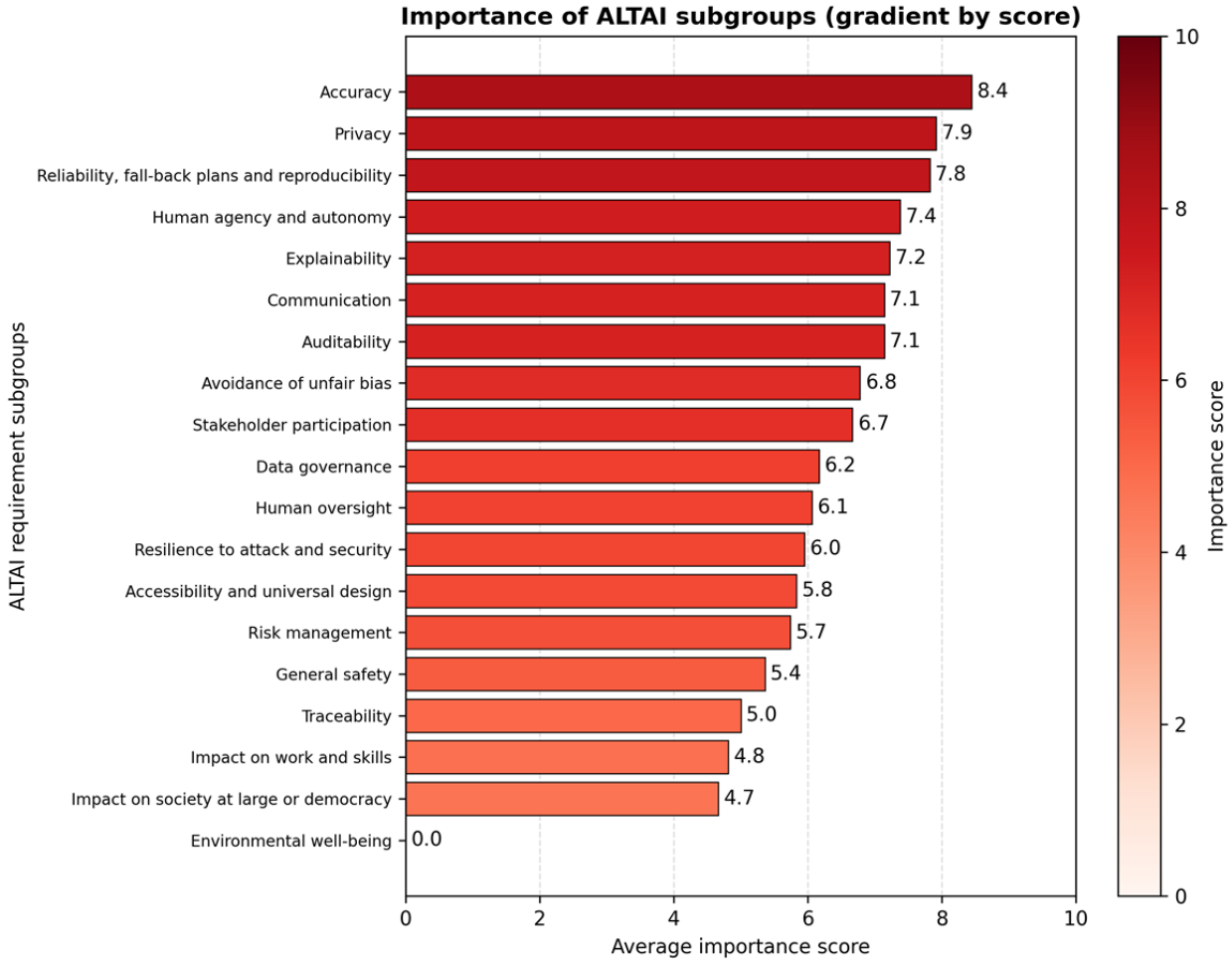

XSL•FO
RenderX

Figure 2. Trustworthy AI components embedded in the AI-PROGNOSIS framework. The graphical representation summarizes how dimensions such as explainability, user experience (UX), user interface (UI), and accountability are operationalized within the AI-PROGNOSIS project, supporting trust by design. AI: Artificial Intelligence.

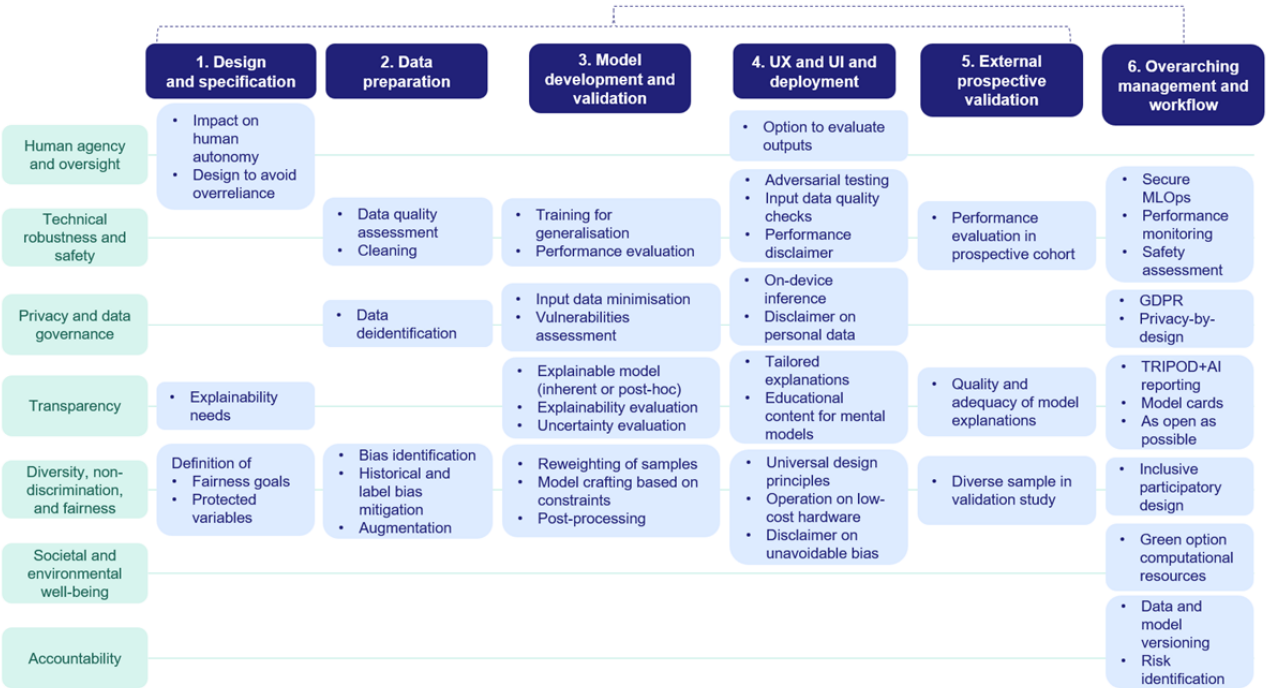

Lifecycle Integration of ALTAI Requirements

The mapping of ALTAI requirements onto the six lifecycle stages of AI-PROGNOSIS reflects how trust-oriented safeguards were embedded as procedural checkpoints rather than post-hoc evaluations. During design and specification, ALTAI's requirements for human agency, accountability, and transparency informed early architectural decisions, including the definition of fallback mechanisms, documentation structures, and user-facing communication strategies. In the data preparation stage, the requirements for privacy, data governance, and fairness guided the establishment of data-quality routines, access-control structures, and reviews to identify potential bias proxies. Within model development and validation, ALTAI's emphasis on technical robustness and safety was operationalized through stress-testing pipelines, adversarial-robustness assessments, and explainability benchmarking, ensuring that model behavior was interrogated under realistic perturbations and uncertainty conditions.

As the system moved toward interface design and deployment, requirements related to human oversight, transparency, and diversity were instantiated through iterative co-design sessions, accessibility reviews, and the integration of explanation elements into the UX and user interface (UI). These activities ensured that model outputs were intelligible, appropriately contextualized, and aligned with user needs. During external validation, ALTAI's focus on fairness, robustness, and accountability shaped the evaluation of model performance across heterogeneous populations and settings, as well as the refinement of explanation strategies for different stakeholder groups. Finally, in the cross-cutting governance and workflow stage, ALTAI's accountability and societal- and environmental-well-being requirements were reflected in the establishment of ML operations pipelines, lineage tracking, dynamic data-governance plans, ethics-board oversight, and monitoring of computational-resource demands over time, enabling continuous, sustainable monitoring and structured decision-making throughout the system's lifecycle. These lifecycle-specific integrations also set the stage for the more detailed examination of practical, ethical, and organizational tensions that emerged during implementation, which are explored in the following section.

Beyond diagnostic value, the ALTAI adaptation fostered internal reflection among development teams, prompting early-stage deliberation on fallback mechanisms, UI accessibility, and the ethical implications of probabilistic risk scoring. These insights informed both system architecture and stakeholder engagement strategies, reinforcing the role of procedural frameworks as catalysts for trust-aware design. Collectively, these findings emphasize ALTAI's function not merely as an evaluative instrument, but as a formative scaffold for embedding trust-oriented design logic throughout the development pipeline, as well. This case study exemplifies how procedural frameworks can facilitate the operationalization of ethical and regulatory objectives in clinical AI, with technical teams serving as key intermediaries in translating abstract principles into implementable system architecture.

Tensions in Pursuit of Trustworthy AI

While procedural frameworks such as ALTAI offer structured guidance for embedding trust into AI development, their translation into clinical practice reveals persistent tensions. These tensions are not solely technical but emerge from the entanglement of ethical, organizational, and regulatory constraints with real-world implementation dynamics. Trust, correspondingly, is not a static attribute in this context, but must

XSL•FO
RenderX

be seen as a negotiated, adaptive property shaped and actively maintained across the AI lifecycle by design choices, deployment decisions, and end user interactions [17,31]. Staking out paths to navigate these tensions should therefore be a procedural undertaking, allowing dynamic adaptation on a case-by-case basis, yet offering a guiding rail. Below, we outline key decision-points and how they were navigated in the AI-PROGNOSIS project, using this approach.

Engineering Clarity Under Complexity

Design-time interventions within AI-PROGNOSIS emphasized risk mitigation through adversarial robustness, constrained model behavior, and transparency mechanisms. Structured data validation pipelines were established to ensure completeness, consistency, and semantic fidelity of clinical inputs. Early-stage vulnerability assessments leveraged tools such as the Adversarial Robustness Toolbox [40] and CleverHans [41] to stress-test model responses under plausible perturbations. These safeguards were necessary but not sufficient, given that trust is also shaped by user perceptions, system intelligibility, and the sociolegal context of deployment.

Explainability was pursued through Shapley additive explanations and local interpretable model-agnostic explanations visualizations [42], integrated into intuitive UX and UI elements, co-developed through co-design sessions with patients, clinicians, and human-computer interaction specialists. The desired output included layout consistency, content hierarchy, accessibility, and performance optimization that can ensure usability across user populations. Additionally, design principles, such as Google's Material Design or Apple's Human Interface Guidelines, offered baseline frameworks for further validation through iterative testing with diverse users. These interfaces incorporated risk disclaimers, scenario-specific warnings, and model output rationales, elements shown to foster situational awareness and calibrate expectations [43]. However, tension emerged between model interpretability and predictive fidelity: simpler, more explainable models sometimes underperformed in capturing longitudinal patterns, while high-dimensional neural architecture offered superior accuracy at the cost of intelligibility [44,45].

Transparency obligations are also intersected with institutional and commercial constraints. Open disclosure of model behavior, data lineage, and source code encountered resistance when proprietary intellectual property, reputational risk, or liability exposure were perceived. These experiences echo concerns documented in broader AI governance literature, where explainability is seen as a boundary object, interpreted differently by legal experts, regulators, engineers, and lay users [17,20,46]. Trust, therefore, cannot rely solely on post hoc visualization tools or interface overlays; it must be cultivated through continuous interaction between technical artifacts and epistemic communities.

Generalization Versus Representativeness

Achieving generalization in clinical AI extends beyond algorithmic optimization. It requires validating performance across subpopulations, health care settings, and temporal shifts, domains where real-world complexity and structural inequities surface. Within AI-PROGNOSIS, data were curated to reflect heterogeneity across age, sex, and disease severity, with feature reviews conducted alongside clinicians to de-risk unintentional bias proxies.

Still, key fairness checks were constrained by unavailable or restricted variables. Under the general data protection regulation [47] and ethical review protocols, the collection of race, ethnicity, and socioeconomic indicators was either prohibited or discouraged, limiting the granularity of bias auditing [9,28,29]. These limitations underscored the tension between privacy-preserving practice and equity-informed auditing. Without disaggregated data, even well-calibrated models can systematically underperform or generate disparate outcomes for minority groups [48]. Moreover, generalizability was not static. Drift monitoring, supported by emerging regulatory requirements such as the EU AI Act's obligations for continuous post-market performance surveillance (Articles 12, 14, 15, and 72) [49], is increasingly implemented through dedicated tools like Evidently AI [50]. These enable temporal performance tracking but still require careful configuration to avoid false alarms or blind spots. Participatory workshops were convened with diverse stakeholders to co-define performance thresholds, identify context-specific harms, and design explanation strategies tailored to each stakeholder type [45,51]. In the same vein, external validation ensures that AI systems generalize beyond development data. Explainable AI plays a central role during external validation. Explanation strategies should be adapted to different user groups, from clinicians to patients to engineers. At the same time, co-creation workshops or equivalent co-creation processes (eg, innovation jams, living laboratories, open innovation platforms, lead user collaborations, and crowdsourced design challenges) are valuable for defining these strategies and ensuring a shared understanding of outputs.

These findings suggest that generalization must be reframed: not only as an empirical measure of cross-sample performance, but as a sociotechnical process of aligning predictive behavior with real-world variance, regulatory constraints, and stakeholder expectations. Fairness, interpretability, and robustness cannot be optimized independently; they must be co-engineered through continuous iteration and value-sensitive design [16,52].

Institutional Scaffolding for Sustainable Trust

Trust, to be enduring, must outlive model deployment. Within AI-PROGNOSIS, governance mechanisms were embedded across the system lifecycle using ML operations pipelines [53] managed via ML flow (an open-source platform for managing ML lifecycle) [54]. These supported reproducibility, lineage tracking, and automated logging of model outputs for compliance and audit purposes. Model documentation, following TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) and Datasheets for Datasets [55] guidelines, facilitates reproducibility and informs external reviewers about data provenance, development conditions, and known limitations. Ethical and legal oversight structures were supported through dynamic data governance plans housed in the European Open Science Cloud's aggregator for open science [56], which offers version-controlled templates for privacy, security, and access control compliance. An internal

XSL•FO
RenderX

ethics board, comprising technical, legal, and clinical representatives, was tasked with monitoring value drift, assessing updates to explainability outputs, and coordinating stakeholder feedback loops.

Sustainability efforts address both environmental impact (eg, energy cost of training pipelines) and downstream clinical implications (eg, deskilling risks or overdependence on AI predictions) [16]. These are increasingly salient concerns as health care systems adopt AI at scale and require not only functional models, but systems that preserve professional autonomy and adapt to evolving sociopolitical contexts [16,52]. Moreover, regulatory compliance, particularly with the general data protection regulation [47], is non-negotiable. Privacy-by-design principles should guide system architecture. Data management teams must ensure ethical oversight, secure processing, and adherence to data-sharing agreements.

Crucially, trust must be institutionally maintained. This requires aligning development workflows with adaptive governance structures capable of incorporating feedback, absorbing policy shifts, and ensuring ethical continuity over time. Static checklists are ill-suited for this role; procedural frameworks, on the other hand, can (and must) evolve into organizational capabilities, rooted in accountability, reflexivity, and stakeholder engagement.

Technical Focus and Societal Impact

Although the lower importance assigned to societal-level impacts, for example, effects on work, skills, democracy, or broader social structures (Figures 1 and 2), could be interpreted as a natural consequence of the AI-PROGNOSIS clinical implementation focus, our findings point to a deeper and well-documented tension in the field of ethical AI. Prior research has shown that technical teams often prioritize system-proximal requirements (eg, accuracy, safety, and reliability) while viewing societal and macro-ethical considerations as diffuse, less actionable, or outside their immediate scope of responsibility [57,58]. This divergence does not indicate that societal impacts are irrelevant; rather, it highlights a persistent gap between optimizing AI performance within a clinical workflow and ensuring that technology aligns with broader social, democratic, and labor-related values. In line with the broader literature on responsible AI, this tension should be understood as an ongoing governance challenge that requires deliberate management rather than acceptance [59,60]. For AI-PROGNOSIS, this means that even if domain experts perceive societal impacts as less central to immediate development tasks, these considerations must still be systematically integrated into oversight structures, stakeholder engagement processes, and long-term governance mechanisms. Addressing this misalignment is essential for ensuring that the system evolves not only as a clinically robust tool but also as one that contributes positively to the wider socio-technical ecosystem in which it will operate.

Taken together, the experience in AI-PROGNOSIS illustrates how ALTAI can function as a starting scaffold rather than an endpoint for procedural trust. Several trust-oriented practices emerged organically during development, not because ALTAI prescribed them, but because real-world clinical, technical, and organizational constraints demanded additional structure. These included the introduction of quantitative lifecycle checks, the systematic integration of stakeholder feedback into design decisions, and the establishment of reproducibility and governance mechanisms that supported continuous oversight. Their emergence demonstrates how ALTAI can trigger deeper procedural adaptations when applied in practice, offering insight into how high-level principles translate into operational routines within complex clinical AI projects.

Outlook: Building Living Systems of Trust

This Viewpoint has outlined a procedural approach to embedding trust in clinical AI, grounded in the ALTAI framework [39] and instantiated through the AI-PROGNOSIS project. By integrating ethical, technical, and regulatory safeguards across the AI lifecycle, we have demonstrated how trust can be operationalized not only as a design goal but as a dynamic property of clinical AI systems. Nevertheless, we acknowledge that the survey was limited by its small sample size (n=10) and restriction to AI developers and data scientists from the AI-PROGNOSIS consortium. This effort represents a pilot internal assessment rather than statistically significant or generalizable evidence of ALTAI domain priorities. Moreover, the survey data were collected exclusively from AI developers and data scientists, whereas other key stakeholders, such as people with Parkinson disease, caregivers, and health care professionals, were not included. This may have biased the prioritization of ALTAI subdomains toward the perspectives and priorities of technical experts, which do not necessarily align with end user values and concerns. Involvement of a more diverse and representative group of stakeholders can ensure that domain weighting better reflects end users' needs and expectations.

Yet, as the field matures, it is increasingly clear that procedural scaffolding alone has its limitations. Trustworthiness must be sustained through adaptive governance, capable of responding to evolving risks, shifting stakeholder expectations, and emerging regulatory mandates. In this regard, the newly introduced European Artificial Intelligence Act (AI Act) [61] represents a pivotal inflection point. Entering into force in August 2024, the AI Act introduces a harmonized, risk-based legal framework for AI across the EU, with specific obligations for high-risk systems, including those deployed in health care [61]. These include requirements for transparency, human oversight, robustness, and postmarket monitoring, many of which align with ALTAI's procedural ethos but now carry legal enforceability.

Translating ALTAI's procedural safeguards into the regulatory obligations of the EU AI Act offers a practical pathway for organizations preparing for compliance with high-risk clinical AI requirements. Several ALTAI mechanisms map directly onto the AI Act's mandated controls, including lifecycle documentation (eg, data sheets and model cards) that support the AI Act's technical documentation and record-keeping duties; structured human-oversight protocols that align with Articles 14 and 29; risk-management workflows that mirror the AI Act's continuous risk-assessment and mitigation obligations; and data governance practices that reinforce requirements for data quality, representativeness, and bias monitoring. Additionally, ALTAI's

XSL•FO
RenderX

emphasis on fallback planning, robustness testing, and postdeployment monitoring corresponds to the Act's provisions for postmarket surveillance and performance drift detection. By adopting ALTAI-like processes early in system design, organizations can establish governance infrastructures that not only enhance trustworthiness but also streamline future conformity assessments under the AI Act.

The AI-PROGNOSIS framework is being continuously adapted to anticipate and respond to these and other regulatory and policy developments. Specifically, future iterations will (1) integrate AI Act compliance checkpoints into development workflows, (2) expand stakeholder engagement to include legal and regulatory experts, and (3) establish mechanisms for continuous postdeployment monitoring and redress. These steps reflect the broader approach, shifting away from principle-based ethics toward procedural trust and institutionalized accountability, where trust is not only designed and cultivated, but governed.

Looking ahead, we argue that trustworthy AI in health care must be conceptualized as a living system, one that evolves through iterative feedback, interdisciplinary and diverse collaboration, and regulatory responsiveness. This requires moving beyond static checklists toward reflexive infrastructures that embed ethical deliberation, stakeholder negotiation, and lifecycle oversight into the core of AI development. As the regulatory landscape crystallizes and clinical adoption accelerates, such infrastructures will be essential to ensure that AI systems remain not only performant but aligned with the values, rights, and expectations of the societies they serve.

Acknowledgments

The authors thank all members of the AI-PROGNOSIS Consortium. The authors declare the use of generative artificial intelligence (GenAI) in the writing process. According to the GAIDeT taxonomy [62], the proofreading task was delegated to GenAI tools under full human supervision. The GenAI tool used was Perplexity (GPT-5.2). Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

The contributors associated with AI-PROGNOSIS Consortium are as follows: Leontios J. Hadjileontiadis, Stelios Hadjidimitriou, Vasileios Charisis, Ioannis Gerasimou, Charalampos Sotirakis, and Apostolos Moustaklis, Aristotle University of Thessaloniki, Greece; Thanasis Kakasis, Ali Saad, Nikos Melanitis, Alex Bensenousi, and Angeliki Zarifi, Ainigma Technologies; Despina Anastasopoulos, Dorine Matzakou, Theodora Brisimi, Marilena Damkali, Petros-Sozon Dimitrakopoulos, and Alexandros Miouglou, Netcompany-Intrasoft, Greece; Ioannis Drivas and Sotirios Michagiannis, Diadikasia Business Consulting Symvouloi Epicheiriseon, Greece; Therese Scott Duncan, Jamie Luckhaus, Anna Clareborn, and Sara Riggare, Uppsala University, Sweden; Olga Sanchez Solino, Abbvie Deutschland GmbH & Co. KG, Spain; Amel Drif, Anna Rybicka, Natalia Del Campo, Margherita Fabbri, and Olivier Rascol, Centre Hospitalier Universitare de Toulouse, France; Sofia B. Dias, Ghada Alhussein, Beatriz Alves, and Filomena Carnide, Faculdade de Motricidade Humana, University of Lisbon, Portugal; Nikos Grammalidis, Kosmas Dimitropoulos, Andreas Stergioulas, and Theocharis Chatzis, Centre for Research and Technology Hellas, Greece; Nikola Goetz and Niloofar Tavakoli, Neurotransdata GmbH, Germany; Helene Huts, Maarten De Vos, Christos Chatzichristos, Thomas Strypsteen, Fan Wang, Aldona Niemiro Sznajder, and Georgios Roussis, Katholieke Universiteit Leuven, Belgium; Elissavet Zogopoulou, Charis Giaralis, John Zaras, and Christos Vasilakis, SquareDev, Greece; Björn Falkenburger, Nils Schnalke, and Tim Feige, Technische Universitaet Dresden, Germany; Kristina Leipuviene, Smartsol SIA, Lithuania; Eleni Zamba-Papanicolaou, Kyriaki Michailidou, Christiana Christodoulou, Paraskevi Chairta, and Kyproula Christodoulou, Kypriako Idryma Erevnon Gia Ti Myiki Distrofia, Cyprus; David Lyreskog, University of Oxford, UK; Maria-Luisa Almarcha-Menargues and Monica Kurtis Urra, Movement Disorders Unit, Neurology Department, Hospital Ruber Internacional, Madrid, Spain.

Funding

This study receives funding from the European Union under Grant Agreement No. 101080581 (AI-PROGNOSIS). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency. Neither the European Union nor the European Health and Digital Executive Agency can be held responsible for them. Moreover, LJH acknowledges support from Khalifa University of Science and Technology, Abu Dhabi, UAE, Provost's Office Grant. The funding sources had no role in study design, data collection, data analysis, data interpretation, the writing of the manuscript, or the decision to submit it for publication.

Data Availability

The datasets generated and analyzed in this study are not publicly available due to access restrictions mandated by the AI-PROGNOSIS project's data governance framework. Access may be granted upon reasonable request to the corresponding author and requires prior authorization through the project's formal data-access procedures.

Authors' Contributions

A Saad contributed to conceptualization, managed data curation, formal analysis, investigation, methodology, visualization, writing – original draft, writing – review and editing. SBD, GA, DL, IG, and BA contributed to investigation, methodology,

XSL•FO
RenderX

validation, writing – review and editing. MdV, ID, JZ, A Stergioulas, and AB contributed to writing – review and editing. CC contributed to conceptualization, methodology, validation, project administration, supervision, writing – review and editing. LJH and SH contributed to conceptualization, methodology, validation, project administration, supervision, funding acquisition, writing – review and editing.

Conflicts of Interest

The authors declare no financial or personal relationships that could have influenced the work reported in this manuscript. Certain authors held affiliations with commercial entities, including AINIGMA Technologies, Diadikasia Business Consulting, and Squaredev, during the study period. These affiliations are disclosed for transparency; the companies had no involvement in the study's design, conduct, analysis, or reporting. No additional competing interests were identified.

Multimedia Appendix 1
Average importance score by ALTAI subgroup and full set of ALTAI checklist responses alongside the corresponding system-generated recommendations.
[DOCX File , 123 KB - jmir_v28i1e85433_app1.docx]

References

1. Rajpurkar P, Lungren MP. The current and future state of AI interpretation of medical images. N Engl J Med 2023;388(21):1981-1990. [doi: 10.1056/NEJMra2301725] [Medline: 37224199]
2. Nong P, Platt J. Patients' trust in health systems to use artificial intelligence. JAMA Netw Open 2025;8(2):e2460628 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.60628] [Medline: 39951270]
3. Tucci V, Saary J, Doyle TE. Factors influencing trust in medical artificial intelligence for healthcare professionals: a narrative review. J Med Artif Intell 2022;5:e000001. [doi: 10.21037/jmai-21-25]
4. Khalighi S, Reddy K, Midya A, Pandav KB, Madabhushi A, Abedalthagafi M. Artificial intelligence in neuro-oncology: advances and challenges in brain tumor diagnosis, prognosis, and precision treatment. NPJ Precis Oncol 2024;8(1):80 [FREE Full text] [doi: 10.1038/s41698-024-00575-0] [Medline: 38553633]
5. Nickel PJ. Trust in medical artificial intelligence: a discretionary account. Ethics Inf Technol 2022;24(1):e000002. [doi: 10.1007/s10676-022-09630-5]
6. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People-an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach (Dordr) 2018;28(4):689-707 [FREE Full text] [doi: 10.1007/s11023-018-9482-5] [Medline: 30930541]
7. Carboni C, Brightwell C, Halpern O, Freyer O, Gilbert S. Reconciling security and care in digital medicine. NPJ Digit Med 2025;8(1):261 [FREE Full text] [doi: 10.1038/s41746-025-01685-0] [Medline: 40341720]
8. Bodnari A, Travis J. Scaling enterprise AI in healthcare: the role of governance in risk mitigation frameworks. NPJ Digit Med 2025;8(1):272 [FREE Full text] [doi: 10.1038/s41746-025-01700-4] [Medline: 40360659]
9. Tran M, Balasooriya C, Jonnagaddala J, Leung GK, Mahboobani N, Ramani S, et al. Situating governance and regulatory concerns for generative artificial intelligence and large language models in medical education. NPJ Digit Med 2025;8(1):315 [FREE Full text] [doi: 10.1038/s41746-025-01721-z] [Medline: 40425695]
10. Reinhardt K. Trust and trustworthiness in AI ethics. AI Ethics 2022;3(3):735-744. [doi: 10.1007/s43681-022-00200-5]
11. Kerasidou C, Kerasidou A, Buscher M, Wilkinson S. Before and beyond trust: reliance in medical AI. J Med Ethics 2022;48(11):852-856 [FREE Full text] [doi: 10.1136/medethics-2020-107095] [Medline: 34426519]
12. Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J Med Ethics 2021;47:e000003. [doi: 10.1136/medethics-2020-106820] [Medline: 33737318]
13. High-level expert group on artificial intelligence. European Commission. URL: https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai [accessed 2025-10-07]
14. Towards Parkinson's risk assessment and prognosis through AI. AI-PROGNOSIS. URL: https://www.ai-prognosis.eu/ [accessed 2026-04-03]
15. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell 2019;1(9):389-399. [doi: 10.1038/s42256-019-0088-2]
16. Mittelstadt B. Principles alone cannot guarantee ethical AI. Nat Mach Intell 2019;1(11):501-507. [doi: 10.1038/s42256-019-0114-4]
17. Dignum V. Ensuring responsible AI in practice. In: Responsible Artificial Intelligence. Artificial Intelligence: Foundations, Theory, and Algorithms. Cham, Switzerland: Springer; 2019:501-507.
18. Hamon R, Junklewitz H, Sanchez Martin J. Robustness and Explainability of Artificial Intelligence. Luxembourg: EUR 30040 EN, Publications Office of the European Union; 2020.
19. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning (still) requires rethinking generalization. Commun. ACM 2021;64(3):107-115. [doi: 10.1145/3446776]

XSL•FO
RenderX

20. Papagni G, de Pagter J, Zafari S, Filzmoser M, Koeszegi ST. Artificial agents' explainability to support trust: considerations on timing and context. AI & Soc 2022;38(2):947-960. [doi: 10.1007/s00146-022-01462-7]
21. ISO/IEC TS 5723:2022-Trustworthiness: vocabulary. International Organization for Standardization. 2022. URL: https://www.iso.org/standard/81608.html [accessed 2025-10-07]
22. Chen P, Wu L, Wang L. AI fairness in data management and analytics: a review on challenges, methodologies and applications. Applied Sciences 2023;13(18):10258. [doi: 10.3390/app131810258]
23. Li B, Qi P, Liu B, Di S, Liu J, Pei J, et al. Trustworthy AI: from principles to practices. ACM Comput Surv 2023;55(9):1-46. [doi: 10.1145/3555803]
24. Retzlaff CO, Angerschmid A, Saranti A, Schneeberger D, Röttger R, Müller H, et al. Post-hoc vs ante-hoc explanations: XAI design guidelines for data scientists. Cognitive Systems Research 2024;86:101243. [doi: 10.1016/j.cogsys.2024.101243]
25. Cui L, Qu Y, Gao L, Xie G, Yu S. Detecting false data attacks using machine learning techniques in smart grid: a survey. Journal of Network and Computer Applications 2020;170:102808. [doi: 10.1016/j.jnca.2020.102808]
26. Geisler S, Vidal M, Cappiello C, Lóscio BF, Gal A, Jarke M, et al. Knowledge-driven data ecosystems toward data transparency. J. Data and Information Quality 2021;14(1):1-12. [doi: 10.1145/3467022]
27. Cockburn A, Dragicevic P, Besançon L, Gutwin C. Threats of a replication crisis in empirical computer science. Commun ACM 2020;63(8):70-79. [doi: 10.1145/3376898]
28. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv 2021;54(6):1-35. [doi: 10.1145/3457607]
29. Binns R. Fairness in machine learning: lessons from political philosophy. 2018 Presented at: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR; February 23-24, 2018; New York University Law School, New York City, USA p. 149-159 URL: https://proceedings.mlr.press/v81/binns18a.html
30. Agarwal S. Trade-offs between fairness and privacy in machine learning. IJCAI Workshop AI for Social Good. 2021. URL: https://crcs.seas.harvard.edu/publications/trade-offs-between-fairness-and-privacy-machine-learning [accessed 2025-10-07]
31. Asilomar AI principles. Future of Life Institute. URL: https://futureoflife.org/ai-principles [accessed 2025-10-07]
32. Ménissier T. A "Machiavellian moment" for artificial intelligence? The montreal declaration for the responsible development of AI [Article in French]. Raisons Polit 2020;77:67-81. [doi: 10.3917/rai.077.0067]
33. High-level expert group on artificial intelligence. European Commission. URL: https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai [accessed 2025-10-07]
34. Governance principles for the new generation artificial intelligence--developing responsible artificial intelligence. China Daily. URL: https://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html [accessed 2025-10-07]
35. Artificial intelligence committee AI in the UK: ready, willing and able? UK Parliament (House of Lords Select Committee on Artificial Intelligence). URL: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm [accessed 2025-10-07]
36. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). United States Food & Drug Administration. 2020. URL: https://www.fda.gov/media/122535/download [accessed 2025-10-07]
37. EU AI Act: first regulation on artificial intelligence. European Parliament. URL: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence [accessed 2025-10-07]
38. Tabassi E. Artificial intelligence risk management framework (AI RMF 1.0). NIST Trustworthy and Responsible AI, National Institute of Standards and Technology. 2023. URL: https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10 [accessed 2026-03-13]
39. Rajamäki J, Gioulekas F, Rocha PAL, Garcia XDT, Ofem P, Tyni J. ALTAI tool for assessing AI-based technologies: lessons learned and recommendations from SHAPES pilots. Healthcare (Basel) 2023;11(10):1454 [FREE Full text] [doi: 10.3390/healthcare11101454] [Medline: 37239739]
40. Nicolae MI. Adversarial robustness toolbox v1.0.0. GitHub. URL: https://github.com/Trusted-AI/adversarial-robustness-toolbox [accessed 2025-10-07]
41. Goodfellow I, Papernot N, McDaniel P. Cleverhans v0.1: an adversarial machine learning library. arXiv.1610.00768v2 2016 Oct 5 [FREE Full text] [doi: 10.1007/978-3-642-27739-9_1635-1]
42. S Band S, Yarahmadi A, Hsu C, Biyari M, Sookhak M, Ameri R, et al. Application of explainable artificial intelligence in medical health: a systematic review of interpretability methods. Informatics in Medicine Unlocked 2023;40:101286. [doi: 10.1016/j.imu.2023.101286]
43. Jawaheer G, Weller P, Kostkova P. Modeling user preferences in recommender systems: a classification framework for explicit and implicit user feedback. ACM Trans Interact Intell Syst 2014;4(2):1-26. [doi: 10.1145/2512208]
44. Watson DS, Krutzinna J, Bruce IN, Griffiths CE, McInnes IB, Barnes MR, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ 2019;364:l886 [FREE Full text] [doi: 10.1136/bmj.l886] [Medline: 30862612]
45. Nazar M, Alam MM, Yafi E, Su'ud MM. A systematic review of human–computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques. IEEE Access 2021;9:153316-153348. [doi: 10.1109/access.2021.3127881]
46. Lipton ZC. The mythos of model interpretability. Commun ACM 2018;61(10):36-43. [doi: 10.1145/3233231]

XSL•FO
RenderX

47. Voigt P, von dem Bussche A. The EU General Data Protection Regulation (GDPR). Cham: Springer; 2017.
48. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018;169(12):866-872 [FREE Full text] [doi: 10.7326/M18-1990] [Medline: 30508424]
49. European Union. Regulation (EU) 2024/1689 of the European parliament and of the council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). European Union EUR-Lex. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj [accessed 2026-02-12]
50. Evidently AI: open'source tools for machine learning monitoring. Evidently AI. URL: https://www.evidentlyai.com [accessed 2026-02-12]
51. Scott IA, van der Vegt A, Lane P, McPhail S, Magrabi F. Achieving large-scale clinician adoption of AI-enabled decision support. BMJ Health Care Inform 2024;31(1):e100971 [FREE Full text] [doi: 10.1136/bmjhci-2023-100971] [Medline: 38816209]
52. Applying sociotechnical approaches to AI governance in practice. Center for Democracy and Technology. URL: https://cdt.org/insights/applying-sociotechnical-approaches-to-ai-governance-in-practice [accessed 2025-10-07]
53. MLOps principles. MLOps. URL: https://ml-ops.org/content/mlops-principles [accessed 2025-10-07]
54. MLflow documentation. MLflow. URL: https://mlflow.org/docs/latest [accessed 2025-10-07]
55. Debray TPA, Collins GS, Riley RD, Snell KIE, Van Calster B, Reitsma JB, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data: TRIPOD-Cluster checklist. BMJ 2023;380:e071018 [FREE Full text] [doi: 10.1136/bmj-2022-071018] [Medline: 36750242]
56. ARGOS: create, link and share data management plans. OpenAIRE. URL: https://www.openaire.eu/argos [accessed 2025-10-07]
57. Mittelstadt BD. Principles alone cannot guarantee ethical AI. Nat Mach Intell 2019;1(11):501-507. [doi: 10.1038/s42256-019-0114-4]
58. Morley J, Floridi L, Kinsey L, Elhalal A. From what to how: an initial review of publicly available ai ethics tools, methods and research to translate principles into practices. Sci Eng Ethics 2020;26(4):2141-2168 [FREE Full text] [doi: 10.1007/s11948-019-00165-5] [Medline: 31828533]
59. Floridi L, Cowls J. A unified framework of five principles for AI in society. Harvard Data Science Review 2019;1(1):2-15. [doi: 10.1162/99608f92.8cd550d1]
60. Fjeld J, Achten N, Hilligoss H, Nagy A, Srikumar M. Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication 2020. [doi: 10.2139/ssrn.3518482]
61. Regulatory framework for artificial intelligence (AI Act). European Commission. URL: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai [accessed 2025-10-07]
62. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (Generative AI Delegation Taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res 2026;33(3):2544331. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]

Abbreviations

AI: artificial intelligence
ALTAI: Assessment List for Trustworthy Artificial Intelligence
EU: European Union
ISO/IEC TS: International Organization for Standardization/International Electrotechnical Commission Technical Specification
ML: machine learning
PD: Parkinson disease
TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis
UI: user interface
UX: user experience

XSL•FO
RenderX

Edited by A Coristine; submitted 07.Oct.2025; peer-reviewed by S Guan, A Amanna; comments to author 04.Dec.2025; revised version received 12.Feb.2026; accepted 13.Feb.2026; published 29.Apr.2026.

Please cite as:
Saad A, Dias SB, Alhussein G, Lyreskog D, Gerasimou I, Alves B, de Vos M, Drivas I, Zaras J, Stergioulas A, Bensenousi I, Hadjileontiadis L, Chatzichristos C, Hadjidimitriou S
Clinical AI is Not (Yet) Trustworthy-But It Could Be
J Med Internet Res 2026;28:e85433
URL: https://www.jmir.org/2026/1/e85433
doi:10.2196/85433
PMID:

©Ali Saad, Sofia B Dias, Ghada Alhussein, David Lyreskog, Ioannis Gerasimou, Beatriz Alves, Maarten de Vos, Ioannis Drivas, John Zaras, Andreas Stergioulas, Iskanter Bensenousi, Leontios Hadjileontiadis, Christos Chatzichristos, Stelios Hadjidimitriou. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Open-Source Large Language Models and AI Health Equity: A Health Service Triangle Model Perspective

Shaolong Wu[1,2,3*], PhD; Wenxin Zou[1*], MPH; Jiong Tu[4], PhD; Chunxiao Wang[5], PhD; Cheng Jin[6], BMed; Jing Liao[3,7], PhD; Kwok Cho Tang[3,7,8], PhD; Ning Liu[9,10*], PhD; Chun Hao[3,7*], PhD

[1]Department of Public Administration, School of Government, Sun Yat-sen University, Guangzhou, China
[2]Sun Yat-Sen Chinese Public Management Research Center, Sun Yat-sen University, Guangzhou, China
[3]Sun Yat-sen Global Health Institute, Institute of State Governance, Sun Yat-sen University, Guangzhou, Guangdong, China
[4]School of Sociology and Anthropology, Sun Yat-sen University, Guangzhou, China
[5]School of Health Management, Guangzhou Medical University, Guangzhou, China
[6]Party Committee, Huizhou Municipal Central People's Hospital, Huizhou, China
[7]Department of Medical Statistics, School of Public Health, Sun Yat-Sen University, Guangzhou, China
[8]Department of Public and International Affairs, City University of Hong Kong, Hong Kong, China (Hong Kong)
[9]School of Management, Lanzhou University, Lanzhou, China
[10]China Research Center for Government Performance-Management, Lanzhou University, Lanzhou, China
[*]these authors contributed equally

Corresponding Author:
Chun Hao, PhD
Department of Medical Statistics
School of Public Health
Sun Yat-Sen University
No. 74 Zhongshan er Road Guangzhou, China
Guangzhou, 510080
China
Phone: 86 02087335706
Email: haochun@mail.sysu.edu.cn

Abstract

This study explores the role of open-source large language models (LLMs) in promoting artificial intelligence (AI) health equity from the perspective of the health service triangle model. First, it defines AI health, categorizes AI-supported decision-making patterns, and assesses the status quo of AI health inequalities. Second, by comparing open-source and closed-source LLMs in terms of patient privacy, data security, accessibility, and use, it demonstrates the distinct advantages of open-source LLMs for AI-enabled health services. Finally, based on the health service triangle model, this study demonstrates how open-source LLMs drive the democratization of AI-enabled health services—particularly benefiting low-resource regions—by expanding service types, improving accessibility, enhancing quality, and reducing costs. This study concludes that, while open-source LLMs must address challenges such as hallucination risks and ethical responsibilities, they ultimately enable AI health equity through technological sharing.

(J Med Internet Res 2026;28:e86769) doi:10.2196/86769

KEYWORDS

artificial intelligence health; AI health; open-source large language models; open-source LLMs; health equity; health service triangle model; democratization; artificial intelligence; AI

Introduction

Artificial intelligence (AI) typically refers to computational technologies that mimic the mechanisms of human intelligence [1-4], such as thinking, deep learning, adaptation, interaction, and understanding [5]. Alternatively, it can be defined as a computational program or machine performing a task that, if accomplished by a human, would be considered intelligent [6]. AI technologies are being extensively used in the health sector to tackle the myriads of challenges confronting it. We use the term “AI health” to describe the application of AI technologies

XSL•FO
RenderX

in the health sector to influence health outcomes. The fundamental distinction between AI health and digital health lies in the fact that, in the former, decisions and actions directly related to health (at both the individual and population level) are made or assisted by AI (computer programs and machines) rather than being independently made by humans. In other words, the defining characteristic of AI health is that AI now assists—or, in some cases, even replaces—human decision-making in matters of health. This represents an unprecedented development in human history.

In the era of AI-assisted health care, decision-making patterns can be categorized into 4 types based on the extent of AI involvement in the health care process (Figure 1): traditional physician-patient interaction (in which physicians make diagnostic and treatment decisions without AI support relying solely on patient-provided information and conventional examinations), patient-AI interaction (where patients use AI tools to support their own health care self-management decisions), physician-AI interaction (in which physicians use AI for clinical decision support), and physician-AI-patient collaborative interaction (where physicians and patients jointly use AI to facilitate shared decision-making throughout the diagnostic and treatment process) [7,8].

Figure 1. Four types of physician–patient–artificial intelligence (AI) decision-making patterns. EMR: electronic medical record.

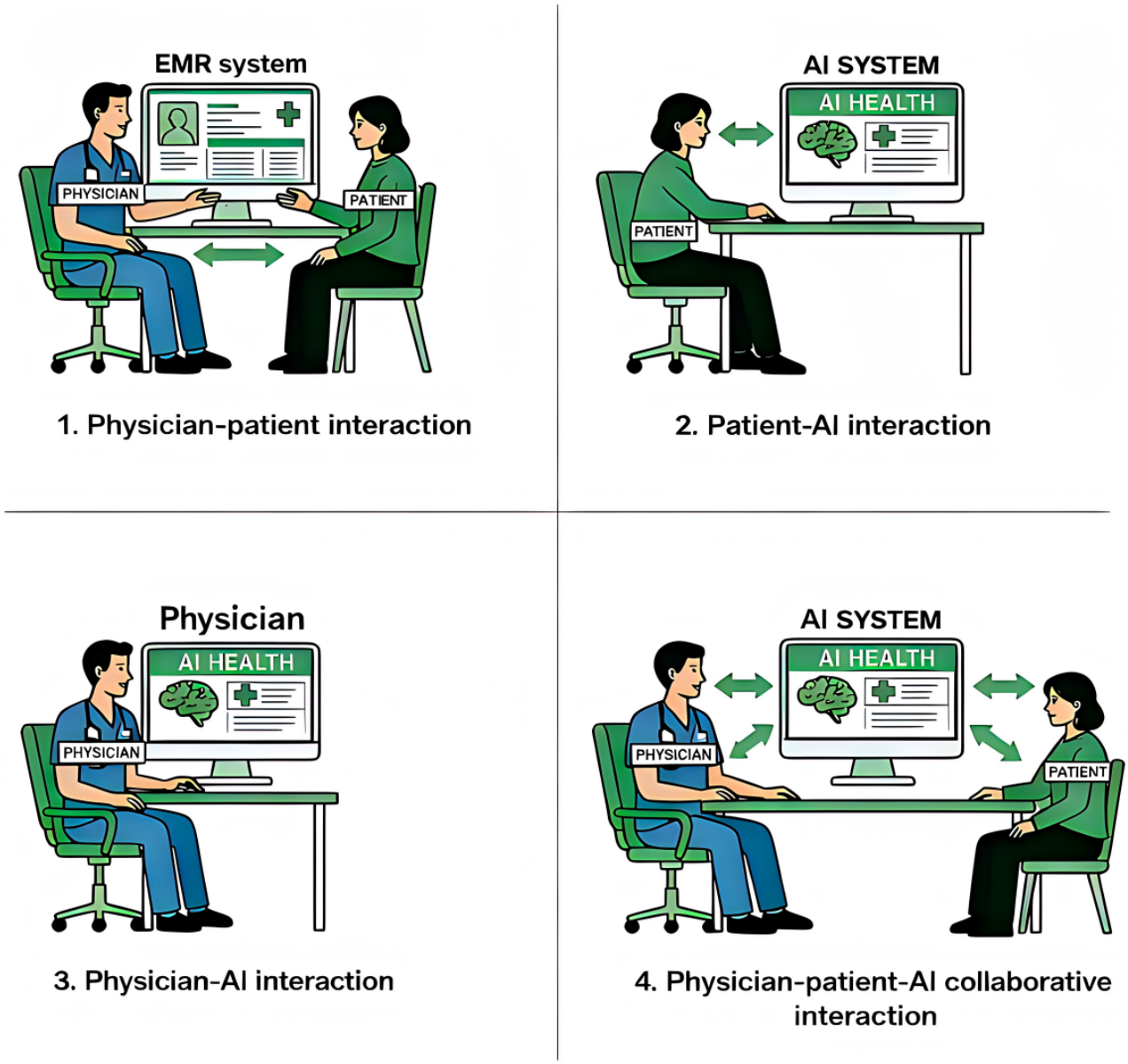

The transition from digital health to AI health represents a paradigm shift in health care, evolving from a physician-patient dyadic model to a triadic physician-patient-AI interaction model. This transformation constitutes an inevitable trend as it holds significant potential to address systemic challenges such as reducing unaffordable costs for patients, alleviating excessive workloads and physician burnout, and enhancing health care quality and operational efficiency [9]. By advancing toward a

XSL•FO
RenderX

more integrated, intelligent, and human-centered system, this shift can ultimately contribute to better health outcomes and sustainable well-being [10].

However, the development of AI technology is uneven across countries, which has led to disparities in its application in the health care sector, resulting in inequalities in health services and health outcomes [11]. According to the Artificial Intelligence Index Report 2025, the United States leads in AI model development with 40 models in 2024 and a dominant cumulative output over the past decade. China produced 15 models in 2024, with its leading model (eg, DeepSeek) achieving near–state-of-the-art performance. Europe released only 3 models, continuing a declining trend, whereas regions such as the Middle East, Latin America, and Southeast Asia contributed sporadically and minimally [5]. In terms of research on AI applications in health care, the United States holds significant influence, and AI is widely used in specialized fields such as radiology and pathology, optimizing clinical decision support systems. Following the release of the DeepSeek model, many of China's hospitals have deployed its public version [12]. In other countries, AI model development remains at an early stage and relies more heavily on international collaboration, resulting in only sporadic outputs; a similar pattern is seen in AI health applications, to which these countries have so far made only limited contributions [5].

In addition to gaps in AI research and health care deployment, access to and use of these technologies remain unequal across demographic groups. As the most advanced frontier of AI, large language models (LLMs) are predominantly controlled by a few leading companies and are close sourced to the public or health care providers. Given the restricted access to these LLMs, the resulting AI health applications are also typically released only to specific user groups by their developers. Similar to gaps observed in digital health [13-15], AI health inequalities stem from disparities in computational power (eg, graphics processing unit [GPU] chips), data resources (eg, standardized datasets), and algorithms (eg, LLM training frameworks). Gaps in access to AI health applications, differences in application skills, and uneven opportunities for use further exacerbate AI health inequity [16,17]. However, with DeepSeek leading a new wave of open-source initiatives [12], the widespread adoption of LLMs has the potential to fundamentally improve the foundation of AI health equity.

Why and how do open-source LLMs, compared to closed-source LLMs, improve AI health equity? Investigating this question helps policymakers, LLM companies and developers, hospitals, physicians, patients, and public health professionals understand the role and trends of AI in the health domain and actively promote the equitable access and application of open-source LLMs in health care. On the basis of a conceptualization of AI health and an analysis of the factors influencing AI health inequality, this study conducted a comparative analysis of the advantages of open-source LLMs over closed-source LLMs in improving AI health equity.. Specifically, using the health service triangle model as an analytical framework, this study compared open-source and closed-source LLMs across health service types, accessibility, quality, and cost. It argues that the local deployment of open-source LLMs offers unparalleled advantages in protecting patient privacy and data security, which in turn facilitates the development of more diverse health services, improves accessibility, enhances quality, and reduces costs. While open-source LLMs face challenges such as hallucinations, technology acceptance, and ethical issues, they are ultimately expected to improve AI health equity through technology sharing.

Closed-Source and Open-Source LLMs and AI Health Equity

The development of AI has historically followed 2 distinct paths: closed-source and open-source. The fundamental tension between these approaches reflects a divergence between technological democratization and commercial monopolization. The closed-source camp advocates for maintaining proprietary control over technology to ensure service quality, achieve commercial returns, safeguard model security and intellectual property, and optimize model performance. In contrast, the open-source approach promotes the sharing of model algorithms, code, and data to engage global developers in improving the models, thereby accelerating innovation and broadening accessibility.

In the early stages of LLM development, many academic institutions and small to medium enterprises embraced open-source principles. However, as LLMs advanced, major technology giants and unicorns increasingly adopted closed-source strategies. Although some leading firms—such as Meta with its Llama series—have pursued open-source releases, their models often lag in performance compared to proprietary alternatives and have not significantly challenged the dominance of closed-source approaches. However, with the emergence of high-performance open-source models such as DeepSeek, the global AI industry is experiencing an unprecedented wave of open-source innovation. Table 1 provides detailed comparisons.

XSL•FO
RenderX

Table 1. Open-source status of major large language models (compiled by the authors from publicly available information [retrieved via Google] as of August 31, 2025).

Model name	Publisher	Parameter scale	Open source announcement date
Gemma	Google DeepMind	2B-7B[a]	February 21, 2024
Mistral 7B version 0.2	Mistral AI	7B	March 24, 2024
DBRX	Databricks	132B	March 27, 2024
Nemotron-4	NVIDIA	340B	June 14, 2024
Llama 3.1	Meta	8B-405B	July 23, 2024
DeepSeek-V3	DeepSeek	671B	December 26, 2024
MiniMax-01	MiniMax	456B	January 15, 2025
DeepSeek-R1	DeepSeek	1.5B-660B	January 20, 2025
Qwen2.5-1 M	Alibaba Cloud	7B-14B	January 27, 2025
Gemma 3	Google DeepMind	1B-27B	March 12, 2025
Llama 4	Meta	109B-2T[b]	April 5, 2025
Qwen3	Alibaba Cloud	0.6B-235B	April 29, 2025
GPT-OSS-120B	OpenAI	120B	August 6, 2025
GPT-OSS-20B	OpenAI	20B	August 6, 2025

[a]Denotes billion parameters.

[b]Denotes trillion parameters.

The openness of LLMs can be categorized into 3 levels: fully open-source, partially open-source, and planned open-source. Fully open-source models not only release model weights but may also include training code, datasets, inference deployment solutions, and more. They allow individuals, enterprises, and institutions to use the models freely, including for commercial purposes, without requiring additional authorization. DeepSeek's series of models are fully open-source, permitting free use, local deployment, closed-source fine-tuning, and commercial applications. Fully open-source models have emerged as a key enabler for democratizing LLMs and fostering innovation in health applications. By lowering the barriers to accessing advanced LLMs, open-source initiatives allow diverse stakeholders—from regional hospitals to community clinics and individuals—to develop and deploy tailored health solutions. This accessibility helps reduce disparities in health care capabilities across different countries and demographic groups, thereby promoting health equity through broader and more equitable adoption of AI health technologies.

The direct factors influencing health include demographic and biological factors, environmental factors, behavioral factors, psychological factors, and health care factors [18,19]. The application of AI can impact most of these factors, thereby exerting a broad influence on health—a concept we define as "AI health," denoting the application of AI technologies within the health sector to shape health outcomes. AI health equity involves multiple levels and dimensions, ranging from individual to regional scales and from behavioral to health care factors. Inequities in foundational resources (computational power, data, and algorithms) and throughout the AI life cycle (research and development, application, access, and use) within health care drive disparities in health outcomes. This constitutes AI health inequity. Thus, compared to closed-source LLMs, open-source LLMs demonstrate greater potential to reduce these disparities and contribute to enhanced AI health equity. At present, AI primarily affects health care services, improving health through health education and promotion, prevention, treatment, and rehabilitation. Given that the impact of AI on health is a relatively recent development and the significant influence of open-source LLMs on health only emerged in early 2025, this paper focuses specifically on the impact of LLMs on health service equity.

For the health service industry, ensuring data security and patient privacy remains one of the most important concerns when applying LLMs [20-22]. A significant advantage of fully open-source LLMs is that they can be deployed locally, enabling health care institutions to maximize the protection of patient privacy and data security [21,23-26]. Therefore, for both the demand side (eg, patients and health care providers) and supply side (eg, developers and institutions), open-source LLMs can help address health service inequalities often associated with closed-source LLMs. By overcoming these key barriers, open-source LLMs pave the way for enhancing the scope, quality, and affordability of health services, contributing to a more comprehensive improvement in AI health equity on a global scale.

Framework for AI Health Equity: Health Service Triangle Model

In the field of health services, 2 models are widely recognized: one is the World Health Organization's health system model focusing on the health care supply side [27], and the other is the model by Andersen [28] emphasizing the health care demand side. By integrating both supply and demand perspectives, Wang

XSL•FO
RenderX

et al [29] developed a health service triangle model based on the essential common elements of these 2 frameworks to analyze online health services (for details, refer to Figure 2). Furthermore, this model can also be applied to evaluate the impact of open-source LLMs on AI health equity.

Figure 2. Health service triangle model.

The health service triangle model highlights health services and their 3 indispensable fundamental elements: accessibility, quality, and cost. Health services form the core of the model, and the type of service determines its accessibility, quality, and cost. What kinds of health services can LLMs actually help provide? Without health services, there is no AI health—the type of health service is the foundation for evaluating AI health equity. For instance, emergency and urgent care services prioritize accessibility. If quality and cost are prioritized instead, delayed treatment could lead to severe health consequences. Accessibility includes geographic, temporal, economic, social, and eligibility dimensions [30,31]. Certain health services delivered via LLMs through online channels hold significant advantages in accessibility. Users can overcome spatiotemporal constraints, accessing services anonymously, inexpensively, or even free of charge around the clock [29].

Health service quality refers to the extent to which health care professionals and health care organizations deliver care to individuals and populations in ways that align with the best expected health outcomes achievable given current medical knowledge and technology [32]. LLMs have the potential to significantly enhance diagnosis, prescription, and treatment and are expected to address issues such as physician shortages, substandard care, and medical errors. The rapid rise in health care costs, which often makes care unaffordable for patients, has long been a fundamental barrier to the sustainable development of health systems. LLM technology, capable of efficiently processing medical images, assisting in diagnosis, discovering drugs, and improving medical education and research, is also seen as a promising tool to curb escalating medical expenses [9]. In terms of indirect costs—such as travel expenses—LLM-powered telemedicine can significantly reduce these burdens.

On the basis of the nature of health services and their triangular attributes, we established a comparative analysis framework for closed-source and open-source LLMs (for details, refer to Table 2). Compared to closed-source LLMs, open-source LLMs generate more positive social outcomes in terms of health service types, accessibility, quality, and cost—leading to a fairer distribution across regions and populations. In summary, the advantages of open-source LLMs over closed-source ones—in terms of health services and their 3 core attributes—enable AI to be widely used to meet various health service needs of both providers and users. This is particularly beneficial for remote and rural areas, financially constrained institutions, and privacy-sensitive applications. While closed-source LLMs can only offer limited generalized services targeted at users with payment capacity, open-source LLMs advance AI health equity by broadening access through more diverse and affordable services.

Table 2. Comparative analysis of artificial intelligence health equity: closed- and open-source large language models (LLMs).

Health service	Closed-source LLMs	Open-source LLMs
Types	Limited availability	Wider variety
Accessibility	Restricted (API[a] dependent and paywall)	Flexible (local deployment and customizable)
Quality	Consistent but generic	Variable but adaptable
Cost	Recurring fees (pay per use)	Low licensing cost and high setup cost

[a]API: application programming interface.

XSL•FO
RenderX

This study used a multimethod approach to compare open-source and closed-source LLMs across 4 dimensions of health services: types, accessibility, quality, and cost. Data were collected through web-based surveys; document research; and a systematic literature review, which included 31 AI model web pages (2020-2025), 7 DeepSeek technical documents, and 80 academic publications. The analysis quantified the relative strengths and weaknesses of each model type, demonstrating how the open-source paradigm promotes the use of AI-supported health services and contributes to improved health outcomes. The analysis shows that open-source LLMs are rapidly empowering both supply-side (eg, developers and institutions) and demand-side (eg, patients and health care professionals) actors, especially those in low- and middle-income countries (LMICs), to adopt AI in addressing their respective challenges. As AI-enabled health services become more widespread, this trend supports the achievement of AI health equity on a broader scale.

Comparative Analysis of AI Health Equity: Closed- and Open-Source LLMs

Service Types

With its advancement into the generative AI era, AI appears capable of performing nearly all tasks that humans can accomplish. In the health care sector, almost all health-related decisions and tasks seem achievable or assistable through LLMs. China's National Health Commission released the Reference Guide for Artificial Intelligence Application Scenarios in the Health Industry in 2024, summarizing AI applications across 13 categories and 84 specific scenarios in the health sector. The 13 application categories include medical services, pharmaceutical services, health insurance services, traditional Chinese medicine management, hospital administration, health management, public health services, older adult care and childcare, medical robots, drug development, traditional Chinese medicine industry, medical education, and medical research. The 84 scenarios further break down each category into specific application areas [33].

In other countries, AI applications in the health sector—primarily in medical diagnosis and decision support (eg, disease classification and prediction, clinical reasoning and planning, and dynamic referral), medical image processing, electronic health record and text analysis, bioinformatics and drug development, public health, and medical education [2,34-36]—vary according to local health challenges and system structures. Despite the broad range of applications, medical diagnosis and treatment remains the core function of AI in health care. However, it is important to note that most AI applications in health care are still in the research and testing phases and have not yet been widely integrated into routine clinical practice [37]. With the rise of open-source LLMs and their local deployment, many of these experimental applications are gradually being implemented in real clinical settings [38]. AI-supported health services, such as other health technologies throughout history, are expected to eventually achieve widespread adoption and popularization.

For individual users, there is no fundamental difference in the types of health services provided by closed-source and open-source LLMs. However, for enterprises and health care institutions, closed-source LLMs (such as GPT-4) are constrained by vendor limitations. Even when used for developing professional health care software, the range of services they can offer remains restricted. In contrast, open-source LLMs (eg, DeepSeek-R1) can be used commercially without vendor-imposed limitations, enabling health care organizations to deploy them locally and develop a wide variety of specialized or personalized health services tailored to local contexts. This flexibility overcomes the application programming interface (API) dependencies and paywalls inherent to closed-source LLMs, allowing institutions to better address specific health care needs.

Accessibility

Accessing closed-source LLMs via the internet typically requires account registration and payment. Some leading LLM companies impose IP blocks on specific countries, preventing users from those regions from registering or using their services. Credit card payments may be prohibited, and API calls might be restricted or return simplified results. Certain countries have also implemented policies to restrict the export of key technologies related to LLMs—such as chips and model weights—to maintain their technological advantage and monopolistic position [39]. In contrast, fully open-source LLMs, such as those released by DeepSeek—which provide both model weights and code, along with distilled versions ranging from 1.5B to 671B parameters—enable worldwide access. Institutions and individuals worldwide can not only directly use these models but also deploy them locally if they possess sufficient computational resources. This increases the likelihood that patients will be able to access AI-provided or AI-assisted decision-making to address their medical needs. In other words, both health care supply-side and demand-side stakeholders benefit from the improved accessibility enabled by open-source LLMs.

Open-source LLMs are typically released under open licenses such as the MIT License or Apache License 2.0, permitting unrestricted commercial and research use. After local deployment, the transparency of model weights and architecture enables autonomous modifications, allowing for secondary development tailored to specific domains—such as model distillation and fine-tuning to adapt to particular health care scenarios and tasks [40]. In addition to enabling offline use, open-source LLMs crucially avoid the need to upload data to external models, thereby ensuring data security and personal privacy. As a result, within just a few months of DeepSeek openly releasing its models, hundreds of hospitals in China announced that they had implemented local deployments. The widespread and decentralized deployment of open-source LLMs also mitigates systemic risks associated with monopolistic control, service failures, or errors from a single provider [41]. In contrast to closed-source LLMs, which rely on proprietary licensing, keep architectural and training parameters confidential, require high-performance hardware, and involve data uploads, open-source LLMs significantly promote the democratization of AI technology. This advancement facilitates

XSL•FO
RenderX

the broader application of AI in health care and contributes to global AI health equity.

Quality of Care

LLMs can evaluate the standardization level of medical images such as x-rays and computed tomography (CT) scans and generate corresponding reports. DeepSeek-R1, for instance, has demonstrated excellent performance in auditing CT reports [42]. By integrating textual descriptions with imaging data, LLMs can also generate diagnostic suggestions. Gemini 2.0 Flash has shown leading capabilities in quality control tasks for chest x-rays [43]. Through analyzing patient symptoms, electronic health records, and laboratory data, LLMs can classify chronic diseases—such as diabetes, cancer, and heart disease—and predict potential risks. For example, DeepSeek-R1 achieved 100% accuracy in mental health and oncology classification tasks [20]. Multimodal AI systems that combine CT imaging with biomarkers (eg, epidermal growth factor receptor gene mutations) have reached a 95.2% accuracy rate in distinguishing benign from malignant pulmonary nodules, surpassing the average performance of human physicians [38].

In terms of clinical reasoning and decision-making, LLMs support multistep diagnostic reasoning, such as generating examination suggestions and treatment plans. When provided with complete examination results, models such as DeepSeek-R1 achieve over 85% accuracy. However, their performance declines significantly in highly complex tasks involving examination recommendations and therapeutic planning. Evaluations on the MedR-Bench dataset reveal that LLMs still have limitations in complex scenarios and require integration with expert knowledge [23]. While LLMs demonstrate high accuracy in medical diagnosis and treatment [20,23,44], it is noteworthy that clinical decisions made solely by LLMs have, in certain cases, outperformed collaborations between physicians and LLMs [45].

Both closed-source and open-source LLMs are trained on publicly available standardized health care datasets and are dynamically updated with newly released medical guidelines and knowledge. Thus, the quality of general health services provided by both types of LLMs should be comparable. However, when open-source LLMs are deployed locally and fine-tuned using regional medical knowledge and case data, they can be adapted to highly specific medical scenarios and specialized cases. With sufficient local medical expertise, the quality of specialized health care services offered by open-source LLMs may even surpass that of closed-source alternatives.

Cost

Closed-source LLMs require ongoing use fees (volume-based or time-based billing), whereas open-source models involve low use and licensing costs but higher initial deployment investment. Individual users can access open-source LLMs free of charge, whereas accessing closed-source models typically requires continuous payment. If APIs are called, both open-source and closed-source models may incur costs, but open-source options are generally much more affordable. Most importantly, open-source models only require a one-time deployment investment with no subsequent use fees.

Accessing the official websites of open-source LLMs is free for individual users; using the basic versions of closed-source LLMs may also be free. However, the advanced versions and complex functionalities of closed-source LLMs generally require payment. Both individuals and enterprises are charged for API calls to LLMs regardless of whether they are open-source or closed-source. However, open-source LLMs are significantly more affordable. For instance, the API cost for DeepSeek-R1 is US $0.55 per million input tokens and US $2.19 per million output tokens, which is only 6.71% of the cost of OpenAI's GPT-o1 model [46]. Unlike the ongoing expenses associated with using closed-source LLMs, locally deploying an open-source LLM involves a one-time deployment cost rather than continuous payments to service providers. The scale of this investment depends on the size of the model being deployed and the required computational resources. Taking DeepSeek-R1 as an example, its 1.5B parameter model can even be deployed on an iPhone. In contrast, deploying the largest 671B parameter version demands substantial hardware resources—including at least 8 GPUs, a server-grade central processing unit, storage equipment, and networking infrastructure—with a minimum cost of US $1 million.

Why can open-source LLMs offer free access to their websites and low-cost API use? This is primarily because open-source LLMs benefit from significantly lower research, development, and training costs, enabling sustainable business practices. For instance, DeepSeek has substantially reduced its research and development expenses through algorithmic innovations and engineering optimizations [47]. In comparison to closed-source LLMs, the pretraining cost of DeepSeek-V3 was approximately US $5.576 million (based on H800 GPU rentals), significantly lower than the US $60 million spent on training Meta's Llama 3.1 [48]. By adopting a mixture-of-experts architecture and efficient training strategies, DeepSeek has compressed training resource demands, reducing its training costs by over 42.5% compared to closed-source LLMs such as GPT-4 or Claude 3.5 [49]. Furthermore, open-source LLMs benefit from continuous improvements driven by the global AI community, which further enhances their efficiency and reduces expenses. Historical trends suggest that, as open-source LLM technology becomes more widely adopted and popularized, AI's impact on health will increasingly promote equity.

Discussion and Conclusions

Since its emergence, AI technology has been applied in the health care domain, evolving from early expert systems to the current variety of diagnostic and treatment support systems, which primarily embody the physician-AI interaction pattern. However, the widespread and direct engagement of patients with AI technology became prominent only with the advent of LLMs. This trend accelerated significantly following the open sourcing of LLMs, facilitating the patient-AI interaction pattern, where patients began using these models extensively for consultation, self-diagnosis, and even prescription information gathering. The proliferation of these interaction patterns has

XSL•FO
RenderX

propelled AI health into a widespread phenomenon, marking a new and distinct phase in the evolution of digital health. In terms of accessibility, quality, and cost, the greatest advantage of open-source LLMs over closed-source alternatives lies in their ability to make both the physician-AI interaction and patient-AI interaction patterns more widespread and equitable.

Open-source LLMs not only promote AI health equity in traditional dimensions—such as across nations, regions, and populations—but also, by enabling shared access for both health care providers and patients, foster equity in a new dimension: a more balanced clinician-patient relationship within the physician-AI-patient interaction pattern. In traditional health care systems, only medical professionals hold the authority to make diagnostic and treatment decisions, whereas patients are typically excluded from participating in choices concerning their own health and disease management. A significant information asymmetry exists between physicians and patients regarding health and medical knowledge, resulting in highly unequal power dynamics during clinical decision-making.

The emergence of LLMs, especially open-source LLMs, has disrupted this imbalance. A growing number of patients now upload their laboratory results, medical images, and even full diagnostic records to LLMs seeking diagnostic interpretations and management recommendations based on their symptoms, history, and clinical data. Although the output of these models is intended for reference only, their high diagnostic accuracy—often surpassing that of average physicians—has led some patients to use LLM-generated insights to facilitate discussions with their physicians regarding diagnosis and treatment plans. Others rely on such suggestions to evaluate the appropriateness of the care recommended by their clinicians. On the health care supply side, medical providers have already begun integrating LLMs into professional workflows, making decisions based on AI-generated assessments or even fully adopting AI-proposed solutions. Introducing open-source LLM–supported diagnostics into medical decision-making enables shared decision-making between clinicians and patients based on AI-informed insights. This transforms the conventional care paradigm, leading to a more equitable relationship between patients and providers throughout the diagnostic and therapeutic process.

The technological democratization driven by open-source LLMs holds significance not only for countries such as China that are leading the open-source movement but also for global health. A core objective of global health is health equity, and addressing the growing disparities in health outcomes worldwide is a shared challenge for all nations. Open-source LLMs substantially enhance the accessibility of AI technology for both supply- and demand-side actors in health care. As long as patients have basic internet-connected devices such as smartphones, they can access open-source LLMs. Similarly, health care providers such as hospitals can deploy suitable models if they possess fundamental computing resources. This is particularly transformative for hospitals, physicians, and patients in low- and middle-income countries as open-source LLMs can effectively improve the accessibility and quality of their health services. Countries previously marginalized by expensive commercial systems can now leverage mobile apps to facilitate initial self-diagnosis [50,51], thereby advancing progress toward universal health coverage.

While open-source LLMs hold significant promise for improving AI health equity, their application in the medical field still faces considerable challenges. Open-source LLMs are prone to a certain degree of AI hallucination, making it crucial to ensure the accuracy and reliability of the medical advice they generate. Some patients may engage in self-medication based on LLM-generated recommendations [50], and errors in such advice could be life-threatening. Although open-source LLMs emphasize that their suggestions are “for reference only” and advise users to “follow professional medical guidance,” legal and ethical questions regarding accountability in cases of medical errors or malpractice remain unresolved [37]. Regarding the openness of LLMs and their diverse users, corresponding ethical and governance strategies should be established. For instance, in the case of deploying fully open-source LLMs, the responsibility for model custodianship, updates, and accountability for application outcomes lies with the deployer rather than the developer.

The comprehensive and standardized recommendations generated by open-source LLMs can mitigate information asymmetry between physicians and patients, yet they may also inadvertently undermine trust in clinical authority. Moreover, integrating even the smallest-parameter open-source LLMs into existing health information systems demands dedicated staff training and infrastructure investment. Meanwhile, the deployment of variably scaled models—such as 70B vs 671B parameter versions—contingent on local hardware capacity may perpetuate or even exacerbate disparities in AI adoption. To address these challenges, global policymakers must collaborate with stakeholders across sectors to establish robust governance frameworks for medical AI. Such frameworks should proactively tackle technical, financial, and ethical barriers, ensuring that AI-driven health services are effectively deployed worldwide and equitably accessible, thereby advancing health outcomes for all.

Acknowledgments

The authors acknowledge that ima.copilot (Tencent and DeepSeek) was used to assist in language editing and improving the clarity of the manuscript.

Funding

The authors declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the following grants: 2025 major project of the National Social Science Fund of China (project 25&ZD151),

XSL•FO
RenderX

2024 National Social Science Fund of China (project 24SGB024), the Guangdong Philosophy and Social Science Foundation Planning Project (General Project) 2025 (project GD25DWQZZ01), and the Major Project of the Key Research Institutes of Humanities and Social Sciences of the Ministry of Education (project 22JJD630022).

Authors' Contributions

SW, NL, and CH contributed substantially to the conception and design of the study. WZ, JT, CW, CJ, GZ, and JL were responsible for data collection and analysis and drafting and revising the manuscript. All authors approved the final version and agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

References

1. Wang F, Preininger A. AI in health: state of the art, challenges, and future directions. Yearb Med Inform 2019 Aug;28(1):16-26 [FREE Full text] [doi: 10.1055/s-0039-1677908] [Medline: 31419814]
2. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. BMC Med Inform Decis Mak 2021 Apr 10;21(1):125 [FREE Full text] [doi: 10.1186/s12911-021-01488-9] [Medline: 33836752]
3. Tagliaferri SD, Angelova M, Zhao X, Owen PJ, Miller CT, Wilkin T, et al. Artificial intelligence to improve back pain outcomes and lessons learnt from clinical classification approaches: three systematic reviews. NPJ Digit Med 2020 Jul 09;3:93 [FREE Full text] [doi: 10.1038/s41746-020-0303-x] [Medline: 32665978]
4. Tran BX, Vu GT, Ha GH, Vuong QH, Ho MT, Vuong TT, et al. Global evolution of research in artificial intelligence in health and medicine: a bibliometric study. J Clin Med 2019 Mar 14;8(3):360 [FREE Full text] [doi: 10.3390/jcm8030360] [Medline: 30875745]
5. Maslej N, Fattorini L, Perrault R, Gil Y, Parli V, Kariuki N, et al. The AI index 2025 annual report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University. 2025. URL: https://hai-production.s3.amazonaws.com/files/hai_ai_index_report_2025.pdf [accessed 2026-04-08]
6. Morley J, Machado CC, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: a mapping review. Soc Sci Med 2020 Sep;260:113172. [doi: 10.1016/j.socscimed.2020.113172] [Medline: 32702587]
7. Lorenzini G, Arbelaez Ossa L, Shaw D, Elger B. Artificial intelligence and the doctor-patient relationship expanding the paradigm of shared decision making. Bioethics 2025 Dec 12 (forthcoming). [doi: 10.1111/bioe.13128] [Medline: 41385350]
8. Huang Y, Tan Z. Ethical reconstruction of doctor-patient relationships through generative artificial intelligence. Med Philos 2024;45(3):49-53.
9. Meskó B, Drobni Z, Bényei É, Gergely B, Győrffy Z. Digital health is a cultural transformation of traditional healthcare. Mhealth 2017 Sep 14;3:38 [FREE Full text] [doi: 10.21037/mhealth.2017.08.07] [Medline: 29184890]
10. Ueda D, Walston SL, Fujita S, Fushimi Y, Tsuboyama T, Kamagata K, et al. Climate change and artificial intelligence in healthcare: review and recommendations towards a sustainable future. Diagn Interv Imaging 2024 Nov;105(11):453-459 [FREE Full text] [doi: 10.1016/j.diii.2024.06.002] [Medline: 38918123]
11. Bright TJ, Norris KC. Artificial intelligence and health equity. Annu Rev Med 2026 Jan;77(1):525-536 [FREE Full text] [doi: 10.1146/annurev-med-043024-125309] [Medline: 41115175]
12. With the open-source wave sweeping the globe, large-scale models urgently need to transform into "commercialization 2.0"? Tencent News. 2025 Apr 08. URL: https://news.qq.com/rain/a/20250408A073KJ00 [accessed 2026-04-14]
13. van Dijk JA. Digital divide research, achievements and shortcomings. Poetics 2006;34(4-5):221-235. [doi: 10.1016/j.poetic.2006.05.004]
14. Riggins FJ, Dewan S. The digital divide: current and future research directions. J Assoc Inf Syst 2005 Dec;6(12):298-337. [doi: 10.17705/1jais.00074]
15. Van Dijk JA. Digital divide: impact of access. In: Rössler P, Hoffner CA, van Zoonen L, editors. The International Encyclopedia of Media Effects. Hoboken, NJ: John Wiley & Sons; 2017.
16. Egger J, De Paiva LF, Luijten G, Krittanawong C, Keyl J, Sallam M, et al. Is DeepSeek-R1 a game changer in healthcare? - A seed review. TechRxiv. Preprint posted online on March 24, 2025 [FREE Full text] [doi: 10.36227/techrxiv.174285174.49544231/v1]
17. Vera CL, Picon IF, del Val Nunez MT, Gandia JA, de Lucas Ancillo A, Arroyo VR, et al. Evaluating large language models on the Spanish Medical Intern Resident (MIR) examination 2024/2025:a comparative analysis of clinical reasoning and knowledge application. arXiv. Preprint posted online on February 25, 2025 . [doi: 10.48550/arXiv.2503.00025]
18. Zhang G, Cai C, Zou W, Jing L, Wu S. Depressive symptoms and socioeconomic status among the labor force: evidence from China's representative sample. PLoS One 2022 Aug 22;17(8):e0272199 [FREE Full text] [doi: 10.1371/journal.pone.0272199] [Medline: 35994489]

XSL•FO
RenderX

19. Solar O, Irwin A. A conceptual framework for action on the social determinants of health. World Health Organization. 2010. URL: https://iris.who.int/bitstreams/ca294183-3263-470f-a5fe-8e124ec48c72/download [accessed 2026-04-08]
20. Gupta GK, Pande P, Acharya N, Singh AK, Niroula S. LLMs in disease diagnosis: a comparative study of DeepSeek-R1 and O3 mini across chronic health conditions. arXiv. Preprint posted online on March 13, 2025 . [doi: 10.48550/arXiv.2503.10486]
21. Meskó B. DeepSeek AI: what does it mean for healthcare? Linkedin. 2025 Feb 25. URL: https://www.linkedin.com/pulse/deepseek-ai-what-does-mean-healthcare-bertalan-mesk%C3%B3-md-phd-cusfe [accessed 2025-03-24]
22. Poo MM. Reflections on DeepSeek's breakthrough. Natl Sci Rev 2025 Mar;12(3):nwaf044. [doi: 10.1093/nsr/nwaf044] [Medline: 40041025]
23. Qiu P, Wu C, Liu S, Fan Y, Zhao W, Chen Z, et al. Quantifying the reasoning abilities of LLMs on real-world clinical cases. arXiv. Preprint posted online on March 6, 2025 . [doi: 10.1038/s41467-025-64769-1]
24. Wu J. The rise of DeepSeek: technology calls for the "catfish effect". J Thorac Dis 2025 Feb 28;17(2):1106-1108 [FREE Full text] [doi: 10.21037/jtd-2025b-02] [Medline: 40083508]
25. Temsah A, Alhasan K, Altamimi I, Jamal A, Al-Eyadhy A, Malki KH, et al. DeepSeek in healthcare: revealing opportunities and steering challenges of a new open-source artificial intelligence frontier. Cureus 2025 Feb 18;17(2):e79221. [doi: 10.7759/cureus.79221] [Medline: 39974299]
26. Will DeepSeek AI drive a healthcare revolution? The Week. 2025 Jan 28. URL: https://www.theweek.in/news/health/2025/01/28/will-deepseek-ai-drive-a-healthcare-revolution.html [accessed 2025-03-24]
27. Monitoring the Building Blocks of Health Systems: A Handbook of Indicators and Their Measurement Strategies. Geneva, Switzerland: World Health Organization; 2011.
28. Andersen RM. Revisiting the behavioral model and access to medical care: does it matter? J Health Soc Behav 1995 Mar;36(1):1-10. [Medline: 7738325]
29. Wang K, Zou W, Lai Y, Hao C, Liu N, Ling X, et al. Accessibility, cost, and quality of an online regular follow-up visit service at an internet hospital in China: mixed methods study. J Med Internet Res 2024 Oct 21;26:e54902 [FREE Full text] [doi: 10.2196/54902] [Medline: 39432365]
30. Gilson L, Russell S, Buse K. The political economy of user fees with targeting: developing equitable health financing policy. J Int Dev 1995;7(3):369-401. [doi: 10.1002/jid.3380070305]
31. Aday LA, Andersen R. A framework for the study of access to medical care. Health Serv Res 1974;9(3):208-220 [FREE Full text] [Medline: 4436074]
32. Donabedian A. The quality of care: how can it be assessed? JAMA 1988 Sep 23;260(12):1743-1748 [FREE Full text] [doi: 10.1001/jama.1988.03410120089033]
33. Notice from the General Office of the National Health Commission on issuing reference guidelines for artificial intelligence application scenarios in the health sector. National Health Commission of the People's Republic of China. 2024. URL: https://www.nhc.gov.cn/guihuaxxs/c100133/202411/3dee425b8dc34f739d63483c4e5c334c.shtml [accessed 2025-05-13]
34. Shaheen MY. Applications of artificial intelligence (AI) in healthcare: a review. ScienceOpen Preprints. Preprint posted online on September 25, 2021 [FREE Full text] [doi: 10.14293/S2199-1006.1.SOR-.PPVRY8K.v1]
35. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. Lancet 2020 May 16;395(10236):1579-1586 [FREE Full text] [doi: 10.1016/S0140-6736(20)30226-9] [Medline: 32416782]
36. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Jun 21;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
37. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ 2025 Feb 05;388:e081554 [FREE Full text] [doi: 10.1136/bmj-2024-081554] [Medline: 39909534]
38. Chen J, Zhang Q. DeepSeek reshaping healthcare in China's tertiary hospitals. ArXiv. Preprint posted online on February 23, 2025 arXiv preprint: arXiv:2502.16732 [FREE Full text] [doi: 10.48550/arXiv.2502.16732]
39. BIS announces new regulatory framework for AI and controls on advanced computing technology and AI models. Wiley Rein. 2025 Jan 17. URL: https://www.wiley.law/alert-BIS-Announces-New-Regulatory-Framework-for-AI-and-Controls-on-Advanced-Computing-Technology-and-AI-Models [accessed 2025-05-14]
40. Sapkota R, Raza S, Karkee M. Comprehensive analysis of transparency and accessibility of ChatGPT, DeepSeek, and other SoTA large language models. ArXiv. Preprint posted online on February 21, 2025 [FREE Full text] [doi: 10.48550/arXiv.2502.18505]
41. Tang X, Shao D, Sohn J, Chen J, Zhang J, Xiang J, et al. MedAgentsBench: benchmarking thinking models and agent frameworks for complex medical reasoning. ArXiv. Preprint posted online on March 10, 2025 [FREE Full text] [doi: 10.48550/arXiv.2503.07459]
42. Qin Z, Gui Q, Bian M, Wang R, Ge H, Yao D, et al. Multimodal human-AI synergy for medical imaging quality control: a hybrid intelligence framework with adaptive dataset curation and closed-loop evaluation. ArXiv. Preprint posted online on March 10, 2025 [FREE Full text] [doi: 10.48550/arXiv.2503.07032]

XSL•FO
RenderX

43. Zhan Z, Zhou S, Zhou H, Deng J, Hou Y, Yeung J, et al. An evaluation of DeepSeek models in biomedical natural language processing. ArXiv. Preprint posted online on March 1, 2025 [FREE Full text] [doi: 10.48550/arXiv.2503.00624]
44. Hussain Z, Delsoz M, Elahi M, Jerkins B, Kanner E, Wright C, et al. Performance of DeepSeek, Qwen 2.5 MAX, and ChatGPT assisting in diagnosis of corneal eye diseases, glaucoma, and neuro-ophthalmology diseases based on clinical case reports. medRxiv. Preprint posted online on March 17, 2025 [FREE Full text] [doi: 10.1101/2025.03.14.25323836] [Medline: 40166547]
45. Hswen Y, Rubin R. An AI chatbot outperformed physicians and physicians plus AI in a trial-what does that mean? JAMA 2025 Jan 28;333(4):273-276. [doi: 10.1001/jama.2024.23860] [Medline: 39729340]
46. Mikhail D, Farah A, Milad J, Nassrallah W, Mihalache A, Milad D, et al. Performance of DeepSeek-R1 in ophthalmology: an evaluation of clinical decision-making and cost-effectiveness. Br J Ophthalmol 2025 Aug 20;109(9):976-981. [doi: 10.1136/bjo-2025-327360] [Medline: 40701781]
47. DeepSeek-AI. DeepSeek-V3 technical report. ArXiv. Preprint posted online on December 27, 2024 [FREE Full text] [doi: 10.48550/arXiv.2412.19437]
48. Gibney E. China's cheap, open AI model DeepSeek thrills scientists. Nature 2025 Jan 23;638(8049):13-14. [doi: 10.1038/d41586-025-00229-6] [Medline: 39849139]
49. Abraham R. Democratizing AI's frontiers: a critical review of DeepSeek AI's open-source ecosystem. OSF Preprints. Preprint posted online on January 31, 2025 [FREE Full text] [doi: 10.31219/osf.io/9r7e3_v1]
50. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors 2023 May 17;10:e47564 [FREE Full text] [doi: 10.2196/47564] [Medline: 37195756]
51. Choudhury A, Shahsavar Y, Shamszare H. User intent to use DeepSeek for health care purposes and their trust in the large language model: multinational survey study. JMIR Hum Factors 2025 May 26;12:e72867 [FREE Full text] [doi: 10.2196/72867] [Medline: 40418796]

Abbreviations

AI: artificial intelligence
API: application programming interface
CT: computed tomography
GPU: graphics processing unit
LLM: large language model

Edited by A Mavragani; submitted 29.Oct.2025; peer-reviewed by O Oluwole, K Khashu, S Saini; comments to author 01.Jan.2026; revised version received 01.Apr.2026; accepted 03.Apr.2026; published 29.Apr.2026.

Please cite as:
Wu S, Zou W, Tu J, Wang C, Jin C, Liao J, Tang KC, Liu N, Hao C
Open-Source Large Language Models and AI Health Equity: A Health Service Triangle Model Perspective
J Med Internet Res 2026;28:e86769
URL: https://www.jmir.org/2026/1/e86769
doi:10.2196/86769
PMID:

©Shaolong Wu, Wenxin Zou, Jiong Tu, Chunxiao Wang, Cheng Jin, Jing Liao, Kwok Cho Tang, Ning Liu, Chun Hao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Pediatric Clinical Images Without Consent: A Governance Gap in the Long-Term Reuse of Health Data in Digital Health Ecosystems

Iwona Bujek[1,2], PhD

[1]Poznan University of Medical Sciences, Poznan, Poland
[2]GGZ Voor Elkaar, Mental Health Institution, Utrecht, The Netherlands

Corresponding Author:
Iwona Bujek, PhD
Poznan University of Medical Sciences
ul. Bukowska 70
Poznan, 61-812
Poland
Phone: 48 61 854 60 00
Email: iwona.bujek@gmail.com

Abstract

Digital health governance frameworks have primarily focused on prospective safeguards, including informed consent at the point of data collection, lawful processing, and data security. Comparatively less attention has been devoted to the long-term circulation of legacy clinical materials, particularly pediatric clinical images reused across educational and digital infrastructures. This viewpoint examines governance challenges associated with the prolonged educational and digital reuse of pediatric clinical images without identifiable evidence of consent. Drawing on a longitudinal case spanning more than 3 decades (1991-2026), this article illustrates how clinical images may continue circulating across textbooks, educational repositories, conference materials, e-books, and online teaching platforms long after their original creation and publication context. The case is informed by archival educational materials, institutional correspondence, publisher communications, and formal regulatory findings, including a decision issued by the Polish Patient Rights Ombudsman confirming continuing violations related to dissemination of intimate pediatric clinical images without identifiable consent. This article argues that current digital health governance frameworks remain insufficiently equipped to address persistence, traceability, provenance, and coordinated withdrawal of legacy clinical materials once they enter distributed educational ecosystems. Fragmented accountability across health care institutions, publishers, educational systems, libraries, repositories, and digital platforms may allow sensitive clinical materials to remain accessible despite regulatory intervention or removal requests. The article further discusses how publicly accessible educational materials may become incorporated into downstream artificial intelligence and machine learning ecosystems through digitization, aggregation, web scraping, and secondary dataset reuse. In this context, unresolved historical consent deficiencies may become embedded within artificial intelligence–enabled infrastructures without effective provenance tracking or remediation mechanisms. To address these limitations, this viewpoint proposes a lifecycle-oriented governance framework emphasizing long-term consent traceability, provenance-aware dissemination systems, verification checkpoints before reuse or republication, periodic review of legacy educational archives, and coordinated cross-platform withdrawal procedures.

(J Med Internet Res 2026;28:e94952) doi:10.2196/94952

KEYWORDS

pediatric clinical images; informed consent; digital health governance; legacy data; medical education; artificial intelligence; patient privacy; data ethics

Introduction

Clinical images are integral to medical education, research, and clinical communication. In digital environments, however, their circulation may extend far beyond the original context of care or publication. Once incorporated into textbooks, conference materials, online repositories, educational platforms, and digital archives, clinical images may persist for decades and move across jurisdictions, institutions, and technological systems [1-6]. Digital health governance frameworks increasingly address the ethical management of health data, secondary data use, and artificial intelligence (AI) applications in medicine

XSL•FO
RenderX

[7-11]. These frameworks primarily emphasize prospective safeguards, including informed consent, lawful processing, data security, and accountability [12-18]. By contrast, comparatively less attention has been devoted to the long-term persistence and downstream reuse of legacy educational materials after their initial publication. This governance gap is particularly significant for clinical images because identifiability may persist even after partial anonymization. Existing publication guidance emphasizes that masking the eye region alone may not adequately protect anonymity and that consent should be obtained whenever identifiability cannot reasonably be excluded [1-6]. Recent studies similarly report uncertainty among clinicians, publishers, and editors regarding acceptable standards for publication and reuse of patient photographs in digital environments [3,4]. Within educational ecosystems, responsibility for consent verification is frequently fragmented. Authors, publishers, universities, libraries, repositories, and digital platforms operate within partially overlapping but insufficiently coordinated mandates. Publishers may rely primarily on author declarations without independent validation, while educational institutions may not systematically evaluate whether consent remains valid across subsequent editions, formats, or digital dissemination channels [1,2,4]. Pediatric clinical images require particular ethical attention because they originate in contexts characterized by vulnerability, dependency, and limited autonomy [19-22]. Individuals depicted in such materials may later encounter their own images in educational or online settings decades after image creation. These situations raise questions regarding privacy, dignity, autonomy, and long-term control over personal data. This viewpoint uses a longitudinal case involving pediatric clinical images disseminated without identifiable consent records as an analytical reference point for examining governance limitations related to persistence, reuse, accountability, provenance, and withdrawal of legacy clinical materials within contemporary digital health ecosystems. Rather than aiming to establish prevalence, the article uses this illustrative case to explore broader governance challenges associated with the long-term circulation of sensitive clinical materials across educational and digital infrastructures.

A Documented Longitudinal Case of Educational Reuse

The governance concerns discussed in this article are illustrated through a longitudinal case spanning more than 3 decades (1991-2026). The discussion draws on educational materials, institutional correspondence, publisher communications, and regulatory findings related to the dissemination of pediatric clinical images across educational and digital infrastructures. The case came to attention in 2023 when an adult medical professional recognized his own pediatric clinical images in publicly accessible educational materials. The images originated from hospital-based clinical care during childhood and adolescence in Poland and had subsequently circulated through academic textbooks, conference presentations, lecture materials, e-books, digital repositories, and online educational platforms. The identified materials included more than 40 editions and reprints of academic medical textbooks in addition to conference resources, downloadable slide collections, and publicly accessible online teaching materials. Dissemination occurred across both print and digital environments, including institutional libraries, educational syllabi, online repositories, and multimedia-sharing platforms. The images included full-body representations of children approximately 3 to 17 years of age. In several instances, full facial features and intimate anatomical regions remained visible. Some materials contained no anonymization, while others used partial masking techniques such as covering the eye region. However, these measures did not reliably prevent recognition where additional facial, bodily, or contextual characteristics remained visible [1,3,5,6]. No identifiable evidence of informed consent for publication or long-term educational reuse could be located across the analyzed materials. The affected individual reported no prior awareness of the existence or dissemination of these materials before 2023. The case is informed by archival materials, institutional correspondence, and regulatory findings, including a decision issued by the Polish Patient Rights Ombudsman on August 27, 2025, concerning continued dissemination of intimate pediatric clinical images without verifiable consent. Identifying institutional and personal details remain intentionally limited because of ongoing legal and privacy considerations. Despite regulatory intervention, the materials remained accessible in 2026 across university libraries, academic reading lists, digital repositories, and previously distributed print editions. Institutional responses were fragmented and inconsistent, and no coordinated mechanism existed to support comprehensive provenance tracking, retrospective consent verification, or coordinated withdrawal across platforms and institutions.

Fragmented Accountability Across Educational Ecosystems

This case illustrates how sensitive clinical materials may persist across educational and digital infrastructures for decades without effective mechanisms for consent traceability, provenance verification, or coordinated remediation. The continued circulation of the identified images reflected fragmentation of responsibility across health care institutions, publishers, educational systems, libraries, repositories, and digital dissemination platforms. No single actor retained both the authority and operational capacity to ensure comprehensive withdrawal once the materials entered distributed educational circulation. The case also highlights the temporal dimension of harm associated with pediatric clinical data. The affected individual became aware of the dissemination only decades later through self-recognition in publicly accessible educational materials. Importantly, identification and subsequent investigation depended heavily on the individual's professional expertise and digital literacy. Without the ability to recognize the clinical context and trace dissemination pathways, the unauthorized reuse might never have been identified. These observations suggest that detection of historical consent deficiencies may depend less on institutional safeguards than on the ability of affected individuals to recognize and investigate unauthorized reuse. They also illustrate how ethical and regulatory recognition of violations does not necessarily translate into effective remediation once materials become embedded

XSL•FO
RenderX

within decentralized educational ecosystems. Historical differences in professional norms, publication practices, archival governance, and documentation standards should also be acknowledged [1,2]. Consent procedures and institutional governance frameworks have varied substantially across jurisdictions and historical periods. Nevertheless, such variability does not eliminate the relevance of contemporary governance responsibilities when legacy materials remain publicly accessible and continue circulating within modern digital and educational environments.

AI-Enabled Secondary Reuse and Provenance Challenges

These governance limitations take on additional significance within contemporary AI ecosystems. Educational clinical materials may enter machine learning pipelines through web scraping of publicly accessible repositories, digitization of legacy textbooks, downstream aggregation, secondary dataset reuse, and incorporation into multimodal training datasets [7-11,23,24]. Publicly accessible educational repositories may therefore become indirect sources of sensitive clinical materials within AI development infrastructures. Where consent provenance remains unclear or undocumented, historical ethical deficiencies may become embedded within AI-enabled systems without effective visibility, traceability, or remediation mechanisms. Current governance initiatives, including the European Health Data Space, the European Union AI Act, and emerging international AI governance frameworks, remain primarily oriented toward prospective data collection and processing [10,11]. Comparatively less attention has been devoted to legacy educational materials already circulating across distributed institutional and technological infrastructures.

The governance issues illustrated by this case therefore extend beyond traditional publication ethics. They raise broader questions concerning downstream reuse, provenance-aware data governance, institutional accountability, and operational mechanisms for retrospective remediation within evolving AI-enabled health data ecosystems.

Toward Lifecycle Governance Frameworks

The governance limitations illustrated by this case suggest the need for lifecycle-oriented approaches capable of addressing persistence, reuse, provenance, reidentification risk, and withdrawal across evolving technological and institutional environments.

Such approaches may include (1) long-term retention of verifiable publication consent documentation; (2) separation of publication consent from treatment consent; (3) verification checkpoints before republication or digital redistribution; (4) provenance-aware dissemination and metadata systems; (5) periodic institutional review of legacy educational archives; and (6) coordinated cross-platform withdrawal procedures involving health care institutions, publishers, universities, libraries, repositories, and digital platforms.

Table 1 summarizes a proposed lifecycle governance framework for clinical images based on governance limitations illustrated through the documented case.

The framework emphasizes that governance responsibilities should not end at the point of initial publication. Instead, accountability mechanisms may need to extend across the full lifecycle of educational clinical materials, including subsequent redistribution, digitization, archival persistence, AI-enabled downstream reuse, and remediation processes.

XSL•FO
RenderX

Table 1. Lifecycle governance framework for clinical images.

Lifecycle stage	Primary responsible actors	Observed governance failures	Proposed governance mechanisms	Accountability/verification mechanism
Creation	Clinicians, hospitals, health care institutions	Absence, loss, or nonretention of documented publication consent; conflation of treatment and publication consent	Separate publication consent from treatment consent; require durable retention of consent documentation; document scope and duration of reuse permissions	Health care institutions maintain auditable consent records subject to regulatory review
Initial publication and dissemination	Authors, publishers, journal editors, educators	Reliance on author declarations without independent consent verification; inadequate review of identifiability risks	Require verifiable consent documentation before publication and republication; implement standardized consent review procedures for clinical images	Publishers and editors verify and retain consent documentation before dissemination
Secondary educational reuse	Universities, libraries, educational repositories, professional societies	Reuse across editions, courses, repositories, and presentations without renewed consent verification or provenance review	Introduce verification checkpoints before republication, digitization, or educational redistribution; require provenance-aware metadata systems	Educational institutions and repositories conduct periodic compliance and provenance review
Digital persistence and redistribution	Digital repositories, e-book platforms, search engines, online educational platforms	Continued circulation after original publication; fragmented dissemination across digital infrastructures; absence of coordinated oversight	Implement consent-linked metadata, provenance tracking, and interoperable removal protocols across platforms	Platform operators maintain traceable provenance and respond to validated removal requests
Withdrawal and remediation	Regulators, publishers, universities, repositories, platform operators	Incomplete withdrawal despite formal takedown orders; inconsistent institutional responses; persistence of distributed copies	Establish coordinated cross-platform withdrawal procedures and remediation protocols spanning print and digital systems	Regulatory authorities coordinate compliance monitoring and institutional remediation
Cross-cutting governance	All institutional actors	Fragmented accountability; unclear distribution of responsibility across systems and jurisdictions	Develop lifecycle-based governance frameworks with clearly defined institutional responsibilities, auditability, and oversight mechanisms	Shared governance agreements and periodic external review across participating institutions

Conclusions

This viewpoint illustrates how pediatric clinical images may persist across educational and digital infrastructures for decades without identifiable consent records, even after formal regulatory intervention. The documented case highlights governance limitations emerging not only at the point of initial publication, but also through the long-term absence of mechanisms capable of supporting consent traceability, provenance verification, coordinated withdrawal, and accountability across distributed systems. Current digital health governance frameworks remain primarily oriented toward prospective data collection and processing. However, the continued circulation of legacy clinical materials suggests the need for lifecycle-based governance approaches capable of addressing persistence, reuse, reidentification risk, and downstream dissemination within evolving educational and AI-enabled ecosystems [13,16-18,23-25]. This viewpoint proposes a lifecycle governance perspective for the long-term management of sensitive clinical materials, emphasizing retrospective consent verification, provenance-aware dissemination, reuse validation, and coordinated remediation mechanisms spanning health care institutions, publishers, educational systems, repositories, and digital platforms.

Acknowledgments

Generative artificial intelligence was used to support language editing, structural refinement, and clarity. All substantive content, analysis, interpretation, and conclusions are the responsibility of the author(s).

XSL•FO
RenderX

Funding

The author(s) received no financial support for the research, authorship, or publication of this article.

Conflicts of Interest

None declared.

References

1. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals. International Committee of Medical Journal Editors Recommendations. 2024. URL: https://www.icmje.org/recommendations/ [accessed 2026-05-11]
2. Guidance on consent for publication. Committee on Publication Ethics. URL: https://publicationethics.org/guidance/Guidelines [accessed 2026-05-11]
3. Roguljić M, Šimunović D, Buljan I, Žuljević MF, Turić A, Marušić A. Publishing identifiable patient photographs in the digital age: focus group study of patients, doctors, and medical students. J Med Internet Res 2025 Mar 05;27:e59970 [FREE Full text] [doi: 10.2196/59970] [Medline: 40053737]
4. Taiwo T, Obiakor B, McClung S, Shinkai K. Informed consent practices for publication of patient images in dermatology journals. JMIR Dermatol 2025 Apr 18;8:e60795 [FREE Full text] [doi: 10.2196/60795] [Medline: 40249653]
5. Harting MT, DeWees JM, Vela KM, Khirallah RT. Medical photography: current technology, evolving issues and legal perspectives. Int J Clin Pract 2015 Apr;69(4):401-409. [doi: 10.1111/ijcp.12627] [Medline: 25708063]
6. Oh C, Koh M, Tan A. Patient photographs in medical publications: ethical considerations. JAMA Dermatol 2016;152(7):759-760. [doi: 10.1001/jamadermatol.2016.1195]
7. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. PLoS Med 2018 Nov;15(11):e1002689 [FREE Full text] [doi: 10.1371/journal.pmed.1002689] [Medline: 30399149]
8. Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: mapping the debate. Big Data & Society 2016 Dec 19;3(2):205395171667967. [doi: 10.1177/2053951716679679]
9. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People-an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach (Dordr) 2018;28(4):689-707 [FREE Full text] [doi: 10.1007/s11023-018-9482-5] [Medline: 30930541]
10. Ethics and Governance of Artificial Intelligence for Health: Large Multi-modal Models. Geneva, Switzerland: World Health Organization; 2024.
11. European Health Data Space (EHDS). European Commission. URL: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en [accessed 2026-05-11]
12. Manson N, O'Neill O. Rethinking Informed Consent in Bioethics. Cambridge, United Kingdom: Cambridge University Press; 2007.
13. Kalkman S, van Delden J, Banerjee A, Tyl B, Mostert M, van Thiel G. Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence. J Med Ethics 2022 Jan;48(1):3-13 [FREE Full text] [doi: 10.1136/medethics-2019-105651] [Medline: 31719155]
14. Shabani M, Borry P. Rules for processing genetic data for research purposes in view of the new EU General Data Protection Regulation. Eur J Hum Genet 2018 Feb;26(2):149-156 [FREE Full text] [doi: 10.1038/s41431-017-0045-7] [Medline: 29187736]
15. Kaye J, Whitley EA, Lund D, Morrison M, Teare H, Melham K. Dynamic consent: a patient interface for twenty-first century research networks. Eur J Hum Genet 2015 Feb;23(2):141-146 [FREE Full text] [doi: 10.1038/ejhg.2014.71] [Medline: 24801761]
16. European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation). Off J Eur Union. 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj [accessed 2026-05-11]
17. Guidelines 05/2020 on consent under Regulation 2016/679. European Data Protection Board. 2020. URL: https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-052020-consent-under-regulation-2016679_en [accessed 2026-05-11]
18. Guidelines 01/2022 on data subject rights-right of access. European Data Protection Board. URL: https://www.edpb.europa.eu/system/files/2022-01/edpb_guidelines_012022_right-of-access_0.pdf [accessed 2026-05-11]
19. Chng SY, Tern MJW, Lee YS, Cheng LT, Kapur J, Eriksson JG, et al. Ethical considerations in AI for child health and recommendations for child-centered medical AI. NPJ Digit Med 2025 Mar 10;8(1):152 [FREE Full text] [doi: 10.1038/s41746-025-01541-1] [Medline: 40065130]
20. Lee Y, Hong S, Yon DK. Artificial intelligence in pediatric healthcare: bridging potential, clinical practice, and ethical considerations. Clin Exp Pediatr 2025 Sep;68(9):652-655 [FREE Full text] [doi: 10.3345/cep.2025.01844] [Medline: 40866147]
21. United Nations. Convention on the Rights of the Child. New York, NY: United Nations; 1989.

XSL•FO
RenderX

22. General comment No. 25 (2021) on children's rights in relation to the digital environment. In: United Nations Committee on the Rights of the Child. Geneva, Switzerland: United Nations; 2021.
23. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, III HD, et al. Datasheets for datasets. Commun. ACM 2021 Dec;64(12):86-92. [doi: 10.1145/3458723]
24. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B. Model cards for model reporting. 2019 Presented at: Conference on Fairness, Accountability, and Transparency; January 29-31, 2019; Atlanta, GA, United States p. 220-229. [doi: 10.1145/3287560.3287596]
25. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3:160018 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]

Abbreviations

AI: artificial intelligence

Edited by M Balcarras; submitted 09.Mar.2026; peer-reviewed by V Palama, M Mansoor; comments to author 22.Apr.2026; revised version received 14.May.2026; accepted 14.May.2026; published 01.Jun.2026.

Please cite as:
Bujek I
Pediatric Clinical Images Without Consent: A Governance Gap in the Long-Term Reuse of Health Data in Digital Health Ecosystems
J Med Internet Res 2026;28:e94952
URL: https://www.jmir.org/2026/1/e94952
doi:10.2196/94952
PMID:

©Iwona Bujek. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Training AI Models for Aesthetic Facial Evaluation: Focused Review and Framework to Mitigate Homogenizing Bias

Anisha R Kumar[1,2], MD; Lav R Varshney[2], PhD

[1]Division of Otolaryngology, Department of Surgery, Stony Brook University, Stony Brook, NY, United States
[2]AI Innovation Institute, Stony Brook University, Stony Brook, NY, United States

Corresponding Author:
Anisha R Kumar, MD
Division of Otolaryngology, Department of Surgery
Stony Brook University
101 Nicolls Road
Stony Brook, NY, 11790
United States
Phone: 1 (631) 444 7875
Email: arkumar@post.harvard.edu

Abstract

As artificial intelligence (AI) models become increasingly integrated into facial aesthetic surgery for attractiveness prediction and surgical outcome simulation, their potential to perpetuate bias poses clinical concerns. Current models trained on limited datasets inaccurately evaluate underrepresented populations and risk promoting aesthetic homogenization that conflicts with patient goals of ethnic feature preservation. Drawing on current literature, this paper examines bias across AI development stages in aesthetic facial evaluation. Benchmark datasets such as SCUT-FBP (South China University of Technology—Facial Beauty Prediction) and the Chicago Face Database underrepresent older adults, non-White, and ethnically diverse populations. Training methodologies lack fairness-aware techniques, and evaluation focuses on overall rather than demographic-stratified accuracy. While individual mitigation strategies exist—including balanced datasets, adversarial debiasing, and fairness metrics—no comprehensive framework integrates these approaches across the entire development lifecycle. We propose a 6-pillar framework spanning the AI development lifecycle: (1) diverse data collection with synthetic augmentation, (2) fairness-aware training techniques, (3) complementary fairness metrics with intersectional assessment, (4) explainable AI for clinical transparency, (5) stakeholder engagement, and (6) continuous monitoring. Despite the challenges of maintaining algorithmic standardization and cultural specificity, this framework provides implementation guidance for AI developers, clinicians, and institutions, with principles applicable beyond aesthetic surgery to broader facial analysis applications.

(J Med Internet Res 2026;28:e95452) doi:10.2196/95452

KEYWORDS

aesthetics; artificial intelligence; fairness; framework; governance; machine learning; medical informatics; model development; plastic surgery; surgery

Introduction

Artificial intelligence (AI) is increasingly integrated into preoperative planning and outcome simulation in facial aesthetic surgery [1,2], typically based on modern machine learning (ML) techniques. Current applications include prediction of attractiveness, simulation of surgical outcomes, and patient assessments [3], offering the potential for objective, standardized aesthetic evaluation [4]. Facial plastic surgeons have preliminarily assessed how established AI-based websites compare to human scoring of facial attractiveness [5], yet standardized and validated AI models for facial aesthetic evaluation have not been established. Before widespread clinical integration, surgeons must understand how these models are trained to ensure accurate, culturally diverse evaluations and avoid perpetuating bias.

The technical challenge of AI model training is compounded by the fact that beauty is a cultural construct rather than a universal fact [6]. Recent large-scale cross-cultural research using geometric morphometrics and Bayesian analysis of 1550 faces from 10 global populations reveals this complexity: distinctiveness (deviation from average facial proportions) negatively affects attractiveness perception universally, and femininity positively influences attractiveness assessments of

XSL•FO
RenderX

female faces across all studied populations [7]. However, 2 traditionally emphasized features showed no robust effects: facial symmetry had no significant association with attractiveness ratings, and masculinity did not consistently influence attractiveness judgments of male faces [7]. These findings challenge conventional assumptions about which features are genuinely universal versus culturally variable.

These universal principles interact with culture-specific preferences in complex ways. Skin coloration demonstrates culturally modulated aesthetic judgments: lighter skin tones associate with attractiveness among Chinese observers judging own-ethnicity faces, whereas European observers associate warmer yellow skin tones with attractiveness in Chinese faces [8]. Cross-cultural studies comparing Japanese and American raters reveal that while overall attractiveness ratings correlate across cultures, specific features driving these judgments differ: Japanese raters emphasize raised eyebrows in attractive male faces and smaller mouths in attractive female faces more than American raters [9]. Similarly, considerable cross-cultural agreement exists regarding Vietnamese facial attractiveness, yet Czech European raters associate attractiveness with averageness significantly more than Vietnamese raters [10]. These findings illustrate that while certain structural features (averageness and femininity) demonstrate universal appeal, the relative importance and specific manifestations of these features vary across cultural contexts.

Historically, plastic surgery has relied on Western aesthetic standards [11], but there is growing patient diversity and a shift toward preservation of features that convey ethnic identity [12,13]. An inherent tension, therefore, exists in training AI models: they must account for genuinely universal principles (such as averageness and femininity in female faces) while also recognizing culture-specific preferences and avoiding the imposition of Eurocentric standards on features that demonstrate regional variation [4].

Furthermore, AI models trained on biased datasets may perpetuate narrow beauty ideals and inaccurate representations of patient populations who vary across race, ethnicity, nationality, language, socioeconomic background, gender, and age [14,15], dimensions that interact to produce compounded disadvantage at their intersections, creating a risk that algorithmic recommendations lead to the elimination of distinctive ethnic characteristics and aesthetic homogenization. Culturally responsive AI training frameworks are therefore needed [4]; without them, a gap will persist between the technical capabilities of AI and its ethical implementation in comprehensive patient care [16].

In this paper, we aim to: (1) review AI training methodologies for aesthetic evaluation, (2) examine sources of bias in training, (3) evaluate current practices for mitigating bias, and (4) propose a framework for artificial intelligence/machine learning (AI/ML) training with recommendations for clinical implementation.

Review of Current Training Methods

Deep learning is a method within AI/ML that uses multilayer neural networks to process large amounts of data and extract complex patterns and features. Within this approach, convolutional neural networks (CNNs) are common for image recognition and processing [16,17]. Such ML models are trained using supervised learning based on human-rated attractiveness scores [18,19], geometric features including symmetry and the Golden Ratio [20], the “rule of thirds” for frontal view analysis [17,20], and extracted facial proportions such as nasofrontal angle, nasolabial angle, and glabella-to-chin angle [17].

Model performance is evaluated using regression tasks, including the Pearson correlation coefficient, which measures the linear relationship between AI-predicted attractiveness scores and human ratings [21,22], and mean absolute error, which measures the average magnitude of prediction errors. These metrics validate how well the AI model replicates human aesthetic judgment. However, recent work in health-related AI has emphasized that evaluation should not rely on task performance metrics alone but also incorporate risk-oriented and context-sensitive assessment frameworks [23].

Beyond attractiveness prediction alone, AI models can simultaneously learn to perform related tasks, including age evaluation, gender identification, ethnicity and race recognition, and facial expression detection [24]. This multitask learning (MTL) approach improves the model’s evaluative capacity across all tasks and develops a more comprehensive understanding of facial features that determine attractiveness. MTL also addresses dataset size limitations (SCUT-FBP [South China University of Technology—Facial Beauty Prediction]: 500 faces; SCUT-FBP5500: 5500 faces) by incorporating auxiliary tasks that provide additional training signals, leveraging natural connections between attractiveness and related attributes [24].

However, recent architectural innovations have moved beyond traditional CNNs to address dataset limitations. Hybrid approaches combining vision transformers and state-space models (such as Mamba) leverage complementary feature extraction capabilities [25]. Vision transformers excel at capturing global facial structure and symmetry through attention mechanisms that process the entire face holistically, while state-space models efficiently model fine-grained local features such as skin quality and texture with linear computational complexity [25].

These architectures use self-supervised pretraining on large, diverse image datasets, followed by task-specific fine-tuning on smaller facial beauty datasets [25]. This transfer learning approach addresses the fundamental challenge that aesthetic datasets, such as SCUT-FBP5500 with 5500 faces, contain orders of magnitude fewer images than typically required for deep learning, while achieving state-of-the-art performance metrics [25].

Several benchmark datasets are currently used to train AI models for attractiveness evaluation. The SCUT-FBP database contains 500 Asian female faces and achieved a Pearson correlation coefficient of 0.8187 between CNN predictions and human ratings [21]. The expanded SCUT-FBP5500 dataset includes 5500 faces with more diverse demographics and achieved correlations of 0.87-0.90 with human evaluators [26]. The Chicago Face Database contains 597 photographs of White and

XSL•FO
RenderX

Black male and female individuals aged 17-65 years [27]. [21,26-32].
Characteristics of each major dataset are detailed in Table 1

XSL•FO
RenderX

Table 1. Comparison of major facial aesthetic datasets. Characteristics of 8 datasets commonly used to train and evaluate artificial intelligence (AI) models for facial attractiveness prediction, organized by publication year. Datasets were identified from the systematic literature review in this paper. Sample size reflects the most current reported demographic information for each dataset.

Database	Sample size (N); age range (years)	Demographics (race, ethnicity, and gender)	Standardized images	Primary use	Key limitations	Validation metrics
SCUT-FBP (2015) [21]	N=500; not reported	• 100% Asian • 100% Female	Yes (controlled laboratory conditions)	Facial attractiveness prediction benchmark	• Small sample • Single demographic (Asian females) • Limited generalizability	Pearson r=0.82
SCUT-FBP5500 (2018) [26]	N=5500; 15-60	• 73% Asian • 27% Caucasian • 50% Male • 50% Female	No (aggregated from multiple sources)	Multiparadigm facial attractiveness prediction	• Limited ethnic diversity (Asian/Caucasian only) • Nonstandardized images • Variable quality	Pearson r=0.87-0.90
Chicago Face Database (2015) [27]	N=597; 17-65	• White • Black • Asian • Latino • Male • Female • Percentages unknown	Yes (laboratory photography, standardized conditions)	Research and AI model training across demographics	• Small sample • May not capture full diversity	Attractiveness ratings by independent raters with interrater reliability; includes mean scores and SDs for multiple attributes
BLINQ (dating site dataset) (not reported) [30]	N>13,000; not reported	• Not specified (unknown demographics)	No (collected from dating websites)	CNN[a] for attractiveness training	• Nonstandardized • Unknown demographics • Selection bias • Rating bias	Not reported
Labeled Faces in the Wild (2007) [28]	N=5749; not reported	• 77.5% Male • 83.5% White • Remaining unknown	No (captured from online articles/press)	Facial recognition[b]	• Severe demographic bias • Not designed for aesthetics • Nonstandardized	Not applicable (facial recognition dataset)
MEBeauty (2022) [29]	N=2550; not reported	• 38% White • 12% Black • 14% Asian • 12% Indian • 12% Middle Eastern • 12% Hispanic • 51% Female • 49% Male	No (in-the-wild collection)	Facial beauty prediction with ethnic diversity	• Small sample • Nonstandardized • Limited validation • Minimal literature usage	Pearson r with human ratings; transfer learning validation
FairFace (2021) [31]	N=108,501; 0-70+	• 50% Female • 50% Male • 14.3% across major racial groups	No (in-the-wild collection)	Facial recognition[b]	• Not designed for aesthetics • Nonstandardized • No attractiveness ratings	Cross-dataset accuracy; fairness evaluation across demographics
Diversity in Faces (2019) [32]	N=1,000,000; not reported	Not categorized by race/ethnicity—uses objective facial coding schemes	No (in-the-wild collection)	Facial recognition[b]	• No race/ethnicity labels • No attractiveness ratings	Not reported

[a]CNN: convolutional neural network.

[b]Labeled Faces in the Wild, FairFace, and Diversity in Faces are facial recognition datasets not designed for attractiveness evaluation. They are included because they are widely used as pretraining resources, bias benchmarks, and demographic diversity references in the facial attractiveness AI literature.

XSL•FO
RenderX

As datasets increase in size, standardization becomes increasingly challenging, with many datasets containing mixed lighting conditions, facial expressions, and makeup [30]. A study comparing CNN models trained on the BLINQ dating site database (containing over 13,000 nonstandardized images) vs models trained on BLINQ and then fine-tuned on the standardized Chicago Face Database [27] demonstrated that the 2-step training approach resulted in less variability in attractiveness scores, as facial expressions were shown to confound assessments [30]. These findings underscore the need for purpose-built training systems that can also account for facial expressions.

A critical challenge in current AI/ML training is the feedback loop problem known as "performativity," where a model's predictions influence future data distribution [33]. Models trained on specific beauty standards inevitably amplify those standards over time [15]. Bias can be introduced at multiple stages of the AI/ML development lifecycle [34], including targeting bias (defining beauty standards), data acquisition bias (using homogeneous datasets), modeling bias, validation and evaluation bias, and deployment and monitoring bias. Without intentional design for objective assessment, these biases lead to inaccurate and culturally inappropriate aesthetic evaluations.

Review of Bias Manifestations

Demographic bias is pervasive in facial recognition and aesthetic AI models, resulting in unreliable predictions for underrepresented groups. In a comprehensive evaluation of facial recognition vendor tests, the National Institute of Standards and Technology (NIST) found significant performance variability across different demographic groups and advised users to be aware of these disparities when selecting algorithms [35]. Facial analysis benchmarks demonstrate particularly unreliable predictions for underrepresented demographics, such as females with higher Fitzpatrick skin types [14]. Beyond racial and gender bias, certain age groups, particularly individuals aged older than 60 years, are underrepresented in training data, further compromising accuracy in aesthetic evaluation for these populations [36].

As noted, Eurocentric beauty standards serve as the default framework in AI models [4]. Features flagged as "flaws" may represent valued ethnic traits: broader nasal bridges characteristic of Arab populations are marked for "correction," while fuller lips common in North African ethnicities are classified as "disproportionate" [4]. Numerous studies have demonstrated that the golden ratio and other neoclassical canons inadequately capture attractiveness across diverse populations [11]. Consequently, AI models built according to singular beauty standards risk inadvertently recommending the westernization of ethnic features without consideration of cultural appropriateness [4].

Regional and cultural aesthetic preferences vary significantly. Patients of Middle Eastern and North African descent often seek nasal tip refinement while preserving other ethnic characteristics such as dorsal height [4,37]. East Asian patients typically prioritize augmentation procedures over reduction [13]. Despite significant cross-cultural variation in motivations, patient expectations frequently include conscious avoidance of "westernized" appearance [37], and high satisfaction correlates strongly with preservation of ethnic features [38].

Review of Existing Bias Mitigation Approaches

Researchers have developed bias mitigation strategies targeting different stages of AI model development. At the dataset level, initiatives like FairFace have created balanced demographic representations across race, gender, and age categories to address training data imbalances [31]. Synthetic data generation using generative adversarial networks has been proposed to augment underrepresented demographic categories.

Multiple algorithmic interventions have been developed to detect and reduce bias during model training. Adversarial debiasing methods have demonstrated improved fairness outcomes by mitigating bias acquired during data collection [39]. Posttraining corrections, such as centroid fairness loss, enable bias measurement and performance alignment across demographic groups without requiring complete model retraining [40]. Skewness-aware reinforcement learning approaches address data distribution imbalances [41], while techniques like debiasing variational autoencoders adjust sampling probabilities for underrepresented categories [42]. Meta-learning approaches enable models to adapt to regional aesthetic standards with limited culture-specific training data [43].

Bias detection has also advanced through evaluation methodologies. Researchers have documented systematic performance disparities across demographic groups in facial analysis systems. Studies examining intersectional accuracy gaps have revealed that these disparities stem from complex structural factors: for example, poor performance on dark-skinned females in gender classification results not from skin tone itself, but from differences in lip, eye, and cheek structure across ethnicities [44].

Gaps in Current Approaches

Despite these advances in bias detection and mitigation, critical gaps remain in their application to aesthetic facial evaluation. First, current approaches address bias mitigation at isolated stages rather than across the complete AI development lifecycle. Techniques such as adversarial debiasing [39], centroid fairness loss [40], skewness-aware reinforcement learning [41], and debiasing variational autoencoders [42] have been validated in isolation primarily for facial recognition tasks—where the objective is identity verification—rather than aesthetic evaluation, where subjective cultural beauty standards introduce fundamentally different fairness challenges.

Recent empirical evidence reveals fundamental inadequacies in current mitigation strategies. Dataset diversification efforts like FairFace [31] balance demographic representation yet fail to address annotation bias, the systematic application of culturally-specific aesthetic judgments by raters during labeling [29]. A facial beauty prediction model trained on the multiethnic MEBeauty dataset exhibited significant prediction disparities

XSL•FO
RenderX

across ethnic groups ($P<.001$) even when evaluated on balanced data, with only 4.8%-9.5% of intergroup comparisons satisfying distributional parity criteria [45]. More concerning, models demonstrated exacerbated bias on balanced demographic datasets compared to training performance, indicating that current approaches may amplify rather than mitigate societal biases when deployed to real-world populations [45].

Second, no comprehensive framework integrates technical solutions with essential nontechnical components across the AI development lifecycle. Existing approaches lack systematic stakeholder engagement, explainability requirements, and governance structures. While participatory design approaches exist for health care AI broadly [46], aesthetic surgery applications do not systematically involve patients, clinicians, and cultural consultants in model development. Postdeployment monitoring, essential for detecting fairness drift and performance degradation across demographic subgroups [4,35], remains absent from aesthetic AI implementations, with validation protocols testing for bias before deployment rarely incorporated into development pipelines [45].

Third, the field lacks consensus on which fairness metrics are most appropriate for aesthetic contexts. While demographic parity, equalized odds, and equal opportunity are well-defined [47], their application to aesthetic evaluation poses unique challenges: achieving demographic parity may conflict with honoring culturally-specific beauty standards [4], and determining appropriate trade-offs requires stakeholder input that current approaches do not systematically incorporate. These gaps necessitate an integrated framework specifically designed for the unique challenges of bias mitigation in aesthetic facial evaluation.

Proposed Framework for AI/ML Training in Aesthetic Facial Evaluation

Overview

To address these gaps, we propose the following framework for the training of AI/ML models in aesthetic evaluation that consists of 6 pillars: data collection and curation, model training methodologies, fairness metrics and evaluation, explainability and transparency, stakeholder engagement, and governance and monitoring.

This framework addresses AI/ML systems used across several distinct aesthetic evaluation tasks, which differ substantially in their fairness stakes and acceptable error thresholds. Attractiveness scoring assigns a rating or ranking to a face and is used primarily for research benchmarking. Preoperative planning uses AI to assess anatomic features and inform surgical approach, where systematic undervaluation of ethnic features could directly influence clinical decision-making. Outcome simulation generates predicted postoperative appearance, where bias has the additional potential to alter rendered ethnicity rather than merely undervalue it, a qualitatively distinct harm addressed further in the framework. Patient counseling involves AI-assisted communication of options, where biased framing may subtly steer patient choices. Fairness requirements and error thresholds should be calibrated to the stakes of the specific use case; the more downstream the application and the more directly it affects patient choice and surgical planning, the more stringent the requirements.

Throughout this framework, recommendations are categorized by their evidence base (Table 2). Practices marked as “established” draw on studies conducted in aesthetic facial evaluation contexts. Practices marked as “adapted” are supported by evidence from adjacent domains, primarily facial recognition, general medical AI, or computer vision, and have been translated to aesthetic evaluation by analogy; these require validation in aesthetic-specific contexts before adoption as standard practice. Practices marked as “proposed” represent conceptual recommendations without empirical validation in any closely related domain and should be treated as research directions.

XSL•FO
RenderX

Table 2. Evidence classification for framework recommendations. Summary of bias mitigation practices included in the proposed framework, organized by strength of evidence. Established evidence base indicates practices supported by studies conducted in aesthetic facial evaluation contexts. Adapted evidence base indicates practices supported by evidence from adjacent domains (facial recognition, general medical AI, or computer vision) that have been translated to aesthetic evaluation by analogy and require validation in aesthetic-specific contexts before adoption as standard practice. Proposed evidence base indicates conceptual recommendations without empirical validation in any closely related domain, to be treated as research directions.

Bias source	Pipeline stage	Mitigation strategy	Evidence base	Residual gap
Targeting bias: narrow beauty standard definition	Data collection	• Pillar 1: ≥7 ethnic categories • Mixed-ancestry probabilistic labeling • Continuous morphometric representation	Established	• Discrete ethnic categories essentialize group-level patterns • Intragroup variation (eg, nationality, socioeconomic background) is rarely captured
Annotation bias: culturally skewed rater judgments	Data collection	• Pillar 1: diverse rater recruitment • Structured training and calibration • Tiered disagreement adjudication • Ongoing score audits	Adapted	• No validated rater calibration protocol exists for aesthetic evaluation • Cultural feature weighting may persist despite diverse panels
Generative adversarial network amplification bias: synthetic augmentation	Data collection	• Pillar 1: quality control gate for synthetic images—fairness audit, feature distribution check, and human review	Adapted	• Mode collapse and feature exaggeration are documented in generative systems • Quality control criteria not validated for aesthetic contexts
Modeling bias: fairness-unaware training	Model training	• Pillar 2: adversarial debiasing • Centroid fairness loss • Skewness-aware reinforcement learning • Debiasing variational autoencoder • Multitask learning	Adapted	• All techniques validated in facial recognition or general computer vision, not aesthetic evaluation • Several rest on preprint evidence • Combined validation absent
Domain shift bias: train/deploy distribution mismatch	Model training → deployment	• Pillar 2: hybrid pretraining on standardized images • Fine-tuning on clinical images • Domain generalization evaluation prerelease	Proposed	• No validated hybrid protocol for aesthetic AI • Clinical image variation not systematically characterized
Evaluation bias: aggregate metrics obscure subgroup disparities	Evaluation	• Pillar 3: layered fairness metrics with prioritization hierarchy • Intersectional assessment • Bayesian hierarchical modeling for rare subgroups	Adapted	• Thresholds are proposed benchmarks without empirical derivation • Metrics can conflict • Intersectional sample sizes often insufficient
Explainability gap: black-box outputs in a cultural context	Evaluation → deployment	• Pillar 4: Grad-CAM[a], LIME[b], SHAP[c] with required human expert review • Geometric/physics-based models as a longer-term goal	Adapted	• Explainable AI tools cannot explain why features are culturally valued • No method validated for cultural appropriateness verification in aesthetic AI
Human–AI decision bias: clinician interpretation and override	Deployment	• Pillar 5 + 6: documentation and audit of AI recommendation override rates by patient demographic • Clinician training on implicit bias	Proposed	• No empirical data on differential override in aesthetic AI • Audit infrastructure absent • Accountability for remediation undefined
Deployment bias: commercial systems without governance	Deployment	• Pillar 6: disclosure-based accountability for commercial developers • FDA SaMD[d] framework alignment	Proposed	• No enforcement mechanism for commercial tools • Patient and clinician verification of compliance is currently impossible

XSL•FO
RenderX

Bias source	Pipeline stage	Mitigation strategy	Evidence base	Residual gap
Drift bias: postdeployment fairness degradation	Monitoring	• Pillar 6: tiered monitoring—continuous process control, quarterly review, annual audit, drift-triggered escalation • Designated AI clinical lead	Adapted	• Drift thresholds not empirically derived for aesthetic AI • Continuous monitoring may not be feasible for community practices

[a]Grad-CAM: gradient-weighted class activation mapping.

[b]LIME: local interpretable model-agnostic explanations.

[c]SHAP: Shapley additive explanations.

[d]FDA SaMD: Food and Drug Administration's Software as a Medical Device.

Data Collection and Curation

As a pragmatic baseline informed by existing benchmark datasets, training data should include a balanced representation of at least 7 racial and ethnic categories [31] (White, Black, East Asian, Southeast Asian, Middle Eastern, Latino, and Indian), with additional stratification by gender, age, nationality, and socioeconomic background. This should include multiregional data collection with noted region-specific aesthetic preferences [4,29]. This scheme is explicitly a minimum starting point, not a definitive classification; implementations should adopt more granular schemes as data availability permits. Individuals of mixed ancestry, a rapidly growing population, should be accommodated through multilabel or probabilistic ancestry representation rather than forced assignment to a single category. As the field matures, continuous morphometric representations of facial ancestry, such as principal components of facial geometry derived from diverse reference populations, offer a more biologically-grounded alternative to discrete ethnic labels and should be pursued to reduce the risk of essentialization.

While standardized photographs in a constrained environment are necessary to reduce training inaccuracies from facial expression confounds [30], they introduce a domain-shift risk at deployment: real-world clinical images routinely involve variation in lighting, angle, makeup, and expression that differs systematically from controlled training conditions. To address this, we recommend a hybrid protocol: initial pretraining on standardized images to establish controlled baseline representations, followed by fine-tuning on a curated set of clinically realistic images incorporating documented augmentation strategies—including geometric transformations such as rotation and translation to simulate angle variation, and color space augmentations to simulate lighting variation—to reduce the gap between training and deployment distributions [48]. Explicit evaluation of domain generalization, measuring performance and fairness metric stability across both standardized and nonstandardized image sets, should be required before clinical release.

Even with standardized imaging protocols, acquiring sufficient photographs across all demographic categories remains challenging. Synthetic data generation can bridge these gaps while also protecting privacy, with pretraining gap analysis using established benchmark datasets such as FairFace [31] or Diversity in Faces [32] used to identify demographic imbalances or biases; these have been established in facial attribute classification, though their applicability to aesthetic evaluation requires confirmation. Generative models are themselves trained on real-world data containing existing biases and are susceptible to mode collapse, where the model produces a narrow range of outputs disproportionately representing dominant features, and to hallucination of exaggerated demographic characteristics when conditioned on ethnic labels. These failure modes risk reintroducing the essentialized representations that the framework is designed to prevent. Accordingly, synthetic images must not enter the training set without explicit quality-control steps: adversarial fairness auditing of synthetic outputs to detect feature exaggeration; statistical comparison of synthetic image feature distributions against reference population benchmarks; and human review by the diverse rater panels proposed elsewhere in this framework. Only images passing all 3 criteria should be incorporated.

Furthermore, photograph raters should be recruited from diverse cultural backgrounds to avoid annotation bias [29]. Annotation bias occurs when raters systematically apply their own cultural aesthetic standards to evaluate faces from different backgrounds, but the patterns of bias are complex and do not reduce to simple in-group favoritism. Research demonstrates that rater ethnicity influences which facial features are emphasized—for instance, Chinese observers associate lighter skin tones with attractiveness in own-ethnicity faces, whereas European observers prefer warmer tones in Chinese faces [8], and Japanese raters emphasize raised eyebrows in attractive male faces and smaller mouths in attractive female faces more than American raters [9]. However, cross-cultural studies comparing attractiveness ratings across European, East Asian, and African faces found no strong own-race preference in overall attractiveness judgments [49], indicating that annotation bias operates through subtle feature weighting rather than categorical group favoritism. For example, raters trained predominantly in Western aesthetic ideals might systematically underweight features like broader nasal bridges or fuller lips that are attractive within specific cultural contexts, not because of explicit racial preference but because their cultural training emphasizes different facial proportions. Recruiting ethnically diverse rater panels is therefore essential to ensure balanced representation of aesthetic preferences rather than assuming any single demographic composition will eliminate bias.

Model Training Methodologies

MTL is central to bias mitigation in AI/ML model training. Models should be simultaneously trained on age, gender, ethnicity, facial expression, and attractiveness ratings to develop

XSL•FO
RenderX

a comprehensive and nuanced assessment of faces [36] where feasible. Rater cultural background, region of upbringing, and socioeconomic status should be recorded as covariates to enable analysis of how these dimensions influence annotation.

Several fairness-aware techniques should be implemented during model training, adapted from the facial recognition literature but not yet validated in aesthetic evaluation. Adversarial learning methods [39] should be applied during the training phase to mitigate bias acquired during data collection. Posttraining, centroid fairness loss [40] enables bias measurement and performance alignment across demographic groups without requiring complex model retraining, a significant practical advantage. Skewness-aware reinforcement learning [41] should be used to recognize and adjust imbalances in data distribution or model performance across demographics. Finally, debiasing variational autoencoder [42] can adjust sampling probabilities for underrepresented categories, balancing the effective training data to enable more equitable performance across patient populations.

Beyond these fairness-aware training techniques, AI/ML models can be designed to adapt to regional aesthetic standards for facial evaluation. A meta-learning approach, supported by preliminary evidence from the beauty prediction literature, enables models to “learn how to learn”: models trained on learning tasks from a range of cultures develop the ability to adapt to other cultural preferences more readily [43]. This methodology can account for the subjective nature of beauty perception across cultures and allow customization based on patient population.

While meta-learning enables adaptation to regional standards, the hierarchical structure of aesthetic preferences, with both universal and ethnicity-specific components, suggests opportunities for more sophisticated architectural approaches. Hierarchical Bayesian models could naturally encode this structure through multilevel parameter sharing, where population-level priors capture universal features while group-specific parameters account for cultural variation. Alternatively, the partial invariance framework extends invariant risk minimization by learning features that are invariant within partitions of training environments rather than globally invariant across all environments [50]. Such approaches may encode averageness-related objectives within cultural partitions rather than across the full training distribution, preserving within-group distinctiveness while limiting cross-group homogenization. Universal structural features would be encoded at the population level only when evidence supports genuine cross-cultural validity, while regionally variable features remain governed by culture-specific parameters. Efficient multigroup equivariant techniques that address intersectional fairness across combinations of protected attributes, such as ethnicity and gender, may offer additional methodological directions [51]. However, these techniques have not yet been validated for subjective aesthetic judgments, and their applicability to facial evaluation remains an open empirical question.

Until such validation is established, these advanced architectural approaches should be considered active research directions rather than recommended clinical components. Several debiasing techniques in this section also rest on preprint or single-study evidence whose reproducibility has not been independently established; responsible clinical adoption requires, at a minimum, prospective studies demonstrating improvement in prespecified fairness metrics across multiple independent ethnic groups, independent reproducibility on held-out datasets, and direct comparison against uncontrolled baseline systems under clinically realistic conditions.

Fairness Metrics and Evaluation

A comprehensive fairness evaluation system integrates group fairness metrics, performance standards, and intersectional assessment. This layered approach prevents the common problem of achieving fairness on average while still having significant disparities in specific subpopulations, which is particularly critical in aesthetic facial evaluation, where cultural considerations vary significantly across intersectional identities.

Fairness metrics are chosen based on the clinical application to facial evaluation. Group fairness metrics measure and reduce bias to ensure that models evaluate different demographic groups equitably [47]. Demographic parity catches outcome bias by ensuring that the models’ aesthetic ratings are independent of ethnicity and equal across different demographic groups. Since the clinical utility of these models requires accuracy as well, equalized odds catches accuracy bias by ensuring that the models perform equally well across all demographic groups, with false positives and false negatives occurring equally. Additionally, the evaluation system should include equal opportunity metrics as an option for clinicians when identifying positive outcomes is the priority and false positives are not an issue. In an aesthetic evaluation scenario, equal opportunity focuses on not missing attractive features while being more flexible about possible overestimation. However, different fairness metrics can oppose each other, which makes it axiomatically impossible to align all of them simultaneously.

We propose the following benchmarks as starting points for community debate and empirical refinement, not as validated thresholds: cultural concordance scores of at least 80%, as reviewed by regional expert review panels [4], feature recognition accuracy of at least 95% [4], and demographic parity in prediction accuracy with no more than 5% variance across groups. These figures represent reasonable aspirational targets informed by analogous fairness benchmarks in other clinical AI applications but require prospective validation in aesthetic evaluation contexts before adoption as standards.

Because fairness metrics will conflict in practice, a decision hierarchy is necessary. Consider the following scenario: a model achieves demographic parity (equal average attractiveness ratings across ethnic groups) but does so by consistently overpredicting attractiveness for underrepresented groups while underperforming on fine-grained feature recognition for those same groups. Demographic parity is satisfied; equalized odds are not. In this scenario, we recommend prioritizing equalized odds because accuracy parity across groups is a prerequisite to the clinical utility of the tool. A model that fails to accurately recognize features in specific ethnic groups cannot serve those patients equitably, regardless of average score distributions. Equal opportunity metrics should then be applied as a secondary

XSL•FO
RenderX

check, specifically where the clinical priority is avoiding false negatives, for example, ensuring that attractive features in underrepresented populations are not systematically missed.

This hierarchy assumes deployment in a pluralistic patient population. In settings where the clinical population is demographically homogeneous, the relevant fairness question shifts: the priority becomes within-population accuracy and avoidance of intragroup bias, rather than cross-group parity. In such contexts, fine-tuning on locally representative data may be both technically appropriate and ethically indicated, provided that the resulting model is transparently scoped to its intended deployment population and not generalized beyond it. A model developed and validated for a specific national or regional context, for example, a system trained primarily on Korean patients for deployment in South Korea, should be evaluated against locally derived aesthetic norms and demographic distributions, and its scope of applicability documented accordingly.

Since the above-mentioned techniques evaluate for biases for individual criteria such as race or age, an intersectional assessment should be introduced into model training to evaluate overlapping biases that may compound. In aesthetic facial evaluation, intersectional bias due to the aggregation of multiple social identities, such as race, ethnicity, nationality, socioeconomic background, gender, and age, can incorrectly influence outcomes and may perpetuate stereotypes [14]. An intersectional assessment provides the framework to apply metrics and standards across complex, overlapping demographic categories such as Black women or older Asian men.

Since individuals hold multiple overlapping identities simultaneously, each with associated social norms and expectations, aesthetic preferences cannot be understood by analyzing demographic categories in isolation, a theoretical foundation reinforcing that intersectional assessment addresses the fundamental mechanism through which identity influences perception [52].

Practically, however, current dataset sizes preclude stable estimates across all intersectional combinations, necessitating a prioritization hierarchy. High-risk intersections with documented performance disparities—such as older women with higher Fitzpatrick skin types—should be evaluated as the primary tier. Compositional approaches offer a promising avenue for addressing certain intersectionality challenges more tractably: multigroup equivariant network designs that use product groups can provide fairness guarantees across intersectional demographic combinations with computational complexity proportional to the sum rather than the product of group sizes, as demonstrated in natural language generation debiasing tasks [51]. Second, Bayesian hierarchical modeling of rare subgroups enables partial pooling of statistical strength from related intersections, providing more stable estimates for low-frequency cells; this approach ties naturally to the hierarchical architectures recommended in the Model Training Methodologies section. Third, multitask regularization can share statistical strength across related intersectional categories during training. Intersectional combinations not covered by the primary tier should be explicitly designated for future work rather than omitted without acknowledgment.

Explainability and Transparency

Explainability and transparency are prerequisites for trustworthy AI/ML systems in clinical use. Three different techniques adapted from general medical AI would be beneficial for training models used in aesthetic facial evaluation: gradient-weighted class activation mapping (Grad-CAM), local interpretable model-agnostic explanations (LIME), and Shapley additive explanations (SHAP) [53]. Grad-CAM provides visual and spatial explanations through heatmaps that highlight which facial regions contribute most to a model's aesthetic predictions. Clinicians can identify which anatomic feature the model prioritizes and can also verify that the model emphasizes culturally appropriate features rather than defaulting to Eurocentric beauty standards. LIME provides case-specific explanations of a model's individual predictions by showing which features influenced that specific assessment. This technique is model-agnostic and therefore flexible and versatile, meaning it can be applied to any AI/ML architecture—whether convolutional neural networks, transformer models, or future technologies—making it adaptable as the field evolves. SHAP provides information about how each evaluated feature contributes to the specific output of a model. It can explain individual predictions and provide a global overview of which features are most important in a dataset. The practical value of SHAP-based interpretation has also been demonstrated in surgical predictive AI, where feature-level explanation was used to identify key perioperative risk factors and improve transparency of model behavior for clinical decision-making [54]. While these methods continue to evolve with ongoing algorithmic refinements, these tools are useful but not sufficient for verifying cultural appropriateness in aesthetic evaluation. They identify which facial regions or features influence a model's output, but do not explain why those features are aesthetically valued within a specific cultural context, which is the central interpretive question this framework is designed to address. Human expert review by culturally knowledgeable clinicians is, therefore, a required complement to explainable AI (XAI) output, not an optional one.

Intrinsically interpretable approaches, which incorporate domain knowledge directly into model architecture rather than applying post hoc explanation methods, represent a more direct path toward clinical-grade cultural interpretability. For facial aesthetic evaluation, this includes explicitly encoding geometric relationships (such as facial proportions, angles, and distances) as structured features within the model, and physics-based models that incorporate established anatomical principles and morphometric relationships.

Transparency standards should ensure that stakeholders have access to essential information about model development and validation. Transparency requirements for AI/ML training models should include the documentation of training data demographics and model architecture, interpretable explanations for aesthetic assessments, and disclosure of identified biases and performance disparities across demographics.

XSL•FO
RenderX

For generative outcome simulation systems specifically, XAI tools must address an additional interpretive requirement: clinicians should be able to verify that simulated postoperative appearances reflect only the intended surgical modifications and do not introduce ethnically incongruent features as artifacts of model bias. This requires comparison of presimulation and postsimulation facial geometry at the feature level, which current Grad-CAM and LIME implementations are not designed to provide; bespoke evaluation protocols for generative systems are needed.

Stakeholder Engagement and Participatory Design

The development of AI models should involve AI developers, patients, clinicians, cultural consultants, and ethicists as equal partners from the conceptualization stage through implementation and postdeployment monitoring [46]. This cocreation process ensures that diverse perspectives shape decision-making at every stage, from selecting training datasets to interpreting model outputs to refining algorithms based on real-world performance. Workshops and iterative feedback sessions should be conducted throughout the development lifecycle to gather input on critical questions such as which facial features should be prioritized for analysis, how to define culturally appropriate aesthetic outcomes, and whether model recommendations align with patient values and clinical judgment. Participant recruitment should deliberately include patients who identify across multiple marginalized dimensions simultaneously, such as older immigrant women from non-Western countries, as their aesthetic priorities and experiences of algorithmic bias are likely to differ from those captured by single-axis demographic sampling. It is vital to recognize and address potential power dynamics to ensure that underrepresented stakeholders' voices are valued equally alongside technical experts and established institutions.

Regional expert review panels convened to evaluate cultural concordance must be constituted with explicit accountability requirements. Panel composition should include representatives from diverse geographic regions within each cultural community, not only urban or elite centers, and should reflect variation in socioeconomic background, age, and gender. Selection criteria should be documented and publicly disclosed. Formal mechanisms for recording minority and dissenting views are required; panel reports should distinguish consensus from majority positions and preserve dissenting opinions for review. Periodic panel rotation and external audit of panel composition guard against the entrenchment of a single institutionalized aesthetic perspective. These panel requirements must be operationalized through rigorous rater training and calibration protocols.

Rater training and calibration protocols are essential to annotation quality. Before scoring, raters should complete structured training that includes: an orientation to the study's cultural equity goals; exposure to diverse face exemplars across all demographic categories to be rated; and explicit instruction to evaluate attractiveness according to within-group cultural standards rather than a universal ideal. Calibration should be conducted using a standardized set of anchor images—rated in advance by a culturally matched expert panel—against which individual rater scores are benchmarked. Raters whose scores diverge systematically from calibration anchors by more than a prespecified threshold (for example, mean absolute deviation greater than 1.0 on a 10-point scale) should receive additional training before contributing to the primary dataset. For images where rater scores span more than 3 points on a 10-point scale, the image should be flagged for adjudication by a culturally matched expert reviewer rather than resolved by averaging. Averaged scores obscure genuine aesthetic disagreement that may itself be informative about cultural variation. Ongoing audit of rater score distributions by demographic subgroup should be conducted throughout data collection to detect systematic drift in individual rater calibration.

A human-centered design framework positions clinicians and patients as essential collaborators while respecting their cultural contexts and prior experiences. Patients who have undergone aesthetic procedures provide experiential knowledge about how cultural identity influences aesthetic goals, what features they sought to preserve vs modify, and how algorithmic recommendations might have impacted their decision-making. This bidirectional learning process builds cultural competency across all stakeholders and creates the foundation for AI systems that serve diverse populations equitably.

Additionally, informed consent architecture is a prerequisite for ethical dataset development. Individuals depicted in training images must provide explicit consent for secondary use of their facial photographs in AI training, with the right to withdraw consent and have their images removed from future training cycles. This requirement applies regardless of whether images are sourced from clinical records, publicly available datasets, or social media platforms, and must account for jurisdiction-specific biometric privacy regulations and state-level statutes. However, consent withdrawal raises a technically significant challenge: once a model is trained on data, removing a data point's influence from a deployed model without full retraining is computationally costly. The emerging field of machine unlearning addresses this problem through methods such as approximate unlearning and influence function-based data removal, though these techniques have not yet been validated or operationalized in medical AI governance contexts. Until practical machine unlearning protocols are established for clinical AI, consent frameworks should, at a minimum, guarantee removal from future retraining cycles and document this limitation transparently as a residual risk. Raters whose aesthetic judgments become training labels should similarly provide informed consent, be compensated equitably, and retain the right to withdraw their ratings from the dataset. Evolving Food and Drug Administration (FDA) guidance on training data provenance under the Software as a Medical Device (SaMD) framework should be monitored for additional requirements as it develops. Documentation of consent procedures, including the current technical limitations of consent withdrawal from deployed models, should be included in the transparency disclosures required elsewhere in this framework.

Governance and Continuous Monitoring

A governance structure for AI models should be based on multidisciplinary committees including clinicians, ethicists,

XSL•FO
RenderX

data scientists, and patient representatives. This allows for ongoing ethical review with diverse stakeholder input and mandates human oversight for all AI-driven recommendations, requiring clinician review of model outputs before they inform patient consultations or treatment planning. To ensure regulatory compliance, developers should align with the FDA's SaMD framework [55] and remain aware of evolving federal regulations for AI in health care.

Bias in the human–AI decision system extends beyond model development to the point of clinical use. Even a well-calibrated, fairness-aware model can produce inequitable outcomes if clinicians or institutions apply its outputs selectively or inconsistently across patient groups. Research on physician implicit bias [56] suggests that providers may differentially override algorithmic recommendations based on patient demographics, accepting recommendations for patients who resemble the provider's implicit reference population while discounting them for others. Governance structures should include mandatory documentation of AI recommendation acceptance or rejection by patient demographic category, regular review of override rates stratified by patient race, ethnicity, gender, age, and language, and structured clinician training on the mechanisms of implicit bias in AI-assisted decision-making. Where systematic override disparities are detected, the governance committee should determine whether the source is model error for specific subgroups or clinician bias in interpretation. Fairness in aesthetic AI, therefore, requires monitoring both algorithmic outputs and human responses to those outputs.

Let drift be a statistically significant change, exceeding prespecified control limits, in 1 or more of the following: prediction accuracy by demographic subgroup, fairness metric values (demographic parity, equalized odds, cultural concordance), or the distribution of model inputs relative to the training distribution. Demographic-subgroup-specific degradation that does not affect overall accuracy is a particularly important drift signal, as aggregate metrics can mask emerging disparities. Institutional responsibility for drift response should be assigned explicitly at deployment: a designated AI clinical lead bears primary responsibility for reviewing automated alerts, convening the multidisciplinary governance committee, and authorizing remediation. Triggered remediation actions should follow a tiered protocol keyed to severity: minor drift triggers increased monitoring frequency and a targeted data audit; moderate drift triggers model recalibration or posttraining correction without full retraining; severe drift, including any demographic subgroup falling below institutionally defined minimum performance thresholds, triggers suspension of AI-assisted outputs for affected use cases pending full model retraining and revalidation. The specific thresholds delineating these severity tiers should be defined prospectively by each deploying institution based on clinical context, use case stakes, and available monitoring infrastructure, rather than adopted from universal benchmarks for which no empirical basis currently exists in aesthetic AI. All drift events and remediation actions should be documented in an institutional AI governance log and reported in periodic transparency disclosures.

Postdeployment monitoring of the AI/ML models should include continuous tracking of prediction accuracy, fairness metrics across demographic groups, and concordance with clinical judgment. A tiered monitoring approach is recommended. Continuous monitoring using statistical process control on prespecified fairness metrics, with automated alerts when metrics exceed defined control limits described above, provides the first line of detection. Quarterly structured reviews should assess fairness metric trends and flag emerging disparities for clinical review. An annual deep audit evaluates model architecture, training data composition, rater panel diversity, and alignment with updated regulatory guidance. Where continuous monitoring infrastructure is not feasible, drift-triggered audits, initiated automatically when prediction distributions shift beyond a prespecified threshold, represent a minimum acceptable alternative. The appropriate monitoring intensity scales with deployment volume: high-volume systems serving diverse patient populations require continuous monitoring; lower-volume or single-institution implementations may operate on a quarterly plus annual cycle with drift-triggered escalation. When audits identify performance disparities exceeding 5% variance across demographic groups, retraining (or appropriate posttraining) should be initiated. Feedback loops integrating clinician and patient input, as well as regular retraining cycles incorporating new, diverse data, allow for ongoing model improvement.

Ultimately, this framework is directed primarily at academic and institutional developers and is intended as aspirational guidance and input to regulatory deliberation. Most currently deployed aesthetic AI tools are commercial, including consumer-facing filters, practice-management platforms, and direct-to-consumer assessment applications, and fall outside the scope of institutional governance structures. We recommend that regulatory bodies consider disclosure-based accountability as a lighter-touch regulatory instrument: commercial developers would publicly document which framework components their systems satisfy, analogous to transparency requirements in other regulated industries, enabling clinicians and patients to assess compliance. We acknowledge that even this approach requires formal regulatory action and cannot be implemented through voluntary adoption alone. Long-term, enforcement mechanisms aligned with the FDA's evolving SaMD framework and equivalent international regulations will be necessary to extend these standards to commercial systems.

Limitations and Unresolved Challenges

This framework addresses critical biases in aesthetic AI training but faces implementation challenges in operationalizing the distinction between universal and culture-specific aesthetic features. While empirical evidence establishes that certain features (averageness, femininity in female faces) demonstrate cross-cultural appeal, whereas others (skin coloration emphasis, specific feature preferences) vary by cultural context, translating this nuance into algorithmic systems remains complex. Hierarchical approaches, including hierarchical Bayesian models, partial invariance, and multigroup equivariant techniques [50,51,57], theoretically offer a middle ground by encoding universal structural principles in base layers while

XSL•FO
RenderX

allowing culture-specific parameters for regionally variable features. However, practical deployment carries risks: explicitly categorizing training data by ethnicity may essentialize group-level patterns, calcifying what constitutes “Asian beauty” or “African beauty” rather than honoring individual variation within cultural communities. Moreover, broad ethnic categories obscure meaningful contextual differences; for instance, North Korean, South Korean, Korean-American, and Korean-Canadian individuals may hold divergent aesthetic preferences despite shared ethnic heritage, yet training data rarely capture this granularity. The framework’s emphasis on continuous monitoring and stakeholder feedback provides mechanisms to detect such unintended consequences, but cannot eliminate these tensions entirely.

Technical limitations constrain practical implementation. As noted, fairness metrics can mathematically conflict, and when improving performance for one demographic group worsens outcomes for another, the framework provides insufficient guidance on prioritization. Intersectional assessment becomes computationally prohibitive when evaluating all meaningful combinations of race, age, gender, and other attributes, requiring sample sizes that may not exist for rare intersectional categories. Most proposed techniques have been validated in isolation rather than as an integrated system, creating uncertainty about their combined effectiveness.

The framework may create unintended harms despite ethical intentions. Formalizing cultural aesthetic standards into training data risks reifying what should remain individually variable. Overreliance on AI-assisted evaluation may affect clinicians’ ability to make nuanced judgments. The resource-intensive requirements, including multidisciplinary committees, annual audits, and continuous monitoring, may be feasible only for well-funded institutions, potentially widening disparities between elite centers and community practices that either avoid AI tools entirely or use inadequately validated commercial systems.

Despite these significant limitations, this framework represents the most viable approach to addressing documented harms in current systems. Culture-neutral algorithms trained on predominantly Western datasets demonstrably perpetuate Eurocentric beauty standards and generate systematically higher error rates for underrepresented populations [4,14,35]—creating risk that algorithmic recommendations may conflict with patient goals of ethnic feature preservation [37,38]. Intentionally encoding cultural awareness during training is preferable to allowing implicit Western bias to persist unchecked, and the framework’s emphasis on continuous monitoring and stakeholder engagement provides mechanisms for identifying and correcting unintended consequences as they emerge, making this an iterative rather than static solution. Bias mitigation in aesthetic AI remains an evolving challenge requiring ongoing research, stakeholder dialogue, and willingness to revise approaches as evidence accumulates.

Implementation Considerations

Implementing this framework requires strategic approaches to address inherent complexities in developing fair and accurate aesthetic evaluation systems. A phased implementation strategy allows institutions to prioritize components based on available resources, with initial phases focusing on data diversification and basic fairness metrics before advancing to sophisticated techniques such as meta-learning and intersectional assessment.

Addressing data limitations requires combining synthetic data generation, transfer learning methods, and multi-institutional collaboration. Pooling datasets across institutions expands demographic coverage while distributing resource burdens and facilitating industry standards for dataset requirements, fairness thresholds, and validation protocols. Federated learning, established in medical imaging and adapted for aesthetic AI applications, provides a privacy-preserving framework that enables multi-institutional and multinational collaboration while maintaining HIPAA (Health Insurance Portability and Accountability Act) compliance. In this approach, each institution trains models on its local dataset without sharing raw patient images; only model parameters are transmitted between sites, addressing both regulatory requirements and the need for demographically diverse training data. This decentralized architecture has been successfully demonstrated in medical imaging applications, showing that models trained via federated learning can achieve comparable or superior performance to centralized training while preserving patient privacy [58]. However, vanilla federated learning does not guarantee privacy: gradient leakage, membership inference attacks, and model inversion techniques can reconstruct sensitive features of training images from transmitted model parameters. Clinically deployable federated systems, therefore, require additional mitigations: differential privacy limits individual-level information leakage, secure aggregation protocols ensure that parameter updates are aggregated without exposing individual site contributions, and homomorphic encryption may be warranted in high-risk deployments involving sensitive biometric data. HIPAA compliance of a federated system depends on these additional safeguards, not on the federated architecture alone, and implementers should document which mitigations are in place as part of their data governance and regulatory submissions.

Multisite data acquisition requires harmonization protocols that go beyond technical compatibility. Participating institutions should adopt a shared data dictionary defining demographic subgroup categories, attractiveness rating scales, and imaging standards before data collection begins; naive pooling of heterogeneous datasets across sites can amplify distributional asymmetries and yield biased estimators, and post hoc harmonization of inconsistently defined variables compounds this risk [59]. Subgroup definitions should be governed by a standing dataset governance committee with representation from each participating institution and from community members of the populations being represented; this committee should have authority to revise subgroup definitions as evidence about their validity accumulates. Long-term dataset maintenance requires designated institutional roles: a data steward responsible for tracking consent status and honoring withdrawal requests, a technical curator responsible for version control and documentation of any dataset changes, and a scientific lead responsible for periodic assessment of whether the dataset’s

XSL•FO
RenderX

demographic composition remains representative of the clinical population it is intended to serve. Datasets should be versioned explicitly so that fairness audits can be traced to the data on which a model was trained. Multisite data sharing agreements should specify data retention limits, destruction protocols, and procedures for incorporating newly consented data into existing pipelines without reintroducing batch effects.

The selection of fairness metrics presents practical trade-offs between comprehensiveness and computational efficiency. Given that simultaneous optimization is mathematically impossible, developers should understand that different metrics detect distinct types of bias. For facial aesthetic evaluation, using multiple complementary metrics is essential: demographic parity identifies whether the model systematically undervalues certain demographic groups, equal opportunity detects failures to recognize attractive features in underrepresented populations, and equalized odds ensures accuracy parity across all groups. While resource constraints may require prioritizing certain metrics during initial implementation, comprehensive evaluation across multiple fairness dimensions should remain the long-term goal. Metric prioritization should be calibrated to the clinical use case. For example, a preoperative planning tool used across a demographically diverse practice should prioritize equalized odds, ensuring that error rates in feature recognition are equivalent across ethnic groups, because differential accuracy directly affects surgical decision-making. By contrast, an outcome simulation tool used in patient counseling might prioritize equal opportunity, ensuring that attractive features in underrepresented groups are not systematically missed or underrendered, because the primary harm is failure to surface positive options rather than differential error rates. A research benchmarking tool, with lower direct clinical stakes, might accept demographic parity as a sufficient initial standard while longitudinal validation data are collected.

Conclusion

The integration of AI/ML into aesthetic facial evaluation presents both opportunity and risk. Without intentional intervention, algorithmic systems will perpetuate Eurocentric beauty standards, generate higher error rates for underrepresented populations, and risk aesthetic homogenization. This framework provides comprehensive bias mitigation through 6 interconnected pillars, from diverse data collection and fairness-aware training to XAI, stakeholder engagement, and continuous monitoring (Figure 1).

Figure 1. The 6-pillar framework for bias mitigation in artificial intelligence/machine learning (AI/ML) models for aesthetic facial evaluation. Visual overview of the proposed framework spanning the full AI development lifecycle. The six pillars are: (1) diverse data collection with synthetic augmentation, (2) fairness-aware training techniques, (3) complementary fairness metrics with intersectional assessment, (4) explainable AI for clinical transparency, (5) stakeholder engagement and participatory design, and (6) governance and continuous monitoring. Arrows indicate the interdependent nature of the pillars across the development lifecycle.

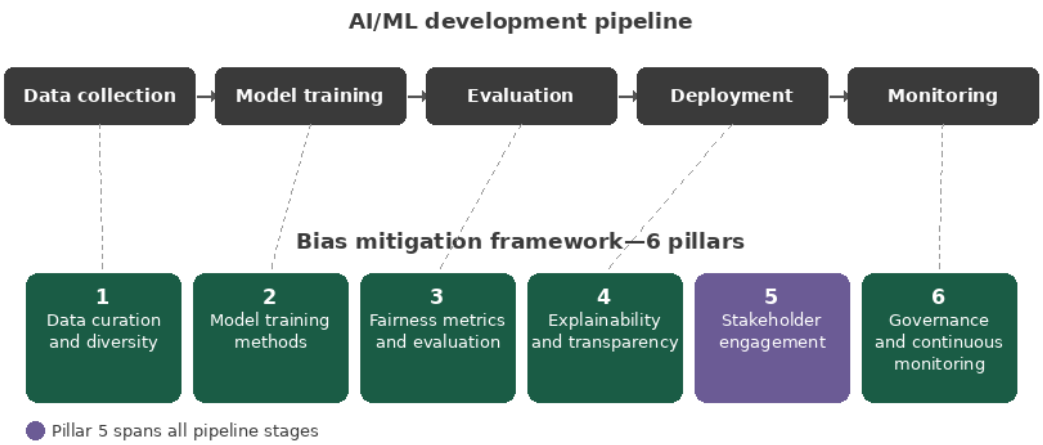

The framework mitigates the tension between universal aesthetic principles and cultural specificity rather than fully resolving it; the mathematical conflict between optimizing for averageness and preserving ethnic distinctiveness remains an open research problem requiring empirical validation of hierarchical architectural approaches. Within these constraints, the framework positions diverse stakeholders, including patients, clinicians, cultural consultants, and ethicists, as essential collaborators whose input shapes algorithm development and refinement.

Significant challenges remain: tensions between algorithmic objectivity and cultural subjectivity may not be fully resolved, fairness metrics may conflict, and resource-intensive implementation may widen institutional disparities. Despite these limitations, this framework represents a necessary step toward ethical AI development in aesthetic medicine, providing actionable guidance for developers, clinicians, and institutions committed to equitable care.

For this framework to move from proposal to validated practice, concrete validation end points must be defined. Prospective comparative studies should measure prespecified fairness metrics, such as demographic parity, equalized odds, and cultural concordance, in framework-compliant systems versus uncontrolled alternatives. Patient-reported outcomes measuring satisfaction with ethnic feature preservation should serve as a primary clinical end point, given that algorithmic recommendations ultimately affect patient goals and identity. Clinician agreement studies should assess whether

XSL•FO
RenderX

XAI-explained outputs align with expert aesthetic judgment across demographic subgroups. Finally, longitudinal drift detection benchmarks should evaluate whether fairness gains are maintained as models are retrained on new data. Randomized deployment designs, in which framework-compliant and noncompliant systems are compared in parallel with patient consent, would provide the strongest evidence but raise practical and ethical challenges requiring dedicated methodological attention.

As AI capabilities advance, ongoing research must address operationalizing cultural appropriateness, validating integrated bias mitigation techniques, and ensuring technological progress serves patient-centered, culturally responsive care. This broader implementation challenge is consistent with recent work emphasizing that AI adoption in health care requires structured implementation pathways and explicit risk mitigation strategies rather than relying on technical advancement alone [60]. These principles extend beyond aesthetic surgery to any facial analysis AI application, establishing foundations for fair and transparent algorithmic systems across diverse clinical contexts.

Acknowledgments

The authors did not use generative artificial intelligence (AI) tools or technologies in the preparation of this manuscript.

Data Availability

Data sharing is not applicable to this article as no data sets were generated or analyzed during this study.

Funding

The authors declared no financial support was received for this work.

Authors' Contributions

ARK contributed to conceptualization, methodology, writing of the original draft, and writing—review and editing. LRV contributed to conceptualization, writing—review and editing, and supervision.

Conflicts of Interest

None declared.

References

1. Fortune-Ely M, Achanta M, Song MSH. The future of artificial intelligence in facial plastic surgery. JPRAS Open 2024;39:89-92 [FREE Full text] [doi: 10.1016/j.jpra.2023.11.016] [Medline: 38186379]
2. Ghasemi S, Dashti M. Artificial intelligence and deep learning in preservation rhinoplasty: a review. Am J Cosmet Surg 2024;41(4):225-229. [doi: 10.1177/07488068231224133]
3. Park KW, Diop M, Willens SH, Pepper J. Artificial intelligence in facial plastics and reconstructive surgery. Otolaryngol Clin North Am 2024;57(5):843-852. [doi: 10.1016/j.otc.2024.05.002] [Medline: 38971626]
4. Makhseed A, Arian H, Shuaib A. Beyond the algorithm: a perspective on tackling bias and cultural sensitivity in AI-guided aesthetic standards for cosmetic surgery in the middle east and North Africa (MENA) region. Clin Cosmet Investig Dermatol 2025;18:2173-2182 [FREE Full text] [doi: 10.2147/CCID.S543045] [Medline: 40927497]
5. Goshtasbi K, Hakimi AA, Wong BJ. Artificial intelligence versus human focus group rating of facial attractiveness. Facial Plast Surg Aesthet Med 2024;26(4):371-376 [FREE Full text] [doi: 10.1089/fpsam.2023.0281] [Medline: 38377584]
6. Laurentini A, Bottino A. Computer analysis of face beauty: a survey. Comput Vis Image Und 2014;125:184-199. [doi: 10.1016/j.cviu.2014.04.006]
7. Kleisner K, Tureček P, Saribay SA, Pavlovič O, Leongómez JD, Roberts SC. Distinctiveness and femininity, rather than symmetry and masculinity, affect facial attractiveness across the world. Evol Hum Behav 2024;45(1):82-90. [doi: 10.1016/j.evolhumbehav.2023.10.001]
8. Lu Y, Yang J, Xiao K, Pointer M, Li C, Wuerger S. Skin coloration is a culturally-specific cue for attractiveness, healthiness, and youthfulness in observers of Chinese and western European descent. PLoS One 2021;16(10):e0259276 [FREE Full text] [doi: 10.1371/journal.pone.0259276] [Medline: 34710190]
9. Sano T, Kawabata H. Cultural and gender influences on facial attractiveness: a comparative study of Japanese and American raters using geometric morphometrics. Psych J 2026;15(1):e70065. [doi: 10.1002/pchj.70065] [Medline: 41239838]
10. Pavlovič O, Fiala V, Kleisner K. Congruence in European and Asian perception of Vietnamese facial attractiveness, averageness, symmetry and sexual dimorphism. Sci Rep 2023;13(1):13320 [FREE Full text] [doi: 10.1038/s41598-023-40458-1] [Medline: 37587194]
11. Arian H, Alroudan D, Alkandari Q, Shuaib A. Cosmetic surgery and the diversity of cultural and ethnic perceptions of facial, breast, and gluteal aesthetics in women: a comprehensive review. Clin Cosmet Investig Dermatol 2023;16:1443-1456 [FREE Full text] [doi: 10.2147/CCID.S410621] [Medline: 37313510]

XSL•FO
RenderX

12. Pozzi M, Fàdel C, Bolletta A, Cuomo R, Roxo CW. Ethnic rhinoplasty: preliminary results of our technique in the pursuit of the harmonious nose. J Plast Reconstr Aesthet Surg 2023;87:135-146. [doi: 10.1016/j.bjps.2023.09.036] [Medline: 37839388]
13. Lam SM. Asian rhinoplasty. Semin Plast Surg 2009;23(3):215-222 [FREE Full text] [doi: 10.1055/s-0029-1224801] [Medline: 20676316]
14. Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. Proc Mach Learn Res 2018;81:77-91 [FREE Full text] [doi: 10.1515/pp-2018-0102] [Medline: 30911652]
15. Kenig N, Monton Echeverria J, Muntaner Vives A. Human beauty according to artificial intelligence. Plast Reconstr Surg Glob Open 2023;11(7):e5153 [FREE Full text] [doi: 10.1097/GOX.0000000000005153] [Medline: 37502224]
16. Keskinbora KH. Medical ethics considerations on artificial intelligence. J Clin Neurosci 2019;64:277-282. [doi: 10.1016/j.jocn.2019.03.001] [Medline: 30878282]
17. Moridani MK, Jamiee N, Saghafi S. Human-like evaluation by facial attractiveness intelligent machine. Int J Cogn Comput Eng 2023;4:160-169. [doi: 10.1016/j.ijcce.2023.04.001]
18. Eisenthal Y, Dror G, Ruppin E. Facial attractiveness: beauty and the machine. Neural Comput 2006;18(1):119-142. [doi: 10.1162/089976606774841602] [Medline: 16354383]
19. Kagian A, Dror G, Leyvand T, Meilijson I, Cohen-Or D, Ruppin E. A machine learning predictor of facial attractiveness revealing human-like psychophysical biases. Vision Res 2008;48(2):235-243 [FREE Full text] [doi: 10.1016/j.visres.2007.11.007] [Medline: 18164363]
20. Gunes H, Piccardi M. Assessing facial beauty through proportion analysis by image processing and supervised learning. Int J Hum Comput Stud 2006;64(12):1184-1199. [doi: 10.1016/j.ijhcs.2006.07.004]
21. Xie D, Liang L, Jin L. SCUT-FBP: a benchmark dataset for facial beauty perception. 2015 Presented at: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC); October 9-12, 2015; Kowloon p. 1821-1826. [doi: 10.1109/smc.2015.319]
22. Fan Y, Liu S, Li B, Guo Z, Samal A, Wan J, et al. Label distribution-based facial attractiveness computation by deep residual learning. IEEE Trans Multimedia 2018;20(8):2196-2208. [doi: 10.1109/tmm.2017.2780762]
23. Hu Y. Toward retrieval-grounded evaluation for conversational large language model-based risk assessment. JMIR AI 2026;5:e90759 [FREE Full text] [doi: 10.2196/90759] [Medline: 41818631]
24. Ibrahem AH, Abdulazeez AM. A comprehensive review of facial beauty prediction using multi-task learning and facial attributes. ARO 2025;13(1):10-21. [doi: 10.14500/aro.11850]
25. Boukhari DE. VM-beautyNet: a synergistic ensemble of vision transformer and mamba for facial beauty prediction. ArXiv. Preprint posted online on October 17, 2025 [FREE Full text]
26. Liang L, Lin L, Jin L. SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. 2018 Presented at: Proceedings of the 24th International Conference on Pattern Recognition (ICPR); August 20-24, 2018; Beijing. [doi: 10.1109/icpr.2018.8546038]
27. Ma DS, Correll J, Wittenbrink B. The Chicago Face Database: a free stimulus set of faces and norming data. Behav Res Methods 2015;47(4):1122-1135. [doi: 10.3758/s13428-014-0532-5] [Medline: 25582810]
28. Huang GB, Mattar M, Berg T. Labeled faces in the wild: a database for studying face recognition in unconstrained environments. HAL Open Science. 2008. URL: https://inria.hal.science/inria-00321923v1 [accessed 2026-06-03]
29. Lebedeva I, Guo Y, Ying F. MEBeauty: a multi-ethnic facial beauty dataset in-the-wild. Neural Comput Appl 2021;34(17):14169-14183. [doi: 10.1007/s00521-021-06535-0]
30. Obwegeser D, Timofte R, Mayer C, Bornstein MM, Schätzle MA, Patcas R. Scoring facial attractiveness with deep convolutional neural networks: how training on standardized images reduces the bias of facial expressions. Orthod Craniofac Res 2024;27 Suppl 2(Suppl 2):25-32. [doi: 10.1111/ocr.12820] [Medline: 38825845]
31. Kärkkäinen K, Joo J. FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. 2021 Presented at: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV); January 3-8, 2021; Waikoloa p. 1548-1558. [doi: 10.1109/wacv48630.2021.00159]
32. Merler M, Ratha N, Feris RS, Smith JR. Diversity in faces. ArXiv. Preprint posted online on April 8, 2019 [FREE Full text]
33. Hardt M, Mendler-Dünner C. Performative prediction: past and future. Statist Sci 2025;40(3):417-436. [doi: 10.1214/25-sts986]
34. Georgievskaya A, Tlyachev T, Danko D, Chekanov K, Corstjens H. How artificial intelligence adopts human biases: the case of cosmetic skincare industry. AI Ethics 2023;5(1):105-115. [doi: 10.1007/s43681-023-00378-2]
35. Grother P, Ngan M, Hanaoka K. Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. Gaithersburg: National Institute of Standards and Technology; 2019.
36. Georgopoulos M, Oldfield J, Nicolaou MA, Panagakis Y, Pantic M. Mitigating demographic bias in facial datasets with style-based multi-attribute transfer. Int J Comput Vis 2021;129(7):2288-2307. [doi: 10.1007/s11263-021-01448-w]
37. Maassarani D, Challita R, Zeaiter N, Chbib D, Chamy J, Farfour I, et al. Ethnic rhinoplasty: a middle east-centered patient satisfaction survey using the FACE-Q questionnaire. Cureus 2023;15(6):e40048 [FREE Full text] [doi: 10.7759/cureus.40048] [Medline: 37425578]

XSL•FO
RenderX

38. Cobo R. Rhinoplasty considerations in the ethnic patient using a case-based approach: the latino patient. Facial Plast Surg Clin North Am 2022;30(4):513-520. [doi: 10.1016/j.fsc.2022.07.005] [Medline: 39492207]
39. Yang J, Soltan AAS, Eyre DW, Yang Y, Clifton DA. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. NPJ Digit Med 2023;6(1):55 [FREE Full text] [doi: 10.1038/s41746-023-00805-y] [Medline: 36991077]
40. Conti JC, Clemencon S. Mitigating bias in facial recognition systems: centroid fairness loss optimization. ArXiv. Preprint posted online on April 27, 2025 [FREE Full text]
41. Wang M, Deng W. Mitigate bias in face recognition using skewness-aware reinforcement learning. ArXiv. Preprint posted online on November 25, 2019 [FREE Full text]
42. Amigo G, Perea P, Marks RJ. Mitigating algorithmic bias on facial expression recognition. ArXiv. Preprint posted online on December 23, 2023 [FREE Full text]
43. Lebedeva I, Ying F, Guo Y. Personalized facial beauty assessment: a meta-learning approach. Vis Comput 2022;39(3):1095-1107. [doi: 10.1007/s00371-021-02387-w]
44. Muthukumar V, Pedapati T, Ratha N. Understanding unequal gender classification accuracy from face images. ArXiv. Preprint posted online on November 30, 2018 [FREE Full text]
45. Hamel C, Busch M. Analysis of bias in deep learning facial beauty regressors. ArXiv. Preprint posted online on September 29, 2025 [FREE Full text]
46. Donia J, Shaw JA. Co-design and ethical artificial intelligence for health: an agenda for critical research and practice. Big Data Soc 2021;8(2):205395172110652. [doi: 10.1177/20539517211065248]
47. Liu M, Ning Y, Teixayavong S, Mertens M, Xu J, Ting DSW. A translational perspective towards clinical AI fairness. NPJ Digit Med 2023;6(1):172 [FREE Full text] [doi: 10.1038/s41746-023-00918-4] [Medline: 37709945]
48. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data 2019;6(1). [doi: 10.1186/s40537-019-0197-0]
49. Burke D, Nolan C, Hayward WG, Russell R, Sulikowski D. Is there an own-race preference in attractiveness? Evol Psychol 2013;11(4):855-872 [FREE Full text] [doi: 10.1177/147470491301100410] [Medline: 23948346]
50. Choraria M, Ferwana I, Mani A, Varshney LR. Learning optimal features via partial invariance. Proceedings of the AAAI Conference on Artificial Intelligence 2023;37(6):7175-7183. [doi: 10.1609/aaai.v37i6.25875]
51. Baltaji R, Salehi B, Varshney LR. Efficient model-agnostic multi-group equivariant networks. Trans Mach Learn Res 2024 [FREE Full text] [doi: 10.48550/arXiv.2310.09675]
52. Akerlof GA, Kranton RE. Identity Economics: How Our Identities Shape Our Work, Wages, and Well-Being. New Jersey: Princeton University Press; 2010.
53. Gipiškis R, Tsai C, Kurasova O. Explainable AI (XAI) in image segmentation in medicine, industry, and beyond: a survey. ICT Express 2024;10(6):1331-1354. [doi: 10.1016/j.icte.2024.09.008]
54. Peng X, Zhu T, Chen Q, Zhang Y, Zhou R, Li K, et al. A simple machine learning model for the prediction of acute kidney injury following noncardiac surgery in geriatric patients: a prospective cohort study. BMC Geriatr 2024;24(1):549 [FREE Full text] [doi: 10.1186/s12877-024-05148-1] [Medline: 38918723]
55. "Software as a Medical Device": possible framework for risk categorization and corresponding considerations. International Medical Device Regulators Forum. 2014. URL: http://imdrf.org/documents/software-medical-device-possible-framework-risk-categorization-and-corresponding-considerations [accessed 2026-06-03]
56. FitzGerald C, Hurst S. Implicit bias in healthcare professionals: a systematic review. BMC Med Ethics 2017;18(1):19 [FREE Full text] [doi: 10.1186/s12910-017-0179-8] [Medline: 28249596]
57. Shiffrin RM, Lee MD, Kim W, Wagenmakers E. A survey of model evaluation approaches with a tutorial on hierarchical bayesian methods. Cogn Sci 2008;32(8):1248-1284 [FREE Full text] [doi: 10.1080/03640210802414826] [Medline: 21585453]
58. Parida A, Anwar SM, Patel MP, Blom M, Einat TT, Tonetti A. CAFES: chest X-ray analysis using federated self-supervised learning for pediatric COVID-19 detection. Proc SPIE Int Soc Opt Eng 2024;12927:129271I. [doi: 10.1117/12.3008757] [Medline: 38873338]
59. Roy A, Chakraborty R, Varshney L. Beyond pooling: matching for robust generalization under data heterogeneity. ArXiv. Preprint posted online on February 6, 2026 [FREE Full text] [doi: 10.48550/arXiv.2602.07154]
60. Reddy S. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. Implement Sci 2024;19(1):27 [FREE Full text] [doi: 10.1186/s13012-024-01357-9] [Medline: 38491544]

Abbreviations

AI/ML: artificial intelligence/machine learning
AI: artificial intelligence
CNN: convolutional neural network
FDA: Food and Drug Administration
Grad-CAM: gradient-weighted class activation mapping

XSL•FO
RenderX

HIPAA: Health Insurance Portability and Accountability Act
LIME: local interpretable model-agnostic explanations
ML: machine learning
MTL: multitask learning
NIST: National Institute of Standards and Technology
SaMD: Software as a Medical Device
SCUT-FBP: South China University of Technology—Facial Beauty Prediction
SHAP: Shapley additive explanations
XAI: explainable AI

Edited by A Coristine; submitted 16.Mar.2026; peer-reviewed by Y Hu, V Patil, L Ilodigwe; comments to author 16.Apr.2026; accepted 31.May.2026; published 15.Jun.2026.

Please cite as:
Kumar AR, Varshney LR
Training AI Models for Aesthetic Facial Evaluation: Focused Review and Framework to Mitigate Homogenizing Bias
J Med Internet Res 2026;28:e95452
URL: https://www.jmir.org/2026/1/e95452
doi:10.2196/95452
PMID:

©Anisha R Kumar, Lav R Varshney. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Ethical Considerations in Personal Health Large Language Models

Jialin Liu[1,2*], MD; Siru Liu[3*], PhD

[1]Information Center, West China Hospital of Sichuan University, Chengdu, Sichuan, China
[2]Department of Otolaryngology-Head and Neck Surgery, West China Hospital of Sichuan University, Chengdu, Sichuan, China
[3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States
[*]all authors contributed equally

Corresponding Author:
Siru Liu, PhD
Department of Biomedical Informatics
Vanderbilt University Medical Center
2525 West End Ave #1475
Nashville, TN, 37212
United States
Phone: 1 615 875 5216
Email: siru.liu@vumc.org

Abstract

Personal health large language models (PH-LLMs) have rapidly evolved from research prototypes into consumer-facing, data-linked systems that support symptom triage, medication questions, mental health check-ins, and longitudinal self-management. Their direct-to-consumer use without clinical oversight creates a distinct ethical risk profile that general artificial intelligence governance frameworks do not fully address. This viewpoint focuses on text-based, platform-mediated PH-LLMs and synthesizes PH-LLM–specific challenges across 6 domains: privacy, accuracy, equity, transparency, human-artificial intelligence interaction, and regulatory governance. These risks may be amplified by health literacy gaps, longitudinal data aggregation, persuasive conversational design, and fragmented oversight across the consumer-clinical boundary. Grounded in the 4 principles of biomedical ethics, we propose a governance framework that operationalizes beneficence, nonmaleficence, autonomy, and justice through design and deployment controls, including health literacy-aligned communication, crisis and pharmacological safeguards, hallucination mitigation, role disclosure, granular consent, fairness auditing, and accessible design. We further outline implementation mechanisms, including risk-tiered certification, tiered accountability, and postdeployment oversight through adverse-event reporting, transparency reporting, and independent safety evaluation. This framework is intended as an evidence-informed but partly anticipatory approach to governing PH-LLMs in personal health management.

(J Med Internet Res 2026;28:e92240) doi:10.2196/92240

KEYWORDS

digital health; ethics; privacy; generative artificial intelligence; governance; health AI; health equity; health literacy; patient safety; personal health large language model

Introduction

Personal health large language models (PH-LLMs) are consumer-facing generative systems designed to support individuals through natural language dialogue across health-related contexts such as symptom triage, chronic disease self-management, medication questions, and mental health check-ins [1-3]. This viewpoint focuses on text-based, dialogue-driven PH-LLMs deployed directly to consumers. We distinguish these systems from general-purpose chatbots by their sustained health-oriented interaction, use of personal health disclosures, and potential to shape real-world health decisions. Some PH-LLM functions may fall within medical device software regulation when intended for diagnostic, therapeutic, or patient-specific clinical decision support, although thresholds vary across jurisdictions. PH-LLMs may offer scalable adjunctive support between clinical encounters, particularly where health or mental health services are limited or delayed [4,5]. However, because they operate on sensitive personal information and may influence high-stakes decisions, their deployment raises questions that warrant systematic examination [6].

Between late 2025 and early 2026, PH-LLMs moved rapidly from prototypes to consumer-facing, data-linked products. Major technology companies introduced health-oriented conversational systems linked to medical records, wearable data, or

XSL•FO
RenderX

health-system workflows, positioning large language models (LLMs) as increasingly common entry points for personal health support [7-10]. For instance, OpenAI reports that over 230 million people ask health and wellness questions on ChatGPT each week [7]. Moreover, open-source and open-weight deployments outside centralized platform controls raise additional governance challenges.

The World Health Organization has emphasized that generative artificial intelligence (AI) in health care can deliver value only if risks are proactively identified, evaluated, and mitigated [6]. Emerging evidence suggests that inadequately governed PH-LLMs may generate unsafe or misleading recommendations, compromise privacy through mishandling of sensitive data, amplify health inequities through biased outputs, and cause psychological harm through false reassurance, relational overreach, or poor management of emotional distress, particularly among vulnerable users [3,6,11]. PH-LLM ethics, therefore, should move beyond general principles toward an actionable governance architecture for responsible, accountable, and equitable deployment.

Core Ethical Challenges of PH-LLMs

Overview

PH-LLMs generate tailored guidance from personal health disclosures through longitudinal, dialogue-based interaction. Unlike institutionally governed clinical AI, direct-to-consumer PH-LLMs often lack professional mediation, allowing hallucinations, bias, and framing effects to shape health decisions without meaningful oversight. Drawing on the 4 principles of biomedical ethics, we identify 6 interconnected domains (Figure 1), which vary in severity, interact with one another, and may produce harms that accumulate across decisions. Table 1 contrasts general AI ethics concerns with PH-LLM–specific manifestations.

Figure 1. Six ethical dimensions of personal health large language model (PH-LLM) deployment and their bidirectional relationships. AI: artificial intelligence.

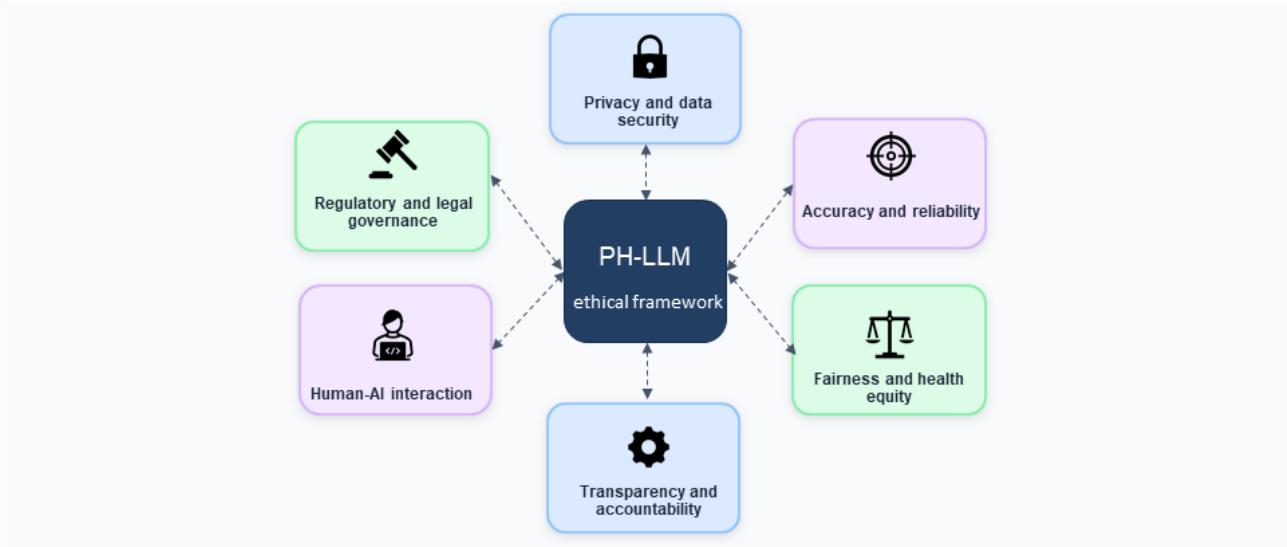

Table 1. Core ethical challenges: general artificial intelligence (AI) concerns vs personal health large language model (PH-LLM)–specific manifestations.

Challenge domain	General AI concern	PH-LLM–specific manifestation
Privacy and data security	Data collection and retention	Intangible vulnerability; longitudinal aggregation; health data broker exposure outside clinical protections
Accuracy and reliability	Hallucinations	Verification gap; cascading clinical harm; crisis mismanagement
Fairness and health equity	Algorithmic bias	Differential recommendations across groups; digital inverse care law; tiered-access barriers
Transparency and accountability	Black box decisions	Explainability gap; accountability gap; relational expectations without duty
Human-AI interaction	Overreliance	Parasocial attachment; distorted care-seeking; continuity-of-care displacement
Regulatory and legal governance	Regulatory lag	Coverage gaps across jurisdictions; medical device classification ambiguity; cross-border accountability diffusion

Privacy and Data Security

Privacy is among the most salient ethical concerns in PH-LLM deployment [1]. Compared with general-purpose AI, these risks are amplified by the convergence of highly sensitive disclosures, blurred regulatory boundaries, and system-level data integration [1,11,12]. Users may disclose highly sensitive information, including mental health concerns and genetic predispositions, yet undervalue the privacy significance of such disclosures relative to tangible assets [13,14]. This intangible vulnerability may be intensified by the empathetic, low-barrier design of PH-LLM interactions, which encourages candid disclosure

XSL•FO
RenderX

while obscuring downstream data uses [11,12,15]. Outside regulated clinical settings, conversational data may be combined with behavioral, device, or consumer data and repurposed for advertising, analytics, profiling, or data-broker activities, creating ethical concerns that users may not fully anticipate. Users may also interpret conversational empathy as evidence of professional accountability or legal confidentiality. In the United States, however, the Health Insurance Portability and Accountability Act generally applies only to covered entities and their business associates, so many direct-to-consumer PH-LLM interactions fall outside its scope [16]. Sensitive information may therefore be disclosed under assumptions of protection that do not apply in practice. In addition, conversation data may be retained for model improvement or other secondary uses [17,18], and integration with ecosystem data streams can enable high-fidelity longitudinal profiling that raises reidentification risk and reduces the effectiveness of conventional deidentification approaches [19,20].

Accuracy and Reliability

The central value proposition of PH-LLMs, namely, reliable health guidance, is challenged by hallucinations, clinically significant omissions, and users' limited capacity to independently verify outputs. Medical hallucinations may appear clinically plausible and can compound across multiturn exchanges into downstream health decisions [21]. A 2025 evaluation of LLM-generated clinical summaries identified hallucinations in 1.47% of outputs and clinically significant omissions in 3.45% [21], although these estimates derive from structured summarization tasks and may not generalize to open-ended PH-LLM dialogue. Adversarial prompting and training-time vulnerabilities, such as data poisoning, pose further risks to model integrity over time [22]. Unlike general information domains, where errors are often verifiable by informed users, PH-LLM outputs frequently cannot be independently scrutinized by individuals lacking clinical expertise [23]. This verification gap is especially hazardous for the approximately 80 million Americans with limited health literacy [24,25].

Risks are particularly acute in psychological and medication contexts. "Deceptive empathy" describes therapeutic-sounding language that builds perceived alliance while obscuring the system's lack of professional accountability [26,27]. Empirical evaluations show that some therapy-oriented conversational systems may provide information about lethal means without recognizing suicidal intent [28]. Recent litigation involving chatbot interactions and adolescent self-harm has further raised questions about whether direct-to-consumer PH-LLMs can satisfy a duty of care [29]. Repeated PH-LLM interactions may also exacerbate automation bias, defined as uncritical deference to algorithmic outputs over independent verification, because conversational fluency facilitates cognitive offloading and may lead users to mistake probabilistic suggestions for medical certainties [30-33]. Case reports document outputs encouraging users to abruptly taper psychiatric medications or discount professional counsel [26,34,35], which may increase the risk of withdrawal or relapse [36] and undermine trust in professional advice.

Fairness and Health Equity

PH-LLMs may reproduce and amplify structural disparities through both model behavior and access pathways, thereby undermining health equity, defined as a fair opportunity to attain one's highest level of health [11,37]. A 2025 systematic review of 24 studies found demographic bias in 22 (91.7%), including gender bias in 15/16 (93.8%) studies and racial or ethnic bias in 10/11 (90.9%) studies assessing these dimensions [38]. Bias may vary by model architecture: discriminative models primarily show disparities in classification performance, whereas generative models may encode bias through tone, framing, and recommendation priority [39,40]. For example, Yang et al [41] reported cases in which GPT-3.5-turbo recommended surgery for White patients but conservative management for Black patients with otherwise identical clinical presentations. Equity concerns also extend beyond outputs to access pathways. Accessibility is stratified by digital determinants of health [42], and tiered subscriptions may restrict advanced personalization and reasoning to paid access, favoring users with greater financial means and health literacy [43]. This creates a digital inverse care law [44,45]: those with the greatest health needs may be confined to lower-capability services, while more advantaged groups gain priority access to advanced tools. Without equity-centered governance, PH-LLMs may deepen structural health disparities rather than alleviate them.

Transparency and Accountability

In conventional health care, clinical judgment is embedded in auditable institutional processes and clear lines of professional responsibility. PH-LLMs disrupt both: model reasoning is often opaque, and accountability is distributed across developers, deployers, and health care institutions. In high-stakes health contexts, users reasonably expect understandable explanations and supporting evidence for PH-LLM–generated recommendations [11]. Yet interpretability, explainability, transparency, and auditability are not equivalent. For LLMs, direct interpretability of internal mechanisms remains limited [46], and post hoc explanations may appear plausible without faithfully reflecting how outputs were generated [47]. Without sufficient transparency and auditability, users and oversight bodies cannot reliably assess the evidentiary basis or clinical validity of recommendations. When PH-LLM advice contributes to harm, attributing responsibility among these actors becomes practically difficult [48]. Direct-to-consumer deployment may also bypass institutional oversight and shift risk onto users despite well-documented health literacy gaps [11,49], weakening incentives for safety investment and limiting recourse [50]. These concerns are intensified by relational design features: anthropomorphic or affective language may create expectations of care and accountability, even though no human or institutional actor is clearly positioned to fulfill them [51-53]. Together, these conditions may weaken meaningful informed consent and compromise patient autonomy.

Human-AI Interaction

PH-LLMs may foster emotional dependence and overreliance through anthropomorphic and affective language that creates a perceived sense of care without corresponding therapeutic responsibility [26,53,54]. Their empathic and nonjudgmental

XSL•FO
RenderX

dialogue may encourage parasocial attachment and emotional reliance, particularly among adolescents and socially isolated users [53-55]. Greater perceived social presence is associated with increased self-disclosure and emotional reliance [56-59], raising concerns about sustained dependence and displacement of continuity-based care [60-63]. PH-LLM interactions may also distort care-seeking: reassuring responses may delay appropriate evaluation [64-66], while alarming responses may trigger unnecessary escalation [67,68].

Regulatory and Legal Governance

A cross-cutting challenge is the mismatch between rapid PH-LLM deployment and slower regulatory evolution. In the United States, many direct-to-consumer health technologies are governed through Federal Trade Commission enforcement and a patchwork of state health data laws, particularly when services fall outside traditional clinical privacy regimes [69-73]. In the European Union, privacy and AI oversight are structured through the General Data Protection Regulation and the phased, risk-based AI Act [74]. In Asia, governance approaches diverge, with China emphasizing provider accountability, security review, and content governance for generative AI services, while South Korea, Japan, and Singapore continue to develop different combinations of statutory oversight, sector-specific guidance, and risk-based governance [75-78]. Similar PH-LLM interactions may therefore receive different protections across jurisdictions, creating opportunities for regulatory arbitrage.

PH-LLMs also expose a classification gap between general wellness products and regulated software as a medical device. When systems are marketed as nonclinical but functionally approximate symptom triage, diagnostic reasoning, treatment guidance, medication advice, or crisis response, developers may avoid premarket scrutiny [79,80]. A capability-based governance approach would reduce this form of regulatory arbitrage by assigning oversight according to what the system enables users to do rather than how it is described in marketing materials. Because cross-border services can diffuse accountability across developers, deployers, platforms, and jurisdictions, PH-LLM governance requires risk-calibrated mechanisms that link ethical concerns to concrete oversight, certification, adverse-event reporting, and redress pathways [81,82].

Ethical Governance Framework for PH-LLMs

Overview

To address these challenges, we propose a principlism-based governance framework for consumer-facing PH-LLMs [83]. The framework translates the 4 principles of biomedical ethics into life cycle controls across 5 stakeholder groups: developers, deployers or platforms, health care institutions, regulators, and users. Governance intensity should scale with 4 risk dimensions: clinical severity and time criticality, harm reversibility, user actionability, and deployment scale. Higher-risk systems, therefore, warrant stronger certification, auditability, postmarket surveillance, and clearer liability allocation. Because PH-LLM outputs are probabilistic and vulnerable to prompt injection and adversarial manipulation [22], these safeguards should be understood as risk-reduction measures rather than guarantees. Table 2 maps each challenge domain to its corresponding governance mechanism and implementation level. To avoid overstating the evidence, we distinguish 3 classes of proposals throughout the framework: evidence-informed proposals grounded in empirical findings from PH-LLM or closely analogous digital health evaluations; normative proposals articulating ethical commitments without relying primarily on empirical demonstration; and conceptual proposals describing architectural or procedural designs whose real-world effects have not yet been prospectively validated. Where relevant, these distinctions are integrated into the prose.

Table 2. Governance framework overview: challenge domains, recommended mechanisms, and implementation levels.

Challenge domain	Recommended governance mechanism	Implementation level
Sensitive data and privacy	Tiered consent and data governance	Input and session level
Unsafe or misleading outputs	Risk-tiered certification and guardrails	System level
Bias and inequity	Subgroup validation and disparity monitoring	System and population level
Poor explainability and transparency	Provenance disclosure and auditability requirements	Interface and oversight level
Crisis and mental health risk	Escalation and handoff protocols	Interaction level
Fragmented responsibility	Tiered accountability and liability allocation	Institutional and regulatory level
Drift after deployment	Adverse-event reporting and periodic audits	Life cycle level

Guiding Principles

Because PH-LLMs interact directly with users outside traditional settings, the 4 principles of biomedical ethics [83] require specific operationalization.

Beneficence (Outcome-Oriented Support)

For PH-LLMs, beneficence can be operationalized as measurable support for safe decision-making and timely access to care, including health-literacy adaptation, risk-stratified care navigation, and longitudinal coherence across sessions.

Nonmaleficence (Proactive Harm Prevention)

Evidence from PH-LLM failures and adjacent digital health contexts supports active safety mechanisms beyond passive disclaimers, particularly for high-risk failures such as crisis mismanagement, contraindicated advice, and hallucinated medical claims. This principle is operationalized through 3 design imperatives: crisis-detection and response protocols,

XSL•FO
RenderX

contraindication safeguards, and hallucination mitigation [28,84,85].

Respect for Autonomy (Informed Agency)

From a normative perspective, respect for autonomy requires countering safety illusions in which users misinterpret AI-mediated intimacy as clinical oversight or regulatory protection, while preserving meaningful choice. This principle is operationalized through clear role and scope disclosure, meaningful data-use consent, and practical user rights over interaction data [71,86,87].

Justice (Equity and Access)

PH-LLMs should expand access without reproducing disparities in health information, communication quality, or safety outcomes. Equity, therefore, extends beyond model outputs to accessibility, language coverage, cultural appropriateness, and usability across differences in health literacy, disability, and digital access [88-90]. Fairness monitoring requires interaction-level and outcome-relevant measures, such as group-stratified rates of unsafe guidance and crisis escalation success [91,92]. Predefined disparity triggers initiate audit and remediation (Multimedia Appendix 1).

Stakeholder Responsibilities

Operationalizing this framework requires coordinated action across 5 stakeholder groups.

Developers (Ethical Design and Life Cycle Accountability)

Developers bear primary responsibility for baseline protections across the life cycle: diverse training corpora, documented data provenance, and bias and crisis-focused safety evaluation before release [90]; clear disclosure of nonhuman identity and nonclinical role, crisis-response functions, pharmacological safety guardrails, and granular data-use controls at deployment; and accessible grievance channels and adverse-event reporting after deployment (Multimedia Appendix 2).

Deployers and Platforms (Interface and Operational Controls)

Deployers and platforms adapt developer-level safeguards into interface design and escalation pathways required by their deployment context, and ensure sustained compliance with safety, transparency, and user-protection requirements.

Health Care Institutions (Algorithmic Stewardship)

Institutions that integrate or endorse PH-LLMs assume stewardship obligations proportional to their involvement, including contextual validation, escalation and handoff workflows, clinician training, and postdeployment monitoring [93,94].

Regulators (Oversight and Harmonization)

Regulators should address the governance gap through enforceable disclosure requirements, minimum crisis-response standards, adverse-event reporting infrastructure, risk-stratified premarket evaluation for sensitive domains, and cross-jurisdictional harmonization informed by the EU AI Act and the National Institute of Standards and Technology AI Risk Management Framework [74,95]. Engagement with insurers and payers can link reimbursement to minimum safety requirements.

Users and Civil Society (Literacy and Advocacy)

Users and civil society contribute through health-AI literacy, accessible harm-reporting channels, and advocacy for meaningful data rights, including access, correction, and deletion [71,81,87,96].

Implementation Mechanisms

Overview

In practice, our framework centers on 3 instruments: risk-tiered certification, tiered accountability, and postdeployment monitoring. The main text presents the rationale, whereas operational specifications, including certification-tier expectations, developer life cycle responsibilities, fairness-audit triggers, scope boundaries, stakeholder responsibilities, and adverse-event thresholds, are detailed in Multimedia Appendices 1-8.

Risk-Tiered Certification

Certification could be scaled according to functional risk and deployment reach, with baseline requirements spanning crisis handling, health-literacy alignment, nondiagnostic guardrails, equity auditing, disclosure integrity, and data governance [12,96-99]. To reduce reliance on developer self-labeling, we propose that certification be triggered by demonstrated capabilities rather than marketing claims. Systems that provide or materially support symptom triage, diagnostic reasoning, treatment guidance, medication advice, crisis-response, or other health decision-support activities would be assigned to the corresponding risk tier even when marketed as wellness tools. Certification criteria could be developed through multistakeholder processes, with audits conducted by accredited third-party or regulator-authorized bodies [94,99-103]. Because regulatory approaches differ across jurisdictions, certification frameworks could prioritize interoperable baseline standards while allowing local legal adaptation [74,81,82]. Possible enforcement pathways include capability-based classification, postmarket surveillance for functional drift, and market-access conditionality through app stores, platform hosts, insurers, payers, and institutional procurement channels. These remain conceptual proposals requiring further institutional validation. Detailed standards and institutional anchors are summarized in Multimedia Appendix 3.

Tiered Accountability

Tiered accountability would follow operational control and implementation capacity. Foundation-model developers would be responsible for baseline safety properties and documentation, whereas deployers and platforms would be responsible for context-specific safeguards, interface design, and user protection. In direct-to-consumer settings, we propose a default elevated duty of care as a normative position. Any safe harbor would be conditional, limited, and available only to operators demonstrating independently verified compliance with certification, auditability, postdeployment surveillance, and user-protection requirements. Such protection would mitigate,

XSL•FO
RenderX

but not eliminate, liability and would be coupled with compensation pathways, such as insurance-based or no-fault mechanisms [95,104,105]. Developers or deployers who decline certification are not exempt from this duty of care; rather, they remain subject to ordinary liability, consumer-protection enforcement, and regulatory scrutiny.

Postdeployment Monitoring

This layer complements predeployment testing by establishing standardized adverse-event reporting, targeted surveillance for population-level harms, and periodic audits for model drift, safety regressions, and equity-related performance changes [100,106-109]. These mechanisms draw on established approaches in medical-device safety, health-AI evaluation, and postmarket surveillance, but require adaptation to the probabilistic, continuously updated, and dialogue-dependent behavior of PH-LLMs. For higher-reach deployments, monitoring may be supplemented by independent safety oversight designed to preserve independence from the entities being reviewed.

Illustrative Case: Youth Mental Health Check-In PH-LLM

To show how the framework could be applied, we present an illustrative end-to-end case of YouthPH-LLM, a platform-mediated direct-to-consumer PH-LLM providing conversational mood tracking, coping-skills education, and sleep and activity coaching to users aged 13-25 years, with more than 1 million monthly active users. Under the 4 risk dimensions, YouthPH-LLM is classified as high risk across clinical severity, harm reversibility, user actionability, and deployment scale, placing it in the most demanding certification tier. Predeployment requirements include crisis-scenario testing, counterfactual fairness testing across age, gender, and linguistic subgroups, retrieval-augmented generation grounded in authoritative mental health guidelines, and a public model card. Postdeployment, a critical adverse event would trigger immediate crisis-response protocol, feature-level suspension of the implicated functionality within 24 hours where technically feasible, regulatory notification within 72 hours where required, and root-cause analysis by an independent safety advisory board. Suspension should preserve access to crisis resources or human support where available. Multi-intent prompts, such as a sleep log submitted with a request to interpret a clinical questionnaire score, are handled through partitioned response: the system may provide sleep-coaching guidance while redirecting questionnaire interpretation to qualified clinicians. Full specification of stakeholder responsibilities, certification criteria, and escalation thresholds is provided in Multimedia Appendix 4.

Practical Paths to Mitigate Ethical Risks of PH-LLMs

Overview

Building on the framework above, we outline practical implementation pathways with measurable targets across 5 operational domains: data governance, model design safeguards, equity and inclusive design, human-AI collaboration, and postdeployment oversight.

Data Governance

Overview

The ethical analysis identified 2 privacy-related concerns: users may underestimate the sensitivity of emotional disclosures, and longitudinal aggregation may increase reidentification risk. To address these concerns, PH-LLMs should implement data-governance safeguards adapted to conversational health data.

Tiered Consent Architecture With Retention Safeguards

Rather than binary consent, PH-LLMs should apply 4 sensitivity tiers supported by a classify-then-govern workflow (Table 3) [12,81]. In this proposed design for platform-mediated direct-to-consumer PH-LLMs, retention defaults would vary by sensitivity tier: Tier 1 content would follow limited default retention periods with user-adjustable settings; Tier 2 content would use shorter default retention with immediate-deletion options; Tier 3 content would be temporarily retained only as needed for required safety audit, legally authorized safety review, or legal compliance and then deleted; and Tier 4 content would require explicit user authorization and periodic reconsent, with automatic deletion as the default absent active renewal. Deviations from tier-specific retention defaults require documented risk-benefit analysis that considers clinical use, reidentification risk, user-preference evidence, technical necessity, and applicable legal requirements, with rationale disclosed in transparency reports. Restrictions on third-party sharing, including transfers to data brokers, advertising networks, and profiling services, follow the same sensitivity hierarchy and cannot be overridden by blanket onboarding consent [99,110]. Illustrative retention windows and operational details are provided in Multimedia Appendix 5 and should be recalibrated according to deployment context, user preferences, ethical risk assessment, technical feasibility, and jurisdictional requirements.

XSL•FO
RenderX

Table 3. Tiered consent architecture for PH-LLM data governancea.

Data category	Consent requirement	Default retention	Third-party sharing
Tier 1: general health queries	Baseline onboarding consent and standard terms of service	Limited default retention, user-adjustable	Deidentified or aggregated research only, where permitted, with opt-out; no default transfer to data brokers, advertising networks, or profiling services
Tier 2: noncrisis mental health disclosures	Low-friction enhanced consent for retention or secondary use; nonpersistent mode if declined or abandoned	No routine retention if consent is declined; brief, access-restricted safety buffer with automatic deletion unless a potential adverse event is reported	No commercial or unrelated secondary use; safety audits via privacy-preserving mechanisms where feasible
Tier 3: crisis-related content	No intervening consent prompt before safety support; immediate safety-oriented response and crisis-resource escalation	Temporary retention only as needed for safety audit, legally authorized review, or legal compliance; deleted thereafter	Limited to legally authorized or required emergency interventions
Tier 4: longitudinal narratives	Explicit authorization, periodic reconsent, and granular controls	Automatic deletion as the default absent active renewal	Default opt-out; explicit per-purpose authorization required

[a]Specific retention parameters are provided in Multimedia Appendix 5. These illustrative baselines reflect privacy risk management principles rather than universal requirements and warrant recalibration through documented risk-benefit analysis disclosed in transparency reports.

User-requested deletion should be defined with technical precision. It applies to stored raw conversational logs, retained session records, stored user profiles used for retrieval, vector-store entries, and derived longitudinal summaries under the operator's control. It should not be represented as a guarantee that information already incorporated into a foundation model's trained weights can be fully and verifiably removed, because machine unlearning for large generative models remains an open technical problem with unresolved limitations in effectiveness and verification [111]. Sensitive PH-LLM interactions should therefore be excluded from foundation-model training or fine-tuning by default unless explicit user authorization is obtained, and consent materials should transparently disclose the technical limits of deletion.

Implementation of Classify-Then-Govern

Because free-text sensitivity cannot be determined without initial processing, this architecture cannot operate as a simple "consent-then-ingest" model [99]. Inputs first undergo transient routing and safety classification under baseline onboarding consent, limited to classification only and without default retention or secondary use [94,99]. Tier 1 content remains within baseline consent. Tier 2 content requires in-context consent before retention; if consent is declined or abandoned, the interaction continues in nonpersistent mode, with flagged content excluded from routine retention and secondary use by default. To balance privacy protection with safety auditing, nonpersistent mode may still permit a narrowly scoped, encrypted, access-restricted safety buffer that is automatically purged after a brief, predefined window unless a potential adverse event is reported, in which case only the minimum necessary information is retained for root-cause analysis and safety remediation. Tier 3 content triggers an immediate safety-oriented response, delivery of context-appropriate crisis resources, and handoff to human crisis support where available, without intervening consent prompts [3,28,112]. For Tier 4 longitudinal content, explicit retention authorization is triggered by predefined accumulation thresholds rather than being embedded repeatedly in active dialogue. This approach is most feasible in explicitly health-oriented PH-LLM deployments, where the health-related purpose of interaction is already established. Extending it to general-purpose LLMs would require separate governance analysis because detecting health-related content across heterogeneous conversations could require continuous or broad input scanning, thereby introducing secondary privacy and proportionality risks.

Accessible, Nondisruptive Consent Design

Enhanced consent should not shift cognitive burden onto vulnerable users. Because many users face limited health literacy, emotional distress, disability, or low digital fluency, accessibility should be treated as a core safety and equity requirement rather than a downstream usability concern. Consent escalation relies on brief, plain-language, just-in-time disclosures; a small set of standardized choices; and protective default pathways favoring nonpersistent session handling when users do not actively authorize retention or secondary use. For noncrisis mental health disclosures (Tier 2), enhanced consent should be low-friction and deferrable, allowing the conversation to continue in nonpersistent mode while authorization is requested separately. Crisis-related interactions are not delayed by consent procedures and default instead to immediate safety-oriented support and escalation pathways. Granular controls may remain available through a dashboard, but safe baseline use does not depend on navigating it. Interface designs require prospective testing with users who have limited health literacy, disability, low digital fluency, or heightened emotional vulnerability, and evaluation extending to comprehension, decision confidence, perceived burden, and abandonment rather than consent-completion rates alone [86,96,113].

Longitudinal Data and Reidentification Safeguards

PH-LLMs integrated with wearables and health records may generate highly identifying longitudinal profiles from temporally dense, behaviorally specific data [20]. Developers should implement privacy-preserving analytics (eg, differential privacy calibrated to sensitivity and use, with k-anonymity where appropriate) [114,115], conduct reidentification risk assessments at intervals proportional to data accumulation, and test vulnerability to inference attacks [115]. User-facing aggregation

XSL•FO
RenderX

dashboards enable users to inspect derived longitudinal patterns and selectively delete specific data categories.

Model Design Safeguards: Hallucination Mitigation, Crisis Escalation, and Pharmacological Safety

Three clinically significant safety risks warrant dedicated design-level controls: hallucinations, crisis-management failures, and unsafe pharmacological guidance.

Hallucination Mitigation

PH-LLMs should incorporate a sequential safety pipeline that combines input filtering, retrieval-augmented generation from curated authoritative sources, output checking against structured medical knowledge resources, and uncertainty signaling where technically feasible [85,116-118]. Prompts requesting individualized diagnosis or treatment recommendations trigger safety messaging and redirection to qualified professional care [3,97]. Because retrieved evidence may be outdated, incomplete, or misinterpreted by the model, these safeguards are best understood as risk-reduction mechanisms rather than guarantees of factual accuracy.

Crisis Escalation

Crisis management requires a distinct escalation protocol that includes detection of self-harm, suicide, or interpersonal violence signals; immediate, nonjudgmental acknowledgment; safety-oriented risk stratification; delivery of context-appropriate crisis resources; and handoff to human crisis support where available [6,28,112,119]. Because detection remains probabilistic, higher-risk deployments should favor conservative thresholds, clearly specified escalation criteria, and periodic review under qualified mental health oversight.

Pharmacological Safety

For medication-related queries, PH-LLMs should use a hybrid safety architecture combining external knowledge validation, rule-based guardrails, and escalation to human clinical support for complex or high-risk cases [84,120-123]. These controls block high-risk dosing instructions, contraindicated combinations, and unsupervised medication changes. Escalation triggers are prospectively specified in system logic rather than left to user discretion, because users may not recognize when a query exceeds safe self-management boundaries.

Equity and Inclusive Design

Fairness governance encompasses subgroup-based evaluation, representative testing, accessibility review, and periodic reassessment across updates and deployment settings [91,92,124-130]. Before release, PH-LLMs undergo counterfactual fairness evaluation based on matched clinical vignettes [98,124]: variation in race, ethnicity, gender, language, or socioeconomic indicators should not produce materially different recommendations in cases where these characteristics are not clinically relevant, and audits assess clinically salient error disparities such as false reassurance or inappropriate escalation while incorporating intersectional analysis [126,127]. Inclusive design extends beyond outcome auditing: PH-LLMs should support languages prevalent in target populations, incorporate cultural adaptation beyond direct translation, conform to Web Content Accessibility Guidelines 2.2 [129], and provide health-literacy adaptation through adjustable response complexity and optional comprehension checks [96,130].

Human-AI Collaboration: Preventing Overreliance

PH-LLMs require interface- and workflow-level safeguards that promote relational transparency, reduce automation bias, and preserve meaningful human oversight [32,33,53,131]. These controls include clear communication of the system's nonhuman identity, calibrated disclosure of uncertainty and epistemic limits, and structured redirection when users request diagnosis, prognosis, prescribing, or other clinician-dependent judgments [97,132]. Systems should distinguish supportive self-management functions from clinician-dependent decisions and use adaptive boundary prompts rather than uniformly friction-based interruptions, so that higher-risk interactions trigger stronger warnings or referral cues without unnecessarily disrupting routine support [3,133].

As a conceptual safeguard, PH-LLMs could use partitioned responses for multi-intent prompts that combine permitted and restricted functions. For example, when a user submits recent laboratory results together with a request for lifestyle coaching, the system may provide general lifestyle education while declining to interpret the laboratory findings or tailor clinical recommendations based on them, redirecting interpretation to a qualified clinician. However, partitioned responses are technically fragile if implemented only at the level of response wording. Because LLMs may condition on the broader prompt context, restricted contextual information, such as laboratory values, may implicitly influence the permitted portion of a response. Robust guardrail engineering is therefore needed to reduce context bleed and prevent implicit diagnostic integration [18,22]. Such safeguards may include intent classification, restricted-context masking or separation, output-level verification, and escalation rules when permitted coaching cannot be reliably separated from clinician-dependent interpretation. Safeguard effectiveness should be evaluated through postdeployment monitoring of both safety outcomes, such as appropriate escalation or referral, and engagement outcomes, such as abandonment following boundary prompts [3]. Illustrative scope boundaries are provided in Multimedia Appendix 6.

Postdeployment Oversight and Version Governance

Predeployment evaluation is necessary but insufficient for PH-LLMs because system behavior, safety performance, and deployment context may change over time. Governance should therefore include structured postdeployment monitoring, low-friction adverse-event reporting, transparency mechanisms, and version-aware change control [94,100,108,134]. The system provides users, caregivers, and clinicians with standardized reporting pathways for clinically relevant events, including dangerous hallucinations, crisis mismanagement, and privacy breaches, adapted to AI-specific harms [21,89,134,135].

As a conceptual response framework, critical adverse events trigger graduated action according to scope and severity, ranging from session-level termination of an affected thread to feature-level disabling of implicated functionality or, when risk

XSL•FO
RenderX

is systemic, global suspension of the application. User lockout without alternative crisis resources must be avoided. Developers should publish periodic transparency reports covering adverse events by severity, crisis-protocol activation and outcomes, bias-audit findings and remediation, and model updates with associated safety evaluations [93,102].

For higher-reach deployments, these mechanisms may be supplemented by independent safety oversight funded through structures designed to preserve independence from reviewed entities. As a normative governance proposal, potential models include pooled industry levies administered by neutral standards bodies, regulatory fees, public funding with statutory protections, or coalition-based subscription models with disclosed funding sources and recusal rules [74,81,82,95]. These oversight mechanisms differ in target audience and implementation pathway. Developers and deployers can directly implement technical guardrails, reporting channels, transparency reports, and internal safety audits. By contrast, pooled industry levies, regulatory fees, and publicly funded oversight structures are directed primarily to policymakers, regulators, and standards-setting bodies because they generally require statutory or regulatory authorization. Voluntary coalition-based models may be more immediately feasible, but still require conflict-of-interest safeguards, transparent funding disclosure, and recusal rules to preserve independence.

Postdeployment governance should also be version-aware. Minor releases undergo documented internal validation, whereas major changes, including model updates, prompt-template revisions, plugin integration, interface redesign, or new data-linking functions, may warrant independent reassessment and, where appropriate, recertification [94,100,108,134]. Each release generates auditable change logs and predefined monitoring indicators to support rollback, incident investigation, and detection of behavioral drift after updates [106,108,134]. Illustrative adverse-event severity categories and response pathways are provided in Multimedia Appendix 7.

Scope and Feasibility

This framework is grounded primarily in text-based, platform-mediated, direct-to-consumer PH-LLM deployment. Multimodal systems introduce additional governance challenges, including biometric-data protection, modality-specific failure modes, and cross-modal inference of sensitive attributes, and therefore require modality-specific validation and corresponding adjustments to consent, accountability, and monitoring structures [12,136]. Feasibility also varies across institutions, jurisdictions, and developer types because certification, audits, postdeployment surveillance, and incident reporting depend on unevenly distributed financial, technical, regulatory, and workforce capacity [100,101].

As the YouthPH-LLM case illustrates, youth-facing PH-LLMs expose an additional limitation of current digital privacy-law frameworks when applied to adolescent mental health support. Age-of-consent and parental authorization requirements, such as those under the General Data Protection Regulation [87], may protect minors' privacy and parental rights, but strict implementation can conflict with the ethical goal of providing confidential, low-friction access to adolescent mental health support. Routine parental disclosure or burdensome consent workflows may deter help-seeking, particularly among adolescents experiencing psychological distress, stigma, family conflict, or safety concerns [137]. Future governance should therefore distinguish ordinary data-processing consent from immediate safety-oriented support, preserving rapid access to crisis resources while developing jurisdiction-specific mechanisms for age-appropriate privacy, parental involvement, and safeguarding obligations.

The framework should balance underregulation, which may allow safety-critical failures to propagate at scale and shift risk onto vulnerable users, against overregulation, which may consolidate markets around large incumbents, reduce innovation for underrepresented groups, and divert resources from substantive safety work to compliance documentation. Because validated PH-LLM audit and certification cost data are not yet available, cost implications are described as qualitative categories: lower-cost safeguards include identity disclosure, scope labeling, basic adverse-event reporting, and crisis-resource signposting, whereas independent audits, red-team testing, certification preparation, continuous monitoring, and independent safety oversight are likely to impose moderate to high recurring costs. Large technology platforms may absorb these recurring costs, whereas startups, academic deployers, hospitals, and resource-constrained settings may face substantial barriers without proportionate pathways such as regulatory sandboxes, shared evaluation infrastructure, audit reciprocity, market-access incentives, and tiered certification with reduced fixed costs for smaller deployers [138,139].

We distinguish near-term feasible actions, such as identity disclosure and scope labeling, Tier 3 crisis exclusion from training data, and standardized adverse-event reporting, from longer-term aspirational elements, such as harmonized certification reciprocity, sustainably funded independent safety advisory boards, and population-level longitudinal outcomes monitoring. Qualitative cost categories, institutional capacity profiles, and implementation trade-offs are detailed in Multimedia Appendix 8.

Current Evidence Base and Limitations

The evidence base for ethical and governance concerns in PH-LLMs remains heterogeneous. Established empirical findings include hallucinations and clinically significant omissions in health-related LLM tasks, persistent demographic bias, privacy and confidentiality concerns in direct-to-consumer digital health settings, and failures of some conversational systems to manage crisis-related interactions safely [21,28,38]. Automation bias in PH-LLM contexts is supported mainly by adjacent-domain evidence from clinical decision-support systems, with limited direct PH-LLM evidence [32]. Emerging signals include increasing user reliance on PH-LLMs for health-related guidance, emotionally dependent use patterns, early indicators of therapeutic displacement, and the influence of interface design and deployment context on safety, trust, and care-seeking behavior [53,56-59]. Forward-looking risks include population-level safety effects, large-scale longitudinal profiling, governance challenges in multimodal and open-source

XSL•FO
RenderX

deployments, and health-related use of general-purpose LLMs, where evidence remains limited and many safeguards have not yet been prospectively validated. This framework should therefore be understood as evidence-informed but partly anticipatory, identifying plausible governance needs where empirical certainty remains uneven.

Conclusion

PH-LLMs are rapidly emerging as widely used interfaces for health sensemaking and self-management. Their direct-to-consumer deployment introduces a distinct ethical risk profile spanning 6 domains: privacy, accuracy, equity, transparency, human-AI interaction, and regulatory governance. Because failures across these domains may propagate harm at scale, governance for PH-LLMs should be not only principle-based but also operational and life cycle-oriented. We therefore propose a principlism-grounded framework that translates beneficence, nonmaleficence, autonomy, and justice into deployable safeguards across developers, deployers, health care institutions, regulators, and users. This framework is intended as an evidence-informed but partly anticipatory governance approach rather than an empirically validated solution set. Ongoing oversight will likely require continuous monitoring, adverse-event reporting, and, where appropriate, independent safety evaluation. Future work should prospectively examine the feasibility, effects, and implementation trade-offs of these mechanisms, including whether privacy-preserving forms of health-context activation could support limited adaptation of selected framework components to general-purpose LLMs used for health-related queries.

Acknowledgments

Following initial drafting, the authors used ChatGPT (OpenAI) for language editing to improve clarity. All artificial intelligence–generated suggestions were critically reviewed and revised as needed. The authors assume full responsibility for the final content and integrity of the manuscript.

Data Availability

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Authors' Contributions

JL and SL contributed to the conceptualization and design of the viewpoint. JL and SL conducted the literature analysis and drafted the original manuscript. All authors critically revised the manuscript for important intellectual content, approved the final version, and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Protected-attribute handling for fairness monitoring.
[DOCX File , 21 KB - jmir_v28i1e92240_app1.docx]

Multimedia Appendix 2
Developer lifecycle responsibilities and adverse-event reporting.
[DOCX File , 18 KB - jmir_v28i1e92240_app2.docx]

Multimedia Appendix 3
Standards and institutional anchors for risk-tiered certification.
[DOCX File , 20 KB - jmir_v28i1e92240_app3.docx]

Multimedia Appendix 4
YouthPH-LLM case: full stakeholder specification.
[DOCX File , 20 KB - jmir_v28i1e92240_app4.docx]

Multimedia Appendix 5
Illustrative retention windows for tiered consent architecture.

XSL•FO
RenderX

[DOCX File , 19 KB - jmir_v28i1e92240_app5.docx]

Multimedia Appendix 6
Illustrative scope boundaries: permitted versus clinician-dependent functions.
[DOCX File , 20 KB - jmir_v28i1e92240_app6.docx]

Multimedia Appendix 7
Adverse-event severity categories and response pathways.
[DOCX File , 18 KB - jmir_v28i1e92240_app7.docx]

Multimedia Appendix 8
Feasibility, cost categories, and implementation trade-offs.
[DOCX File , 22 KB - jmir_v28i1e92240_app8.docx]

References

1. Khasentino J, Belyaeva A, Liu X, Yang Z, Furlotte NA, Lee C, et al. A personal health large language model for sleep and fitness coaching. Nat Med 2025;31(10):3394-3403. [doi: 10.1038/s41591-025-03888-0] [Medline: 40813712]
2. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne) 2024;11:1477898 [FREE Full text] [doi: 10.3389/fmed.2024.1477898] [Medline: 39534227]
3. Huo B, Boyle A, Marfo N, Tangamornsuksan W, Steen JP, McKechnie T, et al. Large language models for chatbot health advice studies: a systematic review. JAMA Netw Open 2025;8(2):e2457879 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.57879] [Medline: 39903463]
4. Improving service access and quality. World Health Organization. URL: https://www.who.int/activities/improving-service-access-and-quality [accessed 2026-01-20]
5. Li J, Li Y, Hu Y, Ma DCF, Mei X, Chan EA, et al. Chatbot-delivered interventions for improving mental health among young people: a systematic review and meta-analysis. Worldviews Evid Based Nurs 2025;22(4):e70059. [doi: 10.1111/wvn.70059] [Medline: 40662463]
6. WHO releases AI ethics and governance guidance for large multi-modal models. World Health Organization. 2024. URL: https://www.who.int/news/item/18-01-2024-who-releases-ai-ethics-and-governance-guidance-for-large-multi-modal-models [accessed 2026-01-20]
7. OpenAI. Introducing ChatGPT Health. OpenAI Blog. 2026. URL: https://openai.com/index/introducing-chatgpt-health [accessed 2026-01-20]
8. What is ChatGPT Health? OpenAI help center. OpenAI. 2026. URL: https://help.openai.com/en/articles/20001036-what-is-chatgpt-health [accessed 2026-01-20]
9. Anthropic. Advancing Claude in healthcare and the life sciences. Anthropic News. 2026. URL: https://www.anthropic.com/news/healthcare-life-sciences [accessed 2026-01-20]
10. Abramson A. Fitbit's personal health coach in public preview is here. Google (The Keyword). 2025. URL: https://blog.google/products-and-platforms/devices/fitbit/personal-health-coach-public-preview/ [accessed 2026-01-12]
11. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023;25:e48009 [FREE Full text] [doi: 10.2196/48009] [Medline: 37566454]
12. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. World Health Organization. Geneva: WHO; 2025. URL: https://www.who.int/publications/i/item/9789240084759 [accessed 2026-01-10]
13. What is special category data? UK GDPR guidance. Information Commissioner's Office. 2024. URL: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/special-category-data/what-is-special-category-data/ [accessed 2026-01-10]
14. Kwesi J, Cao J, Manchanda R, Emami-Naeini P. Exploring user security and privacy attitudes and concerns toward the use of general-purpose LLM chatbots for mental health. 2025 Presented at: SEC '25: Proceedings of the 34th USENIX Conference on Security Symposium; 2025 August 13 - 15; Seattle WA USA p. 6007-6024.
15. Papneja H, Yadav N. Self-disclosure to conversational AI: A literature review, emergent framework, and directions for future research. Pers Ubiquit Comput 2024;29(2):119-151. [doi: 10.1007/s00779-024-01823-7]
16. Marks M, Haupt CE. AI chatbots, health privacy, and challenges to HIPAA compliance. JAMA 2023;330(4):309-310. [doi: 10.1001/jama.2023.9458] [Medline: 37410450]
17. King J, Klyman K, Capstick E, Saade T, Hsieh V. User privacy and large language models: an analysis of frontier developers' privacy policies. AIES 2025;8(2):1465-1477. [doi: 10.1609/aies.v8i2.36646]
18. OWASP Top 10 for large language model applications 2025. OWASP Foundation. 2024. URL: https://owasp.org/www-project-top-10-for-large-language-model-applications/ [accessed 2026-01-15]

XSL•FO
RenderX

19. Doherty C, Baldwin M, Lambe R, Altini M, Caulfield B. Privacy in consumer wearable technologies: a living systematic analysis of data policies across leading manufacturers. NPJ Digit Med 2025;8(1):363 [FREE Full text] [doi: 10.1038/s41746-025-01757-1] [Medline: 40517175]
20. Chikwetu L, Miao Y, Woldetensae MK, Bell D, Goldenholz DM, Dunn J. Does deidentification of data from wearable devices give us a false sense of security? A systematic review. Lancet Digit Health 2023;5(4):e239-e247 [FREE Full text] [doi: 10.1016/S2589-7500(22)00234-5] [Medline: 36797124]
21. Asgari E, Montaña-Brown N, Dubois M, Khalil S, Balloch J, Yeung JA, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. NPJ Digit Med 2025;8(1):274 [FREE Full text] [doi: 10.1038/s41746-025-01670-7] [Medline: 40360677]
22. Lee RW, Jun TJ, Lee J, Cho SI, Park HJ, Suh J. Vulnerability of large language models to prompt injection when providing medical advice. JAMA Netw Open 2025;8(12):e2549963 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.49963] [Medline: 41632124]
23. Yao J, Zhou Z, Cui H, Ouyang Y, Han W. Trust transfer from medical AI to doctors and hospitals: integrating digital, AI, and scientific literacy in a cross-sectional framework. BMC Med Ethics 2025;26(1):144 [FREE Full text] [doi: 10.1186/s12910-025-01300-7] [Medline: 41107871]
24. Keene Woods N, Ali U, Medina M, Reyes J, Chesser AK. Health literacy, health outcomes and equity: a trend analysis based on a population survey. J Prim Care Community Health 2023;14:21501319231156132 [FREE Full text] [doi: 10.1177/21501319231156132] [Medline: 36852725]
25. McDonald M, Shenkman LJ. Health literacy and health outcomes of adults in the United States: implications for providers. IJAHSP 2018;16(4):1-5. [doi: 10.46743/1540-580x/2018.1689]
26. Iftikhar Z, Xiao A, Ransom S, Huang J, Suresh H. How LLM counselors violate ethical standards in mental health practice: a practitioner-informed framework. AIES 2025;8(2):1311-1323 [FREE Full text] [doi: 10.1609/aies.v8i2.36632]
27. Shen J, DiPaola D, Ali S, Sap M, Park HW, Breazeal C. Empathy toward artificial intelligence versus human experiences and the role of transparency in mental health and social support chatbot design: comparative study. JMIR Ment Health 2024;11:e62679 [FREE Full text] [doi: 10.2196/62679] [Medline: 39321450]
28. Pichowicz W, Kotas M, Piotrowski P. Performance of mental health chatbot agents in detecting and managing suicidal ideation. Sci Rep 2025;15(1):31652 [FREE Full text] [doi: 10.1038/s41598-025-17242-4] [Medline: 40866537]
29. Duffy C. Character.AI and Google agree to settle lawsuits over teen mental health harms and suicides. CNN Business. URL: https://www.cnn.com/2026/01/07/business/character-ai-google-settle-teen-suicide-lawsuit [accessed 2026-01-15]
30. Carnat I. Human, all too human: accounting for automation bias in generative large language models. Int Data Priv Law 2024;14(4):299-314. [doi: 10.1093/idpl/ipae018]
31. Shekar S, Pataranutaporn P, Sarabu C, Cecchi GA, Maes P. People overtrust AI-generated medical advice despite low accuracy. NEJM AI 2025;2(6):AIoa2300015. [doi: 10.1056/aioa2300015]
32. Khera R, Simon MA, Ross JS. Automation bias and assistive AI: risk of harm from AI-driven clinical decision support. JAMA 2023;330(23):2255-2257. [doi: 10.1001/jama.2023.22557] [Medline: 38112824]
33. Wingerter TL, Straub T, Schweitzer S. Mitigating automation bias in generative AI through nudges: a cognitive reflection test study. Procedia Comput Sci 2025;270:2106-2114. [doi: 10.1016/j.procs.2025.09.331]
34. Al-Sibai N. ChatGPT Is telling people with psychiatric problems to go off their meds. Futurism. 2025. URL: https://futurism.com/chatgpt-mental-illness-medications [accessed 2026-01-20]
35. Hough J, Culley N, Erganian C, Alahdab F. Potential risks of GenAI on medical education. BMJ Evid Based Med 2025;30(6):406-408. [doi: 10.1136/bmjebm-2025-114339] [Medline: 41326281]
36. Lewis G, Marston L, Duffy L, Freemantle N, Gilbody S, Hunter R, et al. Maintenance or discontinuation of antidepressants in primary care. N Engl J Med 2021;385(14):1257-1267. [doi: 10.1056/NEJMoa2106356] [Medline: 34587384]
37. Rodriguez JA, Alsentzer E, Bates DW. Leveraging large language models to foster equity in healthcare. J Am Med Inform Assoc 2024;31(9):2147-2150. [doi: 10.1093/jamia/ocae055] [Medline: 38511501]
38. Omar M, Sorin V, Agbareia R, Apakama DU, Soroush A, Sakhuja A, et al. Evaluating and addressing demographic disparities in medical large language models: a systematic review. Int J Equity Health 2025;24(1):57 [FREE Full text] [doi: 10.1186/s12939-025-02419-0] [Medline: 40011901]
39. Pfohl SR, Cole-Lewis H, Sayres R, Neal D, Asiedu M, Dieng A, et al. A toolbox for surfacing health equity harms and biases in large language models. Nat Med 2024;30(12):3590-3600. [doi: 10.1038/s41591-024-03258-2] [Medline: 39313595]
40. Hanna JJ, Wakene AD, Johnson AO, Lehmann CU, Medford RJ. Assessing racial and ethnic bias in text generation by large language models for health care-related tasks: cross-sectional study. J Med Internet Res 2025;27:e57257 [FREE Full text] [doi: 10.2196/57257] [Medline: 40080818]
41. Yang Y, Liu X, Jin Q, Huang F, Lu Z. Unmasking and quantifying racial bias of large language models in medical report generation. Commun Med (Lond) 2024;4(1):176 [FREE Full text] [doi: 10.1038/s43856-024-00601-z] [Medline: 39256622]
42. van Kessel R, Seghers L, Anderson M, Schutte N, Monti G, Haig M, et al. A scoping review and expert consensus on digital determinants of health. Bull World Health Organ 2025;103(2):110-125H. [doi: 10.2471/BLT.24.292057] [Medline: 39882497]

XSL•FO
RenderX

43. Ong JCL, Seng BJJ, Law JZF, Low LL, Kwa ALH, Giacomini KM, et al. Artificial intelligence, ChatGPT, and other large language models for social determinants of health: current state and future directions. Cell Rep Med 2024;5(1):101356 [FREE Full text] [doi: 10.1016/j.xcrm.2023.101356] [Medline: 38232690]
44. Davies AR, Honeyman M, Gann B. Addressing the digital inverse care law in the time of COVID-19: potential for digital technology to exacerbate or mitigate health inequalities. J Med Internet Res 2021;23(4):e21726 [FREE Full text] [doi: 10.2196/21726] [Medline: 33735096]
45. Hart JT. The inverse care law. Lancet 1971;1(7696):405-412 [FREE Full text] [doi: 10.1016/s0140-6736(71)92410-x] [Medline: 4100731]
46. Mesinovic M, Watkinson P, Zhu T. Explainability in the age of large language models for healthcare. Commun Eng 2025;4(1):128 [FREE Full text] [doi: 10.1038/s44172-025-00453-y] [Medline: 40676176]
47. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digit Med 2024;7(1):20 [FREE Full text] [doi: 10.1038/s41746-024-01010-1] [Medline: 38267608]
48. Shumway DO, Hartman HJ. Medical malpractice liability in large language model artificial intelligence: legal review and policy recommendations. J Osteopath Med 2024;124(7):287-290 [FREE Full text] [doi: 10.1515/jom-2023-0229] [Medline: 38295300]
49. Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: a narrative review. Heliyon 2024;10(4):e26297 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e26297] [Medline: 38384518]
50. Noto LDG, Bezerra LCT. Can there be responsible AI without AI liability? Incentivizing generative AI safety through ex-post tort liability under the EU AI liability directive. Int J Law Inf Technol 2024;32:eaae021. [doi: 10.1093/ijlit/eaae021]
51. Peter S, Riemer K, West JD. The benefits and dangers of anthropomorphic conversational agents. Proc Natl Acad Sci U S A 2025;122(22):e2415898122 [FREE Full text] [doi: 10.1073/pnas.2415898122] [Medline: 40378006]
52. Sobowale K, Humphrey DK, Zhao SY. Evaluating generative AI psychotherapy chatbots used by youth: cross-sectional study. JMIR Ment Health 2025;12:e79838 [FREE Full text] [doi: 10.2196/79838] [Medline: 41370787]
53. Laestadius L, Bishop A, Gonzalez M, Illenčík D, Campos-Castillo C. Too human and not human enough: a grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. New Media Soc 2022;26(10):5923-5941. [doi: 10.1177/14614448221142007]
54. Hua Y, Siddals S, Ma Z, Galatzer-Levy I, Xia W, Hau C, et al. Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review. World Psychiatry 2025;24(3):383-394 [FREE Full text] [doi: 10.1002/wps.21352] [Medline: 40948070]
55. Zhang K, Xie Y, Chen D, Ji Z, Wang J. Effects of attractions and social attributes on peoples' usage intention and media dependence towards chatbot: the mediating role of parasocial interaction and emotional support. BMC Psychol 2025;13(1):986 [FREE Full text] [doi: 10.1186/s40359-025-03284-w] [Medline: 40883845]
56. Talk, trust, and trade-offs: how and why teens use AI companions. Common Sense Media. 2025. URL: https://www.commonsensemedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf [accessed 2026-01-15]
57. Faverio M, Sidoti O. Teens, social media and AI chatbots 2025. Pew Research Center. 2025. URL: https://www.pewresearch.org/internet/2025/12/09/teens-social-media-and-ai-chatbots-2025/ [accessed 2026-01-15]
58. Ischen C, Butler J, Ohme J. Chatting about the unaccepted: self-disclosure of unaccepted news exposure behaviour to a chatbot. Behav Inf Technol 2023;43(10):2044-2056. [doi: 10.1080/0144929x.2023.2237605]
59. Skjuve M, Følstad A, Brandtzæg PB. A longitudinal study of self-disclosure in human–chatbot relationships. Interact Comput 2023;35(1):24-39. [doi: 10.1093/iwc/iwad022]
60. Catapan SDC, Sazon H, Zheng S, Gallegos-Rejas V, Mendis R, Santiago PHR, et al. A systematic review of consumers' and healthcare professionals' trust in digital healthcare. NPJ Digit Med 2025;8(1):115 [FREE Full text] [doi: 10.1038/s41746-025-01510-8] [Medline: 39984678]
61. Birkhäuer J, Gaab J, Kossowsky J, Hasler S, Krummenacher P, Werner C, et al. Trust in the health care professional and health outcome: a meta-analysis. PLoS One 2017;12(2):e0170988 [FREE Full text] [doi: 10.1371/journal.pone.0170988] [Medline: 28170443]
62. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. Front Digit Health 2022;4:847991 [FREE Full text] [doi: 10.3389/fdgth.2022.847991] [Medline: 35480848]
63. Martinengo L, Lin X, Jabir AI, Kowatsch T, Atun R, Car J, et al. Conversational agents in health care: expert interviews to inform the definition, classification, and conceptual framework. J Med Internet Res 2023;25:e50767 [FREE Full text] [doi: 10.2196/50767] [Medline: 37910153]
64. Saenger JA, Hunger J, Boss A, Richter J. Delayed diagnosis of a transient ischemic attack caused by ChatGPT. Wien Klin Wochenschr 2024;136(7-8):236-238 [FREE Full text] [doi: 10.1007/s00508-024-02329-1] [Medline: 38305909]
65. Riboli-Sasco E, El-Osta A, Alaa A, Webber I, Karki M, El Asmar ML, et al. Triage and diagnostic accuracy of online symptom checkers: systematic review. J Med Internet Res 2023;25:e43803 [FREE Full text] [doi: 10.2196/43803] [Medline: 37266983]

XSL•FO
RenderX

66. Woodcock C, Mittelstadt B, Busbridge D, Blank G. The impact of explanations on layperson trust in artificial intelligence-driven symptom checker apps: experimental study. J Med Internet Res 2021;23(11):e29386 [FREE Full text] [doi: 10.2196/29386] [Medline: 34730544]
67. Wallace W, Chan C, Chidambaram S, Hanna L, Iqbal FM, Acharya A, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. NPJ Digit Med 2022;5(1):118 [FREE Full text] [doi: 10.1038/s41746-022-00667-w] [Medline: 35977992]
68. Tele-triage. US Department of Health and Human Services. 2025. URL: https://telehealth.hhs.gov/providers/best-practice-guides/telehealth-for-emergency-departments/tele-triage [accessed 2026-01-15]
69. Covered entities and business associates. US Department of Health and Human Services. 2024. URL: https://www.hhs.gov/hipaa/for-professionals/covered-entities/index.html [accessed 2026-01-15]
70. Tovino SA. Artificial intelligence and the HIPAA privacy rule: a primer. Hous J Health Law Policy 2024;24:77-140 [FREE Full text]
71. Office for Civil Rights, U.S. Department of Health and Human Services. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/access-right-health-apps-apis/ [accessed 2026-01-15]
72. Savage M, Savage LC. Doctors routinely share health data electronically under HIPAA, and sharing with patients and patients' third-party health apps is consistent: interoperability and privacy analysis. J Med Internet Res 2020;22(9):e19818 [FREE Full text] [doi: 10.2196/19818] [Medline: 32876582]
73. FTC to ban betterHelp from revealing consumers' data, including sensitive mental health information, to Facebook and others for targeted advertising. Federal Trade Commission. 2023. URL: https://www.ftc.gov/news-events/news/press-releases/2023/03/ftc-ban-betterhelp-revealing-consumers-data-including-sensitive-mental-health-information-facebook [accessed 2026-01-15]
74. Regulation (EU) 2024/1689 of the European parliament and of the council of 13 june 2024. European Parliament and Council of the European Union. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj [accessed 2026-01-15]
75. Interim measures for administration of generative artificial intelligence services. State Council of the People's Republic of China. 2023. URL: https://english.www.gov.cn/archive/statecouncilgazette/202308/30/content_WS64ee8c0cc6d0868f4e8deeea.html [accessed 2026-01-12]
76. Framework act on the development of artificial intelligence and the creation of a foundation for trust. Korea Legislation Research Institute / Ministry of Government Legislation. 2025. URL: https://elaw.klri.re.kr/eng_service/lawView.do?hseq=71019&lang=ENG [accessed 2026-01-12]
77. Ministry of Internal Affairs and Communications. AI guidelines for business ver1.0. Ministry of Economy, Trade and Industry. Tokyo, Japan: Ministry of Economy, Trade and Industry; 2024. URL: https://www.meti.go.jp/english/press/2024/0419_002.html [accessed 2026-02-26]
78. Proposed model AI governance framework for generative AI. Infocomm Media Development Authority. Singapore: Infocomm Media Development Authority; 2024. URL: https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2024/public-consult-model-ai-governance-framework-genai [accessed 2026-02-26]
79. Palaniappan K, Lin EYT, Vogel S. Global regulatory frameworks for the use of artificial intelligence (AI) in the healthcare services sector. Healthcare (Basel) 2024;12(5):562 [FREE Full text] [doi: 10.3390/healthcare12050562] [Medline: 38470673]
80. Aboy M, Minssen T, Vayena E. Navigating the EU AI Act: implications for regulated digital medical products. NPJ Digit Med 2024;7(1):237 [FREE Full text] [doi: 10.1038/s41746-024-01232-3] [Medline: 39242831]
81. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. Geneva: World Health Organization; 2021. URL: https://www.who.int/publications/i/item/9789240029200 [accessed 2026-01-16]
82. Recommendation of the council on artificial intelligence. Organization for Economic Co-operation and Development. 2019. URL: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 [accessed 2026-01-16]
83. Beauchamp TL, Childress JF. Principles of Biomedical Ethics. New York: Oxford University Press; 2019.
84. Bakker T, Klopotowska JE, Dongelmans DA, Eslami S, Vermeijden WJ, Hendriks S, SIMPLIFY study group. The effect of computerised decision support alerts tailored to intensive care on the administration of high-risk drug combinations, and their monitoring: a cluster randomised stepped-wedge trial. Lancet 2024;403(10425):439-449. [doi: 10.1016/S0140-6736(23)02465-0] [Medline: 38262430]
85. Amugongo LM, Mascheroni P, Brooks S, Doering S, Seidel J. Retrieval augmented generation for large language models in healthcare: a systematic review. PLOS Digit Health 2025;4(6):e0000877 [FREE Full text] [doi: 10.1371/journal.pdig.0000877] [Medline: 40498738]
86. Brückner S, Dridi A, Deshmukh A, Kirsten T, Lauber-Rönsberg A, Riedel R, et al. A user-driven consent platform for health data sharing in digital health applications. NPJ Digit Med 2025;8(1):699 [FREE Full text] [doi: 10.1038/s41746-025-02147-3] [Medline: 41298895]
87. Regulation (EU) 2016/679 of the European parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). European Parliament and Council of the European Union. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj [accessed 2026-04-30]

XSL•FO
RenderX

88. Alderman JE, Palmer J, Laws E, McCradden MD, Ordish J, Ghassemi M, et al. Tackling algorithmic bias and promoting transparency in health datasets: the STANDING together consensus recommendations. Lancet Digit Health 2025;7(1):e64-e88 [FREE Full text] [doi: 10.1016/S2589-7500(24)00224-3] [Medline: 39701919]
89. Hatef E, Hudson Scholle S, Buckley B, Weiner J, Austin J. Development of an evidence- and consensus-based digital healthcare equity framework. JAMIA Open 2024;7(4):ooae136. [doi: 10.1093/jamiaopen/ooae136] [Medline: 39553827]
90. Babic B, Glenn Cohen I, Stern AD, Li Y, Ouellet M. A general framework for governing marketed AI/ML medical devices. NPJ Digit Med 2025;8(1):328 [FREE Full text] [doi: 10.1038/s41746-025-01717-9] [Medline: 40450160]
91. Liu M, Ning Y, Teixayavong S, Liu X, Mertens M, Shang Y, et al. A scoping review and evidence gap analysis of clinical AI fairness. NPJ Digit Med 2025;8(1):360 [FREE Full text] [doi: 10.1038/s41746-025-01667-2] [Medline: 40517148]
92. Templin T, Fort S, Padmanabham P, Seshadri P, Rimal R, Oliva J, et al. Framework for bias evaluation in large language models in healthcare settings. NPJ Digit Med 2025;8(1):414 [FREE Full text] [doi: 10.1038/s41746-025-01786-w] [Medline: 40624264]
93. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, DECIDE-AI expert group. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ 2022;377:e070904 [FREE Full text] [doi: 10.1136/bmj-2022-070904] [Medline: 35584845]
94. Health Canada, Medicines and Healthcare products Regulatory Agency. Good machine learning practice for medical device development: guiding principles. U.S. Food and Drug Administration. 2021. URL: https://www.fda.gov/media/153486/download [accessed 2026-01-15]
95. Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology. 2023. URL: https://doi.org/10.6028/NIST.AI.100-1 [accessed 2026-01-15]
96. What is health literacy? Centers for Disease Control and Prevention. 2024. URL: https://www.cdc.gov/health-literacy/php/about/index.html [accessed 2026-01-15]
97. Clinical decision support software: guidance for industry and food and drug administration staff. US Food and Drug Administration. 2026. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software [accessed 2026-02-10]
98. Bouguettaya A, Stuart EM, Aboujaoude E. Racial bias in AI-mediated psychiatric diagnosis and treatment: a qualitative comparison of four large language models. NPJ Digit Med 2025;8(1):332 [FREE Full text] [doi: 10.1038/s41746-025-01746-4] [Medline: 40467886]
99. Use of online tracking technologies by HIPAA covered entities and business associates. US Department of Health and Human Services. 2024. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/hipaa-online-tracking/index.html [accessed 2026-01-15]
100. Regulatory considerations on artificial intelligence for health. World Health Organization. Geneva, Switzerland: World Health Organization; 2023. URL: https://www.who.int/publications/i/item/9789240078871 [accessed 2026-03-10]
101. Rozenblit L, Price A, Solomonides A, Joseph AL, Koski E, Srivastava G, et al. Toward responsible AI governance: balancing multi-stakeholder perspectives on AI in healthcare. Int J Med Inform 2025;203:106015 [FREE Full text] [doi: 10.1016/j.ijmedinf.2025.106015] [Medline: 40680319]
102. AI system evaluations: independent evaluations. Developing AI accountability: inputs and a deeper dive (AI accountability policy report series). National Telecommunications and Information Administration. 2024. URL: https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report/developing-accountability-inputs-a-deeper-dive/ai-system-evaluations/independent-evaluations [accessed 2026-01-15]
103. Kiseleva A, Kotzinos D, De Hert P. Transparency of AI in healthcare as a multilayered system of accountabilities: between legal requirements and technical limitations. Front Artif Intell 2022;5:879603 [FREE Full text] [doi: 10.3389/frai.2022.879603] [Medline: 35707765]
104. Sullivan HR, Schweikart SJ. Are current tort liability doctrines adequate for addressing injury caused by AI? AMA J Ethics 2019;21(2):E160-E166 [FREE Full text] [doi: 10.1001/amajethics.2019.160] [Medline: 30794126]
105. Longpre S, Kapoor S, Klyman K, Ramaswami A, Bommasani R, Blili-Hamelin B. A safe harbor for AI evaluation and red teaming. arXiv. Preprint posted online on March 7, 2024 . [doi: 10.48550/arXiv.2403.04893]
106. Davis S, Dorn C, Park D, Matheny M. Emerging algorithmic bias: fairness drift as the next dimension of model maintenance and sustainability. J Am Med Inform Assoc 2025;32(5):845-854. [doi: 10.1093/jamia/ocaf039] [Medline: 40079820]
107. The importance of pharmacovigilance: safety monitoring of medicinal products. World Health Organization. Geneva: WHO; 2002. URL: https://iris.who.int/server/api/core/bitstreams/002b78d5-4dfc-433c-a518-f92ebdee8706/content [accessed 2026-01-15]
108. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions. US Food and Drug Administration. Silver Spring, MD: US Food and Drug Administration; 2024. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence [accessed 2026-03-10]
109. Council for International Organizations of Medical Sciences (CIOMS). Artificial Intelligence in Pharmacovigilance: CIOMS Working Group XIV Report. Geneva: CIOMS (Council for International Organizations of Medical Sciences); 2025.

XSL•FO
RenderX

110. Protecting the privacy and security of your health information when using your personal cell phone or tablet. US Department of Health and Human Services. 2022. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/cell-phone-hipaa/index.html [accessed 2026-01-15]
111. Le-Khac UN, Truong VNX. A survey on large language models unlearning: taxonomy, evaluations, and future directions. Artif Intell Rev 2025;58(12):399. [doi: 10.1007/s10462-025-11376-7]
112. 2025 national guidelines for a behavioral health coordinated system of crisis care. SAMHSA. 2025. URL: https://www.chcs.org/resource-center-item/2025-national-guidelines-for-a-behavioral-health-coordinated-system-of-crisis-care/ [accessed 2026-01-15]
113. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018;52(6):446-462 [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
114. Zhang S, Li X. Differential privacy medical data publishing method based on attribute correlation. Sci Rep 2022;12(1):15725 [FREE Full text] [doi: 10.1038/s41598-022-19544-3] [Medline: 36131115]
115. Garfinkel S, Guttman B, Near J, Dajani A, Singer P. De-Identifying Government Datasets: Techniques and Governance. Gaithersburg, MD: National Institute of Standards and Technology; 2023.
116. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature 2024;630(8017):625-630 [FREE Full text] [doi: 10.1038/s41586-024-07421-0] [Medline: 38898292]
117. Roustan D, Bastardot F. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. Interact J Med Res 2025;14:e59823 [FREE Full text] [doi: 10.2196/59823] [Medline: 39874574]
118. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst 2020;33:9459-9474.
119. Arowosegbe A, Oyelade T. Application of natural language processing (NLP) in detecting and preventing suicide ideation: a systematic review. Int J Environ Res Public Health 2023;20(2):1514 [FREE Full text] [doi: 10.3390/ijerph20021514] [Medline: 36674270]
120. Pais C, Liu J, Voigt R, Gupta V, Wade E, Bayati M. Large language models for preventing medication direction errors in online pharmacies. Nat Med 2024;30(6):1574-1582 [FREE Full text] [doi: 10.1038/s41591-024-02933-8] [Medline: 38664535]
121. National Library of Medicine. RxNorm. National Institutes of Health. URL: https://www.nlm.nih.gov/research/umls/rxnorm/index.html [accessed 2026-01-15]
122. FDA adverse event reporting system (FAERS). U.S. Food and Drug Administration. URL: https://www.fda.gov/drugs/drug-approvals-and-databases/fda-adverse-event-reporting-system-faers-database [accessed 2026-01-15]
123. Ong JCL, Jin L, Elangovan K, Lim GYS, Lim DYZ, Sng GGR, et al. Large language model as clinical decision support system augments medication safety in 16 clinical specialties. Cell Rep Med 2025;6(10):102323. [doi: 10.1016/j.xcrm.2025.102323] [Medline: 40997804]
124. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health 2024;6(1):e12-e22 [FREE Full text] [doi: 10.1016/S2589-7500(23)00225-X] [Medline: 38123252]
125. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366(6464):447-453 [FREE Full text] [doi: 10.1126/science.aax2342] [Medline: 31649194]
126. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? AMA J Ethics 2019;21(2):E167-E179 [FREE Full text] [doi: 10.1001/amajethics.2019.167] [Medline: 30794127]
127. Chin MH, Afsar-Manesh N, Bierman AS, Chang C, Colón-Rodríguez CJ, Dullabh P, et al. Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. JAMA Netw Open 2023;6(12):e2345050 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.45050] [Medline: 38100101]
128. Naderbagi A, Loblay V, Zahed IUM, Ekambareshwar M, Poulsen A, Song YJC, et al. Cultural and contextual adaptation of digital health interventions: narrative review. J Med Internet Res 2024;26:e55130 [FREE Full text] [doi: 10.2196/55130] [Medline: 38980719]
129. Web Content Accessibility Guidelines (WCAG) 2.2. World Wide Web Consortium (W3C). 2023. URL: https://www.w3.org/TR/WCAG22/ [accessed 2026-01-15]
130. Ayre J, Bonner C, Muscat DM, Cvejic E, Mac O, Mouwad D, et al. Online plain language tool and health information quality: a randomized clinical trial. JAMA Netw Open 2024;7(10):e2437955 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.37955] [Medline: 39378036]
131. Shi R, Petrakaki D. User-AI intimacy in digital health. Soc Sci Med 2026;389:118853 [FREE Full text] [doi: 10.1016/j.socscimed.2025.118853] [Medline: 41453777]
132. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digit Med 2021;4(1):4 [FREE Full text] [doi: 10.1038/s41746-020-00367-3] [Medline: 33402680]
133. Boucher EM, Harake NR, Ward HE, Stoeckl SE, Vargas J, Minkel J, et al. Artificially intelligent chatbots in digital mental health interventions: a review. Expert Rev Med Devices 2021;18(sup1):37-49 [FREE Full text] [doi: 10.1080/17434440.2021.2013200] [Medline: 34872429]

XSL•FO
RenderX

134. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. U.S. Food and Drug Administration. Silver Spring, MD: FDA URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices [accessed 2026-01-20]
135. Fossouo Tagne J, Yakob RA, Mcdonald R, Wickramasinghe N. Linking activity theory within user-centered design: novel framework to inform design and evaluation of adverse drug reaction reporting systems in pharmacy. JMIR Hum Factors 2023;10:e43529 [FREE Full text] [doi: 10.2196/43529] [Medline: 36826985]
136. Huang SC, Jensen M, Yeung-Levy S, Lungren MP, Poon H, Chaudhari AS. A systematic review and implementation guidelines of multimodal foundation models in medical imaging. Res Sq Preprint 2025. [doi: 10.21203/rs.3.rs-5537908/v1]
137. Chung R, Lee J, Hackell J, Alderman EM, COMMITTEE ON ADOLESCENCE. Confidentiality in the care of adolescents: policy statement. Pediatrics 2024;153(5):e2024066326. [doi: 10.1542/peds.2024-066326] [Medline: 38646690]
138. OECD. Artificial intelligence and competitive dynamics in downstream markets. OECD Roundtables on Competition Policy Papers No. 331. Paris, France: OECD Publishing; 2025. URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/11/artificial-intelligence-and-competitive-dynamics-in-downstream-markets_c6e81d0e/ccf0624a-en.pdf [accessed 2026-03-10]
139. OECD. Regulatory sandboxes in artificial intelligence. OECD Digital Economy Papers No. 356. Paris, France: OECD Publishing; 2023. URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/07/regulatory-sandboxes-in-artificial-intelligence_a44aae4f/8f80a0e6-en.pdf [accessed 2026-03-10]

Abbreviations

AI: artificial intelligence
LLM: large language model
PH-LLM: personal health large language model

Edited by A Coristine; submitted 27.Jan.2026; peer-reviewed by L Mauri, P Lacroix, A El Ammari; comments to author 26.Feb.2026; revised version received 27.May.2026; accepted 29.May.2026; published 17.Jun.2026.

Please cite as:
Liu J, Liu S
Ethical Considerations in Personal Health Large Language Models
J Med Internet Res 2026;28:e92240
URL: https://www.jmir.org/2026/1/e92240
doi:10.2196/92240
PMID:

©Jialin Liu, Siru Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Integration of Federated Learning and Blockchain in Health Care: Tutorial on Medical Data, Architectures, Privacy, Security, and Regulatory Compliance

Yahya Shahsavari[1], PhD; Yaser Baseri[1], PhD; Abdelhakim Hafid[1], PhD; Oussama Abderrahmane Dambri[2], PhD; Dimitrios Makrakis[2], PhD

[1]Department of Computer Science and Operations Research, Université de Montréal, 2900 Bd Édouard-Montpetit, Montréal, QC, Canada
[2]School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON, Canada

Corresponding Author:
Yahya Shahsavari, PhD
Department of Computer Science and Operations Research, Université de Montréal, 2900 Bd Édouard-Montpetit, Montréal, QC, Canada

Abstract

The convergence of artificial intelligence (AI), blockchain technology, and health care represents one of the most transformative yet technically challenging frontiers in computational medicine. As health care systems adopt data-driven paradigms for precision medicine and clinical decision support, the need for secure, privacy-preserving, and collaborative learning frameworks has become critical. This tutorial introduces a comprehensive, clinically oriented, and compliance-aware framework integrating federated learning (FL) and blockchain for secure and privacy-preserving health care analytics. FL enables collaborative training across distributed institutions without raw data sharing, in alignment with privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR). However, FL remains vulnerable to model poisoning and gradient leakage. To address these risks, we introduce blockchain-based FL (BCFL), which leverages blockchain's immutable ledger and decentralized consensus to enhance trust, verifiability, and auditability. The tutorial's main contributions include (1) a taxonomy of diverse medical data types and their FL requirements; (2) three integration architectures (fully coupled, semicoupled, and loosely coupled) analyzed for security, scalability, and regulatory compliance; (3) a security analysis of health care–specific vulnerabilities and mitigation strategies using advanced cryptography, such as zero-knowledge proofs, homomorphic encryption, and differential privacy; and (4) a regulatory compliance framework addressing HIPAA, GDPR, and United States Food and Drug Administration guidelines for AI-enabled medical devices. We demonstrate BCFL's relevance across major health care applications, including disease prediction, medical imaging, patient monitoring, and drug discovery, and highlight emerging research directions such as quantum-resilient cryptography, scalable interoperability, and automated compliance. This tutorial serves as a foundational resource for advancing secure, compliant, and collaborative AI in health care; fostering privacy-preserving analytics; and improving patient outcomes.

(J Med Internet Res 2026;28:e80178) doi:10.2196/80178

KEYWORDS

blockchain technology; health care security; medical data privacy; IoT health care; machine learning; HIPAA compliance

Introduction

Background

The convergence of artificial intelligence (AI), blockchain technology, and health care represents one of the most transformative yet technically challenging frontiers in computational medicine. As health care systems worldwide transition toward data-driven paradigms for precision medicine, clinical decision support, and population health management, the exponential growth in medical data generation—from electronic health records (EHRs) and medical imaging to genomic sequencing and Internet of Things (IoT)-enabled continuous monitoring—presents unprecedented opportunities for advancing health care through machine learning (ML) [1-3].

Federated learning (FL) has emerged as a promising paradigm for collaborative model training across distributed health care institutions while maintaining data locality [4,5]. However, standard FL architectures face challenges, including model poisoning, gradient leakage, and coordination failures [6-8]. Blockchain technology offers complementary capabilities through immutable audit trails, decentralized trust, and automated smart contract execution [9], yet integration with FL in health care introduces unique scalability, efficiency, and regulatory compliance challenges.

XSL•FO
RenderX

This tutorial addresses a critical gap by providing a structured, cross-disciplinary framework that integrates blockchain with FL in the context of health care, with explicit coverage of medical data taxonomies, clinical workflows, and regulatory compliance.

While prior surveys have explored the integration of FL and blockchain applications in health care and beyond [10-13], this tutorial differs in scope and emphasis. In particular, we provide (1) a systematic medical data taxonomy that classifies diverse health care data types (EHRs, imaging, genomics, biometrics, Internet of Medical Things [IoMT], clinical trials, etc) and analyzes their specific FL requirements; (2) a set of three integration architectures (fully coupled, semicoupled, and loosely coupled) with detailed evaluation of security, scalability, and compliance tradeoffs tailored to health care deployments; (3) comprehensive treatment of advanced cryptographic defenses (zero-knowledge proofs [ZKPs], homomorphic encryption [HE], and differential privacy [DP]) with health care–specific vulnerability analysis; and (4) a structured regulatory compliance framework mapping Health Insurance Portability and Accountability Act (HIPAA), General Data Protection Regulation (GDPR), and United States Food and Drug Administration (FDA) guidelines to blockchain-FL system design. Moreover, while prior surveys have typically focused on high-level concepts, our tutorial adopts a more comprehensive and tutorial-oriented scope by combining blockchain fundamentals, FL methodologies, health care–specific applications, security analysis, and forward-looking research directions. This extensive focus allows us to introduce a comprehensive, clinically oriented, and compliance-aware framework that integrates FL and blockchain for health care, providing actionable insights for both researchers and practitioners implementing secure, privacy-preserving, and regulation-compliant collaborative analytics systems. Table 1 contrasts this tutorial with earlier blockchain-based FL (BCFL) surveys and frameworks to clarify the contributions.

Table . Comparison of our tutorial with prior work on the integration of blockchain and federated learning in health care.

Reference	Medical data taxonomy	Integration architecture	Cryptography and privacy	Regulatory compliance
Abbas et al [10]	General overview	No specific architecture	Privacy mechanisms	Limited overview
Cheng et al [11]	Partial coverage	No specific architecture	Basic techniques	GDPR[a]
Ngoupayou Limbepe et al [12]	Limited scope	No specific architecture	Basic techniques	HIPAA[b] and GDPR
Myrzashova et al [13]	Not included	No specific architecture	Security analysis	Not addressed
This paper	Comprehensive taxonomy	Three distinct architectures	Security and privacy analysis	HIPAA, GDPR, and FDA[c]

[a]GDPR: General Data Protection Regulation.

[b]HIPAA: Health Insurance Portability and Accountability Act.

[c]FDA: Food and Drug Administration.

Motivation and Problem Statement

Contemporary health care faces a fundamental paradox: the increasing need for large-scale, collaborative data analytics to advance medical knowledge is juxtaposed with stringent privacy regulations and the inherent sensitivity of medical data. Traditional centralized approaches to health care ML create critical vulnerabilities where patient information aggregated in single repositories becomes an attractive target for sophisticated attacks [14,15]. Recent breaches affecting millions of patient records underscore the inadequacy of current security models, particularly when regulatory frameworks, such as HIPAA and GDPR, and emerging national data protection laws impose severe penalties for data mishandling [16].

While FL addresses data locality concerns by enabling distributed model training, it introduces health care–critical vulnerabilities. Model poisoning attacks could compromise diagnostic accuracy [6,7], potentially leading to misdiagnosis or incorrect treatment recommendations. Gradient leakage exploits can enable patient reidentification from shared model updates [8,17], violating privacy guarantees that health care institutions must maintain. Coordination failures in distributed learning processes could disrupt critical clinical workflows [18], affecting real-time decision support systems.

Blockchain technology addresses these vulnerabilities through immutable audit trails that ensure accountability, decentralized trust establishment that eliminates single points of failure, automated smart contract execution for verifiable protocol compliance, and comprehensive data provenance tracking [9]. However, health care–specific integration challenges remain, including scalability constraints when processing high-frequency medical data updates from IoMT devices, energy efficiency requirements for sustainable long-term deployment in resource-constrained clinical environments, compliance with medical device regulations for AI/ML systems, and real-time performance requirements for clinical decision support applications.

Existing BCFL approaches [19-23] are predominantly designed for generic data types and fail to address health care–specific requirements. Critical gaps include a lack of regulatory compliance frameworks for HIPAA and GDPR, insufficient integration with clinical workflows and existing health IT systems, inadequate support for multimodal medical data with heterogeneous privacy requirements, and absence of explainability mechanisms required for medical decision-making. This tutorial fills these gaps by providing a comprehensive framework tailored specifically for health care applications.

XSL•FO
RenderX

Contributions and Novelty

This tutorial advances the state-of-the-art at the intersection of blockchain technology, FL, and health care informatics through several methodological and theoretical contributions:

1. Comprehensive medical data taxonomy for FL: We develop a systematic classification of medical data types (EHRs, imaging, genomics, biometrics, IoT sensors, clinical trials, etc) and their specific requirements for FL implementations, including privacy sensitivity levels, regulatory considerations, and technical constraints for distributed processing.
2. Novel integration architecture framework: We introduce and formally define 3 distinct BCFL integration architectures (fully coupled, semicoupled, and loosely coupled) specifically designed for health care environments, with detailed analysis of their security guarantees, scalability characteristics, and regulatory compliance capabilities.
3. Health care–specific security and privacy analysis: We provide a comprehensive security analysis of BCFL systems in health care contexts, identifying unique vulnerabilities related to medical data sensitivity, regulatory requirements, and clinical workflow dependencies, along with corresponding mitigation strategies.
4. Regulatory compliance framework: We systematically examine how BCFL architectures can satisfy complex health care regulations (HIPAA, GDPR, and FDA guidelines for AI/ML-based medical devices), providing practical guidance for implementation in real-world health care settings.
5. Research roadmap and future directions: We identify critical research gaps at the intersection of blockchain, FL, and health care, proposing specific research directions that could accelerate adoption and address current technical limitations.

Tutorial Scope and Target Audience

This tutorial is designed for researchers and practitioners at the intersection of health care informatics, distributed systems, and ML. Our target audience includes (1) health care informaticians seeking to implement privacy-preserving collaborative analytics, (2) computer scientists developing secure FL systems, (3) blockchain researchers exploring health care applications, (4) clinical researchers interested in multi-institutional collaborations, and (5) regulatory and policy experts working on health care data governance. The tutorial assumes familiarity with basic ML concepts but provides a comprehensive background on FL and blockchain technologies. We emphasize practical implementation considerations while maintaining rigorous technical depth appropriate for publication in top-tier venues.

Organization

The remainder of this tutorial is structured as illustrated in Figure 1. The Medical Data and Applications in ML section introduces a comprehensive taxonomy of medical data types, including EHRs, imaging, genomics, and sensor data, and discusses their relevance to ML. The AI and FL in Health Care section explores the evolution of AI in health care, highlighting FL paradigms, their challenges, and future trends. The Blockchain section reviews fundamental blockchain concepts, such as consensus, data structures, and cross-chain communication, with a focus on health care applicability. The Integration of FL With Blockchain section presents integration architectures that combine FL and blockchain, along with their security, privacy, and regulatory implications in health care. The Empirical Validation and Performance Evaluation section synthesizes quantitative evidence from clinical deployments, benchmarks, and adversarial robustness studies to provide empirical validation, assess performance tradeoffs, and establish the framework's real-world feasibility. The Related Work section offers a critical review of previous work on the integration of blockchain and FL to contextualize the contributions of this tutorial. The Future Research Directions section synthesizes emerging research challenges and opportunities, grouped under cryptographic resilience, infrastructure scalability, health care–specific consensus and incentives, and regulatory compliance in system integration. Finally, the Conclusion section summarizes the key findings and reflects on the real-world translation of BCFL systems into clinical practice.

XSL•FO
RenderX

Figure 1. Organizational structure of the tutorial.

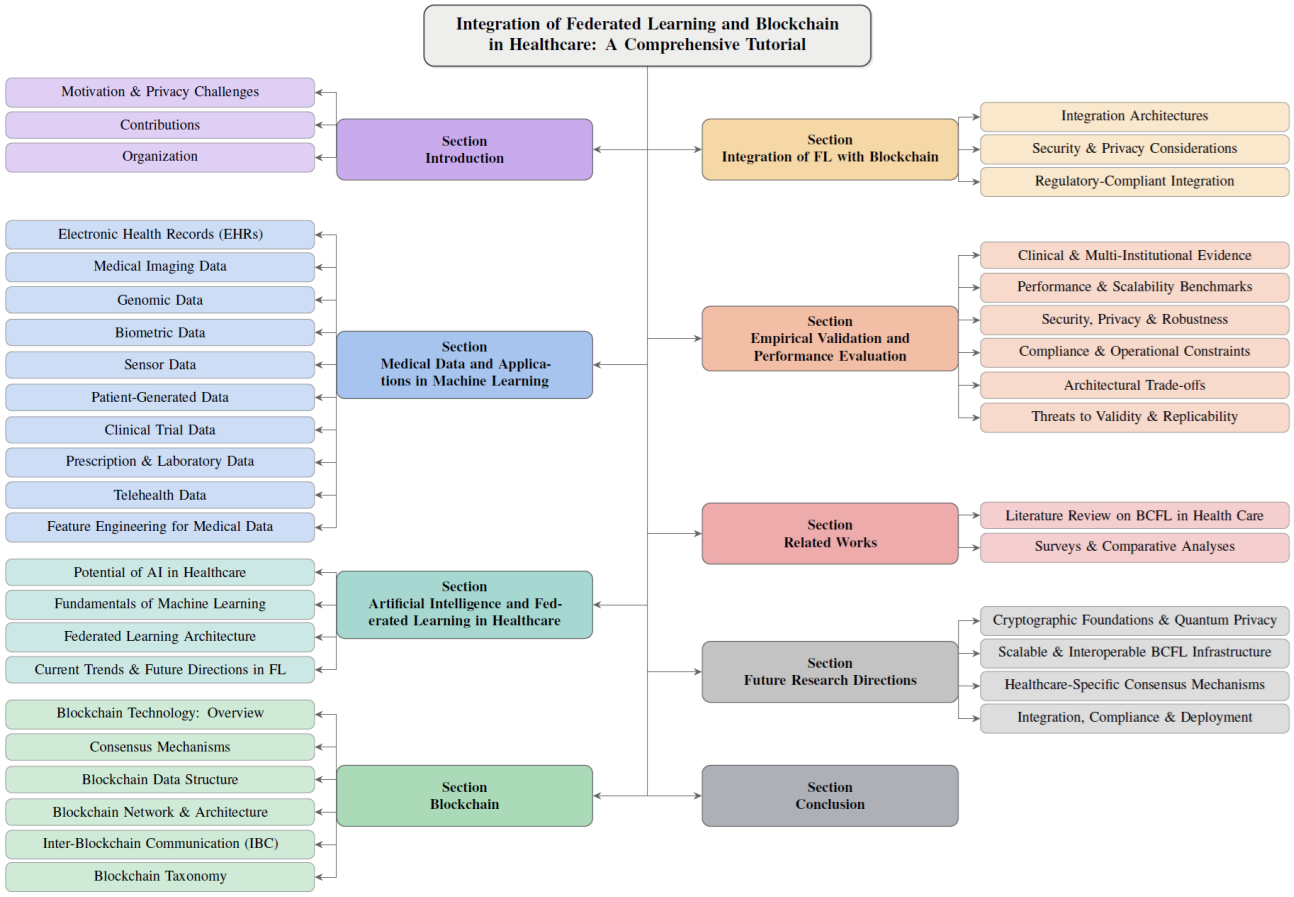

Medical Data and Applications in ML

Overview

Medical data encompass a diverse array of information crucial for understanding, diagnosing, and treating various health conditions. These data, ranging from patient demographics and medical history to diagnostic images and genomic sequences, hold immense potential for advancing health care through ML applications. By harnessing the power of ML algorithms, medical data can be analyzed to extract valuable insights, predict patient outcomes, personalize treatments, and optimize health care delivery. However, the utilization of medical data for ML requires careful consideration of data storage and management practices to ensure compliance with privacy regulations, maintain data integrity, and facilitate seamless access for research and clinical purposes. This section explores the different types of medical data and their applications in ML.

EHR Data

EHRs are digital patient charts containing medical history, diagnoses, medications, treatment plans, immunization dates, allergy information, medical images, and lab results, providing a comprehensive, up-to-date view for informed decision-making [24,25]. The key EHR features include the following: (1) *digital format*, electronic records replace paper for easier storage, access, and sharing; (2) *interoperability*, designed for sharing among different providers and organizations, facilitating better care coordination; (3) *real-time access*, authorized professionals get quick access to critical information, which is crucial for emergencies; (4) *patient engagement*, features allow patients to access data, schedule appointments, and communicate with providers; (5) *decision support*, tools offer alerts for drug interactions, screening reminders, and clinical guidelines; and (6) *data security and privacy*, security measures protect confidentiality, with access restricted to authorized personnel [26]. EHRs are a rich source of structured and unstructured data for ML applications. ML algorithms analyze this comprehensive patient information to extract insights, predict outcomes, identify patterns, and improve clinical decision-making [27,28]. In ML, EHR data are used for tasks, including the following: (1) *predictive analytics*, where models forecast medical events (eg, hospital readmission, disease onset, and mortality) for proactive, personalized care [29,30]; (2) *disease identification and diagnosis*, where algorithms assist in early detection by identifying subtle patterns and anomalies, aiding accurate and timely diagnoses [31,32]; and (3) *treatment recommendations*, where models suggest personalized plans based on history, demographics, genetics, and past responses to optimize outcomes [33].

Medical Imaging Data

Medical imaging data encompass diverse visual representations of internal body structures, acquired via distinct imaging modalities [34,35] for clinical scrutiny, diagnosis, and ongoing assessment. The key components of medical imaging data are (1) patient information, (2) imaging modalities, (3) image files, and (4) reports. First, patient information includes basic demographic and identifying details (eg, name, ID, age, and gender). Second, among imaging modalities, each modality

XSL•FO
RenderX

uses specific physical principles [36] to generate images adapted to clinical needs. The imaging modalities include (1) radiography for projection imaging of osseous structures [37]; (2) computed tomography (CT) for high-resolution cross-sectional imaging using x-ray and computational reconstruction [38]; (3) magnetic resonance imaging (MRI) for detailed soft tissue visualization via magnetic fields and radiofrequency pulses [39]; (4) ultrasonography for real-time acoustic wave imaging, which is ideal for soft tissues and dynamic processes [40]; (5) nuclear medicine imaging for visualizing organ function via radiopharmaceuticals [41]; (6) positron emission tomography for quantitative functional imaging, especially in oncology, neurology, and cardiology [42]; and (7) fluoroscopic imaging for dynamic, real-time x-ray visualization during interventional and functional studies [43]. Third, regarding image files, medical data, including metadata (patient information, acquisition parameters, and image details) [44], are stored, shared, and transmitted according to technical standards [45], notably Digital Imaging and Communications in Medicine (DICOM) [46]. Fourth, reports include narrative documents generated by radiologists summarizing image analysis findings for the treatment planning of referring physicians. Medical imaging data have a wide range of ML applications in health care [47], including the following: (1) *disease diagnosis and classification*, ML algorithms assist in the diagnosis and classification of diseases like cancer and neurological, cardiovascular, and musculoskeletal disorders [48]; (2) *computer-aided detection*, computer-aided detection systems use ML to help radiologists detect abnormalities (eg, tumors, lesions, and fractures), improving diagnostic accuracy and efficiency [49]; (3) *image-based biomarker discovery*, ML identifies imaging biomarkers associated with diseases or treatment responses, which is valuable for prognosis, efficacy assessment, and personalized medicine [50,51]; (4) *treatment planning and monitoring*, imaging data are used to develop personalized plans, and ML predicts outcomes, monitors progression, and optimizes treatment strategies [52]; (5) *image reconstruction and enhancement*, ML techniques (eg, deep learning) improve image quality from various modalities (MRI, CT, and ultrasonography) by reducing artifacts and enhancing resolution for better interpretation [53]; (6) *image registration and fusion*, ML algorithms automatically align and combine multiple images (from different modalities or time points) for comprehensive visualization and analysis; and (7) *drug discovery and development*, ML analyzes imaging data to evaluate drug effects on disease progression, identify potential drug targets, and optimize delivery methods [54].

Genomic Data

Genomic data pertain to information concerning the organization and operation of the genome within an organism [55,56]. In health care ML, this refers to comprehensive information from an individual's genome (entire DNA/genetic composition), offering profound insights into genetic predispositions, treatment responses, and overall health [57]. The key components for medical use cases include the following: (1) *genome sequences*, the complete set of genetic material (DNA), consisting of nucleotides (A, T, C, and G) that encode genetic information; (2) *genes*, specific DNA sequences that encode instructions for protein synthesis, including their location, structure, and function; (3) *single nucleotide polymorphisms*, single-nucleotide variations that can influence traits, disease susceptibility, and drug responses, along with their associated phenotypic and disease information [58]; (4) *copy number variations*, genomic alterations involving changes in the number of copies of DNA segments, affecting gene dosage and leading to phenotypic variations and disease susceptibility [59]; (5) *gene expression profiles*, information on the activity levels of different genes (often generated via microarrays or RNA sequencing [60]), providing insights into protein synthesis; (6) *epigenetic modifications*, heritable changes in gene expression (eg, DNA methylation and histone modifications) that do not alter the DNA sequence but influence gene activity and phenotype [61]; and (7) *genetic variation databases*, compilations of genomic data, genetic variants, annotations, and associated phenotypic information from various sources (eg, population studies and disease databases) [62]. These components are essential for understanding the genetic basis of diseases, identifying risk factors, and developing personalized health care approaches. The utilization of genomic data in health care ML includes the following: (1) *disease prediction and risk stratification*, ML algorithms scrutinize genomic data to discern patterns and variations linked to specific diseases, assessing associated health risks [63]; (2) *personalized medicine*, genomic data form the basis for personalized treatment plans [64,65], and ML models predict individual medication responses for tailored strategies [66-68]; (3) *drug discovery and computational genomics*, ML analyzes genomic data to expedite discovery, identifying drug targets and comprehending genetic underpinnings for more efficacious therapeutic solutions [69]; (4) *genomic counseling*, ML algorithms decipher complex genomic data, assisting health care professionals in communicating intricate genetic details, risks, and familial implications [70]; (5) *early detection and diagnostics*, integrating genomic data and ML algorithms facilitates early detection by discerning subtle genetic variations indicative of specific conditions, enabling timely intervention [71,72]; (6) *research and population health informatics*, aggregated genomic data subjected to ML can advance the understanding of disease genetic foundations at the population level, informing public health initiatives and epidemiological studies [73]; and (7) *genomic sequencing and computational analysis*, ML plays a crucial role in interpreting extensive genomic datasets from sequencing technologies, identifying genetic mutations, variations, and other health-impacting information [74].

Biometric Data

Biometric data in the medical context refer to unique physical or behavioral characteristics used for individual identification and monitoring. These data are crucial in health care for patient identification, medical record access control, personalized treatments, and health condition monitoring. The components of biometric data include the following: (1) *physiological biometrics*, data types like fingerprints, facial recognition, iris/retina scans, DNA, smell recognition, and hand geometry data [75]; (2) *behavioral biometrics*, data types such as voice recognition, gait analysis, and typing dynamics data [75]; (3) *health-related biometrics*, data types such as heart rate, blood

XSL•FO
RenderX

pressure, blood oxygen levels, electrocardiography (ECG) patterns [76], and brain wave patterns (electroencephalography [EEG] patterns) [77]; and (4) *analytical biometric technologies*, advanced modalities like microbial biometrics, which analyze unique microbiome compositions for identification and health assessment [78], and olfactory biometrics, which utilize distinctive body odor profiles for identification and disease detection [79].

Biometric data integrated into health care ML encompass quantifiable physiological or behavioral attributes [75,80] for accurate identification [81,82], secure access [83,84], and health assessment [85]. The primary applications of biometric data in health care ML include the following: (1) *biometric identification and access control*, biometric features (eg, fingerprints and facial characteristics) are pivotal for precise identification and heightened security, governing access to sensitive areas, EHRs, and medical devices [80,86]; (2) *patient identification and record matching*, biometric identifiers (eg, fingerprints and iris scans) ensure precise linkage of patients to their health records, minimizing errors, with ML algorithms enhancing matching accuracy and elevating care quality [87-90]; (3) *biometric monitoring for health assessment*, continuous monitoring of biometric data (eg, heart rate and ECG data) via wearables/sensors facilitates real-time health assessment, and ML analyzes dynamic data for the early detection of anomalies, supporting timely intervention and personalized management [91-93]; (4) *behavioral biometrics for mental health monitoring*, behavioral biometrics (eg, typing and voice modulation) contribute to assessing mental health and detecting behavioral changes [94,95], and ML models discern patterns indicative of conditions, aiding targeted interventions [96]; (5) *biometric data in clinical trials*, biometric data are used for participant identification, monitoring, and data integrity, and ML assists in efficient management and analysis, ensuring study validity [97,98]; (6) *voice and speech analysis for diagnostics*, ML algorithms process voice patterns to detect potential markers for conditions like Parkinson disease [99,100] or respiratory disorders [101,102], contributing to diagnostic capabilities [103]; and (7) *facial recognition for patient monitoring*, facial recognition is used for monitoring patient well-being and distress, and ML analyzes facial expressions, offering insights into comfort levels and enhancing care quality [104,105].

Sensor Data

Sensor data in medical contexts represent information collected from various devices (wearable, implantable, or environmental devices) to monitor health, detect condition changes, and assist diagnoses. These data continuously track physiological parameters, activity levels, and environmental conditions. The components of medical sensor data include the following: (1) *wearable sensors*, data from devices like heart rate monitors, blood pressure monitors, blood glucose monitors, activity trackers, and pulse oximeters; (2) *implantable sensors*, data from devices like cardiac monitors, glucose monitors, and intraocular pressure sensors; (3) *environmental sensors*, data gathered from air quality monitors (for pollutants and allergens affecting respiratory conditions) and temperature/humidity sensors; (4) *specialized medical sensors*, data from ECG, EEG, electromyography, gait sensors, and sleep monitors; and (5) *smart health homes*, data extracted from fall detectors and bed sensors that monitor sleep patterns, bed occupancy, and vital signs.

Sensor data used in health care ML are collected from various sources (wearables, medical equipment, and monitoring tools) [106,107], providing real-time health and activity insights. ML algorithms analyze these data to make predictions, identify patterns, and offer personalized insights. The utilization of sensor data in health care ML includes the following: (1) vital sign monitoring, (2) wearable devices for activity tracking, (3) blood glucose monitoring and continuous glucose monitoring, (4) ECG monitoring, (5) sleep monitoring, (6) environmental sensors, (7) medication adherence monitoring, (8) fall detection and activity recognition, and (9) biometric sensors for stress and emotion monitoring. First, in vital sign monitoring, ML analyzes continuous vital sign data (heart rate, blood pressure, etc) to detect anomalies, predict health deteriorations, and offer early warnings [108,109]. Second, regarding wearable devices for activity tracking, ML models analyze data from accelerometers and gyroscopes to assess physical health, detect abnormalities, and provide personalized insights for fitness and rehabilitation [110]. Third, in blood glucose monitoring and continuous glucose monitoring, ML algorithms analyze continuous glucose monitoring data to predict glucose level trends, recommend insulin dosages, and enhance diabetes management [111]. Fourth, in ECG monitoring, ML interprets ECG data to identify cardiac abnormalities, predict cardiovascular risk, and recommend early interventions [112,113]. Fifth, in sleep monitoring, ML algorithms analyze sleep data (from wearables and bed sensors) to identify disorders, provide insights into sleep hygiene, and recommend personalized interventions [114,115]. Sixth, regarding environmental sensors, ML correlates environmental data (air quality and temperature) with health outcomes, aiding in identifying triggers for respiratory conditions or allergies [116]. Seventh, in medication adherence monitoring, ML analyzes adherence patterns tracked by smart sensors, sending reminders and providing providers with compliance insights [117,118]. Eighth, regarding fall detection and activity recognition, ML models use motion sensor data for fall risk assessment, accident prediction, and adapting care plans for individuals with mobility challenges [119]. Ninth, regarding biometric sensors for stress and emotion monitoring, ML analyzes biometric signals (eg, skin conductance and heart rate variability) to provide insights into mental health, stress management, and emotional well-being [120,121].

Patient-Generated Data

Patient-generated data (PGD) in health care ML refer to health-related information actively contributed by patients [122,123], distinct from traditional clinical records. Sourced directly via wearables, mobile apps, and patient-reported outcomes (PROs), PGD foster a patient-centric, data-driven approach, enabling personalized interventions, early detection, and improved patient-provider communication [124]. Beyond sensor data, PGD utilization in health care ML includes the following: (1) *mobile health apps and surveys*, patients input health information and feedback via apps, and ML processes these data to derive insights into treatment effectiveness,

XSL•FO
RenderX

medication adherence, and satisfaction, informing personalized care plans [125]; (2) *social media and online communities*, patients share health experiences and concerns online [126], and ML conducts analyses on social media data for health-related trends, sentiment, and public health monitoring, contributing to population health research and patient perspectives [127,128]; (3) *genomic and genetic data sharing*, patients voluntarily contribute genetic information for research, and ML analyzes aggregated genomic data to identify genetic factors associated with diseases, fostering precision medicine advancements [129]; and (4) *telehealth and virtual visits*, patients offer health updates during virtual consultations, and ML algorithms analyze PGD from these visits to support clinical decision-making, monitor treatment progress, and enhance virtual health care quality [130].

Clinical Trial Data

Clinical trial data refer to information systematically collected during trials designed to evaluate the safety, efficacy, or effectiveness of new medical interventions (drugs, devices, and procedures) in human participants [131]. These data, governed by a structured protocol, inform medical decision-making, regulatory approvals, and medical knowledge advancements. Integrating these data into ML models enables the development of predictive algorithms, risk assessment tools, and decision support systems, contributing to evidence-based medicine, personalized treatments, and enhanced clinical research efficiency [132]. The key components of clinical trial data are as follows: (1) demographic information, (2) informed consent, (3) medical history, (4) intervention details, (5) clinical assessments, (6) adverse events, (7) efficacy endpoints, (8) follow-up data, and (9) protocol deviations. First, demographic information includes details on study participants (age, gender, race, and ethnicity) incorporated into ML models to assess intervention response across groups and develop personalized treatment plans. Second, informed consent involves documentation confirming that participants were informed of risks/benefits and voluntarily agreed to participate, ensuring ethical standards and regulatory compliance, with ML incorporating this to restrict the analysis to consented data [133,134]. Third, medical history includes information on pre-existing conditions, relevant history, and concurrent medications, which is used to identify comorbidities that may impact treatment outcomes. Fourth, intervention details include specifics about the product/procedure (dosage, administration, and protocol), with ML analyzing these details to identify patterns associated with treatment success or failure, predicting efficacy in future cases [135]. Fifth, clinical assessments include physical examinations, lab tests, imaging, and measurements to assess the health status and intervention response, with ML processing this information to identify trends, correlations, or anomalies indicative of treatment responses or adverse events, aiding early detection and prediction [136]. Sixth, adverse events involve records of side effects experienced, including severity and relation to the intervention, with ML learning from these historical data to predict the likelihood of adverse events for new interventions, supporting risk assessment and proactive management [137,138]. Seventh, efficacy endpoints include measurements (eg, symptom relief and disease marker improvement) used to determine intervention effectiveness, with ML analyzing these data to develop predictive models for treatment success/failure and identify key contributing factors. Eighth, follow-up data include information on long-term outcomes, adherence, and sustained effects collected during posttrial visits, which is essential for longitudinal analyses, with ML using this information to predict the long-term effects of interventions, including sustained efficacy or potential relapse [139]. Ninth, protocol deviations involve documentation of deviations from the original plan and their reasons, with ML analyzing the data to identify how deviations impact study outcomes, allowing for adjustment or prediction of their potential effects [140].

Prescription and Medication Data

Prescription and medication data in medical records relate to prescribed drugs, dosage, frequency, and related details. These data are crucial for patient safety, medication management, monitoring public health and treatment efficacy, research, regulatory compliance, and facilitating communication. Integrated into EHRs, they provide a comprehensive medication history, allowing health care professionals to make informed decisions and avoid potential adverse events. Aggregated and anonymized data are used in research to broadly assess medication safety and effectiveness. The key components of prescription and medication data are as follows: (1) patient information; (2) prescriber information; (3) prescription date; (4) medication name; (5) dosage, frequency, and route of administration; (6) duration of treatment; (7) instructions for use; (8) refill information; (9) allergies and contraindications; (10) adverse reactions; (11) medication changes; (12) medication discontinuation; and (13) medication administration records (MARs). First, patient information includes identification details (name, date of birth, and demographics), which are used in ML for personalized treatment. Second, prescriber information includes details about the prescribing provider, with ML models being trained to recognize prescribing practices associated with positive patient outcomes [141]. Third, prescription date is the date the prescription was issued, and temporal analysis assists in predicting medication adherence and treatment outcomes, with ML identifying patterns related to issuance timing and patient behavior [142]. Fourth, medication name involves generic and brand names, with ML models categorizing medications by therapeutic class, aiding in identifying commonalities in treatment outcomes [143]. Fifth, dosage, frequency, and route of administration include the prescribed amount/strength, how often it should be taken, and the administration method, which are crucial details for predicting adherence and adverse events, with ML models identifying optimal regimens [144]. Sixth, duration of treatment is the prescribed period, with ML models analyzing this to predict long-term outcomes, including treatment success and potential drug resistance [145]. Seventh, instructions for use include additional guidance (eg, taken with food and specific times), with natural language processing (NLP) approaches extracting insights from free-text instructions to identify nuances in patient guidance [146]. Eighth, refill information includes details on authorized refills, which are essential for predicting adherence and persistence, with ML identifying factors influencing refill

XSL•FO
RenderX

behavior and predicting discontinuation likelihood [147]. Ninth, allergies and contraindications involve records of known patient allergies or contraindications, with ML identifying associations among these records, adverse drug reactions, and specific medications to predict safety and suitability for individual patients [148]. Tenth, adverse reactions involve documentation of experienced side effects, with data being analyzed to develop predictive models identifying patients at higher risk for specific side effects, enabling proactive management. Eleventh, medication changes include information about dosage adjustments or switches, with ML analyzing historical changes to predict future treatment modifications, assisting in personalized planning [129,149]. Twelfth, medication discontinuation involves the reason and date of stopping medication, with ML models predicting factors contributing to discontinuation, helping providers intervene to improve adherence [150]. Thirteenth, MARs include records of actual medication administration in a health care setting, with MAR data being used to train ML models for predicting administration patterns and identifying deviations from the prescribed regimen [151].

Laboratory Data

Laboratory data in medical records are the results of tests and analyses conducted on patient samples, which are essential for diagnosing, monitoring, and managing various medical conditions. The types of data collected vary by patient presentation and provider assessment. ML techniques (supervised, unsupervised, and deep learning) are applied depending on the nature of the data and analysis goals. The common types of laboratory data are the results of the following tests: (1) blood tests, (2) urinalysis, (3) microbiology tests, (4) pathology tests, (5) hematology tests, (6) immunology tests, (7) endocrine tests, (8) serology tests, (9) genetic tests, and (10) radiology and imaging studies. First, blood tests include assessments of the complete blood count (cell counts), blood chemistry panel (electrolytes, glucose, and organ function tests), and lipid profile, with ML models identifying patterns in complete blood count results associated with specific diseases (eg, anemia and infections) [152]. Second, urinalysis involves examination of the physical/chemical properties of urine, with ML algorithms detecting patterns indicative of kidney disorders, urinary tract infections, or diabetes [153]. Third, microbiology tests involve identifying microorganisms and their antibiotic sensitivity, with ML assisting in microorganism identification from culture data and predicting antibiotic susceptibility for personalized treatment [154,155]. Fourth, pathology tests include tissue biopsy and cytology (cell examination) performed for disease/abnormality diagnosis (eg, cancer), with image recognition ML models trained on pathology slides assisting in identifying abnormal tissue or cancerous cells [50,156]. Fifth, hematology tests include coagulation studies (blood clotting) and erythrocyte sedimentation rate (an inflammation measure), with ML models analyzing coagulation data to predict the risk of bleeding or clotting disorders [157]. Sixth, immunology tests include antibody tests (detecting immune system antibodies) and viral load (measuring the virus in blood), with ML algorithms identifying patterns in antibody levels to diagnose autoimmune or infectious conditions [158,159]. Seventh, endocrine tests include assessments of hormone levels (eg, thyroid and insulin), with ML models analyzing these levels to predict and monitor endocrine disorders like thyroid dysfunction or diabetes [160,161]. Eighth, serology tests include analyses of serum components (proteins, enzymes, and electrolytes), with ML identifying markers associated with specific diseases, aiding early detection and monitoring [158]. Ninth, genetic tests involve the identification of specific genetic markers for condition diagnosis, with ML analyzing these data to identify disease risk, predict treatment response, or diagnose genetic disorders [70]. Tenth, radiology and imaging studies include diagnostic tests contributing to medical data, with image recognition/segmentation ML models being trained on radiology images to assist in diagnosing conditions like tumors or fractures.

Telehealth Data

Telehealth data refer to information collected during remote health care interactions between patients and providers [162]. These data are crucial in modern health care, utilizing technology (video, phone, and online platforms) to deliver services remotely and contributing to the overall patient record [163]. The key components of telehealth data are as follows: (1) audio and video recordings, (2) text-based communication, (3) diagnostic and monitoring device data, (4) EHR integration, (5) appointment and scheduling data, (6) patient demographics and consent, (7) prescription and medication data, and (8) PROs. First, audio and video recordings include recordings of virtual consultations used for documentation and quality assurance, with ML analyzing these for sentiment or clinical insights, speech recognition transcribing audio for verbal analysis, and facial recognition/sentiment analysis assessing patient emotions and engagement [164]. Second, text-based communication involves chat logs, texts, or emails containing symptom and treatment information, with NLP extracting and categorizing information (symptoms and treatment discussions) and sentiment analysis gauging patient satisfaction and well-being [165]. Third, diagnostic and monitoring device data include data from remote devices (eg, blood pressure monitors, glucose meters, and wearables) that enable continuous remote health monitoring, with ML models analyzing trends and patterns, and time series analysis and predictive modeling detecting trends/anomalies in vital signs and predicting exacerbations of chronic conditions [166,167]. Fourth, EHR integration involves integration into EHRs/electronic medical records for a comprehensive patient view, with ML analyzing combined data to identify correlations between in-person and virtual interactions, improving diagnosis and treatment planning. Fifth, appointment and scheduling data include information on scheduling, duration, and attendance, with analysis optimizing appointment availability and patient access, and predictive analytics optimizing scheduling and resource allocation by anticipating demand. Sixth, patient demographics and consent include details about the patient and consent for services, ensuring privacy compliance and personalized care context, with ML analyzing demographic data for population health trends, targeted interventions, and service personalization. Seventh, prescription and medication data include information on prescriptions, management, and adherence, supporting virtual refills, with predictive modeling

XSL•FO
RenderX

analyzing adherence patterns to inform personalized interventions and identifying medication-related risks and interactions [168]. Eighth, PROs include patient-reported data on symptoms, well-being, and effectiveness, with ML analyzing PROs via text mining and sentiment analysis to predict treatment responses and assess correlations with clinical outcomes to inform decisions [169].

Feature Engineering for Medical Data

While raw medical data provide the foundation for downstream ML applications, they often require substantial transformation to become analytically useful. Feature engineering refers to the process of extracting, selecting, and constructing meaningful features from raw data to enhance model accuracy, interpretability, and generalizability [170]. This step is particularly critical in the health care domain due to the inherent heterogeneity, noise, and sparsity of clinical data.

The feature engineering process encompasses several essential techniques tailored to different data modalities. Normalization and encoding transform categorical variables and scale numerical values to comparable ranges. Temporal aggregation captures time-dependent patterns from longitudinal records such as EHR data. NLP embedding methods convert unstructured clinical notes into dense vector representations. Signal processing techniques extract meaningful patterns from physiological sensors and medical imaging. Dimensionality reduction methods address the curse of dimensionality while preserving informative variance in high-dimensional genomic or imaging data.

Effective feature engineering not only improves model performance but also enhances interpretability and supports reproducibility across diverse health care settings [171]. In FL scenarios, harmonizing feature spaces across distributed clients is crucial to ensure compatible model training and aggregation [171]. Figure 2 illustrates a high-level pipeline that uses systematic preprocessing transformations for mapping heterogeneous raw medical data (detailed taxonomy in Figure 3) to structured feature vectors suitable for downstream ML and FL modeling.

Figure 2. Feature engineering pipeline for heterogeneous medical data. Raw health care data from 10 diverse sources (detailed taxonomy in Figure 3) undergo systematic preprocessing and transformation to generate unified feature vectors for downstream machine learning (ML) or federated learning (FL) applications. NLP: natural language processing.

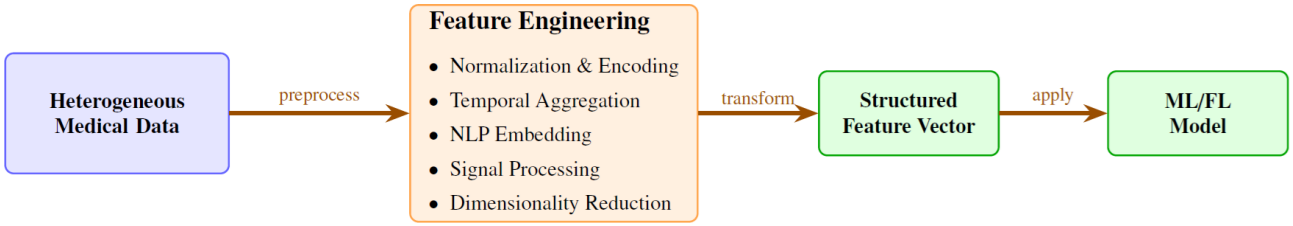

XSL•FO
RenderX

Figure 3. Medical data taxonomy for health care machine learning.

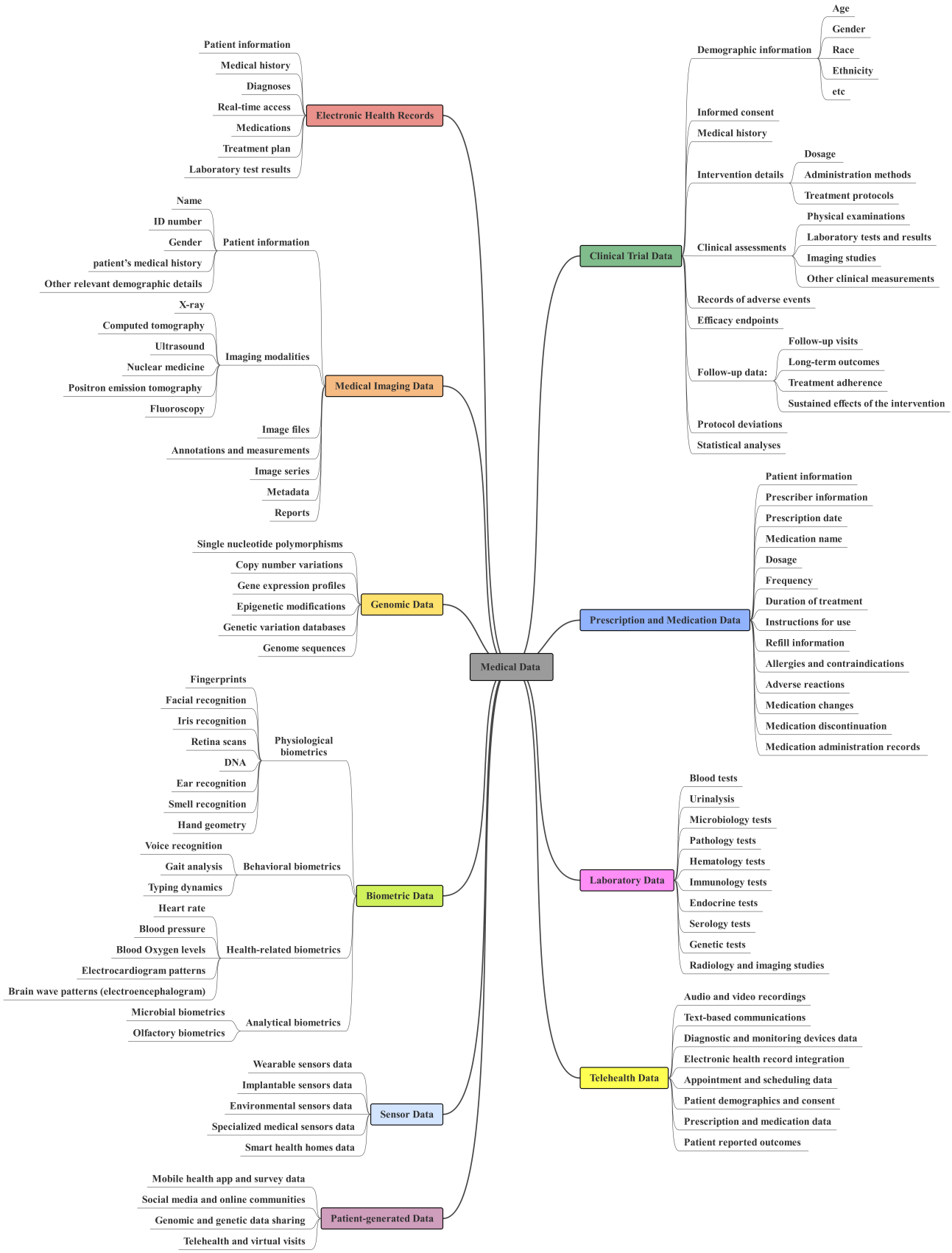

XSL•FO
RenderX

AI and FL in Health Care

Overview

AI is revolutionizing health care, impacting diagnosis, treatment, and patient care. This section explores the evolution of AI in health care, from its historical roots to cutting-edge developments, showcasing its potential across various health care domains. We will further delve into FL as a key technology for the future. FL's collaborative approach enables privacy-preserving data analysis, making it a powerful tool for health care research and delivery. We will discuss its applications, case studies, current challenges, and future directions.

Potential of AI in Health Care

AI is transforming health care across a wide spectrum, from disease diagnosis to patient care management, making significant contributions in several key areas (Figure 4). The first area is *disease diagnosis and imaging*. AI has revolutionized medical imaging, providing more accurate and faster diagnoses [172]. Deep learning models, trained on large datasets (eg, x-ray images [173], MRI scan images [174], and CT scan images [175]), can identify patterns undetectable to the human eye, aiding in the early detection of diseases like cancer, cardiovascular abnormalities, and neurological disorders. The second area is *drug development and personalized medicine*. AI algorithms streamline drug development by predicting molecular behavior [176] and identifying potential drug candidates [177]. In personalized medicine, AI analyzes patient data, including genetic information, to tailor treatments, improving efficacy and reducing side effects [123,178]. The third area is *predictive analytics in patient care*. AI's predictive analytics are crucial to preventive medicine. EHRs are analyzed [179] to predict patient risks for diseases [180], hospital readmission [181], and adverse events, enabling proactive care [180]. The fourth area is *robotics and surgical assistance*. AI-integrated robotics improve surgical precision and outcomes [182]. AI-driven robots assist surgeons in complex procedures, reducing human error and recovery time, and AI supports surgeon training through virtual reality simulations [183]. The fifth area is *patient engagement and telemedicine*. AI-powered chatbots and virtual health assistants provide 24/7 support and health monitoring [184], enhancing patient engagement and adherence. In telemedicine, AI tools assist in remote diagnosis and consultation, making health care more accessible [185]. The sixth area is *health care administration and management*. AI streamlines administrative tasks (scheduling, billing, and claims) [186] and optimizes hospital operations, resource allocation, and patient flow, improving overall efficiency and reducing costs [187]. The seventh area is *global health and epidemic response*. AI is pivotal in tracking and predicting the spread of infectious diseases [188]. During the COVID-19 pandemic, AI models were instrumental in analyzing virus transmission [189], assessing vaccine development [190], and managing health care resources [191]. The key application domains of AI in health care are presented in Figure 5.

Figure 4. Exploring the convergence of artificial intelligence (AI) with health care: trends, applications, and future perspectives. FL: federated learning.

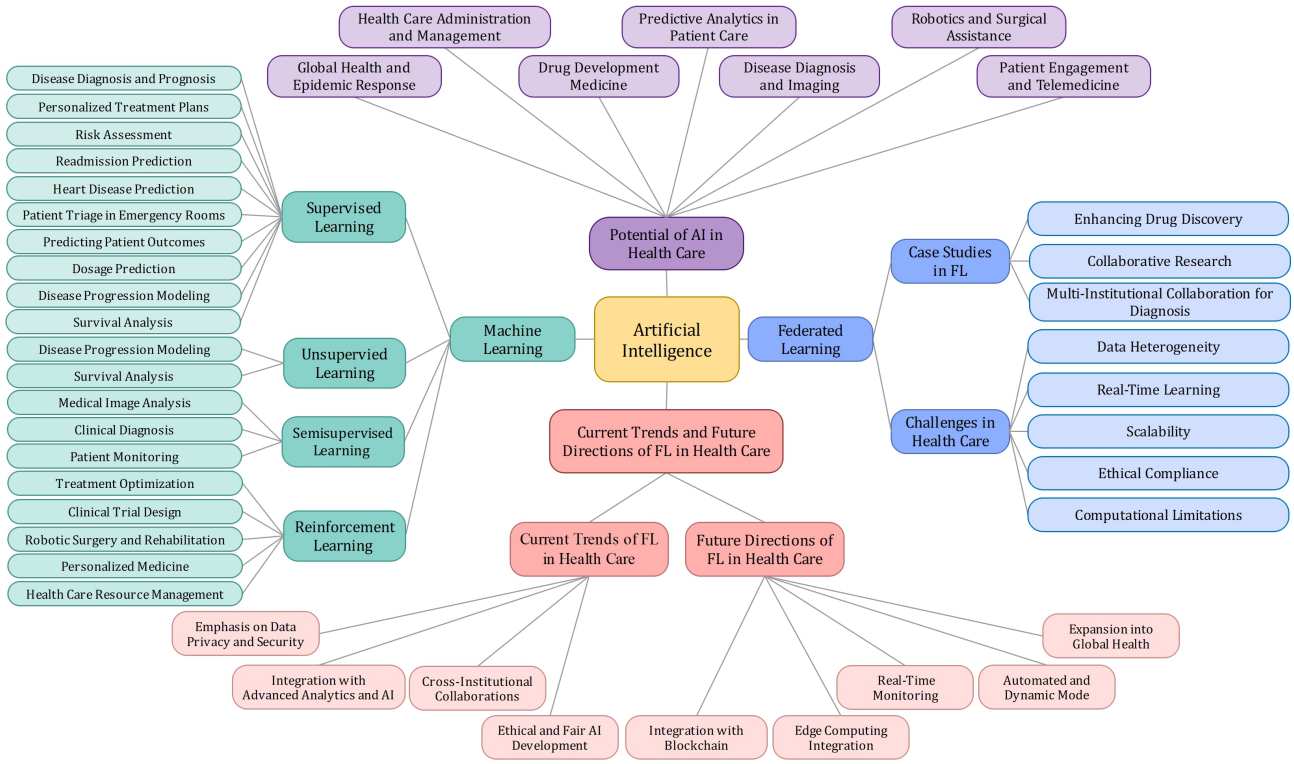

XSL•FO
RenderX

Figure 5. Six key application domains of artificial intelligence (AI) in health care. AI technologies enable advances across drug discovery, surgical precision, diagnostic accuracy, patient rehabilitation, risk assessment, and virtual health assistance.

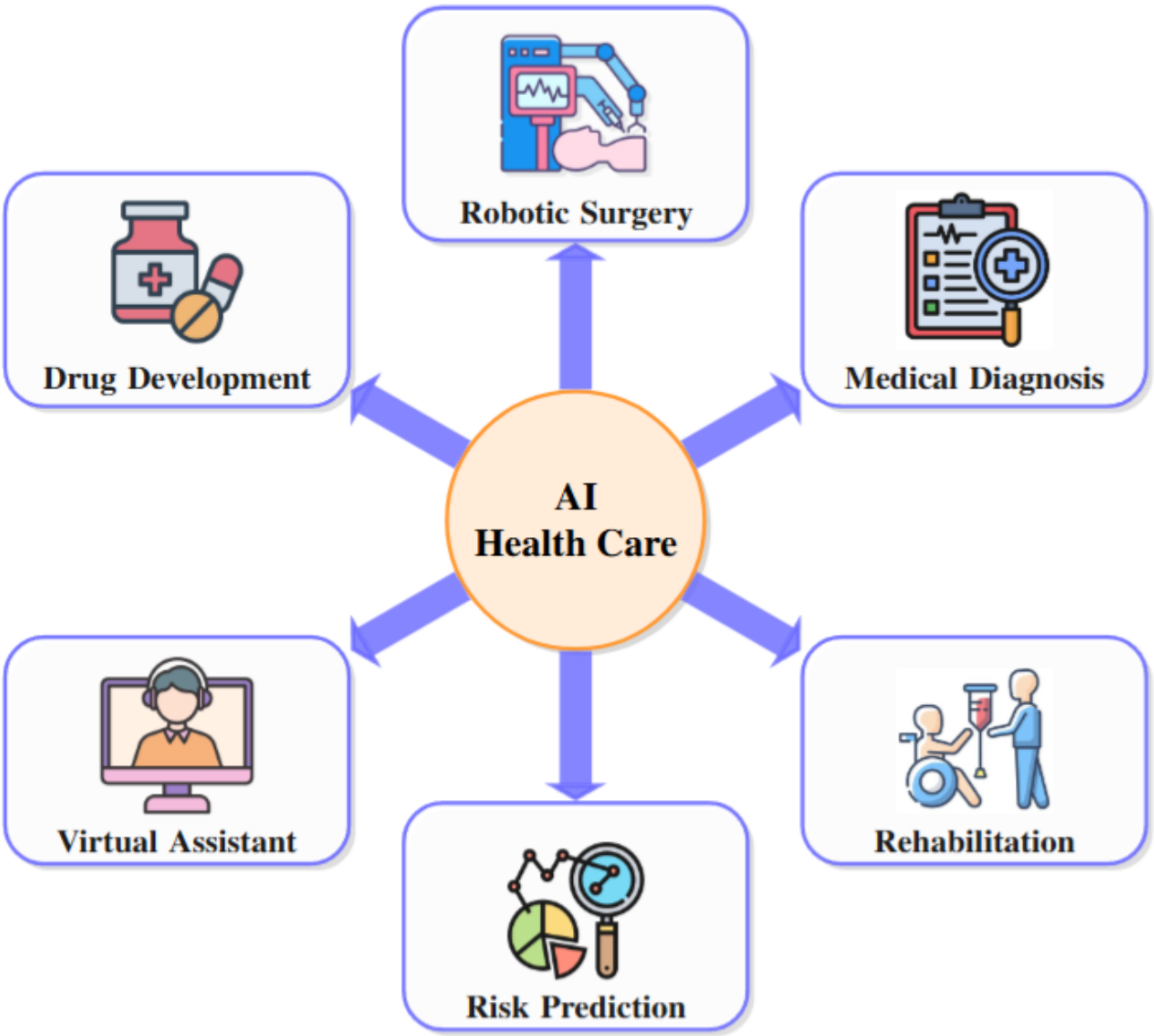

Fundamentals of ML

ML Background

ML, a pivotal branch of AI, is fundamentally reshaping our approach to problem-solving across various domains, including health care. At its core, ML involves the development and application of algorithms that enable computers to learn from and make decisions or predictions based on data. This capacity for self-improvement and adaptation without explicit programming is what sets ML apart.

The cornerstone of ML is data. Algorithms learn from data patterns, and the quality and quantity of these data significantly influence their performance, as explained in the previous section. ML algorithms are sets of rules or instructions given to computers to help them learn from data. These algorithms can be broadly categorized into supervised learning [192-197], unsupervised learning [198-205], semisupervised learning [206-209], and reinforcement learning (RL) [210-215].

Supervised Learning

Supervised learning, a dominant branch of ML, is crucial in health care, leveraging labeled datasets to train models for making predictions or categorizing data [192]. This approach is powerful where the relationship between the input and output is known and can be modeled [216]. The key characteristics include the following: (1) *labeled data*, training data have known outcomes, guiding the algorithm to learn the relationship between input features and the output; (2) *classification and regression*, the two primary tasks are classification (predicting discrete outcomes, eg, diagnosing a disease) and regression (predicting continuous outcomes, eg, recovery time); and (3) *model training and validation*, the model is trained on one portion of data and validated on a separate, unseen dataset to ensure good generalization.

Supervised learning has a significant impact on health care applications, including the following: (1) *disease diagnosis and prognosis*, ML models are trained on clinical data (symptoms, lab results, and imaging) to identify diseases (eg, models trained on imaging data can detect abnormalities like tumors in

XSL•FO
RenderX

radiographic images with high accuracy) [193,194]; (2) *personalized treatment plans*, algorithms analyze patient data to predict individual responses to different treatments, which is effective in fields like oncology [195] where plans are tailored based on tumor genetics; (3) *risk assessment*, models predict the risk of developing conditions (eg, diabetes and heart disease) based on lifestyle, genetics, and other factors [196]; and (4) *readmission prediction*, supervised learning predicts a patient's likelihood of hospital readmission, which is vital for improving patient care and reducing costs [197].

ML offers invaluable tools through its classification and regression tasks. Classification entails categorizing data into predetermined classes, and this is crucial for the following: (1) *disease diagnosis*, models classify patient data into disease categories (eg, using convolutional neural networks to classify dermatological images as benign or malignant skin lesions) [217]; (2) *heart disease prediction*, algorithms analyze patient data (age, blood pressure, and cholesterol level) to classify individuals into risk categories for heart disease, aiding early intervention [218]; and (3) *patient triage in emergency rooms*, models classify patients based on condition severity (analyzing symptoms, vitals, and history) to assist in determining urgency, optimizing patient flow and resource allocation [219].

Regression tasks deal with predicting continuous outcomes in health care, and they are applied to the following: (1) *predicting patient outcomes*, models predict quantitative outcomes like length of hospital stay, surgical recovery time, or disease progression (eg, predicting blood sugar levels in diabetic patients based on lifestyle/medication) [220]; (2) *dosage prediction*, regression algorithms predict the optimal drug dosage for individual patients [144], which is particularly important in treatments like chemotherapy to balance efficacy and toxicity [221]; (3) *disease progression modeling*, models forecast the rate of progression for chronic diseases (eg, Alzheimer disease and Parkinson disease) by analyzing patient data over time, aiding treatment planning [222]; and (4) *survival analysis*, regression models are crucial in oncology for predicting patient survival times after diagnosis or treatment, which is vital for planning and management [223].

Unsupervised Learning

Unsupervised learning is a fundamental ML category that analyzes and groups unlabeled data based on similarities and differences, without predefined labels [198]. Two critical techniques are clustering and dimensionality reduction [199-201,203,205,224], and both are vital in genomics and medical imaging.

Clustering groups objects so that those in the same cluster are more similar to each other than to those in other groups. Its health care applications include the following: (1) *genomic data analysis*, clustering categorizes genes with similar expression patterns, aiding in understanding gene functions, identifying disease markers, and revealing biological pathways (eg, it identifies co-expressed gene groups in diseases like cancer, revealing potential therapeutic targets) [199]; (2) *patient stratification*, algorithms segment patients into groups based on similarities in medical records or genetic information [224], and this stratification identifies disease subtypes with distinct clinical outcomes or treatment responses, facilitating personalized medicine; and (3) *medical imaging*, clustering is used for image segmentation (partitioning images into pixel sets) to identify regions of interest, such as tumors in MRI or CT scans, which is crucial for accurate diagnosis and treatment planning [201].

Dimensionality reduction decreases the number of random variables by obtaining a set of principal variables, which is important for dealing with high-dimensional data common in health care as follows: (1) *genomic data analysis*, genomic data are inherently high-dimensional, with techniques like principal component analysis [202] and t-distributed stochastic neighbor embedding [203] reducing their complexity, and this simplification aids in data visualization, identifying genetic markers, and understanding the genetic architecture of diseases; and (2) *medical imaging*, high-resolution images are computationally intensive, and dimensionality reduction techniques reduce the number of features while retaining essential information [204]. This is crucial for efficient storage, processing, and analysis of medical images, facilitating the development of more efficient diagnostic algorithms [205].

Semisupervised Learning

Semisupervised learning is a fundamental ML category uniquely positioned in health care because it improves model performance by leveraging both labeled and unlabeled data [206,208,225]. This approach is cost-effective, utilizing the abundance of unlabeled data in health care where labeling is often expensive or time-consuming. The key features include the following: (1) *utilizing unlabeled data*, algorithms exploit the vast amounts of unlabeled data in health care databases, which, despite lacking explicit annotations, contain valuable information that complements labeled data and enhances the model's understanding of complex medical phenomena [226]; (2) *combining labeled and unlabeled data*, by incorporating both data types during training, semisupervised algorithms learn more robust representations of the underlying data distribution [227], improving the model's generalization capabilities and leading to more accurate predictions; and (3) *semisupervised techniques*, methods like self-training, co-training, and semisupervised support vector machines iteratively refine predictions using labeled data while leveraging unlabeled data to enhance overall performance [207].

In health care, semisupervised learning applies to diverse areas as follows: (1) *medical image analysis*, algorithms analyze large volumes of unlabeled medical images to identify subtle patterns or anomalies, and combining this unsupervised analysis with labeled data improves the accuracy of tasks such as tumor detection, organ segmentation, and disease classification [209]; (2) *clinical diagnosis*, it assists diagnosis and outcome prediction by leveraging both labeled patient data and unlabeled population health data [208], and this integrated approach enhances diagnostic accuracy and reliability, leading to more informed clinical decisions; and (3) *patient monitoring*, techniques analyze large streams of unlabeled patient data (eg, EHRs and physiological signals) to detect deviations from normal health patterns [225], and incorporating these data into predictive models allows providers to proactively identify and intervene

XSL•FO
RenderX

in adverse health events. Overall, semisupervised learning offers a powerful framework to leverage the wealth of unlabeled health care data, advancing ML model performance, medical research, diagnosis, and treatment strategies [228,229].

Reinforcement Learning

RL is an ML type particularly suited for situations where an "agent" must make a sequence of decisions to achieve a goal [230]. Approaches to learning optimal policies include the following: (1) *dynamic programming*, breaks down decision-making into simpler, recursively solved subproblems, which is effective in environments with a perfect, known model (states and transitions); and (2) *Monte Carlo methods*, model-free approaches relying on repeated random sampling to approximate the optimal policy, which is useful for problems with stochastic dynamics and rewards.

In health care, RL offers innovative ways to approach complex, dynamic decision-making problems [210], operating on reward and penalty principles to learn optimal actions through trial and error. The key concepts include the following: (1) *agent and environment*, an "agent" (eg, a health care model) interacts with its "environment" (eg, patient data/scenarios), taking actions and receiving feedback in the form of rewards or penalties based on the outcomes; (2) *policy*, the strategy the agent uses to determine the next action based on the current environment state (eg, choosing a treatment plan); (3) *reward signal*, guides the agent's actions, with positive rewards encouraging similar decisions and negative rewards signaling adjustment; (4) *value function*, estimates the expected cumulative reward of taking an action in a given state, helping the agent predict long-term outcomes; and (5) *exploration versus exploitation*, balancing trying new actions (exploration) with using known high-reward actions (exploitation), which is analogous to balancing experimental and tried-and-tested treatments.

RL is highly useful in health care applications, including the following: (1) *treatment optimization*, RL adjusts treatment strategies over time based on patient response [211], and for chronic diseases like diabetes, models suggest adaptive insulin dosages, diet, and exercise plans; (2) *clinical trial design*, RL helps determine the most effective trial structures, treatment regimens, and patient selection criteria, enhancing trial efficiency and success rates [231-233]; (3) *robotic surgery and rehabilitation*, used in training robotic systems to adapt to patient-specific conditions and improve based on feedback from surgical outcomes or recovery rates [212]; (4) *personalized medicine*, RL models analyze patient data over time to predict the most effective treatment plans, considering the unique health trajectory and response patterns of each patient [213,214]; and (5) *health care resource management*, algorithms assist in managing resources (hospital bed allocation, staff scheduling, and equipment usage) by learning optimal allocation strategies based on demand and resource availability [215].

Case Studies of ML in Health Care

ML has made significant inroads into health care, transforming patient care, diagnostics, treatment planning, and disease management. Notable real-world examples highlight both successes and challenges as follows: (1) *diagnostic imaging and radiology*, ML (especially deep learning) has revolutionized medical image interpretation (eg, Google Health's ML model outperformed human radiologists in detecting breast cancer in mammograms [234]), and the challenges include integrating these systems into clinical workflows, dealing with diverse data quality, and ensuring consistent performance across different populations and equipment; (2) *drug discovery and development*, ML accelerates drug discovery, reducing cost and time (eg, Atomwise used AI to predict effective molecules, even identifying promising COVID-19 compounds in 2020 [235]), and the primary challenge is the time-consuming and expensive process of validating AI-discovered drugs in clinical trials; (3) *predictive analytics in patient care*, ML models are increasingly used for predictive analytics (eg, the Johns Hopkins team used ML to predict sepsis in hospitalized patients, enabling early intervention [236]), and the challenges include ensuring data privacy, overcoming data silos, and dealing with potential biases in training data; (4) *personalized medicine*, ML aids in tailoring treatments to individual genetic profiles (eg, a team at Earle A Chiles Research Institute used ML to analyze the genetic data of cancer patient for identifying the most effective treatment plans [237]), and the challenges include managing vast amounts of genetic data, ensuring accurate interpretations, and integrating these insights into routine clinical practice; and (5) *mental health applications*, ML models monitor and diagnose mental health conditions (eg, apps like Ginger.io use ML to analyze user interaction and provide personalized mental health support [238]), and the challenges include addressing privacy concerns, ensuring algorithm sensitivity and specificity across diverse populations, and integrating these tools with traditional mental health services.

FL Details

FL Background

Within the field of ML, data security and privacy are paramount concerns. Traditional ML approaches often require centralized data storage, which can raise privacy issues and limit participation due to data ownership restrictions. FL has emerged as a groundbreaking solution, offering a decentralized paradigm for collaborative ML [239]. In FL, multiple entities, such as health care institutions or research centers, collaborate to train a model without sharing their raw data. Each entity trains a local model on its own data and shares only model updates, such as gradients or parameters, with a central server. This collaborative approach allows for distributed learning while preserving data privacy and security.

FL Types

FL encompasses three prominent implementations tailored to diverse scenarios and constraints: (1) horizontal FL (HFL); (2) vertical FL (VFL); and (3) federated transfer learning (FTL) (Figure 6). The first implementation *HFL* involves training models across multiple devices or clients that possess similar data distributions (same features) but cannot share raw data due to privacy concerns. Each client trains a local model and shares only model updates (gradients/parameters) with a central server [240]. The server aggregates these updates to refine a global model, which is then redistributed. HFL is suitable where data are distributed with similar characteristics, such as mobile

XSL•FO
RenderX

phones, IoT edge devices [241], and EEG data [242]. The challenges involve privacy-preserving techniques, communication efficiency, and aggregation strategies [243]. The second implementation *VFL* addresses scenarios where data are distributed across multiple parties with complementary features (different features) that cannot be shared due to privacy/proprietary concerns [244]. Each party holds a subset of features. The goal is collaborative model training without sharing raw data, commonly using secure multiparty computation (SMPC) and HE to enable computation on encrypted data while preserving privacy [245]. VFL is applied in health care where different institutions hold complementary patient data (eg, medical records and lab results) that are crucial for accurate models while preserving privacy [246]. The third implementation *FTL* extends traditional transfer learning to federated settings, where knowledge from related tasks or domains is leveraged across multiple decentralized datasets [247]. Unlike traditional transfer learning, FTL aggregates knowledge from decentralized clients to refine a base model (initialized from pretraining or scratch). This accommodates variations in data distributions and characteristics across clients [248]. FTL is beneficial where labeled data are scarce or unevenly distributed, such as electrocardiogram signal analysis [249], improving performance by leveraging knowledge from related domains.

XSL•FO
RenderX

Figure 6. Types of federated learning (FL): (A) horizontal FL; (B) vertical FL; (C) federated transfer learning.

XSL•FO
RenderX

Case Studies of FL in Health Care

FL holds promise in health care by facilitating the collaborative development of robust ML models across institutions (eg, hospitals and research centers) while safeguarding patient data privacy [250,251]. This approach trains ML models on local datasets, sharing only model updates (not raw data) to a central server for aggregation, thus mitigating privacy concerns and reducing data transfer costs [252]. FL offers enhanced privacy and utilization of diverse datasets without centralization. Real-world examples of FL in health care include the following: (1) *multi-institutional collaboration for disease diagnosis*, FL's most significant success involves collaborative projects for disease diagnosis, and an international consortium (including institutions like Massachusetts General Hospital, University of Cambridge, Lahey Hospital, Assuta Medical Centers, Dasa S.A., National Taiwan University, and Seoul National University Hospital) successfully used FL to develop highly accurate models for predicting critical care patient outcomes (eg, mortality and length of stay) by leveraging diverse populations while maintaining data privacy; (2) *enhancing drug discovery and development*, FL has been used in pharmaceutical research to create predictive models for drug response and toxicity, and a notable instance is a project where multiple pharmaceutical companies shared algorithmic models, rather than data, to predict the success of drug compounds, expediting discovery while maintaining confidentiality of proprietary data [253]; and (3) *collaborative research*, FL enables institutions to collaborate on cancer research without sharing sensitive patient data, and a notable example is the collaboration facilitated by Intel and Penn Medicine, utilizing FL to identify brain tumors [254]. These cases demonstrate how FL advances medical research while maintaining data confidentiality.

Challenges of FL in Health Care

FL offers a groundbreaking opportunity to collaboratively develop potent ML models while safeguarding patient data privacy and minimizing transfer costs. However, realizing its full potential in health care requires overcoming significant challenges, including the following: (1) *data heterogeneity and model generalizability*, variability in medical data across institutions (different formats and diverse populations) complicates model convergence and performance, necessitating models that are generalizable to diverse patient populations and data types [255]; (2) *technical and computational limitations*, successful FL implementation requires substantial computational resources and technical expertise that may be unequally distributed among institutions, and balancing these disparities is crucial for success [256]; (3) *regulatory and ethical compliance*, FL must adhere to complex regulations like HIPAA (United States) and GDPR (Europe) governing data privacy and security, and ethical considerations require obtaining patient consent and ensuring equitable distribution of research benefits, necessitating clear data governance and ethical frameworks [257]; and (4) *scalability and real-time learning*, scaling FL models to accommodate real-time learning and large datasets poses technical hurdles, making the efficient management and real-time updating of models challenging [258]. Possible solutions include establishing standardized data formats across entities for consistency, implementing comprehensive data preprocessing pipelines to enhance quality, and designing FL algorithms adept at handling diverse data types and operating effectively across varying computational resources [259].

Future prospects involve integrating FL with emerging technologies like IoT devices and real-time health monitoring systems. This integration could enable the continuous improvement of ML models with real-time data from diverse, distributed sources, leading to more dynamic and responsive health care solutions. Additionally, advancements in edge computing could further enhance FL's efficiency and scalability in health care [260].

Current Trends and Future Directions of FL in Health Care

Overview

The landscape of FL in health care is rapidly evolving, driven by technological advancements and the growing need for collaborative and privacy-preserving data analysis. This section outlines the current trends shaping FL in health care and forecasts its future trajectory.

Current Trends of FL in Health Care

The prevailing trends shaping FL in health care reflect a dynamic evolution toward enhanced analytics, privacy, collaboration, and ethical considerations as follows: (1) *integration with advanced analytics and AI*, FL is being increasingly integrated with sophisticated AI techniques, such as deep learning, to enhance its analytical capabilities [261,262], and this allows for more complex and accurate models capable of addressing intricate health care challenges like personalized medicine and predictive analytics [250]; (2) *emphasis on data privacy and security*, due to heightened data privacy concerns, FL is gaining traction as a preferred method for collaborative health care research [263], and its inherent design, which allows for model training without sharing raw data, aligns well with stringent data privacy regulations like HIPAA and GDPR; (3) *cross-institutional collaborations*, there is a growing trend of cross-institutional collaborations, as explained previously, which is facilitated by FL, and these collaborations unite hospitals, research centers, and academic institutions, enabling them to pool knowledge and data resources for collective model improvement while maintaining data sovereignty; and (4) *ethical and fair AI development*, there is an increased focus on ethical AI development as FL evolves, ensuring that FL models are fair, unbiased, and representative of all patient demographics, thus addressing concerns around algorithmic bias [19,264,265].

Future Directions of FL in Health Care

FL in health care is poised for significant expansion, driven by emerging technological advancements and evolving health care landscapes. This forthcoming evolution encompasses the following: (1) *expansion into global health initiatives*, FL has the potential to significantly impact global health research, particularly in areas with stringent privacy laws or limited data-sharing capabilities, facilitating the analysis of global health trends and the development of models representative of diverse

XSL•FO
RenderX

populations [250]; (2) *automated and dynamic model updating*, the future will likely see more automated and dynamic updating of FL models [266], enabling health care systems to respond quickly to new data or changing health trends, making the models more adaptive and responsive; (3) *use in remote and real-time monitoring*, with the proliferation of wearable devices and IoT in health care, FL is set to play a significant role in real-time patient monitoring and remote health care, providing personalized insights and treatments based on data from diverse patient populations [267]; (4) *edge computing integration*, integrating FL with edge computing could decentralize computational workload, allowing for faster and more efficient model training and updates, especially in real-time applications [265,268,269]; and (5) *integration with blockchain for enhanced security*, the integration of FL with blockchain technology is a promising development that bolsters data security and adds a layer of transparency and traceability to the FL process, ensuring immutable record-keeping and verifiable model updates in FL networks [13]. In summary, FL in health care is at a dynamic juncture, poised to reshape research and delivery. Its alignment with current needs for privacy, collaboration, and advanced analytics, coupled with its adaptability for future technological trends, positions FL as a key player in the future landscape of health care technology, paving the way for more equitable, secure, and efficient use of health care data globally.

Blockchain

Blockchain Background

Blockchain technology has undergone a remarkable evolution since its inception in 2009 with the creation of Bitcoin by an individual or group using the pseudonym Satoshi Nakamoto [270]. The primary purpose of Bitcoin was to establish a decentralized digital currency, and the innovation that made this possible was blockchain—a distributed ledger that records transactions across a network of computers securely and transparently.

In the following years, the potential applications of blockchain technology expanded beyond cryptocurrency. Vitalik Buterin introduced Ethereum [271] in 2015, introducing the concept of smart contracts—self-executing contracts with the terms of the agreement directly written into code. This development opened up a broader spectrum of decentralized applications (DApps) and laid the foundation for blockchain's role in facilitating not only peer-to-peer (P2P) transactions but also complex programmable interactions.

The years that followed witnessed a surge in blockchain projects and platforms, each aiming to address specific challenges across various industries. The technology gained recognition for its potential to enhance transparency, security, and efficiency. Consortia and collaborations emerged, with enterprises exploring how blockchain could optimize supply chains [272-275], streamline financial transactions, and enhance data integrity.

Blockchain continues to evolve, with ongoing efforts to address scalability issues, energy consumption concerns, and regulatory considerations. From its humble beginnings as the underlying technology for Bitcoin, blockchain has grown into a versatile tool with the potential to reshape how industries manage and verify data. The technology's journey reflects an ongoing quest for innovative solutions to long-standing challenges in the digital realm.

In this section, we provide a concise overview of fundamental concepts, features, structure, and taxonomy within the realm of blockchain technology.

Blockchain Technology: An Overview

Blockchain technology is a decentralized and distributed ledger system designed to facilitate secure and transparent transactions without the need for a central authority. At its core, blockchain consists of a chain of blocks, each containing a list of transactions. These blocks are linked together in a chronological and immutable manner, forming a continuous chain. One of the key features of blockchain is its decentralization, meaning that the ledger is maintained by a network of nodes rather than a single central entity. This distributed nature enhances security, reduces the risk of fraud, and ensures transparency in the transaction process.

Blockchain technology possesses several distinctive features that contribute to its popularity and versatility across various industries. The features of blockchain technology are as follows: (1) decentralization, (2) immutability, (3) transparency, (4) security, (5) distributed ledger, (6) consensus mechanism, (7) anonymity and privacy, (8) efficiency and speed, (9) interoperability, and (10) ability to support smart contracts. First, regarding decentralization, blockchain operates on a P2P network, eliminating the need for a central authority or intermediary, and this enhances security, reduces the risk of a single point of failure, and promotes trust among mutually untrusted participants [276]. Second, immutability is the capability of a blockchain ledger to remain unchanged. Once a block is added to the blockchain, it becomes virtually impossible to alter or delete the information within it. Immutability ensures the integrity of the transaction history and builds trust in the accuracy of recorded data. Third, regarding transparency, the entire transaction history is visible to all participants in the network, and this fosters trust and accountability as participants can independently verify transactions and the state of the blockchain. Fourth, regarding security, blockchain uses cryptographic techniques to secure transactions and control access to the network. Consensus mechanisms, such as proof of work (PoW) and proof of stake (PoS), enhance security by preventing unauthorized changes to the blockchain [277-279]. Fifth, the ledger is distributed among the nodes over the network, and each node in the network holds a copy of the blockchain. This distribution ensures redundancy, resilience, and a shared source of trust among the participants. Sixth, consensus is a mechanism that gives the network the ability to agree upon the validity of transactions (and blocks) and the order in which they can be added to the blockchain. Seventh, regarding anonymity and privacy, while all transactions in the blockchain network are transparent, participants remain pseudonymous due to the use of public/private key pairs. Eighth, regarding efficiency and speed, blockchain reduces the need for intermediaries and manual processes, leading to faster and more efficient transactions. In some cases, however, the speed

XSL•FO
RenderX

of transactions may depend on the specific consensus mechanism used. For instance, PoW blockchains, like Bitcoin, tend to be slower (ie, with less throughput) compared to traditional payment systems like Visa and Mastercard. The primary reasons for this include the inefficiency of the underlying consensus mechanisms and the way transactions are processed [280]. Ninth, blockchain interoperability refers to the capacity of various blockchain networks to interact seamlessly, facilitating the exchange of messages, data, and tokens among them [281-285]. Standards and protocols are evolving to enable such communication and data exchange between disparate blockchain platforms. The Inter-Blockchain Communication (IBC) protocol [286] is designed to facilitate this interoperability by providing a standardized way for independent blockchains to transfer and communicate with each other. Tenth, regarding the ability to support smart contracts, smart contracts are self-executing contracts with the terms directly written into code. These contracts automate and enforce predefined rules and agreements, reducing the need for intermediaries and streamlining processes [287,288].

Consensus Mechanism in Blockchain

Consensus in the context of blockchain refers to the mechanism by which a distributed network of nodes agrees on the state of the system or the validity of transactions [289-291]. Since blockchain operates in a decentralized and trustless environment, consensus is crucial to ensure that all participants have a consistent view of the blockchain's history and current state. The consensus mechanism is responsible for preventing double-spending (where the same digital asset is spent more than once) and maintaining the integrity of the blockchain. Different blockchain networks use various consensus algorithms, each with its own set of rules and processes. The most prominent and currently existing consensus protocols are as follows: PoW; PoS; Byzantine fault tolerance (BFT); direct acyclic graph tangle (DAG); and hybrid consensus, including PoS and PoW hybrid, proof of authority (PoA) and PoW hybrid, delegated PoS (DPoS) and PoA hybrid, PoW and BFT hybrid, and hybrid voting system.

PoW Protocol

This is the original consensus algorithm used by Bitcoin [270] and many more cryptocurrencies. In PoW, participants (miners) solve complex mathematical puzzles to validate transactions and create new blocks. The first miner to solve the puzzle gets the right to add a new block to the blockchain. PoW is resource-intensive and requires a significant amount of computational power [292]. Blockchains that rely on PoW are more prone to forks. A fork in blockchain technology refers to a split in the blockchain's transaction history, resulting in two or more separate paths. This can occur for various reasons, such as changes in protocol rules, disagreements among participants, or software upgrades [293].

PoS Protocol

In PoS [294], validators (ie, block proposer participants) are chosen to create new blocks and validate transactions based on the amount of cryptocurrency they hold and are willing to "stake" as collateral. This eliminates the need for energy-intensive mining and aims to provide a more energy-efficient alternative to PoW. Examples include Ethereum's [295] transition to Ethereum 2.0, Cardano [296], and Algorand [297]. DPoS is an improvement over traditional PoS in terms of scalability and efficiency [298].

BFT Protocols

BFT consensus algorithms are a class of protocols designed to achieve consensus in distributed systems, even in the presence of faulty or malicious nodes. In a BFT-based consensus algorithm, a network of nodes collaborates to agree on the state of the system or the validity of transactions. The term "Byzantine fault" originates from the Byzantine generals' problem [299], a theoretical scenario where a group of generals must come to a unanimous agreement on a coordinated action, despite the possibility of some generals being traitors and sending conflicting messages. Practical BFT (pBFT) is the most prominent variant of BFT-based consensus protocols [300]. This algorithm is designed to tolerate up to one-third of the total number of nodes being faulty or malicious. This means that as long as no more than one-third of the nodes in the network exhibit Byzantine behavior (ie, they may fail arbitrarily or behave maliciously), pBFT can still reach consensus and continue to operate correctly. Some variants of BFT (eg, Hotstuff [301], pBFT, and improved BFT) are supported by Hyperledger Fabric [302].

DAG Protocols

DAG consensus protocols are a class of distributed consensus algorithms that use a data structure called a directed acyclic graph to achieve agreement on the order of transactions or events in a decentralized network. Unlike traditional blockchain-based consensus protocols where transactions are organized into linear blocks, DAG-based protocols organize transactions in a more flexible graph structure [303]. One of the most well-known implementations of DAG consensus is the tangle [304], which is used in the IOTA cryptocurrency network [305]. In the tangle, each transaction directly references and approves 2 previous transactions, forming a directed acyclic graph structure. The most prominent blockchains that run on proof of capacity include Signum, Chia, and SpaceMint.

Hybrid Consensus Protocols

Overview

Hybrid consensus models in blockchain combine elements of multiple traditional consensus mechanisms to leverage their respective strengths and mitigate their weaknesses. These models aim to achieve a balance among decentralization, security, scalability, and energy efficiency [306]. Some hybrid consensus models in blockchain are described below.

PoS and PoW Hybrid

Some blockchain networks combine PoS and PoW mechanisms to achieve consensus (eg, TwinsCoin [307]). For example, a PoW component may be used for initial block creation, while PoS is utilized for subsequent block validation or as a way to elect validators. This hybrid approach aims to maintain security through PoW while improving scalability and energy efficiency with PoS.

XSL•FO
RenderX

PoA and PoW Hybrid

In this hybrid model, a network may utilize PoW for initial block creation and PoA for block validation. PoW ensures the initial distribution of tokens and secures the network against Sybil attacks, while PoA provides fast finality and scalability by relying on known and trusted validators.

DPoS and PoA Hybrid

DPoS allows token holders to vote for a limited number of delegates who are responsible for block validation. In a hybrid approach, DPoS can be combined with PoA, where the initial set of validators is determined through PoA, and then, token holders can vote for additional delegates using DPoS. This hybrid model aims to achieve both decentralization and scalability.

PoW and BFT Hybrid

This hybrid model combines the energy-intensive PoW with a BFT-based consensus algorithm such as pBFT or Tendermint [308]. PoW is used for block creation, while BFT consensus ensures finality and fault tolerance. This approach aims to achieve both security and efficiency in blockchain networks [309].

Hybrid Voting Systems

Some blockchain networks combine different voting mechanisms, such as direct voting by token holders and voting by elected delegates [310]. This hybrid voting system aims to balance the influence of token holders with the expertise and accountability of elected representatives.

Blockchain Data Structure

In blockchain technology, the data structure plays a pivotal role in ensuring the integrity, security, and immutability of the distributed ledger [278,311]. At its core, a blockchain is composed of a series of blocks, each containing a bundle of transactions. These blocks are cryptographically linked together sequentially, forming a continuous chain. The data structure of a block typically includes several key components: a header, a list of transactions, and a cryptographic hash. The header contains metadata, such as the block's unique identifier (block hash), a timestamp, and a reference to the previous block's hash, thus establishing the chronological order of blocks. The transaction list records the details of all transactions included in the block, such as sender and receiver addresses, transaction amounts, and cryptographic signatures for verification. Additionally, each block is assigned a cryptographic hash, which is computed based on its contents using a hashing algorithm like SHA-256. This hash serves as a unique identifier for the block and is crucial for maintaining the integrity of the blockchain. Any alteration to the data within a block would result in a change in its hash, thereby breaking the chain's continuity and signaling tampering. The inherent immutability and tamper-resistance of the data structure in blockchain ensure that once recorded, transactions cannot be altered or deleted without consensus from the network participants, establishing a reliable and transparent system for recording and verifying transactions [312].

Blockchain Network and Architecture

The network architecture in blockchain is a distributed and decentralized system that enables the secure and transparent exchange of data and value across a network of interconnected nodes. At its core, a blockchain operates as a P2P network where each participant, or node, maintains a copy of the entire blockchain ledger. This distributed architecture ensures that there is no single point of failure, as the data are replicated and synchronized across multiple nodes. Nodes communicate with each other through a consensus mechanism. Depending on the consensus algorithm used, nodes may take on different roles, such as miners in the PoW system or validators in the PoS system. Transactions are broadcast to the network and validated by consensus, typically requiring confirmation from a majority of nodes before being added to the blockchain. This network architecture provides several benefits, including resilience against censorship and tampering, increased transparency and accountability, and enhanced security through cryptographic techniques.

Additionally, the decentralized nature of blockchain networks promotes trust among participants by eliminating the need for intermediaries and central authorities, thereby fostering a more inclusive and democratic ecosystem for conducting transactions and exchanging value [313].

Blockchain technology typically consists of 6 common layers, each serving a specific purpose in the network's function and security (Figure 7). The six layers are as follows: (1) *network layer,* the foundation facilitating communication between nodes using protocols like TCP/IP, HTTP, and P2P protocols, responsible for transmitting data (transactions and blocks) across the network; (2) *data layer*, stores the actual blockchain data, including blocks, transactions, and smart contracts, utilizing structures and storage mechanisms optimized for secure and efficient retrieval; (3) *consensus layer*, ensures all nodes agree on the validity and order of transactions added to the blockchain using various mechanisms, such as PoW, PoS, DPoS, and pBFT; (4) *smart contract layer*, enables the creation and execution of programmable, self-executing contracts with terms written directly into code, powering various DApps; (5) *incentive layer*, provides mechanisms (typically block rewards and transaction fees, eg, Bitcoin's block rewards and Ethereum's gas fees) to incentivize participants (miners or validators) to contribute resources and maintain the network's security and integrity; and (6) *application layer*, encompasses the user-facing applications and interfaces (eg, DApps, wallets, smart contracts, and other software) that interact with the blockchain protocol.

XSL•FO
RenderX

Figure 7. Six-layer blockchain architecture illustrating increasing abstraction from the network layer to the application layer.

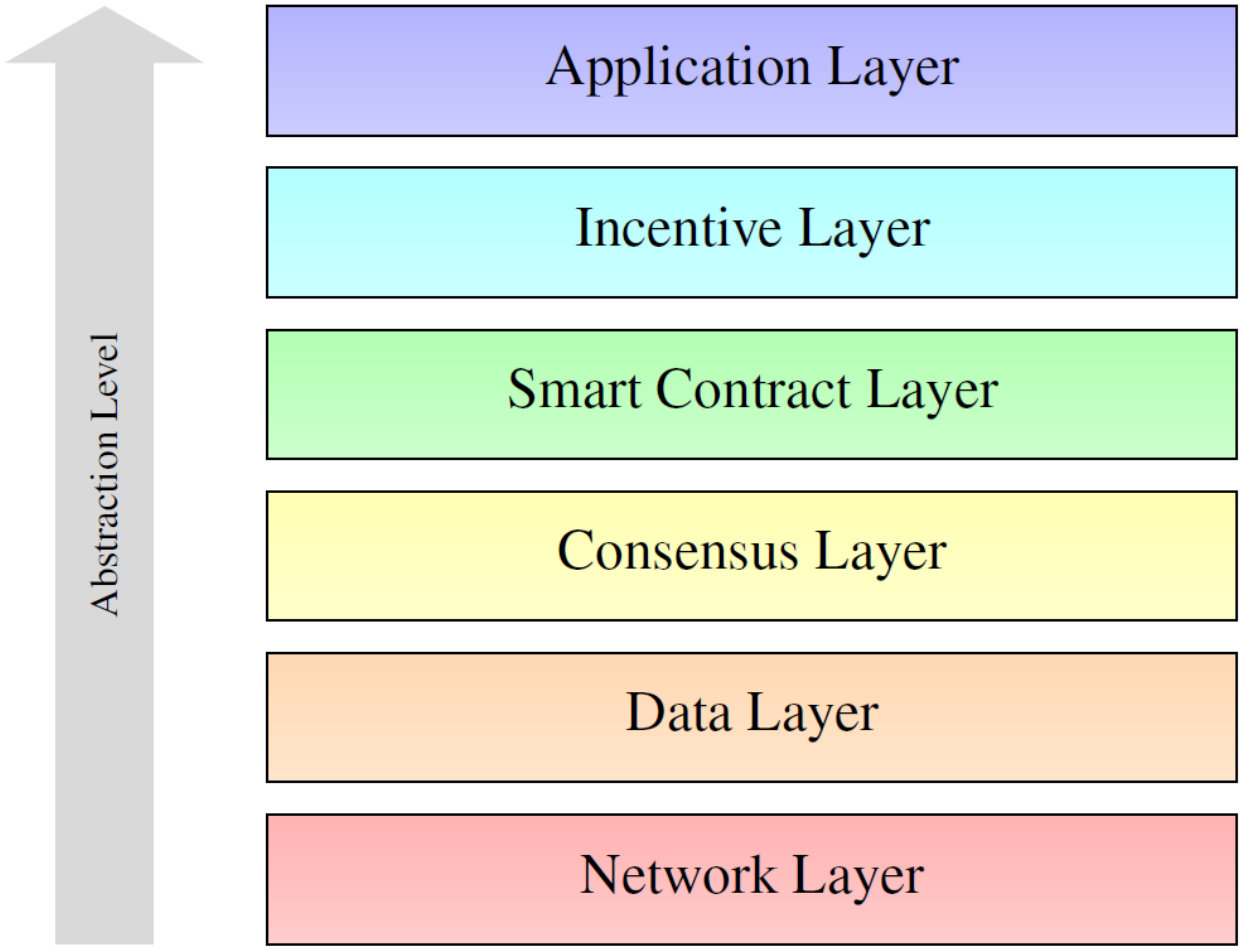

IBC Protocol

Interoperability is one of the most important features of next-generation blockchain networks and refers to the ability of different blockchain platforms to communicate, share data, and transact with each other seamlessly. It enables interoperability between disparate blockchain networks, allowing them to interact and exchange information or assets without the need for intermediaries or centralized exchanges. Interoperability is essential for realizing the full potential of blockchain technology by facilitating cross-chain transactions, asset transfers, and data sharing between different blockchain ecosystems [286].

The IBC protocol is a set of standards and protocols designed to enable communication and interoperability between independent blockchain networks. IBC facilitates the secure and trustless transfer of assets and data across different blockchains, allowing them to interact and transact with each other directly. The protocol defines a standardized messaging format and a set of rules for validating and verifying transactions between participating blockchains. By implementing the IBC protocol, blockchain networks can establish interconnectivity, enabling cross-chain transactions, decentralized exchanges, and interoperable DApps [283,314,315].

IBC is a pillar of the Internet of Blockchains [311,316], a vision where blockchain networks are interconnected like the global internet, creating a decentralized and open ecosystem where data, assets, and services flow freely between different blockchains. Examples of interoperability solutions include the following: (1) *Cosmos*, a decentralized network utilizing the IBC protocol to enable communication and interoperability between its various blockchain platforms, with the Cosmos Hub serving as the primary connection point [317]; (2) *Polkadot*, a multichain blockchain platform that enables interoperability between different parachains (parallel blockchains) within its network via its relay chain, allowing them to share data and assets seamlessly [318]; and (3) *Wanchain*, a cross-chain blockchain platform focused on interoperability, enabling the secure and decentralized exchange of assets between different blockchain networks, including Bitcoin and Ethereum [319].

Blockchain Taxonomy

Overview

At a high level, blockchain networks are classified into 3 main categories: private, public, and consortium blockchains. These are briefly explained in the following subsections.

XSL•FO
RenderX

Private Blockchain

A private blockchain is a permissioned blockchain network where access and participation are restricted to authorized entities only. These entities typically have known identities and are granted permission to join the network by a central authority or administrator. Private blockchains are often used by enterprises and organizations to build internal blockchain solutions for specific use cases such as supply chain management, document verification, and intercompany transactions. They offer enhanced privacy, control, and scalability compared to public blockchains [320,321]. For instance, Hyperledger Fabric is a private blockchain framework developed by the Linux Foundation's Hyperledger project [302,322,323]. It is designed for enterprise use cases and enables organizations to create permissioned blockchain networks with customizable features and governance models.

Public Blockchain

A public blockchain is a permissionless blockchain network that is open to anyone to join, participate, and transact without requiring permission or identification. Public blockchains are decentralized networks where transactions are transparent, immutable, and verifiable by anyone. They offer high levels of transparency, censorship resistance, and security but may sacrifice scalability and privacy due to their open nature [324-327]. Public blockchains are often used for cryptocurrencies, DApps, and tokenized assets. For instance, Bitcoin operates as a decentralized P2P network for sending and receiving the Bitcoin cryptocurrency.

Consortium Blockchain

A consortium blockchain is a semidecentralized blockchain network governed by a consortium or group of organizations rather than a single centralized entity. Consortium blockchains are permissioned networks where the consensus process and governance are shared among a predefined set of participants. Consortium blockchains are commonly used in industries or sectors where multiple organizations collaborate on shared processes or infrastructure while still maintaining some level of control and privacy [328-331]. They offer a balance between the decentralization of public blockchains and the control of private blockchains. For instance, R3 Corda is a consortium blockchain platform developed by the enterprise blockchain consortium R3. Corda is designed for use cases that require privacy, scalability, and interoperability in multiple organizations from sectors such as finance, health care, and supply chain.

Integration of FL With Blockchain

Overview

Integrating blockchain technology with FL has emerged as a novel approach to address inherent data privacy, security, and trust challenges within distributed ML systems [21,23,332,333]. FL, characterized by training ML models across decentralized devices without centrally aggregating raw data, offers significant advantages in preserving user privacy and data confidentiality. In other words, in federated ML, the exchange occurs at the parameter level instead of transmitting raw data. This approach mitigates the risk associated with centralized computing architectures, which are susceptible to targeted attacks and potential denial of service due to their single-point-of-failure vulnerability. A detailed discussion on FL is presented in the FL Details section, including its different types (HFL, VFL, and FTL), applications (particularly in health care), and potential future directions (integration with advanced analytics and cross-institutional collaborations), highlighting FL's pivotal role in reshaping health care technology.

However, concerns persist regarding the integrity of FL systems, particularly regarding data tampering or manipulation by malicious or compromised nodes. Blockchain, renowned for its immutable and transparent ledger capabilities, presents a compelling solution to these challenges. By leveraging blockchain's decentralized consensus mechanisms and cryptographic primitives, FL systems can ensure data integrity, traceability, and transparency throughout the ML model training process [334-336]. Moreover, blockchain's smart contract functionality enables the establishment of auditable and self-executing agreements among participants, further enhancing the trustworthiness of FL collaborations [23]. This integration not only addresses privacy and security concerns but also fosters a more collaborative and inclusive environment for distributed ML research and applications. As such, the motivation behind the integration of blockchain and FL lies in the pursuit of enhancing data privacy, security, and trust in decentralized ML ecosystems, ultimately advancing the adoption and efficacy of FL methodologies in various domains such as health care [13,336,337].

Blockchain technology holds immense potential to significantly enhance the security, transparency, and trustworthiness of FL systems by addressing key challenges like verifying local model updates, aggregating the global model, and incentivizing participants. The integration of blockchain into FL is explored as follows: (1) verifying local model updates, (2) global model aggregation, and (3) incentivizing participants. First, regarding verifying local model updates, blockchain offers an immutable and tamper-proof record for local model updates from participants. These updates are recorded as transactions, ensuring their validity and preventing unauthorized modifications. Furthermore, smart contracts can be used to define specific rules, and they function as verification mechanisms, executing algorithms to validate the integrity and accuracy of updates [338]. Second, regarding global model aggregation, blockchain empowers FL with decentralized consensus mechanisms (eg, PoW and PoS) that allow participants to collectively agree on the process of aggregating the global model. The entire aggregation process is transparently recorded on the blockchain, allowing participants to verify the fairness and accuracy of the final model [339]. Third, regarding incentivizing participants, a key strength of blockchain is its ability to create tokenized incentives (tokens or cryptocurrencies) to reward participants who contribute data or computational resources [340]. Smart contracts automate the distribution of these incentives based on predefined criteria (eg, quality of contribution and computational resources provided) [341]. Crucially, the entire distribution process is transparently recorded on the blockchain, ensuring accountability and fairness [342].

XSL•FO
RenderX

By leveraging the unique strengths of blockchain technology, FL systems can achieve a new level of security, transparency, and trust, ultimately fostering a more collaborative and efficient environment for AI development.

Integration Architecture

Several research efforts have identified distinct architectures for integrating blockchain and FL. This paper proposes a framework similar to that in a previous report [343], with slight variations in terminology. Based on the level of interaction between blockchain and FL entities, we categorize these architectures as fully coupled, semicoupled, and loosely coupled.

Fully-Coupled Architecture

In the fully-coupled architecture, blockchain nodes (ie, miners or validators) perform dual roles as FL clients. Within this architecture, FL clients engage in computing local model updates as well as validating these updates as blockchain nodes. Notably, blockchain nodes not only partake in training local models but also participate as the global model aggregator. The aggregator, which may be a selected node, a designated leader, or a collection of nodes based on the predefined protocol, is responsible for gathering local model updates. Every node in this model has the opportunity to function as a blockchain validator, a local model trainer, and a global aggregator concurrently. Consequently, both local model updates and global model updates are contained within the blockchain. Importantly, the absence of a necessity to transmit the global model to a central server mitigates the risk of a single point of failure within this architecture. A schematic of this architecture is depicted in Figure 8.

Figure 8. Fully-coupled blockchain-based federated learning (FL) architecture: (A) detailed architecture; (B) schematic diagram. All local updates are committed on-chain via a smart contract that performs aggregation. Clients act as validators within a unified blockchain trust domain.

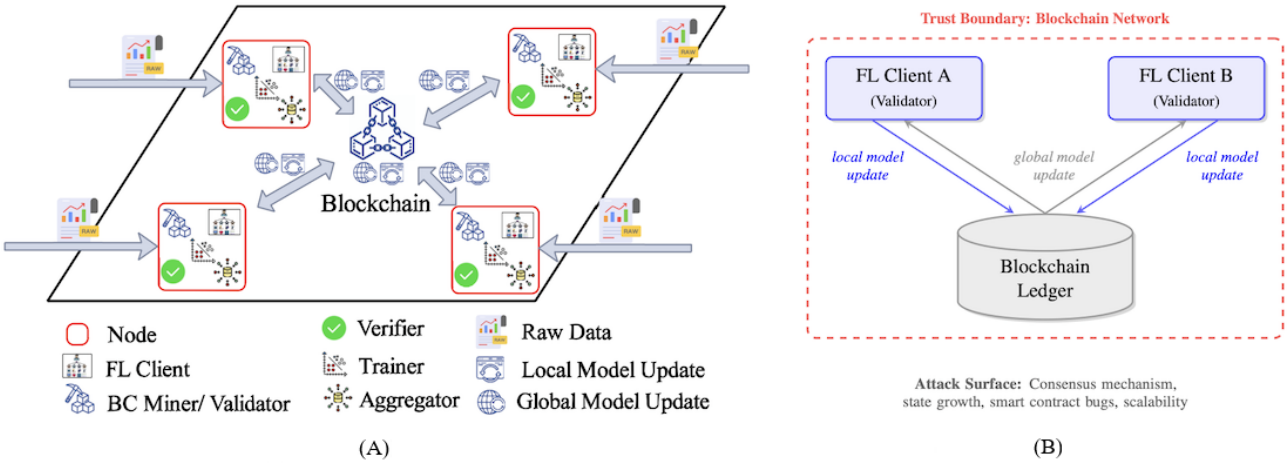

Semicoupled Architecture

In the semicoupled architecture, blockchain and FL clients inhabit separate networks, although FL clients retain the capability to interact with the blockchain and manipulate the distributed ledger. FL clients gather data from diverse sources, train local models, and subsequently upload local model updates to the blockchain. Blockchain nodes (ie, miners or validators) are tasked to validate the uploaded local model updates that will be used for training the global model. Upon the preparation of the global model, blockchain nodes will store it within the blockchain. Participant rewards are allocated based on a predefined incentive mechanism. This architecture also circumvents the potential for a single point of failure. A schematic of this architecture is depicted in Figure 9.

XSL•FO
RenderX

Figure 9. Semicoupled blockchain-based federated learning (FL) architecture: (A) detailed architecture; (B) schematic diagram. Validators perform aggregation off-chain with an on-chain registry and incentives.

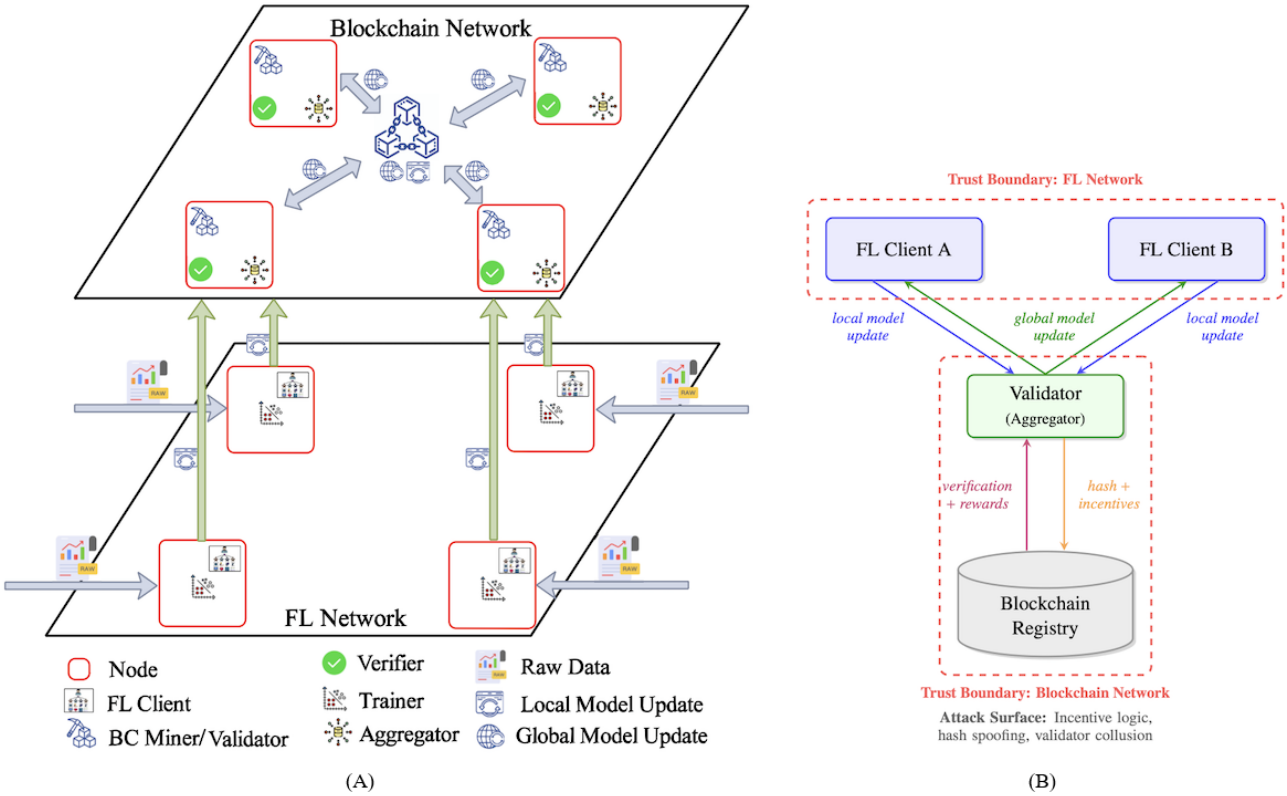

Loosely-Coupled Architecture

In the loosely-coupled architecture, blockchain nodes and FL clients are in 2 distinct networks. This architecture introduces reputation as a criterion for measuring the reliability of the clients. In this architecture, the primary function of the blockchain is to furnish a coordination mechanism for clients, manage their reputation, handle authentication, manage validation of local model updates, and manage incentives (ie, contributions are managed to ascertain reputation and incentivize participation). While the blockchain validates local model updates, it refrains from storing them. In contrast, it stores data related to the reputation of the participants. The responsibility of the FL clients is to train local models and upload updates to the blockchain for validation. After validation, these updates are sent to an aggregator, which can be a distinct server or a cloud space. A schematic of this architecture is depicted in Figure 10.

XSL•FO
RenderX

Figure 10. Loosely-coupled blockchain-based federated learning (FL) architecture: (A) detailed architecture; (B) schematic diagram. Central aggregator with blockchain used only for audit logs.

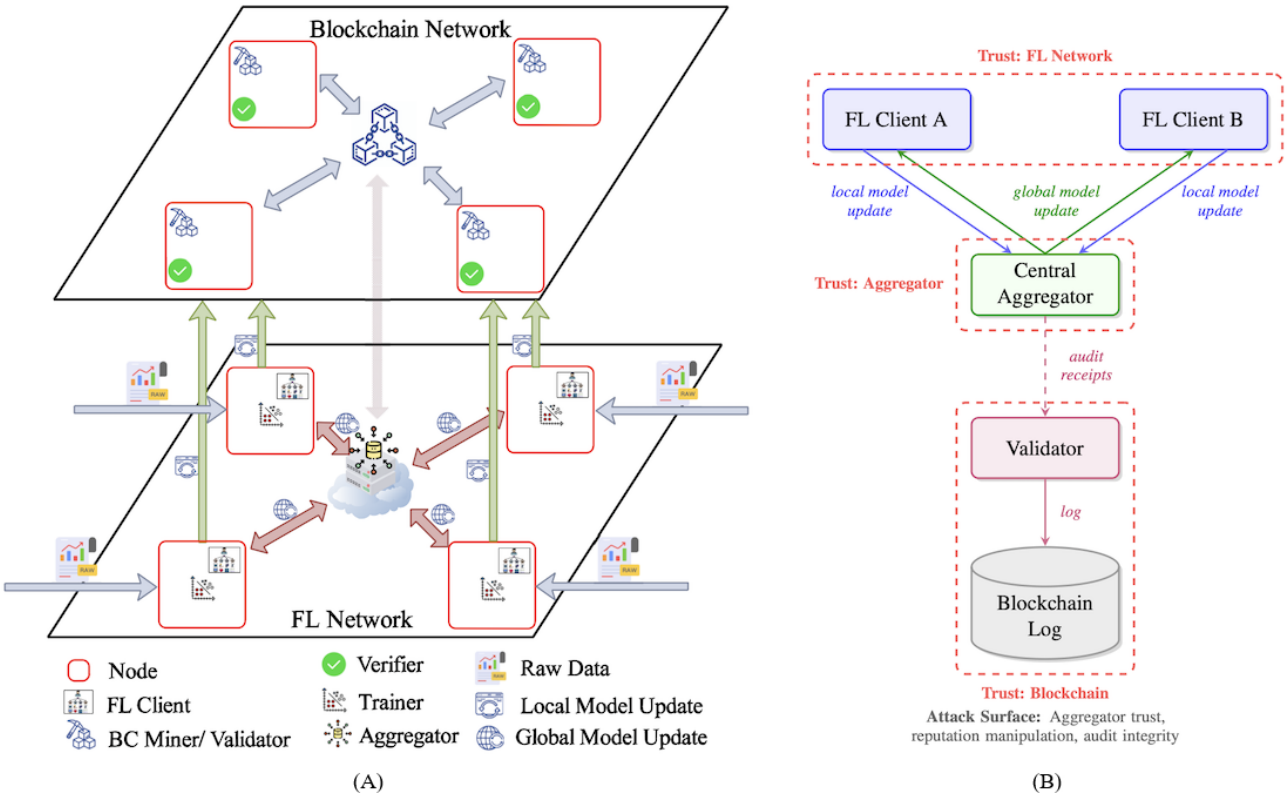

Cross-Architecture Synthesis and Quantitative Cost Analysis

The 3 BCFL integration patterns examined above represent distinct points on a design spectrum trading decentralization for performance and operational simplicity. The fully-coupled design maximizes auditability by executing all operations on-chain, but incurs significant latency (bounded by the blockchain finality time t_f) and ledger state growth, limiting scalability to small consortia (≤50 clients). The semicoupled design achieves a pragmatic balance by performing model aggregation off-chain while recording verifiable provenance and incentive information on-chain, supporting medium-scale health care deployments (≤100 clients). The loosely-coupled design minimizes blockchain interaction, using it primarily for coordination and audit, offering high throughput for large IoMT or regional networks (≥200 clients). Practitioners can map deployment size, regulatory constraints, and trust requirements to these design points to select an optimal integration pattern, as summarized in Table 2.

XSL•FO
RenderX

Table . Comparative characteristics and quantitative cost trends across blockchain-based federated learning integration architectures.[a]

Characteristic	Fully-coupled architecture	Semicoupled architecture	Loosely-coupled architecture
Trust model	Fully decentralized (all clients act as validators)	Partially decentralized (off-chain aggregation, on-chain proofs)	Federated trust (central aggregator with optional audit)
Latency per round	High (~t_f[b]; consensus limited)	Medium (off-chain aggregation, on-chain logging)	Low (minimal blockchain interaction)
Throughput/scale	≤50 clients	≤100 clients	≥200 clients
Compute overhead Δ_{BC}	12%-15% (cryptographic + consensus tasks)	5%-8% (hashing + verification)	2%-4% (logging + coordination)
Network cost pattern	Full model updates transmitted on-chain	Partial metadata and incentive records	Receipts and coordination only
On-chain storage/round	C[c]q[d]$\|W\|$[e] (hundreds of MBs)	10-100 kB	C×0.3 to 1 kB
Auditability	Very high (complete lifecycle recorded)	High (key model lineage and incentives)	Moderate (hash receipts only)
Best use	High-security, small-scale consortia	Balanced health care networks	Large-scale IoMT[f] or regional systems

[a]Typical health care configuration: clients per round (C) = 50, model size ($|W|$) = 80 MB (FP32), compression ratio (q) = 0.1, global rounds (R) = 200, finality time (t_f) = 1 to 10 s, client churn rate (λ) ≈ 0.05.

[b]t_f: finality time.

[c]C: clients per round.

[d]q: compression ratio.

[e]$|W|$: model size.

[f]IoMT: Internet of Medical Things.

The quantitative cost model formalizes these tradeoffs. The per-client computational load combines local training and blockchain overhead as follows: $\text{FLOPs}_{\text{client}} \approx E\, n\, f\,(1 + \Delta_{BC})$, where E represents local training epochs per round, n represents samples in the local dataset, f represents floating-point operations per sample, and Δ_{BC} represents blockchain overhead from cryptographic signing, consensus participation, and data serialization. Empirical studies report Δ_{BC} values of approximately 2%-4% for the loosely-coupled design, 5%-8% for the semicoupled design, and 12%-15% for the fully-coupled design [339,344].

Network usage grows linearly with communication rounds as follows: $\text{Bytes}_{\text{client}} \approx 2q\,|W|\,R(1 + \lambda)$, where $|W|$ represents model parameter size, q represents the compression ratio ($0<q\leq1$), R represents total rounds, and λ represents the client churn rate. The factor of 2 accounts for upload and download traffic. Storage requirements differ sharply for the loosely-coupled design (C×0.3 to 1 kB/round), semicoupled design (10 to 100 kB/round), and fully-coupled design ($Cq|W|$/round). For typical health care configurations, fully-coupled deployments require approximately 400 MB on-chain per round and approximately 3.2 GB total transfer per client, which are considered impractical for most BFT blockchains [302,345], whereas semicoupled and loosely-coupled designs reduce bandwidth by 40%-50% with a submegabyte on-chain state. Round latency is lower-bounded by finality time t_f, which increases with validator set size N_v due to $O(N_v^2)$ message complexity of the BFT consensus [345,346]. Consequently, the fully-coupled design suits small, high-security networks, while the loosely-coupled design offers practical scalability for large federations.

Assumptions and Scope

All values are order-of-magnitude estimates derived under synchronous federated averaging [347] and BFT blockchain configurations (eg, Hyperledger Fabric [302] and Tendermint [348]). While real-world performance may vary with model architecture, bandwidth, or consensus protocol, the relative scaling behavior and architectural tradeoffs among computation, communication, and on-chain costs remain consistent across implementations, including asynchronous or adaptive variants.

Security and Privacy Considerations

Overview

Integrating FL and blockchain addresses security and privacy challenges inherent in collaborative health care learning. While FL enhances privacy by training models on local datasets, it introduces vulnerabilities like model poisoning, gradient leakage (potential for patient reidentification), and coordination failures [349].

Blockchain mitigates these risks via its immutable, decentralized framework, strengthening FL through the following: (1) *immutable audit trails*, every model update, contribution, and access event is cryptographically recorded and timestamped on the blockchain, creating a verifiable and transparent history for forensic investigation [350]; (2) *decentralized trust and consensus*, blockchain eliminates reliance on a central authority, distributing trust across nodes to enhance resilience against single points of failure; and (3) *cryptographic security*, mechanisms ensure data integrity, authenticity, and nonrepudiation for all transactions and model updates, reinforcing the overall trustworthiness.

XSL•FO
RenderX

To further enhance privacy and compliance in BCFL systems, the following advanced cryptographic and security techniques are used: (1) ZKPs, (2) HE, (3) SMPC, (4) trusted execution environments (TEEs), and (5) DP. First, ZKPs enable participants to prove compliance (eg, data quality and consent) without revealing underlying sensitive patient data, allowing verifiable protocol adherence while safeguarding privacy [351,352]. However, ZKPs face substantial computational challenges, limiting practical applicability in real-time clinical systems [353,354]. Second, HE permits computation on encrypted data, enabling secure aggregation of model updates [352,355]. Although improved, HE introduces computational overhead (eg, 3.7 times to 8.2 times in health datasets) that presents scalability concerns and requires specialized hardware in resource-constrained settings [356]. Third, SMPC enables joint computation of aggregate functions (eg, model averaging) over private inputs without revealing individual contributions, offering strong privacy during aggregation [357,358]. SMPC protocols introduce significant computational and communication overhead through intensive cryptographic operations, posing a challenge in bandwidth-constrained health care networks [359,360]. Fourth, regarding TEEs, hardware-based secure enclaves provide isolated computation for sensitive operations (eg, aggregation) with cryptographic attestation [361,362]. However, TEEs suffer from memory limitations necessitating layer-wise training (performance overhead) and remain vulnerable to side-channel attacks (eg, cache timing and speculative execution) that can compromise confidentiality [363,364]. Fifth, DP introduces controlled noise to model updates or gradients for quantifiable privacy guarantees [365]. DP introduces critical privacy-utility tradeoffs (eg, 1% - 3% accuracy reduction in COVID-19 models), and the required privacy budget calibration is challenging, especially in non-independent and identically distributed health care data settings where noise and heterogeneity can slow model convergence [366,367].

Threat Model for BCFL in Health Care

We define a comprehensive threat model for BCFL systems in health care, identifying adversaries, capabilities, attack objectives, and mitigation strategies. This model provides a formal foundation for understanding security guarantees and tradeoffs, enabling practitioners to select appropriate defenses for specific deployment scenarios.

Adversaries and Capabilities

The BCFL ecosystem faces four adversary classes as follows: (1) *malicious clients*, these are compromised health care institutions or IoMT devices executing data poisoning, model poisoning [6,368], free-riding, or inference attacks with full access to local training processes; (2) *malicious aggregators*, these are present in loosely-coupled or semicoupled architectures, and they may infer private data from updates, bias global models through selective exclusion, or manipulate aggregation; (3) *Byzantine validators*, they comprise colluding blockchain nodes approving invalid updates, double-spending incentive tokens, or disrupting consensus; and (4) *external adversaries*, they include passive eavesdroppers intercepting communications, performing traffic analysis, or executing membership inference attacks.

Attack Objectives

Adversaries pursue distinct objectives through various attacks as follows: (1) *integrity attacks*, corrupt model reliability through model poisoning [6,368,369], backdoor insertion [370,371], and Sybil attacks, potentially causing misdiagnosis; (2) *privacy attacks*, extract patient data via gradient inversion [372-374], membership inference [375,376], and model inversion, violating HIPAA and GDPR guidelines; (3) *availability attacks*, disrupt services through denial-of-service on FL coordination, blockchain consensus, or participants, impacting clinical decision support; and (4) *fairness attacks*, introduce demographic biases or exclude underrepresented populations [377], exacerbating health care disparities.

Threat Mitigation

Table 3 maps threats to mitigation techniques with quantified performance implications from the Empirical Validation and Performance Evaluation section. These mitigations provide complementary protection layers deployable based on application requirements. Byzantine-robust aggregation [368,378] maintains integrity with up to 30% compromised participants but incurs 3%-5% accuracy degradation. HE provides semantic security with 3.7 to 8.2 times computational overhead [379]. DP [380] offers formal (,δ)-privacy guarantees with 1%-3% accuracy reduction, as observed in COVID-19 models.

XSL•FO
RenderX

Table . Threat mitigation mapping for blockchain-based federated learning in health care.

Threat	Mitigation	Security guarantee	Cost
Model poisoning	Byzantine-robust aggregation (Krum, FLTrust) [368]; Reputation systems	Resilient to ≤30% malicious clients	3%-5% accuracy drop
Gradient leakage	HE[a] [352]; SMPC[b] [379]	Semantic security for updates	HE: 3.7 to 8.2 times overhead
Membership inference	DP[c] [380,381]; Gradient masking	(,δ)-privacy	1%-3% accuracy loss
Free-riding	Proof of federated work [382,383]; Tokenized incentives	Verifiable contribution	On-chain verification overhead
Model inversion	ZKPs[d] [351]; TEEs[e]	Verifiable computation	ZKPs: 8-15 s proof time
Consensus attacks	BFT[f] protocols (PBFT[g], IBFT[h]) [300,345]	Tolerates less than one-third Byzantine nodes	$O(N^2)$[i] message complexity

[a]HE: homomorphic encryption.
[b]SMPC: secure multiparty computation.
[c]DP: differential privacy.
[d]ZKPs: zero-knowledge proofs.
[e]TEEs: trusted execution environments.
[f]BFT: Byzantine fault tolerance.
[g]PBFT: practical Byzantine fault tolerance.
[h]IBFT: improved Byzantine fault tolerance.
[i]$O(N^2)$: quadratic time complexity.

Architectural Security Tradeoffs

The 3 BCFL architectures provide distinct security-performance tradeoffs. Fully-coupled architectures maximize security through on-chain execution but incur approximately t_f latency per round and 12%-15% overhead, and they are suitable for high-security consortia with stringent audit requirements. Semicoupled architectures balance security and performance through selective on-chain logging, maintaining strong auditability with 5%-8% overhead and supporting up to 100 participants. Loosely-coupled architectures optimize scalability with minimal blockchain interaction, and they are suitable for large-scale IoMT deployments but require stronger trust in the central aggregator.

Empirical validation shows that these layered mitigations maintain accuracy within 2%-3% of nonprivate baselines. Combining DP (<0.1) with encrypted updates sustains utility, with an approximately 2.8 times compute overhead, while Byzantine robust aggregation maintains less than 5% degradation with up to 30% compromised clients. Mitigation selection should be guided by regulatory compliance (HIPAA/GDPR), data sensitivity, and clinical impact on patient outcomes.

Practical Deployment Challenges

The integration of privacy-enhancing technologies into health care involves technical intricacies regarding compatibility, efficiency, and communication overhead [12]. FL requires uniform infrastructure and capabilities across all sites, and unreliable infrastructure can disrupt training [384]. Cross-institutional coordination is amplified by complex data governance, regulatory requirements (eg, HIPAA and GDPR), and heterogeneous resource constraints [344]. Methodological flaws (eg, privacy issues, generalization, and communication costs) have rendered most recent FL health care studies inappropriate for clinical use [385].

Integration architecture significantly impacts tradeoffs as follows: (1) the fully-coupled architecture provides maximum auditability by logging all FL operations on-chain (highest security) but faces scalability challenges due to blockchain throughput limitations and consensus inefficiency, particularly with traditional PoW or basic PoS [344,386]; (2) the semicoupled architecture balances security and performance by selectively recording critical events (eg, model updates) on-chain while keeping routine operations off-chain; and (3) the loosely-coupled architecture offers the greatest scalability by using blockchain primarily for coordination and incentives, relying on off-chain security for the learning process, but it requires additional trust assumptions. Emerging architectures utilize sidechains to address scalability by enabling parallel processing and faster verification [386]. Addressing implementation challenges requires continued advancement in specialized hardware acceleration (eg, DARPA's FHE-focused DPRIVE project [387]), adaptive privacy mechanisms, and standardized protocols for cross-institutional deployments.

Operational Resilience and Failure Modes

Beyond cryptographic security, BCFL operational resilience requires predefined failure protocols with consortium-approved, measurable triggers. We define the following policy parameters: τ_{quorum} (minimum client quorum), Θ_{round} (round deadline), τ_{byz} (maximum tolerated Byzantine clients for robust aggregation [388]), δ_{AUROC} (maximum allowed area under the receiver operating characteristic curve drop), δ_{ECE} (maximum allowed expected calibration error increase), τ^{major}_{PSI} (population stability index bound for major drift), τ_{txfail} (transaction failure rate

XSL•FO
RenderX

bound), and ε_{post} (postupgrade error rate threshold). All thresholds are established during pilot deployments, documented in consortium standard operating procedures (SOPs), and calibrated to each clinical task's criticality and site capabilities. Table 4 summarizes these operational playbooks, specifying for each failure mode its trigger condition, automated response procedures, and monitoring signals, and the associated on-chain audit events. These playbooks link directly to the verifiability and auditability layer discussed in the Auditability and Verifiability in BCFL section, ensuring that every remediation event is immutably logged and cryptographically attestable on-chain.

Table . Operational failure protocols for blockchain-based federated learning in health care (thresholds set via governance).

Failure mode	Trigger condition	Response playbook	Monitoring signals	Audit events
Client dropout	Quorum < $_{quorum}$[a] after Θ_{round}[b]; Byzantine clients > $_{byz}$[c]	Proceed with partial aggregation using Byzantine-robust methods [388]; checkpoint current state; suspend chronic offenders; re-enroll upon verification	Participation rate, round latency, missing sites, straggler ratio	Round degraded, site suspended, site reenrolled
Model drift	AUROC[d] drop > AUROC[e]; ECE[f] increase > ECE[g]; PSI[h] > τ^{major}_{PSI}[i]	Pause promotion; perform shadow evaluation; rollback if unresolved; initiate root cause analysis	AUROC by cohort, ECE by cohort, PSI, drift alerts	Drift alert, model freeze, model promote, model rollback
Contract upgrade	Security vulnerability, policy change, or tx[j] failure rate > txfail[k]	Execute time locked multisig proxy upgrade; perform staged rollout; autorevert if postupgrade errors > post[l]	tx failure rate, version mismatch, event gaps	Upgrade proposed, upgrade executed, upgrade reverted
Key compromise	Unexpected signer behavior, attestation failure, or CRL[m] entry detected	Revoke key; rotate via threshold MPC[n] [389]; reattest using HSM[o]/TEE[p]; reaggregate any tainted data	Signature verify fail rate, unexpected signer count, anomalous payouts	Key revoked, key rotated, site re-enrolled

[a]τ_{quorum}: minimum client quorum.

[b]Θ_{round}: round deadline.

[c]τ_{byz}: maximum tolerated Byzantine clients for robust aggregation.

[d]AUROC: area under the receiver operating characteristic curve.

[e]δ_{AUROC}: maximum allowed AUROC drop.

[f]ECE: expected calibration error.

[g]δ_{ECE}: maximum allowed ECE increase.

[h]PSI: population stability index.

[i]τ^{major}_{PSI}: PSI bound for major drift.

[j]tx: transaction.

[k]τ_{txfail}: transaction failure rate bound.

[l]ε_{post}: postupgrade error rate threshold.

[m]CRL: certificate revocation list.

[n]MPC: multiparty computation.

[o]HSM: hardware security module.

[p]TEE: trusted execution environment.

Case Study: Multi-Institutional Brain-Tumor Segmentation

To concretely illustrate these operational protocols, we consider BCFL deployment for glioblastoma segmentation across 3 academic medical centers (hospitals A, B, and C) utilizing the federated tumor segmentation benchmark [390]. Hospital A serves as the semicoupled aggregator in this consortium. The workflow encompasses the following: (1) *data types and preparation*, where each institution contributes multimodal DICOM MRI scans (T1, T1Gd, T2, and FLAIR sequences) accompanied by expert-annotated segmentation masks for enhancing, necrotic, and edema tumor subregions; (2) *consent and governance*, managed through smart contracts encoding tiered patient permissions that allow model training, permit secondary research (opt-in), and prohibit commercial use; (3) *training execution*, where each hospital trains a local 3D U-Net model on its data, computing gradient hashes $h(\nabla W_i)$ submitted to the blockchain via the ModelUpdate smart contract after each federated round; and (4) *operational resilience*, demonstrated when hospital B experiences a network outage mid-round—the system automatically proceeds using the quorum threshold $\tau_{quorum} = 0.67$ (the quorum threshold $\tau_{quorum} = 0.67$ ensures BFT consistent with standard BFT safety guarantees [$n \geq 3f + 1$],

XSL•FO
RenderX

allowing global aggregation to proceed only when at least two-thirds of clients contribute authenticated updates), and after 3 consecutive missed rounds, hospital B is suspended until manual infrastructure verification. This end-to-end scenario demonstrates how BCFL protocols maintain continuity in collaborative clinical research despite real-world operational challenges.

Regulatory-Compliant Integration

Overview

The integration of blockchain with FL in health care environments requires strict adherence to comprehensive regulatory frameworks governing patient data protection, medical device validation, and cross-jurisdictional data exchange. This section analyzes architectural approaches for achieving regulatory compliance while preserving the security and privacy benefits of BCFL.

Regulatory Framework Overview

Health care data processing operates under multiple regulatory frameworks with stringent requirements for data handling, storage, and sharing. In the United States, HIPAA [391] establishes comprehensive privacy and security standards for protected health information (PHI), mandating administrative, physical, and technical safeguards [15]. Similarly, the GDPR [392] in the European Union requires explicit consent for data processing, adherence to data minimization principles, and provision of data portability and erasure rights. These regulatory requirements create significant challenges for conventional centralized ML approaches, often requiring detailed audit trails, comprehensive data lineage tracking, and the ability to selectively remove individual patient contributions from trained models [393].

The proliferation of AI and ML in medical applications has introduced additional regulatory considerations. The FDA framework for AI/ML-based medical devices emphasizes transparent model development, continuous performance monitoring, and the ability to track model behavior across diverse patient populations [394]. These requirements align well with BCFL capabilities, where immutable audit trails and decentralized validation mechanisms provide the transparency and accountability necessary for regulatory compliance in distributed settings [394].

Table 5 maps these key regulatory obligations to concrete FL and blockchain design mechanisms, highlighting how BCFL architectures can be engineered to support compliance across HIPAA, GDPR, and FDA requirements.

XSL•FO
RenderX

Table . Mapping of HIPAA[a], GDPR[b], and FDA[c] requirements to FL[d] and blockchain design mechanisms.

Guideline and key requirement	FL plus blockchain design mechanisms
HIPAA	
Privacy: PHI[e] confidentiality; Security: electronic PHI, integrity/availability; Minimum necessary use	FL data locality, encrypted model updates, blockchain stores only hashed metadata (not PHI); Permissioned blockchain with BFT[f] consensus, tamper-evident logs, redundant FL storage; Local feature selection, DP[g] and gradient clipping, blockchain smart contracts enforce data policies
Audit controls	Immutable blockchain ledger for access/update logs, fine-grained access control lists, cryptographic signatures
Transmission security	TLS[h] for FL channels, authenticated blockchain overlays, health care public key infrastructure integration
GDPR	
Lawfulness, fairness, transparency; Purpose limitation; Data accuracy; Right to erasure	Blockchain consent smart contracts, transparent logging, explainable FL models; Task-specific FL models, local data filtering, blockchain policies restrict model reuse; Cross-institutional FL validation, blockchain model lineage, rollback capability; Off-chain personal data, blockchain pseudonymous references, key revocation and model forgetting
Privacy by design	FL with DP/secure aggregation, minimal on-chain data, role-based blockchain permissions
Data protection impact assessment	Blockchain logs as impact assessment evidence, threat modeling, anomaly detection for poisoning/leakage
FDA	
Software as a medical device lifecycle control; Good machine learning practice and validation	Blockchain-anchored model versioning, immutable training/deployment records; Federated evaluation across sites, blockchain-backed data/model provenance
Postmarket monitoring	Continuous FL updates, blockchain logs for drift/adverse events, auto-rollback triggers
Change management	Blockchain registry for approved versions, on-chain governance for model rollout
Cybersecurity integrity	Signed FL/blockchain nodes, remote attestation, blockchain-based software bill of materials/model bill of materials tracking

[a]HIPAA: Health Insurance Portability and Accountability Act.
[b]GDPR: General Data Protection Regulation.
[c]FDA: Food and Drug Administration.
[d]FL: federated learning.
[e]PHI: protected health information.
[f]BFT: Byzantine fault tolerance.
[g]DP: differential privacy.
[h]TLS: transport layer security.

Compliance Challenges in FL

Traditional FL implementations face several compliance challenges in health care environments [395]. Data governance becomes complex when training spans multiple institutions with different privacy policies and jurisdictional requirements. Cross-border data sharing restrictions, exemplified by national data localization laws, can limit the scope of federated collaborations [396]. The distributed nature of FL can complicate compliance auditing, as conventional centralized monitoring approaches are inadequate for tracking model

XSL•FO
RenderX

updates and validating data usage across multiple participating nodes. Model explainability requirements present another challenge, given increasing regulatory demands for transparent decision-making processes in clinical AI systems [397]. The aggregated nature of federated model updates can obscure individual data source contributions, hindering the provision of detailed explanations required for regulatory compliance and clinical validation.

Blockchain-Enhanced Compliance Architecture

Blockchain technology provides fundamental mechanisms that directly address regulatory compliance requirements in federated health care systems. The immutable distributed ledger creates comprehensive audit trails that record all model updates, participant contributions, and data access events, satisfying requirements for detailed recordkeeping and accountability. Smart contracts can automate compliance verification by embedding regulatory requirements directly into the blockchain protocol, ensuring that data sharing and model training activities automatically conform to predefined privacy and security standards.

The 3 integration architectures offer varying degrees of compliance capabilities. First, for the fully-coupled architecture, all FL operations are recorded on the blockchain ledger, providing maximum transparency and auditability. Each model update is cryptographically signed and timestamped, creating an immutable record of the entire training process. This architecture suits environments requiring the highest levels of regulatory oversight but may face scalability constraints with high-frequency model updates. Second, the semicoupled architecture is a hybrid approach that selectively records critical compliance events on the blockchain while maintaining routine operations off-chain. Compliance-relevant activities, such as participant authentication, consent management, and model validation, are tracked on-chain, while frequent model updates occur off-chain with periodic on-chain checkpoints. This balance provides robust compliance capabilities while addressing scalability concerns. Third, for the loosely-coupled architecture, the blockchain primarily serves as a coordination and reputation management layer, with detailed compliance tracking implemented through conventional audit mechanisms. This approach offers maximum flexibility for integration with existing health care IT infrastructure but requires careful design to ensure adequate regulatory compliance.

Auditability and Verifiability in BCFL

Beyond confidentiality and integrity guarantees, BCFL must also provide transparent mechanisms for independent verification and traceability of all learning activities. To ensure transparent operations, verifiability and auditability in BCFL are defined as complementary properties that enable independent validation of all training activities without compromising privacy. Verifiability denotes the ability of any authorized entity to cryptographically confirm the authenticity and integrity of model updates, aggregation results, and participating institutions through signed digital evidence recorded on-chain. Auditability extends this notion by enabling authorized auditors to reconstruct and verify the entire lifecycle of the learning process (including the provenance of training rounds, participant consent scope, and compliance checks) using immutable ledger entries and corresponding off-chain records.

In the proposed framework, each local model update (ΔW_i) is hashed (eg, SHA-256) and digitally signed by the participant before being submitted to the blockchain. The on-chain record contains the following: (1) the participant's pseudonymous identifier, (2) the signed update hash, (3) the round number and timestamp, (4) a consent-proof token (linking to consent metadata stored off-chain), and (5) the result of automated policy-compliance checks executed via smart contracts. Full model parameters, gradients, raw medical data, and detailed computational logs remain off-chain within secure institutional storage to preserve privacy and reduce blockchain overhead. This hybrid design allows auditors to verify consistency between on-chain commitments and off-chain artifacts, providing cryptographically verifiable evidence of model provenance, policy adherence, and regulatory compliance while maintaining confidentiality of protected health information.

Case Study: Regulatory Verification

Building upon brain-tumor segmentation deployment, a health care regulator conducts comprehensive compliance verification through systematic analysis of on-chain evidence across the following four critical dimensions: (1) *consent verification*, achieved by querying the ConsentRegistry smart contract to cryptographically validate that all training data possess active consent tokens with appropriate scope limitations and no revocation events during the model training period; (2) *model provenance audit*, tracing the complete lineage of the global segmentation model through immutable, timestamped update records that capture each hospital's contribution sequence and aggregation proofs, thereby creating an auditable chain of custody from raw gradients to the final model; (3) *policy compliance validation*, verifying through smart contract execution logs that every model update underwent automated HIPAA security rule validation and Institutional Review Board (IRB) requirement checks prior to integration into the global model; and (4) *data governance assurance*, confirming through ZKP attestation that all patient data processing remained within institutional boundaries, with only cryptographic commitments—never raw medical images or sensitive patient information— traversing the network. This comprehensive verification process, fully reconstructible from blockchain evidence, demonstrates how the semicoupled BCFL architecture (Figure 9) enables rigorous regulatory oversight. The on-chain audit trail provides concrete verification artifacts, as follows:

Block #18345: $\text{hospital}_A \rightarrow h(\nabla W_A)$ + timestamp + consent proof + HIPAA check; Block #18346: $\text{hospital}_C \rightarrow h(\nabla W_C)$ + timestamp + consent proof + HIPAA check; Block #18347: aggregator $\rightarrow h(W_{\text{global}})$ + aggregation proof + ZKP attestation

This immutable record preserves patient privacy through cryptographic guarantees while maintaining institutional autonomy through verifiable local computation.

Cross-Border Compliance Considerations

International health care collaborations must navigate different regulatory requirements across jurisdictions [398]. BCFL

XSL•FO
RenderX

architectures can incorporate jurisdiction-specific compliance rules through modular smart contract designs. Participants from different regulatory environments can implement their local compliance requirements while engaging in collaborative model training. The blockchain ledger can maintain separate compliance tracks for different jurisdictions [399], ensuring each participant's regulatory obligations are met while enabling global collaboration. Data localization requirements are addressed through FL's inherent data locality principle, where sensitive patient data remain within its originating jurisdiction. The blockchain component facilitates cross-border collaboration by managing model updates and coordination without requiring direct data sharing, satisfying both privacy regulations and data sovereignty requirements.

Implementation Guidelines for Regulatory Compliance

Successful deployment of regulatory-compliant BCFL requires careful attention to several implementation considerations. Organizations must establish clear data governance frameworks that define roles, responsibilities, and accountability measures for all participating entities [400]. Consent management systems must be integrated into the blockchain architecture to ensure patient permissions are tracked and honored throughout the FL process [12]. The technical infrastructure must support both the scalability demands of FL and the audit trail requirements of regulatory compliance [401]. This includes implementing robust key management systems for cryptographic operations, establishing secure communication channels between participating nodes, and designing backup and recovery procedures that maintain compliance during system failures. Regular compliance auditing must be embedded in the system design, with automated monitoring capabilities to detect potential regulatory violations and alert administrators to compliance issues. The blockchain's immutable audit trail provides the foundation for these monitoring systems, enabling continuous compliance verification and rapid response to potential violations.

Ethical and Patient-Centric Considerations

Beyond regulatory and technical safeguards, ethical considerations are essential for responsible BCFL deployment in health care. The decentralized nature of FL and blockchain complicates the acquisition and management of patient consent across autonomous institutions with diverse governance frameworks [402,403]. Traditional consent models, designed for single-institution data use, are inadequate when patient data contribute to shared models with different stakeholders and objectives [402]. To address this, consent management should leverage verifiable cryptographic mechanisms, such as consent tokens, ZKPs, and smart contract–based registries, to enable patients to grant, track, and revoke permissions without disclosing sensitive data [12,404]. These mechanisms uphold autonomy and transparency while aligning with evolving interpretations of HIPAA and GDPR.

Blockchain immutability also conflicts with the GDPR's "right to be forgotten" (Article 17) [405,406]. To reconcile this, BCFL designs should store personally identifiable data off-chain while recording only hash-based consent attestations and model metadata [407,408]. This preserves auditability and enables consent withdrawal without breaching immutability.

Fairness is equally critical, as institutions with greater computational capacity or larger datasets may dominate model updates, biasing outcomes against underrepresented groups [409,410]. Mitigation requires fair aggregation [403], representative datasets [411], and fairness-aware or distributionally robust optimization ensuring consistent performance across demographics [412,413].

Ethical governance mechanisms, including IRBs [414,415] and data ethics councils [416,417], should oversee consent processes, fairness evaluations, and the protection of vulnerable populations [414,418]. Clear accountability across AI developers, health care providers, and blockchain participants is vital [419]. Finally, explainable AI approaches [420-422] must ensure that model decisions are interpretable by clinicians and understandable to patients, reinforcing trust and autonomy. Ethical BCFL thus requires a shift from a compliance-driven design to a patient-centric design, where transparency, fairness, and accountability become foundational principles for building trust and ensuring sustainable adoption of collaborative health care AI systems [402,423].

Ethics and Patient Voice

Ethical deployment of BCFL in health care requires patient-centric design, translating consent, autonomy, and transparency into actionable system features. Consent should evolve from binary opt-in to granular, preference-based models, allowing patients to specify which data types (eg, EHRs, genomic data, and imaging data) and purposes are permitted through layered interfaces [424,425]. A revocation mechanism records consent withdrawal as an on-chain event, preventing further participation. While revocation cannot retroactively remove previously aggregated updates, federated unlearning protocols mitigate historical influence by efficiently retraining global models without revoked participants [426,427]. Meaningful transparency extends beyond cryptographic verifiability to interpretable communication, and systems should notify contributors when models trained on their data are deployed and provide accessible indicators of model performance, bias, and fairness. When BCFL outcomes inform clinical decisions, explainable AI (XAI) techniques (eg, Shapley Additive Explanations [SHAP] and confidence decomposition) support instance-level explanations that bridge distributed model complexity with clinician and patient understanding [422,428,429].

Embedding granular consent, verifiable revocation, and interpretable transparency within the BCFL architecture ensures that distributed health care AI respects patient autonomy; aligns with regulatory obligations under HIPAA, GDPR, and the FDA's AI/ML-based SaMD action plan [430]; and fosters sustained public trust in collaborative health systems [431].

Future Regulatory Developments

The regulatory landscape for AI in health care continues to evolve, with new guidelines and requirements emerging regularly. BCFL architectures must be designed with flexibility to accommodate changing regulatory requirements without

XSL•FO
RenderX

requiring a complete system redesign. The modularity of smart contract–based compliance frameworks provides this adaptability, enabling organizations to update their compliance mechanisms as regulatory requirements evolve. Emerging regulatory trends, such as algorithmic accountability requirements and mandatory AI explainability, are well-suited to blockchain-based approaches that provide transparent audit trails and verifiable model development processes. As health care organizations increasingly adopt AI and ML technologies, the compliance advantages of BCFL will become increasingly important for regulatory acceptance and clinical adoption.

Empirical Validation and Performance Evaluation

Overview

This section grounds our BCFL integration in evidence from clinical deployments, consortia-scale studies, and controlled benchmarks. We synthesize outcomes along four axes: (1) accuracy versus centralized baselines; (2) scalability and latency/throughput under permissioned consensus; (3) robustness and privacy under realistic threat models; and (4) compliance, provenance, and operational constraints. Where possible, we report representative ranges from cited studies rather than single-point claims.

Large-Scale Clinical and Multi-Institutional Evidence

Table 6 summarizes emblematic deployments. Across drug discovery (MELLODDY [253]), neuro-oncology (federated tumor segmentation [390]), pandemic diagnostics [432], and IoMT monitoring [251,433], BCFL consistently delivers accuracy within approximately 2%-3% of centralized training while providing auditability and coordination via permissioned chains or distributed ledgers.

Table . Representative blockchain-based federated learning deployments in health care.

Setting/cohort	Scale	Task	Blockchain role	Accuracy	Latency/overhead	Source
MELLODDY (10 pharma companies)	>2.6 billion data points, 21 million compounds	Drug prediction	Distributed ledger for auditability, consensus, traceability	Median ~2% RIPtoP[a] gain	Comparable to centralized	[253]
FeTS[b] (71 sites)	71 international sites	Brain tumor segmentation	Coordination and registry	Within 2% - 3% of centralized	Real-world federation	[434,435]
COVID-19 consortium	Multinational	CT[c] scan diagnosis	Aggregation logging, verification	94.2% versus 95.1% centralized	~3.2 s/round	[433]
IoMT[d] monitoring	50 patients	Anomaly detection	Consent, notarization	92.7%	<5 s for alerts	[436,437]
EHR[e] management	Multi-institutional	Clinical analytics	Model validation, audit trails	95.2%	<150 ms blockchain latency (10,000 transactions)	[344]

[a]RIPtoP: relative improvement of proximity to perfection.

[b]FeTS is primarily a federated learning platform; blockchain integration is demonstrated in other health care applications listed here.

[c]CT: computed tomography.

[d]IoMT: Internet of Medical Things.

[e]EHR: electronic health record.

Performance and Scalability Benchmarks

Controlled experiments report stable training under 50 - 200 participants with permissioned consensus, and recent empirical studies demonstrate that blockchain-enabled FL can reduce communication overhead by 40% - 60% through communication-aware aggregation schemes [339,344]. Specifically, there was a 43% communication overhead reduction and a 37% lower computational cost while maintaining 95.2% model accuracy [344]. Sharded or committee-based designs sustain 80-120 TPS at the consortium scale while keeping the consensus overhead under 5% for ≤50 nodes, and at 100 nodes, PBFT-style protocols exhibit 12%-15% overhead [438,439].

Security, Privacy, and Robustness

Against gradient inversion and membership inference, combining DP (eg, <0.1) with encrypted or masked updates sustains utility within 3% - 5% of nonprivate baselines at an approximately 2.8 times compute overhead, while Byzantine-robust aggregation maintains <5% degradation under up to 30% compromised clients [19,352]. Recent work has demonstrated that blockchain-enabled frameworks maintain robustness against multiple adversarial attack vectors with accuracy levels above 93% [344]. For verifiable compliance, ZKPs amortize to proof generation of 8 - 15 s and verification of 2 - 4 s per minute-scale update [440].

XSL•FO
RenderX

Compliance, Provenance, and Operational Considerations

Empirical audits indicate technical adherence to approximately 93% of HIPAA/GDPR provisions, with 70% - 80% manual audit effort reduction when provenance is automated via smart contracts [395]. For FDA transparency obligations on AI/ML devices, immutable lineage reduces evidence cycles from an initial 6 - 9 months to 2 - 3 months in reported pipelines [397]. Operational deployment of blockchain-enabled FL systems faces several infrastructure challenges, including maintaining high system availability, managing blockchain storage growth as institutions scale, balancing computational overhead from cryptographic operations, and ensuring adequate network bandwidth for model parameter exchange [9,441].

Architectural Tradeoffs

In permissioned health care networks, PoA achieves 200-300 TPS and approximately 95% lower energy than PoW, whereas PBFT offers stronger Byzantine resilience at 40% - 50% lower throughput [442]. Comparative syntheses show that BCFL improves privacy metrics by 40% - 60% over vanilla FL while staying within approximately 3% of centralized accuracy [344,384]. Recent systematic reviews covering literature from 2023 - 2024 confirm these patterns across diverse health care applications [12,400]. Fully-coupled blockchain-FL designs minimize latency (reduction of 30% - 40%) but demand 2 - 3 times the resources, while semicoupled architectures offer a balanced middle ground [443].

Threats to Validity and Replicability

Information

A crucial aspect of interpreting empirical studies on BCFL involves recognizing their inherent threats to validity. Specifically, researchers must account for the impact of various factors. First, non-independent and identically distributed data and silo imbalance are important aspects, where skewed label and cohort distributions across participating institutions necessitate the reporting of per-site metrics and federated confusion matrices. Second, WAN effects must be meticulously reported for wide-area deployments, including round times, network latency, bandwidth, and specific block/endorsement parameters, to avoid misleading claims based on LAN-only results. Third, attack realism requires results to be reproducible with adaptive rather than static poisoning, along with full disclosure of clipping rules, aggregation mechanisms, and defense strategies. Fourth, from a privacy perspective, privacy budgets must be transparent, demanding the publication of (,δ) values, sensitivity assumptions, and composition accounting across all training rounds. Fifth, achieving determinism is essential for reproducibility, requiring the release of seeds, exact model/optimizer configurations, client sampling methods, and chaincode/smart contract versions. Finally, accurate compute/power accounting must distinguish between the computational overhead dedicated to model training versus the overhead attributed to blockchain operations, detailing the validator count and hardware thermal design power (TDP). Recent comprehensive reviews have systematically identified these validity concerns across the literature from 2018 - 2024 [384,400].

Takeaway

Across clinically meaningful tasks, BCFL achieves near-centralized utility with measurable but manageable overheads when engineered with communication-efficient aggregation and permissioned consensus. Recent empirical evidence from 2024 - 2025 confirms consistent patterns: accuracy within 2% - 3% of centralized baselines, communication overhead reductions of 40% - 50%, and successful deployment across 50 - 200 participating institutions. The primary value additions in health care are auditability, provenance for regulatory processes, and robustness under adversarial and privacy constraints, while the primary bottlenecks are large-scale coordination (≥500 nodes) and legacy-system interoperability.

Related Work

Overview

In this section, we comprehensively explore existing research endeavors focusing on integrating BCFL frameworks within health care contexts. Initially, we delve into the corpus of literature dedicated to elucidating the practical applications, methodologies, and outcomes of utilizing blockchain technology in conjunction with FL for health care use cases. In addition, we aim to provide a taxonomy and categorize the existing work based on the insights garnered throughout this paper. To achieve this, we classify the research according to integration architecture (ie, fully coupled, semicoupled, and loosely coupled), blockchain platform, FL type, and data type. Furthermore, we highlight their primary contributions as well as their limitations. Later, we examine existing surveys and reviews focusing on blockchain-assisted FL in health care. This will provide insights into the overall research landscape and help identify any gaps or areas for further exploration.

Literature Review

Samantray and Reddy [444] proposed a hybrid blockchain architecture with quantum key encryption for Healthcare 5.0. Bhasker et al [445] presented a Healthcare 5.0 smart health care system for addressing confidentiality, trust, and compliance standards for safe health monitoring. Bhardwaj et al [428] introduced a privacy-preserving Federated Blockchain Explainable Artificial Intelligence Optimization (PPFBXAIO) framework for the integration of blockchain and XAI, and optimization techniques to ensure privacy, traceability, and robustness in FL-based systems. Ali et al [446] presented a ZKP and HE approach for EHR security. Das et al [447] presented a meta-learning approach for improved model generalization.

Chen et al [19] proposed a trustworthy and fair blockchain-based FL framework to address challenges like malicious attacks (model/data poisoning) and free-rider behavior. In another study [448], a quality-aware BCFL with a secure key-sharing mechanism to protect local model parameters and ensure data security was proposed. Mazid et al [436] proposed a blockchain-driven FL framework for secure health care services using IoMT devices. Liang et al [449] proposed a software

XSL•FO
RenderX

architecture integrating FL and blockchain to mitigate bias and fairness issues in health care predictive modeling while safeguarding patient privacy. Om Kumar et al [450] introduced a mechanism to reward organizations participating in the FL process, ensuring privacy-preserving model transfer using loosely-coupled integration with the Ethereum blockchain. Moulahi et al [437] integrated FL and blockchain to develop a trusted system for predicting diabetes risk while ensuring data privacy and model integrity.

Ali et al [451] focused on integrating blockchain with FL for secure and decentralized analysis of electronic medical records in precision medicine. Chang et al [442] proposed an integration of adaptive DP and gradient verification-based consensus protocols in a fully-coupled architecture for health care analytics. Lian et al [452] presented a blockchain-based personalized FL system for ensuring security and privacy in IoMT.

Farooq et al [453] developed an automated system for analyzing patients' live data within a fully-coupled architecture, while Zhang et al [454] proposed a blockchain-enabled FL framework for health care data privacy protection. Aich et al [336] introduced a blockchain-assisted FL framework for personal data preservation, and Passerat-Palmbach et al [335] proposed a novel architecture for FL integrated with blockchain within health care systems. A taxonomy of existing research studies, along with additional relevant work, has been compiled in Table 7.

XSL•FO
RenderX

Table . Related work on blockchain-assisted FL[a] in health care.

Reference	Main contribution	Architecture	Blockchain platform	FL type	Data type	Limitations
Chen et al [19]	Introduces a trustworthy and fair blockchain-based FL framework (FedCFB) to address challenges like malicious attacks (model/data poisoning) and free-rider behavior	Semicoupled	Custom-designed	HFL[b]	Unknown (MNIST)	No discussion on throughput/latency tradeoffs for large-scale FL deployments.
Samantray and Reddy [444]	Hybrid blockchain architecture with quantum key encryption for futuristic health care in smart cities	Loosely coupled	Ethereum	HFL	Medical images and documents	Performance and scalability issues in large-scale implementations.
Bhasker et al [445]	A health care 5.0 smart health care system for addressing confidentiality, trust, and compliance standards for safe health monitoring	Loosely coupled	Unknown	HFL	Wearable devices and sensors	Lack of real-world deployment scenarios and practical validation in actual health care settings.
Bhardwaj and Sumangali [428]	Proposes a privacy-preserving federated blockchain explainable artificial intelligence optimization (PPFBX-AIO) framework, which integrates blockchain, explainable artificial intelligence, and optimization techniques to ensure privacy, traceability, and robustness in FL-based systems	Fully coupled	Unknown	HFL	EHRs[c]	Real-world deployment gaps (ie, does not address practical deployment challenges, such as regulatory compliance with health care standards (HIPAA[d] or GDPR[e]) and integration with existing hospital information systems.
Ali et al [446]	Leverages zero-knowledge proofs and homomorphic encryption for EHR security	Semicoupled	Unknown	HFL	EHRs	Limited empirical validation or real-world testing of the proposed framework.
Das et al [447]	Meta-learning approach for improved model generalization	Semicoupled	Hyperledger Besu	HFL	IoMT[f] data	No real-world implementation (the entire study is simulation-based with no actual deployment and no real medical IoT[g] devices involved), incomplete blockchain implementation, and FL implementation is unclear.

XSL•FO
RenderX

Reference	Main contribution	Architecture	Blockchain platform	FL type	Data type	Limitations
Munusamy and Jothi [344]	Proposes an enhanced privacy-preserving blockchain-enabled federated learning (EPP-BCFL) framework that integrates blockchain with hybrid privacy mechanisms and intelligent aggregation strategies for secure EHR	Semicoupled	Hyperledger Fabric	HFL	Medical imaging data (CIFAR-10)	Evaluation limited to CIFAR-10 rather than clinical datasets.
Lo et al [455]	Proposes a BCFL[h] architecture to enable accountability in FL systems.	Loosely coupled	Ethereum	HFL	Medical imaging data (COVID-19 CT[i])	Lack of real-world deployment analysis.
Kumar et al [448]	Proposes a quality-aware BCFL with a secure key-sharing mechanism to protect local model parameters and ensure data security	Loosely coupled	Unknown	HFL	Unknown	Limited scale and real-world testing (eg, not tested in actual hospital environments, no real health care institutions involved, etc).
Mazid et al [436]	Proposes a blockchain-driven FL framework for secure health care services using IoMT devices	Loosely coupled	Unknown	HFL	Medical imaging (colon pathology, breast tumor)	Validated only through simulations, using benchmark datasets (2D colon pathology, breast tumor, and CIFAR-10). There is no real-world hospital or IoMT deployment, which limits conclusions.
Liang et al [449]	Software architecture integrating FL and blockchain to tackle bias and fairness issues in health care predictive modeling while safeguarding patient privacy	Semicoupled	Rahasak	Unknown	Unknown	Limited real-world validation. This study was conducted in a simulated environment with only 5 simulated medical centers rather than actual health care institutions. This significantly limits the generalizability of findings to real-world health care settings.
Om Kumar et al [450]	Introduces a mechanism to reward organizations participating in the FL process, ensuring privacy-preserving model transfer between users and organizations using blockchain	Loosely coupled	Ethereum	HFL	Medical imaging data (COVID-19 CT)	Not explicitly stated.

XSL•FO
RenderX

Reference	Main contribution	Architecture	Blockchain platform	FL type	Data type	Limitations
Moulahi et al [437]	Integrates FL and blockchain technology to develop a trusted system for predicting diabetes risk while ensuring data privacy and model integrity	Semicoupled	Ethereum	HFL	Sensor data (IoT)	Not explicitly stated.
Ali et al [451]	Integrates blockchain with FL to enable secure and decentralized analysis of EMRs in precision medicine	Loosely coupled	Ethereum, Hyperledger Fabric (simulation)	HFL	EMR	The paper appears to be primarily theoretical and simulation-based without actual deployment in real health care settings (eg, there is no evidence of testing with actual health care institutions or real EMR) and limited experimental validation (eg, no specific dataset characteristics are provided).
Chang et al [442]	Integrates adaptive differential privacy and gradient verification-based consensus protocols	Fully coupled	Ethereum	HFL/VFL[j]	IoMT sensor data	Scalability challenges, increased complexity, potential computational overhead, and the need for further validation across diverse medical conditions and datasets.
Lian et al [452]	Blockchain-based personalized FL system to address security and privacy concerns in IoMT	Fully coupled	Unknown (consortium PoS[k] blockchain)	VFL	IoMT sensor data (Fashion-MNIST)	Limited experimental validation (eg, only simulated experiments were used rather than real-world deployment, tested on Fashion-MNIST and CIFAR-10 [image classification datasets]; no actual medical data, etc).
Farooq et al [453]	Develops an automated system for analyzing patients' live data	Fully coupled	Ethereum	HFL	IoMT sensor data	Limited empirical validation or real-world testing of the proposed framework.
Zhang et al [454]	Proposes a BCFL framework for health care data privacy protection	Fully coupled	Unknown (conceptual)	HFL	EHR (MNIST)	Theoretical model only, not supported via experimental implementations.
Aich et al [336]	Proposes a blockchain-assisted FL framework for personal data preserving	Loosely coupled	Unknown (conceptual)	__[l]	No data	Reliance on conceptual assumptions without real-world application.

XSL•FO
RenderX

Reference	Main contribution	Architecture	Blockchain platform	FL type	Data type	Limitations
Passerat-Palmbach et al [335]	Proposes a novel architecture for FL integrated with blockchain within a health care system	Loosely coupled	Ethereum	—	No data	Not explicitly stated.
Rahman et al [438]	Proposes a BCFL framework for COVID-19 applications to classify IoHT[m] data	Loosely coupled	Ethereum	FL type	Sensor data (IoMT)	Not explicitly stated.
Singh et al [456]	Integrates blockchain technology with FL to enhance privacy preservation and scalability in health care data management	Fully coupled	Unknown	Unknown	Sensor data (IoT)	Theoretical model only, not supported via experimental implementations.
Nguyen et al [457]	Proposes a new blockchain-enabled FL-based generative adversarial network framework for secure COVID-19 data analytics	Semicoupled	Unknown (conceptual)	HFL	Medical imaging data (COVID-19 CT)	The blockchain model is conceptual, and a real-world implementation was not reported.
Liu et al [458]	Proposes a framework of blockchain-empowered FL in health care–based cyber-physical systems	Fully coupled	Ethereum	HFL	EHR (MNIST HAM10000)	Not explicitly stated.
Otoum et al [459]	Proposes a novel solution for revolutionizing health care systems by considering concepts like distributivity, self-learnability, and autonomy.	—	Unknown (conceptual)	Unknown	No data	Only introduces a theoretical framework without an empirical validation.
Kumar et al [433]	Blockchain-empowered method to detect patterns of COVID-19 from lung CT scans	Fully coupled	Unknown	HFL	Medical imaging data (COVID-19 CT)	The paper does not extensively discuss the generalization capabilities of the proposed model to handle variations in COVID-19 manifestations across different patients.
Durga and Poovammal [460]	A concise review of BCFL in health care, concepts, and taxonomy	—	Unknown	Unknown	Unknown	The article is very brief and does not cover all concepts.

XSL•FO
RenderX

Reference	Main contribution	Architecture	Blockchain platform	FL type	Data type	Limitations
Lakhan et al [461]	Proposes a FL-based blockchain-enabled task scheduling (FL-BETS) framework with different dynamic heuristics for health care applications	Loosely coupled	Unknown	Unknown	Sensor data (IoMT)	Dynamic and runtime unknown attacks against IoMT were not considered in this work.
Samuel et al [462]	Introduces FedMedChain as a BCFL framework for medical data privacy preservation	Loosely coupled	Unknown	Unknown	Sensor data (Unknown)	Not explicitly stated.
Yang and Xing [463]	Introduces a privacy protection framework for medical data using blockchain and FL for secure and auditable data sharing among medical institutions using a secure aggregation scheme based on homomorphic encryption	Loosely coupled	Ethereum	HFL	EHR (MNIST different datasets)	Not explicitly stated.

[a]FL: federated learning.
[b]HFL: horizontal federated learning.
[c]EHRs: electronic health records.
[d]HIPAA: Health Insurance Portability and Accountability Act.
[e]GDPR: General Data Protection Regulation.
[f]IoMT: Internet of Medical Things.
[g]IoT: Internet of Things.
[h]BCFL: blockchain-based federated learning.
[i]CT: computed tomography.
[j]VFL: vertical federated learning.
[k]PoS: proof of stake.
[l]Not applicable/not available.
[m]IoHT: Internet of Healthcare Things.

Existing Surveys

To the best of our knowledge, comprehensive surveys or reviews focusing on the integration of blockchain and FL for health care use cases are scarce. While individual studies have explored the potential of each technology independently within health care settings, there is a notable lack of resources delving into the synergistic benefits and challenges of combining blockchain's immutable ledger capabilities with FL's decentralized model training approach.

Ngoupayou Limbepe et al [12] presented a taxonomy of FL-based privacy mechanisms categorized into privacy-enhancing technologies and hybrid techniques. This survey integrates privacy-enhancing technologies with blockchain in FL frameworks for smart health care systems. Noteworthy initial explorations, such as those conducted by Myrzashova et al [13] and Nguyen et al [464], offer valuable insights but often lack a broader perspective. Myrzashova et al [13], for instance, analyzed the advantages and disadvantages of BCFL integration in health care, but they overlooked the importance of a data type taxonomy. Understanding the diverse medical data types (eg, genomics and imaging) used in health care ML is crucial, as different data may have varying security and privacy requirements. Similarly, Nguyen et al [464] introduced a new conceptual architecture that integrates blockchain and AI for combating the COVID-19 pandemic. While offering valuable insights into addressing specific challenges posed by the pandemic, its scope is limited to COVID-19 and related data, lacking a broader analysis of the integration's potential for various health care use cases beyond this specific context. Hemdan et al [465] examined the convergence of digital twin technology, blockchain, and FL in

XSL•FO
RenderX

the medical field, with a specific focus on their technical architecture and real-world applications. Orabi et al [400] presented a literature review of BCFL applications in health care for protecting sensitive data. This paper investigated how integrating FL with blockchain can enhance its security, performance, and reliability.

In contrast to these existing reports, this tutorial offers a more comprehensive perspective. We present a taxonomy of medical data used for ML, providing a foundational understanding of the diverse data types relevant to this integration. Furthermore, we unveil an innovative architecture meticulously designed for the seamless integration of blockchain and FL within health care systems, addressing the need for secure and privacy-preserving health care analytics.

Future Research Directions

Overview

The integration of FL and blockchain in health care introduces a transformative paradigm for secure, privacy-preserving, and collaborative medical AI systems. Despite notable theoretical advances, several research frontiers remain underexplored. These can be broadly categorized into 4 interrelated areas: cryptographic foundations and quantum-resilient privacy, scalable and interoperable infrastructure, health care–specific consensus and incentivization, and full-stack integration with regulatory automation.

Cryptographic Foundations and Quantum-Resilient Privacy

A core requirement for the secure deployment of BCFL systems in health care is the development of advanced cryptographic mechanisms that ensure long-term privacy, data integrity, and resilience against quantum adversaries. Postquantum cryptographic primitives (such as lattice-based schemes like CRYSTALS-Kyber and Dilithium, and hash-based schemes like SPHINCS+) must be integrated into the BCFL stack to secure signatures and encryption layers [466]. In parallel, ZKPs and zero-knowledge virtual machines offer promising approaches to verifiably audit FL operations and regulatory compliance without exposing raw medical data. Furthermore, HE schemes, particularly multikey variants of the CKKS (Cheon-KimKim-Song) cryptosystem [467], enable secure aggregation of encrypted model parameters from multiple health care institutions without revealing individual contributions. The substantial computational overhead of such operations can be alleviated through specialized hardware accelerators, notably smart network interface cards, which offload cryptographic computations to dedicated processing units, thereby improving efficiency and scalability [468]. Complementarily, DP mechanisms require refinement to handle medical-specific feature types, with research needed into adaptive noise strategies and budget allocation tailored to data sensitivity. Together, these innovations will enable secure, auditable, and privacy-preserving collaborative learning in health care environments.

Scalable and Interoperable BCFL Infrastructure

As clinical systems generate increasingly voluminous and heterogeneous data, BCFL infrastructure must address throughput, latency, and interoperability challenges to support real-time learning and inference [401,439]. Health care–specific sharding architectures (such as parallel blockchains for EHRs, medical imaging, and genomics) can significantly improve throughput by distributing model updates across application domains. Complementary to this, layer-2 scaling solutions, including rollups [469] and state channels, offer reductions in transaction costs and ledger bloat, making frequent model updates feasible without congesting the blockchain [470]. Additionally, cross-chain interoperability mechanisms, such as those based on the Cosmos IBC protocol or Polkadot's relay chain model, are essential for enabling collaboration among health care institutions operating on heterogeneous blockchain platforms. Moreover, edge-centric FL will become increasingly important as IoMT and wearable devices proliferate. Furthermore, split learning and energy-efficient blockchain clients will be crucial to bring BCFL capabilities to constrained edge nodes while preserving security guarantees. These advancements will enable a scalable and interoperable ecosystem capable of supporting global health care collaborations.

Health Care–Specific Consensus and Incentive Mechanisms

Generic blockchain consensus protocols are ill-suited for the privacy, latency, and regulatory demands of health care. There is a growing need to design health care–oriented consensus protocols, such as medical and practical BFT [471], which can minimize communication overhead and incorporate privacy-aware authorization and medical rule validation into the consensus process. Similarly, reputation-based BFT [472] can ensure the credibility of participating institutions and practitioners, dynamically adjusting voting power based on behavior and contribution history. To sustain active participation in FL processes, especially among resource-constrained or data-rich stakeholders, incentive mechanisms must be tailored to health care environments. The proof-of-federated-work paradigm introduces an innovative strategy to quantify participant contributions using metrics like local accuracy, data diversity, and compliance adherence [473]. These are integrated into token-based or hybrid compensation schemes, further reinforced by game-theoretic models that optimize incentives based on Nash equilibrium to deter free-riding and encourage high-quality model updates. Such mechanisms are foundational to the long-term viability of decentralized health care learning networks.

System Integration, Compliance Automation, and Real-World Translation

For BCFL frameworks to be deployed at scale, they must integrate seamlessly with health care systems, satisfy regulatory requirements, and maintain usability in real-time clinical contexts. Smart contracts should evolve to support automated, upgradeable compliance verification that is capable of adapting to dynamic legal frameworks such as HIPAA, GDPR, and FDA AI guidelines [474]. By incorporating ZKP-based attestations, institutions can demonstrate compliance without revealing sensitive patient-level operations [440]. Simultaneously, embedding explainable AI techniques, such as SHAP values and attention heatmaps, into the FL pipeline is essential to ensure

XSL•FO
RenderX

clinician trust and support regulatory transparency [475]. Algorithmically, innovations in personalized FL, multitask learning (eg, MOCHA [476]), and community-based learning [477] can help address patient heterogeneity and improve predictive power in distributed environments. Real-time FL capabilities, triggered by event-driven smart contracts (eg, emergency alerts), are crucial for responsive applications like ICU triage and chronic condition monitoring [432]. To ensure long-term sustainability, quantum-resistant blockchain architectures should be explored using crypto-agile protocols and hybrid classical/postquantum stacks. These systems must also support interoperability, modular compliance, and secure cross-border data flows, forming the backbone of globally federated, privacy-aware health care AI infrastructure.

Conclusion

This tutorial introduced a comprehensive, clinically oriented, and compliance-aware framework integrating FL and blockchain for secure and privacy-preserving health care analytics. We demonstrated how FL enables decentralized model training across health care institutions while maintaining data locality and how blockchain enhances trust, integrity, and auditability through immutable ledgers and decentralized consensus mechanisms. Our key contributions include the following: (1) a systematic taxonomy of diverse medical data types and their FL requirements; (2) three novel integration architectures (fully coupled, semicoupled, and loosely coupled) with rigorous analysis of security, scalability, and regulatory compliance tradeoffs; (3) comprehensive security analysis identifying health care–specific vulnerabilities and mitigation strategies using advanced cryptographic techniques, including ZKPs, HE, and DP; and (4) a practical regulatory compliance framework addressing HIPAA, GDPR, and FDA guidelines for AI/ML-based medical devices. We validated BCFL effectiveness across critical health care applications, including disease prediction, medical imaging analysis, patient monitoring, and drug discovery. Looking ahead, crucial research frontiers involve quantum-resilient cryptography, scalable interoperable infrastructure, health care–specific consensus mechanisms, and automated compliance frameworks. This tutorial provides a foundational reference for developing trustworthy and patient-centric AI systems in health care. By integrating blockchain and FL, these systems can transform health care delivery while safeguarding privacy, ensuring regulatory compliance, and ultimately driving improved patient outcomes and accelerating medical discoveries in an increasingly connected health care ecosystem.

Acknowledgments

The authors would like to thank the anonymous reviewers and the academic editor for their constructive comments, which helped improve the quality and clarity of this tutorial. Artificial intelligence was used for refining, correcting, and editing the manuscript to improve language clarity.

Funding

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Data Availability

No new datasets were generated or analyzed during this study. All information discussed in this tutorial is derived from previously published literature and publicly available sources.

Authors' Contributions

Conceptualization: YS
Funding acquisition: AH, DM
Investigation: YS
Methodology: YS, YB
Supervision: YS, AH
Writing – original draft: YS, YB, OAD
Writing – review & editing: YS, YB

Conflicts of Interest

None declared.

References

1. Ometov A, Shubina V, Klus L, et al. A survey on wearable technology: history, state-of-the-art and current challenges. Computer Networks 2021 Jul;193:108074. [doi: 10.1016/j.comnet.2021.108074]
2. Kang HS, Exworthy M. Wearing the future-wearables to empower users to take greater responsibility for their health and care: scoping review. JMIR Mhealth Uhealth 2022 Jul 13;10(7):e35684. [doi: 10.2196/35684] [Medline: 35830222]
3. Wu M. Wearable technology applications in healthcare: a literature review. Online J Nurs Inform 2019;23(3) [FREE Full text]

XSL•FO
RenderX

4. Zhang C, Xie Y, Bai H, Yu B, Li W, Gao Y. A survey on federated learning. Knowl Based Syst 2021 Mar;216:106775. [doi: 10.1016/j.knosys.2021.106775]
5. Babar M, Qureshi B, Koubaa A. Review on federated learning for digital transformation in healthcare through big data analytics. Future Generation Computer Systems 2024 Nov;160:14-28. [doi: 10.1016/j.future.2024.05.046]
6. Fang M, Cao X, Jia J, Gong N. Local model poisoning attacks to byzantine-robust federated learning. Presented at: 29th USENIX Security Symposium; Aug 12-14, 2020 URL: https://www.usenix.org/system/files/sec20-fang.pdf [accessed 2026-05-07]
7. Kumar KN, Mohan CK, Cenkeramaddi LR, Awasthi N. Minimal data poisoning attack in federated learning for medical image classification: an attacker perspective. Artif Intell Med 2025 Jan;159:103024. [doi: 10.1016/j.artmed.2024.103024] [Medline: 39591879]
8. Bouacida N, Mohapatra P. Vulnerabilities in federated learning. IEEE Access 2021;9:63229-63249. [doi: 10.1109/ACCESS.2021.3075203]
9. Cai Z, Chen J, Fan Y, Zheng Z, Li K. Blockchain-empowered federated learning: benefits, challenges, and solutions. IEEE Trans Big Data 2025 Feb 13;11(5):2244-2263. [doi: 10.1109/TBDATA.2025.3541560]
10. Abbas SR, Abbas Z, Zahir A, Lee SW. Federated learning in smart healthcare: a comprehensive review on privacy, security, and predictive analytics with IoT integration. Healthcare (Basel) 2024 Dec 22;12(24):2587. [doi: 10.3390/healthcare12242587] [Medline: 39766014]
11. Cheng H, Qu Y, Liu W, Gao L, Zhu T. Decentralized federated learning for private smart healthcare: a survey. Mathematics 2025;13(8):1296. [doi: 10.3390/math13081296]
12. Ngoupayou Limbepe Z, Gai K, Yu J. Blockchain-based privacy-enhancing federated learning in smart healthcare: a survey. Blockchains 2025;3(1):1. [doi: 10.3390/blockchains3010001]
13. Myrzashova R, Alsamhi SH, Shvetsov AV, Hawbani A, Wei X. Blockchain meets federated learning in healthcare: a systematic review with challenges and opportunities. IEEE Internet Things J 2023;10(16):14418-14437. [doi: 10.1109/JIOT.2023.3263598]
14. Datta P, Namin AS, Chatterjee M. A survey of privacy concerns in wearable devices. Presented at: 2018 IEEE International Conference on Big Data (Big Data); Dec 10-13, 2018. [doi: 10.1109/BigData.2018.8622110]
15. Qayyum A, Qadir J, Bilal M, Al-Fuqaha A. Secure and robust machine learning for healthcare: a survey. IEEE Rev Biomed Eng 2021;14:156-180. [doi: 10.1109/RBME.2020.3013489] [Medline: 32746371]
16. Cilliers L. Wearable devices in healthcare: privacy and information security issues. Health Inf Manag 2020;49(2-3):150-156. [doi: 10.1177/1833358319851684] [Medline: 31146589]
17. Lyu L, Yu H, Ma X, et al. Privacy and robustness in federated learning: attacks and defenses. IEEE Trans Neural Netw Learning Syst 2024;35(7):8726-8746. [doi: 10.1109/TNNLS.2022.3216981]
18. Ye H, Liang L, Li GY. Decentralized federated learning with unreliable communications. IEEE J Sel Top Signal Process 2022;16(3):487-500. [doi: 10.1109/JSTSP.2022.3152445]
19. Chen L, Zhao D, Tao L, et al. A credible and fair federated learning framework based on blockchain. IEEE Trans Artif Intell 2024;6(2):301-316. [doi: 10.1109/TAI.2024.3355362]
20. Gupta M, Kumar M, Dhir R. Unleashing the prospective of blockchain-federated learning fusion for IoT security: a comprehensive review. Computer Science Review 2024 Nov;54:100685. [doi: 10.1016/j.cosrev.2024.100685]
21. Qu Y, Uddin MP, Gan C, Xiang Y, Gao L, Yearwood J. Blockchain-enabled federated learning: a survey. ACM Comput Surv 2023 Apr 30;55(4):1-35. [doi: 10.1145/3524104]
22. Zhu J, Cao J, Saxena D, Jiang S, Ferradi H. Blockchain-empowered federated learning: challenges, solutions, and future directions. ACM Comput Surv 2023 Nov 30;55(11):1-31. [doi: 10.1145/3570953]
23. Qammar A, Karim A, Ning H, Ding J. Securing federated learning with blockchain: a systematic literature review. Artif Intell Rev 2023;56(5):3951-3985. [doi: 10.1007/s10462-022-10271-9] [Medline: 36160367]
24. Ramakrishnaiah Y, Macesic N, Webb GI, Peleg AY, Tyagi S. EHR-ML: a data-driven framework for designing machine learning applications with electronic health records. Int J Med Inform 2025 Apr;196:105816. [doi: 10.1016/j.ijmedinf.2025.105816] [Medline: 39891983]
25. Shen Y, Yu J, Zhou J, Hu G. Twenty-five years of evolution and hurdles in electronic health records and interoperability in medical research: comprehensive review. J Med Internet Res 2025 Jan 9;27:e59024. [doi: 10.2196/59024] [Medline: 39787599]
26. Nowrozy R, Ahmed K, Kayes ASM, Wang H, McIntosh TR. Privacy preservation of electronic health records in the modern era: a systematic survey. ACM Comput Surv 2024 Aug 31;56(8):1-37. [doi: 10.1145/3653297]
27. Arbet J, Brokamp C, Meinzen-Derr J, Trinkley KE, Spratt HM. Lessons and tips for designing a machine learning study using EHR data. J Clin Trans Sci 2021;5(1):e21. [doi: 10.1017/cts.2020.513]
28. Wu H, Yamal JM, Yaseen A, Maroufy V. Statistics and Machine Learning Methods for EHR Data: From Data Extraction to Data Analytics: CRC Press; 2020. [doi: 10.1201/9781003030003]
29. Johnston SS, Morton JM, Kalsekar I, Ammann EM, Hsiao CW, Reps J. Using machine learning applied to real-world healthcare data for predictive analytics: an applied example in bariatric surgery. Value Health 2019 May;22(5):580-586. [doi: 10.1016/j.jval.2019.01.011] [Medline: 31104738]

XSL•FO
RenderX

30. Muniasamy A, Tabassam S, Hussain MA, Sultana H, Muniasamy V, Bhatnagar R. Deep learning for predictive analytics in healthcare. In: Hassanien A, Azar A, Gaber T, Bhatnagar R, Tolba M, editors. The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019) AMLTA 2019 Advances in Intelligent Systems and Computing: Springer; 2020:32-42. [doi: 10.1007/978-3-030-14118-9_4]
31. Singh P, Singh N, Singh KK, Singh A. Diagnosing of disease using machine learning. In: Singh KK, Elhoseny M, Singh A, Elngar AA, editors. Machine Learning and the Internet of Medical Things in Healthcare: Academic Press; 2021:89-111. [doi: 10.1016/B978-0-12-821229-5.00003-3]
32. Ahsan MM, Luna SA, Siddique Z. Machine-learning-based disease diagnosis: a comprehensive review. Healthcare (Basel) 2022 Mar 15;10(3):541. [doi: 10.3390/healthcare10030541] [Medline: 35327018]
33. Komal Kumar N, Vigneswari D. A drug recommendation system for multi-disease in health care using machine learning. In: Hura GS, Singh AK, Siong Hoe L, editors. Advances in Communication and Computational Technology ICACCT 2019 Lecture Notes in Electrical Engineering: Springer; 2021:1-12. [doi: 10.1007/978-981-15-5341-7_1]
34. Abhisheka B, Biswas SK, Purkayastha B, Das D, Escargueil A. Recent trend in medical imaging modalities and their applications in disease diagnosis: a review. Multimed Tools Appl 2023;83(14):43035-43070. [doi: 10.1007/s11042-023-17326-1]
35. Lalitha S, Sanjana T, Bhavana H, Bhan I, Harshith G. Medical imaging modalities and different image processing techniques: state of the art review. In: Shinde SV, Mahalle PN, Bendre V, Castillo O, editors. Disruptive Developments in Biomedical Applications: CRC Press; 2022:17-36. [doi: 10.1201/9781003272694-3]
36. Santhi K. A survey on medical imaging techniques and applications. J Innov Image Process 2022;4(3):173-182. [doi: 10.36548/jiip.2022.3.005]
37. Mustapha MT, Uzun B, Ozsahin DU, Ozsahin I. A comparative study of x-ray based medical imaging devices. In: Ozsahin I, Ozsahin DU, Uzun B, editors. Applications of Multi-Criteria Decision-Making Theories in Healthcare and Biomedical Engineering: Academic Press; 2021:163-180. [doi: 10.1016/B978-0-12-824086-1.00011-6]
38. Withers PJ, Bouman C, Carmignato S, et al. X-ray computed tomography. Nat Rev Methods Primers 2021;1(1):18. [doi: 10.1038/s43586-021-00015-4]
39. De Pietro S, Di Martino G, Caroprese M, et al. The role of MRI in radiotherapy planning: a narrative review “from head to toe”. Insights Imaging 2024 Oct 23;15(1):255. [doi: 10.1186/s13244-024-01799-1] [Medline: 39441404]
40. Avola D, Cinque L, Fagioli A, Foresti G, Mecca A. Ultrasound medical imaging techniques. ACM Comput Surv 2022 Apr 30;54(3):1-38. [doi: 10.1145/3447243]
41. Cullen A, O’Connell M. Tutorial 14: introduction to nuclear medicine. In: Redmond C, Lee M, editors. Tutorials in Diagnostic Radiology for Medical Students: Springer; 2020:225-233. [doi: 10.1007/978-3-030-31893-2_14]
42. Wernick MN, Aarsvold JN. Emission Tomography: The Fundamentals of PET and SPECT: Academic Press; 2004.
43. Jones AK, Balter S, Rauch P, Wagner LK. Medical imaging using ionizing radiation: optimization of dose and image quality in fluoroscopy. Med Phys 2014 Jan;41(1):014301. [doi: 10.1118/1.4835495] [Medline: 24387534]
44. Ulrich H, Kock-Schoppenhauer AK, Deppenwiese N, et al. Understanding the nature of metadata: systematic review. J Med Internet Res 2022 Jan 11;24(1):e25440. [doi: 10.2196/25440] [Medline: 35014967]
45. Bhuiyan MN, Rahman MM, Billah MM, Saha D. Internet of things (IoT): a review of its enabling technologies in healthcare applications, standards protocols, security, and market opportunities. IEEE Internet Things J 2021;8(13):10474-10498. [doi: 10.1109/JIOT.2021.3062630]
46. Larobina M. Thirty years of the DICOM standard. Tomography 2023 Oct 6;9(5):1829-1838. [doi: 10.3390/tomography9050145] [Medline: 37888737]
47. Wang J, Wang S, Zhang Y. Deep learning on medical image analysis. CAAI Trans on Intel Tech 2025 Feb;10(1):1-35. [doi: 10.1049/cit2.12356]
48. Rana M, Bhushan M. Machine learning and deep learning approach for medical image analysis: diagnosis to detection. Multimed Tools Appl 2023 Jul;82(17):26731-26769. [doi: 10.1007/s11042-022-14305-w]
49. Jeyaraj PR, Samuel Nadar ER. Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm. J Cancer Res Clin Oncol 2019 Apr;145(4):829-837. [doi: 10.1007/s00432-018-02834-7] [Medline: 30603908]
50. Echle A, Rindtorff NT, Brinker TJ, Luedde T, Pearson AT, Kather JN. Deep learning in cancer pathology: a new generation of clinical biomarkers. Br J Cancer 2021 Feb;124(4):686-696. [doi: 10.1038/s41416-020-01122-x] [Medline: 33204028]
51. Savadjiev P, Chong J, Dohan A, et al. Image-based biomarkers for solid tumor quantification. Eur Radiol 2019 Oct;29(10):5431-5440. [doi: 10.1007/s00330-019-06169-w] [Medline: 30963275]
52. Xu Y, Hosny A, Zeleznik R, et al. Deep learning predicts lung cancer treatment response from serial medical imaging. Clin Cancer Res 2019 Jun 1;25(11):3266-3275. [doi: 10.1158/1078-0432.CCR-18-2495] [Medline: 31010833]
53. Ahishakiye E, Van Gijzen MB, Tumwiine J, Wario R, Obungoloch J. A survey on deep learning in medical image reconstruction. Intelligent Medicine 2021 Sep;1(3):118-127. [doi: 10.1016/j.imed.2021.03.003]
54. Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov 2019 Jun;18(6):463-477. [doi: 10.1038/s41573-019-0024-5] [Medline: 30976107]

XSL•FO
RenderX

55. Ginsburg GS, Willard HF. Genomic and personalized medicine: foundations and applications. Transl Res 2009 Dec;154(6):277-287. [doi: 10.1016/j.trsl.2009.09.005] [Medline: 19931193]
56. Chen YM, Hsiao TH, Lin CH, Fann YC. Unlocking precision medicine: clinical applications of integrating health records, genetics, and immunology through artificial intelligence. J Biomed Sci 2025 Feb 7;32(1):16. [doi: 10.1186/s12929-024-01110-w] [Medline: 39915780]
57. Gupta S, Janu N, Nawal M, Goswami A. Genomics and machine learning: ML approaches, future directions and challenges in genomics. In: Choudhary S, Kumar S, Gowroju S, Gulhane M, Sri Lakshmi R, editors. Genomics at the Nexus of AI, Computer Vision, and Machine Learning 2025:437-457. [doi: 10.1002/9781394268832.ch20]
58. Suissa JS, De La Cerda GY, Graber LC, et al. Data-driven guidelines for phylogenomic analyses using SNP data. Appl Plant Sci 2024;12(6):e11611. [doi: 10.1002/aps3.11611] [Medline: 39628540]
59. Hollox EJ, Zuccherato LW, Tucci S. Genome structural variation in human evolution. Trends Genet 2022 Jan;38(1):45-58. [doi: 10.1016/j.tig.2021.06.015] [Medline: 34284881]
60. Audic S, Claverie JM. The significance of digital gene expression profiles. Genome Res 1997 Oct;7(10):986-995. [doi: 10.1101/gr.7.10.986] [Medline: 9331369]
61. Dai W, Qiao X, Fang Y, et al. Epigenetics-targeted drugs: current paradigms and future challenges. Sig Transduct Target Ther 2024;9(1):332. [doi: 10.1038/s41392-024-02039-0]
62. Schmitz MJ, Bashar A, Soman V, et al. Leveraging diverse genomic data to guide equitable carrier screening: Insights from gnomAD v.4.1.0. Am J Hum Genet 2025 Jan 2;112(1):181-195. [doi: 10.1016/j.ajhg.2024.11.004] [Medline: 39615480]
63. Gulamali FF, Sawant AS, Nadkarni GN. Machine learning for risk stratification in kidney disease. Curr Opin Nephrol Hypertens 2022 Nov 1;31(6):548-552. [doi: 10.1097/MNH.0000000000000832] [Medline: 36004937]
64. Jamalinia M, Weiskirchen R. Advances in personalized medicine: translating genomic insights into targeted therapies for cancer treatment. Ann Transl Med 2025 Apr 30;13(2):18. [doi: 10.21037/atm-25-34] [Medline: 40438512]
65. Mani S, Lalani SR, Pammi M. Genomics and multiomics in the age of precision medicine. Pediatr Res 2025 Mar;97(4):1399-1410. [doi: 10.1038/s41390-025-04021-0] [Medline: 40185865]
66. Vadapalli S, Abdelhalim H, Zeeshan S, Ahmed Z. Artificial intelligence and machine learning approaches using gene expression and variant data for personalized medicine. Brief Bioinform 2022 Sep 20;23(5):bbac191. [doi: 10.1093/bib/bbac191] [Medline: 35595537]
67. Quazi S. Artificial intelligence and machine learning in precision and genomic medicine. Med Oncol 2022 Jun 15;39(8):120. [doi: 10.1007/s12032-022-01711-1] [Medline: 35704152]
68. Khan M. Bioinformatics and machine learning: analyzing genomic data for personalized medicine. Open Science Framework. Preprint posted online on Aug 8, 2023. [doi: 10.31219/osf.io/93584]
69. Dara S, Dhamercherla S, Jadav SS, Babu CM, Ahsan MJ. Machine learning in drug discovery: a review. Artif Intell Rev 2022;55(3):1947-1999. [doi: 10.1007/s10462-021-10058-4] [Medline: 34393317]
70. Kearney E, Wojcik A, Babu D. Artificial intelligence in genetic services delivery: utopia or apocalypse? J Genet Couns 2020 Feb;29(1):8-17. [doi: 10.1002/jgc4.1192] [Medline: 31749317]
71. Benning L, Peintner A, Peintner L. Advances in and the applicability of machine learning-based screening and early detection approaches for cancer: a primer. Cancers (Basel) 2022 Jan 26;14(3):623. [doi: 10.3390/cancers14030623] [Medline: 35158890]
72. Tao K, Bian Z, Zhang Q, et al. Machine learning-based genome-wide interrogation of somatic copy number aberrations in circulating tumor DNA for early detection of hepatocellular carcinoma. EBioMedicine 2020 Jun;56:102811. [doi: 10.1016/j.ebiom.2020.102811] [Medline: 32512514]
73. Mullen M, Zhang A, Lui GK, Romfh AW, Rhee JW, Wu JC. Race and genetics in congenital heart disease: application of iPSCs, omics, and machine learning technologies. Front Cardiovasc Med 2021;8:635280. [doi: 10.3389/fcvm.2021.635280] [Medline: 33681306]
74. Cortés-Ciriano I, Gulhan DC, Lee JJK, Melloni GEM, Park PJ. Computational analysis of cancer genome sequencing data. Nat Rev Genet 2022 May;23(5):298-314. [doi: 10.1038/s41576-021-00431-y] [Medline: 34880424]
75. Dargan S, Kumar M. A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities. Expert Syst Appl 2020 Apr;143:113114. [doi: 10.1016/j.eswa.2019.113114]
76. Uwaechia AN, Ramli DA. A comprehensive survey on ECG signals as new biometric modality for human authentication: recent advances and future challenges. IEEE Access 2021;9:97760-97802. [doi: 10.1109/ACCESS.2021.3095248]
77. Paranjape RB, Mahovsky J, Benedicenti L, Koles' Z. The electroencephalogram as a biometric. Presented at: Canadian Conference on Electrical and Computer Engineering 2001; May 13-16, 2001. [doi: 10.1109/CCECE.2001.933649]
78. Franceschetti L, Lodetti G, Blandino A, Amadasi A, Bugelli V. Exploring the role of the human microbiome in forensic identification: opportunities and challenges. Int J Legal Med 2024 Sep;138(5):1891-1905. [doi: 10.1007/s00414-024-03217-z] [Medline: 38594499]
79. Sharma A, Kumar R, Varadwaj P. Smelling the disease: diagnostic potential of breath analysis. Mol Diagn Ther 2023 May;27(3):321-347. [doi: 10.1007/s40291-023-00640-7] [Medline: 36729362]
80. Abdulrahman SA, Alhayani B. A comprehensive survey on the biometric systems based on physiological and behavioural characteristics. Mater Today 2023;80:2642-2646. [doi: 10.1016/j.matpr.2021.07.005]

XSL•FO
RenderX

81. Kisku DR, Gupta P, Sing JK. Design and Implementation of Healthcare Biometric Systems: IGI Global; 2019. [doi: 10.4018/978-1-5225-7525-2]
82. Mason J, Dave R, Chatterjee P, Graham-Allen I, Esterline A, Roy K. An investigation of biometric authentication in the healthcare environment. Array 2020 Dec;8:100042. [doi: 10.1016/j.array.2020.100042]
83. Kaul SD, Murty VK, Hatzinakos D. Secure and privacy preserving biometric based user authentication with data access control system in the healthcare environment. Presented at: 2020 International Conference on Cyberworlds (CW); Sep 29 to Oct 1, 2020. [doi: 10.1109/CW49994.2020.00047]
84. Barka E, Al Baqari M, Kerrache CA, Herrera-Tapia J. Implementation of a biometric-based blockchain system for preserving privacy, security, and access control in healthcare records. J Sens Actuator Netw 2022;11(4):85. [doi: 10.3390/jsan11040085]
85. Choi JH, Khamraev K, Cheriyan D. Hybrid health risk assessment model using real-time particulate matter, biometrics, and benchmark device. J Clean Prod 2022 May;350:131443. [doi: 10.1016/j.jclepro.2022.131443]
86. Baseri Y, Hafid A, Firoozjaei MD, Cherkaoui S, Ray I. Statistical privacy protection for secure data access control in cloud. J Inf Secur Appl 2024 Aug;84:103823. [doi: 10.1016/j.jisa.2024.103823]
87. Riplinger L, Piera-Jiménez J, Dooling JP. Patient identification techniques - approaches, implications, and findings. Yearb Med Inform 2020 Aug;29(1):81-86. [doi: 10.1055/s-0040-1701984] [Medline: 32823300]
88. Sohn JW, Kim H, Park SB, et al. Clinical study of using biometrics to identify patient and procedure. Front Oncol 2020;10:586232. [doi: 10.3389/fonc.2020.586232] [Medline: 33335855]
89. Fatimah B, Singh P, Singhal A, Pachori RB. Biometric identification from ECG signals using Fourier decomposition and machine learning. IEEE Trans Instrum Meas 2022;71:1-9. [doi: 10.1109/TIM.2022.3199260]
90. Prakash AJ, Patro KK, Samantray S, Pławiak P, Hammad M. A deep learning technique for biometric authentication using ECG beat template matching. Information 2023;14(2):65. [doi: 10.3390/info14020065]
91. Mohsin AH, Zaidan AA, Zaidan BB, et al. Real-time remote health monitoring systems using body sensor information and finger vein biometric verification: a multi-layer systematic review. J Med Syst 2018 Oct 16;42(12):1-36. [doi: 10.1007/s10916-018-1104-5] [Medline: 30327939]
92. Ismail SNA, Nayan NA, Jaafar R, May Z. Recent advances in non-invasive blood pressure monitoring and prediction using a machine learning approach. Sensors (Basel) 2022 Aug 18;22(16):6195. [doi: 10.3390/s22166195] [Medline: 36015956]
93. Fei C, Liu R, Li Z, Wang T, Baig FN. Machine and deep learning algorithms for wearable health monitoring. In: Manocha AK, Jain S, Singh M, Paul S, editors. Computational Intelligence in Healthcare Health Information Science: Springer; 2021:105-160. [doi: 10.1007/978-3-030-68723-6_6]
94. Gomes N, Pato M, Lourenço AR, Datia N. A survey on wearable sensors for mental health monitoring. Sensors (Basel) 2023 Jan 25;23(3):1330. [doi: 10.3390/s23031330] [Medline: 36772370]
95. Killoran J, Cui Y(, Park A, van Esch P, Kietzmann J. Can behavioral biometrics make everyone happy? Bus Horiz 2023 Sep;66(5):585-591. [doi: 10.1016/j.bushor.2023.02.001]
96. Garcia-Ceja E, Riegler M, Nordgreen T, Jakobsen P, Oedegaard KJ, Tørresen J. Mental health monitoring with multimodal sensing and machine learning: a survey. Pervasive Mob Comput 2018 Dec;51:1-26. [doi: 10.1016/j.pmcj.2018.09.003]
97. Harrer S, Shah P, Antony B, Hu J. Artificial intelligence for clinical trial design. Trends Pharmacol Sci 2019 Aug;40(8):577-591. [doi: 10.1016/j.tips.2019.05.005] [Medline: 31326235]
98. Weissler EH, Naumann T, Andersson T, et al. The role of machine learning in clinical research: transforming the future of evidence generation. Trials 2021 Aug 16;22(1):537. [doi: 10.1186/s13063-021-05489-x] [Medline: 34399832]
99. Moro-Velazquez L, Gomez-Garcia JA, Arias-Londoño JD, Dehak N, Godino-Llorente JI. Advances in Parkinson's disease detection and assessment using voice and speech: a review of the articulatory and phonatory aspects. Biomed Signal Process Control 2021 Apr;66:102418. [doi: 10.1016/j.bspc.2021.102418]
100. Ngo QC, Motin MA, Pah ND, Drotár P, Kempster P, Kumar D. Computerized analysis of speech and voice for Parkinson's disease: a systematic review. Comput Methods Programs Biomed 2022 Nov;226:107133. [doi: 10.1016/j.cmpb.2022.107133] [Medline: 36183641]
101. Lella KK, Pja A. Automatic COVID-19 disease diagnosis using 1D convolutional neural network and augmentation with human respiratory sound based on parameters: cough, breath, and voice. AIMS Public Health 2021;8(2):240-264. [doi: 10.3934/publichealth.2021019] [Medline: 34017889]
102. Idrisoglu A, Dallora AL, Anderberg P, Berglund JS. Applied machine learning techniques to diagnose voice-affecting conditions and disorders: systematic literature review. J Med Internet Res 2023 Jul 19;25:e46105. [doi: 10.2196/46105] [Medline: 37467031]
103. Lyakso E, Frolova O, Nikolaev A. Voice and speech features as a diagnostic symptom. In: Psychological Applications and Trends 2021:359-363 URL: https://inpact-psychologyconference.org/wp-content/uploads/2021/05/2021inpact074.pdf [accessed 2026-05-07] [doi: 10.36315/2021inpact074]
104. Onyema EM, Shukla PK, Dalal S, Mathur MN, Zakariah M, Tiwari B. Enhancement of patient facial recognition through deep learning algorithm: ConvNet. J Healthc Eng 2021;2021:5196000. [doi: 10.1155/2021/5196000] [Medline: 34912534]
105. Jeon B, Jeong B, Jee S, et al. A facial recognition mobile app for patient safety and biometric identification: design, development, and validation. JMIR Mhealth Uhealth 2019 Apr 8;7(4):e11472. [doi: 10.2196/11472] [Medline: 30958275]

XSL•FO
RenderX

106. Ghazal TM, Hasan MK, Alshurideh MT, et al. IoT for smart cities: machine learning approaches in smart healthcare—a review. Future Internet 2021;13(8):218. [doi: 10.3390/fi13080218]
107. Balakrishna S, Thirumaran M, Solanki VK. Iot sensor data integration in healthcare using semantics and machine learning approaches. In: Balas V, Solanki V, Kumar R, Ahad M, editors. A Handbook of Internet of Things in Biomedical and Cyber Physical System Intelligent Systems Reference Library: Springer; 2020:275-300. [doi: 10.1007/978-3-030-23983-1_11]
108. Lv Z, Li Y. Wearable sensors for vital signs measurement: a survey. J Sens Actuator Netw 2022;11(1):19. [doi: 10.3390/jsan11010019]
109. Junaid SB, Imam AA, Shuaibu AN, et al. Artificial intelligence, sensors and vital health signs: a review. Appl Sci (Basel) 2022;12(22):11475. [doi: 10.3390/app122211475]
110. Gupta N, Gupta SK, Pathak RK, Jain V, Rashidi P, Suri JS. Human activity recognition in artificial intelligence framework: a narrative review. Artif Intell Rev 2022;55(6):4755-4808. [doi: 10.1007/s10462-021-10116-x] [Medline: 35068651]
111. Zhu T, Uduku C, Li K, Herrero P, Oliver N, Georgiou P. Enhancing self-management in type 1 diabetes with wearables and deep learning. NPJ Digit Med 2022 Jun 27;5(1):78. [doi: 10.1038/s41746-022-00626-5] [Medline: 35760819]
112. Liyakat KKS. Heart health monitoring using IoT and machine learning methods. In: Shaik A, editor. AI-Powered Advances in Pharmacology: IGI Global; 2025:257-282. [doi: 10.4018/979-8-3693-3212-2.ch010]
113. Kwon SH, Dong L. Flexible sensors and machine learning for heart monitoring. Nano Energy 2022 Nov;102:107632. [doi: 10.1016/j.nanoen.2022.107632]
114. Lin SY, Tsai CY, Majumdar A, et al. Combining a wireless radar sleep monitoring device with deep machine learning techniques to assess obstructive sleep apnea severity. J Clin Sleep Med 2024 Aug 1;20(8):1267-1277. [doi: 10.5664/jcsm.11136] [Medline: 38546033]
115. Arora A, Chakraborty P, Bhatia MPS. Analysis of data from wearable sensors for sleep quality estimation and prediction using deep learning. Arab J Sci Eng 2020 Dec;45(12):10793-10812. [doi: 10.1007/s13369-020-04877-w]
116. Zhang W, Ram S. A comprehensive analysis of triggers and risk factors for asthma based on machine learning and large heterogeneous data sources. MIS Q 2020 Mar 1;44(1):305-350. [doi: 10.25300/MISQ/2020/15106]
117. Bohlmann A, Mostafa J, Kumar M. Machine learning and medication adherence: scoping review. JMIRx Med 2021 Nov 24;2(4):e26993. [doi: 10.2196/26993] [Medline: 37725549]
118. Roh H, Shin S, Han J, Lim S. A deep learning-based medication behavior monitoring system. Math Biosci Eng 2021 Jan 28;18(2):1513-1528. [doi: 10.3934/mbe.2021078] [Medline: 33757196]
119. Meyer BM, Tulipani LJ, Gurchiek RD, et al. Wearables and deep learning classify fall risk from gait in multiple sclerosis. IEEE J Biomed Health Inform 2020;25(5):1824-1831. [doi: 10.1109/JBHI.2020.3025049]
120. Kyamakya K, Al-Machot F, Haj Mosa A, Bouchachia H, Chedjou JC, Bagula A. Emotion and stress recognition related sensors and machine learning technologies. Sensors (Basel) 2021 Mar 24;21(7):2273. [doi: 10.3390/s21072273] [Medline: 33804987]
121. Gedam S, Paul S. A review on mental stress detection using wearable sensors and machine learning techniques. IEEE Access 2021;9:84045-84066. [doi: 10.1109/ACCESS.2021.3085502]
122. Fountzilas E, Pearce T, Baysal MA, Chakraborty A, Tsimberidou AM. Convergence of evolving artificial intelligence and machine learning techniques in precision oncology. NPJ Digit Med 2025 Jan 31;8(1):75. [doi: 10.1038/s41746-025-01471-y] [Medline: 39890986]
123. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med 2022 Sep;28(9):1773-1784. [doi: 10.1038/s41591-022-01981-2] [Medline: 36109635]
124. Hsueh PYS, Dey S, Das S, Wetter T. Making sense of patient-generated health data for interpretable patientcentered care: the transition from “more” to “better. In: Gundlapalli AV, Jaulent MC, Zhao D, editors. MEDINFO 2017: Precision Healthcare through Informatics: IOS Press; 2017:113-117. [doi: 10.3233/978-1-61499-830-3-113]
125. Mendo IR, Marques G, de la Torre Díez I, López-Coronado M, Martín-Rodríguez F. Machine learning in medical emergencies: a systematic review and analysis. J Med Syst 2021 Aug 18;45(10):88. [doi: 10.1007/s10916-021-01762-3] [Medline: 34410512]
126. van der Boon RMA, Camm AJ, Aguiar C, et al. Risks and benefits of sharing patient information on social media: a digital dilemma. Eur Heart J Digit Health 2024 May;5(3):199-207. [doi: 10.1093/ehjdh/ztae009] [Medline: 38774369]
127. Gupta A, Katarya R. Social media based surveillance systems for healthcare using machine learning: a systematic review. J Biomed Inform 2020 Aug;108:103500. [doi: 10.1016/j.jbi.2020.103500] [Medline: 32622833]
128. Hasib KM, Islam MR, Sakib S, Akbar MA, Razzak I, Alam MS. Depression detection from social networks data based on machine learning and deep learning techniques: an interrogative survey. IEEE Trans Comput Soc Syst 2023;10(4):1568-1586. [doi: 10.1109/TCSS.2023.3263128]
129. Johnson KB, Wei WQ, Weeraratne D, et al. Precision medicine, AI, and the future of personalized health care. Clin Transl Sci 2021 Jan;14(1):86-93. [doi: 10.1111/cts.12884] [Medline: 32961010]
130. Verma S, Malviya R, Alam MA, Tripathi BD. Tele-health monitoring using artificial intelligence deep learning framework. In: Malviya R, Ghinea G, Dhanaraj RK, Balusamy B, Sundram S, editors. Deep Learning for Targeted Treatments: Transformation in Healthcare: Scrivener Publishing; 2022:199-228. [doi: 10.1002/9781119857983]

XSL•FO
RenderX

131. Moorthy V, Abubakar I, Qadri F, et al. The future of the global clinical trial ecosystem: a vision from the first WHO Global Clinical Trials Forum. The Lancet 2024 Jan;403(10422):124-126. [doi: 10.1016/S0140-6736(23)02798-8]
132. Pettit RW, Fullem R, Cheng C, Amos CI. Artificial intelligence, machine learning, and deep learning for clinical outcome prediction. Emerg Top Life Sci 2021 Dec 20;5(6):729-745. [doi: 10.1042/ETLS20210246] [Medline: 34927670]
133. Cohen IG. Informed consent and medical artificial intelligence: what to tell the patient? SSRN Journal 2019;108. [doi: 10.2139/ssrn.3529576]
134. McKeown A, Mourby M, Harrison P, Walker S, Sheehan M, Singh I. Ethical issues in consent for the reuse of data in health data platforms. Sci Eng Ethics 2021 Feb 4;27(1):9. [doi: 10.1007/s11948-021-00282-0] [Medline: 33538942]
135. Chien I, Enrique A, Palacios J, et al. A machine learning approach to understanding patterns of engagement with internet-delivered mental health interventions. JAMA Netw Open 2020 Jul 1;3(7):e2010791. [doi: 10.1001/jamanetworkopen.2020.10791] [Medline: 32678450]
136. Benke K, Benke G. Artificial intelligence and big data in public health. Int J Environ Res Public Health 2018 Dec 10;15(12):2796. [doi: 10.3390/ijerph15122796] [Medline: 30544648]
137. Song L, Li Y, Nie S, et al. Using machine learning to predict adverse events in acute coronary syndrome: a retrospective study. Clin Cardiol 2023 Dec;46(12):1594-1602. [doi: 10.1002/clc.24127] [Medline: 37654030]
138. Yang J, Wan J, Feng L, et al. Machine learning algorithms for the prediction of adverse prognosis in patients undergoing peritoneal dialysis. BMC Med Inform Decis Mak 2024 Jan 2;24(1):8. [doi: 10.1186/s12911-023-02412-z] [Medline: 38166909]
139. Badwan BA, Liaropoulos G, Kyrodimos E, Skaltsas D, Tsirigos A, Gorgoulis VG. Machine learning approaches to predict drug efficacy and toxicity in oncology. Cell Rep Methods 2023 Feb 27;3(2):100413. [doi: 10.1016/j.crmeth.2023.100413] [Medline: 36936080]
140. Feijoo F, Palopoli M, Bernstein J, Siddiqui S, Albright TE. Key indicators of phase transition for clinical trials through machine learning. Drug Discov Today 2020 Feb;25(2):414-421. [doi: 10.1016/j.drudis.2019.12.014] [Medline: 31926317]
141. MacEachern SJ, Forkert ND. Machine learning for precision medicine. Genome 2021 Apr;64(4):416-425. [doi: 10.1139/gen-2020-0131] [Medline: 33091314]
142. Chalasani SH, Syed J, Ramesh M, Patil V, Pramod Kumar TM. Artificial intelligence in the field of pharmacy practice: a literature review. Explor Res Clin Soc Pharm 2023 Dec;12:100346. [doi: 10.1016/j.rcsop.2023.100346] [Medline: 37885437]
143. Askr H, Elgeldawi E, Aboul Ella H, Elshaier Y, Gomaa MM, Hassanien AE. Deep learning in drug discovery: an integrative review and future challenges. Artif Intell Rev 2023;56(7):5975-6037. [doi: 10.1007/s10462-022-10306-1] [Medline: 36415536]
144. Li Q, Tang B, Wu Y, et al. Machine learning: a new approach for dose individualization. Clin Pharma and Therapeutics 2024 Apr;115(4):727-744. [doi: 10.1002/cpt.3049]
145. Del Fabro L, Bondi E, Serio F, Maggioni E, D'Agostino A, Brambilla P. Machine learning methods to predict outcomes of pharmacological treatment in psychosis. Transl Psychiatry 2023 Mar 2;13(1):75. [doi: 10.1038/s41398-023-02371-z] [Medline: 36864017]
146. D'Costa A, Zatale A. AI and the cardiologist: when mind, heart and machine unite. Open Heart 2021 Dec;8(2):e001874. [doi: 10.1136/openhrt-2021-001874] [Medline: 34949649]
147. Hasan MM, Young GJ, Shi J, et al. A machine learning based two-stage clinical decision support system for predicting patients' discontinuation from opioid use disorder treatment: retrospective observational study. BMC Med Inform Decis Mak 2021 Nov 26;21(1):331. [doi: 10.1186/s12911-021-01692-7] [Medline: 34836524]
148. Kim HR, Sung M, Park JA, et al. Analyzing adverse drug reaction using statistical and machine learning methods: a systematic review. Medicine (Baltimore) 2022 Jun 24;101(25):e29387. [doi: 10.1097/MD.0000000000029387] [Medline: 35758373]
149. Zhang S, Bamakan SMH, Qu Q, Li S. Learning for personalized medicine: a comprehensive review from a deep learning perspective. IEEE Rev Biomed Eng 2019;12:194-208. [doi: 10.1109/RBME.2018.2864254] [Medline: 30106692]
150. Meng W, Zhang X, Ru B, Guan Y. A machine learning approach to real‐world time to treatment discontinuation prediction. Advanced Intelligent Systems 2023 Apr;5(4):2200254 [FREE Full text] [doi: 10.1002/aisy.202200254]
151. Gu Y, Zalkikar A, Liu M, et al. Predicting medication adherence using ensemble learning and deep learning models with large scale healthcare data. Sci Rep 2021 Sep 23;11(1):18961. [doi: 10.1038/s41598-021-98387-w] [Medline: 34556746]
152. Khan S, Sajjad M, Hussain T, Ullah A, Imran AS. A review on traditional machine learning and deep learning models for WBCs classification in blood smear images. IEEE Access 2020;9:10657-10673. [doi: 10.1109/ACCESS.2020.3048172]
153. De Bruyne S, De Kesel P, Oyaert M. Applications of artificial intelligence in urinalysis: is the future already here? Clin Chem 2023 Dec 1;69(12):1348-1360. [doi: 10.1093/clinchem/hvad136] [Medline: 37708293]
154. Goodswen SJ, Barratt JLN, Kennedy PJ, Kaufer A, Calarco L, Ellis JT. Machine learning and applications in microbiology. FEMS Microbiol Rev 2021 Sep 8;45(5):fuab015. [doi: 10.1093/femsre/fuab015] [Medline: 33724378]
155. Ghannam RB, Techtmann SM. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. Comput Struct Biotechnol J 2021;19:1092-1107. [doi: 10.1016/j.csbj.2021.01.028] [Medline: 33680353]

XSL•FO
RenderX

156. Unger M, Kather JN. Deep learning in cancer genomics and histopathology. Genome Med 2024 Mar 27;16(1):44. [doi: 10.1186/s13073-024-01315-6] [Medline: 38539231]
157. Obstfeld AE. Hematology and machine learning. J Appl Lab Med 2023 Jan 4;8(1):129-144. [doi: 10.1093/jalm/jfac108] [Medline: 36610431]
158. Danieli MG, Brunetto S, Gammeri L, et al. Machine learning application in autoimmune diseases: state of art and future prospectives. Autoimmun Rev 2024 Feb;23(2):103496. [doi: 10.1016/j.autrev.2023.103496] [Medline: 38081493]
159. Usategui I, Barbado J, Torres AM, Cascón J, Mateo J. Machine learning, a new tool for the detection of immunodeficiency patterns in systemic lupus erythematosus. J Investig Med 2023 Oct;71(7):742-752. [doi: 10.1177/10815589231171404] [Medline: 37158077]
160. Thomasian NM, Kamel IR, Bai HX. Machine intelligence in non-invasive endocrine cancer diagnostics. Nat Rev Endocrinol 2022 Feb;18(2):81-95. [doi: 10.1038/s41574-021-00543-9] [Medline: 34754064]
161. Hong N, Park H, Rhee Y. Machine learning applications in endocrinology and metabolism research: an overview. Endocrinol Metab 2020 Mar;35(1):71-84. [doi: 10.3803/EnM.2020.35.1.71]
162. Amjad A, Kordel P, Fernandes G. A review on innovation in healthcare sector (telehealth) through artificial intelligence. Sustainability 2023;15(8):6655. [doi: 10.3390/su15086655]
163. Gajarawala SN, Pelkowski JN. Telehealth benefits and barriers. J Nurse Pract 2021 Feb;17(2):218-221. [doi: 10.1016/j.nurpra.2020.09.013] [Medline: 33106751]
164. Schünke LC, Mello B, da Costa CA, et al. A rapid review of machine learning approaches for telemedicine in the scope of COVID-19. Artif Intell Med 2022 Jul;129:102312. [doi: 10.1016/j.artmed.2022.102312] [Medline: 35659388]
165. Hou Y, Huang J. Natural language processing for social science research: a comprehensive review. Chin J Sociol 2025 Jan;11(1):121-157. [doi: 10.1177/2057150X241306780]
166. Hughes A, Shandhi MMH, Master H, Dunn J, Brittain E. Wearable devices in cardiovascular medicine. Circ Res 2023 Mar 3;132(5):652-670. [doi: 10.1161/CIRCRESAHA.122.322389] [Medline: 36862812]
167. Liu Y, Wang B. Advanced applications in chronic disease monitoring using IoT mobile sensing device data, machine learning algorithms and frame theory: a systematic review. Front Public Health 2025;13:1510456. [doi: 10.3389/fpubh.2025.1510456]
168. Segal G, Segev A, Brom A, Lifshitz Y, Wasserstrum Y, Zimlichman E. Reducing drug prescription errors and adverse drug events by application of a probabilistic, machine-learning based clinical decision support system in an inpatient setting. J Am Med Inform Assoc 2019 Dec 1;26(12):1560-1565. [doi: 10.1093/jamia/ocz135] [Medline: 31390471]
169. Verma D, Bach K, Mork PJ. Application of machine learning methods on patient reported outcome measurements for predicting outcomes: a literature review. Informatics (MDPI) 2021;8(3):56. [doi: 10.3390/informatics8030056]
170. Björneld O, Carlsson M, Löwe W. Case study - Feature engineering inspired by domain experts on real world medical data. Intelligence-Based Medicine 2023;8:100110. [doi: 10.1016/j.ibmed.2023.100110]
171. Xu X, Li J, Zhu Z, et al. A comprehensive review on synergy of multi-modal data and AI technologies in medical diagnosis. Bioengineering (Basel) 2024 Feb 25;11(3):219. [doi: 10.3390/bioengineering11030219] [Medline: 38534493]
172. Zhou SK, Greenspan H, Davatzikos C, et al. A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. Proc IEEE 2021;109(5):820-838. [doi: 10.1109/JPROC.2021.3054390]
173. Khalifa M, Albadawy M. AI in diagnostic imaging: revolutionising accuracy and efficiency. Computer Methods and Programs in Biomedicine Update 2024;5:100146 [FREE Full text]
174. Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med 2021 Apr 7;4(1):65. [doi: 10.1038/s41746-021-00438-z] [Medline: 33828217]
175. Prasad VK, Verma A, Bhattacharya P, et al. Revolutionizing healthcare: a comparative insight into deep learning's role in medical imaging. Sci Rep 2024 Dec 4;14(1):30273. [doi: 10.1038/s41598-024-71358-7] [Medline: 39632902]
176. Zhang K, Yang X, Wang Y, et al. Artificial intelligence in drug development. Nat Med 2025 Jan;31(1):45-59. [doi: 10.1038/s41591-024-03434-4] [Medline: 39833407]
177. Kant S, Roy S. Artificial intelligence in drug discovery and development: transforming challenges into opportunities. Discov Pharm Sci 2025;1(1):7. [doi: 10.1007/s44395-025-00007-3]
178. Gomes B, Ashley EA. Artificial intelligence in molecular medicine. N Engl J Med 2023 Jun 29;388(26):2456-2465. [doi: 10.1056/NEJMra2204787] [Medline: 37379136]
179. Davis S, Zhang J, Lee I, et al. Effective hospital readmission prediction models using machine-learned features. BMC Health Serv Res 2022 Nov 24;22(1):1415. [doi: 10.1186/s12913-022-08748-y] [Medline: 36434628]
180. Dixon D, Sattar H, Moros N, et al. Unveiling the influence of AI predictive analytics on patient outcomes: a comprehensive narrative review. Cureus 2024 May;16(5):e59954. [doi: 10.7759/cureus.59954] [Medline: 38854327]
181. Oh EG, Oh S, Cho S, Moon M. Predicting readmission among high-risk discharged patients using a machine learning model with nursing data: retrospective study. JMIR Med Inform 2025 Mar 5;13:e56671. [doi: 10.2196/56671] [Medline: 40106364]
182. Varghese C, Harrison EM, O'Grady G, Topol EJ. Artificial intelligence in surgery. Nat Med 2024 May;30(5):1257-1268. [doi: 10.1038/s41591-024-02970-3] [Medline: 38740998]

XSL•FO
RenderX

183. Antel R, Abbasgholizadeh-Rahimi S, Guadagno E, Harley JM, Poenaru D. The use of artificial intelligence and virtual reality in doctor-patient risk communication: a scoping review. Patient Educ Couns 2022 Oct;105(10):3038-3050. [doi: 10.1016/j.pec.2022.06.006] [Medline: 35725526]
184. Wah JNK. Revolutionizing e-health: the transformative role of AI-powered hybrid chatbots in healthcare solutions. Front Public Health 2025;13:1530799. [doi: 10.3389/fpubh.2025.1530799] [Medline: 40017541]
185. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
186. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019 Jun;6(2):94-98. [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
187. Bravo F, Braun M, Farias V, et al. Optimization-driven framework to understand health care network costs and resource allocation. Health Care Manag Sci 2021 Sep;24(3):640-660. [doi: 10.1007/s10729-021-09565-1] [Medline: 33942227]
188. Wong F, de la Fuente-Nunez C, Collins JJ. Leveraging artificial intelligence in the fight against infectious diseases. Science 2023 Jul 14;381(6654):164-170. [doi: 10.1126/science.adh1114] [Medline: 37440620]
189. Alimadadi A, Aryal S, Manandhar I, Munroe PB, Joe B, Cheng X. Artificial intelligence and machine learning to fight COVID-19. Physiol Genomics 2020 Apr 1;52(4):200-202. [doi: 10.1152/physiolgenomics.00029.2020] [Medline: 32216577]
190. Bhardhwaj HS, Tutika S, Kaushik L.S. S, Gupta PK, Reddy K.S. T. Artificial intelligence in vaccine development. In: Zarrintaj P, Yazdi MK, Bencherif SA, Saeb MR, Mozafari M, editors. Artificial Intelligence in Biomaterials Design and Development: Woodhead Publishing; 2026:185-223. [doi: 10.1016/B978-0-323-95464-8.00012-8]
191. Wang L, Zhang Y, Wang D, et al. Artificial intelligence for COVID-19: a systematic review. Front Med 2021;8:704256. [doi: 10.3389/fmed.2021.704256]
192. Lo Vercio L, Amador K, Bannister JJ, et al. Supervised machine learning tools: a tutorial for clinicians. J Neural Eng 2020 Nov 19;17(6):062001. [doi: 10.1088/1741-2552/abbff2] [Medline: 33036008]
193. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. Radiographics 2017;37(2):505-515. [doi: 10.1148/rg.2017160130] [Medline: 28212054]
194. Kukreja S, Kumar A, Khan GA. A review paper on the diagnosis of lung cancer using machine learning. In: Dagur A, Singh K, Mehra PS, Shukla DK, editors. Artificial Intelligence, Blockchain, Computing and Security: CRC Press; 2023:15-19. [doi: 10.1201/9781032684994-3]
195. Bertsimas D, Wiberg H. Machine learning in oncology: methods, applications, and challenges. JCO Clin Cancer Inform 2020 Oct;4:885-894. [doi: 10.1200/CCI.20.00072] [Medline: 33058693]
196. Asri H, Mousannif H, Moatassime HA, Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Comput Sci 2016;83:1064-1069. [doi: 10.1016/j.procs.2016.04.224]
197. Talwar A, Lopez-Olivo MA, Huang Y, Ying L, Aparasu RR. Performance of advanced machine learning algorithms overlogistic regression in predicting hospital readmissions: a meta-analysis. Explor Res Clin Soc Pharm 2023 Sep;11:100317. [doi: 10.1016/j.rcsop.2023.100317] [Medline: 37662697]
198. Colace F, Gupta BB, Lorusso A, Troiano A, Santaniello D, Valentino C. Unsupervised learning techniques for vibration-based structural health monitoring systems driven by data: a general overview. In: Gupta BB, Colace F, editors. Handbook of Research on AI and ML for Intelligent Machines and Systems: IGI Global; 2024:305-347. [doi: 10.4018/978-1-6684-9999-3.ch013]
199. Makino M, Shimizu K, Kadota K. Enhanced clustering-based differential expression analysis method for RNA-seq data. MethodsX 2024 Jun;12:102518. [doi: 10.1016/j.mex.2023.102518] [Medline: 38179066]
200. Kauffman J, Miotto R, Klang E, et al. Embedding methods for electronic health record research. Annu Rev Biomed Data Sci 2025 Aug;8(1):563-590. [doi: 10.1146/annurev-biodatasci-103123-094729] [Medline: 40312284]
201. Wu J, Cui Y, Sun X, et al. Unsupervised clustering of quantitative image phenotypes reveals breast cancer subtypes with distinct prognoses and molecular pathways. Clin Cancer Res 2017 Jul 1;23(13):3334-3342. [doi: 10.1158/1078-0432.CCR-16-2415] [Medline: 28073839]
202. Ringnér M. What is principal component analysis? Nat Biotechnol 2008 Mar;26(3):303-304. [doi: 10.1038/nbt0308-303] [Medline: 18327243]
203. van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579-2605 [FREE Full text]
204. Smith SM. Fast robust automated brain extraction. Hum Brain Mapp 2002 Nov;17(3):143-155. [doi: 10.1002/hbm.10062] [Medline: 12391568]
205. Menze BH, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging 2015 Oct;34(10):1993-2024. [doi: 10.1109/TMI.2014.2377694] [Medline: 25494501]
206. Qiu L, Cheng J, Gao H, Xiong W, Ren H. Federated semi-supervised learning for medical image segmentation via pseudo-label denoising. IEEE J Biomed Health Inform 2023 Oct;27(10):4672-4683. [doi: 10.1109/JBHI.2023.3274498] [Medline: 37155394]
207. Yang X, Song Z, King I, Xu Z. A survey on deep semi-supervised learning. IEEE Trans Knowl Data Eng 2022;35(9):8934-8954. [doi: 10.1109/TKDE.2022.3220219]
208. Eckardt JN, Bornhäuser M, Wendt K, Middeke JM. Semi-supervised learning in cancer diagnostics. Front Oncol 2022;12:960984. [doi: 10.3389/fonc.2022.960984] [Medline: 35912249]

XSL•FO
RenderX

209. Jiao R, Zhang Y, Ding L, et al. Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. Comput Biol Med 2024 Feb;169:107840. [doi: 10.1016/j.compbiomed.2023.107840] [Medline: 38157773]
210. Yu C, Liu J, Nemati S, Yin G. Reinforcement learning in healthcare: a survey. ACM Comput Surv 2023 Jan 31;55(1):1-36. [doi: 10.1145/3477600]
211. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 2018 Nov;24(11):1716-1720. [doi: 10.1038/s41591-018-0213-5] [Medline: 30349085]
212. Chi W. Context-aware learning for robot-assisted endovascular catheterization [Dissertation]. : Imperial College London; 2019 URL: https://spiral.imperial.ac.uk/entities/publication/17ec29f1-20bb-4171-95c5-7249267a5478 [accessed 2026-05-07] [doi: 10.25560/80437]
213. Wang Y, Zhao Y, Petzold L. Predicting the need for blood transfusion in intensive care units with reinforcement learning. Presented at: 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics; Aug 7-10, 2022. [doi: 10.1145/3535508.3545523]
214. McLaverty B. Unifying data-driven modeling with machine learning to improve personalized treatment of critical care patients [Dissertation]. : University of Pittsburgh; 2023 URL: https://d-scholarship.pitt.edu/concern/etds/1325d248-7852-4eba-aa0e-0a43f2fa4243 [accessed 2026-05-07]
215. Almagrabi AO, Ali R, Alghazzawi D, AlBarakati A, Khurshaid T. A reinforcement learning-based framework for crowdsourcing in massive health care Internet of Things. Big Data 2022 Apr;10(2):161-170. [doi: 10.1089/big.2021.0058] [Medline: 34319812]
216. Rashidi HH, Albahra S, Robertson S, Tran NK, Hu B. Common statistical concepts in the supervised machine learning arena. Front Oncol 2023;13:1130229. [doi: 10.3389/fonc.2023.1130229] [Medline: 36845729]
217. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature New Biol 2017 Feb 2;542(7639):115-118. [doi: 10.1038/nature21056]
218. Shah D, Patel S, Bharti SK. Heart disease prediction using machine learning techniques. SN Comput Sci 2020 Nov;1(6):1-6. [doi: 10.1007/s42979-020-00365-y]
219. Levin S, Toerper M, Hamrock E, et al. Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index. Ann Emerg Med 2018 May;71(5):565-574. [doi: 10.1016/j.annemergmed.2017.08.005] [Medline: 28888332]
220. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Front Genet 2018;9:515. [doi: 10.3389/fgene.2018.00515] [Medline: 30459809]
221. Mucaki EJ, Zhao JZL, Lizotte DJ, Rogan PK. Predicting responses to platin chemotherapy agents with biochemically-inspired machine learning. Signal Transduct Target Ther 2019;4(1):1. [doi: 10.1038/s41392-018-0034-5] [Medline: 30652029]
222. Battineni G, Sagaro GG, Chinatalapudi N, Amenta F. Applications of machine learning predictive models in the chronic disease diagnosis. J Pers Med 2020 Mar 31;10(2):21. [doi: 10.3390/jpm10020021] [Medline: 32244292]
223. Dhiman P, Ma J, Andaur Navarro CL, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. BMC Med Res Methodol 2022 Apr 8;22(1):101. [doi: 10.1186/s12874-022-01577-x] [Medline: 35395724]
224. Palomino-Echeverria S, Huergo E, Ortega-Legarreta A, et al. A robust clustering strategy for stratification unveils unique patient subgroups in acutely decompensated cirrhosis. J Transl Med 2024 Jun 27;22(1):599. [doi: 10.1186/s12967-024-05386-2] [Medline: 38937846]
225. Sinha A, Aljrees T, Pandey SK, et al. Semi-supervised clustering-based DANA algorithm for data gathering and disease detection in healthcare wireless sensor networks (WSN). Sensors (Basel) 2023 Dec 19;24(1):18. [doi: 10.3390/s24010018] [Medline: 38202880]
226. Ren Z, Yeh RA, Schwing AG. Not all unlabeled data are equal: learning to weight data in semi-supervised learning. Presented at: 34th International Conference on Neural Information Processing System; Dec 6-12, 2020. [doi: 10.5555/3495724.3497552]
227. Huang L. Combination of information in labeled and unlabeled data via evidence theory. IEEE Trans Artif Intell 2023;5(5):2179-2192. [doi: 10.1109/TAI.2023.3316194]
228. Qiu S, Chen Y, Yang Y, et al. A review on semi-supervised learning for EEG-based emotion recognition. Information Fusion 2024 Apr;104:102190. [doi: 10.1016/j.inffus.2023.102190]
229. Christopoulou SC. Machine learning models and technologies for evidence-based telehealth and smart care: a review. BioMedInformatics 2024;4(1):754-779. [doi: 10.3390/biomedinformatics4010042]
230. Naeem M, Rizvi STH, Coronato A. A gentle introduction to reinforcement learning and its application in different fields. IEEE Access 2020;8:209320-209344. [doi: 10.1109/ACCESS.2020.3038605]
231. Sverdlov O, Ryeznik Y, Wong WK. Opportunity for efficiency in clinical development: an overview of adaptive clinical trial designs and innovative machine learning tools, with examples from the cardiovascular field. Contemp Clin Trials 2021 Jun;105:106397. [doi: 10.1016/j.cct.2021.106397] [Medline: 33845209]

XSL•FO
RenderX

232. Kumar CUO, Singh I, Suguna M. Optimizing patient recruitment for clinical trials: a hybrid classification model and game-theoretic approach for strategic interaction. IEEE Access 2024;12:10254-10280. [doi: 10.1109/ACCESS.2024.3351688]
233. Malheiro V, Santos B, Figueiras A, Mascarenhas-Melo F. The potential of artificial intelligence in pharmaceutical innovation: from drug discovery to clinical trials. Pharmaceuticals (Basel) 2025 May 25;18(6):788. [doi: 10.3390/ph18060788] [Medline: 40573185]
234. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature New Biol 2020 Jan 2;577(7788):89-94. [doi: 10.1038/s41586-019-1799-6]
235. Zhavoronkov A, Ivanenkov YA, Aliper A, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol 2019 Sep;37(9):1038-1040. [doi: 10.1038/s41587-019-0224-x] [Medline: 31477924]
236. Adams R, Henry KE, Sridharan A, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat Med 2022 Jul;28(7):1455-1460. [doi: 10.1038/s41591-022-01894-0] [Medline: 35864252]
237. Piening B, Bapat B, Weerasinghe RK, et al. Improved outcomes from reflex comprehensive genomic profiling-guided precision therapeutic selection across a major US healthcare system. JCO 2023 Jun 1;41(16_suppl):6622-6622. [doi: 10.1200/JCO.2023.41.16_suppl.6622]
238. Iyortsuun NK, Kim SH, Jhon M, Yang HJ, Pant S. A review of machine learning and deep learning approaches on mental health diagnosis. Healthcare (Basel) 2023 Jan 17;11(3):285. [doi: 10.3390/healthcare11030285] [Medline: 36766860]
239. Konečný J, McMahan HB, Ramage D, Richtárik P. Federated optimization: distributed machine learning for on-device intelligence. arXiv. Preprint posted online on Oct 8, 2016 URL: https://arxiv.org/abs/1610.02527 [accessed 2026-05-07]
240. Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: concept and applications. ACM Transactions on Intelligent Systems and Technology 2019;10(2):1-19. [doi: 10.1145/3298981]
241. Zhang X, Mavromatis A, Vafeas A, Nejabati R, Simeonidou D. Federated feature selection for horizontal federated learning in IoT networks. IEEE Internet Things J 2023;10(11):10095-10112. [doi: 10.1109/JIOT.2023.3237032]
242. Gao D, Ju C, Wei X, Liu Y, Chen T, Yang Q. HHHFL: hierarchical heterogeneous horizontal federated learning for electroencephalography. arXiv. Preprint posted online on Sep 11, 2019 URL: https://arxiv.org/abs/1909.05784 [accessed 2026-05-07]
243. Huang W, Li T, Wang D, Du S, Zhang J, Huang T. Fairness and accuracy in horizontal federated learning. Inf Sci (Ny) 2022 Apr;589:170-185. [doi: 10.1016/j.ins.2021.12.102]
244. Liu Y, Kang Y, Zou T, et al. Vertical federated learning: concepts, advances, and challenges. IEEE Trans Knowl Data Eng 2024;36(7):3615-3634. [doi: 10.1109/TKDE.2024.3352628]
245. Feng S, Yu H, Zhu Y. MMVFL: a simple vertical federated learning framework for multi-class multi-participant scenarios. Sensors (Basel) 2024 Jan 18;24(2):619. [doi: 10.3390/s24020619] [Medline: 38257712]
246. Gupta M, Sharma P, Kalra R. Federated learning and artificial intelligence in e-healthcare. In: Hassan A, Prasad VK, Bhattacharya P, Dutta P, Damaševičius R, editors. Federated Learning and AI for Healthcare 5.0: IGI Global; 2024:104-118. [doi: 10.4018/979-8-3693-1082-3.ch006]
247. Liu Y, Kang Y, Xing C, Chen T, Yang Q. A secure federated transfer learning framework. IEEE Intell Syst 2020;35(4):70-82. [doi: 10.1109/MIS.2020.2988525]
248. Chen Y, Qin X, Wang J, Yu C, Gao W. FedHealth: a federated transfer learning framework for wearable healthcare. IEEE Intell Syst 2020;35(4):83-93. [doi: 10.1109/MIS.2020.2988604]
249. Chorney W, Wang H. Towards federated transfer learning in electrocardiogram signal analysis. Comput Biol Med 2024 Mar;170:107984. [doi: 10.1016/j.compbiomed.2024.107984] [Medline: 38244469]
250. Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. NPJ Digit Med 2020;3(1):119. [doi: 10.1038/s41746-020-00323-1] [Medline: 33015372]
251. Nguyen DC, Pham QV, Pathirana PN, et al. Federated learning for smart healthcare: a survey. ACM Comput Surv 2023 Mar 31;55(3):1-37. [doi: 10.1145/3501296]
252. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. IEEE Signal Process Mag 2020;37(3):50-60. [doi: 10.1109/MSP.2020.2975749]
253. Heyndrickx W, Mervin L, Morawietz T, et al. MELLODDY: cross-pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information. J Chem Inf Model 2024 Apr 8;64(7):2331-2344. [doi: 10.1021/acs.jcim.3c00799] [Medline: 37642660]
254. Foley P, Sheller MJ, Edwards B, et al. OpenFL: the open federated learning library. Phys Med Biol 2022 Oct 19;67(21):214001. [doi: 10.1088/1361-6560/ac97d9] [Medline: 36198326]
255. Mora A, Bujari A, Bellavista P. Enhancing generalization in federated learning with heterogeneous data: a comparative literature review. Future Generation Computer Systems 2024 Aug;157:1-15. [doi: 10.1016/j.future.2024.03.027]
256. Almanifi ORA, Chow CO, Tham ML, Chuah JH, Kanesan J. Communication and computation efficiency in federated learning: a survey. Internet of Things 2023 Jul;22:100742. [doi: 10.1016/j.iot.2023.100742]
257. Dhatterwal JS, Malik K, Kaswan KS, Elngar AA. Federated learning: overview, challenges, and ethical considerations. In: Elngar AA, Oliva D, Balas VE, editors. Artificial Intelligence Using Federated Learning: CRC Press; 2024:1-15. [doi: 10.1201/9781003482000-1]

XSL•FO
RenderX

258. Mohammadi S, Balador A, Sinaei S, Flammini F. Balancing privacy and performance in federated learning: a systematic literature review on methods and metrics. J Parallel Distrib Comput 2024 Oct;192:104918. [doi: 10.1016/j.jpdc.2024.104918]
259. Saha S, Hota A, Chattopadhyay AK, Nag A, Nandi S. A multifaceted survey on privacy preservation of federated learning: progress, challenges, and opportunities. Artif Intell Rev 2024;57(7):184. [doi: 10.1007/s10462-024-10766-7]
260. Narula M, Meena J, Vishwakarma DK. A comprehensive review on federated learning for data-sensitive application: open issues & challenges. Eng Appl Artif Intell 2024 Jul;133:108128. [doi: 10.1016/j.engappai.2024.108128]
261. Malik H, Anees T. Federated learning with deep convolutional neural networks for the detection of multiple chest diseases using chest x-rays. Multimed Tools Appl 2024;83(23):63017-63045. [doi: 10.1007/s11042-023-18065-z]
262. Dasaradharami Reddy K, Gadekallu TR. A comprehensive survey on federated learning techniques for healthcare informatics. Comput Intell Neurosci 2023;2023(1):8393990. [doi: 10.1155/2023/8393990] [Medline: 36909974]
263. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell 2020;2(6):305-311. [doi: 10.1038/s42256-020-0186-1]
264. Rafi TH, Noor FA, Hussain T, Chae DK. Fairness and privacy preserving in federated learning: a survey. Information Fusion 2024 May;105:102198. [doi: 10.1016/j.inffus.2023.102198]
265. Djebrouni Y, Benarba N, Touat O, et al. Bias mitigation in federated learning for edge computing. Proc ACM Interact Mob Wearable Ubiquitous Technol 2023 Dec 19;7(4):1-35. [doi: 10.1145/3631455]
266. You L, Guo Z, Zuo B, Chang Y, Yuen C. SLMFed: a stage-based and layerwise mechanism for incremental federated learning to assist dynamic and ubiquitous IoT. IEEE Internet Things J 2024;11(9):16364-16381. [doi: 10.1109/JIOT.2024.3353793]
267. Mazzocca C, Romandini N, Montanari R, Bellavista P. Enabling federated learning at the edge through the IOTA tangle. Future Generation Computer Systems 2024 Mar;152:17-29. [doi: 10.1016/j.future.2023.10.014]
268. Ji J, Shu Z, Li H, et al. Edge-computing-based knowledge distillation and multitask learning for partial discharge recognition. IEEE Trans Instrum Meas 2024;73:1-11. [doi: 10.1109/TIM.2024.3351239]
269. Elhattab F, Bouchenak S, Boscher C. PASTEL: privacy-preserving federated learning in edge computing. Proc ACM Interact Mob Wearable Ubiquitous Technol 2024;7(4):1-29. [doi: 10.1145/3633808]
270. Nakamoto S. Bitcoin: a peer-to-peer electronic cash system. : Bitcoin.org; 2008 URL: https://bitcoin.org/bitcoin.pdf [accessed 2026-05-07]
271. Wood G. Ethereum: a secure decentralised generalised transaction ledger. : Ethereum; 2025 URL: https://ethereum.github.io/yellowpaper/paper.pdf [accessed 2026-06-01]
272. Dutta P, Choi TM, Somani S, Butala R. Blockchain technology in supply chain operations: applications, challenges and research opportunities. Transp Res E Logist Transp Rev 2020 Oct;142:102067. [doi: 10.1016/j.tre.2020.102067] [Medline: 33013183]
273. Casado-Vara R, Prieto J, la Prieta FD, Corchado JM. How blockchain improves the supply chain: case study alimentary supply chain. Procedia Comput Sci 2018;134:393-398. [doi: 10.1016/j.procs.2018.07.193]
274. Korpela K, Hallikas J, Dahlberg T. Digital supply chain transformation toward blockchain integration. Presented at: 50th Hawaii International Conference on System Sciences; Jan 4-7, 2017. [doi: 10.24251/HICSS.2017.506]
275. Queiroz MM, Telles R, Bonilla SH. Blockchain and supply chain management integration: a systematic review of the literature. Supply Chain Management 2019 Aug 22;25(2):241-254. [doi: 10.1108/SCM-03-2018-0143]
276. Raval S. Decentralized Applications: Harnessing Bitcoin's Blockchain Technology: O'Reilly Media, Inc; 2016. [doi: 10.5555/3074145]
277. Joshi AP, Han M, Wang Y. A survey on security and privacy issues of blockchain technology. Mathematical Foundations of Computing 2018;1(2):121-147. [doi: 10.3934/mfc.2018007]
278. Mohanta BK, Jena D, Panda SS, Sobhanayak S. Blockchain technology: a survey on applications and security privacy challenges. Internet of Things 2019 Dec;8:100107. [doi: 10.1016/j.iot.2019.100107]
279. Ul Hassan M, Rehmani MH, Chen J. Anomaly detection in blockchain networks: a comprehensive survey. IEEE Commun Surv Tutorials 2022;25(1):289-318. [doi: 10.1109/COMST.2022.3205643]
280. Shahsavari Y, Zhang K, Talhi C. A theoretical model for block propagation analysis in bitcoin network. IEEE Trans Eng Manage 2020;69(4):1459-1476. [doi: 10.1109/TEM.2020.2989170]
281. Belchior R, Vasconcelos A, Guerreiro S, Correia M. A survey on blockchain interoperability: past, present, and future trends. ACM Comput Surv 2022 Nov 30;54(8):1-41. [doi: 10.1145/3471140]
282. Lafourcade P, Lombard-Platet M. About blockchain interoperability. Inf Process Lett 2020 Sep;161:105976. [doi: 10.1016/j.ipl.2020.105976]
283. Schulte S, Sigwart M, Frauenthaler P, Borkowski M. Towards blockchain interoperability. In: Di Ciccio C, editor. Business Process Management: Blockchain and Central and Eastern Europe Forum BPM 2019 Lecture Notes in Business Information Processing: Springer; 2019:3-10. [doi: 10.1007/978-3-030-30429-4_1]
284. Hardjono T, Lipton A, Pentland A. Toward an interoperability architecture for blockchain autonomous systems. IEEE Trans Eng Manage 2019;67(4):1298-1309. [doi: 10.1109/TEM.2019.2920154]
285. Zhang P, White J, Schmidt DC, Lenz G. Applying software patterns to address interoperability in blockchainbased healthcare apps. arXiv. Preprint posted online on Jun 5, 2017 URL: https://arxiv.org/abs/1706.03700 [accessed 2026-05-07]

XSL•FO
RenderX

286. Qasse IA, Abu Talib M, Nasir Q. Inter blockchain communication. Presented at: 6th Annual International Conference Research Track; Mar 7-9, 2019. [doi: 10.1145/3333165.3333167]
287. Zou W, Lo D, Kochhar PS, et al. Smart contract development: challenges and opportunities. IIEEE Trans Software Eng 2019;47(10):2084-2106. [doi: 10.1109/TSE.2019.2942301]
288. Delmolino K, Arnett M, Kosba A, Miller A, Shi E. Step by step towards creating a safe smart contract: lessons and insights from a cryptocurrency lab. In: Clark J, Meiklejohn S, Ryan P, Wallach D, Brenner M, Rohloff K, editors. Financial Cryptography and Data Security FC 2016 Lecture Notes in Computer Science: Springer; 2016:79-94. [doi: 10.1007/978-3-662-53357-4_6]
289. Xu J, Wang C, Jia X. A survey of blockchain consensus protocols. ACM Comput Surv 2023 Dec 31;55(13s):1-35. [doi: 10.1145/3579845]
290. Nguyen GT, Kim K. A survey about consensus algorithms used in blockchain. J Inf Process Syst 2018;14(1):101-128. [doi: 10.3745/JIPS.01.0024]
291. Monrat AA, Schelen O, Andersson K. A survey of blockchain from the perspectives of applications, challenges, and opportunities. IEEE Access 2019;7:117134-117151. [doi: 10.1109/ACCESS.2019.2936094]
292. Gervais A, Karame GO, Wüst K, Glykantzis V, Ritzdorf H, Capkun S. On the security and performance of proof of work blockchains. Presented at: 2016 ACM SIGSAC Conference on Computer and Communications Security; Oct 24-28, 2016. [doi: 10.1145/2976749.2978341]
293. Shahsavari Y, Zhang K, Talhi C. A theoretical model for fork analysis in the bitcoin network. Presented at: 2019 IEEE International Conference on Blockchain (Blockchain); Jul 14-17, 2019. [doi: 10.1109/Blockchain.2019.00038]
294. Nguyen CT, Hoang DT, Nguyen DN, Niyato D, Nguyen HT, Dutkiewicz E. Proof-of-stake consensus mechanisms for future blockchain networks: fundamentals, applications and opportunities. IEEE Access 2019;7:85727-85745. [doi: 10.1109/ACCESS.2019.2925010] [Medline: 33777591]
295. Buterin V. A next-generation smart contract and decentralized application platform. : Ethereum; 2014 URL: https://cryptorating.eu/whitepapers/Ethereum/Ethereum_white_paper.pdf [accessed 2026-06-01]
296. Introduction. Cardano Documentation. URL: https://docs.cardano.org/introduction [accessed 2024-02-22]
297. Gilad Y, Hemo R, Micali S, Vlachos G, Zeldovich N. Algorand: scaling byzantine agreements for cryptocurrencies. Presented at: 26th Symposium on Operating Systems Principles; Oct 28, 2017. [doi: 10.1145/3132747.3132757]
298. Wang Q, Xu M, Li X, Qian H. Revisiting the fairness and randomness of delegated proof of stake consensus algorithm. Presented at: 2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom); Dec 17-19, 2020. [doi: 10.1109/ISPA-BDCloud-SocialCom-SustainCom51426.2020.00064]
299. Lamport L, Shostak R, Pease M. The Byzantine generals problem. In: Malkhi D, editor. Concurrency: The Works of Leslie Lamport 2019:203-226. [doi: 10.1145/3335772.3335936]
300. Zhong W, Yang C, Liang W, et al. Byzantine fault-tolerant consensus algorithms: a survey. Electronics (Basel) 2023;12(18):3801. [doi: 10.3390/electronics12183801]
301. Shahsavari Y, Zhang K, Talhi C. Performance modeling and analysis of hotstuff for blockchain consensus. Presented at: Fourth International Conference on Blockchain Computing and Applications (BCCA); Sep 5-7, 2022. [doi: 10.1109/BCCA55292.2022.9922238]
302. Androulaki E, Barger A, Bortnikov V, et al. Hyperledger fabric: a distributed operating system for permissioned blockchains. Presented at: Thirteenth EuroSys Conference; Apr 23-26, 2018. [doi: 10.1145/3190508.3190538]
303. Li Y, Cao B, Peng M, et al. Direct acyclic graph-based ledger for Internet of Things: performance and security analysis. IEEE/ACM Trans Networking 2020;28(4):1643-1656. [doi: 10.1109/TNET.2020.2991994]
304. Popov S. The tangle. Semantic Scholar. 2018. URL: https://www.semanticscholar.org/paper/The-Tangle-Popov/43586b34b054b48891d478407d4e7435702653e0 [accessed 2026-05-07]
305. Saa O, Cullen A, Vigneri L. IOTA 2.0 incentives and tokenomics whitepaper. : IOTA; 2023 URL: https://files.iota.org/papers/IOTA_2.0_Incentives_And_Tokenomics_Whitepaper.pdf [accessed 2026-05-07]
306. Pass R, Shi E. Hybrid consensus: efficient consensus in the permissionless model. Cryptology ePrint Archive. 2016. URL: https://eprint.iacr.org/2016/917 [accessed 2026-05-07]
307. Chepurnoy A, Duong T, Fan L, Zhou HS. Twinscoin: a cryptocurrency via proof-of-work and proof-of-stake. Cryptology ePrint Archive. 2017. URL: https://eprint.iacr.org/2017/232 [accessed 2026-05-07]
308. Kwon J. Tendermint: consensus without mining. : Tendermint; 2014 URL: https://tendermint.com/static/docs/tendermint.pdf [accessed 2026-05-07]
309. Cheng Z, Wu G, Wu H, Zhao M, Zhao L, Cai Q. A new hybrid consensus protocol: deterministic proof of work. arXiv. Preprint posted online on Aug 13, 2018 URL: https://arxiv.org/abs/1808.04142 [accessed 2026-05-07]
310. Jayakumari B, Sheeba SL, Eapen M, et al. E-voting system using cloud-based hybrid blockchain technology. J Saf Sci Resil 2024 Mar;5(1):102-109. [doi: 10.1016/j.jnlssr.2024.01.002]
311. Zarrin J, Wen Phang H, Babu Saheer L, Zarrin B. Blockchain for decentralization of internet: prospects, trends, and challenges. Cluster Comput 2021;24(4):2841-2866. [doi: 10.1007/s10586-021-03301-8] [Medline: 34025209]

XSL•FO
RenderX

312. Dinh TTA, Liu R, Zhang M, Chen G, Ooi BC, Wang J. Untangling blockchain: a data processing view of blockchain systems. IEEE Trans Knowl Data Eng 2018;30(7):1366-1385. [doi: 10.1109/TKDE.2017.2781227]
313. Decker C, Wattenhofer R. Information propagation in the bitcoin network. Presented at: International Conference on Peer-to-Peer Computing; Sep 9-11, 2013. [doi: 10.1109/P2P.2013.6688704]
314. Kan L, Wei Y, Hafiz Muhammad A, Siyuan W, Gao LC, Kai H. A multiple blockchains architecture on inter-blockchain communication. Presented at: 2018 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C); Jul 16-20, 2018. [doi: 10.1109/QRS-C.2018.00037]
315. Chen ZD, Yu Z, Duan ZB, Hu K. Inter-blockchain communication. DEStech Transactions on Computer Science and Engineering 2017:12539. [doi: 10.12783/dtcse/cst2017/12539]
316. Tam Vo H, Wang Z, Karunamoorthy D, Wagner J, Abebe E, Mohania M. Internet of blockchains: techniques and challenges ahead. Presented at: 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData); Jul 30 to Aug 3, 2018. [doi: 10.1109/Cybermatics_2018.2018.00264]
317. Kwon J, Buchman E. Cosmos: a network of distributed ledgers. GitHub. URL: https://github.com/cosmos/cosmos/blob/master/WHITEPAPER.md [accessed 2026-05-07]
318. Wood G. Polkadot: vision for a heterogeneous multi-chain framework. : Polkadot; 2016 URL: https://assets.polkadot.network/Polkadot-whitepaper.pdf [accessed 2026-05-07]
319. Wanchain. URL: https://www.wanchain.org/ [accessed 2024-03-11]
320. Mahajan HB, Junnarkar AA. Smart healthcare system using integrated and lightweight ECC with private blockchain for multimedia medical data processing. Multimed Tools Appl 2023 Apr 15;82(28):1-24. [doi: 10.1007/s11042-023-15204-4] [Medline: 37362704]
321. Pahlajani S, Kshirsagar A, Pachghare V. Survey on private blockchain consensus algorithms. Presented at: 1st International Conference on Innovations in Information and Communication Technology (ICIICT); Apr 25-26, 2019. [doi: 10.1109/ICIICT1.2019.8741353]
322. Antwi M, Adnane A, Ahmad F, Hussain R, Habib ur Rehman M, Kerrache CA. The case of HyperLedger fabric as a blockchain solution for healthcare applications. Blockchain: Research and Applications 2021 Mar;2(1):100012. [doi: 10.1016/j.bcra.2021.100012]
323. Al-Sumaidaee G, Alkhudary R, Zilic Z, Swidan A. Performance analysis of a private blockchain network built on hyperledger fabric for healthcare. Inf Process Manag 2023 Mar;60(2):103160. [doi: 10.1016/j.ipm.2022.103160]
324. Irresberger F, John K, Mueller P, Saleh F. The public blockchain ecosystem: an empirical analysis. : NYU Stern School of Business; 2020. [doi: 10.2139/ssrn.3592849]
325. Ferdous MS, Chowdhury MJM, Hoque MA. A survey of consensus algorithms in public blockchain systems for crypto-currencies. J Netw Comput Appl 2021 May;182:103035. [doi: 10.1016/j.jnca.2021.103035]
326. Benhamouda F, Gentry C, Gorbunov S, et al. Can a public blockchain keep a secret? In: Pass R, Pietrzak K, editors. Theory of Cryptography TCC 2020 Lecture Notes in Computer Science: Springer; 2020:260-290. [doi: 10.1007/978-3-030-64375-1_10]
327. Rebello GAF, Camilo GF, de Souza LAC, et al. A survey on blockchain scalability: from hardware to layer-two protocols. IEEE Commun Surv Tutorials 2024;26(4):2411-2458. [doi: 10.1109/COMST.2024.3376252]
328. Dib O, Brousmiche KL, Durand A, Thea E, Hamida EB. Consortium blockchains: overview, applications and challenges. Int J Adv Telecommun 2018;11(1):51-64 [FREE Full text]
329. Li Z, Kang J, Yu R, Ye D, Deng Q, Zhang Y. Consortium blockchain for secure energy trading in industrial Internet of Things. IEEE Trans Ind Inf 2017;14(8):1-1. [doi: 10.1109/TII.2017.2786307]
330. Yao W, Ye J, Murimi R, Wang G. A survey on consortium blockchain consensus mechanisms. arXiv. Preprint posted online on Feb 24, 2021 URL: https://arxiv.org/abs/2102.12058v2 [accessed 2026-05-07]
331. Zhang A, Lin X. Towards secure and privacy-preserving data sharing in e-health systems via consortium blockchain. J Med Syst 2018 Jun 28;42(8):140. [doi: 10.1007/s10916-018-0995-5] [Medline: 29956061]
332. Li D, Han D, Weng TH, et al. Blockchain for federated learning toward secure distributed machine learning systems: a systemic survey. Soft Comput 2022;26(9):4423-4440. [doi: 10.1007/s00500-021-06496-5] [Medline: 34840525]
333. Houda ZAE, Hafid AS, Khoukhi L, Brik B. When collaborative federated learning meets blockchain to preserve privacy in healthcare. IEEE Trans Netw Sci Eng 2022;10(5):2455-2465. [doi: 10.1109/TNSE.2022.3211192]
334. Issa W, Moustafa N, Turnbull B, Sohrabi N, Tari Z. Blockchain-based federated learning for securing Internet of Things: a comprehensive survey. ACM Comput Surv 2023 Sep 30;55(9):1-43. [doi: 10.1145/3560816]
335. Passerat-Palmbach J, Farnan T, McCoy M, et al. Blockchain-orchestrated machine learning for privacy preserving federated learning in electronic health data. Presented at: 2020 IEEE International Conference on Blockchain (Blockchain); Nov 2-6, 2020. [doi: 10.1109/Blockchain50366.2020.00080]
336. Aich S, Sinai NK, Kumar S, et al. Protecting personal healthcare record using blockchain & federated learning technologies. Presented at: 24th International Conference on Advanced Communication Technology (ICACT); Feb 13-16, 2022. [doi: 10.23919/ICACT53585.2022.9728772]

XSL•FO
RenderX

337. El Rifai O, Biotteau M, de Boissezon X, Megdiche I, Ravat F, Teste O. Blockchain-based federated learning in medicine. In: Michalowski M, Moskovitch R, editors. Artificial Intelligence in Medicine AIME 2020 Lecture Notes in Computer Science: Springer; 2020:214-224. [doi: 10.1007/978-3-030-59137-3_20]
338. Kim YJ, Hong CS. Blockchain-based node-aware dynamic weighting methods for improving federated learning performance. Presented at: 20th Asia-Pacific Network Operations and Management Symposium (APNOMS); Sep 18-20, 2019. [doi: 10.23919/APNOMS.2019.8893114]
339. Lu Y, Huang X, Zhang K, Maharjan S, Zhang Y. Communication-efficient federated learning and permissioned blockchain for digital twin edge networks. IEEE Internet Things J 2020;8(4):2276-2288. [doi: 10.1109/JIOT.2020.3015772]
340. Pandey SR, Nguyen LD, Popovski P. FedToken: tokenized incentives for data contribution in federated learning. arXiv. Preprint posted online on Sep 20, 2022 URL: https://arxiv.org/abs/2209.09775 [accessed 2026-05-07]
341. Behera MR, Upadhyay S, Shetty S. Federated learning using smart contracts on blockchains, based on reward driven approach. arXiv. Preprint posted online on Jul 19, 2021 URL: https://arxiv.org/abs/2107.10243 [accessed 2026-05-07]
342. Ma S, Cao Y, Xiong L. Transparent contribution evaluation for secure federated learning on blockchain. Presented at: 37th International Conference on Data Engineering Workshops (ICDEW); Apr 19-22, 2021. [doi: 10.1109/ICDEW53142.2021.00023]
343. Wang Z, Hu Q. Blockchain-based federated learning: a comprehensive survey. arXiv. Preprint posted online on Oct 5, 2021 URL: https://arxiv.org/abs/2110.02182v1 [accessed 2026-05-07]
344. Munusamy S, Jothi KR. Blockchain-enabled federated learning with edge analytics for secure and efficient electronic health records management. Sci Rep 2025 Jul 28;15(1):27524. [doi: 10.1038/s41598-025-12225-x] [Medline: 40721766]
345. Castro M, Liskov B. Practical Byzantine fault tolerance. Presented at: Third Symposium on Operating Systems Design and Implementation; Feb 22-25, 1999. [doi: 10.5555/296806.296824]
346. Castro M, Liskov B. Practical Byzantine fault tolerance and proactive recovery. ACM Trans Comput Syst 2002 Nov;20(4):398-461. [doi: 10.1145/571637.571640]
347. McMahan B, McMahan E, Ramage D, Hampson S, Aguera y Arcas B. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics: PMLR; 2017:1273-1282 URL: https://proceedings.mlr.press/v54/mcmahan17a [accessed 2026-05-07]
348. Buchman E. Tendermint: Byzantine fault tolerance in the age of blockchains [Dissertation]. : University of Guelph; 2016 URL: https://atrium.lib.uoguelph.ca/items/5459099e-67aa-4a23-83ae-d3471d8d8336 [accessed 2026-05-07]
349. Hu K, Gong S, Zhang Q, Seng C, Xia M, Jiang S. An overview of implementing security and privacy in federated learning. Artif Intell Rev 2024;57(8):204. [doi: 10.1007/s10462-024-10846-8]
350. Xie M, Zhang Z, Hong H, Zhang G, Qin Y. Secure medical data sharing featuring traceable data usage and automatic audit mechanism. IEEE Internet Things J 2025;12(13):25587-25600. [doi: 10.1109/JIOT.2025.3559926]
351. Li C, Xing Z, Liu J, et al. Integrating zero-knowledge proofs into federated learning: a path to on-chain verifiable and privacy-preserving federated learning frameworks. Int J Web Inf Syst 2025 May 15;21(3):275-297. [doi: 10.1108/IJWIS-01-2025-0024]
352. Han B, Li B, Jurdak R, et al. PBFL: a privacy-preserving blockchain-based federated learning framework with homomorphic encryption and single masking. IEEE Internet Things J 2025;12(10):14229-14243. [doi: 10.1109/JIOT.2024.3524632]
353. Lavin R, Liu X, Mohanty H, Norman L, Zaarour G, Krishnamachari B. A survey on the applications of zero-knowledge proofs. arXiv. Preprint posted online on Aug 1, 2024 URL: https://arxiv.org/abs/2408.00243 [accessed 2026-05-07]
354. Zhou Z, Li Y, Wang Y, et al. ZHE: efficient zero-knowledge proofs for HE evaluations. Presented at: 2025 IEEE Symposium on Security and Privacy (SP); May 12-15, 2025. [doi: 10.1109/SP61157.2025.00199]
355. Wang X, Chen T, Dai HN, et al. A privacy-enhanced method for privacy-preserving and verifiable federated learning. IEEE Internet Things J 2025;12(14):26768-26781. [doi: 10.1109/JIOT.2025.3561262]
356. Munjal K, Bhatia R. A systematic review of homomorphic encryption and its contributions in healthcare industry. Complex Intell Syst 2023 Aug;9(4):3759-3786. [doi: 10.1007/s40747-022-00756-z]
357. Kalapaaking AP, Stephanie V, Khalil I, Atiquzzaman M, Yi X, Almashor M. SMPC-based federated learning for 6G-enabled Internet of Medical Things. IEEE Netw 2022;36(4):182-189. [doi: 10.1109/MNET.007.2100717]
358. Liu F, Zheng Z, Shi Y, Tong Y, Zhang Y. A survey on federated learning: a perspective from multi-party computation. Front Comput Sci 2024 Feb;18(1):181336. [doi: 10.1007/s11704-023-3282-7]
359. Naresh VS, Venkata Raju A, Srinivasa Rao O. Secure multiparty computation for privacy‐preserving machine learning in healthcare: a comprehensive survey. WIREs Computational Stats 2025 Sep;17(3):e70046 [FREE Full text] [doi: 10.1002/wics.70046]
360. Hosseini SM, Sikaroudi M, Babaei M, Tizhoosh HR. Cluster based secure multi-party computation in federated learning for histopathology images. In: Albarqouni S, editor. Distributed, Collaborative, and Federated Learning, and Affordable AI and Healthcare for Resource Diverse Global Health DeCaF FAIR 2022 2022 Lecture Notes in Computer Science: Springer; 2022:110-118. [doi: 10.1007/978-3-031-18523-6_11]
361. Kalapaaking AP, Khalil I, Rahman MS, Atiquzzaman M, Yi X, Almashor M. Blockchain-based federated learning with secure aggregation in trusted execution environment for Internet-of-Things. IEEE Trans Ind Inf 2023;19(2):1703-1714. [doi: 10.1109/TII.2022.3170348]

XSL•FO
RenderX

362. Zhu L, Hu S, Zhu X, Meng C, Huang M. Enhancing the security and privacy in the IoT supply chain using blockchain and federated learning with trusted execution environment. Mathematics 2023;11(17):3759. [doi: 10.3390/math11173759]
363. Jiang J, Soriente C, Karame G. On the challenges of detecting side-channel attacks in SGX. Presented at: 25th International Symposium on Research in Attacks, Intrusions and Defenses; Oct 26-28, 2022. [doi: 10.1145/3545948.3545972]
364. Guo J, Vaswani K, Paverd A, Pietzuch P. VerifiableFL: verifiable claims for federated learning using exclaves. arXiv. Preprint posted online on Dec 13, 2025 URL: https://arxiv.org/abs/2412.10537 [accessed 2026-05-07]
365. Zhang L, Fang G, Tan Z. FedCCW: a privacy-preserving Byzantine-robust federated learning with local differential privacy for healthcare. Cluster Comput 2025 Jun;28(3):182. [doi: 10.1007/s10586-024-04894-6]
366. Tayyeh HK, AL-Jumaili ASA. Balancing privacy and performance: a differential privacy approach in federated learning. Computers 2024;13(11):277. [doi: 10.3390/computers13110277]
367. Shukla S, Rajkumar S, Sinha A, Esha M, Elango K, Sampath V. Federated learning with differential privacy for breast cancer diagnosis enabling secure data sharing and model integrity. Sci Rep 2025 Apr 16;15(1):13061. [doi: 10.1038/s41598-025-95858-2] [Medline: 40240790]
368. Cao X, Fang M, Liu J, Gong NZ. FLTrust: Byzantine-robust federated learning via trust bootstrapping. Presented at: NDSS Symposium 2021 Program; Feb 21-25, 2021. [doi: 10.14722/ndss.2021.24434]
369. Yan G, Wang H, Yuan X, Li J. DeFL: defending against model poisoning attacks in federated learning via critical learning periods awareness. AAAI 2023;37(9):10711-10719. [doi: 10.1609/aaai.v37i9.26271]
370. Bagdasaryan E, Veit A, Hua Y, Estrin D, Shmatikov V. How to backdoor federated learning. In: Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics: PMLR; 2020:2938-2948 URL: https://proceedings.mlr.press/v108/bagdasaryan20a.html [accessed 2026-05-07]
371. Xie C, Chen M, Chen PY, Li B. CRFL: certifiably robust federated learning against backdoor attacks. In: Proceedings of the 38th International Conference on Machine Learning: PMLR; 2021:11372-11382 URL: https://proceedings.mlr.press/v139/xie21a.html [accessed 2026-05-07]
372. Geiping J, Bauermeister H, Droge H, Moeller M. Inverting gradients–how easy is it to break privacy in federated learning? In: Advances in Neural Information Processing Systems 33 (NeurIPS 2020): NeurIPS; 2020:16937-16947 URL: https://proceedings.neurips.cc/paper/2020/hash/c4ede56bbd98819ae6112b20ac6bf145-Abstract.html [accessed 2026-05-07]
373. Hatamizadeh A, Yin H, Roth H, et al. GradViT: gradient inversion of vision transformers. Presented at: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 18-24, 2022. [doi: 10.1109/CVPR52688.2022.00978]
374. Geng J, Mou Y, Li Q, et al. Improved gradient inversion attacks and defenses in federated learning. IEEE Trans Big Data 2024;10(6):839-850. [doi: 10.1109/TBDATA.2023.3239116]
375. Nguyen T, Lai P, Tran K, Phan N, Thai MT. Active membership inference attack under local differential privacy in federated learning. In: Proceedings of The 26th International Conference on Artificial Intelligence and Statistics: PMLR; 2023:5714-5730 URL: https://proceedings.mlr.press/v206/nguyen23e.html [accessed 2026-05-07]
376. Bai L, Hu H, Ye Q, Li H, Wang L, Xu J. Membership inference attacks and defenses in federated learning: a survey. ACM Comput Surv 2025 Apr 30;57(4):1-35. [doi: 10.1145/3704633]
377. Commey D, Hounsinou SG, Crosby GV. Post-quantum secure blockchain-based federated learning framework for healthcare analytics. IEEE Netw Lett 2025;7(2):126-129. [doi: 10.1109/LNET.2025.3563434]
378. Li S, Ngai ECH, Voigt T. An experimental study of Byzantine-robust aggregation schemes in federated learning. IEEE Trans Big Data 2024;10(6):975-988. [doi: 10.1109/TBDATA.2023.3237397]
379. Bonawitz K, Ivanov V, Kreuter B, et al. Practical secure aggregation for privacy-preserving machine learning. Presented at: 2017 ACM SIGSAC Conference on Computer and Communications Security; Oct 30 to Nov 3, 2017. [doi: 10.1145/3133956.3133982]
380. McMahan HB, Ramage D, Talwar K, Zhang L. Learning differentially private recurrent language models. arXiv. Preprint posted online on Oct 18, 2017 URL: https://arxiv.org/abs/1710.06963 [accessed 2026-05-07]
381. Geyer RC, Klein T, Nabi M. Differentially private federated learning: a client level perspective. arXiv. Preprint posted online on Dec 20, 2017 URL: https://arxiv.org/abs/1712.07557 [accessed 2026-05-07]
382. Rangwala M, Venugopal KR, Buyya R. Blockchain-enabled federated learning. arXiv. Preprint posted online on Aug 8, 2025 URL: https://arxiv.org/abs/2508.06406 [accessed 2026-07-07]
383. Fang F, Feng L, Xie J, et al. BCFL: a trustworthy and efficient federated learning framework based on blockchain in iot. Presented at: 2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD); May 8-10, 2024. [doi: 10.1109/CSCWD61410.2024.10580415]
384. Teo ZL, Jin L, Li S, et al. Federated machine learning in healthcare: a systematic review on clinical applications and technical architecture. Cell Rep Med 2024 Feb 20;5(2):101419. [doi: 10.1016/j.xcrm.2024.101419] [Medline: 38340728]
385. Li M, Xu P, Hu J, Tang Z, Yang G. From challenges and pitfalls to recommendations and opportunities: implementing federated learning in healthcare. Med Image Anal 2025 Apr;101:103497. [doi: 10.1016/j.media.2025.103497] [Medline: 39961211]
386. Salim MM, Yang LT, Park JH. Privacy-preserving and scalable federated blockchain scheme for healthcare 4.0. Computer Networks 2024 Jun;247:110472. [doi: 10.1016/j.comnet.2024.110472]

XSL•FO
RenderX

387. Moore SK. Chips to compute with encrypted data are coming: fully homomorphic encryption could make data unhackable. IEEE Spectr 2024;61(1):38-40. [doi: 10.1109/MSPEC.2024.10380468]
388. Karimireddy SP, He L, Jaggi M. Learning from history for Byzantine robust optimization. In: Proceedings of the 38th International Conference on Machine Learning: PMLR; 2021:5311-5319 URL: https://proceedings.mlr.press/v139/karimireddy21a.html [accessed 2026-05-07]
389. Komlo C, Goldberg I. FROST: flexible round-optimized Schnorr threshold signatures. In: Dunkelman O, Jacobson MJ, O'Flynn C, editors. Selected Areas in Cryptography SAC 2020 Lecture Notes in Computer Science: Springer; 2021:34-65. [doi: 10.1007/978-3-030-81652-0_2]
390. Pati S, Baid U, Edwards B, et al. The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research. Phys Med Biol 2022 Oct 12;67(20):204002. [doi: 10.1088/1361-6560/ac9449] [Medline: 36137534]
391. The HIPAA privacy rule. US Department of Health and Human Services. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/index.html [accessed 2026-05-07]
392. European Parliament and Council of the European Union. General Data Protection Regulation (GDPR), Regulation (EU) 2016/679. EUR-Lex. 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng [accessed 2026-05-07]
393. Altameem A, Kovtun V, Al-Ma'aitah M, Altameem T, H F, Youssef AE. Patient's data privacy protection in medical healthcare transmission services using back propagation learning. Computers and Electrical Engineering 2022 Sep;102:108087. [doi: 10.1016/j.compeleceng.2022.108087]
394. Artificial intelligence in software as a medical device. US Food and Drug Administration. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-software-medical-device [accessed 2026-05-07]
395. Sinaci AA, Gencturk M, Alvarez-Romero C, et al. Privacy-preserving federated machine learning on FAIR health data: a real-world application. Comput Struct Biotechnol J 2024 Dec;24:136-145. [doi: 10.1016/j.csbj.2024.02.014] [Medline: 38434250]
396. Tölle M, Garthe P, Scherer C, et al. Real world federated learning with a knowledge distilled transformer for cardiac CT imaging. NPJ Digit Med 2025 Feb 6;8(1):88. [doi: 10.1038/s41746-025-01434-3] [Medline: 39915633]
397. Shick AA, Webber CM, Kiarashi N, et al. Transparency of artificial intelligence/machine learning-enabled medical devices. NPJ Digit Med 2024 Jan 26;7(1):21. [doi: 10.1038/s41746-023-00992-8] [Medline: 38273098]
398. Leslie K, Moore J, Robertson C, et al. Regulating health professional scopes of practice: comparing institutional arrangements and approaches in the US, Canada, Australia and the UK. Hum Resour Health 2021 Jan 28;19(1):15. [doi: 10.1186/s12960-020-00550-3] [Medline: 33509209]
399. Tan E, Lerouge E, Du Caju J, Du Seuil D. Verification of education credentials on European blockchain services infrastructure (EBSI): action research in a cross-border use case between Belgium and Italy. BDCC 2023;7(2):79. [doi: 10.3390/bdcc7020079]
400. Orabi MM, Emam O, Fahmy H. Adapting security and decentralized knowledge enhancement in federated learning using blockchain technology: literature review. J Big Data 2025;12(1):55. [doi: 10.1186/s40537-025-01099-5]
401. Kalapaaking AP, Khalil I, Yi X, Lam KY, Huang GB, Wang N. Auditable and verifiable federated learning based on blockchain-enabled decentralization. IEEE Trans Neural Netw Learning Syst 2024;36(1):102-115. [doi: 10.1109/TNNLS.2024.3407670]
402. Kostick-Quenet KM, Compagnucci MC, Aboy M, Minssen T. Patient-centric federated learning: automating meaningful consent to health data sharing with smart contracts. J Law Biosci 2025;12(1):lsaf003. [doi: 10.1093/jlb/lsaf003] [Medline: 40321304]
403. Siniosoglou I, Argyriou V, Sarigiannidis P, et al. Post-processing fairness evaluation of federated models: an unsupervised approach in healthcare. IEEE/ACM Trans Comput Biol and Bioinf 2023;20(4):2518-2529. [doi: 10.1109/TCBB.2023.3269767]
404. McKay F, Williams BJ, Prestwich G, Bansal D, Treanor D, Hallowell N. Artificial intelligence and medical research databases: ethical review by data access committees. BMC Med Ethics 2023 Jul 8;24(1):49. [doi: 10.1186/s12910-023-00927-8] [Medline: 37422629]
405. Zafar A. Reconciling blockchain technology and data protection laws: regulatory challenges, technical solutions, and practical pathways. Journal of Cybersecurity 2025 Jan 17;11(1). [doi: 10.1093/cybsec/tyaf002]
406. Belen-Saglam R, Altuncu E, Lu Y, Li S. A systematic literature review of the tension between the GDPR and public blockchain systems. Blockchain: Research and Applications 2023 Jun;4(2):100129. [doi: 10.1016/j.bcra.2023.100129]
407. Balistri E, Casellato F, Giannelli C, Stefanelli C. BlockHealth: blockchain-based secure and peer-to-peer health information sharing with data protection and right to be forgotten. ICT Express 2021 Sep;7(3):308-315. [doi: 10.1016/j.icte.2021.08.006]
408. Makhdoom I, Zhou I, Abolhasan M, Lipman J, Ni W. PrivySharing: a blockchain-based framework for privacy-preserving and secure data sharing in smart cities. Computers & Security 2020 Jan;88:101653. [doi: 10.1016/j.cose.2019.101653]
409. Zhang F, Shuai Z, Kuang K, Wu F, Zhuang Y, Xiao J. Unified fair federated learning for digital healthcare. Patterns (N Y) 2024 Jan 12;5(1):100907. [doi: 10.1016/j.patter.2023.100907] [Medline: 38264718]
410. Poulain R, Tarek MFB, Beheshti R. Improving fairness in AI models on electronic health records: the case for federated learning methods. arXiv. Preprint posted online on May 19, 2023 URL: https://arxiv.org/abs/2305.11386 [accessed 2026-05-07]

XSL•FO
RenderX

411. Liu M, Ning Y, Teixayavong S, et al. A scoping review and evidence gap analysis of clinical AI fairness. NPJ Digit Med 2025 Jun 14;8(1):360. [doi: 10.1038/s41746-025-01667-2] [Medline: 40517148]
412. Li S, Wu Q, Zhou D, et al. FairFML: fair federated machine learning with a case study on reducing gender disparities in cardiac arrest outcome prediction. NPJ Health Syst 2025;2(1):29. [doi: 10.1038/s44401-025-00035-2]
413. Xing H, Sun R, Ren J, et al. Achieving flexible fairness metrics in federated medical imaging. Nat Commun 2025 Apr 8;16(1):3342. [doi: 10.1038/s41467-025-58549-0] [Medline: 40199877]
414. Friesen P, Douglas‐Jones R, Marks M, et al. Governing AI‐driven health research: are IRBs up to the task? Ethics & Human Research 2021 Mar;43(2):35-42. [doi: 10.1002/eahr.500085]
415. Capili B, Anastasi JK. Ethical research and the institutional review board: an introduction. Am J Nurs 2024 Mar 1;124(3):50-54. [doi: 10.1097/01.NAJ.0001008420.28033.e8] [Medline: 38386835]
416. Bouderhem R. Shaping the future of AI in healthcare through ethics and governance. Humanit Soc Sci Commun 2024;11(1). [doi: 10.1057/s41599-024-02894-w]
417. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240029200 [accessed 2026-07-05]
418. Secretary's Advisory Committee on Human Research Protections. IRB considerations on the use of artificial intelligence in human subjects research. US Department of Health and Human Services. 2022. URL: https://www.hhs.gov/ohrp/sachrp-committee/recommendations/irb-considerations-use-artificial-intelligence-human-subjects-research/index.html [accessed 2026-05-07]
419. Kiseleva A, Kotzinos D, De Hert P. Transparency of AI in healthcare as a multilayered system of accountabilities: between legal requirements and technical limitations. Front Artif Intell 2022;5:879603. [doi: 10.3389/frai.2022.879603] [Medline: 35707765]
420. Holzinger A, Saranti A, Molnar C, Biecek P, Samek W. Explainable AI methods - a brief overview. In: Holzinger A, Goebel R, Fong R, Moon T, Müller KR, Samek W, editors. xxAI - Beyond Explainable AI xxAI 2020 Lecture Notes in Computer Science: Springer; 2022:13-38. [doi: 10.1007/978-3-031-04083-2_2]
421. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1):310. [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
422. Sadeghi Z, Alizadehsani R, Cifci MA, et al. A review of explainable artificial intelligence in healthcare. Computers and Electrical Engineering 2024 Aug;118:109370. [doi: 10.1016/j.compeleceng.2024.109370]
423. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell 2019;1(9):389-399. [doi: 10.1038/s42256-019-0088-2]
424. Rauter CM, Wöhlke S, Schicktanz S. My data, my choice? - German patient organizations' attitudes towards big data-driven approaches in personalized medicine. an empirical-ethical study. J Med Syst 2021 Feb 22;45(4):43. [doi: 10.1007/s10916-020-01702-7] [Medline: 33616768]
425. Shen N, Kassam I, Zhao H, et al. Foundations for meaningful consent in Canada's digital health ecosystem: retrospective study. JMIR Med Inform 2022 Mar 31;10(3):e30986. [doi: 10.2196/30986] [Medline: 35357318]
426. Li N, Zhou C, Gao Y, et al. Machine unlearning: taxonomy, metrics, applications, challenges, and prospects. IEEE Trans Neural Netw Learning Syst 2025;36(8):13709-13729. [doi: 10.1109/TNNLS.2025.3530988]
427. Liu Z, Jiang Y, Shen J, et al. A survey on federated unlearning: challenges, methods, and future directions. ACM Comput Surv 2025 Jan 31;57(1):1-38. [doi: 10.1145/3679014]
428. Bhardwaj T, Sumangali K. An explainable federated blockchain framework with privacy-preserving AI optimization for securing healthcare data. Sci Rep 2025 Jul 1;15(1):21799. [doi: 10.1038/s41598-025-04083-4] [Medline: 40595873]
429. Blackman J, Veerapen R. On the practical, ethical, and legal necessity of clinical artificial intelligence explainability: an examination of key arguments. BMC Med Inform Decis Mak 2025 Mar 5;25(1):111. [doi: 10.1186/s12911-025-02891-2] [Medline: 40045339]
430. Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. US Food and Drug Administration. 2021. URL: https://www.fda.gov/media/145022/download [accessed 2026-05-07]
431. Kauttonen J, Rousi R, Alamäki A. Trust and acceptance challenges in the adoption of AI applications in health care: quantitative survey analysis. J Med Internet Res 2025 Mar 21;27:e65567. [doi: 10.2196/65567] [Medline: 40116853]
432. Dayan I, Roth HR, Zhong A, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. Nat Med 2021 Oct;27(10):1735-1743. [doi: 10.1038/s41591-021-01506-3] [Medline: 34526699]
433. Kumar R, Khan AA, Kumar J, et al. Blockchain-federated-learning and deep learning models for COVID-19 detection using CT imaging. IEEE Sensors J 2021;21(14):16301-16314. [doi: 10.1109/JSEN.2021.3076767]
434. Pati S, Baid U, Edwards B, et al. Federated learning enables big data for rare cancer boundary detection. Nat Commun 2022 Dec 5;13(1):7346. [doi: 10.1038/s41467-022-33407-5] [Medline: 36470898]
435. Zenk M, Baid U, Pati S, et al. Towards fair decentralized benchmarking of healthcare AI algorithms with the federated tumor segmentation (FeTS) challenge. Nat Commun 2025 Jul 8;16(1):6274. [doi: 10.1038/s41467-025-60466-1] [Medline: 40628696]

XSL•FO
RenderX

436. Mazid A, Kirmani S, Abid M, Pawar V. A secure and efficient framework for internet of medical things through blockchain driven customized federated learning. Cluster Comput 2025 Aug;28(4):225. [doi: 10.1007/s10586-024-04896-4]
437. Moulahi W, Jdey I, Moulahi T, Alawida M, Alabdulatif A. A blockchain-based federated learning mechanism for privacy preservation of healthcare IoT data. Comput Biol Med 2023 Dec;167:107630. [doi: 10.1016/j.compbiomed.2023.107630] [Medline: 37952305]
438. Rahman MA, Hossain MS, Islam MS, Alrajeh NA, Muhammad G. Secure and provenance enhanced Internet of Health Things framework: a blockchain managed federated learning approach. IEEE Access 2020;8:205071-205087. [doi: 10.1109/ACCESS.2020.3037474] [Medline: 34192116]
439. Ahmed AA, Alabi OO. Secure and scalable blockchain-based federated learning for cryptocurrency fraud detection: a systematic review. IEEE Access 2024;12:102219-102241. [doi: 10.1109/ACCESS.2024.3429205]
440. Petrosino L, Masi L, D'Antoni F, Merone M, Vollero L. A zero-knowledge proof federated learning on DLT for healthcare data. J Parallel Distrib Comput 2025 Feb;196:104992. [doi: 10.1016/j.jpdc.2024.104992]
441. Jiang Y, Ma B, Wang X, et al. Blockchained federated learning for Internet of Things: a comprehensive survey. ACM Comput Surv 2024 Oct 31;56(10):1-37. [doi: 10.1145/3659099]
442. Chang Y, Fang C, Sun W. A blockchain-based federated learning method for smart healthcare. Comput Intell Neurosci 2021;2021(1):4376418. [doi: 10.1155/2021/4376418] [Medline: 34868289]
443. Zhang H, Jiang S, Xuan S. Decentralized federated learning based on blockchain: concepts, framework, and challenges. Comput Commun 2024 Feb;216:140-150. [doi: 10.1016/j.comcom.2023.12.042]
444. Samantray BS, Reddy KHK. A federated learning approach towards hybrid blockchain, quantum-key-encryption based distributed system: a futuristic healthcare architecture for smart cities. Blockchain: Research and Applications 2025 Sep;100385:100385. [doi: 10.1016/j.bcra.2025.100385]
445. Bhasker B, Rao PM, Saraswathi P, et al. Blockchain framework with IoT device using federated learning for sustainable healthcare systems. Sci Rep 2025 Jul 23;15(1):26736. [doi: 10.1038/s41598-025-06539-z] [Medline: 40701986]
446. Ali AA, Gunavathie MA, Srinivasan V, Aruna M, Chennappan R, Matheena M. Securing electronic health records using blockchain-enabled federated learning for IoT-based smart healthcare. Clinical eHealth 2025 Dec;8:125-133. [doi: 10.1016/j.ceh.2025.04.002]
447. Das P, Kumar N, Jain C, Singh M. Intelligent IoT-enabled healthcare solutions implementing federated meta-learning with blockchain. J Ind Inf Integr 2025 May;45:100797. [doi: 10.1016/j.jii.2025.100797]
448. Kumar R, Bernard CM, Ullah A, et al. Privacy-preserving blockchain-based federated learning for brain tumor segmentation. Comput Biol Med 2024 Jul;177:108646. [doi: 10.1016/j.compbiomed.2024.108646] [Medline: 38824788]
449. Liang X, Zhao J, Chen Y, Bandara E, Shetty S. Architectural design of a blockchain-enabled, federated learning platform for algorithmic fairness in predictive health care: design science study. J Med Internet Res 2023 Oct 30;25:e46547. [doi: 10.2196/46547] [Medline: 37902833]
450. Om Kumar CU, Gajendran S, Balaji V, Nhaveen A, Sai Balakrishnan S. RETRACTED ARTICLE: Securing health care data through blockchain enabled collaborative machine learning. Soft comput 2023;27(14):9941-9954. [doi: 10.1007/s00500-023-08330-6] [Medline: 37287568]
451. Ali A, Al-Rimy BAS, Tin TT, Altamimi SN, Qasem SN, Saeed F. Empowering precision medicine: unlocking revolutionary insights through blockchain-enabled federated learning and electronic medical records. Sensors (Basel) 2023 Aug 28;23(17):7476. [doi: 10.3390/s23177476] [Medline: 37687931]
452. Lian Z, Wang W, Han Z, Su C. Blockchain-based personalized federated learning for Internet of Medical Things. IEEE Trans Sustain Comput 2023;8(4):694-702. [doi: 10.1109/TSUSC.2023.3279111]
453. Farooq K, Syed HJ, Alqahtani SO, Nagmeldin W, Ibrahim AO, Gani A. Blockchain federated learning for in-home health monitoring. Electronics (Basel) 2022;12(1):136. [doi: 10.3390/electronics12010136]
454. Zhang H, Li G, Zhang Y, Gai K, Qiu M. Blockchain-based privacy-preserving medical data sharing scheme using federated learning. In: Qiu H, Zhang C, Fei Z, Qiu M, Kung SY, editors. Knowledge Science, Engineering and Management KSEM 2021 Lecture Notes in Computer Science: Springer; 2021:634-646. [doi: 10.1007/978-3-030-82153-1_52]
455. Lo SK, Liu Y, Lu Q, et al. Toward trustworthy AI: blockchain-based architecture design for accountability and fairness of federated learning systems. IEEE Internet Things J 2022;10(4):3276-3284. [doi: 10.1109/JIOT.2022.3144450]
456. Singh S, Rathore S, Alfarraj O, Tolba A, Yoon B. A framework for privacy-preservation of IoT healthcare data using federated learning and blockchain technology. Future Generation Computer Systems 2022 Apr;129:380-388. [doi: 10.1016/j.future.2021.11.028]
457. Nguyen DC, Ding M, Pham QV, et al. Federated learning meets blockchain in edge computing: opportunities and challenges. IEEE Internet Things J 2021;8(16):12806-12825. [doi: 10.1109/JIOT.2021.3072611]
458. Liu Y, Yu W, Ai Z, Xu G, Zhao L, Tian Z. A blockchain-empowered federated learning in healthcare-based cyber physical systems. IEEE Trans Netw Sci Eng 2022;10(5):2685-2696. [doi: 10.1109/TNSE.2022.3168025]
459. Otoum S, Al Ridhawi I, Mouftah HT. Preventing and controlling epidemics through blockchain-assisted AI-enabled networks. IEEE Netw 2021;35(3):34-41. [doi: 10.1109/MNET.011.2000628]

XSL•FO
RenderX

460. Durga R, Poovammal E. Federated learning model for healthchain system. Presented at: 2021 6th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE); Dec 1-3, 2021. [doi: 10.1109/ICRAIE52900.2021.9703948]
461. Lakhan A, Mohammed MA, Nedoma J, et al. Federated-learning based privacy preservation and fraud-enabled blockchain IoMT system for healthcare. IEEE J Biomed Health Inform 2022;27(2):664-672. [doi: 10.1109/JBHI.2022.3165945]
462. Samuel O, Omojo AB, Onuja AM, et al. IoMT: a COVID-19 healthcare system driven by federated learning and blockchain. IEEE J Biomed Health Inform 2022;27(2):823-834. [doi: 10.1109/JBHI.2022.3143576]
463. Yang X, Xing C. Federated medical learning framework based on blockchain and homomorphic encryption. Wireless Communications and Mobile Computing 2024 Jan 5;2024:1-15. [doi: 10.1155/2024/8138644]
464. Nguyen DC, Ding M, Pathirana PN, Seneviratne A. Blockchain and AI-based solutions to combat coronavirus (COVID-19)-like epidemics: a survey. IEEE Access 2021;9:95730-95753. [doi: 10.1109/ACCESS.2021.3093633] [Medline: 34812398]
465. Hemdan EED, Sayed A. Smart and secure healthcare with digital twins: a deep dive into blockchain, federated learning, and future innovations. Algorithms 2025;18(7):401. [doi: 10.3390/a18070401]
466. Baseri Y, Hafid A, Shahsavari Y, Makrakis D, Khodaiemehr H. Blockchain security risk assessment in quantum era, migration strategies, and proactive defense. IEEE Commun Surv Tutorials 2025;28:2925-2964. [doi: 10.1109/COMST.2025.3621113]
467. Cheon JH, Kim A, Kim M, Song Y. Homomorphic encryption for arithmetic of approximate numbers. In: Takagi T, Peyrin T, editors. Advances in Cryptology – ASIACRYPT 2017 ASIACRYPT 2017 Lecture Notes in Computer Science: Springer; 2017:409-437. [doi: 10.1007/978-3-319-70694-8_15]
468. Doga H, Bose A, Sahin ME, et al. How can quantum computing be applied in clinical trial design and optimization? Trends Pharmacol Sci 2024 Oct;45(10):880-891. [doi: 10.1016/j.tips.2024.08.005] [Medline: 39317621]
469. Thibault LT, Sarry T, Hafid AS. Blockchain scaling using rollups: a comprehensive survey. IEEE Access 2022;10:93039-93054. [doi: 10.1109/ACCESS.2022.3200051]
470. Hafid A, Hafid AS, Samih M. Scaling blockchains: a comprehensive survey. IEEE Access 2020;8:125244-125262. [doi: 10.1109/ACCESS.2020.3007251]
471. You J, Yang R, Zhan Y, Song B, Zhang Y, Wang Z. BR-MTFL: a novel Byzantine resilience-enhanced multitask federated learning framework for high-speed train fault diagnosis. IEEE Trans Instrum Meas 2025;74:1-13. [doi: 10.1109/TIM.2025.3550635]
472. Yang N, Tang C, Deng Z, He D. A Gaussian reputation-based hybrid BFT consensus with a formal security framework. IEEE Trans Dependable and Secure Comput 2025;22(5):5397-5414. [doi: 10.1109/TDSC.2025.3567356]
473. Wang Y, Peng H, Su Z, Luan TH, Benslimane A, Wu Y. A platform-free proof of federated learning consensus mechanism for sustainable blockchains. IEEE J Select Areas Commun 2022;40(12):3305-3324. [doi: 10.1109/JSAC.2022.3213347]
474. Ajmal CS, Yerram S, Abishek V, et al. Innovative approaches in regulatory affairs: leveraging artificial intelligence and machine learning for efficient compliance and decision-making. AAPS J 2025 Jan 7;27(1):22. [doi: 10.1208/s12248-024-01006-5] [Medline: 39776314]
475. Dubey P, Kumar M. Integrating explainable AI with federated learning for next-generation IoT: a comprehensive review and prospective insights. Computer Science Review 2025 May;56:100697. [doi: 10.1016/j.cosrev.2024.100697]
476. Smith V, Chiang CK, Sanjabi M, Talwalkar A. Federated multi-task learning. Presented at: 31st International Conference on Neural Information Processing Systems; Dec 4-9, 2017. [doi: 10.5555/3294996.3295196]
477. Li K, Xiao C. CBFL: a communication-efficient federated learning framework from data redundancy perspective. IEEE Systems Journal ;16(4):5572-5583. [doi: 10.1109/JSYST.2021.3119152]

Abbreviations

AI: artificial intelligence
BCFL: blockchain-based federated learning
BFT: Byzantine fault tolerance
CT: computed tomography
DAG: direct acyclic graph tangle
DApps: decentralized applications
DICOM: Digital Imaging and Communications in Medicine
DP: differential privacy
DPoS: delegated proof of stake
ECG: electrocardiography
EEG: electroencephalography
EHR: electronic health record
FDA: Food and Drug Administration
FL: federated learning

XSL•FO
RenderX

FTL: federated transfer learning
GDPR: General Data Protection Regulation
HE: homomorphic encryption
HFL: horizontal federated learning
HIPAA: Health Insurance Portability and Accountability Act
IBC: Inter-Blockchain Communication
IoMT: Internet of Medical Things
IoT: Internet of Things
MAR: medication administration record
ML: machine learning
MRI: magnetic resonance imaging
NLP: natural language processing
P2P: peer-to-peer
pBFT: practical Byzantine fault tolerance
PGD: patient-generated data
PoA: proof of authority
PoS: proof of stake
PoW: proof of work
PRO: patient-reported outcomes
RL: reinforcement learning
SHAP: Shapley Additive Explanations
SMPC: secure multiparty computation
VFL: vertical federated learning
ZKP: zero-knowledge proof

Edited by A Coristine; submitted 05.Jul.2025; peer-reviewed by N Acharya, O Oluwole, O Oloruntoba; revised version received 18.Jan.2026; accepted 20.Jan.2026; published 15.Jun.2026.

Please cite as:
Shahsavari Y, Baseri Y, Hafid A, Dambri OA, Makrakis D
Integration of Federated Learning and Blockchain in Health Care: Tutorial on Medical Data, Architectures, Privacy, Security, and Regulatory Compliance
J Med Internet Res 2026;28:e80178
URL: https://www.jmir.org/2026/1/e80178
doi:10.2196/80178

© Yahya Shahsavari, Yaser Baseri, Abdelhakim Hafid, Oussama Abderrahmane Dambri, Dimitrios Makrakis. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Building EVA (Educación, Vinculación, y Autoayuda): Tutorial for the Development of a Digital Mental Health Chatbot for Adolescents Living With HIV

Diego Humberto Vasquez[1*], BA; Neil Rupani[2*], BA; Carmen Contreras[1,3], MA; Lenka Kolevic[4,5], MD; Molly Forrest Franke[6], SCD; Jerome Timothy Galea[3], MSW, PhD

[1]Socios En Salud Sucursal Peru, Lima, Peru

[2]Morsani College of Medicine, University of South Florida, Tampa, FL, United States

[3]School of Social Work, College of Behavorial and Community Sciences, University of South Florida, 13301 Bruce B Downs Blvd, MHC 1400, Tampa, FL, United States

[4]Instituto Nacional de Salud del Niño, Lima, Peru

[5]Departamento Académico de Pediatría, Universidad Nacional Mayor de San Marcos, Lima, Peru

[6]Department of Global Health and Social Medicine, Harvard Medical School, Harvard University, Boston, MA, United States

[*]these authors contributed equally

Corresponding Author:
Jerome Timothy Galea, MSW, PhD
School of Social Work, College of Behavorial and Community Sciences, University of South Florida, 13301 Bruce B Downs Blvd, MHC 1400, Tampa, FL, United States

Abstract

Adolescents living with HIV face higher rates of mental health morbidity compared to other age groups, particularly depressive symptoms, while access to specialized services remains limited in many low- and middle-income countries. Chatbots offer a promising, low-cost approach to delivering structured mental health support in resource-constrained settings; however, practical guidance on how to develop such tools with end-user involvement remains scarce. This tutorial provides a step-by-step guide to the human-centered design and development of a mental health chatbot for adolescents living with HIV, illustrated through the creation of EVA (*Educación, Vinculación, y Autoayuda*), a chatbot cocreated with adolescents living with HIV in Peru. Guided by human-centered design principles, the tutorial covers three key steps: (1) understanding user needs through qualitative research, (2) iterative chatbot development with a Youth Advisory Board across five structured sessions, and (3) external review and testing before launch. Throughout the process, adolescents contributed to defining the chatbot's tone, visual identity, navigation structure, and content priorities. The chatbot was built using a low-cost messaging platform and incorporated multimedia components, including brief animated videos, to enhance engagement. Throughout development, iterative testing with the Youth Advisory Board and external health care professionals informed refinements in usability, content clarity, emotional tone, and accessibility. This tutorial synthesizes key methodological decisions, lessons learned, and challenges encountered, providing practical guidance for researchers and practitioners seeking to develop similar adolescent-centered digital mental health tools in low- and middle-income countries. Key takeaways include the importance of early and sustained adolescent involvement, the value of iterative prototyping, and the feasibility of building functional chatbots with limited resources to help address the health service gap experienced by adolescents living with HIV.

International Registered Report Identifier (IRRID): RR2-10.2196/55559

(J Med Internet Res 2026;28:e75351) doi:10.2196/75351

KEYWORDS

chatbot; mHealth; adolescent; HIV; mental health; depression; mobile health

Introduction

In 2024, approximately 1.57 million adolescents aged 0-19 years were living with HIV worldwide, accounting for about 11% of new HIV infections [1]. Adolescents living with HIV face unique challenges that place them at greater risk of mental health problems compared to their peers without HIV, as well as to children and adults living with and without HIV [2,3]. The coexistence of a chronic, potentially life-altering illness with the pressures of transitioning from pediatric to adult care settings can heighten psychological distress, undermine adherence to antiretroviral therapy (ART), and compromise viral suppression and overall health [4-6]. These vulnerabilities are reflected in the high prevalence of depression among adolescents living

XSL•FO
RenderX

with HIV, estimated at nearly 26%, which is markedly higher than the rate observed in the general adolescent population, where fewer than 10.5% experience depression [7,8].

Left unaddressed, mental health problems—especially depression and anxiety—among adolescents living with HIV are associated with suboptimal adherence to ART, leading to negative treatment outcomes [9]. At the same time, however, there is a global shortage of mental health services, with an estimated 70% of people in need worldwide lacking access [10]. For adolescents living with HIV, who predominantly live in low- and middle-income countries (LMICs), this gap results in a severe lack of mental health services, which can, in turn, impact HIV treatment outcomes and overall health.

Multiple factors at the personal, health care system, and policy levels undermine adolescents living with HIV's access to mental health services. Stigma around HIV and mental health can discourage adolescents from disclosing symptoms or seeking care [11,12], while dependence on caregivers adds logistical and privacy concerns [13]. Mental health services are typically not integrated with HIV care, and facilities may lack trained personnel, resulting in delayed or suboptimal care [14]. In many LMICs, including Peru, a shortage of mental health care professionals, limited financial resources, and insufficiently integrated policies further constrain access to adolescent-focused interventions [15,16].

Against this backdrop, depression emerges as one of the most common and pressing mental health concerns among adolescents living with HIV [17]. Although many such adolescents experience mild to moderate symptoms rather than severe depression, even low-intensity symptoms can escalate into more serious issues if left unaddressed [6,18]. Indeed, preliminary data in Peru indicate that up to 92% of adolescents living with HIV presenting with depression have symptoms in the mild-to-moderate range [19]. These youth receive few—if any—routine mental health interventions, such as education, self-help strategies, or linkage to community-based resources, missing a critical opportunity to prevent progression to severe depressive episodes and further ART nonadherence [20].

Given these converging challenges, digital solutions have gained traction to fill the "presevere depression service gap" [21]. Chatbots—automated conversational agents—show promise for delivering accessible, anonymous, and scalable mental health support, requiring no special software and operating via widely used platforms such as SMS text messaging, WhatsApp (WhatsApp Inc), and Facebook Messenger (Meta Platforms, Inc) [21,22]. In Peru, chatbots facilitated mass depression screening, reaching more than 20,000 adults during the COVID-19 pandemic [23,24], suggesting the potential for a chatbot tailored to the needs of adolescents living with HIV that offers psychoeducation, coping strategies, and linkage to care.

Developing such a chatbot, however, requires active end-user participation to maximize relevance and acceptability. Human-centered design (HCD)—also known as participatory or user-centered design—integrates end-user voices at every stage, including content creation, prototype testing, and refinement [25]. Previous digital mental health interventions have experienced high attrition rates among young users when they fail to address factors such as cultural context, stigma, and interface usability [26]. By involving adolescents living with HIV from the outset, we can better tailor the chatbot's functions and content, improving engagement and mitigating the risk of dropout [27]. Yet, despite the growing popularity of chatbots in health care, few published studies describe a step-by-step approach to applying HCD in chatbot development.

This tutorial describes the development of EVA (*Educación, Vinculación, y Autoayuda;* in English: Education, Linkage to Care, and Self-Help), a chatbot created during August-December 2023 as part of a broader mixed-methods study assessing the feasibility of chatbots for Peruvian adolescents living with HIV [20], to illustrate a practical, replicable approach to mental health chatbot development. The larger study comprised three phases (Figure 1): (1) qualitative data collection to understand how adolescents living with HIV conceptualized depression and their initial views and opinions of a mental health chatbot, (2) human-centered development of the chatbot (focus of this tutorial), and (3) feasibility and acceptability testing of the chatbot.

Figure 1. Phases of the principal study.

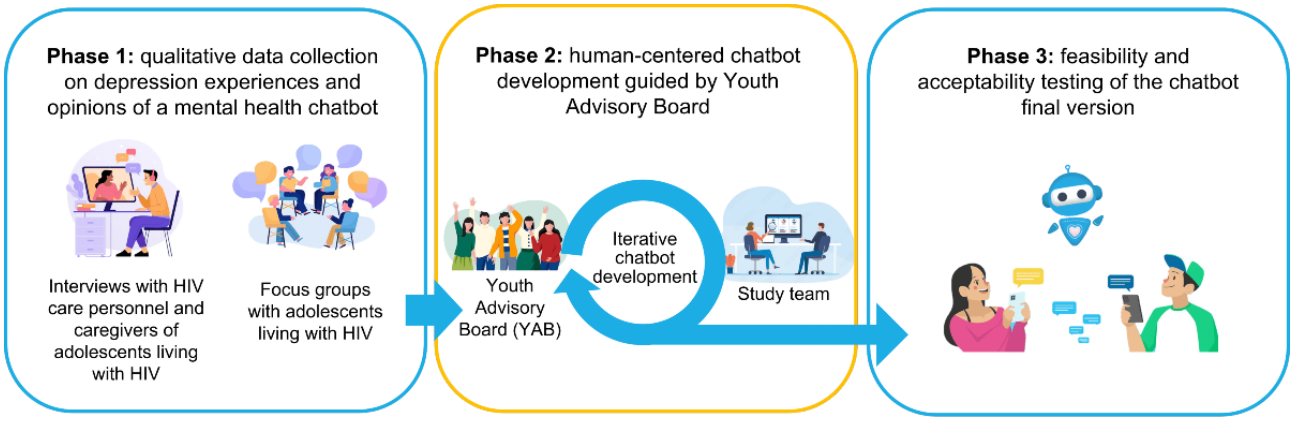

XSL•FO
RenderX

Ethical Considerations

The principal study, which includes the procedures described here, was reviewed and approved by an Institutional Bioethics Committee (Comité Institucional de Bioética) of VÍA LIBRE in Peru. The Institutional Review Board (IRB) approval number is CIB No 8591 (2023a).

The IRB at the University of South Florida formally agreed to rely on VÍA LIBRE for the review, approval, and continuing oversight of the research project under an interagency IRB Authorization Agreement (University of South Florida IRB study no 005124). For participants aged younger than 18 years, informed assent was obtained as well as informed consent from their legal guardians. For participants older than 18 years, informed consent was obtained.

Tutorial Objectives

In this tutorial, we outline our human-centered approach to creating a novel chatbot designed to deliver low-intensity, supportive interventions for adolescents living with HIV in Peru experiencing mild to moderate depressive symptoms, as defined by commonly used clinical instruments such as the Patient Health Questionnaire-9 (PHQ-9). Specifically, this tutorial aims to (1) describe how HCD can be meaningfully applied with adolescents throughout the development of a digital mental health tool; (2) explain how chatbot modules were structured to deliver mental health information and supportive content in a fluid, engaging, and well-organized manner; (3) illustrate how iterative review cycles with a Youth Advisory Board (YAB) can be conducted to refine content, tone, usability, and cultural relevance; and (4) describe how ethical considerations, multimedia elements, and clinical validation processes can be integrated into chatbot development.

Chatbot Development: Key Considerations

HCD Methodology

Chatbot development should follow HCD methodology. HCD focuses on designing conversational experiences centered on users' needs, expectations, and emotions. Its iterative approach enables the creation of more intuitive, efficient, and empathetic chatbots by adhering to 4 main principles: being people-centered, solving the right problem, considering everything as a system, and making simple and small interventions [28].

Central to this approach is co-design, which positions end users not merely as participants but as active contributors to decision-making, ensuring that the intervention reflects their realities, language, and priorities. To operationalize this participatory framework, a YAB composed of 6 Peruvian adolescents living with HIV was established to guide chatbot development.

Ethical Safeguards of the Design

Designing a mental health chatbot for adolescents—particularly adolescents living with HIV—requires integrating ethical, privacy, and clinical safeguards from the earliest stages of development. These considerations are not secondary features but core design components that directly influence trust, safety, and acceptability.

- Privacy and anonymity: the EVA chatbot was designed to minimize personal data collection, following recommendations from adolescents living with HIV consulted before the design phase, which were later reinforced and validated by the YAB. Users were only asked to enter a first name or nickname, and no information regarding HIV status, treatment history, or other personal identifiers was requested at any point. This approach reduces barriers to engagement and is especially important in contexts where stigma related to HIV or mental health remains high. User data were stored securely on Smartbot360 servers under the chosen display name, in a system that is the Health Insurance Portability and Accountability Act–compliant and suitable for sensitive health data [29].
- Transparency about the chatbot's nature: upon initiation, the chatbot clarified that it was a virtual tool and not a human being. Explicitly communicating the system's nonhuman nature is an important ethical practice, helping prevent misunderstanding about the type of support provided.
- Boundaries of support and linkage to care: because chatbots are not substitutes for professional care, EVA incorporated a linkage module. After addressing topics such as anxiety, depression, and eating-related concerns, users were asked whether they wished to connect with a health care professional. Communication through the linkage module occurred entirely in written form within the chatbot interface, without requiring personal identifiers. All conversations between users and health care professionals were securely recorded on Smartbot360 servers and were accessible only to the study team.
- Clinical accuracy and destigmatizing language: the chatbot's educational content was adapted from reliable international sources, including the World Health Organization (WHO) and the Joint United Nations Program on HIV/AIDS (UNAIDS), and reviewed by the YAB, research team, as well as external health care professionals with experience in adolescent HIV care, ensuring accuracy and the use of appropriate, destigmatizing language.
- Informed consent considerations: when developing similar tools, teams should decide early whether consent will be handled digitally, in person, or through a hybrid model. For EVA, digital informed consent was not incorporated into the chatbot structure during development, as a printed informed consent document was planned for phase 3 of the principal study. During that testing phase, participants will also receive detailed instructions and guidance from a mental health care professional, who will supervise the chatbot's use and ensure that all ethical considerations are addressed.

XSL•FO
RenderX

Study Team and Software Tools

The development team should include a mix of health care professionals, researchers, and technical staff with complementary expertise. For EVA, the team included 4 health care professionals: 2 psychologists (CC and DHV), a mental health researcher and clinician with experience in chatbot development (JTG), and a medical student (NR). DHV and NR were primarily responsible for software development, while CC and JTG oversaw the content and construction of the chatbot.

The chatbot was developed using Smartbot360 software [30]. JTG used this software for chatbot development in previous studies aimed at developing tools to improve mental health for other populations [31,32]. In June 2023, JTG conducted virtual and face-to-face training sessions on the use of Smartbot360.

Building EVA, Step-by-Step

Step 1: Understanding User Needs

The first step should involve collecting information on the perceptions, acceptability, and feasibility of a mental health chatbot among the target adolescent population and the adults involved in adolescents' care, to gather preliminary insights. For EVA, we conducted 5 focus groups with 28 adolescents living with HIV, complemented by interviews with health care professionals providing pediatric HIV services, and caregivers of adolescents living with HIV (for details of this step, refer to [33]). Some examples of questions from this phase include "What opinions do you have about the idea of a mental health chatbot?" and "What would be the barriers or difficulties to accessing a chatbot (access to a device, internet, etc)?" These questions were derived from Sekhon's theoretical framework of health care intervention acceptability [34].

The preliminary findings on chatbot acceptability provided valuable insights into users' expectations and perceived challenges before development. Participants emphasized the importance of a secure, confidential, and easily accessible tool, which they believed could serve as an educational resource to strengthen self-help skills, given the long wait times for professional care in public health services, which can range from weeks to months. Therefore, the chatbot's mobile-friendly design was seen as advantageous, given that adolescents naturally engage with technology [35].

At the same time, potential drawbacks or challenges of a chatbot can be identified early. Some adolescents felt that a chatbot could be perceived as impersonal, artificial, and with limited emotional scope for dialogue on sensitive topics that teens may wish to discuss. Furthermore, adolescents' limited access to the internet and mobile devices needs to be considered, as many have limited resources. There were mixed views on the degree of parental involvement, with teens concerned about who might have access to the information they enter and whether parents or guardians would be monitoring their conversations. Adolescents' familiarity with technology could subsequently become a challenge. If they find the chatbot uninteresting or unentertaining, they are unlikely to use it, despite the useful information it may contain.

With these strengths and challenges in mind, the chatbot's structure and key functions can be defined. Insights from the focus groups can guide decisions on the most appropriate access platform (eg, web-based vs social media). For us, it became evident that the most viable option for accessing the chatbot would be a web-based interface rather than social media or WhatsApp, given ethical, privacy, and development resource considerations.

Step 2: Human-Centered, Iterative Development Sessions

Overview

To identify key characteristics that can improve the acceptability of a chatbot for adolescents, it is essential to collaborate directly with this population. For this reason, we convened a YAB comprised of adolescents representative of the end-user population. YAB members can be selected from Step 1 participants (in our case, adolescents living with HIV), prioritizing individuals who were particularly engaged during focus groups—those who openly shared their opinions, demonstrated an understanding of adolescent mental health, and recognized the challenges faced by other adolescents living with HIV while adhering to treatment. Small advisory groups are recommended to be diverse yet compact enough to ensure meaningful participation and sustainability [36,37], and research suggests that groups of 5-7 members optimize discussion, problem-solving, and consensus building [38]. Ultimately, our YAB consisted of 4 females and 2 males, aged 11-19 years (Table 1).

XSL•FO
RenderX

Table . Sociodemographic characteristics of adolescents on the Youth Advisory Board (YAB).

Characteristic	Age (11-14 years)	Age (15-19 years)	Total
Sex			
Male	2	2	4
Female	0	2	2
Sexual orientation			
Heterosexual	2	2	4
Homosexual	0	2	2
HIV acquisition			
At birth	0	2	2
Sexually	2	2	4
Parents who died from HIV			
At least one	1	1	2
None	1	3	4

The YAB's primary task should be to test the chatbot during in-person sessions, providing feedback on its design, content, and functionality, and suggesting ideas for improvement. Sessions should be held at regular intervals, allowing sufficient time for the team to discuss, develop, and test changes between sessions. Each session should last 90 - 120 minutes, including breaks, in accessible, comfortable locations for participants.

Adolescents should access the chatbot individually on their own smartphones, when possible, while additional devices can be provided for those without access. After the interaction period, participants should share their recorded opinions and discuss points of agreement and disagreement to identify key modifications for the next version of the chatbot. With guidance from a facilitator, the YAB can then proceed to evaluate the next feature or content area selected for that session. Structured feedback tools may also be incorporated. For example, we developed a Likert-scale questionnaire (1=very dissatisfied and 5=very satisfied) to assess satisfaction with videos, usefulness and clarity of information, ease of use, and overall experience, with additional items added to evaluate specific modules such as self-help activities or pathways for seeking professional support.

Plan a series of group sessions spaced approximately 30 days apart, scheduled at times convenient for adolescent participants. Sessions should be held in a safe, youth-friendly environment. For EVA, we conducted 5 group sessions between August 2023 and January 2024. All sessions took place on Saturdays to accommodate participants' availability and were held at a nongovernmental organization that promotes health among the lesbian, gay, bisexual, transgender, queer or questioning, intersex, and asexual populations through research and programs focused on the prevention, diagnosis, and treatment of sexually transmitted infections, HIV, and AIDS. During the sessions, YAB members accessed the chatbot through a web link and spent an average of 20 minutes interacting with the content selected for that session, writing down their impressions and ideas on paper.

Because YAB meetings can generate a wide range of ideas and expectations, teams should prioritize refining prototypes to enhance functionality and implement practical solutions to address challenges and suggestions that emerge during the development process.

Session 1: Foundations of Chatbot Development

The first YAB session should introduce the study and foster participant engagement. Group activities, such as icebreaker exercises to introduce the YAB and study team members, and informal conversations, can help to build rapport and establish a trusting environment that encourages adolescents to share opinions, offer critiques, and stay meaningfully involved throughout chatbot development. Establishing trust and supportive group dynamics has been shown to enhance engagement and make digital health interventions more relatable and trustworthy to young people when they are actively involved in cocreation processes, improving the relevance and impact of the resulting resource [39].

The session should also provide an overview of the study goals and planned activities for future meetings and introduce preliminary concepts related to the chatbot's content, its proposed functionality, and structural design, to gather adolescents' perspectives and feedback before developing the first version. The session may also include opportunities for the YAB to shape visual and identity-related elements, such as the chatbot's name and avatar. Inviting adolescents to name the chatbot is especially valuable, as it can strengthen their sense of ownership in the development process.

A brainstorming activity should be conducted with the YAB to discuss key chatbot features, such as confidentiality, health content, and ideas for making the chatbot engaging for adolescent users. For EVA, the YAB provided feedback on the chatbot's avatar design (a graphic character representing the chatbot), reviewed the script for an introductory video, and decided to name the chatbot "EVA," in accordance with its main functions.

XSL•FO
RenderX

An important early activity in chatbot development is creating a flowchart that maps the chatbot's structure to its core themes and objectives. Once key modules are defined, content can be added based on each module's purpose. The first chatbot prototype, developed after Session 1, should prioritize structure over visual polish or complex interactions, allowing the YAB to explore how the chatbot communicates and what types of interactions are possible. Even a simple version helps adolescents better understand the tool's potential than explanations or external examples alone. Including basic but functional content enables users to navigate the system and provide informed feedback on information needs, usability, interaction styles, and content organization, which can then be refined in later iterations.

After the first session, our team developed an initial web-based prototype EVA (version 0.1) guided by YAB priorities and with a strong focus on confidentiality: the chatbot only asked users to enter a name and did not store conversations on their devices. Upon entry, 3 buttons directed users to the EVA modules. The "Education" module allowed users to select topics or type queries, though content was limited to 8 mental health and HIV-related terms. The "Linkage to Care" module simulated real-time written communication with a study team member as an example of connection to professional support, while the "Self-Help" module was not yet active.

Session 2: Testing the EVA Prototypes

After a workable, limited-feature prototype chatbot is developed, repeat consultation with the advisory board is essential. Two important reasons for early consultation with the advisory board, albeit with an underdeveloped chatbot, are to confirm the core features of the chatbot planned during Step 1 and to strengthen the collaborative relationship with the advisory board by demonstrating how their input shaped the chatbot programming (and explaining why some features could not be included). For the first chatbot testing experience with a YAB, ensure that access to the chatbot is simple and reliable. Provide direct entry (eg, via web link) and verify beforehand that the platform runs smoothly on common mobile devices. Facilitators should ideally be members of the development team so they can monitor real-time use, identify navigation or performance issues, and resolve technical problems immediately. Stable Wi-Fi access is essential to prevent connectivity disruptions, and backup devices (eg, smartphones or tablets) should be available for participants whose personal devices are incompatible or have limited functionality.

Although adolescents should be given time for independent exploration proportional to the amount of content under review, the first session should also include a guided walkthrough. Early prototypes are often not fully intuitive, so facilitators should briefly explain the purpose, structure, and function of each module before or during use. This helps align expectations and ensures that feedback reflects the tool's intended design rather than confusion about navigation. The criteria for which suggestions to implement, adapt, or exclude should consider the development team's technical capacity, available time, and financial resources, and the relevance of each modification to the intervention's goals.

For EVA's second session, YAB members accessed EVA (version 0.1) via a web link and explored it for 20 minutes. Afterward, technical issues were documented, and group feedback was collected on content and functionality. Adolescents confirmed that only essential information, such as name and age, should be required to access the chatbot, avoiding sensitive data such as phone numbers and ensuring that conversations are deleted after use.

For content, the YAB prioritized anxiety and depression as key topics and suggested expanding the content to include self-esteem, bullying, eating and nutrition issues, and HIV, along with practical self-help resources (eg, relaxation, social skills, stress management, and coping with negative thoughts). They also recommended using simple language to make the content more accessible for adolescents of different ages, reducing text, and adding more visual elements.

Following each session, the team systematically reviewed the feedback, identifying priority improvements and defining key technical or design challenges. Then, we developed an updated version of the chatbot that incorporated feasible recommendations while maintaining its core structure and objectives (Table 2). To enhance the chatbot's presentation and navigation, an interactive web environment was created using the Wix (Wix.com Ltd) platform [40]. This transition gave the study team full control over the chatbot's interface, enabling real-time error corrections during subsequent sessions with the YAB.

XSL•FO
RenderX

Table . Development of key features of the EVA (*Educación, Vinculación, y Autoayuda*) chatbot.

Feature	Version 0.1 (YAB[a] Session 2)	Version 0.2 (YAB Session 3)	Version 0.3 (YAB Session 4)	Version 1.0 (YAB Session 5)
Chatbot presentation	• No welcome screen: initially, the chatbot displayed an image of the EVA[b] avatar along with a description clarifying that it was a chatbot, not a human.	• Two welcome screens were created on the Wix platform: Screen 01 displays EVA's avatar and a "Hello!" button, which can be clicked to proceed to the next screen, and Screen 02 introduces EVA through vignettes, images, and a video. Clicking the "Continue" button grants access to the chatbot.	• Screen 01: improved by animating EVA's full-body appearance and Screen 02: no change.	• Screen 01: changed the "Hello!" button to "Start" and Screen 02: no change.
Modules built	• Introduction: the chatbot greets users, asks for their name and age, and provides 3 options: Education, Linkage to Care, and Self-Help. • Education: users can choose between 2 topics (depression and anxiety) or enter a custom query. The chatbot recognizes 8 keywords (anxiety, depression, sadness, treatment, HIV, healthy living, alcohol, stress) to provide more relevant information. • Linkage to Care: if users opt to speak to a professional, a real-time chat is initiated. • Self-Help: no content.	• Introduction: the welcome greeting was simplified and made more conversational. • Education: the module now introduces depression and anxiety, allowing navigation through buttons. Users are then presented with additional topics (stress and HIV) with submenus containing key information. • Linkage to Care: users can now access this feature via a permanent button on the chatbot screen, which indicates whether a live agent is available. • Self-help: no content.	• Introduction: after entering information, users are redirected to a new "Introduction to Depression" module. • Introduction to Depression: provides basic information about depression in a conversational manner. Once completed, users can navigate through the 3 modules via swipeable cards. • Education: topics of depression, anxiety, and additional issues (stigma and treatment adherence) are now displayed on swipeable cards. • Linkage to Care: the permanent on-screen button was removed. Instead, users can access options such as "Seek Professional Help" to start a real-time conversation or browse an emergency directory with contact information for crisis support. • Self-Help: now includes 3 activities: deep breathing exercises, managing emotions (with 3 subactivities), and a game about facts and myths regarding HIV.	• Introduction: no changes. • Introduction to Depression: content was refined to be more concise, with fewer images and more accurate information. Emojis were added. • Education: now includes eating disorders under additional issues • Linkage to Care: the number of steps required to start a real-time conversation was reduced. • Self-Help: a game about facts and myths regarding mental health in adolescents was added.

[a]YAB: Youth Advisory Board.

[b]EVA: *Educación, Vinculación, y Autoayuda*.

Session 3: Chatbot Interface Refinement and Assessment

For the third session, focus on reviewing the first improved version of the chatbot, highlighting updates made after initial feedback, including content integration, visual design adjustments, usability improvements, and correction of previously detected errors. At this stage, conduct a detailed and paced exploration of each selected chatbot section. Facilitators

XSL•FO
RenderX

should guide navigation to ensure all components are reviewed and understood. Use simple guiding questions such as: “What do you think about this section? Is the information easy to understand? What would you change to improve it?” After reviewing each section, hold a brief group discussion before moving on to the next one. This structure allows feedback to remain specific and helps the group reach agreement on priority modifications related to navigation, presentation, and content.

In our project, each YAB member received blank sheets to record ideas and observations while reviewing each section of EVA (version 0.2). These notes were then used during the section-by-section discussions to help participants recall key impressions, compare perspectives, identify similarities and differences, and agree on priority changes for the next chatbot iteration. Participants found EVA (version 0.2) more visually appealing but noted that the initial screen was not intuitive, recommending clearer user instructions (Figure 2). The welcome message and introductory video—developed in the Wix web environment—were well received (Figure 3). While navigation buttons were considered useful, adolescents suggested limiting their number and using them mainly for simple choices, as excessive button use felt stressful, with a preference for more text-based interaction to increase engagement. They also reviewed 2 new videos created for the chatbot and recommended adding subtitles to all videos to improve accessibility and comprehension.

The “Education” module was viewed as clear and useful, though participants suggested a more engaging and optimistic tone, particularly for sensitive topics such as depression and anxiety. They further recommended including short, relatable stories about common adolescent experiences (eg, bullying, school stress, body image, eating concerns, HIV, and home situations) to make the content more meaningful.

Incorporating a quantitative method to monitor progress helps identify which chatbot features adolescents find most and least acceptable, enabling the team to track improvements and prioritize refinements. For EVA, a brief satisfaction scale using Likert-type items (1=very dissatisfied and 5=very satisfied) assessed the videos, the usefulness of information, content clarity, and engagement (Table 3). Additional items can be incorporated in later sessions to evaluate emerging features as the chatbot evolves. Results showed generally positive perceptions of usefulness and educational value, while lower ratings identified issues such as unintuitive design and repetitive navigation, both of which required further improvement.

Figure 2. Improving first-screen usability through clear instructions, replacing the greeting (“Hello!”) with a direct-action cue (“Start”) to help users intuitively understand how to continue. EVA: *Educación, Vinculación, y Autoayuda*.

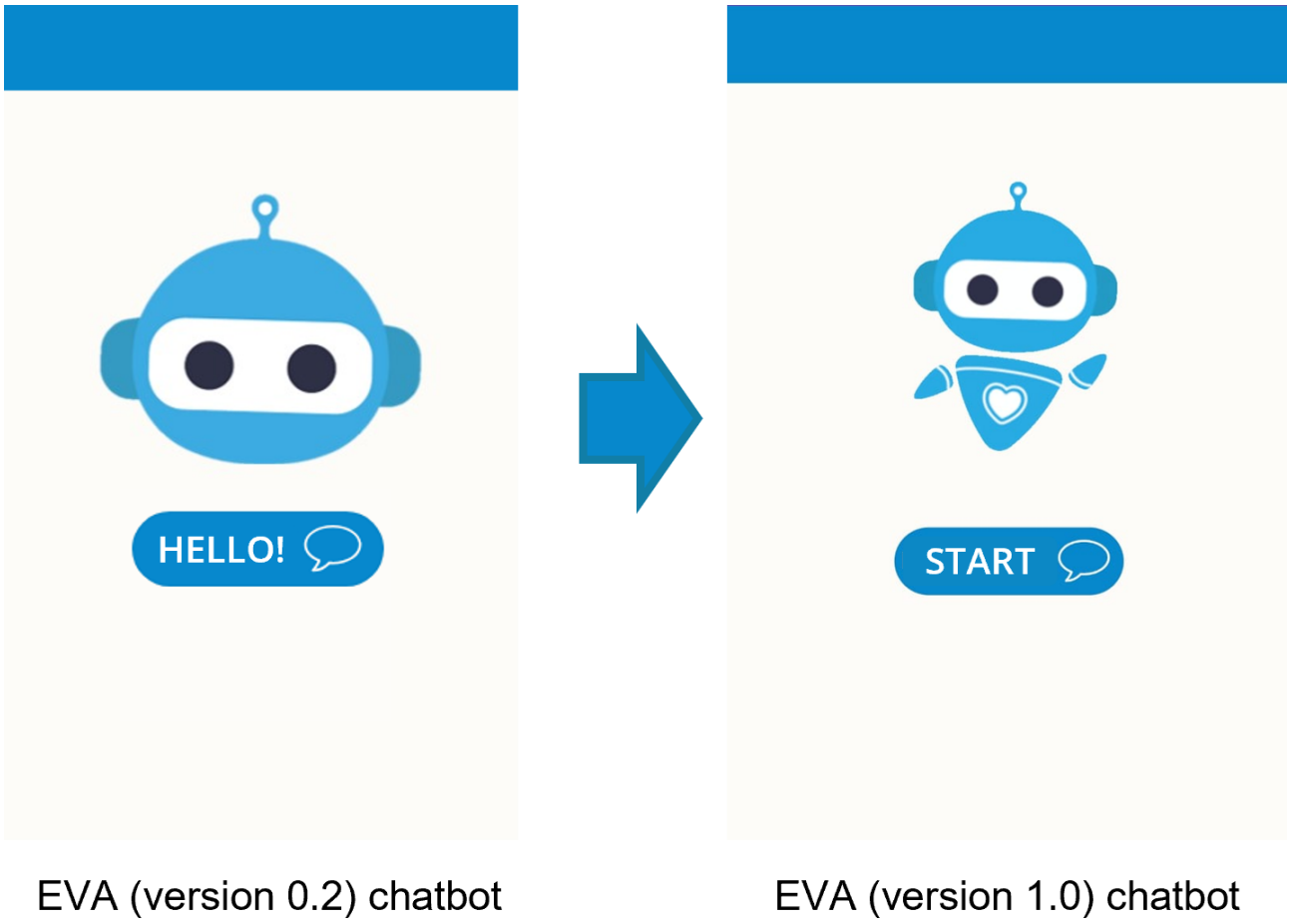

XSL•FO
RenderX

Figure 3. Audiovisual welcome screen to enhance early engagement. EVA (Educación, Vinculación, y Autoayuda) introduces itself as a virtual assistant and uses video and interactive elements to preview what users will experience.

XSL•FO
RenderX

Table . Summary of chatbot characteristics evaluated in Sessions 3 and 4.

Characteristic evaluated	Session 3, mean (SD; range)	Session 4, mean (SD; range)
Like on videos	4 (0; 4)	4.5 (0.55; 4-5)
Useful information	4 (0.45; 4-5)	4 (0.52; 4-5)
Clear information	4 (0.45; 4-5)	4 (0.41; 4-5)
Entertaining chatbot	3 (1.14; 1-4)	4 (0.75; 3-5)
Easy to use	3 (0.55; 3-4)	4 (0.52; 3-5)
Overall assessment	4 (0.45; 4-5)	4 (0.75; 4-5)
Self-help module utility	Not measured	4 (0.52; 4-5)
Use to seek help	Not measured	4 (0.52; 4-5)

Session 4: Chatbot Module Expansion and Usability Improvement

At this stage of development, once key design principles have been defined—such as appropriate text length, simple and age-appropriate language, the chatbot's tone when addressing users, preferred input and interaction formats, and priority mental health topics—it becomes easier to expand the chatbot with new modules or content while maintaining coherence with its objectives. Teams should draft scripts for each new section to maintain a structured overview of content and flow, which facilitates revisions and consistency. When sharing scripts with adolescents, adapt them with visual aids (eg, illustrations or diagrams), especially when the text is extensive, as this supports comprehension and allows for more focused feedback. Reviewing scripts before implementation is particularly effective once participants are familiar with the chatbot's structure.

For this step, prioritize accessibility, intuitive navigation, and engaging presentation. Introduce new ways of displaying information—even for previously reviewed topics—and invite feedback on these interaction formats. Since both the development team and YAB members become increasingly familiar with the chatbot over time, emphasize presenting new or updated content rather than repeating unchanged sections. Temporary navigation shortcuts (eg, quick-access buttons to specific modules) can help streamline testing sessions.

During EVA (version 0.3) testing, quick-access buttons were added to the home screen to streamline navigation during the session. The YAB reviewed a new "Introduction to Depression" module (Figure 4), first annotating feedback on a printed script and then testing its functionality within the chatbot. Alternative navigation formats, including sliding cards, were introduced to improve content presentation and reduce repetitive button use (Figure 5), alongside updated self-help activities and new videos.

Participants reported that the interface felt more intuitive and preferred text-based interaction rather than repeatedly selecting buttons. The videos were well received, though members suggested incorporating additional visual elements. While the new module was considered useful, it was perceived as too lengthy. Extended use of the chatbot also revealed additional programming issues. The YAB recommended expanding educational topics (eg, eating disorders, bullying, and the roles of mental health care professionals). Satisfaction scores indicated improvements in navigation and engagement (Table 3). Final refinements focused on condensing content to enhance readability and resolving remaining technical issues.

XSL•FO
RenderX

Figure 4. Introduction to Depression module. In a conversational way, it introduces and connects basic notions about emotions, mental health, and depression.

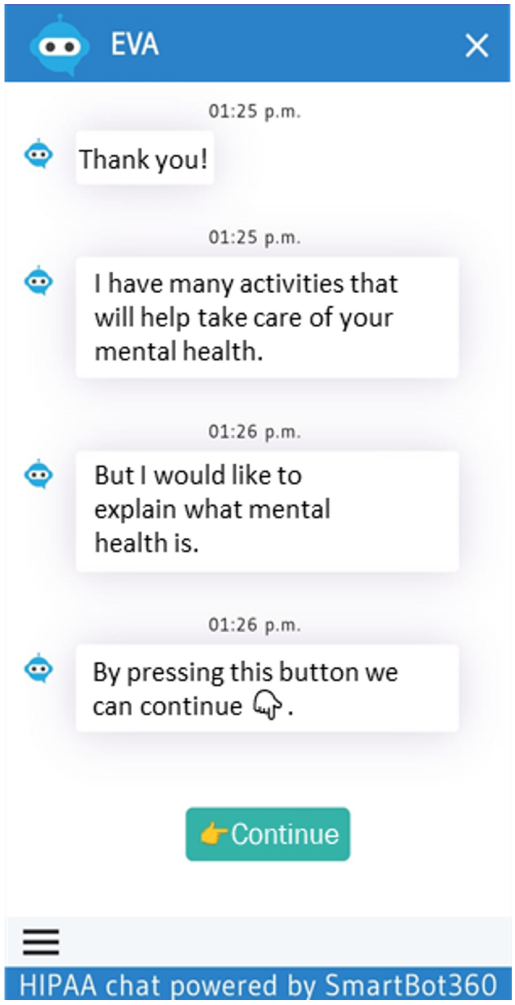

Button to start the module after collecting the user's name and age.

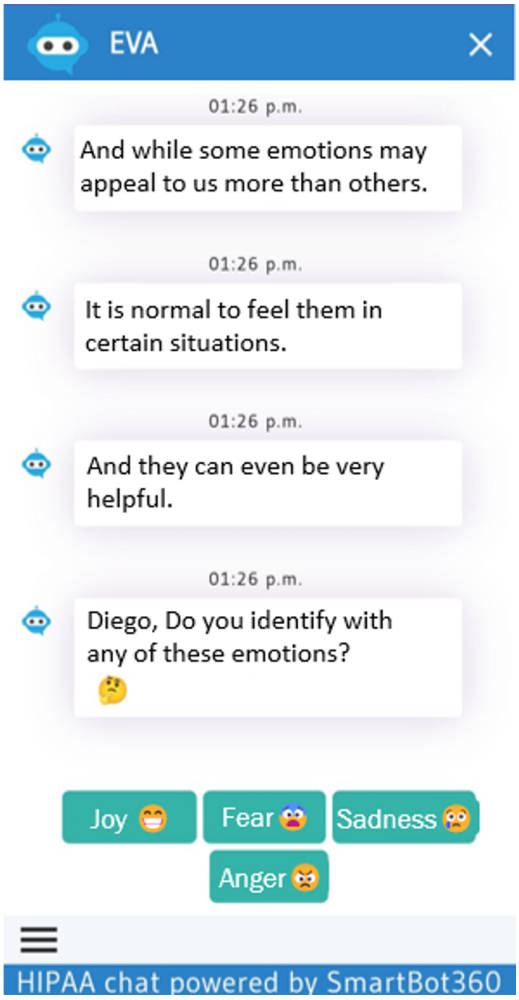

Some interaction resources: user name, emojis, multiple choice.

XSL•FO
RenderX

Figure 5. Changes in the presentation of EVA (*Educación, Vinculación, y Autoayuda)* chatbot modules. A cleaner and more pleasant appearance was achieved through shorter, more precise texts and improved visual resources (sliding cards).

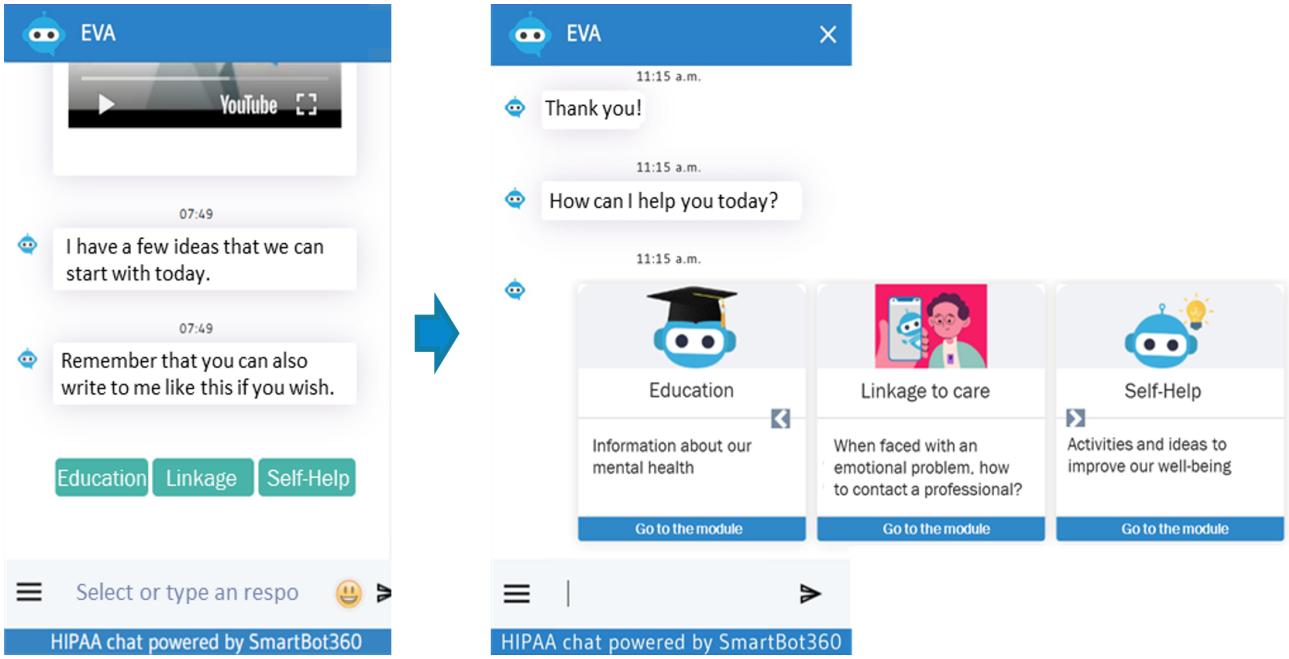

EVA (version 0.2) chatbot EVA (version 0.3) chatbot

Session 5: Final Review and Validation

In the final validation session, present the chatbot in a format that closely reflects how a new user will experience it. Remove temporary testing features such as quick-access buttons or shortcuts and allow adolescents to navigate the tool under realistic conditions. Provide sufficient time for free exploration of all content; if the material is extensive, organize the review by modules or sections while maintaining a natural flow of use.

After each review segment, collect adolescents' feedback. Because participants are already familiar with the chatbot, facilitators should probe more deeply into their responses, asking why they hold certain opinions or suggest specific changes. Comparing the current version with earlier prototypes can help identify which modifications were most positively received and which areas may still need refinement.

Validation is considered achieved when YAB members show broad agreement that the content, tone, and functionality are appropriate for the chatbot's intended purpose and future implementation. To support this, clearly explain how the chatbot will be used in the next phase of the project. If significant suggestions or unresolved issues remain, schedule an additional review session before finalizing the tool.

The YAB approved the overall interface and audiovisual materials, suggesting only minor navigation improvements, such as adding text labels (eg, "Back") to icon-only buttons to make navigation more intuitive. They responded positively to the shortened "Introduction to Depression" module, noting that it delivered key information without overlapping with other content in the "Education" module. The self-help activities were found to be engaging and clear, particularly an interactive emotion-regulation story. The "Linkage to Care" module was also considered easy to access and use (Figure 6).

XSL•FO
RenderX

Figure 6. Linkage to Care module screen. EVA (*Educación, Vinculación, y Autoayuda*) shows whether a professional is available and provides buttons to start a chat or return to previous or main menus.

XSL•FO
RenderX

Including multimedia components is strongly recommended when developing digital tools for adolescents, as brief audiovisual materials can enhance engagement, comprehension, and information retention, particularly when addressing sensitive topics. Evidence from digital health interventions [21,41] shows that short, structured multimedia content improves user engagement and perceived usefulness in youth-focused eHealth tools. Accordingly, videos integrated into chatbots, or similar applications, should be concise (ideally under 1 minute), use clear and age-appropriate language, and present relatable scenarios or examples.

From a user-centered design perspective, decisions about audiovisual tone, style, and priority topics should be explored in early stages and validated with adolescent advisors before chatbot development begins. Because chatbot content often evolves during development, a practical first step is to create a short introductory video that explains the chatbot's purpose, structure, and main modules. The design, characters, and visual style defined in this first video are crucial, as they establish the aesthetic and narrative coherence for subsequent audiovisual materials. As content is defined, additional short videos can be developed to introduce key concepts, explain unfamiliar terms, or highlight essential features, such as self-help strategies or linkage to professional care, ensuring alignment between multimedia elements and the tool's overall objectives.

For EVA, an audiovisual consultant produced 5 animated videos of approximately 30 seconds each (Table 4), based on scripts and visual guidelines from the research team, which were subsequently reviewed with the YAB. The animation style was selected following adolescent consultation and in consideration of the broad target age range (10 - 19 years) to ensure relevance and acceptability across developmental stages. The videos were stored in a dedicated YouTube (Google LLC) [42] account and embedded within the chatbot, allowing users to view them directly through the chatbot interface.

Table . Videos developed for the EVA (*Educación, Vinculación, y Autoayuda*) chatbot.

No	Video title (Spanish)	Chatbot location	Main objective
1	Hello! I am EVA (*"¡Hola!, soy EVA"*)	Welcome screen	Introduce the chatbot and its structure.
2	Self-Help (*"Autoayuda"*)	Self-Help module	Explain the purpose of the self-help section.
3	Why seek help? (*"¿Por qué buscar ayuda?"*)	Linkage to Care module	Promote help-seeking behavior.
4	Adherence (*"Adherencia"*)	Education module	Highlight the link between mood and health behaviors.
5	Stigma (*"Estigma"*)	Education module	Address stigma related to HIV and mental health.

Step 3: Testing Before Launch

An external review by professionals who were not involved in the development process is recommended to mitigate familiarity bias that may emerge as the development team and adolescent advisors become accustomed to the tool during iterative construction. Methodological research on co-design development highlights the importance of involving diverse stakeholders to enhance content validity, detect blind spots, and strengthen credibility before implementation [43-45]. In digital health design, incorporating external expert review is also considered a strategy to improve rigor, transferability, and content accuracy. Ideally, these reviewers should have experience with the target population and the intervention's thematic focus, as this allows feedback not only on usability but also on clinical relevance and appropriateness.

In our study, after YAB approval of EVA (version 1.0), 20 - 30-minute test sessions were conducted with 6 external HIV health care professionals (4 physicians, 1 nurse, and 1 psychologist) working with adolescents living with HIV. Their feedback led to simplifying navigation elements (eg, clearer button labels, shorter welcome messages, standardized "Continue" buttons), refining slide-out menus, and reducing textual redundancy. The "Introduction to Depression" module was reframed to begin with emotions rather than abstract mental health definitions to enhance relatability. Content was further tailored to adolescents living with HIV by condensing general prevention information and expanding topics such as sexual health, relationships, HIV transmission, and treatment adherence. Final internal testing sessions were then conducted to verify functionality and content coherence before proceeding to formal feasibility and acceptability evaluation with a new sample of adolescents.

EVA's Final Version: System and Modules Structure

A final chatbot should integrate the core modules identified through user research and iterative testing, organized into a clear navigational structure. For EVA, the chatbot was developed using Smartbot360 and structured into 5 sections: "Introduction," "Introduction to Depression," "Education," "Linkage to Care," and "Self-Help" (Figure 7). A final video that describes and promotes EVA was also produced (refer to Multimedia Appendix 1).

After 2 welcome screens featuring an introductory video, the chatbot prompts the user for their name and age. Age is collected to confirm eligibility within the target range of 10 - 19 years, detect out-of-range users, inform health care professionals in an eventual follow-up, and for future analyses. The "Introduction

XSL•FO
RenderX

to Depression" module provides brief, adolescent-friendly information on mental health and depression. It is displayed only during the user's first access to the chatbot, after which full access to the core modules is granted.

Figure 7. Structure and user flow of the EVA (*Educación, Vinculación, y Autoayuda*) chatbot.

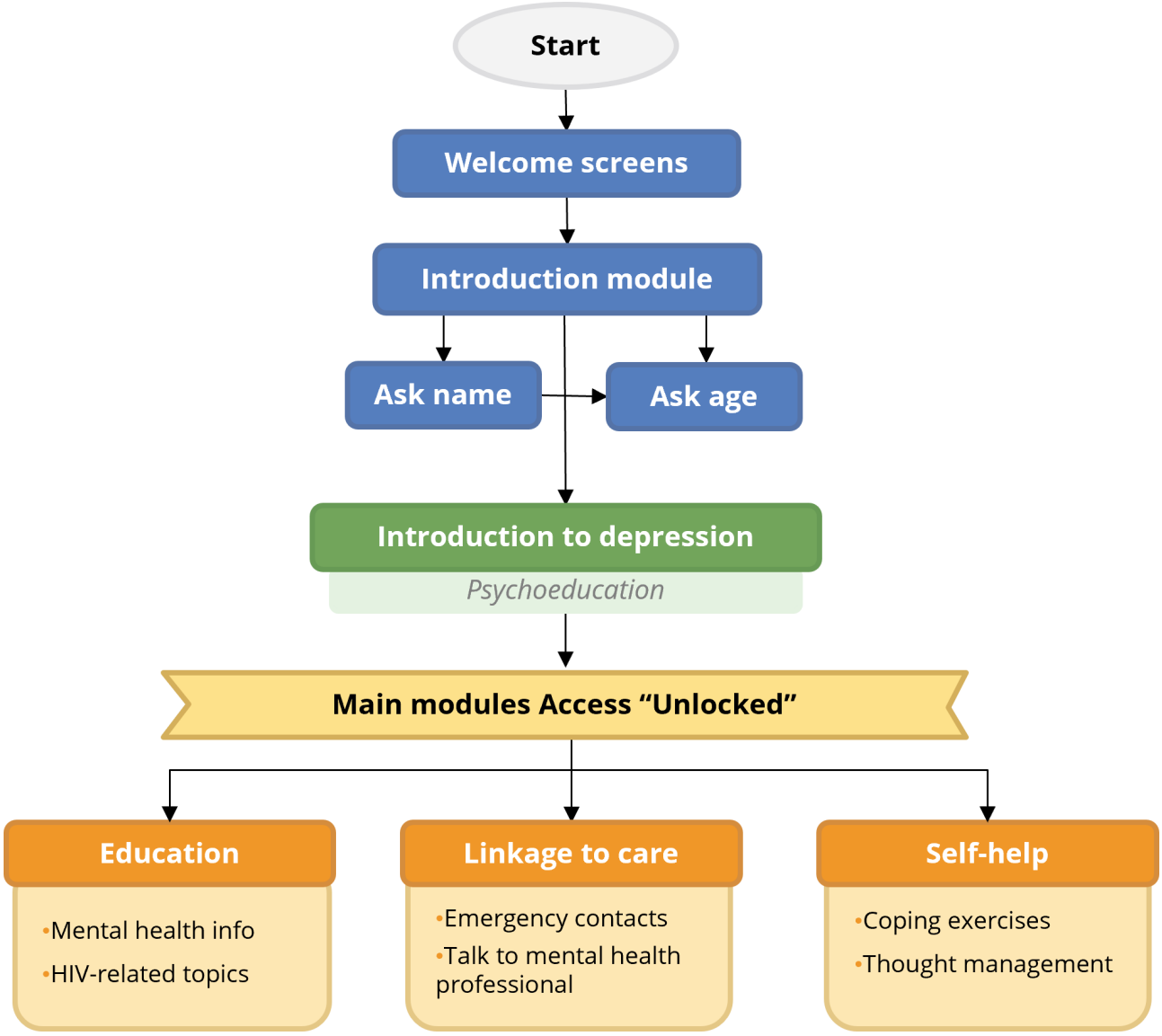

The "Education" module covers depression, anxiety, eating disorders, HIV-related topics, treatment adherence, and discrimination. The "Linkage to Care" module includes emergency contacts and a text-based feature to connect with health care professionals. The "Self-Help" module offers interactive activities focused on relaxation, emotional regulation, and myth clarification related to HIV and mental health.

When budgeting for chatbot development, teams should account for platform subscription fees, multimedia production, personnel time, and logistical costs for participatory sessions. For EVA, development costs included Smartbot360 subscription (US $99/month for up to 10,000 messages) and production of 5 animated videos (US $2710). Personnel time and logistical expenses related to YAB sessions (eg, refreshments, transportation, and incentives) were excluded from this estimate. Teams should develop a detailed budget early in the planning process, as actual costs may vary significantly depending on design complexity, content volume, and the inclusion of additional human resources.

Lessons Learned

This tutorial provides a step-by-step guide to the recommended and implemented decisions during the development of the EVA chatbot.

First, actively involve adolescents from the earliest conceptual stages. Early qualitative interviews, focus groups, and the creation of a YAB ensured that the chatbot's tone, structure, aesthetics, and priority topics reflected adolescents' lived experiences. Similar participatory approaches have been used in the development of mental health chatbots such as [32], Woebot for Teens (Woebot Health) [21], and Zuri [46], underscoring the role of co-design in cultural relevance, acceptability, and engagement. Prior qualitative exploration is particularly useful to define conversational style and interface

XSL•FO
RenderX

decisions before programming begins [47-49]. Transparent discussions about technical constraints, such as limitations in advanced artificial intelligence (AI), are also recommended to maintain trust and realistic expectations [21,35,50].

Second, adopt an explicitly iterative development model. Continuous feedback cycles led to improvements in readability, emotional tone, and module length in response to adolescents' preferences for autonomy and brevity [47]. Such user-driven refinement is essential, as adolescents must perceive the intervention as not only useful but also motivating and hopeful, particularly when discussing sensitive topics such as depression and HIV-related stigma. These principles align with iterative testing approaches described in chatbots such as Wysa (WYSA Ltd) and Woebot (Woebot Health) [21,45,46]. Ensuring an encouraging, supportive, and culturally sensitive tone aligns with evidence that highlights the importance of fostering optimism and self-efficacy in adolescent digital interventions [21,51].

Third, tailor content to the specific psychosocial context of the target population. EVA's "Introduction to Depression" module, for instance, uses destigmatizing language to normalize conversations about emotional distress and reduce barriers to help-seeking, reflecting the reality that many adolescents living with HIV face limited family support and restricted access to reliable mental health information due to socioeconomic constraints. Comparable design choices have been described in chatbots such as Zuri in Kenya [46]. Framing depression through emotions and common symptoms, rather than solely as a clinical condition, promotes a more accessible understanding of mental health and may strengthen adolescents' motivation to seek care.

Finally, maintain direct control over the development tools whenever feasible. The ability to implement real-time corrections facilitates immediate responses to YAB feedback and transparent dialogue about technical possibilities and limitations. Similar hands-on, low-cost development models have been reported in LMICs, where research teams favored flexibility and transparency over sophisticated automation [21]. Involving health care professionals with expertise in HIV and adolescent mental health validated and refined EVA's content, ensuring accuracy, clinical relevance, and practical utility [48].

Costs and Sustainability

The EVA development used relatively low-cost resources: an affordable messaging platform, free web hosting, and open-access images, demonstrating the feasibility of building a functional mental health chatbot in LMICs. Overall costs will vary significantly depending on the complexity of the design, the software required, the audiovisual production quality, and the volume of content included. Comparatively, digital mental health chatbots that integrate advanced AI require substantial investments—particularly when integrating regulatory compliance or large-scale technical deployment—and may even reach tens of thousands of dollars for full development [21,41,52]. In contrast, development teams can minimize costs by foregoing specialized technical staff and relying on free hosting and basic content production.

Nevertheless, replications or expanded versions of this study may require considerably higher budgets if professional developers, graphic designers, robust web infrastructure, or professional audiovisual production are included. Additionally, other costs—such as hiring part-time support staff, organizing meetings with new YAB members, providing incentives, transportation, and testing devices—would only be relevant if the participatory phases of the study are explicitly replicated. Thus, the planning of similar initiatives in LMICs should account for a baseline of moderate costs based on functional platforms and accessible content, while also preparing scalable scenarios for more complex versions that integrate professional technical development, secure infrastructure, and expanded community participation.

Challenges and Limitations

Practitioners should anticipate several challenges when developing similar tools. Although participatory approaches and iterative testing with a YAB can yield significant improvements, some limitations—such as software constraints and an initial reliance on button-based navigation—may not fully meet adolescents' desire for more advanced, AI-driven conversations found in freely available digital products they are already familiar with. Future efforts should explore more sophisticated platforms that offer greater flexibility, interactivity, and personalization, as well as improved data collection methodologies to track and prioritize user feedback systematically. Standardizing evaluation methods across development stages and expanding content on related issues, such as bullying, eating disorders, and distinctions between mental health care professionals, may increase a chatbot's relevance and uptake.

Conclusions

This tutorial describes the step-by-step development of a mental health chatbot through sustained collaboration between adolescents, researchers, and clinicians. The process illustrates how structured adolescent engagement, clear goal-setting, and iterative refinement are fundamental to producing digital tools that are acceptable and feasible for end users. Through continuous feedback cycles and validation by mental health and HIV specialists, EVA achieved a balance between scientific accuracy, emotional resonance, and cultural sensitivity.

Our experience also highlights that participatory, low-cost, and contextually grounded design processes can produce relevant and sustainable digital interventions in resource-limited settings. Future versions of similar tools may incorporate more advanced conversational features, digital assent mechanisms, and scalable infrastructure to enhance sustainability. Although challenges remain, the lessons presented throughout this tutorial provide valuable guidance for researchers and practitioners seeking to develop adolescent-centered digital health interventions, contributing to improved quality, accessibility, and impact of digital support for adolescents living with HIV and beyond. We encourage teams to adapt the steps outlined here to their own contexts and populations, using participatory methods to ensure

XSL•FO
RenderX

that the tools they create are both evidence-informed and responsive to the needs of the communities they serve.

Acknowledgments

We are grateful to the adolescents who participated in this study. The authors declare the use of generative AI (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing, translation, and reformatting. The GAI tool used was ChatGPT-4. Responsibility for the final paper lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by Diego Vasquez.

Funding

Funding for this project was provided in full by a CIPHER (Collaborative Initiative for Pediatric HIV Education and Research) grant from the International AIDS Society. The views expressed in this paper do not necessarily reflect the official policies of the International AIDS Society.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Video made at the end of the study to explain the content and operation of the EVA (*Educación, Vinculación, y Autoayuda*) chatbot. Reproduced from Socios En Salud [53].
[MP4 File, 92997 KB - jmir_v28i1e75351_app1.mp4]

References

1. Adolescent HIV treatment. UNICEF. 2025. URL: https://data.unicef.org/topic/hivaids/adolescent-hiv-treatment/ [accessed 2025-08-20]
2. Mellins CA, Malee KM. Understanding the mental health of youth living with perinatal HIV infection: lessons learned and current challenges. J Int AIDS Soc 2013 Jan;16(1):18593. [doi: 10.7448/IAS.16.1.18593]
3. Laurenzi CA, Skeen S, Gordon S, et al. Preventing mental health conditions in adolescents living with HIV: an urgent need for evidence. J Int AIDS Soc 2020 Sep;23 Suppl 5(Suppl 5):e25556. [doi: 10.1002/jia2.25556] [Medline: 32869530]
4. Adejumo OA, Malee KM, Ryscavage P, Hunter SJ, Taiwo BO. Contemporary issues on the epidemiology and antiretroviral adherence of HIV-infected adolescents in sub-Saharan Africa: a narrative review. J Int AIDS Soc 2015 Jan;18(1):20049. [doi: 10.7448/IAS.18.1.20049] [Medline: 26385853]
5. Agwu AL, Fairlie L. Antiretroviral treatment, management challenges and outcomes in perinatally HIV‐infected adolescents. J Int AIDS Soc 2013 Jan;16(1):18579. [doi: 10.7448/IAS.16.1.18579] [Medline: 23782477]
6. Naar-King S, Templin T, Wright K, Frey M, Parsons JT, Lam P. Psychosocial factors and medication adherence in HIV-positive youth. AIDS Patient Care STDS 2006 Jan;20(1):44-47. [doi: 10.1089/apc.2006.20.44] [Medline: 16426155]
7. Ayano G, Demelash S, Abraha M, Tsegay L. The prevalence of depression among adolescent with HIV/AIDS: a systematic review and meta-analysis. AIDS Res Ther 2021 Apr 27;18(1):23. [doi: 10.1186/s12981-021-00351-1] [Medline: 33906698]
8. Peru: Demographic and Family Health Survey 2022 (ENDES 2022). : National Institute of Statistics and Informatics (INEI); 2023 URL: https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1898/libro.pdf [accessed 2026-03-20]
9. Claborn KR, Meier E, Miller MB, Leffingwell TR. A systematic review of treatment fatigue among HIV-infected patients prescribed antiretroviral therapy. Psychol Health Med 2015;20(3):255-265. [doi: 10.1080/13548506.2014.945601] [Medline: 25110152]
10. Wainberg ML, Scorza P, Shultz JM, et al. Challenges and opportunities in global mental health: a research-to-practice perspective. Curr Psychiatry Rep 2017 May;19(5):28. [doi: 10.1007/s11920-017-0780-z] [Medline: 28425023]
11. Lee RS, Kochman A, Sikkema KJ. Internalized stigma among people living with HIV-AIDS. AIDS Behav 2002 Dec;6(4):309-319. [doi: 10.1023/A:1021144511957]
12. Steward WT, Herek GM, Ramakrishna J, et al. HIV-related stigma: adapting a theoretical framework for use in India. Soc Sci Med 2008 Oct;67(8):1225-1235. [doi: 10.1016/j.socscimed.2008.05.032] [Medline: 18599171]
13. Mburu G, Hodgson I, Kalibala S, et al. Adolescent HIV disclosure in Zambia: barriers, facilitators and outcomes. J Int AIDS Soc 2014 Jan;17(1):18866. [doi: 10.7448/IAS.17.1.18866]
14. Fusar-Poli P, Estradé A, Esposito CM, et al. The lived experience of mental disorders in adolescents: a bottom-up review co-designed, co-conducted and co-written by experts by experience and academics. World Psychiatry 2024 Jun;23(2):191-208. [doi: 10.1002/wps.21189] [Medline: 38727047]

XSL•FO
RenderX

15. Mental Health Atlas 2020: Country Profile (Peru). : World Health Organization; 2020 URL: https://cdn.who.int/media/docs/default-source/mental-health/mental-health-atlas-2020-country-profiles/per.pdf [accessed 2026-03-20]
16. Toyama M, Castillo H, Galea JT, et al. Peruvian mental health reform: a framework for scaling-up mental health services. Int J Health Policy Manag 2017 Sep 1;6(9):501-508. [doi: 10.15171/ijhpm.2017.07] [Medline: 28949462]
17. Vreeman RC, McCoy BM, Lee S. Mental health challenges among adolescents living with HIV. J Int AIDS Soc 2017 May 16;20(Suppl 3):21497. [doi: 10.7448/IAS.20.4.21497] [Medline: 28530045]
18. Haines C, Loades ME, Coetzee BJ, Higson-Sweeney N. Which HIV-infected youth are at risk of developing depression and what treatments help? A systematic review focusing on Southern Africa. Int J Adolesc Med Health 2021 Oct 13;33(5):20190037. [doi: 10.1515/ijamh-2019-0037]
19. Galea JT, Contreras C, Wong M, et al. A non-specialist depression care pathway for adolescents living with HIV and transitioning into adult care in Peru: a nested, proof of concept pilot study. Glob Ment Health 2021;8:e17. [doi: 10.1017/gmh.2021.16]
20. Galea JT, Vasquez DH, Rupani N, et al. Development and pilot-testing of an optimized conversational agent or “Chatbot” for Peruvian adolescents living with HIV to facilitate mental health screening, education, self-help, and linkage to care: protocol for a mixed methods, community-engaged study. JMIR Res Protoc 2024 May 7;13:e55559. [doi: 10.2196/55559] [Medline: 38713501]
21. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 6;4(2):e19. [doi: 10.2196/mental.7785] [Medline: 28588005]
22. Errea RA, Wong M, Senador L, et al. Impacto de la pandemia por SARS-CoV-2 en adolescentes que viven con el VIH en Lima, Perú. Rev Peru Med Exp Salud Publica 2021;38(1):153-158. [doi: 10.17843/rpmesp.2021.381.6296]
23. Santa-Cruz J, Moran L, Tovar M, et al. Mobilizing digital technology to implement a population-based psychological support response during the COVID-19 pandemic in Lima, Peru. Glob Ment Health 2022;9:355-365. [doi: 10.1017/gmh.2022.36]
24. Tzelios C, Contreras C, Istenes B, et al. Using digital chatbots to close gaps in healthcare access during the COVID-19 pandemic. Public Health Action 2022 Dec 21;12(4):180-185. [doi: 10.5588/pha.22.0046] [Medline: 36561900]
25. Auernhammer J. Human-centered AI: the role of human-centered design research in the development of AI. 2020 Presented at: Design Research Society (DRS) International Conference 2020: Synergy; Aug 11-14, 2020. [doi: 10.21606/drs.2020.282]
26. Borghouts J, Eikey E, Mark G, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387. [doi: 10.2196/24387] [Medline: 33759801]
27. Lyon AR, Munson SA, Renn BN, et al. Use of human-centered design to improve implementation of evidence-based psychotherapies in low-resource communities: protocol for studies applying a framework to assess usability. JMIR Res Protoc 2019 Oct 9;8(10):e14990. [doi: 10.2196/14990] [Medline: 31599736]
28. Tahvanainen L, Tetri B, Ahonen O. Exploring and extending human-centered design to develop AI-enabled wellbeing technology in healthcare. In: Särestöniemi M, editor. Digital Health and Wireless Solutions: Springer; 2024:288-306. [doi: 10.1007/978-3-031-59091-7_19]
29. What is HIPAA compliance? Compliancy Group. 2024. URL: https://compliancy-group.com/what-is-hipaa-compliance/ [accessed 2025-08-22]
30. AI virtual health agents. 24/7 patient support. Capacity. 2024. URL: https://smartbot360.com/ [accessed 2025-12-10]
31. Kosyluk K, Baeder T, Greene KY, et al. Mental distress, label avoidance, and use of a mental health chatbot: results from a US survey. JMIR Form Res 2024 Apr 12;8:e45959. [doi: 10.2196/45959] [Medline: 38607665]
32. Sanabria G, Greene KY, Tran JT, et al. “A great way to start the conversation”: evidence for the use of an adolescent mental health chatbot navigator for youth at risk of HIV and other STIs. J Technol Behav Sci 2023 May 11;8:1-10. [doi: 10.1007/s41347-023-00315-4] [Medline: 37362063]
33. Rupani N, Vasquez DH, Contreras C, et al. “Un amigo secreto”: perceived acceptability of a chatbot for depression education, self-help and linkage to care among adolescents living with HIV in Peru. AIDS Care 2025 Oct;37(10):1718-1731. [doi: 10.1080/09540121.2025.2518731] [Medline: 40743294]
34. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017 Jan 26;17(1):88. [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
35. Mummah SA, Robinson TN, King AC, Gardner CD, Sutton S. IDEAS (Integrate, Design, Assess, and Share): a framework and toolkit of strategies for the development of more effective digital interventions to change health behavior. J Med Internet Res 2016 Dec 16;18(12):e317. [doi: 10.2196/jmir.5927] [Medline: 27986647]
36. Jacquez F, Vaughn LM, Wagner E. Youth as partners, participants or passive recipients: a review of children and adolescents in community-based participatory research (CBPR). Am J Community Psychol 2013 Mar;51(1-2):176-189. [doi: 10.1007/s10464-012-9533-7] [Medline: 22718087]
37. Wong NT, Zimmerman MA, Parker EA. A typology of youth participation and empowerment for child and adolescent health promotion. Am J Community Psychol 2010 Sep;46(1-2):100-114. [doi: 10.1007/s10464-010-9330-0] [Medline: 20549334]

XSL•FO
RenderX

38. Hackman JR. Leading Teams: Setting the Stage for Great Performances: Harvard Business Press; 2002. URL: https://hbsp.harvard.edu/product/3332-PDF-ENG [accessed 2026-04-07]
39. Ito-Jaeger S, Perez Vallejos E, Logathasan S, Curran T, Crawford P. Young people's trust in cocreated web-based resources to promote mental health literacy: focus group study. JMIR Ment Health 2023 Jan 9;10:e38346. [doi: 10.2196/38346] [Medline: 36622752]
40. Homepage. Wix. 2024. URL: https://www.wix.com/ [accessed 2026-02-12]
41. Sadeh-Sharvit S, Camp TD, Horton SE, et al. Effects of an artificial intelligence platform for behavioral interventions on depression and anxiety symptoms: randomized clinical trial. J Med Internet Res 2023 Jul 10;25:e46781. [doi: 10.2196/46781] [Medline: 37428547]
42. Homepage. YouTube. URL: https://www.youtube.com/ [accessed 2026-02-10]
43. Duffy A, Boroumandzad N, Sherman AL, Christie G, Riadi I, Moreno S. Examining challenges to co-design digital health interventions with end users: systematic review. J Med Internet Res 2025 Mar 14;27:e50178. [doi: 10.2196/50178] [Medline: 40085834]
44. Skivington K, Matthews L, Simpson SA, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061. [doi: 10.1136/bmj.n2061] [Medline: 34593508]
45. Kilfoy A, Hsu TCC, Stockton-Powdrell C, Whelan P, Chu CH, Jibb L. An umbrella review on how digital health intervention co-design is conducted and described. NPJ Digit Med 2024 Dec 23;7(1):374. [doi: 10.1038/s41746-024-01385-1] [Medline: 39715947]
46. Phiri M, Munoriyarwa A. Health chatbots in Africa: scoping review. J Med Internet Res 2023 Jun 14;25:e35573 [FREE Full text] [doi: 10.2196/35573] [Medline: 35584083]
47. Ospina-Pinillos L, Davenport TA, Ricci CS, Milton AC, Scott EM, Hickie IB. Developing a mental health eClinic to improve access to and quality of mental health care for young people: using participatory design as research methodologies. J Med Internet Res 2018 May 28;20(5):e188. [doi: 10.2196/jmir.9716] [Medline: 29807878]
48. Orlowski S, Lawn S, Matthews B, et al. The promise and the reality: a mental health workforce perspective on technology-enhanced youth mental health service delivery. BMC Health Serv Res 2016 Oct 10;16(1):562. [doi: 10.1186/s12913-016-1790-y] [Medline: 27724951]
49. Radovic A, Vona PL, Santostefano AM, Ciaravino S, Miller E, Stein BD. Smartphone applications for mental health. Cyberpsychol Behav Soc Netw 2016 Jul;19(7):465-470. [doi: 10.1089/cyber.2015.0619] [Medline: 27428034]
50. Kreuter MW, Lukwago SN, Bucholtz RDDC, Clark EM, Sanders-Thompson V. Achieving cultural appropriateness in health promotion programs: targeted and tailored approaches. Health Educ Behav 2003 Apr;30(2):133-146. [doi: 10.1177/1090198102251021] [Medline: 12693519]
51. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and opinions of patients about mental health chatbots: scoping review. J Med Internet Res 2021 Jan 13;23(1):e17828. [doi: 10.2196/17828] [Medline: 33439133]
52. Integration of mental health and HIV interventions: key considerations. World Health Organization. 2022. URL: https://iris.who.int/server/api/core/bitstreams/831587a7-3fab-4a70-86c1-6e179a112638/content [accessed 2025-11-12]
53. Salud SE. EVA 10 – ENG Produced as part of the study on the development and implementation of a mental health chatbot for adolescents living with HIV in Peru. 2025. URL: https://www.youtube.com/watch?v=I1qSwZdd7j0 [accessed 2026-04-12]

Abbreviations

AI: artificial intelligence
ART: antiretroviral therapy
EVA: *Educación, Vinculación, y Autoayuda*
HCD: human-centered design
IRB: Institutional Review Board
LMIC: low- and middle-income country
PHQ-9: Patient Health Questionnaire-9
UNAIDS: Joint United Nations Program on HIV/AIDS
WHO: World Health Organization
YAB: Youth Advisory Board

XSL•FO
RenderX

Edited by A Mavragani; submitted 02.Apr.2025; peer-reviewed by A Adeoye, A Adekola, M Al-Mujtaba, R Chigevenga; revised version received 24.Feb.2026; accepted 27.Feb.2026; published 28.Apr.2026.

Please cite as:
Vasquez DH, Rupani N, Contreras C, Kolevic L, Franke MF, Galea JT
Building EVA (Educación, Vinculación, y Autoayuda): Tutorial for the Development of a Digital Mental Health Chatbot for Adolescents Living With HIV
J Med Internet Res 2026;28:e75351
URL: https://www.jmir.org/2026/1/e75351
doi:10.2196/75351

© Diego Humberto Vasquez, Neil Rupani, Carmen Contreras, Lenka Kolevic, Molly Forrest Franke, Jerome Timothy Galea. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Co-Develop-IT! Unifying Methodological Guideline for the Co-Design, Development, and Evaluation of Individually Tailored Technology-Enhanced Training and Rehabilitation Concepts: Consensus Development Study and Tutorial

Patrick Manser[1], Dr Sc; Lotte ES Hardeman[2], MSc; Andreas Wallin[1,3], PhD; David Moulaee Conradsson[1,3], PhD; Breiffni Leavy[1,4], PhD; Franziska Albrecht[1,3], Dr rer nat; Melvyn Roerdink[2,5,6], PhD; Jean-Jacques Temprado[7], PhD; Eling D de Bruin[1,8,9], PhD; Erika Franzén[1,3,4], PhD

[1]Division of Physiotherapy, Department of Neurobiology, Care Sciences, and Society, Karolinska Institutet, Alfred Nobels Allé 23, Huddinge, Sweden

[2]Department of Human Movement Sciences, Faculty of Behavioural and Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[3]Medical Unit Allied Health Professionals, Women's Health and Allied Health Professionals Theme, Karolinska University Hospital, Stockholm, Sweden

[4]Research and Development Unit, Stockholm Sjukhem Foundation, Stockholm, Sweden

[5]Department of Nutrition and Movement Sciences, NUTRIM Institute of Nutrition and Translational Research in Metabolism, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands

[6]Department of Nutrition and Movement Sciences, MHeNs Institute of Mental Health and Neurosciences, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands

[7]Institut des Sciences du Mouvement (ISM), UMR 7287, CNRS, Aix Marseille Université, Marseille, France

[8]Motor Control and Learning Group, Institute for Biomechanics, Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland

[9]Department of Health, OST, Eastern Switzerland University of Applied Sciences, St Gallen, Switzerland

Corresponding Author:
Patrick Manser, Dr Sc
Division of Physiotherapy, Department of Neurobiology, Care Sciences, and Society, Karolinska Institutet, Alfred Nobels Allé 23, Huddinge, Sweden

Abstract

Background: Applying digital health technologies (DHTs) for health promotion and disease prevention is recommended by official bodies such as the World Health Organization. User-centered co-design with systematic patient and public involvement is considered best practice for developing such complex interventions. Although well-established methodological guides and frameworks are available, an important gap is that they are either holistic but generic, offering minimal operational guidance, or context-specific and operational, but focusing only on subphases of establishing DHT-enhanced interventions.

Objective: This paper presents a unifying consensus-based methodological guideline directed toward multidisciplinary expert teams coordinating projects on individually tailored DHTs. It delineates best practices with operational guidance for each step along the full lifecycle of DHT-enhanced training and rehabilitation concepts—from contextualization, through codevelopment, and evaluation to implementation.

Methods: The Co-Develop-IT guideline was cocreated through a structured expert consensus process that integrated, refined, and expanded on well-established existing guides and frameworks to delineate holistic and context-specific, yet flexible enough, best practices. The process consisted of biweekly 90-minute hybrid meetings between August 2024 and February 2025, in combination with written elaboration, feedback, and revisions between meetings to gradually develop a consensus on best practice recommendations.

Results: The Co-Develop-IT guideline consists of 8 iterative phases. It is applicable to any type of end users, exercise types, intended contexts of use (eg, primary health care, community health services, and telemedicine), and overarching goals (eg, health promotion and primary through tertiary disease prevention, including rehabilitation). The Co-Develop-IT guideline introduces 5 distinct preparatory contextual research phases preceding generative codevelopment. These phases are dedicated to the structured establishment of a more robust foundation to better tailor and steer codevelopment efforts toward successful implementation. In 2 application examples, we provide proof of concept that the resulting guideline fulfills its primary purpose of providing comprehensive, context-specific, and operational, yet flexible enough best practice recommendations.

Conclusions: The unifying Co-Develop-IT guideline provides comprehensive best practices with actionable operational guidance for establishing an appropriate balance between scientific theories and frameworks and the real-world needs of interest-holders in the establishment of individually tailored DHT-enhanced training and rehabilitation concepts. Applying Co-Develop-IT

XSL•FO
RenderX

contributes to overcoming the lingering evidence-to-practice gap by consistently establishing a shared mission with relevant interest-holders and ensuring that all codevelopment steps are directed toward addressing an unmet need in (clinical) practice—ultimately promoting the practical application and impact of purpose-developed DHTs.

(J Med Internet Res 2026;28:e84163) doi:10.2196/84163

KEYWORDS

community participation; digital health; disease prevention; exergaming; health promotion; participatory research; patient participation; rehabilitation; user-centered design

Introduction

Innovations in digital health technologies (DHTs)—such as exergames or electronic and mobile health apps—have attracted significant interest as powerful tools to advance health promotion and disease prevention (primary through tertiary prevention; including rehabilitation) [1-4]. On a system's level, DHTs stand out as scalable and cost-effective tools that can contribute to health system improvements via hybrid intervention delivery in a well-controlled environment [2,4-6]. On the individual level, such DHTs have been mainly used to promote engagement and adherence [2,3]. Recent research, however, shows that their benefits go beyond fun and engagement, as they can be designed to provide superior adherence to relevant principles of behavior change, neuroscience, and exercise science compared to conventional exercise and rehabilitation modalities [2].

User-centered methodologies, which continuously involve end users and other multidisciplinary interest-holders in a participatory process, are considered best practice for designing and evaluating complex interventions such as DHT-enhanced interventions. Several guides and frameworks have been published in recent years to methodologically guide these procedures. However, an important gap in the existing guides and frameworks is that they are either holistic but generic, offering minimal operational guidance, or context-specific and operational, but focusing only on subphases of establishing DHT-enhanced interventions.

The UK Medical Research Council's (MRC) guidance for complex interventions [7] is the most frequently used methodological guide for rehabilitation interventions for older adults [8]. It holistically covers intervention development, evaluation, and implementation with core elements to be considered at each phase. These core elements provide valuable guidance for navigating the challenges of the development and evaluation of complex interventions but may lead to difficulties in applying recommendations due to minimal operational guidance [9]. Furthermore, the MRC guidance applies to all complex interventions (ie, all interventions with several interacting components, from public health interventions, over new surgical techniques, to neurorehabilitation) [7]. Consequently, this broad guidance has often been complemented with more specific frameworks to better fit project-specific contexts [8].

In the broad field of electronic health system development, user-centered design [10] has been the most frequently applied framework [11]. More specific guidance for health care contexts is provided by the generative co-design framework for health care innovation [12]. For DHTs specifically, the nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) framework [13], the Multidisciplinary Iterative Design of Exergames (MIDE) framework [14], or its refined methodology from the "Brain-IT" project [15] provide guidance on how to structure the conceptualization, co-design, and evaluation process. These key frameworks represent important steps to advance research methodologies. However, they only cover 1 or multiple subphases relevant to the establishment of DHT-enhanced interventions, leading to a fragmented landscape of frameworks with varying recommendations. Therefore, the research community would benefit from a unifying methodological guideline that delineates more context-specific, yet flexible enough, best practices [8].

Our aim was to cocreate and present a unifying consensus-based methodological guideline directed toward multidisciplinary expert teams coordinating projects on individually tailored DHTs. It delineates best-practice recommendations and operational guidance for each step along the full lifecycle of individually tailored DHT-enhanced training concepts—from contextualization through codevelopment and evaluation to implementation.

Methods

Overview of the Consensus Process

The Co-Develop-IT guideline was cocreated through an expert consensus process. It consisted of biweekly 90-minute hybrid (in-person+videoconference) meetings beginning in August 2024. As per the nominal group technique [16], each consensus group member was asked to critically reflect on and note down their ideas before each meeting for the topic to be discussed. These ideas were shared at the start of each meeting in a brainstorming phase, followed by clarifications of the proposed ideas where necessary [17]. Similar to round 4 of a Delphi approach [18], the consensus group discussed the relevance of proposed items, but with a focus on giving deviating ideas room for justification and discussion to ensure no elements are overlooked. Finally, the consensus group consolidated proposed ideas by cocreating checklist items and discussing the order and overall structure of the best-practice recommendations.

After each meeting, the first author prepared meeting minutes, synthesized the outcomes, and drafted the corresponding sections of the guideline, checklist, and explanation and elaboration sections. All authors provided written elaboration, feedback, and revisions between meetings. Drafts were

XSL•FO
RenderX

version-controlled, and between-meeting changes were ratified at the subsequent session to finalize content.

For each step (steps 1 to 3; "Co-Ccreation of Co-Develop-IT" section), this process continued until (1) saturation (no new items emerging across 2 consecutive meetings) and (2) consensus on a finalized Co-Develop-IT guideline package, which was achieved in February 2025 and is presented in the Results section. Consensus was defined as unanimous agreement among the consensus group or, where this could not be achieved and conflicting recommendations remained, a supermajority decision (threshold of ≥80%) during a cocreation meeting.

Disciplinary and Professional Background of the Consensus Group

We recognize that the Co-Develop-IT guideline is shaped by the authors' disciplinary and professional backgrounds. To provide a nuanced and integrative guideline, the consensus group was carefully selected to represent experts with hands-on experience in applying the methodological guides and frameworks we sought to combine, refine, and expand. The group includes researchers focused on the contextualization and codevelopment of conventional (8 of 10 authors [PM, AW, DMC, BL, MR, JJT, EDB, and EF]) and DHT-enhanced (7 authors [PM, LEHS, DMC, BL, MR, JJT, and EDB]) training and rehabilitation interventions, efficacy or effectiveness (all authors), and mechanistic (6 authors [PM, AW, FA, JJT, EDB, and EF]) evaluations, and implementation in clinical practice (7 authors [PM, LEHS, AW, DMC, BL, MR, and EDB]). In total, 5 authors (PM, DMC, MR, JJT, and EDB) bring experience with public-private partnerships as well as commercialization and related regulatory efforts (eg, medical device and other certifications) relevant to sustainable projects on DHTs. We also bring dual appointments bridging academic research with (1) clinical vocational practice (3 authors [LEHS, BL, and EF]) and (2) leadership roles in technology development companies (2 authors [MR and JJT]). The educational backgrounds span health sciences and technology, physiotherapy, movement sciences, neuroimaging, and neuropsychology.

Cocreation of Co-Develop-IT

Step 1: Identification of Guides and Frameworks to Build On

To ensure that we provide a unifying methodological guideline that appropriately builds on and expands previous research, we started with a critical discussion on which guides and frameworks were useful to guide previous projects [19,20]. Consensus was reached to build on the MIDE framework [14] and its refined methodology from the "Brain-IT" project [15] and unify and expand it with the MRC guidance for complex interventions [7], the generative co-design framework for health care innovation [12], as well as the NASSS [13] and the Practical Planning for Implementation and Scale-Up (PRACTIS) [21] frameworks.

Step 2: Identification of Focus Areas for a Unifying Methodological Guideline

We critically exchanged methodological learnings from our projects. For example, despite Brain-IT's success in maximizing efficacy and user experience [19,22,23], preparatory focus groups for follow-up projects identified gaps in telemedicine workflow implementability that were overlooked during multidisciplinary codevelopment since only design requirements but not implementation requirements were systematically considered. Combined with the identification of limitations of existing guides and frameworks, we defined the following focus areas for the cocreation of our unifying methodological guideline:

- provide clearer recommendations to harmonize the interests and roles of different contributors—particularly across research, industry, health care, and end users;
- clarify the level of involvement of different interest-holders;
- guide additional methodological steps for the validation of DHT components and to facilitate the implementation, scalability, and sustainability of the solutions to be developed; and
- harmonize the interpretation and implementation of the recommended methodological steps and establish a common terminology between different contributors, who often refer to similar issues with different wordings and theoretical references.

Step 3: Cocreation of Unifying Best Practices With Operational Guidance

Cocreation of the Co-Develop-IT guideline package was achieved by focusing our discussions and the consensus-finding process on how well-established guides and frameworks can be integrated, revised, and, most importantly, be expanded to address the identified focus areas for improvement. The process followed the chronological nature of the relevant phases for the establishment of DHT-enhanced training or rehabilitation interventions, but went through several iterative loops to refine best practice recommendations in each phase until consensus was established on a finalized Co-Develop-IT guideline package as per the general consensus handling process defined in the Overview of the Consensus Process section.

Proof of Concept for Co-Develop-IT

To verify whether the resulting guideline fulfills its primary purpose of providing comprehensive, context-specific, and operational, yet flexible enough best practice recommendations, it was applied to guide the methodological development of 2 ongoing projects—"Park-MOVE" (Karolinska Institutet, Sweden) and "Better Together" (Vrije Universiteit Amsterdam, The Netherlands). The application examples are provided to complement the Co-Develop-IT guideline package for additional operational guidance with project-specific examples.

Ethical Considerations

Given the cocreation of this unifying methodological guideline was done within the author team and we completed no experiments involving human subjects as participants, no ethics approval was necessary. Ethics approval for studies as part of

XSL•FO
RenderX

the "Park-MOVE" and "Better Together" application examples was provided by the Swedish Ethical Review Authority (2024-06688-01) and the Research Ethics Committees United, The Netherlands (R22.076, NL82441.100.22), respectively.

Results

Overview

The *Co-Develop-IT guideline package* consists of the following complementary elements:

1. Co-Develop-IT core guideline (see "The Co-Develop-IT Guideline" section),
2. explanation and elaboration sections of the Co-Develop-IT core guideline (see "Explanation and Elaboration of the Co-Develop-IT Guideline"),
3. Co-Develop-IT checklist (Supplementary File 1 in Multimedia Appendix 1), and
4. explanations and elaborations of each checklist item with guiding examples (Supplementary File 2 in Multimedia Appendix 1).

For Which Contexts Are the Co-Develop-IT Guideline Intended?

The Co-Develop-IT guideline was developed for the context of individually tailored DHT-enhanced training and rehabilitation concepts that are purposefully codeveloped with adequate theoretical underpinnings.

With "DHT-enhanced training or rehabilitation concepts," we refer to:

- a conceptual guideline that delineates all algorithmic decision trees regarding the structure, content, tailoring, delivery, supervision, and monitoring of a DHT-enhanced exercise, training, or rehabilitation program (in adherence with the definitions from Manser et al [2] and Herold et al [24]), combined with:
- a purpose-developed DHT to implement and deliver the exercises, training, or rehabilitation program according to this concept.

For the sake of terminological simplicity, we will refer to "DHT-enhanced training concepts" from here on out.

With "individually tailored" training, we refer to integrating both "personalization" and "individualized progression," understood as follows in the context of this guideline:

- Personalization refers to the systematic selection of the type, content, components, and characteristics of a training or rehabilitation program for a specific person based on their (clinical) characteristics, performance, or preferences at baseline.
- Individualized progression refers to the gradual and systematic adaptation of training dose and stimuli to an individual in response to their training or rehabilitation participation to maintain a training load deemed sufficient to induce targeted health- or disease-related adaptations and/or to promote user experience, engagement, and adherence.

Who Is the Co-Develop-IT Guideline Intended For?

The Co-Develop-IT guideline is directed toward multidisciplinary expert teams coordinating projects that involve the contextualization, codevelopment, evaluation, and implementation of individually tailored DHTs and DHT-enhanced training concepts. It is applicable for any types of end users, exercise types, intended contexts of use (eg, primary health care, community health services, and digital health), and overarching goals (eg, health promotion and disease prevention [primary through tertiary prevention; including rehabilitation]).

How Is the Co-Develop-IT Guideline Intended to Be Used?

To facilitate and support the effective implementation of the conceptually discussed Co-Develop-IT core guideline, we provide the Co-Develop-IT checklist (Supplementary File 1 in Multimedia Appendix 1) that details the operational steps to be performed and documented in projects following this guideline. We recommend that the checklist along with its item-specific explanation and elaboration sections (Supplementary File 2 in Multimedia Appendix 1) be considered the core instruments to work with when implementing the guideline.

What Are Key Elements of the Co-Develop-IT Checklist?

Apart from step-by-step instructions for the procedures to be completed and documented when adhering to the Co-Develop-IT checklist, it defines the following "contextual requirements" for each item: (1) the minimum level of interest-holder involvement according to the participation choice points defined by Vaughn and Jacquez [25], (2) which key interest-holders should be responsible for each item, and (3) item relevance when adhering and reporting a project according to the Co-Develop-IT guideline.

How Does the Co-Develop-IT Guideline Balance Rigor With Practical Flexibility?

The guideline's cocreation was centered around finding a good balance between methodological rigor and practical feasibility, which was confirmed successful in 2 application examples. Accordingly, "recommended" and "beneficial" items are provided in the Co-Develop-IT checklist. While every project would benefit from adhering to all items, we acknowledge that, depending on the resources available, projects might not be able to do so. Therefore, we offer flexibility by the opportunity to choose a nuanced approach between the "beneficial" best practice approach and a short-track option that only adheres to the "recommended" items.

The Co-Develop-IT Guideline

The Co-Develop-IT guideline is structured in 8 phases. An overview of its key structures and iterative steps that systematically guide projects toward making a real-world impact is illustrated in Figure 1.

XSL•FO
RenderX

Figure 1. The Co-Develop-IT guideline. This figure illustrates an overview of the key components and phases of the Co-Develop-IT guideline, which systematically guides multidisciplinary expert teams coordinating projects on individually tailored technology-enhanced training and rehabilitation concepts and related digital health technologies toward making a real-world impact. It is applicable to any type of end users, exercise types, intended contexts of use (eg, primary health care, community health services, and telemedicine), and overarching goals (eg, health promotion and primary through tertiary disease prevention; including rehabilitation). UX: user experience.

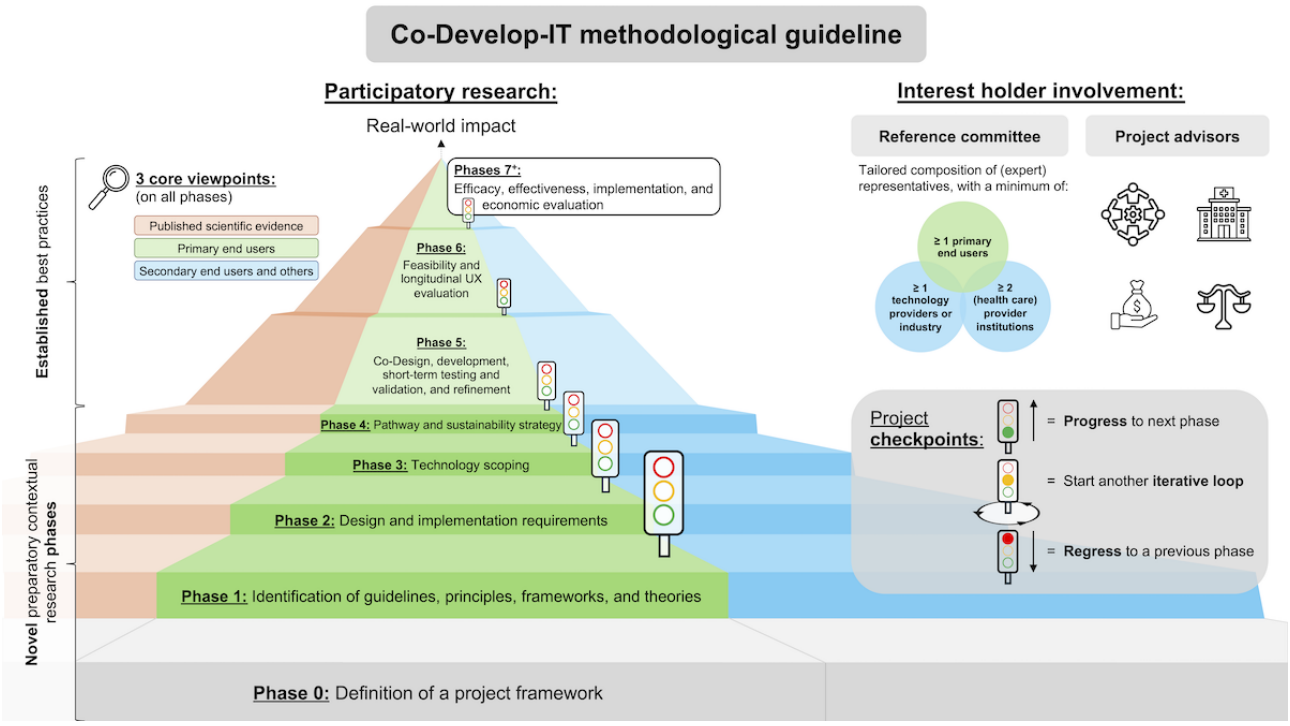

Importantly, while both the guideline and each of the phases that build on each other may be perceived as presented rather linearly, they represent iterative project steps, meaning that a project moves up and down the phases—which may each involve multiple iterative loops—depending on goal achievement and project progress. A core component of the guideline is the establishment of project checkpoints with a traffic light system–based assessment framework with clear benchmarks and progression criteria to adequately guide project progression, as well as the number of required iterations within and across phases. Thereby, a systematic and transparent approach to assess goal achievement is established, leading to iteratively looping a phase (=amber light) until the criteria to progress to the next phase (=green light) or regress to a previous phase to establish a more robust foundation (=red light) are met. Of course, this is also applicable to subprojects.

We acknowledge that details of specific steps and their order within a phase might not be applicable to every subproject and might run to a certain extent in parallel. If necessary, details of specific steps or their order may be adapted. However, the order of completing each phase must be followed, as each phase builds on the findings of the previous phases.

Of note, all findings derived from projects adhering to the Co-Develop-IT guideline should be reported following the appropriate checklists provided on the EQUATOR (Enhancing the Quality and Transparency of Health Research) network [26] to ensure reproducibility, transparency, and comparability to other research.

Explanation and Elaboration of the Co-Develop-IT Guideline

Phase 0: Definition of Project Framework

Overview

The initial phase of the process of applying the Co-Develop-IT guideline is dedicated (1) to defining nonnegotiable key elements of the project (ie, phase 0.1); (2) to defining requirements on interest-holder involvement, resources, and the regulatory environment for the project (ie, phase 0.2); (3) to establishing a project consortium including all relevant interest-holders and defining their roles and responsibilities throughout the project for effective collaboration and accountability (ie, phase 0.3); and (4) to collaboratively agreeing on specific project checkpoints and a preliminary project time plan (ie, phase 0.4).

These steps are to establish a collaboratively agreed-upon framework within which the project should take place.

Phase 0.1: Define Overall Context and Goals of the Project

Initiate the project by outlining its:

- overarching goals and broader intended context (Co-Develop-IT checklist -item 3; eg, which structures or functions the training should impact and the type and multidisciplinary expertise of involved institutions),
- target populations (-item 4; eg, individuals at risk for or living with a specific disease or condition as primary end users, along with health care professionals and relatives or supporters as secondary end users),

XSL•FO
RenderX

- intervention type (-item 5; eg, DHT-enhanced multidomain [physical, motor, and cognitive] training), and
- targeted outcome domain (-item 6; eg, health-related quality of life).

Outlining these aspects allows delineating a clear path to follow, sets clear boundaries for the freedom of the subsequent codevelopment process toward fulfilling a project's larger purpose, and guides the identification of relevant interest-holders to be recruited (phase 0.2). Therefore, make sure these definitions are backed on a clearly defined theoretical rationale for how the development of an individually tailored DHT-enhanced training concept can address a specific problem or knowledge gap (-items 1 and 2). Leave sufficient room for innovations and novel ideas that may arise during the codevelopment process. To do so, no specifics such as exercise or training variables, specific DHTs to be used to implement the training, or any design elements of the DHTs and the training are to be defined at this stage. Once a project consortium is established (phase 0.4), these initial definitions are to be refined and agreed on collaboratively with all relevant interest-holders (-item 14)—thereby providing a structured approach to ensure that the overall project context and goals are backed by both a theoretical rationale and real-world end user and public needs.

Phase 0.2: Define Requirements on Interest-Holder Involvement, Resources, and Regulatory Environment for the Project

In the second subphase, identify:

- contextual requirements and required multidisciplinary interest-holders (-item 7),
- required resources for each interest-holder (-item 8; eg, time, staff, funding, expertise, facilities, and data resources), and
- regulatory requirements (-item 9) to successfully complete the project (eg, institutional, ethics, and good clinical practice regulations and requirements for collaboration agreements).

Table 1 provides a list of potential interest-holders to be considered to ensure that the project is grounded in a realistic and competent environment.

Table . Definitions and key role of multidisciplinary interest-holders to be considered to be involved in research projects in the co-design, development, and evaluation of individually tailored digital health technology (DHT)–enhanced training concepts.

Interest-holder	Definition and role
Research institutions	Academic or research institutions that provide the scientific and methodological expertise necessary for the project.
Health care or other provider institutions	Health care facilities or other provider organizations that facilitate access to the target population, provide clinical insights and support, and implement the resulting DHT-enhanced training concept.
Technology developers	Companies or teams responsible for the development of DHTs (ie, software, hardware, or both).
Technology providers	Companies or teams responsible for supplying the technological infrastructure (software, hardware, computing power, etc) and providing support for the implementation of the DHT-enhanced training concept (eg, installation of devices and ongoing technological maintenance and support for DHTs).
Legal offices	Legal experts who advise the core project team to ensure that all aspects of the project comply with relevant regulations and ethical standards and support and guide efforts toward getting the final DHT-enhanced training concept patented, certified as a medical device, and commercialized.
Additional interest-holders	Depending on the project's scope, additional interest-holders might include patient advocacy groups, funding bodies, policymakers, and end-user representatives.

Moreover, include legal experts at this preparatory stage to outline the steps required to patent the final training concept or obtain medical device certification for the DHTs (ie, software and hardware). This will ensure that all relevant aspects are considered during the project. This may include adhering to standards set by regulatory bodies such as the Food and Drug Administration in the United States or the European Medicines Agency in Europe. The recently published Common European Classification Grid for digital medical devices is helpful to facilitate the definition of the taxonomy and evidence requirements for assessment of DHTs and offers the possibility to have a common reference on a European level against which national classifications and evidence requirements can be mapped [27]. Key considerations include device classification, safety and efficacy testing, cost-effectiveness evaluation, data security, labeling, and postmarket surveillance.

Phase 0.3: Establish a Project Consortium and Define Each Interest-Holder's Roles Throughout the Project

In the third subphase, recruit interest-holders to establish a project consortium based on the drafted project goals and context from subphase 0.1 and the requirements for a complete project consortium defined in phase 0.2. Ensure that at least 1

XSL•FO
RenderX

interest-holder representing each target population (-item 4) and multidisciplinary context (-item 7) is involved.

Establishing a project consortium entails setting up a project reference committee (-item 10b; recommended item) and including project advisors (-item 10c; beneficial item) to continuously guide the core project team (-item 10a; recommended item). In the case of multinational projects, establish a project reference team and recruit project advisors for each country involved to ensure that local and sociocultural nuances are appropriately considered. Tailor the project consortium's constellation to the specific research project, adhering to the minimum key interest-holders and responsibilities listed in Table 2. Of note, the involvement of members of the project consortia might change throughout the conduct of a project, depending on the project-specific requirements, needs, or relevance in different phases.

XSL•FO
RenderX

Table . Recommendations on key interest-holders and responsibilities of core groups to be involved in research projects following the Co-Develop-IT guideline.

Core groups	Recommended members	Responsibilities	Level of involvement[a]
Core project team	The core project team typically consists of the team who initiated the project. As a minimum requirement, this project team includes: • the project coordinators or the local principal investigators of the project and their research teams (-item 7a), • interest-holders from the health care or other provider institutions (-item 7b), and • if the technology development is not done within the research teams the technology developers (-item 7c).	• To coordinate and monitor the successful conduct of the project.	• Minimum condition: "involving" level • Ideal condition: "collaborating" or "empowering" level
Project reference committees	While the constellation of each of the project reference committees should be tailored to the specific research project, we recommend that the project reference committees should include: • ≥1 (expert) representative of the primary end users (eg, patient advocacy groups or end-user representatives; -item 7f), • ≥1 (expert) representative of the technology providers or industry (-item 7c), and • ≥2 (expert) representatives of the health care or other provider institutions (-item 7b): 1 on the organizational level (-item 4e) and 1 on the individual level (-item 4c).	At the beginning of the project: • To collaboratively work out and define specific subgoals and progression criteria (-item 15) with the project core team. • To advise the project core team when working out a preliminary time plan (-item 16). In later stages of the project: • To guide important decisions on the project plan and execution. • To predetermine, together with the core project team, the weighting of how input from different interest-holders should be prioritized when making design decisions in phase 5 (-item 28). • To determine, together with the core project team, how to prioritize and weight the contribution between the 3 core viewpoints when making specific design decisions in the case of potential incongruent findings or conflicting results (eg, phase 2 [-items 19 - 23] or phase 5 [-item 34]).	• Minimum condition: "involving" level • Ideal condition: "collaborating" or "empowering" level

XSL•FO
RenderX

Core groups	Recommended members	Responsibilities	Level of involvement[a]
Project advisors	The project advisors may include representatives for each of all remaining defined levels under -items 4 and 10 and may primarily consist of: • communities and systems in which the training concept should be implemented (-item 4e), • legal offices (-item 7c), and • additional interest-holders, such as funding bodies or policymakers (-item 7d) depending on the project-specific requirements and as defined under -item 7.	To provide nonbinding strategic advice to the core project team on: • the project plan and execution, including regulatory and ethical compliance, • how to deal with potential incongruent findings or conflicting results between the 3 core viewpoints, • to take and refine measures to comply with recommendations on diversity and inclusion, anticipation and reflection, openness and transparency, responsiveness, and adaptive change, and • to support commercialization, patenting, or medical device certification of the DHT[b]-enhanced training concept.	• Minimum condition: "consulting" role on the individual level

[a]Level of participatory research involvement according to the participation choice points outlined by Vaughn and Jacquez [25].

[b]DHT: digital health technology.

To ensure project feasibility and mitigate potential risks of noncompletion, comprehensive resource mapping and comparing the available resources for all interest-holders (-item 11) with the required resources (-item 8) helps to identify any shortcomings to be addressed to support smooth project execution and completion (-item 12). Critically reflect on any relevant additional project-specific resources and risks to ensure thorough risk mitigation.

Establish formal collaboration agreements that delineate the responsibilities, rights, obligations, and degree of involvement of each collaborator (-item 13) to facilitate the management of expectations and ensure accountability. Regularly revisit and update the collaboration agreements as needed to reflect any changes in the project scope or interest-holder roles.

Phase 0.4: Collaboratively Agree on Specific Subgoals and Progression Criteria as well as a Preliminary Time Plan for the Project

Finally, collaboratively refine and agree on the specific goals and subgoals of the project (-items 14 and 15a; as drafted in phase 0.1) and define project checkpoints (-item 15) that are aligned with the overall goals of the project and its phases—both based on a consensus process with all interest-holders of the project consortium (ie, core project team and reference committee and [optionally] project advisors). Each project checkpoint requires a traffic light system–based assessment framework with clear benchmarks (-item 15c) and progression criteria (- item 15d) to ensure transparency in the assessment of agreed-upon goals (-item 15b). The project checkpoints are usually to be positioned toward the end of each phase to inform whether the project can progress to the next phase (=green light), requires further iterative loops with refinements before proceeding to the next phase (=amber light), or requires the establishment of a more robust foundation by regressing to a previous phase (=red light).

Descriptive examples are provided in the "Phase 5.6: Iterative Refinements" section, Supplementary File 2 in Multimedia Appendix 1, and in previous literature of the author team in relation to feasibility and user experience testing in preparation for a randomized controlled trial [23,28]. For transparency in the interpretation of project results and progression, we recommended that the traffic light system–based assessment framework be published (as part of a project protocol) or at a minimum publicly preregistered (eg, on the Open Science Framework) prior to the start of data collection.

Collaboratively agreeing on these goals and checkpoints establishes a unified direction and commitment to the project's success and helps prevent the emergence of misunderstandings or conflicts at a later stage of the project. To support these efforts, adopt techniques such as empathetic target group analysis [29] to foster a better understanding of the diverse needs and perspectives among interest-holders before starting the co-design or cocreation process [30,31]. Moreover, have the added value and expectations of each partner in collaboratively achieving these goals communicated to transparently foster fruitful collaboration.

Finally, a major pitfall in DHT-related projects with public-private partnerships is the potential misalignment in expected rates of progress and output per unit of time in public-private partnerships, given that public research institutions may be perceived by the industry as slow-moving "oil tankers", whereas the industry may be seen by academia as agile "speedboats" rushing toward marketable outputs. Therefore, we recommend collaboratively developing and agreeing on a preliminary project time plan (-item 16) to

XSL•FO
RenderX

align interest-holder expectations and update this time plan regularly to harmonize expectations between partners.

Phase 1: Identification of Guidelines, Principles, Frameworks, and Theories

Phase 1 is dedicated to identifying guidelines and evidence-based recommendations for the overall goals of the project (-item 17) along with principles, frameworks, or theories (-item 18). This phase is key to defining a robust backbone and guiding future steps of projects, particularly decisions on the design, characteristics, and content of the DHT-enhanced training concept to be developed. A few broadly applicable examples that the consensus group recommends following are included in Textbox 1.

Textbox 1. Examples of broadly applicable guidelines, principles, frameworks, and theories the consensus group recommends to follow.

Guidelines and evidence-based recommendations

- Clinical or best practice guidelines for a specific target population
- Consensus statements, such as the global consensus for optimal exercise recommendations for enhancing healthy longevity in older adults [32]

Principles

- General training principles [24,33]
- Neuroplasticity principles [34,35]
- Principles for neurorehabilitation (ie, that integrate motor learning and brain plasticity mechanisms) [36]

Frameworks

- A behavior change framework applicable to physical activity or training [37]
- Guided plasticity facilitation framework [38] for motor-cognitive training
- The adaptive capacity model [39]
- The Beyond “Just” Fun of Exergames framework [2]

Theories

- Theory of effort minimization in physical activity [40,41]
- Optimizing performance through intrinsic motivation and attention for learning theory of motor learning [42]
- A behavior change theory supporting initiation and consolidation of physical activity behavior changes [43]

Phase 2: Determine Design and Implementation Requirements

The overarching aim of phase 2 is to delineate a comprehensive set of design and implementation requirements for the DHT-enhanced training concept. This encompasses a range of considerations to be adhered to throughout the subsequent codevelopment phases, including:

- generating user models (-item 19),
- defining requirements for core components of individually tailored training concepts (-item 20),
- defining environmental requirements (eg, training equipment, space requirements, and connectivity such as Wi-Fi or Bluetooth; -item 21), and
- defining hardware (-item 22) and software (-item 23) requirements for the DHTs.

Elaboration of the design and implementation requirements should be built on phase 1 by integrating identified guidelines and evidence-based recommendations (-item 17) and taking into account relevant principles, frameworks, and theories (item 18). Consider the set of core questions provided in Supplementary File 5 in Multimedia Appendix 1 along with the checklist facilitating DHT adoption provided by Hamasaki et al [44] and a comprehensive (early) health technology assessment (eg, see exemplary question in Table 2 of Grutters et al [45]) to explicitly evaluate the potential value of a health technology [45] to ensure that all potentially relevant aspects are considered. All decisions are to be made by integrating the findings from three core viewpoints: (1) a synthesis of published scientific evidence, together with the findings from performing qualitative research on the perspectives of (2) the intended primary end users, and (3) the intended secondary end users and other relevant interest-holders.

The current state of the scientific evidence should be derived from conducting an umbrella review, meta-analysis, systematic, scoping, or narrative review based on the established levels of evidence [46]. If such high-level evidence is available in recent scientific literature, a narrative synthesis of the current state of evidence suffices. The perspectives of intended primary and secondary end users and all other relevant interest-holders should be derived from conducting qualitative research (based on, eg, semistructured interviews, focus groups, or a [modified] Delphi approach [18]). Refer to -item 29 on how to set the stage to facilitate reflective dialog and productive workshops

XSL•FO
RenderX

and apply these recommendations also to the qualitative research components in this phase (-items 19 - 23).

At this stage of a project, we recommend conducting qualitative research with primary end users separately from the remaining interest-holders. While all interest-holders are later on recommended to codevelop solutions together, this structured approach builds the foundation for productive codevelopment. Specifically, it ensures that primary end users—who might be hesitant to share their insight in the presence of field experts on DHTs or exercise and rehabilitation approaches—get their own platform to express all their insights fully and in the absence of potential power dynamics that might limit their participation [47]. By aligning data collection on design and implementation requirements to reflect the 3 core viewpoints from which insights are to be combined ensures that no important insights are missed, particularly regarding barriers to and facilitators of sustained adoption, including considerations for training and associated learning curve when using DHTs and establishing a supportive organizational culture [48,49]. Furthermore, it fosters a better understanding of the diverse needs and perspectives among different interest-holder groups and facilitates that all those perspectives are presented to empower all codevelopment participants in phase 5 to contribute fully [30,31].

To provide a more nuanced understanding of how inclusivity and accessibility can be ensured for all relevant intended end users, consider examining the current state of evidence and interview or focus group transcripts through the lens of critical discourse analysis [50]. These insights should be aligned with broader frameworks for sustainable innovation and ethical design in digital health [51].

Based on the project checkpoints, phase 2 evaluations should be repeated iteratively (=amber light; eg, due to insufficient coverage of relevant themes as per feedback from the project reference committee) until the green light criteria for project progression are met, at which point, the project can progress to phase 3. If red light criteria are met (eg, due to a strong mismatch between defined project goals [-item 14] and "real-world" needs as per findings from the qualitative studies), it is suggested that the project requires a more robust foundation to be established by returning to phase 2, as per the progression criteria defined in phase 0 (-item 15).

Phase 3: Technology Scoping

The third phase focuses on identifying (-item 24a) and critically appraising (-item 24b) existing (partial) DHT solutions in both research and the market that align with the project's overall goals (phase 0). Apply techniques such as trend analysis and the Strengths, Weaknesses, Opportunities, and Threats matrix [14] to gain a comprehensive understanding of the current state of knowledge and to identify areas for further improvement. Moreover, assess each identified solution against the design and implementation requirements defined in phase 2 (-item 24c). Understanding the strengths, limitations, and areas for improvement of current solutions in relation to the design and implementation requirements is essential for building a robust foundation for advancing the field with credible, innovative, and evidence-based DHT-enhanced training concepts. Finally, identify and analyze emerging trends (-item 24d) to ensure that the project identifies and integrates relevant technological advancements that may be relevant to better fulfill the design and implementation requirements. This is especially important for rapidly evolving technological opportunities, such as augmented or mixed reality, artificial intelligence, bio- or neurofeedback, and brain-computer interfaces [2].

Phase 4: Define Pathway and Sustainability Strategy

The aim of phase 4 is to make a strategic decision on the project's development path (-item 25) and establish a sustainability strategy for the DHT-enhanced training concepts to be (further) developed (-item 26). Specifically, on the basis of the findings of phases 1 - 3, determine whether the project strives to develop a novel DHT-enhanced training concept from scratch (path 1) or further develops and builds on existing solutions (path 2).

A purpose-developed software builds the heart of every DHT, whereas the conceptual decisions and algorithmic decision trees provided in a training concept build the heart of a DHT-enhanced training concept [2]. Therefore, to maximize scalability and transferability to other application scenarios or use cases, focus on developing training concepts and software to implement the training while relying on (1) well-established, off-the-shelf hardware and (2) ensuring the software is universally applicable with different hardware peripherals. This adheres to the "training (instead of product) first" approach [52] and facilitates the sustainability strategy, as it reduces the complexity of providing necessary equipment and improves scalability.

In this regard, under -item 26, a sustainability strategy is to be developed to ensure that the solution (DHT-enhanced training concept) to be (further) developed will be made available to the intended target populations and remain available after the project's completion. This strategy should address its long-term availability, which could be provision and maintenance of open-access to software, developing a business plan and outlining a plan to commercialize the solution, or transferring intellectual property rights to an existing company. Depending on this, mechanisms for ongoing maintenance and support, scalability to reach a broader audience or adapt to different contexts, and identifying resources to support the long-term sustainability of solutions are to be worked out. This step is crucial to ensure that the project's outcomes have a lasting impact and continue to benefit the target populations in the long term.

With these extensive preparatory contextual research steps from phases 1 to 4, a robust conceptual foundation for the targeted codevelopment toward successful implementation phases of the project is laid.

XSL•FO
RenderX

Phase 5: Co-Design, Development, Short-Term Testing and Validation, and Refinement

Overview

Phase 5 consists of multiple iterative cycles of (1) co-designing and developing prototypes of the DHT-enhanced training concept (phases 5.2 to 5.4), (2) short-term user experience and safety testing and validation of components of the DHT-enhanced training concept (phase 5.5), and (3) refining these prototypes by further co-design and development (phase 5.6).

This iterative cycle is to be repeated until an "acceptable" solution (=all components of the DHT-enhanced training concepts have been successfully validated, have a good user experience, and are safe) is achieved. The result of this phase is an "original" DHT-enhanced training concept that enters the next phases of longitudinal evaluations.

Phase 5.1: Build a Framework for the Co-Design

First, identify congruent aspects from previous project phases (-item 27a) and agree on which of the remaining aspects for the DHT-enhanced training concepts require co-design procedures (-item 27b). Subsequently, work out and define, together with the project reference committees, the weighting of how inputs from different interest-holders are to be prioritized when collaboratively making design decisions (-item 28). For example, it might be agreed that design decisions on the graphical user interface should primarily be driven by primary and secondary end users, whereas algorithms for individualized tailoring of the training should mainly rely on scientific theories with input from researchers and health care professionals. Clear definitions of each interest-holder's role and expected contributions help streamline and guide the complex task of integrating various, potentially conflicting perspectives and viewpoints from different interest-holders—increasing the likelihood of successfully generating conceptual prototypes that fulfill all the design and implementation requirements defined in phase 2 of the project and all involved contributors can agree on.

Phase 5.2: Co-Design Workshops

The diversity in background, education levels, and experiences of interest-holders can be expected to lead to power dynamics that influence or limit the interactions during the co-design process [47]. Therefore, a stepwise process is recommended to set the stage for a well-functioning collaborative effort.

Specifically, start the co-design workshops with a short presentation to share the rationale and overall goals of the project (-item 29a) and preparatory findings from earlier phases of the project (-item 29b), and clarify the planned procedures (-item 29c) and expectations (-item 29d) with all co-design workshop participants. With respect to the latter, transparently share the weighting and prioritization of input from different interest-holders (-item 28) so that all participants have a clear understanding of their expected contributions while mitigating potential power imbalances. Finally, invest time in establishing mutual trust, empathy, and comfort with all participants (-item 29e). This step is critical to foster a better understanding of the diverse needs and perspectives among interest-holders and facilitate reflective dialogue before starting the co-design or cocreation process [30,31]. Use techniques such as empathetic target group analysis [29] to achieve this aim.

If the project chooses path 2—to further develop existing DHT-enhanced training concepts (-item 27)—present these solutions to all workshop participants to provide them with a clear understanding of what the project builds on and what preparatory steps led to the chosen starting point (-item 30). The participants should also be given the opportunity to try these existing DHTs to help them identify opportunities for improvements that can be addressed in generative co-design workshops.

Finally, the "generative" part of co-design is performed (-item 31). During the co-design process, use techniques such as 6-3-5 brainwriting [53] to ensure that all interest-holders can express their ideas without judgment [30,31]. To facilitate collaborative idea generation, use different paper- and pencil-based, embodied, or technology-supported prototyping techniques. Simple methods may include, but are not limited to, empathy mapping [29,54], user journey mapping [55], bodystorming [56,57], collaborative sketching and drawings for the rapid development of initial concepts and designs [14], or—when building on and further developing existing solutions—the cognitive walkthrough method [58]. Technology-enhanced prototyping may include the use of middleware interfacing software tools that allow for the rapid development of initial concepts and designs to the creation of prototypes of the proposed concepts (eg, game mechanics and theme) in game engines. Such tools allow early identification of possible modifications that must be performed to facilitate user interaction (eg, removing the need to press buttons) and are particularly useful for the development of gamified DHTs [14].

Phases 5.3 and 5.4: Data Synthesis and Development

Analyze data generated during the "generative" part of co-design, such as audio (and video) recordings combined with generated sketches and drawings or virtual conceptual prototypes, using qualitative research methodologies (-item 32), such as qualitative content analysis [59-61]. Use the frameworks defined in phase 1 to guide the interpretation of findings. For example, the Beyond "Just" Fun of Exergames framework could be useful in this stage to ensure appropriate theoretical backing of DHT designs to relevant principles in motor-cognitive learning, neurorehabilitation, and behavior change [2].

In the next step, present the results from each workshop to the project reference committee to collaboratively integrate the findings and rank-order complementary solutions. It is likely that there are incongruent findings or conflicting results between different workshops and complementary solutions for the same aspects of the DHT-enhanced training concepts from different

XSL•FO
RenderX

workshops. In this case, present these results to the project reference committees and collaboratively make consensus-guided decisions on how to integrate the findings into decisions on the design of the DHT-enhanced training concepts. Ideally, consider the advice of project advisors as well.

Based on the derived rank-ordering (-item 33) and considering resource availability (-item 11), derive a list of DHT development tasks (-item 34) to establish testable prototypes of the DHT-enhanced training concept (-item 35).

Phase 5.5: Short-Term Testing and Validation

The first prototypic components of the DHT-enhanced training concepts are then thoroughly tested on safety and user experience (-item 36), and relevant components of the DHT-enhanced training concept are validated (-item 37).

User experience refers to a complex characteristic that results from the perception of many distinct quality aspects of a product [62]. The relevance of these quality aspects varies between different contexts and products under evaluation [63-65]. It has been defined as a person's perceptions and responses resulting from the use and anticipated use of a product, system, or service [66]. For a comprehensive user experience assessment, combine the advantages of both quantitative and qualitative research methodologies. Consider Campbell's [67] step-by-step instructions and best practices.

For quantitative user experience evaluations, research has often relied on questionnaires for specific subconstructs of user experience, such as usability, which is a prominently evaluated subconstruct in the field of DHTs [65]. For a more comprehensive user experience assessment, we recommend the validated modular framework introduced by Schrepp and Thomaschewski [62]. This framework contains several scales that measure different user experience aspects and allows the construction of customized questionnaires tailored to the specific research question and context of use in a uniform format. While the shortcoming of such a modular questionnaire is the lack of standardized benchmarks, its application is mainly recommended when multiple measurements of the same product are compared over time [62], matching the iterative process proposed by the Co-Develop-IT guideline. This scale may be complemented with other—more specific—scales for certain new use cases and product types to provide a more comprehensive assessment (eg, see item-specific explanation and elaboration statement) [62]. For qualitative user experience evaluations, obtain data complementary to quantitative evaluations to inform further developments and refinements. Specifically, collect specific suggestions for improvements in the components of the DHT-enhanced training concepts to optimize the user experience via semistructured interviews with primary and secondary end users.

Assess the safety of components of the DHT-enhanced training concepts and interaction with DHTs in relation to the characteristics and capabilities of the targeted end users. Consider the perspectives of both primary and secondary end users. We recommend having trained professionals closely observing play-test sessions to assess safety (eg, risk of falls or injuries) along with qualitative research with primary and secondary end users to assess perceived safety and collect suggestions for improvements.

Thoroughly validate the relevant components of the DHT-enhanced training concepts according to the requirements defined under -item 15b. This may include but is not limited to software algorithms for player movement detection [68], (game) performance metrics [69-73], (gamified) assessments [74], or mechanisms for individualized tailoring [69].

At the end of this subphase, check whether the quality criteria defined under 15c are reached (-item 38), share the findings with all co-design workshop participants (-item 39), and progress according to the mechanism outlined under -item 15d.

Phase 5.6: Iterative Refinements

Iteratively repeat subphases 5.1 to 5.5 until an "acceptable" solution is achieved (-item 40) according to the specific subgoals and quality criteria defined under -items 15a and 15c. For example, the criteria for an "acceptable" (=green light) solution could be (1) session adherence and compliance ≥70%, (2) mean user experience score ≥6 [75], (3) mean system usability score ≥70% [76], (4) mean exergame enjoyment score ≥75% [77], and (5) ≥70% of (a) primary end users (eg, individuals with mild neurocognitive disorder) and (b) health care professionals perceive the DHT-enhanced training program as (i) accessible, (ii) useful, and (iii) are willing to use it in practice.

Details on recommended assessments and more elaborate examples with cutoffs for a green, amber, and red checkpoint light are provided under -item 15 and its corresponding explanation and elaboration section. Once a green light has been achieved, progress with longitudinal evaluations. Otherwise, reiterate this phase (in case of an amber light) or return to a previous phase (in case of a red light) to establish a more robust foundation for co-design as per your defined progression criteria (-item 15).

Phase 6: Feasibility and Longitudinal User Experience Evaluation

For the phases introduced thus far, the Co-Develop-IT checklist provides best practice recommendations that are to a large extent novel because they expand existing methodological guides and frameworks with context-specific guidance for DHT-enhanced training. Subsequent phases 6 and 7+, covering longitudinal evaluations and implementation, are more generally applicable and well covered by established guidance. Therefore, we provide (1) recommendations which established procedures to follow and (2) which additional elements to consider to make these guides and frameworks better tailored to DHT-enhanced training.

XSL•FO
RenderX

In phase 6, test the feasibility of the full DHT-enhanced training concept and the study procedures for subsequent full-scale trials, along with a more in-depth investigation of the user experience and safety of the full DHT-enhanced training concept (-items 41 and 42). This phase is again iterative, meaning that it is repeated according to prespecified progression criteria until a solution with acceptable feasibility and user experience among primary and secondary end users is achieved (-item 43).

Evaluate the feasibility of the study procedures and the DHT-enhanced training along with its user experience. We recommend doing these evaluations on the basis of the conceptual framework of Eldridge et al [78], following the terminology and recommendations of the MRC guidance [7], and considering mixed methods evaluations of both feasibility and user experience following O'Cathain et al [79] and Campbell [67]. Like in phase 5, collect complementary qualitative data that provide broader contextual information and obtain data on specific suggestions for improving the feasibility and user experience from both primary and secondary end users. Extend the abovementioned methodological guidance by the following items (-item 42):

First, a major pitfall in feasibility research is the overreliance on narrative interpretations, which may obscure unfavorable outcomes to prematurely progress to efficacy evaluations. To achieve a more transparent, objective, and reproducible approach, work with the project reference committee to establish a traffic light system–based assessment framework with predetermined progression criteria (-item 42a). This step should be finalized and publicly documented prior to data collection, building upon the foundational requirements detailed in -item 15.

Second, in agreement with the MIDE framework [14], systematically evaluate and report on technology performance (-item 42b; eg, downtime, type and frequency of occurrence of technical problems, and effectiveness of supportive strategies for end users) to provide a robust basis for further iterative DHT refinements.

Phases 7^{+}: Efficacy, Effectiveness, Implementation, and Economic Evaluation

Finally, perform efficacy, effectiveness, implementation, and economic evaluations of the resulting DHT-enhanced training concept, depending on the context of the project (-item 44). In any context, we advocate for moving beyond efficacy evaluations to address the lingering evidence-to-practice gap—which is in line with the recommendations by Owoeye et al [80]. This could be achieved by effectiveness and implementation studies following efficacy trials or with hybrid effectiveness-implementation designs [81-83]. Moreover, assess the DHT-enhanced training concept's value (eg, risk-benefit profile and cost-effectiveness) for health care systems to provide data that allow informed decisions on the uptake of such approaches in future clinical practice guidelines.

For these evaluations, we recommend following the UK MRC guidance [7] for efficacy and effectiveness terminology and evaluations, the PRACTIS guide [21] for implementation evaluations, and the World Health Organization guide [84] for cost-effectiveness analyses (-item 44). Like in phase 6, expand the abovementioned guides to accurately reflect the context of evaluations of individually tailored DHT-enhanced training concepts (-item 45) as follows.

First, while randomized controlled trials are considered the gold standard for evaluating efficacy and effectiveness [85], consider alternative study designs that may better accommodate the individualized and adaptive nature of tailored DHT-enhanced training [86] (-item 45a). Nahum-Shani et al [87] provide a pragmatic framework for selecting such alternative designs that are particularly relevant when interventions are multicomponent or tailored and when there are research questions regarding the timing, sequencing, or responsiveness of different intervention components. For example, a factorial design is a type of randomized trial in which 2 or more independent variables are manipulated simultaneously, allowing researchers to assess both the main effects of each factor and potential interactions between them. This makes it an efficient approach for studying multiple intervention components within a single trial. Furthermore, the sequential multiple assignment randomized trial design includes multiple randomizations, enabling participants to be rerandomized based on their response or adherence to the initial treatment, which is particularly useful when behaviors or conditions evolve slowly over time. In contrast, microrandomized trials involve frequent, often daily, randomizations to assess the short-term effects of brief interventions—such as prompts or notifications—on rapidly changing outcomes. Microrandomized trials are valuable for real-time optimization and adaptation of DHT-enhanced training. Despite the promise of these alternative approaches in addressing the need for tailoring and adaptability in digital health, they remain underused in health promotion and disease prevention interventions [87].

Second, continue with a mixed methods approach also at this stage of the project. Specifically, collect complementary participant-reported outcomes and qualitative data on perceived efficacy or effectiveness and acceptance of the implementation of primary and secondary end users and providers, as well as suggestions on how these could be (further) improved (-item 45b). These evaluations can provide insights not ascertainable through physiological, laboratory, clinician-reported, observer-reported, or performance outcomes alone and may thereby bring additional value and optimize the impact on clinical practice and health policy [88].

Discussion

Principal Findings

This methodological guideline development study gradually cocreated unifying consensus-based best practice recommendations with operational guidance for multidisciplinary expert teams coordinating projects on

XSL•FO
RenderX

individually tailored DHTs and DHT-enhanced training concepts. It consists of 8 phases and is applicable to any type of end users, exercise types, intended contexts of use, and overarching goals.

Strengths and Limitations

This project invested considerable time and effort in collaboratively working out a consensus on best practice recommendations—which was driven by a project team that spanned 5 research groups across 4 countries with 10 researchers representing a range of career stages, multidisciplinary backgrounds, and research interests. Nevertheless, we acknowledge that the key limitations of our approach were (1) the reliance on group decision-making and (2) consensus-finding procedures involving a relatively small number of researchers. Given the project's scope, we believe that our deliberate choice to work intensively in a small group of experts is well-suited to our research aims. While a larger number of consensus group participants could have been included through more structured techniques such as the (modified) Delphi method [18], such an approach shows high reliance on researcher bias in formulating initial, preset questionnaires and relies on an often predetermined, limited number of rounds that build on elements agreed on by a supermajority only. In contrast, our process included a high frequency (biweekly) of meetings that enabled creative brainstorming and in-depth critical discussions over more than half a year to gradually propose, discuss, and develop a consensus on new best practice recommendations—which was necessary given the comprehensiveness of this work. Furthermore, the literature suggests that 10 - 15 participants may suffice when the background of such a consensus process is homogeneous [18], which applies to our research goal and methods, thereby providing justification for the number of involved experts.

Nevertheless, we recognize inherent limitations of group decision-making, including the potential influence of dominant individuals, the risk of tangential discussions, and group pressure for compromise. These limitations could be mitigated through alternative methods that incorporate anonymity, controlled feedback, and statistical aggregation of responses [18]. Therefore, we advocate for future iterations to collaboratively consolidate this guideline through broader expert, patient, and public involvement. A (modified) Delphi method [18]—which is recommended [89,90] but underused [91] in the related field of the development of health research reporting guidelines—could be particularly valuable in this regard.

In addition to the above limitations, we acknowledge that all involved researchers are affiliated with European universities. Broader expert involvement with more diverse geographical, sociocultural, and institutional backgrounds—particularly from low- and middle-income countries, nonacademic sectors, and underrepresented communities—would strengthen the consolidation of such broadly applicable best practice recommendations. Nevertheless, the diversity of our consensus group (see "Disciplinary and Professional Background of the Consensus Group" section) sparked numerous critical discussions around the terminology used in different contexts, which in turn facilitated the development of a harmonized terminology that enhances the guideline's applicability across the various interest-holders it is intended to serve.

Comparison With Prior Work

The Co-Develop-IT guideline is an extension of the MIDE framework [14] and its refined methodology from the "Brain-IT" project [15]. The guideline integrates and expands the MRC guidance [7] and the Generative Co-Design Framework for Healthcare Innovation [12] and aspects to inform the design of new DHTs from the NASSS framework [13] and the PRACTIS guide [21]. Given that the Co-Develop-IT guideline integrated and refined these established methodological guides and frameworks, it can be seen in part as a unified approach combining the advantages of previous guides and frameworks that are expanded with refined recommendations specifically for the context of individually tailored DHT-enhanced training. However, the main novelty of this guideline lies in the introduction of multiple aspects to advance current best practices in the field.

First, the Co-Develop-IT guideline provides 5 distinct preparatory contextual research phases preceding generative codevelopment. These phases are dedicated to the structured establishment of a more robust foundation to better tailor and steer codevelopment efforts toward successful implementation. One of the key merits of the Co-Develop-IT guideline is that implementation requirements for DHT-enhanced training are considered alongside its design requirements prior to launching and continuously during the codevelopment and evaluation process. Such extensive preparatory contextual research steps extend the currently conventional approach and establish a solid foundation that facilitates the transfer of the DHT-enhanced training concept into real-world applications and clinical practice.

Second, the Co-Develop-IT guideline supports harmonizing the interests of different contributors (particularly between research, industry, health care, and end users). This is done through definitions of minimum requirements for interest-holder involvement in each project step and recommendations for systematically defining their specific roles and responsibilities.

Finally, the Co-Develop-IT guideline provides a corresponding checklist with an explanation and elaboration section to operationally guide effective implementation, which has been confirmed effective and practically feasible in 2 application examples.

Implications for Research

While collaborative research practices are generally considered best practice, with patient and public involvement nowadays a routine section on most grant proposals, concerns have been raised about whether the costs of collaborative research practices outweigh the benefits for health research [92]. Specifically, Oliver et al [92] provided an overview of the potential risks and costs of coproduction and related participatory research practices such as co-design. Such risks and costs include, but are not limited to, increased time and resource demands, emotional and professional burdens on both researchers and interest-holders, challenges in managing conflicting expectations, risks to academic credibility, and potential dilution of research quality

XSL•FO
RenderX

due to compromises made during collaboration. To overcome these risks, they recommend "[...] clarifying exactly which outcomes are required for whom for any particular piece of research [and] selecting strategies specifically designed to enable these outcomes to be achieved, and properly evaluated" [92]. Furthermore, they called for a cautious approach to codevelopment in the absence of strong evidence supporting its process and impact [92]. While there is now moderate-quality evidence that cocreation is effective by resulting in interventions that have the intended effect in the context of the secondary prevention of noncommunicable diseases [93], recent research agrees that standardized definitions and reporting frameworks for codevelopment are needed to facilitate comparisons, and ensure consistency [93,94].

The Co-Develop-IT guideline provides unifying best practice recommendations for implementing, among others, the earlier-mentioned recommendations. Specifically, it entails collaboratively—with all relevant interest-holders—(1) establishing project checkpoints consisting of a traffic light system–based assessment framework with quantitative thresholds and clear progression criteria, and (2) agreeing on the weighting of how inputs from different interest-holders should be prioritized. Furthermore, the Co-Develop-IT checklist guides standardized reporting and defines the minimum level of participatory research involvement for each item. It thereby provides operational guidance on which project steps, which interest-holders should be involved, to what extent, to ensure successful goal achievement. These strategies support projects in finding a good balance between the benefits and increased costs of collaborative research practices. Ultimately, adopting these recommendations will help advance the field to get the most out of such codevelopment projects for advancing health research that is disseminated and implemented in practical settings.

Implications for Practice

Credible DHT-enhanced interventions and related DHTs should be evidence-based and grounded in health research that provides data to support whether these solutions achieve their intended outcomes. However, a recent review revealed a significant gap between commercial offerings and scientific validation of DHT-enhanced lifestyle interventions in the example of dementia prevention. Specifically, all reviewed studies focused on proxy measures of dementia prevention, and none demonstrated any change in the targeted condition or risk factors [95]. This observation is likely to be transferable to other use cases and highlights the need for rigorous empirical validation—rather than solely equivalence claims—to provide evidence to support the often-advertised claims on clinical effectiveness of such DHT-enhanced interventions.

Following the Co-Develop-IT guideline in future projects will result in the (1) codevelopment of well-grounded DHTs and DHT-enhanced training concepts that are well aligned with all interest-holders' requirements, (2) provision of robust data whether the overarching goals of the specific DHT-enhanced training concept are supported by appropriate scientific evidence, and (3) proof-of-concept for the success of implementation of the resulting DHT-enhanced training concepts. Thereby, the Co-Develop-IT guideline systematically guides projects toward making a real-world impact.

Conclusions

The Co-Develop-IT guideline provides comprehensive best practices with actionable operational guidance for establishing an appropriate balance between scientific theories and frameworks and the real-world needs of interest-holders in the establishment of individually tailored DHT-enhanced training or rehabilitation concepts. Given that this guideline integrated and refined previous methodological guides and frameworks [7,12-15,21], it can be seen as a unified approach combining the advantages of previous guides and frameworks that are expanded with refined recommendations specifically for the context of individually tailored DHT-enhanced training. Its main novelty lies in guiding the structured establishment of a more robust conceptual foundation through extensive preparatory contextual research phases, aimed at better targeting tailored codevelopment efforts toward successful implementation.

Applying Co-Develop-IT contributes to overcoming the lingering evidence-to-practice gap by consistently establishing a shared mission with relevant interest-holders and ensuring that all codevelopment steps are directed toward addressing an unmet need in (clinical) practice—ultimately promoting the practical application and impact of purpose-developed DHTs. We are providing a platform on the Open Science Framework [96] to encourage and guide future initiatives to consolidate the Co-Develop-IT guideline based on critical discussions about experiences in their application in different projects and contexts. These efforts will help the field strike a better balance between maximizing the benefits and mitigating the increased resource demands of collaborative research practices—ultimately maximizing its real-world impact.

Acknowledgments

The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript. A nongenerative artificial intelligence (DeepL) specialized to improve and clarify language was used for refinement and correction to the manuscript.

Funding

PM's time dedicated to this project was funded by the Swedish Research Council (Dnr 2022-00636) as well as the Strategic Research Area Health Care Science at Karolinska Institutet (Blue Sky Grant Dnr 2025-01126). The contribution of MR was supported by the collaboration project LSHM24016, cofunded by the Public-Private Partnerships Allowance made available by

XSL•FO
RenderX

Health~Holland, Top Sector Life Sciences & Health. The fees for open access publication were covered by the Bibsam-agreement. The funders played no role in conceptualizing or designing the study, collecting or interpreting the data, writing the paper, or decision to submit the report for publication.

Data Availability

All data analyzed in this study are included in the paper and Multimedia Appendix 1. Further inquiries can be directed to the corresponding author or the Research Data Office at Karolinska Institutet (rdo@ki.se).

Authors' Contributions

Conceptualization and project administration: PM initiated the project in June 2024. He led and coordinated the project. PM developed the initial project proposal with feedback from EF. PM and EF jointly selected and invited members to the expert panel.
Methodology: All authors contributed feedback on the project proposal and participated in its revision.
Investigation and data curation: All authors participated in hybrid meetings to collaboratively cocreate this consensus-based methodological guideline between August 2024 and February 2025. PM took meeting protocols, synthesized the outcomes of each meeting, and drafted the corresponding sections of the guideline, checklist, and explanation and elaboration documents. All authors provided written feedback and revisions between meetings.
Visualization: The Co-Develop-IT guideline figure was cocreated by PM and LESH.
Writing—original draft: PM drafted the main manuscript and Supplementary Files 1, 2, and 5 in Multimedia Appendix 1. The first application example (Supplementary File 3 in Multimedia Appendix 1) was developed by PM, EF, AW, BL, and FA. The second application example (Supplementary File 4 in Multimedia Appendix 1) was developed by LESH and MR.
Writing—review and editing: All authors contributed to revisions of the manuscript and supplementary files and approved the final submitted version.

Conflicts of Interest

The research project involving MR and LESH that provided the second application example covering the development of Reality DTx is a collaborative initiative between the Vrije Universiteit research team and the manufacturer of Reality DTx, Strolll Limited, United Kingdom, formalized in a collaboration agreement as part of a shared research and development grant (EUreka Eurostars Grant E115506) attracted by MR and Strolll Limited. As part of the arrangement, Vrije Universiteit Amsterdam transferred intellectual property related to augmented reality cueing and data science to Strolll Limited in exchange for share options. MR is a scientific adviser with share options for Strolll Limited alongside his full-time academic positions at Vrije Universiteit Amsterdam and Maastricht University. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

Multimedia Appendix 1
The supplementary file provides the following elements from the Co-Develop-IT guideline package: (1) Co-Develop-IT checklist (Supplementary File 1), (2) explanations and elaborations of each checklist item with guiding examples (Supplementary File 2), (3) application example of the Co-Develop-IT guideline in the “Park-MOVE” project (Supplementary File 3), and (4) application example of the Co-Develop-IT guideline in the “Better Together” project (Supplementary File 4).
[PDF File, 803 KB - jmir_v28i1e84163_app1.pdf]

References

1. Gajardo Sánchez AD, Murillo-Zamorano LR, López-Sánchez J, Bueno-Muñoz C. Gamification in health care management: systematic review of the literature and research agenda. SAGE Open 2023 Oct;13(4):21582440231218834. [doi: 10.1177/21582440231218834]
2. Manser P, de Bruin ED, Temprado JJ, Bherer L, Herold F. Beyond “Just” Fun: the role of exergames in advancing health promotion and disease prevention. Neurosci Biobehav Rev 2025 Sep;176:106260. [doi: 10.1016/j.neubiorev.2025.106260]
3. Maggio MG, Baglio F, Maione R, et al. The overlooked role of exergames in cognitive-motor neurorehabilitation: a systematic review. NPJ Digit Med 2025 Jul 9;8(1):419. [doi: 10.1038/s41746-025-01843-4] [Medline: 40634592]
4. Singh B, Ahmed M, Staiano AE, et al. A systematic umbrella review and meta-meta-analysis of eHealth and mHealth interventions for improving lifestyle behaviours. NPJ Digit Med 2024 Jul 5;7(1):179. [doi: 10.1038/s41746-024-01172-y] [Medline: 38969775]
5. Recommendations on digital interventions for health system strengthening. World Health Organization. 2019. URL: https://www.who.int/publications/i/item/9789241550505 [accessed 2026-04-27]
6. Herold F, Theobald P, Gronwald T, et al. The best of two worlds to promote healthy cognitive aging: definition and classification approach of hybrid physical training interventions. JMIR Aging 2024 Jul 31;7:e56433. [doi: 10.2196/56433] [Medline: 39083334]

XSL•FO
RenderX

7. Skivington K, Matthews L, Simpson SA, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061. [doi: 10.1136/bmj.n2061] [Medline: 34593508]
8. Booth V, Hood-Moore V, Hancox JE, Logan P, Robinson KR. Systematic scoping review of frameworks used to develop rehabilitation interventions for older adults. BMJ Open 2019 Feb;9(2):e024185. [doi: 10.1136/bmjopen-2018-024185]
9. Connell L, Ackerley S, Rycroft-Malone J. Applying the updated MRC framework for developing and evaluating complex interventions with integrated implementation conceptual knowledge: an example using NeuroRehabilitation OnLine. Front Health Serv 2025;5:1562627. [doi: 10.3389/frhs.2025.1562627] [Medline: 40395560]
10. Abras C, Maloney-Krichmar D. User-centered design. In: Bainbridge W, editor. Encyclopedia of Human-Computer Interaction: Sage Publications; 2004, Vol. 4. URL: https://aim.johnkeston.com/wp-content/uploads/2012/01/User-centered_design_encyclopedia_chapter.pdf [accessed 2026-04-27]
11. Moore G, Wilding H, Gray K, Castle D. Participatory methods to engage health service users in the development of electronic health resources: systematic review. J Participat Med 2019;11(1):e11474. [doi: 10.2196/11474]
12. Bird M, McGillion M, Chambers EM, et al. A generative co-design framework for healthcare innovation: development and application of an end-user engagement framework. Res Involv Engagem 2021 Mar 1;7(1):12. [doi: 10.1186/s40900-021-00252-7] [Medline: 33648588]
13. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
14. Li Y, Muñoz J, et al. Multidisciplinary Iterative Design of Exergames (MIDE): a framework for supporting the design, development, and evaluation of exergames for health. Presented at: International Conference on Human-Computer; Jul 19-24, 2020. [doi: 10.1007/978-3-030-50164-8_9]
15. Manser P, de Bruin ED. Making the Best Out of IT: design and development of exergames for older adults with mild neurocognitive disorder—a methodological paper. Front Aging Neurosci 2021;13:734012. [doi: 10.3389/fnagi.2021.734012] [Medline: 34955806]
16. Harb SI, Tao L, Peláez S, Boruff J, Rice DB, Shrier I. Methodological options of the nominal group technique for survey item elicitation in health research: a scoping review. J Clin Epidemiol 2021 Nov;139:140-148. [doi: 10.1016/j.jclinepi.2021.08.008] [Medline: 34400255]
17. McMillan SS, King M, Tully MP. How to use the nominal group and Delphi techniques. Int J Clin Pharm 2016 Jun;38(3):655-662. [doi: 10.1007/s11096-016-0257-x] [Medline: 26846316]
18. Taylor E. We agree, don't we? The Delphi method for health environments research. HERD 2020 Jan;13(1):11-23. [doi: 10.1177/1937586719887709]
19. Manser P, de Bruin ED. "Brain-IT": exergame training with biofeedback breathing in neurocognitive disorders. Alzheimers Dement 2024 Jul;20(7):4747-4764. [doi: 10.1002/alz.13913] [Medline: 38809948]
20. Huber SK, Manser P, de Bruin ED. PEMOCS: theory derivation of a concept for PErsonalized MOtor-Cognitive exergame training in chronic Stroke—a methodological paper with an application example. Front Sports Act Living 2024;6:1397949. [doi: 10.3389/fspor.2024.1397949] [Medline: 38915297]
21. Koorts H, Eakin E, Estabrooks P, Timperio A, Salmon J, Bauman A. Implementation and scale up of population physical activity interventions for clinical and community settings: the PRACTIS guide. Int J Behav Nutr Phys Act 2018 Jun 8;15(1):51. [doi: 10.1186/s12966-018-0678-0] [Medline: 29884236]
22. Manser P, Rosio M, Schmidt A, Michels L, de Bruin ED. Structural brain improvements following individually tailored serious exergame-based training in mild neurocognitive disorders: exploratory randomized controlled trial. Alz Res Therapy 2025;17(1):190. [doi: 10.1186/s13195-025-01835-2]
23. Manser P, Poikonen H, de Bruin ED. Feasibility, usability, and acceptance of "Brain-IT"—a newly developed exergame-based training concept for the secondary prevention of mild neurocognitive disorder: a pilot randomized controlled trial. Front Aging Neurosci 2023;15:1163388. [doi: 10.3389/fnagi.2023.1163388] [Medline: 37810620]
24. Herold F, Müller P, Gronwald T, Müller NG. Dose-response matters!—A perspective on the exercise prescription in exercise-cognition research. Front Psychol 2019;10:2338. [doi: 10.3389/fpsyg.2019.02338] [Medline: 31736815]
25. Vaughn LM, Jacquez F. Participatory research methods—choice points in the research process. J Partic Res Methods 2020;1(1). [doi: 10.35844/001c.13244]
26. Enhancing the QUAlity and Transparency Of health Research. EQUATOR Network. 2025. URL: https://www.equator-network.org/ [accessed 2025-05-30]
27. Boers M, Rochereau A, Stuwe L, et al. Classification grid and evidence matrix for evaluating digital medical devices under the European Union landscape. NPJ Digit Med 2025 May 24;8(1):304. [doi: 10.1038/s41746-025-01697-w] [Medline: 40413360]
28. Wallin A, Franzén E, Ekman U, Piehl F, Johansson S. A highly challenging balance training intervention for people with multiple sclerosis: a feasibility trial. Pilot Feasibility Stud 2023 Mar 15;9(1):41. [doi: 10.1186/s40814-023-01265-7] [Medline: 36922859]
29. Cairns P, Pinker I, Ward A, Watson E, Laidlaw A. Empathy maps in communication skills training. Clin Teach 2021 Apr;18(2):142-146. [doi: 10.1111/tct.13270] [Medline: 33034104]

XSL•FO
RenderX

30. Rohner SL, Stadtmann MP, Gattinger H, et al. Co-creation for the development and implementation of a competence centre for mental health in Eastern Switzerland: a participatory approach. BMC Psychiatry 2025 Mar 18;25(1):254. [doi: 10.1186/s12888-025-06703-9] [Medline: 40102800]
31. Guidelines for co-creation in mental health. Mental Health Europe. 2023. URL: https://www.mentalhealtheurope.org/wp-content/uploads/2024/04/Mental-Health-Europes-Guidelines-for-Cocreation.pdf [accessed 2024-09-21]
32. Izquierdo M, de Souto Barreto P, Arai H, et al. Global consensus on optimal exercise recommendations for enhancing healthy longevity in older adults (ICFSR). J Nutr Health Aging 2025 Jan;29(1):100401. [doi: 10.1016/j.jnha.2024.100401] [Medline: 39743381]
33. American College of Sports Medicine. ACSM's Guidelines for Exercise Testing and Prescription: Wolters Kluwer; 2025. URL: https://acsm.org/education-resources/books/guidelines-exercise-testing-prescription/ [accessed 2026-04-27]
34. Kleim JA, Jones TA. Principles of experience-dependent neural plasticity: implications for rehabilitation after brain damage. J Speech Lang Hear Res 2008 Feb;51(1):S225-S239. [doi: 10.1044/1092-4388(2008/018)] [Medline: 18230848]
35. Kleim JA. Neural plasticity and neurorehabilitation: teaching the new brain old tricks. J Commun Disord 2011;44(5):521-528. [doi: 10.1016/j.jcomdis.2011.04.006] [Medline: 21600589]
36. Maier M, Ballester BR, Verschure P. Principles of neurorehabilitation after stroke based on motor learning and brain plasticity mechanisms. Front Syst Neurosci 2019;13:74. [doi: 10.3389/fnsys.2019.00074] [Medline: 31920570]
37. Rhodes RE, McEwan D, Rebar AL. Theories of physical activity behaviour change: a history and synthesis of approaches. Psychol Sport Exerc 2019 May;42:100-109. [doi: 10.1016/j.psychsport.2018.11.010]
38. Herold F, Hamacher D, Schega L, Müller NG. Thinking while moving or moving while thinking—concepts of motor-cognitive training for cognitive performance enhancement. Front Aging Neurosci 2018;10(228):228. [doi: 10.3389/fnagi.2018.00228] [Medline: 30127732]
39. Raichlen DA, Alexander GE. Adaptive capacity: an evolutionary neuroscience model linking exercise, cognition, and brain health. Trends Neurosci 2017 Jul;40(7):408-421. [doi: 10.1016/j.tins.2017.05.001] [Medline: 28610948]
40. Cheval B, Boisgontier MP. The theory of effort minimization in physical activity. Exerc Sport Sci Rev 2021 Jul 1;49(3):168-178. [doi: 10.1249/JES.0000000000000252] [Medline: 34112744]
41. Gerber M, Cheval B, Cody R, et al. Psychophysiological foundations of human physical activity behavior and motivation: theories, systems, mechanisms, evolution, and genetics. Physiol Rev 2025 Jul 1;105(3):1213-1290. [doi: 10.1152/physrev.00021.2024] [Medline: 39854639]
42. Wulf G, Lewthwaite R. Optimizing performance through intrinsic motivation and attention for learning: the OPTIMAL theory of motor learning. Psychon Bull Rev 2016 Oct;23(5):1382-1414. [doi: 10.3758/s13423-015-0999-9] [Medline: 26833314]
43. El Kirat H, van Belle S, Khattabi A, Belrhiti Z. Behavioral change interventions, theories, and techniques to reduce physical inactivity and sedentary behavior in the general population: a scoping review. BMC Public Health 2024 Aug 3;24(1):2099. [doi: 10.1186/s12889-024-19600-9] [Medline: 39097741]
44. Hamasaki T, Briand C, Mahroug A, Sauvageau A. Digital health technology adoption factors: a rapid review of systematic reviews and checklist development. Disabil Rehabil Assist Technol 2026 Jan;21(1):95-112. [doi: 10.1080/17483107.2025.2526175] [Medline: 40622303]
45. Grutters JPC, Bouttell J, Abrishami P, et al. Defining early health technology assessment: building consensus using Delphi technique. Int J Technol Assess Health Care 2025 Jun 2;41(1):e34. [doi: 10.1017/S0266462325100123] [Medline: 40452233]
46. Burns PB, Rohrich RJ, Chung KC. The levels of evidence and their role in evidence-based medicine. Plast Reconstr Surg 2011 Jul;128(1):305-310. [doi: 10.1097/PRS.0b013e318219c171] [Medline: 21701348]
47. Farr M. Power dynamics and collaborative mechanisms in co-production and co-design processes. Crit Soc Policy 2018 Nov;38(4):623-644. [doi: 10.1177/0261018317747444]
48. Luc A, Slatman S, Everard G, Huygelier H, Pitance L, Detrembleur C. Knowledge, practices, and acceptance of physiotherapists regarding the use of immersive virtual reality in clinical practice: a cross-sectional online survey. Disabil Rehabil 2026 Feb 2;2026:1-17. [doi: 10.1080/09638288.2026.2622898] [Medline: 41622875]
49. Alt Murphy M, Pradhan S, Levin MF, Hancock NJ. Uptake of technology for neurorehabilitation in clinical practice: a scoping review. Phys Ther 2024 Feb 1;104(2):pzad140. [doi: 10.1093/ptj/pzad140] [Medline: 37856528]
50. Gavi B, Daniel H. Analysing Health Communication—Discourse Approaches: Springer; 2021. [doi: 10.1007/978-3-030-68184-5]
51. Marta B, Maria Laura I, et al. Healthcare in the Digital Age—Perspectives for Sustainable Innovation and Assessment: Springer; 2025. [doi: 10.1007/978-981-96-1437-0]
52. Torre MM, Temprado JJ. Effects of exergames on brain and cognition in older adults: a review based on a new categorization of combined training intervention. Front Aging Neurosci 2022;14:859715. [doi: 10.3389/fnagi.2022.859715] [Medline: 35431905]
53. Wilson C. Chapter 2—Brainwriting. In: Brainstorming and Beyond: Morgan Kaufmann; 2013. [doi: 10.1016/B978-0-12-407157-5.00002-6]
54. Gray D, Brown S, et al. Gamestorming: A Playbook for Innovators, Rulebreakers, and Changemakers: O'Reilly Media, Inc; 2010. URL: https://gamestorming.com [accessed 2026-04-27]

XSL•FO
RenderX

55. Chang WL, Shao YC. Co-creating user journey map—a systematic approach to exploring users' day-to-day experience in participatory design workshops. In: Kurosu M, Hashizume A, editors. Human-Computer Interaction: Springer Nature Switzerland; 2023:3-17. [doi: 10.1007/978-3-031-35596-7_1]
56. Márquez Segura E, Turmo Vidal L, Rostami A. Bodystorming for movement-based interaction design. Human Technol 2016;12(2):193-251. [doi: 10.17011/ht/urn.201611174655]
57. Schleicher D, Jones P, Kachur O. Bodystorming as embodied designing. interactions 2010 Nov;17(6):47-51. [doi: 10.1145/1865245.1865256]
58. Mahatody T, Sagar M, Kolski C. State of the art on the cognitive walkthrough method, its variants and evolutions. Int J Hum-Comput Interact 2010 Jul 30;26(8):741-785. [doi: 10.1080/10447311003781409]
59. Mayring P. Qualitative Content Analysis: A Step-by-Step Guide: SAGE Publications Ltd; 2021. URL: https://uk.sagepub.com/en-gb/eur/qualitative-content-analysis/book269922 [accessed 2026-04-27]
60. Krippendorff K. Content Analysis: An Introduction to Its Methodology, th edition: SAGE Publications Ltd; 2019. [doi: 10.4135/9781071878781]
61. Mayring P. Qualitative content analysis: theoretical foundation, basic procedures and software solution. In: Approaches to Qualitative Research in Mathematics Education: Springer; 2014. [doi: 10.1007/978-94-017-9181-6_13]
62. Schrepp M, Thomaschewski J. Design and validation of a framework for the creation of user experience questionnaires. IJIMAI 2019;5(7):88-95. [doi: 10.9781/ijimai.2019.06.006]
63. Winter D, Hinderks A, et al. Welche UX faktoren sind für mein produkt wichtig. Presented at: Mensch und Computer 2017—Usability Professionals; Sep 10-13, 2017. [doi: 10.18420/muc2017-up-0002]
64. Winter D, Schrepp M, et al. Faktoren der user experience-systematische übersicht über produktrelevante UX-qualitätsaspekte. 2015 Presented at: Mensch und Computer 2015—Usability Professionals; Oct 6-9, 2015. [doi: 10.1515/9783110443882-005]
65. Perez FMP, Bellei EA, Nascimento JR, Bombarda LT, Belger VDF, De Marchi ACB. Decoding user experience in exergames: a systematic scoping review of assessment methods. MethodsX 2025 Jun;14:103054. [doi: 10.1016/j.mex.2024.103054] [Medline: 39866200]
66. Johnson RM. User experience research and usability of health information technology. Techn Commun Q 2026 Jan 2;35(1):119-122. [doi: 10.1080/10572252.2025.2455553]
67. Campbell JL. User Experience Research and Usability of Health Information Technology: Auerbach Publications; 2024. [doi: 10.1201/9781003460886]
68. Guggenberger B, Jocham AJ, Jocham B, Nischelwitzer A, Ritschl H. Instrumental validity of the motion detection accuracy of a smartphone-based training game. Int J Environ Res Public Health 2021 Aug 9;18(16):8410. [doi: 10.3390/ijerph18168410] [Medline: 34444160]
69. Kaiser W, de Bruin ED, Manser P. Domain-specific evaluation of exergame metrics among older adults with mild neurocognitive disorder: secondary analysis of 2 randomized controlled trials. JMIR Serious Games 2025 May 21;13:e65878. [doi: 10.2196/65878] [Medline: 40397948]
70. Guimarães V, Sousa I, de Bruin ED, Pais J, Correia MV. Using shoe-mounted inertial sensors and stepping exergames to assess the motor-cognitive status of older adults: a correlational study. Digit Health 2023;9:20552076231167001. [doi: 10.1177/20552076231167001] [Medline: 37009304]
71. Konstantinidis EI, Bamidis PD, Billis A, Kartsidis P, Petsani D, Papageorgiou SG. Physical training in-game metrics for cognitive assessment: evidence from extended trials with the Fitforall exergaming platform. Sensors (Basel) 2021 Aug 26;21(17):5756. [doi: 10.3390/s21175756] [Medline: 34502647]
72. Petsani D, Konstantinidis E, Katsouli AM, et al. Digital biomarkers for well-being through exergame interactions: exploratory study. JMIR Serious Games 2022 Sep 13;10(3):e34768. [doi: 10.2196/34768] [Medline: 36099000]
73. Litz E, Ball C, Jansen CP, Werner C, de Bruin E, Hauer K. Validation of a motor-cognitive assessment for a stepping exergame in older adults: use of game-specific, internal data stream. Games Health J 2020 Apr;9(2):95-107. [doi: 10.1089/g4h.2019.0081] [Medline: 31651191]
74. Wiloth S, Lemke N, Werner C, Hauer K. Validation of a computerized, game-based assessment strategy to measure training effects on motor-cognitive functions in people with dementia. JMIR Serious Games 2016 Jul 18;4(2):e12. [doi: 10.2196/games.5696] [Medline: 27432746]
75. Laugwitz B, Held T, Schrepp M. Construction and evaluation of a user experience questionnaire. In: Holzinger A, editor. HCI and Usability for Education and Work: Springer; 2008:63-76. [doi: 10.1007/978-3-540-89350-9_6]
76. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud 2009;4(3):114-123. [doi: 10.5555/2835587.2835589]
77. Manser P, Huber S, Seinsche J, de Bruin ED, Giannouli E. Development and initial validation of the German version of the Exergame Enjoyment Questionnaire (EEQ-G). PLoS One 2023;18(6):e0286556. [doi: 10.1371/journal.pone.0286556] [Medline: 37289701]
78. Eldridge SM, Lancaster GA, Campbell MJ, et al. Defining feasibility and pilot studies in preparation for randomised controlled trials: development of a conceptual framework. PLoS One 2016;11(3):e0150205. [doi: 10.1371/journal.pone.0150205]

XSL•FO
RenderX

79. O'Cathain A, Hoddinott P, Lewin S, et al. Maximising the impact of qualitative research in feasibility studies for randomised controlled trials: guidance for researchers. Pilot Feasibility Stud 2015 Dec;1(1):32. [doi: 10.1186/s40814-015-0026-y]
80. Owoeye OBA, Rauvola RS, Brownson RC. Dissemination and implementation research in sports and exercise medicine and sports physical therapy: translating evidence to practice and policy. BMJ Open Sport Exerc Med 2020 Nov;6(1):e000974. [doi: 10.1136/bmjsem-2020-000974]
81. Landes SJ, McBain SA, Curran GM. An introduction to effectiveness-implementation hybrid designs. Psychiatry Res 2019 Oct;280:112513. [doi: 10.1016/j.psychres.2019.112513] [Medline: 31434011]
82. Curran GM, Bauer M, Mittman B, Pyne JM, Stetler C. Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. Med Care 2012 Mar;50(3):217-226. [doi: 10.1097/MLR.0b013e3182408812] [Medline: 22310560]
83. Curran GM, Landes SJ, McBain SA, et al. Reflections on 10 years of effectiveness-implementation hybrid studies. Front Health Serv 2022;2:1053496. [doi: 10.3389/frhs.2022.1053496] [Medline: 36925811]
84. Edejer TT. Making choices in health: WHO guide to cost-effectiveness analysis. World Health Organization. 2003. URL: https://iris.who.int/handle/10665/42699 [accessed 2026-04-27]
85. Hariton E, Locascio JJ. Randomised controlled trials—the gold standard for effectiveness research. BJOG 2018 Dec;125(13):1716-1716. [doi: 10.1111/1471-0528.15199]
86. Daniels K, Quadflieg K, Bonnechère B. Mobile health interventions for active aging: a systematic review and meta-analysis on the effectiveness of physical activity promotion. Mhealth 2025;11:4. [doi: 10.21037/mhealth-24-41] [Medline: 39944853]
87. Nahum-Shani I, Dziak JJ, Wetter DW. MCMTC: a pragmatic framework for selecting an experimental design to inform the development of digital interventions. Front Digit Health 2022;4:798025. [doi: 10.3389/fdgth.2022.798025] [Medline: 35355685]
88. Crossnohere NL, Schuster ALR, Bruckel J, et al. Patient-reported outcome measures add value as clinical trial endpoints. Nat Med 2025 Nov;31(11):3601-3604. [doi: 10.1038/s41591-025-03906-1] [Medline: 40855193]
89. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. PLoS Med 2010 Feb 16;7(2):e1000217. [doi: 10.1371/journal.pmed.1000217] [Medline: 20169112]
90. Black N, Murphy M, Lamping D, et al. Consensus development methods: a review of best practice in creating clinical guidelines. J Health Serv Res Policy 1999 Oct;4(4):236-248. [doi: 10.1177/135581969900400410] [Medline: 10623041]
91. Banno M, Tsujimoto Y, Kataoka Y. The majority of reporting guidelines are not developed with the Delphi method: a systematic review of reporting guidelines. J Clin Epidemiol 2020 Aug;124:50-57. [doi: 10.1016/j.jclinepi.2020.04.010] [Medline: 32302679]
92. Oliver K, Kothari A, Mays N. The dark side of coproduction: do the costs outweigh the benefits for health research? Health Res Policy Syst 2019 Mar 28;17(1):33. [doi: 10.1186/s12961-019-0432-3] [Medline: 30922339]
93. Anieto EM, Dall PM, Abaraogu U, et al. The effectiveness of co-created lifestyle interventions in improving health behaviour, physical and mental health in adults with non-communicable diseases: a systematic review with meta-analysis. Public Health (Fairfax) 2025 Nov;248:105929. [doi: 10.1016/j.puhe.2025.105929]
94. Longworth GR, de Boer J, Goh K, et al. Navigating process evaluation in co-creation: a Health CASCADE scoping review of used frameworks and assessed components. BMJ Glob Health 2024 Jul 4;9(7):e014483. [doi: 10.1136/bmjgh-2023-014483] [Medline: 38964878]
95. Vinay R, Probst J, Huynh P, Schlögl M, Kowatsch T, Nißen M. Top-funded digital health companies offering lifestyle interventions for dementia prevention: company overview and evidence analysis. PLoS One 2025;20(5):e0323390. [doi: 10.1371/journal.pone.0323390] [Medline: 40323948]
96. Manser P, Hardeman L, Wallin A, et al. Platform to encourage and guide future initiatives to consolidate "Co-Develop-IT! Unifying methodological guideline for the co-design, development, and evaluation of individually tailored technology-enhanced training and rehabilitation concepts: consensus development study and tutorial". Open Science Framework. 2026. URL: https://doi.org/10.17605/OSF.IO/W2JD3 [accessed 2026-05-09]

Abbreviations

DHT: digital health technology
EQUATOR: Enhancing the Quality and Transparency of Health Research
MIDE: Multidisciplinary Iterative Design of Exergames
MRC: Medical Research Council
NASSS: nonadoption, abandonment, scale-up, spread, and sustainability
PRACTIS: Practical Planning for Implementation and Scale-Up

XSL•FO
RenderX

Edited by A Stone; submitted 17.Sep.2025; peer-reviewed by F Elkourdi, M Rosenlund; revised version received 10.Apr.2026; accepted 10.Apr.2026; published 22.May.2026.

Please cite as:
Manser P, Hardeman LES, Wallin A, Moulaee Conradsson D, Leavy B, Albrecht F, Roerdink M, Temprado JJ, de Bruin ED, Franzén E
Co-Develop-IT! Unifying Methodological Guideline for the Co-Design, Development, and Evaluation of Individually Tailored Technology-Enhanced Training and Rehabilitation Concepts: Consensus Development Study and Tutorial
J Med Internet Res 2026;28:e84163
URL: https://www.jmir.org/2026/1/e84163
doi:10.2196/84163

© Patrick Manser, Lotte E S Hardeman, Andreas Wallin, David Moulaee Conradsson, Breiffni Leavy, Franziska Albrecht, Melvyn Roerdink, Jean-Jacques Temprado, Eling D de Bruin, Erika Franzén. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Using an Adult-Designed Wearable for Pediatric Monitoring: Practical Tutorial and Application in School-Aged Children With Obesity

Alessandra Angelucci[1], PhD; Andrea Aliverti[1], PhD; Matteo Vandoni[2], PhD; Daniela Lucini[3,4], MD, PhD; Cristiana Larizza[5], PhD; Vittoria Carnevale Pellino[2], PhD; Gianvincenzo Zuccotti[6,7*], MD; Valeria Calcaterra[7,8*], MD

[1]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan, Italy

[2]Laboratory of Adapted Motor Activity (LAMA), Department of Public Health, Experimental Medicine and Forensic Science, University of Pavia, Pavia, Italy

[3]BIOMETRA Department, University of Milan, Milan, Italy

[4]Exercise Medicine Unit, Istituto di Ricovero e Cura a Carattere Scientifico Istituto Auxologico Italiano, Milan, Italy

[5]Department of Computer, Electrical and Biomedical Engineering, University of Pavia, Pavia, Italy

[6]Department of Biomedical and Clinical Science, University of Milan, Milan, Italy

[7]Pediatric Department, Buzzi Children's Hospital, Milan, Italy

[8]Department of Internal Medicine and Therapeutics, University of Pavia, Pavia, Italy

[*]these authors contributed equally

Corresponding Author:
Andrea Aliverti, PhD
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan, Italy

Abstract

This tutorial presents a step-by-step guide on how to use an adult-oriented wearable (Fitbit) to collect and analyze activity and cardiovascular data in a pediatric population of school-aged children with obesity. After outlining the current landscape of commercial wearables for adults and children, the tutorial illustrates the distinct considerations required for accurate pediatric monitoring, especially for cardiovascular metrics and derived features. The text provides a clinical application, highlighting how data from these devices were gathered and integrated with standard clinical measurements (ie, 1 week of monitoring with the wearable compared with the 6-minute walk test). The tutorial also discusses potential correlations, which should be interpreted as exploratory, given the small sample size (n=16), as well as limitations and future perspectives on using wearables for long-term pediatric monitoring of school-aged children, aiming to inform clinicians, researchers, and other stakeholders about the additional considerations that are needed to use wearables designed for adults to monitor this age group.

(J Med Internet Res 2026;28:e76166) doi:10.2196/76166

KEYWORDS

childhood obesity; wearable devices; activity monitoring; cardiovascular monitoring; wearable data quality

Introduction

Wearable Devices

Wearables are considered key elements of future health care because they enable ecological measurements without interfering with daily life. Feedback provided by consumer devices, especially those equipped with displays, can induce behavioral reactivity, potentially nudging toward active lifestyles [1]. Wearables find many applications in the field of telemonitoring of patients [2], consumer health and well-being, sports performance assessment (particularly in endurance disciplines and predominantly during training rather than competition [3]), environmental monitoring [4], and emergency settings [5]. Examples of wearables include watches (such as smartwatches and activity trackers), eyeglasses [6], visors [7], rings [8], patches [9,10], sensorized garments (such as shirts, gloves, and socks) [11], elastic bands applied to the torso [12,13], earbuds [14], and devices clipped to the shoes [15].

Even if it is often unclear to what extent nonmedical-grade commercial devices are validated, such devices have been proposed in the medical field as self-assessment tools [16,17] and may provide data associated with information of interest, such as activity levels and physiological parameters.

Activity metrics range from the number of steps performed in a fixed period [18], most commonly in a day, to a wide variety of parameters, such as distance, minutes in different heart rate zones (HRZ), and the number of stand hours [19].

Physiological measurements can consider a single parameter, for example, heart rate (HR) or pulse rate (PR) [20], peripheral

XSL•FO
RenderX

blood oxygen saturation [20], respiratory rate [21], skin temperature, and galvanic skin response or electrodermal activity, or could be more complex measures combining data from several sensors, such as in the case of sleep analysis, estimation of energy expenditure, and analysis of physiological parameters with respect to activity [22-25].

Devices for School-Aged Children

A first distinction must be made between wearable systems for neonates, infants, and school-aged children. Certain US Food and Drug Administration–cleared medical-grade patches, such as ANNE One (Sibel Health Inc) [26,27], have been successfully deployed for continuous vital sign monitoring in neonates and infants, particularly in intensive care settings [27]. These devices are designed for short-term clinical use under medical supervision and are optimized for the anatomical and physiological characteristics of newborns (eg, very small body size, high resting HR, thin skin, and limited mobility). However, such solutions are not directly applicable to ambulatory or home-based pediatric monitoring in school-aged children. Children exhibit larger body dimensions, mobility, and behavioral variability, which require wearables that are robust, reusable, and comfortable for daily use outside the hospital environment.

Another distinction should be made between devices for adults and children. Physiological monitoring in the school-aged child population is inherently more complex than in adults due to several factors. First, children have markedly different cardiovascular physiology; for instance, normal HRs in healthy children can reach upward of 180 beats per minute [28]. Second, anatomical scaling poses challenges in terms of device design, as wearables must be smaller and adapt to smaller vessels, thinner tissues, variable subcutaneous fat, and smaller chest circumference. Third, children's patterns of movement and physical activity are more dynamic and unpredictable, which increases motion artifacts, signal noise, and intermittent sensor loss [29]. Failure to monitor cardiovascular metrics continuously in children can delay early detection of hemodynamic compromise, mask subtle progression of dysfunction, and limit timely interventions, with consequences for morbidity and long-term outcomes [30]. The limitations of wearables for pediatric monitoring add up to the already existing limitations of wearable devices, even more so when commercial devices are involved.

In general, the results of studies about the accuracy and precision of commercial smartwatches used by adults are inconsistent. Regarding PR measurement from a Fitbit wrist-worn device, a study [31] found that the accuracy of PR estimation from a photoplethysmographic (PPG) sensor (<10% mean absolute error with respect to a gold-standard electrocardiogram) across 24 hours and during various activities was acceptable. Conversely, other studies [32] reported significant accuracy variations depending on the activity type, and a notable proportion of PR measurements were substantially inaccurate. Focusing on physical activity metrics, a 2015 study [33] found that Fitbit's step counts strongly correlated with those obtained by gold-standard methods. However, they observed low accuracy at slower walking speeds. Additionally, such results are valid in the adult population and cannot be immediately translated to pediatric applications.

It becomes even more complicated when children are involved. In fact, most research on the use of wearables in children is focused on activity tracking [34]. At the market level, popular models of wearables for children are limited in the number of integrated sensors and, therefore, in functionality. Most models only perform the function of activity trackers, such as the Fitbit Ace 3 [35] and the Garmin vívofit jr. 3 [36]. The Fitbit Ace 3, for instance, is only equipped with an accelerometer and not with a PPG sensor, thus making it not usable to assess cardiovascular features. Brands such as Apple and Polar do not offer any consumer product that is specifically designed for children. Polar used to offer the Polar Active [37] device with a business-to-business model for physical activity programs in school-aged children [38].

Nash et al [39] evaluated the accuracy of Apple Watch–derived cardiac features in children, both for the PPG and electrocardiography (ECG). In terms of PPG, they assessed whether PR estimates were accurate during both sinus rhythm and arrhythmias and found that there was a good agreement between the PPG and the gold-standard ECG during sinus rhythm (intraclass correlation coefficient [ICC] 0.98 - 0.99) and poor agreement during arrhythmias (ICC 0.24 - 0.27).

In a prospective cohort of children referred for 24-hour Holter monitoring (n=31, mean age 13.2, SD 3.6 y), a PPG-based device with the European Union Medical Device Regulation approval for use in adults, CardioWatch 287-2B (Corsano Health BV), achieved a mean accuracy of 84.8% (SD 8.7%) of readings; accuracy was significantly higher at lower HR (mean 90.9%, SD 9.3%) than at higher HR (mean 79.0%, SD 10.6%), and Bland–Altman analysis showed a bias of −1.4 bpm with 95% limits of agreement from −18.8 to +16.0 bpm [40].

In addition, ECG has been explored in smartwatches for pediatric applications. Unlike optical PPG-based wearables, the ScanWatch (Withings) uses a one-lead ECG approach. In a study of 100 children aged 5 to 17 years, the ICC for HR was 0.97, 0.86 for the PR interval, and 0.80 for the QT interval, whereas QRS and QTc interval agreement was only moderate [41].

Rationale and Aim of This Tutorial

The market for wearables for pediatric monitoring of school-aged children presents a gap in the offer and availability of monitors of cardiovascular function, as the most popular commercial wearable brands do not offer PPG-equipped (or ECG-equipped) devices for children or do not offer devices for children at all. It is nowadays necessary to use adult-designed wearables if there is a need for cardiovascular monitoring. The rationale for this tutorial focuses on the methodological adaptation required to use adult-grade devices equipped with PPG sensors and other integrated modules for cardiovascular monitoring in school-aged children. Dedicated child-specific devices with comparable sensing capabilities are largely unavailable, and even when they are marketed as child-specific devices, it is often not disclosed whether the changes are only aesthetic and in shape or also from a sensors and algorithms

XSL•FO
RenderX

point of view. This tutorial, therefore, highlights both the opportunities and the technical and physiological limitations of repurposing such devices for pediatric research and clinical applications. Rather than testing clinical hypotheses, the primary aim of this work is to demonstrate methodological feasibility by showing how an adult-designed wearable can be deployed, processed, and interpreted when used in school-aged children and by providing a transparent, reproducible workflow for researchers and clinicians. The clinical example included in the manuscript serves to illustrate the application of these methods, not to evaluate the effectiveness of the wearable or to draw population-level clinical inferences.

Step-by-Step Tutorial

Overview

The workflow is organized into 12 steps that readers can replicate in pediatric studies using adult-designed wearables. Each step describes what to do and how to do it, including defining the clinical context (steps 1‑4), managing pediatric-specific reference values (step 5), selecting and importing Fitbit files (steps 6 and 7), extracting activity and cardiovascular metrics (steps 8‑10), assessing data quality (step 11), and conducting statistical analysis (step 12).

Step 1: Framing the Clinical Case

The first step must be taken before collecting any data and consists of framing the clinical case that will be investigated by using adult-designed wearables in a pediatric population. This is a necessary step that guides the future choices of the parameters that will be retained or discarded and the clinical outcomes.

This tutorial includes a practical example based on data collected from children with obesity to demonstrate how wearable-derived metrics can be dealt with and compared with conventional clinical measures when adult-grade devices are used in children. Childhood obesity, defined according to the World Health Organization growth reference as a BMI greater than +2 SDs above the age- and sex-specific mean [42], represents a major global health concern [43]. BMI remains the most widely used indicator for classification, with equivalent percentile-based criteria also adopted by the Centers for Disease Control and Prevention [44]. By including this clinical case, the tutorial aims to illustrate how data extracted from an adult-designed wearable can be contextualized and analyzed in relation to standard clinical outcomes, such as the 6-minute walk test (6MWT) and blood pressure percentiles.

Exercise, as a nonpharmacological approach, can help delay obesity-related conditions, improve cardiovascular health, and regulate inflammation in both children and adults [45]. In addition, exercise promotes improvement in functional capacity, which is also reduced in the pediatric population with obesity [46].

Traditional tests and measures of functional capacity evaluate different aspects of physical activity and exercise tolerance; some are subjective (eg, evaluation scales), while others provide objective or quantitative data. Among the objective tests, the 6MWT [47] is considered a simple field-based approximation of submaximal exercise capacity, whereas cardiopulmonary exercise testing represents the gold standard for assessing maximal aerobic capacity but is expensive and time-consuming and thus reserved for selected cases [48,49]. In particular, the 6MWT evaluates exercise tolerance and consists of making the patient walk at a self-paced speed up and down a 30-m-long corridor for a fixed period of 6 minutes [47], and the total distance traveled in meters (6MWT distance or 6-minute walk test distance) is the primary result of this test.

Intermediate approaches such as exercise electrocardiogram stress testing [50] offer additional physiological insight but require clinical supervision. It is also worth noting that in nonclinical or educational contexts, such as school-based physical activity programs, functional capacity is often assessed using simpler field tests (eg, shuttle run or step tests), which provide practical but less accurate estimates of exercise tolerance than laboratory-based evaluations (eg, 6MWT and cardiopulmonary exercise testing) [51].

Step 2: Ethics Committee Approval

The second step consists of obtaining approval to collect data with adult-designed wearables in a cohort of children chosen based on the specific clinical case, which has already been defined in step 1. Approval from an ethics committee or institutional review board is required, and compliance with local laws on the privacy of minors is up to the investigators. To obtain this approval, it is necessary to outline the cohort and the experimental procedures to be followed. The same dataset is generally used for more than one research aim.

In this case, data were obtained from a cohort of children with obesity studied before and after a 12-week online exercise training program, the clinical outcomes of which were published by Calcaterra et al [52]. The study was approved by the institutional ethics committee (Milano Area 1 protocol number 2020/ST/298, approval date December 2, 2020); the study was performed in line with the principles of the Declaration of Helsinki. Written informed consent for participation and continuous wearable monitoring was obtained from the parents or legal guardians of all children. All data were pseudonymized before analysis, stored on secure institutional servers, and processed in full compliance with the General Data Protection Regulation (EU GDPR 2016/679). Given that the dataset includes health- and activity-related data from children (a vulnerable population), sharing individual-level data is subject to strict legal and ethical safeguards under the GDPR and the Italian Personal Data Protection Code. Owing to these restrictions and because this paper presents a tutorial and is not intended as a clinical validation, individual-level data will not be publicly shared.

Step 3: Experimental Data Collection

The third step consists of collecting experimental data, which come not only from the wearable device but also from different sources.

In this application, 2 device models (Fitbit Charge 2 and Charge 3) were used for wearable data collection; assignment to participants was determined by device availability and was not logged at the individual level.

XSL•FO
RenderX

Personal data used in this study included age (computed in months), sex, standing height, weight, BMI, and BMI *z* score (as explained later). A Harpenden stadiometer (Holtain Ltd) with a fixed vertical backboard and an adjustable headpiece was used to measure standing height. Weight was measured with the patients standing upright at the center of the scale platform (Seca) without shoes and wearing light clothing. They faced the recorder, kept their hands at their sides, and looked straight ahead.

The clinical outcomes evaluated in this study were the 6MWT and its *z* score with respect to the normative data reported by Geiger et al [53] and the percentiles of systolic arterial pressure (SAP) and diastolic arterial pressure (DAP). The latter two were measured in the supine position using an electronic mercury sphygmomanometer (A&D Medical) with an appropriately sized cuff on the right arm after 5 minutes of rest.

Step 4: Aim of the Research Work

As mentioned in step 2, the same dataset is generally used for more than one research aim. For this reason, a specific research project often requires further selection of the data collected that are needed to be retained for that specific aim.

Of the wearable data available in this study, only those regarding the first week of the program were selected and compared with clinical outcomes in the research work, so that the values could be considered as baseline because the effect of the exercise training program would be negligible after 1 week.

Conversely, this tutorial goes deeper into showing the opportunities that can be exploited by using the whole activity and cardiovascular dataset, without necessarily having used them in this specific research work. Additionally, the tutorial explains how to deal with anthropometric data in the case of children. Specifically, this tutorial is primarily intended for researchers, biomedical engineers, and clinicians interested in the scientific use of commercial wearables for pediatric monitoring, specifically for monitoring school-aged children. Its purpose is not to provide consumer guidance but to illustrate, step-by-step, how to extract, preprocess, and interpret wearable-derived physiological and activity data in a research context. By combining practical technical explanations (eg, data file structure and feature computation) with clinical interpretation, the tutorial aims to bridge the gap between data science and pediatric health care applications.

Step 5: *z* Score and Percentile Data in Children

In the case of pediatric patients, it must be highlighted that parameters are generally expressed relative to normative data for a given age and sex. This is mostly expressed in terms of *z* score or percentile. The fifth step consists of adjusting all relevant parameters with respect to normative data so that children of different ages and sexes can still be compared and analyzed together.

Age was used in this study to obtain relevant reference data for the computation of *z* scores and percentiles. When the mean of the sample μ and the SD σ are made available, equation (1) is used to compute the *z* score:

$$(1) z=x-\mu\sigma$$

Here, z denotes the *z* score, and x represents the observed value.

When μ is available, but not σ, z can be computed by knowing the SE of the mean and the numerosity of the sample n (ie, the number of people from which data were collected), equation (2) can be used instead:

$$(2) z=x-\mu SE\cdot n$$

To obtain the BMI *z* score, the BMI was then converted into its *z* scores using World Health Organization reference values [54]. In this scenario, because μ and σ are provided as reference values, equation (1) can be applied.

Resting PR (RPR) *z* scores are determined using national reference data from the 1999 - 2008 US population across all age groups [55]. In this case, μ, SE, and n are available, so equation (2) is used.

The blood pressure percentile was determined for each child following recent guidelines [56-58].

Step 6: Device Characteristics

The previously listed steps can be applied to data from all adult-designed wearables. The tutorial can also be used in the case of other devices collecting similar variables but requires more adaptations. The sixth step consists of knowing specific model characteristics and data formats, which may present slight changes even in the case of subsequent versions of the same device. It is always important to thoroughly consult all available documents and manuals before starting to analyze data.

Although some technical details of the considered devices are publicly available and downloadable from the company's website, a concise summary is included here to guide researchers in identifying which data files and parameters are most relevant for pediatric monitoring studies. The purpose of this section is not to reproduce the device manual but to provide practical instructions on how to locate, interpret, and preprocess key Fitbit outputs for subsequent scientific analysis. Although Fitbit labels its optical measurement as HR in device interfaces and documentation, this value is in fact a PR estimate derived from a peripheral PPG sensor; the use of the term "HR" reflects commercial and user-facing conventions rather than physiological accuracy. For this reason, throughout this tutorial, we refer to this signal as PR when we talk about the physiological variable, while we maintain the original nomenclature when we refer to the file or the Fitbit-computed parameter (eg, in the case of HRZ). We refer to PR over steps as PROS and to resting PR as RPR, as previously defined.

While some parameters—such as PR and step counts—are calculated directly by the Fitbit, other metrics can be derived or estimated from the downloadable data. This section is dedicated to the parameters that can be extracted directly from the Fitbit device.

Throughout this protocol, the Fitbit Charge 2 [59] and Charge 3 [60] were used interchangeably as stated before in the manuscript. These models were selected because they have been validated previously in several peer-reviewed studies assessing the accuracy of PPG-based PR and step count estimation [61,62], ensuring data comparability with existing literature.

XSL•FO
RenderX

Furthermore, they were the devices available to our research group at the time of data collection (in 2021), which allowed maintaining consistency across participants.

Although we used both the Fitbit Charge 2 and Charge 3 devices and did not perform a direct head-to-head comparison, existing independent validation studies show that the Charge 2 provides acceptable PR measurement versus ECG in 24-hour ecological recordings in adult populations (mean bias −3.47 bpm) [63], while the Charge 3 showed a larger bias (mean bias −7 bpm) but in a protocol focused on more dynamic activities when compared to a chest - strap reference [64]. On this basis, one can assume a functional equivalence of the used models for the primary metrics (step count and PR), acknowledging that minor differences in sensor or firmware may exist. Newer Fitbit generations were not used to avoid potential discrepancies arising from firmware or algorithmic updates that could compromise data uniformity and comparability with validated models.

Both devices are equipped with a micro-electromechanical system 3-axis accelerometer to track motion patterns, an altimeter to track altitude changes, and an optical PR tracker based on PPG. As stated previously, in Fitbit documentation and app interfaces, this parameter is referred to as HR for commercial and communication purposes; however, as it is derived from peripheral PPG rather than ECG, it technically represents PR rather than HR [65].

Both devices contain a Bluetooth 4.0 radio transceiver to send data to a mobile app and can store 7 days of minute-by-minute statistics and 30 days of summary totals.

Step 7: Import Data

The seventh step consists of a first selection of the files to be imported or discarded and of importing the files of interest for the analysis.

Different files are available for analysis for each patient, all in .json format. Data structures vary between files: some contain time series of specific parameters with associated time stamps, others contain daily recaps, and still others contain weekly recaps. Synchronization between different files was performed using the time stamps in the files that were downloaded from the Fitbit device.

Before importing the data, the first essential filtering step is required. Any file whose content depends on age-based proprietary algorithms must be excluded because Fitbit profiles must be set to 18 years or older, and several derived metrics (eg, sleep stages, sleep score, caloric expenditure, and V˙O2 max–related fitness scores) are computed using adult-specific thresholds and models. For this reason, files related to sleep analysis, HRZ, calories, or fitness level should be discarded, as their values would be systematically biased in a pediatric cohort. More details regarding discarded metrics are reported in Figure 1.

Providing a fully reproducible data import script is not useful to readers because Fitbit intraday files do not follow a fixed header structure; field names differ across device models, firmware versions, application programming interface generations and types (eg, consumer vs research interface), and, more recently, Google-managed data export formats, requiring manual inspection and adjustment at import. More details on this can be found in the Data Import section of Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. Flowchart illustrating the process from all wearable data available from Fitbit to the list of retained variables and the variables obtained with further signal processing (highlighted in purple). The purple tab is dedicated to the additional processing pipeline, that is, to explain how additional variables can be obtained. HRZ: heart rate zone; PR: pulse rate; PROS: pulse rate over steps; RPR: resting pulse rate.

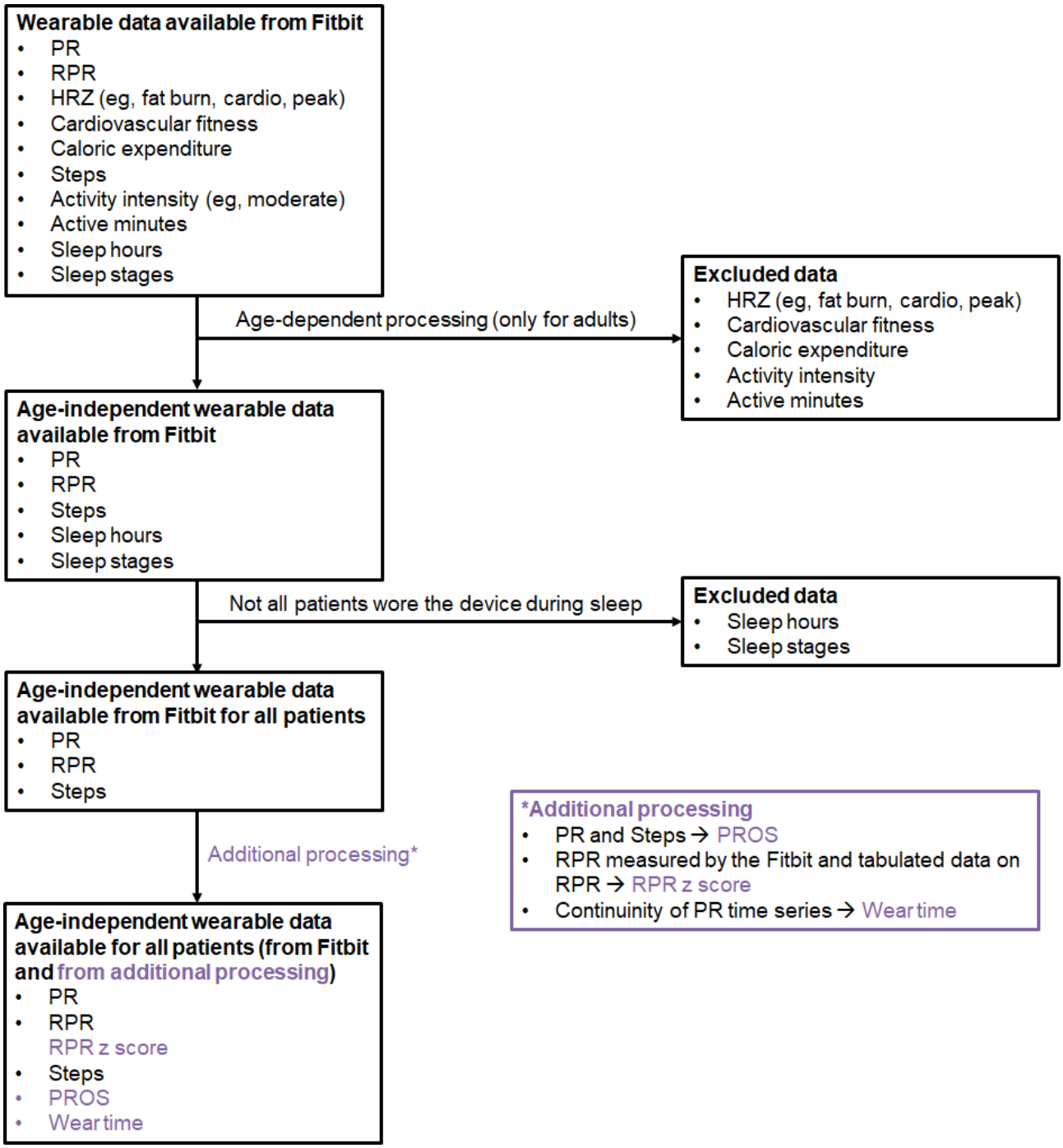

Step 8: Activity Metrics

The eighth step consists of exploring files with parameters related to the activity and selecting age-independent activity-related metrics.

"Activities" File and Derived

The "activities" file contains a daily summary of Fitbit activity data and goals, covering metrics such as steps, floors, active minutes, distance, and calories burned. The main fields include "activities," "goals," and "summary."

Specifically, the field "activities" is an array listing any detailed logged activities. If this array is empty, this indicates that there were no manually logged activities on that day.

The field "goals" describes the user's target goal for the day in terms of "activeMinutes" (target active minutes), "caloriesOut" (target calories to be burned), "distance" (target distance), "floors" (target floors), and "steps" (target steps).

The field "summary" reports the actual daily data, specifically "activeScore" (an internal scoring metric used by Fitbit), "activityCalories" (calories burned during active minutes), "caloriesBMR" (basal metabolic rate—calories burned at rest),

XSL•FO
RenderX

"caloriesOut" (total calories burned, including basal metabolic rate and activity), "distances" (an array of distance values broken down by activity levels), "elevation" (floors climbed in terms of elevation gain; may be 0 if not measured), "fairlyActiveMinutes" (minutes of moderate activity), "floors" (floors climbed), "heartRateZones" (an array showing the time spent in each HRZ—out of range, fat burn, cardio, and peak—along with estimated calories burned in each zone), "lightlyActiveMinutes" (minutes of light activity), "marginalCalories" (extra calories burned while being lightly active), "sedentaryMinutes" (minutes spent in a sedentary state), "steps" (total number of steps taken), and "veryActiveMinutes" (minutes of very active exercise).

All the fields regarding the count of calories ("activeCalories," "caloriesOut," "caloriesBMR," and "marginalCalories"), as well as the "calories" file, contain estimations of caloric consumption that are likely based on age, sex, weight, and height used as input in the Fitbit app. As the Fitbit device is designed for adults, it was only possible to set the age to 18 years or older to have the app working. For this reason, these metrics cannot be considered reliable in the present application, as already mentioned in step 7. The weekly recap files associated with these variables must also be excluded from the analysis.

The same can be said for the information related to the intensity of the activities based on HRZ values, which are still the subject of research in adults and are computed considering the maximum HR for a given age (which was set as 18 y in the application). In addition, in this case, the weekly recaps have been excluded.

For this reason, the only reliable metric is "steps." The number of walked steps is computed using the 3-axis accelerometer embedded in the Fitbit with dedicated processing algorithms. A traditional threshold of 10,000 daily steps is well accepted as indicative of an active lifestyle in adults [66]. To evaluate patient performance, one can consider either the mean daily steps or the maximum daily steps, representing the highest number of steps achieved during the test period. The latter is more representative of the ability to walk a certain distance [67].

The file "range_activities_steps" reports the total number of daily steps for a week. Visualizing this information, extracted either from the daily metrics or from the weekly recap files, can be used to immediately visualize behavioral shifts of patients. Figure 2 shows the distribution of daily steps in the 12 weeks of the intervention in the case of one patient. The higher step counts observed during weeks 7 to 9 likely coincide with the central weeks of August, when many families in Italy are on summer vacation, possibly leading to increased physical activity; however, this interpretation remains speculative.

Figure 2. Weekly steps distribution over the 12-wk online exercise training program (boy, aged 10 y; BMI=23.7 kg/m^2; BMI *z* score=1.74).

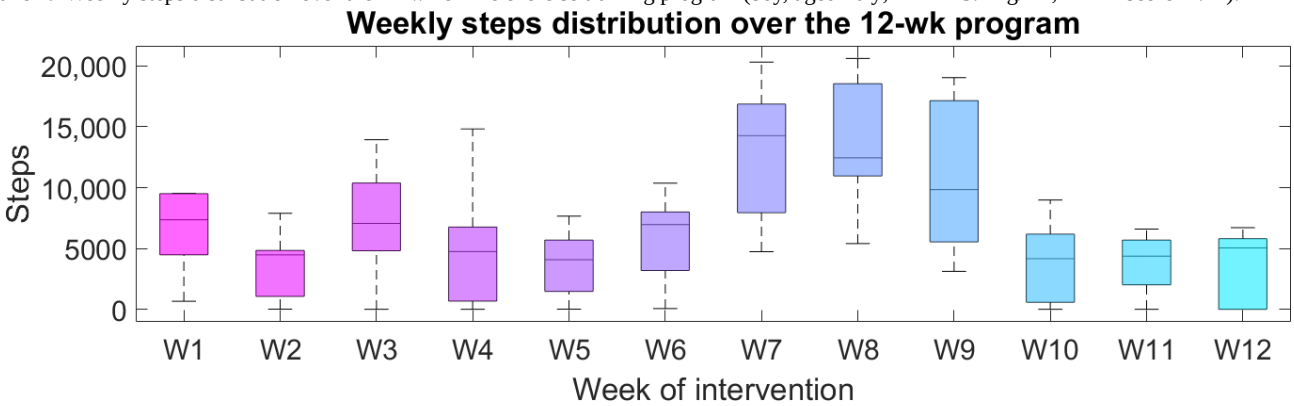

"Steps" File

Unlike the steps derived from the "activities" files family, the daily "steps" file shows the distribution of the steps during each day, in addition to the total number of steps. This allows to see when steps were taken during the day and better track the times of day when there has been the most activity. Examples of this can be seen in Figure 3.

In the considered clinical case example, there were 3 cases when an overall daily step count was provided, but no data on the distribution of the steps. This lack of detailed information on the distribution prevents the computation of other parameters (PROS, explained in a subsequent section).

XSL•FO
RenderX

Figure 3. Example of pulse rate (PR), steps, and pulse rate over steps (PROS) obtained in 1 wk of acquisitions (boy, aged 12 y; BMI=27.1 kg/m^2; BMI *z* score=1.86). Areas with the gray background correspond to night hours (22:00-06:59), while areas with the white background correspond to day hours (07:00-21:59). bpm: beats per minute.

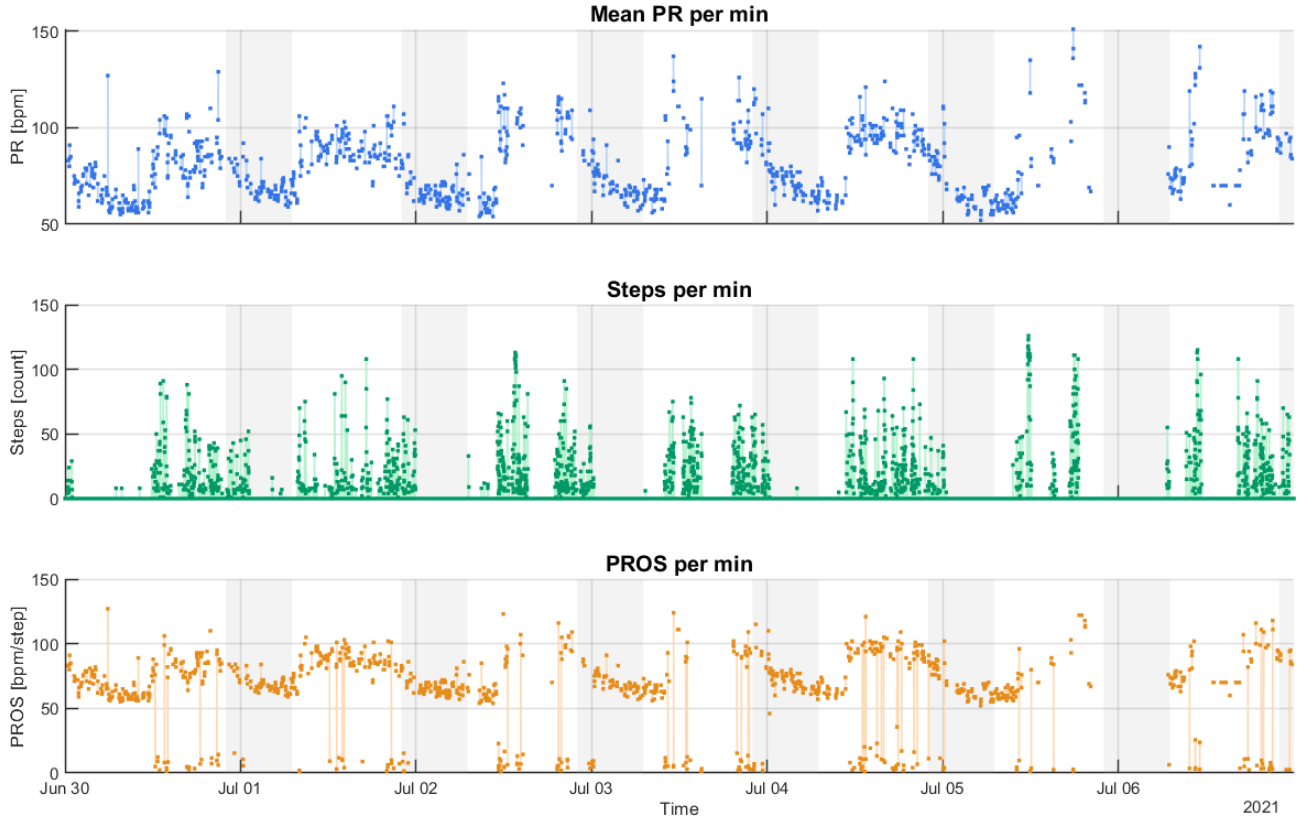

Step 9: Cardiovascular Metrics

The ninth step consists of exploring files with parameters related to the cardiovascular metrics and selecting age-independent metrics. As was mentioned earlier, some data, such as RPR, need to be adjusted for age and sex even if acquired from the wearable devices; in all relevant cases, step 5 should be repeated.

PR is continuously estimated via the embedded PPG sensor, with the raw data processed by an internal proprietary algorithm [67].

"hr" File

The daily "hr" file contains information regarding the already processed PR. There is an array ("activities-heart") containing the summary of the daily information, and then there is the "activities-heart-intraday" array with more details. This array has "dataset" as the primary subfield, which is an array of time-stamped entries; each entry has time and value. Additionally, there are 2 metadata fields, "datasetInterval" and "datasetType," which describe how the intraday data are structured.

PR measurement is affected by the location of the device and movement, as both factors can alter optical coupling, blood perfusion, and motion-induced artifacts, leading to transient inaccuracies or missing data—limitations common to all wrist-based PPG systems [68]. Furthermore, data are available at different sampling rates, with an interval in the order of seconds, one from the other at rest, and an increased sampling rate when exercise is detected. RPR is calculated over a 24-hour period and is known from the literature to reach its lowest value at night [69]. Consequently, nighttime RPR values are directly comparable across patients, whereas daytime values may not reflect the true minimum.

PR variability (PRV) is highly correlated with the HR variability, which quantifies the extent to which the R-R interval or HR changes from one cardiac cycle to the next. This measure can provide a lot of information about how the autonomic nervous system regulates the heart. HR variability has been correlated with cardiac diseases [70], fitness, and functional reserve [71] and is positively affected by exercise [72]. During our study, no PRV values were made available for download, nor could they be derived as no raw signal was given. As we reported in a previous study [67], PR data retrieved from a Fitbit device presented considerable gaps in the recordings, even for continuous recordings. As stated by the manufacturer, the normal sampling period should be between 5 and 15 seconds; however, in the results of that study, lack of data for 15 seconds or more was common. This makes it challenging to compute PRV reliably, so this metric should be analyzed with caution if obtained from this type of device.

Fitbit Charge 2 and 3 also provide information regarding HRZ and consequently the number of minutes spent in each zone. In adults, these data can be used to estimate the physical activity level of the person. However, in the present pediatric application, their reliability cannot be verified because the underlying proprietary algorithms were developed and validated on adult populations, and their exact computation methods are not publicly available.

XSL•FO
RenderX

Step 10: Combining Activity and Cardiovascular Metrics

The tenth step consists of combining activity and cardiovascular metrics. The parameters computed in this work were selected based on the clinical case and research aims, but other parameters can be considered as well.

Combining the metrics obtained from the activity and the cardiovascular domain is possible, and one example is reported in this tutorial. Mishra et al [73] used a Fitbit device to detect SARS-CoV-2 infection before symptom onset. To achieve this, the authors introduced and calculated the HR over steps index to identify potential anomalies in cardiac response relative to physical activity. In this study, this index is referred to as PROS to maintain homogeneity with the rest of the nomenclature.

PROS is computed as the instantaneous ratio between PR and the number of steps recorded within the corresponding time window. As PR is sampled approximately every 5 seconds and steps are aggregated per minute, each minute-level step count was matched to all PR values recorded within that same minute. To avoid division by zero during sedentary periods, 1 step was added to the denominator. The resulting value can therefore be expressed as in equation (3):

$$\text{PROS}(t)=\frac{\text{PR}(t)}{\text{steps}(t)+1} \quad (3)$$

where PR(t) represents PR at time t (beats per minute) and steps(t) is the number of steps detected within that minute. PROS provides a minute-by-minute estimate of the patient's cardiovascular effort and functional capacity. A high PROS value indicates a rapid heartbeat during moderate physical activity, while a low value suggests a restrained heartbeat despite challenging physical activity.

Mishra et al [73] developed a structured algorithm for computing PROS and identifying anomalous values, leveraging a dispersion-based approach that detects deviations from a normal distribution using the elliptic envelope method. The code used to process the data was retrieved from the GitHub repository associated with the original study. The anomalies are detected using a contamination of 0.1, where the contamination means the number of outliers in the dataset. In this manner, the algorithm identifies some outliers in the computed feature, corresponding to approximately 10% of the total. Following the method in the original paper, an iterative approach was used, incrementally increasing contamination values until a stable number of outliers was identified.

A possibility is to evaluate PROS at low quantiles, such as the 1% or the 5% quantile, to account for the possibility that the minimum PROS is an outlier.

As was mentioned before in the section dedicated to steps, it was not possible to compute PROS in 3 patients, as Fitbit provided the overall step count but not the details of the distribution of the steps during the day.

Step 11: Data Quality Assessment

The eleventh step of this tutorial consists of assessing the quality of the collected data, especially in terms of completeness of the dataset. Wearable data collection is unsupervised, and for this reason, patient adherence to the data collection protocol, that is, wear time, must be carefully evaluated in each study. In addition, signal loss is an important parameter to be evaluated, as explained later in the section.

To assess signal completeness and compliance, wear time can be quantified based on PR data continuity. In this dataset, the time intervals between consecutive PR samples were computed for each participant; intervals of 2 minutes or less were considered "worn," while longer gaps indicated device removal. Wear time was then expressed as the total duration of "worn" periods divided by the total duration of the analyzed time window. The 2-minute PR gap threshold was selected as a conservative compromise, as short interruptions of a few seconds are common even when the device is correctly worn. Adult Fitbit studies have proposed different definitions: Dominick et al [74] considered 60 minutes or more of missing signal as nonwear; Orstad et al [75] allowed up to 90 minutes of interruption to consider a day valid (with ≥10 h of recording including the interruption); Claudel et al [76] used minute-level PR gaps to classify nonwear episodes. Given the higher motion and frequent repositioning typical of children, the 2-minute criterion minimizes false nonwear detection while maintaining sensitivity to genuine removal periods.

To quantify how this threshold influenced data retention, it is advised to perform a sensitivity analysis. We performed such calculations using all 16 participants and recalculated wear time with nonwear gap thresholds of 60, 90, 120, 180, and 300 seconds. Weekly wear time changed only minimally across thresholds (60 s: 100.8 h; 90 s: 100.9 h; 120 s: 101.1 h; 180 s: 101.4 h; and 300 s: 101.8 h), with differences never exceeding 1.0 hour per week on average. Similarly, the percentage of days achieving 70% or greater completeness remained constant at 41% for all thresholds. These results indicate that the 120-second rule does not artificially inflate wear time estimates and that more permissive thresholds (eg, 180‑300 s) offer no meaningful advantage in terms of data retention in this cohort. Day and night were defined as 07:00 to 21:59 and 22:00 to 06:59, respectively, to align with typical waking and sleeping schedules in school-aged children. The theoretical acquisition window corresponded to 168 hours per week, split into approximately 105 hours daytime and 63 hours nighttime.

In addition to wear time, a second essential component of data quality is the assessment of signal loss, which reflects gaps in acquisitions even when the device is being worn. Unlike nonwear episodes, signal loss is typically caused by transient motion artifacts, insufficient optical coupling, or abrupt changes in perfusion—phenomena that are especially common in children due to rapid, irregular movements. To quantify signal loss, it is advised to examine the distribution of intersample intervals across the PR time series and identify gaps that exceed the expected sampling pattern (typically 5‑15 s at rest, with denser sampling during activity). These gaps should be flagged as within-wear missing segments, which differ clinically and analytically from nonwear periods. For example, brief clusters of missing samples may be acceptable when computing aggregated daily features (eg, daily steps and mean PR) but can severely affect features that rely on continuous sampling, such as PROS, nocturnal minima, or any attempt to estimate PR

XSL•FO
RenderX

variability. Figure 1 reports a flow diagram of the Fitbit-derived metrics retained or excluded from the analysis, the reasons for exclusion, a list of the metrics that have been computed with additional processing, and a tab dedicated to how additional processing is performed for each variable.

Step 12: Statistical Analysis

The final step is the statistical analysis and is again generalizable like steps 1 to 5. First, traditional statistical metrics such as mean and SD should be computed for each parameter, possibly further divided into subpopulations as done in Table 1, to characterize the population. Then, further steps of statistical analysis must be tailored to the specifics of the study.

In this clinical case, correlation analyses were performed to obtain the Pearson correlation coefficient ρ, the coefficient of determination R^2, and the P value; then a multiple linear regression analysis [77] was conducted to predict the 6 clinical outcomes listed in Textbox 1. Both the raw resting RPR and its age- and sex-adjusted z score were included as predictors in the regression models. Although these variables are mathematically related, they were retained together to explore whether absolute RPR and its standardized deviation from pediatric norms contribute differently to the prediction of functional outcomes. This approach was adopted intentionally to capture both absolute physiological values and relative, age-adjusted variability, acknowledging that some degree of correlation between the two variables may occur.

To mitigate overfitting due to the small sample size, an additional model selection analysis was conducted using leave-one-out cross-validation (LOOCV) combined with least absolute shrinkage and selection operator (LASSO) regularization. This approach automatically shrinks or removes redundant predictors, allowing a more parsimonious set of variables to emerge for each clinical outcome. The models were then refitted using only the selected predictors to verify their explanatory power.

Table . Summary of demographic, cardiovascular, and physical activity metrics of the whole cohort (N=16) and then divided by boys (n_{boys}=11) and girls (n_{girls}=5).

Parameter	Overall (N=16), mean (SD)	Boys (n_{boys}=11), mean (SD)	Girls (n_{girls}=5), mean (SD)
Age (y)	10.2 (1.7)	10.2 (1.3)	10.2 (2.6)
BMI (kg/m²)	28.1 (3.7)	27.1 (3.0)	30.2 (4.5)
BMI z score	2.2 (0.3)	2.1 (0.3)	2.4 (0.4)
RPR[a] (bpm)	79.4 (9.4)	79.3 (8.9)	79.5 (11.3)
RPR z score	−0.06 (0.45)	−0.03 (0.47)	−0.15 (0.42)
SAP[b] (mm Hg)	111.7 (11.1)	109.9 (12.6)	116.5 (2.4)
SAP percentile (%)	76 (27)	72 (30)	88 (8)
DAP[c] (mm Hg)	73.1 (8)	71.5 (8)	77.5 (7)
DAP percentile (%)	82 (16)	79 (17)	91 (10)
Daily steps	7742 (3025)	7838 (3629)	7530 (1148)
Maximum daily steps	12,895 (4964)	13,279 (5873)	12,050 (2210)
6MWT[d] distance (m)	490.5 (55.0)	487.2 (55.6)	517.0 (50.4)
6MWT z score	−2.7 (1.1)	−3.0 (0.9)	−2.0 (1.3)
Total wear time (weekly) (h)	101.1 (44.6)	105.5 (43.8)	91.4 (49.8)
Total wear time (weekly) (%)	60.2 (26.5)	62.8 (26.1)	54.4 (29.6)
Day wear time (weekly) (h)	72.3 (27.3)	73.8 (22.9)	68.8 (38.2)
Day wear time (weekly) (%)	68.8 (26.0)	70.3 (21.8)	65.5 (36.4)
Night wear time (weekly) (h)	28.8 (23.0)	31.6 (23.6)	22.6 (22.8)
Night wear time (weekly) (%)	45.7 (36.5)	50.2 (37.4)	35.9 (36.2)

[a]RPR: resting pulse rate
[b]SAP: systolic arterial pressure
[c]DAP: diastolic arterial pressure
[d]6MWT: 6-minute walk test

XSL•FO
RenderX

Textbox 1. Input data used for the analysis and the clinical outcomes of the presented clinical application.

Input data

- Personal data (anthropometric)
  - Age
  - BMI
  - BMI *z* score
- Wearable data (week 1)
  - 6-minute walk test
  - 6-minute walk test *z* score
  - Systolic arterial pressure
  - Systolic arterial pressure percentile
  - Diastolic arterial pressure
  - Diastolic arterial pressure percentile

Clinical outcomes

- Mean daily steps
- Maximum daily steps
- Resting pulse rate
- Resting pulse rate *z* score
- Pulse rate over steps

Results of the Considered Clinical Case

Population Characteristics

A summary of demographic, cardiovascular, and physical activity metrics for the cohort is presented in Table 1, including data divided by boys and girls. These values are expressed in terms of mean (SD).

Quantitative wear time statistics per participant can be reported in greater detail, as proposed in Table S1 in Multimedia Appendix 1.

Collected Data

An example of time series plots of PR, steps, and PROS obtained in a week of acquisitions is shown in Figure 3.

To further illustrate the temporal pattern of PR recordings and the occurrence of signal loss, Figure 4 shows a representative 15-minute segment extracted from the same week of acquisition plotted in Figure 3. This high-resolution view highlights the irregular sampling of PR data and the short interruptions occurring when the signal was not acquired by the Fitbit device.

XSL•FO
RenderX

Figure 4. Representative 15-min window of PR data from the same participant as in Figure 3 (boy, aged 12 y; BMI=27.1 kg/m^2, BMI *z* score=1.86). Blue dots represent individual PR samples, while light orange shaded areas indicate missing intervals where the time between consecutive samples exceeded 1.5×the expected sampling period. bpm: beats per minute.

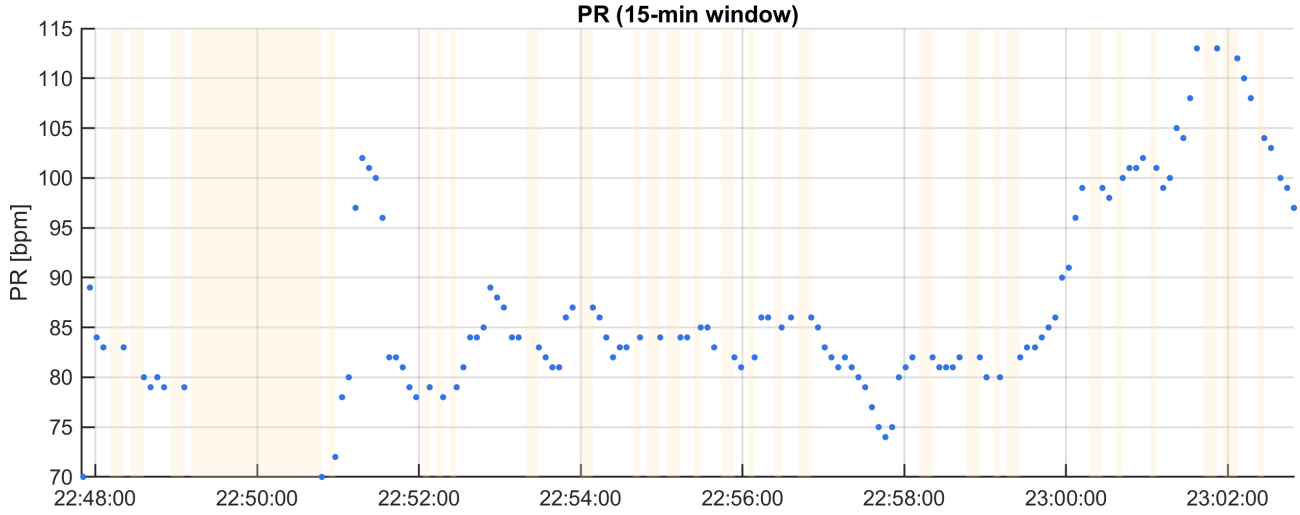

Statistical Analysis

Table 2 reports Pearson correlation coefficients (ρ) to explore the relationships among wearable data and clinical outcomes.

It is noteworthy that RPR *z* score has a significant correlation with 6MWT results (ρ=−0.53; R^2=0.28; *P*=.04), and it is a negative correlation as is expected from physiology. These results directly show the utility of wearable-derived data in clinical outcome assessment.

Two additional statistically significant correlations are observed between maximum weekly steps and the 6MWT (ρ=−0.52; R^2=0.27; *P*=.04) and between maximum weekly steps and the 6MWT *z* score (ρ=−0.60; R^2=0.36; *P*=.01); however, such results are not what one would intuitively expect, nor do they confirm what was obtained in a population of older patients [67]; therefore, it is likely a spurious correlation in both cases, as is discussed in more detail later in the paper.

The mean daily steps showed a moderate negative correlation with the DAP percentile (ρ=−0.42; R^2=0.18; *P*=.12) but did not reach statistical significance.

PROS displayed only weak correlations with functional capacity measures, indicating that this variable alone might have limited predictive value in this context. RPR and RPR *z* score showed a moderate, nonsignificant correlation with DAP (ρ=0.41; R^2=0.17; *P*=.13; and ρ=0.43; R^2=0.19; *P*=.12, respectively), and the RPR *z* score also showed a moderate correlation with DAP percentile (ρ=0.42; R^2=0.17; *P*=.14).

Multiple linear regression models were constructed to examine the associations between wearable data and clinical outcomes. Figure 5 shows the observed and the fitted values of each of the variables, with statistical significance defined as *P*<.05.

with The regression model for fitting 6MWT distance yielded an adjusted R^2 of 0.51, indicating a moderate fit of the model; the regression model fitting the 6MWT *z* score had a stronger fit (adjusted R^2=0.79), and, overall, the best performance among the 6 clinical outcomes considered in the study. In this case, 2 variables, that is, RPR and its *z* score, made a statistically significant contribution to the model (*P*=.03 and *P*=.02, respectively), indicating the strong dependence of the 6MWT *z* score on parameters relating to cardiac activity.

SAP, SAP percentile, DAP, and DAP performed poorly in the case of adjusted R^2, with the value being negative because there are too many variables. Given the small cohort size (n=16), all regression models should be considered underpowered; therefore, their estimates are intended to be illustrative of the methodological workflow and not to provide confirmatory inference. Still, all predictors in the full multivariable models, with coefficient, SE, *P* value, R^2, adjusted R^2, and n, are reported for each outcome in Table S2 of Multimedia Appendix 1.

The LASSO procedure identified age, BMI, maximum daily steps, and RPR *z* score as predictors of the 6MWT distance (adjusted R^2=0.29) and age, BMI *z* score, mean and maximum daily steps, RPR, RPR *z* score, and PROS for the 6MWT *z* score (adjusted R^2=0.75). Details on the predictors selected by LASSO with LOOCV for each outcome are reported in Table S3 of Multimedia Appendix 1. For blood pressure–related outcomes, the regularization excluded most predictors, retaining only maximum daily steps for systolic measures and BMI for diastolic pressure, both with negligible explanatory value (adjusted R^2≈0).

XSL•FO
RenderX

Table . Correlation of wearable data with clinical outcomes.

Variable	Clinical outcomes																	
	SAP[a]			SAP percentile			DAP[b]			DAP percentile			6MWT[c]			6MWT z score		
	ρ	R^2	*P* value	ρ	R^2	*P* value	ρ	R^2	*P* value	ρ	R^2	*P* value	ρ	R^2	*P* value	ρ	R^2	*P* value
Wearable data																		
Mean steps	−0.04	0.00	.90	−0.13	0.02	.64	−0.22	0.05	.44	−0.42	0.18	.12	0.07	0.01	.79	0.05	0.00	.85
Maximum steps	0.12	0.01	.67	0.14	0.02	.63	0.20	0.04	.47	0.16	0.03	.56	−0.52	0.27	*.04*[d]	−0.60	0.36	.01[d]
RPR[e]	0.15	0.03	.57	0.12	0.01	.67	0.41	0.17	.13	0.35	.12	.21	−0.38	0.14	.15	0.03	0.00	.90
RPR z score	0.24	0.06	.40	0.23	0.06	.44	0.43	0.19	.12	0.42	0.17	.14	*−0.53*	*0.28*	*.04*[d]	−0.34	0.12	.21
PROS[f]	−0.14	0.02	.66	−0.17	0.03	.59	0.08	0.01	.81	0.20	0.04	.52	0.20	0.04	.52	0.21	0.04	.49

[a]SAP: systolic arterial pressure.

[b]DAP: diastolic arterial pressure.

[c]6MWT: 6-min walk test.

[d]Statistically significant correlations.

[e]RPR: resting pulse rate.

[f]PROS: pulse rate over steps.

These results confirm that, in this limited cohort, 6MWT and 6MWT *z* score remain the only variables moderately predictable from wearable-derived metrics, while SAP, SAP percentile, DAP, and DAP percentile show minimal dependence on these predictors. The analysis demonstrates that the negative adjusted R^2 values observed in the unregularized multiple regressions shown in Figure 5 are primarily due to sample size limitations rather than methodological flaws.

XSL•FO
RenderX

Figure 5. Observed versus fitted clinical outcomes (6-min walk test [6MWT], 6MWT *z* score, systolic arterial pressure [SAP], SAP percentile, diastolic arterial pressure [DAP], and DAP percentile). The scatter plot illustrates the relationship between observed and fitted distances, with the red dashed line indicating an ideal fit.

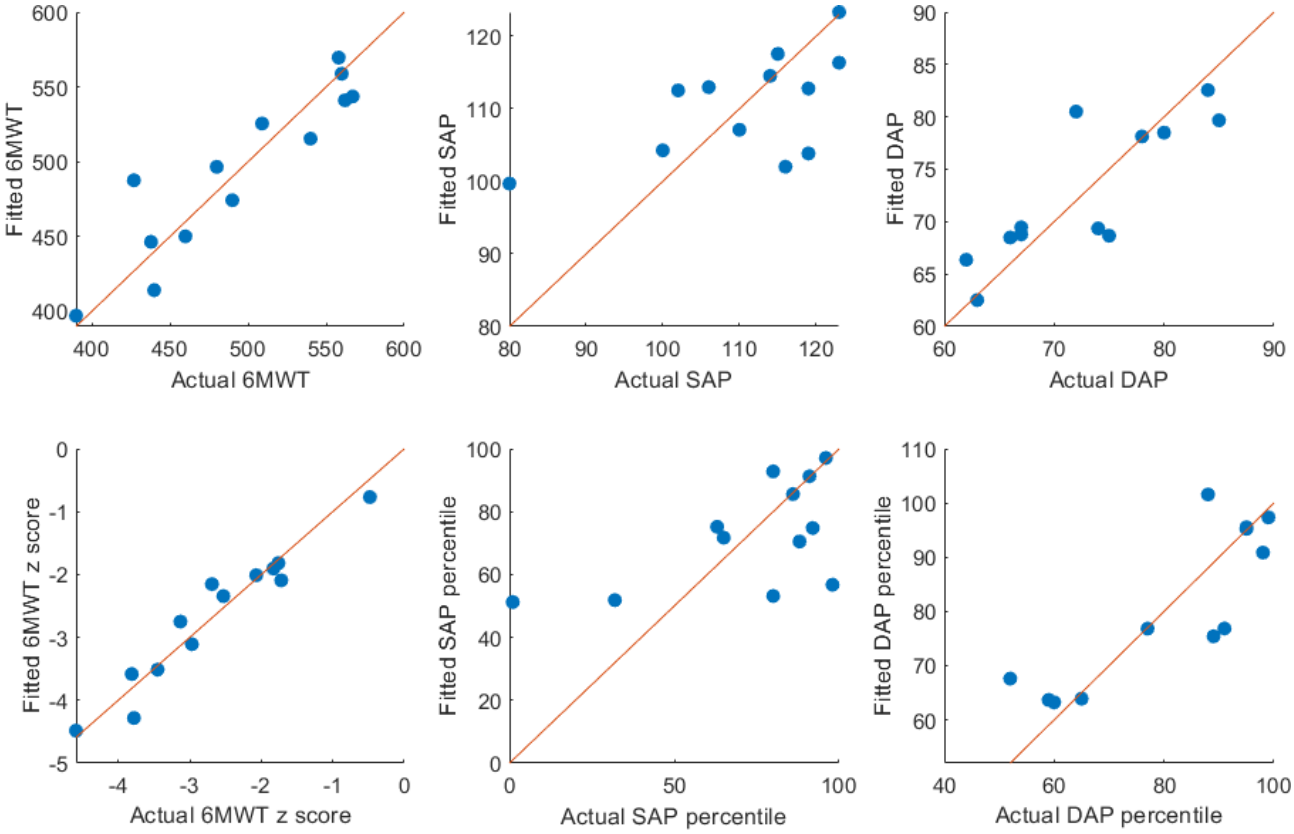

Discussion

This tutorial explains step by step how to process and further engineer activity and cardiovascular data obtained from Fitbit when used in pediatric patients and presents a clinical application in children with obesity.

Study-Specific Findings

The data gathered from Fitbit Charge 2 and Charge 3 in children with obesity demonstrate that consumer-grade wearable devices can provide clinically relevant information about physical activity (eg, mean or maximum daily steps) and cardiovascular function (eg, RPR and PROS) in this population. As shown, these metrics alone do not exhibit strong correlations with standard reference tests (eg, the 6MWT in children or systolic or diastolic blood pressure percentiles). However, by supplementing them with anthropometric indicators and appropriate pediatric reference values, a more robust framework can be built for understanding functional capacity and exercise tolerance. Such results are also expected to improve with larger datasets available. Among the observed statistically significant associations, there were inverse correlations between maximum daily steps and both the 6MWT distance and its *z* score. These associations should be interpreted as exploratory, as the analysis is constrained by the small sample size (n=16), which limits statistical power and the generalizability of the findings. Additionally, the counterintuitive finding of having inverse correlations may have several explanations. First, given the small sample size and multiple comparisons, a spurious correlation cannot be excluded. Second, the 6MWT reflects submaximal exercise tolerance measured under clinical supervision, whereas maximum daily step counts reflect spontaneous physical activity in free-living conditions. Children with lower functional capacity may have been more engaged to increase their daily steps during the monitoring week, potentially due to parental encouragement after clinical evaluation. Measurement variability may also contribute: step count accuracy in children can be affected by shorter stride length, irregular gait, and differences in device placement, leading to over- or under-counting at certain speeds. Additionally, the week chosen for wearable data collection may not coincide with typical activity patterns—for example, periods of school holidays or illness could bias step count data independently of actual fitness.

Overall, our findings indicate that consumer-grade wearable devices, such as the Fitbit Charge 2 and Charge 3, can provide meaningful information on both physical activity and cardiovascular parameters in children with obesity. These results suggest a potential role for such devices as complementary tools in clinical and research settings to monitor behavior and physiological responses outside the laboratory environment, hence the need to provide practical guidelines and indications on possible data processing pipelines.

An additional limitation of our work stems from the fact that the Fitbit app enforces a minimum user age of 18 years or older, which inevitably introduces bias in the estimation of HRZ and caloric output and led us to discard such parameters. As the algorithmic assignment of HRZ is age-based (often through formulas such as “220 − age”), pediatric users are forced to

XSL•FO
RenderX

adopt adult parameterization. Consequently, derived metrics such as time in “cardio” or “peak” zones and estimated energy expenditure are likely to be inaccurate when applied to cohorts of school-aged children. Further complicating matters, children’s basal metabolic rates and activity energy expenditure differ systematically from adult models (eg, children exhibit relatively higher mass-specific metabolic rates) [78]. Thus, caloric estimates based on adult-tuned algorithms may misrepresent true energy usage in minors.

Setting all profiles to 18 years or older affects multiple Fitbit-derived metrics beyond calories. Cardiovascular fitness score (a proxy of $\dot{V}O_2$ max) is explicitly age-dependent and sex-dependent [79], so values in minors will be misestimated. Finally, activity classification and intensity inference are trained on adult patterns or rely on adult cutpoints [80]; studies show poor transfer to children and wide variability when applying adult or nonpediatric thresholds, motivating pediatric-specific models [81]. In practical terms, these biases limit the validity of continuous cardiovascular monitoring using standard Fitbit commercial pipelines in children: improvements or deteriorations in cardiovascular load may be masked or exaggerated.

Some issues appear easier to overcome than others; for instance, the lack of PRV data and the presence of inconsistencies in the sampling could be dealt with by using Fitbit smartwatches instead of activity trackers. It is, in fact, possible to design custom apps for Fitbit smartwatches, but not for trackers, by exploiting the software development kit available. Nevertheless, our experience suggests that, when used in a controlled protocol and combined with rigorous data cleaning, even simple trackers can yield reliable step counts and consistent PR data—even in a cohort of children.

A key limitation of this study is the small sample size (n=16), which substantially restricts the statistical power of correlation and regression analyses. With multiple predictors included in the models, the degrees of freedom are limited, increasing the likelihood of overfitting and unstable coefficient estimates. Accordingly, the regression outputs presented here primarily demonstrate the feasibility of integrating wearable-derived features into statistical models, rather than serving as robust or generalizable predictors. Owing to this, the absence of significant associations should be interpreted cautiously, as these results may reflect insufficient power rather than a true lack of relationship between wearable-derived and clinical variables. In addition, the negative adjusted R^2 values observed in several models, together with the presence of multicollinearity among personal and wearable data, indicate that the LASSO and LOOCV results should not be overinterpreted. With such a small sample and correlated predictors, penalized regression can select unstable subsets of variables, and cross-validation may provide overly optimistic estimates of predictive value. Future studies with larger, more heterogeneous samples will be required to confirm these findings, refine the regression models, and enhance generalizability.

While data acquisition from the Fitbit devices was successful in this cohort, this analysis did not assess broader aspects of feasibility, such as long-term adherence, cost, or regulatory considerations, nor are such aspects the focus of this tutorial.

General Considerations

The 12 steps of the tutorial can be adapted to other studies with Fitbit devices with minimal adjustments and with other devices by maintaining some of the steps (steps 1-5 and step 12) and adapting the remaining steps to the specific file format and available data. This pipeline could be easily replicated in other studies using similar devices. In a previous study conducted by our research group [67], we enrolled 31 older patients (mean age 76.1 years) to assess whether parameters that could be obtained from a Fitbit Inspire 2 device were correlated with 6MWT results. We obtained statistically significant correlation in the cases of nonexercise testing cardiorespiratory fitness (ρ=0.68; P<.001), RPR (ρ=−0.39; P=.03), HR over steps at 1% quantile (ρ=−0.39; P=.04), mean number of steps (ρ=0.59; P<.001), logarithmic transformation of the mean number of steps (ρ=0.56; P=.001), and $\dot{V}O_2$ max (ρ=0.58; P=.006).

On the basis of our findings across multiple studies, data from commercial wearable devices could offer a low-cost alternative to in-hospital assessments such as the 6MWT. This approach is particularly advantageous for patients who live far from the hospital or when hospital access is restricted, as was the case during the recent COVID-19 pandemic. Furthermore, the presented technology allows tracking one’s progress over time. Results obtained in children demonstrate that even models designed for adults could be useful; however, further considerations should be applied when devices are used on children, particularly for company-provided metrics that are not immediately connected to the raw values of parameters.

Fitbit Charge 2 and Charge 3 are low-cost, basic trackers. More recent models (eg, Fitbit Charge 6) can be purchased with less than €150 (about US $173) per unit. An advantage of such devices is that they are widespread, specifically among young people and thus already present in a large part of the population, constituting a mine of data that is constantly supplied with even more data. Beyond childhood obesity, wrist-worn PR activity workflows have relevance across multiple clinical domains. In congenital heart disease, an American Heart Association science advisory supports wearable devices for continuous home monitoring and outlines clinical, ethical, and integration considerations [82]; pediatric studies also show consumer smartwatches (in this case, the Apple Watch) can capture clinically relevant arrhythmias not detected by standard ambulatory monitors [83]. In respiratory disease, cystic fibrosis cohorts have derived individualized PR thresholds from activity trackers (the so-called HRZ) to quantify moderate-to-vigorous activity [84], while asthma programs using multimodal wearable home monitoring have shown feasibility for assessing compliance outside the clinic [85]. These represent only a few examples of possible clinical applications of wearables in school-aged children.

The main limitations of currently available commercial devices are that they generally do not provide raw data, and their metrics (such as the number of minutes of an activity of a given intensity) are computed based on adult data and a black box, generally artificial intelligence-based [86] model from the point

XSL•FO
RenderX

of view of the end user. This makes it difficult, if not impossible, to understand how the algorithm was designed and on which assumptions it is based, particularly because private companies do not disclose their algorithms. If this lack of transparency could still be acceptable in the case of adult populations, where the assumption holds that substantial amounts of data have been collected in the past by major companies, this is no longer true in the case of pediatric applications.

As noted in the state-of-the-art section, most commercial devices for children are focused on activity metrics, but there are no devices that perform cardiovascular measurements. Both raw data and dedicated studies on continuous monitoring of cardiovascular parameters would enhance our ability to evaluate data quality and accuracy.

Conclusions

This tutorial primarily aims to establish the methodological feasibility of using adult-oriented wearable devices for pediatric monitoring, rather than to test clinical hypotheses or generate clinical conclusions. Our findings show that, when appropriate preprocessing and data quality assessments are applied, low-cost wrist-worn wearables can provide some usable data in school-aged children. The clinical example illustrates how such methodological steps translate into practice, but the interpretation remains exploratory and is not intended as a validation study. Although consumer devices may not offer raw signals or entirely pediatric-calibrated metrics, the overall experience we describe indicates that step counting and PR data can still be leveraged in clinical or research contexts involving children. By applying careful methods to manage and analyze the data, clinicians and investigators can minimize inaccuracies and still benefit from remote monitoring capabilities.

To perform reliable clinical trials using wearables on the pediatric population, activity trackers have demonstrated their performance, while further research is needed to assess the accuracy of cardiovascular parameters. The methodology proposed in this tutorial, however, gives indications on how to process such data with caution and maintaining scientific rigor, which is necessary for applications in clinical trials.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT (Generative Artificial Intelligence Delegation Taxonomy) (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing. The GAI tool used was GPT-5.

Responsibility for the final manuscript lies entirely with the authors.

GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was funded by: (1) the PODiaCar Project 101128946-PODiaCar-EU4H-2022-PJ-3 co-funded by European Union; (2) the National Recovery and resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3-Call for proposal No. 341 of 15 March 2022 of Italian Ministry of University and Research funded by the European Union-NextGenerationEU. Project code PE00000003, Concession Decree No. 1550 of 11 October 2022 adopted by the Italian Ministry of University and Research, CUP D93C22000890001, Project title “ON Foods-Research and innovation network on food and nutrition Sustainability, Safety and Security-Working ON Foods”; (3) the National Plan for National Recovery and Resilience Plan (NRRP) Complementary Investments (PNC, established with the decree-law May 6, 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of the Research Initiatives for Technologies and Innovative Trajectories in the health and care sectors (Directorial Decree n. 931 of 06–06-2022)–AdvaNced Technologies for Hu-man-centrEd Medicine (ANTHEM) under Project PNC0000003. This work reflects only the authors’ views and opinions, neither the Ministry for University and Research nor European Commission can be considered responsible for them. The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data import and supplementary tables.
[DOCX File, 58 KB - jmir_v28i1e76166_app1.docx]

References

1. Laranjo L, Ding D, Heleno B, et al. Do smartphone applications and activity trackers increase physical activity in adults? Systematic review, meta-analysis and metaregression. Br J Sports Med 2021 Apr;55(8):422-432. [doi: 10.1136/bjsports-2020-102892]

XSL•FO
RenderX

2. Angelucci A, Aliverti A. Telemonitoring systems for respiratory patients: technological aspects. Pulmonology 2020 Jul 30;26(4):221-232. [doi: 10.1016/j.pulmoe.2019.11.006]
3. Aliverti A, Evangelisti M, Angelucci A. Wearable tech for long-distance runners. In: Canata GL, Jones H, Krutsch W, Thoreux P, Vascellari A, editors. The Running Athlete: A Comprehensive Overview of Running in Different Sports: Berlin, Heidelberg: Springer Berlin Heidelberg; 2022. [doi: 10.1007/978-3-662-65064-6_10]
4. Bernasconi S, Angelucci A, Aliverti A. A scoping review on wearable devices for environmental monitoring and their application for health and wellness. Sensors (Basel) 2022;22(16):5994. [doi: 10.3390/s22165994]
5. Curone D, Secco EL, Tognetti A, et al. Smart garments for emergency operators: the ProeTEX project. IEEE Trans Inform Technol Biomed 2010;14(3):694-701. [doi: 10.1109/TITB.2010.2045003]
6. Solbiati S, Charkas S, Mozzini F, et al. Comparison of ECG-free algorithms for heart rate computation from head-BCG signals obtained with smart eyewear. Presented at: 2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE); Oct 21-23, 2024; St Albans, United Kingdom p. 213-218. [doi: 10.1109/MetroXRAINE62247.2024.10795866]
7. Floris C, Solbiati S, Landreani F, et al. Feasibility of heart rate and respiratory rate estimation by inertial sensors embedded in a virtual reality headset. Sensors (Basel) 2020;20(24):7168. [doi: 10.3390/s20247168]
8. Kinnunen H, Rantanen A, Kenttä T, Koskimäki H. Feasible assessment of recovery and cardiovascular health: accuracy of nocturnal HR and HRV assessed via ring PPG in comparison to medical grade ECG. Physiol Meas 2020 May 7;41(4):MID. [doi: 10.1088/1361-6579/ab840a] [Medline: 32217820]
9. Chu M, Nguyen T, Pandey V, et al. Respiration rate and volume measurements using wearable strain sensors. NPJ Digital Med 2019;2(1):1-9. [doi: 10.1038/s41746-019-0083-3]
10. Angelucci A, Villamar OWP, Agostoni P, Aliverti A. Design and evaluation of a wearable single-lead ECG for continuous monitoring. Presented at: 2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE); Oct 25-27, 2023. [doi: 10.1109/MetroXRAINE58569.2023.10405813]
11. Angelucci A, Cavicchioli M, Cintorrino IA, et al. Smart textiles and sensorized garments for physiological monitoring: a review of available solutions and techniques. Sensors (Basel) 2021;21(3):814. [doi: 10.3390/s21030814]
12. Angelucci A, Kuller D, Aliverti A. A home telemedicine system for continuous respiratory monitoring. IEEE J Biomed Health Inform ;25(4):1247-1256. [doi: 10.1109/JBHI.2020.3012621]
13. Angelucci A, Aliverti A. Detrended fluctuation analysis of day and night breathing parameters from a wearable respiratory holter. Comput Biol Med 2025 Apr;188:109907. [doi: 10.1016/j.compbiomed.2025.109907]
14. Röddiger T, Wolffram D, Laubenstein D, Budde M, Beigl M. Towards respiration rate monitoring using an in-ear headphone inertial measurement unit. Presented at: UbiComp '19: The 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing; Sep 9, 2019; London United Kingdom p. 48-53. [doi: 10.1145/3345615.3361130]
15. García-Pinillos F, Roche-Seruendo LE, Marcén-Cinca N, Marco-Contreras LA, Latorre-Román PA. Absolute reliability and concurrent validity of the Stryd system for the assessment of running stride kinematics at different velocities. Journal of Strength & Conditioning Research 2021;35(1):78-84. [doi: 10.1519/JSC.0000000000002595]
16. Evenson KR, Goto MM, Furberg RD. Systematic review of the validity and reliability of consumer-wearable activity trackers. Int J Behav Nutr Phys Act 2015 Dec;12(1):1-22. [doi: 10.1186/s12966-015-0314-1]
17. Jachymek M, Jachymek MT, Kiedrowicz RM, Kaźmierczak J, Płońska-Gościniak E, Peregud-Pogorzelska M. Wristbands in home-based rehabilitation-validation of heart rate measurement. Sensors (Basel) 2021 Dec 23;22(1):60. [doi: 10.3390/s22010060] [Medline: 35009603]
18. Angelucci A, Canali S, Aliverti A. Digital technologies for step counting: between promises of reliability and risks of reductionism. Front Digit Health 2023;5:1330189. [doi: 10.3389/fdgth.2023.1330189] [Medline: 38152629]
19. Lyons K, Hei Man AH, Booth D, Rena G. Defining activity thresholds triggering a "stand hour" for Apple Watch Users: cross-sectional study. JMIR Form Res 2024 Jun 10;8(1):e53806. [doi: 10.2196/53806]
20. Angelucci A, Bernasconi S, D'Andrea M, et al. Integration of a body sensor network of wearable devices for cardio-respiratory monitoring. Presented at: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Jul 24-27, 2023; Sydney, Australia p. 1-4. [doi: 10.1109/EMBC40787.2023.10340495]
21. Antonelli A, Guilizzoni D, Angelucci A, et al. Comparison between the Airgo™ Device and a Metabolic Cart during Rest and Exercise. Sensors (Basel) 2020 Jul 15;20(14):3943. [doi: 10.3390/s20143943] [Medline: 32679882]
22. Angelucci A, Camuncoli F, Galli M, Aliverti A. A wearable system for respiratory signal filtering based on activity: a preliminary validation. Presented at: 2022 IEEE International Workshop on Sport, Technology and Research (STAR); Jul 6-8, 2022; Trento - Cavalese, Italy p. 19-23. [doi: 10.1109/STAR53492.2022.9860001]
23. Angelucci A, Camuncoli F, Dotti F, et al. A wireless body sensor network for cardiorespiratory monitoring during cycling. Presented at: 2023 IEEE International Workshop on Sport, Technology and Research (STAR); Sep 14-16, 2023; Cavalese - Trento, Italy p. 1-4. [doi: 10.1109/STAR58331.2023.10302438]
24. Angelucci A, Aliverti A. An IMU-based wearable system for respiratory rate estimation in static and dynamic conditions. Cardiovasc Eng Technol 2023 Jun;14(3):351-363. [doi: 10.1007/s13239-023-00657-3] [Medline: 36849621]
25. Angelucci A, Camuncoli F, Dotti F, et al. Validation of a body sensor network for cardiorespiratory monitoring during dynamic activities. Biocybern Biomed Eng 2024 Oct;44(4):794-803. [doi: 10.1016/j.bbe.2024.09.002]

XSL•FO
RenderX

26. Clinical care - Sibel Health, Inc. SIbEL. URL: https://sibelhealth.com/clinical-care/ [accessed 2025-10-14]
27. Chung HU, Rwei AY, Hourlier-Fargette A, et al. Skin-interfaced biosensors for advanced wireless physiological monitoring in neonatal and pediatric intensive-care units. Nat Med 2020 Mar;26(3):418-429. [doi: 10.1038/s41591-020-0792-9]
28. Tandon A, de Ferranti SD. Wearable biosensors in pediatric cardiovascular disease. Circulation 2019 Jul 30;140(5):350-352. [doi: 10.1161/CIRCULATIONAHA.119.038483]
29. Peng DM, Edelson JB, Gureddygari HR, et al. Capturing physiological data in children with heart failure using wearable digital technology: lessons from pilot project. CJC Pediatric and Congenital Heart Disease 2025 Aug;4(4):226-230. [doi: 10.1016/j.cjcpc.2025.03.007]
30. Drummond CK, Tandon A. Advancing wearable technology for monitoring heart activity in paediatric populations. CJC Pediatric and Congenital Heart Disease 2023 Aug;2(4):196-197. [doi: 10.1016/j.cjcpc.2023.06.003]
31. Nelson BW, Allen NB. Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: intraindividual validation study. JMIR Mhealth Uhealth 2019 Mar 11;7(3):e10828. [doi: 10.2196/10828] [Medline: 30855232]
32. Etiwy M, Akhrass Z, Gillinov L, et al. Accuracy of wearable heart rate monitors in cardiac rehabilitation. Cardiovasc Diagn Ther 2019 Jun;9(3):262-271. [doi: 10.21037/cdt.2019.04.08]
33. Diaz KM, Krupka DJ, Chang MJ, et al. Fitbit®: an accurate and reliable device for wireless physical activity tracking. Int J Cardiol 2015 Apr;185:138-140. [doi: 10.1016/j.ijcard.2015.03.038]
34. Au WW, Recchia F, Fong DY, et al. Effect of wearable activity trackers on physical activity in children and adolescents: a systematic review and meta-analysis. Lancet Digit Health 2024 Sep;6(9):e625-e639. [doi: 10.1016/S2589-7500(24)00139-0] [Medline: 39112110]
35. Fitbit Ace 3: user manual version 10. Google Store for Google Made Devices & Accessories. URL: https://staticcs.fitbit.com/content/assets/help/manuals/manual_ace_3_en_US.pdf [accessed 2025-05-26]
36. VÍVOFIT® JR. 3: owner's manual. Garmin Official Site. URL: https://www8.garmin.com/manuals/webhelp/GUID-BE237499-CE69-4300-B563-57DAA26CF813/EN-US/vivofit_JR3_OM_EN-US.pdf [accessed 2025-05-26]
37. Polar active user manual. Polar Support | Polar Global. URL: https://support.polar.com/e_manuals/Active/Polar_Active_user_manual_English/manual.pdf [accessed 2025-03-26]
38. Kim Y, Lochbaum M. Comparison of polar active watch and waist- and wrist-worn ActiGraph accelerometers for measuring children's physical activity levels during unstructured afterschool programs. IJERPH 2018 ;15(10):2268. [doi: 10.3390/ijerph15102268]
39. Nash D, Shah MJ, Shehab O, et al. "But for the blind spot": accuracy and diagnostic performance of smart watch cardiac features in pediatric patients. Heart Rhythm 2024 May;21(5):581-589. [doi: 10.1016/j.hrthm.2024.01.021]
40. Hardon HJ, Van Kerkhof YN, Bartelds B, Kammeraad JAE, Van Deutekom AW. The validation and accuracy of wearable heart rate trackers in children with heart disease: prospective cohort study. JMIR Form Res 2025 Sep 30;9(1):e70835-e70835. [doi: 10.2196/70835]
41. Ernstsson J, Svensson B, Liuba P, Weismann CG. Validation of smartwatch electrocardiogram intervals in children compared to standard 12 lead electrocardiograms. Eur J Pediatr 2024 Sep;183(9):3915-3923. [doi: 10.1007/s00431-024-05648-7] [Medline: 38918230]
42. Obesity and overweight. World Health Organization. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [accessed 2025-03-23]
43. Kumar S, Kelly AS. Review of childhood obesity: from epidemiology, etiology, and comorbidities to clinical assessment and treatment. Mayo Clin Proc 2017 Feb;92(2):251-265. [doi: 10.1016/j.mayocp.2016.09.017] [Medline: 28065514]
44. Ogden CL. Defining overweight in children using growth charts. Md Med 2004;5(3):19-21. [Medline: 15495854]
45. Juonala M, Magnussen CG, Berenson GS, et al. Childhood adiposity, adult adiposity, and cardiovascular risk factors. N Engl J Med 2011 Nov 17;365(20):1876-1885. [doi: 10.1056/NEJMoa1010112]
46. Özgen İ, Çakır E, Torun E, Güleş A, Hepokur MN, Cesur Y. Relationship between functional exercise capacity and lung functions in obese chidren. Jcrpe 2015 Sep 5;7(3):217-221. [doi: 10.4274/jcrpe.1990]
47. Enright PL. The six-minute walk test. Respir Care 2003 Aug;48(8):783-785. [Medline: 12890299]
48. Wijeysundera DN, Pearse RM, Shulman MA, et al. Assessment of functional capacity before major non-cardiac surgery: an international, prospective cohort study. The Lancet 2018 Jun;391(10140):2631-2640. [doi: 10.1016/S0140-6736(18)31131-0]
49. Contini M, Angelucci A, Aliverti A, et al. Comparison between PtCO2 and PaCO2 and derived parameters in heart failure patients during exercise: a preliminary study. Sensors (Basel) 2021 Oct 1;21(19):6666. [doi: 10.3390/s21196666]
50. Ward KM, White DA, McMahon M, Griffith G. Pediatric Exercise Testing: Springer; 2024:993-1040. [doi: 10.1007/978-3-031-07563-6_38]
51. Cruz-León C, Expósito-Carrillo P, Sánchez-Parente S, et al. Feasibility and safety of field-based physical fitness tests: a systematic review. Sports Med Open 2025 Jan 24;11(1):8. [doi: 10.1186/s40798-024-00799-1] [Medline: 39853552]
52. Calcaterra V, Bernardelli G, Malacarne M, et al. Effects of endurance exercise intensities on autonomic and metabolic controls in children with obesity: a feasibility study employing online exercise training. Nutrients 2023 Feb 20;15(4):1054. [doi: 10.3390/nu15041054] [Medline: 36839416]

XSL•FO
RenderX

53. Geiger R, Strasak A, Treml B, et al. Six-minute walk test in children and adolescents. J Pediatr 2007 Apr;150(4):395-399. [doi: 10.1016/j.jpeds.2006.12.052] [Medline: 17382117]
54. Child growth standards. World Health Organization. URL: https://www.who.int/tools/child-growth-standards [accessed 2024-09-02]
55. Ostchega Y, Porter KS, Hughes J, Dillon CF, Nwankwo T. Resting pulse rate reference data for children, adolescents, and adults: United States, 1999-2008. Natl Health Stat Report 2011 Aug 24(41):1-16. [Medline: 21905522]
56. Rosner B, Cook N, Portman R, Daniels S, Falkner B. Determination of blood pressure percentiles in normal-weight children: some methodological issues. Am J Epidemiol 2008 Mar 15;167(6):653-666. [doi: 10.1093/aje/kwm348] [Medline: 18230679]
57. Flynn JT, Kaelber DC, Baker-Smith CM, et al. Clinical practice guideline for screening and management of high blood pressure in children and adolescents. Pediatrics 2017 Sep;140(3):e20171904. [doi: 10.1542/peds.2017-1904] [Medline: 28827377]
58. Age-based pediatric blood pressure reference charts. Baylor College of Medicine. URL: https://www.bcm.edu/bodycomplab/BPappZjs/BPvAgeAPPz.html [accessed 2024-09-02]
59. Fitbit charge 2: user manual version 13. Google Store for Google Made Devices & Accessories. URL: https://staticcs.fitbit.com/content/assets/help/manuals/manual_charge_2_en_US.pdf [accessed 2024-10-28]
60. Fitbit charge 3: user manual version 35. Google Store for Google Made Devices & Accessories. URL: https://staticcs.fitbit.com/content/assets/help/manuals/manual_charge_3_en_US.pdf [accessed 2024-10-28]
61. Evenson KR, Goto MM, Furberg RD. Systematic review of the validity and reliability of consumer-wearable activity trackers. Int J Behav Nutr Phys Act 2015 Dec;12(1). [doi: 10.1186/s12966-015-0314-1]
62. Shcherbina A, Mattsson C, Waggott D, et al. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. JPM 2017 Jun 1;7(2):3. [doi: 10.3390/jpm7020003]
63. Nelson BW, Allen NB. Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: intraindividual validation study. JMIR Mhealth Uhealth 2019 Mar 11;7(3):e10828. [doi: 10.2196/10828]
64. Muggeridge DJ, Hickson K, Davies AV, et al. Measurement of heart rate using the Polar OH1 and Fitbit Charge 3 wearable devices in healthy adults during light, moderate, vigorous, and sprint-based exercise: validation study. JMIR Mhealth Uhealth 2021 Mar 25;9(3):e25313. [doi: 10.2196/25313] [Medline: 33764310]
65. Wong JS, Lu WA, Wu KT, Liu M, Chen GY, Kuo CD. A comparative study of pulse rate variability and heart rate variability in healthy subjects. J Clin Monit Comput 2012 Apr;26(2):107-114. [doi: 10.1007/s10877-012-9340-6]
66. Le-Masurier GC, Sidman CL, Corbin CB. Accumulating 10,000 steps: does this meet current physical activity guidelines? Res Q Exerc Sport 2003 Dec;74(4):389-394. [doi: 10.1080/02701367.2003.10609109]
67. Angelucci A, Greco M, Canali S, et al. Fitbit data to assess functional capacity in patients before elective surgery: pilot prospective observational study. J Med Internet Res 2023;25:e42815. [doi: 10.2196/42815]
68. Charlton PH, Marozas V, Mejía-Mejía E, Kyriacou PA, Mant J. Determinants of photoplethysmography signal quality at the wrist. PLOS Digit Health 2025 Jun 1;4(6):e0000585. [doi: 10.1371/journal.pdig.0000585]
69. Nanchen D. Resting heart rate: what is normal? Heart 2018 Jul;104(13):1048-1049. [doi: 10.1136/heartjnl-2017-312731]
70. Monmeneu JV, Chorro FJ, Bodí V, et al. Relationships between heart rate variability, functional capacity, and left ventricular function following myocardial infarction: an evaluation after one week and six months. Clin Cardiol 2001 Apr;24(4):313-320. [doi: 10.1002/clc.4960240411] [Medline: 11303700]
71. Lewis MJ, Kingsley M, Short AL, Simpson K. Rate of reduction of heart rate variability during exercise as an index of physical work capacity. Scandinavian Med Sci Sports 2007 Dec;17(6):696-702. [doi: 10.1111/j.1600-0838.2006.00616.x]
72. Ståhle A. Aerobic group training improves exercise capacity and heart rate variability in elderly patients with a recent coronary event. A randomized controlled study. Eur Heart J 1999 Nov;20(22):1638-1646. [doi: 10.1053/euhj.1999.1715]
73. Mishra T, Wang M, Metwally AA, et al. Pre-symptomatic detection of COVID-19 from smartwatch data. Nat Biomed Eng 2020 Dec;4(12):1208-1220. [doi: 10.1038/s41551-020-00640-6] [Medline: 33208926]
74. Dominick GM, Winfree KN, Pohlig RT, Papas MA. Physical activity assessment between consumer- and research-grade accelerometers: a comparative study in free-living conditions. JMIR Mhealth Uhealth 2016 Sep 1;4(3):e110. [doi: 10.2196/mhealth.6281]
75. Orstad SL, Gerchow L, Patel NR, et al. Defining valid activity monitor data: a multimethod analysis of weight-loss intervention participants' barriers to wear and first 100 days of physical activity. Informatics (MDPI) 2021 Jun 1;8(2):39. [doi: 10.3390/informatics8020039]
76. Claudel SE, Tamura K, Troendle J, et al. Comparing methods to identify wear-time intervals for physical activity with the Fitbit Charge 2. J Aging Phys Act 2020 Dec 16;29(3):529-535. [doi: 10.1123/japa.2020-0059]
77. Jobson JD. Multiple Linear Regression: Springer; 1991:219-398. [doi: 10.1007/978-1-4612-0955-3_4]
78. Zhou F, Yin X, Hu R, et al. Energy expenditure estimation in children, adolescents and adults by using a respiratory magnetometer plethysmography system and a deep learning model. Nutrients 2022 Oct 1;14(19):4190. [doi: 10.3390/nu14194190]
79. Klepin K, Wing D, Higgins M, Nichols J, Godino JG. Validity of Cardiorespiratory Fitness Measured with Fitbit Compared to V˙O2max. Medicine & Science in Sports & Exercise 2019;51(11):2251-2256. [doi: 10.1249/MSS.0000000000002041]

XSL•FO
RenderX

80. Hosseini A, Fazeli S, Vliet EV, et al. Children activity recognition: challenges and strategies. Presented at: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); Jul 18-21, 2018; Honolulu, HI p. 4331-4334. [doi: 10.1109/EMBC.2018.8513320]
81. Howe CA, Clevenger KA, Leslie RE, Ragan MA. Comparison of accelerometer-based cut-points for children's physical activity: counts vs. steps. Children (Basel) 2018 Aug 3;5(8):105. [doi: 10.3390/children5080105] [Medline: 30081457]
82. Tandon A, Avari Silva JN, Bhatt AB, et al. Advancing wearable biosensors for congenital heart disease: patient and clinician perspectives: a science advisory from the American Heart Association. Circulation 2024 May 7;149(19):e1134-e1142. [doi: 10.1161/CIR.0000000000001225] [Medline: 38545775]
83. Zahedivash A, Chubb H, Giacone H, et al. Utility of smart watches for identifying arrhythmias in children. Commun Med (Lond) 2023 Dec 13;3(1):167. [doi: 10.1038/s43856-023-00392-9] [Medline: 38092993]
84. Tanriver G, Stanojevic S, Filipow N, et al. Using heart rate data from wrist worn activity trackers to define thresholds for moderate to vigorous physical activity in children and young people with cystic fibrosis. J Cyst Fibros 2025 Mar;24(2):412-417. [doi: 10.1016/j.jcf.2024.10.014]
85. Kouis P, Michanikou A, Anagnostopoulou P, et al. Use of wearable sensors to assess compliance of asthmatic children in response to lockdown measures for the COVID-19 epidemic. Sci Rep 2021 Dec 1;11(1). [doi: 10.1038/s41598-021-85358-4]
86. Angelucci A, Li Z, Stoimenova N, Canali S. The paradox of the artificial intelligence system development process: the use case of corporate wellness programs using smart wearables. AI Soc 2022 Sep 26:1-11. [doi: 10.1007/s00146-022-01562-4] [Medline: 36185063]

Abbreviations

6MWT: 6-minute walk test
DAP: diastolic arterial pressure
ECG: electrocardiography
GAI: generative artificial intelligence
GAIDeT: Generative Artificial Intelligence Delegation Taxonomy
HR: heart rate
HRZ: heart rate zone
ICC: intraclass correlation coefficient
LASSO: least absolute shrinkage and selection operator
LOOCV: Leave-one-out cross-validation
PPG: photoplethysmography
PR: pulse rate
PROS: pulse rate over steps
RPR: resting pulse rate
SAP: systolic arterial pressure

Edited by A Mavragani; submitted 17.Apr.2025; peer-reviewed by A Martinko, J Rogers, M Mansoor; revised version received 31.Dec.2025; accepted 20.Jan.2026; published 20.Mar.2026.

Please cite as:
Angelucci A, Aliverti A, Vandoni M, Lucini D, Larizza C, Carnevale Pellino V, Zuccotti G, Calcaterra V
Using an Adult-Designed Wearable for Pediatric Monitoring: Practical Tutorial and Application in School-Aged Children With Obesity
J Med Internet Res 2026;28:e76166
URL: https://www.jmir.org/2026/1/e76166
doi:10.2196/76166

© Alessandra Angelucci, Andrea Aliverti, Matteo Vandoni, Daniela Lucini, Cristiana Larizza, Vittoria Carnevale Pellino, Gianvincenzo Zuccotti, Valeria Calcaterra. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

SMARTCLOTH Prototype for Dietary Management in Patients With Diabetes Mellitus: Tutorial on Human-Centered Design Methodology for Health Care Hardware Development

Jose M Palomares[1,2,3], PhD; Rafael Molina-Luque[1,4], RN, MSc, PhD; Fernando León-García[1,2,3], PhD; Irene Casares-Rodríguez[1], MSc; María García-Rodríguez[1,5], MSc, DN, PhD; María Pilar Villena Esponera[1,6], PhD; Guillermo Molina-Recio[1,4], RN, MSc, PhD

[1]Lifestyles, Innovation and Health Associated Group, Maimonides Institute for Biomedical Research (IMIBIC), Avd. Menéndez Pidal, N/N, Córdoba, Spain

[2]Advanced Informatics Research Group (GIIA), University of Cordoba (UCO), Córdoba, Spain

[3]Department of Electronic and Computer Engineering, University of Cordoba (UCO), Córdoba, Spain

[4]Department of Nursing, Pharmacology, and Physiotherapy, University of Cordoba (UCO), Córdoba, Spain

[5]Faculty of Health Sciences, Atlantic-Mediterranean Technological University (UTAMED), Málaga, Spain

[6]Faculty of Arts and Social Sciences, International University of La Rioja (UNIR), Logroño, Spain

Corresponding Author:
Rafael Molina-Luque, RN, MSc, PhD
Lifestyles, Innovation and Health Associated Group, Maimonides Institute for Biomedical Research (IMIBIC), Avd. Menéndez Pidal, N/N, Córdoba, Spain

Abstract

Background: Developing user-centered digital health hardware requires systematic design methods applicable across clinical contexts. As diabetes mellitus continues to rise globally and contributes to morbidity, mortality, and costs, effective nutritional management remains essential—yet adherence is often poor. Digital health interventions grounded in human-centered design may enhance adherence by better aligning solutions with patients' real needs.

Objective: This tutorial aims to provide replicable guidance on applying the design thinking approach to health care hardware development, illustrated through the design, development, and preliminary usability evaluation of SMARTCLOTH (GA-16: Lifestyles, Innovation, and Health), a smart tablecloth prototype intended to facilitate dietary management and support adherence to nutritional recommendations among individuals with diabetes.

Methods: We demonstrate a systematic design thinking approach adaptable to other hardware contexts, using the Double Diamond model. In mapping, we performed a structured preassessment to define project scope and feasible functionalities. To characterize end user needs, we conducted 6 in-depth interviews with health care professionals and applied persona, empathy map, and customer journey map tools. In exploring, 5 focus groups (patients and diabetes educators) identified barriers, facilitators, and desired functionalities for dietary self-management. In building, we created low- and high-fidelity wireframes and interactive web prototypes using Phaser 3 (HTML5/JS) to simulate a kitchen workspace for meal assembly. In testing, 7 patients with different diabetes profiles participated in 3 iterative usability sessions. Using think-aloud, video analysis, and structured tasks, we documented completion times, errors, and the level of required assistance, enabling refinements. Development progressed through 15 internal versions and 3 user-tested prototypes with real-time adjustments when feasible.

Results: Interviews and focus groups yielded three user profiles guiding design: (1) adolescents with type 1 diabetes navigating social and dietary challenges, (2) working-age adults with type 2 diabetes who were motivated but inconsistent, and (3) older adults with type 2 diabetes showing low adherence due to entrenched habits. Iterative usability testing indicated that the system was intuitive, with improvements in layout, labeling, and navigation. Quantitative metrics showed refinement, with simple tasks being completed in under 1 minute in later iterations, while complex meal simulations took longer. Error rates and required guidance decreased as prototypes evolved. Qualitative feedback highlighted clarity, motivational value, and educational potential, while older participants requested larger text and simplified controls. Despite usability gains, motivational barriers persisted among low-adherence older adults.

Conclusions: This tutorial demonstrates that systematic human-centered design can yield feasible and well-accepted digital health hardware. SMARTCLOTH emerged as a promising tool for dietary management in diabetes, though effectiveness and clinical outcomes were not evaluated. The methodology can be adapted by teams developing hardware for chronic disease management.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e75744) doi:10.2196/75744

KEYWORDS

type 1 diabetes mellitus; type 2 diabetes mellitus; dietary adherence; user-centered design: usability; digital health

Introduction

Developing effective digital health hardware—such as wearable devices, smart medical equipment, or assistive technologies—presents unique methodological challenges that distinguish it from software-based interventions. Hardware development requires balancing user needs with physical constraints, manufacturing feasibility, and cost considerations, while ensuring usability across diverse populations. Human-centered design (HCD) and design thinking (DT) methodologies offer structured approaches to navigate these challenges, yet practical, step-by-step guidance on applying these methods to health care hardware development remains limited in the literature. This tutorial addresses this gap by providing detailed methodological guidance, illustrated through a real-world development case.

Diabetes mellitus (DM) is a chronic disease in which glucose metabolism is altered due to increased resistance to insulin action or low insulin production [1]. The prevalence of this pathology has increased significantly over the last few decades and will continue to do so in the coming decades. Data show that in 2024, an estimated 589 million adults were living with DM, a figure that could reach 853 million by 2050. More than 90% of these cases are type 2 diabetes mellitus (T2DM), making this disease a global public health challenge [2,3].

In addition to the significance of prevalence figures, it is worth noting that DM places a high demand on health care personnel due to the multiple comorbidities it can cause, including nephropathies, vasculopathies, and neuropathies, among others [4], with disease management playing a crucial role in this context. Diabetes management is a dynamic and long-term process that begins at diagnosis, when initial education and treatment are introduced, and continues with daily self-management tasks such as diet, medication adherence, and glucose monitoring. Periodic follow-up is also essential to prevent complications and adapt treatment as the disease progresses. In this context, diabetes self-management education and support should be provided at diagnosis, during critical transitions, and on an ongoing basis throughout the care process [5]. Diabetes self-management education and support contributes to both daily management and the acquisition of self-care skills, which are crucial for long-term adherence. Among lifestyle behaviors, nutrition is particularly influential [6,7], yet adherence to healthy dietary patterns remains inadequate in many patients [8,9]. Our research group focuses on the stage of learning and consolidating dietary self-management skills and on supporting sustained adherence through innovative tools adapted to patients' contexts that improve their clinical situation.

Among dietary strategies, carbohydrate counting stands out as the cornerstone of nutritional management in type 1 diabetes mellitus (T1DM), as it enables the adjustment of insulin doses [10]. However, its complexity often hinders adherence. The rapid growth of digital technologies has created new opportunities in this field. Mobile health apps have been shown to support lifestyle changes and chronic disease management [11-13]. In diabetes, both generalist apps (eg, MyFitnessPal [Mike Lee] and CalorieMama [Azumio Inc]) and T1DM-specific apps (eg, CarbAndMove, SocialDiabetes [SocialDiabetes SL], and Carb Manager [Wombat Apps LLC]) are widely available. Nevertheless, they share essential limitations, including low usability and adherence [14,15], limited involvement of health care professionals in their development [16], technical problems, and poor integration with glucose-monitoring devices [17,18]. This fragmentation often forces patients to use multiple apps simultaneously, complicating management and increasing the risk of abandonment.

Recent reviews also highlight that evidence of their effectiveness remains inconsistent, and many tools fail to adequately meet patients' real needs [19,20]. Successful solutions must therefore actively involve end users in the design process and incorporate features such as educational content, monitoring tools, goal setting, gamification, reminders, and social interaction options [21-23]. Positioning these initiatives within health behavior change frameworks [23-25] provides a comprehensive foundation for understanding how capability, opportunity, and motivation interact to enable sustained dietary change. In this regard, technological tools should not only support day-to-day management but also strengthen patients' ability to learn, consolidate, and adhere to effective self-care strategies over time [26].

This complex clinical challenge—requiring sustained user engagement, multidisciplinary collaboration, and hardware-software integration—provides an ideal context for demonstrating the systematic application of HCD methodology. The following sections present a replicable approach that other teams can adapt to their own hardware development challenges in chronic disease management.

In this scenario, HCD emerges as an innovation method particularly suited to the development of solutions in the highly ambiguous field of health care. It focuses on creating high-value solutions through a people-centered approach that leverages diversity of thought, creativity, and ethnographic insights, which are refined through iterative testing [27,28]. The iterative nature of HCD reduces risks in terms of time and budget, mitigates cognitive biases among researchers, empowers individuals and teams, and enhances both creative confidence and solution quality [29-33]. Despite these advantages, HCD remains relatively new and is inconsistently applied in health care innovation [34]. This fact is particularly relevant for addressing health care disparities, including the digital divide, which can be mitigated through collaborative solutions developed in partnership with patients and stakeholders [35,36].

HCD also complements classical scientific approaches by improving the translation of evidence into practice and

XSL•FO
RenderX

identifying opportunities for intervention in target populations [33,37]. Although it is often used interchangeably with DT [27,30,34,38], important distinctions exist between the two. While both are iterative and user-centered, HCD applies more rigorous methods aimed at usability and user satisfaction in digital health products. In contrast, DT has a broader scope that emphasizes innovation across diverse contexts [37-39]. In short, HCD can be considered a comprehensive framework, while DT operates as a flexible strategy for problem-solving within that framework [40].

At the same time, designing any logical or physical system is inherently complex, shaped by both enabling and constraining factors. The classical system design cycle outlines 5 main stages, namely objectives, research, requirements, design, and evaluation, yet traditionally involves users only in the final evaluation phase [41]. In practice, the process is iterative, with later phases informing adjustments to earlier ones. However, excluding users from early stages often results in products that lack adequate user experience, making people feel uncomfortable when using them with tools that were not developed according to their real needs and requirements.

For this reason, some methodologies have evolved to involve end users and stakeholders throughout the design cycle [29-31,34,38,39]. HCD has proven particularly effective for developing systems and interfaces tailored to users' needs while aligning with institutional objectives [32,33,37,38]. Because users are deeply engaged throughout the process, the resulting interfaces provide a better fit and more satisfying usability experience [27,28]. User input can be collected through both direct methods (eg, questionnaires and interviews) and indirect approaches (eg, observation of prototype interactions and session recordings), complemented by quantitative measures, such as task completion time or error frequency [33].

This tutorial demonstrates the practical application of these principles throughout a complete hardware development cycle.

In this context, a novel initiative called SMARTCLOTH has been developed. SMARTCLOTH is a 3-year project funded by the Spanish 2021 Health Research Projects of the Strategic Action in Health 2017‑2020 call. It aims to develop and preliminarily test the usability of a "smart tablecloth" designed to enhance dietary management and adherence in individuals with diabetes, supporting both the portion diet system (standard in type 1 diabetes) and the plate method (recommended in type 2 diabetes). This tutorial provides step-by-step guidance on applying HCD methodology and iterative prototyping to health care hardware development, illustrated through the SMARTCLOTH project. The approach emphasizes usability testing strategies that ensure accessibility for patients or caregivers of any age and level of technological literacy, principles that can be adapted to other digital health hardware development contexts.

Methods

Overview

This tutorial provides step-by-step guidance on applying HCD methodology and iterative prototyping to health care hardware development, illustrated through the SMARTCLOTH project. The approach emphasizes usability testing strategies that ensure accessibility for patients or caregivers of any age and level of technological literacy, principles that can be adapted to other digital health hardware development contexts. For each phase, we describe (1) the purpose and key objectives of the phase, (2) specific tools and techniques used, (3) practical implementation details, and (4) decision-making processes when challenges arose.

We present the phases sequentially for clarity but acknowledge that in practice, iteration between phases is common and often necessary. Timelines and resource allocations are provided as reference points but will vary based on project scope and available resources.

The research team applied the DT method, following the Double Diamond model proposed by the UK Design Council [42] and adapted according to Gasca and Zaragozá [43,44]. This Double Diamond structure (Multimedia Appendix 1) identifies 4 general stages (mapping, exploring, building, and testing). In turn, the first three are subdivided into another 6. In their proposal, these authors offer specific tools, which are not mutually exclusive, to complete each stage and substage. All the tools used in the development of SMARTCLOTH are detailed in Table 1.

XSL•FO
RenderX

Table . Stages, substages, and tools in the development of SMARTCLOTH.

Stage and substage	Description	Tools
Mapping	It is used to delimit the context of the work (what is known, what is unknown, and what needs to be known) and the scope and objectives of the project.	—[a]
Mapping the team	Map the internal context of the interdisciplinary research team.	• Strengths, Weaknesses, Opportunities, and Threats (SWOT) • In/Out
Mapping customers	This tool helps know the profiles of patients who would benefit from using SMARTCLOTH (external context).	• In-depth interviews • Person • Empathy map • Customer journey map
Exploring	It focuses on using qualitative techniques to understand and delimit the problem. It also gathers information on what is currently being done to address it.	—
Research	It is essential to understand the problem and its related factors in depth.	• Focus group (patients) • Focus group (experts)
Synthesis	The specification of the problem in the design challenge, in addition to synthesizing the research into a challenge, orients the team and focuses on devising solutions on something concrete.	• Design challenge
Building	It consists of embodying ideas in a prototype to define and transmit an idea quickly.	—
Devise	Through divergent thinking, the devise substage allows a broad spectrum of possible solutions to the design challenge to be explored before converging on the most effective and feasible proposals.	• Brainstorming • Cocreation sessions
Prototyping	It consists of materializing ideas. It is the process of quickly defining and transmitting an idea and materializing it into something physical that can be tested.	• Wireframes and web prototyping • 3D printing and physical prototyping
Testing	It is used to get feedback on each idea presented to end users. This stage allows us to learn what works and what doesn't and to test functionalities without the need for the final product.	• User test on the web • User test on a physical prototype • Usability test

[a]Not applicable.

Phase 1: Mapping (January-February 2022)

Scope Definition and User Profiling

The key aspects of this phase include:

1. Purpose: Systematic problem assessment, team alignment, and user characterization
2. Adaptable duration: 1 - 3 months depending on project scope
3. Key outputs: User profiles, project scope definition, and initial requirements
4. Application to SMARTCLOTH: According to the Double Diamond DT, the initial stage is the mapping, in which the problem is assessed, and the available resources are determined. This phase started in January 2022 and finished by the end of February 2022.

Team

In the team mapping phase, conducted in February 2022, we used the SWOT (Strengths, Weaknesses, Opportunities, and Threats) and IN/OUT tools to identify and categorize the team members' strengths, weaknesses, opportunities, and threats and clearly define the elements within or outside the project's scope. Both health care professionals and engineers participated in these sessions.

User and Market

Subsequently, for patient mapping, 6 in-depth interviews were conducted with health care professionals who regularly deal with patients with diabetes (4 nurses and 2 nutritionists), selected by nonrandom convenience sampling between February and March 2022. We used user-centered tools to synthesize the information collected, including "persona," "empathy map" and "customer journey map." The persona tool allows us to create

XSL•FO
RenderX

detailed profiles of fictitious users based on actual data. Each "persona" represents a specific profile of patients who might use SMARTLOTH, detailing their demographics, behaviors, needs, and goals [45]. The "Empathy Map" was used to capture and organize information about what users think, feel, say, and do about diabetes dietary management. This tool helped us understand users' emotions and perspectives and identify their main frustrations and desires [46]. Finally, the "Customer Journey Map" was used to map the complete journey of people with diabetes through the health care system, from the first contact to the end of their interaction. It highlighted the touchpoints, emotions, and potential barriers users face throughout their experience [47]. By identifying opportunities for improvement at each stage of the journey, we were able to define the most effective and successful functionalities of SMARTCLOTH, as well as the point in education of patients with diabetes where it could be used. Ultimately, these tools allowed us to develop detailed patient profiles, deeply understand their needs, emotions, and experiences, and comprehensively visualize their interaction with health services over time, thus facilitating the identification of critical points and opportunities for improvement that SMARTCLOTH could represent.

Data Analysis for Mapping

A phenomenological approach was used to analyze the data from the in-depth interviews with health care professionals specialized in diabetes education. The interviews were transcribed and meticulously reviewed. Subsequently, initial coding of the transcripts was undertaken, identifying emerging themes related to perceptions and experiences of dietary management and technology use. The codes were grouped into broad themes, such as experiences with technology, challenges in diabetes education, and recommendations for technology improvements, which are essential for identifying common patterns and developing detailed user profiles.

User-centered design tools, such as "persona," "empathy mapping," and "consumer journey mapping," were developed using the information extracted from the transcripts and themes. Different user profiles were identified based on familiar patterns, and their demographic, psychographic, and target characteristics were detailed. The phenomenological approach was the most appropriate for this analysis, as it allowed capturing the subjective and emotional experiences of health care professionals and their patients, providing a comprehensive and richly detailed view that facilitated the identification of pain points and opportunities for improvement in the use of diet management hardware in patients with diabetes.

Phase 2: Exploring (April-June 2022)

Research, Feasibility, and Challenge Formulation

The main aspects of this phase are as follows:

1. Purpose: Deep investigation of user needs and technical feasibility
2. Adaptable duration: 2 - 4 months depending on complexity
3. Key outputs: Design challenge formulation, technical framework selection, and detailed requirements
4. Application to SMARTCLOTH: The research stage was carried out after establishing the initial set of objectives. We noticed that at least 2 research areas, clinical and technical, had to be considered.

Clinical Research

During the substage of the research, conducted between April and June 2022, we completed 5 focus groups to gain an in-depth understanding of the needs, experiences, barriers, and facilitators of different patient profiles regarding adherence to the recommended dietary pattern in diabetes. These included 3 focus groups with patients from the 3 profiles previously identified in the mapping stage and 2 focus groups with nurses responsible for diabetes education and management in both primary and specialized care.

Technical Research

We studied different software and hardware design methodologies. Hardware design requires determining the components, integrating all of them, placing each element in the device, building the device, interconnecting everything, testing it, and checking the conformance with the objectives and the experience of use of the product by different users. This task is rather long and time-consuming. Therefore, most companies do not build the final product until all the elements have been tested and approved. The users do all the testing, and in this sense, as the HCD requires large interactions with the users, it is highly suitable for combining both.

For this reason, we decide to include a quick and easy mechanism for creating a simulated device to ask users about the product without building it physically. This tool would reduce the costs of the overall design and development process, and users would provide accurate usage information for the designer. Therefore, the physical product will finally be built with the most accepted design from the users.

We analyzed several options for prototyping, including (1) creating a simulator from scratch using Python (Python Software Foundation) or other similar programming language, (2) building a mobile or tablet app, or (3) making a dynamic web page. The first option is the most powerful, as the programmer can design any behavior, but it requires much effort to develop the simulator. Moreover, adapting the simulator for every platform where it will be executed would be required. The second option has been widely used because tablets and smartphones are ubiquitous. Therefore, the app is developed once and used on lots of devices. The last option was creating a dynamic response web page. This third is the most versatile option, as any device connected to the internet with a web browser is suitable. Simulating the behavior of the SMARTCLOTH device could be like playing a game. In fact, many games are developed for HTML5-compliant web browsers, executing the code on the device and not on the server. Regarding HTML5 development, some of the most relevant game-creation frameworks are Phaser 3, Matter.js, Kiwi.js, Quintus, and CreateJS. All of them are similar, with slightly different variations and compatibilities with various web browsers and platforms. Phaser 3 is probably the most versatile HTML5 game development framework, as

XSL•FO
RenderX

it has fast rendering and multitouch input capabilities on mobile and desktop browsers.

Synthesis

We use the design challenge tool in the synthesis substage to analyze and organize the information gathered. This tool is used to frame the core design problem and helps define the scope and objectives of the project, establishing a common starting point for the team. Typically, the challenge is formulated as an open-ended question that encourages creativity and the exploration of multiple possible solutions [48,49]. While using this tool, we focus on identifying and articulating the primary needs and desires of the users, as well as the constraints and opportunities of the project context. The formulation of the design challenge was agreed upon in a collaborative and interdisciplinary process between the health care professionals and the engineering team, trying to synthesize all the data and insights collected in the previous stages and substages.

Data Analysis for Mapping

Similarly, for the exploration phase, all focus group sessions with different profiles, which were audio- and video-recorded, were transcribed verbatim by a specialized company. Ricoeur's method of hermeneutic analysis[50] a 4-stage inductive approach, was used with the support of NVivo (Lumivero) software to analyze qualitative data. First, 2 researchers reviewed the focus group data and transcripts multiple times to gain a general understanding and compile a list of main ideas. Then, the data were coded, and nodes with predefined categories were identified, establishing 13 analytical subcategories. In the third stage, new nodes were generated based on the affinity of meanings, creating 4 codes and 18 subcodes to achieve greater abstraction and discriminate the characteristics of the same idea. In the fourth stage, a greater understanding of the initial categories was achieved through the "hermeneutic arc," considering both the initial experience and the rereading of the corpus of data. This analysis enabled us to identify which elements or characteristics act as barriers or facilitators from the perspectives of both patients and diabetes education nurses.

Phase 3: Building (Iterative Timeline)

From Low-Fidelity Wireframes to Interactive Web Simulation

The relevant aspects of this phase include:

1. Purpose: Translation of requirements into tangible prototypes
2. Adaptable approach: Can use various prototyping tools based on resources
3. Key outputs: Low-fidelity wireframes, high-fidelity prototypes, and interactive simulations
4. Application to SMARTCLOTH: The team built a functional system for users to test at this stage.

Devise

The devise (ideation) substage was based on the implementation of brainstorming techniques and cocreation sessions and focused on the definition and integration in a single device of the functionalities that SMARTCLOTH should present to respond to the design challenge. During the brainstorming, the interdisciplinary team proposed creative and varied solutions, regardless of their initial feasibility, to maximize creativity and divergent thinking. This process allowed for exploring multiple approaches before selecting the most feasible ideas for further development [51]. The cocreation sessions also involved the interdisciplinary group to harness the diversity of thinking and experience, enrich the design process, and ensure that solutions were comprehensive and well-founded, facilitating a holistic and user-centered approach [52]. Ultimately, both sessions were characterized by interdisciplinarity, seeking solutions from the perspective of health care workers and patients but accepting the technological and time constraints for the completion of the project.

Prototyping

Subsequently, the prototyping substage started with the creation of low- and high-fidelity wireframes and then moved on to web prototypes. The low-fidelity wireframes were initially used to outline the basic structure and layout of the elements in our interfaces without focusing on design details. These simple, quick sketches allowed us to agilely explore different configurations and functionalities.

Finally, we developed high-fidelity wireframes, incorporating finer details, including typography, colors, and graphics, providing a more realistic representation of the final product [53,54]. Once the high-fidelity wireframes were approved, we created interactive web prototypes, which allowed us to simulate the end-user experience and evaluate the functionality and fluidity of the interface in a realistic digital environment, screen and button locations, icons, and general usability ideas. This strategy ensured the design was visually appealing, intuitive, and functional for users [22]. Phaser 3 framework enables drawing 2D objects, adding event-handling hook functions, and managing different functionalities using JavaScript. Several assumptions and simplifications had to be made. For example, we need to simulate the part of preparing the meal by humans. In real life, this process would take place at the kitchen table, by cutting portions of food or using a plate or bowl to serve each item. However, these actions fall outside the scope of the SMARTCLOTH physical device. The design will only deal with the device and its internal functioning. However, to simulate the behavior of the SMARTCLOTH virtual device, the web was split into 2 areas. On the right part, we add an area named "YOUR KITCHEN," in which a dish and several different foods could be dragged to place the dish above the balance or add or remove pieces of food to the dish or from the dish, or even directly above the weight. As a simplification, 2 buttons in the "YOUR KITCHEN" area were added to add 25 g of a given previously activated food or to remove 25 g from the piece of food. The piece of food grows bigger or smaller according to the number of 25 g clicks the user has pressed. If the piece of food goes to 0 g, it disappears from "YOUR KITCHEN." We decided that the "YOUR KITCHEN" area should be placed on the left part of the screen. On the right side of the screen, the SMARTCLOTH device was simulated.

The SMARTCLOTH web prototype was designed and developed using the Phaser 3 JavaScript framework. The initial

XSL•FO
RenderX

prototype, shown in Figure 1, was created using simple forms (circles and rectangles). Members of the project team tested this prototype internally to check the conceptual functionality of the elements.

Figure 1. SMARTCLOTH initial web prototype – not shown to users, internal testing only.

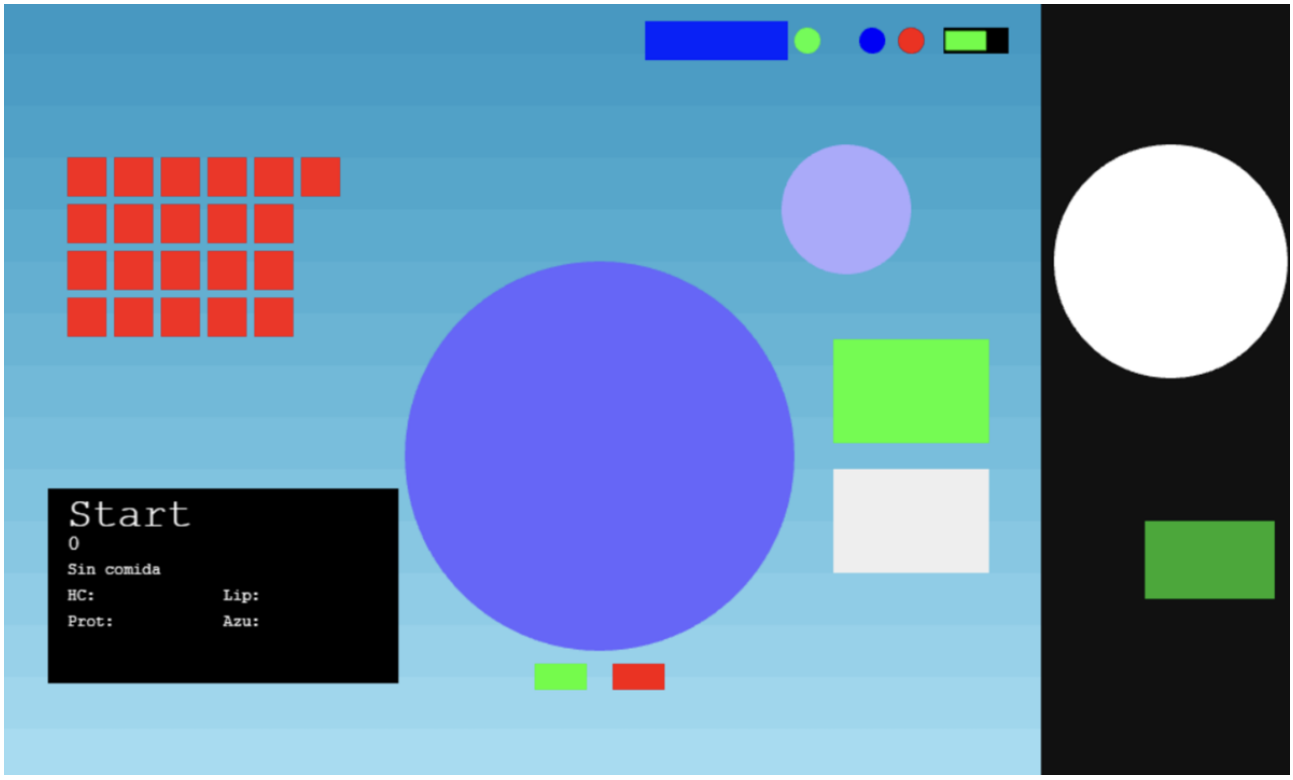

Phase 4: Testing (October-December 2022)

Iterative Usability Evaluation and Refinement

The main elements of this phase include:

1. Purpose: Iterative usability evaluation and refinement
2. Adaptable cycles: 2 - 5 testing iterations recommended
3. Key outputs: Validated design, usability metrics, user feedback, and refined prototype
4. Application to SMARTCLOTH: SMARTCLOTH web prototypes designed in the previous phase were used for the testing stage. Five testing sessions were conducted between October and December 2022 (Multimedia Appendix 2), involving 7 patients with diabetes representing the 3 previously identified profiles. Each prototype was tested by users and supervised by project team members. During these sessions, participants performed a series of specific tasks, detailed in Table 2, to evaluate the functionality and usability of the prototype. Each test was recorded with a video camera so that their behavior could be checked in the future. The users provided some annotated feedback, which was used to make all the changes for the following prototype evolution.

XSL•FO
RenderX

Table . Tasks applied in the SMARTCLOTH user tests.

Test	Elements involved	Difficulty
Weigh one portion of grilled salmon (300 g)	• Dish • Sliced salmon	Easy
Weigh a 150 g apple	• Apple	Easy
Weigh a plate of 150 g of pasta, 50 g of tomato, and 50 g of canned tuna.	• Dish • Pasta (boiled)[a] • Tomato, natural • Canned tuna (raw)	Moderate
Prepare a chickpea salad and a plate of 150 g of pasta, 50 g of tomato, and 50 g of canned tuna.	• Chickpeas (boiled) • Tomato, natural • Canned tuna (raw)	Moderate
Simulate two meals on the same day: • Breakfast: 150 g whole milk + 100 g whole wheat bread + 50 g butter • Lunch: 200 grams of boiled zucchini + 50 grams of oil + 150 g of breast + 1 boiled egg + 150 grams of banana	• Glass • Plate • Whole milk • Butter • Wholemeal bread • Dish[b] • Courgette (boiled) • Olive oil • Breast • Hard-boiled egg • Banana	Hard

[a]SMARTCLOTH can differentiate between raw and boiled foods.

[b]When having 2 meals, 2 dishes are needed.

The think-aloud technique was used to collect qualitative data on the user experience during the development of the tests. This technique involves participants verbalizing their thoughts, emotions, and actions as they interact with the prototype, providing insight into their cognitive process and identifying usability issues [55,56]. The choice of the think-aloud technique, widely recognized in usability research, was based on its ability to reveal critical aspects of user behavior that would not be evident through other evaluation methods [57]. The data collected was analyzed to identify common patterns, recurring problems, and SMARTCLOTH usability, providing the basis for subsequent design iterations and ensuring that the final product was intuitive and functional for end users.

We designed a set of questions for the users after each test. Each user was asked about how many kilocalories that meal had, how many rations of carbohydrates that meal had, and some other similar questions. We took note of the errors they committed and the overall interaction with the simulator. Finally, we asked them to express their feelings and suggestions about using that simulator (Multimedia Appendix 3).

For each recorded test, we wrote down several direct data points extracted from the videos, including number of attempts for each test, total time to complete the test, best time try, average time of the attempts, number of errors, number of explanations and positive reinforcements, and stops in the attempts (direct instructions and indirect indications). Some other indirect data were extracted from the videos of the answers after the test, including responses completed without help, with help, and unanswered or incorrectly answered. In the following, we describe every term extracted from the videos:

1. Attempt: Time interval in which the user works with the prototype to accomplish the proposed task.
2. Explanation: Time interval in which the user receives a more or less detailed explanation of the task, operation of the device, or some other information related to the experiment.
3. Error: Detection of malfunctioning of the prototype. Depending on its severity, it can generate a restart of the attempt.
4. Positive reinforcement: Affirmative expression confirming to the user that he or she is proceeding correctly.
5. Direct help: Explicit indication of the next step to follow.
6. Indirect help: Implicit indication, usually in the form of a question, to give the user clues on how to continue.

In each test, we showed the SMARTCLOTH web prototype on a large (50 inches) screen connected to a computer with an internet connection. We explained every button and element to users, describing their use and behavior. After that, one by one, each user was asked to manage different types of meals with increasing complexity. Each user entered the room where the test was carried out at a given time without interaction with the rest of the users so that no bias could arise when exchanging information among them.

Data Analysis for Testing

A qualitative approach based on content analysis was used to analyze the data obtained through the think-aloud methodology

XSL•FO
RenderX

and the annotations made during testing. The sessions were also videotaped, transcribed, and coded. User comments and behaviors were analyzed and grouped into categories related to usability, understanding of the hardware, difficulties encountered, and suggestions for improvement.

Positionality Statement

The research team comprised health care professionals (nurses and nutritionists) and engineers (computer and electronic engineering) affiliated with academic and biomedical research institutions in Spain. Our backgrounds combine expertise in diabetes education, nutrition therapy, digital health innovation, and HCD methodologies. Several members have extensive clinical experience supporting individuals with diabetes and are committed to enhancing patient education and dietary adherence. Others contribute technical expertise in prototyping, usability, and interactive system design.

We acknowledge that our values—particularly a belief in the potential of technology to enhance lifestyle management and our commitment to patient-centered care—have influenced the research process. These perspectives informed our emphasis on usability, patient engagement, and contextual adaptation of SMARTCLOTH. At the same time, we recognize that our orientation toward health care innovation could predispose us to highlight the positive aspects of digital interventions.

To address these influences, we used reflexive practices, triangulated perspectives within an interdisciplinary team, and integrated continuous feedback from patients and health care professionals throughout the design and testing process. This collaborative and iterative approach aimed to ensure that the findings reflect not only our assumptions but also the lived experiences and challenges of people with diabetes.

Ethical Considerations

The project complied with current Spanish regulations on bioethical research, personal data protection, and bioethics, following Law 3/Dec 5, 2018. In addition, the fundamental principles of the Declaration of Helsinki (1964), the Council of Europe Convention on Human Rights and Biomedicine (1997), and the United Nations Educational, Scientific and Cultural Organization (UNESCO) Universal Declaration on the Human Genome and Human Rights (1997) were respected. The research was approved by the Ethics and Research Committee of Cordoba (Act n° 273, ref. 3754). Since video and audio recordings were made of the participants, they were also asked to sign a consent for transferring image rights. In the case of minors, these consents were signed by their parents.

This study was conducted in accordance with ethical standards on privacy and confidentiality. No identifying information of participants is included in the manuscript or supplementary materials, and all potentially identifiable data were omitted unless scientifically essential. All participants signed informed consent and were able to withdraw it at any time if they considered it appropriate; in the case of minors, informed consent forms were signed by their parents. Since video and audio recordings were made of the participants, they were also asked to sign a consent form for transferring image rights. No compensation was provided to participants.

Results

Methodological Insights and Case Example Outcomes

This section presents the outcomes of applying the Double Diamond methodology to SMARTCLOTH development, organized to highlight both the specific results obtained in our case example and the generalizable methodological insights that emerged from the process. We emphasize decision points, adaptations, and lessons learned that can inform other hardware development projects.

Mapping

Team Mapping

The team considered that the main threats identified included the recent pandemic (and its long-term consequences), the difficulty in changing dietary habits, and the increasing number of similar products on the market. Significant weaknesses included limited time availability and lack of human resources, the disparity of interests among the researchers, and the fact that they were involved in several projects (multitasking). However, the project had important strengths, such as skills in web technologies, support from various institutions, and a strong motivation for health education. Notable opportunities included (1) direct access to patients and diabetes care professionals, (2) a positive view of technology in the wake of the COVID-19 pandemic, (3) the availability of funding, and (4) the opportunity to apply to different national calls for hiring professionals to reinforce the team, all of which supported the feasibility and potential success of the project.

In addition, it was crucial to determine which technological challenges could be addressed by the team. Among the functionalities of SMARTCLOTH that the team felt they were capable of developing were food weighing, barcode reading, design and physical printing of prototypes, design of buttons for user interactions, estimation of calories and macronutrients ingested at each meal with the ability to make longitudinal records of intake, and a website where each user could submit and consult this information, among others. However, it was decided that other functionalities (shopping list, voice recognition, and offering feedback on the diet quality) exceeded the time and economic capabilities of the equipment. A graphical summary of the results of these dynamics is available in Multimedia Appendix 4.

User Profiles Mapping

In the in-depth interviews conducted to identify user profiles, 6 professionals (4 diabetes nurse educators and 2 nutritionists) with varying experience in diabetes management but all over 3 years participated. The analysis of these interviews helped us identify potential users of SMARTCLOTH, identifying 3 profiles.

The first profile (given the name Kevin) is an adolescent with T1DM. Kevin manages his insulin well and understands his condition, but he faces challenges related to social life and eating out. He represents teenagers who think changing their diet is complicated and do not want to be the “weirdo” among their friends. They perceive their health as out of control and

XSL•FO
RenderX

sometimes feel frustrated because they cannot eat like everyone else. These users seek to enjoy and have fun while eating but face occasional family conflicts and struggle to adjust their eating habits. In their journey as users, they face (together with their family) initial doubts and fears, mistrust and loneliness, but show interest and confidence in seeking new tools to manage their condition.

The second profile, Julia, represents mature adults (between 50 and 65 years old) with T2DM. This profile corresponds to active workers, which makes it difficult for them to maintain a regular diet due to their schedules. They are motivated to improve their dietary knowledge and are concerned about their health but find it difficult to plan and prepare healthy meals. Overall, they maintain acceptable control of their disease. However, they perceive diet as a significant sacrifice and feel uncertainty and decreasing motivation when faced with limited consultation time and the confusion of scattered and, in many cases, contrary information they can find on other media, such as the internet. They often commit dietary transgressions linked to family gatherings, parties, and work shifts. In their journey as users, they experience an internal struggle to follow the diet and improve their eating habits while seeking to maintain an active social life and enjoy food with family and friends.

The third profile (Paco) corresponds to older people (aged more than 65 years) with T2DM and low adherence to dietary treatment. Generally speaking, they have a long history of the disease and face insecurities and a lack of interest in following the diet. In addition, they have no objection to adhering to pharmacological treatment, although they say that insulin injections frighten them a little. They prefer traditional food and do not want to give up what they enjoy. This group of users thinks it is difficult to change their eating habits at their age, and they feel distrustful and uncertain about the effectiveness of the diet. In the case of men, it is common for them to delegate responsibility for their diet to their wives (which can lead to family conflicts), and in general, they have not changed their lifestyle and do not follow the recommended diet. In their journey as users, this resistance to change stands out; they argue with their partners about meals and feel overwhelmed by the amount of information about diet despite understanding the importance of taking care of themselves. SMARTCLOTH could be a solution for this group, provided they are motivated. However, the fact that they are not motivated and have poor disease control makes it unlikely that they will ultimately be users of this device. These profiles helped design SMARTCLOTH to meet the specific needs of each user group, providing solutions tailored to their particular contexts and challenges.

Multimedia Appendix 5 (in Spanish) summarizes the tools used to define these profiles (Persona, Empathy Map, and Consumer Journey Map).

Exploring

Once these 3 profiles were identified, 5 focus groups were conducted, 1 for each profile, and 2 other groups that included primary and specialized care professionals responsible for diabetes education. The characteristics of the participants are detailed in Table 3 for the patients and Table 4 for the nurses.

XSL•FO
RenderX

Table . Descriptive variables of the patients participating in the focus groups.

Code	Age (years)	Sex	Marital status	Area	Employment status	Educational level	Socioeconomic status	Diagnosis DM[a] (year)
Patients with T2DM[b]								
T2DMP1	71	F[c]	MR[d]	U[e]	HW[f]	P[g]	ID[h]	2018
T2DMP2	65	F	MR	U	RT[i]	P	IE2[j]	2007
T2DMP3	70	F	MR	U	HW	P	IE1[k]	2004
T2DMP4	77	F	MR	U	RT	P	IE1	2006
T2DMP5	61	F	MR	U	AJS[l]	P	IE1	2009
T2DMP6	68	M[m]	MR	U	RT	P	IE1	__[n]
T2DMP7	67	M	MR	U	RT	P	ID	2012
T2DMP8	53	F	MR	U	A[o]	Sec[p]	IE1	2018
Patients with T1DM[q] and parents								
T1DMP1	16	F	SN[r]	R[s]	ST[t]	Sec	ID	2014
T1DMP2	17	M	SN	U	ST	Sec	ID	2014
T1DMP3	15	M	SN	U	ST	Sec	IA1[u]	2019
T1DMP4	15	M	SN	U	ST	Sec	IC[v]	2018
PART1DMP1[w]	55	M	MR	R	A	P	ID	—
PART1DMP2	50	F	MR	R	HW	P	ID	—
PART1DMP3	45	F	SD[x]	U	A	S	ID	—
PART1DMP4	48	F	MR	U	A	US[y]	IA1	—
PART1DMP5	46	F	MR	R	HW	P	IC	—
PART1DMP6	45	M	MR	R	A	Sec	IC	—

[a]DM: diabetes mellitus.
[b]T2DM: type 2 diabetes mellitus.
[c]F: female.
[d]MR: married.
[e]U: urban.
[f]HW: housewife.
[g]P: primary.
[h]ID: 1313 - 1602€/month (1€=US $1.16).
[i]RT: retired.
[j]IE2: less than 745€/month.
[k]IE1: 745 - 1312€/month.
[l]AJS: active job search.
[m]M: male.
[n]Not available.
[o]A: active
[p]S: secondary.
[q]T1DM: type 1 diabetes mellitus.
[r]SN: single.
[s]R: rural.
[t]ST: student.
[u] IA1: more than 3005€/month.
[v]IC: 1603 - 2145€/month.
[w]FAMXX: A family member or primary caregiver of a patient with type 1 diabetes mellitus.

XSL•FO
RenderX

[x]SD: separated or divorced.

[y]US: university studies.

[z]PAR: parents.

Table . Descriptive variables of the diabetes nurse educators participating in the focus groups.

Code	Age (years)	Sex	Educational level	Socioeconomic status	Years of experience in diabetes education
DNE1[a]	42	F[b]	U[c]	IB[d]	16 (as a primary nurse, although in health centers, they are not exclusively dedicated to this).
DNE 2	55	F	U	IB	17 (as a primary nurse, although in health centers, they are not exclusively dedicated to this). She was diagnosed with diabetes 30 years ago.
DNE 3	58	F	U	IB	32 (as a primary nurse, although in health centers, they are not exclusively dedicated to this).
DNE 4	55	F	U	IB	6 years as a diabetes nurse educator in the endocrinology department.
DNE 5	55	F	U	IB	6 years as a diabetes nurse educator in the endocrinology department.
DNE6	59	F	U	IB	4 years as a diabetes nurse educator in the endocrinology department.

[a]DNE: diabetes nurse educators.

[b]F: female.

[c]U: university studies.

[d]IB: 2146 - 2451€/month (1€=US $1.16).

Patients with T1DM and their caregivers (fathers, mothers, or both) indicated that diabetes management is more difficult in situations outside the daily routine, such as when eating out or at social events. They also mentioned that a lack of knowledge and confidence in portion measurement and carbohydrate counting can be a significant barrier. Other factors highlighted were the influence of the social environment and the need for flexibility in dietary choices. However, motivations for maintaining adherence included the desire for future well-being and effective diabetes management to avoid long-term complications. Possible functionalities we extracted for this group included food scanning and carbohydrate counting, personalized alerts and reminders, a healthy recipe database, real-time monitoring and feedback on glucose levels, adjusting dietary recommendations based on current readings, compatibility with glucose sensors, and educational and motivational functionalities.

Participants belonging to profile 2 (with T2DM and with acceptable metabolic control) highlighted that diabetes management is affected by factors, such as the social environment, the availability of healthy food at home, and the lack of structured support to follow a proper diet. Many mentioned that they find it difficult to follow dietary recommendations due to the temptation to consume unhealthy foods readily available in their environment. In addition, some patients expressed that the lack of clear and personalized information about recommended diets complicates their adherence. On the other hand, motivations for maintaining adherence include avoiding major complications and maintaining a good quality of life. Functionalities for this group included personalized dietary recommendations with personalized meal plans, food scanning and analysis, reminders and alerts (for food intake, medication, and glucose level monitoring), interactive nutrition education, connection to monitoring devices, community and motivational support

XSL•FO
RenderX

(creating a community support platform to share experiences and advice, and receive support), and physical activity recording.

Finally, from the group of patients belonging to profile 3, participants mentioned the difficulty of following a diet due to the constant temptation to consume unhealthy foods available at home, the lack of social and motivational support, and the perception that measuring and weighing food is tedious and difficult to maintain in the long term. Some patients expressed that physical exercise was used to justify consuming nonrecommended foods, believing that physical activity would compensate for dietary excesses. Lack of consistency and additional health problems were also important barriers. On the other hand, the main motivation for following the diet was to improve quality of life and avoid serious health complications. However, there was little motivation to adhere to the recommended dietary pattern. Continuous monitoring of glucose levels, recording of physical activity, alerts and reminders, nutritional and motivational education, food scanning, and carbohydrate counting were also highlighted as functionalities for managing their disease.

According to specialist care nurses for patients with diabetes, patients with T1DM are often newcomers, those who are newly diagnosed, or those with years of evolution who require ongoing education on the management of their condition, including insulin administration and carbohydrate counting. On the other hand, patients with T2DM, especially those poorly controlled (profile 3), tend to be more resistant to changing habits and face additional complications that make adherence to dietary guidelines complex and are the reason for referral from primary care. According to these nurses, the main barrier for patients with T1DM is the lack of adequate initial education, especially at the time of diagnosis, when both patients and their families feel overwhelmed. Adolescents with T1DM (profile 1) sometimes have low adherence to dietary recommendations due to rebelliousness and preference for unhealthy foods. In addition, constant carbohydrate counting and weighing is tedious for many, which can lead to errors in insulin management. However, structured and ongoing diabetes education is a crucial facilitator, as are family support and the use of technological tools, such as mobile apps and continuous glucose monitoring devices, which help to improve adherence. For patients with T2DM, the most prominent barriers include ingrained eating habits and the perception that following a diabetic diet is complicated and tedious. Cultural and social factors also play an important role in dietary adherence. Many patients prioritize medication over diet, which hinders adequate glycemic control. On the other hand, facilitators for these patients include ongoing education and awareness of the importance of diet in diabetes management. Social and family support can motivate patients to follow an appropriate diet, and using personalized meal plan tools is also beneficial.

As a result, the specialized care nurses believe that SMARTCLOTH could include functionalities such as the implementation of a food scanning and analysis system that provides detailed nutritional information that would help patients with carbohydrate counting, personalized meal plans, and detailed menus to avoid the monotony and perceived complication of following a diabetic diet, and automatic alerts and reminders for taking medications, meals, and glucose measurements. In addition, they comment that mobile apps that help to calculate portions and adjust insulin intake according to the meal would be a valuable tool, as well as interactive and ongoing nutrition education, including modules on nutrition and diabetes management, which can improve knowledge and adherence.

Primary care nurses mainly care for patients with T2DM and, to a lesser extent, patients with T1DM, with one of the participating nurses having T1DM, which brought a unique perspective to the session. According to the professionals, patients with T1DM (profile 1) often arrive well educated and controlled from diagnosis, especially if they debut in childhood, but face barriers such as the perception of diet as tedious and lack of adequate initial education. At the same time, facilitators include good diabetes education and family support and engagement. Patients with T2DM face barriers, such as ingrained dietary habits, cultural influences, and a passive attitude toward diabetes, with facilitators including continuing education, social support, and the use of practical tools. For all these reasons, they agree with the other nurses that functionalities, such as food scanning, personalized meal plans, automatic alerts and reminders, carbohydrate counting apps, interactive nutrition education, and a community support platform could improve dietary adherence.

Building

The qualitative assessment of patients' perspectives revealed a set of needs that were systematically translated into essential functionalities, which necessarily had to be incorporated into the SMARTCLOTH device. This process ensured that the design responded directly to real user requirements, providing core features, such as real-time conversion of food weights into macronutrient portions, integrated calorie estimation, food group identification, and barcode or QR code scanning (Table 5). However, other needs could not be satisfied due to technical constraints, challenges encountered during the prototyping phase, or because they exceeded the project's scope and objectives. These unmet needs, also documented in Table 5, included, for example, the provision of insulin dosing guidelines, auditory feedback when pressing keys, and additional nutritional recommendations, such as specific weight-loss or lactose-free diets. All this information was subsequently transferred to a SMARTCLOTH simulation website that supported the first round of user testing.

XSL•FO
RenderX

Table . Identified needs and functionalities developed to meet them[a].

Need identified	Functionality developed to respond to the need
• Difficulty converting food weight in grams into macronutrient portions. • Lengthy calculations, leading to disengagement and discontinuation. • Uncertainty regarding carbohydrate intake when estimating the insulin dose to be administered (T1DM)	The screen provides real-time display of both weight in grams and the corresponding macronutrient portions.
• Need to rely on multiple devices (scales, calculators, plates, depending on the food) to calculate each dish.	SMARTCLOTH integrates all required tools into a single device.
• Lack of information on caloric intake (particularly relevant for some patients with T2DM and obesity)	The screen provides real-time estimation of the caloric content of each dish.
• Lack of knowledge and difficulty identifying foods, the group they belong to, and their main macronutrient	SMARTCLOTH includes a panel of 20 food groups illustrated with images; selecting a group displays a list of foods within it.
• Difficulty estimating macronutrient portions in packaged foods.	SMARTCLOTH incorporates barcode and QR code scanning.
• Lack of overview of daily and weekly caloric and macronutrient intake.	SMARTCLOTH stores consumption data, accessible via a web platform.

[a]Needs that have not been met: (1) guidelines for insulin dosing based on carbohydrate intake; (2) auditory feedback when pressing device keys; (3) provision of healthy recipes or meal plans; (4) additional nutritional recommendations (eg, weight-loss and lactose-free diets); (5) touchscreen interface (requested primarily by younger patients).

The SMARTCLOTH web prototype had 15 internal versions, which provided 3 prototypes to be tested by the users. All the versions were tested internally to check whether the system behaved correctly. After those internal prototypes, users tested each new version and modified it accordingly. Main modifications came from the clinical part as new meals were designed, and thus, the “YOUR KITCHEN” had to be altered to add and remove food to create the meal.

Some other modifications were required because of the users’ feedback about the position or size of the elements (screen, balance, or buttons). These modifications were the most interesting from the design point of view, as they provided firsthand feedback from the users’ usage experience.

Besides, the researchers performed express modification sprints during testing sessions to change some behaviors or correct minor errors that would not affect the rest of the users in the testing. Some of these modifications were done in less than 10 minutes, from detecting the requirement to be modified or added to uploading the new code to the web server. These modifications could be done in real-time as the modified file was on the web server. Therefore, each new prototype version was available just after reloading the web page.

The following version, shown in Figure 2, represented a significant change. All the simple forms were substituted by images representing real buttons, screens, etc, so users could feel the usage experience as real as possible.

Figure 2. SMARTCLOTH’s first web prototype shown to users.

XSL•FO
RenderX

All the users in the test were surprised about the SMARTCLOTH prototype. They all liked it very much and were committed to helping in the test the best they could. We noticed that the users did not "read" the screen. The screen's position in the lower-left corner of the SMARTCLOTH prototype was not the best choice. All the users expected it in the upper left corner. They did not recognize most foods in the buttons representing food groups. The quality or size of those buttons did not allow the users to understand the foods included in every food group well. Some of them suggested adding examples on the screen showing the foods included in each group. Some of them did not like the position of the buttons for adding a new dish or removing a dish from the meal. They preferred all the control buttons related to the meals to be in the same area. The older users asked us to add text to the control buttons. Therefore, we added "TARA" (in Spanish) to the tare button.

The second version of the SMARTCLOTH web prototype, shown in Figure 3, changed the composition. As suggested in the previous test, the screen and the group of food buttons were swapped. Besides, we added new buttons related to cooked or raw foods. The clinical team requested this for the project. We decided to move the add dish and remove dish buttons below the weight, beside the Tare and Save buttons, as indicated in the previous test. Those two new buttons (cooked and raw) were placed on top of the weight. We added text inside the buttons describing the functionality of each of them, as mentioned in the previous test.

Figure 3. The second evolution of the SMARTCLOTH web prototype shown to users.

In the second evolution test, we noticed that the Tare button was useless in how users interacted with the prototype. The system did the tare when a new group of food buttons was pressed. Therefore, we removed the Tare button and the necessity of pressing 2 buttons.

Besides, users asked us to remove them from the final prototype, as neither barcode nor Bluetooth functionalities were included. In the same way, as the power consumption was not simulated in the web prototype, the battery icon was removed in the SMARTCLOTH final prototype. All these changes led us to the final SMARTCLOTH web prototype, shown in Figure 4.

XSL•FO
RenderX

Figure 4. Final testing of SMARTCLOTH HTML5-compliant web prototype.

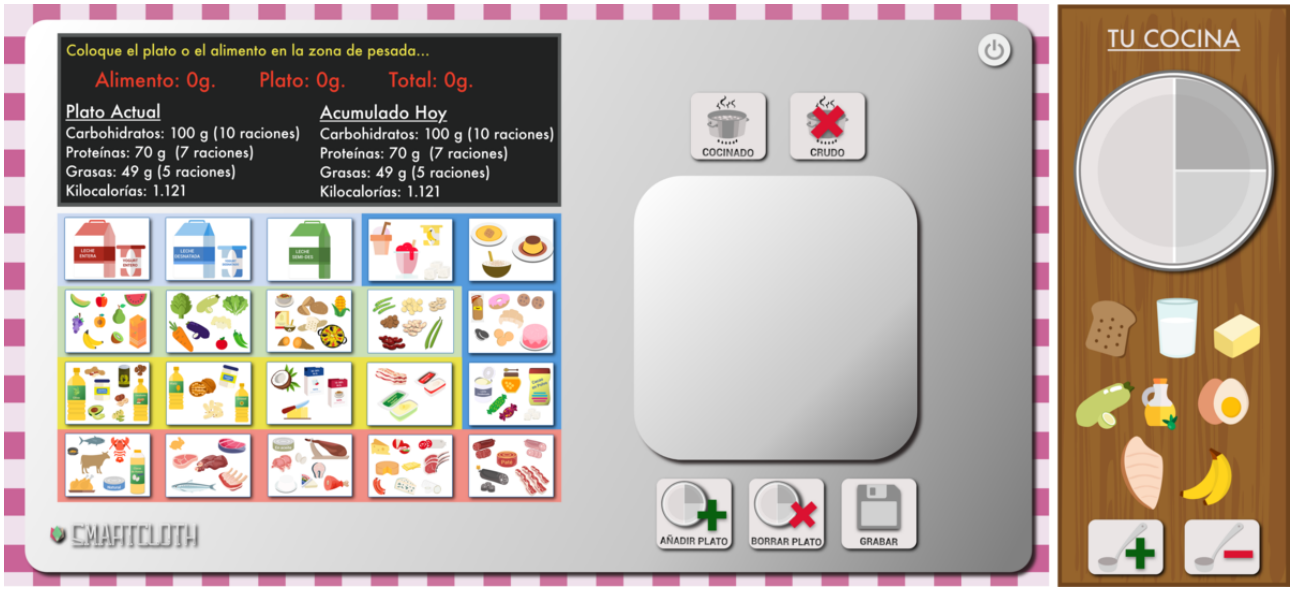

Testing

Three sessions were carried out. Every session was recorded and analyzed afterwards.

Tests 1 and 2

The SMARTCLOTH first web prototype, shown in Figure 2, was used in tests 1 and 2. The first test was a meal of 150 g of apple, and the second was 300 g of salmon. The analysis of the recorded video provided the results shown in Multimedia Appendix 6. Each user was anonymized by the acronym profile (P1/P2/P3) and user (1-5). For instance, P1.3 represents user 3 of profile 1 (T1DM).

Tests 3 and 4

The second version of the SMARTCLOTH web prototype, shown in Figure 3, was used in tests 3 and 4. Test 3 consisted of 150 g of apple, 250 g of yogurt, and 100 g of nuts. Test 4 consisted of 150 g of pasta, 50 g of tomato sauce, and 50 g of tuna. The analysis of the recorded video provided the results shown in Multimedia Appendix 7.

Test 5

The final version of the SMARTCLOTH web prototype, shown in Figure 4, was used in test 5, a complete meal composed of 3 dishes. The analysis of the recorded video provided the results shown in Multimedia Appendix 8.

Figure 5 shows the average time of the last try and the average time of the tries for each test, according to user profiles.

XSL•FO
RenderX

Figure 5. Average completion times for user tests. Bars represent the average time of the tries for each test and user type, while the lines represent the average time of the last try for each test and user type. T1DM: type 1 diabetes mellitus; T2DM: type 2 diabetes mellitus.

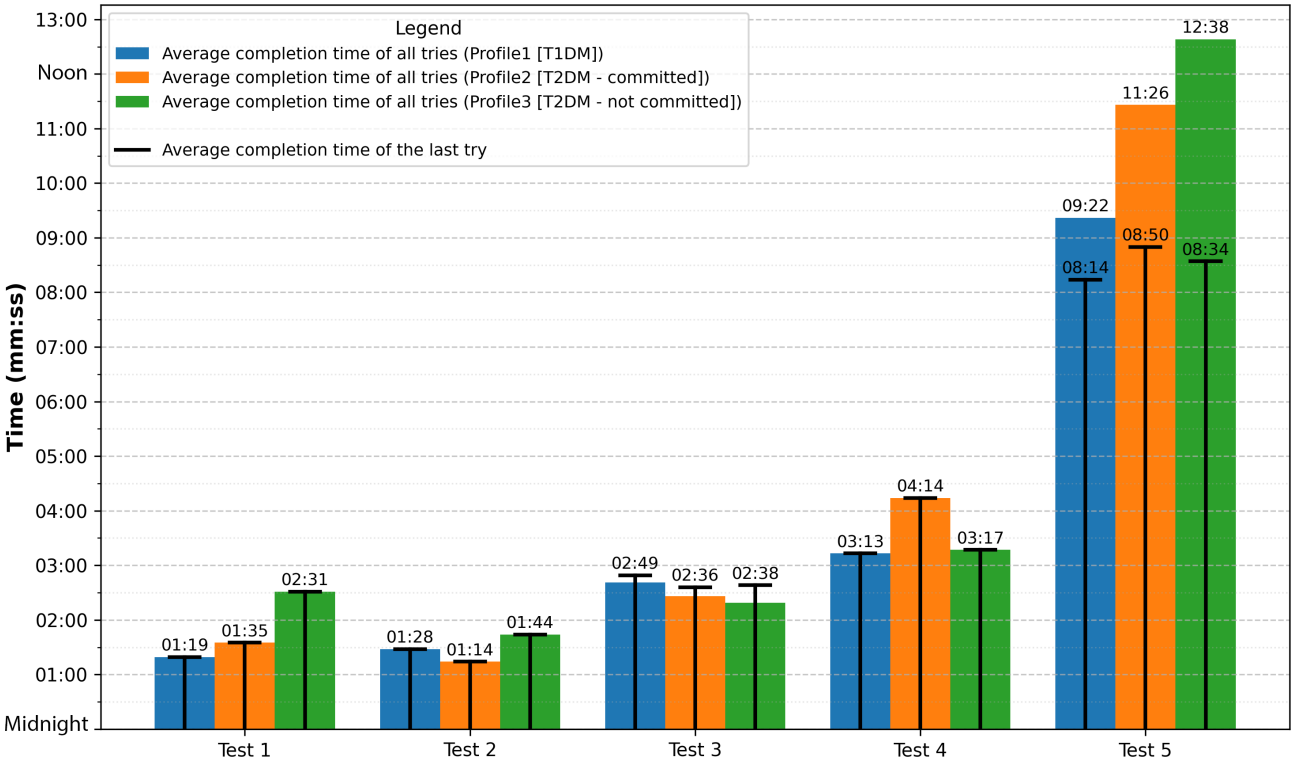

Discussion

Principal Findings

This tutorial has demonstrated the systematic application of HCD and DT methodology to develop digital health hardware, using SMARTCLOTH as an illustrative case example.

This DT methodology is widely used to create technological solutions for chronic health problems. For example, the study by Petersen and Hempler et al [58] used it to develop a mobile app that helps patients with diabetes manage their disease more effectively. The researchers found that the app significantly improved glycemic control and adherence to treatment. Another study used this methodology to design a telemedicine platform for patients with hypertension, resulting in a marked improvement in blood pressure monitoring and patient satisfaction [59]. Like our work, both studies showed optimal results, highlighting the effectiveness of DT in developing technological solutions for managing chronic diseases, where treatment adherence has been shown to be a key element.

Reflections on the HCD Process

Mapping

Team Mapping

Concerning the research team's preassessment, despite the threats and weaknesses identified, this technique offered an interesting approach for researchers to maximize their strengths and take advantage of the opportunities available. The recent pandemic presented significant challenges, coupled with the difficulty in changing dietary habits and competition with similar products on the market. In this sense, an honest and strategic assessment of the project's human, temporal, and economic capacities beforehand allowed the development of all stages of the DT methodology to be adjusted. Despite this, a dearth of studies uses or reports the results of these preassessment strategies. However, the orientation and focus that originated in the team underlines the need for robust methods to identify the potential of interdisciplinary groups and concretize the work in technological projects to improve adherence to lifestyles in chronic diseases. Recent studies have highlighted similar challenges and proposed similar solutions to overcome identified weaknesses in health-related technology projects [60-62].

User Profiles Mapping

Using end-user-centered design tools, such as "persona," "empathy mapping," and "customer journey mapping" has been particularly interesting. Traditionally used in business and the private sector to improve the benefits of business propositions, these tools are applied novelly in the health sector. However, this approach has been evidenced to help increase usability and adherence to use by patients with chronic conditions. For example, a recent study used the persona tool to design a medication-tracking app for patients with cardiovascular disease, significantly improving treatment adherence [63]. Another study implemented the "customer journey map" in developing a telerehabilitation program for patients with chronic obstructive pulmonary disease, finding that the tool helped identify friction points and improve the user experience, thereby increasing patient engagement and satisfaction [64]. Similarly, using the "empathy map" in developing a support platform for patients with diabetes allowed developers to better understand users' emotional and functional needs, leading to a more intuitive and effective design [65]. Like the studies discussed above, our work has also found these tools very useful in understanding and segmenting our users, having identified three profiles (1)

XSL•FO
RenderX

First profile comprised one of adolescents and their caregivers with T1DM, who manage their insulin well but face social and dietary challenges; (2) a second profile, of mature adults with T2DM who, while having difficulty maintaining a diet, try to engage with treatment; and (3) a third profile, corresponding to older people with T2DM and with low adherence to treatment and severe difficulties in modifying lifestyles.

Comparing these profiles with those reported in the literature, we found important similarities and differences. For example, a study on medication adherence in patients with diabetes identified that levels of family support play a crucial role in adherence to treatment. Patients who perceive greater instrumental support from family members, such as help with diet and medical appointments, tend to show better adherence to treatment regimens [66,67]. This finding aligns with our second profile, where family support is a key facilitator. Another study highlighted that barriers, such as lack of knowledge about diabetes and unsupportive family behaviors, such as negative interference with medication adherence, are significantly associated with poorer glycemic control outcomes [68]. This finding is particularly relevant for our third profile of older people with T2DM, who often have difficulty following dietary recommendations due to ingrained dietary habits and lack of structured support [69].

In addition, implementing continuous quality improvement programs in diabetes management has been shown to improve treatment adherence, promote self-care, and increase patients' knowledge about their disease [66,70]. This strategy could benefit all the profiles identified in our study, providing a constant support structure and ongoing education.

Ultimately, these tools have allowed us to identify each group's specific needs and barriers, facilitating the design of the functionalities that SMARTCLOTH should present as the most effective and personalized solutions to improve adherence and dietary management of diabetes. These were the key issues addressed in the next stage.

Exploring

A determining factor in the design of a technological solution is its adaptation to the needs of the end users. In this sense, during the exploring stage, factors that facilitate or hinder adherence to a diet that helps glycemic control in patients with diabetes were detected. At this point, we emphasized those identified barriers or facilitators relevant to the design of the proposed technological solution.

Among the 3 profiles identified, the environment (social events, eating outside the home, family pressure, constant exposure to unhealthy foods...) appears as a common element that gets in the way of following an adequate diet, something described in studies carried out in other countries [71,72]. This situation makes glycemic control difficult, either by eating high-calorie meals, high consumption of proteins and/or fats that alter glucose absorption, or difficulty counting carbohydrates ingested [73,74].

Of the elements mentioned in the previous paragraph, the high availability and exposure to high-calorie foods, unhealthy fats, and free sugars, especially ultraprocessed foods, are particularly important today. Studies show that consumption of these foods is very high [75], so it must be assumed that SMARTCLOTH users will incorporate them into their diet. In this context, a barcode scanner was envisaged to monitor the intake of these products accurately and to increase control and awareness of what is involved in eating them [76].

On the other hand, the difficulty in counting carbohydrates is not only limited to social life but also at specific times (the first months after diagnosis of the disease, adolescence, after years with the disease...), there is insecurity and/or relaxation due to lack of training, in the first case, or fatigue due to the tediousness of the process required to do so, in the second [77,78]. The counting of this macronutrient, especially in patients with T1DM, is fundamental in glycemic control, as it defines the insulin dose [79]. Therefore, the mainstay of SMARTCLOTH is the calculation of grams and servings of the 3 immediate principles and the monitoring of intakes throughout the day. Farooqi et al [80] found that this type of functionality is highly relevant in diabetes management and in increasing adherence to the necessary therapeutic approaches, as also found by other authors [81-83].

Furthermore, the ability to structure an adequate diet is an important limitation for most of the population. During the focus groups, patients with T2DM openly expressed this problem [71,84]. This difficulty is a major problem, as the correct dietary therapeutic approach (quality food, reduction of free sugars, moderate carbohydrate diet, and reduction of caloric intake) is essential to improve metabolic control [85]. In this context, SMARTCLOTH presents itself as an ideal tool for dietary management. In addition to the aforementioned functions, the digital tablecloth has to allow the monitoring of caloric intake at each meal and accumulated caloric intake over a day, which will allow moderation of energy intake and loss of body fat, which has emerged as the main goal set by health care workers for patients with T2DM [86]. However, SMARTCLOTH should be accompanied by dietary recommendations made by qualified personnel tailored to the individual patient's needs.

A lack of culinary knowledge also limits the ability to structure and follow an appropriate diet [87,88]. This shortage means that most meals are not organoleptically comparable to those offered by the food industry and require more preparation time [89,90]. Therefore, taking advantage of the barcode reader, the researchers decided to design a book and a database of standardized healthy recipes that would (1) improve the quality of meals, thanks to a better choice of foods; (2) reduce the time spent in the kitchen, through simple recipes; (3) increase the palatability of meals, through the introduction of very tasty recipes; and (4) reduce the difficulty of controlling meals, because the simple reading of the barcode associated with the recipe will allow the portion to be weighed directly without the need to consider the ingredients separately.

Monitoring blood glucose and its coupling with dietary intake is of particular concern to health care staff caring for this group of patients [91-93]. In this case, SMARTCLOTH will upload all intakes recorded on the device to a database. This information will be uploaded to a website, allowing the patient to enter their blood glucose and graphically observe the correlation between

XSL•FO
RenderX

blood glucose and intake. This functionality can help to establish a pattern or detect errors in the patient's self-care and help to intervene more precisely.

Finally, it should be noted that although SMARTCLOTH is designed and developed according to the needs expressed by patients, motivation will play a key role in its use, as, like any intervention, it will require commitment from users [94-96]. As evidenced throughout this research, it is unlikely that those who do not want to change their habits (profile 3) are likely to do so.

Undoubtedly, SMARTCLOTH will not address all the barriers that were identified, especially some related to the environment. This is because the final prototype will not be of a size that allows it to be used outside the home. However, a solution may be offered by future iterations of SMARTCLOTH, which may be more portable so that it can be used outside the home.

Building

Initially, the prototype was planned to be developed physically. However, adopting the web-based implementation strategy has greatly benefited the project. Web technologies have made it possible to speed up prototyping and make changes at no economic cost.

Three significant evolutions of the prototype (with multiple small internal modifications) have occurred. Agile methodologies guided these evolutions, and mini prototypes were rapidly developed. The development team tested these internally, and if they passed the quality standards, they were integrated into the prototype to be tested by the users.

Two factors limited the set of changes applied to the prototypes. The main limiting factor was the time until the next user testing session. The next factor that recommended limiting the set of changes in each prototype was the need to evaluate the impact on the users while maintaining the functionality, as including many changes could make detailed evaluation too tricky.

Due to this prototyping process's exploratory and discovery nature, it cannot be separated from the testing stage. The experience gained from this project has revealed that the methodology for developing the best prototype cannot be separated into watertight stages. Being user-centered, the process of specification, development, testing, and respecification must be continuous and cannot be disaggregated. Users' own evaluations generate proposals for modifications, which, in turn, require agility in development, with a very rapid process of specifying new requirements. In many cases, this entire process was reduced to 10 minutes. This extreme speed did not allow for a standard process of requirements gathering, analysis, design, and reimplementation. This section followed a software development paradigm [41] that did not precisely conform to the spiral or rapid prototyping paradigm, although it took many elements from both approaches.

Because of this rapidity, many of the usual thorough error-checking processes cannot be applied. This means that the developed prototypes should not be used for unsupervised external deployments but only for controlled testing. In other words, prototypes should be discarded and only be used to extract fundamental information for constructing the final system.

In short, all these evolutions have made it possible to substantially change the functioning and the layout of the different visual and information input elements.

Testing

At this point, it is necessary to differentiate between testing and user evaluation. Testing aims to verify the code. Therefore, tests ensure that the code does not contain errors and that the execution provides consistent results. User evaluation seeks validation of the product, that is, the degree of user acceptance of the system. This user evaluation does not seek the correctness of the results, as this is a later phase. Once the product has been developed per the requirements gathered from the users, it can be used by the research team to achieve the project's objectives.

Focusing on the internal tests, we must distinguish between those carried out before and during the user evaluations. The developments carried out before the evaluations had a short time frame, less than a month, but with the capacity to carry out a sufficient set of tests on the code. These tests detected lexical and syntactic coding faults and even validated results with output verification and acceptable ranges. The testing procedure followed the usual software engineering standards [97]. The code developed before the user evaluations had minor coding errors in specific performances that were difficult to trace. It can be considered good quality code with respect to the specifications previously established for each prototype evolution. However, due to time constraints, code testing was not performed for code modifications made during the evaluations. In this respect, it is essential that the developer involved have extensive experience in programming, in the JavaScript programming language, in Phaser 3, and in the development of web-based systems. The quality of this code varied considerably, although, in general, the code developed could pass the minimum quality requirements.

With regard to user evaluations, focusing on the product development aspect and the acquisition and adjustment of user requirements associated with the system, the procedure had to be more systematic. Follow-up documents were produced and filled in with the subjective evaluation of the user tests. In addition, video recordings were made to allow for a more analytical review of the user tests. In the documents and the video, we did not ask the same questions in all cases, so the users' answers may have been somewhat biased in some cases. However, we think this bias did not affect the sense of the suggested modifications, since the indications were mostly oriented to help users with the assignment of the different foods to the food group buttons. Only when users got stuck in the procedure for too long was full guidance provided until the process was completed. The evaluation of the reasons for the impossibility of completing the process was analyzed. In some cases, the size of the buttons and the lack of knowledge about the assignment of foods to food groups were problematic.

In general, user tests allowed the prototypes to evolve into the final design, which saved the project a lot of time and money.

XSL•FO
RenderX

Although SMARTCLOTH is conceived as a hardware device rather than a mobile app, its integrated intelligent functionalities—such as real-time conversion of food weights into macronutrient portions, barcode or QR scanning, and automated calorie estimation—are expected to reinforce both glycemic control and dietary adherence. From a clinical perspective, improved precision in dietary monitoring is anticipated to contribute to reductions in HbA_{1c}, as demonstrated in studies evaluating technology-based wearable interventions compared with standard care [98]. Beyond glycemic outcomes, SMARTCLOTH could also increase adherence by simplifying complex dietary tasks, reducing the cognitive burden of carbohydrate counting, and making the process more understandable. Usability and clarity of device interfaces have been shown to influence treatment adherence in diabetes care, with user-friendly designs being associated with better long-term engagement [99]. This characteristic is particularly relevant during the early stages after diagnosis, when patients often report errors and insecurity in carbohydrate counting, and an assistive hardware device may function as both an educational and behavioral reinforcement tool. Nevertheless, the translation of these functionalities into measurable clinical outcomes will depend on sustained device engagement, data completeness, and seamless integration into daily routines.

Methodological Reflections: Limitations and Strengths of the HCD Approach

These reflections on the limitations and strengths of our HCD application provide important considerations for other teams adapting this methodology to their own hardware development projects.

This project presents several strengths that enhance the validity and applicability of its findings. First, the use of an HCD methodology ensured that the prototype was developed based on the real needs and experiences of patients and health care professionals, thereby increasing usability and acceptance. Second, the iterative prototyping and testing process enabled continuous refinements, contributing to a functional and user-friendly system. Third, the inclusion of diverse patient profiles (adolescents with T1DM, working-age adults with T2DM, and older adults with low adherence) provided a comprehensive understanding of different user needs and improved the generalizability of the results.

Nevertheless, several limitations should be acknowledged. The user testing involved a relatively small sample, which may restrict representativeness. Evaluations were conducted in controlled settings rather than during everyday use, limiting inferences about long-term adoption. Finally, the current prototype is oriented toward home use, which may limit its applicability in outdoor contexts where dietary management can be especially challenging.

Although the HCD approach enabled us to gather a wide range of ideas and proposals from patients and professionals, not all of them proved feasible—whether due to technological or economic constraints of the project—or guaranteed a significant clinical impact. This fact is an inherent limitation of participatory processes in complex areas, such as diabetes management, particularly when patients express needs that are difficult to address without a highly active role on their part (eg, motivation). Nevertheless, the iterative nature of the design process and the triangulation with health care professionals allowed us to refine these contributions, prioritizing those that were more feasible and had greater potential applicability. In this way, HCD was useful not only for capturing perceptions and needs but also for guiding a process of selection and refinement oriented toward realistic technological solutions aligned with the precise needs of users.

Limitations and Adaptability Considerations

The methodology presented reflects our experience with a specific clinical context (diabetes dietary management) and available resources (an academic research team with engineering and clinical expertise). Teams adopting this approach should consider the following:

1. Resource allocation: Our timeline (8 months from mapping to testing) may need adjustment based on team size and funding.
2. Technical expertise: We chose web-based prototyping (Phaser 3) for accessibility; teams with different capabilities might select alternative frameworks.
3. User recruitment: Access to patient engagement boards or clinical partnerships will influence the feasibility of iterative testing.
4. Cultural and linguistic adaptation: Our Spanish-speaking population required specific interface considerations that will differ in other contexts.

The methodology is broadly applicable to other chronic disease management contexts and health care hardware development scenarios. By sharing our experiences, decision-making processes, and lessons learned, we aim to support other multidisciplinary teams in developing user-centered digital health solutions. Future work should explore adaptations of this approach for resource-constrained settings, integration with established clinical workflows, and methods for scaling from prototype to clinical implementation and evaluation.

Conclusions

This tutorial has provided step-by-step methodological guidance for applying HCD and the Double Diamond model to health care hardware development. Through the SMARTCLOTH case example, we have demonstrated how systematic user engagement, iterative prototyping, and structured usability testing can inform the development of digital health solutions tailored to patient needs. Key methodological contributions of this tutorial include (1) a replicable framework for user needs assessment combining multiple HCD tools (persona, empathy mapping, and customer journey mapping) that can be adapted to different chronic disease contexts; (2) strategies for synthesizing multidisciplinary input from clinical and technical team members through structured cocreation sessions; (3) cost-effective prototyping approaches using web-based frameworks (Phaser3/HTML5) that enable rapid iteration before committing to physical manufacturing; (4) structured usability testing protocols combining quantitative metrics (task completion time and error rates) with qualitative feedback

XSL•FO
RenderX

(think-aloud technique) across diverse user populations; and (5) practical guidance on adapting design requirements for users with varying technological literacy levels, including specific accommodations for older adults and users with limited digital experience.

In our specific application to diabetes dietary management, the resulting SMARTCLOTH prototype demonstrated strong usability and user acceptance across 3 distinct user profiles. The iterative development process, guided by continuous patient and professional input, yielded a tool perceived as intuitive and potentially valuable for dietary self-management support.

The methodology presented in this tutorial is broadly applicable to other chronic disease management contexts and health care hardware development scenarios where user-centered solutions are needed. By sharing our experiences, decision-making processes, and lessons learned—including both successes and limitations—we aim to support other multidisciplinary teams in developing effective digital health hardware.

Future applications of this methodology could explore adaptations for different resource settings, integration with existing clinical workflows and electronic health record systems, methods for scaling from prototype to clinical implementation, and approaches for conducting effectiveness evaluations that measure both usability and clinical outcomes. For SMARTCLOTH specifically, next steps include larger-scale usability testing, evaluation of clinical effectiveness on dietary adherence and glycemic control, and development of portable versions suitable for use beyond the home environment.

Acknowledgments

The authors thank all patients, caregivers, and health care professionals who participated in this study for their time and valuable contributions.

Funding

This research (PI21/01602) has been made possible because of the Strategic Health Action 2021 call, funded by the "Instituto de Salud Carlos III (Carlos III Health Institute)" and cofunded by the European Union.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the sensitive and potentially identifiable nature of these materials (qualitative data, including transcripts of interviews, focus groups, and video recordings) and to protect participant privacy and confidentiality but are available from the corresponding author on reasonable request and subject to appropriate ethical approval and data-sharing agreements.

Authors' Contributions

Conceptualization: GMR (lead), RML (equal)
Data curation: MPVE (focus groups), JMP (prototype testing)
Formal analysis: MGR (focus groups), FLG (prototype testing; lead), ICR (prototype testing; equal)
Funding acquisition: GMR
Investigation: GMR (focus groups, lead), RML (focus groups, equal), MPVE (focus groups, equal), JMP (prototype software design), FLG (prototype hardware design)
Methodology: GMR (lead), JMP (equal), FLG (equal)
Project administration: GMR (lead), RML (equal)
Resources: JMP (lead), FLG (equal), MGR (equal)
Software: JMP (lead), JMP (equal); IGR (supporting)
Supervision: GMR
Validation: GMR (lead), JMP (equal), RML (supporting)
Visualization: GMR (lead), FLG (supporting)
Writing–original draft: JMP, RML, FLG, and GMR (all four equally)
Writing–review & editing: RML and GMR

Conflicts of Interest

None declared.

Multimedia Appendix 1
General phases in the SMARTCLOTH development project. Adapted from Gasca and Zaragoza´, 2021.
[PNG File, 250 KB - jmir_v28i1e75744_app1.png]

Multimedia Appendix 2
Different user tests and patients using SMARTCLOTH.

XSL•FO
RenderX

[PNG File, 908 KB - jmir_v28i1e75744_app2.png]

Multimedia Appendix 3
Grid of actions for user test 2 task using SMARTCLOTH.
[PNG File, 248 KB - jmir_v28i1e75744_app3.png]

Multimedia Appendix 4
Results of In/Out and Strengths, Weaknesses, Opportunities, and Threats.
[PNG File, 1777 KB - jmir_v28i1e75744_app4.png]

Multimedia Appendix 5
Persona, empathy map, and consumer journey map.
[PDF File, 111657 KB - jmir_v28i1e75744_app5.pdf]

Multimedia Appendix 6
Tests 1 and 2 results.
[DOCX File, 17 KB - jmir_v28i1e75744_app6.docx]

Multimedia Appendix 7
Tests 3 and 4 results.
[DOCX File, 17 KB - jmir_v28i1e75744_app7.docx]

Multimedia Appendix 8
Test 5 results.
[DOCX File, 16 KB - jmir_v28i1e75744_app8.docx]

References

1. Brutsaert EF. Diabetes mellitus. MSD Manual. 2023 Oct. URL: https://acortar.link/jyD5sn [accessed 2024-05-29]
2. Diabetes. World Health Organziation. 2023 Apr. URL: https://acortar.link/EGwDd [accessed 2024-05-29]
3. IDF Diabetes Atlas 2025.: International Diabetes Federation URL: https://diabetesatlas.org/resources/idf-diabetes-atlas-2025/ [accessed 2025-09-22]
4. Nowakowska M, Zghebi SS, Ashcroft DM, et al. The comorbidity burden of type 2 diabetes mellitus: patterns, clusters and predictions from a large English primary care cohort. BMC Med 2019 Jul 25;17(1):145. [doi: 10.1186/s12916-019-1373-y] [Medline: 31345214]
5. ElSayed NA, McCoy RG, Aleppo G, et al. Summary of revisions: standards of care in diabetes—2025. Diabetes Care 2025 Jan 1;48(Supplement_1):S6-S13. [doi: 10.2337/dc25-SREV]
6. Lemieux I. Reversing type 2 diabetes: the time for lifestyle medicine has come!. Nutrients 2020 Jul 3;12(7):1974. [doi: 10.3390/nu12071974] [Medline: 32635141]
7. Sami W, Ansari T, Butt NS, Hamid MRA. Effect of diet on type 2 diabetes mellitus: a review. Int J Health Sci (Qassim) 2017;11(2):65-71. [Medline: 28539866]
8. Pérez Unanua MP, Alonso Fernández M, López Simarro F, et al. Adherence to healthy lifestyle behaviours in patients with type 2 diabetes in Spain. SEMERGEN 2021 Apr;47(3):161-169. [doi: 10.1016/j.semerg.2020.08.009] [Medline: 33160855]
9. Chong S, Ding D, Byun R, Comino E, Bauman A, Jalaludin B. Lifestyle changes after a diagnosis of type 2 diabetes. Diabetes Spectr 2017 Feb;30(1):43-50. [doi: 10.2337/ds15-0044] [Medline: 28270714]
10. Gillespie SJ, Kulkarni KD, Daly AE. Using carbohydrate counting in diabetes clinical practice. J Am Diet Assoc 1998 Aug;98(8):897-905. [doi: 10.1016/S0002-8223(98)00206-5] [Medline: 9710660]
11. Hanlon P, Daines L, Campbell C, McKinstry B, Weller D, Pinnock H. Telehealth interventions to support self-management of long-term conditions: a systematic metareview of diabetes, heart failure, asthma, chronic obstructive pulmonary disease, and cancer. J Med Internet Res 2017 May 17;19(5):e172. [doi: 10.2196/jmir.6688] [Medline: 28526671]
12. Moreno-Ligero M, Moral-Munoz JA, Salazar A, Failde I. mHealth intervention for improving pain, quality of life, and functional disability in patients with chronic pain: systematic review. JMIR Mhealth Uhealth 2023 Feb 2;11:e40844. [doi: 10.2196/40844] [Medline: 36729570]
13. Sakane N, Suganuma A, Domichi M, et al. The effect of a mHealth app (KENPO-app) for specific health guidance on weight changes in adults with obesity and hypertension: pilot randomized controlled trial. JMIR Mhealth Uhealth 2023 Apr 12;11:e43236. [doi: 10.2196/43236] [Medline: 37043287]
14. Vlahu-Gjorgievska E, Burazor A, Win KT, Trajkovik V. mHealth apps targeting obesity and overweight in young people. App review and analysis. JMIR Mhealth Uhealth 2023 Jan 19;11:e37716. [doi: 10.2196/37716] [Medline: 36656624]

XSL•FO
RenderX

15. Pulman A, Taylor J, Galvin K, Masding M. Ideas and enhancements related to mobile applications to support type 1 diabetes. JMIR Mhealth Uhealth 2013 Jul 25;1(2):e12. [doi: 10.2196/mhealth.2567] [Medline: 25100684]
16. Lamprinos I, Papadaki C, Schmuhl HH, Demski H, Hildebrand C, Plößnig M. Mobile personal health application for empowering diabetic patients. J Int Soc Telemed eHealth 2014;2(1):3-11 [FREE Full text]
17. Jakob R, Harperink S, Rudolf AM, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371. [doi: 10.2196/35371] [Medline: 35612886]
18. Wang X, Shu W, Du J, et al. Mobile health in the management of type 1 diabetes: a systematic review and meta-analysis. BMC Endocr Disord 2019 Feb 13;19(1):21. [doi: 10.1186/s12902-019-0347-6] [Medline: 30760280]
19. Kebede MM, Zeeb H, Peters M, Heise TL, Pischke CR. Effectiveness of digital interventions for improving glycemic control in persons with poorly controlled type 2 diabetes: a systematic review, meta-analysis, and meta-regression analysis. Diabetes Technol Ther 2018 Nov;20(11):767-782. [doi: 10.1089/dia.2018.0216] [Medline: 30257102]
20. Martínez-Pérez B, de la Torre-Díez I, López-Coronado M. Mobile health applications for the most prevalent conditions by the World Health Organization: review and analysis. J Med Internet Res 2013 Jun 14;15(6):e120. [doi: 10.2196/jmir.2600] [Medline: 23770578]
21. Molina-Recio G, Molina-Luque R, Romero-Saldaña M. The importance of knowing and listening to all those involved in the design and use of nutrition mobile apps. Getting to know the Great GApp. Nutr Hosp 2021 Jun 10;38(3):555-562. [doi: 10.20960/nh.03385] [Medline: 33813833]
22. Saparamadu A, Fernando P, Zeng P, et al. User-centered design process of an mHealth app for health professionals: case study. JMIR Mhealth Uhealth 2021 Mar 26;9(3):e18079. [doi: 10.2196/18079] [Medline: 33769297]
23. Lövestam E, Vivanti A, Steiber A, et al. The International Nutrition Care Process and Terminology Implementation Survey: towards a Global Evaluation Tool to assess individual practitioner implementation in multiple countries and languages. J Acad Nutr Diet 2019 Feb;119(2):242-260. [doi: 10.1016/j.jand.2018.09.004] [Medline: 30552017]
24. Atkins L, Michie S. Designing interventions to change eating behaviours. Proc Nutr Soc 2015 May;74(2):164-170. [doi: 10.1017/S0029665115000075] [Medline: 25998679]
25. Classification of digital interventions, services and applications in health: a shared language to describe the uses of digital technology for health, 2nd ed. World Health Organization. 2023. URL: https://www.who.int/publications/i/item/9789240081949 [accessed 2025-12-08]
26. Daly A, Hovorka R. Technology in the management of type 2 diabetes: present status and future prospects. Diabetes Obes Metab 2021 Aug;23(8):1722-1732. [doi: 10.1111/dom.14418] [Medline: 33950566]
27. Levander XA, VanDerSchaaf H, Barragán VG, et al. The role of human-centered design in healthcare innovation: a digital health equity case study. J Gen Intern Med 2024 Mar;39(4):690-695. [doi: 10.1007/s11606-023-08500-0] [Medline: 37973709]
28. Dopp AR, Parisi KE, Munson SA, Lyon AR. Aligning implementation and user-centered design strategies to enhance the impact of health services: results from a concept mapping study. Implement Sci Commun 2020;1(1):17. [doi: 10.1186/s43058-020-00020-w] [Medline: 32885179]
29. Kaur N, Pluye P. Delineating and operationalizing the definition of patient-oriented research: a modified e-Delphi study. J Patient Cent Res Rev 2019;6(1):7-16. [doi: 10.17294/2330-0698.1655] [Medline: 31414019]
30. Crowe B, Gaulton JS, Minor N, et al. To improve quality, leverage design. BMJ Qual Saf 2022 Jan;31(1):70-74. [doi: 10.1136/bmjqs-2021-013605] [Medline: 34510018]
31. Chen E, Neta G, Roberts MC. Complementary approaches to problem solving in healthcare and public health: implementation science and human-centered design. Transl Behav Med 2021 May 25;11(5):1115-1121. [doi: 10.1093/tbm/ibaa079] [Medline: 32986098]
32. Dopp AR, Parisi KE, Munson SA, Lyon AR. Integrating implementation and user-centred design strategies to enhance the impact of health services: protocol from a concept mapping study. Health Res Policy Sys 2019 Dec;17(1):1-13. [doi: 10.1186/s12961-018-0403-0]
33. Hookway S, Johansson MF, Svensson A, Heiden B. The problem with problems: reframing and cognitive bias in healthcare innovation. Des J 2019 Apr 1;22(sup1):553-574. [doi: 10.1080/14606925.2019.1595438]
34. Altman M, Huang TTK, Breland JY. Design thinking in health care. Prev Chronic Dis 2018 Sep 27;15:E117. [doi: 10.5888/pcd15.180128] [Medline: 30264690]
35. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119. [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
36. Sieck CJ, Sheon A, Ancker JS, Castek J, Callahan B, Siefer A. Digital inclusion as a social determinant of health. NPJ Digit Med 2021 Mar 17;4(1):52. [doi: 10.1038/s41746-021-00413-8] [Medline: 33731887]
37. Larusdottir M, Cajander Å, Roto V. User-centered design approaches and software development processes. In: Vanderdonckt J, Palanque P, Winckler M, editors. Handbook of Human Computer Interaction: Springer; 2023:1-4. [doi: 10.1007/978-3-319-27648-9_104-1]
38. Baker FW III, Moukhliss S. Concretising design thinking: a content analysis of systematic and extended literature reviews on design thinking and human-centred design. Rev Educ 2020 Feb;8(1):305-333. [doi: 10.1002/rev3.3186]

XSL•FO
RenderX

39. Rösch N, Tiberius V, Kraus S. Design thinking for innovation: context factors, process, and outcomes. Eur J Innov Manag 2023 Dec 18;26(7):160-176. [doi: 10.1108/EJIM-03-2022-0164]
40. Pomar P. Differences between design thinking and human centered design. Thinkernautas. 2017 May 27. URL: https://thinkernautas.com/diferencias-design-thinking-human-centered-design [accessed 2025-12-08]
41. Pressman R, Maxim B. Software Engineering: A Practitioner's Approach, 9th edition: McGraw-Hill Education; 2020.
42. The double diamond. Design Council. 2019. URL: https://www.designcouncil.org.uk/our-resources/the-double-diamond/ [accessed 2024-05-27]
43. Gasca J. Designpedia: 80 Herramientas Para Construir Tus Ideas [Designpedia: 80 Tools to Build Your Ideas], 5th edition: LID editorial; 2021.
44. Gasca J, Zaragozá R. El Workbook de Designpedia [Workbook of Designpedia], 1st edition: LID editorial.
45. Pruitt J, Adlin T. The Persona Lifecycle: Keeping People in Mind Throughout Product Design, 2nd edition: Morgan Kaufmann; 2006. [doi: 10.1016/B978-012566251-2/50002-2]
46. Gray D. The empathy map canvas. Gamestorming. 2017. URL: https://gamestorming.com/wp-content/uploads/2017/07/Empathy-Map-Canvas-006.pdf [accessed 2025-12-23]
47. Kalbach J. Mapping Experiences: A Complete Guide to Creating Value through Journeys, Blueprints, and Diagrams, 2nd edition: O'Reilly Media; 2020.
48. Lewrick M, Link P, Leifer L. The Design Thinking Toolbox: A Guide to Mastering the Most Popular and Valuable Innovation Methods: Wiley; 2020.
49. Stickdorn M, Hormess ME, Lawrence A, Schneider J. This Is Service Design Doing: Applying Service Design Thinking in the Real World, 2nd edition: O'Reilly Media; 2020. URL: https://www.thisisservicedesigndoing.com/ [accessed 2026-01-12]
50. Tan H, Wilson A, Olver I. Ricoeur's theory of interpretation: an instrument for data interpretation in hermeneutic phenomenology. Int J Qual Methods 2009 Dec;8(4):1-15. [doi: 10.1177/160940690900800401]
51. Osborn AF. Applied Imagination: Principles and Procedures of Creative Thinking, 4th edition: Creative Education Foundation; 2020.
52. Sanders EBN, Stappers PJ. Convivial Toolbox: Generative Research for the Front End of Design, 2nd edition: BIS Publishers; 2020.
53. McElroy K. Prototyping for Designers: Developing the Best Digital and Physical Products, 2nd edition: O'Reilly Media; 2017.
54. Rutter T. Designing Interfaces: Patterns for Effective Interaction Design, 3rd edition: O'Reilly Media; 2021.
55. Fonteyn ME, Kuipers B, Grobe SJ. A description of think aloud method and protocol analysis. Qual Health Res 1993 Nov;3(4):430-441. [doi: 10.1177/104973239300300403]
56. Ericsson KA, Simon HA. Protocol Analysis: Verbal Reports as Data Revised Ed: MIT Press; 2023.
57. Nielsen J, Molich R. Heuristic evaluation of user interfaces. 1990 Presented at: CHI '90: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; Apr 1-5, 1990; Seattle, WA p. 249-256. [doi: 10.1145/97243.97281]
58. Petersen M, Hempler NF. Development and testing of a mobile application to support diabetes self-management for people with newly diagnosed type 2 diabetes: a design thinking case study. BMC Med Inform Decis Mak 2017 Sep 12;17(1):133. [doi: 10.1186/s12911-017-0525-2] [Medline: 28899375]
59. Griffin AC, Khairat S, Bailey SC, Chung AE. A chatbot for hypertension self-management support: user-centered design, development, and usability testing. JAMIA Open 2023 Jul 4;6(3):ooad073. [doi: 10.1093/jamiaopen/ooad073]
60. Blandford A, Gibbs J, Newhouse N, Perski O, Singh A, Murray E. Seven lessons for interdisciplinary research on interactive digital health interventions. Digit Health 2018 Jan;4:2055207618770325. [doi: 10.1177/2055207618770325]
61. Krause-Jüttler G, Weitz J, Bork U. Interdisciplinary collaborations in digital health research: mixed methods case study. JMIR Hum Factors 2022 May 4;9(2):e36579. [doi: 10.2196/36579] [Medline: 35507400]
62. Fowe IE. Evaluating organizational readiness for change in the implementation of telehealth and mobile health interventions for chronic disease management. AMIA Jt Summits Transl Sci Proc 2021;2021(1):210-219. [Medline: 34457135]
63. Haldane V, Koh JJK, Srivastava A, et al. User preferences and persona design for an mHealth intervention to support adherence to cardiovascular disease medication in Singapore. JMIR Mhealth Uhealth 2019 May 28;7(5):e10465. [doi: 10.2196/10465] [Medline: 31140445]
64. Liu Y, Dickerson T, Early F, Fuld J, Jiang C, Clarkson PJ. Understanding the influences of COPD Patient's capability on the uptake of pulmonary rehabilitation in the UK through an inclusive design approach. Int J Chron Obstruct Pulmon Dis 2021;16(2):1717-1740. [doi: 10.2147/COPD.S305145] [Medline: 34168438]
65. Imhemed SM, Kamaruddin A, Atan R. Using empathy approach to design type 2 diabetic user persona. J Adv Sci Eng Res 2021;7(1):27-42 [FREE Full text]
66. Horta C, Quaresma G, Lucas P. Adherence to therapeutic regimen for people with diabetes through the implementation of continuous quality improvement projects – Scoping Review. New Trends Qual Res 2022;13:1-11. [doi: 10.36367/ntqr.13.2022.e678]
67. Mayberry LS, Osborn CY. Family involvement is helpful and harmful to patients' self-care and glycemic control. Patient Educ Couns 2014 Dec;97(3):418-425. [doi: 10.1016/j.pec.2014.09.011] [Medline: 25282327]

XSL•FO
RenderX

68. Mayberry LS, Osborn CY. Family support, medication adherence, and glycemic control among adults with type 2 diabetes. Diabetes Care 2012 Jun 1;35(6):1239-1245. [doi: 10.2337/dc11-2103]
69. Beverly EA, Ritholz MD, Brooks KM, et al. A qualitative study of perceived responsibility and self-blame in type 2 diabetes: reflections of physicians and patients. J Gen Intern Med 2012 Sep;27(9):1180-1187. [doi: 10.1007/s11606-012-2070-0] [Medline: 22549299]
70. Edupuganti S, Bushman J, Maditz R, Kaminoulu P, Halalau A. A quality improvement project to increase compliance with diabetes measures in an academic outpatient setting. Clin Diabetes Endocrinol 2019 Dec;5(1):15. [doi: 10.1186/s40842-019-0084-9]
71. Siopis G, Colagiuri S, Allman-Farinelli M. People with type 2 diabetes report dietitians, social support, and health literacy facilitate their dietary change. J Nutr Educ Behav 2021 Jan;53(1):43-53. [doi: 10.1016/j.jneb.2020.09.003] [Medline: 33077370]
72. Swarna Nantha Y, Kalasivan A, Ponnusamy Pillai M, et al. The validation of the Malay Yale Food Addiction Scale 2.0: factor structure, item analysis and model fit. Public Health Nutr 2020 Feb;23(3):402-409. [doi: 10.1017/S1368980019002684] [Medline: 31538554]
73. Al-Adwi ME, Al-Haswsa ZM, Alhmmadi KM, et al. Effects of different diets on glycemic control among patients with type 2 diabetes: a literature review. Nutr Health 2023 Jun;29(2):215-221. [doi: 10.1177/02601060221112805] [Medline: 35795964]
74. Paterson M, Bell KJ, O'Connell SM, Smart CE, Shafat A, King B. The role of dietary protein and fat in glycaemic control in type 1 diabetes: implications for intensive diabetes management. Curr Diab Rep 2015 Sep;15(9):61. [doi: 10.1007/s11892-015-0630-5] [Medline: 26202844]
75. Mertens E, Colizzi C, Peñalvo JL. Ultra-processed food consumption in adults across Europe. Eur J Nutr 2022 Apr;61(3):1521-1539. [doi: 10.1007/s00394-021-02733-7] [Medline: 34862518]
76. Sob C, Siegrist M, Hartmann C. The Positive Eating Scale: Associations with eating behavior, food choice, and body mass index. Eat Behav 2023 Jan;48:101706. [doi: 10.1016/j.eatbeh.2023.101706] [Medline: 36773373]
77. Bayram S, Kızıltan G, Akın O. Effect of adherence to carbohydrate counting on metabolic control in children and adolescents with type 1 diabetes mellitus. Ann Pediatr Endocrinol Metab 2020 Sep;25(3):156-162. [doi: 10.6065/apem.1938192.096] [Medline: 32871653]
78. Tascini G, Berioli MG, Cerquiglini L, et al. Carbohydrate counting in children and adolescents with type 1 diabetes. Nutrients 2018 Jan 22;10(1):109. [doi: 10.3390/nu10010109] [Medline: 29361766]
79. Uliana GC, Carvalhal M, Berino TN, et al. Adherence to carbohydrate counting improved diet quality of adults with type 1 diabetes mellitus during social distancing due to COVID-19. Int J Environ Res Public Health 2022 Aug 9;19(16):9776. [doi: 10.3390/ijerph19169776] [Medline: 36011412]
80. Farooqi MH, Abdelmannan DK, Al Buflasa MM, et al. The impact of telemonitoring on improving glycemic and metabolic control in previously lost-to-follow-up patients with type 2 diabetes mellitus: a single-center interventional study in the United Arab Emirates. Int J Clin Pract 2022;2022:6286574. [doi: 10.1155/2022/6286574] [Medline: 35685530]
81. Russell-Minda E, Jutai J, Speechley M, Bradley K, Chudyk A, Petrella R. Health technologies for monitoring and managing diabetes: a systematic review. J Diabetes Sci Technol 2009 Nov 1;3(6):1460-1471. [doi: 10.1177/193229680900300628] [Medline: 20144402]
82. Eberle C, Löhnert M, Stichling S. Effectiveness of disease-specific mhealth apps in patients with diabetes mellitus: scoping review. JMIR Mhealth Uhealth 2021 Feb 15;9(2):e23477. [doi: 10.2196/23477] [Medline: 33587045]
83. Stephen DA, Nordin A, Nilsson J, Persenius M. Using mHealth applications for self-care - an integrative review on perceptions among adults with type 1 diabetes. BMC Endocr Disord 2022 May 25;22(1):138. [doi: 10.1186/s12902-022-01039-x] [Medline: 35614419]
84. Moutou KE, England C, Gutteridge C, Toumpakari Z, McArdle PD, Papadaki A. Exploring dietitians' practice and views of giving advice on dietary patterns to patients with type 2 diabetes mellitus: a qualitative study. J Human Nutrition Diet 2022 Feb;35(1):179-190. [doi: 10.1111/jhn.12939]
85. Kahleova H, Znayenko-Miller T, Smith K, et al. Effect of a dietary intervention on insulin requirements and glycemic control in type 1 diabetes: a 12-week randomized clinical trial. Clin Diabetes 2024;42(3):419-427. [doi: 10.2337/cd23-0086] [Medline: 39015168]
86. Zhang S, Jiang H, Wang L, et al. Longitudinal relationship between body fat percentage and risk of type 2 diabetes in Chinese adults: evidence from the China Health and Nutrition Survey. Front Public Health 2022;10:1032130. [doi: 10.3389/fpubh.2022.1032130]
87. Stotz SA, Ricks KA, Eisenstat SA, Wexler DJ, Berkowitz SA. Opportunities for interventions that address socioeconomic barriers to type 2 diabetes management: patient perspectives. Sci Diabetes Self Manag Care 2021 Apr;47(2):153-163. [doi: 10.1177/0145721721996291] [Medline: 34078177]
88. Moore SE, McEvoy CT, Prior L, et al. Barriers to adopting a Mediterranean diet in Northern European adults at high risk of developing cardiovascular disease. J Human Nutrition Diet 2018 Aug;31(4):451-462. [doi: 10.1111/jhn.12523]
89. Monsivais P, Aggarwal A, Drewnowski A. Time spent on home food preparation and indicators of healthy eating. Am J Prev Med 2014 Dec;47(6):796-802. [doi: 10.1016/j.amepre.2014.07.033] [Medline: 25245799]

XSL•FO
RenderX

90. Vanstone M, Rewegan A, Brundisini F, Giacomini M, Kandasamy S, DeJean D. Diet modification challenges faced by marginalized and nonmarginalized adults with type 2 diabetes: a systematic review and qualitative meta-synthesis. Chronic Illn 2017 Sep;13(3):217-235. [doi: 10.1177/1742395316675024] [Medline: 27884930]
91. Adu MD, Malabu UH, Malau-Aduli AEO, Malau-Aduli BS. Enablers and barriers to effective diabetes self-management: a multi-national investigation. PLoS One 2019;14(6):e0217771. [doi: 10.1371/journal.pone.0217771] [Medline: 31166971]
92. Moström P, Ahlén E, Imberg H, Hansson PO, Lind M. Adherence of self-monitoring of blood glucose in persons with type 1 diabetes in Sweden. BMJ Open Diabetes Res Care 2017;5(1):e000342. [doi: 10.1136/bmjdrc-2016-000342] [Medline: 28611921]
93. Sousa C, Neves JS, Dias CC, Sampaio R. Adherence to glucose monitoring with intermittently scanned continuous glucose monitoring in patients with type 1 diabetes. Endocrine 2023 Mar;79(3):477-483. [doi: 10.1007/s12020-022-03288-1] [Medline: 36574148]
94. Martinez K, Frazer SF, Dempster M, Hamill A, Fleming H, McCorry NK. Psychological factors associated with diabetes self-management among adolescents with Type 1 diabetes: a systematic review. J Health Psychol 2018 Nov;23(13):1749-1765. [doi: 10.1177/1359105316669580] [Medline: 27663288]
95. Koenigsberg MR, Corliss J. Diabetes self-management: facilitating lifestyle change. Am Fam Physician 2017 Sep 15;96(6):362-370. [Medline: 28925635]
96. Vilafranca Cartagena M, Tort-Nasarre G, Rubinat Arnaldo E. Barriers and facilitators for physical activity in adults with type 2 diabetes mellitus: a scoping review. Int J Environ Res Public Health 2021 May 18;18(10):5359. [doi: 10.3390/ijerph18105359] [Medline: 34069859]
97. Ammann P, Offutt J. Introduction to Software Testing, 2nd edition: Cambridge University Press; 2016. [doi: 10.1017/9781316771273]
98. Luo J, Zhang K, Xu Y, Tao Y, Zhang Q. Effectiveness of wearable device-based intervention on glycemic control in patients with type 2 diabetes: a system review and meta-analysis. J Med Syst 2021 Dec 24;46(1):11. [doi: 10.1007/s10916-021-01797-6] [Medline: 34951684]
99. Toletti G, Boaretto A, Di Loreto C, Fornengo R, Gigante A, Perrone G. Enhancing diabetes therapy adherence: a comprehensive study on glucometer usability for type 2 diabetes patients. Front Clin Diabetes Healthc 2024;5:1328181. [doi: 10.3389/fcdhc.2024.1328181] [Medline: 38807703]

Abbreviations

DM: diabetes mellitus
DT: design thinking
HCD: human-centered design
SWOT: Strengths, Weaknesses, Opportunities, and Threats
T1DM: type 1 diabetes mellitus
T2DM: type 2 diabetes mellitus
UNESCO: United Nations Educational, Scientific and Cultural Organization

Edited by N Cahill; submitted 25.Apr.2025; peer-reviewed by C Hueso-Montoro, P Brauer; revised version received 07.Nov.2025; accepted 10.Nov.2025; published 21.Jan.2026.

Please cite as:
Palomares JM, Molina-Luque R, León-García F, Casares-Rodríguez I, García-Rodríguez M, Villena Esponera MP, Molina-Recio G SMARTCLOTH Prototype for Dietary Management in Patients With Diabetes Mellitus: Tutorial on Human-Centered Design Methodology for Health Care Hardware Development
J Med Internet Res 2026;28:e75744
URL: https://www.jmir.org/2026/1/e75744
doi:10.2196/75744

© Jose M Palomares, Rafael Molina-Luque, Fernando León-García, Irene Casares-Rodríguez, María García-Rodríguez, María Pilar Villena Esponera, Guillermo Molina-Recio. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Developing a Trauma-Informed Social Media Campaign to Disseminate Endometriosis-Specific Qualitative Art-Based Research Findings: Tutorial

Kerry Marshall[1,2*], RN, MN; Hargun Dhillon[3*]; A Fuchsia Howard[1,2], RN, PhD; Heather Noga[2], MA; Grace J Yang[3], BSc; William Zhu[3], BSc; Jessica Sutherland[4], BA; Sarah Lett[4]; Anna Leonova[2,3], MSc; Paul J Yong[2,3], MD, PhD; Natasha L Orr[1,2,3], PhD

[1]School of Nursing, University of British Columbia, Gateway Health Building, 5955 University Blvd., Vancouver, BC, Canada
[2]Women's Health Research Institute, BC Women's Hospital & Health Centre, Vancouver, BC, Canada
[3]Department of Obstetrics and Gynecology, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada
[4]Patient Research Advisory Board, Endometriosis and Pelvic Pain Lab, University of British Columbia, Vancouver, BC, Canada
[*]these authors contributed equally

Corresponding Author:
A Fuchsia Howard, RN, PhD
School of Nursing, University of British Columbia, Gateway Health Building, 5955 University Blvd., Vancouver, BC, Canada

Abstract

Trauma-informed approaches can promote the creation of systems that prioritize safety and empowerment to improve patient well-being. These approaches are especially important in sexual and reproductive health care, where patients are often asked to disclose sensitive and personal information. This disclosure is particularly relevant in the context of endometriosis, a condition that affects 10% of reproductive-aged women and causes debilitating pelvic pain. Our team led a trauma-informed social media campaign to raise awareness and improve the understanding of endometriosis by sharing research findings from a photovoice study focusing on Asian women's experiences of endometriosis during the COVID-19 pandemic in Canada (*EndoPhoto Study*). In this paper, we describe how we adapted and applied trauma-informed approaches to the development and implementation of the social media campaign. To do this, we followed five adapted trauma-informed principles: (1) support and collaboration, (2) trustworthiness and transparency, (3) safety, (4) empowerment and voice, and (5) cultural and gender sensitivity, and four steps: (1) frame the campaign, (2) create content and manage the campaign, (3) measure campaign impact, and (4) conduct postcampaign reflections. We co-designed this campaign with patient partners having lived experience of endometriosis to facilitate support and collaboration. Additionally, we shared details about the funders of this study to increase trust and transparency, moderated comments and deidentified images to promote participant safety, chose safer platforms to enhance empowerment and voice, avoided stereotypes, and shared authentic experiences of Asian women with endometriosis to support cultural and gender sensitivity. The campaign launched on Instagram and Pinterest in March 2025 to coincide with Endometriosis Awareness Month. The social media campaign received 8,540,528 total impressions over the course of the month and had engagement rates of 6.23% and 1.4% on Instagram and Pinterest, respectively.

(J Med Internet Res 2026;28:e83491) doi:10.2196/83491

KEYWORDS

trauma-informed approach; social media; knowledge translation; endometriosis; information dissemination; content creation

Background and Rationale

Overview

Endometriosis is a chronic inflammatory condition characterized by the presence of endometrial-like tissue outside the uterus [1]. The symptoms may vary, but they often include severe pelvic pain, painful periods, painful sexual intercourse, and infertility [2]. Despite affecting approximately 10% of reproductive-aged women and girls, and an unmeasured number of gender diverse people, endometriosis remains significantly underdiagnosed and misunderstood [1,3]. Although diagnostic delays average 5 years in Canada, some individuals have reported a formal diagnosis taking up to 20 years [3,4]. The invisibility of symptoms, stigma surrounding sexual and menstrual health, and dismissal of women's pain all contribute to misinformation and present barriers to timely diagnosis and treatment [3,5,6]. Ultimately, these aspects all affect the mental and physical health of those with endometriosis. Furthermore, racialized populations may experience additional barriers to endometriosis diagnosis and care [7]. For instance, one study found that East and/or Southeast Asian women were 8 times

XSL•FO
RenderX

more likely than their White counterparts to experience severe disease before being referred to more specialized care [8].

Globally, the COVID-19 pandemic further exacerbated the gaps in endometriosis care as it upended the health care system, causing resource redirection toward patients with COVID-19 and interrupting the continuity of care for patients with chronic conditions like endometriosis [9,10]. In Canada, appointments and surgeries for people with endometriosis were postponed or canceled as hospitals became overwhelmed and health care providers transitioned to virtual environments [11]. Concurrently, mandatory self-isolation measures dramatically altered people's levels of social support, contributing to worsening psychological symptoms such as depression and anxiety [12]. Additionally, the COVID-19 pandemic was marked by a global rise in anti-Asian sentiment, with people of Asian descent reporting increasing episodes of violence and feelings of vulnerability to discrimination [13].

Given these compounding factors, our team conducted a study—the *EndoPhoto Study*—with 22 South, East, and/or Southeast Asian cisgender women with endometriosis in Canada to better understand the experiences of people in these communities during the COVID-19 pandemic. This study used photovoice, an art-based methodology that provides opportunities to use photos to share experiences and emotions related to stigmatized or hidden conditions [14]. Results from the *EndoPhoto Study* are published elsewhere [15,16] and highlight several key themes. These themes include the ways in which the pandemic exacerbated feelings of isolation and created additional challenges in accessing health care for those living with endometriosis. Participants also built resilience during the pandemic by accepting social support from peers, advocating for themselves in health care interactions, and taking empowering actions to self-manage their conditions. More details on the methodology and data analysis of the *EndoPhoto Study* are available in other studies [15,16]. The *EndoPhoto Study* was approved by the University of British Columbia Children's and Women's Research Ethics Board (reference number: H22-02390).

Findings from the *EndoPhoto Study* and our team's previous research highlight the importance of sharing evidence that validates the experiences of people affected by endometriosis, helps people feel they are not alone, fosters hope, and recognizes the strengths of those affected. Our team's pre-existing website showcased *EndoPhoto* results via images and quotes (*EndoPhoto* website [17]). The original website was co-created by researchers, clinicians, and patient partners to disseminate information and resources related to endometriosis. As guided by our Patient Research Advisory Board (PRAB; a group of people with lived experience of endometriosis), we chose to disseminate the *EndoPhoto Study* findings and *EndoPhoto* website to a public audience through a social media campaign. The goal of the campaign was to amplify the stories shared by Asian women regarding their experiences during the COVID-19 pandemic while focusing on disrupting silence related to the medical dismissal, social isolation, and cultural stigma of pelvic pain and endometriosis. As such, we recognized the relevance of using a trauma-informed approach to develop and implement the campaign. See Figure 1 for a project overview and the campaign development process.

XSL•FO
RenderX

Figure 1. Project overview and campaign development process. PRAB: Patient Research Advisory Board.

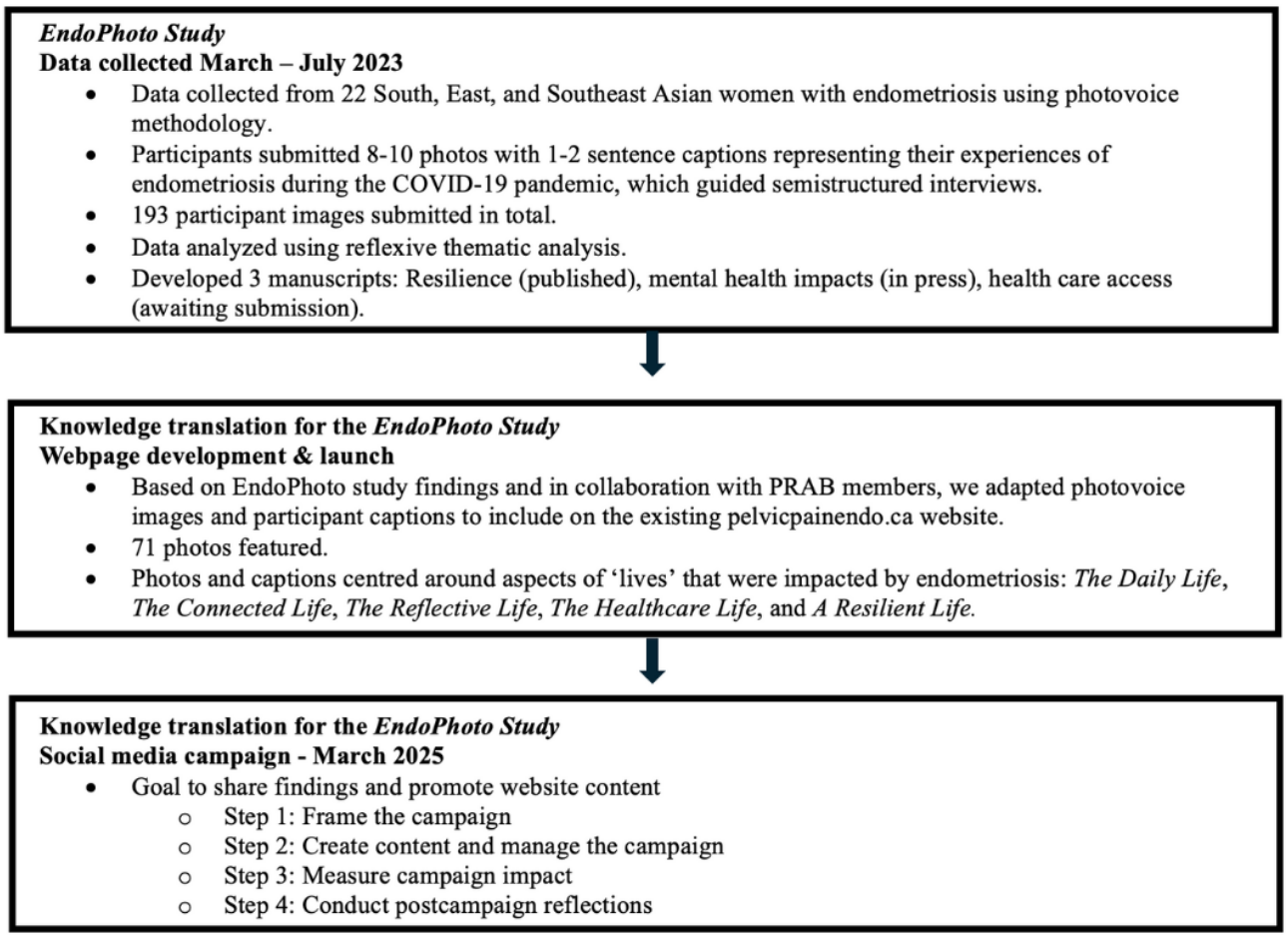

Objective

Our primary objective is to provide information on the steps we took when developing a social media campaign informed by the principles of a trauma-informed approach. Our secondary objective is to share the engagement results of the social media campaign. The target audience includes individuals and teams interested in trauma-informed social media campaigns, particularly those disseminating health-adjacent research findings. We begin by sharing information about our team, followed by information on trauma-informed approaches and social media dissemination. Lastly, we share the four steps that can be taken when developing a trauma-informed social media campaign: (1) frame the campaign, (2) create content and manage the campaign, (3) measure campaign impact, and (4) conduct postcampaign reflections.

Our Team

We are a multigenerational team with diverse genders, sexualities, and ethnicities, and are committed to improving the understanding and awareness of endometriosis through cutting-edge interdisciplinary research and knowledge translation. We recognize the importance of disseminating intentionally curated, evidence-based, and nuanced research findings to the endometriosis community and the public. Our team includes researchers, clinicians, health care trainees, and patient partners who are part of our PRAB. We are affiliated with the Endometriosis and Pelvic Pain Laboratory at the University of British Columbia, Canada.

What Are Trauma-Informed Approaches?

The formal conceptualization of trauma-informed care was first introduced by Harris and Fallot [18] in the context of mental health and substance use treatment systems. However, these principles have long-standing roots in community-based practices, including those within Indigenous traditions [19]. Since its inception, trauma-informed care has been adapted to various disciplines, with the framework of the Substance Abuse and Mental Health Services Administration (SAMHSA) often cited [20]. Trauma-informed approaches often acknowledge that trauma is widespread, and they actively support creating systems that promote physical and psychological safety [20].

SAMHSA defines trauma broadly, encompassing experiences at individual and structural levels that can be considered emotionally harmful [20]. SAMHSA's trauma-informed approach rests on four key assumptions: (1) realizing that trauma is widespread and can deeply affect individuals, communities, and societies; (2) recognizing the signs of trauma; (3) responding to trauma by integrating trauma-informed approaches; and (4) resisting retraumatization [20]. These assumptions are operationalized through six guiding principles: (1) safety; (2) trustworthiness and transparency; (3) peer support; (4) collaboration; (5) empowerment, voice, and choice; and (6) cultural, historical, and gender issues [20].

In health care settings, trauma-informed approaches in the provision of care have been shown to improve negative mental health symptoms and increase patient satisfaction, especially

XSL•FO
RenderX

among populations with histories of trauma or medical dismissal [21]. Trauma-informed approaches are particularly relevant in sexual and reproductive health, where patients are frequently asked to disclose sensitive information and are more likely to have experienced prior health care–related trauma [22]. Trauma-informed approaches are relevant for people with endometriosis as they have described feelings of shame and emotional distress related to their health care encounters where their symptoms have been diminished, normalized, or dismissed [23]. Interactions with health care systems and providers—as well as with broader public discourses that minimize people's experiences of endometriosis—have been further characterized as harmful, disempowering, and socially isolating. These experiences highlight the importance of using a trauma-informed approach that prioritizes safety, empowerment, and collaboration [23].

Social Media, Knowledge Dissemination, and Trauma-Informed Approaches

Social media is broadly defined as a digital space centered around information sharing and human connection [24]. Social media has increasingly become a pervasive aspect of everyday life and a powerful knowledge dissemination tool where various social media platforms, such as Facebook and TikTok, have been used by the health care community for patient education, peer support, and advocacy [25-27]. Social media platforms are participatory and easy to access, allowing the dissemination of information and rapid engagement of large, globally-connected audiences [28]. Content shared on social media platforms can provide unique opportunities to build a community, share experiences, and influence public health discourse [26].

There are also challenges with social media. For example, the nature of instant access to information and a lack of fact-verifying measures can lead to the unchecked and rapid spread of misinformation and disinformation to the public. Moreover, algorithms may incidentally lead to retraumatization and feelings of stigmatization [29-31]. Furthermore, considering people often consume social media content in isolation, it is difficult for content creators to recognize potential traumatization or retraumatization of viewers, highlighting the importance of intentionally creating and sharing content.

As knowledge dissemination of health-adjacent information increasingly moves onto social media platforms, trauma-informed approaches appear particularly relevant and potentially useful when working in these virtual environments. While there is limited guidance for applying trauma-informed approaches in digital spaces [32], the literature is emerging. We drew upon 3 frameworks that highlighted the potential of these approaches to reduce harm when sharing health-adjacent information digitally.

First, Josephs et al [33] emphasized 3 key pillars for digital trauma-informed design specific to sexual and reproductive health: privacy and confidentiality, intuitive and representative designs, and inclusive language. Second, *trauma-informed computing*, introduced by Chen et al [34], presents a framework guiding the adaptation of trauma-informed principles to digital design. This framework recognizes that digital tools can cause or exacerbate trauma and seeks to enable safer technological experiences [34]. Key adaptations of trauma-informed principles for online settings include safety, trustworthiness, peer support and collaboration, empowerment and choice, and cultural sensitivity.

Lastly, Scott et al [30] built upon the framework from Chen et al [34], adding specific aspects and examples to consider when applying trauma-informed approaches to social media engagement. The framework from Scott et al [30] outlines six guiding principles: (1) safety (eg, safe data collection and storage, and relaxing colors); (2) trustworthiness and transparency (eg, transparent about what user data are collected and why); (3) peer support (eg, protection for those sharing their unique stories); (4) collaboration and mutuality (eg, co-design with people having lived experience); (5) empowerment, voice, and choice (eg, no real names); and (6) cultural and historical gender issues (eg, acknowledge algorithmic biases).

Developing Our Trauma-Informed Social Media Campaign

Step 1: Frame the Campaign

Determine Relevant Trauma-Informed Principles

Based on the previously mentioned frameworks, we adapted our campaign to focus on the following principles of a trauma-informed approach: (1) support and collaboration, (2) trustworthiness and transparency, (3) safety, (4) empowerment and voice, and (5) cultural and gender sensitivity. These principles were used throughout the campaign development process. Table 1 provides an overview of how we approached incorporating trauma-informed principles throughout the campaign.

XSL•FO
RenderX

Table . Overview of our approach.

Trauma-informed guiding principle	Our approach and examples	Considerations

Support and collaboration	• Step 1: Frame the campaign
	  • Co-designed with PRAB[a] members and a diverse interdisciplinary team of experts
	  • Incorporated feedback throughout the process
	  • Shared content from lived experiences
	  • Engaged with influencers we had a previous relationship with to promote content
	• Step 4: Conduct postcampaign reflections
	  • Reflected as a team on successes, challenges, and learnings for future social media engagement
		• More time may be needed to include the feedback and ideas of all team members, and thus, there may be a longer timeline to project completion

Trustworthiness and transparency	• Step 1: Frame the campaign
	  • Informed participants of the purpose/content when consenting to the original study
	  • Obtained explicit and ongoing consent related to the use of data
	• Step 2: Create content and manage the campaign
	  • Created content highlighting our research team's positionality statement
	  • Created content highlighting transparency around funding and what this meant
	• Step 3: Measure campaign impact
	  • Intentionally collected metrics that were not overly invasive
	  • Intentionally collected metrics that were inclusive of multiple ways of engaging with content, recognizing that people may engage with content differently
		• Viewers may have personal negative feelings about funders
		• Additional human resources are needed to gather confirmatory consent from participants
		• Opt-out versus opt-in could incidentally include photos that participants did not want to share, but they did not see the email

Safety	• Step 1: Frame the campaign
	  • Engaged with specific platforms (Instagram and Pinterest)
	• Step 2: Create content and manage the campaign
	  • Created content using gentle, muted color palettes in content creation (eg, using light yellow)
	  • Created content in grouped images in collage format and shared select images
	  • Created content with deidentified images by blurring faces or including images where people were masked
	  • Moderated comments
	  • Included content warnings regarding sensitive topics
		• Participants may provide images that violate design principles and gentle colors
		• Additional human resources are required for moderating comments
		• Not including all experiences captured, as many images were not selected
		• Not as wide a reach due to including only select friendly platforms
		• Temporary stories reduce reach
		• Deidentifying images or using the photos without the original captions may alter the intended goal and impact of the image, limiting participant creativity

XSL•FO
RenderX

Trauma-informed guiding principle	Our approach and examples	Considerations

Empowerment and voice	• Step 1: Frame the campaign
	  • Leveraged platforms to amplify voices
	• Step 2: Create content and manage the campaign
	  • Created content that used friendly, nonstigmatizing, and everyday language when captioning photos
	  • Made changes to the content based on algorithm limitations while staying grounded in authentic patient voice and experiences
		• Potential for less reach when only using select platforms, and the reach to certain groups (eg, older people who more often use Facebook) might be reduced; demographics can vary across platforms
		• Participant caption is not always included with a photo, which could change the intended meaning
		• Focus on "positive" aspects due to the algorithm could represent a one-sided or skewed representation of experiences
		• Sharing content that is "trauma-informed" may incidentally portray the information and experiences as being neutral or positive

Cultural and gender sensitivity	• Step 2: Create content and manage the campaign
	  • Created content that ensured sharing only "positive" language and experiences due to the Meta algorithm
	  • Created content that avoided stereotypes and stigmatization
	  • Created content that avoided hypergendered content
		• Excluding images that depict too much pain or have identifying information may reduce the transparency of people's experiences
		• Some people might feel that the nonhypergendered colors and content do not relate to them as much

[a]PRAB: Patient Research Advisory Board.

Identify a Theoretical Approach

Theoretical approaches provide the foundation for a research study, offering structure and guidance when developing objectives, methods, and analysis [35]. Throughout the development of the campaign, including content creation, data interpretation, and dissemination strategies, we were primarily guided by the principles of intersectional feminism and integrated knowledge translation (IKT). An intersectional feminist perspective informed our understanding of how overlapping identities, such as race, gender, and country of origin, shape individuals' health care experiences. IKT is a collaborative process that emphasizes partnership between researchers and knowledge users throughout all stages of research [36]. IKT informed how we identified priorities, designed methods, interpreted data, and shared results [36]. Unlike traditional models that position researchers as the primary producers of knowledge, IKT recognizes the expertise of both researchers and community partners, aiming to minimize power differentials and promote equitable, contextually relevant knowledge creation [37].

Identify Campaign Goals and Messaging

Our main goal for the social media campaign was to share key research findings from the *EndoPhoto Study*, increase awareness of Asian women's experiences living with endometriosis, and direct viewers to our newly developed interactive *EndoPhoto* website [17]. In conducting the campaign, a second goal was to foster a sense of validation, emotional safety, and support for both previous research participants and audiences, and minimize their risk of retraumatization.

The campaign's central message—*your pain is real, you are believed, and you are not alone*—matched the overall messaging of content produced by our team and shared on the website. Additionally, this message was consistently emphasized across all digital platforms. The team deliberately declined to create a new hashtag for the campaign, given the objective of using social media as a mechanism to reach a wide audience of users rather than share isolated content with limited reach. As such, all the posts incorporated the widely recognized community hashtag #ThisIsEndo, supplemented by topic-relevant hashtags.

Determine When and Where to Launch the Campaign

We launched the campaign in March 2025 to coincide with Endometriosis Awareness Month. The Endometriosis and Pelvic Pain Laboratory had a previously established Instagram profile (@pelvicpainendo) with approximately 1000 followers before the campaign and had a Pinterest account (@pelvicpainendo), which was created for this and future campaigns. These platforms aligned with the campaign's visual and trauma-informed goals, offering features such as content warnings and comment moderation. The content shared was similar for both platforms but adapted in format to best use each platform's features. For example, Instagram's reels, stories, and carousels supported a balance of educational content and personal narratives, while Pinterest enabled thematic curation through boards and infographics. Pinterest has a different set of users and is more aligned with artistic communities.

Considering that Pinterest policies restrict the use of paid advertisements for new accounts, a soft launch on the platform began on February 19, 2025. Because of this, we published several posts prior to the full campaign, which allowed us to generate early interest, establish baseline engagement, and be considered an established account.

XSL•FO
RenderX

Step 2: Create Content and Manage the Campaign

Overview

The content that was shared during the social media campaign was initially designed by HD (quote posts), GJY (image posts), and WZ (videos/reels), who drew upon findings from the *EndoPhoto Study* and website content under the guidance of the announcement post creator and content lead (HN). Before final approval, all content was reviewed and discussed with the broader research team during team meetings every 2 weeks and PRAB meetings every 2 months. In total, the campaign featured 41 posts between March 1 and March 31, 2025, across Instagram and Pinterest. In this paper, we chose to share the images of content only from Instagram owing to the easy viewability of the entire post. In the following section, we discuss how we enacted these aspects in content creation and when managing the campaign. It is important to note that although these aspects are presented as being separate ideas, many have overlapping and related actions.

Incorporating Support and Collaboration

We defined support and collaboration as intentional actions taken to meaningfully engage with those having lived experience and others in the community doing similar work. As such, PRAB members defined campaign goals; reviewed and co-developed content; and ensured that materials were safe, empowering, and contextually relevant. The PRAB involvement created an opportunity for safety by foregrounding lived experience and ensuring that social media content reflected their values and perspectives. The campaign was also managed by an external communications agency, where the partner and social media specialist (SL) was someone with lived experience of endometriosis who simultaneously acted as a member of the PRAB.

We also took a collaborative approach to promoting the campaign, leveraging our known and existing networks. This involved approaching familiar social media accounts, including science communicators, nonprofit organizations, news outlets, and independent influencers, and asking them to share and promote our content. The campaign also partnered with advocacy organizations, including The Endometriosis Network Canada, to broaden outreach and ensure alignment with existing efforts in the endometriosis advocacy landscape.

Incorporating Trustworthiness and Transparency

When creating content, we considered trustworthiness and transparency aspects that involved disclosing who is behind the campaign, our goals, the funders, and where the content came from through “announcement” posts (Figure 2). Prior to the campaign, we were transparent with the participants of the *EndoPhoto Study* about how we were using their data.

Figure 2. Example of an announcement-style post on Instagram.

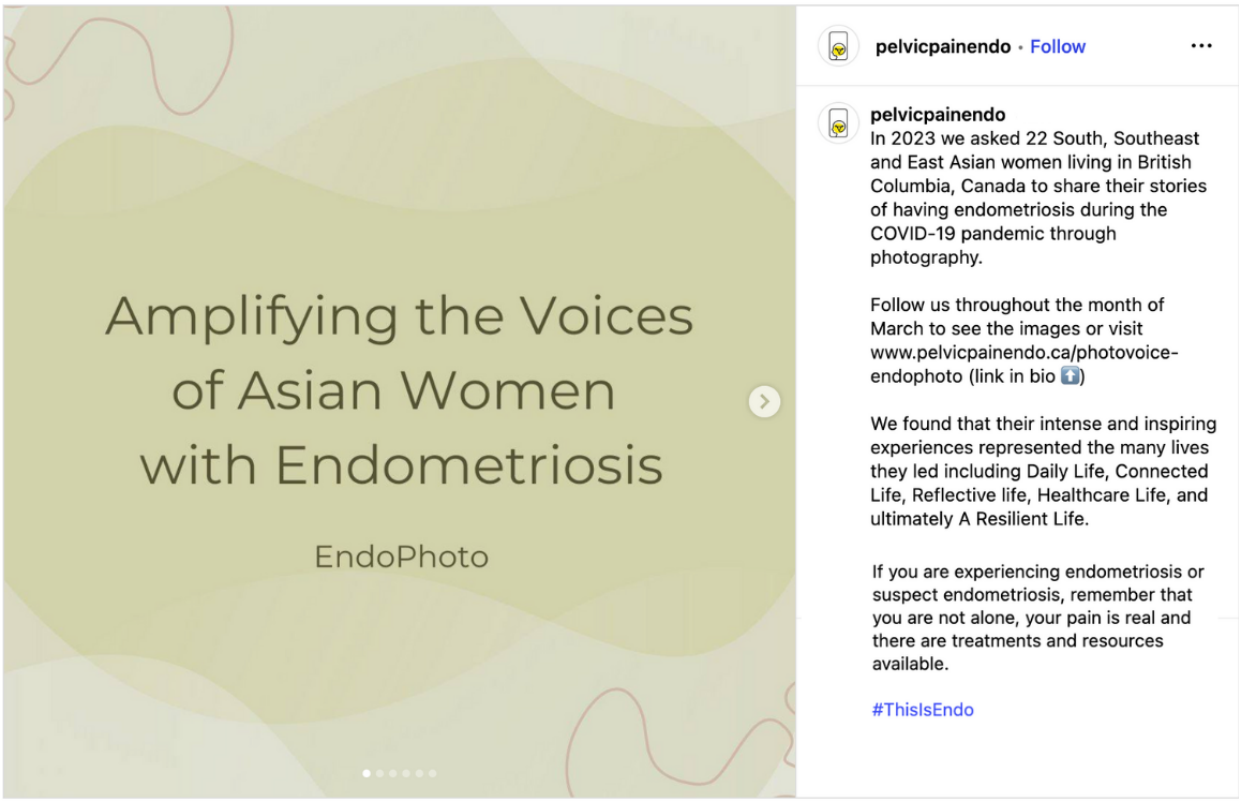

All participants in the *EndoPhoto Study* provided explicit informed consent to have their photos used in a social media campaign. We also provided participants with a lay summary of the results via email that included an “action required” message, showed participants the website [17] where photos had been included in a virtual gallery, and gave participants the ability to withdraw their photos and quotes at any point. We also strongly encouraged *EndoPhoto Study* participants to review their photos to ensure they were comfortable with these being shared. No participants opted out of their photos being shared.

XSL•FO
RenderX

Incorporating Safety

When considering safety, we focused on the principles of privacy and confidentiality from Josephs et al [33], and prioritized safety and preventing retraumatization of participants of the *EndoPhoto Study* whose images we shared. First, when creating content, to maintain the emotional safety of those who participated in the research and shared their photos, we intentionally curated visual content in a way that still honored participants' lived experiences. We did this by choosing images that were not intimately personal or overtly medical in nature, or that did not depict individual people in moments of visible distress. Instead, the campaign showcased strength-based visuals, such as nature scenes, symbolic objects, and comforting moments like participants' pets offering support (Figures 3 and 4).

Figure 3. Example of a collage of nature and social support.

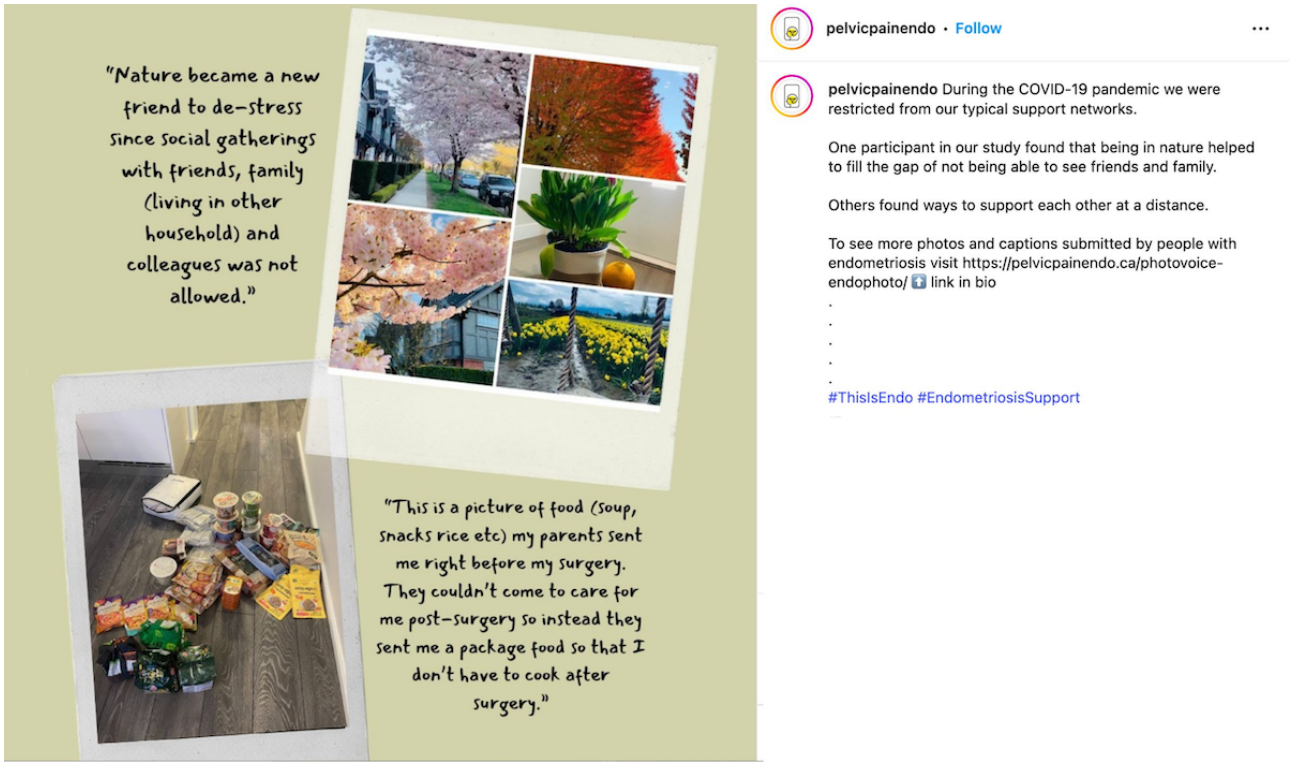

Figure 4. Example of a post about pets providing comfort.

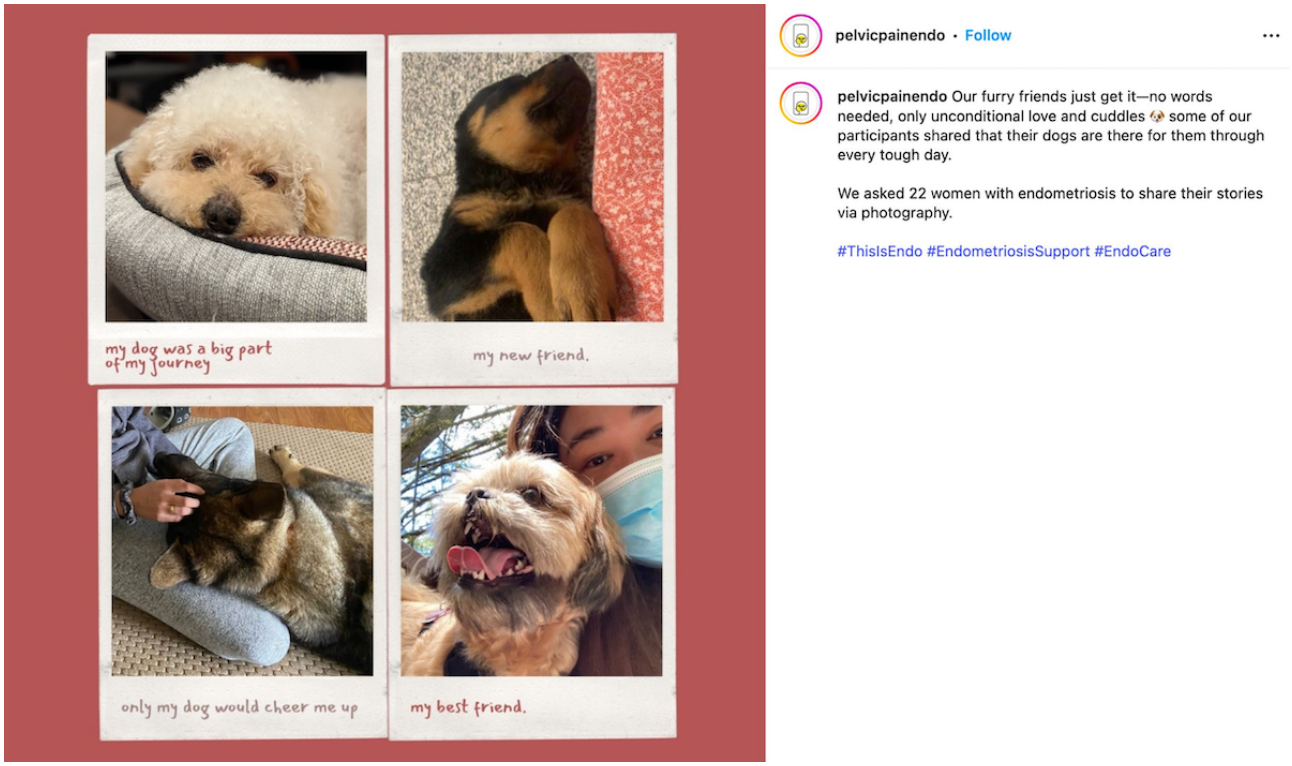

XSL•FO
RenderX

When possible, in content creation, images were paired with participant-authored captions that emphasized resilience, healing, and other personally meaningful themes to balance narrative authenticity with emotional safety. We also intentionally blurred faces to protect identities or selected images that further supported anonymity, such as those in which individuals used a surgical mask (Figures 4 and 5).

Figure 5. Example of identity protection on Instagram.

Second, when creating content, careful consideration was given to how participant photos were shared. We opted to present images as collages or grouped images rather than posting them individually (Figure 6). This approach was chosen to minimize the risk of certain photos receiving disproportionately more “likes” or “shares” than others, which could cause distress among some participants if they noticed their photos were less “liked.”

XSL•FO
RenderX

Figure 6. Example of a shared collage of participant photos from an Instagram reel.

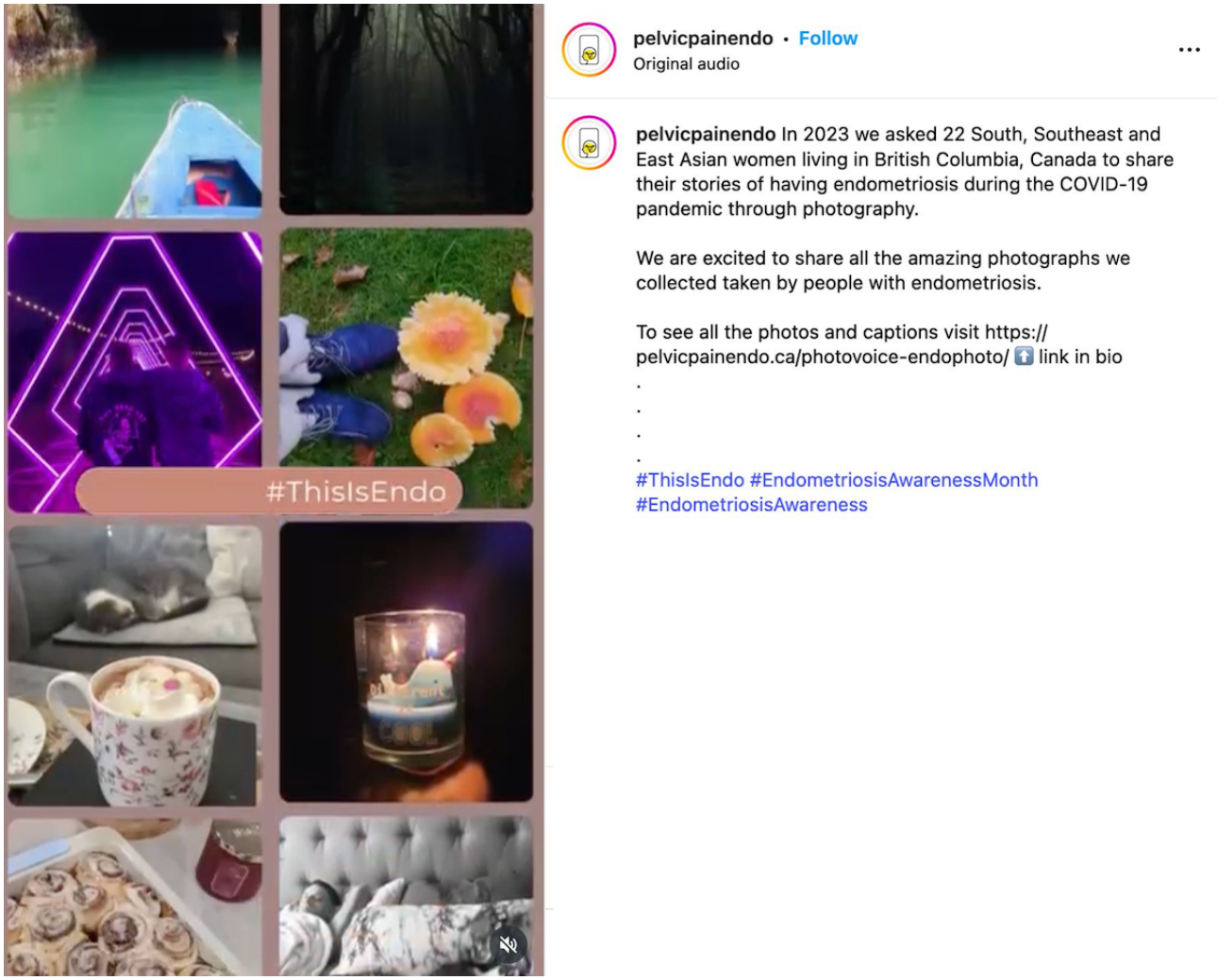

Third, based on recommendations from our social media specialist and patient partners, Instagram and Pinterest were specifically chosen as platforms for this campaign, given that they are well-suited to image-based storytelling and, anecdotally, were considered less volatile during the dates of the campaign, with lower rates of reproductive health online harassment compared with other well-known platforms.

Fourth, elements of safety were further considered through active moderation of comments to identify and remove hate speech, trolling, and unsolicited medical advice; however, we did not find that these were issues in this campaign.

Incorporating Empowerment and Voice

We considered empowerment and voice to focus on ensuring that we were truthfully representing participant experiences of the *EndoPhoto Study*, opting to create content that was more strength-based and showcased resilience and empowerment while also sharing the reality of experiencing endometriosis (Figures 7 and 8). In order to accomplish this, we created content that used nonstigmatizing, everyday language and often incorporated participants' own words in explaining the context of the photos (Figure 9).

XSL•FO
RenderX

Figure 7. Example of quote-based content on strength while living with endometriosis.

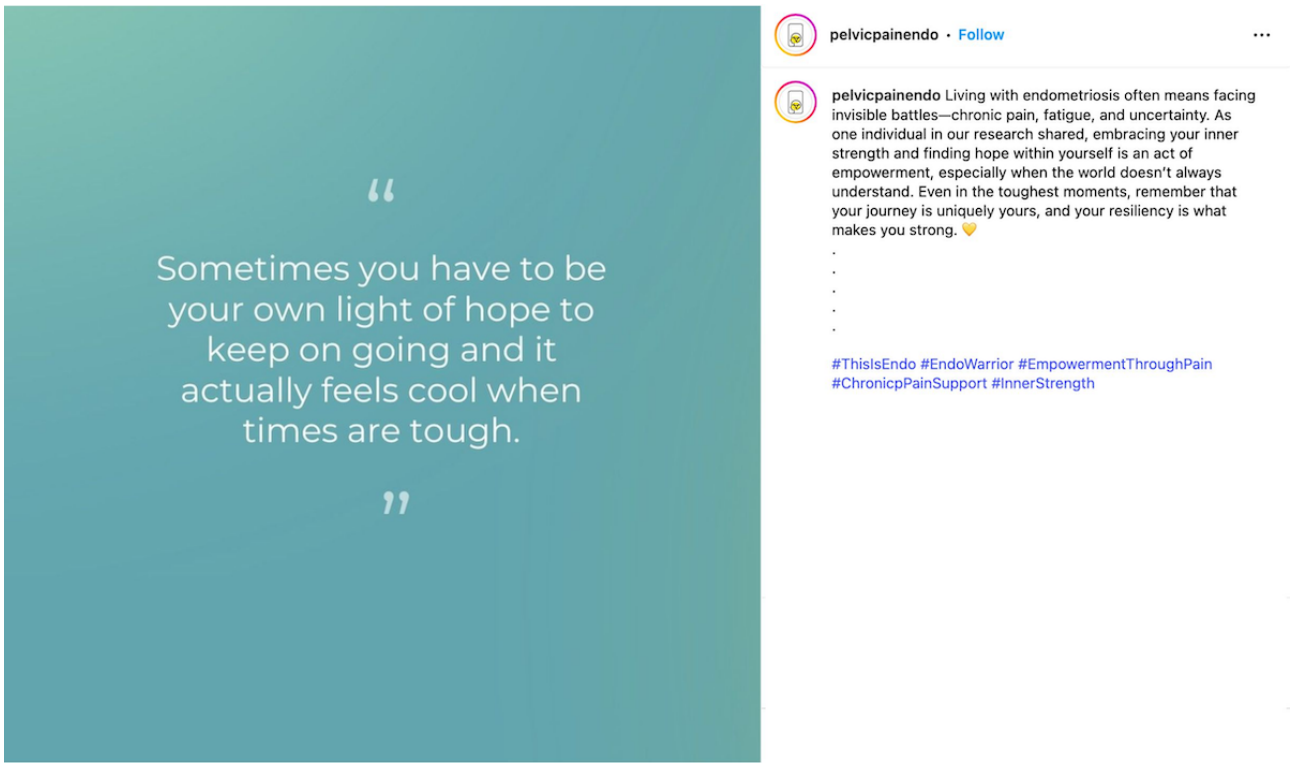

Figure 8. Example of quote-based content on living with endometriosis.

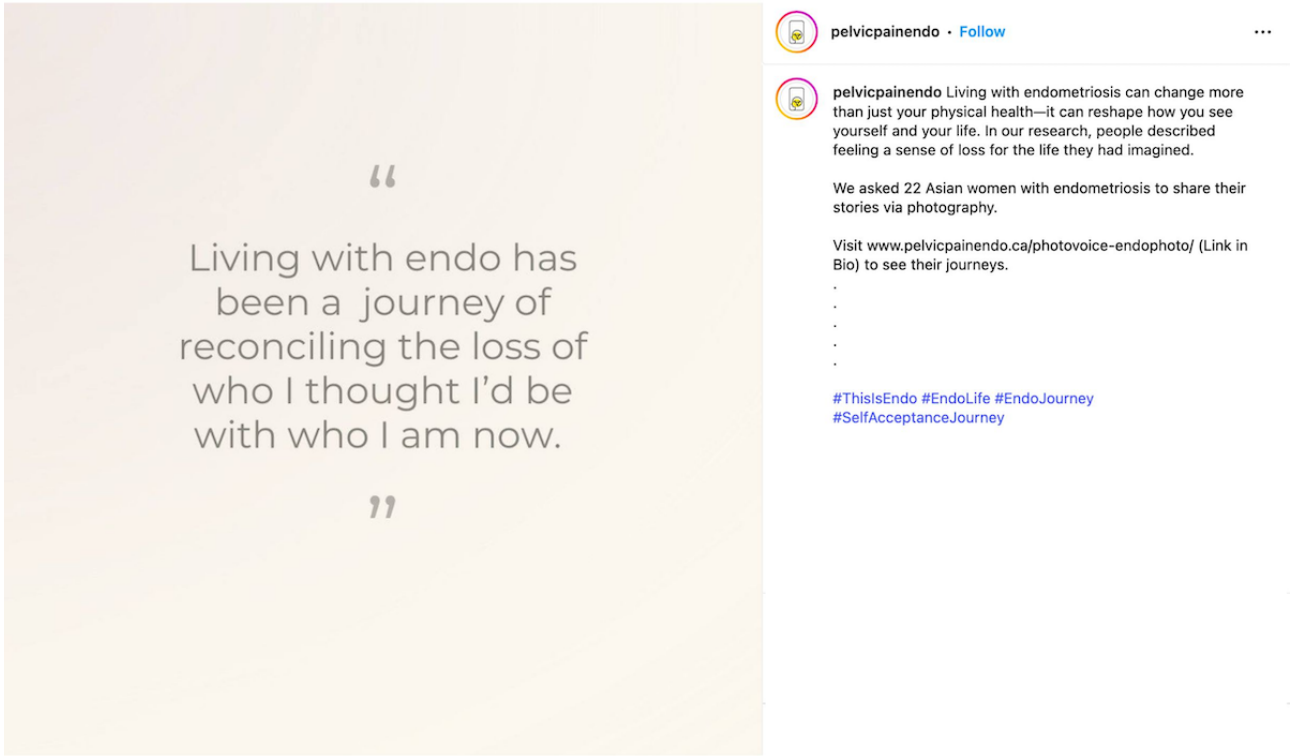

XSL•FO
RenderX

Figure 9. Example of everyday language use in a caption.

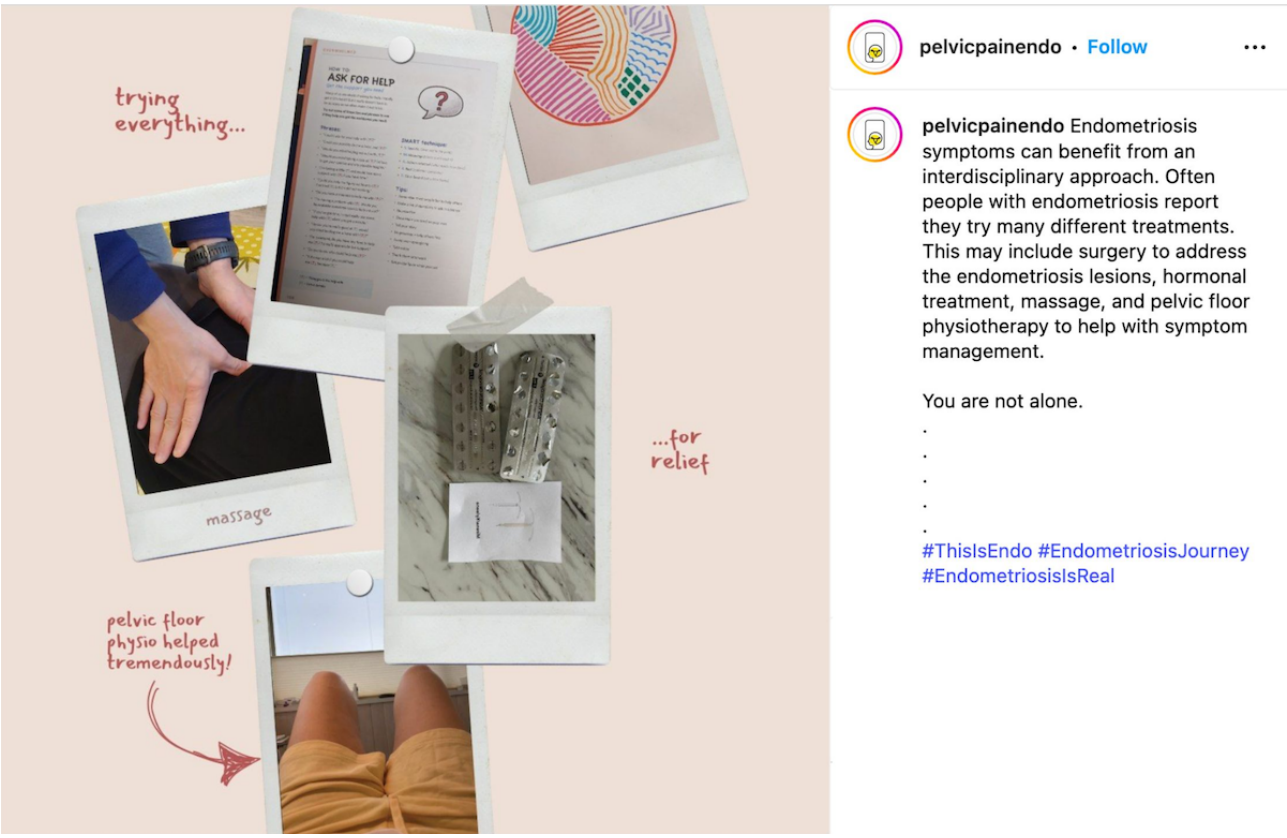

By integrating visual narratives with participant-authored captions, the campaign created opportunities for individuals to reclaim agency in narrating their health care experiences, particularly where medical and workplace systems had previously been invalidating. For example, 1 post focused on how cultural taboos surrounding menstruation can lead to a lack of communication and discussion of pain. This caption drew attention to the compounding effects of stress, isolation, and reduced access to care (Figure 10). Together, we intended these posts to help humanize the lived realities of people with endometriosis while fostering empathy, reducing stigma, and encouraging public dialogue.

XSL•FO
RenderX

Figure 10. Example of content highlighting cultural taboos around menstruation.

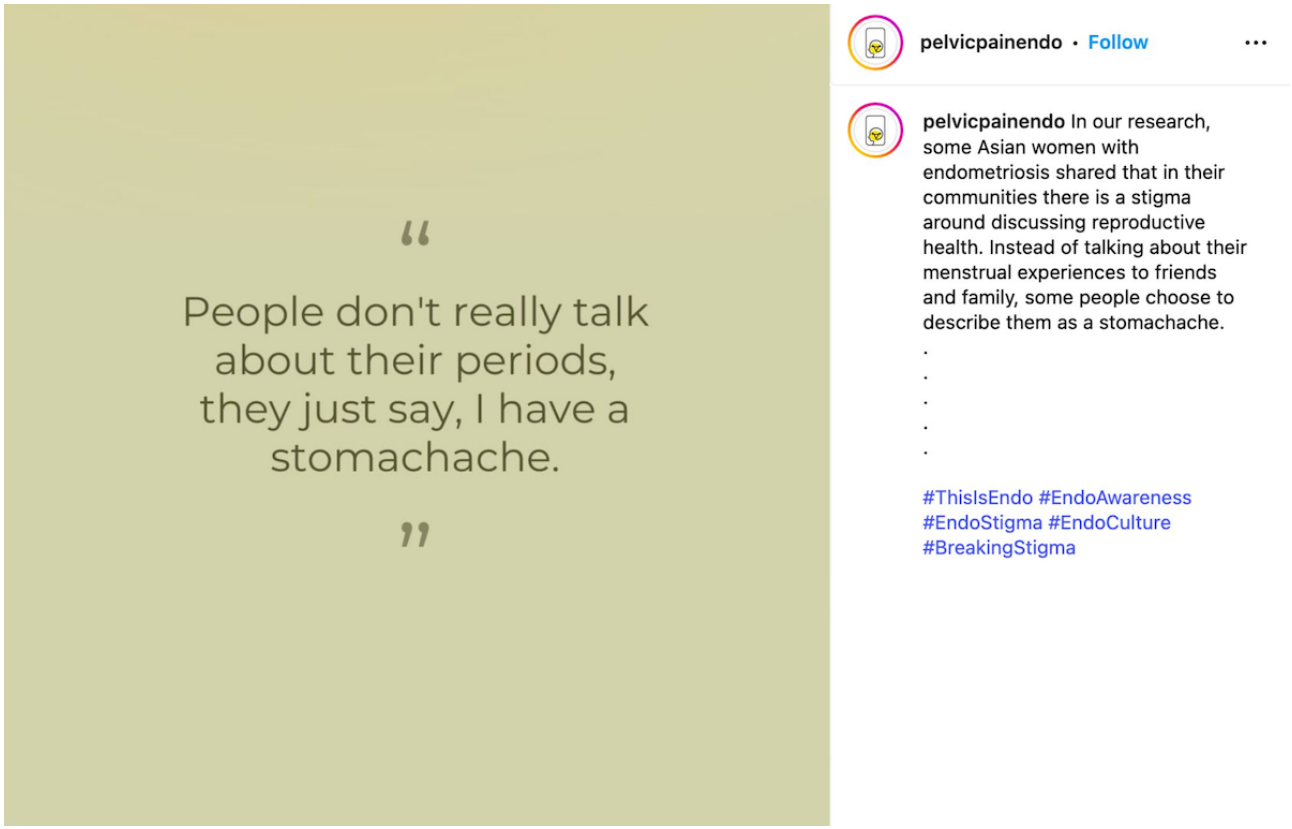

Incorporating Cultural and Gender Sensitivity

Striving for cultural and gender sensitivity, we carefully curated content to avoid sensationalism, incidental stigmatization, stereotypes, clinical or diagnostic language, and potentially distressing imagery. The campaign content aimed to disrupt the silence surrounding pelvic pain and endometriosis, particularly the effects of medical dismissal, social isolation, and cultural stigma. Considering the gendered nature of endometriosis—and although all the participants whose photos we shared identified as cisgender women—we intentionally avoided making the content hyperfeminized or gendered toward women exclusively. We also aimed to avoid perpetuating stereotypes and hyperfeminized content by choosing a color palette that was intentionally calming while not overly gendered (Figure 11).

XSL•FO
RenderX

Figure 11. Example of the avoidance of hypergendered or stigmatizing language or colors.

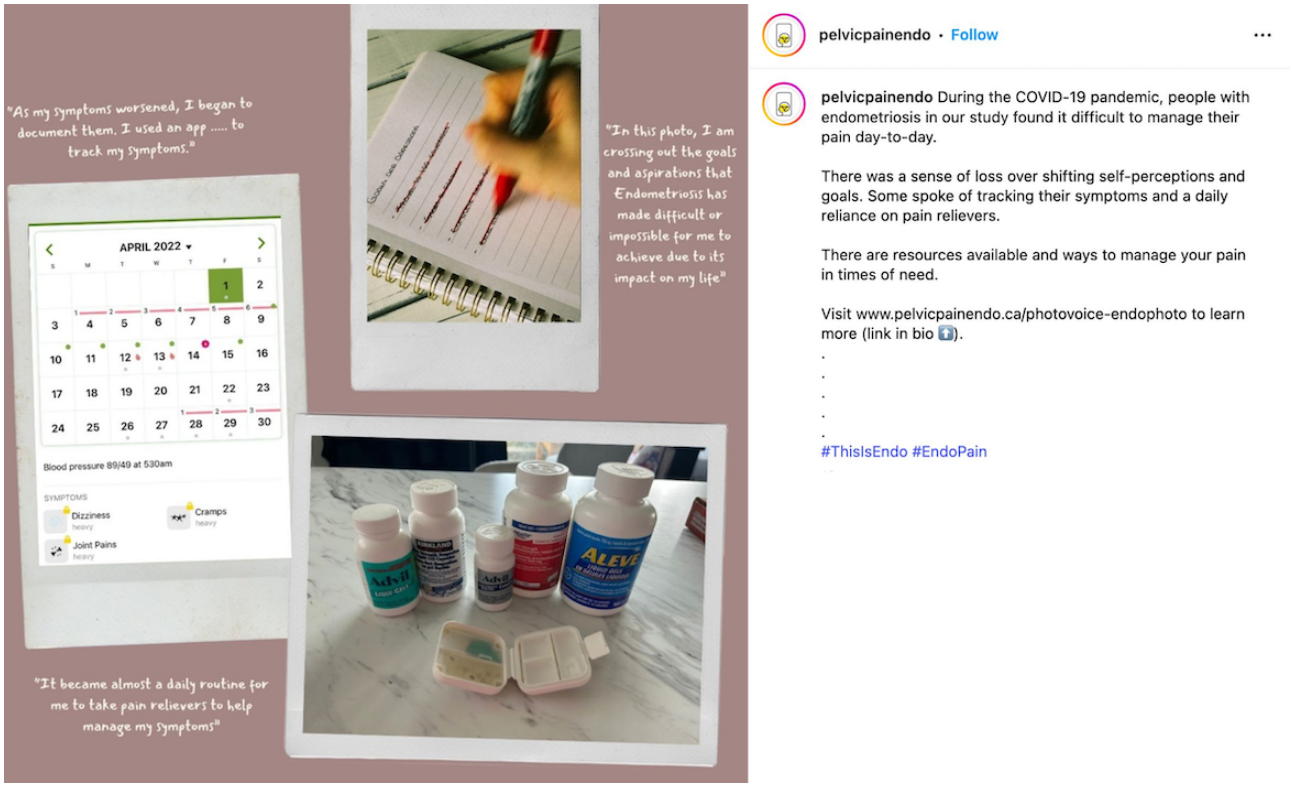

Step 3: Measure Campaign Impact

We used platform-integrated analytics (Instagram Insights and Pinterest Analytics) to monitor primary performance indicators. The metrics included reach, engagement (likes, shares, comments, profile visits, and link clicks), and website page visits (for definitions, see Table 2). We reviewed these metrics throughout the campaign to enable real-time optimization of posting frequency and timing, as well as advertising spend based on platform recommendations and observed audience behavior. Over the 31-day campaign, the website attracted 6326 unique users (for additional engagement measures, see Table 3).

Table . Consolidated definitions of terms used to describe campaign engagement.

Term	Definition
Reach	Reach is defined as the number of unique users who have seen the online content at least once
Engagement	Likes, shares, comments, profile visits, and link clicks
Impressions	The total number of times the content is presented to potential users on a screen
Volume of sessions	Number of visits to the website
Dwell time	The duration of time people view the content
Click-through rate	The percentage of people who click on a link to the website within the content

XSL•FO
RenderX

Table . Engagement results from Instagram and Pinterest.

Platform and metric	Value
Overall	
Total social impressions, n	8,540,528
Instagram	
Engagement[a], %	6.23
Advertisement impressions, n	7,941,457
Advertisement reach, n	3,550,309
Pinterest	
Total impressions, n	581,081
Total engagement, n	5528

[a]With regard to engagement, on Instagram, anything above 6% is considered high engagement, and on Pinterest, anything between 1% and 2% is considered average engagement [38,39].

Instagram generated both the greatest volume of sessions and the longest dwell time. Pinterest's shorter dwell time likely reflected the account's infancy (new profile, first campaign, and algorithmic learning period). The high number of impressions was largely attributed to 3 posts that "went viral" during the campaign, meaning they received over 1 million views each (Figures 12-14). These posts reflected messages of support, resiliency, and healing. Advertisements drew in the largest number of platform users to the page, accounting for most of the sessions and engagement, with arguably the least human resources.

Figure 12. Still image from a reel that "went viral" during the campaign.

XSL•FO
RenderX

Figure 13. A post that "went viral" during the campaign.

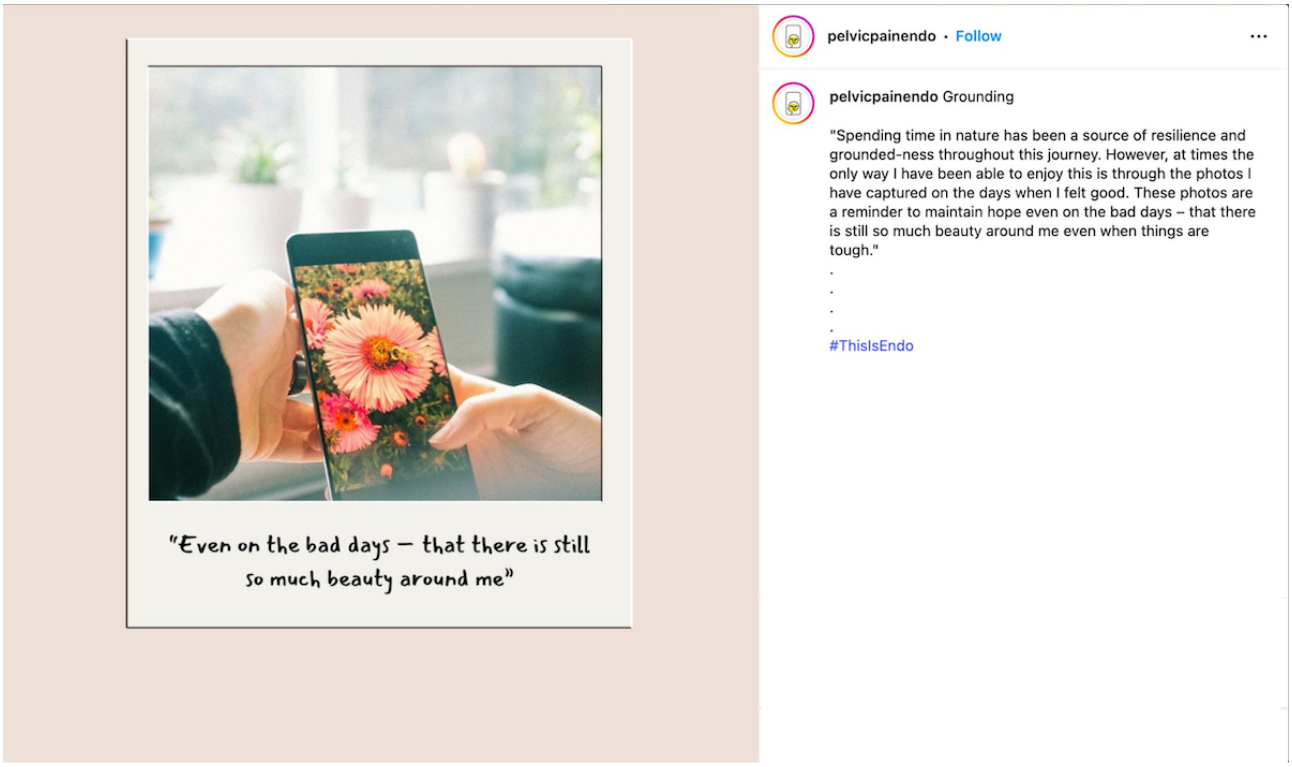

Figure 14. Still image taken from the second reel that "went viral" during the campaign.

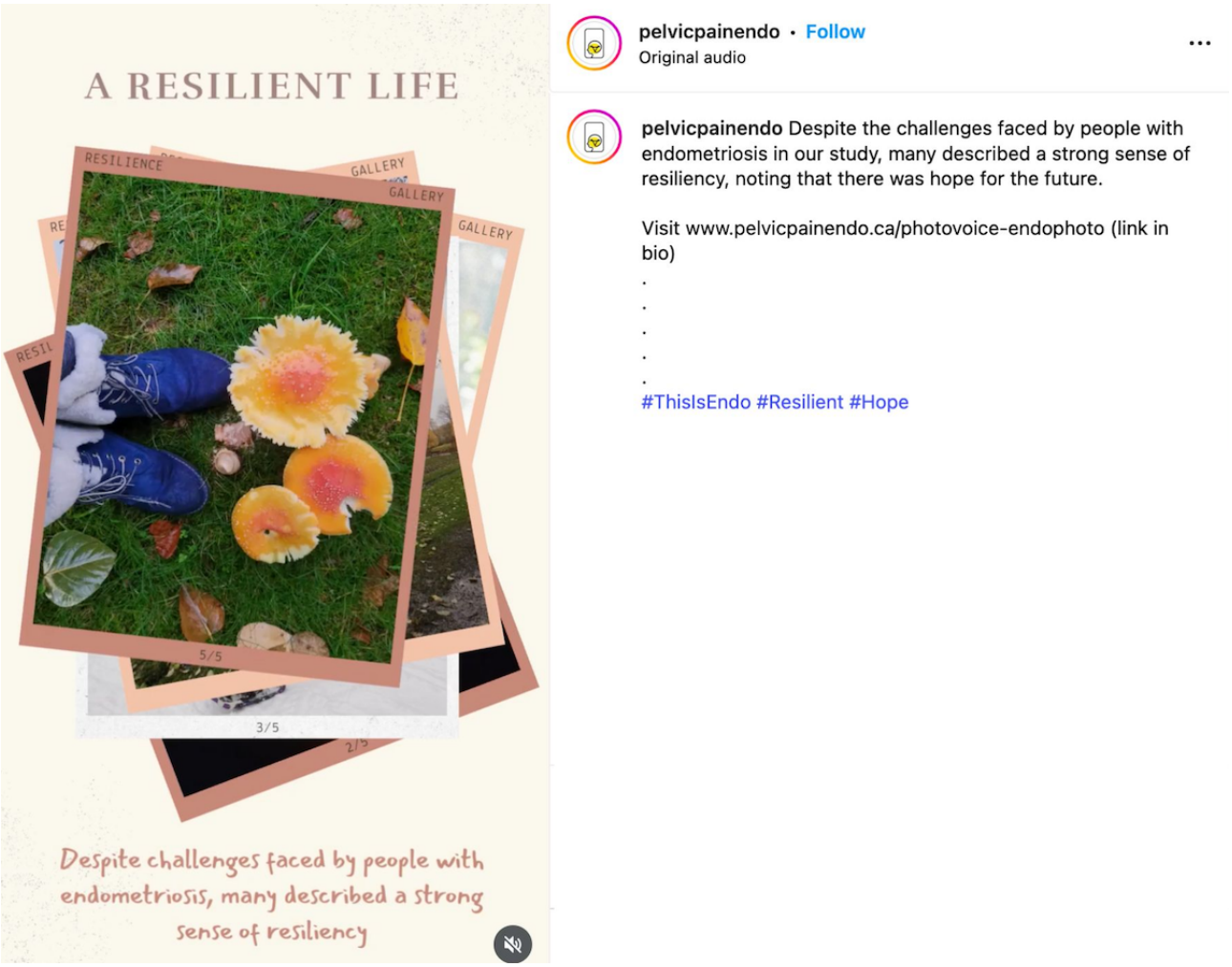

XSL•FO
RenderX

Step 4: Conduct Postcampaign Reflections

Identify What Went Well

To increase the reach of the social media campaign and the number of people seeing the online content at least once, we paid for advertisements on both platforms, with a total budget of CAD $3000 (US $2160). We found that advertisements required few human resources and were a relatively inexpensive way to increase reach. Posts that were advertised resulted in greater reach than posts that were not advertised. For our paid advertisements, a detailed audience profile was developed to guide content creation and advertisement targeting. The intended audience included individuals who either had a confirmed or suspected endometriosis diagnosis, were assigned female at birth, were of reproductive age (inclusive of all gender identities, sexual orientations, and relationship statuses), had moderate levels of health literacy, and understood English. We did not explicitly target Asian women and people living with endometriosis, as race and disease data are not captured for users by social media platforms explicitly. High-frequency search terms related to pelvic pain, endometriosis, and Asian experiences were identified to optimize advertisement discoverability (eg, pain, symptoms, self-care, journey, resilience, and reflection).

One of the campaign's primary strengths was its collaborative, interdisciplinary approach. The involvement of patient partners through the PRAB ensured content authenticity and emotional safety. In addition, working with a social media specialist was invaluable for navigating the landscape of digital media. The campaign also benefited from sufficient funding, allowing strategic investment in high-performing advertisement formats and continuous optimization based on analytics.

Our team's expertise, including content development led by those with lived, clinical, and research experience, added legitimacy, credibility, and multiple perspectives that countered misinformation in digital health spaces. Importantly, the campaign filled a representational void by centering the narratives of South, East, and Southeast Asian women with endometriosis, a demographic historically underrepresented in both research and advocacy.

Identify Challenges

Early in the campaign (March 3, 2025), our Instagram account was flagged by the Meta algorithm and included in a category called "Health and Wellness" owing to a post that was identified as being "negative" (Figure 15). This category was designed by Meta to reduce "negative" advertising and monetization of organizations that use advertising to sell products and services. Although we were not advertising products, our placement in this category limited what content could be promoted. Posts that included any features flagged as "negative" by the algorithm were prohibited from being advertised. Considering that endometriosis experiences often involve challenges, our preplanned content required significant changes in order to meet the algorithm's criteria.

Figure 15. An original post on Instagram that was flagged by the algorithm as being "negative".

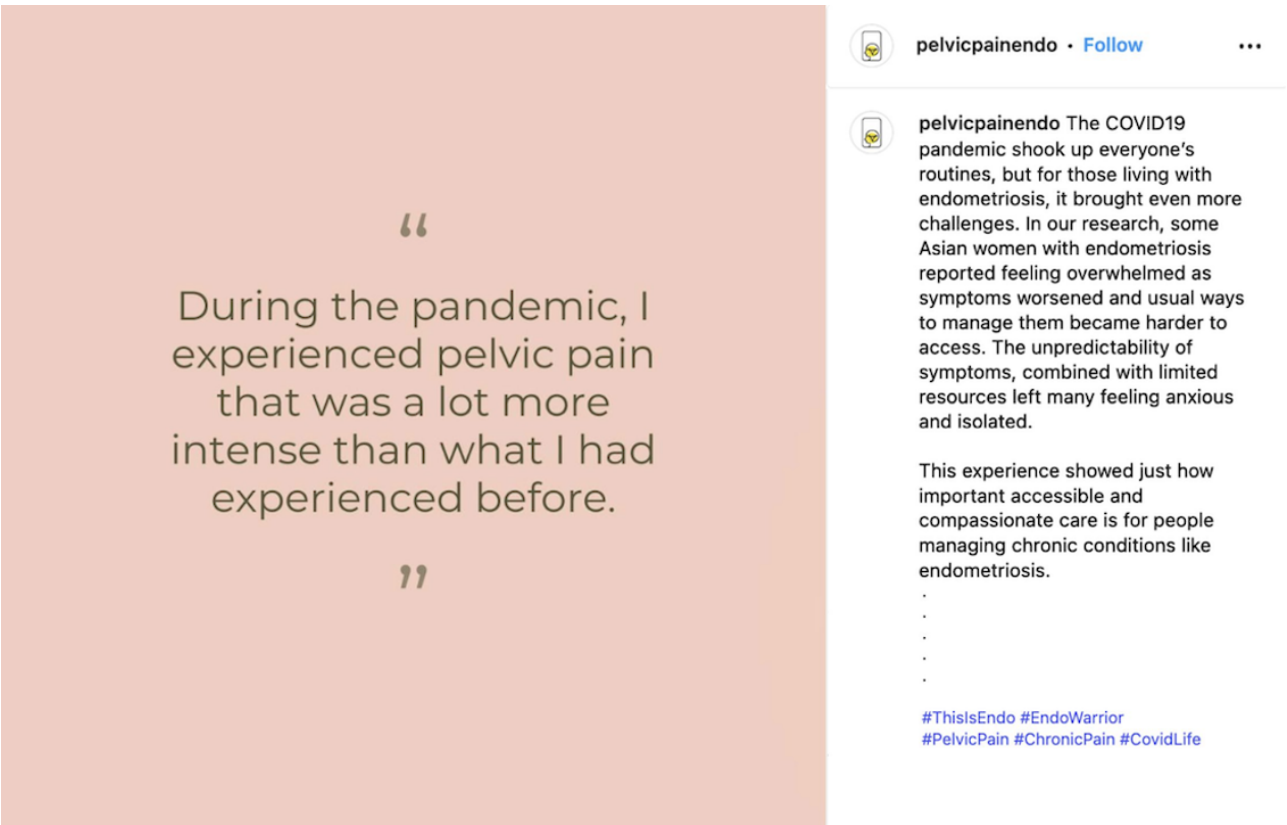

XSL•FO
RenderX

Discussion

Lessons Learned

This manuscript describes the development of a trauma-informed social media campaign designed to disseminate findings from the *EndoPhoto Study*, which explored the experiences of South, East, and Southeast Asian women living with endometriosis in Canada during the COVID-19 pandemic. We intend for this work to serve as a guide for others seeking to share their research findings through social media, with broader applicability for those interested in trauma-informed campaign development. By integrating trauma-informed principles, the campaign not only centered the voices of underrepresented communities but also demonstrated the potential for digital platforms to promote trauma-informed knowledge dissemination.

Upon reviewing our engagement metrics, we found that images with quote-based content (as opposed to image- or quote-only–based content) produced the highest click-through rates on both platforms; announcement posts (eg, study overview) generated the greatest engagement; and Instagram advertisements, particularly images with quotes, outperformed other creative formats (Table 4). Although it is challenging to measure how the content truly impacted viewers, some of these indicators may provide insights into the meaning viewers garnered from the content. For example, the option to "save" content on Pinterest may suggest that some of the posts resonated enough for viewers to want to share or review the content at a later date, perhaps indicating a feeling of added value.

Table . Examples of posts and engagement.

Type of content and example in the manuscript (figure)	Instagram reach	Instagram impressions	Pinterest impressions
Image-based			
Figure 5	484 unique viewers	658 presentations	__[a]
Quote-based			
Figure 8	295	422	259; 96,371 presentations with advertisements
Announcement style			
Figure 2	1396 unique viewers	2361 presentations	—
Reel			
Figure 6	509 unique viewers	840 presentations	—
Advertisement			
Figure 14	1,924,605 unique viewers	2,506,671 presentations	—
Figure 12	1,044,746 unique viewers	1,099,259 presentations	—
Figure 9	1,043,201 unique viewers	1,455,368 presentations	—
Figure 13	1,007,878 unique viewers	1,559,976 presentations	—
Figure 10	—	—	81,523 presentations
Figure 3	14,172 unique viewers	16,963 presentations	23,674 presentations

[a]Not applicable.

At this time, we are unable to say definitively as to why some post types had higher levels of reach and impressions (outside of advertisements). To the best of our knowledge, there is no research literature that shows which specific types of social media content tend to perform better, as it is highly based on the content and target audience. Some social media blogs have suggested that reels, carousel-type posts or those that have multiple images (our announcement posts), and relatable content tend to have higher engagement [40]. Although these content categories are broad, they could indicate that working as a team that has lived knowledge, clinical knowledge, and social media knowledge proved beneficial in understanding the types of content that may be engaging for our audience.

While trauma-informed principles are increasingly recognized in clinical and community settings, their application to digital media, particularly social media, remains largely underexplored. This tutorial highlights how principles, such as collaboration, safety, trustworthiness, voice, and cultural and gender sensitivity, can be applied in online spaces to mitigate harm and increase engagement. Additionally, this project helps to address the critical gap in the representation of racialized individuals, particularly Asian women, in endometriosis advocacy and online discourse.

One unanticipated lesson learned was the suppression of women's sexual and reproductive health information on Meta platforms. A recent report published by the Center for Intimacy Justice [41] highlighted that a bias exists on major Meta platforms, where organizations felt that their content and advertisements related to women's sexual and reproductive health, including fertility and pelvic health, were being censored and overmoderated. Social media algorithms, driven by artificial intelligence, limited content visibility when posts included

XSL•FO
RenderX

"sensitive" health-related terms or were deemed to violate vague platform policies such as Meta's "Personal Health and Appearance" guidelines. This report aligned with our experience, as our content was flagged as not aligning with community standards owing to its "negative" nature and association with health care. This flag necessitated a shift toward resilience-focused and positively framed messaging only, which may have constrained the full scope of participant narratives.

Limitations of the Campaign

Considering that platform selection was intentional, using only Instagram and Pinterest (chosen for their visual nature and perceived safety) may have excluded audiences who primarily engage with platforms like TikTok, X (formerly Twitter), or Facebook. Additionally, the metrics used to gauge success, such as views, impressions, and likes, offer limited insight into true impact. While some posts featuring animals gained viral traction, it is unclear whether the viral nature was due to their relevance to endometriosis or due to the important role that animals can play in people's lives, prompting questions about whether high engagement with content truly reflected increased awareness or understanding of endometriosis specifically.

In addition, because the campaign relied on Meta-owned platforms, our content was influenced by algorithmic restrictions that often suppressed posts containing sexual or reproductive health terms. This may have unintentionally narrowed the range of experiences we were able to highlight. While the campaign focused on the voices of Asian women, our sample does not capture the full diversity of people living with endometriosis, including those with different cultural backgrounds, gender identities, or varying levels of access to digital spaces. Finally, even with trauma-informed strategies in place, sharing health-related stories online carries the risk of re-exposure or secondary trauma for some viewers, especially when content reflects their own experiences.

Key Takeaways

The key takeaways are as follows:

- Trauma-informed principles can be adapted for digital health communication and effectively applied in social media campaigns.
- Posts that featured images with quotes, support networks (including pets), and announcements consistently maximized engagement, suggesting that these formats may be prioritized to ensure engagement.
- Early budget reallocation toward viral creative assets improved cost-efficiency.
- Paid advertisements created opportunities for ensuring wider reach and may be helpful in providing opportunities for content to be viewed by a larger audience.
- Inclusive online storytelling that prioritizes participant voice and emotional safety resonates with audiences and supports effective knowledge translation.
- Algorithmic biases targeted toward women's sexual and reproductive health may necessitate creativity to avoid messages being flagged or framing messaging in "positive" ways. This can ensure a wider reach but highlights gender bias within social media platforms.

Acknowledgments

We would like to thank the participants of the EndoPhoto Study who shared their stories and experiences with us. Moreover, we are grateful to have worked with Mass Velocity Media to develop this social media campaign. Generative artificial intelligence was not used in any form while writing this manuscript.

Funding

This work was funded by a 2024 Michael Smith Health Research BC Reach Award (RA-2024-04266).

Data Availability

Data sharing is not applicable to this article as no data sets were generated or analyzed during this study.

Authors' Contributions

Conceptualization: AFH, HN, NLO
Data curation: HN
Formal analysis: SL
Funding acquisition: KM, AFH, HN, JS, AL, PJY, NLO
Investigation: HN, SL
Methodology: KM, HD, AFH, HN, GJY, WZ, JS, SL, AL, PJY, NLO
Project administration: HN, NLO
Resources: HN
Supervision: KM, AFH, HN, NLO
Visualization: KM, HD, HN, GJY, WZ, AL, NLO
Writing – original draft: KM, HD, GJY, WZ, NLO
Writing – review & editing: KM, HD, AFH, HN, GJY, WZ, JS, SL, AL, PJY, NLO

XSL•FO
RenderX

Conflicts of Interest

There are no major conflicts of interest. However, SL acted as a patient partner and a paid consultant from Mass Velocity Media who contributed to the development and strategies of the social media campaign.

References

1. Zondervan KT, Becker CM, Missmer SA. Endometriosis. N Engl J Med 2020 Mar 26;382(13):1244-1256. [doi: 10.1056/NEJMra1810764] [Medline: 32212520]
2. Johnson NP, Hummelshoj L, Adamson GD, et al. World Endometriosis Society consensus on the classification of endometriosis. Hum Reprod 2017 Feb;32(2):315-324. [doi: 10.1093/humrep/dew293] [Medline: 27920089]
3. Wahl KJ, Yong PJ, Bridge-Cook P, Allaire C, EndoAct C. Endometriosis in Canada: it is time for collaboration to advance patient-oriented, evidence-based policy, care, and research. J Obstet Gynaecol Can 2021 Jan;43(1):88-90. [doi: 10.1016/j.jogc.2020.05.009] [Medline: 32753352]
4. Singh S, Soliman AM, Rahal Y, et al. Prevalence, symptomatic burden, and diagnosis of endometriosis in Canada: cross-sectional survey of 30 000 women. J Obstet Gynaecol Can 2020 Jul;42(7):829-838. [doi: 10.1016/j.jogc.2019.10.038] [Medline: 32001176]
5. Greene R, Stratton P, Cleary SD, Ballweg ML, Sinaii N. Diagnostic experience among 4,334 women reporting surgically diagnosed endometriosis. Fertil Steril 2009 Jan;91(1):32-39. [doi: 10.1016/j.fertnstert.2007.11.020] [Medline: 18367178]
6. Sims OT, Gupta J, Missmer SA, Aninye IO. Stigma and endometriosis: a brief overview and recommendations to improve psychosocial well-being and diagnostic delay. Int J Environ Res Public Health 2021 Aug 3;18(15):8210. [doi: 10.3390/ijerph18158210] [Medline: 34360501]
7. Bougie O, Nwosu I, Warshafsky C. Revisiting the impact of race/ethnicity in endometriosis. Reprod Fertil 2022 Apr 1;3(2):R34-R41. [doi: 10.1530/RAF-21-0106] [Medline: 35514542]
8. Williams C, Long AJ, Noga H, et al. East and South East Asian ethnicity and moderate-to-severe endometriosis. J Minim Invasive Gynecol 2019;26(3):507-515. [doi: 10.1016/j.jmig.2018.06.009] [Medline: 29935381]
9. Kabani Z, Ramos-Nino ME, Ramdass P. Endometriosis and COVID-19: a systematic review and meta-analysis. Int J Mol Sci 2022 Oct 26;23(21):12951. [doi: 10.3390/ijms232112951] [Medline: 36361745]
10. Demetriou L, Cox E, Lunde CE, et al. The global impact of COVID-19 on the care of people with endometriosis. Front Glob Womens Health 2021;2:662732. [doi: 10.3389/fgwh.2021.662732] [Medline: 34816218]
11. Leonardi M, Horne AW, Vincent K, et al. Self-management strategies to consider to combat endometriosis symptoms during the COVID-19 pandemic. Hum Reprod Open 2020;2020(2):hoaa028. [doi: 10.1093/hropen/hoaa028] [Medline: 32509977]
12. Schwab R, Stewen K, Kottmann T, et al. Mental health and social support are key predictors of resilience in German women with endometriosis during the COVID-19 pandemic. J Clin Med 2022 Jun 26;11(13):3684. [doi: 10.3390/jcm11133684] [Medline: 35806968]
13. Leigh JP, Moss SJ, Tiifu F, et al. Lived experiences of Asian Canadians encountering discrimination during the COVID-19 pandemic: a qualitative interview study. CMAJ Open 2022;10(2):E539-E545. [doi: 10.9778/cmajo.20220019] [Medline: 35700997]
14. Han CS, Oliffe JL. Photovoice in mental illness research: a review and recommendations. Health (London) 2016 Mar;20(2):110-126. [doi: 10.1177/1363459314567790] [Medline: 25673051]
15. Marshall K, Howard AF, Marshall N, Noga H, Rojas HE, Leonova A, et al. Impacts of the COVID-19 pandemic on the mental health of asian women with endometriosis in canada: a photovoice study. SAGE Women's Health (forthcoming) 2025.
16. Marshall N, Howard AF, Marshall K, et al. Endometriosis and expressions of self-management and resilience among asian women living in canada during the COVID-19 pandemic: a photovoice study. J Public Health Res 2026 Jan;15(1). [doi: 10.1177/22799036251407192]
17. EndoPhoto. Endometriosis and Pelvic Pain Laboratory. URL: https://pelvicpainendo.ca/photovoice-endophoto [accessed 2025-07-15]
18. Harris M, Fallot RD. Envisioning a trauma-informed service system: a vital paradigm shift. New Dir Ment Health Serv 2001(89):3-22. [doi: 10.1002/yd.23320018903] [Medline: 11291260]
19. Makosis P, Greenwood M. What's new is really old: trauma-informed health practices through an understanding of historic trauma. National Collaborating Centre for Indigenous Health (NCCIH). 2017. URL: https://www.nccih.ca/495/Webinar__What_s_new_is_really_old__Trauma_informed_health_practices_through_an_understanding_of_historic_trauma.nccah?id=205 [accessed 2026-01-14]
20. SAMHSA's concept of trauma and guidance for a trauma-informed approach. : Substance Abuse and Mental Health Services Administration; 2014 URL: https://www.health.ny.gov/health_care/medicaid/program/medicaid_health_homes/docs/samhsa_trauma_concept_paper.pdf [accessed 2026-01-14]

XSL•FO
RenderX

21. Raja S, Hasnain M, Hoersch M, Gove-Yin S, Rajagopalan C. Trauma informed care in medicine: current knowledge and future research directions. Fam Community Health 2015;38(3):216-226. [doi: 10.1097/FCH.0000000000000071] [Medline: 26017000]
22. Caring for patients who have experienced trauma: ACOG Committee opinion, number 825. Obstet Gynecol 2021 Apr 1;137(4):e94-e99. [doi: 10.1097/AOG.0000000000004326] [Medline: 33759830]
23. Parmar G, Howard AF, Noga H, et al. Pelvic pain & endometriosis: the development of a patient-centred e-health resource for those affected by endometriosis-associated dyspareunia. BMC Med Inform Decis Mak 2025 Feb 13;25(1):79. [doi: 10.1186/s12911-025-02907-x] [Medline: 39948529]
24. Burgess J, Marwick A, Poell T. Editors' introduction. In: Burgess J, Marwick A, Poell T, editors. The SAGE Handbook of Social Media: SAGE Publications; 2018:1-10. [doi: 10.4135/9781473984066.n1]
25. Davis JL. Social media. In: Mazzoleni G, editor. The International Encyclopedia of Political Communication: John Wiley & Sons, Inc; 2016. [doi: 10.1002/9781118541555.wbiepc004]
26. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85. [doi: 10.2196/jmir.1933] [Medline: 23615206]
27. Ezeilo CO, Leon N, Jajodia A, Han HR. Use of social media for health advocacy for digital communities: descriptive study. JMIR Form Res 2023 Nov 14;7:e51752. [doi: 10.2196/51752] [Medline: 37962914]
28. Shawky S, Kubacki K, Dietrich T, Weaven S. Using social media to create engagement: a social marketing review. JSOCM 2019 Apr 8;9(2):204-224. [doi: 10.1108/JSOCM-05-2018-0046]
29. Aïmeur E, Amri S, Brassard G. Fake news, disinformation and misinformation in social media: a review. Soc Netw Anal Min 2023;13(1):30. [doi: 10.1007/s13278-023-01028-5] [Medline: 36789378]
30. Scott CF, Marcu G, Anderson RE, Newman MW, Schoenebeck S. Trauma-informed social media: towards solutions for reducing and healing online harm. Presented at: 2023 CHI Conference on Human Factors in Computing Systems; Apr 23-28, 2023. [doi: 10.1145/3544548.3581512]
31. Abdulai AF, Howard AF, Yong PJ, Currie LM. Defining destigmatizing design guidelines for use in sexual health-related digital technologies: a Delphi study. PLOS Digit Health 2023 Jul;2(7):e0000223. [doi: 10.1371/journal.pdig.0000223] [Medline: 37436972]
32. Zheng W, Walquist E, Datey I. "It's not what we were trying to get at, but I think maybe it should be": learning how to do trauma-informed design with a data donation platform for online dating sexual violence. Presented at: 2024 CHI Conference on Human Factors in Computing System; May 11-16, 2024. [doi: 10.1145/3613904.3642045]
33. Josephs JC, Bungay V, Guta A, Gilbert M, Abdulai AF. Trauma-informed technology design in digital sexual health interventions. Stud Health Technol Inform 2024 Jul 24;315:773-774. [doi: 10.3233/SHTI240323] [Medline: 39049423]
34. Chen JX, McDonald A, Zou Y, et al. Trauma-informed computing: towards safer technology experiences for all. Presented at: 2022 CHI Conference on Human Factors in Computing System; Apr 29 to May 5, 2022. [doi: 10.1145/3491102.3517475]
35. Heale R, Noble H. Integration of a theoretical framework into your research study. Evid Based Nurs 2019 Apr;22(2):36-37. [doi: 10.1136/ebnurs-2019-103077] [Medline: 30894364]
36. Kothari A, Wathen CN. A critical second look at integrated knowledge translation. Health Policy 2013 Feb;109(2):187-191. [doi: 10.1016/j.healthpol.2012.11.004] [Medline: 23228520]
37. Crosschild C, Huynh N, De Sousa I, Bawafaa E, Brown H. Where is critical analysis of power and positionality in knowledge translation? Health Res Policy Syst 2021 Jun 11;19(1):92. [doi: 10.1186/s12961-021-00726-w] [Medline: 34116685]
38. What's a good engagement rate on Pinterest? Hudson Design Company. URL: https://www.hudsondesigncompany.com/whats-a-good-engagement-rate-on-pinterest-exploring-the-industry-standards [accessed 2025-07-15]
39. Polishchuk D. What is a good Instagram engagement rate? Promo Republic. 2022. URL: https://promorepublic.com/en/blog/what-is-a-good-instagram-engagement-rate [accessed 2025-07-15]
40. Rose-Collins F. Content types that perform best on Instagram. Ranktracker. 2025. URL: https://www.ranktracker.com/blog/content-types-that-perform-best-on-instagram [accessed 2025-12-12]
41. The digital gag: supression of sexual and reproductive health on Meta, TikTok, Amazon and Google. : Center for Intimacy Justice; 2025 URL: https://www.intimacyjustice.org/ [accessed 2026-01-14]

Abbreviations

IKT: integrated knowledge translation
PRAB: Patient Research Advisory Board
SAMHSA: Substance Abuse and Mental Health Services Administration

XSL•FO
RenderX

Edited by A Mavragani; submitted 04.Sep.2025; peer-reviewed by SS Yilmaz, T Harpel; accepted 23.Dec.2025; published 27.Jan.2026.

Please cite as:
Marshall K, Dhillon H, Howard AF, Noga H, Yang GJ, Zhu W, Sutherland J, Lett S, Leonova A, Yong PJ, Orr NL
Developing a Trauma-Informed Social Media Campaign to Disseminate Endometriosis-Specific Qualitative Art-Based Research Findings: Tutorial
J Med Internet Res 2026;28:e83491
URL: https://www.jmir.org/2026/1/e83491
doi:10.2196/83491

© Kerry Marshall, Hargun Dhillon, A Fuchsia Howard, Heather Noga, Grace J Yang, William Zhu, Jessica Sutherland, Sarah Lett, Anna Leonova, Paul J Yong, Natasha L Orr. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Tutorial

Team-Based Analysis of Large-Scale Qualitative Data: Tutorial Using a Nationwide SMS Text Messaging Poll of Youth

Melissa DeJonckheere[1,2], PhD; Samantha A Chuisano[1], MPH; Marika Waselewski[1], MPH; Kendrin Sonneville[3], SCD, RD; Tammy Chang[1,2], MD, MPH, MS

[1]Department of Family Medicine, University of Michigan, Ann Arbor, MI, United States

[2]Institute for Healthcare Policy and Innovation, University of Michigan, Ann Arbor, MI, United States

[3]School of Public Health, Department of Nutritional Sciences, University of Michigan, Ann Arbor, MI, United States

Corresponding Author:
Melissa DeJonckheere, PhD
Department of Family Medicine
University of Michigan
1018 Fuller Street
Ann Arbor, MI, 48105
United States
Phone: 1 7349361927
Email: mdejonck@umich.edu

Abstract

With the growing use of technology in qualitative data collection and analysis, there is an opportunity to gather rich and varied perspectives to improve health and well-being. However, large-scale qualitative datasets can be difficult to manage using traditional qualitative methods, and there are few examples of the application of large-scale qualitative analysis. In the context of digital health, large qualitative datasets are increasingly made up of short text segments, which need to be analyzed differently from lengthy transcripts from interviews or focus groups. Therefore, this tutorial describes the use of traditional qualitative methods to analyze a large corpus of qualitative text data. We use examples from a nationwide SMS text messaging poll of youth to highlight the opportunities to use this team-based analysis approach, which has been accessible and meaningful to youth researchers and novice qualitative researchers. These large-scale qualitative strategies may benefit novice researchers analyzing large volumes of qualitative data and short text segments, including SMS text messaging, social media posts, medical notes, and open-ended survey questions, among others.

International Registered Report Identifier (IRRID): RR2-10.2196/resprot.8502

(J Med Internet Res 2026;28:e72526) doi:10.2196/72526

KEYWORDS

qualitative; large-scale; survey methods; data analysis; adolescence

Introduction

Large-scale qualitative methods can gather the diverse perspectives of many people to augment our understanding of complex health and social issues. Traditional qualitative methods have focused on depth rather than breadth of analysis, resulting in smaller samples with rich detail. However, large-scale qualitative methods can emphasize both depth and breadth, responding to common critiques of qualitative research related to external validity (ie, generalizability in quantitative research or transferability in qualitative research). Large-scale qualitative approaches are appropriate for some studies and research questions, including those that use large primary or secondary text data sources (eg, social media posts, documents, and medical records) [1], multiple sites for data collection [2], longitudinal studies [3], or mixed-methods studies where understanding experiences across all participants is desirable (eg, clinical trials) [4-6]. Advances in technology have also increased access to large qualitative data sources through data mining techniques, such as social media sites, electronic health records [7,8], and patient portal messages [4]. As a result, mobile health and other digital health researchers have access to vast qualitative datasets of short text data.

Much of the existing literature on large-scale qualitative studies has focused on collaborative research teams that collect data across multiple sites, institutions, states, or nations. Traditional approaches for qualitative analysis leverage a line-by-line review

XSL•FO
RenderX

of each interview transcript or open-ended text response. However, the time and resources required to apply these approaches to a dataset with thousands of qualitative datapoints would be restrictive for nearly all research teams. Little practical guidance exists for researchers on how to approach data management and analysis of these large datasets [6,9]. As a result, some teams may intentionally narrow their analytic focus using purposive sampling techniques [3,5], including a subsample of responses into their analysis, or use rapid approaches to qualitative analysis [10,11]. Furthermore, many teams leverage software to assist with data management, coding, or summarization [12-14]. However, these approaches may introduce the need for additional time and resources for learning and acquiring new tools or new methodologies [13,15].

The growth of artificial intelligence (AI) has shaped some approaches to qualitative data analysis as some researchers seek methods to improve efficiency or increase objectivity [16]. For example, natural language processing has been used to support text analysis of large text datasets through the initial development of a coding structure or validation of themes [4,17,18]. Large language models, like ChatGPT (OpenAI), have been used to explore the use of qualitative analysis through summarization and theme development [19-21]. However, AI-supported qualitative analysis has thus far been limited in its ability to understand nuance and context in qualitative data and develop interpretive themes [18,21]. Furthermore, some qualitative researchers have rejected the use of AI in interpretive and reflexive qualitative approaches, which are intentionally human and subjective processes used to make meaning of human thoughts and experiences [22].

As human interpretation is still integral to qualitative research, practical guidance for researchers seeking to conduct analysis of increasingly large and complex studies in digital health is needed [6,9]. While analysis can be burdensome, human-based qualitative processes can be conducted efficiently on large datasets without sacrificing methodological rigor. Likewise, analysis of large qualitative datasets can be completed without the need for advanced methodologies or acquisition of new software [15]. Therefore, the purpose of this tutorial is to describe an example of team-based qualitative analysis of a large-scale SMS text message poll using traditional qualitative methods and freely available, easy to learn collaboration tools, including cloud-based spreadsheets. While spreadsheets and matrices have been used to organize and analyze transcripts from interviews and focus groups, [23,24] there is limited guidance on strategies for datasets with a large volume of short text responses common in digital health studies. We describe our experience conducting team-based qualitative analysis of open-ended survey responses from a nationwide SMS text messaging poll of youth across the United States [25]. Our approach has been used to efficiently and effectively engage both new and experienced researchers in qualitative research, including youth, students, and trainees.

Methods

Example Study

To illustrate our analysis approach, we use MyVoice, a nationwide poll of individuals aged 14 to 24 years across the United States that is administered via SMS text messaging [25]. The goal of MyVoice is to gather diverse youth perspectives across the United States to inform policies and programs that impact youth. Incorporating youth voices can create more relevant, effective, sustainable, and impactful interventions that ultimately improve the health of young people [26,27]. Increasingly, research has included the perspectives of adolescents, rather than adult proxies, while also considering the unique needs of this age group [28]. Participatory approaches to research have also successfully integrated adolescents into the research team to increase the relevance of research to youth populations while also promoting youth development and capacity [29-31]. However, engaging adolescents in research remains challenging when conducted in traditional settings and when using methods that are not accessible or approachable to young people.

MyVoice is designed to address challenges related to engaging youth in research through familiar and easy-to-use methods, like SMS text messaging to collect personal opinions and stories. Each MyVoice poll consists of 4-5 primarily open-ended questions designed to elicit youth perspectives on complex health and social issues. Participants respond to questions directly via SMS text messaging, rather than following a link to an external platform. The MyVoice research protocol was previously described after enrolling approximately 100 participants [25]. While the study rationale and overall methodology have not changed since 2016, there are currently approximately 800 MyVoice participants in each cohort to maximize understanding of variations in their lived experiences to better inform policy and practice. A brief description of the overall methods is presented below, highlighting considerations for large-scale data collection relevant to this tutorial. For additional details on recruitment, screening, and enrollment of MyVoice participants, see Multimedia Appendix 1.

Weekly Polls

SMS text messaging polls are administered using a cloud-based platform to send and receive SMS messages directly, which includes Textizen and the University of Michigan instance of REDCap (Research Electronic Data Capture; Vanderbilt University) with Twilio integration. REDCap is a secure, web-based software platform designed to support validated data capture, data management, and integration for research studies [32,33]. Questions are sent one at a time, and respondents must send an SMS text message to receive the next question. Participants are given 7 days to complete the poll before the next poll is automatically sent. Participants who have not responded after 2-3 days are sent a reminder to complete the poll. Each week, approximately 2750-3250 open-ended responses are collected. Since 2016, we have collected over 120 SMS text messaging polls and 45,000 unique responses.

XSL•FO
RenderX

Ethical Considerations

The MyVoice study was approved by the Institutional Review Board at the University of Michigan Medical School (HUM00119982). All participants provide online informed consent or assent (for those under the age of 18 years) prior to enrollment in the study. Parental consent is waived for minors. Participants receive US $1 for each poll that they answer (each poll consists of 4-5 questions) and US $5 for completion of a demographic survey upon enrollment. All study data are collected and stored in secure web-based platforms. Data are deidentified before shared with team members and any external collaborators.

Data Analysis of Large-Scale Qualitative Data

Analysis Overview

Below, we outline the efficient and reliable method we have used for team-based coding of large qualitative datasets. The overall approach includes five distinct steps: (1) data cleaning and management, (2) codebook development, (3) coding, (4) reconciliation, and (5) presentation of findings, which are described in detail below. Our approach leverages traditional qualitative approaches (ie, manual human coding) and enables collaboration among teams at various experience levels by using a simple spreadsheet – a freely accessible tool that can be used with minimal to no training. While spreadsheets have been used to structure coding of interview and focus group data, few studies [15,23,24,34] have described the use of spreadsheets for qualitative analysis of short open-ended text responses. Worked examples can be particularly beneficial in highlighting strategies that have been successfully implemented by other researchers to manage qualitative data and make study design decisions [15,35]. In the example provided, we highlight specific considerations for large-scale short text responses from the MyVoice study; however, this method could be easily applied to any study where data are available in a table or a .csv format (ie, social media data, unstructured medical record text, open-ended survey responses, etc).

Step 1: Data Cleaning and Management

SMS text messaging responses are downloaded from the data collection platform as a .csv file. Data are deidentified when downloaded and identified by a participant identification number. Cleaning includes merging responses with participant demographic data, rearranging lines of text that were separated, preliminary identification of junk or skipped responses, and filtering out nonresponses (ie, when a respondent does not answer a question).

The resulting deidentified spreadsheet is organized with the following columns: participant study ID, responses separated by question, and demographic survey responses, each in a separate column. Each row in the file includes the data for a unique participant.

For all MyVoice studies, analyses include line-by-line coding of deidentified participant responses in a spreadsheet. Teams primarily use a cloud-based platform (eg, Google Sheets) to ensure that all members of the research team (including high school students, undergraduate students, and other trainees) have access to the data throughout the analysis process and are able to work synchronously and asynchronously.

Step 2: Codebook Development

Codebooks are used to systematically record and organize qualitative data analysis by providing a list of key ideas in a dataset (ie, codes) and ensuring consistency and transparency in analysis [13]. MyVoice teams typically use a structured codebook to record the emerging ideas in their large-scale text data, including codes, definitions, and examples [13]. The codebook development process includes three overarching and iterative steps: (1) data familiarization through open coding and memoing, (2) identifying shared concepts across team members, and (3) defining and refining codes. Our codebooks are developed iteratively and collaboratively, with multiple team members offering interpretations of the important ideas that are present in the dataset and revising the list of codes throughout the analysis process [36].

Data Familiarization

Within each study team, at least 2 members independently read a portion of participant responses. Team members read approximately 100 responses per question, which in most datasets is approximately 10-20% of the responses. Multiple research team members independently use open coding to list the main concepts in the responses in an analytic memo, a document reflecting the researcher's thinking throughout the analysis process [13,37]. In our team, memos typically include a running list of concepts that are repeated in the data that will inform future codes. Some researchers include personal reflections, emerging patterns, and connections to the literature in their analytic memos [13].

Identifying Shared Concepts

Team members compare analytic memos to create a synthesized list of concepts that were identified in the responses. These concepts form the basis of the codes, although they may be renamed or modified in subsequent steps.

Defining and Refining Codes

Each concept is defined in a structured codebook that includes the code (ie, label for the concept), definition, and example response [13,38]. Some teams also include inclusion and exclusion criteria in their codebook, depending on the type of question and desired overlap between codes [38]. In studies with poll questions that are close-ended or when using a more deductive analysis approach, responses are typically categorized into discrete codes with clear inclusion and exclusion criteria. Teams may also structure their codebook with rules for co-coding responses, which can be helpful when analyzing responses that include both categorical and narrative responses.

Figure 1 displays an excerpt from a structured codebook that was used to analyze a dataset on youth perspectives on telemedicine before and after the start of the COVID-19 pandemic [39]. In total, 5 open-ended questions were asked in October 2019, and a new set of 5 questions was asked in October 2020. A separate coding scheme was created for all 10 questions.

XSL•FO
RenderX

Figure 1. Excerpt from structured codebook organized by poll question (Waselewski et al, 2022 [39]).

Q2: If you were to use telemedicine to see a healthcare provider, would you prefer a phone call or video visit? Why?

1. **Video**
   a. Definition: A response that indicates they would **prefer a video** visit for their telemedicine care
2. **Phone Call**
   a. Definition: A response that indicates they would **prefer a phone** visit for their telemedicine care
3. **Depends/Either**
   a. Definition: A response that indicates their **preference depends** on various situations
   b. Examples: "I would be okay with either." or "I think it would depend on the reason"
4. **Neither/never use**
   a. Definition: A response that indicates they would **not use either** or would never use telemedicine
   b. Examples: "I don't believe I would ever use it."
5. **Idk/Not sure**
   a. Definition: A response that indicates they are **not sure** what their preference is; not used if they say it depends or if they give a preference with an exception
   b. Examples: "Not sure" or "Never even heard of it"
6. **Skip/Junk**

Video reasons

7. **More personal**
   a. Definition: A response that indicates they prefer **video** because it is more personal, they feel more connected, or would generally be more comfortable
   b. Examples: "face to face seems more personal" or "i prefer a video visit, it's seems most personal" or "because it would make me more comfortable to see my doctors face"
8. **Better communication**
   a. Definition: A response that indicates they prefer **video** because it is easier to read facial expressions, body language, and just generally communicate more effectively via video
   b. Examples: "I feel it's easier to talk to a doctor when I can see them" or "Easier to understand with the facial expression and non body language"
9. **Visual**
   a. Definition: A response that indicates they prefer **video** because there are opportunities for more visual issues or assessments to be addressed
   b. Examples: "Video visit in case I need to show them anything" or "I would want them to be able to look at me."
10. **Natural or real**
   a. Definition: A response that indicates they prefer **video** because it feels more real, more natural, or is closer or more similar to in person
   b. Examples: "Video so it feels more normal!" or "being able to interact in a way that feels closer to reality would at least make me feel much more secure in the type of care I felt I was receiving."
11. **Attention or engaged**
   a. Definition: A response that indicates they prefer **video** because they feel like they are able get more attention, it creates more engagement, they want to be seen/see the provider, or it creates more rapport
   b. Examples: "In order to physically see them and having their undivided attention." or "I think a video visit because then they could actually see me" or "Video visit I want them to see me"

Phone reasons

12. **More personal**
   a. Definition: A response that indicates they prefer **phone calls** because they think they are more personal or are more comfortable with them
   b. Examples: "I'd be more comfortable not being on camera" or "it's more personal obviously"
2. **Easiest**
   a. Definition: A response that indicates they prefer **phone calls** because it is easier to talk on the phone, generally easiest method, or more convenient than video
   b. Examples: "Phone call because it's easier to talk when only working about one thing" or "Probably a phone call because it's less exhausting" or "Phone because that is just easiest"
3. **Natural**
   a. Definition: A response that indicates they prefer phone calls because they are more natural or normal for them
   b. Examples: "Phone call bc its more natural"
4. **Dislike video**
   a. Definition: A response that indicates they prefer **phone calls** because they generally dislike video chats or find them more awkward
   b. Examples: "Phone call. It's awkward sometimes with videos" or "Phone call because I don't like video visits" or "I feel really awkward on video."
5. **Anxious on video**
   a. Definition: A response that indicates they prefer **phone calls** because they are self conscious, nervous, uncomfortable, or anxious about being on video
   b. Examples: "Phone call, i get self conscious on video" or "A phone call! Video calls make me nervous"
6. **Appearance concerns**
   a. Definition: A response that indicates they prefer **phone calls** because they are less formal, they don't have to worry about how they look, or worry about what they wear
   b. Examples: "less concerned about appearance" or "It's nice to be able to have a doctor's visit in PJs!"
7. **Tech issues**
   a. Definition: A response that indicates they prefer **phone calls** because they think video is a clunkier process, more likely to have tech issues, and generally less reliable
   b. Examples: "Phone, video visits usually mean a clunky app" or "probably cut out less"
8. **No point to video**
   a. Definition: A response that indicates they prefer **phone calls** because they don't see a point or added benefit to using video visits
   b. Examples: "I really don't need to see the person unless it's something they need to see." or "Because a video visit seems pointless. Everything can be said over the phone."

Caveats (Co-coded with "Depends on..." or "Video but…" or "Phone but…")

9. **Depends on visit type**
   a. Definition: A response that indicates their preference for video or phone **depends** on the type of visit they are having or that they prefer one option except for a specific type of visit
   b. Examples: "Phone call unless I had someone in mind I needed to show them - wound, rash, etc" or "Phone call for regular doctor talks and visits, video for something more specialized to allow the doctor to get a better picture and build better rapport with me"
10. **Depends on provider relationship**
   a. Definition: A response that indicates their preference for video or phone **depends** on how well they know their provider; note that some prefer video for new providers to get to know them or build rapport while others would prefer a phone call for someone they do not know

For example, in the October 2020 poll, 1 question read, "If you were to use telemedicine to see a health care provider, would you prefer a phone call or video visit? Why?" The research team inductively developed codes for this question that included categorical codes that captured the first part of the question (ie, preference for phone or video visits) and descriptive codes for the subsequent explanation. Categorical codes included "Video," "Phone call," "Depends or either," "Neither or never use," or "I don't know or not sure" to understand participants' preferences for video or phone visits when using telemedicine. Here, there is a clear distinction in the categories, and responses cannot be categorized into multiple codes.

Each categorical code was also accompanied by a list of reasons that were derived from the text responses explaining their preference between phone and video visits (ie, descriptive codes). As seen in Figure 1, respondents who indicated a preference for telemedicine using video often indicated in their responses that video visits were more personal, improved communication, included visuals that could help with assessment, better approximated an in-person visit, and improved engagement with the provider. These reasons were then translated into descriptive codes that could be co-coded alongside "Video" (eg, "more personal," and "better communication"). Respondents who preferred phone visits described them as more personal, easy, and natural, among other factors. These reasons were translated into descriptive codes that could only be co-coded with "Phone call." While this team had rules for co-coding such that the concept of "more personal" had a code for preferring video visits and a separate code for preferring phone calls, an alternative strategy would be to have a single code of "more personal" that could be applied across categorical codes. Decisions are made by the research team to determine the scope of their codebook and direction for analysis. In this example, having separate codes allowed for quick calculations of frequency for each code in the content analysis.

Most MyVoice teams create a separate coding scheme for each question (as in Figure 1). A separate codebook for each question is helpful when conducting a content analysis, when the questions and, therefore, responses vary significantly, or to focus the analysis. It is also possible to create an overarching coding scheme that can be applied across questions in a single dataset. A single, overarching coding scheme would be appropriate when the poll questions are closely tied, and content from multiple responses can be captured with related codes. Alternatively, if analyzing all the data by individual respondent, rather than by question, a single coding scheme would also be optimal. This approach would closely align with most thematic analysis approaches that proceed with line-by-line coding using an individual participant or individual interview transcript, observation, or other textual data as the unit of analysis.

See, for example, an excerpt from a codebook developed by high school student co-researchers [40]. During the MYHealth research training program, high school students developed a

XSL•FO
RenderX

study to understand youth perspectives on the impact of social media on adolescents and young adults. Poll questions included the following: (1) How do you typically use social media (posting, watching videos, scrolling, messaging, commenting, etc.)? (2) What benefits do you get from social media? (3) How do you create a positive experience for yourself while using social media? (4) Who or what do you enjoy seeing or interacting with on social media? Why? (5) Tell us about a time when social media made you feel good about yourself. This team elected to analyze responses across individual participants, rather than by question, to track patterns across all 5 questions. See Figure 2 for an excerpt from the resulting codebook.

Figure 2. Example excerpt from MYHealth social media dataset with single codebook.

Code Name	Code Definition	Examples
SM Breaks	limits the amount of time they spend on social media, takes breaks from social media, or deletes social media apps	"I try to follow creators that are beneficial to my mental health and try to limit my screen time" "i recently decided to do a social media cleanse so as of right now i don't use a ton of social media..." "By being selective in what I allow on my screen"
Tailor content	Makes an effort to follow, view, post positive content or unfollow/block negative content.	"By blocking accounts that make me unhappy and only viewing specific things so that the algorithm suggests related posts" "I block toxic people" "I don't really look at celebrities or stuff that'll make me feel bad about myself, I don't spend a lot of time on social media either. And I don't post content for others to judge good or bad "
Maintain personal circle	This person tries to use SM to connect with, post content to, or share content with ***ONLY*** people they know personally. This code can also be used if the person says they avoid content from celebrities and influencers	"...I only let people I personally know or have known follow me." "I am mostly a lurker and I rarely post anything or interact beyond liking unless I personally know the poster" (in response to Q4) "My friends because it's personal" "With personal social media like Snapchat and instagram I only follow and let people follow me that are close"
Don't worry about other people's opinions	When the person says they try not to worry about what others think about them or avoid comments on their own posts	"i focus on myself, and don't let what others say, both negatively and positively, impact my self respect and self esteem" "Avoiding getting caught up being concerned about likes or comments" "I don't read the negative comments or most comments on anything I post...."
Reduce loneliness	Using social media to feel less isolated or alone or to feel accepted (that it is okay to be themselves)	'I have severe scoliosis and I found a community of other women who struggle with the same issue and it makes me **feel less isolated**" "laughter, sharing with friends, and realizing **I'm not alone** in thoughts/feelings/actions" "I enjoy seeing posts about mental health because **it makes me feel less alone and** about my favorite books/tv shows because it's fun to see content about things I enjoy" "It has helped me a lot to see other trans and queer people. It's hard to feel accepted in a conservative state, but being able to connect with others reminds me **that I am okay, and that I'm not alone**. "
Community	Uses social media to connect with people or connect with a community or people with similar backgrounds or interests	"I enjoy interacting with my friends as well as online users who are in the same community as I am such as shared hobbies" "I love sports content (especially related to my favorite teams and fans I've made friends with) and creative content (photography, design, etc.) because it helps create a sense of community" "Feeling a sense of community and staying informed" Feel apart of a community" ""I enjoy seeing people of color, women, and queer people. I share identities with and they are also widely underrepresented and promoted on social media"
Distraction	Using social media as a way to distract themselves or as a way to pass or fill time. This can be positive distraction or avoidance of negative things	"Scrolling on my phone because **I'm bored**" "It's a nice **distraction** and I use it to keep in contact with friends" "**Distraction** in case I need it." "**Turns my brain off**"
Receives Attention	When the respondent says they have received positive attention or support from SM	"feeling plugged in, humor, a mind numbing way to **pass the time**" "when people like/comment/share their posts/**receiving attention**" "**positive reinforcement**/positivity" "When I was included in an **appreciation post**"
Entertainment	This person uses social media as a way to find amusement, boost their mood, or get inspired. This can also include someone saying they use SM to find funny content or to laugh	"I find **entertainment** from watching videos about dogs, Scrolling through videos and watching inspiration from them give me a mood booster and is entertaining to me." "I enjoy interacting with niche influencers that align with my passions and morals. they **inspire** me to further shoot for my goals & also give me **inspo** for different aspects of my life" "... **inspiration** from other people" "I post reels on my story, text my friends , scroll through posts, send funny or **inspiring** or relatable things to my friends"

Step 3: Coding

In Google Sheets, Microsoft Excel, or another spreadsheet, each code from the finalized codebook is entered into a new, separate column. Within each research team, at least 2 members independently perform line-by-line coding of each participant's response. Cloud-based tools allow team members to collaboratively code data in real-time.

Team members use the codebook developed in the previous step to apply codes to each participant's response. Where a concept is described by the participant, a “1” is entered into the spreadsheet cell. All other cells remain blank (or could be entered as “0”). In most cases, multiple codes may be applied to a participant’s response to reflect the many ideas that are present. Many responses include multiple sentences, phrases, or lists of ideas that are unique concepts.

For example, in response to the question about preference between video or phone visits for telemedicine, one respondent shared:

> *Video visit, for sure. Because the non-verbal cues that come along with being able to see someone are extremely important. And with my primary, I wouldn’t feel satisfied with a phone call because being able to interact in a way that feels closer to reality would at least make me feel more secure in the type of care I was receiving.*

This response included several important concepts, which should each be coded to accurately reflect the data. The research team

XSL•FO
RenderX

identified the following codes: (1) video, (2) better communication, (3) natural or real, and (4) works better (represented in the highlighted row in Figure 3).

Figure 3. Example coding sheet and completed coding.

Study ID	If you were to use telemedicine to see a healthcare provider, would you prefer a phone call or video visit? Why?	Video	Call	Depends /either	More personal	Better com	Visual	Natural /real	Attention /engaged	Dislike phone	Anxious on phone	Tech issues	Works better	More personal	Easiest	Natural	Dislike video
95	Phone call for regular doctor talks and visits, video for something more specialized to allow the doctor to get a better picture and build better rapport with me			1			1		1								
98	Video call because it feels more personal to me	1			1												
108	Video visit. It seems more personal to me	1			1												
111	Video visit, for sure. Because the non-verbal cues that come along with being able to see someone are extremely important. And with my primary, I wouldn't feel satisfied with a phone call because being able to interact in a way that feels closer to reality would at least make me feel much more secure in the type of care I felt I was receiving.	1				1		1					1				
112	No because I want them to see my issues in person																
131	the first time i'd want a video call to see who i'm talking to, but after that i'd be fine to do it over a phone call			1													
141	Video to get the nonverbals	1				1											
145	Video visit so I can also show what I am talking about	1					1										

Step 4: Reconciliation

Depending on the analysis strategy, teams often review and reconcile discrepancies in coding to reach agreement on the meaning of each response and final assignment of codes. Teams will use a spreadsheet formula to quickly identify discrepancies in applied codes across the large dataset. Disagreements are labeled in a new column so they can easily be filtered and reviewed. To resolve disagreements and reach consensus, the 2 independent coders compare their applied codes, discuss the rationale for the codes, and address the discrepancy by either agreeing on the final application of codes or asking a third member of the research team to review. With online tools, reconciliation can take place in real-time to support collaboration and integration of multiple perspectives.

Step 5: Development of Findings (eg, Themes or Descriptions)

Content Analysis

After the coding process is complete, MyVoice teams often summarize codes with frequencies to determine the prevalence of codes and categories (ie, groups of related codes) across participants. Teams will often present code frequencies (or the proportion of responses that are aligned with a code) in their findings. Presentations of frequencies are often grouped by question or overarching main ideas across the dataset, supported by sample responses. For example, Schmid et al [41] examined the impact of COVID-19 on weight and shape control behaviors among youth. The authors conducted a content analysis that involved developing initial themes, constructing a structured codebook with codes and definitions, independently applying codes to responses for each survey question, comparing discrepancies between each pair of coders, and finalizing themes and subthemes that summarized the main ideas in responses to each question. Content analysis of the prompt, “Since the pandemic, have you changed your eating or exercise habits because you were worried about your weight or shape. Tell us about it,” resulted in 2 themes: changes for weight or shape control and changes in overall attitude or concerns. Overall, 4 subthemes that detail the responses common in each theme were presented alongside the frequency of responses. By capturing the frequencies in responses, the authors concluded that one-third of the sample implemented weight or shape control behaviors during the COVID-19 pandemic.

Thematic Analysis

Theme development occurs by looking for relationships across codes, both within and across questions in the dataset. MyVoice teams often develop themes through iterative team discussion where codes are grouped into related categories or concepts, patterns are identified across categories or concepts, and variation in participant experiences is described. For example, a MyVoice study examined youth perspectives on use and access to primary care [42]. Questions included:

1. Where do you go for a health problem you can’t solve on your own (family doctor, ER, urgent care, school clinic, etc)? Why do you go there?
2. Do you have a regular doctor you go to for check-ups or when you feel sick? If you don’t have a regular doctor, why not?
3. Do you think it’s important to have a regular doctor you go to for check-ups or when you feel sick? Why or why not?
4. How could your doctor’s office make your visits easier?
5. Have you had a bad experience with a doctor or clinic visit? What happened?

Using a thematic analysis approach, the research team followed the analysis steps outlined above, including data familiarization, codebook development, line-by-line coding, code reconciliation,

XSL•FO
RenderX

and consolidation of conceptually related categories. To develop themes, the team held many discussions where 6 faculty and program staff reflected on patterns in the data. They considered prevalent categories (eg, convenience) in addition to those that were salient but less prevalent (eg, experiences with physical and emotional harm in health care settings). One prominent theme was "Feeling respected, valued, and heard was essential to youth having positive experiences in a primary care health care setting." In the description of this theme and in the selection of representative quotes in the final manuscript, the authors presented the range of experiences (ie, both positive and negative previous experiences with primary care) among MyVoice participants to fully elaborate on the breadth of responses [42].

Quantitative Analysis

Occasionally, teams will pursue a quantitative or mixed methods approach that incorporates statistical analyses, in addition to the methods described above [43,44]. Quantitative analysis of large-scale qualitative data may be tempting because of the volume of data and the need to efficiently summarize responses. In addition, the use of spreadsheets for coding facilitates frequency counts and other descriptive methods that are less common in many approaches to qualitative research [35,38].

Several considerations should be made when performing quantitative analysis of qualitative data. First, the goal of qualitative research is typically not to generalize to a population. Decisions in the study design, including sampling, data collection, and data analysis, will likely impose limitations for a quantitative analysis. SMS text messaging responses in MyVoice are limited in that follow-up and clarifying questions cannot be asked. Therefore, analysis is restricted to the words participants choose to share. When questions are framed as very broad and open-ended, participants may only share some of their perspectives. Researchers cannot be sure that participants would not have endorsed a theme, just that they did not share it in their written responses, which creates potential missing data. As a result, quantitative analyses should align with the scope of the research question and poll question.

For example, Schuiteman et al [44] investigated self-management of health care among youth in a mixed methods study. The four MyVoice poll questions were:

1. Who schedules your doctor's appointments? Why?
2. Who usually goes with you to your doctor's appointments? Why?
3. Who usually picks up your prescription medicine? Why?
4. Do you wish you were more or less involved in your health care? How?

The primary outcome was whether respondents desired to be more, less, or equally involved in their health care (ie, responses to the fourth question), and "independence scores" were calculated to assess variation in this outcome by summing participants' levels of independence in the first 3 questions (ie, through identification of self or help from others). Differences in the primary outcome were compared by demographic characteristics using chi-square tests. In this example, questions were worded in a way that supported coding of responses into categorical variables. Qualitative analyses explored the "why" and "how" narrative responses provided across questions, and resulting themes were compared between participants with high independence and low independence scores.

Second, careful consideration should also be given to what differences by demographics represent, particularly demographic characteristics that are socially constructed [45]. Analyses should have sufficient justification in the literature or existing theory, particularly in cases where testing for differences by demographic characteristics is desired. For example, one could expect, based on the literature, that nonbinary and transgender youth have more negative experiences with health care providers than cisgender youth [46]. In the MyVoice study of youth perspectives on primary care use and access, the research team conducted exploratory quantitative analyses to compare endorsement of each theme by different demographic groups using chi-square (for gender, race, and income level) and *t* tests (1-tailed; for age). In the study sample, more than one-third (284/789) reported previous bad experiences in a healthcare setting. Among these participants, 47% of nonbinary or transgender participants and 44% of cisgender women reported negative experiences with health care providers compared to 24% of men, which was statistically significant ($P<.001$) [42].

Discussion

Although technology has enabled the data collection and analysis of larger volumes of text data, there has been limited practical guidance on how researchers can manage large databases and rigorously analyze large datasets using qualitative methods [3,5]. In this tutorial, we presented steps and tools we have used for team-based analysis of large-scale text data collected through a nationwide SMS text messaging poll of youth. These steps have been used by teams that include youth and trainees of all levels, supporting the development of research interest, experiences, and skills through meaningful and authentic participation in research. Novice researchers can apply the methods described here using familiar tools, like spreadsheets, to reduce the time and resources spent using other qualitative data software tools. In addition to open-ended survey responses, the data analysis methods described can be applied by other digital and mobile health researchers conducting qualitative studies with substantial samples of short text data.

This tutorial describes the application of a qualitative analysis approach to short text responses, like those in SMS text messaging, open-ended survey questions, social media posts, and more. SMS text messaging polls allow youth to share their perspectives using an easy, familiar, and accessible tool, which may support youth to feel more comfortable engaging in research [47,48]. However, responses can vary in length, detail, and quality. Unlike some other qualitative data collection methods, like semistructured interviews, MyVoice SMS text message polls are limited to a predetermined set of questions. Teams are unable to ask follow-up questions or probe for more detail, which likely limits some of the richness of the data and therefore understanding of participant perspectives and experiences [49]. Indeed, Tran et al [50] found that significantly more responses were needed in open-ended survey responses

XSL•FO
RenderX

to reach the point of data sufficiency than is commonly recommended in qualitative analysis of interviews or focus groups. As a result, studies using short text responses can require larger datasets that can apply the team-based approach detailed in this tutorial. Further research is needed to illustrate analysis methods for large-scale qualitative research with longer text responses, such as interviews or observations.

Our initial study protocol proposed the use of natural language processing to support analysis of SMS text message responses. We used natural language processing to analyze several datasets and to try to streamline research processes for novice researchers [4,18]. However, time and resource limitations (eg, sufficient expertise among team members and updating algorithms for the context of each unique dataset) and concerns about the utility and validity of outputs restricted our regular use of this approach to date. Advancements in AI, including large language models and tools within qualitative data analysis software, will likely improve the efficiency and validity of AI-assisted qualitative analysis over time.

Our team includes youth as active team members in the research process to enhance the validity and relevance of the research [29,51,52]. We have found that line-by-line coding and pattern identification are processes and tools that are familiar to youth and similar to strategies taught in classrooms, enabling youth researchers to quickly engage in meaningful research opportunities. Through the MYHealth research training program, over 150 high school students and recent high school graduates have successfully used the methods in this tutorial to understand youth perspectives on future planning, social media, academic expectations, health education, and health care access [25,53]. In addition, novice researchers and researchers new to qualitative methods have successfully used our team-based approach to collaboratively and efficiently analyze large text datasets on a variety of health and health policy topics.

Despite the recent proliferation of AI tools, manual analysis procedures are still preferred in many qualitative and mixed methods studies to ensure understanding of the nuance and context available through qualitative data and limiting the bias introduced by AI that can influence findings [16,18]. This tutorial has provided an example of using a team-based qualitative analysis approach to analyze a large corpus of data collected through SMS text messaging polls of youth with commonly available tools. This tutorial adds to the limited research describing the step-by-step application of qualitative analysis techniques to large-scale qualitative data, particularly short text responses. Our approach can be used by research teams seeking to analyze short response text data using qualitative methods.

Funding

Establishment of the MyVoice study was supported by the Michigan Institute for Clinical and Health Research, the University of Michigan MCubed program, and the University of Michigan Department of Family Medicine and continues to receive funding through contracts with collaborating teams. The University of Michigan Research Electronic Data Capture (REDCap) instance used to collect data throughout MyVoice is maintained by the Michigan Institute for Clinical & Health Research. Michigan Institute for Clinical & Health Research is supported by the National Center for Advancing Translational Sciences at the National Institutes of Health (grant number UM1TR004404). A portion of MD's time is supported by a career development award funded through the National Institute of Diabetes and Digestive and Kidney Diseases and the National Institutes of Health (grant number K01DK134766). The MYHealth program is funded by a Science Education Partnership Award through the National Institute of General Medical Sciences and the National Institutes of Health (grant number R25GM137361). The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

Data sharing is not applicable to this article as no data sets were generated or analyzed during this study.

Authors' Contributions

Conceptualization: MD (lead), SAC (supporting), MW (supporting), KS (supporting), TC (supporting)
Writing – original draft: MD (lead), SAC (supporting), MW (supporting)
Writing – review & editing: MD (lead), SAC (supporting), MW (supporting), KS (supporting), TC (supporting)

Conflicts of Interest

MD is an Associate Editor for *JMIR mHealth and uHealth*.

Multimedia Appendix 1
Recruitment, Screening and Enrollment Process.
[DOCX File , 22 KB - jmir_v28i1e72526_app1.docx]

References

XSL•FO
RenderX

1. Davidson E, Edwards R, Jamieson L, Weller S. Big data, qualitative style: a breadth-and-depth method for working with large amounts of secondary qualitative data. Qual Quant 2019;53(1):363-376 [FREE Full text] [doi: 10.1007/s11135-018-0757-y] [Medline: 30930496]
2. Nallamothu BK, Guetterman TC, Harrod M, Kellenberg JE, Lehrich JL, Kronick SL, et al. How do resuscitation teams at top-performing hospitals for in-hospital cardiac arrest succeed? A qualitative study. Circulation 2018;138(2):154-163 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.118.033674] [Medline: 29986959]
3. Winskell K, Singleton R, Sabben G. Enabling analysis of big, thick, long, and wide data: Data management for the analysis of a large longitudinal and cross-national narrative data set. Qual Health Res 2018;28(10):1629-1639 [FREE Full text] [doi: 10.1177/1049732318759658] [Medline: 29557295]
4. Chang T, DeJonckheere M, Vydiswaran VGV, Li J, Buis LR, Guetterman TC. Accelerating mixed methods research with natural language processing of big text data. Journal of Mixed Methods Research 2021;15(3):398-412. [doi: 10.1177/15586898211021196]
5. Robins CS, Eisen K. Strategies for the effective use of NVivo in a large-scale study: qualitative analysis and the repeal of don't ask, don't tell. Qualitative Inquiry 2017;23(10):768-778. [doi: 10.1177/1077800417731089]
6. Yang Y, Pankow J, Swan H, Willett J, Mitchell SG, Rudes DS, et al. Preparing for analysis: a practical guide for a critical step for procedural rigor in large-scale multisite qualitative research studies. Qual Quant 2018;52(2):815-828 [FREE Full text] [doi: 10.1007/s11135-017-0490-y] [Medline: 37904883]
7. Vydiswaran VGV, Romero DM, Zhao X, Yu D, Gomez-Lopez I, Lu JX, et al. Uncovering the relationship between food-related discussion on twitter and neighborhood characteristics. J Am Med Inform Assoc 2020;27(2):254-264 [FREE Full text] [doi: 10.1093/jamia/ocz181] [Medline: 31633756]
8. Liu T, Waselewski EA, Dougherty A, Lee JD, Hill NE, Marzolf BA, et al. Use of and medical decision-making in portal messages among patients with type 2 diabetes: mixed methods study. J Med Internet Res 2025;27:e79413 [FREE Full text] [doi: 10.2196/79413] [Medline: 40997326]
9. White DE, Oelke ND, Friesen S. Management of a large qualitative data set: establishing trustworthiness of the data. Int J Qual Methods 2012;11(3):244-258. [doi: 10.1177/160940691201100305]
10. St George SM, Harkness AR, Rodriguez-Diaz CE, Weinstein ER, Pavia V, Hamilton AB. Applying rapid qualitative analysis for health equity: lessons learned using "EARS" with latino communities. Int J Qual Methods 2023;22:10.1177/16094069231164938 [FREE Full text] [doi: 10.1177/16094069231164938] [Medline: 38463016]
11. Watkins DC. Rapid and rigorous qualitative data analysis: the "RADaR" technique for applied research. Int J Qual Methods 2017;16(1):160940691771213. [doi: 10.1177/1609406917712131]
12. Parkinson S, Eatough V, Holmes J, Stapley E, Midgley N. Framework analysis: a worked example of a study exploring young people's experiences of depression. Qual Res Psychol 2015;13(2):109-129. [doi: 10.1080/14780887.2015.1119228]
13. Saldana J. The Coding Manual for Qualitative Researchers. Thousand Oaks, CA: SAGE Publications; 2021.
14. Woods M, Paulus T, Atkins DP, Macklin R. Advancing qualitative research using qualitative data analysis software (QDAS)? Reviewing potential versus practice in published studies using ATLAS.ti and NVivo, 1994–2013. Soc Sci Comput Rev 2016;34(5):597-617. [doi: 10.1177/0894439315596311]
15. Ose SO. Using excel and word to structure qualitative data. J Appl Soc Sci 2016;10(2):147-162. [doi: 10.1177/1936724416664948]
16. Prescott MR, Yeager S, Ham L, Rivera Saldana CD, Serrano V, Narez J, et al. Comparing the efficacy and efficiency of human and generative AI: qualitative thematic analyses. JMIR AI 2024;3:e54482 [FREE Full text] [doi: 10.2196/54482] [Medline: 39094113]
17. Abram MD, Mancini KT, Parker RD. Methods to integrate natural language processing into qualitative research. Int J Qual Methods 2020;19:1609406920984608. [doi: 10.1177/1609406920984608]
18. Guetterman TC, Chang T, DeJonckheere M, Basu T, Scruggs E, Vydiswaran VGV. Augmenting qualitative text analysis with natural language processing: methodological study. J Med Internet Res 2018;20(6):e231 [FREE Full text] [doi: 10.2196/jmir.9702] [Medline: 29959110]
19. Hamilton L, Elliott D, Quick A, Smith S, Choplin V. Exploring the use of AI in qualitative analysis: a comparative study of guaranteed income data. Int J Qual Methods 2023;22:16094069231201504. [doi: 10.1177/16094069231201504]
20. Hitch D. Artificial intelligence augmented qualitative analysis: the way of the future? Qual Health Res 2024;34(7):595-606 [FREE Full text] [doi: 10.1177/10497323231217392] [Medline: 38064244]
21. Morgan DL. Exploring the use of artificial intelligence for qualitative data analysis: the case of ChatGPT. Int J Qual Methods 2023;22:16094069231211248. [doi: 10.1177/16094069231211248]
22. Jowsey T, Braun V, Clarke V, Lupton D, Fine M. We reject the use of generative artificial intelligence for reflexive qualitative research. Qualitative Inquiry 2025. [doi: 10.1177/10778004251401851]
23. Meyer DZ, Avery LM. Excel as a qualitative data analysis tool. Field Methods 2008;21(1):91-112. [doi: 10.1177/1525822x08323985]
24. Ndungo I, Balimuttajjo S, Akugizibwe E. Excel-based qualitative data analysis: methodological guidelines for systematic organization, iterative coding, and thematic visualization. Int J Educ Technol Sci 2025;5(3):271-288. [doi: 10.5281/zenodo.17007908]

XSL•FO
RenderX

25. DeJonckheere M, Nichols LP, Moniz MH, Sonneville KR, Vydiswaran VGV, Zhao X, et al. MyVoice national text message survey of youth aged 14 to 24 years: study protocol. JMIR Res Protoc 2017;6(12):e247 [FREE Full text] [doi: 10.2196/resprot.8502] [Medline: 29229587]
26. Aceves-Martins M, Aleman-Diaz AY, Giralt M, Solà R. Involving young people in health promotion, research and policy-making: practical recommendations. Int J Qual Health Care 2019;31(2):147-153. [doi: 10.1093/intqhc/mzy113] [Medline: 29788085]
27. Patton GC, Sawyer SM, Santelli JS, Ross DA, Afifi R, Allen NB, et al. Our future: a Lancet commission on adolescent health and wellbeing. Lancet 2016;387(10036):2423-2478 [FREE Full text] [doi: 10.1016/S0140-6736(16)00579-1] [Medline: 27174304]
28. Schelbe L, Chanmugam A, Moses T, Saltzburg S, Williams LR, Letendre J. Youth participation in qualitative research: challenges and possibilities. Qualitative Social Work 2014;14(4):504-521. [doi: 10.1177/1473325014556792]
29. Anyon Y, Bender K, Kennedy H, Dechants J. A systematic review of youth participatory action research (YPAR) in the United States: methodologies, youth outcomes, and future directions. Health Educ Behav 2018;45(6):865-878. [doi: 10.1177/1090198118769357] [Medline: 29749267]
30. Ballonoff Suleiman A, Ballard PJ, Hoyt LT, Ozer EJ. Applying a developmental lens to youth-led participatory action research: a critical examination and integration of existing evidence. Youth & Society 2021;53(1):26-53. [doi: 10.1177/0044118x19837871]
31. Kennedy H, DeChants J, Bender K, Anyon Y. More than data collectors: a systematic review of the environmental outcomes of youth inquiry approaches in the United States. Am J Community Psychol 2019;63(1-2):208-226. [doi: 10.1002/ajcp.12321] [Medline: 30843254]
32. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
33. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
34. Bree RT, Gallagher G. Using Microsoft Excel to code and thematically analyse qualitative data: a simple, cost-effective approach. AISHE-J: The All Ireland Journal of Teaching and Learning in Higher Education 2016;8. [doi: 10.62707/aishej.v8i2.281]
35. Rosen RK, Gainey M, Nasrin S, Garbern SC, Lantini R, Elshabassi N, et al. Use of framework matrix and thematic coding methods in qualitative analysis for mHealth: NIRUDAK study data. Int J Qual Methods 2023;22:10.1177/16094069231184123 [FREE Full text] [doi: 10.1177/16094069231184123] [Medline: 38817641]
36. Reyes V, Bogumil E, Welch LE. The living codebook: documenting the process of qualitative data analysis. Sociol Methods Res 2021;53(1):89-120. [doi: 10.1177/0049124120986185]
37. Charmaz K. Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis. Thousand Oaks, CA: SAGE Publications; 2006.
38. MacQueen KM, McLellan E, Kay K, Milstein B. Codebook development for team-based qualitative analysis. CAM Journal 1998;10(2):31-36. [doi: 10.1177/1525822X980100020301]
39. Waselewski ME, Waselewski EA, Wasvary M, Wood G, Pratt K, Chang T, et al. Perspectives on telemedicine from a national study of youth in the United States. Telemed J E Health 2022;28(4):575-582. [doi: 10.1089/tmj.2021.0153] [Medline: 34415810]
40. Chuisano SA, Rafferty J, Allen A, Chang T, Diemer M, Harris K, et al. Increasing representation and diversity in health research: a protocol of the MYHealth research training program for high school students. PLoS One 2023;18(9):e0281790 [FREE Full text] [doi: 10.1371/journal.pone.0281790] [Medline: 37768968]
41. Schmid JC, Rose KL, Hadler NL, Amaro X, Frank A, Wilkie E, et al. Content analysis of the impact of COVID-19 on weight and shape control behaviors and social media content of U.S. adolescents and young adults. Eat Behav 2022;45:101635 [FREE Full text] [doi: 10.1016/j.eatbeh.2022.101635] [Medline: 35567879]
42. DeJonckheere M, Zhao A, Lane J, Plegue MA, Ranalli L, Wagner E, et al. Results of a national text message poll of youth: perspectives on primary care. J Adolesc Health 2020;67(6):814-820. [doi: 10.1016/j.jadohealth.2020.04.017] [Medline: 32620346]
43. DeJonckheere M, Waselewski M, Amaro X, Frank A, Chua KP. Views on COVID-19 and use of face coverings among U.S. youth. J Adolesc Health 2021;68(5):873-881 [FREE Full text] [doi: 10.1016/j.jadohealth.2021.02.015] [Medline: 33896552]
44. Schuiteman S, Chua KP, Plegue MA, Ilyas O, Chang T. Self-management of health care among youth: implications for policies on transitions of care. J Adolesc Health 2020;66(5):616-622 [FREE Full text] [doi: 10.1016/j.jadohealth.2020.01.009] [Medline: 32113903]
45. Lett E, Asabor E, Beltrán S, Cannon AM, Arah OA. Conceptualizing, contextualizing, and operationalizing race in quantitative health sciences research. Ann Fam Med 2022;20(2):157-163 [FREE Full text] [doi: 10.1370/afm.2792] [Medline: 35045967]

XSL•FO
RenderX

46. Goldenberg T, Gamarel KE, Reisner SL, Jadwin-Cakmak L, Harper GW. Gender affirmation as a source of resilience for addressing stigmatizing healthcare experiences of transgender youth of color. Ann Behav Med 2021;55(12):1168-1183 [FREE Full text] [doi: 10.1093/abm/kaab011] [Medline: 33761531]
47. French RS, McCarthy O, Baraitser P, Wellings K, Bailey JV, Free C. Young people's views and experiences of a mobile phone texting intervention to promote safer sex behavior. JMIR Mhealth Uhealth 2016;4(2):e26 [FREE Full text] [doi: 10.2196/mhealth.4302] [Medline: 27083784]
48. Perry RCW, Kayekjian KC, Braun RA, Cantu M, Sheoran B, Chung PJ. Adolescents' perspectives on the use of a text messaging service for preventive sexual health promotion. J Adolesc Health 2012;51(3):220-225. [doi: 10.1016/j.jadohealth.2011.11.012] [Medline: 22921131]
49. Robinson OC. Probing in qualitative research interviews: theory and practice. Qual Res Psychol 2023;20(3):382-397. [doi: 10.1080/14780887.2023.2238625]
50. Tran VT, Porcher R, Falissard B, Ravaud P. Point of data saturation was assessed using resampling methods in a survey with open-ended questions. J Clin Epidemiol 2016;80:88-96. [doi: 10.1016/j.jclinepi.2016.07.014] [Medline: 27492788]
51. Balazs CL, Morello-Frosch R. The three R's: How community based participatory research strengthens the rigor, relevance and reach of science. Environ Justice 2013;6(1):10.1089/env.2012.0017 [FREE Full text] [doi: 10.1089/env.2012.0017] [Medline: 24260590]
52. Malorni A, Lea CH, Richards-Schuster K, Spencer MS. Child Youth Serv Rev 2022;136:106399 [FREE Full text] [doi: 10.1016/j.childyouth.2022.106399]
53. Durden A, Lucien J, Chuisano SA, Palazzolo B, Ahmed Z, Ceaser N, et al. Youth perspectives of school-based health education: findings from a US youth-led nationwide qualitative survey. Health Educ J 2025:10.1177/00178969251376859. [doi: 10.1177/00178969251376859] [Medline: 41393786]

Abbreviations

AI: artificial intelligence
REDCap: Research Electronic Data Capture

Edited by T de Azevedo Cardoso, A Stone; submitted 11.Feb.2025; peer-reviewed by D Levine, J Cornelius; comments to author 01.Aug.2025; revised version received 04.Feb.2026; accepted 10.Feb.2026; published 27.Feb.2026.

Please cite as:
DeJonckheere M, Chuisano SA, Waselewski M, Sonneville K, Chang T
Team-Based Analysis of Large-Scale Qualitative Data: Tutorial Using a Nationwide SMS Text Messaging Poll of Youth
J Med Internet Res 2026;28:e72526
URL: https://www.jmir.org/2026/1/e72526
doi:10.2196/72526
PMID:

©Melissa DeJonckheere, Samantha A Chuisano, Marika Waselewski, Kendrin Sonneville, Tammy Chang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Translating Telehealth Communication Research Into Patient-Centered, Implementable Practice

Rachel Pittmann[1*], PhD, CCC-SLP; Paula D Koppel[2*], NBC-HWC, AHN-BC, GNP-BC, RNC, PhD; David Barrett[3], RN, PhD

[1]Center for Interprofessional Education and Practice, MGH Institute of Health Professions, 36 1st Avenue, Charlestown, MA, United States

[2]School of Nursing, Duke University, Durham, NC, United States

[3]Department of Health Sciences, University of York, Heslington, York, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Rachel Pittmann, PhD, CCC-SLP
Center for Interprofessional Education and Practice, MGH Institute of Health Professions, 36 1st Avenue, Charlestown, MA, United States

Related Article:

https://www.jmir.org/2026/1/e79399

Abstract

Understanding both patient and clinician perspectives on communication challenges in virtual primary care consultations is important to ensure safe and effective care. This commentary reviews the important work of Alboksmaty and colleagues, highlighting their contributions and noting some limitations. To further support translation of clinical practice–based recommendations from this original research, this commentary relates the study findings to patient-centered communication frameworks and evidence-based virtual communication strategies. This commentary also extends the helpful mitigation strategies offered in the original research, demonstrating how they apply to the Expert Recommendations for Implementing Change (ERIC) implementation framework to facilitate organization-based change. The need for additional research differentiating phone-based and videoconferencing-based primary care consultation communication strategies is highlighted.

(J Med Internet Res 2026;28:e93690) doi:10.2196/93690

KEYWORDS

telemedicine; telehealth; patient-centered communication; videoconferencing; phone-based virtual care

Introduction

We congratulate Alboksmaty and colleagues [1] on their interesting study analyzing the communication challenges that occur during virtual consultations and thank them for generating mitigation strategies to address these challenges. Using 2 frameworks, the Shannon Weaver communication model (SWCM) and the capability, opportunity, motivation, and behavior (COM-B) model, to support their analyses, the researchers analyzed the data they collected during focus groups with UK-based general practitioners and patients to understand the communication barriers and gaps in virtual consultations. The SWCM and COM-B provided an interesting perspective when applied to telehealth communication. After reviewing this study closely, we offer some guidance for viewing and implementing the findings in practice.

Disaggregate Phone-Based and Videoconferencing-Based Virtual Consultation Results

In Alboksmaty and colleagues' study [1], phone-based and videoconferencing-based consultations were considered together under the umbrella heading of "virtual" consultations. Our view would be that treating virtual consultations as a homogeneous entity and combining these two modalities into one analysis limits the ability to fully interpret the communication challenges. For example, the first theme in the findings highlights the patients' and clinicians' challenges in interpreting emotional and clinical cues, resulting in feelings of disconnection related primarily to phone-based virtual care. Similarly, the second theme of compromised clarity when discussing complex or sensitive issues was also more relevant to phone-based virtual care, as was the third theme, relating to challenges in trust and confidence centered around the absence of nonverbal interactions. We propose that separately analyzing phone-based

XSL•FO
RenderX

and videoconferencing-based virtual care could provide a clearer understanding of the unique strengths and challenges of each modality. Such an approach might have led to different interpretations of the fifth and sixth themes—aligning virtual modalities with clinical needs and balancing contextual expectations with patient preferences. Disaggregating the data into phone-based and videoconferencing-based modalities could better support shared decision-making between clinicians and patients by helping to determine which option is most appropriate for a given clinical scenario. This approach aligns with recent calls for researchers to directly compare the use of phones and videoconferencing in primary care to inform a structured process for modality selection [2].

Use Patient-Centered Communication Models and Principles in Telehealth Research

We would have found it helpful to understand the rationale for the researchers' selection of the SWCM to support their analysis. The SWCM was originally a model developed in the mathematical theory of communications and, though it has been used in health care research and education, we do not believe it enables a holistic analysis of the complex nature of human interactions between clinicians and patients. Using the Roter interaction analysis system, often leveraged to measure verbal and nonverbal communication patterns in health care, including telehealth [3], or the patient-centered communication (PCC) framework, associated with quality in health care [4], may have supported more meaningful interpretation of the data.

We offer an expansion of Alboksmaty et al's [1] findings by aligning them with PCC principles whereby clinicians empathically engage with patients' needs and values and share in decision-making. PCC, as described by Moser et al [5], is organized around 7 core functions, which include ensuring adequate time and attention for patients to ask questions and address their emotional needs (functions 1, 2, and 6), sharing decision-making (function 3), enabling patients to understand how to take care of their health (function 4), providing clinical explanations that are understandable (function 5), and supporting patients in managing feelings of uncertainty (function 7). PCC has been shown to enhance patient trust and medication adherence as well as overall patient satisfaction [6].

In Table 1, we align Alboksmaty et al's [1] study findings with the PCC core functions and evidence-based, patient-centered virtual communication tools. As can be seen in the table, training and skill development for virtual consultations relate to 4 of the 6 study themes, reflecting the importance of building capacity (ie, knowledge and skill) for clinicians, patients, and organizations. Clinicians should consider that phone-based and videoconferencing-based consultations require distinct communication adaptations, and that specific needs and preferences of patients may also require accommodation. We encourage clinicians to consider their role in the development of the patient-clinician relationship and reference evidence-based telehealth PCC strategies [7-9] in the table to ensure the virtual environment supports patients' preferences and needs.

XSL•FO
RenderX

Table . Study findings in relation to patient-centered communication (PCC) challenges, mitigation strategies, and evidence-based virtual communication tools.

Communication challenges and associated mitigation strategies identified by Alboksmaty et al [1]	PCC functions at risk if mitigation strategies are not used[a]	Evidence-based, patient-centered virtual communication tools [7-9]
• Theme 1 challenges: missed cues and misinterpretations (sender-encoder) • Theme 1 strategies: training and skill development for virtual consultations (capability)	• Question (function 1) • Decisions (function 3) • Understood (function 4) • Explain (function 5)	• Create a safe, private, and comfortable environment for the visit in which the patient and clinician can adequately see and hear one another. • Maintain attentiveness and presence. • Personalize communication based on patient's information needs (literacy, preferred learning style, desired level of detail, sensory challenges). • Use summarizing statements and open-ended questions. • Avoid medical jargon. • Check for understanding (teach-back and follow-up questions). • Ensure adequate time for discussion and questions. • Demonstrate empathy using verbal and, whenever possible, nonverbal means.
• Theme 2 challenge: compromises related to lack of clarity in virtual communication (message) • Theme 2 strategy: addressing barriers to effective communication (opportunity)	• Question (function 1) • Decisions (function 3) • Understood (function 4) • Explain (function 5)	• Maintain attentiveness and presence. • Personalize communication based on patient's information needs (literacy, preferred learning style, desired level of detail, sensory challenges). • Use summarizing statements and open-ended questions. • Avoid medical jargon. • Check for understanding (teach-back and follow-up questions). • Ensure adequate time for discussion and questions. • Demonstrate and clearly state what is being asked of patients during assessment.
• Theme 3 challenges: technical issues, poor connectivity, and accessibility issues (channel) • Theme 3 strategies: training and skill development for virtual consultations (capability); triage and risk mitigation (opportunity); addressing barriers to effective communication (opportunity); training and resources to support patient preparedness (opportunity and motivation)	• Understood (function 4) • Time (function 6)	• Have a working knowledge of telehealth technologies and problem-solving skills. • Demonstrate a willingness to adapt to changing situations with ease and inform patients of a backup plan for technology failures. • Demonstrate respect for the patient challenges that are associated with virtual care. • Ensure patients know how to proceed when technology issues arise.
• Theme 4 challenge: lack of trust in discussing complex needs (receiver-decoder feedback) • Theme 4 strategies: patient and clinician preferences in consultation modalities (motivation); training and skill development for virtual consultations (capability)	• Question (function 1) • Attention (function 2) • Decision (function 3) • Understood (function 4) • Explain (function 5) • Time (function 6) • Feelings (function 7)	• Develop a joint agenda and an agreement on roles and expectations. • Maintain attentiveness and presence. • Take time to ensure a safe and comfortable environment (ensure privacy, create an interpersonal connection with the patient). • Demonstrate empathy using verbal and, whenever possible, nonverbal means.

XSL•FO
RenderX

Communication challenges and associated mitigation strategies identified by Alboksmaty et al [1]	PCC functions at risk if mitigation strategies are not used[a]	Evidence-based, patient-centered virtual communication tools [7-9]
• Theme 5 challenge: feeling less empathy and emotional engagement from clinician (receiver-decoder feedback) • Theme 5 strategies: training and skill development for virtual consultations (capability)	• Attention (function 2) • Understood (function 4) • Time (function 6) • Feelings (function 7)	• Maintain attentiveness and presence. • Intentionally create an interpersonal connection at the beginning of each visit. • Ensure the technology and the virtual space facilitate interpersonal connection (ie, eye contact, minimal distractions, adequate lighting). • Magnify use of active listening techniques (verbal and nonverbal). • Heighten self-awareness to ensure behaviors reflect empathy and care. • Demonstrate empathy using verbal and, whenever possible, nonverbal means.
• Theme 6 challenge: social norms, cultural factors, and age (context)	• Attention (function 2) • Decision (function 3) • Understood (function 4) • Explain (function 5) • Time (function 6) • Feelings (function 7)	• Demonstrate respect for the patient challenges associated with virtual care. • Demonstrate willingness to adapt to changing situations with ease and assist patients with technology. • Personalize communication approach as noted above to accommodate patient needs and/or preferences.
• Theme 7 challenges: lack of patient autonomy and inclusivity leading to decreased engagement and willingness to participate in virtual care	• Question (function 1) • Attention (function 2) • Decision (function 3) • Understood (function 4) • Explain (function 5) • Time (function 6) • Feelings (function 7)	• Demonstrate respect for the patient challenges associated with virtual care. • Demonstrate a willingness to adapt to changing situations with ease and assist patients with technology. • Personalize communication approach as noted above to accommodate patient needs and/or preferences.

[a]Full description of the PCC core functions [5]: (1) question—adequate time for patient to ask all health-related questions; (2) attention—adequate attention to patients' feelings and emotions; (3) decision—shared decision-making; (4) understood—ensuring patient understands how to manage their health; (5) explain—ensuring patient understands what clinician explains; (6) time— ensuring there is adequate time spent with the patient; (7) feelings—adequate time to manage patients' feelings of uncertainty about their health.

Use Implementation Frameworks to Facilitate Transfer of Research Findings to Clinical Practice

To further facilitate the translation of Alboksmaty et al's [1] findings and the associated evidence-based, patient-centered virtual communication tools into clinical practice, we propose using strategies from the Expert Recommendations for Implementing Change (ERIC) implementation framework [10]. Given the challenges, mitigation strategies, and evidence-based tools shown above, we propose using several ERIC strategies at both the individual clinician level and the organizational level for implementation of mitigation strategies. Specifically, creating a community of telehealth practitioners composed of champion clinicians, who will develop educational materials and outreach initiatives and who will model and simulate change prior to the implementation phase, supports the effective implementation of strategies into practice. Additionally, inviting patients as partners in the community of practice enables their voices to be central to the patient-centered change implementation.

Alboksmaty et al [1] analyzed telehealth communication challenges identified by clinicians and patients using unique frameworks. This commentary found it was feasible to apply their findings to the PCC and ERIC models, providing additional support to students and clinicians adapting their communication to telehealth. Future research focused on better differentiating communication strategies for phone-based and videoconferencing-based consultations is necessary to maximize the quality of virtual care.

Funding

The authors declare that no financial support was received for this work.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Alboksmaty A, Lunova T, Darzi A, Neves AL. Communication challenges and mitigation strategies in primary care virtual consultations: qualitative study. J Med Internet Res 2026 Jan 20;28(1):e79399. [doi: 10.2196/79399] [Medline: 41557949]
2. Byambasuren O, Greenwood H, Bakhit M, et al. Comparison of telephone and video telehealth consultations: systematic review. J Med Internet Res 2023 Nov 17;25(1):e49942. [doi: 10.2196/49942] [Medline: 37976100]
3. Agha Z, Roter DL, Schapira RM. An evaluation of patient-physician communication style during telemedicine consultations. J Med Internet Res 2009 Sep 30;11(3):e36. [doi: 10.2196/jmir.1193] [Medline: 19793720]
4. Institute of Medicine (US) Committee on Quality of Health Care in America. Crossing the Quality Chasm: A New Health System for the 21st Century: National Academies Press (US); 2001. URL: http://www.ncbi.nlm.nih.gov/books/NBK222274 [accessed 2026-03-10]
5. Moser RP, Trivedi N, Murray A, Jensen RE, Willis G, Blake KD. Patient-centered communication (PCC) scale: psychometric analysis and validation of a health survey measure. PLOS ONE 2022;17(12):e0279725. [doi: 10.1371/journal.pone.0279725] [Medline: 36584146]
6. Street RL, Makoul G, Arora NK, Epstein RM. How does communication heal? Pathways linking clinician-patient communication to health outcomes. Patient Educ Couns 2009 Mar;74(3):295-301. [doi: 10.1016/j.pec.2008.11.015] [Medline: 19150199]
7. Pittmann R, Danaher-Garcia N, Adair White BA, Thompson A. Development and validation of the Telehealth Etiquette Competency Checklist: a Delphi study. J Telemed Telecare 2025 Oct;31(9):1308-1316. [doi: 10.1177/1357633X241279494] [Medline: 39311041]
8. Koppel PD, De Gagne JC, Webb M, et al. Guidelines for rapport-building in telehealth videoconferencing: interprofessional e-Delphi study. JMIR Med Educ 2025 Aug 7;11:e76260. [doi: 10.2196/76260] [Medline: 40773683]
9. Gustin TS, Kott K, Rutledge C. Telehealth etiquette training: a guideline for preparing interprofessional teams for successful encounters. Nurse Educ 2020;45(2):88-92. [doi: 10.1097/NNE.0000000000000680] [Medline: 31022072]
10. Powell BJ, Waltz TJ, Chinman MJ, et al. A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. Implement Sci 2015 Feb 12;10:21. [doi: 10.1186/s13012-015-0209-1] [Medline: 25889199]

Abbreviations

COM-B: capability, opportunity, motivation, and behavior
ERIC: Expert Recommendations for Implementing Change
PCC: patient-centered communication
SWCM: Shannon Weaver communication model

Edited by T Leung; submitted 17.Feb.2026; this is a non–peer-reviewed article;accepted 25.Feb.2026; published 19.Mar.2026.

Please cite as:
Pittmann R, Koppel PD, Barrett D
Translating Telehealth Communication Research Into Patient-Centered, Implementable Practice
J Med Internet Res 2026;28:e93690
URL: https://www.jmir.org/2026/1/e93690
doi:10.2196/93690

© Rachel Pittmann, Paula D Koppel, David Barrett. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Beyond GPT-4: The Rapidly Evolving Potential of Large Language Models for Clinical Guideline Improvement

Scott D Nelson, PharmD, MS; Adam Wright, PhD

Department of Biomedical Informatics, School of Medicine, Vanderbilt University Medical Center, 3401 West End Ave, Nashville, TN, United States

Corresponding Author:
Scott D Nelson, PharmD, MS
Department of Biomedical Informatics, School of Medicine, Vanderbilt University Medical Center, 3401 West End Ave, Nashville, TN, United States

Related Article:

https://www.jmir.org/2026/1/e81915

Abstract

This commentary reviews the study by Jones et al, which evaluated whether GPT-4 could improve the readability of injectable medication guidelines while preserving important safety information. The study found that GPT-4 produced modest readability gains comparable to manual revision, but also introduced omissions and meaning changes in a minority of sections. These findings highlight both the potential and limitations of early large language models (LLMs) in clinical contexts. However, this study reflects the capabilities of a specific model in a rapidly evolving domain. Since the release of GPT-4, advances in multistep reasoning, model-critique workflows, and structured validation have substantially improved the ability of newer systems to detect omissions, maintain factual fidelity, and support controlled editing. As a result, some documented limitations may stem from the constraints of a single-model, single-pass workflow rather than intrinsic flaws in LLM-assisted guideline revision. This commentary highlights the need for evaluation frameworks that can keep pace with LLM progress and emphasizes that clinical oversight and user-centered testing remain essential. Updated research using contemporary models is needed to determine how emerging architectures can more safely support clarity, consistency, and maintenance of clinical guidelines.

(J Med Internet Res 2026;28:e95004) doi:10.2196/95004

KEYWORDS

artificial intelligence; clinical guidelines; large language model; patient safety; readability; clinical decision support

Introduction

In their recent study, Jones et al [1] evaluated a GPT-4–based pipeline for improving the readability of 20 guidelines from the United Kingdom's National Health Service Injectable Medicines Guide (IMG). The authors found that GPT-4 produced modest but statistically significant readability improvements, and expert pharmacist reviewers rated the revised versions as easier to understand for 26 of 60 (43%) of the ratings. The gains in readability were comparable to those achieved by manual revisions by guideline authors; however, the greatest readability improvements were for two guidelines (aminophylline and voriconazole) using iterative user testing from a previous study.

Notably, using the large language model (LLM) was not without risk. At least one pharmacist reviewer identified omissions in 30 of 153 subsections (20%), additions in 7 subsections (5%), and changes in meaning in 18 subsections (12%). Eight subsections had omissions identified by all 3 reviewers, but no additions or changes in meaning were unanimously flagged. Overall, 65% of all identified issues were flagged by only a single reviewer. The authors concluded that GPT-4 could help augment, rather than replace, manual expert review and user-centered testing to improve guideline readability.

The Challenge of Evaluating a Moving Target

Interpreting these findings requires an appreciation for how rapidly LLM technology has advanced since the release of GPT-4 in March 2023, an extraordinarily long time ago in the current field of artificial intelligence advancements. By the time the study was published, newer models had already introduced major improvements in error checking, multistep reasoning, and structured critique workflows [2,3]. This creates a temporal mismatch where clinical research and guideline development cycles operate on the scale of *years*, while LLM development cycles operate on the scale of *months* and could continue to accelerate [4]. Thus, this study should be viewed as an evaluation of a specific model at a fixed point in time, not a judgment on the overall potential of LLMs in guideline development and review workflows. This is not a criticism of the study itself, as the authors designed a careful, well-controlled

XSL•FO
RenderX

evaluation and their findings are valuable precisely because they document specific failure modes that future systems must address. Rather, it is a call to develop more agile evaluation frameworks that can keep pace with technological change, so that evidence generation does not perpetually lag behind the tools available for deployment.

For example, the GPT-4 pipeline relied on a single model for both the editing and quality assurance steps. This architecture creates an inherent tension, since simplification commonly results in removing content, so omissions could be somewhat expected. Newer multiagent systems separate generation from critique, allowing an editor model to propose revisions while a critic model checks for completeness, consistency, and factual accuracy [5]. Structured reasoning frameworks, such as tree-of-thought and self-consistency, enable models to justify edits and cross-check them before finalizing. Skill-based architectures allow explicit function calls to validate medication names, units, values, and section completeness, replacing soft prompts with enforceable programmatic safeguards, while dynamic prompt optimization can iteratively refine instructions to prevent prompt-induced errors. These advances do not eliminate the need for clinician oversight, but they offer more robust mechanisms for preserving informational fidelity while improving readability.

Contextualizing the Use Case

The IMG guidelines represent a relatively favorable use case for LLM revision: procedural, deterministic instructions with clear ground truth. This contrasts with other clinical practice guidelines, such as those for disease management, which must synthesize heterogeneous evidence, navigate gaps, and rely on expert consensus across nonrepresentative study populations, posing far greater challenges. In this example, the study evaluated GPT-4's performance on the IMG, which is a nationally curated, professionally edited resource. Improving readability on an already well-crafted guideline is inherently challenging, yet the model sill produced modest readability improvements. In practice, LLMs may offer even greater value for locally developed clinical guidelines and documents, which are often produced under time pressure with less editorial rigor. Enhancing the clarity and consistency of these documents could improve staff comprehension and ultimately lower the risk of downstream errors. This study tested the LLM against the hardest version of the problem, improving something already well-crafted, while the real-world opportunity may lie in lifting up the documents that need it most.

The Continued Need for User-Centered Testing

User testing remains the guideline improvement technique with the strongest evidence base, and the authors' own data confirm this. Future studies should also include the intended end users. For example, the IMG is primarily used by nurses, who may prioritize quick scanability and visual hierarchy over the pharmacological completeness that pharmacist reviewers would naturally emphasize. The interaction between content accuracy and practical usability can only be fully understood by the people who use these documents at the point of care.

Beyond Readability

Improving readability is only one part of making guidelines more usable. Simplifying text naturally risks omitting information, while providing excessive detail poses its own risks: dense guidelines can cause clinicians to overlook or misinterpret critical information. In the study, 65% of errors were identified by only a single pharmacist, suggesting that many issues were subtle and difficult to detect. The real question is how we can preserve and present essential information in the clearest, most usable form. Newer multimodal models offer approaches beyond text alone. They can generate diagrams, flowcharts, and annotated step-by-step visuals, which may communicate procedural information, such as reconstitution or infusion setup, more effectively than narrative text [6]. These multimodal formats can help reduce cognitive load and help clinicians understand complex instructions more intuitively.

Furthermore, LLMs have broader potential in the guideline ecosystem. They could support translation into other languages, improving access and equity in multilingual care settings [7], though clinical translation would require rigorous verification [8]. LLMs may also enable just-in-time guidance by retrieving and tailoring the relevant portion of a guideline to a clinician's immediate question, which would often be more valuable than improving long documents clinicians may not have time to read. In addition, LLMs could assist with clinical decision support (CDS) maintenance, helping translate updated recommendations into structured CDS logic, or flag conflicts between new evidence and existing rules, reducing alert fatigue and easing the burden of keeping CDS systems current [9]. These applications warrant dedicated study using current-generation models.

Conclusion

Jones et al [1] provide a rigorous, timely evaluation of GPT-4's capabilities and limitations in revising medication guidelines. Their findings identify clear failure modes that future systems must overcome. As LLM architectures continue to advance, updated evaluations are essential to determine how well newer systems address the documented issues and how they can safely support clinicians, guideline authors, and health care organizations. None of these advances eliminate the need for clinician oversight or user-centered testing. The goal is to equip guideline authors and informatics teams with powerful tools for improving how clinical knowledge is communicated and delivered at the point of care. The evidence base must evolve alongside the technology.

XSL•FO
RenderX

Acknowledgments

The authors declare the use of generative AI (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: text generation, proofreading and editing, and summarizing text. The GAI tool used was Microsoft Copilot. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

The authors declared no financial support was received for this work.

Authors' Contributions

SDN conceptualized the study and was responsible for writing the original draft. SDN and AW contributed to writing, review, and editing of the manuscript.

Conflicts of Interest

SDN serves on the advisory board for Merative Micromedex and Baxter Healthcare. AW declares no conflicts of interest.

References

1. Jones MD, Torgbi M, Tayyar Madabushi H. Improving the understandability of clinical guidelines: development and evaluation of a GPT-4-based pipeline. J Med Internet Res 2026 Feb 23;28:e81915. [doi: 10.2196/81915] [Medline: 41730207]
2. Introducing GPT-5. OpenAI. 2025. URL: https://openai.com/index/introducing-gpt-5 [accessed 2026-04-06]
3. Stanciuc AM. OpenAI's GPT-54 sets new records on professional benchmarks. The Next Web. 2026. URL: https://thenextweb.com/news/openai-gpt-54-launch-computer-use-benchmarks [accessed 2026-04-06]
4. Aschenbrenner L. Situational Awareness: The Decade Ahead. 2024. URL: https://situational-awareness.ai [accessed 2026-04-06]
5. Yuan Y, Xie T. Reinforce LLM reasoning through multi-agent reflection. arXiv. Preprint posted online on 2025. [doi: 10.48550/arXiv.2506.08379]
6. Benito MD, Diana-Albelda C, García-Martín Á, Bescos J, Viñolo ME, SanMiguel JC. MIRAGE: retrieval and generation of multimodal images and texts for medical education international workshop on applications of medical AI. In: Wu S, Shabestari B, Xing L, editors. Applications of Medical Artificial Intelligence. AMAI 2025. Lecture Notes in Computer Science, Vol 16206: Springer; 2026. [doi: 10.1007/978-3-032-09569-5_11]
7. Pavithra RS. Bridging health literacy gaps in Indian languages: multilingual LLMs for clinical text simplification. In: Zhao W, D'Souza J, Eger S, et al, editors. Proceedings of The First Workshop on Human–LLM Collaboration for Ethical and Responsible Science Production: Association for Computational Linguistics; 2025. [doi: 10.18653/v1/2025.sciprodllm-1.1]
8. Schlicht IB, Sayin B, Zhao Z, Labonté FM, Barbera C, Viviani M, et al. Disparities in multilingual LLM-based healthcare Q&A. arXiv. Preprint posted online on 2025. [doi: 10.48550/arXiv.2510.17476]
9. Liu S, Wright AP, Patterson BL, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 2023 Jun 20;30(7):1237-1245. [doi: 10.1093/jamia/ocad072] [Medline: 37087108]

Abbreviations

CDS: clinical decision support
IMG: Injectable Medicines Guide
LLM: large language model

Edited by T Leung; submitted 09.Mar.2026; this is a non–peer-reviewed article;accepted 20.Mar.2026; published 10.Apr.2026.

Please cite as:
Nelson SD, Wright A
Beyond GPT-4: The Rapidly Evolving Potential of Large Language Models for Clinical Guideline Improvement
J Med Internet Res 2026;28:e95004
URL: https://www.jmir.org/2026/1/e95004
doi:10.2196/95004

© Scott D Nelson, Adam Wright. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License

XSL•FO
RenderX

(https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Why We Need Patients and Community at the Center of AI Health Communication Research

Julie Ayre[1], PhD; Lexuan Shao[2], MPhil; Adam G Dunn[2], PhD

[1]Sydney Health Literacy Lab, Sydney School of Public Health, Faculty of Medicine and Health, The University of Sydney, Edward Ford Building (A27), Sydney, New South Wales, Australia

[2]Sydney School of Public Health, Faculty of Medicine and Health, The University of Sydney, Sydney, New South Wales, Australia

Corresponding Author:
Julie Ayre, PhD
Sydney Health Literacy Lab, Sydney School of Public Health, Faculty of Medicine and Health, The University of Sydney, Edward Ford Building (A27), Sydney, New South Wales, Australia

Related Article:

https://www.jmir.org/2026/1/e81243

Abstract

Holderried and colleagues tested whether artificial intelligence (AI)–generated, patient-centered information can help people understand what they need to do after being discharged from a hospital. Participants demonstrated stronger comprehension when they viewed the simplified, patient-centered information rather than a standard letter. This work adds to the available early-phase evidence of AI supporting hospital discharge communication. To meaningfully progress this area of research, we now need to carefully consider how to enhance the design and evaluation of patient-facing AI health communication tools. In this commentary, we argue that equity and consumer engagement remain underrepresented in studies on patient-facing AI health communication tools and describe possible approaches to address this issue.

(J Med Internet Res 2026;28:e97577) doi:10.2196/97577

KEYWORDS

artificial intelligence; health communication; health literacy; patient education; discharge communication

Introduction

Artificial intelligence (AI) has great potential to support effective health communication in complex environments, such as hospital discharge [1]. The challenge is clear: hospitals are often unfamiliar places, and patients try to simultaneously manage stress, anxiety, and fatigue while learning about their medical situation and what they need to do next. Miscommunication at discharge can have serious consequences, including readmission to the hospital or even death, and these risks are greater for health priority groups, such as people with low health literacy [2]. Meta-analyses suggest that discharge communication often fails to meet patient needs, with only 47% of patients in emergency departments correctly recalling verbal advice about discharge and only 58% correctly recalling written advice [3]. If AI can meaningfully improve discharge communication, this scalable technology stands to benefit millions of people and the health systems that serve them.

Holderried et al [1] used a novel experimental design to test whether AI-generated, patient-centered information can help people understand what they need to do after leaving a hospital. Participants showed stronger comprehension when they viewed simplified, patient-centered information rather than a standard letter. The study contributes to a growing body of evidence suggesting that AI could meaningfully contribute to more effective communication during this important transition of care [4,5], and it provides a useful example of what a patient-centered evaluation might look like in practice. In this commentary, we discuss the importance of this recently published study and discuss how a stronger focus on equity and consumer engagement can help shift future research further along the pathway toward real-world implementation.

Start With Evidence-Based Health Literacy Practices

Health literacy is at the heart of equitable communication practices. However, to date, research on AI for patient communication has focused on evaluating accuracy and safety, with limited examples of health literacy evaluation. In the Holderried et al [1] study, patient-centered discharge letters generated by AI scored more favorably on readability than the original discharge letters [6]. Readability is a widely used outcome for estimating the complexity of health information [7]. Despite the improvement in AI-generated letters, readability was still at a level that most people find difficult to understand.

XSL•FO
RenderX

This finding is consistent with a recent review of 47 studies showing that although AI tools can significantly improve the readability of health information, prompts to date often fail to produce outputs that meet established readability targets [7]. This may reflect limitations of the AI models themselves and the need for more effective prompts.

Most plain language prompts for AI-generated health advice primarily rely on readability scores to benchmark performance [7]. This approach fails to recognize that readability is a narrow construct that should be supported by other complementary plain language outcomes [8]. Holderried et al [1] recognized this limitation and therefore also assessed the use of medical jargon and acronyms. Researchers could also consider other relevant linguistic features, such as passive voice, vocabulary size, lexical density, and textual cohesion [9]. Beyond automated measures, validated tools, such as the Patient Education Materials Assessment Tool, support a more holistic evaluation, including whether health advice is easy to act on [10]. Discharge letters may also need strategies that are more specific to their content. For example, the Universal Medication Schedule is an established format for medicine instructions that can improve medication adherence, particularly for older patients and people with lower health literacy [11].

Diverse Consumer Engagement Is Essential

Research stands to benefit from authentic partnerships with diverse patients, carers, and communities and from methodological approaches that embed these voices throughout the research process. For example, consumer engagement can help a research team understand what matters most to patients at discharge, how these needs might vary, and the implications for the design of AI-generated discharge letters and their subsequent evaluation.

In the Holderried et al [1] study, patient actors were trained to be “standardized patients.” Although this is encouraging, given that so few studies of this nature include any consumer or patient perspective [7], there is a clear need to more meaningfully involve consumers with relevant lived experience throughout the research process. This could include early discussions about preferred formats, such as interactive chat, video, or voice AI, and important discussions about when AI advice is not appropriate. Discussions about grounding AI chatbot responses in external knowledge sources (eg, via retrieval-augmented generation) are highly relevant, as these approaches afford greater interactivity and tailoring for patients but also limit the opportunity for human oversight, which poses a potential risk to safety. Lastly, as evidence of an AI tool’s benefits accumulates, studies must seek to evaluate the tool with real patients and in more realistic contexts.

To avoid exacerbating health inequities, researchers should also partner with consumers and stakeholders from priority groups who stand to benefit the most from AI discharge communication tools. These groups may include, for example, patients from culturally and linguistically diverse communities, First Nations communities, people living with severe mental illness or chronic health conditions, older adults, and people with intellectual disabilities [2].

We Need Frameworks to Guide Decisions About Balancing Risk and Opportunity

As Holderried et al [1] discuss, their AI-generated discharge summaries were not perfect, and five of the learning objectives from the original summaries could not be evaluated because they had been omitted in the patient-centered forms. Although research continues to evaluate the safety and accuracy of clinical AI tools, we also need frameworks to guide decisions about which tools to implement and how to feasibly and ethically do so, including appropriate evaluation methods. These decisions should involve diverse consumers and representative bodies to ensure that judgments about acceptable risk and appropriate use reflect patient needs and priorities. Discussions can be further supported by data reporting on the fairness and equity of these AI tools.

Conclusion

Holderried et al [1] provide new evidence showing the feasibility of an AI tool for improving discharge communication. This is one of few published examples where evaluation methods incorporate multiple health literacy assessments and explore learning outcomes. Now is the time to design studies that make full use of evidence-based health literacy practices and meaningfully involve diverse consumers throughout the research process. This is critical for ensuring that research meets patient needs and meaningfully improves the health care experience.

Funding

JA was supported by a National Health and Medical Research Council (NHMRC) Investigator Grant (APP 2017278). The NHMRC is not involved in any aspect of development of this paper.

Authors' Contributions

Writing – original draft: JA
Writing – review & editing: JA, AGD, LS

XSL•FO
RenderX

Conflicts of Interest

The Sydney Health Literacy Lab (SHeLL) Health Literacy Editor is a research tool owned by the University of Sydney. It is sublicensed to Health Literacy Solutions PTY Ltd to enable wider public use. JA is a codirector of Health Literacy Solutions Pty Ltd. No other conflicts of interest declared.

References

1. Holderried F, Sonanini A, Stegemann-Philipps C, et al. Impact of GPT-4-generated discharge letters on patients' medical comprehension: prospective crossover study. J Med Internet Res 2026 Feb 26;28:e81243. [doi: 10.2196/81243] [Medline: 41746691]
2. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med 2011 Jul 19;155(2):97-107. [doi: 10.7326/0003-4819-155-2-201107190-00005] [Medline: 21768583]
3. Hoek AE, Anker SCP, van Beeck EF, Burdorf A, Rood PPM, Haagsma JA. Patient discharge instructions in the emergency department and their effects on comprehension and recall of discharge instructions: a systematic review and meta-analysis. Ann Emerg Med 2020 Mar;75(3):435-444. [doi: 10.1016/j.annemergmed.2019.06.008] [Medline: 31439363]
4. Kumar A, Wang H, Muir KW, Mishra V, Engelhard M. A cross-sectional study of GPT-4–based plain language translation of clinical notes to improve patient comprehension of disease course and management. NEJM AI 2025 Jan 23;2(2). [doi: 10.1056/AIoa2400402]
5. Stanceski K, Zhong S, Zhang X, et al. The quality and safety of using generative AI to produce patient-centred discharge instructions. NPJ Digit Med 2024 Nov 20;7(1):329. [doi: 10.1038/s41746-024-01336-w] [Medline: 39567722]
6. Eisinger F, Holderried F, Mahling M, et al. What's going on with me and how can I better manage my health? The potential of GPT-4 to transform discharge letters into patient-centered letters to enhance patient safety: prospective, exploratory study. J Med Internet Res 2025 Jan 21;27:e67143. [doi: 10.2196/67143] [Medline: 39836954]
7. Ugas M, Huynh J, Lenarcik-Packham A, et al. The utility of artificial intelligence in plain language writing: a scoping review. Patient Educ Couns 2026 Apr;145:109483. [doi: 10.1016/j.pec.2026.109483] [Medline: 41564566]
8. Mac O, Muscat DM, Ayre J, McCaffery K. Fundamentally flawed or functional and feasible? The use of readability metrics in healthcare. Patient Educ Couns 2025 Jun;135:108741. [doi: 10.1016/j.pec.2025.108741] [Medline: 40081155]
9. Ayre J, Bonner C, Muscat DM, et al. Multiple automated health literacy assessments of written health information: development of the SHeLL (Sydney Health Literacy Lab) Health Literacy Editor v1. JMIR Form Res 2023 Feb 14;7:e40645. [doi: 10.2196/40645] [Medline: 36787164]
10. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014 Sep;96(3):395-403. [doi: 10.1016/j.pec.2014.05.027] [Medline: 24973195]
11. Davis TC, Federman AD, Bass PF 3rd, et al. Improving patient understanding of prescription drug label instructions. J Gen Intern Med 2009 Jan;24(1):57-62. [doi: 10.1007/s11606-008-0833-4] [Medline: 18979142]

Abbreviations

AI: artificial intelligence

Edited by S Law, T Leung; submitted 08.Apr.2026; this is a non–peer-reviewed article;accepted 29.Apr.2026; published 07.May.2026.

Please cite as:
Ayre J, Shao L, Dunn AG
Why We Need Patients and Community at the Center of AI Health Communication Research
J Med Internet Res 2026;28:e97577
URL: https://www.jmir.org/2026/1/e97577
doi:10.2196/97577

© Julie Ayre, Lexuan Shao, Adam G Dunn. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Rethinking Postdischarge Intervention Evaluation

Brigitte Fong Yeong Woo, BSc (Hons), MClinRes, PhD

Nursing Administration, Ng Teng Fong General Hospital, 1 Jurong East Street 21, Singapore, Singapore

Corresponding Author:
Brigitte Fong Yeong Woo, BSc (Hons), MClinRes, PhD
Nursing Administration, Ng Teng Fong General Hospital, 1 Jurong East Street 21, Singapore, Singapore

Related Article:

https://www.jmir.org/2026/1/e80529

Abstract

This commentary argues that for low-intensity postdischarge interventions, emergency department use may be a more sensitive and appropriate indicator of transitional care quality than readmission. It also positions nurse-led telephone follow-up as interpretive, equity-sensitive transitional care work that helps patients make discharge plans actionable in the home context while highlighting the value of accessible, scalable digital modalities such as telephone outreach.

(J Med Internet Res 2026;28:e98435) doi:10.2196/98435

KEYWORDS

postdischarge care; telephone calls; health service evaluation; patient readmission; hospitalization; emergency department; continuity of care; digital health

Du et al [1] reported that a nurse-led telephone call delivered approximately 48 hours after discharge was associated with lower emergency department (ED) use at 7 and 30 days but without statistically significant reductions in hospital readmission. Rather than interpreting this as a partial success, I would argue that the pattern of findings raises a more important conceptual point: for low-intensity postdischarge interventions, ED use may be a more appropriate and sensitive indicator of transitional care quality than readmission alone.

This argument matters because the hospital readmission metric continues to dominate how postdischarge interventions are judged, despite longstanding concerns about its conceptual and methodological limitations. Readmission is influenced not only by discharge quality but also by disease severity, frailty, multimorbidity, caregiver capacity, community service availability, and broader social disadvantage. A recent commentary in *JAMA Network Open* [2] has further argued that an exclusive focus on 30-day readmissions is increasingly inadequate because it omits other clinically meaningful revisits, particularly ED presentations and observation stays, which constitute a substantial proportion of postdischarge acute care use. Earlier policy analyses have similarly questioned the reliability of readmission rates as stand-alone indicators of hospital quality [3]. Against that background, Du et al's [1] findings may be less surprising than they first appear. A single follow-up call may reasonably reduce uncertainty-driven ED attendance without being sufficiently intensive to alter the more complex causal pathways that produce rehospitalization.

This interpretation is also consistent with the broader transitional care literature. A recent systematic review of interventions supporting transition from acute care to home in older adults found that benefits are most often observed with multicomponent interventions, especially those combining early follow-up, family involvement, and multidisciplinary input, and even then, effects tend to be modest and short-term [4]. Put differently, readmission is not an outcome that can be expected to shift easily in response to a single contact point, however well designed. When readmission becomes the dominant evaluative standard, there is a risk that lower-intensity but operationally scalable interventions are undervalued, even when they meaningfully improve continuity, reassurance, symptom management, or service navigation.

A second contribution of Du et al's [1] study is that it draws attention to the mechanism by which seemingly modest interventions may work. The authors found that 40% of completed calls identified at least one gap in discharge understanding or follow-up care. This finding should be interpreted in light of the wider evidence showing that patient understanding at discharge is often overestimated by clinicians and by health systems. Studies in emergency and hospital settings have repeatedly shown deficits in patients' comprehension of diagnosis, medication instructions, follow-up plans, and return precautions [5]. Recent work on discharge quality has similarly suggested that written instructions, while important, do not eliminate inequities in preparedness or self-management capacity after hospitalization [6]. The implication is that postdischarge follow-up should not be viewed simply as reinforcement of information already given. It functions as a second-stage clinical assessment of whether the

XSL•FO
RenderX

discharge plan is intelligible and actionable in the patient's actual home context.

From a nursing perspective, this is highly significant. Transitional care is often framed administratively, but the telephone call in this study is better understood as interpretive clinical work. The nurse is not merely checking whether instructions were received; the nurse is assessing whether they can be operationalized, whether medication regimens are manageable, whether symptoms are being interpreted appropriately, and whether follow-up arrangements are viable. This aligns with longstanding evidence that transitional care is most effective when it addresses coordination, continuity, and patient capability rather than information transfer alone [7]. In that respect, Du et al [1] contribute to a more realistic account of what postdischarge nursing support accomplishes: it manages uncertainty at the boundary between institutional care and patient self-management.

The paper is also relevant to current debates in digital health. Du et al [1] explicitly position telephone outreach as a digital health intervention. That framing is important because the digital health literature increasingly recognizes that technologically complex interventions can widen inequities when access, literacy, language, and infrastructure are unevenly distributed. A recent scoping review concluded that digital health technologies can reinforce inequalities and called for implementation strategies that explicitly address digital inclusion [8]. Empirical studies of telemedicine use have likewise shown that older age, limited English proficiency, public insurance, and other markers of disadvantage are associated with lower uptake of video visits relative to telephone care [9]. The significance of the intervention studied by Du et al [1], therefore, is not merely that it was effective but that it was delivered through a modality with comparatively low access barriers. In an environment where digital innovation is often equated with increasing technical sophistication, this study usefully supports a different principle: the value of a digital intervention should be judged partly by its reach and success in implementation, not only by its novelty.

The study's quasi-randomized, operational design also deserves attention. As it was embedded in routine service delivery, the findings speak to effectiveness under real service conditions rather than efficacy under idealized trial circumstances [1]. That does not remove concerns about allocation, exposure, or residual confounding, but it increases the policy relevance of the findings. The more important question is whether a service can be standardized, integrated into existing workflows, and delivered at scale without excessive burden. On that question, this study is informative.

As such, the central insight from Du et al [1] is not simply that a nurse-led telephone call reduces ED use. It is that evaluation of transitional care may need recalibration. If the intervention target is postdischarge self-management, then ED use may be the more responsive outcome because it captures a form of potentially preventable acute care use that precedes or substitutes for readmission. Future work should, therefore, move beyond a binary success-or-failure assessment based on readmission and instead evaluate transitional care using a broader set of outcomes, including ED revisits, patient preparedness, unresolved care gaps, and the distribution of benefit across patient groups. That approach would be more aligned with both the complexity of discharge and the realities of contemporary health system use.

Conflicts of Interest

None declared.

References

1. Du S, Anada A, Montenegro M, et al. Effectiveness of postdischarge telephone calls in reducing hospital utilization: quasi-randomized controlled trial. J Med Internet Res 2026 Mar 17;28:e80529. [doi: 10.2196/80529] [Medline: 41843752]
2. Wright B. In pursuit of a better hospital quality metric for Medicare. JAMA Netw Open 2025 Jul 1;8(7):e2524138. [doi: 10.1001/jamanetworkopen.2025.24138] [Medline: 40736740]
3. Press MJ, Scanlon DP, Ryan AM, et al. Limits of readmission rates in measuring hospital quality suggest the need for added metrics. Health Aff (Millwood) 2013 Jun;32(6):1083-1091. [doi: 10.1377/hlthaff.2012.0518] [Medline: 23733983]
4. Steiner LM, Osmancevic S, Hahn S, Bonetti L, Zwakhalen S. The effectiveness of interventions on clinical and patient-reported outcomes in hospital-to-home transitions of older adults: a systematic review. Eur J Ageing 2025 Nov 4;22(1):58. [doi: 10.1007/s10433-025-00890-w] [Medline: 41186737]
5. Sheikh H, Brezar A, Dzwonek A, Yau L, Calder LA. Patient understanding of discharge instructions in the emergency department: do different patients need different approaches? Int J Emerg Med 2018 Feb 8;11(1):5. [doi: 10.1186/s12245-018-0164-0] [Medline: 29423767]
6. Austad K, Lee JH, Lanney H, et al. Evaluating the quality and equity of patient hospital discharge instructions. BMC Health Serv Res 2025 Feb 21;25(1):291. [doi: 10.1186/s12913-025-12410-8] [Medline: 39980004]
7. Fernando RL, Karnon J, Crotty M, et al. Optimising older people's transition from acute care into residential aged care through multidisciplinary assessment and liaison (OPTIMAL): protocol for a stepped wedge cluster randomised controlled trial with embedded process evaluation. BMC Geriatr 2025 Jul 28;25(1):550. [doi: 10.1186/s12877-025-06187-y] [Medline: 40722056]
8. Badr J, Motulsky A, Denis JL. Digital health technologies and inequalities: a scoping review of potential impacts and policy recommendations. Health Policy 2024 Aug;146:105122. [doi: 10.1016/j.healthpol.2024.105122] [Medline: 38986333]

XSL•FO
RenderX

9. Rodriguez JA, Betancourt JR, Sequist TD, Ganguli I. Differences in the use of telephone and video telemedicine visits during the COVID-19 pandemic. Am J Manag Care 2021 Jan;27(1):21-26. [doi: 10.37765/ajmc.2021.88573] [Medline: 33471458]

Abbreviations

ED: emergency department

Edited by S Law, T Leung; submitted 16.Apr.2026; this is a non–peer-reviewed article;accepted 22.Apr.2026; published 07.May.2026.

Please cite as:
Woo BFY
Rethinking Postdischarge Intervention Evaluation
J Med Internet Res 2026;28:e98435
URL: https://www.jmir.org/2026/1/e98435
doi:10.2196/98435

© Brigitte Fong Yeong Woo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Moving From Keywords to Contextual Meaning: A Commentary on Hybrid Bibliometric Synthesis in Health Research

Dimitrios Zikos, BSN, MSc, PhD

Department of Healthcare Management and Leadership, College of Health Professions, Texas Tech University Health Sciences Center, 3601 4th Street, Lubbock, TX, United States

Corresponding Author:
Dimitrios Zikos, BSN, MSc, PhD
Department of Healthcare Management and Leadership, College of Health Professions, Texas Tech University Health Sciences Center, 3601 4th Street, Lubbock, TX, United States

Related Article:

https://www.jmir.org/2026/1/e86200

Abstract

The fast growth of social media mining in health research has contributed to an invaluable but quite fragmented body of literature. As the amount of unstructured patient-reported data grows, traditional bibliometric analyses face methodological limitations, particularly regarding synonym fragmentation and arbitrary parameter selection. In their recent publication, "Thematic Mapping and Evolution of Social Media Mining in Health Research: Hybrid Bibliometric Synthesis," Yang and Bohnet-Joschko attempt to address these flaws by introducing a semantic-structural (hybrid) bibliometric framework. This commentary evaluates the methodological innovations of their study and its departure from traditional syntactic keyword-matching tools. By combining citation-informed transformers (SPECTER2) and biomedical language models (PubMedBERT) and dimensionality reduction and density-based clustering, the authors created a reproducible pipeline. In their architecture, they start with foundational machine learning (statistical validity) before transitioning into large language models for qualitative synthesis. I will attempt to explain how this transition from syntactic mapping to semantic vector representation solves known challenges in evidence synthesis, naturally grouping conceptual synonyms without artificially forcing boundaries on the literature. Furthermore, I examine the practical implications of their temporal findings. Such real-time social media mining applications can be very useful for retrospective reporting and evaluating targeted public health interventions. While this pipeline offers high generalizability across disciplines, it also introduces a computational literacy barrier to some, and this re-emphasizes the need for data literacy for health professions. Ultimately, the study provides a transparent approach to informatics because mathematically validated frameworks are foundational for the future of evidence-driven public health policy and clinical decision-making.

(J Med Internet Res 2026;28:e102159) doi:10.2196/102159

KEYWORDS

bibliometrics; machine learning; semantic mapping; public health surveillance; social media

Introduction

The fast growth of social media mining (SMM) in health research has provided new opportunities to study public health. However, the volume of this literature has created a fragmented knowledge base that resists traditional evidence synthesis. In their recent article, "Thematic Mapping and Evolution of Social Media Mining in Health Research: Hybrid Bibliometric Synthesis," Yang and Bohnet-Joschko [1] go beyond descriptive field mapping and introduce a more semantically focused framework for evidence synthesis by extending traditional bibliometric methods with a hybrid semantic-structural pipeline. Since the health care data resources and knowledge base have already become massive and complex, synthesizing literature needs to go beyond just "cataloging" publications. This commentary explores how their methodology attempts to overcome statistical limitations, optimizes analytical architecture, and supports contextually relevant public health applications.

The Limitations of Traditional Keyword Matching

For over a decade, bibliometric reviews have mostly relied on out-of-the-box platforms like VOSviewer, but in some ways, those may reduce transparency regarding parameter sensitivity and clustering behavior [2]. They are created around prebuilt clustering algorithms (such as the Louvain method) that do not provide the researcher control over cluster thresholding, network normalizations, and parameter sensitivity. Additionally, these traditional methods are constrained by syntactic mapping: they treat keywords, titles, and abstracts of articles as isolated strings

XSL•FO
RenderX

of text. For example, traditional tools group papers by vocabulary (linking a Twitter flu paper with a bird flu paper), while the hybrid model groups them by scientific intent. Unless a researcher manually creates a detailed thesaurus, the algorithm fragments the literature based on vocabulary rather than meaning. Similarly, legacy topic modeling methods like latent Dirichlet allocation (LDA) rely on a rigid "bag-of-words" assumption that ignores contextual syntax and forces researchers to predefine the number of topics, artificially "bounding" the data [3].

Yang and Bohnet-Joschko [1] attempted to solve this by moving from syntactic to semantic mapping (Table 1). By choosing to use SPECTER2 and PubMedBERT embeddings, they ingest the entire context of a text and convert documents into high-dimensional vectors [4,5]. Because these models understand semantic meaning, they group conceptual synonyms together, making fewer errors than the traditional string-matching tools. Furthermore, because SPECTER2 is pretrained on very large amounts of citation data, it makes the approach an interesting hybrid of text mining and citation analysis.

Table . Comparison of traditional bibliometric methods to Yang and Bohnet-Joschko's [1] pipeline.

Dimension	Traditional bibliometric methods	Yang and Bohnet-Joschko's [1] hybrid pipeline
Data processing	Syntactic mapping of isolated text strings	Semantic vector representation using context
Core algorithms	Out-of-the-box platforms (eg, VOSviewer) using Louvain or latent Dirichlet allocation	Citation-informed transformers (SPECTER2), biomedical models (PubMedBERT), UMAP[a], and HDBSCAN[b]
Clustering	Relies on forced boundaries, often requiring predefined topic counts	Density-based clustering that mathematically isolates unassigned noise
Vocabulary handling	Fragments literature based on terminology unless a manual thesaurus is built	Naturally groups conceptual synonyms together based on scientific intent

[a]UMAP: uniform manifold approximation and projection.

[b]HDBSCAN: hierarchical density-based spatial clustering of applications with noise.

Optimization of the Architecture

The strength of Yang and Bohnet-Joschko's [1] methodology is in the sequencing of its analytical architecture. In the current era of artificial intelligence, it is tempting to feed raw data directly into a large language model. The study argues for an emerging informatics principle: structurally validated machine learning pipelines should precede large language model–assisted interpretation.

Yang and Bohnet-Joschko [1] validate their data structure using uniform manifold approximation and projection (UMAP) for dimensionality reduction [6] and hierarchical density-based spatial clustering of applications with noise (HDBSCAN) for structural clustering [7]. Unlike LDA, HDBSCAN does not require a predefined cluster count. It identifies naturally occurring dense regions of literature and mathematically isolates outlier papers rather than forcing them into irrelevant groups. This way, the subsequent thematic synthesis is grounded in reproducible data science.

From Methodology to Public Health Practice

Beyond the methodological advantages, the study's [1] temporal slicing is, for SMM, an example of moving from computational experimentation to real-world public health engagement. The prominence of application-driven clusters, such as infodemiology and sociopsychological determinants, shows well how SMM can be a useful tool for community health surveillance [8]. Accurate SMM is essential for evaluating localized, real-world health initiatives. For instance, evaluating colorectal cancer screening prevention campaigns in Texas requires an understanding of localized, real-time community sentiment, barriers to access, and sociopsychological hesitation. SMM captures this narrative, augmenting structured retrospective federal tools that often lag by months or years. Such real-time sentiment surveillance may also support decision-making for health systems, including targeted outreach, misinformation monitoring, and fast evaluation of intervention uptake.

Furthermore, the Yang and Bohnet-Joschko [1] treatment of HDBSCAN's unassigned noise (cluster 1) is a very reasonable analytical choice. Rather than dismissing this noise, they identify it as a candidate "incubator pool." I believe that embracing these topics can help researchers who want to predict the next wave of sociopsychological determinants before they become new established literature.

Limitations and Future Work

Despite its advantages over legacy systems, this pipeline introduces some challenges. The transition from graphical user interface–based tools to code-based learning models introduces a computational barrier to entry. There is a risk that evidence synthesis becomes gated behind advanced data science skills. Furthermore, while Yang and Bohnet-Joschko's [1] dual-level validation ensures semantic coherence, the study lacks an empirical comparison against older baselines (such as LDA or Louvain clustering) on the identical dataset, which would be necessary to quantify the reduction in fragmentation. Additionally, future validation across multiple bibliographic databases (eg, PubMed, Scopus, and Web of Science) would help determine the stability of the pipeline.

XSL•FO
RenderX

Researchers must be careful not to treat these new machine learning algorithms as new inherently opaque systems. While HDBSCAN eliminates the need to guess cluster counts, it introduces hypersensitivity to the minimum cluster size and the minimum sample size. Similarly, UMAP is dependent on neighbor and distance parameters. Future researchers who adopt this methodology will need to optimize and report their parameter selections to ensure their underlying models match their specific discipline. As a researcher and instructor, I would like to stress that addressing this requires shifts in health informatics education, promoting learning that relies on algorithmic logic and mechanics, and not just software operation.

Conclusion

Yang and Bohnet-Joschko's [1] study moved successfully from syntactic keyword-matching to semantic vector representation, in an attempt to avoid the limitations of traditional bibliometrics. More importantly, they provided a transparent, reproducible blueprint for future studies. As the volume of medical literature and patient-reported data continues to increase, adopting machine learning pipelines is an imperative for any research designed to extract evidence-driven insights to guide patient care and public health policy.

Acknowledgments

Generative artificial intelligence was not used in any capacity during the preparation, writing, or editing of this manuscript.

Funding

The author declared no financial support was received for this work.

Conflicts of Interest

None declared.

References

1. Yang MJ, Bohnet-Joschko S. Thematic mapping and evolution of social media mining in health research: hybrid bibliometric synthesis. J Med Internet Res 2026 May 8;28:e86200. [doi: 10.2196/86200] [Medline: 42115141]
2. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 2010 Aug;84(2):523-538. [doi: 10.1007/s11192-009-0146-3] [Medline: 20585380]
3. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Machine Learning Res 2003;3:993-1022 [FREE Full text]
4. Cohan A, Feldman S, Beltagy I, Downey D, Weld DS. SPECTER: document-level representation learning using citation-informed transformers. arXiv. Preprint posted online on Apr 15, 2022. [doi: 10.48550/arXiv.2004.07180]
5. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare 2022 Jan 31;3(1):1-23. [doi: 10.1145/3458754]
6. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv. Preprint posted online on Feb 9, 2018. [doi: 10.48550/arXiv.1802.03426]
7. Campello R, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates. In: Pei J, Tseng VS, Cao L, Motoda H, Xu G, editors. Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II 2013:160-172. [doi: 10.1007/978-3-642-37456-2_14]
8. Eysenbach G. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. Am J Prev Med 2011 May;40(5 Suppl 2):S154-S158. [doi: 10.1016/j.amepre.2011.02.006] [Medline: 21521589]

Abbreviations

HDBSCAN: hierarchical density-based spatial clustering of applications with noise
LDA: latent Dirichlet allocation
SMM: social media mining
UMAP: uniform manifold approximation and projection

XSL•FO
RenderX

Edited by S Law, T Leung; submitted 22.May.2026; this is a non–peer-reviewed article;accepted 22.May.2026; published 03.Jun.2026.

Please cite as:
Zikos D
Moving From Keywords to Contextual Meaning: A Commentary on Hybrid Bibliometric Synthesis in Health Research
J Med Internet Res 2026;28:e102159
URL: https://www.jmir.org/2026/1/e102159
doi:10.2196/102159

© Dimitrios Zikos. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 3.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

From Data Stewardship to Model Stewardship: Extending Governance Frameworks for AI Era Health Data Use

Leon Rozenblit[1,2], JD, PhD; Steven Labkoff[2,3], MD; Charles Safran[2,4], MS, MD

[1]Q.E.D. Institute, New Haven, CT, United States

[2]Division of Clinical Informatics, Beth Israel Deaconess Medical Center, 133 Brookline Avenue, HVMA Annex, Suite 2200, Boston, MA, United States

[3]Luminant Consulting, Stamford, CT, United States

[4]Department of Medicine, Harvard Medical School, Boston, MA, United States

Corresponding Author:
Leon Rozenblit, JD, PhD
Q.E.D. Institute, New Haven, CT, United States

Abstract

Maris et al document important ethical challenges at the intersection of electronic health record data and artificial intelligence development, but existing governance frameworks designed for secondary data use are categorically insufficient for artificial intelligence model training, which creates persistent deployable artifacts that encode local clinical patterns as generalizable knowledge. Drawing on two decades of stewardship framework development, we propose extending governance from data stewardship to model stewardship.

(J Med Internet Res 2026;28:e97859) doi:10.2196/97859

KEYWORDS

data stewardship; model stewardship; AI governance; electronic health records; decontextualization; secondary use; clinical tropism; health data ethics

Introduction

Maris et al [1] make an empirical contribution to the ethics of health data use for artificial intelligence (AI), grounding four cross-cutting themes (privacy, public trust, fair representation, and responsible integration) in stakeholder perspectives from the LEAPfROG project. Their identification of "decontextualization" as a central challenge deserves particular attention. We write from the vantage of two decades of work on stewardship frameworks for health data. The American Medical Informatics Association (AMIA) national framework for secondary use [2], the National Committee on Vital and Health Statistics (NCVHS) stewardship report to the Department of Health and Human Services [3], and the elaboration of data stewardship principles [4] established core principles (accountability, chain of trust, transparency, data quality) that Maris et al's [1] stakeholders independently rediscover. This convergence is validating but concerning: the principles hold, yet remain unoperationalized. We argue that AI model training represents a fundamentally new form of data use requiring a shift from data stewardship to model stewardship.

AI Model Training Is Not Your Grandfather's Secondary Use

Nearly 20 years ago, a national expert panel defined the secondary use of health data as uses beyond direct patient care, including research, quality measurement, public health surveillance, and commercial applications [2]. AI model training falls under this broad umbrella, but it differs from every use the framework's architects envisioned. Traditional secondary uses analyze data and produce bounded findings; AI training creates persistent, deployable artifacts: models that may be commercialized globally, influence clinical decisions at scale, and embed the assumptions of their training context into every future prediction. A research study produces conclusions bounded by its methods and sample; a model trained on the same data produces an artifact with unbounded downstream reach and no expiration date.

The NCVHS [3] recommended abandoning "secondary use" as too imprecise for meaningful governance; advice even more apt today. AI training should be recognized as a qualitatively distinct category of secondary use, with stewardship requirements reflecting its unique characteristics: persistence, scalability, commercial deployment, and the encoding of institutional context as generalizable knowledge.

The Challenge of Decontextualization

Maris et al [1] identify decontextualization as a cross-cutting ethical concern. We argue it is more fundamental than their analysis suggests: not one challenge among several, but the mechanism through which the others arise.

XSL•FO
RenderX

Electronic health record data encode not only clinical facts but institutional workflows, documentation practices, coding conventions, billing incentives, and resource constraints. Van der Lei's [5] first law of medical informatics, that data should be used only for the purpose for which they were collected, takes on new force when the reuse creates persistent, deployable artifacts rather than bounded research findings.

Two distinct challenges are at work. The first is data quality, and it is improvable: advances in ambient documentation, terminology that captures clinical intent, and better problem list governance will strengthen electronic health record data over time. The second is structural and persists regardless of data quality: every dataset carries the institutional fingerprint of its origin. What Maris et al [1], following Alami et al [6], term "clinical tropism," the tendency of AI to reproduce narrow training environment practices, is a symptom of this structural layer. A model trained at an academic medical center with aggressive sepsis protocols learns different signals than one at a community hospital, not because the data are poor, but because they faithfully reflect different contexts (Figure 1).

Deploying models trained in one context across settings that differ systematically risks disadvantaging patients in predictable, preventable ways, the kind of harm that stewardship frameworks were designed to address.

Figure 1. Progressive decontextualization of clinical data. Panel A: a recognizable scene representing the rich context of clinical reality. Panel B: the same scene reduced to a grayscale grid of discrete tiles. Panel C: the tiles resorted by value, severing all spatial relationships—a visual metaphor for how electronic health record data lose institutional context when extracted for model training. AI: artificial intelligence.

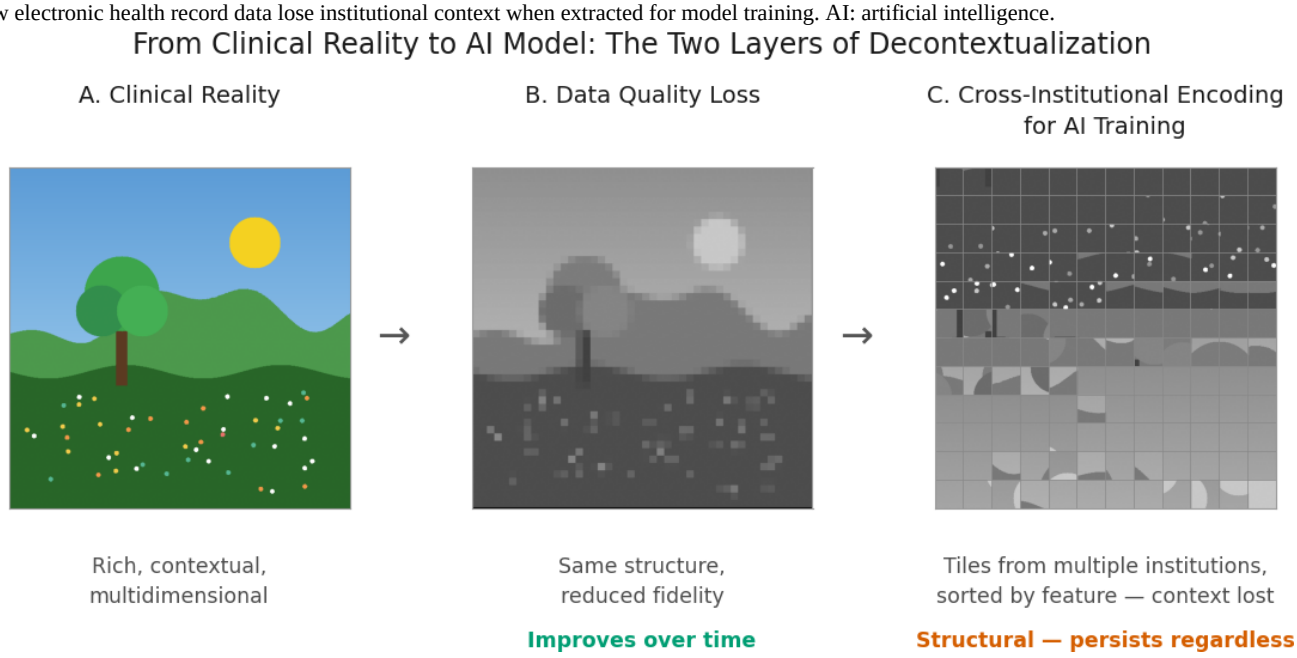

From Data Stewardship to Model Stewardship

The AMIA and NCVHS frameworks established stewardship principles for health data: accountability, chain of trust, transparency, oversight, data quality, and individual participation [3,4]. These principles must now extend to AI models and the datasets used to train them.

Consider the chain of trust, a core NCVHS concept. When data flow from hospital to aggregator to AI company to commercial model to clinical deployment across institutions, with the training data never represented, the chain does not merely stretch; it breaks. Who is the steward of a model trained on data from five health systems and deployed in 50?

Recent work offers concrete starting points that are achievable with existing infrastructure. Multistakeholder governance frameworks [7,8] propose domain-specific approaches: clinical decision support, real-world evidence generation, and consumer health AI each require distinct governance structures. The Safe, Effective, Equitable, Trustworthy (SEET) framework provides organizing principles [8], while recommendations for AI-enabled clinical decision support specify validation, certification, safety monitoring, adverse event reporting, and provenance documentation requirements [9,10]. Real-world data governance standards, including metadata requirements and bias documentation [9], offer complementary infrastructure.

Model stewardship, we propose, should encompass at minimum training data provenance documentation, so downstream users know what populations and practice settings a model reflects; (2) cross-institutional validation before deployment beyond the training context; (3) ongoing monitoring for context drift as clinical practices evolve; and (4) accountability structures that follow the model through its life cycle, not merely the data at its origin. All these requirements are technically feasible: provenance documentation and validation protocols exist in other regulated domains. Do we have the will to mandate them?

Conclusion

Maris et al [1] are right that stakeholder-led governance is essential. But governance must evolve to match its target. AI models are not simply a new use of data; they are new artifacts with their own life cycle, risks, and accountability requirements. The stewardship frameworks built over two decades provide a

XSL•FO
RenderX

proven foundation; extending them is the challenge. The immediate task is clear: require training data provenance and cross-institutional validation as preconditions for clinical AI deployment, just as we require evidence of efficacy before deploying therapeutics.

Acknowledgments

Generative artificial intelligence tools (Anthropic Claude, Opus 4.6 model) were used to assist with literature organization, outline structuring, prose drafting, and generating illustrative figures during the preparation of this commentary. All intellectual content, arguments, and conclusions are the authors' own. The authors reviewed, edited, and take full responsibility for the final manuscript.

Funding

The authors declared that no financial support was received for this work.

Data Availability

This commentary reports no original research data.

Authors' Contributions

LR conceived the commentary, developed the argument structure, and wrote the first draft. SL contributed to the historical stewardship framework analysis and codeveloped the thesis. CS reviewed the outline and draft, contributed to the data quality improvement perspective, and provided critical revisions. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Maris MT, Klopotowska JE, Cornet R, et al. The ethics of leveraging routinely collected patient data for AI development: mixed methods study. J Med Internet Res 2026 Mar 2;28:e79863. [doi: 10.2196/79863] [Medline: 41814967]
2. Safran C, Bloomrosen M, Hammond WE, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. J Am Med Inform Assoc 2007;14(1):1-9. [doi: 10.1197/jamia.M2273] [Medline: 17077452]
3. National Committee on Vital and Health Statistics. Enhanced protections for uses of health data: a stewardship framework for "secondary uses" of electronically collected and transmitted health data. : Department of Health & Human Services; 2007 URL: https://ncvhs.hhs.gov/wp-content/uploads/2013/12/071221lt.pdf [accessed 2026-04-09]
4. Bloomrosen M, Detmer DE. Advancing the framework: use of health data--a report of a working conference of the American Medical Informatics Association. J Am Med Inform Assoc 2008;15(6):715-722. [doi: 10.1197/jamia.M2905] [Medline: 18755988]
5. van der Lei J. Use and abuse of computer-stored medical records. Methods Inf Med 1991 Apr;30(2):79-80. [doi: 10.1055/s-0038-1634831] [Medline: 1857252]
6. Alami H, Lehoux P, Auclair Y, et al. Artificial intelligence and health technology assessment: anticipating a new level of complexity. J Med Internet Res 2020 Jul 7;22(7):e17707. [doi: 10.2196/17707] [Medline: 32406850]
7. Rozenblit L, Price A, Solomonides A, et al. Towards a multi-stakeholder process for developing responsible AI governance in consumer health. Int J Med Inform 2025 Mar;195:105713. [doi: 10.1016/j.ijmedinf.2024.105713] [Medline: 39642592]
8. Rozenblit L, Price A, Solomonides A, et al. Toward responsible AI governance: balancing multi-stakeholder perspectives on AI in healthcare. Int J Med Inform 2025 Nov;203:106015. [doi: 10.1016/j.ijmedinf.2025.106015] [Medline: 40680319]
9. Koski E, Das A, Hsueh PYS, et al. Towards responsible artificial intelligence in healthcare-getting real about real-world data and evidence. J Am Med Inform Assoc 2025 Nov 1;32(11):1746-1755. [doi: 10.1093/jamia/ocaf133] [Medline: 40999782]
10. Labkoff S, Oladimeji B, Kannry J, et al. Toward a responsible future: recommendations for AI-enabled clinical decision support. J Am Med Inform Assoc 2024 Nov 1;31(11):2730-2739. [doi: 10.1093/jamia/ocae209] [Medline: 39325508]

Abbreviations

AI: artificial intelligence
AMIA: American Medical Informatics Association
NCVHS: National Committee on Vital and Health Statistics
SEET: Safe, Effective, Equitable, Trustworthy

XSL•FO
RenderX

Edited by T Leung; submitted 10.Apr.2026; this is a non–peer-reviewed article;accepted 03.May.2026; published 05.Jun.2026.

Please cite as:
Rozenblit L, Labkoff S, Safran C
From Data Stewardship to Model Stewardship: Extending Governance Frameworks for AI Era Health Data Use
J Med Internet Res 2026;28:e97859
URL: https://www.jmir.org/2026/1/e97859
doi:10.2196/97859

© Leon Rozenblit, Steven Labkoff, Charles Safran. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Maturity, Safety, and Equity of AI-Enabled Systems and Triage in Integrated Primary Care

Siaw-Teng Liaw, MBBS, PhD, Prof Dr

Discipline of General Practice, School of Clinical Medicine and School of Population Health, UNSW Sydney, 24 Kelvinside St, Balwyn North, Victoria, Australia

Corresponding Author:
Siaw-Teng Liaw, MBBS, PhD, Prof Dr
Discipline of General Practice, School of Clinical Medicine and School of Population Health, UNSW Sydney, 24 Kelvinside St, Balwyn North, Victoria, Australia

Related Article:

https://www.jmir.org/2026/1/e88396

Abstract

Artificial intelligence (AI)–enabled systems must simultaneously improve the Quintuple Aim and digital health maturity, including equitable access to and quality and interoperability of data, tools, agents, and services. This requires a comprehensive sociotechnical and global approach to cocreation, management, and governance for individuals and organizations in the ecosystem.

(J Med Internet Res 2026;28:e97341) doi:10.2196/97341

KEYWORDS

artificial intelligence; AI; triage; primary care; general practice; patient safety; health equity; real-world evidence; clinical decision support; digital maturity; Quintuple Aim

Alamoudi et al [1] described the potential of artificial intelligence (AI)–enabled triage, lamented the lack of credible evidence, criticized the lack of real-world and equity-stratified outcomes, and proposed a sociotechnical approach to a "practical agenda for evaluating and governing AI-enabled triage in primary care that integrates real-world safety outcomes, equity-stratified performance reporting, and sociotechnical implementation and surveillance." While commendable, equity aims must be set within the Quintuple Aim, which includes the cost-effective care of the individual and population, patient satisfaction, provider well-being, and equitable health for all [2], and the AI ethics and governance principles proposed by the World Health Organization (WHO) [3].

AI-enabled triage systems collect and analyze patient data to determine the urgency of care required, provide a risk assessment, and recommend a triage level, including self-care, routine review, urgent care, or emergency referral. This definition means that the management, governance, and evaluation of AI-enabled triage systems in primary care require an integrated approach where data, tools, agents, services, and systems are integrated and interoperable with those in other parts of the health ecosystem [4].

The learning and reasoning models underlying AI-enabled triage in general/family practice and primary care must reflect their patient-centered biopsychosocial diagnostic process, including making a probability diagnosis, excluding dangerous things, identifying differentials, and elucidating "hidden" patient agendas. This requires good quality models, algorithms, and data that are consistent with local community and institutional values, benchmarks, and priorities.

Therefore, this sociotechnical agenda must be implemented within a conceptual framework that includes improving the maturity, quality, and interoperability of the essential digital health foundations to enable the individual, organization, community, and ecosystem to adopt and integrate AI-enabled systems to achieve the Quintuple Aim successfully and sustainably.

However, the rapid deployment of AI-enabled systems despite a low evidence base for safety, bias, and quality [1,5], and the low digital health maturity (DHM) of individuals, organizations, and ecosystems raises concerns about inadequate digital transformation for health services adopting AI-enabled systems to achieve the Quintuple Aim [6].

The Digital Health Profile and Maturity Assessment Toolkit (DHPMAT) was developed to create a digital health profile and assess the DHM of health organizations and ecosystems. The digital health profile and DHM are then aligned with national, subnational, or organizational priorities to develop a realistic road map for adopting, managing, governing, evaluating, improving, and sustaining digital health interventions [4,7].

Figure 1 illustrates the use of DHPMAT to assess the sociotechnical and financial readiness of citizens, digital health professionals, and organizations to adopt AI-enabled systems,

XSL•FO
RenderX

including triage, in the integrated health ecosystem. Readiness to implement digital health systems is determined by assessing the maturity of five essential digital health foundations (Figure 1, rows): information and communication technology, Internet-of-Things, or AI infrastructure; digital public goods; platform for information sharing; enabling environment; and quality improvement, measurement, monitoring, and evaluation. The five linear stages of DHM (Figure 1, columns) describe an initial adapting and assessing stage, which matures to controlling, standardizing, optimizing, and innovating stages. Standardizing at the center of Figure 1 (bolded in red) is the keystone stage [4].

Figure 1. Indicators to assess readiness to support artificial intelligence (AI)–enabled triage and systems. incl: including; Mx: management; R&D: research and development; re: regarding.

Essential Digital Foundations to support and sustain AI-enabled systems incl triage	Digital health maturity stages with sociotechnical indicators				
	ADAPTING *Unpredictable Knowledge & data not shared* ***Assessing & adapting***	**CONTROLLED** *Predictable & consistent Knowledge sharing started* ***Problem driven***	**STANDARDISED** *Organization-wide knowledge & data sharing* ***Request driven***	**OPTIMIZING** *Learning health organization Enterprise-wide knowledge sharing* ***Service driven***	**INNOVATING** *Learning health ecosystem System-wide knowledge sharing* ***Value driven***
ICT/Internet of Things (IoT)/AI "compute" infrastructure	Enterprise architecture (EA) plan for secure, reliable, & equitable ICT/IoT/AI infrastructure	Standard Operating Procedures (SOPs) for equity & standards for ICT/IoT/AI infrastructure	EA Platform with SOPs for security & equity of ICT/IoT/AI infrastructure incl data centers	SOPs for CQI of EA platform to support data & agents across organization.	SOPs to support system-wide algorithm-driven EA platform
Essential AI-enabled digital public goods (DPG)	Evidence-based open-source AI Triage system shortlisted as per EA Plan SOPs	AI Triage system tested for quality benchmarks and interoperability	AI Triage system performing consistently on request on the EA Platform.	CQI of AI Triage system inter-operable with other AI-DPGs shared across organisation.	Enterprise-wide AI Triage system driving innovation in integrated primary care.
Readiness for AI-enabled information sharing	EA plan with SOPs to manage algorithm-driven data analytics & equitable sharing	SOPs to manage & govern interoperable algorithm-driven data, tools, agents & services	**SOPs about quality & interoperability of data, agents & services on EA Platform in the organisation**	CQI of data, agents & services shared on EA platform across the enterprise	Enterprise-wide AI-driven data linkage within a common data model for innovation
Enablers of trust & adoption of AI-enabled digital transformation	EA plan with road maps to develop staff & organisation in ethical & legal compliance requirements	SOPs for transparency, reciprocity, and change Mx to address ELSI & promote trust.	EA Platform compliant to regulatory, privacy, security, & education benchmarks for digitalization & AI	CQI of compliance & education support system on EA platform with citizen engagement	Enterprise-wide AI-driven innovative governance, management, & citizen engagement
Quality improvement, measurement, monitoring & evaluation (QIMME)	EA plan with road map to achieve a CER program with AI-driven data, agents & services	SOPs to monitor CER program of AI-driven tools & agents in the organization	EA Platform supporting CER to cocreate evidence for DHM & Quintuple Aim in the learning organization	CQI of the CER program & QIMME culture across the enterprise learning system	Enterprise-wide AI-driven R&D in equitable digitalized clinical & population health services

NOTES: CER=Comparative Effectiveness Research; CQI=Continuous Quality Improvement; DPG=Digital public goods; EA=Enterprise architecture, ELSI=Ethical, legal & social issues; ICT=Information & Communication Technology; IoT=Internet of Things; Mx=management; R&D=research & development; SOP=Standard Operating Procedures

XSL•FO
RenderX

Many organizations often skip maturity stages, usually because politicians favor "shiny new toys" over upstream essential but unglamorous standards and quality work to achieve an enterprise architecture that supports interoperability. This lack of well-implemented standards, along with proprietary systems in a competitive business environment, explains why interoperability remains a longstanding challenge and digital health interventions frequently fail to realize their potential benefits. It is also why the DHPMAT encourages the cocreation and use of open-source digital public goods.

Cocreation, including co-design, coimplementation, comonitoring, co-evaluation, and coimprovement over the development life cycle, is encouraged as a sociotechnical process to ensure that individual digital competencies and organizational DHM are improved concurrently with the Quintuple Aim [1,2,4]. Cocreating safe, useful, and usable cost-effective AI-enabled systems requires mutually trusting relationships among consumers, technology vendors, and health professionals in organizations and ecosystems. A fair and just regulatory framework that covers intellectual property (IP) and social capital is essential to cocreating, managing, governing, and sustaining digital assets equitably in local organizations and communities.

The balance between proprietary (private good) and open-source (public good) business models for digital health should be managed, especially in less-resourced countries. Proprietary organizations such as OpenAI have scraped the internet for diverse personal and public information, without formal consent from or payment to IP owners, to develop the generative pretrained transformer and large language model that underpin ChatGPT. Ironically, ChatGPT is now proprietary, and OpenAI is accusing DeepSeek, an open-source AI technology, of IP theft. OpenAI also illustrates the need for caution about AI-driven organizations with questionable business and humanitarian ethics [8].

A problem with terminology standards is the commodification of SNOMED-CT (Systematized Nomenclature of Medicine—Clinical Terms) and pharmaceutical classifications/vocabularies to be sold as proprietary terminology servers, adding financial and access barriers to semantically interoperable health information systems. In contrast, the WHO *International Statistical Classification of Diseases, 11th Revision* (*ICD-11*) is an open-source digital terminology public good and an alternative to SNOMED-CT. This private versus public good tension is inherent in every sector of the global economy [9].

Cocreating AI technology as a digital public good in a networked multilateral world is attractive. However, the corruption of efforts to develop digital public goods during COVID-19 by profit-seeking "big tech" and "big pharma" led to expensive suboptimal "shiny new toys" as well as vaccine, antiviral, and food apartheid. This continues to exacerbate inequity, morbidity, and mortality among people on the wrong side of the race, age, gender, poverty, and cultural divides [10].

While capitalist economies have greater inequality than socialist ones, both must deal with inequity as unintended sociotechnical consequences of development. Will the capitalist approach with privately owned corporations and "trickle-down-economics" or the socialist approach with state-controlled enterprises be more successful in minimizing digital health inequity [10]?

Equity-stratified AI-enabled health systems, including triage, must simultaneously improve the Quintuple Aim and DHM, especially access to high-quality and interoperable data, tools, agents, and services. This requires comprehensive sociotechnical and intersectoral approaches to investment, cocreation, management, governance, and quality improvement of digital tools to benefit individuals, organizations, and global ecosystems.

Conflicts of Interest

None declared.

References

1. Alamoudi A, Kontopantelis E, Zghebi SS, Brown B. AI triage in primary care: building safer and more equitable real-world evidence. J Med Internet Res 2026 Mar 4;28:e88396. [doi: 10.2196/88396] [Medline: 41780919]
2. Nundy S, Cooper LA, Mate KS. The Quintuple Aim for health care improvement: a new imperative to advance health equity. JAMA 2022 Feb 8;327(6):521-522. [doi: 10.1001/jama.2021.25181] [Medline: 35061006]
3. Ethics and Governance of Artificial Intelligence for Health: WHO Guidance: World Health Organization; 2021.
4. Liaw ST. Chapter 1 - digital health maturity. In: Liaw ST, Jonnagaddala J, Godinho MA, Wilkins-Wong ZSY, editors. Digital Health Maturity: Quality, Interoperability, and Innovation: Academic Press; 2026:3-55. [doi: 10.1016/B978-0-323-95260-6.00001-0]
5. Laranjo L, Tudor Car L, Payne RE, Neves AL, Kidd M, Jaime Miranda J. Artificial intelligence in primary care: innovation at a crossroads. Lancet Prim Care 2026 Mar;2(3). [doi: 10.1016/j.lanprc.2025.100078] [Medline: 41969643]
6. Katonai G, Arvai N, Mesko B. AI and primary care: scoping review. J Med Internet Res 2025 Aug 15;27:e65950. [doi: 10.2196/65950] [Medline: 40836794]
7. Kuziemsky C, Zhou R, Liaw ST. Chapter 7 - managing health systems to improve digital health maturity. In: Liaw ST, Jonnagaddala J, Godinho MA, Wilkins-Wong ZSY, editors. Digital Health Maturity: Quality, Interoperability, and Innovation: Academic Press; 2026:179-194. [doi: 10.1016/B978-0-323-95260-6.00007-1]
8. Hao K. Empire of AI: Inside the Reckless Race for Total Domination: Allen Lane; 2025.

XSL•FO
RenderX

9. WHO explores collaboration with SNOMED International for linkage between ICD-11 and SNOMED CT. World Health Organization. 2024 Oct 22. URL: https://tinyurl.com/47t9j7wn [accessed 2024-11-12]
10. Varoufakis Y. Technofeudalism: What Killed Capitalism: Penguin Random House; 2024.

Abbreviations

AI: artificial intelligence
DHM: digital health maturity
DHPMAT: Digital Health Profile and Maturity Assessment Toolkit
***ICD-11*:** *International Statistical Classification of Diseases, 11th Revision*
IP: intellectual property
SNOMED-CT: Systematized Nomenclature of Medicine—Clinical Terms
WHO: World Health Organization

Edited by S Law, T Leung; submitted 06.Apr.2026; this is a non–peer-reviewed article;accepted 27.Apr.2026; published 14.May.2026.

Please cite as:
Liaw ST
Maturity, Safety, and Equity of AI-Enabled Systems and Triage in Integrated Primary Care
J Med Internet Res 2026;28:e97341
URL: https://www.jmir.org/2026/1/e97341
doi:10.2196/97341

© Siaw-Teng Liaw. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Hospitalists Are Already Using AI—Why Implementation Will Determine Its Impact

Anna Maw*, MS, MD; Aakriti Pandita*, MD; Marisha Burden, MBA, MD

Division of Hospital Medicine, University of Colorado School of Medicine, Leprino Building, 4th Floor, Mailstop F-782, 12401 E 17th Avenue, Aurora, CO, United States

*these authors contributed equally

Corresponding Author:
Aakriti Pandita, MD
Division of Hospital Medicine, University of Colorado School of Medicine, Leprino Building, 4th Floor, Mailstop F-782, 12401 E 17th Avenue, Aurora, CO, United States

Related Article:

https://www.jmir.org/2026/1/e85973

Abstract

The adoption of artificial intelligence (AI) into clinical practice is accelerating, outpacing the development of organizational guidance, training, and governance. A recent study indicated that two-thirds of hospitalists are using AI, particularly large language model (LLM)–based platforms, in their clinical work. However, as with prior disruptive health technologies, adoption alone does not ensure meaningful improvement in care. Drawing on lessons from electronic health record implementation, we argue that AI's ultimate impact will be determined not by use rates, but by implementation quality and fit. Poorly implemented digital tools have been shown to increase clinician workload and burnout, despite their intended benefits. Early evidence on LLM-based diagnostic AI further underscores this risk: clinical-decision making supported by AI may be suboptimal when integration, training, and workflow design are inadequate. To provide value, AI tools must be thoughtfully embedded into clinical reasoning processes through evidence-informed training, intentional workflow design, and supportive organizational culture. As AI technologies are rapidly adopted, three priorities come into focus: training clinicians on AI inputs and interpreting outputs, applying implementation science frameworks for AI deployment in clinical environments, and establishing strategies for ongoing evaluation of the impact of AI tools over time. Implementation science frameworks offer practical guidance to assess workflow integration, training needs, infrastructure, and potential unintended consequences that can then inform adaptation of implementation strategies to enhance contextual fit. In parallel, learning health system infrastructure can enable continuous monitoring and iterative adaptation using routinely collected clinical and workflow data that reflect the value of the intervention across the quintuple aim of clinical outcomes, health equity, cost, and patient and clinician experience. AI adoption in hospital medicine is likely inevitable. Its ability to advance the quintuple aim will depend on how effectively these tools are implemented, supported, evaluated, and adapted in practice.

(J Med Internet Res 2026;28:e97419) doi:10.2196/97419

KEYWORDS

artificial intelligence; generative AI; hospital medicine; clinical decision support systems; health care technology adoption; implementation science

The integration of artificial intelligence (AI) into clinical practice is accelerating, with clinicians adopting AI tools faster than health systems can guide and govern their use, often without structured training or guidance to interpret outputs or manage potential risk. In a recent study by Bagla and colleagues [1], two-thirds of hospitalists reported using AI platforms in clinical practice, most commonly large language model (LLM)–based tools, and were doing so largely in the absence of health system integration. Hospitalists reported using these tools primarily to support clinical decision-making, including generating differential diagnoses and management options, with a strong preference for medical-specific platforms over general-purpose applications. Yet as with any disruptive health technology, adoption alone should not be considered success. Prior experience has shown that widespread uptake of technology does not guarantee progress toward the quintuple aim of optimizing patient outcomes, health equity, cost, and patient and clinician experience [2]. The work by Bagla et al [1] provides an important foundation by highlighting how AI is already being used in practice, with broader applications and tools likely to continue expanding. The critical next question is whether and under what conditions these patterns of use translate into meaningful improvements across the quintuple aim,

XSL•FO
RenderX

particularly as health systems begin to consider how these tools should be implemented, supported, and governed.

Whether clinician-initiated (as with many current AI tools) or system-led (as with electronic health record [EHR] implementation), health technologies often fall short when implementation, training, and workflow redesign lag behind adoption. The history of EHR implementation provides a cautionary example of adopting technology without careful attention to implementation and contextual fit. While EHRs have contributed to important gains in standardization and safety, they have also introduced risks, including interoperability challenges, clinician inefficiencies and cognitive burden, and even downstream patient safety concerns [3]. Importantly, these harms largely reflect design and implementation failures rather than what the technology itself can offer. As emerging AI technologies are rapidly adopted, three priorities come into focus: training clinicians on prompt engineering and interpreting outputs, applying implementation science frameworks for AI deployment in clinical environments, and establishing strategies for ongoing evaluation of the impact of AI tools over time.

As health systems move toward integration of various AI technologies (including LLM-based generative AI solutions and beyond), emerging evidence from randomized controlled trials highlights how implementation and training critically shape their impact. For example, when comparing diagnostic performance of LLMs across three groups—clinicians alone, clinicians with AI assistance, and AI alone—Goh and colleagues [4] reported that AI assistance did not improve diagnostic reasoning of clinicians and AI alone outperformed clinicians. This does not necessarily imply that clinicians should be replaced; rather, it reflects a critical gap in implementation. Simply adding AI to clinical practice is insufficient. Without evidence-informed training and structured approaches to use, clinicians may fail to appropriately prompt, interpret, contextualize, and act on AI outputs, limiting the clinical benefits that could be achieved. Additional evidence from randomized controlled trials further demonstrates that it is not only *whether* AI is used, but *how* and *when* it is integrated into clinical reasoning that shapes outcomes. Everett et al [5] demonstrated that when clinicians received targeted training and used intentionally designed AI-first or AI-second workflows, diagnostic accuracy improved markedly, highlighting that *how* AI is integrated influences outcomes. Similarly, early studies of AI-enabled documentation tools, such as AI scribes, show promise but also demonstrate variable benefit across care settings and workflows [6].

To advance the quintuple aim, AI use in clinical care must be approached as more than simply deploying a new tool, with careful attention to how it is introduced, supported, and used in practice [7]. This is particularly important for LLM-based platforms being used to support clinical decision-making, whose outputs vary by prompt input and can include overly confident, sometimes erroneous output that could negatively impact clinician decision-making. Implementation science frameworks provide practical scaffolding for operational decision-making, particularly as AI tools are rapidly deployed into clinical environments. For instance, the pragmatic robust implementation and sustainability model (PRISM) can help health system leaders systematically assess key factors, such as workflow integration, organizational culture, infrastructure (such as training), and external pressures, and anticipate and measure how these elements influence outcomes and unintended consequences [8]. In practice, this means moving beyond “Does it work?” to more actionable questions: “Where in our workflows should this tool be used?” “Who needs training?” “What could go wrong?” “How will we monitor and improve performance over time?” and “What are the impacts on patients, the workforce, and the health system?” At the same time, responsible use extends beyond organizational strategy to individual clinicians who are already incorporating AI tools into their practice. Clinicians must understand the strengths and limitations of the AI tools they use clinically, and how the outputs of these tools influence their decision-making. Clinicians must also ensure that their tool use aligns with Health Insurance Portability and Accountability Act (HIPAA) requirements and organizational policies.

Finally, to understand whether training and deployment strategies are working in practice, implementation of AI tools at scale will require ongoing health system evaluation and adaptation rather than one-time deployment. Learning health system infrastructure, which allows system leaders to leverage data collected through routine clinical care, can enable continuous, near-real-time monitoring of process and outcome measures, making timely iterative improvement feasible at scale [9]. Fortunately, much of these data already exist within the EHR, from audit logs that reveal workflow patterns to clinical data that track patient outcomes [10]. However, measurement must extend beyond what is easily captured: periodic assessments of clinician experience and workflow burden are critical, as are economic evaluations to ensure that AI tools deliver on their investment.

Ultimately, AI adoption in hospital medicine (and beyond) will only continue, but its impact on critical outcomes remains uncertain. The central challenge is no longer whether these tools are adopted but whether they are implemented in ways that meaningfully improve outcomes for patients, clinicians, and health systems.

Acknowledgments

The authors used Microsoft CoPilot and ChatGPT versions 5 and 5.3 for editing of original content to improve readability. All information and materials in the manuscript are original.

Funding

The authors declared no financial support was received for this work.

XSL•FO
RenderX

Authors' Contributions

All authors contributed to conceptualization, writing, review, and editing of the manuscript. All authors reviewed and approved the final manuscript.

Conflicts of Interest

MB reports funding from the Agency for Healthcare Research and Quality, the National Institute for Occupational Health and Safety, a University of Colorado Innovations digiSPARK award, and the American Medical Association not related to this work. MB contributed to the development of GrittyWork, a digital workforce application and a registered trademark of the University of Colorado not related to this work. MB reports an honorarium from Med-IQ not related to this work. AP reports funding from a University of Colorado Innovations digiSPARK award not related to this work. AP contributed to the development of an AI-powered application not related to this work.

References

1. Bagla P, Hanna J, Marthambadi B, Watkins S. Patterns of AI use in clinical work by hospitalists: survey study. J Med Internet Res 2026 Mar 3;28:e85973. [doi: 10.2196/85973] [Medline: 41581898]
2. Nundy S, Cooper LA, Mate KS. The quintuple aim for health care improvement: a new imperative to advance health equity. JAMA 2022 Feb 8;327(6):521-522. [doi: 10.1001/jama.2021.25181] [Medline: 35061006]
3. Shanafelt TD, Dyrbye LN, Sinsky C, et al. Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. Mayo Clin Proc 2016 Jul;91(7):836-848. [doi: 10.1016/j.mayocp.2016.05.007] [Medline: 27313121]
4. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open 2024 Oct 1;7(10):e2440969. [doi: 10.1001/jamanetworkopen.2024.40969] [Medline: 39466245]
5. Everett SS, Bunning BJ, Jain P, et al. From tool to teammate in a randomized controlled trial of clinician-AI collaborative workflows for diagnosis. NPJ Digit Med 2026 Mar 18. [doi: 10.1038/s41746-026-02545-1] [Medline: 41851268]
6. Rotenstein LS, Holmgren AJ, Thombley R, et al. Changes in clinician time expenditure and visit quantity with adoption of artificial intelligence-powered scribes: a multisite study. JAMA 2026 Apr 28;335(16):1408-1417. [doi: 10.1001/jama.2026.2253] [Medline: 41920565]
7. Skivington K, Matthews L, Simpson SA, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061. [doi: 10.1136/bmj.n2061] [Medline: 34593508]
8. Glasgow RE, Trinkley KE, Ford B, Rabin BA. The application and evolution of the practical, robust implementation and sustainability model (PRISM): history and Innovations. Glob Implement Res Appl 2024;4(4):404-420. [doi: 10.1007/s43477-024-00134-6] [Medline: 39568619]
9. Maw AM, Trinkley KE, Glasgow RE. The role of pragmatic implementation science methods in achieving equitable and effective use of artificial intelligence in healthcare. J Gen Intern Med 2024 May;39(7):1242-1244. [doi: 10.1007/s11606-023-08580-y] [Medline: 38172408]
10. Burden M, Dyrbye L. Evidence-based work design - bridging the divide. N Engl J Med 2025 Mar 13;392(11):1044-1046. [doi: 10.1056/NEJMp2412389] [Medline: 40062670]

Abbreviations

AI: artificial intelligence
EHR : electronic health record
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
PRISM: pragmatic robust implementation and sustainability model

Edited by T Leung; submitted 06.Apr.2026; this is a non–peer-reviewed article;accepted 27.Apr.2026; published 01.Jun.2026.

Please cite as:
Maw A, Pandita A, Burden M
Hospitalists Are Already Using AI—Why Implementation Will Determine Its Impact
J Med Internet Res 2026;28:e97419
URL: https://www.jmir.org/2026/1/e97419
doi:10.2196/97419

XSL•FO
RenderX

© Anna Maw, Aakriti Pandita, Marisha Burden. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 1.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Beyond Time Saved: Implementation, Equity, and the Utility Threshold for Nursing AI Scribes

Charlene E Ronquillo, MSN, PhD, RN

School of Nursing, Faculty of Health and Social Development, University of British Columbia, Okanagan Campus, 1147 Research Road, Kelowna, BC, Canada

Corresponding Author:
Charlene E Ronquillo, MSN, PhD, RN
School of Nursing, Faculty of Health and Social Development, University of British Columbia, Okanagan Campus, 1147 Research Road, Kelowna, BC, Canada

Abstract

Schwabe et al's pre-post time-motion study of a domain-specific artificial intelligence (AI) speech assistant used by nurses in German long-term care provides one of the few real-world, full-shift evaluations of an AI scribe deployed to a nonphysician workforce, with paired objective observation and self-reported outcomes. This commentary points to the implications of these findings that extend well beyond the time savings headline. The study reports substantial reduction in self-reported documentation time and increased satisfaction with the documentation system, yet workplace satisfaction and the perception that AI scribes are "a good idea to implement" did not improve. Taken together, these findings show three undertheorized issues for AI scribe implementation in nursing and long-term care. First, postimplementation increases in time spent reviewing entries and retrieving information indicate that AI scribes redistribute cognitive effort from authoring to verification, with unknown consequences for satisfaction, mastery, and error detection. Second, the apparent paradox of rising documentation satisfaction alongside falling expectations of AI quality represents user calibration. Third, the substantial equity considerations of automatic speech recognition documentation reflect a broader trend of AI scribe studies that treat equity as a caveat, rather than treating equitable performance as empirically measurable and testable across variations in linguistic styles, dialects, and social linguistic dimensions. To advance the field, the next generation of nursing AI scribe research must treat documentation as a heterogeneous bundle of authoring, reviewing, retrieving, and verifying activities with distinct satisfaction and error profiles; specify and validate end-user–defined anchor utilities, rather than having a narrow focus on diffuse improvement; and treat equity testing and reporting of both automatic speech recognition systems and workforce adoption as standard reporting expectations, rather than caveats.

(J Med Internet Res 2026;28:e101190) doi:10.2196/101190

KEYWORDS

artificial intelligence; long-term care; nursing informatics; speech recognition; health equity; digital health

Schwabe et al's [1] study is among the few carefully designed real-world evaluations of an automatic speech recognition (ASR) artificial intelligence (AI) scribe in nursing to date, a domain substantially underevaluated relative to physician documentation [2]. The full-shift, pre-post time-motion design and the choice of a domain-specific speech recognition system trained on nursing language are methodological advances over the small physician-focused pilots that dominate the AI scribe literature [2]. Like much of that literature, Schwabe et al [1] is a vendor-conducted evaluation: the first author and all observers were GmbH employees, developers of the ASR AI scribe used in the study. The methodological advance credited here lies in evaluation setting, duration, and analytic transparency, not independence from the vendor. In this commentary, attention is pointed to three findings that generate research questions the field has yet to substantively engage: (1) documentation work is reshaped, not only reduced; (2) the paradox of rising satisfaction alongside falling quality expectations can be interpreted as a utility threshold; and (3) the equity considerations of ASR-based documentation are treated as a caveat rather than empirically tested.

Documentation Work Is Reshaped, Not Only Reduced

The reduction in long-term care nurses' self-reported documentation time (Δ=−31.14 min) is accompanied by an acknowledged but less-discussed compositional change from authoring toward verification and information retrieval: time spent reviewing entries rose from 0.82 to 1.47 minutes, and information retrieval rose from 3.48 to 4.85 minutes [1]. The authors note the inability to directly link reduced documentation time to changes in documentation quality. This surfaces an open empirical question that the existing AI scribe literature has rarely disaggregated [2]: whether the review and verification of ASR-produced documentation produces the same satisfaction, mastery, and confidence as clinician authoring, or whether it changes error detection, omission, or copy-forward propagation. The next generation of evaluation should treat documentation

XSL•FO
RenderX

as a heterogeneous bundle of authoring, reviewing, retrieving, and verifying activities, each potentially carrying distinct satisfaction and error profiles.

The Utility Threshold and the "Good Idea" Paradox

Nurses' expectations that AI scribe use would improve documentation quality were lower post test than at baseline, even as satisfaction with the documentation system itself rose, and perceptions that it was a "good idea to implement" slightly decreased [1]. Building on technology acceptance literature [3], this pattern can be interpreted as signaling calibration rather than disappointment, from which the construct of a *utility threshold* emerges: the minimum demonstrable day-to-day clinical value an AI scribe must produce for nurses to endorse its sustained use. Perhaps counter to developer perceptions, the utility threshold may not be for an AI scribe to "do everything well" but "do at least one thing well enough that I notice it in my shift." The reduction in overall documentation time per shift may indeed be sufficient for clinicians to continue using it, even where documentation quality is not markedly improved. The findings on reduced daily interruptions deserve separate emphasis: interruptions to nursing work are associated with loss of concentration and focus, delays in planned tasks, and incomplete work, and are major contributors to medication errors [4,5]. Any tool that meaningfully reduces interruptions warrants a dedicated patient safety study. Finally, the privacy and confidentiality work that nurses absorb when dictating at the bedside, in corridors, and in shared spaces was not addressed in the study and offers one explanation for why "good idea" perceptions did not rise alongside other satisfaction measures. AI scribe design should validate explicitly identified anchor utilities such as overall time savings, reduced interruptions, privacy trade-offs, or yet-to-be-identified gains, rather than diffuse improvement. The greatest promise of clinical benefit will arguably stem from these anchor utilities being defined by end users.

Equity as an Underexamined Dimension

The equity stakes of ASR-based documentation are substantial but underexplored. In a US home health care setting, an evaluation of four ASR systems (two commercial: AWS General, AWS Medical; two open-source: Whisper, Wav2Vec 2.0) against gold-standard transcriptions of patient-nurse encounters found that all systems performed significantly worse for Black patients than White patients, with the largest discrepancies in affective and social linguistic dimensions, particularly relevant for nursing assessment of social cues [6]. The mechanism is structural. Training data underrepresent dialect and register variation [6], resulting in a documented pattern across major commercial ASR systems with differences in average word error rates among speakers of different racial backgrounds [7]. It is useful to distinguish speaker-level ASR bias (dialect, accent, vocal aging) from system-level training bias (sampled at training): speaker-level bias can sometimes be mitigated through fine-tuning, whereas system-level bias requires upstream intervention beyond downstream implementers' reach. Acknowledgment by Schwabe et al [1] of differential performance across dialects, accents, and terminology, while a good start, illustrates a pattern among AI scribe studies that treat equity as a caveat rather than an empirically testable and measurable dimension of technical performance. For example, the German nursing workforce in the study [1] includes regional dialect, generational, and language-of-origin variation that can serve as equitable ASR performance axes for empirical testing and measurement. A second equity dimension lies at the intersection of workforce characteristics and AI scribe adoption. The authors acknowledge that continued tool engagement may differ by career stage, training status, and role-related workload [1], and can extend to age, language-of-origin, prior digital health experience, and other established predictors of technology acceptance. Addressing the ASR equity gap rather than treating it as an afterthought requires conducting subgroup analyses by context-specific axes of linguistic variation, including accent, dialect, and professional characteristics as suggested [1], and disaggregated reporting on adoption, dropout, and equity metrics of ASR performance.

A Research Agenda

Table 1 summarizes research priorities following each thread. The Schwabe et al [1] study, despite limitations, contributes what hopefully becomes a precedent for stronger designs and longer time horizons of AI scribe and nursing AI studies. While the 8-week pilot is relatively brief, it offers important insight into implementation factors that determine sustained use, including workflow fit, staffing stability, and organizational readiness [8]. The time-motion methodology and pairing of objective and self-reported outcomes are the kind of evidence nursing AI scribe research needs more of. The implementation, equity, and utility-specification work, however, remains.

XSL•FO
RenderX

Table . Schwabe et al [1] findings, commentary interpretation, and research priorities for nursing artificial intelligence scribe research.

Schwabe et al [1] finding	Commentary interpretation	Research priority
↓ Self-reported documentation time (Δ=–31.14, SE 6.57 min)	Time savings real; sustainability beyond 8 weeks unknown	Longitudinal (≥12 mo) implementation studies anchored in implementation science framework (eg, CFIR[a] inner-setting constructs [8,9])
↑ Reviewing entries (0.82 → 1.47 min); ↑ information retrieval (3.48 → 4.85 min)	Clinical documentation task composition shifts from authoring to review and verification	Disaggregate documentation as a heterogeneous bundle of activities; examine satisfaction, mastery, and error implications of each
↑ Documentation satisfaction; ↓ quality expectations; "good idea" not improved	Utility threshold reached; unaddressed privacy and confidentiality labor a possible explanation for stagnant "good idea" perceptions	Specify and validate end-user–defined anchor utilities; investigate privacy and confidentiality protocols for bedside dictation
↓ Daily interruptions	Patient safety relevance, not workflow nicety [4,5]	Test whether interruption reductions translate into measurable reductions in nursing error, omission, or delay
Workplace satisfaction and "good idea to implement" unchanged	Adoption gains ≠ workforce benefit; possible utility threshold not yet reached for sustained use	Mechanism studies of what drives sustained use vs initial adoption
ASR[b] equity not empirically tested	Bias likely along regional, generational, language-of-origin axes [6,7]	Equity audits as standard reporting: ASR performance disaggregated by context- and domain-relevant patient/speaker characteristics; adoption disaggregated by workforce demographic

[a]CFIR: Consolidated Framework for Implementation Research.

[b]ASR: automatic speech recognition.

Acknowledgments

During the preparation of this work, the author used Claude Sonnet 4.6 to identify and review relevant literature and make the text more concise and readable. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Funding

The author declares that no financial support was received for this work.

Conflicts of Interest

None declared.

References

1. Schwabe K, Ferizaj D, Neumann S, Strube-Lahmann S, Lahmann N. Time savings through an AI speech assistant for nursing documentation: pre-post time-motion study in German long-term care. J Med Internet Res 2026 Apr 8;28:e86078. [doi: 10.2196/86078] [Medline: 41950503]
2. Kanaparthy NS, Villuendas-Rey Y, Bakare T, et al. Real-world evidence synthesis of digital scribes using ambient listening and generative artificial intelligence for clinician documentation workflows: rapid review. JMIR AI 2025 Oct 10;4:e76743. [doi: 10.2196/76743] [Medline: 41071988]
3. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
4. McGillis Hall L, Pedersen C, Hubley P, et al. Interruptions and pediatric patient safety. J Pediatr Nurs 2010 Jun;25(3):167-175. [doi: 10.1016/j.pedn.2008.09.005] [Medline: 20430277]
5. Stratton KM, Blegen MA, Pepper G, Vaughn T. Reporting of medication errors by pediatric nurses. J Pediatr Nurs 2004 Dec;19(6):385-392. [doi: 10.1016/j.pedn.2004.11.007] [Medline: 15637579]
6. Zolnoori M, Vergez S, Xu Z, et al. Decoding disparities: evaluating automatic speech recognition system performance in transcribing Black and White patient verbal communication with nurses in home healthcare. JAMIA Open 2024 Dec 10;7(4):ooae130. [doi: 10.1093/jamiaopen/ooae130] [Medline: 39659993]
7. Koenecke A, Nam A, Lake E, et al. Racial disparities in automated speech recognition. Proc Natl Acad Sci U S A 2020 Apr 7;117(14):7684-7689. [doi: 10.1073/pnas.1915768117] [Medline: 32205437]

XSL•FO
RenderX

8. Damschroder LJ, Reardon CM, Opra Widerquist MA, Lowery J. Conceptualizing outcomes for use with the Consolidated Framework for Implementation Research (CFIR): the CFIR Outcomes Addendum. Implement Sci 2022 Jan 22;17(1):7. [doi: 10.1186/s13012-021-01181-5] [Medline: 35065675]
9. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated Consolidated Framework for Implementation Research based on user feedback. Implement Sci 2022 Oct 29;17(1):75. [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]

Abbreviations

AI: artificial intelligence
ASR: automatic speech recognition

Edited by S Law, T Leung; submitted 12.May.2026; this is a non–peer-reviewed article;accepted 13.May.2026; published 27.May.2026.

Please cite as:
Ronquillo CE
Beyond Time Saved: Implementation, Equity, and the Utility Threshold for Nursing AI Scribes
J Med Internet Res 2026;28:e101190
URL: https://www.jmir.org/2026/1/e101190
doi:10.2196/101190

© Charlene E Ronquillo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

From Design to Accountable Impact for Data Dashboards in Health Care

Barbara-Jo Achuff, MD

Department of Pediatrics, Department of Biomedical Informatics, Vanderbilt School of Medicine, Vanderbilt University Medical Center, 2200 Children's Way, 5121 DOT, Nashville, TN, United States

Corresponding Author:
Barbara-Jo Achuff, MD
Department of Pediatrics, Department of Biomedical Informatics, Vanderbilt School of Medicine, Vanderbilt University Medical Center, 2200 Children's Way, 5121 DOT, Nashville, TN, United States

Related Article:

https://www.jmir.org/2026/1/e77361

Abstract

Vornhagen and colleagues synthesize design practices for the development of health care dashboards and provide a timely reference for a rapidly expanding class of tools that increasingly mediate clinical decisions and quality improvement. The authors have defined 4 pillars of design with associated practices, establishing a practical approach to the thoughtful development of useful dashboards. Building on these pillars, this commentary proposes accountable dashboarding, defined as making explicit the causal chain linking data sources and governance to visualization, interpretation, action, and measurable outcomes. Dashboards could be treated as sociotechnical interventions in which upstream choices are inseparable from the user interface and downstream decision-making. Evaluation should extend beyond usability and satisfaction to include decision quality and behavioral proxies, unintended consequences, and patient-centered outcomes of dashboard-informed interventions.

(J Med Internet Res 2026;28:e98272) doi:10.2196/98272

KEYWORDS

health care dashboards; data visualization; visual analytics; clinical decision support; workflow integration; data governance; data quality; implementation science; quality improvement

Design Practices for Data Dashboards

The scoping review by Vornhagen and colleagues [1] is a welcome synthesis of dashboard design practices in health care, drawing together guidance scattered across visualization, human factors, clinical informatics, and quality improvement literature. Their work is timely: dashboards now span bedside decision support, unit-level operations, and longitudinal quality surveillance, often built ad hoc under real-world constraints and evaluated inconsistently.

At the same time, the heterogeneity captured in a scoping review can be a double-edged sword. Implementers may adopt design practices as a checklist of features without clarity about which practices matter for which settings, workflows, and outcomes. As dashboards increasingly influence clinical decisions and resource allocation, the field should move toward accountable dashboarding, where dashboards are designed, implemented, and reported as interventions with explicit mechanisms and measurable effects.

Design Is Upstream of the Interface: Dashboards Are Sociotechnical Systems

Vornhagen and colleagues [1] describe themes emerging from the review's secondary analysis and group them into pillars. The first, pillar 1, includes the design approach and planning work that precedes dashboard building and deployment. A dashboard's most consequential design choices often occur before the interface is built and are intimately tied to the other themes, or pillars. Metric definitions, data provenance, missingness, refresh cadence, and transformation logic determine whether clinicians trust what they see and whether the dashboard fits the decision tempo of care. In practice, users experience these upstream elements as part of the product because they determine timeliness, stability, and credibility, core attributes of Tufte's "clear thinking made visible" in information design [2].

Building on this first pillar, this commentary suggests that accountable dashboarding requires describing sociotechnical components as first-class design elements, including metric governance and versioning, which concerns who owns measures and how changes are updated and communicated (ie, control

XSL•FO
RenderX

and governance); data provenance and quality, which encompasses the source systems that feed the dashboard, known gaps, and the handling of missing or delayed data (ie, truth and veracity); alingment of refresh cadence to decision moments (ie, fidelity and alignment); and workflow integration and role specificity, which refers to where dashboarding is used (eg, prerounds, rounds, huddles, or quality improvement meetings) and what actions should follow (ie, usability, orchestration, and integration). These are not merely technical details; they can shape behavior, adoption, and sustainability and are essential for others to interpret results and reproduce success.

Linking Design to Outcomes: Evaluation Should Match the Stakes

Vornhagen and colleagues [1] highlight evaluation approaches found in the literature and note that dashboard evaluation too often ends at usability or satisfaction. Those outcomes matter but may be insufficient for tools intended to change behavior and improve outcomes; usability principles remain necessary but not definitive evidence of effectiveness [3]. A usable dashboard can be clinically inert; conversely, a clinically valuable dashboard can fail if it increases workload or creates new sources of error.

A practical next step could include a required simple logic model that follows an evaluation chain of (1) data, (2) visualization, (3) interpretation, (4) decision and action, and (5) outcome. At a minimum, evaluation reports need to be developed and should include adoption and reach (who used it, how often, and in which workflow moments); behavioral proxies (changes in ordering, documentation, protocol adherence, or other care processes); workload and unintended consequences (time cost, duplication, workarounds); outcomes (patient-centered outcomes where feasible, otherwise tightly linked intermediate outcomes); and sustainability (maintenance burden and performance over time, including measure drift).

Numerous methods exist to measure and improve health care quality, safety, and value. By providing a structured framework, implementation science and quality improvement reporting guidelines enhance the rigor of these efforts without requiring new trials in every setting [4-6].

Bedside and Longitudinal Value: Lessons From a Pediatric Intensive Care Unit

In our tertiary academic pediatric intensive care unit, the need arose to make sedation medication exposure visible at both the bedside and unit level because existing electronic record views were difficult to interpret longitudinally and did not reliably support practice standardization. A dashboard that refreshed daily was developed with multidisciplinary stewardship using electronic medication administration record data to visualize sedation exposure for intensive care unit patients in two ways: (1) patient-level longitudinal profiles across hospitalization and/or postoperative day, and (2) aggregate unit-level trends over months for quality metrics. To make data interpretable and comparable, upstream transformation logic (eg, conversion to standardized opioid and benzodiazepine equivalents and calculation rules for continuous infusions) was treated as a core design component.

In practice, the dashboard enabled the identification of improvable patterns at scale, including dosing variation and spikes at specific hours—signals difficult to detect using static reports or manual review—and supported integration into bedside rounds and ordering behaviors. The dashboard was also used to monitor practice after the implementation of standardized sedation assessment and education. Over time, benzodiazepine exposure trended downward with attenuation of day-and-night variation. Although these changes occurred within a multicomponent improvement effort, the dashboard functioned as both a diagnostic tool (identifying targets) and a sustainability tool (monitoring after change). We identified that a more frequent refresh could clinically support more timely clinical discussions and action. This dual role, supporting bedside discussion while enabling longitudinal surveillance, may be underemphasized in the dashboard design literature and deserves explicit reporting and evaluation.

Equity and Accessibility Should Be Baseline Requirements

A persistent gap in dashboard development is that equity and accessibility are treated as optional enhancements rather than requirements. Dashboards influence attention, prioritization, and standardization, and can differentially affect patient subgroups and clinical roles. Three baseline commitments are pragmatic and feasible: accessibility by default using design informed by the Web Content Accessibility Guidelines (contrast, typography, and color choices accommodating color vision deficiencies) [7], role-specific usability testing across nurses, pharmacists, trainees, and attendings, and lastly, equity-sensitive evaluation to assess whether dashboard-driven metrics or interventions perform differently across patient subgroups or reflect documentation bias [8].

Extending the Pillars

Vornhagen and colleagues [1] should be congratulated for providing a strong foundation in their scoping review. To enhance generalizability, dashboard research and implementation reports should add three expectations: (1) publish the logic model and the moment of use, defined as when and where in the real workflow the dashboard is accessed and what action it is intended to trigger; (2) report governance and data-quality processes as part of design (sources, refresh cadence, missingness handling, and metric ownership and versioning); and (3) adopt a minimum evaluation and reporting set combining usability with adoption, workflow burden, unintended consequences, sustainability, outcomes, and equity and accessibility checks.

Dashboards are now part of the clinical infrastructure. The goal is not only to design them well, but to make mechanisms and impacts knowable so others can reproduce benefits, avoid predictable failures, and ensure improvements are trustworthy and equitable.

XSL•FO
RenderX

Acknowledgments

The author declares the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following task was delegated to GAI tools under full human supervision: publication support, including proofreading, editing, and reformatting. The GAI tool used was GPT-5.2. The tool was used solely to assist with rewriting text for clarity and consistency and providing clarification. GAI was not used for ideation or interpretation. All content generated with artificial intelligence assistance was reviewed, edited, and approved by the author. Responsibility for the final manuscript lies entirely with the author. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

The author declares no financial support was received for this work.

Conflicts of Interest

None declared.

References

1. Vornhagen H, Barrett S, Carroll C, et al. Design practices for data dashboards in health care: scoping review. J Med Internet Res 2026 Feb 25;28:e77361. [doi: 10.2196/77361] [Medline: 41740151]
2. Tufte ER. The Visual Display of Quantitative Information, 2nd edition: Graphics Press; 2001.
3. Nielsen J. Usability Engineering: Morgan Kaufmann; 1994.
4. Ogrinc G, Davies L, Goodman D, Batalden P, Davidoff F, Stevens D. SQUIRE 2.0 (Standards for Quality Improvement Reporting Excellence): revised publication guidelines from a detailed consensus process. BMJ Qual Saf 2016 Dec;25(12):986-992. [doi: 10.1136/bmjqs-2015-004411] [Medline: 26369893]
5. Bernal JL, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol 2017 Feb 1;46(1):348-355. [doi: 10.1093/ije/dyw098] [Medline: 27283160]
6. Carayon P, Schoofs Hundt A, Karsh BT, et al. Work system design for patient safety: the SEIPS model. Qual Saf Health Care 2006 Dec;15 Suppl 1(Suppl 1):i50-i58. [doi: 10.1136/qshc.2005.015842] [Medline: 17142610]
7. Web Content Accessibility Guidelines (WCAG) 22. W3C. 2024. URL: https://www.w3.org/TR/WCAG22/ [accessed 2026-05-14]
8. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]

Edited by S Law, T Leung; submitted 15.Apr.2026; this is a non–peer-reviewed article;accepted 30.Apr.2026; published 25.May.2026.

Please cite as:
Achuff BJ
From Design to Accountable Impact for Data Dashboards in Health Care
J Med Internet Res 2026;28:e98272
URL: https://www.jmir.org/2026/1/e98272
doi:10.2196/98272

© Barbara-Jo Achuff. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Health Interventions for Family Members of ICU Patients: Current Evidence and Future Directions

Tanya Abenaim[1]; Dan Poenaru[2], MA, MD, PhD; Michael Goldfarb[3], MD, MSc

[1]Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada

[2]Division of Pediatric Surgery, Montreal Children's Hospital, Montreal, QC, Canada

[3]Division of Cardiology, Jewish General Hospital, McGill University, Office E-212, 3755 Cote Ste Catherine Road, Montreal, QC, Canada

Corresponding Author:
Michael Goldfarb, MD, MSc
Division of Cardiology, Jewish General Hospital, McGill University, Office E-212, 3755 Cote Ste Catherine Road, Montreal, QC, Canada

Related Article:

https://www.jmir.org/2026/1/e83294

Abstract

A recent systematic review and meta-analysis of randomized trials evaluating digital health interventions for family members of intensive care unit (ICU) patients found no significant improvements in anxiety, depression, posttraumatic stress, quality of life, or communication quality. Rather than concluding that digital approaches are inherently ineffective, we argue that these null findings reflect identifiable and remediable limitations in intervention design, outcome measurement, and trial methodology. In this commentary, we examine four structural barriers that currently constrain the evidence base and outline the conditions that next-generation trials must meet to adequately address the questions raised by this review.

(J Med Internet Res 2026;28:e95876) doi:10.2196/95876

KEYWORDS

family engagement; digital health interventions; intensive care units; post-intensive care syndrome-family; PICS-F

Intensive care admissions rarely affect only the patient. Family members of critically ill patients also carry a substantial and well-documented psychological burden. Post–intensive care syndrome–family (PICS-F) affects a significant proportion of relatives of patients in the intensive care unit (ICU), with symptoms such as anxiety, depression, and posttraumatic stress that can persist well beyond ICU discharge. Distressed family members may be less able to provide the emotional and practical support that facilitates patient recovery, making PICS-F not only a humanitarian concern but also an important factor in patient-outcomes.

Against this backdrop, digital health interventions (DHIs) have attracted substantial interest as scalable solutions—encompassing mobile apps, web-based platforms, videoconferencing, virtual reality, and internet-delivered cognitive behavioral therapy. Zhang and colleagues [1] have now published a systematic review with meta-analysis of randomized controlled trials (RCTs) evaluating DHIs specifically for family members of ICU patients, synthesizing 17 trials and 1864 participants across three continents. No statistically significant effects were found for any of the primary outcomes: anxiety, depression, posttraumatic stress disorder (PTSD), quality of life, or communication quality. Wide prediction intervals and substantial statistical heterogeneity indicated marked variability in effects across settings. However, those intervals warrant closer examination than a simple null verdict would suggest.

Interpreting the Null Result

A null meta-analytic result is neither proof of no effect nor license for the field to continue unchanged. Both interpretations remain plausible: that DHIs are conceptually sound but poorly executed and that they are genuinely ineffective for this population. The available evidence cannot adjudicate between them. What it does allow is a clearer diagnosis of why the evidence remains uninformative.

The 95% prediction interval for anxiety (−1.46 to 0.79) is the most consequential figure in the paper. Its upper boundary indicates the possibility of clinically meaningful harm. However, Zhang et al [1] report that no adverse outcomes were prospectively monitored across any of the 17 included trials—a striking omission given that all intervened in a population already at elevated risk for PTSD, depression, and complicated grief. The same adverse-event reporting standards applied to pharmaceutical RCTs should apply here. An undetected signal cannot be assumed benign; it may be neutral, heterogeneous

XSL•FO
RenderX

across subgroups, or—in a subset of vulnerable families—harmful.

The heterogeneity is not merely statistical. A 14-minute virtual reality exposure [2] is being pooled with a 5-week internet-based cognitive behavioral writing program [3]. These are fundamentally different interventions, illustrating the absence of a coherent intervention taxonomy. This is not a criticism of the meta-analysis itself, but a call for theory-driven subgroup classification before the next generation of trials is designed.

Four Structural Barriers to Conclusive Evidence

Taken together, these issues point to four structural barriers that currently prevent the field from generating conclusive evidence.

Barrier 1: Underdeveloped Intervention Theory

Most included DHIs were information-centric: unidirectional, static, and content-driven. Qualitative evidence consistently attributes PICS-F persistence to uncertainty, helplessness, traumatic witnessing, and disrupted attachment [4]—psychological processes for which passive information delivery is likely insufficient. The gap between what families experience and what current DHIs address is not incidental; it is a design failure that must be corrected before outcome data can be meaningfully interpreted.

Barrier 2: Mismatch Between Measurement and mechanism

Before the next generation of trials is designed, the field requires a clearer account of what DHIs are hypothesized to change and through what pathway. For instance, *reduced anxiety at 12 weeks* is an outcome, not a mechanism; it cannot distinguish whether an intervention succeeded by reducing informational uncertainty, increasing perceived control, facilitating emotional processing, or simply providing structured human contact through a digital medium. Without mechanistic specificity, outcome selection remains arbitrary and follow-up windows cannot be meaningfully calibrated.

PTSD is a delayed-onset condition; its prevention requires follow-up windows matched to its natural history, not 3 months or fewer that most included studies employed. Notably, several trials reported positive subjective family feedback despite nonsignificant quantitative outcomes, a pattern Zhang et al interpret as suggesting that current instruments may lack the sensitivity to capture experiential gains that families themselves perceive. If families are experiencing meaningful benefit that validated scales cannot detect, that is a measurement failure as much as an intervention failure, and it strengthens the case for multimodal outcome assessment, including patient- and family-reported experience measures alongside standardized psychological instruments.

Barrier 3: Near-Total Absence of Engagement Data

Petrinec et al [5] reported a mean of 11.4 application logins and 50 minutes of total use across a study period, a dose unlikely to produce meaningful behavioral change. Whether current DHIs fall short because they are conceptually inadequate or simply not used enough cannot be determined from outcomes data alone. Answering this question requires dose-response analyses and engagement-stratified outcome reporting, neither of which was conducted in the included trials.

Barrier 4: Outcome Measure Heterogeneity

Across 17 included studies, Zhang et al identified 18 distinct validated instruments, rendering cross-trial comparisons interpretively fragile. However, no PICS-F-specific core outcome set initiative has yet been registered with Core Outcome Measures in Effectiveness Trials (COMET). developing one should be a methodological priority [6].

Requirements for the Next Generation of Trials

Future trials require interventions grounded in behavioral theory, delivered longitudinally, with pre-specified engagement thresholds, prospective adverse-event monitoring, and follow-up windows matched to the outcomes being targeted. Equity must be a design requirement, not an afterthought: older adults, non-English-speaking families, and low-income caregivers bear a disproportionate PICS-F burden yet face the greatest barriers to DHI access [7]. If equity considerations are not embedded in eligibility criteria and delivery design, effective interventions will reach only the most resourced families, deepening the disparities they might otherwise address.

Table 1 traces the evolution of DHI design toward a third generation defined by artificial intelligence (AI)–driven personalization, adaptive delivery, and bidirectional clinician–family communication. These tools represent a genuine conceptual advance over what has been tested—but not yet a proven one. The Society of Critical Care Medicine has articulated the case for family-centered ICU care [8]; building the evidence base to deliver it is the work that remains.

XSL•FO
RenderX

Table . Evolution of digital health intervention design for families of ICU patients.

Generation	Characteristics	Examples	Limitations
First generation	• Static information delivery • One-way communication	• Educational websites • SMS updates	• Low engagement • Generic content
Second generation	• Interactive platforms • Basic personalization	• Mobile apps with feedback forms • Video calls	• Limited artificial intelligence (AI) integration • Reactive rather than proactive
Third generation (emerging)	• AI–driven personalization • Predictive analytics • Real-time adaptation • Culturally adaptive content • Integration with clinical workflows	• Chatbots with natural language processing • Adaptive content delivery	• Nascent evidence base • Requires rigorous evaluation before adoption

What Zhang et al actually demonstrate is not that DHIs for ICU families do not work. It is that we do not yet know whether they work, in whom, at what dose, or through what mechanism. Answering these questions will require a fundamentally different approach to trial design than the field has so far adopted.

Funding

The authors declare no financial support was received for this work.

Conflicts of Interest

None declared.

References

1. Zhang C, Sui W, Jiang W, et al. The effects of digital health interventions for family members in intensive care units: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2026 Feb 25;28:e83294. [doi: 10.2196/83294] [Medline: 41740153]
2. Drop DLQ, Vlake JH, Wils EJ, et al. Effect of an intensive care unit virtual reality intervention on relatives´ mental health distress: a multicenter, randomized controlled trial. Crit Care 2025 Feb 5;29(1):62. [doi: 10.1186/s13054-025-05281-2] [Medline: 39910674]
3. Gawlytta R, Kesselmeier M, Scherag A, et al. Internet-based cognitive-behavioural writing therapy for reducing post-traumatic stress after severe sepsis in patients and their spouses (REPAIR): results of a randomised-controlled trial. BMJ Open 2022 Mar 9;12(3):e050305. [doi: 10.1136/bmjopen-2021-050305] [Medline: 35264337]
4. Kean S, Donaghy E, Bancroft A, Clegg G, Rodgers S. Theorising survivorship after intensive care: A systematic review of patient and family experiences. J Clin Nurs 2021 Sep;30(17-18):2584-2610. [doi: 10.1111/jocn.15766] [Medline: 33829568]
5. Petrinec AB, Wilk C, Hughes JW, Zullo MD, George RL. Self-care mental health app intervention for post-intensive care syndrome-family: a randomized pilot study. Am J Crit Care 2023 Nov 1;32(6):440-448. [doi: 10.4037/ajcc2023800] [Medline: 37907376]
6. Core outcome measures in effectiveness trials. COMET Initiative. URL: https://www.comet-initiative.org/ [accessed 2026-03-25]
7. Lam K, Lu AD, Shi Y, Covinsky KE. Assessing telemedicine unreadiness among older adults in the united states during the COVID-19 pandemic. JAMA Intern Med 2020 Oct 1;180(10):1389-1391. [doi: 10.1001/jamainternmed.2020.2671] [Medline: 32744593]
8. Hwang DY, Oczkowski SJW, Lewis K, et al. Society of critical care medicine guidelines on family-centered care for adult ICUs: 2024. Crit Care Med 2025 Feb 1;53(2):e465-e482. [doi: 10.1097/CCM.0000000000006549] [Medline: 39982184]

Abbreviations

AI: aritfiical intelligence
DHI: digital health intervention
ICU: intensive care unit
PICS-F: post–intensive care syndrome–family
PTSD: posttraumatic stress disorder

XSL•FO
RenderX

RCT: randomized controlled trial

Edited by T Leung; submitted 22.Mar.2026; this is a non–peer-reviewed article;accepted 24.Mar.2026; published 30.Apr.2026.

Please cite as:
Abenaim T, Poenaru D, Goldfarb M
Digital Health Interventions for Family Members of ICU Patients: Current Evidence and Future Directions
J Med Internet Res 2026;28:e95876
URL: https://www.jmir.org/2026/1/e95876
doi:10.2196/95876

© Tanya Abenaim, Dan Poenaru, Michael Goldfarb. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Approaches to Collect Comprehensive Electronic Patient Data Across Multiple Providers and Payers for Research: Landscape Analysis

Robert S Rudin[1], PhD; Nihar Chhatiawala[2], PhD; Joshua C Mandel[3,4,5], MD; William J Gordon[5,6], MD, MBI; David McCallie[7], MD; Daniel Gottlieb[4,8], MPA

[1]Health Division, RAND Corporation, 20 Park Plaza, 9th Floor, Suite 910, Boston, MA, United States
[2]Health Division, RAND Corporation, Santa Monica, CA, United States
[3]Microsoft Research, Redmond, WA, United States
[4]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, United States
[5]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, United States
[6]Department of Medicine, Brigham and Women's Hospital, Boston, MA, United States
[7]Formerly with Cerner Corporation, Kansas, MO, United States
[8]Central Square Solutions, Boston, MA, United States

Corresponding Author:
Robert S Rudin, PhD
Health Division, RAND Corporation, 20 Park Plaza, 9th Floor, Suite 910, Boston, MA, United States

Abstract

Background: The digitization of medical data and advances in interoperability have opened opportunities for research studies to use more comprehensive, longitudinal patient data from multiple sources. As patients often interact with many providers and payers over time, collecting data across organizations may have critical implications for accuracy and bias in study results. US policy has promoted exchanging health information among providers, payers, and patients, but less attention has focused on facilitating data collection for research, which presents unique challenges.

Objective: This study aimed to identify and evaluate existing and emerging approaches for collecting comprehensive provider and payer data for research in the United States, with the goals of informing researchers of possible methods and generating evidence to inform policy initiatives. Our focus was on electronic approaches to data aggregation for studies requiring patient consent.

Methods: We conducted a landscape analysis through interviews with subject matter experts (SMEs). SMEs were selected based on expertise. We created a list of evaluation criteria, identified existing and emerging approaches, and described the benefits and limitations of each approach by applying the evaluation criteria. We interviewed SMEs until saturation was achieved. Data collection was limited to the United States.

Results: A total of 20 SMEs helped identify 8 distinct approaches: (1) general-purpose smartphone app, (2) commercial app, (3) research community app, (4) structured data export, (5) Trust Exchange Framework and Common Agreement Individual Access Service, (6) regional study query, (7) national study query, and (8) aggregated data source. Participant-mediated exchange approaches (1-5) leveraged patients' right of access. Three approaches leveraged existing data exchange services (5-7). To evaluate these approaches, we identified 12 criteria, including perspectives of participants, research teams, and broader stakeholders. Each approach had benefits and limitations; no single approach emerged as superior for all use cases. While currently available approaches for participant-mediated exchange bypass the need for complex governance arrangements, they are limited by participant burden, effort needed by research teams, and data gaps, especially from payers. Some regional health information exchanges and aggregated data sources address governance challenges and can provide services such as preparing analytic datasets but are restricted to specific locations and/or data-source coverage. National networks currently do not allow queries for research and confront challenges in establishing trust and enforcing compliance with data-sharing requirements among network sites.

Conclusions: Collecting comprehensive health data from multiple providers and payers in the United States is a complex and evolving process. The suitability of an approach may vary based on the needs of a study. Given the numerous barriers and the lack of a clear dominant method, further exploration and benchmark comparisons of all identified approaches are necessary. Ongoing public policy efforts will likely play an important role in progress.

(J Med Internet Res 2026;28:e86330) doi:10.2196/86330

XSL•FO
RenderX

KEYWORDS

research data; interoperability; health information policy; landscape analysis; subject matter expert; SME

Introduction

Overview

The large-scale digitization of medical data and widespread adoption of electronic health records (EHRs) in the United States provide researchers with the opportunity to use longitudinal data on participants from across their sites of care [1,2]. Yet, today, typical health care research efforts in the United States lack access to complete patient data [3-6]. US policy efforts, such as the 21st Century Cures Act and related regulations, have promoted approaches for exchanging health information among providers and payers, and with patients, including for purposes of facilitating data collection for research purposes [7-11]. To better understand existing and emerging approaches, we conducted a landscape analysis with the aim of informing researchers and policy efforts.

Background

The capacity to conduct studies that extend beyond a single institution and are not confined to a limited network of health care providers is crucial. Without access to comprehensive data, researchers may overlook important information regarding participants' prior health-related encounters, potentially resulting in biased outcomes [3-6]. The absence of complete data can also impede the feasibility of some kinds of studies, including those aimed at evaluating the long-term value of health care and those seeking a holistic understanding of participants' health [12].

Data for most patients are fragmented across multiple organizations. In the United States, for example, a typical Medicare beneficiary sees a median of 2 primary care physicians and 5 specialists working in 4 different practices [13]. Beneficiaries with multiple chronic conditions have even greater levels of fragmentation. Studies that recruit patients of only 1 primary care physician will find that, on average, every 100 of their Medicare patients receive care from 53 different practices [14]. The effort required to access and link these data can be prohibitively large, even for studies enrolling only a small number of patients. Payer data also require substantial effort to collect. Patients may change health plans during a study period—one study found that 21.5% of members turned over each year [15]—further adding to the fragmentation and effort needed to assemble complete data.

The goal of assembling longitudinal patient data has long been recognized [2,16]. However, previous attempts to achieve this goal in the United States have failed. For example, one effort to create a patient-controlled "health record bank" closed due to limited enrollment [17]. Some health information exchange (HIE) organizations have assembled data from multiple health care providers within geographic regions, but many face challenges with sustainability [18]. National data exchange networks focus primarily on clinical care; their potential to facilitate data collection for research has not been examined.

We interviewed experts to identify the range of possible approaches that research teams might use to collect multi-institutional, comprehensive provider and payer data on patients enrolled in studies and to evaluate their benefits and limitations. By assessing the current landscape, this work ultimately aims to help research teams select among the options available to them for collecting data on their studies' enrolled patients, help research teams prepare for emerging options, and inform efforts to further facilitate the complete collection of patient data for research.

Methods

Objective and Scope

Landscape analyses aim to comprehensively describe and map the current state of a domain of interest and provide an overview to inform decision-making [19-21]. This approach is particularly useful for rapidly evolving domains where formal literature may not reflect current practice or emerging innovations. This US landscape analysis had three objectives:

1. Identify evaluation criteria for assessing approaches that research teams might use to collect comprehensive provider and payer data on study participants.
2. Identify existing and emerging approaches to the problem.
3. Evaluate the benefits and limitations of the identified approaches by applying the evaluation criteria where applicable.

We focused on approaches that would support research studies in which patients provide consent for their data to be collected from multiple sources and used for research. This would include many kinds of clinical trials, implementation studies of clinical interventions, and observational studies. Approaches that do not require patient consent (eg, using data from a deidentified repository for population-level studies), involve a single source of data (eg, from one payer or one provider), or only make a narrow subset of the medical record data available (eg, clinical registries) were not included, as they would involve different research workflows.

We included health care providers and payers as data sources only because of their prevalent use in health care research studies; other sources (eg, wearables, social media, community-based organizations, and direct-to-consumer genetic testing services) were not included. We also excluded imaging data (eg, magnetic resonance imaging) because of differences in the technology and workflows needed to access these data compared with other data derived from health care encounters. We included approaches that addressed the following activities:

- Identifying the source and location of the data: determine where the participant previously received care or may have electronic health data available, including through patient self-identification, following chains of records (eg, from a payer to the providers that billed for care), and query tools that access metadata describing data sources (eg, name,

XSL•FO
RenderX

address, Fast Health Interoperability Resources [FHIR] endpoint) and match using demographic characteristics.
- Retrieving the data: collect electronic health data from the data source through manual or automated download processes and the use of specific software tools to facilitate the data collection.
- Sharing the data with research studies: make electronic health data available to research team members, including automated approaches and those that require an extra step on the part of an individual research participant.

Approaches to identify patients for eligibility or recruitment into studies required separate research workflows and were out of scope.

We limited our scope to electronic approaches, particularly those that retrieve data using application programming interfaces (APIs) such as FHIR, because of their potential to facilitate data collection; paper- or fax-based approaches were considered out of scope. Additionally, only approaches that were considered by the research team or experts to be reasonably feasible for implementation within the next few years were included.

Recruitment of Subject Matter Experts and Interviews

We identified subject matter experts (SMEs) through publications, recommendations from our professional network, and recommendations from other SMEs (ie, snowball sampling). We conducted each interview via videoconference and recorded all interviews, which lasted roughly 1 hour. If we had additional questions, we requested a second interview with the SME or exchanged follow-up emails. In some cases, 2 SMEs attended the same interview. Interviews were conducted between February 2024 and July 2025.

To prepare for the interviews, we prepared preliminary lists of approaches and evaluation criteria, and a table applying the evaluation criteria to each approach. The research team developed these based on its existing knowledge of the landscape. We also prepared an interview guide that covered our 3 objectives. Specifically, we asked SMEs to identify and describe in detail all existing and emerging approaches with which they were familiar for collecting provider and payer data for research, identify evaluation criteria for assessing these approaches, and evaluate the approaches based on the evaluation criteria. As SMEs had expertise with different approaches, we focused the discussion on approaches with which the SME was most familiar. For SMEs with broader knowledge of the landscape, we reviewed our preliminary lists of approaches and evaluation criteria, asked them to identify anything missing, and asked them about what policy or other health care system changes may influence the assessment of approaches. Finally, we asked the SMEs to identify relevant publications or other documents that should be reviewed to understand the evolving landscape of approaches. After each interview, we updated and refined our preliminary lists of approaches, evaluation criteria, and application of evaluation criteria to approaches. We continued interviewing SMEs until we achieved saturation in terms of evaluation criteria and approaches. We aimed to include at least 1 SME with detailed knowledge of each identified approach.

Analysis

We used conventional content analysis to analyze the qualitative data [22,23], inductively identifying approaches, evaluation criteria, and applications of evaluation criteria to the approaches. For each interview, one research team member summarized key points discussed and extracted relevant quotes, and a second research team member reviewed each recording and the summaries and notes (RSR and NC). Both research team members had domain experience with EHRs and digital health. Differences were resolved by consensus. To generate overall findings, after each interview, we updated a working matrix that mapped evaluation criteria against approaches, noting areas of agreement, disagreement, and uncertainty across SMEs. For evaluation criteria, we compiled all factors mentioned by SMEs, grouped conceptually similar factors, and refined descriptions through discussion until we achieved a comprehensive list, aiming to exhaustively cover all important factors for assessing the benefits and limitations of the various approaches. Evaluation criteria were not required to be mutually exclusive. For approaches, we created detailed written descriptions that included technical details, workflows, and perspectives from relevant interviews. We then applied each evaluation criterion to each approach, synthesizing assessments from SMEs. We noted where SMEs provided conflicting assessments or expressed a lack of sufficient evidence or experience to make an assessment.

To ensure rigor and trustworthiness of our findings, we used the following methods during the analysis: 2 researchers independently reviewed recordings and summaries, resolving discrepancies through consensus discussion; SMEs were selected to include a range of perspectives (eg, technology developers, policy experts, and researchers); we reviewed our preliminary findings with SMEs who had broad knowledge of the landscape; and we continued sampling until we achieved saturation. Additionally, all authors have relevant domain expertise in EHRs, interoperability, and digital health, and contributed to interpreting technical content and synthesizing results.

We strengthened the rigor of our findings and supplemented our interview data by incorporating 3 types of documents suggested by SMEs into this analysis. First, we reviewed technical documentation (eg, FHIR implementation guides and Trust Exchange Framework and Common Agreement [TEFCA] policies) to verify and elaborate on technical details described by SMEs. Second, we reviewed gray literature sources (eg, blog posts and white papers) to understand emerging perspectives. Third, we reviewed relevant peer-reviewed publications to contextualize SME assessments. These documents helped confirm factual claims and provide supplemental information.

Ethical Considerations

This study was reviewed and approved by both the RAND Human Subjects Protection Committee and the Mass General Brigham Institutional Review Board (IRB) as exempt. All written and audiovisual data were stored on secure servers accessible only to the research team. Verbal consent was obtained before conducting the interviews, which included permission for identifying the individual in the publication but not attributing any quote or information to any specific

XSL•FO
RenderX

individual. We chose to identify SMEs to enhance transparency and allow readers to evaluate the credibility and relevance of our sources [24]. Participants were not compensated.

Results

SME Characteristics

We interviewed 20 SMEs (Table 1). SME expertise covered a wide range of aspects relevant to this topic, including smartphone-based approaches, FHIR, and federal and state policy.

Table . Names and characteristics of subject matter experts interviewed in the study.

Subject matter expert	Title (at time of interview)
Ricky Bloomfield	Chief Medical Officer at ŌURA; formerly Clinical and Informatics Lead at Apple
Dan Drozd	Chief Medical Officer at PicnicHealth
Andrew Cress	Chief Executive Officer and Cofounder at HealthVerity
Andrew Goldberg	Chief Operating Officer and Cofounder at HealthVerity
Ryan Howells	Lead of CARIN Alliance; Principal at Leavitt Partners; digital health policy and interoperability expert
Jocelyn Keegan	Program Manager of the HL7 Da Vinci Project; Payer-Practice Lead at Point-of-Care Partners
Brenden Keeler	Head of Product at Flexpa
David Kendrick	Chief Executive Officer of MyHealth Access Network, the Oklahoma state-designated entity for health information exchange
Vik Kheterpal	Principal at Care Evolution, Inc
Mark Knee	Director, Interoperability Division, Office of the National Coordinator for Health Information Technology
Greg Liptak	Vice President for Research and Innovation at HealthShare Exchange, the regional health information exchange for greater Philadelphia
Joel Montavon	Clinical Informaticist at RTI International
Michele Mottini	FHIR[a] software developer at Care Evolution, Inc
Christopher Muir	Director, Network and Scalability Division, Office of the National Coordinator for Health Information Technology
JP Pollack	Cocreator of CommonHealth; Cofounder of The Commons Project Foundation
Molly Prieto	Deputy Director, Standards Division, Office of the National Coordinator for Health Information Technology
Marc Rabner	Chief Medical Officer of Chesapeake Regional Information System for Our Patients (CRISP) Shared Services
Raheel Sayed	Lead of the People Heart Study, which uses Apple Health
Jeff Smith	Deputy Director, Certification and Testing Division, Office of the National Coordinator for Health Information Technology
Mariann Yeager	Chief Executive Officer of the Sequoia Project, which supports TEFCA[b]

[a]FHIR: Fast Health Interoperability Resources.

[b]TEFCA: Trust Exchange Framework and Common Agreement.

Evaluation Criteria for Assessing Approaches

We achieved saturation by identifying 12 evaluation criteria (Table 2). We took an inclusive approach and included even those criteria that might be challenging to apply.

XSL•FO
RenderX

Table . Evaluation criteria for assessing approaches to collect comprehensive provider and payer data on study participants.

Evaluation criterion	Description
Reach and accessibility	Volume of participants whose data can be accessed (includes requirements for devices such as smartphones).
Data availability and quality	Extent to which records and their contents are available and usable for research (includes consideration of whether variables are in designated fields, completeness due to factors such as geographic coverage and accuracy).
Study team burden	Effort or cost required to implement and operate a fully developed approach (includes required technical support).
Feasibility	Effort or changes required to make the approach widely available (includes technical effort needed by stakeholders).
Participant burden	Effort or cost required of participants to use the approach.
Participant benefit	Potential benefits that might motivate use.
Security	Risk or perceived risk of breach.
Privacy	Nature of patient control over data.
Impact of industry trends	Potential for technology or other trends to affect the approach (eg, adoption of technical standards).
Sensitivity to changes	Potential for stakeholders to make decisions that affect the approach (eg, market changes in service providers).
Alignment and synergies with policies and policy goals	The extent to which the approach supports or is supported by policies and policy goals and lacks regulatory barriers.
Configurability and innovation enablement potential	Potential for the approach to foster innovation in research and practice (eg, customizability).

Evaluation of Approaches to Collect Comprehensive Data for Research

We identified 8 approaches, which achieved saturation within the scope of the study (Table 3 contains summary assessments; approaches are labeled 1 through 8; Multimedia Appendix 1 contains full application of all evaluation criteria to all approaches). Five are participant-mediated approaches (1-5), in which research participants are directly involved in retrieving their data. In the remaining approaches, which included study queries of existing data exchange infrastructure (6 and 7) and aggregated data sources (8), participants' roles were limited to providing consent. Many SMEs emphasized the benefits of participant-mediated approaches because they leverage patients' right of access to their own data. Four of those approaches required study participants to manually identify their sources of data and use the patient or member portals at each site (1-4). In contrast, the other approaches leveraged existing national qualified health information networks (QHINs), regional HIEs (also called regional health data utilities), or aggregated data sources to identify sites of care and pull data. For these latter approaches, SMEs stated that although providers and payers, as Health Insurance Portability and Accountability Act (HIPAA)–covered entities, are legally required to release data with patient consent, rigorous governance processes are often required because of concerns about liability from errors and the lack of financial incentives to streamline data sharing.

XSL•FO
RenderX

Table . Approaches that research teams might use to collect comprehensive provider and payer data on study participants and summary assessment.

Approach name	Description	Summary assessments[a]
Participant-mediated approaches		
General purpose smartphone app	Participants use Apple Health (iOS) or Common Health (Android) to collect data and share it manually or through a purpose-built study app	• Benefits: widely available as an installed app (Apple Health) or an installable app (CommonHealth); management of FHIR[b] endpoints is handled by the app. • Limitations: participant burden for manual data sharing or research team burden to create and maintain a study app, not all FHIR resources may be supported, gaps in FHIR endpoint coverage, potential perceived privacy concerns, limitations in customizability, and potential sampling bias due to smartphone requirement.
Commercial app	Study pays for an app allowing participants to collect and share data	• Benefits: management of the FHIR endpoints directory handled by the app, processing of data can minimize study team burden of analysis, additional features of value offered to patients, and some apps have the potential to be accessed from a web browser. • Limitations: cost to the research team, potential perceived privacy concerns, and dependency on the app company for customizations.
Research community app	A study or institution uses a homegrown or open-source app allowing participants to collect and share data	• Benefits: potentially lower cost for the research team, reduced perceived privacy concerns, potential for additional features of value to patients, and potential for access from a web browser. • Limitations: requires development effort to become widely available and is a burden on the study team or organization to operate and to manage the FHIR endpoint directory and client registrations.
Structured data export	Participants use EHI[c] export or a similar feature to collect data	• Benefits: may become widely available in the near future, likely accessible from any web browser, potential reduced perceived privacy concerns, and potentially be comprehensive in terms of data. • Limitations: not widely available today, burden on the participant to manually download and share data, burden on the study team to set up a process to receive data shared by the participant, and lack of data format standardization.
TEFCA[d] Individual Access Service	Participants access data directly from a national network (QHIN[e])	• Benefits: participant burden is minimal because they would need to use only one app or portal and is supported by current policies. • Limitations: not usable today, burden on the study team to set up a process to receive data shared by the participant, and participants may still need to log into sites of care.
Queries of data exchange infrastructure		

XSL•FO
RenderX

Approach name	Description	Summary assessments[a]
Regional study query through a HIE[f]	Participants consent for the study team to query an HIE	• Benefits: minimal or no burden on participants and processing of data can minimize the study team's burden of analysis. • Limitations: unavailable today in most regions, gaps in data coverage, especially among payer claims, and cost to the research team.
National study query through QHINs	Participants consent for the study team to query a QHIN	• Benefits: minimal or no burden on participants. • Limitations: research is not currently an authorized exchange purpose, and the timeline for adding it is uncertain (although it is aligned with policy goals); there may be gaps in QHIN coverage of sites, the burden on the study team is unknown, and potential perceived privacy concerns.
Use of aggregated data sources		
Aggregated data source	Participants consent for the study team to collect data from a company or organization that has already collected data from multiple sources	• Benefits: minimal or no burden on participants, data sources are nationwide, and processing of data can minimize the study team's burden of analysis. • Limitations: may not cover all data sources or elements needed for the study and cost to the research team.

[a]Summary assessment is based on the criteria listed in Table 2 and includes selected key findings.

[b]FHIR: Fast Health Interoperability Resources.

[c]EHI: electronic health information.

[d]TEFCA: Trust Exchange Framework and Common Agreement.

[e]QHIN: qualified health information network.

[f]HIE: health information exchange.

Approaches Using Patient and Member Portals

Three of the approaches we identified required study participants to access their patient and members' portals to retrieve data using Substitutable Medical Applications and Reusable Technologies (SMART) on FHIR (Table 3; approaches 1-3). To avoid repetition, we describe the assessment of this component here. Benefits of this approach include the high availability of patient and member portals. Patient portals have become increasingly used since the COVID-19 pandemic and are required to support patient access to data via SMART on FHIR, which addresses security concerns and allows patients some granular control over their data. Payer member portals are also widely available, although usage rates are not known. Additionally, these approaches align with federal policy goals. Facilitating patient access to their own health data has been a primary policy objective and a driver of certification programs and functionality for many years [25,26] and continues to be a priority [27]. Some SMEs pointed to more recent plans by Centers for Medicare & Medicaid Services (CMS) to advance digital identity solutions in health care and were optimistic that substantial progress toward the use of single sign-on (SSO) to patient portals, and potentially member portals, was occurring as complex issues such as price transparency and antitrust issues were being addressed [26,28].

There are 2 limitations to relying on patient and member portals for collecting data for research. First, the burden on study participants to log in to each portal is substantial. Patients may not remember prior locations of care, may have trouble finding login information, or lack portal access at some locations, and may be unwilling to spend time registering. This may result in incomplete data. However, adoption of SSO technologies could reduce this barrier substantially and is part of the CMS Interoperability Framework [29-31]. Additionally, studies may need to rely on participants for the timing of data retrieval, which may introduce variability.

Second, approaches that rely on patient and member portals must register with FHIR endpoints and establish a searchable directory for participants. The process of registering varies by EHR vendor or, in many cases, by individual site. One SME described the challenges: "It's kind of a nightmare. We rely on a combination of reaching out to providers where the data is really messy and communicating with the EHR vendors who are only partially motivated to have that information be correct." One SME estimated that "maybe about 10% of providers actually care about this… make sure that [metadata] is correct." A recent FHIR standard called "User-access Brands and Endpoint" is designed to help users locate their data sources but has not been widely adopted, although recent regulation requires certified EHR vendors to make endpoint lists available

XSL•FO
RenderX

in FHIR format [32,33]. Support from vendors and providers is also variable: "Sending email into the void is a common thing." Some vendors have needed to develop automated tools and dedicate teams to monitoring endpoints. Recent CMS commitments to "building a dynamic, interoperable national provider directory" may help address these barriers [34].

Payer member portals have their own unique challenge, such as a lack of compliance with FHIR standards, missing data fields, delays in making data available, and technical issues resulting in downtime. SMEs noted that only CMS-regulated plans are required to support individual access to payer data, and only a minority of commercial plans have opted to offer this service voluntarily, though several states have begun efforts to extend this requirement to all payers. Some SMEs suggested that, for payers operating both commercial and CMS-regulated plans, it would be technically easier to offer all their members access to their data, but some are investing substantial resources to make this service available only to members with plans that require it. Even for those that require it, SMEs suggested that many payers were not investing sufficiently, resulting in poor service: "They were more concerned about compliance and getting penalized. And now that they have done a shoddy job for 2 years and nothing has happened, they have developed comfort to say 'whatever.'" SMEs hoped that when payers are required to share their data with other payers, they will also make that data available to individuals because the same FHIR endpoints would be used: "It is one thing when 5 or 6 app makers call you up and complain that your API is non-conformant but when 300 payers call you and complain I think it will probably be a different story."

General-Purpose Smartphone App

Description

Study participants can use general-purpose smartphone apps to pull their data from patient or member portals, store that data in the apps on their smartphones, and share the data with research teams. Both major smartphone operating systems allow for this functionality. On iOS, the Apple Health app enables users to retrieve their health records and store them on the device in Apple's HealthKit [35]. On Android, a third-party app called CommonHealth enables users to retrieve their health records and store them in the app or in the device's HealthConnect repository [36]. In both cases, users retrieve their data by logging into patient and member portals using SMART on FHIR protocols. However, the FHIR endpoint directory functionality, which allows users to search for their providers and payers, is different: in iOS/Apple Health, providers must register with Apple and provide metadata, including a logo, and payers are not supported; in Android/CommonHealth, a larger directory of providers and payers is supported, and sites do not need to register. To share data with research studies, both iOS/Apple Health and Android/CommonHealth offer APIs that allow purpose-built apps to integrate with these tools, and there are some open-source examples available [37,38]. Absent a purpose-built app, users could manually share an export of their data for iOS/Apple Health by uploading it to a study server, or create a clinical summary using Android/CommonHealth and allow the research team to access it with a 6-digit passcode [39], but these features were not designed for the research use case and would be more burdensome for participants compared with a purpose-built app.

Assessment

Refer to the "Approaches Using Patient and Member Portals" subheading for the benefits and limitations of that component of this approach. We found benefits and limitations in the way Apple Health and CommonHealth managed FHIR endpoint directories: Apple Health's directory excludes all payers and many providers but may have more accurate and usable metadata to allow for easier search, whereas CommonHealth's includes payers and a larger number of providers, but the metadata is not curated. There are also differences in available data—for example, only CommonHealth includes the FHIR Patient or Encounter resources. Both projects' prior conformance testing on the FHIR data helped identify issues that the site could address to make the data more usable for research and other purposes. One benefit of this approach is that these data, once collected, have the potential to also be used by participants for self-care health apps [40].

Limitations of this approach identified by SMEs include the burden on the research team to establish a mechanism for participants to share their data with the research team—either a burdensome manual process or the development of a custom app; the perception of privacy concerns among participants regarding having their data accessible to technology companies without any affiliation with their health care; dependency on the company for functionality and customizations (eg, inclusion of payer data); and possible sampling bias due to the requirement that participants own and use a smartphone.

Commercial App

Description

SMEs identified several companies that offer products designed to help facilitate research participants sharing their electronic health data with research teams [41]. These products allow users to retrieve their data by logging into their patient and member portals using SMART on FHIR protocols. Some use other, less standardized methods for retrieving the data. The companies that develop these products curate site metadata for their FHIR endpoint directories. Some products are browser-based and therefore not tied to smartphones. We found some products that targeted specific niches within this area, such as focusing exclusively on payer data or supporting clinical trials sponsored by pharmaceutical companies.

Assessment

Refer to the "Approaches Using Patient and Member Portals" subheading for the benefits and limitations of that component of this approach. Unlike approach 1, these products are not limited to smartphones and therefore have a broader reach. Because these products are designed specifically for collecting data for research, they minimize burden on the study team and provide technical support, but they come with a financial cost. Some products provide features that may be helpful depending on the study's needs, such as converting a FHIR payload to the Observational Medical Outcomes Partnership Common Data Model [42] or identifying gaps in data (eg, data elements that

XSL•FO
RenderX

would be expected for a study of early breast cancer) to inform the need for further data collection. Participants who use these apps may benefit from other features offered by them that leverage their aggregated health data [40].

Limitations include study participants' potential privacy concerns with sharing data with a commercial company with which they have no direct relationship. However, the extent of this concern is unknown. An SME from one of these companies said, "We wouldn't have a business if we weren't able to build and maintain that trust with patients." Commercial apps also may have limitations in their ability to be customized for specific study needs.

Research Community App

Description

We identified two noncommercial software tools developed to facilitate data collection for research studies: Sync-for-Science Procure [43] (developed by co-author DG and available as open-source) and Wellmine [44] (developed by RTI International and available for studies conducted by that organization). This approach allows users to retrieve their data by logging into their patient and member portals using SMART on FHIR protocols. Research studies, or their institutions, using these products would need to deploy an instance of the product and create and manage their own FHIR endpoint directories, registering the instance with each endpoint. As with several of the commercial apps (approach 2), both of these tools are browser-based.

Assessment

Refer to the "Approaches Using Patient and Member Portals" subheading for the benefits and limitations of that component of this approach. Benefits of this approach compared with approach 2 (commercial apps) include potentially reducing dependency on an external company, the potential for reduced privacy concerns if the product is viewed by participants as affiliated with a trusted organization (eg, their health care provider), the potential to more easily add customizations for study-specific needs, and the potential to foster a community to innovate in this area, which may involve developing features that provide value to patients from their aggregated health data. Limitations include the lack of polish that's often present in open-source or homegrown software products, which may necessitate additional development, and the burden on studies or their institutions to establish and maintain an instance and FHIR endpoint directory; however, once established, a single instance of the app may be used across many studies.

Structured Data Export

Description

In this approach, the study participant logs into each of their patient and member portals, downloads their health data onto their local device, and then shares it using a mechanism set up by the research team (eg, uploading the data to a study server). EHR vendors are required by the 21st Century Cures Act Final Rule to support electronic health information (EHI) export functionality for patient data [45]. The participant would need to request these data and then share them through a method established by the research team. EHI export may also be requested separately from a patient portal, such as by submitting a request form or email; such manual approaches are out of scope for this landscape analysis.

Assessment

As with other approaches that rely on the use of patient and member portals, there is a burden on study participants to remember their sites of care and health plans, establish portals, and log into each one, which may result in incomplete data. There may also be data completeness issues due to relying on the participants for the timing of the data sharing.

EHI export has generally not been made directly available to patients through portals, but SMEs suggested that it would become more available if requested by patients, as providing this capability is required under existing regulations to be supported by EHRs. Although the data may be more comprehensive compared with data available through FHIR-based approaches, it may be challenging to work with [40] because of nonstandardized data formats that vary by EHR vendor and site of care: "We left it up to developers to determine for themselves how to export information. We basically said the EHI export has to come with a data dictionary so people can figure out how to kind of piece things back together. But I have no doubt that there are some developers who do a fairly nice job with it and then other developers who do not." Similar to EHI but less standardized, many portals currently allow patients to download some or all of their data in PDF format, which may be increasingly usable for research because they can be processed into structured data by large language models.

This approach does not involve SMART on FHIR and therefore no FHIR endpoint directory is required, reducing related study burden and/or costs. However, the study would still need to establish a method to allow for secure sharing of the data, and the manual effort required on the part of study participants may be challenging for many participants.

TEFCA Individual Access Service

Description

In this approach, participants would use the Individual Access Service (IAS) required by the TEFCA [7]. TEFCA is designed to create a "network of networks" to facilitate the exchange of health information [7]. The IAS allows apps to register with participating QHINs (a designated network under TEFCA) and query for and aggregate all of an individual's health data across QHINs. The data would then be shared with research studies either using functionality offered by the app or by asking the participant to export the data and share it using a method established by the research team (eg, uploading the data to a study server). This approach would require an app and, therefore, would need to be bundled with approach 1, 2, or 3.

Assessment

Benefits of this approach are that it would leverage TEFCA's mechanism for searching for and retrieving individuals' health records, obviating the need for research teams and app companies to establish FHIR endpoint directories; facilitate identification of data sources by using TEFCA's record location

XSL•FO
RenderX

and identity matching methods so that research participants would not have to remember their sites of care; potentially offer participants additional functionality based on the aggregated data; provide privacy controls if offered by the app; and align with existing policy supporting individual access to health data. Currently, some QHINs require users to authenticate with and approve an access request within the patient portal of each potential site of care to ensure correct patient identity matching and consent when using IAS; while the QHIN can still serve as a record locator, this greatly limits the benefits of this approach by adding burden to the patient. At other QHINs, patients may only be matched to sites of care that have a very high level of confidence in the accuracy of the match, potentially excluding sites of care with patient data. Active efforts are trying to increase support for shared login credentials across sites [31]. Apps that use the IAS must meet HIPAA privacy and security requirements through their signed participation agreements with QHINs (even though they are not obligated under HIPAA as covered entities) [7]. Additionally, apps that adhere to the CARIN code of conduct may increase the security safeguards even further [46].

The major limitation of this approach is that although it is technically available and required as part of TEFCA today, sites that receive IAS requests often refuse to respond to them. SMEs believed this was because sites do not believe that the data would always be released to the correct patient and that they would be blamed for HIPAA violations if the query resulted in an improper disclosure. One SME said that providers would not share records "until they know darn good and well that they're not going to get sued for giving the wrong person the wrong information." SMEs suggested that the adoption of FHIR-based queries would allow for more individual identity data to be used for record matching, thereby reducing the odds of improper disclosures. However, some SMEs believed competitive concerns were also a factor—sites did not want to share data with entities that did not also provide data back into the network. SMEs suggested that some method of enforcement would be needed. Use of SSO across patient and member portals may help with matching concerns by standardizing and improving the quality of identity information [29] as well as automating portal logins [47]. Other limitations include the lack of some data sources available on QHINs (eg, payers and many small EHRs do not participate), performance issues when trying to query large geographic areas, imperfect record matching resulting in missed data, lack of analytic dataset preparation (unless supported by the app), and lack of support for querying directly from an app on an individual's device, which would be more privacy preserving.

Regional Study Query Through HIE

Description

HIEs exchange electronic health data for multiple purposes within a specific geographic region or at a statewide level [48,49]. There is a high level of interest among HIEs in supporting research, but only some currently do so [50,51]. One barrier is that some HIEs that did not initially include research use cases are required to update their member participation agreements, which can be a cumbersome process. In this approach, the research team works with the HIE staff to specify their data needs for consented patients under an agreed-upon scope of work, which may include preparing the data in a specific format, for a fee. HIEs' governance processes address state and federal legal concerns (eg, the sharing of sensitive health data) as well as issues raised by HIE members. The processes differ across HIEs. Some require IRB approval from several member organizations; others use a single IRB. Some allow access to claims data from multiple payers; others lack claims data. Some share clinical notes for research; others do not. Members may have the option to opt out of sharing their data for specific requests. There is an interest among at least some HIEs in providing patients with access to a portal for managing their consents for research projects. The HIE staff create the datasets and work with the research team to resolve issues, such as fixing demographic information if no record matches are found for specific patients.

Assessment

This approach would be widely accessible to all participants regardless of their ability to use technology, involve minimal burden on the participant, allow research teams to control the timing of the queries to reduce variability in retrieved data, broadly align with policy goals supporting data exchange, and may foster innovation in how HIEs can better support research needs. It leverages years of investment in infrastructure and governance processes in many communities. Some HIEs also include social determinants of health data from community-based organizations, which would be valuable for many research studies. Another benefit that SMEs pointed to is that HIEs can ensure compliance with state laws for data sharing centrally, which may be much more challenging for national-level queries. The major limitation of this approach is that it is currently available only in a limited number of geographic areas, though rigorous data are not available. Also, not all HIEs include claims data, and very few have access to Medicare or Medicaid data; their coverage of providers varies, and some states require patient opt-in, resulting in less available data.

National Study Query

Description

In this approach, the participant consents to allowing the research team to query a national network of health care providers, such as CareQuality, Common Well Health Alliance, or other QHINs participating in TEFCA. The research team then performs the query and assembles the data. For these networks, queries are required to specify the "purpose of use" of the data [7]. Most queries today specify "treatment," which requires that the querier be a provider of care and offer data back to the network for others to query (ie, reciprocity). Currently, participants are not required to accept requests that specify "research" as a purpose of use under TEFCA or any national network. SMEs described preliminary conversations among policymakers about making research a purpose of use in TEFCA, including plans to conduct a review of the landscape of laws across states and to gather other information from stakeholders.

XSL•FO
RenderX

Assessment

This approach would be widely accessible to all participants regardless of their ability to use technology, involve minimal burden on the participant or the research team, allow research teams to control the timing of the queries to reduce variability in retrieved data, and broadly aligns with policy goals. FHIR resources subscription standards under development could allow research teams to receive automatic updates for new data, which may be useful to some studies [52].

Many limitations are the same as those of approach 5 (IAS) and include the lack of some data sources available on QHINs (eg, payers and many small EHRs are not participating), performance issues when trying to query large geographic areas, and imperfect record matching resulting in missed data.

However, the major limitation is that, unlike the queries in approach 5, support for research queries is not currently a requirement for TEFCA participants, and the effort required to make research a purpose of use that is answered reliably by sites is substantial. SMEs described how participating providers and EHR vendors would require processes to be in place to ensure that a query is legitimate (eg, participants are consented, the study is IRB approved), and an enforcement mechanism to ensure that the data are used only for the designated purpose and in compliance with HIPAA regulations. This will be challenging because rules governing consent for such queries vary across states. Some have proposed that QHINs offer digital identity services to patients and allow them to control and view access to their data—by putting the patient in charge, this would obviate trust issues among network participants [53]. Yet, one SME suggested that research queries “may never happen at scale.”

SMEs described how national queries were being used for research today. Some existing companies query these networks for treatment purposes and then use the data for research as a secondary use. For example, one company justified its use of the treatment purpose of use by establishing itself as a telemedicine provider that helps patients navigate the health system, and then it uses those data for research with participant consent. In other cases, some forms of clinical trials might be considered both treatment and research. SMEs hoped that a future legal pathway for research queries would clarify this situation.

Aggregate Data Source

Description

In this approach, the participant consents to allowing the research team to collect data from a company that has aggregated datasets from multiple providers (directly or through EHR vendors or national laboratories) and/or payers (directly or through claims clearinghouses or pharmacy aggregators). The company serves as a marketplace connecting research teams with the data sources. The research team works with the organization to specify the data they need and pays a fee. The organization handles all the governance, business, and operational issues, including securing approval from the data sources, identifying the records associated with the consented patients, and preparing analytic datasets.

Assessment

This approach is available today, widely accessible to all participants regardless of their ability to use technology, involves minimal burden on the participant, and allows research teams to control the timing of the queries to reduce variability in retrieved data. Limitations include cost to the research team, limits in the extent of available data sources, and lack of patient control over or awareness of the data available. Many data sources (eg, providers and payers) will only agree to release data anonymously. Research teams conducting studies with consented patients could meet this requirement by bifurcating their team into those with access to limited identifiable data (eg, for recruiting) and those with access to deidentified data for analyses. This may add burden to the study team and may not be feasible for some research teams.

Discussion

Principal Findings

We identified 8 approaches that would facilitate research team access to data from multiple providers and payers for participants in US studies, and 12 evaluation criteria. There were clear benefits and limitations for each approach. Participant-mediated approaches have the major advantage of bypassing the need for complex governance and privacy management processes, but they burden the participant, require effort or costs for the study team, and have incomplete connectivity to data sources, especially payers. Approaches involving queries of data exchange infrastructure are robust in some regions but vary substantially in terms of the data available and their ability to support research, and they may be challenging to implement at a national scale. Aggregate data sources allow for queries of national-level data in a way that addresses privacy management and governance issues but have limits in the scope of data sources. The benefits and limitations of each approach will likely change as the landscape evolves, driven largely by public policy.

In addition to the three steps of our focus (identifying the location of the data, retrieving the data, and sharing the data with research studies), SMEs emphasized the value of a fourth step: preparing the data for research. Many SMEs emphasized the value of these services, including normalization, conformance testing, and quality checks. As FHIR is a relatively new format for health data, most researchers are not accustomed to analyzing it. Many researchers may benefit from working with organizations familiar with the data’s idiosyncrasies, and that can produce an analytic dataset tailored to the needs of the research project. There are active efforts to make FHIR data easier to validate [54] and analyze [55], which may mitigate this concern.

The value of synergies between payer and provider data (eg, leveraging payer data to identify and retrieve data from providers that a participant has visited previously) is only beginning to be explored and has the potential to provide a feedback loop in this process. For example, payer data can identify gaps in provider data and locations to target for data collection.

XSL•FO
RenderX

This complex landscape is the result of the fragmented structure of the US health care system, which produces the data used in research, wherein patients receive care from many different providers who use different EHR systems and are covered by different payers. Within this landscape, each research study will need to consider the best option to address its data needs within study budgets. Studies may consider multiple approaches to obtaining as complete a dataset as possible. For example, a study might use a participant-mediated approach for Medicare claims data in combination with working with a local HIE or aggregate data source for EHR data.

Contribution

To our knowledge, this is the first study attempting to describe the US landscape of approaches to collect data from multiple providers and payers for research. Our results provide a comprehensive comparison of the benefits and limitations of each approach, allowing research teams to understand existing and anticipated future options and policymakers to target areas for improvement.

We found only 1 publication that discusses how a participant-mediated approach was used to collect data for a study [41]. Other studies may have used this approach but have not described or assessed it. Sayeed et al [37] assessed the potential for using FHIR standards for multiple research tasks, including collecting EHR data, as part of the People Heart Study. We did not find any assessments of the use of TEFCA IAS. HIEs and aggregated data sources have been used for studies involving consented patients, but we did not find a systematic examination of these approaches.

Limitations

This study was limited to a US focus; the landscape in other countries regarding the collection of comprehensive data for research will likely differ and would be worth investigating in future work. Our data relied on SMEs and may not have included all relevant perspectives. The landscape of approaches is rapidly evolving, and some of our findings may become outdated, especially as new policies are implemented. Additionally, we did not conduct a systematic literature review, as our primary objective was to gather expert views to inform our analysis. However, we did review relevant literature identified by the research team and SMEs, who likely were aware of influential publications. Our scope was limited to one use case: the collection of provider and payer data for consented participants. We excluded other study workflows (eg, recruitment) and other types of studies (eg, population-level analyses of deidentified data).

Conclusions

Complex and evolving trade-offs exist among the approaches we identified for collecting comprehensive electronic health data for research in the United States. We did not find a clear "winner"—the optimal approaches will likely vary considerably based on factors specific to a study. Continued development, exploration, and evaluation of all approaches are warranted.

Policies supporting participant-mediated exchange could greatly improve those approaches by making FHIR endpoints more findable (eg, through a national public directory), making it easier for researchers to broadly register an app (eg, requiring a central location for all of an EHR vendor's customers), and ensuring that FHIR APIs are conformant with standards (eg, automated testing). Conformance requirements for FHIR APIs and data standards are particularly needed for private payer endpoints and are more efficient than expecting each research team to deal with the data issues individually. Advances in digital identity and the use of SSO across providers and patients may facilitate portal signup and the ability for study participants to sign in and pull records. Ongoing conformance evaluation of FHIR endpoints and assessments of participant ability to find and share data from prior sites of care would help inform improvement efforts.

Policies may also support queries of the data exchange infrastructure. Some regional HIEs have demonstrated the feasibility and value of using their existing infrastructure to pull data for research [50,56-58]. These HIEs have distinct advantages, such as coverage of most relevant local data sources, familiarity with state and local privacy laws, established governance processes, and knowledge of their members' data quality issues. Some are more advanced in supporting research than others, and there is room for additional innovation. Policies could support establishing HIEs in regions that lack them, developing more capabilities to support research, and facilitating the sharing of best practices and evaluation. National efforts to make data available for research are still in their infancy, and it is not clear if or how the trust issues among data holders can be resolved, especially considering recent litigation [59,60]. The IAS may be the most promising approach that uses the national networks, but it will require clarifying the liability rules for data holders when they respond to requests. Evaluations should assess retrieved data in comparison with data obtained using a treatment purpose of use.

As these approaches evolve, ongoing rigorous evaluation and benchmarking of the benefits and limitations of each approach would empower research teams to make better decisions for study data collection.

Funding

Research reported in this publication was supported by the National Institute on Aging of the National Institutes of Health (NIH) under award R01AG083035. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This paper is subject to the NIH Public Access Policy. Through acceptance of this federal funding, the NIH has been granted the right to make this paper publicly available in PubMed Central upon the official date of publication, as defined by NIH. No generative artificial intelligence tools were used in the preparation of this paper.

XSL•FO
RenderX

Data Availability

The datasets generated and analyzed during this study are not publicly available because of participant confidentiality concerns but are available from the corresponding author on reasonable request.

Authors' Contributions

RSR led the conceptualization, data collection, formal analysis, methodology, supervision, funding acquisition, and writing of the original draft. DG, NC, JCM, WJG, and DM contributed to conceptualization. DG and NC contributed to the formal analysis, and NC also contributed to data collection and methodology.

Conflicts of Interest

RSR reports a small revenue share with GreenCape Health, a small digital health company, unrelated to this work. William J Gordon is an employee at Manatt Health. Joshua C Mendel is an employee at Microsoft Corporation.

Multimedia Appendix 1
Evaluation of approaches to collecting comprehensive provider and payer data for research.
[XLSX File, 15 KB - jmir_v28i1e86330_app1.xlsx]

Checklist 1
SRQR checklist.
[PDF File, 69 KB - jmir_v28i1e86330_app2.pdf]

References

1. Adoption of electronic health records by hospital service type 2019-2021. Office of the National Coordinator for Health Information Technology. 2022. URL: https://healthit.gov/data/quickstats/adoption-electronic-health-records-hospital-service-type-2019-2021/#overview [accessed 2026-06-08]
2. Riva A, Mandl KD, Oh DH, et al. The personal internetworked notary and guardian. Int J Med Inform 2001 Jun;62(1):27-40. [doi: 10.1016/s1386-5056(00)00136-2] [Medline: 11340004]
3. Al-Sahab B, Leviton A, Loddenkemper T, Paneth N, Zhang B. Biases in electronic health records data for generating real-world evidence: an overview. J Healthc Inform Res 2024 Mar;8(1):121-139. [doi: 10.1007/s41666-023-00153-2] [Medline: 38273982]
4. Gurupur V, Hooshmand S, Prabhu DF, Trader E, Salvi S. Incompleteness of electronic health records: an impending process problem within healthcare. Healthcare (Basel) 2025 Nov 13;13(22):2900. [doi: 10.3390/healthcare13222900] [Medline: 41302288]
5. Kar S, Bessette LG, Wyss R, Kesselheim AS, Lin KJ. The impact of electronic health record discontinuity on prediction modeling. PLoS One 2023;18(7):e0287985. [doi: 10.1371/journal.pone.0287985] [Medline: 37410777]
6. Martin S, Wagner J, Lupulescu-Mann N, et al. Comparison of EHR-based diagnosis documentation locations to a gold standard for risk stratification in patients with multiple chronic conditions. Appl Clin Inform 2017 Aug 2;8(3):794-809. [doi: 10.4338/ACI-2016-12-RA-0210] [Medline: 28765864]
7. Advancing nationwide interoperability with TEFCA®. Office of the National Coordinator for Health Information Technology. URL: https://healthit.gov/policy/tefca/ [accessed 2026-06-08]
8. Black JR, Hulkower RL, Ramanathan T. Health information blocking: responses under the 21st Century Cures Act. Public Health Rep 2018;133(5):610-613. [doi: 10.1177/0033354918791544] [Medline: 30134128]
9. Lye CT, Forman HP, Daniel JG, Krumholz HM. The 21st Century Cures Act and electronic health records one year later: will patients see the benefits? J Am Med Inform Assoc 2018 Sep 1;25(9):1218-1220. [doi: 10.1093/jamia/ocy065] [Medline: 30184156]
10. Majumder MA, Guerrini CJ, Bollinger JM, Cook-Deegan R, McGuire AL. Sharing data under the 21st Century Cures Act. Genet Med 2017 Dec;19(12):1289-1294. [doi: 10.1038/gim.2017.59] [Medline: 28541278]
11. Gordon WJ, Rudin RS. Why APIs? Anticipated value, barriers, and opportunities for standards-based application programming interfaces in healthcare: perspectives of US thought leaders. JAMIA Open 2022 Jul;5(2):ooac023. [doi: 10.1093/jamiaopen/ooac023] [Medline: 35474716]
12. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care 2012 Jul;50 Suppl:S21-S29. [doi: 10.1097/MLR.0b013e318257dd67] [Medline: 22692254]
13. Pham HH, Schrag D, O'Malley AS, Wu B, Bach PB. Care patterns in Medicare and their implications for pay for performance. N Engl J Med 2007 Mar 15;356(11):1130-1139. [doi: 10.1056/NEJMsa063979] [Medline: 17360991]

XSL•FO
RenderX

14. Pham HH, O'Malley AS, Bach PB, Saiontz-Martinez C, Schrag D. Primary care physicians' links to other physicians through Medicare patients: the scope of care coordination. Ann Intern Med 2009 Feb 17;150(4):236-242. [doi: 10.7326/0003-4819-150-4-200902170-00004] [Medline: 19221375]
15. Fang H, Frean M, Sylwestrzak G, Ukert B. Trends in disenrollment and reenrollment within US commercial health insurance plans, 2006-2018. JAMA Netw Open 2022 Feb 1;5(2):e220320. [doi: 10.1001/jamanetworkopen.2022.0320] [Medline: 35201308]
16. Szolovits P, Long WJ, Kohane I, Pauker SG. Guardian angel: patient-centered health information systems. : Massachusetts Institute of Technology, Laboratory for Computer Science; 1994 URL: https://dspace.mit.edu/entities/publication/f1e3abf1-73d7-4c08-8851-b23bf5e0bcba [accessed 2026-06-08]
17. Yasnoff WA, Shortliffe EH. Lessons learned from a health record bank start-up. Methods Inf Med 2014;53(2):66-72. [doi: 10.3414/ME13-02-0030] [Medline: 24477917]
18. Everson J, Chang W, Patel V, Adler-Milstein J. The state of health information organizations and plans to participate in the federal exchange framework. Health Aff Sch 2024 Aug;2(8):qxae098. [doi: 10.1093/haschl/qxae098] [Medline: 39188926]
19. Chimanuka Murhima'alika C, Mary M, Chiribagula Zalinga C, et al. Exploring opportunities to strengthen maternal and perinatal death surveillance and response: a landscape analysis of surveillance and health information systems in the Eastern Democratic Republic of Congo. Confl Health 2025 Nov 5;19(1):80. [doi: 10.1186/s13031-025-00720-x] [Medline: 41194246]
20. Meek JY, Nelson JM, Hanley LE, Onyema-Melton N, Wood JK. Landscape analysis of breastfeeding-related physician education in the United States. Breastfeed Med 2020 Jun;15(6):401-411. [doi: 10.1089/bfm.2019.0263] [Medline: 32320260]
21. Shalash A, Nemer M, Abu-Rmeileh N, Kittaneh M, Kelly D, Elmusharaf K. Mapping stakeholders, services, data, and the information system for adolescent health in the West Bank. Reprod Health 2025 May 31;22(Suppl 1):81. [doi: 10.1186/s12978-025-01991-6] [Medline: 40448084]
22. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
23. Berg BL. Qualitative Research Methods for the Social Sciences, 6th edition: Pearson; 2007. URL: https://openlibrary.org/books/OL21527186M/Qualitative_research_methods_for_the_social_sciences [accessed 2026-06-08]
24. Saunders B, Kitzinger J, Kitzinger C. Anonymising interview data: challenges and compromise in practice. Qual Res 2015 Oct;15(5):616-632. [doi: 10.1177/1468794114550439] [Medline: 26457066]
25. Medicare and Medicaid programs; Electronic Health Record Incentive Program-Stage 3 and modifications to meaningful use in 2015 through 2017. Federal Register. 2015. URL: https://www.federalregister.gov/documents/2015/10/16/2015-25595/medicare-and-medicaid-programs-electronic-health-record-incentive-program-stage-3-and-modifications [accessed 2026-06-03]
26. CMS seeks public input on improving technology to empower medicare beneficiaries. CMS. 2025. URL: https://www.cms.gov/newsroom/press-releases/cms-seeks-public-input-improving-technology-empower-medicare-beneficiaries [accessed 2026-06-03]
27. Health tech ecosystem categories. CMS. 2025. URL: https://www.cms.gov/health-technology-ecosystem/categories [accessed 2026-06-03]
28. Rudin RS, Hillestad R, Ridgely MS, Qureshi NS, Davis JS 2nd, Fischer SH. Defining and evaluating patient-empowered approaches to improving record matching. Rand Health Q 2019 May;8(3):3. [Medline: 31205803]
29. Tiered OAuth for user authentication. HL7. 2025. URL: https://hl7.org/fhir/us/udap-security/user.html [accessed 2026-06-03]
30. CARIN Alliance and HHS release digital identity federation report about their test proof of concept. Leavitt Partners. 2023. URL: https://leavittpartners.com/carin-alliance-and-hhs-release-digital-identity-federation-report-about-their-test-proof-of-concept [accessed 2026-06-08]
31. Interoperability framework. CMS. URL: https://www.cms.gov/health-technology-ecosystem/interoperability-framework [accessed 2026-06-08]
32. User-access brands and endpoints. HL7. 2024. URL: https://www.hl7.org/fhir/smart-app-launch/brands.html [accessed 2026-06-08]
33. Health data, technology, and interoperability: certification program updates, algorithm transparency, and information sharing. Federal Register. 2024. URL: https://www.federalregister.gov/documents/2024/01/09/2023-28857/health-data-technology-and-interoperability-certification-program-updates-algorithm-transparency-and [accessed 2026-06-08]
34. CMS building foundational infrastructure for digital healthcare ecosystem. Centers for Medicare & Medicaid Services. 2025. URL: https://www.cms.gov/newsroom/press-releases/cms-building-foundational-infrastructure-digital-healthcare-ecosystem [accessed 2026-06-08]
35. HealthKit. Apple Developer. 2026. URL: https://developer.apple.com/documentation/healthkit [accessed 2026-06-08]
36. Medical records. Android Developers. 2026. URL: https://developer.android.com/health-and-fitness/guides/medical-records [accessed 2026-06-08]
37. Sayeed R, Kreda D, Mandel JC, et al. A standards-based approach to digital health research: implementing the people heart study. J Am Med Inform Assoc 2025 Dec 1;32(12):1811-1821. [doi: 10.1093/jamia/ocaf163] [Medline: 41032388]

XSL•FO
RenderX

38. Truslow J, Spillane A, Lin H, et al. Understanding activity and physiology at scale: the Apple Heart & Movement Study. NPJ Digit Med 2024 Sep 10;7(1):242. [doi: 10.1038/s41746-024-01187-5] [Medline: 39256546]
39. SHL protocol specification. SMART Health Links. URL: https://docs.smarthealthit.org/smart-health-links/spec/ [accessed 2026-06-08]
40. Keeler B. Indiana Jones and the personal health record. Health API Guy. 2021. URL: https://healthapiguy.substack.com/p/indiana-jones-and-the-personal-health [accessed 2026-06-08]
41. Krumholz HM, Sawano M, Bhattacharjee B, et al. The PAX LC trial: A decentralized, phase 2, randomized, double-blind study of nirmatrelvir/ritonavir compared with placebo/ritonavir for long COVID. Am J Med 2025 May;138(5):884-892. [doi: 10.1016/j.amjmed.2024.04.030] [Medline: 38735354]
42. Standardized data: the OMOP common data model. Observational Health Data Sciences and Informatics. URL: https://www.ohdsi.org/data-standardization [accessed 2026-06-08]
43. Procure-wip. GitHub. URL: https://github.com/sync-for-science/procure-wip [accessed 2026-06-08]
44. Homepage. RTI Wellmine. 2025. URL: https://wellmine.rti.org [accessed 2026-06-08]
45. 21st Century Cures Act: interoperability, information blocking, and the ONC Health IT Certification Program. Federal Register. 2020. URL: https://www.federalregister.gov/documents/2020/05/01/2020-07419/21st-century-cures-act-interoperability-information-blocking-and-the-onc-health-it-certification [accessed 2026-06-08]
46. Code of conduct. CARIN Alliance. URL: https://www.carinalliance.com/code-of-conduct [accessed 2026-06-03]
47. Kulatunga J. TEFCA IAS - deep dive into patient matching & RLS responses. Fasten Health Blog. 2025. URL: https://blog.fastenhealth.com/tefca-patient-matching-rls [accessed 2026-06-08]
48. Health data utilities. CIVITAS: Networks for Health. URL: https://www.civitasforhealth.org/health-data-utilities [accessed 2026-06-08]
49. Gryczynski J, Nordeck CD, Welsh C, Mitchell SG, O'Grady KE, Schwartz RP. Preventing hospital readmission for patients with comorbid substance use disorder: a randomized trial. Ann Intern Med 2021 Jul;174(7):899-909. [doi: 10.7326/M20-5475] [Medline: 33819055]
50. Clinical data drives insight and innovation to transform our healthcare system. CRISP Health. URL: https://www.crisphealth.org/research/ [accessed 2026-06-08]
51. EHR Good Neighbor. Scaling research data access statewide with bulk FHIR. URL: https://good-neighbor.smarthealthit.org/case-studies/scaling-research-data-access-statewide-with-bulk-fhir [accessed 2026-06-08]
52. Resource subscription. HL7. URL: https://www.hl7.org/fhir/subscription.html [accessed 2026-06-08]
53. Mandel JC, Pollak JP, Mandl KD. The patient role in a federal national-scale health information exchange. J Med Internet Res 2022 Nov 4;24(11):e41750. [doi: 10.2196/41750] [Medline: 36331535]
54. Qualifier. GitHub. URL: https://github.com/sync-for-science/qualifier/tree/master [accessed 2026-06-08]
55. SQL on FHIR. HL7 FHIR Implementation Guides. URL: https://build.fhir.org/ig/HL7/sql-on-fhir/ [accessed 2026-06-08]
56. Homepage. HealthShare Exchange (HSX). URL: https://healthshareexchange.org/homepage [accessed 2026-06-08]
57. How we bring the data to you. Regenstrief Institute. URL: https://www.regenstrief.org/rds/data [accessed 2026-06-03]
58. Homepage. MyHealth Access Network. URL: https://myhealthaccess.net [accessed 2026-06-08]
59. Adler S. Epic sues health information exchange network alleging improper record access. The HIPAA Journal. 2026. URL: https://www.hipaajournal.com/epic-sues-health-information-exchange-network-improper-record-access/ [accessed 2026-06-03]
60. Keeler B. Epic v. particle 2: the problem of secondary use. Health API Guy. 2024. URL: https://healthapiguy.substack.com/p/epic-v-particle-2-the-problem-of [accessed 2026-06-08]

Abbreviations

API: application programming interface
CMS: Centers for Medicare & Medicaid Services
EHI: electronic health information
EHR: electronic health record
FHIR: Fast Health Interoperability Resources
HIE: health information exchange
HIPAA: Health Insurance Portability and Accountability Act
IAS: Individual Access Service
IRB: Institutional review board
QHIN: qualified health information network
SMART: Substitutable Medical Applications and Reusable Technologies
SME: subject matter expert
SSO: single sign-on
TEFCA: Trust Exchange Framework and Common Agreement

XSL•FO
RenderX

Edited by A Stone; submitted 27.Oct.2025; peer-reviewed by B Ward, O Steichen; revised version received 02.Apr.2026; accepted 02.Apr.2026; published 15.Jun.2026.

Please cite as:
Rudin RS, Chhatiawala N, Mandel JC, Gordon WJ, McCallie D, Gottlieb D
Approaches to Collect Comprehensive Electronic Patient Data Across Multiple Providers and Payers for Research: Landscape Analysis
J Med Internet Res 2026;28:e86330
URL: https://www.jmir.org/2026/1/e86330
doi:10.2196/86330

© Robert S Rudin, Nihar Chhatiawala, Joshua C Mandel, William J Gordon, David McCallie, Daniel Gottlieb. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Association Between Telemedicine Adoption and Physician Job Satisfaction: Cross-Sectional Study

Xiaohui Zhai[1], PhD; Zhongliang Zhou[2], PhD; Zhichao Wang[2], PhD; Guanping Liu[2], PhD; Peter C Coyte[3], PhD

[1]School of Public Health, Health Science Center, Xi'an Jiaotong University, Xi'an, Shaanxi, China

[2]School of Public Policy and Administration, Xi'an Jiaotong University, No. 28 Xianning West Road, Xi'an, Shaanxi, China

[3]Institute of Health Policy, Management & Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Zhongliang Zhou, PhD
School of Public Policy and Administration, Xi'an Jiaotong University, No. 28 Xianning West Road, Xi'an, Shaanxi, China

Abstract

Background: Telemedicine has expanded rapidly in recent years, with particularly pronounced growth following the COVID-19 pandemic. By improving access to care and offering greater flexibility in service delivery, it has become an important component of health care. Although the benefits of telemedicine for patients are well documented, its effects on physician job satisfaction remain insufficiently understood. Given the importance of job satisfaction for workforce stability, physician well-being, and quality of care, further examination of how telemedicine affects physician job satisfaction is warranted.

Objective: This study aims to examine the association between telemedicine adoption and physician job satisfaction and to assess whether the physician-patient relationship mediates this association.

Methods: A cross-sectional survey was conducted among health care professionals in Xi'an, China. Data were collected between November 7 and December 8, 2023, via an online questionnaire administered using the REDCap (Research Electronic Data Capture; Vanderbilt University) platform. A total of 12,052 physicians were included in the analysis. Physician job satisfaction was measured using a validated 6-point Likert scale. Telemedicine adoption was assessed through self-report. A partial proportional odds model was used to examine the association between telemedicine adoption and job satisfaction, adjusting for a comprehensive set of potential confounders. Additionally, the Karlson-Holm-Breen (KHB) decomposition method was used to explore the mediating role of physician-patient relationship quality in this association.

Results: Among 12,052 surveyed physicians, 1642 (13.62%) reported adopting telemedicine, whereas 10,410 (86.38%) did not. After adjusting for demographic characteristics, work-related factors, psychological factors, and physician-patient relationship, telemedicine adoption was significantly associated with higher job satisfaction (odds ratio [OR] 1.17, 95% CI 1.05 - 1.30). Findings were robust across multiple sensitivity analyses. Subgroup analyses indicated that the association did not vary across physician subgroups, and no significant interaction effects were observed. Mediation analysis revealed a total effect of telemedicine on job satisfaction of 0.33 (95% CI 0.17 - 0.50), with an indirect effect of 0.10 (95% CI 0.07 - 0.13) through improved physician-patient relationships, accounting for 30.30% of the total effect.

Conclusions: These findings suggest that telemedicine adoption is positively associated with physician job satisfaction, partially mediated by the physician-patient relationship. Policies should promote telemedicine adoption while prioritizing platform designs that support effective physician-patient interactions to enhance provider well-being and care outcomes.

(J Med Internet Res 2026;28:e82285) doi:10.2196/82285

KEYWORDS

telemedicine; job satisfaction; technology adoption; cross-sectional study; physician patient relationship

Introduction

Telemedicine refers to the provision of health care services over a distance using information and communication technologies, enabling remote diagnosis, consultation, treatment, and patient monitoring without the need for physical co-location [1]. Telemedicine has demonstrated significant potential to address longstanding challenges in health care systems. It effectively mitigates barriers such as limited access to care in rural and underserved populations, inefficiencies caused by overcrowded facilities and prolonged patient wait times, uneven distribution of health care resources, and escalating costs associated with traditional in-person care [2]. Globally, the use of telemedicine has increased significantly, especially following the COVID-19 pandemic [3]. According to a World Health Organization (WHO) survey, by 2022, over 70% of member states in the WHO European Region had implemented some form of telemedicine [4].

XSL•FO
RenderX

Telemedicine has become increasingly integrated into diverse clinical contexts, including primary care [5], chronic disease management [6], mental health services [7,8], and specialist consultations [9]. For physicians, this development has significantly altered traditional modes of clinical work. Remote consultations alter the pace and structure of daily workflows, reduce face-to-face interactions, and often require adaptation to new technologies [10]. Clinical decision-making is increasingly supported by digital tools [11], and communication with both patients and colleagues often takes place via electronic platforms.

These shifts have important implications for physicians' work experience [12], particularly in relation to job satisfaction. Job satisfaction is widely recognized as a critical component of physician well-being and a key determinant of health care system performance. Higher levels of physician satisfaction are consistently associated with improved clinical performance [13,14], lower turnover intentions [15,16], reduced burnout [17], and better patient care [18]. Nevertheless, empirical evidence on the association between telemedicine and physicians' job satisfaction remains mixed. In Australia, a large national survey of over 7000 physicians demonstrated that those using digital health technologies were approximately 14 percentage points more likely to report higher job satisfaction versus nonusers [19]. Similarly, a US study found that the use of specific telemedicine modalities, particularly video consultations and electronic health record (EHR) integrated platforms, was positively associated with higher physician job satisfaction [20]. However, evidence from a large ambulatory cohort indicated potential drawbacks, as higher telemedicine intensity was associated with increased after-hours EHR workload, which may negatively affect job satisfaction and work-life balance [21]. These divergent findings highlight the need for further investigation into the relationship between telemedicine adoption and physicians' overall job satisfaction.

Although prior studies have explored the effects of telemedicine on physician job satisfaction, limited attention has been paid to whether this relationship varies across different physician subgroups. Given the substantial variation in physicians' demographic and professional characteristics, it is plausible that telemedicine may not affect all physicians in the same way [22]. Examining these potential subgroup differences can improve our understanding of how telemedicine influences physician satisfaction and help inform more targeted and effective implementation strategies.

Moreover, few studies have examined the mechanisms underlying the relationship between telemedicine adoption and physician job satisfaction. One potential mediator is the physician-patient relationship. Telemedicine transforms how doctors and patients communicate by enabling clinical interactions to occur on virtual platforms. The adoption of telemedicine can modify key aspects of clinical communication, enabling more flexible information exchange, reduced time burden, and more efficient clarification and follow-up of clinical issues [23]. Such improvements can strengthen the quality of physician-patient interactions and foster a more positive relationship. In turn, this improved relationship is plausibly linked to physicians' job satisfaction, according to Relational Job Design Theory [24], which posits that more frequent, prolonged, and meaningful interactions with beneficiaries strengthen employees' perceptions of their positive impact and affective commitment, thereby increasing their motivation to make a prosocial difference. For physicians, stronger prosocial motivation is likely to enhance their engagement in patient care and reinforce their sense of competence and social worth, both of which are strongly linked to higher job satisfaction [25,26]. Taken together, these insights suggest that the physician-patient relationship may serve as a key pathway linking telemedicine adoption to physicians' job satisfaction.

Given the increasing integration of telemedicine into routine clinical practice, understanding its impact on physician well-being has become increasingly important. This study aims to evaluate the impact of telemedicine adoption on physicians' job satisfaction and to examine whether this association varies across different physician subgroups. Moreover, the study explores the underlying mechanisms of this relationship by assessing the mediating role of the physician-patient relationship. The findings are expected to provide practical guidance for improving telemedicine implementation and enhancing physician well-being.

Methods

Study Design and Participants

This study used data from a large-scale cross-sectional survey conducted among health care professionals in Xi'an, China. A stratified cluster sampling design based on probability proportional to size was used. In total, 47 hospitals and 59 primary health care facilities were selected. All employees within the selected institutions were invited to participate in the survey. Data were collected between November 7 and December 8, 2023, using an online survey administered through REDCap (Research Electronic Data Capture; version 15.2.1; Vanderbilt University). The questionnaire encompassed multiple domains, including demographic characteristics, health status, work-related experiences, attitudes toward internet-based health care, and perceptions of workplace changes.

For the purposes of this study, only physicians were included in the analysis, as they are the primary providers of telemedicine-based clinical consultations. Physicians were identified based on professional licensure status, including those registered as licensed physicians or assistant physicians in either Western or Traditional Chinese medicine. A total of 20,396 physicians responded to the survey. After excluding incomplete and ineligible responses, 12,052 valid cases remained in the final analytic sample. A detailed overview of the data cleaning and selection process is presented in Figure 1. The reporting of this cross-sectional study adheres to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist (Checklist 1).

XSL•FO
RenderX

Figure 1. Flowchart of the sample selection process.

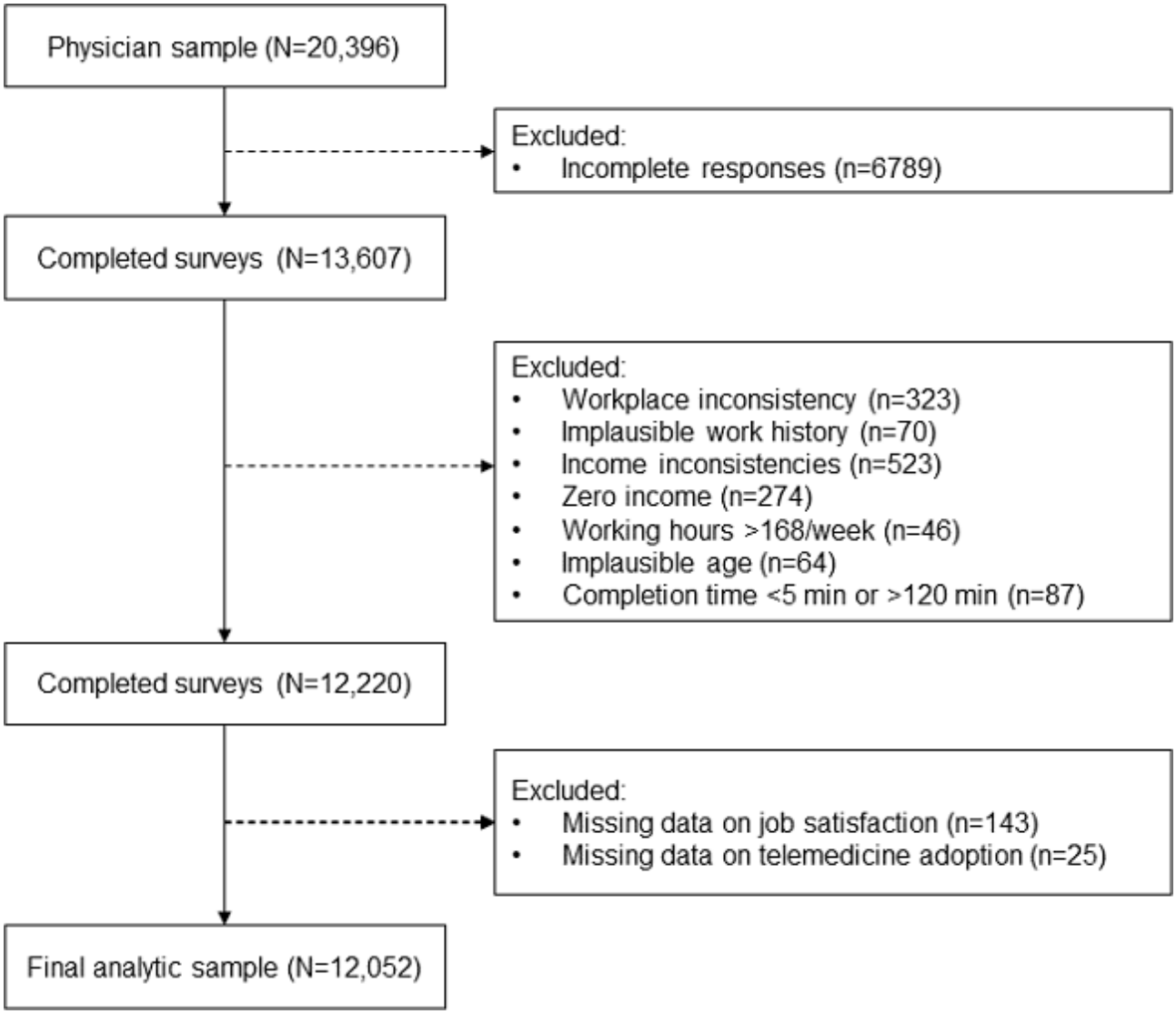

Measures

Job Satisfaction

Job satisfaction was measured using a single-item statement, "Overall, I am very satisfied with my current job." Respondents rated their agreement using a 6-point Likert-type scale with the following response options: (1) strongly disagree, (2) disagree, (3) somewhat disagree, (4) somewhat agree, (5) agree, and (6) strongly agree. The use of a single-item measure to capture overall job satisfaction has been validated in prior studies, demonstrating adequate levels of reliability and convergent validity [27,28]. The decision to adopt a 6-point scale, rather than a traditional 5-point or 7-point format, was based on evidence suggesting that even-point Likert scales can reduce central tendency bias by eliminating the neutral option and may provide greater discriminatory power across response categories [29,30].

Telemedicine Adoption

The primary independent variable in this study was physicians' adoption of telemedicine services, which reflects the integration of digital technologies into clinical practice. This variable was measured by asking physicians whether they had provided medical consultations or treatments via internet-based platforms. Responses were binary, with physicians answering "yes" if they had delivered services online and "no" if they had not.

Physician-Patient Relationship

To measure the physician-patient relationship from the physician's viewpoint, a 5-item physician-patient relationship scale was developed based on core relational dimensions identified in prior research and validated instruments such as the Patient-Doctor Relationship Questionnaire (PDRQ-9) [31]. The scale focused on 5 key relational components, including trust, respect, recognition, interpersonal satisfaction, and overall relational quality. Specifically, physicians were asked to rate their perceptions on the following aspects: (1) To what extent do you feel patients respect you? (2) To what extent do you feel patients trust the services you provide? (3) How would you rate the physician-patient relationship? (4) To what extent do you feel patients recognize the value of your work? (5) In most cases, how satisfied are patients with the care you provide? All items were rated on a 5-point Likert scale ranging from 1 ("very poor") to 5 ("very good"). The average score of the 5 items was calculated to represent the overall level of physician-patient

XSL•FO
RenderX

relationship, with higher scores indicating more positive relationships.

To assess the measurement validity of the 5-item physician-patient relationship scale, we conducted an exploratory factor analysis, which yielded a Kaiser-Meyer-Olkin (KMO) value of 0.82 and a significant Bartlett test of sphericity (χ^2_{10}=19,569.89; P<.001). All 5 items loaded strongly on the factor (loadings ranged from 0.59 to 0.74). The internal consistency of the scale was good, with a Cronbach α of 0.81.

Control Variables

We included three sets of control variables in the analysis: demographic, work-related characteristics, and psychological factors. Physician demographics comprised sex (male or female), age (categorized as ≤30, 31‑44, 45‑59, and ≥60 years), marital status (married or unmarried), and educational attainment (categorized as associate degree or below, bachelor's degree, master's degree, or doctoral degree).

Work-related characteristics included professional title, hospital type, employment status, medical specialty, managerial position, years in practice, years of service at the current institution, monthly income, and weekly working hours. In China, physicians hold professional titles that indicate their career stage. Professional titles were used as a proxy for career seniority and categorized into 4 levels according to the Chinese medical title system, including no title, junior, mid-level, and senior. As hierarchical levels are defined and routinely applied within the public hospital system in China, and such information was not collected in our data for private hospitals, hospitals were categorized into 4 types, including tertiary, secondary, primary, and private hospitals. Employment status was measured by asking whether physicians held a permanent (tenured) position at their institution. Responses were coded as 1 for permanent employee and 0 for contract employee. Medical specialty was classified according to the physician's department and included the following categories: clinical, medical technology, general practitioner, and others. The managerial position was coded as a binary variable indicating whether the physician held any administrative or leadership roles within their department, with 1 representing "yes" and 0 representing "no." Years in practice were measured by asking physicians how many years they have been practicing medicine since completing their training. Years of service at the current institution were assessed by asking physicians how long they had been employed at their current workplace.

Psychological factors were measured using indicators of work stress, depression, and anxiety. Work stress was assessed using "Overall, I feel a high level of work-related stress," with responses on a 6-point Likert scale ranging from strongly disagree to strongly agree. For analysis, responses were grouped into 3 categories: low stress (combining strongly disagree and disagree), moderate stress (combining somewhat disagree and somewhat agree), and high stress (combining agree and strongly agree). Depression and anxiety were assessed by self-reported clinical diagnosis, with participants indicating whether they have ever been diagnosed with depression or anxiety disorders.

Statistical Analysis

All statistical analyses were performed using STATA (version 17 MP; StataCorp LLC) and R (version 4.5.0; R Foundation for Statistical Computing). Statistical significance was set at a 2-sided P<.05.

Descriptive statistics were first calculated to summarize the characteristics of physicians. Given the skewed distribution of continuous variables, medians and IQRs were reported. Categorical variables were summarized using frequencies and percentages. To compare differences between physicians who provided telemedicine services and those who did not, the Mann-Whitney U test was applied for continuous variables, and the chi-square test was used for categorical variables.

Given the ordered nature of the job satisfaction outcome, we initially fitted an ordered logistic regression model to examine the association between telemedicine adoption and physicians' job satisfaction. The proportional odds assumption was evaluated using the Brant test [32]. The overall test indicated that the assumption was violated (χ^2_{116}=357.04; P<.001), although telemedicine adoption met the proportionality requirement (χ^2_4=4.16; P=.39). The detailed Brant test results are presented in Table S1 in Multimedia Appendix 1. To address these violations and avoid biased estimates, we used a partial proportional odds model, which relaxes the proportional odds constraint for variables that do not meet the assumption while maintaining the constraint for variables that do [33]. The final model included telemedicine adoption as the primary independent variable and adjusted for all control variables, including physicians' demographic characteristics, work-related factors, psychological factors, and the physician-patient relationship. Subgroup analyses were performed to examine potential variations in the relationship between telemedicine adoption and job satisfaction across all control variables. Continuous variables (eg, years in practice and weekly working hours) were dichotomized at their median values to ensure balanced group sizes for subgroup comparisons.

We used the Karlson-Holm-Breen (KHB) method to examine whether the physician-patient relationship mediated the association between telemedicine adoption and job satisfaction. The KHB approach allows for the decomposition of total effects into direct and indirect components in non-linear probability models [34]. This method effectively corrects for the rescaling problem that arises when comparing coefficients across nested nonlinear models, thereby providing unbiased estimates of mediation effects [35].

We conducted additional analyses to verify the robustness of our main finding. First, we re-estimated the main models using ordinary least squares regression as an alternative specification. Second, we used an alternative measure of job satisfaction based on a 5-point Likert item from the questionnaire, "Overall, I am very satisfied with my current job," with responses ranging from strongly disagree to strongly agree, to test the consistency of results across different satisfaction measures. Third, we used physicians' weekly online patient volume as a continuous measure of telemedicine adoption to assess the robustness of our findings.

XSL•FO
RenderX

Ethical Considerations

This study was approved by the Biomedical Ethics Committee of Xi'an Jiaotong University (XJTUAE2646). Electronic informed consent was obtained from all participants. All responses were anonymous, and no personal identifying information was collected. No compensation was provided to participants.

Results

Descriptive Statistics

The basic characteristics of physicians are displayed in Table 1. Among the 12,052 physicians, 1642 (13.62%) reported adopting telemedicine services, while 10,410 (86.38%) did not use telemedicine. The sample comprised 7351 out of 12,052 (61.08%) female physicians. Over half of the participants (6318/12,052; 52.42%) were aged 31-44 years. Most physicians were married. Regarding educational attainment, 5380 out of 12,052 (44.64%) held a bachelor's degree and 4512 out of 12,052 (37.44%) held a master's degree. Physicians with mid-level professional titles accounted for 39.17% (4712/12,052) of the sample, followed by those with a junior title (3850/12,052; 32.00%). More than half of the physicians (6765/12,052; 56.20%) were employed at tertiary hospitals. About 65.73% (7282/12,052) of the physicians held permanent positions. Clinicians accounted for 53.71% (6465/12,052) of the sample. The proportion of physicians holding managerial positions was 24.89% (3000/12,052). The median years of practice and years of service at the current institution were 12 and 8, respectively. The median monthly income was 8000 yuan (US $1159.34). The median number of weekly working hours was 50. Physicians who perceived work stress as "high" and "moderate" accounted for 48.14% (5800/12,052) and 42.98% (5179/12,052), respectively. The prevalence of depression and anxiety among participants was 4.32% (521/12,052) and 10.33% (1245/12,052), respectively.

XSL•FO
RenderX

Table . Characteristics of physicians stratified by telemedicine adoption.

Variables[a]	Total (N=12,052)	Telemedicine adoption		*P* value
		Yes (n=1642)	No (n=10,410)	
Sex, n (%)				<.001
Female	7351 (61.08)	797 (48.60)	6554 (63.05)	
Male	4684 (38.92)	843 (51.40)	3841 (36.95)	
Age (years), n (%)				<.001
≤30	2409 (19.99)	113 (6.88)	2296 (22.06)	
31-44	6318 (52.42)	858 (52.25)	5460 (52.45)	
45-59	3077 (25.53)	612 (37.27)	2465 (23.68)	
≥60	248 (2.06)	59 (3.59)	189 (1.82)	
Marital status, n (%)				<.001
Married	9862 (82.00)	1481 (90.30)	8381 (80.69)	
Unmarried[b]	2165 (18.00)	159 (9.70)	2006 (19.31)	
Education, n (%)				<.001
Associate degree or below	1449 (12.02)	137 (8.34)	1312 (12.60)	
Bachelor's degree	5380 (44.64)	658 (40.07)	4722 (45.36)	
Master's degree	4512 (37.44)	615 (37.45)	3897 (37.44)	
Doctoral degree	711 (5.90)	232 (14.13)	479 (4.60)	
Professional titles, n (%)				<.001
No title	266 (2.21)	13 (0.79)	253 (2.43)	
Junior	3850 (32.00)	220 (13.41)	3630 (34.93)	
Mid-level	4712 (39.17)	564 (34.39)	4148 (39.92)	
Senior	3203 (26.62)	843 (51.40)	2360 (22.71)	
Hospital type, n (%)				<.001
Tertiary hospitals	6765 (56.20)	1063 (64.78)	5702 (54.84)	
Secondary hospitals	2242 (18.62)	203 (12.37)	2039 (19.61)	
Primary hospitals	2233 (18.55)	163 (9.93)	2070 (19.91)	
Private hospitals	798 (6.63)	212 (12.92)	586 (5.64)	
Employment status, n (%)				<.001
Permanent employee	7282 (65.73)	1067 (75.46)	6215 (64.30)	
Contract employee	3797 (34.27)	347 (24.54)	3450 (35.70)	
Medical specialty, n (%)				<.001
Clinical	6465 (53.71)	1079 (65.75)	5386 (51.81)	
Medical technology	376 (3.12)	28 (1.71)	348 (3.35)	
General practitioner	1224 (10.17)	117 (7.13)	1107 (10.65)	
Others	3971 (32.99)	417 (25.41)	3554 (34.19)	
Managerial position, n (%)				<.001
No	9052 (75.11)	1033 (62.91)	8019 (77.03)	
Yes	3000 (24.89)	609 (37.09)	2391 (22.97)	
Years in practice, median (IQR)	12 (6 - 21)	17 (10 - 26)	11 (5 - 20)	<.001
Years of service at the current institution, median (IQR)	8 (3-15)	12 (6 - 20)	7 (3-14)	<.001

XSL•FO
RenderX

Variables[a]	Total (N=12,052)	Telemedicine adoption		*P* value
		Yes (n=1642)	No (n=10,410)	
Monthly income[c], median (IQR)	8000 (5000 - 10,000)	10,000 (7000 - 15,000)	7500 (5000 - 10,000)	<.001
Weekly working hours, median (IQR)	50 (40 - 60)	50 (45 - 60)	50 (40 - 60)	<.001
Work stress, n (%)				<.001
Low	1070 (8.88)	132 (8.04)	938 (9.01)	
Moderate	5179 (42.98)	633 (38.57)	4546 (43.68)	
High	5800 (48.14)	876 (53.38)	4924 (47.31)	
Depression, n (%)				.009
No	11,531 (95.68)	1551 (94.46)	9980 (95.87)	
Yes	521 (4.32)	91 (5.54)	430 (4.13)	
Anxiety, n (%)				<.001
No	10,807 (89.67)	1423 (86.66)	9384 (90.14)	
Yes	1245 (10.33)	219 (13.34)	1026 (9.86)	
Physician-patient relationship, median (IQR)	3.8 (3.4 - 4)	3.8 (3.4 - 4.2)	3.8 (3.4 - 4)	<.001

[a]Counts are based on valid (nonmissing) responses for each variable.

[b]Individuals who are single (never married), divorced, or widowed.

[c]Income is measured in Chinese Yuan (RMB), in 2023, the average exchange rate was approximately US $1=7.08 RMB.

Significant differences were observed between physicians who adopted telemedicine and those who did not across multiple characteristics. Compared to those who did not adopt telemedicine, physicians who adopted telemedicine were more likely to be male, older, married, hold a doctoral degree and senior professional titles, work in tertiary hospitals, be permanent employees, work in clinical departments, possess managerial roles, have longer years of practice and tenure at their current institution, and report higher monthly income (all *P*<.001). Additionally, telemedicine adopters reported higher work stress and were more likely to experience symptoms of depression and anxiety (all *P*<.01).

Association Between Telemedicine Adoption and Physician Job Satisfaction

We used a partial proportional odds model to estimate the association between telemedicine adoption and job satisfaction among physicians. The telemedicine adoption met the proportional odds assumption, indicating a consistent effect across cumulative logits. As shown in Table 2, telemedicine adoption was positively associated with higher job satisfaction. Specifically, physicians who adopted telemedicine had 17% higher odds of reporting a higher satisfaction level compared with those who did not (odds ratio [OR] 1.17, 95% CI 1.05 - 1.30).

XSL•FO
RenderX

Table . Association between telemedicine adoption and physician job satisfaction.

Variables	Physician job satisfaction, OR[a] (95% CI)				
	Very dissatisfied	Dissatisfied	Somewhat dissatisfied	Somewhat satisfied	Satisfied
Telemedicine adoption (reference=no)					
Yes	1.17[b] (1.05-1.30)	1.17[b] (1.05-1.30)	1.17[b] (1.05-1.30)	1.17[b] (1.05-1.30)	1.17[b] (1.05-1.30)
Sex (reference=female)					
Male	0.70[b] (0.55-0.90)	0.87 (0.74-1.02)	1.13[c] (1.02-1.25)	1.13[b] (1.03-1.23)	1.24[d] (1.11-1.39)
Age (years), (reference ≤30)					
31-44	0.97 (0.85-1.10)	0.97 (0.85-1.10)	0.97 (0.85-1.10)	0.97 (0.85-1.10)	0.97 (0.85-1.10)
45-59	1.13 (0.78-1.63)	1.08 (0.82-1.42)	0.99 (0.79-1.25)	1.10 (0.88-1.36)	1.26 (1.00-1.59)
≥60	1.45 (0.98-2.14)	1.45 (0.98-2.14)	1.45 (0.98-2.14)	1.45 (0.98-2.14)	1.45 (0.98-2.14)
Marital status (reference=unmarried)					
Married	1.78[d] (1.33-2.37)	1.40[b] (1.15-1.71)	1.25[b] (1.09-1.43)	1.08 (0.96-1.22)	1.05 (0.90-1.23)
Education (reference=associate degree or below)					
Bachelor's degree	0.80[b] (0.68-0.93)	0.80[b] (0.68-0.93)	0.80[b] (0.68-0.93)	0.80[b] (0.68-0.93)	0.80[b] (0.68-0.93)
Master's degree	0.73[b] (0.60-0.88)	0.73[b] (0.60-0.88)	0.73[b] (0.60-0.88)	0.73[b] (0.60-0.88)	0.73[b] (0.60-0.88)
Doctoral degree	0.71[b] (0.56-0.90)	0.71[b] (0.56-0.90)	0.71[b] (0.56-0.90)	0.71[b] (0.56-0.90)	0.71[b] (0.56-0.90)
Professional titles (reference=no title)					
Junior	0.80 (0.62-1.03)	0.80 (0.62-1.03)	0.80 (0.62-1.03)	0.80 (0.62-1.03)	0.80 (0.62-1.03)
Mid-level	0.58[d] (0.45-0.76)	0.58[d] (0.45-0.76)	0.58[d] (0.45-0.76)	0.58[d] (0.45-0.76)	0.58[d] (0.45-0.76)
Senior	0.60[d] (0.45-0.79)	0.60[d] (0.45-0.79)	0.60[d] (0.45-0.79)	0.60[d] (0.45-0.79)	0.60[d] (0.45-0.79)
Hospital type (reference =tertiary hospitals)					
Secondary hospitals	0.83[b] (0.75-0.93)	0.83[b] (0.75-0.93)	0.83[b] (0.75-0.93)	0.83[b] (0.75-0.93)	0.83[b] (0.75-0.93)
Primary hospitals	0.71[d] (0.62-0.81)	0.71[d] (0.62-0.81)	0.71[d] (0.62-0.81)	0.71[d] (0.62-0.81)	0.71[d] (0.62-0.81)
Private hospitals	4.67e-11[d] (4.36e-11 to 5.00e-11)	23.41 (0.01-590922.3)	0.48 (0.19-1.20)	1.69 (0.79-3.64)	2.30[c] (1.14-4.65)
Employment status (reference=contract employee)					
Permanent employee	0.86[d] (0.79-0.93)	0.86[d] (0.79-0.93)	0.86[d] (0.79-0.93)	0.86[d] (0.79-0.93)	0.86[d] (0.79-0.93)
Medical specialty (reference=clinical)					
Medical technology	1.25[c] (1.02-1.53)	1.25[c] (1.02-1.53)	1.25[c] (1.02-1.53)	1.25[c] (1.02-1.53)	1.25[c] (1.02-1.53)
General practitioner	1.02 (0.89-1.16)	1.02 (0.89-1.16)	1.02 (0.89-1.16)	1.02 (0.89-1.16)	1.02 (0.89-1.16)
Others	1.05 (0.97-1.13)	1.05 (0.97-1.13)	1.05 (0.97-1.13)	1.05 (0.97-1.13)	1.05 (0.97-1.13)
Managerial position (reference=no)					
Yes	1.44[d] (1.32-1.58)	1.44[d] (1.32-1.58)	1.44[d] (1.32-1.58)	1.44[d] (1.32-1.58)	1.44[d] (1.32-1.58)

XSL•FO
RenderX

Variables	Physician job satisfaction, OR[a] (95% CI)				
	Very dissatisfied	Dissatisfied	Somewhat dissatisfied	Somewhat satisfied	Satisfied
Years in practice	1.00 (0.99-1.01)	1.00 (0.99-1.01)	1.00 (0.99-1.01)	1.00 (0.99-1.01)	1.00 (0.99-1.01)
Years of service at the current institution	0.99[c] (0.99-1.00)	0.99[c] (0.99-1.00)	0.99[c] (0.99-1.00)	0.99[c] (0.99-1.00)	0.99[c] (0.99-1.00)
ln(Income)[e]	1.03 (0.99-1.08)	1.03 (0.99-1.08)	1.03 (0.99-1.08)	1.03 (0.99-1.08)	1.03 (0.99-1.08)
Weekly working hours	1.00[d] (0.99-1.00)	1.00[d] (0.99-1.00)	1.00[d] (0.99-1.00)	1.00[d] (0.991.00)	1.00[d] (0.99-1.00)
Work stress (reference=low)					
Moderate	4.76[d] (2.74-8.26)	2.44[d] (1.78-3.33)	0.73[b] (0.58-0.91)	0.24[d] (0.20-0.29)	0.21[d] (0.18-0.25)
High	1.04 (0.67-1.64)	1.05 (0.79-1.40)	0.57[d] (0.46-0.72)	0.38[d] (0.32-0.45)	0.33[d] (0.28-0.38)
Depression (reference=no)					
Yes	0.49[d] (0.34-0.71)	0.49[d] (0.37-0.65)	0.52[d] (0.41-0.65)	0.73[b] (0.58-0.92)	1.14 (0.80-1.63)
Anxiety (reference=no)					
Yes	0.42[d] (0.31-0.58)	0.62[d] (0.49-0.77)	0.79[b] (0.67-0.92)	0.66[d] (0.57-0.77)	0.72[b] (0.57-0.92)
Physician-patient relationship	3.05[d] (2.47-3.76)	3.03[d] (2.61-3.51)	2.79[d] (2.53-3.08)	3.41[d] (3.13-3.71)	4.54[d] (4.06-5.08)

[a]OR: odds ratio.

[b]$P<.01$.

[c]$P<.05$.

[d]$P<.001$.

[e]ln(Income) refers to the natural logarithm of monthly income.

In addition to telemedicine adoption, several covariates were significantly associated with job satisfaction, and many exhibited clear nonproportional effects across the satisfaction thresholds. For example, in the lowest satisfaction category, males are less likely to report higher satisfaction levels than females (OR 0.70, 95% CI 0.55 - 0.90), whereas at higher thresholds the direction of the association reversed, with males more likely to report higher satisfaction (eg, OR 1.24, 95% CI 1.11 - 1.39). Married physicians were more likely to report higher satisfaction than unmarried physicians at the lower thresholds, but the association diminished and was not statistically significant at the higher satisfaction levels. Physicians with moderate stress were more likely to report higher satisfaction at the lower thresholds but were less likely to report higher levels of satisfaction at the higher thresholds. High stress was associated with lower satisfaction at the higher thresholds. Physicians with depressive symptoms were less likely to report higher satisfaction at the lower and middle thresholds, with the association weakening at the highest satisfaction level. Physicians with anxiety were less likely to report higher satisfaction across all thresholds. Better physician-patient relationships were strongly and consistently associated with higher satisfaction across all thresholds, with stronger effects observed at the highest satisfaction level.

For covariates meeting the proportional-odds assumption, the estimated effects were consistent across all satisfaction thresholds. Physicians with higher education and those holding mid-level or senior professional titles exhibited significantly lower job satisfaction. Employment in secondary and primary hospitals was associated with reduced satisfaction relative to tertiary hospitals. Permanent employees reported lower levels of job satisfaction than contract employees (OR 0.86, 95% CI 0.79 - 0.93). Physicians in medical technology specialties reported higher satisfaction than those in clinical specialties (OR 1.25, 95% CI 1.02 - 1.53). Holding a managerial position showed a positive association with job satisfaction (OR 1.44, 95% CI 1.32 - 1.58). Longer tenure was associated with slightly higher odds of reporting lower satisfaction (OR 0.99, 95% CI 0.99 - 1.00). Longer weekly working hours were negatively associated with job satisfaction (OR 1.00, 95% CI 0.99 - 1.00).

Subgroup Analyses

Subgroup analyses were conducted to examine whether the association between telemedicine adoption and physician job satisfaction differed by physician characteristics. The association was generally consistent across groups, with no significant interaction effects detected (all P for interaction >.05), as illustrated in Figure 2. Notably, the association was statistically significant among female physicians (OR 1.20, 95% CI 1.03 - 1.40), those aged 31 - 44 years (OR 1.22, 95% CI 1.05 - 1.42), married physicians (OR 1.19, 95% CI 1.06 - 1.33), those with a master's degree (OR 1.21, 95% CI 1.02 - 1.44), permanent employees (OR 1.17, 95% CI 1.03 - 1.33), and physicians in the "others" specialty group (OR 1.29, 95% CI 1.05 - 1.59). In addition, significant associations were found among physicians with managerial positions (OR 1.25, 95% CI 1.03 - 1.50), those with fewer than 12 years of practice (OR

XSL•FO
RenderX

1.27, 95% CI 1.05 - 1.55), those with less than 8 years of service at the current institution (OR 1.25, 95% CI 1.04 - 1.51), those with monthly income below 8000 RMB (US $1159.34; OR 1.30, 95% CI 1.06 - 1.60), and those working fewer than 50 hours per week (OR 1.32, 95% CI 1.11 - 1.57). Finally, telemedicine adoption was positively associated with job satisfaction among physicians reporting moderate work stress (OR 1.27, 95% CI 1.07 - 1.51), those without depression (OR 1.15, 95% CI 1.02 - 1.28), and those without anxiety symptoms (OR 1.14, 95% CI 1.02 - 1.29).

Figure 2. Forest plot of subgroup analyses on the association between telemedicine adoption and physician job satisfaction. Income is measured in Chinese Yuan (RMB). In 2023, the average exchange rate was approximately US $1=7.08 RMB. OR: odds ratio.

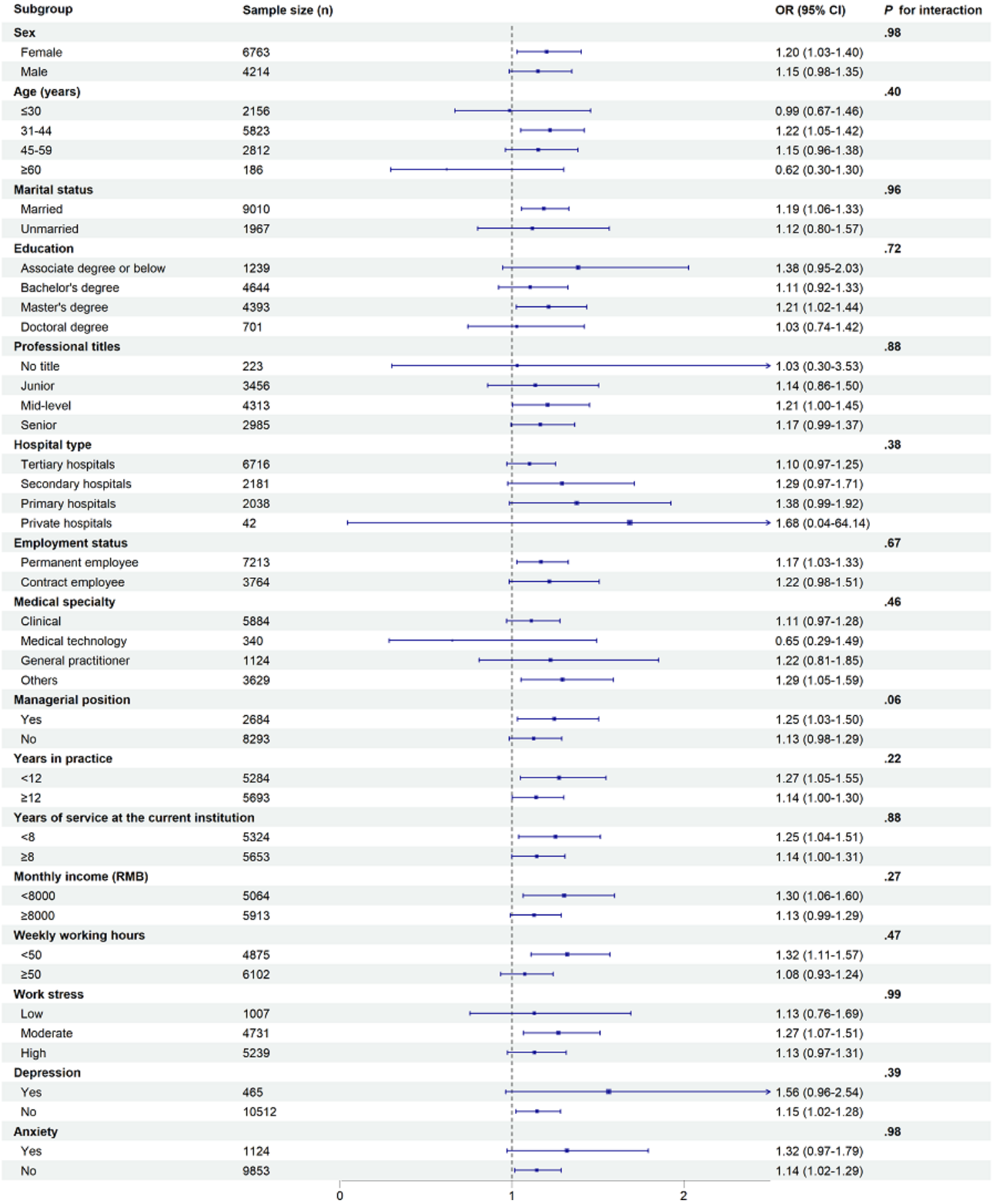

XSL•FO
RenderX

Mediation Analyses

Table 3 presents the results of the KHB decomposition analysis. The results indicated a significant partial mediation effect. The total effect of telemedicine adoption on job satisfaction was 0.33 (95% CI 0.17-0.50). After accounting for the physician-patient relationship, the direct effect decreased to 0.23 (95% CI 0.07-0.39), while the indirect effect through the physician-patient relationship was 0.10 (95% CI 0.07-0.13). This indirect pathway accounted for 30.30% of the total effect, suggesting that the physician-patient relationship plays an important mediating role between telemedicine adoption and job satisfaction.

Table . Mediation effect of physician-patient relationship on telemedicine and job satisfaction.

Effect	β (95% CI)	*P* value	Proportion of mediating effect
Total effect	.33 (0.17-0.50)	<.001	__[a]
Direct effect	.23 (0.07-0.39)	.006	69.70%
Indirect effect	.10 (0.07-0.13)	<.001	30.30%

[a]Not available.

Sensitivity Analysis

The sensitivity analyses supported the robustness of our main findings. First, we re-estimated the models using ordinary least squares regression as a robustness check, and the results remained consistent with the original models in terms of direction and statistical significance (Table S2 in Multimedia Appendix 1). Second, when job satisfaction was measured using an alternative 5-point Likert item, the association between telemedicine adoption and job satisfaction remained robust (Table S3 in Multimedia Appendix 1). Third, we replaced the binary indicator of telemedicine adoption with a continuous measure reflecting the number of patients a physician consulted online per week. The results remained robust, further supporting the validity of our main findings (Table S4 in Multimedia Appendix 1).

Discussion

Principal Findings

In this cross-sectional study of physicians, telemedicine adoption was significantly associated with higher levels of job satisfaction. Physicians who adopted telemedicine reported higher levels of job satisfaction compared with those who did not, and this association remained robust across multiple sensitivity analyses. Mediation analysis further suggested that the physician-patient relationship partially explained this association. These findings contribute to a growing body of literature examining how digital health technologies affect physicians' work experiences and underscore the importance of preserving strong interpersonal relationships in virtual care settings.

Comparison With Prior Work

The telemedicine adoption rate in our study was 13.62%, which appears relatively low compared to telemedicine adoption rates reported in other countries, where physician uptake during the COVID-19 pandemic ranged from approximately 30% to 87% [36-38]. This difference may be partly attributable to variations in the measurement of telemedicine adoption. In our study, adoption was defined specifically as whether physicians had provided medical consultations or treatments via internet-based platforms, whereas some international studies use broader definitions of telemedicine that include not only direct patient consultations but also activities such as exchanging referrals with other providers, ordering and reviewing diagnostic tests, or other forms of remote care, thereby yielding higher reported rates. Additionally, regional differences and health care system characteristics in Xi'an and broader China may play a role.

Consistent with previous research, our study found that telemedicine adoption is positively associated with physician job satisfaction [19,20]. This association may reflect several potential advantages of telehealth services that directly influence physicians' work experiences. Previous studies have found that telemedicine enhances scheduling flexibility and work-life balance among physicians [19], reduces administrative burdens [39], and increases practice productivity [40], all of which are closely linked to lower stress and higher job satisfaction. Additionally, telemedicine may support more efficient patient follow-up and continuity of care, particularly in primary care [41] and chronic disease management settings [42,43], thereby improving the clinical workflow. Taken together, these features may contribute to a more sustainable and fulfilling clinical practice.

Subgroup analyses demonstrated that the positive association between telemedicine adoption and physician job satisfaction was generally consistent across various demographic and professional groups, with no significant interaction effects observed. These findings imply that telemedicine adoption is broadly beneficial for physician job satisfaction, regardless of demographic or work-related characteristics. Nonetheless, the significant associations within certain subgroups indicate that the effect may be more detectable among physicians sharing specific characteristics.

The mediation analysis revealed a significant partial mediating effect, indicating that telemedicine adoption is associated with a more positive physician-patient relationship, which in turn contributes to higher job satisfaction. Specifically, physicians who used telemedicine tended to report higher physician-patient relationship scores. Empirical studies similarly show that telemedicine enables more frequent, timely, and continuous contact, which strengthens physicians' sense of connection and familiarity with their patients, thereby fostering a more trusting and collaborative physician-patient relationship [44,45]. Such positive relational experiences have been associated with higher

XSL•FO
RenderX

job satisfaction [46], partly because strong physician-patient relationships enhance physicians' sense of professional efficacy, reduce workplace stress, and reinforce the meaningfulness of their clinical work. Overall, these findings suggest that improvements in physician-patient relationships represent an important pathway through which telemedicine may enhance physicians' job satisfaction.

Building on these findings, several policy implications can be drawn. Health care systems should facilitate telemedicine adoption through targeted training and resource allocation tailored to physicians' diverse needs. The design of telemedicine platforms should explicitly incorporate features aimed at improving the quality of the physician-patient relationship. Additionally, workload management strategies and support mechanisms are necessary to mitigate potential burnout associated with telemedicine use.

Limitations

Several limitations of this study warrant careful consideration. First, the cross-sectional design inherently restricts the ability to draw causal inferences regarding the relationship between telemedicine use and physicians' job satisfaction. Longitudinal studies are needed to establish temporal dynamics and causal pathways. Second, job satisfaction was assessed using a single-item measure. While this approach has been validated in previous research and is suitable for large-scale surveys with space constraints, multi-item instruments could offer a more detailed understanding of satisfaction dimensions. Third, the survey measured telemedicine use based on whether physicians reported using telemedicine for diagnosis or treatment, without distinguishing specific modalities such as video consultations, asynchronous messaging, or phone visits. Future studies with more detailed information on specific telemedicine modalities could clarify how different forms of telemedicine influence physicians' job satisfaction. Finally, although a stratified probability proportional to size sampling approach was used to enhance representativeness, the sample is limited to physicians within Xi'an, a specific urban context in China. This geographic concentration may limit the external validity of findings, as organizational cultures, telemedicine infrastructure, and health care policies vary substantially across regions and countries.

Conclusions

In this cross-sectional study, telemedicine adoption was positively associated with physician job satisfaction. This association was partially mediated by the physician-patient relationship, suggesting that enhanced communication and connection through telemedicine contribute to higher satisfaction. To optimize these benefits, physicians need to use telemedicine appropriately to balance its benefits and challenges. Meanwhile, health care institutions should provide targeted support, including communication training and stress management, to promote effective and sustainable telemedicine practice.

Acknowledgments

The authors would like to thank all participating physicians for their time and valuable contributions to this study. We are also grateful to the administrative staff and coordinators at the participating hospitals and primary health care facilities in Xi'an for their support during data collection.

Funding

This study was funded by leading talents project in philosophy and social sciences of the National Social Science Foundation of China (2022LJRC02).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: XZ
Formal analysis: XZ
Funding acquisition: ZZ
Methodology: XZ
Supervision: ZZ, PCC
Visualization: XZ
Writing – original draft: XZ
Writing – review & editing: XZ, ZZ, ZW, GL, PCC

Conflicts of Interest

None declared.

Multimedia Appendix 1
Results of Brant test and robustness checks.
[DOCX File, 43 KB - jmir_v28i1e82285_app1.docx]

XSL•FO
RenderX

Checklist 1
The STROBE checklist.
[DOCX File, 33 KB - jmir_v28i1e82285_app2.docx]

References

1. World Health Organization. Telemedicine: Opportunities and Developments in Member States: Report on the Second Global Survey on eHealth: World Health Organization; 2010:93 URL: https://iris.who.int/handle/10665/44497 [accessed 2026-03-19]
2. Bashshur RL, Shannon GW, Smith BR, et al. The empirical foundations of telemedicine interventions for chronic disease management. Telemed J E Health 2014 Sep;20(9):769-800. [doi: 10.1089/tmj.2014.9981] [Medline: 24968105]
3. Suran M. Increased use of medicare telehealth during the pandemic. JAMA 2022 Jan 25;327(4):313. [doi: 10.1001/jama.2021.23332]
4. Gullslett MK, Ronchi E, Lundberg L, et al. Telehealth development in the WHO European region: results from a quantitative survey and insights from Norway. Int J Med Inform 2024 Nov;191:105558. [doi: 10.1016/j.ijmedinf.2024.105558] [Medline: 39084085]
5. Orozco-Beltran D, Sánchez-Molla M, Sanchez JJ, Mira JJ, ValCrònic Research Group. Telemedicine in primary care for patients with chronic conditions: the ValCrònic quasi-experimental study. J Med Internet Res 2017 Dec 15;19(12):e400. [doi: 10.2196/jmir.7677] [Medline: 29246881]
6. Lewinski AA, Walsh C, Rushton S, et al. Telehealth for the longitudinal management of chronic conditions: systematic review. J Med Internet Res 2022 Aug 26;24(8):e37100. [doi: 10.2196/37100] [Medline: 36018711]
7. Fortney JC, Bauer AM, Cerimele JM, et al. Comparison of teleintegrated care and telereferral care for treating complex psychiatric disorders in primary care: a pragmatic randomized comparative effectiveness trial. JAMA Psychiatry 2021 Nov 1;78(11):1189-1199. [doi: 10.1001/jamapsychiatry.2021.2318] [Medline: 34431972]
8. Wu M, Li C, Hu T, et al. Effectiveness of telecare interventions on depression symptoms among older adults: systematic review and meta-analysis. JMIR Mhealth Uhealth 2024 Jan 17;12:e50787. [doi: 10.2196/50787] [Medline: 38231546]
9. Di Cerbo A, Morales-Medina JC, Palmieri B, Iannitti T. Narrative review of telemedicine consultation in medical practice. Patient Prefer Adherence 2015;9:65-75. [doi: 10.2147/PPA.S61617] [Medline: 25609928]
10. Greenhalgh T, Rosen R, Shaw SE, et al. Planning and evaluating remote consultation services: a new conceptual framework incorporating complexity and practical ethics. Front Digit Health 2021;3:726095. [doi: 10.3389/fdgth.2021.726095] [Medline: 34713199]
11. Shah KM, Shah MM, Shah IP, Patel HH. The role of technology in enhancing clinical decision-making for health professiona. Front Health Inform 2024 Dec 26:920-929. [doi: 10.53555/fhi.v13i7.1577]
12. Kaihlanen AM, Laukka E, Nadav J, et al. The effects of digitalisation on health and social care work: a qualitative descriptive study of the perceptions of professionals and managers. BMC Health Serv Res 2023 Jun 30;23(1):714. [doi: 10.1186/s12913-023-09730-y] [Medline: 37386423]
13. Scheepers RA. Physicians' professional performance: an occupational health psychology perspective. Perspect Med Educ 2017 Dec;6(6):425-428. [doi: 10.1007/s40037-017-0382-9] [Medline: 29064072]
14. Linn LS, Brook RH, Clark VA, Davies AR, Fink A, Kosecoff J. Physician and patient satisfaction as factors related to the organization of internal medicine group practices. Med Care 1985 Oct;23(10):1171-1178. [doi: 10.1097/00005650-198510000-00006] [Medline: 4058071]
15. Buchbinder SB, Wilson M, Melick CF, Powe NR. Primary care physician job satisfaction and turnover. Am J Manag Care 2001 Jul;7(7):701-713. [Medline: 11464428]
16. Hall CB, Brazil K, Wakefield D, Lerer T, Tennen H. Organizational culture, job satisfaction, and clinician turnover in primary care. J Prim Care Community Health 2010 Apr 1;1(1):29-36. [doi: 10.1177/2150131909360990] [Medline: 23804066]
17. Hodkinson A, Zhou A, Johnson J, et al. Associations of physician burnout with career engagement and quality of patient care: systematic review and meta-analysis. BMJ 2022 Sep 14;378:e070442. [doi: 10.1136/bmj-2022-070442] [Medline: 36104064]
18. Williams ES, Skinner AC. Outcomes of physician job satisfaction: a narrative review, implications, and directions for future research. Health Care Manage Rev 2003;28(2):119-139. [doi: 10.1097/00004010-200304000-00004] [Medline: 12744449]
19. Zaresani A, Scott A. Does digital health technology improve physicians' job satisfaction and work-life balance? A cross-sectional national survey and regression analysis using an instrumental variable. BMJ Open 2020 Dec 12;10(12):e041690. [doi: 10.1136/bmjopen-2020-041690] [Medline: 33310807]
20. Sengupta A, Sarkar S, Bhattacherjee A. The relationship between telemedicine tools and physician satisfaction, quality of care, and patient visits during the COVID-19 pandemic. Int J Med Inform 2024 Oct;190:105541. [doi: 10.1016/j.ijmedinf.2024.105541] [Medline: 38996654]
21. Lawrence K, Nov O, Mann D, Mandal S, Iturrate E, Wiesenfeld B. The impact of telemedicine on physicians' after-hours electronic health record "Work Outside Work" during the COVID-19 pandemic: retrospective cohort study. JMIR Med Inform 2022 Jul 28;10(7):e34826. [doi: 10.2196/34826] [Medline: 35749661]

XSL•FO
RenderX

22. Leigh JP, Kravitz RL, Schembri M, Samuels SJ, Mobley S. Physician career satisfaction across specialties. Arch Intern Med 2002 Jul 22;162(14):1577-1584. [doi: 10.1001/archinte.162.14.1577] [Medline: 12123400]
23. Haleem A, Javaid M, Singh RP, Suman R. Telemedicine for healthcare: capabilities, features, barriers, and applications. Sens Int 2021;2:100117. [doi: 10.1016/j.sintl.2021.100117] [Medline: 34806053]
24. Grant AM. Relational job design and the motivation to make a prosocial difference. AMR 2007 Apr;32(2):393-417. [doi: 10.5465/amr.2007.24351328]
25. Gkliati A, Saiti A. Work engagement and job satisfaction in the medical sector: two aspects of psychological well-being among medical staff. Int J Health Sci 2022 Jul 25;6(S6):7753-7761. [doi: 10.53730/ijhs.v6nS6.11070]
26. Bailey C, Yeoman R, Madden A, Thompson M, Kerridge G. A review of the empirical literature on meaningful work: progress and research agenda. Hum Resour Dev Rev 2019 Mar;18(1):83-113. [doi: 10.1177/1534484318804653]
27. Wanous JP, Reichers AE, Hudy MJ. Overall job satisfaction: how good are single-item measures? J Appl Psychol 1997 Apr;82(2):247-252. [doi: 10.1037/0021-9010.82.2.247] [Medline: 9109282]
28. Dolbier CL, Webster JA, McCalister KT, Mallon MW, Steinhardt MA. Reliability and validity of a single-item measure of job satisfaction. Am J Health Promot 2005;19(3):194-198. [doi: 10.4278/0890-1171-19.3.194] [Medline: 15693347]
29. Quality of psychology test between Likert scale 5 and 6 points. J Soc Sci 2010 Mar 1;6(3):399-403. [doi: 10.3844/jssp.2010.399.403]
30. Leung SO. A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-Point Likert Scales. J Soc Serv Res 2011 Jul;37(4):412-421. [doi: 10.1080/01488376.2011.580697]
31. Van der Feltz-Cornelis CM, Van Oppen P, Van Marwijk HWJ, De Beurs E, Van Dyck R. A patient-doctor relationship questionnaire (PDRQ-9) in primary care: development and psychometric evaluation. Gen Hosp Psychiatry 2004;26(2):115-120. [doi: 10.1016/j.genhosppsych.2003.08.010] [Medline: 15038928]
32. Long JS, Freese J. Regression Models for Categorical Dependent Variables Using Stata, 3rd edition: College Station, Texas: Stata Press; 2014.
33. Williams R. Generalized ordered logit/partial proportional odds models for ordinal dependent variables. Stata J 2006 Feb;6(1):58-82. [doi: 10.1177/1536867X0600600104]
34. Breen R, Karlson KB, Holm A. Total, direct, and indirect effects in logit and probit models. Sociol Methods Res 2013 May;42(2):164-191. [doi: 10.1177/0049124113494572]
35. Karlson KB, Holm A, Breen R. Comparing regression coefficients between same-sample nested models using logit and probit. Sociol Methodol 2012 Aug;42(1):286-313. [doi: 10.1177/0081175012444861]
36. Majority of UK physicians see continued rise in telemedicine usage, reveals globaldata. GlobalData. 2024. URL: https://www.globaldata.com/media/pharma/majority-uk-physicians-see-continued-rise-telemedicine-usage-reveals-globaldata/ [accessed 2025-08-11]
37. Scott A, Bai T, Zhang Y. Association between telehealth use and general practitioner characteristics during COVID-19: findings from a nationally representative survey of Australian doctors. BMJ Open 2021 Mar 24;11(3):e046857. [doi: 10.1136/bmjopen-2020-046857] [Medline: 33762248]
38. Pylypchuk Y, Barker W. Use of telemedicine among office-based physicians. In: ASTP Health IT Data Brief: Office of the Assistant Secretary for Technology Policy; 2021. [Medline: 39475515]
39. Hopkins L, Pedwell G. The COVID PIVOT - re-orienting child and youth mental health care in the light of pandemic restrictions. Psychiatr Q 2021 Sep;92(3):1259-1270. [doi: 10.1007/s11126-021-09909-8] [Medline: 33754240]
40. Stefos T, Carey K, Shen ML, Poe S, Oh DH, Moran E. The effect of telehealth services on provider productivity. Med Care 2021 May 1;59(5):456-460. [doi: 10.1097/MLR.0000000000001529] [Medline: 33821831]
41. Carrillo de Albornoz S, Sia KL, Harris A. The effectiveness of teleconsultations in primary care: systematic review. Fam Pract 2022 Jan 19;39(1):168-182. [doi: 10.1093/fampra/cmab077] [Medline: 34278421]
42. Mabeza RMS, Maynard K, Tarn DM. Influence of synchronous primary care telemedicine versus in-person visits on diabetes, hypertension, and hyperlipidemia outcomes: a systematic review. BMC Prim Care 2022 Mar 21;23(1):52. [doi: 10.1186/s12875-022-01662-6] [Medline: 35313804]
43. Guarino M, Cossiga V, Fiorentino A, Pontillo G, Morisco F. Use of telemedicine for chronic liver disease at a single care center during the COVID-19 pandemic: prospective observational study. J Med Internet Res 2020 Sep 21;22(9):e20874. [doi: 10.2196/20874] [Medline: 32896833]
44. Sandberg J, Trief PM, Izquierdo R, et al. A qualitative study of the experiences and satisfaction of direct telemedicine providers in diabetes case management. Telemed J E Health 2009 Oct;15(8):742-750. [doi: 10.1089/tmj.2009.0027] [Medline: 19780691]
45. Maria ARJ, Serra H, Castro MG, Heleno B. Interaction at the primary-secondary care interface: patients' and physicians' perceptions of teleconsultations. Digit Health 2022;8:20552076221133698. [doi: 10.1177/20552076221133698] [Medline: 36465985]
46. Li T, Guan L, Zhang R, et al. Roles of doctor-patient relationship perception and job satisfaction in the impact of workplace violence on medical professionals' turnover intentions in the early phase of COVID-19: a cross-sectional study in China. BMJ Open 2023 Sep 4;13(9):e074744. [doi: 10.1136/bmjopen-2023-074744] [Medline: 37666559]

XSL•FO
RenderX

Abbreviations

EHR: electronic health record
KHB: Karlson-Holm-Breen
KMO: Kaiser-Meyer-Olkin
OR: odds ratio
PDRQ-9: Patient-Doctor Relationship Questionnaire-9 item
REDCap: Research Electronic Data Capture
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
WHO: World Health Organization

Edited by A Stone; submitted 12.Aug.2025; peer-reviewed by GY Han, R Guo; revised version received 20.Jan.2026; accepted 20.Jan.2026; published 31.Mar.2026.

Please cite as:
Zhai X, Zhou Z, Wang Z, Liu G, Coyte PC
Association Between Telemedicine Adoption and Physician Job Satisfaction: Cross-Sectional Study
J Med Internet Res 2026;28:e82285
URL: https://www.jmir.org/2026/1/e82285
doi:10.2196/82285

© Xiaohui Zhai, Zhongliang Zhou, Zhichao Wang, Guanping Liu, Peter C Coyte. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Facilitators of and Barriers to Global Digital Oral Health: Mixed Methods Study

Elham Emami[1], DDS, MSc, PhD; Pascaline Kengne Talla[1], DDS, MPH, PhD; Camille Inquimbert[2], DDS, PhD; Nicolas Giraudeau[3], DDS, PhD

[1]Faculty of Dental Medicine and Oral Health Sciences, McGill University, Montreal, QC, Canada

[2]Public Health Department, Faculty of Dental Medicine, Institut Desbrest d'épidémiologie et de santé publique, IDESP UMR UA11 INSERM, University of Montpellier, Montpellier, Occitanie, France

[3]Centre d'études politiques et sociales, Centre national de la recherche scientifique, University of Montpellier, Montpellier, France

Corresponding Author:
Elham Emami, DDS, MSc, PhD
Faculty of Dental Medicine and Oral Health Sciences
McGill University
2001 McGill College Ave
Montreal, QC, H3A 1G1
Canada
Phone: 1 5143987222
Email: elham.emami@mcgill.ca

Abstract

Background: Digital oral health builds on the broader framework of eHealth, leveraging digital technologies to improve patient care, increase access to dental services, and enhance oral health outcomes. However, health care organizations and institutions encounter challenges in implementing digital oral health interventions across various levels. Addressing these challenges requires a comprehensive understanding of the barriers and facilitators that influence its successful adoption.

Objective: This study aimed to explore the facilitators of and barriers to the implementation of digital oral health programs from the perspective of chief dental officers from countries across the World Health Organization (WHO) regions.

Methods: This study is part of a broader investigation into global readiness for digital oral health. Participants were the 144 chief dental officers or designated oral health officials within ministries of health across the 6 WHO regions. An explanatory sequential mixed methods design was used across 2 phases. In the quantitative phase, an online survey was administered using the WHO's global survey on eHealth instrument. Some items were modified slightly to be applied to the field of dentistry. Descriptive statistics were used to present the quantitative data. In the qualitative phase, data were collected through virtual interviews, using an interview guide developed based on preliminary findings from the quantitative phase, the technology acceptance model, and the eHealth readiness assessment tool. The qualitative data were analyzed using thematic analysis.

Results: The survey response rate was 70.1% (101/144). The qualitative phase involved in-depth interviews with 15 participants. The findings were integrated under 2 broad themes of facilitators and barriers. Perceived facilitators included the existence of national policies and guidelines on eHealth. Approximately 63.9% (53/83) of the respondents indicated the presence of a national oral health policy in their countries. Capacity building, motivation of health care providers and academic leadership, digital health training for students or professionals, and WHO support to implement the mOral Health program were the other facilitators. The strongest barriers were a lack of funding to develop and support digital health programs, lack of norms and standards to guarantee application interoperability, and lack of equipment and/or connectivity. Approximately 45.1% (37/82) of the participants reported having government-sponsored mobile health programs, while 31.7% (26/82) reported having no financial support for the implementation of national digital oral health programs. Furthermore, lack of evidence on the effectiveness and cost-effectiveness of programs was highlighted as a barrier by 73.8% (59/80) and 73% (57/78) of the participants, respectively.

Conclusions: The results of this study enabled the identification of key barriers to and enablers of the implementation of digital oral health programs in WHO member countries. Supportive governmental policies and adequate funding and investment in digital infrastructure and technologies are essential to mitigate digital oral health–related challenges.

(J Med Internet Res 2026;28:e76236) doi:10.2196/76236

XSL•FO
RenderX

KEYWORDS

digital health; mobile health; mHealth; oral health; health policy; World Health Organization

Introduction

Digital health has been defined as "the cultural transformation of how disruptive technologies that provide digital and objective data accessible to both caregivers and patients lead to an equal level doctor-patient relationship with shared decision-making and the democratization of care" [1]. Digital health supports the Quintuple Aim [2] in health care by improving patients' experiences and the population's health, and increasing accessibility and equity in care while reducing the costs of the health care system [3]. This definition applies to dentistry as well, as oral health is an integral part of overall health [4]. In this context, digital oral health reflects the same principles, leveraging tools such as tele-dentistry, artificial intelligence (AI)–powered diagnostics, digital imaging, and electronic health records to enhance access, precision, and patient engagement in oral health care.

Over the past decade, the World Health Organization (WHO) has actively supported the integration of digital technologies into health care systems, most notably through its global strategy on digital health 2020-2025 [2]. Furthermore, the global strategy and action plan on oral health recognizes digital health interventions as catalysts to optimize oral health equity and increase environmental sustainability to achieve universal coverage. The WHO has also issued implementation guidelines for mobile technologies as part of the mOral Health Program [5]. Although digital innovations such as computer-aided design and manufacturing, digital imaging, and AI [5,6] have progressed rapidly in dentistry, their integration into oral health care systems has not kept the pace. This gap is influenced by several key factors. Among these, shared values, professional attitudes and behaviors, and organizational readiness, along with enablers such as technological infrastructure, workforce capacity, and available funds, are essential for the successful adoption and sustained use of digital oral health interventions [6,7].

The recent global COVID-19 pandemic has significantly accelerated the adoption of tele-dentistry across health care systems worldwide [7]. Faced with the urgent need to provide care while minimizing the risk of virus transmission, many dental practices and health care systems rapidly integrated telehealth technologies [7]. This shift allowed for the partial fulfillment of emergency dental care needs, enabling patients to receive timely consultations and guidance remotely [8,9]. Despite this progress, evidence indicates that the integration of digital technologies in oral health care organizations and institutions continues to face various barriers at the micro, meso, and macro levels and across the 4 domains of integration: clinical, professional, organizational, and systemic [10,11].

This indicates the need for the global scale-up of a digital oral health plan and a transformative shift in the delivery of dental care worldwide through the adoption of digital platforms that facilitate remote consultations, modernize electronic oral health care records and data management, and enable virtual and AI-driven diagnostic systems [12]. This global scale-up of the digital plan will only be successful if it is grounded in a system-based approach and if it is universally applicable to all countries. This also depends on international collaboration, universal standardization, and universal protocols [13]. This necessitates an understanding of organizational readiness, or "the degree to which users, healthcare institutions, and the healthcare system itself are prepared to participate and succeed with e-health implementation" [14-16]. The dimensions of e-readiness include governmental readiness, or the presence of relevant policies and funding; organizational readiness, or the presence of policies and management support; and, finally, societal readiness, or the interaction of health care institutions with their government and communities, as well as the readiness of health care providers and patients [17-19].

To understand the state of readiness of countries for digital oral health interventions and inform a comprehensive digital oral health strategy, a good understanding of global readiness, challenges, opportunities, capacities, and obstacles is required. Therefore, built on a strong partnership among the WHO, the University of Montpellier, and McGill University Faculty of Dental Medicine and Oral Health Sciences, and with the overarching goal of optimizing patient-centered care and value-based health care [20], a broad study on global readiness for digital oral health was conducted. This paper specifically addresses the following research question: *what are the barriers to and facilitators of implementing digital oral health programs as perceived by chief dental officers (CDOs) across WHO regions?*

Methods

Overview

This study is part of a broader study on global readiness for digital oral health, and it focuses on presenting the results regarding facilitators of and barriers to global digital oral health. The reporting of this study adheres to the GRAMMS (Good Reporting of A Mixed Methods Study) guidelines [21]. The GRAMMS checklist is available in Multimedia Appendix 1.

Study Design and Methods

An explanatory sequential mixed methods [22] study was conducted in 2 phases. The initial phase used a quantitative approach. Subsequently, a qualitative phase was undertaken aimed at contextualizing and elaborating on the quantitative findings [23]. This design is valuable for gaining an in-depth understanding of the context of the study and supports policy implementation. The quantitative and qualitative data were collected sequentially from April 2022 to May 2023 but were analyzed simultaneously and in an iterative way to validate and confirm the results (triangulation) [24].

Conceptual Frameworks

The technology acceptance model [25] and the eHealth readiness assessment tool by Khoja et al [17] were used as conceptual frameworks for the qualitative study. The technology acceptance

XSL•FO
RenderX

model is a well-known theory widely used in research on health IT, explaining how factors such as perceived ease of use, perceived usefulness, and behavioral intention influence technology acceptance. The eHealth readiness assessment tool evaluates e-readiness among managers and health care providers across the categories of core, societal, and policy readiness, as well as technological readiness and learning readiness.

Study Participants

The potential study participants were invited through a collaboration with the WHO. One of the study team members (NG), who acts as a WHO expert in digital health, facilitated the communication with CDOs via email. The invitation email included an informative letter summarizing the study objectives and methods, including the anonymous online survey. The email addresses we contacted were valid as none of the messages were returned as undeliverable.

The eligibility criteria were being a CDO or the responsible oral health officer in the ministry of health of countries that were WHO members and represented the 6 WHO regions. The 6 WHO regions encompass the African Region, the Region of the Americas, the Eastern Mediterranean Region, the European Region, the South-East Asia Region, and the Western Pacific Region. These countries varied with regard to income level and included high, upper-middle, lower-middle, and low income levels. The income level was defined using the World Development Indicators database gross national income per capita, which categorizes economies based on their gross national income per capita into different income groups [26]. The invitation was sent only to 144 CDOs or responsible officers from 6 WHO regions and 194 WHO member countries. The FDI World Dental Federation maintains a section dedicated to CDOs and dental public health worldwide.

Data Collection

Phase 1: Quantitative Data

An online survey was conducted using the WHO's global survey on eHealth instrument [27]. The items of this questionnaire were modified slightly to be applied to the field of dentistry without any change in the constructs, which would have required further validation. Translation services were used to translate the modified questionnaire into other languages according to the participants' language preference.

The survey instrument comprised nine sections: (1) demographic characteristics of the participants, (2) digital health foundations, (3) mobile health (mHealth) and telehealth, (4) big data, (5) electronic health records, (6) use of e-learning in health sciences, (7) social media, (8) barriers to implementing digital health programs, and (9) digital health network. However, for this study, data related only to facilitators and barriers were included (Multimedia Appendix 2). These included responses on the existence of national digital and oral health policies, telehealth policies, mHealth programs, funding, infrastructure, and WHO support. In addition, items on training and the interoperability, effectiveness, and cost-effectiveness of the programs were included.

The modified version was pilot-tested among a group of clinicians and oral health administrators, who evaluated the relevance, length, and overall comprehensiveness of the items. To enhance the response rate, reminders were sent to participants 2 and 3 weeks after the initial distribution of the survey. Participants were also asked to indicate their willingness to take part in the qualitative phase of the study. Participants were asked to sign a consent form before filling out the survey, which was launched through the WHO platform. Participants who agreed to take part in the study were asked to indicate their preferred language for the survey.

Phase 2

One-on-one, virtual, face-to-face 60- to 90-minute interviews were conducted to collect the qualitative data. Maximum variation and purposeful sampling techniques were used to represent the WHO's 6 regions and varying income levels [28]. This approach allowed for capturing the perspectives of participants from different countries with diverse policies, cultural contexts, socioeconomic backgrounds, and educational systems [28]. The interviews were conducted virtually using secure platforms such as Zoom (Zoom Video Communications), Webex Meetings (Cisco Systems), or Microsoft Teams, depending on the accessibility, preferences, and convenience of the study participants. The interviews were conducted in English or French as requested and selected by the study participants. They were carried out by 2 members of the research team (PKT and CI) fluent in both languages and with previous training and expertise in qualitative research. Using the study's conceptual frameworks and the results of the quantitative study, an interview guide with open-ended questions was developed to collect "information-rich" data and capture perspectives on the facilitators and barriers related to the implementation of digital oral health solutions, including the mOral Health program [29] (Multimedia Appendix 3). Data collection and analysis were conducted concurrently. Interviews continued until saturation was reached to ensure that no new information would be obtained from further interviews [30]. All interviews were recorded and transcribed verbatim.

Ethical Considerations

Ethics approval for this study was granted by the institutional review board (A04-E17-21A) of McGill University. Participation in the study was completely voluntary, and informed consent was obtained from all participants. Participants were clearly informed that any data collected during both the quantitative and qualitative phases would be deidentified to maintain confidentiality and ensure anonymity. No monetary or other forms of compensation were provided for participation.

Data Analyses

Quantitative Phase

Non–English- or French-language surveys were translated into English. For the purpose of this study, only items relevant to the study objectives were analyzed. Surveys that did not specify the country of origin and those that were less than 90% complete were excluded. Furthermore, as there were minimal missing data, a modified complete case analysis was performed [31]. Descriptive statistics were used to present the data using the

XSL•FO
RenderX

SPSS Statistics software (version 29.0.0.0; IBM Corp). Sample size estimation was not conducted for this descriptive study because the aim was to summarize and describe the facilitators and barriers rather than hypothesis testing or predictions about associations between variables.

Qualitative Phase

Thematic analysis was conducted using both deductive and inductive approaches, including data transcription, coding, categorization, analysis, and interpretation [30]. Data transcription was carried out through repeated examination of the recorded interviews by 2 researchers (PKT and CI). The research team engaged in regular, in-depth discussions to resolve uncertainties in coding and interpretation, reaching consensus to uphold the study's credibility.

During the categorization phase, codes with similar conceptual content were grouped to identify emerging themes. Each stage was reviewed by 3 members of the research team (PKT, EE, and CI) and discussed. This approach facilitated shared interpretation, resolution of discrepancies, and consensus building around codes and themes. It also ensured trustworthiness and analytic accuracy, minimized individual bias and misinterpretation, and prevented the exclusion of any relevant data, thus strengthening the confirmability and credibility of the findings [32,33].

Data analysis was conducted manually, supported by the use of the MAXQDA 2022 software (VERBI GmbH) [34].

The research team represented diverse sociocultural backgrounds, aligning with the principles of equity, diversity, and inclusion and contributing to a more nuanced and inclusive interpretation of the data. They included senior and mid- and early-career researchers with several years of experience in qualitative research. Throughout the research process, the researchers engaged in reflexivity and acknowledged how their background, experience, and personal beliefs could shape the interpretation of the data.

Results

Characteristics of the Study Participants

The online survey was received by all 144 invited CDOs or responsible officers, of whom 101 (70.1%) returned the questionnaires. Approximately 17.8% (18/101) of the filled-out questionnaires had incomplete responses and were excluded from the analysis, leaving 83 questionnaires for the final analysis.

The mean age of the participants was 50.7 (SD 9.3) years. The majority of the participants were female (48/80, 60%). Approximately 69.9% (n=58) spoke English, 19.3% (n=16) spoke French, 9.6% (n=8) spoke Spanish, and 1.2% (n=1) spoke Russian. In total, 93% (n=77) worked full-time as CDOs or WHO associate members.

Of the 83 participants, 38.6% (n=32) were representatives of the African Region, 8.4% (n=7) were representatives of the Eastern Mediterranean Region, 13.3% (n=11) were representatives of the Region of the Americas, 16.9% (n=14) were representatives of the European Region, 8.4% (n=7) were representatives of the South-East Asia Region, and 14.5% (n=12) were representatives of the Western Pacific Region. The participants were from high-income (n=25, 30.1%), upper-middle–income (n=23, 27.7%), lower-middle–income (n=23, 27.7%), and low-income countries (n=12, 14.5%). More than three-quarters of the respondents (80/83, 84.3%) used a computer to complete the survey. Smartphones were mostly used by respondents from low- or lower-middle–income countries, whereas those from high-income countries (HICs) predominantly used computers.

For the qualitative study, interviews were conducted with 15 participants (n=14, 93.3% female) aged between 34 and 69 years, representing countries of all income levels: 2 (13.3%) from HICs, 3 (20%) from upper-middle–income countries, 6 (40%) from lower-middle–income countries, and 4 (26.7%) from low-income countries. Data saturation was reached after the 10th interview; however, we continued data collection to achieve diversity and at least one representative from all WHO regions.

Study Findings

Overview

Both the quantitative and qualitative findings were integrated under 2 broad themes of facilitators and barriers. Perceived facilitators and barriers represented well the macro, meso, and micro domains of e-readiness. The main facilitators and barriers are illustrated in Figure 1 and Figure 2, respectively. Four subthemes (2 per theme) emerged from the analysis.

XSL•FO
RenderX

Figure 1. Facilitators of the implementation of digital oral health programs. WHO: World Health Organization.

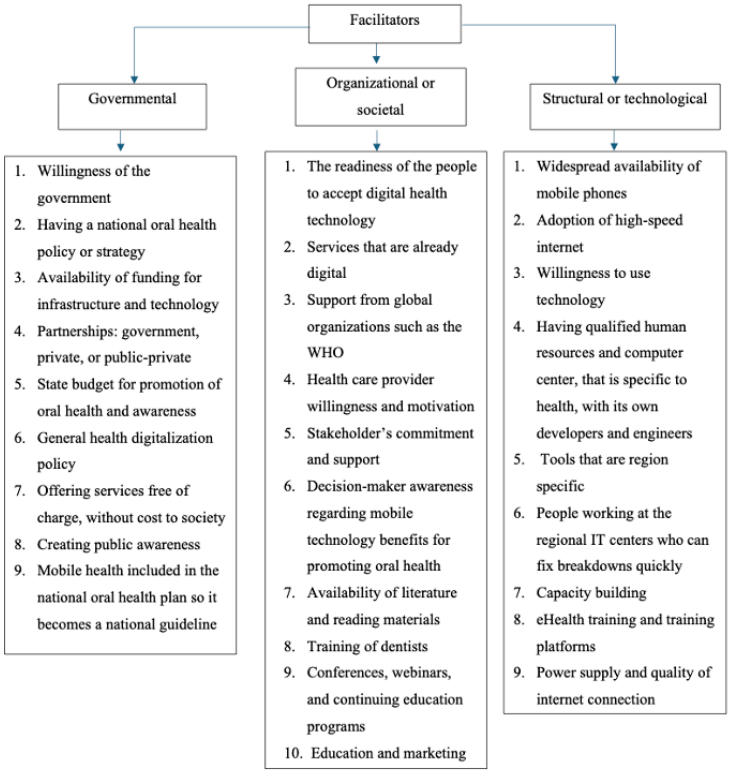

Figure 2. Barriers to the implementation of digital oral health programs. mHealth: mobile health; WHO: World Health Organization.

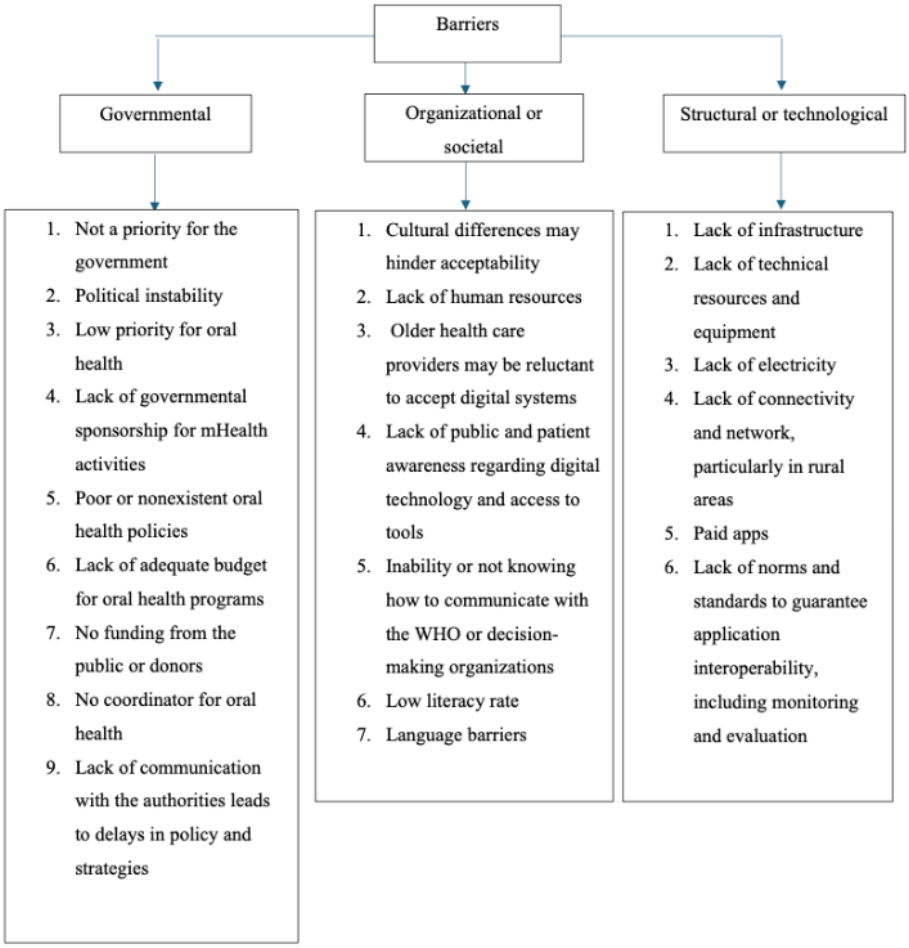

XSL•FO
RenderX

Facilitators

Building Digital Health Foundations Through Policy Frameworks and Motivation

Perceived enablers included the presence of national policies and guidelines on eHealth, social acceptance, a desire for digital health, access to necessary digital tools, capacity building on existing e-platforms, and the motivation of health care providers and academic leadership. The theme highlighted the perceived usefulness and positive attitude toward using health technology. Governmental readiness was reflected in the presence of national digital health policies:

> *We have a very high potential of starting a mOralHealth program because don't forget that we have something like 143 million active subscribers to mobile phones and there are 173 million people with mobile phones on them. So, it's almost 90% of the population having access to mobile phones.* [CDO 10]

According to the survey, 63.9% (n=53) of the participants stated that their country had a national oral health policy, of whom 32.1% (17/53) reported that this policy or strategy clearly referred to digital health to support oral health. Additionally, 60.3% (32/53) of those who indicated that their country had a national oral health policy also had a national digital health or eHealth policy. Digital oral health policies were more common in the WHO European Region. In countries that had an eHealth policy or strategy, 78.1% (25/32) of the participants reported that their countries had adopted a national health information system policy or strategy:

> *I think the government is very much into digitalization and using artificial intelligence. As you know, digital health is one of the priorities.* [CDO 9]

With regard to the global implementation of the mOral Health program, the WHO was perceived as a powerful organization that could influence behavioral intentions at the macro level. Moreover, there was a general expectation of support, assistance, and methodological input as 80% (64/80) of the participants reported the need for WHO support to implement e–oral health programs:

> *The WHO just needs to make very strong recommendations to the government because the government is very, very receptive towards the WHO recommendation. [CDO 8]*

The key drivers behind the mOral Health program were working with those who were "the most motivated" and global competition:

> *It is necessary to work with the most motivated countries, the most motivated providers, the most motivated patient groups. Positive competition must be established.* [CDO 2]

Capacity Building Through Training and Technology

The availability of e-training and e-education, as well as the funding allocated for digital health initiatives, was considered advantageous for the successful implementation and use of digital oral health. The participants expressed their views on the importance of programs and resources to develop skills among health care providers in using digital technologies:

> *We have a platform for health workers. It's mostly for their training, like continuing education programs. If we can put it online, if we can integrate it in those programs, that will be one way of implementing it.* [CDO 7]

> *In the new generation, there's almost no problem in adopting mobile health. But for the older generation, we need some training, and we have to socialize the program before it is implemented.* [CDO 13]

Approximately 46.9% (38/81) of the participants reported that universities or technical colleges provided training for students or professionals on the use of information and communications technologies and eHealth. However, only 34.5% (28/81) reported that institutions (universities or technical colleges) offered training to dentistry students on the use of digital health. More than two-thirds of the respondents (n=60, 72%) reported that they had e-learning courses accredited by continuing medical education or professional licensing bodies.

Technological readiness and health care provider readiness were perceived as strong enablers:

> *As long as the application is not complicated and you can use very low data, you don't need high-speed internet. I think we are ready to receive this type of program.* [CDO 13]

> *We are able to give information on what we need to do in oral health, in e-health. Whatever the platform that has been set. So, if you have information that is packaged already, then we have a platform.* [CDO 2]

Barriers

Limited Prioritization and Resources

Neglected e-oral health priorities and programs; failure to recognize the value of eHealth in health care delivery; and lack of infrastructure, qualified human resources, and funding were highlighted as barriers and societal challenges.

Among the countries included in our study, 45.1% (37/82) had government-sponsored mHealth programs, and only a small percentage, 25.9% (21/81), had a national telehealth policy. In total, 28.9% (24/83) of the participants reported that health authorities provided incentives and guidance for innovation, research, and evaluation regarding eHealth:

> *We actually do not have an oral health policy in place. So, talking about mOralHealth program would be slightly premature because there is no oral health policy or strategy that we are following.* [CDO 10]

The scarcity of financial support was one of the main concerns regarding digital oral health implementation. The survey revealed that 31.7% (26/82) of the participants reported having no financial support (public, private, both, or from donors) available for the implementation of their national digital oral health programs. Among those who did have funding, 36.6% (30/82) received it from the public sector, 19.5% (16/82)

XSL•FO
RenderX

received it from the private sector, 18.3% (15/82) received it from donors, and 18.3% (15/82) received it from public-private partnerships. Only 10.9% (9/82) of the participants reported all 4 types of funding. Furthermore, 39% (32/82) of the participants reported that their national digital health policy funding did not support oral health.

Public funding was the most commonly reported funding type among participants from countries in the European Region (9/30, 30%). Approximately 68.7% (11/16) of the higher-income countries had adequate financial resources through private organizations or public-private partnerships (8/15, 53.3%):

> *As far as the funding is concerned, again, let me remind you that we spend less than 3% or 2% of our capital, our GDP, on health. And since oral health is very low on priority, there is practically nothing left for oral health when it comes to a national policy. So funding, again, is an issue as far as the country's readiness.* [CDO 10]

Inadequate technological access, infrastructure, and integration with existing services; the limited support for implementing mHealth initiatives; and the lack of norms and standards to guarantee application interoperability were perceived as important barriers. At the meso level, issues with structural and technological readiness, technical interoperability between existing systems, and a lack of equipment or connectivity were also identified as barriers:

> *There are definitely issues related to specifically Wi-Fi and internet connectivity where everyone seems to have some kind of cellphone or tablet kind of mechanism for social media access. When it comes to actual connectivity data, etc., it costs money for them to download information, download material.* [CDO 11]

> *Sometimes in big cities, we can implement this type of program without any problem. But when we go to the rural area, to the remote island, they cannot do this because there is no signal or enough internet to do that.* [CDO 13]

> *A lot of healthcare workers are keen and willing to be part of a program. But I think again. If there are problems, say, with the ongoing lack of connectivity, when they need to get in touch with a consultant and if they cannot at that moment or if they do not have efficient and ready ICT support...the healthcare providers lose interest if there isn't a good two-way feedback when we implement these programs.* [CDO 11]

Hesitation and Uncertainty

The lack of demand for eHealth programs among health professionals was a significant barrier for upper-middle–, low-middle–, and low-income countries. Moreover, only approximately one-quarter of all participants (n=18, 21.6%) reported that health authorities provided regulatory oversight for mHealth apps in terms of quality, safety, and reliability. Additionally, only one-third (n=24, 29%) of all participants stated that health authorities offered incentives and guidance for innovation, research, and evaluation regarding eHealth.

Furthermore, concerns related to the lack of evidence on the effectiveness and cost-effectiveness of programs, as well as a lack of confidence in digital and data management (eg, cyber risks), were expressed as barriers by 73.8% (59/80), 73% (57/78), and 69.6% (55/79) of the participants, respectively:

> *My comment would be around the availability of reading materials that could help to get people prepared. Is there literature people can access online to familiarize themselves with what this is and what it involves, so that we can share them to disseminate and start getting that feeling with people.* [CDO 7]

Discussion

Principal Findings

This study identified key factors influencing the adoption of digital oral health in countries belonging to the 6 WHO regions from the decision-makers' perspective. The key enablers identified were the presence of national eHealth policies, availability of funding for infrastructure and technology, and health care provider willingness. On the other hand, the barriers identified were the lack of governmental support for digital health, lack of partnerships, neglected oral health programs, lack of infrastructure and funding, and inadequate human resources. This study revealed that the frequency of these barriers was higher in low- and middle-income countries (LMICs), where policy gaps, competing health care priorities, and limited resources often exacerbate these issues compared to HICs.

Comparison With Prior Work

Our findings are in agreement with those of previous systematic reviews [35,36] that have pointed to insufficient infrastructure, outdated equipment, technology gaps, and human resource challenges as major barriers to mHealth implementation in LMICs. Additionally, health care executives in Ethiopia have emphasized further challenges, including economic inflation, low digital literacy, and a lack of system interoperability as significant challenges to the adoption of mHealth solutions [37]. These barriers collectively pose significant threats to the performance of the health workforce, undermining their efforts to deliver high-quality care and further perpetuating health disparities.

Underuse and slow progression of digital health initiatives and mHealth programs despite the growing adoption of mobile phones and recognition of their potential benefits was a finding in LMICs. One key factor contributing to this issue could be the low mobile internet penetration in these regions, which stands at 69% and only 48% when China is excluded despite an estimated 2.9 billion current users of mobile phones [38,39]. This is in sharp contrast to 87% mobile internet penetration in HICs [39]. The lack of mobile internet access can prevent the effective use of mHealth tools that depend on real-time communication, data transmission, and access to digital health platforms. Consequently, the government's role is crucial to improve the telecommunications infrastructure and enhance

XSL•FO
RenderX

mobile internet access by making it more affordable through low costs of mobile devices and data services [40].

Digital technology offers an opportunity to enhance education, incentives, aid, supervision, and support for universal health coverage [41]. However, these efforts require adequate funding, the lack of which was reported as a significant barrier in our study and has also been reported as a barrier in other research [35,36], affecting both HICs and LMICs. The successful implementation of digital health interventions requires investment not only in technology but also in operational costs; training for dental health care providers; and the development of health care workflows, including data management and security systems. Securing sustained support from the government and relevant local funders and institutions is crucial for the success of digital interventions [42]. Moreover, to facilitate both scaling up and financial sustainability, diverse funding sources, such as private sector investments and public-private partnerships, should be sought rather than relying on a single source [40,43].

In addition to funding, proactive government engagement and the development of clear policies are essential, with well-defined roles and responsibilities at all levels, from government bodies to health care managers and dental health care providers [35]. Engaging local stakeholders and collaborating with mobile technology developers early in the design phase can help ensure that the interventions are both suitable and embraced [35]. Furthermore, strong partnerships between the national and local governments, nongovernmental organizations, private entities, and global organizations such as the WHO must be established for active leadership toward the successful implementation and long-term sustainability of the programs [40]. Within routine dental practices, ensuring successful implementation of digital health would require reforms in dental education toward the development of relevant competencies, positive attitudes, and behavioral intentions, ensuring easy access to digital technologies and establishing supportive policies and regulatory frameworks [44,45].

Strengths and Limitations

To our knowledge, this is the first study conducted in partnership with the WHO on global e–oral health readiness that involved participants from many countries working in different settings and with various profiles. The results of this study provide a basis for comparison and future planning. The mixed methods design used in this study enabled the combination of 2 complementary datasets to generate meaningful information and gain a strong understanding of factors that can be considered in developing a global strategic plan and improving the success of mOral Health implementation. However, this study has some limitations. First, while the participation rate was acceptable, the survey was self-administered, which may have resulted in an over- or underestimation of information. Second, many surveys were incomplete, resulting in the impossibility of analyzing nonrespondents. However, the purposeful sampling technique for the qualitative study provided rich data due to the inclusion of diverse participants from the 6 WHO regions. Third, the participants in this study represented WHO member countries from the 6 WHO regions. Therefore, the results may not be generalizable to nonmember countries, whose national health and digital health policies might differ from those of the countries included in our study.

Conclusions

This study identified key barriers to and enablers of implementing digital oral health programs in WHO member states through an e-readiness assessment across multiple levels. The analysis revealed gaps in funding support, the establishment of norms and regulations, and the availability of robust infrastructure. As readiness is essential for the successful adoption of new policies, practices, and technologies, its systematic assessment is crucial for integrating digital oral health into oral health care systems and practice to advance global oral health outcomes. Governments play a central role in this process by developing concrete policies, investing in infrastructure, and allocating adequate budgets for e–oral health technologies to overcome the limitations of traditional health care delivery models.

Acknowledgments

The authors are grateful to Benoit Varenne, Yuriko Harada, and Nicole Rendell from the World Health Organization Oral Health Programme, Department of Noncommunicable Diseases, Geneva, Switzerland, for their support in conducting this study. They also thank Anisha Rodrigues, research assistant, for her assistance in preparing the manuscript.

Funding

The project was funded by the Borrow Foundation

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

EE and NG contributed to the development of the research idea and partnership with key stakeholders. EE, NG, and PKT developed the protocol and study design. EE, NG, PKT, and CI contributed to data acquisition, analysis, and interpretation and drafting and revision of the manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
GRAMMS checklist.
[DOCX File , 15 KB - jmir_v28i1e76236_app1.docx]

Multimedia Appendix 2
Questionnaire related to facilitators and barriers.
[PDF File (Adobe PDF File), 79 KB - jmir_v28i1e76236_app2.pdf]

Multimedia Appendix 3
Interview guide.
[DOCX File , 133 KB - jmir_v28i1e76236_app3.docx]

References

1. Meskó B, Drobni Z, Bényei É, Gergely B, Győrffy Z. Digital health is a cultural transformation of traditional healthcare. Mhealth 2017 Sep;3:38 [FREE Full text] [doi: 10.21037/mhealth.2017.08.07] [Medline: 29184890]
2. Nundy S, Cooper LA, Mate KS. The quintuple aim for health care improvement: a new imperative to advance health equity. JAMA 2022 Feb 08;327(6):521-522. [doi: 10.1001/jama.2021.25181] [Medline: 35061006]
3. Coleman K, Wagner E, Schaefer J, Reid R, LeRoy L. Redefining primary care for the 21st century. Agency for Healthcare Research and Quality. 2016. URL: https://www.ahrq.gov/sites/default/files/wysiwyg/professionals/systems/primary-care/workforce-financing/white_paper.pdf [accessed 2026-04-02]
4. Peres MA, Macpherson LM, Weyant RJ, Daly B, Venturelli R, Mathur MR, et al. Oral diseases: a global public health challenge. Lancet 2019 Jul 20;394(10194):249-260. [doi: 10.1016/S0140-6736(19)31146-8] [Medline: 31327369]
5. Wang J, Wang B, Liu YY, Luo YL, Wu YY, Xiang L, et al. Recent advances in digital technology in implant dentistry. J Dent Res 2024 Jul;103(8):787-799. [doi: 10.1177/00220345241253794] [Medline: 38822563]
6. Lee SJ, Poon J, Jindarojanakul A, Huang CC, Viera O, Cheong CW, et al. Artificial intelligence in dentistry: exploring emerging applications and future prospects. J Dent 2025 Apr;155:105648. [doi: 10.1016/j.jdent.2025.105648] [Medline: 39993553]
7. Mahdavi A, Atlasi R, Naemi R. Teledentistry during COVID-19 pandemic: scientometric and content analysis approach. BMC Health Serv Res 2022 Sep 01;22(1):1111 [FREE Full text] [doi: 10.1186/s12913-022-08488-z] [Medline: 36050678]
8. Gurgel-Juarez N, Torres-Pereira C, Haddad AE, Sheehy L, Finestone H, Mallet K, et al. Accuracy and effectiveness of teledentistry: a systematic review of systematic reviews. Evid Based Dent 2022 Jul 08:1-8 [FREE Full text] [doi: 10.1038/s41432-022-0257-8] [Medline: 35804195]
9. Kengne Talla P, Allison P, Bussières A, Rodrigues A, Bergeron F, Giraudeau N, et al. Teledentistry for improving access to, and quality of oral health care: overview of systematic reviews and meta-analyses. J Med Internet Res 2025 Jul 30;27:e65211 [FREE Full text] [doi: 10.2196/65211] [Medline: 40737673]
10. Cresswell K, Sheikh A. Organizational issues in the implementation and adoption of health information technology innovations: an interpretative review. Int J Med Inform 2013 May;82(5):e73-e86. [doi: 10.1016/j.ijmedinf.2012.10.007] [Medline: 23146626]
11. Desveaux L, Soobiah C, Bhatia RS, Shaw J. Identifying and overcoming policy-level barriers to the implementation of digital health innovation: qualitative study. J Med Internet Res 2019 Dec 20;21(12):e14994 [FREE Full text] [doi: 10.2196/14994] [Medline: 31859679]
12. Draft global oral health action plan (2023–2030). World Health Organization. 2023. URL: https://cdn.who.int/media/docs/default-source/ncds/mnd/oral-health/eb152-draft-global-oral-health-action-plan-2023-2030-en.pdf?sfvrsn=2f348123_19&download=true [accessed 2026-04-02]
13. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2026-04-02]
14. Jagde AK, Shrivastava R, Feine J, Emami E. Patients' e-readiness to use e-health technologies for oral health. PLoS One 2021 Jul 12;16(7):e0253922 [FREE Full text] [doi: 10.1371/journal.pone.0253922] [Medline: 34252096]
15. Nilsen ER, Stendal K, Gullslett MK. Implementation of eHealth technology in community health care: the complexity of stakeholder involvement. BMC Health Serv Res 2020 May 11;20(1):395 [FREE Full text] [doi: 10.1186/s12913-020-05287-2] [Medline: 32393265]
16. Tossaint-Schoenmakers R, Versluis A, Chavannes N, Talboom-Kamp E, Kasteleyn M. The challenge of integrating eHealth into health care: systematic literature review of the donabedian model of structure, process, and outcome. J Med Internet Res 2021 May 10;23(5):e27180 [FREE Full text] [doi: 10.2196/27180] [Medline: 33970123]

XSL•FO
RenderX

17. Khoja S, Scott RE, Casebeer AL, Mohsin M, Ishaq AF, Gilani S. e-Health readiness assessment tools for healthcare institutions in developing countries. Telemed J E Health 2007 Aug;13(4):425-431. [doi: 10.1089/tmj.2006.0064] [Medline: 17848110]
18. Mauco KL, Scott RE, Mars M. Validation of an e-health readiness assessment framework for developing countries. BMC Health Serv Res 2020 Jun 23;20(1):575 [FREE Full text] [doi: 10.1186/s12913-020-05448-3] [Medline: 32576174]
19. Yusif S, Hafeez-Baig A, Soar J. An exploratory study of the readiness of public healthcare facilities in developing countries to adopt health information technology (HIT)/e-Health: the case of Ghana. J Healthc Inform Res 2020 Jan 31;4(2):189-214 [FREE Full text] [doi: 10.1007/s41666-020-00070-8] [Medline: 35415442]
20. Jivraj A, Barrow J, Listl S. Value-based oral health care: implementation lessons from four case studies. J Evid Based Dent Pract 2022 Jan;22(1S):101662 [FREE Full text] [doi: 10.1016/j.jebdp.2021.101662] [Medline: 35063180]
21. O'Cathain A, Murphy E, Nicholl J. The quality of mixed methods studies in health services research. J Health Serv Res Policy 2008 Apr 21;13(2):92-98. [doi: 10.1258/jhsrp.2007.007074] [Medline: 18416914]
22. Ivankova NV, Creswell JW, Stick SL. Using mixed-methods sequential explanatory design: from theory to practice. Field Methods 2006 Feb 01;18(1):3-20. [doi: 10.1177/1525822X05282260]
23. Palinkas LA, Aarons GA, Horwitz S, Chamberlain P, Hurlburt M, Landsverk J. Mixed method designs in implementation research. Adm Policy Ment Health 2011 Jan 22;38(1):44-53 [FREE Full text] [doi: 10.1007/s10488-010-0314-z] [Medline: 20967495]
24. Creswell JW. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Thousand Oaks, CA: SAGE Publications; 2003.
25. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: four longitudinal field studies. Manag Sci 2000 Feb;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
26. Indicator metadata registry list. World Health Organization. URL: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/1144 [accessed 2026-04-02]
27. Atlas of eHealth country profiles: the use of eHealth in support of universal health coverage: based on the findings of the third global survey on eHealth 2015. World Health Organization. 2016. URL: https://www.who.int/publications/i/item/9789241565219 [accessed 2026-04-03]
28. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health 2015 Sep 6;42(5):533-544 [FREE Full text] [doi: 10.1007/s10488-013-0528-y] [Medline: 24193818]
29. Dillman DA. Why choice of survey mode makes a difference. Public Health Rep 2006;121(1):11-13 [FREE Full text] [doi: 10.1177/003335490612100106] [Medline: 16416693]
30. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907 [FREE Full text] [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
31. Ross RK, Breskin A, Westreich D. When is a complete-case approach to missing data valid? The importance of effect-measure modification. Am J Epidemiol 2020 Dec 01;189(12):1583-1589 [FREE Full text] [doi: 10.1093/aje/kwaa124] [Medline: 32601706]
32. Creswell JW, Creswell JD. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Thousand Oaks, CA: SAGE Publications; 2017.
33. Korstjens I, Moser A. Series: practical guidance to qualitative research. Part 4: trustworthiness and publishing. Eur J Gen Pract 2018 Dec;24(1):120-124 [FREE Full text] [doi: 10.1080/13814788.2017.1375092] [Medline: 29202616]
34. VERBI software. MAXQDA. URL: https://www.maxqda.com/about [accessed 2026-04-03]
35. Aranda-Jan CB, Mohutsiwa-Dibe N, Loukanova S. Systematic review on what works, what does not work and why of implementation of mobile health (mHealth) projects in Africa. BMC Public Health 2014 Feb 21;14(1):188 [FREE Full text] [doi: 10.1186/1471-2458-14-188] [Medline: 24555733]
36. Kruse C, Betancourt J, Ortiz S, Valdes Luna SM, Bamrah IK, Segovia N. Barriers to the use of mobile health in improving health outcomes in developing countries: systematic review. J Med Internet Res 2019 Oct 09;21(10):e13263. [doi: 10.2196/13263] [Medline: 31593543]
37. Aboye GT, Simegn GL, Aerts J. Assessment of the barriers and enablers of the use of mHealth systems in sub-Saharan Africa according to the perceptions of patients, physicians, and health care executives in Ethiopia: qualitative study. J Med Internet Res 2024 Mar 27;26:e50337 [FREE Full text] [doi: 10.2196/50337] [Medline: 38536231]
38. Bahia K. The state of mobile internet connectivity report 2020. GSMA Intelligence. 2020. URL: https://www.gsmaintelligence.com/research/the-state-of-mobile-internet-connectivity-report-2020 [accessed 2026-04-03]
39. Bahia K. The state of mobile internet connectivity 2023. GSMA Intelligence. 2023. URL: https://www.gsmaintelligence.com/research/the-state-of-mobile-internet-connectivity-2023 [accessed 2026-04-03]
40. Barkman C, Weinehall L. Policymakers and mHealth: roles and expectations, with observations from Ethiopia, Ghana and Sweden. Glob Health Action 2017 Jun;10(sup3):1337356. [doi: 10.1080/16549716.2017.1337356] [Medline: 28838303]

XSL•FO
RenderX

41. Long LA, Pariyo G, Kallander K. Digital technologies for health workforce development in low- and middle-income countries: a scoping review. Glob Health Sci Pract 2018 Oct 10;6(Suppl 1):S41-S48 [FREE Full text] [doi: 10.9745/GHSP-D-18-00167] [Medline: 30305338]
42. McCool J, Dobson R, Muinga N, Paton C, Pagliari C, Agawal S, et al. Factors influencing the sustainability of digital health interventions in low-resource settings: lessons from five countries. J Glob Health 2020 Dec;10(2):020396 [FREE Full text] [doi: 10.7189/jogh.10.020396] [Medline: 33274059]
43. The MAPS toolkit. World Health Organization. 2015. URL: https://www.who.int/publications/i/item/9789241509510 [accessed 2026-04-03]
44. Schnitzler C, Bohnet-Joschko S. Technology readiness drives digital adoption in dentistry: insights from a cross-sectional study. Healthcare (Basel) 2025 May 15;13(10):1155 [FREE Full text] [doi: 10.3390/healthcare13101155] [Medline: 40427991]
45. Galazzi A, Fonda F, Chiappinotto S, Justi L, Sønderskov Frydensberg M, Lehmann Boesen R, et al. Recommendations to promote the digital healthcare transformation in the clinical practice: findings from an international consensus development method. BMC Health Serv Res 2025 Jul 04;25(1):929 [FREE Full text] [doi: 10.1186/s12913-025-13079-9] [Medline: 40616118]

Abbreviations

AI: artificial intelligence
CDO: chief dental officer
GRAMMS: Good Reporting of a Mixed Methods Study
HIC: high-income country
LMIC: low- and middle-income country
mHealth: mobile health
WHO: World Health Organization

Edited by T Leung, N Cahill; submitted 21.Apr.2025; peer-reviewed by P Milgrom, GRM La Rosa; comments to author 22.Jul.2025; revised version received 10.Nov.2025; accepted 30.Nov.2025; published 30.Apr.2026.

Please cite as:
Emami E, Kengne Talla P, Inquimbert C, Giraudeau N
Facilitators of and Barriers to Global Digital Oral Health: Mixed Methods Study
J Med Internet Res 2026;28:e76236
URL: https://www.jmir.org/2026/1/e76236
doi:10.2196/76236
PMID:42060912

©Elham Emami, Pascaline Kengne Talla, Camille Inquimbert, Nicolas Giraudeau. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

User Profiles and Engagement in a Hypertension Self-Management App: Cross-Sectional Survey

Felix Muehlensiepen[1,2,3,4], MPH, PhD, Dr rer medic habil; Susann May[1,2,3], MPH, Dr rer medic; Frances Seifert[1,2,3], MSc; Eileen Wengemuth[1], Dr rer nat; Olen Johannsen[5], MSc; Martin Middeke[6], Prof, Dr; Martin Heinze[1,7], Prof, Dr; Pascal Petit[4], PhD; Nicolas Vuillerme[4], Prof, Dr; Sebastian Spethmann[2,3*], Prof, Dr; Dunja Bruch[8*], MSc

[1]Center for Health Services Research, Faculty of Health Sciences Brandenburg, Brandenburg Medical School, Rüdersdorf, Seebad 82/83, Ruedersdorf bei Berlin, Germany

[2]Department of Cardiology, Angiology and Intensive Care Medicine, Deutsches Herzzentrum der Charité, Berlin, Germany

[3]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[4]Université Grenoble Alpes, CNRS, Grenoble INP, LIG Sangria, Grenoble, France

[5]Hypertension Care UG, Ingolstadt, Germany

[6]Hypertension Center Munich, Munich, Germany

[7]Center for Mental Health, Brandenburg Medical School, Immanuel Klinik Rüdersdorf, Ruedersdorf bei Berlin, Germany

[8]Department of Cardiovascular Surgery, Heart Center Brandenburg, Faculty of Health Sciences, Brandenburg Medical School Theodor Fontane, Bernau, Germany

[*]these authors contributed equally

Corresponding Author:
Felix Muehlensiepen, MPH, PhD, Dr rer medic habil
Center for Health Services Research, Faculty of Health Sciences Brandenburg, Brandenburg Medical School, Rüdersdorf, Seebad 82/83, Ruedersdorf bei Berlin, Germany

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2026/1/e93515

Abstract

Background: Mobile health (mHealth) technologies can improve hypertension self-management, yet real-world adoption remains limited and unequally distributed.

Objective: This study aimed to characterize the profiles, usage patterns, and engagement of active users of a hypertension self-management app (*Hypertension.APP*) in Germany, with a focus on user engagement and potential digital divides.

Methods: We conducted a cross-sectional online survey among adult users of *Hypertension.APP* in Germany between January and September 2023. An 88-item questionnaire assessed app usage patterns, perceived utility, integration into clinical care, sociodemographic and clinical data, and digital health literacy (eHealth Literacy Scale; scores 16 - 40). Digital health literacy was categorized as low (16 - 23.99), moderate (24 - 31.99), or high (32-40). Descriptive statistics and univariable ordinal logistic regression were used to explore associations between sociodemographic and clinical variables and app usage frequency.

Results: Of 254 respondents, the mean age was 53.6 years, and 54.3% (138/254) were male. A total of 44.5% (113/254) had a university or technical college degree, and 44.5% (113/254) reported a monthly net income higher than €2500 (US $2950). Most participants (224/254, 88.2%) reported access to at least two digital devices. Overall, 88.2% (224/254) had moderate or high digital health literacy (eHealth Literacy Scale ≥24). App engagement was high: 80.7% (205/254) reported using the app at least weekly, and 52.4% (133/254) reported using the app to prepare for medical visits. However, only 20.1% (51/254) reported that the app was formally integrated into their medical care, and 11.8% (30/254) indicated that medication had been adjusted based on app data. In univariable ordinal logistic regression analyses, higher education, longer duration of hypertension, and living in a small town (5000 - 20,000 inhabitants) were associated with more frequent app use, whereas systolic blood pressure of 140 mm Hg or higher was associated with less frequent use. Digital health literacy was not clearly associated with app usage frequency among current users.

Conclusions: Users of this hypertension self-management app were predominantly well-educated, digitally literate individuals with established hypertension, reinforcing concerns about a persistent digital divide. While app usability and engagement were high, formal clinical integration remained limited. Simply making an app available is insufficient; strategies to promote equitable

XSL•FO
RenderX

access, strengthen clinical integration, and support patients with lower digital health literacy are needed for mHealth to contribute effectively to hypertension management.

Trial Registration: German Clinical Trials Register DRKS00029761; https://drks.de/search/en/trial/DRKS00029761/details

(J Med Internet Res 2026;28:e83075) doi:10.2196/83075

KEYWORDS

digital health; mHealth; digital divide; hypertension; cardiology; chronic disease; questionnaire; health services research; user perspectives; equity; digital public health; mobile health

Introduction

Cardiovascular diseases remain the leading cause of death worldwide, despite the availability of evidence-based strategies for prevention and treatment [1]. Arterial hypertension—affecting more than 1 billion people globally—is a major modifiable risk factor [2]. In addition to genetic predisposition, the development and progression of high blood pressure are strongly influenced by individual lifestyle factors, health literacy, and behaviors [3], as well as by structural determinants such as access to and quality of medical care [4].

In recent years, digital health technologies have gained increasing relevance in the prevention and management of chronic conditions, including hypertension [5]. Mobile health (mHealth) applications, wearable devices, and other digital tools offer new opportunities for individuals to monitor their health, adopt healthier behaviors, and take a more active role in managing their condition. These tools also provide health professionals with novel ways to support daily care routines and improve treatment adherence [6].

Multiple studies have shown that mHealth interventions can contribute to better blood pressure control [7]. Most recently, Prendergast et al [8] demonstrated in a landmark single-center randomized clinical trial that an emergency department–based education and mHealth empowerment intervention for patients with hypertension led to significantly greater reductions in systolic blood pressure (SBP) at 6 months compared with standard discharge care.

However, findings from controlled trials may not translate to routine practice. Evidence from other areas of digital health suggests that adherence to mHealth interventions in everyday practice tends to be low [9-11]. Common barriers include limited awareness of mHealth benefits among both patients and health care professionals [12-14], insufficient reimbursement structures [15], and difficulties integrating digital tools into existing care pathways [16]. While these challenges have been widely discussed from the perspective of health systems and providers, less is known about the individuals who actively use such tools. Gaining insight into who uses hypertension apps, how they are used, and what users expect from them can support the development of more effective and user-centered digital health solutions.

In a preceding qualitative study [17], we explored patient users' perspectives on a hypertension self-management app (*Hypertension.APP*), focusing on perceived benefits, potential risks, and the app's role in everyday care. Patients valued features such as continuous monitoring and personalized feedback but reported limited integration into formal health care and expressed concerns about data privacy. Moreover, the findings suggested that users tended to be digitally literate, self-effective, and relatively well-educated. These qualitative insights provided important initial evidence but were based on a small sample (n=20). The present study aims to build on these findings using a quantitative approach to provide a broader and more generalizable understanding of app users in hypertension care.

The objective of this cross-sectional study was to investigate the profiles, usage patterns, and engagement of active users of a hypertension self-management app. Using a standardized survey, we examined user demographics, usage patterns, usability, and clinical integration of *Hypertension.APP*.

Methods

Study Design

This study was conducted as a cross-sectional study. It examines how users of *Hypertension.APP* evaluate their experiences and the app's perceived impact on their hypertension care. The data provide insights into user attitudes, app engagement, and the role of digital health technologies in hypertension management.

This study is part of the Digital Preventive Measures for Arterial Hypertension project [18], which examines the structural and individual factors of patients and stakeholders in the use of digital preventive measures among patients with arterial hypertension in Germany.

Setting

The study was conducted via online recruitment, targeting adult users of *Hypertension.APP* across Germany. Recruitment and data collection were carried out between January 1 and September 30, 2023. Users were approached within their real-world use context of the app. Potential participants were invited via 3 digital channels: (1) in-app notifications within *Hypertension.APP*, (2) email invitations sent to registered users, and (3) social media posts on channels maintained by the app provider and project partners. All invitations briefly described the purpose of the study, stated the eligibility criteria, and provided a link to the online questionnaire.

Participants

Eligibility criteria included individuals aged 18 years or older who had used the app for at least 1 month. Exclusion criteria included individuals without access to a smartphone or tablet and those unable to provide informed consent. After clicking on the survey link, interested individuals were first presented

XSL•FO
RenderX

with an online information sheet and an electronic informed consent form. Only participants who provided consent were forwarded to the actual questionnaire. Participation was voluntary and anonymous, and respondents could skip individual questions if they did not wish to answer them.

Sample Size

The planned sample size was set at N=200, based on an a priori power analysis (G*Power) for the chi-square test with the following assumptions: medium effect (w=0.2), α=.05, power (1-β)=.80, *df*=1 (required sample size n=197). Due to the high willingness of app users to participate, the sample size was exceeded.

mHealth App: *Hypertension.APP*

Based on expert recommendations (Prof Dr Martin Middeke and Prof Dr med Miklós Illyés), *Hypertension.APP* was selected as an example and examined through a standardized user survey. *Hypertension.APP* was developed and is maintained by Hypertension Care UG (Ingolstadt) [19]. The application had been downloaded in app stores nearly 11,000 times as of July 2025. It includes a diary function to document blood pressure data, a personalized guide that provides information based on the user's health profile, an export function to generate reports for medical consultation, a reminder function, and interactive exercises. The app is freely accessible in various app stores. There is also a purchasable premium version that includes additional features.

Two authors (OJ and MM) are the founders and/or chief executive officers of Hypertension Care UG, the company that developed *Hypertension.APP*, the app evaluated in this study. To minimize potential conflicts of interest, their role in the research process was restricted. The academic research team (FM, SM, DB, PP, NV, SS) independently developed the questionnaire, conducted data collection, managed the dataset, and performed all statistical analyses. The app developers did not have access to raw data and were not involved in data analysis or interpretation. All authors contributed to manuscript drafting, but final decisions on analysis and interpretation were made by the independent academic team.

Questionnaire

The development of the 88-item questionnaire followed a structured, multistep process. First, an initial item pool was generated based on preliminary findings from our preceding qualitative interview study on hypertension app use [17], existing literature on mHealth adoption, and theoretical frameworks related to self-management and digital health literacy. To ensure content validity, we integrated established measurement instruments, including the German version of the eHealth Literacy Scale (eHEALS) [20] and system usability scale (SUS) [21], and supplemented these with newly developed items capturing app-specific use cases, clinical integration, and user expectations. Second, the research team (FM, SM, EW, DB) iteratively refined item wording and domain structure to improve face validity, comprehensiveness, and consistency. This process included expert review by clinicians and digital health researchers with experience in hypertension management and questionnaire development (OJ, MM, SS). Third, the questionnaire was pilot-tested with a sample of 10 adult hypertension app users to evaluate clarity, item relevance, and response burden. Feedback from the pilot phase resulted in minor adjustments to wording, response options, and the sequence of questions to enhance comprehensibility and usability.

The final instrument (German version: Multimedia Appendix 1; English version: Multimedia Appendix 2) covered multiple domains: app usage patterns, integration into medical care, perceived utility of specific app features, usability (SUS), digital health literacy (eHEALS), clinical characteristics, health behaviors, and sociodemographic information.

Variables

Primary variables included user demographics, eHealth literacy, app engagement patterns, and perceptions of app utility. eHealth literacy was assessed using the German version of the eHEALS scale [20] and categorized into "Low" (16 - 23.99), "Moderate" (24 - 31.99), and "High" (32–40) levels. These thresholds follow the categorization approach proposed by Monkman et al [22], who examined conceptual distinctions within the scale and recommended these score ranges to differentiate levels of digital health literacy. App utility was measured using the German version of the SUS [21]. App engagement metrics, such as frequency and duration of use, were assessed alongside subjective evaluations of the app's role in hypertension care. Sociodemographic data, including age, gender, education, and place of residence, were also collected.

Data Sources and Measurement

All questionnaire data were self-reported, including blood pressure values. Participants provided their most recent blood pressure measurement, and no standardized measurement protocol was imposed. Participants were instructed to report values documented in *Hypertension.APP* or measured through their usual home monitoring routines. Accordingly, systolic and diastolic values reflect self-reported blood pressure at or near the time of survey completion rather than standardized study measurements.

Bias

Efforts to mitigate bias included the use of validated instruments, such as eHEALS scale [20] and SUS [22], and standardized survey procedures. Missing data were handled by including only observations with less than 20% missing values.

Statistical Methods

Data analysis was performed using R statistical software (version 2023.06.2; R Core Team, 2023). Descriptive statistics were used to summarize participant characteristics, app usage patterns, and SUS scores. Subgroup analyses were conducted for app usage frequency to explore variations by demographic factors (age, gender, place of residence, and education level). SUS scores were reported descriptively for the overall sample without subgroup analyses. To examine whether sociodemographic variables and digital health literacy (eHEALS) predicted app usage frequency, a univariable ordinal logistic regression analysis was performed using the MASS package [23].

XSL•FO
RenderX

Ethical Considerations

The study was approved by the Ethics Committee of Brandenburg Medical School Theodor Fontane (E-02‐20220620). The study was registered as part of the Digital Preventive Measures for Arterial Hypertension project at the German Clinical Trials Register (DRKS00029761). Primary data collection was conducted in compliance with the current data protection regulations of the General Data Protection Regulation and the Declaration of Helsinki. Participants provided informed consent before being forwarded the actual questionnaire. Data collection was anonymous, and no conclusions can be drawn about individuals in the presentation of results. Participants received €15.00 (US $17.53) as an incentive and compensation for participation. The anonymity of the survey data during incentivization was ensured by using a separate database and individual passwords, with the two datasets stored separately. Personal data used for incentivization were deleted. There was no formal affiliation or relationship between the hypertension app examined in this study and the overarching research project. The app was developed and operated independently, and the collection and analysis of usage data were conducted solely for scientific purposes within the scope of this study.

Results

Demographic Characteristics

A total of 254 individuals completed the survey. The mean age was 53.6 (SD 13.9) years, ranging from 18 to 83 years. The majority of participants identified as male (139/254, 54.7%) or female (109/254, 42.9%), with a small proportion identifying as nonbinary (2/254, 0.8%). Most respondents reported having completed secondary or higher education, with 113/254 (44.5%) holding a university or technical college degree. Regarding monthly net income, 113/254 (44.5%) reported earnings above €2500 (US $3000), while 36/254 (14.2%) earned less than €1500 (US $1800) per month. Income was not specified by 49/254 (19.3%) of participants.

Participants were geographically diverse. Over two-fifths (105/254, 41.3%) reported living in large cities (>100,000 inhabitants), while 38/254 (15%) resided in rural areas (<5000 inhabitants). Digital health literacy was measured using the eHEALS scale. Among participants, 183/254 (72%) demonstrated moderate or higher levels of digital health literacy. Specifically, 20/254 (8%) of users fell into the “Low” category, 163/254 (64%) were in the “Moderate” category, and 71/254 (28%) were classified as having “High” eHealth literacy. Overall, 216/254 (85%) of participants reported knowing how to find health-related information online, but only 90/254 (35.4%) felt confident making health decisions based on online information.

An overview of sociodemographic characteristics is provided in Table 1.

XSL•FO
RenderX

Table . Sociodemographic characteristics of adult users of *Hypertension.APP* participating in a cross-sectional online survey in Germany (N=254; January–September 2023). The table includes age, sex, education, income, place of residence, and digital health literacy.

Characteristic	Values, n (%)
Sex	
Male	139 (54.7)
Female	109 (42.9)
Nonbinary	2 (0.8)
Missing	4 (1.6)
Education level	
No degree	2 (0.8)
Secondary school certificate	27 (10.6)
Intermediate school diploma	42 (16.5)
High school diploma	56 (22)
University/technical college degree	113 (44.5)
Other degree	7 (2.8)
Missing	7 (2.8)
Income (monthly net income)	
<€1000 (US $1200)	16 (6.3)
€1000 - 1500 (US $1200-1800)	20 (7.9)
€1500 - 2500 (US $1800-3000)	51 (20.1)
>€2500 (US $3000)	113 (44.5)
Not specified	49 (19.3)
Missing	5 (2)
Size of place of residence	
Rural (<5000 inhabitants)	38 (15)
Small town (5000-20,000)	64 (25.2)
Medium-sized town (20,000-100,000)	42 (16.5)
Large city (>100,000)	105 (41.3)
Missing	5 (2)
Digital health literacy (eHEALS[a])	
Low (16 - 23.99)	26 (10.2)
Moderate (24 - 31.99)	98 (38.6)
High (32-40)	126 (49.6)
Missing	4 (1.6)
Availability of digital technologies	
Smartphone	244 (96.1)
Computer or laptop	223 (87.8)
Tablet	173 (68.1)
Landline phone	164 (64.6)
Smartwatch/fitness tracker	124 (48.8)

[a]eHEALS: eHealth literacy scale.

Clinical Characteristics

Table 2 summarizes the clinical profile of participants. The majority of respondents reported SBP values between 120 and 139 mm Hg (158/254, 62.2%), and 24.8% (63/254) reported values at or above 140 mm Hg. Diastolic blood pressure was most frequently reported between 80 and 99 mm Hg (185/254,

XSL•FO
RenderX

72.8%), with 24.4% (62/254) indicating levels of 100 mm Hg or higher. Most participants (192/254, 75.6%) were currently taking antihypertensive medication. Regarding hypertension history, 21.3% (54/254) reported a diagnosis within the past year, while 35% (89/254) had been diagnosed more than 5 years ago. Notably, 6.3% (16/254) indicated that they had never been diagnosed with high blood pressure or no longer had the condition. In terms of smoking behavior, 9.1% (23/254) of respondents were current smokers, 32.3% (82/254) were former smokers, and 57.9% (147/254) had never smoked regularly.

XSL•FO
RenderX

Table . Clinical characteristics of adult users of *Hypertension.APP* in Germany, collected as part of a cross-sectional online survey (January–September 2023). Variables include blood pressure categories, comorbidities, medication use, duration of hypertension, and smoking status.

Characteristic	Values, n (%)
Systolic blood pressure	
<120 mm Hg	31 (12.2)
120 - 129 mm Hg	67 (26.4)
130 - 139 mm Hg	91 (35.8)
140 - 159 mm Hg	54 (21.3)
160 - 179 mm Hg	9 (3.5)
>180 mm Hg	0 (0)
Missing	2 (0.8)
Diastolic blood pressure	
<80 mm Hg	3 (1.2)
80 - 84 mm Hg	68 (26.8)
85 - 89 mm Hg	57 (22.4)
90 - 99 mm Hg	60 (23.6)
100 - 109 mm Hg	56 (22)
>110 mm Hg	6 (2.4)
Missing	4 (1.6)
Antihypertensive medication use	
Yes	192 (75.6)
No	60 (23.6)
Missing	2 (0.8)
Time since hypertension diagnosis	
<1 y	54 (21.3)
1 - 5 y	93 (36.6)
5 - 10 y	29 (11.4)
>10 y	60 (23.6)
Never diagnosed or no longer hypertensive	16 (6.3)
Missing	2 (0.8)
Smoking status	
Current smoker	23 (9.1)
Former smoker	82 (32.3)
Never smoked regularly	147 (57.9)
Missing	2 (0.8)
Comorbidities	
Elevated cholesterol levels	69 (27.2)
Other chronic or serious medical conditions	48 (18.9)
Physical disability	28 (11)
History of heart attack or stroke	22 (8.66)
History of cardiac or vascular surgery	12 (4.72)
Diagnosed with diabetes mellitus	8 (3.15)
Diagnosed with chronic kidney disease	6 (2.36)
Means of transportation to go to the doctor	

XSL•FO
RenderX

Characteristic	Values, n (%)
Car	180 (70.9)
Public transportation	116 (45.7)
Bike	109 (42.9)
Other means of transport	32 (12.6)

The mean reported travel time to the doctor was 18.4 (SD 18.8) minutes, ranging from 0 to 159 minutes.

App Usage and User Experience

Figure 1 summarizes the results of the 10-item SUS for all 254 participants. *Hypertension.APP* achieved a mean SUS score of 82 (SD 11.98; range 38 - 100), indicating "excellent" usability according to established benchmarks.

Figure 1. System usability scale (SUS) ratings of *Hypertension.App* among adult users in Germany participating in a cross-sectional survey (January-September 2023). The figure displays item-level responses (left) and the distribution of total SUS scores (right). Values represent mean (range)

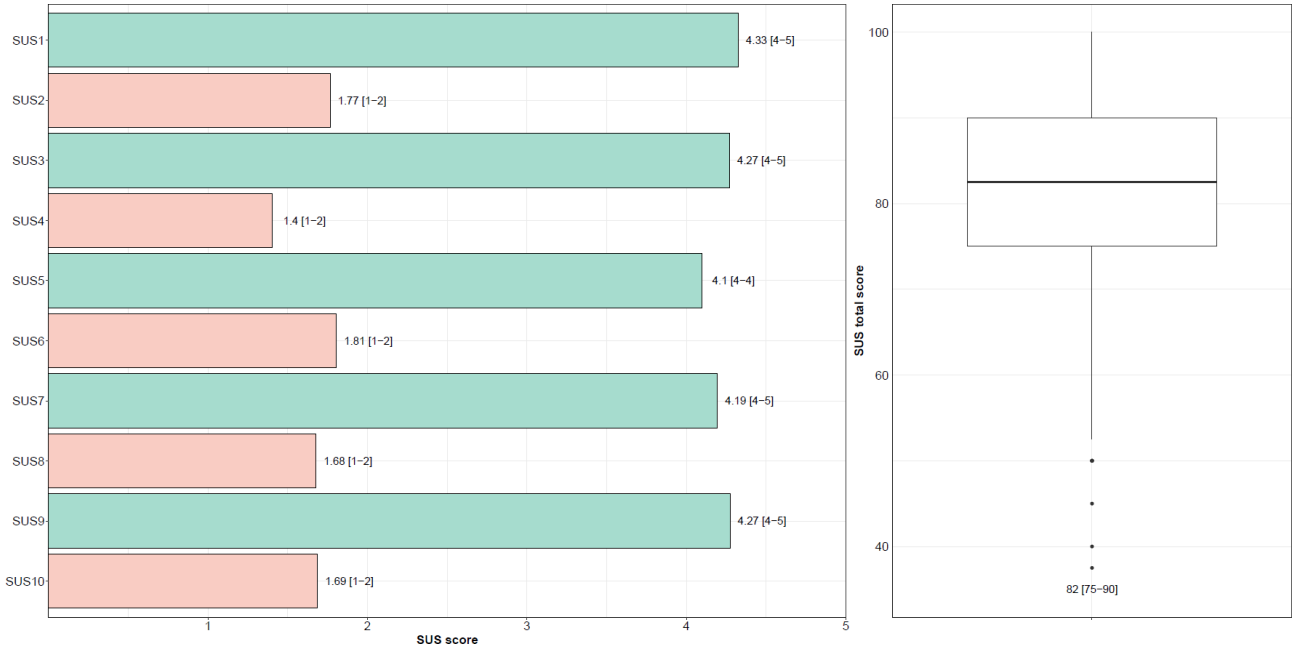

Most respondents (239/254, 94.1%) reported actively using the app at the time of the survey; 52.4% (133/254) used it daily and 80.7% (205/254) at least once per week. Table 3 provides an overview of motivations for app use, specific use cases, and the extent to which the app was integrated into medical care.

XSL•FO
RenderX

Table . App usage patterns, reasons for use, integration into medical care, and perceived usefulness among adult users of *Hypertension.APP* in Germany, based on a cross-sectional online survey conducted between January and September 2023 (N=254).

Characteristic (multiple responses possible/top 5)	Values, n (%)
Source of app awareness	
Independently searched for blood pressure apps online or in the app store	112 (44.1)
Recommendation by a doctor	45 (17.7)
Advertisement in print media	30 (11.8)
Recommendation from friends/family	27 (10.6)
Online advertisement	20 (7.9)
Learned about the app in another way	31 (12.2)
Reasons for using the app (user motivations)	
To document blood pressure	220 (86.6)
To get information about high blood pressure	107 (42.1)
To do something for one's health	76 (29.9)
Own curiosity	75 (29.5)
To try deep breathing function	34 (13.4)
App use cases (features actually used)	
Documentation blood pressure values	242 (95.3)
Personalized health information suggestions	169 (66.5)
Diary function	102 (40.2)
General health information	96 (37.8)
To create reports for doctor's visits	86 (33.9)
App integrated into medical treatment	
Using blood pressure records to prepare for doctor's visits	133 (52.4)
App not yet integrated, but planned for future use	76 (29.9)
App not currently integrated into care	51 (20.1)
Medication adjusted by physician based on patient-reported entries	30 (11.8)
Discussing advice or health information with physician	14 (5.5)
Best feature of the app	
Blood pressure documentation	230 (90.6)
Direct feedback on blood pressure values	138 (54.3)
Preparation of reports for doctor's visit	111 (43.7)
Personalized health information	98 (38.6)
Diary entries	77 (30.3)

To explore potential predictors of usage frequency, we conducted univariable ordinal logistic regression analyses (full results are provided in Multimedia Appendix 3). Of the 10 sociodemographic and clinical variables examined, 4 were associated with app usage frequency. Having a SBP of 140 mm Hg or higher was associated with less frequent app use, whereas living in a small town (5000 - 20,000 inhabitants), having received an initial hypertension diagnosis more than 1 year ago, and having a higher level of education were associated with more frequent app use.

Participants reported using the app for multiple purposes. The most common use cases were documentation of blood pressure values (242/254, 95.3%), accessing personalized health information (169/254, 66.5%), and using the diary function (102/254, 40.2%). More than half of participants (133/254,

XSL•FO
RenderX

52.4%) indicated that they used the app to prepare for doctor's visits. However, only a minority reported that the app had been formally integrated into their medical care (51/254, 20.1%) or that their physicians had adjusted medications based on app data (30/254, 11.8%). When asked about the app's most helpful features, participants most frequently cited blood pressure documentation (230/254, 90.6%), direct feedback on blood pressure values (138/254, 54.3%), and preparation of reports for medical visits (111/254, 43.7%).

Awareness of the app was most commonly gained through self-initiated searches in the app store or online (112/254, 44.1%), followed by recommendations from health care professionals (45/254, 17.7%), advertisements in print media (30/254, 11.8%), and recommendations from friends or family (27/254, 10.6%).

Discussion

Overview

This cross-sectional survey provides insight into the profiles, usage patterns, and engagement of active users of a mobile app in hypertension care. Most users were digitally literate, middle-aged or older adults with higher education levels and above-average income. While app utility ratings and engagement were generally high, especially for documentation and preparation of doctor visits, formal integration of the app into clinical care remained limited.

Who Uses the App? The Usual Suspects

Taken together, our data underscore a persistent digital divide [24]. The user base of the app largely reflects the "usual suspects" of digital health adoption: individuals with higher levels of education and income, often residing in urban or semiurban areas, and possessing moderate to high digital health literacy. These users are typically proactive in seeking out digital tools, show high engagement, and report favorable usability experiences (SUS). In contrast, individuals with lower digital competencies or limited access to technology were underrepresented, even though such groups often bear a disproportionate hypertension burden and face greater challenges in self-management [25].

This pattern aligns closely with prior research. A 2025 Dutch study demonstrated that people with a low socioeconomic position had significantly less access to and use of eHealth, whereas individuals with a higher socioeconomic position showed substantially greater utilization [26]. One key explanatory factor is eHealth literacy: highly educated individuals generally have stronger eHealth literacy and comfort with technology [26], which lowers barriers to using health apps. Conversely, individuals from disadvantaged backgrounds often face multiple obstacles, including limited skills, fewer resources, or lower confidence, and are therefore more likely to adopt new digital health tools later—or not at all [26].

Interestingly, digital health literacy was not associated with app usage frequency in our sample. At first glance, this appears to contradict previous research. However, our sample consisted exclusively of active app users, which implies a pronounced self-selection process: individuals with lower digital health literacy are less likely to adopt such tools in the first place. Among those who already use an app, variability in digital literacy is typically reduced, and literacy may no longer influence usage intensity. Moreover, eHEALS measures perceived rather than actual digital skills, which may further attenuate associations. Taken together, these factors likely explain why digital health literacy did not emerge as a significant predictor of usage frequency in this study.

Consistent with our findings, prior surveys in the United States and Europe have similarly found that health app users are typically younger or middle-aged adults with higher income and education levels [27,28]. Thus, despite the widespread availability of mHealth apps, their use remains socially patterned and disproportionately concentrated among individuals with higher socioeconomic status and digital competencies, as noted by multiple researchers [24,29]. Our study reinforces this point: access alone is not uniform, and digital interventions in hypertension care may be inadvertently catering to an already empowered subset of patients.

Notably, participants with SBP of 140 mm Hg or higher used the app less frequently, suggesting that worse current control may correlate with lower digital engagement. Interestingly, a noteworthy proportion of participants reported diastolic blood pressure values above recommended thresholds. While we did not perform subgroup analyses exploring associations between location, sociodemographic features, or other predictors and blood pressure levels, this finding merits attention. Elevated diastolic readings may reflect limited hypertension control among specific subgroups, which could affect both engagement with self-monitoring tools and clinical outcomes. Future studies should explore whether such patterns are associated with geographic, educational, or behavioral factors and assess whether app usage may contribute to improved blood pressure control over time.

While causal inferences cannot be drawn, similar paradoxical patterns have been observed elsewhere. For example, a German survey reported lower health app use among individuals who perceived themselves at high diabetes risk [30]. Even after adjusting for demographics, high perceived risk was associated with lower app usage, suggesting that individuals with greater health concerns or disease burden might rely more on traditional care or face motivational barriers to using apps. These findings highlight the need for targeted strategies to better support and engage higher-risk patients in digital self-management.

Implications for Research and Practice

Our results highlight an unmet need to make digital health technologies more inclusive and accessible. Currently, mHealth interventions often serve the "tech-savvy" but miss those with limited digital experience or lower health literacy, a pattern that can exacerbate health disparities [29]. To ensure these tools benefit a broader patient population, developers and health care providers should consider strategies to lower the barriers to entry for less digitally literate users. Recommended approaches include the following:

- Usability and language: Apps should use clear navigation and plain language. Short sentences improve

XSL•FO
RenderX

comprehension, and intuitive icons and visuals can support users who struggle with text-heavy information. Furthermore, providing content and instructions in multiple languages is crucial in diverse populations [31].

- Low-threshold onboarding and education: It should be easy for new users to get started. This could involve in-app tutorials, community workshops, or integration with primary care so that health care staff can introduce and demonstrate the app. Such onboarding can build confidence for those who are less technologically experienced. Digital navigators [32] could serve as an important bridge between patients and technology and could play a key role in this process; however, they have yet to be widely adopted in routine care.
- Collaboration with health care providers (clinical integration): Stronger links between app data and primary care workflows would benefit both patients and clinicians. Integrating the app with electronic health records and enabling data sharing with physicians would enable routine use of these data for monitoring and decision-making during clinic visits.

From a health systems perspective, clinical integration is pivotal. Although patient engagement was high, incorporation of the app into routine care by providers was low—a disconnect commonly seen with mHealth tools. Experts argue that integrating patient-facing apps into standard clinical practice is essential for digital health to reach its full potential [33]. At present, most consumer health apps operate in isolation—patients may track their data, but it often stays on their phone and is not mandatorily reviewed by their clinician. In line with this, few physicians today formally recommend or prescribe health apps to their patients, often due to limited time and lack of information, reimbursement, or guidelines [12-16]. Addressing these barriers will require collaboration among developers, health care providers, and policy makers to ensure apps meet clinical standards, fit seamlessly into workflows, and address patient needs. Ultimately, stronger collaboration and integration are necessary for mHealth to move from a personal self-management aid to a recognized element of health care delivery.

Limitations and Generalizability

To our knowledge, this is one of the first quantitative studies to systematically characterize real-world users of a hypertension self-management app in Germany. Prior work in this field has primarily relied on small qualitative samples or clinical trial settings, which often do not reflect everyday app usage. By combining detailed sociodemographic, clinical, and digital literacy data with real-world app engagement patterns, our study advances understanding of who uses hypertension apps and how these tools are integrated—or not integrated—into routine care. The finding that a digital divide persists even among active app users, as well as the inverse association between elevated SBP and frequency of app use, provides novel insights that have not been reported in previous hypertension mHealth research.

Yet, it is important to acknowledge the study's limitations. First, our data come from a cross-sectional survey based on self-reported app use, behaviors, and blood pressure values. Self-reported data are susceptible to recall bias (participants might not remember their usage accurately) and social desirability bias (participants might overstate "good" behaviors, like app use, or underreport problems). In addition, blood pressure measurements were self-reported by users and not clinically validated, which may have introduced measurement inaccuracies or reporting bias. This is a common caveat in mHealth surveys, and while we took steps to anonymize responses, some bias cannot be ruled out.

Second, the sample is not representative of all individuals with hypertension or of all app users. Due to the self-selection recruitment process, the study inherently attracted participants who were receptive to research, which might also be associated with academic backgrounds. Furthermore, the convenience sample—recruited within *Hypertension.APP*—limits representativeness for all people with hypertension or all hypertension app users.

Third, Germany's high-resource, digitally mature context may not reflect lower-resource settings. What we observed in an urban European context (high smartphone penetration, moderate-to-high digital literacy) might not directly translate to settings where access to smartphones, the internet, or health education is more limited.

A further limitation is that important sociodemographic factors known to influence mHealth adoption, such as ethnicity, migration background, and primary language, were not collected. This restricts our ability to assess disparities along these dimensions and may obscure differential patterns of digital engagement. Finally, income was measured with a coarse, self-developed scale that bunched many participants into the highest category, limiting socioeconomic resolution. A more granular income measure would have provided a clearer picture of the socioeconomic background of app users.

Conclusion

In summary, users of the hypertension self-management app in Germany were predominantly well-educated, digitally literate individuals with a history of hypertension. While the app was rated as useful and showed high engagement, particularly for tracking health data and preparing for medical appointments, its formal integration into clinical care remained limited. Furthermore, user profiles point to a persistent digital divide, suggesting that socioeconomically disadvantaged groups are underrepresented. Simply making an app available is insufficient; who uses it, how it is used, and whether it is integrated into care are critical. Future research should test strategies to engage underrepresented groups, strengthen confidence in digital health information, and link patient-facing apps more effectively with clinical care so that mHealth can be more inclusive and effective in managing hypertension.

XSL•FO
RenderX

Acknowledgments

The authors would like to deeply thank the participants of this study. For language editing and readability, the authors used ChatGPT (version 5) during the preparation of this manuscript. All content was subsequently reviewed and revised by the authors, who take full responsibility for the final version of the publication.

Funding

This project was supported by a grant of the Federal Joint Committee (Gemeinsamer Bundesausschuss, G-BA) in Germany from May 2022 to April 2025 (Innovation fund grant number 01VSF21042). This work was also partially supported by funding from the French government, managed by the National Research Agency (ANR) under the France 2030 program (ANR-23-IACL-0006) and under the "Investissements d'avenir" program (ANR-10-AIRT-0005 and ANR-15-IDEX-0002). The funders of the study had no influence in the study design.

Data Availability

The datasets generated or analyzed during this study are not publicly available as participants did not provide consent for data sharing. For further questions regarding the reuse of data, please contact the corresponding author.

Authors' Contributions

All authors were involved in drafting the article and critically revising it for important intellectual content, and all authors approved the final version to be submitted for publication.
FM had full access to all the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis.
Study conception and design: FM, SM, EW, OJ, SS, DB.
Acquisition of data: FM, DB, OJ.
Analysis and interpretation of data: FM, PP, NV.

Conflicts of Interest

OJ is CEO of *Hypertension Care UG*, which offers *Hypertension.APP*. OJ and MM are founders of *Hypertension.APP*. These authors contributed to the conception of the study and supported recruitment but were not involved in the statistical analysis or interpretation of the data. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Multimedia Appendix 1
German questionnaire.
[PDF File, 637 KB - jmir_v28i1e83075_app1.pdf]

Multimedia Appendix 2
English version of the questionnaire.
[PDF File, 649 KB - jmir_v28i1e83075_app2.pdf]

Multimedia Appendix 3
Association between app usage frequency and sociodemographic variables.
[PDF File, 210 KB - jmir_v28i1e83075_app3.pdf]

References

1. Roth GA, Mensah GA, Johnson CO, et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. J Am Coll Cardiol 2020 Dec 22;76(25):2982-3021. [doi: 10.1016/j.jacc.2020.11.010] [Medline: 33309175]
2. Forouzanfar MH, Liu P, Roth GA, et al. Global burden of hypertension and systolic blood pressure of at least 110 to 115 mm Hg, 1990-2015. JAMA 2017 Jan 10;317(2):165-182. [doi: 10.1001/jama.2016.19043] [Medline: 28097354]
3. Williams B, Mancia G, Spiering W, et al. 2018 ESC/ESH Guidelines for the management of arterial hypertension. Eur Heart J 2018 Sep 1;39(33):3021-3104. [doi: 10.1093/eurheartj/ehy339] [Medline: 30165516]
4. Egan BM, Li J, Sutherland SE, Rakotz MK, Wozniak GD. Hypertension control in the United States 2009 to 2018: factors underlying falling control rates during 2015 to 2018 across age- and race-ethnicity groups. Hypertension 2021 Sep;78(3):578-587. [doi: 10.1161/HYPERTENSIONAHA.120.16418] [Medline: 34120453]

XSL•FO
RenderX

5. Parati G, Goncalves A, Soergel D, et al. New perspectives for hypertension management: progress in methodological and technological developments. Eur J Prev Cardiol 2023 Jan 11;30(1):48-60. [doi: 10.1093/eurjpc/zwac203] [Medline: 36073370]
6. Kuwabara A, Su S, Krauss J. Utilizing digital health technologies for patient education in lifestyle medicine. Am J Lifestyle Med 2019;14(2):137-142. [doi: 10.1177/1559827619892547] [Medline: 32231478]
7. Mikulski BS, Bellei EA, Biduski D, De Marchi ACB. Mobile health applications and medication adherence of patients with hypertension: a systematic review and meta-analysis. Am J Prev Med 2022 Apr;62(4):626-634. [doi: 10.1016/j.amepre.2021.11.003] [Medline: 34963562]
8. Prendergast H, Kitsiou S, Petzel Gimbar R, et al. Emergency department–based education and mHealth empowerment intervention for hypertension: the TOUCHED randomized clinical trial. JAMA Cardiol 2025 Jul 1;10(7):657-665. [doi: 10.1001/jamacardio.2025.0675] [Medline: 40266598]
9. Labinsky H, Gupta L, Raimondo MG, Schett G, Knitza J. Real-world usage of digital health applications (DiGA) in rheumatology: results from a German patient survey. Rheumatol Int 2023 Apr;43(4):713-719. [doi: 10.1007/s00296-022-05261-7] [Medline: 36543961]
10. Boy K, May S, Labinsky H, et al. Back on track - digital health applications to treat back pain of rheumatic patients? Results of a qualitative interview study. Rheumatol Int 2024 Dec;44(12):2977-2987. [doi: 10.1007/s00296-024-05726-x] [Medline: 39340666]
11. Albrecht A, Taubmann J, Minopoulou I, et al. Real-world-evidence of digital health applications (DiGAs) in rheumatology: insights from the DiGAReal registry. Rheumatol Ther 2025 Apr;12(2):267-282. [doi: 10.1007/s40744-025-00744-y] [Medline: 39836353]
12. Peng W, Kanthawala S, Yuan S, Hussain SA. A qualitative study of user perceptions of mobile health apps. BMC Public Health 2016 Nov 14;16(1):1158. [doi: 10.1186/s12889-016-3808-0] [Medline: 27842533]
13. May S, Muehlensiepen F, Wengemuth E, et al. Benefits and barriers to mHealth in hypertension care: qualitative study with German health care professionals. JMIR Hum Factors 2025 Mar 10;12:e52544. [doi: 10.2196/52544] [Medline: 40063928]
14. May S, Seifert F, Bruch D, Heinze M, Spethmann S, Muehlensiepen F. Insights into how mHealth applications could be introduced into standard hypertension care in Germany: qualitative study with German cardiologists and general practitioners. JMIR mHealth uHealth 2025 Mar 28;13:e56666. [doi: 10.2196/56666] [Medline: 40153776]
15. Melchiorre MG, Lamura G, Barbabella F, ICARE4EU Consortium. eHealth for people with multimorbidity: results from the ICARE4EU project and insights from the “10 e’s” by Gunther Eysenbach. PLoS One 2018;13(11):e0207292. [doi: 10.1371/journal.pone.0207292] [Medline: 30427924]
16. Ardito V, Golubev G, Ciani O, Tarricone R. Evaluating barriers and facilitators to the uptake of mHealth apps in cancer care using the consolidated framework for implementation research: scoping literature review. JMIR Cancer 2023 Mar 30;9:e42092. [doi: 10.2196/42092] [Medline: 36995750]
17. Muehlensiepen F, Bruch D, Seifert F, et al. mHealth apps for hypertension self-management: interview study among patient-users. JMIR Form Res 2024 Sep 27;8:e56162. [doi: 10.2196/56162] [Medline: 39331954]
18. Bruch D, Muehlensiepen F, May S, et al. Digital Preventive Measures for Arterial Hypertension (DiPaH) - a mixed-methods study protocol for health services research. Front Cardiovasc Med 2022;9:1089968. [doi: 10.3389/fcvm.2022.1089968] [Medline: 36704456]
19. Hypertension Care UG. HypertensionAPP. 2025. URL: https://www.hypertonie.app/?lang=en [accessed 2026-01-27]
20. Soellner R, Huber S, Reder M. The concept of eHealth literacy and its measurement: German translation of the eHEALS. J Media Psychol 2014;26(1):29-38. [doi: 10.1027/1864-1105/a000104]
21. Gao M, Kortum P, Oswald FL. Multi-language toolkit for the system usability scale. Int J Hum Comput Interact 2020 Dec 13;36(20):1883-1901. [doi: 10.1080/10447318.2020.1801173]
22. Monkman H, Kushniruk AW, Barnett J, Borycki EM, Greiner LE, Sheets D. Are health literacy and eHealth literacy the same or different? Stud Health Technol Inform 2017;245:178-182. [doi: 10.3233/978-1-61499-830-3-178] [Medline: 29295077]
23. Venables WN, Ripley BD. Modern Applied Statistics with S, 4th edition: Springer; 2002. [doi: 10.1007/978-0-387-21706-2]
24. Equity within digital health technology within the WHO European Region: a scoping review. : World Health Organization Regional Office for Europe; 2022 URL: https://iris.who.int/server/api/core/bitstreams/c0b312f6-4877-4d2a-aa66-769e84625226/content [accessed 2026-01-27]
25. Leng B, Jin Y, Li G, Chen L, Jin N. Socioeconomic status and hypertension: a meta-analysis. J Hypertens 2015 Feb;33(2):221-229. [doi: 10.1097/HJH.0000000000000428] [Medline: 25479029]
26. Standaar L, van Tuyl L, Suijkerbuijk A, Brabers A, Friele R. Differences in eHealth access, use, and perceived benefit between different socioeconomic groups in the Dutch context: secondary cross-sectional study. JMIR Form Res 2025 Jan 7;9:e49585. [doi: 10.2196/49585] [Medline: 39773883]
27. Ernsting C, Dombrowski SU, Oedekoven M, et al. Using smartphones and health apps to change and manage health behaviors: a population-based survey. J Med Internet Res 2017 Apr 5;19(4):e101. [doi: 10.2196/jmir.6838] [Medline: 28381394]

XSL•FO
RenderX

28. Krebs P, Duncan DT. Health app use among US mobile phone owners: a national survey. JMIR mHealth uHealth 2015 Nov 4;3(4):e101. [doi: 10.2196/mhealth.4924] [Medline: 26537656]
29. Hengst TM, Lechner L, Dohmen D, Bolman CA. The facilitators and barriers of mHealth adoption and use among people with a low socio-economic position: a scoping review. Digit Health 2023;9:20552076231198702. [doi: 10.1177/20552076231198702] [Medline: 37691766]
30. Stühmann LM, Paprott R, Heidemann C, et al. Health app use and its correlates among individuals with and without type 2 diabetes: nationwide population-based survey. JMIR Diabetes 2020 May 20;5(2):e14396. [doi: 10.2196/14396] [Medline: 32432555]
31. eHealth literacy. Centers for Disease Control and Prevention. 2024 Jun 21. URL: https://www.cdc.gov/health-literacy/php/research-summaries/ehealth.html [accessed 2026-01-27]
32. DigiNavi.: MHB IAG Mental Health Policy & Digitization; 2025. URL: https://www.diginavi.de/en/ [accessed 2025-07-27]
33. Gordon WJ, Landman A, Zhang H, Bates DW. Beyond validation: getting health apps into clinical practice. NPJ Digit Med 2020;3:14. [doi: 10.1038/s41746-019-0212-z] [Medline: 32047860]

Abbreviations

eHEALS: eHealth Literacy Scale
mHealth: mobile health
SBP: systolic blood pressure
SUS: system usability scale

Edited by A Stone; submitted 27.Aug.2025; peer-reviewed by G Kerr, K Lakshminarayan; accepted 16.Dec.2025; published 11.Feb.2026.

Please cite as:
Muehlensiepen F, May S, Seifert F, Wengemuth E, Johannsen O, Middeke M, Heinze M, Petit P, Vuillerme N, Spethmann S, Bruch D
User Profiles and Engagement in a Hypertension Self-Management App: Cross-Sectional Survey
J Med Internet Res 2026;28:e83075
URL: https://www.jmir.org/2026/1/e83075
doi:10.2196/83075

© Felix Muehlensiepen, Susann May, Frances Seifert, Eileen Wengemuth, Olen Johannsen, Martin Middeke, Martin Heinze, Pascal Petit, Nicolas Vuillerme, Sebastian Spethmann, Dunja Bruch. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Associations Between Short-Video Platform Use and Health Across Health Distribution and Usage Behaviors in China: Cross-Sectional Questionnaire Study

Yangyang Pan[1], MSc; Kangkang Zhang[1], MSc, PhD; Yilin Wei[1], MA, PhD; Yangzhen Huang[1], MSc; Chengxu Long[1,2], MSc, PhD; Chenxin Yang[1], MSc, PhD; Shangfeng Tang[1,2], MSc, PhD

[1]School of Medicine and Health Management, Huazhong University of Science and Technology, No. 13, Aviation Road, Qiaokou District, Wuhan, China

[2]Research Center for Rural Health Service, Key Research Institute of Humanities & Social Sciences of Hubei Provincial Department of Education, Wuhan, China

Corresponding Author:
Shangfeng Tang, MSc, PhD
School of Medicine and Health Management, Huazhong University of Science and Technology, No. 13, Aviation Road, Qiaokou District, Wuhan, China

Abstract

Background: Short-video platforms, characterized by algorithmic curation and passive consumption, have emerged as dominant components of digital life. However, the associations between short-video platform use and health across different groups and usage behaviors remain understudied.

Objective: This study investigates associations between short-video platform use and health, examining whether these relationships vary across health status, usage behaviors, and socioeconomic status.

Methods: A cross-sectional study was conducted using multistage stratified sampling across eastern, central, and western China from July to September 2024. The inclusion criteria were age 18 years or older, ability to communicate effectively, and no cognitive disorders or mental disturbance. Of 7725 participants enrolled, 46.96% (n=3628) were male, and the average age was 65.49 (SD 8.39) years. The data were collected via face-to-face interviews using a structured questionnaire. Self-rated health and relative health deprivation (Kakwani index) were used to measure health. Quantile regression explored associations between whether using short-video platform and health varies across the health distribution, while linear regression examined associations of years, frequency, daily duration, and purpose diversity of short-video platform use with health. Moderating effect analysis explored the role of socioeconomic status in the relationship between the daily duration of use and health.

Results: Coefficients were tested using 2-tailed *t* tests, and statistical significance was defined as a 2-sided *P* value less than .05. Quantile regression revealed heterogeneous associations. Compared to nonusers, short-video platform users had better self-rated health at the 70th to 90th quantiles and lower relative health deprivation at the 10th to 30th quantiles. However, the users at the 10th quantile of self-rated health had worse self-rated health (β=−2.224, 95% CI −3.835 to −0.613). Longer engagement (≥3 y) correlated with lower relative health deprivation (β=1.970, 95% CI 0.308-3.632), while daily use of 1 - 4 hours was associated with poorer self-rated health (β=−3.385, 95% CI −4.872 to −1.898; β=−3.038, 95% CI −5.054 to −1.022) and higher relative health deprivation (β=0.035, 95% CI 0.021-0.050; *P*<.001; β=0.034, 95% CI 0.014-0.054). Compared to no purposeful use, using with 2 purposes was associated with better self-rated health (β=6.082, 95% CI 0.250-11.914) and lower relative health deprivation (β=−0.063, 95% CI −0.120 to −0.005). The association was stronger for use with 3 or more purposes. Socioeconomic status moderated the relationship between daily duration of use and health.

Conclusions: This study provides a more specific investigation of how these associations vary across health strata and usage patterns. The findings reveal patterns of benefit and risk across population subgroups, underscoring that how and why individuals engage with platforms matter more than mere access or frequency. These insights necessitate targeted digital well-being policies that protect vulnerable groups, particularly those in poor health or with lower socioeconomic status. Furthermore, policies should actively encourage intentional, functionally grounded use to reduce health inequities and advance equitable digital inclusion.

(J Med Internet Res 2026;28:e86526) doi:10.2196/86526

KEYWORDS

short-video platform; self-rated health; quantile regression; health disparities; socioeconomic status

XSL•FO
RenderX

Introduction

The rapid proliferation of digital technologies has fundamentally reconfigured human behavior and social dynamics, establishing digital engagement as a pivotal social determinant of health [1]. Emerging evidence suggests that digital engagement patterns exhibit complex relationships with physical and mental well-being [2-6]. Some studies indicate that digital engagement promotes social support and information access [7,8], while others suggest that it correlates with anxiety, depression, and sedentary behavior [9,10]. This paradoxical finding underscores the necessity to consider technological characteristics and user experience variability when assessing the relationship between digital engagement and health.

Among digital platforms, short-video platforms, such as TikTok, Douyin, and Reels, represent a unique research frontier given their algorithmic-driven content delivery, passive consumption modalities, and entertainment-centric design [11]. In China, over 1 billion users engage with short-video content, accounting for 95.2% of internet users as of June 2023 [12]. However, high-frequency passive browsing and performative interactions features of short-video platforms can amplify the potential benefits of digital engagement, as well as exacerbate their risks. These platforms mediate health-related behaviors through 3 intersecting pathways: information, behavior, and psychology. Information pathways primarily shape individuals' health knowledge and literacy. Studies have shown that internet use for health information was associated with higher health literacy and health service use [13]. But perceived health information overload discourages enhancing health literacy [14]. Behavioral pathways encompass habitual usage behaviors and health-promoting behaviors. Access to health information facilitates the development of health behaviors [15,16], while sedentary behavior and addictive behavior during internet use may adversely affect health [17,18]. Psychological pathways involve platform-based social interactions and psychological symptoms. On the one hand, the internet can expand social networks, providing social support and entertainment [19,20]. On the other hand, problematic internet use may lead to symptoms of anxiety and depression [21]. Crucially, the relationship between these pathways and health is unlikely to be uniform across all individuals but rather varies according to their existing resources and health status.

However, existing studies predominantly focus on aggregate associations between digital engagement and health outcomes. The heterogeneous association between short-video platform use and health regarding different population subgroups remains understudied. One key issue is how its multidimensional characteristics, including the time, frequency, duration, and purpose of short-video platform use, are related to individual health. Moreover, engagement with short-video platforms has been an emerging behavioral determinant nested within broader socioeconomic and environmental structures. The social determinants of health theory holds that health and health inequalities are determined by social, economic, and environmental factors [22]. The digital health paradox further sharpens this lens, highlighting that people excluded by digital technology face the risk of being abandoned by society in most cases. So, people who have very limited access to the internet are precisely those who need the most health services and bear the heaviest health burden, such as low-income populations and long-term ill individuals [23]. Therefore, it is necessary to consider whether the association between short-video platform use and health exhibits heterogeneity across different health distributions and socioeconomic status (SES) groups.

To comprehensively understand the relationship between short-video platform use and individual health, the measurement of health also requires a multidimensional perspective. Self-rated health has been validated to comprehensively reflect overall health status and effectively predict mortality [24], yet it fails to capture the distribution and accumulation of health conditions. To address this limitation, we incorporate relative health deprivation, operationalized through the Kakwani index [25]. Relative deprivation theory holds that an individual's health level is negatively correlated with health disadvantages within a specific group, which means the lower the health level, the greater the health disadvantage and the higher the degree of relative deprivation of health faced. By calculating the health deprivation index, it is possible to reflect the long-term and structural accumulation of health disadvantages.

This study aims to investigate the heterogeneous associations between short-video platform use and individual health across different health status and usage behaviors. Specifically, we aim to (1) examine the association between whether using short-video platform and health varies across the health distribution, (2) analyze the associations between multidimensional usage behaviors (years, frequency, daily duration, and purpose diversity) and health, and (3) explore the moderating role of SES in the relationship between daily duration of use and health. By identifying vulnerable subgroups involved in digital participation, this study provides critical, actionable insights for designing targeted digital health interventions and equitable platform governance policies.

Methods

Inclusion and Exclusion

Individuals who met the following criteria were included in this survey: (1) 18 years of age or older and (2) were able to communicate effectively with investigators. The exclusion criteria included (1) no cognitive disorders and (2) no mental disturbance. The data were collected using a structured questionnaire. Questionnaires with completion times below 5 minutes or above 50 minutes were excluded, as pilot testing indicated these thresholds captured inattentive or interrupted responses. Questionnaires with a response missing rate exceeding 50% for the purpose of quality control were also excluded.

Participant Characteristics

Of the 7725 participants included in the final analysis, 3628 (46.96%) were male, and the average age was 65.49 (SD 9.39) years. A total of 4403 (57.0%) participants were users of short-video platforms. Further detailed characteristics of the participants, including demographic, socioeconomic,

XSL•FO
RenderX

health-related, and social support factors, are presented in Table 1.

XSL•FO
RenderX

Table . Characteristics of the participants before propensity matching.

Characteristics	All	Short-video platforms		Self-rated health		Relative health deprivation	
		Users	Nonusers	Mean (SD)	*P* value	Mean (SD)	*P* value
N	7725 (100)	4403 (57.00)	3322 (43.00)	__[a]	—	—	—
Age (y), mean (SD)	65.49 (9.39)	61.92 (8.68)	70.21 (8.12)	65.59 (17.17)	<.001	0.14 (0.17)	<.001
Sex, n (%)					.007		.09
Male	3628 (46.96)	2270 (51.56)	1358 (40.88)	66.15 (17.40)		0.14 (0.17)	
Female	4097 (53.04)	2133 (48.44)	1964 (59.12)	65.09 (16.96)		0.15 (0.17)	
Marriage, n (%)					<.001		<.001
Married or co-habiting	6492 (84.04)	3966 (90.07)	2526 (76.04)	66.17 (16.98)		0.14 (0.16)	
Other marital status[b]	1233 (15.96)	437 (9.93)	796 (23.96)	62.52 (17.85)		0.17 (0.18)	
Urban-rural residence, n (%)					<.001		<.001
Urban	2342 (30.32)	1332 (30.25)	1010 (30.40)	67.93 (16.02)		0.12 (0.15)	
Rural	5383 (69.68)	3071 (69.75)	2312 (69.60)	65.59 (17.56)		0.15 (0.17)	
Socioeconomic status, n (%)					<.001		<.001
Low	2880 (37.28)	1622 (36.84)	1258 (37.87)	61.07 (18.19)		0.19 (0.19)	
Middle	4310 (55.79)	2416 (54.87)	1894 (57.01)	68.00 (15.86)		0.12 (0.14)	
High	535 (6.93)	365 (8.29)	170 (5.12)	70.39 (16.49)		0.11 (0.15)	
Education, n (%)					<.001		<.001
Illiterate or semiliterate	2960 (38.32)	1104 (25.07)	1856 (55.87)	63.45 (16.86)		0.16 (0.17)	
Primary school	2277 (29.48)	1336 (30.34)	941 (28.33)	65.42 (17.56)		0.15 (0.17)	
Middle school	1619 (20.96)	1281 (29.09)	338 (10.17)	67.46 (16.96)		0.13 (0.16)	
High school or vocational school	686 (8.88)	535 (12.15)	151 (4.55)	69.40 (16.72)		0.11 (0.16)	
College degree / Bachelor's degree or above	183 (2.37)	147 (3.34)	36 (1.08)	71.30 (15.26)		0.10 (0.14)	
Employment status, n (%)					<.001		<.001
Unemployed	4539 (58.76)	2337 (53.08)	2202 (66.29)	64.22 (17.51)		0.16 (0.18)	
Employed	3186 (41.24)	2066 (46.92)	1120 (33.71)	67.53 (16.49)		0.13 (0.15)	
Convenience of obtaining medical services, n (%)					<.001		<.001
Low	22 (0.29)	15 (0.34)	7 (0.21)	55.18 (27.16)		0.28 (0.32)	
Medium low	284 (3.68)	104 (2.36)	180 (5.42)	55.10 (19.23)		0.25 (0.21)	
Middle	900 (11.67)	378 (8.59)	522 (15.71)	59.49 (18.08)		0.20 (0.19)	
Medium high	4394 (57.00)	2448 (55.60)	1946 (58.58)	66.55 (16.46)		0.13 (0.16)	
High	2109 (27.36)	1458 (33.11)	667 (20.08)	67.74 (16.72)		0.13 (0.16)	
Number of persons living together, n (%)					.008		<.001
0	740 (9.58)	319 (7.25)	421 (12.67)	63.92 (16.88)		0.16 (0.17)	
1-3	3961 (51.28)	2331 (52.94)	1630 (49.07)	65.22 (17.48)		0.15 (0.17)	
4-6	2657 (34.39)	1543 (35.04)	1114 (33.53)	66.36 (16.96)		0.14 (0.16)	
>6	367 (4.75)	210 (4.77)	157 (4.73)	67.34 (15.42)		0.12 (0.14)	

XSL•FO
RenderX

Characteristics	All	Short-video platforms		Self-rated health		Relative health deprivation	
		Users	Nonusers	Mean (SD)	*P* value	Mean (SD)	*P* value
Number of persons contacting with regularly, n (%)					<.001		<.001
0	307 (3.97)	151 (3.43)	156 (4.70)	58.57 (19.56)		0.22 (0.21)	
1	1348 (17.45)	583 (13.24)	765 (23.03)	64.94 (17.34)		0.15 (0.18)	
2-3	2855 (3.96)	1600 (36.34)	1255 (37.78)	65.76 (1660)		0.14 (0.16)	
4-5	1567 (20.28)	915 (20.78)	652 (19.63)	66.44 (15.91)		0.13 (0.15)	
≥6	1648 (21.33)	1154 (26.21)	494 (14.87)	66.31 (18.36)		0.14 (0.18)	
Multimorbidity-weighted index, mean (SD)	2.51 (2.02)	2.40 (1.96)	2.66 (2.09)	65.59 (17.17)	<.001	0.14 (0.17)	<.001
Depression, mean (SD)	5.87 (4.78)	5.28 (4.50)	6.65 (5.02)	65.59 (17.17)	<.001	0.14 (0.17)	<.001
Anxiety, mean (SD)	2.07 (3.45)	1.82 (3.22)	2.39 (3.71)	65.59 (17.17)	<.001	0.14 (0.17)	<.001
Loneliness, mean (SD)	3.70 (1.29)	3.54 (1.19)	3.90 (1.39)	65.59 (17.17)	<.001	0.14 (0.17)	<.001

[a]Not applicable.

[b]Other marital status includes being widowed, separated, divorced, or never married.

Sampling Procedures

Multistage stratified sampling was applied to select participants in the survey from July to September 2024. First, initial stratification was conducted based on geographical region. Fujian province, Hubei province, and Chongqing municipality were chosen as the survey areas, which are respectively located in the eastern, central, and western regions of China. Second, within each province, 6 - 12 urban subdistricts and rural townships were randomly selected as secondary sampling units, ultimately covering 24 subdistricts or townships. Finally, through the community registries, 300 - 400 participants were randomly selected from each subdistrict or township. Sampling weights were not applied in the analyses, and all estimates are presented as unweighted statistics. The sampling design aimed to ensure structural representativeness across regional and urban-rural strata within the selected provinces. The data were collected through face-to-face interviews with the respondents by uniformly trained investigators, and the questionnaires were filled out simultaneously. This study was conducted and reported following the Journal Article Reporting Standards for quantitative research [26].

Sample Size, Power, and Precision

A total of 7725 valid samples were included from the survey, yielding an effective recovery rate of 85.48%. All the analyses were conducted on this sample of valid questionnaires, and there were no missing data for the variables required by this study. Sample size estimation followed the standard formula for proportion-based calculations: [n=uα2pq⁄d2], where uα represents the critical value for a 95% CI (uα=1.96), *p* is the estimated proportion of short-video platform users based on the 52nd Statistical Report on Internet Development in China—Research on Internet Development [12], *q* is the complementary proportion (q=1-p), and d denotes the allowable error (1.2%). Using these parameters, the minimum required sample size was calculated as 5283.

Measures and Covariates

Self-Rated Health

Self-rated health was applied to reflect an individual's health status. Self-rated health was assessed using a visual analog scale. Participants were asked the following question: "Overall, how would you rate your current health status?" They responded by selecting a point on a horizontal scale ranging from 0 to 100, with a higher score indicating a better level of health.

Relative Health Deprivation

We measured relative health deprivation by calculating the Kakwani relative deprivation index using individuals' self-rated health, referring to the research of Kakwani [25]. The first step is to assume *G* as a reference group containing *k* sample sizes and rank the self-rated health levels of individuals in this group in ascending order. Then we can obtain $G = (g_1,g_2,g_3, \ldots ,g_{k-1},g_k)$, which is the distribution vector of self-rated health. Therefore, compared with the self-rated health of individual *i* with that of *j*, the relative health deprivation of individual *i* is as follows:

(1)RD(gj−gi)={gj−giif gj≥gi0if gj<gi

On this basis, the average relative health deprivation for individual *i* can be calculated as follows:

(2)RD(gi)=1kλG(kg+gi+−kg+gi)

where λ_G is the average self-rated health of all samples within the reference group *G*. ***kgi+*** is the number of individuals whose self-rated health exceeds that of individual *i*. λgi+ is the average self-rated health of individuals who have higher self-rated health than *i*.

XSL•FO
RenderX

The reference group for calculating relative health deprivation was defined as all participants within the same geographically stratified sampling group. This allows the deprivation measure to reflect within-region health comparisons.

Short-Video Platform Use

We took "whether to use short video software" as one of the independent variables, with the options containing yes and no. Besides, due to the multidimensional characteristics of digital engagement, years, frequency, daily duration, and purpose diversity of short-video platform use were selected to measure the usage patterns. The assessment methods for the four dimensions are as follows: (1) years of use included "≤1 year," "1-3 years," and "≥3 years," with the assigned values being 1, 2, and 3; (2) frequency of use included "several times a year," "several times a month," "1-4 times a week," and "almost every day," with values ranging from 1 to 4; (3) daily duration of use included: "≤30 minutes," "31-60 minutes," "61-120 minutes," "120-240 minutes," and "≥240 minutes," with values ranging from 1 to 5; and (4) the purpose diversity of use mainly measures the number of intended uses, regarding information acquisition, entertainment, social interaction, creative sharing, and economic acquisition. The options of number of purposes included "0," "1," "2," and "≥3," with values ranging from 1 to 4.

SES

To further analyze the role of socioeconomic factors in the relationship between short-video platform use and health, SES was incorporated into the moderating effect analysis, and it was set at 3 levels: low, middle, and high. The SES level was reported by the participants themselves.

Covariates

We used a series of demographic and socioeconomic factors, social support factors, and health-related factors as covariates. Demographic and socioeconomic factors included age, gender, urban-rural residence, SES, education, and employment status. Social support factors measured marriage, number of persons living together, and number of persons contacted regularly, following the framework proposed by van den Brink et al [27]. Among health-related factors, the Chinese multimorbidity–weighted index was used to control multimorbidity burden, which was calculated according to Hu et al [28]. Moreover, the convenience of obtaining medical services and mental health, including depression (as assessed with 10-item Center for Epidemiologic Studies Depression Scale) [29], anxiety (as assessed with 7-item Generalized Anxiety Disorder) [30], and loneliness (as assessed with University of California at Los Angeles Loneliness Scale) [31], was also controlled. The options included by these variables are shown in Table 1.

Statistical Analysis

First, 2-tailed *t* test and variance analysis were used to analyze the differences in self-rated health and relative health deprivation among different groups. In the second step, propensity score matching analysis with no replacement was performed to control for selection bias based on the probability of short-video platform use (users and nonusers). Afterward, we fitted quantile regression models on the 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, and 90th quantiles of the self-rated health and relative health deprivation distribution to investigate the differences in health between short-video platform users and nonusers. Quantile indicates the location of individuals within the outcome distribution [32]. For example, the individual in the 80th quantile of the conditional distribution has a self-rated health larger than 79% of those with similar short-video platform use behaviors. Quantile regression is a flexible and powerful tool that can assess how health relates to short-video platform use at any quantile distribution. Then, a linear regression model was applied to explore the association of different dimensions of short-video platform use with health, using the full sample. Independent variables consisted of the years, frequency, daily duration, and purpose diversity of use. Finally, we used moderating effect analysis to explore the moderating role of SES between the daily duration of use and health. A series of sensitivity analyses was conducted to test the robustness of the results. Quantile regression was conducted again based on propensity score matching with replacement. In the linear regression model, years, frequency, daily duration, and purpose diversity of use were regarded as continuous variables to be analyzed. Further, we limited the analysis of these dimensions to users of short-video platforms, so as to make the coverage more targeted. The urban-rural residence was used to replace the self-assessed SES to evaluate the moderating role of socioeconomic factors. In all models, coefficients were tested using 2-tailed *t* tests.

Stata 18.0 was used for all analyses. Statistical significance was defined as a 2-sided *P* value less than .05.

Ethical Considerations

Ethics Review and Approval

This study involved human participants and was conducted in accordance with the Declaration of Helsinki. The study protocol was reviewed and approved by the Medical Ethics Committee of Tongji Medical College, Huazhong University of Science and Technology (approval No. [2023] Lun Shen Zi (S193)).

Informed Consent

Prior to participation, all individuals were provided with detailed information about the study's purpose, procedures, potential risks, and benefits. Written informed consent was obtained from each participant.

Privacy and Confidentiality

To protect participant privacy, all collected data were anonymized at the point of collection. Personal identifiers were removed, and each participant was assigned a unique study code. The data were stored on password-protected secure servers accessible only to the authorized research team.

Participant Compensation

Participants did not receive financial compensation for participation. However, small tokens of appreciation (toothpaste of RMB ¥10 [approximately US $1.42]) were provided in accordance with local ethical norms to acknowledge participants' time.

XSL•FO
RenderX

Image and Identifiable Data Protection

This manuscript and its supplementary materials do not contain any images, videos, or personal details that could lead to the identification of an individual participant.

Results

Characteristics of Participants

The final sample consisted of 7725 participants in this study (Figure 1). Characteristics of eligible participants are presented in Table 1. Of the 7725 participants, 4403 (57%) used short-video platforms, 3628 (46.96%) were male, 6492 (84.04%) were married or cohabiting, and 2342 (30.32%) lived in urban areas. Participants who were spousal (n=3966, 90.07% vs n=2526, 76.04%) and had received education (n=6621, 74.93% vs n=5869, 44.13%) were more likely to use short-video platforms. Compared with nonusers of short-video platforms, users had lower scores of multimorbidity-weighted index (2.40, SD 1.96 vs 2.66, SD 2.09), depression (5.28, SD 4.50 vs 6.65, SD 5.02), anxiety (1.82, SD 3.22 vs 2.39, SD 3.71), and loneliness (3.54, SD 1.19 vs 3.90, SD 1.39). The differences in scores of self-rated health and relative health deprivation among different groups are shown in Table 1.

Figure 1. Flow diagram of participant recruitment and data exclusion.

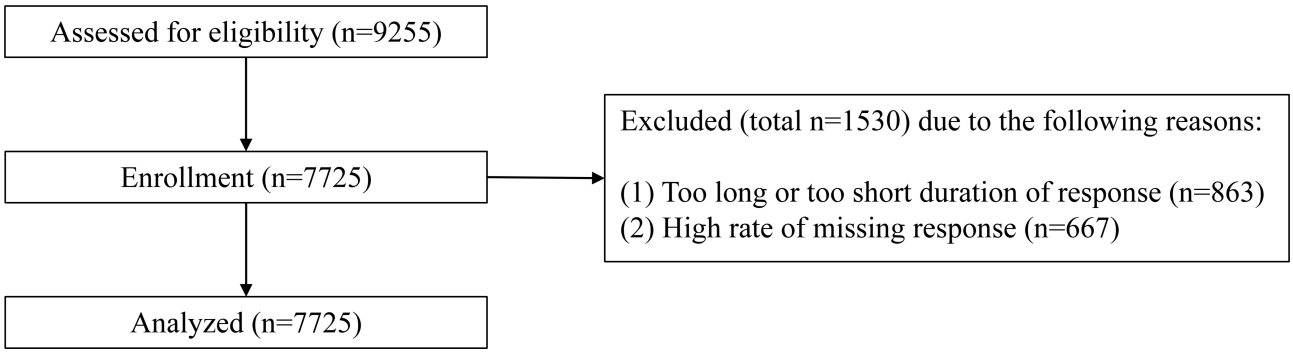

Quantile Regression Results

Figures S1-S2 in Multimedia Appendix 1 show the distribution of self-rated health and relative health deprivation in quantiles. The 2 purple curves fluctuate to a large extent near the gray uniform line, indicating that the scores of self-rated health and relative health deprivation are not normally distributed. So, it is necessary to apply quantile regression. In total, 3322 pairs of users of short-video platforms versus nonusers of short-video platforms were formed based on the propensity matching scores. The characteristics of the participants in the propensity-matched dataset are provided in Table S1 in Multimedia Appendix 1.

The quantile regression results in the propensity-matched dataset are reported in Table 2. Compared with the nonusers of short-video platforms, the self-rated health of users was higher at the 70th to 90th quantile ($P<.05$), and the score of the relative health deprivation of users was lower at the 10th to 30th quantile ($P<.05$). For self-rated health, the correlation was the strongest at the 90th quantile (β=1.931, 95% CI 1.299-2.563; $P<.001$). Compared to the users at the 20th and 30th quantiles, the score of relative health deprivation decreased to a lesser extent among those at the 10th percentile (β=−0.003, 95% CI −0.005 to −0.001; $P<.001$). Moreover, the self-rated health of users is lower than nonusers below the 10th quantile (β=−2.224, 95% CI −3.835 to −0.613; $P=.007$), while no significant association was observed at the 90th quantile of relative health deprivation. There was no significant association at the 20th to 60th quantile of self-rated health and the 40th to 90th quantile of relative health deprivation.

XSL•FO
RenderX

Table . Results for quantile regression analysis of the association between the use of short-video platforms and self-rated health, and relative health deprivation in propensity-matched dataset.

Quantile (reference: nonusers)	Self-rated health[a]		Relative health deprivation[b]	
	Coefficient (95% CI)	*P* value	Coefficient (95% CI)	*P* value
0.1	–2.224 (–3.835 to –0.613)	.007	–0.003 (–0.005 to –0.001)	<.001
0.2	–0.013 (–1.375 to 1.348)	.99	–0.005 (–0.007 to –0.002)	.001
0.3	0.813 (–0.355 to 1.982)	.17	–0.005 (–0.007 to –0.002)	<.001
0.4	0.544 (–0.521 to 1.608)	.32	–0.002 (–0.007 to 0.003)	.38
0.5	0.772 (–0.600 to 2.145)	.27	–0.001 (–0.006 to 0.003)	.58
0.6	0.627 (–0.304 to 1.558)	.19	–0.005 (–0.012 to 0.002)	.15
0.7	0.990 (0.067 to 1.913)	.04	–0.008 (–0.020 to 0.003)	.15
0.8	1.423 (0.600 to 2.246)	.001	–0.002 (–0.019 to 0.014)	.78
0.9	1.931 (1.299 to 2.563)	<.001	0.021 (–0.001 to 0.043)	.07

[a]Adjusted R^2: 0.179; *F*-value: 97.67; *F* significance: <.001.

[b]Adjusted R^2: 0.17; *F*-value: 91.78; *F* significance: <.001

Linear Regression Results

Correlation analysis indicated that there was a correlation among years, frequency, daily duration, and the number of use purposes of short-video platforms usage. The results of the multicollinearity test showed that there was no collinearity among these variables (variance inflation factor<5). The specific results are shown in Tables S2-S3 in Multimedia Appendix 1.

As shown in Table 3, compared with participants who have used short-video platforms for ≤1 year, those who have used them for ≥3 years showed significantly better self-rated health (β=1.970, 95% CI 0.308-3.632; *P*<.05). Yet, after adjusting for the control variables, there was no significant correlation between the years of use and relative health deprivation. There was no significant correlation between frequency of use and self-rated health and relative health deprivation. Participants who spent 61 - 120 minutes and 120 - 240 minutes per day had significantly worse self-rated health (β=−3.385, 95% CI −4.872 to −1.898; *P*<.001; β=−3.038, 95% CI −5.054 to −1.022; *P*<.01) and higher score of relative health deprivation (β=0.035, 95% CI 0.021-0.050; *P*<.001; β=0.034, 95% CI 0.014-0.054; *P*<.01). However, there was no significant relationship between the participants' daily duration of use exceeding 240 minutes and their self-rated health or relative health deprivation. Compared to no purposeful use, using short-video platforms for 2 purposes was associated with better self-rated health (β=6.082, 95% CI 0.250-11.914; *P*<.05) and lower relative health deprivation (β=−0.063, 95% CI −0.120 to −0.005; *P*<.05). The association was stronger for use with ≥3 purposes (self-rated health: β=6.213, 95% CI 0.276-12.151; *P*<.05; relative health deprivation: β=−.068, 95% CI −0.127 to −0.009; *P*<.05).

XSL•FO
RenderX

Table . Association between the usage behaviors of short-video platforms and self-rated health and relative health deprivation.

Variables	Self-rated health		Relative health deprivation	
	Model 1, coefficient (95% CI)	Model 2, coefficient (95% CI)	Model 3, coefficient (95% CI)	Model 4, coefficient (95% CI)
Years of use (reference: ≤1)				
1-3	1.775 (–0.012 to 3.562)	1.156 (–0.493 to 2.805)	–0.011 (–0.029 to 0.006)	–0.006 (–0.023 to 0.010)
≥3	4.116 (2.332 to 5.900)[a]	1.970 (0.308 to 3.632)[b]	–0.032 (–0.050 to –0.015)[a]	–0.014 (–0.031 to 0.002)
Frequency of use (reference: Several times a year)				
Several times a month	–4.718 (–11.378 to 1.941)	–5.033 (–11.13 to 1.074)	0.050 (–0.016 to 0.115)	0.053 (–0.007 to 0.114)
1-4 times a week	–4.399 (–10.485 to 1.687)	–4.673 (–10.255 to 0.907)	0.039 (–0.020 to 0.099)	0.043 (–0.012 to 0.098)
Almost every day	–4.983 (–10.997 to 1.032)	–5.250 (–10.763 to 0.262)	0.049 (–0.010 to 0.108)	0.052 (–0.002 to 0.107)
Daily duration of use (min; reference: ≤30)				
31-60	1.021 (–0.396 to 2.437)	0.046 (–1.259 to 1.350)	–0.008 (–0.022 to 0.006)	0.002 (–0.011 to 0.015)
61-120	–2.451 (–4.067 to –0.834)[c]	–3.385 (–4.872 to –1.898)[a]	0.026 (0.010 to 0.042)[c]	0.035 (0.021 to 0.050)[a]
120-240	–2.856 (–5.055, –0.657)[b]	–3.038 (–5.054 to –1.022)[c]	0.033 (0.011 to 0.054)[c]	0.034 (0.014 to 0.054)[c]
≥240	0.202 (–3.220 to 2.815)	–0.352 (–3.119 to 2.415)	0.008 (–0.022 to 0.038)	0.009 (–0.018 to 0.037)
Numbers of purposes of use (reference: 0)				
1	5.694 (–0.564 to 11.953)	4.970 (–0.773 to 10.712)	–0.055 (–0.117 to 0.006)	–0.051 (–0.108 to 0.006)
2	7.263 (0.910 to 13.616)[b]	6.082 (0.250 to 11.914)[b]	–0.071 (–0.133 to –0.008)[b]	–0.063 (–0.120 to –0.005)[b]
≥3	8.349 (1.890 to 14.807)[b]	6.213 (0.276 to 12.151)[b]	–0.084 (–0.147 to –0.020)[b]	–0.068 (–0.127 to –0.009)[b]
Control variables	No	Yes	No	Yes
Adjusted R^2	0.021	0.18	0.016	0.166
F-value	14.82	66.1	11.21	59.95
F significance	<.001	<.001	<.001	<.001

[a]*P*<.001.
[b]*P*<.05.
[c]*P*<.01.

Moderating Effect Analysis Results

As shown in Figures 2 and 3, SES was a significant moderator of the association between daily duration of use and health (all *P* for interaction <.05). The corresponding coefficients and *P* for interaction can be found in Table S4 in Multimedia Appendix 1. Specifically, compared to the high SES group, participants with middle SES showed significantly lower self-rated health (β=−2.257, 95% CI −3.540 to −0.975; *P*=.001) and higher relative health deprivation (β=0.018, 95% CI 0.005-0.031; *P*=.005). Similarly, the low SES group showed significantly poorer self-rated health (β=−2.022, 95% CI −3.349 to −0.695; *P*=.003) and a higher score of health deprivation (β=0.014, 95% CI 0.001-0.027; *P*=.04).

XSL•FO
RenderX

Figure 2. Results for moderating effect analysis between daily duration of short-video platforms usage and self-rated health. "≤30 minutes," "31-60 minutes," "61-120 minutes," "120-240 minutes," and "≥240 minutes" indicate daily duration of use. Error bars represent 95% CI.

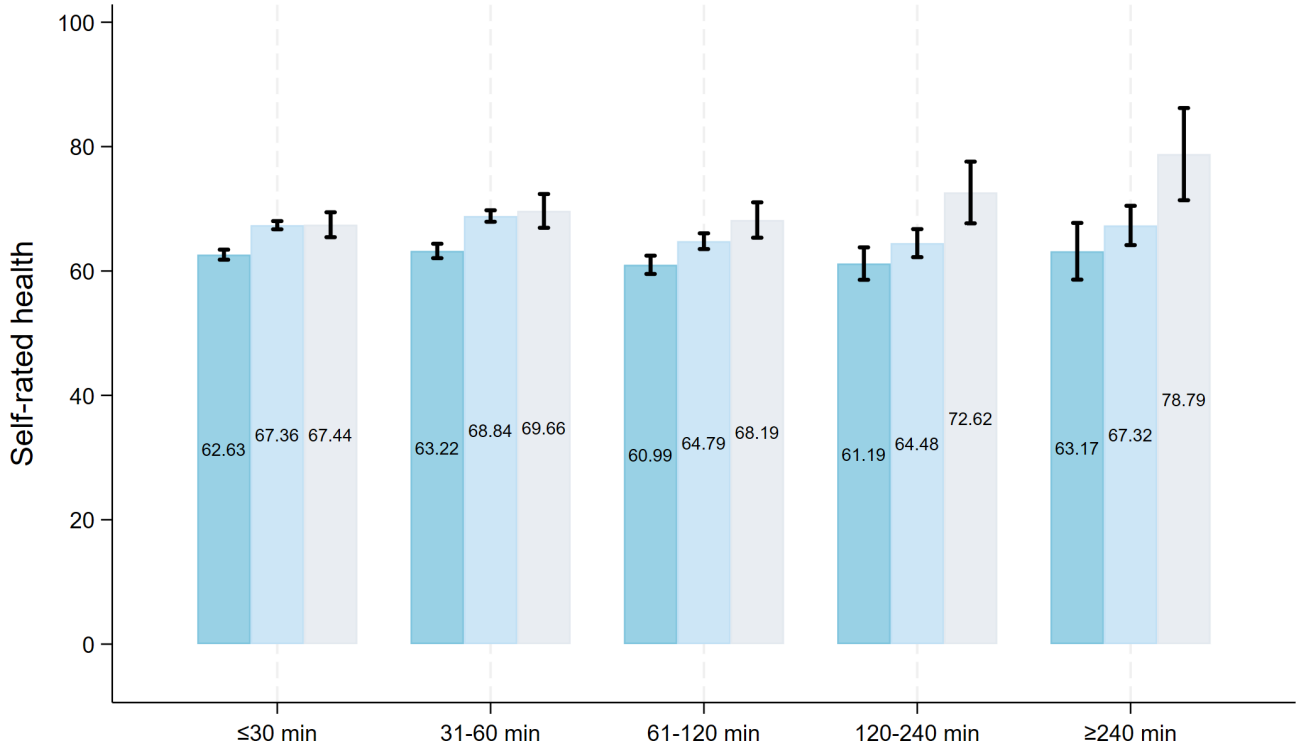

Figure 3. Results for moderating effect analysis between daily duration of short-video platforms usage and relative health deprivation. "≤30 minutes," "31-60 minutes," "61-120 minutes," "120-240 minutes," and "≥240 minutes" indicate daily duration of use. Error bars represent 95% CI.

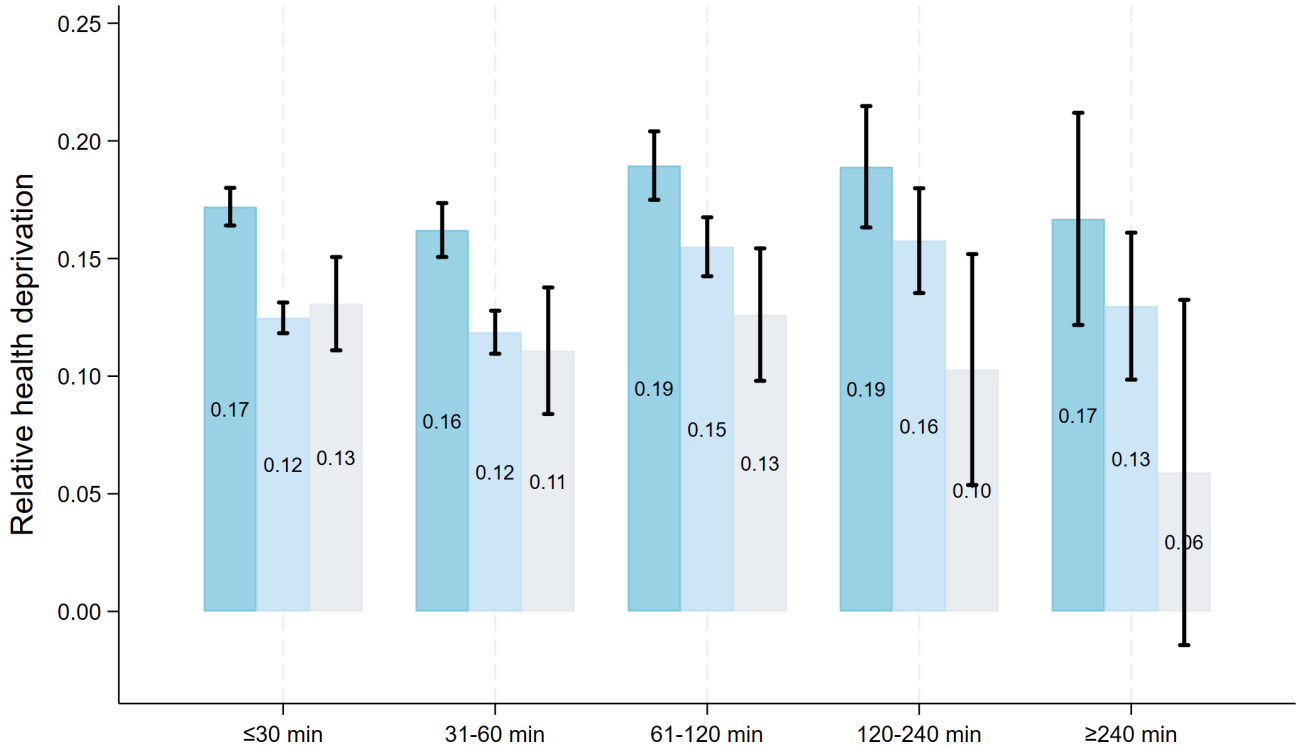

Sensitivity Analysis Results

After changing the propensity score matching method, the conclusions of the quantile regression remained unchanged. Regarding years, frequency, daily duration, and purpose diversity of usage as continuous variables in linear regression, the relevant conclusions are broadly consistent. The analysis results for the users of short-video platforms also found similar results. Moreover, compared to the urban residence group, the rural residence group had significantly lower self-rated health (β=−0.890, 95% CI −1.667 to −.113; *P*=0.03) and higher relative health deprivation (β=0.008, 95% CI 0.000-0.016; *P*=.038). This is also consistent with the previous analysis of the moderating role. All sensitivity analysis results are provided in Tables S5-S8 in Multimedia Appendix 1.

XSL•FO
RenderX

Discussion

Principal Results

This study explored the heterogeneous association between multidimensional short-video platform use and individual health among Chinese populations. The results reveal a series of nuanced findings. Only among the health-advantaged group with relatively high self-rated health (70th-90th quantile) and low relative deprivation (10th-30th quantile), using short-video platforms was associated with better self-rated health and lower relative health deprivation compared to nonuse. Moreover, short-video platform use was linked to worse self-rated health in the least healthy group (10th percentile) yet showed no significant association with health deprivation in the most relatively deprived group (90th percentile). Another finding showed that greater purpose diversity (≥2 purposes) and longer engagement (≥3 y) with short-video platforms were positively associated with better self-rated health. However, frequency of use showed no significant link to health. A daily use duration of 1 to 4 hours was associated with poorer self-rated health and higher relative deprivation, yet no such negative association was found for duration exceeding 4 hours. SES significantly moderated the association between daily use duration and relative health deprivation, with the strength of this association being more pronounced among middle- and low-SES groups than in the high-SES group.

Interpretation

Our findings demonstrate theoretical congruence with the diffusion of innovations framework. This theory posits that technology adoption follows a social gradient, often privileging early adopters who possess greater resources [33]. The health-advantaged group has preexisting health and social resources that are better equipped to navigate digital environments selectively. They can make use of these advantages for information, social connection, and stress relief, thereby further enhancing their well-being [34]. Their lower baseline deprivation may also buffer them against negative social comparison, a known risk of short-video platforms use. Conversely, the significant negative association observed in the least healthy group (10th percentile of self-rated health) suggests potential pathways of harm. Individuals in poor health may be more vulnerable to algorithmic traps and engage in passive consumption of content. Their use may exacerbate sedentary behavior or increase exposure to distressing or misleading health information [35-37]. Another noteworthy finding is that there is no significant correlation between short-video platform use and relative health deprivation among the group with the highest level of health deprivation (10th percentile). It indicates that for individuals who perceive themselves as severely disadvantaged compared to others, engagement with short-video platforms does little to relieve this sense of inequity [38]. Their deprivation likely depends on structural factors, such as the severity of the disease and socioeconomic disadvantage. These cannot be easily relieved by the fragmented social support and information on short-video platforms. For individuals at the 20th to 60th percentiles of self-rated health and at the 40th to 90th percentiles of relative health deprivation, no significant association was observed. This group typically fits the description of the early majority as defined in diffusion of innovations theory [33]. Their habits of use are more likely to be influenced by the inherent characteristics of short videos, which are typically characterized by frequent but brief, passive, and entertainment-oriented consumption [39]. Together, this pattern of use tends to make short videos a relatively insignificant supplementary tool. Its association with cognitive, emotional, and behavioral remains diffuse and thus below the threshold of detectable statistical association in a cross-sectional design. Notably, our finding of a modest reduction in relative health deprivation ($\beta=-0.003$) at the 10th percentile requires careful interpretation. While this change is small at the individual level, its consistent detection at the lower tail of the deprivation distribution holds broader significance. It suggests that even among those who already feel relatively less deprived, short-video platform use may contribute a marginal, yet statistically robust, psychological sense of gain. From a population health perspective, a small shift that is reliably present across a large subgroup can translate into a meaningful aggregate benefit [40].

The differentiated associations observed between behaviors of short-video platform use and health suggest that how and why individuals engage with these platforms may matter more than how often they do so. The positive association of purpose diversity (≥2 purposes) with health underscores the importance of an active, agentic attitude for short-video platforms use. Short-video platforms are engineered for high engagement through infinite scroll and personalized feeds, which typically promote passive viewing [41]. Users who could actively use the platform for information seeking, social interaction, and content creation likely mitigate risks of mindless scrolling [42-44]. The association linked to longer engagement (≥3 y) may reflect digital literacy increases with the accumulation of use time. New users need to spend time exploring and mastering the functions of the short-video platforms, while long-term users may develop more regular and proactive habits. They might be able to use the platform's functions more flexibly and benefit from it [45,46]. No significant correlation between frequency of use and health is observed, which is inconsistent with the previous studies regarding internet usage [10,47,48]. It might be due to the inherent characteristics of the short-video platforms. Unlike traditional media consumption, where each session constitutes a discrete, sustained period of attention, short-video use is characterized by brief and repetitive engagements [49]. Therefore, the frequency of use may not accurately reflect its association with health, as it fails to distinguish between these short usage periods and more prolonged, immersive usage periods. For the daily duration of use, using short-video platforms for 1 to 4 hours per day is associated with poorer health. This might indicate the psychological fatigue caused by excessive use and the adverse physical activities, such as sedentary behavior [5,50,51]. The daily duration of use exceeding 4 hours was not significantly associated with self-rated health and relative health deprivation. This might be related to the user's occupation, pattern of use, and so on. This requires further analysis in subsequent studies.

The moderating role of SES observed in this study also aligns with the innovation diffusion theory and the digital divide theory

XSL•FO
RenderX

[33]. From the perspective of innovation diffusion, high-SES individuals often act as early adopters. Their greater material and cognitive resources facilitate not only access but also more adaptive and beneficial engagement with new technologies, such as short-video platforms. This translates into higher digital health literacy, better self-regulatory capacity, and the ability to use these platforms more selectively [52-54]. Conversely, the theory of the second-level digital divide explains the heightened vulnerability of middle- and low-SES groups [55]. Disparities in digital skills and resources often confine their engagement to more passive, entertainment-focused consumption patterns [56]. This pattern of use, characterized by prolonged but less intentional exposure, makes them more susceptible to the platforms' risks, such as unfavorable social comparison, algorithmic traps, and sedentary behavior [57]. Thus, preexisting socioeconomic inequalities are reproduced and potentially amplified in the digital realm, exacerbating feelings of relative health deprivation among the already disadvantaged.

Limitations

Several limitations of this study should be acknowledged. First, despite the application of propensity score matching to mitigate selection bias, the inherent constraints of cross-sectional observational design preclude definitive causal inferences regarding short-video platform use and health outcomes [24]. Second, while multistage stratified sampling was implemented to ensure regional diversity, potential sampling biases may exist due to nonprobability sampling methods and self-reported participation rates, limiting the generalizability of the findings to the broader Chinese adult population. Third, self-rated health as a subjective measure is susceptible to social desirability bias and cultural interpretation variance, potentially underestimating or overestimating actual health disparities. Fourth, although multiple covariates were statistically controlled, residual confounding effects from unmeasured variables, such as cognitive function and digital literacy, cannot be entirely ruled out, particularly given their demonstrated associations with both health behaviors and technology engagement patterns. Fifth, SES relying on self-report may have measurement error. Future studies would benefit from incorporating objective indicators, such as household income brackets, educational attainment levels, and occupational prestige scores to enhance construct validity. Sixth, the categorization of screen time into discrete intervals limits our ability to model dose-response relationships or detect nonlinear threshold effects within usage categories. This limitation could be addressed through device-based time-use monitoring in subsequent research. Seven, the lack of content type differentiation obscures the understanding of how specific platform features link to health, an issue that could be resolved through qualitative content analysis. Finally, the study focused on short-video platforms and did not capture the potential association between other digital media forms and health. Future longitudinal studies should incorporate specific contents of usage to better identify for whom and under what conditions digital engagement supports health.

Conclusions

This study advances the field by moving beyond the examination of overall associations in the novel context of short-video platforms. It investigates for whom and under what conditions the use of these platforms is linked to health. By exploring multidimensional usage behaviors, it identifies distinct patterns of benefit and risk across population subgroups, providing a nuanced understanding of the relationship between digital engagement and health. Self-rated health and relative health deprivation help to clarify an individual's subjective perception of health, as well as the sense of health inequality derived from social comparisons. This study carries significant associations between short-video platform use and individual health, which are heterogeneous across population subgroups. Policy must therefore shift from ensuring universal access to actively governing for equitable digital well-being. This requires targeted protections for vulnerable groups, especially those in poor health or with lower SES. They are most susceptible to harm from passive or prolonged use. Meanwhile, the observed disparities in relative health deprivation suggest public policy should aim not only to improve material health access but also to intentionally shape digital environments to address both material and perceived health disparities.

Acknowledgments

We thank all the volunteer participants for their patience and participation. The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing. The GAI tool used was DeepSeek-V3.2. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was supported by National Key R&D Program in China (2022YFE0133000), National Nature Science Foundation of China (72374079), Hubei Provincial Department of Science and Technology (JCZRYB202500352). The funders had no involvement in the study design, data collection, data analysis, interpretation of results, or the writing of the manuscript.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to institutional data protection policies and participant confidentiality regulations. The data may be made available from the corresponding author upon reasonable request by qualified researchers for academic purposes.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: YP
Data curation: KZ
Formal analysis: YP
Funding acquisition: ST
Investigation: KZ, YH, YW
Methodology: ST
Project administration: ST
Resources: ST
Software: YP
Supervision: ST
Validation: YW
Visualization: YH
Writing – original draft: YP
Writing – review and editing: CL, CY, YP

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional results.
[DOCX File, 201 KB - jmir_v28i1e86526_app1.docx]

References

1. Altanlar A, Eren İ, Aktuğlu Aktan E. Digital technologies and public life in Istanbul: socio-demographic patterns and urban implications. Cities 2025 Oct;165:106179. [doi: 10.1016/j.cities.2025.106179]
2. Tavares AI. Self-assessed health among older people in Europe and internet use. Int J Med Inform 2020 Sep;141:104240. [doi: 10.1016/j.ijmedinf.2020.104240] [Medline: 32739610]
3. Li X, Liu J, Huang N, Zhao W, He H. Association between internet use and sleep health among middle-aged and older Chinese individuals: nationwide longitudinal study. J Med Internet Res 2025;27:e71030. [doi: 10.2196/71030] [Medline: 40239202]
4. Wen W, Zhang Y, Shi W, Li J. Association between internet use and physical health, mental health, and subjective health in middle-aged and older adults: nationally representative cross-sectional survey in China. J Med Internet Res 2023 Mar 21;25:e40956. [doi: 10.2196/40956] [Medline: 36943368]
5. Choudhury A, Asan O, Alelyani T. Exploring the role of the internet, care quality and communication in shaping mental health: analysis of the Health Information National Trends Survey. IEEE J Biomed Health Inform 2022 Jan;26(1):468-477. [doi: 10.1109/JBHI.2021.3087083] [Medline: 34097623]
6. Ma JQ, Sheng L. Internet use time and mental health among rural adolescents in China: a longitudinal study. J Affect Disord 2023 Sep 15;337(18-26):18-26. [doi: 10.1016/j.jad.2023.05.054] [Medline: 37236271]
7. Tirado-Morueta R, Aguaded-Gómez JI, Ortíz-Sobrino M, Rodríguez-Martín A, Álvarez-Arregui E. Determinants of social gratifications obtained by older adults moderated by public supports for Internet access in Spain. Telemat Inform 2020 Jun;49:101363. [doi: 10.1016/j.tele.2020.101363]
8. Shim H, Ailshire J, Zelinski E, Crimmins E. The Health and Retirement Study: analysis of associations between use of the internet for health information and use of health services at multiple time points. J Med Internet Res 2018 May 25;20(5):e200. [doi: 10.2196/jmir.8203] [Medline: 29802088]
9. Thom RP, Bickham DS, Rich M. Internet use, depression, and anxiety in a healthy adolescent population: prospective cohort study. JMIR Ment Health 2018 May 22;5(2):e44. [doi: 10.2196/mental.8471] [Medline: 29789282]
10. Rana J. 1156 Associations between internet use and overweight/obesity in Nepal: do gender and urbanity modify these associations? Int J Epidemiol 2021 Sep 1;50(Supplement_1). [doi: 10.1093/ije/dyab168.559]
11. Guan JL, Xia SH, Zhao K, et al. Videos in short-video sharing platforms as sources of information on colorectal polyps: cross-sectional content analysis study. J Med Internet Res 2024 Oct 29;26:e51655. [doi: 10.2196/51655] [Medline: 39470708]
12. The 52nd statistical report on internet development in China. : China Internet Network Information Center; 2023 URL: https://www.cnnic.com.cn/IDR/ReportDownloads/202311/P020231121355042476714.pdf [accessed 2026-02-20]
13. Wu J, Peng H, Fan Z, Zhao H. Family guidance weakens an effect of old people's educational attainment on eHealth literacy by internet experience. Innov Aging 2023 Dec 21;7(Supplement_1):953-954. [doi: 10.1093/geroni/igad104.3063]
14. Jiang S, Beaudoin CE. Health literacy and the internet: an exploratory study on the 2013 HINTS survey. Comput Human Behav 2016 May;58:240-248. [doi: 10.1016/j.chb.2016.01.007]

XSL•FO
RenderX

15. Mitsutake S, Shibata A, Ishii K, Oka K. Associations of eHealth literacy with health behavior among adult internet users. J Med Internet Res 2016 Jul 18;18(7):e192. [doi: 10.2196/jmir.5413] [Medline: 27432783]
16. Nakagomi A, Shiba K, Kawachi I, et al. Internet use and subsequent health and well-being in older adults: an outcome-wide analysis. Comput Human Behav 2022 May;130:107156. [doi: 10.1016/j.chb.2021.107156]
17. Szreder E, Krysta K. Internet and mobile phone addiction in individuals aged 15-50 years. Eur Psychiatr 2025 Apr;68(S1):S465-S465. [doi: 10.1192/j.eurpsy.2025.963]
18. Kovačić Petrović Z, Peraica T, Kozarić-Kovačić D, Palavra IR. Internet use and internet-based addictive behaviours during coronavirus pandemic. Curr Opin Psychiatry 2022 Sep 1;35(5):324-331. [doi: 10.1097/YCO.0000000000000804] [Medline: 35855504]
19. Mitchell ME, Lebow JR, Uribe R, Grathouse H, Shoger W. Internet use, happiness, social support and introversion: a more fine grained analysis of person variables and internet activity. Comput Human Behav 2011 Sep;27(5):1857-1861. [doi: 10.1016/j.chb.2011.04.008]
20. Zhao H, Peng HM, Fan ZY, Wu JX. Different types of Internet use and social adaptation among older adults. Innov Aging 2023 Dec 21;7(Supplement_1):669. [doi: 10.1093/geroni/igad104.2175]
21. Hernández C, Cottin M, Parada F, et al. Watching the world from my screen: a longitudinal evaluation of the influence of a problematic use of the internet on depressive symptomatology. Comput Human Behav 2022 Jan;126:106995. [doi: 10.1016/j.chb.2021.106995]
22. Dahlgren G, Whitehead M. The Dahlgren-Whitehead model of health determinants: 30 years on and still chasing rainbows. Public Health 2021 Oct;199:20-24. [doi: 10.1016/j.puhe.2021.08.009] [Medline: 34534885]
23. Gann B. Combating digital health inequality in the time of coronavirus. J Consum Health Internet 2020 Jul 2;24(3):278-284. [doi: 10.1080/15398285.2020.1791670]
24. Garbarski D. Research in and prospects for the measurement of health using self-rated health. Public Opin Q 2016;80(4):977-997. [doi: 10.1093/poq/nfw033] [Medline: 27833212]
25. Kakwani N. The relative deprivation curve and its applications. J Bus Econ Stat 1984 Oct;2(4):384-394. [doi: 10.1080/07350015.1984.10509412]
26. Levitt HM, Bamberg M, Creswell JW, Frost DM, Josselson R, Suárez-Orozco C. Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):26-46. [doi: 10.1037/amp0000151] [Medline: 29345485]
27. van den Brink RHS, Schutter N, Hanssen DJC, et al. Prognostic significance of social network, social support and loneliness for course of major depressive disorder in adulthood and old age. Epidemiol Psychiatr Sci 2018 Jun;27(3):266-277. [doi: 10.1017/S2045796017000014] [Medline: 28183368]
28. Hu WH, Liu YY, Yang CH, et al. Developing and validating a Chinese multimorbidity-weighted index for middle-aged and older community-dwelling individuals. Age Ageing 2022 Feb 2;51(2):afab274. [doi: 10.1093/ageing/afab274] [Medline: 35211718]
29. Andresen EM, Malmgren JA, Carter WB, Patrick DL. Screening for depression in well older adults: evaluation of a short form of the CES-D. Am J Prev Med 1994 Mar;10(2):77-84. [doi: 10.1016/S0749-3797(18)30622-6]
30. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
31. Russell D, Peplau LA, Cutrona CE. The revised UCLA Loneliness Scale: concurrent and discriminant validity evidence. J Pers Soc Psychol 1980 Sep;39(3):472-480. [doi: 10.1037//0022-3514.39.3.472] [Medline: 7431205]
32. Koenker R, Bassett G. Regression quantiles. Econometrica 1978 Jan;46(1):33. [doi: 10.2307/1913643]
33. Rogers EM. Diffusion of Innovations, 5th edition: Free Press; 2003. URL: https://www.simonandschuster.com/books/Diffusion-of-Innovations-5th-Edition/Everett-M-Rogers/9780743222099?utm_source=chatgpt.com [accessed 2026-02-20]
34. Rademakers J, Heijmans M. Beyond reading and understanding: health literacy as the capacity to act. Int J Environ Res Public Health 2018 Aug 7;15(8):1676. [doi: 10.3390/ijerph15081676] [Medline: 30087254]
35. Theodoratou M, Patiri I, Katsarou D, et al. The link between problematic internet use, physical activity, and mental health: implications for depression and anxiety. Eur Psychiatr 2025 Apr;68(S1):S806-S807. [doi: 10.1192/j.eurpsy.2025.1641]
36. Cline RJ, Haynes KM. Consumer health information seeking on the Internet: the state of the art. Health Educ Res 2001 Dec;16(6):671-692. [doi: 10.1093/her/16.6.671] [Medline: 11780707]
37. Ang S, Lim E, Malhotra R. Digital independence: health-related difficulty in internet use and how it affects older adults' quality of life. Innov Aging 2019 Nov 8;3(Supplement_1):S442-S443. [doi: 10.1093/geroni/igz038.1661]
38. McCloud RF, Okechukwu CA, Sorensen G, Viswanath K. Beyond access: barriers to internet health information seeking among the urban poor. J Am Med Inform Assoc 2016 Nov;23(6):1053-1059. [doi: 10.1093/jamia/ocv204] [Medline: 27206459]
39. Pinho C, Franco M, Mendes L. Application of innovation diffusion theory to the E-learning process: higher education context. Educ Inf Technol 2021 Jan;26(1):421-440. [doi: 10.1007/s10639-020-10269-2]
40. Thompson C. Rose's prevention paradox. J Applied Philosophy 2018 May;35(2):242-256. [doi: 10.1111/japp.12177]

XSL•FO
RenderX

41. Xu J, Ismail AB, Jamal JB. The impact of mobile short video platform characteristics on users' continuous use behavior in China: a stimulus-organism-response perspective [Article in Russian]. Multidiscip Sci J 2024 Jul;7(1):2025019. [doi: 10.31893/multiscience.2025019]
42. Ghweeba M, Lindenmeyer A, Shishi S, Abbas M, Waheed A, Amer S. What predicts online health information-seeking behavior among Egyptian adults? A cross-sectional study. J Med Internet Res 2017 Jun 22;19(6):e216. [doi: 10.2196/jmir.6855] [Medline: 28642216]
43. Vošner HB, Bobek S, Kokol P, Krečič MJ. Attitudes of active older Internet users towards online social networking. Comput Human Behav 2016 Feb;55:230-241. [doi: 10.1016/j.chb.2015.09.014]
44. Lee S, Kim BG. User, system, and social related factors affecting perceived usefulness for continuance usage intention of mobile apps. Int J Mob Commun 2021;19(2):190-217. [doi: 10.1504/IJMC.2021.113457]
45. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
46. Arcury TA, Sandberg JC, Melius KP, et al. Older adult internet use and eHealth literacy. J Appl Gerontol 2020 Feb;39(2):141-150. [doi: 10.1177/0733464818807468] [Medline: 30353776]
47. Falk Erhag H, Ahlner F, Rydberg Sterner T, Skoog I, Bergström A. Internet use and self-rated health among Swedish 70-year-olds: a cross-sectional study. BMC Geriatr 2019 Dec 23;19(1):365. [doi: 10.1186/s12877-019-1392-8] [Medline: 31870313]
48. Luo Y, Yip PSF, Zhang Q. Positive association between Internet use and mental health among adults aged ≥50 years in 23 countries. Nat Hum Behav 2025 Jan;9(1):90-100. [doi: 10.1038/s41562-024-02048-7] [Medline: 39558112]
49. Yang Y, Liu RD, Ding Y, Lin J, Ding Z, Yang X. Time distortion for short-form video users. Comput Human Behav 2024 Feb;151:108009. [doi: 10.1016/j.chb.2023.108009]
50. Kraut R, Patterson M, Lundmark V, Kiesler S, Mukopadhyay T, Scherlis W. Internet paradox. A social technology that reduces social involvement and psychological well-being? Am Psychol 1998 Sep;53(9):1017-1031. [doi: 10.1037//0003-066x.53.9.1017] [Medline: 9841579]
51. Bessière K, Pressman S, Kiesler S, Kraut R. Effects of internet use on health and depression: a longitudinal study. J Med Internet Res 2010 Feb 28;12(1):e6. [doi: 10.2196/jmir.1149] [Medline: 20228047]
52. Anrijs S, Mariën I, De Marez L, Ponnet K. Excluded from essential internet services: examining associations between digital exclusion, socio-economic resources and internet resources. Technol Soc 2023 May;73:102211. [doi: 10.1016/j.techsoc.2023.102211]
53. Kobul MK. Socioeconomic status influences Turkish digital natives' internet use habitus. Behav Inf Technol 2023 Apr 4;42(5):624-642. [doi: 10.1080/0144929X.2022.2034970]
54. Shin YM. Internet addiction in children and adolescents. In: Internet Addiction in Children and Adolescents: Risk Factors, Assessment, and Treatment: Springer Publishing Company; 2017:29-45. [doi: 10.1891/9780826133731.0002]
55. Van Dijk J. The Digital Divide: Polity Press; 2019. URL: https://www.researchgate.net/publication/336775102_The_Digital_Divide [accessed 2026-02-20]
56. Ndasauka Y, Kayange GM. Introduction: making a case for addiction in Africa. In: Ndasauka Y, Kayange GM, editors. Addiction in South and East Africa: Interdisciplinary Approaches: Palgrave Macmillan Cham; 2019:1-6. [doi: 10.1007/978-3-030-13593-5_1]
57. Scheerder A, van Deursen A, van Dijk J. Determinants of Internet skills, uses and outcomes. A systematic review of the second- and third-level digital divide. Telemat Inform 2017 Dec;34(8):1607-1624. [doi: 10.1016/j.tele.2017.07.007]

Abbreviations

SES: socioeconomic status

Edited by S Brini; submitted 26.Oct.2025; peer-reviewed by A Famotire, B Najand, PA Okooboh, WS Babalola; accepted 29.Jan.2026; published 12.Mar.2026.

Please cite as:
Pan Y, Zhang K, Wei Y, Huang Y, Long C, Yang C, Tang S
Associations Between Short-Video Platform Use and Health Across Health Distribution and Usage Behaviors in China: Cross-Sectional Questionnaire Study
J Med Internet Res 2026;28:e86526
URL: https://www.jmir.org/2026/1/e86526
doi:10.2196/86526

XSL•FO
RenderX

© Yangyang Pan, Kangkang Zhang, Yilin Wei, Yangzhen Huang, Chengxu Long, Chenxin Yang, Shangfeng Tang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Health Equity Analysis of Awareness and Use of GetCheckedOnline, British Columbia's Digital Intervention for Sexually Transmitted and Blood-Borne Infection Testing in 5 Urban, Suburban, and Rural Communities: Cross-Sectional Survey Study

Rodrigo Sierra Rosales[1,2], MSc, MD; Aidan Ablona[3], MPH; Hsiu-Ju Chang[2], MSc; Devon Haag[2], MSc; Heather Pedersen[2], MPH; Catherine Worthington[4], PhD; Daniel Grace[5], PhD; Rod Knight[6], PhD; Devon Greyson[1], PhD; Mark Gilbert[1,2], MD

[1]School of Population and Public Health, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada
[2]Clinical Prevention Services, BC Centre for Disease Control, 655 West 12th Avenue, Vancouver, BC, Canada
[3]Toronto Public Health, Toronto, ON, Canada
[4]School of Public Health & Public Policy, University of Victoria, Victoria, BC, Canada
[5]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
[6]École de Santé Publique, Université de Montréal, Montréal, QC, Canada

Corresponding Author:
Mark Gilbert, MD
School of Population and Public Health, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

Abstract

Background: Digital sexually transmitted and blood-borne infection (STBBIs) testing services are used to improve testing access, but might replicate existing social inequities. Previous research has shown that the digital STBBI testing service *GetCheckedOnline* has improved access to testing in British Columbia (BC), Canada. As part of the program's continuous evaluation, we examined awareness and use of the service in 5 urban, suburban, and rural communities where the program has expanded.

Objective: This study aimed to determine if social location is associated with differences in awareness and use of the service in 5 communities outside Vancouver, BC.

Methods: From July to September 2022, we conducted a cross-sectional survey recruiting (in-person and online) sexually active people aged 16 years or older in 5 urban, suburban, and rural communities where *GetCheckedOnline* had sample collection sites available at the time. We examined differences in awareness and use by age, gender identity, sexual identity, race/ethnicity, education, and income using logistic regression models informed by the Health Equity Measurement Framework.

Results: Of the 1658 participants (n=1058, 63.8% in-person and n=600, 36.2% online), 35.3% (586/1658) were aware of *GetCheckedOnline* and 19.5% (324/1658) had used it. Awareness and use were lower in the first and last age quartiles compared to the second quartile (>38 years: awareness odds ratio [OR] 0.23, 95% CI 0.17 - 0.32; use OR 0.19, 95% CI 0.12 - 0.28; <25 years: awareness OR 0.39, 95% CI 0.28 - 0.53; use OR 0.28, 95% CI 0.18 - 0.41). Awareness and use were also lower in the lowest income group compared to the highest (awareness OR 0.39, 95% CI 0.24 - 0.65; use OR 0.36, 95% CI 0.20 - 0.65). Awareness and use were higher among genderfluid, genderqueer, and nonbinary participants compared to men (awareness OR 2.27, 95% CI 1.63 - 3.18; use OR 1.97, 95% CI 1.36 - 2.84), transgender compared to cisgender participants (awareness OR 2.17, 95% CI 1.54 - 3.06; use OR 2.15, 95% CI 0.46 - 3.13), and nonheterosexual compared to heterosexual participants (awareness OR 2.37, 95% CI 1.89 - 2.97; use OR 2.53, 95% CI 1.91 - 3.38). People of color had higher awareness and use vs White participants (awareness OR 1.74, 95% CI 1.34 - 2.26; use OR 2.01, 95% CI 1.48 - 2.72). Indigenous participants had higher awareness than White participants (OR 1.65, 95% CI 1.19 - 2.20) but no difference in use. Women had similar awareness but lower use compared to men (OR 0.68, 95% CI 0.50 - 0.92).

Conclusions: *GetCheckedOnline* is an equitable means of access to STBBI testing for some but not all equity-owed groups in BC. Further adaptations should consider factors such as differences in material circumstances to improve its accessibility for all.

(J Med Internet Res 2026;28:e78561) doi:10.2196/78561

XSL•FO
RenderX

KEYWORDS

testing; health equity; digital health; sexually transmitted infections; human immunodeficiency virus (HIV)

Introduction

Digital interventions for sexually transmitted and blood-borne infections (STBBIs) testing have been established as a strategy to improve access to testing services [1]. These interventions have been designed using diverse service models to provide access to STBBI testing without needing to visit a health care provider [2]. Compared to services offered at clinics, these interventions are expected to provide lower-cost and more convenient testing options [3]. However, after initial general optimism towards digital interventions as a means to equitably improve health care access, equity concerns have been raised over the past decade regarding the performance of these interventions in the real-world setting [4].

Increasing evidence has suggested that digital interventions can potentially reinforce and even create inequities mediated by complex interactions between intervention characteristics and social and structural determinants of health [5,6]. For example, social and structural determinants affect access to the internet and the affordability of technologies. As such, implementation research on digital STBBI testing interventions has increasingly focused on digital health equity, with reports finding differential use and benefit of these interventions according to systems generating social stratification [7]. Studies reporting differential effects have revealed that age, gender, sexual identity, race and ethnicity, educational attainment, and income can be associated with the uptake of digital STBBI testing services [8-20].

Overall patterns show how uptake is generally higher among women, White people with higher socioeconomic status, urban residents, and heterosexual people [7]. However, most studies are based on program statistics of use and do not collect data within communities. Relying exclusively on program use statistics also does not allow for evaluating intermediate implementation outcomes, such as program awareness, which are essential to understanding uptake patterns (ie, to understand the extent to which not using the service is linked to unawareness) [21].

GetCheckedOnline is a free online tool that facilitates STBBI testing by allowing users to generate a laboratory requisition based on their reported sexual activity without visiting a health care provider in British Columbia (BC) [2]. Negative results are available through the website, while those testing positive are contacted by a nurse who connects them to appropriate treatment and care and supports partner notification. It was launched in Vancouver in September 2014, and research suggests *GetCheckedOnline* increases testing frequency and is used by populations facing testing barriers [3,11]. For example, a survey of gay and bisexual men suggested that the service provided a culturally appropriate and safe service to those who had not discussed their sexual identity with providers [22]. *GetCheckedOnline* has since expanded to other, less populated, urban, suburban, and rural communities where implementation outcomes may differ from Vancouver.

In this study, we conducted an analysis of awareness and use of *GetCheckedOnline* in communities outside Vancouver where *GetCheckedOnline* is available. Our objective was to determine if sociodemographic characteristics as markers of social location—including age, gender identity, sexual identity, race/ethnicity, educational attainment, and income—were associated with differences in *GetCheckedOnline* awareness and use. We hypothesized that such differences existed and that higher awareness and use would be found for the most privileged group of each characteristic. We hoped the findings from this survey would identify groups where further promotion or adaptation of the service may be needed.

Methods

Theoretical Framework

Our study design was informed by the Health Equity Measurement Framework, which emphasizes the role of social stratification processes at the root of inequities and avoids inferring causal relationships as inherited by individuals [23]. Sociodemographic characteristics were used as markers of social location resulting from social stratification, which supported interpretations of *GetCheckedOnline* adoption in different groups.

Study Design and Setting

We conducted a survey to evaluate awareness and use of *GetCheckedOnline* in communities where the service expanded with laboratory partners outside Vancouver to other urban, suburban, and rural communities in the province: Victoria, Langford, Kamloops, and Nelson in 2016, Kimberly in 2019, and Maple Ridge in 2020 (see map in Multimedia Appendix 1). The survey instrument included questions about *GetCheckedOnline*, participants' experiences with STBBI testing, participants' sexual health experiences, sociodemographic characteristics, and participants' use of technology. The survey was piloted with 17 people from across BC, including 9 members of our community advisory board. We requested feedback on the questionnaire's comprehension, readability, and flow and revised it accordingly.

Recruitment and Data Collection

Recruitment took place from July to September 2022 using in-person and online versions of the survey (to maximize recruitment of people with lower digital literacy or access to technology). Our inclusion criteria were residents of BC, 16 years of age or older, sexually active (defined as oral, vaginal, or anal sex with ≥1 partner in the past year), able to complete the survey in English, and not having completed the survey before. Public health and community organization STBBI leads were identified, and peer researchers were hired for each community. For in-person recruitment, these leads and peer researchers identified common gathering places and venues likely to include individuals facing barriers to accessing STBBI testing. These included 1-time and recurrent community events (eg, Pride festivals and markets) and fixed venues (eg,

XSL•FO
RenderX

recreational center). At each venue, peer researchers approached potential participants, described the inclusion criteria, and obtained verbal informed consent with participants self-assessing their eligibility and self-administering the paper survey. Online recruitment occurred through community leaders and local agencies circulating study recruitment information by email, social media (eg, Facebook), and peer researchers putting up posters in venues with QR codes directing to the online survey. Additional recruitment occurred through geotargeted online advertising and promotion in online venues relevant to each community (eg, online forums). The online survey was administered through REDCap (Research Electronic Data Capture; Vanderbilt University) and included measures to prevent automated responses or bots (eg, requiring manual entry of a passphrase included in the survey). The survey was designed to take 10-15 minutes to complete. In-person completion times were not recorded. For the online survey, the median completion time was 12 (IQR 8-20) minutes.

Study Variables

The implementation outcomes for our study were self-reported awareness and use of *GetCheckedOnline*. Awareness was assessed with the question, "With *GetCheckedOnline*, you can get tested for sexually transmitted infections by printing a lab form or downloading an electronic version on your phone, that you then take to a lab, and then get your results online or by phone. Before today, did you know about *GetCheckedOnline*?" (responses: yes/no). Use was assessed by asking, "Have you been tested through *GetCheckedOnline*?" (responses: yes/no/not sure). The independent equity-relevant sociodemographic variables examined in relation to these outcomes were age, gender identity, sexual identity, race/ethnicity, educational attainment, and self-reported annual income. For annual income, a currency exchange rate of CAD $1=US $0.78 is applicable.

Data Management

Paper surveys were entered into a database using a double-entry procedure by 2 independent researchers, with discrepancies resolved by a third independent researcher (question discrepancy rate ranged between 0% and 4.9%). In-person and online survey data were combined. Participants reporting age <16 years, 0 sexual partners in the past year, and living in locations outside BC were excluded. Online participants with missing or incorrect passphrases or discrepancies in start and completion times inconsistent with survey time zones were considered likely to be bots or ineligible participants and excluded.

Statistical Analyses

We modeled awareness and use of *GetCheckedOnline* using logistic regression models to estimate the effect of each of our variables of interest. To build these models, we categorized numeric variables, built mutually exclusive categories for multiple-response survey questions, and selected reference groups. Under our hypothesis, those likely to experience the least barriers to digital STBBI testing services, such as *GetCheckedOnline*, were considered the most privileged group and served as the reference for comparisons. Age categories were based on sample quartiles (<25 years, 25-29 years [reference], 30 - 37 years, and >38 years). While older age is typically associated with a privileged social location, we selected the 25 - 29 years age group as the reference, as an age range favored by assumptions of digital health literacy (positioning younger people as "digital natives") and targeting of technology marketing [24]. Gender was studied at 2 levels: the first (dimension 1: gender identity) included the categories of men only (reference), women only, and those identifying as genderfluid, genderqueer, agender, nonbinary, or selecting multiple gender identities (GGANB+); the second level (dimension 2: gender modality [25]) included cisgender (reference) and transgender categories. Sexual identity included those identifying as heterosexual only (reference) and lesbian, gay, bisexual, or other nonexclusively heterosexual identities (LGB+). Race/ethnicity included those identifying as White only (reference), Indigenous, and non-Indigenous racialized people (people of color). Additional details of the question to collect this data and how categories were built are included in Multimedia Appendix 1. Education (graduate degree; reference) and income (≥$80,000; reference) did not require further categorization.

Informed by the Health Equity Measurement Framework and the conceptualization of sociodemographic characteristics as proxy variables for multiple social stratification processes [23], we built the directed acyclic graph in Figure 1. We used the DAGitty software (Johannes Textor, Institute for Computing and Information Sciences, Radboud University) to identify the minimal sufficient adjustment set for each of our exposures of interest [26,27]. We built variable-specific models to obtain total direct effects for each exposure [28]. Missing responses were excluded from the analyses. Sensitivity analyses included the assessment of awareness and use distribution comparing those with complete vs missing sociodemographic information using the Fisher exact test to test for differences and identify the potential effects of their exclusion. We report estimated total direct effects through odds ratios (ORs) or adjusted odds ratios (aORs) with 95% CIs. All statistical analyses were conducted using R software (version 4.1.3; R Foundation for Statistical Computing) and its *tidyverse* package [29,30].

XSL•FO
RenderX

Figure 1. Directed acyclic graph depicting the hypothesized causal pathways between age, gender and sexual identities, race/ethnicity, educational attainment, income, and awareness and use of *GetCheckedOnline*.

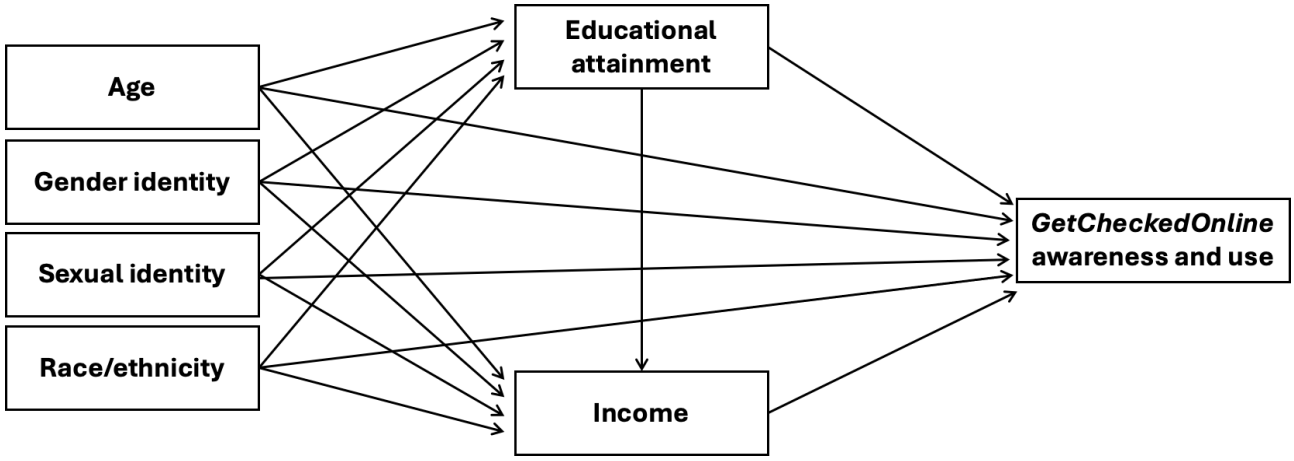

Ethical Considerations

The informed consent procedure was carried out by peer researchers at each venue, who explained the study and obtained verbal consent before handing out surveys to participants. Online and in-person participants completing the survey could enter a draw for 1 of 5 CAD $100 (US $78) VISA gift cards by providing contact information that was not linked to their responses. Participants recruited in person were also offered *GetCheckedOnline* branded promotional materials. Paper survey responses were digitized by 2 core team members and kept with restricted access at the BC Centre for Disease Control. Compiled electronic data from paper and online surveys was stored in secure servers at the BC Centre for Disease Control Central Data Repository. The execution and analysis of the *GetCheckedOnline* community survey were examined and approved by the Behavioral Research Ethics Board of the University of British Columbia (Vancouver, BC, Canada) under the *GetCheckedOnline* Implementation Science Grant (ID H18-00437).

Results

Overall, 2720 individuals responded to the survey (n=1205, 44.3% in-person and n=1515, 55.7% online). The final sample included 1658 individual participants (n=1058, 63.8% in-person and n=600, 36.2% online; Figure 2). The median age of participants was 30 (IQR 25-38) years (Table 1). Almost half of the participants identified only as a woman (784/1658, 47.3%) and 27.7% (460/1658) only as a man. Three-quarters of the sample identified as cisgender (1246/1658, 75.2%) and 9.1% (151/1658) identified as transgender. More than half of participants identified only as White (945/1658, 57%), followed by Indigenous (209/1658, 12.6%). A total of 39.7% (656/1658) of the sample identified only as heterosexual. Any form of postsecondary education was most common (1062/1658, 64%), and 23.6% (392/1658) reported a personal annual income of less than $22,000 in 2021. More than half of participants (1006/1658, 59.7%) had experienced at least 1 barrier to STBBI testing in the past year (Table 2). Details of participants' sociodemographic characteristics before mutually exclusive group assignments are included in Multimedia Appendix 1.

Overall, 35.3% (586/1658) of participants were aware of *GetCheckedOnline*, and 19.5% (324/1658) had used the service for STBBI testing. Among those aware (n=586), 55.3% (n=324) had used the service. All independent variables showed statistically significant associations with awareness and use of *GetCheckedOnline* (Table 3). Age and income showed differences in awareness and use, favoring the reference group. For age, when compared to the 25 - 29 years group, the remaining age quartiles—particularly the oldest and youngest quartiles—had lower odds of awareness and use (>38 years: awareness OR 0.23, 95% CI 0.17 - 0.32; use OR 0.19, 95% CI 0.12 - 0.28; <25 years: awareness OR 0.39, 95% CI 0.28 - 0.53; use OR 0.28, 95% CI 0.18 - 0.41). For income, adjusted for age, gender, sexual identity, race/ethnicity, and education, 2 groups showed statistically significant differences. Compared to the highest income group, those with an income of more than $20,000/year had decreased odds of awareness (aOR 0.39, 95% CI 0.24 - 0.65) and use of *GetCheckedOnline* (use aOR 0.36, 95% CI 0.20 - 0.65). Compared to the same reference, those with an income between $20,000 and $40,000 per year had similar odds of awareness but lower odds of use (aOR 0.53, 95% CI 0.30 - 0.94).

XSL•FO
RenderX

Figure 2. Flowchart of participants included in the analysis of awareness and use of *GetCheckedOnline* in 5 communities outside Vancouver, British Columbia (BC). FSA: Forward Sortation Area of postal code; MT: Mountain Time; PST: Pacific Standard Time.

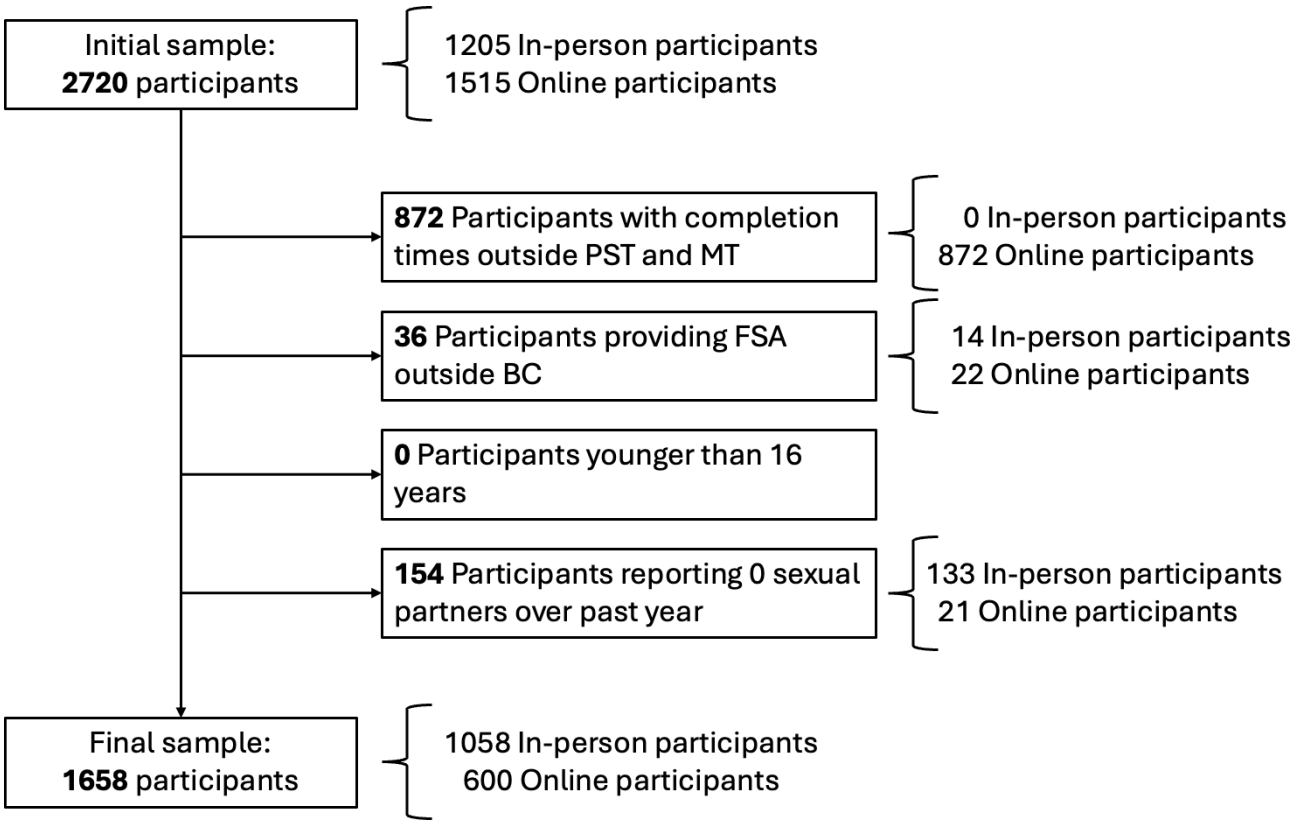

XSL•FO
RenderX

Table . Sample description and awareness and use of *GetCheckedOnline* by subgroups of participants recruited in 5 communities in British Columbia where the service is available (n=1658).

Variable and subgroup	Sample size, n (%)	Awareness of *GetCheckedOnline*, n/N (%)	Use of *GetCheckedOnline*, n/N (%)
Age			
Q[a]1: <25 years	352 (21.2)	102/352 (29)	40/352 (11.4)
Q2: 25 - 29 years	352 (21.2)	182/352 (51.7)	113/352 (32.1)
Q3: 30 - 38 years	368 (22.2)	146/368 (39.7)	95/368 (25.8)
Q4: >38 years	403 (24.3)	79/403 (19.6)	33/403 (8.2)
Missing	183 (11)	—[b]	—
Gender—dimension 1 (gender identity)			
Women (only)	784 (47.3)	233/784 (29.7)	114/784 (14.5)
Men (only)	460 (27.7)	159/460 (34.6)	93/460 (20.2)
Genderqueer, genderfluid, agender, nonbinary, and those with multiple gender identities	213 (12.8)	115/213 (54)	70/213 (28.9)
Missing	201 (12.1)	—	—
Gender—dimension 2 (gender modality)			
Cisgender	1246 (75.2)	401/1246 (32.2)	212/1246 (17)
Transgender	151 (9.1)	76/151 (50.3)	45/151 (29.8)
Missing	261 (15.7)	—	—
Sexual Identity			
Heterosexual (only)	656 (39.6)	161/656 (24.5)	77/656 (11.7)
Lesbian, gay, bisexual, and other nonexclusively heterosexual identities	818 (49.3)	354/818 (43.3)	205/818 (25.1)
Missing	184 (11.1)	—	—
Race/ethnicity			
Indigenous	209 (12.6)	88/209 (42.1)	43/209 (20.6)
People of color	311 (18.8)	133/311 (42.8)	86/311 (27.7)
White (only)	945 (57)	288/945 (30.5)	152/945 (16.1)
Missing	193 (11.6)	—	—
Educational attainment			
Less than high school	70 (4.2)	27/70 (38.6)	14/70 (20)
High school	350 (21.1)	100/350 (28.6)	50/350 (14.3)
Postsecondary certificate	517 (31.2)	175/517 (33.8)	98/517 (19)
Bachelor's degree	395 (23.8)	165/395 (41.8)	95/395 (24.1)
Graduate degree	150 (9)	46/150 (30.7)	24/150 (16)
Missing	176 (10.6)	—	—
Income			
Less than $20,000	392 (23.6)	89/392 (22.7)	41/392 (10.5)
20,000 to 40,000	319 (19.2)	124/319 (38.9)	62/319 (19.4)
40,000 to 60,000	284 (17.1)	123/284 (43.3)	70/284 (24.6)
60,000 to 80,000	179 (10.8)	85/179 (47.5)	59/179 (33)

XSL•FO
RenderX

Variable and subgroup	Sample size, n (%)	Awareness of *GetCheckedOnline*, n/N (%)	Use of *GetCheckedOnline*, n/N (%)
$80,000 and up	175 (10.6)	57/175 (32.6)	34/175 (19.4)
Missing	309 (18.6)	—	—

[a]Quartile.
[b]Not available.

Table . Usual testing frequency and barriers to provider-based testing for sexually transmitted and blood-borne infections (STBBIs) in a sample of participants from 5 communities in British Columbia where the service is available (n=1658).

Sample descriptor	Value, n (%)
Usual STBBI testing frequency	
Never had an STBBI test	319 (19.2)
Have only tested once	179 (10.8)
No set pattern (including when having symptoms or new sexual partners)	318 (19.2)
Every few years	226 (13.6)
Once a year	227 (13.7)
Twice a year	92 (5.5)
A few times per year (eg, every 3 - 4 months)	161 (9.7)
Once a month	18 (1.1)
Missing responses	118 (7.1)
Barriers to provider-based STBBI testing (past year)	
The wait was too long	332 (20)
Needed an appointment	312 (18.8)
Didn’t know where to go	242 (14.6)
The clinic wasn’t open when I could test	212 (12.8)
Concerned about privacy	183 (11)
The clinic was too far away	164 (9.9)
The usual place closed because of COVID-19	149 (9)
Didn’t want to see a doctor or nurse	141 (8.5)
Worried about getting COVID-19 at a testing site	111 (6.7)
Other nonlisted barriers	141 (8.5)
Did not experience barriers to testing in the past year	492 (29.7)

XSL•FO
RenderX

Table . Associations between sociodemographic characteristics and awareness and use of *GetCheckedOnline* in a sample of participants recruited in 5 communities in British Columbia where the service is available (n=1658).

Sociodemographic characteristic	Awareness, estimate (95% CI)	Use, estimate (95% CI)
Age[a]		
Q[b]1: <25 years	0.39 (0.28 - 0.53)[c]	0.28 (0.18 - 0.41)[c]
Q2: 25 - 29 years	Reference	Reference
Q3: 30 - 37 years	0.60 (0.45 - 0.81)[c]	0.73 (0.53 - 1.01)
Q4: >38 years	0.23 (0.17 - 0.32)[c]	0.19 (0.12 - 0.28)[c]
Gender[a]—dimension 1 (gender identity)		
Man (only)	Reference	Reference
Woman (only)	0.82 (0.64 - 1.04)	0.68 (0.50 - 0.92)[c]
GGANB+[d]	2.27 (1.63 - 3.18)[c]	1.97 (1.36 - 2.84)[c]
Gender[a]—dimension 2 (gender modality)		
Cisgender	Reference	Reference
Transgender	2.17 (1.54 - 3.06)[c]	2.15 (1.46 - 3.13)[c]
Race/ethnicity[a]		
White (only)	Reference	Reference
Indigenous	1.65 (1.19 - 2.20)[c]	1.33 (0.91 - 1.93)
People of color	1.74 (1.34 - 2.26)[c]	2.01 (1.48 - 2.72)[c]
Sexual identity[a]		
Heterosexual	Reference	Reference
LGB+[e]	2.37 (1.89 - 2.97)[c]	2.53 (1.91 - 3.38)[c]
Educational attainment[f]		
Elementary or some high school	0.79 (0.38 - 1.62)	1.39 (0.56 - 3.34)
High school	0.84 (0.52 - 1.38)	1.09 (0.58 - 2.08)
Postsecondary certificate	0.97 (0.62 - 1.53)	1.34 (0.76 - 2.43)
Bachelor's degree	1.48 (0.94 - 2.35)	1.81 (1.03 - 3.31)[c]
Graduate degree	Reference	Reference
Income[g]		
Less than $20,000 per year	0.39 (0.24 - 0.65)[c]	0.32 (0.17 - 0.58)[c]
$20,000 to <$40,000 per year	0.77 (0.48 - 1.25)	0.53 (0.30 - 0.94)[c]
$40,000 to <$60,000 per year	0.94 (0.59 - 1.49)	0.65 (0.38 - 1.13)
$60,000 to <$80,000 per year	1.28 (0.79 - 2.09)	1.27 (0.74 - 2.20)
$80,000 or more per year	Reference	Reference

[a]Univariate models: the estimate is an unadjusted odds ratio.

[b]Quartile.

[c]Statistically significant results.

[d]GGANB+: genderqueer, genderfluid, agender, nonbinary people, and those with multiple gender identities.

[e]LGB+: lesbian, gay, bisexual, and other nonexclusively heterosexual identities.

[f]Adjusted for age, gender, sexual identity, and race/ethnicity. Estimate is an adjusted odds ratio.

[g]Adjusted for age, gender, sexual identity, race/ethnicity, and educational attainment. The estimate is an adjusted odds ratio.

XSL•FO
RenderX

In contrast, sexual identity and some comparisons of gender and race/ethnicity showed differences favoring less privileged groups. For sexual identity, higher odds of awareness and use were found in those in the LGB+ group (OR 2.37, 95% CI 1.89 - 2.97 for awareness and OR 2.53, 95% CI 1.91 - 3.38 for use) as compared to those identifying only as heterosexual. Gender identity (dimension 1) showed higher awareness and use in the GGANB+ group (OR 2.27, 95% CI 1.63 - 3.18 for awareness and OR 1.97, 95% CI 1.36 - 2.84 for use) compared to men (only). Gender modality (dimension 2) showed higher odds of awareness and use in those identifying as transgender compared to cisgender (OR 2.17, 95% CI 1.54 - 3.06 for awareness and OR 2.15, 95% CI 0.46 - 3.13 for use). These differences favoring less privileged groups in gender comparisons contrast with what was observed for those identifying only as women, who showed lower odds of use (OR 0.68, 95% CI 0.50 - 0.92) but similar levels of awareness compared to men (only).

A similar pattern favoring a less privileged group was observed in the category of people of color as compared to White (only) individuals (OR 1.74, 95% CI 1.34 - 2.26 for awareness and OR 2.01, 95% CI 1.48 - 2.72 for use). This pattern was not observed among Indigenous individuals who had higher odds of being aware of *GetCheckedOnline* (OR 1.65, 95% CI 1.19 - 2.20) compared to White (only) but no difference in use. Lastly, in education, adjusted for age, gender, sexual identity, and race/ethnicity, a difference in use was found between the 2 most privileged groups (bachelor's degree compared to graduate degree: aOR 1.81, 95% CI 1.03 - 3.31) but not lower education groups with no differences in awareness.

In sensitivity analyses, we first compared those who provided complete sociodemographic information (n=1166) to those without sociodemographic information (n=136). Awareness of *GetCheckedOnline* in those who did not give any sociodemographic information (58/136, 42.6%) was not significantly different than in those who provided complete sociodemographic characteristics (420/1166, 36%; Fisher exact test, *P*=.13). Use of *GetCheckedOnline* did differ between these groups (233/1166, 20% vs 39/136, 28.7%; Fisher exact test, *P*=.03). Since more participants skipped the income question than other sociodemographic questions, we compared this group to those who provided complete sociodemographic information in a second sensitivity analysis. Compared to those who provided complete sociodemographic details, those who omitted only the income question (n=109) differed by educational attainment (n=9, 8.3% vs n=45, 3.9% with elementary or some high school education; Fisher exact test, *P*=.02) and race/ethnicity (n=62, 56.9% vs n=796, 68.3% identifying as White; Fisher exact test, *P*=.048). Other sociodemographic characteristics did not significantly differ between these 2 groups. However, the awareness level did not differ between these 2 groups (32/109, 29.4% in those who only omitted income vs 233/1166, 36% in those with complete sociodemographic information; Fisher exact test, *P*=.17). Use of *GetCheckedOnline* did not differ between these groups (14/109, 12.8% vs 233/1166, 20%; Fisher exact test; *P*=.08).

Discussion

Principal Results

Awareness and use of *GetCheckedOnline* were high (approximately 1 in 3 and 1 in 5, respectively) in this sample of people in 5 communities where *GetCheckedOnline* is available. We found that differences in sociodemographic characteristics were associated with awareness and use of the service. Some of these differences favored more privileged groups (those in the central quartiles of age and those with higher income). In contrast, other differences favored equity-owed groups (people of color, transgender and gender-diverse participants, and nonheterosexual participants). Two groups—those identifying only as women and Indigenous people—showed similar or higher levels of awareness but not higher use of the service. This study makes an important contribution to the scant but growing literature on the equitable impacts of digital STBBI testing services by being one of the first to measure implementation outcomes across groups at a community level and to use a theoretically informed approach to health equity analyses. This approach has also allowed us to identify the specific equity-owed groups where further promotion of *GetCheckedOnline* may be needed to increase awareness, and where service adaptations may be required to increase use.

Comparison With Prior Work

Our results align with other research evaluating health equity indicators of digital STBBI testing services, in which differential uptake patterns have been observed with complexity in the directions and magnitude of the effects [7]. Our findings related to age and income are consistent with an overall tendency of interventions favoring privileged groups [7]. However, our findings related to gender, sexual identity, and race/ethnicity follow a different direction, with at least some oppressed groups showing higher uptake. Other studies have reported differences favoring transgender and nonbinary individuals and nonheterosexual men [9,13,20], but all studies finding differences associated with race/ethnicity show higher uptake in White people [9,11,14].

The diffusion of innovations theory proposes concepts that can further support the interpretation of these differences, such as the concept of relative advantage and communication networks [31]. In this case, we examined the relative advantages of *GetCheckedOnline* over provider-based testing in the context of the social stratification processes generating differences in living conditions. Digital access might provide some individuals with an alternative to avoid contact with potential experiences of discrimination in the health care system [32,33], which is a known issue affecting sexual and gender minorities and people of color, even more so in areas away from larger urban centers [34]. Previous research on *GetCheckedOnline* has highlighted the benefits that gay, bisexual, and other men who have sex with men (GBMSM) describe in terms of privacy, control of sample collection, and access to testing by men who live outside of larger urban areas [32]. Transgender and gender-nonconforming individuals have also provided insights

XSL•FO
RenderX

into their experiences testing, supporting *GetCheckedOnline*'s adaptation to offer a more inclusive service [35].

In addition, promotional efforts might also explain the higher awareness and use of *GetCheckedOnline* observed in some groups. Early promotion of *GetCheckedOnline* intentionally included representation of and directed marketing to GBMSM [36]. Historical targeting of HIV testing campaigns on GBMSM has also been related to a higher uptake of testing interventions by nonheterosexual men [11]. A higher perception of relative advantage and promotion efforts influences the diffusion through communication networks. This diffusion is more likely to occur among individuals who perceive themselves as similar [31], more so when topics related to sexual health and STBBI still carry a significant social stigma.

However, the service's relative advantage is impacted by its requirement for digital devices and access to the internet, which has been shown to limit access to digital health services for those with lower incomes [5]. Similarly, *GetCheckedOnline* might be less appealing to those looking for additional services, such as human papillomavirus testing, vaccines, or contraception. In other digital STBBI testing programs, lower uptake in women has been associated with their access to testing in the context of other health-related visits [9]. In other cases, provider-based testing may appear as a more acceptable health care pathway, which might relate to older individuals' trust in and preference for in-person services. In contrast, *GetCheckedOnline*, as a provincial service, might be perceived as part of the broader health care system, which in turn carries a sense of mistrust from Indigenous communities, coming from the historical and present harms that the colonial systems have perpetrated on them [37]. In addition, they might not see themselves reflected in the service promotion, as *GetCheckedOnline* has not been promoted directly to Indigenous people in the locations included in this study.

Limitations

Our study had the following limitations. Digital STBBI testing services include diverse service models that might not share essential features of *GetCheckedOnline* (eg, relying on external laboratory partners); therefore, findings might not be generalizable to other digital STBBI testing interventions. Our approach allowed us to estimate the differences associated with the selected sociodemographic characteristics as markers of social location. However, these differences should not be interpreted as the only determinants of the outcomes of our study. Although previous research on *GetCheckedOnline* has shown that variables related to sexual history (eg, behaviors, relationship status, and prior STBBI diagnoses) were not as strongly associated with awareness and use as sociodemographic characteristics [22], these variables may still account for differences in the service implementation outcomes. Similarly, our findings describe the differences associated with markers of social location in these urban, suburban, and rural communities. However, given that geographic location and urbanity function both as materialization and reproduction of social stratification processes [23,38], further analyses and comparisons with larger urban areas may be needed to better understand the role of geographic location on these implementation outcomes.

Furthermore, we relied on self-reported data, which may be subject to volunteer or recall bias, particularly related to implementation outcomes, and some participants did not provide complete sociodemographic information. Sensitivity analyses showed that *GetCheckedOnline* use significantly differed only between those who provided complete sociodemographic information and those who did provide some or no sociodemographic details. However, given the size effect of our found associations and the size of this group, we believe that missing data exclusion would be unlikely to bias the direction of these associations.

Future Research

Future research can support understanding this complex distribution of service uptake by sociodemographic characteristics. Critical to understanding these uptake patterns is exploring the interactions between different social stratification processes acting simultaneously through an intersectionality lens. Understanding these social processes and focusing on modifiable factors could inform program adjustments. Local context factors at the community level, such as the availability, acceptability, and appropriateness of testing services, play a significant role in service uptake. More research is needed to understand how social stratification processes affect these factors and inform service adaptation as a response. Longitudinal studies can show if the patterns we observed in this study are maintained over time and describe the dynamics of service penetration.

Conclusions

In conclusion, our study found that patterns of awareness and use of BC's digital STBBI testing service were associated with sociodemographic characteristics that reflect social stratification processes. Our findings suggest that *GetCheckedOnline* seems to respond equitably to the needs of some equity-owed groups (eg, sexual and gender minorities and people of color). However, further understanding of the lower uptake in other equity-owed groups (eg, lower-income people) is needed. These findings support the need for continuous evaluation of digital STBBI testing programs, and developers should evaluate the factors affecting equity in their local context. Efforts to promote and adapt digital STBBI testing interventions must consider the existing uptake pattern and the social and material circumstances determining accessibility to the service to ensure services reach those who experience a higher burden of these infections.

Acknowledgments

The authors would like to acknowledge the contribution of our community advisory board and our community partners, ANKORS and ASK Wellness Society. We are also grateful to the research assistants in each community, as well as to Hayley Watt for

XSL•FO
RenderX

leading the online promotion of the survey and data entry. We would also like to thank Dr Ihoghosa Iyamu, Dr Travis Salway, Dr Paul Flowers, Dr Nathan Lachowsky, and the CHER lab for their contribution to the design and implementation of the study.

Funding

Funding for this study was provided by the Canadian Institutes of Health Research (FRN: #CTW-1553878). The funding agency had no participation in the design, analysis, or decision to publish this study.

Authors' Contributions

AA, HJC, DH, HP, CW, D Grace, and MG designed the survey and data collection methods. RSR designed the study and data analysis plan. RSR conducted the data analysis with support from AA. RSR led the interpretation of findings with contributions from D Greyson, RK, and MG. RSR wrote the manuscript. MG, D Greyson, and RK supervised all stages of the study and crucially reviewed various versions of the study protocol and manuscript. All authors approved the final version of the manuscript. MG is responsible for the overall content of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
GetCheckedOnline service availability, GCO Community Survey Questionnaire, and nonmutually exclusive sociodemographic information from survey participants.
[DOCX File, 243 KB - jmir_v28i1e78561_app1.docx]

References

1. Exten C, Pinto CN, Gaynor AM, et al. Direct-to-consumer sexually transmitted infection testing services: a position statement from the American Sexually Transmitted Diseases Association. Sexual Trans Dis 2021;48(11):e155-e159. [doi: 10.1097/OLQ.0000000000001475]
2. Gilbert M, Salway T, Haag D, et al. Use of GetCheckedOnline, a comprehensive web-based testing service for sexually transmitted and blood-borne infections. J Med Internet Res 2017;19(3):e81. [doi: 10.2196/jmir.7097] [Medline: 28320690]
3. Gilbert M, Thomson K, Salway T, et al. Differences in experiences of barriers to STI testing between clients of the internet-based diagnostic testing service GetCheckedOnline.com and an STI clinic in Vancouver, Canada. Sex Transm Infect 2019;95(2):151-156. [doi: 10.1136/sextrans-2017-053325]
4. McAuley A. Digital health interventions: widening access or widening inequalities? Public Health (Fairfax) 2014;128(12):1118-1120. [doi: 10.1016/j.puhe.2014.10.008]
5. Lawrence K. Digital Health Equity: Exon Publications; 2022:121-130. [doi: 10.36255/exon-publications-digital-health-health-equity]
6. Lyles CR, Wachter RM, Sarkar U. Focusing on digital health equity. JAMA 2021;326(18):1795-1796. [doi: 10.1001/jama.2021.18459] [Medline: 34677577]
7. Iyamu I, Sierra-Rosales R, Estcourt CS, Salmon A, Koehoorn M, Gilbert M. Differential uptake and effects of digital sexually transmitted and bloodborne infection testing interventions among equity-seeking groups: a scoping review. Sex Transm Infect 2023;99(8):554-560. [doi: 10.1136/sextrans-2023-055749]
8. van den Broek IVF, van Bergen J, Brouwers E, et al. Effectiveness of yearly, register based screening for chlamydia in the Netherlands: controlled trial with randomised stepped wedge implementation. BMJ 2012;345(jul05 1):e4316-e4316. [doi: 10.1136/bmj.e4316]
9. Barnard S, Free C, Bakolis I, Turner KME, Looker KJ, Baraitser P. Comparing the characteristics of users of an online service for STI self-sampling with clinic service users: a cross-sectional analysis. Sex Transm Infect 2018;94(5):377-383. [doi: 10.1136/sextrans-2017-053302]
10. van Bergen JE, Fennema JS, van den Broek IV, et al. Rationale, design, and results of the first screening round of a comprehensive, register-based, chlamydia screening implementation programme in the Netherlands. BMC Infect Dis 2010;10(1). [doi: 10.1186/1471-2334-10-293]
11. Gilbert M, Salway T, Haag D, et al. A cohort study comparing rate of repeat testing for sexually transmitted and blood-borne infections between clients of an internet-based testing programme and of sexually transmitted infection clinics in Vancouver, Canada. Sex Transm Infect 2019;95(7):540-546. [doi: 10.1136/sextrans-2019-054071]
12. Grov C, Westmoreland DA, Carneiro PB, et al. Recruiting vulnerable populations to participate in HIV prevention research: findings from the Together 5000 cohort study. Ann Epidemiol 2019;35:4-11. [doi: 10.1016/j.annepidem.2019.05.003]
13. Fistonich GM, Troutman KM, Visconti AJ. A pilot of mail-out HIV and sexually transmitted infection testing in Washington, District of Columbia during the COVID-19 pandemic. Am J Prev Med 2021;61(5):S16-S25. [doi: 10.1016/j.amepre.2021.06.009]

XSL•FO
RenderX

14. Ricca AV, Hall EW, Khosropour CM, Sullivan PS. Factors associated with returning at-home specimen collection kits for HIV testing among internet-using men who have sex with men. J Int Assoc Provid AIDS Care 2016;15(6):463-469. [doi: 10.1177/2325957416668579] [Medline: 27635015]
15. Zuure FR, Davidovich U, Coutinho RA, et al. Using mass media and the internet as tools to diagnose hepatitis C infections in the general population. Am J Prev Med 2011;40(3):345-352. [doi: 10.1016/j.amepre.2010.11.019]
16. Götz HM, van den Broek IVF, Hoebe C, et al. High yield of reinfections by home-based automatic rescreening of Chlamydia positives in a large-scale register-based screening programme and determinants of repeat infections. Sex Transm Infect 2013;89(1):63-69. [doi: 10.1136/sextrans-2011-050455]
17. Op de Coul ELM, Götz HM, van Bergen J, et al. Who participates in the Dutch Chlamydia screening? A study on demographic and behavioral correlates of participation and positivity. Sex Transm Dis 2012;39(2):97-103. [doi: 10.1097/OLQ.0b013e3182383097] [Medline: 22249297]
18. Phanuphak N, Anand T, Jantarapakde J, et al. What would you choose: online or offline or mixed services? Feasibility of online HIV counselling and testing among Thai men who have sex with men and transgender women and factors associated with service uptake. J Intern AIDS Soc 2018;21(S5). [doi: 10.1002/jia2.25118]
19. Day S, Smith J, Perera S, Jones S, Kinsella R. Beyond the binary: sexual health outcomes of transgender and non-binary service users of an online sexual health service. Int J STD AIDS 2021;32(10):896-902. [doi: 10.1177/0956462420982830]
20. Veronese V, Ryan KE, Hughes C, Lim MS, Pedrana A, Stoové M. Using digital communication technology to increase HIV testing among men who have sex with men and transgender women: systematic review and meta-analysis. J Med Internet Res 2020;22(7):e14230. [doi: 10.2196/14230] [Medline: 32720902]
21. Proctor E, Silmere H, Raghavan R, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011;38(2):65-76. [doi: 10.1007/s10488-010-0319-7]
22. Montiel A, Ablona A, Klassen B, et al. Reach of GetCheckedOnline among gay, bisexual, transgender and queer men and two-spirit people and correlates of use 5 years after program launch in British Columbia, Canada. Sex Transm Infect 2024;100(4):208-215. [doi: 10.1136/sextrans-2023-056007]
23. Dover DC, Belon AP. The health equity measurement framework: a comprehensive model to measure social inequities in health. Int J Equity Health 2019;18(1):36. [doi: 10.1186/s12939-019-0935-0]
24. World Health Organization. Global report on ageism. 2021 URL: https://apps.who.int/iris/rest/bitstreams/1336324/retrieve [accessed 2023-02-22]
25. Ashley F, Brightly-Brown S, Rider GN. Beyond the trans/cis binary: introducing new terms will enrich gender research. Nature New Biol 2024;630(8016):293-295. [doi: 10.1038/d41586-024-01719-9]
26. Textor J, van der Zander B, Gilthorpe MS, Liśkiewicz M, Ellison GTH. Robust causal inference using directed acyclic graphs: the R package 'DAGitty'. Int J Epidemiol 2017:dyw341. [doi: 10.1093/ije/dyw341]
27. Greenland S, Pearce N. Statistical foundations for model-based adjustments. Annu Rev Public Health 2015;36(1):89-108. [doi: 10.1146/annurev-publhealth-031914-122559] [Medline: 25785886]
28. Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. Am J Epidemiol 2013;177(4):292-298. [doi: 10.1093/aje/kws412] [Medline: 23371353]
29. Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. JOSS 2019;4(43):1686. [doi: 10.21105/joss.01686]
30. R Core Team. R: a language and environment for statistical computing. URL: https://www.R-project.org/ [accessed 2026-03-19]
31. Rogers EM. Diffusion of Innovations: Free Press; 2003. URL: http://ebookcentral.proquest.com/lib/ubc/detail.action?docID=4935198 [accessed 2023-06-21]
32. Knight RE, Chabot C, Carson A, et al. Qualitative analysis of the experiences of gay, bisexual and other men who have sex with men who use GetCheckedOnline.com: a comprehensive internet-based diagnostic service for HIV and other STIs. Sex Transm Infect 2019;95(2):145-150. [doi: 10.1136/sextrans-2018-053645]
33. Stewart M, Ryu H, Blaque E, et al. Cisnormativity as a structural barrier to STI testing for trans masculine, two-spirit, and non-binary people who are gay, bisexual, or have sex with men. PLoS ONE 2022;17(11):e0277315. [doi: 10.1371/journal.pone.0277315] [Medline: 36441729]
34. Apodaca C, Casanova-Perez R, Bascom E, et al. Maybe they had a bad day: how LGBTQ and BIPOC patients react to bias in healthcare and struggle to speak out. J Am Med Inform Assoc 2022;29(12):2075-2082. [doi: 10.1093/jamia/ocac142]
35. Farrell J, Haag D, Bondyra M, Thomson K, Gilbert M. Should your visuals be wearing blue, pink or yellow pants (or no pants at all)? Developing specimen self-collection instructions that are more sensitive to the needs of traditionally marginalized and underserved client populations. 2015 Presented at: Gay & Lesbian Medical Association 33rd Annual Conference; Oct 17-20, 2015.
36. Gilbert M, Salway T, Haag D, et al. Assessing the impact of a social marketing campaign on program outcomes for users of an internet-based testing service for sexually transmitted and blood-borne infections: observational study. J Med Internet Res 2019;21(1):e11291. [doi: 10.2196/11291] [Medline: 30664456]
37. Loppie S, Reading C, de Leeuw S. Indigenous experiences with racism and its impacts. : National Collaborating Centre for Aboriginal Health; 2014.

XSL•FO
RenderX

38. Bernroider L, Miro Born A, Kulz C, Gang SU. Introduction: exploring urban violence and inequality from intersectional perspectives. In: Intersectionality and the City: Routledge.

Abbreviations

aOR: adjusted odds ratio
BC: British Columbia
GBMSM: gay, bisexual, and other men who have sex with men
GGANB+: genderfluid, genderqueer, agender, nonbinary, and people with multiple gender identities
LGB+: lesbian, gay, bisexual, or other nonexclusively heterosexual identities
OR: odds ratio
REDCap: Research Electronic Data Capture
STBBI: sexually transmitted and blood-borne infections

Edited by M Balcarras; submitted 17.Jun.2025; peer-reviewed by G Kothawar, NS Wong; revised version received 23.Feb.2026; accepted 24.Feb.2026; published 10.Apr.2026.

Please cite as:
Sierra Rosales R, Ablona A, Chang HJ, Haag D, Pedersen H, Worthington C, Grace D, Knight R, Greyson D, Gilbert M
Health Equity Analysis of Awareness and Use of GetCheckedOnline, British Columbia's Digital Intervention for Sexually Transmitted and Blood-Borne Infection Testing in 5 Urban, Suburban, and Rural Communities: Cross-Sectional Survey Study
J Med Internet Res 2026;28:e78561
URL: https://www.jmir.org/2026/1/e78561
doi:10.2196/78561

© Rodrigo Sierra Rosales, Aidan Ablona, Hsiu-Ju Chang, Devon Haag, Heather Pedersen, Catherine Worthington, Daniel Grace, Rod Knight, Devon Greyson, Mark Gilbert. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Prevalence of Social Media Addiction and Associations With Usage Patterns, Burnout, and Health Conditions Among Medical Trainees in China: Cross-Sectional Study

Zexu Guan[1], PhD; Ni Tang[2,3,4,5], MPH; Guoshuai Luo[6], PhD; Xiao Zhang[2,3,4,5], MD, PhD

[1]School of International Politics and Communication, Beijing Language and Culture University, Beijing, China
[2]Peking University Sixth Hospital, Huayuanbei Road 51, Haidian District, Beijing, China
[3]Peking University Institute of Mental Health, Beijing, China
[4]NHC Key Laboratory of Mental Health, Peking University, Beijing, China
[5]National Clinical Research Center for Mental Disorders, Peking University Sixth Hospital, Beijing, China
[6]Tianjin Anding Hospital, Tianjin, China

Corresponding Author:
Xiao Zhang, MD, PhD
Peking University Sixth Hospital, Huayuanbei Road 51, Haidian District, Beijing, China

Abstract

Background: Medical residency is a demanding training stage characterized by high levels of stress and burnout. As digital natives, current medical trainees (ie, residents) are frequent users of social media; however, little is known about how their personal (nonprofessional) use relates to burnout and social media addiction (SMA).

Objective: This study aims to characterize the prevalence of SMA among Chinese medical trainees and explore its complex relationships with social media use patterns, occupational burnout, and related risk and protective factors.

Methods: A nationwide cross-sectional survey was deployed through Wenjuanxing and disseminated via WeChat between August 29 and September 10, 2024. Data included demographics, physical and psychiatric health history, work variables (eg, training year and night shifts), personality traits, and social media use. SMA was assessed using the Bergen Social Media Addiction Scale. Logistic regression was performed to identify predictors of addiction, and mediation and moderation analyses were conducted to clarify the role of occupational burnout.

Results: Of 3621 medical trainees, 211 (5.8%) met the criteria for SMA (Bergen Social Media Addiction Scale ≥24, indicating addiction). Second-year medical trainees reported the highest addiction prevalence (92/1159, 7.9%). Logistic regression analysis revealed that higher burnout (odds ratio [OR] 1.41, 95% CI 1.23-1.62; $P<.001$), longer daily use (OR 1.39, 95% CI 1.23-1.56; $P<.001$), physical health problems (OR 1.56, 95% CI 1.13-2.16; $P=.006$), and psychiatric history (OR 2.00, 95% CI 1.41-2.84; $P<.001$) significantly increased the odds of addiction, whereas conscientiousness was protective (OR 0.92, 95% CI 0.86-0.99; $P=.02$). Social media use showed significant U-shaped associations with burnout, physical health problems, psychiatric history, personality characteristics, and mental health outcomes. For example, medical trainees using social media 1 hour or less (104/404, 25.7% with psychiatric history) and more than 4 hours daily (97/419, 23.2% with psychiatric history) both had higher risk profiles than moderate users. Mediation analysis showed that occupational burnout explained 28.1% of the effect of psychiatric history and 29.6% of the effect of physical health problems on addiction risk.

Conclusions: This large-scale survey provides the first systematic characterization of SMA among Chinese medical trainees and elucidates its associated risks and protective factors. Burnout consistently emerged as a key and pervasive predictor of SMA, functioning both as an independent risk factor and as a mediator amplifying the impact of health-related vulnerabilities. Moreover, the findings highlight that both minimal and excessive daily social media use may signal distinct behavioral manifestations of distress, potentially reflecting different clinical phenotypes: digital disengagement under acute stress versus compulsive engagement driven by chronic burnout. Notably, while mental health symptoms exhibited U-shaped associations with usage, SMA risk increased progressively with daily duration. These results underscore the need for interventions that extend beyond simply monitoring usage duration, emphasizing strategies to reduce burnout and enhance the overall well-being of medical trainees.

(J Med Internet Res 2026;28:e75675) doi:10.2196/75675

KEYWORDS

social media addiction; SMA; Chinese medical trainees; occupational burnout; social media usage time; U-shaped relationship

XSL•FO
RenderX

Introduction

Social media has become deeply embedded in the everyday lives of young adults and is increasingly intertwined with medical education, communication, and professional identity formation [1]. A recent systematic review of social media in undergraduate medical education showed that these platforms are now widely used by medical learners for communication, knowledge exchange, and informal learning, while also raising concerns regarding professionalism and potential adverse effects on mental and physical well-being [2]. However, growing evidence suggests that problematic or addictive patterns of social media use are associated with depression, anxiety, sleep disturbances, and poorer mental well-being, as highlighted in recent reviews and large-scale studies [3].

Social media addiction (SMA) is characterized not only by excessive time spent online but also by withdrawal symptoms (eg, irritability when access is restricted) and functional impairment in daily or occupational domains [4,5]. Despite growing evidence linking SMA to adverse mental health outcomes in medical learner populations, there remains limited comprehensive evidence on its prevalence, determinants, and potential mechanisms among medical trainees in China, a population uniquely vulnerable to chronic stress [6]. Given the rapid digitalization and widespread integration of social media into daily life in China [7], understanding SMA in this context is particularly important.

Medical trainees face a significant transition from an educational environment to clinical scenarios, including long working hours, frequent night shifts, high patient loads, and continuous exposure to tasks that demand substantial physical and emotional effort, as well as constant engagement with complex clinical situations [8]. Under these pressures, medical trainees may be particularly vulnerable to problematic social media use as a maladaptive coping strategy, which has been linked—along with poor work-life balance—to increased stress, depression, and burnout [9]. In parallel, prolonged work-related stress is a well-established driver of occupational burnout among medical trainees [10-14], characterized by a triad of symptoms: emotional exhaustion, depersonalization (cynicism), and a reduced sense of personal accomplishment [15]. However, the role of burnout in the development of SMA remains unclear, particularly whether it functions as an independent predictor or as a mediator linking vulnerability factors to addictive social media use.

To address these gaps, we conducted a data-driven, nationwide cross-sectional survey of Chinese medical trainees. This study aimed to (1) estimate the prevalence of SMA and identify associated risk and protective factors, including occupational burnout, health conditions (both physical and mental), and usage patterns; (2) characterize patterns of social media use (eg, daily usage time and number of platforms); and (3) examine the direct and indirect roles of burnout in the development of SMA.

Methods

Study Design and Participants

This study used a cross-sectional survey design to explore the factors contributing to SMA among medical trainees in China. Data were collected using an anonymous online questionnaire. A completed CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist is provided in Checklist 1 to ensure the transparency and reproducibility of the e-survey results.

Medical Training Pathway in China

In this study, the medical trainees were residents enrolled in China's standardized residency training (SRT) program. They had completed a 5-year bachelor's degree in medicine. Participants represented three common SRT pathways: (1) social track (independent medical students entering residency directly after undergraduate training), (2) integrated master's track (students completing a coordinated master's degree and residency program), and (3) other track (individuals with atypical educational routes, eg, PhD holders who subsequently entered residency).

Survey Development and Pilot Testing

The questionnaire was developed by experienced psychologists and psychiatrists, drawing on validated instruments commonly used in mental health and behavioral research. The final questionnaire consisted of 36 major questions and 84 items, including matrix subitems. It had four sections: (1) basic information (15 major questions; eg, demographics and physical and psychiatric health conditions), (2) workload and lifestyle factors (9 major questions; eg, number of night shifts and daily social media use), (3) mental health status (3 major questionnaires), and (4) psychological characteristics (9 major questions; eg, personality traits and occupational burnout). The draft survey was pilot-tested with 20 graduate students to ensure clarity, usability, and technical functionality. Minor revisions were made based on the feedback.

Survey Content and Completion

Respondents were required to complete all 80 core items before submission to ensure data completeness. The 4 noncore items were optional and included preferred social media influencers, positive experiences during residency training, negative experiences during residency training, and suggestions for educational or health authorities. The average completion time was 495 (SD 966) seconds (approximately 8 min).

Survey Dissemination

The online survey was hosted by Wenjuanxing, the largest online survey platform in China. Recruitment was conducted via WeChat, the most widely used social media platform in China. Medical trainees could directly access and complete the survey by clicking the Wenjuanxing link on WeChat. The survey link was disseminated through professional groups and residency trainee networks. Data were collected for 2 weeks (from August 29, 2024, to September 10, 2024).

XSL•FO
RenderX

Eligibility and Data Quality Assurance

Medical trainees enrolled in SRT programs were eligible. One mandatory item required participants to report their specialty, and responses outside the national residency catalog were excluded during data cleaning. To ensure authenticity, cookies were used to prevent duplicate entries, and questionnaires with implausibly short completion times (<2 min) were excluded. No statistical weighting or propensity score adjustments were applied because the sample size was sufficiently large and the sex distribution was balanced. Furthermore, as all mandatory items were required for submission, the final dataset contained no missing values.

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki and was approved by the ethics committee of Tianjin Anding Hospital (2023‐27). The informed consent form was presented on the first page of the online questionnaire, and only participants who checked the consent box were instructed to complete the remainder of the questionnaire. All mandatory questions were nonidentifiable and contained no items that could be used to trace participants. Participants who completed the survey received monetary compensation of 5 RMB (<US $1), which was considered unlikely to introduce response bias.

Data Collection and Instruments

Social media use was assessed based on the type of platform used and the average daily duration of use. SMA was evaluated using the Bergen Social Media Addiction Scale (BSMAS), a brief self-report instrument derived from the component model of behavioral addiction. The BSMAS consists of 6 items assessing core addiction components (salience, mood modification, tolerance, withdrawal, conflict, and relapse), each rated on a 5-point Likert scale (1="very rarely" to 5="very often"), yielding a total score of 6 to 30 points. The scale has demonstrated good reliability and validity in Chinese populations, and a cutoff score of 24 or more was used to indicate probable SMA [16,17]. Importantly, a large meta-analysis of the prevalence of SMA across 32 countries reported that all included epidemiological studies used the BSMAS framework, including the original and translated versions, supporting its widespread adoption and cross-study comparability [18].

In addition to social media–related information, the questionnaire covered basic information such as age, sex, financial burden, physical health problems, psychiatric history, and prior psychological treatment. Physical health problems, psychiatric history, prior psychological treatment, and financial burden were assessed using binary (yes or no) self-report items. Participants were asked whether they had any chronic physical health conditions requiring long-term attention, had ever been diagnosed with mental health conditions such as anxiety or depression, had received psychological counseling or psychotherapy, and had family-related financial responsibilities (ie, providing financial support to parents or other family members or supporting children). For all items, a "yes" response indicated the presence of the corresponding condition or experience. The exact wording of all corresponding questionnaire items is provided in Multimedia Appendix 1.

Work-related variables included training year, average monthly night shifts, and occupational burnout, with burnout assessed on a 5-point Likert scale (1="no burnout" to 5="extremely high burnout"). Personality traits were measured using the Chinese version of the 10-Item Personality Inventory, which has demonstrated good reliability in assessing the Big Five personality dimensions in Chinese populations [19]. Mental health outcomes included depressive symptoms, anxiety, and sleep problems, which were assessed using the Patient Health Questionnaire-9, the Generalized Anxiety Disorder-7 scale, and the Insomnia Severity Index, respectively.

Statistical Analysis

Descriptive Statistics

Descriptive statistics were used to summarize participants' basic characteristics, work variables, personality traits, and social media usage. This step provided a clear picture of the key attributes of the sample and a foundation for further analysis. As training year was a categorical variable with three levels, this factor was not included in the subsequent regression and interaction analyses for the BSMAS scores to maintain model parsimony and ease of interpretation. However, to elucidate its impact on medical trainees' social media behaviors, we examined its relationship with the number of frequently used social media platforms and BSMAS scores. Specifically, one-way analyses of variance were conducted to determine whether differences existed in both the platform numbers and BSMAS scores across the 3 training levels (first, second, and third years). Chi-square tests were used to compare the prevalence of SMA among these groups.

Correlation and Regression Analyses

To provide a preliminary overview of the interrelationships among all variables, a Spearman correlation analysis was conducted, and the results were visualized using a comprehensive heatmap that included demographic, personality, and mental health indicators. To identify the factors associated with SMA, a univariate regression analysis was used to assess the relationship between each independent variable (age, sex, personality traits, and social media usage) and the dependent variable (BSMAS score). Mental health outcomes (Patient Health Questionnaire-9, Generalized Anxiety Disorder-7, and Insomnia Severity Index) were excluded as independent variables to prevent multicollinearity and to ensure that the analysis focused on underlying predictors rather than co-occurring psychological symptoms. Following this, logistic regression was used to refine the model by selecting the most relevant variables affecting SMA (BSMAS ≥24). This approach used the full information of the scale while clarifying the most important diagnostic indicators. Stratified analyses were conducted by residency year (first, second, and third), repeating the logistic regression analysis within each subgroup.

Mediation and Moderation Analyses

Predictors of addictive behaviors were classified according to the interaction of the person-affect-cognition-execution model

XSL•FO
RenderX

for addictive behaviors [20,21]. Accordingly, variables were categorized into relatively stable predisposing factors corresponding to the "person" component (including psychiatric history, physical health problems, and personality traits) and more dynamic factors (such as occupational burnout and night-shift frequency). We hypothesized that these dynamic factors mediate or moderate the association between stable predictors and the severity of SMA. To test this hypothesis, two sets of analyses were conducted: (1) mediation analyses to examine whether dynamic factors mediated the effects of stable predictors on BSMAS scores (using 5000 bootstrap samples) and (2) moderation analyses to evaluate whether dynamic factors modified the relationships between stable predictors and BSMAS scores.

Results

Demographics and Descriptive Statistics: Study Population Characteristics

A total of 3621 medical trainees were included in the analysis. The mean age was 25.3 (SD 2.5) years, and the sample included 1567 (43.2%) males and 2054 (56.8%) females. Participants were well distributed across residency years. Overall, 1065 (29.4%) participants reported physical health problems, 707 (19.5%) reported a psychiatric history, 472 (13%) had received prior psychological treatment, and 1176 (32.5%) reported a financial burden. Nearly 60% (n=2173) of participants were only children. Detailed demographic characteristics are presented in Table 1, and post hoc results are provided in Tables S1-S9 and Figures S1-S3 in Multimedia Appendix 1.

Table . Demographic and baseline health characteristics of Chinese medical trainees enrolled in standardized residency training programs who participated in a nationwide cross-sectional anonymous online survey conducted from August 29, 2024, to September 10, 2024 (N=3621).

Characteristics	Values
Age (years), mean (SD)	25.3 (2.5)
Sex, n (%)	
Male	1567 (43.2)
Female	2054 (56.8)
Training year, n (%)	
First year	1657 (45.7)
Second year	1159 (32)
Third year	805 (22.2)
Physical health problems (yes), n (%)	1065 (29.4)
Psychiatric history (yes), n (%)	707 (19.5)
Financial burden (yes), n (%)	1176 (32.5)
Only child status (yes), n (%)	2139 (59)
Prior psychological treatment (yes), n (%)	472 (13)

Demographics and Work-Related Characteristics

Sex distribution differed significantly across social media usage groups (χ^2_4=49.786; P<.001). When distributions were examined within sex, a higher proportion of female medical trainees reported social media use of 2 hours or more per day (1233/2054, 60.1%) compared with male medical trainees (728/1567, 46.5%). Usage duration was significantly associated with physical health problems (χ^2_4=29.894; P<.001) and psychiatric history (χ^2_4=20.447; P<.001), with higher prevalence observed at both extremes of social media use, defined as minimal use (≤1 h/d) and excessive use (>4 h/d), compared with moderate use. Financial burden also differed significantly across usage groups (χ^2_4=33.250; P<.001). No significant differences were detected for age, prior psychological treatment, or only-child status. Regarding work-related variables, both minimal and excessive users reported higher occupational burnout scores than moderate users ($F_{4, 3616}$=14.393; P<.001). Night-shift frequency varied across groups ($F_{4, 3616}$=5.088; P<.001), with higher averages observed among minimal users. Training year distribution differed significantly across usage categories (χ^2_8=25.383; P=.001), with excessive social media use becoming more prevalent across advancing residency years. Details are presented in Table 2.

XSL•FO
RenderX

Table . Demographic and work-related characteristics of Chinese medical trainees, stratified by daily social media use duration (N=3621)[a].

Characteristic	≤1 h (n=404)	1 - 2 hours (n=1256)	2 - 3 hours (n=1069)	3 - 4 hours (n=473)	>4 hours (n=419)	Statistics	*P* value
Age (years), mean (SD)	25.5 (2.64)	25.39 (2.58)	25.19 (2.36)	25.24 (2.30)	25.34 (2.39)	1.638[b] (4, 3616)	.16
Sex, n (%)						49.786[c] (4)	<.001
Male	225 (14.4)	741 (47.3)	387 (24.7)	174 (11.1)	167 (10.7)		
Female	179 (8.7)	515 (25.1)	682 (33.2)	299 (14.6)	252 (12.3)		
Physical health problems, n (%)	153 (37.9)	334 (26.6)	308 (28.8)	121 (25.6)	149 (35.6)	29.894[c] (4)	<.001
Psychiatric history, n (%)	104 (25.7)	213 (17)	193 (18.1)	88 (18.6)	97 (23.2)	20.447[c] (4)	<.001
Prior psychological treatment, n (%)	48 (11.9)	146 (11.6)	150 (14)	69 (14.6)	59 (14.1)	5.027[c] (4)	.29
Only child, n (%)	239 (59.2)	747 (59.5)	613 (57.3)	290 (61.3)	255 (60.9)	2.915[c] (4)	.57
Financial burden, n (%)	179 (44.3)	412 (32.8)	309 (28.9)	143 (30.2)	133 (31.7)	33.250[c] (4)	<.001
Training year, n (%)						25.383[c] (8)	.001
First year	191 (11.5)	619 (37.4)	469 (28.3)	213 (12.9)	165 (10)		
Second year	114 (9.8)	386 (33.3)	376 (32.4)	144 (12.4)	139 (12)		
Third year	99 (12.3)	251 (31.2)	224 (27.8)	116 (14.4)	115 (14.3)		
Monthly night shifts, mean (SD)	6.47 (6.08)	5.44 (5.10)	5.23 (5.24)	4.98 (4.79)	5.53 (5.89)	5.088[b] (4, 3616)	<.001
Occupational burnout, mean (SD)	2.64 (1.31)	2.28 (0.99)	2.36 (0.96)	2.21 (0.86)	2.51 (1.14)	14.393[b] (4, 3616)	<.001

[a]Data were derived from a nationwide cross-sectional anonymous online survey conducted from August 29, 2024, to September 10, 2024.

[b]*F* test.

[c]Chi-square test.

Personality Traits, Social Media Engagement, and Mental Health Outcomes

Among the Big Five personality traits, openness, conscientiousness, agreeableness, and emotional stability showed significant between-group differences (all F=3.09 - 8.32; all P<.05), whereas extraversion did not. Medical trainees with moderate social media use tended to exhibit more favorable personality profiles. The number of frequently used social media platforms ($F_{4, 3616}$=39.476; P<.001), BSMAS scores ($F_{4, 3616}$=48.023; P<.001), and the prevalence of SMA (χ^2_4=43.160; P<.001) increased with the duration of social media use. Mental health outcomes also differed significantly by usage duration. Depression, anxiety, and insomnia scores followed a U-shaped pattern (all F=20.82 - 24.83; all P<.001), with higher symptom levels observed among both minimal and excessive users compared with moderate users. Details are presented in Table 3.

XSL•FO
RenderX

Table . Personality traits, social media engagement indicators, and mental health scale scores among Chinese medical trainees, stratified by daily social media use duration (N=3621)[a].

Characteristics	≤1 hour	1 - 2 hours	2 - 3 hours	3 - 4 hours	>4 hours	Statistics	*P* value
Openness, mean (SD)	0.44 (2.31)	0.69 (2.21)	0.72 (2.14)	0.74 (2.21)	0.38 (2.23)	3.091[b] (4, 3616)	.02
Conscientiousness, mean (SD)	0.65 (2.53)	1.13 (2.32)	0.86 (2.29)	0.78 (2.36)	0.54 (2.32)	6.995[b] (4, 3616)	<.001
Extraversion, mean (SD)	–0.47 (2.66)	–0.37 (2.58)	–0.19 (2.61)	–0.42 (2.65)	–0.36 (2.71)	1.255[b] (4, 3616)	.29
Agreeableness, mean (SD)	2.10 (2.24)	2.59 (2.04)	2.50 (2.01)	2.58 (2.06)	2.30 (2.17)	5.344[b] (4, 3616)	<.001
Emotional stability, mean (SD)	0.82 (2.49)	1.23 (2.39)	1.04 (2.35)	1.22 (2.40)	0.52 (2.47)	8.319[b] (4, 3616)	<.001
Number of platforms, mean (SD)	2.32 (1.66)	2.80 (1.42)	3.14 (1.43)	3.29 (1.38)	3.36 (1.70)	39.476 (4, 3616)	<.001
BSMAS[c] score, mean (SD)	13.07 (5.57)	14.18 (4.78)	15.64 (4.69)	15.92 (4.93)	16.90 (5.36)	48.023 (4, 3616)	<.001
SMA[d], n (%)	19 (4.7)	46 (3.7)	63 (5.9)	32 (6.8)	51 (12.2)	43.160[e] (4)	<.001
PHQ-9[f], mean (SD)	9.54 (8.01)	6.63 (5.78)	6.6 (5.27)	6.4 (4.96)	7.89 (6.3)	24.825 (4, 3616)	<.001
GAD-7[g], mean (SD)	7.07 (6.56)	4.82 (4.97)	4.81 (4.54)	4.39 (4.31)	5.78 (5.49)	21.607 (4, 3616)	<.001
ISI[h], mean (SD)	9.99 (8.02)	7.09 (5.94)	7.03 (5.81)	6.84 (5.66)	7.91 (6.68)	20.822 (4, 3616)	<.001

[a]Data were derived from a nationwide cross-sectional anonymous online survey conducted from August 29, 2024, to September 10, 2024.

[b]*F* test.

[c]BSMAS: Bergen Social Media Addiction Scale.

[d]SMA: social media addiction; defined as a BSMAS score ≥24.

[e]Chi-square test.

[f]PHQ-9: 9-item Patient Health Questionnaire.

[g]GAD-7: 7-item Generalized Anxiety Disorder.

[h]ISI: Insomnia Severity Index.

Correlation Analysis

The Spearman correlation heatmap (Figure 1) provides a comprehensive overview of the interrelationships among demographic, personality, and mental health variables. Specifically, BSMAS scores, as an index of SMA, showed significant positive correlations with occupational burnout (r=0.23; P<.001), the number of frequently used platforms (r=0.15; P<.001), and monthly night shifts (r=0.04; P=.029). Regarding mental health outcomes, BSMAS scores exhibited robust positive associations with depression (r=0.38; P<.001), anxiety (r=0.37; P<.001), and insomnia (r=0.33; P<.001). Furthermore, significant correlations were observed between BSMAS and several categorical risk factors: physical health problems (r=0.19; P<.001), psychiatric history (r=0.19; P<.001), and prior psychological treatment (r=0.15, P<.001) were all positively associated with addiction severity. In terms of personality traits, conscientiousness (r=–0.23; P<.001), stability (r=–0.29; P<.001), and agreeableness (r=–0.25; P<.001) exhibited significant negative correlations with BSMAS scores. These preliminary findings suggest that SMA is closely associated with systemic occupational stress and individual psychological vulnerabilities.

XSL•FO
RenderX

Figure 1. Spearman correlation heatmap of demographic, workload, personality, and mental health variables among Chinese medical trainees (N=3621). Data were derived from a nationwide cross-sectional anonymous online survey conducted from August 29, 2024, to September 10, 2024. The heatmap illustrates intercorrelations among all analyzed variables, with the color gradient representing the strength and direction of associations (red: positive; blue: negative). Numbers within cells indicate correlation coefficients. Categorical and ordinal variables were coded as follows: physical health problems, psychiatric history, prior psychological treatment, and financial burden (1="no" and 2="yes"); only-child status (1="yes" and 2="no"); and training year (1="year 1," 2="year 2," and 3="year 3"). BSMAS: Bergen Social Media Addiction Scale; ISI: Insomnia Severity Index; SM: social media. *P<.05, **P<.01, ***P<.001.

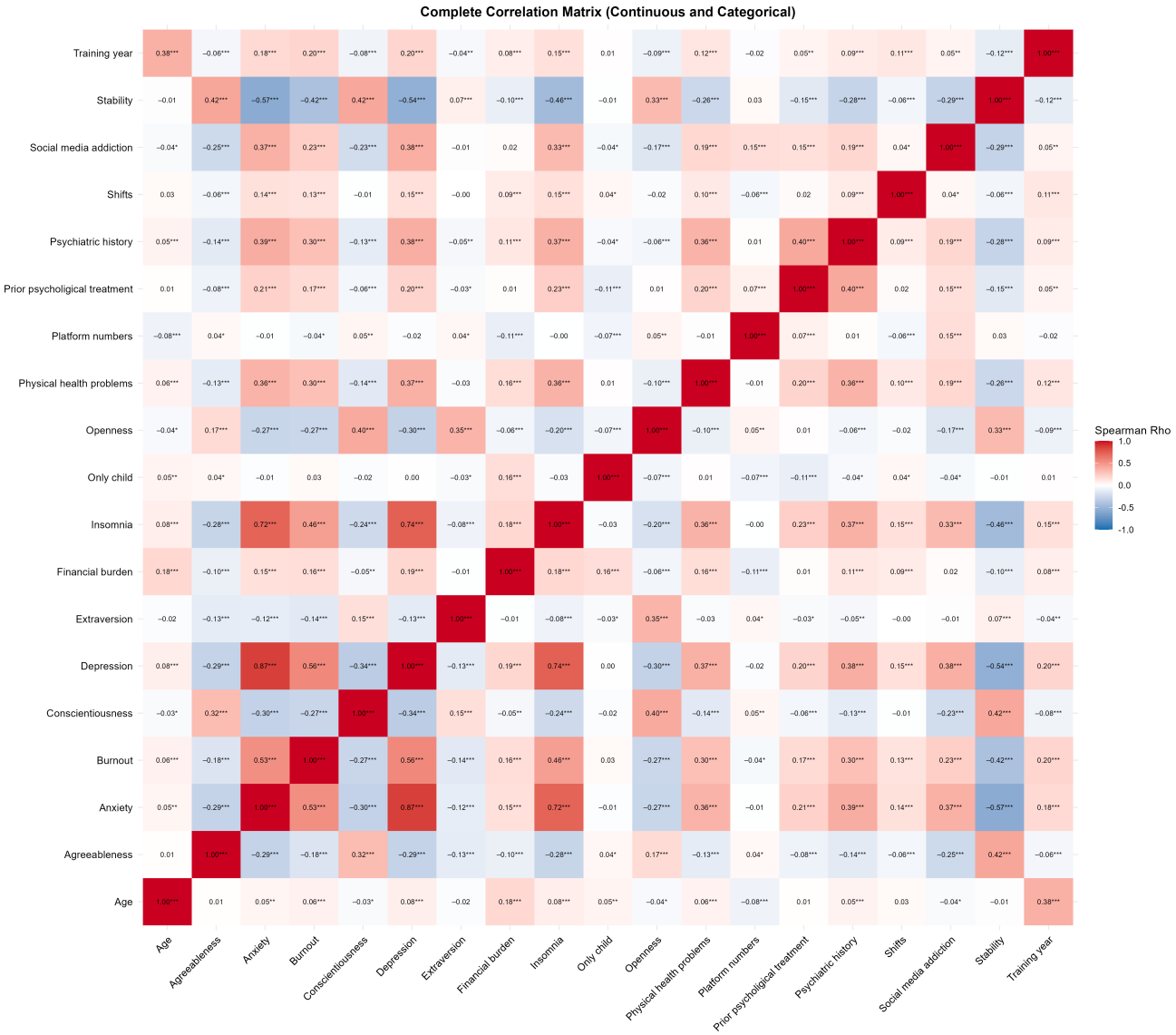

Regression Analysis

Univariate linear regression analyses identified several significant predictors of BSMAS scores (Figure 2; Table S10 in Multimedia Appendix 1). Higher occupational burnout, more frequent night shifts, longer daily social media use, female sex, physical health problems, and psychiatric history were positively associated with BSMAS scores (β=.15 to .35; all P<.05). In contrast, conscientiousness was inversely associated with BSMAS scores (β=–.22; P=.003). No other variables were significantly associated with BSMAS scores.

XSL•FO
RenderX

Figure 2. Forest plot of univariable linear regression coefficients for predictors of Bergen Social Media Addiction Scale (BSMAS) scores among Chinese medical trainees (N=3621). Data were derived from a nationwide cross-sectional anonymous online survey conducted from August 29, 2024, to September 10, 2024. Regression coefficients (solid points) and 95% CIs (horizontal lines) are shown for demographic, occupational, personality, and health-related predictors. The dashed vertical line indicates a coefficient of zero (no association). Categorical variables were coded as follows: physical health problems, psychiatric history, prior psychological treatment, and financial burden (1="no" and 2="yes") and only-child status (1="yes" and 2="no").

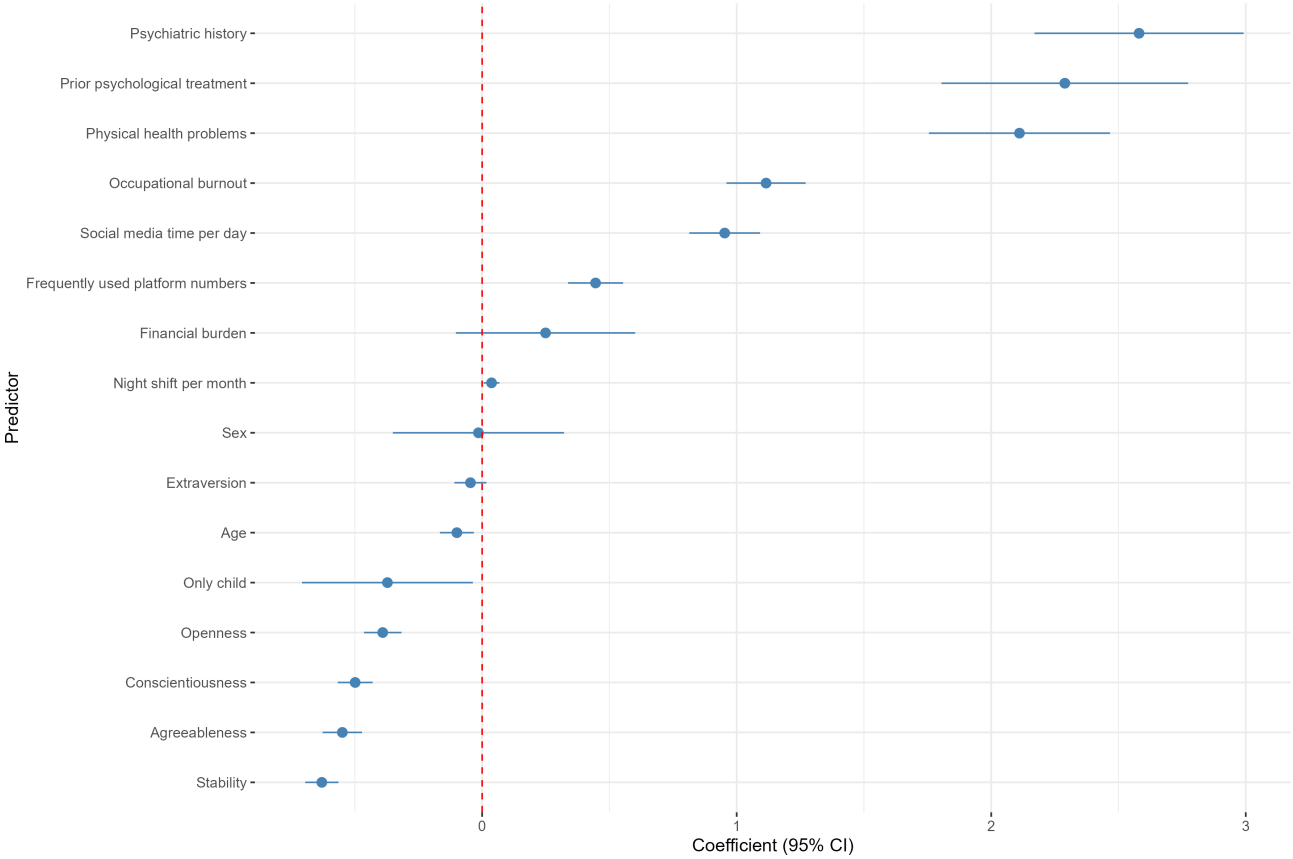

When using a BSMAS score of 24 or more to define SMA, only a subset of predictors remained significant in the logistic regression model (Figure 3; Table S11 in Multimedia Appendix 1). Specifically, higher occupational burnout was significantly associated with increased odds of developing SMA (odds ratio [OR] 1.409, 95% CI 1.226 - 1.617; *P*<.001). Similarly, longer daily social media use (OR 1.388, 95% CI 1.233 - 1.562; *P*<.001), physical health problems (OR 1.564, 95% CI 1.133 - 2.157; *P*=.006), and a positive psychiatric history (OR 2.000, 95% CI 1.406 - 2.838; *P*<.001) were all associated with an elevated risk of addiction. In contrast, higher conscientiousness showed a protective effect (OR 0.922, 95% CI 0.860 - 0.989; *P*=.02).

XSL•FO
RenderX

Figure 3. Forest plot of multivariable logistic regression predictors of probable social media addiction (Bergen Social Media Addiction Scale [BSMAS] ≥24) among Chinese medical trainees (N=3621). Data were derived from a nationwide cross-sectional anonymous online survey conducted from August 29, 2024, to September 10, 2024. Odds ratios (ORs; solid points) and 95% CIs (horizontal lines) are displayed for each predictor. The dashed vertical line denotes a null effect (OR 1.00). Categorical variables were coded as follows: physical health problems, psychiatric history, prior psychological treatment, and financial burden (1="no" and 2="yes") and only-child status (1="yes" and 2="no").

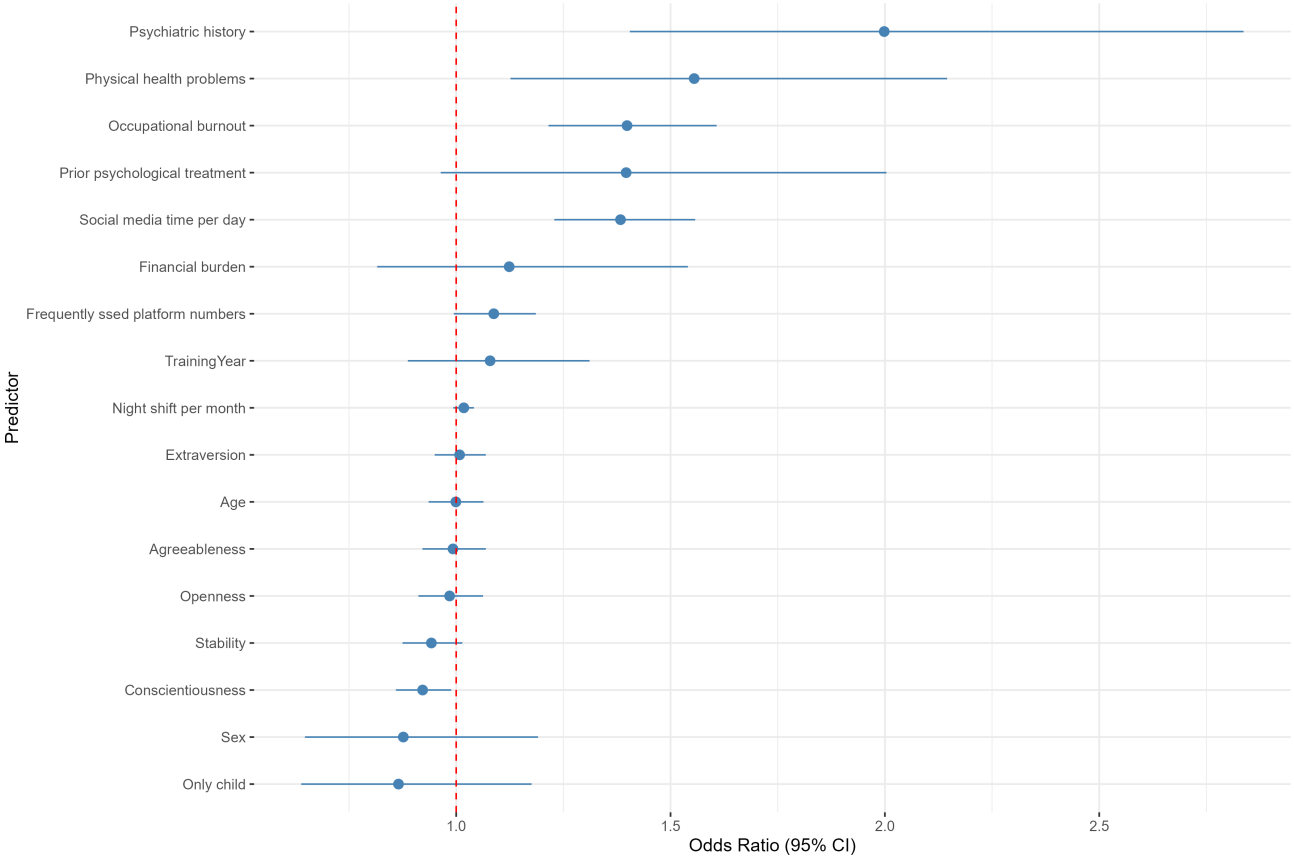

Mediation Analysis

Occupational burnout was positively correlated with both psychiatric history (β=.809; *P*<.001) and physical health problems (β=.681; *P*<.001) and was negatively associated with conscientiousness (β=–.119; *P*<.001). Mediation analysis revealed that burnout significantly mediated the effects of psychiatric history and physical health problems on BSMAS scores, with mediation proportions of 28.1% (indirect effect β=.726; bootstrap standard error [BootSE]=0.090; *P*<.001) and 29.6% (indirect effect β=.626; BootSE=0.074; *P*<.001), respectively. Furthermore, the protective effect of conscientiousness on BSMAS scores was partially mediated by burnout, contributing a negative mediation effect with a proportion of 20.9% (indirect effect β=–.104; BootSE=0.013; *P*<.001).

In contrast, daily social media use time was not significantly associated with psychiatric history or physical health problems but was negatively correlated with conscientiousness (β=–.023; *P*=.008). Mediation analysis indicated that daily social media use time modestly mediated the association between conscientiousness and BSMAS scores, with a mediation proportion of 4.1% (indirect effect β=–.021; BootSE=0.008; *P*=.009). For more detailed results, refer to Figure 4 and Table S12 in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 4. Parallel mediation models evaluating indirect associations between health/personality factors and social media addiction severity via occupational burnout and daily social media use duration among Chinese medical trainees (N=3621). Data were derived from a nationwide cross-sectional anonymous online survey conducted from August 29, 2024, to September 10, 2024. The models illustrate how psychiatric history, physical health problems, and conscientiousness are associated with Bergen Social Media Addiction Scale scores through two mediators: (1) occupational burnout and (2) daily social media use duration. Values on the paths represent standardized regression coefficients, with significance levels denoted by *$P<.05$, **$P<.01$, ***$P<.001$.

XSL•FO
RenderX

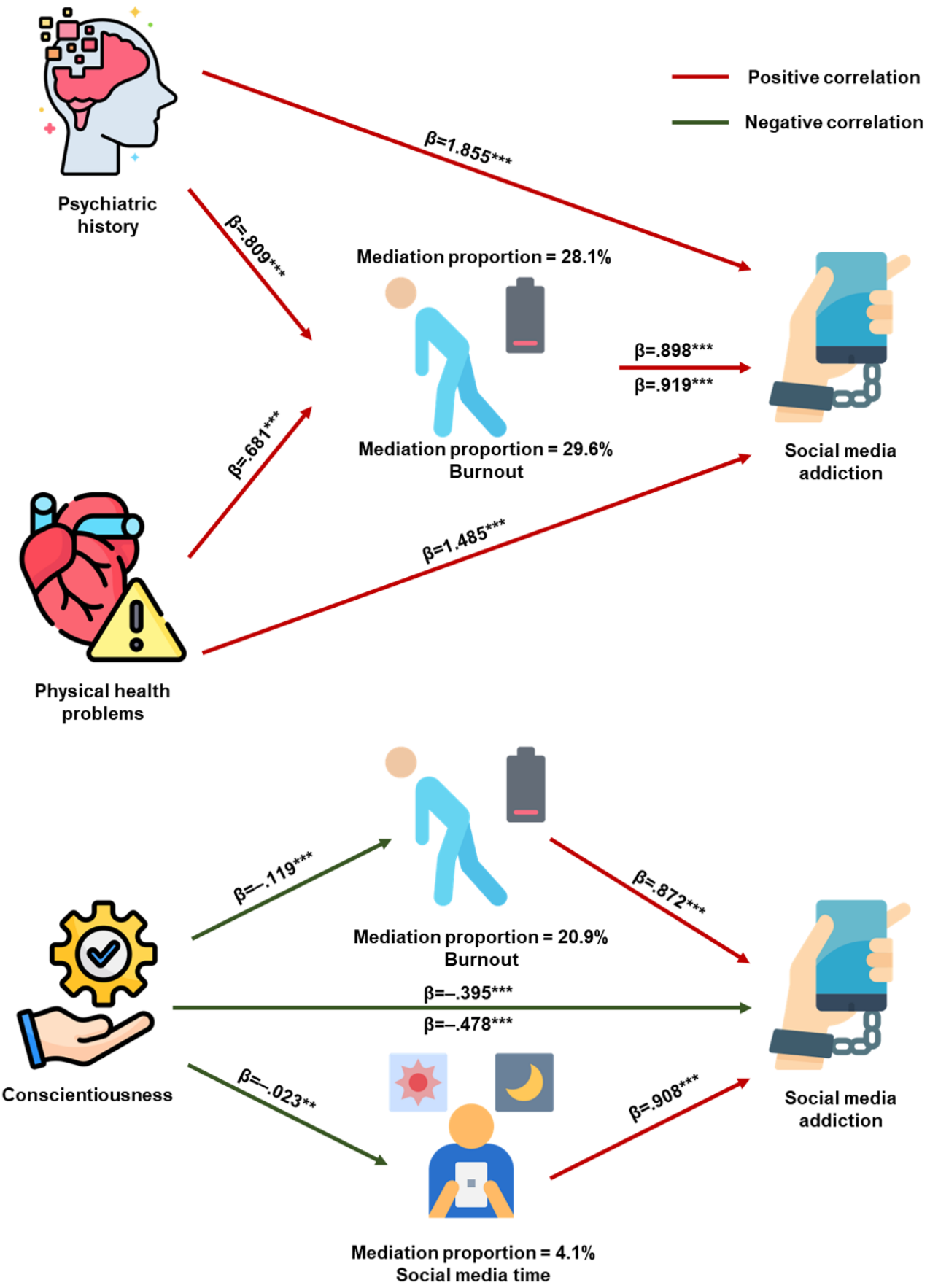

XSL•FO
RenderX

Moderation Analysis

Moderation analyses revealed significant interaction effects in both models (Table S13 and Figure S4 in Multimedia Appendix 1). Burnout significantly moderated the association between psychiatric history and BSMAS scores (β=−.795; *P*<.001), such that the effect of psychiatric history weakened at higher levels of burnout; correspondingly, the proportion of SMA increased across burnout levels among individuals with a psychiatric history (eg, 14.4%, 10.3%, 15%, 17%, and 19.6%). A similar moderating pattern was observed for physical health problems (β=−.513; *P*=.002), with SMA prevalence likewise increasing across burnout levels among affected individuals (eg, 8.5%, 7.8%, 11.4%, 14%, and 16.1%). No other interaction effects were statistically significant.

Stratified Analysis

In stratified analyses by training year, occupational burnout consistently predicted higher odds of SMA across all subgroups (year 1: OR 1.327, 95% CI 1.112 - 1.582, *P*=.002; year 2: OR 1.442, 95% CI 1.201 - 1.729, *P*<.001; year 3: OR 1.391, 95% CI 1.123 - 1.722, *P*=.002), fully aligning with the overall sample. Physical health problems were associated with addiction only in year 2 medical trainees (OR 1.675, 95% CI 1.154 - 2.429; *P*=.007). Psychiatric history showed significant effects in both year 2 (OR 1.675, 95% CI 1.154 - 2.429; *P*=.007) and year 3 medical trainees (OR 1.586, 95% CI 1.039 - 2.420; *P*=.03) but not in year 1 medical trainees. Daily social media use time was not significant in the stratified models (all *P*>.100). Conscientiousness was protective only in year 1 (OR 0.918, 95% CI 0.850 - 0.991; *P*=.03), while showing null associations in years 2 and 3. Beyond these 5 primary predictors, night-shift frequency was an additional significant factor in year 2 medical trainees (OR 1.041, 95% CI 1.008 - 1.076; *P*=.02), suggesting a unique role of workload intensity at this training stage. For more detailed results, refer to Tables S14-S16 in Multimedia Appendix 1.

Discussion

Principal Findings

In this large cross-sectional survey of 3621 medical trainees in China, we found that the prevalence of SMA (BSMAS ≥24) was 5.8% (211/3621), with the highest proportion among second-year medical trainees (92/1159, 7.9%) and among those using social media for more than 4 hours per day (12.2%). Higher burnout, longer daily use, physical health problems, and psychiatric history significantly increased the odds of developing SMA, whereas conscientiousness had a protective effect. In addition, we observed U-shaped associations between daily social media use and occupational burnout, physical health problems, psychiatric history, and mental health outcomes (depression, anxiety, and insomnia), all of which were linked to less favorable outcomes than moderate use. Exploratory analyses further suggested that burnout and time spent on social media functioned as key pathways linking health vulnerabilities and personality traits to SMA.

SMA Prevalence and Its Relationship With Mental Health Outcomes

Meta-analytic evidence using the BSMAS with a clinical cutoff score of 24 indicates a pooled global prevalence of SMA of approximately 8% (95% CI 4% - 12%) [18]. In this study, the prevalence of SMA among medical trainees was 5.8%. Although numerically lower than the pooled global estimate, this rate remains clinically concerning given the professional demands and vulnerability of this population. Notably, this prevalence was substantially higher than the 3.49% reported among Chinese adolescents when identical diagnostic criteria were applied [17]. Furthermore, we observed that the prevalence of SMA doubled or even tripled as daily usage transitioned from moderate to excessive levels (3.7% - 6.8% for ≤4 h vs 12.2% for >4 h). This nonlinear escalation suggests that excessive social media use represents a critical tipping point at which behavioral engagement shifts from routine activity to clinically significant addiction. Importantly, irrespective of the threshold or classification approach, SMA has been consistently linked to adverse mental health outcomes. Recent longitudinal studies and meta-analyses have demonstrated robust associations between SMA and impaired concentration, chronic sleep disturbances (particularly insomnia), and a markedly increased risk of depression and suicidal ideation [22,23]. Our findings further corroborate these associations, providing additional evidence that SMA is closely related to poor mental health outcomes among medical trainees.

U-Shaped Associations Between Social Media Use and Mental Health Outcomes

Social media usage has long been considered a potential correlate of mental health problems [3]. Based on large-scale survey data from Chinese medical trainees, both excessive and minimal social media use were closely associated with higher levels of burnout, depression, anxiety, and insomnia. Notably, the ≤1-hour usage group reported greater financial burden and more frequent night shifts, suggesting substantial occupational constraints. These U-shaped patterns may reflect 2 behavioral phenotypes under occupational stress: minimal use associated with withdrawal or disengagement, and excessive use potentially reflecting emotion-focused or avoidant coping in response to distress. Both patterns were linked to poorer mental health outcomes than moderate use, warranting longitudinal validation. These findings argue against a simple linear relationship between the duration of social media use and psychological and physical health outcomes.

Occupational Burnout as a Pathway and Amplifier in SMA

In both the overall sample and the stratified analyses, occupational burnout consistently emerged as a significant predictor of SMA. Recognizing burnout as a modifiable risk factor creates opportunities for targeted interventions in medical training environments [24,25]. Furthermore, our findings indicate that burnout functions as a mediating factor, amplifying the adverse effects of psychiatric history and physical health problems on SMA while attenuating the protective role of conscientiousness. This dual role underscores its centrality as

XSL•FO
RenderX

a target for interventions aimed at reducing digital overreliance and safeguarding the mental health of medical trainees [26].

The association between burnout and SMA may be driven by a dysregulated neurobiological stress response involving the hypothalamic-pituitary-adrenal axis and the mesocorticolimbic reward system [27]. Chronic stress related to burnout can create prolonged hypothalamic-pituitary-adrenal axis hyperactivity, eventually leading to neuroendocrine exhaustion and an "antireward" brain state [28]. In this state, natural rewards (eg, clinical success) become less satisfying, and the brain becomes sensitized to dopamine surges from digital microrewards such as "likes" and notifications. Consequently, burned-out medical trainees may engage in compulsive social media use as a maladaptive form of "self-medication" to counteract anhedonia and emotional numbness, creating a pathological cycle of addiction [29].

Interpretation of Stratified Findings

Stratified analyses confirmed that burnout remained a consistent predictor of SMA across all residency years. In contrast, other predictors showed stage-specific effects. Physical health problems and psychiatric history were significant only in years 2 and 3, suggesting cumulative vulnerability to prolonged stress exposure. Furthermore, personality traits played a pivotal role in mental health [30]; specifically, conscientiousness was protective only in year 1, indicating an early buffering effect. Night-shift frequency uniquely predicted addiction in year 2, highlighting this stage as a critical inflection point at which workload and circadian disruption may increase risk.

Comparison With Prior Studies and Cross-Cultural Perspectives

Our findings are broadly consistent with and extend evidence from other cultural settings. For example, studies in Saudi Arabia have linked SMA among medical students to higher anxiety, poorer academic performance, and greater burnout [31,32]. Tunisian data indicate associations with reduced self-esteem [33], whereas Ugandan data show strong links between problematic digital use and depression [34]. Multicountry research has further confirmed that increased social media use is associated with psychological distress and impaired academic outcomes through maladaptive coping [35]. Collectively, these studies suggest that the interplay between social media use, addiction risk, and burnout transcends cultural boundaries, reinforcing burnout as a universal intervention target while underscoring the influence of contextual factors, such as training demands and social support.

Limitations

First, the cross-sectional design prevents the establishment of definitive cause-and-effect relationships. Although our findings suggest that burnout increases the risk of SMA development, a bidirectional relationship is also possible. Future research should use longitudinal designs to determine causality more precisely. Second, the data were based on self-reports, which are susceptible to recall and social desirability biases and may have affected the accuracy of the prevalence and usage findings. Future studies could enhance data accuracy by incorporating objective measures of social media use. Third, the focus on Chinese medical trainees restricts the generalizability of the findings to other populations. Social and cultural backgrounds, health care systems, and social media usage habits may vary significantly across countries. Future studies should conduct cross-cultural comparisons to validate these findings.

Future Directions

To build on the foundation of this study, future research should adopt longitudinal designs to establish causal relationships among occupational burnout, physical health problems, and the development of SMA. Qualitative research could provide valuable context by exploring the lived experiences of both minimal and excessive social media users, helping to validate and enrich the "2 phenotypes" hypothesis. Intervention studies are needed to test the efficacy of burnout-focused programs in reducing the prevalence and severity of SMA in this population. These studies would provide evidence of effective strategies for addressing this emerging health issue.

Conclusions

This study provides a systematic characterization of SMA among Chinese medical trainees. Occupational burnout emerged as the most consistent correlate of SMA, operating both as an independent risk factor and as a pathway that amplifies the effects of psychiatric history, physical health problems, and personality traits. A U-shaped association was identified between the duration of social media use and adverse mental health outcomes, with both minimal and excessive use associated with less favorable profiles than moderate use, suggesting distinct but maladaptive patterns of engagement under occupational stress. Stratified analyses further indicated stage-specific vulnerabilities across residency years. Together, these findings suggest that SMA in medical trainees reflects the interaction among systemic workload stressors, individual vulnerability, and behavioral coping patterns. Accordingly, interventions should move beyond monitoring screen time and prioritize burnout reduction and supportive training environments to mitigate addiction risk.

Acknowledgments

The authors express their sincere appreciation to all participants for their valuable contributions. Generative artificial intelligence (Gemini 3) was used in this study solely for language editing.

Funding

This study was supported by the Bejing High-Level Innovation and Entrepreneurship Talent Support Program (G202532215; XZ) and the Tianjin Health Research Project (TJWJ2025MS023; GL).

XSL•FO
RenderX

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

All authors conducted the study and take full responsibility for the study design, data analysis, interpretation of findings, and all final content decisions. ZG, XZ, NT, and GL conceived and designed the study. XZ and NT contributed to data acquisition. NT, XZ, and GL supervised the entire project. ZG drafted the manuscript. ZG and XZ performed the methodology, statistical analyses, and data interpretation, and provided key intellectual contributions.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional documentation and detailed statistical evidence supporting the study findings.
[DOCX File, 517 KB - jmir_v28i1e75675_app1.docx]

Checklist 1
CHERRIES checklist.
[DOCX File, 18 KB - jmir_v28i1e75675_app2.docx]

References

1. Clavier T, Chevalier E, Demailly Z, Veber B, Messaadi IA, Popoff B. Social media usage for medical education and smartphone addiction among medical students: national web-based survey. JMIR Med Educ 2024 Oct 22;10:e55149. [doi: 10.2196/55149] [Medline: 39437450]
2. Guckian J, Utukuri M, Asif A, et al. Social media in undergraduate medical education: a systematic review. Med Educ 2021 Nov;55(11):1227-1241. [doi: 10.1111/medu.14567] [Medline: 33988867]
3. Santos RMS, Mendes CG, Sen Bressani GY, et al. The associations between screen time and mental health in adolescents: a systematic review. BMC Psychol 2023 Apr 20;11(1):127. [doi: 10.1186/s40359-023-01166-7] [Medline: 37081557]
4. Griffiths MD, Kuss DJ, Demetrovics Z. Social networking addiction: an overview of preliminary findings. In: Behavioral Addictions: Criteria, Evidence, and Treatment: Academic Press; 2014:119-141. [doi: 10.1016/B978-0-12-407724-9.00006-9]
5. Schou Andreassen C, Billieux J, Griffiths MD, et al. The relationship between addictive use of social media and video games and symptoms of psychiatric disorders: a large-scale cross-sectional study. Psychol Addict Behav 2016 Mar;30(2):252-262. [doi: 10.1037/adb0000160] [Medline: 26999354]
6. Feng S, Li X, Huang Z, et al. The relationship between burnout and sense of school belonging among the resident physicians in the standardization training in China. Med Educ Online 2024 Dec 31;29(1):2343515. [doi: 10.1080/10872981.2024.2343515] [Medline: 38660991]
7. Wei L, Mei Z, Jian C, et al. Unpacking social media usage divides in Mainland China: the interplay of demographics and personality across six user archetypes. Int J Soc Sci Lang Linguist 2025;05(1):01-08. [doi: 10.55640/ijssll-05-01-01]
8. Ishak WW, Lederer S, Mandili C, et al. Burnout during residency training: a literature review. J Grad Med Educ 2009 Dec;1(2):236-242. [doi: 10.4300/JGME-D-09-00054.1] [Medline: 21975985]
9. Suliman S, Allen M, Chivese T, de Rijk AE, Koopmans R, Könings KD. Is medical training solely to blame? Generational influences on the mental health of our medical trainees. Med Educ Online 2024 Dec 31;29(1):2329404. [doi: 10.1080/10872981.2024.2329404] [Medline: 38488138]
10. Feng T, Narayanan S. Understanding stress, burnout, and behavioral patterns in medical residents using large-scale longitudinal wearable recordings. Presented at: 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); Jul 15-19, 2024. [doi: 10.1109/EMBC53108.2024.10782414]
11. Rodrigues H, Cobucci R, Oliveira A, et al. Burnout syndrome among medical residents: a systematic review and meta-analysis. PLoS One 2018;13(11):e0206840. [doi: 10.1371/journal.pone.0206840] [Medline: 30418984]
12. He Y, Lin S, Wang Y, et al. A survey on mental health among resident physicians: psychological resilience as a mediator. BMC Psychiatry 2025 Jan 31;25(1):87. [doi: 10.1186/s12888-025-06517-9] [Medline: 39891091]
13. Muhamad NA, Maamor NH, et al. Prevalence and associated risk factors of self-harm among health care workers: protocol for systematic review and meta-analysis. JMIR Res Protoc 2025 Sep 17;14:e67059. [doi: 10.2196/67059] [Medline: 40961495]
14. Guo X, Gong S, Chen Y, et al. Weight and lifestyle behavior changes in Chinese health care workers during the COVID-19 pandemic: 3-year retrospective survey. Interact J Med Res 2024 Dec 10;13:e50754. [doi: 10.2196/50754] [Medline: 39657182]

XSL•FO
RenderX

15. Nagarajan R, Ramachandran P, Dilipkumar R, Kaur P. Global estimate of burnout among the public health workforce: a systematic review and meta-analysis. Hum Resour Health 2024 May 21;22(1):30. [doi: 10.1186/s12960-024-00917-w] [Medline: 38773482]
16. Chen IH, Ahorsu DK, Pakpour AH, Griffiths MD, Lin CY, Chen CY. Psychometric properties of three simplified Chinese online-related addictive behavior instruments among Mainland Chinese primary school students. Front Psychiatry 2020;11(875):875. [doi: 10.3389/fpsyt.2020.00875] [Medline: 33101070]
17. Luo T, Qin L, Cheng L, et al. Determination the cut-off point for the Bergen Social Media Addiction (BSMAS): diagnostic contribution of the six criteria of the components model of addiction for social media disorder. J Behav Addict 2021 Jul 15;10(2):281-290. [doi: 10.1556/2006.2021.00025] [Medline: 34010148]
18. Cheng C, Lau YC, Chan L, Luk JW. Prevalence of social media addiction across 32 nations: meta-analysis with subgroup analysis of classification schemes and cultural values. Addict Behav 2021 Jun;117:106845. [doi: 10.1016/j.addbeh.2021.106845] [Medline: 33550200]
19. Thørrisen MM, Sadeghi T. The Ten-Item Personality Inventory (TIPI): a scoping review of versions, translations and psychometric properties. Front Psychol 2023;14:1202953. [doi: 10.3389/fpsyg.2023.1202953] [Medline: 37434881]
20. Brand M, Young KS, Laier C, Wölfling K, Potenza MN. Integrating psychological and neurobiological considerations regarding the development and maintenance of specific Internet-use disorders: an Interaction of Person-Affect-Cognition-Execution (I-PACE) model. Neurosci Biobehav Rev 2016 Dec;71:252-266. [doi: 10.1016/j.neubiorev.2016.08.033] [Medline: 27590829]
21. Brand M, Wegmann E, Stark R, et al. The Interaction of Person-Affect-Cognition-Execution (I-PACE) model for addictive behaviors: update, generalization to addictive behaviors beyond internet-use disorders, and specification of the process character of addictive behaviors. Neurosci Biobehav Rev 2019 Sep;104:1-10. [doi: 10.1016/j.neubiorev.2019.06.032] [Medline: 31247240]
22. Xiao Y, Meng Y, Brown TT, Keyes KM, Mann JJ. Addictive screen use trajectories and suicidal behaviors, suicidal ideation, and mental health in US youths. JAMA 2025 Jul 15;334(3):219-228. [doi: 10.1001/jama.2025.7829] [Medline: 40531519]
23. Banodha H, Saini P. Consequences of social media addiction: a triadic perspective on mental, social, and physical health. Int J Indian Psychol 2025 Sep;13(3). [doi: 10.25215/1303.373]
24. Liu C, Ma J. Social media addiction and burnout: the mediating roles of envy and social media use anxiety. Curr Psychol 2020;39(6):1883-1891. [doi: 10.1007/s12144-018-9998-0]
25. Prins JT, Gazendam-Donofrio SM, Tubben BJ, van der Heijden FMMA, van de Wiel HBM, Hoekstra-Weebers JEHM. Burnout in medical residents: a review. Med Educ 2007 Aug;41(8):788-800. [doi: 10.1111/j.1365-2923.2007.02797.x] [Medline: 17661887]
26. Bonsaksen T, Kleppang AL. Social media and mental health: lessons learned from the psychology research and behavior management article collection. Psychol Res Behav Manag 2025;18:2039-2052. [doi: 10.2147/PRBM.S549588] [Medline: 41017847]
27. McEwen BS. Physiology and neurobiology of stress and adaptation: central role of the brain. Physiol Rev 2007 Jul;87(3):873-904. [doi: 10.1152/physrev.00041.2006] [Medline: 17615391]
28. Nunez SG, Rabelo SP, Subotic N, Caruso JW, Knezevic NN. Chronic stress and autoimmunity: the role of HPA axis and cortisol dysregulation. Int J Mol Sci 2025 Oct 14;26(20):9994. [doi: 10.3390/ijms26209994] [Medline: 41155288]
29. Blum K, Bowirrat A, Braverman ER, et al. Reward deficiency syndrome (RDS): A cytoarchitectural common neurobiological trait of all addictions. Int J Environ Res Public Health 2021 Nov 2;18(21):11529. [doi: 10.3390/ijerph182111529] [Medline: 34770047]
30. Zhang X, Yan H, Yu H, et al. Childhood urbanicity interacts with polygenic risk for depression to affect stress-related medial prefrontal function. Transl Psychiatry 2021 Oct 12;11(1):522. [doi: 10.1038/s41398-021-01650-x] [Medline: 34642305]
31. Alfaya MA, Abdullah NS, Alshahrani NZ, et al. Prevalence and determinants of social media addiction among medical students in a selected university in Saudi Arabia: a cross-sectional study. Healthcare (Basel) 2023 May 10;11(10):1370. [doi: 10.3390/healthcare11101370] [Medline: 37239655]
32. Badri HM, Aboalshamat KT, Abdouh IM, et al. The effect of social media addiction on burnout among health-care students and professionals in Saudi Arabia. Int J Acad Med 2023;9(1):11-17. [doi: 10.4103/ijam.ijam_72_22]
33. Guermazi F, Abid W, Baati I, et al. Social media addiction and personality dimensions among Tunisian medical students. Front Psychiatry 2024;15:1471425. [doi: 10.3389/fpsyt.2024.1471425] [Medline: 39351331]
34. Sserunkuuma J, Kaggwa MM, Muwanguzi M, et al. Problematic use of the internet, smartphones, and social media among medical students and relationship with depression: an exploratory study. PLoS One 2023;18(5):e0286424. [doi: 10.1371/journal.pone.0286424] [Medline: 37235547]
35. Shiraly R, Roshanfekr A, Asadollahi A, Griffiths MD. Psychological distress, social media use, and academic performance of medical students: the mediating role of coping style. BMC Med Educ 2024 Sep 13;24(1):999. [doi: 10.1186/s12909-024-05988-w] [Medline: 39272179]

XSL•FO
RenderX

Abbreviations

BootSE: bootstrap standard error
BSMAS: Bergen Social Media Addiction Scale
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
OR: odds ratio
SMA: social media addiction
SRT: standardized residency training

Edited by A Mavragani, N Cahill; submitted 08.Apr.2025; peer-reviewed by H Yu, V Dauletova; revised version received 28.Mar.2026; accepted 30.Mar.2026; published 04.May.2026.

Please cite as:
Guan Z, Tang N, Luo G, Zhang X
Prevalence of Social Media Addiction and Associations With Usage Patterns, Burnout, and Health Conditions Among Medical Trainees in China: Cross-Sectional Study
J Med Internet Res 2026;28:e75675
URL: https://www.jmir.org/2026/1/e75675
doi:10.2196/75675

© Zexu Guan, Ni Tang, Guoshuai Luo, Xiao Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 4.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Health Literacy, Technology Acceptance, and Competence Among Older Adults Aged ≥65 Years: Cross-Sectional Study Investigating Differences Between Women and Men

Franziska Ulrike Jung[1], Dr rer nat; Melanie Luppa[1], Dr rer med; Matthias Reusche[2,3], Dipl-Inf; Kerstin Wirkner[3], PD Dr; Melanie Eberl[3]; Yvonne Dietz[3]; Christoph Engel[2,3], PD Dr; Steffi G Riedel-Heller[1], MPH, MD, Prof Dr

[1]Institute of Social Medicine, Occupational Health and Public Health (ISAP), Faculty of Medicine, Leipzig University, Philipp-Rosenthal-Str. 55, Leipzig, Germany

[2]Institute for Medical Informatics, Statistics and Epidemiology, Leipzig University, Leipzig, Germany

[3]LIFE—Leipzig Research Centre for Civilization Diseases, Leipzig University, Leipzig, Germany

Corresponding Author:
Franziska Ulrike Jung, Dr rer nat
Institute of Social Medicine, Occupational Health and Public Health (ISAP), Faculty of Medicine, Leipzig University, Philipp-Rosenthal-Str. 55, Leipzig, Germany

Abstract

Background: Digital health literacy (DHL) has the potential to improve health among older adults by enhancing access to health-related information and health care services.

Objective: The aim of this study was to analyze the relationship between DHL and technology commitment in adults aged 65 years and older, while also investigating possible gender differences.

Methods: The analytical sample consisted of 1824 individuals. The analysis included descriptive comparisons in terms of DHL, technology acceptance, competency, support, and internet use. Multivariate regression models (generalized linear models) were applied in order to test the association between DHL and technology commitment, controlling for internet use as well as health-related and sociodemographic characteristics.

Results: Male and female participants did not differ in terms of DHL (mean score: 3.5, SD 1.2 [men] and 3.5, SD 1.3 [women]; P=.70); however, male participants reported significantly higher technology acceptance (P<.001) and higher technology competencies (P<.001), but less support with regard to technology use (P<.001). Within regression models, only higher technology acceptance (coefficient=0.023, 95% CI 0.006 - 0.041; P=.01) and support (coefficient=0.027, 95% CI 0.014 - 0.040; P<.001) were significantly linked to greater DHL. The subgroup analysis revealed that DHL was significantly associated with technology acceptance among men (coefficient=0.036, 95% CI 0.012 - 0.060; P=.003) but not women (coefficient=0.024, 95% CI 0.008 - 0.040; P=.44).

Conclusions: According to the current results, DHL is highly related to technology commitment. Gender differences should be taken into account when developing and evaluating appropriate interventions to improve DHL by addressing the acceptance of technologies and optimizing support infrastructures.

(J Med Internet Res 2026;28:e85846) doi:10.2196/85846

KEYWORDS

digital health literacy; technology; acceptance; competence; older adults

Introduction

The association between digital health literacy (DHL) and technology use or acceptance among older adults is a critical area of research, given the increasing reliance on digital platforms for health information and services. This relationship reflects how DHL informs technology acceptance, as well as technology competencies, and facilitates a greater use of health technologies.

Previous research highlights the importance of internet access and participation in order to improve health outcomes and prevent health care disparity [1,2]. A study investigating internet use in old age in 2016 has found that only 27% reported using the internet on a daily basis and 57% never used the internet [3]. However, the increasing proportion of internet users aged 65 years and older is a well-documented demographic trend [4]. Multiple studies found evidence for the gradual rise of internet engagement among the senior population, driven by factors including health-related motivations, social connectivity, and technological advancements. In this context, 81% (62/242) of

XSL•FO
RenderX

internet users aged 60 years and above use the internet for health- or disease-related searches [5].

In order to successfully apply health-related information in decision-making and health behavior, it is important to understand and critically evaluate digital information and services. Although there is now a clear definition, DHL is often described as the skills required to seek, understand, and use health information from digital sources to make health-related decisions [6-8]. According to a recent systematic review, higher levels of DHL among older adults were related to being younger, a better health status, a higher socioeconomic status, higher education levels, and higher income [9]. In addition, social connectivity and participation, as well as positive attitudes toward internet and digital health information, have also been associated with higher digital literacy [9]. Especially among the older adults, age-related limitations with regard to vision or hearing may influence the motivation to adopt digital technologies and therefore influence DHL [10,11]. According to a study by Liu et al, perceived usefulness, ease of use, and reliability of digital health information may not only impact the DHL of older adults. Those who view these factors positively are also more likely to use digital health services or devices [12]. This relationship is particularly rooted in the technology acceptance model (TAM) [13], which posits that users' attitudes toward technology may stem from perceived attributes and personal needs [12,14]. Similarly, Wei and Zhou [15] found that higher DHL correlates with more frequent engagement in health-promoting activities, suggesting that improved literacy not only facilitates the use of technology but also leads to effective integration into health management practices.

Moreover, the willingness of older adults to engage with digital health resources has been shown to be influenced by the degree of technology acceptance or commitment. Kim et al [16] reported that older adults' intentions to use digital health applications are significantly shaped by their attitudes toward new technologies, arguing that negative perceptions may hinder technology adoption. The fear of using complex technologies often leads to resistance, compounded by anxiety about using these tools. In this context, higher anxiety levels correlated with lower eHealth literacy, as indicated by Kim et al [16]. These findings align with those of Rasekaba et al [17], who highlighted the impact of low digital literacy on older adults' ability to engage with such technologies due to less confidence and limited access, consequently creating a digital divide that can affect their health management. However, technology acceptance, competencies, and support are understudied mechanisms, which link technology and DHL in older adults. While technology use and acceptance are often treated as prerequisites for digital engagement, DHL represents a distinct, domain-specific construct that extends beyond general digital skills or device use. DHL requires the ability to locate, critically evaluate, and apply health information in digital environments—competencies that are not necessarily captured by measures of technology adoption or acceptance alone.

The exploration of gender-specific findings in the context of DHL and technology acceptance among older adults reveals several nuances that warrant consideration. Research indicates that gender differences exist, influencing both levels of DHL and the willingness to engage with technology. According to a recent review, older women had lower eHealth literacy compared to older men [18]. However, the authors suggest that the gender-related disparity may gradually diminish and, in addition, may be explained by differences regarding (lower) socioeconomic status. This emphasizes the complexity of social determinants affecting DHL and suggests that factors such as income, education, or social support may also be important, by influencing access to digital health resources [19].

The aim of this study was to investigate the relationship between DHL and technology use and acceptance among individuals aged 65 years and older. Importantly, rather than assuming gender differences in DHL itself, this study investigates whether the factors associated with DHL differ by gender, thereby addressing inconsistencies in the existing literature on adults aged 65 years and older. Based on previous studies, we hypothesize that a higher degree of acceptance and higher competencies as well as support in relation to technology use may be related to higher DHL. Furthermore, we aim to explore possible gender differences using the subgroup analysis. In this context, the impact of sociodemographic characteristics will be taken into account.

Methods

Recruitment

The current data analysis is based on a cross-sectional study on DHL among old and oldest people in Saxony (Germany). For the purpose of recruitment, 3000 participants from the participant pool of the LIFE (Leipzig Research Centre for Civilization Diseases) Adult Cohort were randomly chosen, recontacted, and surveyed using a paper-and-pencil questionnaire. Overall, a sample of 2227 returned the questionnaire, representing a response rate of 74.2%. The only inclusion criteria were being at least 65 years old at the time of the survey (May 2024). The study and recruitment procedure of the overall LIFE Adult Study has been already provided elsewhere [20,21]. The LIFE Adult Study is a population-based cohort study and includes data of 10,000 randomly selected participants from the City of Leipzig (Germany). The objective of the overall LIFE Adult Study is to investigate prevalences, predispositions, and determining factors of major civilization diseases over time.

Ethical Considerations

The LIFE Adult Study was conducted in accordance with the Declaration of Helsinki, and ethical approval was granted by the Ethics Committee of the Medical Faculty of Leipzig University (approval numbers 263–2009 - 14122009, 263/09-ff, 201/17-ek). Each participant provided written informed consent. At the start of the survey, all participants were informed about the aim of the study, as well as data handling and protection. Participation was anonymous and voluntary, and withdrawal of consent was possible at any time during the study. Participants did not receive incentives for participating in this study. The data privacy and confidentiality was approved by the responsible data protection officer.

XSL•FO
RenderX

Study Design

Apart from sociodemographic and health-related information (ie, age, sex, income, and general health status) and DHL, the questionnaire assessed internet use, technology acceptance, and competency. The instruments and constructs are described in more detail as follows:

DHL was investigated using the 8-item revised German eHealth Literacy Scale (GR-eHEALS) [22]. Exploratory and confirmatory factor analyses in previous studies confirmed the validity and factorial structure of the GR-eHEALS [22,23]. The items have been described in greater detail elsewhere [22]. DHL is measured using a 5-point Likert scale ranging from 1 (*do not agree at all*) to 5 (*do fully agree*), with higher scores indicating greater DHL (Cronbach α 0.970). The scale consists of 4 items that assess how information is searched on the internet (example item: "I know where I can find helpful health information on the internet"). The other 4 items focus on how the information found may be evaluated (example item: "I can distinguish between trustworthy and dubious websites with health information").

Eight items of the short scale on technology commitment were used to assess technology acceptance and self-perceived competence on a 5-point Likert scale [24]. A translation of the items and scale can be found in Multimedia Appendix 1. In this study, Cronbach α was 0.895 for the subscale self-perceived technology competence and 0.927 for the subscale technology acceptance. The scale has been used widely and validated among older age groups [25,26]. Furthermore, 3 items were used to assess social support in relation to technology use [27]. The items and the answer options can be found in Multimedia Appendix 1. Cronbach α was 0.954.

In order to survey internet usage, the following items were included in the survey: "how often they were using the internet?" (ranging from 0=never to 7=several times a day); "how familiar they are with using the internet" (ranging from 0=not very familiar to 3=very familiar); and "how often do you use the internet to search for information that may affect your health" (ranging from 0=never to 4=always) for both items. These items have been used in similar studies before [28].

In order to survey the sex, the answer categories, male or female, were provided. Educational attainment was assessed by sociodemographic items on the highest school qualification and highest job training qualification, resulting in low, intermediate, or high educational attainment, known as CASMIN (Comparative Analysis of Social Mobility in Industrial Nations) categories. The procedure and calculation of the CASMIN categories have been described elsewhere [29,30]. Family status was surveyed using the following 5 categories: single, married, divorced, widowed, and married, but separated. Household net income (monthly) was assessed using the following answer categories: <€1000, €1000 to <€1500, €1500 to <€2000, €2000 to <€2500, €2500 to <€3000, €3000 to <€3500, and more than €3500.

Perceived health status was surveyed using the EuroQol Visual Analogue Scale. Participants were asked to rate their current health status on a scale between 0 (*worst health status*) and 100 (*best health status*) [31]. In addition, participants were asked to rate their eyesight using the following question: "Do you have trouble seeing?" Possible answers included no impairment, mild impairment, significant impairment, and most serious impairment [32]. Similarly, they were asked to rate their hearing skills using the following item: "Do you have any trouble hearing?" (answer categories: no impairment, mild hearing loss, significant hearing loss, and extreme hearing loss/deafness) [32].

Individual competence expectations when dealing with difficulties in everyday life settings (self-efficacy) were measured using the ASKU (Allgemeine Selbstwirksamkeit Kurzskala) scale [33]. The instrument consists of 3 items and can be answered using a 5-point Likert scale. The items can be found in Multimedia Appendix 1. Cronbach α was 0.886.

Statistical Analysis

Overall, 3000 individuals were contacted by mail in which 2227 individuals replied. After removing cases with missing information, the analytical sample contained 1824 individuals. The analytical sample was significantly younger (P<.001; 75.3 y vs 77.9 y) and more highly educated (P<.001). However, no significant differences were found with regard to sex.

During a first step, differences between male and female participants were analyzed descriptively and compared using χ^2 tests, t tests, and Kruskal-Wallis tests, as appropriate. In addition, multivariate regression models were applied to investigate the relationship between DHL (outcome variable, continuous) and technology commitment and support (predictor variables) across the overall sample (model 1). Before the analysis, we assessed multicollinearity and tested for normality. No multicollinearity between variables was found. Due to the skewness of the outcome variable (DHL), generalized linear models with robust variance estimators were calculated (γ family, link-log [34]). Generalized linear models—regression models with γ distribution—are widely used in modeling continuous, nonnegative, and positive-skewed data. The subgroup analysis was conducted to investigate possible sex-related differences (model 2: female subsample; model 3: male subsample). All regression models were controlled for sociodemographic characteristics, relevant health parameters (health status and eyesight), self-efficacy, and aspects related to internet use. Covariates were chosen based on relevance for DHL or the relationship between DHL and internet use or technology commitment. Statistical analyses were performed using the software package STATA, version 16 (StataCorp), assuming an α level of .05 (2-tailed) as an indicator for statistical significance.

Results

Participants Characteristics

Overall, the majority in this sample (N=1824) was female (n=1016, 53%) and married (n=1247, 65%), and the mean age of the participants was 75 (SD 6.4) years. Further sociodemographic characteristics and differences between female and male participants are summarized in Table 1. In this study, men and women differed in terms of marital status

XSL•FO
RenderX

(P<.001), education (P<.001), income (P<.001), and overall health status (P=.02). No significant differences were found with regard to general self-efficacy. No significant differences between female and male individuals were found in terms of DHL. However, men reported using the internet more frequently and feeling more confident while using the internet (P<.001 for both).

XSL•FO
RenderX

Table . Characteristics of the overall study sample, the female subsample, and the male subsample (N=1824).

Characteristics	Overall (N=1824)	Women (n=973)	Men (n=851)	Significance (*P* value)
Age (y), mean (SD; range)	75.3 (6.4; 65 - 91)	74.7 (6.3; 65 - 91)	75.9 (6.4; 65 - 90)	<.001
Marital status, n (%)				<.001
Married	1.191 (65.3)	544 (56)	647 (76)	
Married, but separated	35 (1.9)	13 (1.3)	22 (2.6)	
Single	71 (3.9)	34 (3.5)	37 (4.3)	
Divorced	213 (11.7)	136 (14)	77 (9.1)	
Widowed	314 (17.2)	246 (25.3)	68 (8)	
Education (CASMIN[a]), n (%)				<.001
Low	187 (10.2)	103 (10.6)	84 (9.8)	
Intermediate	799 (43.9)	503 (51.8)	296 (34.9)	
High	838 (45.9)	367 (37.7)	471 (55.3)	
Household income (€), monthly, n (%)				<.001
<1499	206 (11.3)	125 (12.9)	81 (9.5)	
1500 - 1999	252 (13.8)	169 (17.4)	83 (9.8)	
2000 - 2499	396 (21.7)	224 (23)	172 (20.1)	
2500 - 2999	352 (19.3)	169 (17.4)	183 (21.6)	
3000 - 3499	294 (16.1)	144 (14.8)	150 (17.7)	
>3500	324 (17.7)	142 (14.6)	182 (21.3)	
Overall health status[b], mean (SD)	69.4 (18.6)	70.3 (18.4)	68.3 (18.8)	.02
Eyesight, n (%)				.67
No impairment	1.564 (85.7)	831 (85.3)	733 (86.2)	
Mild impairment	221 (12.1)	118 (12.1)	103 (12.1)	
Significant impairment	37 (2.1)	23 (2.5)	14 (1.6)	
Most serious impairment	2 (0.1)	1 (0.1)	1 (0.1)	
Hearing, n (%)				—[c]
No impairment	1.247 (68.4)	719 (73.9)	528 (62)	
Mild hearing loss	515 (28.2)	224 (23)	291 (34.2)	
Significant hearing loss	58 (3.2)	27 (2.8)	31 (3.6)	
Extreme hearing loss/deafness	4 (0.2)	3 (0.3)	1 (0.1)	
Digital health literacy[d], mean (SD)	3.5 (1.2)	3.5 (1.3)	3.5 (1.2)	.70
Self-efficacy[e], mean (SD)	12.4 (2.1)	12.4 (2.1)	12.4 (2.2)	.80
Internet use				
Frequency[f], mean (SD)	5.2 (2.4)	5 (2.5)	5.5 (2.3)	<.001
Confidence[g], mean (SD)	1.7 (0.9)	1.6 (0.9)	1.8 (0.9)	<.001
Health-related[h], mean (SD)	1.6 (1.2)	1.6 (1.2)	1.6 (1.1)	.05

[a]CASMIN: Comparative Analysis of Social Mobility in Industrial Nations.

[b]A higher score indicates better overall health, range: 0-100.

XSL•FO
RenderX

[c]Not available.

[d]A higher score indicates greater digital health literacy, range: 2-10.

[e]A higher score indicates greater self-efficacy, range: 3-15.

[f]A higher score indicates more frequent use, range: 0-7.

[g]A higher score indicates feeling more familiar using the internet, range: 0-3.

[h]A higher score indicates greater health-related internet use, range: 0-4.

Characteristics of technology use in terms of acceptance, competence, and social support have also been analyzed (Figure 1). In this context, male participants reported higher technology acceptance (*P*=.001), higher technology competencies (*P*<.001), but less support with regard to technology use (*P*<.001; Table 1).

Figure 1. Characteristics of technology use among the older adults (overall, women, and men, n=1824). Acceptance: the higher score indicates greater acceptance of technology; competence: the higher score indicates higher self-perceived competence with regard to technology use; support: the higher score indicates greater support with regard to technology use.

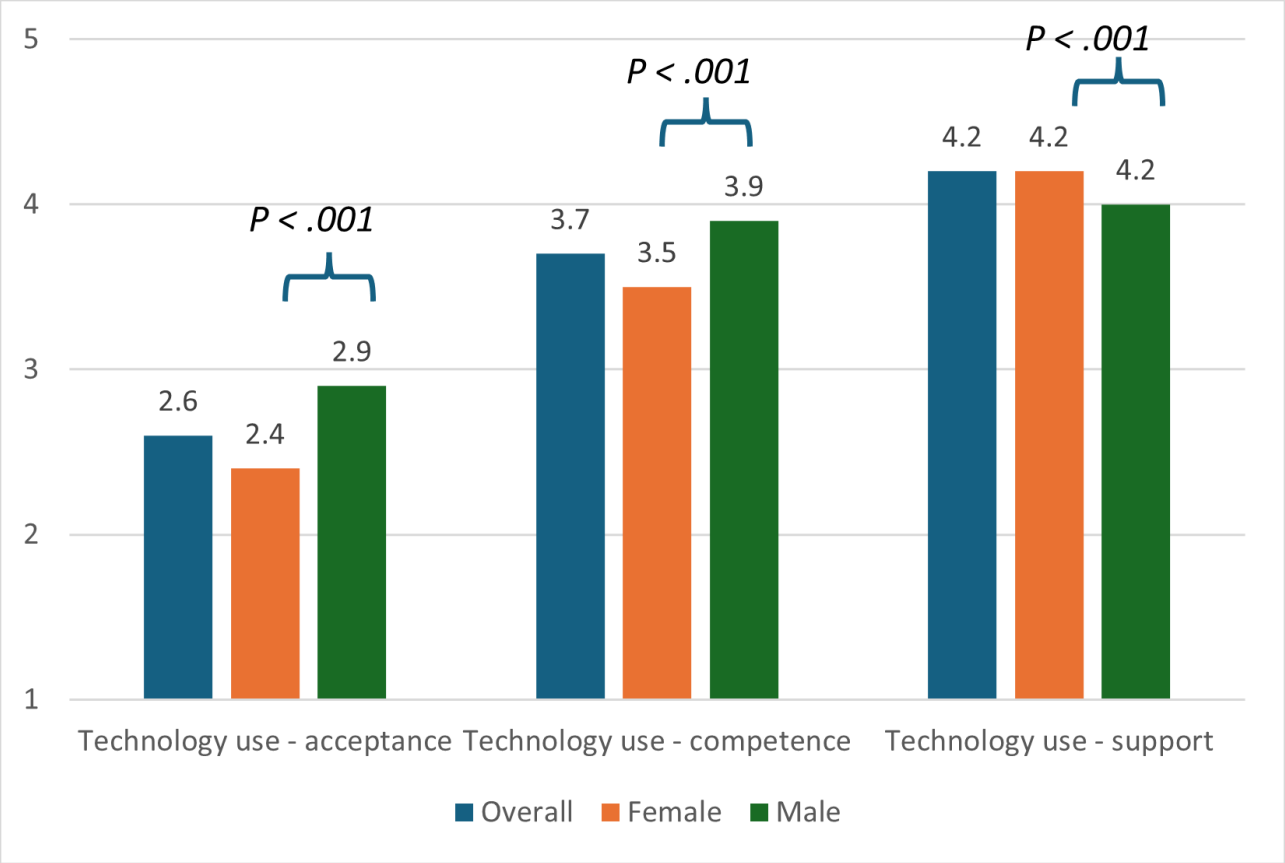

General Regression Analysis

In addition to descriptive analysis, regression analysis was conducted. Table 2 shows the results of the multivariate generalized regression model with DHL as the outcome variable (model 1). When controlling for sociodemographic and health-related variables, higher technology acceptance (coefficient=0.023, *P*=.01) and support were linked to greater DHL (coefficient=0.027, *P*<.001), indicating that motivational readiness and access to assistance play an important role in older adults' ability to navigate digital health information. Similarly, feeling more confident and greater frequency of internet use (in general and with regard to health-related aspects) were also significantly linked to better DHL (*P*<.001 for all). Interestingly, self-perceived technology competence was not related to DHL (coefficient=0.005, *P*=.61). This finding suggests that the self-assessment of overall technology competence captures a broader construct that may not be directly related to health-specific digital literacy skills within this study population. Taken together, the included predictors explained a substantial proportion of the variance in DHL (61% of 100% variance explained), underscoring the importance of motivational, experiential, and support-related factors beyond sociodemographic characteristics alone.

XSL•FO
RenderX

Table . Regression analysis with digital health literacy as the outcome and internet use and technology affinity as predictor variables, controlling for sociodemographic characteristics and health-related factors (model 1[a], N=1824).

Variables	Coefficient (95% CI)	*P* value
Internet use and technology affinity		
Technology use		
Acceptance	0.023 (0.006 to 0.041)	.01
Competence	0.005 (−0.014 to 0.024)	.61
Support	0.027 (0.014 to 0.040)	<.001
Internet use		
Confidence	0.153 (0.125 to 0.182)	<.001
Frequency	0.057 (0.046 to 0.069)	<.001
Health-related search	0.080 (0.067 to 0.094)	<.001
Health and self-efficacy		
Self-efficacy	0.016 (0.008 to 0.024)	<.001
Overall health	0.001 (−0.001 to 0.002)	.48
Eyesight (reference: no impairment), χ^2 (*df*)=6.69 (3), *P*=.08		
Mild impairment	−0.022 (−0.062 to 0.019)	.29
Significant impairment	−0.090 (−0.204 to 0.024)	.12
Most serious impairment	−0.278 (−0.561 to 0.006)	.10
Hearing (reference: no impairment), χ^2 (*df*)=1.97 (3), *P*=.58		
Mild hearing loss	−0.019 (−0.050 to 0.012)	.22
Significant hearing loss	0.009 (−0.075 to 0.093)	.83
Extreme hearing loss/deafness	0.110 (−0.287 to 0.507)	.59
Sociodemographic characteristics		
Age	0.001 (−0.001 to 0.003)	.44
Sex (reference: male)	0.029 (0.001 to 0.057)	.045
Marital status (reference: married), χ^2 (*df*)=6.54 (4), *P*=.16		
Married, but separated	−0.018 (−0.125 to 0.089)	.74
Single	−0.055 (−0.123 to 0.013)	.11
Divorced	−0.009 (−0.051 to 0.034)	.70
Widowed	−0.047 (−0.090 to −0.005)	.03
Education[b] (reference: low), χ^2 (*df*)=3.70 (2), *P*=.16		
Mediate	0.040 (−0.023 to 0.102)	.22
High	0.016 (−0.046 to 0.078)	.61
Income	−0.008 (−0.017 to 0.002)	.11

[a]R^2=0.610.

[b]Comparative Analysis of Social Mobility in Industrial Nations (CASMIN).

Regression Analysis With Regard to Gender

Gender-stratified analyses revealed both shared and gender-specific patterns in the factors associated with DHL. The results can be found in Table 3, which shows that DHL was significantly associated with technology acceptance (coefficient=0.036, *P*=.003) among men (model 3) but not among women (model 2), suggesting that motivational readiness may play a more prominent role in shaping DHL for older men. The other aspects related to technology and internet use, such as frequency of use and received support, appear to be relevant to both genders, as indicated by *P* levels <.001. In other words, experiential use and access to assistance represent common enabling factors across genders. Additionally, 61% of DHL could be explained by our set of predictors.

XSL•FO
RenderX

Table . Subgroup analysis: regression analysis with digital health literacy as the outcome and internet use and technology affinity as predictor variables, controlling for sociodemographic characteristics and health-related factors (model 2: female subsample; model 3: male subsample).

Variables	Model 2[a]: women (n=973)		Model 3[b]: men (n=851)	
	Coefficient (95% CI)	*P* value	Coefficient (95% CI)	*P* value
Internet use and technology affinity				
Technology use				
Acceptance	0.010 (−0.015 to 0.035)	.44	0.036 (0.012 to 0.060)	.003
Competence	0.004 (−0.022 to 0.029)	.78	0.004 (0.025 to 0.033)	.79
Support	0.028 (0.009 to 0.048)	.005	0.024 (0.008 to 0.040)	.003
Internet use				
Confidence	0.198 (0.155 to 0.242)	<.001	0.112 (0.075 to 0.150)	<.001
Frequency	0.048 (0.032 to 0.063)	<.001	0.069 (0.052 to 0.085)	<.001
Health-related search	0.071 (0.051 to 0.091)	<.001	0.086 (0.069 to 0.104)	<.001
Health and self-efficacy				
Self-efficacy	0.012 (0.001 to 0.024)	.03	0.018 (0.006 to 0.029)	.002
Overall health	4.2×10^{5} (−0.001 to 0.001)	.95	0.001 (−0.001 to 0.002)	.39
Eyesight (reference: no impairment), model 2: χ^2 (*df*)=5.61 (3), *P*=.13; model 3: χ^2 (*df*)=30.30 (3), *P*<.001				
Mild impairment	−0.016 (−0.077 to 0.046)	.62	−0.027 (−0.079 to 0.025)	.31
Significant impairment	−0.145 (−0.305 to 0.014)	.07	0.001 (−0.150 to 0.152)	.99
Most serious impairment	−0.061 (−0.129 to 0.006)	.08	−0.498 (−0.676 to −0.320)	<.001
Hearing (reference: no impairment), model 2: χ^2 (*df*)=2.74 (3), *P*=.43; model 3: χ^2 (*df*)=47.26 (3), *P*<.001				
Mild hearing loss	−0.016 (−0.065 to 0.033)	.53	−0.025 (−0.064 to 0.014)	.21
Significant hearing loss	0.041 (−0.105 to 0.188)	.58	−0.023 (−0.115 to 0.070)	.63
Extreme hearing loss/deafness	0.272 (−0.109 to 0.653)	.16	−0.449 (−0.580 to −0.319)	<.001
Sociodemographic characteristics				
Age	−0.001 (−0.005 to 0.003)	.65	0.003 (−0.001 to 0.006)	.12
Marital status (reference: married), model 2: χ^2 (*df*)=1.15 (4), *P*=.89; model 3: χ^2 (*df*)=7.23 (4), *P*=.12				
Married, but separated	−0.004 (−0.175 to 0.168)	.97	−0.015 (−0.149 to 0.118)	.82
Single	−0.018 (−0.136 to 0.100)	.77	−0.075 (−0.154 to 0.005)	.07
Divorced	−0.003 (−0.058 to 0.063)	.93	0.007 (−0.065 to 0.052)	.82
Widowed	−0.026 (−0.081 to 0.280)	.34	−0.074 (−0.144 to −0.004)	.04
Education[c] (reference: low), model 2: χ^2 (*df*)=2.78 (2), *P*=.25; model 3: χ^2 (*df*)=1.76 (2), *P*=.41				
Mediate	0.062 (−0.028 to 0.152)	.18	0.030 (−0.055 to 0.115)	.48
High	0.040 (−0.052 to 0.132)	.40	0.004 (−0.076 to 0.084)	.92
Income	−0.007 (−0.022 to 0.007)	.31	−0.006 (−0.018 to 0.007)	.37

[a]R^2=0.609.

[b]R^2=0.614.

[c]Comparative Analysis of Social Mobility in Industrial Nations (CASMIN).

XSL•FO
RenderX

Discussion

Main Findings

The aim of this study was to shed light on technology commitment and support among individuals aged 65 years and older. Moreover, we investigated the relationship between DHL and several aspects related to technology and internet use, as well as how this association may differ between men and women.

According to our results, DHL was slightly lower than other studies, but similar to studies focusing on individuals aged 65 years and older, indicating medium levels of DHL and generally lower DHL scores compared to younger individuals [35-37]. Moreover, there were no significant differences between male and female participants, aligning with another study that did not reveal any significant gender differences in eHealth literacy levels [38,39]. In contrast, it has also been argued by a recent meta-analysis that women may exhibit lower DHL than men [18]. This inconsistency emphasizes the complexity of social determinants affecting DHL and suggests that factors such as socioeconomic status or geographical location might interact with gender, influencing access to digital health resources [19]. Therefore, future studies may want to investigate whether gender differences in terms of DHL may be mediated by loneliness and living situation (in terms of social network) or whether they depend on anxiety, social pressure, and perceived risks associated with digital health services or platforms, thereby explaining if gender differences may sometimes be shaded by the impact of other factors.

In terms of technological commitment, technology acceptance scores were slightly different compared to other studies with individuals aged 60 years and older, whereas technology competence was similar to previous findings [25,26,40]. However, this study replicates previous findings that men exhibit higher technological acceptance and competence [26]. Interestingly, our findings also show that women may receive more support with regard to technology use. Our finding that significant differences were found regarding technology acceptance could be meaningfully explained by the TAM. According to TAM, technology use is not primarily determined by competencies and skills but may rather depend on perceived usefulness as well as perceived ease of use [13,41]. In other words, similar DHL levels between men and women suggest similar competencies with regard to accessing, understanding, and using health information, whereas gender differences in terms of technology acceptance reflect gender-related differences in social dynamics (an important mediator in TAM). In general, older adults often rely on social support from family members or relatives when engaging with digital technologies [27], and studies suggest that such support mechanisms may differ between men and women. Previous research on social networks and relationships has shown that women are increasingly embedded socially [27,42] and may therefore be more likely to interact with friends or community members than men when it comes to technology issues or seeking advice and assistance, thereby receiving more support. In addition, older men are often expected to be more “technically competent” due to lifelong gendered stereotypes around technology, which may lead relatives to assume they *don’t need help* [43,44].

Based on the regression analysis, we further aimed to investigate the link between DHL and technology commitment and support, controlling for sociodemographic and health-related variables. Among the overall sample, both higher technology acceptance and support were positively linked to greater DHL, confirming our prior hypotheses. Moreover, feeling more confident and frequency of internet use (general and health-related) were significantly linked to greater DHL. However, self-perceived technology competence was not significantly linked to DHL. It may be that technology acceptance and support may be stronger predictors of DHL because it promotes active engagement with digital tools and therefore motivates to learn and develop new skills, increasing DHL [45]. In contrast, self-perceived technology competence only captures a subset of abilities to develop better DHL. In other words, being able to use technical devices is a general technological competence and may not be needed for all facets of DHL (ie, being able to understand or interpret digital health information does not need technical skills). Moreover, attitudes toward technology and actual engagement—elements captured by technology acceptance and frequency of use—are supposed to facilitate DHL by promoting active interaction with digital health resources, consistent with models of technology use that highlight the relevance of behavioral engagement and perceived usefulness [13,41,46].

Interestingly, the subgroup regression analysis regarding gender revealed some differences between men and women. In other words, being interested in using new technological developments (ie, technology acceptance) was associated with greater DHL among men. These significant associations were not found among women. Men may be more likely to approach new technology through independent exploration, meaning that their curiosity or interest translates into actual learning and skill-building, hence greater DHL. Women, by contrast, may use collaborative or supported strategies to overcome barriers, which dilute the direct effect of curiosity on literacy. A study by Kim et al [16] found a connection between gender and digital literacy, revealing that older women often encounter more significant challenges and demonstrate lower self-confidence in using technology compared to their male counterparts. These disparities in digital literacy can hinder women’s ability to effectively utilize digital health resources. Therefore, gender-responsive interventions to enhance digital literacy among older females could positively impact health management outcomes [16]. However, the frequency of internet use, feeling competent and confident in using technology, and having people around who can support with technical problems were associated with greater DHL among both men and women.

According to previous and current findings, it may be a complex interplay of sociodemographic characteristics, internet use and technology commitment, and health literacy that affects older adults’ participation in digital health initiatives. Tailoring efforts to promote DHL for both genders, while considering the unique barriers faced by the older adults, could bridge the gap and lead to improved health outcomes in individuals aged 65 years and older. Importantly, the impact of improving DHL extends

XSL•FO
RenderX

beyond immediate technology use; it enhances overall life satisfaction and quality of life among older adults, as indicated by studies illustrating the mediating effect of social capital in this relationship [47,48]. Increasing digital literacy may empower older individuals to take more proactive roles in managing their health, fostering a greater sense of agency and improving overall health outcomes. Future research could further investigate whether DHL is related to utilization of care, as it has been shown that general health literacy has been found to be associated with satisfaction with health care and unmet needs regarding various aspects of care [49]. Although this study did not differentiate between specific digital health devices, previous studies indicate that older adults' engagement with digital technologies increasingly involves activities, such as searching for health- or disease-related information, accessing electronic health records, using appointment systems, and, to a lesser extent, mobile health applications for self-management or monitoring. Technology acceptance in this context may therefore be understood as being open toward engaging with different kinds of digital health content, which require not only possibilities to access but also the ability to understand and interpret electronic health information. Prospective studies could go further by investigating the acceptance of everyday digital health activities that are increasingly relevant for older adults. These may include the use of smartphones or tablets to search for health information, use patient portals, schedule appointments using web-based tools, communicate with health care providers, or use health-related applications for medication management and monitoring.

Limitations

The aim of this study was to investigate the relationship between DHL and internet use and technology affinity. The study has some methodological constraints. Due to the cross-sectional characteristics of the study, it is not possible to draw a conclusion with regard to the direction of this relationship. In other words, better technology competence and acceptance as well as greater frequency of internet use—especially with regard to health-related aspects—may help to improve DHL. On the other hand, greater DHL due to other factors, such as generally higher self-efficacy, may lead to individuals feeling more prepared to use the internet or feel more confident, therefore using the internet more often. This, in turn, may increase opportunities for learning how to find health-related information more effectively and efficiently, therefore increasing DHL. Furthermore, the study is based on self-reported information. Participants may overestimate or underestimate their own abilities regarding (health-related) internet use. In fact, women have been found to be harder on themselves with regard to self-perceived technology capabilities, even if this does not match with their actual performance [50]. In addition, there are sample-specific issues that should be taken into account when interpreting the results. Although a heterogeneous sample in terms of education and income was analyzed, participants from rural areas or participants not living in their own (ie, living in nursing homes) may be underrepresented in this study. In this context, these individuals may experience difficulties with regard to internet access, lack possibilities to receive support in terms of technology use, or are generally more vulnerable to digital exclusion. Therefore, future studies may want to expand their sample and take into account individual living situations to further investigate possible interactions. Unfortunately, no nonresponder analysis in terms of reasons for nonparticipation could be conducted due to privacy restrictions and anonymization procedures. However, compared to similar studies, the response rate was very high (74.2%). In addition, we conducted case-wise deletion; therefore, possible selection effects should be considered regarding the current findings.

Implication and Conclusion

The interplay between DHL and technology acceptance among older adults highlights the urgent need for focused educational interventions and supportive frameworks that facilitate the adoption of digital health technologies. Enhancing DHL may not only promote the acceptance of new technologies but also potentially transform the health management landscape for older individuals. The current results highlight the importance of individual interventions that take into account specific gender differences. Future studies may want to investigate what form of support older people desire, what needs they have (taking into account individual resources and health status), and how certain fears in using digital devices or digital sources can be addressed to improve DHL, prevent the digital divide, and reduce (age-related) health care disparities. The importance of targeted interventions is clear; effective training programs must address these gender differences to ensure that both men and women benefit equally from advancements in digital health technologies. Tailored programs may need to address specific fears and anxieties related to technology use in older women, as highlighted by Kim et al [16], which focus on technology-use anxiety. Moreover, qualitative insights from Leong et al [51] indicated that older adults' trust in technology varies by gender and can be influenced by their experiences and their family's input. Women, who may be more cautious toward adopting new technologies, often require validation from trusted sources before engaging [51]. In fact, prior research suggested that community or peer learning, as well as intergenerational learning, may have positive long-term effects on DHL, as well as overall health status and management among older adults [18,52,53].

Acknowledgments

The authors would like to thank the participants of the LIFE (Leipzig Research Centre for Civilization Disease) Adult Study for participation. Generative artificial intelligence was not used as part of the study procedure.

XSL•FO
RenderX

Funding

The study received funding from the Saxon State Ministry for Science, Culture and Tourism (SMWK) and was cofinanced with tax revenue based on the budget approved by the Saxon State Parliament. The SMWK was not involved in the study design, data collection, analysis, interpretation, or the writing of the manuscript. Publication was supported by the Open Access Publishing Fund of Leipzig University.

Data Availability

The underlying data are available from the corresponding author upon reasonable request due to ethical and privacy restrictions.

Authors' Contributions

Conceptualization: FUJ (lead), SGR-H (supporting), ML (supporting)
Data curation: FUJ, SGR-H (equal), ML (equal)
Project administration: MR, KW (equal), ME (equal), YD (equal)
Writing – original draft: FUJ
Writing – review and editing: MR, KW (equal), ME (equal), YD (equal), SGR-H (equal), ML (equal)
The final manuscript has been read and approved by all mentioned authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Instruments.
[DOCX File, 17 KB - jmir_v28i1e85846_app1.docx]

References

1. Gibbons MC, Fleisher L, Slamon RE, Bass S, Kandadai V, Beck JR. Exploring the potential of Web 2.0 to address health disparities. J Health Commun 2011;16 Suppl 1:77-89. [doi: 10.1080/10810730.2011.596916] [Medline: 21843097]
2. Sadah SA, Shahbazi M, Wiley MT, Hristidis V. A study of the demographics of web-based health-related social media users. J Med Internet Res 2015 Aug 6;17(8):e194. [doi: 10.2196/jmir.4308] [Medline: 26250986]
3. Quittschalle J, Stein J, Luppa M, et al. Internet use in old age: results of a German population-representative survey. J Med Internet Res 2020 Nov 23;22(11):e15543. [doi: 10.2196/15543] [Medline: 33226351]
4. Yoon H, Jang Y, Kim S, Speasmaker A, Nam I. Trends in internet use among older adults in the United States, 2011-2016. J Appl Gerontol 2021 May;40(5):466-470. [doi: 10.1177/0733464820908427] [Medline: 32131670]
5. Bujnowska-Fedak MM, Mastalerz-Migas A. Usage of medical internet and e-health services by the elderly. Adv Exp Med Biol 2015;834:75-80. [doi: 10.1007/5584_2014_74] [Medline: 25315621]
6. van der Vaart R, Drossaert C. Development of the digital health literacy instrument: measuring a broad spectrum of Health 1.0 and Health 2.0 skills. J Med Internet Res 2017 Jan 24;19(1):e27. [doi: 10.2196/jmir.6709] [Medline: 28119275]
7. Ban S, Kim Y, Seomun G. Digital health literacy: a concept analysis. Digit Health 2024;10:20552076241287894. [doi: 10.1177/20552076241287894] [Medline: 39381807]
8. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9. [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
9. Shi Z, Du X, Li J, Hou R, Sun J, Marohabutr T. Factors influencing digital health literacy among older adults: a scoping review. Front Public Health 2024;12:1447747. [doi: 10.3389/fpubh.2024.1447747] [Medline: 39555039]
10. Western MJ, Smit ES, Gültzow T, et al. Bridging the digital health divide: a narrative review of the causes, implications, and solutions for digital health inequalities. Health Psychol Behav Med 2025;13(1):2493139. [doi: 10.1080/21642850.2025.2493139] [Medline: 40276490]
11. Mason AM, Compton J, Bhati S. Disabilities and the digital divide: assessing web accessibility, readability, and mobility of popular health websites. J Health Commun 2021 Oct 3;26(10):667-674. [doi: 10.1080/10810730.2021.1987591] [Medline: 34657585]
12. Liu S, Zhao H, Fu J, et al. Current status and influencing factors of digital health literacy among community-dwelling older adults in Southwest China: a cross-sectional study. BMC Public Health 2022 May 17;22(1):996. [doi: 10.1186/s12889-022-13378-4] [Medline: 35581565]
13. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
14. Hafner J, Schönfeld S, Tokgöz P, Hrynyschyn R, Schlubach A, Dockweiler C. Akzeptanz von digitalen Gesundheitstechnologien bei leichter bis mittelgradiger Depression – Eine qualitative Analyse aus der Perspektive von

XSL•FO
RenderX

Betroffenen, Angehörigen und Leistungserbringern [Article in German]. Psychiatr Prax 2022 Sep;49(6):313-321. [doi: 10.1055/a-1472-5388] [Medline: 34015851]
15. Wei Y, Zhou Y. The mediating impact of exercise frequency and duration on the relationship between digital health literacy and health among older adults: a cross-sectional study. Medicine (Baltimore) 2024 Oct 18;103(42):e39877. [doi: 10.1097/MD.0000000000039877] [Medline: 39432656]
16. Kim J, Jeon SW, Byun H, Yi E. Exploring e-Health literacy and technology-use anxiety among older adults in Korea. Healthcare (Basel) 2023;11(11):1556. [doi: 10.3390/healthcare11111556] [Medline: 37297696]
17. Rasekaba TM, Pereira P, Rani GV, Johnson R, McKechnie R, Blackberry I. Exploring telehealth readiness in a resource limited setting: digital and health literacy among older people in rural India (DAHLIA). Geriatrics (Basel) 2022 Mar 1;7(2):28. [doi: 10.3390/geriatrics7020028] [Medline: 35314600]
18. Jiang X, Wang L, Leng Y, et al. The level of electronic health literacy among older adults: a systematic review and meta-analysis. Arch Public Health 2024 Nov 7;82(1):204. [doi: 10.1186/s13690-024-01428-9] [Medline: 39511667]
19. Shao Y, Yang X, Chen Q, et al. Determinants of digital health literacy among older adult patients with chronic diseases: a qualitative study. Front Public Health 2025;13:1568043. [doi: 10.3389/fpubh.2025.1568043] [Medline: 40231178]
20. Loeffler M, Engel C, Ahnert P, et al. The LIFE-Adult-Study: objectives and design of a population-based cohort study with 10,000 deeply phenotyped adults in Germany. BMC Public Health 2015 Jul 22;15(1):691. [doi: 10.1186/s12889-015-1983-z] [Medline: 26197779]
21. Engel C, Wirkner K, Zeynalova S, et al. Cohort profile: the LIFE-Adult-Study. Int J Epidemiol 2023 Feb 8;52(1):e66-e79. [doi: 10.1093/ije/dyac114] [Medline: 35640047]
22. Marsall M, Engelmann G, Skoda EM, Teufel M, Bäuerle A. Measuring electronic health literacy: development, validation, and test of measurement invariance of a Revised German Version of the eHealth Literacy Scale. J Med Internet Res 2022 Feb 2;24(2):e28252. [doi: 10.2196/28252] [Medline: 35107437]
23. Bäuerle A, Marsall M, Jahre LM, et al. Psychometric properties of the German Revised Version of the Ehealth Literacy Scale in individuals with cardiac diseases: validation and test of measurement invariance. Digit Health 2023;9:20552076231194915. [doi: 10.1177/20552076231194915] [Medline: 37588160]
24. Kurzskala technikbereitschaft (TB, technology commitment) [Web page in German]. ZIS - GESIS Leibniz Institute for the Social Sciences. 2016. URL: https://zis.gesis.org/skala/Neyer-Felber-Gebhardt-Kurzskala-Technikbereitschaft-%28TB,-technology-commitment%29?lang=en [accessed 2026-04-30]
25. Latendorf A, Runde LM, Salminen T, Steinert A. Digitization of neuropsychological diagnostics: a pilot study to compare three paper-based and digitized cognitive assessments. Aging Clin Exp Res 2021 Jun;33(6):1585-1597. [doi: 10.1007/s40520-020-01668-z] [Medline: 33001402]
26. Ferizaj D, Perotti L, Dahms R, Heimann-Steinert A. Technologienutzung im Alter: Zusammenhänge zwischen Akzeptanz, Kompetenz, Kontrolle, Interesse und sozialen Indikatoren bei Personen über 60 Jahre [Article in German]. Z Gerontol Geriat 2024 May;57(3):227-234. [doi: 10.1007/s00391-023-02225-9]
27. Kamin ST, Beyer A, Lang FR. Social support is associated with technology use in old age [Article in German]. Z Gerontol Geriatr 2020 May;53(3):256-262. [doi: 10.1007/s00391-019-01529-z] [Medline: 30877360]
28. Welzel FD, Löbner M, Quittschalle J, et al. Loss and bereavement in late life (60+): study protocol for a randomized controlled trial regarding an internet-based self-help intervention. Internet Interv 2021 Dec;26:100451. [doi: 10.1016/j.invent.2021.100451] [Medline: 34540595]
29. Lechert Y, Schroedter J, Lüttinger P. Die Umsetzung der Bildungsklassifikation CASMIN für die Volkszählung 1970, die Mikrozensus-Zusatzerhebung 1971 und die Mikrozensen 1976-2004 [Report in German]. : GESIS - Leibniz-Institut für Sozialwissenschaften; 2006 URL: https://www.ssoar.info/ssoar/handle/document/26235 [accessed 2026-04-30]
30. Brauns H, Steinmann S. Educational reform in France, West-Germany and the United Kingdom: updating the CASMIN educational classification. ZUMA Nachrichten 1999;23(44):7-44 [FREE Full text]
31. EQ-5D-5L user guide. : EuroQol Group; 2019 URL: https://euroqol-domain.ams3.digitaloceanspaces.com/wp-content/uploads/2025/01/12124516/EQ-5D-5L-Userguide-2025-04.pdf [accessed 2026-04-30]
32. Zülke A, Luck T, Pabst A, et al. AgeWell.de - study protocol of a pragmatic multi-center cluster-randomized controlled prevention trial against cognitive decline in older primary care patients. BMC Geriatr 2019 Aug 1;19(1):203. [doi: 10.1186/s12877-019-1212-1] [Medline: 31370792]
33. Allgemeine selbstwirksamkeit kurzskala (ASKU) [Website in German]. GESIS Leibniz Institute for the Social Sciences. 2014. URL: https://zis.gesis.org/skala/Beierlein-Kovaleva-Kemper-Rammstedt-Allgemeine-Selbstwirksamkeit-Kurzskala-%28ASKU%29?lang=de [accessed 2026-04-30]
34. Manning WG, Basu A, Mullahy J. Generalized modeling approaches to risk adjustment of skewed outcomes data. J Health Econ 2005 May;24(3):465-488. [doi: 10.1016/j.jhealeco.2004.09.011] [Medline: 15811539]
35. König L, Kuhlmey A, Suhr R. Digital health literacy of the population in Germany and its association with physical health, mental health, life satisfaction, and health behaviors: nationally representative survey study. JMIR Public Health Surveill 2024 Feb 21;10:e48685. [doi: 10.2196/48685] [Medline: 38381497]

XSL•FO
RenderX

36. Marsall M, Weigl M, Lüttel D, Müller H. Digital health literacy: a cross-sectional survey study among patients after hospitalization in Germany. Z Evid Fortbild Qual Gesundhwes 2025 Apr;193(18-25):18-25. [doi: 10.1016/j.zefq.2025.01.004] [Medline: 40023719]
37. Muellmann S, Wiersing R, Zeeb H, Brand T. Digital health literacy in adults with low reading and writing skills living in Germany: mixed methods study. JMIR Hum Factors 2025 May 22;12:e65345. [doi: 10.2196/65345] [Medline: 40402567]
38. Kyaw MY, Aung MN, Koyanagi Y, et al. Sociodigital determinants of ehealth literacy and related impact on health outcomes and ehealth use in Korean older adults: community-based cross-sectional survey. JMIR Aging 2024 Aug 13;7:e56061. [doi: 10.2196/56061] [Medline: 39140239]
39. Qiu CS, Lunova T, Greenfield G, et al. Determinants of Digital Health Literacy: International Cross-Sectional Study. J Med Internet Res ;27:e66631. [doi: 10.2196/66631]
40. Tahden MAS, Gieseler A, Meis M, Wagener KC, Colonius H. What keeps older adults with hearing impairment from adopting hearing aids? Trends Hear 2018;22:2331216518809737. [doi: 10.1177/2331216518809737] [Medline: 30451099]
41. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. Manage Sci 1989 Aug;35(8):982-1003. [doi: 10.1287/mnsc.35.8.982]
42. Schwartz E, Litwin H. Social network changes among older Europeans: the role of gender. Eur J Ageing 2018 Dec;15(4):359-367. [doi: 10.1007/s10433-017-0454-z] [Medline: 30532673]
43. Hargittai E, Shafer S. Differences in actual and perceived online skills: the role of gender. Soc Sci Q 2006 Jun;87(2):432-448. [doi: 10.1111/j.1540-6237.2006.00389.x]
44. Morahan-Martin J, Schumacher P. Attitudinal and experiential predictors of technological expertise. Comput Human Behav 2007 Sep;23(5):2230-2239. [doi: 10.1016/j.chb.2006.03.003]
45. Seifert A, Schelling HR. Mobile use of the internet using smartphones or tablets by Swiss people over 65 years. Gerontechnology 2015;14(1):57-62. [doi: 10.4017/gt.2015.14.1.006.00]
46. Adnan A, Irvine RE, Williams A, Harris M, Antonacci G. Improving acceptability of mhealth apps-the use of the technology acceptance model to assess the acceptability of mHealth apps: systematic review. J Med Internet Res 2025 May 7;27:e66432. [doi: 10.2196/66432] [Medline: 40334265]
47. Lee SY, Kim Y, Kim B, Lee SG, Jang SY, Kim TH. Digital literacy and its association with subjective health status and healthy lifestyle behaviors among Korean older adults: cross-sectional study. JMIR Aging 2025 Jun 5;8:e64974. [doi: 10.2196/64974] [Medline: 40472263]
48. Jung H, Lee H, Nam EW. Mediating effect of social capital on the association between digital literacy and life satisfaction among older adults in South Korea: cross-sectional study. JMIR Form Res 2025 Feb 6;9:e68163. [doi: 10.2196/68163] [Medline: 39916363]
49. Jung F, Zülke A, Wirkner K, et al. Gesundheitskompetenz, Anlaufstellen, ungedeckte subjektive Bedarfe und Behandlungszufriedenheit von Long-COVID-Betroffenen mit langanhaltender neuropsychiatrischer Symptomatik [Article in German]. Psychiatr Prax 2024 Nov;51(8):418-425. [doi: 10.1055/a-2334-1822] [Medline: 39079687]
50. Sobieraj S, Krämer NC. Similarities and differences between genders in the usage of computer with different levels of technological complexity. Comput Human Behav 2020 Mar;104:106145. [doi: 10.1016/j.chb.2019.09.021]
51. Leong QY, Lee VV, Ng WY, et al. Older adults' perspectives and experiences with digital health in singapore: qualitative study. JMIR Hum Factors 2024 Nov 11;11:e58641. [doi: 10.2196/58641] [Medline: 39527787]
52. Dong Q, Liu T, Liu R, Yang H, Liu C. Effectiveness of digital health literacy interventions in older adults: single-arm meta-analysis. J Med Internet Res 2023 Jun 28;25:e48166. [doi: 10.2196/48166] [Medline: 37379077]
53. Wale A, Everitt J, Ayres T, et al. Effectiveness of interventions for addressing digital exclusion in older adults in the social care domain: rapid review. JMIR Aging 2025 Dec 30;8:e70377. [doi: 10.2196/70377] [Medline: 41468583]

Abbreviations

CASMIN: Comparative Analysis of Social Mobility in Industrial Nations
DHL: digital health literacy
GR-eHEALS: German eHealth Literacy Scale
LIFE: Leipzig Research Centre for Civilization Diseases
TAM: technology acceptance model

XSL•FO
RenderX

Edited by A Mavragani; submitted 14.Oct.2025; peer-reviewed by CK Tan, K Mouloudj; revised version received 26.Jan.2026; accepted 30.Mar.2026; published 14.May.2026.

Please cite as:
Jung FU, Luppa M, Reusche M, Wirkner K, Eberl M, Dietz Y, Engel C, Riedel-Heller SG
Digital Health Literacy, Technology Acceptance, and Competence Among Older Adults Aged ≥65 Years: Cross-Sectional Study Investigating Differences Between Women and Men
J Med Internet Res 2026;28:e85846
URL: https://www.jmir.org/2026/1/e85846
doi:10.2196/85846

© Franziska Ulrike Jung, Melanie Luppa, Matthias Reusche, Kerstin Wirkner, Melanie Eberl, Yvonne Dietz, Christoph Engel, Steffi G Riedel-Heller. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Structural Inequalities in Online Health Information Seeking: Cross-National Multilevel Study

Petra Raudenská[1], PhD; Elena Link[2], Prof Dr

[1]Institute of Sociology, Czech Academy of Sciences, Jilská 1, Prague, Czech Republic

[2]Department of Communication, Johannes Gutenberg-University Mainz, Jakob-Welder-Weg 12, Mainz, Germany

Corresponding Author:
Petra Raudenská, PhD
Institute of Sociology, Czech Academy of Sciences, Jilská 1, Prague, Czech Republic

Abstract

Background: Online health information–seeking behavior (OHISB) has become an increasingly common component of contemporary health self-management. Individuals use a wide range of digital sources, including websites, social media platforms, and mobile apps, to obtain health-related information. However, substantial disparities persist in who seeks health information online, and which populations benefit from digital health resources. While previous research has largely focused on individual-level determinants, cross-national evidence on structural influences remains limited.

Objective: This study aims to assess between-country variation in OHISB, examine associations between individual-level characteristics and OHISB, and investigate how country-level structural conditions are associated with cross-national differences in OHISB, net of individual-level characteristics.

Methods: Data were drawn from the Health and Health Care II module of the International Social Survey Programme (ISSP 2021 - 2024; n=35,592; 32 countries). OHISB was measured as any use of the Internet to search for health-related information during the past 12 months. Multilevel logistic regression models were estimated. Country-level indicators were reduced using principal component analysis into 4 composite indices. Robustness checks included analyses excluding respondents without internet access and models incorporating survey weights.

Results: OHISB varied substantially across countries (intraclass correlation coefficient=0.177). At the individual level, younger age, higher education, female respondents, recent health problems, doctor visits, unmet medical needs, and perceived usefulness of the internet were associated with higher odds of OHISB. At the macro level, the socioeconomic and health development showed the strongest association (odds ratio=1.52 per SD; *P*=.003) and explained a substantial share of between-country variation. Cultural hierarchy–individualism was associated with OHISB in separate models but attenuated when adjusted for development. Cross-level interactions indicated that the gender gap and the role of perceived usefulness were more pronounced in higher-development contexts, although these findings were exploratory.

Conclusions: OHISB is associated with both individual characteristics and broader structural conditions. Socioeconomic and health development appear to play a key contextual role in shaping cross-national differences in digital health engagement, highlighting the importance of addressing both individual and structural dimensions of digital health inequalities.

(J Med Internet Res 2026;28:e88110) doi:10.2196/88110

KEYWORDS

online health information seeking; digital health; digital divide; multilevel analysis; cross-national comparison; socioeconomic development; structural determinants; health inequalities; ISSP; International Social Survey Programme

Introduction

Background

The rapid expansion of digital health information has transformed how individuals access and manage health-related knowledge. Today, individuals obtain health information from a wide range of online sources, including search engines, institutional and news websites, mobile apps, online forums, social media platforms, or video-based content, and increasingly also artificial intelligence (AI)-assisted search tools and conversational interfaces [1,2]. These resources provide access to diverse types of health content, ranging from preventive advice and lifestyle guidance to information on symptoms, medical treatments, and health care services, which individuals use to understand symptoms, evaluate treatment options, manage chronic conditions, regulate emotions, and support health-related decision-making [3-6]. While digital technologies may enhance access to information and support more active participation in health-related decision-making [7,8], they also introduce new challenges, including information overload, misinformation,

XSL•FO
RenderX

and inequalities in individuals' ability to access and benefit from online health information [9,10].

Social Stratification and Online Health Information Seeking

These disparities reflect broader patterns of social and digital inequality. Stratification theory suggests that existing social hierarchies tend to be reproduced in digital environments rather than eliminated by technological access alone [11-13]. Accordingly, the benefits of digital technologies are unevenly distributed across population groups. Engagement in online health information seeking varies with socioeconomic resources, education, age, digital skills, health literacy, and trust in online information sources, reflecting both structural and motivational barriers to online engagement. Non-engagement may therefore result either from lack of internet access or from active disengagement driven by low perceived utility, distrust, or preferences for alternative sources [14]. Moreover, the types of digital platforms and information sources used for health information seeking also differ across demographic and cultural groups, reflecting variations in media preferences, language constraints, and health information practices [15,16].

From this perspective, online health information–seeking behavior (OHISB) can be understood as a behavioral manifestation of digital stratification. Empirical research shows that individuals with greater socioeconomic resources, education, and digital skills are more likely to access and benefit from online health information, whereas disadvantaged groups face persistent barriers to engagement. Women, younger individuals, and those with higher levels of education or income tend to seek health information online more frequently, whereas older adults, individuals with lower socioeconomic status, rural residents, and racial minorities are less likely to engage in such behaviors [4,17-19].

In addition to structural resources, health-related circumstances and psycho-motivational factors also shape online health information seeking. Individuals experiencing health problems, perceived health risks, or higher levels of health anxiety are more likely to search for health information online [20,21]. Information-seeking models further highlight the importance of perceived usefulness and credibility of information sources, knowledge gaps, and self-efficacy in shaping engagement with online health information [22-24].

From Individual Determinants to Cross-National Contexts

Inequalities in OHISB extend beyond individuals and are also reflected in substantial differences between countries. In high-income countries, between 70% and 80% of adults report searching for health information online, compared to substantially lower rates in many lower-income settings [4,25,26]. Even within Europe, prevalence ranges from approximately 43% in Bulgaria to more than 80% in Finland [27].

These patterns suggest that engagement with online health information is associated not only with individual characteristics but also with broader structural conditions [28-30]. National differences in health care systems, digital infrastructure, public health conditions, and socioeconomic development can affect both the availability of online health resources and individuals' capacity to benefit from them [31-34]. For example, constraints in health care access, such as limited availability of medical services, high costs, or long waiting times, may encourage individuals to seek complementary information online, whereas well-functioning health care systems with high levels of trust and satisfaction may reduce this need [35].

Cultural contexts may further shape norms of health-related help-seeking, trust in institutions, and attitudes toward autonomy in decision-making. Societies characterized by higher individualism and lower power distance may encourage more autonomous health information seeking and greater willingness to verify or supplement medical advice through independent sources, whereas more hierarchical or collectivist contexts may rely more strongly on interpersonal networks or professional guidance [36-39].

Digital infrastructure defines the technological opportunity structure for online engagement. Countries with higher levels of internet access, connectivity, and digital adoption provide individuals with greater opportunities to access and use online health information, while differences in digital skills and information literacy influence individuals' ability to navigate online information environments and evaluate the credibility of health content [40-42].

Finally, broader socioeconomic development forms the structural foundation for digital engagement and health information access. Higher levels of development are associated with better health care infrastructure, higher educational attainment, stronger digital skills, and greater digital connectivity, all of which facilitate engagement with online health information [42-44]. Conversely, lower levels of development are often linked to weaker infrastructure, lower health literacy, and limited trust in digital sources, potentially constraining engagement with online health information. Regulatory and institutional contexts may further shape the online health information environment by influencing the credibility, visibility, and governance of digital health content across platforms [45-47].

Despite growing recognition of these disparities, research on OHISB remains fragmented. Existing studies have primarily focused on individual-level determinants, while the role of macro-level contexts has received comparatively less systematic attention in large-scale cross-national analyses [9,48]. Moreover, many studies rely on descriptive comparisons or analyses based on a limited number of countries, often without applying multilevel analytical frameworks, making it difficult to disentangle the relative contributions of individual and contextual factors. As a result, the relative contribution of individual and contextual factors to cross-national variation in OHISB remains insufficiently understood.

Study Objectives

This study addresses this gap by examining cross-national variation in OHISB using a comparative multilevel framework. The objectives of this study are threefold: (1) to assess between-country variation in OHISB, (2) to examine associations between individual-level characteristics and

XSL•FO
RenderX

OHISB, and (3) to investigate how country-level structural conditions are associated with cross-national differences in OHISB. The analysis draws on data from the Health and Health Care II module of the International Social Survey Programme (ISSP 2021-2024; 32 countries; n=35,592), a long-running cross-national survey program that collects harmonized social science data using probabilistic national samples [49]. To address these objectives, we use multilevel logistic regression models that account for individuals nested within countries, using data from 32 nationally representative samples.

Methods

Data and Sample

The study combines individual-level and country-level data. Individual-level data were drawn from the Health and Health Care II module of the ISSP, collected between 2021 and 2024 across 32 countries [49-51]. The ISSP is a long-running cross-national collaboration that provides harmonized survey data based on probabilistic national samples of individuals aged 16 years and older. The integrated dataset used in this study includes countries from Europe, the Americas, Asia, Africa, and Oceania, allowing for comparative analysis across a diverse range of institutional and socioeconomic contexts. National sample sizes ranged from 1001 in Mexico to 3349 in Switzerland, yielding an initial sample of 48,302 respondents. Of these, 52.2% (25,213/48,302) were female, 24.9% (12,027/48,302) had tertiary education, and the mean age was 47.9 (SD 17.6) years. Further details on country coverage and fieldwork timing are provided in Table S1 of Multimedia Appendix 1.

Missing data were present across several variables, resulting in the exclusion of approximately 26.3% (12,703/48,302) of the initial sample in the final analytical dataset (n=35,592). The highest rates of missingness were observed for attitudinal measures of perceived usefulness of the internet (3767/48,302, 7.8%) and reliability of online health information (4781/48,302, 9.9%). These missing values were largely driven by the ISSP survey design, in which respondents without internet access or those who selected “can’t choose” were coded as missing.

Because imputing values for respondents without internet access would be conceptually inappropriate, these cases were not imputed. For the remaining variables, item nonresponse was generally low. However, comparison of respondents included in the final analytical sample with those excluded due to missing data revealed systematic differences (Table S2 in Multimedia Appendix 1). Excluded respondents were more likely to be older, female, less educated, economically inactive, and in poorer health, consistent with established patterns of digital exclusion. This indicates that missingness was unlikely to be fully random.

Multiple imputation was considered but not implemented due to the structurally induced missingness in internet-related variables and the risk of introducing cross-national artifacts. Country-specific imputation was also not feasible due to small numbers of missing observations within countries, which led to model instability and convergence issues. Given the relatively low level of item nonresponse and the stability of results across sensitivity analyses (including models excluding respondents without internet access), listwise deletion was retained as a parsimonious analytical strategy.

Measurements

OHISB

OHISB was measured using four items from the ISSP Health and Health Care II module. Respondents were asked how often during the past 12 months they had used the internet on any device (eg, computer, tablet, or smartphone) to search for (1) health or medical information for themselves or others; (2) information about a healthy lifestyle; (3) information related to anxiety, stress, or similar issues; and (4) vaccination information. Response categories ranged from “never or almost never” to “several times a week.”

Because the ISSP question captures general internet searching behavior rather than the use of specific platforms, it reflects a broad range of digital environments through which individuals may obtain health information, including diverse digital channels such as search engines, apps, social media, and online forums.

For the main analyses, a binary indicator was constructed to distinguish between respondents who reported any engagement in online health information seeking (“health onliners”) and those who did not (“health offliners”) [33]. Respondents who indicated “never or almost never, I do not have access to the internet” across all items were classified as health offliners, while those reporting any higher frequency on at least one item were classified as health onliners. This operationalization reflects the conceptualization of OHISB as a behavioral outcome, where the absence of online health information seeking indicates non-engagement regardless of the underlying reason.

At the same time, this definition combines individuals who are unable to access the internet (ie, constrained nonuse) with those who have access but do not seek health information online (ie, voluntary nonuse), introducing heterogeneity in the outcome. Sensitivity analyses excluding respondents without internet access yielded substantively similar results, indicating that the main findings are not driven solely by access constraints. Descriptive comparisons further showed that OHISB prevalence is consistently higher when restricted to individuals with internet access (Table S3 in Multimedia Appendix 1).

Additionally, the use of a binary indicator reflects considerations of cross-national measurement comparability. Prior research suggests that single-item measures of online health information seeking may lack measurement invariance across countries, whereas composite indicators provide more robust comparability in cross-national settings [52]. The combined binary measure used in this study therefore captures engagement across multiple domains while reducing potential bias arising from cross-national differences in item functioning.

Individual-Level Variables

Individual-level variables were selected to capture sociodemographic characteristics, health status, health care experiences, and psycho-motivational factors associated with OHISB.

XSL•FO
RenderX

Sociodemographic variables included gender (female=1), age groups (eight ten-year categories), education (eight-level degree classification based on International Standard Classification of Education [ISCED] categories), employment status, and partnership status. These variables reflect established dimensions of social stratification associated with digital inequality. Income and minority status were excluded due to cross-national data limitations. Ethnicity was inconsistently coded across countries and entirely missing for Switzerland, while income data were available only for the economically active population, which would have substantially reduced the analytical sample.

Health-related variables comprised indicators of health status and health behaviors, including the presence of health problems, smoking, alcohol consumption, physical activity, disability, subjective health, and well-being. These variables capture both objective and subjective aspects of health that may be associated with engagement in health information seeking.

Health care-related variables included frequency of doctor visits, satisfaction with health care use, unmet health care needs, and trust in health care providers and systems (eg, doctor trust, confidence in the health care system, satisfaction with health care, perceived quality, and access). These indicators reflect individuals' experiences with and perceptions of health care systems, which may also be associated with their engagement in online health information seeking.

Finally, psycho-motivational factors included perceived usefulness of the internet for health-related purposes and perceived reliability of online health information. These variables capture individuals' familiarity with digital environments and their evaluation of online information sources.

To account for temporal differences in data collection, fieldwork year (2021 - 2024) was included as an individual-level control variable. Because countries participated in different waves of data collection, this variable also captures potential temporal variation associated with post-pandemic changes in digital health engagement (Table S1 in Multimedia Appendix 1 for country-level fieldwork timing). It was modeled as a grand-mean–centered continuous variable (year minus 2021) to capture linear temporal variation. Sensitivity analyses using alternative specifications (dummy-coded years and quadratic term) produced substantively similar results and did not improve model fit.

Full definitions, coding details, and descriptive statistics are provided in Tables S4a–S4c of Multimedia Appendix 1.

Country-Level Variables

Country-level variables were selected to capture structural characteristics of national contexts that may be associated with cross-national variation in OHISB. Building on prior research and the theoretical framework of this study, four domains were considered: health care system characteristics, digital infrastructure, socioeconomic development, and cultural context.

Indicators of health care system characteristics included measures of health care resources and system performance, such as the number of physicians [53-55] and hospitals per capita [56-63], travel time to health facility [64], life expectancy, healthy life expectancy [65,66], and selected health-related behaviors at the population level (eg, smoking [67-69] and alcohol consumption [70,71]). Evaluative indicators were derived from aggregated survey measures capturing trust in doctors and perceptions of the health care system, including trust, satisfaction, perceived quality, and accessibility.

Digital infrastructure was captured through indicators of internet access and internet use [72-74]. Socioeconomic development was measured using macroeconomic and human development indicators, including gross domestic product [75,76], the human development index [77,78], and health expenditure [79-81]. Cultural context was represented using Hofstede's cultural dimensions (eg, individualism and power distance) [82].

Given the large number of correlated macro-level indicators and the relatively small number of countries, these variables were aggregated into composite indices using principal component analysis (PCA). This approach reduces dimensionality, limits multicollinearity, and captures broader structural dimensions rather than relying on highly collinear single indicators. The analysis yielded four components representing (1) health system evaluation and trust, (2) socioeconomic and health development, (3) cultural hierarchy–individualism, and (4) unhealthy lifestyle patterns.

The resulting components were standardized and included in the multilevel models as country-level variables. As a robustness check, we also estimated alternative models using selected individual country-level indicators in place of the composite indices. These specifications produced substantively similar results, supporting the robustness of the PCA-derived indices used in the main analysis. Correlations between country-level variables, factor loadings, and component diagnostics are reported in Tables S5–S6 in Multimedia Appendix 1. Detailed country-level data are provided in the Open Science Framework repository [83].

Analytical Strategy

To examine associations between individual- and country-level factors and OHISB, we estimated multilevel logistic regression models with individuals nested within countries. This approach accounts for the hierarchical structure of the data and allows for the separation of within- and between-country variation.

We first estimated a null model (Model 0) with a random intercept for countries to assess between-country variation [84]. The intraclass correlation coefficient (ICC) was calculated using the latent variable approach, assuming a level-1 variance of $\pi^2/3$. This provided an estimate of the proportion of total variance attributable to differences between countries.

Individual-level variables were then introduced stepwise: sociodemographic characteristics (Model 1), health-related factors (Model 2), health care-related variables (Model 3), and psycho-motivational factors (Model 4). Country-level variables were subsequently added to the full individual-level model to assess contextual effects. PCA-derived indices were first introduced in separate models and then combined in a final model if they improved model fit.

XSL•FO
RenderX

As a robustness check, country-level indicators were also examined individually in models including all individual-level covariates to assess consistency of effects and potential multicollinearity. These alternative specifications yielded substantively similar results (Table S9a–S9b in Multimedia Appendix 1).

Cross-level interactions between selected country-level indices and theoretically relevant individual-level variables were explored to assess whether individual-level associations varied across country contexts. Only statistically significant interactions were retained.

Model fit was evaluated using the Bayesian information criterion and likelihood-ratio tests [85]. The proportional change in variance (PCV) was used to assess the contribution of successive models.

Different centering strategies were applied to distinguish within- and between-country effects and to reduce cross-level confounding. For macro-level variables derived from aggregated individual data, individual-level variables were group-mean centered and country-level aggregates grand-mean centered. For externally sourced country-level variables, both levels were grand-mean centered [86].

For descriptive analyses, the original ISSP design weights were applied to ensure national representativeness. Multilevel models were estimated without rescaled weights; robustness checks using rescaled weights yielded substantively similar results, indicating that findings were not sensitive to weighting strategy [87,88].

Multicollinearity was assessed using variance inflation factors, with all values below conventional thresholds (maximum variance inflation factor=2.07; Table S7 in Multimedia Appendix 1). Model assumptions, including linearity of continuous variables in the logit, were assessed, with no substantial violations detected. All annotated code and replication materials are available in the Open Science Framework repository [83].

Ethical Considerations

This study is based on a secondary analysis of data from the ISSP (ISSP 2021 - 2024). The dataset is fully anonymized and contains no personally identifiable information. No new data collection involving human participants was conducted by the authors.

The original ISSP data collection was conducted in accordance with national and institutional ethical standards in each participating country. Participation was voluntary and informed consent was obtained from all respondents. In addition, ISSP survey instruments are reviewed and approved by the ISSP General Assembly with regard to scientific quality and ethical appropriateness, and all national datasets are anonymized prior to inclusion in the ISSP archive to prevent identification of individual participants [89].

The authors had permission to use the data for research purposes in accordance with ISSP data access policies.

As this study involved only secondary analysis of fully anonymized data and did not involve interaction with human participants or access to identifiable personal data, it did not constitute human participants research as defined by applicable research ethics standards. Therefore, formal ethics approval was not required. No compensation was provided to participants in the context of this secondary analysis.

All analyses were conducted in accordance with established principles of research integrity and ethical data handling, including the Code of Ethics for Researchers of the Czech Academy of Sciences [90] and internationally recognized guidelines for secondary data use. Data privacy and confidentiality were maintained throughout the study.

Results

Descriptive Findings

The prevalence of OHISB varied substantially across countries, ranging from 60.3% (1085/1787) in the Philippines to 98.1% (1164/1186) in Israel (Figure 1). Countries with lower levels of internet penetration generally exhibited lower prevalence, whereas countries with near-universal access showed consistently high levels of engagement.

XSL•FO
RenderX

Figure 1. Prevalence of online health information–seeking behavior across countries (ranked), based on weighted data.

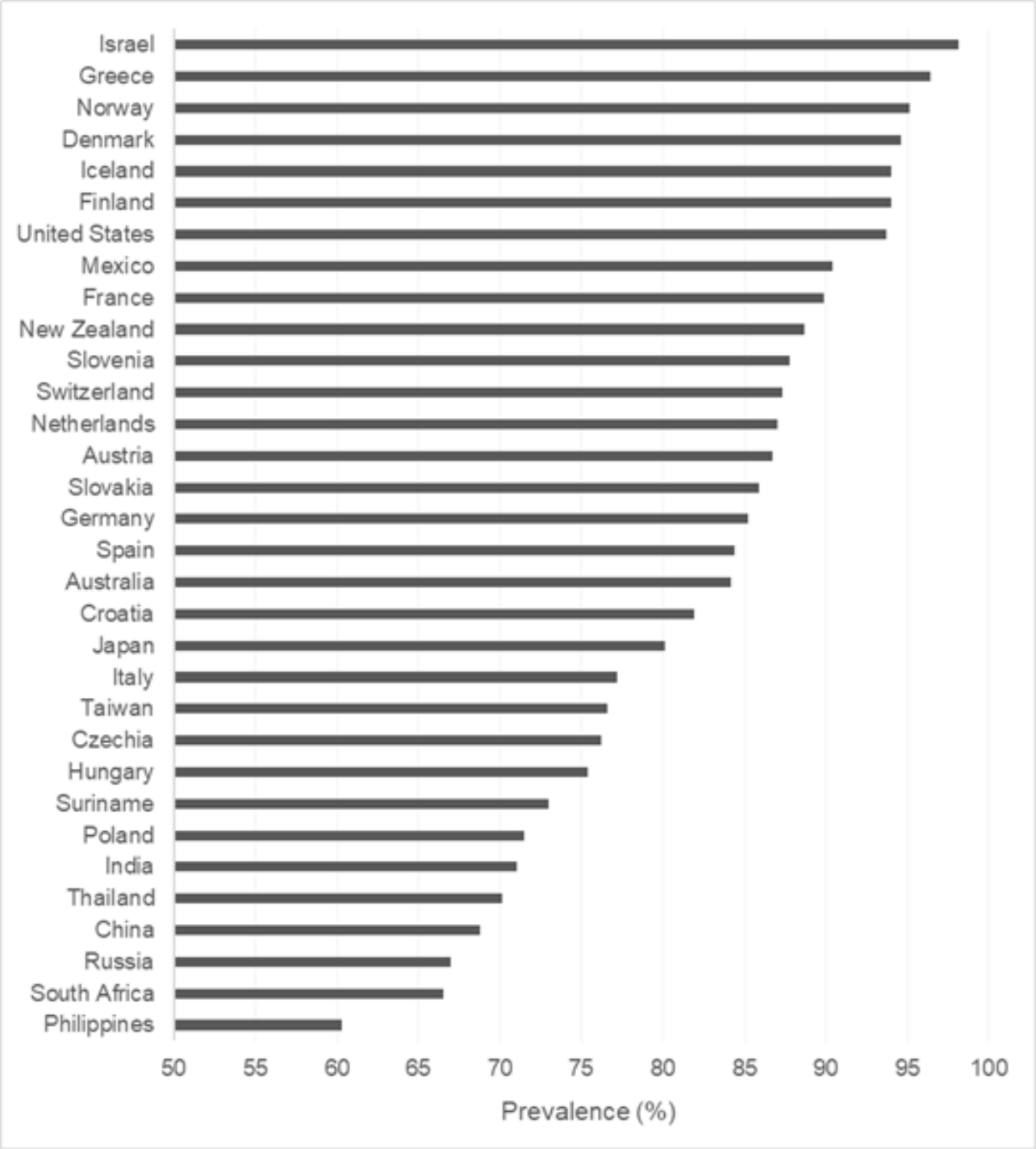

When restricted to individuals with internet access, prevalence levels were higher across all countries, with particularly large differences observed in lower-access contexts (Table S3 in Multimedia Appendix 1).

Individual-Level Associations

A null model indicated substantial between-country variation in OHISB. The ICC was 0.177 (95% CI 0.115 - 0.264), corresponding to approximately 18% of the total variance attributable to differences between countries.

Sequential inclusion of individual-level variables improved model fit. Individual-level factors explained approximately one quarter of the between-country variance in OHISB (Table S8 in Multimedia Appendix 1).

No significant overall temporal effects were observed; fieldwork year was not systematically associated with OHISB.

Sociodemographic factors showed consistent patterns. Individuals with higher education had significantly higher odds of OHISB (odds ratio [OR] 1.46, 95% CI 1.43 - 1.49), while older respondents showed lower odds compared to younger age groups (OR 0.66, 95% CI 0.64 - 0.67). Women reported higher engagement than men (OR 1.38, 95% CI 1.28 - 1.49). Being partnered or economically active also had modest positive effects, as shown in Figure 2.

XSL•FO
RenderX

Health-related characteristics were also associated with OHISB. Individuals reporting poorer subjective health (OR 1.09, 95% CI 1.04 - 1.13), the presence of health problems (OR 1.62, 95% CI 1.48 - 1.77), and lower levels of well-being (OR 0.85, 95% CI 0.81 - 0.90) were more likely to seek health information online. Health behaviors such as smoking, alcohol consumption, and physical activity showed more modest and less consistent associations.

Health care-related variables revealed that individuals with more frequent contact with health care services (OR 1.15, 95% CI 1.10 - 1.20), unmet health care needs (OR 1.38, 95% CI 1.24 - 1.53), and higher levels of trust in the health care system (OR 1.05, 95% CI 1 - 1.10) were more likely to engage in OHISB. In contrast, individuals reporting higher satisfaction with their most recent doctor visit or with health care overall were less likely to seek health information online.

Psycho-motivational factors exhibited some of the strongest associations. Perceiving the internet as a useful health-related tool (OR 1.56, 95% CI 1.50 - 1.62) and higher perceived reliability of online health information (OR 1.07, 95% CI 1.04 - 1.11) were both strongly associated with a greater likelihood of OHISB.

Figure 2. Odds ratios with 95% CIs for individual-level factors associated with online health information–seeking behavior. Estimates are derived from multilevel logistic regression models including all individual-level variables. Odds ratios greater than 1 indicate a higher likelihood of engaging in online health information–seeking behavior.

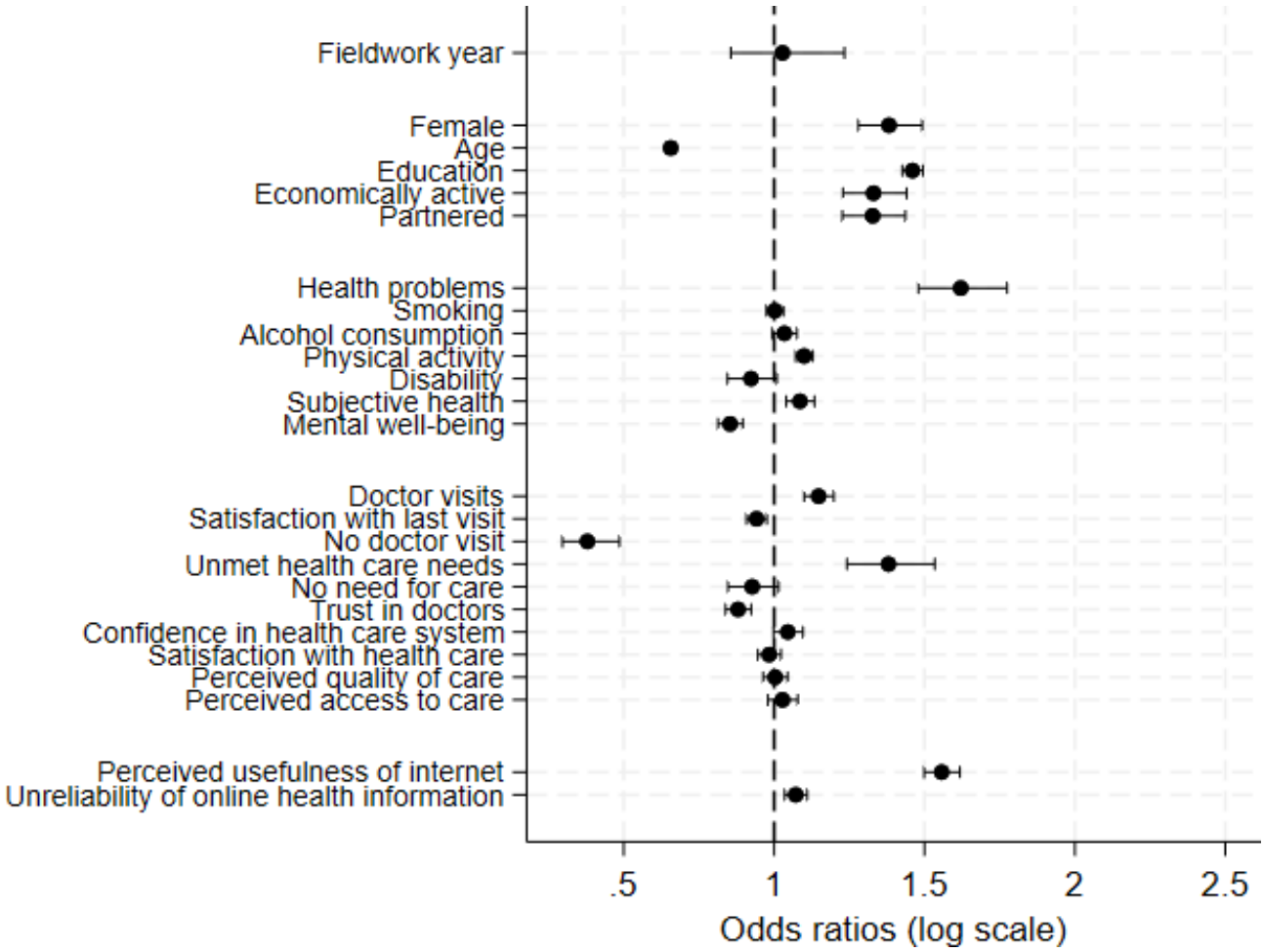

Country-Level Factors

Country-level structural conditions were associated with variation in OHISB across countries (Table 1). In particular, higher levels of socioeconomic and health development were linked to a greater likelihood of OHISB, with a one standard deviation increase in the index corresponding to substantially higher odds (OR=1.99, 95% CI 1.63 - 2.45), with a PCV of 63%.

XSL•FO
RenderX

Table . Associations between country-level factors and online health information–seeking behavior.

Model	Country-level factor	OR[a,b] (95% CI)	*P* value	BIC[c]	PCV[d]
Individual model	—[e]	—	—	20737.6	—
Health system evaluation	—	1.24 (0.94-1.63)	.13	20745.9	7
Socioeconomic and health development	—	1.99 (1.63-2.45)	<.001	20719.3	63
Cultural hierarchy–individualism index	—	1.89 (1.51-2.37)	<.001	20721.2	59
Unhealthy lifestyle	—	1.13 (0.83-1.54)	.42	20747.3	3
Final model	Socioeconomic and health development	1.52 (1.15-2)	.003	20726.0	76
	Cultural hierarchy–individualism index	1.37 (0.99-1.89)	.06	—	—

[a]OR: odds ratio.

[b]Note: Odds ratios represent the change in odds associated with a one SD increase in each country-level index. Proportional change in variance indicates the proportion of between-country variance explained relative to the individual-level model.

[c]BIC: Bayesian information criterion.

[d]PCV: proportional change in variance.

[e]Not available.

The cultural hierarchy–individualism index was also positively associated with OHISB (OR 1.89, 95% CI 1.51 - 2.37).

In contrast, health system evaluation and unhealthy lifestyle patterns showed weak or non-significant associations with OHISB and did not substantially improve model fit. However, the health system evaluation index accounted for a small proportion of between-country variance (PCV=7%) (Table 1 and Tables S10a–S10b in Multimedia Appendix 1).

When both indices were included simultaneously, socioeconomic and health development remained associated with OHISB (OR 1.52, 95% CI 1.15 - 2.00), whereas the association of cultural orientation was attenuated and no longer statistically significant (OR 1.37, 95% CI 0.99 - 1.89). Including the cultural index provided only limited additional improvement in model fit beyond socioeconomic and health development, increasing the explained between-country variance from 63% to 76% (Table 1 and Table S11 in Multimedia Appendix 1).

Cross-Level Interactions

Exploratory analyses identified cross-level interactions between selected country-level indices and individual-level variables (Table S12 in Multimedia Appendix 1).

First, the effect of gender differed by national development. The gender gap in OHISB, favoring women, increased with higher levels of socioeconomic and health development (OR 1.17, 95% CI 1.03 - 1.31). As illustrated in Figure 3, gender differences were minimal in low-development contexts but became substantially larger in highly developed countries, where women were markedly more likely than men to seek health information online.

XSL•FO
RenderX

Figure 3. Cross-level interaction between socioeconomic and health development and gender: marginal effects on OHISB with 95% CIs. Predicted probabilities are derived from multilevel logistic regression models including individual-level variables and the socioeconomic and health development index. Error bars indicate 95% CIs. OHISB: online health information–seeking behavior.

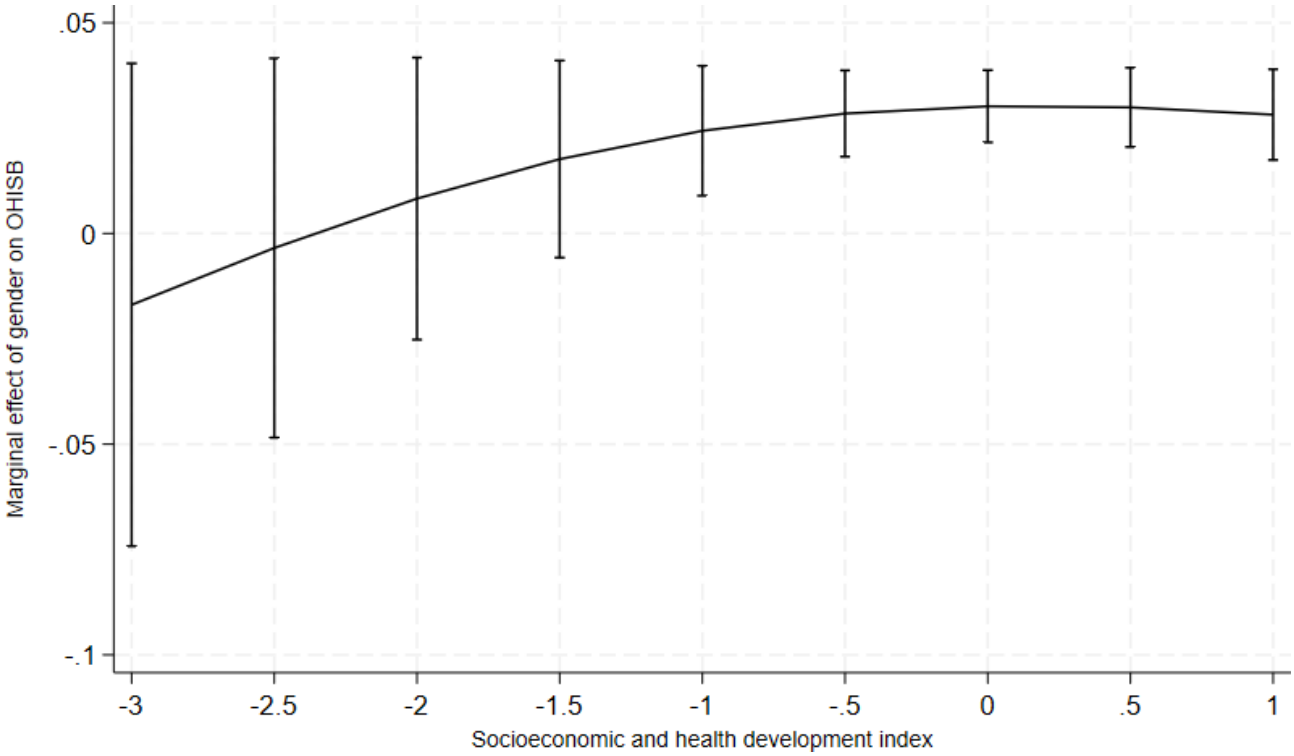

Second, perceived usefulness of the internet showed a context-dependent association. Its relationship with OHISB generally strengthened with increasing levels of national development but was not strictly monotonic (OR 1.09, 95% CI 1.05 - 1.14). As shown in Figure 4, the effect of perceived usefulness was positive across all contexts, peaked in moderately developed countries, and slightly attenuated at the highest levels of development.

Figure 4. Cross-level interaction between socioeconomic and health development and perceived usefulness of the internet: marginal effects on online health information–seeking behavior with 95% CIs. Predicted probabilities are derived from multilevel logistic regression models including individual-level variables and the socioeconomic and health development index. Error bars indicate 95% CIs. OHISB: online health information–seeking behavior.

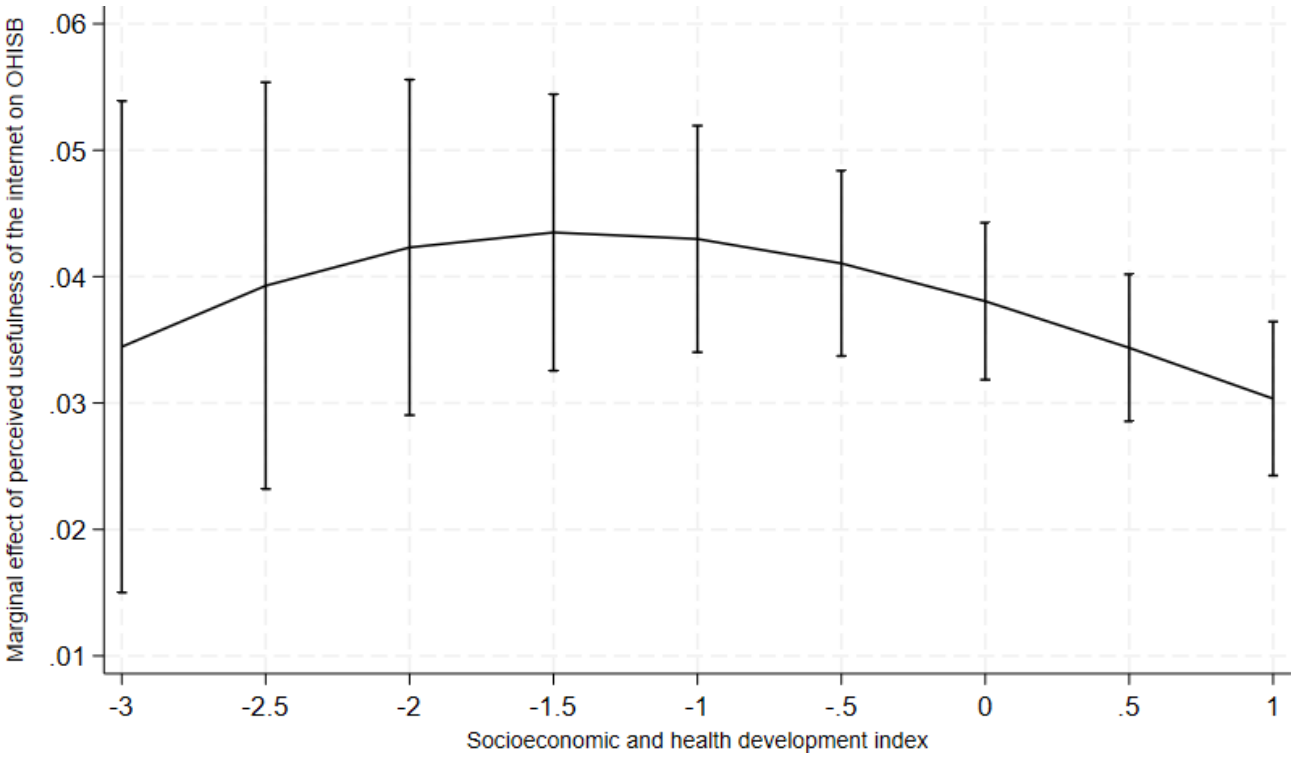

XSL•FO
RenderX

Robustness Checks

To evaluate the stability of the findings, a series of robustness checks were conducted.

First, all analyses were replicated after excluding respondents without internet access. This restriction also reduced missingness in psycho-motivational variables to below 5% (typically around 2%). The re-estimated models produced substantively identical results. The direction and significance of individual-level associations remained unchanged, and country-level effects were stable. Although between-country variance and the ICC increased slightly, the overall pattern of results was unaffected.

Second, all models were estimated with and without the combined ISSP weight (WEIGHT_COM). For weighted analyses, the weights were normalized within countries (mean=1) and applied at the individual level, while countries remained equally weighted. The results were substantively identical across specifications: coefficients remained stable in both magnitude and significance, and the pattern of between-country variation was unchanged.

Third, individual country-level indicators were examined separately in models including all individual-level variables to assess the robustness of contextual effects. These analyses yielded consistent results: key indicators related to socioeconomic development, digital access, and population health were positively associated with OHISB and explained a substantial share of the between-country variation.

Discussion

Principal Findings

This study shows that OHISB follows well-established social gradients in digital and health-related behavior. Across 32 countries, younger, more educated, and female respondents were more likely to search for health information online, consistent with prior research on generational, educational, and gendered patterns in digital engagement [4,18]. These findings align with broader research on digital inequality.

Health-related factors were also associated with OHISB. Individuals reporting health problems or poorer mental well-being were more likely to search for health information online, suggesting that OHISB may reflect both coping-related and self-management–oriented responses to health needs. This pattern is consistent with previous research showing that health concerns and perceived vulnerability are important motivations for online health information seeking [3,21].

While individuals with more frequent contact with health care services were more likely to engage in OHISB, higher satisfaction with health care experiences was associated with a lower likelihood of online health information seeking. This pattern suggests that online health information seeking may serve both complementary and substitutive functions. Individuals dissatisfied with health care encounters may be more likely to seek additional information online, either to verify medical advice or to compensate for unmet informational needs, whereas those with more satisfactory health care experiences may rely more on formal sources of care and information [91,92].

Psycho-motivational variables showed some of the strongest associations. Perceived usefulness of the internet and higher perceived reliability of online health information were both positively associated with OHISB, underscoring the importance of digital familiarity and trust in shaping engagement with online health resources. These findings are consistent with information-seeking models that emphasize perceived utility and credibility of information sources as relevant factors of health information behavior [22,24].

At the country level, socioeconomic and health development emerged as the most robust contextual factor associated with OHISB. Individuals in more developed contexts were more likely to engage in online health information seeking, suggesting that structural conditions such as digital infrastructure, educational attainment, and population health are associated with environments that facilitate engagement with online health resources. In addition, this factor accounted for a substantial share of between-country variation.

These results align with research on the digital divide, which emphasizes that inequalities in digital engagement are not solely determined by individual characteristics but are also embedded in societal contexts [12]. Even when individuals have access to the internet, differences in national development shape the opportunities and conditions under which online health information can be accessed and used.

Cultural context also showed a meaningful association with OHISB when considered separately, with higher levels of individualism associated with greater engagement in online health information seeking. This finding suggests that cultural norms related to autonomy, self-reliance, and individual responsibility for health may play a role in shaping health information behavior. However, when considered jointly with socioeconomic development, the association of cultural orientation was attenuated, likely reflecting the empirical overlap between cultural and structural dimensions, as more economically developed countries also tend to exhibit higher levels of individualism [37].

Health system evaluation and unhealthy lifestyle patterns were not statistically significantly associated with OHISB. However, the health system evaluation index accounted for a small proportion of between-country variance, which may reflect limited statistical power at the country level rather than the absence of contextual relevance.

The analysis of cross-level interactions suggests that structural conditions may be associated with variation in the strength of selected individual-level factors. In particular, gender differences were more pronounced in highly developed contexts, with women showing higher levels of engagement relative to men. Similarly, the association between perceived usefulness of the internet and OHISB was generally stronger in higher-development settings. These findings suggest that structural conditions may be associated with differences in behavioral patterns between population groups and in the salience of motivational factors. However, these findings should be interpreted with caution, given the relatively small number of country-level units.

XSL•FO
RenderX

By combining nationally representative data from 32 countries with a multilevel framework, this study extends prior descriptive cross-national research and provides a more integrated assessment of how individual- and country-level factors are jointly associated with OHISB.

Limitations and Future Research

However, several limitations should be acknowledged. First, the measurement of OHISB is necessarily broad and does not capture the heterogeneity of online health information seeking. The binary classification of respondents into health-onliners and health-offliners aggregates a wide range of behaviors, including searching for information related to prevention, symptoms, treatment, and mental health. These different forms of engagement may reflect distinct motivations and stages of health-related decision-making, ranging from primary prevention to disease management [93]. This aggregation may also obscure differences in the frequency and intensity of online health information seeking, which may vary systematically across social groups and country contexts. Although more fine-grained measures could capture additional behavioral variation, their cross-national comparability is limited [52]. Future research may benefit from examining specific domains of OHISB in more homogeneous contexts where measurement comparability can be more readily established.

Second, the binary operationalization of OHISB does not distinguish between structurally constrained nonseekers (eg, individuals without internet access) and motivationally disengaged non-seekers who have access but choose not to seek health information online. Although sensitivity analyses excluding respondents without internet access yielded substantively similar results, suggesting that the main findings are not driven solely by access barriers, these groups may differ in underlying mechanisms and responsiveness to policy interventions.

Third, the data do not allow differentiation between specific platforms or sources of online health information. Individuals may rely on diverse digital environments, including search engines, apps, social media, and online forums, which differ in credibility, accessibility, and informational quality. This heterogeneity may influence both the likelihood of seeking information and the potential benefits derived from it, but cannot be directly assessed in this study. Future research could also examine whether the observed patterns extend to emerging forms of digital health engagement, such as telemedicine, digital therapeutics, or AI-assisted health tools.

Fourth, the cross-sectional design of the study precludes causal interpretations of the observed associations. Relationships between OHISB and individual-level factors, particularly psycho-motivational variables such as perceived usefulness, may be bidirectional [94,95], and reverse causality cannot be ruled out.

Fifth, although the analysis includes a large number of countries, the number of level-2 units remains relatively limited for detecting complex cross-level interactions [84]. In addition, cross-level interactions were explored across multiple specifications, which increases the risk of Type I error. As a result, findings related to contextual moderation should be considered tentative.

Sixth, the analysis relies on listwise deletion, which resulted in the exclusion of a substantial proportion of cases. Although additional analyses indicated that the main findings were robust across alternative specifications, the assumption that missing data are random cannot be fully verified, and some degree of selection bias may remain. In particular, excluded respondents were more likely to belong to socially disadvantaged groups, which may lead to a conservative estimate of inequalities in OHISB.

Finally, the data were collected between 2021 and 2024, a period marked by the COVID-19 pandemic and its aftermath. Changes in health care access and increased reliance on digital resources during this period may have influenced patterns of online health information seeking. Because countries participated in different data collection waves, country-specific temporal variation related to differences in fieldwork timing cannot be fully ruled out. In addition, the rapid emergence and growing use of AI-assisted health information tools postdates much of the data collection and may affect the generalizability of the findings to current digital health environments.

Future research should extend these findings using longitudinal or repeated cross-sectional designs to better capture temporal dynamics in digital health engagement and strengthen causal inference. More granular macro-level indicators, such as digital skills distributions, health literacy environments, institutional trust, and digital governance, may help to identify additional contextual mechanisms underlying cross-national differences. Future research should also consider regulatory and legislative contexts shaping digital health information environments, including policies governing online health content and the control of misinformation.

Practical Implications

The findings have several practical implications for efforts to reduce digital health inequalities. First, they highlight that improving access to digital health information alone is insufficient. Structural conditions, such as socioeconomic development, digital infrastructure, and population health, are associated with levels of engagement with online health resources. Policies aimed at reducing digital health inequalities should therefore combine improvements in access with broader investments in education, digital skills, and health literacy.

Second, the findings suggest that contextual conditions related to institutional trust, health system performance, and broader structural development may be relevant for engagement with online health information. This indicates that strengthening trust in health institutions and ensuring the availability of reliable, high-quality online health information may support more effective engagement with digital health resources.

Third, the findings on cross-level interactions indicate that the strength of individual-level associations may vary across contexts. In particular, gender differences and the role of perceived usefulness were more pronounced in more developed settings. These patterns suggest that differences in engagement between population groups may vary across contexts, indicating

XSL•FO
RenderX

that interventions should be sensitive to contextual conditions rather than assuming uniform patterns across countries.

Finally, given the increasing importance of digital health technologies, including telemedicine and AI-driven tools, efforts to promote equitable digital health engagement should address both individual-level barriers and structural inequalities. As these technologies continue to evolve rapidly, they may further reshape how individuals access and evaluate health information, highlighting the need for policies that support both digital inclusion and the quality and trustworthiness of online health content. A comprehensive approach that integrates digital inclusion, health system development, and trust-building may be necessary to ensure that the benefits of digital health are distributed more evenly across populations.

Conclusions

This study demonstrates that OHISB is associated with both individual characteristics and broader structural conditions. By integrating a large and geographically diverse set of countries within a multilevel framework, this study provides a more comprehensive assessment of how individual- and country-level factors jointly contribute to cross-national variation in digital health engagement. These findings highlight that digital health inequalities are embedded in structural contexts and underscore the importance of addressing both individual capabilities and broader contextual conditions to support more equitable engagement with digital health information.

Funding

The work of the PR was supported by [Research of Excellence on Digital Technologies and Wellbeing / CZ.02.01.01/00/22_008/0004583], which is co-financed by the European Union. EL declares that no external funding was received for her contribution.

Data Availability

The datasets analyzed in this study are publicly available from the GESIS Data Archive [49-51]. All datasets contain fully anonymized data and comply with applicable data protection standards. Additional materials, including analysis scripts, replication files, and output tables, are available via the Open Science Framework [83].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables presenting detailed sample characteristics, descriptive statistics, country-level indicators, and full results of multilevel regression and sensitivity analyses.
[PDF File, 792 KB - jmir_v28i1e88110_app1.pdf]

References

1. Qiu J, Zhang T, Xu Z, Fu Y, Mi Q, Li S. Understanding users' online health information seeking switching intention: from traditional health information platforms to GAI. Online Inf Rev 2026 Jan 6;50(1):18-36. [doi: 10.1108/OIR-11-2024-0706]
2. Zimmerman MS, Shaw G. Health information seeking behaviour: a concept analysis. Health Info Libr J 2020 Sep;37(3):173-191. [doi: 10.1111/hir.12287] [Medline: 32052549]
3. Bachl M, Link E, Mangold F, Stier S. Search engine use for health-related purposes: behavioral data on online health information-seeking in Germany. Health Commun 2024 Jul;39(8):1651-1664. [doi: 10.1080/10410236.2024.2309810] [Medline: 38326714]
4. Wang X, Shi J, Kong H. Online health information seeking: a review and meta-analysis. Health Commun 2021 Sep;36(10):1163-1175. [doi: 10.1080/10410236.2020.1748829] [Medline: 32290679]
5. Bujnowska-Fedak MM, Węgierek P. The Impact of online health information on patient health behaviours and making decisions concerning health. Int J Environ Res Public Health 2020 Jan 31;17(3):880. [doi: 10.3390/ijerph17030880] [Medline: 32023828]
6. Ayers SL, Kronenfeld JJ. Chronic illness and health-seeking information on the Internet. Health (London) 2007 Jul;11(3):327-347. [doi: 10.1177/1363459307077547] [Medline: 17606698]
7. Nickel PJ, Loosman I, Frank L, Vinnikova A. Justice and empowerment through digital health: ethical challenges and opportunities. DISO 2023 Dec;2(3):42. [doi: 10.1007/s44206-023-00075-x]
8. Sharma S, Kar AK, Gupta MP, Dwivedi YK, Janssen M. Digital citizen empowerment: a systematic literature review of theories and development models. Information Technology for Development 2022 Oct 2;28(4):660-687. [doi: 10.1080/02681102.2022.2046533]
9. Jia X, Pang Y, Liu LS. Online health information seeking behavior: a systematic review. Healthcare (Basel) 2021 Dec 16;9(12):1740. [doi: 10.3390/healthcare9121740] [Medline: 34946466]

XSL•FO
RenderX

10. Jahnel T, Dassow HH, Gerhardus A, Schüz B. The digital rainbow: digital determinants of health inequities. Digit Health 2022;8:20552076221129093. [doi: 10.1177/20552076221129093] [Medline: 36204706]
11. DiMaggio P, Hargittai E, Celeste C, Shafer S. From unequal access to differentiated use: a literature review and agenda for research on digital inequality. EconPapers. URL: https://EconPapers.repec.org/RePEc:pri:cpanda:29 [accessed 2026-05-02]
12. Van Dijk J. The Digital Divide: Polity Press; 2020.
13. Hargittai E. The digital divide and what to do about it. In: Jones DC, editor. New Economy Handbook: Emerald Group Publishing Limited; 2003:821-839.
14. Powell J, Deetjen U. Characterizing the digital health citizen: mixed-methods study deriving a new typology. J Med Internet Res 2019 Mar 5;21(3):e11279. [doi: 10.2196/11279] [Medline: 30835238]
15. Al Shamsi H, Almutairi AG, Al Mashrafi S, Al Kalbani T. Implications of language barriers for healthcare: a systematic review. Oman Med J 2020 Mar;35(2):e122-e122. [doi: 10.5001/omj.2020.40] [Medline: 32411417]
16. Turavinina D, Amornkitvikai Y. Internet health information-seeking behavior and the use of traditional and complementary medicine: the role of online engagement and perceived information usefulness and reliability. BMC Complement Med Ther 2025 Nov 24;25(1):431. [doi: 10.1186/s12906-025-05167-4] [Medline: 41286849]
17. Stifjell K, Sandanger TM, Wien C. Exploring online health information-seeking behavior among young adults: scoping review. J Med Internet Res 2025 Sep 9;27:e70379. [doi: 10.2196/70379] [Medline: 40925001]
18. Zhao YC, Zhao M, Song S. Online health information seeking behaviors among older adults: systematic scoping review. J Med Internet Res 2022 Feb 16;24(2):e34790. [doi: 10.2196/34790] [Medline: 35171099]
19. Renahy E, Parizot I, Chauvin P. Health information seeking on the Internet: a double divide? results from a representative survey in the Paris metropolitan area, France, 2005-2006. BMC Public Health 2008 Feb 21;8(1):69. [doi: 10.1186/1471-2458-8-69] [Medline: 18291024]
20. Link E, Baumann E, Klimmt C. Explaining online information seeking behaviors in people with different health statuses: German representative cross-sectional survey. J Med Internet Res 2021 Dec 10;23(12):e25963. [doi: 10.2196/25963] [Medline: 34890348]
21. McMullan RD, Berle D, Arnáez S, Starcevic V. The relationships between health anxiety, online health information seeking, and cyberchondria: systematic review and meta-analysis. J Affect Disord 2019 Feb 15;245:270-278. [doi: 10.1016/j.jad.2018.11.037] [Medline: 30419526]
22. Kahlor LA. PRISM: a planned risk information seeking model. Health Commun 2010 Jun;25(4):345-356. [doi: 10.1080/10410231003775172] [Medline: 20512716]
23. Basnyat I, Nekmat E, Jiang S, Lin J. Applying the modified comprehensive model of information seeking to online health information seeking in the context of India. J Health Commun 2018;23(6):563-572. [doi: 10.1080/10810730.2018.1493058] [Medline: 29979921]
24. Ahadzadeh AS, Pahlevan Sharif S, Ong FS, Khong KW. Integrating health belief model and technology acceptance model: an investigation of health-related internet use. J Med Internet Res 2015 Feb 19;17(2):e45. [doi: 10.2196/jmir.3564] [Medline: 25700481]
25. Demirci Ş, Uğurluoğlu Ö, Konca M, Çakmak C. Socio‐demographic characteristics affect health information seeking on the internet in Turkey. Health Info Libr J 2021 Dec;38(4):304-312. [doi: 10.1111/hir.12358]
26. Nangsangna RD, da-Costa Vroom F. Factors influencing online health information seeking behaviour among patients in Kwahu West Municipal, Nkawkaw, Ghana. Online J Public Health Inform 2019;11(2):e13. [doi: 10.5210/ojphi.v11i2.10141] [Medline: 31632607]
27. Individuals using the internet for seeking health-related information. Eurostat. 2024. URL: https://ec.europa.eu/eurostat/databrowser/view/tin00101/default/table?lang=en&category=t_isoc.t_isoc_i.t_isoc_iiu [accessed 2024-12-09]
28. Alvarez-Galvez J, Salinas-Perez JA, Montagni I, Salvador-Carulla L. The persistence of digital divides in the use of health information: a comparative study in 28 European countries. Int J Public Health 2020 Apr;65(3):325-333. [doi: 10.1007/s00038-020-01363-w] [Medline: 32277246]
29. Andreassen HK, Bujnowska-Fedak MM, Chronaki CE, et al. European citizens' use of E-health services: a study of seven countries. BMC Public Health 2007 Apr 10;7(1):53. [doi: 10.1186/1471-2458-7-53] [Medline: 17425798]
30. Kummervold PE, Chronaki CE, Lausen B, et al. eHealth trends in Europe 2005-2007: a population-based survey. J Med Internet Res 2008 Nov 17;10(4):e42. [doi: 10.2196/jmir.1023] [Medline: 19017584]
31. Kummervold PE, Wynn R. Health information accessed on the internet: the development in 5 European countries. Int J Telemed Appl 2012;2012:297416. [doi: 10.1155/2012/297416] [Medline: 23304133]
32. Santana S, Lausen B, Bujnowska-Fedak M, Chronaki CE, Prokosch HU, Wynn R. Informed citizen and empowered citizen in health: results from an European survey. BMC Fam Pract 2011 Apr 16;12(1):20. [doi: 10.1186/1471-2296-12-20] [Medline: 21496309]
33. Link E, Baumann E, Linn A, Fahr A, Schulz P, Abuzahra ME. Influencing factors of online health information seeking in selected European countries. EJHC 2021;2(1):29-55. [doi: 10.47368/ejhc.2021.002]
34. Bachl M. Online health information seeking in Europe: do digital divides persist? SC|M|SC|M 2016;5(4):427-453. [doi: 10.5771/2192-4007-2016-4-427]

XSL•FO
RenderX

35. Bundorf MK, Wagner TH, Singer SJ, Baker LC. Who searches the internet for health information? Health Serv Res 2006 Jun;41(3 Pt 1):819-836. [doi: 10.1111/j.1475-6773.2006.00510.x] [Medline: 16704514]
36. Lin HC, Ho WH. Cultural effects on use of online social media for health-related information acquisition and sharing in Taiwan. International Journal of Human–Computer Interaction 2018 Nov 2;34(11):1063-1076. [doi: 10.1080/10447318.2017.1413790]
37. Hofstede G. The business of international business is culture. Int Bus Rev 1994 Mar;3(1):1-14. [doi: 10.1016/0969-5931(94)90011-6]
38. Siau K, Erickson J, Nah FFH. Effects of national culture on types of knowledge sharing in virtual communities. IEEE Trans Profess Commun 2010 Sep;53(3):278-292. [doi: 10.1109/TPC.2010.2052842]
39. Kim HS, Sherman DK, Taylor SE. Culture and social support. Am Psychol 2008 Sep;63(6):518-526. [doi: 10.1037/0003-066X] [Medline: 18793039]
40. Büchi M, Just N, Latzer M. Modeling the second-level digital divide: a five-country study of social differences in internet use. New Media & Society 2016 Dec;18(11):2703-2722. [doi: 10.1177/1461444815604154]
41. Van Deursen A, Helsper EJ. The third-level digital divide: who benefits most from being online? In: Communication and Information Technologies Annual: Emerald Group Publishing Limited; 2015:29-52. [doi: 10.1108/S2050-206020150000010002]
42. Van Dijk J, Van Deursen A. Digital Skills: Unlocking the Information Society: Palgrave Macmillan; 2014. [doi: 10.1057/9781137437037]
43. Scheerder A, van Deursen A, van Dijk J. Determinants of internet skills, uses and outcomes. A systematic review of the second- and third-level digital divide. Telematics and Informatics 2017 Dec;34(8):1607-1624. [doi: 10.1016/j.tele.2017.07.007]
44. Liobikienė G, Bernatonienė J. The determinants of access to information on the internet and knowledge of health related topics in European countries. Health Policy 2018 Dec;122(12):1348-1355. [doi: 10.1016/j.healthpol.2018.09.019] [Medline: 30337158]
45. Denniss E, Lindberg R. Social media and the spread of misinformation: infectious and a threat to public health. Health Promot Int 2025 Mar 5;40(2):daaf023. [doi: 10.1093/heapro/daaf023] [Medline: 40159949]
46. Fernandes FA, Chaltikyan G, Adib K, Caton-Peters H, Novillo-Ortiz D. The role of governance in the digital transformation of healthcare: results of a survey in the WHO Europe region. Int J Med Inform 2024 Sep;189:105510. [doi: 10.1016/j.ijmedinf.2024.105510] [Medline: 38901269]
47. Wang M, Lu X, Du Y, et al. Digital health governance in China by a whole-of-society approach. NPJ Digit Med 2025 Aug 2;8(1):496. [doi: 10.1038/s41746-025-01876-9] [Medline: 40750663]
48. Ou M, Ho SS. A meta-analysis of factors related to health information seeking: an integration from six theoretical frameworks. Communic Res 2022 Jun;49(4):567-593. [doi: 10.1177/00936502211043024]
49. ISSP Research Group. International Social Survey Programme: Health and Health Care II - ISSP 2021 (ZA8000; version 2.0.0) [Dataset]. GESIS, Cologne. 2024. URL: https://doi.org/10.4232/5.ZA8000.2.0.0 [accessed 2026-05-07]
50. Andreadis I. International Social Survey Programme: Health and Health Care II - ISSP 2021 (Greece) [Dataset]. GESIS, Cologne. 2024.
51. Centro de Investigaciones Sociológicas. International Social Survey Programme: Health and Health Care II - ISSP 2021 (Spain) (ZA8872; version 1.1.0) [Dataset]. GESIS, Cologne. 2024. URL: https://doi.org/10.4232/1.14454 [accessed 2026-05-07]
52. Raudenská P. Cross-national validation of digital health engagement scales: evidence from 30 countries. Front Public Health 2026;14:1783814. [doi: 10.3389/fpubh.2026.1783814] [Medline: 41923762]
53. Health and social employment. OECD. URL: https://data-explorer.oecd.org/?fs[0]=Topic%2C0%7CHealth%23HEA%23&pg=0&fc=Topic&bp=true&snb=85 [accessed 2024-10-26]
54. Number of physicians in Taiwan from 2014 to 2024. Statista. 2024. URL: https://www.statista.com/statistics/860217/taiwan-physician-number/ [accessed 2024-10-26]
55. Physicians (per 1000 people). The World Bank. URL: https://data.worldbank.org/indicator/SH.MED.PHYS.ZS?locations=SK-PH-RU-TH&name_desc=false [accessed 2024-10-26]
56. General data on Danish healthcare system. Ministry of Foreign Affairs of Denmark. 2024. URL: https://www.jmari.med.or.jp/wp-content/uploads/2024/05/Basic-Information-Danish-Healthcare-1.pdf [accessed 2024-10-26]
57. Hospitals. OECD. URL: https://data-explorer.oecd.org/?fs[0]=Topic%2C0%7CHealth%23HEA%23&pg=0&fc=Topic&bp=true&snb=85 [accessed 2024-10-26]
58. Number of hospitals in China from 2000 to 2024. Statista. 2025. URL: https://www.statista.com/statistics/279322/number-of-hospitals-in-china/ [accessed 2025-01-08]
59. Estimated number of public and private hospitals across India in 2019. Statista. 2025. URL: https://www.statista.com/statistics/1128425/india-number-of-public-and-private-hospitals-estimated/ [accessed 2025-01-08]
60. Number of hospitals in Taiwan from 2014 to 2024. Statista. URL: https://www.statista.com/statistics/934717/taiwan-number-of-hospitals/ [accessed 2025-01-08]

XSL•FO
RenderX

61. Number of hospitals in Russia from 2014 to 2029. Statista. 2025. URL: https://www.statista.com/forecasts/1166252/hospital-count-forecast-in-russia [accessed 2025-01-08]
62. Number of hospitals in Thailand from 2014 to 2023. Statista. 2025. URL: https://www.statista.com/statistics/1114483/thailand-number-of-hospitals/ [accessed 2025-01-08]
63. Density of hospitals (per 100,000 inhabitants) in South Africa in 2021. Statista. 2025. URL: https://www.statista.com/outlook/hmo/hospitals/south-africa [accessed 2025-01-08]
64. Weiss DJ, Nelson A, Vargas-Ruiz CA, et al. Global maps of travel time to healthcare facilities. Nat Med 2020 Dec;26(12):1835-1838. [doi: 10.1038/s41591-020-1059-1] [Medline: 32989313]
65. Dattani S, Rodés-Guirao L, Ritchie H, Ortiz-Ospina E, Roser M. Life expectancy. Our World in Data. 2023. URL: https://ourworldindata.org/life-expectancy [accessed 2026-05-02]
66. Healthy life expectancy (HALE) at birth (years). World Health Organization. 2024. URL: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/gho-ghe-hale-healthy-life-expectancy-at-birth [accessed 2024-10-26]
67. Results of a survey on smoking behavior among Chinese people [Web page in Chinese]. Health Promotion Administration. 2024. URL: https://www.hpa.gov.tw/Pages/Detail.aspx?nodeid=1718&pid=9913 [accessed 2024-10-26]
68. WHO report on the Global Tobacco Epidemic, 2021. : World Health Organization; 2021 URL: https://cdn.who.int/media/docs/default-source/country-profiles/tobacco/who_rgte_2021_suriname.pdf? [accessed 2025-01-08]
69. SDG target 3.a tobacco control. World Health Organization. 2024. URL: https://www.who.int/data/gho/data/themes/topics/sdg-target-3_a-tobacco-control [accessed 2024-10-26]
70. Chen RH, Chang HY, Hsu YT, Chen WJ, Chen CY. Harm from others' drinking among young adults in Taiwan: predictors and deteriorating quality of life. Drug Alcohol Rev 2024 Sep;43(6):1483-1492. [doi: 10.1111/dar.13903] [Medline: 38982724]
71. Total alcohol consumption per capita (liters of pure alcohol, projected estimates, 15+ years of age). World Bank. 2024. URL: https://data.worldbank.org/indicator/SH.ALC.PCAP.LI [accessed 2024-06-27]
72. Households with internet access at home. DataHub. 2024. URL: https://datahub.itu.int/data/?i=12047 [accessed 2024-10-26]
73. Number of internet users in Taiwan from January 2015 to January 2025. Statista. URL: https://www.statista.com/statistics/1296415/taiwan-online-population/ [accessed 2024-10-26]
74. Individuals using the internet (% of population). World Bank. URL: https://data.worldbank.org/indicator/IT.NET.USER.ZS [accessed 2024-10-26]
75. Taiwan GDP per capita. CEIC Data. URL: https://www.ceicdata.com/en/indicator/taiwan/gdp-per-capita [accessed 2026-05-02]
76. GDP per capita (current US$). World Bank. URL: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD [accessed 2024-10-28]
77. Taiwan country report 2024. BTI Transformation Index. URL: https://bti-project.org/en/reports/country-report/TWN?utm_source=chatgpt.com [accessed 2026-05-02]
78. Human development insights. UNDP Human Development Reports. 2022. URL: https://hdr.undp.org/data-center/country-insights#/ranks [accessed 2025-01-09]
79. National health expenditure. Ministry of Health and Welfare. 2022. URL: https://www.mohw.gov.tw/lp-130-2.html [accessed 2024-10-28]
80. Health expenditure as a percentage of gross domestic product (GDP) in selected countries as of 2024. Statista. URL: https://www.statista.com/statistics/268826/health-expenditure-as-gdp-percentage-in-oecd-countries/ [accessed 2024-10-28]
81. Current health expenditure (% of GDP). World Bank. URL: https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS [accessed 2024-10-28]
82. Country comparison tool. The Culture Factor Group. 2021. URL: https://www.theculturefactor.com/country-comparison-tool [accessed 2024-10-26]
83. Structural inequalities in online health information seeking: a cross-national multilevel study. Open Science Framework. 2026. URL: https://doi.org/10.17605/OSF.IO/XBC7T
84. Bryan ML, Jenkins SP. Multilevel modelling of country effects: a cautionary tale. Eur Sociol Rev 2016 Feb;32(1):3-22. [doi: 10.1093/esr/jcv059]
85. Raftery AE. Bayesian model selection in social research. Sociol Methodol 1995;25:111. [doi: 10.2307/271063]
86. Enders CK. Centering predictors and contextual effects. In: Simonoff JS, Scott MA, Marx BD, editors. The SAGE Handbook of Multilevel Modeling: SAGE; 2013:89-107. [doi: 10.4135/9781446247600.n6]
87. Ganzeboom H. Why weights do not matter, but do harm. Zenodo.: Department of Sociology VUA; 2023. URL: https://zenodo.org/records/8024951 [accessed 2026-05-06]
88. Gelman A. Struggles with survey weighting and regression modeling. Statist Sci 2007 May 1;22(2). [doi: 10.1214/088342306000000691]
89. International social survey programme ethical statement. International Social Survey Programme (ISSP). 2022. URL: https://issp.org/wp-content/uploads/2022/02/ethical_statement_issp.pdf [accessed 2026-05-02]
90. Code of ethics for researchers of the Czech Academy of Sciences. Czech Academy of Sciences. 2024. URL: https://www.avcr.cz/en/about-us/legal-regulations/code-of-ethics-for-researchers-of-the-czech-academy-of-sciences/ [accessed 2026-05-02]

XSL•FO
RenderX

91. Lee SY, Hawkins R. Why do patients seek an alternative channel? the effects of unmet needs on patients' health-related internet use. J Health Commun 2010 Mar;15(2):152-166. [doi: 10.1080/10810730903528033] [Medline: 20390984]
92. Tustin N. The role of patient satisfaction in online health information seeking. J Health Commun 2010 Jan;15(1):3-17. [doi: 10.1080/10810730903465491] [Medline: 20390974]
93. AbdulRaheem Y. Unveiling the significance and challenges of integrating prevention levels in healthcare practice. J Prim Care Community Health 2023;14:21501319231186500. [doi: 10.1177/21501319231186500] [Medline: 37449436]
94. Song Y, Min D. The logic behind cyberchondria: longitudinal relations among risk perception, health anxiety, and online health information seeking. J Assoc Inf Sci Technol 2025 Mar;76(3):545-562. [doi: 10.1002/asi.24946]
95. Wang T, Dong Q, Sun Q, et al. Trajectories and influencing factors of online health information-seeking behaviors among community-dwelling older adults: longitudinal mixed methods study. J Med Internet Res 2025 Nov 5;27:e77549. [doi: 10.2196/77549] [Medline: 41191927]

Abbreviations

AI: artificial intelligence
ICC: intraclass correlation coefficient
OHISB: online health information–seeking behavior
OR: odds ratio
PCA: principal component analysis
PCV: proportional change in variance

Edited by M Balcarras; submitted 19.Nov.2025; peer-reviewed by C Lei, C Baxter; revised version received 01.Apr.2026; accepted 01.Apr.2026; published 19.May.2026.

Please cite as:
Raudenská P, Link E
Structural Inequalities in Online Health Information Seeking: Cross-National Multilevel Study
J Med Internet Res 2026;28:e88110
URL: https://www.jmir.org/2026/1/e88110
doi:10.2196/88110

© Petra Raudenská, Elena Link. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Women's Engagement With Different Internet-Enabled Technologies to Access Digital Menopause Information: Mixed Methods, Multiphase Sequential Study

Alison K Osborne[1], PhD; Elizabeth Sillence[1,2], PhD; Caroline Claisse[3], PhD; Abigail C Durrant[3], PhD

[1]Department of Psychology, Faculty of Health and Life Sciences, Northumbria University, Northumberland Building, Newcastle upon Tyne, United Kingdom

[2]School of Psychology, Faculty of Health and Wellbeing, Northumbria University, Newcastle upon Tyne, United Kingdom

[3]Open Lab, School of Computing, Newcastle University, Newcastle upon Tyne, United Kingdom

Corresponding Author:
Elizabeth Sillence, PhD
Department of Psychology, Faculty of Health and Life Sciences, Northumbria University, Northumberland Building, Newcastle upon Tyne, United Kingdom

Abstract

Background: Information on menopause can come from a variety of sources, from friends and family to health care professionals, but increasingly, digital information has become a significant source. Digital information can be accessed online, through websites, social media, podcasts, and online groups or forums. The extent to which digital information on menopause is accessed and consumed can vary widely depending on individual circumstances. Existing literature has focused on investigating a single technology at a time in terms of digital menopause information, rather than exploring a wider ecosystem that people may use. As a result, existing evidence is perhaps limited in the understanding of the role these internet-enabled digital technologies play holistically, specifically for digital menopause information.

Objective: This study explored how women engage with different internet-enabled technologies to access digital menopause information and to understand how the technologies relate to feelings of competence, autonomy, and relatedness. Using a feminist new materialism lens, participants' experiences of different technologies are considered in terms of the affordances of the technology, the associated affective factors, and the agential capacities created.

Methods: Sixteen women aged between 40 and 62 years, who were going through the menopause, took part in a multiphase, mixed methods study. Initially, participants completed an online survey to capture demographics, their perceived knowledge of menopause, and information regarding their existing digital engagement. This was followed by an entry interview to establish the context of their experiences. Participants then completed 3 menopause information gathering tasks using websites, podcasts, and online groups or forums. After each task, participants took part in an interview or completed an online survey to report and reflect on their experiences. Following this, 4 focus groups were run to gather a broader understanding of the role of technology in information seeking.

Results: Websites afforded the greatest accessibility to digital menopause information, with participants reporting higher levels of competence and autonomy, primarily due to familiarity. Podcasts were the most novel of the 3 internet-enabled technologies for participants and also led to greater levels of competence and autonomy, particularly in comparison to the online groups or forums. Most participants found online groups or forums to be overwhelming and difficult to navigate. Across the 3 technologies, affective factors varied, markedly regarding levels of trust in information, but several key sources were relied upon. Engaging with internet-enabled technology for menopause information opened opportunities for participants to access peer experiences, which helped to normalize, validate, and understand their menopause experiences.

Conclusions: It is important to consider how individuals engage with several different internet-enabled technologies for menopause information rather than investigating one at a time. This study highlights the nuances across websites, podcasts, and online groups or forums in terms of familiarity, accessibility, trust, and lived experience.

(J Med Internet Res 2026;28:e78215) doi:10.2196/78215

KEYWORDS

menopause; digital health; women's health; health communication; digital technology; websites; podcasts; online forums; health information seeking behavior; internet; mixed methods research; qualitative research

XSL•FO
RenderX

Introduction

Menopause is characterized by the point in time at which a woman's menstrual periods have stopped for 12 consecutive months [1]. This usually occurs naturally between 44 and 55 years of age but can be highly variable, and menopause can also be induced as a result of surgery, serious illness, or medication [2-4]. In an aging population, women spend one-third of their life perimenopausal, menopausal, or postmenopausal [5]. Perimenopause is the period before this point in time, and women often experience physical and physiological symptoms that can be confusing and affect their quality of life [6-8], including hot flashes, night sweats, weight gain, anxiety, and fatigue [9,10].

Despite the commonality of these experiences, many women report poor knowledge of the menopause transition and are less likely to seek or receive the medical assistance they require [7,8,11]. While menopause information can come from a variety of sources, such as friends, family, and health care professionals, women are increasingly turning to internet-enabled technologies for menopause information to increase their knowledge and understanding. Research has shown that women are higher users of online health information compared to men [12], reflecting a broader shift toward proactive and digitally mediated health engagement for women. This trend underscores the importance of digital health literacy, defined as the ability to seek, understand, and evaluate health information from electronic sources [13], and raises questions about accessibility and equity.

This is in line with rising expectations that individuals take ownership of their health [14], use their own judgment to seek medical advice, question health care professionals about their health [15,16], and increase their own knowledge and awareness [17]. The Health Belief Model suggests that well-informed individuals are more likely to participate in health care decisions and engage in health preventative and treatment behaviors [18]. This also aligns with the National Health Service's (NHS's) commitment to patient-centered, patient-led care, encouraging individuals to be active participants in their health management and decision-making [19]. Consequently, the internet is now often the first place people look for health information compared to friends, family, and health care professionals [20].

Since the COVID-19 pandemic, there has been a further surge in digital health resources to improve health communication. Specific to menopause, these resources, online information, and mHealth (mobile health) apps are often termed "Femtech" or "menotech" [21,22]. New technologies have changed the way we communicate; however, the abundance of menopause-related media content—ranging from podcasts and social media posts to online groups and message boards—is rarely governed by regulations [23]. While this content is often accessible and engaging, the quality and reliability can vary significantly, which may exacerbate inequalities in digital health literacy and access.

Existing evidence is perhaps limited, as noted in a recent literature review by Osborne and Sillence [24] that identified only 14 papers that considered the use of internet-enabled technologies for accessing menopause information. When engaging with these digital technologies for menopause information, women reported finding them accessible, convenient, and easy to use [22,25-28]. The main motivation for engaging with this content was centered on the desire for women to normalize, validate, and legitimize their own menopause experience as well as to understand their symptoms [22,26,29-31]. However, the extent to which someone trusted a source impacted whether an internet-enabled technology was accessed and how subsequent information on menopause was appraised [22,26,28,30-33].

A small body of research has found that general health information presented verbally increased health knowledge [34,35], more than written content [36], suggesting that audio-based materials may be an accessible and understandable means of communicating health information [25]. However, little research has explored how audio media can be specifically used for menopause or how they may serve dual roles in education and entertainment. Only a few papers have specifically considered the role of podcasts in relation to menopause [25,26], highlighting the importance of peer experiences in menopause information. Similarly, online group spaces and forums place peer experiences at the center, but to date, research that has considered this has focused on the data collection of posts and content rather than on women's experiences of consuming the media content on menopause information in these settings [37-39]

Despite the pressure to be informed about one's own health and the breadth of digital resources available on menopause, what remains poorly understood is how different internet-enabled technologies collectively shape women's access to information. Existing research has largely examined individual technologies in isolation, overlooking the interplay between platforms and the contextual factors that influence engagement [24,40]. Examining women's use of different technologies to access menopause information allows the exploration of commonalities in value and approach, as well as the potential to highlight challenges in digital engagement across the provision as a whole. It is important to consider the nuances between how the technologies are perceived and used to understand experiences and better inform future digital solutions regarding accessibility, representation, and sustainability of digital menopause information that facilitates meaningful and informative engagement.

One study that has attempted to examine women's use of different digital technologies is that of Lupton and Maslen [41]. Although not focusing on menopause specifically, Lupton and Maslen [41] explored how Australian women use different types of digital technologies regularly for health-related purposes. Using a feminist new materialism perspective, they focused on the affordances, the associated affective factors, and the agential capacities created by these different technologies (eg, the ability to act confidently and make informed choices when engaging), including health information websites and online discussion forums. This approach to this study ensured rich findings into the nuances and complexity of technology use when engaging in online health information. This methodological perspective illuminates the pervasive nature of internet-enabled technologies for engaging with health information in the context of everyday life.

XSL•FO
RenderX

In this paper, we take a similar holistic approach to examine women's engagement with websites, podcasts, and online forums as part of a digital ecosystem. Drawing from feminist new materialism and self-determination theory [42-44], we explore the affordances, affective factors, and the agential capacities of these technologies, alongside motivations and challenges related to autonomy (agency), competence (ability and effectiveness), and relatedness (connection). This dual theoretical lens enables a nuanced understanding of engagement and informs the design of inclusive, accessible, and trustworthy menopause resources.

This paper aims to explore how women engage with different internet-enabled technologies to access digital menopause information.

Methods

Design

A mixed methods, multiphase sequential design was used, split into 3 main parts: preassessment, primary data collection, and postassessment. Data collection was carried out using interviews, surveys, and focus groups at various stages throughout this study. See Figure 1 for an overview of this study's design. This was an exploratory pilot study designed to generate rich, nuanced insights rather than generalizable findings. The focus was on depth of engagement across multiple digital technologies.

Figure 1. Study design overview.

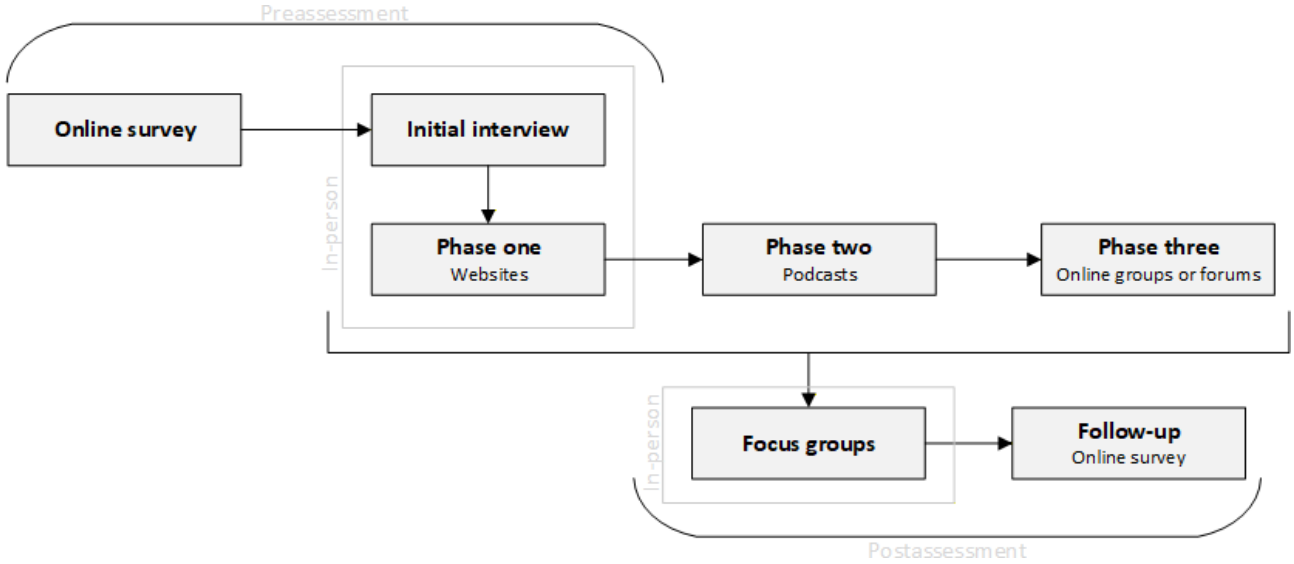

Participant Recruitment

Women who identified as going through the menopause were recruited from the North East of England through purposive sampling (led by AKO), local organizations focusing on menopause and/or women's health (eg, health centers or menopause cafes), university campus, and staff networks.

All participants took part in every stage of the research. This comprehensive participation ensured that detailed and nuanced experiences of accessing menopause information across an ecosystem of internet-enabled technologies were captured. Recruitment was ongoing in the early stages of the research and was stopped when data saturation was achieved during the preassessment interview. This saturation was achieved at each interview stage of this study.

Preassessment

Participants initially completed a short 10-minute online survey to capture demographics, menopause experience, including their perceived knowledge of menopause, and self-reported digital engagement and ability. Following this, participants took part in an in-person 15-minute semistructured interview exploring the topics from the survey in greater detail (see Multimedia Appendix 1 for survey and interview schedule).

Primary Data Collection

Participants engaged in 3 sequential phases of menopause information-seeking tasks using different internet-enabled technologies:

- Websites: 30 minutes searching for menopause-related information on websites using a desktop computer (in-person).
- Podcasts: listening to menopause-related podcast episodes for a minimum of 2 hours over 2 weeks using any personal device (at-home).
- Online groups or forums: engaging with menopause-focused online groups or forums for at least 1 hour over 1 week using any personal device (at-home).

Participants took part in an in-person 20-minute interview after the website's task and completed an online survey after the podcasts and online groups or forums tasks, exploring participants' experiences, motivations, challenges, perceived knowledge, and likelihood of future use through primarily textual responses.

Data on the sources of information that participants engaged with were also collected. To assess engagement beyond usability, the TENS-Interface (Technology-Based Experience of Need Satisfaction–Interface) [45] was administered after

XSL•FO
RenderX

each task. This 15-item scale maps onto the 3 factors associated with self-determination theory, autonomy (agency), competence (ability and effectiveness), and relatedness (connection), to capture mastery, need satisfaction, and immersion (see Multimedia Appendix 1 for task sheets, interview schedule, and surveys).

Postassessment

Participants attended 1 of 4 focus groups in-person (approximately 60 minutes) to reflect on their experiences across all technologies, considering accessibility, motivations, and complementary use. Further, 6 weeks later, participants completed a follow-up online survey to assess the lasting impact of engaging in these information-seeking tasks (see Multimedia Appendix 1 for focus group schedule and survey).

Ethical Considerations

All participants were provided with an information sheet and asked to read it before completing a consent form, giving their full informed consent to take part. Following completion of this study, a debrief sheet was provided reiterating what this study was about, participants' right to withdraw, and contact details for support if needed. All data from this study were anonymized. Participants received £60 (US $81.24) on completion of this whole study. This study was approved through Northumbria University's ethical approval system (REF: 5289).

Analysis

Audio files from interviews and focus groups were transcribed and collated with the qualitative survey data. Qualitative data were analyzed manually using visual and tactile techniques (eg, printed transcripts, annotation, Post-it [3M] notes, and paper-based thematic mapping) to support an iterative and collaborative analytic process. Participants' quotes were physically extracted from transcripts and organized alongside codes on large-format thematic maps to aid the development and refinement of themes. Qualitative data were analyzed using reflexive thematic analysis [46,47] following a deductive, theory-informed approach grounded in feminist new materialism [41]. Data were analyzed for each technology individually before synthesizing cross-cutting themes with a focus on how participants described the affordances and affective factors of different internet-enabled technologies and how they impact the agential capacities of each of these. The 6-phase process outlined by the guided analysis of Braun and Clarke [46,47]: familiarization with the data, generation of initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the report. In the final phase, 3 cross-cutting themes were identified that synthesized key patterns across all technologies and were guided by the theoretical lens. Coding was conducted collaboratively by AKO, ES, and CC, with iterative refinement and team review.

All quantitative data were entered into SPSS (IBM Corp) and analyzed. Descriptive statistics were collated to provide an overall understanding of participants' digital skills, technology engagement, and TENS-Interface scores. ANOVAs were conducted on the TENS-Interface scores to assess statistical differences between the technologies.

Qualitative and quantitative data were collected across phases and analyzed separately, then integrated at the interpretation stage through triangulation. Quantitative findings informed cross-technology comparisons, while qualitative data provided contextual depth and explanatory insight into these patterns.

Results

Overview

The findings begin with an overview of participants' demographics, then their experiences engaging with technology for menopause information, including their self-reported scores on competency, autonomy, and relatedness for each internet-enabled technology. The findings are then presented by internet-enabled technology—websites, podcasts, and online groups or forums—with a focus on the reported affordances, affective factors, and agential capacities of each. The Discussion section explores these technologies as part of a wider digital ecosystem, examining key similarities, differences, and emergent cross-cutting themes.

Participant Demographics

A total of 16 women, aged between 40 and 62 (mean 49.9, SD 6.30) years, took part in every stage of this study. Most participants reported being at the perimenopause stage (n=10) and were currently taking hormone replacement therapy (HRT; n=9). For the participants, perimenopause began between the ages of 37 and 51 (mean 44.4, SD 4.40) years.

Engaging With Internet-Enabled Technology

The initial interview and online survey explored participants' experience with internet-enabled technologies for menopause information. All participants had experience of accessing menopause information online, describing their digital engagement as between "learning the ropes" and being "confident" in their abilities; all self-reported basic digital skills.

Participants reported which internet-enabled technologies they had used previously to access digital information on menopause and other health issues. Websites were the most commonly used to access menopause and other health information, followed by social media. More participants used podcasts, mobile apps, and social media for menopause information than for other health issues, whereas more had experience with websites for other health issues.

Beyond access to digital menopause information, 81% (n=13) of the participants had seen a health care professional at some point for menopause. The sources of menopause information accessed by participants in the information-seeking tasks were recorded throughout this study. The most commonly accessed resources were the UK NHS for websites, podcasts hosted by health care professionals, and charity-based online forums.

After each information-seeking task, participants completed the TENS-Interface to explore their motivations to engage with the different technologies. This resulted in a score for competence, autonomy, and relatedness for websites, podcasts, and online groups (Figure 2). ANOVAs were then carried out with scores from all 16 participants.

XSL•FO
RenderX

Figure 2. Competence, autonomy, and relatedness scores for websites, podcasts, and online groups, based on the TENS-Interface (n=16). TENS-Interface: Technology-Based Experience of Need Satisfaction–Interface.

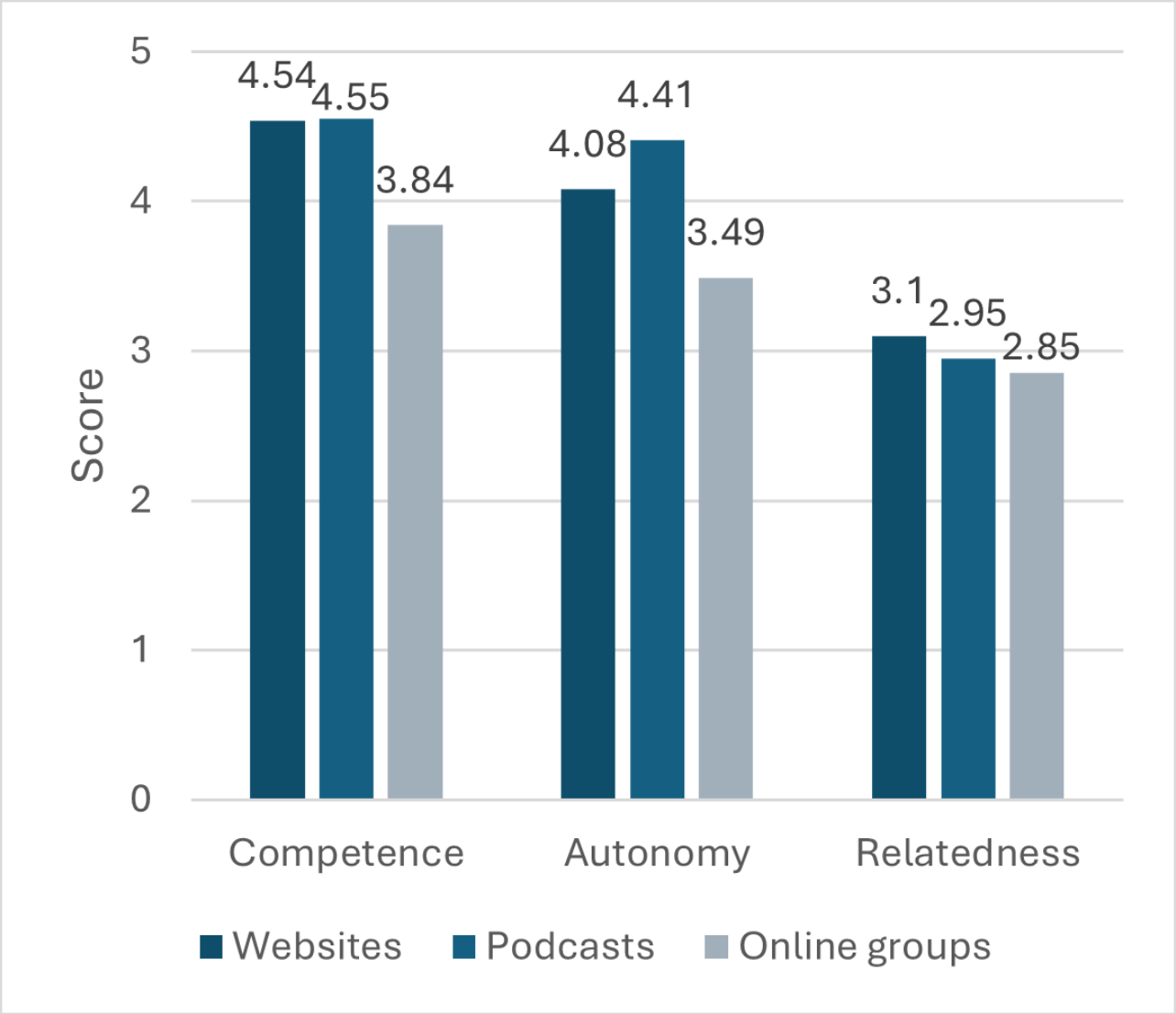

There was a statistically significant difference between groups for competence, as determined by 1-way ANOVA ($F_{2,45}$=6.883, P=.002). The effect size, eta squared (η^2), was 0.234, indicating a large effect size. A Tukey post hoc test revealed competence scores were statistically higher for the websites (mean 4.54, SD 0.494, P=.007) and the podcasts (mean 4.55, SD 0.514, P=.006) compared to the online groups (mean 3.84, SD 0.807). There was no statistically significant difference between the scores for the websites and podcasts.

There was a statistically significant difference between groups for autonomy as determined by 1-way ANOVA ($F_{2,45}$=7.816, P=.001). The effect size, eta squared (η^2), was 0.258, indicating a large effect size. A Tukey post hoc test revealed autonomy scores were statistically higher for the websites (mean 4.08, SD 0.708, P=.04) and the podcasts (mean 4.41, SD 0.376, P<.001) compared to the online groups (mean 3.49, SD 0.839). There was no statistically significant difference between the scores for the websites and podcasts.

There were no statistically significant differences between groups for relatedness as determined by 1-way ANOVA ($F_{2,45}$=0.461, P=.63). The effect size, eta squared (η^2), was 0.020, indicating a small effect size.

Menopause-Related Websites

In this study, menopause-related websites were defined as any website containing menopause information, ranging from single web pages to whole websites dedicated to menopause. These websites were very familiar to participants and were often the first resource used to find any health information, particularly for menopause. Participants described extensive experience using the internet, often daily, as part of their job or as a habit.

> *It's something I'm used to, it's territory I'm familiar with, using the internet, browsing, you know, computer etc, so that was really familiar for me.* [P14]
>
> *I think it's habit, and my habit has always been to just Google.* [P13]

All participants used Google (Google LLC) as their search engine and found websites very accessible, easy to use, and had a lot of information on them.

XSL•FO
RenderX

> *The internet, you're straight in, you're right, you just go straight to what you're looking for, you find it.* [P05]

Despite the participants' ease of access, they emphasized the importance of approaching searches with a plan. Having a specific question was necessary to focus, manage the volume of information, and avoid becoming overwhelmed.

> *Google is very clever. It is very, very clever, but sometimes it's just what you put in. You've got to be very specific.* [P07]

Descriptions of the risks of getting lost in "a rabbit hole" (P05 and P07) were common, highlighting the negative consequences of easily accessible digital information with heavily hyperlinked website designs.

Regarding design, participants discussed how different websites are from each other and that how they engage with them depends on the "success" of this design. Poorly designed websites often result in a negative experience. A couple of participants specifically explored the incongruence of the use of color to convey serious information, the connotations of certain colors, and issues around language:

> *If you go onto a website that looks – that purports to be – talking about something serious, "Menopause, the truth," but you go on and it's all pink pastels, and flowers and dream catchers, instantly you're going to go, "That's a load of bollocks," and move away from it, for me.* [P14]

> *I find some of those websites have quite patronising language sometimes, which might make me sort of go hmm. It's usually about weight gain.* [P02]

Using a search engine such as Google left participants needing to make further decisions regarding which websites to access (tasks were carried out before the introduction of Google artificial intelligence in searches). Affective factors such as the familiarity of website names and the perceived trustworthiness of these, such as the NHS, were key in deciding which sites to visit.

> *Everybody has heard of the NHS, and it's really well trusted it just kind of made me want to visit first just to kind of see what they said first. Just for reputation really.* [P09]

An "expert" or professional feel to the website also contributed to the perceived credibility, particularly in the case of the NHS. Trust extended to any websites that were linked from the NHS, as they were seen as "recommended" sources:

> *If it is linked to the NHS website, you think, "oh, well, that must be a good resource to go to if they're…" It's almost like a recommendation isn't it, that it's there.* [P04]

Participants valued a balance between professional expertise and personal experiences. The voice of the source was important, and P01 specified that it "always feels better when it has got a bit of a personal thing" on the websites. There was a dislike of commercialization on websites, advertisements, particularly linked to pharmaceutical companies or any organization that would make a profit, which led to distrust of the information available:

> *I would steer clear of anybody who was asking me, I wouldn't buy any product from a website… I just wouldn't trust that at all.* [P12]

> *The top five or six [results] were ads for specific medications, it's just like selling snake oil. It's just like, "you know, this'll change your life. You don't need HRT," so I avoided those kinds of things.* [P02]

Most importantly, relevance was central to engagement. Participants sought information that best represented themselves, particularly UK-based content or the way in which menopause and women were portrayed. They expressed distrust of sources that were not aligned with UK health care or presented narrow representations of menopause.

> *I normally put UK on the end because you often get a load of American stuff, which isn't particularly relevant.* [P04]

> *We're underrepresented, basically, because people think the menopause happens to people with long hair, husbands and children at home.* [P12]

Engaging with menopause-related websites increased participants' knowledge and confidence. Many reported that the information reinforced what they already knew about menopause; for others, they were able to validate their symptoms and experiences. This upskill in knowledge meant that participants felt that they were more prepared for what was to come, for those in the earlier stages of menopause, and could share their knowledge with others.

> *Just thought "well I'll read what they have to say about the things I possible already know." They didn't tell me anything new, in lots of ways, but they reinforced that the knowledge that I already have, basically, is okay.* [P08]

> *I was right to look, not before this session but over the last year… It's good to have the foreknowledge, it's not scary.* [P05]

Accessing menopause information online also fostered a sense of shared experience and reduced feelings of isolation. This was evidenced by P02, who was unaware of the extent to which others shared their experience before engaging with technology:

> *It's been invaluable, a really good resource to look at what other people are saying about the menopause and realising you are not alone. Menopause with younger women is not common but it's not wild, it's not crazy.* [P02]

Menopause Podcasts

Menopause podcasts were defined as any podcast episode that focused on menopause. Almost all participants had little experience with podcasts and were new to the menopause-related content in this format. Engaging with podcasts was overwhelmingly positive, with participants often expressing surprise at how informative podcasts were:

XSL•FO
RenderX

> *There is a lot that hasn't crossed the mind regarding menopause related information and podcasts seem to be a great way to gain that, especially regarding current issues.* [P13]

> *I found the episodes I listened very engaging. I have not really considered listening podcasts on the topics previously and was positively supposed how informative the podcasts were.* [P16]

Participants found podcasts to be very accessible, particularly the ability to listen while completing other activities such as household chores or driving. This also afforded participants the choice to be active or passive in engaging with the podcasts.

> *Podcasts are the easiest to fit into your life because you can do it as you're either driving or in the house pottering, so it's an easy thing to fit in.* [P03]

> *Probably one of the things about podcasts as well is that you're kind of allowing people's expertise to filter the internet for you. And you can just sit back and be educated passively, you don't have to do any of that searching.* [P01]

However, some felt less engaged with audio information, often when the content was complex or required focused attention.

> *There were a few that were too complicated to listen to whilst walking – felt like I needed a paper and pen and a few Google searches to follow the narrative.* [P01]

For others, podcasts provided a sense of intimacy and personal connection when using headphones:

> *I think the intimacy of a voice, if you're wearing headphone in your ear or something there's the confessional kind of element brought in and it does, yeah, it feels really intimate doesn't it as opposed to just reading a story on the page.* [P12]

Searching for specific menopause information within podcast platforms, whether through Spotify (Spotify Service), Apple Podcasts (Apple Inc), or a podcast-specific website, was challenging with search functions that were not intuitive. This difficulty sometimes led to unplanned engagement with broader or unexpected menopause-related topics.

> *I was a bit overwhelmed by the lack of sorting or whatever algorithm Google does for me to make things easier to search for on Spotify and BBC. That was the main challenge.* [P01]

> *Some of the content I would not have initially looked to listen to, however this gave me a much wider understanding of the menopause across many different areas.* [P07]

Regardless of the information participants accessed through the podcasts, decisions to continue listening were influenced primarily by the tone, style, and speaker credibility. A professional approach was favored with some personal experience or personal stories to "break up" the podcast and keep the format "light" and "conversational." Medical expertise was considered essential for trust, even when personal narratives were present.

> *What I preferred, I loved the funny remarks within some of the episodes, they made me smile, whereas soe seemed to concentrate on the negative aspects of the menopause.* [P15]

> *It's a conversation, rather than you've gone to a presentation or something like that, it's more a conversation.* [P01]

> *I though the content of them was good, liked that they had medical professionals providing information and not a self-appointed expert… I just wanted someone who definitely knew the science behind the process properly.* [P06]

As with websites, participants favored podcasts that covered topics that were relevant to the participant, especially those with a national or local application. There was a lack of interest or engagement with US-based content, often due to commercial undertones and the differences in health care between the 2 countries.

> *I was less engaged with the US podcast so probably would avoid that particular one in the future.* [P04]

Participants reported they were able to digest tangible information and share this with friends. Several affective factors were linked with the usage of the podcasts, specifically not feeling alone. This was most prominent in reflections centered on the practice of sharing peer experiences of menopause, which helped to normalize menopause.

> *Women are not alone and what we are going through is normal.* [P09]

> *Until I listened to that podcast, I was generally worried. I thought, "my God, what's happening?"* [P10]

Overall, the menopause podcasts left participants feeling "less alone and more empowered to talk about the menopause with other people" (P01) and to be proactive in their own health, self-monitoring potential changes.

> *I thought I was well informed and surprise I am not, so they have made me feel more empowered to do something about my menopause.* [P06]

Online Groups and Forums for Menopause

Participants encountered several online groups and forums dedicated to menopause. Experiences varied depending on their familiarity with platform design. P14, a Facebook (Meta) user, found the groups on this platform to be "easy to engage." However, the most common response to accessibility and navigation of online groups or forums was more negative. There was a lot of content available, and navigation was frequently described as difficult and overwhelming, leading to difficulties in processing the information and going "down menopause rabbit holes" (P05) for extended periods of time.

> *Just being overwhelmed by the sheer tonne of information and stream of people's personal horror stories that flood your brain.* [P12]

XSL•FO
RenderX

> *There was a wealth of info to read and digest and I spent so much time on that I didn't actually progress to anything more involved.* [P04]

The volume of information available often led to extended time spent reading, with participants reporting difficulty identifying relevant or trustworthy information. Many felt that engagement was time-consuming and offered limited educational benefit. P02 reflected on "loads of conflicting information" that was "difficult to sift through."

> *It was awful, such a nightmare to try and negotiate. Like I said, I am not savvy on them, so I may not have been used to the way to find things… it's so time consuming.* [P15]

> *I found it took a while to sift through information and do an awful lot of reading which is sometimes challenging after being at work on a computer all day.* [P02]

Online groups or forums afforded the participating women with numerous ways to interact with posts, whether this be through liking, sharing, or commenting. This is different compared to general websites or even podcasts. Many participants could see a benefit of this approach for certain individuals, but did not feel comfortable engaging or interacting in this way themselves. This was often due to the public nature of the forums, reporting they "didn't feel "brave" enough to ask or answer any questions" (P05). Overall, the group conveyed a preference to only read the posts.

> *I am very wary of the risks of engaging online… I have never commented online, I am a "lurker" – I read only. I feel reading is enough for me, I don't need more engagement as yet.* [P12]

Primarily, the posts that participants encountered were of women sharing their experiences of menopause, whether this be to ask questions, give tips, or to "vent." The participants valued these stories; they were "comforted by the stories in the community" (P12), they seemed "real and the answers heartfelt" (P05). There was a focus on "the humanity of it, the stories" (P12). Reading about peer experiences of menopause provided a measure of connection, which ultimately made participants feel as though they were not alone in their own menopause journey and that there was a supportive community.

This was particularly evident for participant P12, who reflected on how online groups or forums could be "good for calming you down when you feel you are the only one suffering from a particular symptom," as it can be "very reassuring to read others' experiences and it usually introduces you to some new medication/approach which you can follow up." Participants did not always feel that the online groups or forums worked for them, but reflected a lot on the benefits for others. It was important to them that any information gained through this mode was considered in the context of information gathered through other sources. The very personal nature of the content on the posts sometimes contradicted their own experience or knowledge of menopause.

> *In the groups, people are putting their individual experience on, so maybe contradicts with what you thought you had learned.* [P01]

Concerns were raised about the credibility of information shared in forums, particularly due to the absence of identifiable medical expertise and the anonymity of contributors.

> *You don't know how you're talking to as well. You trust the podcast because you know it's Doctor Newson, or whoever's online, you could be talking to Bill who's 59.* [P10]

> *It just felt like everyone was almost looking for help and trying to find an answer, but nobody there had the answer. There wasn't a doctor there or a medical professional or an expert or someone, even your Davina's of the world.* [P14]

The large volume of information was overwhelming; P15 explained that "the challenge is getting to the source of what you want." Participants described feeling less confident in their knowledge after engaging with online groups or forums, sometimes doubting previously trusted information. Emotional responses were also notable, with many describing feelings of sadness or distress when reading others' experiences.

> *This was one thing I picked up from Mumsnet was there was so much stuff on there that you think, "well perhaps I don't know as much as I thought I knew."* [P04]

> *There was almost quite despairing, quite sad feel to them if I put it in those terms, whereas the rest of the information I felt in control of.* [P14]

Participants voiced a lack of control regarding seeing this information, and some felt it was important to use strategies to deal with encountering negative posts and posts that were not relevant. They also expressed wanting to stop the experience from becoming overwhelming for them.

Discussion

Principal Results

This study explored how women engage with different internet-enabled technologies to access and consume digital menopause information. Unlike previous research that has focused on single technologies in isolation, this study adopted a holistic approach, examining websites, podcasts, and online groups or forums as part of a wider digital ecosystem. With the exception of the study by Lupton and Maslen [41], few studies have taken this comparative approach. Drawing on this paper, feminist new materialism and self-determination theory, this study examined how the affordances, affective factors, and agential capacities of different technologies shaped engagement, alongside motivations and challenges relating to autonomy, competence, and relatedness.

A total of 16 women experiencing menopause participated across all phases of this study, providing rich comparative insights. The exploratory design allowed for depth and detail. Each internet-enabled technology offered distinct affordances, affective factors, and agential capacities, shaped by participants'

XSL•FO
RenderX

familiarity, digital skills, and preferences for accessing and consuming health information. Websites and podcasts were associated with higher competence and autonomy scores than online groups or forums, suggesting these technologies may better support self-directed and confident information-seeking. Relatedness scores did not differ significantly across technologies, likely reflecting the general nature of this study's tasks rather than a focus on support-seeking.

Overall, the findings highlight how different internet-enabled technologies shape menopause information-seeking in nuanced ways and point to important implications for the design of inclusive, trustworthy digital health resources. Discussed below, this work begins to develop our understanding of the interplay between different internet-enabled technologies, the format of media content, and the consumer's personal and contextual preferences.

Familiar, Accessible, and Navigable Technology: Overcoming Information Overload

Across internet-enabled technologies, participants discussed the overwhelming nature of menopause information, frequently referencing going "down a rabbit hole." This is unsurprising as internet users are often presented with a large amount of information, which is often irrelevant [48]. However, in terms of menopause information, this contradicts existing research suggesting there is a paucity of readily available, relevant, reliable, and accessible resources of menopause information [27].

Issues arise when searchers are unable to effectively filter and evaluate content, leading to information overload that can negatively affect engagement and information processing [49]. This is particularly prominent with health-related media, and research has acknowledged that this has become a widespread problem in patients and physicians [49,50]. Zhong et al [51] further demonstrated the effect of information overload on mHealth users' perceived symptom severity, susceptibility, treatment benefits, barriers, self-efficacy, and action cues.

In this study, participants appeared better able to manage overload when using websites and podcasts. Familiarity with search engines and prior digital experience enabled participants to navigate information effectively, illustrating the role of digital health literacy in mitigating overload. Websites, in particular, were the most familiar due to the frequency of use in their everyday lives and were perceived as accessible, contributing to higher autonomy and competence scores. This aligns with evidence that the internet is often a first source for health information [20,52] and can support empowerment and informed decision-making [53]. Specific to menopause information, research has indicated that women have the greatest experience with health websites and the lowest experience and awareness of digital health apps [29], but that technology as a tool to access menopause information is accessible, convenient, and easy to use [24].

The way in which a technology is seen as accessible can be different for each person. In comparison to the familiarity participants had with websites, podcasts were the most novel technology. Only 21% of participants had any experience of using podcasts for menopause information, mirroring findings in the study by Harper et al [54], where 20% had used podcasts. Podcasts afforded participants the option to engage with menopause content actively or passively, while carrying out other tasks such as driving or housework. This was a key motivation for engaging with podcasts, as well as the convenience and accessibility. Few research papers have considered the role of podcasts in menopause information, but these findings support qualitative work by Shaw et al [26] and broader literature highlighting the portability as a strength of audio media (eg, the study by Weiner [55]).

In contrast, navigability was more problematic for podcasts and especially online groups or forums. Participants valued Google's ability to filter and prioritize relevant information—an affordance not replicated in podcast platforms or forums. While competence scores for podcasts remained high, online groups or forums elicited significantly lower autonomy and competence scores. Information overload in these spaces reduced participants' sense of control, trust, and perceived knowledge, consistent with research showing that overload can lead to disengagement and negative attitudes toward digital technologies [56]. This was evidenced in participants' reflections on future engagement with this technology, their lack of trust in the authenticity of the content, and their feeling less knowledgeable about menopause. In this case, the information overload compromised participants' perceived control over searching for relevant and reliable content more than improving it [57] and was reflected in the autonomy scores. In comparison to websites and podcasts, online groups or forums reported significantly lower scores on autonomy and competence, and reported feeling less knowledgeable about menopause after engaging with the platform.

Assessing the Trustworthiness of Menopause Information Sources

Motivations for engaging with the different internet-enabled technologies were shaped not only by functional affordances but also by affective factors, particularly perceived trustworthiness. This trust extended beyond the information source to the technology itself as a mediator of content. Participants favored familiar and recognized sources, particularly evidence-based, medical providers. For websites, trust was strongly associated with reputable organizations such as the UK NHS website. Participants relied on established visual and institutional cues to assess credibility, and issues arose when this was incongruent. For example, the use of certain colors, such as pink, to convey serious health information, and the connotations of these colors. As noted in the literature review by Osborne and Sillence [24], menopause information is most trusted when sourced from medical or expert organizations [28,33]. This reliance upon major sources of health content, such as hospitals, national health services, and health care professional websites for trustworthy information, is not specific to menopause and has been evidenced in wider women's health research [41,58]. Importantly, participants appeared to trust websites as a technology before evaluating individual sources, suggesting a preexisting confidence in this technology.

XSL•FO
RenderX

Emphasis on trust was slightly different for podcasts. While familiarity with presenters (eg, celebrities or influencers) remained important, participants also considered topic relevance and popularity. The podcasts facilitated a sense of intimacy, fostering a connection with the presenters and stories, a characteristic feature extensively documented in podcast research [59-62]. Similarly, Shaw et al [26] noted that relatability and storytelling were central to engagement with menopause podcasts. Podcast listeners may also form stronger parasocial relationships with the presenters when they share interests and personal experiences [63].

The development of this trust in podcasts was a direct contrast to the difficulties participants experienced with establishing trust in online groups or forums. Concerns centered on anonymity, authenticity, and lack of visible expertise. Considering the engagement framework from Sillence et al [64], these findings suggest that without baseline trust in the platform, engagement and positive outcomes are compromised. It is not just the media content source that participants did not find credible, but the affordances of the internet-enabled technology itself—the poor navigation and prior reputation of online groups or forums.

Despite the qualitative data in this study emphasizing the value of lived experiences, relatedness scores across the 3 technologies were low. This may reflect the general nature of the tasks, which focused on information-seeking rather than explicit support. Participants' varied interests and motivations likely contributed to this pattern.

However, regardless of which internet-enabled technology, participants expressed strong mistrust of commercialized content. This has been a consistent finding across the literature for digital menopause information, where any advice and information linked to pharmaceutical companies or for-profit organizations was mistrusted [26,28,33].

Access to Lived Experience Menopause Stories

Lived experience narratives played an important role in shaping trust and engagement, particularly within podcasts and online groups or forums. Podcasts were favored for the "humanness" of hearing women's stories, making them easier to connect to, supporting previous work highlighting the role of narratives in menopause communication [26,28].

Engaging with lived experiences enhanced agential capacities by reducing feelings of isolation and increasing empowerment to discuss menopause openly. To some extent, these stories became a source of validation, allowing participants to reframe their own menopause experiences. This was less evident for websites, where information-seeking focused primarily on trusted, medical content and sources (eg, NHS). These findings align with those of Osborne and Sillence [24], who note that motivations for engaging with digital menopause resources often include a desire to validate and legitimize menopause experiences through diverse stories [22,25,26,31]. Peer experiences not only promote normalization but also frame menopause in a relatable way, fostering a sense of community and reassurance that others are navigating similar challenges [22,25].

Despite this, participants in this study predominantly reported negative experiences with online groups or forums. These internet-enabled platforms elicited several negative affective factors, with participants reporting dissatisfaction with the format, emotional overload, and distrust in the credibility of content. This contrasts with existing menopause research suggesting online peer support can be supportive and reduce feelings of loneliness [22,27]. A prior literature review by Eysenbach et al [65] might provide some insight into these mixed results, suggesting the benefits of online communities depend on an individual's intrinsic motivation to engage socially. Participants in the current study often acknowledged the potential benefits of online groups or forums for others but did not feel such spaces suited their own needs.

Recent research by Dallolio et al [66] further supports this nuance, illustrating that while online groups or forums can provide valuable support and facilitate emotional exchanges, particularly for women who lack proper health care and support during menopause, they may not be universally beneficial. Importantly, podcasts were perceived as more reliable sources of lived experience because narratives were typically curated and facilitated by experts, enhancing their credibility and coherence. This professional yet personal approach aligns with earlier work on online health communication, where trust and relevance are key factors in determining the success of health-related media content [67]. Sillence et al [68] found that while people value the authenticity of peer experiences, they often question their reliability [68,69]. This suggests that professional oversight or facilitation might mitigate concerns over the credibility of shared information.

The tone and ethos of an online community were also critical. In this study, participants described some of these groups as platforms primarily used for venting about menopause experiences. While some participants found therapeutic value in these exchanges, others felt they were overwhelming or unhelpful. The value of these communities depends on individual preferences—some may seek emotional support and validation, while others may prefer more structured or evidence-based discussions. Armstrong et al [70] emphasize the importance of community identity and tone in online health spaces. When the tone of a community does not align with an individual's needs—such as focusing on venting or specific treatments (eg, HRT)—it can detract from the perceived usefulness of the stories shared, leading to disengagement. This misalignment may explain the negative affective factors reported by some participants in this study. Consequently, the efficacy of online peer communities in menopause health communication may be influenced by how well the platform matches the users' expectations and emotional needs.

Ultimately, the evidence suggests that the format and credibility of menopause-related digital content, particularly when it includes both professional guidance and lived experience narratives, play significant roles in shaping engagement and trust.

Limitations

All participants self-reported good digital literacy skills and familiarity with digital technology for health

XSL•FO
RenderX

information-seeking. As a result, findings may not reflect the experiences of individuals facing challenges with usability, navigation, or accessibility. Future research should explore these barriers, considering factors such as technical access, autonomy of use, social support networks, and experience [71]. Building on the in-depth insights provided here, there is clear value in scaling this work to a larger survey. A broader, quantitative study could examine patterns of access and engagement across websites, podcasts, and online forums within a more diverse population, complementing the nuanced findings presented here and offering a more comprehensive understanding of digital health behaviors in the menopause context. Additionally, not enough detail is understood regarding the potential role of online groups or forums. It is possible that the conflicting information in this study is a result of the nature of the tasks and the focus on upskilling participants' menopause knowledge rather than a more specific purpose. Research with a greater emphasis on more organic information searching methods may help to mitigate this limitation.

Conclusions

Different internet-enabled technologies play a distinct role in accessing menopause information. Websites remain the most familiar and trusted internet-enabled technology, supporting effective evaluation of sources and navigation of information. Podcasts, though novel for many participants, were highly valued for their accessibility, intimacy, and facilitation of content by professionals. Trusting in the source of menopause content was key in the motivations to engage with a technology and use menopause information, presenting challenges for online groups or forums where participants reported distrust, navigational difficulties, and emotional overload. While online forums may offer benefits for women lacking support networks, they were not universally perceived as helpful.

Across all internet-enabled technologies, lived experience stories were valued, though engagement depended on personal preferences, experience, trust, and availability of wider support networks. By examining multiple technologies together, this study provides a nuanced understanding of how format, credibility, and lived experience shape engagement. These insights inform the design of inclusive, trustworthy menopause resources and highlight opportunities to scale this work to larger studies for broader population-level insights.

Acknowledgments

The authors would like to thank all the participants for their time and support of this study. Generative artificial intelligence tools were not used at any stage in the preparation of this paper.

Funding

The authors declare financial support was received for the research, authorship, and publication of this paper. This work was funded by the Engineering and Physical Sciences Research Council (EPSRC) as part of the Centre for Digital Citizens (EP/T022582/1). The funder had no involvement in this study's design, data collection, analysis, interpretation, or the writing of this paper.

Data Availability

Data sharing is not available for this paper to protect the anonymity of participants; however, quotations from participants are included in this paper.

Authors' Contributions

Conceptualization: AKO, ES
Data curation: AKO
Formal analysis: AKO (lead), ES (equal), CC (supporting)
Funding acquisition: ACD
Investigation: AKO (lead), ES (supporting), CC (supporting)
Methodology: AKO, ES
Project administration: AKO
Resources: AKO, ES
Writing – original draft: AKO
Writing – review & editing: ES (lead), CC (supporting), ACD (supporting).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data collection materials.
[DOCX File, 41 KB - jmir_v28i1e78215_app1.docx]

XSL•FO
RenderX

References

1. Menopause. World Health Organization. 2024 Oct 16. URL: https://www.who.int/news-room/fact-sheets/detail/menopause [accessed 2026-04-30]
2. Hunter MS, Gentry-Maharaj A, Ryan A, et al. Prevalence, frequency and problem rating of hot flushes persist in older postmenopausal women: impact of age, body mass index, hysterectomy, hormone therapy use, lifestyle and mood in a cross-sectional cohort study of 10,418 British women aged 54-65. BJOG 2012 Jan;119(1):40-50. [doi: 10.1111/j.1471-0528.2011.03166.x] [Medline: 22008610]
3. Hoga L, Rodolpho J, Gonçalves B, Quirino B. Women's experience of menopause: a systematic review of qualitative evidence. JBI Database System Rev Implement Rep 2015 Sep 16;13(8):250-337. [doi: 10.11124/jbisrir-2015-1948] [Medline: 26455946]
4. Hardy C, Hunter MS, Griffiths A. Menopause and work: an overview of UK guidance. Occup Med (Chic Ill) 2018 Dec 26;68(9):580-586. [doi: 10.1093/occmed/kqy134]
5. Gold EB. The timing of the age at which natural menopause occurs. Obstet Gynecol Clin North Am 2011 Sep;38(3):425-440. [doi: 10.1016/j.ogc.2011.05.002] [Medline: 21961711]
6. O'Reilly K, McDermid F, McInnes S, Peters K. An exploration of women's knowledge and experience of perimenopause and menopause: an integrative literature review. J Clin Nurs 2023 Aug;32(15-16):4528-4540. [doi: 10.1111/jocn.16568] [Medline: 36336832]
7. Duffy O, Iversen L, Hannaford PC. The menopause "it's somewhere between a taboo and a joke". A focus group study. Climacteric 2011 Aug;14(4):497-505. [doi: 10.3109/13697137.2010.549974] [Medline: 21395452]
8. Marnocha SK, Bergstrom M, Dempsey LF. The lived experience of perimenopause and menopause. Contemp Nurse 2011 Feb;37(2):229-240. [doi: 10.5172/conu.2011.37.2.229] [Medline: 21692594]
9. Hunter MS. Menopause and midlife: psychosocial perspectives and interventions. In: Routledge International Handbook of Women's Sexual and Reproductive Health: Routledge; 2019:83-98. [doi: 10.4324/9781351035620-7]
10. Marlatt KL, Beyl RA, Redman LM. A qualitative assessment of health behaviors and experiences during menopause: a cross-sectional, observational study. Maturitas 2018 Oct;116:36-42. [doi: 10.1016/j.maturitas.2018.07.014] [Medline: 30244777]
11. Koyuncu T, Unsal A, Arslantas D. Evaluation of the effectiveness of health education on menopause symptoms and knowledge and attitude in terms of menopause. J Epidemiol Glob Health 2018 Dec;8(1-2):8-12. [doi: 10.2991/j.jegh.2018.08.103] [Medline: 30859781]
12. Bidmon S, Terlutter R. Gender differences in searching for health information on the internet and the virtual patient-physician relationship in Germany: exploratory results on how men and women differ and why. J Med Internet Res 2015 Jun 22;17(6):e156. [doi: 10.2196/jmir.4127] [Medline: 26099325]
13. Febretti A, Garzotto F. Usability, playability, and long-term engagement in computer games. CHI '09 2009 Apr 4:4063-4068. [doi: 10.1145/1520340.1520618]
14. Timmermans S. The engaged patient: the relevance of patient-physician communication for twenty-first-century health. J Health Soc Behav 2020 Sep;61(3):259-273. [doi: 10.1177/0022146520943514] [Medline: 32723112]
15. Timmermans S, Freidin B. Caretaking as articulation work: the effects of taking up responsibility for a child with asthma on labor force participation. Soc Sci Med 2007 Oct;65(7):1351-1363. [doi: 10.1016/j.socscimed.2007.05.027] [Medline: 17590253]
16. Lareau A. Unequal childhoods: class, race, and family life. In: Inequality in the 21st Century: Routledge; 2018:444-451.
17. Munn C, Vaughan L, Talaulikar V, Davies MC, Harper JC. Menopause knowledge and education in women under 40: results from an online survey. Womens Health (Lond) 2022;18:17455057221139660. [doi: 10.1177/17455057221139660] [Medline: 36533635]
18. Janz NK, Becker MH. The health belief model: a decade later. Health Educ Q 1984;11(1):1-47. [doi: 10.1177/109019818401100101] [Medline: 6392204]
19. What is digital health technology and what can it do for me? Health and social care services research. National Institute for Health and Care Research. 2022 Sep 27. URL: https://evidence.nihr.ac.uk/collection/what-is-digital-health-technology [accessed 2026-04-30]
20. Jacobs W, Amuta AO, Jeon KC. Health information seeking in the digital age: an analysis of health information seeking behavior among US adults. Cogent Soc Sci 2017 Jan 1;3(1):1302785. [doi: 10.1080/23311886.2017.1302785]
21. Wiederhold BK. Femtech: digital help for women's health care across the life span. Cyberpsychol Behav Soc Netw 2021 Nov;24(11):697-698. [doi: 10.1089/cyber.2021.29230.editorial] [Medline: 34806914]
22. Sillence E, Hardy C, Kemp E. "This app just gets me": assessing the quality, features and user reviews of menopause smartphone apps. J Consum Health Internet 2023 Apr 3;27(2):156-172. [doi: 10.1080/15398285.2023.2204287]
23. Aleksova J, Kuczynska-Burggraf M, Ranasinha S, Vincent A. Information on early menopause: is the internet the place to search? Climacteric 2017 Jun;20(3):248-255. [doi: 10.1080/13697137.2017.1301920] [Medline: 28333562]
24. Osborne AK, Sillence E. Accessing information on menopause transition and the role of digital health technologies: a narrative review. Women Health 2025 Jul;65(6):508-521. [doi: 10.1080/03630242.2025.2523258] [Medline: 40568889]

XSL•FO
RenderX

25. Edwards AL, Shaw PA, Halton CC, et al. “It just makes me feel a little less alone”: a qualitative exploration of the podcast menopause: unmuted on women’s perceptions of menopause. Menopause 2021 Sep 7;28(12):1374-1384. [doi: 10.1097/GME.0000000000001855] [Medline: 34469933]
26. Shaw PA, Sumner AL, Halton CC, et al. “You’re more engaged when you’re listening to somebody tell their story”: a qualitative exploration into the mechanisms of the podcast ‘menopause: unmuted’ for communicating health information. Patient Educ Couns 2022 Dec;105(12):3494-3500. [doi: 10.1016/j.pec.2022.09.003] [Medline: 36109275]
27. Aljumah R, Phillips S, Harper JC. An online survey of postmenopausal women to determine their attitudes and knowledge of the menopause. Post Reprod Health 2023 Jun;29(2):67-84. [doi: 10.1177/20533691231166543] [Medline: 36994487]
28. Sillence E, Briggs P, Harris PR, Fishwick L. How do patients evaluate and make use of online health information? Soc Sci Med 2007 May;64(9):1853-1862. [doi: 10.1016/j.socscimed.2007.01.012]
29. Richard-Davis G, Singer A, King DD, Mattle L. Understanding attitudes, beliefs, and behaviors surrounding menopause transition: results from three surveys. Patient Relat Outcome Meas 2022;13:273-286. [doi: 10.2147/PROM.S375144] [Medline: 36540377]
30. Backonja U, Taylor-Swanson L, Miller AD, Jung SH, Haldar S, Woods NF. “There’s a problem, now what’s the solution?”: suggestions for technologies to support the menopausal transition from individuals experiencing menopause and healthcare practitioners. J Am Med Inform Assoc 2021 Feb 15;28(2):209-221. [doi: 10.1093/jamia/ocaa178] [Medline: 33582820]
31. Genuis SK. Constructing “sense” from evolving health information: a qualitative investigation of information seeking and sense making across sources. J Am Soc Inf Sci Tec 2012 Aug;63(8):1553-1566. [doi: 10.1002/asi.22691]
32. Huston SA, Jackowski RM, Kirking DM. Women’s trust in and use of information sources in the treatment of menopausal symptoms. Womens Health Issues 2009;19(2):144-153. [doi: 10.1016/j.whi.2009.01.004] [Medline: 19272565]
33. Henwood F, Wyatt S, Hart A, Smith J. ‘Ignorance is bliss sometimes’: constraints on the emergence of the ‘informed patient’ in the changing landscapes of health information. Sociol Health Illn 2003 Sep;25(6):589-607. [doi: 10.1111/1467-9566.00360] [Medline: 12919447]
34. Krawczyk A, Lau E, Perez S, Delisle V, Amsel R, Rosberger Z. How to inform: comparing written and video education interventions to increase human papillomavirus knowledge and vaccination intentions in young adults. J Am Coll Health 2012;60(4):316-322. [doi: 10.1080/07448481.2011.615355] [Medline: 22559091]
35. Ross L, Ashford AD, Bleechington SJ, Dark T, Erwin DO. Applicability of a video intervention to increase informed decision making for prostate-specific antigen testing. J Natl Med Assoc 2010 Mar;102(3):228-236. [doi: 10.1016/s0027-9684(15)30529-0] [Medline: 20355352]
36. Armstrong AW, Idriss NZ, Kim RH. Effects of video-based, online education on behavioral and knowledge outcomes in sunscreen use: a randomized controlled trial. Patient Educ Couns 2011 May;83(2):273-277. [doi: 10.1016/j.pec.2010.04.033] [Medline: 20570081]
37. Halliday L, Boughton M. Premature menopause: exploring the experience through online communication. Nurs Health Sci 2009 Mar;11(1):17-22. [doi: 10.1111/j.1442-2018.2008.00415.x]
38. Lazar A, Su NM, Bardzell J. Parting the red sea: sociotechnical systems and lived experiences of menopause. Proc 2019 CHI Conf Hum Factors Comput Syst 2019 May 2:1-16. [doi: 10.1145/3290605.3300710]
39. Brown P, Britton V, Leclerc M. PCR119 exploring women’s needs and information seeking behaviors on online menopause UK forums: implications for health communication and support. Value Health 2024 Dec;27(12):S529. [doi: 10.1016/j.jval.2024.10.3362]
40. Stawarz K, Preist C, Coyle D. Use of smartphone apps, social media, and web-based resources to support mental health and well-being: online survey. JMIR Ment Health 2019 Jul 12;6(7):e12546. [doi: 10.2196/12546] [Medline: 31301126]
41. Lupton D, Maslen S. How women use digital technologies for health: qualitative interview and focus group study. J Med Internet Res 2019 Jan 25;21(1):e11481. [doi: 10.2196/11481] [Medline: 30681963]
42. Ryan RM, Deci EL. Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness: Guilford Publications; 2018. [doi: 10.1521/978.14625/28806]
43. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000 Jan;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
44. Ryan RM, Deci EL. Self-determination theory. In: Maggino F, editor. Encyclopedia of Quality of Life and Well-Being Research: Springer, Cham; 2024:6229-6235. [doi: 10.1007/978-3-031-17299-1_2630]
45. Peters D, Calvo RA, Ryan RM. Designing for motivation, engagement and wellbeing in digital experience. Front Psychol 2018;9(797):797. [doi: 10.3389/fpsyg.2018.00797] [Medline: 29892246]
46. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
47. Braun V, Clarke V. Thematic Analysis: A Practical Guide, 1st edition: SAGE; 2021. [doi: 10.53841/bpsqmip.2022.1.33.46]
48. Jiang S, Beaudoin CE. Health literacy and the internet: an exploratory study on the 2013 HINTS survey. Comput Human Behav 2016 May;58:240-248. [doi: 10.1016/j.chb.2016.01.007]
49. Kim K, Lustria MLA, Burke D, Kwon N. Predictors of cancer information overload: findings from a national survey. Inf Res 2007;12(4):12-14 [FREE Full text]

XSL•FO
RenderX

50. Nijor S, Rallis G, Lad N, Gokcen E. Patient safety issues from information overload in electronic medical records. J Patient Saf 2022 Sep 1;18(6):e999-e1003. [doi: 10.1097/PTS.0000000000001002] [Medline: 35985047]
51. Zhong L, Cao J, Xue F. The paradox of convenience: how information overload in mHealth apps leads to medical service overuse. Front Public Health 2024;12:1408998. [doi: 10.3389/fpubh.2024.1408998] [Medline: 39668954]
52. Wang X, Shi J, Kong H. Online health information seeking: a review and meta-analysis. Health Commun 2021 Sep;36(10):1163-1175. [doi: 10.1080/10410236.2020.1748829] [Medline: 32290679]
53. Lu Q, Schulz PJ. Physician perspectives on internet-informed patients: systematic review. J Med Internet Res 2024 Jun 6;26:e47620. [doi: 10.2196/47620] [Medline: 38842920]
54. Harper JC, Phillips S, Biswakarma R, et al. An online survey of perimenopausal women to determine their attitudes and knowledge of the menopause. Womens Health (Lond) 2022;18:17455057221106890. [doi: 10.1177/17455057221106890] [Medline: 35758176]
55. Weiner J. The voices: toward a critical theory of podcasting. Slate. 2014 Dec 14. URL: https://www.slate.com/articles/arts/ten_years_in_your_ears/2014/12/what_makes_podcasts_so_addictive_and_pleasurable.html [accessed 2026-04-30]
56. Swar B, Hameed T, Reychav I. Information overload, psychological ill-being, and behavioral intention to continue online healthcare information search. Comput Hum Behav 2017 May;70:416-425. [doi: 10.1016/j.chb.2016.12.068]
57. Edmunds A, Morris A. The problem of information overload in business organisations: a review of the literature. Int J Inf Manage 2000 Feb;20(1):17-28. [doi: 10.1016/S0268-4012(99)00051-1]
58. Sparidaens EM, Beerendonk CCM, Fleischer K, Nelen W, Braat DDM, Hermens R. Exploration of fertility and early menopause related information needs and development of online information for young breast cancer survivors. BMC Womens Health 2022 Aug 3;22(1):329. [doi: 10.1186/s12905-022-01901-z] [Medline: 35922784]
59. Perks LG, Turner JS. Podcasts and productivity: a qualitative uses and gratifications study. Mass Commun Soc 2019 Jan 2;22(1):96-116. [doi: 10.1080/15205436.2018.1490434]
60. Heshmat Y, Yang L, Neustaedter C, editors. Quality “alone” time through conversations and storytelling: podcast listening behaviors and routines. Graphics Interface (GI) ’18 Proc 2018:76-83. [doi: 10.20380/GI2018.11]
61. Chan-Olmsted S, Wang R. Understanding podcast users: consumption motives and behaviors. New Media Soc 2022 Mar;24(3):684-704. [doi: 10.1177/1461444820963776]
62. Tobin SJ, Guadagno RE. Why people listen: motivations and outcomes of podcast listening. PLoS ONE 2022;17(4):e0265806. [doi: 10.1371/journal.pone.0265806] [Medline: 35385493]
63. Schlütz D, Hedder I. Aural parasocial relations: host-listener relationships in podcasts. J Radio Audio Media 2022 Jul 3;29(2):457-474. [doi: 10.1080/19376529.2020.1870467]
64. Sillence E, Hardy C, Harris PR, Briggs P. Modeling patient engagement in peer-to-peer healthcare. WWW ’14 2014 Apr 7:481-486. [doi: 10.1145/2567948.2576937]
65. Eysenbach G, Powell J, Englesakis M, Rizo C, Stern A. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. BMJ 2004 May 15;328(7449):1166. [doi: 10.1136/bmj.328.7449.1166] [Medline: 15142921]
66. Dallolio AS, Abdalla CC, Ferraz SB. Bytes of affect: unveiling the emergence of the responsible menopause consumer on social media. J Mark Manage 2025 Feb 12;41(3-4):312-342. [doi: 10.1080/0267257X.2025.2463496]
67. Sillence E, Briggs P. Examining the role of the internet in health behaviour. In: Joinson A, McKenna KYA, Postmes T, Reips UD, editors. Oxford Handbook of Internet Psychology: Oxford University Press; 2012:347-360. [doi: 10.1093/oxfordhb/9780199561803.013.0022]
68. Sillence E, Hardy C, Briggs P. Why don’t we trust health websites that help us help each other?: An analysis of online peer-to-peer healthcare. WebSci ’13: Proc 5th Ann ACM Web Sci Conf 2013 May 2:396-404. [doi: 10.1145/2464464.2464488]
69. Fan H, Lederman R. Online health communities: how do community members build the trust required to adopt information and form close relationships? Eur J Inf Syst 2018 Jan 2;27(1):62-89. [doi: 10.1080/0960085X.2017.1390187]
70. Armstrong N, Koteyko N, Powell J. “Oh dear, should I really be saying that on here?”: Issues of identity and authority in an online diabetes community. Health (London) 2012 Jul;16(4):347-365. [doi: 10.1177/1363459311425514] [Medline: 22067915]
71. Hargittai E. The digital divide and what to do about it. In: New Economy Handbook 2003:821-839 URL: https://www.webuse.org/webuse.org/pdf/Hargittai-DigitalDivideWhatToDo2007.pdf [accessed 2026-05-14]

Abbreviations

HRT: hormone replacement therapy
mHealth: mobile health
NHS: National Health Service
TENS-Interface: Technology-Based Experience of Need Satisfaction–Interface

XSL•FO
RenderX

Edited by A Stone; submitted 28.May.2025; peer-reviewed by C Cronin, M Patrick; revised version received 19.Mar.2026; accepted 20.Mar.2026; published 20.May.2026.

Please cite as:
Osborne AK, Sillence E, Claisse C, Durrant AC
Women's Engagement With Different Internet-Enabled Technologies to Access Digital Menopause Information: Mixed Methods, Multiphase Sequential Study
J Med Internet Res 2026;28:e78215
URL: https://www.jmir.org/2026/1/e78215
doi:10.2196/78215

© Alison K Osborne, Elizabeth Sillence, Caroline Claisse, Abigail C Durrant. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Operationalizing Digital Health Equity in Artificial Intelligence–Enabled Patient Decision Aids for Older Adults: Mixed Methods Study

Cindy Yue Tian[1,2], PhD; Xiaochen Yang[1], MPhil; Kailu Wang[1,2], PhD; Annie Wai-Ling Cheung[1,2], MPhil; Jonathan Chun-Hei Ma[1,2], MPH; Canjie Lu[1,2], MA; Jasmine Cheuk-Ying Yu[3]; Crystal Ying Chan[1,2], PhD; Jiamin Chen[4], BCS; Kun Ouyang[5], PhD; Ivan Wai-Kiu Lin[6], BSW; Tim Hung-Cheong Pang[6], BSW; Shi Zhao[7], PhD; Yingwei Wang[8,9], MD, PhD; Eliza Lai-Yi Wong[1,2], RN, PhD

[1]JC School of Public Health and Primary Care, Faculty of Medicine, Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

[2]Centre for Health Systems and Policy Research, JC School of Public Health and Primary Care, Faculty of Medicine, Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

[3]Faculty of Medicine, Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

[4]Department of Computer Science and Engineering, Faculty of Engineering, Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

[5]School of Computing, National University of Singapore, Singapore, Singapore

[6]Society for Community Organization, Hong Kong, China (Hong Kong)

[7]School of Public Health, Tianjin Medical University, Tianjin, China

[8]Department of Medical Humanities, Tzu Chi University, Taiwan, Taiwan

[9]Center for Palliative Care, Hualien Tzu Chi Hospital, Taiwan, Taiwan

Corresponding Author:
Eliza Lai-Yi Wong, RN, PhD
JC School of Public Health and Primary Care
Faculty of Medicine
Chinese University of Hong Kong
Rm 418, School of Public Health Building,
Prince of Wales Hospital, Shatin, N.T.
Hong Kong, 999077
China (Hong Kong)
Phone: 852 22528772
Email: lywong@cuhk.edu.hk

Abstract

Background: Artificial intelligence–enabled patient decision aids (AI-PDAs) hold promise for supporting older adults with chronic diseases in accessing personalized health information, clarifying preferences, and engaging in shared decision-making. Achieving equity in their design requires attention to the complex health care and digital contexts in which these tools are used. While the Digital Health Equity Framework (DHEF) provides a conceptual foundation, practical strategies for its application remain limited.

Objective: This study aimed to identify equity-related determinants and generate actionable design strategies for applying the DHEF to AI-PDAs for older adults.

Methods: A mixed methods study was conducted. Semistructured interviews were conducted with older adults living with hypertension and/or diabetes, health care providers, and medical students to explore equity determinants relevant to AI-PDAs. In parallel, a review of reviews synthesized existing evidence on approaches to addressing these determinants. Interview findings and review findings were integrated through an iterative mapping process conducted by the research team and refined through multidisciplinary expert consultation involving medicine, public health, social services, and computer science.

Results: A total of 33 stakeholders were interviewed, including 15 older adults, 8 health care providers, and 10 medical students. Thirteen reviews were included in the umbrella review. The integrated synthesis identified equity determinants spanning individual, interpersonal, community, and societal levels across both the health care and digital environments, together with cross-level concerns related to algorithmic fairness. These findings informed 5 recommendations for equitable AI-PDA development: (1)

XSL•FO
RenderX

co-design with end users to address their needs, (2) embrace relationship-centered design, (3) leverage community resources to improve support, (4) promote accessible and equitable artificial intelligence (AI) governance in society, and (5) enhance equitable AI through algorithmic fairness. Together, these recommendations provide practical guidance for design, pilot testing, implementation, and evaluation.

Conclusions: By integrating stakeholder perspectives with synthesized review evidence, this study extends the DHEF from a primarily conceptual framework toward a more practice-oriented approach for AI-PDAs for older adults with chronic disease. Health care settings serve as a mediating sociotechnical context where AI tools may either support or constrain equitable care participation. The findings underscore the need for interdisciplinary collaboration to align technological innovation with equity-oriented design. Future work should focus on co-designed prototypes, real-world testing, and measurable equity outcomes.

(J Med Internet Res 2026;28:e89011) doi:10.2196/89011

KEYWORDS

artificial intelligence; co-design; digital health equity; older adult care; patient decision aids; umbrella review

Introduction

Chronic diseases, including hypertension and diabetes, are major contributors to global morbidity and mortality [1]. Effective management of these conditions requires coordinating lifestyle modifications, medication adherence, and ongoing monitoring, highlighting the need for approaches that integrate patients' values into evidence-based care. Shared decision-making (SDM) is widely promoted to achieve this goal. It is a collaborative process in which patients and clinicians jointly discuss and evaluate health care options to determine the most appropriate course of action for each patient [2].

Patient decision aids are structured, evidence-based tools that facilitate SDM in chronic illness management [3]. They outline available treatment options, present information on benefits and risks, and assist patients in evaluating these options against their personal values and preferences. Artificial intelligence (AI) now offers a paradigm to enhance the functionality and impact of traditional patient decision aids. AI-enabled patient decision aids (AI-PDAs) embed advanced computational techniques into digital platforms to synthesize individual patient data with large-scale clinical evidence. Moving beyond static information delivery, these tools offer dynamic, personalized support by tailoring explanations, generating individualized risk-benefit profiles, and strengthening clinician-patient dialogue [4-6].

Older adults may particularly benefit from AI-PDAs, given their high burden of multiple chronic diseases and challenges in processing complex health information and applying it to their own context. However, without inclusive and thoughtful design, these benefits may fail to be realized. As noted in the inverse care law [7], younger individuals are more likely to adopt digital interventions, drawing greater attention in digital health development. Consequently, the needs of older adults have received limited consideration. Even when included, older adults are often treated as a homogeneous group, with insufficient recognition of differences in digital literacy, cognitive ability, and readiness to adopt new technologies [8]. Such oversights risk reducing adoption and widening existing digital inequities in this population [9].

To address these concerns, there is a need for equitable AI-PDAs for older adults, which are intentionally designed to ensure fair access, usability, and benefit across diverse aging populations. The Digital Health Equity Framework (DHEF) [10] offers a useful conceptual structure for this purpose, guiding the identification of digital determinants of health at the individual, interpersonal, community, and societal levels. Previous discussions of this framework have mainly focused on the digital environment (eg, whether people have access to technology) and on a market environment (eg, whether the market supports digital health) [11-13]. Much less attention has been paid to the health care environment itself, especially to the factors that influence the health problems these tools target [14]. Addressing equity determinants of AI-PDAs in the health care and digital environments ensures that these tools are not only technically accessible but also effectively support equitable patient participation in health care decisions. Furthermore, while the DHEF has advanced conceptual understanding, practical approaches for operationalizing identified determinants to inform the development of digital tools remain limited [15,16]. This gap makes it challenging for designers to translate equity principles into concrete development strategies for these tools.

This study reports the predesign phase of developing an AI-PDA to support SDM among older adults with hypertension and/or diabetes in Hong Kong. We aimed to systematically identify equity-related determinants and generate actionable design strategies that can inform the AI-PDAs for older adult care. Specifically, the study pursued three objectives: (1) to identify equity determinants related to AI-PDAs in health care and digital environments through stakeholder interviews with older adult patients, health care providers (HCPs), and medical students; (2) to synthesize evidence-based strategies for addressing these determinants through an umbrella review; and (3) to integrate insights from both the interviews and the literature review into actionable recommendations that operationalize the DHEF. Although these recommendations were developed for older adults with hypertension/diabetes in Hong Kong, they may also inform the design and implementation of equitable patient-facing AI tools in similar settings and populations.

Methods

This study adopted a mixed methods design combining stakeholder interviews, an umbrella review, and expert consultations to develop contextually grounded, evidence-informed recommendations for operationalizing the DHEF for AI-PDA for older adults.

XSL•FO
RenderX

Stakeholder Interviews

From July to August 2025, we conducted semistructured interviews with 3 key stakeholder groups using purposive sampling. First, older adults aged ≥65 years living with hypertension and/or diabetes were recruited through outpatient clinics, community health centers, previous project contacts, supplementary social media advertisements, and community activities serving residents of deprived housing estates in Hong Kong. Recruitment sought variation in socioeconomic and demographic characteristics relevant to digital health equity, including sex, educational attainment, and household income. Second, health care professionals involved in chronic disease management, including early-career (<6 years of clinical experience) and senior practitioners (≥6 years), were recruited through the research team's professional network. Third, senior medical students were recruited through the university's medical school, as they may offer emerging perspectives shaped by recent exposure to digital and AI-enabled tools in medical training. Recruitment and preliminary analysis proceeded concurrently. Saturation was assessed separately within each stakeholder group through iterative review of interview notes, coding, and team discussion. Interviews were concluded when later interviews generated few or no new codes and yielded largely repetitive accounts that did not substantially alter the interpretation of emerging themes.

To distinguish determinants in health care vs digital environments, the interview guide comprised 2 sections (Multimedia Appendix 1). The first examined participants' understanding and experiences of SDM. As AI-PDAs are intended to facilitate SDM, we treated the factors that shape equitable participation in SDM as relevant determinants of equitable engagement with AI-PDAs in the health care environment. The second explored views on AI-PDAs, including expected benefits and concerns about using these tools in the digital environment. All interviews were conducted in person in Cantonese, audio-recorded, and transcribed using OneDrive transcription software. The transcripts were then manually reviewed for accuracy by 3 researchers (JCHM, CL, and AWLC). Audio recordings did not include participants' names and other potentially identifying information. Electronic files were stored on password-protected systems accessible only to the research team.

Thematic analysis was conducted using a hybrid deductive-inductive approach. The key terms under each level from the DHEF served as the main coding structure, with equity determinants mapped onto its 4 levels (individual, interpersonal, community, and societal) within 2 analytic domains: the health care environment and the digital environment. Where data revealed determinants not anticipated by the DHEF's original formulation, additional codes were generated inductively and subsequently mapped onto the most appropriate DHEF level. Transcript coding was conducted independently by 2 researchers (CYT and XY), with coding differences discussed and resolved through consensus. Interviewers and coders hold public health backgrounds and had no prior personal relationship with participants before recruitment. We acknowledge that this disciplinary background may have oriented interpretation toward structural and system-level determinants. To mitigate this, codes and emerging themes were discussed within the research team to consider alternative interpretations and ensure balanced attention to both individual and structural factors.

Umbrella Review

This targeted umbrella review was conducted to complement the qualitative interviews by synthesizing evidence-based strategies for addressing the equity determinants identified through the study. The literature search was conducted to identify reviews published from January 2022 to August 2025 in PubMed, MEDLINE, and ACM Digital Library. The search focused on 2 domains: AI-PDAs for chronic conditions (hypertension/diabetes) and barriers/solutions to digital health equity. For screening purposes, AI-enabled tools were defined as digital health tools incorporating an AI-related component, such as machine learning, natural language processing, predictive modeling, or adaptive decision logic, rather than conventional static or nonadaptive digital tools. Digital health equity was operationalized as issues related to equitable access, use, usability, engagement, or benefit from digital health tools, including barriers, facilitators, disparities, or strategies to support underserved populations. Studies were eligible if digital health equity was addressed as either a primary focus or a clearly identifiable secondary focus. Search terms combined controlled vocabulary and free-text keywords related to chronic disease management/hypertension/diabetes, digital health equity/health equity, and review using database-specific Boolean operators and syntax. Search algorithms are detailed in Multimedia Appendix 2. The search was supplemented with a general internet search via Google Scholar, a focused search on key authors, and citation and reference tracking.

Studies were included if they met the following criteria: (1) review articles using a transparent methodology (eg, systematic review, scoping review, meta-analysis, realist review, integrative review, or narrative review); (2) discussed AI-enabled digital tools for chronic disease management; (3) discussed digital health equity in the context of chronic disease management, including hypertension or diabetes; (4) published in English; and (5) available in full text. All records were imported into Rayyan (Rayyan Systems Inc) for screening. Two researchers (CYT and XY) independently screened titles and abstracts and then assessed potentially eligible articles at the full-text stage. Disagreements were resolved through discussion and consensus. The included studies were then independently reviewed and summarized by 2 researchers (CYT and XY). We did not conduct a formal methodological quality appraisal of the included reviews, as the purpose of this targeted umbrella review was to support thematic synthesis and framework operationalization rather than to assess intervention effectiveness or the certainty of the evidence. Similar to the interview process, thematic analysis was conducted, guided by the DHEF, to identify suggested strategies at each framework level. Where strategies were found to operate simultaneously across multiple levels rather than primarily at a single level, these were identified as cross-level interactions.

Data Analysis and Recommendation Generation

We identified determinants of AI-PDA equity from interviews and strategies from an umbrella review, both organized by

XSL•FO
RenderX

DHEF levels (health care and digital environments). Similar to previous stakeholder-informed, evidence-integrated studies [17,18], the research team iteratively integrated the 2 sources (Figure 1). One researcher (CYT) first mapped the review-derived strategies onto the interview-derived determinants to identify alignment, additional considerations, and remaining gaps, and then prepared draft recommendations organized according to DHEF levels and cross-level considerations. These drafts were circulated to multidisciplinary experts by email for feedback and refinement. Feedback was incorporated through iterative discussion within the research team, with further consideration of expert comments where needed. After finalizing the recommendations, CYT added illustrative implementation priorities, primary stakeholder groups, and example indicators to support interpretation and practical application. These classifications were then shared with the multidisciplinary experts for final review and confirmation. The recommendations were therefore developed through iterative synthesis, team discussion, and expert input rather than through a formal consensus procedure.

Figure 1. Study integration process for recommendation generation.

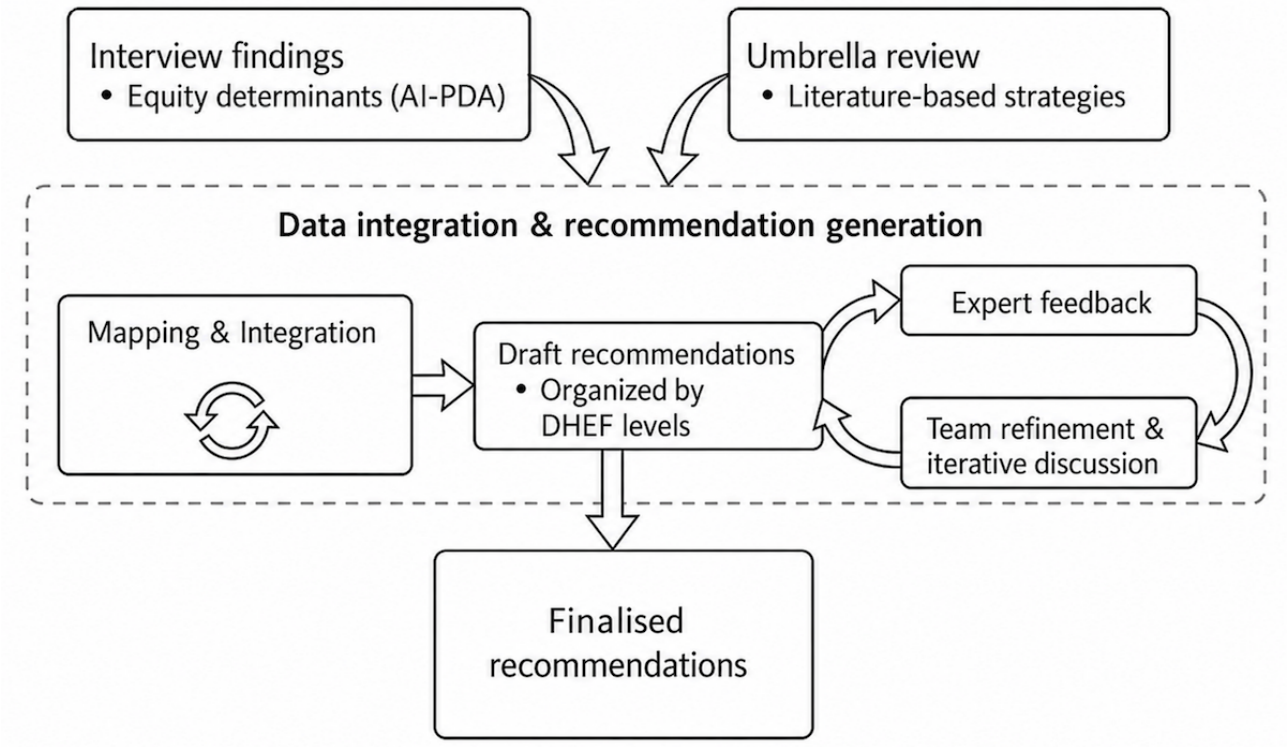

Ethical Considerations

Ethics approval for the stakeholder interviews of this study was granted by The Chinese University of Hong Kong Survey and Behavioural Research Ethics (reference number SBRE-24-0708). Written informed consent was obtained from all interview participants before data collection. Each participant received an HKD 100 (US $12.8) incentive in recognition of their time and contribution, consistent with local research practice and as approved by the ethics committee. The umbrella review and recommendation-generation involved evidence synthesis rather than direct human-subject research and therefore did not require separate ethics approval.

Results

Participant Demographics

Table 1 summarizes the sociodemographic characteristics of the participants, including 15 older adult patients with hypertension and/or diabetes (12/15, 80% with secondary education or below), 10 medical students, and 8 HCP. Except for 1 patient, all participants managed their conditions within the public health care system, so the identified determinants primarily reflect the public health care context. Table 2 presents the identified key equity determinants related to AI-PDAs from the interviews. Drawing on these qualitative insights and evidence from the 13 included reviews [13,19-30], we further engaged with a multidisciplinary advisory team (n=8; 2 experts each from medicine, public health, computer science, and social services) to propose actionable recommendations for operationalizing DHEF for older adults' AI-PDAs. No major disagreements arose during the recommendation development process.

XSL•FO
RenderX

Table 1. Sociodemographic characteristics of the participants.

	Patients (n=15)	Medical students (n=10)	Health care professionals (n=8)
Age (years), mean (SD)	69.9 (4.4)	21.0 (0.9)	41.6 (12.0)
Sex, n			
Male	9	5	6
Female	6	5	2
Educational attainment, n			
Secondary and below	12	0	0
Postsecondary	3	10	7
Household income, n			
Below median monthly household income	6	—[a]	—
At the median monthly household income	5	—	—
Over the median monthly household income	4	—	—
Major, n			
Medicine	—	5	—
Nursing	—	2	—
Public health/gerontology/community health practice	—	3	—
Expertise, n			
Specialist—family medicine	—	—	2
Specialist—geriatric medicine	—	—	2
Nurse	—	—	2
Public health practitioners	—	—	2
Clinical experience, n			
Less than 6 years	—	—	2
6 years or more	—	—	6

[a]Not available.

Table 2. Key equity determinants of artificial intelligence–enabled patient decision aids (AI-PDAs) based on the Digital Health Equity Framework[a].

Equity determinants of AI-PDAs	Health care environment	Digital environment
Individual	Limited health literacy; low self-efficacy; age-related cognitive decline	Limited digital literacy; low digital efficacy; distrust of AI[b]
Interpersonal	Paternalistic communication; weak patient-clinician relationship	Limited clinician engagement; provider technology bias
Community	Limited community support for SDM[c]	Limited community support for digital tools
Societal	Health system constraints; paternalistic medical culture	Affordability; regulatory gaps; poor health care integration

[a]This table is based on findings from the stakeholder interviews.

[b]AI: artificial intelligence.

[c]SDM: shared decision-making.

Equity Determinants Related to AI-PDAs in the Health Care Environment

Key equity determinants affecting older adults' use of AI-PDAs in health care are summarized below.

Individual Level

At this level, patients frequently reported limited health literacy (eg, "I don't know how to ask my doctor question" [P2]), low self-efficacy (eg, "I don't have much to say to the doctor because your level is different from the doctor's" [P7]), and age-related cognitive decline (eg, "I forgot to ask the doctor... I had a lot of medications last time" [P9]), as determinants related to

XSL•FO
RenderX

participation in SDM. Health care professionals and medical students also recognized these issues. They frequently described a subset of older adults who were more adept at seeking health information, asking relevant questions, and actively participating in consultations as smart. They noted that this group was more likely to be involved in SDM.

Interpersonal Level

Older adult patients often attend medical consultations without family accompaniment, making the patient-clinician relationship central in SDM. Two key determinants emerged at this level. First, clinicians' paternalistic communication. Patients reported that doctors rarely presented multiple options; instead, communication often took the form of informing patients of a decision that had already been made (eg, "The doctor just informed you of the treatment, no space for me to make a choice" [P8]). Second, a weak patient-clinician relationship. Some patients expressed reluctance to voice preferences or concerns due to fear of negative attitudes from doctors. For example, one patient (P10) explained, "If I ask too many questions, the doctors will get annoyed." By contrast, a patient treated by a physician recommended by a trusted church member described a more positive experience. She shared: "Because I knew the doctor quite well, I usually asked more questions" (P3).

Community Level

Patients faced challenges at the community level, where infrastructure to support SDM was largely absent (eg, patient education workshops related to SDM). With limited guidance on how to participate in decision-making, all patients often turned to YouTube, newspapers, and TV to address their health concerns, while community centers and official government resources were seldom consulted. This lack of SDM training was also observed by the medical community. Medical students reported limited exposure to SDM concepts, noting that curricula emphasized biomedical knowledge over patient engagement: "In class, the topic was not explained in depth, just briefly mentioned" (MD1).

Societal Level

Systemic issues related to the health care system and social norms emerged as important determinants at this level. Limited consultation time and fragmented continuity of care constrained both patients and providers. For example, one patient described the time pressure during consultations: "You wait for an hour, and then just get five or six minutes for the consultation" (P11). Health care professionals (HCP2, 3, 4) similarly reported frustration at being unable to provide comprehensive care under these constraints. Social norms, particularly a historically paternalistic medical culture, also influenced experiences. The majority of patients reported that they respected authority and simply followed doctors' instructions.

Equity Determinants Related to AI-PDAs in the Digital Environment

Key equity determinants influencing older adult patients' adoption of AI-PDAs in the digital environment are summarized below.

Individual Level

Patients often reported that their use of AI-PDAs was strongly shaped by limited digital literacy (eg, "I didn't know a mobile app could record blood pressure" [P4]), low digital self-efficacy (eg, "I'm just not good with technology" [P9]), and age-related cognitive decline (eg, "I always forgot my log in password" [P10]). A few patients (P3, P5, P10) tended to express their distrust in AI, particularly regarding the reliability of data sources and the readiness of new technologies for clinical use.

Interpersonal Level

Limited clinician engagement may hinder older adults' use of AI-PDAs, as clinicians may be skeptical of AI and concerned that AI-generated information. As one nurse explained, "While the AI might provide information about ten different treatment medication options, then the patients have 10 questions, if so... we won't use this tool because we don't have time to address all these questions" (HCP4). One patient (P10) expressed concern about no feedback from the clinicians if he used the tool. In addition, implicit technology bias can shape both provider attitudes and patient experiences. Some health care professionals assume that older adults are inherently unable to use digital tools, as reflected in one comment: "Even if you teach them how to use an app once, they might forget by the second time" (HCP6). However, other professionals recognize this bias and emphasize that older adults can engage digitally when provided with adequate support. "There's this bias that they can't use digital tools, but in reality, they can—it's just that they need support and facilitation. With that, they might actually manage it" (HCP5). Finally, the presence or lack of technological assistance from family members or caregivers also plays an important role. One patient described how her daughter, a nurse, taught her how to use the government's digital tool to make a medical appointment, which enabled her to navigate the system effectively (P2).

Community Level

Limited community support for using digital tools is a crucial determinant. Among the digital health tools most familiar to older adults is the government's official apps. Patients reported that they usually downloaded and learned to use the app during community-based promotion activities or workshops, which provided them with initial guidance and confidence in adoption. A public health care practitioner (HCP6) noted that assisting older adults with a digital platform for booking medical consultations was particularly beneficial for this population.

Societal Level

Many patients reported that affordability (eg, "I wouldn't use the app if it's not free" [P7]) would affect their use of this tool. HCPs emphasized that poor incorporation of AI-PDAs into health care systems would lead to underuse and limit equitable adoption. As one clinician (HCP1) shared, he collaborated with a research team to use an app for tracking patients' hypertension. Typically, he would ask patients to show their blood pressure records through the app during consultation, which helped patients gradually become familiar with using it. He commented, "You really need to engage doctors and embed the tool into clinical workflows; otherwise, patients won't use it." Regulatory

XSL•FO
RenderX

gaps represent another crucial determinant. Inadequate regulatory oversight may lead to variability in the quality of health information provided by these tools, undermining stakeholders' trust and fair uptake. As one professional cautioned, "Be careful with AI, as its outputs may not always be consistent. You might get different answers today compared to tomorrow." (HCP1).

Strategies for Advancing Digital Health Equity From an Umbrella Review

In the umbrella review, 13 identified reviews [13,19-30] were included (Figure 2). Strategies were synthesized across DHEF levels (Multimedia Appendix 3). Although the DHEF does not explicitly include a technology level, technology-related factors intersect with individual, interpersonal, community, and societal determinants, supporting a cross-level perspective in examining digital health equity. Technology-related determinants were most frequently represented in the review literature (n=26). In contrast, determinants at the individual (n=6) and interpersonal levels (n=4) were comparatively underrepresented, while community-level (n=10) and societal level (n=12) determinants received moderate attention. These frequencies indicate the degree of emphasis placed on each DHEF level in the existing literature rather than the relative importance of the corresponding strategies.

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

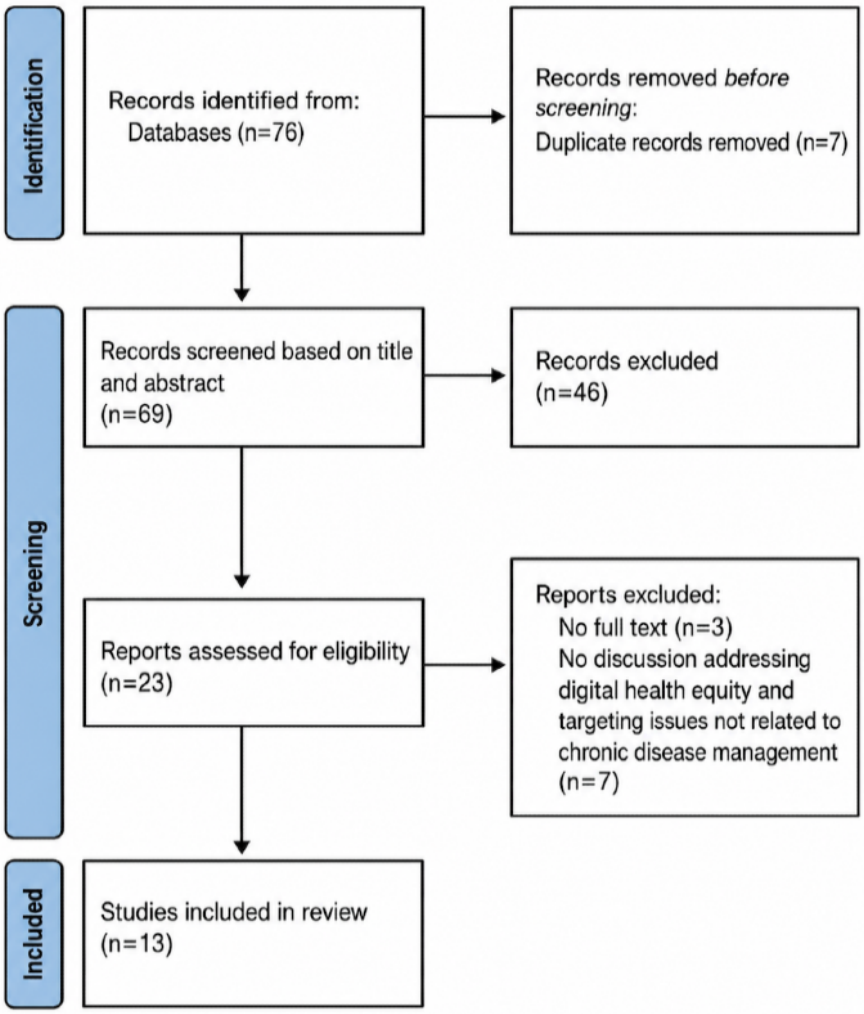

Recommendations on Operationalizing the Digital Health Equity Framework

Overview

By mapping interview-derived determinants onto a synthesized evidence base from reviews, and engaging a multistakeholder team, we generated a set of actionable recommendations for applying the DHEF to the design of AI-PDAs for older adults (Table 3).

XSL•FO
RenderX

Table 3. Key recommendations on operationalizing Digital Health Equity Framework (DHEF) in artificial intelligence–enabled patient decision aids (AI-PDAs) for older adults[a].

DHEF level	Recommendation	Key actions	Implementation priority	Primary stakeholders	Example indicators
Individual	Co-design with end users to address their needs	Tailor content to varying health and digital literacy levels; incorporate features that support self-efficacy and digital confidence; accommodate age-related cognitive needs through accessible design and communication.	Early-stage (design and pilot testing)	Developers, clinicians, patients	Subgroup uptake, onboarding completion, comprehension, need for assistance, self-efficacy, and sustained use
Interpersonal	Embrace relationship-centered design	Involve clinicians in design and workflow integration; maintain human oversight of AI[b]-generated recommendations; support participation of caregivers and family members where appropriate.	Early-stage (design and pilot testing)	Developers, clinicians, service managers, patients, caregivers	Clinician and caregiver acceptability, quality of communication during consultations, and perceived workflow burden
Community	Leverage community resources to improve support	Engage community partners in co-design and evaluation; provide community-based digital health support programs; connect users with trusted local resources.	Medium- to long-term (implementation and scale-up)	Community organizations, clinicians, service managers, policymakers	Participation in community-based support activities, sustained use with community assistance, perceived cultural relevance, and feedback from community stakeholders
Societal	Promote accessible and equitable AI governance	Reduce financial and technical barriers; establish standards for safety, transparency, and accountability; integrate AI-PDAs into routine care pathways; foster multidisciplinary collaboration.	Medium- to long-term (implementation and scale-up)	Policymakers, regulators, health system leaders, service managers	Affordability, staff training completion, integration into routine workflows, reported safety or escalation events, and institutional readiness for sustained implementation
Cross-level	Enhance equitable AI through algorithmic fairness	Ensure representative and inclusive datasets; implement bias mitigation throughout model development and deployment; support transparent reporting and equity monitoring.	Foundational technical standard (from design onward and remain central during postdeployment evaluation)	Developers, data scientists, regulators, institutional leaders	Subgroup-specific uptake and outcomes, fairness performance reporting across key older adult subgroups, documentation of bias mitigation steps, transparency of model limitations, and stakeholder-informed review of equity implications

[a]Content is based on integrated insights from stakeholder interview findings and the umbrella review, synthesized by the research team.

[b]AI: artificial intelligence.

Recommendation 1: Co-Design With End Users to Address Their Needs

Primarily directed to developers, clinicians, and patients, this recommendation emphasizes co-design and participatory approaches to ensure that the digital health solution reflects older adults' needs, capabilities, and preferences [21]. Key strategies include:

- Tailor content to patient health and digital literacy: health literacy and digital literacy are key determinants of effective AI-PDAs, yet they are often overlooked in personalized interventions for chronic disease management [19,20,31]. Designers should first understand end users' health and digital literacy levels, then adapt content to match these varying levels [20]. For example, provide simplified explanations of medical terms, step-by-step guidance, and longer training sections for beginners [13], while offering advanced features and data visualization for more experienced users.
- Psychological support to build self-efficacy and digital self-efficacy: psychological support is a potential mediator in building strong patient-clinician relationships and fostering sustained engagement with digital health tools for older adults [28,30]. Empathetic AI-enabled tools can

XSL•FO
RenderX

provide encouraging feedback, a sense of progress, and tailored emotional guidance, enhancing older adult patients' confidence in using these tools and helping alleviate stress associated with chronic illness and digital technologies.
- Support cognitive needs: AI-PDAs should accommodate age-related cognitive challenges through simplified interfaces, larger fonts, reminder messages, and voice guidance [28]. Additionally, present information in plain language and embed it in practical, everyday contexts to enhance comprehension and usability [22,24,28].

This recommendation may be considered an early-stage implementation priority as these strategies are often the most immediately actionable during design and pilot testing, although they remain important during implementation, and illustrative indicators of success include subgroup uptake, onboarding completion, comprehension, need for assistance, self-efficacy, and sustained use.

Recommendation 2: Embrace Relationship-Centered Design

This recommendation requires close collaboration among designers, clinicians, service managers, patients, and caregivers, as AI-PDAs are inherently situated within a network of interactions among patients, HCPs, and caregivers. Their design must explicitly account for these relational dynamics. Key strategies include:

- Engage clinicians in AI tool design: involve HCPs in co-design to ensure AI-PDAs complement clinical practice [28,30]. This engagement also raises awareness of the diverse digital capabilities of older adults, helping address clinicians' implicit technology bias and strengthening clinician-patient relationships.
- Reinforce human-in-the-loop integration: clinicians are embedded within the AI-PDA workflow, actively interacting with patients while reviewing and interpreting AI-generated recommendations [19]. This ensures that AI-PDAs support, rather than replace, critical clinical judgment, thereby enhancing clinicians' and patients' trust in these tools.
- Support social interdependence: acknowledge that digital tool usage is often a collaborative activity for older adults. AI-PDAs' architectures should therefore facilitate supported use by integrating functionalities for caregivers and family members to participate in data entry, progress monitoring, and decision support [28,30].

This recommendation, such as Recommendation 1, may also be considered an early-stage implementation priority, and illustrative indicators of success include clinician and caregiver acceptability, quality of communication during consultations, and perceived workflow burden associated with integrating the AI-PDA into routine care.

Recommendation 3: Leverage Community Resources to Improve Support

Primarily relevant to clinicians, service managers, community organizations, and policymakers, this recommendation recognizes that community-level determinants shape the adoption and effective use of AI-PDAs. Key strategies include:

- Engage community partners in co-design: engage community representatives (eg, nongovernmental organizations, faith-based organizations, and lived experience advisors) in co-design and usability testing, thereby linking AI-PDAs to trusted community resources. Community members can also contribute to study recruitment and ensure that diverse voices are represented [22].
- Mobilize community-based adoption support: facilitate adoption of AI tools through community-based initiatives, including health literacy, digital literacy, and AI literacy programs for patients, caregivers, and health care workers (eg, workshops or "tech buddy" programs) [22,30].
- Include community perspectives in ongoing evaluation: engage community representatives in the ongoing evaluation of AI-PDAs and longitudinal design to ensure cultural relevance and equitable outcomes [21,25,32].

This recommendation is more likely to be considered as medium- to longer-term priorities as these strategies depend on external coordination with community partners and sustained engagement beyond the initial pilot phase, and illustrative indicators of success include participation in community-based support activities, sustained use with community assistance, perceived cultural relevance, and feedback from community stakeholders.

Recommendation 4: Promote Accessible and Equitable AI Governance in Society

This recommendation depends particularly on policymakers, regulators, health system leaders, and service managers, as equitable implementation requires social support and governance mechanisms to reduce access barriers and guarantee the ethical integrity of AI tools. Key strategies include:

- Ensure affordability and accessibility: ensure AI-PDAs are financially and physically accessible to all patients, including those from disadvantaged groups [13]. This may involve securing financial resources or policy-driven subsidies to provide low-cost or free access, minimize data or download requirements, and reduce complex log-in procedures with particular attention to the needs of older adults [32].
- Design standards to ensure credibility and equity: the design standards aim to guarantee responsible and technically reliable AI tools by mandating safety, reliability, transparency, and accountability to prevent disproportionate harms across diverse populations [23]. For example, these standards should require clear communication about information sources, provide references to high-quality medical resources [25], data protection [28], and emphasize equitable health care access [24].
- Integrate into health care system workflows: ensuring systemic connections and clinical sustainability is crucial [24]. Identify suitable adoption settings, such as primary, secondary, or tertiary care, and embed AI-PDAs into existing workflows. Examples may include use before the consultation during nurse-led preassessment, in waiting areas through tablets, and during chronic disease follow-up in community health care centers [28]. Additionally,

XSL•FO
RenderX

implementation should be accompanied by training and guidance for health care professionals to ensure effective adoption while maintaining workflow efficiency and high-quality care [28].

- Multidisciplinary collaboration: AI designers must recognize how technology intersects with this complex historical and social context [21]. Effective digital implementation in AI tools needs a strong public and private partnership and multidisciplinary collaboration, where nongovernmental organizations, private companies, academic institutes (eg, engineering, informatics, and social sciences), and government can collaborate to leverage each other's strengths [19,32]

This recommendation is also more likely to be considered a medium- to longer-term priority as it depends on broader institutional, regulatory, and system-level changes, and illustrative indicators of success include affordability, staff training completion, workflow integration, safety or escalation events, and institutional readiness for sustained implementation.

Recommendation 5: Enhance Equitable AI Through Algorithmic Fairness

Primarily directed to designers, engineers, data scientists, regulators, and institutional leaders, this recommendation addresses data practices, model development, and reporting and governance processes. While technical implementation of fairness methods may be led by developers and data scientists, clinicians, patients, caregivers, and community partners contribute primarily through oversight, contextual interpretation, and identification of equity concerns in real-world use. The strategies below align with existing AI governance and fairness guidance, while highlighting considerations that are particularly relevant to AI-PDAs for older adults with chronic disease:

- Representative and inclusive data practices (developers, data scientists, and institutions): datasets should represent diverse populations across age, sex, race/ethnicity, and socioeconomic status, particularly including historically disadvantaged subgroups [19,25]. As the field evolves to use novel data from wearables and sensors, proactive bias mitigation is essential to identify and counteract new forms of digital usage bias inherent in these sources [19]. Sampling limitations should be transparently documented [20], and data privacy and security should be safeguarded throughout model development and implementation [29].
    - Distinctive consideration for older adults' AI-PDAs: datasets should capture heterogeneity among older adults, including age subgroup, health and digital literacy, caregiver dependence, health status, and living setting [20,24]. Recruitment and data collection should avoid single-mode strategies (eg, digital-only and clinic-only) that may systematically exclude more vulnerable groups [26].
- Fairness-aware model development and evaluation (developers and data scientists): this begins with preprocessing techniques such as oversampling underrepresented groups or applying importance weighting to mitigate initial data biases [21]. In-process techniques are then critical, directly embedding fairness into the model's training through methods like parity constraints, fairness regularization, and adversarial debiasing to prevent the algorithm from learning discriminatory patterns [21]. Finally, postprocessing adjustments, such as calibrating model thresholds or test-retest for different subgroups, ensure equitable outcomes [21,23].
    - Distinctive consideration for older adults' AI-PDAs: in chronic disease management for older adults, false negatives and false positives may carry different clinical and psychosocial consequences across subgroups. Fairness assessment should therefore extend beyond statistical parity to include subgroup-specific evaluation of clinical impact, stakeholder-informed equity review, and threshold calibration based on differential risks, benefits, and implementation burden [26,29].
- Transparent equity reporting and governance (developers, institutions, and regulators): this includes (1) adherence to established AI-specific reporting guidelines (eg, Consolidated Standards of Reporting Trials-Artificial Intelligence [33] and Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis-Artificial Intelligence [34]) in all publications [19]; (2) public disclosure of equity impact assessments, including subgroup-specific performance metrics [24,26]; (3) formally incorporating and documenting feedback from diverse stakeholders to identify and address potential biases; and (4) explicitly communicating the AI's intended use, capabilities, and limitations to end users to support appropriate use and prevent misuse [20].
    - Distinctive consideration for older adults' AI-PDAs: reporting should also support understanding and trust through plain-language explanations, disclosure of developers and evidence sources, documentation of input from older adults and caregivers, and attention to cultural, linguistic, and dialect diversity [25].

Algorithmic fairness should be embedded throughout the AI life cycle, with ongoing evaluation of subgroup performance, transparency of model limitations, and stakeholder-informed review of equity implications.

Discussion

Principal Findings

Although broader equity-oriented checklists have been proposed for AI and digital health, few studies have translated the DHEF into practical recommendations for developing AI-PDAs for older adults. This study addresses this gap by providing DHEF-informed recommendations tailored to this context. Drawing on stakeholder interviews and an umbrella review, we generated recommendations across 5 socioecological levels to guide more equitable design and implementation. This framing reflects the growing recognition that advancing equity-oriented digital health design requires moving beyond purely technology-centered or user-centered approaches [14,35,36]

Across the recommendations, a key finding was that equity risks emerged at both relational and structural levels. Stakeholder interviews gave particular prominence to individual- and

XSL•FO
RenderX

interpersonal-level determinants, including trust, communication, caregiver involvement, emotional support, and confidence in using digital tools. These issues received comparatively less attention in the identified reviews, consistent with previous evidence that equity-related factors at individual and interpersonal levels are often underexamined in digital health research [11]. This gap is particularly important for older adults, for whom encouragement, trust, and support from clinicians, caregivers, and social networks can shape whether digital tools are perceived as usable, relevant, and safe [8,37].

At the same time, the umbrella review more strongly captured upstream community-, societal-, and technical equity risks, including governance, affordability, participatory design, data representation, and algorithmic fairness. We interpreted this divergence between stakeholder interviews and literature review as reflecting the comparative strengths of each evidence source rather than a conflict to resolve: stakeholders captured equity barriers as experienced in lived care interactions, while the literature captured upstream structural risks. These map onto distinct levels within the DHEF. During the iterative mapping process, literature-derived upstream concerns were retained and formalized into Recommendations 3, 4, and 5, even where stakeholders did not raise them. Conversely, stakeholder-identified relational determinants, such as human-in-the-loop integration and caregiver role-based access, were prioritized in Recommendations 1 and 2, even where literature coverage was limited. This level-sensitive integration is important because equitable AI-PDAs for older adults require both relationally responsive design and broader institutional, community, and policy support [38-41].

Implications

Building on this multilevel interpretation, a central contribution of this work is to reconceptualize digital health equity in patient-facing AI as a sociotechnical and relational problem, not simply a question of access, connectivity, or algorithmic performance. Prior digital health equity frameworks, including the DHEF and related multilevel syntheses [42,43], have been valuable in identifying key determinants across policy, system, community, interpersonal, and individual levels, particularly with respect to general foci on access, connectivity, digital literacy, and digital inclusion. However, they have provided less explicit guidance on the mechanisms through which these determinants shape the real-world use of patient-facing AI in clinical encounters already marked by unequal communication, limited consultation time, workflow constraints, and variable caregiver involvement. Our findings suggest that the health care environment is not merely a background context, but the mechanism through which digital tools are either converted into meaningful participation in care or fail to generate benefit. In this sense, AI-PDAs do not enter an equity-neutral setting; rather, they are introduced into encounters already structured by disparities in health literacy, clinician-patient communication, trust, caregiver dependence, and system-level access barriers affecting older adults. This shifts the equity question from whether patients can access AI to whether they can actually use it to gain understanding, exercise agency, and influence decisions within real consultations. The implication is that equitable AI implementation requires more than technically robust tools. It requires workflow-sensitive, relationship-aware, and institutionally grounded design that accounts for how benefits are mediated, amplified, or constrained in practice [44,45].

Importantly, we position AI algorithmic fairness as a cross-level domain, rather than a separate level, because the DHEF does not specify a stand-alone “technology” level [10]. Rather than functioning independently, algorithmic fairness is shaped by and operates across social and structural conditions throughout the framework. For AI-PDAs for older adults, fairness cannot be treated as a downstream concern to be checked after deployment. Instead, it requires early and ongoing attention to data representation, bias mitigation, subgroup-specific evaluation, and transparent reporting across the full AI life cycle. This is particularly important because older adults are heterogeneous, and inequities may arise not only across conventional demographic categories but also across differences in health literacy, digital literacy, caregiver dependence, health status, and living setting [46,47]. By integrating user-centered and technical equity considerations, this work provides practical, cross-disciplinary guidance for advancing equitable AI implementation in health care [48-50].

While the multilevel logic of the DHEF provides a transferable organizing structure, the specific recommendations were developed in the Hong Kong context and therefore require contextual adaptation before implementation elsewhere. Hong Kong’s predominantly public-led health system, marked by high service demand, time-constrained consultations, and particular care pathways, likely shaped both the need for and the potential role of AI-PDAs. Transferability may also be influenced by local cultural norms surrounding medical authority, family involvement in care, and attitudes toward technology. For example, in settings where health decisions are shaped more collectively through family involvement, the adaptation of AI-PDAs should consider who is expected to use the tool, whether caregivers should be included in the decision-support process, and how patient autonomy and caregiver input can be balanced. In contexts where patients are less accustomed to questioning clinicians or using self-directed support tools, implementers should consider whether AI-PDAs need to include additional guidance on preparing questions, clarifying preferences, and initiating SDM conversations. In addition, Hong Kong’s dense urban environment and relatively strong digital connectivity may facilitate access and implementation in ways that do not readily translate to settings with more dispersed populations, weaker digital infrastructure, or greater barriers to continuity of care. In such settings, implementers should consider whether older adults have reliable access to devices, connectivity, and support, and how AI-PDAs can be embedded into existing community or primary care workflows.

Accordingly, the key issue is not whether these recommendations are transferable in principle, but how their relative priority, sequencing, and feasibility should be adapted across settings. For example, in low- and middle-income countries, Recommendations 3 and 4 may need to place greater emphasis on foundational infrastructure and community health worker integration before institutional deployment, with explicit

XSL•FO
RenderX

attention to the training and governance arrangements this requires [51,52]. Similarly, in settings with different cultural attitudes toward aging, medical authority, and family involvement, Recommendations 1 and 2 may need recalibration to reflect local relational norms [53]. For instance, where family members play a more central role in health decision-making, or where deference to clinical authority shapes how patients engage with self-management tools. Future research should therefore test and refine these recommendations across diverse health systems, sociocultural, and digital contexts.

Limitations

Limitations of this study include the following. First, although we followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines for our literature review, some relevant studies may have been omitted due to subjective judgments during the selection process. Second, we only included studies published in English, potentially limiting the comprehensiveness of our analysis. Third, we did not formally assess the overlap of primary studies across the included reviews; accordingly, the reported frequencies reflect how often strategies appeared in the review literature rather than the number of unique supporting studies. Fourth, we did not conduct a formal methodological quality appraisal of the included reviews; therefore, the synthesized strategies should be interpreted as descriptive review-level findings rather than as evidence of equal strength across recommendations. Fifth, the classification of suggested strategies by level involved subjective judgment by the research team. While we applied systematic procedures and multidisciplinary discussions to enhance consistency, some degree of subjectivity may have influenced the categorization. Future work could involve in-depth consultations with AI tool developers to gain additional insights on equity determinants. Sixth, these recommendations, although informed by stakeholder perspectives and the umbrella review, have not yet been empirically tested as formal design requirements or evaluated for their effects on digital health equity outcomes. They should therefore be considered preliminary, evidence-informed design considerations. Seventh, the recommendations were refined through iterative email feedback from multidisciplinary experts rather than through formal consensus meetings or a structured consensus method. Although this pragmatic approach supported multidisciplinary input, it may have limited reproducibility and reduced transparency in how differing perspectives were weighed and how priorities were established. Finally, social desirability bias may have influenced the findings, as clinicians may have underreported paternalistic practices and patients may have minimized dissatisfaction or hesitated to express criticism. Although we sought to mitigate this through semistructured interviews and cross-stakeholder comparison, this bias cannot be fully excluded.

Conclusion

This study operationalizes the DHEF by translating its concepts into practical strategies for designing inclusive AI-PDAs for older adult care. The findings suggest that health care settings function as a mediating sociotechnical context through which AI tools may either support or constrain equitable participation in care. Consequently, the study underscores the necessity of interdisciplinary collaboration encompassing technology designers, clinicians, public health professionals, social service providers, and policymakers to reconcile technological innovation with user-centered and equity-oriented design. Future work should focus on translating these recommendations into co-designed AI-PDA prototypes, testing them in real-world clinical settings, and establishing measurable equity-relevant outcomes and process indicators to evaluate whether implementation is both effective and equitable.

Acknowledgments

We sincerely thank all participants for their time, insights, and contributions to this study. The authors acknowledge the use of ChatGPT (GPT-4 [OpenAI], 2026) solely to improve the clarity and readability of selected text in the manuscript. AI tools were not used to generate study data, conduct analyses, or develop core interpretations. All artificial intelligence–generated outputs were carefully reviewed, revised, and verified by the authors, who take full responsibility for the final content.

Data Availability

The datasets generated and analyzed during the study are not publicly available due to the inclusion of potentially identifiable information from stakeholder interviews, but are available from the corresponding author on reasonable request, subject to ethical approval and data sharing agreements. The review of reviews included publicly available literature, as cited in the manuscript.

Funding

This work was funded by the Knowledge Transfer Project Fund of The Chinese University of Hong Kong (KPF25GWP47).

Authors' Contributions

Conceptualization: CYT, ELYW
Methodology: CYT, KW, XY
Formal analysis: CYT, KW, XY
Investigation: CYT, AWLC, JCHM, CL, JCYY
Data Curation: CYT, AWLC, JCHM, CL, JCYY
Funding acquisition: ELYW

XSL•FO
RenderX

Project administration: ELYW
Writing—original draft: CYT
Writing—review and editing: CYT, KW, XY, JC, KO, IWKL, THCP, SZ, YW, ELYW
Supervision: ELYW

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of interview domains and illustrative questions.
[DOCX File , 30 KB - jmir_v28i1e89011_app1.docx]

Multimedia Appendix 2
Search algorithms of the umbrella review.
[DOCX File , 21 KB - jmir_v28i1e89011_app2.docx]

Multimedia Appendix 3
Summary of suggested strategies for addressing digital health equity from the included reviews.
[DOCX File , 33 KB - jmir_v28i1e89011_app3.docx]

References

1. Yen FS, Wei JC, Chiu LT, Hsu CC, Hwu CM. Diabetes, hypertension, and cardiovascular disease development. J Transl Med 2022;20(1):9. [doi: 10.1186/s12967-021-03217-2] [Medline: 34980154]
2. Barry MJ, Edgman-Levitan S. Shared decision making--pinnacle of patient-centered care. N Engl J Med 2012;366(9):780-781. [doi: 10.1056/NEJMp1109283] [Medline: 22375967]
3. Wu CJ, Yeh TP, Chu G, Ho YF. Effectiveness of computerised decision aids for patients with chronic diseases in shared decision-making: a systematic review and meta-analysis. J Clin Nurs 2024;33(7):2732-2754. [doi: 10.1111/jocn.17095] [Medline: 38553843]
4. Wang Y, Li PF, Tian Y, Ren JJ, Li JS. A shared decision-making system for diabetes medication choice utilizing electronic health record data. IEEE J Biomed Health Inform 2017;21(5):1280-1287. [doi: 10.1109/jbhi.2016.2614991]
5. Sheng B, Pushpanathan K, Guan Z, Lim QH, Lim ZW, Yew SME, et al. Artificial intelligence for diabetes care: current and future prospects. Lancet Diabetes Endocrinol 2024;12(8):569-595. [doi: 10.1016/s2213-8587(24)00154-2]
6. Cho JS, Park JH. Application of artificial intelligence in hypertension. Clin Hypertens 2024;30(1):11 [FREE Full text] [doi: 10.1186/s40885-024-00266-9] [Medline: 38689376]
7. Hart JT. The inverse care law. Lancet 1971;1(7696):405-412. [doi: 10.1016/S0140-6736(71)92410-X]
8. Kokorelias KM, Nelson M, Tang T, Steele Gray C, Ellen M, Plett D, et al. Inclusion of older adults in digital health technologies to support hospital-to-home transitions: secondary analysis of a rapid review and equity-informed recommendations. JMIR Aging 2022;5(2):e35925 [FREE Full text] [doi: 10.2196/35925] [Medline: 35475971]
9. Veras M, Auger LP, Sigouin J, Gheidari N, Nelson ML, Miller WC, et al. Ethics and equity challenges in telerehabilitation for older adults: rapid review. JMIR Aging 2025;8:e69660. [doi: 10.2196/69660] [Medline: 40802966]
10. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022;5(1):119 [FREE Full text] [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
11. Abejirinde IO, Kishimoto V, Pfisterer KJ, Uddin N, McGuire K, Brual J, et al. An equity analysis of remote patient monitoring programs unveils assumptions on digital health equity. NPJ Digit Med 2025;8(1):320 [FREE Full text] [doi: 10.1038/s41746-025-01731-x] [Medline: 40442266]
12. Narayan SM, Chung MK, Adedinsewo D, Brant LCC, Davis LL, Duncker D, et al. Access to digital health technologies: personalized framework and global perspectives. Nat Rev Cardiol 2026;23(1):9-22. [doi: 10.1038/s41569-025-01184-5] [Medline: 40670723]
13. Salmi EM, Basile FW, Khan FA, Watt L, Song R, Bijker EM. Facilitators and barriers affecting the implementation of e-health for chronic respiratory diseases in remote settings: a qualitative evidence synthesis. BMC Health Serv Res 2025;25(1):19 [FREE Full text] [doi: 10.1186/s12913-024-12050-4] [Medline: 39754241]
14. Rodriguez NM, Burleson G, Linnes JC, Sienko KH. Thinking beyond the device: an overview of human- and equity-centered approaches for health technology design. Annu Rev Biomed Eng 2023;25(1):257-280 [FREE Full text] [doi: 10.1146/annurev-bioeng-081922-024834] [Medline: 37068765]
15. Rowland S, Brewer LC, Rosas LG. Digital health equity: a call to action for clinical and translational scientists. J Clin Transl Sci 2024;8(1):e145. [doi: 10.1017/cts.2024.564] [Medline: 39478785]

XSL•FO
RenderX

16. Pratt B, Parker M, Bull S. Equitable design and use of digital surveillance technologies during COVID-19: norms and concerns. J Empir Res Hum Res Ethics 2022;17(5):573-586 [FREE Full text] [doi: 10.1177/15562646221118127] [Medline: 36069118]
17. Gaebel W, Lukies R, Kerst A, Stricker J, Zielasek J, Diekmann S, et al. Upscaling e-mental health in Europe: a six-country qualitative analysis and policy recommendations from the eMEN project. Eur Arch Psychiatry Clin Neurosci 2021;271(6):1005-1016. [doi: 10.1007/s00406-020-01133-y] [Medline: 32393997]
18. Haig M, Main C, Chávez D, Kanavos P. A value framework to assess patient-facing digital health technologies that aim to improve chronic disease management: a Delphi approach. Value Health 2023;26(10):1474-1484 [FREE Full text] [doi: 10.1016/j.jval.2023.06.008] [Medline: 37385445]
19. Clifford N, Tunis R, Ariyo A, Yu H, Rhee H, Radhakrishnan K. Trends and gaps in digital precision hypertension management: scoping review. J Med Internet Res 2025;27:e59841 [FREE Full text] [doi: 10.2196/59841] [Medline: 39928934]
20. Harrington C, Martin-Hammond A, Bray KE. Examining identity as a variable of health technology research for older adults: a systematic review. 2022 Presented at: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems; 2026 June 9; New Orleans. [doi: 10.1145/3491102.3517621]
21. Iloanusi NJ, Chun SA. AI impact on health equity for marginalized, racial, and ethnic minorities. 2024 Presented at: Proceedings of the 25th Annual International Conference on Digital Government Research; 2026 June 9; Taipei, Taiwan. [doi: 10.1145/3657054.3657152]
22. Katz ME, Mszar R, Grimshaw AA, Gunderson CG, Onuma OK, Lu Y, et al. Digital health interventions for hypertension management in US populations experiencing health disparities: a systematic review and meta-analysis. JAMA Netw Open 2024;7(2):e2356070 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.56070] [Medline: 38353950]
23. Oikonomou EK, Khera R. Machine learning in precision diabetes care and cardiovascular risk prediction. Cardiovasc Diabetol 2023;22(1):259 [FREE Full text] [doi: 10.1186/s12933-023-01985-3] [Medline: 37749579]
24. Osonuga A, Osonuga AA, Fidelis SC, Osonuga GC, Juckes J, Olawade DB. Bridging the digital divide: artificial intelligence as a catalyst for health equity in primary care settings. Int J Med Inform 2025;204:106051 [FREE Full text] [doi: 10.1016/j.ijmedinf.2025.106051] [Medline: 40706198]
25. Qin H, Tong Y. Opportunities and challenges for large language models in primary health care. J Prim Care Community Health 2025;16:21501319241312571 [FREE Full text] [doi: 10.1177/21501319241312571] [Medline: 40162893]
26. Senteio C, Murdock PJ. The efficacy of health information technology in supporting health equity for Black and Hispanic patients with chronic diseases: systematic review. J Med Internet Res 2022;24(4):e22124 [FREE Full text] [doi: 10.2196/22124] [Medline: 35377331]
27. Sylla B, Ismaila O, Diallo G. 25 years of digital health toward universal health coverage in low- and middle-income countries: rapid systematic review. J Med Internet Res 2025;27:e59042 [FREE Full text] [doi: 10.2196/59042] [Medline: 40440696]
28. Feng G, Weng F, Lu W, Xu L, Zhu W, Tan M, et al. Artificial intelligence in chronic disease management for aging populations: a systematic review of machine learning and NLP applications. Int J Gen Med 2025;18:3105-3115 [FREE Full text] [doi: 10.2147/IJGM.S516247] [Medline: 40529344]
29. Guan Z, Li H, Liu R, Cai C, Liu Y, Li J, et al. Artificial intelligence in diabetes management: advancements, opportunities, and challenges. Cell Rep Med 2023;4(10):101213 [FREE Full text] [doi: 10.1016/j.xcrm.2023.101213] [Medline: 37788667]
30. Parker AG, Vardoulakis LM, Alla J, Harrington CN. Participatory AI considerations for advancing racial health equity. In; 2025 Presented at: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems; 2025 April 26; Yokohama. [doi: 10.1145/3706598.3713165]
31. Hao Y, Liu Z, Riter RN, Kalantari S. Advancing patient-centered shared decision-making with AI systems for older adult cancer patients. 2024 Presented at: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems; 2024 May 11-16; Honolulu. [doi: 10.1145/3613904.3642353]
32. Zainal NH, Liu X, Leong U, Yan X, Chakraborty B. Bridging innovation and equity: advancing public health through just-in-time adaptive interventions. Annu Rev Public Health 2025;46(1):43-68 [FREE Full text] [doi: 10.1146/annurev-publhealth-071723-103909] [Medline: 39656954]
33. Liu X, Cruz Rivera S, Moher D. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health 2020;2(10):e537-e548 [FREE Full text] [doi: 10.1038/s41591-020-1037-7] [Medline: 32908284]
34. Collins GS, Moons KGM, Dhiman P. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:q902 [FREE Full text] [doi: 10.1136/bmj.q902] [Medline: 38636956]
35. MacDonald L, Thomas E, Javernick-Will A, Austin-Breneman J, Aranda I, Salvinelli C, et al. Aligning learning objectives and approaches in global engineering graduate programs: review and recommendations by an interdisciplinary working group. Dev Eng 2022;7:100095. [doi: 10.1016/j.deveng.2022.100095]

XSL•FO
RenderX

36. Abbasgholizadeh Rahimi S, Cwintal M, Huang Y, Ghadiri P, Grad R, Poenaru D, et al. Application of artificial intelligence in shared decision making: scoping review. JMIR Med Inform 2022;10(8):e36199 [FREE Full text] [doi: 10.2196/36199] [Medline: 35943793]
37. Wilson S, Tolley C, Mc Ardle R, Lawson L, Beswick E, Hassan N, et al. Recommendations to advance digital health equity: a systematic review of qualitative studies. NPJ Digit Med 2024;7(1):173 [FREE Full text] [doi: 10.1038/s41746-024-01177-7] [Medline: 38951666]
38. Balakrishnan P, Leema AA, Sangaiah AK. Empathetic AI: multimodal interaction framework for emotional resilience and cognitive engagement in elderly care. New York: IEEE; 2025 Presented at: 2025 International Conference on Next Generation Information System Engineering (NGISE); 2025 March 28-29; Ghaziabad, Delhi, India p. 1-6. [doi: 10.1109/ngise64126.2025.11085267]
39. Kim DS, Eltahir AA, Ngo S, Rodriguez F. Bridging the gap: how accounting for social determinants of health can improve digital health equity in cardiovascular medicine. Curr Atheroscler Rep 2024;27(1):9. [doi: 10.1007/s11883-024-01249-9] [Medline: 39576395]
40. Park NY, Jang S. App-based digital health equity determinants according to ecological models: scoping review. Sustainability 2024;16(6):2232. [doi: 10.3390/su16062232]
41. Beauchamp A, Backholer K, Magliano D, Peeters A. The effect of obesity prevention interventions according to socioeconomic position: a systematic review. Obes Rev 2014;15(7):541-554. [doi: 10.1111/obr.12161] [Medline: 24629126]
42. Lyles CR, Nguyen OK, Khoong EC, Aguilera A, Sarkar U. Multilevel determinants of digital health equity: a literature synthesis to advance the field. Annu Rev Public Health 2023;44(1):383-405 [FREE Full text] [doi: 10.1146/annurev-publhealth-071521-023913] [Medline: 36525960]
43. Kim KK, Backonja U. Digital health equity frameworks and key concepts: a scoping review. J Am Med Inform Assoc 2025;32(5):932-944 [FREE Full text] [doi: 10.1093/jamia/ocaf017] [Medline: 39936843]
44. Salwei ME, Carayon P. A sociotechnical systems framework for the application of artificial intelligence in health care delivery. J Cogn Eng Decis Mak 2022;16(4):194-206 [FREE Full text] [doi: 10.1177/15553434221097357] [Medline: 36704421]
45. Song P, Jiang D, Zhou J, Zhu Y, Manaf RA, Bojude DA, et al. Research priorities for data science and artificial intelligence in global health: an international consensus exercise. Lancet Glob Health 2026;14(3):e455-e465 [FREE Full text] [doi: 10.1016/S2214-109X(25)00473-5] [Medline: 41713447]
46. Bertolazzi A, Quaglia V, Bongelli R. Barriers and facilitators to health technology adoption by older adults with chronic diseases: an integrative systematic review. BMC Public Health 2024;24(1):506 [FREE Full text] [doi: 10.1186/s12889-024-18036-5] [Medline: 38365698]
47. Kebede AS, Ozolins LL, Holst H, Galvin K. Digital engagement of older adults: scoping review. J Med Internet Res 2022;24(12):e40192 [FREE Full text] [doi: 10.2196/40192] [Medline: 36477006]
48. Chin MH, Afsar-Manesh N, Bierman AS, Chang C, Colón-Rodríguez CJ, Dullabh P, et al. Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. JAMA Netw Open 2023;6(12):e2345050 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.45050] [Medline: 38100101]
49. Char DS, Abràmoff MD, Feudtner C. Identifying ethical considerations for machine learning healthcare applications. Am J Bioeth 2020;20(11):7-17. [doi: 10.1080/15265161.2020.1819469] [Medline: 33103967]
50. Agarwal R, Bjarnadottir M, Rhue L, Dugas M, Crowley K, Clark J, et al. Addressing algorithmic bias and the perpetuation of health inequities: an AI bias aware framework. Health Policy Technol 2023;12(1):100702. [doi: 10.1016/j.hlpt.2022.100702]
51. Okuzu O, Malaga R, Okereafor K, Amos U, Dosunmu A, Oyeneyin A, et al. Role of digital health insurance management systems in scaling health insurance coverage in low- and middle-income countries: a case study from Nigeria. Front Digit Health 2022;4:1008458 [FREE Full text] [doi: 10.3389/fdgth.2022.1008458] [Medline: 36204711]
52. Babili A, Nsanzimana S, Rwagasore E, Lester RT. SMS-based digital health intervention in Rwanda's home-based care program for remote management of COVID-19 cases and contacts: a qualitative study of sustainability and scalability. Front Digit Health 2022;4:1071790 [FREE Full text] [doi: 10.3389/fdgth.2022.1071790] [Medline: 36714610]
53. Doekhie KD, Buljac-Samardzic M, Strating MMH, Paauwe J. Elderly patients' decision-making embedded in the social context: a mixed-method analysis of subjective norms and social support. BMC Geriatr 2020;20(1):53 [FREE Full text] [doi: 10.1186/s12877-020-1458-7] [Medline: 32050911]

Abbreviations

AI: artificial intelligence
AI-PDA: artificial intelligence–enabled patient decision aid
DHEF: Digital Health Equity Framework
HCP: health care provider
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SDM: shared decision-making

XSL•FO
RenderX

Edited by A Stone; submitted 05.Dec.2025; peer-reviewed by S Kakraba, M Veras; comments to author 18.Mar.2026; revised version received 03.Jun.2026; accepted 03.Jun.2026; published 29.Jun.2026.

Please cite as:
Tian CY, Yang X, Wang K, Cheung AWL, Ma JCH, Lu C, Yu JCY, Chan CY, Chen J, Ouyang K, Lin IWK, Pang THC, Zhao S, Wang Y, Wong ELY
Operationalizing Digital Health Equity in Artificial Intelligence–Enabled Patient Decision Aids for Older Adults: Mixed Methods Study
J Med Internet Res 2026;28:e89011
URL: https://www.jmir.org/2026/1/e89011
doi:10.2196/89011
PMID:

©Cindy Yue Tian, Xiaochen Yang, Kailu Wang, Annie Wai-Ling Cheung, Jonathan Chun-Hei Ma, Canjie Lu, Jasmine Cheuk-Ying Yu, Crystal Ying Chan, Jiamin Chen, Kun Ouyang, Ivan Wai-Kiu Lin, Tim Hung-Cheong Pang, Shi Zhao, Yingwei Wang, Eliza Lai-Yi Wong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

An Internet-Based Mindfulness- and Compassion-Based Intercare Program for Reducing Parental Burnout: Randomized Controlled Trial

Francisco Javier Villalón López[1,2], MD; Maximiliano Escaffi-Schwarz[3], PhD

[1]Centro de Educación Médica y Simulación Clínica (CEMSIC), Faculty of Medicine, Universidad Diego Portales, avenida Ejercito 141, Santiago, Santiago Metropolitan, Chile

[2]Programa de Postgrado en Psiquiatría y Salud Mental, Complejo Asistencial Sótero del Río, Faculty of Medicine, Universidad Diego Portales, Santiago, Santiago Metropolitan, Chile

[3]Facultad de Administración y Economía, Universidad Diego Portales, Santiago, Santiago Metropolitan, Chile

Corresponding Author:
Francisco Javier Villalón López, MD
Centro de Educación Médica y Simulación Clínica (CEMSIC), Faculty of Medicine, Universidad Diego Portales, avenida Ejercito 141, Santiago, Santiago Metropolitan, Chile

Abstract

Background: Parental burnout is an underrecognized syndrome characterized by emotional exhaustion, detachment from children, and reduced parental efficacy. It is associated with sleep disturbance, addictive behaviors, suicidal ideation, and increased risk of child neglect and family conflict. Despite its public health relevance, evidence-based interventions remain limited, particularly in low- and middle-income contexts.

Objective: This study aims to evaluate the efficacy and safety of a mindfulness- and compassion-based intercare program for parental burnout (IBAP-BP)—designed to reduce burnout symptoms in teleworking mothers.

Methods: A 3-arm randomized controlled trial (IBAP-BP, active control, and waitlist) was conducted across Chile (December 2022-March 2023) with a 9-month follow-up. Participants (N=593) were women aged 18 years or older, teleworking ≥1 day/week, and living with at least 1 child. Exclusion criteria were self-reported severe psychiatric disorders. Randomization was computer-generated and centrally concealed; data analysts were blinded. The IBAP-BP group attended 8 weekly 2-hour internet-based sessions plus daily home practice integrating mindfulness and compassion. The active control performed relaxation and reflective journaling matched for duration and structure. The primary outcome was parental burnout at 9 months measured using the average-item score of the Parental Burnout Assessment; secondary outcomes were mindfulness, balance of risks/resources, and adverse effects. Modified intention-to-treat analyses and multilevel structural models assessed effects over time.

Results: Of 593 randomized participants, 343 contributed at least 1 postbaseline assessment and comprised the modified intention-to-treat sample. At 9 months, the prespecified primary analysis showed greater reductions in parental burnout for IBAP-BP compared with the waitlist (mean difference=0.62, 95% CI 0.09-1.14; Cohen $d \approx 0.6$, 95% CI 0.41-0.77). No significant difference was found between IBAP-BP and the active control, which showed transient improvements up to 3 months. Effects remained robust in sensitivity analyses. Self-reported adverse events were rare and mild across IBAP-BP and active control. Mediation analyses showed inconsistent associations between mindfulness facets and outcomes.

Conclusions: The culturally adapted, internet-based IBAP-BP program was feasible, safe, and effective compared with the waitlist condition for reducing parental burnout in working mothers, with effects sustained over 9 months, although it did not demonstrate superiority over the active control at the primary end point. Internet-based mindfulness- and compassion-based interventions may offer scalable, equitable options for preventing and treating parental burnout in resource-limited settings.

Trial Registration: ClinicalTrials.gov NCT05833269; https://clinicaltrials.gov/study/NCT05833269

(J Med Internet Res 2026;28:e87416) doi:10.2196/87416

KEYWORDS

burnout; parents; psychology; mindfulness; compassion; teleworking; mothers; randomized controlled trials as topic; internet-based intervention; mental health; developing countries; psychological

XSL•FO
RenderX

Introduction

Parental burnout (PB) is a critical yet underrecognized condition, affecting an estimated 0.3% to 8.9% of parents globally. This syndrome, characterized by emotional exhaustion, detachment from children, lack of fulfillment in their role as a parent, and diminished parenting efficacy [1], has profound implications for individual health and family well-being. PB is associated with increased somatic complaints, sleep disturbance, addictive behavior, and suicidal ideation. Furthermore, it has been linked to higher couple conflicts and higher risk of child neglect and violence [2,3]. Therefore, interventions to reduce PB are urgent.

Mindfulness-based interventions (MBIs) have shown promising effects in reducing PB [4-6]. Mindfulness refers to practices and processes that foster present-moment awareness with curiosity and acceptance [7,8], enhancing attention, emotional regulation, and self-regulation, all of which are key mechanisms in stress reduction [7,9]. Providing preliminary support for the effects of mindfulness in reducing PB levels, the Bayot et al [10] trial (n=76) compared a mindfulness–self-compassion program to an active control (AC) and showed reductions in PB in both groups, but with a stronger reduction in the mindfulness–self-compassion group (η^2=0.45). Furthermore, the 8-week informal mindfulness intervention proposed by Urbanowicz et al [11] (n=30) led to large within-group reductions in PB (η^2=0.24).

However, there remain important practical and methodological limitations. MBIs are typically time-intensive, which may be challenging for parents already experiencing a high demand-resource imbalance [12,13]. Most studies have been conducted in high-income settings, with limited cultural adaptation and accessibility, particularly in contexts where commuting and economic barriers constrain participation [10,11,14,15]. Finally, in addition to the above practical issues, a Cochrane review of 11 trials (n=2118) found low to very low-certainty evidence for MBIs, stressing the need for further research [6]. Existing trials, including Bayot et al [10] and Urbanowicz et al [11], provide preliminary support but are limited by small samples, short follow-up, and methodological constraints (eg, lack of preregistration or inactive control groups), restricting causal inference.

To address the above issues, we then examine whether an adapted internet-based mindfulness- and compassion-based intercare program (Intercare for Parental Burnout [IBAP]) can reduce PB levels in mothers who telework who are particularly at risk of experiencing PB [2,16]. The IBAP, which was developed in Chile [17], was developed with cultural and modular components, flexible delivery, accelerated instructor training, and an emphasis on belonging through intercare. Prior studies have shown its potential to reduce burnout in physicians during COVID-19 [18] and depressive symptoms in students [19]. IBAP targets key factors linked to PB, including perfectionism, emotion regulation, and social support [13]. To provide a more robust examination of the potential benefits of the IBAP, we implemented a 3-arm randomized controlled trial (RCT) with pre-post measurements and a 9-month follow-up.

The aim of this study is to evaluate the efficacy, feasibility, and safety of an internet-based mindfulness- and compassion-based intercare program for parental burnout (IBAP-BP) in reducing PB among teleworking mothers through a 3-arm RCT (IBAP-BP, active control [AC], and waitlist [WL]). Secondary aims are to examine potential mediators (mindfulness and parental risk-resource balance) and to assess acceptability and adverse effects.

Methods

Design

A 3-arm randomized, parallel-group superiority trial with blinded data analysts was conducted in community settings across Chile. Participants were randomly assigned to one of three arms: (1) the mindfulness- and compassion-based intercare program adapted to reduce PB (IBAP-BP), (2) an AC group, and (3) a WL group.

Participant enrollment took place between December 22, 2022, and March 8, 2023. The trial was retrospectively registered at ClinicalTrials.gov (NCT05833269) on April 15, 2023, after enrollment had commenced but prior to the primary outcome assessment period.

Due to an administrative oversight, the third arm was not included in the initial registration and was later added through a registry update on August 15, 2023. All study arms and outcomes had been prespecified in the ethics-approved protocol prior to the start of enrollment, and no primary or secondary outcomes were modified after registration.

A detailed reconciliation of the originally approved design, subsequent amendments, timing of changes, and any protocol deviations is provided in Multimedia Appendix 1 (the *Trial Registration Acknowledgement* section). Reporting adheres to CONSORT (Consolidated Standards of Reporting Trials) (Checklist 1), CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth), and TIDieR (Template for Intervention Description and Replication) (Checklist 2) guidelines.

Participants

Eligible participants were women aged 18 years or older, teleworking ≥1 day/week, cohabiting with at least 1 child; criteria reflected funder priorities and prior evidence, suggesting telework as a potential risk context for PB due to work-family boundary blurring and increased role demands [2,16]. There were no treatment restrictions. Exclusion criteria included self-reported severe depression, substance abuse, or psychosis.

Participants were recruited in Chile between December 22, 2022, and March 8, 2023, nationwide through social media advertising and institutional partnerships coordinated by the Chilean Occupational Safety Institute.

Randomization and Blinding

Participants were centrally randomized by an independent researcher using a computer-generated sequence, with allocation concealed until assignment. Stratification was performed based

XSL•FO
RenderX

on baseline PB severity (PBA ≥84), and no blocking was applied.

Due to logistical constraints related to scheduling group sessions, the intervention arms (IBAP-BP and AC) had fixed capacities (n=125 each), whereas the WL group had no capacity restrictions. Randomization was therefore implemented across the full sample, with assignment to intervention arms limited by the available capacity and the remaining participants allocated to the WL group.

Participants were required to confirm their availability before the disclosure of allocation. A subset of participants withdrew prior to the disclosure of allocation and before confirming their participation. These vacant intervention slots were filled with participants from the WL pool, following the original randomization sequence, to ensure the preservation of allocation procedures.

After the 9-month assessment, participants in the AC and WL groups were offered IBAP-BP in an open-label phase. Full randomization and scheduling procedures are detailed in Multimedia Appendix 1.

Procedure

The IBAP-BP intervention consisted of 8 weekly 2-hour therapist-guided, internet-delivered group sessions led by trained instructors, complemented by 20 to 30 minutes of daily home practice incorporating guided meditation and structured journaling. The program was standardized across 5 cohorts of 25 participants each, with no individual tailoring. Participants received weekly digital materials containing session content, practice reminders, and audio recordings. The intervention content was fixed during the trial and did not undergo modifications.

IBAP-BP was adapted from a validated mindfulness- and compassion-based program previously applied in diverse populations [17-19]. It targeted key risk factors for PB—perfectionism, limited emotion regulation, and low social support—while fostering self-compassion, emotional regulation, and effective parenting practices [13]. The intervention followed a structured manual ensuring uniform delivery across cohorts. Platform, connectivity, and communication details are provided in Multimedia Appendix 1, and the full module content and practices are summarized in Table S13 in Multimedia Appendix 1.

The AC was a self-guided, home-based program matched in structure and duration. Participants practiced approximately 20 to 30 minutes per day for 8 weeks, combining progressive muscle relaxation with structured journaling and psychoeducational content addressing parental stress, perfectionism, coparenting, adaptive coping, and enjoyment in parenting. Weekly materials and reminders were distributed electronically, and a group chat was available for optional peer support (file formats and delivery logistics are reported in Multimedia Appendix 1).

Adherence in the IBAP-BP group was assessed through session attendance and self-reported home practice, whereas in the AC group, only self-reported practice frequency and duration were collected, given the absence of platform-based tracking.

The WL group received no intervention during the initial 9-month period but was offered IBAP-BP afterward. No restrictions were imposed regarding participation in other interventions during this time.

Data collection and intervention delivery were conducted through secure platforms (Zoom Pro [Zoom Video Communications], WhatsApp [Meta Platforms], and Qualtrics), and participant confidentiality was ensured.

Instructors

Each group was led by an instructor, supported by an assistant managing the internet-based platform. The team comprised 2 psychologists, 2 psychiatrists, and 1 physician. Four instructors had prior training as mindfulness teachers: 1 in mindfulness-based stress reduction (MBSR), 1 in both MBSR and mindfulness-based cognitive therapy (MBCT), and 2 in MBSR, MBCT, and compassion-based programs. One facilitator did not have formal mindfulness teacher training but had completed an 8-week MBCT program as a participant. This facilitator underwent 6 months of preparation for delivering IBAP, following the Mindfulness-Based Intervention Teacher Assessment Criteria [20]. All instructors completed 8 weekly 2-hour sessions of training on IBAP-BP. All materials, including session guides, content, and audio instructions, were standardized and provided by an external team to ensure consistency across groups.

Data Collection and Follow-Up

Participants completed internet-based surveys at baseline (preintervention), 3 (end of intervention), 6, and 9 months, with follow-ups extended to 12, 18, and 24 months during an open-label phase after the WL and AC groups were offered. IBAP-BP completion rates were enhanced through email reminders and telephone check-ins conducted by independent assistants. Surveys gathered demographic information (eg, gender, age, educational level, occupation, job role, income range, workplace hierarchy, marital status, and number and gender of children) and assessed primary and secondary outcomes.

The primary outcome was the *Parental Burnout Assessment* (PBA) total score at 9 months, validated in Chile. This 23-item scale measures 4 dimensions: emotional exhaustion, loss of pleasure in the parental role, emotional distancing from children, and self-perception as a parent. Raw scores range from 0 to 138. However, analyses and tables report the average-item metric (range 0 - 6), with higher scores indicating greater PB. The use of the average score, rather than the total, avoids the assumption that missing values are equivalent to zero, providing a more accurate representation [21].

Secondary outcomes assessed at 3, 6, or 9 months and during the open phase at 12, 18, and 24 months included PBA, the Five Facets of Mindfulness Questionnaire–15 items [22], and Balance Between Risks and Resources [13]. Self-reported adverse experiences were assessed at 3-, 6-, and 9-month follow-ups using the Unwanted Effects of Meditation Checklist in the

XSL•FO
RenderX

IBAP-BP and AC groups only; adverse effects were not assessed in the WL group [23]. Participants reporting significant distress were advised to seek clinical support, and contact information for mental health services was provided. No formal stopping rules were required, given the low-intensity nature of the intervention. In addition, the self-reported number of home practice days and minutes per week, as well as session attendance, were recorded. Other outcomes will be published as secondary data in a separate article (see Multimedia Appendix 1 for details).

Statistical Analysis

Sample size, power calculation, reliability, and assumption checks are detailed in Multimedia Appendix 1.

Due to scheduling constraints inherent to group-based delivery, a subset of randomized participants withdrew prior to allocation disclosure and before confirming their participation schedule. These individuals were not followed longitudinally and were excluded from the outcome dataset. Accordingly, analyses were conducted using a modified intention-to-treat (mITT) approach, defined a priori as including all randomized participants who did not withdraw before confirming participation and who provided at least 1 postbaseline assessment. Baseline characteristics of included and excluded participants were compared to assess potential bias.

To mitigate potential bias, baseline characteristics of excluded and included participants were compared, and sensitivity analyses were conducted. No interim analyses were prespecified.

The primary end point was PB (PBA total score) at 9 months after the baseline measurement (T3), corresponding to 6 months post intervention. The primary confirmatory analysis consisted of a between-group comparison at 9 months using one-way ANOVA under the mITT framework. To conduct the post hoc analyses, we used the Games-Howell test, rather than the Tukey test, as it is robust to unequal variance and unequal sample size between groups [24]. Effect sizes are reported as mean differences with 95% CI and Cohen *d*.

Missing data in the primary analysis were handled using complete-case data within the mITT sample. Accordingly, the operational primary analysis consisted of a complete-case analysis within the mITT sample, whereas multiple imputation and mixed-effects repeated-measures models were conducted as sensitivity analyses to assess robustness.

Secondary analyses examined between-group differences at 3 (T1) and 6 months (T2), as well as PBA subdimensions, mindfulness, parental risk-resource balance, adverse effects, and adherence-related measures. Sensitivity analyses included (1) a per-protocol analysis including participants in the mindfulness group who attended ≥50% of sessions; (2) an assessment of baseline differences related to follow-up nonresponse; (3) a repeated-measures analysis with imputed data using a generalized mixed model with random coefficients; and (4) a subgroup comparison based on baseline severity, using a cutoff score of ≥84 on the PBA to distinguish between participants classified as treatment (PBA≥84) and prevention (PBA<84) groups.

Exploratory analyses included multilevel structural equation modeling to assess longitudinal intervention effects and cross-lagged panel models to examine potential mediation pathways [25-27]. These analyses were hypothesis-generating and should be interpreted with caution.

Statistical significance was set at $P<.05$. The 9-month PB comparison constituted the prespecified primary confirmatory end point and was therefore not subjected to multiplicity correction. Holm-Bonferroni correction was applied exclusively to secondary and exploratory outcome families to control for type I error [28]. All analyses were conducted in R (R Foundation, v4.3.1). No analysis plan was prospectively preregistered; however, the reported statistics adhere to the recommendations of SAP (Statistical Analysis Plan) guidelines for RCT [29]. Details are provided in Multimedia Appendix 1.

Ethical Considerations

Ethical approval for this study was granted by the Diego Portales University Ethics Committee on October 27, 2022 (Code 38 - 2022). An amendment was subsequently approved on December 20, 2022, to incorporate a third study arm. All participants provided electronic informed consent before enrollment after receiving detailed information regarding the study objectives, procedures, potential risks and benefits, voluntary participation, and the right to withdraw at any time without consequences. Participant privacy and confidentiality were protected throughout the study. Data were collected through secured password-protected platforms, stored in deidentified form, and accessible only to authorized members of the research team. No personally identifiable information was included in the analyses or publication materials. Participants did not receive financial compensation for participation in the study.

Results

Baseline Characteristics

A total of 665 participants were enrolled between December 22, 2022, and March 8, 2023. After exclusions and withdrawals, 593 participants were randomized. Of these 593 participants, we found that 39% (n=231) of them presented with high levels of PB (84 points or above), while 61% (n=362) reported low levels of PB (<84 points).

Prior to allocation disclosure, 65 participants withdrew due to scheduling conflicts (37 in the IBAP-BP group and 28 in the AC group) and were excluded from the mITT sample. Their slots were dynamically reallocated using participants from the WL group, following the original randomization sequence. The final randomized allocation included 161 participants in the IBAP-BP group (124 initiated the intervention), 157 participants in the AC group (129 initiated the intervention), and 275 participants in the WL group. The mITT analytical sample (n=343) therefore comprised participants who confirmed scheduling after randomization and provided at least 1 postbaseline assessment. A detailed participant flow is provided in the CONSORT diagram (Figure 1).

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram. mITT: modified intention-to-treat; PP: per-protocol.

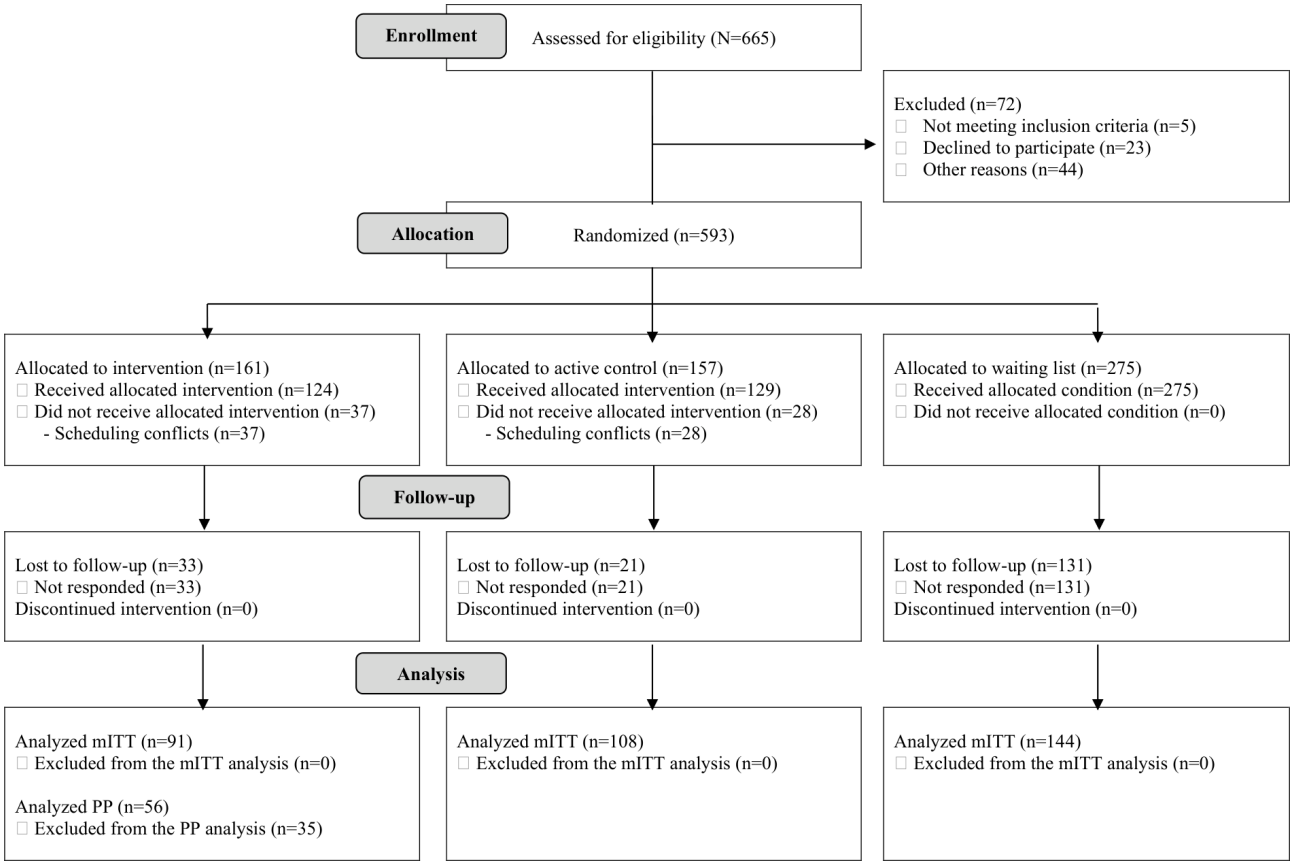

Participants had a mean age of 38 (SD 5.28) years. Baseline characteristics (age, number of children, workdays, education, and relationship status) were balanced across groups (all $P>.05$ [0.07-0.69]; Table 1), although none followed a normal distribution (all $P<.001$; Table S1 in Multimedia Appendix 1).

XSL•FO
RenderX

Table . Demographic characteristics of participants in the total sample and by study group: mindfulness- and compassion-based intercare program for parental burnout (IBAP-BP), active control, and waiting list[a].

Variable	Total sample (N=343)	IBAP-BP (n=91)	Active control (n=108)	Waiting list (n=144)
Age (y), mean (SD)	38.84 (5.28)	38.95 (6.38)	39.18 (4.66)	38.55 (5.01)
Number of children living at home, mean (SD)	1.65 (0.77)	1.73 (0.76)	1.65 (0.78)	1.61 (0.78)
Average age of children (y), mean (SD)	4.96 (4.45)	5.32 (4.85)	4.8 (3.89)	4.86 (4.6)
Number of workdays, mean (SD)	5.01 (0.67)	4.99 (0.69)	5.06 (0.59)	4.99 (0.71)
Number of days working from home, mean (SD)	3.67 (1.46)	3.49 (1.44)	3.53 (1.58)	3.9 (1.36)
Share of child caring responsibilities, mean (SD)	78.72 (16.41)	78.67 (17.42)	79.46 (14.96)	78.19 (16.89)
Educational level, n/N (%)				
Complete school education	7/343 (2%)	2/91 (2%)	1/108 (1%)	4/144 (3%)
Professional education	41/343 (12%)	13/91 (14%)	13/108 (12%)	15/144 (10%)
Bachelor's degree	164/343 (48%)	32/91 (35%)	50/108 (46%)	82/144 (57%)
Postgraduate degree (eg, MSc)	131/343 (38%)	44/91 (48%)	44/108 (41%)	43/144 (30%)
Relationship status, n/N (%)				
Married	166/343 (48%)	35/91 (38%)	59/108 (55%)	72/144 (50%)
In partnership, but not married	89/343 (26%)	28/91 (31%)	18/108 (17%)	43/144 (30%)
Widow	0/343 (0%)	0/91 (0%)	0/108 (0%)	0/144 (0%)
Divorced	24/343 (7%)	7/91 (8%)	9/108 (8%)	8/144 (6%)
Single	60/343 (17%)	19/91 (21%)	22/108 (20%)	19/144 (13%)
Prefer not to say	2/343 (1%)	2/91 (2%)	0/108 (0%)	2/144 (1%)
Parental burnout, mean (SD)				
Total score	3.38 (1.28)	3.31 (1.39)	3.47 (1.27)	3.37 (1.21)
Parental exhaustion	4.29 (1.39)	4.15 (1.49)	4.36 (1.39)	4.34 (1.33)
Parental identity conflict	2.8 (1.47)	2.84 (1.55)	2.94 (1.49)	2.67 (1.39)
Parental fed up	2.73 (1.38)	2.64 (1.41)	2.77 (1.46)	2.76 (1.31)
Parental emotional distancing	2.91 (1.48)	2.81 (1.59)	3.03 (1.44)	2.89 (1.44)

[a]Data include age, number of children, child age, workdays, remote workdays, child-caring responsibilities, educational attainment, and relationship status. Values are presented as mean (SD) for continuous variables and percentage for categorical variables. Parental burnout values are presented using the Parental Burnout Assessment (PBA) average-item metric (range 0 - 6) rather than the raw total score (0 - 138).

Primary Outcome

The mITT analysis showed a significant between-group difference in PBA scores at 9 months among participants who completed the 9-month follow-up assessment (n=203; IBAP-BP: n=66, AC: n=73, WL: n=64; ANOVA: $F_{(2, 200)}$=4.01, P=.02). Post hoc analysis using the Games-Howell test revealed a significant reduction for IBAP-BP versus WL (mean difference 0.62, 95% CI 0.09-1.14), but no significant differences were found between IBAP-BP and AC groups (mean difference −0.25, 95% CI −0.75 to 0.25) or between AC and WL groups (mean difference 0.37, 95% CI −0.14 to 0.88). Reductions from baseline to 9 months were moderate and significant for IBAP-BP (Cohen d=0.67, 95% CI 0.40-0.93) and AC (d=0.75, 95% CI 0.49-1.01), and they were small and nonsignificant for WL (d=0.18, 95% CI −0.07 to 0.41) (Figure 2; Tables S2–S4 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 2. Mean trajectories of parental burnout average-item scores (PBA range 0 - 6) per group. Asterisks represent the post hoc Games-Howell test between each intervention group and the waiting list mean level at each wave. Lines represent the CIs. *P<.05, ** P<.01. IBAP-BP: mindfulness- and compassion-based intercare program for parental burnout; PBA: Parental Burnout Assessment.

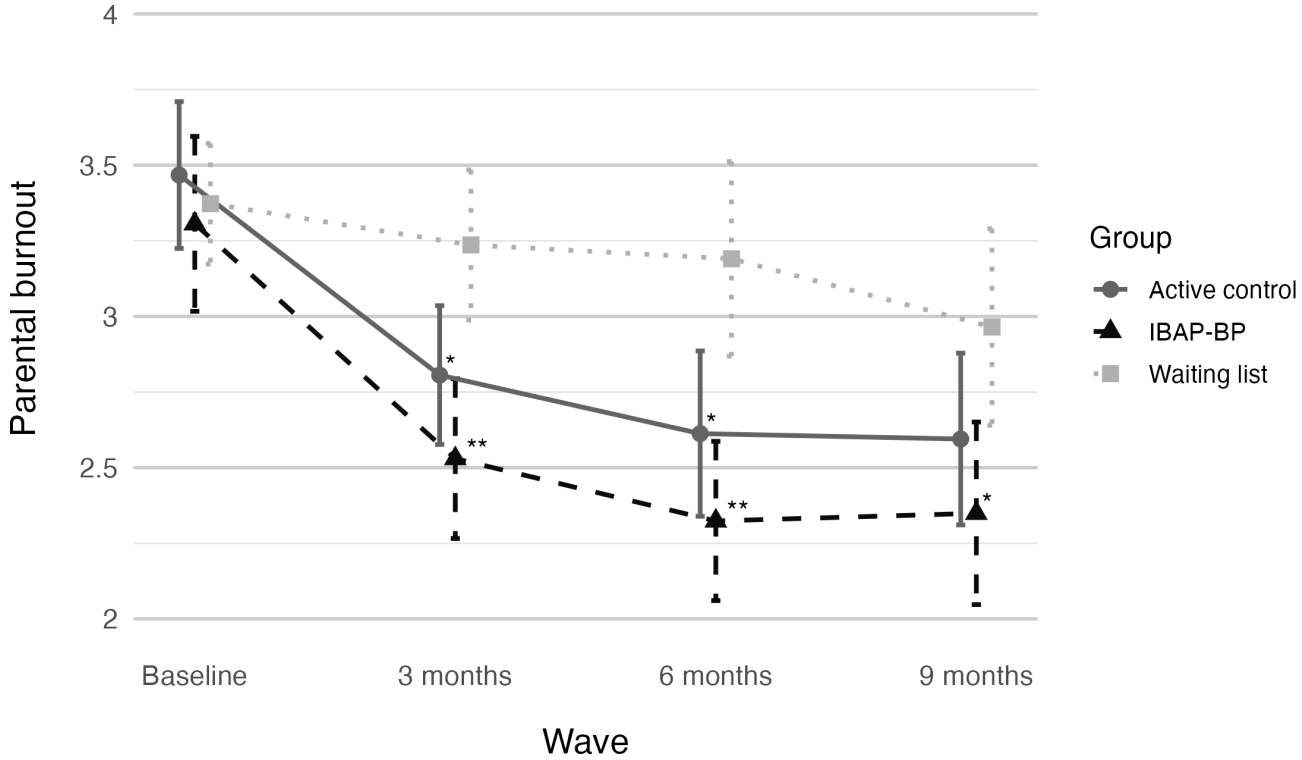

Attrition analysis showed no baseline PBA differences between completers and dropouts in IBAP-BP and WL (P=.96 and P=.50 respectively), indicating random losses. In contrast, AC dropouts had lower baseline PBA (P=.03), possibly underestimating the effect. No significant differences were observed by age, education, or other baseline sociodemographic variables (P>.05 [P value ranges between 0.22 and 0.99]).

Multilevel structural equation modeling for the whole mITT sample confirmed the benefits of both interventions over time (N=343; IBAP-BP: n=91; AC: n=108; WL: n=144; IBAP-BP: b=−0.20, P<.001; AC: b=−0.24, P<.001). Furthermore, we found a negative main effect of time on PBA (b=−0.09, P=.01) after controlling for the main effects of both interventions over time, suggesting gradual, but lesser improvement in the WL group (Table S9 in Multimedia Appendix 1). Sensitivity analyses, including per-protocol and multiple imputation for handling missing data, yielded results consistent with the ones already presented (Tables S6-S10 in Multimedia Appendix 1). When examining the sample who complied with the IBAP-BP intervention protocol (>50% of attendance to the sessions), the results show significant differences between the IBAP-BP group and the WL group (mean difference 0.73, 95% CI 0.21-1.26) at 9 months. Furthermore, when using imputed data for handling missing data, we find support for the benefits of the IBAP-BP group (b=0.12, P=.04), such that there is a significant decrease of PBA over time. We did not find a significant effect of the AC on PBA over time (b=−0.11, P=.06).

When the sample was divided according to PB levels (231/593, 39% with high PB vs 362/593, 61% without), the intervention showed significant effects only in the group with lower burnout. Specifically, IBAP-BP participants reported greater reductions than those in the WL group (mean difference 0.795, 95% CI 0.275-1.31]), while no significant differences emerged between IBAP-BP and AC (mean difference −0.467, 95% CI −0.939 to 0.004) or between AC and WL (mean difference 0.328, 95% CI −0.261 to 0.916). It is worth noting, however, that reductions from baseline to 9 months were moderate and significant for IBAP-BP in the high-burnout group (Cohen d=0.88, 95% CI 0.39-1.35), and these improvements were comparable to those observed in the lower-burnout group (Cohen d=0.74, 95% CI 0.40-1.07).

Secondary Outcome

Significant group differences in PBA scores were detected at 3 months (ANOVA: $F_{2,289}$=8.03, P<.001) and 6 months ($F_{2,221}$=8.57, P<.001). Post hoc analysis demonstrated significant improvements in PBA scores for IBAP-BP (third month: mean difference 0.71, 95% CI 0.28-1.14; sixth month: mean difference 0.87, 95% CI 0.37-1.36) and AC (third month: mean difference 0.43, 95% CI 0.03-1.08; sixth month: mean difference 0.58, 95% CI 0.08-1.08) compared to WL at both time points. These findings were consistent in per-protocol analyses (see Table S7 in Multimedia Appendix 1).

Subdimension analyses revealed that IBAP-BP significantly reduced parental exhaustion and emotional distancing at 3, 6, and 9 months compared to WL (third month: mean difference 0.78, 95% CI 0.27-1.29; sixth month: mean difference 1.06, 95% CI 0.49-1.63; ninth month: mean difference 0.74, 95% CI 0.14-1.35). When comparing the IBAP-BP to the WL group, improvements in parental identity conflict and “Fed up” scores were observed at 3 and 6 months (parental identity conflict—third month: mean difference 0.61, 95% CI 0.13-1.08; sixth month: mean difference 0.73, 95% CI 0.18-1.29 and “Fed

XSL•FO
RenderX

up”—third month: mean difference 0.64, 95% CI 0.21-1.07; sixth month: mean difference 0.76, 95% CI 0.30-1.22), but these effects were not sustained at 9 months. Only IBAP-BP maintained significant differences across all subdimensions at 9 months in the per-protocol analysis (see Table S7 in Multimedia Appendix 1). Following the Holm-Bonferroni correction applied to secondary outcome families, differences in secondary subdimension analyses at 9 months were no longer statistically significant.

Safety Analysis and Adherence

Adverse experiences were assessed at the 3-, 6-, and 9-month follow-up waves in the 2 active intervention groups. The intervention demonstrated a favorable safety profile. Self-reported adverse event rates were low and comparable across groups: 0.0%, 1.4%, and 1.5% for IBAP-BP at 3, 6, and 9 months, respectively, compared to 3.6%, 2.5%, and 2.7% for AC. No severe adverse events were reported (P=.62, Table 2).

XSL•FO
RenderX

Table . Self-reported adverse experiences assessed at 3-, 6-, and 9-month follow-up[a].

Variable	IBAP-BP[b]			Active control		
	3rd month (n=80), n (%)	6th month (n=71), n (%)	9th month (n=66), n (%)	3rd month (n=102), n (%)	6th month (n=79), n (%)	9th month (n=73), n (%)
Reported adverse effects	—[c]	1 (1.4)	1 (1.5)	3 (2.9)	2 (2.5)	2 (2.7)
Greater emotional pain	—	1 (1.4)	1 (1.5)	—	1 (1.3)	—
Increased self-criticism	—	—	1 (1.5)	1 (1)	—	1 (1.4)
Greater fear/anxiety/depression	—	1 (1.4)	1 (1.5)	—	2 (2.5)	2 (2.7)
Lack of direction in life	—	1 (1.4)	1 (1.5)	—	—	1 (1.4)
Lower motivation in life	—	1 (1.4)	—	—	—	—
Feeling of needing something more, of lacking something, etc	—	1 (1.4)	1 (1.5)	1 (1)	1 (1.3)	1 (1.4)
Increased mental confusion	—	1 (1.4)	1 (1.5)	—	1 (1.3)	—
Lack of interest in your surroundings	—	—	—	1 (1)	—	—
Boredom	—	—	—	—	—	—
Need to meditate continuously	—	—	—	—	—	—
Feeling that time not spent meditating is wasted	—	—	—	—	—	—
Restlessness/anxiety when not practicing formal meditation	—	1 (1.4)	—	—	1 (1.3)	—
Increased criticism of others	—	—	—	—	—	—
Greater awareness of one's negative traits	—	—	1 (1.5)	1 (1)	—	—
Feeling of being superior to others/better than them	—	—	—	—	—	—
Feelings of boredom caused by people	—	—	—	—	—	—
Feelings of lack of interest in others	—	—	—	1 (1)	—	1 (1.4)
Lack of interest in people's conversations	—	1 (1.4)	—	1 (1)	—	—
Feeling that only people who meditate have value	—	—	—	—	—	—

XSL•FO
RenderX

Variable	IBAP-BP[b]			Active control		
	3rd month (n=80), n (%)	6th month (n=71), n (%)	9th month (n=66), n (%)	3rd month (n=102), n (%)	6th month (n=79), n (%)	9th month (n=73), n (%)
Feeling of being alienated from society	—	—	—	—	—	—
Hypersensitivity/rejection of urban life	—	—	1 (1.5)	1 (1)	—	—
Difficulty feeling comfortable in the world	—	1 (1.4)	1 (1.5)	1 (1)	—	—
Other (eg, body pain when performing the exercises, guilt for not performing the exercises)	—	1 (1.4)	—	1 (1)	2 (2.5)	—

[a]The data reflect the number of participants reporting adverse effects across 3 assessment waves in the IBAP-BP and active control groups. Values represent percentages of participants reporting each self-reported adverse experience at each assessment wave. Adverse experiences were assessed only in the IBAP-BP and active control groups at 3-, 6-, and 9-month follow-up.

[b]IBAP-BP: mindfulness- and compassion-based intercare program for parental burnout.

[c]Not applicable.

Assistant attendance of the 124 participants enrolled in the IBAP-BP group averaged 3.48 (SD 2.93) sessions, with a range between 0 and 8 sessions. In addition, 48% (n=59) of participants completed more than 50% of the sessions, and 10% (n=13) completed 100% of the program. Self-reported adherence to home practice averaged 2.92 (SD 2.02) days per week and 28.51 (SD 20.38) minutes per week in the IBAP-BP group immediately after the intervention, 1.83 (SD 1.73) days per week and 21.78 (SD 19.81) minutes per week at 6 months, and 1.97 (SD 1.98) days per week and 22.09 (SD 25.69) minutes per week at 9 months. In turn, for the AC, the self-reported practice averaged 2.11 (SD 1.90) days per week and 31.06 (SD 59.49) minutes per week immediately after the intervention, 1.88 (SD 1.97) days per week and 24.53 (SD 28.67) minutes per week at 6 months, and 1.36 (SD 1.53) days per week and 16.75 (SD 17.44) minutes per week at 9 months. In both groups, we observe a steady decline of practice at home over time.

Exploratory Mediation Analysis

We tested 2 classes of mediators: parental risks or resources and the 5 mindfulness facets. IBAP-BP significantly increased common and specific parental resources (b=0.50, P=.03 and b=0.52, P<.01 respectively). In contrast, AC had no effect on either common (b=0.20, P=.93) or specific resources (b=0.33, P=.10). However, neither changes in common nor specific resources were associated with changes in PBA scores, suggesting that they did not mediate the intervention effects (Table S11 in Multimedia Appendix 1).

Regarding the 5 facets of mindfulness, IBAP-BP increased the describe (b=0.24, P=.04) and observe (b=0.30, P=.008) dimensions and reduced present-moment awareness (b=−0.39, P<.002). AC increased describe (b=0.24, P=.03) and reduced awareness (b=−0.31, P=.01). Nonjudgment postintervention predicted higher PBA scores at 3 months (b=0.16, P=.02) and observe at 3 months predicted higher PBA at 9 months (b=0.34, P=.01). The only potential mediation found was an indirect effect of IBAP-BP on PBA reduction at 9 months via postintervention increase in observe—though in the opposite direction of expectations (Table S12 in Multimedia Appendix 1).

Discussion

Main Result

This study evaluated an internet-based mindfulness- and compassion-based intervention for PB against AC and WL groups. IBAP-BP showed moderate-to-large improvements versus the WL at 3, 6, and 9 months, but not consistent superiority over the AC. The AC—comprising relaxation, journaling, psychoeducation, and group support—was associated with reductions in burnout up to 3 months after the intervention. However, these effects were not sustained at a longer 9-month follow-up, suggesting a potentially low-cost and scalable, but transient, alternative.

Despite an expected 50% attrition rate, sensitivity analyses confirmed the robustness of the results. Effect sizes increased in per-protocol and imputed analyses, and no systematic attrition patterns were observed. Self-reported adverse events were rare, mild, and comparable across IBAP-BP and AC.

Exploratory mediation analysis showed increased mindfulness among IBAP-BP participants, but effects were inconsistent and inversely related to the nonjudgment and observe facets. No mediation effects were found for parental risks or resources.

Previous Evidence

Our findings align with prior research demonstrating the effectiveness of mindfulness- and compassion-based interventions in reducing PB. Meta-analyses of related constructs, such as parental stress, have yielded comparable

XSL•FO
RenderX

results. Notably, less structured approaches such as those used in our AC group have also shown utility in stress reduction [4,10,14].

This study builds on previous trials by addressing methodological limitations, including the absence of AC and WL, limited follow-up durations (<3 mo), and self-enrollment randomization procedures [10,14,15]. The AC group showed short-term benefits up to 3 months, which diminished over time, contrasting with the sustained effects of IBAP-BP. The inclusion of a WL group clarified the natural progression of PB, revealing modest improvements even without intervention.

To enhance internal validity, we implemented centralized randomization with concealed allocation, reducing the potential for selection bias. In contrast, earlier studies that relied on self-enrollment or site-based randomization may have introduced uncertainty in sequence generation and allocation concealment, which can modestly inflate effect sizes by 7%10% [30,31].

Attrition rates (30% - 50%) were similar to those reported in comparable internet-based interventions [18,19]. A streamlined single-stage recruitment process facilitated accessibility. Among participants who initiated the allocated intervention, retention defined as contributing at least 1 follow-up assessment was 83.7% in the AC group and 73.3% in the IBAP-BP group. These rates exceed those typically observed in multistage recruitment designs and are consistent with prior studies, possibly reflecting proactive follow-up procedures (email and telephone reminders) [10,14,32]. Notably, attrition in the AC group was associated with lower baseline burnout, suggesting potential differential dropout by symptom severity, which may have biased estimates toward greater apparent effectiveness in this group.

The low rate and mild nature of adverse events are consistent with prior literature on MBIs. Transient experiences, such as increased emotional distress, anxiety, or self-criticism, are commonly reported, particularly during the early stages of practice. These effects are generally short-lived and rarely lead to discontinuation or the need for clinical intervention [23].

The study also introduces innovations in delivery and focus. While prior interventions were primarily face-to-face [10,14,15], our internet-based approach improved accessibility while maintaining efficacy. In Chile, where average commute times are lengthy (44-54 min per trip) and inversely correlated with income (R^2=0.74), internet-based delivery addresses logistical and equity challenges [33].

The AC showed benefits up to 3 months after the intervention, particularly among participants with higher baseline burnout. This suggests that flexible, peer-supported approaches may serve as an initial step within a stepped-care framework. Its components—relaxation, journaling, psychoeducation, and group support—represent established stress-reduction strategies, supporting its role as a therapeutic comparator rather than a neutral control [34-37].

Adherence patterns may have contributed to these differences. IBAP-BP, which combines structured group sessions with home practice, showed more stable engagement, whereas the self-guided AC demonstrated greater variability and decline over time. Prior evidence indicates that strategies such as text reminders and peer coaching can enhance adherence and support sustained engagement [38], which may be particularly relevant for less structured interventions.

The focus on teleworking mothers—100% employed at baseline, compared to 76% to 77% in prior studies—reflects a pragmatic approach aligned with funding priorities. Teleworking mothers may face a higher-risk context for PB. This includes blurred work-family boundaries, increased role conflict, and reduced recovery time—factors consistently linked to emotional exhaustion and parental stress. These conditions may elevate baseline burnout and influence responsiveness to self-regulation–focused interventions [2,16]. While the Latin American socioeconomic context differs from European samples, participants had similar educational levels and marital statuses [10,14]. Although generalizability to lower socioeconomic groups may be limited, effect sizes remained consistent.

IBAP-BP is distinguished by unique components: the “moment of care” priming technique (validated in meta-analyses [39]), compassion exercises for individuals and dyads, habit-formation strategies, and an “inter-care plan” targeting multiple life domains (physical, emotional, social, economic, ethical, and spiritual) [17]. These elements support a comprehensive and sustainable model for reducing PB.

Mechanisms

Both interventions shared core behavioral ingredients—structured daily practice, psychoeducation, and elements of social support—which are established mechanisms for stress reduction and may account for the comparable outcomes between groups [34-37]. Contrary to standard models of MBIs—which emphasize acceptance, present-moment awareness, and decentering—no consistent mediation effects were found across the 5 mindfulness facets [40-42]. Unexpectedly, increases in nonjudgment and observe were associated with higher PB at follow-up. This may reflect measurement limitations, as certain facets (particularly *observe*) may capture heightened awareness of distress rather than adaptive regulation [43]. Alternatively, unmeasured mechanisms, such as compassion, may account for these associations. The 3-month assessment intervals may have been insufficient to capture the temporal dynamics of mediation, potentially obscuring short-term or delayed effects, as previous research suggests that changes in mindfulness may occur over shorter periods. Statistical suppression or multicollinearity among mindfulness facets may have contributed to directionally inconsistent associations [44,45]. Finally, prior studies have not conclusively demonstrated that mindful parenting mediates changes in PB or in the balance of risks and resources [10], and they do not provide clear support for mindfulness facets as mediators in this context.

Limitations

Strengths include its design with AC and WL, high retention, culturally tailored content, and real-world recruitment. This is also the first Latin American study on PB, expanding regional representation. Limitations include a lack of adherence monitoring in the AC group and the inability to blind

XSL•FO
RenderX

participants. Additionally, loss to follow-up may have reduced confidence in the robustness of the results.

The use of an mITT approach represents a key limitation. Participants who withdrew prior to confirming participation scheduling (n=65) were excluded from analyses, which deviate from strict intention-to-treat principles and may introduce selection bias. In particular, this procedure may limit comparability between randomized groups and reduce generalizability, as individuals with greater logistical constraints or lower availability may be underrepresented in the analytical sample. However, baseline characteristics did not differ significantly between excluded and included participants, and sensitivity analyses yielded consistent results, supporting the robustness of the findings.

Importantly, the third study arm was prospectively approved by the ethics committee prior to participant enrollment through a formal protocol amendment. However, the trial was registered after enrollment had begun, and the third arm was not included in the initial registry entry due to an administrative oversight, being added only in a subsequent update. No primary or secondary outcomes reported in this paper were modified following registration, and all analyses were conducted according to the prespecified protocol. A detailed reconciliation of the original protocol, subsequent amendments, and registration timeline is provided in Multimedia Appendix 1. Nevertheless, this deviation from prospective registration standards may affect perceived transparency and introduce a potential risk of selective reporting bias.

The unexpectedly strong response in the AC group, combined with a sample size powered for larger effects (26%, d=0.71), may have limited detection of smaller differences (eg, the observed 10%). Comorbid conditions (eg, depression and posttraumatic stress disorder) or concomitant treatments were not controlled.

Finally, the exclusive inclusion of highly educated, teleworking mothers may limit generalizability. This group may present a distinct burnout profile related to work-family boundary blurring and role overload. Findings may not extend to fathers, unemployed caregivers, or parents of younger children, who may experience different stressors and intervention responses [2,16]. In addition, participation may have been shaped by socioeconomic, digital, and cultural factors, as higher education, income, and digital literacy are associated with greater uptake of digital interventions, while cultural context may influence engagement with mindfulness-based approaches, potentially limiting the inclusion of more vulnerable groups and widening health disparities [46-48].

Future Research

Future studies should use multicenter designs across diverse settings, test different delivery formats (eg, in-person), and assess cost-effectiveness. Shorter or more flexible programs may improve adherence. Flexible, peer-supported, or stepped-care approaches also appear feasible and warrant further investigation. Including fathers and broader employment contexts would enhance generalizability, while more frequent assessments could help elucidate mechanisms of change.

Clinical Implications

The internet-based IBAP-BP offers a scalable, accessible, and effective intervention compared with no-treatment conditions, making it especially valuable in resource-limited settings. Its internet-based format addresses barriers such as facilitator scarcity and geographical constraints. From a public health perspective, the program could enhance family well-being and reduce the burden of PB. However, given the absence of clear superiority over an active comparator, flexible, peer-supported, low-cost options—such as the AC program combining relaxation, journaling, psychoeducation, and group support—may represent clinically meaningful alternatives. A stepped-care approach could help bridge the gap toward more intensive and sustainable programs such as IBAP-BP.

Acknowledgments

ChatGPT-4o was used to assist in improving the readability and language of the paper. The specific prompt used was "If you were the editor of an excellence journal, correct the spelling, grammar, and wording of the following text." The full manuscript, as originally written, was input into the model, and subsequent differences were carefully reviewed and corrected by the authors.

Funding

This study was funded by the Convocatoria de Proyectos de Investigación e Innovación en Prevención de Accidentes y Enfermedades Profesionales 2022 of the Superintendencia de Seguridad Social (Chile) and financed by the Instituto de Seguridad del Trabajo (IST) with resources from the Social Insurance Fund under Law Number 16.744 on Occupational Accidents and Occupational Diseases. The funding source had no role in the study design, implementation, data collection, analysis, interpretation, or manuscript writing. The authors had full access to all data and take full responsibility for the integrity and accuracy of the analyses.

Data Availability

The data collected for this study, including deidentified individual participant data and a data dictionary defining each field in the dataset, will be made available upon reasonable request. Additional documents, such as the study protocol, statistical analysis plan, and informed consent form, will also be accessible. Data will be available upon publication and will remain accessible for 5 years. Researchers interested in accessing the data can request it by contacting francisco.villalon@mail.udp.cl. Data sharing

XSL•FO
RenderX

will be subject to the approval of a research proposal and the signing of a data access agreement. Investigator support may be provided upon request.

Conflicts of Interest

The funds were received by Diego Portales University. Both authors received compensation for the design, implementation, analysis, and publication of this study. FJVL has received payments for delivering lectures, presentations, classes, and workshops on mindfulness and compassion from the institutions mentioned, as well as from individuals.

Editorial Notice

This randomized study was retrospectively registered because participant enrollment had already begun before trial registration due to an administrative oversight. In addition, the third study arm was inadvertently omitted from the initial registry entry and was subsequently incorporated through a registry update. Importantly, the 3-arm design, as well as all primary and secondary outcomes, had already been prespecified in the ethics-approved protocol prior to the initiation of participant enrollment, and no outcomes were modified after registration. The editor granted an exception to the ICMJE (International Committee of Medical Journal Editors) requirement for prospective registration of randomized trials. However, readers are advised to carefully interpret any explicit or implicit claims regarding primary outcomes or intervention effectiveness, as retrospective registration does not preclude the possibility of post hoc outcome modifications.

Multimedia Appendix 1
Supplementary materials, including detailed methods; protocol deviations and trial registration reconciliation; recruitment and randomization procedures; intervention content and delivery; statistical and sensitivity analyses; mediation models; supplementary tables and figures; and additional outcome data.
[DOCX File, 168 KB - jmir_v28i1e87416_app1.docx]

Checklist 1
CONSORT checklist.
[PDF File, 195 KB - jmir_v28i1e87416_app2.pdf]

Checklist 2
TIDieR checklist.
[PDF File, 239 KB - jmir_v28i1e87416_app3.pdf]

References

1. Roskam I, Aguiar J, Akgun E, et al. Parental burnout around the globe: a 42-country study. Affect Sci 2021;2(1):58-79. [doi: 10.1007/s42761-020-00028-4] [Medline: 33758826]
2. Griffith AK. Parental burnout and child maltreatment during the COVID-19 pandemic. J Fam Violence 2022;37(5):725-731. [doi: 10.1007/s10896-020-00172-2] [Medline: 32836736]
3. Mikolajczak M, Brianda ME, Avalosse H, Roskam I. Consequences of parental burnout: its specific effect on child neglect and violence. Child Abuse Negl 2018 Jun;80:134-145. [doi: 10.1016/j.chiabu.2018.03.025] [Medline: 29604504]
4. Burgdorf V, Szabó M, Abbott MJ. The effect of mindfulness interventions for parents on parenting stress and youth psychological outcomes: a systematic review and meta-analysis. Front Psychol 2019;10:1336. [doi: 10.3389/fpsyg.2019.01336] [Medline: 31244732]
5. Turpyn CC, Fischer S, et al. Parenting-focused mindfulness intervention reduces stress and improves parenting in highly-stressed mothers of adolescents. Mindfulness (N Y) 2021 Feb;12(2):450-462. [doi: 10.1007/s12671-018-1026-9] [Medline: 33737987]
6. Featherston R, Barlow J, Song Y, et al. Mindfulness-enhanced parenting programmes for improving the psychosocial outcomes of children (0 to 18 years) and their parents. Cochrane Database Syst Rev 2024 Jan 10;1(1):CD012445. [doi: 10.1002/14651858.CD012445.pub2] [Medline: 38197473]
7. Villalón López FJ. Mindfulness, compasión e intercuidado: su marco conceptual [Article in Spanish]. Pinelatinoamericana 2023;3(1):42-53 [FREE Full text]
8. Bishop SR, Lau M, Shapiro S, et al. Mindfulness: a proposed operational definition. Clin Psychol Sci Pract 2004;11(3):230-241. [doi: 10.1093/clipsy.bph077]
9. Tang YY, Hölzel BK, Posner MI. The neuroscience of mindfulness meditation. Nat Rev Neurosci 2015 Apr;16(4):213-225. [doi: 10.1038/nrn3916] [Medline: 25783612]
10. Bayot M, Brianda ME, van der Straten N, Mikolajczak M, Shankland R, Roskam I. Treating parental burnout: impact and particularities of a mindfulness- and compassion-based approach. Children (Basel) 2024 Jan 27;11(2):168. [doi: 10.3390/children11020168] [Medline: 38397280]

XSL•FO
RenderX

11. Urbanowicz AM, Shankland R, Rance J, Bennett P, Gauchet A. Informal mindfulness practices: a new approach to the prevention and treatment of parental burnout. Fam J 2024;32(4):580-586. [doi: 10.1177/10664807241230048]
12. Goldberg SB, Tucker RP, Greene PA, et al. Mindfulness-based interventions for psychiatric disorders: a systematic review and meta-analysis. Clin Psychol Rev 2018 Feb;59(2018):52-60. [doi: 10.1016/j.cpr.2017.10.011] [Medline: 29126747]
13. Mikolajczak M, Roskam I. A theoretical and clinical framework for parental burnout: the Balance Between Risks and Resources (BR²). Front Psychol 2018;9:886. [doi: 10.3389/fpsyg.2018.00886] [Medline: 29946278]
14. Brianda ME, Roskam I, Gross JJ, et al. Treating parental burnout: impact of two treatment modalities on burnout symptoms, emotions, hair cortisol, and parental neglect and violence. Psychother Psychosom 2020;89(5):330-332. [doi: 10.1159/000506354] [Medline: 32155634]
15. Urbanowicz AM, Shankland R, Rance J, Bennett P, Leys C, Gauchet A. Cognitive behavioral stress management for parents: prevention and reduction of parental burnout. Int J Clin Health Psychol 2023;23(4):100365. [doi: 10.1016/j.ijchp.2023.100365] [Medline: 36908695]
16. Mendonça I, Coelho F, Ferrajão P, Abreu AM. Telework and mental health during COVID-19. Int J Environ Res Public Health 2022 Feb 24;19(5):2602. [doi: 10.3390/ijerph19052602] [Medline: 35270294]
17. Villalón López FJ. Mindfulness, compasión e Inter cuidado: El programa de Inter cuidado basado en atención Plena (IBAP) [Article in Spanish]. Pinelatinoamericana 2023;3(2):103-114 [FREE Full text]
18. Physicians burnout: cause, consequence and mindfulness and intercare based intervention for physicians burnout reduction. ClinicalTrials.gov. 2022. URL: https://clinicaltrials.gov/ct2/show/NCT05013489 [accessed 2022-07-12]
19. Moreno MI, Rivera R, Villalón López FJ, et al. Brief online mindfulness- and compassion-based inter-care program for students during COVID-19 pandemic: a randomized controlled trial. Mindfulness (N Y) 2023;14(8):1918-1929. [doi: 10.1007/s12671-023-02159-8]
20. Crane RS, Eames C, Kuyken W, et al. Development and validation of the Mindfulness-Based Interventions—Teaching Assessment Criteria (MBI:TAC). Assessment 2013 Dec;20(6):681-688. [doi: 10.1177/1073191113490790] [Medline: 23794181]
21. Manrique-Millones D, Vasin GM, Dominguez-Lara S, et al. Parental burnout assessment (PBA) in different Hispanic countries: an exploratory structural equation modeling approach. Front Psychol 2022;13:827014. [doi: 10.3389/fpsyg.2022.827014] [Medline: 35465477]
22. Villalón López FJ, Escaffi-Schwarz M, Mundt A. Validation of the Five Facet Mindfulness Questionnaire scale in medical students and doctors in Chile [Article in Spanish]. Rev Med Chil 2023 Apr;151(4):435-445. [doi: 10.4067/s0034-98872023000400435] [Medline: 38687518]
23. Cebolla A, Demarzo M, Martins P, Soler J, Garcia-Campayo J. Unwanted effects: is there a negative side of meditation? A multicentre survey. PLoS One 2017;12(9):e0183137. [doi: 10.1371/journal.pone.0183137] [Medline: 28873417]
24. Lee S, Lee DK. What is the proper way to apply the multiple comparison test? Korean J Anesthesiol 2018 Oct;71(5):353-360. [doi: 10.4097/kja.d.18.00242] [Medline: 30157585]
25. Kline RB. Principles and Practice of Structural Equation Modeling, 4th edition: The Guilford Press; 2015.
26. Little TD. Longitudinal Structural Equation Modeling: The Guilford Press; 2013.
27. MacKinnon DP, Krull JL, Lockwood CM. Equivalence of the mediation, confounding and suppression effect. Prev Sci 2000 Dec;1(4):173-181. [doi: 10.1023/a:1026595011371] [Medline: 11523746]
28. Perneger TV. What's wrong with Bonferroni adjustments. BMJ 1998 Apr 18;316(7139):1236-1238. [doi: 10.1136/bmj.316.7139.1236] [Medline: 9553006]
29. Gamble C, Krishan A, Stocken D, et al. Guidelines for the content of statistical analysis plans in clinical trials. JAMA 2017 Dec 19;318(23):2337-2343. [doi: 10.1001/jama.2017.18556] [Medline: 29260229]
30. Page MJ, Higgins JPT, Clayton G, Sterne JAC, Hróbjartsson A, Savović J. Empirical evidence of study design biases in randomized trials: systematic review of meta-epidemiological studies. PLoS One 2016;11(7):e0159267. [doi: 10.1371/journal.pone.0159267] [Medline: 27398997]
31. Balk EM, Bonis PAL, Moskowitz H, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 2002 Jun 12;287(22):2973-2982. [doi: 10.1001/jama.287.22.2973] [Medline: 12052127]
32. Gul RB, Ali PA. Clinical trials: the challenge of recruitment and retention of participants. J Clin Nurs 2010 Jan;19(1-2):227-233. [doi: 10.1111/j.1365-2702.2009.03041.x] [Medline: 20500260]
33. Moreno Alba DF, Figueroa Ó, Gurdon C, Moreno Alba DF, Figueroa Ó, Gurdon C. Desigualdades urbanas: costos y tiempos de viaje en el Área Metropolitana de Santiago [Article in Spanish]. Rev INVI 2021;36(102):54-79. [doi: 10.4067/S0718-83582021000200054]
34. Urbanowicz AM, Verger NB, Shankland R, Rance J, Bennett P, Gauchet A. A meta-analysis of parental burnout interventions. J Affect Disord 2026 Apr 15;399:121022. [doi: 10.1016/j.jad.2025.121022] [Medline: 41443317]
35. Lindström C, Åman J, Anderzén-Carlsson A, Lindahl Norberg A. Group intervention for burnout in parents of chronically ill children—a small-scale study. Scand J Caring Sci 2016 Dec;30(4):678-686. [doi: 10.1111/scs.12287] [Medline: 26395446]

XSL•FO
RenderX

36. Whitney RV, Smith G. Emotional disclosure through journal writing: telehealth intervention for maternal stress and mother-child relationships. J Autism Dev Disord 2015 Nov;45(11):3735-3745. [doi: 10.1007/s10803-014-2332-2] [Medline: 25503483]
37. Guo L. The delayed, durable effect of expressive writing on depression, anxiety and stress: a meta-analytic review of studies with long-term follow-ups. Br J Clin Psychol 2023 Mar;62(1):272-297. [doi: 10.1111/bjc.12408] [Medline: 36536513]
38. Klimczak KS, Twohig MP, Peacock GG, Levin ME. Using peer-support coaching to improve adherence to online ACT self-help for college mental health: a randomized controlled trial. Behav Res Ther 2023 Jan;160:104228. [doi: 10.1016/j.brat.2022.104228] [Medline: 36455430]
39. Gillath O, Karantzas GC, Romano D, Karantzas KM. Attachment security priming: a meta-analysis. Pers Soc Psychol Rev 2022 Aug;26(3):183-241. [doi: 10.1177/10888683211054592] [Medline: 35209765]
40. Stein E, Witkiewitz K. Dismantling mindfulness-based programs: a systematic review to identify active components of treatment. Mindfulness (N Y) 2020 Nov;11(11):2470-2485. [doi: 10.1007/s12671-020-01444-0] [Medline: 36816712]
41. Feldman G, Greeson J, Senville J. Differential effects of mindful breathing, progressive muscle relaxation, and loving-kindness meditation on decentering and negative reactions to repetitive thoughts. Behav Res Ther 2010 Oct;48(10):1002-1011. [doi: 10.1016/j.brat.2010.06.006] [Medline: 20633873]
42. Garland EL, Hanley AW, Baker AK, Howard MO. Biobehavioral mechanisms of mindfulness as a treatment for chronic stress: an RDoC perspective. Chronic Stress (Thousand Oaks) 2017 Feb;1:2470547017711912. [doi: 10.1177/2470547017711912] [Medline: 28840198]
43. Kinnunen SM, Puolakanaho A, Tolvanen A, Mäkikangas A, Lappalainen R. Improvements in mindfulness facets mediate the alleviation of burnout dimensions. Mindfulness (N Y) 2020 Dec;11(12):2779-2792. [doi: 10.1007/s12671-020-01490-8]
44. PaulVogt W. Suppressor variables in path models: definitions and interpretations. In: SAGE Quantitative Research Methods: SAGE Publications; 2011. [doi: 10.4135/9780857028228]
45. Vatcheva KP, Lee M, McCormick JB, Rahbar MH. Multicollinearity in regression analyses conducted in epidemiologic studies. Epidemiology (Sunnyvale) 2016 Apr;6(2):227. [doi: 10.4172/2161-1165.1000227] [Medline: 27274911]
46. Li WW, McIntyre R, Reid C, Chao M. A systematic review on the state of the art of culturally adapted mindfulness-based interventions for First Nations peoples: cultural adaptation, effectiveness, and feasibility. Mindfulness (N Y) 2024 Sep;15(9):2147-2172. [doi: 10.1007/s12671-024-02440-4]
47. Ruotsalainen I, Valtanen M, Kärsämä R, et al. Unraveling the factors associated with digital health intervention uptake: cross-sectional study. JMIR mHealth uHealth 2025 Dec 9;13:e63896. [doi: 10.2196/63896] [Medline: 41364913]
48. Estrela M, Semedo G, Roque F, Ferreira PL, Herdeiro MT. Sociodemographic determinants of digital health literacy: a systematic review and meta-analysis. Int J Med Inform 2023 Sep;177:105124. [doi: 10.1016/j.ijmedinf.2023.105124] [Medline: 37329766]

Abbreviations

AC: active control
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
IBAP-BP: mindfulness- and compassion-based intercare program for parental burnout
MBCT: mindfulness-based cognitive therapy
MBI: mindfulness-based intervention
MBSR: mindfulness-based stress reduction
mITT: modified intention-to-treat
PB: parental burnout
PBA: Parental Burnout Assessment
RCT: randomized controlled trial
SAP: Statistical Analysis Plan
TIDieR: Template for Intervention Description and Replication
WL: waitlist control

XSL•FO
RenderX

Edited by M Balcarras; submitted 08.Nov.2025; peer-reviewed by T Jimoh, Y Chen; revised version received 11.May.2026; accepted 12.May.2026; published 25.Jun.2026.

Please cite as:
Villalón López FJ, Escaffi-Schwarz M
An Internet-Based Mindfulness- and Compassion-Based Intercare Program for Reducing Parental Burnout: Randomized Controlled Trial
J Med Internet Res 2026;28:e87416
URL: https://www.jmir.org/2026/1/e87416
doi:10.2196/87416

© Francisco Javier Villalón López, Maximiliano Escaffi-Schwarz. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Understanding University Students' Experiences of Engaging With AI and Apps for Their Mental Health and Well-Being: Qualitative Study

Julia Groot[1], PhD; Ben Ainsworth[2], PhD; Nathalia Gjersoe[1], PhD; Christopher Clarke[3], PhD; Richard Hamshaw[1], PhD; Mark Brosnan[4], PhD

[1]Department of Psychology, University of Bath, Claverton Down, Bath, United Kingdom

[2]School of Psychology, Faculty of Environmental and Life Sciences, University of Southampton, Southampton, United Kingdom

[3]Department of Computer Science, University of Bath, Bath, United Kingdom

[4]Centre for Applied Autism Research, Department of Psychology, University of Bath, Bath, United Kingdom

Corresponding Author:
Julia Groot, PhD
Department of Psychology, University of Bath, Claverton Down, Bath, United Kingdom

Abstract

Background: Fully automated digital technologies, such as artificial intelligence (AI) and apps, provide a particularly promising way to promote and support mental health and well-being in university students due to their accessibility, scalability, and cost-effectiveness, among other factors. Nevertheless, they are currently impeded by suboptimal engagement and high dropout rates, limiting their effectiveness to promote and support mental health and well-being.

Objective: This study aimed to understand university students' experiences of engaging with AI and apps to promote and support mental health and well-being.

Methods: University students who did not currently experience a mental health condition were recruited to ensure a nonclinical sample. Qualitative semistructured interviews were adopted and focused on students' experiences of engagement with AI and apps for their mental health and well-being. These interviews were conducted in April and May 2023 and lasted 30 minutes, 2 seconds (SD 9 min, 49 s), on average. Interviews were transcribed verbatim and analyzed using a reflexive thematic analysis.

Results: A total of 21 interviews were conducted, and 4 main themes and 4 subthemes were constructed. The first main theme refers to the "need" to engage with AI and apps for mental health and well-being. Specifically, this theme describes how nonclinical students would primarily use these technologies as a support strategy when their mental health and well-being deteriorate, and their preexisting mental health and well-being strategies are insufficient. The second theme refers to AI and apps as both a barrier and solution to stigma; while students are less inclined to access mental health apps due to stigma, they also consider apps to be less intrusive compared with other forms of support. The third theme considers a lack of trust in AI and apps. This lack of trust primarily exists due to skepticism about the capabilities of AI and apps supporting and promoting mental health and well-being, and skepticism about their ability to safeguard mental health and well-being. The final theme describes how usage is dependent on unique AI and app characteristics. Students may engage more in AI and apps when humanity, warmth, and care are considered less crucial, and when a lack of judgment and pressure is considered imperative.

Conclusions: Overall, nonclinical university students were more likely to engage with AI and apps when they experienced a decline in their mental health and well-being. Thus, it could be more beneficial to adopt apps as a support strategy rather than as a promotional strategy in a nonclinical sample. Furthermore, future policy and practice should implement strategies to safeguard mental health and well-being and provide open and honest communication about the capabilities of AI and apps in order to build trust and enhance engagement with digital technologies for mental health support.

(J Med Internet Res 2026;28:e75381) doi:10.2196/75381

KEYWORDS

AI; apps; mental health apps; MHapps; mental health; well-being; support; promotion; engagement; artificial intelligence

XSL•FO
RenderX

Introduction

Background

Mental well-being is a complex construct that is commonly defined to include 2 perspectives—the hedonic and the eudaimonic perspective [1]. The hedonic perspective refers to a subjective experience (often summarized as happiness), which includes satisfaction with life, the presence of a positive mood, and the absence of a negative mood [2]. The eudaimonic perspective is defined as a process of self-realization that contains six dimensions, including (1) autonomy, (2) purpose in life, (3) personal growth, (4) positive relations with others, (5) environmental mastery, and (6) self-acceptance [3]. Mental well-being can be promoted through activities that aim to enhance, build, or strengthen individuals' strengths, resourcefulness, or resiliency [4]. The main aim of mental well-being promotion is to enhance or maintain good levels of mental well-being. However, mental well-being and mental health are closely related, and therefore, mental well-being promotion often simultaneously improves mental health outcomes (such as anxiety and depression) [5].

A large proportion of nonclinical university students experience suboptimal levels of mental well-being [6-8]. This is likely due to the significant lifestyle changes university students experience, such as moving away from home, academic and financial pressures, and a loss of social and emotional support networks [9]. Additionally, research has suggested that approximately 25% of university students experience depression and anxiety [10], and this prevalence of mental health conditions in university students continues to rise. While many universities identify a duty of care for students' mental health and well-being, support services at universities are often overstretched [9], thereby limiting students' access to mental health and well-being support. This highlights the importance of accessible care in a nonclinical student population to prevent the onset of mental health conditions and improve their mental well-being.

A particularly promising way to deliver accessible mental health and well-being support to nonclinical university students is through fully automated digital technologies—digital technologies that do not require any form of ongoing human input to be delivered [11]. Automated digital technologies include apps that can be accessed through technological devices, such as smartphones and tablets [11]. Apps are designed to perform specific tasks, such as the delivery of activities for mental health and well-being. However, apps can also include generative artificial intelligence (GenAI), which allows for greater personalization of content, or they can rely entirely upon artificial intelligence (AI; eg, large language models) [12]. The automated delivery through apps and AI allows for mental health and well-being support to be highly scalable and cost-effective, thereby providing an easily accessible way to access support [13]. Unsurprisingly, mental health and well-being apps are becoming increasingly prevalent. The apps market size of US $6.25 billion in 2023 is expected to continue growing at 15.2% annually, with young adults (including university students) being its primary user [7,14].

A recent systematic review and meta-analysis indicates that automated apps and AI can effectively improve mental well-being in the nonclinical population [15]. Systematic reviews and meta-analyses also indicate that apps can effectively reduce mental health symptoms alongside improving mental well-being [16]. Additionally, AI has been shown to effectively support mental health [17]. Despite this great potential, studies have shown that these technologies suffer from suboptimal levels of engagement and high levels of dropout, thereby not reaching their full potential [15,18].

Some research has been conducted to understand engagement with specific mental health and well-being apps. Specifically, research investigated student engagement after providing them with access to a mental health app "Thought Spot" (Thought Spot was created through participatory design research methods and usability testing with college and university students [19]) for 6 months [20]. They found important app-specific determinants of continued engagement, such as detailed and inclusive content enhancing engagement, and technical errors and a lack of integration with other apps reducing engagement. Further research aimed to understand participants' engagement with mood-tracking apps [21]. They found further context-specific determinants of engagement, such as enjoying keeping track of trends in their moods through calendars. Similarly, research found app-specific determinants of engagement with a mental health promotion intervention, such as engaging and accessible app content [22].

Borghouts et al [23] conducted a review on barriers to engagement across different mental health and well-being apps, investigating studies from 2010 to 2019. However, this review did not adopt a nonclinical university student sample, which is particularly important since the review found mental health status to be an important determinant of engagement [23]. Furthermore, tremendous digital developments have occurred since then, and the review does not include perspectives regarding engagement with more advanced technologies, such as AI. Melcher et al [24] further conducted an empirical study to understand engagement across different mental health apps in college students, but did not conduct an in-depth qualitative analysis and may therefore provide less depth toward the contextual understanding of engagement [25]. Jardine et al [26] conducted an exploratory analysis of questionnaire and interview responses regarding engagement with mental health apps. However, this study focused on nonengagers. Overall, there is a need for research into engagement with AI and apps for mental health and well-being in a nonclinical student population, adopting in-depth qualitative analyses.

Research Aim

Therefore, this study aimed to understand university students' experiences of engaging with AI and apps for their mental health and well-being.

Methods

Study Design

This study adopted semistructured qualitative interviews.

XSL•FO
RenderX

Epistemological Position

A constructivist epistemological stance was taken in this study. This approach was adopted as it identifies that the researcher, interviewing and analyzing the data, plays a role in the construction of truth and meaning [27]. The approach identifies that people can construct the truth or meaning differently in similar situations. Therefore, it was considered important to reflect on the researchers' subjectivity and its impact on the data and findings by adopting a reflexive thematic analysis (TA) [28].

Participants and Recruitment

Participants were recruited via convenience sampling at the researchers' institution. A study advert (with relevant information regarding the study and a link to sign up) was promoted at sites at the researchers' institution, and via online university channels. University channels used to advertise the study targeted students in the Department of Psychology and Computer Science.

To ensure the confidentiality of participants, no demographic information was collected. Particularly, since demographic characteristics (such as age, gender, year of study, and department) in combination with qualitative interview data were considered to lead to highly identifiable data. Therefore, the general demographics of students in both departments were obtained instead.

Sample Size

The sample size was considered using the information power principle. This suggests that the more information a sample holds, the smaller the sample is needed [29]. This study adopted a relatively broad study aim and cross-case analysis, but a dense sample specificity, use of an established theory, and strong dialogue. Therefore, the initial aim was to recruit 15 - 20 participants.

Eligibility Criteria

A convenience student sample was recruited. Eligible participants for this study were (1) university students, (2) aged more than 18 years. Participants were excluded if they (1) were receiving treatment or psychopharmacological drugs for a mental health condition at the time of the study, (2) they scored ≥10 on the Patient Health Questionnaire-8, and (3) they scored ≥8 on the Generalized Anxiety Disorder-7. This was to ensure participants' mental health could be considered "nonclinical," as fully automated digital technologies are considered particularly beneficial as universal interventions targeted at the general nonclinical population [30].

There were no specific inclusion or exclusion criteria regarding students' experiences with mental health and well-being apps. Therefore, some students were currently using or had used several mental health and well-being apps, while other students had never used them before. Furthermore, the apps that students had used included a range of different ones, and the results are therefore not specific to an app. This was to ensure general perspectives and experiences regarding engagement across mental health and well-being apps were considered. Nevertheless, inductive coding was adopted to identify the mental health and well-being apps (including AI) that students referred to during the interviews.

Materials

A semistructured interview schedule was developed for the purpose of this study (Part S1 in Multimedia Appendix 1). This interview schedule adopted several open-ended questions aimed at understanding students' experiences of using AI and apps for their mental health and well-being. Since the initial focus of the interview was entirely on mental well-being promotion, participants were asked about their understanding of "mental well-being promotion" and provided a definition by the researcher (refer to Introduction section) to ensure participants had a consistent awareness of the meaning of this construct.

Procedure

Students expressed their interest in participating in the study by signing up via an online questionnaire link on the study adverts. This link provided participants with the information sheet (Multimedia Appendix 2), after which they were taken to the screening questionnaires (Patient Health Questionnaire-8 and Generalized Anxiety Disorder-7). After filling out the screening questionnaires, participants who were not eligible were provided with relevant information and resources for their mental health and well-being. Participants who were eligible were contacted by the researcher (JG) via email to schedule a convenient date and time for the interview to take place. The interviews took place in the laboratories at the researchers' institution between April 19 and May 31, 2023. Upon arrival in the laboratory, informed consent was obtained (Multimedia Appendix 3). Following this, a brief explanation was provided, and the interview was conducted. The interview consisted of 2 separate parts, lasting a maximum of 1 hour in total. The first part included a semistructured interview regarding participants' views of apps and AI for mental health and well-being support; this part of the interview was analyzed and is presented in this study. The second part of the interview involved participants looking at app content and sharing their thoughts out loud; this part of the interview was analyzed in a separate study. Refer to Multimedia Appendix 1 for the interview schedule, including both parts of the interview. Once the interviews finished, participants were provided with a debrief form (Multimedia Appendix 4) and student credits or a £10 (US $13.52) gift voucher to thank them for their time.

Ethical Considerations

Ethical approval was obtained from the Psychology Research Ethics Committee at the University of Bath (reference 23 042). Participants were first presented with an information sheet when they accessed the online expression of interest questionnaire. They were presented with the information sheet again at the start of their interview in the laboratory. Participants were provided with an opportunity to ask questions, and informed consent was obtained in writing before data collection. Participants had the opportunity to withdraw during the interview by informing the interviewer. After the interview was completed, participants could withdraw their data until May 31, 2023. Participants were informed about this on the information sheet. To ensure confidentiality, any identifiable information

XSL•FO
RenderX

was anonymized during transcription, and participants were assigned pseudonyms, which are used in this publication. Furthermore, to mitigate participant burden, combined duration of Part 1 and Part 2 of the interview (with Part 2 reported elsewhere) remained within a 1-hour time limit. Aftercare was provided if necessary. Specifically, if sensitive topics were discussed and/or concerns regarding mental health and well-being were raised during the interview, a discussion between the researcher and the supervisory team was conducted to identify the most appropriate form of aftercare—such as the university well-being apps and/or other mental health support services.

Data Analysis

Data was analyzed using a reflexive TA [28]. This was considered the most appropriate method of analysis to identify different themes and patterns of meaning regarding students' engagement with AI and apps for their mental health and well-being. An inductive approach was adopted to analyze the data. The interviews were audio-recorded using an Olympus VN-541 PC high-quality digital voice recorder. The part of the semistructured interviews reported in this study lasted 30 minutes 2 seconds (SD 9 min 49 s) on average, ranging from 16 minutes 20 seconds to 49 minutes 37 seconds. The recordings were transcribed using an orthographic (verbatim) transcription method. This method was adopted as it was decided that only information relevant to understanding and conceptualizing the meaning would be necessary to answer the research question and would therefore be transcribed. Interview transcripts were fully anonymized, and each participant was assigned a pseudonym.

The six phases of TA proposed by Braun and Clarke [28] were used to analyze the primary research question: (1) familiarization with the data took place, (2) transcripts were coded, (3) themes were generated, (4) themes were reviewed, (5) final defining and naming of themes, and (6) write-up took place. Specifically, a reflexive journal was kept throughout the analytical process, and continuing discussions took place between author JG and the research team to further conceptualize meaning. The transcripts were coded by author JG using NVivo 14 (Lumivero), adopting an inductive interpretive approach, which at times was more descriptive in nature. Codes were regularly revisited by JG to conceptualize and construct meaning. Initial themes were constructed by JG, which were further developed and reviewed in discussion with coauthors.

Reflexivity

Research Team

This study was undertaken by an interdisciplinary research team. The research team consisted of a PhD student in digital mental health and well-being interventions (JG), an associate professor in digital health and behavior change (BA), a professor in cognitive developmental psychology and an associate pro-vice-chancellor in student experiences (NG), a lecturer in human-computer interactions (CC), a senior lecturer in social psychology and qualitative research methods (RH), and a professor in autism spectrum disorder adopting digital technologies (MB).

Positionality

Author JG adopts a position as both an outsider (researcher) and an insider (peer or student) in this study. The remaining research team could be considered outsider researchers. Adopting a position as an insider researcher inevitably creates preconceptions regarding students' experiences of engaging with apps for mental health and well-being, shaping the construction of themes in this study. Furthermore, research found that important factors for self-disclosure include the recipient being considered trustworthy and confidential—qualities which may have been more likely to be instilled in an (outsider) researcher [31]. Therefore, it was prioritized to instill trust through the study environment (university laboratories), the interviewer (JG) introducing herself as an (outsider) researcher, and highlighting data confidentiality at the start of the interview.

Results

Overview

A total of 21 participants took part in this study. The initial estimation of sample size (guided by information power) was continually appraised during the analytic process [29]. Specifically, the authors identified that participants frequently conflated mental well-being promotion and mental health support during interviews. While there is considerable overlap between mental well-being promotion and mental health support, important differences exist. For example, mental well-being promotion primarily aims to maintain or improve mental well-being outcomes and is often delivered through universal interventions targeting the entire nonclinical population, while mental health support primarily aims to support mental health (such as anxiety and depression) and is often delivered to targeted or at-risk populations [30]. Despite this, conflation highlighted the need to revise the initial study aim to include mental health support alongside mental well-being promotion. This revision expanded the scope of the study aim, and therefore, it was agreed that additional participants were required to capture the thematic complexity. The final sample size was then determined using code saturation [32]. Author JG identified code saturation (whereby no new codes were identified) after 21 interviews. Code saturation was then verified through discussion with the research team.

Demographic characteristics of the population from which participants were recruited are presented in Table 1.

XSL•FO
RenderX

Table . Demographic characteristics for the 2022-2023 student cohort in the departments of Psychology and Computer Science.

Characteristics	Department of Psychology (N=1120), n (%)	Department of Computer Science (N=1725), n (%)
Sex		
Male	171 (15.25)	1353 (78.43)
Female	949 (84.75)	372 (21.57)
Age (years)		
18-20	777 (69.33)	640 (37.1)
21-24	199 (17.78)	295 (17.1)
25+	144 (12.89)	790 (45.8)
Level of education		
Undergraduate	780 (69.64)	645 (37.39)
Postgraduate taught	200 (17.86)	995 (57.68)
Research	140 (12.5)	85 (4.93)
Total number of students	1120 (100)	1725 (100)

Inductive coding identified a range of mental health and well-being apps and AI technologies referred to by participants during the interviews. These include habit and mood trackers, journaling, exercise, sleep, mindfulness, productivity, university well-being apps, mental health support via rule-based and AI-based chatbots, as well as AI therapists, AI to tailor or personalize content within apps, and general large language models, such as ChatGPT (OpenAI). For a full overview of the apps and AI participants (explicitly) referred to during interviews, refer to Multimedia Appendix 5. Some of the below findings referencing these apps and AI tools for mental well-being promotion and mental health support are discussed hypothetically, while others may reflect participants' actual experiences.

Summary of Themes and Subthemes

A total of 4 main themes and 4 subthemes were constructed by the authors to provide an understanding of students' experiences of engaging with AI and apps to promote and support their mental health and well-being (Table 2).

XSL•FO
RenderX

Table . Overview of constructed themes and subthemes.

Themes and subthemes	Characteristics
Experiencing a "need" to engage	This theme refers to the importance of having a mental health and well-being "need" to engage with AI[a] and apps. This "need" primarily occurs when preexisting mental health and well-being strategies are considered insufficient. Therefore, AI-based apps and apps are often not considered as "promotional tools" but instead as "support" for mental health and well-being.
A barrier and solution to stigma	This theme identifies stigma as a barrier to engagement with AI and apps for mental health and well-being support, potentially due to these tools primarily being considered as a "support" rather than a "promotion" strategy. It further identifies that AI and apps could potentially help reduce this stigma and "normalize" accessing support through stepped care.
Lack of trust	
Insufficient capability to support and promote mental health and well-being?	This subtheme refers to the skepticism students hold which works as a barrier toward engagement with AI and apps to promote their mental health and well-being. Specifically, it refers to the skepticism students have about these technologies' current and general ability to promote and support mental health and well-being.
An inability to safeguard mental health and well-being?	This subtheme refers to students' skepticism toward AI and apps due to the perceived inability to, and prioritization of, safeguarding of mental health and well-being in AI and apps, thereby reducing their engagement with apps.
Usage dependent on unique AI and app characteristics	
No judgment and pressure	This subtheme highlights how the unique AI and app characteristic "no judgment and pressure" facilitates engagement in certain situations. It identifies how an inability to judge and less perceived pressure in AI and apps provides nonintrusive access to mental health and well-being support.
A lack of humanity, warmth, and care	This subtheme highlights situations in which individuals may be more and less likely to engage with AI and apps depending upon the unique characteristic "lack of humanity, warmth, and care." It highlights that although humanity, warmth, and care are lacking, this could be considered a barrier as well as a facilitator to engagement, depending on each specific situation.

[a]AI: artificial intelligence.

Theme 1: Experiencing a "Need" to Engage

Most students identified that they did not engage with AI and apps for their mental health and well-being, as they did not currently experience a mental health or well-being "need." Students mentioned that this lack of "need" was due to their mental health and well-being being "good" at that moment.

> *I don't feel the need to, I guess … Just because I think that, uhm, I do think that I could gain sort of a greater sense of well-being from it, but … I have sort of a general quite high level of well-being. And so although I think that it could promote my well-being further, it's not the most sort of pressing concern for me currently.* [Jack]

Jack's account was interpreted as positioning engagement with AI and apps as conditional on experiencing lower well-being; since he currently feels his well-being is "quite high," Jack does not perceive a "need" to engage at this time. Interestingly, this contradicts theory that considers self-actualization a "growth need"—a need that occurs out of a desire to grow rather than a deficiency [33]. However, perhaps students perceive mental health and well-being as a deficiency need instead of a growth need. Deficiency needs motivate people when they are unmet, and once a deficiency need has been satisfied, it goes away [34]. Indeed, students identified that they used AI and apps in times of "need," but discontinued their use once this "need" was satisfied and went away.

> *If I feel like … after the first session or few sessions I feel like I've- already getting better like, like a bit improving, and I'll probably like stop using it for a bit, I guess like oh, maybe I don't need any more because I'm like, become I guess more stable enough [Interviewer: "mm-hmm"] to stop using it until I need it next time.* [Brenda]

Brenda identified that once the "need" to engage disappeared as her mental health improved, she would discontinue using the intervention. This may suggest students are not motivated to self-actualize through AI and apps and perceive these as relevant tools to satisfy a deficiency need instead.

Nevertheless, it may be consistent with the technology acceptance model (TAM). This model suggests that engagement with AI and apps occurs when this is considered "useful" [35]. Perhaps engaging when there is no mental health "need" might be considered less useful. Instead, most students mentioned that they would rather engage in other activities to promote their mental well-being and self-actualize.

XSL•FO
RenderX

> *If I want to promote my mental well-being, you know, I'll just go for a run after my lecture or something.* [Elsa]

> *[Playing squash] is quite good to like switch off and it's quite a social thing as well. … I'm enjoying it.* [Holly]

Holly and Elsa identified that when there is no immediate "need" to work on their mental health and well-being, they prefer engaging in other (nondigital) activities for their mental health and well-being. This might be because most students identified that they experience more joy and further benefits, such as connectedness and improvements to their physical health, when pursuing other activities. This may suggest that automated digital activities do not lead to the same satisfaction of basic psychological needs. According to the basic psychological needs theory (BPNT), the satisfaction of (1) autonomy (volition and willingness which lead to a sense of authenticity), (2) relatedness (connection, experiencing warmth, bonding, and care), and (3) competence (the experience of mastery to use and extend skills and expertise) are essential for optimal mental health and well-being [36]. Hence, perhaps students pursue other well-being activities as they lead to greater satisfaction of basic psychological needs.

Since students had already implemented these other well-being activities in their daily lives, they would also often prefer to stick to these activities instead of implementing a novel app or AI activity in their daily lives. They mentioned that only when these activities did not provide the necessary support or effect for their mental health and well-being, would they consider adopting novel mental well-being strategies delivered through automated digital technologies.

> *If there is an issue with my friend group, [Interviewer: "Mm-hmm"] and I only share my problem with my friends in a situation like that, the app would be useful where I could share it with someone who is not in that situation.* [Molly]

Molly highlights that her friend group would be her mental health and well-being support strategy. Whenever she is unable to access this support strategy, she would consider looking at novel support strategies such as apps and AI. Molly's account was interpreted to position apps as a secondary option—useful when her primary support system (friends) is unavailable—supporting the construction of the theme of need-dependent engagement.

Although students mention they are interested and open to using automated digital technologies for their mental health and well-being in times of "need," they also identified that adopting these self-help strategies when in "need" could be challenging. Primarily, these interventions require individuals to self-manage their condition, rather than receiving more intensive support and guidance, such as through traditional face-to-face support methods. Adopting self-help in times of "need" could be particularly challenging when experiencing reduced mental capacity.

> *If … I feel quite stressed, I feel like having to then, perform my [app activity], might feel unreachable, not realistic?* [Henry]

Henry identifies that engaging with automated digital technologies to perform a self-guided activity could feel "unreachable" and "unrealistic" when in times of "need." Similarly, research has found that the majority of students are interested in using an app for their mental health, but that only a small proportion of students (16/149, 10.7%) have actually engaged with apps during times of need [37]. This might indicate the challenge to engage with automated digital technologies when in times of "need."

Theme 2: A Barrier and Solution to Stigma

Another important factor impacting student engagement is public and self-stigma. While it has been suggested that AI and apps are associated with less stigma compared with traditional (nondigital) mental health support services [13], it is important to recognize that stigma still plays an important role, thereby reducing engagement. Several students identified that accessing any psychological intervention—including ones delivered digitally—is stigmatized. Perhaps this is unsurprising as stigma is considered a prominent barrier to help-seeking generally [38].

> ***Interviewer:*** *Is there anything that stops you, uhm, at the moment from using a mental well-being promotion app?*
>
> ***Jack:*** *Probably the stigma of using it. Uhm, as I go back to, it doesn't seem normalized, I don't know anyone who uses it.*

Jack highlights that stigma, mental health and well-being apps, and AI not being normalized, are important reasons for him not to engage. This is in line with research, which has found that self-stigma (someone's internalized attitudes) [39] has been linked to reduced help-seeking behaviors [40,41]. Jack shows self-stigma by mentioning that he personally does not think it is normalized, as he does not see anyone using it. Nevertheless, this contradicts recent findings that mental health and well-being apps are downloaded over a million times monthly [42].

However, stigma is understood as more prominent when apps have been developed for specific mental health purposes (such as an AI-based therapist) as opposed to low-key well-being activities (such as journaling and mindfulness meditation). It was also understood as a more prominent barrier when digital technologies are developed by recognized mental health organizations (such as universities, government, and mental health charities) compared with the private sector.

> *Uhm, there might be a s-s-stigma? Like using the university one [university mental health app]? Opposed to like using … like headspace, and stuff like that.* [Elsa]

> *Depends a little bit as well on what exactly it is that you're doing in that app. [in overlap] If it's- if it's for example, something like journaling, and there might be less stigma to than more of an AI focused kind of therapist.* [Jack]

XSL•FO
RenderX

Elsa and Jack suggest that the barrier "stigma" to engage with AI and apps for mental health and well-being is perhaps not uniform, suggesting that students perceive greater stigma when tools appear more clinical or institutionally linked. This likely arises from the perception that accessing more intensive psychological support, such as AI and apps developed by recognized mental health organizations, is associated with weakness. Yap et al [43] refer to this specific kind of stigma as "weak-not-sick beliefs," which is further associated with lower intentions to seek support. Indeed, several students mentioned that admitting to a decline in their mental health and well-being to an extent where they "need" additional support is stigmatized. Therefore, students adopt other strategies before admitting to this "need" for support and accessing AI and apps.

> *I'll like do an exercise or something like, uhm, like try to like calm down before I use any of these apps, because I feel like they're still a bit of … a bit of a pressure for me to use these apps. … I feel like I try to stay away from like, a professional kind of help, like, I'm quite an independent person.* [Harry]

Harry mentions he would try another activity first, as he tries to stay away from "professional help." Harry's emphasis on independence demonstrates how internalized stigmatizing beliefs about help-seeking shape reluctance to engage with digital tools. Harry mentions that he is an "independent person," which indicates that he perceives individuals who access support as not being independent and could relate to the "weak-not-sick belief." This again highlights the stigma associated with accessing mental health support generally, and therefore, apps for mental health and well-being support too. Nevertheless, mental well-being promotion might also suffer from this barrier, as promotion and support were often considered interchangeably by students.

Additionally, stigma was further shown in students hiding that they were accessing support. Students identified that it is easier to hide that they are accessing support when accessing AI and apps, but even then, they expressed concerns that someone else could find out.

> *So if, for example, someone has your phone and sees that, "oh, that person has a mental health app." That could potentially make someone less likely to download that app because they will think "oh, what if people think that there's something wrong with me?"* [Ralph]

Ralph mentions that someone finding a mental health and well-being app on their phone could be considered a barrier to engagement. This highlights the perceived public stigma associated with accessing any psychological intervention or support, and therefore, not wanting others to be aware of accessing this support. However, this could increase engagement relative to traditional face-to-face support methods, as it may be easier to hide that you are accessing support when accessing AI and apps.

Although students highlighted stigma would prevent them from engaging with apps and AI generally, students also identified that they could potentially provide a solution to public and self-stigma.

> *I think sometimes, with things that, with things like this that have a stigma. I think it takes people doing the small things first, to realise: "you know what? This isn't so bad."* [Anne]

> *I think taking that first step of trying something self-guided, might make people more willing to then contact them, use the messaging and contact someone in-, with the app.* [Barbara]

Anne and Barbara highlight the importance of engaging in something low-key for mental health and well-being first, which may act as a solution to gradually reduce stigma to engage with more intensive mental health and well-being support. Indeed, Ho et al [44] suggest that a stepped-care model could reduce stigma and improve access to those in "need." They suggest a stepped-care model in which different levels of care are provided, from less intensive care (such as self-guided interventions) to more intensive care (such as specialist treatment). Research further identifies several approaches that can effectively reduce stigma by adopting apps. For example, education-based approaches are one of the most commonly used approaches to combat stigma [45]; furthermore, mental well-being activities (such as the positive psychology activity "expressive writing") have been found to reduce self-stigma [46]. Both of these activities could be delivered through digital technologies and thereby help reduce stigma. Thus, AI and apps could potentially provide a crucial way of reducing stigma to alter public and self-stigma and enhance engagement with mental health and well-being support accordingly.

Theme 3: Lack of Trust

Overview

Many students recognized a great potential for AI and apps to promote and support their mental health and well-being. However, they currently do not engage with AI and apps due to a lack of trust. Trust is defined by Mayer et al [47] as: "a willingness of a party to be vulnerable to the action of another party based on the expectation that the party will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party." Indeed, relevant theory suggests that trust is an important factor determining engagement with digital technologies (TAM) [35,48]. Two subthemes have been constructed, identifying prominent reasons students currently lack trust, which prevents them from engaging with AI and apps.

Insufficient Capability to Support and Promote Mental Health and Well-Being?

Students are aware of technologies (particularly AI) developing at an exponential rate, and expressed concerns about whether these rapidly developing technological tools are sufficiently capable to provide support for their mental health and well-being. They identified that, at times, AI can make errors and is unable to tailor to an individual's needs.

> *I seem to get really frustrated [chuckles] when it's like a robot or like erm an AI on the other side because … there are certain things that they don't really understand.* [Harry]

XSL•FO
RenderX

Harry mentions that he gets very frustrated with AI when it does not understand him. Chhabra et al [49] indeed identified that a lack of semantic understanding (an inability to understand the queries), personalization when responding to the query, and competency (inability to provide a fitting solution) are some of the most common limitations in AI's capabilities that can lead to frustration in individuals. Similarly, Molly mentions she is skeptical of apps' abilities to support individuals' needs.

> *I would say in-person, in-person would more likely to be more focused on what the person needs because they're probably trained [Interviewer: "mm-hmm"] so they know what to do exactly and in what conditions, which the app might not be able to do that.* [Molly]

Harry and Molly's views illustrate students' doubts about whether digital tools can offer personalized, context-sensitive support. Nevertheless, their skeptical views on AI and apps' capability to support and promote mental health and well-being may be justified. Recent research by De Freitas et al [50] indeed suggests that AI returns messages to mental health queries that are considered unhelpful in as many as 24.5% (265/1080) of cases, and messages are even considered to be risky in 38.1% (411/1080) of cases. Furthermore, they found that in high-risk situations, such as suicide and self-harm, AI returned unhelpful and risky messages, such as by suggesting talking to people with the same interests. Its capability was found to depend on the clarity of communication by the user. For example, clear, explicit queries for support were more accurately supported compared with vague support queries [50].

While this refers to the limited capabilities when communicating verbally, it does not take into account nonverbal communication. Students identified the importance of nonverbal expressions in mental health and well-being support—information that AI and apps are currently unable to access—and how this might limit the capability to support mental health and well-being. Nonverbal expressions could include facial expressions, eye contact, body movement, and affect [51]. These nonverbal cues may at times more accurately represent true emotional states than verbal cues [52]. Furthermore, Burgoon et al [53] show that approximately 60% - 65% of interpersonal communication is displayed nonverbally. Thus, students are skeptical about AI and apps' ability to support mental health and well-being when they do not have access to any nonverbal expressions currently.

> *You might not have that same level of personal care? That you get from talking to a person? You know, they can sort of see how you're like, reacting to questions in a physical way? Like, I don't know, sometimes you might be talking and you sort of make a face when you're thinking about something, and that might sort of indicate to a GP or a mental healthcare practitioner that, maybe you're not getting to the bottom of what- how you're actually feeling. Whereas if you're sort of messaging on an app, you might not, it might not pick up on those like nuances that an in-person appointment might sort of access.* [Henry]

> *I don't think an app is good enough for a crisis … it's not directly seeing the person. I think that's what app can't mimic, it can't see the person, and directly see what that person, how they're feeling [Interviewer: "mm"]. And I think, if the person's in crisis, I just don't think the app would b-, I don't think it would be good enough?* [Meghan]

While Henry and Meghan may be speaking hypothetically, they express their concern about engaging with AI and apps as they do not directly "see" a person's nonverbal response and how they are feeling, therefore missing important information. While this could be considered a general limitation regarding the capability of AI and apps to provide accurate and effective care, this limited capability might be particularly important in high-risk situations, such as during a crisis (as suggested by Meghan). Indeed, research suggests that due to limited verbal communication in a crisis, non-verbal communication may be of particular importance [54]. Thus, while students expressed general concerns about the capability of AI and apps to provide mental health and well-being support, these concerns might be particularly present for crisis situations.

Thus, students were skeptical about whether mental health and well-being can ever be fully comprehended, supported, and promoted by automated digital technologies. Therefore, leading to a lack of trust, which currently prevents them from engaging with AI and apps to support and promote their mental health and well-being.

An Inability to Safeguard Mental Health and Well-Being?

Another significant concern students expressed is whether AI and apps can safeguard people's mental health and well-being. Safeguarding mental health and well-being includes protecting mental health and well-being and preventing harm [55]. AI and apps' limited capability, and the potential harm that could occur (especially when being used in high-risk situations), raised concerns about the morality and ethicality of using AI for mental health and well-being support. Students would engage with AI and apps for their mental health and well-being more if they were aware of safeguarding systems that are in place and knew that they could trust the apps' intention and ability to safeguard.

Students identified the importance of mental health and well-being being protected and safeguarded and that this should always be a priority. They mentioned that this should be at the forefront for them to trust AI and apps, and to engage with them for their mental health and well-being. Students identified that AI and apps should only be used for mental health and well-being purposes that are safe and have been proven effective. Furthermore, whenever mental health deteriorates, students identified it is important for AI and apps to recognize this, so the individual can be signposted to health care professionals instead.

> *I think that maybe, it should sort of identify a point where, "right this person needs to then be in contact with a person rather than just the app." … I think knowing the point where you then put them in contact with the person, would be really important.* [Carry]

XSL•FO
RenderX

Carry's account was considered to reflect a broader pattern across participants, in which human-delivered professional support is constructed as more trustworthy and preferable compared to AI- and app-based alternatives. This is illustrated by Carry's account that safeguarding should be in place by signposting an individual from the AI or app to a mental health care professional when mental health deteriorates. Many students identified that this was the moral and ethical obligation developers have when creating automated digital technologies for mental health and well-being.

> *If you were realising that someone was really struggling and then you didn't offer them [professional support], like in an ethical, moral standpoint, that wouldn't be very uhm righteous I think, to be almost like "You should stay here and we can solve everything you have."* [Colin]

Carry and Colin's quotes suggest that digital tools should recognize their inabilities and escalate appropriately to human support when users' mental health worsens to safeguard mental health and well-being. Leo et al [56] similarly found that when mental health symptom severity increased, a preference was displayed for human support compared to digital support—potentially due to individuals' enhanced trust in the capability and effectiveness of treatment provided by health care professionals.

Nevertheless, students also identified that safeguarding and signposting might be difficult due to the scalability of AI and apps. The great scalability means that many users from different cultures and backgrounds can access the app. This poses risks to safeguarding mental health, as cultural background is an important factor in the etiology of mental health conditions, as well as the expression and diagnosis associated with mental health symptoms [57].

> *I think problems could be maybe harder to monitor, because … if so many people are using this app around the world at any moment at any time, it would be very hard to have people … monitoring, uhm, whether there are these sort of safe practises ongoing.* [Harry]

Harry identifies that it could be challenging to safeguard mental health when many people are using AI and apps worldwide. Safeguarding by ensuring relevant and tailored support is provided to effectively support individuals and prevent harm across different cultural backgrounds is important. Furthermore, as the expression and diagnosis differ across cultural backgrounds, it is important to consider what signposting could be in place to further safeguard mental health and well-being.

Theme 4: Usage Dependent on Unique AI and App Characteristics

Overview

Students identified that AI and apps contain unique characteristics. These unique characteristics impact the situations in which students would engage with them. This is again consistent with the widely used TAM, which suggests that technological characteristics are important in determining engagement [35,58]. Although students primarily referred to AI when discussing specific technological characteristics, they may apply to general apps too.

No Judgment or Pressure

AI and apps contain unique characteristics that students found made it a favorable option for promotion and support of mental health and well-being in certain situations. Many students identified a concern about accessing professional support, as this takes up the time of health care workers. Furthermore, students are aware that mental health care services are overwhelmed and that waiting lists are long, thereby feeling guilty about taking up health care workers' time. In addition, they considered that taking up another person's time adds a certain "pressure." Similarly, students felt social support also has a "pressure" associated with it, while automated digital technologies do not contain this "pressure."

> *You don't feel like you're sort of wasting a GP's time.* [Robert]

> *… the pressure's off ... I feel like [face to face] is more intimidating actually, like talking to someone in person, like expressing your feelings.* [Sally]

These quotes by Sally and Robert identify the pressure and burden they perceive when accessing professional and social support, taking up another person's time. Sally identified that she feels more comfortable talking to an AI as there is less "pressure." Indeed, this was a view reflected by several students.

Another characteristic that students perceive positively is the perceived inability to judge. This was a reason for many students to engage with AI and apps for their mental health and well-being. Many students described being worried or anxious about judgment from others when seeking social support (eg, friends and family) as well as professional support. Talking to an AI or app that is unable to judge them, therefore provides a less intrusive way to access support.

> *When I first started therapy, I was not as comfortable expressing everything to my therapist, because at the end of the day, it is a human. You're always kind of thinking that "oh, what if they judge me though?" [Interviewer: "mm"] "What if they think certain things about me" [Interviewer: "mm"]. Therefore applying all that to an app … you can be as blunt and honest as you want [Interviewer: "mm-hmm"].* [Hayley]

> *I think it sort of frees you to be quite open about things. Whereas, you know, in front of someone, you might feel that sort of … judgement of maybe like what you're saying.* [Nathalie]

Hayley and Nathalie mention that when using digital technologies, they feel less judged and therefore more comfortable expressing themselves, indicating they feel less disrupted expressing themselves to automated digital technologies compared with people. It has long been established that anxiety and fear of judgment are important factors in determining someone's willingness to seek help [59] as well as someone's willingness to be open and discuss certain topics in therapy [60]. Overall, the AI and app characteristic lacking the

XSL•FO
RenderX

ability to judge enhances students' engagement for mental health and well-being purposes.

Lack of Humanity, Warmth, and Care

Students further identified that a unique characteristic that is lacking in AI and apps is warmth and humanity, which refers to the kindness and compassion that people have for one another. This kindness and compassion are uniquely shared among humans. Similarly, humans can "care" for one another, which is feeling or showing concern for or kindness to others. According to Watson [61], humans who care are "present, open to compassion, loving, merciful, gentle, loving and kind." While AI can convincingly imitate some human feelings and expressions, such as kindness or care, through speech, the motivation for these expressions does not stem from innate and underlying feelings of warmth, kindness, and care. Therefore, humanity and warmth are important determinants for the situations in which students may engage in these digital technologies.

> *… sometimes you just sort of need a person? … I know there's a lot that AI and, sort of, apps can do. But I think that, there's a lot to be said for being able to speak to a person. I think it gives you more of that sort of human element. There's more humanity in it.* [Ellie]

Ellie mentions in this quote that AI can "do a lot," which refers to AI being very capable in creating effective responses. For example, research has found that when individuals are unaware of whether a message is created by AI or a human, AI can at times better detect people's emotions and show empathy than humans can [62]. Nevertheless, Yin et al [62] also found that when individuals are aware that a response has been generated by AI instead of a human, individuals feel more "heard" by a human despite this response reflecting their emotions less accurately and being less empathetic. People adopting AI in the real world are aware they are talking to AI chatbots, and this could lead them to feel less heard. However, this lack of humanity seems imperative only when an individual needs humanity (eg, when they want someone to care for them, or they want to feel understood).

> *It depends on the circumstance and also, uhm, the reason why you decide to go to the app in the first place.* [Sarah]

Sarah identifies that not every goal or purpose might require humanity. This is in line with recent research findings that AI can effectively deliver a range of different strategies that promote and support mental health and well-being that do not require humanity, such as psychoeducation, treatment adherence, and positive psychology techniques [63]. Thus, humanity, warmth, and care can function as a barrier to engagement in certain situations. However, most individuals identified a range of situations in which AI and apps could be beneficial despite the lack of humanity, warmth, and care.

Overall, AI and apps contain unique characteristics that impact the way and situations in which individuals would engage with them for their mental health and well-being. Specifically, it might provide a less intrusive way to access mental health support due to the lack of pressure. Furthermore, individuals might be more likely to engage in situations in which a lack of judgment may be more imperative, or when they are in need of support that does not require humanity, warmth, and care.

Discussion

Summary of the Findings

This study aimed to understand nonclinical students' experiences of engaging with AI and apps to promote and support their mental health and well-being. A reflexive TA of 21 interviews led to the construction of 4 main themes and 4 subthemes. The main themes identified how (1) experiencing a "need" to engage, (2) stigma as a barrier and solution, (3) a lack of trust, and (4) usage dependent on unique AI and app characteristics can impact students' engagement with apps and AI for their mental health and well-being. Furthermore, in the theme "lack of trust," the subthemes "insufficient capability to support and promote mental health and well-being?" and "an inability to safeguard mental health and well-being?" highlight the lack of trust students currently experience in AI and apps to effectively support their mental health and well-being, thereby reducing their engagement. Finally, the theme "usage dependent on unique AI and app characteristics" consisted of subthemes "no judgment and pressure" and "a lack of humanity, warmth, and care," showing the unique characteristics of AI and apps, which determine the situations in which students are likely to engage with these digital technologies for their mental health and well-being.

Positioning of the Study Findings

It is important to reflect on the positioning of the findings of this study to consider its transferability. First, this research was conducted at a university in the United Kingdom, situating the findings within a Western cultural context and a highly educated sample. The majority of students at this university were undergraduate, home-based students, although a growing proportion (28,957/103,618, 23%) were international students [64]. At the time of this study, 65% (12,970/19,770) of students at the university identified as White, while 35% (6800/19,770) identified as belonging to Black, Asian, and ethnic minority groups. According to the TAM, such demographic variables influence engagement with digital interventions [58]. Therefore, the positioning of our participants in a Western university context limits transferability of the findings; for example, to other population subgroups with similar demographics (such as university students in a Western context, which includes hundreds of millions of students).

Second, recruitment strategies focused on the departments of Psychology and Computer Science, further positioning the sample as individuals with greater knowledge of mental health and well-being and/or digital technologies. The TAM suggests how a range of factors, such as self-efficacy, perceived ease of use, and technology anxiety, among others, play an important role in determining engagement with digital interventions. The study population's enhanced knowledge may lead to greater self-efficacy, ease of use, and reduced anxiety of adopting technology and/or mental health and well-being activities. Hence, this limits the transferability of the findings to other

XSL•FO
RenderX

population subgroups with greater knowledge of digital and/or mental health and well-being.

Finally, the rapid and ongoing development of GenAI positions this study temporally. Specifically, findings reflect participant perceptions formed during a specific technological and social context during which GenAI was novel while still rapidly developing. The TAM suggests that subjective norms and attitudes are important in determining engagement with digital technologies [58]. The subjective norms and attitudes participants may have held toward AI and apps may be temporally bounded. Nevertheless, the findings of this study hold relevance for similar populations and contexts, and can provide meaningful contributions to relevant theoretical frameworks, conceptual debates, and practical implications, which are explored in the following sections [65].

Contribution to Knowledge

Experiencing a "Need" to Engage

The first constructed theme suggests that a "need" is required for nonclinical students to engage with fully automated digital technologies for their mental health and well-being. Indeed, research has found that mental health "need" is an important predictor of engagement with digital mental well-being interventions [23]. This finding is also in line with the TAM, suggesting that "usefulness" of a digital intervention determines engagement. Nonclinical university students may find a digital well-being intervention less useful when their mental health and well-being are considered good. However, this contradicts Maslow's hierarchy of needs, suggesting that self-actualization is a growth need [33]. According to this theory, individuals engage in mental well-being activities from a desire to improve upon their mental well-being as opposed to a "need." In line with Maslow's hierarchy of needs, this study found that individuals engage in self-actualization through nondigital activities (such as sports, social activities, hobbies, etc) without experiencing a "need," suggesting self-actualization could be considered a growth need. Thus, perhaps the digital well-being interventions are uniquely perceived as activities that are "useful" to target deficiency needs but not growth needs.

The BPNT could provide an explanation for this. It suggests that individuals might not use fully automated digital interventions for self-actualization as these interventions may not be as beneficial as nondigital interventions. The BPNT suggests that individuals experience greater mental well-being benefits when the need for "relatedness" is met [36]. The need for relatedness may not be met to the same extent by a fully automated digital intervention as compared with nondigital interventions, and therefore, individuals may turn to digital interventions less for self-actualization. Indeed, relevant evidence suggests that fully automated digital well-being interventions are less beneficial compared to nondigital interventions for self-actualization [15]. However, this does not explain why some individuals do access fully automated digital interventions for self-actualization. Research has identified that nonclinical students engage and continue engaging with mental health and well-being apps in the real-world [66]. Therefore, future research should be conducted to understand why some nonclinical students may engage with fully automated digital interventions for the purpose of self-actualization.

A Barrier and Solution to Stigma

A second constructed theme suggests stigma is an important factor in determining engagement. This is consistent with much research on help-seeking for mental health and well-being [40,41,43]. It is important to consider that different factors and indicators relate to stigma across different cultures and countries [67]. Hence, these findings are positioned in a specific setting (Bath, United Kingdom), which has likely influenced the findings. For example, research has found that stigma is more prominent in Eastern compared with Western cultures [68]. Therefore, it is important to consider that stigma may have a different impact across cultural contexts (eg, cultures with increased stigma may be more likely to engage in apps compared with in-person support, as they are even more likely to hide accessing support) [68]. Thus, the findings of this study are consistent with relevant research in Western contexts, but the transferability of the findings to other cultural contexts is limited.

Lack of Trust

The third constructed theme suggests that a lack of trust in the capability of digital interventions to support mental health and well-being, alongside a lack of trust in the ability of these tools to safeguard mental health and well-being, reduces engagement. Indeed, the TAM suggests that trust can impact upon intentions to engage with digital interventions [58]. Interestingly, the updated versions of TAM largely omit the impact of technological capability on engagement, while the findings of this study might suggest that the capability of digital technology could impact trust, which might in turn lead to greater engagement. Additionally, research findings regarding the importance of trust in determining engagement are conflicting. Some research highlights the importance of "trust" in determining engagement with digital technologies [26], whereas other research does not find this [23]. Therefore, additional research is needed to understand the relationship between trust and engagement in digital mental well-being interventions.

Furthermore, it is important to consider that while this study found that digital tools may inherently be perceived as less capable and trustworthy than humans, apps and AI are becoming increasingly capable, which may (in part) diminish the impact of "trust" in determining engagement with digital mental well-being interventions. For example, they are increasingly able to simulate human-like conversations [69]. This might impact upon the ability of apps to deliver mental health and well-being promotion and support strategies, as research has found that "trust" is increased when systems act more human-like. Therefore, this finding may (in part) be temporally bounded.

Unique AI and App Characteristics

The final constructed theme suggests that unique AI and app characteristics, such as "a lack of humanity, warmth, and care," and "a lack of judgment and pressure," determine engagement. The TAM indeed suggests that technological characteristics could impact upon the engagement with digital technology,

XSL•FO
RenderX

although it does not suggest specific characteristics that may be important to consider. Similarly, previous research suggests that technology-related constructs can impact upon engagement [23,24], but this research does not suggest specific factors such as humanity or lack of judgment as important technological characteristics determining engagement in AI and apps for mental health and well-being purposes. Interestingly, while previous research suggests that a need for social connection is an important facilitator to engage in mental health apps [23,26], this study suggests that a lack of social connection may influence the specific contexts in which individuals engage. Perhaps this was found as this study focuses on fully automated digital interventions only, whereas previous research often focuses on nonautomated digital interventions. Nevertheless, this suggests that the situations in which individuals may engage in fully automated digital interventions compared with nonautomated interventions may differ. Future research is needed to explore this further.

Limitations of the Findings

This study has several limitations that should be considered when interpreting the findings. One study limitation is that students referred to mental well-being promotion and mental health support interchangeably. Initially, the interview focused on mental well-being promotion entirely, and therefore, a definition of mental well-being promotion was provided at the start of the interview (Multimedia Appendix 1). Despite this, participants frequently drifted from the definition to discuss the relevance of apps for mental health support instead. Participants' interchangeable use of these terms limits the findings, as it remains unclear whether reported experiences of engaging with apps and AI relate to mental well-being promotion, mental health support, or both. Similarly, participants' interchangeable reference to apps and AI makes it unclear whether these experiences relate specifically to apps, AI, or both.

Furthermore, while the research question was refined to include mental health support (which is considered good practice when adopting a reflexive qualitative approach [70]), the initial interview schedule did not include questions relating to mental health support. Therefore, findings relating to mental health support and crisis situations were entirely emergent and did not derive from systematic questioning across participants. This limits the representativeness of these findings. Therefore, future research should adopt systematic questioning across participants for more representative insights into their perceptions of engagement with apps for mental health support and during periods of crisis.

Another limitation is that complexities and contradictions may have been lost as the dataset was analyzed cross-cases. This is a particular issue in larger samples [28]. When focusing on patterns, the commonalities in the dataset are identified. Nevertheless, it is important to recognize that individual differences exist. For example, the TAM identifies the importance of demographic and personality characteristics impacting engagement with digital technologies [58]. Similarly, we found that while most students would only use AI and apps for mental health and well-being when in "need," some students would continually engage to promote their mental well-being. This loss of data is called the "fragmentation of accounts" [71], and while this means there is a loss of contradiction within the dataset, the study's aim was to provide a general overview of the commonly experienced factors to engage with AI and apps for their mental health and well-being. Nevertheless, it is important to consider that differences in perceptions and engagement exist across individual differences, which should be further investigated in future research.

Finally, a limitation of the study was that no demographic data were collected. While demographic data were not collected in order to protect participants' confidentiality, this limits the contextual understanding of the findings in this study. For example, it remains unclear whether participants were primarily male or female, and research has shown that sex (among other demographic factors) is important in determining the acceptance and engagement with digital technology [58]. Nevertheless, despite this limitation, the National Institute for Health and Care Excellence suggests that clear reporting of the recruitment strategy (eg, recruitment within the department of Computer Science and Psychology) can still contextualize the study findings to some extent [11].

Directions for Future Research

The themes constructed in this inductive reflexive TA have identified many different areas for future research into AI and apps to promote and support mental health and well-being in students. First, since students primarily identified interest in engaging in times of "need," future research could focus on engagement and the effectiveness of adopting these unguided self-help interventions during times of "need." Indeed, a recent systematic review aimed to understand engagement with unguided self-help apps in clinical populations but found insufficient evidence [72]. Thus, it remains unclear how effective these interventions are when in "need."

Second, while this study presents common perceptions and experiences regarding engagement, such as most students primarily being interested in engaging for mental health support, it is important to consider that individual differences exist. Indeed, Borghouts et al [23] found that individual differences impacted engagement with digital technologies. However, it remains unclear which individual differences impact engagement with and the effectiveness of digital mental well-being promotion [15]. Thus, future research should investigate the impact of individual differences in engagement with AI and apps for mental health and well-being promotion.

Third, while stigma has been identified as a barrier to engagement, the current findings propose that AI and apps could help reduce public and self-stigma. Either through a stepped-care model [44], or through effective activities combating stigma being incorporated into apps [45]. Nevertheless, future research is needed to investigate whether a stepped-care model that incorporates apps and engagement with activities within apps that reduce stigma could ease the public and self-stigma in a UK university setting.

Finally, this study identified that building trust in digital health technologies is an important determinant of engagement with AI and apps for mental health and well-being support. However,

XSL•FO
RenderX

it was also identified that mistrust is occasionally warranted due to the limitations in AI and app capabilities [50]. Future research should therefore establish when and under what conditions AI and apps can be "trusted." It should establish when and under what conditions AI and apps are safe to use and effective for mental health and well-being support. This future research would be essential to ensure appropriate recommendations regarding engagement can be made and to foster justified trust in these interventions.

Implications for Policy and Practice

Adopting an inductive thematic approach in this study allowed the generation of unanticipated insights and insights with actionable outcomes that could shape policy development [28]. Several implications for policy and actionable outcomes for AI and app developers are identified.

First, the study found that stigma was perceived as a common barrier to engagement in a UK setting. Therefore, UK policy and practice should consider ways to reduce the stigma associated with AI and apps for mental health and well-being purposes. Ho et al [44] suggest that less intensive support may be considered less stigmatizing. Hence, it may be important for digital mental well-being promotion to be advertised as distinct from more intensive digital mental health support to reduce stigma. Furthermore, other methods could be adopted to reduce public and self-stigma, including adopting education-based programs [73].

Furthermore, developers should consider whether developing AI and apps for the purpose of mental well-being promotion is most beneficial to nonclinical students. The World Health Organization suggests mental well-being promotion can be developed through universal interventions [30] that are aimed at general (nonclinical) population groups who are not specifically "at-risk" due to lower levels of mental health or well-being. Furthermore, developing these interventions digitally provides many benefits. However, this study suggests that mental well-being promotion interventions may be more engaging when delivered nondigitally. It also suggests that students are more willing to engage with AI and apps in times of "need." Thus, perhaps future AI and app development should primarily focus on supporting mental health conditions instead of promoting mental well-being.

Finally, policy and practice should incorporate ways to safeguard mental health when making AI and apps publicly available. Mental health conditions may present and can be identified differently across cultures [57], making it more challenging to safeguard mental health when these automated digital technologies are publicly available worldwide. This may be particularly challenging, as identifying high-risk individuals may have consequences for data protection and security [74]. Nevertheless, to build trust and enhance engagement with AI and apps, further strategies should be adopted by developers to safeguard mental health and well-being.

Conclusion

In conclusion, this study aimed to understand university students' experiences and perspectives of engaging with AI and apps to promote and support their mental health and well-being. Several themes were constructed by the research team, identifying that (1) most students have to experience a "need" to engage; (2) engagement may be impeded by stigma in university students in the UK, although AI and apps might also provide a promising solution to stigma through a stepped-care model and educational activities; (3) a lack of trust in the capability of AI and apps for mental health and well-being and its ability to safeguard are currently perceived as barriers to engagement, and policy and practice should consider ways to build trust to enhance engagement; and (4) unique AI and app characteristics may determine in what situations individuals may engage. Future research should investigate for whom these digital technologies may be engaging in nonclinical populations and when trust in AI and apps for mental health and well-being support may be warranted. Finally, future policy and practice should adopt strategies to reduce public and self-stigma and build trust by incorporating safeguarding strategies for mental health and well-being, particularly when scaling these automated digital technologies worldwide. Nevertheless, these accessible and cost-effective interventions provide a promising method to support mental health and well-being in university students.

Acknowledgments

The authors are grateful to the participants for taking part in this study.

No generative artificial intelligence was used in the write-up of this manuscript.

Funding

This study was funded as part of PhD funding by the Economic and Social Research Council (grant 2572559), and Cyberlimbic Systems Ltd, United Kingdom.

Data Availability

In line with open research practices and data sharing policies by the Economic and Social Research Council (ESRC; part-funder of this research project), interview transcripts from this study are available, where possible, on the University of Bath Research Data Archive (ID 1652).

XSL•FO
RenderX

Authors' Contributions

Conceptualization: JG, BA, CC, MB
Data curation: JG
Formal analysis: JG
Funding acquisition: JG, BA, CC, MB
Investigation: JG
Methodology: JG, BA, CC, MB
Project administration: JG
Resources: JG
Software: JG
Supervision: BA, CC, MB, NG, RH
Validation: JG, BA, CC, MB, NG, RH
Visualization: JG
Writing – original draft: JG
Writing – review and editing: JG, BA, CC, MB, NG, RH

Conflicts of Interest

JG's part-funding of PhD research project is by Cyberlimbic Systems Ltd, an app-developer. BA sits on advisory boards for Medito Foundation and earGym. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Interview schedule.
[DOCX File, 18 KB - jmir_v28i1e75381_app1.docx]

Multimedia Appendix 2
Information sheet.
[DOCX File, 25 KB - jmir_v28i1e75381_app2.docx]

Multimedia Appendix 3
Consent form.
[DOCX File, 25 KB - jmir_v28i1e75381_app3.docx]

Multimedia Appendix 4
Debrief form.
[DOCX File, 22 KB - jmir_v28i1e75381_app4.docx]

Multimedia Appendix 5
Overview of artificial intelligence (AI) and general apps (these apps may include AI features).
[DOCX File, 17 KB - jmir_v28i1e75381_app5.docx]

References

1. Ryan RM, Deci EL. On happiness and human potentials: a review of research on hedonic and eudaimonic well-being. Annu Rev Psychol 2001;52:141-166. [doi: 10.1146/annurev.psych.52.1.141] [Medline: 11148302]
2. Diener E. Subjective well-being. Psychol Bull 1984 May;95(3):542-575. [doi: 10.1037/0033-2909.95.3.542] [Medline: 6399758]
3. Ryff CD. Happiness is everything, or is it? Explorations on the meaning of psychological well-being. J Pers Soc Psychol 1989;57(6):1069-1081. [doi: 10.1037/0022-3514.57.6.1069]
4. Promoting mental health. : World Health Organization; 2004 URL: https://apps.who.int/iris/bitstream/handle/10665/42940/9241591595.pdf [accessed 2026-05-13]
5. Mental health. World Health Organization. URL: http://www.who.int/mediacentre/factsheets/fs220/en/ [accessed 2026-05-13]
6. Helliwell JF, Layard R, Sachs JD, De Neve JE, Aknin LB, Wang S, editors. World happiness report 2024. : University of Oxford: Wellbeing Research Centre; 2024 URL: https://worldhappiness.report/ed/2024/ [accessed 2026-05-13]
7. Martínez-García A, Valverde-Montesino S, García-García M. Promoting mental health in higher education: towards a model of well-being factors in emerging adulthood. Int J Qual Stud Health Well-being 2024 Dec;19(1):2408831. [doi: 10.1080/17482631.2024.2408831] [Medline: 39356675]
8. Keyes CL. Flourishing: Positive Psychology and the Life Well-Lived: American Psychological Association; 2003. [doi: 10.1037/10594-013]

XSL•FO
RenderX

9. Lewis J, Stiebahl S. Student mental health in England: statistics, policy and guidance [Research Briefing]. : UK Parliament; 2023 URL: https://dera.ioe.ac.uk/id/eprint/40985/ [accessed 2026-05-20]
10. Akram U, Irvine K, Gardani M, Allen S, Akram A, Stevenson JC. Prevalence of anxiety, depression, mania, insomnia, stress, suicidal ideation, psychotic experiences, & loneliness in UK university students. Sci Data 2023 Sep 13;10(1):621. [doi: 10.1038/s41597-023-02520-5] [Medline: 37704598]
11. Behaviour change: digital and mobile health interventions | NICE guideline. : National Institute for Health and Care Excellence; 2020 URL: https://www.nice.org.uk/guidance/ng183/resources/behaviour-change-digital-and-mobile-health-interventions-pdf-66142020002245 [accessed 2026-05-13]
12. Stryker C, Kavlakoglu E. What is artificial intelligence (AI)? International Business Machines. URL: https://www.ibm.com/topics/artificial-intelligence [accessed 2026-05-13]
13. Torous J, Jän Myrick K, Rauseo-Ricupero N, Firth J. Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow. JMIR Ment Health 2020 Mar 26;7(3):e18848. [doi: 10.2196/18848] [Medline: 32213476]
14. Leading health and meditation apps worldwide as of november 27, 2025, by downloads. Statista. URL: https://www.statista.com/statistics/1239640/top-health-and-meditation-apps-monthly-downloads/ [accessed 2025-02-10]
15. Groot J, MacLellan A, Butler M, et al. The effectiveness of fully automated digital interventions in promoting mental well-being in the general population: systematic review and meta-analysis. JMIR Ment Health 2023 Oct 19;10:e44658. [doi: 10.2196/44658] [Medline: 37856172]
16. Eisenstadt M, Liverpool S, Infanti E, Ciuvat RM, Carlsson C. Mobile apps that promote emotion regulation, positive mental health, and well-being in the general population: systematic review and meta-analysis. JMIR Ment Health 2021 Nov 8;8(11):e31170. [doi: 10.2196/31170] [Medline: 34747713]
17. Li H, Zhang R, Lee YC, Kraut RE, Mohr DC. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med 2023 Dec 19;6(1):236. [doi: 10.1038/s41746-023-00979-5] [Medline: 38114588]
18. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015 Jan 30;17(1):e30. [doi: 10.2196/jmir.4055] [Medline: 25639757]
19. Wiljer D, Shi J, Lo B, et al. Effects of a mobile and web app (Thought Spot) on mental health help-seeking among college and university students: randomized controlled trial. J Med Internet Res 2020 Oct 30;22(10):e20790. [doi: 10.2196/20790] [Medline: 33124984]
20. Wong HW, Lo B, Shi J, et al. Postsecondary student engagement with a mental health app and online platform (Thought Spot): qualitative study of user experience. JMIR Ment Health 2021 Apr 2;8(4):e23447. [doi: 10.2196/23447] [Medline: 33797395]
21. Schueller SM, Neary M, Lai J, Epstein DA. Understanding people's use of and perspectives on mood-tracking apps: interview study. JMIR Ment Health 2021 Aug 11;8(8):e29368. [doi: 10.2196/29368] [Medline: 34383678]
22. Renfrew ME, Morton DP, Northcote M, Morton JK, Hinze JS, Przybylko G. Participant perceptions of facilitators and barriers to adherence in a digital mental health intervention for a nonclinical cohort: content analysis. J Med Internet Res 2021 Apr 14;23(4):e25358. [doi: 10.2196/25358] [Medline: 33851925]
23. Borghouts J, Eikey E, Mark G, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387. [doi: 10.2196/24387] [Medline: 33759801]
24. Melcher J, Camacho E, Lagan S, Torous J. College student engagement with mental health apps: analysis of barriers to sustained use. J Am Coll Health 2022;70(6):1819-1825. [doi: 10.1080/07448481.2020.1825225] [Medline: 33048626]
25. Saunders CH, Sierpe A, von Plessen C, et al. Practical thematic analysis: a guide for multidisciplinary health services research teams engaging in qualitative analysis. BMJ 2023 Jun 8;381:e074256. [doi: 10.1136/bmj-2022-074256] [Medline: 37290778]
26. Jardine J, Nadal C, Robinson S, Enrique A, Hanratty M, Doherty G. Between rhetoric and reality: real-world barriers to uptake and early engagement in digital mental health interventions. ACM Trans Comput-Hum Interact 2024 Apr 30;31(2):1-59. [doi: 10.1145/3635472]
27. James N, Busher H. Online Interviewing: SAGE Publications; 2009.
28. Braun V, Clarke V. Thematic Analysis: A Practical Guide: SAGE Publications; 2021.
29. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
30. Promotion of mental well-being: pursuit of happiness. World Health Organization. 2013. URL: https://www.who.int/publications/i/item/9789290224228 [accessed 2026-05-13]
31. Ignatius E, Kokkonen M. Factors contributing to verbal self-disclosure. Nord Psychol 2007 Jan;59(4):362-391. [doi: 10.1027/1901-2276.59.4.362]
32. Vasileiou K, Barnett J, Thorpe S, Young T. Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. BMC Med Res Methodol 2018 Nov 21;18(1):148. [doi: 10.1186/s12874-018-0594-7] [Medline: 30463515]

XSL•FO
RenderX

33. Maslow A. Motivation and Personality: Harper and Brothers; 1954. URL: https://www.amazon.com/Motivation-Personality-H-Maslow/dp/B0006ATW94# [accessed 2026-05-20]
34. McLeod SA. Maslow's hierarchy of needs. Simply Psychology. URL: https://www.simplypsychology.org/maslow.html [accessed 2025-02-10]
35. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
36. Ryan RM, Deci EL. Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness: Guilford Press; 2017.
37. Lipschitz J, Miller CJ, Hogan TP, et al. Adoption of mobile apps for depression and anxiety: cross-sectional survey study on patient interest and barriers to engagement. JMIR Ment Health 2019 Jan 25;6(1):e11334. [doi: 10.2196/11334] [Medline: 30681968]
38. Aguirre Velasco A, Cruz ISS, Billings J, Jimenez M, Rowe S. What are the barriers, facilitators and interventions targeting help-seeking behaviours for common mental health problems in adolescents? A systematic review. BMC Psychiatry 2020 Jun 11;20(1):293. [doi: 10.1186/s12888-020-02659-0] [Medline: 32527236]
39. Link BG, Phelan JC. Conceptualizing stigma. Annu Rev Sociol 2001 Aug;27(1):363-385. [doi: 10.1146/annurev.soc.27.1.363]
40. Eisenberg D, Downs MF, Golberstein E, Zivin K. Stigma and help seeking for mental health among college students. Med Care Res Rev 2009 Oct;66(5):522-541. [doi: 10.1177/1077558709335173] [Medline: 19454625]
41. Dyrbye LN, Eacker A, Durning SJ, et al. The impact of stigma and personal experiences on the help-seeking behaviors of medical students with burnout. Acad Med 2015 Jul;90(7):961-969. [doi: 10.1097/ACM.0000000000000655] [Medline: 25650824]
42. UK mental health apps market (2025 - 2030). Grand View Research. URL: https://www.grandviewresearch.com/industry-analysis/uk-mental-health-apps-market-report [accessed 2026-05-13]
43. Yap MBH, Reavley NJ, Jorm AF. Associations between stigma and help-seeking intentions and beliefs: findings from an Australian national survey of young people. Psychiatry Res 2013 Dec 30;210(3):1154-1160. [doi: 10.1016/j.psychres.2013.08.029] [Medline: 24011848]
44. Ho FYY, Yeung WF, Ng THY, Chan CS. The efficacy and cost-effectiveness of stepped care prevention and treatment for depressive and/or anxiety disorders: a systematic review and meta-analysis. Sci Rep 2016 Jul 5;6:29281. [doi: 10.1038/srep29281] [Medline: 27377429]
45. Rao D, Elshafei A, Nguyen M, Hatzenbuehler ML, Frey S, Go VF. A systematic review of multi-level stigma interventions: state of the science and future directions. BMC Med 2019 Feb 15;17(1):41. [doi: 10.1186/s12916-018-1244-y] [Medline: 30770756]
46. Tang MW, Cheng Y, Zhang YH, Liu SJ. Effect of a positive psychology expressive writing on stigma, hope, coping style, and quality of life in hospitalized female patients with schizophrenia: a randomized, controlled trial. Perspect Psychiatr Care 2023 Jan;2023(1):1577352. [doi: 10.1155/2023/1577352]
47. Mayer RC, Davis JH, Schoorman FD. An integrative model of organizational trust. The Academy of Management Review 1995 Jul;20(3):709. [doi: 10.2307/258792]
48. Gefen D. What makes an ERP implementation relationship worthwhile: linking trust mechanisms and ERP usefulness. Journal of Management Information Systems 2004 Jul 1;21(1):263-288. [doi: 10.1080/07421222.2004.11045792]
49. Chhabra S, Kaushal V, Girija S. Determining the causes of user frustration in the case of conversational chatbots. Behav Inf Technol 2025 May 9;44(8):1557-1575. [doi: 10.1080/0144929X.2024.2362956]
50. De Freitas J, Uğuralp AK, Oğuz‐Uğuralp Z, Puntoni S. Chatbots and mental health: insights into the safety of generative AI. J Consum Psychol 2024 Jul;34(3):481-491. [doi: 10.1002/jcpy.1393]
51. Foley GN, Gentile JP. Nonverbal communication in psychotherapy. Psychiatry (Edgmont) 2010 Jun;7(6):38-44. [Medline: 20622944]
52. Philippot P, Feldman RS, Coats EJ, editors. Nonverbal Behavior in Clinical Settings: Oxford University Press; 2003. [doi: 10.1093/med:psych/9780195141092.001.0001]
53. Burgoon JK, Guerrero LK, Manusov V. Nonverbal signals. In: Knapp ML, Daly JA, editors. The SAGE Handbook of Interpersonal Communication, Fourth Edition 2011:239-280.
54. Rytterström P, Lindeborg M, Korhonen S, Sellin T. Finding the silent message: nurses' experiences of non-verbal communication preceding a suicide. PSYCH 2019;10(1):1-18. [doi: 10.4236/psych.2019.101001]
55. Safeguarding people. Care Quality Commission. URL: https://www.cqc.org.uk/what-we-do/how-we-do-our-job/safeguarding-people [accessed 2025-02-10]
56. Leo AJ, Schuelke MJ, Hunt DM, Miller JP, Areán PA, Cheng AL. Digital mental health intervention plus usual care compared with usual care only and usual care plus in-person psychological counseling for orthopedic patients with symptoms of depression or anxiety: cohort study. JMIR Form Res 2022 May 4;6(5):e36203. [doi: 10.2196/36203] [Medline: 35507387]
57. Hwang WC, Myers HF, Abe-Kim J, Ting JY. A conceptual paradigm for understanding culture's impact on mental health: the cultural influences on mental health (CIMH) model. Clin Psychol Rev 2008 Feb;28(2):211-227. [doi: 10.1016/j.cpr.2007.05.001] [Medline: 17587473]

XSL•FO
RenderX

58. Marangunić N, Granić A. Technology acceptance model: a literature review from 1986 to 2013. Univ Access Inf Soc 2015 Mar;14(1):81-95. [doi: 10.1007/s10209-014-0348-1]
59. Deane FP, Chamberlain K. Treatment fearfulness and distress as predictors of professional psychological help-seeking. Br J Guid Counc 1994 May;22(2):207-217. [doi: 10.1080/03069889408260315]
60. Radcliffe K, Masterson C, Martin C. Clients' experience of non‐response to psychological therapy: a qualitative analysis. Couns and Psychother Res 2018 Jun;18(2):220-229 [FREE Full text] [doi: 10.1002/capr.12161]
61. Watson J. Human Caring Science: Jones & Bartlett Learning; 2012.
62. Yin Y, Jia N, Wakslak CJ. AI can help people feel heard, but an AI label diminishes this impact. Proc Natl Acad Sci U S A 2024 Apr 2;121(14):e2319112121. [doi: 10.1073/pnas.2319112121] [Medline: 38551835]
63. Rubin M, Arnon H, Huppert JD, Perry A. Considering the role of human empathy in AI-driven therapy. JMIR Ment Health 2024 Jun 11;11:e56529. [doi: 10.2196/56529] [Medline: 38861302]
64. University of Bath student data. University of Bath. URL: https://www.bath.ac.uk/corporate-information/university-of-bath-student-data/ [accessed 2026-05-13]
65. Drisko JW. Transferability and generalization in qualitative research. Res Soc Work Pract 2025 Jan;35(1):102-110. [doi: 10.1177/10497315241256560]
66. Fürtjes S, Gebel E, Kische H, Beesdo-Baum K. Characteristic of mental health app usage: a cross-sectional survey in the general population. BMC Public Health 2024 Nov 12;24(1):3133. [doi: 10.1186/s12889-024-20500-1] [Medline: 39533246]
67. Weiss MG, Jadhav S, Raguram R, Vounatsou P, Littlewood R. Psychiatric stigma across cultures: local validation in Bangalore and London. Anthropol Med 2001 Apr;8(1):71-87. [doi: 10.1080/13648470120063906]
68. Krendl AC, Pescosolido BA. Countries and cultural differences in the stigma of mental illness: the east–west divide. J Cross Cult Psychol 2020 Feb;51(2):149-167. [doi: 10.1177/0022022119901297]
69. DeSmith C. Making algorithm used in AI more human-like. The Harvard Gazette. 2023. URL: https://news.harvard.edu/gazette/story/2023/07/making-algorithm-used-in-ai-more-human-like/ [accessed 2025-02-10]
70. Agee J. Developing qualitative research questions: a reflective process. Int J Qual Stud Educ 2009 Jul;22(4):431-447. [doi: 10.1080/09518390902736512]
71. King N, Brooks J. Thematic analysis in organisational research. In: SAGE Handbook of Qualitative Business and Management Research Methods 2018:219-236. [doi: 10.4135/9781526430236.n14]
72. Forbes A, Keleher MR, Venditto M, DiBiasi F. Assessing patient adherence to and engagement with digital interventions for depression in clinical trials: systematic literature review. J Med Internet Res 2023 Aug 11;25:e43727. [doi: 10.2196/43727] [Medline: 37566447]
73. Morgan AJ, Reavley NJ, Ross A, Too LS, Jorm AF. Interventions to reduce stigma towards people with severe mental illness: systematic review and meta-analysis. J Psychiatr Res 2018 Aug;103:120-133. [doi: 10.1016/j.jpsychires.2018.05.017] [Medline: 29843003]
74. Olawade DB, Wada OZ, Odetayo A, David-Olawade AC, Asaolu F, Eberhardt J. Enhancing mental health with artificial intelligence: current trends and future prospects. Journal of Medicine, Surgery, and Public Health 2024 Aug;3:100099. [doi: 10.1016/j.glmedi.2024.100099]

Abbreviations

AI: artificial intelligence
BPNT: basic psychological needs theory
GenAI: generative artificial intelligence
TA: thematic analysis
TAM: technology acceptance model

Edited by A Coristine, TDA Cardoso; submitted 02.Apr.2025; peer-reviewed by A Patricia Oliveira, A Eisingerich; revised version received 23.Apr.2026; accepted 24.Apr.2026; published 30.Jun.2026.

Please cite as:
Groot J, Ainsworth B, Gjersoe N, Clarke C, Hamshaw R, Brosnan M
Understanding University Students' Experiences of Engaging With AI and Apps for Their Mental Health and Well-Being: Qualitative Study
J Med Internet Res 2026;28:e75381
URL: https://www.jmir.org/2026/1/e75381
doi:10.2196/75381

XSL•FO
RenderX

© Julia Groot, Ben Ainsworth, Nathalia Gjersoe, Christopher Clarke, Richard Hamshaw, Mark Brosnan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Examining the Association Between Internet Addiction and Nonsuicidal Self-Injury Among Chinese Middle School Students: Prospective Cohort Study

Tianqing Fan, PhD, Dr med; Lintong Song, PhD, Dr med; Sihong Li, MM, Dr med; Xingyue Jing, PhD, Dr rer med; Mireille Twayigira, MM; Chunxiang Huang, PhD, Prof Dr Med; Xuerong Luo, Prof Dr Med; Yanmei Shen, PhD, Dr med

Department of Psychiatry, National Clinical Research Center for Mental Disorders, and National Center for Mental Disorders, The Second Xiangya Hospital of Central South University, No. 139, Renmin Middle Road, Furong District, Changsha, Hunan, China

Corresponding Author:
Yanmei Shen, PhD, Dr med
Department of Psychiatry, National Clinical Research Center for Mental Disorders, and National Center for Mental Disorders, The Second Xiangya Hospital of Central South University, No. 139, Renmin Middle Road, Furong District, Changsha, Hunan, China

Abstract

Background: While cross-sectional studies have consistently reported an association between nonsuicidal self-injury (NSSI) and internet addiction (IA), longitudinal evidence regarding the directionality and dose-response relationship remains limited. Furthermore, the roles of sex and varying degrees of problematic internet use in predicting new-onset NSSI are not fully understood.

Objective: This prospective cohort study aimed to investigate whether baseline IA and its intermediate states predict the subsequent new onset of NSSI among Chinese adolescents over a 6-month period and to explore potential sex differences in this longitudinal association.

Methods: A prospective cohort design was used. A total of 1315 junior high school students without a history of NSSI were recruited at baseline, and 704 (53.5%) students completed the 6-month follow-up. IA and NSSI were assessed using the Chinese Internet Addiction Scale-Revised and a self-report questionnaire from the Adolescent Health-Related Risky Behavior Inventory, respectively. Logistic regression analysis was conducted to examine the predictive value of IA exposure for incident NSSI, adjusting for key covariates, including sex, age, ethnicity, only child status, anxiety, and depression. Restricted cubic spline regression was used to model the dose-response relationship between distinct states of IA and NSSI risk.

Results: The baseline prevalence of IA was 9.09% (64/704). At the 6-month follow-up, the incidence rate of NSSI was 9.8% (69/704). Restricted cubic spline regression revealed a linear dose-response relationship, where the risk of incident NSSI escalated with increasing IA severity. In the fully adjusted model for the total sample, baseline IA was a significant predictor of subsequent NSSI (odds ratio [OR] 2.185, 95% CI 1.031-4.627; *P*=.04). Crucially, stratified analyses revealed significant sex disparities: the longitudinal association between IA and subsequent NSSI was statistically significant among female adolescents (OR 3.271, 95% CI 1.101-9.717; *P*=.03) and the intermediate internet-dependent state (OR 2.593, 95% CI 1.002-6.710; *P*=.049) but not among male adolescents (IA: *P*=.44; internet-dependent state: *P*=.87).

Conclusions: NSSI incidence is notably prevalent among Chinese junior high school students. While IA serves as a robust, independent risk factor for predicting the new onset of NSSI in the overall adolescent population, sex-stratified analyses revealed that this longitudinal association was statistically significant (*P*=.03) only among female students. These findings underscore the critical need to integrate IA assessments into school-based mental health screenings and highlight the necessity of developing sex-specific, emotion-focused prevention strategies to mitigate NSSI risk.

(J Med Internet Res 2026;28:e86427) doi:10.2196/86427

KEYWORDS

internet addiction disorder; self-injurious behavior; adolescents; sex characteristics; longitudinal studies

Introduction

Nonsuicidal self-injury (NSSI) is defined as committing direct and intentional self-harm behaviors without suicidal intent and frequently involves cutting or carving the skin [1]. It is quite common among young people, with a lifetime prevalence ranging from 17% to 60% [2]. It is significantly associated with higher risks of suicide and subsequent repeated NSSI [3]. The more frequent the NSSI, the higher the risk of suicidal behavior [4]. NSSI in adolescence is a strong indicator of mental health

XSL•FO
RenderX

problems later in life [5] and is often comorbid with mental disorders, such as emotional disorders (depression [6], bipolar disorder [7], and others), neurodevelopmental disorders [8], and personality disorders [9]. There has been burgeoning interest in exploring the correlation between internet addiction (IA) and NSSI.

IA, also called problematic internet use, is characterized by excessive and inappropriate human-computer interaction behavior and is considered a behavioral addiction [10]. IA can serve as a coping mechanism for emotional distress, as described by the compensatory internet use theory, which posits that adolescents engage in excessive internet use as a maladaptive strategy to alleviate negative affect or escape real-life difficulties [11]. It might also be associated with serious consequences including suicidal behavior and emotional problems such as depression and anxiety [12], particularly in relation to NSSI in Chinese adolescents [13]. Individuals with IA may isolate themselves from social interactions, escape reality, have a desire for instant satisfaction, and experience challenges in managing negative emotions [14].

NSSI also functions as a maladaptive emotional coping mechanism [15], and reliance on the virtual world for emotional regulation is often insufficient or temporary. According to the experiential avoidance model [16], when low-intensity avoidance strategies, such as internet use, fail to dampen overwhelming distress or mask underlying symptoms of anxiety and depression, adolescents may retreat into the virtual world to temporarily alleviate negative affect. This may obscure underlying symptoms of anxiety or depression that would otherwise trigger clinical attention. When online activities fail to effectively regulate intensifying emotional pain, adolescents may resort to NSSI as a more drastic mechanism for emotional release [17]. This mechanism may involve a desire to avoid or escape distressing situations, as well as difficulties in regulating emotions. Additionally, potential psychological factors, such as impulsivity, low self-esteem, depression, or anxiety, may also contribute to NSSI [18]. Given the abundance of similarities, IA and NSSI may share underlying mechanisms. Theoretically, IA may serve as a maladaptive coping strategy that exacerbates emotional dysregulation and social isolation, thereby increasing vulnerability to NSSI through psychological pathways, such as diminished self-esteem and identity confusion. This theoretical link is supported by emerging longitudinal evidence in related mental health domains. For instance, large-scale prospective studies among Chinese adolescents have demonstrated that internet gaming disorder significantly predicts the incidence and persistence of suicidal ideation and depression over time [19,20]. Furthermore, a dose-response relationship has been observed between internet gaming disorder severity and the risk of first-onset suicide attempts [21]. However, despite these advances in understanding suicidality, specific longitudinal research on IA and NSSI remains limited. Most existing studies have explored this association cross-sectionally. Consequently, strictly longitudinal designs are urgently needed to clarify the directionality and determine whether IA represents a modifiable risk factor for early intervention in NSSI.

Given that previous studies have indicated sex differences in IA [22] and NSSI [23], this pathway may be moderated by sex due to distinct patterns of online behavior and coping mechanisms. Studies have indicated that male individuals with IA exhibit a higher susceptibility to gaming addiction, while female individuals with IA are more prone to challenges related to social networking [24], where upward social comparison and cyberbullying serve as potent triggers for internalizing problems that may precipitate NSSI. Furthermore, women demonstrate a greater tendency to use NSSI as a means of coping with psychological stress and distress [25]. Despite these plausible mechanisms, literature specifically examining sex differences in the longitudinal relationship between IA and NSSI remains limited. This study aims to investigate the effect of baseline IA on incident NSSI through a prospective cohort design and to explore whether there are sex differences in this context. We hypothesized that baseline IA predicts subsequent NSSI and that this association varies between male and female adolescents.

Methods

Study Population and Data Collection

A baseline survey was conducted in November and December 2020 to gather data from an urban public middle school located in Changsha, Hunan Province, China. To ensure a systematic recruitment procedure, the research team collaborated with school administrators and head teachers. Prior to the formal survey, study information sheets and electronic informed consent forms were distributed to all targeted students and their legal guardians. Using a cluster sampling method, all classes from Grades 7 to 9 were included. Initially, 1785 students were invited. However, 94 (5.3%) students declined to participate, and 120 (6.7%) students had incomplete information, resulting in their exclusion from the study. The inclusion criterion was students aged 12 to 18 years with no serious physical illness who could comprehend and respond to the questionnaires. The exclusion criterion was students whose parents and themselves declined to participate or did not provide informed consent for participation. A total of 1571 valid questionnaires (n=834, 53.1% male students and n=737, 46.9% female students) were collected at baseline. First, to establish a prospective cohort for incidence analysis, 256 (16.3%) students with prevalent NSSI at baseline were excluded. The follow-up survey took place in May and June 2021. Among the remaining 1315 eligible participants, 611 (46.5%) were lost to follow-up, primarily because ninth-grade students graduated during the study period. Consequently, a total of 704 (53.5%) students completed the follow-up survey after 6 months. Thus, these 704 middle school students were included in the follow-up analysis. A priori power analysis using G*Power 3.1 (Heinrich Heine University Düsseldorf) indicated that a minimum sample size of 308 was required; therefore, we intentionally oversampled at baseline (N=1785) to ensure sufficient statistical power despite the expected attrition in this longitudinal study.

Main Exposure Variables

The Chinese Internet Addiction Scale-Revised (CIAS-R) was used for IA assessment. It is a 26-item scale designed by Chen et al in 2003 [26] to evaluate IA. Bai and Fan [27] revised it to a 19-item scale in 2005 and verified its validity and reliability in mainland adolescents in China (Cronbach α=0.90). This study

XSL•FO
RenderX

used the validated 19-item version. A Likert scale with a 4-point rating system was used, where each item was assessed on a scale ranging from 1 (extremely inconsistent with my experience) to 4 (very consistent with my experience). The cutoff score is 53; a score above 53 is regarded as having IA, while a score between 46 and 53 is defined as an internet-dependent state, and a score below 46 is defined as normal. Consequently, participants were classified into three categories based on the CIAS-R total score: (1) normal (controls): score <46; (2) internet-dependent state: score 46 to 53, representing an intermediate or at-risk state; and (3) IA: score >53. In this study, the scale showed high internal consistency (Cronbach α=0.969).

Outcome Assessment

NSSI was measured at baseline using the question "In the past 12 months, have you intentionally hurt yourself, but not for the purpose of taking your own life?" from the Adolescent Health-Related Risky Behavior Inventory, with "hit yourself," "banged your head against something hard," "pulled your own hair," "scratched yourself," "pinched yourself," "bit yourself," "cut yourself," "burned yourself," and "other methods" listed as examples. This instrument has been widely validated and extensively used in epidemiological studies among Chinese adolescents, demonstrating satisfactory test-retest reliability and construct validity [28-30]. NSSI was assessed at the 6-month follow-up using the same question, but the period was altered to the last 6 months. Participants were requested to select a yes or no response, and if they selected yes, they were asked to specify the NSSI methods they used.

Assessment of Covariates

Basic sociodemographic information, such as age, sex, only child status, and ethnicity, were collected. Depression and anxiety in the past week were collected by the 2 subscales of the Depression Anxiety Stress Scale (DASS-21) [31]. Each subscale includes 7 items, and each item is rated on a 4-point Likert scale ranging from 0 (extremely inconsistent with my experience) to 3 (very consistent with my experience). Higher scores indicate more severe emotional problems, and the cutoff scores for the presence of depression and anxiety were 5 and 4 (raw score), respectively, based on the original criteria. This scale has been commonly used in Chinese adolescents and has proven to have good reliability, validity, and high practical value [32]. In this study, the Cronbach α coefficients for depression and anxiety subscales were 0.888 and 0.841, respectively.

Procedure

The longitudinal study was implemented in 2 waves. As described in the *Study Population and Data Collection* section, the baseline survey was conducted in November and December 2020, followed by a 6-month follow-up in May and June 2021. Both waves used an online survey distributed via WeChat (Tencent Holdings Ltd). To ensure data matching between baseline and follow-up while maintaining confidentiality, student ID numbers were used as unique identifiers. Before the survey, all the schoolteachers who participated in this study were trained uniformly on how to deliver the survey, how to guide the students to complete the questionnaire, and how to assist when needed. Notably, students completed the survey in classrooms under uniform supervision to ensure standardized conditions and independent completion.

Ethical Considerations

The ethics committee of the Second Xiangya Hospital of Central South University thoroughly reviewed and approved the use of the data, ensuring compliance with ethical standards (approval (2019) Lun-Shen [Ke] 021). Additionally, prior informed consent was obtained from all participant students and their legal guardians. All the students and their caregivers were made aware that participation in the study was entirely voluntary and that they could opt out at any time. To safeguard privacy and confidentiality, all collected data were deidentified; student ID numbers were temporarily used solely for matching baseline and follow-up surveys, after which the dataset was fully anonymized. Finally, participants did not receive any financial compensation for their involvement in this study.

Statistical Analysis

Chi-square tests were used to compare categorical variables. After excluding NSSI at baseline, the association between baseline IA and new-onset NSSI during a 6-month follow-up was examined by logistic regression analysis. We built 3 models for the data. Model 1 calculated the crude odds ratios (ORs) with a 95% CI without adjusting for any covariable; model 2 adjusted for age, sex, only child status, and ethnicity; and model 3 adjusted for age, sex, only child status, ethnicity, and emotional state (including anxiety and depression). Furthermore, sex differences were also explored by constructing models hierarchically by sex. Statistical analyses were conducted using SPSS (version 23.0; IBM Corp) for chi-square tests and logistic regression, while R software (version 4.3.1; R Foundation for Statistical Computing) was used for restricted cubic spline analysis. Significance was set at $P<.05$. The reporting of this study adheres to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for cohort studies (Checklist 1).

Results

Baseline Characteristics of the Final Analytical Sample

A total of 704 participants who were free of NSSI at baseline completed the assessments and were included in the final analytical sample (Figure 1). The cohort comprised 395 (56.1%) male and 309 (43.9%) female participants, with an average age of 12.60 (SD 0.61) years. Based on the CIAS-R criteria, the baseline prevalence of the internet-dependent state was 11.7% (n=82), and the prevalence of IA was 9.1% (n=64). Baseline demographic comparisons revealed no significant intergroup differences among the normal, internet-dependent state, and IA groups concerning age, sex, or only child status (all $P>.05$); a statistically significant difference was observed regarding ethnicity, primarily driven by the overwhelming majority of Han participants, which is representative of the regional demographic profile. However, regarding clinical characteristics, students in the internet-dependent state and IA groups exhibited a significantly higher baseline prevalence of depression ($P<.001$) and anxiety ($P<.001$) compared with the normal control group.

XSL•FO
RenderX

The consolidated demographic and clinical characteristics of the final analytical sample are summarized in Table 1. During the 6-month follow-up period, the overall incidence rate of new-onset NSSI among the final analytical cohort was 9.8% (69/704). To assess potential attrition bias, baseline characteristics were compared between participants who completed the follow-up and those who were lost. Results showed no significant differences in sex or baseline IA scores (all $P>.05$), indicating no systematic bias related to the primary exposure variable. Although the attrition group was significantly older due to graduation ($P<.001$; refer to Multimedia Appendix 1 for details), age was included as a covariate in all subsequent multivariate logistic regression models to statistically adjust for this baseline discrepancy.

Figure 1. Flowchart of participant recruitment and 6-month follow-up in the prospective cohort study investigating incident nonsuicidal self-injury (NSSI) among middle school students (Changsha, China; November 2020 to June 2021).

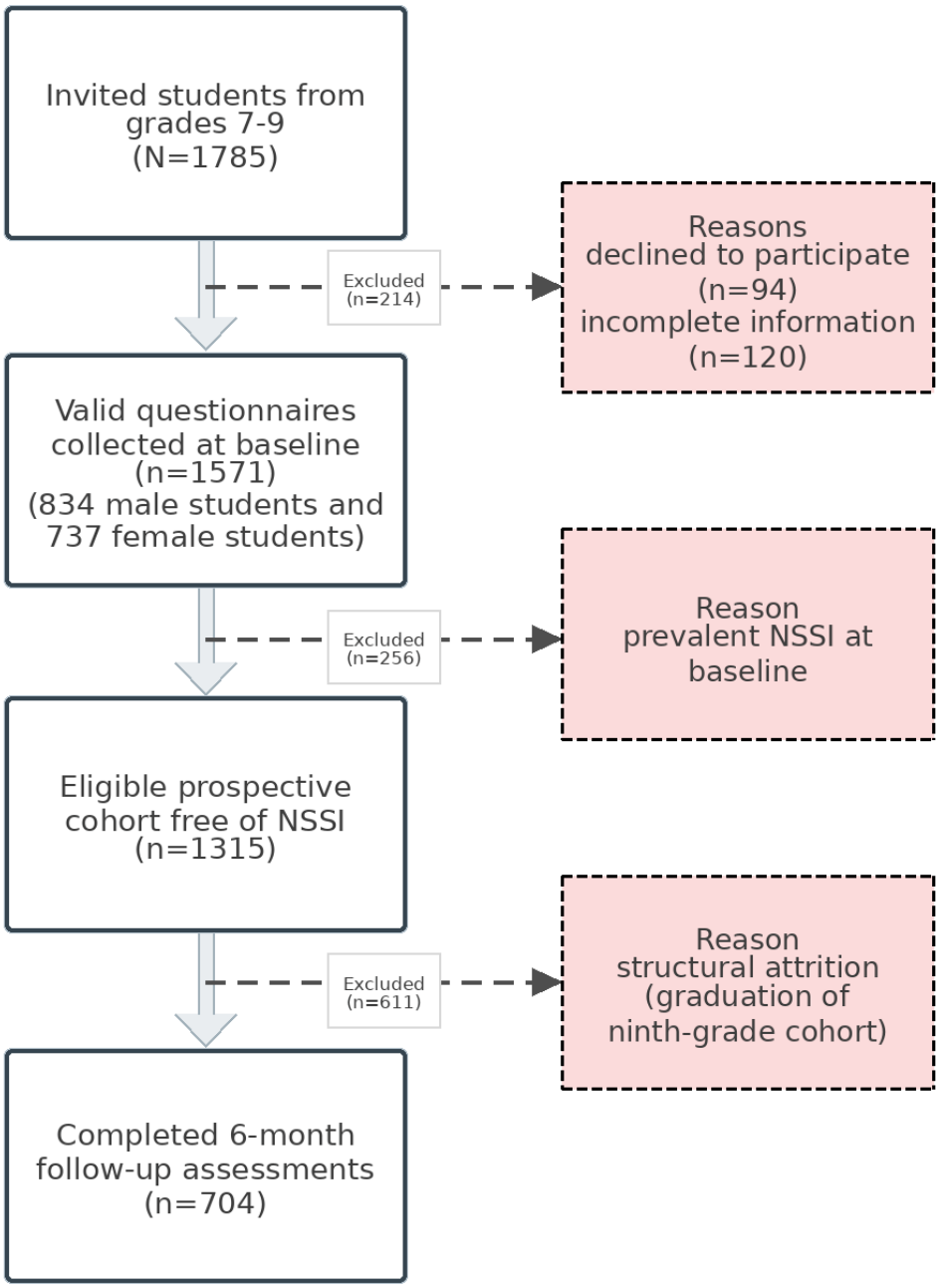

XSL•FO
RenderX

Table . Baseline demographic characteristics of the study population between groups with different internet addiction (IA) status in a prospective cohort of junior high school students (Changsha, China; November to December 2020).

	Normal (n=558)	IDS[a] (n=82)	IA (n=64)	*F* test/χ^{2b}(*df*)	*P* value
Age (y), mean (SD)	12.60 (0.61)	12.63 (0.58)	12.53 (0.76)	0.5 (2,701)	.60
Sex, n (%)				2.8 (2)	.25
Male	306 (54.8)	47 (57.3)	42 (65.6)		
Female	252 (45.2)	35 (42.7)	22 (34.4)		
Ethnicity, n (%)				11.5 (2)	.003
Han	528 (94.6)	81 (98.8)	55 (85.9)		
Other	30 (5.4)	1 (1.2)	9 (14.1)		
Only child status, n (%)				2.1 (2)	.26
Yes	233 (41.8)	32 (39.0)	20 (31.3)		
No	325 (58.2)	50 (61.0)	44 (68.8)		
Depression, n (%)				49.5 (2)	<.001
No	437 (78.3)	43 (52.4)	29 (45.3)		
Yes	121 (21.7)	39 (47.6)	35 (54.7)		
Anxiety, n (%)				36.8 (2)	<.001
No	347 (62.2)	29 (35.4)	21 (32.8)		
Yes	211 (37.8)	53 (64.6)	43 (67.2)		
Nonsuicidal self-injury, n (%)				9.9 (2)	.007
No	513 (91.9)	70 (85.4)	52 (81.3)		
Yes	45 (8.1)	12 (14.6)	12 (18.8)		

[a]IDS: internet-dependent state.

[b]*F* values were used for continuous variables, and chi-square values were used for categorical variables.

Longitudinal Association and Dose-Response Relationship

Significant ORs for NSSI were observed in IA (Table 2: models 1, 2, and 3). The crude OR of IA for NSSI was 2.631 (95% CI 1.309 - 5.286; *P*=.007), and the significance remained after controlling for age, sex, ethnicity, and only child status in model 2, with an adjusted OR of 2.587 (95% CI 1.257 - 5.321; *P*=.01). The association remained significant when further controlling for anxiety and depression in model 3, with an adjusted OR of 2.185 (95% CI 1.031 - 4.627; *P*=.04). In contrast, no significant longitudinal association was observed between the internet-dependent state and subsequent NSSI. Specifically, even in the fully adjusted model (model 3), the internet-dependent state did not significantly predict incident NSSI (adjusted OR 1.810, 95% CI 0.881 - 3.717; *P*=.11). These findings suggest that the elevated risk of incident NSSI is specifically driven by full IA rather than its intermediate state. Furthermore, in the restricted cubic spline regression models, the correlation between continuous IA severity and the risk of subsequent NSSI was displayed linearly (Figure 2).

XSL•FO
RenderX

Table . Longitudinal association between baseline internet addiction (IA) and the 6-month incidence of nonsuicidal self-injury (NSSI) among adolescents: odds ratios (ORs) and 95% CIs from a prospective cohort study (Changsha, China; 2020 - 2021).

	NSSI, n (%)	Model 1[a]		Model 2[b]		Model 3[c]	
		OR (95% CI)[d]	*P* value	OR (95% CI)[e]	*P* value	OR (95% CI)[f]	*P* value
Normal (reference)	45 (8.1)	1.00	__[g]	1.00	—	1.00	—
IDS[h]	12 (14.6)	1.95 (0.99 - 3.87)	.06	2.09 (1.04 - 4.20)	.04	1.81 (0.88 - 3.72)	.11
IA	12 (18.8)	2.63 (1.31 - 5.29)	.007	2.59 (1.26 - 5.32)	.01	2.19 (1.03 - 4.63)	.04

[a]Model 1: crude OR.

[b]Model 2: OR adjusted for sex, age, ethnicity, and only child status.

[c]Model 3: OR adjusted for sex, age, ethnicity, only child status, anxiety, and depression

[d]*P* value for trend<.001.

[e]*P* value for trend=.009.

[f]*P* value for trend=.06.

[g]Not applicable.

[h]IDS: internet-dependent state.

Figure 2. Restricted cubic spline regression analysis of the dose-response relationship between baseline internet addiction (IA) and the risk of incident nonsuicidal self-injury (NSSI) at 6-month follow-up in a prospective cohort of adolescents (Changsha, China; 2020 - 2021). The x-axis represents scores from the Chinese Internet Addiction Scale (CIAS), while the y-axis represents the odds ratio (OR) from logistic regression for the risk of new occurrences of NSSI behavior 6 months later. (A) Model A denotes the entire participant population, (B) model B represents the female cohort, and (C) model C represents the male cohort. All models were adjusted for covariates, including age, ethnicity, anxiety, and depression. Model A was further adjusted for sex.

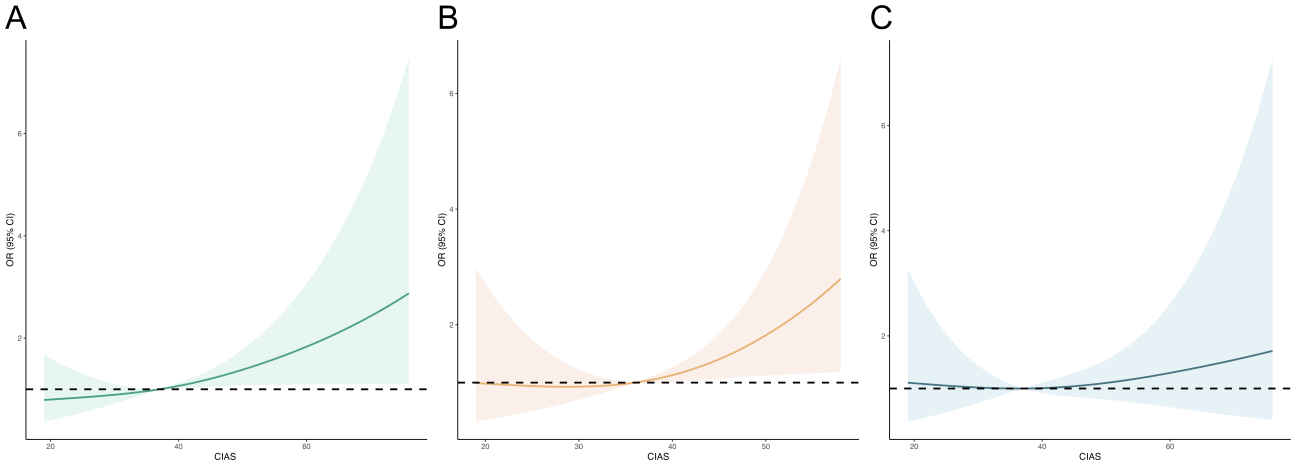

Sex-Stratified Analyses

The sex-based hierarchical model (Figure 3) did not reveal any significant OR for either the internet-dependent state or IA among male adolescents across all models (all *P*>.05). Conversely, for female adolescents, both the internet-dependent state and IA emerged as significant predictors of incident NSSI across all 3 models. Importantly, these results maintained their significance even after adjusting for age, ethnicity, only child status, anxiety, and depression (adjusted OR 3.271, 95% CI 1.101 - 9.717; *P*=.03), and the adjusted OR for the internet-dependent state was also statistically significant (adjusted OR 2.593, 95% CI 1.002 - 6.710; *P*=.049). Furthermore, it is noteworthy that in each of these models, the OR for IA surpassed that of the internet-dependent state.

XSL•FO
RenderX

Figure 3. Sex-stratified analysis of the longitudinal association between baseline internet addiction (IA) and incident nonsuicidal self-injury (NSSI) at 6-month follow-up among a prospective cohort of adolescents (Changsha, China; 2020 - 2021). Model 1: crude odds ratio (OR); model 2: OR adjusted for age, ethnicity, and only child status; and model 3: OR adjusted for age, ethnicity, only child status, anxiety, and depression. IDS: internet-dependent state.

	NSSI	Model 1		Model 2		Model 3	
		OR (95%CI)	*p*	OR (95%CI)	*p*	OR (95%CI)	*p*
Male							
Normal (Ref)	20 (6.5%)	1.00		1.00		1.00	
IDS	4 (8.5%)	1.330 (0.434-4.078)	0.618	1.304 (0.424-4.015)	0.643	1.104 (0.349-3.490)	0.866
IA	5 (11.9%)	1.932 (0.684-5.457)	0.214	1.921 (0.667-5.533)	0.226	1.552 (0.515-4.683)	0.435
p for trend		0.443		0.464		0.737	
Female							
Normal (Ref)	25 (9.9%)	1.00		1.00		1.00	
IDS	8 (22.9%)	2.690 (1.104-6.555)	0.029	2.861 (1.147-7.135)	0.024	2.593 (1.002-6.710)	0.049
IA	7 (17.5%)	4.237 (1.578-11.377)	0.004	3.647 (1.280-10.389)	0.015	3.271 (1.101-9.717)	0.033
p for trend		0.004		0.009		0.031	

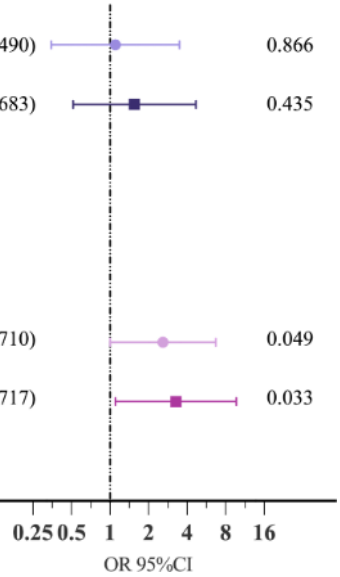

Discussion

This study is the first prospective cohort investigation to emphasize sex differences to examine the longitudinal association between IA exposure and the new onset of NSSI among Chinese adolescents. Our key findings are as follows: (1) baseline IA, but not the intermediate internet-dependent state, was a robust independent predictor of incident NSSI in the overall sample; (2) significant sex disparities were identified, with both the internet-dependent state and IA significantly associated with subsequent NSSI *only* among female adolescents, indicating a sex-specific dose-response relationship; and (3) the lack of overall internet-dependent state significance is likely due to a "sex masking" effect, highlighting the extreme vulnerability of female adolescents to even subclinical internet dependence. Together, these findings highlight the independent predictive value of IA for adolescent NSSI and underscore the critical need for sex-specific prevention strategies.

This study found that baseline IA might predict the risk of new-onset NSSI among adolescents, which is consistent with findings from a previous longitudinal design study [33]. This study used a multivariate logistic regression analysis with IA at baseline as a predictor of new self-harm or suicidal tendencies in the next year. Status classified as IA had a relative risk of new NSSI compared with those without, and this association remained significant after adjusting for several confounding variables (eg, baseline depression and anxiety levels). This suggests that while emotional distress is undoubtedly a contributor, IA appears to add a unique, incremental risk for subsequent NSSI behavior. Consistent with the experiential avoidance model and compensatory internet use theory discussed earlier, when the internet is used as an avoidant coping mechanism for emotional distress, its failure to adequately regulate intensifying negative emotions may precipitate NSSI as a more severe regulatory strategy [15]. It is worth noting that the internet-dependent state did not show a significant association with incident NSSI in the total sample across all models (all $P>.05$). This suggests a "clinical threshold" in the general adolescent population, where intermediate use is generally insufficient to precipitate NSSI, and only full IA acts as a robust independent predictor. However, significant associations were observed for the internet-dependent state across all models in the female subgroup. The overall nonsignificance likely resulted from a "sex masking" effect, concealing the extreme vulnerability of female adolescents. This suggests that the association of the internet-dependent state with NSSI may be highly sex specific. Mechanistically, because female adolescents are more prone to social networking, even moderate problematic use rapidly exposes them to upward social comparison, cyberbullying, and appearance anxiety, potent triggers for self-harm. In contrast, male adolescents predominantly use internet gaming for emotional venting, requiring a much higher addiction threshold (IA) to manifest NSSI. This highlights the critical importance of early sex-targeted monitoring when intervening with female adolescents in the at-risk internet-dependent state stage.

The sex stratification model indicated significant associations between IA and subsequent NSSI only in female adolescents with both an internet-dependent state and IA, which is inconsistent with previous findings. Surprisingly, only 2

XSL•FO
RenderX

cross-sectional studies have investigated the sex difference between IA and NSSI. A study of 11,356 adolescents in 11 European countries showed no significant sex interactions in IA and NSSI [34]. Another study enrolled 15,623 adolescents in China and showed that among adolescents aged 11 to 14 years, the ORs of IA for NSSI were greater in male adolescents (1.53, 95% CI 1.25 - 1.88) than in female adolescents (1.13, 95% CI 0.90 - 1.47) [13]. Various factors might contribute to sex differences, such as biological dissimilarities [35], differences in social status and expectations, variations in perceived social support [36], and varying susceptibilities to stress-related concerns [37]. The inconsistent results could be explained by the following aspects. First, variations in sex distribution among diverse participant samples and the complexity of defining IA may be associated with the final results. Second, the nature of internet use might play a pivotal role. Previous studies have unveiled sex-based IA disparities: male adolescents have a propensity for gaming addiction [38], which may serve as a more temporary distraction from distress rather than a direct source of negative self-evaluation, whereas female adolescents display increased vulnerability to social media–related problems [39], which fosters maladaptive social comparisons and cyberbullying victimization—both of which are potent triggers for body dissatisfaction and subsequent NSSI. Another study found that the sex differences in the association between different forms of internet use and NSSI were significant [40]. This study investigated the correlation between NSSI and IA with or without game play and found that nongamers had higher NSSI risk. The gamers were primarily male (27% female and 73% male), while the nongamers were predominantly female (78.8% female and 21.2% male). This finding may support a higher risk of NSSI in female adolescents with IA. This also suggests that we need to be more cautious in drawing conclusions when analyzing sex differences. Third, both IA and NSSI involve difficulties in emotional regulation and serve as a coping strategy [41,42], and IA in female adolescents is associated with an increased likelihood of emotional issues [43], and female adolescents are more likely to engage in rumination and internalizing coping styles when faced with online stressors. This pattern is strongly linked to both IA and NSSI, and these individuals may be more prone to adopting maladaptive coping strategies during episodes of emotional distress [44]. However, due to the current scarcity of research, further studies on sex differences are needed.

There are some limitations that need to be noted. First, the assessment of NSSI relied on a single binary item (yes or no), and while this measure was sufficient to identify the onset of the behavior in our longitudinal design**,** it constitutes a major limitation because it fails to assess its frequency, severity, methods, or functional motivations. Additionally, the data were based on retrospective self-report scales, which may introduce recall and social desirability biases. Future studies should use diagnostic interview tools to capture the multidimensional nature of NSSI and mitigate these measurement biases. Second, the attrition rate was relatively high (46.46%) due to the graduation of grade 9 students during the follow-up period. While this loss was structural, it may have excluded older adolescents from the final analysis and potentially introduced selection bias. Furthermore, the sample size for the subgroup of female adolescents with IA was relatively small (22/704, 3.1%). Although the association with incident NSSI was statistically significant, this limited sample size reduces statistical power and warrants caution in generalizing the sex-stratified findings. Third, although this study controlled for several confounding variables that might explain the association between IA and NSSI in adolescents such as depression and anxiety, it is noteworthy that other psychosocial variables such as family conflict and bullying victimization were not included. Future studies should include these measures to enhance understanding of the causal pathways.

In conclusion, our study reveals a significant association between IA and new-onset NSSI 6 months later, indicating that NSSI is prevalent in Chinese adolescents and that female adolescents with both IA and an internet-dependent state had a higher NSSI risk, which provides enhanced evidence for clinicians for the early identification and intervention of NSSI. These findings also emphasize the need for further research to better understand the mechanisms underlying this relationship between them. These findings have broader implications for public health policy, suggesting that school-based mental health screenings should integrate assessments of IA. Furthermore, intervention programs targeting NSSI prevention may benefit from sex-specific strategies that address IA as a modifiable risk factor.

Acknowledgments

The authors would like to express their sincere gratitude to all adolescents, their parents, and the schoolteachers in Changsha who actively participated in and facilitated this longitudinal survey. They also thank the administrators of the participating middle school for their tremendous support during the data collection process.

During the preparation of this manuscript, the authors used ChatGPT, developed by OpenAI, strictly for the purposes of language editing, proofreading, and improving the readability of the text. After using this tool, the authors thoroughly reviewed and edited the content as needed and took full responsibility for the final content and the integrity of the published work. No AI tools were used for data generation, statistical analysis, or drawing scientific conclusions.

Funding

This work was supported by the Postdoctoral Foundation of the Second Xiangya Hospital, Central South University.

XSL•FO
RenderX

Data Availability

The datasets generated during and/or analyzed during this study are not publicly available due to the sensitive nature of the research involving minor individuals and mental health issues but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Comparison of baseline characteristics between participants who completed the 6-month follow-up and those lost to follow-up.
[DOCX File, 18 KB - jmir_v28i1e86427_app1.docx]

Checklist 1
STROBE checklist.
[PDF File, 369 KB - jmir_v28i1e86427_app2.pdf]

References

1. Nock MK. Self-injury. Annu Rev Clin Psychol 2010;6:339-363. [doi: 10.1146/annurev.clinpsy.121208.131258] [Medline: 20192787]
2. Brown RC, Plener PL. Non-suicidal self-injury in adolescence. Curr Psychiatry Rep 2017 Mar;19(3):20. [doi: 10.1007/s11920-017-0767-9] [Medline: 28315191]
3. da Silva Bandeira BE, Dos Santos Júnior A, Dalgalarrondo P, de Azevedo RC, Celeri E. Nonsuicidal self-injury in undergraduate students: a cross-sectional study and association with suicidal behavior. Psychiatry Res 2022 Dec;318:114917. [doi: 10.1016/j.psychres.2022.114917] [Medline: 36332506]
4. Miller L, Campo JV. Depression in adolescents. N Engl J Med 2021 Jul 29;385(5):445-449. [doi: 10.1056/NEJMra2033475] [Medline: 34320289]
5. Daukantaitė D, Lundh LG, Wångby-Lundh M, et al. What happens to young adults who have engaged in self-injurious behavior as adolescents? A 10-year follow-up. Eur Child Adolesc Psychiatry 2021 Mar;30(3):475-492. [doi: 10.1007/s00787-020-01533-4] [Medline: 32318877]
6. Qian H, Shu C, Feng L, Xiang J, Guo Y, Wang G. Childhood maltreatment, stressful life events, cognitive emotion regulation strategies, and non-suicidal self-injury in adolescents and young adults with first-episode depressive disorder: direct and indirect pathways. Front Psychiatry 2022;13:838693. [doi: 10.3389/fpsyt.2022.838693] [Medline: 35492724]
7. Esposito-Smythers C, Goldstein T, Birmaher B, et al. Clinical and psychosocial correlates of non-suicidal self-injury within a sample of children and adolescents with bipolar disorder. J Affect Disord 2010 Sep;125(1-3):89-97. [doi: 10.1016/j.jad.2009.12.029] [Medline: 20089313]
8. Sabus A, Feinstein J, Romani P, Goldson E, Blackmer A. Management of self-injurious behaviors in children with neurodevelopmental disorders: a pharmacotherapy overview. Pharmacotherapy 2019 Jun;39(6):645-664. [doi: 10.1002/phar.2238] [Medline: 30793794]
9. Reichl C, Kaess M. Self-harm in the context of borderline personality disorder. Curr Opin Psychol 2021 Feb;37:139-144. [doi: 10.1016/j.copsyc.2020.12.007] [Medline: 33548678]
10. Pan YC, Chiu YC, Lin YH. Systematic review and meta-analysis of epidemiology of internet addiction. Neurosci Biobehav Rev 2020 Nov;118:612-622. [doi: 10.1016/j.neubiorev.2020.08.013] [Medline: 32853626]
11. Kardefelt-Winther D. A conceptual and methodological critique of internet addiction research: towards a model of compensatory internet use. Comput Hum Behav 2014 Feb;31:351-354. [doi: 10.1016/j.chb.2013.10.059]
12. Evli M, Şimşek N, Işıkgöz M, Öztürk H. Internet addiction, insomnia, and violence tendency in adolescents. Int J Soc Psychiatry 2023 Mar;69(2):351-361. [doi: 10.1177/00207640221090964] [Medline: 35470724]
13. Tang J, Ma Y, Lewis SP, et al. Association of internet addiction with nonsuicidal self-injury among adolescents in China. JAMA Netw Open 2020 Jun 1;3(6):e206863. [doi: 10.1001/jamanetworkopen.2020.6863] [Medline: 32496567]
14. Chen CY, Chang SL. Moderating effects of information-oriented versus escapism-oriented motivations on the relationship between psychological well-being and problematic use of video game live-streaming services. J Behav Addict 2019 Sep 1;8(3):564-573. [doi: 10.1556/2006.8.2019.34] [Medline: 31328955]
15. Plener PL, Kaess M, Schmahl C, Pollak S, Fegert JM, Brown RC. Nonsuicidal self-injury in adolescents. Dtsch Arztebl Int 2018 Jan 19;115(3):23-30. [doi: 10.3238/arztebl.2018.0023] [Medline: 29366448]
16. Chapman AL, Gratz KL, Brown MZ. Solving the puzzle of deliberate self-harm: the experiential avoidance model. Behav Res Ther 2006 Mar;44(3):371-394. [doi: 10.1016/j.brat.2005.03.005] [Medline: 16446150]

XSL•FO
RenderX

17. Chen Y, Liang H, Jiang H. Cybervictimization and non-suicidal self-injury among Chinese college students: an exploration through the lens of experiential avoidance model. Psychol Res Behav Manag 2024;17:3133-3145. [doi: 10.2147/PRBM.S476657] [Medline: 39280036]
18. Fan YY, Liu J, Zeng YY, Conrad R, Tang YL. Factors associated with non-suicidal self-injury in Chinese adolescents: a meta-analysis. Front Psychiatry 2021;12:747031. [doi: 10.3389/fpsyt.2021.747031] [Medline: 34916971]
19. Peng P, Chen Z, Ren S, et al. Internet gaming disorder predicts the incidence, persistence, and worsening of suicidal ideation: a population-based cohort study of 96,158 Chinese adolescents. J Affect Disord 2025 Jun 15;379:186-193. [doi: 10.1016/j.jad.2025.03.029] [Medline: 40081581]
20. Peng P, Chen ZM, Ren SL, et al. Internet gaming disorder and depression mediated by impaired resilience and sleep distress: a three-wave longitudinal study among Chinese adolescents. Epidemiol Psychiatr Sci 2025 Feb 19;34:e11. [doi: 10.1017/S2045796025000046] [Medline: 39965936]
21. Peng P, Chen Z, Ren S, et al. Prospective association of internet gaming disorder with subsequent first suicidal attempt: a large-scale school-based study of Chinese adolescents. Prev Med 2025 Oct;199:108396. [doi: 10.1016/j.ypmed.2025.108396] [Medline: 40882707]
22. Peng C, Guo T, Cheng J, et al. Sex differences in association between internet addiction and aggression among adolescents aged 12 to 18 in mainland of China. J Affect Disord 2022 Sep 1;312:198-207. [doi: 10.1016/j.jad.2022.06.026] [Medline: 35728679]
23. Wilkinson PO, Qiu T, Jesmont C, et al. Age and gender effects on non-suicidal self-injury, and their interplay with psychological distress. J Affect Disord 2022 Jun 1;306:240-245. [doi: 10.1016/j.jad.2022.03.021] [Medline: 35304237]
24. Petry NM, Zajac K, Ginley MK. Behavioral addictions as mental disorders: to be or not to be? Annu Rev Clin Psychol 2018 May 7;14:399-423. [doi: 10.1146/annurev-clinpsy-032816-045120] [Medline: 29734827]
25. Fox KR, Millner AJ, Mukerji CE, Nock MK. Examining the role of sex in self-injurious thoughts and behaviors. Clin Psychol Rev 2018 Dec;66:3-11. [doi: 10.1016/j.cpr.2017.09.009] [Medline: 28993102]
26. Chen SH, Weng LJ, Su YJ, Wu HM, Yang PF. Development of a Chinese Internet Addiction Scale and its psychometric study. Chin J Psychol 2003;45(3):279-294 [FREE Full text]
27. Bai Y, Fan FM. A study on the internet dependence of college students: the revising and applying of a measurement. Psychol Dev Educ 2005;21(99):104 [FREE Full text]
28. Xu H, Song X, Wang S, Zhang S, Xu S, Wan Y. Mediating effect of social support in the relationship between childhood abuse and non-suicidal self-injury among Chinese undergraduates: the role of only-child status. Int J Environ Res Public Health 2019 Oct 21;16(20):4023. [doi: 10.3390/ijerph16204023] [Medline: 31640165]
29. Wang MC, Yi J, Cai L, et al. Reliability and validity of the adolescent health related risky behavior inventory. Chin Ment Health J 2012;26(4):287-292. [doi: 10.3969/j.issn.1000-6729.2012.04.010]
30. Wan Y, Chen R, Ma S, et al. Associations of adverse childhood experiences and social support with self-injurious behaviour and suicidality in adolescents. Br J Psychiatry 2019 Mar;214(3):146-152. [doi: 10.1192/bjp.2018.263] [Medline: 30477603]
31. Tonsing KN. Psychometric properties and validation of Nepali version of the Depression Anxiety Stress Scales (DASS-21). Asian J Psychiatr 2014 Apr;8:63-66. [doi: 10.1016/j.ajp.2013.11.001] [Medline: 24655630]
32. Li XL, Tang HB, Guo F, et al. A review on the reliability and validity of the Depression Anxiety Stress Scale. Zhongguo Lin Chuang Xin Li Xue Za Zhi 2012;20(3) [FREE Full text]
33. Pan PY, Yeh CB. Internet addiction among adolescents may predict self-harm/suicidal behavior: a prospective study. J Pediatr 2018 Jun;197:262-267. [doi: 10.1016/j.jpeds.2018.01.046] [Medline: 29550226]
34. Kaess M, Durkee T, Brunner R, et al. Pathological internet use among European adolescents: psychopathology and self-destructive behaviours. Eur Child Adolesc Psychiatry 2014 Nov;23(11):1093-1102. [doi: 10.1007/s00787-014-0562-7] [Medline: 24888750]
35. Zeng N, Wang M, Zheng H, et al. Gender-related differences in frontal-parietal modular segregation and altered effective connectivity in internet gaming disorder. J Behav Addict 2021 Mar 10;10(1):123-134. [doi: 10.1556/2006.2021.00015] [Medline: 33704084]
36. Du G, Lyu H. Future expectations and internet addiction among adolescents: the roles of intolerance of uncertainty and perceived social support. Front Psychiatry 2021;12:727106. [doi: 10.3389/fpsyt.2021.727106] [Medline: 34512423]
37. Albertella L, Rotaru K, Christensen E, et al. The influence of trait compulsivity and impulsivity on addictive and compulsive behaviors during COVID-19. Front Psychiatry 2021;12:634583. [doi: 10.3389/fpsyt.2021.634583] [Medline: 33708147]
38. Meng SQ, Cheng JL, Li YY, et al. Global prevalence of digital addiction in general population: a systematic review and meta-analysis. Clin Psychol Rev 2022 Mar;92:102128. [doi: 10.1016/j.cpr.2022.102128] [Medline: 35150965]
39. Coyne SM, Weinstein E, Sheppard JA, et al. Analysis of social media use, mental health, and gender identity among US youths. JAMA Netw Open 2023 Jul 3;6(7):e2324389. [doi: 10.1001/jamanetworkopen.2023.24389] [Medline: 37486631]
40. Strittmatter E, Kaess M, Parzer P, et al. Pathological internet use among adolescents: comparing gamers and non-gamers. Psychiatry Res 2015 Jul 30;228(1):128-135. [doi: 10.1016/j.psychres.2015.04.029] [Medline: 25959265]
41. Taylor PJ, Jomar K, Dhingra K, Forrester R, Shahmalak U, Dickson JM. A meta-analysis of the prevalence of different functions of non-suicidal self-injury. J Affect Disord 2018 Feb;227:759-769. [doi: 10.1016/j.jad.2017.11.073] [Medline: 29689691]

XSL•FO
RenderX

42. McNicol ML, Thorsteinsson EB. Internet addiction, psychological distress, and coping responses among adolescents and adults. Cyberpsychol Behav Soc Netw 2017 May;20(5):296-304. [doi: 10.1089/cyber.2016.0669] [Medline: 28414517]
43. Shen Y, Wang L, Huang C, et al. Sex differences in prevalence, risk factors and clinical correlates of internet addiction among Chinese college students. J Affect Disord 2021 Jan 15;279:680-686. [doi: 10.1016/j.jad.2020.10.054] [Medline: 33190119]
44. Compas BE, Jaser SS, Bettis AH, et al. Coping, emotion regulation, and psychopathology in childhood and adolescence: a meta-analysis and narrative review. Psychol Bull 2017 Sep;143(9):939-991. [doi: 10.1037/bul0000110] [Medline: 28616996]

Abbreviations

CIAS-R: Chinese Internet Addiction Scale-Revised
DASS-21: Depression Anxiety Stress Scale
IA: internet addiction
NSSI: nonsuicidal self-injury
OR: odds ratio
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani; submitted 24.Oct.2025; peer-reviewed by X Wang, Y Liao, Y Liang; revised version received 10.Mar.2026; accepted 03.Apr.2026; published 17.Jun.2026.

Please cite as:
Fan T, Song L, Li S, Jing X, Twayigira M, Huang C, Luo X, Shen Y
Examining the Association Between Internet Addiction and Nonsuicidal Self-Injury Among Chinese Middle School Students: Prospective Cohort Study
J Med Internet Res 2026;28:e86427
URL: https://www.jmir.org/2026/1/e86427
doi:10.2196/86427

© Tianqing Fan, Lintong Song, Sihong Li, Xingyue Jing, Mireille Twayigira, Chunxiang Huang, Xuerong Luo, Yanmei Shen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Promoting Psychological Resilience and Well-Being in Youth With a Smartphone-Based Ecological Momentary mHealth Intervention: Secondary Analysis of a Microrandomized Trial

Eva Wierzba[1], MSc; Anita Schick[1], Dipl (Psych), PhD; Christian Rauschenberg[1,2], MSc, PhD; Janik Fechtelpeter[3,4,5,6], MSc; Selina Hiller[7,8], MSc; Christian Götzl[9], MA; Daniel Durstewitz[3,10], MSc, PhD; Silvia Krumm[11], MA, PhD; Georgia Koppe[4,5,6,12], MSc, Dipl (Psych), PhD; Ulrich Reininghaus[1,2,13,14], MSc, Dipl (Psych), PhD

[1]Department of Public Mental Health, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, J5, Mannheim, Germany

[10]Interdisciplinary Center for Scientific Computing, Heidelberg University, Heidelberg, Germany

[11]Department of Psychiatry and Psychotherapy, Medical Faculty, Leipzig University, Leipzig, Germany

[12]Hertie Institute for AI (Artificial Intelligence) in Brain Health, University of Tübingen, Tübingen, Germany

[13]Health Service and Population Research Department, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom

[14]ESRC (Economic and Social Research Council) Centre for Society and Mental Health, King's College London, London, United Kingdom

[2]German Center for Mental Health (DZPG), Partner Site Mannheim-Heidelberg-Ulm, Mannheim, Germany

[3]Department of Theoretical Neuroscience, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

[4]Hector Institute for Artificial Intelligence in Psychiatry, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Heidelberg, Germany

[5]Department of Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

[6]RG (Research Group) Machine Learning Human Behavior, Interdisciplinary Center for Scientific Computing, Heidelberg University, Heidelberg, Germany

[7]Department of Psychiatry and Psychotherapy, School of Medicine and Health, Technical University of Munich, TUM University Hospital, Munich, Germany

[8]German Center for Mental Health (DZPG), Partner Site Munich-Augsburg, Munich, Germany

[9]Department of Forensic Psychiatry and Psychotherapy, University of Ulm and BKH Guenzburg, Ulm, Germany

Corresponding Author:
Ulrich Reininghaus, MSc, Dipl (Psych), PhD
Department of Public Mental Health, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, J5, Mannheim, Germany

Abstract

Background: Ecological momentary interventions (EMIs) may be a promising tool for promoting mental well-being in youth, as they allow for targeting resilience and other protective factors in daily life.

Objective: In this secondary analysis of a microrandomized trial, we explored proximal effects of a coach-guided compassion-focused smartphone-based digital training for young people applying principles and techniques of EMI (the AI4U [artificial intelligence for personalized digital mental health promotion in youth] training) on momentary outcomes of mental well-being.

Methods: A convenience sample of participants aged 14 to 25 years was recruited mostly via an open-access website and completed up to 6 self-report ecological momentary assessments (EMA) per day to measure their momentary mental well-being (ie, positive affect, negative affect, and stress) during the 30-day AI4U training phase. Some EMA prompts triggered EMI components to promote their momentary resilience and well-being. Multilevel modeling was used to analyze proximal effects of initiating EMI components on outcomes assessed at the next time point, the potential moderating effect of momentary affect and stress at the time of EMI initiation, and the potential mediating effect of change in momentary resilience on the effect of EMI initiation on changes in momentary affect and stress.

Results: A total of 170 individuals completed 13,059 EMA prompts and initiated 6667 EMI components. No evidence was found that momentary outcomes of mental well-being at a time point differed depending on whether an EMI component was initiated at the previous time point vs when no EMI component was initiated at the previous time point (positive affect: b=0.00, 95% CI −0.04 to 0.04; negative affect: b=−0.03, 95% CI −0.06 to 0.01; stress: b=0.01, 95% CI −0.03 to 0.05). There was no strong difference in the magnitude of this effect when momentary mental well-being at the time of EMI initiation was high vs

XSL•FO
RenderX

low (positive affect: b=0.08, 95% CI 0.01 to 0.16; negative affect: b=−0.09, 95% CI −0.17 to −0.01; stress: b=0.09, 95% CI 0.01 to 0.17). Changes in momentary resilience did not mediate the effect of EMI component initiation (vs EMI component noninitiation) on changes in momentary mental well-being (positive affect: b=0.00, 95% CI −0.01 to 0.02; negative affect: b=−0.00, 95% CI −0.01 to 0.01; stress: b=0.00; 95% CI 0.00 to 0.00).

Conclusions: By investigating proximal effects of EMI components, this secondary analysis analyzes proximal effects of a novel digital training using EMA, which presents an innovative approach to understanding how the digital training leads to long-term improvement of distal outcomes. The analysis contributes to research in the field of EMI by serving as a basis for future investigations on the momentary effects of EMI components, which can support the development of scalable interventions to promote mental well-being in the public.

(J Med Internet Res 2026;28:e85552) doi:10.2196/85552

KEYWORDS

ecological momentary assessment; experience sampling method; ambulatory assessment; ecological momentary intervention; just-in-time adaptive intervention; digital training; mobile health; digital health; mental health; affect; adolescent; mobile phone

Introduction

Background and Rationale

Young individuals are a vulnerable group at risk for mental health conditions, as many mental health conditions first appear in childhood or adolescence [1,2] and contribute significantly to the disease burden of young people [3]. In the past years, young individuals have also been facing a multitude of crises, such as the COVID-19 pandemic and the climate crisis, which are associated with negative mental health outcomes [4]. For instance, a systematic review showed that the prevalence of depressive and anxiety symptoms among children and adolescents increased during the COVID-19 pandemic compared to prepandemic numbers [5]. Hence, there is an urgent need for easily accessible measures of mental health promotion in youth.

A promising approach involves targeting psychological resilience, as numerous studies have demonstrated a strong association between higher levels of resilience and improved mental health and well-being [6,7]. Psychological resilience can be defined as the ability to deal with adversity in different areas of life while maintaining focus on one's personal goals [8,9]. It is therefore an especially important capability to protect mental health in times of crisis. Different approaches exist to increase resilience, such as interventions based on cognitive behavioral therapy (CBT) or mindfulness [10]. However, these interventions are often associated with high costs and time investment [10]. In addition, young people can be hard to reach for interventions aimed at improving their mental well-being due to, for example, stigma or low mental health literacy [11].

Digital interventions have the potential to increase the accessibility of mental health services for young people [12] and to reduce costs [13]. A review showed that digital resilience programs have positive effects on, for example, psychological distress and positive mental health in nonclinical adult populations [14]. A specific type of digital intervention are ecological momentary interventions (EMIs), which deliver intervention components in real time via digital platforms and apps and can therefore be directly incorporated in everyday life [15-17]. They can be tailored to the specific needs of an individual and to the requirements of a given context [18,19] based on intensive longitudinal data collected using ecological momentary assessment (EMA) [20-23].

The AI4U (artificial intelligence for personalized digital mental health promotion in youth) training, a digital mHealth (mobile health) training that was developed as part of the living lab AI4U applies principles of both EMA and EMI. The content and design of the EMI components were informed by EMIs implemented in previous studies, that is, SELFIE [21,24], EMIcompass [20,25], and PerPAIN [26]. The EMI components consist of short exercises that are based on positive refocusing, positive imagery, mindfulness, and compassion-focused techniques, such as breathing exercises and a journal of joyful moments (for a full description of the EMI components, see Multimedia Appendix 1). The components aim to improve resilience, emotion regulation, and a compassionate self-image, and thereby to promote mental health and well-being. Findings from the primary analysis of the living lab AI4U suggest beneficial effects of the AI4U training on distal indicators of mental health and well-being at post intervention (compared to baseline), that is, reduced psychological distress, improved resilience, and increased use of adaptive emotion regulation strategies [27].

EMIs have been found to be effective in the promotion of distal indicators of mental health in other research: a review found that EMI based on CBT can improve mental well-being and reduce symptoms of mental health conditions in both healthy and clinical samples [28]. Another review suggested that EMIs targeting outcomes of mental health and positive psychological well-being have small to medium-sized effects, but that the low quality of the available studies requires more rigorous evidence generated by randomized controlled trials (RCTs) [29]. A recent RCT has shown that a mindfulness-based EMI led to a reduction in anxiety and depressive symptoms in individuals with generalized anxiety disorder after 2 weeks of training [30]. In another RCT, participants' self-esteem after using the SELFIE intervention for 6 weeks was, on average, higher both at postintervention and at 6-month follow-up compared to the control condition that received care as usual [21]. An EMI with compassion-focused exercises improved quality of life and stress-reactivity after 6 weeks of training in a feasibility RCT [20].

XSL•FO
RenderX

Despite the evidence on the effects of EMIs for improving outcomes of mental health and psychological well-being, so far, it is unclear how exactly EMIs yield these beneficial effects. Understanding the exact mechanisms through which EMIs work would offer the opportunity to further adapt EMIs to increase their efficacy. One way to elucidate this further is to shed light on the immediate effects of individual EMI components, often referred to as proximal effects [31]. Proximal effects of EMIs were reported in previous studies: an EMI targeted to reduce smoking had a proximal effect on negative affect [32]. Furthermore, the results of a process evaluation of the EMIcompass intervention indicated that strengthening the soothing system, which may reflect resilience [33], may have accounted for the initial signals of reduced stress reactivity and improved quality of life that were observed in the feasibility RCT [20,33]. In a subsample of participants with early mental health problems, initial signals for proximal effects of the EMIcompass intervention on negative affect, stress, and stress reactivity were found [34].

In the AI4U training, EMI components that aim at improving resilience are easily integrated into daily life to facilitate the ecological translation of the strategies learned. Building on previous research [20,32,33], completing EMI components in the moment is intended to yield a proximal effect on momentary mental well-being, such as affect and stress. Specifically, executing strategies that aim at improving resilience may help individuals to mitigate their negative affect, to amplify their positive affect, and to reduce their perceived stress [6,7,34]. As resilience can be defined as the ability to deal with and overcome adversity [8,9], applying EMI components in the moment may be posited to be beneficial in adverse situations when individuals need to cope with negative feelings or stress.

Objectives

The aim of this secondary analysis of data collected as part of the AI4U living lab was to investigate proximal effects of initiating EMI components, that is, effects of EMI components on outcomes assessed at the next time point using EMA in youths. It was expected that initiating an EMI component, which means applying strategies aimed toward improving resilience, would improve momentary indicators of mental well-being, that is, positive affect, negative affect, and stress at the subsequent EMA time point. We assume that initiating an EMI component aimed at improving resilience is especially useful in situations that require the individual to deal with adversity. Therefore, the effect of the initiation of an EMI component was expected to be greater when mental well-being at the time of initiating the EMI was low compared to when it was high. In addition, it was assumed that the effect of initiating EMI components on mental well-being would be partially mediated by an increase in momentary resilience following the initiation of the EMI. Therefore, three hypotheses were tested.

First, it was hypothesized that (1) positive affect at t_n will be higher following the initiation of an EMI component at t_{n-1} compared with noninitiation of an EMI component at t_{n-1}, (2) negative affect, and (3) perceived stress at t_n will be lower following initiation of an EMI component at t_{n-1} compared with noninitiation of an EMI component at t_{n-1}.

Second, it was hypothesized that the magnitude of this difference in (1) positive affect, (2) negative affect, and (3) perceived stress between initiation and noninitiation of EMI components at t_{n-1} will be modified by levels of (1) positive affect, (2) negative affect, and (3) perceived stress at t_{n-1} such that the magnitude of this difference will be greater following (1) low levels of positive affect, (2) high levels of negative affect, and (3) high levels of perceived stress at t_{n-1} compared with (1) high levels of positive affect, (2) low levels of negative affect, and (3) low levels of perceived stress at t_{n-1}.

Third, the effect of initiating an EMI component at t_{n-1} on (1) positive affect, (2) negative affect, and (3) stress at t_n was hypothesized to be mediated by an increase in momentary resilience from t_{n-1} to t_n, while controlling for (1) positive affect, (2) negative affect, and (3) stress at t_{n-1}.

Methods

Participants and Public Involvement

A convenience sample of participants with a target size of 180 was recruited from the general population and educational counseling services in Germany through a combination of methods, including a call for participation posted on this project's website and information events held in community settings, such as youth clubs. In addition, after receiving an information workshop on the AI4U training, educational counselors were encouraged to suggest participating in this study to their clients and to integrate their clients' experiences with the AI4U training in their counseling sessions. With the living lab AI4U, we aimed to incorporate the views of young people through the entire research process and therefore used a peer researcher and four coresearchers from the target population. They actively participated in the design, implementation, and dissemination of this study and especially contributed to the qualitative research conducted as part of the AI4U living lab, which is published elsewhere [35].

Study Design

The living lab AI4U carried out four transdisciplinary projects involving direct participation of relevant stakeholders, users from the target population, and an interdisciplinary research group. The current study draws on data collected using EMA during three within-subject microrandomized trials (MRT) [36,37] conducted consecutively as part of the second transdisciplinary AI4U project to investigate the effects of applying machine learning (ML) algorithms to deliver EMI components [27]. Findings on the primary outcome of these MRTs and other transdisciplinary projects of the living lab AI4U are published elsewhere [27,35,38,39]. These secondary analyses were preregistered on October 25, 2024, at the Open Science Framework [40] after data collection was completed (June 2024) but before accessing the data (November 2024). Deviations from the preregistration can be found in Multimedia Appendix 2. For reporting this secondary analysis of an MRT, we used the CONSORT (Consolidated Standards of Reporting Trials)

XSL•FO
RenderX

reporting checklist [41] and the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist [42], with the latter being included as Checklist 1.

Each participant took part in a 90-minute training session with a coach, that is, a trained study team member who provided guidance on the content and structure of the training before starting the 40-day AI4U training, consisting of a 10-day introductory phase and a 30-day training phase, on a study smartphone. The AI4U training consisted of an EMI administered using a smartphone-based app (movisensXS app, version 1.6.2-beta.4, movisens GmbH) for adaptive real-time and real-world transfer of mental health promotion principles into daily life. During the AI4U training, participants had continuous access to the web-based AI4U dashboard, which displayed summary statistics of the collected EMA data in various charts and tables, and allowed them to self-monitor assessed EMA variables in real time. Members of this study's team contacted each participant up to two times via phone during the training to address questions and encourage consistent app use. After the AI4U training, participants were invited to a final review session. Data on distal outcomes were collected at baseline and at post intervention. Full details on the procedure and on distal mental health outcomes can be found elsewhere [27].

Study Setting

The data collection took place at the Central Institute of Mental Health in Mannheim, Germany. Participants were recruited from the general population across Germany and from 5 different public educational counseling services in Baden-Württemberg (Southern Germany).

Eligibility Criteria

The inclusion criteria were being aged between 14 and 25 years, willingness to participate in the EMI, and ability to give written informed consent independently or through a legal guardian. The exclusion criteria were insufficient command of German, self-reported current diagnosis of any mental health condition, self-reported current psychiatric or psychotherapeutic treatment, and acute suicidality.

All study team members were required to hold at least a bachelor of science in psychology or a related field.

Ethical Considerations

This study was approved by the Medical Ethics Review Committee II at Heidelberg University (Medical Faculty Mannheim; Ref. No. 2022‑550). Before participation, participants or, for underage participants, their legal guardians gave written informed consent to using the collected data in pseudonymized form for research purposes (see Multimedia Appendix 3). Participation in this study was voluntary, and consent could be withdrawn at any time. Collected data were treated confidentially and were analyzed in a pseudonymized form. Pseudonymized data were stored and processed in accordance with the General Data Protection Regulation. Participants were compensated with €70 (US $81.52) to €120 (US $139.76) for their time, depending on how many times they interacted with the AI4U training. No identification of individual participants in any part of the manuscript or supplementary material is possible.

Intervention

The AI4U training consisted of a 10-day introductory phase, during which the different EMI components were successively introduced, and a subsequent 30-day training phase. Proximal outcomes were assessed using EMA. This study focuses on the proximal effects of EMI component initiation vs noninitiation on momentary indicators of well-being during the training phase. The main findings of the primary analysis are described elsewhere [27].

The EMI components based on positive refocusing, positive imagery, mindfulness, and compassion-focused techniques consist of four types of exercises: breathing exercises ("counting your breath" and "breathing with breaks"), compass of emotions, positive imagery techniques ("my calm and safe place," "my compassionate companion," and "emotion as a wave"), and positive refocusing ("journal of joyful moments" and "positive data log"). A detailed description of the EMI components can be found in Multimedia Appendix 1.

Three types of delivery schemes were used for the delivery of EMI components: enhancing, consolidating, and interactive/adaptive delivery of EMI components, as previously used in other EMIs [24-26,43]. The enhanced delivery of EMI components introduces participants to new EMI components during the introductory phase. The consolidating delivery allows participants to practice previously learned components at user-defined times, both during the introductory phase and the training phase. In addition, during the training phase, the interactive/adaptive delivery scheme provided EMI components after an EMA was completed at a random time. The interactive/adaptive EMI components and the consolidating EMI components during the training phase were provided using repeated randomization with equal probability to either the experimental condition, where one of the seven EMI components was assigned using an ML-algorithm based on EMA ratings of affect and behavior, or the active control condition, where the EMI component was randomly assigned [27,39]. Participants were repeatedly randomized to one of these two conditions up to seven times daily (once for consolidating delivery and six times for interactive/adaptive delivery), resulting in a maximum of 210 decision points. The type of assignment was recorded (ie, random, ML-based, or erroneous), with erroneous occurring in cases where, for example, a technical issue prevented the mobile connection to the server, making an ML-based assignment impossible. While the effect of random vs ML-based assignment was analyzed in the primary analyses, this study only focused on the effect of initiation of EMI components independently of the type of assignment of EMI components. Hence, we controlled for randomization to ML-based assignment of EMI components or random assignment of EMI components in all analyses. As in line with the Multiphase Optimization Strategy [44], the EMI and ML algorithm were optimized after each of the three MRTs; we controlled for the MRT that the participant was assigned to. The optimizations made to the EMI between the MRTs mainly related to the timing of EMA and

XSL•FO
RenderX

EMI prompts: there were at least 45 minutes (in MRT 1) or 90 minutes (in MRT 2 and MRT 3) between two EMA prompts and, thus, the subsequent assignment of interactive/adaptive EMI components. These components could be postponed by the participants by up to 20 minutes (in MRT 1) or up to 300 minutes (in MRT 2 and MRT 3). Participants were able to start the EMI component within 50 seconds after it was suggested. When participants did not start the suggested EMI component, they were reminded to do so by an alarm up to 5 times (in MRT 1) or 3 times (in MRT 2 and MRT 3). The consolidating delivery of EMI components occurred twice per day during the introductory phase and once per day during the training phase at a user-defined time and could be postponed by up to 65 minutes (in MRT 1) or 600 minutes (in MRT 2 and MRT 3). In MRT 2 and MRT 3, participants were able to request ML-based assignment of consolidating delivery of EMI components at any time during the day by button press.

Measures

Sociodemographic data and baseline characteristics were collected with online questionnaires. They included age (in years), gender (male, female, and diverse or not specified), migration history, and psychological distress at baseline, measured by calculating the sum score of the Kessler Psychological Distress Scale (K10) [45].

Outcomes

During the training phase, EMA included six assessments per day, scheduled at random within set blocks of time at user-defined time periods. Participants received a signal (sound and/or vibration) to complete the EMA. For this analysis, an EMA was considered completed when participants completed the items measuring positive affect, negative affect, stress, and resilience. Momentary positive affect was measured with the following three items: “I feel good,” “I feel relaxed,” or “I feel satisfied.” For momentary negative affect, the three items “I feel scared,” “I feel down,” and “I feel sad” were rated. The mean of the respective items was calculated for positive affect and negative affect, respectively. If any items were missing, the mean was considered missing. Momentary stress was measured using the single item “I feel stressed,” and momentary resilience was measured with the single item “I can handle all the difficulties that I may encounter.” All items were rated on a 7-point Likert scale ranging from not at all (rating of 1) to very much (rating of 7). The primary outcome of this secondary analysis was the level of momentary positive affect, negative affect, and stress at a time point t_n in relation to a previous time point t_{n-1}, measured using EMA, depending on whether an EMI component was initiated or not initiated by the participant at t_{n-1}.

Harms

Serious adverse events were nonsystematically assessed in the final review session after the AI4U training by asking the participants to describe their experience with the training.

Sample Size

The required sample size for this secondary analysis was determined based on a study which showed that, to detect a mediation effect with a small effect size (ie, an indirect effect explains at least 4% of the total variance in the outcome variable), the required number of clusters is at least 100 with cluster sizes ranging from 80 to 160 (with power at 80% and $P<.05$) [46]. To detect a medium effect (ie, Cohen $d=0.5$), only 50 clusters would be required (with power at 80% and $P<.05$). In this study, the target size was 180 clusters (ie, participants) with cluster sizes up to 210.

Randomization and Blinding

Each participant received the same EMI and could decide to initiate or not to initiate an EMI component every time it was suggested automatically by the app after completing an EMA, creating a quasi-experimental “within-subject control condition” (ie, occasions where an EMI component was suggested vs occasions when no EMI component was suggested). Therefore, no randomization of participants to initiation vs noninitiation of an EMI component was performed. This study’s personnel were not blinded.

Statistical Methods

For the analysis, R (version 4.5.1; R Foundation) was used. To fit linear mixed-effects models, the packages *lme4* (version 1.1.37) [47] and *nlme* (version 3.1.168) [48] were used. For all analyses, the available data were included, and no imputation methods were applied to account for missing data. Little missing completely at random test using the package *naniar* (version 1.1.0) [49] indicated that data were not missing completely at random ($\chi^2_{65}=4729$, $P<.001$). We assumed that data were missing at random and therefore used restricted maximum likelihood to estimate our models. Generalized logistic mixed models were fitted to identify covariates associated with missingness, which were included in the models.

All level-1 predictor variables (ie, positive affect, negative affect, stress, and resilience) were centered using person-mean centering and z-standardized. Lagged variables for positive affect, negative affect, and stress were calculated for each time point an EMA prompt was completed, using data from the most recently completed EMA prompt before that time point within the same day. To account for the temporal dependence of the data, within-person residual autocorrelation was modeled using an AR(1) structure for the first and second hypotheses.

To test the first hypothesis, linear mixed-effects models were used to account for the nested structure of EMA data (ie, time points nested within participants). Each model was controlled for age, gender, type of assignment of EMI components, psychological distress at baseline, and assigned MRT. To facilitate the interpretation of results, Cohen *d*-type effect sizes were calculated by dividing the beta coefficient by the square root of the total variance of the model. For the first hypothesis, three separate models were fitted, with the outcome variable well-being at t_n being operationalized as (1) positive affect, (2) negative affect, and (3) stress, respectively. The independent variable was the dichotomous variable of whether an EMI component was initiated or not at t_{n-1}. It was controlled for (1) positive affect at t_{n-1}, (2) negative affect at t_{n-1}, or (3) stress at t_{n-1}. Random intercepts for individuals were assumed for (1)

XSL•FO
RenderX

positive affect at t_{n-1}, (2) negative affect at t_{n-1}, and (3) stress at t_{n-1} to allow for interindividual differences in the level of well-being. Random slopes for the effects of well-being at t_{n-1} and initiation of an EMI component at t_{n-1} were initially assumed, but the models were revised to exclude the random slope for well-being at t_{n-1} due to convergence issues.

For the second hypothesis, a moderator variable was added to the model. The moderator was the lagged well-being variable (1) positive affect at t_{n-1}, (2) negative affect at t_{n-1}, or (3) stress at t_{n-1}. It was included in the model as a two-way interaction term with the dichotomous variable indicating whether an EMI component was initiated at t_{n-1}. Random intercepts for the individuals were assumed for (1) positive affect at t_{n-1}, (2) negative affect at t_{n-1}, and (3) stress at t_{n-1} to allow for interindividual differences in the level of well-being. In addition, random slopes for the effects of well-being at t_{n-1} and initiation of an EMI component at t_{n-1}, and the interaction of well-being at t_{n-1} and initiation of an EMI component at t_{n-1} were initially assumed, but due to convergence issues, the models were revised, including only random slopes for initiation of an EMI component at t_{n-1}. To investigate whether there were differences in the magnitude of the effect of EMI initiation at t_{n-1} on well-being at t_n within occasions with high (mean+1 SD) vs low (mean −1 SD) levels of well-being at t_{n-1}, linear combinations of coefficients were calculated using the R package *multcomp* [50], with well-being indicators entered as continuous variables in the model.

For the third hypothesis, three mediation models were fitted with difference scores ($t_n - t_{n-1}$) for the well-being variables (1) positive affect, (2) negative affect, and (3) stress as the outcome variables and the difference score ($t_n - t_{n-1}$) for resilience as the mediator. To examine the mediation effects, a two-step multilevel modeling approach was used: first, a linear mixed-effects model was fitted to predict the mediator (ie, the change in resilience) depending on whether an EMI component was initiated at t_{n-1}. Second, a linear mixed-effects model was fitted to predict the outcome (ie, the change in well-being) depending on the initiation of an EMI component at t_{n-1} and the difference score for resilience. The initial model specification included random slopes for both the mediator and the outcome model. However, the models did not converge properly, and parameter estimates suggested that the random slopes did not explain meaningful variability. To ensure model stability and interpretability, the model was specified, including only random intercepts for the individuals. To test the mediation effect of the change in resilience on the relationship between initiation of an EMI component and the change in well-being, the function “mediate” from the R package *mediation* (version 4.5.1) [51] was used. The mediation effect was evaluated using 1000 bootstrap simulations to estimate indirect, direct, and total effects, along with their statistical significance.

Two sensitivity analyses will be conducted for all analyses: first, we will investigate whether the results remain the same when considering the specific type of EMI component (vs not initiating any EMI component). The type of EMI component (including no EMI component) will be operationalized as a categorical predictor in the models. Second, the reason why an EMI component was not initiated (ie, because none was suggested by the app vs because the participant did not initiate the suggested component) will be considered. The reason for noninitiation will be included as a categorical predictor.

Results

Participant Flow

Of the 408 individuals identified for potential participation, 237 could be screened for eligibility. Of those, 190 individuals met the inclusion criteria and provided written informed consent, but 11 participants discontinued their study participation before starting the AI4U training. The full sample of participants that started the AI4U training, therefore, consisted of 179 participants. Of those, 8 participants did not continue with the training phase after the 10-day introductory phase, and 1 participant completed less than two EMA prompts within any day, and their data thus were not included in the analyses. Therefore, the analytic sample for this analysis comprised 170 individuals (ie, 56 in MRT 1, 56 in MRT 2, and 58 in MRT 3) recruited from the general population (n=156) and educational counseling services (n=14). The participant flow is depicted in Figure 1 according to the CONSORT guidelines [41]. The recruitment and the subsequent data collection took place between July 2022 and June 2024.

XSL•FO
RenderX

Figure 1. Flow diagram of participants according to CONSORT guidelines [41]. The flowchart outlines participant inclusion in (a) the AI4U study and (b) the EMI, after which each participant could repeatedly initiate or not initiate a suggested EMI component after the completion of an EMA prompt. AI4U: artificial intelligence for personalized digital mental health promotion in youth; CONSORT: Consolidated Standards of Reporting Trials; EMA: ecological momentary assessment; EMI: ecological momentary intervention.

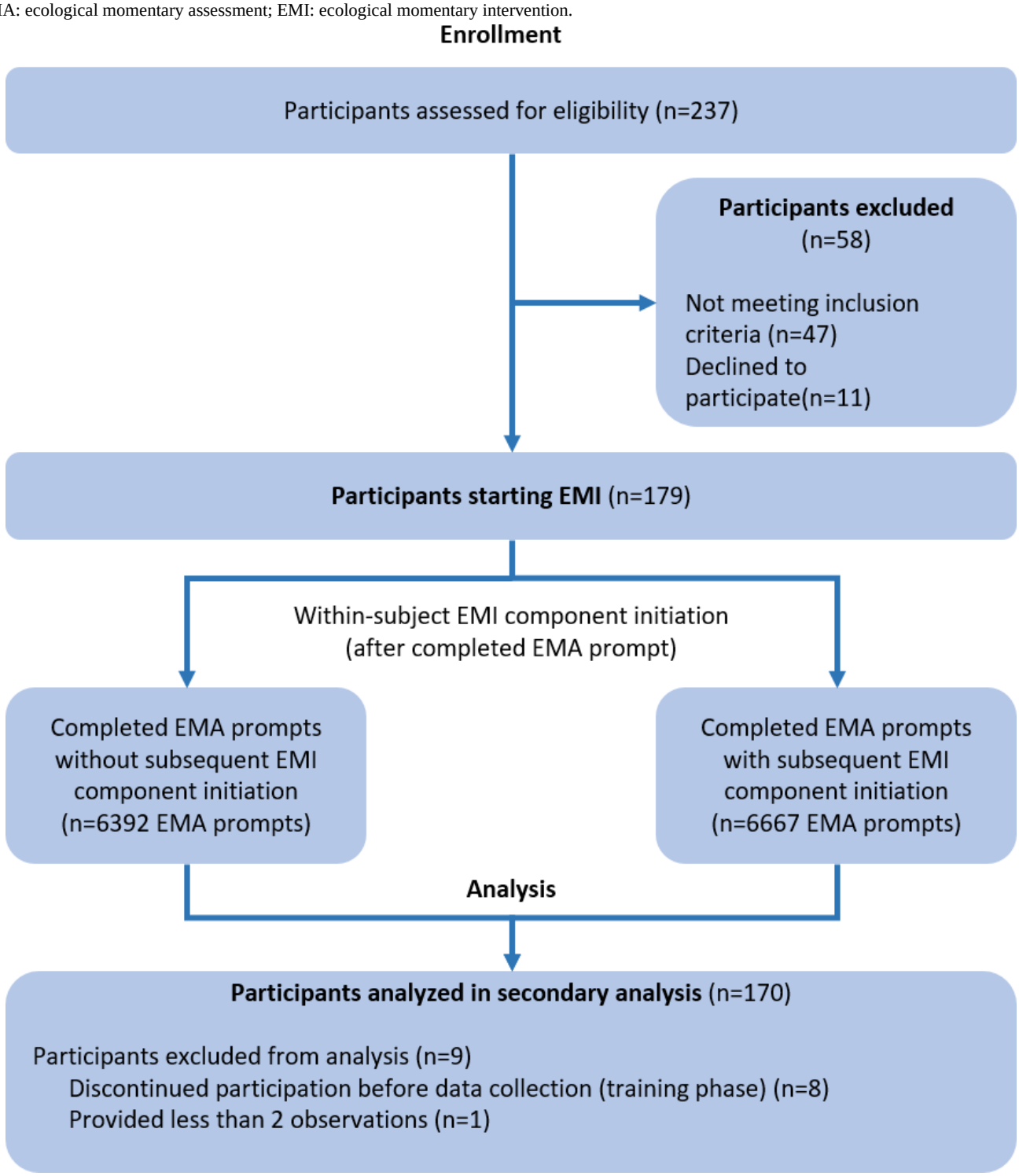

EMI Delivery

During the 30-day training phase of the EMI, both EMA prompts and EMI components were presented to the participants. In total, participants completed 13,059 EMA prompts and missed or did not complete 16,802 EMA prompts, which leads to a completion rate of 43.7% (13,059/29,861). The time of day did not have a significant effect on the completion of an EMA prompt (vs noncompletion); however, the day of training since starting the training phase did have a small effect on the completion of an EMA prompt (Multimedia Appendix 4). Within the same day, the mean time interval between two completed EMA prompts was 178.8 (SD 135.4) minutes, and the mean time interval between an initiated EMI and the next consecutive completed EMA prompt was 167.0 (SD 115.4) minutes. In response to 51.1% (6667/13,059) of the completed EMA prompts, interactive/adaptive EMI components were initiated within 45 minutes after the EMA was completed. The frequency

XSL•FO
RenderX

distribution for initiating EMI components was broadly similar across the eight different EMI components (see Multimedia Appendix 5).

Baseline Data and Sample Characteristics

The analytic sample was mostly female (125/170, 73.5%), with a mean age of 20.41 (SD 3.40) years. Characteristics of the full sample and the analytic sample, as well as aggregated means, skewness, and kurtosis of the proximal outcome variables, can be found in Table 1. The Spearman's rank correlations of the person means of the proximal outcome variables with psychological distress and resilience measured at baseline are reported in Multimedia Appendix 6. No serious adverse events, privacy breaches, or unintended effects were detected during this study.

XSL•FO
RenderX

Table . Sample characteristics (gender, age, migration history, psychological distress, etc) of the full sample and the analytic sample, and descriptive statistics of proximal outcomes assessed by EMA (positive affect, negative affect, stress, and resilience) and time intervals between EMA and EMI of the analytic sample.

Sample characteristics	Full sample (n=179)	Analytic sample (n=170)
Gender, n (%)[a]		
Female	128 (71.5)	125 (73.5)
Male	44 (24.6)	39 (22.9)
Diverse or not specified	3 (1.7)	3 (1.8)
Age (years), mean (SD)	20.36 (3.39)	20.41 (3.40)
Migration history, n (%)[a]		
No migration history	83 (46.4)	79 (46.5)
Foreign born	10 (5.6)	9 (5.3)
First-generation migrant	48 (26.8)	46 (27.1)
Second-generation migrant	25 (14)	24 (14.1)
Psychological distress at baseline, n (%)[a,b]		
None	50 (27.9)	49 (28.8)
Mild	60 (33.5)	58 (34.1)
Moderate	37 (20.7)	34 (20)
Severe	25 (14)	23 (13.5)
Descriptive statistics of proximal outcomes assessed by EMA[c,d]		
Positive affect		
Mean (SD)	—[e]	4.65 (1.17)
Skewness	—	−0.47
Kurtosis	—	0.02
Between-person reliability[f]	—	0.96
Within-person reliability[f]	—	0.78
Negative affect		
Mean (SD)	—	2.15 (1.13)
Skewness	—	1.38
Kurtosis	—	1.88
Between-person reliability[f]	—	0.93
Within-person reliability[f]	—	0.70
Stress		
Mean (SD)	—	3.03 (1.66)
Skewness	—	0.82
Kurtosis	—	−0.13
Resilience		
Mean (SD)	—	4.73 (1.35)
Skewness	—	–0.52
Kurtosis	—	0.03
Descriptive statistics on the time interval between completed EMA and initiated EMI		
Time in minutes between two consecutive completed EMAs within the same day		

XSL•FO
RenderX

Sample characteristics	Full sample (n=179)	Analytic sample (n=170)
Mean (SD)	—	178.8 (135.4)
Median (IQR)	—	140.1 (91.7-218.7)
Time in minutes between an initiated EMI and the consecutive completed EMA within the same day		
Mean (SD)	—	167.0 (115.4)
Median (IQR)	—	134.6 (91.9-202.9)

[a]Number of participants with missing data, n (%), full sample/analytic sample: gender 4 (4/179, 2.2%)/3 (3/170, 1.8%), age 4 (4/179, 2.2%)/3 (3/170, 1.8%), migration history 13 (13/179, 7.3%)/12 (12/170, 7.1%), and psychological distress 7 (7/179, 3.9%)/6 (6/170, 3.5%).

[b]Cutoffs to categorize psychological distress measured with K10: ”none” (range score: 10 to 19); “mild” (range score: 20-24); “moderate” (range score: 25-29); “severe” (range score: 30-50).

[c]EMA: ecological momentary assessment.

[d]All EMA items have been completed on a 7-point scale (1=not at all, 7=very much). All values (except reliability) represent aggregated values over the 30-day training phase.

[e]Not available.

[f]Reliabilty was calculated by applying the function “compRelSem” of the R package *semTools* (version 0.5-7) [52] after estimating a confirmatory factor analysis of the 3-item scales with participants as clusters using *lavaan* (version 0.6.20) [53,54].

Initiation of an EMI Component at t_{n-1} and Well-Being at t_n

To examine whether well-being differed at t_n depending on whether or not an EMI component was initiated after the previous EMA prompt, linear mixed-effects models were specified for each of the three proximal outcome variables (ie, positive affect, negative affect, and stress). All outcomes at t_{n-1} were associated with the respective outcome at t_n (Table 2). Findings did not signal a difference in positive affect at t_n (b=0.00, 95% CI −0.04 to 0.04, d=0.00), negative affect at t_n (b=−0.03, 95% CI −0.06 to 0.01, d=−0.03), or stress at t_n (b=0.01, 95% CI −0.03 to 0.05, d=0.01) for initiating an EMI component at t_{n-1} vs not initiating an EMI component at t_{n-1}, as zero was included in the 95% CI. None of the potential confounding variables was associated to a relevant extent with the proximal outcomes. Sensitivity analyses showed that the results remained largely unchanged when analyzing the effect of initiating a specific type of EMI component vs not initiating any EMI component (Multimedia Appendix 7). The reason why an EMI component was not initiated also did not alter the results (Multimedia Appendix 8).

Table . Regression coefficients, 95% CIs, *P* values, and effect sizes of the linear mixed model for the effect of initiating an EMI[a] component at the previous time point, t_{n-1}, on well-being at the later time point t_n[b].

	b value	95% CI	*P* value	Cohen *d*
Outcome: positive affect at t_n				
Initiation of EMI component at t_{n-1}	0.00	−0.04 to 0.04	.96	0.00
Positive affect at t_{n-1}	0.44	0.42 to 0.46	<.001	0.50
Outcome: negative affect at t_n				
Initiation of EMI component at t_{n-1}	−0.03	−0.06 to 0.01	.17	−0.03
Negative affect at t_{n-1}	0.41	0.39 to 0.43	<.001	0.46
Outcome: stress at t_n				
Initiation of EMI component t_{n-1}	0.01	−0.03 to 0.05	.60	0.01
Stress at t_{n-1}	0.39	0.37 to 0.41	<.001	0.42

[a]EMI: ecological momentary intervention.

[b]Adjusted for potential confounding by age, gender, assignment of ecological momentary intervention component, microrandomized trial, and psychological distress at baseline.

XSL•FO
RenderX

Effects of Initiating an EMI Component at t_{n-1} on Well-Being at t_n by Well-Being at t_{n-1}

Linear mixed-effects models were used to examine whether the magnitude of the effect of initiating an EMI component at t_{n-1} vs not initiating an EMI component at t_{n-1} on well-being at t_n differed depending on the level of well-being at t_{n-1} (Table 3). When taking multiple testing into account, there was no strong difference in the magnitude of the effect of initiating an EMI component vs not initiating an EMI component at t_{n-1} on (1) positive affect at t_n when positive affect at t_{n-1} was high (mean +1 SD) vs low (mean −1 SD, *b*=0.08, 95% CI 0.01 to 0.16, *d*=0.09), (2) negative affect at t_n when negative affect at t_{n-1} was high vs low (*b*=−0.09, 95% CI −0.17 to −0.01, *d*=−0.10), (3) stress at t_n when stress at t_{n-1} was high vs low (*b*=0.09, 95% CI 0.01 to 0.17, *d*=0.10 (with the effect sizes being small or falling just short of being small. Also, findings signaled trivial effect sizes for (1) the difference in positive affect at t_n for initiating vs not initiating an EMI component at t_{n-1} when positive affect at t_{n-1} was high (*d*=0.05) and low (*d*=−0.04), (2) the difference in negative affect at t_n for initiating vs not initiating an EMI component at t_{n-1} when negative affect at t_{n-1} was high (*d*=−0.08) and low (*d*=0.02), and (3) the difference in stress at t_n for initiating vs not initiating an EMI component at t_{n-1} when stress at t_{n-1} was high (*d*=0.06) and low (*d*=−0.04). All outcome variables at t_{n-1} were associated with the respective outcome variables at t_n (Table 3). Sensitivity analyses revealed that the results remained broadly similar when considering the type of EMI component (Multimedia Appendix 7) and the reason for noninitiation of an EMI component into account (Multimedia Appendix 8).

XSL•FO
RenderX

Table . Regression coefficients, 95% CIs, *P* values, and effect sizes of the linear mixed models for the effect of initiating an EMI[a] component at the previous time point t_{n-1} on well-being at the later time point t_n by well-being at t_{n-1}[b], with linear comparisons of coefficients for high (mean +1 SD) vs low (mean −1 SD) levels of well-being at t_{n-1}[b].

	b value	95% CI	*P* value	Cohen *d*
Outcome: positive affect at t_n				
Initiation of EMI component at t_{n-1}	0.00	−0.04 to 0.04	.92	0.00
Positive affect at t_{n-1}	0.42	0.40 to 0.45	<.001	0.48
Initiation of EMI component at t_{n-1} × positive affect at t_{n-1}	0.04	−0.00 to 0.08	.04	0.05
High positive affect	0.04	−0.02 to 0.10	.26	0.05
Low positive affect	−0.04	−0.10 to 0.02	.30	−0.04
High vs low positive affect	0.08	0.01 to 0.16	.04	0.09
Outcome: negative affect at t_n				
Initiation of EMI component	−0.03	−0.06 to 0.01	.17	−0.03
Negative affect at t_{n-1}	0.43	0.41 to 0.46	<.001	0.49
Initiation of EMI component at t_{n-1} × negative affect at t_{n-1}	−0.05	−0.08 to -0.01	.02	−0.05
High negative affect	−0.70	−0.14 to -0.01	.02	−0.08
Low negative affect	0.02	−0.04 to 0.08	.75	0.02
High vs low negative affect	−0.09	−0.17 to -0.01	.02	−0.10
Outcome: stress at t_n				
Initiation of EMI component t_{n-1}	0.01	−0.03 to 0.05	.63	0.01
Stress at t_{n-1}	0.36	0.34 to 0.40	<.001	0.40
Initiation of EMI component at t_{n-1} × stress at t_{n-1}	0.05	−0.01 to 0.09	.02	0.05
High stress	0.06	−0.01 to 0.12	.11	0.06
Low stress	−0.04	−0.10 to 0.03	.39	−0.04
High vs low stress	0.09	0.01 to 0.17	.02	0.10

[a]EMI: ecological momentary intervention.

[b]Adjusted for potential confounding by age, gender, assignment of ecological momentary intervention component, microrandomized trial, and psychological distress at baseline.

Effects of Initiating an EMI Component at t_{n-1} on Changes in Well-Being From t_{n-1} to t_n via Changes in Resilience From t_{n-1} to t_n

Three mediation analyses were conducted to examine the effects of initiation of an EMI component on changes in proximal well-being outcomes from t_{n-1} to t_n via changes in momentary resilience from t_{n-1} to t_n as a potential mediator while controlling for well-being at t_{n-1}. The results of the linear mixed models for the mediator indicated that there was no difference in change in resilience from t_{n-1} to t_n for initiating an EMI component at t_{n-1} vs not initiating an EMI component at t_{n-1} (Multimedia Appendix 9). Change in resilience was associated with changes in well-being in all three models (Multimedia Appendix 9): change in resilience predicted change in positive affect (*b*=0.22, 95% CI 0.21 to 0.24, *d*=0.27), indicating that a greater increase in resilience was associated with a greater increase in positive affect. Similarly, change in resilience predicted change in negative affect (*b*=−0.16, 95% CI −0.18 to −0.15, *d*=−0.19), suggesting that a greater increase in resilience was associated

XSL•FO
RenderX

with a greater decrease in negative affect. Change in resilience was also a predictor of change in stress (b=−0.15, 95% CI −0.17 to −0.13, d=−0.14), indicating that an increase in resilience was associated with a decrease in stress. In all three models, the effect of initiating an EMI component on changes in proximal well-being outcomes was not mediated via changes in resilience (Table 4). Findings were broadly similar in sensitivity analyses taking into account the specific type of EMI component (Multimedia Appendix 7) and the reason for not initiating an EMI component (Multimedia Appendix 8).

Table . Regression coefficients, 95% CIs, and P values of the mediation model for the effect of initiating an EMI[a] component at t_{n-1} on change in well-being from t_{n-1} to t_n via change in resilience from t_{n-1} to t_n[b].

	b value	95% CI	P value
Outcome: change in positive affect			
Total effect	0.00	−0.03 to 0.04	.91
Direct effect	−0.00	−0.04 to 0.03	.86
Indirect effect	0.00	−0.01 to 0.02	.36
Proportion mediated	0.10	−4.69 to 4.35	.84
Outcome: change in negative affect			
Total effect	−0.03	−0.06 to 0.01	.16
Direct effect	−0.02	−0.06 to 0.01	.21
Indirect effect	−0.00	−0.01 to 0.01	.51
Proportion mediated	0.09	−0.80 to 1.17	.55
Outcome: change in stress			
Total effect	0.01	−0.02 to 0.05	.50
Direct effect	0.01	−0.02 to 0.05	.50
Indirect effect	0.00	0.00 to 0.00	>.99
Proportion mediated	0.00	0.00 to 0.00	>.99

[a]EMI: ecological momentary intervention.

[b]Adjusted for potential confounding by age, gender, assignment of ecological momentary intervention component, microrandomized trial, and psychological distress at baseline.

Discussion

Principal Findings

This secondary analysis of data collected in the AI4U project sought to explore proximal effects of initiating EMI components of a compassion-focused smartphone-based digital training for youth on momentary outcomes of well-being (ie, positive affect, negative affect, and stress) assessed at the next time point using EMA. We did not find convincing evidence that well-being at a time point improves compared to the previous time point, depending on whether an EMI component was initiated at the previous time point vs when no EMI component was initiated. There was also no strong indication that this potential effect may be influenced by the level of well-being at the time of EMI initiation. While changes in resilience predicted changes in well-being, the results did not suggest that the effect of initiating an EMI component on mental well-being may be partially mediated by an increase in momentary resilience.

Interpretation

There is growing evidence on the effectiveness of EMI for various outcomes related to mental health [28-30]. However, many previous studies [20,21,28-30] and also the primary analysis of the current study [27] focused on distal outcomes in promoting mental health and well-being. While evidence on these outcomes is important, they are more limited in how exactly specific EMI components exert their effects over time [31,32]. Our study extended previous research by investigating proximal effects of EMI and thereby addressing the theoretical assumption [19], building on prior evidence [20,21,55], that EMIs have an effect at the momentary level. However, so far, there is limited evidence on proximal effects of EMIs on mental health–related outcomes: in one study, by Bernstein et al [56], an EMI was used to promote CBT skills practice in patients admitted for suicidal ideation or behavior during and after their inpatient stay. It showed that when an EMA measuring negative affect was followed by an EMI to engage in CBT skills practice, negative affect measured directly after the skills practice improved compared to when no EMI was prompted [56]. While the sample was small (n=25) and drawn from a specific clinical population, the results highlight that the immediate effects of an EMI can be detected using EMA [56]. Nevertheless, the authors emphasized that as the current study lacked a control group, it may be possible that the observed effects were not exclusively attributable to the EMI, but rather influenced by the multiple sessions conducted during the inpatient stay, in which participants learned CBT skills [56]. In contrast to this study, in our study, the subsequent EMA measuring positive and negative affect and stress did not differ depending on whether

XSL•FO
RenderX

a single EMI component was initiated vs not initiated at the previous time point, and the next EMA did not immediately follow the EMI component. In addition, our sample consisted of young individuals who were not necessarily in contact with mental health services and without a current diagnosis of a mental health condition. The EMI implemented in this study did, accordingly, not serve as a therapeutic intervention but an intervention for mental health promotion and prevention. Hence, our sample is also likely to have experienced a lower intensity or fewer fluctuations of negative affective states [57,58]. This assumption is supported by the distribution of ratings of EMA items (Table 1), which, for all proximal outcome variables, is skewed toward ratings indicating high well-being (ie, low stress, low negative affect, and high positive affect), indicating potential floor and ceiling effects.

In a different study by Vinci et al [32], an EMI for smoking cessation was found to have proximal effects on negative affect at the next timepoint as well as after completion of one week of the training. However, the effect of the EMI could not be observed on other outcome variables (eg, positive affect) [32]. Therefore, while the study used a different intervention and focused on a different population [32], the results align, to some extent, with the results of the current study. In the primary analysis of this study, using data from the first of three MRTs, the authors found that when EMI components were allocated ML-based, levels of momentary resilience at the next time point were, on average, higher compared to when EMI components were allocated at random [27]. In addition, they found evidence for uncontrolled effects of the EMI on distal outcomes, that is, an increase in resilience and the use of adaptive emotion regulation strategies and a decrease in psychological distress [27]. Furthermore, qualitative analyses of the AI4U training indicate that the participants' subjective mental well-being did improve during the AI4U training, as several participants reported improvements in their emotion regulation and self-reflection on emotions [35].

Echoing previous research [6,7], our results showed that resilience was associated with indicators of well-being. This highlights the role resilience may play for mental well-being [6,7]. Previous research on the relationship between resilience and well-being has primarily relied on cross-sectional assessments [7]. With our study, we expand this evidence by investigating the association between changes in momentary resilience, on the one hand, and changes in momentary affect and stress, on the other. Our results showed that an increase in resilience from one EMA to the next was associated with an increase in positive affect and a decrease in negative affect and stress. This finding does not allow us to draw conclusions on a causal relationship between changes in resilience and changes in well-being: an increase in resilience may lead to an increase in well-being, but also increased well-being may lead to more positive ratings of resilience [59]—or changes in both may be due to another, unobserved variable. The findings do, however, indicate that the widely reported association of resilience and well-being in youths and adolescents [6,7] can also be found at the momentary level. Momentary resilience was, however, not affected by the initiation of an EMI component at the previous time point. One possible explanation for this may be that, for digital interventions to be effective, it is necessary that participants engage with them [60,61].

The high variability in and mean number of days of training in this sample (Table 1) may indicate that, while many participants strongly engaged with the training, some engaged only to a limited extent. The duration of this study allowed the overall number of completed EMA prompts representing the proximal outcome of this study to be high (n=13,059). However, 56.3% (16,802/29,861) of EMA prompts were not completed or missed. While we assume that noncompletion of EMA prompts occurred at random and have verified that it is not associated with the time of day (see Multimedia Appendix 4), it is still possible that the available data may be biased due to the participants only responding to the EMA prompts in certain moments. We observed that the likelihood of not completing an EMA prompt increased in the course of the training (Multimedia Appendix 4), indicating that overall, participants tended to engage less with the EMI as the training progressed. According to qualitative feedback of participants, possible reasons for low engagement could be feeling uncomfortable using the AI4U training in public or experiencing the repeated prompts as distracting [35]. As low compliance may reduce exposure to EMI components over time [60,61], it may attenuate its potential impact: Without sufficient interaction with the EMI, the mechanisms leading to an increase in resilience and well-being may not be active to a detectable degree. Thus, the lack of proximal effects of EMI initiation on the next EMA may not necessarily indicate that there is no proximal effect of the EMI, but may rather be due to low adherence to the EMI. Thus, identifying factors associated with low adherence should be the focus of further research [32].

Another explanation for our findings may be that the completion of EMA prompts may have, per se, an effect on well-being and, hence, reflected an active "within-subject control condition" noninferior to proximal effects of EMI components. Prior research has shown that smartphone-based self-monitoring of emotions can have a positive effect on mental health even without the implementation of an EMI [62,63]. The potential mechanism behind this effect may be that emotional self-awareness may increase as a result of repeated self-monitoring [62], which, in turn, may positively impact mental health [62]. It is tempting to speculate whether the EMA delivered as part of the AI4U training may have had an effect on well-being that obscured the effect of the EMI. However, the design of this study does not allow for disentangling this potential masking effect, which highlights the importance of future methodological work on how EMI operates and exerts their beneficial effects.

Limitations

When interpreting the reported results, it is important to consider that all analyses were performed as secondary analyses and that the original study [27] was neither designed nor powered to test the hypotheses of this paper. Sample size considerations for these secondary analyses were based on a previous simulation study [46], which indicates that the available sample size may be sufficient for these secondary analyses. However, it is possible that the sample size was still not large enough to detect effects of EMI component initiation, taking into account that

XSL•FO
RenderX

each of the three MRTs was primarily powered to detect one proximal effect of ML-based EMI delivery and not for additional tests performed as part of current exploratory analyses [27]. In addition, the following five methodological aspects need to be considered.

First, to explore proximal effects of EMI components in the current analysis, completed EMA prompts before (t_{n-1}) and after (t_n) EMI components that were initiated were considered. The time interval between an EMI and the next completed EMA prompt was not fixed and depended on the compliance of participants. Thus, it may be the case that participants missed or ignored EMA prompts after initiation of the EMI. The subsequent EMA prompt was therefore not necessarily in a close temporal relationship with the preceding EMI component. In the time interval between an EMI component and the following EMA, proximal effects may have attenuated and/or external influences may have occurred with a potentially larger effect on affect and stress than the EMI component. While the current approach allowed for some approximation to proximal effects of EMI components, a change in affect or stress from t_{n-1} to t_n does not depict the actual proximal effect of EMI components, as in some cases, several hours could have passed between initiating the EMI component and completing the next EMA prompt. Thus, future studies on proximal effects of EMI components should implement an EMA prompt directly after each EMI component (ie, within minutes).

Second, due to the nature of EMA and EMI, participants completed both the assessments and the interventions in their daily life, that is, without supervision [15,18]. Therefore, it remains unknown whether the EMI component that was initiated was indeed completed in accordance with the instructions or participants were distracted by, or occupied with other tasks [64]. Qualitative data shows that several participants found it challenging to complete EMA and EMI when not at home, but, for example, in public, and that participants were sometimes not conscientious [35]. Likewise, participants were provided with the option to perform EMI exercises that they found helpful on their own (as part of consolidating delivery of EMI components), that is, when no EMI component was suggested by the app. It is also possible that participants performed exercises they learned during the AI4U training independently, as several participants reported that they integrated the breathing intervention into their everyday lives even beyond the duration of the study [35]. This may have masked potential effects of the EMI component and thus biased our findings. Asking participants, for example, as part of the EMA for collecting proximal outcome data or at the end of the EMI training phase, to which degree they complied with the training instructions or implemented EMI exercises into their everyday lives independent of being prompted, may provide more insight into EMI adherence, engagement, and fidelity [65].

Third, this analysis focused on the effects of EMI components that were initiated after EMA completion upon suggestion (ie, interactive/adaptive EMI), which allowed for approximating a quasi-experimental "within-subject control condition" [26] (ie, occasions where an EMI component was suggested vs occasions when no EMI component was suggested). However, the sampling scheme was more complex, including consolidating delivery of EMI components on demand and at prescheduled times based on self-set reminders [27]. Especially, self-initiated EMI components may be very likely to be completed according to instructions, with high fidelity, when deemed helpful by participants. Therefore, especially self-initiated EMI components may have a more marked effect on mental well-being. Future investigations should therefore consider all initiated EMI components in investigating proximal EMI effects, including self-initiated EMI components.

Fourth, the EMI components differed in their content, and not all EMI components were designed and expected to generate immediate effects. For example, the "compass of emotions" component served predominantly psychoeducational purposes that may be unlikely to have a proximal effect. When looking at each of the types of EMI components separately, sensitivity analysis did not detect consistent effects of the type of EMI that was initiated. This contrasts with the findings of other authors [34], who found initial signals that the effect of completing a compassion-focused EMI component (vs not completing an EMI component) was more pronounced when the type of EMI component was either a "breathing exercise" or "soothing imagery" (comparable to "positive imagery" in this study). It may therefore be valuable to further investigate the effects of different EMI types.

Fifth, to keep the burden as low as possible for participants [64], while still ensuring that all relevant aspects were assessed, proximal outcomes were assessed with only one to three items. However, this may result in only a limited aspect of the broader constructs of affect, stress, and resilience being captured. Especially resilience, which is assessed in this study using only a single item, is thought to be actually a very complex and dynamic process comprising various resilience factors [66,67]. The item used in this study, "I can handle all the difficulties that I may encounter," was intended to capture the capacity to withstand or recover quickly from difficulties, as a core aspect of resilience [66], but it neglects other important aspects such as perceived social support and sense of coherence [68,69]. A more comprehensive EMA measure of resilience may facilitate the detection of changes in these specific aspects, but it needs to be carefully considered in the light of the additional burden on the participants.

Conclusions

This secondary analysis did not support the hypothesis that momentary mental well-being (ie, positive affect, negative affect, or stress) at a time point differed depending on whether or not an EMI component was initiated at the previous time point during a 30-day EMI training for mental health promotion in youths. To our knowledge, this is the first study to investigate proximal effects of components of a compassion-focused EMI for mental health promotion on momentary positive affect, negative affect, and stress in a nonclinical sample of youths. This study thereby expands existing research on the effects of EMI by investigating proximal effects of EMI components on momentary outcomes of mental well-being using EMA and by presenting the innovative approach to investigate potential mechanisms of how EMI leads to long-term improvements of

XSL•FO
RenderX

distal outcomes. This can serve as a basis for future research on momentary effects of EMI components, which should consider several aspects relevant for disentangling beneficial proximal and distal effects of EMIs, as reported by prior research [27], such as the implementation of more comprehensive instruments to assess proximal outcomes in closer temporal proximity to the previous EMI component and the importance of compliance with the EMI. Building on and expanding the approach and results presented in this secondary analysis can allow us to further extend our knowledge on proximal effects of EMI and to identify relevant factors for EMI to be effective in supporting mental well-being. Understanding more fully the proximal effects of EMI and how they unfold to translate into previously reported beneficial effects on distal outcomes has the potential to support the development of scalable digital interventions and thereby improve the public's mental health and well-being.

Acknowledgments

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT (Generative Artificial Intelligence Delegation Taxonomy; 2025), the following tasks were delegated to GenAI tools under full human supervision: translation. The GenAI tool used was DeepL. Responsibility for this final paper lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the outcomes. Declaration submitted by EW. Additional note: DeepL was used only for the translation of individual words and phrases from German to English.

Funding

This study was supported by the Ministry of Science, Research and Arts Baden-Württemberg (MWK) from January 2021 to June 2024. We gratefully acknowledge the financial support provided, which made this research possible. The funder had no involvement in this study's design, data collection, analysis, interpretation, or the writing of this paper.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to reasons of data protection, as data was not collected anonymously. Pseudonymized datasets are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: UR, SK, DD, GK, CR, JF, SH, CG, AS, EW
Data curation: JF, GK, CR, EW
Formal analysis: EW (lead), AS (supporting)
Funding acquisition: UR, SK, DD, GK
Investigation: CR, EW
Supervision: UR, AS
Writing – original draft: EW
Writing – review & editing: UR, SK, DD, GK, CR, JF, SH, CG, AS, EW

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of the ecological momentary intervention components.
[DOCX File, 16 KB - jmir_v28i1e85552_app1.docx]

Multimedia Appendix 2
Deviations from the preregistration.
[DOCX File, 24 KB - jmir_v28i1e85552_app2.docx]

Multimedia Appendix 3
Informed consent form for adult participants.
[DOCX File, 5437 KB - jmir_v28i1e85552_app3.docx]

Multimedia Appendix 4
Effect of time of day and day of training on noncompletion vs completion of an ecological momentary assessment prompt.
[DOCX File, 14 KB - jmir_v28i1e85552_app4.docx]

XSL•FO
RenderX

Multimedia Appendix 5
Characteristics of the 30-day training phase.
[DOCX File, 16 KB - jmir_v28i1e85552_app5.docx]

Multimedia Appendix 6
Spearman rank correlations of variables measured using ecological momentary assessment with psychological distress and resilience at baseline.
[DOCX File, 16 KB - jmir_v28i1e85552_app6.docx]

Multimedia Appendix 7
Sensitivity analyses for type of ecological momentary intervention component.
[DOCX File, 28 KB - jmir_v28i1e85552_app7.docx]

Multimedia Appendix 8
Sensitivity analysis for reason of ecological momentary intervention component noninitiation.
[DOCX File, 31 KB - jmir_v28i1e85552_app8.docx]

Multimedia Appendix 9
Mediator and outcome model for hypothesis 3.
[DOCX File, 17 KB - jmir_v28i1e85552_app9.docx]

Checklist 1
CONSORT-EHEALTH checklist.
[PDF File, 1052 KB - jmir_v28i1e85552_app10.pdf]

References

1. Solmi M, Radua J, Olivola M, et al. Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. Mol Psychiatry 2022 Jan;27(1):281-295. [doi: 10.1038/s41380-021-01161-7] [Medline: 34079068]
2. Kessler RC, Angermeyer M, Anthony JC, et al. Lifetime prevalence and age-of-onset distributions of mental disorders in the World Health Organization's World Mental Health Survey Initiative. World Psychiatry 2007 Oct;6(3):168-176. [Medline: 18188442]
3. Erskine HE, Moffitt TE, Copeland WE, et al. A heavy burden on young minds: the global burden of mental and substance use disorders in children and youth. Psychol Med 2015 May;45(7):1551-1563. [doi: 10.1017/S0033291714002888] [Medline: 25534496]
4. Lass-Hennemann J, Sopp MR, Ruf N, et al. Generation climate crisis, COVID-19, and Russia-Ukraine-War: global crises and mental health in adolescents. Eur Child Adolesc Psychiatry 2024 Jul;33(7):2203-2216. [doi: 10.1007/s00787-023-02300-x] [Medline: 37814081]
5. Samji H, Wu J, Ladak A, et al. Review: mental health impacts of the COVID-19 pandemic on children and youth - a systematic review. Child Adolesc Ment Health 2022 May;27(2):173-189. [doi: 10.1111/camh.12501] [Medline: 34455683]
6. Konaszewski K, Niesiobędzka M, Surzykiewicz J. Resilience and mental health among juveniles: role of strategies for coping with stress. Health Qual Life Outcomes 2021 Dec;19(1):33602278. [doi: 10.1186/s12955-021-01701-3]
7. Mesman E, Vreeker A, Hillegers M. Resilience and mental health in children and adolescents: an update of the recent literature and future directions. Curr Opin Psychiatry 2021 Nov 1;34(6):586-592. [doi: 10.1097/YCO.0000000000000741] [Medline: 34433193]
8. Connor KM, Davidson JRT. Development of a new resilience scale: the Connor-Davidson Resilience Scale (CD-RISC). Depress Anxiety 2003;18(2):76-82. [doi: 10.1002/da.10113] [Medline: 12964174]
9. Sisto A, Vicinanza F, Campanozzi LL, Ricci G, Tartaglini D, Tambone V. Towards a transversal definition of psychological resilience: a literature review. Medicina (Kaunas) 2019 Nov 16;55(11):745. [doi: 10.3390/medicina55110745] [Medline: 31744109]
10. Joyce S, Shand F, Tighe J, Laurent SJ, Bryant RA, Harvey SB. Road to resilience: a systematic review and meta-analysis of resilience training programmes and interventions. BMJ Open 2018 Jun 14;8(6):e017858. [doi: 10.1136/bmjopen-2017-017858] [Medline: 29903782]
11. Gulliver A, Griffiths KM, Christensen H. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. BMC Psychiatry 2010 Dec 30;10(113):113. [doi: 10.1186/1471-244X-10-113] [Medline: 21192795]
12. Piers R, Williams JM, Sharpe H. Review: can digital mental health interventions bridge the "digital divide" for socioeconomically and digitally marginalised youth? A systematic review. Child Adolesc Ment Health 2023 Feb;28(1):90-104. [doi: 10.1111/camh.12620] [Medline: 36510375]

XSL•FO
RenderX

13. Díaz-García A, Franke M, Herrero R, Ebert DD, Botella C. Theoretical adequacy, methodological quality and efficacy of online interventions targeting resilience: a systematic review and meta-analysis. Eur J Public Health 2021 Jul 7;31(31 Suppl 1):i11-i18. [doi: 10.1093/eurpub/ckaa255] [Medline: 34240159]
14. Schäfer SK, von Boros L, Schaubruch LM, et al. Digital interventions to promote psychological resilience: a systematic review and meta-analysis. NPJ Digit Med 2024 Feb 8;7(1):30. [doi: 10.1038/s41746-024-01017-8] [Medline: 38332030]
15. Schulte-Strathaus JCC, Rauschenberg C, Baumeister H, Reininghaus U. Ecological momentary interventions in public mental health provision. In: Montag C, Baumeister H, editors. Digital Phenotyping and Mobile Sensing: Springer; 2023:427-439. [doi: 10.1007/978-3-030-98546-2_25]
16. Heron KE, Smyth JM. Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. Br J Health Psychol 2010 Feb;15(Pt 1):1-39. [doi: 10.1348/135910709X466063] [Medline: 19646331]
17. Balaskas A, Schueller SM, Cox AL, Doherty G. Ecological momentary interventions for mental health: a scoping review. PLoS One 2021;16(3):e0248152. [doi: 10.1371/journal.pone.0248152] [Medline: 33705457]
18. Myin-Germeys I, Kasanova Z, Vaessen T, et al. Experience sampling methodology in mental health research: new insights and technical developments. World Psychiatry 2018 Jun;17(2):123-132. [doi: 10.1002/wps.20513] [Medline: 29856567]
19. Myin-Germeys I, Klippel A, Steinhart H, Reininghaus U. Ecological momentary interventions in psychiatry. Curr Opin Psychiatry 2016 Jul;29(4):258-263. [doi: 10.1097/YCO.0000000000000255] [Medline: 27153125]
20. Reininghaus U, Paetzold I, Rauschenberg C, et al. Effects of a novel, transdiagnostic ecological momentary intervention for prevention, and early intervention of severe mental disorder in youth (EMIcompass): findings from an exploratory randomized controlled trial. Schizophr Bull 2023 May 3;49(3):592-604. [doi: 10.1093/schbul/sbac212] [Medline: 36738168]
21. Reininghaus U, Daemen M, Postma MR, et al. Transdiagnostic ecological momentary intervention for improving self-esteem in youth exposed to childhood adversity: the SELFIE randomized clinical trial. JAMA Psychiatry 2024 Mar 1;81(3):227-239. [doi: 10.1001/jamapsychiatry.2023.4590] [Medline: 38019495]
22. Reininghaus U. Ecological momentary interventions in psychiatry: the momentum for change in daily social context. Psychiatr Prax 2018 Mar;45(2):59-61. [doi: 10.1055/s-0044-101986] [Medline: 29495051]
23. Schick A, Rauschenberg C, Ader L, et al. Novel digital methods for gathering intensive time series data in mental health research: scoping review of a rapidly evolving field. Psychol Med 2023 Jan;53(1):55-65. [doi: 10.1017/S0033291722003336] [Medline: 36377538]
24. Daemen M, Postma MR, Lindauer R, et al. Efficacy of a transdiagnostic ecological momentary intervention for improving self-esteem (SELFIE) in youth exposed to childhood adversity: study protocol for a multi-center randomized controlled trial. Trials 2021 Sep 19;22(1):641. [doi: 10.1186/s13063-021-05585-y] [Medline: 34538272]
25. Schick A, Paetzold I, Rauschenberg C, et al. Effects of a novel, transdiagnostic, hybrid ecological momentary intervention for improving resilience in youth (EMIcompass): protocol for an exploratory randomized controlled trial. JMIR Res Protoc 2021 Dec 3;10(12):e27462. [doi: 10.2196/27462] [Medline: 34870613]
26. Ader L, Schick A, Löffler M, et al. Refocusing of attention on positive events using monitoring-based feedback and microinterventions for patients with chronic musculoskeletal pain in the perpain randomized controlled trial: protocol for a microrandomized trial. JMIR Res Protoc 2023 Sep 20;12:e43376. [doi: 10.2196/43376] [Medline: 37728983]
27. Rauschenberg C, Schirmbeck F, Fechtelpeter J, et al. Effects of AI4U training, a machine learning-based, adaptive ecological momentary intervention for personalized mental health promotion in youth: findings from a micro-randomized trial. PsyArXiv. Preprint posted online on Aug 23, 2025. [doi: 10.31234/osf.io/mvyar]
28. Marciniak MA, Shanahan L, Rohde J, et al. Standalone smartphone cognitive behavioral therapy-based ecological momentary interventions to increase mental health: narrative review. JMIR mHealth uHealth 2020 Nov 12;8(11):e19836. [doi: 10.2196/19836] [Medline: 33180027]
29. Versluis A, Verkuil B, Spinhoven P, van der Ploeg MM, Brosschot JF. Changing mental health and positive psychological well-being using ecological momentary interventions: a systematic review and meta-analysis. J Med Internet Res 2016 Jun 27;18(6):e152. [doi: 10.2196/jmir.5642] [Medline: 27349305]
30. Zainal NH, Newman MG. A randomized controlled trial of a 14-day mindfulness ecological momentary intervention (MEMI) for generalized anxiety disorder. Eur Psychiatry 2023 Jan 16;66(1):e12. [doi: 10.1192/j.eurpsy.2023.2] [Medline: 36645098]
31. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462. [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
32. Vinci C, Sutton SK, Yang MJ, Jones SR, Kumar S, Wetter DW. Proximal effects of a Just-in-Time Adaptive Intervention for smoking cessation with wearable sensors: microrandomized trial. JMIR mHealth uHealth 2025 Mar 19;13:e55379. [doi: 10.2196/55379]
33. Paetzold I, Gugel J, Schick A, et al. Exploring the implementation of a novel, transdiagnostic, hybrid ecological momentary intervention for improving resilience in youth (EMIcompass): a process evaluation in the realist framework. Psychol Psychother 2025 Mar;98(1):103-132. [doi: 10.1111/papt.12565] [Medline: 39699715]

XSL•FO
RenderX

34. Gugel J, Paetzold I, Rauschenberg C, et al. Zooming in: proximal effects of a transdiagnostic, ecological momentary intervention for enhancing resilience in help-seeking young people (emicompass). JMIR Preprints. Preprint posted online on May 8, 2025. [doi: 10.2196/preprints.77150]
35. Hiller S, Götzl C, Rauschenberg C, et al. Health-promoting effects and everyday experiences with a mental health app using ecological momentary assessments and AI-based ecological momentary interventions among young people: qualitative interview and focus group study. JMIR mHealth uHealth 2025 Apr 29;13(1):e65106. [doi: 10.2196/65106] [Medline: 40300160]
36. Bidargaddi N, Schrader G, Klasnja P, Licinio J, Murphy S. Designing m-Health interventions for precision mental health support. Transl Psychiatry 2020 Jul 7;10(1):222. [doi: 10.1038/s41398-020-00895-2] [Medline: 32636358]
37. Klasnja P, Hekler EB, Shiffman S, et al. Microrandomized trials: an experimental design for developing just-in-time adaptive interventions. Health Psychol 2015 Dec;34S:1220-1228. [doi: 10.1037/hea0000305] [Medline: 26651463]
38. Götzl C, Hiller S, Rauschenberg C, et al. Artificial intelligence-informed mobile mental health apps for young people: a mixed-methods approach on users' and stakeholders' perspectives. Child Adolesc Psychiatry Ment Health 2022 Nov 17;16(1):86. [doi: 10.1186/s13034-022-00522-6] [Medline: 36397097]
39. Fechtelpeter J, Rauschenberg C, Goetzl C, et al. Computational network models for forecasting and control of mental health trajectories in digital applications. NPJ Digit Med 2025 Dec 30;9(1):70. [doi: 10.1038/s41746-025-02252-3] [Medline: 41461943]
40. Wierzba E, Schick A, Rauschenberg C, Fechtelpeter J, Hiller S, Götzl C, et al. Promoting psychological resilience and well-being in youth with a smartphone-based ecological momentary mhealth intervention. Open Science Framework - Preregistration. 2024 Oct 25. URL: https://osf.io/738u5/?view_only=1745d36e147d497e919d5f9ed8755dc1 [accessed 2026-05-26]
41. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ 2025 Apr 14;389:e081123. [doi: 10.1136/bmj-2024-081123] [Medline: 40228833]
42. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
43. Rauschenberg C, Boecking B, Paetzold I, et al. A compassion-focused ecological momentary intervention for enhancing resilience in help-seeking youth: uncontrolled pilot study. JMIR Ment Health 2021 Aug 5;8(8):e25650. [doi: 10.2196/25650] [Medline: 34383687]
44. Collins LM. Optimization of Behavioral, Biobehavioral, and Biomedical Interventions: The Multiphase Optimization Strategy (MOST): Springer; 2018:297. [doi: 10.1007/978-3-319-72206-1]
45. Kessler RC, Andrews G, Colpe LJ, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002 Aug;32(6):959-976. [doi: 10.1017/s0033291702006074] [Medline: 12214795]
46. McNeish D. Multilevel mediation with small samples: a cautionary note on the multilevel structural equation modeling framework. Struct Equ Modeling 2017 Jul 4;24(4):609-625. [doi: 10.1080/10705511.2017.1280797]
47. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw 2015;67(1):1-48. [doi: 10.18637/jss.v067.i01]
48. Pinheiro J, Bates D, Team RC. R core team. nlme: linear and nonlinear mixed effects models. R package version. 2026 Mar 27. URL: https://cran.r-project.org/web/packages/nlme/index.html [accessed 2026-05-26]
49. Tierney N, Cook D. Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations. J Stat Soft 2023;105(7):1-31. [doi: 10.18637/jss.v105.i07]
50. Hothorn T, Bretz F, Westfall P. Simultaneous inference in general parametric models. Biom J 2008 Jun;50(3):346-363. [doi: 10.1002/bimj.200810425] [Medline: 18481363]
51. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. mediation: R Package for Causal Mediation Analysis. J Stat Softw 2014;59(5):1-38. [doi: 10.18637/jss.v059.i05]
52. Jorgensen TD, Pornprasertmanit S, Schoemann AM, Rosseel Y. SemTools: useful tools for structural equation modeling. 2026 Feb 14. URL: https://CRAN.R-project.org/package=semTools [accessed 2026-05-26]
53. lavaan RY. An R package for structural equation modeling. J Stat Softw 2012;48(2):1-36. [doi: 10.18637/jss.v048.i02]
54. Rosseel Y, Jorgensen TD, lavaan WLD. Latent variable analysis. 2025 Dec 21. URL: https://CRAN.R-project.org/package=lavaan [accessed 2026-05-26]
55. Myin-Germeys I, van Aubel E, Vaessen T, et al. Efficacy of acceptance and commitment therapy in daily life in early psychosis: results from the multi-center INTERACT randomized controlled trial. Psychother Psychosom 2022;91(6):411-423. [doi: 10.1159/000522274] [Medline: 35306502]
56. Bernstein EE, Bentley KH, Nock MK, Stein MB, Beck S, Kleiman EM. An ecological momentary intervention study of emotional responses to smartphone-prompted CBT skills practice and the relationship to clinical outcomes. Behav Ther 2022 Mar;53(2):267-280. [doi: 10.1016/j.beth.2021.09.001] [Medline: 35227403]
57. Garnefski N, Van Den Kommer T, Kraaij V, Teerds J, Legerstee J, Onstein E. The relationship between cognitive emotion regulation strategies and emotional problems: comparison between a clinical and a non‐clinical sample. Eur J Pers 2002 Sep;16(5):403-420. [doi: 10.1002/per.458]

XSL•FO
RenderX

58. Naim R, Shaughnessy S, Smith A, Karalunas SL, Kircanski K, Brotman MA. Real-time assessment of positive and negative affective fluctuations and mood lability in a transdiagnostic sample of youth. Depress Anxiety 2022 Dec;39(12):870-880. [doi: 10.1002/da.23293] [Medline: 36325887]
59. Colombo D, Suso-Ribera C, Fernández-Álvarez J, et al. Affect recall bias: being resilient by distorting reality. Cogn Ther Res 2020 Oct;44(5):906-918. [doi: 10.1007/s10608-020-10122-3]
60. Yardley L, Spring BJ, Riper H, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
61. Milne-Ives M, Homer SR, Andrade J, Meinert E. Potential associations between behavior change techniques and engagement with mobile health apps: a systematic review. Front Psychol 2023;14(1227443):37794916. [doi: 10.3389/fpsyg.2023.1227443]
62. Bakker D, Rickard N. Engagement in mobile phone app for self-monitoring of emotional wellbeing predicts changes in mental health: MoodPrism. J Affect Disord 2018 Feb;227(432-42):432-442. [doi: 10.1016/j.jad.2017.11.016] [Medline: 29154165]
63. Kauer SD, Reid SC, Crooke AHD, et al. Self-monitoring using mobile phones in the early stages of adolescent depression: randomized controlled trial. J Med Internet Res 2012 Jun 25;14(3):e67. [doi: 10.2196/jmir.1858] [Medline: 22732135]
64. Stone AA, Schneider S, Smyth JM. Evaluation of pressing issues in ecological momentary assessment. Annu Rev Clin Psychol 2023 May 9;19(107-31):107-131. [doi: 10.1146/annurev-clinpsy-080921-083128] [Medline: 36475718]
65. Rintala A, Apers S, Eisele G, Verhoeven D. Briefing and debriefing in an experience sampling study. In: Myin-Germeys I, Kuppens P, editors. The Open Handbook of Experience Sampling Methodology: A Step-by-Step Guide to Designing, Conducting and Analysing ESM Studies, 2nd edition: Center for Research on Experience Sampling and Ambulatory Methods Leuven (REAL); 2022:119-134 URL: https://www.scribd.com/document/710205358/The-Open-Handbook-of-Experience-Sampling-Methodology [accessed 2026-06-09]
66. Ader L, Schick A, Vaessen T, et al. The role of childhood trauma in affective stress recovery in early psychosis: an experience sampling study. Schizophr Bull 2024 Jul 27;50(4):891-902. [doi: 10.1093/schbul/sbae004] [Medline: 38366989]
67. Southwick SM, Bonanno GA, Masten AS, Panter-Brick C, Yehuda R. Resilience definitions, theory, and challenges: interdisciplinary perspectives. Eur J Psychotraumatol 2014;5(1):25338. [doi: 10.3402/ejpt.v5.25338] [Medline: 25317257]
68. Schäfer SK, Fritz J, Sopp MR, et al. Interrelations of resilience factors and their incremental impact for mental health: insights from network modeling using a prospective study across seven timepoints. Transl Psychiatry 2023 Oct 23;13(1):328. [doi: 10.1038/s41398-023-02603-2] [Medline: 37872216]
69. Schäfer SK, Kunzler AM, Kalisch R, Tüscher O, Lieb K. Trajectories of resilience and mental distress to global major disruptions. Trends Cogn Sci 2022 Dec;26(12):1171-1189. [doi: 10.1016/j.tics.2022.09.017] [Medline: 36302711]

Abbreviations

AI4U: artificial intelligence for personalized digital mental health promotion in youth
CBT: cognitive behavioral therapy
CONSORT : Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
EMA: ecological momentary assessment
EMI: ecological momentary intervention
K10: Kessler Psychological Distress Scale
mHealth: mobile health
ML: machine learning
MRT: microrandomized trial
RCT: randomized controlled trial

Edited by S Brini; submitted 10.Oct.2025; peer-reviewed by H Li, JM Griffith, P Santangelo; revised version received 07.May.2026; accepted 11.May.2026; published 18.Jun.2026.

Please cite as:
Wierzba E, Schick A, Rauschenberg C, Fechtelpeter J, Hiller S, Götzl C, Durstewitz D, Krumm S, Koppe G, Reininghaus U
Promoting Psychological Resilience and Well-Being in Youth With a Smartphone-Based Ecological Momentary mHealth Intervention: Secondary Analysis of a Microrandomized Trial
J Med Internet Res 2026;28:e85552
URL: https://www.jmir.org/2026/1/e85552
doi:10.2196/85552

XSL•FO
RenderX

© Eva Wierzba, Anita Schick, Christian Rauschenberg, Janik Fechtelpeter, Selina Hiller, Christian Götzl, Daniel Durstewitz, Silvia Krumm, Georgia Koppe, Ulrich Reininghaus. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Emerging Roles of AI in Self-Directed Stress Management: Systematic Review

Mary Kamillah Grace Reyes*, BSocSc; Shauna Sha Min Teo*, BSocSc; Andree Hartanto, PhD

School of Social Sciences, Singapore Management University, Level 4, 10 Canning Rise, Singapore, Singapore

*these authors contributed equally

Corresponding Author:
Mary Kamillah Grace Reyes, BSocSc
School of Social Sciences, Singapore Management University, Level 4, 10 Canning Rise, Singapore, Singapore

Abstract

Background: Stress is widespread and carries substantial mental health, social, and economic burdens. Yet, access to clinician-led stress management remains constrained by service capacity, cost, and stigma. In response, artificial intelligence (AI)–enabled tools have rapidly proliferated as scalable, self-directed options. However, evidence on how these systems support stress management outside formal clinical settings remains fragmented.

Objective: This systematic review aimed to synthesize empirical evidence on how AI-enabled technologies are used for self-directed stress management. We mapped the emerging functions of these tools, the psychological frameworks informing their design, the populations and settings studied, and the outcomes reported.

Methods: We conducted a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)–compliant systematic review of English-language studies published between 2000 and 2025. Six databases were searched (APA PsycINFO, PubMed, MEDLINE, Scopus, Web of Science Core Collection, ProQuest, and Google Scholar).

Results: Of 3008 records identified, 35 studies met the inclusion criteria. The methodological quality of included studies was critically appraised using the Mixed Methods Appraisal Tool (version 2018). Findings illustrated that AI-supported stress management can operate through 5 core functions, including psychological intervention, behavioral support, psychoeducation, companionship, and emotional support, and stress monitoring, detection, and triage. Across the reviewed studies, these functions supported self-directed stress management by helping users identify stress, regulate responses, and engage in coping outside formal clinical care.

Conclusions: AI-enabled systems show preliminary promise for supporting self-directed stress management through multiple user-facing functions grounded in established psychological frameworks.

Trial Registration: PROSPERO CRD420251135780; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251135780

(J Med Internet Res 2026;28:e90709) doi:10.2196/90709

KEYWORDS

artificial intelligence; chatbots; conversational agents; digital mental health; stress management; self-directed stress management; self-guided intervention; psychoeducation; stress monitoring; Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PRISMA; systematic review

Introduction

The Emerging Roles of Artificial Intelligence in Self-Directed Stress Management

Overview

From minor hassles to major crises, everyday life is saturated with stress, commonly understood as a state of worry or mental tension caused by difficult situations [1]. Amidst the COVID-19 pandemic, stress symptomatology was estimated to affect up to 25% of individuals [2,3]. Global estimates indicate that stress prevalence ranges from approximately 13% to nearly 40% across population-level surveys and meta-analytic studies, with approximately 1 in 3 adults reporting daily stress and most countries showing worsening trajectories over the past decade [4-7]. These figures are often magnified in at-risk groups, such as informal caregivers. Meta-analytic evidence indicates that informal caregivers face a substantial mental health burden, with many reported to meet diagnostic criteria for mental health conditions, and nearly half experience clinically significant caregiver burden [8,9]. Therefore, stress can act as a symptom of existing difficulties and as a risk factor that worsens long-term mental health trajectories. However, scalable and flexible support that can accommodate diverse needs remains limited by both system-level and individual-level barriers. For instance, constrained service capacity means that groups such as students experience high stress but low uptake of treatment [10,11].

XSL•FO
RenderX

Meanwhile, financial, logistical, and stigma-related obstacles disproportionately impede access to clinical and counseling care, contributing to poorer mental health outcomes [12-14].

In light of this, self-directed stress management has gained traction as a practical alternative. Such approaches parallel self-help interventions, defined as standardized, evidence-based programs that participants implement independently of professional therapists or guides [10,15]. These self-directed solutions can enhance accessibility while offering greater autonomy and convenience, especially for those facing structural or social barriers to conventional care. Digital mental health interventions targeting youth have shown potential to achieve sustainable improvements in stress and anxiety outcomes by providing personalized, accessible, and culturally responsive resources. However, high dropout rates, limited long-term data, and the inherently low accountability of self-directed digital interventions [16,17] identify a clear need for novel strategies for supporting continued use and intervention efficacy.

AI and Self-Directed Stress Management Interventions

Given these limitations, artificial intelligence (AI) offers capabilities to address gaps in current self-directed stress management programs. Large language models (LLMs), natural language processing (NLP), and related machine learning approaches have become increasingly ubiquitous in many sectors. In this context, a key strength of these intelligent models is their ability to interpret complex user data and translate it into tailored, context-specific responses [18-20]. Leveraging these capabilities could enable more accessible and efficient delivery of stress management solutions that are responsive to individual needs. In broader health care contexts, AI has been proposed and increasingly applied to support personalized monitoring, reduce selected workflow burdens, and assist routine tasks [21-23]. These precedents suggest that AI-enabled models can harness similar efficiencies for independent stress self-management.

At the intervention level, workplace internet-based stress management interventions have demonstrated substantial reductions in perceived stress [24] as well as cost-effective benefits [24,25]. AI-based systems further enhance these interventions by offering greater scalability, consistent availability, and anonymity [26-28], which may lower barriers for individuals who are hesitant to engage in face-to-face stress management programs. Conversational agents exemplify this potential by enabling dialogue that users experience as more therapeutic, personalized, and autonomous [17,29,30]. Some evidence suggests that users may be more willing to disclose sensitive information to nonhuman agents, partly because of reduced perceived judgment [17,31-33]. In addition, intelligent conversational models do not require ongoing clinician involvement, allowing them to scale far more efficiently than human-delivered or clinician-supported digital interventions [34-36]. Collectively, these capabilities indicate that AI-driven self-directed stress management tools can extend beyond static guidance to provide adaptive, ongoing support that responds to technological advances and broader societal needs for accessible, flexible, and scalable approaches to stress management.

Applications of AI for Self-Directed Stress Management

One prominent application of AI in this context is the use of physiological and behavioral signals to monitor stress. Specifically, stress monitoring is based on measuring physiological changes and comparing them with baseline metrics [37-39]. The symptomatology of stress involves consistent and predictable physiological changes, such as heart rate variability; AI may be used to detect and model these physiological signs to infer when an individual is under stress [40]. Machine learning models trained on such health data can classify stress states with promising accuracy [41] and, when embedded in wearable or smartphone-based sensors, could provide continuous, unobtrusive monitoring across daily contexts [38,39]. This addresses a core limitation of traditional self-report measures, which capture only brief, time-bound snapshots of experience [39,42]. Related literature on digital phenotyping shows that passive smartphone-sensor data can be used to infer behavioral and symptom patterns, enabling the matching of individuals to mental health applications that suit specific needs [43,44], offering a route to more personalized and timely self-help support.

Beyond monitoring physiological symptoms, AI can also support behavior change processes that reduce stress in the long term. Intelligent systems are increasingly used to deliver habit tracking tools, adaptive nudges, and tailored coping suggestions, drawing on broader developments in digital behavior change technologies [45-47]. In this context, “user-centered” and “person-based” design approaches have become central to sustaining behavioral patterns in similar interventions [48]. In practice, virtual therapists and chatbots may offer relatively discreet spaces, reducing perceived stigma and the sense of judgment that often deters help-seeking [35,49]. Empathy-oriented conversational agents take this further by analyzing user sentiment and generating emotionally attuned responses [50]. Such intelligent dialogue platforms have shown promise in reducing student stress through supportive, user-centric conversation and increased perceived trust in the conversational agent [50-52].

At the same time, the broader adoption of AI in mental health raises substantial concerns, as generative and predictive models are inherently nondeterministic and may produce unreliable or erroneous outputs. Such risks are compounded when these probabilistic models are trained on noisy, incomplete, or low-fidelity data from consumer-grade devices unable to match the reliability of clinical-grade systems [38,53]. Data privacy concerns are especially salient for multimodal AI models that rely on facial images, voice recordings, or other sensitive user data [54,55]. Under regulations such as the Health Insurance Portability and Accountability Act (HIPAA), patient health information held by covered entities is strongly protected; however, most existing regulatory and data protection frameworks have yet to fully consider how AI developers and vendors collect, use, and share mental health–related data [55-57]. In practice, many of these systems are trained on datasets that are limited in scale and diversity, with systematic underrepresentation of minority groups [58-60]. This pattern of limitations systematically biases model predictions and performance for users in marginalized populations, potentially exacerbating existing mental health care inequalities

XSL•FO
RenderX

[23,35,58-60]. As a result, appraising the value and impact of AI-driven interventions for stress management requires rigorous scrutiny of how these systems are conceptualized, designed, trained, and evaluated for different population needs.

This Research

Despite rapid growth in AI-enabled mental health tools, evidence on AI for self-directed stress management remains fragmented. Existing studies vary widely in the types of AI used, the theoretical frameworks guiding development, and the stress-related functions targeted (eg, monitoring, psychoeducation, and behavior change), and there has been no systematic synthesis focused on autonomous use outside formal clinical settings. This review addressed this gap through a systematic examination of empirical studies published between 2000 and 2025 that used AI-integrated platforms independently by individuals to cope with stress outside formal clinical care. Specifically, we identified the main categories of AI tools evaluated to date and noted the populations and settings they target. Finally, we analyzed the functions by which these tools support self-directed stress coping. In doing so, we aimed to provide an integrated overview of existing approaches and lay a foundation for future work on responsible, user-centered AI in self-directed stress management.

Methods

Transparency and Openness

We conducted this systematic review in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Checklist 1) [61]. The review protocol, including the search strategy and planned synthesis framework, was submitted for registration on PROSPERO (registration ID: CRD420251135780). Furthermore, citation management, record deduplication, and metadata extraction were conducted using Zotero version 7.0.24 (Corporation for Digital Scholarship) [62]. Relevant files, namely screening records and extraction data used in this work, will be made publicly available upon publication.

Search Strategy

A systematic search strategy was jointly developed by the first 2 authors (MKGR and SSMT) and conducted across 6 electronic databases: APA PsycINFO, PubMed or MEDLINE, Scopus, Web of Science Core Collection, ProQuest, and Google Scholar. The long-string series of keywords, (“artificial intelligence” OR “A.I.” OR AI OR GPT OR chatbot OR “large language model” OR LLM*) AND (stress OR “stress management” OR “stress response*” OR (“self-directed” AND stress*) OR “de-stress*” OR “self-care”), was applied to APA PsycINFO, PubMed, Scopus, and Web of Science. All records were extracted from these 4 databases. For supplementary searching in ProQuest and Google Scholar, a shortened query with the same Boolean structure was used, namely (“artificial intelligence” OR AI OR GPT OR LLM OR chatbot) AND (stress OR “stress management” OR “stress response*” OR “de-stress*"). Only the first 100 results were retrieved based on relevance. Records were limited to studies published in English from 2000 to 2025 that involved human participants, including early online publications where available. Exclusion filters were applied to exclude medical, biological, and clinical trial domains unrelated to self-directed stress management. The search and extraction of records were conducted on September 4, 2025.

Screening Criteria and Calibration

This review used a 2-stage screening process consisting of (1) title and abstract screening and (2) full-text screening. Both the first and second authors independently reviewed studies from all 6 sources, with a focus on human subjects who independently use AI or AI-enabled agents for personal stress management. We excluded records that involved AI and stress management in purely clinical settings, AI use in non–stress-related settings, and non–human-related stress (eg, stress on crop rotation). Records were retained according to the following inclusion criteria:

1. The record was written in English.
2. The record was unique and not a duplicate of any prior entry.
3. The record was not retracted or withdrawn at the time of screening.
4. The record explicitly addressed self-directed or personal stress management, rather than clinician-led, medically supervised, or biomarker-based clinical interventions.
5. The record provided sufficient information to obtain full-text access.

A trial set of 100 records was independently screened to calibrate inclusion decisions and ensure consistency. Following this calibration, the remaining records were screened independently. Records with insufficient abstract detail but retrievable full text were advanced, whereas records with no abstract and insufficient retrievable information were excluded. Records meeting inclusion criteria at the title and abstract level proceeded to full-text review. Similarly, at this subsequent full-text screening stage, a trial set of 50 records was undertaken to ensure alignment, after which the same inclusion and exclusion criteria as those used in title and abstract screening were applied, with the addition of the following criteria:

1. The record was peer reviewed and published.
2. The record did not pertain to interventions delivered in conventional clinical mental health settings.
3. The record explicitly involved AI use or AI-assisted technology (eg, chatbots, LLMs, and AI-enabled apps) for stress self-management.
4. The record noted that AI use was “self-directed,” ie, initiated and managed by the user without clinician or therapist oversight.
5. The record presented empirical data (quantitative, qualitative, or mixed methods) and collected data from human participants.

Data Extraction and Synthesis

Data extraction was conducted independently by both the first and second authors, with discrepancies resolved through discussion. As the included records comprised both qualitative and quantitative studies, thematic analysis was used to identify and categorize recurring themes in how AI supports stress management. This approach was appropriate because it enabled

XSL•FO
RenderX

findings from qualitative, quantitative, and mixed method studies to be integrated into a coherent whole, consistent with guidance for mixed methods systematic review [63]. An exploratory coding procedure was applied to identify recurring patterns, emerging categories, and overlapping conceptual discussions within the agreed extraction criteria. A trial set of 20 records was screened to ensure consistency of data for extraction. When consolidating outcome data across heterogeneous study designs, we prioritized textual descriptions of reported effects; where quantitative results were available, these were also mapped onto a common "direction of effect" rubric (improvement, no or minimal change, and worsening), following recommendations for thematic synthesis in systematic reviews of mixed evidence [64,65].

To synthesize data from both qualitative and quantitative records, we followed Thomas and Harden's [66] procedure for thematic synthesis. After confirming the final set of included records, both first and second authors reviewed all records again to develop codes for the text and descriptive themes. We initially analyzed each record against our full-text inclusion criteria, focusing on the forms of AI used to manage stress, the notable frameworks on which AI-enabled agents were trained to provide help, and the significant or nonsignificant results of interventions. The first and second authors then discussed our preliminary findings to produce a list of themes that address the focus of our research: how AI models support stress management. Themes were internally refined to ensure coherence and clear differentiation. Both the first and second authors independently reviewed the final included records and categorized them. In addition, following the data extraction protocols of Aromataris et al [67] for umbrella review, the following were also extracted from each record: (1) author and year of publication, (2) journal or publication, (3) country (of sample), (4) study type, (5) study pool, (6) sample size, (7) mean age (years), (8) age range (years), (9) intervention name and description, (10) intervention framework, and (11) key findings.

Results

Overview of Search and Inclusion Results

The search string yielded 3008 records (Figure 1). The full list of all unique records identified and screened (N=2810), including those advanced to full-text review (n=173) and excluded full-text articles with reasons (n=138), is available in Multimedia Appendix 1. After removing duplicates (198/3008, 6.58%), 2810 (93.42%) unique records remained. This review proceeded in 2 phases. In the first phase, both authors reviewed the titles and abstracts of all 2810 records to determine eligibility for the full-text phase based on the criteria outlined in the "Methods" section. A total of 2637 (93.84%) records were excluded after title and abstract screening, retaining 173 (6.16%) records for full-text screening. Specifically, 27 (1.02%) records lacked an abstract, 354 (13.45%) were clinical studies not explicitly related to stress self-management, 3 (0.11%) were not in English, and 2253 (85.44%) were outside the scope of this review. In the second phase, the 173 included records were reviewed in full by both authors. During this full-text screening phase, 63 (36.42%) records were removed for not being peer reviewed or in preprint, 8 (4.62%) records were studies set in purely clinical settings (ie, set in conventional therapeutic settings), 13 (7.51%) records involved the presence of a clinician or AI use for stress relief that was co-administered with clinical treatment, 40 (23.12%) records had AI use not for stress management, and 5 (2.89%) records did not have independent use of AI tools for stress management. During the data extraction and thematic analysis, a further 9 (5.20%) records were removed due to insufficient data. A total of 35 (20.23%) records were included in the final review. The overall agreement rates for title and abstract screening (2426/2810, 86.30%) and for full-text screening (138/173, 79.77%) were adequate, with data extraction agreement at >70%. Disagreements were resolved at the end of each stage through discussion with both authors.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram outlining the systematic search process. AI: artificial intelligence.

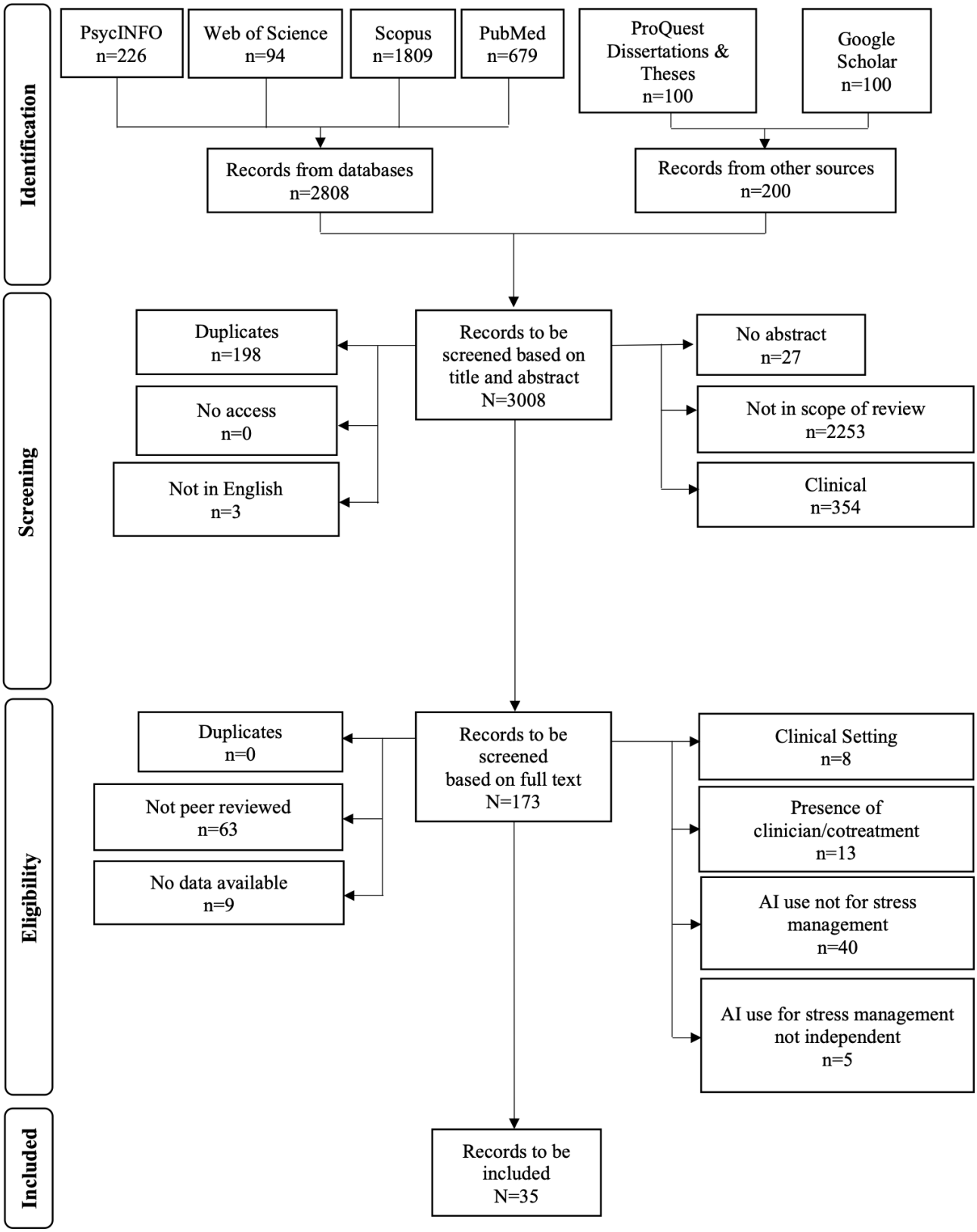

Overview of Included Studies

This review synthesized 35 studies, including 3 (8.57%) qualitative studies, 11 (31.43%) randomized controlled trials, 6 (17.14%) nonrandomized quantitative studies, 4 (11.43%) quantitative descriptive studies, and 11 (31.43%) mixed methods studies. These included studies examined AI-enabled tools for self-directed stress management across chatbots, mobile apps, embodied agents, and other digital platforms. Table 1 summarizes the key characteristics of the included records, including study-level information, intervention frameworks, AI architectures used, and sample characteristics where reported.

XSL•FO
RenderX

Sample characteristics include the study pool or demographic focus of each sample (eg, employed nurses, fathers, and university students), as well as sample size, mean age, and age range. AI architecture classifications for each study (eg, LLM-based, rule-based, and hybrid) are also included in Table 1.

XSL•FO
RenderX

Table . Characteristics of included records.

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Alanazi et al, 2025 [68]	*Scientific Reports*	Saudi Arabia	Mixed methods	Children with Down syndrome and caregivers of children with Down syndrome	N=130 (n=7 and n=123)	—[b]	3 - 18 (children) and ≥18 (caregivers)	Mixed (CBT[c]-, occupational-, and behavior-oriented components)	Several multimodal AI systems, combining computer vision, NLP[d], and machine learning	Moderate-large improvements were observed in children's mobility (d≈0.65), communication (d≈0.72), and daily living or household tasks (d≈0.83); carers reported reduced caregiving stress and greater child independence, engagement, and confidence, although adoption was limited by cost, lack of support, and technical issues, with qualitative reports also indicating subjective stress reduction.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Alanezi, 2024 [69]	*Journal of Multidisciplinary Healthcare*	Saudi Arabia	Qualitative	Mixed adult population	n=24	—	≥18	CBT-based psychoeducation	ChatGPT-3.5 or / LLM[e]-based conversational AI	Participants reported using AI mental health tools primarily for psychoeducation (~60%), emotional support, and feeling heard (~50%), goal setting or motivation, and accessing mental health resources; perceived benefits included CBT-like reframing, coping guidance, and self-monitoring, while concerns centered on limited assessment accuracy, data privacy or ethics, and cultural-linguistic mismatch.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Allan, 2024 [70]	*Behaviour and Information Technology*	New Zealand	Quantitative nonrandomized	SME[f] owners and managers	N=31 (n=14; n=17)	36.9	21 - 57	Stress Mindset Theory [71]	Rule-based decision-tree chatbot with keyword matching	A brief conversational agent intervention produced very large gains in a "stress-is-enhancing" mindset (eg, SMM[g] scores approximately doubled, *d*≈2.19) and large improvements in self-reported work performance (*d*≈1.59), with very high engagement (100% adherence and 86% wanting continued access) and participants attributing productivity gains to stress-reframing processes.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Baek and Cha, 2025 [72]	*Worldviews on Evidence-Based Nursing*	South Korea	Quantitative randomized controlled trials	Employed nurses	N=120 (n=40; n=40; n=40)	31.55; 31.88; 32.05	≥18	WHO[h]-based program integrating mindfulness, ACT[i], storytelling, and laughter therapy	AI recommendation algorithm using similarity scoring and neural-network updating	Across 4 weeks, the AI-supported intervention (and especially self-selected use) produced significant reductions in client-related and personal burnout and stress responses, with job stress decreasing in all conditions; stress-response reductions were strongest in the AI and self-selection groups, but not evident in the minimal-control group.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Cho et al, 2024 [73]	*Journal of Medical Internet Research*	South Korea	Quantitative nonrandomized	Employed nurses	N=603 (n=293; n=10; n=300)	32.81	≥18	WHO-based mindfulness, ACT, storytelling, laughter therapy modules	Collaborative-filtering recommender with artificial neural network updates	Burnout, job stress, and stress responses significantly decreased from baseline through posttest assessments (*P*<.001), with multicomponent modules (mindfulness, ACT, storytelling, laughter therapy) each improving different burnout facets and stress indices; satisfaction increased as the AI optimization progressed, and reductions in job stress and stress response were already evident after 2 weeks.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Fuhrmann et al, 2025 [74]	*Scientific Reports*	Germany	Quantitative randomized controlled trials	Mixed adult population	N=166 (n=56; n=56; n=54)	25.39; 23.25; 23.76	18 - 60	CBT-informed stress management (HRV[j] biofeedback+CBT elements)	Sensor-based signal processing using accelerometer-derived HR[k] biofeedback and facial emotion recognition	Both versions of the stress management app improved perceived stress, emotion regulation, and well-being, but the HR-based biofeedback version yielded significantly greater reductions in perceived stress than the waitlist control (significant time× condition interaction on PSS[l]), with high usability ratings.
Gabrielli et al, 2021 [75]	*JMIR mHealth and uHealth*	Italy	Mixed methods	University students	N=71	20.6	18 - 34	CBT, mindfulness, positive psychology	Scripted rule-based chatbot with predefined response options	Use of the app was associated with significant decreases in perceived stress (PSS-10) and severe anxiety symptoms, alongside improvements in mindfulness facets (describing, nonjudging); stress levels shifted from high to low categories over time, with moderate engagement (≈78 interactions; 58% completion).

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Hiller et al, 2025 [76]	*JMIR mHealth and uHealth*	Germany	Qualitative	Mixed adult population	N=61 (n=27; n=34)	30.5; 20.38	18 - 59	Ecological momentary intervention or just-in-time adaptive intervention with CBT and ACT components	Predictive time-series machine learning for adaptive EMI[m] delivery	A 4-week digital intervention significantly improved well-being and reduced perceived stress and depressive symptoms; qualitative feedback highlighted increased mental health awareness, the ability to identify stress triggers, and the value of ongoing structure, support, and emotional guidance.
Huang et al, 2022 [77]	*International Journal of Social Robotics*	Taiwan	Quantitative nonrandomized	Mixed adult population	N=122 (n=23; n=40; n=59)	25; 22.1; 23.87	—	Study-specific human-robot Interaction framework (ability, appropriateness, and validity)	Hybrid NLP system using knowledge graph, Bayesian network, neural network, and AIML[n] chatbot	Participants rated the robot's stressor inference accuracy highly (~4.4/5) and judged its emotional support responses as more appropriate than those of a baseline chatbot; they perceived the agent as understanding their feelings, providing valid and reliable emotional support, and making self-disclosure easier.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Indrayanti et al, 2025 [78]	*SSM–Mental Health*	Indonesia	Mixed methods	University students	N=32 (n=16; n=16)	20.21	19 - 22	PFA[o]	GPT-3 chatbot with NLP-based sentiment detection	Compared with controls, the PsyBot condition significantly reduced loneliness and improved well-being over 2 weeks, indicating that the chatbot was more effective than controls in alleviating social isolation.
Jiang and Yang, 2025 [79]	*BMC Psychology*	China	Quantitative randomized controlled trials	Middle-school students	n=65	15.24; 15.75	14 - 18	CBT (Elomia) with Mindfulness and PFA-derived components	Hybrid recommendation and chatbot system using NLP and machine learning	Relative to controls, the intervention produced significant and sustained reductions in somatic symptoms, anxiety and insomnia, social dysfunction, and severe depression, reflecting overall alleviation of psychological distress, while the control group showed no meaningful change.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Keung and So, 2025 [80]	*Frontiers in Psychology*	Hong Kong	Quantitative randomized controlled trials	University students	N=110 (n=55; n=55)	29.93; 28.13	18 - 71; 18 - 60	CBT-based supportive chatbot	GPT-4 or LLM-based counseling chatbot	Stress levels decreased from pre- to post-test in both AI and human support conditions, regardless of whether the chatbot's identity was revealed, but higher "perceptual fear" of AI predicted lower perceived support quality; those with lower fear evaluated the chatbot more favorably.
Ko and Woo, 2025 [81]	*JMIR mHealth and uHealth*	South Korea	Quantitative descriptive	Population not specified	N=7 (raters)	–	–	Mixed (CBT, health information provision, and physiological monitoring)	Mixed app ecosystem with algorithmic chat, AI mood checking, and recommender features	Among commercial apps, those providing rich mental health information, counseling services, and multi-provider details achieved the highest uMARS[p] scores (17/41 apps >4.0, range of 3.08 - 4.08) and were preferred by users, who particularly valued 24-hour access and not having to rely on a single provider.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Lai et al, 2025 [82]	*Journal of Nursing Research*	Taiwan	Quantitative randomized controlled trials	Employed nurses	N=80	33.6	≥18	Stress appraisal and coping theory+CBT	Hybrid GPT-supported chatbot embedded in structured intervention content	Relative to controls, the intervention produced significant improvements in sleep disturbance, perceived stress, anxiety, and depressive symptoms by 6-week follow-up (all $P<.001$); perceived stress was already lower in the intervention group by 3 weeks, whereas the control group showed a slight increase in stress over time.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Lee and Hahn, 2024 [83]	*Frontiers in Psychology*	South Korea	Quantitative nonrandomized	University students	N=*137*	23.3	18 - 49	Study-specific mind perception and support-type manipulation framework	Rule-based chatbot with Dialogflow ML[q] intent or emotion classification	Attributing "mind" to the chatbot (explicitly or implicitly) predicted higher perceived message effectiveness (explicit *d*≈1.09; implicit *d*≈0.50); emotional support reduced message effectiveness when users did not implicitly attribute mind, whereas informational support showed no such interaction, and participants described the interaction as facilitating introspection and self-evaluation.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Lee et al, 2025 [84]	*JMIR mHealth and uHealth*	South Korea	Mixed methods	University students	N=170 (n=85; n=47; n=38)	20.6; 20.2	–	CBT-based skills and behavioral support chatbot	Semigenerative chatbot combining KoGPT2 generation, KoBERT classification, CNN[r], and fixed CBT scenarios	Only the chatbot group showed significant improvements in time management skills, perceived stress (eg, PSS reductions of approximately 2.5 - 4 points), and procrastination, with particularly strong effects in a high-engagement cluster; control participants did not exhibit comparable improvements.
Li et al, 2025 [85]	*Journal of the American Psychiatric Nurses Association*	Hong Kong	Quantitative nonrandomized	University students	N=30	23.03	18 - 25	Mindfulness-Based Stress Reduction	Scripted rule-based chatbot using tree-graph dialogue, keyword spotting, conditional logic, and database memory	A fully chatbot-delivered audio mindfulness program produced large reductions in stress (PSS from 34.11 to 24.65; $d \approx -1.95$, $P<.001$) and significant gains in related well-being domains, with high adherence (~87%) and no adverse events; most participants reported improved sleep, mood, and emotional regulation.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Ly et al, 2017 [86]	*Internet Interventions*	Sweden	Mixed methods	University students	N=37 (n=14, n=14, n=9)	21.1, 25.4, 28.8	20 - 49	CBT+positive psychology	Rule-based chatbot using triggers, keywords, NLU[s], and lightweight database	After a 2-week intervention with CBT interface (SHIMS), only the adherence group (consistent use over 2 wk) showed significant interaction effects of group and time on perceived stress ($F_{1,27}$=4.30; P=.048), with a large-medium effect size (d=1.06).
Marwaha et al, 2025 [87]	*The Open Psychology Journal*	India	Mixed methods	University students	N=214	–	18 - 23	CBT+mindfulness	ChatGPT or LLM-based digital intervention	Following chatbot use, COVID-19 stress scores significantly declined (from 16.50 to 15.80; P<.01), and qualitative data indicated reduced fear of being judged, less felt need for counseling or "someone to talk to," and fewer felt stress-related mood swings, with concomitant improvements in sleep quality.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Medeiros et al, 2022 [88]	*IEEE Transactions on Human-Machine Systems*	United Kingdom	Quantitative randomized controlled trials	Mixed adult population	N=188 (n=86; n=102)	28.74	18 - 35	Emotion Regulation Theory (Gross) (study-specific implementation)	Hybrid chatbot using rule-based stressor mapping, NLP/text mining, Watson NLU, and ML classification	Compared with a neutral chatbot, interacting with a supportive chatbot produced modest reductions in perceived stress (~0.7 points; *P*≈.01), increased positive affect, and decreased negative affect. However, arousal-related stress indices did not show robust change.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Meng and Dai, 2021 [89]	*Journal of Computer-Mediated Communication*	United States	Quantitative randomized controlled trials	University students	N=211	20.4	–	Study-specific emotional support versus self-disclosure framework	Rule-based keyword-recognition chatbot with predefined scripts	Emotional support from the chatbot did not directly reduce stress (nonsignificant main effect) but produced significant reductions in worry and an indirect effect on stress via perceived supportiveness; receiving emotional support predicted stress reduction through feeling supported, whereas a chatbot that only self-disclosed (without emotional support) yielded the poorest stress outcomes.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Nelekar et al, 2022 [90]	*British Journal of Educational Technology*	India	Quantitative randomized controlled trials	University students	N=61	20.52	–	BDI[t] model+ explanation-based agent architecture	Explainable rule-based BDI-style embodied agent using option-based dialogue	Stress significantly decreased across explanation conditions (eg, belief group from 7.65 to 6.91, *P*≈.01), and behavioral intentions increased for multiple health behaviors depending on explanation type; participants also reported heightened awareness of their own stress through interacting with the AI explanation system.
Orji et al, 2025 [91]	*Behaviour and Information Technology*	Canada	Mixed methods	Mixed adult population	N=190 (n=95; n=72; n=23)	–	18 - 45	Positive psychology, PERMA[u], flow theory, PSD[v] model	Hybrid app combining smile detection (computer vision), ChatGPT API[w], and sentiment analysis	The intervention led to significant improvements in mood (BMIS[x] pleasantness, calmness, and positivity all increased, *d* ≈ 0.6), with 80% of interviewees reporting stress relief and relaxation; both quantitative and qualitative data indicated that the app effectively helped users feel calmer.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Park et al, 2025 [92]	*International Journal of Educational Technology in Higher Education*	South Korea	Mixed methods	University students	N=78 (n=50; n=28)	22.44; 22.6; 22.28	–	Self-reflection framework (self-disclosure–based design)	GPT-4o LLM chatbot with prompt engineering and self-disclosure manipulation	The self-disclosure chatbot elicited deeper reflective processing. This was seen in longer sessions, richer narrative content, and greater engagement across all reflection substages, which supported users' awareness of stress triggers and appraisal patterns. Although objective assessment accuracy did not differ significantly from the non-self-disclosure version, self-disclosure interactions were perceived as more emotionally resonant and conducive to stress clarification, helping users articulate stressors and evaluate coping options more effectively than the NSD[y] chatbot.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Park et al, 2019 [93]	*Journal of Medical Internet Research*	South Korea	Qualitative	University students	N=30	–	≥18	MI[z]	Rule-based ELIZA[aa]-style chatbot with keyword-based response generation	Participants generally appreciated evocative, context-sensitive questions that facilitated self-reflection and motivation and valued the chatbot's nonjudgmental stance and empathy toward graduate school stress; however, some responses were seen as repetitive or generic, and many participants desired more personalized and informational support.
Qi, 2025 [94]	*International Journal of Human-Computer Interaction*	China	Quantitative nonrandomized	University students	N=206 (n=86; n=120)	19.8	–	CBT+crisis intervention components	Scripted CBT chatbot with preprepared responses	The intervention significantly reduced DASS[ab]-Stress scores (~4.2-point decrease, *P*<.01) and negative affect (SCL-90)[ac], with most students reporting improved emotional clarity, although between-group differences in stress outcomes were not statistically significant in this sample.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Silva et al, 2020 [95]	*Journal of Medical Systems*	Portugal	Quantitative descriptive	Medical school students	N=83	22.13	17 - 38	Study-specific "EU-STRESS" system (physiological stress monitoring and prediction)	Sensor-based ML stress detection using HR/HRV and neural network classification	The system effectively detected examination-related stress: physiological markers (RR[ad] interval, SDNN[ae], RMSSD[af]) and PSS-13 scores all indicated significantly higher stress during examinations than at baseline; a neural network model achieved good discrimination (sensitivity~75%, specificity~78%) between examination and baseline states.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Sun et al, 2023 [96]	*Frontiers in Artificial Intelligence and Applications*	The Netherlands	Quantitative randomized controlled trials	Mixed adult population	N=107 (n=30; n=77)	–	≥16	MI, BCTs[ag], GET[ah], mindfulness	Hybrid chatbot combining scripted expert dialogue, NLU/NLG[ai], sentiment analysis, classification, and generation	Mindfulness-oriented chatbot conversations led to perceived stress reductions (belief in reduced stress increased; eg, stress ratings ≈3.6 vs neutral 3; *P*<.001) and strengthened beliefs in physical activity engagement for most barrier types, although in some cases control participants reported higher PA[aj] engagement beliefs than the intervention group.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Teague et al, 2025 [97]	*Behaviour & Information Technology*	Australia	Mixed methods	Mixed adult population	N=186 (n=53; n=10; n=43; n=80)	31.2; 31.23	24 - 39	CBT, mindfulness, BCT	Rule-based expert-system chatbot using hierarchical decision trees and menus	The father-focused app showed good usability (SUS[ak]≈ 75 - 81; MARS[al]≈3.8 - 4.3) and high engagement (~8.7 interactions per day); fathers and clinicians rated it as having high perceived impact on fathers' mental health awareness and support-seeking, with fathers reporting clearer mood awareness and benefits from mood and goal tracking.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Tong et al, 2025 [98]	*Journal of Medical Internet Research*	Hong Kong	Quantitative randomized controlled trials	Mixed adult population	N=285	26.45	18 - 61	CBT, mindfulness, behavioral activation, positive psychology	Rule-based decision-tree chatbot with predefined categorical choices	The intervention produced small to moderate improvements in self-care behaviors (d≈0.36), mindfulness (d≈0.37), depressive symptoms (d≈−0.26), and positive emotions (d≈0.28); stress decreased over time in both intervention and control groups, yielding a main effect of time but no significant group× time interaction for stress.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Troitskaya and Batkhina, 2022 [99]	*Family Process*	Russia	Quantitative randomized controlled trials	Mixed adult population	N=156 (n=58; n=98)	31.0; 29.24; 31.08	18 - 54	CBT, mindfulness, positive psychology	Decision-tree chatbot with branching scripts and stored user variables	Over 21 days, participants using iCognito showed significant reductions in psychological distress (DASS-stress decreased by~2.7 points, *P*<.01) and qualitative reports of better emotional regulation and communication; there were moderate effects on relationship satisfaction and conflict reduction, although some interaction effects (eg, on quarrels and emotional intelligence) were not significant.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Weng et al, 2024 [100]	*Journal of Medical Internet Research*	Singapore	Quantitative descriptive	Mixed adult population	N=786	–	≥18	CBT, behavioral activation, mindfulness, positive psychology, motivational interviewing	Hybrid platform combining chatbot, screening algorithm, and CBT-based digital tools	Usage analytics indicated high engagement with grounding, sleep, and breathing tools, and ~57% of users reported feeling better after chatbot sessions, with greater improvements linked to completing structured CBT tools; however, outcomes were based on anonymous, uncontrolled usage data, and no RCT[am] evidence was available.
Williams et al, 2021 [101]	*Social Sciences*	New Zealand	Mixed methods	University students	N=256 (n=124; n=132)	19.9	18 - 25	CBT, DBT[an] skills, mindfulness, behavioral activation, positive psychology	Rule-based chatbot using quick reply options and branching paths	Participants showed moderate to large pre-post improvements in stress (PSS-10 reduction of ~5.7, *P*<.001) and well-being (WHO-5), alongside modest sleep improvements and high adherence (median 19.5 of 21 days completed), with effect sizes in the small-medium range (≈0.38 - 0.49).

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Yong, 2025 [102]	*Journal of Technology in Behavioral Science*	France	Quantitative descriptive	University students	n=10	–	17 - 21	Study-specific emotional-AI sentiment-detection+ affective computing framework	Multimodal emotional-AI system combining ML, FACS[ao] computer vision, IoT[ap] sensors, and smart home orchestration	AI-guided sessions improved positive affect (PANAS+6.1) and reduced negative affect (~−4.3), with users reporting relaxation, clearer mood, and interest in more personalized recommendations; however, the evidence was primarily feasibility- and experience-focused, with limited formal hypothesis testing, and the broader work emphasized integrating AI-infused mental health supports into the home.

XSL•FO
RenderX

Author, year of publication	Journal or publication	Country/Region	Study type	Study pool	Sample size	Mean age (y)	Age range (y)	Intervention framework	AI[a] architecture used	Key findings
Yoon et al, 2024 [103]	*BMJ Health & Care Informatics*	Singapore	Mixed methods	Mixed adult population	N=330 (n=278; n=12; n=40)	–	18 - 60	CBT, DBT, mindfulness, behavioral activation, motivational interviewing	AI-enabled platform using chatbot with algorithmic self-assessment and resource recommendation	Platform logs showed strong engagement with breathing, grounding, reframing, and sleep tools; survey and interview data indicated reduced stress, improved emotional regulation, and high perceived usefulness, with most users finding the platform helpful (≈71%) and reporting mood improvements (≈62%); anonymity, convenience, and immediate access were valued, particularly for managing work-related stress.

XSL•FO
RenderX

[a]AI: artificial intelligence.
[b]Not applicable.
[c]CBT: cognitive behavioral therapy.
[d]NLP: natural language processing.
[e]LLM: large language model.
[f]SME: small-to-medium-sized enterprise.
[g]SMM: Stress Mindset Measure.
[h]WHO: World Health Organization.
[i]ACT: Acceptance and Commitment Therapy.
[j]HRV: heart rate variability.
[k]HR: heart rate.
[l]PSS-10: Perceived Stress Scale (10 item).
[m]EMI: ecological momentary intervention.
[n]AIML: Artificial Intelligence Markup Language.
[o]PFA: Psychological First Aid.
[p]uMARS: user version of the Mobile Application Rating Scale.
[q]ML: machine learning.
[r]CNN: convolutional neural network.
[s]NLU: natural language understanding.
[t]BDI: belief-desire-intention.
[u]PERMA: Positive Emotion, Engagement, Relationships, Meaning, Accomplishment model.
[v]PSD: Persuasive Systems Design model.
[w]API: Application Programming Interface.
[x]BMIS: Brief Mood Introspection Scale.
[y]NSD: non-self-disclosure.
[z]MI: motivational interviewing.
[aa]ELIZA: "ELIZA-style chatbot", an early natural language processing computer program, reacted to user input without the capacity for logic-driven decision making, made use of pre-programmed responses.
[ab]DASS: Depression, Anxiety, and Stress Scale.
[ac]SCL-90: Symptom Checklist-90.
[ad]RR: R-wave to R-wave interval.
[ae]SDNN: standard deviation of normal-to-normal intervals.
[af]RMSSD: root mean square of successive differences.
[ag]BCT: behavior change technique.
[ah]GET: Graded Exercise Therapy.
[ai]NLG: natural language generation.
[aj]PA: physical activity.
[ak]SUS: System Usability Scale.
[al]MARS: Mobile Application Rating Scale.
[am]RCT: randomized control trial.
[an]DBT: Dialectical Behavior Therapy.
[ao]FACS: Facial Action Coding System.
[ap]IoT: internet of things.

Several trends are evident across the included records. First, the records were predominantly published after 2020, suggesting an increase in interest in the use of AI-enabled services for self-initiated personal stress management. Second, the included records comprised studies using qualitative research methods, as well as studies incorporating experimentation and randomized controlled trials. This indicates that the field is moving beyond early feasibility work, although the evidence base remains uneven. Subsequent research can thus build on established experimental findings and protocols to assess the efficacy of AI-enabled stress management strategies. Finally, the included records span 20 countries across multiple regions, suggesting that stress management is a concern of broad international relevance and that AI-enabled stress management strategies have been examined across diverse cultural contexts.

Quality Appraisal

The methodological quality of the 35 included studies was assessed using the Mixed Methods Appraisal Tool (MMAT) [104], which appraises qualitative, quantitative, and mixed methods studies using 5 design-specific criteria. Two independent reviewers (the first and second authors) evaluated each study. Initial interrater agreement was 83.8%, and discrepancies were resolved through discussion. The full MMAT

XSL•FO
RenderX

ratings, criteria, and coding framework applied in this review are provided in Multimedia Appendix 2.

What Emerging Roles Do AI-Enabled Tools Play in Self-Directed Stress Management?

Overview

From our included records, we identified the following themes: (1) psychological intervention, use of AI to administer psychological aid for individuals independent of a clinical or medical practitioner; (2) behavioral support, where AI agents support users in making intended behavioral changes to manage stress; (3) psychoeducation, where AI-enabled agents are equipped with resources to provide users access to information about their current stressors, how to manage stress and stressors, and access to multiple sources of medical intervention; (4) companionship and emotional support, where AI-enabled agents can be present for users in moments of high stress or need, and readily provide emotional support for users; and (5) stress monitoring, detection, and triage, where AI-enabled agents assist users in monitoring personal stress levels through language detection, physiological signals when paired with wearables, and provide advice for users' next best steps. For clarity, the term AI-enabled agents is used as an umbrella term to refer to chatbots, conversational agents, and AI-driven applications.

Psychological Intervention

Making use of AI-enabled agents for immediate care in high stress or prolonged stressful situations without needing to wait for a clinician or medical practitioner is a prominent benefit of AI-enabled agents. Thirteen articles found that AI can be used for stress management through providing psychological intervention enabled by AI agents [70,72-74,76,79,87,88,91-94,100]. Within these studies, AI architectures included rule-based or scripted conversational systems [70,93,94]; machine learning, recommendation, or predictive NLP-based systems [72,73,76,79,88]; LLM or generative artificial intelligence (GenAI) systems [87,92]; sensor, computer vision, or multimodal hybrid systems [74,91]; and AI-chatbot plus screening or recommendation systems [100].

In general, these AI agents have been shown to be effective in delivering timely interventions. For example, Allan [70] showed that a brief conversation with an AI chatbot shifted users' mindsets from viewing stress as harmful to viewing stress as enhancing ($d \approx 2.19$), and users attributed productivity gains to this mindset change. In high-stress work environments, such as hospitals, participants who engaged with AI-supported interventions reported significant reductions in burnout, job stress, and stress responses over a 4-week period [72,73]. AI-enabled interventions for high-stress work environments were particularly valuable to users, given the fast-paced nature of their jobs, which made it unlikely that they would seek clinical intervention. Stress-response reductions were the strongest in the AI-supported intervention group [72], which supports the efficacy of immediate intervention.

AI-enabled interventions can also be combined with other forms of technology, such as biofeedback through wearables, where detecting a change in user physiology and administering immediate intervention yielded greater stress reductions in perceived stress [74]. This demonstrates the flexibility in how stress management interventions can be delivered without a clinician present. Immediate use of AI-enabled agents to mitigate stress also has positive effects. Marwaha et al [87] found a significant decline in COVID-19–related student stress scores after interacting with the chatbot compared to baseline values, with several participants qualitatively reporting reduced fear of peer judgment and decreased stress-related mood swings. Moreover, the findings of Hiller et al [76] highlight that the "just-in-time" ability of AI agents to act as a considerable merit, where users interacting with the AI-enabled system, *mila*, also saw significantly improved well-being and reduced perceived stress and depressive symptoms, with qualitative reports of improved ability to identify triggers when supported by the chatbot.

Beyond the immediate deployment of intervention, prolonged support with AI-enabled agents also benefits users by improving mood and affect. Weng et al [100] showed that users who engaged with the AI support chatbot *Wysa*, hosted on the website platform mindline.sg, repeatedly reported feeling better after chatbot sessions. Similarly, users who interacted with *SmileApp* reported significant improvements in mood, experiencing stress relief and relaxation, thereby helping them feel calmer when dealing with stressors [91].

In the context of psychological intervention, a significant number of AI-enabled systems implement evidence-based strategies derived from established frameworks such as cognitive behavioral therapy [74,76,79,87,94], mindfulness [72,73,79,87,94], and Gross's emotion regulation theory [88]. The flexibility of chatbots to be programmed with various psychological aids and frameworks to address different scenarios and stressors demonstrates a significant use case for AI-supported agents in stress management and in catering to each user's needs. The available evidence supports AI-enabled agents as a feasible means of delivering psychological intervention content in the reviewed contexts, particularly when systems are designed around established psychological frameworks.

Behavioral Support

AI-enabled agents can also help users achieve intended behaviors that mitigate stress. Ten studies found uses for AI to assist in helping users achieve behavioral changes [68,75,82,84-86,98,99,102,103]. Among the behavioral support studies, AI architectures included rule-based, scripted, or decision tree chatbot systems [75,85,86,98,99]. In these studies, AI agents primarily took the form of online chatbots, being delivered via websites or mobile apps [103] and even integrated with smart home technology [102]. Other used AI architectures include hybrid GPT or semigenerative systems combined with structured intervention content or ML classifiers [82,84] and multimodal, AI-enabled, or ML-based assistive or recommendation systems [68,102,103]. AI-supported agents were often programmed to address behaviors that contribute to stress or could improve stress outcomes. For example, Lai et al [82] showed that a 6-week intervention using an AI agent based on stress appraisal and coping theory encouraged healthier sleep

XSL•FO
RenderX

habits and adaptive coping strategies, resulting in significant improvements in perceived stress, anxiety, and depressive symptoms. Participants also reported that, following interaction with the chatbot, they experienced improved sleep, mood, and emotion regulation. Another targeted behavior was procrastination. Lee et al [84] reported that engagement with a cognitive behavioral thCognitive Behavioural Therapy (CBT)–based supportive chatbot (*Moa*) was associated with greater improvements in time management skills, perceived stress, and procrastination behaviors compared with a control condition.

Beyond self-improvement domains, AI agents have helped users identify their personal conflict styles and adjust their approaches to interpersonal disagreements. For instance, Troitskaya and Batkhina [99] showed that over 21 days, using an AI-supported application was associated with a considerable reduction in psychological distress for couples. Qualitative reports indicated improved emotional regulation and communication in their relationships [99]. Consistent use of AI support was also associated with significant decreases in perceived stress while yielding improvements in mindfulness and self-care behaviors [75,84,98]. Taken together, these findings suggest that AI interventions can be effective aids in helping users make lasting behavioral changes in targeted areas, supporting them in discontinuing negative habits and encouraging positive ones. Furthermore, immediate access to AI platforms promotes greater accountability and user engagement, ensuring users remain on track to achieve their intended goals.

Psychoeducation

AI-enabled technology has also been applied to help users learn more about stress and stress mindsets, identify stressors, and access resources when experiencing stress. Four articles found that AI-enabled agents helped users through providing education about stress mindsets and assisting users to understand indicators of stress and how to mitigate harmful effects [69,81,97,101]. AI architectures used for psychoeducational interventions included rule-based chatbot or expert system designs [97,101], LLM- or GenAI-based support [69], and mixed app ecosystem architectures involving algorithmic chatting, AI mood checking, or possible AI recommendation features [81]. The psychoeducational functions were primarily delivered through chatbots trained on established psychological frameworks, often including CBT-informed elements, which participants engaged with independently [69,81,97,101]. For instance, Alanezi [69] reported that participants used AI agents to learn about stress and strategies for managing stressors and described experiences such as feeling heard and supported during interactions with the chatbot. The same study further noted that participants valued the ease of access to mental health resources, as well as features such as CBT-informed cognitive reframing, coping guidance, and stress self-monitoring [69]. AI-based psychoeducation can also be tailored to specific contexts or populations. Teague et al [97] reported that fathers of newborn children who interacted with a chatbot (*Rover*) trained in cognitive behavioral therapy, mindfulness, and behavior change techniques described greater awareness of their mental health and increased support-seeking. The chatbot provided information tailored to the perinatal context and was reported to support fathers in identifying, understanding, and processing mood changes associated with caring for a newborn. Participants used these outputs to contextualize their stress levels in relation to demands and plan strategies with the AI to better cope with increased stress at predicted times. Commercial apps with AI support marketed for mass population use also have the potential to impart large amounts of information to the general population. Ko and Woo [81] found that among highly rated mental health apps with AI-enabled support, frequent users reported 24-hour access and multiple sources of support and information, a significant advantage over relying on a single medical provider. Users also had access to extensive mental health information through the app, which could be tailored to their specific needs.

Moreover, findings suggest that AI-enabled technologies can effectively address population-specific stress management needs by customizing information for different groups while simultaneously broadening access to mental health information and support. Williams et al [101] reported that consistent interaction with a chatbot informed by cognitive behavioral therapy, positive psychology, and mindfulness was associated with statistically significant pre-post improvements in well-being (5-item World Health Organization Well-Being Index) and perceived stress (Perceived Stress Scale-10). These changes corresponded to small-to-medium effect sizes (approximately 0.38 - 0.49) and were accompanied by self-reported improvements in adherence to healthy sleep habits and sleep quality. In summary, AI-facilitated psychoeducation can empower users to independently monitor their stress levels and manage their stress responses more effectively by teaching them evidence-based strategies. By customizing educational content to the user's context and providing on-demand support, AI tools may help increase users' literacy about stress and coping, which is a critical component of long-term stress management.

Companionship and Emotional Support

Users can make use of AI-enabled agents for immediate emotional support and companionship during high stress periods. Several studies highlighted the potential for AI agents to serve as readily available sources of emotional support [78,80,83,89,90]. Within these 5 studies, AI architectures were mixed. Two centered around LLM or GenAI conversational systems [78,80], one used a rule-based chatbot augmented by ML-based Dialogflow classification [83], one used a rule-based keyword recognition chatbot with predefined scripts [89], and one used an explainable rule-based belief-desire-intention–style embodied conversational agent [90].

The immediacy of AI-based support makes these agents valuable, especially for individuals who may be unable or reluctant to seek human support through other means. For example, Indrayanti et al [78] reported that the *PsyBot* chatbot supported users experiencing negative emotions during periods of distress, with participants describing the platform as providing a supportive space for emotional expression and feeling heard. Similarly, Nelekar et al [90] reported that participants experienced significant reductions in stress after interacting with an AI chatbot designed to facilitate conversations about users' beliefs, desires, and life goals, alongside increased

XSL•FO
RenderX

awareness and understanding of their own stress through the system's explanation-based features.

Although participants generally report benefits from AI-mediated emotional support, some frameworks do not produce a direct reduction in stress; instead, stress may decrease indirectly via reduced worry when the chatbot increases perceived support and prompts self-disclosure [89]. In Meng and Dai [89], the Chatfuel-based supportive chatbot showed no direct effect on stress but demonstrated an indirect effect via perceived support and associated worry reductions, consistent with its design to provide emotional support or elicit self-disclosure. Within this theme, interventions grounded in established frameworks, such as *PsyBot* (Psychological First Aid) and *ARU* (Belief–Desire–Intention model with an explanation-based architecture), were associated with reductions in negative affect and stress and, in some cases, increases in positive affect; such findings highlight the value of validated psychological frameworks in guiding chatbot design in this context. Overall, the immediacy and availability of AI agents provide users with rapid access to emotional support, which may contribute to improvements in well-being over time.

Stress Monitoring, Detection, and Triage

AI-enabled technology can also enable users to monitor personal stress levels independently, without a clinician or other medical professional, and to detect changes in these levels before users are aware of them. Three studies fall into this category [77,95,96]. Huang et al [77] created a bespoke human-robot interaction system for adults (the *RoBoHoN*) that engaged participants in conversation. Across stress monitoring and triage studies, AI architectures included physiological or sensor-based ML systems [95] and hybrid NLP, ML, or generative systems for stressor and barrier inference [77,96]. As such, the 3 studies in this category used physiological, sensor-based, or conversational data, rather than self-reported symptom scales, to support timely remote self-intervention recommendations. Users perceived this robot as an agent capable of understanding their feelings, providing valid and reliable emotional support in moments of distress, and facilitating self-disclosure to identify causes or effects of stress. Another study by Sun et al [96] tested 6 chatbots each trained with a different therapeutic framework (eg, motivational interviewing and mindfulness). Participants in the mindfulness-oriented chatbot condition showed a significant reduction in perceived stress after the intervention, demonstrating that early detection and mitigation of stressors can be achieved with AI agents [96].

In addition to conversational monitoring approaches, Silva et al [95] described the *EuStress* solution, an AI-enabled information system designed for continuous, real-time stress assessment using wearable-derived physiological signals and machine learning classification. Consistent with a physiological stress monitoring approach, the system assesses the stress levels of the students using a wearable device to collect stress-related data with minimal user interaction. The solution was intended to interpret the stress reactivity patterns of the individual and predict stress states, cumulative effects of stress, and chronic stress, positioning the intervention as a detection-and-triage tool that translates biometric data into actionable stress-state inference [95]. While the independent use of AI-enabled systems for self-monitoring and self-management of stress is limited, the efficacy of AI in stress monitoring is supported, serving as a just-in-time strategy rather than allowing a stressful situation to spiral out of users' control.

All 3 key studies examining AI for stress monitoring, detection, and triage [77,95,96] indicate that AI shows potential for effectively monitoring stress-related indicators and supporting stress detection and triage. *RoBoHoN*, for instance, was an interactive robotic system that inferred users' stressors from conversational cues and was rated highly for inference accuracy [77]. These findings indicate that AI technologies can be trained to detect subtle cues indicative of stress, even before users are consciously aware of their stress. Importantly, user experiences with *RoBoHoN* were positive. Participants reported that the interaction was enjoyable and that the agent appeared to understand their feelings. They also described the agent as providing valid and reliable emotional support, which facilitated emotional self-disclosure and contributed to stress relief. In a study by Sun et al [96], a multichatbot study showed that reductions in perceived stress were accompanied by strengthened beliefs in physical activity as a stress mitigation strategy. This finding suggests that AI-based interventions may also influence users' coping beliefs and attitudes, specifically by reinforcing exercise as a coping strategy when stress is detected. Moreover, the *EuStress* system described by Silva et al [95] further demonstrated that stress-linked physiological patterns could be meaningfully distinguished in real-world contexts, such as baseline periods versus examination periods among students. ML models were able to classify stress states with promising performance using wearable sensor data. The authors noted practical constraints, including limited sample sizes and data gaps, that affected model robustness and generalizability. These observations underscore that although AI models can detect stress, they require sufficiently large and diverse datasets to ensure reliable performance across users and contexts.

Discussion

Overview

The sample characteristics indicate that the included records were drawn from a reasonably broad range of participant groups but were concentrated mainly in university student samples and mixed adult populations. Several studies also focused on more specific groups, including employed nurses, small-to-medium-sized enterprise owners and managers, medical students, fathers, and children with Down syndrome and their caregivers. This suggests that the evidence base spans educational, occupational, family, and caregiving contexts. Sample sizes varied widely across the included records, from small qualitative or feasibility samples to larger intervention, survey, and platform-based samples. This pattern is consistent with the methodological range of the review, which included qualitative, mixed methods, descriptive, and trial-based studies. Where age was reported, the samples were weighted mainly toward young adults, especially in student-focused studies. Student samples were generally concentrated in the late teens to 20s, whereas occupational- and community-based samples

XSL•FO
RenderX

more often fell in the late 20s to 30s. Most studies included participants aged 18 years and older, although a smaller number included adolescents or child-caregiver samples. Overall, the evidence base reflects a range of applied contexts and participant groups while remaining centered largely on students, working adults, and other accessible community populations that are well suited to the use and evaluation of AI-based stress support tools.

This review synthesized 35 studies, comprising qualitative studies (n=3), randomized controlled trials (n=11), nonrandomized quantitative studies (n=6), quantitative descriptive studies (n=4), and mixed methods studies (n=11). Collectively, the included studies examined AI-enabled tools for self-directed stress management, including chatbots, mobile apps, embodied agents, and other digital platforms. Overall, these interventions show preliminary promise as scalable and accessible supports for psychological intervention, behavioral support, psychoeducation, companionship and emotional support, and stress monitoring. However, the evidence base remains early-stage and heterogeneous, so the findings should be interpreted cautiously. Quality appraisal showed that limitations tended to cluster by study design rather than by individual study. All included studies met the 2 MMAT screening criteria (ie, “Are there clear research questions?” and “Do the collected data allow addressing the research questions?”; [104]). However, methodological quality was more mixed across the design-specific criteria. Among the qualitative studies, all 3 met all 5 study type–specific MMAT criteria. Among the randomized controlled trials, most met criteria for randomization, baseline comparability, and outcome completeness, although blinding was less consistently established and participant adherence was unclear in 1 study. For the nonrandomized quantitative studies, the main limitations were inadequate control of confounding and, in some cases, limited representativeness of the target population. For the quantitative descriptive studies, statistical analysis was generally appropriate, but measurement appropriateness and nonresponse bias were less consistently addressed. For several mixed methods studies, the main weaknesses were incomplete integration of qualitative and quantitative components and limited consideration of inconsistencies between the 2 strands. Overall, the studies were sufficient to identify emerging roles of AI in self-directed stress management, but the methodological variability warrants cautious interpretation of effectiveness and generalizability going forward.

Key Findings

Overview

This review synthesized evidence on how AI can assist in self-directed stress management, identifying 5 distinct functional roles of AI-enabled agents. AI-enabled stress management systems took multiple forms, with several studies using chatbots that delivered support based on existing psychological frameworks, including CBT-informed approaches [74,76,79,87,94] and mindfulness-based programs such as Mindfulness-Based Stress Reduction [87]. Other studies evaluated web- and app-based platforms that provided self-guided tools and conversational support [75,91,100,103]. In addition, certain systems used physiological monitoring or biofeedback to support stress regulation [74,95]. These tools are largely designed to function as on-demand support for users in need, providing alerts that inform them when they experience elevated physiological symptoms indicative of stress and offering companionship in moments of emotional distress. AI-based support can be made continuously available, and because many agents are designed around established psychological frameworks, they may support users in managing immediate stressors, although evidence of effectiveness remains mixed and study dependent. Our analysis revealed clear distinctions in how AI-integrated solutions enable self-directed stress management. Each of the 5 identified themes represents distinct mechanisms or use cases for AI agents, assistants, software, and interventions. In synthesizing the emerging field, we illustrated the broad impacts and the specific ways in which AI helps users, from tracking daily habits to coping with persistent workplace stress.

Furthermore, the diversity of AI functions observed across studies was also reflected in the technical architectures and classification frameworks used to describe them. Rule-based and decision tree systems were commonly used for structured tasks, such as CBT-style exercises, mindfulness sessions, psychoeducation, and behavioral prompts. LLM- or GenAI- and NLP-enabled systems appeared more relevant where the intended role involved open-ended conversation, emotional support, or companionship. Sensor-based, physiological, computer-vision, and multimodal ML systems were more prominent when AI was used to infer stress states, emotional cues, or behavioral patterns. Many of the discussed interventions systems also integrated multiple AI approaches, making broad labels such as “chatbot” or “AI app” insufficiently precise. Although existing taxonomies classify AI systems by service design characteristics, human-AI activities [eg, 105,106], or broader risk and evaluation considerations, these frameworks remain wider in scope than this review [eg, 107,108]. By organizing studies according to AI’s functional role in self-directed stress management, this review provides a clearer and more use case–specific basis for interpreting outcomes, clarifying what each system is intended to do and avoiding the treatment of “AI” as a single undifferentiated category. In particular, a psychoeducational system, for example, may be assessed by whether it delivers accurate and understandable information; a behavioral support tool by adherence and behavior change; a psychological intervention by stress-related outcomes; a companionship and emotional support system by perceived support and emotional relief; and a monitoring or triage system by detection accuracy, timeliness, and appropriate feedback [109,110]. Across the included studies, AI served distinct functions in self-directed stress management, delivering psychological interventions, supporting behavior change, providing psychoeducation, offering emotional companionship, and monitoring or triaging stress-related states, rather than functioning as a single intervention type. This role-based structure clarifies each system’s intended purpose and offers a more meaningful basis for evaluating efficacy than AI architecture alone.

XSL•FO
RenderX

Traditionally, stress has been viewed as a negative subject, with extreme stress labeled as the cause of physical ailments and declining mental health [71,95,111]. Our findings, however, suggest that AI-based strategies aid in reframing stress in more constructive ways. For example, the use of AI tools can help individuals to reframe stress as a motivating factor [70] or as a warning sign of burnout [72], with readily available access to AI-enabled agents reducing another layer of stress in seeking clinical intervention or professional help [72]. AI-enabled stress management strategies also serve as a gateway to increasing mental health literacy about personal stressors [97], allowing users to understand their experiences from a different perspective. The functional roles identified in our review provide a clear foundation for further research on AI and stress management and will help develop more targeted solutions for each individual. Furthermore, the impact of stress-related illnesses extends beyond immediate negative affect. Stress, when unmanaged on a widespread scale, can snowball into enormous health care costs, loss in productivity at work, and, in severe cases, amount to physical disabilities [7]. The ability of AI-enabled technology to intervene and provide assistance to individuals before their stress levels reach levels of psychological or physical harm is a significant use case for AI and could be used meaningfully to help individuals manage their stress in healthy ways, while remaining accessible at any time.

Research Gaps and Limitations

Through this systematic synthesis, we identified several gaps in the existing literature on AI-supported self-directed stress management. First, the evidence base is dominated by early-stage pilot and feasibility studies. Although many of these studies report short-term improvements in stress-related outcomes, few evaluate mature or clinically validated systems, and distinctions between AI-enabled interventions and broader digital interventions are often insufficiently specified. As a result, evidence regarding long-term effectiveness remains limited, and this literature does not yet support meaningful estimation of pooled effect sizes. In addition, the available evidence may be skewed toward positive findings, reflecting both the novelty of AI-based interventions and broader publication biases. Second, we observed substantial heterogeneity across AI models, psychological frameworks, stress outcome measures, and evaluation methods. While this diversity reflects the field's interdisciplinary nature, it limits comparability and makes it difficult to determine which approaches are most effective. This challenge is further compounded by the absence of standardized terminology across studies, as rapidly evolving AI capabilities and inconsistent system labeling make it difficult to systematically capture, compare, and track meaningful distinctions between intervention types over time. Third, we identified that the existing study samples are narrow and demographically skewed, with most studies involving young adults or university students. This pattern is expected given typical technology adoption trends [112], but it limits generalizability to older or less technologically proficient populations. Factors such as digital literacy and frequency of use may have differing effects on populations that were not captured within this study. Next, issues of safety, fairness, and reliability have been underexamined. Few studies systematically examined erroneous or inappropriate AI outputs, algorithmic bias, or differential performance across demographic groups. Reporting practices also tended to emphasize favorable outcomes, with comparatively little discussion of system failures, unintended effects, or potential harms. Addressing these gaps will be critical as AI-based stress management tools move toward broader real-world deployment. Finally, our search was limited to English-language publications, which introduce the potential for language bias. Both authors' first language is English, which limits the representativeness of the evidence base outside of English.

Future Directions

To move the field of AI-based stress management beyond the scope of this thematic synthesis, future research should prioritize the development of a cumulative, methodologically rigorous evidence base. First, long-term randomized controlled trials are needed to assess whether stress reduction effects and behavioral changes are sustained beyond the short durations (often less than 3 months) reported in existing studies. Second, as the literature matures, meta-analytic work will be essential for estimating pooled effect sizes and identifying moderators such as age, cultural context, delivery modality, and underlying psychological frameworks. Achieving this will require more standardized outcome measures and longer follow-up intervals. Third, future studies should investigate user-AI interaction processes, including perceived mind attribution, trust, empathy, and fear of AI, as these dimensions appear to influence engagement and stress-related outcomes, in addition to examining the effects of such AI usage patterns on different age groups and demographics. Using validated scales of AI-specific individual differences, such as measures of trust, dependency, or perceived agency, would help clarify these mechanisms. Fourth, methodological standards must be strengthened across disciplines. As observed in this review, reporting practices were highly variable, and only a small subset of studies adhered to established AI-specific guidelines such as CONSORT-AI (Consolidated Standards of Reporting Trials - Artificial Intelligence) [113] and SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials - Artificial Intelligence) [114]. Future work should therefore prioritize transparency, include protocols for error detection and explainability, and adopt mechanisms for real-time error handling to support replicability and user safety. Fifth, improving the diversity of datasets is critical to mitigate algorithmic bias and ensure that AI systems perform equitably across populations, particularly those currently underrepresented. Sixth, the cultural adaptation of AI tools warrants systematic attention. This includes accounting for linguistic variation, culturally specific stressors, and local norms in the expression of stress and coping. The ability of different AI tools to be personalized and scaled to each individual's needs is an area that could be explored in the future. Across the included studies, none tested age-related moderation effects. Digital literacy was not consistently considered, measured, or modeled as a covariate. One study considered digital literacy as familiarity with AI [68] but did not conduct a moderation analysis. Future research should examine whether age, digital literacy, and

XSL•FO
RenderX

technological familiarity moderate engagement, acceptability, and effectiveness, given the current skew toward younger and technologically savvy populations. Seventh, research and health care organizations should begin evaluating AI-based stress interventions in real-world settings. In particular, studies set in workplaces, educational institutions, and community health platforms would be ideal for examining the extent to which emerging human-AI models improve engagement, accountability, and sustainability. Furthermore, multimodal AI systems that integrate conversational data with physiological and behavioral signals hold promise for improving ecological validity and enabling more dynamic, real-time stress detection and intervention. Together, these directions offer a roadmap for advancing the field toward more robust, inclusive, and context-sensitive AI applications in self-directed stress management.

Finally, regulation and legislation surrounding AI applications for stress management should be strengthened, especially concerning data security for clinically vulnerable populations. While AI may be beneficial to individual users, providers and governing bodies must be held accountable, ensuring users' data is not compromised or shared without consent, so as not to harm users in the process. Existing AI guidance frameworks, including the Organisation for Economic Co-operation and Development AI Principles [115], the European Commission's Ethics Guidelines for Trustworthy AI [116], and the ASEAN Guide on AI Governance and Ethics [117], highlight principles such as transparency, safety, privacy, human oversight, fairness, and accountability. However, these frameworks offer broad governance guidance rather than reporting standards specific to AI use in nonclinical psychological research. A potential alternative could be to adapt principles of ethical use of AI from existing sources before deploying AI in research and scaling the use of AI systems in consumer-facing products. At present, psychology-specific reporting guidance for AI use in nonclinical research appears limited. Frameworks such as CONSORT-AI [113] and SPIRIT-AI [114] offer relevant examples for AI-related reporting in clinical trial reports and protocols, but they are not general regulatory or accountability frameworks. Researchers could therefore draw cautiously on existing AI governance principles when designing studies or developing consumer-facing AI tools, particularly where human participants are involved. Human oversight may also be useful to help researchers monitor AI outputs and respond to unexpected issues during deployment.

Conclusions

This systematic review consolidates current evidence on how AI supports self-directed stress management. We illustrated 5 distinct ways in which intelligent systems can assist users in managing stress, ranging from monitoring and intervention to behavior change and emotional support. Across these applications, one central insight emerges: several AI stress management interventions were informed by user-centered design principles or psychological frameworks that emphasize coping, behavior change, mindfulness, or emotional support [72,73,76,80,82,83]. As AI systems learn directly from their training data, model validation and reliability testing are as crucial as development studies on novel systems. Small, homogenous, or biased samples risk producing models that misrepresent or overlook entire segments of users [118,119]. As such, future work must give equal weight to large-scale, diverse data collection and rigorous validation studies, ensuring that AI tools trained for stress management generalize safely, equitably, and across contexts. At the same time, emerging techniques such as digital phenotyping [43,44] offer a promising path forward. By matching users to interventions that align with their behavioral and symptom profiles, such approaches directly address low engagement stemming from perceived unmet needs. Ultimately, these synthesized findings have wide-reaching implications across cultures, settings, and mental health domains. AI systems hold preliminary promise for expanding access and enhancing stress self-management. As these tools increasingly operate without clinician oversight, standards such as transparent communication of model capabilities, safeguards, and reporting requirements are critical to protect users while realizing the full potential of AI in self-directed interventions.

Acknowledgments

MKGR and SSMT should be considered joint first authors. The order of co-first authors provided was randomized. Co-first authors can prioritize their names when adding this paper's reference to their resumes. The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the Generative Artificial Intelligence Delegation Taxonomy (GAIDeT; 2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing. The GenAI tool used was ChatGPT-5.2. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

Andree Hartanto was supported by Singapore Management University research grants from the Ministry of Education Academic Research Fund Tier 1 (25-SOSS-SMU-006) and the Lee Kong Chian Fund for Research Excellence.

Authors' Contributions

MKGR contributed to conceptualization, formal analysis, investigation, methodology, visualization, writing the original draft, and review and editing of the manuscript. SSMT contributed to conceptualization, formal analysis, investigation, methodology,

XSL•FO
RenderX

visualization, writing the original draft, and review and editing of the manuscript. AH contributed to the conceptualization, supervision, and review and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Title and abstract screening log and full-text screening log (record-level decisions), including inclusion or exclusion decisions and reasons for exclusion at full-text stage.
[XLSX File, 803 KB - jmir_v28i1e90709_app1.xlsx]

Multimedia Appendix 2
MMAT (2018) quality appraisal worksheet, documenting study-level methodological quality assessments for included studies.
[XLSX File, 18 KB - jmir_v28i1e90709_app2.xlsx]

Checklist 1
PRISMA checklist.
[DOCX File, 32 KB - jmir_v28i1e90709_app3.docx]

References

1. Stress: questions and answers. World Health Organization. 2023. URL: https://www.who.int/news-room/questions-and-answers/item/stress#:~:text=What%20is%20stress%3F [accessed 2025-11-28]
2. Mahmud S, Mohsin M, Dewan MN, Muyeed A. The global prevalence of depression, anxiety, stress, and insomnia among general population during COVID-19 pandemic: a systematic review and meta-analysis. Trends Psychol 2023;31(1):143-170. [doi: 10.1007/s43076-021-00116-9] [Medline: 40477944]
3. Salari N, Khazaie H, Hosseinian-Far A, et al. The prevalence of stress, anxiety and depression within front-line healthcare workers caring for COVID-19 patients: a systematic review and meta-regression. Hum Resour Health 2020 Dec 17;18(1):100. [doi: 10.1186/s12960-020-00544-1] [Medline: 33334335]
4. Canaletti EF, Lun P, Stutzman LD, Chan M, Cheung F. Rising tide of stress: global trends and structural predictors over 18 years. Wellbeing, Space and Society 2026 Jun;10:100319. [doi: 10.1016/j.wss.2025.100319]
5. Jalali A, Ziapour A, Karimi Z, et al. Global prevalence of depression, anxiety, and stress in the elderly population: a systematic review and meta-analysis. BMC Geriatr 2024 Oct 4;24(1):809. [doi: 10.1186/s12877-024-05311-8] [Medline: 39367305]
6. Piao X, Xie J, Managi S. Continuous worsening of population emotional stress globally: universality and variations. BMC Public Health 2024 Dec 23;24(1):3576. [doi: 10.1186/s12889-024-20961-4] [Medline: 39716139]
7. Smith MD, Wesselbaum D. Global evidence on the prevalence of and risk factors associated with stress. J Affect Disord 2025 Apr 1;374:179-183. [doi: 10.1016/j.jad.2025.01.053] [Medline: 39805499]
8. Cham CQ, Ibrahim N, Siau CS, et al. Caregiver burden among caregivers of patients with mental illness: a systematic review and meta-analysis. Healthcare (Basel) 2022 Nov 30;10(12):2423. [doi: 10.3390/healthcare10122423] [Medline: 36553947]
9. Soh XC, Hartanto A, Ling N, Reyes M, Sim L, Majeed NM. Prevalence of depression, anxiety, burden, burnout, and stress in informal caregivers: an umbrella review of meta-analyses. Archives of Gerontology and Geriatrics Plus 2025 Sep;2(3):100197. [doi: 10.1016/j.aggp.2025.100197]
10. Amanvermez Y, Zhao R, Cuijpers P, et al. Effects of self-guided stress management interventions in college students: a systematic review and meta-analysis. Internet Interv 2022 Apr;28:100503. [doi: 10.1016/j.invent.2022.100503] [Medline: 35242591]
11. Bruffaerts R, Mortier P, Auerbach RP, et al. Lifetime and 12-month treatment for mental disorders and suicidal thoughts and behaviors among first year college students. Int J Methods Psychiatr Res 2019 Jun;28(2):e1764. [doi: 10.1002/mpr.1764] [Medline: 30663193]
12. Arahanthabailu P, Praharaj SK, Purohith AN, Yesodharan R, Rege S, Appaji R. Exploring barriers to seek mental health services among patients with severe mental illness and their caregivers in a modified assertive community treatment program: a qualitative thematic analysis. Indian J Psychiatry 2024 Jul;66(7):621-629. [doi: 10.4103/indianjpsychiatry.indianjpsychiatry_314_24] [Medline: 39257514]
13. Ebert DD, Mortier P, Kaehlke F, et al. Barriers of mental health treatment utilization among first-year college students: first cross-national results from the WHO World Mental Health International College Student Initiative. Int J Methods Psychiatr Res 2019 Jun;28(2):e1782. [doi: 10.1002/mpr.1782] [Medline: 31069905]
14. Keast R, Lengyell M, Finnigan C, Brown J, Jay M. Barriers to mental health care for low-income clients as perceived by counsellors. Journal for Social Action in Counseling & Psychology 2025 Aug 6;17(1):2-21. [doi: 10.33043/r23873x5]

XSL•FO
RenderX

15. Cuijpers P, Schuurmans J. Self-help interventions for anxiety disorders: an overview. Curr Psychiatry Rep 2007 Aug;9(4):284-290. [doi: 10.1007/s11920-007-0034-6] [Medline: 17880859]
16. Chen X, Chen L, Song D, Wang Y, Ryoo J. Digital mental health for young people: sustainably managing stress and anxiety. Mental Health and Digital Technologies 2025 Oct 29;2(3):157-182. [doi: 10.1108/MHDT-01-2025-0009]
17. Gaffney H, Mansell W, Tai S. Conversational agents in the treatment of mental health problems: mixed-method systematic review. JMIR Ment Health 2019 Oct 18;6(10):e14166. [doi: 10.2196/14166] [Medline: 31628789]
18. Ghorbian M, Ghobaei-Arani M. Large language models for mental health diagnosis and treatment: a survey. Artif Intell Rev 2025 Nov 14;59(1):9. [doi: 10.1007/s10462-025-11418-0]
19. Köckritz J, İlgen B, Cohrdes C, Hattab G. Current applications and future directions in natural language processing for news media and mental health. Sci Rep 2025 Sep 15;15(1):32532. [doi: 10.1038/s41598-025-18413-z] [Medline: 40954152]
20. Raza M, Jahangir Z, Riaz MB, Saeed MJ, Sattar MA. Industrial applications of large language models. Sci Rep 2025 Apr 21;15(1):13755. [doi: 10.1038/s41598-025-98483-1] [Medline: 40258923]
21. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
22. Obodo OR. Exploring the role of AI-driven decision support systems in reducing diagnostic errors in rural healthcare settings. International Journal of Scientific Research and Modern Technology 2025 Sep 20:131-135. [doi: 10.38124/ijsrmt.v4i9.816]
23. Pailaha AD. The impact and issues of artificial intelligence in nursing science and healthcare settings. SAGE Open Nurs 2023;9:23779608231196847. [doi: 10.1177/23779608231196847] [Medline: 37691725]
24. Freund J, Smit F, Lehr D, et al. A universal digital stress management intervention for employees: randomized controlled trial with health-economic evaluation. J Med Internet Res 2024 Oct 22;26:e48481. [doi: 10.2196/48481] [Medline: 39437382]
25. Peeters SB, Hilgersom M, van Krugten FCW, et al. Economic evaluations of eHealth interventions targeting mental health problems in the workplace: a systematic review. J Ment Health 2025 Oct;34(5):580-595. [doi: 10.1080/09638237.2024.2390371] [Medline: 39163353]
26. Ali M, Ali S, Abbas Q, Abbas Z, Lee SW. Artificial intelligence for mental health: a narrative review of applications, challenges, and future directions in digital health. Digit Health 2025;11:20552076251395548. [doi: 10.1177/20552076251395548] [Medline: 41262770]
27. Feng Y, Hang Y, Wu W, et al. Effectiveness of AI-driven conversational agents in improving mental health among young people: systematic review and meta-analysis. J Med Internet Res 2025 May 14;27:e69639. [doi: 10.2196/69639] [Medline: 40367506]
28. Jain G, Pareek S, Carlbring P. Revealing the source: how awareness alters perceptions of AI and human-generated mental health responses. Internet Interv 2024 Jun;36:100745. [doi: 10.1016/j.invent.2024.100745] [Medline: 38746873]
29. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. Front Digit Health 2022;4:847991. [doi: 10.3389/fdgth.2022.847991] [Medline: 35480848]
30. Mauriello ML, Lincoln T, Hon G, Simon D, Jurafsky D, Paredes P. SAD: a stress annotated dataset for recognizing everyday stressors in SMS-like conversational systems. Presented at: CHI EA '21: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems; May 8-13, 2021. [doi: 10.1145/3411763.3451799]
31. Croes EAJ, Antheunis ML, van der Lee C, de Wit JMS. Digital confessions: the willingness to disclose intimate information to a chatbot and its impact on emotional well-being. Interact Comput 2024 Aug 15;36(5):279-292. [doi: 10.1093/iwc/iwae016]
32. Lucas GM, Gratch J, King A, Morency LP. It's only a computer: virtual humans increase willingness to disclose. Comput Human Behav 2014 Aug;37:94-100. [doi: 10.1016/j.chb.2014.04.043]
33. Merwin ER, Hagen AC, Keebler JR, Forbes C. Self-disclosure to AI: people provide personal information to AI and humans equivalently. Computers in Human Behavior: Artificial Humans 2025 Aug;5:100180. [doi: 10.1016/j.chbah.2025.100180]
34. Hua Y, Na H, Li Z, et al. A scoping review of large language models for generative tasks in mental health care. NPJ Digit Med 2025 Apr 30;8(1):230. [doi: 10.1038/s41746-025-01611-4] [Medline: 40307331]
35. Olawade DB, Wada OZ, Odetayo A, David-Olawade AC, Asaolu F, Eberhardt J. Enhancing mental health with artificial intelligence: current trends and future prospects. Journal of Medicine, Surgery, and Public Health 2024 Aug;3:100099. [doi: 10.1016/j.glmedi.2024.100099]
36. Pavuluri S, Sangal R, Sather J, Taylor RA. Balancing act: the complex role of artificial intelligence in addressing burnout and healthcare workforce dynamics. BMJ Health Care Inform 2024 Aug 24;31(1):e101120. [doi: 10.1136/bmjhci-2024-101120] [Medline: 39181545]
37. Crosswell AD, Lockwood KG. Best practices for stress measurement: how to measure psychological stress in health research. Health Psychol Open 2020;7(2):2055102920933072. [doi: 10.1177/2055102920933072] [Medline: 32704379]
38. Neri L, Oberdier MT, van Abeelen KCJ, et al. Electrocardiogram monitoring wearable devices and artificial-intelligence-enabled diagnostic capabilities: a review. Sensors (Basel) 2023 May 16;23(10):4805. [doi: 10.3390/s23104805] [Medline: 37430719]
39. Smets E, Rios Velazquez E, Schiavone G, et al. Large-scale wearable data reveal digital phenotypes for daily-life stress detection. NPJ Digit Med 2018;1(1):67. [doi: 10.1038/s41746-018-0074-9] [Medline: 31304344]

XSL•FO
RenderX

40. Castaldo R, Melillo P, Bracale U, Caserta M, Triassi M, Pecchia L. Acute mental stress assessment via short term HRV analysis in healthy adults: a systematic review with meta-analysis. Biomed Signal Process Control 2015 Apr;18:370-377. [doi: 10.1016/j.bspc.2015.02.012]
41. Almadhor A, Sampedro GA, Abisado M, Abbas S. Efficient feature-selection-based stacking model for stress detection based on chest electrodermal activity. Sensors (Basel) 2023 Jul 25;23(15):6664. [doi: 10.3390/s23156664] [Medline: 37571448]
42. Scollon CN, Kim-Prieto C, Diener E. Experience sampling: promises and pitfalls, strengths and weaknesses. J Happiness Stud 2003 Mar;4(1):5-34. [doi: 10.1023/A:1023605205115]
43. Choi A, Ooi A, Lottridge D. Digital phenotyping for stress, anxiety, and mild depression: systematic literature review. JMIR Mhealth Uhealth 2024 May 23;12:e40689. [doi: 10.2196/40689] [Medline: 38780995]
44. Dwyer B, Flathers M, Burns J, et al. Assessing digital phenotyping for app recommendations and sustained engagement: cohort study. JMIR Form Res 2024 Nov 19;8:e62725. [doi: 10.2196/62725] [Medline: 39560976]
45. Loughnane C, Laiti J, O'Donovan R, Dunne PJ. Systematic review exploring human, AI, and hybrid health coaching in digital health interventions: trends, engagement, and lifestyle outcomes. Front Digit Health 2025;7:1536416. [doi: 10.3389/fdgth.2025.1536416] [Medline: 40343215]
46. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. J Med Internet Res 2017 Jun 29;19(6):e232. [doi: 10.2196/jmir.7126] [Medline: 28663162]
47. Zhu Y, Long Y, Wang H, Lee KP, Zhang L, Wang SJ. Digital behavior change intervention designs for habit formation: systematic review. J Med Internet Res 2024 May 24;26:e54375. [doi: 10.2196/54375] [Medline: 38787601]
48. Michie S, Wood CE, Johnston M, Abraham C, Francis JJ, Hardeman W. Behaviour change techniques: the development and evaluation of a taxonomic method for reporting and describing behaviour change interventions (a suite of five studies involving consensus methods, randomised controlled trials and analysis of qualitative data). Health Technol Assess 2015 Nov;19(99):1-188. [doi: 10.3310/hta19990] [Medline: 26616119]
49. Bendig E, Erb B, Schulze-Thuesing L, Baumeister H. The next generation: chatbots in clinical psychology and psychotherapy to foster mental health – a scoping review. Verhaltenstherapie 2022;32(Suppl. 1):64-76. [doi: 10.1159/000501812]
50. Trappey AJC, Lin APC, Hsu KYK, Trappey CV, Tu KLK. Development of an empathy-centric counseling chatbot system capable of sentimental dialogue analysis. Processes (Basel) 2022 May 8;10(5):930. [doi: 10.3390/pr10050930]
51. Kraus M, Wagner N, Callejas Z, Minker W. The role of trust in proactive conversational assistants. IEEE Access 2021;9:112821-112836. [doi: 10.1109/ACCESS.2021.3103893]
52. Sanjeewa R, Iyer R, Apputhurai P, Wickramasinghe N, Meyer D. Empathic conversational agent platform designs and their evaluation in the context of mental health: systematic review. JMIR Ment Health 2024 Sep 9;11:e58974. [doi: 10.2196/58974] [Medline: 39250799]
53. Ho JQH, Hartanto A, Koh A, Majeed NM. Gender biases within artificial intelligence and ChatGPT: evidence, sources of biases and solutions. Computers in Human Behavior: Artificial Humans 2025 May;4:100145. [doi: 10.1016/j.chbah.2025.100145]
54. Alkaeed M, Qayyum A, Qadir J. Privacy preservation in artificial intelligence and extended reality (AI-XR) metaverses: a survey. Journal of Network and Computer Applications 2024 Nov;231:103989. [doi: 10.1016/j.jnca.2024.103989]
55. Mandal A, Paradkar M, Panindre P, Kumar S. AI-based detection of stress using heart rate data obtained from wearable devices. Procedia Comput Sci 2023;230:749-759. [doi: 10.1016/j.procs.2023.12.129]
56. Rezaeikhonakdar D. AI chatbots and challenges of HIPAA compliance for AI developers and vendors. J Law Med Ethics 2023;51(4):988-995. [doi: 10.1017/jme.2024.15] [Medline: 38477276]
57. Tavory T. Regulating AI in mental health: ethics of care perspective. JMIR Ment Health 2024 Sep 19;11:e58493. [doi: 10.2196/58493] [Medline: 39298759]
58. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018 Nov 1;178(11):1544-1547. [doi: 10.1001/jamainternmed.2018.3763] [Medline: 30128552]
59. Horsfall LJ, Bondaronek P, Ive J, Poduval S. Clinical algorithms and the legacy of race-based correction: historical errors, contemporary revisions and equity-oriented methodologies for epidemiologists. Clin Epidemiol 2025;17:647-662. [doi: 10.2147/CLEP.S527000] [Medline: 40673267]
60. Joseph J. Algorithmic bias in public health AI: a silent threat to equity in low-resource settings. Front Public Health 2025;13:1643180. [doi: 10.3389/fpubh.2025.1643180] [Medline: 40771228]
61. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
62. Corporation for digital scholarship. Zotero. 2025. URL: https://www.zotero.org/
63. Stern C, Lizarondo L, Carrier J, et al. Methodological guidance for the conduct of mixed methods systematic reviews. JBI Evid Synth 2020 Oct;18(10):2108-2118. [doi: 10.11124/JBISRIR-D-19-00169] [Medline: 32813460]

XSL•FO
RenderX

64. Rodgers M, Sowden A, Petticrew M, et al. Testing methodological guidance on the conduct of narrative synthesis in systematic reviews: effectiveness of interventions to promote smoke alarm ownership and function. Evaluation (Lond) 2009 Jan;15(1):49-73. [doi: 10.1177/1356389008097871]
65. Popay J, Roberts H, Sowden A, et al. Guidance on the Conduct of Narrative Synthesis in Systematic Reviews: A Product from the ESRC Methods Programme: Lancaster University; 2006. [doi: 10.13140/2.1.1018.4643]
66. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008 Jul 10;8(1):45. [doi: 10.1186/1471-2288-8-45] [Medline: 18616818]
67. Aromataris E, Fernandez R, Godfrey CM, Holly C, Khalil H, Tungpunkom P. Summarizing systematic reviews: methodological development, conduct and reporting of an umbrella review approach. Int J Evid Based Healthc 2015 Sep;13(3):132-140. [doi: 10.1097/XEB.0000000000000055] [Medline: 26360830]
68. Alanazi R, Alanazi AS, Alqazlan S, Benlaria H. Assessing the impact of AI tools on mobility and daily assistance for children with down syndrome in Saudi Arabia. Sci Rep 2025 Aug;15(1). [doi: 10.1038/s41598-025-15085-7]
69. Alanezi F. Assessing the effectiveness of ChatGPT in delivering mental health support: a qualitative study. J Multidiscip Healthc 2024;17:461-471. [doi: 10.2147/JMDH.S447368] [Medline: 38314011]
70. Allan DD. A mobile messaging-based conversational agent-led stress mindset intervention for New Zealand small-to-medium-sized enterprise owner-managers: effectiveness and acceptability study. Behaviour & Information Technology 2024 Sep 9;43(12):2785-2798. [doi: 10.1080/0144929X.2023.2259500]
71. Crum AJ, Akinola M, Martin A, Fath S. The role of stress mindset in shaping cognitive, emotional, and physiological responses to challenging and threatening stress. Anxiety Stress Coping 2017 Jul;30(4):379-395. [doi: 10.1080/10615806.2016.1275585] [Medline: 28120622]
72. Baek G, Cha C. AI-assisted tailored intervention for nurse burnout: a three-group randomized controlled trial. Worldviews Evid Based Nurs 2025 Feb;22(1):e70003. [doi: 10.1111/wvn.70003] [Medline: 39981583]
73. Cho A, Cha C, Baek G. Development of an artificial intelligence-based tailored mobile intervention for nurse burnout: single-arm trial. J Med Internet Res 2024 Jun 21;26:e54029. [doi: 10.2196/54029] [Medline: 38905631]
74. Fuhrmann LM, Lukas CA, Schindler-Gmelch L, Berking M. Evaluating a brief smartphone-based stress management intervention with heart rate biofeedback from built-in sensors in a three arm randomized controlled trial. Sci Rep 2025 Jun 23;15(1):20257. [doi: 10.1038/s41598-025-06588-4] [Medline: 40550871]
75. Gabrielli S, Rizzi S, Bassi G, et al. Engagement and effectiveness of a healthy-coping intervention via chatbot for university students during the COVID-19 pandemic: mixed methods proof-of-concept study. JMIR Mhealth Uhealth 2021 May 28;9(5):e27965. [doi: 10.2196/27965] [Medline: 33950849]
76. Hiller S, Götzl C, Rauschenberg C, et al. Health-promoting effects and everyday experiences with a mental health app using ecological momentary assessments and AI-based ecological momentary interventions among young people: qualitative interview and focus group study. JMIR Mhealth Uhealth 2025 Apr 29;13:e65106. [doi: 10.2196/65106] [Medline: 40300160]
77. Huang YC, Gamborino E, Huang YJ, Qian X, Fu LC, Yeh SL. Inferring stressors from conversation: towards an emotional support robot companion. Int J of Soc Robotics 2022 Sep;14(7):1657-1671. [doi: 10.1007/s12369-022-00902-0]
78. Indrayanti I, Salsabila AK, Amrita V, Alhaddad MM, Saskia AB, Ramadhani DP. PsyBot: a randomized controlled trial of WhatsApp-based psychological first aid to reduce loneliness among 18–22-year-old students in Yogyakarta, Indonesia. SSM - Mental Health 2025 Dec;8:100504. [doi: 10.1016/j.ssmmh.2025.100504]
79. Jiang J, Yang Y. GymBuddy and Elomia, AI-integrated applications, effects on the mental health of the students with psychological disorders. BMC Psychol 2025 Aug;13(1):350. [doi: 10.1186/s40359-025-02640-0]
80. Keung WM, So TY. Attitudes towards AI counseling: the existence of perceptual fear in affecting perceived chatbot support quality. Front Psychol 2025;16:1538387. [doi: 10.3389/fpsyg.2025.1538387] [Medline: 40823408]
81. Ko S, Woo H. Users' needs for mental health apps: quality evaluation using the user version of the mobile application rating scale. JMIR Mhealth Uhealth 2025 Jul 4;13:e64622. [doi: 10.2196/64622] [Medline: 40614100]
82. Lai HL, Chen CI, Liaw GH, Lu LC, Huang CY. Effectiveness of a chatbot-based internet-support-coping-caring intervention in improving sleep quality and work-related outcomes among nurses in taiwan during the COVID-19 pandemic: a pilot study. J Nurs Res 2025 Oct 1;33(5):e411. [doi: 10.1097/jnr.0000000000000691] [Medline: 40748237]
83. Lee I, Hahn S. On the relationship between mind perception and social support of chatbots. Front Psychol 2024;15:1282036. [doi: 10.3389/fpsyg.2024.1282036] [Medline: 38510306]
84. Lee S, Jeong J, Kim M, Lee S, Kim SP, Jung D. Development of a mobile intervention for procrastination augmented with a semigenerative chatbot for university students: pilot randomized controlled trial. JMIR Mhealth Uhealth 2025 Apr 10;13:e53133. [doi: 10.2196/53133] [Medline: 40208664]
85. Li Y, Chung TY, Lu W, et al. Chatbot-based mindfulness-based stress reduction program for university students with depressive symptoms: intervention development and pilot evaluation. J Am Psychiatr Nurses Assoc 2025;31(4):398-411. [doi: 10.1177/10783903241302092] [Medline: 39704086]
86. Ly KH, Ly AM, Andersson G. A fully automated conversational agent for promoting mental well-being: a pilot RCT using mixed methods. Internet Interv 2017 Dec;10:39-46. [doi: 10.1016/j.invent.2017.10.002] [Medline: 30135751]

XSL•FO
RenderX

87. Marwaha EB, Gera J, Thareja R, Gupta S. The digital bridge: using digital intervention to understand psychosocial correlation between COVID-19 stress and sleep quality in undergraduate students before and after COVID-19. TOPSYJ 2025 Feb 28;18(1):e18743501364965. [doi: 10.2174/0118743501364965250220093932]
88. Medeiros L, Bosse T, Gerritsen C. Can a chatbot comfort humans? Studying the impact of a supportive chatbot on users' self-perceived stress. IEEE Trans Human-Mach Syst 2022 Jun;52(3):343-353. [doi: 10.1109/THMS.2021.3113643]
89. Meng J, Dai Y(. Emotional support from AI chatbots: should a supportive partner self-disclose or not? J Comput-Mediat Commun 2021 Sep 9;26(4):207-222. [doi: 10.1093/jcmc/zmab005]
90. Nelekar S, Abdulrahman A, Gupta M, Richards D. Effectiveness of embodied conversational agents for managing academic stress at an Indian University (ARU) during COVID‐19. Brit J Educational Tech 2022 May;53(3):491-511. [doi: 10.1111/bjet.13174]
91. Orji J, Chan G, Orji R. Design, development, and evaluation of an mHealth app to reduce stress and promote happiness through smiling. Behaviour & Information Technology 2026 May 9;45(8):1546-1566. [doi: 10.1080/0144929X.2025.2494278]
92. Park M, Fels S, Seo K. Enhancing academic stress assessment through self-disclosure chatbots: effects on engagement, accuracy, and self-reflection. Int J Educ Technol High Educ 2025;22(1). [doi: 10.1186/s41239-025-00527-z]
93. Park S, Choi J, Lee S, et al. Designing a chatbot for a brief motivational interview on stress management: qualitative case study. J Med Internet Res 2019 Apr 16;21(4):e12231. [doi: 10.2196/12231] [Medline: 30990463]
94. Qi Y. Pilot quasi-experimental research on the effectiveness of the Woebot AI chatbot for reducing mild depression symptoms among athletes. International Journal of Human–Computer Interaction 2025 Jan 2;41(1):452-459. [doi: 10.1080/10447318.2023.2301256]
95. Silva E, Aguiar J, Reis LP, Sá JOE, Gonçalves J, Carvalho V. Stress among Portuguese medical students: the EuStress Solution. J Med Syst 2020 Jan 2;44(2):45. [doi: 10.1007/s10916-019-1520-1] [Medline: 31897774]
96. Sun X, Casula D, Navaratnam A, et al. In: Lukowicz P, Mayer S, Koch J, Shawe-Taylor J, Tiddi I, editors. Virtual Support for Real-World Movement: Using Chatbots to Overcome Barriers to Physical Activity: Front Artif Intell Appl IOS Press; 2023. [doi: 10.3233/FAIA230084]
97. Teague SJ, Shatte ABR, Fuller-Tyszkiewicz M, Hutchinson DM. User-centred design and evaluation of an mHealth app for fathers' perinatal mental health: a feasibility, acceptability, and usability study. Behav Inf Technol 2025 May 9;44(20):1-15. [doi: 10.1080/0144929X.2025.2502474] [Medline: 40385514]
98. Tong ACY, Wong KTY, Chung WWT, Mak WWS. Effectiveness of topic-based chatbots on mental health self-care and mental well-being: randomized controlled trial. J Med Internet Res 2025 Apr 30;27:e70436. [doi: 10.2196/70436] [Medline: 40306635]
99. Troitskaya O, Batkhina A. Mobile application for couple relationships: results of a pilot effectiveness study. Fam Process 2022 Jun;61(2):625-642. [doi: 10.1111/famp.12733] [Medline: 34904235]
100. Weng JH, Hu Y, Heaukulani C, et al. Mental wellness self-care in Singapore with mindline.sg: a tutorial on the development of a digital mental health platform for behavior change. J Med Internet Res 2024 Jun 4;26:e44443. [doi: 10.2196/44443] [Medline: 38833294]
101. Williams R, Hopkins S, Frampton C, Holt-Quick C, Merry SN, Stasiak K. 21-Day stress detox: open trial of a universal well-being chatbot for young adults. Soc Sci (Basel) 2021;10(11):416. [doi: 10.3390/socsci10110416]
102. Yong SC. Integrating emotional AI into mobile apps with smart home systems for personalized mental wellness. J Technol Behav Sci 2025;10(4):761-778. [doi: 10.1007/s41347-025-00508-z]
103. Yoon S, Goh H, Low XC, Weng JH, Heaukulani C. User perceptions and utilisation of features of an AI-enabled workplace digital mental wellness platform 'mindline at work'. BMJ Health Care Inform 2024 Aug 17;31(1):e101045. [doi: 10.1136/bmjhci-2024-101045] [Medline: 39153756]
104. Hong QN, Fàbregues S, Bartlett G, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. EFI 2018 Dec 18;34(4):285-291. [doi: 10.3233/EFI-180221]
105. Blaß M, Gimpel H, Karnebogen P. A taxonomy and archetypes of AI-Based Health Care Services: Qualitative Study. J Med Internet Res 2024 Nov 27;26:e53986. [doi: 10.2196/53986] [Medline: 39602787]
106. Choong YY, Jensen T. AI use taxonomy: a human-centered approach. : National Institute of Standards and Technology; 2024. [doi: 10.6028/NIST.AI.200-1]
107. Uuk R, Gutierrez CI, Guppy D, et al. A taxonomy of systemic risks from general-purpose AI. SSRN. Preprint posted online on 2025. [doi: 10.2139/ssrn.5030173]
108. Slattery P, Saeri AK, Grundy EAC, et al. The AI risk repository: a meta-review, database, and taxonomy of risks from artificial intelligence. Patterns 2026 May;7(5):101517. [doi: 10.1016/j.patter.2026.101517]
109. Rahman M, Victoros E, Davis R, Duaa T, Shanjana Y, Islam M. Use of artificial intelligence in mental healthcare, health psychology, and related research: a narrative review to address challenges and opportunities. Health Sci Rep 2025 Dec;8(12):e71595. [doi: 10.1002/hsr2.71595] [Medline: 41332917]
110. Alekseev A, Strobel C. A taxonomy of AI experiments. Journal of Behavioral and Experimental Economics 2026 Mar;121:102525. [doi: 10.1016/j.socec.2026.102525]

XSL•FO
RenderX

111. Manley AE, Biddle L, Savović J, Moran P. The positive and negative consequences of stress and its relationship with coping in medical students: a qualitative study. Med Teach 2025 Feb;47(2):323-328. [doi: 10.1080/0142159X.2024.2333799] [Medline: 38593839]
112. Faiz F, Le V, Masli EK. Determinants of digital technology adoption in innovative SMEs. Journal of Innovation & Knowledge 2024 Oct;9(4):100610. [doi: 10.1016/j.jik.2024.100610]
113. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 2020 Sep;26(9):1364-1374. [doi: 10.1038/s41591-020-1034-x] [Medline: 32908283]
114. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. BMJ 2020 Sep 9;370:m3210. [doi: 10.1136/bmj.m3210] [Medline: 32907797]
115. Artificial intelligence. Organisation for Economic Co-operation and Development (OECD). 2026. URL: https://www.oecd.org/en/topics/policy-issues/artificial-intelligence.html [accessed 2026-05-02]
116. Ethics guidelines for trustworthy AI. European Commission. 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai [accessed 2026-05-02]
117. ASEAN guide on AI governance and ethics. ASEAN Secretariat. 2024. URL: https://asean.org/book/asean-guide-on-ai-governance-and-ethics/ [accessed 2026-05-02]
118. Chen P, Wu L, Wang L. AI fairness in data management and analytics: a review on challenges, methodologies and applications. Appl Sci (Basel) 2023 Sep 13;13(18):10258. [doi: 10.3390/app131810258]
119. Timmons AC, Duong JB, Simo Fiallo N, et al. A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. Perspect Psychol Sci 2023 Sep;18(5):1062-1096. [doi: 10.1177/17456916221134490] [Medline: 36490369]

Abbreviations

AI: artificial intelligence
CBT: cognitive behavioral therapy
CONSORT-AI: Consolidated Standards of Reporting Trials - Artificial Intelligence
GAIDeT: Generative Artificial Intelligence Delegation Taxonomy
GenAI: generative artificial intelligence
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
MMAT: Mixed Methods Appraisal Tool
NLP: natural language processing
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence
WHO: World Health Organization

Edited by M Balcarras; submitted 02.Jan.2026; peer-reviewed by MZ Hossain, O Ndibe, V Gupta, Z Liu; revised version received 05.May.2026; accepted 05.May.2026; published 24.Jun.2026.

Please cite as:
Reyes MKG, Teo SSM, Hartanto A
The Emerging Roles of AI in Self-Directed Stress Management: Systematic Review
J Med Internet Res 2026;28:e90709
URL: https://www.jmir.org/2026/1/e90709
doi:10.2196/90709

© Mary Kamillah Grace Reyes, Shauna Sha Min Teo, Andree Hartanto. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Mediating Role of Interpersonal Needs in Perceived Parenting Styles and Social Media Addiction Among University Students: Cross-Sectional Study

Yujie Liu[1,2*], PhD; Suping Wang[3*], PhD; Ting Wei[1*], MS; Birong Wu[2], MS; Xin Ge[2], MS; Shangbin Liu[2], MS; Chen Xu[2], MS; Shunyu Tao[2], MS; Xiaohong Fan[4], PhD; Fan Hu[2], PhD; Ying Wang[2], PhD; Xue Yang[5], PhD; Yong Cai[2,6], PhD

[1]Tongji Hospital, School of Medicine, Tongji University, Shanghai, China

[2]Institute of Clinical Research, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[3]Shanghai Jiao Tong University School of Medicine, Shanghai, China

[4]Shanghai Public Health Clinical Center, Shanghai, China

[5]JC School of Public Health and Primary Care, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong

[6]Center for Community Health Care, Shanghai Jiao Tong University China Hospital Development Institute, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Yong Cai, PhD
Institute of Clinical Research, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Abstract

Background: The rapid development of digital technologies has intensified concerns about social media addiction, particularly among university students. The perceived parenting styles they experienced may influence this risk, with unmet interpersonal needs acting as potential mediators.

Objective: This study aimed to investigate whether perceived burdensomeness and thwarted belongingness mediate the relationship between perceived parenting and social media addiction among university students.

Methods: A cross-sectional survey was conducted among 1766 university students from March 2023 to May 2023. Parenting styles were assessed using the short-form Egna Minnen av Barndoms Uppfostran (Swedish for “My memories of upbringing”) for Chinese, unmet interpersonal needs were assessed via the 15-item version of the Interpersonal Needs Questionnaire, and social media addiction was assessed through the Social Network Addiction Tendency Scale. Parenting profiles were identified using latent profile analysis, and mediation models were examined through path analysis.

Results: Three perceived parenting profiles were identified: supportive (1094/1766, 61.9%), emotionally distant (192/1766, 10.9%), and controlling-critical (480/1766, 27.2%). Path analyses indicated indirect effects of parenting profile on social media addiction via unmet interpersonal needs. Compared to the supportive group, the emotionally distant (β=0.107 for perceived burdensomeness; β=0.354 for thwarted belongingness) and controlling-critical (β=0.543 for perceived burdensomeness; β=0.457 for thwarted belongingness) groups reported higher unmet interpersonal needs, which in turn were positively associated with social media addiction (β=0.119 for perceived burdensomeness; β=0.158 for thwarted belongingness). Multigroup analyses further showed that perceived burdensomeness was significantly associated with social media addiction among male students (β=0.180; $P<.001$) but not among female students (β=0.038; $P=.30$).

Conclusions: Unmet interpersonal needs mediate the link between maladaptive parenting and social media addiction. Early interventions that promote emotional warmth and reduce rejecting and controlling behaviors may help prevent digital dependence among young adults, particularly when tailored to specific interpersonal vulnerabilities.

(J Med Internet Res 2026;28:e91861) doi:10.2196/91861

KEYWORDS

social media addiction; parenting styles; university students; unmet interpersonal needs; psychological mediation; latent profile analysis

XSL•FO
RenderX

Introduction

The rapid proliferation of digital technologies has reshaped human interaction, offering unprecedented opportunities for communication. In China, this shift has been accompanied by the widespread use of locally dominant social media platforms such as WeChat and Weibo, particularly among university students. As these technologies have become increasingly embedded in everyday life, concerns about digital dependence have also grown. Social media addiction is a specific manifestation of digital dependence characterized by impaired control over social media use, obsessive preoccupation with online activities, difficulty resisting the urge to engage, and a preference for online over face-to-face interactions [1]. A meta-analysis showed that the pooled prevalence of social media addiction reached 24% across 32 countries, with higher rates observed among young people [2]. University students are particularly vulnerable due to unlimited internet access, lack of parental supervision, flexible schedules, and developmental demands for social interaction [3]. In this population, social media addiction has been linked to adverse consequences, including poor sleep quality, reduced academic performance, and increased risks of mental health problems [3-5]. These concerns underscore the importance of clarifying the pathways underlying the development of social media addiction.

Although numerous studies have identified the roles of proximal individual factors such as attachment styles and personality traits in social media addiction [6,7], its development may also be shaped by broader environmental influences. Early family experiences are particularly crucial in shaping psychological vulnerabilities related to social media addiction [8]. Against the backdrop of rapid digital change, traditional family expectations and relational norms may remain influential [9], whereas young adults are increasingly embedded in highly digitalized social environments. In this context, the relationship between parenting and online engagement warrants particular attention. Previous research among Chinese university students has shown that poor family functioning is associated with higher levels of social networking addiction [10]. Maladaptive parenting practices such as overprotection and psychological control have been positively associated with problematic social media use [11,12], whereas positive parenting may attenuate the association between negative life events and social media addiction [13]. However, the mechanisms through which parenting style influences social media addiction remain poorly understood.

According to the social compensation hypothesis, individuals facing difficulties in offline social interactions may turn to online platforms as a compensatory strategy [14]. Unsatisfied interpersonal needs and social deficits can increase the propensity for excessive use, which may develop into an addictive pattern sustained by negative reinforcement [15]. Thwarted belongingness and perceived burdensomeness represent 2 fundamental unmet interpersonal needs, which are conceptualized within the framework of the interpersonal theory of suicide as critical factors contributing to adverse psychological states [16]. Thwarted belongingness reflects the perception that one's need for social connection is unfulfilled, whereas perceived burdensomeness represents the belief that one is a burden to others. Both conditions have been linked to a wide range of adverse health outcomes, including a higher risk of suicide [17-19]. In collectivist cultural contexts such as China, these interpersonal needs may be particularly relevant given the strong emphasis on relational interdependence and social harmony [20]. There is evidence from Chinese samples further suggesting that individuals experiencing these unmet interpersonal needs may engage in excessive social media use to seek social connection [21], supporting the compensatory pathway toward social media addiction. Moreover, parenting practices such as low emotional warmth and high verbal hostility have been shown to predict the development of unmet interpersonal needs in offspring [22]. These findings highlight thwarted belongingness and perceived burdensomeness as potential pathways linking parenting styles to social media addiction.

Despite the theoretical plausibility of these pathways, empirical evidence examining the mediating effects of thwarted belongingness and perceived burdensomeness remains limited. To address this gap, this study investigates whether these 2 dimensions of unmet interpersonal needs mediate the relationship between perceived parenting styles and social media addiction among Chinese university students. Clarifying these mechanisms can advance understanding of the psychological processes underlying social media addiction and inform the development of targeted, evidence-based interventions for digitally vulnerable emerging adults.

Methods

Participants and Procedure

A cross-sectional survey was conducted between March 2023 and May 2023 in Shanghai, China. Participants were recruited using a multistage sampling method. In the first stage, universities were selected based on type, including 2 medical universities, 2 liberal arts universities, and 1 science and engineering university. In the subsequent stages, participants were further stratified by academic year and major for sampling purposes.

Electronic questionnaires were administered via the Wenjuanxing platform, a professional online survey system commonly used in academic research in China. The questionnaire was designed with mandatory responses for all items and embedded logical validation rules to ensure data quality. Only fully completed questionnaires could be submitted successfully.

Of the 2125 questionnaires distributed, 1768 (83.2%) were returned. Of these 1768 questionnaires, after excluding invalid responses (eg, identical answers across items or completion times under 5 minutes), 1766 (99.9%) valid questionnaires were retained for analysis.

Measurements

Perceived Parenting Style

Perceived parenting style was assessed using the short-form Egna Minnen av Barndoms Uppfostran for Chinese (S-EMBU-C) [23], adapted from the original self-reported Egna

XSL•FO
RenderX

Minnen av Barndoms Uppfostran (Swedish for "My memories of upbringing") [24]. The S-EMBU-C includes 42 items organized into 2 parallel scales for fathers and mothers. Each scale comprises 21 items measuring 3 dimensions: rejection (eg, "My father/mother treat me in such a way that I feel ashamed"), emotional warmth (eg, "My father/mother try to encourage me to become the best"), and overprotection (eg, "My father/mother get overly anxious that something might happen to me"). Each item is rated on a 4-point Likert scale ranging from 1 ("never") to 4 ("always"). The S-EMBU-C has demonstrated good reliability and validity among Chinese university students [23]. In this study, Cronbach α values were 0.949 to 0.953 for rejection, 0.949 to 0.951 for emotional warmth, and 0.778 to 0.800 for overprotection.

Interpersonal Needs

Interpersonal needs were assessed using the 15-item version of the Interpersonal Needs Questionnaire (INQ-15), which is a shortened form of the original scale [25]. The INQ-15 captures 2 dimensions: perceived burdensomeness and thwarted belongingness, which have been validated in prior research as key indicators of suicide-related interpersonal needs [26,27]. Each item is rated on a 7-point Likert scale ranging from 1 ("not true for me at all") to 7 ("very true for me"), with higher total values reflecting greater unmet interpersonal needs. The Chinese version of the INQ-15 has shown satisfactory reliability and validity in Chinese university student samples [28]. In this study, Cronbach α values were 0.974 for perceived burdensomeness and 0.872 for thwarted belongingness.

Social Media Addiction

Social media addiction was assessed using the Social Network Addiction Tendency Scale [29]. The scale includes 8 items, each rated on a 5-point Likert scale from 1 ("strongly disagree") to 5 ("strongly agree"). All items are positively scored (eg, "Using social networks makes it hard for me to focus on my studies"), with higher total scores indicating a greater propensity toward social media addiction. Previous studies have demonstrated that this scale possesses good psychometric properties among Chinese university students [29]. In this study, the scale showed good internal consistency, with a Cronbach α coefficient of 0.910.

Covariates

Covariates included sex (male and female), age, academic year (first year, second year, third year, and fourth year or above), university type (medical, liberal arts, and science and engineering), only child status (yes or no), and parental educational levels (primary school or lower, junior or senior high school, and college or higher).

Statistical Analysis

Descriptive analyses were conducted to summarize participants' demographic characteristics and the distribution of social media addiction. Group differences in social media addiction across demographic groups were examined using 2-tailed independent-sample *t* tests or one-way ANOVA. Pearson correlations were computed to assess the associations among parenting styles, perceived burdensomeness, thwarted belongingness, and social media addiction.

Latent profile analysis (LPA) was performed using the R package *tidyLPA* (R Foundation for Statistical Computing) to identify distinct parenting profiles based on paternal and maternal rejection, emotional warmth, and overprotection. Models with 2 to 5 classes were estimated. Model fit was evaluated using the Akaike information criterion, Bayesian information criterion (BIC), sample size–adjusted BIC, entropy, and bootstrap likelihood ratio test. Class proportions and interpretability were also taken into account.

Hierarchical regression analyses were then conducted to examine the association between parenting profiles and social media addiction. In model 1, only parenting profiles were included. Model 2 further adjusted for demographic covariates (sex, age, university type, academic year, only child status, and parental educational levels). Model 3 additionally incorporated perceived burdensomeness and thwarted belongingness.

Finally, path analysis was used to test the pathways from parenting profiles to social media addiction via perceived burdensomeness and thwarted belongingness. Parenting profiles were dummy coded, with the largest group serving as the reference category. Model fit was assessed using the comparative fit index, Tucker-Lewis index, root mean squared error of approximation, and standardized root mean squared residual. Direct and indirect effects were tested using bootstrapping with 5000 resamples.

Ethical Considerations

All procedures performed involving human participants were in accordance with the 1964 Declaration of Helsinki. The study protocol was approved by the ethics committee of Shanghai Jiao Tong University School of Medicine (SJUPN-201813). Before accessing the survey, participants were required to read and electronically sign an informed consent form explaining the study's purpose, voluntary participation, and confidentiality. Participants received a CN ¥5 (approximately US $0.74) digital incentive after completing the survey.

Results

Descriptive Statistics and Correlations

Table 1 presents the demographic characteristics of the 1766 participants. The sample included 48.5% (n=857) men and 51.5% (n=909) women, with a mean age of 19.9 (SD 1.5) years. Regarding university type, 44.9% (n=793) of the participants were from medical universities, 14.3% (n=252) were from the science and engineering university, and 40.8% (n=721) were from liberal arts universities. Approximately half (n=895, 50.7%) were first-year students. Social media addiction scores did not significantly differ by gender, age, or academic year. However, significant differences were observed across university types (P<.001), with students from liberal arts and science and engineering universities reporting higher scores than those from medical universities. Additional differences in social media addiction were found according to only child status and parental educational levels (P<.001).

XSL•FO
RenderX

Table . Demographic characteristics and social media addiction scores (N=1766).

	Participants, n (%)	Social media addiction score (8-40), mean (SD)	Statistic		*P* value
			t test (*df*)	*F* test (*df*)	
Sex			0.27 (1764)	—[a]	.79
Male	857 (48.5)	21.05 (6.73)			
Female	909 (51.5)	21.13 (5.84)			
Age (y)			—	1.52 (2, 1763)	.22
18-19	839 (47.5)	20.87 (6.10)			
20-21	695 (39.4)	21.17 (6.44)			
22-24	232 (13.1)	21.66 (6.45)			
University type			—	50.91 (2, 1763)	<.001
Medical	793 (44.9)	19.53 (5.69)			
Science and engineering	252 (14.3)	21.39 (6.05)			
Liberal arts	721 (40.8)	22.7 (6.57)			
Academic year			—	0.29 (3, 1762)	.83
First year	895 (50.7)	21.01 (6.23)			
Second year	437 (24.7)	21.29 (6.17)			
Third year	325 (18.4)	21.14 (6.51)			
Fourth year or above	109 (6.2)	20.77 (6.57)			
Only child			3.98 (1764)	—	<.001
Yes	881 (49.9)	21.68 (6.14)			
No	885 (50.1)	20.5 (6.37)			
Father's educational level			—	10.60 (2, 1763)	<.001
Primary school or lower	202 (11.4)	22.59 (6.41)			
Junior or senior high school	761 (43.1)	21.36 (5.94)			
College or higher	803 (45.5)	20.46 (6.49)			
Mother's educational level			—	12.47 (2,1763)	<.001
Primary school or lower	284 (16.1)	22.32 (6.46)			
Middle school	787 (44.6)	21.38 (6.11)			
High school	695 (39.4)	20.26 (6.3)			

[a]Not applicable.

Table S1 in Multimedia Appendix 1 summarizes the descriptive statistics and correlations among parenting styles, perceived burdensomeness, thwarted belongingness, and social media addiction. Significant correlations were found between parenting styles and both perceived burdensomeness and thwarted belongingness (*P*<.001). Social media addiction was positively associated with parental rejection and overprotection but negatively associated with parental emotional warmth (*P*<.001). Furthermore, social media addiction was positively correlated with both perceived burdensomeness and thwarted belongingness (*P*<.001).

Latent Profiles of Parenting Style

LPA was conducted using paternal and maternal rejection, emotional warmth, and overprotection as indicators (Table 2). Entropy values were high for both the 2- and 3-profile models (0.914 - 0.933), indicating clear classification. Although the Akaike information criterion, BIC, and sample size–adjusted BIC values continued to decrease with additional profiles, model selection also considered interpretability. The 4- and 5-profile solutions mainly split existing profiles into smaller intermediate subgroups rather than identifying qualitatively distinct new configurations (Figure S1 in Multimedia Appendix 1). In

XSL•FO
RenderX

contrast, the 3-profile solution identified 3 clear and theoretically meaningful parenting patterns with acceptable class proportions and the smallest class exceeding 10% of the sample. Taken together, the 3-profile model was selected as the optimal solution.

Table . Model fit indexes of latent profile analysis using short-form Egna Minnen av Barndoms Uppfostran for Chinese scores.

Profile model	AIC[a]	BIC[b]	aBIC[c]	Entropy	BLRT[d]	Proportions[e]
1	63,088	63,154	63,115	1	—[f]	1
2	60,507	60,611	60,551	0.933	0.010	0.732/0.268
3	59,434	59,576	59,494	0.914	0.010	0.619/0.109/0.272
4	58,968	59,149	59,044	0.848	0.010	0.370/0.297/0.260/0.072
5	57,756	57,975	57,848	0.884	0.010	0.353/0.298/0.228/0.066/0.061

[a]AIC: Akaike information criterion.

[b]BIC: Bayesian information criterion.

[c]aBIC: sample size–adjusted BIC.

[d]BLRT: bootstrap likelihood ratio test.

[e]Proportions indicate the proportion of each profile in the model.

[f]Not applicable.

Figure 1 shows the mean S-EMBU-C scores across the 3 latent parenting profiles. Group 1, labeled the supportive parenting group (1094/1766, 61.9%), represented the majority and was characterized by high parental emotional warmth and moderate levels of rejection and overprotection. Group 2, the emotionally distant parenting group (192/1766, 10.9%), was marked by lower parental emotional warmth. Group 3, the controlling-critical parenting group (480/1766, 27.2%), showed elevated parental rejection and overprotection. Significant differences were observed across all parenting profiles ($P<.001$). Parental rejection and overprotection were highest in the controlling-critical parenting group compared to the other 2 groups ($P<.001$), whereas parental emotional warmth was highest in the supportive parenting group ($P<.001$; Table S2 in Multimedia Appendix 1).

Figure 1. Latent profiles of parenting styles: mean short-form Egna Minnen av Barndoms Uppfostran for Chinese scores across classes. Group 1: supportive parenting (1094/1766, 61.9%); group 2: emotionally distant parenting (192/1766, 10.9%); group 3: controlling-critical parenting (480/1766, 27.2%); MO: maternal overprotection; MR: maternal rejection; MW: maternal emotional warmth; PO: paternal overprotection; PR: paternal rejection; PW: paternal emotional warmth.

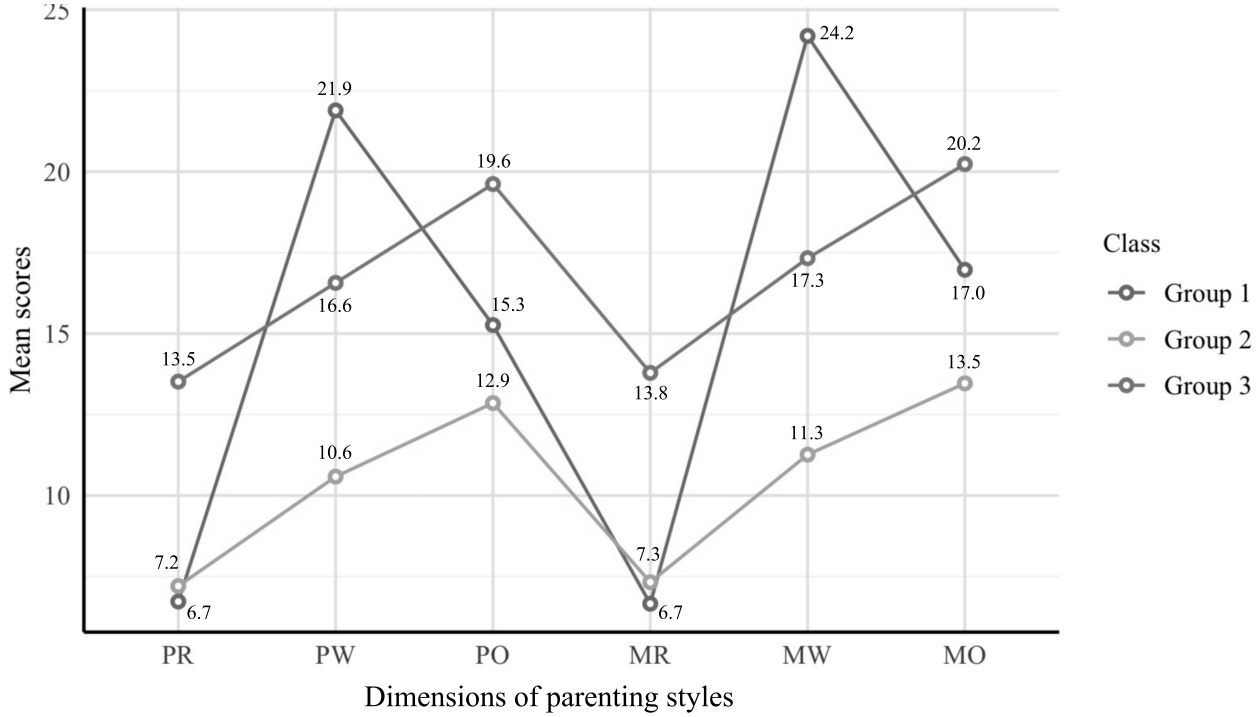

The emotionally distant parenting group was relatively small, which may have reduced power to detect small effects. Therefore, nonsignificant subgroup findings should be interpreted with caution.

XSL•FO
RenderX

Hierarchical Regression Analyses of Social Media Addiction

Table 3 presents the results of the hierarchical regression analyses. In model 1, parenting profiles significantly predicted social media addiction. Compared with the supportive parenting group, participants in the emotionally distant (β=1.417; *P*=.003) and controlling-critical (β=4.438; *P*<.001) parenting groups reported higher levels of social media addiction. After adjusting for demographic variables in model 2, these associations remained significant. In model 3, both perceived burdensomeness (β=0.096; *P*<.001) and thwarted belongingness (β=0.089; *P*<.001) were positively associated with social media addiction. Importantly, even after accounting for these interpersonal need factors, the controlling-critical parenting group continued to report higher levels compared with the supportive parenting group (β=2.081; *P*<.001).

XSL•FO
RenderX

Table . Regression analyses predicting social media addiction from parenting profiles.

Variables	Model 1[a]		Model 2[b]		Model 3[c]	
	β (SE)	*P* value	β (SE)	*P* value	β (SE)	*P* value
Parenting profile						
Supportive parenting	Reference	Reference	Reference	Reference	Reference	Reference
Emotionally distant parenting	1.417 (0.468)	.003	0.774 (0.478)	.11	–0.419 (0.498)	.40
Controlling-critical parenting	4.438 (0.328)	<.001	3.814 (0.347)	<.001	2.081 (0.400)	<.001
Sex						
Male	__[d]	—	Reference	Reference	Reference	Reference
Female	—	—	0.740 (0.287)	.01	0.865 (0.282)	.002
Age	—	—	−0.098 (0.121)	.42	−0.148 (0.119)	.22
University type						
Medical	—	—	Reference	Reference	Reference	Reference
Science and engineering	—	—	1.641 (0.441)	<.001	1.397 (0.433)	.001
Liberal arts	—	—	2.187 (0.343)	<.001	1.823 (0.339)	<.001
Academic year						
First year	—	—	Reference	Reference	Reference	Reference
Second year	—	—	0.133 (0.379)	.73	0.259 (0.372)	.49
Third year	—	—	0.795 (0.457)	.08	0.862 (0.448)	.05
Fourth year or above	—	—	−0.231 (0.697)	.74	−0.173 (0.684)	.80
Only child						
Yes	—	—	Reference	Reference	Reference	Reference
No	—	—	−0.579 (0.311)	.06	−0.681 (0.305)	.03
Father's educational level						
Primary school or lower	—	—	Reference	Reference	Reference	Reference
Junior or senior high school	—	—	−0.421 (0.514)	.41	−0.407 (0.504)	.42
College or above	—	—	−0.090 (0.593)	.88	−0.064 (0.582)	.91
Mother's educational level						
Primary school or lower	—	—	Reference	Reference	Reference	Reference
Middle school	—	—	−0.019 (0.461)	.97	0.176 (0.453)	.70
High school	—	—	−0.314 (0.562)	.58	0.008 (0.552)	.99
Perceived burdensomeness	—	—	—	—	0.096 (0.022)	<.001
Thwarted belongingness	—	—	—	—	0.089 (0.015)	<.001

[a]R^2=0.094.

[b]R^2=0.125.

[c]R^2=0.160.

[d]Not applicable.

XSL•FO
RenderX

Pathways Linking Parenting Profiles to Social Media Addiction

Figure 2 presents the pathways linking parenting profiles to social media addiction through perceived burdensomeness and thwarted belongingness. Model fit indexes indicated a good fit (comparative fit index=0.963; Tucker-Lewis index=0.931; root mean squared error of approximation=0.038; standardized root mean squared residual=0.017).

Figure 2. Mediation model from parenting profiles to social media addiction through perceived burdensomeness and thwarted belongingness. Standardized coefficients (β) are reported. ***P<.001; **P<.01; *P<.05; group 1: supportive parenting; group 2: emotionally distant parenting; group 3: controlling-critical parenting.

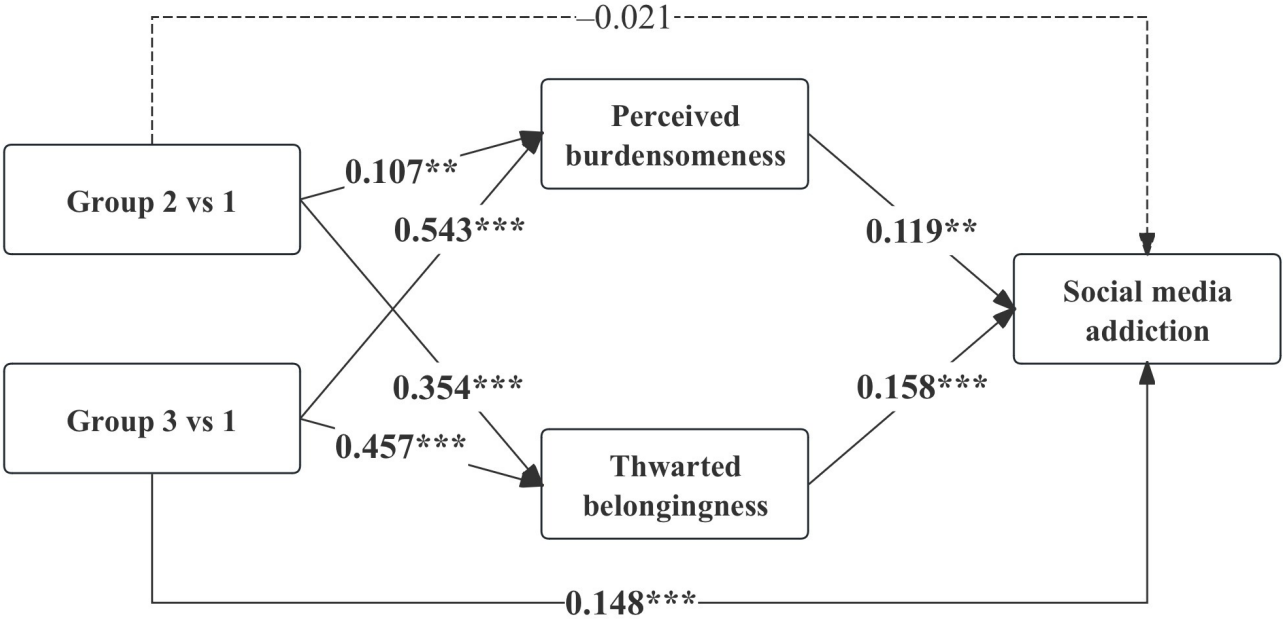

Parenting profiles were significantly associated with both interpersonal need dimensions. Compared with the supportive parenting group, participants in the emotionally distant parenting group reported higher levels of perceived burdensomeness (β=0.107; P<.001) and thwarted belongingness (β=0.354; P<.001). Similarly, students in the controlling-critical parenting group showed positive associations with perceived burdensomeness (β=0.543; P<.001) and thwarted belongingness (β=0.457; P<.001). In turn, both perceived burdensomeness (β=0.119; P=.001) and thwarted belongingness (β=0.158; P<.001) were positively associated with social media addiction. Detailed estimates of all paths in the model can be found in Table S3 in Multimedia Appendix 1.

Mediation analyses further indicated that the total indirect effect was significant for both the emotionally distant (β=0.069; P<.001) and controlling-critical (β=0.137; P<.001) parenting groups compared with the supportive parenting group. Notably, the direct effect of the emotionally distant parenting group was not significant (β=–0.021; P=.43), whereas the controlling-critical parenting group showed a significant direct effect on social media addiction (β=0.148; P<.001).

Multigroup Analyses by Sex and Only Child Status

Multigroup analyses revealed that the association patterns between parenting profiles and the mediating variables were largely consistent across subgroups. Thwarted belongingness was positively associated with social media addiction in all groups, including male students (β=0.150; P<.001), female students (β=0.175; P<.001), only children (β=0.166; P<.001), and non–only children (β=0.155; P<.001). In contrast, perceived burdensomeness was significantly associated with social media addiction among male students (β=0.180; P<.001) but not among female students (β=0.038; P=.30). For only child status, perceived burdensomeness remained significant in both only children (β=0.155; P<.001) and non–only children (β=0.083; P=.02). In addition, the direct effect of the controlling-critical parenting profile on social media addiction remained significant across subgroups, including male students (β=0.147; P<.001), female students (β=0.158; P<.001), only children (β=0.111; P=.009), and non–only children (β=0.192; P<.001). The results of multigroup analyses based on sex and only child status are shown in Figures 3 and 4.

XSL•FO
RenderX

Figure 3. Multigroup analyses stratified by sex. Standardized coefficients (β) are reported for male individuals (A) and female individuals (B). ***P<.001; **P<.01; *P<.05; group 1: supportive parenting; group 2: emotionally distant parenting; group 3: controlling-critical parenting.

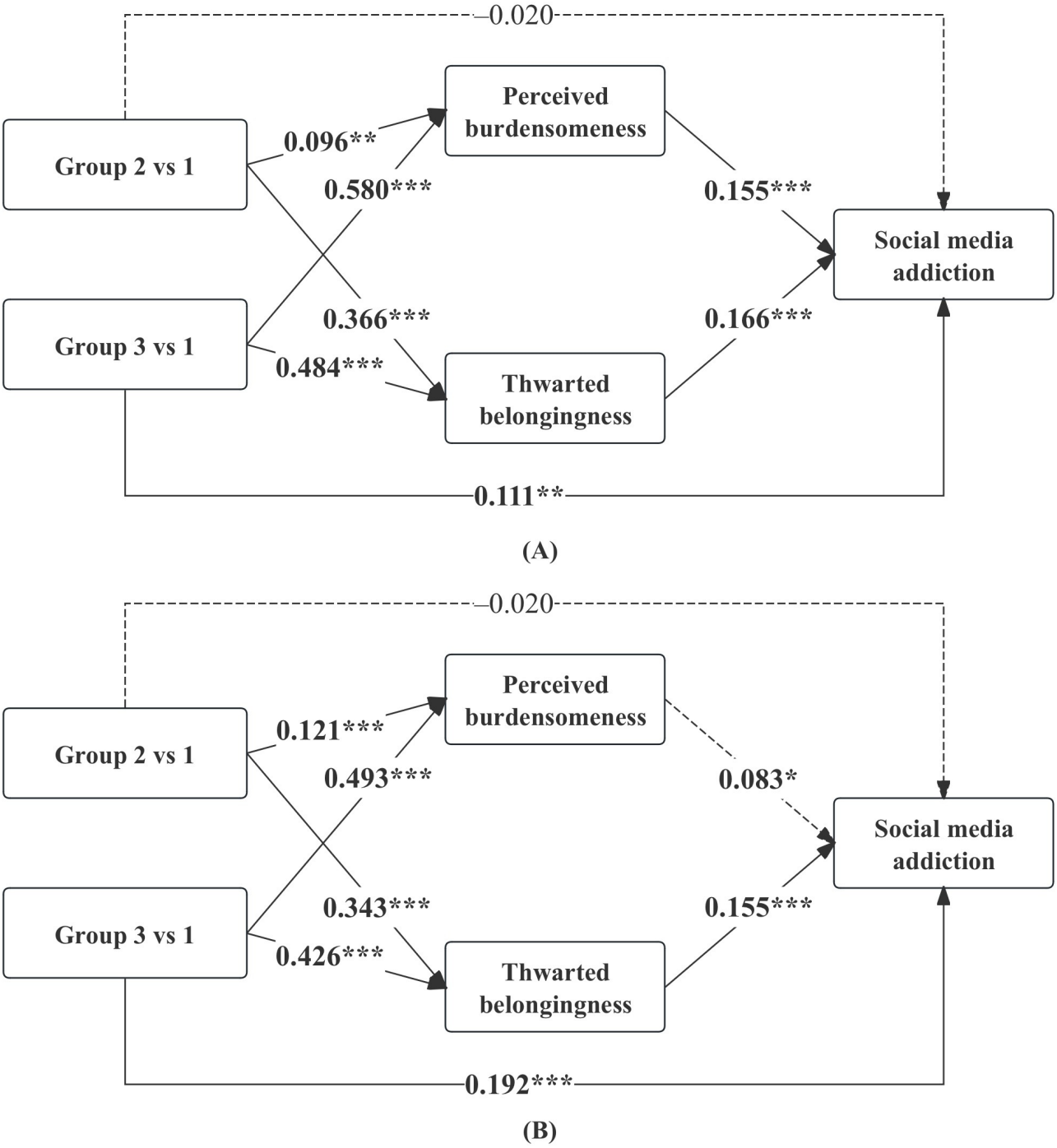

XSL•FO
RenderX

Figure 4. Multigroup analyses stratified by only child status. Standardized coefficients (β) are reported for only children (A) and non–only children (B). ***P<.001; **P<.01; *P<.05; group 1: supportive parenting; group 2: emotionally distant parenting; group 3: controlling-critical parenting.

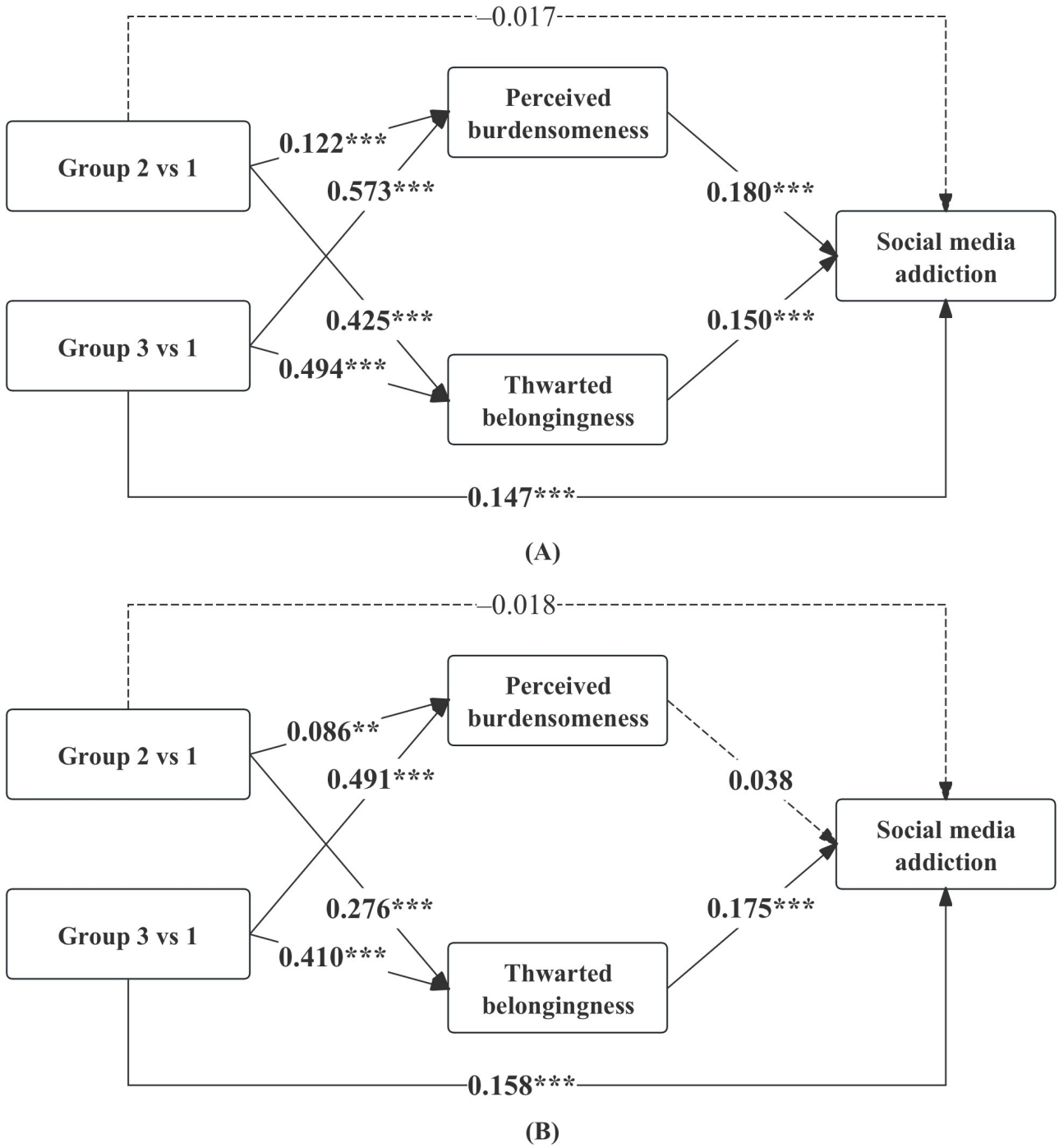

Discussion

Main Findings

This study identified 3 latent profiles of parenting styles among university students: supportive, emotionally distant, and controlling-critical parenting. Compared to the supportive group, both the emotionally distant and controlling-critical groups reported higher levels of social media addiction, a prominent form of digital dependence. These associations were partly mediated by perceived burdensomeness and thwarted belongingness.

Drawing on paternal and maternal rejection, emotional warmth, and overprotection, 3 parenting profiles were identified through LPA. Consistent with previous studies on Chinese families, the largest profile was supportive parenting, characterized by higher levels of emotional warmth and lower levels of control [30,31]. Beyond this parenting style, 2 maladaptive profiles also emerged: an emotionally distant group with limited emotional warmth and a controlling-critical group with elevated rejection and overprotection. Unlike earlier studies that broadly categorized parenting into positive and negative types [32], our findings distinguished 2 qualitatively different forms of maladaptive parenting, with one defined by deficiencies in positive parenting and the other defined by excessive negative

XSL•FO
RenderX

behaviors. This distinction provides a foundation for further understanding how parenting styles contribute to different health risks.

In this study, university students with emotionally distant or controlling-critical parents had higher levels of social media addiction compared to those with supportive parents. Prior studies have indicated that parental rejection and overprotection increase the risk of digital addiction [33,34], whereas a supportive general parenting context provides protection against problematic engagement with social media [35,36]. Researchers have further suggested that maladaptive parenting may undermine children's 3 fundamental psychological needs, namely, relatedness, autonomy, and competence, which subsequently drives individuals to seek compensation through online interactions [37]. In line with this perspective, the attenuation of the association between parenting and social media addiction after accounting for psychological factors highlights the critical role of perceived burdensomeness and thwarted belongingness.

Path analysis further confirmed that perceived burdensomeness and thwarted belongingness mediated the association between parenting profiles and social media addiction. Previous research indicates that maladaptive parenting can contribute to heightened psychological and cognitive vulnerabilities [38,39], which in turn foster unmet interpersonal needs [22]. Although originally conceptualized within the framework of suicidal behavior [16], perceived burdensomeness and thwarted belongingness have been linked to a broader range of adverse outcomes [17,19,40]. Specifically, perceived burdensomeness has been associated with social media addiction [21], suggesting that it may represent an internal manifestation of unsatisfied social needs in offline interactions. To alleviate this deficiency, individuals may increasingly rely on digital technologies as a compensatory mechanism, increasing the risk of problematic use. Building on prior evidence, our findings indicate perceived burdensomeness and thwarted belongingness as psychological pathways linking maladaptive parenting to social media addiction.

The association between parenting and unmet interpersonal needs may vary across different parenting profiles. Previous research suggests that specific parenting styles are linked to distinct personality traits [41], which may subsequently shape psychological states. Consistently, we found that emotionally distant parenting was more strongly related to thwarted belongingness, reflecting insufficient emotional responsiveness that fosters insecure attachment and undermines children's sense of connection and inclusion [42]. In contrast, controlling and critical parenting often involves disapproval of children's feelings and excessive control that can foster perceptions of being a burden to others [37] and, thus, showed a closer association with perceived burdensomeness.

This heterogeneity was further reflected in the different mediation patterns observed across the 2 maladaptive parenting profiles. For the emotionally distant profile, the association with social media addiction became nonsignificant after perceived burdensomeness and thwarted belongingness were included. This pattern suggests that emotionally distant parenting may influence social media addiction primarily through unmet interpersonal needs, particularly reduced belongingness. In contrast, the controlling-critical profile showed evidence of partial mediation. One possible explanation is that, beyond undermining interpersonal needs, controlling-critical parenting may also affect other psychological processes, such as perceived meaning and social support [37,43], which were not captured in this model. Together, these findings indicate that different forms of maladaptive parenting may contribute to social media addiction through partially overlapping yet distinct psychological pathways.

In the multigroup analyses, perceived burdensomeness was associated with social media addiction only among male students, whereas thwarted belongingness remained significant across subgroups. Compared with thwarted belongingness, perceived burdensomeness is more closely related to self-worth and perceived social value [16]. In many sociocultural contexts, male individuals may be more likely to tie self-evaluation to competence, independence, and usefulness. Under such conditions, feeling like a burden to others may more readily translate into compensatory engagement with social media, where recognition or distraction can be more easily obtained. This interpretation is consistent with prior studies suggesting gender differences in the roles of unmet interpersonal needs as risk factors, with perceived burdensomeness appearing particularly relevant for male individuals [44]. In contrast, the stable association with thwarted belongingness across subgroups suggests that unmet needs for social connection may represent a more general pathway to problematic social media use.

Potential conceptual overlap between parenting styles and unmet interpersonal needs should be acknowledged. However, these constructs differ in temporal focus and conceptual level, with parenting reflecting early-life experiences and interpersonal needs capturing current internal states. Therefore, these needs may represent internalized consequences of earlier experiences rather than direct equivalents of parenting. Although shared method variance and partial overlap cannot be ruled out, the findings offer several theoretical contributions. They extend the interpersonal theory of suicide to the domain of digital addiction and integrate it with the social compensation hypothesis, highlighting that compensatory online engagement may operate through specific unmet interpersonal needs. The identification of distinct parenting profiles further shows that different forms of maladaptive parenting may confer risk through distinct psychological pathways. These findings provide a more cohesive framework for understanding how early family experiences shape vulnerability to problematic social media use in emerging adulthood.

Implications

These findings reveal the potential role of early family experiences in the development of social media addiction during emerging adulthood, underscoring the importance of supportive parenting practices. Emotional warmth can be fostered through expressions of care, positive interactions, and open communication [45]. Meanwhile, parental supervision should be aligned with developmental needs, respecting children's autonomy and personal boundaries while minimizing psychological control [45]. Furthermore, unmet interpersonal

XSL•FO
RenderX

needs may represent a fundamental psychological mechanism underlying social media addiction. These states are modifiable through interventions that strengthen interpersonal communication skills and restore self-worth, thereby reducing reliance on online interactions as a compensatory strategy.

The distinct mediation patterns observed across parenting profiles suggest that intervention priorities may differ by family context. For individuals from emotionally distant families, interventions focused on strengthening belongingness and emotional connectedness may be particularly relevant. In contrast, those from controlling-critical families may require broader multicomponent approaches beyond interpersonal needs while still emphasizing the reduction of perceived burdensomeness through greater autonomy support. In addition, approaches may benefit from sensitivity to gender differences, particularly in addressing burdensomeness-related cognitions and self-worth among male students. These findings suggest that screening for parenting-related risk patterns and unmet interpersonal needs may help guide more tailored interventions.

Limitations

This study has several limitations that should be acknowledged. First, the cross-sectional design prevents causal inference, and the observed associations may be influenced by bidirectional or reciprocal effects. Future longitudinal or experimental studies are warranted to clarify the temporal relationship among parenting styles, interpersonal needs, and social media addiction. Second, parenting styles were assessed through participants' retrospective self-reports, which may be subject to recall bias and personal interpretation. Third, although the study examined social media addiction, it did not differentiate between specific platforms or investigate the underlying motives for social media use, which may provide further insights into patterns of digital engagement. Fourth, the study sample was restricted to university students in Shanghai, which may limit the generalizability of the findings to other regions of China and other cultural contexts. In particular, the salience of family interdependence, relational obligations, and unmet interpersonal needs may differ across sociocultural settings. Therefore, cross-cultural studies are needed to examine whether similar parenting profiles and interpersonal pathways operate in other contexts. Fifth, although a multistage sampling approach was used to account for variation across universities, majors, and academic years, the sample cannot be considered fully representative of university students in Shanghai because proportional allocation across strata was not implemented. In addition, because demographic information for the excluded respondents was not retained or was not comparable, potential selection bias cannot be ruled out. Exclusion criteria may also have preferentially excluded less attentive or lower-engagement respondents. Sixth, although the overall sample size was relatively large, some subgroup analyses were based on smaller subsamples, particularly the emotionally distant parenting group and the stratified multigroup models. Accordingly, nonsignificant findings in these analyses should be interpreted cautiously. Seventh, although intergenerational coresidence is relatively common in the Chinese context and may influence relational experiences, this specific aspect of family structure was not assessed in this study. Future research should incorporate measures of intergenerational context to better understand its influence on social media use. Finally, this study did not account for potentially relevant confounders such as mental health problems or personality traits, which may have influenced both interpersonal needs and social media use.

Conclusions

This study identified 3 distinct parenting profiles among university students, namely, supportive, emotionally distant, and controlling-critical parenting, and examined their associations with social media addiction. Students with emotionally distant or controlling-critical parents reported higher levels of social media addiction compared to those with supportive parents. Perceived burdensomeness and thwarted belongingness mediated these associations. Notably, perceived burdensomeness was associated with social media addiction only among male students, highlighting potential gender differences in underlying mechanisms. These findings underscore the importance of early interventions that promote parental emotional warmth, reduce rejecting and controlling behaviors, and address unmet interpersonal needs by strengthening offline social connections. Prevention efforts may further benefit from tailored approaches that consider specific interpersonal vulnerabilities and family contexts. The ultimate goal is to mitigate the risk of social media addiction among young adults.

Funding

This study was supported by the Key Discipline Project of the Shanghai Three-Year Action Plan for Public Health (grant GWVI-11·1-29), Shanghai Municipal Commission of Health (grant 202340019), Shanghai Rehabilitation Medicine Association Health Management Research Fund Project (grant 2023JGKT15), and Year 2023 Shanghai Jiao Tong University Morning Star Outstanding Young Management Personnel Program.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to privacy or ethical restrictions but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: YC
Data curation: SW, BW, XG

XSL•FO
RenderX

Formal analysis: YL, ST, FH
Funding acquisition: YC
Methodology: XY, YC
Writing—original draft: YL, SL, CX
Writing—review and editing: YW, XY, YC
YC and XY are co-corresponding authors of this paper.
All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional tables and figures supporting the main analyses.
[DOCX File, 761 KB - jmir_v28i1e91861_app1.docx]

Checklist 1
STROBE checklist.
[PDF File, 126 KB - jmir_v28i1e91861_app2.pdf]

References

1. Svicher A, Fioravanti G, Casale S. Identifying the central symptoms of problematic social networking sites use through network analysis. J Behav Addict 2021 Aug 25;10(3):767-778. [doi: 10.1556/2006.2021.00053] [Medline: 34437299]
2. Cheng C, Lau YC, Chan L, Luk JW. Prevalence of social media addiction across 32 nations: meta-analysis with subgroup analysis of classification schemes and cultural values. Addict Behav 2021 Jun;117:106845. [doi: 10.1016/j.addbeh.2021.106845] [Medline: 33550200]
3. Yu L, Luo T. Social networking addiction among Hong Kong university students: its health consequences and relationships with parenting behaviors. Front Public Health 2021;8:555990. [doi: 10.3389/fpubh.2020.555990] [Medline: 33569365]
4. Chen XM, Ning YF, Flett GL, et al. The relationship between specific problematic internet use and hope: academic exhaustion as mediator and mattering as moderator among Chinese university students. BMC Psychol 2025 Mar 4;13(1):194. [doi: 10.1186/s40359-025-02500-x] [Medline: 40038829]
5. Luo Y, Liang L, Zhou H, Qi X, Xu J, Meng H. The impact of social media addiction on depression among university students in Wuhan, China: a longitudinal study. BMC Public Health 2025 Jul 2;25(1):2279. [doi: 10.1186/s12889-025-23443-3] [Medline: 40604701]
6. Alshowkan A, Shdaifat E. Fear of missing out, social media addiction, and personality traits among nursing students: cross-sectional study. JMIR Nurs 2025 May 29;8:e71502. [doi: 10.2196/71502] [Medline: 40440638]
7. Eichenberg C, Schneider R, Rumpl H. Social media addiction: associations with attachment style, mental distress, and personality. BMC Psychiatry 2024 Apr 15;24(1):278. [doi: 10.1186/s12888-024-05709-z] [Medline: 38622677]
8. Chen P, Harris KM. Association of positive family relationships with mental health trajectories from adolescence to midlife. JAMA Pediatr 2019 Dec 1;173(12):e193336. [doi: 10.1001/jamapediatrics.2019.3336] [Medline: 31589247]
9. Yan Y. The private life approach to the rise of neo-familism in China. Sociol Rev 2025 Jul;73(4):753-770. [doi: 10.1177/00380261251347745]
10. Lu H, Ren Z, Zhen P, Su Z, Wu Y. The longitudinal association between family functioning and problematic social media use among Chinese university students: mediation via loneliness and a subgroup analysis by sex. Addict Behav 2025 Jul;166:108337. [doi: 10.1016/j.addbeh.2025.108337] [Medline: 40121924]
11. Wei S, Teo T, Malpique A, Lausen A. Parental autonomy support, parental psychological control and Chinese university students' behavior regulation: the mediating role of basic psychological needs. Front Psychol 2022;12:735570. [doi: 10.3389/fpsyg.2021.735570] [Medline: 35250687]
12. Hwang W, Fu X, Kim S, Jung E, Zhang Y. A multidimensional construct of helicopter parenting and college students' game and social media addictive behaviors: a cross-cultural study in South Korea and China. Front Psychol 2023;13:1022914. [doi: 10.3389/fpsyg.2022.1022914] [Medline: 36938126]
13. Yuan XQ, Dou K, Li YY. The longitudinal association between negative life events and problematic social media use among Chinese college students: the mediating role of FoMO and the moderating role of positive parenting. Stress Health 2024 Dec;40(6):e3505. [doi: 10.1002/smi.3505] [Medline: 39572931]
14. Kardefelt-Winther D. A conceptual and methodological critique of internet addiction research: towards a model of compensatory internet use. Comput Hum Behav 2014 Feb;31:351-354. [doi: 10.1016/j.chb.2013.10.059]
15. Wegmann E, Brand M. A narrative overview about psychosocial characteristics as risk factors of a problematic social networks use. Curr Addict Rep 2019 Dec;6(4):402-409. [doi: 10.1007/s40429-019-00286-8]

XSL•FO
RenderX

16. Van Orden KA, Witte TK, Cukrowicz KC, Braithwaite SR, Selby EA, Joiner TEJ. The interpersonal theory of suicide. Psychol Rev 2010 Apr;117(2):575-600. [doi: 10.1037/a0018697] [Medline: 20438238]
17. Amin SM, Khedr MA, Tawfik AF, Gamal Noaman Malek M, El-Ashry AM. The mediating and moderating role of social support on the relationship between psychological well-being and burdensomeness among elderly with chronic illness: community nursing perspective. BMC Nurs 2025 Feb 10;24:156. [doi: 10.1186/s12912-025-02743-4]
18. Lee H, Cho M, Park CH. Impact of childhood trauma on suicidal ideation: sequential mediating effects of interpersonal sensitivity and interpersonal needs. J Affect Disord 2025 Sep 15;385:119361. [doi: 10.1016/j.jad.2025.05.021] [Medline: 40339716]
19. Ooi PB, Khor KS, Tan CC, Ong DL. Depression, anxiety, stress, and satisfaction with life: moderating role of interpersonal needs among university students. Front Public Health 2022;10:958884. [doi: 10.3389/fpubh.2022.958884] [Medline: 36249213]
20. Ma M, Dong S, Lin S, Zhang L. Status speeds decisions? Cultural differences in the impact of subjective socioeconomic status on decision difficulty between China and the United States. Int J Psychol 2026 Feb;61(1):e70165. [doi: 10.1002/ijop.70165] [Medline: 41536048]
21. Wang P, Ouyang M, Yin Y, Li B. Mother phubbing and adolescents' problematic SNS use: the mediating role of perceived burdensomeness and the moderating role of need to belong. Front Psychol 2023;14:1098707. [doi: 10.3389/fpsyg.2023.1098707] [Medline: 37359861]
22. Maultsby K, López RJ, Wolff J, Spirito A, Esposito-Smythers C. Longitudinal relations between parenting practices and adolescent suicidal ideation in a high-risk clinical sample: a moderated mediation model. Res Child Adolesc Psychopathol 2023;51:613-623. [doi: 10.1007/s10802-022-01018-9]
23. Jiang J, Lu ZR, Jiang BJ, Xu Y. Revision of the Short-form Egna Minnenav Barndoms Uppfostran for Chinese. Psychol Dev Educ 2010;26(1):94-99 [FREE Full text]
24. Perris C, Jacobsson L, Lindström H, von Knorring L, Perris H. Development of a new inventory assessing memories of parental rearing behaviour. Acta Psychiatr Scand 1980 Apr;61(4):265-274. [doi: 10.1111/j.1600-0447.1980.tb00581.x] [Medline: 7446184]
25. Van Orden KA, Cukrowicz KC, Witte TK, Joiner TE. Thwarted belongingness and perceived burdensomeness: construct validity and psychometric properties of the Interpersonal Needs Questionnaire. Psychol Assess 2012 Mar;24(1):197-215. [doi: 10.1037/a0025358] [Medline: 21928908]
26. Amini-Tehrani M, Sadeghi R, Nasiri M, Jalali T, Zamanian H. Suicide-related interpersonal needs of young Iranian people: a preliminary validation of thwarted belongingness and perceived burdensomeness constructs. Clin Psychol Psychother 2021 Mar;28(2):422-437. [doi: 10.1002/cpp.2507] [Medline: 32852132]
27. Iliceto P, D'Antuono L, Fino E, et al. Psychometric properties of the Italian version of the Interpersonal Needs Questionnaire-15 (INQ-15-I). J Clin Psychol 2021 Jan;77(1):268-285. [doi: 10.1002/jclp.23026] [Medline: 32662083]
28. Li XM, Xin TG, Yuan J, Lv LX, Tao JY, Liu Y. Validity and reliability of the interpersonal needs questionnaire in Chinese college students. Chin J Clin Psychol 2015;23(4):590-593 [FREE Full text]
29. Wang SP, Cai Y, Zhu R, Li SC, Liu XY, Gong RJ. Development of social network addiction tendency scale for adolescents and its reliability and validity. Chin J Gen Pract 2022;20(2):324-326. [doi: 10.16766/j.cnki.issn.1674-4152.002343]
30. Lan X. Perceived parenting styles, cognitive flexibility, and prosocial behavior in Chinese youth with an immigrant background: a three-group comparison. Curr Psychol 2023 Aug;42:20718-20736. [doi: 10.1007/s12144-022-03140-9]
31. Ren L, Li Y, Li X, Li J, Li L, Fan J. Reconsidering the typology of parenting styles and its association with preschoolers' development in Chinese families. Child Dev 2025;96(4):1326-1342. [doi: 10.1111/cdev.14236] [Medline: 40272119]
32. Cui X, Zhang X, Zhang H. The impact of parenting style on malevolent creativity based on Chinese university students: a latent profile analysis. Front Psychol 2024;15(1363778). [doi: 10.3389/fpsyg.2024.1363778] [Medline: 38988383]
33. Liu W, Zeng S. Negative parenting styles and problematic mobile phone use in university students: adult attachment and fear of missing out as chain mediators. Curr Psychol 2024 Oct;43(38):30259-30267. [doi: 10.1007/s12144-024-06651-9]
34. Yao MZ, He J, Ko DM, Pang K. The influence of personality, parental behaviors, and self-esteem on Internet addiction: a study of Chinese college students. Cyberpsychol Behav Soc Netw 2014 Feb;17(2):104-110. [doi: 10.1089/cyber.2012.0710] [Medline: 24003966]
35. Lukavská K, Hrabec O, Lukavský J, Demetrovics Z, Király O. The associations of adolescent problematic internet use with parenting: a meta-analysis. Addict Behav 2022 Dec;135:107423. [doi: 10.1016/j.addbeh.2022.107423] [Medline: 35933287]
36. Geurts SM, Koning IM, Van den Eijnden R, Vossen HG. Predicting adolescents' problematic social media use from profiles of internet-specific parenting practices and general parenting dimensions. J Youth Adolesc 2023 Sep;52(9):1829-1843. [doi: 10.1007/s10964-023-01816-4] [Medline: 37403002]
37. Kong F, Wei W, Xia Y, et al. Mediating roles of psychological entitlement and presence of meaning in the associations between parental measures (rejection and overprotection) and adolescent internet addiction. J Affect Disord 2025 Aug 1;382:507-517. [doi: 10.1016/j.jad.2025.04.131] [Medline: 40286930]
38. Khaleque A. Perceived parental hostility and aggression, and children's psychological maladjustment, and negative personality dispositions: a meta-analysis. J Child Fam Stud 2017 Apr;26(4):977-988. [doi: 10.1007/s10826-016-0637-9]

XSL•FO
RenderX

39. Oppenheimer CW, Hankin BL, Young J. Effect of parenting and peer stressors on cognitive vulnerability and risk for depression among youth. J Abnorm Child Psychol 2018 Apr;46(3):597-612. [doi: 10.1007/s10802-017-0315-4] [Medline: 28623624]
40. Gallyer AJ, Dougherty SP, Gai AR, et al. Problematic alcohol use and suicidal ideation among firefighters: a multi-study investigation of the explanatory roles of perceived burdensomeness and thwarted belongingness. J Affect Disord 2018 Oct 1;238:281-288. [doi: 10.1016/j.jad.2018.05.045] [Medline: 29898427]
41. Arrindell WA, Sanavio E, Aguilar G, et al. The development of a short form of the EMBU: its appraisal with students in Greece, Guatemala, Hungary and Italy. Pers Individ Dif 1999 Oct;27:613-628. [doi: 10.1016/S0191-8869(98)00192-5]
42. Bowlby J. Attachment and Loss Volume I Attachment: Basic Books; 1969.
43. Lo CK, Ho FK, Emery C, et al. Association of harsh parenting and maltreatment with internet addiction, and the mediating role of bullying and social support. Child Abuse Negl 2021 Mar;113:104928. [doi: 10.1016/j.chiabu.2021.104928] [Medline: 33429283]
44. Donker T, Batterham PJ, Van Orden KA, Christensen H. Gender-differences in risk factors for suicidal behaviour identified by perceived burdensomeness, thwarted belongingness and acquired capability: cross-sectional analysis from a longitudinal cohort study. BMC Psychol 2014;2(1):20. [doi: 10.1186/2050-7283-2-20] [Medline: 25815191]
45. Smetana JG. Current research on parenting styles, dimensions, and beliefs. Curr Opin Psychol 2017 Jun;15:19-25. [doi: 10.1016/j.copsyc.2017.02.012] [Medline: 28813261]

Abbreviations

BIC: Bayesian information criterion
INQ-15: 15-item version of the Interpersonal Needs Questionnaire
LPA: latent profile analysis
S-EMBU-C: short-form Egna Minnen av Barndoms Uppfostran for Chinese

Edited by M Balcarras; submitted 21.Jan.2026; peer-reviewed by A Jameel, M Shamir; revised version received 26.Mar.2026; accepted 14.Apr.2026; published 27.May.2026.

Please cite as:
Liu Y, Wang S, Wei T, Wu B, Ge X, Liu S, Xu C, Tao S, Fan X, Hu F, Wang Y, Yang X, Cai Y
The Mediating Role of Interpersonal Needs in Perceived Parenting Styles and Social Media Addiction Among University Students: Cross-Sectional Study
J Med Internet Res 2026;28:e91861
URL: https://www.jmir.org/2026/1/e91861
doi:10.2196/91861

© Yujie Liu, Suping Wang, Ting Wei, Birong Wu, Xin Ge, Shangbin Liu, Chen Xu, Shunyu Tao, Xiaohong Fan, Fan Hu, Ying Wang, Xue Yang, Yong Cai. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Preventive Social Media Intervention for Perinatal Depression and Anxiety in Regional, Rural, and Remote Communities: Participatory Co-Design Study

Kacey Jane Lynch[1], BPsych; Adrian Brian Royce Shatte[2,3], PhD; Jessica Muller[1,3,4], PhD; Kendall George[5,6], PhD; Gisele Rossini[1], MPH; Angela Anson[7], MBBS MP; Courtney Hala[8], BMid, BSocSc; James Dimmock[1,3], PhD; Delyse Hutchinson[9,10,11,12], PhD; Samantha Teague[1,3,9], PhD

[1]Department of Psychology, College of Healthcare Sciences, James Cook University, 1 James Cook Drive, Townsville, Queensland, Australia
[10]Centre for Adolescent Health, Murdoch Childrens Research Institute, Melbourne Royal Childrens Hospital, Melbourne, Australia
[11]Department of Paediatrics, University of Melbourne, Melbourne, Australia
[12]National Drug and Alcohol Research Centre, University of New South Wales, Sydney, Australia
[2]Department of Information Technology, College of Science and Engineering, James Cook University, Townsville, Queensland, Australia
[3]Margaret Roderick Centre for Mental Health Research, James Cook University, Townsville, Queensland, Australia
[4]National Centre for Indigenous Genomics, Australian National University, Canberra, Australian Capital Territory, Australia
[5]Department of Nursing and Midwifery, College of Healthcare Sciences, James Cook University, Townsville, Queensland, Australia
[6]Women's and Children's Service Group, Townsville Hospital, Townsville, Queensland, Australia
[7]Cairns Perinatal and Infant Mental Health (PIMH), Cairns Hospital, Cairns, Queensland, Australia
[8]Children's Health Queensland Hospital and Health Service, Brisbane, Queensland, Australia
[9]Deakin Lifespan Institute, School of Psychology, Faculty of Health, Deakin University, Melbourne, Australia

Corresponding Author:
Samantha Teague, PhD
Department of Psychology, College of Healthcare Sciences, James Cook University, 1 James Cook Drive, Townsville, Queensland, Australia

Abstract

Background: Perinatal depression and anxiety are significant public health concerns, affecting up to 1 in 5 women globally, with disproportionate burden carried by women in regional, rural, and remote communities where structural and social inequities amplify vulnerability. Access to perinatal mental health support in these settings is severely constrained by geographical isolation, workforce shortages, financial barriers, and a lack of culturally safe services. Prevention is recognized as critical to reducing this burden, with evidence suggesting that effective preventive approaches can reduce population-level illness by up to 40% and alleviate downstream demand on overstretched services. Digital mental health interventions hold promise for improving access to support, yet few are co-designed with underserved perinatal populations.

Objective: This study aimed to identify the mental health needs of perinatal women in regional, rural, and remote communities and to co-design a framework for a preventive social media–based intervention informed by platform-specific affordances and constraints, using Northern Queensland, Australia, as an exemplar.

Methods: Using a participatory co-design approach, 26 perinatal women (21 postnatal and 5 antenatal) and 8 mental health care professionals from regional, rural, and remote Northern Queensland participated in focus groups or interviews, supplemented by ongoing consultation with a community advisory group comprising lived experience representatives, clinicians, and local community leaders. Qualitative data were analyzed using reflexive thematic analysis to identify core community mental health needs. Identified needs were then examined through a needs-affordances framework to determine how specific platform features could address, enable, or constrain those needs in the context of a preventive intervention.

Results: Five core mental health needs were identified: (1) social connection and support; (2) personalized and respectful health care; (3) information that empowers; (4) place-based and culturally safe support; and (5) accessible, low-burden digital formats. Participants viewed social media as a potentially useful platform for fostering peer connection, normalizing perinatal experiences, and providing timely psychoeducation. However, both mothers and professionals expressed concerns about misinformation, harmful social comparison, and privacy risks that must be proactively addressed in program design. These insights were synthesized into a set of prototype design guidelines specifying recommended content, features, tone, and delivery formats to inform subsequent intervention development.

XSL•FO
RenderX

Conclusions: This study provides a place-based, co-designed needs-affordances framework to guide the development of a preventive social media–based intervention for perinatal mental health support in regional, rural, and remote communities. The findings demonstrate that social media is an acceptable and promising platform for preventive perinatal mental health support in these settings, provided that design is driven by community need, platform affordances are systematically analyzed, and known risks are explicitly mitigated. These findings address a significant gap in the literature and offer a replicable methodological approach for co-designing contextually relevant digital mental health interventions with underserved populations.

(J Med Internet Res 2026;28:e91778) doi:10.2196/91778

KEYWORDS

prevention; mental health; perinatal; maternal; depression; anxiety; pregnant; regional; rural; remote; social media; telemedicine; participatory research; co-design; patient and public involvement; PPI

Introduction

Perinatal depression and anxiety are major public health concerns globally, affecting approximately 1 in 5 women across pregnancy and the first postpartum year, with true prevalence likely higher due to stigma and underreporting [1,2]. These conditions impair maternal functioning, strain intimate and family relationships, and reduce women's quality of life [1,3]. Impacts also extend to infants, who face increased risks of poorer birth outcomes, developmental delays, behavioral dysregulation, and long-term social-emotional difficulties [4].

Australia's public health care system, Medicare, provides universal access to a broad range of medical services through a hybrid public-private model; however, the practical availability of specialized mental health and perinatal care varies enormously by geography [5]. In Queensland, perinatal mental health care services are primarily delivered through public maternity hospitals, community health care centers, and primary care providers under the coordination of primary health networks, such as the Northern Queensland primary health network, which covers a vast and geographically dispersed catchment area spanning from St Lawrence to the Torres Strait and west to Croydon and Kowanyama [6]. This region is characterized by small and remote communities, higher proportions of First Nations populations, and a critical shortage of specialist mental health and perinatal services relative to need [6].

Access to timely and culturally appropriate support is therefore particularly limited in regional, rural, and remote communities across this area, where structural inequities create substantial barriers to care, including geographical isolation, workforce shortages, financial strain, and limited specialist services [7]. Women in these settings experience heightened health care needs yet have lower service usage than women in metropolitan areas, including First Nations mothers who face ongoing impacts of colonization and systemic racism within health care [8]. These disparities are compounded by stigma, stoicism, and concerns about confidentiality in close-knit communities, as well as a lack of culturally safe and responsive care [9]. Addressing these inequities requires place-based preventive strategies that respond to local determinants of health and build on existing community infrastructures [10,11]. Preventive approaches are particularly important for reducing downstream service demand, given the limited availability of perinatal mental health treatment in these regions [10,11].

Despite its importance, prevention remains underdeveloped in perinatal mental health care [12]. Treatment alone has not reduced the population burden, and modeling suggests that even full treatment coverage would avert only a minority of cases without concurrent prevention efforts [13]. Evidence-based preventive interventions, including cognitive behavioral and interpersonal approaches, have been demonstrated to reduce perinatal mental illness by up to 40%, generating substantial health care system efficiencies [14]. Yet preventive support rarely reaches women in regional, rural, and remote communities, where access challenges are most pronounced [15].

Digital delivery modes offer a promising avenue for expanding the reach of prevention. Social media platforms, in particular, provide a low-cost, scalable, engaging, and accessible mechanism for reaching regional, rural, and remote populations [13,16]. Women aged 25 - 34 years—the modal age of first-time mothers—are among the highest users of platforms such as Facebook (Meta Platforms, Inc) and Instagram (Meta Platforms, Inc) and increasingly seek health information through these channels [17-20]. Social media affordances, including ubiquity, algorithm-driven content personalization, broadcast communication, and the ability to form peer networks, align closely with known preventive mechanisms such as normalization, early psychoeducation, emotional support, and strengthened social connection. These affordances may be especially valuable in regional, rural, and remote communities, where opportunities for connection and professional support are more limited. However, social media also carries risks for perinatal well-being, including exposure to misinformation, information overload, racism, and maladaptive social comparisons [21-23], underscoring the need for careful, evidence-informed, and culturally appropriate intervention design to maximize benefits while mitigating harm [24].

A growing body of evidence supports the effectiveness of social media and digitally delivered interventions for improving mental health outcomes in perinatal populations. Systematic reviews and meta-analyses have demonstrated that internet- and app-based interventions can produce significant reductions in perinatal depression and anxiety symptoms, with effect sizes comparable to those of face-to-face interventions, particularly when they incorporate structured psychological content such as cognitive behavioral therapy or mindfulness-based components [25,26]. Perinatal mental health interventions delivered via social media platforms have shown particular

XSL•FO
RenderX

promise, with studies reporting improvements in social connectedness, perceived support, parenting self-efficacy, and reductions in loneliness and depressive symptoms among postpartum women alongside high rates of engagement [27,28]. For example, Facebook-based interventions have reported engagement rates of 83% - 96%, with some also reporting decreased depression severity and improved parenting competence [27,29]. Interventions based on WhatsApp (Meta Platforms, Inc) and WeChat (Tencent Holdings Ltd) have demonstrated feasibility and acceptability in diverse cultural contexts, with some reporting reductions in depression, anxiety, and stress [28,30,31]. Despite this promise, the evidence base remains nascent, with only a handful of studies conducted and critical questions about optimal design and generalizability to underserved populations remaining largely unanswered.

Although interest in social media–based perinatal mental health support is growing, evidence for scalable preventive interventions remains limited. Existing interventions have targeted either prenatal or postpartum women via platforms such as Facebook, WhatsApp, or WeChat to deliver psychoeducation, mindfulness training, peer support, or gamified physical activity programs [27-33]. While most studies report high engagement and improvements in select outcomes, such as parenting competence or mindfulness skills, effects are inconsistent across depression, anxiety, stress, and broader psychosocial functioning [29,31,33]. Most programs are treatment-focused programs for mothers with moderate-to-severe symptoms of depression and/or anxiety rather than universal preventive care [27,28,31,33]. Critically, few interventions have incorporated user perspectives during development or systematically examined the affordances and constraints of social media for prevention. Most have been implemented in high-income urban settings, limiting their generalizability to regional, rural, and remote contexts. To our knowledge, no prior study has conducted a structured needs-affordances analysis to inform the co-design of a social media–based mental health intervention for perinatal women, representing a substantial gap in both research and practice.

This study aimed to address the aforementioned limitation by conducting a novel needs-affordances analysis to support the co-design of a preventive social media–based mental health intervention for women living in regional, rural, and remote Northern Queensland, Australia. Specifically, the study sought to (1) identify the mental health needs of perinatal women in regional, rural, and remote areas and (2) develop prototype design guidelines for a social media program informed by platform-specific affordances and constraints.

Methods

Overview

This study represents Phase 1 of a multiphase project that aims to co-design and evaluate a social media–based universal prevention intervention for perinatal mental health in regional, rural, and remote communities, using Northern Queensland, Australia, as an exemplar. The project uses a participatory co-design approach informed by the Double Diamond model [34] through divergent exploration and convergent refinement to develop and evaluate a preventive perinatal mental health intervention program. Specifically, Phase 1 of the project aimed to discover the problems, insights, user needs, and ideas generated by stakeholders and consider how social media may afford or constrain the mental health needs of North Queensland perinatal women, resulting in a set of design principles for the prototype intervention. In Phase 2, these principles will be applied to develop a prototype intervention integrating evidence-based psychological techniques for preventing perinatal depression and anxiety, which will be iteratively refined through additional focus groups and interviews with stakeholders. In Phase 3, the prototype intervention will be evaluated in a pilot trial to assess its feasibility and acceptability with the target population ahead of a future randomized controlled trial.

The research team comprised mental health care researchers and clinicians (a clinical psychologist, a perinatal psychiatrist, a psychologist, a mental health nurse, midwives, and a social worker) as well as a human-computer interaction researcher. A community advisory group of representatives with lived experience of perinatal depression and/or anxiety diagnoses, perinatal mental health clinicians, and local community leaders provided ongoing consultation throughout the design process. The study also adhered to the GRIPP2 (Guidance for Reporting Involvement of Patients and the Public) checklist (Checklist 1) [35].

Participants

Participants were perinatal women and mental health care professionals residing in Northern Queensland, Australia. Northern Queensland is a regional, rural, and remote area within Queensland, Australia, with a population of approximately 732,000 people spanning over 810,000 km^2 from Sarina on the central Queensland coast up through Mackay, Townsville, and Cairns to the Torres Strait islands and westwards into inland and remote areas of the Gulf of Carpentaria and Mount Isa [6,36].

Perinatal women were eligible if they (1) were aged ≥18 years, (2) were pregnant or ≤12 months postpartum, (3) resided in Northern Queensland, (4) were proficient in English, (5) did not currently meet criteria for a mental health diagnosis, (6) had prior experience with social media, and (7) did not report moderate/severe symptoms or suicidal ideation on standardized screening measures. Mental health status was assessed using the Edinburgh Postnatal Depression Scale (EPDS), with scores >11 indicating moderate-to-severe distress and endorsement of the suicidality item prompting exclusion and referral to supports [37]. For Aboriginal and Torres Strait Islander mothers, Part 1 of the Kimberley Mum Mood Scale (KMMS) was offered as a culturally appropriate adaptation of the EPDS, with the same scoring and exclusion threshold [38]. Both the EPDS and KMMS are recommended as universal perinatal mental health screening instruments in Queensland Health clinical guidelines, with evidence suggesting approximately two-thirds of women are screened using the EPDS in pregnancy statewide [39,40]. Mothers were also asked whether they had a previous diagnosis of a mental health condition in the demographics survey, although this did not form part of the eligibility criteria. No

XSL•FO
RenderX

additional prior mental health data were available for participants.

Mental health care professionals were eligible if (1) they were registered as a mental health care professional, (2) they had experience supporting perinatal mental health, and (3) they lived or worked in Northern Queensland. All participants received an AUD $50 (US $31.48) Prezzee voucher to reimburse them for their time.

Sample size was not predetermined but was guided by the principle of information power, which holds that the adequacy of a qualitative sample is determined by its capacity to illuminate the specific phenomenon under study rather than by a fixed numerical threshold [41]. Information power is shaped by the study aims, the specificity of the participant group, the quality of dialogue, and the analytical approach used. Data collection continued until the research team determined that sufficient information power had been achieved, taking into account the clearly defined study aims, the high specificity of the participant groups, and the depth and richness of data generated across sessions. The final sample of 26 perinatal women and 8 mental health care professionals was considered to reflect this determination.

Recruitment

Participants were recruited using convenience and community-based sampling methods. Recruitment advertisements were distributed via social media, flyers, community organization newsletters, and media releases inviting women and professionals to "help design a digital program for Northern Queensland women's perinatal mental health." Social media marketing was targeted using location and user preference information to reach the intended audience and was used to reach existing users. Snowball recruitment was also encouraged through participants' social networks.

Interested individuals were directed to an online survey hosted in Qualtrics (Qualtrics International, Inc). After reviewing the participant information sheet, individuals provided informed consent, completed eligibility screening, and answered questions on their demographic information. Participants also provided a preferred nickname for use in group discussions and selected their availability for online or in-person sessions.

Procedure

Focus groups and individual interviews were conducted from May 6 to July 4, 2025. Eight focus groups (2-4 participants each) and 15 individual interviews were held, with perinatal women and mental health care professionals participating in separate sessions. Sessions lasted approximately 60 minutes on average. Sessions were offered online via Microsoft Teams or in person in Mackay, Townsville, Cairns, and Mt Isa to maximize accessibility.

A semistructured interview guide was developed, consisting of 7 open-ended questions exploring experiences related to perinatal mental health, social support, and digital technologies (Multimedia Appendix 1). Each session began with an acknowledgment of country, an overview of study aims, and a reminder of participant rights, confidentiality, and voluntary participation.

In-person focus groups were cofacilitated by KJL and either ST or JM. Online sessions were facilitated by the lead author (KJL). Sessions lasted 60 - 90 minutes, were audio recorded, and were supplemented by facilitator field notes to capture contextual information.

Data Analysis

Analysis of Quantitative Data

Quantitative demographic data were summarized using descriptive statistics (frequencies, percentages, means, and SDs). Current distress levels were characterized using mean scores and SDs on the EPDS and, for First Nations participants, the KMMS [37,38]. Remoteness was classified using the Modified Monash Model (MM1-MM7), and socioeconomic status was indexed using the Socio-Economic Indexes for Areas quintile measure, both based on participants' postcode [42,43].

Analysis of Qualitative Data

Overview

A 2-stage needs-affordances analysis was undertaken using a hybrid inductive-deductive approach. Transcript data from perinatal women and mental health care professionals were analyzed separately in the first instance, with each dataset coded independently to preserve the integrity of each group's perspectives. The resulting themes from both datasets were then brought together in a synthesizing phase, allowing convergences, divergences, and complementarities between lived experience and professional perspectives to be identified and integrated.

Stage 1: Identification of Community Mental Health Needs

Transcript data were analyzed inductively using reflexive thematic analysis supported by Taguette (version 1.4.1; The Taguette Project) [44,45]. The lead author (KJL) conducted initial coding of each dataset separately, met regularly with the research team to refine theme boundaries, and consulted with the advisory group to ensure interpretations reflected local context and lived experience. Following independent analysis of each dataset, themes were synthesized across the 2 groups to develop an integrated account of perinatal women's and professionals' perspectives on needs, challenges, and priorities relevant to preventive mental health support.

Stage 2: Social Media Affordances Analysis

The synthesized needs were then examined deductively using the Moreno and D'Angelo [46] social media affordances framework, which conceptualizes how user needs intersect with functional platform features (eg, connectivity, visibility, social support, and information exchange). This framework provided a structured lens through which to map identified community needs onto specific social media platform capabilities, exploring how social media may enable or constrain each need.

The translational process from needs to prototype design guidelines occurred across 3 complementary steps. First, the focus group and interview guide included questions specifically designed to elicit participant perspectives on social media affordances and preferred intervention features. Participants

XSL•FO
RenderX

were asked to reflect on what they believed social media was good and bad for in the context of perinatal mental health and were invited to brainstorm features they would find valuable in a preventive intervention—for example, what would motivate daily engagement, what types of content and features would feel supportive during moments of need, and how they would like local service information to be presented and accessed. This ensured that the affordances analysis was grounded in authentic participant preferences rather than researcher assumptions. Second, the research team applied the affordances framework to systematically analyze how the identified needs could be enabled or constrained by social media platform features, producing a preliminary set of design principles. Third, these preliminary principles were presented to the broader chief investigator team and community advisory group, who were asked to provide input on which features and design elements would best address the identified needs within the regional, rural, and remote context. The finalized prototype design guidelines reflect the integration of participant-generated ideas, affordances-informed analysis, and advisory group input across this iterative process.

Reflexivity and Rigor

The study was conducted from a critical realist perspective, recognizing that participants' accounts reflect both lived experience and broader sociocultural structures, while also acknowledging the influence of researcher positionality [47]. The lead author and facilitator (KJL) was an honors psychology student researcher at the time of data collection, rather than an experienced researcher or clinician, and this positionality is acknowledged as a potential influence on participant rapport and the depth of dialogue in some sessions. KJL also has lived experience of rural health care barriers, which informed her sensitivity to participant perspectives and is reflected in her reflexive memos, maintained throughout the analytical process to document assumptions and decision-making. The presence of mental health care professionals as a separate participant group is additionally acknowledged as a potential influence on power dynamics within those sessions.

Several strategies were used to mitigate these potential influences. In-person focus groups were cofacilitated by experienced mental health care researchers (ST and JM), and KJL received ongoing supervision and methodological guidance from the broader research team throughout data collection and analysis. Perinatal women and mental health care professionals participated in separate focus groups and interviews, ensuring that the potential inhibitory effect of professional authority on mothers' openness was structurally mitigated. Participants were additionally reminded at the outset of each session of the voluntary nature of their participation, their right to withdraw, and were encouraged to share only what they felt comfortable disclosing.

Coding and theme development were reviewed through iterative team discussions and consultation with the advisory group to support reflexivity and critical engagement with the data, including examination of assumptions, alternative interpretations, and researcher positionality, rather than to achieve coding consensus or interrater reliability [45]. Trustworthiness was supported through verbatim transcription, an audit trail of coding decisions, team-based interpretation to enhance credibility and dependability, and alignment of findings with study objectives to ensure analytic coherence. This approach ensured that the identified needs and design principles were grounded in authentic participant perspectives and were directly applicable to the development of a social media–based preventive intervention.

Ethical Considerations

Ethics approval was obtained from the James Cook University Human Research Ethics Committee (HREC 24H-9686), and the study was conducted in accordance with the National Statement on Ethical Conduct in Human Research. All participants provided informed electronic consent before completing the demographic survey and attending any focus group or interview session. Consent materials clearly outlined the voluntary nature of participation, the right to withdraw at any time without consequence, and how data would be stored and used. To protect participant privacy and confidentiality, all data were deidentified before analysis. Participants were invited, though not required, to use pseudonyms during focus group sessions, and any identifying information was removed from study records. Sessions were audio recorded solely for transcription, and recordings were destroyed once transcription was complete. No images of individual participants are included in the paper or additional materials. Participants were compensated for their time with an AUD $50 (US $31.48) Prezzee digital gift voucher upon completion of their focus group or interview session.

Results

Participant Characteristics

A total of 26 perinatal women participated in the study (Table 1). Of these, 5 were pregnant and 21 were postpartum. Fifteen participated in focus groups and 11 participated in interviews. Participants ranged in age from 25 to 48 years (mean 32.2, SD 4.6 years). Most were multiparous (n=17, 65.4%), partnered (n=24, 92.3%), and living in regional (n=9, 34.6%), rural (n=8, 30.8%), or remote (n=9, 34.6%) areas classified by Modified Monash Model (MM) categories (MM2-MM7). Nearly half (n=12, 46.2%) reported a previous mental health diagnosis, with mean EPDS and KMMS scores indicating low levels of current distress. Most participants used social media daily, primarily Facebook and Instagram, with typical usage between 1 and 3 hours per day.

XSL•FO
RenderX

Table . Sample characteristics of perinatal women (n=26) across regional, rural, and remote North Queensland.

Characteristic	Perinatal women
Age (years), mean (SD; range)	32.2 (4.6; 25-48)
Perinatal status, n (%)	
Pregnant	5 (19.2)
Postpartum	21 (80.8)
Parity, n (%)	
Primiparity	9 (34.6)
Multiparity	17 (65.4)
Remoteness[a] (location), n (%)	
Regional (MM2)	9 (34.6)
Rural (MM3-MM5)	8 (30.8)
Remote (MM6-MM7)	9 (34.6)
Area-level socioeconomic disadvantage[b], n (%)	
0% - 20% (least advantaged)	2 (7.7)
20% - 40%	5 (19.2)
40% - 60%	14 (53.8)
60% - 80%	5 (19.2)
80% - 100% (most advantaged)	0 (0)
Past mental health diagnosis, n (%)	12 (46.2)
Current distress, mean (SD)	
EPDS[c]	5.0 (2.6)
KMMS[d]	7.5[e]
Aboriginal and/or Torres Strait Islander, n (%)	2 (7.7)
Country of birth, n (%)	
Australia	19 (73.1)
Other	7 (26.9)
Spoken language, n (%)	
English speaking	4 (57.1)
Non-English speaking	3 (42.9)
Employment, n (%)	
Full time	3 (11.5)
Part time	4 (15.4)
Unemployed	1 (3.8)
On maternity leave	18 (69.2)
Relationship status, n (%)	
Partnered	24 (92.3)
Single	2 (7.7)
Social media use, n (%)	
≤1 h/d	0 (0)
>1 - 3 h/d	17 (65.4)
>3 - 5 h/d	7 (26.9)
>5 h/d	2 (7.7)

XSL•FO
RenderX

Characteristic	Perinatal women
Preferred social media platforms, n (%)[f]	
Facebook (Meta Platforms, Inc)	18 (69.2)
Instagram (Meta Platforms, Inc)	15 (57.7)
YouTube (Alphabet Inc)	7 (26.9)
TikTok (ByteDance Ltd)	6 (23.1)
Snapchat (Snap Inc)	5 (19.2)
Primary reason for social media use, n (%)[f]	
Connection	23 (88.5)
Information seeking	21 (80.8)
Entertainment	21 (80.8)
Sharing updates	18 (69.2)
Connecting with other mothers	14 (53.8)

[a]Modified Monash (MM) model classifications [42]: MM1=major cities; MM2=regional centers (>50,000 population); MM3=large rural towns (15,000 - 50,000 population); MM4=medium rural towns (5000 - 15,000 population); MM5=small rural towns; MM6=remote communities; and MM7=very remote communities.

[b]Socio-Economic Indexes for Areas (SEIFA) quintiles [43] (postcode-derived).

[c]EPDS: Edinburgh Postnatal Depression Scale.

[d]KMMS: Kimberley Mum Mood Scale.

[e]SD value was not reported for KMMS because only 2 participants were included in this analysis.

[f]Totals based on participants' top 3 selections.

Eight mental health care professionals also participated (Table 2). Four engaged in focus groups and 4 in interviews. Participants represented a range of professional backgrounds and had a mean of 16.6 years of clinical experience. They worked across public, private, and community settings in MM2-MM7 regions, and 5 (62.5%) had received specialist perinatal mental health training. Most used social media for less than 1 hour per day, primarily Facebook and Instagram, and reported moderate familiarity with using social media as a psychoeducational tool.

XSL•FO
RenderX

Table . Sample characteristics of perinatal mental health care professionals (n=8) across regional, rural, and remote North Queensland.

Characteristic	Mental health care professionals
Clinical experience (years), mean (SD; range)	16.6 (7.5; 7-27)
Profession, n (%)	
Counselor	1 (12.5)
Social worker	3 (37.5)
Mental health nurse	2 (25)
Psychologist	1 (12.5)
Occupational therapist	1 (12.5)
Specialist perinatal mental health training, n (%)	5 (62.5)
Remoteness[a] (location), n (%)	
Regional (MM2)	5 (62.5)
Rural (MM3-MM5)	1 (12.5)
Remote (MM6-MM7)	2 (25)
Area-level socioeconomic disadvantage[b], n (%)	
0% - 20% (least advantaged)	0 (0)
20% - 40%	2 (25)
40% - 60%	6 (75)
60% - 80%	0 (0)
80% - 100% (most advantaged)	0 (0)
Aboriginal and/or Torres Strait Islander, n (%)	0 (0)
Employment, n (%)	
Nongovernment organization	1 (12.5)
Hospital	1 (12.5)
Community health	3 (37.5)
Private practice	3 (37.5)
Social media use, n (%)	
≤1 h/d	5 (62.5)
>1 - 3 h/d	3 (37.5)
>3 - 5 h/d	0 (0)
>5 h/d	0 (0)
Preferred social media platforms[c], n (%)	
Facebook (Meta Platforms, Inc)	5 (62.5)
Instagram (Meta Platforms, Inc)	4 (50)
Pinterest (Pinterest, Inc)	1 (12.5)
WhatsApp (Meta Platforms, Inc)	1 (12.5)
X (formerly Twitter; X Corp)	1 (12.5)
TikTok (ByteDance Ltd)	1 (12.5)
LinkedIn (Microsoft Corporation)	1 (12.5)
YouTube (Alphabet Inc)	1 (12.5)
Familiarity with using social media as a health tool, n (%)	
Not familiar at all	0 (0)
Slightly familiar	2 (25)

XSL•FO
RenderX

Characteristic	Mental health care professionals
Moderately familiar	5 (62.5)
Very familiar	1 (12.5)
Extremely familiar	0 (0)

[a]Modified Monash (MM) model classifications [42]: MM1=major cities; MM2=regional centers (>50,000 population); MM3=large rural towns (15,000 - 50,000 population); MM4=medium rural towns (5000 - 15,000 population); MM5=small rural towns; MM6=remote communities; and MM7=very remote communities.

[b]Socio-Economic Indexes for Areas (SEIFA) quintiles [43] (postcode-derived).

[c]Totals based on participants' top 3 selections.

Stage 1: Needs Assessment via Reflexive Thematic Analysis

Overview

The following 5 themes were identified reflecting the mental health support needs of perinatal women in regional, rural, and remote areas (Table 3): (1) "Women don't have that proverbial village"—the need for social connection and support, (2) "It can feel like you're a number"—the need for personalized and respectful health care, (3) "Am I doing this right? Is this normal?"—the need for information that empowers, (4) "We're the forgotten North"—the need for place-based support, and (5) "Being able to access something on my phone would be so helpful"—the need for accessible digital forms of support.

XSL•FO
RenderX

Table . Five core mental health needs themes for North Queensland women's perinatal mental health were derived from reflexive thematic analysis across perinatal women and mental health care professionals.

Theme	Definition	Example code
"Women don't have that proverbial village"—The need for social connection and support	The need for authentic, consistent relationships with peers, family, and community that provide emotional validation, support, and a sense of belonging.	• "If you don't have any friends yet, because you don't know anyone and you don't have family around, it can be very isolating." (Regional postpartum mother, 29 years) • "Especially first time mums, if they don't yet have that community of other mums, there's no one to get support from [in terms of] people who are going through the same thing." (Regional mental health occupational therapist, 39 years).
"It can feel like you're a number"—The need for personalized and respectful health care	The need for consistent health care interactions that acknowledge and respect the personal values, needs, and preferences of the mother.	• "Go along and get help and you'll feel better...but if you can't click with them, you're not going to get anything out of it." (Regional pregnant mother, 30 years) • "It's just about having a conversation and unfortunately, as a medical profession, we have become very time poor and we don't have those conversations." (Regional mental health nurse, 48 years)
"Am I doing this right? Is this normal"—The need for information that empowers	The need for easily accessible, accurate information that equips mothers to make informed decisions and feel self-confident in their role as a mother.	• "I definitely think pre and post birth information is really important, so everything from. what to look out for [and] what's considered normal." (Remote pregnant mother, 35 years) • "How can we develop and deliver this so that we can reach women who may not have previously reached out. it's all about bridging these gaps, stopping people from falling through the cracks." (Regional mental health nurse, 48 years)
"We're the forgotten North"—The need for place-based support	The need for services and resources that are culturally inclusive and safe and provide practical support within the context of regional, rural, remote communities.	• "People in North Queensland have less access to things than Brisbane or Melbourne. because of locality and population size...we don't get as much stuff happening or available to us." (Regional postpartum mother, 32 years) • "I bring that real North Queenslander, blunt approach...there are differences in the demographic and Indigenous health isn't even on [Brisbane's] radar." (Regional social worker, 53 years)
"Being able to access something on my phone would be so helpful"—The need for accessible digital forms of support	The need for digital support tools that are readily accessible, facilitate social connection, and provide relevant information.	• "There were definitely gaps in the system...being able to access more information myself would have been really helpful." (Regional postpartum mother, 28 years) • "Everyone uses [social media] so we should definitely infiltrate it with healthcare." (Regional mental health nurse, 48 years)

Theme 1: "Women Don't Have That Proverbial Village"—The Need for Social Connection and Support

Mothers described profound loneliness during pregnancy and postpartum, noting that the perinatal period "[took] a big toll on [their] mental health" (regional postpartum mother, 37 years). Many had relocated to Northern Queensland without established social networks and felt geographically and emotionally disconnected from family and friends. Even where partners were present, work commitments or limited understanding left gaps in support. One mother reflected:

> *I do not have any family other than my partner's family here, so that I think having a village is very important.* [Rural postpartum mother, 38 years]

The need for connection extended beyond family and friends to include peer-to-peer support from other mothers navigating similar experiences. While playgroups and community meetups

XSL•FO
RenderX

were helpful, experiences of accessibility, inclusivity, and emotional safety varied. Participants sought nonjudgmental spaces for genuine connection:

> *I think that's just as important for mental health as mental health services postpartum, where you can actually take your baby and meet other women.* [Regional postpartum mother, 27 years]

Mental health care professionals echoed these concerns, emphasizing the critical role of social connections for well-being: "the isolation is huge" (regional social worker, 38 years).

Theme 2: "It Can Feel Like You're a Number"—The Need for Personalized and Respectful Health Care

Mothers frequently reported depersonalization within health care environments, describing interactions as procedural "box ticking" (remote postpartum mother, 39 years) rather than genuine care, fostering mistrust and disengagement:

> *I needed more from them [the healthcare provider]. I needed my own support network to speak up or to help me find that next level of support.* [Regional postpartum mother, 37 years]

Mothers described feeling pressured to conform to a system that did not reflect their needs or values. Limited access to specialist perinatal mental health care services, inflexible birthing options, and difficulty finding supportive providers left them feeling unseen and disempowered:

> *The need to be treated as an individual…not just have to fit into some formula.* [Rural postpartum mother, 43 years]

Continuity of care, particularly midwife-led models, was viewed as vital for fostering trust and validation. Fragmented care, conflicting advice, and inconsistent provider relationships undermined mothers' confidence:

> *We're relying on these providers who we drop in and out of to give us helpful advice when they've met us for 10 minutes.* [Rural postpartum mother, 39 years]

Mental health care professionals acknowledged the impacts that this depersonalization has on the mothers' well-being and the therapeutic outcomes:

> *A bad experience with mental health services makes them even more reluctant to engage with us.* [Regional mental health nurse, 48 years]

Theme 3: "Am I Doing This Right? Is This Normal?"—The Need for Information That Empowers

Mothers felt unprepared for the emotional and identity shifts of the perinatal period. Challenges such as anxiety, depression, eating disorder symptoms, body dissatisfaction, and birth trauma often went unrecognized until symptoms escalated, with little signposting that it "might [become] a reality" (regional postpartum mother, 25 years). Mental health care professionals acknowledged these challenges, describing the perinatal period as "a real transition" that "requires change" (regional social worker, 66 years). This gap between expectations and reality contributed to feelings of confusion, inadequacy, and shame. First-time mothers described struggling with identity loss in moving from being career-focused individuals to being isolated at home with a newborn:

> *Before you're defined by: you are this person, you work here. Whereas now like, what do I do?* [Rural postpartum mother, 38 years]

Societal and self-imposed expectations intensified uncertainty, with many questioning if their experiences were "considered normal" and what "doing the right thing" (remote pregnant mother, 35 years) looked like. Partners' mental health also influenced maternal well-being, yet support for partners was often lacking:

> *I didn't so much have anxiety, but my partner did…another thing that's not really acknowledged.* [Remote postpartum mother, 29 years]

Theme 4: "We're the Forgotten North"—The Need for Place-Based Support

Mothers perceived their mental health needs as being overlooked compared to metropolitan areas:

> *[Northern Queensland] sometimes does feel like a different world…at least to the Southeast, it really does.* [Rural postpartum mother, 28 years]

Mental health care professionals advocated strongly for place-based solutions, noting:

> *The local aspect of [the program] is so important…what works for one does not work for all communities.* [Regional social worker, 30 years]

Both groups emphasized that geographic isolation, workforce shortages, and cultural and social barriers are unique to regional, rural, and remote areas, requiring tailored support that accounts for the diversity within the region. Mothers described systemic inequities, including long waitlists, unaffordable services, and travel burdens:

> *City folks don't understand, they take for granted the ease of access.* [Remote pregnant mother, 28 years]

Socioeconomic constraints further amplified these barriers:

> *I'm on half pay [whilst on maternity leave]…spending money on psychological services is something that we put on the back burner.* [Rural postpartum mother, 28 years]

Small-town social dynamics also hindered help-seeking, with community expectations to "have everything together" (regional postpartum mother, 28 years) contributing to guilt and shame around mental health needs. Further, culturally and linguistically diverse mothers and First Nations women described perinatal mental health care services as minimal or nonexistent and noted that the dominant Westernized health care model did not always "apply to how [they] think" (regional postpartum mother, 28 years):

> *There's always issues with cultural awareness, sensitivity and safety, whether it's migrant or Indigenous culture.* [Rural postpartum mother, 43 years]

XSL•FO
RenderX

Theme 5: "Being Able to Access Something on My Phone Would Be So Helpful"—The Need for Accessible Digital Forms of Support

Participants strongly supported digital mental health interventions, particularly via social media. Mothers envisioned a centralized, curated, "one stop shop" (regional pregnant mother, 30 years) for trusted resources, peer connection, and tailored information:

> *A social media platform that allows you to have all those resources in one place would be really good…sometimes you don't know where to look.* [Rural postpartum mother, 29 years]

Flexibility, user-controlled content, and moderated peer spaces were considered essential to reduce distress and information overload. Leveraging social media to strengthen in-person community connections was seen as a powerful tool to help prevent perinatal mental health challenges:

> *A small community with some meet ups, some social media—based check ins would be the way to go.* [Remote postpartum mother, 39 years]

The mothers had predominantly used Facebook and Instagram previously to access low-level mental health support, praising these platforms for their ease of use and ability to facilitate a sense of autonomy:

> *I like to go to a place that is updated regularly, like Instagram for its link in bio feature or Facebook for having info consistently pinned at the top of a page.* [Remote pregnant mother, 35 years]

Mental health care professionals acknowledged these potential benefits of social media but noted it can be a "mixed bag" (regional psychologist), and particularly emphasized that it cannot replace the therapeutic relationship:

> *It won't be what supports people fully.* [Regional mental health occupational therapist, 39 years]

Stage 2: Social Media Affordances Analysis

Overview

The second stage of analysis mapped the identified perinatal mental health needs onto the affordances and constraints of social media, following Moreno and D'Angelo [46], to inform the design of a tailored digital intervention. Table 4 summarizes each need, associated affordances, and design guidelines for the prototype's content and functionality.

XSL•FO
RenderX

Table . Mapping identified North Queensland women's perinatal mental health needs to social media affordances and prototype design guidelines for a future intervention.

Identified need	Applicable affordances	Social media affordances	Social media constraints	Prototype design guidelines
Social connection and support	Social and functional	Peer interaction, shared lived experience, and emotional validation via comments and posts; asynchronous communication allows flexibility: "A person that you can send a message to...you know how you [need to] get it off your chest?" (Rural postpartum mother, 28 years) and "Having a way for women that want to be linked to other women in their community would be helpful." (Regional Social Worker, 66 years)	Risk of misinformation or online hostility; possible lack of depth in relationships; may reinforce exclusion: "If it's a page where I'm seeing people getting judged or nasty comments, I'm not going to want to come back to it." (Remote pregnant mother, 35 years) and "What people write in social media comments can be abusive. that goes on as well." (Regional Mental Health Occupational Therapist, 39 years)	Moderated peer chat rooms and closed community groups; include peer facilitators; options to join interest-specific pages (eg, postpartum fitness, recipes, and developmental milestones); localized groups based on geography.
Personalized and respectful health care	Identity, emotional, and functional	Ability to tailor content through algorithms or self-selected topics; opportunities for one-on-one messaging or live Q and As[a]: "You can choose based on your needs who you're going to reach out to or what to engage with." (Regional pregnant mother, 30 years) and "You can choose how much you engage on social media." (Regional mental health occupational therapist, 39 years)	Risk of generic or impersonal content; no replacement for therapeutic relationships: "I felt like what actually helped the most was those [psychologist] sessions." (Rural postpartum mother, 29 years) and "I see that as a potential pitfall...you don't want to get too caught in trying to make it universal." (Regional social worker, 38 years)	User profiles with optional anonymity; personalized content feeds; features to block or filter topics; daily self-reflection prompts and affirmations.
Information that empowers	Social, cognitive, emotional, and functional	Visual and written content (videos, stories, and infographics) for education; rapid dissemination of evidence-based information: "You want things evidence-based from healthcare professionals...from the nurse, the psychologist, the OT, the physio." (Rural postpartum mother, 39 years) and "It can grow your confidence rather than just being told that this is what needs to happen." (Regional mental health occupational therapist, 39 years)	Can promote unrealistic standards; risks reinforcing guilt or shame if not carefully framed: "The negative influence of social media at times when you've got people who are eating salads...why can't I crave a salad?" (Regional pregnant mother, 30 years) and "Other parts of it are not helpful.in terms of unrealistic expectations." (Regional psychologist, 42 years)	"Ask Anything" sections; content series from health care experts; ability for health care experts to track client progress; partner targeted psychoeducation; library of bite-sized evidence-informed resources; quick links to services; use strengths-based, nonjudgmental language; include diverse representations of motherhood experiences.

XSL•FO
RenderX

Identified need	Applicable affordances	Social media affordances	Social media constraints	Prototype design guidelines
Place-based support	Social, cognitive, and functional	Overcomes physical distance; accessible via mobile devices; can highlight regional or cultural content: “If you’re a reader, you might not want to call [a service]...it needs to be culturally and linguistically diverse.” (Regional pregnant mother, 30 years) and “I think one of the strengths is that it can be very focused on local news and local people.” (Regional social worker, 38 years)	Poor internet access in remote areas; dominant cultural narratives often exclude First Nations and migrant voices: “Because of my cultural background, sometimes programs tailored to my local region do not apply to how I think.” (Regional postpartum mother, 28 years) and “We have to take into account the different cultures...make sure we aren’t putting a White Australian English perspective on it.” (Regional mental health occupational therapist, 39 years)	Selectable filters by culture, language, location; localized stories, resources, services, event calendars (eg, playgroups and nurse visits); North Queensland-specific psychoeducation (eg, cyclones and other weather events and travel support); integration of Aboriginal and Torres Strait Islander and culturally and linguistically diverse voices.
Accessible digital formats	Functional	On demand, mobile-first design allows access anytime; no need for in-person appointments: “Telehealth can definitely have its advantages for parents.” (Rural postpartum mother, 39 years) and “I think it’s really good when you have a baby because it can be hard to leave the house.” (Regional psychologist, 42 years)	Digital literacy or affordability may limit access; limited reach to those not on social media: “From a remote perspective, do people have access to computers and smartphones?” (Remote postpartum mother, 29 years) and “People needing a device would be a barrier.” (Regional mental health nurse, 48 years)	Mobile-first, low data design; asynchronous access; integration with free-to-access platforms (eg, Facebook [Meta Platforms, Inc]); inclusive user experience (UX) design for varying digital literacy levels.

[a]Q and A: questions and answers.

Theme 1: Social Connection and Support

Mothers’ desires for social connection align with the social and functional affordances of social media. Peer-to-peer interaction, asynchronous communication, and shared lived experiences can provide emotional validation and flexible engagement. Constraints include the risk of misinformation, online hostility, or shallow relationships. To mitigate these, prototypes could include moderated peer chat rooms, closed community groups with peer facilitators, and interest-specific subgroups (eg, postpartum fitness, local parenting communities):

> *Those lived experience stories can be helpful to know that it’s just a season and it will pass.* [Regional postpartum mother, 35 years]

Theme 2: Personalized and Respectful Health Care

The need to feel seen and respected can be supported by identity, emotional, and functional affordances, such as tailored content through self-selected topics or algorithms and one-on-one messaging or live questions and answers. Constraints include generic content and the inability of social media to replace therapeutic relationships. Recommended features include optional anonymity, personalized content feeds, filters, and daily self-reflection prompts:

> *I don’t know if there’s scope to add a filter, but I think it [needs to be] a realistic, nonjudgemental space.* [Regional pregnant mother, 30 years]

Theme 3: Information That Empowers

Mothers highlighted gaps in psychoeducation and the need for accessible, evidence-based information. Social, cognitive, emotional, and functional affordances allow rapid dissemination via videos, stories, and infographics. Constraints include reinforcing unrealistic expectations or guilt if the content is poorly framed. Design guidelines include bite-sized, strengths-based resources, expert-led content series, partner-targeted psychoeducation, and diverse representations of motherhood:

> *I would love more evidence-based helpful information…[also] to understand how social media can make things worse [rather] than better.* [Rural postpartum mother, 29 years]

Theme 4: Place-Based Support

Social media can address geographic and cultural barriers through social, cognitive, and functional affordances. Online tools allow access without travel or long wait times and can be tailored to local cultural and linguistic needs. Constraints include limited internet access and dominant cultural narratives that may exclude First Nations and migrant voices. Recommended features include localized filters, region-specific resources and events, and culturally inclusive content:

> *It has to look like it’s something for me, something for North Queensland…with the pictures and the way it looks.* [Regional postpartum mother, 25 years]

XSL•FO
RenderX

Theme 5: Accessible Digital Formats

The need for flexible, on-demand support maps onto functional affordances, such as asynchronous access, mobile-first design, and replicable content. Constraints include varying digital literacy and affordability. To address this, prototypes should prioritize low-data, mobile-first design, integration with free-to-access platforms, and inclusive user experience design:

> *The more barriers you put in front of people, the less likely people are to engage…it needs to be direct links.* [Regional postpartum mother, 42 years]

Discussion

Overview

This study sought to co-design a universal social media–based preventive intervention for perinatal mental illness among women living in regional, rural, and remote Northern Queensland. We identified 5 core mental health needs: social connection and support, personalized and respectful health care, information that empowers, place-based support, and accessible digital formats. Social media shows promise for addressing these needs through peer support, psychoeducation, and culturally responsive content, yet constraints such as misinformation, limited digital literacy, and inequitable access must be addressed in program design. Our findings demonstrate how digital platforms can be leveraged to deliver preventive mental health support in regional, rural, and remote communities, with implications for clinical practice, service design, and public health policy.

Principal Findings

The first 3 themes mirror the broader perinatal experiences across diverse contexts. Previous literature indicates that mothers value connection and peer support during this period, which is often marked by loneliness [48]. The need for personal and respectful interactions that are clear, evidence-based, and empowering within health care is frequently cited as crucial to perinatal mental health, yet lacking within available supports [48]. Our study extends these findings by demonstrating that these needs are magnified in regional, rural, and remote communities, where social and systemic inequities amplify their salience. While these communities are resilient, this does not deem them exempt from requiring mental health support [9]. Acknowledging and addressing these needs is essential to reducing the compounding effects of inequity and mitigating intergenerational impacts of poor perinatal mental health [4].

Alongside these broader needs, our study highlights the unique systemic and social barriers faced by women in regional, rural, and remote regions. Unlike metropolitan centers, these regions experience workforce shortages that create longer wait times and a lack of individualized care, exacerbated by geographic isolation that has considerable financial and travel-related burdens. The unique social dynamics of smaller communities, characterized by shame, stigma, and a lack of anonymity, created barriers to seeking support beyond that of service availability. Our findings highlight the need for place-based approaches to appropriately address systemic and social inequities and cultural and contextual differences within these communities [11].

Further, our findings suggest that a well-designed social media intervention may help to mitigate some of these challenges. The mothers were supportive of a social media intervention due to its capacity to provide accessible information and act as a “digital village” to foster peer support. This is consistent with previous literature demonstrating that social media is an important source of health information for perinatal women [20]. However, social media cannot address all barriers, including limited digital literacy and structural inequities that remain prominent in these regions. Previous literature demonstrates how social media can perpetuate unhealthy comparison, identity distress, and parenting anxiety, negatively influencing mental health [23]. These concerns were shared by both mothers and mental health care professionals in this study and must be addressed in both the design and functionality of digital platforms. The needs-affordances analysis allowed us to translate these insights into actionable, relevant, and feasible design guidelines. It emphasized the need for a place-based approach that fosters social support while mitigating harmful comparison through moderation. While previous digital health interventions have reduced symptoms of anxiety and depression in perinatal women, the use of social media as a preventive mental health tool remains novel [49]. This research demonstrates that a social media health intervention is a feasible and acceptable platform for delivering preventive, accessible psychoeducational mental health support to perinatal women at a population level.

Comparison With Prior Work

The needs and design guidelines identified in this study align with prior interventions but add nuance that addresses key gaps. Most previous interventions, including those that were effective in reducing depression and anxiety, incorporated peer connection via chat groups, peer worker support, or forums, highlighting the value of social support [27,30]. However, these interventions were predominantly psychoeducation-based, delivered uniform content, and rarely allowed users to self-direct learning and engagement. Only Hatamleh et al [32] explicitly considered diverse cultural needs, leaving most programs limited in their capacity to engage culturally diverse or regionally isolated mothers [32]. Importantly, none provided options for users to filter or select topics of interest. This was a feature that the mothers in this study identified as key to fostering meaningful engagement and supporting mental health outcomes. Our guidelines directly address these limitations by emphasizing a user-led, autonomous approach grounded in place-based content and design. This approach not only increases the likelihood of engagement and relevance to mothers in regional, rural, and remote communities, but also mitigates the risk of inadvertently excluding or disadvantaging mothers from diverse cultural and linguistic backgrounds. In doing so, these findings extend prior work by illustrating how social media interventions might be tailored to a regional, rural, and remote context to maximize both efficacy and equity, providing actionable guidance for future mental health programs.

The social media intervention could integrate into existing health care systems in a way that alleviates, rather than duplicates, existing burdens. Mental health care services remain predominantly reactive and treatment-oriented, constrained by

XSL•FO
RenderX

workforce shortages, increasing health demands, and higher health care costs [50]. In the context of the continuum of care, this program could provide preventive, universal support for perinatal mothers [51]. The potential for a social media intervention to complement existing services was highlighted by the mental health care professionals in this study, which would be particularly beneficial for both their practice and client outcomes. For those experiencing mild distress, peer support and psychoeducation may be sufficient to prevent symptom escalation. The program can also provide a mechanism for mental health screening and referral to formalized supports, complementing existing services and potentially decreasing the need for more intensive services downstream.

Implications for Policy, Practice, and Future Research

The findings of this study challenge the assumption that digital mental health interventions developed in metropolitan or international contexts can be straightforwardly applied to regional, rural, and remote settings. The mental health needs identified in this study were experienced with a distinctly place-based character, shaped by geographic isolation, workforce shortages, small community social dynamics, and systemic inequity, which generic interventions are unlikely to adequately address. Place-based co-design is, therefore, not merely a methodological preference but a practical necessity for intervention relevance and equity in these communities.

For clinicians and health care services, a well-designed social media intervention holds potential to complement existing perinatal mental health care by providing psychoeducation, peer connection, and normalization of help-seeking at a population level, alongside a mechanism for referral to formalized supports where needed. For policymakers and funders, these findings make a case for sustained investment in co-designed digital mental health infrastructure, including the structural enablers of equitable access such as reliable connectivity, device affordability, and digital literacy support.

More broadly, this study demonstrates that social media is a viable and acceptable intervention platform for adult perinatal women—a population largely absent from the existing evidence base for social media interventions [52]. Our findings highlight the potential for social media to function as an equalizer of access to mental health support, offering an engaging, low-cost, geographically unconstrained, and stigma-reducing touchpoint in settings where structural barriers would otherwise prevent timely care. However, the findings also emphasize that social media is not a neutral delivery channel for mental health support. Realizing this potential requires affordance analysis to be adopted as a standard step in social media intervention development, ensuring platform selection and design are driven by user need.

The next phase of this research is a pilot study involving the development and evaluation of the universal preventive intervention informed by the prototype design guidelines generated here. This pilot phase will use a feasibility and acceptability design, examining whether the intervention can be implemented as intended in regional, rural, and remote communities, and generating preliminary data on outcomes including mental health and well-being, social connectedness, and help-seeking behavior, to inform the design of a subsequent fully powered effectiveness trial. Future studies should also explore the potential for predictive and personalized digital tools, including adaptive algorithms and artificial intelligence, to anticipate psychological distress and triage mothers into appropriate care [53,54].

Reflection on Patient and Public Involvement

The co-design process generated rich, contextually grounded data and was strengthened by the active involvement of a community advisory group comprising lived experience representatives, clinicians, and local community leaders, who provided ongoing input throughout the research process. The dual online and in-person format was effective in expanding accessibility across a geographically dispersed region, and participants across both formats engaged substantively with the focus group and interview questions, generating detailed and meaningful contributions to the design process.

Notably, the co-design process substantially shaped the direction of the intervention in ways that were not anticipated at the outset. The original intervention concept envisaged a broadcast-style social media presence through which preventive psychotherapeutic content would be disseminated to a largely anonymous audience (eg, a public Instagram or Facebook page). However, the needs-affordances analysis revealed that peer interaction, mutual support, and personalized connection were central priorities for participants, fundamentally reorienting the intervention toward a more interactive and community-driven model. This shift reflects the genuine influence of patient and public involvement on the research outcomes and underscores the value of co-design approaches for intervention development.

However, several challenges were also encountered. Higher-than-anticipated dropout was observed among mothers attending in-person sessions, with cancellations frequently occurring at short notice, consistent with the practical realities of caring for a newborn or infant, including feeding schedules, sleep deprivation, and unpredictable child health. Future co-design research with perinatal populations should anticipate this from the outset, for example, by offering shorter or more flexible session formats or providing childcare support where feasible.

Strengths and Limitations

This study's strengths lie largely in its co-design approach, which provided contextually grounded insights supported by lived and clinical experience. Using face-to-face and online methods across diverse regions enhanced depth and representation within the data. The inclusion of a community advisory group with lived experience of perinatal depression and anxiety, alongside a sample in which nearly half of participants reported a prior mental health diagnosis, further strengthened the relevance and authenticity of the findings.

However, several limitations should be noted. First, the sample underrepresented First Nations and LGBTQIA+ (lesbian, gay, bisexual, transgender, queer (or questioning), intersex, and asexual (or agender), and other diverse sexual orientations, gender identities, and sex characteristics) voices, with only 2 participants identifying as Aboriginal and/or Torres Strait

XSL•FO
RenderX

Islander despite recruitment through targeted social media advertising, posters in libraries, health care services, community spaces, and media releases. Notably, a number of Aboriginal and Torres Strait Islander mothers expressed interest in participating but were disproportionately screened out at eligibility due to reporting moderate-to-severe symptoms on the KMMS, reflecting the well-documented higher rates of perinatal distress among First Nations mothers in Australia [8,55], and representing an important intersection between our eligibility criteria and First Nations representation. This also highlights a broader tension inherent to universal program co-design: although the intervention is intended to be broadly accessible across diverse communities, the co-design process itself may not have equally captured the perspectives of all subgroups within the target population. First Nations perspectives were nonetheless present at the level of research governance, with Aboriginal and Torres Strait Islander members contributing to both the chief investigator group and community advisory group throughout the study. Future research should prioritize community-embedded recruitment developed in genuine partnership with Aboriginal Community Controlled Health Organizations, employ First Nations researchers and community liaisons, and consider whether supplementary targeted co-design processes are needed to ensure the intervention is culturally safe and relevant for First Nations and LGBTQIA+ communities specifically [56,57].

Second, the exclusion of women currently experiencing moderate-to-severe mental health symptoms or suicidal ideation limits the study's capacity to capture insights from those at highest risk. This exclusion was intentional and reflects the preventive rather than clinical focus of the intervention; the co-design sample was deliberately aligned with the intended end-user population of the program, which is explicitly preventive and psychoeducational in nature rather than a clinical or therapeutic tool. Accordingly, the findings and design guidelines generated by this study should not be generalized to clinical programs aiming to treat moderate-to-severe perinatal mental health presentations; such programs serve a distinct population with distinct needs and would require their own dedicated co-design processes to determine what women in those contexts want and need from a social media-based intervention. Nevertheless, we acknowledge that women who have experienced symptom escalation may have offered valuable perspectives on what earlier digital support could have prevented. This limitation is partially mitigated by the fact that 46% of participants reported a prior diagnosis of depression and/or anxiety, and by the inclusion of lived experience representatives of perinatal depression and/or anxiety on the advisory group. Future work could consider engaging women with current or recent clinical presentations as a distinct consultative group to complement the perspectives captured here.

Third, the eligibility criterion requiring prior social media experience was included to ensure participants could engage meaningfully with the affordances analysis. While this criterion could theoretically filter out the most digitally excluded members of the target population, a tension worth acknowledging given that digital literacy and affordability emerged as barriers in our findings (Theme 5), no participants were excluded on this basis in practice. This reflects the near ubiquity of social media use even among populations with limited digital literacy or constrained financial resources. Recruitment through both online and offline channels further reduced the likelihood that digitally excluded women were unable to learn about or express interest in the study. Nonetheless, future evaluation phases should monitor whether the intervention is equally accessible and acceptable to women with lower levels of digital literacy to ensure equitable reach.

Fourth, social media affordances are not static and may shift over time as platform designs, algorithms, community norms, privacy policies, and access conditions evolve. The design principles generated by this study reflect the affordances landscape at the time of data collection, and future intervention development and evaluation phases will need to remain responsive to changes in the social media environment to ensure continued relevance and safety.

Fifth, recruitment via convenience sampling and snowball methods, including social media advertising and participant referral networks, may have preferentially engaged women who are already digitally active and connected to community networks. Women with lower digital literacy, limited internet connectivity, or reduced engagement with community organizations, including those in the most geographically remote parts of Northern Queensland, may therefore be underrepresented. Future co-design and implementation research should incorporate targeted, community-embedded recruitment strategies to reach these subgroups.

Conclusion

This study aimed to identify the mental health needs of perinatal women in regional, rural, and remote communities and to co-design a framework for a preventive social media–based intervention informed by platform-specific affordances and constraints. Using a participatory co-design approach with 26 perinatal women and 8 mental health care professionals in Northern Queensland, Australia, 5 core mental health needs were identified: social connection and support, personalized and respectful health care, information that empowers, place-based and culturally safe support, and accessible low-burden digital formats. A needs-affordances analysis demonstrated that social media holds promise for addressing these needs through peer support, psychoeducation, normalization, and timely information delivery, provided that known constraints are proactively addressed in program design, including misinformation, privacy risks, harmful social comparison, and inequitable digital access. These findings were synthesized into prototype design guidelines to inform subsequent phases of intervention development. Taken together, these findings address a significant gap in the literature and offer a replicable methodological approach for co-designing digital mental health interventions with underserved perinatal populations.

XSL•FO
RenderX

Acknowledgments

We would like to thank the members of the Advisory Group for their guidance and feedback throughout this study. Academic and clinical members who contributed expertise were Suzanne Riach, Cassie Kemp, Pieta Shakes, Renee McCallister, and Nicki Walsh. Members with lived experience or representing community organizations are acknowledged collectively for their invaluable input, which helped ensure the study was relevant, feasible, and culturally sensitive. We also thank all participants who generously shared their time, knowledge, and experience, which ultimately shaped this project. Finally, we thank the organizations that provided spaces for data collection—Townsville Bebegu Yumba Campus, Mt Isa Centre for Rural and Remote Health, Cairns Nguma-Bada Campus, and the Mackay Mater Clinical Training Centre—for supporting meaningful conversations with participants. No generative artificial intelligence tools were used in the conception, design, analysis, or writing of this paper.

Funding

This research was supported by a Tropical Australian Academic Health Centre (TAAHC) Seed Grant. Author ST is supported by an Australian National Health and Medical Research Council (NHMRC) Investigator Grant (APP2025839), and author DH is supported by an NHMRC Investigator Grant and a Medical Research Future Fund Grant (APP1197488 and MRF2044506, respectively). The funders had no role in the design or conduct of the study; the collection, management, analysis, or interpretation of the data; the preparation, review, or approval of the paper; or the decision to submit the paper for publication.

Data Availability

The datasets generated and/or analyzed during this study are not publicly available due to participant privacy and ethical restrictions. Deidentified data may be made available from the corresponding author upon reasonable request and subject to institutional ethics approval.

Authors' Contributions

Conceptualization was undertaken by ST, KJL, ABRS, DH, JD, JM, KG, GR, AA, and CH. Data curation was conducted by KJL and ST. Formal analysis was performed by KJL, ST, ABRS, DH, JD, JM, KG, GR, AA, and CH. Funding acquisition was led by ST, ABRS, DH, JD, JM, KG, GR, and AA. Investigation was conducted by KJL, ST, JM, ABRS, and DH. Methodology was developed by ST, KJL, ABRS, DH, JD, JM, KG, GR, AA, and CH. Project administration and supervision were undertaken by ST. Resources, validation, and visualization were contributed by KJL and ST. The original draft was prepared by KJL and ST.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Focus group or interview session outline.
[DOCX File, 18 KB - jmir_v28i1e91778_app1.docx]

Checklist 1
GRIPP2 checklist.
[DOCX File, 15 KB - jmir_v28i1e91778_app2.docx]

References

1. Howard LM, Khalifeh H. Perinatal mental health: a review of progress and challenges. World Psychiatry 2020 Oct;19(3):313-327. [doi: 10.1002/wps.20769] [Medline: 32931106]
2. Law S, Ormel I, Babinski S, et al. Dread and solace: talking about perinatal mental health. Int J Mental Health Nurs 2021 Oct;30(S1):1376-1385. [doi: 10.1111/inm.12884]
3. Khajehei M. Mental health of perinatal women. WJOG 2015;4(2):46-51. [doi: 10.5317/wjog.v4.i2.46]
4. Rogers A, Obst S, Teague SJ, et al. Association between maternal perinatal depression and anxiety and child and adolescent development: a meta-analysis. JAMA Pediatr 2020 Nov 1;174(11):1082-1092. [doi: 10.1001/jamapediatrics.2020.2910] [Medline: 32926075]
5. The Australian health system. Australian Government Department of Health, Disability and Ageing. 2026. URL: https://www.health.gov.au/topics/about-the-department/australian-health-system [accessed 2026-05-18]
6. Annual report 2024-2025. : Northern Queensland Primary Health Network; 2025 URL: https://annualreport2025.nqphn.com.au/ [accessed 2026-05-18]
7. Galbally M, Watson SJ, Coleman M, et al. Rurality as a predictor of perinatal mental health and well‐being in an Australian cohort. Australian J Rural Health 2023 Apr;31(2):182-195. [doi: 10.1111/ajr.12934]
8. Fox H, Topp SM, Lindsay D, Callander E. Ethnic, socio‐economic and geographic inequities in maternal health service coverage in Australia. Health Plan Manag 2021 Nov;36(6):2182-2198. [doi: 10.1002/hpm.3277]

XSL•FO
RenderX

9. Kavanagh BE, Corney KB, Beks H, Williams LJ, Quirk SE, Versace VL. A scoping review of the barriers and facilitators to accessing and utilising mental health services across regional, rural, and remote Australia. BMC Health Serv Res 2023 Oct 4;23(1):1060. [doi: 10.1186/s12913-023-10034-4] [Medline: 37794469]
10. Arango C, Díaz-Caneja CM, McGorry PD, et al. Preventive strategies for mental health. Lancet Psychiatry 2018 Jul;5(7):591-604. [doi: 10.1016/S2215-0366(18)30057-9]
11. Klepac B, Mowle A, Riley T, Craike M. Government, governance, and place-based approaches: lessons from and for public policy. Health Res Policy Sys 2023 Nov 29;21(1):126. [doi: 10.1186/s12961-023-01074-7]
12. Fitzgerald L, McNab S, Njau P, et al. Beyond survival: prioritizing the unmet mental health needs of pregnant and postpartum women and their caregivers. PLOS Glob Public Health 2024;4(2):e0002782. [doi: 10.1371/journal.pgph.0002782] [Medline: 38315641]
13. Patel V, Saxena S, Lund C, et al. The Lancet Commission on global mental health and sustainable development. Lancet 2018 Oct 27;392(10157):1553-1598. [doi: 10.1016/S0140-6736(18)31612-X] [Medline: 30314863]
14. O'Connor E, Senger CA, Henninger ML, Coppola E, Gaynes BN. Interventions to prevent perinatal depression: evidence report and systematic review for the US Preventive Services Task Force. JAMA 2019 Feb 12;321(6):588-601. [doi: 10.1001/jama.2018.20865] [Medline: 30747970]
15. Lewis Johnson TE, Clare CA, Johnson JE, Simon MA. Preventing perinatal depression now: a call to action. J Womens Health (Larchmt) 2020 Sep;29(9):1143-1147. [doi: 10.1089/jwh.2020.8646] [Medline: 32749917]
16. Perkins D, Farmer J, Salvador‐Carulla L, Dalton H, Luscombe G. The Orange Declaration on rural and remote mental health. Australian J Rural Health 2019 Oct;27(5):374-379. [doi: 10.1111/ajr.12560]
17. Petkovic J, Duench S, Trawin J, et al. Behavioural interventions delivered through interactive social media for health behaviour change, health outcomes, and health equity in the adult population. Cochrane Database Syst Rev 2021 May 31;5(5):CD012932. [doi: 10.1002/14651858.CD012932.pub2] [Medline: 34057201]
18. Kemp S. Digital 2021: Australia. Data Reportal. 2021. URL: https://datareportal.com/reports/digital-2021-australia [accessed 2026-05-29]
19. Australia's mothers and babies. Australian Institute of Health and Welfare. 2025. URL: https://www.aihw.gov.au/reports/mothers-babies/australias-mothers-babies/contents/overview-and-demographics/maternal-age [accessed 2025-03-15]
20. Gow ML, Henderson M, Henry A, Roberts L, Roth H. Australian postpartum women want reputable health information delivered via social networking sites. Aust NZ J Obst Gynaeco 2024 Dec;64(6):643-646. [doi: 10.1111/ajo.13840]
21. Li B, Jin N, Wang Y, Hou X, Meng J, Zhang Y. Perinatal women's perception of maternal health information quality on digital media: scoping review. J Med Internet Res 2025 Jul 2;27:e67620. [doi: 10.2196/67620] [Medline: 40600605]
22. Tate MK. The impact of social comparison via social media on maternal mental health, within the context of the intensive mothering ideology: a scoping review of the literature. Issues Ment Health Nurs 2023 Sep;44(9):854-870. [doi: 10.1080/01612840.2023.2238813] [Medline: 37713723]
23. Smith M, Mitchell AS, Townsend ML, Herbert JS. The relationship between digital media use during pregnancy, maternal psychological wellbeing, and maternal-fetal attachment. PLoS One 2020;15(12):e0243898. [doi: 10.1371/journal.pone.0243898] [Medline: 33326465]
24. Li SH, Achilles MR, Spanos S, Habak S, Werner-Seidler A, O'Dea B. A cognitive behavioural therapy smartphone app for adolescent depression and anxiety: co-design of ClearlyMe. tCBT 2022;15:e13. [doi: 10.1017/S1754470X22000095]
25. Ansaari N, Rajan SK, Kuruveettissery S. Efficacy of in-person versus digital mental health interventions for postpartum depression: meta-analysis of randomized controlled trials. J Reprod Infant Psychol 2025 Jul;43(4):1067-1087. [doi: 10.1080/02646838.2024.2303470] [Medline: 38221725]
26. Xie H, Cong S, Wang R, et al. Effect of eHealth interventions on perinatal depression: a meta-analysis. J Affect Disord 2024 Jun 1;354:160-172. [doi: 10.1016/j.jad.2024.03.027] [Medline: 38490593]
27. Boyd RC, Price J, Mogul M, Yates T, Guevara JP. Pilot RCT of a social media parenting intervention for postpartum mothers with depression symptoms. J Reprod Infant Psychol 2019 Jul;37(3):290-301. [doi: 10.1080/02646838.2018.1556788] [Medline: 30556428]
28. Yang M, Jia G, Sun S, Ye C, Zhang R, Yu X. Effects of an online mindfulness intervention focusing on attention monitoring and acceptance in pregnant women: a randomized controlled trial. J Midwife Womens Health 2019 Jan;64(1):68-77. [doi: 10.1111/jmwh.12944]
29. Kernot J, Lewis L, Olds T, Maher C. Effectiveness of a Facebook-delivered physical activity intervention for postpartum women: a randomized controlled trial. J Phys Act Health 2019 Feb 1;16(2):125-133. [doi: 10.1123/jpah.2017-0573] [Medline: 30541378]
30. Shorey S, Chee CYI, Ng ED, Lau Y, Dennis CL, Chan YH. Evaluation of a technology-based peer-support intervention program for preventing postnatal depression (part 1): randomized controlled trial. J Med Internet Res 2019 Aug 29;21(8):e12410. [doi: 10.2196/12410] [Medline: 31469084]
31. Zhang X, Lin P, Sun J, et al. Prenatal stress self-help mindfulness intervention via social media: a randomized controlled trial. J Ment Health 2023 Feb;32(1):206-215. [doi: 10.1080/09638237.2021.1952947] [Medline: 34264775]

XSL•FO
RenderX

32. Hatamleh R, AbdelMahdi AbuAbed AS, Abujilban S, Joseph R. Using social media platforms to enhance the delivery of a childbirth education program. J Perinat Neonatal Nurs 2023;37(1):36-43. [doi: 10.1097/JPN.0000000000000700] [Medline: 36707746]
33. Guevara JP, Morales K, Mandell D, et al. Social media-based parenting program for women with postpartum depressive symptoms: an RCT. Pediatrics 2023 Mar 1;151(3):e2022058719. [doi: 10.1542/peds.2022-058719] [Medline: 36808207]
34. The Double Diamond. Design Council. 2005. URL: https://www.designcouncil.org.uk/resources/the-double-diamond/ [accessed 2026-05-30]
35. Staniszewska S, Brett J, Simera I, et al. GRIPP2 reporting checklists: tools to improve reporting of patient and public involvement in research. BMJ 2017 Aug 2;358:j3453. [doi: 10.1136/bmj.j3453] [Medline: 28768629]
36. Annual report 2024-2025. : Western Queensland Primary Health Network; 2025 URL: https://cdn.prod.website-files.com/6498e7a6f8c93e238042358e/69265b6c903c1183f3e4b33e_WQPHN%20FY25%20Annual%20report%20FINAL%20DIGITAL%20SINGLES.pdf [accessed 2026-05-18]
37. Cox JL, Holden JM, Sagovsky R. Detection of postnatal depression. Development of the 10-item Edinburgh Postnatal Depression Scale. Br J Psychiatry 1987 Jun;150(6):782-786. [doi: 10.1192/bjp.150.6.782] [Medline: 3651732]
38. Carlin E, Ferrari K, Spry EP, Williams M, Atkinson D, Marley JV. Implementation of the 'Kimberley Mum's Mood Scale' across primary health care services in the Kimberley region of Western Australia: a mixed methods assessment. PLoS One 2022;17(9):e0273689. [doi: 10.1371/journal.pone.0273689] [Medline: 36054104]
39. San Martin Porter MA, Betts K, Kisely S, Pecoraro G, Alati R. Screening for perinatal depression and predictors of underscreening: findings of the Born in Queensland study. Med J Aust 2019 Jan;210(1):32-37. [doi: 10.5694/mja2.12030] [Medline: 30636310]
40. Queensland clinical guidelines. Queensland Health. 2024. URL: http://www.health.qld.gov.au/qcg [accessed 2025-05-18]
41. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
42. Modified Monash Model. Australian Government Department of Health and Aged Care. 2025. URL: https://www.health.gov.au/topics/rural-health-workforce/classifications/mmm [accessed 2026-05-29]
43. 2033.0.55.001 - Census of Population and Housing: Socio-Economic Indexes for Areas (SEIFA), Australia, 2011. : Australian Bureau of Statistics; 2011 URL: https://www.abs.gov.au/ausstats/abs@.nsf/DetailsPage/2033.0.55.0012011 [accessed 2026-05-29]
44. Rampin R, Rampin V. Taguette: open-source qualitative data analysis. JOSS 2021 Dec 10;6(68):3522. [doi: 10.21105/joss.03522]
45. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Aug 8;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
46. Moreno MA, D'Angelo J. Social media intervention design: applying an affordances framework. J Med Internet Res 2019 Mar 26;21(3):e11014. [doi: 10.2196/11014] [Medline: 30912754]
47. Braun V, Clarke V. Thematic Analysis: A Practical Guide: SAGE Publications; 2021.
48. Schwartz H, McCusker J, Law S, Zelkowitz P, Somera J, Singh S. Perinatal mental healthcare needs among women at a community hospital. J Obstet Gynaecol Can 2021 Mar;43(3):322-328. [doi: 10.1016/j.jogc.2020.08.015] [Medline: 33446473]
49. Chan KL, Chen M. Effects of social media and mobile health apps on pregnancy care: meta-analysis. JMIR Mhealth Uhealth 2019 Jan 30;7(1):e11836. [doi: 10.2196/11836] [Medline: 30698533]
50. Dixit SK, Sambasivan M. A review of the Australian healthcare system: a policy perspective. SAGE Open Med 2018;6:2050312118769211. [doi: 10.1177/2050312118769211] [Medline: 29686869]
51. Baker E, August K, Mangra N, et al. The impact of a continuum of care on social determinants of health and health disparities: an IEEE perspective. THMT 2025 May 28;10(1). [doi: 10.30953/thmt.v10.557]
52. Zhang Q, Huang Z, Sui Y, et al. Social-media-based mental health interventions: meta-analysis of randomized controlled trials. J Med Internet Res 2025 Aug 14;27:e67953. [doi: 10.2196/67953] [Medline: 40811685]
53. Barrett M, Boyne J, Brandts J, et al. Artificial intelligence supported patient self-care in chronic heart failure: a paradigm shift from reactive to predictive, preventive and personalised care. EPMA J 2019 Dec;10(4):445-464. [doi: 10.1007/s13167-019-00188-9] [Medline: 31832118]
54. Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. Psychol Med 2019 Jul;49(9):1426-1448. [doi: 10.1017/S0033291719000151] [Medline: 30744717]
55. Owais S, Faltyn M, Johnson AVD, et al. The perinatal mental health of indigenous women: a systematic review and meta-analysis. Can J Psychiatry 2020 Mar;65(3):149-163. [doi: 10.1177/0706743719877029] [Medline: 31601125]
56. Escobar-Viera CG, Melcher EM, Miller RS, et al. A systematic review of the engagement with social media-delivered interventions for improving health outcomes among sexual and gender minorities. Internet Interv 2021 Sep;25:100428. [doi: 10.1016/j.invent.2021.100428] [Medline: 34401387]
57. Reilly R, Stephens J, Micklem J, et al. Use and uptake of web-based therapeutic interventions amongst Indigenous populations in Australia, New Zealand, the United States of America and Canada: a scoping review. Syst Rev 2020 May 31;9(1):123. [doi: 10.1186/s13643-020-01374-x] [Medline: 32475342]

XSL•FO
RenderX

Abbreviations

EPDS: Edinburgh Postnatal Depression Scale
GRIPP2: Guidance for Reporting Involvement of Patients and the Public
KMMS: Kimberley Mum Mood Scale
LGBTQIA+: lesbian, gay, bisexual, transgender, queer (or questioning), intersex, and asexual (or agender), and other diverse sexual orientations, gender identities, and sex characteristics

Edited by A Coristine; submitted 19.Jan.2026; peer-reviewed by K Baggett, M Gonzalez; revised version received 24.Apr.2026; accepted 24.Apr.2026; published 10.Jun.2026.

Please cite as:
Lynch KJ, Shatte ABR, Muller J, George K, Rossini G, Anson A, Hala C, Dimmock J, Hutchinson D, Teague S
A Preventive Social Media Intervention for Perinatal Depression and Anxiety in Regional, Rural, and Remote Communities: Participatory Co-Design Study
J Med Internet Res 2026;28:e91778
URL: https://www.jmir.org/2026/1/e91778
doi:10.2196/91778

© Kacey Jane Lynch, Adrian Brian Royce Shatte, Jessica Muller, Kendall George, Gisele Rossini, Angela Anson, Courtney Hala, James Dimmock, Delyse Hutchinson, Samantha Teague. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Optimization of University Counseling Consent Forms With Large Language Models: Multidimensional Comparative Evaluation

Jianchen Luo[1,2*], MD; Jing Ma[3*], MM; Danni Zhan[4], MEd; Yuhong Zhou[5], PhD; Jiayu Li[6], PhD; Lan Zhang[2,3], MD; Wentao Wang[1,2], MD, PhD

[1]Department of Liver Surgery, West China Hospital of Sichuan University, 37 Guoxue Alley, Wuhou District, Chengdu, China
[2]West China School of Clinical Medicine, Sichuan University, Chengdu, China
[3]Mental Health Center, West China Hospital of Sichuan University, Chengdu, China
[4]Department of Vocational Education, The Open University of Mianyang, Mianyang, China
[5]Psychological Research and Counseling Center, Southwest Jiaotong University, Chengdu, China
[6]Department of Psychology, Hebei Normal University, Shijiazhuang, China
[*]these authors contributed equally

Corresponding Author:
Wentao Wang, MD, PhD
Department of Liver Surgery, West China Hospital of Sichuan University, 37 Guoxue Alley, Wuhou District, Chengdu, China

Abstract

Background: Mental health problems among university students are a growing global concern, yet limited counseling resources and inadequate understanding of counseling procedures often delay timely help-seeking. Informed consent forms (ICFs) are essential for safeguarding autonomy and clarifying counseling procedures, but many universities' counseling ICFs are incomplete, ambiguous, or overly technical. Large language models (LLMs) may offer scalable assistance for improving clarity and accessibility.

Objective: This study aimed to evaluate whether LLM-based rewriting could improve the structure, readability, content quality, and comprehensibility of university counseling ICFs, and compared 2 advanced models (ChatGPT [GPT-5] and Grok-4).

Methods: We conducted a comparative evaluation of counseling ICFs collected from 33 Chinese universities (original texts) and generated 2 rewritten versions for each ICF using ChatGPT (GPT-5) and Grok-4. A multidimensional framework assessed (1) textual structure and readability, (2) expert-rated content quality from a counselor perspective, and (3) volunteer-rated reading comprehension from a client perspective. Comparisons between original and rewritten texts were performed using Wilcoxon signed rank tests, with linear mixed-effects models used to validate results while accounting for rater variability.

Results: Compared with the originals, both LLM-rewritten ICFs showed significant improvements across all evaluated dimensions. The mean Lee-Yang Readability Index decreased from 28.68 (SD 5.69) to 22.39 (SD 2.13) with ChatGPT (GPT-5) and 24.37 (SD 2.32) with Grok-4 (both $P<.001$), and mean tone friendliness increased from 2.57 (SD 0.29) to 2.67 (SD 0.12) and 2.67 (SD 0.13), respectively. The mean expert-rated content quality improved from 45.33 (SD 8.74) to 52.54 (SD 7.92) and 55.49 (SD 7.81) ($P<.001$), driven mainly by higher completeness and specificity of key information. The mean volunteer-rated reading comprehension scores increased from 19.02 (SD 1.32) to 22.33 (SD 0.81) and 22.05 (SD 0.90) ($P<.001$), indicating improved clarity, readability, and acceptability. Across structural features, Grok-4 tended to produce longer rewritten forms than the originals, highlighting a potential trade-off between added informational content and document length.

Conclusions: In this comparative evaluation of 33 Chinese university counseling ICFs, LLM-based rewriting was associated with improved readability, expert-rated content quality, and volunteer-rated comprehension relative to original forms. These findings suggest that LLMs can support the optimization of counseling documentation; however, implementation should consider practical constraints (eg, document length) and retain human oversight.

(J Med Internet Res 2026;28:e86502) doi:10.2196/86502

KEYWORDS

large language models; informed consent forms; university counseling; mental health accessibility; higher education

Introduction

Mental health problems among adolescents and young adults represent a growing global public health concern, particularly in the context of higher education [1]. Estimates suggest that up to 50% of college freshmen worldwide experience at least 1 mental health problem during their time at university [2]. With the rising prevalence of mental health challenges in this

XSL•FO
RenderX

population, psychological counseling services have become an indispensable component of campus mental health systems [3]. As the initial step in delivering such services, informed consent forms (ICFs) not only embody fundamental ethical and legal principles but also serve as a critical safeguard for students' right to information and autonomy. Nevertheless, substantial variability exists in the structure, content, and language of ICFs across universities [4]: some are excessively lengthy and filled with technical jargon, making them difficult for students to understand, while others are incomplete or ambiguously worded, undermining their protective function [5,6].

The clarity and standardization of ICFs are directly linked to recipients' levels of understanding and the quality of their decision-making [5]. Although relevant efforts have been made in the medical and mental health fields, systematic research specifically targeting ICFs for university counseling services remains scarce [7]. At the institutional level in higher education, many universities lack the capacity to continuously refine these forms due to constraints in manpower and resources, as well as the absence of unified standards and practical technical approaches [8,9].

In recent years, large language models (LLMs) have demonstrated notable strengths in text generation and rewriting [10-13]. Their natural language processing capabilities not only enhance readability and structural clarity but also offer universities lacking professional support a low-cost and scalable auxiliary approach to gradually establish more standardized consent form systems [14-16]. However, empirical evidence remains limited regarding whether LLMs may inadvertently omit critical information during rewriting [17], how to balance professionalism with comprehensibility [18], and whether they can genuinely contribute to the standardization of such documents [19].

To address these gaps, we collected counseling consent forms from 33 universities in China and applied 2 state-of-the-art LLMs, ChatGPT (GPT-5) and Grok-4, to optimize and standardize the texts. The evaluation was conducted across three dimensions: textual structure and readability, content quality, and reading comprehension. By comparing original and LLM-optimized versions, this study aims to examine the feasibility and limitations of using LLMs to improve counseling consent forms and to provide insights for strengthening youth mental health services and exploring human-artificial intelligence (AI) collaboration in clinical communication.

Methods

Study Materials and Sources

This study collected 33 psychological counseling ICFs, including 32 from the official websites of Chinese universities and 1 template issued by the Clinical Psychology Registration Work Committee of the Chinese Psychological Society [20]. The template, initially used for pilot testing of LLM rewriting and prompt design, was later included in the final analysis. The sampled universities covered China's eastern, central, western, and northeastern regions, ensuring geographic representativeness, and encompassed national key, provincial key, and local undergraduate institutions, reflecting variation across higher education tiers. To maintain objectivity and validity, all original documents were analyzed without modification, other than the removal of university identifiers. A full list of universities is provided in Table S1 (see Multimedia Appendix 1).

Model Selection and Text Generation

Two representative LLMs were examined: ChatGPT (GPT-5) (OpenAI; released August 7, 2025) and Grok-4 (xAI; released July 9, 2025) [21,22]. Both are recently released, widely adopted, and high-performing closed-source systems with strong capabilities in Chinese language processing and text generation [23,24]. All model outputs were generated using the publicly available versions of these models accessible during a defined time window (August 21-September 8, 2025), and the reported results reflect model behavior at the time of data generation. Prior to formal experimentation, we developed and tested a structured prompt defining requirements for contextual framing, preservation and supplementation of key information, linguistic style, and textual organization (Supplementary Information S1 in Multimedia Appendix 2). The structured prompts were natively drafted in Chinese to align with the language of the source documents, and no cross-linguistic translation was involved. The template ICF was used for preliminary trials to verify that the models generated content that was both complete and structurally consistent. Following this procedure, each original text was optimized independently by ChatGPT (GPT-5) and Grok-4, yielding three versions—Original, ChatGPT (GPT-5), and Grok-4—for subsequent comparison and evaluation (all documents are provided in Multimedia Appendix 3). When generating optimized ICFs, both ChatGPT (GPT-5) and Grok-4 were prompted using the same instruction template to ensure fair comparison. The prompts required adherence to both international ethical guidelines and domestic regulations relevant to counseling practice, with Chinese laws prioritized in cases of discrepancy. Retrieval-augmented generation or external document retrieval was not used; all outputs were generated through the models' standard web-based interfaces.

Textual Structure and Readability Analysis

To evaluate differences among the 3 ICF versions in structural complexity and linguistic expression, 8 metrics were applied. Five basic structural features were measured: character count, word count, sentence count, words per sentence, and characters per word [4,6,25,26]. Syntactic complexity was captured by the nested sentence ratio [27], reading difficulty by the Lee-Yang Readability Index [28], and tone friendliness was added to reflect approachability and acceptability [29]. All features were derived through a standardized processing pipeline.

The Lee-Yang Readability Index is a weighted composite of average sentence length, word length, and lexical complexity, widely used to assess the readability of Chinese texts [30], especially in science communication, education, and health contexts [28,31]. Tone friendliness was operationalized using a lexical counting approach. Two predefined dictionaries—politeness- or empathy-related words and harsh words—were applied (Supplementary Information S2 in Multimedia Appendix 2) [32]. For each text, the normalized

XSL•FO
RenderX

frequency difference between the 2 word categories (adjusted by sentence count) was linearly mapped from −1 to 1 onto a 0 to 5 scale, with values truncated if outside this range. This metric was intended as a surface-level lexical proxy of linguistic tone rather than a contextual or pragmatic assessment of emotional friendliness. Higher scores indicate a greater relative use of polite or empathic wording and fewer harsh or legalistic terms. The method is interpretable, reproducible, and adaptable through dictionary refinement [33].

Content Quality Assessment

To assess the completeness and standardization of ICFs, we drew on major international and national ethical guidelines. These included the American Psychological Association Ethical Principles and Code of Conduct (2016) [34], the British Psychological Society Code of Ethics and Conduct (2018) [35], the Code of Ethics for Clinical and Counseling Psychology (2018) [36], and the World Health Organization Comprehensive Mental Health Action Plan 2013 - 2030 (2021) [37]. Together, these documents provided the ethical and methodological foundation for the content quality evaluation framework.

Building on this foundation, we established an evaluation framework of 20 core indicators widely considered essential for ICFs: confidentiality; exceptions to confidentiality; client rights; guardian consent; goals and scope; format and frequency; fees and cancelation policy; recording methods; authorization and revocation; crisis procedures; complaints and appeals; data protection; disclaimer of boundaries; language clarity; voluntariness; client obligations; counseling limitations; counselor qualifications; counseling modalities; target population (detailed descriptions and rationales in Multimedia Appendix 1, Table S2). Each indicator was assessed across three dimensions: existence (0 - 1), defined as whether the element was explicitly presented as an independent item (eg, a distinct heading or clearly delineated paragraph); specificity (0 - 2), reflecting whether concrete details were provided; and operability (0 - 2), indicating whether actionable procedures were specified (detailed in Multimedia Appendix 1**,** Table S3). When a relevant indicator was mentioned within other sections without being explicitly listed as an independent item, existence was scored as 0, while specificity and operability were still assessed based on the content provided. Each indicator had a maximum of 5 points, for a total of 100 across all 20 indicators, with subscores of 20 for existence, 40 for specificity, and 40 for operability.

Prior to formal application, 3 senior experts in psychology and ethics reviewed and revised the proposed framework to ensure scientific rigor, comprehensiveness, and operability. A pilot test using the template ICF was then conducted to examine applicability and discriminative capacity. Once feasibility was confirmed, the finalized framework was applied to the remaining 32 ICFs. Scoring was performed independently by 5 experts with backgrounds in psychology or mental health. To maximize fairness and consistency, all texts were randomized and anonymized so that raters were blinded to both the source and version.

Reading Comprehension Assessment

To assess differences in comprehensibility among ICF versions, 10 university volunteers from nonpsychology and nonhealth-related majors participated as readers simulating clients; none had received formal training in counseling or mental health disciplines. They evaluated the texts using a 6D framework: comprehensibility—ease of reading and understanding [8,38]; clarity—whether key information was clearly and unambiguously expressed [39,40]; trustworthiness—perceived reliability of content [41,42]; friendliness—supportive, respectful, and approachable tone [32]; professionalism—appropriate formality and expertise [43]; and acceptability—whether content and style were acceptable to clients [44]. Each dimension was rated on a 5-point Likert scale (1=strongly disagree, 5=strongly agree) [45,46]. The maximum per-dimension score was 5, yielding a total score of 6 - 30. A seventh composite dimension ("dimension means"), defined as the average of the six dimensions (range 1–5), was calculated. Prior to the formal study, a pilot test with the template ICF was conducted to examine the scale's applicability and discriminative capacity. During formal rating, all texts were randomized and anonymized to blind raters to both source and version, ensuring objectivity.

Statistical Analysis

Differences among the 3 ICF versions were examined across textual structure and readability, content quality, and reading comprehension. For textual structure and readability, values for 8 indicators were calculated for each document. For content quality and reading comprehension, ratings from multiple evaluators were averaged per document and then summarized by version. In all cases, means and SDs were reported.

All between-version comparisons were conducted using the Wilcoxon signed rank test. Effect sizes (r) were reported, and P-values were adjusted for multiple testing using the false discovery rate approach (Benjamini-Hochberg procedure). Interrater reliability for total scores of content quality and reading comprehension was assessed using intraclass correlation coefficients (ICCs), based on ratings from 5 experts and 10 readers, respectively.

In addition, to account for individual differences among raters and documents, linear mixed-effects models (LMMs) were fitted with "version" as a fixed effect and "document ID" and "rater" as random effects, allowing robust estimation of version effects under a repeated-measures design. Because content quality and reading comprehension were evaluated by different groups, separate models were constructed for each outcome. Estimation used the MixedLM module. All tests were 2-tailed with significance set at $P<.05$. P values from the LMMs were not adjusted for multiple comparisons; therefore, the possibility of type I error due to multiple testing cannot be excluded. Analyses were performed in Python 3.11 using pandas (v2.2.2; pandas development team), NumPy (v1.26.4; NumPy developers), SciPy (v1.11.4; SciPy developers), jieba (v0.42.1; Jieba development team), statsmodels (v0.14.1; statsmodels developers), scikit-learn (v1.4.0; scikit-learn developers), and Pingouin (v0.5.4; Pingouin developers). All codes are available in Multimedia Appendix 4 for reproducibility.

XSL•FO
RenderX

Ethical Considerations

This study involved human participants and was reviewed and approved by the Ethics Committee of West China Hospital, Sichuan University (Approval 2024‑1431). The study was classified as minimal risk and conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all participants prior to participation, including expert raters and student volunteers who evaluated deidentified university counseling ICFs; no clinical intervention or therapeutic interaction was involved. To protect privacy and confidentiality, all documents were anonymized before analysis, with institutional identifiers removed, and no personally identifiable information was collected or analyzed. Participants received no financial or material compensation, and the manuscript and multimedia appendices do not contain any images or content that could enable identification of individual participants.

Results

Textual Structure and Readability

Compared with the original texts, ChatGPT (GPT-5)–optimized ICFs showed more sentences (mean 51.30, SD 12.27 vs mean 36.79, SD 13.29; $P<.001$), fewer words per sentence (mean 20.23, SD 2.29 vs mean 28.05, SD 6.45; $P<.001$), more characters per word (mean 1.16, SD 0.04 vs mean 1.12, SD 0.03; $P<.001$), and a lower nested sentence ratio (mean 0.21, SD 0.06 vs mean 0.32, SD 0.14; $P<.001$). Character count (mean 1206.42, SD 353.00 vs mean 1117.61, SD 376.58; $P=.17$) and word count (mean 1039.33, SD 297.52 vs mean 996.24, SD 338.14; $P=.44$) did not differ significantly. The Lee-Yang Readability Index was lower (mean 22.39, SD 2.13 vs mean 28.68, SD 5.69; $P<.001$), while tone friendliness was higher (mean 2.67, SD 0.12 vs mean 2.57, SD 0.29; $P=.04$).

Grok-4–optimized ICFs were significantly longer, with higher character counts (mean 1669.52, SD 234.09; $P<.001$) and word counts (mean 1393.21, SD 208.82; $P<.001$). Sentence count also increased (mean 65.33, SD 10.15; $P<.001$), words per sentence decreased (mean 21.49, SD 2.56; $P<.001$), characters per word increased (mean 1.20, SD 0.04; $P<.001$), and the nested sentence ratio decreased (mean 0.23, SD 0.06; $P<.001$). The Lee-Yang Index also decreased (mean 24.37, SD 2.32; $P<.001$), and tone friendliness improved (mean 2.67, SD 0.13; $P=.02$).

Direct comparisons showed that Grok-4 had higher character counts, word counts, sentence counts, and characters per word ($P<.001$). Differences in words per sentence were modest ($P=.02$), nested sentence ratios did not differ ($P=.09$), and Grok-4 had a higher Lee-Yang Readability Index ($P<.001$). Tone friendliness was comparable ($P=.76$). Full results are shown in Table 1 and Figure 1. Detailed statistical analysis results are provided in Table S4 (Multimedia Appendix 1).

XSL•FO
RenderX

Table . Comparative analysis of multidimensional evaluation results across all versions.

Indicator	Original, mean (SD)	ChatGPT-5, mean (SD)	Grok-4, mean (SD)	ChatGPT-5 versus original (*P* value)[a]	Grok-4 versus original (*P* value)	Grok-4 versus ChatGPT-5 (*P* value)
Text structure						
Character count	1117.61 (376.58)	1206.42 (353.00)	1669.52 (234.09)	.17	<.001	<.001
Word count	996.24 (338.14)	1039.33 (297.52)	1393.21 (208.82)	.44	<.001	<.001
Sentence count	36.79 (13.29)	51.30 (12.27)	65.33 (10.15)	<.001	<.001	<.001
Words per sentence	28.05 (6.45)	20.23 (2.29)	21.49 (2.56)	<.001	<.001	.02
Characters per word	1.12 (0.03)	1.16 (0.04)	1.20 (0.04)	<.001	<.001	<.001
Nested sentence ratio	0.32 (0.14)	0.21 (0.06)	0.23 (0.06)	<.001	<.001	.09
Readability						
Lee-Yang Readability Index	28.68 (5.69)	22.39 (2.13)	24.37 (2.32)	<.001	<.001	<.001
Tone friendliness	2.57 (0.29)	2.67 (0.12)	2.67 (0.13)	.04	.02	.76
Content quality						
Existence	11.03 (2.18)	12.72 (1.26)	13.55 (1.24)	<.001	<.001	.004
Specificity	11.55 (3.08)	16.51 (3.38)	18.25 (2.38)	<.001	<.001	.02
Operability	22.75 (5.73)	23.31 (6.12)	23.69 (6.46)	.31	.10	.60
Total	45.33 (8.74)	52.54 (7.92)	55.49 (7.81)	<.001	<.001	.01
Reading comprehension						
Comprehensibility	2.93 (0.45)	3.76 (0.30)	3.69 (0.35)	<.001	<.001	.19
Clarity	2.98 (0.47)	3.78 (0.28)	3.74 (0.28)	<.001	<.001	.32
Trustworthiness	3.30 (0.30)	3.72 (0.26)	3.61 (0.32)	<.001	<.001	.12
Friendliness	3.35 (0.35)	3.75 (0.32)	3.73 (0.27)	<.001	<.001	.79
Professionalism	3.25 (0.25)	3.55 (0.32)	3.64 (0.30)	<.001	<.001	.21
Acceptability	3.20 (0.40)	3.77 (0.37)	3.65 (0.33)	<.001	<.001	.06
Total[b]	19.02 (1.32)	22.33 (0.81)	22.05 (0.90)	<.001	<.001	.11
Dimension means (total[c]/6)	3.17 (0.22)	3.72 (0.13)	3.68 (0.15)	<.001	<.001	.11

[a]*P* values were adjusted for multiple comparisons using the false discovery rate.

[b]Total score of the 3 dimensions in content quality evaluation.

[c]Total score of the 6 dimensions in reading comprehension evaluation.

XSL•FO
RenderX

Figure 1. Text structure and readability indicators across original and large language model–optimized university counseling informed consent forms. Box plots compare 8 structural and readability indicators across original and ChatGPT-5– and Grok-4–rewritten counseling informed consent forms (n=33). Dots represent individual documents. Statistical comparisons were performed using Wilcoxon signed rank tests. Significance thresholds: *P*<.05 (*), *P*<.01 (**), *P*<.001 (***), not significant (ns). (A) Character count, (B) word count, (C) sentence count, (D) words per sentence, (E) characters per word, (F) nested sentence ratio, (G) Lee-Yang Readability Index, and (H) tone friendliness.

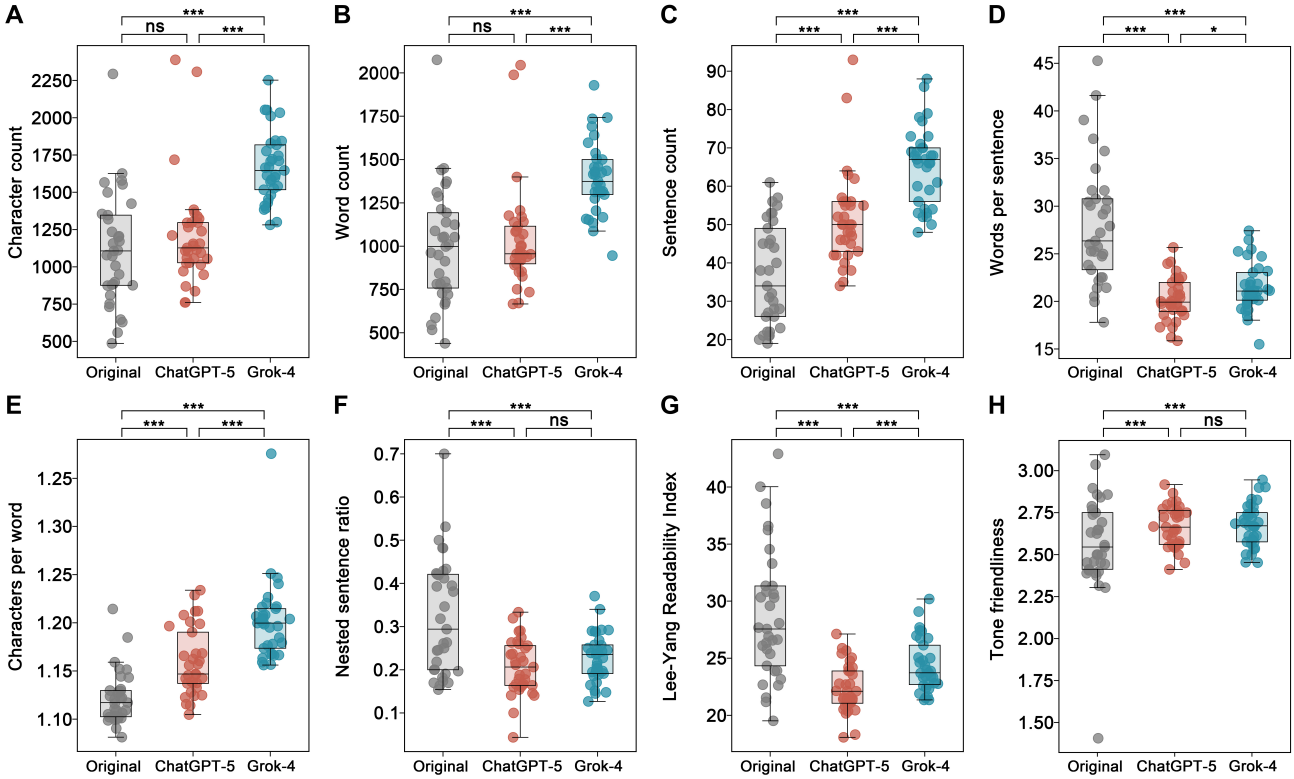

Content Quality

Interrater reliability among 5 expert raters across all documents was high (ICC [2,k]=0.83; Table S5 in Multimedia Appendix 1). Compared with the original texts, ChatGPT (GPT-5)–optimized ICFs showed higher existence (mean 12.72, SD 1.26 vs mean 11.03, SD 2.18; *P*<.001) and specificity scores (mean 16.51, SD 3.38 vs mean 11.55, SD 3.08; *P*<.001), while operability did not differ (mean 23.31, SD 6.12 vs mean 22.75, SD 5.73; *P*=.31). The total score increased significantly (mean 52.54, SD 7.92 vs mean 45.33, SD 8.74; *P*<.001). Grok-4 also showed marked gains, with higher existence (mean 13.55, SD 1.24; *P*<.001) and specificity (mean 18.25, SD 2.38; *P*<.001). Operability increased (mean 23.69, SD 6.46) but was not significant (*P*=.09). The total score rose to a mean of 55.49 (SD 7.81; *P*<.001). Direct comparisons showed Grok-4 scored higher than ChatGPT (GPT-5) in existence (*P*=.004) and specificity (*P*=.02), with no difference in operability (*P*=.60). The overall score was also higher for Grok-4 (*P*=.01) (Table 1). Detailed statistical analysis results are provided in Multimedia Appendix 1, Table S6.

To illustrate performance across the 20 core indicators, a heatmap was generated (Figure 2). The analyses showed that several indicators—confidentiality, exceptions to confidentiality, goals and scope, authorization and revocation, complaints and appeals, disclaimer of boundaries, and counseling modalities—improved significantly under both LLM-optimized versions. These gains were mainly in specificity, with few improvements in existence or operability. Between the 2 LLMs, significant differences appeared only in client rights and language clarity; all other indicators showed no differences. Full statistical results are provided in Tables S7-S11, and rater-level scores in Supplementary Data S1-S5 (Multimedia Appendix 5).

XSL•FO
RenderX

Figure 2. Expert-rated content quality profiles of original and large language models–optimized university counseling informed consent forms. Heatmaps display document-level content quality scores for 33 counseling informed consent forms, evaluated across 20 predefined indicators grouped into three dimensions (existence, specificity, and operability), as well as the total score. Darker colors indicate higher scores. (A) Grok-4–rewritten versions, (B) ChatGPT (GPT-5)–rewritten versions, and (C) original documents. Dimension-level summaries include (i) total score, (ii) existence score, (iii) specificity score, and (iv) operability score.

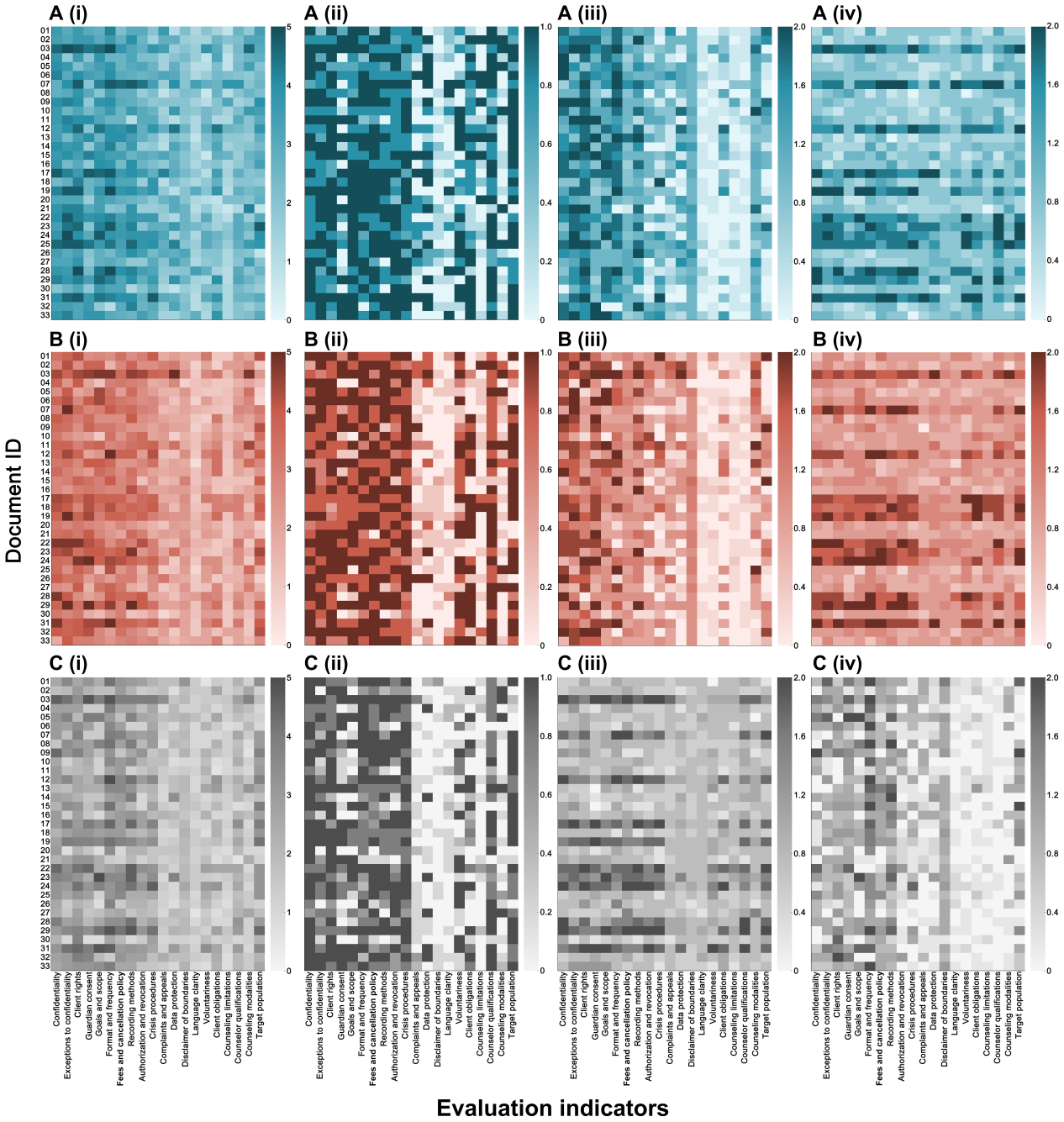

Reading Comprehension

Interrater reliability among 10 readers across all documents was fair (ICC [2,k]=0.59; Table S12 in Multimedia Appendix 1). LLM-optimized ICFs outperformed the originals across all dimensions and in total scores (Table 1). For ChatGPT (GPT-5), scores improved in comprehensibility (mean 3.76, SD 0.30 vs mean 2.93, SD 0.45; *P*<.001), clarity (mean 3.78, SD 0.28 vs mean 2.98, SD 0.47; *P*<.001), trustworthiness (mean 3.72, SD 0.26 vs mean 3.30, SD 0.30; *P*<.001), friendliness (mean 3.75, SD 0.32 vs mean 3.35, SD 0.35; *P*<.001), professionalism (mean 3.55, SD 0.32 vs mean 3.25, SD 0.25; *P*<.001), and acceptability (mean 3.77, SD 0.37 vs mean 3.20, SD 0.40; *P*<.001). The total score also rose (mean 22.33, SD 0.81 vs mean 19.02, SD 1.32; *P*<.001).

Grok-4 showed similar mean gains: comprehensibility 3.69 (SD 0.35; *P*<.001), clarity 3.74 (SD 0.28; *P*<.001), trustworthiness 3.61 (SD 0.32; *P*<.001), friendliness 3.73 (SD 0.27; *P*<.001), professionalism 3.64 (SD 0.30; *P*<.001), and acceptability 3.65 (SD 0.33; *P*<.001), with a total of 22.05 (SD 0.90; *P*<.001).

In direct comparison, ChatGPT (GPT-5) had higher means across most dimensions and the total score, but none reached significance (comprehensibility: *P*=.19; clarity: *P*=.32;

XSL•FO
RenderX

trustworthiness: *P*=.12; friendliness: *P*=.79; professionalism: *P*=.21; and acceptability: *P*=.06). For visualization, dimension means were plotted instead of totals (Figure 3). Detailed statistical analysis results are provided in Table S13 in Multimedia Appendix 1. Complete dimension-level ratings are in Supplementary Data S6-S15 in Multimedia Appendix 5.

Figure 3. Volunteer-rated reading comprehension of original and large language model–optimized university counseling informed consent forms.

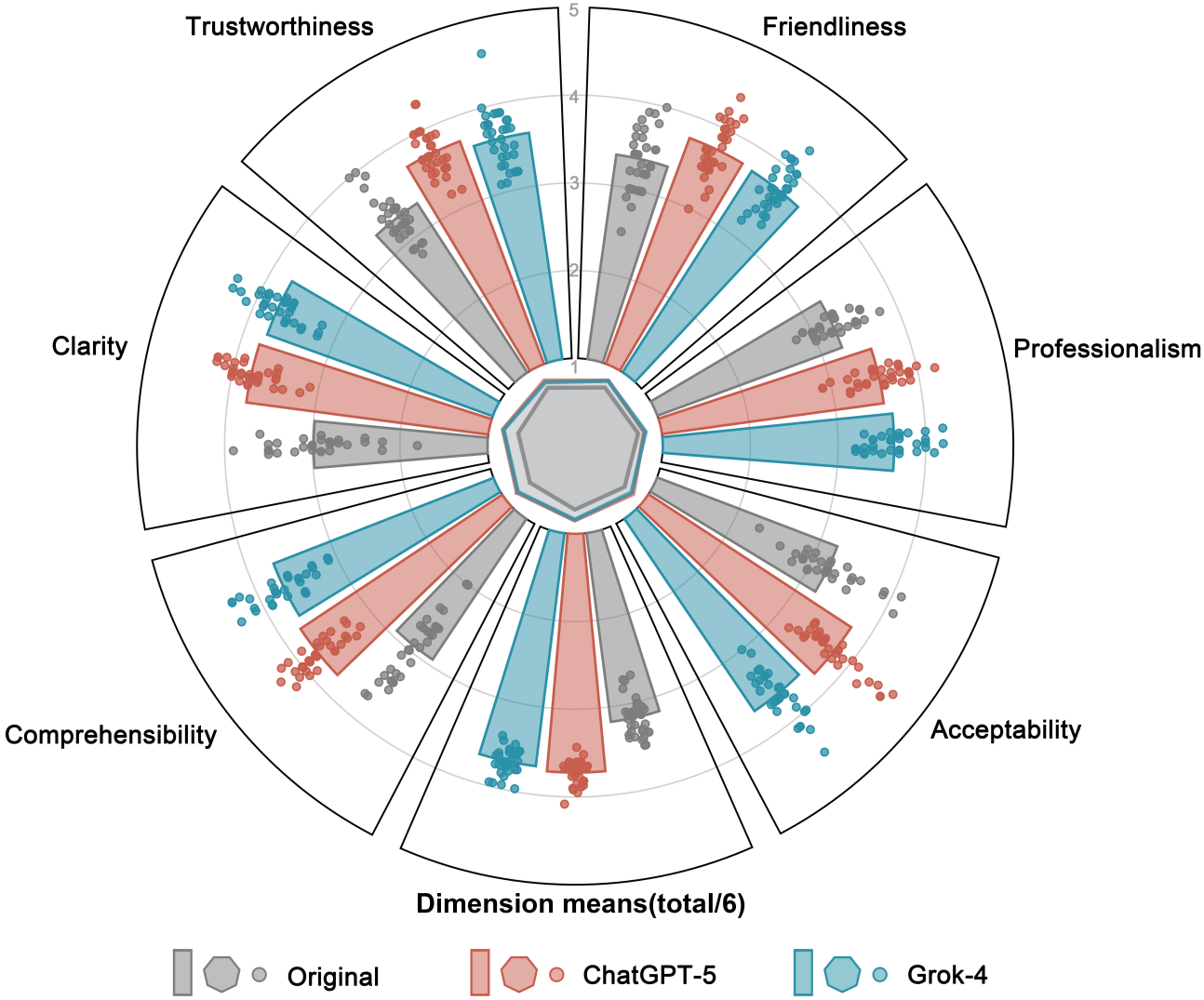

The figure compares reading comprehension scores across original, ChatGPT (GPT-5)–rewritten, and Grok-4–rewritten counseling ICFs (n=33). Mean scores across 6 comprehension dimensions and the total score are shown, with dots indicating document-level distributions. A central radar plot summarizes dimension-level means for comparison.

Linear Mixed-Effects Model

To account for individual differences among raters and texts, LMMs with crossed random effects were constructed, specifying "version" as a fixed effect and including random intercepts for rater and document ID. Separate models were fitted for content quality and reading comprehension (Figure 4). In the content quality model, ChatGPT (GPT-5) yielded higher total scores than the original texts (β=7.20, *P*<.001), with significant gains in existence (β=1.62; *P*<.001) and specificity (β=5.22; *P*<.001), but not operability (β=.36; *P*=.17). Grok-4 also showed significant improvements over the originals in total scores (β=10.15; *P*<.001), existence (β=2.45; *P*<.001), specificity (β=6.96; *P*<.001), and operability (β=.74; *P*=.02). Direct comparison showed Grok-4 scored higher than ChatGPT (GPT-5) in total scores (β=2.95; *P*<.001), existence (β=0.82; *P*<.001), and specificity (β=1.75; *P*<.001), with no difference in operability (β=.38; *P*=.27).

XSL•FO
RenderX

Figure 4. Linear mixed-effects model results for (A) content quality and (B) reading comprehension across original and large language model–optimized university counseling informed consent forms.

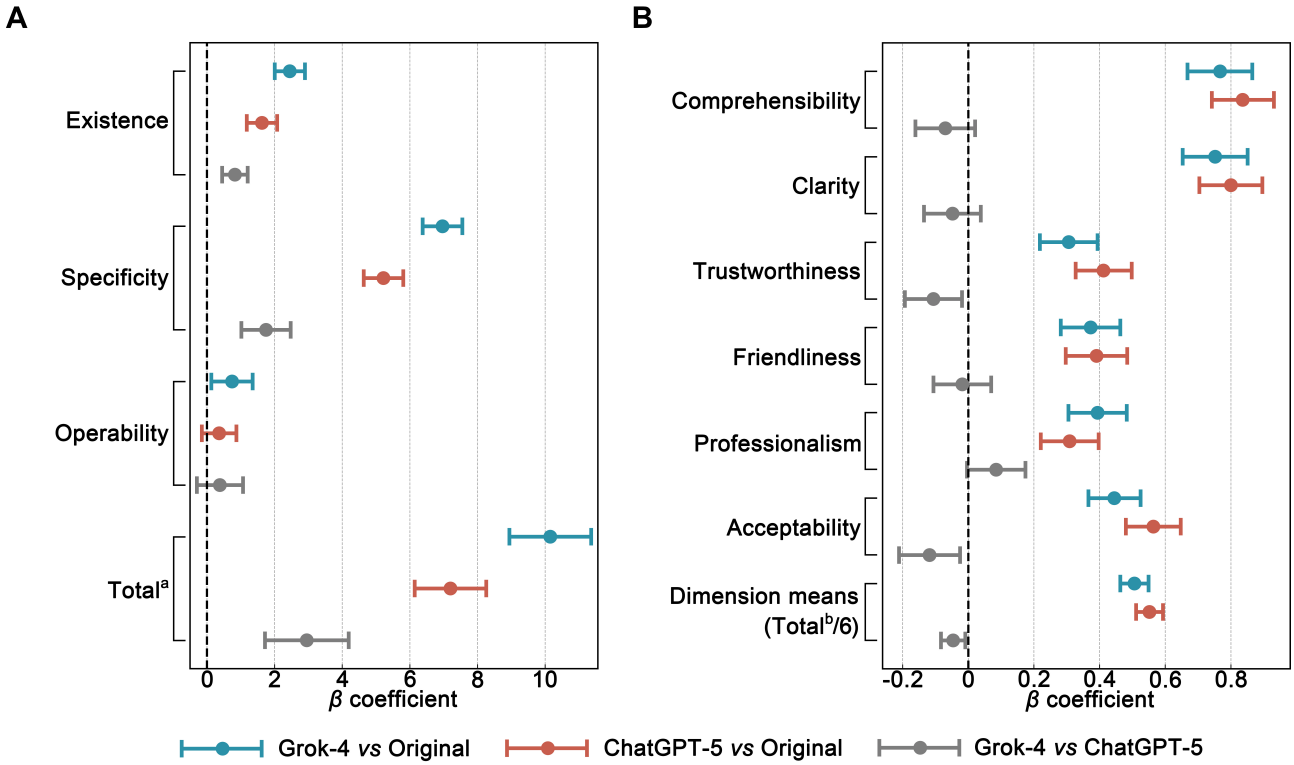

The figure shows LMM estimates comparing original, ChatGPT (GPT-5)–rewritten, and Grok-4–rewritten counseling ICFs (n=33). Models were fitted separately for content quality and reading comprehension outcomes. Positive β coefficients indicate higher scores for the first group in each comparison. The vertical dashed line at β=0 represents no difference between groups. CIs that do not cross the zero line indicate statistically significant differences at *P*<.05.

In the reading comprehension model, ChatGPT (GPT-5) showed significant improvements over the original texts in comprehensibility (β=.84; *P*<.001), clarity (β=.80; *P*<.001), trustworthiness (β=.41; *P*<.001), friendliness (β=.39; *P*<.001), professionalism (β=.31; *P*<.001), acceptability (β=.56; *P*<.001), and dimension means (β=.55; *P*<.001). Grok-4 likewise outperformed the originals, with gains in comprehensibility (β=.77; *P*<.001), clarity (β=.75; *P*<.001), trustworthiness (β=.31; *P*<.001), friendliness (β=.37; *P*<.001), professionalism (β=.39; *P*<.001), acceptability (β=.45; *P*<.001), and dimension means (β=.51; *P*<.001). In direct comparison, Grok-4 scored lower than ChatGPT (GPT-5) in trustworthiness (β=−.11; *P*=.02), acceptability (β=−.12; *P*=.01), and dimension means (β=−.05; *P*=.01). Details are reported in Table 2, with random effects and model fit indices in Multimedia Appendix 1, Tables S14 and S15.

XSL•FO
RenderX

Table . Results of linear mixed-effects model analysis on content quality and reading comprehension.

Dimension and comparison	β (95% CI)	SE	*P* value[a]
Content quality			
Existence			
Grok-4 vs original	2.45 (2.00 to 2.90)	0.23	<.001
ChatGPT-5 vs original	1.62 (1.18 to 2.07)	0.23	<.001
Grok-4 vs ChatGPT-5	0.82 (0.45 to 1.20)	0.19	<.001
Specificity			
Grok-4 vs original	6.96 (6.38 to 7.55)	0.30	<.001
ChatGPT-5 vs original	5.22 (4.63 to 5.80)	0.30	<.001
Grok-4 vs ChatGPT-5	1.74 (1.02 to 2.47)	0.37	<.001
Operability			
Grok-4 vs original	0.74 (0.13 to 1.35)	0.31	.02
ChatGPT-5 vs original	0.36 (–0.15 to 0.87)	0.26	.17
Grok-4 vs ChatGPT-5	0.38 (–0.30 to 1.06)	0.35	.27
Total[b]			
Grok-4 vs original	10.15 (8.94 to 11.36)	0.62	<.001
ChatGPT-5 vs original	7.20 (6.14 to 8.26)	0.54	<.001
Grok-4 vs ChatGPT-5	2.95 (1.71 to 4.19)	0.63	<.001
Reading comprehension			
Comprehensibility			
Grok-4 vs original	0.77 (0.67 to 0.86)	0.05	<.001
ChatGPT-5 vs original	0.84 (0.74 to 0.93)	0.05	<.001
Grok-4 vs ChatGPT-5	–0.07 (–0.16 to 0.02)	0.05	.13
Clarity			
Grok-4 vs original	0.75 (0.65 to 0.85)	0.05	<.001
ChatGPT-5 vs original	0.80 (0.70 to 0.90)	0.05	<.001
Grok-4 vs ChatGPT-5	–0.05 (–0.14 to 0.04)	0.04	.27
Trustworthiness			
Grok-4 vs original	0.31 (0.22 to 0.39)	0.05	<.001
ChatGPT-5 vs original	0.41 (0.33 to 0.50)	0.04	<.001
Grok-4 vs ChatGPT-5	–0.11 (–0.19 to 0.02)	0.04	.02
Friendliness			
Grok-4 vs original	0.37 (0.28 to 0.46)	0.05	<.001
ChatGPT-5 vs original	0.39 (0.30 to 0.48)	0.05	<.001
Grok-4 vs ChatGPT-5	–0.02 (–0.11 to 0.07)	0.05	.69
Professionalism			
Grok-4 vs original	0.39 (0.30 to 0.48)	0.05	<.001
ChatGPT-5 vs original	0.31 (0.22 to 0.40)	0.05	<.001
Grok-4 vs ChatGPT-5	0.08 (0.00 to 0.17)	0.05	.06
Acceptability			
Grok-4 vs original	0.44 (0.37 to 0.53)	0.04	<.001
ChatGPT-5 vs original	0.56 (0.48 to 0.65)	0.04	<.001

XSL•FO
RenderX

Dimension and comparison	β (95% CI)	SE	*P* value[a]
Grok-4 vs ChatGPT-5	–0.12 (–0.21 to 0.03)	0.05	.01
Total[c]			
Grok-4 vs original	3.04 (2.78 to 3.30)	0.13	<.001
ChatGPT-5 vs original	3.31 (3.07 to 3.56)	0.13	<.001
Grok-4 vs ChatGPT-5	–0.28 (–0.50 to 0.06)	0.11	.01
Dimension means (Total[c]/6)			
Grok-4 vs original	0.51 (0.46 to 0.55)	0.02	<.001
ChatGPT-5 vs original	0.55 (0.51 to 0.59)	0.02	<.001
Grok-4 vs ChatGPT-5	–0.05 (–0.08 to 0.01)	0.02	.01

[a]*P* values from the linear mixed-effects models were not adjusted for multiple comparisons.

[b]Total score of the 3 dimensions in content quality evaluation.

[c]Total score of the 6 dimensions in reading comprehension evaluation.

Discussion

Principal Results

This study systematically evaluated the performance of LLMs in rewriting university counseling ICFs. It focused on 2 state-of-the-art closed-source models, ChatGPT (GPT-5) and Grok-4. Both achieved significant improvements over the originals in textual structure and readability, content quality, and reading comprehension. LLM-based rewriting reduced linguistic complexity, enhanced the completeness and specificity of key information, and improved clarity and acceptability, making ICFs easier to understand and apply. Such improvements may help lower barriers to accessing counseling, thereby supporting more timely use of mental health services among young people. LMMs confirmed these advantages across multiple dimensions. This study is also the first to design and apply a multidimensional evaluation framework for university counseling ICFs. The framework deliberately distinguished evaluation perspectives: content quality was assessed by psychology or mental health experts from a counselor's standpoint, while reading comprehension was evaluated by university volunteers simulating clients. This dual-perspective design enabled the assessment of both professional and ethical adequacy and actual acceptability and comprehensibility. Together, the findings underscore the value of LLMs in optimizing counseling ICFs and offer a methodological reference for future research.

Comparison With Prior Work

In terms of textual structure and readability, LLM-based rewriting substantially altered the linguistic features of ICFs. ChatGPT (GPT-5)–generated texts showed shorter words per sentence and lower nested sentence ratios, indicating improved clarity and comprehensibility through sentence splitting and syntactic simplification [47]. The increase in characters per word likely reflects the use of more standardized, professional vocabulary to ensure accuracy while simplifying syntax [4]. The lower Lee-Yang Readability Index and higher tone friendliness suggest reduced overall difficulty with a more approachable style [28]. It should be noted that tone friendliness, as operationalized here, reflects surface-level lexical tendencies and may not fully capture context-dependent or pragmatically appropriate expressions; for example, legally protective wording may serve a positive ethical function despite being classified as strict. Although statistically significant, the absolute increase was small (approximately 0.1 on a 5-point scale), and the linear mapping approach may amplify minor lexical frequency differences. These findings should therefore be interpreted as indicative of modest stylistic shifts rather than substantive changes in empathetic communication. In addition, sentence counts rose significantly, with modest increases in character and word counts, likely reflecting supplementation of missing content [26]. Grok-4 followed a similar strategy but produced longer texts with larger increases in length and sentence count, indicating more active supplementation of absent material, while still improving readability and tone friendliness. However, increased length may also impose practical burdens in counseling settings, highlighting a trade-off between informational completeness and usability. The lack of significant differences between the models in nested sentence ratio and tone friendliness suggests convergence in syntactic and stylistic adjustments [48]. These findings align with prior research. Ali et al [49] reported that GPT-4 simplification of surgical consent forms lowered reading difficulty and enhanced readability, while Beattie et al [50] found LLM-optimized trial consent forms easier to understand. Mirza et al [26] further showed that US trial consent forms are written at overly high reading levels, with each grade-level increase linked to ~16% participant attrition, underscoring the value of simplification. The present study confirms that LLM-based rewriting reduces syntactic complexity and improves readability, with ChatGPT (GPT-5) emphasizing clarity through conciseness and Grok-4 expanding coverage through longer texts, highlighting differences in generation strategies that may shape readers' experiences.

The analysis of content quality indicates that LLM-based rewriting effectively supplemented key information and enhanced the concreteness of items, particularly in the specificity dimension. However, improvements in operability were relatively limited, suggesting that models still face challenges in generating explicit execution pathways and actionable

XSL•FO
RenderX

requirements. Conceptually, operability requires specifying not only what should be done, but also who is responsible, when actions occur, and under what conditions they apply. Such institution-specific procedural details are difficult for LLMs to generate reliably without access to local workflows. Consequently, operability may be less suitable for fully automated optimization and better addressed through human-AI collaboration, with LLMs supporting structural drafting and clinicians refining actionable details. This limited improvement may also reflect the absence of institution-specific "ground truth" information in the source texts and prompts. Because no retrieval-augmented generation or context injection was used, the models lacked access to concrete local procedural details (eg, office hours or emergency contacts), which inherently constrained operability-related gains. Grok-4 showed a more proactive optimization pattern. Taken together, these perspectives were further supported by the findings of the LMM analyses. Previous studies have similarly highlighted the potential of LLMs to improve the quality of ICFs and health-related documents. Shi et al [51] confirmed that LLMs significantly enhanced completeness and specificity in clinical trial consent forms. Oh et al [52] showed that LLMs simplified surgical consent forms while preserving expert-endorsed quality. Rust et al [53] and Romoff et al [54] reported in cardiology discharge summaries and orthopedic educational materials that LLM-based rewriting not only improved readability but also consistently received positive evaluations in content quality. These findings align with the present results, suggesting that LLMs have stable advantages in reinforcing information coverage and refining provisions, while limitations in operability-oriented expressions warrant further investigation.

With regard to practical implementation, operability emerged as a particularly challenging dimension. Indicator-level analyses further revealed that LLM optimization was concentrated on information presentation and provision refinement, with vague or missing clauses supplemented and clarified after rewriting. By contrast, operability and implementation pathways showed only modest gains. While this can be regarded as a limitation at the rewriting level, it also reflects the models' cautious tendency to avoid generating potentially erroneous or misleading operational details in high-risk contexts, thereby reducing the risk of hallucinations. This trend echoes prior research: Alber et al [55] warned of data poisoning risks in medical corpora, underscoring the need for prudence, while Sebastian et al [56] systematically documented hallucinations in LLM-generated medical texts and emphasized risk evaluation mechanisms. Collectively, these observations suggest that the latest generation of models may exhibit enhanced stability in sensitive contexts. Moreover, this pattern leaves scope for improvement through prompt engineering or human-AI collaboration to guide models toward generating more detailed and personalized operational provisions.

It is noteworthy that cases were observed in which existence was scored as 0, while specificity or operability received nonzero scores. This pattern reflects a distinction between structural explicitness and semantic presence incorporated in the evaluation framework. In the present study, the existence dimension was intentionally defined to capture whether information was explicitly listed as an independent item, rather than whether related content appeared anywhere in the document. This design choice was motivated by the intended audience of counseling ICFs, which primarily consists of nonprofessional readers who rely on clear, structured, and itemized presentation to accurately identify key information. Accordingly, information that was mentioned within other sections without explicit listing could still be evaluated for its specificity and operability once identified, but such nonexplicit presentation may reduce retrievability and comprehensibility, as relevant details—even if included—may not be readily recognized without clear structural cues [57]. Although LLM-based optimization improved overall performance in existence and specificity, fully explicit and well-structured presentation was not consistently achieved across all items [49]. These findings underscore the importance of structural clarity in counseling ICFs and suggest that further improvements may benefit from the integration of professional human review and standardized templates to enhance explicitness and ease of information retrieval [26].

With respect to reading comprehension, LLM-based rewriting significantly outperformed the original texts across all 6 dimensions and the total score, indicating that syntactic simplification and structural optimization enhanced clients' understanding of ICFs. The 2 LLMs showed broadly similar patterns, with ChatGPT (GPT-5) slightly higher than Grok-4 on most dimensions. Notably, discrepancies arose between the Wilcoxon test and the LMM in model-to-model comparisons: Wilcoxon signed rank tests suggested no significant differences, whereas LMMs indicated that ChatGPT-5 was significantly superior in trustworthiness and acceptability. This discrepancy may reflect the small sample size (10 readers) and only moderate interrater reliability, which limited the power of nonparametric tests to detect differences. In interpreting these findings, we primarily relied on the LMM results, as they explicitly account for both document- and rater-level variability and are better suited to the hierarchical structure of the data. These observations accord with prior research. Will et al [58] reported that LLM-based rewriting improved the readability of online patient education materials without compromising accuracy, while Ali et al [49] found that optimized surgical consent forms enhanced comprehensibility while preserving professionalism and trustworthiness. Together with the present findings, this evidence suggests that LLMs can refine the structure and language of counseling ICFs to improve client comprehension while maintaining credibility and professional integrity.

Overall, this study is the first to systematically validate the feasibility and value of applying LLMs to optimize ICFs for university mental health counseling. The results demonstrated consistent improvements in text structure, content quality, and comprehensibility, with recognition from both professional evaluators and simulated clients. Notably, the baseline content quality scores of the original documents were relatively low (approximately mid-40s out of 100), and the observed gains therefore represent relative improvements rather than the attainment of an optimal standard. Importantly, the magnitude of these improvements was reflected in the β coefficients derived from the LMMs, which provide interpretable estimates of

XSL•FO
RenderX

practical effect size beyond statistical significance. These improvements suggest that LLMs can function as auxiliary tools in resource-limited higher education settings, enabling the rapid generation of standardized and accessible consent forms to address shortages of trained personnel and professional resources. More importantly, by lowering barriers to comprehension and engagement, LLM-based optimization may facilitate earlier help-seeking, promote more equitable access to counseling services, and strengthen the overall efficiency of youth mental health systems. While rewriting consent forms represents only an initial step, the strong performance observed here underscores the broader potential of LLMs to support standardized [49,51], equitable, and client-centered mental health services for adolescents and young adults worldwide.

In addition, the multidimensional evaluation framework developed in this study offers a methodological reference for future digital mental health research and highlights the value of prompt design and human-AI collaboration in sensitive contexts. Specifically, human-AI collaboration may help address key limitations identified in this study by allowing LLMs to support drafting and structural standardization, while human experts provide contextual judgment, ethical oversight, legal review, and final validation prior to real-world implementation. From a governance perspective, LLM-generated consent materials should therefore be regarded as assistive drafts rather than autonomous outputs, with institutional mechanisms such as mandatory human review and version control playing a central role in mitigating ethical, legal, and accountability risks. In practice, such collaboration may follow a structured workflow, in which LLMs generate an initial draft, counselors or mental health professionals review and refine content for contextual appropriateness and ethical compliance, and institutions provide final oversight and approval prior to deployment.

Limitations

Several limitations should be acknowledged. First, study materials were limited to Chinese-language ICFs from universities in China, which may constrain generalizability. Direct transfer to other settings may be limited by differences in language structure and readability conventions, as well as jurisdiction-specific legal and ethical requirements (eg, confidentiality exceptions, consent norms, and institutional procedures) and broader cultural expectations regarding counseling relationships. Future research should test applicability in English and other languages and adapt the evaluation framework through context-specific calibration of readability metrics and locally relevant ethical and regulatory standards to support cross-cultural use. Second, despite anonymization and randomization, rater blinding may not have been fully preserved. The substantially longer Grok-4–optimized texts may have been visually distinguishable from the originals, potentially introducing unconscious bias. Although efforts were made to reduce this risk, residual expectancy effects cannot be excluded. In addition, potential rater fatigue and sequence bias should be considered. Each expert evaluated 99 documents across 20 indicators, which may have imposed substantial cognitive load. Although the document order was randomized, fatigue-related effects cannot be fully excluded. Third, the reading comprehension assessment included only 10 volunteers, with moderate interrater consistency and no actual clients, underscoring the need for larger, more representative samples, including real audiences. This small sample size also limited statistical power, particularly for detecting subtle differences between LLM models. Future studies should further validate this evaluation framework against real client decision-making processes and counseling outcomes in applied settings. Moreover, university students may not represent individuals with lower health literacy or acute psychological distress, and documents understandable to college volunteers may still pose barriers to more vulnerable populations. Fourth, although our expert ratings captured multidimensional aspects of content quality (eg, completeness, specificity, and operability), the current evaluation did not systematically verify the factual and legal accuracy of LLM-generated content; given the known risk of hallucinations, optimized forms may inadvertently introduce legally incorrect statements, misinterpret jurisdiction-specific regulations (eg, provisions under the Chinese Mental Health Law), or omit critical ethical requirements despite fluent wording. Because no explicit legal texts were provided to the models, such outputs rely entirely on internal training data, which may be outdated or incomplete. This underscores the necessity of mandatory legal and ethical expert review before any real-world implementation. In addition, aggregate mean scores may obscure isolated but clinically significant errors. For example, LLM-generated drafts could omit mandatory clauses or include contextually inaccurate procedural details despite otherwise strong performance. Future studies should incorporate structured qualitative audits to complement quantitative evaluation. Fifth, this study evaluated only 2 closed-source models, ChatGPT (GPT-5) and Grok-4, without considering open-source systems or alternative architectures. In addition, model outputs were generated within a fixed time window, and future updates to these rapidly evolving systems may lead to different results even under identical prompts. As closed-source commercial systems undergo continuous updates, output behavior may shift over time, and future replications may yield different results due to model drift. Moreover, because outputs were generated via consumer-facing web interfaces rather than application programming interfaces with fixed parameters, unrecorded variables may have limited exact replication of the generation environment and further compounded model drift. Future work should therefore broaden model comparisons and incorporate transparent reporting of model versions and access dates to enhance reproducibility. Sixth, although a multidimensional evaluation framework was applied, this study did not conduct an in-depth analysis of specific textual revisions at the case level; future research could incorporate representative before and after examples to complement the quantitative findings. Finally, improvements in operability-related items were limited, suggesting current models remain insufficient in generating explicit execution pathways; this points to directions such as prompt engineering or human-AI collaboration to enhance operability in future work.

XSL•FO
RenderX

Conclusion

This study is the first to systematically evaluate the application of LLMs in optimizing ICFs for university mental health counseling. We found significant improvements in structure, readability, content quality, and comprehension, with consistent recognition from both experts and simulated clients. These findings suggest that LLMs can serve as tools to improve currently deficient counseling consent documentation by enhancing clarity and standardization. However, they may not yet produce fully comprehensive or high-quality consent forms without significant human refinement. For universities or primary institutions with limited professional resources, LLMs may serve as practical auxiliary tools to generate accessible, high-quality materials that facilitate early engagement with counseling. Further research is required to evaluate their safety, accuracy, and practical feasibility in real-world counseling workflows.

Acknowledgments

The authors thank Meiou Wang and Linli Yi for their assistance during the testing phase of this study. We are also grateful to Jiajie Zhang and colleagues for their expert scoring, and to Dongdong Zhu, Yuzhi Zhang, Hongfang Chen, and others for their contributions as readers in the rating process.

Funding

The authors declared no financial support was received for this work.

Data Availability

All data directly supporting the findings of this study are included in Multimedia Appendix 5. Additional underlying data can be made available upon reasonable request to the corresponding author (wwtdoctor02@163.com). Requests from verified academic researchers will be assessed by the Data Access Committee and answered within 1 month, in accordance with institutional and ethical requirements.

Parts of the analysis code used in this study are included in Multimedia Appendix 4. Qualified researchers may request access to the remaining code by contacting the co-corresponding authors (wwtdoctor02@163.com and ljc846702504@163.com). For requests submitted by verified academic researchers, the Data Access Committee will assess the request and grant access within 1 month.

Authors' Contributions

J Luo designed the study; performed code development, statistical analysis, and preliminary testing; and drafted the manuscript. JM coordinated the study and contributed to code development, figure preparation, document collection, preliminary testing, and manuscript writing. DZ, J Li, and YZ collected documents, organized reader and expert ratings, and assisted in manuscript preparation. LZ organized expert evaluations, supervised the study, revised the manuscript, and served as co-corresponding author. WW supervised the study, revised the manuscript, and coordinated submission as corresponding author. J Luo and JM contributed equally as first authors. LZ and WW contributed equally as corresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed results of the analyses.
[PDF File, 822 KB - jmir_v28i1e86502_app1.pdf]

Multimedia Appendix 2
Supplementary information including prompt instructions, vocabulary lists, and bilingual (Chinese-English) design details.
[PDF File, 173 KB - jmir_v28i1e86502_app2.pdf]

Multimedia Appendix 3
Complete set of documents in 3 versions (Original, ChatGPT5-generated, and Grok4-generated).
[PDF File, 23240 KB - jmir_v28i1e86502_app3.pdf]

Multimedia Appendix 4
Source code used for data processing and statistical analysis.
[PDF File, 279 KB - jmir_v28i1e86502_app4.pdf]

XSL•FO
RenderX

Multimedia Appendix 5
Supplementary data including all scoring results across documents and versions.
[PDF File, 4142 KB - jmir_v28i1e86502_app5.pdf]

References

1. Dong J, Chen J, Li Y, Huang X, Rong X, Chen L. Relationship between freshmen's psychological health and family economic status in Chinese universities: a latent profile analysis. Psychol Res Behav Manag 2023;16:3489-3502. [doi: 10.2147/PRBM.S424798] [Medline: 37667735]
2. Osborn TG, Li S, Saunders R, Fonagy P. University students' use of mental health services: a systematic review and meta-analysis. Int J Ment Health Syst 2022 Dec 17;16(1):57. [doi: 10.1186/s13033-022-00569-0] [Medline: 36527036]
3. Deng X, Li C. The impact of university freshmen's mental health on academic performance: an empirical study based on M University in Fujian Province, China. Front Public Health 2024;12:1483031. [doi: 10.3389/fpubh.2024.1483031] [Medline: 39735743]
4. Lin GT, Mitchell MB, Hammack-Aviran C, Gao Y, Liu D, Langerman A. Content and readability of US procedure consent forms. JAMA Intern Med 2024 Feb 1;184(2):214-216. [doi: 10.1001/jamainternmed.2023.6431] [Medline: 38079183]
5. García-Álvarez JM, García-Sánchez A. Readability and comprehension of anesthesia informed consent forms in a Spanish county hospital: an observational study. Nurs Rep 2024 May 24;14(2):1338-1352. [doi: 10.3390/nursrep14020101] [Medline: 38921711]
6. Emanuel EJ, Boyle CW. Assessment of length and readability of informed consent documents for COVID-19 vaccine trials. JAMA Netw Open 2021 Apr 1;4(4):e2110843. [doi: 10.1001/jamanetworkopen.2021.10843] [Medline: 33909052]
7. Gerke L, Ladwig S, Pauls F, Trachsel M, Härter M, Nestoriuc Y. Optimized informed consent for psychotherapy: protocol for a randomized controlled trial. JMIR Res Protoc 2022 Sep 30;11(9):e39843. [doi: 10.2196/39843] [Medline: 36178713]
8. Pietrzykowski T, Smilowska K. The reality of informed consent: empirical studies on patient comprehension-systematic review. Trials 2021 Jan 14;22(1):57. [doi: 10.1186/s13063-020-04969-w] [Medline: 33446265]
9. Fischer AE, Venter WDF, Collins S, Carman M, Lalla-Edward ST. The readability of informed consent forms for research studies conducted in South Africa. S Afr Med J 2021 Feb 1;111(2):180-183. [doi: 10.7196/SAMJ.2021.v111i2.14752] [Medline: 33944731]
10. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA 2023 Sep 5;330(9):866-869. [doi: 10.1001/jama.2023.14217] [Medline: 37548965]
11. Wornow M, Xu Y, Thapa R, et al. The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit Med 2023 Jul 29;6(1):135. [doi: 10.1038/s41746-023-00879-8] [Medline: 37516790]
12. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 2025 Jan 28;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
13. Liu S, McCoy AB, Wright AP, et al. Leveraging large language models for generating responses to patient messages-a subjective analysis. J Am Med Inform Assoc 2024 May 20;31(6):1367-1379. [doi: 10.1093/jamia/ocae052] [Medline: 38497958]
14. Garcia P, Ma SP, Shah S, et al. Artificial intelligence-generated draft replies to patient inbox messages. JAMA Netw Open 2024 Mar 4;7(3):e243201. [doi: 10.1001/jamanetworkopen.2024.3201] [Medline: 38506805]
15. Schmidt J, Lichy I, Kurz T, et al. ChatGPT as a support tool for informed consent and preoperative patient education prior to penile prosthesis implantation. J Clin Med 2024 Dec 10;13(24):7482. [doi: 10.3390/jcm13247482] [Medline: 39768416]
16. Swisher AR, Wu AW, Liu GC, Lee MK, Carle TR, Tang DM. Enhancing health literacy: evaluating the readability of patient handouts revised by ChatGPT's large language model. Otolaryngol Head Neck Surg 2024 Dec;171(6):1751-1757. [doi: 10.1002/ohn.927] [Medline: 39105460]
17. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
18. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne) 2024;11:1477898. [doi: 10.3389/fmed.2024.1477898] [Medline: 39534227]
19. Chung EM, Zhang SC, Nguyen AT, Atkins KM, Sandler HM, Kamrava M. Feasibility and acceptability of ChatGPT generated radiology report summaries for cancer patients. Digit Health 2023;9:20552076231221620. [doi: 10.1177/20552076231221620] [Medline: 38130802]
20. Informed consent form for university counseling [Report in Chinese]. : The Clinical Psychology Registration Working Committee of the Chinese Psychological Society; 2022 URL: https://cspec.hntou.edu.cn/down/zqtys/202204/P020220419320054681253.pdf [accessed 2026-03-10]
21. ChatGPT. URL: https://chatgpt.com [accessed 2026-03-10]
22. Grok. 2025. URL: https://grok.com [accessed 2026-03-10]
23. Edwards A. OpenAI launches GPT-5 and multilingual performance shows little improvement. Slator. 2025. URL: https://slator.com/openai-launches-gpt5 [accessed 2026-03-10]
24. Grok-4 vs Kimi K2 vs Claude 4: a professional comparison of leading AI models in 2025. iWeaver AI. URL: https://www.iweaver.ai/blog/grok-4-vs-kimi-k2-vs-claude-4-professional-comparison-of-leading-ai-models [accessed 2026-03-10]

XSL•FO
RenderX

25. Karimi AH, Guyler MR, Hecht CJ II, Burkhart RJ, Acuña AJ, Kamath AF. Assessing the readability of clinical trial consent forms for surgical specialties. J Surg Res 2024 Apr;296:711-719. [doi: 10.1016/j.jss.2024.01.045] [Medline: 38367522]
26. Mirza FN, Wu E, Abdulrazeq HF, et al. The literacy barrier in clinical trial consents: a retrospective analysis. EClinicalMedicine 2024 Sep;75:102814. [doi: 10.1016/j.eclinm.2024.102814] [Medline: 39763593]
27. Lakretz Y, Hupkes D, Vergallito A, Marelli M, Baroni M, Dehaene S. Mechanisms for handling nested dependencies in neural-network language models and humans. Cognition 2021 Aug;213:104699. [doi: 10.1016/j.cognition.2021.104699] [Medline: 33941375]
28. Chang LC, Sun CC, Chen TH, Tsai DC, Lin HL, Liao LL. Evaluation of the quality and readability of ChatGPT responses to frequently asked questions about myopia in traditional Chinese language. Digit Health 2024;10:20552076241277021. [doi: 10.1177/20552076241277021] [Medline: 39229462]
29. Harrison RM, Lapteva E, Bibin A. Behavioral nudging with generative AI for content development in SMS health care interventions: case study. JMIR AI 2024 Oct 15;3:e52974. [doi: 10.2196/52974] [Medline: 39405108]
30. Ji M, Bouillon P. Predicting the linguistic accessibility of Chinese health translations: machine learning algorithm development. JMIR Med Inform 2021 Oct 7;9(10):e30588. [doi: 10.2196/30588] [Medline: 34617914]
31. Li Y, Zhou X, Zhou Y, et al. Evaluation of the quality and readability of online information about breast cancer in China. Patient Educ Couns 2021 Apr;104(4):858-864. [doi: 10.1016/j.pec.2020.09.012] [Medline: 32988687]
32. Wang Y, Yang X, Liu J. Navigating sensitive conversations: patient-centered communication and politeness markers in Chinese online medical consultations. Healthcare (Basel) 2024 Dec 6;12(23):2465. [doi: 10.3390/healthcare12232465] [Medline: 39685087]
33. Li C, Fu J, Lai J, et al. Construction of an emotional lexicon of patients with breast cancer: development and sentiment analysis. J Med Internet Res 2023 Sep 12;25:e44897. [doi: 10.2196/44897] [Medline: 37698914]
34. Ethical principles of psychologists and code of conduct. American Psychological Association. 2017. URL: https://www.apa.org/ethics/code [accessed 2026-03-10]
35. Code of ethics and conduct. British Psychological Society. 2018. URL: https://www.bps.org.uk/guideline/code-ethics-and-conduct [accessed 2026-03-10]
36. Chinese Psychological Society. URL: http://www.cpsbeijing.org [accessed 2026-03-10]
37. Comprehensive mental health action plan 2013–2030. : World Health Organization; 2021 Sep 10 URL: https://iris.who.int/server/api/core/bitstreams/69921758-6229-49ba-bd3d-c24736e35829/content [accessed 2026-03-10]
38. Steeb T, Reinhardt L, Harlaß M, Heppt MV, Meier F, Berking C. Assessment of the quality, understandability, and reliability of YouTube videos as a source of information on basal cell carcinoma: web-based analysis. JMIR Cancer 2022 Mar 11;8(1):e29581. [doi: 10.2196/29581] [Medline: 35275067]
39. Glaser J, Nouri S, Fernandez A, et al. Interventions to improve patient comprehension in informed consent for medical and surgical procedures: an updated systematic review. Med Decis Making 2020 Feb;40(2):119-143. [doi: 10.1177/0272989X19896348] [Medline: 31948345]
40. Mani NS, Ottosen T, Fratta M, Yu F. A health literacy analysis of the consumer-oriented COVID-19 information produced by ten state health departments. J Med Libr Assoc 2021 Jul 1;109(3):422-431. [doi: 10.5195/jmla.2021.1165] [Medline: 34629971]
41. Sørensen K, Pelikan JM, Röthlin F, et al. Health literacy in Europe: comparative results of the European Health Literacy Survey (HLS-EU). Eur J Public Health 2015 Dec;25(6):1053-1058. [doi: 10.1093/eurpub/ckv043] [Medline: 25843827]
42. Giordano M, Caccavella VM, Tariciotti L, Della Pepa GM, Olivi A, Polli FM. The reliability and quality of online patient education videos for vestibular schwannoma. Br J Neurosurg 2024 Aug;38(4):811-816. [doi: 10.1080/02688697.2021.1973369] [Medline: 34472385]
43. Zhang X, Zhang R. Impact of physicians' competence and warmth on chronic patients' intention to use online health communities. Healthcare (Basel) 2021 Jul 29;9(8):957. [doi: 10.3390/healthcare9080957] [Medline: 34442093]
44. Bucyibaruta JB, Peu D, Bamford L, van der Wath A. Closing the gaps in defining and conceptualising acceptability of healthcare: a qualitative thematic content analysis. Afr Health Sci 2022 Sep;22(3):703-709. [doi: 10.4314/ahs.v22i3.75] [Medline: 36910403]
45. Fang Y, Liu P, Gao Q. Assessment of knowledge, attitude, and practice toward COVID-19 in China: an online cross-sectional survey. Am J Trop Med Hyg 2021 Feb 19;104(4):1461-1471. [doi: 10.4269/ajtmh.20-0452] [Medline: 33606668]
46. Hall CE, Brooks SK, Potts HWW, Greenberg N, Weston D. Rates of, and factors associated with, common mental disorders in homeworking UK Government response employees' during COVID-19: a cross-sectional survey and secondary data analysis. BMC Psychol 2024 Aug 7;12(1):429. [doi: 10.1186/s40359-024-01921-4] [Medline: 39113130]
47. Leroy G, Kauchak D, Haeger D, Spegman D. Evaluation of an online text simplification editor using manual and automated metrics for perceived and actual text difficulty. JAMIA Open 2022 Jul;5(2):ooac044. [doi: 10.1093/jamiaopen/ooac044] [Medline: 35663117]
48. Leroy G, Kauchak D, Harber P, Pal A, Shukla A. Text and audio simplification: human vs. ChatGPT. AMIA Jt Summits Transl Sci Proc 2024;2024:295-304. [Medline: 38827082]
49. Ali R, Connolly ID, Tang OY, et al. Bridging the literacy gap for surgical consents: an AI-human expert collaborative approach. NPJ Digit Med 2024 Mar 8;7(1):63. [doi: 10.1038/s41746-024-01039-2] [Medline: 38459205]

XSL•FO
RenderX

50. Beattie J, Neufeld S, Yang DX, et al. Using large language models to create patient centered consent forms. Int J Radiat Oncol Biol Phys 2024 Oct;120(2):e612. [doi: 10.1016/j.ijrobp.2024.07.1346]
51. Shi Q, Luzuriaga K, Allison JJ, et al. Transforming informed consent generation using large language models: mixed methods study. JMIR Med Inform 2025 Feb 13;13:e68139. [doi: 10.2196/68139] [Medline: 39946400]
52. Oh N, Kim J, Park S, et al. Large language model-assisted surgical consent forms in non-English language: content analysis and readability evaluation. J Med Internet Res 2025 Jun 19;27:e73222. [doi: 10.2196/73222] [Medline: 40537063]
53. Rust P, Frings J, Meister S, Fehring L. Evaluation of a large language model to simplify discharge summaries and provide cardiological lifestyle recommendations. Commun Med (Lond) 2025 May 29;5(1):208. [doi: 10.1038/s43856-025-00927-2] [Medline: 40442348]
54. Romoff M, Brunette M, Peterson MK, Hashmi SZ, Kim MS. The role of large language models in improving the readability of orthopaedic spine patient educational material. J Orthop Surg Res 2025 May 28;20(1):531. [doi: 10.1186/s13018-025-05955-1] [Medline: 40426209]
55. Alber DA, Yang Z, Alyakin A, et al. Medical large language models are vulnerable to data-poisoning attacks. Nat Med 2025 Feb;31(2):618-626. [doi: 10.1038/s41591-024-03445-1] [Medline: 39779928]
56. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature 2024 Jun;630(8017):625-630. [doi: 10.1038/s41586-024-07421-0] [Medline: 38898292]
57. Gelinas L, Morrell W, Tse T, Glazier A, Zarin DA, Bierer BE. Characterization of key information sections in informed consent forms posted on ClinicalTrials.gov. J Clin Transl Sci 2023;7(1):e185. [doi: 10.1017/cts.2023.605] [Medline: 37745937]
58. Will J, Gupta M, Zaretsky J, Dowlath A, Testa P, Feldman J. Enhancing the readability of online patient education materials using large language models: cross-sectional study. J Med Internet Res 2025 Jun 4;27:e69955. [doi: 10.2196/69955] [Medline: 40465378]

Abbreviations

AI: artificial intelligence
ICC: intraclass correlation coefficient
ICF: informed consent form
LLM: large language model
LMM: linear mixed-effects model

Edited by A Coristine; submitted 25.Oct.2025; peer-reviewed by A gopinath, S Singireddy, Z Liu; accepted 24.Feb.2026; published 01.Apr.2026.

Please cite as:
Luo J, Ma J, Zhan D, Zhou Y, Li J, Zhang L, Wang W
Optimization of University Counseling Consent Forms With Large Language Models: Multidimensional Comparative Evaluation
J Med Internet Res 2026;28:e86502
URL: https://www.jmir.org/2026/1/e86502
doi:10.2196/86502

© Jianchen Luo, Jing Ma, Danni Zhan, Yuhong Zhou, Jiayu Li, Lan Zhang, Wentao Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 1.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Well-Being and Cognitive Factors Influencing Health Care Workers' Adherence to Internet-Based Stress Management: Mixed Methods Analysis of a Nonrandomized Controlled Study

Miqi Li[1*], MS; Quan Wang[2*], PhD; Danli Zheng[1], MS; Danni Feng[1], MS; Xiaorong Lang[1], MS

[1]Department of Nursing, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, No.1965 Jiefang Road, Wuhan, Hubei, China

[2]Nursing School, Tong Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China

[*]these authors contributed equally

Corresponding Author:
Danli Zheng, MS
Department of Nursing, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, No.1965 Jiefang Road, Wuhan, Hubei, China

Abstract

Background: High stress levels are common among health care workers (HCWs), threatening their health and workforce stability. Internet-based mobile stress management (MSM) is a promising intervention for reducing work-related stress; however, poor adherence limits effectiveness. Exploring factors influencing HCWs' adherence may thus aid in developing optimal interventions.

Objective: The research aimed to investigate (1) how HCWs' well-being and cognitive factors influenced MSM treatment adherence and (2) what HCWs' specific needs for MSM were.

Methods: This study was a convergent mixed methods secondary analysis of a nonrandomized controlled trial. HCWs who were currently employed, had internet access, had no serious medical problems, and were willing to participate were recruited by convenience sampling through an MSM project in a large Chinese general hospital from August 11, 2021, to January 31, 2022. Those intending to leave the hospital or with insufficient medical condition for follow-up were excluded. Quantitative data were collected from 157 HCWs (n=135, 86% female participants; mean age of 33.7, SD 4.9 y) electronically via Research Electronic Data Capture (REDCap). Measures included sociodemographic characteristics, the Fatigue Assessment Scale, the 14-item Perceived Stress Scale, a user experience questionnaire, an attitudes scale (perceived usefulness, feasibility, and enjoyment), and self-reported practice frequency. Qualitative data were collected via an open-ended question answered by 96 participants. Quantitative data were analyzed using hierarchical regression and structural equation modeling. Qualitative data were analyzed using reflexive-thematic analysis in NVivo (QSR International).

Results: In the quantitative study (n=157), hierarchical regression analyses showed that fatigue was a significant negative predictor of adherence (b=−0.050, 95% CI −0.086 to −0.015; t_{149}=−2.859; 2-tailed P=.005), while user experience (b=0.074, 95% CI 0.042-0.106; t_{148}=4.569; 2-tailed P<.001) and attitudes (b=0.238, 95% CI 0.140-0.336; t_{147}=4.863; 2-tailed P<.001) were positive predictors. Results from the structural equation model demonstrated a significant indirect effect of user experience on adherence through attitudes (indirect effect=0.047, 95% CI 0.021-0.083; P<.001), but no significant direct effect. Separately, qualitative data of responses (n=96) identified four themes about the experience, namely, (1) individualized intervention, (2) effective feedback, (3) reward and constraint mechanisms, and (4) duration of intervention.

Conclusions: This mixed methods study innovatively integrates quantitative and qualitative needs assessment to elucidate the adherence pathway in digital stress management for HCWs. It differentiates itself from prior research by moving beyond merely describing adherence barriers by quantifying the mediating role of attitudes between user experience and treatment adherence. The study indicates that despite fatigue barriers, improving user experience enhances positive attitudes toward MSM, which is vital for ensuring treatment adherence. This contributes to the field by elucidating a key psychological pathway for engagement. For real-world implications, to optimize user experience, future studies should consider focusing on individualized needs, providing effective feedback, developing reward and constraint mechanisms, and designing flexible intervention durations.

Trial Registration: ClinicalTrials.gov NCT05239065; https://clinicaltrials.gov/study/NCT05239065

(J Med Internet Res 2026;28:e73306) doi:10.2196/73306

XSL•FO
RenderX

KEYWORDS

adherence; stress management; internet-based intervention; fatigue; work-related stress; health care workers; user experience

Introduction

Health care workers (HCWs) are often more vulnerable to work-related stress due to multiple stressors that exist in health care contexts, including long working hours, excessive workload, shift work, and inadequate staffing [1-3]. Moreover, the postpandemic era may require HCWs to spend more energy and time on extra work, which places them in a more strained and stressful situation [4,5]. Data released by the World Health Organization in 2025 revealed that one-third of physicians and nurses experience depression or anxiety, with over 10% having experienced suicidal or self-harm ideation [6]. In line with this global pattern, a recent nationwide study among Chinese medical professionals revealed that 28.2% and 24% of participants exhibited major depressive and anxiety symptoms, respectively, with the accumulation of multiple work-related stressors significantly elevating mental health risks [7]. In general, it estimates indicate that the prevalence of stress among HCWs ranges from 29.8% to 63% [8]. The detrimental effects of overwhelming work-related stress on HCWs have been well established, such as mental disorders, sleep disturbance, and physical symptoms [9-11]. Long-term work-related stress may also reduce patient satisfaction and jeopardize patient safety [12,13]. Hence, implementation of stress management interventions is essential in improving the personal health of HCWs and further ensuring the stability of health care human resources.

For now, several stress management techniques, including mindfulness, deep breathing, and regular exercise, have been identified as simple to apply and effective in alleviating both the physical and psychological indices of stress [14-16]. However, adherence to stress management behaviors may not be that easy for HCWs with shift work and irregular working schedules. A recent mixed methods study concluded that cumulative stressors would hinder the application of recommended stress management techniques, resulting in half of them failing to adhere to stress management behaviors [17]. All these findings suggest that it is critical to develop an adaptable and accessible stress management intervention so that HCWs can better use stress management strategies to cope with daily stress.

With the widespread use of digital devices, accumulative research evidence has suggested that mobile stress management (MSM) intervention may produce promising effects on improving stress symptoms and well-being in the workplace [18-20]. However, unlike traditional interventions in which the psychotherapist may supervise the adherence to treatment through face-to-face communications, the adherence of online participants has always been a bottleneck issue in such interventions [21,22]. Recent studies have suggested that personal support might enhance the adherence to internet-based interventions rather than a "pure" self-guided intervention, and the effects varied with the intensity of the intervention [23]. For example, a randomized study found that the adherence rate for stress management in the content-focused guidance sample was 70.5%, higher than both the adherence-focused guidance sample (68.8%) and the administrative guidance sample (42%) [20]. On the one hand, personal guidance may be regarded as an effective approach to improving adherence to internet-based stress management; on the other hand, it would be meaningful to find out factors that influence interindividual differences in nonadherence to such interventions.

Although personal guidance can improve adherence to internet-based stress management, understanding the individual factors underlying nonadherence remains critical. Notably, significant gaps persist regarding how these factors function specifically in MSM interventions targeting HCWs. Previous research identified that well-being might be a critical individual factor that might influence treatment adherence. For example, a qualitative study conducted among general populations identified that high stress levels and fatigue symptoms would hinder individuals from prioritizing and engaging in internet-based interventions [24]. However, another study conducted among HCWs suggested that high workload rather than stress levels may act as a barrier to web-based interventions for modifying their lifestyle, including physical activity [25]. This might be explained by the fact that the appraisal of stress can vary among various people, while some individuals presume stress as a "hindrance," others would identify it as a "challenge" instead [26]. Consequently, the extent to which high stress levels act as a hindrance for health care professionals is not very definite. It would be fascinating to investigate how individual well-being, as expressed by stress and fatigue complaints, impacts HCWs' adherence to MSM.

Meanwhile, cognitive elements of therapy adherence should be considered. Users' perspective on these intervention components or user experience cannot be overlooked. However, research specifically examining how user experience influences adherence to MSM among HCWs remains limited. Because work-related stress can cause a wide range of symptoms, including psychological, physical, and behavioral problems [27-29], those who experience different stress symptoms will require different interventions [30]. According to the Technology Acceptance Model models, individuals will only implement and use an intervention if they believe it is valuable and feasible [31]. To build an effective tailored stress management intervention, it would be required to focus on the user experience and investigate its effects on attitudes and adherence to MSM. This represents a significant gap in the current literature, particularly regarding HCW populations.

To summarize, the current research sought to investigate (1) how HCWs' well-being and cognitive factors influenced their treatment adherence to MSM and (2) what HCWs' requests for MSM were. In particular, we suggested the following research hypotheses:

*H*1*:* Individual well-being has a significant influence on HCWs' treatment adherence, and fatigue has a negative effect, although stress levels do not.

XSL•FO
RenderX

H2: The user experience of HCWs would have significant effects on their perceived feasibility and use.

H3: The attitudes would have significant effects on HCWs' treatment adherence.

H4: The attitudes of HCWs would significantly mediate the impact of their user experience on treatment adherence.

Methods

Study Design

This study was a convergent mixed methods secondary analysis of a nonrandomized controlled trial [32,33]. It aimed to investigate the influence of HCWs' well-being and cognitive variables on their adherence to MSM use. This study was reported in compliance with the Good Reporting of a Mixed Methods Study [34]. We acquired secondary data from HCWs who participated in the MSM program at a large general hospital in China from August 11, 2021, to January 31, 2022 [32,33]. Information on the research intervention can be found on the ClinicalTrials.gov website (NCT05239065).

Inclusion and Exclusion

The inclusion criteria were (1) currently employed, (2) self-reported having internet access, (3) did not have any serious medical problems (eg, cancer or stroke), and (4) were willing to participate in this study and sign the consent form. Participants were excluded if they (1) intended to leave the hospital or units for any reason or (2) were in insufficient medical condition to be followed up on.

Participants Characteristics

Sociodemographic characteristics included sex (male or female), age (y), marriage status (married or unmarried), educational background (undergraduate degree or postgraduate degree), position (physician, nurse, or other), working hours (h), income (Chinese Yuan [CNY]), and BMI.

Sampling Procedures

A convenience sample of HCWs was recruited from a large general hospital in China between August 11, 2021, and January 31, 2022. Participants were enrolled from an ongoing MSM program.

Sample Size, Power, and Precision

This study was a secondary analysis of the original MSM program data. In the primary trial, 179 participants were ultimately allocated to and received the intervention. For the purposes of the current analysis, which designated the Fatigue Assessment Scale (FAS) as a key outcome, we applied a strict criterion of complete data: any record with one or more missing items on the 10-item FAS was excluded to ensure the integrity of the primary outcome measure without imputation. According to the inclusion and exclusion criteria, after applying this data completeness rule, 157 participants were ultimately enrolled and completed the quantitative survey. The sample size for the qualitative component was guided by the principle of information saturation, whereby recruitment continued until no substantially new themes emerged from the data. Ultimately, 96 individuals voluntarily agreed to participate.

Measures and Covariates

Well-Being Factors

Fatigue symptoms were measured using the FAS, a regularly used tiredness measure consisting of 10 questions, was used to assess fatigue symptoms [35]. Each item was graded using a 5-point Likert scale ranging from 0 (never) to 4 (always). Those with a higher FAS total score had more severe fatigue symptoms. The FAS has a Cronbach α of 0.883, suggesting that the internal reliability of this scale was satisfactory in this study.

Perceived stress was assessed using the 14-item Perceived Stress Scale (PSS-14) to gage stress levels at the baseline. PSS-14 is an internationally used self-reported 5-point Likert scale, with a greater total score indicating a higher degree of stress [36]. The Cronbach α was 0.835, indicating that the PSS-14 was a credible instrument in this study for measuring felt stress.

Cognitive Factors

User experience was evaluated using an 8-item questionnaire ("Can you tell us about your experience with this research? [in terms of studying organization, quality of staff for communication or explanation, quality of videos, quality of instructional brochure, quality of education regarding the effects of stress on health, quality of education regarding stress management strategies, quality of weekly feedback reports, and portable device and app applications]") with a 5-point Likert scale (1 [extremely terrible] to 5 [extremely excellent]), and the overall user experience score varied from 8 to 40, with a higher score indicating a better user experience. The Cronbach α coefficient for these 8 items was 0.933, indicating that the 8-item questionnaire was trustworthy for assessing HCW user experience.

Attitudes were assessed on three dimensions: perceived use, perceived feasibility, and perceived enjoyment. Perceived usefulness was defined as "the level of one's belief in using technology," perceived feasibility was defined as "a person's belief in the use of a technology that can be easily used," and perceived enjoyment was defined as "the degree of enjoyment of using a technology," as suggested by the previous study. We specifically investigated HCW sentiments using three items: (1) "Did you find the stress management intervention study beneficial in general?" (perceived use), (2) "Did you find the stress management intervention study to be straightforward to complete?" (feasibility), and (3) "Did you find the stress management intervention study interesting overall?" (perceived enjoyment). The 3 items were graded on a 5-point Likert scale, with higher scores suggesting a more favorable attitude to MSM. The Cronbach α coefficient for these 3 items was 0.713, showing that the 3-item questionnaire was consistent enough to assess attitudes against MSM among HCWs.

Treatment Adherence

Treatment adherence was measured by self-reported practice frequency of three stress management techniques: physical activity, deep breathing exercises, and mindfulness, using a 7-point Likert scale ("Did you practice the stress management

XSL•FO
RenderX

interventions?" 1 [never] to 7 [everyday]). A higher overall score suggested more MSM adherence. The Cronbach α coefficient for these 3 items was 0.793, indicating that the 3-item questionnaire was trustworthy in assessing HCWs' adherence to MSM.

Open-Ended Survey

In order to learn about the unique needs of HCWs when using MSM, we requested HCWs to answer an open-ended question (online interview), "Do you have any suggestions for improving the mobile stress management program?"

Data Collection

Research Electronic Data Capture (REDCap), a web-based online survey tool, was used to capture both quantitative and qualitative data. The MSM was a human-assisted intervention that provided weekly feedback and online stress management education. Stress management training was provided by a neuropsychologist in the form of video clips, posters, and messaging. Participants were encouraged to use at least 1 of the 3 recommended stress management approaches on a weekly basis (physical activity, deep breathing, and mindfulness). Participants were asked to keep a weekly journal that documented their practice frequency and mood state, and they were subsequently given feedback based on their replies. Details of the intervention and data collection have been published previously [32,33].

Statistical Analysis

Overview

This study used a convergent mixed methods design. Qualitative and quantitative data were collected concurrently but analyzed separately. The integration occurred during the interpretation phase through triangulation; the qualitative findings (user-identified needs and preferences) were used to explain and contextualize the quantitative results (statistical predictors of adherence). This approach allowed us to understand not only what factors influence adherence but also why and how they matter from the users' perspective.

Quantitative Data Analysis

IBM SPSS (version 26.0; IBM Corp) and IBM SPSS Amos (version 28.0; IBM Corp) were used to analyze quantitative data. For each categorical item, a descriptive analysis of the response frequencies was performed, and the mean (SD) for the continuous variables was calculated. Then, hierarchical regression analyses were run to examine the relative contributions of demographic variables, well-being factors, and cognitive factors in treatment adherence variances. Four sequential models were tested, with R^2 change indicating the additional variance explained at each step and a 2-side α level of .05.

To test the hypothesized mediation model, structural equation modeling with maximum likelihood estimation was used. The mediation effect of attitudes between user experience and adherence was examined using bootstrapping with 5000 resamples. Model fit was assessed using multiple indices including the χ^2/df ratio (3), root mean square error of approximation (<0.08), standardized root mean square residual (<0.08), comparative fit index (>0.90), and Tucker-Lewis index (>0.90) [17]. The indirect effects were examined, and a significant effect was determined if the 95% CI did not include zero.

Qualitative Data Analysis

QSR NVivo 14.0 was used to analyze qualitative data. The open-ended survey responses were analyzed using a pragmatic approach to reflexive-thematic analysis [37]. The analysis was carried out in six steps: (1) familiarization with the data, (2) code generation, (3) theme construction, (4) review of potential themes, (5) theme definition and naming, and (6) report production. In this stage of the study, 2 experienced researchers (ML and QW) independently read the transcript and coded the findings. Both researchers then met to compare their coding and discuss any discrepancies until a consensus was reached on the thematic structure.

Ethical Considerations

This study was approved by the Institutional Review Board (IRB) of the Medical Ethics Committee of Tongji Medical College, Huazhong University of Science and Technology, China (approval number: IRB number 2021S141). All participants provided informed consent electronically prior to participation. For the secondary analysis reported in this manuscript, the original IRB approval and consent procedures permitted the use of deidentified data. All data were collected and stored anonymously using coded identifiers. No personal identifying information was linked to the survey responses, ensuring participant confidentiality. Participants did not receive any financial compensation for their involvement in the study. This manuscript does not contain any personal images, audio or video recordings, or text excerpts that could lead to the identification of an individual participant.

Results

Participant Characteristics

There were 157 participants in total, with 135 (86%) female participants and 22 (14%) male participants. In terms of age, the mean age was 33.7 (SD 4.9) years, 35 (22.3%) participants were under the age of 30 years, 97 (61.8%) were between the ages of 30 and 40 years, and 25 (15.9%) were over the age of 40 years. In terms of marital status, 115 (73.2%) participants were married, while 42 (26.8%) were single. In terms of position, 30 participants (19.1%) were physicians, 104 (66.2%) were nurses, and 23 (14.7%) were other HCWs. When it came to family income, 18 (11.5%) of the participants had less than 10,000 CNY (1 CNY≈0.155 USD). 81 (51.6%) participants had a family income of 10,000 to 15,000 CNY, 35 (22.3%) had a family income of 15,001 to 20,000 CNY, and 23 (14.6%) had a family income of more than 20,000 CNY. When it came to working hours, 22 (14%) participants worked less than 35 hours per week, 74 (47.1%) worked 35 to 44 hours per week, and 61 (38.9%) worked more than 44 hours per week (Table 1).

XSL•FO
RenderX

Table . Baseline characteristics of participants in the quantitative component of a mixed methods study on mobile stress management adherence (n=157)

Characteristics	Participants, n (%)
Sex	
Female	135 (86)
Male	22 (14)
Age group (y)	
<30	35 (22.3)
30 to 40	97 (61.8)
>40	25 (15.9)
Marital status	
Married	115 (73.2)
Unmarried	42 (26.8)
Educational level	
Undergraduate degree	120 (76.4)
Postgraduate degree	37 (23.6)
Occupation	
Physician	30 (19.1)
Nurse	104 (66.2)
Other health care workers	23 (14.7)
Income (Chinese Yuan)[a]	
<10,000	18 (11.5)
10,000 to 15,000	81 (51.6)
15,001 to 20,000	35 (22.3)
>20,000	23 (14.6)
Working (h)	
<35	22 (14)
35 to 44	74 (47.1)
>44	61 (38.9)
BMI	
≤23.9	120 (76.4)
>23.9	37 (23.6)

[a]Income is reported in Chinese Yuan; the average exchange rate was 1 Chinese Yuan≈0.155 USD during the study period of 2021 - 2022.

Descriptive Statistics

Table 2 displays descriptive results for the main study variables. Participants rated perceived usefulness as 2.96 out of 5.0 (SD 0.85), perceived feasibility as 3.71 out of 5.0 (SD 0.72), and perceived enjoyment as 3.59 out of 5.0 (SD 0.69), in that order. The HCWs rated their experience with MSM as 32.03 out of 40.0 (SD 5.01), and the “quality of weekly feedback reports” received the highest mean score of 4.17 (SD 0.73). Participants rated their adherence to MSM for physical activity as 2.91 out of 5.0 (SD 1.15), deep breathing as 3.25 out of 5.0 (SD 1.34), and mindfulness as 2.92 out of 5.0 (SD 1.33), on average.

XSL•FO
RenderX

Table . Descriptive statistics of attitude, use experience, and adherence to mobile stress management in quantitative component (n=157).

Variables	Mean (SD; range)
Experience[a]	32.03 (5.01; 8-40)
Experience regarding study organization	3.74 (0.71; 1-5)
Quality of staff for communication or explanation	4.08 (0.77; 1-5)
Quality of the videos	4.03 (0.80; 1-5)
Quality of education regarding the effects of stress on health	4.09 (0.77; 1-5)
Quality of education regarding stress management strategies	4.06 (0.77; 1-5)
Quality of weekly feedback reports	4.17 (0.73; 1-5)
Applications of portable device and app	4.01 (0.75; 1-5)
Attitude[a]	10.26 (1.82; 3-15)
Perceived usefulness	2.96 (0.85; 1-5)
Perceived feasibility	3.71 (0.72; 1-5)
Perceived interest	3.59 (0.69; 1-5)
Adherence to mobile stress management[a]	9.15 (3.22; 3-21)
Physical activity	2.97 (1.15; 1-7)
Deep breathing	3.25 (1.34; 1-7)
Mindfulness	2.92 (1.33; 1-7)

[a]Composite score of questionnaire.

Hierarchical Regression Analyses

For the hierarchical regression analysis, 4 models were obtained. As shown in Table 3, the demographic variables (sex, age, marital status, position, income, and working hours) were entered into the model in the first step, and no variable was found to be significant in explaining the variance of treatment adherence. When perceived stress and fatigue symptoms were included in step 2, the percentage of treatment adherence explained by the predictors significantly increased from R^2=0.043 in model 1 to R^2=0.093 in model 2, and fatigue symptom was demonstrated to be a significant predictor (b=−0.050; t_{149}=−2.859; and 2-tailed P=.005). Then, after user experience was included as the predictor in step 3, the percentage of treatment adherence explained by the predictors significantly increased from R^2=0.093 in model 2 to R^2=0.205 in model 3, and user experience was demonstrated to be a significant predictor (b=0.074; t_{148}=4.569; and 2-tailed P<.001). In the last step, when attitude was included as the predictor, the percentage of treatment adherence explained by the predictors significantly increased from R^2=0.205 in model 3 to R^2=0.315 in model 4, while attitude was shown to be a significant predictor (b=0.238; t_{147}=4.863; and 2-tailed P<.001), the significance of user experience was shown to be disappearing in model 4 (b=0.031; t_{147}=1.770; and P=.08). All the predictors of model 4 together explained 31.5% of the variance of treatment adherence (R^2=0.315; P<.001).

XSL•FO
RenderX

Table . Hierarchical regression analysis of mobile stress management adherence (n=157).

Predictors	Model 1 (R^2=0.043; F test (5, 151)=1.385)			Model 2 (R^2=0.093; F test (7, 149)=2.182[a])			Model 3 (R^2=0.205; F test (8, 148)=4.774[b])			Model 4 (R^2=0.315; F test (9, 147)=7.520[b])		
	b [c]	β[d]	2-tailed t test (151)	b	β	2-tailed t test (149)	b	β	2-tailed t test (148)	b	β	2-tailed t test (147)
Constant	1.750	—[e]	3.773[b]	2.505	—[e]	3.789[b]	0.076	—[e]	0.093	−0.814	—[e]	−1.040
Sex	0.095	0.031	0.359	0.100	0.032	0.383	0.174	0.056	0.708	0.082	0.027	0.358
Age	0.185	0.084	0.997	0.170	0.078	0.934	0.078	0.036	0.452	−0.068	−0.031	−0.416
Position	0.113	0.131	0.865	0.051	0.035	0.389	0.002	0.001	0.014	0.033	0.023	0.292
Income	−0.185	−0.151	−1.785	−0.173	−0.141	−1.701	−0.119	−0.097	−1.241	−0.052	−0.042	−0.573
Working hours	0.151	0.069	0.794	0.167	0.076	0.883	0.192	0.088	1.084	0.174	0.079	1.054
Perceived stress	—	—	—	0.026	0.152	1.634	0.017	0.100	1.140	0.013	0.076	0.921
Fatigue symptoms	—	—	—	−0.050	−0.264	−2.859[f]	−0.041	−0.214	−2.452[f]	−0.038	−0.201	−2.468[f]
User experience	—	—	—	—	—	—	0.074	0.345	4.569[b]	0.031	0.144	1.770
Attitudes	—	—	—	—	—	—	—	—	—	0.238	0.402	4.863[b]

[a]P<0.05.

[b]P<0.001.

[c]b=unstandardized coefficient.

[d]β=standardized coefficient.

[e]Not applicable.

[f]P<.01.

Mediation Analysis

According to Hu and Bentler's [38] suggested model fit indicators, the model in this paper had a good fitting degree. Table 4 shows that the χ^2/df values were less than 3, the root mean square error of approximation and standardized root mean square residual values were less than 0.80, and the nonnormed fit index, comparative fit index, and adjusted goodness of fit index values were all greater than 0.90, all of which were within the range of acceptable fit indexes. To summarize, the results of confirmatory factor analysis confirm that the empirical model comprising attitude, experience, and adherence has good validity; the collected data are highly reliable; and the research findings are robust.

XSL•FO
RenderX

Table . Goodness of fit indices for the structural equation modeling testing the mediation effect of attitudes.

Index	Goodness-of-fit measure		Model fit
	Good fit	Accept fit	
χ^2/df	$0 \leq \chi^2/df \leq 2$	$2 < \chi^2/df \leq 3$	1.640
P value	$0.05 < P \leq 1.00$	$0.01 < P \leq 0.05$	0.073
RMSEA[a]	0≤ RMSEA ≤0.05	0.05< RMSEA ≤0.08	0.064
SRMR[b]	0≤ SRMR ≤0.05	0.05< SRMR ≤0.08	0.061
NNFI[c]	0.97< NNFI ≤1.00	0.95≤ NNFI ≤0.97	0.951
CFI[d]	0.97< CFI ≤1.00	0.95≤ CFI ≤0.97	0.980
AGFI[e]	0.90< AGFI ≤1.00	0.95≤ AGFI ≤0.97	0.928

[a]RMSEA: root mean square error of approximation.

[b]SRMR: standardized root mean square residual.

[c]NNFI: nonnormed fit index.

[d]CFI: comparative fit index.

[e]AGFI: adjusted goodness of fit index.

Figure 1 depicts the results of attitude mediation between experience and adherence to MSM. The path coefficient from experience to attitude was 0.079 (*P*<.001), and the path coefficient from attitude to adherence was 0.568 (*P*<.001). Meanwhile, the Bootstrapping estimates revealed that the direct influence of experience on adherence was not significant (direct effect=0.002; *P*=.86), implying that experience had a full indirect effect on adherence via the attitude toward MSM (indirect effect=0.047; *P*<.001).

Figure 1. The mediating role of attitudes in the relationship between user experience and adherence to mobile stress management.

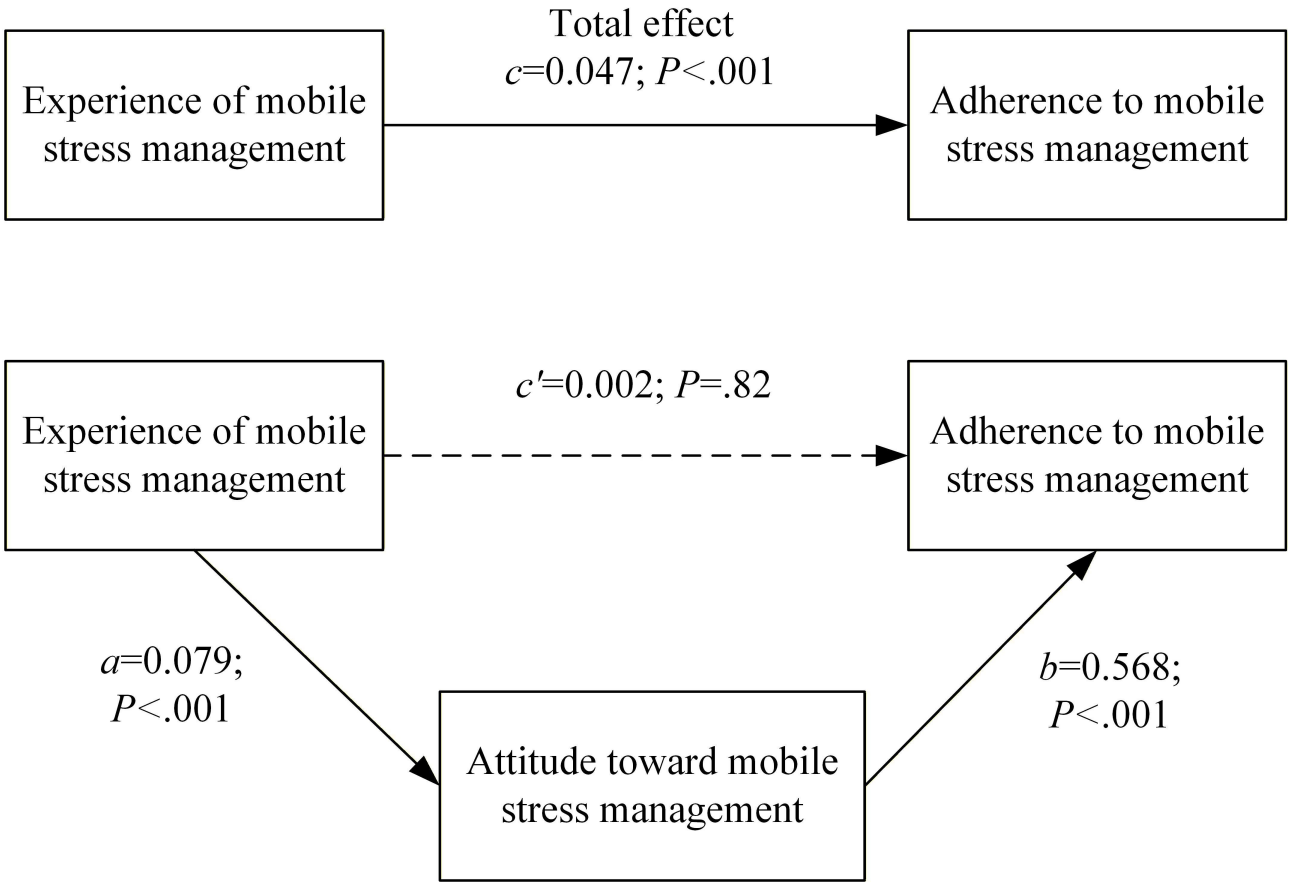

Results of Open-Ended Survey Response

The open-ended survey received 96 responses from 157 participants, which were identified and categorized into four themes: (1) individualized intervention, (2) effective feedback, (3) reward and constraint mechanisms, and (4) duration of intervention. Each theme could have 3 subthemes. Table 5 summarizes the themes, subthemes, and sample quotes.

XSL•FO
RenderX

Table . Themes, subthemes, and sample quotes from the open-ended survey responses regarding needs for improving mobile stress management in qualitative analyses (n=96).

Themes and subthemes	Sample quotes
Individualized intervention	
Diversified modules	• *"Everyone's stress sources are different and require one-on-one counseling from professionals."* (Participant 79) • *"I want to learn more about how to alleviate my anxiety."* (Participant 118) • *"I find it hard to change my sleep quality."* (Participants 83 and 107) • *"I want to know how to reduce negative emotions and deal with doctor-patient conflicts."* (Participant 100)
Adapt to work schedules	• *"I often work too long hours, and I find it difficult to guarantee enough time to practice those techniques."* (Participant 9) • *"I didn't have time to exercise."* (Participant 26)
Staging the intervention	• *"It's hard to get started."* (Participant 137) • *"It's hard to insist on."* (Participant 14, 32, 50, 128) • *"Professionals are required to follow up and guide the practice process."* (Participant 79)
Effective feedback	
Feedback content	• *"The content of the feedback should be more specific to the weekly plan and monthly schedule, which would be more intuitive."* (Participant 68) • *"The weekly questionnaire should be adapted to each individual's situation."* (Participant 41) • *"It would be better if we could get summarized and analyzed feedback data on a weekly or biweekly basis."* (Participant 147)
Feedback frequency	• *"More frequent reminders are needed."* (Participant 143) • *"We need more supervision and feedback."* (Participant 117)
Reward or constraint mechanisms	
Motivations for practice	• *"Giving rewards to stimulate movement would be useful."* (Participant 108) • *"It would work by motivating everyone to exercise through group check-ins."* (Participant 50) • *"It's important to mobilize the wisdom of everyone in the group and share more methods, skills, and experiences together."* (Participant 156)
More restraints	• *"It is completely dependent on self-consciousness and self-restraint."* (Participant 158) • *"It would be best if participants could be supervised and demanded a little more."* (Participant 79) • *"A little more restraint is needed to help us form habits."* (Participant 96)
Duration of intervention	
Short-term intervention	• *"Reduce the experimental cycle."* (Participant 18)
Long-term intervention	• *"I hope the project will continue for a longer time."* (Participant 54, 90)

Discussion

Principal Findings

This research focused on how individual treatment adherence differed depending on their well-being and cognitive factors. By using a convergent mixed methods approach, we specifically examined the direct effects of well-being factors on treatment adherence and the indirect effects of cognitive factors on treatment adherence, as well as learning about the specific requirements of HCWs for MSM interventions.

Overall, adherence to the internet-based MSM was generally modest, with mindfulness and physical activity techniques showing lower adherence compared to deep breathing exercises. The results are consistent with previous studies [20,22], despite

XSL•FO
RenderX

the reduction in dropouts in the internet-based intervention compared to face-to-face, participants often fail to have complete adherence. While physical activity and mindfulness are evidence-based techniques, their lower adherence in our sample may relate to the demanding work schedules of HCWs [8,16,19,39]. A previous study showed that breathing exercises can reduce stress levels even in the short term (mean time of 1.9, SD 1.2 min) [40], which means that breathing exercises may be more available for HCWs who spend most of their time providing health care. HCWs are a special group of people with heavy workloads and limited job autonomy. The time spent setting requirements and the accessibility of management programs were priorities in stress management for this group of people. Therefore, in the future, the stress management of specific interventions needs to be based on the working conditions such as the working hours or environment of different HCWs, taking into account the feasibility of developing a personalized management plan to improve adherence.

On the issue of burnout symptoms, this study revealed that fatigue symptoms were a significant predictor rather than perceived stress. An explanation for this finding is that perceived stress did not reach a statistically significant level in relation to burnout. One possible reason for this could be attributed to insufficient training hours, while another factor may be the nurses' anticipation of encountering stressful work in their future shifts. Additionally, according to Maslow's hierarchy of needs theory, fatigue symptoms represent fundamental and immediate needs, whereas perceived stress can potentially be managed through various methods by HCWs. As professional practitioners, HCWs consciously adapt themselves to reduce stress levels; however, it is important to consider potential biases resulting from psychological effects during the design phase of this study.

Our findings regarding the prominence of fatigue over perceived stress align with broader postpandemic research on psychological well-being, which suggests a shift in the salient components of well-being. Babazadeh [41], using a multidimensional measure of psychological well-being, found that factors like health locus of control and social support were significant predictors, underscoring the importance of targeting specific well-being components and cognitive factors in interventions rather than focusing solely on global stress measures. The significant role of fatigue in our study may reflect its centrality within the somatic symptoms dimension of well-being, which appears highly relevant for HCWs. Furthermore, fatigue and burnout may have intertwined effects, as evidenced by a prospective cohort study reporting that individuals in the "burnout and fatigue" group had a higher proportion of chronic course type compared to those experiencing "pure burnout" or "pure fatigue." Severe burnout and fatigue symptoms become interconnected with each other as well as with other psychological and/or somatic complaints [42]. However, the specific association between fatigue and specific factors remains unclear [43].

The most interesting finding is that the attitude mediates experience and adherence. Our mediation analysis revealed a significant indirect effect of user experience on adherence through attitudes, while the direct effect was not significant. Most existing studies have emphasized the importance of user experience on treatment adherence [8,20]. However, the direct relationship between the experience and adherence is limited to a definite explanation. On the contrary, the discussion of attitude and adherence is a classic issue, and the amount of evidence reports the relationship between attitude and adherence in stress management interventions [8,44,45]. Our research indicates that the direct impact of user experience on treatment adherence disappeared after inputting the variable of attitudes. An implication of the further relationship is the possibility that the improvement of experience with the object of adherence increasing is to make the attitude of the internet-based mobile program or management positive, despite the weak significant association between the experience and attitude. This may provide another perspective on retention in stress management, and a similar relationship in other psychological health of HCWs should be developed in the future.

We also identify the 4 aspects of individualized intervention, effective feedback, reward or constraint mechanisms, and duration of intervention as certain factors to adherence to internet-based MSM. Generally, all the factors highlight the personalized service throughout the whole intervention. The high quality and individual needs of intervention contents are more vital; inadequately supported contents are a hindrance. Furthermore, HCWs, as a special population with a health care education background, face the problem of their personal situations, may not have time to integrate exercises into their daily lives, and may need more personal conversations with the eCoach during the staging and feedback stages. This is in line with previous observations [46-48]. In the reward or constraint mechanisms, HCWs required the increasing demands outwards, which addressed the role of the organization in the interventions. Proving reminders, encouraging each other, and using rewards as the external triggers were taken as effective measures to increase adherence in other evidence [49,50]. To the duration of intervention, extensive studies have pointed out that stress was reduced with the increasing duration of intervention [9,51,52]. Considering that health care often requires demanding physical and psychological efforts, HCWs often have little energy to implement stress management strategies on a daily basis. Therefore, future studies may need to be more focused on less exhausting interventions for those working in health care contexts.

Limitations

Some limitations in the paper also need to be taken into consideration. First, the exploratory study investigated a small sample size with a short-term intervention. Second, the data collection of qualitative research conducted online could result in lacking information. Furthermore, the implication of this paper is that participants in qualitative research may have been positively biased in their support of the intervention because all the interviewees finished the intervention. It would be valuable in future work to extend the sample, focus on those who chose to drop out of the intervention, and explore the reasons for dropping out.

XSL•FO
RenderX

Implications

The findings of this study highlight a central pathway for engagement in digital interventions and point to clear practical directions. By innovatively integrating quantitative modeling with qualitative assessment, this research moves beyond merely describing barriers to adherence. It thereby differentiates itself from prior work by quantitatively establishing the critical mediating role of participant attitudes between user experience and sustained use. This contributes a key insight to the field: successful engagement depends not just on the intervention itself, but on fostering positive attitudes through that experience. Consequently, the real-world implication is straightforward. To optimize adherence, MSM programs for HCWs must prioritize enhancing user experience via personalized content, actionable feedback, incentive mechanisms, and flexible duration, all essential for cultivating the positive attitudes needed for long-term use in high-demand settings.

Conclusion

In summary, by innovatively integrating quantitative modeling with qualitative assessment, this study demonstrates that the value of user experience in stress management adherence to internet-based mobile intervention is fully mediated by participant attitudes, highlighting a key insight that differentiates our work from previous barrier-focused research. For HCWs, especially in high-demand settings, adherence is enhanced by interventions that are perceived as personalized, with actionable feedback, incentive mechanisms, and flexible duration. Future research should incorporate additional mediating variables through larger-scale and longer-term studies to further elucidate these mechanisms.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing. The GAI tool used was DeepSeek-V3.1 Terminus. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors do not bear responsibility for the final outcomes. Declaration submitted by ML on behalf of all the coauthors.

Funding

This work was supported by the National Key R&D Program of China (2021YFC2502200). ML is additionally supported by Scientific Research Fund of Tongji Hospital (TJ2025D05). The funders were not involved in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: ML, QW, DZ
Data curation: ML, QW, DF, XL
Formal analysis: ML, QW
Investigation: ML, QW, DF, XL
Methodology: ML, QW, DF, XL
Project administration: ML, QW, DZ
Supervision: ML, DZ
Validation: ML, QW, DZ
Writing- original draft: ML, QW
Writing-review & editing: ML, QW, DF, XL, DZ

Conflicts of Interest

None declared.

References

1. Joseph B, Joseph M. The health of the healthcare workers. Indian J Occup Environ Med 2016;20(2):71-72. [doi: 10.4103/0019-5278.197518] [Medline: 28194078]
2. Medisauskaite A, Kamau C. Prevalence of oncologists in distress: systematic review and meta-analysis. Psychooncology 2017 Nov;26(11):1732-1740. [doi: 10.1002/pon.4382] [Medline: 28116833]
3. Reith TP. Burnout in United States healthcare professionals: a narrative review. Cureus 2018 Dec 4;10(12):e3681. [doi: 10.7759/cureus.3681] [Medline: 30761233]
4. Caruso CC. Negative impacts of shiftwork and long work hours. Rehabil Nurs 2014;39(1):16-25. [doi: 10.1002/rnj.107] [Medline: 23780784]

XSL•FO
RenderX

5. Rosa D, Terzoni S, Dellafiore F, Destrebecq A. Systematic review of shift work and nurses' health. Occup Med (Lond) 2019 Jun 24;69(4):237-243. [doi: 10.1093/occmed/kqz063] [Medline: 31132107]
6. Healing hands – hurting minds. World Health Organization. 2025. URL: https://www.who.int/europe/news/item/10-10-2025-healing-hands---hurting-minds [accessed 2025-11-17]
7. Chen L, Wu L, Xia J, Cai X, Chen L. Gender differences of the association between work-related stressors and mental health among Chinese medical professionals: a nationwide cross-sectional study. Front Public Health 2025;13:1561588. [doi: 10.3389/fpubh.2025.1561588] [Medline: 40265062]
8. Vizheh M, Qorbani M, Arzaghi SM, Muhidin S, Javanmard Z, Esmaeili M. The mental health of healthcare workers in the COVID-19 pandemic: a systematic review. J Diabetes Metab Disord 2020;19(2):1967-1978. [doi: 10.1007/s40200-020-00643-9] [Medline: 33134211]
9. Bischoff LL, Otto AK, Hold C, Wollesen B. The effect of physical activity interventions on occupational stress for health personnel: a systematic review. Int J Nurs Stud 2019 Sep;97:94-104. [doi: 10.1016/j.ijnurstu.2019.06.002] [Medline: 31234106]
10. George LS, Lais H, Chacko M, Retnakumar C, Krishnapillai V. Motivators and barriers for physical activity among health-care professionals: a qualitative study. Indian J Community Med 2021;46(1):66-69. [doi: 10.4103/ijcm.IJCM_200_20] [Medline: 34035579]
11. Philbrick G, Sheridan NF, McCauley K. An exploration of New Zealand mental health nurses' personal physical activities. Int J Ment Health Nurs 2022 Jun;31(3):625-638. [doi: 10.1111/inm.12981] [Medline: 35166003]
12. Conversano C, Ciacchini R, Orrù G, Di Giuseppe M, Gemignani A, Poli A. Mindfulness, compassion, and self-compassion among health care professionals: what's new? A systematic review. Front Psychol 2020;11:1683. [doi: 10.3389/fpsyg.2020.01683] [Medline: 32849021]
13. Goyal M, Singh S, Sibinga EMS, et al. Meditation programs for psychological stress and well-being: a systematic review and meta-analysis. JAMA Intern Med 2014 Mar;174(3):357-368. [doi: 10.1001/jamainternmed.2013.13018] [Medline: 24395196]
14. Heber E, Ebert DD, Lehr D, et al. The benefit of web- and computer-based interventions for stress: a systematic review and meta-analysis. J Med Internet Res 2017 Feb 17;19(2):e32. [doi: 10.2196/jmir.5774] [Medline: 28213341]
15. Stanic J, Barth J, Danon N, Bondolfi G, Jermann F, Eicher M. Adherence to standardized 8-week mindfulness-based interventions among women with breast or gynecological cancer: a scoping review. J Psychosoc Oncol Res Pract 2021;3(2):e048. [doi: 10.1097/OR9.0000000000000048]
16. Taylor H, Cavanagh K, Field AP, Strauss C. Health care workers' need for headspace: findings from a multisite definitive randomized controlled trial of an unguided digital mindfulness-based self-help app to reduce healthcare worker stress. JMIR mHealth uHealth 2022 Aug 25;10(8):e31744. [doi: 10.2196/31744] [Medline: 36006668]
17. Ebert DD, Kählke F, Buntrock C, et al. A health economic outcome evaluation of an internet-based mobile-supported stress management intervention for employees. Scand J Work Environ Health 2018 Mar 1;44(2):171-182. [doi: 10.5271/sjweh.3691] [Medline: 29144535]
18. Ebert DD, Franke M, Zarski AC, et al. Effectiveness and moderators of an internet-based mobile-supported stress management intervention as a universal prevention approach: randomized controlled trial. J Med Internet Res 2021 Dec 22;23(12):e22107. [doi: 10.2196/22107] [Medline: 34941541]
19. Xia Y, Yang Y. RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: the story they tell depends on the estimation methods. Behav Res Methods 2019 Feb;51(1):409-428. [doi: 10.3758/s13428-018-1055-2] [Medline: 29869222]
20. Zarski AC, Lehr D, Berking M, Riper H, Cuijpers P, Ebert DD. Adherence to internet-based mobile-supported stress management: a pooled analysis of individual participant data from three randomized controlled trials. J Med Internet Res 2016 Jun 29;18(6):e146. [doi: 10.2196/jmir.4493] [Medline: 27357528]
21. Bijelic T, Ekberg E, Willman A, Nilsson IM. Expectations and experiences of internet-based therapy for adolescents with TMD pain. J Oral Facial Pain Headache 2022;36(3-4):237-252. [doi: 10.11607/ofph.3042] [Medline: 36445908]
22. Johansson M, Marcusson-Clavertz D, Gunnarsson C, Olsson I, Kaldo V, Bratt A. Feasibility and preliminary evaluation of internet-based compassion and cognitive-behavioral stress-management courses for health care professionals: a randomized controlled pilot trial. Internet Interv 2022;30:100574. [doi: 10.1016/j.invent.2022.100574] [Medline: 36185345]
23. Todkill D, Powell J. Participant experiences of an internet-based intervention and randomised control trial: interview study. BMC Public Health 2013 Oct 28;13(1):1017. [doi: 10.1186/1471-2458-13-1017] [Medline: 24165325]
24. Zavorotnyy M, Ehrlich F, Nenadic I. Health-related Internet use and treatment adherence: a transdiagnostic comparison of outpatients with major depressive disorder and schizophrenia. Psych J 2020 Apr;9(2):174-184. [doi: 10.1002/pchj.355] [Medline: 32189452]
25. Wu D, Lowry PB, Zhang D, Tao Y. Patient trust in physicians matters-understanding the role of a mobile patient education system and patient-physician communication in improving patient adherence behavior: field study. J Med Internet Res 2022 Dec 20;24(12):e42941. [doi: 10.2196/42941] [Medline: 36538351]

XSL•FO
RenderX

26. Weisel KK, Zarski AC, Berger T, et al. User experience and effects of an individually tailored transdiagnostic internet-based and mobile-supported intervention for anxiety disorders: mixed-methods study. J Med Internet Res 2020 Sep 16;22(9):e16450. [doi: 10.2196/16450] [Medline: 32936085]
27. Cavanagh K. Turn on, tune in and (don't) drop out: engagement, adherence, attrition, and alliance with internet-based interventions oxford guide to low intensity CBT interventions. In: Oxford Guide to Low Intensity CBT Interventions: Oxford University Press; 2010:227-234. [doi: 10.1093/med:psych/9780199590117.003.0021]
28. Coelhoso CC, Tobo PR, Lacerda SS, et al. A new mental health mobile app for well-being and stress reduction in working women: randomized controlled trial. J Med Internet Res 2019 Nov 7;21(11):e14269. [doi: 10.2196/14269] [Medline: 31697244]
29. Li J, Riedel N, Barrech A, et al. Long-term effectiveness of a stress management intervention at work: a 9-year follow-up study based on a randomized wait-list controlled trial in male managers. Biomed Res Int 2017;2017:2853813. [doi: 10.1155/2017/2853813] [Medline: 29181392]
30. Knekt P, Virtala E, Härkänen T, Vaarama M, Lehtonen J, Lindfors O. The outcome of short- and long-term psychotherapy 10 years after start of treatment. Psychol Med 2016 Apr;46(6):1175-1188. [doi: 10.1017/S0033291715002718] [Medline: 26755201]
31. Herr RM, Barrech A, Riedel N, Gündel H, Angerer P, Li J. Long-term effectiveness of stress management at work: effects of the changes in perceived stress reactivity on mental health and sleep problems seven years later. Int J Environ Res Public Health 2018 Feb 3;15(2):255. [doi: 10.3390/ijerph15020255] [Medline: 29401657]
32. Liu W, Wang Q, Zheng D, et al. The effects of a complex interactive multimodal intervention on personalized stress management among health care workers in China: nonrandomized controlled study. J Med Internet Res 2024 Jul 12;26:e45422. [doi: 10.2196/45422] [Medline: 38996333]
33. Wang Q, Collet JP, Mei J, et al. Complex interactive multimodal intervention to improve personalized stress management among healthcare workers in China: a knowledge translation protocol. Digit Health 2023;9:20552076231184052. [doi: 10.1177/20552076231184052] [Medline: 37545629]
34. O'Cathain A, Murphy E, Nicholl J. The quality of mixed methods studies in health services research. J Health Serv Res Policy 2008 Apr;13(2):92-98. [doi: 10.1258/jhsrp.2007.007074] [Medline: 18416914]
35. Mihalits DS, Schiller B, Barrech A, et al. The flipside of work engagement: a qualitative evaluation of a stress management intervention in the workplace. Hu Arenas 2023 Jun;6(2):358-372. [doi: 10.1007/s42087-021-00209-y]
36. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [doi: 10.2307/2136404] [Medline: 6668417]
37. Campbell K, Orr E, Durepos P, et al. Reflexive thematic analysis for applied qualitative health research. Qual Rep 2021;26(6):2011-2028. [doi: 10.46743/2160-3715/2021.5010]
38. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Modeling 1999 Jan;6(1):1-55. [doi: 10.1080/10705519909540118]
39. Zhang M, Murphy B, Cabanilla A, Yidi C. Physical relaxation for occupational stress in healthcare workers: a systematic review and network meta-analysis of randomized controlled trials. J Occup Health 2021 Jan;63(1):e12243. [doi: 10.1002/1348-9585.12243] [Medline: 34235817]
40. Kim J, Gray JA, Johnson H. The effect of a web-based deep breathing app on stress of direct care workers: uncontrolled intervention study. J Altern Complement Med 2021 Oct;27(10):876-883. [doi: 10.1089/acm.2020.0541] [Medline: 34227854]
41. Babazadeh T, Ghaffari-Fam S, Shahnavaz-Yoshanluie F, Ranjbaran S. Psychological well-being and factors affecting it after the COVID-19 pandemic. Front Psychol 2024;15:1295774. [doi: 10.3389/fpsyg.2024.1295774] [Medline: 38390417]
42. Leone SS, Huibers MJH, Knottnerus JA, Kant I. A comparison of the course of burnout and prolonged fatigue: a 4-year prospective cohort study. J Psychosom Res 2008 Jul;65(1):31-38. [doi: 10.1016/j.jpsychores.2008.03.018] [Medline: 18582609]
43. Van Dam A, Keijsers G, Verbraak M, Eling P, Becker E. Level and appraisal of fatigue are not specific in burnout. Clin Psychol Psychother 2015;22(2):133-141. [doi: 10.1002/cpp.1869] [Medline: 24022877]
44. De Kock JH, Latham HA, Cowden RG. The mental health of healthcare workers during the COVID-19 pandemic: a narrative review. Curr Opin Psychiatry 2022 Sep 1;35(5):311-316. [doi: 10.1097/YCO.0000000000000805] [Medline: 35855506]
45. Dong F, Liu HL, Yang M, et al. Immediate psychosocial impact on healthcare workers during COVID-19 pandemic in China: a systematic review and meta-analysis. Front Psychol 2021;12:645460. [doi: 10.3389/fpsyg.2021.645460] [Medline: 34122233]
46. De Sio S, Letizia C, Petramala L, et al. Work-related stress and cortisol levels: is there an association? Results of an observational study. Eur Rev Med Pharmacol Sci 2018 Dec;22(24):9012-9017. [doi: 10.26355/eurrev_201812_16672] [Medline: 30575947]
47. Franck E, Haegdorens F, Goossens E, et al. The role of coping behavior in healthcare workers' distress and somatization during the COVID-19 pandemic. Front Psychol 2021;12:684618. [doi: 10.3389/fpsyg.2021.684618] [Medline: 34367005]
48. Pappa S, Sakkas N, Sakka E. A year in review: sleep dysfunction and psychological distress in healthcare workers during the COVID-19 pandemic. Sleep Med 2022 Mar;91:237-245. [doi: 10.1016/j.sleep.2021.07.009] [Medline: 34334303]

XSL•FO
RenderX

49. Panagioti M, Geraghty K, Johnson J, et al. Association between physician burnout and patient safety, professionalism, and patient satisfaction: a systematic review and meta-analysis. JAMA Intern Med 2018 Oct 1;178(10):1317-1331. [doi: 10.1001/jamainternmed.2018.3713] [Medline: 30193239]
50. Montgomery A, van der Doef M, Panagopoulou E, Leiter MP. Connecting Healthcare Worker Well-Being, Patient Safety and Organisational Change The Triple Challenge: Springer Cham; 2020. [doi: 10.1007/978-3-030-60998-6]
51. Bartlett L, Martin A, Neil AL, et al. A systematic review and meta-analysis of workplace mindfulness training randomized controlled trials. J Occup Health Psychol 2019 Feb;24(1):108-126. [doi: 10.1037/ocp0000146] [Medline: 30714811]
52. Zaccaro A, Piarulli A, Laurino M, et al. How breath-control can change your life: a systematic review on psycho-physiological correlates of slow breathing. Front Hum Neurosci 2018;12:353. [doi: 10.3389/fnhum.2018.00353] [Medline: 30245619]

Abbreviations

FAS: Fatigue Assessment Scale
HCW: health care worker
IRB: Institutional Review Board
MSM: mobile stress management
PSS-14: 14-item Perceived Stress Scale
REDCap: Research Electronic Data Capture

Edited by S Brini; submitted 01.Mar.2025; peer-reviewed by J Maciejewski, Q Wang, S Ranjbaran; revised version received 13.Jan.2026; accepted 16.Jan.2026; published 02.Apr.2026.

Please cite as:
Li M, Wang Q, Zheng D, Feng D, Lang X
Well-Being and Cognitive Factors Influencing Health Care Workers' Adherence to Internet-Based Stress Management: Mixed Methods Analysis of a Nonrandomized Controlled Study
J Med Internet Res 2026;28:e73306
URL: https://www.jmir.org/2026/1/e73306
doi:10.2196/73306

© Miqi Li, Quan Wang, Danli Zheng, Danni Feng, Xiaorong Lang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Changes in Workplace Productivity and Estimated Cost Savings During Internet-Based Cognitive Behavioral Therapy in the Irish National Health Service: Naturalistic, Repeated-Measures, Retrospective Survey Study

Chi Tak Lee[1,2,3,4*], PhD; Siobhán Harty[1,2,3,4*], PhD; Angel Enrique[4], PhD; Alba Jiménez-Díaz[5], MS; Garrett Hisler[4*], PhD; Daniel Duffy[4], PhD; Derek Richards[1,4], PhD

[1]School of Psychology, Trinity College Dublin, Dublin, Ireland

[2]Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland

[3]Global Brain Health Institute, Trinity College Dublin, Dublin, Ireland

[4]Amwell Science, Amwell, 1 Stephen Street Upper, Dublin, Ireland

[5]Department of Personality, Evaluation and Psychological Treatment, Faculty of Psychology, University of Valencia, Valencia, Spain

[*]these authors contributed equally

Corresponding Author:
Garrett Hisler, PhD
Amwell Science, Amwell, 1 Stephen Street Upper, Dublin, Ireland

Abstract

Background: Depression and anxiety can significantly impact workplace productivity, for instance, by increasing absenteeism and presenteeism. This loss of productivity leads to diminished workplace economic outcomes. Internet-based cognitive behavioral therapy (iCBT) has emerged as a cost-effective intervention within workplace settings that improves workplace productivity loss due to depression and anxiety, but more generalizable evidence beyond the workplace, such as in a national health service setting, is lacking.

Objective: This naturalistic, repeated-measures, retrospective study investigated the impact of iCBT on work productivity metrics using nationally representative data from patients enrolled in the Irish national health service (ie, the Health Service Executive).

Methods: We analyzed repeated measures retrospective data from 7125 employed patients enrolled in iCBT at the Health Service Executive between March 2023 and May 2024. The Work Productivity and Activity Impairment questionnaire was used to measure absenteeism, presenteeism, overall productivity loss, and activity impairment. Secondary outcomes included depression (Patient Health Questionnaire-9) and anxiety (Generalized Anxiety Disorder-7). Patients were primarily 25 to 64 years old (n=5578, 78%), female (n=4956, 70%), and met clinical scoring criteria on the Patient Health Questionnaire-9 or Generalized Anxiety Disorder-7 (n=4774, 67%). Missing data were handled using multiple imputation. We used mixed-effects models to assess pre-post treatment changes in outcomes and then utilized Irish national salary estimates from 2022 to derive cost savings (in 2022 € values; €1=approximately US $1.05) based on productivity improvement during use of the iCBT program.

Results: From baseline to follow-up, absenteeism reduced by 6.85% (P<.001, 95% CI 5.79% - 7.91%, Cohen d=0.21), presenteeism reduced by 5.84% (P<.001, 95% CI 4.59% - 7.09%, Cohen d=0.18), productivity loss reduced by 9.48% (P<.001, 95% CI 8.30% - 10.66%, Cohen d=0.27), and activity impairment reduced by 8.34% (P<.001, 95% CI 7.07% - 9.61%, Cohen d=0.30). Depression symptom scores reduced by 2.70 points (P<.001, 95% CI 2.50 - 2.90, Cohen d=0.51) and anxiety symptom scores reduced by 2.71 points (P<.001, 95% CI 2.51 - 2.91, Cohen d=0.52) scores posttreatment. Larger reductions in depression and anxiety symptoms were linked to greater improvements in workplace functioning outcomes (r=0.06-0.34, all P<.001). Patients with higher baseline clinical severity experienced approximately 6% greater improvements in workplace productivity than subclinical patients (P<.001). These improvements in work productivity from baseline to follow-up corresponded to more than €4000 annual savings per patient treated, which equated to an estimated annual savings of €29 million in the sample (in 2022 € values). The average currency exchange rate in 2022 would be €1 to approximately US $1.05.

Conclusions: Building on work that primarily focused on work productivity improvements from iCBT in occupational contexts, this study provides evidence that iCBT for the general adult population in a national routine health care setting is associated with improvements in both Common mental disorders (CMD) and workplace productivity losses associated with CMD. These productivity improvements were further associated with sizable cost savings. These findings suggest that integrating iCBT into

XSL•FO
RenderX

national mental health infrastructure can be a feasible, scalable solution, which could generate clinical and economic benefits at a population level.

(J Med Internet Res 2026;28:e80689) doi:10.2196/80689

KEYWORDS

digital mental health; internet-based cognitive behavioral therapy; iCBT; workplace productivity; depression; anxiety; cost effectiveness; health services

Introduction

Common mental disorders (CMDs) like depression and anxiety impact millions of people worldwide [1], resulting in negative social, functional, and health outcomes for individuals [2], as well as significant economic costs. Indeed, the global economy suffers an annual loss of US $1.15 trillion due to CMD, with one-third of this loss attributed to lost employment and productivity [3]. Compared to healthy individuals, 4.7 billion additional workdays are lost, amounting to a total cost of US $592 billion per year [3]. These significant implications warrant cost-effective interventions for CMD, with cognitive behavioral therapy (CBT) showing promise as an affordable, accessible solution when delivered online. Internet-based cognitive behavioral therapy (iCBT) has consistently demonstrated effectiveness in treating CMD across various settings [4-7], with cost-effectiveness [8] superior to other treatments or no treatment [9,10]. One key area to determine the cost-effectiveness of iCBT is through the indirect costs of CMD, such as the costs of decreased workplace productivity due to CMD [11]. Interestingly, recent systematic reviews on economic evaluations of iCBT for CMD suggest iCBT economic evaluations typically focus on direct costs (eg, costs of treatment, prescriptions, medical care), and indirect costs are frequently lacking or understudied [8,12].

CMD can accrue indirect costs through the impairment of workplace productivity, such as by increasing absenteeism (increased work absences), presenteeism (reduced effectiveness at work), higher rates of disability leaves, and early retirement [13-15]. These effects are observed at both clinical and subclinical levels of CMD [16,17]. A common method of measuring workplace productivity loss due to health conditions is through the Work Productivity and Activity Impairment (WPAI) questionnaire [18-20]. This questionnaire allows for the assessment of absenteeism, presenteeism, and overall productivity loss (the combination of absenteeism and presenteeism), as well as general activity impairment (the degree to which a health problem impairs daily activities outside of work), though the activity impairment index is less relevant for examining economic consequences. Focusing on the WPAI, a recent systematic review found that depression was associated with approximately 26 work absences per year and a 37% to 52% reduction in effectiveness at work [21].

Research conducted in occupational contexts suggests that iCBT can also improve absenteeism, presenteeism, and overall productivity [22-28], which, in turn, generates substantial financial returns for employers [22-24,29]. However, access to effective interventions through an employer is not always guaranteed. Despite growing awareness and actions addressing employee well-being, adoption and utilization rates of workplace mental health programs are low and dependent on company structure and culture [30,31]. Alternatively, iCBT administered to the general population through a public health service could provide a different approach to improving work productivity losses due to CMD. However, whether the positive impact of iCBT on work-related outcomes can generalize beyond workplace settings remains unclear.

The integration of iCBT into public health services can greatly increase the reach and cost-effectiveness of mental health care nationwide [25]. This expanded reach should potentially amplify any associated benefits from improved workplace productivity. In Ireland, the Health Service Executive (HSE) provides comprehensive, publicly funded mental health services for treating CMD. The HSE incorporated iCBT from *SilverCloud Health* by Amwell as part of their offering in 2021, making it one of the first nationally implemented iCBT services. Previous work on SilverCloud has shown that it is effective in treating CMD within the HSE and has yielded acceptable service-level metrics and high levels of user satisfaction [32]. Outside of the HSE, previous work has supported SilverCloud as being cost-effective within the United Kingdom's Improving Access to Talking Therapies, though this prior study was not able to incorporate or specifically examine costs associated with workplace productivity loss [33].

Building on this, we investigated the impact of providing iCBT in a national health service on workplace functioning, including absenteeism, presenteeism, productivity loss, and activity impairment. To do so, we analyzed mental health and work productivity routine care data collected from patients enrolled in SilverCloud at the HSE. We hypothesized that both mental health symptoms and workplace functioning would improve posttreatment. We further estimated the cost savings incurred by observed changes in work-related outcomes.

Methods

Study Setting and Participants

This naturalistic study analyzed repeated measures retrospective data from n=11,131 patients enrolled in SilverCloud within Ireland's HSE during the time period in which the WPAI questionnaire was administered (from March 8, 2023 to May 27, 2024). Note that this study adhered to CHEERS (Consolidated Health Economic Evaluation Reporting Standards) 2022 guidelines for economic evaluations (see Checklist 1) [34]. Within the HSE, patients are referred to SilverCloud through health professionals across 5 referral groups (ie, General Practitioners, Primary Care Psychology, Counseling in Primary Care, Community Mental Health, and an HSE-funded

XSL•FO
RenderX

mental health charity called Jigsaw). The HSE and SilverCloud staff both provided communications to these referral groups detailing the availability of this iCBT service, who would be suitable for it, and how to have a referral conversation with their patients.

When considering a patient for SilverCloud, health professionals perform an initial screening to assess whether they may be suitable for the intervention. Eligible patients for the treatment (and the study) were at least 18 years old and presented with common mental health issues (eg, low mood, anxiety, feelings of isolation, stress, poor sleep). Patients were ineligible if they had severe mental health symptoms such as manic or psychotic episodes and/or suicidal ideation. If a patient is suitable, the health professional submits a referral to SilverCloud, which triggers an email invite sent to the patient through the SilverCloud platform. Within the email, an invitation link redirects the client to the sign-up page on the SilverCloud website. Patients who do not activate their account receive weekly reminder emails for the first 30 days following the initial invitation. During the sign-up process, patients are asked to provide demographic information, complete baseline clinical questionnaires, and can choose to consent for their pseudoanonymized data to be used for clinical research and service evaluation purposes. Once sign-up is complete, patients are directed to the SilverCloud homepage where they can begin the program.

Intervention

Within the HSE, SilverCloud's *Space from Depression*, *Space from Anxiety*, *Space from Depression and Anxiety*, and *Space from Generalized Anxiety Disorder* programs were offered during the study time period. These are general purpose iCBT programs that consist of a series of modules aligned with evidence-based CBT principles for the treatment of depression and anxiety [33,35]. Specifically, each program aims to help the patient develop and increase an understanding of the relationship between one's thoughts, feelings, and behaviors, alongside information about coping strategies and therapeutic techniques such as graded exposure, behavioral activation, cognitive restructuring, problem solving, and activity scheduling. This content is delivered through various interactive media and tools, including quizzes, videos, activities, and personal stories. During program use, patients also complete routine outcome measures (ie, Patient Health Questionnaire-9 [PHQ-9], Generalized Anxiety Disorder-7 [GAD-7], WPAI, see below) at intake, week 2, week 4, and week 8 of treatment. Risk is monitored continuously through item 9 on the PHQ-9.

Patients are also assigned a supporter during their use of SilverCloud. All supporters are graduate psychologists with at least a master's degree or doctorate in progress who have been trained in the delivery of online support. Supporters also receive training specific to providing support on SilverCloud and receive supervision from clinical supervisors. Supporters provide an asynchronous written "review" in the platform to the patient. The primary aim of the review is to provide the patient with personalized guidance and feedback based on their progress and clinical symptomology. Supporter reviews are provided weekly for the first 6 weeks, though additional reviews may be provided at the discretion of the supporter and their clinical supervisor. Once all the reviews are provided, or if the patient does not engage with the program for 3 consecutive review periods, the supporter discharges the patient from the supported service and no longer provides reviews. After patients are discharged, they still have full access to the platform in an unsupported capacity.

Assessments

Demographic Questionnaire

During intake, clients completed demographic questions about gender, education, race, ethnicity, age, work status, and marital status. They are also asked whether they are in concurrent mental health treatment and to rate how likely they think SilverCloud will work for them.

Routine Outcome Measures

The following questionnaires are part of the HSE's routine outcome measures that were administered at intake, week 2, week 4, and week 8 of treatment.

WPAI: Specific Health Problem

The WPAI: Specific Health Problem is a 6-item questionnaire that probes the impact of a specified health problem on workplace functioning over the past 7 days [19]. Outcomes include absenteeism (work time missed), presenteeism (impairment while working), overall productivity loss (absenteeism and presenteeism combined), and nonwork activity impairment (see Multimedia Appendix 1). Each outcome is expressed as a percentage, with higher numbers indicating greater impairment. The measure has shown good construct validity and test-retest reliability [19].

Patient Health Questionnaire-9

The PHQ-9 is a 9-item self-report questionnaire for depressive symptoms, with good reliability, sensitivity, and specificity [36]. The total score ranges from 0 to 27, with each item rated on a 4-point Likert scale (0=*not at all* to 3=*nearly every day*). The scale has 4 cut-off points of 5, 10, 15, and 20, which represent mild, moderate, moderately severe, and severe levels of depression, respectively.

Generalized Anxiety Disorder-7

The GAD-7 is a 7-item self-report questionnaire that identifies individuals with generalized anxiety disorder [37]. Items are scored on a 4-point Likert scale (0=*not at all* to 3=*nearly every day*). The total score can range from 0 to 21, with three cut-off points (5, 10, and 15) representing mild, moderate, and severe anxiety levels, respectively. The scale has also demonstrated good reliability, sensitivity, and specificity [37].

Ethical Considerations

All procedures of this study adhered to the ethical standards of the relevant national and institutional committees on human experimentation and the Declaration of Helsinki (1975, revised in 2008). The protocol-related documents and informed consent were reviewed and approved by the Trinity College Dublin School of Psychology Research Ethics Committee (Approval ID: SPREC112021-04). Approval for users to not provide informed consent was granted, as consent had already been

XSL•FO
RenderX

given for treatment and for therapists to gather and examine clinical outcome measures as part of normal service evaluation procedures. Service users who refused consent to having their data passed on for these purposes were not included in this study. Study data were collected and stored in a deidentified manner and complied with General Data Protection Regulations. Patients were not compensated for this study, as the data were derived from routine service assessments administered during routine treatment. No images of patients were used in this study or are presented in this manuscript or supplemental materials.

Data Analysis

The study sample size was based on all available data from all patients who reported being employed via the WPAI questionnaire. Each patient's last assessment prior to treatment discharge was used as their follow-up measure. We included all patients who completed the baseline assessment and used multiple imputation to estimate follow-up CMD and workplace functioning scores for those missing follow-up data. Performing the Little Missing Completely At Random test suggested WPAI data were not missing at random (χ^2_{27}=3921.24, P<.001). Multiple imputation is a conservative method for handling missing data as it incorporates natural variability into the missing values [38]. Pooled estimates from the imputation are reported. For sensitivity purposes, core analyses were repeated using complete-case data (Multimedia Appendix 2). Descriptive analyses were conducted on patients' demographics, baseline clinical and work-related outcomes, as well as treatment characteristics.

First, univariate regressions were used to test whether CMD (depression and anxiety) (independent variable) and workplace functioning (dependent variable) were associated at baseline. To identify potential confounds, we tested the association between changes in WPAI outcomes and covariates, including demographics (age, gender, education, ethnicity, marital status, long-term conditions) and treatment-related characteristics (concurrent treatment, expectation), which were coded according to Table 1. Due to our naturalistic study design, the time from baseline to follow-up varied considerably (mean 34.89, SD 14.68 d). We controlled for this by evaluating the assessment interval as a potential covariate, which captured the treatment duration relevant to the measured effects. Next, to assess whether workplace functioning and CMD improved posttreatment, we employed mixed-effects models, regressing workplace functioning and CMD outcomes (dependent variable) on time as a fixed effect (independent variable), including significant covariates previously identified. To address repeated observations nested within patients across time, we incorporated random intercepts for patients in our models. To quantify the pre-post change in WPAI outcomes, we calculated Cohen *d* effect sizes [39]. We explored baseline CMD severity and comorbidity as moderators to examine their effects on changes in workplace functioning.

To calculate productivity-related cost savings from the intervention, we adopted the human capital approach, a widely established method for estimating the monetary value associated with productivity losses. The method values 1 hour of productivity loss based on the individual's hourly wages [40]. We obtained median hourly salary estimates (stratified by gender and education) from available Irish census data in the 2022 Structure of Earnings Survey [41]. Thus, estimates are in 2022 € values (€1=approximately US $1.05) [42]. We then multiplied these estimates by the percentages of absenteeism, presenteeism, and overall productivity loss to determine the cost, scaling it accordingly to weekly and yearly estimates, assuming a full-time 39-hour work week. We did not apply a discount rate given the short timeframe of the study. We calculated this per patient and used the mean for analysis.

All analyses were conducted in R (version 4.4.0), and all tests were 2-sided, with statistical significance set at α=.05.

XSL•FO
RenderX

Table . Baseline characteristics of the study sample[a].

Characteristics	Participants
Age (y), n (%)	
18 - 24	1461 (20.64)
25 - 64	5578 (78.82)
65+	38 (0.54)
Gender, n (%)	
Female	4956 (70.23)
Male	2067 (29.29)
Other/Prefer not to say	34 (0.48)
Race and ethnicity, n (%)	
White Irish	6023 (85.32)
Other White European	652 (9.24)
Asian	99 (1.40)
Mixed	74 (1.05)
Black	71 (1.01)
Latino	62 (0.88)
Other	78 (1.10)
Education, n (%)	
Primary to secondary	2411 (34.41)
College/university	3614 (51.58)
Postgraduate	982 (14.01)
Marital status, n (%)	
Single	2134 (30.30)
In a relationship	2332 (33.12)
Married	2155 (30.60)
Separated/divorced	376 (5.34)
Widowed	45 (0.64)
Long-term condition, n (%)	
Yes	1605 (22.87)
No	5413 (77.13)
Concurrent treatment, n (%)	
Yes	3766 (53.34)
No	3295 (46.66)
Depression severity (PHQ-9)[b], mean (SD)	12.82 (6)
Minimal, n (%)	546 (7.66)
Mild, n (%)	1832 (25.71)
Moderate, n (%)	1982 (27.82)
Moderately severe, n (%)	1621 (22.75)
Severe, n (%)	1144 (16.06)
Anxiety severity (GAD-7)[c], mean (SD)	12.62 (5.38)
Minimal, n (%)	457 (6.41)
Mild, n (%)	1830 (25.68)

XSL•FO
RenderX

Characteristics	Participants
Moderate, n (%)	1903 (26.71)
Severe, n (%)	2935 (41.19)
WPAI[d] absenteeism (%), mean (SD)	19.45 (33.88)
WPAI presenteeism (%), mean (SD)	34.17 (30.04)
WPAI productivity loss (%), mean (SD)	49.19 (34.44)
WPAI activity impairment (%), mean (SD)	49.74 (27.87)
Treatment duration (d), mean (SD)	34.89 (14.68)
Treatment expectation (0 - 4), mean (SD)	2.17 (0.65)

[a]Summary descriptives were calculated based on the total responses per variable, which may differ due to missing data. The proportion of missingness for all variables was less than 2.5%, except for baseline presenteeism (6%), productivity loss (6%), and activity impairment (8%).

[b]PHQ-9: Patient Health Questionnaire-9.

[c]GAD-7: Generalized Anxiety Disorder-7.

[d]WPAI: Workplace Productivity and Activity Impairment.

Results

Patient Characteristics

The participant flow chart is shown in Figure 1. Of the 11,810 patients enrolled in the national iCBT service, 11,131 completed the baseline assessment fully or partially, of whom 7125 (64%) reported being employed. These 7125 participants formed the final sample for analyses.

XSL•FO
RenderX

Figure 1. Participant flow chart. Participants may have completed their baseline and follow-up assessments fully or partially, as completion of the assessment was not compulsory for them to progress in the treatment. WPAI: Work Productivity and Activity Impairment.

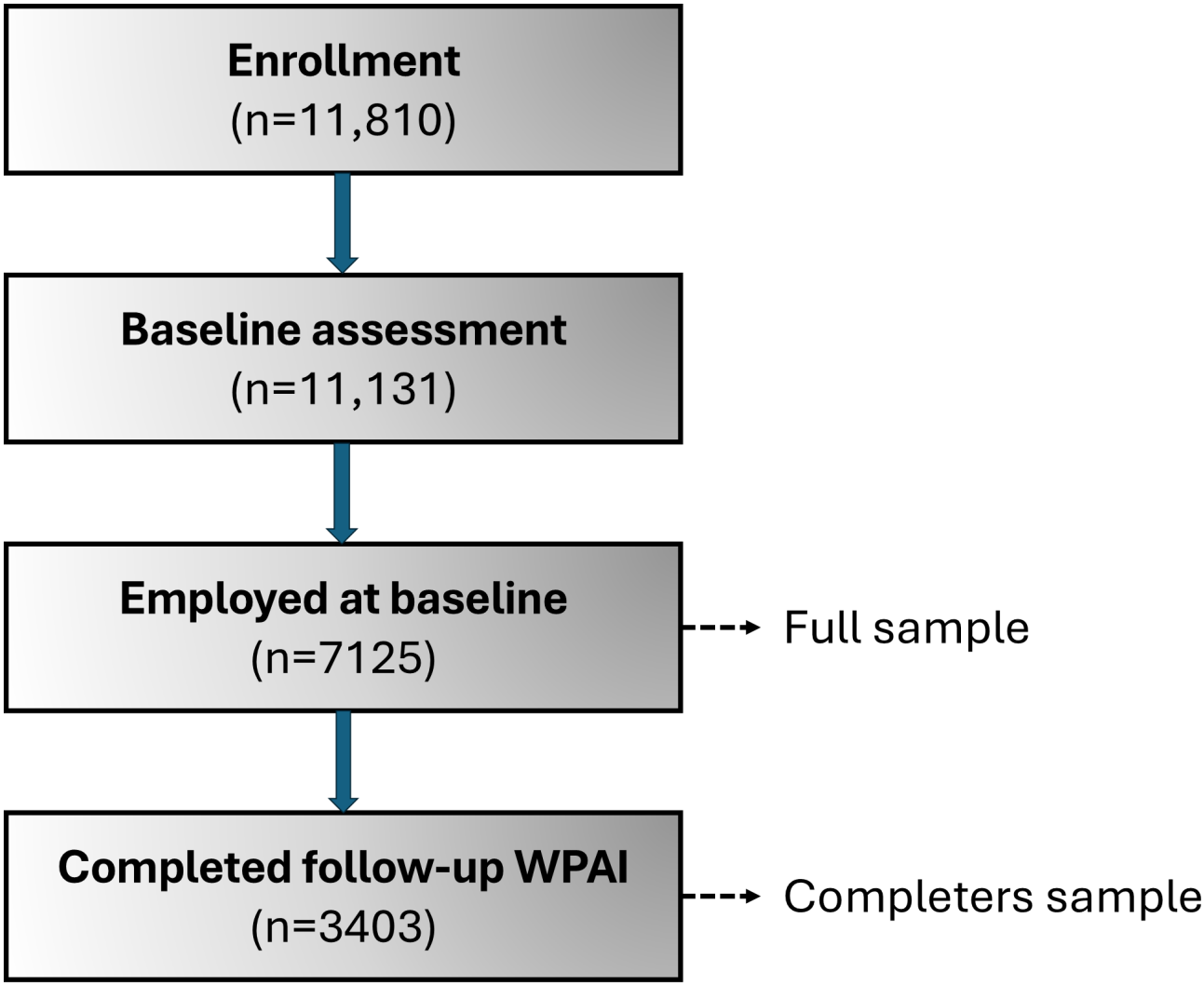

Baseline Analyses

Table 1 outlines the characteristics of our sample. Patients were primarily 25 to 64 years old, female, White Irish, had college or university education, in romantic relationships, and not presenting with a physical long-term condition. Their baseline severity on average fell within the moderate classification of depression (PHQ-9) and anxiety (GAD-7), respectively. Over half were receiving concurrent treatment in addition to iCBT, which most were undertaking for approximately 35 days. Overall, there was a positive expectation regarding the benefit that users believed they would gain from the treatment, with approximately 90% indicating that they believed that it was at least 'somewhat likely' to work for them. Regarding workplace functioning, patients were on average absent from work 20% of the week and had 34% reduced performance while working during the week. Both productivity loss and activity impairment averaged near 50%.

A series of univariate regressions revealed positive associations between both PHQ-9 and GAD-7 and all four WPAI outcomes at baseline (Table 2), where greater levels of depression and anxiety predicted increased absenteeism, presenteeism, overall productivity loss, and activity impairment (all $P<.001$). More specifically, every 1-point increase on the PHQ-9 and GAD-7 predicted approximately a 1% to 2.6% decrease in workplace functioning.

XSL•FO
RenderX

Table . Baseline associations between clinical and work-related outcomes at baseline, as assessed using univariate linear regressions.

Work-related outcomes and CMD[a]	Univariate analyses			
	Unstandardized β (95% CI)	SE	*t* test (*df*)	*P* value
Absenteeism				
PHQ-9[b]	1.36 (1.24 - 1.48)	0.06	20.89 (7119.64)	<.001
GAD-7[c]	1.01 (0.87 - 1.15)	0.07	13.76 (7120.21)	<.001
Presenteeism				
PHQ-9	1.43 (1.31 - 1.54)	0.06	24.28 (2451.73)	<.001
GAD-7	1.56 (1.42 - 1.70)	0.07	23.34 (1358.33)	<.001
Productivity loss				
PHQ-9	2.30 (2.18 - 2.42)	0.06	36.38 (3893.65)	<.001
GAD-7	2.13 (1.99 - 2.27)	0.07	28.84 (1972.90)	<.001
Activity impairment				
PHQ-9	2.55 (2.45 - 2.65)	0.05	52.26 (1090.67)	<.001
GAD-7	2.26 (2.14 - 2.38)	0.06	39.88 (3198.07)	<.001

[a]CMD: common mental disorder.

[b]PHQ-9: Patient Health Questionnaire-9.

[c]GAD-7: Generalized Anxiety Disorder-7.

Pre-Post Changes in Workplace Functioning

Over an average treatment period of 35 days, we observed significant improvements across all WPAI outcomes. Specifically, there was an average reduction of 6.85% in absenteeism (b=−6.85, $t_{71.47}$=−12.75, P<.001, 95% CI −5.79 to −7.91, Cohen d=0.21), 5.84% in presenteeism (b=−5.84, $t_{38.51}$=−9.15, P<.001, 95% CI −4.59 to −7.09, Cohen d=0.18), 9.48% in productivity loss (b=−9.48, $t_{61.18}$=−15.70, P<.001, 95% CI −8.30 to −10.66, Cohen d=0.27), and 8.34% in activity impairment (b=−8.34, $t_{29.13}$=−12.88, P<.001, 95% CI −7.07 to −9.61, Cohen d=0.30) (Figure 2). Covariates such as treatment duration, treatment expectation, gender, concurrent treatment, and marital status were significant predictors of change in one or more WPAI outcomes. Importantly, all observed pre-post changes in WPAI outcomes remained statistically significant after adjusting for these confounders (all P<.001) (Multimedia Appendix 3).

XSL•FO
RenderX

Figure 2. Pooled mean impairment at baseline and follow-up for absenteeism, presenteeism, productivity loss, and activity impairment. Error bars denote the 95% CI.

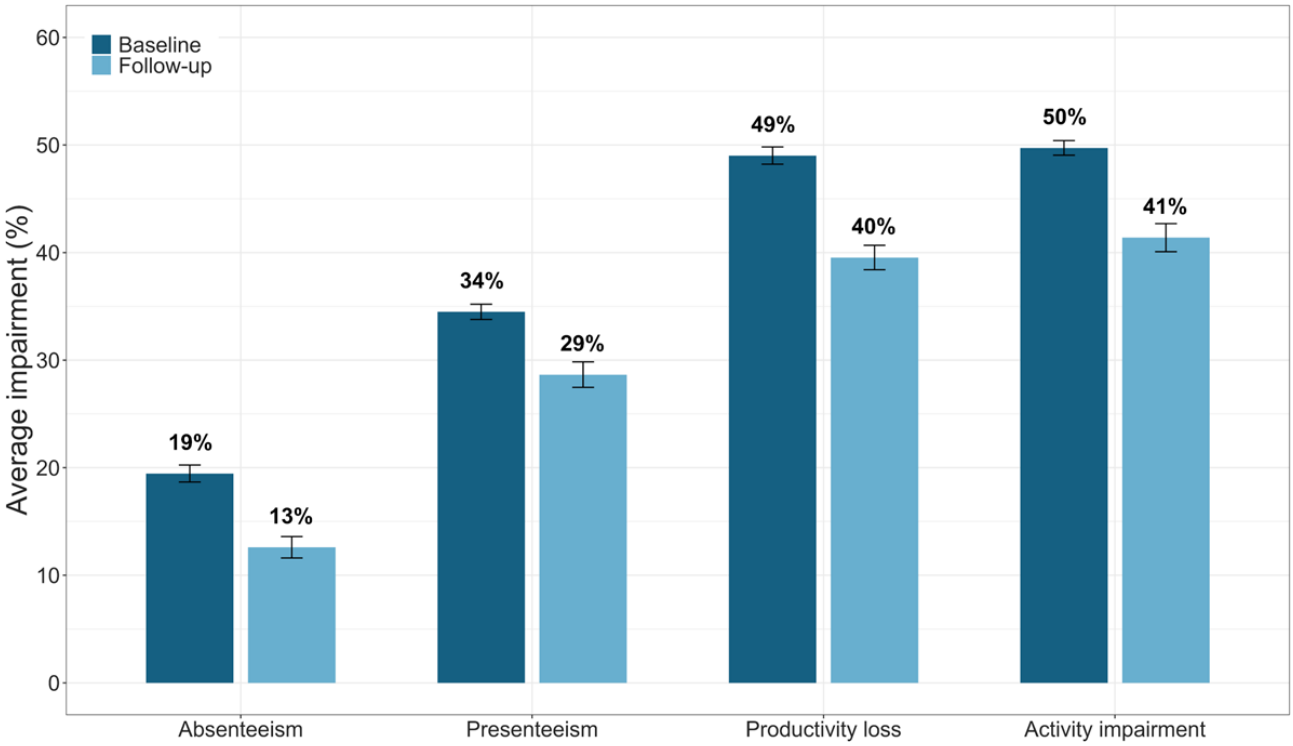

For depression and anxiety, we further observed a significant reduction in both PHQ-9 (b=−2.70, $t_{42.71}$=-27.11, P<.001, 95% CI −2.50 to −2.90, Cohen d=0.51) and GAD-7 (b=−2.71, $t_{39.36}$=−26.52, P<.001, 95% CI −2.51 to −2.91, Cohen d=0.52) scores posttreatment. Partial correlations accounting for significant covariates previously identified revealed that these clinical symptom reductions were also positively associated with changes in absenteeism (PHQ-9: r=0.10, 95% CI 0.07 to 0.13; GAD-7: r=0.06, 95% CI 0.03 to 0.09), presenteeism (PHQ-9: r=0.18, 95% CI 0.14 to 0.21; GAD-7: r=0.21, 95% CI 0.18 to 0.24), productivity loss (PHQ-9: r=0.21, 95% CI 0.18 to 0.24; GAD-7: r=0.21, 95% CI 0.18 to 0.24), and activity impairment (PHQ-9: r=0.34, 95% CI 0.31 to 0.37; GAD-7: r=0.33, 95% CI 0.30 to 0.36; all P<.001). This pattern indicates that greater improvements in depression and anxiety symptoms were associated with greater improvements in each WPAI outcome.

To examine whether baseline CMD severity, that is, clinical (moderate to severe symptoms) vs subclinical (minimal to mild symptoms), influences treatment effects on workplace functioning, we included baseline PHQ-9 and GAD-7 scores as moderators, along with covariates previously identified for each outcome (see Multimedia Appendix 4 for demographic characteristics by CMD severity). Significant time × baseline PHQ-9 interaction effects suggested that patients with higher baseline depression experienced larger reductions in absenteeism (b=−5.68, $t_{130.37}$=-5.55, 95% CI −7.68 to −3.68; P<.001), presenteeism (b=−2.48, $t_{50.02}$=−2.03, 95% CI −4.87 to −0.09; P=.048), productivity loss (b=−5.32, $t_{80.23}$=−4.43, 95% CI −6.35 to −2.87; P<.001), and activity impairment (b=−4.61, $t_{107.97}$=−5.17, 95% CI −9.61 to −7.07; P<.001). More specifically, those who were clinical at baseline experienced larger reductions than their subclinical counterparts in absenteeism (subclinical: b=−3.06, 95% CI −1.59 to −4.53, P<.001; clinical: b=−8.74, 95% CI −7.35 to −10.13, P<.001), presenteeism (subclinical: b=−4.18, 95% CI −2.55 to −5.81, P<.001; clinical: b=−6.67, 95% CI −4.95 to −8.39, P<.001), productivity (subclinical: b=−5.93, 95% CI −4.13 to −7.73, P<.001; clinical: b=−11.25, 95% CI −9.72 to −12.78, P<.001), and activity impairment (subclinical: b=−5.27, 95% CI −3.80 to −6.74, P<.001; clinical: b=−9.88, 95% CI −8.35 to −11.41, P<.001; Figure 3 and Multimedia Appendix 4).

A similar trend was also observed for anxiety, where those entering treatment with higher GAD-7 scores experienced larger reductions in workplace functioning (absenteeism: b=−4.87, $t_{380.35}$=−5.21, 95% CI −6.69 to −3.05, P<.001; presenteeism: b=−4.62, $t_{76.03}$=−4.17, 95% CI −6.80 to −2.44, P<.001; productivity loss: b=−6.61, $t_{102.34}$=−5.70, 95% CI −8.88 to −4.34, P<.001; activity impairment: b=−5.67, $t_{75.85}$=−5.90, 95% CI −7.55 to −3.79, P<.001). Patients with clinical score levels on the GAD-7 had larger decreases than subclinical patients in absenteeism (subclinical: b=−3.54, 95% CI −1.99 to −5.09, P<.001; clinical: b=−8.41, 95% CI −7.16 to −9.66, P<.001), presenteeism (subclinical: b=−2.70, 95% CI −0.96 to −4.44, P=.02; clinical: b=−6.19, 95% CI −4.64 to −7.74, P<.001), productivity loss (subclinical: b=−4.99, 95% CI −3.09 to −6.89, P<.001; clinical: b=−11.60, 95% CI −10.19 to −13.01, P<.001), as well as activity impairment (subclinical: b=−4.50, 95% CI −2.79 to −6.21, P<.001; clinical: b=−10.20, 95% CI −8.73 to −11.67, P<.001) (Figure 3 and Multimedia Appendix 4).

Importantly, we found that patients experiencing comorbid clinical levels of depression and anxiety at baseline also improved more in workplace functioning than those who were noncomorbid. This was evidenced by significant time ×

XSL•FO
RenderX

comorbidity interaction effects observed for all WPAI outcomes (all $P<.001$, Multimedia Appendix 5).

Figure 3. Changes in work-related outcomes from baseline to follow-up, moderated by baseline depression severity (top row) and baseline anxiety severity (bottom row). The points represent unadjusted means, and the error bars represent SE. GAD7: Generalized Anxiety Disorder-7; PHQ9: Patient Health Questionnaire-9.

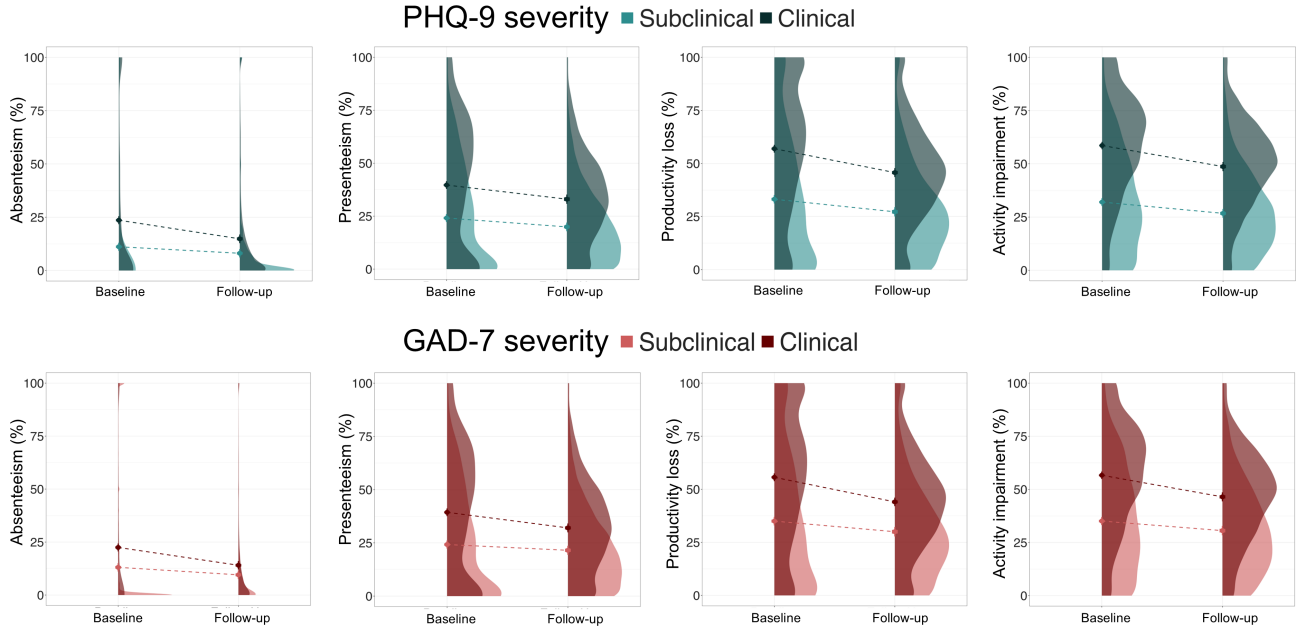

Cost Analysis

Based on 2022 Irish census median salary estimates by gender and educational attainment, the average median salary in the sample was estimated to be €836.06 (95% CI €830.29-€841.83, SD €248.62) per week, which is, €43,475.34 (95% CI €43,175.15-€43,775.43, SD €12,928.16) per year. At baseline, we calculated the cost induced by deficits in workplace functioning for each patient (WPAI % multiplied by salary), before totaling these values to produce an average. The average costs of absenteeism and presenteeism per patient per week were estimated to be €159.82 (95% CI €153.00-€166.64, SD €293.59) and €290.80 (95% CI €284.30-€297.30, SD €280.04), totaling to an annual monetary loss of €8310.64 (95% CI €7955.52-€8664.48, SD €15,266.39) and €15,121.60 (95% CI €14,782.87-€15,459.13, SD €14,562.22). For overall productivity loss, the total cost induced is nearly half of each patient's earnings, estimated at €409.87 (95% CI €402.41-€417.33, SD €321.29) per week and €21,313.26 (95% CI €20,925.32-€21,701.20, SD €16,707.25) per year.

To estimate the annual cost savings due to improvements in workplace functioning, we multiplied the adjusted mean reduction in absenteeism (6.85%), presenteeism (5.84%), and productivity loss (9.48%) by the average salary of our sample. With each percent change in WPAI outcomes, annual costs were estimated to decline by €434.75 per patient. This results in annual cost savings of €2978.06 for absenteeism (95% CI €2546.94-€3409.18), €2538.96 for presenteeism (95% CI €1989.67-€3088.25), and €4121.11 for overall productivity loss (95% CI €3589.11-€4645.11) for each patient treated (Figure 4A). In the context of our study sample (n=7125), this improvement equates to an annual cost saving of €29.33 million.

Importantly, when we consider CMD severity and comorbidity of patients at baseline, cost savings were more substantial for those presenting with more severe symptomatology (Figure 4). Considering overall productivity loss, annual cost savings for patients with clinical depression (11.25% improvement) were 1.90 times higher compared to subclinical patients (5.93% improvement) (average cost savings for clinical depression: €4892.98, 95% CI €4233.28-€5550.72 vs subclinical depression: €2579.04, 95% CI €1762.23-€3395.77; Figure 4B). For anxiety, annual cost savings for clinical patients (11.60% improvement) were 2.33 times higher than subclinical patients (4.99% improvement) for overall productivity loss (average cost savings for clinical anxiety: €5042.36, 95% CI €4402.67-€5682.71 vs subclinical anxiety: €2169.02, 95% CI €1317.28-€3020.76; Figure 4C). Similarly, cost savings are 2.05 times higher for comorbid patients (12.23% improvement) compared to their counterparts (5.98% improvement) (average cost savings for non-comorbid: €2601.92, 95% CI €1890.44-€3313.40 vs comorbid: €5318.41, 95% CI €4623.67-€6013.15; Figure 4D).

XSL•FO
RenderX

Figure 4. (A) Estimated average annual cost savings (in 2022 € values; €1=approximately US $1.05) postintervention induced by improvements in absenteeism, presenteeism, and overall productivity loss for patients with clinical levels versus subclinical levels of (B) depression, (C) anxiety, and for (D) patients with comorbid clinical levels of depression and anxiety vs those without. Monetary figures were rounded to the nearest integer. Error bars denote the 95% CI.

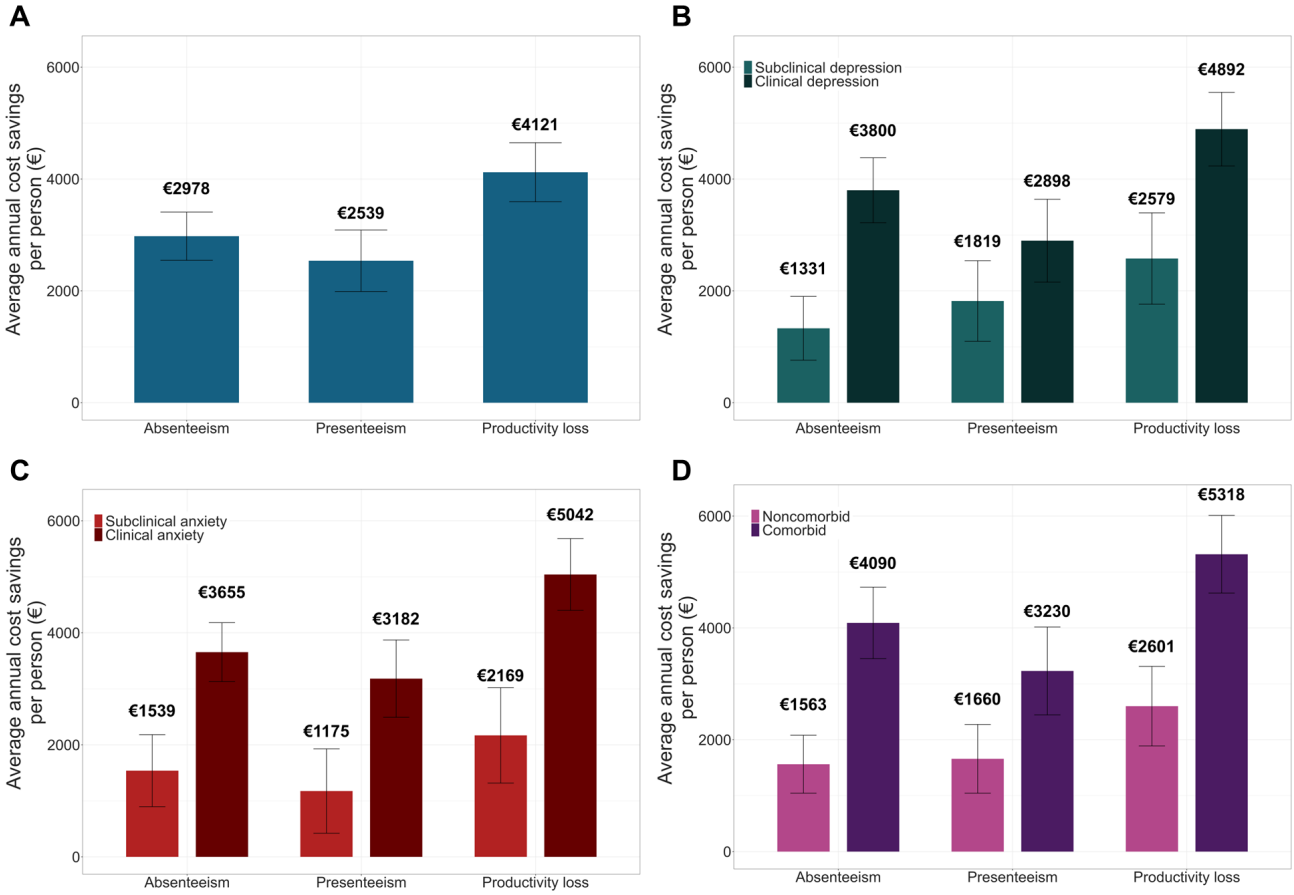

Sensitivity Analysis

To determine whether multiple imputation introduced bias into our findings, we repeated the core analysis, including only patients who completed both baseline and follow-up assessments (n=3402) (Multimedia Appendix 2). The results remained consistent, showing significant improvements in all CMD and WPAI outcomes, with greater improvements among patients with clinical levels of CMD and comorbid CMD compared to their counterparts.

Discussion

This study is the first to use nationwide routine care data to investigate the effectiveness of iCBT, delivered as a public health initiative for treating CMD, in improving workplace functioning in the general population. These improvements in work productivity were estimated to lead to substantial cost savings. As hypothesized, patients reported fewer missed workdays, less impairment during work and activities, and higher overall productivity after 5 weeks of treatment. Although the effects observed here (6% - 10% reductions) were smaller than those in prior studies that included small, completer-only samples with limited generalizability [22,29], they align more closely with those reported in meta-analyses of work-focused CBT or iCBT interventions on work absences and effectiveness (Hedges g=0.22 - 0.25) [25,43]. Work-related improvements were observed alongside reductions in depression and anxiety symptoms, which were consistent with our previous study [32].

Notably, iCBT improved workplace functioning for both clinical and subclinical patients, with greater benefits for those with more severe symptomatology at baseline. This finding not only highlights iCBT as an effective intervention for improving workplace functioning in individuals with clinically apparent symptoms but also for subclinical individuals, helping to prevent their symptoms from reaching clinical levels. Note that these subclinical individuals likely represent the typical working population [22]. Future research should explore whether engagement with standard vs work-focused content differentially impacts improvements in workplace functioning, again using more representative data from the general working population. This is particularly pertinent given the limited evidence supporting the superiority of tailored over nontailored digital interventions [26,27,44].

The consequences and scale of impaired workplace functioning due to mental illness present major public health challenges, necessitating the implementation of cost-effective and accessible interventions [3,44]. Our study builds on the promising evidence of iCBT interventions in improving clinical, work-related, and financial outcomes in occupational contexts [22-24] and demonstrates the scalability of these benefits when iCBT is provided in a national health service. By expanding the intervention's reach beyond industry-specific settings and

XSL•FO
RenderX

barriers, these outcomes have significant implications for the economic health of the nation. This is evident in the measurable cost savings demonstrated in our study, estimated at more than €4000 over the course of a year per patient treated, totaling €29 million annually within our sample. These findings are aligned with the conclusions of the few economic evaluations on workplace iCBT interventions [22-24,29]. The findings also dovetail with other mental health and economic evaluations focusing on iCBT implemented within national health care systems, such as MindSpot in Australia and SilverCloud within Talking Therapies in the United Kingdom, though these prior studies were not focused on cost-effectiveness through workplace productivity [33,45]

These cost-savings estimates should also be seen as conservative, given that they focus on costs due to workplace impairment and do not capture medical costs, costs of providing the treatment, or broader societal benefits (eg, reduced welfare and health care expenditure, better health care utilization, and positive spillover effects on patients' support networks) [3,46]. While it is possible that expanding treatment coverage in the population may increase the average cost per person initially, particularly after reaching previously inaccessible patients in remote areas, the expected financial returns from improved health and work-related outcomes have been shown to far outweigh the treatment costs for depression and anxiety (return ratio of 2.3-3:1 considering only economic benefits, 3.3 - 5.7:1 when including health returns) [3]. Furthermore, our findings highlight that iCBT is uniquely positioned to bridge the substantial treatment gap between demand and availability on a national level. As a feasible, scalable solution, iCBT can alleviate the strain of CMDs on overburdened health care systems and improve workforce productivity, all the while allowing greater access to care at lower cost compared to in-person approaches [47,48].

This study has several limitations. The naturalistic design of our study, without comparison groups, prevents us from determining the causality of the observed treatment effects. Estimates from this study could serve as benchmarks for future studies focused on evaluating outcomes and the cost-effectiveness of mental health treatments within the HSE. Outcomes were also collected at fixed intervals set by the service, with the latest assessment administered at week 8 of treatment. Thus, the extent to which treatment effects on symptoms persist over longer time frames (eg, 6 mo, 1 y) is unknown in this sample. This may particularly bias study estimates projecting annual cost savings. Though prior work investigating the treatment durability of the SilverCloud iCBT program suggests that approximately 70% of patients who recover from symptoms remain in recovery 9 months later [49]. The study also relies on self-report data, uncorroborated by objective and formal assessments, which may be more accurate. Another limitation is the precision of our estimates due to the nature of the available data. As WPAI outcomes can only be calculated for employed individuals [19], this fails to account for those unemployed due to their condition (ie, complete loss of productivity).

Additionally, while using national salary estimates is a common procedure for estimating cost savings [22-24,29], collecting and incorporating data on patient-specific salary, employment sector, and typical hours worked would produce financial return estimates produced directly from actual participant wage and job data. It is also important to note that we assumed a linear transformation of productivity loss to monetary loss, and it is possible that the association may be nonlinear. Data on intervention costs and other medical costs were also unavailable, which prevented a more comprehensive cost-effectiveness analysis. Future studies should incorporate income and medical cost assessments with comparator treatments using representative data to enable a national economic evaluation.

Depression and anxiety significantly impact labor participation and productivity, but this can be effectively managed with accessible treatments through public health services. Building on work that primarily focused on work productivity improvements from iCBT in occupational contexts, this study provides evidence that iCBT for the general adult population in a national routine health care setting is associated with improvements in both CMD and workplace productivity losses associated with CMD. These productivity improvements were further associated with sizable cost savings. These findings suggest that integrating iCBT into national mental health infrastructure can be a feasible, scalable solution, which could generate clinical and economic benefits at a population level.

Acknowledgments

We extend our gratitude to the Health Service Executive leadership, all relevant stakeholders, and clinicians for their roles in the success of this collaboration. We thank our patients for providing their data for this project. Additionally, we would like to thank all the people who played a role in the dissemination and promotion of this collaboration, specifically Elaine Martin, Hannah Farrell, Paulina Chmara, and Liam Ryan. Artificial Intelligence was not used in the writing of this manuscript.

Funding

This study was not supported by a grant from any funding agency, commercial or not-for-profit sectors; however, CTL, SH, AE, GH, DD, and DR are current employees of SilverCloud Health.

Data Availability

The dataset, along with the analysis scripts for the current study, is available upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

CTL, DR, SH, DD, and AE contributed to the conceptualization of the overall study. CTL, SH, GH, and AE analyzed and interpreted the data. CTL, AJ-D, SH, and AE produced the first draft of the manuscript. All authors edited, revised, and approved the final manuscript.

Conflicts of Interest

CTL, SH, AE, GH, DD, and DR are or were employees of SilverCloud Health during the production of this manuscript. SilverCloud Health is a subsidiary of Amwell. SH, GH, and DR hold shares in Amwell.

Multimedia Appendix 1
Calculation of work productivity activity and impairment outcomes.
[DOCX File, 17 KB - jmir_v28i1e80689_app1.docx]

Multimedia Appendix 2
Core data analyses for completers sample.
[DOCX File, 21 KB - jmir_v28i1e80689_app2.docx]

Multimedia Appendix 3
Covariate analyses examining pre-post changes in Work Productivity and Activity Impairment (WPAI).
[DOCX File, 17 KB - jmir_v28i1e80689_app3.docx]

Multimedia Appendix 4
Pre-post analyses of changes in Work Productivity and Activity Impairment (WPAI) for patients with clinical and subclinical common mental disorders.
[DOCX File, 20 KB - jmir_v28i1e80689_app4.docx]

Multimedia Appendix 5
Pre-post analyses of changes in Work Productivity and Activity Impairment (WPAI) for patients with comorbid and noncomorbid common mental disorders.
[DOCX File, 17 KB - jmir_v28i1e80689_app5.docx]

Checklist 1
CHEERS 2022 checklist.
[DOCX File, 19 KB - jmir_v28i1e80689_app6.docx]

References

1. World mental health report: transforming mental health for all. : World Health Organization; 2022 URL: https://www.who.int/publications/i/item/9789240049338 [accessed 2024-08-05]
2. National Collaborating Centre for Mental Health (UK). Common Mental Health Disorders: Identification and Pathways to Care: British Psychological Society; 2011. [Medline: 22536621]
3. Chisholm D, Sweeny K, Sheehan P, et al. Scaling-up treatment of depression and anxiety: a global return on investment analysis. Lancet Psychiatry 2016 May;3(5):415-424. [doi: 10.1016/S2215-0366(16)30024-4] [Medline: 27083119]
4. Andrews G, Basu A, Cuijpers P, et al. Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: an updated meta-analysis. J Anxiety Disord 2018 Apr;55:70-78. [doi: 10.1016/j.janxdis.2018.01.001] [Medline: 29422409]
5. Cuijpers P, Noma H, Karyotaki E, Cipriani A, Furukawa TA. Effectiveness and acceptability of cognitive behavior therapy delivery formats in adults with depression: a network meta-analysis. JAMA Psychiatry 2019 Jul 1;76(7):700-707. [doi: 10.1001/jamapsychiatry.2019.0268] [Medline: 30994877]
6. Hadjistavropoulos HD, Peynenburg V, Thiessen DL, et al. Utilization, patient characteristics, and longitudinal improvements among patients from a provincially funded transdiagnostic internet-delivered cognitive behavioural therapy program: observational study of trends over 6 years. Can J Psychiatry 2022 Mar;67(3):192-206. [doi: 10.1177/07067437211006873] [Medline: 33840264]
7. Titov N, Hadjistavropoulos HD, Nielssen O, Mohr DC, Andersson G, Dear BF. From research to practice: ten lessons in delivering digital mental health services. J Clin Med 2019 Aug 17;8(8):1239. [doi: 10.3390/jcm8081239] [Medline: 31426460]

XSL•FO
RenderX

8. Mitchell LM, Joshi U, Patel V, Lu C, Naslund JA. Economic evaluations of internet-based psychological interventions for anxiety disorders and depression: a systematic review. J Affect Disord 2021 Apr 1;284:157-182. [doi: 10.1016/j.jad.2021.01.092] [Medline: 33601245]
9. Ahern E, Kinsella S, Semkovska M. Clinical efficacy and economic evaluation of online cognitive behavioral therapy for major depressive disorder: a systematic review and meta-analysis. Expert Rev Pharmacoecon Outcomes Res 2018 Feb;18(1):25-41. [doi: 10.1080/14737167.2018.1407245] [Medline: 29145746]
10. Hedman E, Ljótsson B, Lindefors N. Cognitive behavior therapy via the internet: a systematic review of applications, clinical efficacy and cost-effectiveness. Expert Rev Pharmacoecon Outcomes Res 2012 Dec;12(6):745-764. [doi: 10.1586/erp.12.67] [Medline: 23252357]
11. Smith AF, Brown GC. Understanding cost effectiveness: a detailed review. Br J Ophthalmol 2000 Jul;84(7):794-798. [doi: 10.1136/bjo.84.7.794] [Medline: 10873998]
12. König H, König HH, Konnopka A. The excess costs of depression: a systematic review and meta-analysis. Epidemiol Psychiatr Sci 2020;29:e30. [doi: 10.1017/S2045796019000180]
13. Dewa CS, Lin E. Chronic physical illness, psychiatric disorder and disability in the workplace. Soc Sci Med 2000 Jul;51(1):41-50. [doi: 10.1016/S0277-9536(99)00431-1]
14. Hendriks SM, Spijker J, Licht CMM, et al. Long-term work disability and absenteeism in anxiety and depressive disorders. J Affect Disord 2015 Jun 1;178:121-130. [doi: 10.1016/j.jad.2015.03.004] [Medline: 25805404]
15. Knudsen AK, Harvey SB, Mykletun A, Øverland S. Common mental disorders and long-term sickness absence in a general working population. The Hordaland Health Study. Acta Psychiatr Scand 2013 Apr;127(4):287-297. [doi: 10.1111/j.1600-0447.2012.01902.x] [Medline: 22775341]
16. Beck A, Crain AL, Solberg LI, et al. Severity of depression and magnitude of productivity loss. Ann Fam Med 2011;9(4):305-311. [doi: 10.1370/afm.1260] [Medline: 21747101]
17. Erickson SR, Guthrie S, Vanetten-Lee M, et al. Severity of anxiety and work-related outcomes of patients with anxiety disorders. Depress Anxiety 2009;26(12):1165-1171. [doi: 10.1002/da.20624] [Medline: 19842165]
18. Walker-Bone K. The Work Productivity and Activity Impairment (WPAI) questionnaire. Occup Med (Lond) 2025;14:kqaf097. [doi: 10.1093/occmed/kqaf097] [Medline: 40971698]
19. Reilly MC, Zbrozek AS, Dukes EM. The validity and reproducibility of a work productivity and activity impairment instrument. Pharmacoeconomics 1993 Nov;4(5):353-365. [doi: 10.2165/00019053-199304050-00006] [Medline: 10146874]
20. de Oliveira C, Saka M, Bone L, Jacobs R. The role of mental health on workplace productivity: a critical review of the literature. Appl Health Econ Health Policy 2023 Mar;21(2):167-193. [doi: 10.1007/s40258-022-00761-w] [Medline: 36376610]
21. Rojanasarot S, Bhattacharyya SK, Edwards N. Productivity loss and productivity loss costs to United States employers due to priority conditions: a systematic review. J Med Econ 2023;26(1):262-270. [doi: 10.1080/13696998.2023.2172282] [Medline: 36695516]
22. Attridge M. Internet-based cognitive-behavioral therapy for employees with anxiety, depression, social phobia, or insomnia: clinical and work outcomes. Sage Open 2020 Jan;10(1):2158244020914398. [doi: 10.1177/2158244020914398]
23. Bondar J, Babich Morrow C, Gueorguieva R, et al. Clinical and financial outcomes associated with a workplace mental health program before and during the COVID-19 pandemic. JAMA Netw Open 2022 Jun 1;5(6):e2216349. [doi: 10.1001/jamanetworkopen.2022.16349] [Medline: 35679044]
24. Ward EJ, Fragala MS, Birse CE, et al. Assessing the impact of a comprehensive mental health program on frontline health service workers. PLOS ONE 2023;18(11):e0294414. [doi: 10.1371/journal.pone.0294414] [Medline: 37988363]
25. Carolan S, Harris PR, Cavanagh K. Improving employee well-being and effectiveness: systematic review and meta-analysis of web-based psychological interventions delivered in the workplace. J Med Internet Res 2017 Jul 26;19(7):e271. [doi: 10.2196/jmir.7583] [Medline: 28747293]
26. Moe-Byrne T, Shepherd J, Merecz-Kot D, et al. Effectiveness of tailored digital health interventions for mental health at the workplace: a systematic review of randomised controlled trials. PLOS Digit Health 2022 Oct;1(10):e0000123. [doi: 10.1371/journal.pdig.0000123] [Medline: 36812547]
27. Persson Asplund R, Asplund S, von Buxhoeveden H, et al. Work-focused versus generic internet-based interventions for employees with stress-related disorders: randomized controlled trial. J Med Internet Res 2023 Apr 25;25:e34446. [doi: 10.2196/34446] [Medline: 37097739]
28. Kechagias EP, Papadopoulos GA, Rokai I. Evaluating the impact of digital health interventions on workplace health outcomes: a systematic review. Adm Sci 2024;14(6):131. [doi: 10.3390/admsci14060131]
29. Forman-Hoffman VL, Flom M, Montgomery R, Robinson A. Improvements in work productivity and activity impairment among adults with anxiety or depressive symptoms participating in a relational agent-delivered digital mental health intervention. J Occup Environ Med 2024 Mar 1;66(3):e99-e105. [doi: 10.1097/JOM.0000000000003038] [Medline: 38242139]
30. Paterson C, Leduc C, Maxwell M, et al. Barriers and facilitators to implementing workplace interventions to promote mental health: qualitative evidence synthesis. Syst Rev 2024 Jun 7;13(1):152. [doi: 10.1186/s13643-024-02569-2] [Medline: 38849924]

XSL•FO
RenderX

31. Evans-Lacko S, Knapp M. Global patterns of workplace productivity for people with depression: absenteeism and presenteeism costs across eight diverse countries. Soc Psychiatry Psychiatr Epidemiol 2016 Nov;51(11):1525-1537. [doi: 10.1007/s00127-016-1278-4] [Medline: 27667656]
32. Harty S, Enrique A, Akkol-Solakoglu S, et al. Implementing digital mental health interventions at scale: one-year evaluation of a national digital CBT service in Ireland. Int J Ment Health Syst 2023 Oct 10;17(1):29. [doi: 10.1186/s13033-023-00592-9] [Medline: 37817270]
33. Richards D, Enrique A, Eilert N, et al. A pragmatic randomized waitlist-controlled effectiveness and cost-effectiveness trial of digital interventions for depression and anxiety. NPJ Digit Med 2020;3(1):85. [doi: 10.1038/s41746-020-0293-8] [Medline: 32566763]
34. Husereau D, Drummond M, Augustovski F, et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations. MDM Policy Pract 2022;7(1). [doi: 10.1177/23814683211061097] [Medline: 35036563]
35. Richards D, Richardson T, Timulak L, McElvaney J. The efficacy of internet-delivered treatment for generalized anxiety disorder: a systematic review and meta-analysis. Internet Interv 2015 Sep;2(3):272-282. [doi: 10.1016/j.invent.2015.07.003]
36. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
37. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
38. Wijesuriya R, Moreno-Betancur M, Carlin JB, White IR, Quartagno M, Lee KJ. Multiple imputation for longitudinal data: a tutorial. Stat Med 2025 Feb 10;44(3-4):e10274. [doi: 10.1002/sim.10274] [Medline: 39846338]
39. Cohen J. Statistical Power Analysis for the Behavioral Sciences: Routledge Academic; 1988.
40. Lofland JH, Pizzi L, Frick KD. A review of health-related workplace productivity loss instruments. Pharmacoeconomics 2004;22(3):165-184. [doi: 10.2165/00019053-200422030-00003] [Medline: 14871164]
41. Structure of earnings survey. Eurostat. URL: https://ec.europa.eu/eurostat/web/microdata/structure-of-earnings-survey [accessed 2026-03-23]
42. US dollar (USD). European Central Bank. URL: https://www.ecb.europa.eu/stats/policy_and_exchange_rates/euro_reference_exchange_rates/html/eurofxref-graph-usd.en.html [accessed 2026-03-30]
43. Finnes A, Enebrink P, Ghaderi A, Dahl J, Nager A, Öst LG. Psychological treatments for return to work in individuals on sickness absence due to common mental disorders or musculoskeletal disorders: a systematic review and meta-analysis of randomized-controlled trials. Int Arch Occup Environ Health 2019 Apr;92(3):273-293. [doi: 10.1007/s00420-018-1380-x] [Medline: 30474734]
44. Joyce S, Modini M, Christensen H, et al. Workplace interventions for common mental disorders: a systematic meta-review. Psychol Med 2016 Mar;46(4):683-697. [doi: 10.1017/S0033291715002408] [Medline: 26620157]
45. Lee YC, Gao L, Dear BF, Titov N, Mihalopoulos C. The Cost-effectiveness of the online MindSpot Clinic for the treatment of depression and anxiety in Australia. J Ment Health Policy Econ 2017 Dec 1;20(4):155-166. [Medline: 29300702]
46. Penev T, Zhao S, Lee JL, Chen CE, Metcalfe L, Ozminkowski RJ. The impact of a workforce mental health program on employer medical plan spend: an application of cost efficiency measurement for mental health care. Popul Health Manag 2023 Feb;26(1):60-71. [doi: 10.1089/pop.2022.0240] [Medline: 36799934]
47. Torous J, Jän Myrick K, Rauseo-Ricupero N, Firth J. Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow. JMIR Ment Health 2020 Mar 26;7(3):e18848. [doi: 10.2196/18848] [Medline: 32213476]
48. Webb CA, Rosso IM, Rauch SL. Internet-based cognitive-behavioral therapy for depression: current progress and future directions. Harv Rev Psychiatry 2017;25(3):114-122. [doi: 10.1097/HRP.0000000000000139] [Medline: 28475503]
49. Palacios JE, Enrique A, Mooney O, et al. Durability of treatment effects following internet-delivered cognitive behavioural therapy for depression and anxiety delivered within a routine care setting. Clin Psychol Psychother 2022 Sep;29(5):1768-1777. [doi: 10.1002/cpp.2743] [Medline: 35466486]

Abbreviations

CBT: cognitive behavioral therapy
CHEERS: Consolidated Health Economic Evaluation Reporting Standards
CMD: common mental disorder
GAD-7: Generalized Anxiety Disorder-7
HSE: Health Service Executive
iCBT: internet-based cognitive behavioral therapy
PHQ-9: Patient Health Questionnaire-9
WPAI: Work Productivity and Activity Impairment

XSL•FO
RenderX

Edited by S Brini; submitted 16.Jul.2025; peer-reviewed by A Adepoju, R Singh; accepted 05.Jan.2026; published 07.Apr.2026.

Please cite as:
Lee CT, Harty S, Enrique A, Jiménez-Díaz A, Hisler G, Duffy D, Richards D
Changes in Workplace Productivity and Estimated Cost Savings During Internet-Based Cognitive Behavioral Therapy in the Irish National Health Service: Naturalistic, Repeated-Measures, Retrospective Survey Study
J Med Internet Res 2026;28:e80689
URL: https://www.jmir.org/2026/1/e80689
doi:10.2196/80689

© Chi Tak Lee, Siobhán Harty, Angel Enrique, Alba Jiménez-Díaz, Garrett Hisler, Daniel Duffy, Derek Richards. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Goal Setting and Anchoring Effects on Meditation Using a Digital Platform: Large-Scale Digital Field Study

Michael Bowen[1,2], MS; Michael Beam[1], PhD

[1]Kent State University, 1800 E Main Street, Kent, OH, United States

[2]Spotify (Sweden), Stockholm, Sweden

Corresponding Author:
Michael Bowen, MS
Kent State University, 1800 E Main Street, Kent, OH, United States

Abstract

Background: Meditation has grown in popularity in recent years, but many people who try meditation often fail to establish a habit. Goal setting has been demonstrated to be an effective technique in behavior change in other health-related contexts but is understudied in the meditation context.

Objective: This study had 2 objectives: (1) to assess the association between goal setting and the number of days people meditated and (2) to evaluate whether anchoring bias in the goal-setting question (via response option order) influences goal selection and subsequent meditation behavior.

Methods: This large-scale quasi-experimental field study included 18,559 Spotify mobile users aged 18 years or older residing in Australia, Canada, New Zealand, the United Kingdom, or the United States who had listened to at least 5 minutes of meditation content from a specified teacher. The in-app experiment consisted of 2 goal-setting test conditions and an active control. In the test conditions, participants selected the number of days they intended to listen to content from the meditation teacher in the next 7 days. The conditions differed only in the order of goal response options (higher goals listed first vs last). The active control rated how much they liked the teacher but did not set a goal. Because responding was optional, selection bias is possible, and the design is quasi-experimental.

Results: The act of setting any goal had a modest positive association with the number of days people meditated in both treatment condition 1 (β=.08, 95% CI 0.01-0.16) and treatment condition 2 (β=.08, 95% CI 0.002-0.15). People who committed to higher goals were also more likely to meditate more than those who committed to lower goals. Additionally, the distribution of goals between the treatment conditions varied (χ^2_2=84.24; P<.001), and the differences in these distributions subsequently yielded differences in the number of days each group meditated, on average ($t_{2744.1}$=−2.34; P=.02; Cohen d=−0.09). Ultimately, placing the highest goal as the first answer choice yielded higher average active days among those who chose a goal, but many more people opted out of answering the question itself.

Conclusions: Simply offering an optional in-app goal-setting prompt, the intention-to-treat estimate from the experiment, did not change meditation engagement at the population level. However, among users who chose to respond and set a goal, goal selection was associated with a modest increase in active days. Response option order (anchoring) shifted both goal selection and opt-out rates. These findings highlight an uptake-engagement trade-off that is relevant for digital behavior change design.

(J Med Internet Res 2026;28:e85801) doi:10.2196/85801

KEYWORDS

goal setting; commitment; precommitment; anchoring; nudging; meditation; mindfulness; app; mobile

Introduction

Background

Meditation has many demonstrated health benefits [1-3], and in recent years, meditation has been increasing in popularity [4]. Many meditators use meditation apps and streaming services to support their practices [5,6]; however, many people who have tried meditation no longer meditate [6], and retention with digital meditation apps and content, in particular, is low [7-9]. Evidence suggests that greater home meditation practice frequency is associated with better outcomes, highlighting the importance of continued practice [10,11]. Developing a new habit, such as meditation, requires repeated behavior that can develop into automaticity [12-15]. One effective strategy to support this process is planning ahead and setting specific goals, typically expressed in terms of when, where, and how much one intends to engage in the behavior. Goal setting has been shown to increase engagement across a variety of domains,

XSL•FO
RenderX

including physical activity, healthy eating, and other health-promoting behaviors [16-19].

Goal setting is an important determinant of goal commitment and pursuit [20], which is a key component of many behavior change theories, both directly in social cognitive theory [21] and through related constructs, such as perceived behavioral control in the theory of planned behavior [22]. For instance, being actively involved in one's own goal setting can yield positive outcomes [20], and stating a goal publicly can enhance commitment to the goal [23], consequently increasing the belief in one's ability to achieve the goals. As such, understanding the influence of goal setting in behavior change is important to test key theories of behavioral regulation and habit formation.

Within the context of meditation, action and coping planning strategies and attitudinal commitments have been positively associated with the number of days practicing meditation [24,25]. However, explicit goal setting and goal-intention effects in meditation are understudied, particularly through rigorous techniques. This knowledge gap is especially noteworthy because of the role that goal setting can play in prominent behavior change theories (eg, social cognitive theory, theory of planned behavior). Additionally, goal setting is associated with behavioral outcomes, and practice frequency is crucial for realizing the mental health benefits of meditation, including benefits from using meditation apps and digital content [26-30].

When setting goals, the measurement interface used to elicit the goal is also an important and understudied component of the experience, particularly in the context of meditation. Cognitive psychology has demonstrated the influence of anchoring. For example, the starting points and positions of choices and the subsequent effects these have on people's judgments and choices [31]. Along these lines, in surveys and survey question design, anchoring biases are a well-known phenomenon that can impact the response distribution [32-34]. In fact, people even rate their success in life differently in surveys when the numeric values on a scale differ [35]. These effects may be explained by dual-process models of judgment, in which people rely on fast, heuristic processing when making quick decisions [31]. In the case of goal commitment within a mobile app, it is plausible that individuals "satisfice" by selecting the first option they come across that seems reasonable, thus making response order a plausible driver of goal commitment. From the perspective of goal-setting theory, anchors that shift potential goals may also change perceived goal difficulty and feasibility, which may influence commitment, self-efficacy, and intention formation [20].

When the objective of asking a question is not explicitly to gather unbiased data for research purposes but rather to encourage people to meditate more, standard survey best practices, such as choice randomization, may be less important than optimizing the goal-setting intervention to outcomes. In a way, these techniques are commonly found in push polls, which are not scientific surveys but rather political persuasion techniques framed as surveys. These kinds of techniques, in which questions have an overt design bias, have been demonstrated to be effective for shifting both memory and opinions [36,37]. These kinds of techniques have also been used in the health behavior goal-setting context. For example, introducing anchoring has demonstrated that people set higher goals when questions are framed as sacrifices (what to forgo) rather than actions (what to undertake) [38]. Consequently, it is important for theorists, applied researchers, and user experience (UX) designers to further understand the effects of anchoring in goal setting, including in the context of meditation. By doing so, these insights can be leveraged for positive impacts on people's mental health.

Research Purpose

With these 2 streams of literature in mind, the purpose of this research was twofold. The first and primary objective was to identify the behavioral association between goal setting and engaging with meditation content through a large-scale digital platform. In doing so, we assessed whether people who preemptively set a goal are more likely to meditate than people who did not select a meditation goal. This objective resulted in the following hypothesis:

- H1: People in the treatment conditions who preemptively state a meditation goal will have more days listening to guided meditation teachings during the 7 days following the intervention, on average, compared to the control group.

The second objective was to explore the effects of anchoring on goal intentions, specifically, whether anchoring different goals in the goal-setting question changes the ways in which people set their goals and subsequently how the anchoring effect impacts meditation behaviors. This objective resulted in the following research hypotheses:

- H2: The effects of anchoring preemptively stated meditation goals in the goal-setting question will result in a different distribution of goal intentions.
- H3: Among people who preemptively state a goal, the effects of anchoring different meditation goals in the goal-setting question will result in varying levels of active days with the meditation content during the 7 days following the intervention.

By exploring these hypotheses, this research can support meditation and mental health app product managers in considering whether to implement goal-setting features in their UX and to provide guidance on how to design the feature. Additionally, the present study can help future researchers who study meditation habit formation, particularly with digital health platforms, understand the role of goal setting within the broader context of behavior change.

Methods

Sampling and Recruitment

This research was conducted among 18,559 users of the Spotify mobile app. To qualify to participate in the study, individuals needed to have listened to at least 5 minutes of content provided by a popular meditation teacher through the Spotify app during the month of June 2025. The content offered by the meditation teacher consisted of both guided meditations and teachings. Additional qualifications included being aged 18 years or older and residing in the United States, United Kingdom, Canada,

XSL•FO
RenderX

Australia, or New Zealand. These countries were chosen because Spotify is available in these countries, the meditation content is in English, and these countries are predominantly English-speaking. Fieldwork was conducted from July 2, 2025, to July 8, 2025. Table 2 in the Results section includes the total sample sizes, country, age, and gender composition of the research sample for each experimental condition. Note that chi-square tests indicated no differences between cells based on country, age, and gender identification.

As the primary experimental intervention, qualified users were assigned to receive 1 of 3 potential one-tap survey questions. Two of the survey questions were treatment conditions, and one was the active control condition. The one-tap survey questions were embedded within a pop-up message in the Spotify mobile app that appeared when a qualified research participant visited the Spotify homepage. Participants could choose to answer the question or dismiss the message without selecting an answer option. Additionally, participants could select a "Prefer not to say" response option if they did not want to select a meditation goal. Upon answering the question, the message was subsequently dismissed automatically, thus ending the intervention.

Because the survey question was optional and, subsequently, goal intention was dependent on a post random assignment event, there was a selection bias in the design when analyzing only people who clicked on the message itself. We detail how this was handled and the implications on the results and interpretation in the Data Analysis Plan section.

Meditation Content

The meditation content evaluated in this study consisted of a large, publicly available audio catalog produced by a single teacher and distributed as a long-running mindfulness podcast. The catalog includes both (1) guided meditation sessions and (2) didactic "teachings" that provide instruction and reflection intended to support practice. Content primarily reflects mindfulness-based approaches and frequently incorporates self-compassion–oriented instruction (eg, attention to breath/body, present-moment awareness, and working with difficult emotions). Episodes vary in length but tend to be between 20 and 90 minutes. The catalog spans more than a decade and contains over 1500 audio items, often with multiple new items released per week.

Experimental Design

Each test condition comprised a single survey question. The 2 treatment conditions were variants of the goal-setting question. Specifically, each goal-setting message requested the participant to select the number of days in the next 7 days they intended to listen to guided meditations and teachings offered by the teacher on Spotify. The variants were nearly identical, with only the question response options scale being flipped, which was designed to test the anchoring effect of having the highest vs the lowest number of days response option at the top of the response battery.

Specifically, treatment 1 asked how many days in the next 7 days the user intended to listen to meditation and teachings offered by the teacher on Spotify. Seven days were chosen because goal-setting theory argues that proximal (short-term) goals can facilitate the performance of complex tasks [20], of which meditation could reasonably be considered. The question response options were listed from top to bottom as "0 days," "1 day," "2-3 days," "4 days or more," and "Prefer not to say." In treatment 2, the question itself was the same, but the options were reversed, so from top to bottom, the options were listed as "4 days or more," "2-3 days," "1 day," "0 days," and "Prefer not to say." Note that selecting "Prefer not to say" was offered as a survey response and was not the same interaction as dismissing the message without selecting any response. However, selecting "Prefer not to say" was not treated as having selected a goal.

The control condition asked participants how they felt about the meditation and teachings offered by the same meditation teacher on Spotify, with the response options listed from top to bottom as "I love them," "I like them," "I have no opinion," "I dislike them," and "Prefer not to say." In using this type of active control, we ensured that the treatment and control conditions had similar interventions, thus equalizing the cognitive effort of responding to a question, controlling for question-behavior effects (where answering any question can influence subsequent behavior), and avoiding inadvertently inducing goal-setting or planning cognitions.

See Table 1 for sample sizes based on participants' interactions with the goal-setting question.

Table . Sample sizes based on message interactions.

	Saw the message (all participants), n (%)	Survey responders[a], n (%)	Selected a goal[b], n (%)
Control	6012 (100)	2428 (40.4)	2249 (37.4)
Treatment 1	6320 (100)	2065 (32.7)	1542 (24.4)
Treatment 2	6227 (100)	2155 (34.6)	1349 (21.7)

[a]Participants who did not dismiss the message and chose an answer to the one-tap survey question.

[b]Participants who did not dismiss the message and did not select "Prefer not to say" as their response.

Following this brief survey-based intervention, we used trace behavioral data from the Spotify app to assess how often users streamed the specific content offered by the respective meditation teacher on the Spotify app.

Active Days Variable

We define active days as the number of days a participant streamed content from the meditation teacher on Spotify. The

XSL•FO
RenderX

postintervention active-days measure (days 0 - 7 after the intervention) is the dependent variable for H1 and H3. The preintervention active-days measure (the 30 days preceding the intervention) is also included as an independent covariate. To be considered active on a given day, the participant needed to have streamed content from the teacher for any period of time on the respective day (having pressed play and streamed the content for >0 ms). While this is not necessarily indicative of a meditation session having occurred, it avoids otherwise arbitrary definitions of an active day and is a strong indicator of intent. Summary statistics by cohort are shown in Table 3 in the Results section. In all models, both measures were treated as integer counts.

Data Analysis Plan

The analysis for this study was conducted using Python. The click-through rate of the survey question (the percentage of people who saw the survey question and subsequently chose a response option, as opposed to dismissing it without responding) for each of treatment 1, treatment 2, and the control was 32.7% (2065/6320), 34.6% (2155/6227), and 40.4% (2428/6012), respectively. A chi-square test indicates that these click-through rates are different (χ^2_3=85.71; P<.001). As a result, there is a selection bias introduced into the experimental design. So, while the overall study is a natural field experiment, the analysis of the effects of goal setting is quasi-experimental and should be considered associations.

To test H1, specifically whether people in the treatment conditions who preemptively state a meditation goal will have more days listening to guided meditation teachings during the 7 days following the intervention, on average, compared to the control, a linear regression model was used. The dependent variable in the model was the number of active days on which the participant listened to content from the meditation teacher through the Spotify app during the 7 days after the intervention. While active days is a bounded and nonnormally distributed variable, linear regression is appropriate because it provides a parsimonious and interpretable approximation for mean differences and has been demonstrated to be valid in these conditions when sample sizes are large [39]. This was coded as an integer, with values ranging from 0 to 8. The range is 0 to 8, not 0 to 7, because the day of the intervention, at any point after the question was answered, is included in the count as day 0. Each treatment condition was inserted separately as an independent variable in the model, while the control condition was held out as the reference category. Note that only people who selected a goal were included in the model. To control for past engagement with the meditation content, the number of active days during the 30 days preceding the treatment that the participant listened to meditation content from the meditation teacher through the Spotify app was included as an independent variable. Age was binned into 5 groups and one-hot encoded as 18 to 24, 25 to 34, 35 to 44, 45 to 54, and aged 55 years or older. The 55 years and older group was held out of the model as the reference category. Gender values included female, male, and other and were also one-hot encoded for modeling. Female was held out as the reference category. Country was also inserted into the model, with the United States held out as the reference category.

H2 explored the role of anchoring on the response distributions of goals chosen. To test this hypothesis, we explored the percentage of users who selected each response option in the survey and applied a chi-square test of independence to specifically assess whether these distributions varied.

To test H3, specifically that the anchoring effect between treatment cells would result in different levels of active days between treatment conditions, a 2-sample Welch t test was used to assess differences in postintervention active days.

Ethical Considerations

The study was reviewed and approved as exempt by the Kent State University Institutional Review Board (IRB) under reference number 2145. The consent process was governed by Spotify's Terms of Service and Privacy Policy, as the experiment was conducted on the platform with its members. The research team and the IRB carefully reviewed these policies and the ethical conduct of the study, concluding that consent was obtained through the Terms of Service and Privacy Policy. The IRB also determined that this study qualifies as a benign behavioral intervention. Personally identifiable information about Spotify users was not accessible to the researchers and was safeguarded through encryption. It can only be accessed through a review and approval process by the Spotify Data Protection Office. No compensation for participation was provided.

Results

Table 2 provides the demographic composition of the sample. As previously noted, chi-square tests indicated no differences between cells based on country, age, and gender identification.

Table 3 provides the descriptive statistics for the number of active days before and after the treatment based on the varying levels of engagement with the message.

XSL•FO
RenderX

Table . Demographic composition of experimental conditions.

	Control (N=6012), n (%)	Treatment 1 (N=6320), n (%)	Treatment 2 (N=6227), n (%)
Country			
Australia	631 (10.5)	653 (10.3)	632 (10.1)
Canada	756 (12.6)	767 (12.1)	787 (12.6)
United Kingdom	819 (13.6)	897 (14.2)	917 (14.7)
United States	3684 (61.3)	3867 (61.2)	3733 (59.9)
New Zealand	122 (2)	136 (2.2)	158 (2.5)
Age (years)			
18-24	234 (3.9)	233 (3.7)	244 (3.9)
25-34	1520 (25.3)	1570 (24.8)	1520 (24.4)
35-44	1898 (31.6)	1975 (31.3)	1935 (31.1)
45-54	1223 (20.3)	1272 (20.1)	1275 (20.5)
≥55	1137 (18.9)	1270 (20.1)	1253 (20.1)
Gender identification[a]			
Female	4453 (74.1)	4601 (72.8)	4605 (74)
Male	1257 (20.9)	1394 (22.1)	1344 (21.6)
Other/prefer not to say	59 (1)	72 (1.1)	53 (1)

[a]Gender data is collected at the time of Spotify registration. It is provided to Spotify optionally and is not available for all participants.

Table . Summary statistics for "active days."

Time window	Role in models	Cohort	Mean (SD)
7 days postintervention	Dependent (H1, H3)	All participants	0.95 (1.5)
7 days postintervention	Dependent (H1, H3)	Survey responders[a]	1.11 (1.62)
7 days postintervention	Dependent (H1, H3)	Selected a goal[b]	1.18 (1.66)
30 days preintervention	Independent covariate	All participants	3.93 (4.31)
30 days preintervention	Independent covariate	Survey responders[a]	4.27 (4.52)
30 days preintervention	Independent covariate	Selected a goal[b]	4.49 (4.66)

[a]Participants who did not dismiss the message and chose an answer to the one-tap survey question.

[b]Participants who did not dismiss the message and did not select "Prefer not to say" as their response.

Before testing the specific hypotheses listed previously, we first explored whether simple exposure to the commitment message (ie, participants saw the message) affected the likelihood of participants in the control versus treatment conditions to have more active days following the intervention. This could be considered the intention-to-treat effect from the messages and the results of the true field experiment component of the research. A Welch *t* test showed that the overall mean number of active days between the control condition (mean 0.94, SD 1.48 days) and the treatment conditions combined (mean 0.96, SD 1.51 days) was not different ($t_{12,045.6}$=−0.76; *P*=.45).

In the linear regression used to test H1, participants in both treatment condition 1 (β=.08, 95% CI .01-.16) and treatment condition 2 (β=.08, 95% CI .002-.15) had more average active days with the meditation content compared to the control group. The model controlled for preintervention activity levels, as well as the specific goals and answer choices participants selected, among others, indicating that simply selecting a goal is associated with increased active days. Therefore, we can reject the null hypothesis for H1.

Additionally, people who set higher goals for themselves tended to have more active days on average. The adjusted R^2 for this model was 0.396. See Table 4 for more details on the linear regression model of days active after the intervention.

XSL•FO
RenderX

Table . Coefficient, upper and lower bound CI (95%), and *P* value from linear regression of posttreatment days active.

Variable	Coefficient (95% CI)	*P* value
Constant	0.19 (0.08 to 0.30)	<.001
Preexposure days active	0.21 (0.20 to 0.22)	<.001
Experimental condition		
Control	__[a]	—
Treatment 1	0.08 (0.01 to 0.16)	.03
Treatment 2	0.08 (0.002 to 0.15)	.05
Country		
Australia	0.09 (−0.04 to 0.21)	.17
Canada	0.09 (−0.03 to 0.21)	.13
United Kingdom	0.05 (−0.06 to 0.16)	.34
United States	—	—
New Zealand	0.28 (0.06 to 0.50)	.01
Gender identification		
Female	—	—
Male	0.08 (0 to 0.17)	.06
Age (y)		
18 - 24	−0.36 (−0.54 to −0.19)	<.001
25 - 34	−0.22 (−0.33 to −0.11)	<.001
35 - 44	−0.22 (−0.33 to −0.11)	<.001
45 - 54	−0.07 (−0.19 to 0.04)	.22
55 or older	—	—
Goal intention survey response		
0 days	−0.21 (−0.29 to −0.14)	<.001
1 day	−0.05 (−0.12 to 0.03)	.23
2 - 3 days	0.08 (0 to 0.16)	.04
4 days or more	0.34 (0.23 to 0.44)	<.001
Prefer not to say	—	—
Control survey response		
I dislike them	0.01 (−0.34 to 0.37)	.94
I have no opinion	−0.10 (−0.29 to 0.09)	.30
I like them	−0.02 (−0.15 to 0.10)	.70
I love them	0.14 (0.05 to 0.23)	.003
Prefer not to say	—	—

[a]Not applicable.

To test whether the ordering of the goal response options created an anchoring effect that influenced the distribution of responses in H2, a chi-square test of independence was used. Figure 1 includes the distribution of responses from the survey within the 2 test conditions. The test yielded (χ^2_2=84.24; *P*<.001), indicating significant differences in the distribution of responses across treatment conditions. When the "Prefer not to say" response is removed from the distribution and the chi-square statistic recalculated, the result yielded (χ^2_2=13.03; *P*=.005), indicating that the anchoring effect persists. As a result, we can reject the null hypothesis for H2.

XSL•FO
RenderX

Figure 1. Distribution of goal intention one-tap survey responses based on treatment condition (percentage of users who chose their respective meditation goal by treatment condition).

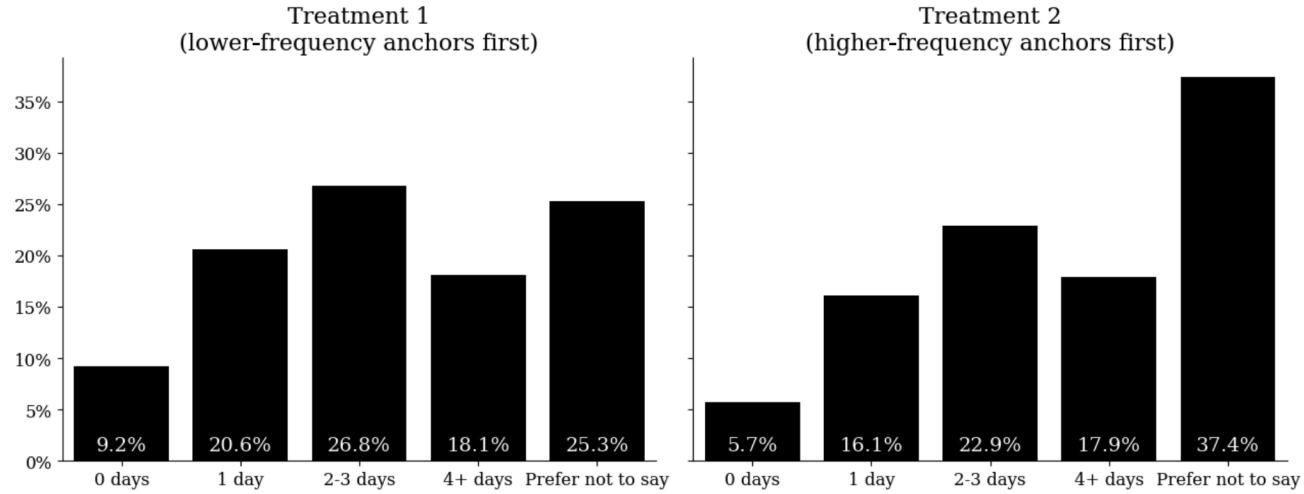

In H3, the Welch *t* test showed that the mean number of active days differed between treatment 1 (mean 1.15, SD 1.62 days) and treatment 2 (mean 1.3, SD 1.78 days; $t_{2744.1}$=−2.34; *P*=.02). A Cohen *d* of −0.09 indicates a modest overall effect of the differences in days active between treatment conditions. As a result, we can reject the null hypothesis for H3. It is important to note, however, that when we included people who did not select a goal (ie, instead they selected "Prefer not to say"), the effect is not significant ($t_{4217.7}$=−0.68; *P*=.50).

Table 5 provides a summary of the hypotheses, test results, and decisions. Additionally, a CONSORT (Consolidated Standards of Reporting Trials) diagram of the full experimental setup is available in Multimedia Appendix 1.

Table . Summary of hypotheses, test results, and decisions.

Hypothesis	Hypothesis (abridged)	Test (*P* value)	Decision
H1	People who preemptively state a meditation goal will have more days listening to guided meditation teachings during the 7 days following the intervention, on average, compared to the control.	Participants who stated a meditation goal were more likely to engage with the meditation content in both Treatment 1 (β=.08, 95% CI 0.01-0.16) and Treatment 2 (β=.08, 95% CI 0.002-0.15), compared to the control.	Reject null hypothesis
H2	The effects of anchoring different meditation goals in the treatment conditions will result in a different distribution of goal intentions.	χ^2 (2) = 84.24 (*P*<.001)	Reject null hypothesis
H3	The effects of anchoring different meditation goals in the goal-setting question will result in varying levels of active days with the meditation content during the 7 days following the intervention.	*t* (2744.1) = −2.34 (*P*=.02)	Reject null hypothesis

Discussion

Principal Results

This study demonstrated that setting short-term goals for guided meditation and teachings through a mobile app was associated with engaging with the meditation content on more days. Specifically, people who preemptively set themselves any goal were more likely to meditate on more days than people in the control condition. Additionally, people who chose higher goals were more likely to engage with meditation content more often, even after controlling for past engagement with the meditation content, as well as age, gender, and country of residence. Conversely, people who selected zero as a goal meditated less often. While the differences were modest, these uptake-conditional results suggest that among participants who elected to respond, goal selection was associated with slightly more active days. The intention-to-treat analysis indicated that simply offering the optional goal-setting prompt (thus including people who chose to ignore the message) did not influence subsequent meditation behavior.

This finding adds to the body of literature that goal setting may be useful in health behavior change in the context of meditation, similar to other health behavior contexts such as physical activity and nutrition [16-19]. Like Epton et al [16] found in their meta-analysis of the effects of goal setting across a range of behaviors, the effect of goal setting from this research was found to be small and should subsequently be considered just one, perhaps small, component of meditation habit formation.

XSL•FO
RenderX

However, while the overall effects may be small, at scale, these kinds of cost-effective nudges can be worthwhile interventions [40]. For meditation content providers and program managers, these findings provide evidence that goal-setting interventions may be useful, if modest, tools to increase engagement with meditation content. Ultimately, for anyone who has a desire to develop a meditation habit, this study provides compelling data that goal setting may be worthwhile in building and maintaining a meditation habit.

This study also provides clear evidence of an anchoring effect based on how the goal-intention question is structured. Specifically, when higher goal response options are anchored at the top of the response battery, the response distributions change substantively. Most notably, many more people selected "Prefer not to say" when the "4+ days per week" option was anchored at the top of the response battery compared to the bottom (37.4% vs 25.3%). This is likely explained by dual-process models of judgment, in which individuals rely on quick, heuristic processing when making low-effort decisions [31]. Within this study environment, individuals likely chose the first plausible goal. Seeing the most ambitious goal at the top may have felt unrealistic or hindered self-efficacy for some people, both of which are important components of goal setting and goal accomplishment [20]. Additionally, people who were in the Treatment 2 condition and selected a meditation goal (ie, did not select "Prefer not to say") subsequently engaged with meditation slightly more than people in Treatment 1 who selected a goal, suggesting that the anchoring effect may not have only affected the response distribution but also had a modest effect on people's propensity to meditate.

These findings indicate that there are trade-offs to anchoring. While it is best practice in survey research to randomize response options or randomly flip ordered scales in survey design [32-34], practitioners and UX designers should consider the outcomes they want to achieve when choosing anchors and whether to flip scales. Push polls are ethically fraught cautionary tales of the impact of asking biased questions [36,37]. However, in the context of health behavior, this technique may lead people to set more ambitious goals [38].

Overall, this study provides theoretical insights about the role of goal setting in meditation habit development and maintenance, adding to the body of literature about goal setting in different contexts. It also offers new insights into the effects of anchoring on goal setting within the meditation context. Finally, the research provides UX designers and product managers practical guidance on the potential impact of goal setting and anchoring on user engagement.

Limitations

There are several limitations to this research that should be considered. First, as described previously, because people could opt out of participating in the one-tap survey, there is a self-selection bias that makes this research a quasi-experimental design and not a full field experiment. Consequently, we cannot confidently conclude that we have established a causal association between goal setting and incremental meditation. While this is an important limitation in drawing conclusions, the nature of the research itself reflects real-world environments and digital platform experiences.

Second, this study was conducted on a single digital platform (Spotify) and includes only a single meditation teacher, which limits the broader generalizability of the findings. Spotify may be unique in that it primarily offers music, podcast, and audiobook content and is not broadly considered a digital mental health services provider. As such, goal-setting exercises on platforms that specialize in digital mental health services may differ. Additionally, the research includes only log behavioral data and does not include important context, such as motivations for meditation and meditation experience. Expanding this research to additional platforms and teachers, to more segments of the meditation population (eg, more experienced meditators who already have strong self-efficacy vs less experienced meditators who need structure), and with additional context (eg, mapping motivations to goal-setting outcomes) could be considered by future researchers.

Additionally, while the single-tap question is a strength of the study because it optimizes response rates and limits nonresponse bias, a single question served as a pop-up has formatting constraints. Consequently, we could not offer more detailed response options and could not probe more deeply into participants' relationship with meditation. Data of this kind may have been helpful to add more context to the insights offered from the study.

Future Directions

The results from this study open up additional avenues for exploration for future researchers. First, future studies may strive to make the research a fully experimental design by making the goal-setting question itself mandatory. This would allow researchers to better understand the full causal effects of goal setting within the meditation technology context. Additionally, as described, it would also be useful to study this phenomenon on other apps and in other contexts, which would allow researchers to generalize the effects of goal setting on meditation more broadly.

This research also looked only at the average effect of the treatment across experimental conditions. Future researchers might further explore how goal setting impacts different cohorts of meditators, such as lighter versus heavier meditators or people with a burgeoning versus an established practice. Doing so would allow theorists to better understand the effects of goal setting throughout the stages of meditation habit formation and for UX designers and product managers to strategically target their goal-setting user growth initiatives.

Conclusion

This research provides insights about the role of goal setting in the meditation context when applied to digital health platforms. Ultimately, goal setting may be a useful and cost-effective tool when applied on a large scale to encourage people to engage with meditation more frequently on digital platforms, although the overall impact may be modest. However, anchoring effects play a significant role in people's willingness to set meditation goals, the goals they set for themselves, and even their incremental meditation engagement. Researchers who study

XSL•FO
RenderX

the theoretical aspects of meditation habit formation can use this research to consider the role of goal setting in their theories and models. Clinical practitioners and UX designers should consider explicit goal setting when encouraging people to practice meditation but think carefully about the effects of anchoring in their designs.

Acknowledgments

This study would not have been possible without the support of Spotify. ChatGPT (GPT-5) was used to help refine the language in some sections of this work in an editorial manner after the original draft was written. The authors reviewed and edited all content and accept full responsibility for the final publication.

Funding

This publication was made possible in part by support from the Kent State University Open Access Publishing Fund and the KSU CCI Research and Creative Fund.

Data Availability

Aggregated outputs used in the manuscript are included in the paper; requests for additional access can be directed to mtbowen@kent.edu. However, approval is unlikely due to contractual and privacy constraints.

Authors' Contributions

Conceptualization: M Bowen
Data curation: M Bowen
Formal analysis: M Bowen
Funding acquisition: M Bowen
Methodology: M Bowen
Project administration: M Bowen
Resources: M Bowen
Software: M Bowen
Supervision: M Beam
Validation: both authors
Visualization: M Bowen
Writing – original draft: M Bowen
Writing – review & editing: both authors

Conflicts of Interest

M Bowen is employed by Spotify, which is a streaming service that primarily offers music, podcasts, and audiobook content. Some content on Spotify is meditation-related. However, the salary and equity of the author are not dependent upon the results of the research. M Beam declares no conflicts of interest.

Multimedia Appendix 1
CONSORT diagram of experimental design and outcomes.
[PNG File, 211 KB - jmir_v28i1e85801_app1.png]

References

1. Galante J, Friedrich C, Dawson AF, et al. Mindfulness-based programmes for mental health promotion in adults in nonclinical settings: a systematic review and meta-analysis of randomised controlled trials. PLoS Med 2021 Jan;18(1):e1003481. [doi: 10.1371/journal.pmed.1003481] [Medline: 33428616]
2. Goldberg SB, Tucker RP, Greene PA, et al. Mindfulness-based interventions for psychiatric disorders: a systematic review and meta-analysis. Clin Psychol Rev 2018 Feb;59:52-60. [doi: 10.1016/j.cpr.2017.10.011] [Medline: 29126747]
3. Luberto CM, Shinday N, Song R, et al. A systematic review and meta-analysis of the effects of meditation on empathy, compassion, and prosocial behaviors. Mindfulness (N Y) 2018 Jun;9(3):708-724. [doi: 10.1007/s12671-017-0841-8] [Medline: 30100929]
4. Davies JN, Faschinger A, Galante J, Van Dam NT. Prevalence and 20-year trends in meditation, yoga, guided imagery and progressive relaxation use among US adults from 2002 to 2022. Sci Rep 2024 Jul 1;14(1):14987. [doi: 10.1038/s41598-024-64562-y] [Medline: 38951149]
5. Lam SU, Xie Q, Goldberg SB. Situating meditation apps within the ecosystem of meditation practice: population-based survey study. JMIR Ment Health 2023 Apr 28;10:e43565. [doi: 10.2196/43565] [Medline: 37115618]

XSL•FO
RenderX

6. Bowen M, Beam M. Mapping mindfulness: assessing the stages of meditation habit formation in the USA using the Sussex Mindfulness Meditation (SuMMed) model. Mindfulness (N Y) 2025 May;16(5):1340-1351. [doi: 10.1007/s12671-025-02555-2]
7. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567. [doi: 10.2196/14567] [Medline: 31573916]
8. Kozlov E, Bantum E, Pagano I, et al. The reach, use, and impact of a free mHealth mindfulness app in the general population: mobile data analysis. JMIR Ment Health 2020 Nov 27;7(11):e23377. [doi: 10.2196/23377] [Medline: 33245289]
9. Bowen M, Beam M, Semb J, Yoo DW. Evaluating theory-driven messaging to overcome the barriers to meditation: large-scale digital field experiment. JMIR Ment Health 2025 Jul 7;12:e71732. [doi: 10.2196/71732] [Medline: 40698564]
10. Parsons CE, Crane C, Parsons LJ, Fjorback LO, Kuyken W. Home practice in mindfulness-based cognitive therapy and mindfulness-based stress reduction: a systematic review and meta-analysis of participants' mindfulness practice and its association with outcomes. Behav Res Ther 2017 Aug;95:29-41. [doi: 10.1016/j.brat.2017.05.004] [Medline: 28527330]
11. Crane C, Crane RS, Eames C, et al. The effects of amount of home meditation practice in mindfulness based cognitive therapy on hazard of relapse to depression in the Staying Well after Depression Trial. Behav Res Ther 2014 Dec;63:17-24. [doi: 10.1016/j.brat.2014.08.015] [Medline: 25261599]
12. Wood W, Neal DT. Healthy through habit: interventions for initiating & maintaining health behavior change. Behav Sci Policy 2016 Apr;2(1):71-83. [doi: 10.1177/237946151600200109]
13. Gardner B. A review and analysis of the use of "habit" in understanding, predicting and influencing health-related behaviour. Health Psychol Rev 2015;9(3):277-295. [doi: 10.1080/17437199.2013.876238] [Medline: 25207647]
14. Gardner B, Lally P. Modelling habit formation and its determinants. In: The Psychology of Habit: Theory, Mechanisms, Change, and Contexts: Springer International Publishing; 2018:207-229. [doi: 10.1007/978-3-319-97529-0_12]
15. Miles E, Matcham F, Strauss C, Cavanagh K. Making mindfulness meditation a healthy habit. Mindfulness (N Y) 2023 Dec;14(12):2988-3005. [doi: 10.1007/s12671-023-02258-6]
16. Epton T, Currie S, Armitage CJ. Unique effects of setting goals on behavior change: systematic review and meta-analysis. J Consult Clin Psychol 2017 Dec;85(12):1182-1198. [doi: 10.1037/ccp0000260] [Medline: 29189034]
17. da Silva MAV, São-João TM, Brizon VC, Franco DH, Mialhe FL. Impact of implementation intentions on physical activity practice in adults: a systematic review and meta-analysis of randomized clinical trials. PLoS One 2018;13(11):e0206294. [doi: 10.1371/journal.pone.0206294] [Medline: 30427874]
18. Adriaanse MA, Vinkers CDW, De Ridder DTD, Hox JJ, De Wit JBF. Do implementation intentions help to eat a healthy diet? A systematic review and meta-analysis of the empirical evidence. Appetite 2011 Feb;56(1):183-193. [doi: 10.1016/j.appet.2010.10.012] [Medline: 21056605]
19. Swann C, Hooper A, Schweickle MJ, et al. Comparing the effects of goal types in a walking session with healthy adults: preliminary evidence for open goals in physical activity. Psychol Sport Exerc 2020 Mar;47:101475. [doi: 10.1016/j.psychsport.2019.01.003]
20. Locke EA, Latham GP. Building a practically useful theory of goal setting and task motivation: a 35-year odyssey. Am Psychol 2002 Sep;57(9):705-717. [doi: 10.1037/0003-066X.57.9.705]
21. Bandura A. Health promotion from the perspective of social cognitive theory. Psychol Health 1998 Jul;13(4):623-649. [doi: 10.1080/08870449808407422]
22. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
23. Hollenbeck JR, Williams CR, Klein HJ. An empirical examination of the antecedents of commitment to difficult goals. J Appl Psychol 1989;74(1):18-23. [doi: 10.1037//0021-9010.74.1.18]
24. Cloughesy JN, Mrazek AJ, Mrazek MD, Schooler JW. Planning to practice: action and coping plans increase days of meditation practiced. Psi Chi J Psychol Res 2020;25(2):203-209. [doi: 10.24839/2325-7342.JN25.2.203]
25. Galla BM, Baelen RN, Duckworth AL, Baime MJ. Mindfulness, meet self-regulation: boosting out-of-class meditation practice with brief action plans. Motiv Sci 2016;2(4):220-237. [doi: 10.1037/mot0000045]
26. Cearns M, Clark SR. The effects of dose, practice habits, and objects of focus on digital meditation effectiveness and adherence: longitudinal study of 280,000 digital meditation sessions across 103 countries. J Med Internet Res 2023 Sep 19;25:e43358. [doi: 10.2196/43358] [Medline: 37725801]
27. Stecher C, Berardi V, Fowers R, Christ J, Chung Y, Huberty J. Identifying app-based meditation habits and the associated mental health benefits: longitudinal observational study. J Med Internet Res 2021 Nov 4;23(11):e27282. [doi: 10.2196/27282] [Medline: 34734826]
28. Zawadzki MJ, Torok ZA, Peña M, Gavrilova L. App-based mindfulness meditation reduces stress in novice meditators: a randomized controlled trial of headspace using ecological momentary assessment. Ann Behav Med 2025 Jan 4;59(1):kaaf025. [doi: 10.1093/abm/kaaf025] [Medline: 40257119]
29. Bostock S, Crosswell AD, Prather AA, Steptoe A. Mindfulness on-the-go: effects of a mindfulness meditation app on work stress and well-being. J Occup Health Psychol 2019 Feb;24(1):127-138. [doi: 10.1037/ocp0000118] [Medline: 29723001]
30. Goleman D, Davidson RJ. Altered Traits: Science Reveals How Meditation Changes Your Mind, Brain, and Body: Avery; 2017.

XSL•FO
RenderX

31. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. Science 1974 Sep 27;185(4157):1124-1131. [doi: 10.1126/science.185.4157.1124] [Medline: 17835457]
32. Krosnick JA, Alwin DF. An evaluation of a cognitive theory of response-order effects in survey measurement. Public Opin Q 1987;51(2):201-219. [doi: 10.1086/269029]
33. Holbrook AL, Krosnick JA, Moore D, Tourangeau R. Response order effects in dichotomous categorical questions presented orally: the impact of question and respondent attributes. Public Opin Q 2007;71(3):325-348. [doi: 10.1093/poq/nfm024]
34. Höhne J, Lenzner T. Investigating response order effects in web surveys using eye tracking. Psihologija 2015;48(4):361-377. [doi: 10.2298/PSI1504361H]
35. Schwarz N, Knauper B, Hippler HJ, Noelle-Neumann E, Clark L. Rating scales: numeric values may change the meaning of scale labels. Public Opin Q 1991;55(4):570-582. [doi: 10.1086/269282]
36. Murphy G, Lynch L, Loftus E, Egan R. Push polls increase false memories for fake news stories. Memory 2021 Jul;29(6):693-707. [doi: 10.1080/09658211.2021.1934033] [Medline: 34080495]
37. Erčulj V, Šulc A. Swayed by leading questions. Qual Quant 2025 Feb;59(1):191-209. [doi: 10.1007/s11135-024-01934-6]
38. Tuk MA, Prokopec S, Van den Bergh B. Do versus don't: the impact of framing on goal-level setting. J Consum Res 2021 Mar 20;47(6):1003-1024. [doi: 10.1093/jcr/ucaa050]
39. Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health 2002;23(1):151-169. [doi: 10.1146/annurev.publhealth.23.100901.140546] [Medline: 11910059]
40. Athey S, Grabarz K, Luca M, Wernerfelt N. Digital public health interventions at scale: the impact of social media advertising on beliefs and outcomes related to COVID vaccines. Proc Natl Acad Sci U S A 2023 Jan 31;120(5):e2208110120. [doi: 10.1073/pnas.2208110120] [Medline: 36701366]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
IRB: institutional review board
UX: user experience

Edited by A Stone; submitted 13.Oct.2025; peer-reviewed by O Akhadelor, O Akinpeloye, T Tanimoto; revised version received 09.Mar.2026; accepted 09.Mar.2026; published 09.Apr.2026.

Please cite as:
Bowen M, Beam M
Goal Setting and Anchoring Effects on Meditation Using a Digital Platform: Large-Scale Digital Field Study
J Med Internet Res 2026;28:e85801
URL: https://www.jmir.org/2026/1/e85801
doi:10.2196/85801

© Michael Bowen, Michael Beam. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Identifying Evidence-Based Strategies in a Digital Mental Health Intervention for Depression: Qualitative Content Analysis

Geneva K Jonathan[1,2], PhD; Jenna Y Sung[1], PhD; Heyli T Arcese[1], BS; Phoebe Holz[1], BA; Kathryn H Bentley[1], PhD; Emily E Bernstein[1], PhD; Hilary M Weingarden[1,2], PhD; Jennifer L Greenberg[1,2], PhD; Oliver T Harrison[3], MD; Sabine Wilhelm[1,2], PhD

[1]Department of Psychiatry, Massachusetts General Hospital, 185 Cambridge Street, 2nd Floor, Boston, MA, United States

[2]Department of Psychiatry, Harvard Medical School, Boston, MA, United States

[3]Koa Health, London, United Kingdom

Corresponding Author:
Geneva K Jonathan, PhD
Department of Psychiatry, Massachusetts General Hospital, 185 Cambridge Street, 2nd Floor, Boston, MA, United States

Abstract

Background: Depression is one of the leading causes of disability worldwide. Cognitive behavioral therapy (CBT) is an effective treatment, but it is difficult to access due to clinician shortages, waitlists, and logistical barriers. Smartphone-based CBT interventions offer a scalable alternative to traditional face-to-face care, but few provide transparency regarding how closely they adhere to evidence-based therapeutic principles. Understanding what therapeutic components are included in interventions helps clinicians and patients determine whether they follow CBT principles and how they might help reduce depressive symptoms.

Objective: This study aimed to characterize the therapeutic content of Mindset (Koa Health), a therapist-guided smartphone intervention for depression, by identifying the core CBT techniques it delivers and the specific behavioral strategies the app uses to put those techniques into practice.

Methods: A qualitative content analysis was conducted on all 393 unique intervention pages of Mindset. Using established CBT strategy definitions and the behavior change technique (BCT) Taxonomy version 1 (BCTTv1), coders independently evaluated each page using a collaborative consensus approach. Interrater agreement was 93.75% for CBT and 93.62% for BCT coding. Descriptive statistics (frequency, mean, and SD) and overlap between the two were calculated.

Results: All 16 core CBT techniques were identified. CBT techniques were used a total of 528 times (mean per module 66.0, SD 56.0). The most frequently used techniques included psychoeducation (164/325, 50.5% of pages), skill building (110/325, 33.8%), cognitive restructuring (46/325, 14.2%), activity scheduling (42/325, 12.9%), and self-monitoring (39/325, 12%). Across modules, 37 of 93 possible BCTs were coded 878 times (mean per module 109.8, SD 92.0) across 13 of 16 BCTTv1 categories. The most frequently applied BCT categories were shaping knowledge (205/325, 63.1% of pages), repetition and substitution (138/325, 42.5%), and feedback and monitoring (113/325, 34.8%). Overlap between the 2 frameworks was common, with the most frequent CBT-BCT pairings being psychoeducation (CBT technique)×Shaping knowledge (BCT category; appearing together on 119 pages), skill building×Shaping knowledge (80 pages), activity scheduling×Shaping knowledge (42 pages), and activity scheduling×Repetition and substitution (42 pages).

Conclusions: Mindset demonstrates coverage of CBT techniques and alignment with evidence-based BCTs. This study is the first to introduce mechanism mapping, a dual-coding approach that describes the presence of therapeutic strategies and how they are behaviorally operationalized, addressing a gap in digital mental health transparency. Unlike existing content evaluations that use presence or absence checklists, our framework captures implementation depth through systematic documentation of behavioral scaffolding. This replicable methodology enables researchers to evaluate therapeutic fidelity, supports clinicians in making evidence-informed recommendations for digital mental health treatments, and provides a foundation for the development of adaptive interventions that can enhance real-world treatment outcomes for individuals with depression.

(J Med Internet Res 2026;28:e84030) doi:10.2196/84030

KEYWORDS

digital mental health; cognitive behavioral therapy; behavior change techniques; depression; content analysis; treatment fidelity; therapist-guided; smartphone intervention; transparency; qualitative

XSL•FO
RenderX

Introduction

Depression is a severe and common illness that affects more than 300 million people worldwide, making it one of the leading causes of disability [1,2]. Its prevalence has risen in recent years [3], increasing the disease's burden on individuals and health care systems [4]. Depression impairs quality of life [5], disrupts cognitive functioning [6-8], and increases suicidal thoughts and behaviors [4,9]. Cognitive behavioral therapy (CBT) is among the most effective treatments for depression, with extensive evidence supporting its ability to reduce symptom severity [10-12] and improve functioning [13] and quality of life [14]. However, access to traditional face-to-face CBT is limited due to clinician shortages, long waitlists, geographic limitations, out-of-pocket costs, stigma, and motivational difficulties inherent to depression [15-19].

Digital interventions have emerged to address barriers to traditional face-to-face care. Multiple meta-analyses of app and web-based interventions for individuals with moderate-to-severe depression demonstrate moderate reductions in symptom severity relative to usual care or minimal control conditions, with pooled effect sizes falling in the moderate range (standardized mean difference 0.50 - 0.62) [20-22]. However, these average effects provide limited insight into what users are actually receiving in digital "CBT-based" interventions. Systematic reviews of these apps consistently show that adherence to evidence-based CBT or behavioral activation protocols is as low as 15%, and that commercially available apps do not fully meet established guideline, protocol, or treatment manual standards [23-29].

The gap between interventions labeled as "CBT-based" and their actual implementation of CBT becomes apparent when examining app content in detail. Content analyses of existing depression apps show that many include only a small number of evidence-based treatment elements (median 3 per app), with some of the most "hallmark" CBT strategies, such as cognitive restructuring, problem-solving, and relapse prevention, absent or underrepresented [25]. Even when CBT components are present, they may be implemented superficially, with limited tailoring and follow-up. For example, a review of 28 unguided depression apps and web programs found that even though behavioral activation was "present" in many programs, the strategy was rarely broken into achievable steps, anticipated barriers were not addressed, and few programs checked whether planned activities were completed, despite behavioral activation being one of the most commonly included components in CBT for depression [27]. These findings suggest that existing methods for evaluating the presence and depth of therapeutic strategies fall short of capturing the concrete implementation features that determine whether a therapeutic component functions as structured skills training (likely to achieve a greater therapeutic effect) or as a brief, one-off exercise.

Despite this limitation, existing reviews and meta-analyses have not expanded their approach to characterizing how CBT is operationalized within these tools. Most syntheses classify interventions using broad labels such as "CBT-based" or by the presence or absence of high-level components, which are often derived from intervention descriptions, stated theoretical orientations, or protocol-level summaries rather than a systematic review of the delivered in-app content and exercises themselves [21,29,30]. As a result, interventions that differ meaningfully in their therapeutic design are often treated as equivalent, obscuring variation in the active ingredients users are exposed to. Even content-focused evaluations typically summarize interventions at the program or app level, focusing on content frequency or presence and providing little information about the distribution, sequencing, or repetition of specific therapeutic techniques [24,25]. Without more granular and methodological characterization of therapeutic content, it is difficult to link intervention design to engagement, behavior change, or clinical outcomes, or to identify which components of digital CBT are most essential for effective depression treatment. However, detailed characterization of therapeutic content only addresses part of the challenge. Knowing what strategies are present does not explain how they produce change.

The mechanisms through which digital interventions exert their effects are poorly understood [22,30]. In traditional face-to-face CBT, symptom improvement is thought to arise from the coordinated enactment of cognitive, behavioral, and self-regulatory processes supported through structured practice, homework, and therapist feedback [31]. However, it is unclear whether digital CBT interventions engage these same change processes, particularly given their asynchronous format, reduced clinician involvement, and reliance on self-directed engagement. A critical but underexplored question is how digital interventions operationalize these strategies; that is, through what behavioral mechanisms do they prompt users to engage in the cognitive and behavioral work that facilitates change. Without understanding how digital interventions translate therapeutic strategies into concrete, repeated user behaviors, it is difficult to determine whether they provide structured practice opportunities that established CBT theory suggests are necessary for skill acquisition and symptom change [32,33].

To understand how digital CBT engages change mechanisms, we must recognize that CBT strategies are behavioral in nature [31]. Cognitive restructuring involves identifying and examining thoughts, activity scheduling requires planning and completing activities, and behavioral experiments entail designing and implementing behavioral tests. Critically, the same strategy can be delivered through very different user tasks, such as written thought record versus brief reflection prompts, or a full weekly activity plan versus a single daily goal. Because CBT strategies can be enacted in multiple ways, we need to move beyond describing whether they are mentioned to specifying how strategies are translated into user actions and practice opportunities.

Behavior change techniques (BCTs) provide a structured way to characterize how interventions translate therapeutic strategies into user actions that support skill enactment and change. BCTs are the smallest identifiable "active ingredients" designed to change behavior, for example, prompting self-monitoring, providing instruction, or encouraging repetition [34]. The BCT Taxonomy version 1 (BCTTv1) has been widely applied across digital behavioral domains, including medication adherence [35], physical activity [36,37], alcohol [38], smoking cessation

XSL•FO
RenderX

[39], and condom use [40]. However, the application of BCT in digital mental health is comparatively limited [41-43]. Pairing CBT strategies with BCT coding enables a systematic approach to understanding how therapeutic components are behaviorally operationalized, for example, how activity scheduling is supported through action planning, self-monitoring, or behavioral prompts. This aligns with growing calls to identify active ingredients and clarify the processes through which digital interventions achieve effects [44,45].

This paper evaluates Mindset (Koa Health), a therapist-guided smartphone CBT intervention for depression, to advance transparency and mechanistic understanding in digital mental health. Mindset was selected because it has demonstrated clinical outcomes in an open trial [46] and uses a therapist-guided delivery model associated with superior effectiveness compared to self-guided approaches [47,48]. Most importantly, we had complete access to all intervention content for comprehensive coding. In the 8-week trial of Mindset (n=28), participants showed significant reductions in clinician-rated depression severity on the Hamilton Depression Rating Scale [49] from baseline (mean 19.1, SD 5.0) to posttreatment (mean 10.8, SD 6.1), with a large effect size (Hedges g=1.47; $P<.001$; mean reduction 7.8 points, 95% CI 5.2 - 10.5), with gains maintained at 3-month follow-up [46]. Large effects were also observed for self-reported depression, functioning, and quality of life [46].

To better understand how Mindset supports change and delivers its clinical effects, we aimed to identify and describe the CBT strategies and BCTs present in Mindset and explore how BCTs operationalize CBT strategies by describing their overlap. Our overarching goal is to propose a methodological framework for describing therapeutic content that can be replicated to support digital mental health transparency.

Methods

Study Design

We conducted a qualitative content analysis of all intervention pages (ie, all app content) in Mindset, a therapist-guided smartphone CBT intervention for depression.

Ethical Considerations

The original open trial of Mindset for Depression took place at Massachusetts General Hospital in Boston, Massachusetts, between May 2022 and February 2023. It was approved by the Institutional Review Board of Massachusetts General Hospital (protocol 2020P001958) and was registered on ClinicalTrials.gov (NCT05386329). All trial participants provided informed consent before the initiation of study procedures and were given the option to withdraw at any time. Participants in the original trial were compensated US $25 at each of 3 assessment time points (midtreatment, posttreatment, and 3-month follow-up). Given that it was a content analysis of the intervention materials only, this study did not include any participant data, user-generated content, or personally identifiable information. Therefore, no additional consent was collected. No images, screenshots, or other materials in this paper or supplementary files contain any participant data or identifiable information; all content reflects only the intervention's standardized therapeutic materials. The full open trial, including a detailed description of study methods and results, is reported elsewhere [46].

Mindset Intervention Delivery

Mindset is a therapist-guided smartphone CBT intervention for depression consisting of 8 modules delivered across 8 weeks [46]. The intervention is grounded in CBT theory, which posits that improvements in depression are facilitated by changes in thoughts, behaviors, and emotional responses [50].

Participants had immediate access to the intervention after completing baseline assessments. They were instructed to use the app independently while attending weekly 15 - to 20-minute therapist sessions conducted via videoconferencing (Health Insurance Portability and Accountability Act [HIPAA]-compliant) to review progress and receive support. Therapists were licensed clinical psychologists trained in CBT for depression, and, because of licensure requirements, participants were required to reside in Massachusetts.

The program began with a brief onboarding sequence that introduced CBT principles and app navigation (Table 1). Modules followed a typical CBT progression, beginning with psychoeducation and self-monitoring, moving into behavioral activation, and ending with cognitive restructuring and relapse prevention [50]. Module 1 could be completed at the participant's own pace, whereas Modules 2 - 8 required a minimum of 7 days per module to support practice and skill consolidation. During Module 2, participants were prompted to self-monitor by logging at least 3 activities they engaged in and the corresponding moods each day. From Module 3 onward, participants implemented behavioral activation by scheduling at least 3 activities per week aligned with 7 value domains (health and wellness, causes and community, spirituality, creative pursuits, career and education, personal relationships, and day-to-day tasks), then logging mood after each activity. The activity library included 102 options, and activity scheduling continued through Modules 4 - 8. Although therapeutic content was standardized, participants' activity choices and the extent to which they revisited the activity scheduling and logging pages allowed for optional individual tailoring or personalization.

XSL•FO
RenderX

Table . Mindset intervention overview, module-by-module.

Module	Description	Number of pages		
		Unique	Repeated	Total
Onboarding	Introduction to intervention and how to use	10	0	10
Module 1: Exploring How We Think	Cognitive restructuring to challenge unhelpful thinking	135	0	135
Module 2: Activities and Mood	Self-monitoring of activities and associated moods	28	0	28
Module 3: Goals and Scheduling	Identify personal values, set goals, and schedule activities aligned with values	24	7	31
Module 4: Establishing a Rhythm	Mindful breathing, continued scheduling of activities	19	7	26
Module 5: Keeping Momentum	Mindful grounding, continued scheduling of activities	9	7	16
Module 6: Planning Around Priorities	Skills to let go of unhelpful thoughts, continued scheduling of activities	9	7	16
Module 7: Examining Core Beliefs	Modification of core beliefs and building self-esteem, continued scheduling of activities	26	7	33
Module 8: Preparing for the Future	Relapse prevention, reflect on progress and learned skills, continued scheduling of activities	23	7	30
Total		283	__[a]	325

[a]Not applicable.

All data and therapeutic materials were sourced directly from the Mindset app, which included 393 unique app pages across onboarding content and modules (Table 1), totaling 943 pages when repeated content was included. No materials or pages from the app were excluded from analysis. Content was accessible only within the smartphone app and was not available as external files or in print. On average, there were 39.3 pages per module (SD 44.8). Content was delivered in the app via text, though some pages, such as the introductory module, mindfulness and grounding exercises, included the option to watch the content via video.

Coding Framework

We used a dual-coding approach to systematically identify what therapeutic strategies were delivered and how. This approach, which we call mechanism mapping, pairs established CBT technique definitions [31,50] with the BCTTv1 [34] to describe the therapeutic content included and its behavioral operationalization.

CBT Techniques

We first identified and labeled CBT techniques present in Mindset to determine which therapeutic strategies were included and to evaluate alignment with established CBT protocols for depression (Table S1 in Multimedia Appendix 1). The CBT coding framework consisted of 16 techniques found in manualized CBT protocols for depression [25,31,50]. Refer to Table S1 in Multimedia Appendix 1 for coding definitions.

BCTs

We applied the BCTTv1, a validated framework comprising 93 techniques organized into 16 hierarchically clustered categories [34,51,52]. The first author (GKJ) completed training through the official online platform [51]. The 16 categories and their definitions are fully described by Michie et al [34] and in Table S2 in Multimedia Appendix 1.

We chose BCTTv1 because of its ability to characterize the "ingredients" that operationalize change in a digital CBT intervention, consistent with a mechanisms-focused approach [34]. It also provides a validated, fine-grained taxonomy for identifying discrete, observable techniques embedded in app-based content, that is, what users see and do on each screen, making it particularly appropriate for therapy apps that translate evidence-based strategies into modular, self-guided activities. Other frameworks, such as the Theoretical Domains Framework and the Capability, Opportunity, Motivation–Behavior (COM-B) theory, describe why behavior changes by organizing determinants, but do not enumerate concrete techniques [53,54]. Although BCTTv1 was originally developed for health behavior interventions, its structure aligns well with CBT's emphasis on observable behavioral and cognitive skills [50].

XSL•FO
RenderX

Coding Procedures

The raw text of the Mindset app was organized page-by-page to preserve the intervention's structure during coding. A collaborative coding approach was used to identify CBT techniques and BCTs [55]. Techniques were coded based on the presence of explicit content, exercises, or instructions; implied techniques were not coded.

Coder Roles, Independence, and Workflow

Coding was conducted by 2 doctoral-level clinical psychologists (GKJ and JYS) and 1 bachelor's-level research assistant (HTA), all familiar with CBT for depression. None were involved in the initial development or testing of the Mindset intervention. The first author (GKJ) independently coded each intervention page in Dedoose (SocioCultural Research Consultants, LLC) [56]. Two additional coders (JYS and HTA) reviewed the codes and documented disagreements in memos. Discrepancies were discussed collaboratively using the memos as well as CBT and BCTTv1 definitions until consensus was reached.

Interrater Agreement

To assess coder agreement, we report percentage agreement as the reliability metric [57]. Percentage agreement is appropriate for qualitative content analysis with multiple potential codes (eg, different pages of an intervention included multiple techniques or skills), because kappa statistics can be overly conservative when the number of possible categories is large relative to the sample size [57,58]. Before consensus, agreement was 93.62% for BCT codes (822/878) and 93.75% for CBT codes (494/528).

Analysis

We calculated the frequency and distribution of CBT techniques and BCT categories across Mindset, including mean applications, SD, range, and percentage of total app pages.

For ease of interpreting results, we summarize BCT categories rather than each technique, and we prioritize CBT techniques and BCT categories that contribute ≥10% of total codes for reporting.

To prevent repeated content from inflating counts, we weighted pages that participants were guided to revisit multiple times. For example, activity scheduling pages appeared in Modules 3 - 8, with instructions to schedule at least 3 activities per week. We applied a weighting factor equal to the average number of activities scheduled per participant divided by the total number of possible activity pages (n=110). This weighting assumption was based on participant instructions provided in the app. After applying this weighting factor, the 110 activity library pages in each of Modules 3-8 were reduced to 7 weighted pages per module, yielding an analytic dataset of 325 pages (Table 1) used as the denominator for all frequency and percentage calculations.

Data aggregation and analysis were conducted using R version 4.4.2 (R Core Team) [59]; a heatmap was developed using the *pheatmap* package [60], and figures were finalized in Canva (Canva Pty Ltd) [61]. Multimedia Appendix 1 includes detailed documentation of coding distributions.

Reporting Standards

This study follows the TIDieR (Template for Intervention Description and Replication) reporting guideline to promote transparency and replicability in describing intervention content and delivery [62]. A completed TIDieR checklist is provided in Checklist 1.

Results

CBT Techniques

All 16 CBT techniques were represented and applied 528 times (mean 66.0 per module, SD 56.0) throughout the intervention. Figure 1 shows the distribution of CBT techniques by module, and Table 2 reports the descriptive summary of identified strategies. Full frequencies and distributions are reported in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. Distribution of cognitive behavioral therapy techniques across intervention modules in the therapist-guided Mindset smartphone program for depression. This figure summarizes when each of the 16 cognitive behavioral therapy techniques was delivered throughout the 8 modules. Horizontal bars represent the percentage of app pages (N=325) within each module that included the cognitive behavioral therapy strategy. CBT: cognitive behavioral therapy.

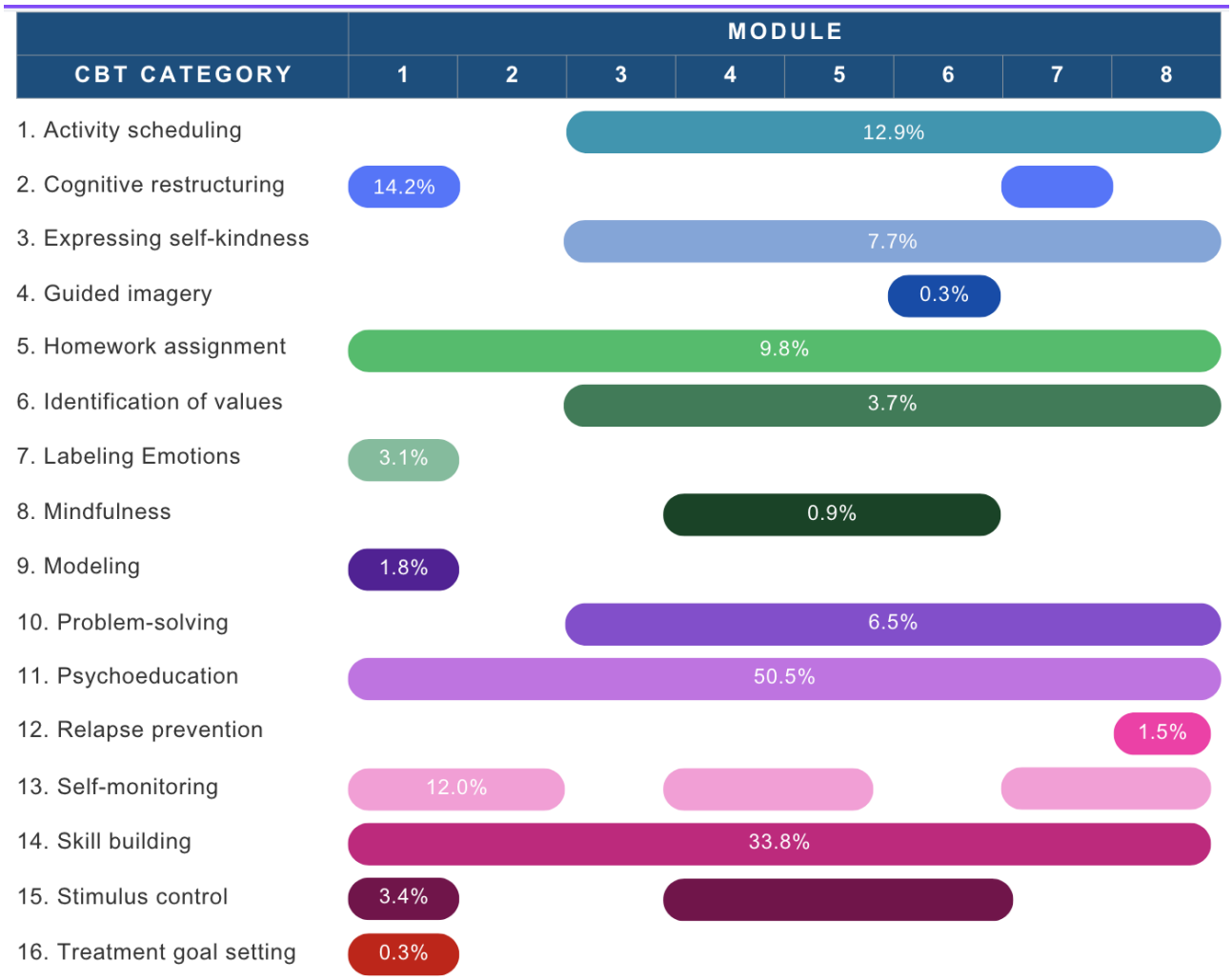

XSL•FO
RenderX

Table . Cognitive behavioral therapy techniques in the Mindset smartphone intervention for depression.[a]

Rank	CBT[b] category	Top module	Pages, n (%)	Mean per module when present (SD)	Examples of implementation
1	Psychoeducation	1	164 (50.5)	20.5 (23.1)	Explanation of the difference between sadness and depression
2	Skill building	1	110 (33.8)	13.8 (12.9)	Prompting participants to log 3 activities a day
3	Cognitive restructuring	1	46 (14.2)	23.0 (21.2)	Identifying a negative thought or belief and typing in a more balanced thought
4	Activity scheduling	3-8	42 (12.9)	7.0 (0.0)	Scheduling 3 activities per day and rating the mood for each activity
5	Self-monitoring	1	39 (12.0)	5.6 (13.8)	Logging 3 activities per day and rating mood for each activity
6	Homework assignment	1	32 (9.8)	4.0 (0.5)	Direction to at least 3 log activities in order to move to next module
7	Expressing self-kindness	1	25 (7.7)	4.2 (2.4)	“Write an encouraging note to self”
8	Problem-solving	1	21 (6.5)	2.6 (0.8)	Ask participants to identify and write down barriers to planned activities in the app
9	Identification of values	1	12 (3.7)	1.5 (1.9)	Selection of activities aligned with personal values
10	Stimulus control	1	11 (3.4)	2.8 (3.5)	“Turn off notifications while practicing mindfulness”
11	Labeling emotions	1	10 (3.1)	10 (0.0)	Encouragement to reflect on and name what emotion they feel after a challenging situation
12	Modeling	1	6 (1.8)	6.0 (0.0)	Demonstration of how to set a SMART[c] goal
13	Relapse prevention	8	5 (1.5)	5.0 (0.0)	Preparing a plan for anticipated stressful events
14	Mindfulness	4-6	3 (0.9)	1.0 (0.0)	Guided breathing exercise with attention to present-moment sensations
15	Guided imagery	6	1 (0.3)	1.0 (0.0)	“Visualize placing each thought on a cloud and watching it drift away”
16	Treatment goal setting	1	1(0.3)	0.1 (0.4)	Set goal in alignment with values

[a]Cognitive behavioral therapy strategies were coded 528 times throughout all 8 intervention modules. As such, percentages are reported as a percentage of the total app pages (N=325).

[b]CBT: cognitive behavioral therapy.

[c]SMART: specific, measurable, achievable, relevant, and time-bound.

XSL•FO
RenderX

High-Frequency Techniques (≥10%)

Psychoeducation was present in all modules and accounted for more than half of the intervention pages (164/325, 50.5%), followed by skill building (110/325, 33.8%) and cognitive restructuring (46/325, 14.2%), both of which were most present in Module 1. Activity scheduling (42/325, 12.9%) appeared consistently across Modules 3 - 8, in alignment with repeated practice of behavioral activation skills. Self-monitoring (39/325, 12.0%) was present intermittently throughout the intervention, but most focused on modules that asked users to log and observe mood changes linked to activities.

Low-Frequency Techniques (<10%)

Homework assignments, self-kindness, problem-solving, and value identification appeared on 3% - 10% of the pages. Labeling emotions, mindfulness, guided imagery, and relapse prevention were less frequent (<5%).

BCTs

The intervention included 37 BCTs (39.8% of the taxonomy), from 13 of 16 categories, applied 878 times (mean 109.8, SD 92.0) across modules. Figure 2 shows the distribution of BCTs by module, and Table 3 provides a descriptive summary of the identified techniques.

Figure 2. Distribution of behavior change technique categories across the therapist-guided Mindset smartphone cognitive behavioral therapy intervention for depression. Each of the 13 behavior change technique categories identified in Mindset was delivered throughout the 8 modules. Horizontal bars represent the percentage of app pages (N=325) within each module that included the cognitive behavioral therapy strategy. BCT: behavior change technique.

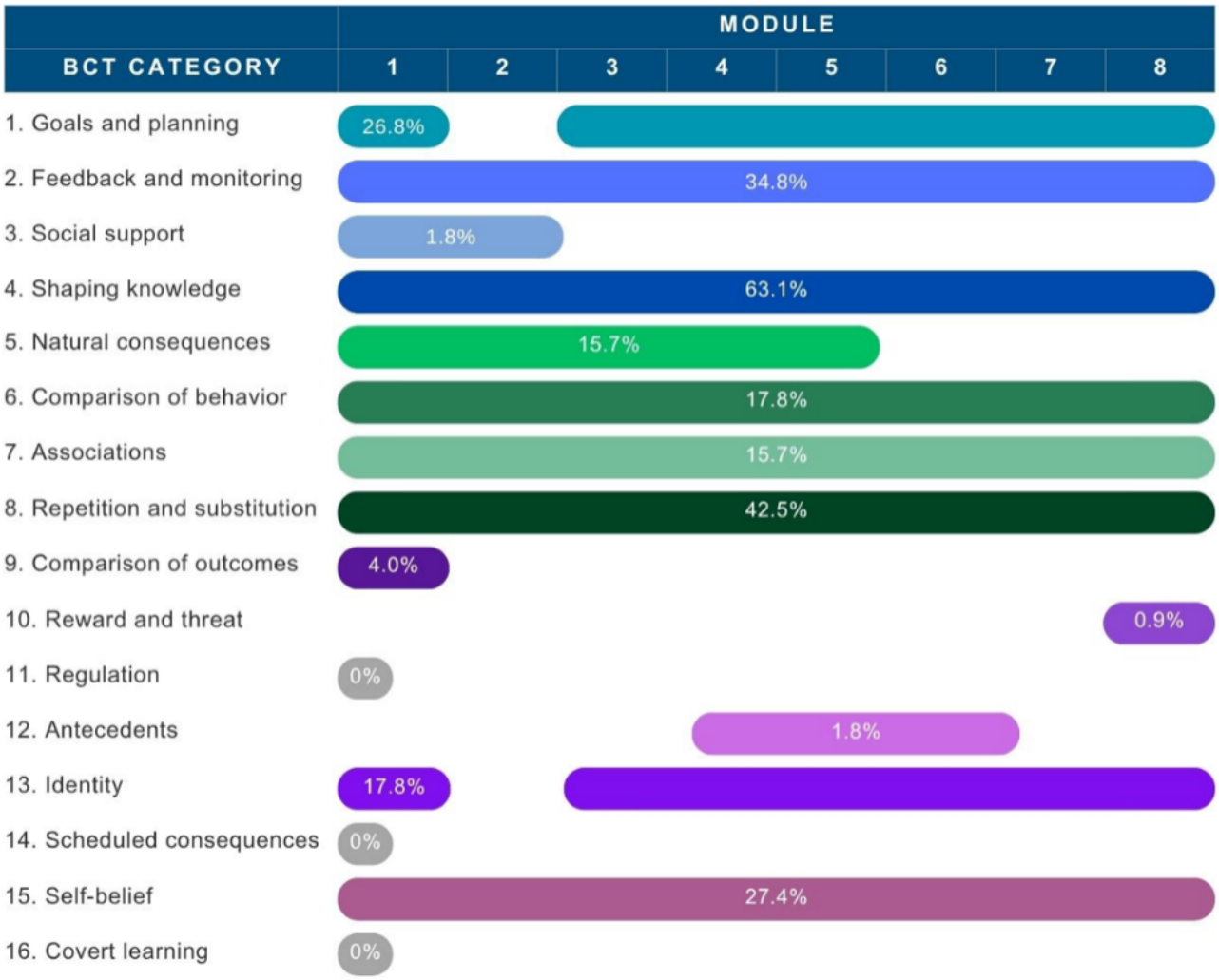

XSL•FO
RenderX

Table . Behavior change techniques in the Mindset smartphone intervention for depression.[a]

Rank	BCT[b] category	Top module	Pages, n (%)	Mean per module when present (SD)	Most common BCTs used	Intervention, n (%)	Examples of implementation
1	Shaping knowledge	1	205 (63.1)	25.6 (22.5)			
					Instruction on how to perform a behavior	140 (43.1)	Description of how to plan activities (eg, check schedule and anticipate barriers)
					Information about antecedents	65 (20.0)	Changes in season can lead to low mood
2	Repetition and substitution	1	138 (42.5)	17.3 (13.9)			
					Behavioral practice or rehearsal	110 (33.8)	Logging 3 activities per day in app
					Habit formation	16 (4.9)	Repeated practice of activity scheduling or planning
3	Feedback and monitoring	1	113 (34.8)	14.1 (12.4)			
					Self-monitoring of behavior or thought	55 (16.9)	Rate mood after activity
					Feedback on behavior	45 (13.8)	Feedback on cognitive restructuring quiz
4	Self-belief	1	89 (27.4)	11.1 (11.0)			
					Self-talk	41 (12.6)	Textbox to write an encouraging note to self
					Verbal persuasion about capability	38 (11.7)	You're well-equipped to handle future challenges
5	Goals and planning	8	87 (26.8)	10.9 (6.1)			
					Action planning	21 (6.5)	Plan for anticipated stressful events
					Goal setting (behavior)	21 (6.5)	Set intentions for use of Mindset
6	Identity	1	58 (17.8)	7.3 (12.2)			
					Framing or reframing	45 (13.8)	Identify negative beliefs, reframe with a balanced perspective
					Valued self-identity	10 (3.1)	Choose positive qualities about self from the list
7	Comparison of behavior	1	58 (17.8)	7.3 (5.4)			

XSL•FO
RenderX

Rank	BCT[b] category	Top module	Pages, n (%)	Mean per module when present (SD)	Most common BCTs used	Intervention, n (%)	Examples of implementation
					Demonstration of behavior	55 (16.9)	Example of negative thinking pattern
					Social comparison	3 (0.9)	Describe another person and select their characteristics from a list
8	Natural consequences	1	51 (15.7)	6.4 (10.8)			
					Information about emotional consequences	35 (10.8)	Negative thoughts may make you feel discouraged
					Information about social or environmental consequences	11 (3.4)	Low mood may lead to social interaction avoidance
9	Associations	1	51 (15.7)	6.4 (5.7)			
					Prompts or cues	51 (15.7)	Encouragement to return to Mindset to refresh skills
10	Comparison of outcomes	1	13 (4.0)	1.6 (4.6)			
					Credible Source	13 (4.0)	Embedded links to evidence-based resources
11	Antecedents	4-6	6 (1.8)	0.8 (1.0)			
					Restructuring the physical environment	3 (0.9)	Enabling "do not disturb" during mindfulness practice
					Body changes	3 (0.9)	Bring attention to breath during mindfulness practice
12	Social support	1	6 (1.8)	0.8 (1.6)			
					Social support (practical)	4 (1.2)	Send messages to ask therapist how to use Mindset
					Social support (emotional)	1 (0.3)	Encouragement to reach out to 911 or 988 in an emergency
13	Reward and threat	8	3 (0.9)	0.4 (1.1)			
					Self-reward	3 (0.9)	Acknowledge how far you've come and skills gained

[a]Behavior change techniques were coded 878 times throughout all 8 intervention modules; as such, percentages are reported as a percentage of the total app pages.

[b]BCT: behavior change technique.

XSL•FO
RenderX

High-Frequency Techniques (≥10%)

Six BCT categories were found in all 8 modules: shaping knowledge (205/325, 63.1%), repetition and substitution (138/325, 42.5%), feedback and monitoring (113/325, 34.8%), self-belief (89/325, 27.4%), comparison of behavior (58/325, 17.8%), and associations (51/878, 15.7%). Goals and planning (87/325 26.8%) and identity (58/325, 17.8%) were present in all modules, except Module 2, where participants began learning to monitor thoughts and activities. Natural consequences (51/325, 15.7%) were present in more than half of the modules.

Low-Frequency Techniques (<10%)

Comparison of outcomes, social support, antecedents, and reward and threat occurred on fewer than 5% of pages. No techniques were coded from the regulation, scheduled-consequences, or covert-learning categories.

CBT and BCT Co-Occurrence

We examined how CBT techniques were behaviorally implemented through the codelivery of BCT mechanisms at the page level. Each co-occurrence corresponds to an intervention page where a CBT technique and a BCT are presented simultaneously, aiming to provide insight into how skills are taught, rehearsed, and reinforced during intervention use (Figure 3).

Psychoeducation showed the largest overlap, co-occurring with all 13 observed BCT categories, most frequently with shaping knowledge (119/325, 36.6% of intervention pages had this overlap). Skill building was also widely scaffolded, most often through repetition and substitution (114/325, 35.1%), feedback and monitoring (83/325, 25.5%), and shaping knowledge (80/325, 24.6%). Activity scheduling was consistently behaviorally reinforced, most prominently through goals and planning (72/325, 22.2%), repetition and substitution (42/325, 12.9%), and shaping knowledge (42/325, 12.9%).

Several core CBT techniques were behaviorally supported at moderate densities. For example, cognitive restructuring was frequently paired with repetition and substitution (43/325, 13.2%) and feedback and monitoring (35/325, 10.8%). Self-monitoring and problem solving were supported by fewer BCT categories (generally 2% - 8% of pages per pairing).

By contrast, mindfulness, guided imagery, and relapse prevention were rarely paired with BCTs, with most co-occurrences appearing on ≤10 pages (≤3.1%). Motivational, consequence-based, and social mechanisms, including reward and threat, social support, comparison of outcomes, and antecedents, were similarly infrequent (≤20 pages, ≤6.2% each).

XSL•FO
RenderX

Figure 3. Mapping of cognitive behavioral therapy techniques to behavior change technique categories in the Mindset digital cognitive behavioral therapy intervention for depression. This heatmap displays how often each cognitive behavioral therapy technique (*x*-axis) was coded with each behavior change technique category (*y*-axis) based on page-level content coding of the Mindset smartphone program for adults with major depressive disorder. Cell shading reflects co-occurrence frequency (0 - 150; darker cells indicate higher frequency).

Discussion

Principal Findings

This study aimed to advance transparency and mechanistic understanding in digital mental health by conducting a comprehensive analysis of Mindset, a therapist-guided smartphone CBT intervention for depression. Three primary findings emerged. First, all 16 CBT techniques were present across the intervention's 8 modules, with a total of 528 instances. Second, Mindset incorporated 37 unique BCTs from 13 of 16 taxonomy categories, coded 878 times throughout the intervention. Shaping knowledge, repetition and substitution, and feedback and monitoring were most prevalent. Third, co-occurrence analysis highlighted the behavioral mechanisms through which CBT strategies were operationalized. High-frequency techniques such as psychoeducation, skill building, and activity scheduling were consistently paired with multiple BCTs that structure practice (repetition and substitution), provide feedback (feedback and monitoring), and support learning (shaping knowledge). In contrast, lower-frequency techniques such as mindfulness and guided imagery showed minimal BCT scaffolding, suggesting they were introduced but not behaviorally reinforced through repeated practice. These findings show how Mindset accomplishes fidelity to CBT protocols through multilayered behavioral mechanisms while establishing a replicable methodology for describing therapeutic content in a digital intervention.

XSL•FO
RenderX

Mechanism Mapping: Showing Implementation Depth Through Behavioral Scaffolding

This study introduces a mechanism mapping approach that directly addresses limitations of prior evaluations of digital CBT by methodologically examining how therapeutic techniques are implemented in user-facing content [25,43,44,63,64]. Specifically, our co-occurrence analysis (Figure 3) builds on previous studies by showing which BCTs scaffold each CBT strategy, making implementation depth visible and quantifiable.

Applying this mechanism mapping approach to Mindset highlighted considerable variation in implementation depth across CBT strategies. High-frequency techniques showed dense scaffolding: psychoeducation drew upon all 13 BCT categories, engaging users through informational content (shaping knowledge), emotional reflection exercises (natural consequences), and repeated practice activities (repetition and substitution). In contrast, mindfulness and guided imagery appeared with minimal BCT support, suggesting in-app introduction without behavioral reinforcement. The content distribution also showed app design choices: Module 1 was the largest (135 pages vs 9 - 28 in other modules) and most densely scaffolded, front-loading introductory skills with instruction, practice, and feedback, laying the groundwork for subsequent modules to build upon.

This variation in implementation would be invisible using CBT-only classification but may help explain variability among apps claiming to deliver “CBT-based” care. Evidence from component-focused meta-analyses reinforces why implementation depth matters: combined cognitive and behavioral processes mediate digital intervention outcomes, but techniques delivered on their own showed no effects [45]. Additionally, skill enactment, actively practicing therapeutic strategies, predicts clinical improvement in most trials, whereas knowledge acquisition alone does not [32]. In Mindset, the BCTs supporting behavioral activation and cognitive change, such as instruction, repeated practice, and ongoing monitoring, appeared consistently across modules, creating structured opportunities for skill rehearsal. The intervention was designed this way so that users actively had to practice and rehearse strategies, not just learn about them. Most importantly, mechanism mapping generates testable hypotheses about content features that may facilitate clinical effects. To our knowledge, no other digital mental health study has visualized or described these relationships, providing a novel analytical framework that is replicable across interventions.

Positioning Mindset Within the Digital Mental Health Landscape

Mindset’s 16 CBT strategies and 37 BCTs exceed typical implementation in digital mental health tools. Prior assessments show that most digital tools include only a median of 3 - 4 CBT-based elements [25,29] and 9 - 15 BCTs, with conversational agents averaging 15 BCTs (range 4 - 30) [63], eating disorder interventions averaging 14 BCTs (range 9 - 18) [43], and general mental health apps containing 1 - 10 techniques [65]. Our higher technique count compared to prior reviews likely reflects methodological differences in content analysis. We examined every page of the actual intervention, where previous studies coded app store descriptions or published summaries [41,44,63,64], which risks overlooking embedded therapeutic elements and overrelying on developer claims. However, technique quantity alone is not a reliable predictor of treatment effects [65]. Interventions grounded in established psychological theory show better outcomes than those lacking theoretical grounding [43], and full CBT packages outperform single-component strategies. Accordingly, Mindset’s extensive technique involvement may matter less than how these techniques are combined and delivered.

The most frequent CBT techniques in Mindset align with what systematic reviews consistently identify in digital interventions: psychoeducation (present in 50% - 75% of apps), behavioral activation including activity planning and scheduling (31% - 68%), and cognitive restructuring (31% - 79%) [25,27,29]. However, many techniques we identified in Mindset are inconsistently reported in large-scale content syntheses. Skill building, for example, appears in 0% - 12% of interventions depending on how broadly it is defined, with some reviews only capturing specific skill types such as social skills training [25,27]. Similarly, techniques such as homework assignment, self-compassion or expressing self-kindness, problem-solving, stimulus control, and emotion labeling were present in Mindset but are rarely systematically documented in existing app reviews. This inconsistency may reflect the absence of standardized CBT technique taxonomies in digital mental health research, making cross-study comparisons difficult and potentially underestimating the therapeutic content actually delivered to users.

The BCTs most strongly associated with engagement in mobile health interventions, feedback on behavior, self-monitoring, and instructions on how to perform a behavior were among Mindset’s most frequently used [41]. Importantly, these behavioral strategies were not features exclusively designed to keep users engaged; they were directly tied to CBT skill enactment (eg, mood monitoring implemented in service of behavioral activation). This integration contrasts sharply with common commercial designs in which self-monitoring or reminders are detached from core therapeutic mechanisms [25,41]. Mindset’s approach of embedding engagement-related BCTs within therapeutic workflows may explain how the intervention maintains both user engagement and therapeutic fidelity.

Our comprehensive content analysis also sheds light on which CBT strategies received minimal behavioral support and which BCT categories were absent entirely. Techniques such as grounding and mindfulness appeared infrequently and were associated with few BCTs. However, this may reflect the implementation format rather than inadequate support: these strategies were delivered through video demonstrations, which our text-based coding may have undercounted compared to written exercises that explicitly name behavioral techniques. Additionally, entire BCT categories were absent from Mindset, including regulation (eg, reduce negative emotions and conserve mental resources) and scheduled consequences (eg, punishment and reward). However, transparency about what is missing must be interpreted cautiously. Identifying absent strategies or mechanisms does not indicate whether their inclusion would

XSL•FO
RenderX

improve outcomes. The value of this analysis lies in enabling transparent evaluation of what tools actually deliver, allowing clinicians and researchers to generate testable hypotheses about which gaps, if any, meaningfully impact outcomes.

Limitations

This analysis has several limitations. First, the study did not statistically link participants' exposure to or adherence with these techniques to clinical outcomes, preventing examination of which techniques may have been associated with these outcomes. Second, the review focused on text-based content, potentially underestimating exposure to interactive audio or video elements that may be more engaging than their script counterparts. Third, despite formal BCT training, content coding may have introduced subjective biases. Fourth, Mindset includes therapist-delivered content that was intentionally excluded to isolate digital components, meaning this analysis may underrepresent the full therapeutic exposure users received. Finally, our analysis concentrated on evidence-based CBT techniques and BCTs, without assessing non–evidence-based or potentially contraindicated content. Some mental health apps include elements that may conflict with recommended approaches, which may also impact safety and effectiveness. Accordingly, mechanism mapping should therefore be seen as an initial foundation for objective fidelity monitoring instead of a complete measure of treatment quality.

Broader Implications

This study demonstrates that transparent evaluation of digital mental health interventions is feasible and necessary. We offer the following actionable recommendations for advancing quality and accountability in the field. For intervention developers, we recommend auditing existing tools using this dual-coding framework to identify which therapeutic techniques lack behavioral scaffolding so that we can confirm techniques are appropriately reinforced through user action. For researchers, we recommend contributing to transparency by reporting the frequency of therapeutic techniques, making this standard practice to enable meta-analyses to examine implementation depth as a moderator of outcomes. Our methodology is fully replicable for any text-based intervention. We encourage clinicians to evaluate apps based on documented content rather than marketing claims, to seek evidence of technique coverage and reinforcement of mechanisms, and transparent reporting of delivered content. This can be used to inform patient-centered recommendations of these tools. Finally, for the field, we hope that this report sets a new precedent for minimum reporting standards when evaluating these tools, so that they can be more comprehensively evaluated and compared.

Conclusions

The rapid global adoption of digital mental health tools has outpaced our ability to evaluate what they actually deliver. This study introduces mechanism mapping as the first systematic approach to describe which therapeutic strategies are present and how they are behaviorally operationalized through user actions. Unlike prior content analyses that rely on presence or absence coding, our dual-framework methodology reveals implementation depth, distinguishing between techniques that are mentioned versus those supported through repeated practice, feedback, and skill consolidation. This innovation addresses a fundamental transparency gap: most digital interventions function as "black boxes," making it impossible for clinicians, patients, or researchers to assess therapeutic fidelity.

This replicable framework brings 3 critical contributions to the field: it enables comparative evaluation of digital interventions based on therapeutic content rather than marketing claims, provides testable hypotheses about which implementation features drive clinical effects, and establishes a methodological standard for transparent reporting. In the real world, this work supports clinicians in making evidence-based referrals, empowers patients to choose interventions aligned with their therapeutic needs, and guides developers in creating tools with demonstrable fidelity to evidence-based practices.

Future work should extend this approach by integrating behavioral measurement and mechanistic outcomes, applying mechanism mapping across multiple digital interventions to identify common implementation signatures, exploring whether implementation patterns could inform adaptive tailoring of behavioral supports, and incorporating therapist-delivered components in hybrid models to capture the full therapeutic ecosystem. Importantly, transparency must become the standard in digital mental health, not the exception. Understanding how apps work, for whom, and through what processes is critical for the next generation of precision mental health care.

Acknowledgments

Artificial intelligence (AI) was used to generate a table-of-contents image for the manuscript.

Funding

This primary outcome study on which this paper is based was supported by Koa Health (Wilhelm, principal investigator), and a NIDA Career Development Program in Substance Use and Addiction Medicine (principal investigator AEE, K12 DA043490) as well as an anonymous donor fund (Wilhelm, principal investigator).

Data Availability

The code-frequency data generated and analyzed during this study are included in this published article (Multimedia Appendix 1). The full intervention pages, underlying code assignments, and codebooks are not publicly available as the digital intervention is commercially licensed. However, summary files of the behavioral categories and cognitive behavioral therapy frequency counts

XSL•FO
RenderX

have been provided to enable transparency and independent verification of findings. Additional data may be made available from the corresponding author on reasonable request, subject to commercial licensing allowances.

Authors' Contributions

Conceptualization: GJ
Formal analysis: GJ, JS, HA
Methodology: GJ
Investigation: KB, EB, JG, SW
Project administration: EB, HW
Resources: OH, SW
Supervision: KB, EB, SW
Funding acquisition: SW
Writing – original draft: GJ, JS, HA
Writing – review & editing: PH, KB, EB, HW, JG, OH, SW

Conflicts of Interest

GKJ has received research support from the National Institute on Drug Abuse. JYS has received research support from the Robert Wood Johnson Foundation as well as the International OCD Foundation. EEB has received research support from Koa Health Digital Solutions LLC. HMW receives salary from HabitAware, Inc, is a paid consultant to APA Labs, and was formerly a paid consultant to Hello Therapeutics, Inc., is a paid scientific advisory board member to Augmend Health, and has received research support from Koa Health, Inc., National Institute of Mental Health, and President and Fellows of Harvard College. She is a member of the APA's Mental Health Technology Advisory Committee. SW is a presenter for the Massachusetts General Hospital Psychiatry Academy in educational programs supported through independent medical education grants. She has received royalties from Guilford Publications, New Harbinger Publications, Springer, and Oxford University Press. SW has also received speaking honoraria from various academic institutions and foundations, including the International Obsessive Compulsive Disorder Foundation, the Tourette Association of America, and the Centers for Disease Control and Prevention. In addition, she received honoraria for her role on the Scientific Advisory Board for One-Mind (PsyberGuide), Koa Health Digital Solutions LLC, and Noom, Inc. SW has received research support from Koa Health Digital Solutions LLC. HMW receives salary from HabitAware, Inc, has been a paid consultant to APA Labs and Hello Therapeutics, Inc, is a paid scientific advisory board member for Augmend Health, and has received research support from National Institute of Mental Health, Koa Health, LLC, and President and Fellows of Harvard College. KHB is a paid consultant to Visible Health Inc, and has received research support from the National Institute of Mental Health. OTH is Founder and CEO of Koa Health Limited, and receives reimbursement as the Royal Society Entrepreneur in Residence in healthcare artificial intelligence at Oxford University. All other authors have no disclosures to report.

Multimedia Appendix 1
Cognitive behavioral therapy and behavior change technique coding descriptive data.
[XLSX File, 22 KB - jmir_v28i1e84030_app1.xlsx]

Checklist 1
TIDieR framework checklist.
[DOCX File, 22 KB - jmir_v28i1e84030_app2.docx]

References

1. Depressive disorder (depression). World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/depression [accessed 2024-11-22]
2. Friedrich MJ. Depression is the leading cause of disability around the world. JAMA 2017 Apr 18;317(15):1517. [doi: 10.1001/jama.2017.3826]
3. Moreno-Agostino D, Wu YT, Daskalopoulou C, Hasan MT, Huisman M, Prina M. Global trends in the prevalence and incidence of depression:a systematic review and meta-analysis. J Affect Disord 2021 Feb 15;281:235-243. [doi: 10.1016/j.jad.2020.12.035] [Medline: 33338841]
4. Lépine JP, Briley M. The increasing burden of depression. Neuropsychiatr Dis Treat 2011;7(Suppl 1):3-7. [doi: 10.2147/NDT.S19617] [Medline: 21750622]
5. Hansson L. Quality of life in depression and anxiety. Int Rev Psychiatry 2002 Jan;14(3):185-189. [doi: 10.1080/09540260220144966]
6. Snyder HR. Major depressive disorder is associated with broad impairments on neuropsychological measures of executive function: a meta-analysis and review. Psychol Bull 2013 Jan;139(1):81-132. [doi: 10.1037/a0028727] [Medline: 22642228]

XSL•FO
RenderX

7. Porter RJ, Bourke C, Gallagher P. Neuropsychological impairment in major depression: its nature, origin and clinical significance. Aust N Z J Psychiatry 2007 Feb;41(2):115-128. [doi: 10.1080/00048670601109881] [Medline: 17464689]
8. Dillon DG, Pizzagalli DA. Mechanisms of memory disruption in depression. Trends Neurosci 2018 Mar;41(3):137-149. [doi: 10.1016/j.tins.2017.12.006] [Medline: 29331265]
9. Hawton K, Casañas I Comabella C, Haw C, Saunders K. Risk factors for suicide in individuals with depression: a systematic review. J Affect Disord 2013 May;147(1-3):17-28. [doi: 10.1016/j.jad.2013.01.004] [Medline: 23411024]
10. Lepping P, Whittington R, Sambhi RS, et al. Clinical relevance of findings in trials of CBT for depression. Eur Psychiatry 2017 Sep;45:207-211. [doi: 10.1016/j.eurpsy.2017.07.003] [Medline: 28957788]
11. Twomey C, O'Reilly G, Byrne M. Effectiveness of cognitive behavioural therapy for anxiety and depression in primary care: a meta-analysis. Fam Pract 2015 Feb;32(1):3-15. [doi: 10.1093/fampra/cmu060] [Medline: 25248976]
12. Oud M, de Winter L, Vermeulen-Smit E, et al. Effectiveness of CBT for children and adolescents with depression: a systematic review and meta-regression analysis. Eur Psychiatry 2019 Apr;57:33-45. [doi: 10.1016/j.eurpsy.2018.12.008] [Medline: 30658278]
13. Renner F, Cuijpers P, Huibers MJH. The effect of psychotherapy for depression on improvements in social functioning: a meta-analysis. Psychol Med 2014 Oct;44(14):2913-2926. [doi: 10.1017/S0033291713003152] [Medline: 24472135]
14. Hofmann SG, Curtiss J, Carpenter JK, Kind S. Effect of treatments for depression on quality of life: a meta-analysis. Cogn Behav Ther 2017 Jun;46(4):265-286. [doi: 10.1080/16506073.2017.1304445] [Medline: 28440699]
15. Arnaez JM, Krendl AC, McCormick BP, Chen Z, Chomistek AK. The association of depression stigma with barriers to seeking mental health care: a cross-sectional analysis. J Ment Health 2020 Apr;29(2):182-190. [doi: 10.1080/09638237.2019.1644494] [Medline: 31373519]
16. Kohn R, Saxena S, Levav I, Saraceno B. The treatment gap in mental health care. Bull World Health Organ 2004 Nov;82(11):858-866. [Medline: 15640922]
17. Chekroud AM, Foster D, Zheutlin AB, et al. Predicting barriers to treatment for depression in a U.S. national sample: a cross-sectional, proof-of-concept study. Psychiatr Serv 2018 Aug 1;69(8):927-934. [doi: 10.1176/appi.ps.201800094] [Medline: 29962307]
18. Andrade LH, Alonso J, Mneimneh Z, et al. Barriers to mental health treatment: results from the WHO World Mental Health surveys. Psychol Med 2014 Apr;44(6):1303-1317. [doi: 10.1017/S0033291713001943] [Medline: 23931656]
19. Cruz M, Pincus HA, Harman JS, Reynolds CF 3rd, Post EP. Barriers to care-seeking for depressed African Americans. Int J Psychiatry Med 2008;38(1):71-80. [doi: 10.2190/PM.38.1.g] [Medline: 18624019]
20. Serrano-Ripoll MJ, Zamanillo-Campos R, Fiol-DeRoque MA, Castro A, Ricci-Cabello I. Impact of smartphone app-based psychological interventions for reducing depressive symptoms in people with depression: systematic literature review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2022 Jan 27;10(1):e29621. [doi: 10.2196/29621] [Medline: 35084346]
21. Wang M, Chen H, Yang F, Xu X, Li J. Effects of digital psychotherapy for depression and anxiety: a systematic review and bayesian network meta-analysis. J Affect Disord 2023 Oct;338:569-580. [doi: 10.1016/j.jad.2023.06.057]
22. Bae H, Shin H, Ji HG, Kwon JS, Kim H, Hur JW. App-based interventions for moderate to severe depression: a systematic review and meta-analysis. JAMA Netw Open 2023 Nov 1;6(11):e2344120. [doi: 10.1001/jamanetworkopen.2023.44120] [Medline: 37983028]
23. Buss JF, Steinberg JS, Banks G, et al. Availability of internet-based cognitive-behavioral therapies for depression: a systematic review. Behav Ther 2024 Jan;55(1):201-211. [doi: 10.1016/j.beth.2023.06.003] [Medline: 38216233]
24. Huguet A, Rao S, McGrath PJ, et al. A systematic review of cognitive behavioral therapy and behavioral activation apps for depression. PLoS One 2016;11(5):e0154248. [doi: 10.1371/journal.pone.0154248] [Medline: 27135410]
25. Wasil AR, Venturo-Conerly KE, Shingleton RM, Weisz JR. A review of popular smartphone apps for depression and anxiety: assessing the inclusion of evidence-based content. Behav Res Ther 2019 Dec;123:103498. [doi: 10.1016/j.brat.2019.103498] [Medline: 31707224]
26. Bowie-DaBreo D, Sünram-Lea SI, Sas C, Iles-Smith H. Evaluation of treatment descriptions and alignment with clinical guidance of apps for depression on app stores: systematic search and content analysis. JMIR Form Res 2020 Nov 13;4(11):e14988. [doi: 10.2196/14988] [Medline: 33185566]
27. Bubolz S, Mayer G, Gronewold N, Hilbel T, Schultz JH. Adherence to established treatment guidelines among unguided digital interventions for depression: quality evaluation of 28 web-based programs and mobile apps. J Med Internet Res 2020 Jul 13;22(7):e16136. [doi: 10.2196/16136] [Medline: 32673221]
28. Stawarz K, Preist C, Tallon D, Wiles N, Coyle D. User experience of cognitive behavioral therapy apps for depression: an analysis of app functionality and user reviews. J Med Internet Res 2018 Jun 6;20(6):e10120. [doi: 10.2196/10120] [Medline: 29875087]
29. Martinengo L, Stona AC, Griva K, et al. Self-guided cognitive behavioral therapy apps for depression: systematic assessment of features, functionality, and congruence with evidence. J Med Internet Res 2021 Jul 30;23(7):e27619. [doi: 10.2196/27619] [Medline: 34328431]

XSL•FO
RenderX

30. Plessen CY, Panagiotopoulou OM, Tong L, Cuijpers P, Karyotaki E. Digital mental health interventions for the treatment of depression: a multiverse meta-analysis. J Affect Disord 2025 Jan 15;369:1031-1044. [doi: 10.1016/j.jad.2024.10.018] [Medline: 39419189]
31. López-López JA, Davies SR, Caldwell DM, et al. The process and delivery of CBT for depression in adults: a systematic review and network meta-analysis. Psychol Med 2019 Sep;49(12):1937-1947. [doi: 10.1017/S003329171900120X] [Medline: 31179960]
32. Jackson HM, Calear AL, Batterham PJ, Ohan JL, Farmer GM, Farrer LM. Skill enactment and knowledge acquisition in digital cognitive behavioral therapy for depression and anxiety: systematic review of randomized controlled trials. J Med Internet Res 2023 May 31;25:e44673. [doi: 10.2196/44673] [Medline: 37256673]
33. Hundt NE, Mignogna J, Underhill C, Cully JA. The relationship between use of CBT skills and depression treatment outcome: a theoretical and methodological review of the literature. Behav Ther 2013 Mar;44(1):12-26. [doi: 10.1016/j.beth.2012.10.001] [Medline: 23312423]
34. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
35. Bobrow K, Brennan T, Springer D, et al. Efficacy of a text messaging (SMS) based intervention for adults with hypertension: protocol for the StAR (SMS Text-message Adherence suppoRt trial) randomised controlled trial. BMC Public Health 2014 Jan 11;14(1):1-9. [doi: 10.1186/1471-2458-14-28] [Medline: 24410738]
36. Devi R, Powell J, Singh S. A web-based program improves physical activity outcomes in a primary care angina population: randomized controlled trial. J Med Internet Res 2014 Sep 12;16(9):e186. [doi: 10.2196/jmir.3340] [Medline: 25217464]
37. Yang CH, Maher JP, Conroy DE. Implementation of behavior change techniques in mobile applications for physical activity. Am J Prev Med 2015 Apr;48(4):452-455. [doi: 10.1016/j.amepre.2014.10.010] [Medline: 25576494]
38. Garnett C, Crane D, West R, Brown J, Michie S. Identification of behavior change techniques and engagement strategies to design a smartphone app to reduce alcohol consumption using a formal consensus method. JMIR Mhealth Uhealth 2015 Jun 29;3(2):e73. [doi: 10.2196/mhealth.3895] [Medline: 26123578]
39. Struik L, Rodberg D, Sharma RH. The behavior change techniques used in Canadian online smoking cessation programs: content analysis. JMIR Ment Health 2022 Mar 1;9(3):e35234. [doi: 10.2196/35234] [Medline: 35230253]
40. Webster R, Michie S, Estcourt C, Gerressu M, Bailey JV, Group MT. Increasing condom use in heterosexual men: development of a theory-based interactive digital intervention. Transl Behav Med 2016 Sep;6(3):418-427. [doi: 10.1007/s13142-015-0338-8] [Medline: 27528531]
41. Milne-Ives M, Lam C, De Cock C, Van Velthoven MH, Meinert E. Mobile apps for health behavior change in physical activity, diet, drug and alcohol use, and mental health: systematic review. JMIR Mhealth Uhealth 2020 Mar 18;8(3):e17046. [doi: 10.2196/17046] [Medline: 32186518]
42. Goulding EH, Dopke CA, Rossom RC, et al. A smartphone-based self-management intervention for individuals with bipolar disorder (LiveWell): empirical and theoretical framework, intervention design, and study protocol for a randomized controlled trial. JMIR Res Protoc 2022 Feb 21;11(2):e30710. [doi: 10.2196/30710] [Medline: 35188473]
43. Thomas PC, Curtis K, Potts HWW, et al. Behavior change techniques within digital interventions for the treatment of eating disorders: systematic review and meta-analysis. JMIR Ment Health 2024 Aug 1;11:e57577. [doi: 10.2196/57577] [Medline: 39088817]
44. Chiang CP, Hayes D, Panagiotopoulou E. Apps targeting anorexia nervosa in young people: a systematic review of active ingredients. Transl Behav Med 2023 Jun 9;13(6):406-417. [doi: 10.1093/tbm/ibad003] [Medline: 36753537]
45. Watkins E, Newbold A, Tester-Jones M, Collins LM, Mostazir M. Investigation of active ingredients within internet-delivered cognitive behavioral therapy for depression: a randomized optimization trial. JAMA Psychiatry 2023 Sep 1;80(9):942-951. [doi: 10.1001/jamapsychiatry.2023.1937] [Medline: 37378962]
46. Wilhelm S, Bernstein EE, Bentley KH, et al. Feasibility, acceptability, and preliminary efficacy of a smartphone app-led cognitive behavioral therapy for depression under therapist supervision: open trial. JMIR Ment Health 2024 Apr 9;11(1):e53998. [doi: 10.2196/53998] [Medline: 38592771]
47. Karyotaki E, Efthimiou O, Miguel C, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021 Apr 1;78(4):361-371. [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
48. Gratzer D, Khalid-Khan F. Internet-delivered cognitive behavioural therapy in the treatment of psychiatric illness. CMAJ 2016 Mar 1;188(4):263-272. [doi: 10.1503/cmaj.150007] [Medline: 26527829]
49. Williams JBW. A structured interview guide for the Hamilton Depression Rating Scale. Arch Gen Psychiatry 1988 Aug 1;45(8):742. [doi: 10.1001/archpsyc.1988.01800320058007]
50. Gautam M, Tripathi A, Deshmukh D, Gaur M. Cognitive behavioral therapy for depression. Indian J Psychiatry 2020 Jan;62(Suppl 2):S223-S229. [doi: 10.4103/psychiatry.IndianJPsychiatry_772_19] [Medline: 32055065]
51. Michie S, Johnston M, Richardson M, Francis J, Abraham C, Wood CE, et al. BCTTv1 online training. BCT Taxonomy. URL: https://www.bct-taxonomy.com [accessed 2026-03-14]

XSL•FO
RenderX

52. Carey RN, Connell LE, Johnston M, et al. Behavior change techniques and their mechanisms of action: a synthesis of links described in published intervention literature. Ann Behav Med 2019 Jul 17;53(8):693-707. [doi: 10.1093/abm/kay078] [Medline: 30304386]
53. Cane J, O'Connor D, Michie S. Validation of the theoretical domains framework for use in behaviour change and implementation research. Implement Sci 2012 Apr 24;7(1):37. [doi: 10.1186/1748-5908-7-37] [Medline: 22530986]
54. Willmott TJ, Pang B, Rundle-Thiele S. Capability, opportunity, and motivation: an across contexts empirical examination of the COM-B model. BMC Public Health 2021 May 29;21(1):1014. [doi: 10.1186/s12889-021-11019-w] [Medline: 34051788]
55. Richards KAR, Hemphill MA. A practical guide to collaborative qualitative data analysis. J Teach Phys Educ 2018 Apr;37(2):225-231. [doi: 10.1123/jtpe.2017-0084]
56. Salmona M, Lieber E, Kaczynski D. Qualitative and Mixed Methods Data Analysis Using Dedoose: A Practical Approach for Research across the Social Sciences: Sage Publications; 2019.
57. O'Connor C, Joffe H. Intercoder reliability in qualitative research: debates and practical guidelines. Int J Qual Methods 2020 Jan 1;19:1609406919899220. [doi: 10.1177/1609406919899220]
58. Burla L, Knierim B, Barth J, Liewald K, Duetz M, Abel T. From text to codings: intercoder reliability assessment in qualitative content analysis. Nurs Res 2008;57(2):113-117. [doi: 10.1097/01.NNR.0000313482.33917.7d] [Medline: 18347483]
59. R Core Team. R: a language and environment for statistical computing. : R Foundation for Statistical Computing; 2020 URL: https://www.r-project.org/ [accessed 2026-04-10]
60. Kolde R, Kolde MR. Package 'pheatmap'. R package 2015;1(7):790 [FREE Full text]
61. Gehred AP. Canva. JMLA 2020;108(2):338. [doi: 10.5195/jmla.2020.923]
62. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: Template for Intervention Description and Replication (TIDieR) checklist and guide. BMJ 2014 Mar 7;348:g1687. [doi: 10.1136/bmj.g1687] [Medline: 24609605]
63. Lin X, Martinengo L, Jabir AI, et al. Scope, characteristics, behavior change techniques, and quality of conversational agents for mental health and well-being: systematic assessment of apps. J Med Internet Res 2023 Jul 18;25:e45984. [doi: 10.2196/45984] [Medline: 37463036]
64. Martinengo L, Jabir AI, Goh WWT, et al. Conversational agents in health care: scoping review of their behavior change techniques and underpinning theory. J Med Internet Res 2022 Oct 3;24(10):e39243. [doi: 10.2196/39243] [Medline: 36190749]
65. Alqahtani F, Al Khalifah G, Oyebode O, Orji R. Apps for mental health: an evaluation of behavior change strategies and recommendations for future development. Front Artif Intell 2019;2:30. [doi: 10.3389/frai.2019.00030] [Medline: 33733119]

Abbreviations

BCT: behavior change technique
BCTTv1: behavior change technique Taxonomy version 1
CBT: cognitive behavioral therapy
COM-B: Capability, Opportunity, Motivation–Behavior
HIPAA: Health Insurance Portability and Accountability Act
TIDieR: Template for Intervention Description and Replication

Edited by S Brini; submitted 12.Sep.2025; peer-reviewed by B Najand, J Angel; revised version received 12.Jan.2026; accepted 19.Jan.2026; published 16.Apr.2026.

Please cite as:
Jonathan GK, Sung JY, Arcese HT, Holz P, Bentley KH, Bernstein EE, Weingarden HM, Greenberg JL, Harrison OT, Wilhelm S
Identifying Evidence-Based Strategies in a Digital Mental Health Intervention for Depression: Qualitative Content Analysis
J Med Internet Res 2026;28:e84030
URL: https://www.jmir.org/2026/1/e84030
doi:10.2196/84030

© Geneva K Jonathan, Jenna Y Sung, Heyli T Arcese, Phoebe Holz, Kathryn H Bentley, Emily E Bernstein, Hilary M Weingarden, Jennifer L Greenberg, Oliver T Harrison, Sabine Wilhelm. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly

XSL•FO
RenderX

cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Long-Term Outcomes, Moderators, and Predictors in Online Mindfulness–Based Cognitive Therapy for People With Cancer: Secondary Analysis of a Randomized Controlled Trial

Nasim Badaghi[1], MSc; Jelle Lubbers[1], PhD; Judith Prins[1], Prof Dr; Saskia Kelders[2,3], PhD; Anne Speckens[1], Prof Dr; Linda Kwakkenbos[1,4], PhD

[1]Department of Psychiatry, Radboud University Medical Center, Geert Grooteplein Zuid 10, Nijmegen, The Netherlands

[2]Department of Psychology, Health and Technology, University of Twente, Twente, The Netherlands

[3]Optentia Research Focus Area, North-West University, Vanderbijlpark, South Africa

[4]Department of Clinical Psychology, Behavioural Science Institute, Radboud University, Nijmegen, The Netherlands

Corresponding Author:
Nasim Badaghi, MSc
Department of Psychiatry, Radboud University Medical Center, Geert Grooteplein Zuid 10, Nijmegen, The Netherlands

Abstract

Background: A 3-armed randomized controlled trial (RCT) demonstrated that 2 formats of online mindfulness–based cognitive therapy (eMBCT)—group-blended and individual-unguided—effectively reduced psychological distress and improved positive health outcomes in people with cancer and survivors, when compared to care as usual, up to 3 months posttreatment. However, the long-term effectiveness and possible predictors and moderators of treatment outcomes remain unknown.

Objective: This study examined the long-term effects (up to 9 months posttreatment) of group-blended and individual-unguided eMBCTs on psychological distress and other mental health outcomes in people with cancer. Additionally, it explored possible predictors and moderators of treatment effects across the 2 delivery formats.

Methods: The study population consisted of people with cancer who were initially randomized to group-blended or individual-unguided eMBCT within a 3-arm RCT, augmented by those who completed the waitlist condition and were subsequently randomly allocated to one of the eMBCT formats. Both groups were assessed over a 9-month follow-up period. Outcomes completed at baseline, posttreatment, and 3-, 6-, and 9-month follow-up assessments included psychological distress (primary), fear of cancer recurrence, rumination, fatigue, mindfulness skills, decentering, self-compassion, and well-being. Linear mixed-effects models examined changes over time, while linear mixed-effects models and binary logistic regression analyzed potential predictors and moderators of psychological distress and dropout.

Results: Of the 186 participants enrolled in the 3-arm RCT, 161 participants were randomly assigned to either group-blended or individual-unguided eMBCTs after adding those initially assigned to the waiting condition (group-blended: n=71; individual-unguided: n=90). The majority of participants were female (n=129, 80%), diagnosed with breast cancer (n=78, 48%), and were undergoing or had completed treatment with curative intent (n=124, 77%). The mean age was 52.8 (SD 11.4) years. Both eMBCT formats resulted in significant reductions in psychological distress, fear of cancer recurrence, rumination, and fatigue, alongside improvements in mindfulness skills, decentering, and self-compassion, up to 9 months posttreatment. Higher baseline rumination, as well as lower mindfulness skills and self-compassion at baseline, predicted larger reductions in psychological distress from baseline to the 9-month follow-up period. Additionally, highly distressed participants in the group-blended eMBCT arm were more prone to dropout than those with lower distress scores, whereas psychological distress was not associated with dropout in the individual-unguided format. No other significant moderators were identified.

Conclusions: Group-blended and individual-unguided eMBCTs effectively reduced psychological distress and improved well-being among people with cancer and survivors, with greater benefits for those with fewer psychological resources. However, individuals experiencing higher levels of distress were more likely to discontinue group-blended eMBCT. These findings highlight the importance of considering individual preferences and pragmatic factors in treatment decisions. Larger, fully powered RCTs are needed to confirm these results and provide more definitive guidance on treatment format selection.

Trial Registration: Dutch Registry CCMO NL73117.091.20; https://onderzoekmetmensen.nl/en/trial/49163 and ClinicalTrials.gov NCT05336916; https://clinicaltrials.gov/study/NCT05336916

(J Med Internet Res 2026;28:e79928) doi:10.2196/79928

XSL•FO
RenderX

KEYWORDS

telehealth; cancer; clinical trials; long-term effect; mindfulness; oncology

Introduction

Background

Cancer can strongly affect a person's life, beginning at the moment of diagnosis and possibly lasting a lifetime. Although the growing number of cancer survivors reflects improvements in medical progress [1], it also presents a challenge: more people must live with the long-term consequences of the disease. Research indicates that people with cancer can experience detrimental psychological effects that may persist after completing cancer treatment, significantly impacting the quality of life and mental health of cancer survivors for years [2]. In addition, it has been shown that long-term cancer survivors (≥5 y after diagnosis) experience reduced quality of life, fatigue, and neuropsychological issues, such as memory and attention deficits [3].

Mindfulness-based interventions (MBIs) have proven to be effective in alleviating the psychological challenges encountered by people with cancer [4,5]. There is robust evidence showing that, in oncological settings, MBIs are not only effective in symptom reduction [4,5] but also in fostering positive health outcomes such as mindfulness skills, self-compassion, and personal growth [6]. At the core of MBIs lies mindfulness, which is defined as the awareness that emerges from paying attention to the present moment [7]. MBIs help people recognize their emotions, thoughts, bodily sensations, and automatic reactions with nonjudgmental awareness and kindness, allowing them to respond in a more adaptive and beneficial way. MBIs are traditionally delivered through in-person group sessions, typically lasting around 2.5 hours, and led by a trained mindfulness teacher.

While MBIs have been shown to be effective for people with cancer, attending face-to-face group sessions can be challenging due to travel demands and limited scheduling flexibility [8]. To address these barriers, online MBIs (eMBIs) have been developed, showing strong evidence for reducing distress, depressive symptoms, stress, and sleep problems, while improving quality of life [9-11]. However, little is known about their long-term impact, or possible predictors and moderators of treatment outcomes in different eMBI formats.

Long-Term Effects

According to a recent meta-analysis by Chayadi et al [12], MBIs can significantly reduce symptoms of depression (Hedges g=0.43), anxiety (Hedges g=0.55), and cancer-related fatigue (Hedges g=0.43) in oncology settings up to at least 3 months post-MBI. In addition, at longer follow-ups (6 months on average), MBIs also demonstrate statistically significant effects on psychological distress, anxiety, depression, and fatigue, with effect sizes ranging from Hedges g=0.19 to 0.36 [4]. Research on the long-term effects of eMBIs, however, remains relatively scarce. Current evidence does not yet clarify whether the benefits of eMBIs are sustained over extended periods [11]. Given the increasing prevalence of cancer survivors and the growing demand for more flexible and accessible interventions, establishing the long-term effectiveness of eMBIs is highly relevant.

Predictors and Moderators

Although MBIs or eMBIs have been shown to be beneficial, not everyone may benefit in the same way. Understanding these differences is essential for tailoring interventions to meet personal needs and enhance their effectiveness. Research on predictors and moderators can provide valuable insights into determining what works best for whom. A predictor is a variable that forecasts an outcome regardless of intervention condition, while a moderator influences the strength or direction of the relationship between a predictor and an outcome. Prior research found that lower baseline levels of psychological distress, rumination, and neuroticism, as well as higher levels of extraversion and agreeableness, predicted lower levels of distress at a 9-month follow-up after face-to-face or online mindfulness–based cognitive therapy (eMBCT), a specific form of eMBI [13]. Additionally, individuals who were less mindful and conscientious benefited more from therapist-assisted eMBCT than face-to-face MBCT [13]. However, the results were derived from 1 study only. Predictors and moderators of more scalable eMBIs such as individual-unguided formats remain unidentified. Recent studies highlight the need to explore these variables to enhance the effectiveness and accessibility of eMBIs for people with cancer [10,14].

Current Research

We recently completed a 3-arm randomized controlled trial (RCT) that evaluated the effects of 2 formats of eMBCT (ie, group-blended and individual-unguided) on psychological distress in people with cancer [15]. At posttreatment, only the group-blended eMBCT significantly reduced psychological distress. However, at the 3-month follow-up, both intervention formats appeared to be effective in comparison to care as usual (CAU). Additionally, both formats led to reductions in rumination and improvements in mindfulness skills and decentering, whereas only the group-blended eMBCT also increased self-compassion at the 3-month follow-up.

This study aimed to examine the long-term effects of group-blended and individual-unguided eMBCTs on individuals with cancer over the course of the 9-month follow-up. Specifically, we aimed to explore the effects of both formats on psychological distress, fear of cancer recurrence, fatigue, rumination, mindfulness skills, decentering, self-compassion, and well-being.

Additionally, we investigated potential predictors and moderators that may influence treatment effects. Among the predictors and moderators examined, this study focused on rumination, mindfulness, decentering, and self-compassion. A recent scoping review of mediators of MBIs identified mindfulness skills, decentering, and mindfulness-related attitudes (such as self-compassion) as the most consistent mechanisms linking MBIs to improvements in mental health

XSL•FO
RenderX

and well-being across adult clinical and nonclinical populations [16]. These mechanisms primarily explained reductions in symptoms, such as stress, anxiety, and depression, and to a lesser extent improvements in well-being [16]. In people with cancer, rumination has also been shown to mediate MBI effects, including in the trial underlying the follow-up study [17,18], and different components of mindfulness skills have been identified as mediators in this population [18]. Additionally, individuals with lower baseline mindfulness skills appear to benefit more from therapist-assisted eMBCT than from traditional MBCT in the long term [13]. Decentering has likewise been proposed as a mediator of mindfulness in people with cancer [19] and may help counteract rumination by offering a protective mechanism. Together, these factors are important for understanding how individual characteristics interact with different intervention formats.

Methods

Follow-up data from a 3-arm, parallel RCT were used [15]. Initially, participants were randomized to group-blended eMBCT, individual-unguided eMBCT, or CAU. After completing their 3-month CAU follow-up period, participants in the CAU condition were randomized to either group-blended or individual-unguided eMBCT.

Ethical Considerations

This study was approved by the ethical review board of Radboud University Medical Center (Central Committee on Research Involving Human Subjects Arnhem-Nijmegen; NL73117.091.20), registered in the Dutch Registry CCMO (NL73117.091.20) and on ClinicalTrials.gov (NCT05336916). We published a protocol in advance of trial completion [15] and conducted an uncontrolled feasibility study before running the full RCT [20]. Eligible participants were sent comprehensive written information about the study and an informed consent form via email and regular mail. Once the signed consent form was returned, participants were enrolled in the trial and received a link via email to complete baseline assessments. All participants provided informed consent prior to participation. Their privacy and confidentiality were fully protected, and no compensation was provided for their participation.

Participants

Because the methods used in studies based on our RCT data are similar, we have reused some methodological descriptions from previous publications [15,20]. In doing so, we followed the guidelines provided by the Text Recycling Research Project [21]. Participants were recruited between January 2021 and September 2023 via posters in health care settings, flyers handed out by health care professionals, posts in regional newspapers, online posts placed on websites of cancer-related organizations, and social media platforms. Eligible participants were adults who had been diagnosed with cancer (irrespective of type or stage of cancer and time since diagnosis), had internet access, the ability to use a computer, and a good command of the Dutch language. Participants were excluded from the study if they had previously attended more than 4 sessions of any MBI, had a severe psychiatric comorbidity, drug or alcohol dependence, significant cognitive impairment, or were unwilling to be randomized to 1 of the 3 study conditions. Participants could continue any ongoing medical, psychological, or paramedical care during the study, except for participation in other MBIs outside the study context.

Procedure

The study was conducted at the Radboudumc Expertise Center for Mindfulness in Nijmegen, The Netherlands. Interested participants could reach out via email, phone, or the contact form on the trial website. Eligible participants were sent comprehensive written information about the study and an informed consent form via email and regular mail. Once the signed consent form was returned, participants were enrolled in the trial and received a link via email to complete baseline assessments. Baseline assessments consisted of online self-reported questionnaires, including demographic characteristics and various patient-reported outcome measures. After participants completed the questionnaires, they were randomized into one of three conditions (group-blended, individual-unguided, or CAU) by an independent research assistant using the secure data management platform Castor EDC. Randomization was stratified based on 2 variables: cancer type (breast cancer vs other types) and treatment intent (curative vs palliative).

For participants in the CAU group who were randomized a second time, the same procedure as the first randomization in the RCT was followed. After completing the CAU period, a research assistant contacted them by email or phone to ask about their willingness to participate in eMBCT. If they agreed, they completed the T2 outcome assessments (new baseline) and were then randomized to either group-blended or individual-unguided eMBCT. While this rerandomization allowed for comparisons between the 2 eMBCT interventions, it is important to note that there is no CAU control group after the 3-month follow-up period.

Intervention

Overview

The group-blended and individual-unguided eMBCT programs were adaptations of the MBCT protocol developed by Segal et al [22]. The intervention formats followed the standard 8-week MBCT structure, with weekly 2.5-hour sessions incorporating formal and informal mindfulness practices, reflective exercises, daily home practice (30 min to 1 h), and a silent day. The program content was adapted to include psychoeducation on mindfulness for cancer and grief, and the movement exercises were modified to accommodate people with cancer. Detailed descriptions of the intervention content, development process, and conditions have been published elsewhere [15].

Intervention Delivery

Participants could access the eMBCT program in a private section of the Minddistrict platform (Minddistrict Development BV), which provided all necessary program information and materials. Participants created accounts and accessed the program through an internet browser on a website or by downloading the application onto their devices. While the session content was consistent across both conditions, the

XSL•FO
RenderX

delivery methods differed as follows: (1) group-blended eMBCT, where sessions 1, 3, 5, and 8 were conducted as live group sessions via Zoom (Zoom Communications, Inc), led by a mindfulness teacher; and (2) individual-unguided eMBCT, where all sessions were fully self-guided through the Minddistrict platform, where participants engaged with an avatar rather than a physical teacher.

Mindfulness Teachers

The mindfulness teachers in the group-blended condition were 2 health care professionals (a psychologist and a social worker) with expertise in psycho-oncology and several years of experience teaching mindfulness- and compassion-based courses. They met the qualification standards set by the Association of Mindfulness Teachers in the Netherlands and Flanders, which are aligned with the UK Network for Mindfulness-Based Teachers' criteria (2015) [23]. To safeguard consistency, adherence to the protocol, and high-quality program delivery, the teachers attended regular peer supervision sessions with a senior mindfulness teacher (AS). Additionally, they participated in weekly trial research meetings to stay aligned with the study's objectives and address any challenges encountered during intervention delivery.

Outcome Measures

Participants completed online questionnaires at 6 time points: baseline (T0); midtreatment; posttreatment (T1); and 3 (T2), 6 (T3), and 9 (T4) months after treatment [15]. Participants in the CAU condition were randomized into either the group-blended or individual-unguided eMBCT after the 3-month follow-up period (T2). For this study, T2 completed by CAU participants was used as baseline, as it better reflected their present state at the start of the intervention. Therefore, participants in the CAU group completed all the assessments mentioned above, as well as 2 additional ones (T5 and T6), which corresponded to the follow-up period for their intervention. The data analyses include results from all 5 time points—T0-T4 for the intervention group and T2-T6 for the participants in the CAU group, who were rerandomized into either the group-blended or individual-unguided eMBCT after 6 months (T2). For the follow-up analyses, T2 became their new T0 measurement.

The primary outcome measure was psychological distress, assessed using the Hospital Anxiety and Depression Scale [24]. Secondary outcomes included the Fear of Cancer Recurrence Inventory [25], the Rumination subscale of the Rumination and Reflection Questionnaire (RRQ) [26], the Fatigue Severity subscale of the Checklist Individual Strength [27], the Five Facet Mindfulness Questionnaire–Short Form (FFMQ-SF) [28], the Decentering subscale of the Experiences Questionnaire (EQ) [29], the Self-Compassion Scale–Short Form (SCS-SF) [30], and the Mental Health Continuum–Short Form [31]. Detailed information about each questionnaire has been published previously [15].

Several constructs (rumination, mindfulness skills, decentering, and self-compassion) were examined in different analytic roles in this study. When analyzed as secondary outcomes, their repeated measurements across all assessment time points were used to evaluate changes over time following the intervention. In contrast, in the prediction and moderation analyses, only baseline values of these constructs were included as predictors or moderators of treatment dropout and psychological distress trajectories. Time-varying values of these variables were not used as predictors in any model.

Predictors and Moderators

The selection of predictors and moderators was informed by prior empirical studies [4,13]. We examined predictors and moderators of psychological distress (measured by the Hospital Anxiety and Depression Scale) over a 9-month follow-up period, as well as dropout at posttreatment (defined as completing fewer than 4 sessions). It is important to note that dropout was treated as a dependent variable because the study aimed to explore whether baseline characteristics could predict dropout both generally and in relation to the 2 intervention conditions. The variables tested included sex, age, cancer diagnosis (breast cancer vs other types), treatment intent (palliative vs curative), time since diagnosis, and baseline levels of rumination (RRQ), mindfulness skills (FFMQ-SF), decentering (EQ), and self-compassion (SCS-SF).

Statistical Analyses

Overview

All analyses were conducted in SPSS version 29 (IBM Corp). For the CAU group, baseline scores on the primary and secondary outcomes were substituted with post-CAU scores (corresponding to T2 in the initial 3-armed RCT), as these scores were temporally aligned with the start of the intervention. For all analyses, the data from participants who followed group-blended or individual-unguided eMBCT after CAU were combined with the data from participants who were initially randomized to these 2 conditions. After analyzing baseline differences in participant characteristics and outcome measures across the full sample, and comparing baseline scores between initially randomized participants and those who were rerandomized, the only variable that significantly differed between individuals who completed the primary outcome measure at posttreatment and those who did not was fatigue. Noncompleters reported significantly higher fatigue levels, as shown by elevated Fatigue Severity subscale of the Checklist Individual Strength scores (P=.02). This indicates that missingness was related to an observed participant characteristic. Assuming missing at random, linear mixed-effects models with (restricted) maximum likelihood estimation were used, which can accommodate incomplete data when missingness depends on observed variables.

Long-Term Effects

We built separate linear mixed-effects models for each intervention condition and dependent variable to explore the long-term effects on our primary and secondary outcomes. These linear mixed-effects models used the scores of the primary (psychological distress) or secondary (fear of cancer recurrence, rumination, fatigue, mindfulness skills, decentering, self-compassion, or well-being) outcomes as the dependent variable. Independent variables included time (baseline, posttreatment, 3-month follow-up, 6-month follow-up, and 9-month follow-up), intervention condition (group-blended or

XSL•FO
RenderX

individual-unguided eMBCT), and the time×intervention condition interaction. Time was treated as a continuous variable. Random intercepts and random slopes for participants were added. Furthermore, a heterogeneous first-order autoregressive (ARH[1]) covariance structure was applied to model stronger correlations between measurements taken closer together in time compared to those further apart. Restricted maximum likelihood was used to calculate the estimates [32]. In addition, effect sizes over the 9-month follow-up were calculated using Cohen *d* [33], and according to the conventions suggested by Cohen [33], effect sizes of 0.2, 0.5, and 0.8 were considered small, medium, and large, respectively.

Prediction and Moderation

Prediction and moderation analyses of psychological distress were conducted using linear mixed models [34]. The dependent variable was psychological distress over the 9-month follow-up period, measured at 5 time points: baseline, posttreatment, and at 3-, 6-, and 9-month follow-ups. Independent variables included time and intervention condition (group-blended eMBCT vs individual-unguided eMBCT). Baseline characteristics used as predictors and moderators were sex, age, cancer diagnosis, cancer treatment intent, time since diagnosis, psychological distress, rumination (RRQ), mindfulness skills (FFMQ-SF), decentering (EQ), and self-compassion (SCS-SF). To assess moderation, interaction terms between intervention condition, moderator variables, and time were included. Random effects were specified for participants to account for within-subject correlations over time [35]. As the analyses were exploratory, original *P* values uncorrected for multiple comparisons were reported.

We defined intervention dropout as attending fewer than 4 of the 8 sessions, which is consistent with other MBI trials [36-38]. An adequate dose of MBCT is typically considered to be participation in at least 4 of the 8 sessions, which captures those participants who have meaningfully engaged with the program [39]. Typical MBCT attendance rates show that over 85% of participants attend all or nearly all of the sessions, with dropouts generally occurring between the orientation and first group session or within the initial few sessions [39]. This attendance distribution justifies the use of the 4-session threshold, aligning with established practices in the field and maintaining consistency with the main RCT of these data. Prediction and moderation analyses of dropout were conducted using binary logistic regression [40]. The dependent variable was dropout status, and the independent variable was the intervention condition. The same baseline characteristics were used as predictors and moderators. Moderation was assessed by including interaction terms between the intervention condition and each moderator.

For a detailed overview of the analyses conducted, please refer to the statistical analysis plan provided in Multimedia Appendix 1.

It is important to note that no power calculation was conducted for the long-term effects or moderation analyses, and the study's sample size (N=161), along with substantial dropout, may have limited the statistical power to detect small or moderate effects, particularly for 3-way interactions (time× condition × moderator). Therefore, these analyses should be considered exploratory.

Results

Overview

A total of 186 participants were originally randomized to group-blended, individual-unguided, and CAU. Of the 54 participants randomized to CAU, 29 were willing to be rerandomized and were assigned to group-blended (n=14) and individual-unguided (n=15) eMBCTs (Table 1; Multimedia Appendix 2). There were no differences in baseline characteristics between those originally randomized to the 2 formats and those who first completed CAU. Of the 161 participants randomized to group-blended or individual-unguided eMBCT, 48 (30%) dropped out after completing fewer than 4 sessions. Dropout rates were significantly higher in the individual-unguided group (n=37, 41%) compared to the group-blended group (n=11, 15%; $P<.001$).

XSL•FO
RenderX

Table . Baseline sociodemographic and clinical characteristics (N=161).

Characteristic	All (N=161)	Group-blended (n=71)	Individual-unguided (n=90)
Sex, n (%)			
Female	129 (80.1)	55 (77.5)	74 (82.2)
Male	32 (19.9)	16 (22.5)	16 (17.8)
Age (y), mean (SD)	52.8 (11.4)	54.6 (10.9)	51.4 (11.6)
Diagnosis, n (%)			
Breast cancer	78 (48.4)	33 (46.5)	45 (50)
Blood cancer	14 (8.7)	7 (9.9)	7 (7.8)
Prostate cancer	8 (5.0)	4 (5.6)	4 (4.4)
Intestine cancer	7 (4.3)	3 (4.2)	4 (4.4)
Ovarian cancer	7 (4.3)	3 (4.2)	4 (4.4)
Skin cancer	7 (4.3)	4 (5.6)	3 (3.3)
Colon cancer	4 (2.5)	2 (2.8)	2 (2.2)
Other	36 (22.5)	15 (21.2)	21 (23.3)
Treatment intent, n (%)			
Curative	124 (77.0)	55 (77.5)	69 (76.7)
Palliative	37 (23.0)	16 (22.5)	21 (23.3)
Occupation, n (%)			
Employed	99 (61.6)	42 (59.2)	57 (63.3)
Retired	26 (16.1)	12 (16.9)	14 (15.6)
On disability	16 (9.9)	6 (8.5)	10 (11.1)
Houseman or housewife	10 (6.2)	6 (8.5)	4 (4.4)
Unemployed	5 (3.1)	2 (2.8)	3 (3.3)
Other (not specified)	5 (3.1)	3 (4.2)	2 (2.2)
Level of education, n (%)			
High	97 (60.2)	43 (60.6)	54 (60)
Middle	53 (32.9)	22 (31)	31 (34.5)
Low	7 (4.3)	4 (5.6)	3 (3.3)
Other (not specified)	4 (2.5)	2 (2.8)	2 (2.2)
Time since diagnosis (y), n (%)			
0 - 1	3 (1.9)	2 (2.8)	1 (1.1)
3	38 (23.6)	16 (22.5)	22 (24.4)
5	62 (38.5)	26 (36.6)	36 (40)
>5	58 (36)	27 (38)	31 (34.4)
Married or in a relationship, n (%)			
Yes	108 (67.1)	46 (64.8)	62 (68.9)
No	53 (32.9)	25 (35.2)	28 (31.1)

Long-Term Effects

Results from the separate linear mixed-effects models examining the within-group long-term effects of group-blended and individual-unguided eMBCTs showed that both formats led to significant improvements in all primary and secondary psychological outcomes from baseline to the 9-month follow-up. For every outcome and in both intervention conditions, *P* values were below .05. Table 2 shows the descriptive statistics per outcome and intervention condition across time.

XSL•FO
RenderX

Table . Descriptive statistics and effects of the interaction between time and intervention condition on primary and secondary outcomes in a linear mixed-effects model.

Outcome measure and time	Group-blended		Individual-unguided		Time × intervention condition (*P* value)
	Score, mean (SD)	Participants, n	Score, mean (SD)	Participants, n	
Psychological distress					.76
T0	15.50 (6.32)	69	15.05 (6.94)	90	
T1	10.94 (5.72)	51	11.52 (7.30)	65	
T2	10.58 (5.97)	51	10.94 (5.52)	66	
T3	10.12 (9.50)	50	10.30 (5.85)	53	
T4	10.88 (6.99)	43	9.72 (4.53)	53	
Fear of cancer recurrence					.19
T0	77.84 (22.82)	69	77.06 (24.04)	90	
T1	69.31 (21.16)	49	70.56 (23.53)	63	
T2	65.98 (19.83)	51	70.17 (23.32)	65	
T3	64.40 (16.92)	50	66.87 (23.48)	52	
T4	64.76 (19.79)	42	67.00 (20.79)	51	
Fatigue severity					.19
T0	37.12 (11.16)	69	35.51 (12.01)	90	
T1	32.37 (11.59)	51	32.35 (12.47)	63	
T2	31.49 (12.74)	51	31.86 (11.03)	65	
T3	29.78 (12.19)	50	30.61 (13.79)	52	
T4	29.19 (11.37)	42	29.94 (12.16)	52	
Rumination					.14
T0	3.42 (0.66)	69	3.27 (0.83)	89	
T1	3.06 (0.72)	49	2.98 (0.77)	62	
T2	3.08 (0.78)	51	2.95 (0.78)	64	
T3	3.01 (0.77)	48	2.87 (0.74)	51	
T4	3.05 (0.81)	41	2.91 (0.77)	49	
Mindfulness skills					.11
T0	78.91 (11.21)	69	78.60 (11.26)	90	
T1	84.79 (10.89)	49	84.86 (12.52)	63	
T2	87.37 (10.75)	51	85.26 (11.50)	65	
T3	87.12 (10.26)	50	86.29 (10.13)	51	
T4	87.64 (10.14)	42	86.26 (10.85)	50	
Decentering					.18
T0	33.12 (6.60)	69	33.45 (6.93)	89	
T1	38.86 (5.76)	49	38.55 (6.34)	62	
T2	38.59 (5.93)	51	38.62 (6.68)	65	
T3	39.16 (5.29)	50	38.14 (6.22)	51	
T4	39.07 (6.41)	41	38.54 (6.28)	50	
Self-compassion					.09
T0	48.49 (11.83)	69	50.61 (15.00)	89	
T1	57.39 (11.46)	49	57.81 (13.12)	62	
T2	56.75 (13.60)	51	57.37 (11.73)	64	

XSL•FO
RenderX

Outcome measure and time	Group-blended		Individual-unguided		Time × intervention condition (*P* value)
	Score, mean (SD)	Participants, n	Score, mean (SD)	Participants, n	
T3	56.77 (11.31)	48	58.00 (11.63)	51	
T4	57.49 (13.42)	41	57.08 (12.47)	49	
Well-being					.65
T0	2.89 (0.86)	69	2.87 (0.92)	90	
T1	3.39 (0.74)	49	3.10 (0.93)	63	
T2	3.25 (0.94)	51	3.12 (0.94)	65	
T3	3.26 (0.91)	50	3.09 (0.89)	52	
T4	3.25 (1.01)	42	3.04 (0.94)	51	

Group-blended eMBCT resulted in significant reductions with moderate effect size in psychological distress (Cohen *d*=0.73, 95% CI 0.14-1.32), fear of cancer recurrence (Cohen *d*=0.60, 95% CI 0.02-1.18), rumination (Cohen *d*=0.51, 95% CI 0.09-1.11), and fatigue (Cohen *d*=0.71, 95% CI 0.13-1.29). The analyses also showed improvements with large effect size in mindfulness skills (Cohen *d*=0.81, 95% CI 0.21-1.41), decentering (Cohen *d*=0.81, 95% CI 0.21-1.41), and self-compassion (Cohen *d*=0.73, 95% CI 0.10-1.36). Additionally, well-being was improved, though to a lesser extent (Cohen *d*=0.22, 95% CI 0.32-0.77).

For individual-unguided eMBCT, psychological distress (Cohen *d*=0.83, 95% CI 0.24-1.42) and rumination (Cohen *d*=0.44, 95% CI 0.16-1.04) showed improvements with large ES, while fear of cancer recurrence (Cohen *d*=0.44, 95% CI 0.18-1.07) and fatigue (Cohen *d*=0.46, 95% CI 0.15-1.07) demonstrated smaller effects. Mindfulness skills (Cohen *d*=0.69, 95% CI 0.08-1.30), decentering (Cohen *d*=0.09, 95% CI 0.46-0.65), and self-compassion (Cohen *d*=0.47, 95% CI 0.13-1.07) were also enhanced with medium to large effect sizes (Multimedia Appendix 1).

A third model, including both intervention conditions, showed that the interaction between time and intervention condition was not significant for any of the outcomes, indicating that the degree of change from baseline to follow-up was similar across the 2 delivery formats.

Prediction

Table 3 and Table S1 in Multimedia Appendix 1 show the results of the prediction and moderation analyses of psychological distress and dropout, respectively. Regarding psychological distress, several baseline variables were identified as significant predictors over time, irrespective of the eMBCT format. Higher rumination, as well as lower mindfulness skills and self-compassion at baseline, predicted a greater reduction in psychological distress from baseline to the 9-month follow-up period (Multimedia Appendix 3). Decentering and cancer diagnosis did not predict treatment outcomes over time, nor did age, sex, cancer treatment, or time since diagnosis. Regarding dropout, none of the predictors were identified as significant.

XSL•FO
RenderX

Table . Relationship between predictor or moderator, intervention condition[a], and psychological distress over time. The moderators were coded as follows: sex (0="male" and 1="female"), cancer treatment (0="palliative" and 1="curative"), and cancer diagnosis (0="breast cancer" and 1="no breast cancer").

Predictor or moderator and variables and interaction terms	Estimate	SE	95% CI	*F* test (*df*)	*P* value
Age					
Age	−0.11	0.08	−0.27 to 0.05	2.52 (1, 278)	.11
Age×time	0.04	0.02	−0.00 to 0.08	1.84 (1, 161)	.18
Age×time×intervention	−0.05	0.03	−0.11 to 0.01	3.73 (1, 161)	.07
Sex					
Sex	−2.03	2.18	−6.30 to 2.24	0.56 (1, 271)	.45
Sex×time	−0.51	0.48	−1.45 to 0.43	1.32 (1, 147)	.25
Male×time×intervention	0.24	0.59	−0.92 to 1.40	0.08 (2, 148)	.09
Cancer treatment					
Cancer treatment	−1.41	2.25	−3.00 to 5.81	1.47 (1, 281)	.23
Cancer treatment×time	0.15	0.50	−0.83 to 1.13	1.68 (1, 159)	.19
Cancer treatment×time×intervention	0.58	0.69	−0.79 to 1.95	0.72 (1, 159)	.39
Cancer diagnosis					
No breast cancer	1.56	1.79	−1.95 to 5.07	4.65 (1, 275)	.03
No breast cancer×time	−0.08	0.40	−1.58 to −0.01	2.02 (1, 155)	.16
No breast cancer×time×intervention	−0.61	0.54	−1.67 to 0.45	1.27 (1, 155)	.26
Time since diagnosis					
Time since diagnosis	0.16	0.14	−0.11 to 0.43	0.16 (1, 277)	.69
Time since diagnosis×time	0.02	0.03	−0.03 to 0.07	0.88 (1, 157)	.35
Time since diagnosis×time×intervention	0.00	0.04	−0.08 to 0.08	0.00 (1, 157)	.95
Rumination (RRQ[b])					
T0 RRQ	3.06	1.35	0.41 to 5.71	15.09 (1, 258)	<.001
T0 RRQ×time	−0.28	0.33	−0.92 to 0.36	4.17 (1, 155)	.04
T0 RRQ×time×intervention	−0.27	0.41	−1.07 to 0.53	0.42 (1, 155)	.52
Mindfulness skills (FFMQ-SF[c])					
T0 FFMQ-SF	−0.29	0.08	−0.45 to −0.13	33.31 (1, 256)	<.001
T0 FFMQ-SF×time	0.06	0.02	0.02 to 0.09	10.44 (1, 159)	.00
T0 FFMQ-SF×time×intervention	−0.04	0.03	−0.09 to −0.02	3.02 (1, 159)	.08
Decentering (EQ[d])					
T0 EQ	−0.29	0.14	−0.56 to −0.02	10.03 (1, 264)	.02
T0 EQ×time	0.05	0.03	−0.10 to 0.00	2.88 (1, 151)	.09
T0 EQ×time×intervention	−0.02	0.04	−0.09 to 0.06	0.23 (1, 151)	.63
Self-compassion (SCS-SF[e])					
T0 SCS-SF	−0.20	0.07	−0.34 to −0.06	21.75 (1, 259)	<.001
T0 SCS-SF×time	0.04	0.02	0.00 to 0.07	9.87 (1, 143)	.02
T0 SCS-SF×time×intervention	−0.02	0.02	−0.07 to 0.01	0.77 (1, 143)	.38

[a]Intervention condition was coded as follows: group-blended eMBCT=1 and individual-unguided eMBCT=0.

[b]RRQ: Rumination and Reflection Questionnaire.

[c]FFMQ-SF: Five Facet Mindfulness Questionnaire–Short Form.

XSL•FO
RenderX

[d]EQ: Experiences Questionnaire.

[e]SCS-SF: Self-Compassion Scale–Short Form.

Binary logistic regression revealed a significant model, including psychological distress and intervention condition. Participants in the group-blended eMBCT were significantly less likely to drop out than those in the individual-unguided format. However, this effect was moderated by psychological distress: within the group-blended condition, participants with higher distress were more likely to drop out compared to those with lower distress. In the individual-unguided condition, psychological distress did not influence dropout rates.

Additionally, results from the linear mixed-effects models showed that none of the 3-way interactions between time, intervention condition, and baseline variables were significant. This suggests that none of the tested baseline characteristics moderated the effect of the eMBCT format on psychological distress over the 9-month follow-up period.

Discussion

Principal Findings

This study explores the long-term effects of group-blended and individual-unguided eMBCTs for people with cancer, building on our previous research that focused on short-term (3 months) outcomes in a 3-armed RCT. We now present findings on long-term effects (up to 9 months), including predictors and moderators. As the number of cancer survivors grows, there is an increasing demand for interventions that provide sustained mental health support. Our results show that both eMBCT formats led to lasting improvements in psychological outcomes, reducing distress, fear of recurrence, rumination, and fatigue, while enhancing mindfulness, decentering, and self-compassion. These findings support eMBCT as a scalable, cost-effective approach for addressing the long-term mental health needs of cancer survivors. These findings are consistent with previous research showing that MBIs or eMBIs can have benefits for up to a year in similar populations [13,41]. In addition, beyond reducing symptoms, MBIs have also been shown to support well-being in the long term for people with cancer [6]. Combined with our findings, this evidence underscores MBIs or eMBIs as a meaningful and lasting source of support for this group of individuals.

The results from this study also suggest that individuals with higher levels of rumination or lower levels of mindfulness and self-compassion tended to show greater reductions in distress over time. While those with greater internal resources also experienced improvement, their gains were generally smaller, likely because they started from a more favorable baseline. This suggests that those with more severe initial symptoms may benefit more from eMBCT. These findings are consistent with previous research on standard (in-person) MBCT, which showed that people struggling with depression or anxiety, particularly those who ruminate heavily and have low self-compassion, often experience the most significant improvements [42,43]. Notably, the benefits observed in this research were present for both intervention conditions. This challenges the common assumption that unguided interventions are primarily suitable for individuals with milder symptoms [44].

No differential moderators of treatment effect were identified, suggesting that both eMBCT interventions appear to have similar effects across the diverse sample. However, these findings should be interpreted with caution, given the exploratory nature of the analyses and the potential limitations in statistical power.

Clinical Implications

As noted, baseline psychological vulnerability (characterized by high rumination, low mindfulness, and low self-compassion) was associated with greater reductions in distress, indicating that those most vulnerable may benefit most from eMBCT. Both group-blended and individual-unguided formats showed comparable effects in reducing psychological distress over the 9-month follow-up, regardless of demographic or clinical factors such as age, sex, cancer type, treatment, or time since diagnosis. Given this, both formats may be broadly offered, enhancing accessibility. While the choice between them can be guided by shared decision-making based on personal preferences, the individual-unguided format may be a more feasible and affordable option when resources are limited.

The dropout rate in the individual-unguided condition was 41%. This aligns with findings from a recent meta-analysis of eMBIs, which reported dropout rates of around 38.7% (95% CI 29.0-49.4) in larger trials [45]. The meta-analysis identified several factors that influenced attrition, such as monetary compensation and the enrollment method, which may partially explain the dropout rates observed in our study. Specifically, we did not offer monetary compensation, and our enrollment process did not involve in-person contact, which previous research has shown can improve retention [45]. Additionally, the higher engagement in the group-blended condition, where participants had personal contact with both the mindfulness teacher and peers, likely fostered a greater sense of accountability, contributing to lower dropout rates in this group. Incorporating more interactive or supportive elements, such as live sessions or compensation strategies, could improve retention in eMBI studies.

The dropout rate was lower in the group-blended eMBCT compared to the individual-unguided format, suggesting that group support may enhance retention. However, within the group-blended condition, participants with higher psychological distress were more likely to drop out. This pattern was not observed in the individual-unguided group. While group formats can offer structure and connection, they may feel overwhelming for individuals in acute distress [46], indicating a need for more tailored or flexible support options.

Despite the higher dropout rate in the individual-unguided format, both versions showed similar effectiveness. It is possible that those who discontinued the unguided format may have derived limited benefit even if they had completed the intervention. This raises the question of whether improving adherence alone is sufficient. Instead, enhancing subjective

XSL•FO
RenderX

engagement (how useful participants perceive the intervention to be) may be a more meaningful target for improving outcomes.

As cancer survival rates increase and recurrence remains common [1], long-term support is crucial. MBIs offer practical tools that can be merged into daily life, encouraging habits that promote lasting well-being. Both group-blended and individual-unguided eMBCTs could be implemented and offered by regional or national centers of expertise or networks of qualified mindfulness teachers specialized in psycho-oncological care. Although exploratory, these findings suggest the clinical value of both eMBCTs as scalable interventions that may help people with cancer manage symptoms and maintain psychological well-being over the long term.

Research Implications

Some important areas for future research are highlighted. First, the absence of assessments of personality traits and user preferences limits our understanding of how individual characteristics may influence engagement with and response to the intervention. For example, traits such as openness, conscientiousness, and neuroticism may impact receptivity to mindfulness practices, adherence to self-guided formats, and overall outcomes. Future studies should incorporate these factors to better tailor and optimize eMBCT for a diverse range of patients.

Second, the study design did not allow us to identify which specific components of eMBCT contributed to the observed outcomes. While the intervention included elements such as guided meditations, homework, and group discussions, their individual roles remain unclear. Further research should investigate the active components of eMBCT to refine and enhance its effectiveness for individuals with varying psychological needs. The eMBCT formats integrate a range of therapeutic elements (eg, guided meditations, reflections, homework, and group discussions), yet their individual or synergistic contributions remain unclear. Understanding the active ingredients of eMBCT is essential for refining and optimizing the intervention, particularly for individuals with specific psychological profiles or needs. Furthermore, designs such as the multiphase optimization strategy or sequential multiple assignment randomized trials could also be used to optimize intervention components, delivery sequences, and intensities for greater personalization and scalability [47].

Strengths and Limitations

This trial has several key methodological strengths. First, it included a long-term follow-up period, allowing for the assessment of sustained effects. Second, it used 2 eMBCT interventions, enabling a comparison of delivery formats. Third, the group-blended eMBCT was delivered by qualified teachers. Finally, the inclusion of a wide range of outcome measures, beyond psychological distress, provides a holistic understanding of treatment effects. This approach offers valuable insights into both symptom reduction and positive psychological growth.

This study also has several limitations that should be taken into account when interpreting the results. First, the study sample was predominantly female, with nearly half of the participants diagnosed with breast cancer, which means that our findings cannot simply be generalized to other types of cancer. Second, self-selection of participants may have introduced bias, reducing generalizability to less motivated or less digitally inclined cancer survivors. Third, although the primary RCT was adequately powered to detect overall intervention effects from pretreatment to posttreatment, it was not powered to detect small to moderate predictor or moderator effects. Consequently, the analyses were exploratory in nature and involved the examination of a relatively large set of baseline variables without correction for multiple testing; this increases the risk of type I error [48]. Therefore, these findings should be interpreted as hypothesis-generating rather than confirmatory and warrant replication in larger, adequately powered studies. Finally, long-term engagement and how participants integrated mindfulness into their lives posttreatment were not explored.

Conclusions

Both group-blended and individual-unguided eMBCTs seemed to be effective in reducing psychological distress and enhancing positive mental health outcomes over the course of the intervention and the 9-month follow-up period in people with cancer. Noticeably, individuals with higher rumination and lower mindfulness and self-compassion at baseline showed greater long-term benefits. These findings suggest that both eMBCT formats may have similar effects across the sample, though the absence of significant moderators of treatment effect should be interpreted with caution. The exploratory nature of the analysis and the limited power may have constrained our ability to detect potential moderators. Further studies with larger samples and more robust designs are needed to confirm whether both formats are equally effective for people with cancer. With cost-effectiveness studies underway, our goal is to make these accessible, flexible, and impactful eMBCTs available within the health care system, empowering people with cancer to face their trajectory with greater well-being.

Funding

This research was funded by the Dutch Cancer Society (grant 12125).

Data Availability

The data for this study may be made available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: NB, JP, SK, AS, LK

XSL•FO
RenderX

Data curation: NB
Formal analysis: NB, JL
Funding acquisition: JP, SK, AS
Investigation: NB, LK
Methodology: JP, SK, AS
Project administration: AS, LK
Supervision: JP, SK, AS, LK
Validation: NB, JL, LK
Visualization: NB
Writing – original draft: NB
Writing – review & editing: NB, JP, SK, AS, LK

Conflicts of Interest

None declared.

Multimedia Appendix 1
Statistical analysis plan, supplemental long-term effects figures, and supplemental table on the relationship between the predictors/moderators, intervention condition, and dropout.
[DOCX File, 660 KB - jmir_v28i1e79928_app1.docx]

Multimedia Appendix 2
Participant flowchart showing randomization, completion of intervention conditions, and assessments based on the primary outcome.
[DOCX File, 53 KB - jmir_v28i1e79928_app2.docx]

Multimedia Appendix 3
Supplementary graphs.
[DOCX File, 419 KB - jmir_v28i1e79928_app3.docx]

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019 Jan;69(1):7-34. [doi: 10.3322/caac.21551] [Medline: 30620402]
2. Fereidouni Z, Dehghan Abnavi S, Ghanbari Z, et al. The impact of cancer on mental health and the importance of supportive services. Galen Med J 2024;13:e3327. [doi: 10.31661/gmj.v13i.3327] [Medline: 39224547]
3. Schmidt ME, Goldschmidt S, Hermann S, Steindorf K. Late effects, long-term problems and unmet needs of cancer survivors. Int J Cancer 2022 Oct 15;151(8):1280-1290. [doi: 10.1002/ijc.34152] [Medline: 35657637]
4. Cillessen L, Johannsen M, Speckens AEM, Zachariae R. Mindfulness-based interventions for psychological and physical health outcomes in cancer patients and survivors: a systematic review and meta-analysis of randomized controlled trials. Psychooncology 2019 Dec;28(12):2257-2269. [doi: 10.1002/pon.5214] [Medline: 31464026]
5. Goldberg SB, Riordan KM, Sun S, Davidson RJ. The empirical status of mindfulness-based interventions: a systematic review of 44 meta-analyses of randomized controlled trials. Perspect Psychol Sci 2022 Jan;17(1):108-130. [doi: 10.1177/1745691620968771] [Medline: 33593124]
6. Badaghi N, Buskbjerg C, Kwakkenbos L, Bosman S, Zachariae R, Speckens A. Positive health outcomes of mindfulness-based interventions for cancer patients and survivors: a systematic review and meta-analysis. Clin Psychol Rev 2024 Dec;114:102505. [doi: 10.1016/j.cpr.2024.102505] [Medline: 39316940]
7. Kabat-Zinn J. Full Catastrophe Living (Revised Edition): Using the Wisdom of Your Body and Mind to Face Stress, Pain, and Illness: Bantam; 2013.
8. Compen FR, Bisseling EM, Schellekens MP, Jansen ET, van der Lee ML, Speckens AE. Mindfulness-based cognitive therapy for cancer patients delivered via internet: qualitative study of patient and therapist barriers and facilitators. J Med Internet Res 2017 Dec 18;19(12):e407. [doi: 10.2196/jmir.7783] [Medline: 29254912]
9. Compen F, Bisseling E, Schellekens M, et al. Face-to-face and Internet-based mindfulness-based cognitive therapy compared with treatment as usual in reducing psychological distress in patients with cancer: a multicenter randomized controlled trial. J Clin Oncol 2018 Aug 10;36(23):2413-2421. [doi: 10.1200/JCO.2017.76.5669] [Medline: 29953304]
10. Wang T, Tang C, Jiang X, Guo Y, Zhu S, Xu Q. Effectiveness of web-based mindfulness–based interventions for patients with cancer: systematic review and meta-analyses. J Med Internet Res 2024 Jun 25;26:e47704. [doi: 10.2196/47704] [Medline: 38917445]

XSL•FO
RenderX

11. Fan M, Wang Y, Zheng L, Cui M, Zhou X, Liu Z. Effectiveness of online mindfulness–based interventions for cancer patients: a systematic review and meta-analysis. Jpn J Clin Oncol 2023 Nov 5;53(11):1068-1076. [doi: 10.1093/jjco/hyad101] [Medline: 37567587]
12. Chayadi E, Baes N, Kiropoulos L. The effects of mindfulness-based interventions on symptoms of depression, anxiety, and cancer-related fatigue in oncology patients: a systematic review and meta-analysis. PLoS One 2022;17(7):e0269519. [doi: 10.1371/journal.pone.0269519] [Medline: 35834503]
13. Cillessen L, Schellekens MPJ, Van de Ven MOM, et al. Consolidation and prediction of long-term treatment effect of group and online mindfulness-based cognitive therapy for distressed cancer patients. Acta Oncol 2018 Oct;57(10):1293-1302. [doi: 10.1080/0284186X.2018.1479071] [Medline: 29932784]
14. Matis J, Svetlak M, Slezackova A, Svoboda M, Šumec R. Mindfulness-based programs for patients with cancer via ehealth and mobile health: systematic review and synthesis of quantitative research. J Med Internet Res 2020 Nov 16;22(11):e20709. [doi: 10.2196/20709] [Medline: 33196452]
15. Badaghi N, van Kruijsbergen M, Prins J, et al. Effect of blended and unguided online delivery of mindfulness-based cognitive therapy versus care as usual on distress among cancer patients and survivors: protocol for the three-arm parallel randomized controlled buddy trial. BMC Psychol 2023 Jan 25;11(1):21. [doi: 10.1186/s40359-023-01052-2] [Medline: 36698197]
16. Maloney S, Kock M, Slaghekke Y, et al. Target mechanisms of mindfulness-based programmes and practices: a scoping review. BMJ Ment Health 2024 Aug 24;27(1):e300955. [doi: 10.1136/bmjment-2023-300955] [Medline: 39181568]
17. Badaghi N, Kwakkenbos L, Prins J, Donders R, Kelders S, Speckens A. Randomized controlled trial of group-blended and individual-unguided online mindfulness–based cognitive therapy to reduce psychological distress in people with cancer. Psychooncology 2025 Sep;34(9):e70286. [doi: 10.1002/pon.70286] [Medline: 40968786]
18. Labelle LE, Campbell TS, Faris P, Carlson LE. Mediators of Mindfulness-Based Stress Reduction (MBSR): assessing the timing and sequence of change in cancer patients. J Clin Psychol 2015 Jan;71(1):21-40. [doi: 10.1002/jclp.22117] [Medline: 25099479]
19. Gecht J, Kessel R, Forkmann T, et al. A mediation model of mindfulness and decentering: sequential psychological constructs or one and the same? BMC Psychol 2014;2(1):18. [doi: 10.1186/2050-7283-2-18] [Medline: 25815189]
20. Badaghi N, van Kruijsbergen M, Speckens A, et al. Group, blended and individual, unguided online delivery of mindfulness-based cognitive therapy for people with cancer: feasibility uncontrolled trial. JMIR Form Res 2024 Feb 21;8(1):e52338. [doi: 10.2196/52338] [Medline: 38381493]
21. Hall S, Moskovitz C, Pemberton M. Understanding text recycling: a guide for researchers. : Text Recycling Research Project; 2021 URL: https://textrecycling.org/files/2021/06/Understanding-Text-Recycling_A-Guide-for-Researchers-V.1.pdf [accessed 2026-03-21]
22. Segal Z, Williams M, Teasdale J. Mindfulness-Based Cognitive Therapy for Depression: Guilford Press; 2012.
23. Good practice guidance for teachers. : UK Network of Mindfulness Teacher Training Organisations; 2015 URL: https://www.mindfulnessassociation.net/wp-content/uploads/2018/04/UK-MB-teacher-GPG-2015-final-2.pdf [accessed 2026-03-21]
24. Mitchell AJ, Meader N, Symonds P. Diagnostic validity of the Hospital Anxiety and Depression Scale (HADS) in cancer and palliative settings: a meta-analysis. J Affect Disord 2010 Nov;126(3):335-348. [doi: 10.1016/j.jad.2010.01.067] [Medline: 20207007]
25. Smith AB, Costa D, Galica J, et al. Spotlight on the Fear of Cancer Recurrence Inventory (FCRI). Psychol Res Behav Manag 2020;13:1257-1268. [doi: 10.2147/PRBM.S231577] [Medline: 33376421]
26. Luyckx K, Schwartz SJ, Berzonsky MD, et al. Capturing ruminative exploration: extending the four-dimensional model of identity formation in late adolescence. J Res Pers 2008 Feb;42(1):58-82. [doi: 10.1016/j.jrp.2007.04.004]
27. Worm-Smeitink M, Gielissen M, Bloot L, et al. The assessment of fatigue: psychometric qualities and norms for the Checklist Individual Strength. J Psychosom Res 2017 Jul;98:40-46. [doi: 10.1016/j.jpsychores.2017.05.007] [Medline: 28554371]
28. Bohlmeijer E, ten Klooster PM, Fledderus M, Veehof M, Baer R. Psychometric properties of the Five Facet Mindfulness Questionnaire in depressed adults and development of a short form. Assessment 2011 Sep;18(3):308-320. [doi: 10.1177/1073191111408231] [Medline: 21586480]
29. Fresco DM, Moore MT, van Dulmen MHM, et al. Initial psychometric properties of the Experiences Questionnaire: validation of a self-report measure of decentering. Behav Ther 2007 Sep;38(3):234-246. [doi: 10.1016/j.beth.2006.08.003] [Medline: 17697849]
30. Raes F, Pommier E, Neff KD, Van Gucht D. Construction and factorial validation of a short form of the Self-Compassion Scale. Clin Psychol Psychother 2011;18(3):250-255. [doi: 10.1002/cpp.702] [Medline: 21584907]
31. Lamers SMA, Westerhof GJ, Bohlmeijer ET, ten Klooster PM, Keyes CLM. Evaluating the psychometric properties of the Mental Health Continuum–Short Form (MHC-SF). J Clin Psychol 2011 Jan;67(1):99-110. [doi: 10.1002/jclp.20741] [Medline: 20973032]
32. Newman DA. Missing data: five practical guidelines. Organ Res Methods 2014 Oct;17(4):372-411. [doi: 10.1177/1094428114548590]
33. Cohen J. Statistical Power Analysis for the Behavioral Sciences: Routledge; 2013.

XSL•FO
RenderX

34. West BT. Analyzing longitudinal data with the linear mixed models procedure in SPSS. Eval Health Prof 2009 Sep;32(3):207-228. [doi: 10.1177/0163278709338554] [Medline: 19679634]
35. Nick TG, Campbell KM. Logistic regression. In: Ambrosius WT, editor. Topics in Biostatistics: Humana Press; 2007:273-301.
36. Kuyken W, Byford S, Taylor RS, et al. Mindfulness-based cognitive therapy to prevent relapse in recurrent depression. J Consult Clin Psychol 2008 Dec;76(6):966-978. [doi: 10.1037/a0013786] [Medline: 19045965]
37. Ma SH, Teasdale JD. Mindfulness-based cognitive therapy for depression: replication and exploration of differential relapse prevention effects. J Consult Clin Psychol 2004 Feb;72(1):31-40. [doi: 10.1037/0022-006X.72.1.31] [Medline: 14756612]
38. Teasdale JD, Segal ZV, Williams JM, Ridgeway VA, Soulsby JM, Lau MA. Prevention of relapse/recurrence in major depression by mindfulness-based cognitive therapy. J Consult Clin Psychol 2000 Aug;68(4):615-623. [doi: 10.1037//0022-006x.68.4.615] [Medline: 10965637]
39. Kuyken W, Watkins E, Holden E, et al. How does mindfulness-based cognitive therapy work? Behav Res Ther 2010 Nov;48(11):1105-1112. [doi: 10.1016/j.brat.2010.08.003] [Medline: 20810101]
40. Lorah JA. Interpretation and visualization of moderation effects and random slopes in multilevel models. Quant Methods Psychol 2022 Mar 1;18(1):111-127. [doi: 10.20982/tqmp.18.1.p111]
41. Carlson LE, Speca M, Faris P, Patel KD. One year pre-post intervention follow-up of psychological, immune, endocrine and blood pressure outcomes of mindfulness-based stress reduction (MBSR) in breast and prostate cancer outpatients. Brain Behav Immun 2007 Nov;21(8):1038-1049. [doi: 10.1016/j.bbi.2007.04.002] [Medline: 17521871]
42. Lubbers J, Geurts DEM, Spinhoven P, et al. Rumination and self-compassion moderate mindfulness-based cognitive therapy for patients with recurrent and persistent major depressive disorder: a controlled trial. Depress Anxiety 2024;2024(1):3511703. [doi: 10.1155/da/3511703] [Medline: 40226644]
43. Frostadottir AD, Dorjee D. Effects of mindfulness based cognitive therapy (MBCT) and compassion focused therapy (CFT) on symptom change, mindfulness, self-compassion, and rumination in clients with depression, anxiety, and stress. Front Psychol 2019;10:1099. [doi: 10.3389/fpsyg.2019.01099] [Medline: 31164849]
44. Johansson R, Andersson G. Internet-based psychological treatments for depression. Expert Rev Neurother 2012 Jul;12(7):861-869. [doi: 10.1586/ern.12.63] [Medline: 22853793]
45. Linardon J. Rates of attrition and engagement in randomized controlled trials of mindfulness apps: systematic review and meta-analysis. Behav Res Ther 2023 Nov;170:104421. [doi: 10.1016/j.brat.2023.104421] [Medline: 37862854]
46. Tang WZ, Mangantig E, Iskandar YHP, Cheng SL, Yusuf A, Jia K. Prevalence and associated factors of psychological distress among patients with breast cancer: a systematic review and meta-analysis. BMJ Open 2024 Sep 26;14(9):e077067. [doi: 10.1136/bmjopen-2023-077067] [Medline: 39327054]
47. Collins LM, Murphy SA, Strecher V. The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): new methods for more potent eHealth interventions. Am J Prev Med 2007 May;32(5 Suppl):S112-S118. [doi: 10.1016/j.amepre.2007.01.022] [Medline: 17466815]
48. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol 1995 Jan 1;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]

Abbreviations

CAU: care as usual
eMBCT: online mindfulness–based cognitive therapy
eMBI: online mindfulness–based intervention
EQ: Experiences Questionnaire
FFMQ: Five Facet Mindfulness Questionnaire–Short Form
MBCT: mindfulness-based cognitive therapy
MBI: mindfulness-based intervention
RCT: randomized controlled trial
RRQ: Rumination and Reflection Questionnaire
SCS-SF: Self-Compassion Scale–Short Form

Edited by A Schwartz, M Balcarras; submitted 01.Jul.2025; peer-reviewed by N Zhu, SF Fung; revised version received 18.Feb.2026; accepted 23.Feb.2026; published 17.Apr.2026.

Please cite as:
Badaghi N, Lubbers J, Prins J, Kelders S, Speckens A, Kwakkenbos L
Long-Term Outcomes, Moderators, and Predictors in Online Mindfulness–Based Cognitive Therapy for People With Cancer: Secondary Analysis of a Randomized Controlled Trial
J Med Internet Res 2026;28:e79928
URL: https://www.jmir.org/2026/1/e79928
doi:10.2196/79928

XSL•FO
RenderX

© Nasim Badaghi, Jelle Lubbers, Judith Prins, Saskia Kelders, Anne Speckens, Linda Kwakkenbos. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Stimulating Future-Oriented Thinking and Goal-Achievement Through the Future Self Using Virtual Reality and a Smartphone App: Randomized Controlled Trial

Jean-Louis van Gelder[1,2], PhD; Tiffany Tettero[1,2], MSc; Aniek Siezenga[1,2], MSc; Esther Mertens[1,3], PhD

[1]Department of Criminology, Max Planck Institute for the Study of Crime, Security and Law, Günterstalstraße 73, Freiburg, Germany

[2]Institute of Education and Child Studies, Leiden University, Leiden, The Netherlands

[3]Netherlands Institute for the Study of Crime and Law Enforcement, Amsterdam, The Netherlands

Corresponding Author:
Jean-Louis van Gelder, PhD
Department of Criminology, Max Planck Institute for the Study of Crime, Security and Law, Günterstalstraße 73, Freiburg, Germany

Abstract

Background: Research suggests that identification with one's future self—encompassing vividness, connectedness, and valence—plays a key role in motivating future-oriented choices and goal pursuit. Interventions aiming to strengthen future self-identification have been shown to reduce maladaptive behaviors and promote well-being, but traditional approaches often rely heavily on imagination. Emerging technologies offer novel opportunities to make the future self more vivid and tangible, potentially reducing cognitive burden, and enhancing intervention effectiveness.

Objective: This randomized controlled trial evaluated the effectiveness of a digital intervention on the following primary outcomes: future self-identification, future orientation, consideration of future consequences, self-defeating behavior, goal commitment, and goal achievement.

Methods: In this parallel, 3-arm randomized controlled trial, 321 first-year students from a large public university in the Netherlands were randomized (1:1:1) with blocks of 9 to a smartphone app–based future-self intervention, an immersive virtual reality (VR) version of the same intervention, or an active goal-setting control condition. Participants assigned to the smartphone or VR condition engaged with the intervention over a 3-week period, during which they interacted with their 10-year-old self, their future self. Participants in the control condition received no further support. Each of the 3 conditions consisted of 107 participants, who were all included in the analyses following the intention-to-treat principle. Due to the nature of the intervention, blinding was not possible.

Results: Compared to the goal-setting control condition, both intervention conditions yielded significant short-term improvements on all 3 aspects of future self-identification (smartphone: $d_{vividness}$=0.49; $d_{valence}$=0.44; $d_{connectedness}$=0.43; VR: $d_{vividness}$=0.35; $d_{valence}$=0.44; $d_{connectedness}$=0.43), and a small decline in vividness (smartphone: d=−0.36; VR: d=−0.23) and connectedness (smartphone: d=−0.36; VR: d=−0.32) at 6-month follow-up. Additionally, the intervention buffered declines in future orientation during the study period (smartphone: d=.16; VR: d=0.18), and VR delivery led to significantly higher weekly goal achievement (d=0.88). However, effects on future orientation were not sustained over time (smartphone: d=−0.18; VR: d=−0.25), resulting in similar levels across conditions at 6-month follow-up. Furthermore, no significant effects emerged for other primary or secondary outcomes, such as self-defeating behavior (smartphone: $d_{overall}$=−0.10; VR: $d_{overall}$=−0.03), impulsivity (smartphone: $d_{overall}$=−0.07; VR: $d_{overall}$=−0.09), or academic performance ($\eta^2_{partial}$=0.00). No adverse events were reported in any of the conditions.

Conclusions: This randomized controlled trial innovatively compares smartphone and immersive VR future-self interventions. Findings suggest that digital interventions leveraging visual and interactive representations of the future self can strengthen future self-identification and future orientation, and support (short-term) goal pursuit. Together, they highlight the potential of scalable digital future self approaches for educational and preventive health contexts.

Trial Registration: ClinicalTrials.gov NCT05578755; https://clinicaltrials.gov/study/NCT05578755

(J Med Internet Res 2026;28:e84420) doi:10.2196/84420

KEYWORDS

future self-identification; future orientation; future self-continuity; intervention; mental time travel; goal achievement; virtual reality; smartphone app

XSL•FO
RenderX

Introduction

Background

Considering the future is a key component of psychosocial functioning and a ubiquitous part of mental life [1,2]. Individuals who are more future-oriented tend to make more informed trade-offs between the short-term and long-term consequences of their decisions [3] and are more likely to set goals they aspire to achieve over time [4]. This orientation is also linked to various beneficial outcomes, including heightened feelings of competence, better health, increased savings, and stronger academic performance [4,5]. In contrast, individuals who are more present-focused more often act on impulse and display a preference for immediate gratification [6,7]. Such a focus has been associated with behaviors that offer short-term benefits, but often entail severe long-term costs—that is, are self-defeating—such as delinquency, gambling, and substance use [6,8]. Thus, enhancing future-oriented thinking could contribute to psychosocial functioning in important ways. Here, we examine the effectiveness of an intervention aiming to increase future-oriented thinking and goal achievement by strengthening identification with the future self, using a smartphone app and immersive virtual reality (VR) as delivery technologies.

A Future-Oriented Mindset and the Future Self

Research leveraging the future self to stimulate future-oriented behavior and successful goal pursuit is premised on the idea that the degree to which individuals consider temporally distant outcomes is related to how strongly they identify with, and feel connected to, their future self [9]. Tendencies to disregard or downplay possible long-term consequences of current actions and decisions are believed to result from an inability to project one's self into the future [10]. Rather than emphasizing how current behaviors may carry (negative) future consequences in a more generalized way, which is unlikely to resonate with those who have a strong here-and-now orientation, this approach emphasizes the idea of a shared fate between the present and the future self [11]. Just as people experience varying degrees of closeness to others—and these differences affect how willing we are to make sacrifices on their behalf—varying degrees of identification with our future selves can shape to what extent we act in ways that benefit those future selves.

Experiencing a stronger sense of identification with one's future self is assumed to contribute to a greater awareness that the possible costs of current actions will ultimately be borne by oneself, even if they materialize only at a later point in time, and also makes the burden of such costs more salient [10]. Conversely, it also promotes the realization that the potential benefits of current effort and sacrifice will get to be enjoyed by oneself, even if such benefits are reaped only in the future. Thus, stronger identification with one's future self should make it more likely that the interests of this self are factored into decisions made in the present [12].

In support of these assumptions, empirical studies have shown that enhancing identification with the future self promotes positive behaviors, such as increased physical activity [9], addiction recovery [13], and savings behavior [12,14], while reducing negative behaviors, such as delinquency [15,16], cheating [17], temporal discounting [18], and procrastination [19]. For example, in a study using a large nationally representative sample in the United States, Hershfield et al [14] found that individuals with a higher perceived connection to their future self demonstrated better financial well-being and more robust saving behavior. A study among convicted offenders showed that exposure to their future self using a VR paradigm resulted in a decrease in self-defeating behavior [16].

Furthermore, identification with the future self has recently been found to be a protective factor in mental health. Research involving adolescents has shown that a lack of possible selves or negative valence toward the future self is a key characteristic of depressive symptoms [17]. Other work found that people struggling with depression or suicidal ideation often describe their future selves as distant, dissimilar, and emotionally disconnected—a pattern reflecting low future self-identification [18,19]. Strengthening identification with the future self could therefore have cascading effects in different domains. Recent work has started leveraging the future self as an intervention mechanism for improving people's ability to think about the future and to reduce symptoms of depression and suicidal ideation [20]. In short, interventions that target the future self may offer transdiagnostic benefits.

Future Self-Identification as Vividness, Connectedness, and Valence

Future self-identification (also referred to as future self-continuity in the research literature) is often conceived as a tripartite structure consisting of the extent to which individuals can clearly and concretely imagine and describe their self in the future (vividness), how similar and connected they feel to it (connectedness), and the positive or negative affect they associate with it (valence) [21,22]. Future self vividness and connectedness, in particular, have been shown to predict intertemporal choice; however, they are believed to operate through different mechanisms.

Vividness is related to psychological distance, which shapes how events are mentally represented. Events anticipated in the distant future tend to be construed in more abstract and decontextualized terms and with less detail, making them less likely to prompt action [23,24]. Events perceived as temporally near, in contrast, are represented more concretely and are more likely to motivate behavior. One important mechanism accounting for this is that more concrete representations enhance the perceived likelihood and immediacy of those events; the ability to vividly imagine specific, personal future scenarios can make distant goals feel more immediate and is thus more likely to prompt action [2,25]. Furthermore, vivid mental representations also intensify the emotional salience of potential outcomes, which can influence behavior over and above the subjective likelihoods of those outcomes [6,26]. Thus, enhancing the vividness of the future self reduces the psychological distance to the current self and increases the emotional salience of remote outcomes and thereby motivates goal-directed action.

Future self connectedness denotes the degree of overlap between the current and future selves one experiences in terms of factors such as shared goals, values, identity, and personality

XSL•FO
RenderX

characteristics [22,27]. People experiencing high levels of connectedness are more likely to view their future self as an extension of their current identity. In contrast, people who experience only a weak sense of connection with their future self may regard it as distinct and unfamiliar, akin to perceiving a stranger [12,28]. The sense of connection people experience with their future self is consequential, as it affects whether people will act in its best interest (vs acting only with the well-being of their present self in mind). That is, similar to how individuals experience varying degrees of psychological connectedness to others—which in turn influences their willingness to make sacrifices on others' behalf—variability in one's sense of connection to the future self can predict the extent to which they engage in behavior that serves the interests of their future selves [10].

Valence of the future self regards the positive and negative affect people associate with it. Although it stands to reason that people with more positive views of their future self are likely to invest more in their future, and positive evaluations of one's future self have also been empirically linked to current self-esteem, mental health, and overall well-being, the empirical relationship between future-self valence and future-oriented behaviors is not well established [10]. Research on episodic future thinking also shows mixed findings regarding the relation between valence and temporal discounting, an indicator of short-term thinking [29].

Prior research has tended to either use an index measure of future self-identification that includes all three aspects—vividness, connectedness, and valence—or has focused on only a single aspect, leaving out the other two. Index measures are unable to identify which of the three aspects is driving the effect on outcomes and may mask meaningful relations between specific aspects and outcomes. Research focusing on only one aspect leaves the possible effects of the other two unaccounted for. In this study, we include all three aspects of future self-identification and examine the effects of the intervention on each of them separately.

Modes of Delivery: Smartphone App Versus VR

Traditional interventions to promote future-oriented thinking and behavior often rely heavily on people's imaginative abilities and use exercises that can carry a significant cognitive burden, such as asking people to imagine a possible future event [30] or to write a letter to their future self [31]. Visual and immersive technologies, such as smartphone apps and VR, provide not only visual support through graphic renderings of a future self but also allow for a sense of interaction with, and, in the case of VR, the possibility of embodiment of, the future self. Hence, such interventions can reduce cognitive load and the dependency on people's imaginative abilities.

Aside from their shared ability to provide interactive visual content, as delivery methods, apps and VR also have several distinct characteristics, and they differ in the way users can interact with them [32]. Smartphones tend to be integrated in people's daily lives, which provides the opportunity to deliver an intervention in their natural environment through frequent exposure and interaction [33,34]. Users can access the intervention content wherever and whenever they want, as they carry the intervention with them "in their back pocket."

VR, in contrast, offers the opportunity to provide participants with highly immersive experiences tailored to the goals of an intervention [32]. For example, VR allows users to embody avatars that differ meaningfully from their own physical or identity characteristics in perceptually convincing ways [35]. Such alterations in avatar design can unconsciously influence users' perceptions or behaviors, even when they remain consciously aware of the artificial nature of the virtual setting, and thereby change attitudes and behaviors [36,37]. To capitalize on this affordance, in this study, participants embody avatar versions of their future (and current) selves in a virtual environment.

This Study

The goal of this study was to evaluate the effectiveness of an intervention developed in the context of the FutureU research program. The intervention aims to foster future-oriented thinking and future self-identification through three broad strategies: (1) prompting participants to imagine their future and desired goals, (2) encouraging reflection on their future self, and (3) enabling interaction with digital renderings of this temporally remote self. These renderings are intended to instill a more vivid image of the future self, to encourage participants to think about their future self in daily life, and to stimulate a sense of connection to it. Through various assignments, participants are exposed to, and interact with, a digital version of their future self 10 years from now (see Table 1 for a description of the different intervention modules).

XSL•FO
RenderX

Table . Description and features of the 3 intervention modules (future self-identification, future self-perspective, and goal setting and achievement), including underlying theory and core features implemented across conditions[a].

Module	Theory	Core features smartphone app	Core features VR[b] sessions
Future self-identification			
Stimulating vividness, familiarity, and identification with the future self	• Exposure to and vividness of the future self increases future orientation [38]. Additionally in smartphone intervention • Incremental personality theory: The belief that personality can change over time can reduce problematic behaviors [39]. • People's willingness to change on personality traits in socially desirable ways increases after feedback on their current trait levels [40].	• Complete personal profile of the future self (eg, work experience, skills, and accomplishments). • Current scores on personality traits with an indication of norm scores. • Short animation with psychoeducation that personality can change over time. • Set scores of personality traits of the future self.	• Time travel portal facilitating mental time traveling in order to "prelive" events. • Embodiment of the avatar representing future self, bolstering vividness of and identification with the future self. • Interview future self about personal profile (eg, work experience, skills, and accomplishments). • A grid containing participants' answers showing a personal profile of the future self.
Future self-perspective			
Cultivate future-oriented choices and increase self-insight by distanced perspective taking, aiming to stimulate attitudes and behaviors favoring the future self	• People make more future-oriented choices (1) for others (ie, Solomon's paradox) [41], (2) when they have a vivid perception of the future self [38], and (3) when they can psychologically or temporally distance themselves from the situation (ie, construal level theory) [23]. • Wise reasoning is enhanced with third-person self-reflection [42].	• Short animation clip with psychoeducation that people make more future-oriented choices when they distance themselves from the situation, and when they think about the long-term consequences. • Time portal to take a future self-perspective for giving advice. • Participants address themselves in the third person. • Future self-interaction portal: The future self emphasizes that personality can change over time and stimulates decision-making with the future self in mind.	• Verbal psychoeducation that people make more future-oriented choices when they distance themselves from the situation and when they think about the long-term consequences. • Ask the future self for advice on a challenge in the study domain and on a freely chosen domain. Switching perspectives is used to mimic a conversation and facilitates clarification of the challenge or problem and giving advice. • A grid containing participant's answers showing the posed challenges and the future self's advice.
Goal setting and achievement			
Bolster goal setting and achievement by teaching a growth mindset and mental contrasting and implementation intentions	• Growth mindset: The belief that people's abilities can develop over time. This mindset aids engagement in thoughts and behaviors to work toward goals [43]. • Mental contrasting and implementation intentions [44]: A method in which the desired future is contrasted with the current reality and then reflected upon obstacles in the way of attaining the desired future. Subsequently, a plan is formulated to implement behaviors to overcome obstacles, that is, implementation intentions, in the format: If situation X, then I will do Y.	• Short animation clip with psychoeducation that abilities can develop over time. • Short animation clip explaining mental contrasting and implementation intentions. • Practice with mental contrasting and implementation intentions to work toward goals via filling in a scheme. • Writing a letter to the future self with goals. • Future self-interaction portal: The future self stimulates taking the perspective of the future self (in decision-making and in goal achievement) and provides a guided episodic future thinking exercise.	• Verbal psychoeducation that abilities can develop over time. • Practice mental contrasting and implementation intentions via interviewing the future self. • A grid containing participants' answers providing a scheme of goals, obstacles, and plans to overcome the obstacles.

[a]Reprinted from the protocol paper of the research project [45].

[b]VR: virtual reality.

XSL•FO
RenderX

We examined intervention effects on future orientation, goal achievement, and self-defeating behaviors, comparing the smartphone and VR experimental conditions to an active, goal-setting control condition. Specifically, we assessed proximal effects on the 3 aspects of future self-identification (ie, vividness, connectedness, and valence), as well as distal intervention effects on primary outcomes (ie, future orientation, consideration of future consequences, self-defeating behavior, goal commitment, and goal achievement) and secondary outcomes (ie, self-efficacy, academic performance, and impulsivity). In case of significant intervention effects, we compared the 2 delivery methods (ie, app vs VR) with each other to establish whether one generated stronger effects than the other. We hypothesized that the intervention would enhance future self-identification and result in higher scores on the primary outcomes. No specific hypotheses regarding the most effective delivery method were formulated.

Methods

Trial Design and Procedure

The intervention was examined by means of a parallel randomized controlled trial, with three conditions: (1) a smartphone condition in which participants set goals and received the intervention via the app (an iterated version of the app examined in Mertens et al [46]), (2) a VR condition in which participants set goals and received the intervention via immersive VR, and (3) an active control condition in which participants set goals—as in the other conditions—but received no further intervention.

In all 3 conditions, participants started out with an intake session at the faculty's research laboratory. During the intake, participants set personal goals and completed a baseline questionnaire. Participants completed online questionnaires during intake (T1 or baseline), at weekly intervals during the intervention (T2 and T3), immediately after the intervention (T4), and 3 and 6 months after the end of the intervention (T5 and T6), or at parallel time points in the control condition. Questionnaires not completed in time (ie, within 4 days for T2 and T3, 8 days for T4, 16 days for T5, and 32 days for T6) were treated as missing data. Data were collected between October 2022 and January 2024.

Avatar Creation

In both the smartphone and the VR conditions, avatars representing participants' 10-year older self, that is, the future self, were created with multiple plug-in services and software specifically developed for the research project (all custom software development, design, and illustrations by Orb Amsterdam and Studio Barbaar). The avatar creation process differed slightly between conditions. In the smartphone condition, participants took a "selfie" using the integrated camera of their own smartphone when opening the FutureU app for the first time. This photo was age-progressed by 10 years using a custom-made server and the online service of ChangeMyFace. Subsequently, the aged image was converted into a 3D digital representation via software developed by Avatar SDK, version "Head 2.0."

In the VR condition, a (full body) avatar was created at the start of the first VR session (T2). This was a 2-step procedure. First, a photo of the participant's face was made using a webcam that was connected to the computer running the VR simulation. The rest of the procedure for creating the avatar's head was identical to the smartphone condition. Second, the avatar's body was created using custom-made software. The software allows for adjusting the proportions of a generic male or female virtual body using a set of sliders to match the participant's actual body and skin color. The upper-body clothing of the present self-avatar and of the future self-avatar differed in color to emphasize the difference between the 2.

Participants and Recruitment

Research suggests that the most effective time to intervene is during transformational events and contextual shifts [47]. Such transitions typically involve changes in environment, routines, and social networks. Therefore, we evaluated the effectiveness of the intervention among first-year university students—a group undergoing transition from secondary school to university. This period often involves impactful changes, such as moving away from home, relocating to a new city, and adopting new roles, making it an opportune moment for intervention.

Participants were 321 first-year university students in the Netherlands who were predominantly enrolled in pedagogical sciences (n=187, 58%) and psychology (n=133, 41%). Most participants were female (n=287, 89%) with an average age of 19.59 (SD 2.25) years. Each of the 3 conditions consisted of 107 participants. There were no differences between the conditions regarding sex distribution (χ^2_2=1.1; P=.59; ϕ=.057; female: smartphone: n=97, VR: n=93, control: n=97) and age ($F_{2,318}$=0.55; P=.58; $\eta^2_{partial}$=0.003; smartphone: mean 19.40, SD 2.46; VR: mean 19.70, SD 2.02; control: mean 19.65, SD 2.24).

Randomization and Allocation Concealment

Participants were randomly assigned to the conditions on a 1:1:1 ratio with blocks of 9. The random sequence was generated with an online tool by the project manager. Participants signed up via the university's online portal for study participation and were assigned to a condition in the order of enrollment by the project manager in accordance with the generated random sequence. Allocation was known to the researcher at the start of the intake, given that the intake procedure slightly differed between conditions (ie, installing the app and scheduling the VR sessions). Participants were unaware of the different conditions, the intervention's content, and the study hypotheses.

Intervention and Comparator

Overview

Participants started with setting personal goals; one for the coming year and another one for the coming month. Additionally, they set a goal for the upcoming week as a first step toward reaching their monthly goal. The researcher supported the formulation of these goals and ensured that goals set were specific, measurable, and challenging (though attainable) following the SMART-goal model and Zimmerman's criteria [48]. In the smartphone and control conditions,

XSL•FO
RenderX

participants independently set weekly goals during each week of the intervention period to attain their monthly goal. In the VR condition, participants received additional guidance for setting the weekly goals from the experimenter at the start of each VR session. After setting their goals, participants in the control condition received no further support to achieve their goals. Participants in the intervention conditions received the intervention.

Intervention Conditions

After setting their goals, participants received 3 consecutive intervention modules. The first module primarily focused on increasing vividness of the future self, the second module aimed to encourage future-oriented decision-making, and the third module was directed at goal achievement. The theoretical foundation of the 3 modules is presented in Table 1 (see Mertens et al [45] for more information). During development, these modules went through extensive user tests involving both experts and the target population, and a pilot randomized controlled trial [46], to gather qualitative and quantitative feedback for improvements.

Smartphone Condition

In the app (see Figure 1 for screenshots; reprinted from protocol paper of the current research project [45]), participants were asked to interact with the app on a daily basis for a period of 3 weeks (ie, 21 days) for approximately 5 minutes a day. As a reminder to interact with the app, participants received push notifications. When opening the app, they were directed to the chat feature where they interacted with a chatbot (Figure 1A). This chatbot provided psychoeducation, asked targeted questions, and gave instructions for the interaction or assignment of that day.

Figure 1. Screenshots of the FutureU smartphone app: (A) chat, (**B**) the connection mechanic, (**C**) future self-interaction, (**D**) home screen, (**E**) personal profile, (**F**) personality menu, and (**G**) goal scheme.

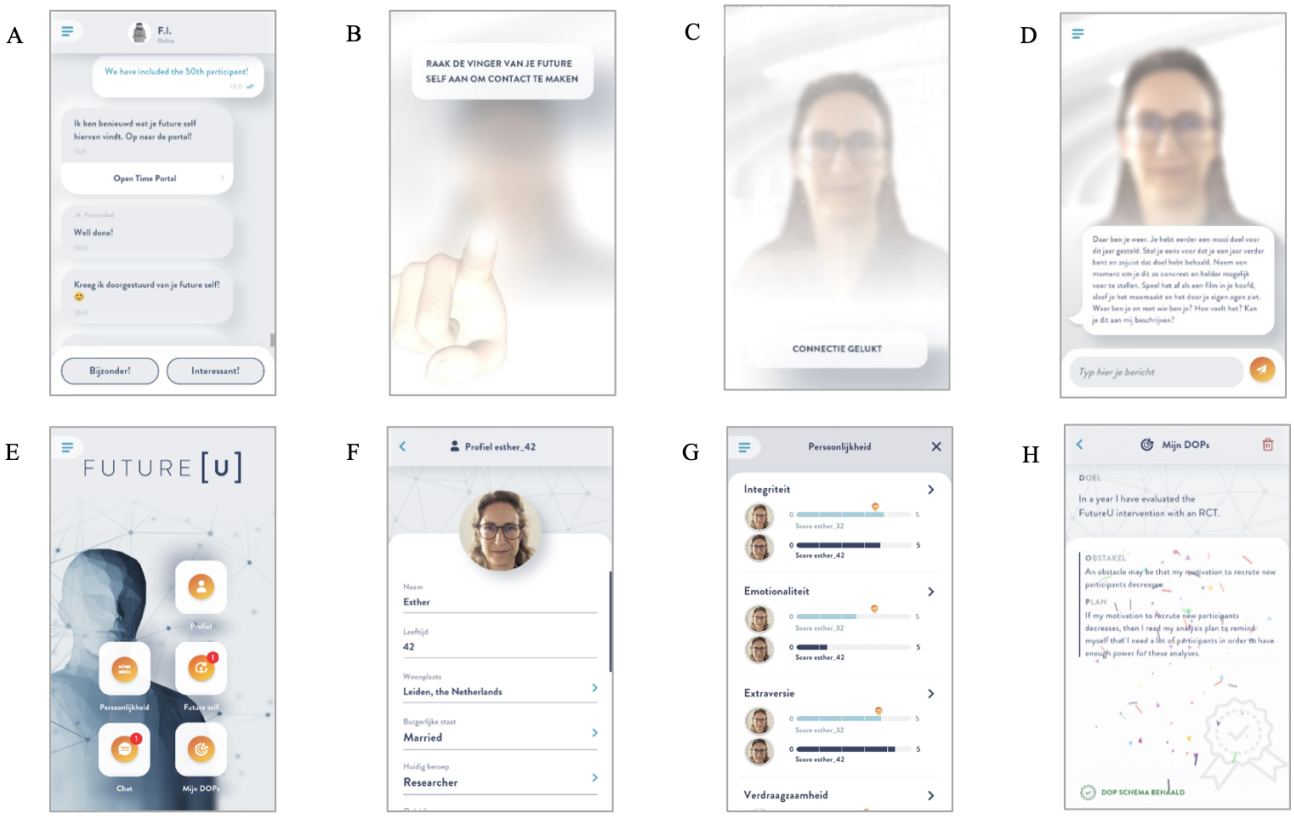

Roughly every other day, the app sent an additional push notification from the future self. This notification directed participants to a "future self-interaction" feature. After "connecting" with their future self by touching the virtual finger of their future self (Figure 1B), the screen unblurred, and an interaction with their future self-avatar started (Figure 1C). The consecutive future self-interactions aimed to instill a more vivid image of the future self, encourage participants to think about their future self in their daily life, and stimulate a sense of connection with the future self.

VR Condition

The VR intervention consisted of 3 approximately 30-minute sessions scheduled 1 week apart from each other. The virtual environment consisted of a room with a table and a time machine allowing for "time travel" (Figure 2). Sitting on opposite sides of the table were 2 avatars, one representing the present self, the other representing the future self. The interaction was guided by a researcher-controlled, hovering robot (Figure 2; reprint from protocol paper [45]). This robot provided psychoeducation and explained the controls and interactions to the participants.

During the VR sessions, participants alternated between embodying their present self- and their future self-avatar. While embodying the present self, there were multiple cards with questions lying in front of the participant. They were instructed to read the questions out loud one by one. This process was recorded and, after they finished reading them, they pulled a

XSL•FO
RenderX

virtual lever to activate the time machine in order to "time travel" to the future. Following the time travel, they embodied their future self-avatar seated on the other side of the table. The recorded questions were then played back one by one and the participant was invited to respond to them. Again, responses were recorded and, after traveling back to the present, played to the participant who once again was now embodied as their present self. In each session, the first 2 rounds of interaction between the present and future self-avatars were structured using cards. In the third round, participants were free to ask their own questions to their future self. At the end of each session, participants reflected on their answers, which were shown on a floating grid completed by the researcher during the session (The virtual environment was developed with the Unity Pro engine. An HTC Vive head-mounted display with a 110° field-of-view and a resolution of 1080×1200 pixels per eye displayed at 90 Hz was used to deliver the experience. Participants used 2 controllers to move the arms and torso of their avatar.).

Figure 2. General overview of the FutureU VR environment. 1: robot FI controlled by researcher; 2: present self-avatar; 3: future self-avatar; 4: cards with questions; 5: recording light; 6: handle to travel through time; 7: time machine; VR: virtual reality.

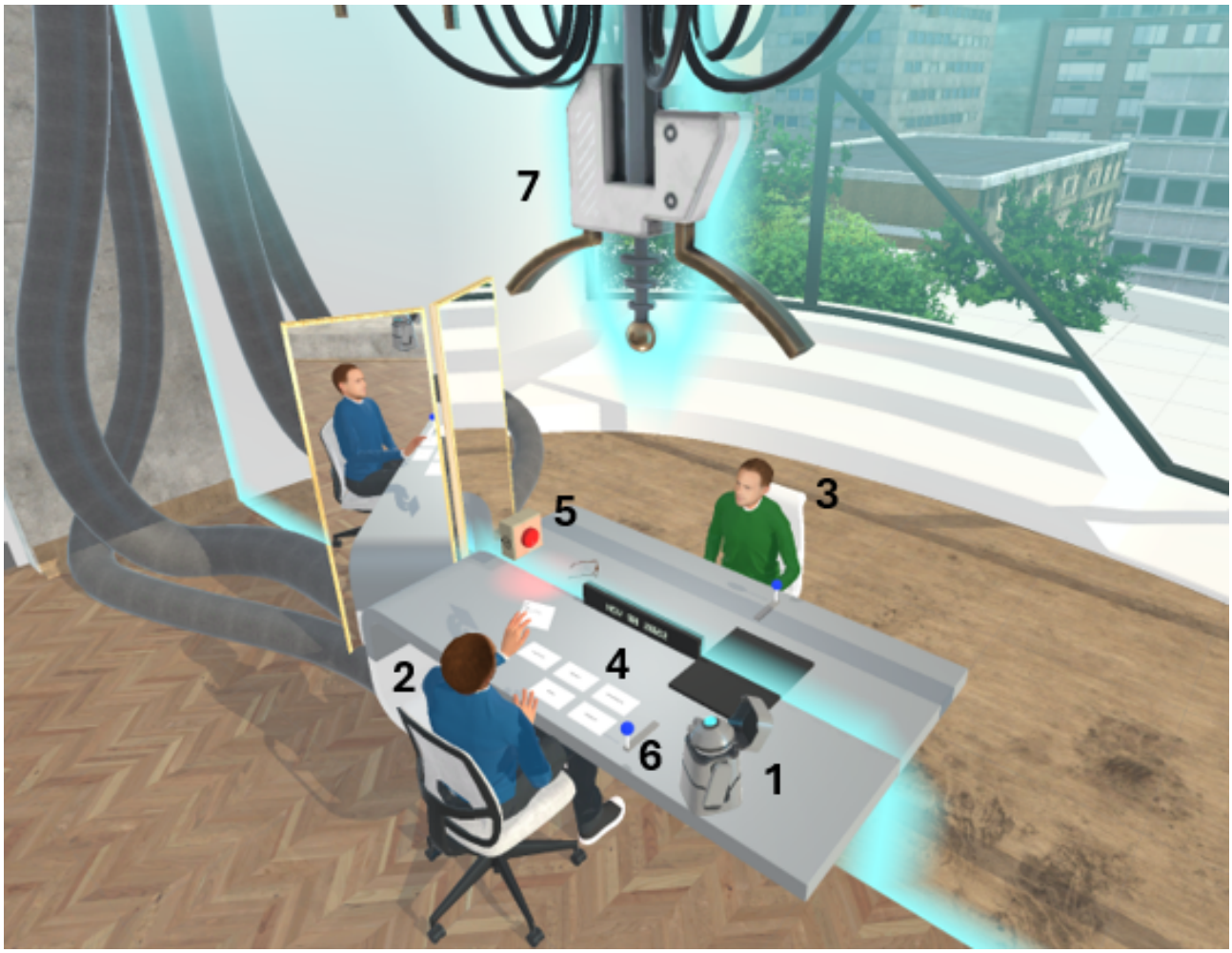

Outcomes

Proximal Outcomes

Future self-identification, ie, the extent to which people identify with their future self, was assessed with 3 subscales: vividness, valence, and connectedness. These were assessed at each time point.

Vividness of the future self was assessed with 5 items (eg, "I have a clear image of myself in 10 years from now.") on a 7-point Likert scale (1=completely disagree to 7=completely agree; T1-T6 α=.91-.93) based on Van Gelder et al [10].

Valence of the future self was measured with a single item: "How do you feel when you think about your future?" [49], answered with the 9-point Self-Assessment Manikin (1=negative to 9=positive).

Connectedness to the future self was measured with the 2-item Future Self-Continuity Measure [49]. These items each consist of a set of 7 increasingly overlapping circles representing the present and the future self. Reliability was adequate at T2 through T6 (α=.70-.79), but somewhat low at T1 (α=.54).

Distal Outcomes

Primary Outcomes

Future orientation is assessed with the Future Orientation Scale [3] and the Consideration of Future Consequences scale [50]. This scale contains 3 subscales, namely time perspective, anticipation of future consequences, and planning ahead. The

XSL•FO
RenderX

scale contains 15 items that each consist of a present-oriented and a future-oriented statement (eg, “Some people spend very little time thinking about how things might be in the future, but other people spend a lot of time thinking about how things might be in the future.”). When choosing the present-oriented statement, participants score a 1=completely true or 2=a little bit true. When choosing the future-oriented statement, participants score a 3=a little bit true or a 4=completely true. This measurement was assessed at all time points (T1-T6: α=.74-.87). At T2 and T3, a selection of 6 items based on factor loadings and face validity were used.

Consideration of future consequences, that is, the degree to which people take immediate versus distant consequences into account in their potential behaviors, was measured by the Consideration of Future Consequences scale by Strathman et al [50]. The scale contained 9 items (eg, “I consider how things might be in the future.”) answered on a 5-point Likert scale (1=completely disagree to 5=completely agree). It was assessed at T1 and T4-T6 (α=.79-.82).

Self-defeating behavior concerns behaviors with immediate gains but potential long-term costs. These were assessed with 15 items (based on Van Gelder et al [10]) assessing self-defeating behaviors (eg, “How often in the last week have you missed classes or work?”) on a 5-point scale (1=never to 5=more than 10 times). Prior research has indicated higher reliability and validity and lower skewness of dichotomized variety scales compared to ordinal frequency scales [51,52]. We therefore dichotomized responses to 0=never and 1=at least once and subsequently summed them to form a scale. Self-defeating behavior was assessed at all time points (T1-T6: α=.60-.68).

Goal commitment to the yearly goal set by the participants was assessed with the Goal Commitment questionnaire [53] consisting of 7 items (eg, “I think this goal is a good goal to shoot for.”) answered on a 7-point Likert scale (1=completely disagree to 7=completely agree). Reliability was good at T4 through T6 (α=.73-.82), though somewhat low at T1 (α=.56).

Weekly and monthly goal achievement measured the degree to which participants reached their weekly and monthly goals. Both were assessed with 3 items on a 5-point Likert scale (1=completely disagree to 5=completely agree) developed for this study: “I have often thought about my goal,” “I have worked hard towards my goal,” and “I have achieved my goal.” Weekly goal achievement was assessed after each intervention week at T2, T3, and T4 (α=.72-.79). Monthly goal achievement was assessed immediately after the intervention at T4 (α=.75).

Secondary Outcomes

Self-efficacy assesses the degree to which people feel competent to deal with life’s stressors effectively and was measured with the General Self-Efficacy questionnaire [54] consisting of 10 items (eg, “I can always manage to solve difficult problems if I try hard enough.”) answered on a 4-point Likert scale (1=completely disagree to 4=completely agree). It was assessed at T1 and T4 through T6 (α=.79-.81).

The average academic results of the participants were obtained from the university’s records at the end of the academic year. Impulsiveness was assessed with the Barratt Impulsiveness Scale short form [55]. This questionnaire consists of 10 items (eg, “I do things without thinking.”) answered on a 4-point Likert scale (1=completely disagree to 4=completely agree) and was assessed at T1 and T4 through T6 (α=.84-.86).

Statistical Methods

The data were analyzed with an intention-to-treat approach, in which all participants assigned to the intervention were included in the analyses regardless of whether they received the intervention or not. To include all participants in the analyses, we used full information maximum likelihood procedures with maximum likelihood with robust standard errors estimation. Interim analyses and stopping guidelines were not applied.

The effectiveness of the intervention was examined using a series of latent growth curve (LGC) models in R with the *LAVAAN* package [56]. LGC models estimate an individual growth curve for each participant based on their initial level (ie, intercept) and change over time (ie, slope). These individual growth curves are used as indicators of latent variables that describe the average growth trajectories of the group while allowing for differences in trajectories between participants (when the variance of a slope is fixed to 0, the average growth trajectories of the group are modeled, but not the differences in trajectories between participants) [57]. Intervention effects show when the slope alters in the desired direction compared to the control condition.

In this study, growth rates of slopes were specified to indicate the differences in time intervals between the time points (ie, number of weeks since baseline, including the 3- and 6-month follow-ups), that is, 0, 1, 2, 3, 15, and 27. We modeled 3 different types of LGC models, as not all outcomes were assessed at each time point. For outcomes measured at all time points (ie, T1-T6), we modeled linear piecewise LGC models with a slope for change over time during the intervention (T1-T4) and a slope for change over time during the follow-up period (T4-T6) for each outcome variable separately. For outcomes measured at baseline, immediately after the intervention, and at follow-up (T1, T4-T6), we modeled LGC models, in which T4 was unspecified to allow for nonlinear growth. For weekly goal achievement, measured at T2, T3, and T4, we modeled an LGC model, in which T3 was unspecified. Monthly goal achievement and academic results were assessed at only 1 time point and therefore analyzed with an ANOVA. Sensitivity analyses were conducted to test the robustness of the results (Multimedia Appendix 1). No mediation analyses were conducted, as these were beyond the scope of the paper.

The effects of the intervention were assessed by creating 2 dummy variables (ie, smartphone and VR condition) with the control condition as a reference group and regressing the intercept and the slopes on these 2 dummy variables. In case of intervention effects in the piecewise LGC models, the effect sizes were calculated using the formula: $d=((b_{slope\ 1}*duration_{phase\ 1})+(b_{slope\ 2}*duration_{phase\ 2}))/SD_{outcome\ T1\ pooled}$. In case of intervention effects in the unconditional LGC models, we used the formula: $d=(b_{slope}*duration_{Total})/SD_{outcome\ T1\ pooled}$ [58]. Additionally, when intervention effects appeared for both

XSL•FO
RenderX

intervention conditions, these 2 conditions were subsequently compared by selecting the participants in the intervention conditions and rerunning the model with the smartphone condition as the reference group.

Ethical Considerations

The study was approved by the independent ethics board of the Institute of Education and Child Studies at Leiden University (ECPW2021-320) and was prospectively registered on ClinicalTrials.gov (NCT05578755) on October 13, 2022, with first participant enrollment on October 19, 2022 (see Mertens et al [45] for the study protocol). No changes to the trial were made after registration. The study is reported following the CONSORT (Consolidated Standards of Reporting Trials) 2025 statement [59] and CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) guidelines [60]. All participants provided active informed consent, after reading the information letter and asking any questions they had, at the beginning of the intake before any data were collected. To ensure the privacy of the participants, data were handled according to the study's Data Management Plan and Data Processing Inventory Analysis, approved by the data manager and privacy officer of Leiden University. Data were anonymized and stored on the secured servers of Leiden University, and access to the data was limited to the research team. The pictures of the participants' faces were automatically, and permanently, deleted from the secured servers after 24 hours. Participants received either 8 course credits or approximately US $40 for completing the questionnaires of T1 to T4. For completing both follow-up questionnaires (T5 and T6), they received an additional approximately US $23. Figure 1 presents screenshots of the app, showing the future self-rendering of EM, coauthor of this paper. Consent has been granted for the publication of these images.

Attrition Analyses

In total, 10% of the data were missing. Results of the attrition analyses were based on an α level of .01 to correct for multiple testing. A series of Little's missing completely at random tests indicated that there was no strong evidence against missing completely at random, though note that one *P* value was .04, which would be considered significant at an α level of .05 (χ^2_2=0.3 - 13.5; *P*=.04-.86).

Attrition analyses showed that participants with missing data at T3 (n=13; multivariate analysis of variance [MANOVA] $F_{9,311}$=1.08; *P*=.38; $\eta^2_{partial}$=0.030), T4 (n=10; MANOVA $F_{9,311}$=1.85; *P*=.06; $\eta^2_{partial}$=0.051), T5 (n=67; MANOVA $F_{9,311}$=1.78; *P*=.07; $\eta^2_{partial}$=0.049), and T6 (MANOVA $F_{9,311}$=2.31; *P*=.02; $\eta^2_{partial}$=0.063) did not significantly differ at baseline from participants with complete data at the corresponding time point on age and the outcome variables. At T2, there were differences (n=10; MANOVA $F_{9,311}$=2.60; *P*=.007; $\eta^2_{partial}$=0.070) with participants with missing data scoring lower at baseline on valence (mean 5.70, SD 1.83) than participants with no missing data (mean 6.84, SD 1.21). There were no differences in sex distribution between participants with and without missing data, except at T6. At T6, participants with missing data (n=88) were more often male (χ^2_1=7.4; *P*=.007; ϕ=−0.152; n=16, 18% vs n=18, 8%).

Intervention Delivery

In the smartphone condition, treatment adherence was moderate. During the 21-day intervention period, participants were asked to check into the app daily. In total, 12 (11%) participants checked in each day, and 53 (50%) participants checked in on 16 or more days. A total of 17 (16%) participants checked in on 7 or fewer days. Treatment adherence decreased during the intervention. During module 1, a total of 44 (41%) participants checked in daily; during module 2, a total of 30 (28%) participants did so; and during module 3, a total of 18 (17%) participants did so. The number of participants who checked in at least 5 of the 7 days of a module was 91 (85%), 60 (56%), and 52 (49%) respectively. On average, participants used the app for 14 (SD 5.44) days, with 5.56 (SD 1.75) minutes spent on the app on average per day. Furthermore, 13 (4%) participants experienced technical problems with the app, of which 1 (0.3%) was not able to install it. No adverse events or discomfort were reported.

In the VR condition, treatment adherence was adequate; all participants received the 3 VR sessions, except for 3 participants (1 VR session: n=2; 2 VR sessions: n=1). The sessions were intended to be scheduled between 6 and 10 days after intake or the preceding session. In total, 3 (3%) participants did not receive all 3 VR sessions. Of all sessions, 1 (0.3%) was scheduled with a shorter time span between 1 of the 3 sessions, and 43 (14%) were scheduled with a longer time span between 1 of the 3 sessions. The longer time spans ranged from 11 to 14 days between the sessions (n=32, 10%) to 15 to 18 days between the sessions (n=9, 3%). One participant had more than 19 days between sessions 2 and 3. Furthermore, 17 (16%) participants experienced technical problems, such as a distorted voice, during 1 or more sessions. After each VR session, participants were asked by the researcher how they were feeling as well as reported on cybersickness. No adverse events or discomfort were reported.

Results

Descriptives and Flowchart

The descriptives of the outcomes per condition at the different time points are presented in Multimedia Appendix 2, and the model fit statistics of the (different types of) LGC models are provided in Table 2. See Figure 3 for flowchart.

XSL•FO
RenderX

Table . Model fit statistics (CFI[a], RMSEA[b], and SRMR[c]) of the latent growth curve models of the outcome variables.

	CFI	RMSEA	SRMR
Proximal outcomes			
Vividness	0.971	0.079	0.036
Valence	0.996	0.022	0.024
Connectedness	0.985	0.055	0.042
Distal outcomes			
Primary outcomes			
Future orientation	0.969	0.095	0.047
Consideration of future consequences	0.996	0.032	0.025
Self-defeating behavior	0.964	0.072	0.038
Goal commitment	0.827	0.148	0.064
Weekly goal achievement[d]	0.915	0.126	0.053
Secondary outcomes			
Self-efficacy[d]	1.000	0.000	0.021
Impulsiveness[e]	1.000	0.010	0.015

[a]CFI: comparative fit index.

[b]RMSEA: root mean square error of approximation.

[c]SRMR: standardized root mean square residual.

[d]Variance S1 constrained to 0 due to small negative variances.

[e]Unspecified T4 growth model did not converge, so modeled a linear model with time specified as 0, 3, 15, and 27.

XSL•FO
RenderX

Figure 3. CONSORT flow diagram showing participant recruitment, randomization into the 3 conditions (smartphone, VR, and control), and retention across all 6 time points (from baseline to follow-up). CONSORT: Consolidated Standards of Reporting Trials; VR: virtual reality.

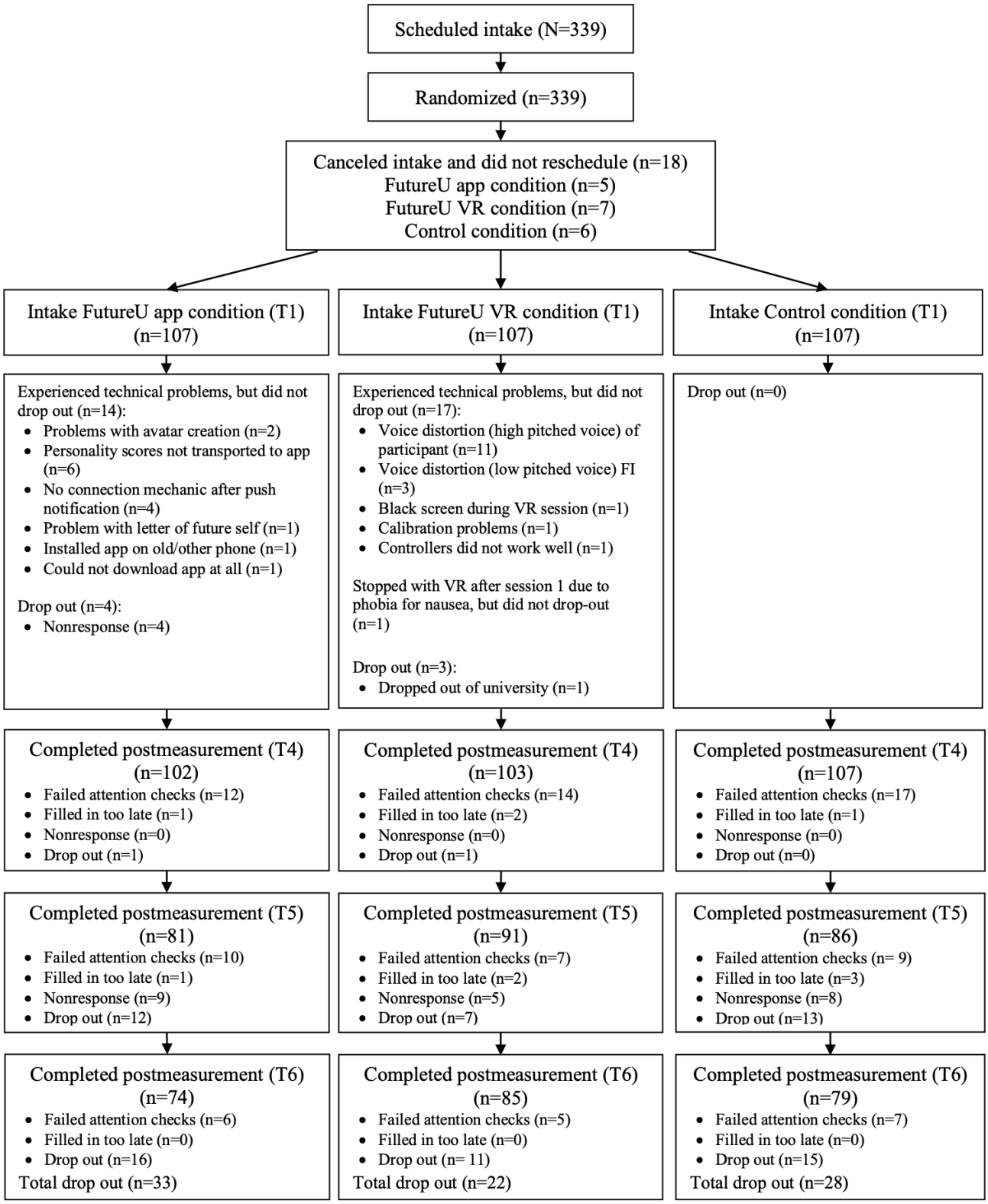

Intervention Effects on Proximal Outcomes

During the intervention period, the LGC models showed moderate positive intervention effects on all 3 of the proximal outcomes (Table 3). Participants in both the smartphone and the VR conditions increased more on vividness, connectedness, and valence than participants in the control condition. At follow-up, there was a negative effect on vividness and connectedness, in which participants in both intervention conditions decreased more than participants in the control condition. Valence remained stable at follow-up. The effect sizes of the decrease during the 6-month follow-up period were small to moderate and smaller than the increase during the 3-week intervention period. The intervention effects did not differ significantly nor relevantly between the 2 intervention conditions.

XSL•FO
RenderX

Table . Parameter estimates and effect sizes of the smartphone and virtual reality (VR) conditions compared to the control condition[a].

	Smartphone condition				VR condition			
	B (SE)	95% CI	*P* value	*d*	B (SE)	95% CI	*P* value	*d*
Proximal outcomes								
Vividness								
I[b]	−0.02 (0.18)	−0.37 to 0.34	.92	__[c]	0.24 (0.17)	−0.10 to 0.57	.17	—
S1[d]	0.23 (0.05)	0.14 to 0.32	<.001	0.49	0.17 (0.05)	0.07 to 0.26	<.001	0.35
S2[e]	−0.02 (0.01)	−0.03 to −0.01	<.001	−0.36	−0.01 (0.01)	−0.02 to −0.00	.01	−0.23
Valence								
I	−0.11 (0.16)	−0.43 to 0.20	.48	—	−0.06 (0.15)	−0.36 to 0.23	.68	—
S1	0.18 (0.05)	0.09 to 0.28	<.001	0.44	0.18 (0.05)	0.09 to 0.28	<.001	0.44
S2	−0.01 (0.01)	−0.02 to 0.00	.17	−0.17	−0.01 (0.01)	−0.02 to 0.01	.27	−0.13
Connectedness								
I	−0.06 (0.12)	−0.30 to 0.18	.62	—	0.14 (0.12)	−0.10 to 0.39	.25	—
S1	0.13 (0.04)	0.06 to 0.21	.001	0.43	0.13 (0.04)	0.05 to 0.21	.001	0.43
S2	−0.01 (0.01)	−0.02 to −0.00	.009	−0.36	−0.01 (0.01)	−0.02 to −0.00	.02	−0.32
Distal outcomes								
Primary outcomes								
Future orientation								
I	0.05 (0.07)	−0.09 to 0.18	.51	—	0.07 (0.07)	−0.06 to 0.21	.29	—
S1	0.03 (0.01)	0.00 to 0.05	.048	0.16	0.03 (0.01)	0.01 to 0.05	.01	0.18
S2	−0.00 (0.00)	−0.01 to −0.00	.01	−0.18	−0.01 (0.00)	−0.01 to −0.00	<.001	−0.25
Consideration of future consequences								
I	0.02 (0.09)	−0.16 to 0.20	.83	—	0.13 (0.08)	−0.02 to 0.28	.09	—
S1	−0.01 (0.00)	−0.01 to 0.00	.29	−0.21	−0.00 (0.00)	−0.01 to 0.01	.55	−0.12
Self-defeating behavior								
I	−0.19 (0.24)	−0.65 to 0.28	.43	—	−0.24 (0.25)	−0.74 to 0.26	.36	—
S1	0.01 (0.08)	−0.15 to 0.17	.93	−0.02	0.04 (0.08)	−0.12 to 0.20	.64	−0.07
S2	0.01 (0.01)	−0.01 to 0.03	.52	−0.08	−0.00 (0.01)	−0.02 to 0.02	.79	0.04
Goal commitment								
I	0.09 (0.07)	−0.05 to 0.23	.20	—	0.01 (0.07)	−0.13 to 0.14	.92	—
S1	−0.00 (0.01)	−0.01 to 0.01	.85	−0.06	−0.00 (0.01)	−0.01 to 0.01	.54	−0.17
Weekly goal achievement								
I	0.10 (0.11)	−0.12 to 0.32	.38	—	0.00 (0.10)	−0.20 to 0.20	.99	—
S1	−0.03 (0.07)	−0.17 to 0.11	.63	−0.08	0.37 (0.08)	0.20 to 0.53	<.001	0.88
Secondary outcomes								
Self-efficacy								
I	0.02 (0.04)	−0.07 to 0.10	.73	—	0.09 (0.05)	−0.01 to 0.18	.08	—
S1	−0.00 (0.00)	−0.01 to 0.00	.43	−0.11	−0.00 (0.00)	−0.01 to 0.00	.19	−0.17
Impulsiveness								
I	−0.01 (0.05)	−0.11 to 0.09	.84	—	−0.06 (0.05)	−0.17 to 0.04	.25	—
S1	0.00 (0.00)	−0.00 to 0.00	.40	−0.07	0.00 (0.00)	−0.00 to 0.00	.27	−0.09

[a]Positive effect sizes indicate changes in the desired direction.

XSL•FO
RenderX

[b]I: intercept latent growth curve model.

[c]Not applicable.

[d]S1: slope 1 latent growth curve model.

[e]S2: slope 2 latent growth curve model.

Intervention Effects on Distal Outcomes: Primary Outcomes

During the intervention period, there was a small positive intervention effect on future orientation. Participants in both intervention conditions remained stable on future orientation, whereas participants in the control condition decreased on this outcome. At follow-up, there was a small negative effect on future orientation for the smartphone and VR conditions compared to the control condition. Participants in the intervention conditions remained stable on future orientation, where participants in the control condition showed a small increase. The effect sizes of the positive effects during the intervention period were almost equal to the effect sizes of the negative effects at follow-up. These effects were similar in both intervention conditions.

There was also a large positive effect on weekly goal achievement in the VR condition compared to the control condition. Participants in the VR condition increased more on weekly goal achievement than participants in the control condition.

No significant nor relevant intervention effects on the other primary outcomes emerged (Table 3). Note, however, that the model fit of the LGC model of goal commitment was moderate (Table 2), which may have reduced sensitivity for significant differences between conditions. Furthermore, there was no intervention effect on monthly goal achievement (ANOVA $F_{1,308}$=0.13; P=.72; $\eta^2_{partial}$=0.00).

Intervention Effects on Distal Outcomes: Secondary Outcomes

There were no significant or relevant intervention effects on the secondary outcomes (Table 3). In addition, there was no effect on academic results (ANOVA $F_{1,316}$=1.02; P=.31; $\eta^2_{partial}$=0.00).

Discussion

Summary of Main Findings

This study provides evidence that future-oriented thinking and goal achievement can be strengthened through a digital intervention focused on fostering identification with the future self. Consistent with expectations, the intervention showed robust effects on proximal outcomes, with both delivery modes significantly increasing vividness of, connectedness with, and valence toward the future self during the intervention period. At follow-up, the intervention effect on valence remained stable, whereas a partial reversal of the effects on vividness and connectedness was observed. The finding that intervention effects slightly decreased over time, albeit less than the initial increase, indicates that the gains in future self-identification may attenuate over time without continued reinforcement. This suggests that booster sessions or continued exposure to the future self could be necessary to maintain the effects. An intervention app may be best suited for this, as an app can remain installed on participants' smartphones after the intervention has concluded to deliver intervention content, eg, push notifications and/or future self-avatar messages, or allow participants to use the app on their own initiative [32,33].

The intervention had more limited effects on the distal outcomes, although there was a small positive effect on future orientation in both intervention conditions and a strong positive effect on weekly goal achievement when the intervention was delivered via VR. There was no effect on monthly goal achievement, indicating that short-term goal engagement may have been boosted but not more durable goal progress. No effects were found for the other primary outcomes (ie, consideration of future consequences, self-defeating behavior, goal commitment, and monthly goal achievement) nor for the secondary outcomes (ie, self-efficacy, impulsiveness, and academic performance).

The positive effect on future orientation may indicate that enhanced future self-identification indeed cultivates a future-oriented mindset, as theorized [10,12]. However, more distal and often more deeply embedded behavioral patterns related to such a mindset, such as self-defeating behavior, may need a stronger, more focused, and/or more sustained intervention. Importantly, these null findings were consistent across measures and time points, suggesting that while the intervention momentarily influenced self-perception and weekly goal engagement within the intervention context, such changes did not translate into sustained or generalized behavioral change over the longer term, signaling room for further optimization.

Regarding weekly goal achievement, the intervention had no effect when delivered via the app, but there was a strong effect when delivered in VR. The VR intervention created an immersive experience, in which people talked to their future self, embodied this self, and took its perspective. As such, the VR condition created a first-person experience of how it could feel to have achieved personal goals. Imagination of outcomes has been shown to be an important step toward goal achievement [61]. Nonetheless, the difference between delivery modes cannot be attributed unambiguously to immersive VR technology per se. The VR condition differed from the smartphone condition on several dimensions beyond immersion that may have also affected goal achievement. Specifically, because FutureU VR was provided in person, people in the VR condition had direct interaction with a researcher and received explicit checks of weekly goals against SMART criteria. In contrast, participants using the app engaged independently with brief daily app interactions, without personal contact, external accountability, or synchronous feedback on SMART criteria. The possibility that these differences had effects over and above the immersion in VR can therefore not be excluded. As such, the observed advantage of the VR condition should be interpreted as hypothesis-generating rather than as evidence that immersive

XSL•FO
RenderX

VR delivery is inherently more effective than app-based delivery for goal achievement.

Taken together, these results underscore a key design challenge for future self-interventions: brief, standalone VR experiences and app interactions can be effective for eliciting short-term psychological shifts, but driving enduring behavioral or academic change can be more challenging. Rather than a limitation unique to this study, the absence of distal effects aligns with broader evidence that translating changes in future-oriented thinking into lasting real-world behavior likely requires repeated exposure, scaffolding, and integration with ongoing self-regulatory support [33]. As such, the present findings contribute to a more nuanced understanding of what VR- and app-based future self-interventions can—and cannot—be expected to achieve in the short term.

Strengths and Limitations

A key strength of this study is its rigorous randomized controlled design and relatively large sample size. Prior work on future self-identification has often relied on cross-sectional data and/or smaller samples (for exceptions, see Hershfield et al [14] and Ganschow et al [62]). Additionally, the inclusion of both a smartphone and a VR modality enabled direct comparison of these delivery formats. Furthermore, research on future self-identification has generally targeted only one of the aspects of its tripartite structure or used index measures that do not allow for disentangling individual effects [33], whereas our study included all three simultaneously and examined them separately.

Several limitations of this study must also be acknowledged. First, and foremost, delivery modality was confounded with other design features. The VR condition involved in-person researcher involvement and structured goal review, whereas the smartphone condition relied on brief, self-guided engagement. While more extensive personal contact can be seen as a secondary characteristic of this type of technology, our design makes it difficult to isolate the specific contribution of immersive VR technology from the effects of human support. To disentangle these factors, future research could, for example, provide equivalent in-person guidance across smartphone and VR conditions or test fully automated VR interventions without researcher involvement. Such designs would allow clearer conclusions about whether embodiment and immersion independently contribute to goal achievement.

Second, although we deliberately opted for first-year university students as this sample allowed us to capitalize on a major life transition phase, which may increase receptivity to interventions, it also limits the generalizability of our findings. The sample was predominantly female and drawn from a relatively narrow university context, a population that may already be relatively future-focused and hence have limited room for growth. Additionally, participants received compensation for completing the questionnaires, which could have affected the results and limited generalization to uncompensated populations. Furthermore, self-reported measures, especially for constructs like self-defeating behavior and impulsivity, may be susceptible to bias or lack of sensitivity to subtle behavioral changes, which may therefore not have been picked up. Finally, some effects may have been masked by measurement timing; for instance, positive gains in future orientation during the intervention period may have waned by follow-up, while longer-term benefits (eg, academic performance) may not yet have materialized.

Conclusions

Given the transdiagnostic potential of the future self—promoting positive behavior [9,13], reducing negative behavior [15,16,19], and improving mental health [20]—it is increasingly leveraged as an intervention mechanism. Building on previous research, this study evaluated a future self-intervention and not only examined all three aspects of future self-identification but also a novel implementation method with new technologies. Where more traditional future self-interventions rely on imagination [30,31], visual and immersive technologies, such as smartphone apps versus VR, can relieve this cognitive burden by presenting graphic renderings of the future self and enabling a sense of interaction with it. The results reported in this study suggest that visual and interactive technologies may serve as promising tools for strengthening future self-identification, cultivating a future-oriented mindset, and contributing to goal achievement. To enhance and sustain effects, future work could explore the integration of booster sessions, more personalized content delivery (eg, through the integration of artificial intelligence–powered conversational agents), application in more diverse and at-risk populations, and more targeted interventions. Indeed, recent work has already started to leverage novel possibilities offered by generative artificial intelligence, including voice cloning and personalized biographical narratives [63-65], to augment future-oriented thinking. All in all, the results of these studies and our own underscore the potential of future self-based interventions using novel technologies as scalable strategies to support long-term goal pursuit and positive psychosocial development.

Acknowledgments

The authors would like to thank Orb Amsterdam and Studio Barbaar for developing the smartphone app and the virtual reality environment and for providing helpful suggestions.

Funding

This study was financially supported by the ERC Consolidator Grant (772911-CRIMETIME). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are available in the Open Science Framework repository [66].

Authors' Contributions

Conceptualization: JLG
Data curation: EM, TT, AS
Formal analysis: EM
Funding acquisition: JLG
Investigation: EM, TT, AS
Methodology: JLG, EM
Project administration: EM
Supervision: JLG
Writing—original draft: JLG
Writing—review and editing: JLG, EM, TT, AS

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sensitivity analyses.
[DOCX File, 29 KB - jmir_v28i1e84420_app1.docx]

Multimedia Appendix 2
Descriptives of the outcomes per condition per time point.
[DOCX File, 20 KB - jmir_v28i1e84420_app2.docx]

Checklist 1
CONSORT checklist.
[DOCX File, 49 KB - jmir_v28i1e84420_app3.docx]

Checklist 2
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File, 10342 KB - jmir_v28i1e84420_app4.pdf]

References

1. Roepke AM, Seligman MEP. Depression and prospection. Br J Clin Psychol 2016 Mar;55(1):23-48. [doi: 10.1111/bjc.12087]
2. Suddendorf T, Corballis MC. The evolution of foresight: what is mental time travel, and is it unique to humans? Behav Brain Sci 2007 Jun;30(3):299-313. [doi: 10.1017/S0140525X07001975] [Medline: 17963565]
3. Steinberg L, Graham S, O'Brien L, Woolard J, Cauffman E, Banich M. Age differences in future orientation and delay discounting. Child Dev 2009;80(1):28-44. [doi: 10.1111/j.1467-8624.2008.01244.x] [Medline: 19236391]
4. Schmid KL, Phelps E, Lerner RM. Constructing positive futures: modeling the relationship between adolescents' hopeful future expectations and intentional self regulation in predicting positive youth development. J Adolesc 2011 Dec;34(6):1127-1135. [doi: 10.1016/j.adolescence.2011.07.009] [Medline: 22118506]
5. Rolison JJ, Hanoch Y, Wood S. Saving for the future: dynamic effects of time horizon. J Behav Exp Econ 2017 Oct;70:47-54. [doi: 10.1016/j.socec.2017.07.006]
6. Loewenstein G. Out of control: visceral influences on behavior. Organ Behav Hum Decis Process 1996 Mar;65(3):272-292. [doi: 10.1006/obhd.1996.0028]
7. van Gelder JL, Frankenhuis WE. Short-term mindsets and crime. Annu Rev Criminol 2025 Jan 29;8(1):333-358. [doi: 10.1146/annurev-criminol-022422-124536]
8. Robbins RN, Bryan A. Relationships between future orientation, impulsive sensation seeking, and risk behavior among adjudicated adolescents. J Adolesc Res 2004 Jul 1;19(4):428-445. [doi: 10.1177/0743558403258860] [Medline: 16429605]
9. Rutchick AM, Slepian ML, Reyes MO, Pleskus LN, Hershfield HE. Future self-continuity is associated with improved health and increases exercise behavior. J Exp Psychol Appl 2018 Mar;24(1):72-80. [doi: 10.1037/xap0000153] [Medline: 29595304]
10. Van Gelder J, Luciano EC, Weulen Kranenbarg M, Hershfield HE. Friends with my future self: longitudinal vividness intervention reduces delinquency. Criminology 2015 May;53(2):158-179. [doi: 10.1111/1745-9125.12064]

XSL•FO
RenderX

11. Ganschow B, Zebel S, van Gelder JL, Cornet LJM. Feeling connected but dissimilar to one's future self reduces the intention-behavior gap. PLoS One 2024;19(7):e0305815. [doi: 10.1371/journal.pone.0305815] [Medline: 39042617]
12. Hershfield HE. Future self-continuity: how conceptions of the future self transform intertemporal choice. Ann N Y Acad Sci 2011 Oct;1235(1):30-43. [doi: 10.1111/j.1749-6632.2011.06201.x] [Medline: 22023566]
13. Shen YI, Nelson AJ, Oberlin BG. Virtual reality intervention effects on future self-continuity and delayed reward preference in substance use disorder recovery: pilot study results. Discov Ment Health 2022;2(1):19. [doi: 10.1007/s44192-022-00022-1] [Medline: 36128578]
14. Hershfield HE, Brimhall CI, Kerbel S. Exploring the distribution and correlates of future self-continuity in a large, nationally representative sample. Judgm Decis Mak 2025;20(29):1-14. [doi: 10.1017/jdm.2024.23]
15. van Gelder JL, Hershfield HE, Nordgren LF. Vividness of the future self predicts delinquency. Psychol Sci 2013 Jun;24(6):974-980. [doi: 10.1177/0956797612465197] [Medline: 23592649]
16. van Gelder JL, Cornet LJM, Zwalua NP, Mertens ECA, van der Schalk J. Interaction with the future self in virtual reality reduces self-defeating behavior in a sample of convicted offenders. Sci Rep 2022 Feb 10;12(1):2254. [doi: 10.1038/s41598-022-06305-5] [Medline: 35145177]
17. Hards E, Hsu TC, Joshi G, Ellis J, Reynolds S. "Who will I become?": possible selves and depression symptoms in adolescents. Behav Cogn Psychother 2024 Jul;52(4):414-425. [doi: 10.1017/S1352465823000619] [Medline: 38284269]
18. Levin C, Nenninger S, Freundlich D, Glatt S, Sokol Y. How future self-continuity mediates the impact of job loss on negative mental health outcomes among transitioning veterans. Mil Psychol 2024 Sep 2;36(5):491-503. [doi: 10.1080/08995605.2023.2215697] [Medline: 37252891]
19. Sokol Y, Serper M. Temporal self appraisal and continuous identity: associations with depression and hopelessness. J Affect Disord 2017 Jan 15;208:503-511. [doi: 10.1016/j.jad.2016.10.033] [Medline: 27832930]
20. Webb MA, van Gelder JL. A virtual-reality based intervention on thoughts of the future self to reduce negative affect, depression, and suicidal ideation: protocol for a feasibility and acceptability randomized controlled pilot trial (FutureU for Mental Health). Pilot Feasibility Stud 2025 Nov 15;11(1):143. [doi: 10.1186/s40814-025-01709-2] [Medline: 41257809]
21. Bixter MT, McMichael SL, Bunker CJ, et al. A test of a triadic conceptualization of future self-identification. PLoS One 2020;15(11):e0242504. [doi: 10.1371/journal.pone.0242504] [Medline: 33232353]
22. Hershfield HE, Bartels D. The future self. In: Oettingen G, Sevincer AT, Gollwitzer P, editors. The Psychology of Thinking About the Future: Guilford Press; 2018:89-109.
23. Trope Y, Liberman N. Temporal construal. Psychol Rev 2003;110(3):403-421. [doi: 10.1037/0033-295X.110.3.403]
24. Trope Y, Liberman N. Construal-level theory of psychological distance. Psychol Rev 2010;117(2):440-463. [doi: 10.1037/a0018963]
25. Schacter DL, Benoit RG, Szpunar KK. Episodic future thinking: mechanisms and functions. Curr Opin Behav Sci 2017 Oct;17:41-50. [doi: 10.1016/j.cobeha.2017.06.002] [Medline: 29130061]
26. Renner F, Murphy FC, Ji JL, Manly T, Holmes EA. Mental imagery as a "motivational amplifier" to promote activities. Behav Res Ther 2019 Mar;114:51-59. [doi: 10.1016/j.brat.2019.02.002] [Medline: 30797989]
27. Sokol Y, Serper M. Development and validation of a future self-continuity questionnaire: a preliminary report. J Pers Assess 2020;102(5):677-688. [doi: 10.1080/00223891.2019.1611588] [Medline: 31107602]
28. Parfit D. Reasons and Persons: Oxford University Press; 1984.
29. Brown JM, Stein JS. Putting prospection into practice: methodological considerations in the use of episodic future thinking to reduce delay discounting and maladaptive health behaviors. Front Public Health 2022;10:1020171. [doi: 10.3389/fpubh.2022.1020171] [Medline: 36408004]
30. Hollis-Hansen K, Seidman J, O'Donnell S, et al. An ecological momentary episodic future thinking intervention on mother's weekly food purchases. Health Psychol 2020 Feb;39(2):159-167. [doi: 10.1037/hea0000817] [Medline: 31682147]
31. Chishima Y, Wilson AE. Conversation with a future self: a letter-exchange exercise enhances student self-continuity, career planning, and academic thinking. Self Identity 2021 Jul 4;20(5):646-671. [doi: 10.1080/15298868.2020.1754283]
32. Mertens ECA, Van Gelder JL. The DID-guide: a guide to developing digital mental health interventions. Internet Interv 2025;39:100794. [doi: 10.1016/j.invent.2025.100794] [Medline: 39720336]
33. Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry 2019 Oct;18(3):325-336. [doi: 10.1002/wps.20673] [Medline: 31496095]
34. Schoeppe S, Alley S, Van Lippevelde W, et al. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review. Int J Behav Nutr Phys Act 2016 Dec 7;13(1):127. [doi: 10.1186/s12966-016-0454-y] [Medline: 27927218]
35. Slater M, Sanchez-Vives MV. Enhancing our lives with immersive virtual reality. Front Robot AI 2016;3. [doi: 10.3389/frobt.2016.00074]
36. Segovia KY, Bailenson JN. Virtual imposters: responses to avatars that do not look like their controllers. Soc Influ 2012 Oct;7(4):285-303. [doi: 10.1080/15534510.2012.670906]
37. Gonzalez-Franco M, Lanier J. Model of illusions and virtual reality. Front Psychol 2017;8:1125. [doi: 10.3389/fpsyg.2017.01125] [Medline: 28713323]

XSL•FO
RenderX

38. McMichael SL, Bixter MT, Okun MA, et al. Is seeing believing? A longitudinal study of vividness of the future and its effects on academic self-efficacy and success in college. Pers Soc Psychol Bull 2022 Mar;48(3):478-492. [doi: 10.1177/01461672211015888] [Medline: 34018855]
39. Yeager DS. Dealing with social difficulty during adolescence: the role of implicit theories of personality. Child Dev Perspect 2017 Sep;11(3):196-201. [doi: 10.1111/cdep.12234] [Medline: 28983325]
40. Thielmann I, de Vries RE. Who wants to change and how? On the trait-specificity of personality change goals. J Pers Soc Psychol 2021 Nov;121(5):1112-1139. [doi: 10.1037/pspp0000304] [Medline: 33475400]
41. Grossmann I, Kross E. Exploring Solomon's paradox: self-distancing eliminates the self-other asymmetry in wise reasoning about close relationships in younger and older adults. Psychol Sci 2014 Aug;25(8):1571-1580. [doi: 10.1177/0956797614535400] [Medline: 24916084]
42. Grossmann I, Dorfman A, Oakes H, Santos HC, Vohs KD, Scholer AA. Training for wisdom: the distanced-self-reflection diary method. Psychol Sci 2021 Mar;32(3):381-394. [doi: 10.1177/0956797620969170] [Medline: 33539229]
43. Dweck CS, Yeager DS. Mindsets: a view from two eras. Perspect Psychol Sci 2019 May;14(3):481-496. [doi: 10.1177/1745691618804166] [Medline: 30707853]
44. Oettingen G, Gollwitzer PM. Strategies of setting and implementing goals: mental contrasting and implementation intentions. In: Maddux JE, editor. Social Psychological Foundations of Clinical Psychology: Guilford; 2010:114-135.
45. Mertens ECA, Siezenga AM, Tettero T, van Gelder JL. A future orientation intervention delivered through a smartphone application and virtual reality: study protocol for a randomized controlled trial. BMC Psychol 2022 Dec 20;10(1):315. [doi: 10.1186/s40359-022-01025-x] [Medline: 36539846]
46. Mertens ECA, Siezenga AM, van der Schalk J, van Gelder JL. A novel smartphone-based intervention aimed at increasing future orientation via the future self: a pilot randomized controlled trial of a prototype application. Prev Sci 2024 Feb;25(2):392-405. [doi: 10.1007/s11121-023-01609-y] [Medline: 37976010]
47. Dai H, Milkman KL, Riis J. The fresh start effect: temporal landmarks motivate aspirational behavior. Manage Sci 2014 Oct;60(10):2563-2582. [doi: 10.1287/mnsc.2014.1901]
48. Ogbeiwi O. General concepts of goals and goal-setting in healthcare: a narrative review. J Manag Organ 2021 Mar;27(2):324-341. [doi: 10.1017/jmo.2018.11]
49. Ersner-Hershfield H, Garton MT, Ballard K, Samanez-Larkin GR, Knutson B. Don't stop thinking about tomorrow: individual differences in future self-continuity account for saving. Judgm Decis Mak 2009 Jun 1;4(4):280-286. [doi: 10.1017/S1930297500003855] [Medline: 19774230]
50. Strathman A, Gleicher F, Boninger DS, Edwards CS. The consideration of future consequences: weighing immediate and distant outcomes of behavior. J Pers Soc Psychol 1994;66(4):742-752. [doi: 10.1037/0022-3514.66.4.742]
51. Bendixen M, Endresen IM, Olweus D. Variety and frequency scales of antisocial involvement: which one is better? Legal Criminol Psychol 2003 Sep;8(2):135-150. [doi: 10.1348/135532503322362924]
52. Sweeten G. Scaling criminal offending. J Quant Criminol 2012 Sep;28(3):533-557. [doi: 10.1007/s10940-011-9160-8]
53. Hollenbeck JR, Klein HJ, O'Leary AM, Wright PM. Investigation of the construct validity of a self-report measure of goal commitment. J Appl Psychol 1989;74(6):951-956. [doi: 10.1037/0021-9010.74.6.951]
54. Schwarzer R, Jerusalem M. Generalized self-efficacy scale. In: Weinman J, Wright S, Johnston M, editors. Measures in Health Psychology: A User's Portfolio Causal and Control Beliefs: NFER-NELSON; 1995:35-37.
55. Spinella M. Normative data and a short form of the Barratt Impulsiveness Scale. Int J Neurosci 2007 Mar;117(3):359-368. [doi: 10.1080/00207450600588881] [Medline: 17365120]
56. Rosseel Y. lavaan: an R package for structural equation modeling. J Stat Softw 2012;48(2):1-36. [doi: 10.18637/jss.v048.i02]
57. Muthén LK, Muthén BO. Mplus User's Guide, 8th edition: Muthén & Muthén; 2017.
58. Feingold A. New approaches for estimation of effect sizes and their confidence intervals for treatment effects from randomized controlled trials. Quant Method Psychol 2019;15(2):96-111. [doi: 10.20982/tqmp.15.2.p096] [Medline: 32775313]
59. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ 2025;389:e081123. [doi: 10.1136/bmj-2024-081123]
60. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
61. Adriaanse MA, Oettingen G, Gollwitzer PM, Hennes EP, de Ridder DTD, de Wit JBF. When planning is not enough: fighting unhealthy snacking habits by mental contrasting with implementation intentions (MCII). Eur J Soc Psychol 2010 Dec;40(7):1277-1293. [doi: 10.1002/ejsp.730]
62. Ganschow B, Zebel S, van der Schalk J, Hershfield HE, van Gelder JL. Adolescent stressful life events predict future self-connectedness in adulthood. J Early Adolesc 2024 Nov;44(9):1188-1218. [doi: 10.1177/02724316231216380]
63. Albrecht C, Archiwaranguprok C, Poonsiriwong R, et al. Future you: designing and evaluating multimodal AI-generated digital twins for strengthening future self-continuity. Presented at: IUI '26: Proceedings of the 31st International Conference on Intelligent User Interfaces; Mar 23-26, 2026. [doi: 10.1145/3742413.3789455]
64. Dechant M, Lash E, Shokr S, O'Driscoll C. Future Me, a prospection-based chatbot to promote mental well-being in youth: two exploratory user experience studies. JMIR Form Res 2025 Jul 18;9:e74411. [doi: 10.2196/74411] [Medline: 40680255]

XSL•FO
RenderX

65. Pataranutaporn P, Winson K, Yin P, et al. Future you: a conversation with an AI-generated future self reduces anxiety, negative emotions, and increases future self-continuity. Presented at: 2024 IEEE Frontiers in Education Conference (FIE); Oct 13-16, 2024. [doi: 10.1109/FIE61694.2024.10893443]
66. Mertens ECA, Tettero T, Siezenga AM, Van Gelder JL. A smartphone- and VR-based intervention aimed at increasing future orientation via the future self: a pilot RCT of a prototype application and an RCT of the smartphone- and VR-based intervention. Open Science Framework. URL: https://osf.io/9jbrp [accessed 2026-04-28]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
LGC: latent growth curve
MANOVA: multivariate analysis of variance
VR: virtual reality

Edited by S Brini; submitted 19.Sep.2025; peer-reviewed by A Shivanna, BG Oberlin, CY Su; revised version received 10.Mar.2026; accepted 10.Mar.2026; published 13.May.2026.

Please cite as:
van Gelder JL, Tettero T, Siezenga A, Mertens E
Stimulating Future-Oriented Thinking and Goal-Achievement Through the Future Self Using Virtual Reality and a Smartphone App: Randomized Controlled Trial
J Med Internet Res 2026;28:e84420
URL: https://www.jmir.org/2026/1/e84420
doi:10.2196/84420

© Jean-Louis van Gelder, Tiffany Tettero, Aniek Siezenga, Esther Mertens. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Self-Guided App-Based Mindfulness Intervention for Racially and Ethnically Minoritized Individuals Who Experience Discrimination-Related Mental Health Symptoms: Randomized Controlled Trial

Giovanni Ramos[1], PhD; Amanda K Montoya[2], PhD; Adrian Aguilera[3,4], PhD; Anna S Lau[2], PhD; Craig K Enders[2], PhD; Yinyin Wen[2], BA; Denise A Chavira[2], PhD

[1]Department of Psychology, University of California, Berkeley, Berkeley, 2121 Berkeley Way, Berkeley, CA, United States

[2]Department of Psychology, University of California, Los Angeles, Los Angeles, CA, United States

[3]School of Social Welfare, University of California, Berkeley, Berkeley, CA, United States

[4]Department of Psychiatry and Behavioral Sciences, University of California, San Francisco, San Francisco, CA, United States

Corresponding Author:
Giovanni Ramos, PhD
Department of Psychology, University of California, Berkeley, Berkeley, 2121 Berkeley Way, Berkeley, CA, United States

Abstract

Background: Racially and ethnically minoritized individuals (REMs) who experience discrimination are at risk of developing stress, anxiety, and depression, and digital mental health interventions (DMHIs) can make evidence-based treatments such as mindfulness available to these groups. However, REMs are significantly underrepresented in the overall DMHI and mindfulness-based DMHI literature, limiting our understanding of the effectiveness and feasibility of these digital tools to advance mental health equity for these populations.

Objective: This randomized controlled trial evaluated the effectiveness of a 4-week self-guided app-based mindfulness DMHI versus treatment as usual (TAU) in reducing discrimination-related stress, anxiety, and depressive symptoms among REMs, while assessing feasibility metrics, including uptake, engagement, dropout, and program satisfaction.

Methods: A total of 155 participants (mean age 27.28, SD 9.6 years) were randomized to either receive the DMHI (n=80) or TAU (n=75). Participants in the DMHI group underwent an onboarding procedure and were asked to complete 1 meditation daily for 4 weeks. Participants in the TAU group were encouraged to seek mental health services on their own. The racial and ethnic composition of the sample was 38.7% (60/155) Latinx, 37.4% (58/155) Asian, 11.6% (18/155) Black, 10.3% (16/155) Multiracial, and 2% (3/155) Native American. Mean differences between groups were examined using multilevel regression models, applying a false discovery rate (FDR) correction across all fixed effects. Treatment effect sizes and clinical significance were calculated using pseudo-R^2s and the percentage of participants who achieved a minimal clinically significant difference (MCID), respectively. Feasibility outcomes were examined descriptively and included the percentage of participants who accessed the DMHI, days using it, meditations completed, total time meditated, the percentage of participants who dropped out of the study, and self-reported treatment satisfaction.

Results: Compared with participants in the TAU group, participants in the DMHI group experienced greater reductions in stress (β=−4.52, 95% CI –6.54 to –2.51; FDR-adjusted P<.001; pseudo-R^2=.41; MCID=61%, 49/80), anxiety (β=−3.31, 95% CI –4.89 to –1.74; FDR-adjusted P<.001; pseudo-R^2=.45; MCID=48%, 38/80), and depression (β=−2.84, 95% CI –4.57 to –1.12; FDR-adjusted P=.004; pseudo-R^2=.53; MCID=54%, 43/80) by the end of the program. All participants in the DMHI group downloaded the app (80/80, 100%) and used it, on average, 16.83 (SD 7.8) days, completed 24.85 (SD 23.3) mediations, and meditated 238.67 (SD 260.7) minutes. Of all participants, 12 (12/155, 8%) dropped out of this study. Most dropouts occurred in the intervention group (11/80, 14%). Participants perceived the DMHI program as satisfactory by the end of treatment (mean treatment satisfaction 23.14, SD 4.9).

Conclusions: These findings position this self-guided app-based mindfulness DMHI as an effective and feasible strategy for mitigating the deleterious mental health consequences of exposure to discrimination among REMs.

Trial Registration: ClinicalTrials.gov NCT05027113; https://clinicaltrials.gov/ct2/show/NCT05027113

International Registered Report Identifier (IRRID): DERR1-10.2196/35196

(J Med Internet Res 2026;28:e84328) doi:10.2196/84328

XSL•FO
RenderX

KEYWORDS

digital mental health intervention; discrimination; mindfulness; racially and ethnically minoritized; stress; anxiety; depression; feasibility; equity; app

Introduction

In the United States, approximately 60% of racially and ethnically minoritized individuals (REMs) report experiencing daily instances of discrimination [1]. According to the Perceived Discrimination and Health Model [2], exposure to this chronic and unpredictable stressor triggers physiological (eg, cardiovascular reactivity), cognitive-affective (eg, rumination), and behavioral (eg, avoidance) responses that heighten the individual's vulnerability to developing mental health problems. Exposure to discrimination is most strongly associated with increases in stress, anxiety, and depression symptoms [3]. This discrimination-related distress is likely to contribute to the high prevalence rates of these mental health problems in REMs [3,4]. While acknowledging that the most potent strategy to reduce this mental health burden is to dismantle the societal systems that perpetuate discrimination, there is still a need for individual-level interventions that could help REMs cope [5].

Among mental health treatments available to help REMs cope with discrimination-related mental health problems, mindfulness meditation may be particularly well-suited. In mindfulness meditation, practitioners cultivate present-moment awareness of their experience, including physical sensations, thoughts, and emotions, with an open and nonjudgmental attitude [6]. In the context of dealing with the mental health consequences of experiencing discrimination, mindfulness meditation reduces physiological reactivity, breaks repetitive negative thinking patterns, and promotes adaptive emotion regulation strategies, potentially disrupting some of the mechanisms by which discrimination "gets under the skin," leading to stress, anxiety, and depression [5,7]. Numerous randomized controlled trials (RCTs) show that mindfulness interventions are effective in decreasing stress, anxiety, and depression [8]. However, mindfulness effectiveness research to date has been conducted with almost exclusively White samples [9], with only modest preliminary evidence in REMs [10].

Given the limited representation of REMs in mindfulness research, some researchers have theorized that, without careful cultural adaptation, these interventions may not be engaging or satisfying for these groups [11]. Concerns include incompatibility between REMs' religious beliefs and the Buddhist origins of mindfulness, negative attitudes toward experiencing and expressing certain emotions, and perceived mismatches between collectivistic values and the individualistic nature of meditation. Trying to address these issues, mindfulness interventions have been culturally adapted for several REM groups [11-14]. However, whether those culturally adapted interventions are more effective than the original protocols remains an empirical question. In fact, one of the few meta-analyses examining the gains associated with culturally adapting mindfulness for REMs found that the number of cultural adaptations performed was not associated with treatment gains [14]. Nevertheless, these findings should be considered preliminary, given the small sample sizes in these studies, the heterogeneous changes implemented, and the lack of distinction between true cultural adaptations and implementation adaptations. Considering that culturally adapting interventions is a complex, time-consuming, and resource-intensive process, data on non–culturally adapted programs' feasibility and effectiveness are needed to determine the need to, or lack thereof, make such adaptations [15].

Culturally adapted or not, REMs are less likely to access mindfulness interventions in the community [16]. These inequities in access seem to be driven mostly by provider shortages, transportation and time constraints, and treatment costs that disproportionally affect these groups [17]. Digital mental health interventions (DMHIs) could reduce this treatment gap by providing evidence-based treatments without the involvement of mental health professionals, making those interventions accessible anywhere at any time, and reducing costs [18]. Although in the United States REMs are still affected by the digital divide and are less likely than White individuals to have broadband internet at home [19], 85% - 96% of them still own a smartphone with internet access [20]. Furthermore, a growing body of evidence supports the effectiveness of smartphone-based DMHIs [21], including mindfulness apps [22]. However, like traditional mindfulness interventions, there is limited representation of REMs in DMHI research [23], including mindfulness-based DMHIs [22].

Including REMs in DMHI research is necessary to determine whether these approaches are truly able to reduce inequities in mental health. For instance, in mindfulness-based DMHI RCTs, up to 55% of participants never access the app, and those who do so only use it a couple of times before approximately 25% drop out of treatment [24]. These rates are even higher in real-world app usage, with up to 79% of users never opening the mental health app even after downloading it [25] and only 4% of those who downloaded it continue using it after 2 weeks [26]. Considering that REMs face numerous barriers to engaging with traditional mental health services [17], it may be possible that they also face significant difficulties engaging in DMHIs. Indeed, not having enough time, forgetting to use the digital tool, a lack of motivation, skepticism about the need for treatment and treatment effectiveness, and cost were the most commonly reported barriers to using DMHIs among REMs [27,28]. Therefore, examining the uptake of, engagement with, dropout in, and program satisfaction with DMHIs among REMs is a pressing need in the field.

Considering the significant gaps in both traditional and DMHI mindfulness research with REM groups, this RCT examines the effectiveness and feasibility of a self-guided app-based mindfulness DMHI for REMs who experience elevated levels of discrimination.

Aim 1: Test the effectiveness of a mindfulness-based DMHI in reducing stress, anxiety, and depression among REMs who experience elevated levels of discrimination. Hypothesis 1: Based on results from traditional mindfulness interventions [8]

XSL•FO
RenderX

and mindfulness-based DMHIs [22], we hypothesized that participants in the DMHI group would experience greater reductions in stress, anxiety, and depression compared with their peers in the treatment-as-usual (TAU) group.

Aim 2: Describe the feasibility of this DMHI program, including its uptake, engagement, dropout, and self-reported program satisfaction. Hypothesis 2: Given the lack of consensus on the most important engagement metrics in DMHIs [29] and benchmarks for what constitutes "optimal" engagement with these tools [30], we did not have specific hypotheses.

Methods

Study Context

This parallel-group RCT took place at the intersection of 3 major political and social events in the United States from 2020 to 2022 that made discrimination against REMs especially salient. First, a hostile political discourse against REMs, which was associated with increases in discrimination against these groups [31]. Second, mental health problems caused by exposure to discrimination among REMs who were blamed for the COVID-19 pandemic [32]. Third, multiple instances of police violence against REMs, which were associated with poor mental health outcomes at the population level [33]. Considering this context, there was a significant need for mental health services that could help REMs cope with symptoms of stress, anxiety, and depression associated with experiencing discrimination.

Ethical Considerations

The University of California, Los Angeles (UCLA) Institutional Review Board approved all procedures in this study (IRB 21 - 001426). The study hypotheses and data analytic plan were preregistered (NCT05027113), and the study protocol was peer-reviewed and published as a registered report (DERR1-10.2196/35196) [31]. We report any deviations from the registered report in each pertinent section. While this study focuses on the main clinical and feasibility outcomes of this RCT, additional measures of mechanisms of action were collected and will be reported in future publications. There are no previous publications coming from this dataset. A member of the research team contacted individuals who met eligibility criteria over the phone to discuss the study, answer questions, and obtain their informed consent to participate. Electronic raw data were stored in the Health Insurance Portability and Accountability Act (HIPAA)-compliant UCLA's Qualtrics server. Data were deidentified by assigning each participant a unique ID number and keeping consent forms and other digital documents containing identifiable information separate from study data. All participants received a free 2-month subscription to the DMHI as an incentive to participate. Additionally, participants were compensated with US $10, US $15, and US $25 for completing questionnaires and sharing app usage data at baseline, midtreatment, and posttreatment, respectively. Reporting in this trial adhered to the Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth (CONSORT-EHEALTH) [34].

Procedure

Nationwide recruitment took place from October 2021 to April 2022. Recruitment materials depicted REMs and used language that emphasized the stressful nature of experiencing discrimination. Materials were posted on social media, emailed to relevant social and professional groups, and physically distributed in communities with a high representation of REMs in Los Angeles, California. Individuals interested in participating in the study completed an online screening questionnaire. For study inclusion, participants were required to (1) identify as a REM, (2) consider themselves fluent in English, (3) not receive psychological services currently, (4) not have practiced mindfulness meditation for more than 2 hours in the month prior to study commencement, (5) own a smartphone with access to the internet, (6) be willing to install the mindfulness app and receive daily notifications and text reminders, (7) be at least 18 years old, and (8) report experiencing elevated levels of discrimination (ie, Multicultural Discrimination Module [MDM] score ≥12 [35]). Originally, participants who scored 55 or greater on the Index of Race-Related Stress-Brief (IRRS-B) [36] were also eligible. However, both measures seemed redundant as evidenced by their high correlation (r=0.72) and virtually equivalent correlations with the clinical outcomes at baseline—stress (r=0.24 and r=0.26), anxiety (r=0.60 and r=0.60), and depression (r=0.73 and r=0.73). Considering also that the MDM was developed for multiple REM groups while the IRRS-B was originally intended for one specific REM group (ie, Black), and the better psychometric properties of the MDM (ω=.93) compared with the IRRS-B (ω=.71) seen in this sample, we opted to only use the MDM inclusion criterion in this study. Nevertheless, this decision did not affect the final number of participants deemed eligible, enrolled, or included in the analyses.

Participants were randomly assigned to either the DMHI (n=80) or TAU (n=75) condition using computer-generated block randomization with a fixed block size of 5. The randomization sequence was generated automatically, and allocation was concealed until assignment. Due to the nature of the intervention, participants were not blinded to condition assignment. However, masking occurred at the levels of care provider (ie, the intervention was delivered via a fully automated DMHI), investigators (ie, group assignment was determined by a computer program without investigators' involvement), and outcome assessment (ie, all outcomes were collected via self-administered online assessments). Participants randomized to the DMHI group underwent a short, standardized onboarding procedure over the phone (approximately 30 min) in which they were assisted with installing the app, learning to use it, and setting reminders and notifications. Participants in the TAU condition were reminded of the scheduled assessments, encouraged to seek mental health services or use a DMHI on their own if interested, and given access to the study's DMHI 4 weeks after randomization. Data were collected entirely through self-assessments online at baseline (ie, randomization or wk 0), midtreatment (ie, wk 2), and posttreatment (ie, wk 4).

XSL•FO
RenderX

Group Conditions

DMHI Group

The Mindfulness for Us, or Mind-Us for short, program consisted of using the commercially available Ten Percent Happier app (now Happier Meditation) to complete at least 1 meditation daily, with the possibility of exceeding this goal by choosing additional meditations, for 4 weeks. More specifically, participants were asked to complete "The Basics," "The Basics II," and "Essential Advice" courses to learn the basics of mindfulness meditation. Each session begins with a short video providing psychoeducation on mindfulness principles (eg, awareness of the present moment) followed by teaching a specific mindfulness technique (eg, noting thoughts and emotions). These introductory courses also dispel common myths about mindfulness meditation (eg, meditators empty their minds) and teach users how to deal with common meditation obstacles (eg, strong negative emotions). The courses comprise 29 sessions, in which the meditation length gradually increases from 5 to 18 minutes. For additional practice, participants were allowed to select any meditation from the app. We chose this mindfulness app, given its high user ratings (ie, 4.8 stars out of 5 in app stores) and previous evidence of effectiveness with mostly White samples [37,38].

We chose to use a commercially available app instead of a culturally adapted one, given evidence showing that DMHIs based on evidence-based principles are effective for REMs, regardless of whether they are culturally adapted or not [15,18]. This approach has the potential to advance mental health equity for REM by making existing DMHIs immediately available to these groups, rather than waiting for culturally adapted versions to be developed [15]. Furthermore, using commercially available DMHIs with REMs can allow researchers to document potential areas of improvement for these tools, effectively gathering the data necessary to inform data-driven cultural adaptations of DMHIs that are more likely to improve treatment outcomes than when assuming a priori that these adaptations are needed [15,18,39].

In an effort to minimize engagement issues commonly observed in self-guided app-based DMHIs [26,40], the Mind-Us program used a combination of engagement-enhancement strategies. First, participants were encouraged to substitute longer meditations with shorter ones as needed to make meditation fit their routine and facilitate daily practice. The Ten Percent Happier app also allowed users to set a daily notification that can be personalized. Similarly, participants received 3 daily text messages in the morning, noon, and evening with psychoeducational content, prompts encouraging them to reflect on their progress, and reminders to meditate and complete study assessments. During the onboarding call, study staff also helped participants generate their own strategies to deal with potential barriers to treatment engagement (eg, integrating the use of the app into daily routines such as before going to bed).

Following recommendations on how to improve the cultural fit of mindfulness-based interventions [5,11-14] and DMHIs for REMs [18], during onboarding participants also discussed any reservations about using an app not explicitly designed for REM groups and received psychoeducation about how mindfulness meditation could help them cope with the mental health consequences of experiencing discrimination. In this discussion, study staff emphasized that mindfulness meditation would not teach participants to "accept" or "reappraise" the discriminatory behavior of others but rather to gain awareness of the distress associated with the experience, validate their emotional responses, and gain distance between negative messages about REM groups and their own sense of self [5]. All participants in the DMHI group (80/80, 100%) completed this onboarding procedure.

TAU Group

Originally, we had proposed to use a waitlist group as the control condition. Instead, participants randomized to the control group were encouraged to seek professional mental health services or use commercially available DMHIs on their own, verified via self-report at midtreatment and posttreatment, effectively making this control condition TAU. Given our overarching goal of making mental health services available to REMs exposed to discrimination, participants in the TAU group were also given access to the DMHI after completing the postintervention assessment. Then, they received an email with the same instructions for downloading and using the app that participants in the intervention group received during the onboarding call. We chose this control condition following the Pragmatic Model for Comparator Selection in Health-Related Behavioral Trials framework [41], which indicated that, given the limited effectiveness evidence for both face-to-face [8,10] and DMHI-based mindfulness programs for REM groups [22], a TAU control was adequate.

Measures

Screening

Discrimination

The 8-item MDM [35] is a measure of discrimination for different racial, ethnic, and cultural groups. Participants rate items on a scale ranging from 1 (never) to 4 (often), with higher scores indicating higher levels of discrimination over the last year. Items include "How often have you been treated with less respect than other people" and "How often have you been threatened or harassed." The MDM was developed using a nationally representative sample of REMs in the United States and has shown validity and good internal consistency (α=0.81-0.88) [35]. For study inclusion, participants needed to obtain at least a score of 12 at screening, which falls within the 75th percentile of the population distribution scores. In this study, the MDM reliability was excellent (ω=.93).

Effectiveness Outcomes

Stress

The 10-item Perceived Stress Scale (PSS) [42] is a measure of perceived stress. Participants rate items on a scale ranging from 0 (never) to 4 (very often), with higher scores indicating more severe stress symptoms over the last 2 weeks. The measure includes items, such as "How often have you felt that you were unable to control the important things in your life?" and "How often have you felt confident about your ability to handle your personal problems?" The PSS has shown good reliability

XSL•FO
RenderX

(α=0.84-0.86) and validity with REM samples [43]. The PSS was administered at baseline, midtreatment, and posttreatment. In this study, the PSS reliability was excellent (ω=.87-.91).

Anxiety

The 7-item General Anxiety Disorder-7 (GAD-7) [44] is a measure of anxiety symptoms. Participants rate items on a scale ranging from 0 (not at all) to 3 (nearly every day), with higher scores indicating more severe anxiety symptoms over the last 2 weeks. The scale includes items, such as "Feeling nervous, anxious or on edge," and "Not being able to stop or control worrying." The GAD-7 has shown good internal consistency (α=0.89-0.90) and validity with REM samples [45]. The GAD-7 was administered at baseline, midtreatment, and posttreatment. In this study, the GAD-7 reliability was excellent (ω=.88-.93).

Depression

The 8-item Patient Health Questionnaire-8 (PHQ-8) is a measure of depressive symptoms derived from the original 9-item Patient Health Questionnaire-9 (PHQ-9) [46], which omits the suicidal ideation item while still being virtually equivalent [47]. Participants rate items on a scale ranging from 0 (not at all) to 3 (nearly every day), with higher scores indicating more severe depressive symptoms over the last 2 weeks. This measure includes items, such as "Little interest or pleasure in doing things" and "Feeling down, depressed, or hopeless." The PHQ-8 has shown good reliability (α=0.86-0.89) and validity across different racial and ethnic groups [48]. The PHQ-8 was administered at baseline, midtreatment, and posttreatment. In this study, the PHQ-8 reliability was excellent (ω=.87-.90).

Feasibility Outcomes

Uptake

This feasibility outcome was operationalized as the percentage of participants who downloaded the app and created a user profile during the 4-week intervention.

Engagement

The number of days using the app, the number of meditations completed, and the total time meditated in minutes were used as behavioral measures of DMHI engagement. These implementation outcomes were tracked by the DMHI. Participants were asked to share their app usage data via screenshots when completing midtreatment and posttreatment assessments. Previous studies have shown that this method provides accurate behavioral indicators of treatment engagement [37,49].

Dropout

This feasibility outcome was operationalized as the percentage of participants who left the program and did not complete the posttreatment assessment, regardless of their level of engagement with the DMHI.

Program Satisfaction

The 6-item Satisfaction with Therapy subscale of the Satisfaction with Therapy and Therapist Scale-Revised (STTS-R) [50] was used to assess program satisfaction with the DMHI program. Participants rate items on a scale ranging from 1 (strongly disagree) to 5 (strongly agree), with higher scores indicating more program satisfaction in the last 2 weeks. Scores above 18 on this subscale are considered satisfactory. Examples of items include "I am satisfied with the quality of the program I received" and "I would recommend this program to a friend." The Satisfaction with Therapy subscale has shown excellent reliability (α=0.92) and validity in DMHIs with REM groups [51]. The Satisfaction with Therapy subscale was administered at midtreatment and posttreatment. In this study, the Satisfaction with Therapy subscale reliability was excellent (ω=.91-.93).

Data Analytic Plan

The original preregistered power analysis was updated based on more recent estimates of the treatment effect sizes of app-based mindfulness DMHIs [22] and published test-retest estimates for repeated measures of stress [43], anxiety [52], and depression [53]. A power analysis using G*Power (version 3.1.9.6; Faul, Erdfelder, Lang, and Buchner) [54] indicated that, based on 2 groups, 3 repeated measurements, a 2-week correlation of r=.60, and an average treatment effect size of d=0.30 (range 0.24 - 0.42), a sample of n=94 would provide 80% power to detect potential group-by-time interactions. Expecting an approximate 25% dropout rate similar to previous app-based mindfulness DMHI trials [24], we originally planned to recruit 120 participants. However, given the high demand for this program (ie, more than 50% of the original sample was recruited within the first month of study commencement), we expanded recruitment until resources to provide financial compensation to participants were exhausted.

All statistical analyses were conducted using R (version 4.3.2; R Core Team) [55]. Model assumptions were evaluated, including the distribution of outcomes, linearity of relationships, residual normality and homoscedasticity, and multicollinearity among predictors. Inspection of diagnostic plots and indices indicated no apparent violations, supporting the adequacy of the specified models. To test mean differences between groups, we conducted multilevel regression models with group condition as the predictor, time as a within-group factor, and a group-by-time interaction. We examined potential group differences from baseline to posttreatment (ie, wk 0 to wk 4), baseline to midtreatment (ie, wk 0 to wk 2), and midtreatment to posttreatment (ie, wk 2 to wk 4). We included participant age, gender, and perceived financial stress as covariates in all analyses, considering previous studies with REM populations showing differences in the frequency and impact of discrimination on mental health outcomes [1,2], as well as differences in the prevalence of mental health problems based on these demographics [4]. We fitted all models using the *nlme* package in R (version 3.1-158; R Core Team) [56], including random intercepts, an autoregressive error structure, and restricted maximum likelihood as the estimation method. To control for multiple testing in each model, we applied a Benjamini-Hochberg false discovery rate (FDR) correction across all fixed effects using the *p.adjust* function in R (version 4.3.2; R Core Team). This type I correction included all predictors and covariates, which provided a more conservative control of false discoveries by accounting for all inferential tests that could influence model interpretation. To determine the clinical significance of treatment effects, we compared

XSL•FO
RenderX

reductions in the PSS, GAD-7, and PHQ-8 scores observed in this study with the minimum clinically important difference (MCID) established for these measures. The MCID represents the smallest clinically relevant reduction in symptoms, regardless of whether the difference is statistically significant [57]. Previous studies have established MCIDs of 4 points for the PSS [58] and 3 points for both the GAD-7 and PHQ-9 [59]. Although the PHQ-8 and PHQ-9 are virtually equivalent [47], we prorated the PHQ-8 scores by multiplying them by 1.125 to align them with the PHQ-9 score range and facilitate comparisons with previous studies and future meta-analytic work. This is a common practice in DMHI studies using the PHQ-8 [60]. Nonprorated results were virtually equivalent and are publicly available elsewhere [61]. To determine treatment effect sizes, we calculated pseudo-R^2 statistics for all models using the *r2mlm* package in R (version 0.3.2; R Core Team) [62], which provides estimates of the total variance explained by the model. We also examined the feasibility outcomes of this DMHI (ie, uptake, engagement, dropout, and program satisfaction) using descriptive statistics. Originally, we had proposed to use the last-observation-carried-forward method to handle missing data. Instead, we tested the robustness of these results under a variety of missing data models and assumptions about the missing data mechanisms, using a Bayesian latent variable estimation modeling and imputation approach using Blimp (version 4.0; Keller and Enders) [63]. Only 8% (13/155) of participants had missing outcome data in this trial. Similarly, only 12% (9/80) of participants had missing engagement data in the DMHI group. None of the results changed based on the missing data model used. Therefore, we present results using an imputed dataset following the model with the best performance (Multimedia Appendix 1). Finally, we opted not to include per-protocol analyses as originally planned. Instead, we only conducted and reported intention-to-treat analyses to provide a more accurate estimate of DMHI effectiveness.

Results

The flowchart of participants' recruitment, allocation, and analysis is presented in Figure 1.

Table 1 shows the demographics of the final sample (N=155).

XSL•FO
RenderX

Figure 1. Study design flowchart.

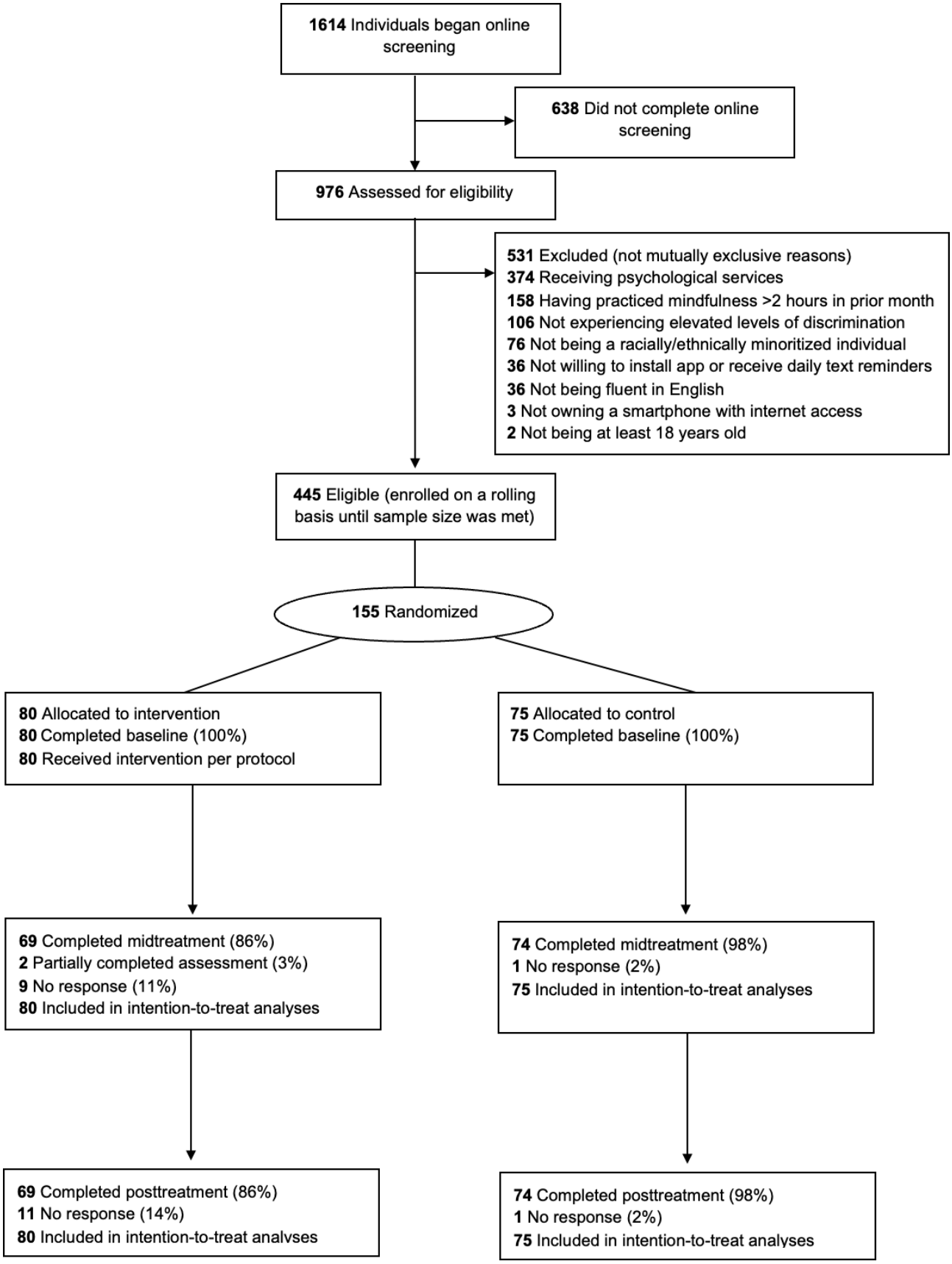

XSL•FO
RenderX

Table . Baseline characteristics of participants in the TAU[a] group, DMHI[b] group, and total sample.

Variable	TAU (n=75)	DMHI (n=80)	Total (N=155)
Age (years), mean (SD)	27.49 (10.1)	27.09 (9.2)	27.28 (9.6)
Race and ethnicity, n (%)			
Latinx	29 (38.7)	31 (38.8)	60 (38.7)
Asian	28 (37.3)	30 (37.5)	58 (37.4)
Black	9 (12)	9 (11.2)	18 (11.6)
Multiracial	7 (9.3)	9 (11.2)	16 (10.3)
Native American	2 (2.7)	1 (1.3)	3 (2)
Gender, n (%)			
Female	59 (78.6)	69 (86.2)	128 (82.6)
Male	11 (14.7)	11 (13.8)	22 (14.2)
Nonbinary	5 (6.7)	0 (0)	5 (3.2)
Educational level, n (%)			
Less than high school	0 (0)	1 (1.3)	1 (0.6)
High school or General Educational Development	10 (13.3)	4 (5)	14 (9)
Trade school	0 (0)	1 (1.3)	1 (0.6)
Some college, no degree	18 (24)	27 (33.8)	45 (29)
College	21 (28)	25 (31.2)	46 (29.8)
Graduate degree	26 (34.7)	22 (27.4)	48 (31)
Occupation, n (%)			
Unemployed	1 (1.3)	2 (2.5)	3 (2)
Employed	27 (36)	38 (47.5)	65 (41.9)
Student	46 (61.4)	40 (50)	86 (55.5)
Retired	1 (1.3)	0 (0)	1 (0.6)
Relationship status, n (%)			
Single	45 (60)	43 (53.7)	88 (56.8)
In a long-term relationship	27 (36)	36 (45)	63 (40.6)
Separated or divorced	3 (4)	1 (1.3)	4 (2.6)
Financial stress, n (%)			
Yes	14 (18.7)	10 (12.5)	24 (15.5)
No	61 (81.3)	70 (87.5)	131 (84.5)
Previous mindfulness experience, n (%)			
Yes	33 (44)	41 (51.2)	74 (47.7)
No	42 (56)	39 (48.8)	81 (52.3)
MDM[c], mean (SD)	19.29 (4.7)	19.25 (4.6)	19.27 (4.6)
PSS[d], mean (SD)	20.15 (6)	21.55 (5)	20.87 (5.5)
GAD-7[e], mean (SD)	7.91 (4.5)	9.12 (4.6)	8.53 (4.6)
PHQ-9[f,g], mean (SD)	8.96 (5.5)	9.95 (5.5)	9.4 (5.5)

[a]TAU: treatment as usual.

[b]DMHI: digital mental health intervention.

[c]MDM: Multicultural Discrimination Module.

XSL•FO
RenderX

[d]PSS: Perceived Stress Scale.

[e]GAD-7: General Anxiety Disorder-7.

[f]PHQ-9: Patient Health Questionnaire-9.

[g] PHQ-8 scores were prorated by multiplying them by 1.125 such that they were in the same score range as those in the PHQ-9.

In the model for stress, there was a significant group-by-time interaction showing a greater decrease in stress among participants in the DMHI group compared with participants in the TAU group at posttreatment (β=−4.52, 95% CI −6.54 to−2.51, SE 1.03, t_{7511}=−4.40; FDR-adjusted *P*<.001). By posttreatment, 61% (49/80) of participants in the DMHI group achieved the 4-point MCID reduction, compared with 23% (17/75) in the TAU group. As seen in Figure 2, these differences in change over time emerged from baseline to midtreatment (β=−2.41, 95% CI −4.14 to −0.67, SE 0.89, t_{10441}=−2.72; FDR-adjusted *P*=.01; MCID in DMHI group=50%, 40/80, MCID in TAU group=29%, 22/75) and continued from midtreatment to posttreatment (β=−2.11, 95% CI −3.90 to –0.32, SE 0.91, t_{1977}=−2.32; FDR-adjusted *P*=.03; MCID in DMHI group=37%, 30/80, MCID in TAU group=21%, 16/75). The total variance (pseudo-R^2) explained by the model was 0.41.

Figure 2. Trajectories of change in stress symptoms across groups. PSS: Perceived Stress Scale.

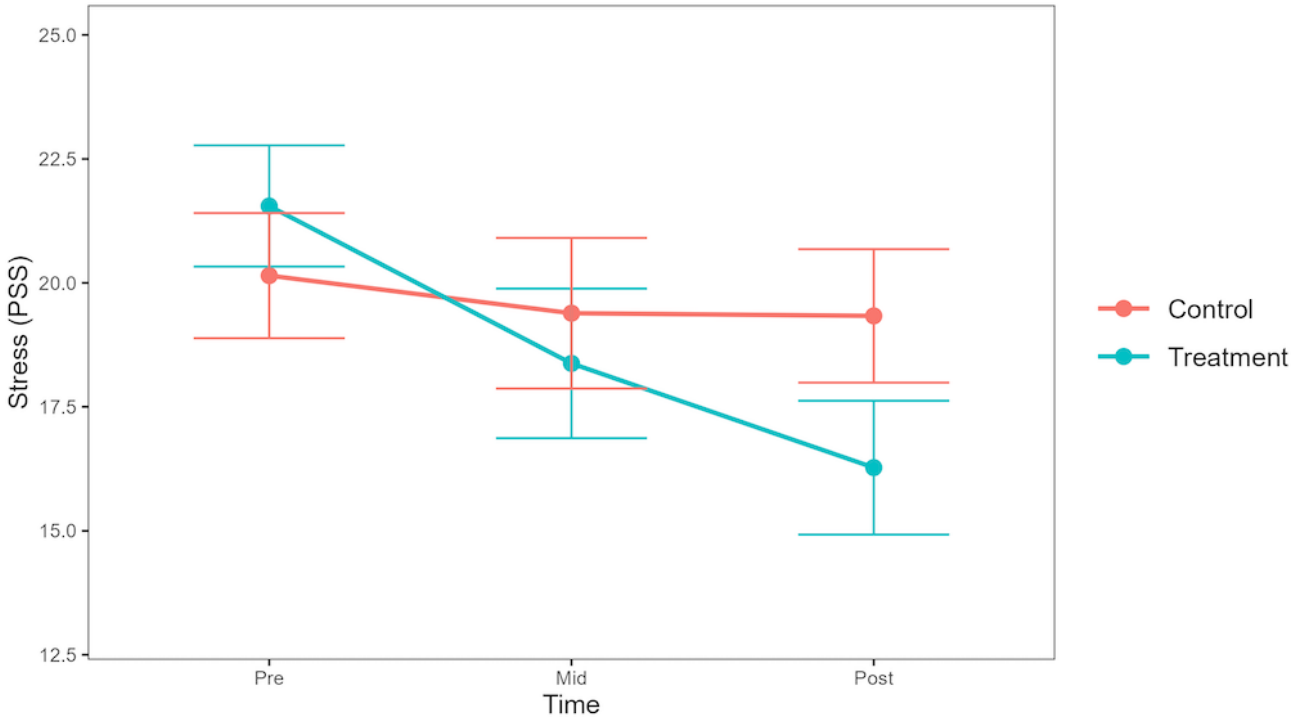

In the model for anxiety, there was a significant group-by-time interaction showing a greater decrease in anxiety among participants in the DMHI group compared with participants in the TAU group at posttreatment (β=−3.31, 95% CI −4.89 to −1.74, SE 0.80, t_{22128}=−4.13; FDR-adjusted *P*<.001). By posttreatment, 48% (38/80) of participants in the DMHI group achieved the 3-point MCID reduction, compared with 27% (20/75) in the TAU group. As seen in Figure 3, these differences in change over time emerged from baseline to midtreatment (β=−2.31, 95% CI −3.77 to –0.85, SE 0.75, t_{6191}=−3.10; FDR-adjusted *P*=.005; MCID in DMHI group=50%, 40/80, MCID in TAU group=22%, 17/75), and the reduction in symptoms continued the same trajectory from midtreatment to posttreatment but without reaching statistical significance (β=−1.00, 95% CI −2.47 to 0.47, SE 0.75, t_{4626}=−1.33; FDR-adjusted *P*=.25; MCID in DMHI group=30%, 24/80, MCID in TAU group=22%, 17/75). The total variance (pseudo-R^2) explained by the model was 0.45.

XSL•FO
RenderX

Figure 3. Trajectories of change in anxiety symptoms across groups. GAD-7: General Anxiety Disorder-7.

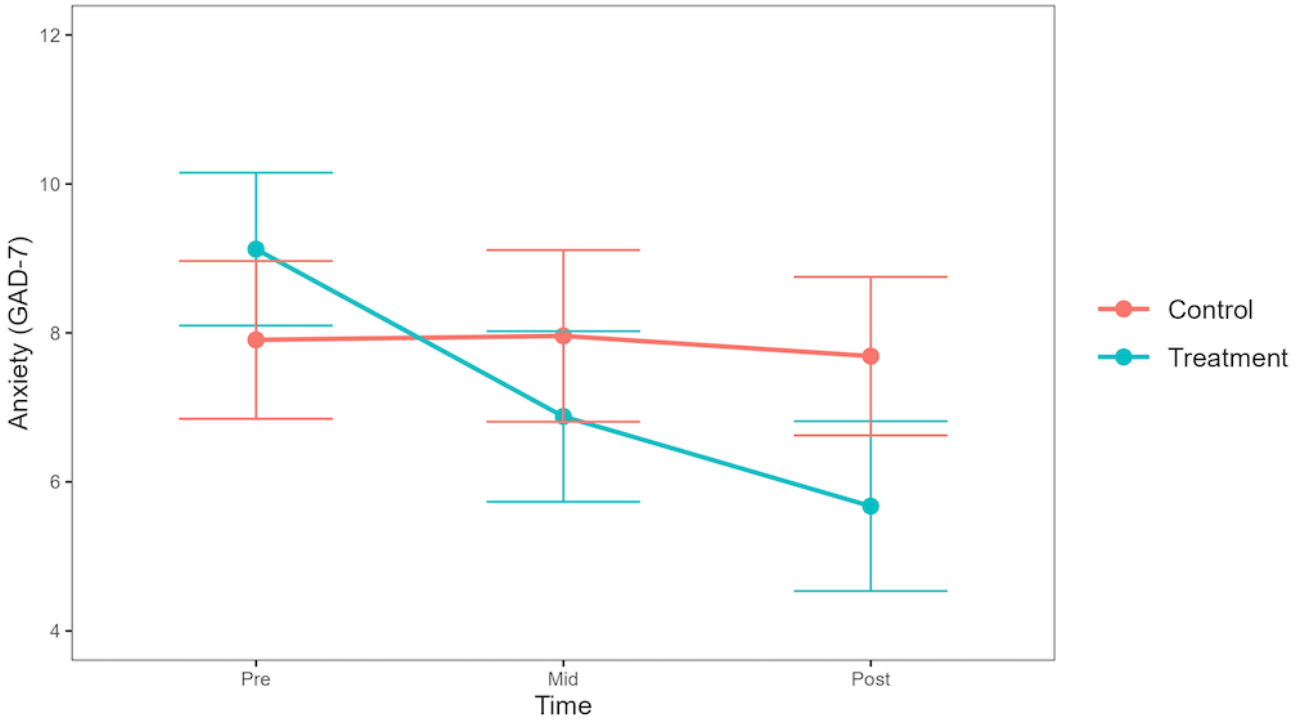

In the model for depression, there was a significant group-by-time interaction showing a greater decrease in depression among participants in the DMHI group compared with participants in the TAU group at posttreatment (β=−2.84, 95% CI −4.57 to −1.12, SE 0.88, t_{11660}=−3.23; FDR-adjusted *P*=.004). By posttreatment, 54% (43/80) of participants in the DMHI group achieved the 3-point MCID reduction, compared with 31% (23/75) in the TAU group. As seen in Figure 4, these differences in change over time emerged from baseline to midtreatment (β=−1.86, 95% CI −3.38 to –0.33, SE 0.78, t_{8235}=−2.38; FDR-adjusted *P*=.04; MCID in DMHI group=46%, 37/80, MCID in TAU group=24%, 18/75), and the reduction in symptoms continued the same trajectory from midtreatment to posttreatment but without reaching statistical significance (β=−0.99, 95% CI −2.52 to 0.54, SE 0.78, t_{7566}=−1.26; FDR-adjusted *P*=.24; MCID in DMHI group=39%, 31/80, MCID in TAU group=30%, 23/75). The total variance (pseudo-R^2) explained by the model was 0.53. Table 2 shows descriptive statistics of clinical outcomes at each assessment point.

Figure 4. Trajectories of change in depressive symptoms across groups. PHQ-9: Patient Health Questionnaire-9.

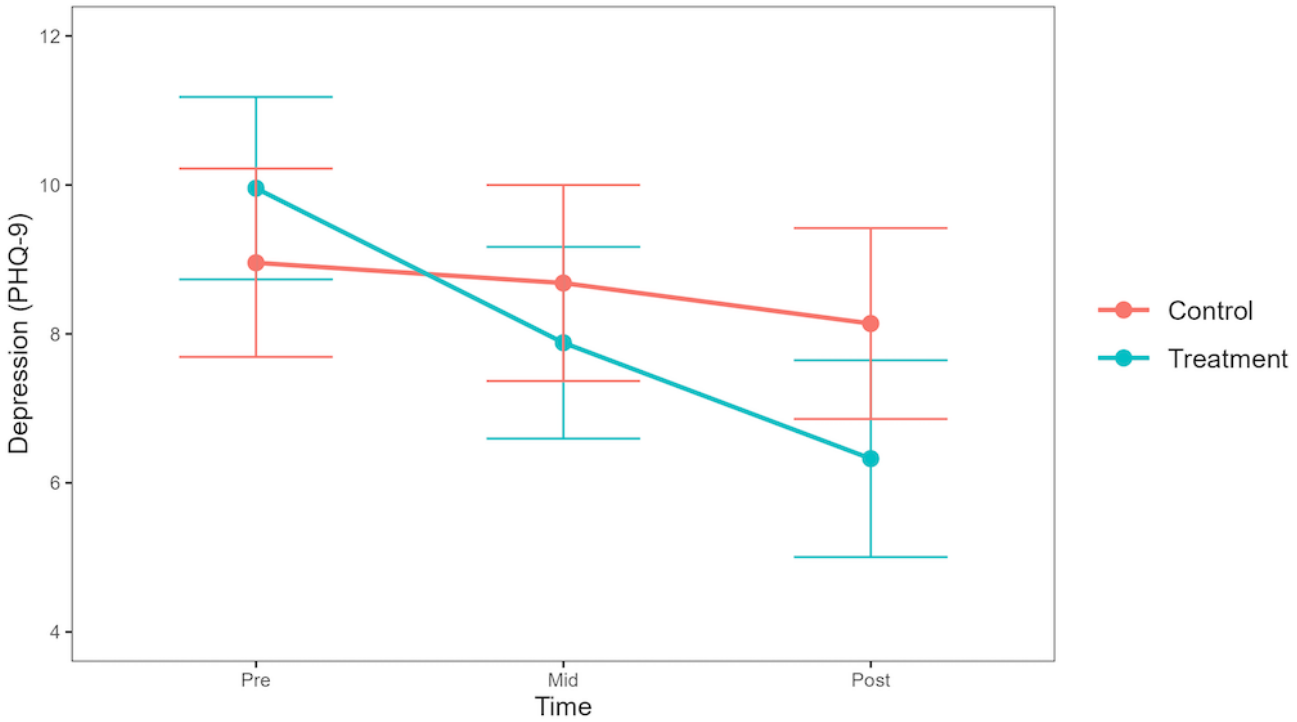

XSL•FO
RenderX

Table . Descriptive statistics of clinic outcomes with missing data imputed.

Outcome and group	Baseline, mean (SD)	Midtreatment, mean (SD)	Posttreatment, mean (SD)
Stress (PSS)[a] (range 0 - 40)			
Control	20.15 (6)	19.39 (6.3)	19.33 (6.1)
Intervention	21.55 (5)	18.38 (6.8)	16.28 (5.6)
Whole sample	20.87 (5.5)	18.87 (6.6)	17.76 (6)
Anxiety (GAD-7)[b] (range 0 - 21)			
Control	7.91 (4.5)	7.96 (5.3)	7.69 (5)
Intervention	9.12 (4.6)	6.88 (4.7)	5.67 (4.2)
Whole sample	8.54 (4.6)	7.40 (5)	6.65 (4.7)
Depression (PHQ-9)[c,d] (range 0 - 27)			
Control	8.96 (5.5)	8.68 (5.8)	8.14 (5.9)
Intervention	9.96 (5.5)	7.88 (5.5)	6.33 (5.2)
Whole sample	9.47 (5.5)	8.27 (5.7)	7.20 (5.6)

[a]PSS: Perceived Stress Scale.

[b]GAD-7: Generalized Anxiety Disorder-7.

[c]PHQ-9: Patient Health Questionnaire-9.

[d]PHQ-8 scores were prorated by multiplying them by 1.125 such that they were in the same score range as those in the PHQ-9.

In terms of feasibility, all participants in the DMHI group (80/80, 100%) downloaded the app and created a user profile. By the end of the 4-week program, participants used the app for a mean of 16.83 (SD 7.8; range 0 - 28) days, completed a mean of 24.85 (SD 23.3; range 0 - 28) meditations, and meditated for a mean of 238.67 (SD 260.7; range 0 - 2195) minutes. Only 2% of participants (2/80) did not use the DMHI at all. Participants used the app for a mean of 9.01 (SD 3.6; range 0 - 14) days, completed a mean of 13.27 (SD 7.7; range 0 - 52) meditations, and meditated for a mean of 113.76 (SD 90.5; range 0 - 652) minutes from baseline to midtreatment. During these first 2 weeks of the program, 2% of participants (2/80) did not use the DMHI at all. Participants also used the app for a mean of 7.85 (SD 4.7; range 0 - 14) days, completed a mean of 11.58 (SD 17.2; range 0 - 156) meditations, and meditated a mean of 124.90 (SD 179.7; range 0 - 1543) minutes from midtreatment to posttreatment. During these last 2 weeks of the program, 7% of participants (6/80) did not use the DMHI at all. Only 8% of participants (12/80) dropped out of this study. Of note, 11 of them (11/80, 14%) came from the DMHI group. In terms of program satisfaction, participants reported being satisfied with the DMHI program at both posttreatment (program satisfaction mean 23.14, SD 4.9) and midtreatment (program satisfaction mean 21.97, SD 5.1). No participant in the TAU group (0/80, 0%) reported accessing professional mental health services or a commercially available DMHI during the 4-week program.

Discussion

Principal Findings

The goal of this study was to examine the effectiveness and feasibility of a self-guided app-based DMHI for REMs who experience significant levels of discrimination. Results showed that the Mind-Us program was effective in reducing symptoms of stress, anxiety, and depression in as little as 2 weeks. Crucially, all participants assigned to the intervention group accessed the DMHI, engagement was consistent throughout 4 weeks, the dropout rate was low, and participants perceived the program as satisfactory. Taken together, these findings suggest that DMHIs can target the unique mental health needs of REMs. As such, low-intensity DMHI approaches hold significant promise to provide REMs with evidence-based treatments to cope with the deleterious mental health consequences of experiencing discrimination.

Based on results from 158 studies of traditional mindfulness interventions [8,10] and 43 studies of mindfulness-based DMHIs [22], this RCT is one of the first studies to support the effectiveness of a low-intensity mindfulness-based DMHI in reducing discrimination-related mental health symptoms across multiple REM groups. Consistent with our hypothesis, REMs who received the DMHI experienced significant short-term reductions in stress, anxiety, and depression. Although not in the context of experiencing discrimination and mostly with White samples, these results align with previous mindfulness-based DMHIs, showing these interventions are effective in targeting these mental health problems [22]. Notably, these reductions in stress, anxiety, and depression are not only statistically significant but also clinically meaningful. By the end of treatment, 61%, 48%, and 54% of participants experienced a decrease in stress, anxiety, and depression, respectively, that can be considered clinically significant according to empirically established cutoffs [58,59]. Considering that most mindfulness-based DMHI trials to date have failed to examine whether treatment effects translate into meaningful clinical changes [22], this study contributes to the mindfulness and DMHI literatures by demonstrating that this approach is effective for REMs exposed to discrimination. Taken together,

XSL•FO
RenderX

these results help establish self-guided mindfulness-based DMHIs as a potentially effective approach to help REM groups cope with mental health problems associated with being exposed to discrimination.

Another important finding was that the Mind-Us program led to significant symptom reductions even after only 2 weeks of treatment. Most mindfulness-based DMHIs to date have examined high-intensity programs lasting 8 weeks or longer, requiring intensive meditation practice, or including some contact with a mindfulness facilitator [22]. In contrast, this DMHI yielded significant improvement in symptoms after 2 weeks of using the mindfulness app for less than 10 minutes a day and without the support of a mindfulness facilitator. Further, these reductions in stress, anxiety, and depressive symptoms were clinically meaningful for approximately 50% of participants who received the DMHI by MCID benchmarks [58,59]. This treatment response pattern is consistent with other mindfulness-based DMHIs, showing that most of the clinical gains seem to be made in the first 10 days of treatment [64]. This treatment response is especially encouraging as DMHIs that involve a low time commitment may be able to engage vulnerable populations who are unable to participate in more intensive mental health services due to competing demands and logistical barriers.

Even if participants experienced significant treatment gains in the first 2 weeks, the Mind-Us program led to even greater reductions in symptoms after 4 weeks of treatment, which suggests that additional mindfulness practice is important to promote further improvement. An important caveat is that, except for stress symptoms that continued to decrease, participants did not experience significant reductions in anxiety or depressive symptoms from midtreatment to posttreatment (ie, from wk 2 to wk 4). As previously discussed, other mindfulness-based DMHIs have found that improvements in symptoms occur at a slower rate after 10 days of treatment [64]. The difference in treatment response for stress symptoms versus anxiety and depressive symptoms could be due to the nature of these mental health problems. Stress symptoms seem to be more short-lived and situational compared with anxiety and depressive symptoms that are more pervasive and persistent [65]. Therefore, stress may be a more modifiable clinical target in DMHIs. In fact, mindfulness-based DMHIs often lead to larger treatment effect sizes when targeting stress compared with anxiety or depression [8,66]. Together, our findings suggest that there may be varying durations of mindfulness-based DMHIs that could be beneficial for REMs, especially based on the clinical target (ie, stress vs anxiety or depression). Future work still needs to determine the optimal mindfulness dosage in DMHIs that maximizes symptom reduction while maintaining the tolerability of the program, whether dosage optimization differs across subgroups (eg, different REM groups and individuals with different symptomatology profiles), and potential dose-response interactions.

Notably, the Mind-Us program used a commercially available mindfulness app not designed specifically for REMs, suggesting that non–culturally adapted DMHIs can be effective and well-accepted among these groups. Showing proof of DMHI effectiveness and program satisfaction among REM groups is essential, considering ongoing concerns about potential mismatches between mindfulness interventions and the cultural background of REMs [11], the lack of DMHI content that responds to the unique needs of REMs [67], and the need (or not) to culturally adapt DMHIs [15,18]. The results of this trial support using DMHIs with a strong evidence base and high treatment satisfaction profiles, even if not previously established for specific REM populations. This is an approach that can advance mental health equity more immediately than waiting for REM-specific DMHIs to be developed [15,39]. While DMHIs often fail to reach REMs, when they do, these programs are effective and well accepted, regardless of whether they are culturally adapted [15,18]. In fact, emerging research suggests that culturally adapted DMHIs do not outperform their original counterparts [68], raising questions about the need to undergo this complex, time-consuming, and resource-intensive process [15,18]. Therefore, future DMHI studies must collect information on who accesses these tools, engages with the content, and benefits from the intervention, with an emphasis on identifying real inequities (eg, limited uptake among Black individuals, poor engagement among Latinx individuals, more attenuated reductions in symptoms among Asian individuals) and using that data to inform precise adaptations as needed [15,39]. In traditional mental health services, this data-driven approach to culturally adapting treatments has proven more effective than assuming a priori adaptations are needed [69].

An important caveat regarding the effectiveness of non-culturally adapted DMHIs for REMs is that the Mind-Us program still leveraged previously proposed strategies to improve the cultural robustness of DMHIs through human support [18]. During the onboarding procedure, study staff discussed potential perceived mismatches between the content of the DMHI and the participant's cultural background and identity, provided psychoeducation about mindfulness and DMHIs, and helped the participant see how mindfulness could be helpful in dealing with discrimination-related stressors. Since few researchers and clinicians are in a position to culturally adapt existing DMHIs or design new ones for specific REM groups, using human support to improve the cultural fit of DMHIs represents an alternative with significant implementation and dissemination potential. Onboarding procedures, weekly check-ins, and wrap-up strategies could be used as a flexible, idiographic, and personalized approach to address the current shortcomings of commercially available DMHIs when used with REM populations [18,70]. The potential of this approach is supported by numerous studies showing that DMHIs with some form of human support have better engagement [71] and larger treatment effects [72] compared with completely self-guided DMHIs. Regardless of their promise, human support models need to be further investigated through dismantling studies to determine which components of these implementation strategies drive their effects (eg, technical troubleshooting, motivational strategies, or discussions of cultural fit).

In contrast with the poor uptake found in previous app-based DMHI trials and real-world cases, we found a promising adoption of this program. This DMHI had an uptake rate of 100%, which contrasts with average uptake rates of 45% in RCTs [24] and 21% in real-world settings [25]. We attribute

XSL•FO
RenderX

this success to the short onboarding call at the beginning of the intervention, in which participants were assisted in downloading the app, learning how to use it, and troubleshooting any technical issues. In designing this onboarding procedure, we followed recommendations to address common barriers faced by REMs when using DMHIs [18]. The effectiveness of onboarding procedures is supported by a previous study showing that a similar approach was effective in addressing barriers related to digital literacy and promoting DMHI uptake among REMs [73]. Thus, onboarding procedures that address issues with digital literacy show promise to promote uptake among REMs using DMHIs. However, this interpretation needs to be considered cautiously, given that our study design did not allow us to determine whether this high uptake was truly driven by the onboarding procedure or some other aspect or combination of factors in this DMHI program.

Participants in the DMHI group also used the app to complete approximately 1 session a day for a total of 1 hour of meditation per week for 4 weeks, reporting high program satisfaction. This consistent engagement and high program satisfaction throughout the program are significantly different from app usage patterns found in previous RCTs, where users only use the app a couple of times early in treatment [24], and app usage in the real world, where only 4% of participants continue using these digital tools after 2 weeks [26]. Similarly, 25% of participants in RCTs [24] and 96% of real-world app users [26] drop out after 4 weeks. Engagement in DMHI is crucial because it is often associated with better treatment outcomes [15]. These consistent engagement, low dropout, and high treatment satisfaction are especially encouraging because REMs face significant barriers to engaging with face-to-face mental health services [17,74]. Considering recent work showing that DMHI users also face similar barriers to those of traditional mental health services [27,28], this program shows promise to circumvent these problems, allowing REMs to engage in and benefit from mental health services via DMHIs. Despite our inability to infer the exact factors driving these effects, we believe that the use of multiple engagement-enhancement strategies, such as daily text messages [40], an onboarding procedure with study staff [18], and a short program length [40] could have contributed to these outcomes. Among these strategies, the onboarding procedure might have been especially beneficial for REMs, as participants problem-solved issues that could interfere with their everyday use of the app in advance. The potential effectiveness of providing support at the beginning of DMHI programs aligns with studies showing that app-based DMHIs with some type of human guidance tend to have better engagement and lower dropout compared with fully self-guided DMHIs [24]. Future studies are needed to determine the most effective engagement-enhancement strategies (eg, human support vs automated reminders), the mechanisms by which these strategies affect engagement and dropout (eg, self-efficacy), factors promoting program satisfaction (eg, therapeutic alliance with human supporter), and the specific cost-benefit of supplementing DMHIs via human support (eg, fully self-guided DMHIs being cheaper but leading to significantly smaller treatment effect sizes than guided DMHIs).

Limitations and Future Directions

Findings in this study should be interpreted in the context of existing limitations. First, this study was conducted from 2020 to 2022, during the COVID-19 pandemic, nationwide lockdowns, and a social and political environment in the United States that made discrimination against REM groups more common and prominent. This unique intersection of conditions was likely to influence participants' mental health needs, willingness to participate in this trial, and motivation to stay engaged with the DMHI. Another limitation is that this RCT did not include a measure of discrimination-based distress as an outcome. Instead, we considered high exposure to discrimination as a "risk factor" associated with increases in mental health symptoms [3]. While enrolling "at risk" samples is a common practice in DMHI research, we recognize that discrimination-based distress may be a meaningful clinical outcome over and above stress, anxiety, and depression. Although our sample was uniquely diverse in terms of race and ethnicity, participants were still mostly young, highly educated, English-speaking females. DMHI research to date has been dominated by these types of samples [15], and additional efforts are needed to recruit a more diverse set of participants across age, socioeconomic status, language preference, and gender. For instance, partnering with trusted community members (eg, pastors) and key organizations (eg, schools) [15] and using recruitment materials that directly target populations underrepresented in DMHI research [75] have proven effective strategies for engaging more diverse populations in DMHI trials. In this study, we also relied on participants' willingness to share app usage data with the research team, and future work can benefit from obtaining app usage data directly from DMHIs to obtain more accurate estimates of treatment engagement. Similarly, information on participants' reasons for stopping the use of the DMHI was not collected. Future studies examining the experiences of REM users qualitatively are warranted to better understand DMHI engagement and attrition among these groups. Even though our sample size and the number of assessments were similar to previous app-based mindfulness DMHI trials [22], future studies with larger sample sizes and long-term follow-ups are necessary to replicate these results, potentially identify smaller treatment effect sizes, and elucidate the duration of treatment effects. A final limitation is the lack of an active control condition, as none of the participants randomized to TAU received mental health services during the duration of the trial, effectively making this control condition a waitlist. Although considering the current state of the literature in app-based mindfulness DMHIs for REMs may minimize this concern [22,41], active control conditions are necessary to advance both mindfulness and DMHI effectiveness research.

Conclusions

This study shows that self-guided DMHIs that combine commercially available digital tools with some human support and automated text messages hold significant promise to advance mental health equity for REMs. Among participants who received this non-culturally adapted DMHI, 48% - 61% experienced a clinically significant reduction in symptoms by the end of treatment, suggesting that low-intensity DMHIs can improve the well-being of REMs exposed to discrimination.

XSL•FO
RenderX

Futher supporting the promise of this program, all participants accessed the DMHI, used it consistently throughout 4 weeks, reported high treatment satisfaction, and only a small number left the intervention prematurely. Crucially, the dissemination and implementation potential of this approach is substantial, considering the recent introduction of new Centers for Medicare and Medicaid Services codes (ie, G0552, G0553, and G0554) that has made both DMHIs and the human support necessary for their deployment eligible for insurance reimbursement, effectively creating the financial infraestructure necessary for the long-term viability of hybrid models of care in the United States.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the writing of this manuscript. The following tasks were delegated to GAI tools under full human supervision: proofreading and editing. The GAI tool used was Grammarly.

Funding

This study was supported by funding from the APAGS Psychological Science Research Grant, the APF COGDOP Graduate Student Scholarship, the APF Queen-Nellie Evans Scholarship, the UCLA Eugene V. Cota-Robles Fellowship, the Ford Foundation Predoctoral and Postdoctoral Fellowships, and the UC Chancellor's Postdoctoral Fellowship awarded to Giovanni Ramos. Research reported in this publication was also supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under award number T32HL166114 awarded to Giovanni Ramos. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The dataset analyzed during this study as well as statistical analyses code are available on GitHub [76].

Authors' Contributions

Conceptualization: GR, DC, AL, AM, AG
Data curation: GR, YW, AM, CE
Formal analysis: GR, AM, CE
Funding acquisition: GR, DC
Investigation: GR, YW
Methodology: GR, DC, AL, AM, AG
Project administration: GR, YW
Resources: GR, DC
Software: GR, AM, CE
Supervision: GR, DC
Validation: GR, AM
Visualization: GR, AM
Writing–original draft: GR
Writing–review & editing: GR, DC, AM, AL, AG

Conflicts of Interest

None declared.

Multimedia Appendix 1
Missing data analytic plan.
[DOCX File, 42 KB - jmir_v28i1e84328_app1.docx]

Checklist 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File, 1160 KB - jmir_v28i1e84328_app2.pdf]

References

1. Lee RT, Perez AD, Boykin CM, Mendoza-Denton R. On the prevalence of racial discrimination in the United States. PLoS One 2019;14(1):e0210698. [doi: 10.1371/journal.pone.0210698] [Medline: 30629706]
2. Pascoe EA, Smart Richman L. Perceived discrimination and health: a meta-analytic review. Psychol Bull 2009 Jul;135(4):531-554. [doi: 10.1037/a0016059] [Medline: 19586161]

XSL•FO
RenderX

3. Carter RT, Johnson VE, Kirkinis K, Roberson K, Muchow C, Galgay C. A meta-analytic review of racial discrimination: relationships to health and culture. Race Soc Probl 2019 Mar;11(1):15-32. [doi: 10.1007/s12552-018-9256-y]
4. Vilsaint CL, NeMoyer A, Fillbrunn M, et al. Racial/ethnic differences in 12-month prevalence and persistence of mood, anxiety, and substance use disorders: variation by nativity and socioeconomic status. Compr Psychiatry 2019 Feb;89:52-60. [doi: 10.1016/j.comppsych.2018.12.008] [Medline: 30594752]
5. Martinez JH, Suyemoto KL, Abdullah T, Burnett-Zeigler I, Roemer L. Mindfulness and valued living in the face of racism-related stress. Mindfulness (N Y) 2022 May;13(5):1112-1125. [doi: 10.1007/s12671-022-01826-6] [Medline: 36969085]
6. Kabat-Zinn J. Coming to Our Senses: Healing Ourselves and the World Through Mindfulness: Hachette; 2005.
7. Gu J, Strauss C, Bond R, Cavanagh K. How do mindfulness-based cognitive therapy and mindfulness-based stress reduction improve mental health and wellbeing? A systematic review and meta-analysis of mediation studies. Clin Psychol Rev 2015 Apr;37:1-12. [doi: 10.1016/j.cpr.2015.01.006] [Medline: 25689576]
8. Galante J, Friedrich C, Dawson AF, et al. Mindfulness-based programmes for mental health promotion in adults in nonclinical settings: a systematic review and meta-analysis of randomised controlled trials. PLoS Med 2021 Jan;18(1):e1003481. [doi: 10.1371/journal.pmed.1003481] [Medline: 33428616]
9. Waldron EM, Hong S, Moskowitz JT, Burnett-Zeigler I. A systematic review of the demographic characteristics of participants in US-based randomized controlled trials of mindfulness-based interventions. Mindfulness (N Y) 2018 Dec;9(6):1671-1692. [doi: 10.1007/s12671-018-0920-5]
10. Sun S, Goldberg SB, Loucks EB, Brewer JA. Mindfulness-based interventions among people of color: a systematic review and meta-analysis. Psychother Res 2022 Mar;32(3):277-290. [doi: 10.1080/10503307.2021.1937369] [Medline: 34098859]
11. Proulx J, Croff R, Oken B, et al. Considerations for research and development of culturally relevant mindfulness interventions in American minority communities. Mindfulness (N Y) 2018 Apr;9(2):361-370. [doi: 10.1007/s12671-017-0785-z] [Medline: 29892321]
12. Hall GCN, Hong JJ, Zane NWS, Meyer OL. Culturally competent treatments for Asian Americans: the relevance of mindfulness and acceptance - based psychotherapies. Clin Psychol Sci Pract 2011 Sep;18(3):215-231. [doi: 10.1111/j.1468-2850.2011.01253.x]
13. Biggers A, Spears CA, Sanders K, Ong J, Sharp LK, Gerber BS. Promoting mindfulness in African American communities. Mindfulness (N Y) 2020 Oct;11(10):2274-2282. [doi: 10.1007/s12671-020-01480-w] [Medline: 33584869]
14. Castellanos R, Yildiz Spinel M, Phan V, Orengo-Aguayo R, Humphreys KL, Flory K. A systematic review and meta-analysis of cultural adaptations of mindfulness-based interventions for Hispanic populations. Mindfulness (N Y) 2020 Feb;11(2):317-332. [doi: 10.1007/s12671-019-01210-x]
15. Ramos G, Hernandez-Ramos R, Taylor M, Schueller SM. State of the science: using digital mental health interventions to extend the impact of psychological services. Behav Ther 2024 Nov;55(6):1364-1379. [doi: 10.1016/j.beth.2024.04.004] [Medline: 39443071]
16. Olano HA, Kachan D, Tannenbaum SL, Mehta A, Annane D, Lee DJ. Engagement in mindfulness practices by U.S. adults: sociodemographic barriers. J Altern Complement Med 2015 Feb;21(2):100-102. [doi: 10.1089/acm.2014.0269] [Medline: 25685958]
17. Mojtabai R, Olfson M, Sampson NA, et al. Barriers to mental health treatment: results from the National Comorbidity Survey Replication. Psychol Med 2011 Aug;41(8):1751-1761. [doi: 10.1017/S0033291710002291] [Medline: 21134315]
18. Ramos G, Chavira DA. Use of technology to provide mental health care for racial and ethnic minorities: evidence, promise, and challenges. Cogn Behav Pract 2022 Feb;29(1):15-40. [doi: 10.1016/j.cbpra.2019.10.004]
19. Internet, broadband fact sheet. Pew Research Center. 2025. URL: https://www.pewresearch.org/internet/fact-sheet/internet-broadband/ [accessed 2026-04-21]
20. Mobile fact sheet. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2025-09-16]
21. Linardon J, Torous J, Firth J, Cuijpers P, Messer M, Fuller-Tyszkiewicz M. Current evidence on the efficacy of mental health smartphone apps for symptoms of depression and anxiety. A meta-analysis of 176 randomized controlled trials. World Psychiatry 2024 Feb;23(1):139-149. [doi: 10.1002/wps.21183] [Medline: 38214614]
22. Linardon J, Messer M, Goldberg SB, Fuller-Tyszkiewicz M. The efficacy of mindfulness apps on symptoms of depression and anxiety: an updated meta-analysis of randomized controlled trials. Clin Psychol Rev 2024 Feb;107:102370. [doi: 10.1016/j.cpr.2023.102370] [Medline: 38056219]
23. De Jesús-Romero R, Holder-Dixon AR, Buss JF, Lorenzo-Luaces L. Race, ethnicity, and other cultural background factors in trials of internet-based cognitive behavioral therapy for depression: systematic review. J Med Internet Res 2024 Feb 1;26:e50780. [doi: 10.2196/50780] [Medline: 38300699]
24. Linardon J. Rates of attrition and engagement in randomized controlled trials of mindfulness apps: systematic review and meta-analysis. Behav Res Ther 2023 Nov;170:104421. [doi: 10.1016/j.brat.2023.104421] [Medline: 37862854]
25. Fleming T, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018 Jun 6;20(6):e199. [doi: 10.2196/jmir.9275] [Medline: 29875089]

XSL•FO
RenderX

26. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567. [doi: 10.2196/14567] [Medline: 31573916]
27. Ramos G, Montoya AK, Hammons HR, Smith D, Chavira DA, Rith-Najarian LR. Digital Intervention Barriers Scale-7 (DIBS-7): development, evaluation, and preliminary validation. JMIR Form Res 2023 Apr 6;7:e40509. [doi: 10.2196/40509] [Medline: 37023417]
28. Starvaggi I, Lorenzo-Luaces L. Psychotherapy access barriers and interest in digital mental health interventions among adults with treatment needs: survey study. JMIR Ment Health 2025 Apr 1;12:e65356. [doi: 10.2196/65356] [Medline: 40168039]
29. Lipschitz JM, Pike CK, Hogan TP, Murphy SA, Burdick KE. The engagement problem: a review of engagement with digital mental health interventions and recommendations for a path forward. Curr Treat Options Psychiatry 2023 Sep;10(3):119-135. [doi: 10.1007/s40501-023-00297-3] [Medline: 38390026]
30. Boucher EM, Raiker JS. Engagement and retention in digital mental health interventions: a narrative review. BMC Digit Health 2024 Aug 8;2(1):52. [doi: 10.1186/s44247-024-00105-9]
31. Rushin S, Edwards GS. The effect of President Trump's election on hate crimes. SSRN J 2018. [doi: 10.2139/ssrn.3102652]
32. Chae DH, Yip T, Martz CD, et al. Vicarious racism and vigilance during the COVID-19 pandemic: mental health implications among Asian and Black Americans. Public Health Rep 2021;136(4):508-517. [doi: 10.1177/00333549211018675] [Medline: 34034574]
33. Curtis DS, Washburn T, Lee H, et al. Highly public anti-Black violence is associated with poor mental health days for Black Americans. Proc Natl Acad Sci U S A 2021 Apr 27;118(17):e2019624118. [doi: 10.1073/pnas.2019624118] [Medline: 33875593]
34. Eysenbach G, Group CE. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
35. Shariff-Marco S, Breen N, Landrine H, et al. Measuring everyday racial/ethnic discrimination in health surveys. Du Bois Rev 2011;8(1):159-177. [doi: 10.1017/S1742058X11000129]
36. Utsey SO. Development and validation of a short form of the Index of Race-Related Stress (IRRS)—Brief Version. Meas Eval Couns Dev 1999 Oct 1;32(3):149-167. [doi: 10.1080/07481756.1999.12068981]
37. Acabchuk RL, Simon MA, Low S, Brisson JM, Johnson BT. Measuring meditation progress with a consumer-grade EEG device: caution from a randomized controlled trial. Mindfulness (N Y) 2021 Jan;12(1):68-81. [doi: 10.1007/s12671-020-01497-1] [Medline: 34149957]
38. Huberty J, Eckert R, Larkey L, et al. Smartphone-based meditation for myeloproliferative neoplasm patients: feasibility study to inform future trials. JMIR Form Res 2019 Apr 29;3(2):e12662. [doi: 10.2196/12662] [Medline: 31033443]
39. Ramos G, Karyotaki E, Mohr DC, Schueller SM. Digital mental health treatments for depression. Psychiatric Clin North Am 2025 Oct. [doi: 10.1016/j.psc.2025.08.023]
40. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: a systematic and meta-analytic review. J Consult Clin Psychol 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
41. Freedland KE, King AC, Ambrosius WT, et al. The selection of comparators for randomized controlled trials of health-related behavioral interventions: recommendations of an NIH expert panel. J Clin Epidemiol 2019 Jun;110:74-81. [doi: 10.1016/j.jclinepi.2019.02.011] [Medline: 30826377]
42. Cohen S, Kamarck T, Mermelstein RJ. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [Medline: 6668417]
43. Lee EH. Review of the psychometric evidence of the Perceived Stress Scale. Asian Nurs Res (Korean Soc Nurs Sci) 2012 Dec;6(4):121-127. [doi: 10.1016/j.anr.2012.08.004] [Medline: 25031113]
44. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
45. Parkerson HA, Thibodeau MA, Brandt CP, Zvolensky MJ, Asmundson GJG. Cultural-based biases of the GAD-7. J Anxiety Disord 2015 Apr;31:38-42. [doi: 10.1016/j.janxdis.2015.01.005]
46. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x]
47. Wu Y, Levis B, Riehm KE, et al. Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis. Psychol Med 2020 Jun;50(8):1368-1380. [doi: 10.1017/S0033291719001314] [Medline: 31298180]
48. Patel JS, Oh Y, Rand KL, et al. Measurement invariance of the Patient Health Questionnaire-9 (PHQ-9) depression screener in U.S. adults across sex, race/ethnicity, and education level: NHANES 2005-2016. Depress Anxiety 2019 Sep;36(9):813-823. [doi: 10.1002/da.22940] [Medline: 31356710]
49. Smith JL, Allen JW, Haack C, et al. The impact of app-delivered mindfulness meditation on functional connectivity and self-reported mindfulness among health profession trainees. Mindfulness (N Y) 2021;12(1):92-106. [doi: 10.1007/s12671-020-01502-7] [Medline: 33052251]

XSL•FO
RenderX

50. Oei TPS, Green AL. The Satisfaction With Therapy and Therapist Scale--Revised (STTS-R) for group psychotherapy: psychometric properties and confirmatory factor analysis. Prof Psychol Res Pr 2008 Aug;39(4):435-442. [doi: 10.1037/0735-7028.39.4.435]
51. Littleton H, Grills AE, Kline KD, Schoemann AM, Dodd JC. The From Survivor to Thriver program: RCT of an online therapist-facilitated program for rape-related PTSD. J Anxiety Disord 2016 Oct;43:41-51. [doi: 10.1016/j.janxdis.2016.07.010] [Medline: 27513363]
52. Bischoff T, Anderson SR, Heafner J, Tambling R. Establishment of a reliable change index for the GAD-7. Psychol Community Health 2020 Aug;8(1):176-187. [doi: 10.5964/pch.v8i1.309]
53. Test review: Patient Health Questionnaire–9 (PHQ-9). Rehabil Couns Bull 2014 Jul;57(4):246-248. [doi: 10.1177/0034355213515305]
54. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
55. The R project for statistical computing. R Foundation. 2025. URL: https://www.R-project.org/ [accessed 2026-04-21]
56. Pinheiro JC, Bates DM. Mixed-Effects Models in S and S-PLUS: Springer-Verlag; 2000. [doi: 10.1007/b98882]
57. Mouelhi Y, Jouve E, Castelli C, Gentile S. How is the minimal clinically important difference established in health-related quality of life instruments? Review of anchors and methods. Health Qual Life Outcomes 2020 May 12;18(1):136. [doi: 10.1186/s12955-020-01344-w] [Medline: 32398083]
58. Eskildsen A, Dalgaard VL, Nielsen KJ, et al. Cross-cultural adaptation and validation of the Danish consensus version of the 10-item Perceived Stress Scale. Scand J Work Environ Health 2015 Sep 1;41(5):486-490. [doi: 10.5271/sjweh.3510] [Medline: 26111225]
59. Kroenke K, Wu J, Yu Z, et al. Patient Health Questionnaire Anxiety and Depression Scale: initial validation in three clinical trials. Psychosom Med 2016;78(6):716-727. [doi: 10.1097/PSY.0000000000000322] [Medline: 27187854]
60. Lorenzo-Luaces L, Howard J. Efficacy of an unguided, digital single-session intervention for internalizing symptoms in web-based workers: randomized controlled trial. J Med Internet Res 2023 Jul 7;25:e45411. [doi: 10.2196/45411] [Medline: 37418303]
61. Ramos G. App-based mindfulness meditation for people of color who experience race-related stress: a randomized controlled trial [dissertation]. : UCLA; 2023 URL: https://escholarship.org/content/qt5gn2g9gp/qt5gn2g9gp.pdf [accessed 2026-05-06]
62. Shaw M, Rights JD, Sterba SS, Flake JK. r2mlm: an R package calculating R-squared measures for multilevel models. Behav Res Methods 2023 Jun;55(4):1942-1964. [doi: 10.3758/s13428-022-01841-4] [Medline: 35798918]
63. Keller BT, Enders CK. Blimp User's Guide (Version 4) 2026. URL: https://www.appliedmissingdata.com/blimp [accessed 2026-04-21]
64. Champion L, Economides M, Chandler C. The efficacy of a brief app-based mindfulness intervention on psychosocial outcomes in healthy adults: a pilot randomised controlled trial. PLoS One 2018;13(12):e0209482. [doi: 10.1371/journal.pone.0209482] [Medline: 30596696]
65. Bystritsky A, Kronemyer D. Stress and anxiety. Psychiatr Clin North Am 2014 Dec;37(4):489-518. [doi: 10.1016/j.psc.2014.08.002]
66. Sommers-Spijkerman M, Austin J, Bohlmeijer E, Pots W. New evidence in the booming field of online mindfulness: an updated meta-analysis of randomized controlled trials. JMIR Ment Health 2021 Jul 19;8(7):e28168. [doi: 10.2196/28168] [Medline: 34279240]
67. Ramos G, Ponting C, Labao JP, Sobowale K. Considerations of diversity, equity, and inclusion in mental health apps: a scoping review of evaluation frameworks. Behav Res Ther 2021 Dec;147:103990. [doi: 10.1016/j.brat.2021.103990] [Medline: 34715396]
68. Balci S, Spanhel K, Sander LB, Baumeister H. Culturally adapting internet- and mobile-based health promotion interventions might not be worth the effort: a systematic review and meta-analysis. NPJ Digit Med 2022 Mar 23;5(1):34. [doi: 10.1038/s41746-022-00569-x] [Medline: 35322172]
69. Ramos G, Woller M, Quetsch L, et al. Trajectories of change in parent-child interaction therapy outcomes in Latinx families: implications for cultural adaptation. J Clin Child Adolesc Psychol 2025;54(6):684-698. [doi: 10.1080/15374416.2024.2395272] [Medline: 39287980]
70. Bear HA, Ayala Nunes L, Ramos G, et al. The acceptability, engagement, and feasibility of mental health apps for marginalized and underserved young people: systematic review and qualitative study. J Med Internet Res 2024 Jul 30;26:e48964. [doi: 10.2196/48964] [Medline: 39078699]
71. Musiat P, Johnson C, Atkinson M, Wilksch S, Wade T. Impact of guidance on intervention adherence in computerised interventions for mental health problems: a meta-analysis. Psychol Med 2022 Jan;52(2):229-240. [doi: 10.1017/S0033291721004621] [Medline: 34802474]
72. Werntz A, Amado S, Jasman M, Ervin A, Rhodes JE. Providing human support for the use of digital mental health interventions: systematic meta-review. J Med Internet Res 2023 Feb 6;25:e42864. [doi: 10.2196/42864] [Medline: 36745497]
73. Hernandez-Ramos R, Aguilera A, Garcia F, et al. Conducting internet-based visits for onboarding populations with limited digital literacy to an mHealth intervention: development of a patient-centered approach. JMIR Form Res 2021 Apr 29;5(4):e25299. [doi: 10.2196/25299] [Medline: 33872184]

XSL•FO
RenderX

74. Cummings JR, Allen L, Clennon J, Ji X, Druss BG. Geographic access to specialty mental health care across high- and Low-income US communities. JAMA Psychiatry 2017 May 1;74(5):476-484. [doi: 10.1001/jamapsychiatry.2017.0303] [Medline: 28384733]
75. Rith-Najarian L, Sun W, Chen A, et al. What's in a name? Branding of online mental health programming for university students. J Consult Clin Psychol 2019;87(4):380-391. [doi: 10.1037/ccp0000383.supp]
76. Mind-us intervention. GitHub. URL: https://github.com/Quantitative-Research-Collaboratory/Mind-Us-Intervention [accessed 2026-05-07]

Abbreviations

CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
DMHI: digital mental health intervention
FDR: false discovery rate
GAD-7: Generalized Anxiety Disorder-7
HIPAA: Health Insurance Portability and Accountability Act
IRRS-B: Index of Race-Related Stress-Brief
MCID: minimal clinically significant difference
MDM: Multicultural Discrimination Module
PHQ-8: Patient Health Questionnaire-8
PHQ-9: Patient Health Questionnaire-9
PSS: Perceived Stress Scale
RCT: randomized controlled trial
REM: racially and ethnically minoritized individual
STTS-R: Satisfaction with Therapy and Therapist Scale-Revised
TAU: treatment-as-usual
UCLA: University of California, Los Angeles

Edited by A Stone; submitted 17.Sep.2025; peer-reviewed by E Carrotte, R Browne, R Nooripour; revised version received 25.Mar.2026; accepted 26.Mar.2026; published 14.May.2026.

Please cite as:
Ramos G, Montoya AK, Aguilera A, Lau AS, Enders CK, Wen Y, Chavira DA
A Self-Guided App-Based Mindfulness Intervention for Racially and Ethnically Minoritized Individuals Who Experience Discrimination-Related Mental Health Symptoms: Randomized Controlled Trial
J Med Internet Res 2026;28:e84328
URL: https://www.jmir.org/2026/1/e84328
doi:10.2196/84328

© Giovanni Ramos, Amanda K Montoya, Adrian Aguilera, Anna S Lau, Craig K Enders, Yinyin Wen, Denise A Chavira. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Bidirectionality Between Perceived Immediate and Long-Term Benefits and Losses and Internet Gaming Disorder Among Chinese Adolescent Gamers: Prospective Longitudinal Study

Siman Li[1], BMed; Jianxin Zhang[2], BMed; Ji-bin Li[3], PhD; Joseph TF Lau[4,5*], PhD; Yanqiu Yu[1*], PhD

[1]Department of Preventive Medicine and Health Education, School of Public Health, Fudan University, Dong'an Road 130, Xuhui District, Shanghai, China

[2]West China School of Public Health, Sichuan University, Chengdu, Sichuan, China

[3]Department of Clinical Research, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangdong, Guangdong, China

[4]Public Mental Health Center, School of Mental Health, Wenzhou Medical University, Wenzhou, China

[5]Zhejiang Provincial Clinical Research Center for Mental Disorders, Affiliated Wenzhou Kangning Hospital of Wenzhou Medical University, Wenzhou, China

[*]these authors contributed equally

Corresponding Author:
Yanqiu Yu, PhD
Department of Preventive Medicine and Health Education, School of Public Health, Fudan University, Dong'an Road 130, Xuhui District, Shanghai, China

Abstract

Background: Adolescents perceive both immediate and long-term benefits and losses related to internet gaming, affecting their risk of internet gaming disorder (IGD). These perceptions could also be shaped and reinforced by IGD, indicating potential bidirectionality.

Objective: This study aimed to investigate the bidirectional relationships between perceived immediate and long-term benefits in 3 domains (mental health, social relationships, and personal achievement) and IGD, and between perceived immediate and long-term losses in 6 domains (mental health, sleep quality, academic performance, family relationships, social relationships, and personal achievement) and IGD.

Methods: A 12-month 2-wave prospective longitudinal study was conducted among junior middle school students who had played internet games in the past 12 months in Guangzhou and Chengdu, China, with a baseline survey (T1, December 2018) and the other identical follow-up survey conducted 1 year later (T2, December 2019). The participating schools were conveniently selected; all Grade 7 and 8 students were invited to self-administer the questionnaires in a classroom setting without the presence of the schoolteachers. The final sample size was 1173 students (mean age 12.5, SD 0.6 y; male: 693/1173, 59.1%). IGD was assessed by using the 9-item *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* IGD checklist.

Results: Cross-lagged panel analysis (adjusting for background factors) showed (1) stronger perceived immediate benefits of mental health (β=.08, 95% CI 0.01-0.15) and personal achievement (β=.10, 95% CI 0.01-0.20) at T1 significantly predicted more IGD symptoms at T2; (2) more IGD symptoms at T1 significantly predicted stronger perceived immediate and long-term benefits of social relationships (immediate: β=.09, 95% CI 0.03-0.15; long-term: β=.11, 95% CI:0.05-0.17) and personal achievement (immediate: β=.12, 95% CI 0.06-0.18; long-term: β=.10, 95% CI 0.04-0.16) at T2; (3) more IGD symptoms at T1 significantly predicted stronger perceived immediate and future losses in mental health (immediate: β=.09, 95% CI 0.03-0.15; long-term: β=.08, 95% CI 0.02-0.14), sleep quality (immediate: β=.10, 95% CI 0.04-0.16; long-term: β=.13, 95% CI 0.07-0.19), academic performance (immediate: β=.09, 95% CI 0.04-0.15; long-term: β=.07, 95% CI 0.01-0.13), and family relationships (immediate: β=.11, 95% CI 0.05-0.17; long-term: β=.10, 95% CI 0.04-0.16) at T2, as well as perceived long-term losses in social relationships at T2 (β=.08, 95% CI 0.02-0.14).

Conclusions: This study was innovative in integrating time perspective into both perceived benefits and losses of internet gaming, a cognitive dimension previously overlooked in literature. The current findings advance the field by revealing the unidimensional predictive effects of IGD on perceived immediate and long-term benefits and losses, with 2 exceptions of perceived immediate and long-term benefits of mental health and personal achievement conversely predicting IGD. These results contribute to the development of effective interventions: the cognitive components should go beyond the general pros and cons of gaming and target the potential temporal bias gamers hold.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e74030) doi:10.2196/74030

KEYWORDS

internet gaming disorder; time perspective; perceived benefits; perceived losses; cross-lagged panel model; adolescents

Introduction

Internet gaming is increasingly popular; the number of internet gamers worldwide reached 3.31 billion in 2023, with a growing rate of 4.3% year-on-year [1]. China had the largest number of internet gamers (668 million as of December 2023) [2]. The internet penetration rate of Chinese minors (aged below 18 y) was 97.2%, and 67.8% of them played internet games regularly [3]. Internet gaming could lead to internet gaming disorder (IGD), a mental disorder included by *the International Classification of Diseases, 11th Revision (ICD-11)* [4], and negative consequences of social (eg, poor parent-child relationships), psychological (eg, depression), and functional (eg, impaired cognitive functions) problems [5-14]. A meta-analysis of 96 studies reported the pooled prevalence of IGD among adolescents ranged from 7.5% to 10% globally [15]; it ranged from 2.4% to 21.5% in China [16,17]. The potentially high adolescent IGD prevalence underscores the need for a better understanding of the precedents and consequences of IGD.

Numerous theoretical and empirical studies highlight the role of gaming-specific cognitions in understanding IGD development and maintenance [18], among which outcome expectancy forms a typical domain. According to social cognitive theory, outcome expectancy refers to tangible or intangible outcomes anticipated as results of a behavior [19]. Outcome expectancy could be manifested in positive (ie, perceived benefits) and negative (eg, perceived losses) domains [19]. Regarding perceived benefits of internet gaming, a recent study summarized 3 key dimensions: mental health (eg, emotional boost), social relationships (eg, social interactions), and personal achievement (eg, improved self-esteem) [20]; all these perceived benefits were positively associated with IGD in multiple cross-sectional studies [21-26]. To our knowledge, these associations have not yet been demonstrated longitudinally. Regarding perceived losses due to internet gaming, internet gaming can entail potential negative consequences [27-30] that constitute losses to gamers. The literature generally identified 6 dimensions of these losses. First, IGD could exacerbate mental problems [5,8]. Second, IGD predicted strongly sleep problems (eg, sleep deprivation [6] and poor sleep quality [31,32]). Third, IGD had negative impacts on family relationships, including worsened parent-child relationships [33] and decreased family functioning [34]. Fourth, although internet gaming provides a platform for online social interactions [35], it could undermine real-life social interactions [36,37] and increase social anxiety or phobia [5,38-40], generating losses in social relationships. Last, internet gaming may lead to declines in academic performance and personal achievement. IGD was longitudinally associated with worsened academic achievement [41,42] and personal achievement [43,44]. Adolescent gamers may be aware of these losses, potentially changing their gaming behaviors and affecting IGD risk. Notably, while existing literature has examined associations between IGD and negative consequences, few studies have directly investigated perceived losses per se, their predictive effects on IGD, or the reverse association.

It is also novel to integrate the time perspective into the concept of perceived benefits and losses of internet gaming. The temporal frames of past, present, and future affect one's decision-making and a variety of addictive behaviors [45-48]. Accordingly, perceived benefits and losses of internet gaming could occur at present (immediate benefits and losses) or in the future (long-term benefits and losses). Our literature review identified only 1 cross-sectional study looking at both perceived immediate and future benefits of internet gaming simultaneously, and it reported positive associations between these 2 perceived benefits and IGD with a comparable effect size [20]. It suggests that perceived future benefits might be as important as perceived immediate benefits and should not be neglected in future studies. In contrast, it was expected that these perceived benefits might be the outcomes of IGD. Our literature search identified a sole study reporting that IGD significantly predicted positive outcome expectancies of using internet gaming as a self-regulatory strategy [49]. In addition, no studies looked at the reciprocity between perceived immediate and long-term losses and IGD.

The reciprocal determinism construct of Social Cognitive Theory postulates that personal cognitive factors are key drivers of health behavior and could be shaped and reinforced by the behavior itself [19], supporting the bidirectional relationships between perceived immediate and long-term benefits and losses and IGD. There is additional theoretical support. The expectancy-value theories postulate that a positive outcome expectancy would increase the likelihood of performing a behavior, while a negative outcome expectancy would decrease the likelihood of performing the behavior [50]. Accordingly, perceived benefits would be expected to increase IGD risk while perceived losses were expected to mitigate it. Conversely, the predictive effect of IGD on perceived benefits and losses could be explained by Cognitive Dissonance Theory [51]. There may be cognitive dissonance when one's behavior contradicts one's well-being or values. To reduce this dissonance without changing behavior, the individual may alter their perceptions, potentially increasing perceived benefits and reducing perceived losses. Conversely, experiencing the negative consequences of IGD could plausibly reduce perceived benefits and heighten perceived losses.

Given the background, this 2-wave prospective longitudinal study aimed to (1) investigate the levels of perceived immediate and long-term benefits of internet gaming in 3 dimensions (mental health, social relationships, and personal achievement) and perceived immediate and long-term losses in 6 dimensions (mental health, sleep quality, academic performance, family relationships, social relationships, and personal achievement) among Chinese adolescents, and (2) examine the bidirectional

XSL•FO
RenderX

relationships between perceived immediate and long-term benefits and losses and IGD. It was hypothesized that (1) perceived immediate and long-term benefits at baseline would be positively associated with IGD at follow-up, (2) perceived immediate and long-term losses at baseline would be negatively associated with IGD at follow-up, (3) IGD at baseline would be significantly associated with perceived immediate and long-term benefits and losses at follow-up, and the directions would be either positive or negative.

Methods

Participants and Data Collection

A 12-month 2-wave prospective longitudinal study was conducted among junior middle school students in Guangzhou and Chengdu, China; baseline and follow-up surveys were performed in December 2018 and 2019, respectively. The inclusion criteria at baseline included full-time grade 7 and grade 8 students, aged 12 - 14 years, who had played internet games in the past 12 months at baseline, and were willing to participate in this survey. The selection of Grade 7 and 8 students was driven by their behavioral risks and environmental shifts specific to early adolescence. Behaviorally, this period involves heightened sensation-seeking and emotional instability, making the immediate gratification of gaming highly attractive [52]. Environmentally, these students are navigating the stressful transition to secondary education and experiencing reduced parental supervision regarding digital device use compared to primary school. Those Grade 9 students were excluded as they would leave school after 1 year and could not be followed up.

With the assistance of local education sectors, 4 and 2 junior middle schools were conveniently selected, and all grade 7 and 8 students of the selected schools were invited for participation; the in-school surveys were approved by school principals. The anonymous survey was self-administered by the students in the classroom setting in the absence of schoolteachers. Before the survey, well-trained field workers prebriefed the students on the objectives, content, logistics, and voluntary nature of this study, and that submitting a completed questionnaire implied informed consent for participation. Written informed consent was not collected to maintain anonymity. Such information was also printed on the cover page of the structured questionnaire. In addition, field workers assisted the students (eg, answering inquiries) and did the quality check during and after the survey, respectively. After 12 months, the students completed the same questionnaire with the same data collection procedures. The reporting of this study follows the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [53], and the checklist is included in the supplementary material.

At baseline, 2463 completed the questionnaire, and a total of 77.4% (1906/2463) questionnaires were matched for both baseline (T1) and follow-up (T2) surveys. The lost to follow-up (530/1906, 21.8%) group was more likely than those followed-up to be older, Chengdu participants, male, not living with both parents, having father's and mother's educational levels of junior middle school or below, and having higher scores of IGD and some domains of perceived immediate and long-term benefits and losses (Table S1 in Multimedia Appendix 1). Among the matched questionnaires, 25.9% (494/1906) were excluded due to not having played internet games in the past 12 months at T1, 5.4% (103/1906) due to missing data in key variables (eg, gaming behaviors and IGD), and 7.1% (136/1906) due to age below12 or above 14 years. The final sample size was 1173 in this study.

Measurements

Background Variables

Background information was collected, including age, sex, city of study, whether living with both parents, father's and mother's educational levels (junior middle school or below, senior middle school or equal, or college or above), and perceived family financial status (very good, good, moderate, poor, or very poor).

IGD

The 9-item *DSM-5* Checklist was used to screen for the presence of IGD symptoms in the past 12 months, including (1) preoccupation, (2) withdrawal symptoms, (3) tolerance, (4) inability to control internet gaming, (5) prioritizing internet games over other activities, (6) continuing to play internet games excessively despite psychosocial problems, (7) deceiving others about gaming time, (8) using internet games for escapism, and (9) impaired functions (eg, social and academic performance) due to internet gaming. All the items were rated with binary response options (0=No and 1=Yes). A total score was calculated by summing the responses, resulting in a range from 0 to 9; higher scores indicate a higher severity of IGD symptoms. The Chinese version of the *DSM-5* Checklist has been validated among adolescents in China with satisfactory reliability and validity indices [54]. In this study, the Cronbach α of this checklist at T1 and T2 was 0.75 and 0.77, respectively.

Perceived Immediate and Long-Term Benefits of Internet Gaming

Perceived immediate and perceived long-term benefits of internet gaming are assessed using the Perceived Overall Immediate Benefits Scale (POIBS) and the Perceived Overall Future Benefits Scale (POFBS; Fudan University), respectively, both of which were developed in Chinese adolescents and comprised 3 dimensions: mental health, social relationships, and personal achievement [20]. Perceived immediate benefits were rated as the extent to which the participants believed that internet gaming had positive impacts AT PRESENT on each of the above 3 dimensions (eg, "How much positive impact do you think internet gaming has AT PRESENT on your mental health?"). Perceived long-term benefits were rated as the extent to which the participants believed that internet gaming had positive impacts IN THE FUTURE on each of the above 3 dimensions (eg, "How much positive impact do you think internet gaming has IN THE FUTURE on your mental health?"). The 6 items were rated by using 5-point Likert scales (0=None to 4=extremely strong); higher scores indicate higher levels of perceived immediate and long-term benefits. The Cronbach α of POIBS and POFBS were 0.76/0.75 and 0.78/0.78 at T1/T2, respectively.

XSL•FO
RenderX

Perceived Immediate and Long-Term Losses Due to Internet Gaming

Overall, 6 dimensions of perceived losses due to internet gaming were considered in this study based on a thorough literature review and expert panel discussion, including (1) mental health, (2) sleep quality, (3) academic performance, (4) family relationships, (5) social relationships, and (6) personal achievement. Following previous publications and taking reference from the above POIBS and POFBS scale [20], perceived immediate losses were rated as the extent to which the participants believed that internet gaming had negative impacts AT PRESENT on each of the above 6 dimensions (eg, "How much negative impacts do you think internet gaming has AT PRESENT on your mental health?"), while perceived long-term losses were rated as the extent to which the participants believed that internet gaming had negative impacts IN THE FUTURE on each of the above 6 dimensions (eg, "How much negative impacts do you think internet gaming has IN THE FUTURE on your mental health?"). The 12 items were rated by using 5-point Likert scales (0=None to 4=extremely strong); higher scores indicate higher levels of perceived immediate and long-term losses. The Cronbach α of perceived immediate and long-term losses scales were 0.87/0.88 and 0.89/0.90 at T1/T2, respectively. Confirmatory factor analyses confirmed the 6-item structure of perceived immediate and long-term losses at both waves, demonstrating satisfactory model fit indices [perceived immediate losses (T1/T2): all factor loadings over 0.67/0.66 (all $P<.001$), comparative fit index (CFI)=0.97/0.97, Tucker-Lewis Index (TLI)=0.95/0.95, and standardized root mean squared residual (SRMR)=0.03/0.03; perceived long-term losses (T1/T2): all factor loadings over 0.71/0.69 (all $P<.001$), CFI=0.98/0.97, TLI=0.96/0.95, and SRMR=0.02/0.03].

Data Analysis

Descriptive statistics and normality tests (via Kolmogorov-Smirnov and Shapiro-Wilk) were performed. Little's Missing Completely at Random (MCAR) test was conducted to evaluate the pattern of missing data; Multiple Imputation was used to handle missing data. Paired-sample *t* tests were used to test for within-individual differences in levels of perceived immediate and long-term benefits and losses at both T1 and T2. Such data analysis was performed by using SPSS software (version 26.0; IBM Corp). A 2-sided *P* value <.05 was defined as statistically different.

Cross-lagged panel analyses with the maximum likelihood robust estimator were performed using Mplus 8.3; the analyses were conducted across 20 imputed datasets, and parameters were pooled according to Rubin's rules. In total, 9 models were fit to examine the potential bidirectional associations between each pair of perceived immediate versus long-term benefits and losses and IGD. These models had 3 types of correlations: (1) cross-sectional correlations between variables at the same time point, (2) autocorrelations between the level of the same variable at both T1 and T2, and (3) cross-lagged correlations between the variable at T1 and the other variables at T2. To better reveal the cross-lagged correlations (eg, perceived short-term benefits of mental health on IGD), both cross-sectional correlations and autocorrelations, as well as background factors (ie, age, sex, city of study, whether living with both parents, parental educational levels, and perceived family financial status) at T1 were controlled. Satisfactory model fit indices of the models included CFI≥0.90, TLI≥0.90, and SRMR≤0.08 [55].

Ethical Considerations

This study was approved by the Survey and Behavioral Research Ethics Committee of the Chinese University of Hong Kong (number SBRE-18 - 430). Participation in this study was entirely voluntary. Prior to data collection, all participants were informed of the study objectives, procedures, and their rights as research participants. Given the anonymous nature of the survey, written informed consent was not obtained. Instead, informed consent was implied by the voluntary completion and submission of the questionnaire. Such information was also printed on the cover page of the structured questionnaire. All data were anonymous and no direct personal identifiers were collected. To enable longitudinal matching while preserving anonymity, instead, the last 4 digits of the father's and mother's phone numbers and the last phonetic letters of the father's and mother's given names were collected for matching purposes. No identifiable information about participants is hence possible in any part of this study or supplementary material. No financial or material incentives were provided to participants for their involvement in this study. In addition, parental approval was sought, and the parental opt-out procedure was exercised.

Results

Descriptive Statistics

Among all participants, the mean age was 12.5 (SD 0.6; range=12-14) years; 72.1% (846/1173) were from Guangzhou and 27.9% (327/1173) from Chengdu; over half (693/1173, 59.1%) were male; 15.6% (183/1173) did not live with both parents; 29.9% (350/1173) perceived that their family financial level was poor or very poor; nearly one-third of the participants' fathers (361/1173, 30.8%) or mothers (335/1173, 28.6%) had received college or above education (Table 1). The mean IGD scores were 2.1 (SD 2.1; range=0-9) and 1.7 (SD 2.1; range=0-9) at T1 and T2, respectively.

XSL•FO
RenderX

Table . Frequencies and proportions of background variables in the 2-wave cross-lagged panel study on bidirectionality between perceived immediate and long-term benefits and losses and internet gaming disorder among adolescent internet gamers in Chengdu and Guangzhou, China (December 2018 to December 2019).

Frequencies and proportions	n (%) (N=1173)
City of study	
Guangzhou	846 (72.1)
Chengdu	327 (27.9)
Sex	
Male	693 (59.1)
Female	480 (40.9)
Living with both parents	
Yes	975 (83.1)
No	183 (15.6)
Missing data	15 (1.3)
Family financial status	
Very good or good	112 (9.5)
Moderate	709 (60.4)
Poor or very poor	350 (29.9)
Missing data	2 (0.2)
Father's education level	
Junior middle school or below	427 (36.4)
Senior middle school or equivalent	316 (26.9)
College or above	361 (30.8)
Missing data	69 (5.9)
Mother's education level	
Junior middle school or below	451 (38.4)
Senior middle school or equivalent	322 (27.5)
College or above	335 (28.6)
Missing data	65 (5.5)

Within-Individual Comparison of Perceived Immediate and Long-Term Benefits and Losses at Both T1 and T2

The results (Table 2) showed (1) among the 3 dimensions of perceived immediate versus long-term benefits, all within-individual differences (except that between perceived immediate and long-term benefits of personal achievement at T2) were statistically significant, but the effect sizes were small (Cohen *d* ranged from 0.06 to 0.12), and (2) among the 6 dimensions of perceived immediate versus long-term losses, all within-individual differences (except those regarding sleep quality and social relationships at T1) were statistically significant with small effect sizes (Cohen *d* ranged from 0.08 to 0.22).

XSL•FO
RenderX

Table . The within-individual comparisons in perceived immediate and long-term benefits and losses in the 2-wave cross-lagged panel study on bidirectionality between perceived immediate and long-term benefits and losses and internet gaming disorder among adolescent internet gamers in Chengdu and Guangzhou, China (December 2018 to December 2019).

Dimension	T1[a], mean (SD)	Cohen *d*	*P* value	T2[b], mean (SD)	Cohen *d*	*P* value
Perceived benefits						
Mental health		−0.08	.004		0.06	.04
Immediate impacts	1.4 (1)			1.5 (1)		
Long-term impacts	1.4 (1.1)			1.4 (1.1)		
Social relationships		0.07	.01		0.09	.002
Immediate impacts	1.5 (1.3)			1.5 (1.2)		
Long-term impacts	1.5 (1.3)			1.4 (1.2)		
Personal achievement		−0.12	<.001		−0.03	.30
Immediate impacts	1.1 (1.2)			1.1 (1.1)		
Long-term impacts	1.2 (1.3)			1.1 (1.2)		
Perceived losses						
Mental health		−0.22	<.001		−0.2	<.001
Immediate impacts	1.1 (1.1)			1.3 (1.1)		
Long-term impacts	1.3 (1.2)			1.4 (1.2)		
Sleep quality		−0.05	.13		−0.08	.01
Immediate impacts	1.1 (1.2)			1.2 (1.2)		
Long-term impacts	1.1 (1.3)			1.3 (1.3)		
Academic performance		−0.08	.007		−0.08	.004
Immediate impacts	1.3 (1.2)			1.3 (1.2)		
Long-term impacts	1.3 (1.3)			1.4 (1.3)		
Family relationships		0.03	.25		−0.1	.001
Immediate impacts	1 (1.2)			1 (1.3)		
Long-term impacts	0.9 (1.2)			1.1 (1.2)		
Social relationships		−0.15	<.001		−0.13	<.001
Immediate impacts	0.7 (1)			0.8 (1)		
Long-term impacts	0.8 (1.1)			0.9 (1.1)		

XSL•FO
RenderX

Dimension	T1[a], mean (SD)	Cohen *d*	*P* value	T2[b], mean (SD)	Cohen *d*	*P* value
Personal achievement		−0.13	<.001		−0.14	<.001
Immediate impacts	0.8 (1.1)			0.9 (1.1)		
Long-term impacts	0.9 (1.2)			1 (1.2)		

[a]T1: assessment at baseline.

[b]T2: assessment at follow-up.

Cross-Lagged Panel Analysis

Bidirectional Relationships Between Perceived Immediate Versus Long-Term Benefits and IGD

Normality tests demonstrated that all key variables (perceived benefits and losses and IGD) at T1 and T2 violated the assumption of normality. Little's MCAR test was statistically significant (χ^2_{1353}=1737.65; *P*<.001), suggesting that data were not missing completely at random. Consequently, the cross-lagged panel models were estimated using a maximum likelihood robust estimator combined with multiple imputation to account for nonnormality and missing patterns. In this section, 3 cross-lagged panel models were fit, each containing a pair of perceived immediate versus long-term benefits and IGD and adjusting for background variables (ie, age, sex, city of study, whether living with both parents, parental educational levels, and perceived family financial status); all models showed satisfactory model fit indices (Figure 1). Higher levels of perceived immediate benefits of mental health and personal achievement (but not social relationships) at T1 predicted more IGD symptoms at T2, while the predictive effects of perceived long-term benefits in the 3 dimensions on IGD were statistically nonsignificant. Conversely, more IGD symptoms at T1 significantly predicted higher levels of perceived immediate and long-term benefits of social relationships and personal achievement (but not mental health) at T2.

Figure 1. Cross-lagged panel model testing the bidirectional relationships between perceived immediate and long-term benefits and internet gaming disorder in the 2-wave cross-lagged panel study among adolescent internet gamers in Chengdu and Guangzhou, China (December 2018 to December 2019). Standardized coefficients were presented. Dashed lines represent nonsignificant paths and solid lines represent significant paths. The models were adjusted for background factors including the city of study, sex, age, father's and mother's educational level, and whether living with both parents. The cross-sectional correlations and autocorrelations were not presented for simplicity. CFI: Comparative Fit Index; IGD: internet gaming disorder; SRMR: standardized root mean square residual; T1: assessment at baseline; T2: assessment at follow-up; TLI: Tucker-Lewis Index.

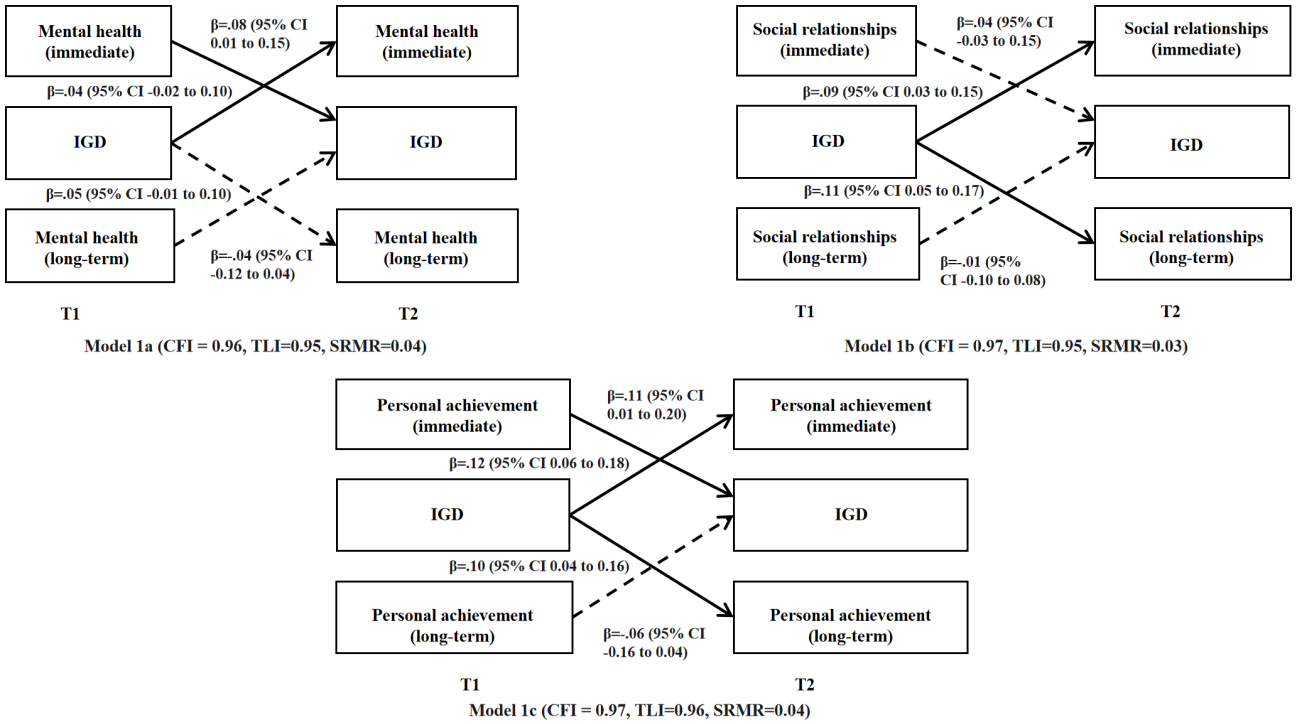

Bidirectional Relationships Between Perceived Immediate Versus Long-Term Losses and IGD

Similarly, 6 cross-lagged panel models were fitted, each containing a pair of perceived immediate versus long-term losses and IGD and adjusting for background variables; all models showed satisfactory model fit indices (Figure 2). All predictive effects of perceived immediate and long-term losses in mental health, sleep quality, academic performance, family relationships, social relationships, and personal achievement at

XSL•FO
RenderX

T1 on IGD at T2 were statistically nonsignificant. Conversely, more IGD symptoms at T1 significantly predicted higher levels of perceived immediate and long-term losses in mental health, sleep quality, academic performance, family relationships, and social relationships (only the long-term losses) at T2; the predictive effect of IGD at T1 on perceived losses in personal achievement was statistically nonsignificant.

Figure 2. Cross-lagged panel model testing the bidirectional relationships between perceived immediate and long-term losses and internet gaming disorder in the 2-wave cross-lagged panel study among adolescent internet gamers in Chengdu and Guangzhou, China (December 2018 to December 2019). Standardized coefficients were presented. Dashed lines represent nonsignificant paths and solid lines represent significant paths. The models were adjusted for background factors including the city of study, sex, age, father's and mother's educational level, and whether living with both parents. The cross-sectional correlations and autocorrelations were not presented for simplicity. CFI: Comparative Fit Index; IGD: internet gaming disorder; SRMR: standardized root mean square residual; T1: assessment at baseline; T2: assessment at follow-up; TLI: Tucker-Lewis Index.

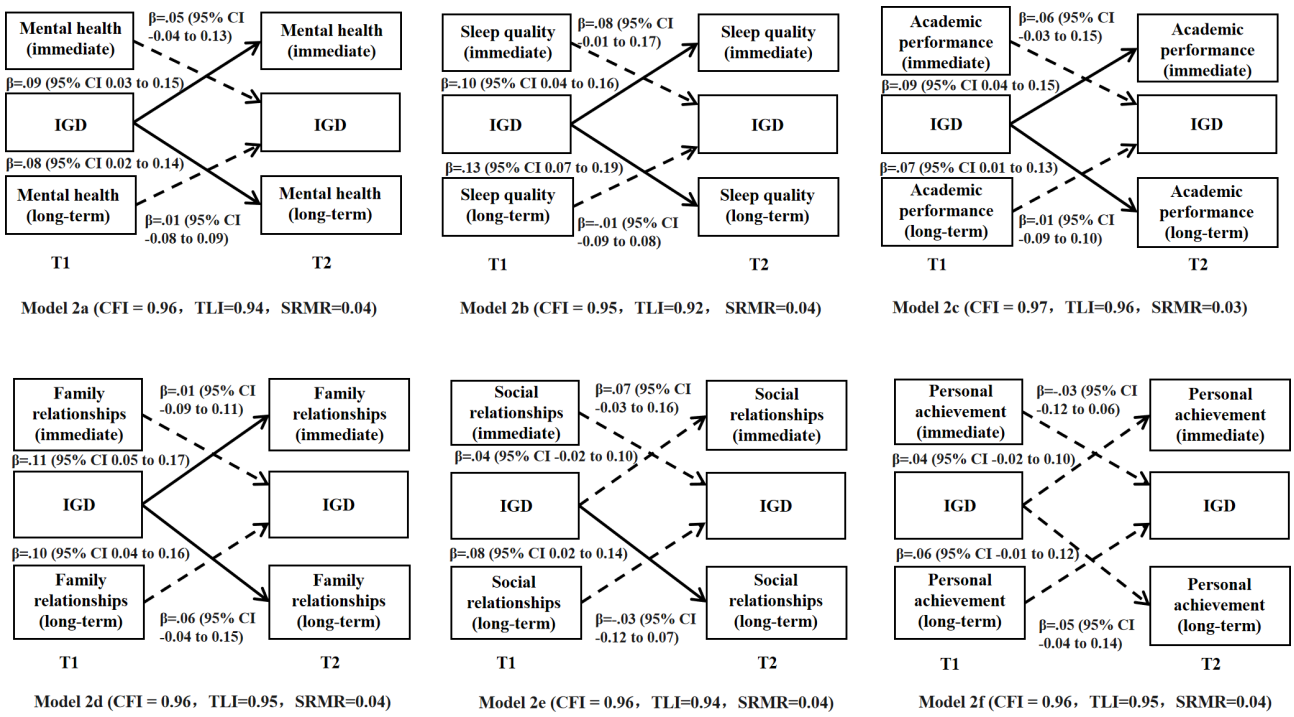

Discussion

Principal Findings

This study was novel to investigate the levels of perceived immediate and long-term benefits in 3 dimensions (mental health, social relationships, and personal achievement) and perceived immediate and long-term losses in 6 dimensions (mental health, sleep quality, academic performance, family relationships, social relationships, and personal achievement). All these levels were comparably low (within a range of 0 - 4); the mean scores for perceived benefits ranged from 1.1 to 1.5, while those for perceived losses ranged from 0.7 to 1.4. These findings reflect the nature of the general adolescent sample, where the majority of participants engage in gaming recreationally without reporting extreme levels of both positive and negative experiences. Nonetheless, the observations confirmed a previous study reporting comparable levels of perceived immediate versus long-term benefits [22] and further underscore the importance of integrating the time perspective into perceived benefits and losses. As the first study to investigate perceived losses, our findings suggest that adolescents may perceive both benefits and losses equally, and perceived losses should be a nonnegligible target in future interventions; future studies in this regard are greatly warranted for confirmation and elaboration.

This longitudinal study was also the first to reveal the interesting bidirectional relationships between perceived immediate and long-term benefits and IGD. Although existing cross-sectional studies reported that perceived benefits of improved mental health, social relationships, and personal achievement were positively associated with IGD [21-26], this longitudinal study only partially supports these findings and revealed the differential predictive effects of these perceptions on IGD. The significant results highlight the importance of perceived immediate mental health and personal achievement benefits. Plausibly, the immediate mental health benefits could offer quick and easily accessible relief from negative emotions for gamers. Furthermore, the instant gratification model postulates that behaviors (eg, internet gaming) providing immediate emotional relief are more likely to be repeated [56], potentially contributing to increased IGD risk. On the other hand, the self-determination theory postulates that satisfaction with the psychological need for competence is powerful in motivating behavior [57]. Accordingly, the sense of accomplishment gained from internet gaming could supply immediate feedback and rewards (ie, immediate personal achievement benefits), motivating repetitive gaming, thereby increasing IGD risk. In addition, the significant predictive effects involving immediate but not long-term benefits corroborate previous publications emphasizing immediate rewards over delayed gratification in the context of IGD [58]. Speculatively, long-term benefits on mental health and personal achievement from internet gaming

XSL•FO
RenderX

might occur gradually and remain less tangible to adolescent gamers, leading to the nonsignificant results. The nonsignificant predictive effects regarding perceived social relationships benefits might be associated with the complexity of social interactions, which can be context-dependent and influenced by external factors outside gaming (eg, peer relationships and behavioral inhibition) [59,60]. It is speculated that social interactions in gaming might not necessarily translate into meaningful or fulfilling relationships, which might reduce their impacts on IGD development.

Conversely, IGD at T1 significantly predicted perceived immediate and long-term benefits of social relationships and personal achievement at T2, but not those of mental health. It suggests that those with more IGD symptoms might amplify perceived benefits related to social relationships and personal achievement. Internet gaming is often designed with structured opportunities for social interactions (eg, multiplayer engagement) and measurable achievements (eg, leveling up) [61]; these aspects might be more pronounced among those with more IGD symptoms, leading to stronger perceptions of these benefits over time. In addition, those with more IGD symptoms may prioritize social and achievement benefits from internet gaming over those from other sources over time [21,62]. In contrast, individuals with more IGD symptoms might experience negative mental health outcomes, such as anxiety and depression [5,8]. These negative psychological conditions may overshadow the mental health benefits and lead to the nonsignificant predictive effect of IGD on perceived mental health benefits.

Unexpectedly, all predictive effects of perceived immediate and long-term losses in 6 domains at T1 on IGD at T2 were statistically nonsignificant. While perceived benefits and losses might be perceived at comparable levels, the psychological impacts of immediate benefits might have stronger influences on developing IGD due to their ability to fulfill psychological needs like competence as aforementioned [20]. This could overshadow the importance of perceived losses in terms of affecting IGD. This speculation was also supported by the concept of reinforcement that positive reinforcement (eg, perceived benefits) tends to have more immediate and stronger effects on motivating behavior in comparison with negative reinforcement or punishment (eg, perceived losses) [63]. Nonetheless, such speculations should be tested in future studies.

Reversely, this study observed differential predictive effects of IGD at T1 on the 6 domains of perceived immediate and long-term losses at T2. First, IGD significantly predicted perceived immediate and long-term losses in mental health, sleep quality, academic performance, and family relationships, suggesting that IGD could lead to noticeable harm in these domains. These findings align with empirical evidence that IGD cases face common negative consequences, such as depression, disturbed sleep patterns, worsened academic performance, and strained family relationships [27,33,41]. Notably, these domains of perceived losses might be more directly observable or harder to rationalize over time and are recommended to be emphasized in IGD intervention programs. Second, IGD at T1 significantly predicted perceived long-term, but not immediate, losses in social relationships, indicating a potential lag effect. It is speculated that, initially, the social benefits of internet gaming might buffer against perceived immediate losses, but, over time, those with more IGD symptoms might experience weakened offline social ties and relationships [64], making the long-term losses more apparent and significant. Last, the predictive effects of IGD at T1 on perceived immediate and long-term losses in personal achievement were statistically nonsignificant, suggesting that individuals with more IGD symptoms might not associate their gaming habits with detriments in personal achievement. Plausibly, as aforementioned, internet gaming may fulfill one's psychological needs of competence [65] and provide satisfactory substitutes for real-life achievements [66], decreasing perceived losses.

The current findings have practical implications, as gaming-specific cognitions are important modifiable constructs in IGD interventions (eg, cognitive-behavioral therapy) [67]. IGD prevention interventions may focus more on the immediate benefits on mental health and personal achievement. For instance, alternatives for immediate rewards of mental health and personal achievements (eg, outdoor physical activities) should be provided for adolescents to reduce their reliance on and sensitivity toward these gaming rewards. It might also be helpful to separate gaming achievement from real-world achievement and then educate adolescents on how to fulfill the latter [68]. In contrast, IGD treatment intervention may underscore the cognitions of potential losses in mental health, sleep quality, academic performance, and family relationships, as well as the long-term losses in social relationships, while perceived rewards of social relationships and personal achievement are recommended to be downplayed.

There are several limitations of this study. First, the attrition rate of this study was about 20%. Attrition analysis revealed that those lost to follow-up demonstrated stronger perceived benefit and losses in some domains, besides specific background differences (eg, age and sex). In addition, Little's MCAR test showed that the missing data were not completely random. It suggests that the study sample may have stronger perceived benefits and losses than the baseline population. Nonetheless, multiple imputation was used in the cross-lagged panel analyses to address the missing data patterns, increasing the robustness of the current findings. Furthermore, the study population was selected from 2 cities in China through convenience sampling, which may lead to selection bias. Cautions are needed when extrapolating the results to other regions or countries. Second, as participants were asked to self-administer the questionnaire, there might be reporting bias, including recall bias and social desirability bias. Third, although this study provides valuable insights on the bidirectional relationships between perceived immediate and long-term benefits and losses and IGD, it comprised only 2 time-points within a 12-month follow-up period. Future longitudinal studies with more time lags and longer study periods are needed to confirm the findings. In addition, while several predictive effects were statistically significant, their associated effect size was modest. It suggests that, despite statistical significance, these associations might have limited practical impacts on individuals. However, it is important to note that small effect sizes could still bear practical implications at the population level. Modest increases or

XSL•FO
RenderX

reductions in each predictor might yield substantial benefits for the population, though future intervention studies are needed for confirmation. Last, there might be other potential losses due to internet gaming that were not included in this study (eg, losses in physical health and the financial domain).

Conclusions

This study was novel in integrating the time perspective into perceived benefits and losses related to internet gaming and investigating their bidirectional relationships with IGD. The results are interesting in that only perceived immediate benefits of mental health and personal achievement predicted IGD, while IGD predicted perceived immediate and long-term benefits on social relationships and personal achievement, as well as perceived immediate and long-term losses in all 6 domains (except personal achievement). These findings enhance the understanding of the relationships between gaming-specific cognitions and IGD and shed new insights for relevant studies, taking into consideration the time perspective. The differential bidirectional relationships are also important for the design of IGD prevention and treatment interventions. Specifically, relevant cognitive components should target the specific temporal bias adolescent gamers possess, rather than focusing merely on the general benefits and losses of internet gaming.

Acknowledgments

We would like to thank the participants for their contribution to the present. The authors declare that no generative artificial intelligence was used in the drafting of this manuscript.

Funding

This study was funded by Shanghai Pujiang Programme (No. 23PJC018) and the National Natural Science Foundation of China (No. 82304258). The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the article.

Data Availability

The data sets generated and/or analysed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization and validation: YY
Methodology and writing-review and editing: SL, YY, and JTFL
Investigation: JZ and JBL
Software, formal analysis, data curation, and writing-original draft: SL
Resources: JTFL
Supervision: YY and JTFL

Conflicts of Interest

None declared.

Multimedia Appendix 1
Attrition analysis in the 2-wave cross-lagged panel study on bidirectionality between perceived immediate and long-term benefits/losses and internet gaming disorder among adolescent internet gamers in Chengdu and Guangzhou, China (December 2018 to December 2019).
[DOCX File, 24 KB - jmir_v28i1e74030_app1.docx]

References

1. Newzoo's global games market report 2023. Newzoo. URL: https://newzoo.com/resources/trend-reports/newzoo-global-games-market-report-2023-free-version [accessed 2026-02-11]
2. China game industry report 2023. CGIGC. URL: https://www.cgigc.com.cn/details.html?id=374087&type=report [accessed 2026-02-11]
3. (CNNIC). Report on the 5th national survey on internet use by minors. 2023.
4. ICD-11 for mortality and morbidity statistics. URL: https://icd.who.int/browse/2025-01/mms/en [accessed 2026-02-11]
5. Gentile DA, Choo H, Liau A, et al. Pathological video game use among youths: a two-year longitudinal study. Pediatrics 2011 Feb;127(2):e319-e329. [doi: 10.1542/peds.2010-1353] [Medline: 21242221]
6. Kim SY, Kim MS, Park B, Kim JH, Choi HG. Lack of sleep is associated with internet use for leisure. PLoS ONE 2018;13(1):e0191713. [doi: 10.1371/journal.pone.0191713] [Medline: 29360882]
7. Schneider LA, King DL, Delfabbro PH. Family factors in adolescent problematic Internet gaming: a systematic review. J Behav Addict 2017 Sep 1;6(3):321-333. [doi: 10.1556/2006.6.2017.035] [Medline: 28762279]

XSL•FO
RenderX

8. Ahmadi J, Amiri A, Ghanizadeh A, et al. Prevalence of addiction to the internet, computer games, DVD, and video and its relationship to anxiety and depression in a sample of Iranian High School students. Iran J Psychiatry Behav Sci 2014;8(2):75-80. [Medline: 25053960]
9. Henchoz Y, Studer J, Deline S, N'Goran AA, Baggio S, Gmel G. Video gaming disorder and sport and exercise in emerging adulthood: a longitudinal study. Behav Med 2016;42(2):105-111. [doi: 10.1080/08964289.2014.965127] [Medline: 25258243]
10. De Pasquale C, Dinaro C, Sciacca F. Relationship of internet gaming disorder with dissociative experience in Italian university students. Ann Gen Psychiatry 2018;17:28. [doi: 10.1186/s12991-018-0198-y] [Medline: 29983724]
11. Ma HK. Internet addiction and antisocial internet behavior of adolescents. ScientificWorldJournal 2011;11:2187-2196. [doi: 10.1100/2011/308631] [Medline: 22125466]
12. Yamada T, Moshier SJ, Otto MW. No site unseen: predicting the failure to control problematic internet use among young adults. Cogn Behav Ther 2016 Nov;45(6):496-500. [doi: 10.1080/16506073.2016.1205657] [Medline: 27426432]
13. Dong G, Potenza MN. Risk-taking and risky decision-making in Internet gaming disorder: implications regarding online gaming in the setting of negative consequences. J Psychiatr Res 2016 Feb;73:1-8. [doi: 10.1016/j.jpsychires.2015.11.011] [Medline: 26656573]
14. Wang Q, Ren H, Long J, Liu Y, Liu T. Research progress and debates on gaming disorder. Gen Psychiatr 2019;32(3):e100071. [doi: 10.1136/gpsych-2019-100071] [Medline: 31423477]
15. Gao YX, Wang JY, Dong GH. The prevalence and possible risk factors of internet gaming disorder among adolescents and young adults: systematic reviews and meta-analyses. J Psychiatr Res 2022 Oct;154:35-43. [doi: 10.1016/j.jpsychires.2022.06.049] [Medline: 35926424]
16. Paulus FW, Ohmann S, von Gontard A, Popow C. Internet gaming disorder in children and adolescents: a systematic review. Develop Med Child Neuro 2018 Jul;60(7):645-659 [FREE Full text] [doi: 10.1111/dmcn.13754]
17. Liao Z, Huang Q, Huang S, et al. Prevalence of internet gaming disorder and its association with personality traits and gaming characteristics among Chinese adolescent gamers. Front Psychiatry 2020;11:598585. [doi: 10.3389/fpsyt.2020.598585] [Medline: 33312143]
18. Billieux J, Potenza MN, Maurage P, et al. Cognitive Factors Associated with Gaming Disorder: Academic Press; 2020:221-230. [doi: 10.1016/B978-0-12-815298-0.00016-2]
19. Nabavi RT. Bandura's social learning theory & social cognitive learning theory. In: Theory of Developmental Psychology 2012:1-24 URL: https://www.researchgate.net/profile/Mohammad-Bijandi/publication/267750204_Bandura's_Social_Learning_Theory_Social_Cognitive_Learning_Theory/links/67536801b558f41d0fbe2fb1/Banduras-Social-Learning-Theory-Social-Cognitive-Learning-Theory.pdf
20. Yu Y, Fong VWI, Li JB, Zhang J, Lau JTF. The independent impacts of perceived immediate and future benefits of internet gaming on internet gaming disorder among adolescent internet gamers in China: mediation via internet gaming time. Comput Human Behav 2023 Dec;149:107965. [doi: 10.1016/j.chb.2023.107965]
21. Wang HY, Cheng C. The associations between gaming motivation and internet gaming disorder: systematic review and meta-analysis. JMIR Ment Health 2022 Feb 17;9(2):e23700. [doi: 10.2196/23700] [Medline: 35175204]
22. Cabeza-Ramírez LJ, Sánchez-Cañizares SM, Fuentes-García FJ, Santos-Roldán LM. Exploring the connection between playing video games and watching video game streaming: relationships with potential problematic uses. Comput Human Behav 2022 Mar;128:107130. [doi: 10.1016/j.chb.2021.107130]
23. Yildiz Durak H, Haktanir A, Saritepeci M. Examining the predictors of video game addiction according to expertise levels of the players: the role of time spent on video gaming, engagement, positive gaming perception, social support and relational health indices. Int J Ment Health Addict 2023 May 9:1-26. [doi: 10.1007/s11469-023-01073-3] [Medline: 37363766]
24. Cabeza-Ramírez LJ, Rey-Carmona FJ, Del Carmen Cano-Vicente M, Solano-Sánchez M. Analysis of the coexistence of gaming and viewing activities in twitch users and their relationship with pathological gaming: a multilayer perceptron approach. Sci Rep 2022 May 12;12(1):7904. [doi: 10.1038/s41598-022-11985-0] [Medline: 35551493]
25. Zeng W, Wei H, Liu M. Need for distinctiveness leads to pathological internet use? The perspective of cognitive behavioral model. Int J Environ Res Public Health 2023 Jan 16;20(2):1609. [doi: 10.3390/ijerph20021609] [Medline: 36674362]
26. Johannes N, Vuorre M, Przybylski AK. Video game play is positively correlated with well-being. R Soc Open Sci 2021 Feb 17;8(2):202049. [doi: 10.1098/rsos.202049] [Medline: 33972879]
27. Chen YL, Gau SSF. Sleep problems and internet addiction among children and adolescents: a longitudinal study. J Sleep Res 2016 Aug;25(4):458-465. [doi: 10.1111/jsr.12388] [Medline: 26854132]
28. Lemmens JS, Valkenburg PM, Peter J. The effects of pathological gaming on aggressive behavior. J Youth Adolesc 2011 Jan;40(1):38-47. [doi: 10.1007/s10964-010-9558-x] [Medline: 20549320]
29. Gentile D. Pathological video-game use among youth ages 8 to 18: a national study. Psychol Sci 2009 May;20(5):594-602. [doi: 10.1111/j.1467-9280.2009.02340.x] [Medline: 19476590]
30. Yu Y, Yang X, Wang S, et al. Serial multiple mediation of the association between internet gaming disorder and suicidal ideation by insomnia and depression in adolescents in Shanghai, China. BMC Psychiatry 2020 Sep 23;20(1):460. [doi: 10.1186/s12888-020-02870-z] [Medline: 32967648]
31. Lam LT. Internet gaming addiction, problematic use of the internet, and sleep problems: a systematic review. Curr Psychiatry Rep 2014 Apr;16(4):1-9. [doi: 10.1007/s11920-014-0444-1] [Medline: 24619594]

XSL•FO
RenderX

32. Smyth JM. Beyond self-selection in video game play: an experimental examination of the consequences of massively multiplayer online role-playing game play. Cyberpsychol Behav 2007 Oct;10(5):717-721. [doi: 10.1089/cpb.2007.9963]
33. Charlie CWD, HyeKyung C, Khoo A. Role of parental relationships in pathological gaming. Procedia - Social and Behavioral Sciences 2011;30:1230-1236. [doi: 10.1016/j.sbspro.2011.10.238]
34. Ko CH, Wang PW, Liu TL, Yen CF, Chen CS, Yen JY. Bidirectional associations between family factors and internet addiction among adolescents in a prospective investigation. Psychiatry Clin Neurosci 2015 Apr;69(4):192-200. [doi: 10.1111/pcn.12204] [Medline: 24836367]
35. Liu CC, Chang IC. Model of online game addiction: the role of computer-mediated communication motives. Telematics and Informatics 2016 Nov;33(4):904-915. [doi: 10.1016/j.tele.2016.02.002]
36. Kaczmarek LD, Drążkowski D. MMORPG escapism predicts decreased well-being: examination of gaming time, game realism beliefs, and online social support for offline problems. Cyberpsychol Behav Soc Netw 2014 May;17(5):298-302. [doi: 10.1089/cyber.2013.0595] [Medline: 24605951]
37. Chappell D, Eatough V, Davies MNO, Griffiths M. EverQuest—it's just a computer game right? An interpretative phenomenological analysis of online gaming addiction. Int J Ment Health Addiction 2006 Jul;4(3):205-216. [doi: 10.1007/s11469-006-9028-6]
38. Marino C, Canale N, Vieno A, Caselli G, Scacchi L, Spada MM. Social anxiety and internet gaming disorder: the role of motives and metacognitions. J Behav Addict 2020 Oct 12;9(3):617-628. [doi: 10.1556/2006.2020.00044] [Medline: 32750032]
39. Sioni SR, Burleson MH, Bekerian DA. Internet gaming disorder: social phobia and identifying with your virtual self. Comput Human Behav 2017 Jun;71:11-15. [doi: 10.1016/j.chb.2017.01.044]
40. Gioia F, Colella GM, Boursier V. Evidence on problematic online gaming and social anxiety over the past ten years: a systematic literature review. Curr Addict Rep 2022 Mar;9(1):32-47. [doi: 10.1007/s40429-021-00406-3]
41. Brunborg GS, Mentzoni RA, Frøyland LR. Is video gaming, or video game addiction, associated with depression, academic achievement, heavy episodic drinking, or conduct problems? J Behav Addict 2014 Mar;3(1):27-32. [doi: 10.1556/JBA.3.2014.002] [Medline: 25215212]
42. Nie Q, Teng Z, Yang C, Griffiths MD, Guo C. Longitudinal relationships between school climate, academic achievement, and gaming disorder symptoms among chinese adolescents. J Youth Adolescence 2024 Jul;53(7):1646-1665. [doi: 10.1007/s10964-024-01952-5]
43. Teng Z, Pontes HM, Nie Q, Xiang G, Griffiths MD, Guo C. Internet gaming disorder and psychosocial well-being: a longitudinal study of older-aged adolescents and emerging adults. Addict Behav 2020 Nov;110:106530. [doi: 10.1016/j.addbeh.2020.106530] [Medline: 32683173]
44. Nie Q, Griffiths MD, Teng Z. The role of self-esteem in protecting against cyber-victimization and gaming disorder symptoms among adolescents: a temporal dynamics analysis. J Youth Adolesc 2024 Apr;53(4):863-876. [doi: 10.1007/s10964-023-01890-8] [Medline: 37904058]
45. Zimbardo PG, Boyd JN. Putting time in perspective: a valid, reliable individual-differences metric. In: Time Perspective Theory; Review, Research and Application: Essays in Honor of Philip G Zimbardo: Springer; 2014:17-55.
46. Apostolidis T, Fieulaine N, Soulé F. Future time perspective as predictor of cannabis use: exploring the role of substance perception among French adolescents. Addict Behav 2006 Dec;31(12):2339-2343. [doi: 10.1016/j.addbeh.2006.03.008] [Medline: 16626883]
47. Barnett E, Spruijt-Metz D, Unger JB, Rohrbach LA, Sun P, Sussman S. Bidirectional associations between future time perspective and substance use among continuation high-school students. Subst Use Misuse 2013 Jun 18;48(8):574-580. [doi: 10.3109/10826084.2013.787092]
48. Kim J, Hong H, Lee J, Hyun MH. Effects of time perspective and self-control on procrastination and internet addiction. J Behav Addict 2017 Jun 1;6(2):229-236. [doi: 10.1556/2006.6.2017.017] [Medline: 28494615]
49. Dang L, Yang HM, Spada MM, Wu AMS. A three-wave longitudinal study on the underlying metacognitive mechanism between depression and Internet gaming disorder. J Behav Addict 2024 Mar 26;13(1):215-225. [doi: 10.1556/2006.2023.00072] [Medline: 38183452]
50. Oei TP, Baldwin AR. Expectancy theory: a two-process model of alcohol use and abuse. J Stud Alcohol 1994 Sep;55(5):525-534. [doi: 10.15288/jsa.1994.55.525] [Medline: 7990462]
51. Chiou WB, Wan CS. Using cognitive dissonance to induce adolescents' escaping from the claw of online gaming: the roles of personal responsibility and justification of cost. Cyberpsychol Behav 2007 Oct;10(5):663-670. [doi: 10.1089/cpb.2007.9972]
52. Crone EA, Dahl RE. Understanding adolescence as a period of social-affective engagement and goal flexibility. Nat Rev Neurosci 2012 Sep;13(9):636-650. [doi: 10.1038/nrn3313] [Medline: 22903221]
53. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. The Lancet 2007 Oct;370(9596):1453-1457. [doi: 10.1016/S0140-6736(07)61602-X]

XSL•FO
RenderX

54. Yang X, Jiang X, Wu AMS, et al. Validation of the internet gaming disorder symptoms checklist based on the fifth edition of the diagnostic and statistical manual of mental disorders in Chinese adolescents. Child Psychiatry Hum Dev 2023 Feb;54(1):26-33. [doi: 10.1007/s10578-021-01213-7] [Medline: 34338962]
55. Kline RB. Principles and Practice of Structural Equation Modeling: Guilford Publications; 2023. URL: https://books.google.co.in/ [illegible] [accessed 2026-02-16]
56. Harris C, Laibson D. Instantaneous gratification. Q J Econ 2013 Feb;128(1):205-248. [doi: 10.1093/qje/qjs051]
57. Deci EL, Ryan RM. Self-determination theory. In: Handbook of Theories of Social Psychology: Sage Publications; 2012:416-436. [doi: 10.4135/9781446249215.n21]
58. Weinstein A, Abu HB, Timor A, Mama Y. Delay discounting, risk-taking, and rejection sensitivity among individuals with internet and video gaming disorders. J Behav Addict 2016 Dec;5(4):674-682. [doi: 10.1556/2006.5.2016.081] [Medline: 27958761]
59. Essex MJ, Klein MH, Slattery MJ, Goldsmith HH, Kalin NH. Early risk factors and developmental pathways to chronic high inhibition and social anxiety disorder in adolescence. Am J Psychiatry 2010 Jan;167(1):40-46. [doi: 10.1176/appi.ajp.2009.07010051] [Medline: 19917594]
60. Tillfors M, Persson S, Willén M, Burk WJ. Prospective links between social anxiety and adolescent peer relations. J Adolesc 2012 Oct;35(5):1255-1263. [doi: 10.1016/j.adolescence.2012.04.008] [Medline: 22595770]
61. Wang H, Sun CT. Game reward systems: gaming experiences and social meanings. Presented at: Proceedings of DiGRA 2011 Conference; Sep 14-17, 2011. [doi: 10.26503/dl.v2011i1.594]
62. Collins E, Freeman J. Do problematic and non-problematic video game players differ in extraversion, trait empathy, social capital and prosocial tendencies? Comput Human Behav 2013 Sep;29(5):1933-1940. [doi: 10.1016/j.chb.2013.03.002]
63. Shteingart H, Loewenstein Y. Reinforcement learning and human behavior. Curr Opin Neurobiol 2014 Apr;25:93-98. [doi: 10.1016/j.conb.2013.12.004] [Medline: 24709606]
64. Longman H, O'Connor E, Obst P. The effect of social support derived from world of warcraft on negative psychological symptoms. Cyberpsychol Behav 2009 Oct;12(5):563-566. [doi: 10.1089/cpb.2009.0001]
65. Wan CS, Chiou WB. Why are adolescents addicted to online gaming? An interview study in Taiwan. Cyberpsychol Behav 2006 Dec;9(6):762-766. [doi: 10.1089/cpb.2006.9.762]
66. Snodgrass JG, Dengah HJF 2nd, Lacy MG, Fagan J. A formal anthropological view of motivation models of problematic MMO play: achievement, social, and immersion factors in the context of culture. Transcult Psychiatry 2013 Apr;50(2):235-262. [doi: 10.1177/1363461513487666] [Medline: 23690445]
67. King DL, Delfabbro PH. The cognitive psychopathology of internet gaming disorder in adolescence. J Abnorm Child Psychol 2016 Nov;44(8):1635-1645. [doi: 10.1007/s10802-016-0135-y] [Medline: 26875565]
68. Allen JJ, Anderson CA. Satisfaction and frustration of basic psychological needs in the real world and in video games predict internet gaming disorder scores and well-being. Comput Human Behav 2018 Jul;84:220-229. [doi: 10.1016/j.chb.2018.02.034]

Abbreviations

CFI: comparative fit index
DSM-5: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
ICD-11: *International Classification of Diseases, 11th Revision*
IGD: internet gaming disorder
MCAR: missing completely at random
POFBS: Perceived Overall Future Benefits Scale
POIBS: Perceived Overall Immediate Benefits Scale
SRMR: standardized root mean squared residual
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TLI: Tucker-Lewis Index

Edited by S Brini; submitted 16.Mar.2025; peer-reviewed by T Baranowski, TY Wang; accepted 26.Jan.2026; published 26.Feb.2026.

Please cite as:
Li S, Zhang J, Li JB, Lau JTF, Yu Y
Bidirectionality Between Perceived Immediate and Long-Term Benefits and Losses and Internet Gaming Disorder Among Chinese Adolescent Gamers: Prospective Longitudinal Study
J Med Internet Res 2026;28:e74030
URL: https://www.jmir.org/2026/1/e74030
doi:10.2196/74030

XSL•FO
RenderX

© Siman Li, Jianxin Zhang, Ji-bin Li, Joseph TF Lau, Yanqiu Yu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Interactions of Technology and Obsessive-Compulsive Disorder Symptomatology in Adults: Qualitative Interview Study

Lucas Occhino-Moede, MD, MPH; Kaitlyn Sulivan-Pascual; Kendall Phelan, MPH; Harrison Wang, BS; Daniel Mokhtar, BS; Elisa Liu; Erica Schug, BS; Megan Mirkis, BS; Thomas Baek, BS; Tamerlane Visher, BA; Ujjwal Pasupulety, MS; Adam Charles Frank, MD, PhD

Department of Psychiatry and Behavioral Sciences, Keck School of Medicine, University of Southern California, 2250 Alcazar St., Suite 2200, Los Angeles, CA, United States

Corresponding Author:
Adam Charles Frank, MD, PhD
Department of Psychiatry and Behavioral Sciences, Keck School of Medicine, University of Southern California, 2250 Alcazar St., Suite 2200, Los Angeles, CA, United States

Abstract

Background: Obsessive-compulsive disorder (OCD) affects 1% - 3% of the population and is marked by intrusive obsessions and compulsive behaviors that impair daily functioning. As digital technologies have become ubiquitous, their features may interact with OCD symptom dimensions in ways that both exacerbate and alleviate symptoms. While case reports and clinical anecdotes suggest such interactions, systematic investigation of patients' lived experiences with technology remains limited.

Objective: This study aimed to explore how individuals with OCD perceive and navigate their interactions with modern technologies, and to identify how specific features of technology may contribute to, reinforce, or relieve obsessive-compulsive symptom cycles.

Methods: We conducted semistructured interviews (n=24) with adults self-reporting a diagnosis of OCD, recruited through online OCD communities and advocacy networks. Interviews were conducted via the HIPAA (Health Insurance Portability and Accountability Act)-compliant platform Zoom (Zoom Communications) between May and December 2024 (median duration 51, IQR 6.5 minutes). Transcripts were coded in Dedoose (version 9.2.22; SocioCultural Research Consultants) using a constructivist grounded theory approach. Coding proceeded iteratively through open and focused coding, with theoretical saturation reached after 15 interviews. Constant comparison and analytic memoing guided the development of a conceptual framework linking technology features to OCD symptom dimensions.

Results: Participants (median age 26, IQR 12.8, range 20 - 64 years; 67%, 16/24 women, 29%, 7/24 men, and 4%, 1/24 nonbinary) described technology as both a trigger for and a coping tool against OCD symptoms. Analysis produced four central technology-related categories: (1) information-provision platforms (eg, social media, search engines, large language models, etc) that triggered disturbing-thought obsessions and enabled compulsive checking and reassurance-seeking; (2) gamification and quantification features (eg, streaks, progress bars, and tracking metrics) that reinforced "not-just-right" and symmetry-based compulsions; (3) notifications that provoked urges to clear, check, and maintain control, spanning both disturbing-thought and symmetry domains; and (4) user interfaces whose complexity and customizability elicited compulsive ordering, avoidance behaviors, and digital overwhelm.

Conclusions: This study characterizes how interactions between OCD and digital technologies manifest across established symptom domains, most notably disturbing-thought and "not-just-right" categories. Participants overwhelmingly experienced compulsive checking, reassurance-seeking, and ordering behaviors reinforced by features such as information-provision, gamification, notifications, and user interfaces. These findings highlight the clinical relevance of technology-related compulsions and suggest value in their systematic assessment, incorporation into psychoeducation, and consideration in digital design.

(J Med Internet Res 2026;28:e85033) doi:10.2196/85033

KEYWORDS

obsessive-compulsive disorder; OCD; technology; wearables; qualitative; patient perspectives

Introduction

Obsessive-compulsive disorder (OCD) occurs in 1% - 3% of the population [1] and is most commonly characterized by the presence of obsessions and compulsions. Obsessions are intrusive, unwanted thoughts, images, or impulses that are typically experienced as ego-dystonic and distressing. Compulsions are repetitive actions or mental rituals in which an individual engages to relieve anxiety related to obsessions.

XSL•FO
RenderX

Obsessions and compulsions, in the context of OCD, are time consuming and often decrease functioning in social, occupational, and other aspects of life.

OCD is a heterogeneous disorder with variety in the types of obsessions and compulsions experienced, severity of symptoms, and presence of comorbidities. While no two individuals are likely to experience the same pattern of symptoms, factor analytic approaches have been used to parse OCD heterogeneity. These studies have consistently identified the following four broad dimensions: (1) forbidden or taboo thoughts, (2) symmetry and ordering, (3) contamination and cleaning, and (4) hoarding [2,3]. More recent research has used expanded symptom checklists and refined analytic techniques to explore the relationships between these dimensions. The higher-level groupings of incompleteness and disturbing thoughts (distressing images that typically follow themes like contamination, harmful acts, or religion) strongly associate with the other dimensions and thereby are postulated to represent core phenotypes within OCD. Subdimensions within these two categories were also delineated and noted to be structured in a hierarchy. The first order domain of incompleteness was divided into accuracy, mental and perceptual, and "not-just-right" (NJR) behaviors. NJR experiences are typically an internal sense of disturbance related to qualities such as order, placement, or frequency and drive individuals to engage in actions like tapping, counting, or reviewing. Meanwhile, disturbing thoughts encompassed symptoms relating to harm and checking and forbidden thoughts [4].

Excessive or inappropriate use of technology is associated with mental health issues. Children and adolescents are particularly susceptible; higher amounts of screen time and social media use have been linked to a range of mental health symptoms due to disturbances in regular sleep-wake cycles [5]. Technology can also facilitate maladaptive behaviors such as compulsive internet use and gambling disorder [6]. Patients with psychotic disorders have expressed that online platforms can exacerbate their paranoia and delusions through misinformation or targeted harassment [7]. Within OCD, the increasing digitization of society over the past several decades has provided new mediums for the manifestation of symptoms [8]. Recent case reports describe patients with OCD whose symptoms have been exacerbated by technology use. One individual described obsessions about inappropriately reacting to social media posts and would consequently spend hours per day scrolling through her feed to confirm that she had not mistakenly done so. Another patient worried that she had posted shameful content online and would screen record her internet use to later verify appropriate use [9].

However, advances in digital health have also introduced a wide range of technologies for the assessment and management of OCD. Passive monitoring tools, such as actigraphy and wearable devices, have been used to unobtrusively track sleep and activity patterns, often aligning closely with patient reports [10]. Active approaches, including ecological momentary assessment, allow for real-time symptom capture through smartphone prompts, offering insights into daily fluctuations that may be missed by retrospective recall [11]. Beyond monitoring, emerging interventions incorporate smartphone-based apps, cognitive training tools, and even multimodal systems that combine neurophysiological measures with behavioral data [12].

There exists a need to further understand patients' perspectives toward technology and how its design or features might interact with their OCD. It remains unclear what forms of technology are more likely to trigger obsessions or compulsions, how these symptoms manifest, and whether patients find such interactions distressing or relieving. This qualitative study aims to address these questions by examining the mechanisms through which technology may contribute to or reinforce obsessive-compulsive behaviors.

Methods

Study Design: Constructivist Grounded Theory

We conducted a phenomenological study aimed at understanding the ways in which individuals with lived experience of OCD uniquely interface and interact with technology. Methodology was guided by constructivist grounded theory, which frames analysis as dialogic and iterative, constructing an empirically grounded model of the phenomenon under study [13]. We used the COREQ (Consolidated Criteria for Reporting Qualitative Research) framework, a validated checklist for qualitative reporting, to guide and document important aspects of the research team, methodology, findings, and analysis [14]. The study was approved by the University of Southern California Institutional Review Board (UP-23-01094).

Recruitment and Sample

Participants were recruited through a convenience sampling method between May and December 2024 through online posts in OCD-related communities, including the International OCD Foundation (IOCDF) website, OCD SoCal (a local affiliate of the IOCDF), and existing OCD-focused research participant registries maintained by the laboratory. These online spaces included both moderated forums and institutional web pages aimed at providing information, treatment resources, or research participation opportunities for those with OCD. Recruitment materials included a standardized study announcement containing the study overview, eligibility requirements, and compensation details. Potential participants were routed to contact the study team via email for screening through a REDCap (Research Electronic Data Capture; Vanderbilt University) survey that gathered demographic and self-report data relevant to the inclusion criteria. Study personnel reviewed survey responses in order of submission and contacted participants who met study criteria to schedule an initial virtual meeting to confirm survey responses and obtain informed consent. Following completion of the study, participants were compensated with a US $50 gift card for their time and effort. Inclusion criteria included self-reported age of 18 years and older and a diagnosis of OCD. A total of 27 participants completed the consent process and were interviewed. All interviews were reviewed by study personnel trained in the phenomenology of OCD for content consistent with knowledge of and experience with OCD. A total of 3 interviews were excluded due to credible concerns regarding eligibility (eg, no clear indication of an OCD diagnosis and suspected repeat participation by a single individual). To mitigate this risk, we

XSL•FO
RenderX

subsequently implemented procedures requiring camera-on interviews and a brief discussion of participants' OCD symptoms with the interviewer prior to proceeding. No participants withdrew or declined participation after enrollment. A total of 7 participants had previous contact with laboratory personnel through other unrelated research studies.

Research Team

Data collection and analysis were conducted by a multidisciplinary team, including the principal investigator (ACF; MD, PhD, psychiatrist) and graduate and undergraduate research assistants. Multimedia Appendix 1 provides an overview of team demographics, training, and specific roles. Interviewers disclosed their role in the laboratory to participants at the beginning of each interview. Team members brought varied clinical, academic, and personal perspectives related to OCD and technology and had differing levels of qualitative research experience.

To ensure consistency and competency, all interviewers completed a minimum of 6 weeks of qualitative research training. Standardization efforts included review sessions of the interview guide, observation of senior interviewers, and postinterview debriefings.

Data Collection and Setting

Interviews were conducted via the HIPAA (Health Insurance Portability and Accountability Act)-compliant Zoom software (Zoom Communications) between May 16, 2024, and December 20, 2024. A primary facilitator led each interview, with additional team members present to assist with technical support or occasional follow-up questions. Only study participants and research team members were present during interviews. Interviews ranged from 35 to 65 minutes, with a median duration of 51 minutes.

All interviews were audio- and video-recorded, and transcripts were generated using Zoom's auto-transcription feature. Research assistants verified transcript accuracy and performed deidentification. All audio and video recordings were permanently deleted after transcript verification. No repeat interviews or member checking of transcripts were conducted.

Interview Guide

The semistructured interview guide was developed to explore participants' lived experiences with technologies used for health or health care, particularly in relation to OCD. The guide was informed by phenomenological and user-centered perspectives, emphasizing discovery, motivation, use patterns, benefits, concerns, and perceived interactions with OCD symptoms. While not based on a single theoretical model, the structure reflects elements common to sociotechnical and technology adoption frameworks, with a focus on how individuals make sense of and navigate these tools in daily life [15,16]. Interview questions were piloted within the research team and refined through early interviews as part of our grounded theory approach. Revisions were made iteratively as team discussions and early analysis revealed new areas of interest. A sample of the interview guide is provided in Multimedia Appendix 2.

Data Analysis: Coding

The first 5 transcripts were independently open-coded by pairs of researchers, where open-coding refers to the initial process of breaking data into discrete parts and assigning conceptual labels to segments of interest. Each pair met to discuss initial findings and reach consensus on a shared set of open codes for each transcript. These were then reviewed through group meetings to identify patterns and groupings, generating a set of focused codes informed by emerging theoretical insights that constructed the initial codebook. The coding structure is outlined in Multimedia Appendix 3. Transcripts were uploaded to Dedoose (version 9.2.22; Sociocultural Research Consultants), and the codebook was implemented in the software. The codebook was then applied to each transcript by a pair of researchers, and consensus meetings between coders followed to ensure agreement on code applications.

We operationalized theoretical saturation in our dataset by the inclusion of a "new finding" code in the codebook to capture data relevant to the research question but not represented by existing codes. After all transcripts were coded, the "new finding" excerpts were reviewed collectively to assess whether additional conceptual categories had emerged. New categories were tracked and subsequent emergence of related insights was collapsed, such that "new finding" applications represented only unique and uncaptured theoretical insights [17,18]. We considered interview number 15 the point of theoretical saturation as no subsequent "new finding" applications identified categories or relationships that would meaningfully alter the emerging theory (Multimedia Appendix 4). These discussions led to final adjustments to the codebook. All transcripts were then recoded retrospectively using the updated coding structure.

Data Analysis: Theory Generation

After applying the final codebook, the team engaged in an inductive analytic process to develop a theoretical framework describing a salient emerging topic from the data: understanding interactions between technology and OCD symptoms. Using constant comparison across transcripts and memos, we identified key conceptual categories, examined their interrelationships, and developed an articulation of the processes shaping participants' experiences.

Analysis began with targeted review of excerpts coded with conceptually related codes ("Exacerbating and enabling OCD symptoms with technology" and "Alleviating mental health and OCD symptoms with technology"). These were used to write analytical memos, which guided weekly team discussions and supported theoretical elaboration. Through this iterative process, we developed a conceptual framework that articulates relationships between specific types of technology, the symptom domains they interact with, and distinct stages of the OCD symptom cycle.

Ethical Considerations

The study was approved by the University of Southern California Institutional Review Board (UP-23 - 01094). All study procedures involving human participants adhered to the ethical standards of the institutional review board and the 1964 Declaration of Helsinki and its subsequent amendments. All

XSL•FO
RenderX

subjects received written information regarding study aims, procedures, potential risks, and anticipated benefits, which were reviewed in discussion with study personnel during the onboarding process. All participants provided informed consent before study enrollment. Participation was voluntary, and participants could withdraw at any time without penalty. Participant privacy and confidentiality were strictly protected, and all data were deidentified before analysis. Participants were compensated with a US $50 Tango gift card upon study completion.

Results

Participants

A description of the participants (n=24) is provided in Table 1. No participants refused to participate in the study or withdrew their participation.

Table . Participant demographics (N=24).

Characteristics	Results
Age (years), median (range)	26.2 (20.1 - 63.6)
Gender, n (%)	
Men	7 (29)
Women	16 (67)
Nonbinary	1 (4)
State, n (%)	
California	14 (58)
Pennsylvania	3 (13)
Illinois	2 (8)
Ohio	2 (8)
Florida	1 (4)
Maryland	1 (4)
Virginia	1 (4)

A coding tree with 5 parent codes and 7 child codes was developed during analysis. Among these, the parent codes Exacerbating and enabling OCD symptoms with technology and Alleviating OCD symptoms with technology were most central to addressing the research question and guided the organization of findings. From this coding structure, we identified four pertinent qualities of technology in relation to participants' symptomatic experiences: (1) information provision, (2) gamification and quantification, (3) notifications, and (4) user interfaces. Information-provision technologies had strong interactions with OCD symptoms consistent with the disturbing-thought domain, whereas the 3 other technology characteristics had interactions with both disturbing-thought and symmetry domains. Notably, the application of OCD symptom domains (eg, symmetry and disturbing–thought-based) was not deductively imposed during the interview process or initial codebook development. Rather, these domains were inductively derived from participants' narratives and symptom descriptions. While the interactions between technological characteristics and symptom manifestations emerged organically from the data, the terminology used to describe these domains is drawn from the existing OCD literature to ensure consistency and scientific clarity in reporting [2,3].

Result 1: Information-Provision Technologies and the Disturbing-Thought Domain of OCD

Across interviews, participants described information-providing technologies as implicated in the emergence and reinforcement of symptoms within the disturbing thought-based domain of OCD. These technologies, which range from social media and search engines to health portals and messaging apps, triggered intrusive, ego-dystonic thoughts, while also enabling compulsive behaviors such as checking, reassurance-seeking, and mental review.

Many participants emphasized the role of unsolicited and algorithm-driven content in triggering disturbing-thought symptoms. These triggers were often delivered unexpectedly via video thumbnails, pop-up ads, or headlines, bypassing user intention and presenting emotionally charged or ambiguous material. This dynamic was especially distressing for individuals with harm, health, or moral scrupulosity themes:

> *It just pops up on your page. It can be very triggering….I started getting videos of this one specific food that I really like… This patient that was eating this food a lot developed fatty liver disease and my OCD brain was like, 'oh no'… I've been spending hours every day since seeing that, you know, researching correct dietary choices, researching health.* [P3]

The above quote also highlights the role of information-providing technology in inducing compulsive checking or reviewing to gain certainty or reassurance in response to obsessional distress. In our data, obsessional distress was related to intrusive fears of past wrongdoing, undiagnosed illness, or social harm. Commonly implicated platforms used

XSL•FO
RenderX

for checking or reviewing included search engines, social media timelines, court databases, and health information portals:

> *I will get on my phone…look on social media for like dates and look up people and you know, think that I hurt this person or did I do this? Did I do that?…I'll look my name up on the sex offender registry and I'll be like, oh no, you know, am I on here?…all that information being out there just makes me want to search and search more.* [P2]

Technologies that allowed participants to upload information for others to interact with produced similar patterns in both triggering disturbing-thought obsessions and enabling compulsions. Written communication platforms such as email and messaging apps were particularly salient for participants with social or moral scrupulosity, a subtype of disturbing-thought OCD. Participants reported that the informational permanence of these technologies often led to obsessive worries about their perception by others, including harm, misunderstanding, or being morally inappropriate. Participants then described compulsively reviewing sent messages, analyzing others' tone, and fearing reputational or ethical harm:

> *Yes. I had, I guess I had some OCD or some anxiety anyway about the Teams chat at work that we use for workplace communication…I would go back and check to see…could this be taken [as] inappropriate by this person? And I'll, you know, check, check a few times.* [P1]

Participants who used information-providing technology compulsively often acknowledged that the boundaries between reassurance-seeking, educational inquiry, and compulsive overuse are often opaque:

> *I just compulsively look up health care information…I think initially like all compulsions, when they're carried out, it provides some reassurance, although that's not always healthy.* [P1]

In some cases, participants reported that an obsession-compulsion cycle would subside once they encountered information that felt sufficiently reassuring. However, the threshold for what constituted a "sufficient" or trustworthy source varied across individuals. Both participants P8 and P9 described using search engines to manage health-related obsessions but differed in their sense of reassurance from the information they found:

> *So, I guess just try to, like, put it out of my mind until I can actually talk with someone from the office. So just, like, distract myself, or I'll find something on the internet that will maybe explain it, if I feel like it, is the answer. But usually I'm never like. Usually I'm never satisfied, because, like different websites can say different things. So I mean, that can sometimes work like, if I feel like, okay, I have a reliable website, and you know it says this. Then I'll feel like I found, you know I have the answer or the explanation to it.* [P8]

> *So like, some websites are more trusted than others…But again…I have trust issues deep down. And so the information I'm getting, I don't feel confident in even though I go to town. As soon as I see it, I kind of feel better because, you know, it answers my question. Like I just got a question answered, which is an anxiety, like to not know…Then it's like, I need to keep going and look, look, look, look. And I go to multiple different lengths. So I guess there's really not a point where I feel confident in it. And because the point is that my brain goes into more and more and more possibilities…And then I end up just having a bunch of different things that could be wrong with me. Um, and then I message my doctor and…that's pretty much how that goes.* [P9]

The duality of relief and exacerbation that informational technologies evoke was complex and individualized. The type of clarifying or validating information was also varied and included peer connection, digital monitoring, and access to authoritative sources. Another nuanced experience of participants regarded their deliberate limiting or avoidance of technology platforms perceived as triggering. For some, this involved active efforts to reduce exposure to social media or other content-driven apps:

> *Yeah, social media, like, I have Instagram, kind of only of the social medias, and it's a constant fight. I try to use it as little as I can, down to like, I delete the app every few days…That interacts for sure with my OCD.* [P7]

However, participants also reflected on the complexity of avoidance itself. While reducing exposure to distressing content was sometimes framed as protective, others were careful to distinguish between harmful avoidance (which can reinforce OCD) and what they considered healthy boundary-setting:

> *I guess if we want to talk about media… I try to avoid the amount. And I shouldn't really say 'avoid,' I guess, 'cause that makes it sound like counterproductive avoidance, but more like a self-care kind of avoidance, if you will…I will kind of doom spiral and ruminate and obsess about things like that if I focus on it too much…I just wanted to clarify, not be counterproductive. As we all know, like, avoiding things to heal your OCD doesn't work. That just makes it worse in the long run. So that's why I don't mean like hiding away from it, I mean just having my own healthy boundaries with things.* [P5]

Across both positive and cautionary accounts, participants consistently emphasized how the structure and function of information-providing technologies, whether through unsolicited content or user-driven inquiry, directly shaped the phenomenology of their disturbing-thought–based OCD symptoms.

XSL•FO
RenderX

Result 2: Gamified and Quantified Technologies in Symmetry- and Disturbing-Thought–Based OCD

Several participants described how technologies that quantify behavior or incorporate gamified features—such as streaks, scores, and progress tracking—interacted with their OCD symptoms. These interactions most commonly aligned with symmetry- or NJR-based obsessions. In these cases, discomfort was often triggered by an unbalanced number, an incomplete goal, or an interrupted streak and was accompanied by a vague but compelling internal sense that something was unresolved. Several participants expressed this interaction with health and fitness tracking apps:

> *Sometimes, you are too addicted to getting the perfect [step count] scores…I need to achieve this target, although you might have been drained out and tired throughout the day. But you want to. You want that number to be complete in order to just, like, tick off. It feels like ticking off one of the checkboxes.* [P13]

Participants frequently acknowledged that the distress did not always stem from a clearly articulated fear or consequence. Instead, their behaviors were driven by an ambiguous internal pressure, with goals or numbers taking on a rigid symbolic weight:

> *I would say I'm kind of obsessed with trying to have at least the same or more number of steps every day… I didn't think about that, but that's absolutely true… So I have like, let's say 10,000 or 11,000 steps a day and when I see that it's less, I'm like, oh, no, you know, I need to go for a walk and just go get some groceries just to make it even neater…I check it very often… so that gives me some anxiety probably you know internally…I can't figure out, I mean, the thoughts behind that, like the obsession…* [P15]

The automatic provision of a step count number to the user both initiates distress and enables individuals to perform checking compulsions. This highlights the ambiguity often described by participants in relation to NJR feelings with a fixation on numbers and achievements. One participant powerfully captured how app-based gamification and quantification intersected with OCD perfectionism in a way that borders on overwhelming:

> *Anytime when there's a streak, like you've meditated 9 times in a row, or like the way the apps are kind of gamified. The worst for this is Duolingo…just to give an example where it's like, you have all the cauldrons of stars and bonuses and streaks, and to me, that adds a layer of gamification and stress and numbers and scoring. And related to my OCD perfectionism, that's not as constructive to me…I would say it interacts…when it seems to be a thing I can pass and fail at. Particularly because there are numbers involved and or a kind of like, gamified system…when it feels like that kind of contest…It makes it more complicated.* [P7]

While symmetry-related themes were most common in response to these technologies, some participants described gamified or quantified features as triggering distress related to disturbing-thought–based obsessions. In these cases, the compulsion was not driven by a need for balance or completion, but rather by intrusive fears about consequences if the behavior was not completed:

> *I used the Headspace app to just…calm the thoughts…, but it ironically turned into kind of an OCD pattern where I felt like I couldn't not do the Headspace thing…I remember I had like a 130-day streak, which is great. But I realized that there were some days where I didn't want to do it…[My OCD] kind of left me in a pattern where it was like, you need to do it, or you're gonna fall back, or you're gonna retreat to your old self…The first therapist I met with was the one who kind of helped me be like, it's okay if you miss a day or two—like, the world's not gonna end.* [P14]

This development of intrusive fears and compulsive use of a mental health app underscores both the seemingly innocuous sources of symptom onset and the dual potential of mental health technologies to be either helpful or harmful, depending on how they interact with a user's symptom profile.

Some participants note that there are qualities of quantification technology that exacerbate obsessive-compulsive symptomatology in ways that nondigital alternatives might not:

> *MyFitnessPal was the one where I realized, okay, I need to stop using this…everything I ate, I had to put in. I spent a bunch of time making sure I found as accurate as possible of a submission for what I ate that day…because if you're just journaling, you're not going to see your progress towards this [amount of calories], like if you've gone over by this amount…So it was just very automatic and specific. And I think that's kind of what led to it. Versus if I was journaling, I don't think I would have been as likely to be tracking it so tight.* [P11]

> *I can find myself like, really obsessed with my reading speed and the percentage that I finish a book. So, you can track both your reading speed and the percentage that you finish a book on Kindle versus, you know, paper, and I can find myself feeling stressed or, like, feeling unable to do anything else unless I complete my reading goal. Like, I make my reading goal to reach like 5% of book, for example. And I feel like I cannot stop reading until that's done and that's only on kindle. It's such a quantified thing.* [P4]

Together, these accounts illustrate how features like quantification, gamification, and automated feedback can interact with both symmetry-based and disturbing-thought-based OCD symptoms, transforming everyday technologies into sources of obsessional triggers and compulsive pressure.

Result 3: Notifications and the Urge to Clear, Know, and Control

Participants frequently described notifications as a trigger for compulsive interactions that spanned both NJR- and disturbing-thought–based OCD experiences. Participants described two distinct ways notifications interacted with OCD

XSL•FO
RenderX

symptoms. For some, notifications triggered a need for visual or numerical resolution such as clearing unread counts to zero or removing banners that disrupted the screen, which aligned with NJR experiences. For others, notifications provoked disturbing-thought–related fears about neglecting responsibilities or missing important information, reinforcing compulsive checking behaviors.

For participants whose distress aligned with symmetry-based experiences, notifications disrupted a visual or emotional sense of order:

> *I also have to make sure all the notification numbers on the apps are cleared every time I check it…my wallpaper is a graduation photo with my parents. If a notification pops up and covers my mom, I have to remove it to make sure the important parts of the photo are not covered.* [P6]

> *I get really overwhelmed by my computer and my phone and in relation to my work. It's a space of like, Slack messages that I'm obsessive about, and clearing my inbox to 0 and, you know, checking my email again and again… It becomes like interface [number] 13, that I need to have like cleared and sort of dealt with.* [P7]

For this participant, the presence of notifications created a sense of visual clutter or incompleteness that became intolerable, prompting frequent checking and clearing rituals.

Other participants described notification-related distress that aligned more closely with disturbing-thought–based OCD, particularly around fears of neglecting responsibilities or missing something socially or professionally important. The spontaneity and unpredictability of notifications compounded the sense of being overwhelmed.

> *I just feel a bit burdened about it. Oh, I have a lot of notifications to clear, I need to catch up on, just so that I'm informed. Oh, okay, all of these are there, and I've read through them. So at least I have the knowledge of what was in my notifications.* [P10]

This compulsion of checking and clearing reflects a drive to restore a sense of control by accounting for and resolving the uncertainty that notifications generate.

Result 4: User Interfaces and Experiences of Ordering and Overwhelm

Participants described how the interfaces of personal devices frequently triggered obsessive-compulsive symptoms. Features typically celebrated for enhancing usability, such as customization, integration, and expansive functionality, were often experienced as overwhelming or destabilizing. For individuals with OCD, these features presented countless opportunities to engage in compulsions related to control, ordering, and avoidance, particularly when the interface felt overly stimulating or difficult to contain. These experiences aligned with both symmetry and disturbing-thought–based OCD symptom domains.

For example, some participants described compulsive ordering rituals related to app layout and usage patterns:

> *So, in my OCD, there's a certain way I do things, like there's a certain way I wash my hands. So similarly for social media, I have a certain order in which I use the apps and a certain order and amount of time. On every app there is a routine for it, like, for example, Whenever I wake up, I'll first open Instagram, then LinkedIn, then Handshake, then maybe Snapchat. So, there is a particular order.* [P10]

While certain participants engaged compulsively with interfaces in ways that clearly aligned with symmetry-based symptoms, such as imposing order, others described distress in response to the layered, limitless nature of digital environments. These experiences were characterized by feelings of cognitive overload or a sense of losing control. However, the underlying sources of this distress were more difficult to categorize. For some participants, it appeared to reflect a symmetry-based desire for structure, simplicity, or containment. A participant described this experience while engaging with cloud-based storage systems:

> *If it's my file, it's just abstract. It's weirdly hard for me how everything is virtual, just like not having that order…I work with one organization where it's all through Google Drive. There are hundreds and hundreds of sub folders…Getting lost in these online labyrinths somehow overwhelms my OCD. I struggle when I can't, like, control. And maybe this speaks to apps in terms of simplicity of interface for me. I think that would be more soothing to my OCD, because it wouldn't have that effect like, "this is a thing with like 10 subfolders, and so many different places to explore"…Yeah, if there's too much to it, it can be overwhelming.* [P7]

This sense of digital overload prompted some participants to deliberately restrict the functionality of their devices or to seek out simplified alternatives. The goal was not merely to reduce screen time, but to avoid the sense of overstimulation caused especially by vast integration of technology. A participant reflected on the difference between two wearable devices, seemingly driven by a disturbing-thought fear such as missing important information or falling behind:

> *There's so many different ways to reach me. So that is something that I get very, internally stressed and overwhelmed with…That's one of my main symptoms, is that feeling of always having to catch up and keep up with all the side conversations [on my phone]…The Apple watch reflected my computer, my phone, everything, on my wrist, which was just too much for me. I find the Fitbit to be less invasive because I haven't fully integrated my whole phone into it. I really enjoy the Fitbit for the limited functions that I've got on it right now. I know that it has more capabilities, but I like that I haven't turned those on.* [P12]

Across interviews, participants described how the boundless, integrated nature of digital interfaces transformed their devices into overwhelming, even invasive, objects. Abundance itself became a source of distress as the increasing complexity and

XSL•FO
RenderX

functionality of user interfaces created countless opportunities for engagement that could both facilitate compulsions and trigger anxiety. For some, these technologies reinforced ordering or avoidance rituals; for others, they produced a chaotic sense of overstimulation and loss of control. Several individuals noted limiting or simplifying device functions as a self-protective strategy to reduce compulsive urges and restore a sense of control.

Discussion

Principal Findings

Our findings show that everyday technologies shape the expression, maintenance, and phenomenology of OCD in ways that are both domain-specific and cross-cutting. Information-providing technologies often amplified disturbing-thought symptoms by delivering unexpected or ambiguous content and by enabling compulsive reassurance-seeking, checking, and mental reviewing. Participants described a fluid boundary between appropriate information gathering and compulsive overuse, noting that algorithmic content, the permanence of written communication, and the abundance of online information all contributed to escalating cycles of uncertainty and distress.

Gamified and quantified technologies interacted with symmetry- and NJR-based symptoms by attaching symbolic weight to numbers, streaks, and completion metrics. These features provoked both an internal drive for balance or perfection and, in some cases, intrusive fears about the consequences of breaking a streak. Participants described how automatic feedback loops transformed neutral metrics into triggers for compulsive behavior.

Across symptom domains, notifications acted as potent cues for urges to clear, control, and know. Some participants experienced notifications as visual disorders that required resolution, while others felt compelled to check notifications out of fear they might overlook or miss essential information. These experiences were reflective of both symmetry-based obsessions and disturbing-thought-based obsessions.

Finally, the structure of the user interface itself emerged as a significant source of OCD-related distress. Highly integrated or complex digital environments elicited compulsive ordering, avoidance, or feelings of cognitive overload. Many participants sought simplified devices, simplified digital environments, or limited device functionality to regain a sense of control and reduce compulsions.

Together, these results illustrate how contemporary technologies deliver content and structure interactions in ways that directly shape OCD phenomenology. The design of digital environments, including informational density, automated feedback systems, and levels of integration across platforms, creates conditions in which obsessions can be triggered and compulsions reinforced. These findings point to technology as an active context in which OCD symptoms unfold, rather than a neutral backdrop. This perspective helps clarify why existing research has only begun to capture the breadth and nuance of these experiences, and it provides a foundation for situating our results within the emerging literature.

Previous Research

There is a relative dearth of literature exploring stakeholder perspectives on how OCD interacts with technology, both as a mental health disorder and a lived experience. Previous work has highlighted how symptoms could manifest in the digital realm, including intrusive worries about posting offensive content to the Internet or social media, with associated compulsive checking, screen capture, and reassurance seeking [9]. Expert opinion now suggests assessing for “digital obsessions and compulsions,” when evaluating patients with OCD [8]. A number of studies that used technology in the assessment and treatment of OCD also asked participants about their experiences using this technology; however, this information was not collected systematically and remained broadly focused on acceptability and tolerability [19-21]. Finally, one study explored experiences with online peer-support communities in individuals with OCD; thematic analysis revealed factors such as social comparison and misinformation as contributing to negative experiences [22].

This study sought to explore how individuals with OCD perceive and navigate their interactions with modern technology. Notably, neither our interview guide nor codebook was structured to map onto formal OCD dimensions such as those delineated by the Yale-Brown Obsessive Compulsive Scale Symptom Checklist [23]. Yet, clear and consistent patterns emerged that aligned with established symptom clusters, most notably the disturbing-thought (eg, harm, sexual, religious, and moral scrupulosity) and NJR domains. Participants described experiences that mirrored clinical subtypes, underscoring the validity of these dimensions and the extent to which they shape real-world behavior, including digital behavior.

Interestingly, symptoms in the domains of cleanliness, contamination, and excessive or ritualized washing or grooming were less prominent in our data. This absence may reflect that contamination concerns often center on physical contact with contaminants and health-related concerns generally have a focus on the physical body. However, mental contamination is a recognized phenomenon in OCD and shares features with contact contamination symptoms and symmetry and incompleteness symptoms [24]. While this was not present in our data, interactions between technology and mental contamination should be considered for further exploration. Finally, our findings complicate assumptions that certain domains are purely physical. For example, other studies have found that digital clutter can provoke similar emotional distress and difficulty discarding as does physical hoarding [25].

Below, we elaborate on the implications of these findings for 4 key stakeholder groups: patients, clinicians, technology developers, and researchers.

Implications for Patients

Technology can serve as an important source of community, a method for accessing educational and treatment resources, and an approach to monitoring symptoms for individuals with OCD. However, these same tools can exacerbate symptoms and

XSL•FO
RenderX

complicate treatment. Patients may benefit from intentionally monitoring how digital tools affect their emotional state and symptom patterns over time. We recommend that individuals reflect on the distinction between purposeful use of technology and patterns of behavior that drift into compulsive checking, reassurance seeking, or overmonitoring.

In practical terms, this may involve setting time- or context-based limits, pausing engagement with certain platforms during periods of heightened vulnerability, or discussing emerging online behaviors with a therapist to identify early signs of compulsive use. Patients should also be mindful that customization features, achievement systems, and constant access to information can shift from motivating to destabilizing without clear warning. We suggest open and ongoing discussions between patients and their treating clinicians regarding technology in daily life and OCD treatment.

We suggest framing technology not as inherently helpful or harmful, but as a continuum of use that can shift depending on symptom severity, emotional needs, and the design of specific platforms. Approaching digital engagement with this flexible, self-reflective mindset may help patients cultivate healthier, more sustainable relationships with the technologies that shape their daily lives.

Implications for Clinicians

For clinicians, assessment of OCD symptomatology is important, and we concur with previous studies that recommend explicit evaluation for digital behaviors [8]. This can be achieved by adding targeted questions (eg, “Do you feel compelled to check, post, or search online to relieve distress, even when it interferes with your life?”) or by supplementing existing measures with a brief checklist addressing technology-related compulsions, such as repetitive searching, reassurance-seeking, or monitoring notifications. Because current instruments do not systematically probe for these behaviors, clinicians should remain attentive to technology use that functions as a compulsion. It also remains unclear what proportion of OCD symptoms can be attributed uniquely to digital compulsions, and future studies should quantify this domain to inform updates of assessment tools.

With regard to treatment, clinicians should incorporate psychoeducation about how app design, social media features, or online communities can inadvertently reinforce obsessive-compulsive cycles. We suggest clinicians provide concrete examples to patients: streaks and push notifications can induce checking behaviors, social media can reinforce negative stereotypes and foster misinformation, and user interfaces can lead to a feeling of digital overwhelm. These examples can help patients recognize their own experiences and better understand these processes as environmental triggers rather than personal failings. Clinicians might also help patients identify strategies for more adaptive engagement, such as disabling notifications, setting time limits, or designating “offline” periods.

Cognitive-behavioral strategies, particularly exposure and response prevention, can be readily adapted to digital contexts. For instance, therapists might collaborate with patients to create graded exposure hierarchies that include delaying responses to notifications, resisting the urge to refresh a feed, or intentionally interrupting a streak or posting schedule. These exercises can be integrated into therapy sessions or assigned as structured homework.

Finally, clinicians should maintain flexibility as technology evolves. Integrating a brief review of digital behaviors into routine follow-up visits can help identify emerging risks, prevent relapse, and reinforce adaptive digital habits. Tracking these observations in deidentified case notes may also contribute valuable data for refining diagnostic instruments and treatment protocols.

Implications for Technology Developers

Our findings raise important considerations for those who design and develop digital platforms. Many features intended to enhance engagement, such as algorithmic personalization, goal-setting interfaces, streak counters, and push notifications, were described by participants as direct triggers of OCD symptoms.

Rather than removing these features entirely, developers could implement optional design accommodations that give users greater control over their interaction patterns. For example, adjustable interface settings could allow users to disable streak counters, hide quantified feedback (eg, progress bars and daily goals), or consolidate notifications into scheduled digests. Content algorithms might include transparency dashboards where users can view or modify personalization parameters, thereby increasing predictability and reducing uncertainty-driven compulsions. Similarly, introducing a “low-stimulation” or “minimal-feedback” mode could help reduce inadvertent reinforcement of repetitive checking behaviors.

An “OCD-friendly” design, in this context, would prioritize user autonomy, predictability, and control over feedback loops that otherwise promote compulsive engagement. These accommodations could be incorporated without removing core functionality but by expanding the range of interaction options available to all users.

Finally, the mental health technology sector, particularly apps targeting wellness, mindfulness, or productivity, could benefit from evaluating whether their engagement metrics inadvertently reinforce obsessive-compulsive symptom cycles. Collaborating with clinicians, behavioral scientists, and individuals with lived experience of OCD could help establish design standards that promote sustained engagement without exacerbating compulsive behaviors.

Implications for Researchers

This study opens several avenues for future research. There is a clear need to develop and validate new measurement tools that specifically assess the intersection of OCD symptoms and technology use. These tools could help clinicians differentiate between typical digital habits and pathological compulsions, especially in younger or digitally native populations.

Future studies might also explore how different OCD subtypes respond to specific forms of technology in more controlled, quantitative settings. For example, are individuals with NJR symptoms more sensitive to gamified or goal-based interfaces

XSL•FO
RenderX

than those with contamination concerns? Are certain design features (eg, infinite scroll and intermittent rewards) more strongly associated with symptom exacerbation in OCD than in other clinical populations?

Finally, longitudinal studies can address several outstanding questions. First, for individuals with a diagnosis of OCD, what is the time course and progression of technology-related symptoms? Cohort studies of individuals with OCD that include careful assessment for this symptom domain could improve our understanding of the evolution of technology-related obsessions and compulsions. Relatedly, our results demonstrate positive and negative aspects of technology use in adults with OCD; understanding how, and over what period, a neutral or beneficial use of technology transitions into problematic or compulsive use can aid in assessment and evaluation in clinical settings, as well as inform technology design. Finally, tracking technology use in the general population and monitoring for transition to compulsive use could help in identifying specific aspects of the technology and person that may predispose them to these behaviors. Studies in this domain could also assess for conversion to meeting formal OCD diagnostic criteria and evaluate progression of broader OCD symptoms in this population. Overall, understanding these dynamics will be essential as technology continues to evolve and embed itself more deeply into everyday life.

Limitations

This study has several limitations that should be considered when interpreting its findings. First, the demographics of our participants skewed young and were geographically concentrated in California. The median age of participants was 26.2 (range 20.1 - 63.6) years, 14 of the 24 participants resided in California, and 10 participants resided in other states. This geographic concentration likely reflects our recruitment strategies. Seven participants were drawn from other studies conducted at the University of Southern California that required in-person participation in Los Angeles. Additional participants were recruited online through local and national chapters of OCD advocacy organizations. The relatively young age of our sample may, in part, reflect these online recruitment strategies, as previous research has shown that younger adults spend more time online than older adults [26]. Other studies using online recruitment methods have also reported similarly young average participant ages [27-29]. The convenience sampling strategy we used may have contributed to this demographic profile. Regardless of the age profile and geographic distribution of our sample, qualitative research is designed to explore experiences, perspectives, and contexts in depth, rather than to produce findings that are statistically generalizable to a larger population. Its value lies in uncovering rich, detailed insights and generating understanding of complex phenomena, which can then inform theory, practice, or further research.

Second, our participant sample needed a baseline level of technological fluency. All participants were able to access digital recruitment materials, use Zoom for interviews, and speak fluently about their technology use. As a result, the findings may not fully capture the experiences of individuals with lower digital literacy, limited access to technology, or differing generational relationships with digital tools. Individuals for whom technology is so distressing that they cannot consistently use the internet or Zoom would also not be captured in our study.

Additionally, while participants self-identified as having a diagnosis of OCD, we did not conduct formal clinical assessments or structured diagnostic interviews to confirm a diagnosis of OCD or assess specific symptom subtypes. As such, some participant narratives may reflect overlapping symptomatology with related conditions, such as generalized anxiety. However, the use of self-reported diagnosis is an established approach in qualitative mental health research, where the analytic focus is on lived experience, identity, and help-seeking rather than diagnostic validation. Qualitative studies routinely use a self-reported mental health diagnosis as an inclusion criterion for participation, with these diagnoses ranging from OCD to depression to bipolar disorder [30-33]. In line with these precedents, our sampling strategy is appropriate for experiential inquiry, while still requiring cautious interpretation regarding the boundaries between OCD and overlapping or comorbid forms of distress. Additionally, research personnel reviewed transcripts for content consistent with obsessive and compulsive behaviors.

Our analysis was guided by a constructivist grounded theory approach, and results reflect the perspectives and interpretive lens of a research team composed primarily of nonclinicians. While this multidisciplinary team allowed for rich interpretive dialog, the extrapolation of findings was based on analytic interpretation that considered the existing literature on OCD but did not rely on a structured, validated metric. Specifically, aligning participant accounts with disturbing-thought–based or symmetry-based domains was an interpretive process shaped by reflexivity of researchers. Additionally, apparent ambiguity in the data may reflect other barriers to participant disclosure, such as limited insight or the stigmatization of OCD symptoms, rather than the absence of an experience.

Despite these limitations, the study provides novel insights into how individuals with OCD perceive and engage with technology and offers a foundation for future research to more systematically examine these dynamics.

Conclusions

This study highlights how digital technologies can both trigger and sustain OCD symptom patterns. Participants' accounts reveal that the design of these systems often mirrors OCD's own dynamics of uncertainty, reassurance-seeking, and control, blurring the boundary between pathology and platform. Clinically, these findings underscore the importance of assessing technology use as part of symptom formulation and helping patients develop strategies for intentional, rather than compulsive engagement. For designers and digital health developers, we highlight the need for interfaces that minimize reinforcement of compulsive behaviors and allow for user control over triggering features. More broadly, understanding OCD within its technological context reframes it as a disorder increasingly expressed through the architectures of modern attention, emphasizing the shared responsibility of clinicians,

XSL•FO
RenderX

researchers, and technologists in shaping environments that support, rather than exploit, cognitive vulnerability.

Acknowledgments

We would like to thank Dr Rachel Ceasar for invaluable consultation on initial formulations of our interview guide and study protocol. We would like to thank the following research assistants for their effort in data collection, interview transcription, and focused coding: Natalie Bakken, Trevor Bailey, Jenna Kim, Dejan Shakya, and Zoe Elliot. Emma Garland assisted with formatting, copy editing, and manuscript review.

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: literature search and systematization, proofreading and editing, and reformatting. The GenAI tools used were: GPT-4, GPT-4.5, and OpenEvidence. Responsibility for the final manuscript lies entirely with the authors. Gen AI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted with collective responsibility.

Funding

This work was funded by a Brain and Behavior Research Foundation Young Investigator Award to ACF (32015). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: ACF
Data curation: LO, ACF
Formal analysis: LO, KSP, KP, HW, DM, EL, ES, MM, TB, TV, UP, ACF
Funding acquisition: ACF
Investigation: LO, KSP, KP, EL, ES, MM, TB, TV, UP, ACF
Supervision: LO, UP, ACF
Writing – original draft: LO, KSP, KP, HW, DM
Writing – review and editing: ACF

Conflicts of Interest

None declared.

Multimedia Appendix 1
Study team demographics.
[DOCX File, 1811 KB - jmir_v28i1e85033_app1.docx]

Multimedia Appendix 2
Example semistructured interview guide.
[DOCX File, 32 KB - jmir_v28i1e85033_app2.docx]

Multimedia Appendix 3
Sample coding tree.
[DOCX File, 1809 KB - jmir_v28i1e85033_app3.docx]

Multimedia Appendix 4
Table of thematic saturation.
[DOCX File, 1808 KB - jmir_v28i1e85033_app4.docx]

References

1. Ruscio AM, Stein DJ, Chiu WT, Kessler RC. The epidemiology of obsessive-compulsive disorder in the National Comorbidity Survey Replication. Mol Psychiatry 2010 Jan;15(1):53-63. [doi: 10.1038/mp.2008.94] [Medline: 18725912]

XSL•FO
RenderX

2. Bloch MH, Landeros-Weisenberger A, Rosario MC, Pittenger C, Leckman JF. Meta-analysis of the symptom structure of obsessive-compulsive disorder. Am J Psychiatry 2008 Dec;165(12):1532-1542. [doi: 10.1176/appi.ajp.2008.08020320] [Medline: 18923068]
3. Mataix-Cols D, Rosario-Campos MD, Leckman JF. A multidimensional model of obsessive-compulsive disorder. Am J Psychiatry 2005 Feb;162(2):228-238. [doi: 10.1176/appi.ajp.162.2.228] [Medline: 15677583]
4. Cervin M, Miguel EC, Güler AS, et al. Towards a definitive symptom structure of obsessive-compulsive disorder: a factor and network analysis of 87 distinct symptoms in 1366 individuals. Psychol Med 2022 Oct;52(14):3267-3279. [doi: 10.1017/S0033291720005437] [Medline: 33557980]
5. Nagata JM, Lee CM, Hur JO, Baker FC. What we know about screen time and social media in early adolescence: a review of findings from the Adolescent Brain Cognitive Development Study. Curr Opin Pediatr 2025 Aug 1;37(4):357-364. [doi: 10.1097/MOP.0000000000001462] [Medline: 40172268]
6. Monteith S, Glenn T, Geddes JR, et al. Importance of patient online activities. Br J Psychiatry 2025 Oct;227(4):718-720. [doi: 10.1192/bjp.2025.32] [Medline: 40369901]
7. Torous J, Bucci S, Bell IH, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021 Oct;20(3):318-335. [doi: 10.1002/wps.20883] [Medline: 34505369]
8. Carmi L, Zohar J, Arush OB, Morein-Zamir S. From checking the door to checking the app: assessment and treatment implications for obsessive-compulsive disorder in the digital era. CNS Spectr 2021 Oct;26(5):457-458. [doi: 10.1017/S1092852920001509] [Medline: 32600488]
9. van Bennekom MJ, de Koning PP, Denys D. Social media and smartphone technology in the symptomatology of OCD. BMJ Case Rep 2018 Aug 28;2018:bcr2017223662. [doi: 10.1136/bcr-2017-223662] [Medline: 30154174]
10. Drummond LM, Wulff K, Rani RS, et al. How should we measure delayed sleep phase shift in severe, refractory obsessive-compulsive disorder? Int J Psychiatry Clin Pract 2012 Oct;16(4):268-276. [doi: 10.3109/13651501.2012.709866] [Medline: 22809128]
11. Gloster AT, Richard DCS, Himle J, et al. Accuracy of retrospective memory and covariation estimation in patients with obsessive-compulsive disorder. Behav Res Ther 2008 May;46(5):642-655. [doi: 10.1016/j.brat.2008.02.010] [Medline: 18417100]
12. Arevian AC, O'Hora J, Rosser J, Mango JD, Miklowitz DJ, Wells KB. Patient and provider cocreation of mobile texting apps to support behavioral health: usability study. JMIR Mhealth Uhealth 2020 Jul 29;8(7):e12655. [doi: 10.2196/12655] [Medline: 32723714]
13. Charmaz K. Constructivist grounded theory. J Posit Psychol 2017 May 4;12(3):299-300. [doi: 10.1080/17439760.2016.1262612]
14. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
15. Mohr DC, Schueller SM, Montague E, Burns MN, Rashidi P. The behavioral intervention technology model: an integrated conceptual and technological framework for eHealth and mHealth interventions. J Med Internet Res 2014 Jun 5;16(6):e146. [doi: 10.2196/jmir.3077] [Medline: 24905070]
16. Rahimi B, Nadri H, Lotfnezhad Afshar H, Timpka T. A systematic review of the technology acceptance model in health informatics. Appl Clin Inform 2018 Jul;9(3):604-634. [doi: 10.1055/s-0038-1668091] [Medline: 30112741]
17. Hennink MM, Kaiser BN, Marconi VC. Code saturation versus meaning saturation: how many interviews are enough? Qual Health Res 2017 Mar;27(4):591-608. [doi: 10.1177/1049732316665344] [Medline: 27670770]
18. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: a systematic review of empirical tests. Soc Sci Med 2022 Jan;292:114523. [doi: 10.1016/j.socscimed.2021.114523] [Medline: 34785096]
19. Rupp C, Falke C, Gühne D, Doebler P, Andor F, Buhlmann U. A study on treatment sensitivity of ecological momentary assessment in obsessive-compulsive disorder. Clin Psychol Psychother 2019 Nov;26(6):695-706. [doi: 10.1002/cpp.2392] [Medline: 31365952]
20. Olbrich H, Stengler K, Olbrich S. Smartphone based Geo-Feedback in obsessive compulsive disorder as facilitatory intervention: a case report. J Obsessive Compuls Relat Disord 2016 Jan;8:75-78. [doi: 10.1016/j.jocrd.2016.01.001]
21. Klein CS, Alt AK, Pascher A, et al. Cognitive behavioral therapy for pediatric obsessive-compulsive disorder delivered via internet videoconferencing: a manualized sensor-assisted feasibility approach. Child Adolesc Psychiatry Ment Health 2024 Dec 4;18(1):154. [doi: 10.1186/s13034-024-00844-7] [Medline: 39633405]
22. Tan YT, Rehm IC, Stevenson JL, De Foe A. Social media peer support groups for obsessive-compulsive and related disorders: understanding the predictors of negative experiences. J Affect Disord 2021 Feb 15;281:661-672. [doi: 10.1016/j.jad.2020.11.094] [Medline: 33234279]
23. Goodman WK, Price LH, Rasmussen SA, et al. The Yale-Brown Obsessive Compulsive Scale. I. Development, use, and reliability. Arch Gen Psychiatry 1989 Nov;46(11):1006-1011. [doi: 10.1001/archpsyc.1989.01810110048007] [Medline: 2684084]

XSL•FO
RenderX

24. Jacoby RJ, Blakey SM, Reuman L, Abramowitz JS. Mental contamination obsessions: an examination across the obsessive-compulsive symptom dimensions. J Obsessive Compuls Relat Disord 2018 Apr;17:9-15. [doi: 10.1016/j.jocrd.2017.08.005]
25. Luxon AM, Hamilton CE, Bates S, Chasson GS. Pinning our possessions: associations between digital hoarding and symptoms of hoarding disorder. J Obsessive Compuls Relat Disord 2019 Apr;21:60-68. [doi: 10.1016/j.jocrd.2018.12.007]
26. Gelles-Watnick R. Americans' use of mobile technology and home broadband. Pew Research Center. 2024 Jan 31. URL: https://www.pewresearch.org/wp-content/uploads/sites/20/2024/01/PI_2024.01.31_Home-Broadband-Mobile-Use_FINAL.pdf [accessed 2025-11-11]
27. Milczarski W, Borkowska A, Białek M. No evidence of risk aversion or foreign language effects in incentivized verbal probability gambles. Judgm decis mak 2025;20:e17. [doi: 10.1017/jdm.2025.3]
28. Gregory SEA. Investigating facilitatory versus inhibitory effects of dynamic social and non-social cues on attention in a realistic space. Psychol Res 2022 Jul;86(5):1578-1590. [doi: 10.1007/s00426-021-01574-7] [Medline: 34374844]
29. Le TP, Bradshaw BT, Pease M, Kuo L. An intersectional investigation of Asian American men's muscularity-oriented disordered eating: associations with gendered racism and masculine norms. Eat Disord 2022;30(5):492-514. [doi: 10.1080/10640266.2021.1924925] [Medline: 33998395]
30. Staiger T, Stiawa M, Mueller-Stierlin AS, et al. Masculinity and help-seeking among men with depression: a qualitative study. Front Psychiatry 2020;11:599039. [doi: 10.3389/fpsyt.2020.599039] [Medline: 33329149]
31. Morton E, Hole R, Murray G, Buzwell S, Michalak E. Experiences of a web-based quality of life self-monitoring tool for individuals with bipolar disorder: a qualitative exploration. JMIR Ment Health 2019 Dec 4;6(12):e16121. [doi: 10.2196/16121] [Medline: 31799936]
32. Arnáez S, Roncero M, López-Santiago J, et al. Fighting against self-stigma in adults with self-reported diagnosis of OCD: a single-arm pilot study using a mobile app-based intervention. Br J Clin Psychol 2025 Sep;64(3):788-805. [doi: 10.1111/bjc.12537] [Medline: 40065191]
33. Wairauch Y. Compulsive rituals in obsessive-compulsive disorder - a qualitative exploration of thoughts, feelings and behavioral patterns. J Behav Ther Exp Psychiatry 2024 Sep;84:101960. [doi: 10.1016/j.jbtep.2024.101960] [Medline: 38513433]

ABBREVIATIONS

COREQ: Consolidated Criteria for Reporting Qualitative Research
HIPAA: Health Insurance Portability and Accountability Act
IOCDF: International OCD Foundation
NJR: not-just-right
OCD: obsessive-compulsive disorder
REDCap: Research Electronic Data Capture

Edited by A Stone; submitted 30.Sep.2025; peer-reviewed by H Maheshwari, R Marshall; revised version received 19.Nov.2025; accepted 17.Dec.2025; published 05.Feb.2026.

Please cite as:
Occhino-Moede L, Sulivan-Pascual K, Phelan K, Wang H, Mokhtar D, Liu E, Schug E, Mirkis M, Baek T, Visher T, Pasupulety U, Frank AC
Interactions of Technology and Obsessive-Compulsive Disorder Symptomatology in Adults: Qualitative Interview Study
J Med Internet Res 2026;28:e85033
URL: https://www.jmir.org/2026/1/e85033
doi:10.2196/85033

© Lucas Occhino-Moede, Kaitlyn Sulivan-Pascual, Kendall Phelan, Harrison Wang, Daniel Mokhtar, Elisa Liu, Erica Schug, Megan Mirkis, Thomas Baek, Tamerlane Visher, Ujjwal Pasupulety, Adam Charles Frank. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessing Usage and Usability of a Narrative-Based Psychoeducational Digital Intervention to Improve Medication Adherence Among Individuals With Schizophrenia in a Stable Phase: Mixed Methods Study

Dian Zhu[1*], PhD; Fangyuan Chang[2*], PhD; Hongyi Yang[2], PhD; Yiwen Wei[2], MEng; Zhao Liu[2], PhD

[1]School of Art Design and Media, East China University of Science and Technology, Shanghai, China

[2]School of Design, Shanghai Jiao Tong University, No.800, Dongchuan RD, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Zhao Liu, PhD
School of Design, Shanghai Jiao Tong University, No.800, Dongchuan RD, Shanghai, China

Abstract

Background: Nonadherence to antipsychotic medication remains one of the most substantial challenges in the management of schizophrenia, contributing to relapse, rehospitalization, and functional decline. Although psychoeducational interventions are a key intervention for relapse prevention, traditional formats often lack interactivity and cultural resonance, thereby limiting engagement and sustained impact. Digital health innovations offer an opportunity to improve both treatment adherence and user experience, but evidence in schizophrenia populations remains limited.

Objective: This study aimed to evaluate the usage patterns, usability, and effectiveness of a narrative-based psychoeducational digital intervention designed to enhance medication adherence among individuals with schizophrenia in the maintenance phase. By employing a mixed methods design, the study integrated quantitative measures of adherence and functioning with qualitative insights into participants' experiences and perceptions.

Methods: A 6-month parallel mixed methods randomized controlled trial was conducted in community mental health settings in Shanghai. Seventy individuals with schizophrenia in a stable phase were randomly assigned (1:1) to the intervention group, which received the digital narrative-based psychoeducation application (Healing Town) in addition to routine community care, or to the control group, which received routine community rehabilitation only. Quantitative evaluation focused on medication adherence, drug attitude, social functioning, and psychiatric symptoms. In parallel, qualitative data were collected through semistructured interviews with patients, caregivers, and clinicians to examine intervention usage, usability, engagement, and perceived impact.

Results: Seventy participants (mean age 44.2, SD 8.057 y; 61% male) were enrolled, and 69 (98.6%) completed the 6-month trial, with one dropout during the intervention period. At 6 months, the intervention group showed significantly higher medication adherence (mean difference 1.27, 95% CI 0.30 - 2.24; P=.02) and more positive drug attitudes (mean difference 3.41, 95% CI 1.18 - 5.65; P=.002) compared with controls. Improvements in social functioning were significant within the intervention group (P=.03) but not between groups. No significant group differences were observed in psychiatric symptoms. Qualitative findings identified three overarching themes: (1) adherence and usability—patients reported enhanced treatment knowledge, confidence, and motivation, though some described challenges with feedback tone and pacing; (2) experiences and attitudes—users valued cultural relevance, immersive narratives, and gamified elements but noted occasional overstimulation; and (3) expectations and recommendations—participants expressed demand for personalized features, reminders, and dynamic content to sustain engagement.

Conclusions: This mixed methods study provides preliminary evidence that a narrative-based digital psychoeducational intervention may enhance medication adherence and attitudes toward medication among individuals with schizophrenia in the maintenance phase, while being perceived as engaging, usable, and culturally relevant. Furthermore, the qualitative findings suggest that supportive feedback, adaptive difficulty, and personalized features may enhance user motivation and optimize future scalability. Overall, this narrative-based digital psychoeducation represents a promising and potentially cost-effective approach to supporting community-based psychiatric rehabilitation, meriting further longitudinal and multisite investigation.

Trial Registration: ClinicalTrials.gov NCT06175559; https://clinicaltrials.gov/study/NCT06175559

(J Med Internet Res 2026;28:e59175) doi:10.2196/59175

XSL•FO
RenderX

KEYWORDS

schizophrenia; psychoeducation; medication adherence; cognitive behavior therapy; mixed methods study

Introduction

Schizophrenia is a chronic and severe mental disorder characterized by pervasive impairments in cognition, emotion, perception, and social functioning. Long-term use of antipsychotic medication is typically required to manage symptoms, prevent relapse, and reduce the risk of hospitalization [1-3]. However, despite advances in pharmacological treatments, approximately 50% of outpatients experience suboptimal medication adherence, which contributes to symptom exacerbation, frequent hospitalizations, and an elevated risk of suicide [4-6]. Improving adherence is not only critical for therapeutic success but also essential for enhancing long-term functional outcomes and quality of life [7].

Patients' subjective understanding of their illness and treatment plays a pivotal role in medication-taking behavior. Misconceptions about drug efficacy or side effects, limited coping strategies, and insufficient illness insight may all diminish motivation to adhere to prescribed regimens. Psychoeducation, a core intervention aimed at enhancing illness awareness and promoting adherence, has been widely implemented in schizophrenia care. It has demonstrated effectiveness in supporting symptom recognition, self-management, and functional recovery. Nevertheless, traditional psychoeducation programs often rely heavily on didactic instruction and lack interactivity, making them less suited for individuals who may experience impairments in attention, memory, and metacognition [8]. Moreover, impaired insight—often associated with prefrontal dysfunction—can lead to resistance toward biomedical explanations, especially when these conflict with patients' personal experiences or cultural beliefs [9,10].

In recent years, digital psychoeducation has shown great promise in enhancing treatment adherence among individuals with mental health conditions. Its advantages lie in offering highly interactive, contextualized, and personalized feedback. By integrating cognitive-behavioral strategies, gamified elements, and multimedia delivery, digital interventions have improved the accessibility of health information while enhancing user engagement and behavioral follow-through [11-13]. A particularly promising development is the emergence of narrative-based digital psychoeducation, which has demonstrated unique potential in reshaping health beliefs, eliciting emotional resonance, and strengthening treatment motivation [14]. Rooted in narrative psychology, this approach emphasizes that when individuals engage emotionally and cognitively in realistic, emotionally rich scenarios and characters, their beliefs about health, attitudes toward illness, and behavioral intentions can be reconstructed through simulated experience and observational learning [15]. Narrative interventions have shown encouraging outcomes in other areas of health education, particularly in sustaining behavior change over time [16].

Although some recent interventions targeting individuals with schizophrenia have begun to incorporate immersive, graphical narratives to enhance engagement and comprehension, this area of research remains in its early stages [17]. As Keats once suggested, narrative analysis should integrate spoken, written, and visual modalities to enable multidimensional expressions of complex psychological experiences [18]. However, many existing digital interventions still emphasize information delivery and sensory presentation, with limited attention to how patients construct meaning, express themselves, or reshape their attitudes toward treatment through narrative. This lack of integration between narrative structures and individual psychological mechanisms may undermine the effectiveness of such interventions in fostering motivation and medication adherence. In the context of long-term schizophrenia management, medication adherence is not only a key predictor of relapse and prognosis but also a reflection of the individual's cognitive, emotional, and attitudinal engagement with treatment. Therefore, conventional educational strategies alone may fall short in activating patients' intrinsic motivation and participatory engagement.

To address this gap, we developed an interactive, narrative-based digital application and conducted a 6-month randomized controlled trial (RCT) combined with qualitative interviews. Using a mixed methods design, we integrated quantitative and qualitative findings to evaluate the usability and user experience of the narrative-based digital psychoeducational application among individuals with schizophrenia in the maintenance phase. In addition, we examined how these user-centered outcomes were related to key 6-month end points, including medication adherence, social functioning, and clinical symptoms.

Methods

Study Design

A mixed methods approach was adopted in this study [19,20] following the ethical guidelines for human research outlined by the American Psychological Association [21].

In January 2023, five experienced psychiatrists contributed to the story script design, while three individuals with schizophrenia evaluated the visual style. After finalizing the script and visuals, the project moved to the application development phase.

Participants were recruited after providing written informed consent and screened for eligibility. Potential participants were identified through the Shanghai Mental Health Information Management System. Community psychiatrists screened eligible individuals and provided study information during routine follow-up visits. Those who expressed interest were subsequently contacted by research staff, who obtained written informed consent. The primary intervention phase was conducted between March 2024 and October 2024, lasting 6 months in total. During this period, participants attended five supervised in-person sessions each week, in which they completed the narrative-based psychoeducation program under therapist guidance (Figure 1). Baseline and postintervention

XSL•FO
RenderX

assessments were conducted. At baseline, participants completed demographic surveys and medication adherence scales. All participants were trained to use the application developed by the research team. The same assessments were repeated at subsequent time points. After the intervention, structured interviews were conducted using standardized procedures. These included think-aloud protocols and semistructured interviews, with responses recorded and later digitized. Audio recordings were made for further analysis.

Figure 1. Study design of the mixed methods study to evaluate a narrative-based psychoeducational digital intervention. QUAL: qualitative; QUANT: quantitative.

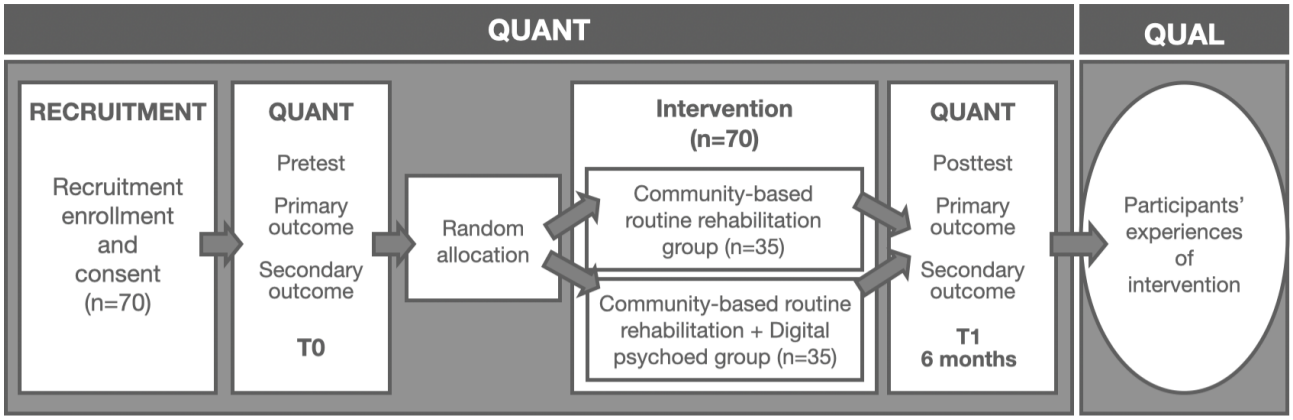

The quantitative component of this study was derived from a previously conducted parallel, two-arm RCT in Shanghai, implemented in accordance with the CONSORT (Consolidated Standards of Reporting Trials; Checklist 1) guidelines [22]. Participants were randomly assigned in a 1:1 ratio to either a community-based usual care group or a digital psychoeducational intervention group, with a 6-month intervention period. The usual care group received standard public health services, including regular home visits, participation in community mental health center rehabilitation programs, and education on maintaining medication adherence. In addition to these services, the intervention group participated in 5 weekly sessions of the narrative-based digital psychoeducation, each lasting 15 to 20 minutes. Each intervention site was staffed with one to two mental health professionals whose role was limited to resolving technical issues with the devices.

Following the 6-month quantitative assessment, a qualitative study was conducted in the intervention group. Semistructured interviews were carried out with all 35 users of the narrative-based digital psychoeducational intervention and 5 mental health professionals (with expertise spanning psychology, public health management, and nursing), ensuring that thematic saturation was achieved. Interviews took place between October 2024 and November 2024, conducted face-to-face in community mental health centers, and lasted an average of 15 minutes. All interviews were audio-recorded with consent, transcribed verbatim, deidentified prior to analysis, and guided by an interview protocol developed based on prior literature and expert consultation. The guide focused on usability, acceptability, user engagement, barriers and facilitators, as well as perceived impacts on medication-related behaviors.

Participant Recruitment

The study recruited individuals registered in Shanghai's community mental health system who had been diagnosed with schizophrenia and were receiving maintenance medication. The trial was conducted across 7 community rehabilitation centers administered by the Shanghai Mental Health Center. Shanghai has been recognized as a national leader in the management of schizophrenia, and digital technologies are increasingly being integrated into community-based services to support patient rehabilitation. Eligible participants were formally enrolled after providing written informed consent.

Stable-phase schizophrenia patients were selected as participants by psychiatrists from the Shanghai Mental Health Center. Trained psychiatrists conducted prospective monthly assessments of schizophrenia patients using the Positive and Negative Syndrome Scale (PANSS), with an evaluation period of at least 1 year. Patients were considered to have a stable condition if their total PANSS score did not change by more than 3 points and if there had been no changes in their medication treatment for at least 6 months prior to the neuropsychological assessment [23]. The inclusion and exclusion criteria are summarized in Table 1.

Participants were further required to be in a clinically stable phase, as determined by their PANSS scores.

XSL•FO
RenderX

Table . Inclusion and exclusion criteria for participant recruitment.

Domain	Inclusion criteria	Exclusion criteria
Registration and diagnosis	• Registered in the Shanghai Mental Health Information Management System • Diagnosed with schizophrenia according to the *International Classification of Diseases, 10th Revision* and determined to be in a clinically stable phase based on PANSS[a] assessment	• Planning to relocate outside of Shanghai • Presence of severe physical illness or organic brain disease • Comorbid with other psychotic disorders
Demographics and capacity	• Aged 18 - 60 y, with at least primary school education • Normal or corrected-to-normal vision and hearing • Ownership of and ability to independently operate a smartphone or other electronic device	• Participation in any intervention or treatment other than medication or basic public health services in the past 6 mo
Treatment status	• Receiving maintenance treatment with second-generation antipsychotics.	—[b]
Consent	• Participant or family member provided informed consent and signed the consent form.	—

[a]PANSS: Positive and Negative Syndrome Scale.

[b]Not available.

Sample Size Calculation

Sample size was calculated using G*Power 3.1 with a 2-tailed independent samples *t* test. Medication adherence scores were designated as the primary outcome measure. Based on a literature review and pilot study results, the expected between-group mean difference with a pooled SD was 0.78 (0.93), yielding an estimated effect size of approximately 0.83. With a significance level of α=.05 and a statistical power of 90%, the required minimum sample size was calculated to be 31 participants per group. Accounting for a 1:1 randomization ratio and a 20% dropout or loss to follow-up rate, the study required at least 76 participants (38 per group). Proportional sampling was conducted across districts based on the number of individuals with schizophrenia in each district.

Randomization, Allocation Concealment, and Blinding

After providing informed consent, participants were randomly assigned in a 1:1 ratio to the community-based routine rehabilitation group or the digital psychoeducation intervention group using a computer-generated random number table. Randomization ensured that baseline characteristics were evenly distributed between groups. These assessments aimed to evaluate the effectiveness of digital psychoeducation interventions in improving medication adherence and other health outcomes.

Intervention

Application Introduction

This study developed a graphic narrative-based digital psychoeducation application that integrates core concepts from narrative psychology and cognitive-behavioral theory. The intervention aims to enhance medication adherence and treatment motivation among individuals living with chronic-phase schizophrenia. Technologically, the application is grounded in interactive graphics, storyline-driven engagement, and behavioral tracking. It consists of three core modules: narrative storylines, cognitive training games, and self-monitoring logs. The intervention is set within fictional yet realistic life scenarios, allowing for high ecological validity. All user interactions—including choice selections, response times, and behavioral paths—are recorded in real time on the backend for subsequent behavioral analysis and feedback generation. The narrative pathways are dynamically open-ended, meaning that different choices lead to varying outcomes, thereby enhancing immersion and personalization (Multimedia Appendix 1).

Narrative Storytelling Psychoeducation

The narrative module is informed by the foundational principle of narrative psychology: individuals construct meaning, integrate experiences, and facilitate behavioral change through storytelling [24]. The application adopts a fictional framework wherein a "nonclinical character supports an individual experiencing schizophrenia," inviting users to engage in role-playing from a third-person perspective. Users observe the story's protagonist and assist them in making treatment-related decisions [25]. Although users are not directly narrating their own experiences, they project aspects of their own reality through simulated interactions, enabling externalization of internal struggles and fostering reappraisal of illness and self-identity. This indirect narrative mechanism encourages emotional involvement and character identification, aligning with the dual persuasive processes of *transportation* and *identification* proposed in narrative communication theory, which are known to support behavior change and motivational development [26,27].

XSL•FO
RenderX

To strengthen the educational impact and structural coherence of the storylines, the narratives were designed based on the classic 3-act structure and were iteratively refined under the guidance of psychiatric rehabilitation specialists and clinical psychiatrists. In total, 30 narrative threads were developed, each targeting specific objectives related to medication education (Multimedia Appendix 2).

Cognitive Games

The cognitive training game is designed to enhance individuals' ability to apply cognitive skills in daily life, thereby supporting functional recovery. Codeveloped by clinical psychiatrists and psychiatric rehabilitation experts, this module includes a series of real-world scenario-based tasks—such as grocery shopping or ordering food at a restaurant—that simulate daily challenges and promote skills such as information integration, planning, and decision-making (Multimedia Appendix 2). The games adopt an adaptive learning mechanism in which task difficulty is dynamically adjusted based on the user's performance. This ensures an optimal balance between challenge and attainability, thereby maintaining engagement and reinforcing learning outcomes.

Beyond strengthening executive function, attention, and problem-solving abilities, this module promotes the transfer of cognitive strategies to real-world environments, offering targeted support for rehabilitation in ecologically relevant contexts.

Self-Management Support

The self-management module enables users to schedule medication intake, access educational content related to medication adherence, and input personalized data—such as the daily dosage and timing of medications. The platform facilitates routine tracking of medication usage and provides accessible information on potential side effects and recommended coping strategies. It also includes educational materials highlighting the risks associated with medication nonadherence, including symptom relapse.

Through structured daily logs, users are encouraged to monitor their behavioral patterns, emotional states, and medication routines, thereby fostering self-awareness and self-regulation. This module supports the development of autonomous health behaviors, which are critical for sustaining long-term treatment engagement and improving functional outcomes.

Outcome Measures

Quantitative Outcomes

Morisky Medication Adherence Scale

The Morisky Medication Adherence Scale (MMAS) [28-31] comprises eight items. The total score ranges from 0 to 8, with higher scores indicating better medication adherence.

Drug Attitude Inventory

The Drug Attitude Inventory (DAI) [32] consists of 10 items, of which 6 are positively scored (correct answers score 1 point and incorrect answers score −1 point), while 4 are reverse-scored (correct answers score −1 point and incorrect answers score 1 point). Higher scores indicate a more favorable attitude toward medication.

Social Disability Screening Schedule

The Social Disability Screening Schedule (SDSS) [33] includes 10 items rated on a 3-point scale (0="no impairment," 2="severe impairment," and 9="not applicable"). The total score ranges from 0 to 20, with higher scores indicating greater social disability. The evaluation will be performed by trained psychiatric professionals at the time of assessment.

Brief Psychiatric Rating Scale

The Brief Psychiatric Rating Scale (BPRS) [34] evaluates 5 factors: anxiety and depression, anergia, thought disturbance, activation, and hostility or suspiciousness. It consists of 18 items scored on a 7-point Likert scale (1="not present" and 7="extremely severe"), with a total score ranging from 18 to 126. Higher scores indicate greater symptom severity. The evaluation will be performed by trained psychiatric professionals at the time of assessment.

Qualitative Outcomes

Participants engaged in semistructured and unstructured interviews accompanied by professional physicians and caregivers. These interviews aimed to explore participants' attitudes toward the application and its impact on medication adherence. Professional health care providers were also interviewed to evaluate the effectiveness of this kind of intervention. At the end of the intervention, patients were surveyed on their satisfaction with the psychoeducation based on several specific questions (Textbox 1).

XSL•FO
RenderX

Textbox 1. Semistructured interview guide used to explore usability and user experience of a narrative-based psychoeducational digital intervention.

Survey on the impact of psychoeducation on medicine adherence (participants)

1. What is the frequency at which the application is utilized?
2. Did you experience any moments during the course of this intervention when you felt the urge to quit up? Why?
3. What motivates or hinders you?
4. Has your medication usage changed as a result of the intervention? How so, if at all. If not, then explain.

Attitudes toward the digital psychoeducation (participants)

1. What is your opinions on such intervention measures?
2. Do you find narratives that are fundamental to be engaging?
3. Which of the chapters that you most impressive?
4. What aspect did you appreciate the most? (scenes, voice-overs, mini-digital psychoeducation interactions, narratives, and characters)

Evaluation of the effectiveness of the digital psychoeducation (mental health professionals)

1. Evaluate the level of patient allure exhibited by the intervention.
2. Kindly assess the efficacy in promoting adherence to medication.
3. In your opinion, which elements are linked to medication adherence? Why?
4. Would you recommend the intervention to your patients?

During each visit, brief semistructured interviews were conducted and audio-recorded to evaluate participants' attitudes toward the digital psychoeducation. During the baseline visit, interview questions focused on general feedback regarding the digital psychoeducation, including its narrative and visual elements. After the intervention, these questions were revisited, along with additional questions addressing specific digital psychoeducation features, such as the self-management check-in system.

Additionally, backend application usage data were collected, including the date and time of app usage and the story paths chosen within the intervention. Researcher YW analyzed each participant's experimental sessions by linking the time of app usage to their final story path selections.

Data Analysis

Data Integration

A sequential explanatory mixed methods design was adopted. Quantitative data from the RCT were first analyzed to assess changes in medication adherence, attitudes, and clinical outcomes. Subsequently, qualitative interviews were conducted to explore participants' experiences and to explain the quantitative results. Integration occurred during the interpretation phase through a triangulation process, in which findings from both data sources were compared and merged to identify areas of convergence, complementarity, and divergence.

Quantitative Analysis

All quantitative survey data were analyzed using SPSS version 26.0 (IBM Corporation). Categorical variables were presented as frequencies and percentages, while continuous variables were expressed as means (SD). Comparisons of categorical variables between groups were conducted using χ^2 tests or Fisher exact tests. For continuous variables, independent sample *t* tests or Mann-Whitney *U* tests were used for between-group comparisons, depending on data distribution. Within-group comparisons were performed using paired *t* tests or Wilcoxon signed rank tests. Baseline characteristics between the intervention and control groups were compared using chi-square tests or independent sample *t* tests. The effects of the intervention were evaluated by comparing baseline and postintervention data using independent sample *t* tests or nonparametric alternatives. All statistical tests were 2-tailed, and a *P* value of <.05 was considered statistically significant. No missing data were observed in this study.

Qualitative Analysis

The transcripts were analyzed in NVivo (version 12; QSR International) through an inductive form of content analysis [35]. At the beginning of the analysis, DZ and FC extracted text segments from 35 participants' transcripts to ensure consistency in coding. They then turned to analyze the rest of the transcripts separately. The records were initially coded to gain insights into the participants' thoughts and behaviors during the intervention process. This involved line-by-line review, extraction of relevant text segments, and identification of codes (n=49), with a focus on participants' attempts at taking action and their choices. Subsequently, subthemes (n=27) related to different attitudes toward adopting digital psychoeducation to enhance compliance were identified through discussions with participants and experts. The authors then collectively discussed the codes and subthemes, further refining them into broader themes (n=3). Examples of data extraction are listed in Multimedia Appendix 2. This approach allowed us to discern participants' varying attitudes toward the digital psychoeducation and their implementation of medication practices and processes. Any discrepancies that emerged during the analysis were discussed with the rest of the authors. To ensure the validity of the codes, the authors employed member checking by reaching out to

XSL•FO
RenderX

participants to validate interpretations of the data [36]. Specifically, comprehensive results were provided to participants to verify the consistency and accuracy of their experiences with the data.

Ethical Considerations

This study was conducted in accordance with the ethical principles of the Declaration of Helsinki and the American Psychological Association's guidelines for research with human participants. The research protocol was reviewed and approved by the Shanghai Jiao Tong University Human Research Ethics Committee (H20230207I). The trial was prospectively registered on ClinicalTrials.gov (NCT06175559). No exemption from ethical review was sought or applied.

All participants, or their legal guardians when appropriate, were provided with detailed verbal and written information about the study procedures, potential risks, and benefits. Written informed consent was obtained prior to enrollment. For participants with limited decision-making capacity, consent was additionally confirmed by a family member or caregiver. The informed consent process covered both primary data collection and the use of deidentified data for secondary analyses without requiring further consent. To protect privacy and confidentiality, all personal identifiers were removed from the study records. The digital psychoeducational application logged only anonymized behavioral data, which were stored in encrypted servers with access restricted to the research team. Interview transcripts were deidentified prior to the analysis. Participants received a gift card valued at approximately US $30 upon completing the study as compensation for their time and effort.

Results

Descriptive Analysis

From March 2024 to October 2024, a total of 82 community-dwelling individuals with schizophrenia undergoing rehabilitation were recruited. Following screening and assessment, 3 participants did not meet the inclusion criteria, and 7 were excluded due to their transition to or planned transition to long-acting injectable medications. The final expected enrollment number was 72 participants (Figure 2). Based on the residential communities of these 72 participants, local intervention sites were identified. Two participants voluntarily withdrew from the study, citing the distance to the intervention site, leaving 70 participants (35 in the community-based rehabilitation group and 35 in the digital psychoeducation intervention group), who completed baseline assessments. During the intervention period, no participants dropped out, and all 70 participants completed the postintervention assessment. The exceptionally low dropout rate can likely be explained by the enforcement of national health policies in China, which require individuals with a confirmed diagnosis of schizophrenia to regularly attend community rehabilitation services as part of their standard care.

Figure 2. CONSORT (Consolidated Standards for Reporting Trials) flow diagram of participant recruitment of a narrative-based psychoeducational digital intervention. BPRS: Brief Psychiatric Rating Scale; DAI: Drug Attitude Inventory; MMAS-8: 8-item Morisky Medication Adherence Scale; SDSS: Social Disability Screening Schedule.

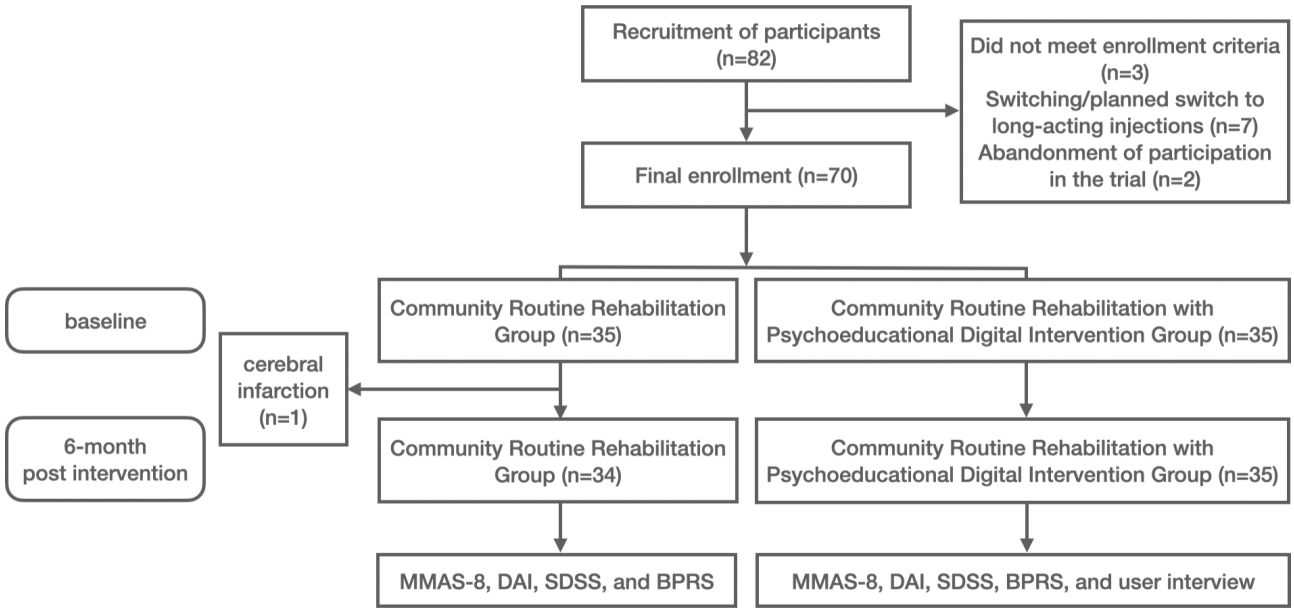

Characteristics of the Study Participants

A total of 70 participants were included in the study, with 43 male patients and 27 female patients. The mean (SD) age of the participants was 44.20 (8.057) years. The distribution of demographic characteristics is shown in Tables 2 and 3. The participants were randomly assigned to either the control group or the intervention group, with 35 individuals in each group. Independent samples *t* test results showed no significant difference in age between the 2 groups ($P>.05$). Chi-square test results indicated no significant differences between the two groups in terms of gender, education level, marital status, or living situation ($P>.05$). Baseline data are provided in Multimedia Appendix 3.

XSL•FO
RenderX

Table . Baseline demographic and clinical characteristics of participants with stable-phase schizophrenia.

Participants	All (N=70)	Control group (n=35)	Experimental group (n=35)	χ^2/*t* test[a] (df)	*P* value
Sex, n				0.54 (1)	.46
Male	43	20	23		
Female	27	15	12		
Age (y), mean (SD)	44.20 (8.06)	44.29 (9.10)	11 (6.99)	0.09 (68)	.93
Education, n				3.29 (2)	.19
Primary school	2	0	2		
Middle school	45	21	24		
College or above	23	14	9		
Marital status, n				0.51 (2)	.77
Married	11	5	6		
Divorced and widowed	10	6	4		
Unmarried	49	24	25		
Residence status, n				0.22 (1)	.64
Living alone	5	3	2		
Living with family	65	32	33		

[a]*t* tests were used for continuous variables (eg, age), and chi-square tests were used for categorical variables (eg, gender, education level, marital status, and residence status).

Table . Demographic characteristics of two participating mental health professionals.

Mental health professionals	Major	Gender	Age (y)	Years of experience
D1	Nursing	Female	39	15
D2	Public health management	Female	32	6
D3	Public health management	Male	41	12
D4	Psychology	Male	30	2
D5	Psychology	Female	42	15

Analysis of Intervention Results

This study evaluated differences between the intervention and control groups across multiple outcome measures, including medication adherence, attitudes toward medication, social functioning, and clinical symptoms (Table 4). The intervention group demonstrated significant improvement in medication adherence. At postintervention, a statistically significant between-group difference was observed, with a mean difference of 1.27 (95% CI 0.30-2.24; *P*=.02). Regarding attitudes toward medication, the intervention group also showed a significant improvement compared to the control group, with a mean difference of 3.41 (95% CI 1.18-5.65; *P*=.002).

XSL•FO
RenderX

Table . Changes in medication adherence, medication attitude, social functioning, and clinical symptoms from baseline to 6 months in a randomized controlled trial.

Characteristics	Intervention group, mean (SD)	Within-group *P* value	Control group, mean (SD)	Within-group *P* value	Mean difference (95% CI)	Between-group *P* value
Medication adherence		.03[a]		.11	1.27 (0.30 to 2.24)	.02[a]
Pre	6.91 (1.66)		7.23 (1.22)			
Post	7.66 (0.49)		6.68 (2.05)			
Medication attitude		<.001[b]		.47	3.41 (1.18 to 5.65)	.002[a]
Pre	4.97 (3.64)		5.83 (3.12)			
Post	8.91 (1.90)		6.35 (3.89)			
Social functioning		.03[a]		.85	−1.91 (−8.55 to 4.73)	.34
Pre	2.20 (2.83)		2.20 (2.79)			
Post	1.06 (1.97)		2.88 (3.68)			
Clinical symptoms		.06		.66	−7.88 (−34.90 to 19.14)	.12
Pre	28.37 (16.04)		25.77 (11.08)			
Post	21.37 (3.49)		26.68 (8.48)			

[a]*P*<.05.

[b]*P*<.001.

Qualitative Feedback and User Experience

Among the secondary outcomes, participants in the intervention group exhibited a notable within-group improvement in social functioning (*P*=.03). However, no statistically significant between-group differences were found in social functioning or clinical symptom severity (*P*>.05).

Application Usage (Back-End Data)

Analysis of back-end data indicated that 34 of 35 (97.14%) participants completed all game content. However, participants exhibited relatively high error rates in decisions related to certain topics, highlighting areas for targeted education in future interventions.

Alcohol Consumption and Medication

Although 25 of 35 (71.43%) participants selected the correct response—“Alcohol should be avoided, as it interacts with antipsychotic medications causing central nervous system depression, and excessive drinking may trigger other severe drug side effects”—a notable proportion (13/35, 37.14%) chose incorrect options, such as “All medicines are toxic, better not take them at all” or “Life is short, it’s fine to skip once in a while.” This suggests persistent misconceptions regarding alcohol–medication interactions.

Medication When Going Out

Some participants (8/35, 22.86%) initially chose unsafe options when deciding whether to take medication while going out, including “Skipping occasionally is fine” or “It’s okay to skip while traveling.” After experiencing negative story consequences, participants corrected their choices to the appropriate response: “Ensure to bring and take medications on time to avoid missed doses.”

Medication Side Effects

Most participants (31/35, 88.57%) correctly responded to in-story side effects by selecting “Consult a professional promptly if experiencing discomfort.” Nevertheless, a subset (7/35, 20%) initially chose the incorrect response, “It’s just lack of exercise, so I can endure it.”

Overall, these back-end data reveal both high engagement with the application and specific knowledge gaps, particularly regarding alcohol use, medication adherence while traveling, and interpretation of side effects. These findings underscore key areas for future educational emphasis.

Result of the Interview

Overview

The quantitative analysis revealed that the intervention significantly improved medication adherence and attitudes toward medication compared with standard community rehabilitation. The qualitative findings complemented these results by illustrating how narrative immersion and interactive gameplay enhanced users’ motivation and engagement, thereby reinforcing the behavioral changes observed in the quantitative phase. The demographic and clinical characteristics of the respondents are shown in Table 2. Thematic saturation was achieved after approximately 25 interviews, when no substantially new codes emerged. The remaining 10 interviews confirmed and enriched existing themes, particularly regarding variations in family support and illness duration. The specifics of the qualitative findings are displayed in Multimedia Appendix 4. Three main higher-order themes emerged from the participants’ and experts’ accounts (Multimedia Appendix 5).

XSL•FO
RenderX

Adherence and Usability

This theme encompassed factors related to medication adherence and the usability of the narrative-based digital intervention. Adherence was strengthened through increased medication knowledge, enhanced confidence in recovery outcomes, and active engagement with interactive narratives. Participants generally reported that the narrative intervention promoted adherence by enhancing knowledge about medication and building awareness of illness consequences, going beyond the functions of traditional reminder tools. This cognitive restructuring process appeared to activate intrinsic motivation. One participant explicitly noted that the intervention content changed their pattern of avoiding medication: "I used to avoid taking my medication due to side effects. After using this app, I realized that failing to reach the prescribed dose could have real consequences. Now I remind myself more consistently to follow the plan." Another participant stated, "I used to reduce doses or skip medication on my own… but now I feel more capable of managing my condition and trusting my treatment plan."

Positive reinforcement, such as the perceived "increase in medication knowledge" and "improvement in health status" after completing each narrative scenario, significantly enhanced participants' self-efficacy and encouraged continued engagement. Additionally, the intervention's self-management interface, which clearly displayed daily progress in medication-taking and task completion, provided ongoing external motivation: "Every time I log in, I feel like I'm working toward my health, which really motivates me." However, feedback tone and clarity were sensitive factors that could influence continued engagement. When participants made incorrect choices, the perceived harshness of outcomes sometimes triggered overstimulation and negative emotions, leading to tension and feelings of being scolded ("A bad ending made me angry, and I didn't understand why I was wrong"), highlighting the importance of strategic and tactful feedback.

Experiences and Attitudes Toward the Digital Psychoeducation

This theme revealed that participants expressed positive attitudes toward the narrative-based intervention, particularly appreciating its novelty and immersive quality. They described this learning approach as "not like a class, more like living inside a story," which significantly enhanced the learning experience. High acceptance was closely related to cultural sensitivity in the intervention, such as familiar home environments, clothing styles, and locally grounded story backgrounds. Participants noted, "The character design feels like people I really know… it is very relatable and comforting." Psychologically, the narrative also provided participants with important emotional connection and a sense of inclusion. Some participants reported feeling socially isolated in real life, but nonplayer characters within the intervention offered companionship and acceptance: "I always feel like people around me don't like me or want to talk to me… only they (in the app) are willing to talk to me."

Despite the overall cultural relevance of the narrative content, limitations remained at the individualized level. Some participants indicated that the narrative did not fully reflect their specific clinical or demographic characteristics, resulting in reduced relevance for certain content. One male participant noted, "Not all themes relate to my personal experiences, and not all side effects happened to me," suggesting that highly themed or specific content may create cognitive distance or misalignment with other user groups.

Expectations for the Digital Psychoeducation

This theme explored participants' suggestions regarding design elements and interactive features of the intervention, aimed at enhancing understanding of medication use. While initial experiences were positive, both participants and therapists emphasized the importance of ongoing content updates and functional completeness for maintaining long-term adherence. One participant remarked, "After completing all the stories, there's nothing new. I hope it could include more everyday situations, like working or traveling with friends." Due to the relatively simple content, one participant completed all narrative scenarios within 6 weeks, resulting in diminished interest later. The rapid consumption of content led some participants to discontinue usage after story updates ceased: "Towards the end, only the daily check-in feature was available, no new storylines. So, I just stopped engaging with the intervention."

From a functional perspective, participants repeatedly emphasized the need for external reminders, not only for new story releases but also for synchronized medication taking, requesting that the app "send notifications like other apps." Therapists highlighted the feasibility of future technological integration, suggesting "It would be great if AI could generate more interactive stories rather than having doctors spend time creating scripts," to ensure a continuous content supply and reduce the burden on clinical staff, thereby enabling sustainable integration of the system. Finally, some participants noted frustration with the lack of a review function: "Sometimes I have to interrupt the narrative due to other things happening, and I forget what happened. When I come back to the story, I realize that progress from before that level wasn't saved…" indicating the need to optimize memory and review mechanisms.

Discussion

Principal Findings

This study evaluated the usability, user experience, and behavioral effects of a narrative-based digital psychoeducational intervention developed for individuals with chronic schizophrenia in community settings. Using a mixed methods design, our findings suggest that, compared with standard community rehabilitation, the intervention significantly improved medication adherence and attitudes toward medication, while no significant changes were observed in clinical symptoms. Complementary qualitative findings further highlighted that participants valued the narrative elements for enhancing knowledge, motivation, and engagement, while also reporting challenges such as negative emotional reactions to corrective feedback. Most participants emphasized that maintaining long-term content engagement remained a concern, underscoring the need for future digital designs to incorporate personalized, accessible, and supportive features.

XSL•FO
RenderX

Specifically, following the narrative-driven digital psychoeducational intervention, participants exhibited notable improvements in medication adherence and medication attitude. Qualitative interviews provided contextual explanations for these quantitative findings, illustrating how the intervention influenced behavior through narrative mechanisms. Immersive storylines and interactive feedback enhanced participants' understanding of medication routines and their sense of control, leading to better adherence and fewer unplanned treatment interruptions. For example, some users mentioned that after completing the story, they "felt responsible for the character's health," reflecting a motivational shift from passive medication-taking to active self-management. This indicates that the digital format not only increased engagement but also fostered emotional resonance and cognitive involvement, supporting the observed behavioral improvements [37,38]. Care providers noted that integrating the intervention into repetitive cognitive rehabilitation sessions encouraged sustained participation in regular health-promoting behaviors. These improvements may be attributed to embedded mechanisms of emotional engagement and cognitive participation within the system [39]. Consistent with prior research, digital technologies such as smartphone applications have been found acceptable and feasible for individuals with psychotic disorders [40]. Existing mobile applications targeting medication adherence in schizophrenia mainly address logistical and memory-related barriers through reminders, dose tracking, and external reinforcement [41,42]. However, for individuals with schizophrenia, key barriers to treatment adherence are often internal, psychological, and cognitive, such as limited insight, misconceptions about side effects, perceived stigma, and reduced intrinsic motivation for long-term therapy [43]. Thus, interventions targeting patients' deeper cognitive representations of illness and treatment offer a promising direction for narrative-based approaches [44]. Immersive narratives provided a judgment-free, low-pressure environment in which participants could temporarily shift from the passive patient role to that of an active, rational decision-maker, engaging in simulated choices about medication and recovery. Many participants perceived the scenarios as familiar and personally relevant, which appeared to reduce resistance toward illness-related content and strengthen their willingness to engage in medication behaviors [41,42]. Although no significant between-group differences were observed in clinical symptoms, within-group improvements were identified. Qualitative narratives illuminated these patterns, as some participants described transient negative emotions during feedback or difficulties sustaining motivation over time, suggesting that behavioral and attitudinal improvements may precede measurable symptom changes [43,44]. Also, given the long-term trajectory of schizophrenia recovery, these findings collectively suggest that increased persistence and motivation could represent intermediate mechanistic changes that may facilitate long-term clinical benefits [45-47].

Regarding user experience and attitudes, most participants expressed positive feedback toward the modular narrative structure, describing the intervention as immersive and engaging. These qualitative findings complemented the quantitative results on adherence and medication attitude, suggesting that enhanced engagement and motivation may have mediated the observed behavioral improvements. By framing participants' self-representation as part of a healthy population, the interactive design reduced perceived stigma and promoted active knowledge acquisition and health behavior. Therapists noted that the intervention embedded medication-related knowledge through a sequence of stories, situational decision points, and real-time feedback, enabling participants to rehearse treatment-related decisions in a safe, low-pressure environment [48]. This narrative-based approach not only facilitated knowledge acquisition but also preserved users' autonomy and sense of realism. The usability and emotional engagement of the intervention appear to underpin the cognitive and motivational mechanisms that supported adherence improvement. As Stephens [49] emphasized, narratives serve not only as a means of organizing experience but also as a tool for identity construction and social positioning. By persuading a virtual character to "take medication on time," participants effectively constructed a socially valued identity—one characterized by cognitive competence and responsibility. This process represents a form of narrative repositioning, in which individuals psychologically shift from passive recipients of care to active agents of self-management, thereby enhancing perceived control and treatment motivation [50].

However, some users reported that certain narratives reflecting specific cultural or gendered contexts limited identification and engagement, highlighting the need for more personalized and culturally adaptive content in future iterations. Additionally, motivational activation was influenced not only by the narrative content but also by functional design elements such as pacing, reward systems, and feedback mechanisms. Participants reported confusion when storylines became overly abstract or fragmented, particularly during periods of cognitive fatigue, emphasizing the importance of clarity and simplicity in interface design for individuals with psychotic disorders. Thus, in addition to ensuring psychological relevance and cultural appropriateness, system design must prioritize simplicity and error tolerance [48]. Future development should continue to involve diverse stakeholders—including therapists, patients, and family members—for iterative discussion and codesign [51]. Mechanistically, participants indicated that the cognitive game components were seamlessly integrated with the narrative elements, enabling repeated practice in recognizing and reframing unhelpful automatic thoughts [42]. The task-based structure and immediate feedback combined behavioral activation with cognitive restructuring, helping participants modify maladaptive thinking patterns and rebuild medication-related cognitive frameworks. This mechanism aligns with the principles of cognitive behavioral therapy, which seeks to enhance adherence by reducing cognitive resistance and emotional conflict while strengthening behavioral consistency and motivation [52,53]. Finally, achievement-based feedback, milestone rewards, and task-driven structures created a sustained motivational environment that supported engagement even when immediate reinforcement was limited [54].

This mixed methods study provides preliminary evidence that the narrative-based digital psychoeducational intervention tested in this study demonstrated good usability and acceptability and may support short-term improvements in medication adherence

XSL•FO
RenderX

and attitudes toward medication among individuals with chronic schizophrenia. Although short-term symptom outcomes did not show significant improvement, the intervention addressed key behavioral and motivational processes critical for long-term rehabilitation. These findings emphasize the potential of this narrative-based, culturally contextualized digital tool to complement traditional community rehabilitation and to inform the development of more sustainable models of psychiatric care. Future research should extend follow-up periods to examine delayed clinical effects, incorporate adaptive and AI-driven narrative content to enhance cultural applicability, and test scalability across different health systems and populations.

Limitations

This study has several limitations. First, the financial incentives provided to participants may have facilitated their continued engagement. However, the primary purpose of these incentives was to compensate participants for their time and travel expenses. The amount of compensation was limited and unlikely to produce long-term changes in medication-taking behavior, suggesting that observed improvements in adherence are more likely attributable to the intervention mechanisms themselves. Nevertheless, we acknowledge that financial incentives may have modestly increased engagement, representing a potential confounding factor. Future studies could consider designing conditions without incentives or employing stratified analyses to control for this potential confounder. Second, all outcome measures relied on self-report, introducing the risk of subjective bias. The involvement of experienced psychiatrists may help mitigate such bias. Third, participants' familiarity and comfort with using smartphones to play the game varied, which could have influenced the effectiveness of the intervention. Additionally, as individuals in the intervention group were aware that they were receiving a digital intervention aimed at improving medication adherence, the absence of participant blinding may have introduced expectancy bias. Finally, while cultural customization likely enhanced the acceptability of the intervention within Chinese community settings, it may limit generalizability to more diverse or international populations. Future iterations of narrative-based psychoeducation should consider incorporating cross-cultural adaptation processes or leveraging AI-driven adaptive narratives to dynamically generate culturally relevant storylines.

Despite these limitations, the findings offer meaningful insights into the potential of narrative-driven digital psychoeducation to enhance medication adherence among individuals with schizophrenia in the long-term chronic phase. The intervention's integration of immersive storytelling, cognitive training, and self-management features presents a promising direction for future psychosocial support tools.

Conclusion

Our mixed methods findings indicate that the narrative-based digital psychoeducational intervention developed in this study demonstrated good usability and acceptability among individuals with chronic schizophrenia and may support short-term improvements in medication adherence and attitudes toward medication. Mobile health interventions of this kind could serve as a valuable complement to traditional community rehabilitation by enhancing behavioral and motivational mechanisms that contribute to long-term recovery. Nevertheless, careful attention should be paid to feedback design and cultural adaptation of narrative content, and tasks should be tailored to users' preferences and cognitive characteristics to ensure sustained usability and long-term effectiveness of the intervention.

Acknowledgments

©MMAS 2006 was used with permission of www.adherence.cc. The Morisky Medication Adherence Scale was used under a paid license granted by Dr. Donald E. Morisky and his colleagues (License Number 4525-8490-4067-7009-3166).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Video of Healing Town.
[MP4 File, 50167 KB - jmir_v28i1e59175_app1.mp4]

Multimedia Appendix 2
Content introduction of Healing Town.
[DOCX File, 3520 KB - jmir_v28i1e59175_app2.docx]

Multimedia Appendix 3
Data at baseline.
[DOCX File, 12 KB - jmir_v28i1e59175_app3.docx]

Multimedia Appendix 4
Themes of the interview.
[DOCX File, 24 KB - jmir_v28i1e59175_app4.docx]

XSL•FO
RenderX

Multimedia Appendix 5
High-order themes emerged from narrations.
[DOCX File, 16 KB - jmir_v28i1e59175_app5.docx]

Checklist 1
CONSORT-EHEALTH checklist (V 1.6.1).
[PDF File, 1222 KB - jmir_v28i1e59175_app6.pdf]

References

1. McCutcheon RA, Reis Marques T, Howes OD. Schizophrenia-an overview. JAMA Psychiatry 2020 Feb 1;77(2):201-210. [doi: 10.1001/jamapsychiatry.2019.3360] [Medline: 31664453]
2. van Os J, Kenis G, Rutten BPF. The environment and schizophrenia. Nature 2010 Nov;468(7321):203-212. [doi: 10.1038/nature09563]
3. Lançon C, Auquier P, Reine G, Bernard D, Addington D. Relationships between depression and psychotic symptoms of schizophrenia during an acute episode and stable period. Schizophr Res 2001 Mar 1;47(2-3):135-140. [doi: 10.1016/s0920-9964(00)00002-5] [Medline: 11278130]
4. Masand PS, Roca M, Turner MS, Kane JM. Partial adherence to antipsychotic medication impacts the course of illness in patients with schizophrenia. Prim Care Companion J Clin Psychiatry 2009;11(4):147-154. [doi: 10.4088/PCC.08r00612] [Medline: 19750066]
5. Cañas F, Alptekin K, Azorin JM, et al. Improving treatment adherence in your patients with schizophrenia: the STAY initiative. Clin Drug Investig 2013 Feb;33(2):97-107. [doi: 10.1007/s40261-012-0047-8] [Medline: 23288695]
6. Higashi K, Medic G, Littlewood KJ, Diez T, Granström O, De Hert M. Medication adherence in schizophrenia: factors influencing adherence and consequences of nonadherence, a systematic literature review. Ther Adv Psychopharmacol 2013 Aug;3(4):200-218. [doi: 10.1177/2045125312474019] [Medline: 24167693]
7. Ben-Zeev D, Scherer EA, Gottlieb JD, et al. mHealth for schizophrenia: patient engagement with a mobile phone intervention following hospital discharge. JMIR Ment Health 2016 Jul 27;3(3):e34. [doi: 10.2196/mental.6348] [Medline: 27465803]
8. Lincoln TM, Wilhelm K, Nestoriuc Y. Effectiveness of psychoeducation for relapse, symptoms, knowledge, adherence and functioning in psychotic disorders: a meta-analysis. Schizophr Res 2007 Nov;96(1-3):232-245. [doi: 10.1016/j.schres.2007.07.022] [Medline: 17826034]
9. Herrera SN, Sarac C, Phili A, et al. Psychoeducation for individuals at clinical high risk for psychosis: a scoping review. Schizophr Res 2023 Feb;252:148-158. [doi: 10.1016/j.schres.2023.01.008] [Medline: 36652831]
10. Çetin N, Aylaz R. The effect of mindfulness-based psychoeducation on insight and medication adherence of schizophrenia patients. Arch Psychiatr Nurs 2018 Oct;32(5):737-744. [doi: 10.1016/j.apnu.2018.04.011] [Medline: 30201202]
11. Al-Shashani A, Abu Sabra MA, Al-Gamal E. The Impact of using digital health interventions and psychoeducation on medication adherence among patients with schizophrenia: a scoping review. Issues Ment Health Nurs 2025 Jul;46(7):735-745. [doi: 10.1080/01612840.2025.2492694] [Medline: 40300193]
12. El-Mallakh P, Findlay J. Strategies to improve medication adherence in patients with schizophrenia: the role of support services. Neuropsychiatr Dis Treat 2015;11:1077-1090. [doi: 10.2147/NDT.S56107] [Medline: 25931823]
13. Cahaya N, Kristina SA, Widayanti AW, Green J. Interventions to improve medication adherence in people with schizophrenia: a systematic review. Patient Prefer Adherence 2022;16:2431-2449. [doi: 10.2147/PPA.S378951] [Medline: 36072918]
14. Li PWC, Yu DSF, Yan BP, Wong CW, Yue SCS, Chan CMC. Effects of a narrative-based psychoeducational intervention to prepare patients for responding to acute myocardial infarction: a randomized clinical trial. JAMA Netw Open 2022 Oct 3;5(10):e2239208. [doi: 10.1001/jamanetworkopen.2022.39208] [Medline: 36306128]
15. Bellack AS, Gold JM, Buchanan RW. Cognitive rehabilitation for schizophrenia: problems, prospects, and strategies. Schizophr Bull 1999;25(2):257-274. [doi: 10.1093/oxfordjournals.schbul.a033377] [Medline: 10416730]
16. Hinyard LJ, Kreuter MW. Using narrative communication as a tool for health behavior change: a conceptual, theoretical, and empirical overview. Health Educ Behav 2007 Oct;34(5):777-792. [doi: 10.1177/1090198106291963] [Medline: 17200094]
17. Landa Y, Mueser KT, Wyka KE, et al. Development of a group and family-based cognitive behavioural therapy program for youth at risk for psychosis. Early Interv Psychiatry 2016 Dec;10(6):511-521. [doi: 10.1111/eip.12204] [Medline: 25585830]
18. Keats PA. Multiple text analysis in narrative research: visual, written, and spoken stories of experience. Qual Res 2009 Apr;9(2):181-195. [doi: 10.1177/1468794108099320]
19. Johnson RB, Onwuegbuzie AJ, Turner LA. Toward a definition of mixed methods research. J Mix Methods Res 2007 Apr;1(2):112-133. [doi: 10.1177/1558689806298224]
20. Creswell JW, Clark VLP. Designing and Conducting Mixed Methods Research: Sage Publications; 2017:434-436.
21. Ethical principles of psychologists and code of conduct. : American Psychological Association; 2016 URL: https://www.apa.org/ethics/code/ethics-code-2017.pdf [accessed 2025-11-25] [doi: 10.1037/14805-030]

XSL•FO
RenderX

22. Liu Z, Zhu D, Zhang Z, et al. Effects of a narrative-based psychoeducational intervention on medication adherence in individuals with schizophrenia: a multicentre, parallel-group randomised controlled trial. EClinicalMedicine 2025 Oct;88:103483. [doi: 10.1016/j.eclinm.2025.103483] [Medline: 40979217]
23. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr Bull 1987;13(2):261-276. [doi: 10.1093/schbul/13.2.261] [Medline: 3616518]
24. Adams DM, Mayer RE, MacNamara A, Koenig A, Wainess R. Narrative games for learning: testing the discovery and narrative hypotheses. J Educ Psychol 2012;104(1):235-249. [doi: 10.1037/a0025595]
25. Cutting JE. Narrative theory and the dynamics of popular movies. Psychon Bull Rev 2016 Dec;23(6):1713-1743. [doi: 10.3758/s13423-016-1051-4] [Medline: 27142769]
26. Zhou C, Occa A, Kim S, Morgan S. A meta-analysis of narrative game-based interventions for promoting healthy behaviors. J Health Commun 2020;25(1):54-65. [doi: 10.1080/10810730.2019.1701586] [Medline: 31829829]
27. Levesque DA, Van Marter DF, Schneider RJ, et al. Randomized trial of a computer-tailored intervention for patients with depression. Am J Health Promot 2011;26(2):77-89. [doi: 10.4278/ajhp.090123-QUAN-27] [Medline: 22040388]
28. Oliveira-Filho AD, Barreto-Filho JA, Neves SJF, Lyra Junior DD. Association between the 8-item Morisky Medication Adherence Scale (MMAS-8) and blood pressure control. Arq Bras Cardiol 2012 Jul;99(1):649-658. [doi: 10.1590/s0066-782x2012005000053] [Medline: 22688844]
29. Moon SJ, Lee WY, Hwang JS, Hong YP, Morisky DE. Accuracy of a screening tool for medication adherence: a systematic review and meta-analysis of the Morisky Medication Adherence Scale-8. PLoS ONE 2017;12(11):e0187139. [doi: 10.1371/journal.pone.0187139] [Medline: 29095870]
30. Wang J, Bian RW, Mo YZ. Validation of the Chinese version of the eight-item Morisky medication adherence scale in patients with type 2 diabetes mellitus. J Clin Gerontol Geriatr 2013 Dec;4(4):119-122. [doi: 10.1016/j.jcgg.2013.06.002]
31. Krousel-Wood M, Islam T, Webber LS, Re RN, Morisky DE, Muntner P. New medication adherence scale versus pharmacy fill rates in seniors with hypertension. Am J Manag Care 2009 Jan;15(1):59-66. [Medline: 19146365]
32. Hogan TP, Awad AG, Eastwood R. A self-report scale predictive of drug compliance in schizophrenics: reliability and discriminative validity. Psychol Med 1983 Feb;13(1):177-183. [doi: 10.1017/s0033291700050182] [Medline: 6133297]
33. Shen YC. Epidemiological study of mental disorders in 12 regions of China: methodology and data analysis. Zhonghua Shen Jing Jing Shen Ke Za Zhi 1986 Apr;19(2):65-69. [Medline: 3743252]
34. Overall JE, Gorham DR. The Brief Psychiatric Rating Scale (BPRS): recent developments in ascertainment and scaling. Psychopharmacol Bull 1988;24(1):97-99. [Medline: 3387516]
35. Graneheim UH, Lindgren BM, Lundman B. Methodological challenges in qualitative content analysis: a discussion paper. Nurse Educ Today 2017 Sep;56:29-34. [doi: 10.1016/j.nedt.2017.06.002] [Medline: 28651100]
36. Carter N, Bryant-Lukosius D, DiCenso A, Blythe J, Neville AJ. The use of triangulation in qualitative research. Oncol Nurs Forum 2014 Sep;41(5):545-547. [doi: 10.1188/14.ONF.545-547] [Medline: 25158659]
37. Grey M, Whittemore R, Jeon S, et al. Internet psycho-education programs improve outcomes in youth with type 1 diabetes. Diabetes Care 2013 Sep;36(9):2475-2482. [doi: 10.2337/dc12-2199] [Medline: 23579179]
38. Alvarez-Jimenez M, Alcazar-Corcoles MA, González-Blanch C, Bendall S, McGorry PD, Gleeson JF. Online, social media and mobile technologies for psychosis treatment: a systematic review on novel user-led interventions. Schizophr Res 2014 Jun;156(1):96-106. [doi: 10.1016/j.schres.2014.03.021] [Medline: 24746468]
39. Killikelly C, He Z, Reeder C, Wykes T. Improving adherence to web-based and mobile technologies for people with psychosis: systematic review of new potential predictors of adherence. JMIR Mhealth Uhealth 2017 Jul 20;5(7):e94. [doi: 10.2196/mhealth.7088] [Medline: 28729235]
40. Lim MH, Penn DL. Using digital technology in the treatment of schizophrenia. Schizophr Bull 2018 Aug 20;44(5):937-938. [doi: 10.1093/schbul/sby081] [Medline: 29878251]
41. Gumley AI, Bradstreet S, Ainsworth J, et al. The EMPOWER blended digital intervention for relapse prevention in schizophrenia: a feasibility cluster randomised controlled trial in Scotland and Australia. Lancet Psychiatry 2022 Jun;9(6):477-486. [doi: 10.1016/S2215-0366(22)00103-1] [Medline: 35569503]
42. Garety P, Ward T, Emsley R, et al. Effects of SlowMo, a blended digital therapy targeting reasoning, on paranoia among people with psychosis: a randomized clinical trial. JAMA Psychiatry 2021 Jul 1;78(7):714-725. [doi: 10.1001/jamapsychiatry.2021.0326] [Medline: 33825827]
43. Maurin KD, Girod C, Consolini JL, et al. Use of a serious game to strengthen medication adherence in euthymic patients with bipolar disorder following a psychoeducational programme: a randomized controlled trial. J Affect Disord 2020 Feb 1;262:182-188. [doi: 10.1016/j.jad.2019.10.008] [Medline: 31668996]
44. Pan MR, Huang F, Zhao MJ, Wang YF, Wang YF, Qian QJ. A comparison of efficacy between cognitive behavioral therapy (CBT) and CBT combined with medication in adults with attention-deficit/hyperactivity disorder (ADHD). Psychiatry Res 2019 Sep;279:23-33. [doi: 10.1016/j.psychres.2019.06.040] [Medline: 31280035]
45. Robinson D, Woerner MG, Alvir JM, et al. Predictors of relapse following response from a first episode of schizophrenia or schizoaffective disorder. Arch Gen Psychiatry 1999 Mar;56(3):241-247. [doi: 10.1001/archpsyc.56.3.241] [Medline: 10078501]

XSL•FO
RenderX

46. Ascher-Svanum H, Faries DE, Zhu B, Ernst FR, Swartz MS, Swanson JW. Medication adherence and long-term functional outcomes in the treatment of schizophrenia in usual care. J Clin Psychiatry 2006 Mar;67(3):453-460. [doi: 10.4088/jcp.v67n0317] [Medline: 16649833]
47. Onitsuka T, Hirano Y, Nakazawa T, et al. Toward recovery in schizophrenia: current concepts, findings, and future research directions. Psychiatry Clin Neurosci 2022 Jul;76(7):282-291. [doi: 10.1111/pcn.13342] [Medline: 35235256]
48. Ma Z, Zytko D. Designing immersive stories for health: choosing character perspective based on the viewer's modality. Int J Hum Comput Interact 2021 Sep 14;37(15):1423-1435. [doi: 10.1080/10447318.2021.1886486]
49. Stephens C. Narrative analysis in health psychology research: personal, dialogical and social stories of health. Health Psychol Rev 2011 Mar;5(1):62-78. [doi: 10.1080/17437199.2010.543385]
50. Bai GN, Wang YF, Yang L, Niu WY. Effectiveness of a focused, brief psychoeducation program for parents of ADHD children: improvement of medication adherence and symptoms. Neuropsychiatr Dis Treat 2015;11:2721-2735. [doi: 10.2147/NDT.S88625] [Medline: 26604761]
51. Howard GS. Culture tales: a narrative approach to thinking, cross-cultural psychology, and psychotherapy. Am Psychol 1991 Mar;46(3):187-197. [doi: 10.1037//0003-066x.46.3.187] [Medline: 2035929]
52. Garety PA, Fowler D, Kuipers E. Cognitive-behavioral therapy for medication-resistant symptoms. Schizophr Bull 2000;26(1):73-86. [doi: 10.1093/oxfordjournals.schbul.a033447] [Medline: 10755670]
53. Westra HA. Comparing the predictive capacity of observed in-session resistance to self-reported motivation in cognitive behavioral therapy. Behav Res Ther 2011 Feb;49(2):106-113. [doi: 10.1016/j.brat.2010.11.007] [Medline: 21159325]
54. Olivet J, Haselden M, Piscitelli S, et al. Results from a pilot study of a computer-based role-playing game for young people with psychosis. Early Interv Psychiatry 2019 Aug;13(4):767-772. [doi: 10.1111/eip.12556] [Medline: 29542863]

Abbreviations

BPRS: Brief Psychiatric Rating Scale
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
DAI: Drug Attitude Inventory
MMAS: Morisky Medication Adherence Scale
PANSS: Positive and Negative Syndrome Scale
RCT: randomized controlled trial
SDSS: Social Disability Screening Schedule

Edited by A Stone; submitted 02.Jul.2025; peer-reviewed by AAL Sawafi, M Tusconi; revised version received 10.Nov.2025; accepted 10.Nov.2025; published 20.Jan.2026.

Please cite as:
Zhu D, Chang F, Yang H, Wei Y, Liu Z
Assessing Usage and Usability of a Narrative-Based Psychoeducational Digital Intervention to Improve Medication Adherence Among Individuals With Schizophrenia in a Stable Phase: Mixed Methods Study
J Med Internet Res 2026;28:e59175
URL: https://www.jmir.org/2026/1/e59175
doi:10.2196/59175

© Dian Zhu, Fangyuan Chang, Hongyi Yang, Yiwen Wei, Zhao Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effectiveness of and Mechanisms of Change in a Self-Help Web- and App-Based Resilience Intervention on Perceived Stress in the General Working Population: Randomized Controlled Trial

Sandy Hannibal[1], MSc; Dörte Behrendt[1], MSc; Michèle Wessa[2,3,4], PhD; Sarah K Schäfer[5,6], PhD; Nina Dalkner[7], PhD; Dirk Lehr[1], PhD

[1]Department of Health Psychology and Applied Biological Psychology, Institute of Sustainability Psychology, Leuphana University of Lüneburg, Universitätsallee 1, Lüneburg, Germany

[2]Department of Neuropsychology and Psychological Resilience Research, Central Institute of Mental Health, Mannheim, Germany

[3]Research Division Cancer Survivorship and Psychological Resilience, German Cancer Research Center (DKFZ)-Hector Cancer Institute at the University Medical Center Mannheim, Mannheim, Germany

[4]Division Cancer Survivorship and Psychological Resilience, German Cancer Research Center (DKFZ) Heidelberg, Heidelberg, Germany

[5]Department of Clinical Psychology and Psychotherapy for Children and Adolescents, Technische Universität Braunschweig, Braunschweig, Germany

[6]Leibniz Institute for Resilience Research, Mainz, Germany

[7]Division of Psychiatry and Psychotherapeutic Medicine, Medical University of Graz, Graz, Austria

Corresponding Author:
Sandy Hannibal, MSc
Department of Health Psychology and Applied Biological Psychology, Institute of Sustainability Psychology, Leuphana University of Lüneburg, Universitätsallee 1, Lüneburg, Germany

Abstract

Background: Promoting individual resilience—that is, maintaining or regaining mental health despite stressful circumstances—is regarded as an important endeavor to prevent mental illness. However, digital resilience interventions designed to enhance mental health have yielded mixed results. Such heterogeneous effects reflect a variety of unsolved conceptual challenges in interventional resilience research. These range from grounding interventions in resilience frameworks, using theory or targeting etiologically important resilience factors as intervention content, to a lack of knowledge about the mechanisms underlying effects, and using techniques specifically developed to foster psychosocial resources. The web- and app-based resilience intervention RESIST was designed to address these challenges, mainly by using both the Positive Appraisal Style Theory of Resilience as its theoretical foundation and interventional techniques from Strengths-Based Cognitive Behavioral Therapy.

Objective: This study's primary aim was to evaluate the effectiveness of RESIST on perceived stress in a general working population as a means of universal prevention, relative to a waitlist control group. A secondary study aim was to explore the resilience factors of self-efficacy, optimism, self-compassion, and perceived social support, the intervention targets as potential mediators of its effect on stress and self-perceived resilience.

Methods: In total, 352 employees were randomly assigned to either a self-help version of RESIST or a waitlist control group. Data were collected via the web at baseline, postintervention, and at 3- and 6-month (intervention group [IG] only) follow-ups. The primary outcome was perceived stress, measured with the Perceived Stress Scale-10. Secondary outcomes included self-perceived resilience, the resilience factors targeted, and other mental and work-related health outcomes.

Results: The IG reported significantly less stress than controls postintervention ($\Delta=-3.14$; $d=-0.54$, 95% CI −0.75 to −0.34, and $P<.001$) and at 3-month follow-up ($\Delta=-2.79$; $d=-0.47$, 95% CI −0.71 to −0.22, and $P=.002$). These improvements in the IG were maintained at 6-month follow-up. Favorable between-group differences also were detected for self-perceived resilience and the resilience factors. IG participants completed on average 2.2 (SD 2.3) web-based sessions and used the app's core feature a median of 14 times (IQR 4.00-33.75, range 1-220). The positive effects of the intervention on stress and resilience were primarily mediated by changes in optimism and self-compassion. No evidence was found that self-efficacy and social support also acted as mediators.

Conclusions: In a sample of employees experiencing heightened work-burden levels, RESIST was effective in reducing perceived stress and increasing self-perceived resilience as well as the targeted resilience factors. Mediation analyses suggested that developing a positive future outlook and a self-compassionate attitude toward oneself may be key drivers to enhance resilience. Changing the quality of social relationships and strengthening the belief in one's abilities may require more time, the involvement of others, or personal support from an eCoach to ensure sufficient learning opportunities.

Trial Registration: German Clinical Trials Register DRKS00017605; https://drks.de/search/de/trial/DRKS00017605

XSL•FO
RenderX

(J Med Internet Res 2026;28:e78335) doi:10.2196/78335

KEYWORDS

stress; resilience factor; resilience mechanism; resilience training; digital mental health intervention; internet-based intervention; mobile intervention; occupational eMental health; prevention; RCT; randomized controlled trial; mobile phone

Introduction

Background

Stressors are an integral part of life. By challenging an individual to adapt to new circumstances, they provide opportunities for growth and development. Most people can maintain or quickly regain good mental health despite experiencing stress—a phenomenon referred to as resilience [1-3]. However, when adaptation fails, stressors can contribute to the development of stress-related mental disorders [4]. Given that disorders such as depression and anxiety rank among the top 25 leading causes of global health-related burden [5,6], finding effective strategies to help individuals adapt to stress and prevent mental illness is of high importance. The promotion of individual resilience can, therefore, be an important endeavor to mitigate the potentially harmful effects of stress for individuals and, in turn, societies.

Resilience, however, is a complex phenomenon and the subject of extensive research. Resilience research has not yet reached consensus on how to define or conceptualize it. Although definitions vary, recent conceptual approaches, such as the Positive Appraisal Style Theory of Resilience (PASTOR) [2,3], converge on the idea that resilience reflects a dynamic, active, and multidimensional process of adaptation to stressors. Since this process is potentially modifiable, this opens avenues for resilience-promoting interventions.

Evidence for Resilience-Promoting Interventions

Published evidence on the effectiveness of both nondigital and digital resilience interventions, however, remains mixed, with highly heterogeneous effects observed between primary studies and inconsistent conclusions drawn in systematic reviews and meta-analyses [7-11]. Concerning digital interventions, one meta-analysis identified small-to-moderate effects on self-perceived resilience, but not on stress reduction [7], while another found no effect on self-perceived resilience [8]. Conversely, a recent meta-analysis by Schäfer et al [11] detected small, favorable effects both on perceived stress and self-perceived resilience. Favorable effects were also found for various resilience factors, including self-compassion, optimism, and social support [11]. Overall heterogeneity in the literature may stem from the general lack of consensus on what constitutes a resilience intervention [9,12], resulting in inconsistent inclusion and exclusion decisions between reviews.

Classification of Resilience Interventions

Overview

One approach has been to classify any intervention as a resilience intervention, regardless of more specific characteristics, as long as its primary objective is to modify resilience-related outcomes. These outcomes can include both self-perceived resilience, operationalized as the perceived ability to bounce back and recover from stress, or via related mental health outcomes (eg, stress and well-being).

An alternative approach is to define resilience interventions based on more specific intervention characteristics, including (1) the theory or rationale underlying the intervention; (2) the intervention's content; (3) the interventional techniques applied; and (4) the timing of the intervention relative to stressor exposure. These dimensions contribute to significant variability in the design of resilience interventions and represent a variety of unsolved conceptual challenges in interventional resilience research.

Underlying Theory

Concerning underlying theories and rationales behind interventions, many resilience interventions take a pragmatic, atheoretical approach, often without reference to any specific theoretical foundation. Conversely, some interventions are informed by established theories drawn from related domains, which are then applied to the context of resilience [12]; for example, the Transactional Model of Stress [13]. Notably rare, however, are interventions explicitly grounded in a specific theory of resilience or a genuine resilience framework. As introduced earlier, PASTOR [2,3] provides such a framework. It emphasizes that resilience as an outcome from a dynamic, active, and multicausal process of stressor adaptation is shaped by both internal and external resilience factors. The theory further highlights that these resilience factors might exert their protective effects through common cognitive appraisal processes, and that the core mechanism underlying resilience may be a positive appraisal style of stressors, that is, a general tendency to appraise stressors in a nonnegative, nonthreatening, but positive way.

Intervention Content

Further, an intervention may be considered a resilience intervention based on its content, regardless of the outcomes studied. The most common approach involves targeting intervention content toward building or strengthening certain resilience factors. These factors predominantly consist of a broad range of psychosocial resources, including self-efficacy, locus of control, problem-solving, optimism, cognitive flexibility, mindfulness, self-compassion, and perceived social support [7-9]. Such resources have been shown to be statistically associated with resiliency outcomes and are thought to facilitate adaptive responses to stress [2], thereby helping individuals to bounce back from adversity. Consistent with the idea that access to a broad repertoire of resources facilitates flexible responses to diverse challenging situations [14], resilience interventions should, nonetheless, aim to target multiple resilience factors rather than a single one. For instance, enhancing self-efficacy may have limited value in situations where an individual is unable to influence outcomes through their own actions. However, certain resilience factors are also commonly targeted

XSL•FO
RenderX

in other well-established intervention approaches, including problem-solving approaches [15] and mindfulness-based practices [16]. In contrast, if resilience interventions are categorized as a distinct class of intervention, it is essential to delineate the specific factors that should be targeted to meaningfully differentiate them from other existing approaches. In this regard, the resilience factors addressed within such interventions may be those that reflect more specific and overarching resilience processes and include, for example, factors that may serve as correlates or indicators of higher-order resilience mechanisms, such as a positive appraisal style [2].

The web- and app-based resilience training RESIST is an example of an intervention that is grounded in PASTOR [2], as the intervention's underlying theoretical framework. It aims to foster a set of specific resilience factors, namely self-efficacy, optimism, self-compassion, and social support, associated with a positive appraisal style as a higher-order resilience mechanism [17]. A pilot randomized controlled trial (RCT) has already demonstrated its effectiveness [17], providing preliminary evidence that an intervention based on these principles might be a promising approach to fostering self-perceived resilience.

Mechanisms of Change

Despite certain resilience factors being targeted within given interventions, it nonetheless remains largely unknown whether favorable intervention outcomes can be directly attributed to increases in these resilience factors. In such cases, the targeted resilience factors would function as mediators of the intervention, that is, the intervening variables that statistically account for the relationship between an intervention and its outcome [18]. They may therefore point to mechanisms of change through which the intervention achieves its effects [19]. Notwithstanding the growing number of resilience interventions, empirically based understanding about their mechanisms of change remains scarce [11]; with only a few published studies, often inadequately powered, having explored this [20,21]. Examining the mechanisms underlying resilience interventions could offer valuable insights into core processes promoting resilience. Such insights could help to identify the most appropriate selection of resilience factors for interventions to target, ultimately contributing to their optimization and increased effectiveness.

Interventional Techniques

Concerning the intervention techniques used, which determine how intervention content is delivered, there appears to be a disconnect between the psychosocial resources targeted by resilience interventions and the use of techniques explicitly designed to foster such resources (eg, keeping a diary with a focus on successful experiences to foster resilience by enhancing self-efficacy). Indeed, many resilience interventions adopt a problem-reducing approach [22], rooted in psychotherapeutic approaches via, for example, cognitive restructuring techniques. One contrasting example of an approach explicitly developed to foster individuals' protective resources to build resilience is Strengths-Based Cognitive Behavioral Therapy (Strengths-Based CBT) by Padesky and Mooney [23]. Among the first of its kind, the RESIST intervention applied this approach as the intervention technique used it to target the resilience factors as intervention content [17]. Since RESIST targets employees experiencing workplace stressors, it can be classified, in terms of the timing of the intervention relative to stressor exposure (4), as an intervention for use during stressor exposure [12].

Aims of This Study

Building on the findings of the pilot study assessing RESIST in an indicated preventive setting [17], we conducted a sufficiently powered study in the general working population to compare the effectiveness of RESIST against a waitlist control group (WL) in reducing subjective stress levels, which served as the primary outcome. We conducted further exploratory analyses to examine its effects on various secondary outcomes, including self-perceived resilience and the resilience factors targeted within the intervention. In light of the urgent need to address the gap in knowledge regarding resilience interventions' mechanisms of change, we also examined, as a secondary aim, the mediating effects of the resilience factors emphasized in RESIST on stress and self-perceived resilience.

Methods

Study Design

This study was conducted as a two-arm RCT from November 2019 to August 2021 in Germany. To investigate the intervention effects on the primary outcome of perceived stress, participants were randomly assigned either to the intervention group (IG) with immediate access to the resilience intervention RESIST in the form of a self-help format or to a WL whose members were offered the resilience intervention after 3 months of waiting time.

The study is reported in compliance with the CONSORT-eHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) guidelines for improving and standardizing the reporting of web-based and mobile health interventions [24] and following the recommendations of Chmitorz et al [12] for reporting trials on resilience interventions. The intervention itself is reported in accordance with the TIDieR (Template for Intervention Description and Replication) checklist [25]; an overview of the checklist, along with information on where to find the corresponding details in the main text, is provided in Multimedia Appendix 1.

Ethical Considerations

This study received ethical approval from the Ethics Committee of Leuphana University of Lüneburg. Before initiating subject enrollment, this study was registered at the German Clinical Trials Registry (DRKS00017605), a primary registry of the World Health Organization. All participants provided written informed consent before taking part in this study. Participants were informed that they could withdraw their consent at any time and request the deletion of their data without experiencing any negative consequences.

All data were deidentified before analysis; personal identifiers were replaced with unique codes, and the coding key was stored securely and separately. Access to identifiable data was restricted to authorized members of the research team.

XSL•FO
RenderX

Participation in this study was entirely voluntary and not financially compensated.

Participant Recruitment

This study was conducted in a universal preventive setting, addressing the general working population. Participants were recruited via (1) articles on resilience containing a link to the study's landing page in two well-known popular German science magazines ("Spektrum der Wissenschaft" and "Spektrum der Psychologie"); (2) newsletters from several statutory German health insurance companies (eg, Daimler BKK) that included a section referring to this study; and (3) social media posts on various platforms (Xing [New Work SE], Instagram [Instagram from Meta], and Facebook [Meta]).

Individuals who entered their email address in a form on this study's landing page or contacted us directly via email received detailed information about this study and were provided access to an online screening questionnaire via email. Inclusion criteria for study participation were (1) being aged at least 18 years, (2) having employment (part-time or full-time), (3) having steady access to the internet, and (4) possessing a smartphone. Exclusion criteria were (1) having an earlier or current diagnosis of a severe psychological disorder (eg, psychosis); (2) undergoing current psychotherapy, including being on a waiting list for or having the intention to receive psychotherapy; (3) current participation in another resilience intervention or stress management program; (4) changes in current medication for a stress disorder, anxiety disorder, or depression within the past four weeks; and (5) current suicidal ideation (agreeing with either the statement "I would like to kill myself" or "I would kill myself if I had the chance" on the related item in the Beck Depression Inventory-II [26]).

Individuals eligible for study participation received a link to the baseline assessment. After completing the baseline assessment and giving informed consent for study participation, participants were randomly assigned to one of the two study groups using a computer-generated randomization list with a ratio of 2:2 and a block size of four. The list was generated by a member of the department of the first and last author, who was not otherwise involved in this study. Randomization was conducted anonymously, without any personal contact between study personnel and participants. Blinding of participants was infeasible because participants were aware of whether they were randomized to the IG or the WL. Participants were informed about their group allocation via email and, depending on the randomization outcome, either were granted immediate access to the web- and app-based intervention (IG) or were promised access after 3 months of waiting time (WL).

Intervention

The feasibility and preliminary effectiveness of the web- and app-based resilience intervention RESIST were positively evaluated in a pilot RCT [17]. As a genuine resilience framework, PASTOR was considered the underlying theoretical foundation of the intervention [2,17]. Overall, the intervention aims to promote a positive appraisal style of stressors and stressful situations to help users adjust to work-related stress. Thus, this study design followed an approach focused on delivering the intervention during ongoing stressor exposure [12]. According to PASTOR, both a positive appraisal style as a resilience mechanism and successful adaptation to stress are facilitated by several resilience factors. Thus, promoting a set of specific resilience factors was at the intervention's core. However, PASTOR is a conceptual framework for this study of resilience, aiming for comprehensiveness in explaining the mechanisms underlying resilient adaptation. Consequently, it does not provide specific guidelines on which factors should be strengthened or how this should be achieved. Thus, the selection of the resilience factors targeted in the training was guided by several considerations: (1) they should be etiologically relevant for fostering adaptive responses to adversity; (2) they should potentially be competencies fundamental to resilience; (3) they should encompass different facets of resilience, including behavioral (eg, self-efficacy), cognitive (eg, optimism), emotional (eg, self-compassion), and interpersonal (eg, social support) dimensions; (4) they should reflect varying temporal orientations (eg, self-efficacy grounded in past achievements and optimism oriented toward the future); (5) they should be distinct from factors commonly targeted in other well-established intervention approaches (eg, mindfulness or active coping, which are typically addressed in the context of mindfulness and stress-management training); and (6) the number of selected factors should remain manageable to enable the design of medium-length interventions, which have been shown to be most effective [27]. Based on these considerations, self-efficacy, optimism, self-compassion, and social support [28-31] were selected as resilience factors targeted in the training.

RESIST was designed as a hybrid intervention to be delivered individually to participants, combining a web-based intervention and a mobile app component. The intervention's format allows participants to complete it anywhere they prefer. The web-based component accessible via a digital platform [32], consisting of longer intervention sessions, is designed for completion on a one-per-week basis. Once one session is completed, the next session is unlocked automatically. The app component that can be downloaded from the Google Play Store and Apple App Store (no longer available) consists of daily exercises. The two components are intentionally interrelated, with daily app exercises informing web-based sessions and web sessions guiding app-based practice. The intervention period for each study participant was 8 weeks. A detailed overview of the interventions' content and exercises can be found in Multimedia Appendix 2. Strengths-Based CBT [23] is applied as the interventional technique addressing *how* the resilience factors are targeted and is comprised of four steps: (1) searching for strengths; (2) building a personal model of resilience (PMRe), that is, a positive self-concept including positive imagery or metaphors; (3) applying the PMRe; and (4) practicing the PMRe.

Using the app component was step 1 (search for strengths) of the Strengths-Based CBT approach (Multimedia Appendix 2). As the app's central feature, users can document experiences of resilience, so-called moments of resilience, throughout the day (see Figure S2A in Multimedia Appendix 3). These moments capture instances when users overcome minor or major obstacles or experience positive events despite ongoing stressors

XSL•FO
RenderX

or challenging life circumstances. During a daily review, these moments are categorized into one of the four resilience factors of self-efficacy, optimism, self-compassion, and social support, based on the user's descriptions (see Figure S2D in Multimedia Appendix 3). Details on other app components are provided in Multimedia Appendix 3.

The web-based component has six sessions. The first session introduces the resilience intervention. Each of the second through fifth sessions addresses step 2 (construct PMRe), step 3, and step 4 (apply+practice PMRe) of the Strengths-Based CBT approach (Multimedia Appendix 2), which are incorporated as follows. The moments of resilience collected with the app are displayed in the web component. Based on step 2 of the Strengths-Based CBT approach, an exercise named resilience self-image was designed to further work with the moments of resilience. Within this exercise, users can select one of the displayed moments of resilience and develop a positive imagery or metaphor describing how they felt in the respective moment (eg, "I felt like a tribal leader, aware of my position and bravely taking responsibility"). This was intended to help participants develop a positive and resilient self-concept. A third exercise, named the resilience project, was developed based on steps 3 and 4 (apply+practice PMRe) of the Strengths-Based CBT approach. Within this exercise, users can plan how to approach and overcome an upcoming stressful situation by making use of their resilience self-image and resilience factors. The sixth session recaps the content of all the previous sessions.

Besides the exercises constructed on the Strengths-Based CBT, sessions include educational elements via text or video, as well as further written or audio exercises on the four resilience factors, each targeted in a single session. The sessions further feature two personas, which represent fictional users who are also completing the resilience training, thereby providing examples of how to complete the exercises. Example screenshots of the web component are available in Multimedia Appendix 3. Each session consists of 8‑10 web pages and requires 45‑60 minutes to complete. Users may complete the intervention at their own pace; however, it is recommended that they finish one session per week. No external reminders were sent to participants to complete the web-based component; however, in-app reminders could be set within the app component. The intervention was not tailored or modified during the course of this study. Access to the intervention can be provided to readers upon request.

Measures

Overview

Web-based data assessment via patient-reported outcome measures took place before randomization for screening purposes (T0), at baseline (T1), immediately postintervention, which was eight weeks after randomization (postintervention assessment [T2]), and three months after randomization (3 mo follow-up [T3]). We further included a 6-month follow-up (T4) assessment for the IG only, guided by existing evidence indicating that between-group intervention effects of similar programs remain largely stable over this period [10,33]. Thus, for ethical reasons, the control group did not continue without intervention beyond the initial 3 months.

Primary Outcome Measure

The primary outcome of this study was subjective stress referring to the last 7 days, measured using the Perceived Stress Scale-10 (PSS-10) [34,35], a well-established self-report measurement that assesses the degree to which people appraise situations in life as stressful—that is, unpredictable, uncontrollable, or overloaded. It consists of 10 items (eg, "In the last month, how often have you felt nervous and 'stressed'?") with each item rated on a scale ranging from 0 to 4. The total score ranges from 0 to 40, with higher scores indicating higher levels of perceived stress. In the present sample, the scale demonstrated high internal consistency, with an α=.89 at baseline.

Secondary Outcome Measures

Overview

Secondary explorative outcomes included resilience-, mental health-, work-, and intervention-related outcomes. Further details and the reliability of all outcome measures are listed in Multimedia Appendix 4.

Resilience and Resilience Factors

Self-perceived resilience—the perceived ability to recover from stress—was measured using the Brief Resilience Scale [36]. The resilience factor self-efficacy was assessed using the General Self-Efficacy Short Scale-3 [37]. As related constructs, internal and external control beliefs were estimated using the Short Scale for the Assessment of Locus of Control [37] for explorative purposes. The resilience factor, optimism, was assessed using the revised version of the Life Orientation Test [38]. The resilience factor, self-compassion, was assessed using the Self-Compassion Scale Short Form [39]. The resilience factor, social support, was measured with the perceived available social support subscale and the support-seeking subscale of the Berlin Social Support Scales [40].

Mental Health Outcomes

Symptoms of depression were measured using the Epidemiologic Studies Depression Scale [41]. Values ≥18 indicate clinically significant levels of depressive symptoms [42]. Psychological distress, including anxiety and psychosomatic symptoms, was measured using the short version of the Brief Symptom Inventory [43].

Consistent with the notion that resilience per se is defined as good mental health despite adversity [44], exposure to different types of stressors was assessed. Critical life events (eg, death of a family member) were measured using an adapted version of the Life History Calendar [45]. Daily hassles were assessed using the Mainz Inventory of Microstressors [46].

Work-Related Outcomes

Adverse circumstances at work were assessed using the short version of the Effort-Reward-Imbalance Questionnaire [47,48], which contains the subscales effort, reward, and overcommitment. For the purpose of describing this study's sample baseline characteristics, the ratio of effort (numerator) over reward (denominator) at T1 was calculated to capture the imbalance between costs and gains experienced at work [49].

XSL•FO
RenderX

According to Lehr et al [49], a ratio >0.715 indicates an adversarial workplace environment.

Concerning productivity in the workplace, days of absenteeism and presenteeism were measured using the respective items of the Trimbos and Institute of Medical Technology Assessment Cost Questionnaire for Psychiatry [50]. Ability to work effectively was measured using the single-item Work Ability Index [51], assessing "current work ability compared with life-time best."

Due to data management errors, data on the secondary outcomes, psychological distress, effort, and rewards are missing for T2 and T4, while data for life events and daily hassles are missing from T2 to T4. These missing data represent minor deviations from this study's protocol.

Intervention Usage and Client Satisfaction

Intervention usage was assessed in two ways: (1) by the number of web-based sessions completed per IG participant (extracted from the online platform); and (2) by the collected number of moments of resilience, the core feature of the app, per participant (as tracked by the native application). Satisfaction with the web- and app-based intervention was assessed postintervention (only for IG) using an adapted version of the Client Satisfaction Questionnaire adjusted to the online format of the intervention [52]. Participants' perception of the app's attractiveness and quality was assessed using the short version of the AttrakDiff [53].

To characterize this study's sample, data on demographic variables such as age, gender, marital status, educational level, employment status, income, and the previous use of health services were collected as part of the screening procedure.

Statistical Analysis

Overview

In accordance with CONSORT-eHEALTH guidelines [24], data were analyzed following an intention-to-treat procedure. Specific details of the analysis are provided below. All statistical analysis was performed using R Studio (R version 4.3.2) [54] with a two-tailed significance level set at $P \leq .05$.

Power Calculation

This study's sample size calculation was based on identifying a realistic and practically meaningful effect size. First, regarding a realistically achievable effect, meta-analytic evidence on unguided web-based interventions for stress reduction was considered [27]. In the meta-analysis by Heber et al [27], an effect size of $d=-0.33$ was reported for unguided stress interventions. Second, by discussion, the research team agreed on 2.5 points as the minimal practically important difference between the IG and WL immediately after the intervention. Assuming an SD of 6.42 points on the PSS-10, as observed in a representative German sample [34], this resulted in an anticipated intervention effect of $d=-0.4$ for the primary outcome in this study. To detect an effect of that size with 80% power and 95% significance, an a priori power-analysis calculation yielded 352 participants required for this study.

Missing Data

All participants provided baseline data. Missing data at T2-T4 were imputed through multivariate imputation by chained equations using the *mice* package [55]. Twenty imputation sets were generated. Variables were included as predictors in the imputation model if they showed a correlation of at least 0.4 with the variable to be imputed and had at least 60% available data, ensuring that only predictors carrying a substantial amount of meaningful information were used [56]. The imputed datasets were analyzed separately, and the parameter estimates and hypothesis tests were ultimately pooled using the rule by Rubin [57].

Intervention Effect

Between-group differences at T2 and T3 were analyzed using analysis of covariance (ANCOVA), as recommended by O'Connell et al [58], to evaluate the interventions' effectiveness in influencing the primary and secondary outcomes. Baseline scores of the respective outcomes were considered covariates. Following recommendations by Harrer et al [56], between-group Cohen *d* and corresponding CIs were derived directly from the covariate-adjusted models used to calculate the magnitude of treatment effect, incorporating the pooled SD (based on Cohen formula) into the models.

Sensitivity analyses for the primary outcome were performed to test the robustness of the intention-to-treat analyses in multiple ways. First, study completers (participants providing data at all measurement points) were analyzed. Second, intervention completers were analyzed, defined as participants having completed at least the first five sessions of the web-based component (incorporating all exercises on the targeted four resilience factors) and collected moments of resilience with the app at least twice per week (≥16) within the training period. Third, the intervention effect was measured using linear mixed models.

Response and Deterioration Rates

Concerning better communicating the results to help-seekers and policy makers, we followed the recommendation by Cuijpers [59] and calculated response and deterioration rates. For all response rates, we calculated the numbers needed to treat to achieve one additional beneficial outcome (number needed to treat for benefit; NNTB) and one additional harmful outcome (numbers needed to treat for harm; NNTH) in the IG, relative to the WL.

We analyzed response rates from various perspectives. First, reliable improvement and deterioration were calculated based on Heber et al [33], who reported a change of ±5.16 points in the PSS-10 following the recommendations of Jacobson and Truax [60]. Second, using anchor-based criteria was advised for determining what constitutes a practically meaningful change [61]. Bauer-Staeb et al [62] highlighted that the magnitude of a practically meaningful change depends on the baseline level of distress and provided reference criteria. Given that the sample was expected to be primarily mildly to moderately distressed, a change of 20% from baseline to postintervention was considered a minimally practical meaningful difference. Third, remission was determined using the preferred method described

XSL•FO
RenderX

by Jacobson and Truax [60] for operationalizing clinically significant change—defined as having a score closer to the mean of the functional than dysfunctional population. The mean for a representative German sample (mean 12.57) served as an estimate for the functional population [34], while the mean score of a highly stressed, help-seeking sample (mean 22.65) was used to define the dysfunctional population [63]. Consequently, we defined remission as scoring 17 points or lower on the PSS-10, excluding participants already considered to have no significant symptoms by this definition at baseline.

Longer-Term Effects

An extended T4 assessment evaluated longer-term interventional effects among participants in the IG. Those effects were analyzed using within-subject comparisons via repeated-measures ANOVA between T1 and T4. Within-group Cohen *d* and respective CIs for the IG were calculated by dividing the change score (T1-T4 difference) by the SD of the change.

Mediation Analyses

To investigate the targeted resilience factors self-efficacy, optimism, perceived support, and self-compassion as potential mediators of the intervention's effect on stress and resilience, multiple single (including one mediator) and parallel mediation analyses (including all mediators at once) were conducted using the *semTools* package [64]. Single mediation analyses were conducted to assess the unique effect of each mediator in isolation, while parallel models provided a more comprehensive view of how multiple mediators operate concurrently and allowed for comparison of their relative contributions. To establish temporal precedence, T2 scores of the potential mediators and T3 scores of stress and resilience were used [18]. Mediator and baseline scores for stress and self-perceived resilience were included as covariates in the model, as recommended by Hayes and Rockwood [65]. A statistically significant effect of the mediation was achieved if the estimated 95% CI of the indirect effect did not cross zero. We conducted all mediation analyses with the study completer sample as additional sensitivity analyses.

Results

Participants and Baseline Characteristics

Figure 1 illustrates the flow of participants. A total of 352 individuals were randomized and assigned to either the IG or WL. Two participants initially allocated to the WL were subsequently excluded from this study due to concurrent involvement in another parallel research project. As a result, the final participant count was 176 participants for the IG and 174 participants for the WL.

XSL•FO
RenderX

Figure 1. Study flow of participants. IG: intervention group; PSS: Perceived Stress Scale; WL: waitlist control group.

XSL•FO
RenderX

Table 1 summarizes the study samples' baseline characteristics. On average, participants were aged 42.8 years; 65.7% (230/350) were female; and 74.9% (262/350) worked full-time. At screening, participants reported an average perceived stress level of 21.58 (SD 6.03). The average depression level among participants at baseline was 11.09 (SD 6.42). A total of 15.7% (55/350) of the participants reported clinically relevant levels of depressive symptoms. Participants reported an average sum of 68.45 (SD 32.72) daily hassles in the preceding 7 days (eg, a conflict or disagreement with a close person), along with 9.24 (SD 5.08) experienced life events (eg, serious illness of self or family member). Almost nine in ten (309/350, 88.3%) reported working under adversarial workplace conditions (Effort Reward Imbalance Scale ratio >0.715) at baseline. Participants' work ability—compared to life-time best indicated as 10 points—was reduced (mean 6.83, SD 1.86).

Table . Baseline characteristics of this study's sample.

	Total	IG[a] (n=176)	WL[b] (n=174)
Sociodemographics			
Age (y), mean (SD)	42.8 (11.2)	41.9 (11.2)	43.8 (11.1)
Female, n (%)	230 (65.7)	123 (69.9)	107 (61.5)
Single, n (%)	116 (33.1)	67 (38.1)	49 (28.2)
Married, n (%)	169 (48.3)	77 (43.8)	92 (52.9)
Relationship, n (%)	40 (11.4)	18 (10.2)	22 (12.6)
Divorced, n (%)	25 (7.1)	14 (8)	11 (6.3)
Education, n (%)			
No university degree	81 (23.1)	41 (23.3)	40 (23)
University degree	269 (76.9)	135 (76.7)	134 (77)
Employment, n (%)			
Full-time	262 (74.9)	143 (81.3)	119 (68.4)
Part-time	88 (25.1)	33 (18.8)	55 (31.6)
Prior experience with a mental health intervention, n (%)			
No	310 (88.6)	152 (86.4)	158 (90.8)
Yes	40 (11.4)	24 (13.6)	16 (9.2)
Experience with psychotherapy, n (%)			
No, never	243 (69.4)	123 (69.9)	120 (69)
Yes, in the past	103 (29.4)	52 (29.6)	51 (29.3)
Perceived stress levels, mean (SD)			
T0 PSS-10[c]	21.6 (6.03)	21.8 (5.7)	21.4 (6.4)
Clinically significant levels of depressive symptoms[d], n (%)			
Yes	55 (15.7)	22 (12.5)	33 (19)
No	295 (84.3)	154 (87.5)	141 (81)
Stressor exposure before intervention			
Adverse workplace situation (ERI-S[e]>0.715), n (%)	309 (88.3)	153 (86.9)	156 (89.7)
Macrostressors[f], mean (SD)	9.24 (5.1)	9.1 (4.9)	9.4 (5.2)
Microstressors[g], mean (SD)	68.5 (32.7)	68.1 (32.2)	68.8 (33.3)

[a]IG: intervention group.
[b]WL: waitlist control group.
[c]PSS-10: Perceived Stress Scale 10.
[d]Scores ≥18 on the Epidemiologic Studies Depression Scale (short form).
[e]ERI-S: Effort-Reward-Imbalance Scale (short form).
[f]Life events (counts).
[g]Daily hassles (counts).

XSL•FO
RenderX

Primary Outcome Measure

Table 2 lists the means and SDs of the primary and all secondary outcomes at all assessment points. At T2, individuals in the IG reported significantly lower levels of perceived stress than individuals in the WL, as demonstrated by ANCOVA ($F_{1,518}$=21.31, P<.001, d=−0.54, 95% CI −0.75 to −0.34, between-group difference: −3.14 points, see Table 3). The between-group effect remained significant and comparable in size at 3-month follow-up ($F_{1,79}$=10.38, P=.002, d=−0.47, 95% CI −0.71 to −0.22, between-group difference: −2.79, see Table 3).

XSL•FO
RenderX

Table . Means and SDs of the intention-to-treat approach's sample (intervention group=176; waitlist control group=174).

Outcome	T1[a]		T2[b]		T3[c]		T4[d]
	IG[e]	WL[f]	IG	WL	IG	WL	IG
Primary outcome, mean (SD)							
Stress	21.35 (6.21)	20.52 (6.71)	16.22 (6.47)	19.36 (6.63)	16.63 (7.33)	19.42 (6.92)	16.11 (6.86)
Secondary outcomes, mean (SD)							
Self-perceived resilience	16.80 (4.70)	16.61 (4.44)	18.84 (3.40)	17.19 (3.31)	19.34 (4.88)	17.86 (4.36)	19.67 (4.48)
Resilience factors, mean (SD)							
Self-efficacy	10.52 (2.13)	10.30 (2.23)	11.25 (2.22)	10.85 (2.02)	11.20 (2.25)	10.62 (2.01)	11.56 (1.98)
Internal control	3.58 (0.76)	3.50 (0.75)	3.81 (0.69)	3.56 (0.75)	3.70 (0.83)	3.46 (0.77)	3.77 (0.79)
External control	2.72 (0.06)	2.65 (0.80)	2.40 (0.80)	2.63 (0.77)	2.33 (0.96)	2.61 (1.05)	2.32 (0.70)
Optimism	14.06 (4.63)	13.90 (4.48)	15.59 (4.62)	14.46 (4.44)	15.89 (4.86)	14.35 (4.72)	16.11 (4.37)
Self-compassion	2.75 (0.72)	2.76 (0.68)	3.27 (0.71)	2.87 (0.69)	3.19 (0.73)	2.87 (0.69)	3.25 (0.74)
Perceived support	26.90 (4.83)	26.16 (5.20)	27.92 (4.38)	26.96 (4.89)	27.63 (4.71)	26.68 (4.90)	27.99 (4.70)
Support seeking	13.48 (3.84)	13.40 (3.76)	14.55 (3.60)	13.62 (3.54)	14.58 (3.70)	13.39 (3.47)	14.75 (3.42)
Mental health, mean (SD)							
Depressive symptoms	10.85 (6.19)	11.33 (6.64)	9.61 (6.16)	11.41 (6.41)	9.61 (6.16)	11.12 (6.61)	9.68 (6.06)
Psychological distress	0.83 (0.53)	0.82 (0.55)	—[g]	—	0.68 (0.52)	0.74 (0.51)	—
Work-related health, mean (SD)							
Work ability	6.80 (1.99)	6.86 (1.73)	7.26 (1.98)	6.75 (1.90)	6.73 (2.27)	6.84 (1.85)	7.17 (2.15)
Effort	8.61 (2.30)	8.83 (2.11)	—	—	7.82 (2.47)	8.59 (2.35)	—
Reward	18.79 (3.75)	18.02 (3.93)	—	—	18.74 (3.92)	17.96 (4.04)	—
Over-commitment	15.99 (3.73)	16.06 (3.44)	14.74 (3.68)	15.61 (3.69)	14.33 (3.55)	15.69 (3.77)	14.28 (3.78)
Absenteeism	7.01 (7.75)	8.05 (11.35)	5.80 (7.88)	8.20 (12.44)	3.69 (4.48)	4.63 (5.75)	7.34 (9.29)
Presenteeism	6.34 (4.90)	7.32 (9.05)	6.53 (6.30)	6.67 (5.67)	6.66 (6.62)	5.26 (4.17)	4.08 (2.89)

[a]T1: baseline.

[b]T2: postintervention (8 weeks after randomization).

[c]T3: 3-month follow-up (3 months after randomization).

[d]T4: 6-month follow-up (6 months after randomization, intervention group only).

[e]IG: intervention group.

[f]WL: waitlist control group.

[g]Missing due to data management errors.

XSL•FO
RenderX

Table . Between-group differences postintervention and at 3-month follow-up for primary and resilience-related secondary outcome measures.

Outcome		Differences between the intervention group and the waitlist control group					
		T2[a]			T3[b]		
		F test (*df*)	*P* value	Cohen *d* (95% CI)	*F* test (*df*)	*P* value	Cohen *d* (95% CI)
Primary outcome							
	Stress	21.31 (1, 518)	<.001	−0.54 (−0.75 to −0.34)	10.38 (1, 79)	.002	−0.47 (−0.71 to −0.22)
Secondary outcomes							
	Self-perceived resilience	30.46 (1, 253)	<.001	0.47 (0.29 to 0.64)	9.33 (1, 72)	.003	0.29 (0.08 to 0.49)
	Resilience factors						
	Self-efficacy	4.15 (1, 248)	.04	0.12 (0.06 to 0.30)	4.87 (1, 69)	.03	0.21 (−0.03 to 0.45)
	Internal control	11.23 (1, 191)	<.001	0.27 (0.06 to 0.47)	7.82 (1, 181)	.006	0.24 (0.03 to 0.45)
	External control	6.82 (1, 132)	.01	−0.35 (−0.57 to −0.12)	4.76 (1, 81)	.03	−0.33 (−0.57 to −0.08)
	Optimism	6.35 (1, 49)	.02	0.22 (0.16 to 0.60)	12.38 (1, 89)	.001	0.30 (0.11 to 0.48)
	Self-compassion	36.23 (1, 106)	<.001	0.57 (0.39 to 0.76)	11.36 (1, 46)	.002	0.42 (0.18 to 0.66)
	Perceived support	5.09 (1, 74)	.03	0.09 (−0.09 to 0.27)	4.18 (1, 78)	.04	0.09 (−0.10 to 0.27)
	Support seeking	11.92 (1, 380)	.001	0.25 (−0.09 to 0.40)	12.89 (1, 92)	.001	0.32 (0.14 to 0.49)

[a]T2: postintervention (8 weeks after randomization).

[b]T3: 3-month follow-up (3 months after randomization).

Sensitivity Analyses

Sensitivity analysis based on study completers (T2: d=−0.64, 95% CI −0.89 to −0.39; T3: d=−0.55, 95% CI −0.82 to −0.28) corroborated the results obtained by intention-to-treat analysis. The group difference between intervention completers and the WL at T2, however, was only close to being statistically significant (P=.05, d=−0.42, 95% CI −0.79 to −0.06). In contrast, the group difference at T3 was significant and similar in size to that obtained by the intention-to-treat approach (d=−0.46, 95% CI −0.84 to −0.09). A linear mixed-effects model revealed results similar to those obtained by ANCOVA using imputed data. Equivalently, the interactions between time and intervention at T2 (P<.001, SE=0.75) and T3 (P<.001, SE=0.80) were significant, indicating that the change in stress differed between groups. Specifically, participants in the IG showed a significantly greater reduction in stress from T1 to T2 (β=−4.02, 95% CI −5.48 to −2.56), as well as from T1 to T3 (β=−3.67, 95% CI −5.24 to −2.11), compared to those on the WL.

Response and Deterioration Rates

At T2 and T3, significantly more participants in the IG than WL controls reported reliable and practically meaningful improvement, with NNTB ranging from 3.42 (95% CI 2.56 to 5.16) to 4.35 (95% CI 3.05 to 7.58). The rates of reliable and practically meaningful deterioration were higher in controls than in the IG at both postintervention points, with NNTH values ranging from 7.51 (95% CI 4.80 to 17.21) to 10.81 (95% CI 6.57 to 30.57). However, 4.6% (8/174) to 10.9% (19/174) of participants in the IG also experienced a worsening of stress symptoms. Multimedia Appendix 5 summarizes the results of the response and deterioration analysis.

Secondary Outcome Measures

Table 3 summarizes the results of resilience-related secondary outcome analysis. For self-perceived resilience, there was a significant effect in favor of active treatment at T2 (d=0.47) and T3 (d=0.29). For the resilience factors, there were significantly better outcomes among participants in the IG for all outcomes at T2 and T3. Immediately postintervention, effect sizes were 0.12 for self-efficacy, 0.27 for internal control, −0.35 for external control, 0.22 for optimism, 0.57 for self-compassion, 0.09 for perceived social support, and 0.25 for support seeking.

Multimedia Appendix 6 summarizes the findings of the other mental health- and work-related outcomes assessed. For depression, there was a significant between-group effect at T2 (d=−0.24), but not at T3. Regarding work-related outcomes, significant effects in favor of the IG were found for work ability at T2 (d=0.28), and for overcommitment at T2 (d=−0.22) and T3 (d=−0.36), but not for absenteeism or presenteeism at either of these postintervention measurement points.

XSL•FO
RenderX

Longer-Term Effects

Concerning the intervention's long-term effectiveness (T1-T4), outcomes remained stable, indicating that the beneficial effects were maintained. Within-group effect sizes for the primary and secondary outcomes were all significant, except for absenteeism. The effect size for stress was d=−0.80 (95% CI −0.97 to −0.63), for self-perceived resilience d=1.19, for self-efficacy d=0.80, for optimism d=0.87, for self-compassion d=1.25, and for perceived social support d=0.37. A table summarizing the results of the within-group comparisons obtained by repeated-measures ANOVA is available in Multimedia Appendix 7.

Mediation Analyses

Perceived Stress

Single mediation analysis revealed that indirect effects through optimism (ab=−0.46, 95% CI −0.80 to −0.12) and self-compassion (ab=−0.95, 95% CI −1.49 to −.41) were statistically significant, indicating that these variables mediated the effect of the intervention on perceived stress at T3. In contrast, indirect effects through self-efficacy (ab=−0.11; 95% CI −0.34 to 0.12) and perceived support (ab=−0.04; 95% CI −0.18 to 0.10) were not significant.

As depicted in Figure 2, the results of parallel mediation analysis aligned with the results obtained with single mediation analyses. In a joint model, the indirect effects through optimism (a_2b_2=−0.34, 95% CI −0.63 to −0.06) and self-compassion (a_4b_4=−0.66, 95% CI −1.15 to −0.17) were again statistically significant, while the indirect effects through self-efficacy (a_1b_1=−0.02, 95% CI −.14 to 0.10) and perceived support (a_3b_3=0.04, 95% CI −0.09 to 0.16) were not. The direct effect of the intervention on perceived stress remained significant after incorporating the mediators in the model (c'=−2.25, 95% CI −3.79 to −0.70).

Figure 2. Parallel multiple mediation model with 3-month follow-up (T3) stress severity scores as the outcome variable (Y), posttreatment (T2) resilience factors scores as mediators, and baseline values of mediators and outcome as covariates. Intervention (X) is coded 1=IG, 0=WL. Unstandardized beta coefficients are shown with 95% CIs in brackets. IG: intervention group; WL: waitlist control group.

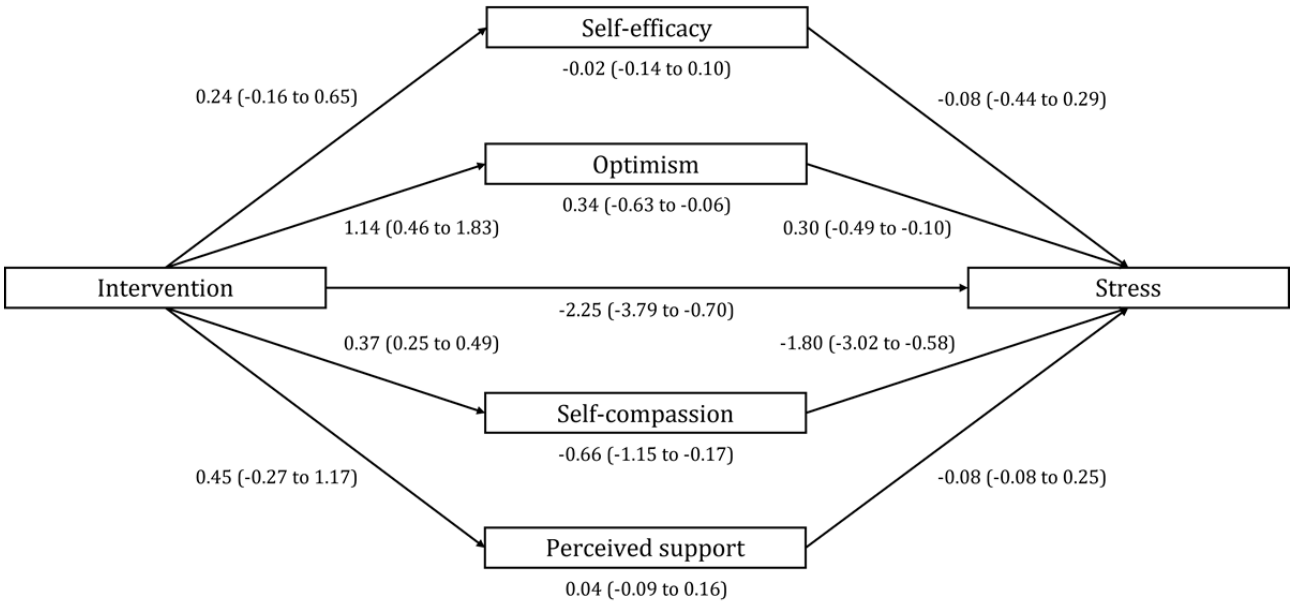

Self-Perceived Resilience

Mediation analyses with self-perceived resilience—that is, the ability to recover from stress—as a dependent outcome revealed a pattern similar to that observed with perceived stress as the outcome. Single mediation analyses showed that the indirect effects through optimism (ab=0.27, 95% CI 0.07 to 0.46) and self-compassion (ab=0.46, 95% CI 0.16 to 0.76) were statistically significant, indicating that these variables mediated the intervention's effect on self-perceived resilience at T3. In contrast, indirect effects through self-efficacy (ab=0.08; 95% CI −0.06 to 0.22) and perceived support (ab=0.01; 95% CI −0.07 to 0.08) were not.

On parallel mediation analysis, the indirect effect through optimism (a_2b_2=0.21, 95% CI 0.06 to 0.37) significantly mediated the intervention's effect on resilience at T3, while the indirect effect through self-compassion (a_4b_4=0.26, 95% CI −0.01 to 0.53), though close, failed to achieve statistical significance. The indirect effects achieved through self-efficacy (a_1b_1=0.04, 95% CI −0.03 to 0.11) and perceived support (a_3b_3=−0.03, 95% CI −0.10 to 0.04) were also not statistically significant. The direct effect of the intervention on self-perceived resilience remained significant after incorporating mediators into the model (c'=0.86, 95% CI 0.04 to 1.65). The corresponding mediation model is shown in Figure S1 in Multimedia Appendix 8.

Repeating all single and parallel mediation analyses with this study completer sample as additional sensitivity analyses revealed no differences in the results relative to those obtained with the intention-to-treat approach (see Figures S2 and S3 in Multimedia Appendix 8). The only exception was in the joint model with self-perceived resilience as the dependent variable, in which the indirect effect via self-compassion reached statistical significance (a_4b_4=0.43, 95% CI 0.13 to 0.73).

XSL•FO
RenderX

Intervention Usage and Client Satisfaction

Most participants started with the web component of the intervention (156/176, 88.6%), completing an average of 2.2 (SD 2.3) sessions. Completion rates ranged from 70.5% (110/156) for session 1 to 18.6% (29/156) for all 6 sessions.

Those who used the app component (104/176, 59.1%) collected a median of 14 moments of resilience (range 1 to 220), indicating that half collected these moments almost twice per week over the 8-week intervention period. Based upon participant reports, users collected a total of 3064 moments of resilience, most frequently for self-compassion. Detailed engagement patterns are summarized in Multimedia Appendix 9. Of those participants reporting their intervention usage at T2, 38.1% (32/84) indicated they had not yet finished and expressed an intention to continue.

Participants' satisfaction with the intervention (considering both the web and app components) was high (mean 24.02, SD 6.18). Of those participants providing questionnaire data about their satisfaction with the intervention, 87/99 (87.9%) indicated satisfaction in an "overall, general sense" ("very satisfied" or "mostly satisfied"), and 80/99 (80.5%) indicated satisfaction with the amount of help they received. In addition, 73/99 (73.7%) agreed that the intervention helped them to deal with problems more effectively, and 77/99 (77.8%) said they would recommend the intervention to a friend in need of similar help. The overall user experience of the app was rated as 4.71 (SD 1.04; range 1 to 7; n=99), indicating an overall good level of user experience [53].

Discussion

Principal Results

Overview

For the first time, this study identified a favorable effect of the RESIST intervention, relative to being on a waiting list, on perceived stress in a universal prevention setting. Improvements were evident immediately after the intervention and remained stable 3 and 6 months after randomization. A similar pattern was observed for self-perceived resilience and the resilience factors targeted within the intervention. Effects on other health- and work-related outcomes were mixed.

Concerning the second study aim, mediation analyses suggested that the resilience factors targeted within the training played distinct roles in how the intervention exerted its effects on stress and self-perceived resilience. The strongest evidence for mediation emerged for optimism and self-compassion, with significant indirect effects observed for both variables. In contrast, indirect effects through self-efficacy and social support were not statistically significant.

Effects on Stress and Self-Perceived Resilience

The observed effect on stress reduction immediately postintervention (d=−0.54) slightly exceeded the effect size expected from the a priori sample size calculation (d=−0.40). Building on the findings from the pilot study [17], this study confirmed the intervention's effectiveness in a larger sample within a universal preventive setting, demonstrating that the intervention is effective not only when delivered with guidance, as in the pilot study, but also in a self-help format that requires fewer resources.

When compared with interventions targeting a similar set of resilience factors that may be unique to resilience interventions [21,66], the effect sizes for stress and self-perceived resilience (d=0.47) observed in this study fall between those reported for a web-based resilience intervention for college students [66] (d=−0.34 for stress) and a multicomponent positive psychology intervention for the general population [21] (d=0.67 for resilience).

Concerning intervention techniques, this study's findings align with those of 2 previous studies that also used Strengths-Based CBT [67,68]. In one of these studies [68], distressed college students showed significant postintervention improvements in resilience (d=0.34).

Compared to meta-analytic findings, the stress reduction effect observed for the present intervention postassessment was substantially larger than the nonsignificant effect (g=0.14) reported by Ang et al [7], who used a broad resilience factor approach. In contrast, the findings related to self-perceived resilience closely align with those of Ang et al [7] (g=0.54). The effect of RESIST on self-perceived resilience is also similar in magnitude to that reported for the meta-analysis by Díaz-García et al [8], which adopted an outcome-focused resilience approach, including interventions aimed at modifying resilience or related constructs regardless of the interventions' characteristics. Furthermore, the stress reduction effect of RESIST was slightly larger than the effect size for stress reported in the meta-analysis by Schäfer et al [11] (standard mean difference=0.33, 95% CI −0.41 to −0.24), which applied a broad conceptualization of resilience interventions, also including established intervention formats such as stress-management and mindfulness programs. The effects for self-perceived resilience found in the present study also surpassed those reported by Schäfer et al [11] (standard mean difference=0.22).

Although these comparisons do not establish superiority of interventions that target a specific set of resilience factors that may reflect more overarching resilience processes and may be correlates of higher-order resilience mechanisms, such as a positive appraisal style, the findings suggest that it is valuable to continue developing and investigating interventions based on these principles. This is particularly important if resilience interventions are to be seen as a distinct form of intervention.

The comparisons further underscore the challenges of adequately evaluating and comparing resilience interventions, stemming from the lack of a clear understanding of what constitutes a resilience intervention. The variability in the design of resilience interventions makes it difficult to identify comparable interventions, especially when investigators fail to adequately describe the intervention's characteristics. As mentioned earlier, such conceptual ambiguity is reflected in meta-analyses, where differing definitions of resilience interventions can lead to substantially heterogeneous results and conclusions between meta-analyses. This emphasizes the need for a standardized classification system for resilience interventions. Building on

XSL•FO
RenderX

prior considerations by Chmitorz et al [12], (1) the theory or rationale underlying an intervention, (2) its specific content, (3) the intervention techniques applied, and (4) the timing of the intervention relative to stressor exposure should be considered as dimensions for building a systematic classification system of resilience interventions. This may contribute to identifying the core elements of effective intervention designs.

Response and Deterioration Rates

The NNTB suggests that, on average, three to four individuals must be given access to RESIST for a single individual to experience 20% symptom improvement or reliable improvement in stress symptoms immediately after the intervention. This aligns with the results of other digital mental health interventions such as self-guided stress management training [69]. Corresponding to findings from prior research [70,71], it must be noted, however, that some individuals (12/174, 6.9%) in our IG met the criterion for reliable deterioration. This highlights safety issues within resilience interventions. While resilience interventions labeled positively might seem more appealing than, for example, stress management interventions, this does not guarantee symptom improvement, as some users may experience adverse effects. Potential risks may include heightened frustration or self-blame if participants do not meet perceived expectations of "being resilient" [72]. The positively framed concept of "resilience" may furthermore create implicit pressure to bounce back, potentially reinforcing feelings of inadequacy when setbacks persist despite effort. In the systematic review by Chmitorz et al [12], none of the 43 RCTs included in the analysis assessed adverse effects. Hence, the potential harms of so-called "positive" interventions remain largely overlooked [71], underscoring the need for more conceptual and empirical research in this area.

Effects on Resilience Factors

Besides examining RESIST's effectiveness in reducing stress as a primary outcome and resilience as an important secondary outcome, we also investigated whether the intervention promoted the resilience factors of self-efficacy, optimism, self-compassion, and social support that the intervention was designed to target. The effects observed postintervention were favorable, yet variable in magnitude.

For self-efficacy, we observed a small but statistically significant between-group effect after the intervention (d=0.12). This is numerically similar to the effect reported by Schäfer et al [11] in their meta-analysis (d=0.01). Regarding optimism (ie, the general expectation that one's own outcomes will be positive [73]), our findings indicated a more sizeable intervention effect (d=0.22), in line with results from meta-analyses by Schäfer et al [11] and Malouff and Schutte [74]. The most pronounced effect at T2 was observed for self-compassion (d=0.57; ie, a supportive attitude toward oneself in times of suffering [75]). This finding mirrors results reported by Schäfer et al [11], where self-compassion also showed the largest effect among the resilience factors examined. For perceived social support, we found small between-group differences at T2 (d=0.09), which also aligns with published meta-analyses [31]. Despite the statistically significant findings, the practical significance of these effects is difficult to interpret due to the lack of established criteria for meaningful change in resilience factors. Future research should investigate such criteria, following the methodological recommendations of Cook et al [76].

Mediation Analyses

The mediation analyses further revealed that RESIST not only favorably fostered optimism and self-compassion, but that these factors also contributed to how the intervention exerted its positive effects. With one exception—self-compassion mediated the effect on self-perceived resilience only in this study's completer sample—the mediation results were consistent, irrespective of whether stress or self-perceived resilience was assessed as the outcome, underscoring that the same resilience factors may matter for both positive and negative mental health outcomes. Comparing these mediation findings to prior research results is limited by the scarcity of studies that have investigated the mechanisms of change underlying resilience interventions. Our mediation findings align with isolated prior findings of the previously mentioned multicomponent email-guided positive psychology intervention [21], for which both optimism and self-compassion also emerged as mediators of the intervention's effects.

The findings shed light on the importance of optimism and self-compassion as potential key processes in resilience promotion. In the case of optimism, enhancing the belief that one's own outcomes will be positive may both strengthen someone's confidence that a currently stressful situation can improve over time and foster the perception that such situations are manageable and temporary, thereby reducing perceived stress and strengthening self-perceived resilience. This, in turn, might reflect a positive appraisal style of stressors. This finding aligns with a substantial body of prior research, including both meta-analyses and longitudinal studies, that have revealed the importance of optimism for positive mental health outcomes [30,77-80]. For instance, in a meta-analysis published by Gallagher et al [30], a negative association was identified between optimism and posttraumatic stress disorder. Meanwhile, Romswinkel et al [80] identified optimism both as a predictor of reduced depressive symptoms and as a mediator in the relationship between job stress and depression.

Regarding self-compassion, fostering a kinder and warmer attitude toward oneself during times of stress might have enabled IG participants to better care for themselves and respond with reduced self-criticism, particularly when facing stressors rooted in personal mistakes or feelings of inadequacy [75]. By promoting a more accepting and supportive internal response, the intervention may have helped them to reinterpret such stressors more constructively, thereby establishing a link to a positive appraisal style and ultimately reducing perceived stress. Empirical evidence supports this strong link between self-compassion and mental health [81-85]. For example, one meta-analysis has demonstrated an inverse association between self-compassion and stress [82], while a longitudinal study by Lee et al [81] identified self-compassion as a significant predictor of mental well-being over five years.

In contrast to self-compassion and optimism, no significant mediation effects were observed for self-efficacy and social support. However, it is premature to discount the roles of

XSL•FO
RenderX

self-efficacy and social support in interventions aiming to promote resilience. Differences in the control and ease of implementation, as well as the timescale of observable change, may partly explain the pattern of effects. Self-compassion, for example, can be cultivated through self-directed practices, as evidenced by participants' predominant selection of moments of resilience related to this factor, according to app data. In contrast, fostering self-efficacy could rely more on sufficient learning opportunities and external feedback [86], which may be provided by a mental health professional, such as an eCoach. Increases in perceived social support might depend more on social interactions and environmental factors, potentially requiring more time to become manifest. Including peer support in the intervention could serve as one valuable option to enhance these social interactions.

Strengths and Limitations

This study has several key strengths that help it contribute to the body of research in the field. Its first strength is that it was conducted in accordance with the recommendations for resilience intervention studies outlined by Chmitorz et al [12], including an outcome-oriented resilience definition, and the separate assessment of various mental health outcomes and resilience factors, as well as adverse effects.

Second, to the best of our knowledge, the resilience intervention RESIST examined in this study is among the first to be explicitly grounded in a genuine resilience framework—namely, PASTOR [2], which served as the theoretical foundation of the intervention.

Third, the intervention's content was informed by resilience factors considered etiologically relevant for fostering adaptive responses to adversity [87] and which may reflect resilience-specific competencies.

Fourth, the intervention specifically targeted resilience factors that are typically not addressed in established mental health programs, such as stress management or mindfulness-based interventions. Lastly, by studying mediators, this study addresses the call to study mechanisms of change to optimize intervention development and treatment outcomes from digital resilience interventions [11].

Despite this study's strengths, the results should be interpreted in light of some limitations, with the first two relating to generalizability, the next two to methodological aspects, and the final two to theoretical considerations.

First, the findings may have limited generalizability to other implementation settings; for example, RESIST's effectiveness could differ if the intervention is employer-provided rather than individually initiated. Employees may view employer-led resilience training as contradictory if it ignores structural causes of stress. This concern is supported by a meta-analysis that detected smaller-than-expected effects for organization-implemented e-mental health interventions [88].

Second, this study's sample was predominantly female and highly educated, which may also limit generalizability. However, systematic reviews suggest that sociodemographic factors such as gender or education rarely modify intervention effects [89-91]. The main challenge may therefore be promoting uptake in other groups. Whether this may be achieved, for example, via approaches such as participatory design, tailored recruitment, or interventions addressing mental health indirectly (eg, physical exercise [92]), is not clear yet.

Third, the dropout rate across both intervention arms (82/350, 23.4%) could have negatively affected the results' validity. However, this extent of attrition falls well within the range of similar digital self-help interventions [69,71], and the multiple imputations approach used to address this shortfall is both a robust and widely accepted method for handling missing data [93]. In addition, a range of sensitivity analyses, including mixed model analyses, supported the robustness of the intention-to-treat sample's findings.

Fourth, nearly 40% (32/83) of participants indicated postintervention that they had not yet completed the intervention but intended to continue engaging with it. This may reflect the nature of the self-help format, which typically allows for greater flexibility. It also suggests that the intended pacing, one session and at least 2 app entries per week, may not have been well-matched to participants' preferences or natural usage patterns.

Fifth**,** the applied mediation model implicitly assumes that all mediators change simultaneously over time, thereby limiting its capacity to account for temporal variability in their effects—some mediators may exert their influence more quickly, while others do so more slowly. This limitation is common in studies assessing such models.

Lastly, the interpretability of the mediation analyses' results is limited with regard to drawing conclusions about any potential causal link between specific mediators and individual training components. For instance, it cannot be concluded that the content related to the resilience factors within the training necessarily encompasses the active ingredients driving the intervention's effects. Component studies are needed to further investigate the active ingredients of RESIST.

Theoretical Implications

Based on the study's insights, various theoretical implications and perspectives for future research emerge. First, almost 90% (309/350) of the participants reported experiencing an effort-reward imbalance in their work life and, accordingly, trained resilience in the context of ongoing stressor exposure. A next step would be to more precisely assess the specific stressors experienced by participants and relate them to mental health outcomes, as proposed by Kalisch et al [94]. This would, for instance, enable meaningful comparisons between individuals who show similar improvements in mental health after training, despite facing different levels of adversity.

Second, while the mediation analyses provided initial and valuable insights into the intervention's mechanisms of change, they also highlight several avenues for future research. The precise interrelations, causal directions, and underlying mechanisms linking the resilience factors optimism, self-compassion, and a positive appraisal style as potential resilience mechanisms warrant further investigation to deepen our understanding of their contribution to psychological

XSL•FO
RenderX

adaptation and resilience. Further, for the investigation of RESIST, a logical next step will be to include an assessment of positive appraisal style [95] as an additional construct (which has been published just recently) in future assessments and to examine it as a potential mediator in sequential mediation, onto which the various resilience factors may converge. Additionally, as already discussed, our findings suggest a lack of understanding regarding the varying timeframes required for different resilience factors to develop and change, as well as the degree to which their change can be regulated internally. Regarding the further development of resilience interventions and assuming that resilience factors require more time or external support to develop, so that change becomes better understood, such components could be positioned strategically at the beginning or end of an intervention. For research on mechanisms of change, this underscores the need for a more fine-grained level of investigation of mediators that also respects strict temporal precedence. One promising approach would be to assess the mediating role of resilience factors during the course of the intervention itself. To this end, methods such as session-by-session assessments or ecological momentary assessments in daily life between sessions may be particularly useful [96,97]. An alternative approach could involve using dynamic network models to capture the complexity inherent in change processes within interventions [98].

Third, by targeting multiple resilience factors, our intervention implicitly emphasized the importance of having access to a broad repertoire of resources to draw on various strategies, so individuals are able to respond flexibly to different challenging situations. This idea is reflected in the concept of regulatory flexibility [14], which emphasizes the importance of a good strategy-situation fit, as no single strategy is effective for all stressors. Explicitly integrating elements reflecting such a fit-based perspective into training could be a promising direction for future intervention development, as it could help individuals to learn how to select and apply the most suitable strategies depending on the specific demands of each situation. Examining changes in individuals' repertoire could be a further meaningful outcome for future intervention studies.

Lastly, given that individualized resilience interventions have not yet been explored sufficiently [11], tailoring the content of RESIST based on individual resource profiles could be a promising avenue for future research. For one study, Fassnacht et al [70] adopted a similar approach by integrating an assessment tool into their intervention, which provided participants with a profile of personal strengths and vulnerabilities to support informed decisions about which areas they should focus on. Building on this idea, incorporating a resilience profile assessment to identify participants' existing psychosocial resources—upon which the training might then be constructed—could personalize the approach and potentially enhance its effectiveness.

Practical Implications

The present findings yield several practical implications for implementing digital resilience training, such as RESIST, in occupational settings.

First, adherence to the self-help version of RESIST was low, reflecting a common challenge in digital interventions [99-101]. To enhance adherence, employers should consider recognizing participation in trainings such as RESIST as working time [102]. Low-cost technical measures, such as automated reminders, may also mitigate this issue [103]. More resource-intensive strategies, including personal recruitment, financial incentives, or professional guidance, can further improve engagement and adherence [104,105]. The use of adaptive systems—for example, offering on-demand support during self-help programs or enabling participants in guided formats to opt out of support when preferring independent training—can help limit associated costs. Cost-effectiveness analyses may assist in determining the optimal balance between resource investment to enhance adherence and the resulting health benefits in the population [106].

Second, the observed deterioration rates highlight the need for safety measures, such as providing personal support alongside the intervention or offering clear guidance on where to seek professional external help. Even if the reasons for deterioration remain unclear, as in this study, occupational health care professionals should be aware of potential deterioration and take ethical responsibility into account when offering mental health interventions.

Third, when inviting employees to participate in interventions, occupational health care professionals should communicate potential health benefits appropriately, given the importance of expectations for usage intentions, actual usage, and outcomes [107,108]. The observed response and deterioration rates, together with NNTB and NNTH, provide a solid basis for managing such expectations. Indicators based on standardized mean differences (eg, Cohen *d*) are frequently used but can be difficult for nonresearchers to understand [59] and may not be effective for expectation management.

Fourth, promoting the resilience of individual employees is an important but selective aspect of mental well-being. Interventions such as RESIST should be embedded in a comprehensive occupational health promotion strategy that includes a balanced set of measures addressing both individual and structural, workplace-related aspects [102]. Work-directed interventions—such as improving leadership culture, optimizing workloads, or enhancing effective communication within teams—should complement individual programs. This dual focus helps prevent the impression that employees are solely responsible for their well-being and reinforces the shared responsibility between staff and employers [102].

Conclusions

This study is among the first to evaluate the effectiveness of a self-help digital resilience intervention that was designed to enhance mental health during stressful times by promoting a set of theory- and evidence-informed selection of resilience factors representing trainable competencies. According to the PASTOR framework, they may serve as correlates of a positive appraisal style of stressors as a higher-order resilience mechanism. The interventional techniques used in the intervention, according to Strengths-Based CBT, proved to be feasible and effective, which confirms the results of the pilot

XSL•FO
RenderX

study [17]. Especially when intervention developers find the idea convincing that the methods used to promote resilience should be different from the methods used in psychotherapy aiming at reducing distress, then Strengths-Based CBT appears to be a promising option.

By studying resilience factors as mediators, this trial further addressed the recent call to investigate mechanisms of change within digital resilience interventions [11] to optimize interventional design and outcomes. Mediation analysis suggested that optimism and self-compassion may be important drivers promoting resilient outcomes, while redesigning the intervention seems to be needed to achieve stronger effects for self-efficacy and perceived social support. Future research is needed to investigate the relationships between resilience factors and higher-order resilience mechanisms, the ease of implementation, and the temporal dynamics of how resilience factors exert their protective effects. Finally, a standardized classification system for resilience interventions and a consensus on what constitutes such interventions are urgently needed to enable meaningful comparisons between resilience interventions. This should form the basis of a structured process to improve resilience interventions and make them more effective.

Acknowledgments

Generative artificial intelligence was used to assist in improving spelling and grammar in the initial draft of this paper. The final version was subsequently proofread by a professional editor.

Funding

This publication was funded by the Open Access Publication Fund of Leuphana University Lüneburg. The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of this paper.

Data Availability

The dataset generated and analyzed during this study is available in PubData, the institutional repository of Leuphana University for archiving and publishing datasets [109]. Access is provided by the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: DB (lead), DL (supporting), MW (supporting)
Data curation: SH
Formal analysis: SH
Investigation: DB
Methodology: DB (equal), DL (equal)
Project administration: DB
Supervision: DL
Visualization: SH
Writing – original draft: SH
Writing – review & editing: SH (lead), DB (equal), ND (supporting), SKS (supporting), MW (supporting), DL (equal)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Information on intervention description according to the TIDieR (Template for Intervention Description and Replication) checklist.
[DOCX File, 31 KB - jmir_v28i1e78335_app1.docx]

Multimedia Appendix 2
Intervention content of the web- and app-based resilience training RESIST.
[DOCX File, 17 KB - jmir_v28i1e78335_app2.docx]

Multimedia Appendix 3
Screenshots of the web- and app-based components of RESIST.
[DOCX File, 3645 KB - jmir_v28i1e78335_app3.docx]

Multimedia Appendix 4
Response formats and reliability of patient-reported outcomes measures.
[DOCX File, 15 KB - jmir_v28i1e78335_app4.docx]

XSL•FO
RenderX

Multimedia Appendix 5
Response and deterioration rates for the primary outcome measure of stress.
[DOCX File, 16 KB - jmir_v28i1e78335_app5.docx]

Multimedia Appendix 6
Between-group differences postintervention and at 3-month follow-up for mental health- and work-related secondary outcome measures.
[DOCX File, 18 KB - jmir_v28i1e78335_app6.docx]

Multimedia Appendix 7
Within-subject comparisons for the intervention group from baseline to 6-month follow-up.
[DOCX File, 18 KB - jmir_v28i1e78335_app7.docx]

Multimedia Appendix 8
Additional mediation analyses' results.
[DOCX File, 265 KB - jmir_v28i1e78335_app8.docx]

Multimedia Appendix 9
Engagement with the web and app components of the intervention.
[DOCX File, 14 KB - jmir_v28i1e78335_app9.docx]

Checklist 1
CONSORT-eHEALTH checklist (V1.6).
[PDF File, 1219 KB - jmir_v28i1e78335_app10.pdf]

References

1. Bonanno GA, Westphal M, Mancini AD. Resilience to loss and potential trauma. Annu Rev Clin Psychol 2011;7(1):511-535. [doi: 10.1146/annurev-clinpsy-032210-104526] [Medline: 21091190]
2. Kalisch R, Müller MB, Tüscher O. A conceptual framework for the neurobiological study of resilience. Behav Brain Sci 2015;38:e92. [doi: 10.1017/S0140525X1400082X]
3. Kalisch R, Baker DG, Basten U, et al. The resilience framework as a strategy to combat stress-related disorders. Nat Hum Behav 2017 Nov;1(11):784-790. [doi: 10.1038/s41562-017-0200-8] [Medline: 31024125]
4. Köhler CA, Evangelou E, Stubbs B, et al. Mapping risk factors for depression across the lifespan: an umbrella review of evidence from meta-analyses and Mendelian randomization studies. J Psychiatr Res 2018 Aug;103:189-207. [doi: 10.1016/j.jpsychires.2018.05.020]
5. Vos T, Lim SS, Abbafati C, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct;396(10258):1204-1222. [doi: 10.1016/S0140-6736(20)30925-9]
6. Santomauro DF, Herrera AMM, Shadid J, et al. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 2021 Nov;398(10312):1700-1712. [doi: 10.1016/S0140-6736(21)02143-7]
7. Ang WHD, Chew HSJ, Dong J, Yi H, Mahendren R, Lau Y. Digital training for building resilience: systematic review, meta-analysis, and meta-regression. Stress Health 2022 Dec;38(5):848-869. [doi: 10.1002/smi.3154] [Medline: 35460533]
8. Díaz-García A, Franke M, Herrero R, Ebert DD, Botella C. Theoretical adequacy, methodological quality and efficacy of online interventions targeting resilience: a systematic review and meta-analysis. Eur J Public Health 2021 Jul 7;31(31 Suppl 1):i11-i18. [doi: 10.1093/eurpub/ckaa255] [Medline: 34240159]
9. Ferreira M, Marques A, Gomes PV. Individual resilience interventions: a systematic review in adult population samples over the last decade. Int J Environ Res Public Health 2021 Jul 16;18(14):7564. [doi: 10.3390/ijerph18147564] [Medline: 34300018]
10. Joyce S, Shand F, Tighe J, Laurent SJ, Bryant RA, Harvey SB. Road to resilience: a systematic review and meta-analysis of resilience training programmes and interventions. BMJ Open 2018 Jun 14;8(6):e017858. [doi: 10.1136/bmjopen-2017-017858] [Medline: 29903782]
11. Schäfer SK, von Boros L, Schaubruch LM, et al. Digital interventions to promote psychological resilience: a systematic review and meta-analysis. NPJ Digit Med 2024 Feb 8;7(1):30. [doi: 10.1038/s41746-024-01017-8] [Medline: 38332030]
12. Chmitorz A, Kunzler A, Helmreich I, et al. Intervention studies to foster resilience – a systematic review and proposal for a resilience framework in future intervention studies. Clin Psychol Rev 2018 Feb;59:78-100. [doi: 10.1016/j.cpr.2017.11.002]
13. Folkman S. Stress: Appraisal and Coping: Springer; 1984. [doi: 10.1007/978-1-4419-1005-9_215]

XSL•FO
RenderX

14. Bonanno GA, Chen S, Galatzer-Levy IR. Resilience to potential trauma and adversity through regulatory flexibility. Nat Rev Psychol 2023;2(11):663-675. [doi: 10.1038/s44159-023-00233-5]
15. Harrer M, Adam SH, Fleischmann RJ, et al. Effectiveness of an internet- and app-based intervention for college students with elevated stress: randomized controlled trial. J Med Internet Res 2018;20(4):e136. [doi: 10.2196/jmir.9293]
16. Aikens KA, Astin J, Pelletier KR, et al. Mindfulness goes to work: impact of an online workplace intervention. J Occup Environ Med 2014 Jul;56(7):721-731. [doi: 10.1097/JOM.0000000000000209] [Medline: 24988100]
17. Behrendt D, Boß L, Hannibal S, Kunzler AM, Wessa M, Lehr D. Feasibility and efficacy of a digital resilience training: a pilot study of the strengths-based training RESIST. Internet Interv 2023 Sep;33:100649. [doi: 10.1016/j.invent.2023.100649] [Medline: 37545556]
18. Kazdin AE. Mediators and mechanisms of change in psychotherapy research. Annu Rev Clin Psychol 2007;3(1):1-27. [doi: 10.1146/annurev.clinpsy.3.022806.091432] [Medline: 17716046]
19. Kraemer HC, Wilson GT, Fairburn CG, Agras WS. Mediators and moderators of treatment effects in randomized clinical trials. Arch Gen Psychiatry 2002 Oct;59(10):877-883. [doi: 10.1001/archpsyc.59.10.877] [Medline: 12365874]
20. Cieslak R, Benight CC, Rogala A, et al. Effects of internet-based self-efficacy intervention on secondary traumatic stress and secondary posttraumatic growth among health and human services professionals exposed to indirect trauma. Front Psychol 2016;7:1009. [doi: 10.3389/fpsyg.2016.01009] [Medline: 27458407]
21. Schotanus-Dijkstra M, Pieterse ME, Drossaert CHC, Walburg JA, Bohlmeijer ET. Possible mechanisms in a multicomponent email guided positive psychology intervention to improve mental well-being, anxiety and depression: a multiple mediation model. J Posit Psychol 2019 Mar 4;14(2):141-155. [doi: 10.1080/17439760.2017.1388430]
22. Forbes S, Fikretoglu D. Building resilience: the conceptual basis and research evidence for resilience training programs. Rev Gen Psychol 2018 Dec;22(4):452-468. [doi: 10.1037/gpr0000152]
23. Padesky CA, Mooney KA. Strengths-based cognitive-behavioural therapy: a four-step model to build resilience. Clin Psychol Psychother 2012;19(4):283-290. [doi: 10.1002/cpp.1795] [Medline: 22653834]
24. Eysenbach G, CONSORT-eHEALTH Group. CONSORT-eHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
25. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: Template for Intervention Description and Replication (TIDieR) checklist and guide. BMJ 2014 Mar 7;348:g1687. [doi: 10.1136/bmj.g1687] [Medline: 24609605]
26. Beck AT, Steer RA, Brown G. Beck Depression Inventory–II: American Psychological Association; 1996. [doi: 10.1037/t00742-000]
27. Heber E, Ebert DD, Lehr D, et al. The benefit of web- and computer-based interventions for stress: a systematic review and meta-analysis. J Med Internet Res 2017 Feb 17;19(2):e32. [doi: 10.2196/jmir.5774] [Medline: 28213341]
28. Stewart DE, Yuen T. A systematic review of resilience in the physically ill. Psychosomatics 2011;52(3):199-209. [doi: 10.1016/j.psym.2011.01.036] [Medline: 21565591]
29. Khazanov GK, Ruscio AM. Is low positive emotionality a specific risk factor for depression? A meta-analysis of longitudinal studies. Psychol Bull 2016 Sep;142(9):991-1015. [doi: 10.1037/bul0000059] [Medline: 27416140]
30. Gallagher MW, Long LJ, Phillips CA. Hope, optimism, self-efficacy, and posttraumatic stress disorder: a meta-analytic review of the protective effects of positive expectancies. J Clin Psychol 2020 Mar;76(3):329-355. [doi: 10.1002/jclp.22882] [Medline: 31714617]
31. Wang Y, Chung MC, Wang N, Yu X, Kenardy J. Social support and posttraumatic stress disorder: a meta-analysis of longitudinal studies. Clin Psychol Rev 2021 Apr;85:101998. [doi: 10.1016/j.cpr.2021.101998] [Medline: 33714168]
32. GetOn Platform. URL: https://coach.geton-training.de/ [accessed 2025-12-04]
33. Heber E, Lehr D, Ebert DD, Berking M, Riper H. Web-based and mobile stress management intervention for employees: a randomized controlled trial. J Med Internet Res 2016 Jan 27;18(1):e21. [doi: 10.2196/jmir.5112] [Medline: 26818683]
34. Klein EM, Brähler E, Dreier M, et al. The German version of the Perceived Stress Scale - psychometric characteristics in a representative German community sample. BMC Psychiatry 2016 May 23;16(1):159. [doi: 10.1186/s12888-016-0875-9] [Medline: 27216151]
35. Reis D, Lehr D, Heber E, Ebert DD. The German Version of the Perceived Stress Scale (PSS-10): evaluation of dimensionality, validity, and measurement invariance with exploratory and confirmatory bifactor modeling. Assessment 2019 Oct;26(7):1246-1259. [doi: 10.1177/1073191117715731] [Medline: 28627220]
36. Chmitorz A, Wenzel M, Stieglitz RD, et al. Population-based validation of a German version of the Brief Resilience Scale. PLoS One 2018;13(2):e0192761. [doi: 10.1371/journal.pone.0192761] [Medline: 29438435]
37. Beierlein C, Kovaleva A, Kemper CJ, Rammstedt B. Ein Messinstrument zur Erfassung subjektiver Kompetenzerwartungen: Allgemeine Selbstwirksamkeit Kurzskala (ASKU) [Website in German]. 2012. URL: https://www.ssoar.info/ssoar/bitstream/handle/document/29235/ssoar-2012-beierlein_et_al-ein_messinstrument_zur_erfassung_subjektiver.pdf?sequence=1 [accessed 2025-12-04]
38. Glaesmer H, Hoyer J, Klotsche J, Herzberg PY. Die Deutsche version des Life-Orientation-Tests (LOT-R) zum dispositionellen optimismus und pessimismus [Article in German]. Z Gesundheitspsychologie 2008 Jan;16(1):26-31. [doi: 10.1026/0943-8149.16.1.26]

XSL•FO
RenderX

39. Hupfeld J, Ruffieux N. Validierung einer Deutschen version der Self-Compassion Scale (SCS-D) [Article in German]. Z Klin Psychol Psychother 2011 Apr;40(2):115-123. [doi: 10.1026/1616-3443/a000088]
40. Schulz U, Schwarzer R. Soziale unterstützung bei der krankheitsbewältigung: die Berliner Social Support Skalen (BSSS) [Article in German]. Diagnostica 2003 Apr;49(2):73-82. [doi: 10.1026//0012-1924.49.2.73]
41. Hautzinger M, Bailer M, Hofmeister D. Center for Epidemiological Studies Depression Scale (CES-D; Radloff, LS, 1977)-German adaptation. Psychiatr Prax 2012;39(6):302-304. [doi: 10.1055/s-0032-1326702]
42. Lehr D, Hillert A, Schmitz E, Sosnowsky N. Screening depressiver Störungen mittels Allgemeiner Depressions-Skala (ADS-K) und State-Trait Depressions Scales (STDS-T). Diagnostica 2008 Apr;54(2):61-70 [FREE Full text] [doi: 10.1026/0012-1924.54.2.61]
43. Spitzer C, Hammer S, Löwe B, et al. Die kurzform des Brief Symptom Inventory (BSI -18): erste befunde zu den psychometrischen kennwerten der Deutschen version [Article in German]. Fortschr Neurol Psychiatr 2011 Sep;79(9):517-523. [doi: 10.1055/s-0031-1281602]
44. Mancini AD, Bonanno GA. Predictors and parameters of resilience to loss: toward an individual differences model. J Pers 2009 Dec;77(6):1805-1832. [doi: 10.1111/j.1467-6494.2009.00601.x] [Medline: 19807863]
45. Canli T, Qiu M, Omura K, et al. Neural correlates of epigenesis. Proc Natl Acad Sci U S A 2006 Oct 24;103(43):16033-16038. [doi: 10.1073/pnas.0601674103] [Medline: 17032778]
46. Chmitorz A, Kurth K, Mey LK, et al. Assessment of microstressors in adults: questionnaire development and ecological validation of the Mainz inventory of microstressors. JMIR Ment Health 2020 Feb 24;7(2):e14566. [doi: 10.2196/14566] [Medline: 32130154]
47. Rödel A, Siegrist J, Hessel A, Brähler E. Fragebogen zur messung beruflicher gratifikationskrisen [Article in German]. Z Differentielle Diagnostische Psychol 2004 Jan;25(4):227-238. [doi: 10.1024/0170-1789.25.4.227]
48. Siegrist J, Wege N, Pühlhofer F, Wahrendorf M. A short generic measure of work stress in the era of globalization: effort-reward imbalance. Int Arch Occup Environ Health 2009 Aug;82(8):1005-1013. [doi: 10.1007/s00420-008-0384-3] [Medline: 19018554]
49. Lehr D, Koch S, Hillert A. Where is (im)balance? Necessity and construction of evaluated cut - off points for effort - reward imbalance and overcommitment. J Occup O Psychol 2010 Mar;83(1):251-261. [doi: 10.1348/096317909X406772]
50. Bouwmans C, De Jong K, Timman R, et al. Feasibility, reliability and validity of a questionnaire on healthcare consumption and productivity loss in patients with a psychiatric disorder (TiC-P). BMC Health Serv Res 2013 Jun 15;13(1):217. [doi: 10.1186/1472-6963-13-217] [Medline: 23768141]
51. Ahlstrom L, Grimby-Ekman A, Hagberg M, Dellve L. The work ability index and single-item question: associations with sick leave, symptoms, and health – a prospective study of women on long-term sick leave. Scand J Work Environ Health 2010 Sep;36(5):404-412. [doi: 10.5271/sjweh.2917]
52. Boß L, Lehr D, Reis D, et al. Reliability and validity of assessing user satisfaction with web-based health interventions. J Med Internet Res 2016 Aug 31;18(8):e234. [doi: 10.2196/jmir.5952] [Medline: 27582341]
53. Hassenzahl M, Monk A. The inference of perceived usability from beauty. Hum-Comput Interact 2010 Jul;25(3):235-260. [doi: 10.1080/07370024.2010.500139]
54. R Core Team. R: a language and environment for statistical computing. The R Foundation. 2023. URL: https://www.R-project.org/ [accessed 2025-12-04]
55. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw 2011;45:1-67. [doi: 10.18637/jss.v045.i03]
56. Harrer M, Cuijpers P, Schuurmans LKJ, et al. Evaluation of randomized controlled trials: a primer and tutorial for mental health researchers. Trials 2023 Aug 30;24(1):562. [doi: 10.1186/s13063-023-07596-3] [Medline: 37649083]
57. Rubin DB. Multiple Imputation for Nonresponse in Surveys: John Wiley & Sons; 2004. [doi: 10.1002/9780470316696]
58. O'Connell NS, Dai L, Jiang Y, et al. Methods for analysis of pre-post data in clinical research: a comparison of five common methods. J Biom Biostat 2017 Feb 24;8(1):1-8. [doi: 10.4172/2155-6180.1000334] [Medline: 30555734]
59. Cuijpers P. Has the time come to stop using the "standardised mean difference"? Clin Psychol Eur 2021 Sep;3(3):e6835. [doi: 10.32872/cpe.6835] [Medline: 36398102]
60. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 1991;59(1):12-19. [doi: 10.1037/0022-006X.59.1.12]
61. Carrasco-Labra A, Devji T, Qasim A, et al. Minimal important difference estimates for patient-reported outcomes: a systematic survey. J Clin Epidemiol 2021 May;133:61-71. [doi: 10.1016/j.jclinepi.2020.11.024] [Medline: 33321175]
62. Bauer-Staeb C, Kounali DZ, Welton NJ, et al. Effective dose 50 method as the minimal clinically important difference: evidence from depression trials. J Clin Epidemiol 2021 Sep;137:200-208. [doi: 10.1016/j.jclinepi.2021.04.002] [Medline: 33892086]
63. Ebert DD, Franke M, Zarski AC, et al. Effectiveness and moderators of an internet-based mobile-supported stress management intervention as a universal prevention approach: randomized controlled trial. J Med Internet Res 2021 Dec 22;23(12):e22107. [doi: 10.2196/22107] [Medline: 34941541]
64. Jorgensen TD, Pornprasertmanit S, Schoemann AM, et al. SemTools: useful tools for structural equation modeling. The Comprehensive R Archive Network. 2022. URL: https://CRAN.R-project.org/package=semTools [accessed 2025-12-04]

XSL•FO
RenderX

65. Hayes AF, Rockwood NJ. Regression-based statistical mediation and moderation analysis in clinical research: observations, recommendations, and implementation. Behav Res Ther 2017 Nov;98:39-57. [doi: 10.1016/j.brat.2016.11.001] [Medline: 27865431]
66. Roig AE, Mooney O, Salamanca-Sanabria A, Lee CT, Farrell S, Richards D. Assessing the efficacy and acceptability of a web-based intervention for resilience among college students: pilot randomized controlled trial. JMIR Form Res 2020 Nov 11;4(11):e20167. [doi: 10.2196/20167] [Medline: 33174530]
67. Victor P, Krug I, Vehoff C, Lyons N, Willutzki U. Strengths-based CBT: internet-based versus face-to-face therapy in a randomized controlled trial. J Depress Anxiety 2018;07(2):1000301. [doi: 10.4172/2167-1044.1000301]
68. Victor PP, Teismann T, Willutzki U. A pilot evaluation of a strengths-based CBT intervention module with college students. Behav Cogn Psychother 2017 Jul;45(4):427-431. [doi: 10.1017/S1352465816000552] [Medline: 28347377]
69. Ebert DD, Heber E, Berking M, et al. Self-guided internet-based and mobile-based stress management for employees: results of a randomised controlled trial. Occup Environ Med 2016 May;73(5):315-323. [doi: 10.1136/oemed-2015-103269] [Medline: 26884049]
70. Fassnacht DB, Ali K, van Agteren J, et al. A group-facilitated, internet-based intervention to promote mental health and well-being in a vulnerable population of university students: randomized controlled trial of the be well plan program. JMIR Ment Health 2022 May 5;9(5):e37292. [doi: 10.2196/37292] [Medline: 35471196]
71. Lehr D, Freund H, Sieland B, et al. Effectiveness of a guided multicomponent internet and mobile gratitude training program - a pragmatic randomized controlled trial. Internet Interv 2024 Dec;38:100787. [doi: 10.1016/j.invent.2024.100787] [Medline: 39635229]
72. Polivy J, Herman CP. The false-hope syndrome: unfulfilled expectations of self-change. Curr Dir Psychol Sci 2000;9(4):128-131. [doi: 10.1111/1467-8721.00076]
73. Carver CS, Scheier MF. Dispositional optimism. Trends Cogn Sci 2014 Jun;18(6):293-299. [doi: 10.1016/j.tics.2014.02.003] [Medline: 24630971]
74. Malouff JM, Schutte NS. Can psychological interventions increase optimism? A meta-analysis. J Posit Psychol 2017 Nov 2;12(6):594-604. [doi: 10.1080/17439760.2016.1221122]
75. Neff KD. Self-compassion: theory, method, research, and intervention. Annu Rev Psychol 2023 Jan 18;74(1):193-218. [doi: 10.1146/annurev-psych-032420-031047] [Medline: 35961039]
76. Cook JA, Julious SA, Sones W, et al. Practical help for specifying the target difference in sample size calculations for RCTs: the DELTA2 five-stage study, including a workshop. Health Technol Assess 2019 Oct;23(60):1-88. [doi: 10.3310/hta23600] [Medline: 31661431]
77. Alarcon GM, Bowling NA, Khazon S. Great expectations: a meta-analytic examination of optimism and hope. Pers Individ Dif 2013 May;54(7):821-827. [doi: 10.1016/j.paid.2012.12.004]
78. Kaida N, Kaida K. Positive associations of optimism-pessimism orientation with pro-environmental behavior and subjective well-being: a longitudinal study on quality of life and everyday behavior. Qual Life Res 2019 Dec;28(12):3323-3332. [doi: 10.1007/s11136-019-02273-y] [Medline: 31422540]
79. Prati G, Pietrantoni L. Optimism, social support, and coping strategies as factors contributing to posttraumatic growth: a meta-analysis. J Loss Trauma 2009 Aug 26;14(5):364-388. [doi: 10.1080/15325020902724271]
80. Romswinkel EV, König HH, Hajek A. The role of optimism in the relationship between job stress and depressive symptoms. Longitudinal findings from the German Ageing Survey. J Affect Disord 2018 Dec 1;241:249-255. [doi: 10.1016/j.jad.2018.08.005] [Medline: 30138809]
81. Lee EE, Govind T, Ramsey M, et al. Compassion toward others and self-compassion predict mental and physical well-being: a 5-year longitudinal study of 1090 community-dwelling adults across the lifespan. Transl Psychiatry 2021;11(1):1-9. [doi: 10.1038/s41398-021-01491-8]
82. MacBeth A, Gumley A. Exploring compassion: a meta-analysis of the association between self-compassion and psychopathology. Clin Psychol Rev 2012 Aug;32(6):545-552. [doi: 10.1016/j.cpr.2012.06.003]
83. Stutts LA, Leary MR, Zeveney AS, Hufnagle AS. A longitudinal analysis of the relationship between self-compassion and the psychological effects of perceived stress. Self Identity 2018 Nov 2;17(6):609-626. [doi: 10.1080/15298868.2017.1422537]
84. Suh H, Jeong J. Association of self-compassion with suicidal thoughts and behaviors and non-suicidal self injury: a meta-analysis. Front Psychol 2021;12:633482. [doi: 10.3389/fpsyg.2021.633482] [Medline: 34122224]
85. Zessin U, Dickhäuser O, Garbade S. The relationship between self-compassion and well-being: a meta-analysis. Appl Psychol Health Well Being 2015 Nov;7(3):340-364. [doi: 10.1111/aphw.12051] [Medline: 26311196]
86. Bandura A. Self-efficacy: toward a unifying theory of behavioral change. Adv Behav Res Ther 1978 Jan;1(4):139-161. [doi: 10.1016/0146-6402(78)90002-4]
87. Kunzler AM, Helmreich I, Chmitorz A, et al. Psychological interventions to foster resilience in healthcare professionals. Cochrane Database Syst Rev 2020 Jul 5;7(7):CD012527. [doi: 10.1002/14651858.CD012527.pub2] [Medline: 32627860]
88. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on Internet-based mental health interventions — a systematic review. Internet Interv 2014 Oct;1(4):205-215. [doi: 10.1016/j.invent.2014.08.003]

XSL•FO
RenderX

89. Haller K, Becker P, Niemeyer H, Boettcher J. Who benefits from guided internet-based interventions? A systematic review of predictors and moderators of treatment outcome. Internet Interv 2023 Sep;33:100635. [doi: 10.1016/j.invent.2023.100635] [Medline: 37449052]
90. Reins JA, Buntrock C, Zimmermann J, et al. Efficacy and moderators of internet-based interventions in adults with subthreshold depression: an individual participant data meta-analysis of randomized controlled trials. Psychother Psychosom 2021;90(2):94-106. [doi: 10.1159/000507819] [Medline: 32544912]
91. Thielecke J, Kuper P, Lehr D, et al. Who benefits from indirect prevention and treatment of depression using an online intervention for insomnia? Results from an individual-participant data meta-analysis. Psychol Med 2024 Jul;54(10):2389-2402. [doi: 10.1017/S0033291724000527] [Medline: 38469832]
92. Kazdin AE. Indirect interventions: lifestyle options to treat mental disorders. Healthcare (Basel) 2025 Feb 26;13(5):505. [doi: 10.3390/healthcare13050505] [Medline: 40077067]
93. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009 Jun 29;338:b2393. [doi: 10.1136/bmj.b2393] [Medline: 19564179]
94. Kalisch R, Köber G, Binder H, et al. The frequent stressor and mental health monitoring-paradigm: a proposal for the operationalization and measurement of resilience and the identification of resilience processes in longitudinal observational studies. Front Psychol 2021;12:710493. [doi: 10.3389/fpsyg.2021.710493] [Medline: 34539510]
95. Petri-Romão P, Engen H, Rupanova A, et al. Self-report assessment of Positive Appraisal Style (PAS): development of a process-focused and a content-focused questionnaire for use in mental health and resilience research. PLoS One 2024;19(2):e0295562. [doi: 10.1371/journal.pone.0295562] [Medline: 38306328]
96. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4(1):1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
97. Trull TJ, Ebner-Priemer U. Ambulatory assessment. Annu Rev Clin Psychol 2013;9(1):151-176. [doi: 10.1146/annurev-clinpsy-050212-185510] [Medline: 23157450]
98. Hofmann SG, Curtiss JE, Hayes SC. Beyond linear mediation: toward a dynamic network approach to study treatment processes. Clin Psychol Rev 2020 Mar;76:101824. [doi: 10.1016/j.cpr.2020.101824] [Medline: 32035297]
99. Boucher EM, Raiker JS. Engagement and retention in digital mental health interventions: a narrative review. BMC Digit Health 2024;2(1):52. [doi: 10.1186/s44247-024-00105-9]
100. Gan DZQ, McGillivray L, Larsen ME, Christensen H, Torok M. Technology-supported strategies for promoting user engagement with digital mental health interventions: a systematic review. Digit Health 2022;8:20552076221098268. [doi: 10.1177/20552076221098268] [Medline: 35677785]
101. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: a systematic and meta-analytic review. J Consult Clin Psychol 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
102. Boß L, Ross J, Reis D, et al. Effectiveness of an integrated platform-based intervention for promoting psychosocial safety climate and mental health in nursing staff: a pragmatic cluster randomised controlled trial. Int J Nurs Stud 2025 Jul;167:105076. [doi: 10.1016/j.ijnurstu.2025.105076] [Medline: 40215602]
103. Alkhaldi G, Hamilton FL, Lau R, Webster R, Michie S, Murray E. The effectiveness of prompts to promote engagement with digital interventions: a systematic review. J Med Internet Res 2016 Jan 8;18(1):e6. [doi: 10.2196/jmir.4790] [Medline: 26747176]
104. Apolinário-Hagen J, Fritsche L, Bierhals C, Salewski C. Improving attitudes toward e-mental health services in the general population via psychoeducational information material: a randomized controlled trial. Internet Interv 2018 Jun;12:141-149. [doi: 10.1016/j.invent.2017.12.002] [Medline: 30135778]
105. Werntz A, Amado S, Jasman M, Ervin A, Rhodes JE. Providing human support for the use of digital mental health interventions: systematic meta-review. J Med Internet Res 2023 Feb 6;25(1):e42864. [doi: 10.2196/42864] [Medline: 36745497]
106. Freund J, Smit F, Lehr D, et al. A universal digital stress management intervention for employees: randomized controlled trial with health-economic evaluation. J Med Internet Res 2024 Oct 22;26:e48481. [doi: 10.2196/48481] [Medline: 39437382]
107. Blut M, Chong AYL, Tsigna Z, et al. Meta-analysis of the Unified Theory of Acceptance and Use of Technology (UTAUT): challenging its validity and charting a research agenda in the red ocean. JAIS 2022;23(1):13-95. [doi: 10.17705/1jais.00719]
108. Thielecke J, Kuper P, Ebert D, et al. Does outcome expectancy predict outcomes in online depression prevention? Secondary analysis of randomised-controlled trials. Health Expect 2024 Feb;27(1):e13951. [doi: 10.1111/hex.13951] [Medline: 39102655]
109. Hannibal S, Behrendt D, Wessa M, Schäfer SK, Dalkner N, Lehr D. RCT Outcome Data RESIST Unguided 2025. [doi: 10.48548/pubdata-2437]

Abbreviations

ANCOVA: analysis of covariance
CBT: cognitive behavioral therapy

XSL•FO
RenderX

CONSORT-eHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
IG: intervention group
NNTB: numbers needed to treat for benefit
NNTH: numbers needed to treat for harm
PASTOR: Positive Appraisal Style Theory of Resilience
PMRe: personal model of resilience
PSS-10: Perceived Stress Scale-10
RCT: randomized controlled trial
TIDieR: Template for Intervention Description and Replication
WL: waitlist control group

Edited by A Mavragani, TDA Cardoso; submitted 31.May.2025; peer-reviewed by RK Kanji, X Liang; revised version received 28.Oct.2025; accepted 29.Oct.2025; published 05.Jan.2026.

Please cite as:
Hannibal S, Behrendt D, Wessa M, Schäfer SK, Dalkner N, Lehr D
Effectiveness of and Mechanisms of Change in a Self-Help Web- and App-Based Resilience Intervention on Perceived Stress in the General Working Population: Randomized Controlled Trial
J Med Internet Res 2026;28:e78335
URL: https://www.jmir.org/2026/1/e78335
doi:10.2196/78335

© Sandy Hannibal, Dörte Behrendt, Michèle Wessa, Sarah K Schäfer, Nina Dalkner, Dirk Lehr. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Predictors of Professional Responses in Nonprofit Mental Health Forums: Interpretable Machine Learning Analysis

Shuang Geng[1], DPhil; Yanghui Li[2], MEng; Jie Wang[2], MEng; Peixuan Chen[3], MSci; Xusheng Wu[4], MEng; Zhiqun Zhang[5], MSci

[1]College of Management, Shenzhen University, Shenzhen, Guangdong, China
[2]College of Artificial Intelligence, Shenzhen University, Shenzhen, Guangdong, China
[3]School of Life Sciences, Central South University, Changsha, Hunan, China
[4]Shenzhen Health Development Research and Data Management Center, Shenzhen, Guangdong, China
[5]Shenzhen Traditional Chinese Medicine Hospital, Shenzhen, Guangdong, China

Corresponding Author:
Xusheng Wu, MEng
Shenzhen Health Development Research and Data Management Center
5 Yuanling Road, Futian
Shenzhen, Guangdong, 518000
China
Phone: 86 19928750177
Email: jysgwuxusheng@163.com

Abstract

Background: Online mental health communities increase access and equity for patients seeking psychological support. User demand and professional contributions are key to their sustainability. While previous research has examined factors influencing physicians' participation in online consultation platforms, limited attention has been given to how post characteristics affect the quantity and length of professional responses in nonprofit mental health communities.

Objective: This study aims to examine how textual (ie, topic, sentiment, title length, and content length) and contextual (ie, page views and posting time) characteristics of inquiries in nonprofit mental health forums influence the quantity and length of responses from mental health professionals, providing insights for enhancing community interactions.

Methods: We collected 18,572 question-and-answer records from a Chinese online mental health platform (August 2024-July 2025). Topic features were extracted using BERTopic, and sentiment features were obtained through a distilled Bidirectional Encoder Representations from Transformers–based sentiment classification model. Additional features were derived from post metadata. We compared 5 machine learning models and identified Light Gradient Boosting Machine as the best performer. We then applied Shapley Additive Explanations (SHAP) analysis to it to evaluate the feature contributions to the prediction of response quantity and length.

Results: In virtual mental health communities, user inquiries fall into 7 topic categories: work, love, depression, boyfriends or girlfriends, school, marriage, and family. Depression-related topics negatively predict response quantity, whereas interpersonal, school, marriage, or family topics are positively correlated. SHAP analysis revealed that page views (SHAP value=0.187) and title length (SHAP value=0.073) are key factors in predicting response quantity, and content length (SHAP value=0.274), sentiment category (SHAP value=0.054), and title length (SHAP value=0.053) are key factors in predicting response length. Posts exhibiting negative emotions are positively related to both the predicted quantity and length of responses, and this effect becomes more pronounced as the degree of emotional intensity increases. Titles with 15-20 characters and content with more than 60 characters are positively correlated with responses, whereas titles with fewer than 7 characters have negative effects. Higher view counts and weekday posts also increase response likelihood.

Conclusions: This study provides insights into how textual and contextual features of patient posts influence professional responses in nonprofit mental health forums. It enhances understanding of voluntary knowledge contribution behaviors in online mental health communities and offers practical guidance for optimizing platform functional design and user posting strategies. Future researchers are encouraged to address the limitations of this study, which focuses solely on response quantity and length, and to explore details of professional responses, such as by developing a comprehensive measure of response quality.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e74359) doi:10.2196/74359

KEYWORDS

influencing factors; online mental health community; predictive analysis; response length; response quantity; sentiment analysis; theme analysis

Introduction

Background

Mental health disorders are critical global health concerns that pose major challenges to both individual well-being and health care systems [1-4]. According to the World Health Organization, approximately 1 billion individuals (more than 12.5% of adults and adolescents) are affected by mental health disorders, which account for approximately 5% of disability-adjusted life years [5]. Individuals living with mental health conditions experience a high burden of illness and face an increased risk of mortality [6]. Therefore, strengthening mental health services has emerged as a notable research concern [7].

In recent years, an increasing number of online mental health communities (OMHCs) have been established, allowing mental health professionals from offline hospitals to offer services on these platforms. These online mental health platforms connect mental health service providers with patient users seeking psychological support and are gaining popularity [8,9]. The benefits of OMHCs are manyfold. For example, OMHCs effectively mitigate social prejudice and stigma associated with mental health disorders. Many individuals delay or avoid seeking professional help because of concerns about being labeled. The anonymity setting of these platforms alleviates such concerns and encourages people with mental health challenges to actively seek psychological support [10,11]. Moreover, OMHCs help address the time and geographical limitations of offline clinical services. This also addresses the uneven distribution of mental health resources and enhances the accessibility and equity of mental health services [12]. Given its importance, this study focuses on OMHCs to better understand the contributing behavior of health professionals on these platforms.

Prior studies on online mental health can be classified into 3 main categories: studies that focus on patient users (eg, user-generated content, user engagement patterns) [13-17], studies that focus on health professionals (eg, contribution behaviors) [18,19], and studies that focus on communities (eg, community quality and value cocreation) [20-22]. Existing studies on the contribution behavior of health professionals suggest that their active participation within the community determines the effectiveness of counseling services [18]. For OMHC managers, understanding the factors that influence the contribution behaviors of health professionals is crucial for improving both the efficiency and quality of community responses, promoting sustained user participation, and enhancing user retention and engagement [23]. From the perspective of patient users, when they seek medical advice and feelings of expression in OMHCs, their posts signal their need for informational or emotional support from professionals. Responses from professional community members may make them feel valued and accepted and enhance their sense of belonging to the community. Moreover, the information shared by professionals benefits not only help-seekers but also lurkers and other community members [24]. Together, the engagement of both patients and professionals helps cultivate a sustainable ecosystem for health knowledge exchange [25]. Therefore, it is imperative to focus on mental health professionals and examine the factors influencing their knowledge contributions in OMHCs.

Previous studies have provided insights into the factors influencing health professionals' knowledge contribution behaviors [20,26-33]. These factors include content-related factors, such as question type, information quality, and readability [26-28]; community-related factors, such as reward mechanisms [26]; and health professional–related factors, such as social relationships, self-efficacy, and reputation needs [30-33]. For example, Srivastava et al [27] analyzed the intent, criticism, readability, and emotion of user posts on Reddit, and found the prominence of "self-criticism" as the most prevalent form of criticism expressed by help-seekers. They also found that individuals who explicitly express their need for help are more likely to receive assistance. These studies explore these factors primarily through the lens of signal theory, motivational theory, self-determination theory, and social exchange theory. For instance, Chen et al [28] reported that emotional and informational language signals increase the likelihood of professionals providing informational and emotional support. Drawing on social exchange theory, Wang et al [29] reported that material and psychological rewards significantly increase the online contributions of health professionals. Additionally, several studies have explored the impact of the intrinsic and extrinsic motivations of professionals' contribution behaviors. Imlawi and Gregg [32] noted that factors such as helping motivator, reputation motivator, and moral obligation motivator influence professionals' contribution continuance intentions. Maheshwari et al [33] found that self-efficacy and reciprocity positively influence the attitude toward knowledge sharing; however, the rewards' moderating effect is not significant.

While there is a wide range of factors that influence professionals' knowledge contribution, post characteristics serve as the most direct medium for patient users to share stories and engage with the community, and they are closely associated with social interaction outcomes [27]. The informational cues (eg, topic) and emotional cues (eg, sentiment) in a post are crucial signals for consultants to understand the patient users' problem and to decide whether to respond. For example, posts with clear problem descriptions may reduce the cognitive effort required by consultants, which is especially important when the professionals are volunteers with limited time. Research in computer-mediated communication suggests that emotional cues are important drivers of social support [34]. Beyond content, posting time is a critical factor that determines a post's visibility. For example, a post published late at night may be quickly buried under a flood of newer posts, making it receive

XSL•FO
RenderX

fewer replies. Investigating the influences of these post features can provide actionable insights beyond those gained from focusing on user or professional factors. Demographic or professional attributes of users and professionals are often static and difficult to change within a platform's context. In contrast, content features are dynamic. Community managers can develop posting guidelines and platform functionalities to help patients craft more effective queries. Therefore, this study aims to address this research gap by investigating diverse post features and providing practical implications that empower users to effectively seek and obtain professional responses. This study conducts a microlevel analysis of post features, which complements the user-level and community-level research in the OMHC landscape. Our work offers granular and actionable explanations for knowledge contribution in OMHCs.

Previous online health community research has usually adopted structural equation modeling, multiple linear regression, and fixed effect modeling to analyze the relationships between focal factors [29,35,36]. The rise of artificial intelligence and machine learning has provided powerful tools for modeling, partitioning, and interpreting the complex relationships between factors. A few recent mental health studies have used machine learning algorithms, such as logistic regression, decision trees, and ensemble models [37-39]. Light Gradient Boosting Machine (LightGBM), an ensemble learning algorithm based on gradient boosting decision trees, is widely used for classification, regression, and ranking tasks because of its efficiency and outstanding performance, particularly in structured and tabular data problems [38]. However, LightGBM, like other ensemble models (eg, Extreme Gradient Boosting [XGBoost]), is often regarded as a "black-box" model and lacks sufficient transparency and interpretability. Explainable machine learning techniques such as Shapley Additive Explanations (SHAP) [39] can visualize the importance of factors in driving model decisions, enabling stakeholders to understand the logic behind the model's outputs and make informed decisions. Therefore, this study combines these 2 techniques to detect important post features that affect the quantity and length of professional responses in OMHCs, aiming to improve community interaction quality and enhance the effectiveness of mental health services.

Objective

The research question for this study is as follows: "In nonprofit OMHCs, how do patient posts' textual features (eg, topic, sentiment, title length, and content length) and contextual features (eg, page view and posting time) influence the response quantity and length of health professionals?" To answer this question, this paper constructs an interpretable machine learning model for analysis. This study aims to deepen the understanding of knowledge contribution in OMHCs and to inform management strategies for building more supportive and efficient mental health communities.

Methods

Overview

The research design for this study is shown in Figure 1. This framework consists of 4 main phases: data collection, thematic and sentiment analysis, predictive model construction, and predictive model interpretation. In the data collection process, we collected a total of 18,572 post-level records using a web crawler. Subsequently, we performed preliminary data cleaning to remove posts with empty titles or empty content. We then analyzed the post topics using BERTopic and removed noisy clusters. A total of 11,154 data entries were obtained after noise (topic cluster index=–1) removal. Sentiment analysis was conducted using distilled Bidirectional Encoder Representations from Transformers (DistilBERT) to determine both the category and intensity of sentiment for each post. During the modeling phase, we trained and compared 5 models (ie, LightGBM, Support Vector Machine [SVM], XGBoost, random forest [RF], ridge regression) to select the best one to develop the final model. These models were trained using textual and contextual features of the posts as predictive variables. The models were designed to predict 2 key outcome variables: response quantity and response length. Finally, we used the SHAP method to interpret the model. It provided global interpretations via summary plots to determine overall feature importance and local interpretations using dependence plots to elucidate how individual features influence the model's predictions.

XSL•FO
RenderX

Figure 1. Study flowchart. Panels 1-2 depict the data collection and processing workflow, and panels 3-4 illustrate the development of the interpretable machine learning model. LightGBM: Light Gradient Boosting Machine; MAE: mean absolute error; MSE: mean squared error; Q&A: question-and-answer; RF: random forest; SHAP: Shapley Additive Explanations; SVM: Support Vector Machine; XGBoost: Extreme Gradient Boosting.

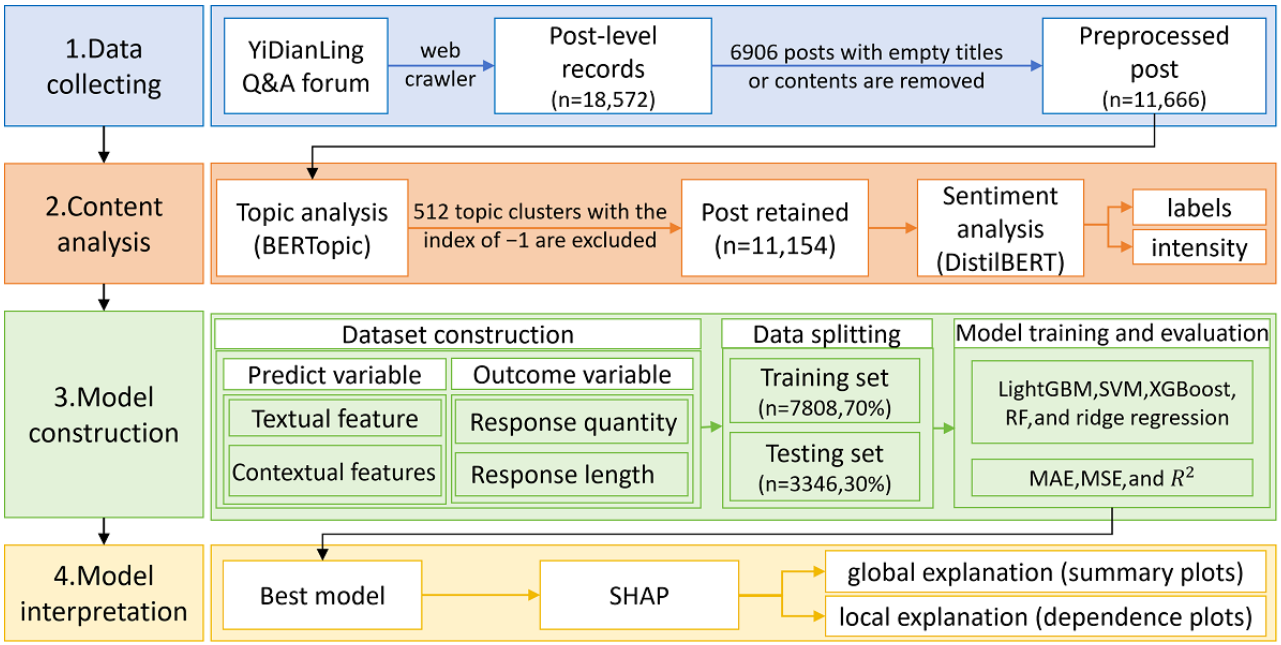

Data Collection

We collected data from the YiDianLing platform [40]. YiDianLing is a leading nonprofit mental health service provider in China, which hosts approximately 50 million registered users and 60,000 professional psychological consultants. The large and diverse user base ensures the representativeness of the data. Moreover, the platform's nationwide coverage provides broad geographic representation. YiDianLing provides a dedicated public question-and-answer forum to facilitate interaction between patient users and certified psychological consultants. In this forum, users can anonymously post their mental health concerns, and psychological consultants can provide free responses. Therefore, the forum generates rich data, including user posts, consultant responses, post view counts, and response volume.

Data from the YiDianLing, which comprised 18,572 entries from August 2024 to July 2025, were collected using a crawler program. An example of the data source page is depicted in Figure 2. Each data sample includes information about the user's post, such as the post title, post content, date of post, number of page views, number of responses, and number of responses provided by psychological consultants. The temporal evolution of the number of posts and the number of responses on the platform is presented in Multimedia Appendix 1. The volume of posts and responses exhibits a high degree of stability, with the number of replies consistently exceeding the number of posts.

XSL•FO
RenderX

Figure 2. Example of extracted features from the post.

Content Analysis

Topic Analysis

BERTopic is a topic modeling technique based on Bidirectional Encoder Representations from Transformers (BERT), which is a pretrained language model based on the transformer architecture that reads text in both directions (left-to-right and right-to-left) to understand the full context [41]. This architecture is based on multilayer neural networks called encoders, and it uses a self-attention mechanism to capture word relationships and context. BERTopic can effectively add interoperability challenges between density-focused clustering and centroid-oriented methods. This study uses the BERTopic model for thematic analysis, as it has demonstrated advantages in various topic modeling benchmark tests [42].

BERTopic facilitates consistent topic identification by leveraging a category-specific version of the term frequency-inverse document frequency (TF-IDF). In this technique, all text in a cluster is considered one entity, and TF-IDF is applied to determine the relevance scores for words within that cluster. By extracting important words in each cluster, descriptions of topics are obtained. This method is known as class-based TF-IDF:

where $f_{x,c}$ represents the frequency of word x in cluster c, f_x denotes the frequency of word x across all clusters, and A signifies the average number of words contained in each cluster.

Sentiment Analysis

In this study, we used the DistilBERT method to classify post sentiment into 5 categories: very negative, negative, neutral, positive, and very positive [43]. This model is a lightweight pretrained language model built on BERT through knowledge distillation techniques [44]. DistilBERT retains BERT's performance in capturing the sentence context and reduces the number of parameters to achieve higher computational efficiency. The used DistilBERT model has been fine-tuned on a multilingual corpus, including Chinese, and has been successfully applied to a wide range of tasks, such as product review classification, social media sentiment analysis, and customer feedback analysis [45,46].

In the sentiment analysis, we first segmented each user post into a set of sentences and applied DistillBERT analysis separately. The predicted sentiment labels were then mapped to numerical scores, with "very negative" as –1, "negative" as –0.5, "neutral" as 0, "positive" as 0.5, and "very positive" as 1. The sentiment scores of the set of sentences were then summed to obtain an overall sentiment score for the post. On the basis of this overall sentiment score, we classified each post into one of three sentiment polarities: posts with sentiment scores less than 0 were classified as negative, posts with a score equal to 0 were classified as neutral, and posts with a score greater than 0 were classified as positive. The sentiment intensity of each post was measured as the logarithm of the absolute value of the overall sentiment score plus 1.

XSL•FO
RenderX

Model Construction

LightGBM is an efficient ensemble method built on gradient boosting decision trees, which construct multiple classifiers and integrate their outputs to obtain the final prediction [47]. It is widely applied to classification, regression, and ranking tasks and can model structured and tabular data [38,48]. Unlike RFs and XGBoost, LightGBM adopts a histogram-based splitting technique that splits data and scans the statistics to determine the best split point, which enables less memory consumption and more efficient training [48].

We compared 5 popular machine learning models: LightGBM, SVM, XGBoost, RF, and ridge regression. We used a 70/30 train-test split, and the data are randomly divided 5 times to reduce the randomness introduced by data splits. The average performance of the 5 trained models is used for final model evaluation using mean absolute error, mean squared error, and R-squared. Based on these evaluations (Multimedia Appendix 2), LightGBM showed the best performance and was selected for subsequent regression prediction and interpretation.

During model training, hyperparameters were tuned on the basis of the official LightGBM documentation [49]. Specifically, the maximum number of leaves per weak learner was set to 40 to mitigate overfitting; the learning rate was set to 0.05 to accelerate convergence and improve prediction accuracy; and feature_fraction was set to 0.8, enabling the model to randomly select a subset of features when constructing each tree, thereby reducing training time. The detailed hyperparameter settings are provided in Multimedia Appendix 3.

Model Interpretation

The black-box nature of traditional machine learning models, such as ensemble methods and neural networks, limits their clinical application in the mental health domain, as stakeholders require transparent and trustworthy decision-making processes [50]. SHAP is a model interpretation method based on cooperative game theory [51]; it provides a unified measure of feature importance by attributing the model's prediction to the marginal contributions of each feature, known as SHAP values. It works as follows.

Consider the ith sample as x_i, where x_{ij} represents the jth feature of the ith sample, and y_i denotes the model's forecast for this sample. The baseline model prediction (often the predicted mean of all samples) is denoted y_{base}. The SHAP value is then derived based on the following formula:

where $f(x_{ij})$ denotes the SHAP value for x_{ij}, indicating the influence of the jth feature of the ith sample on the ultimate prediction y_i. A positive value suggests that the feature enhances the prediction, whereas a negative $f(x_{ij})$ indicates a diminishing effect on the predicted outcome.

We conducted SHAP analysis and visualized the contribution of each feature to the model's prediction using importance ranking summary plots. The dependence plots show the relationship between the changes in a feature's value and its impact on the prediction. The quantified contributions of features, either positive or negative, enhance the transparency of the prediction model.

Ethical Considerations

This study did not involve human participants. The data were publicly accessible information on the YiDianLing platform. All user posts in the question-and-answer forum were published anonymously by patient users, and no identifiable or reidentifiable personal information was collected or processed throughout the research. Therefore, there is no risk to individual privacy or foreseeable harm to users. We conducted our data collection in accordance with the platform's data authorization agreement and ensured that all procedures fully complied with the relevant ethical standards. During the handling of the dataset, we also took steps to maintain data security. This research project received formal approval from the Institutional Review Board of Shenzhen University (approval number PN-202500199).

Results

Topic Analysis and Sentiment Analysis

The results of the topic modeling visualization revealed 8 themes of user posts, as shown in Figure 3. The gray areas in Figure 3 are clusters with a category index of –1 and are considered noise and were excluded from the analysis. Representative keywords for each topic are provided in Multimedia Appendix 4.

XSL•FO
RenderX

Figure 3. Topic clustering results.

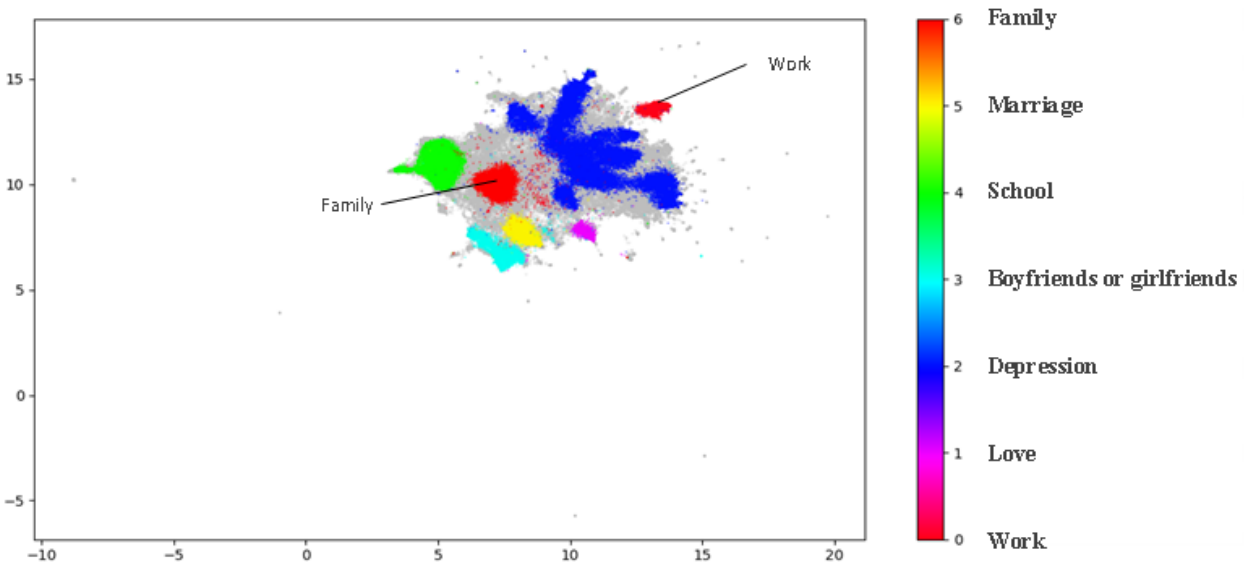

On the basis of the results of thematic clustering, we visualize the frequency of different themes over the observation period in Figure 4. The results indicate that depression consistently remained the predominant theme. Its proportion significantly exceeds that of the other categories, followed by family, school, and work. The overall topic frequency demonstrated stability throughout the period.

Figure 4. Topic frequency evolution.

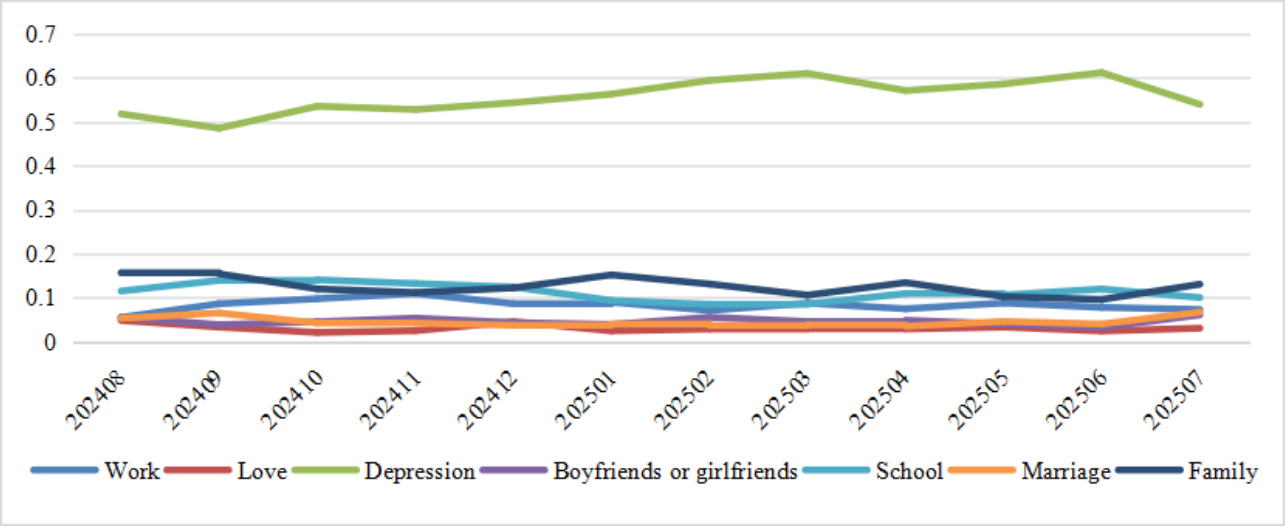

The proportions of sentiment categories at different times are depicted in Figure 5, which shows that the proportion of posts with negative sentiment is greater than those with positive and neutral sentiment. The distribution of sentiment categories remains relatively stable over time, indicating that patient users primarily express negative emotions in the mental health community.

XSL•FO
RenderX

Figure 5. Emotional map for different dates.

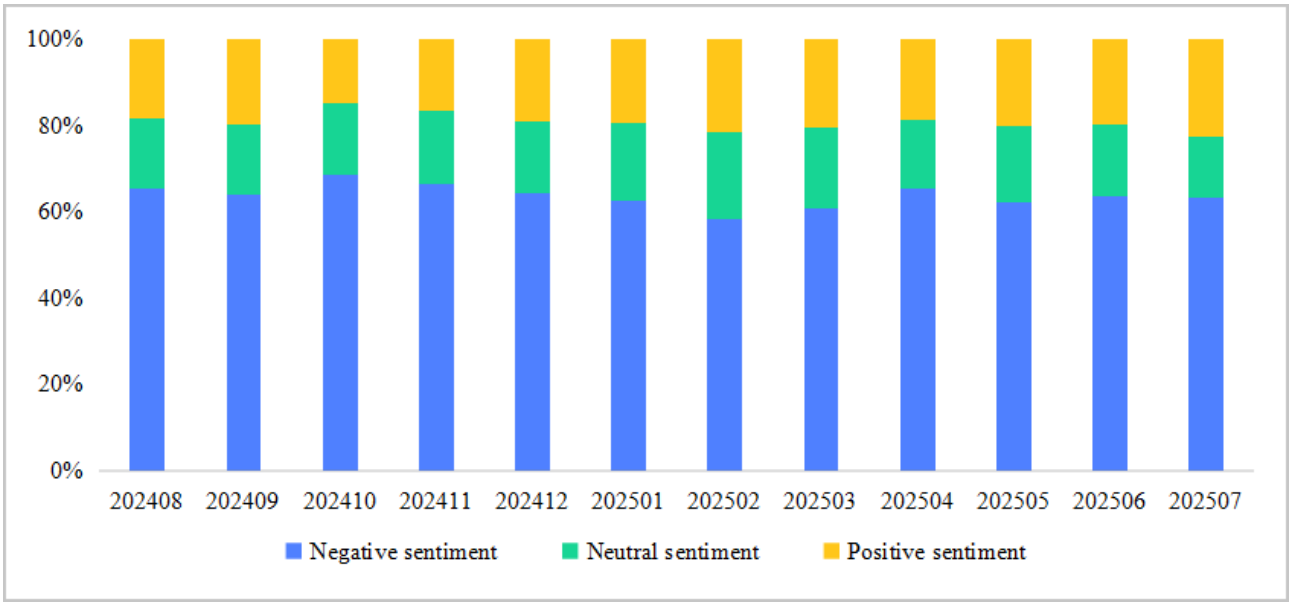

Descriptive Statistics

A total of 11,154 data entries were obtained after topic clustering and noise removal and were used for analysis. The predictor variables included post theme, sentiment category, sentiment intensity, title length, content length, posting year, posting month, posting date, public holiday status, time of day, day of the week, and page view count. The outcome variables comprise the quantity and length of replies, and prediction models are constructed separately for each. We acknowledge that a full picture of community interaction also depends on the quality of responses, a multifaceted construct that encompasses aspects like relevance, empathy, and supportiveness. Capturing this richness quantitatively poses a distinct methodological challenge. We therefore view our work as a critical first step that sets the stage for, and thereby invites, subsequent research to delve into the nuanced quality of professional contributions. A description of the features is provided in Table 1. Descriptive statistics of these selected variables are presented in Table 2, excluding the temporal characteristics of the posts.

XSL•FO
RenderX

Table 1. Description of features.

Feature	Feature description	Variable type
Topic	The topic of the content; 0=work; 1=love; 2=depression; 3=boyfriends or girlfriends; 4=school; 5=marriage; 6=family	Categorical variable
Sentiment category	0=negative; 1=neutral; 2=positive	Categorical variable
Sentiment intensity	The logarithm of the absolute value of the sentiment score plus 1	Continuous variable
Page view	The logarithm of the number of page views posted on a given date	Continuous variable
Year	Year of the post	Categorical variable
Month	Month of the post	Categorical variable
Day	Day of the post	Categorical variable
Title length	The natural logarithm (base e) of the number of Chinese characters in the post title (note: raw character counts are used for result interpretation)	Continuous variable
Content length	The natural logarithm (base e) of the number of Chinese characters in the post content (note: raw character counts are used for result interpretation)	Continuous variable
Hour	0=00:00~00:59, 1=01:00~01:59, 2=02:00~02:59, 3=03:00~03:59, 4=04:00~04:59, 5=05:00~05:59, 6=06:00~06:59, 7=07:00~07:59, 8=08:00~08:59, 9=09:00~09:59, 10=10:00~10:59, 11=11:00~11:59, 12=12:00~12:59, 13=13:00~13:59, 14=14:00~14:59, 15=15:00~15:59, 16=16:00~16:59, 17=17:00~17:59, 18=18:00~18:59, 19=19:00~19:59, 20=20:00~20:59, 21=21:00~21:59, 22=22:00~22:59, 23=23:00~23:59	Categorical variable
Week	0=Monday; 1=Tuesday; 2=Wednesday; 3=Thursday, 4=Friday; 5=Saturday; 6=Sunday	Categorical variable
Holiday	0=no; 1=yes	Categorical variable
Reply quantity	The logarithm of the number of replies to the post	Continuous variable
Reply length	The logarithm of the average reply length of the post	Continuous variable

Table 2. Descriptive statistics.

Features	Value (n=11,154)
Page view, mean (SD)	5.09 (0.67)
Sentiment category, n (%)	
Positive	2117 (19)
Neutral	1904 (17.1)
Negative	7133 (63.9)
Sentiment intensity, mean (SD)	0.69 (0.48)
Topic, n (%)	
Work	917 (8.2)
Love	355 (3.2)
Depression	6139 (55)
Boyfriends or girlfriends	519 (4.7)
School	1269 (11.4)
Marriage	501 (4.5)
Family	1454 (13)
Title length, mean (SD)	2.92 (0.40)
Content length, mean (SD)	4.29 (1.30)
Reply quantity, mean (SD)	0.22 (0.66)
Reply length, mean (SD)	5.25 (0.83)

XSL•FO
RenderX

Comparisons of Model Performance

We evaluated 5 machine learning models: LightGBM, SVM, XGBoost, RF, and ridge regression for predicting response quantity and response length. Tables 3 and 4 present the performance of these models. Results show that LightGBM achieved the lowest mean absolute error (mean 0.2859, SD 0.0072) and mean squared error (mean 0.3100, SD 0.0142), along with the highest R^2 (mean 0.2754, SD 0.0323), statistically outperforming SVM, XGBoost, RF, and ridge regression. Similarly, for response length prediction (Table 4), LightGBM demonstrated superior overall performance. These metrics provide a comprehensive assessment of each model's capabilities. Additionally, calibration curves for the LightGBM model in both prediction tasks are provided in Multimedia Appendix 2, both of which indicate a good model fit.

Table 3. Response quantity prediction performance of the compared models.

Model	MAE[a], mean (SD)	MSE[b], mean (SD)	R^2, mean (SD)
LightGBM[c]	0.2859 (0.0072)	0.3100 (0.0142)	0.2754 (0.0323)
SVM[d]	0.3101 (0.0161)	0.3931 (0.0391)	0.0855 (0.0187)
XGBoost[e]	0.3155 (0.0061)	0.3223 (0.0190)	0.2476 (0.0182)
RF[f]	0.2962 (0.0065)	0.3116 (0.0146)	0.2717 (0.0302)
Ridge regression	0.3693 (0.0049)	0.3516 (0.0195)	0.1789 (0.0221)

Table 4. Response length prediction performance of the compared models.

Model	MAE[a], mean (SD)	MSE[b], mean (SD)	R^2, mean (SD)
LightGBM[c]	0.6988 (0.0079)	0.8367 (0.0155)	0.2766 (0.0221)
SVM[d]	0.6941 (0.0054)	0.8905 (0.0149)	0.2302 (0.0145)
XGBoost[e]	0.7219 (0.0072)	0.8688 (0.0168)	0.2490 (0.0129)
RF[f]	0.7058 (0.0069)	0.8637 (0.0146)	0.2532 (0.0218)
Ridge regression	0.7171 (0.0070)	0.8804 (0.0161)	0.2390 (0.0154)

[a]MAE: mean absolute error.
[b]MSE: mean squared error.
[c]LightGBM: Light Gradient Boosting Machine.
[d]SVM: Support Vector Machine.
[e]XGBoost: Extreme Gradient Boosting.
[f]RF: random forest.

Model Interpretability

Global Interpretability

We applied the SHAP method to the LightGBM model. The global interpretation graphs of LightGBM for predicting the number of replies and the length of replies are shown in Figure 6. The average SHAP value for each feature is detailed in Multimedia Appendix 5. When a SHAP value of 0 is used as the dividing line, the points on the left indicate the features contributing negatively to the prediction, whereas the points on the right indicate positive contributions. The relationship between each feature and the prediction of the number of replies is shown in Figure 6A. This indicates that positively correlated features include page views, content length, and sentiment intensity. Higher values of these features correspond to a greater number of responses received by the posts. The relationship between each feature and the prediction of response length is shown in Figure 6B. These findings indicate that the length of the question content has a positive effect on response length. Most other features are categorical variables, whose effects are not clearly discernible from the figure.

XSL•FO
RenderX

Figure 6. Summary plots of Light Gradient Boosting Machine. (A) Prediction for response quantity. (B) Prediction for response length. SHAP: Shapley Additive Explanations.

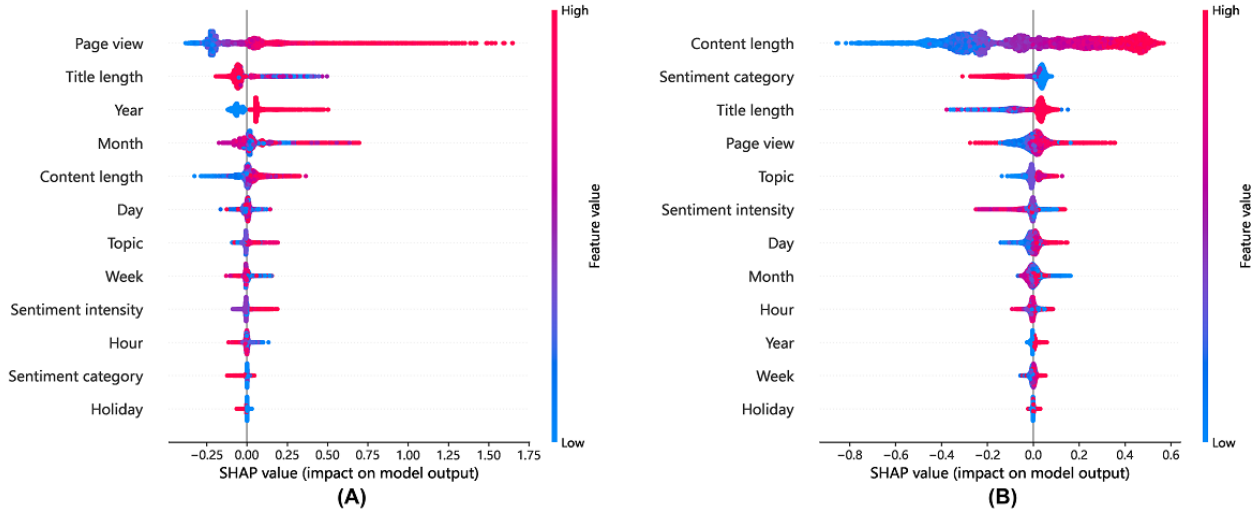

Local Interpretability in Response Quantity Prediction

We constructed feature-SHAP value scatter plots for each feature to analyze its impact on the response quantity (Figure 7). Each dot in the scatter plot represents a single post in our dataset. They illustrate the relationship between feature values (x-axis) and their corresponding SHAP values (y-axis). The SHAP value is a direct measure of how much that specific feature value pushed the model's prediction toward receiving more (positive SHAP value) or fewer (negative SHAP value) replies. These plots answer a critical question: how a specific post characteristic influences a consultant's likelihood to reply, and whether this influence is consistently positive, negative, or more complex?

XSL•FO
RenderX

Figure 7. Shapley Additive Explanations (SHAP) dependence plots of Light Gradient Boosting Machine for predicting response quantity. Panels A-L respectively show the SHAP value distributions for the following features: topic, sentiment category, sentiment intensity, page views, title length, content length, year, month, day, hour, week, and holiday.

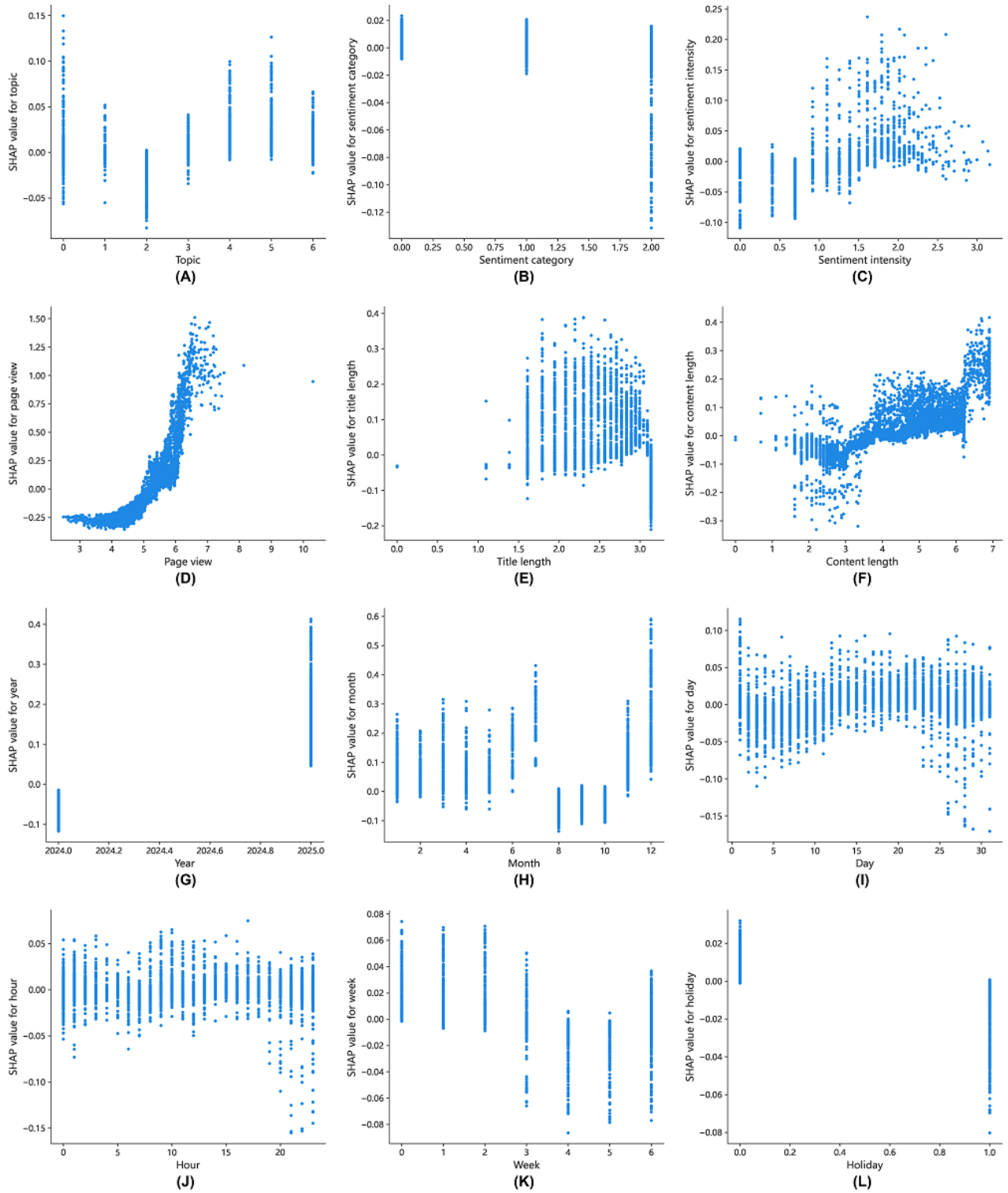

The impact of topic features on response quantity is illustrated in Figure 7A. Topic 2 (depression) has a SHAP value less than 0, whereas topics 3-6 (3: boyfriends or girlfriends; 4: school; 5: marriage; and 6: family) have SHAP values greater than 0. These findings suggest that the topic of depression has a negative effect on response quantity, whereas those related to boyfriends or girlfriends, school, marriage, and family have positive effects. The impact of other topics on response quantity is not clearly defined.

The relationships between post sentiment and response quantity are shown in Figures 7B and 7C. The SHAP values for negative sentiment are greater than 0, indicating a positive contribution to response quantity. In contrast, the SHAP values for positive sentiment are less than 0, indicating a negative contribution. Moreover, when the sentiment intensity exceeds the threshold

XSL•FO
RenderX

of 1.7 (equivalent to a sentiment intensity of 5.5), the post sentiment contributes positively to the response quantity. These findings suggest that mental health professionals are more likely to respond to posts that express negative emotions and provide support to high-risk patient users.

The effects of other features on the prediction of response quantity are presented in Figures 7D-7L. The results indicate that when page views exceed the threshold of 5.5 (approximately 244 views), the title length is between 2.75 and 2.95 (approximately 15 to 20 characters), and the content length exceeds the threshold of 4.85 (approximately 127 characters), the posts are more likely to receive professional responses. With respect to the posting time, when the post is published between December and February, in the middle of the month (days 18 to 21), during the period between Monday and Thursday, and on nonpublic holidays, it is more likely to receive a professional response. Moreover, when the page view is less than 5.5 (approximately 244 views), the title length exceeds 2.95 (approximately 20 characters), or when the post is published from August to October, on weekends, or on public holidays, the SHAP values are less than 0, indicating an inhibitory effect on the prediction of response quantity. Additionally, the posting time (hour of day) does not have a significant effect.

Local Interpretability in Response Length Prediction

We report SHAP dependence plots for each feature to explain the response length predicted by the LightGBM model (Figure 8). According to Figure 8A, the topic of love has a negative effect on the prediction of response length. The results from Figures 8B and 8C show that negative sentiment has a positive effect on predicting response length, whereas positive emotions have an inhibitory effect. When the sentiment intensity is less than 1.5 (equivalent to the original sentiment intensity of 4.5), it negatively affects the prediction of the reply length. As shown in Figure 8E, when the length of the post's title is less than the threshold of 2.0 (approximately 7 characters), the predicted response length decreases. As shown in Figure 8F, the critical value for content length is 4.1 (approximately 60 characters), and posts with lengths exceeding 60 characters are more likely to receive longer responses.

XSL•FO
RenderX

Figure 8. Shapley Additive Explanations (SHAP) dependence plots of Light Gradient Boosting Machine for predicting response length. Panels A-L respectively show the SHAP value distributions for features: topic, sentiment category, sentiment intensity, page views, title length, content length, year, month, day, hour, week, and holiday.

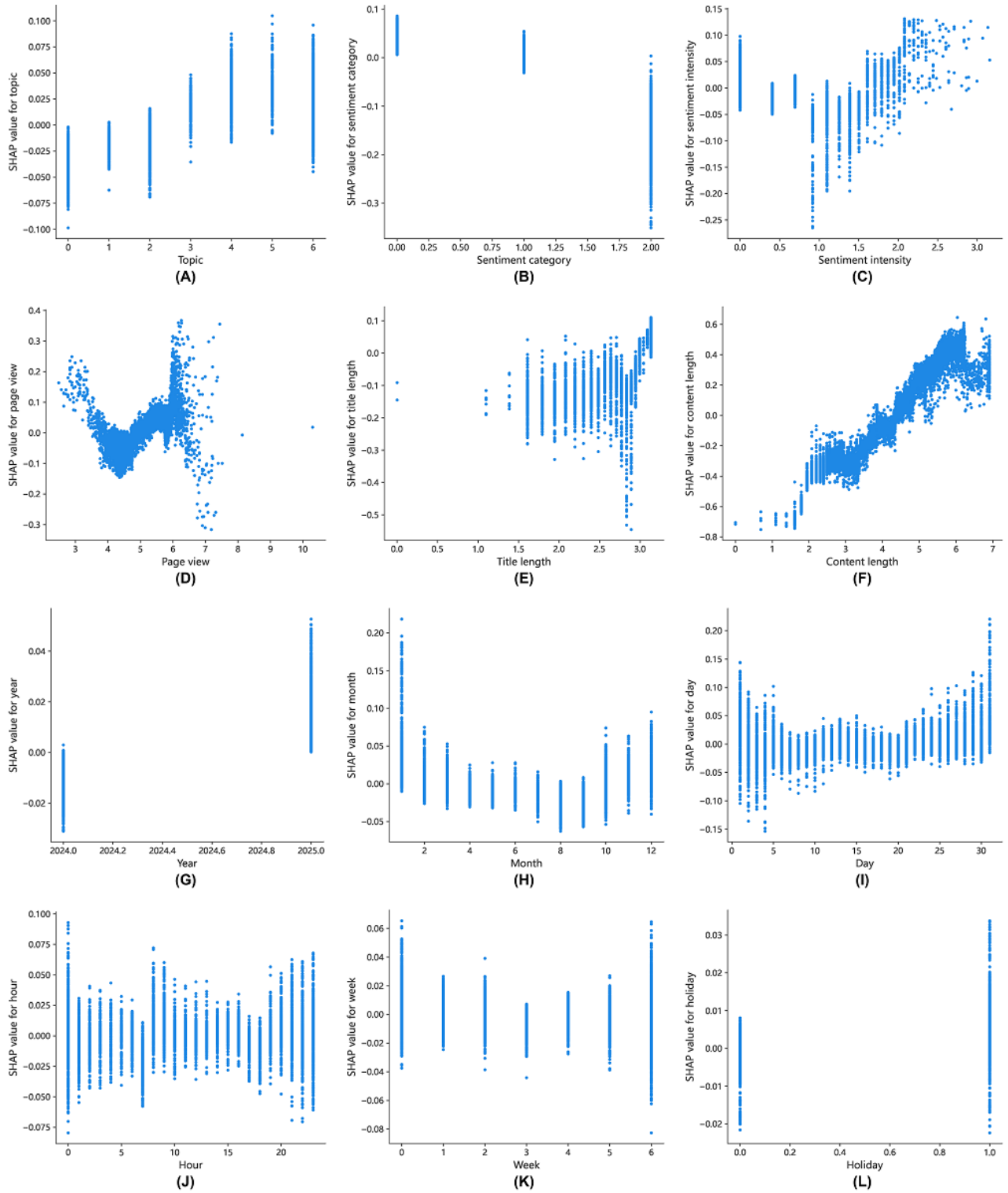

Robustness Analysis

In the robustness analysis, we collected and analyzed forum data from another online mental health platform, YiXinli [52], covering the period from December 2024 to July 2025, with the aim of examining the generalizability of the model interpretations through cross-platform comparison. All the analytical procedures were kept consistent with those applied to the primary dataset. Multimedia Appendix 6 reports the descriptive statistics, topic and sentiment evolution results, model evaluation results, and SHAP-based interpretation outcomes. The findings indicate that the topic distributions, sentiment distributions, and their temporal trends are consistent with those observed on the YiDianLing platform. Moreover, the SHAP interpretations of the LightGBM models for predicting the quantity and length of responses indicate that the

XSL•FO
RenderX

effects of key features are relatively consistent with those observed in the primary dataset. These results provide robust support for our findings.

Discussion

Principal Findings

This study uses interpretable machine learning techniques to analyze question-and-answer posts from a nonprofit OMHC. The findings highlight the pivotal role of various features of the posts in shaping professionals' contribution behaviors. First, patient users' demands for psychological services primarily fall into 7 topic categories: work, love, depression, boyfriends or girlfriends, school, marriage, and family. The majority of user posts are related to depression, which aligns with previous research [14,53]. Users also frequently express concerns related to daily life, work stress, and social relationships [54]. Posts with greater response volumes are often associated with themes such as boyfriends or girlfriends, school, marriage, and family, whereas depression-related posts receive fewer replies. One possible explanation is that depression-related posts are more likely to be posted by diagnosed patients. The forum's professionals are composed mainly of psychological consultants who are not therapists capable of providing clinical treatment. Therefore, they tend to be more cautious when addressing depression-related issues, thus avoiding the risk of inadvertently harming these patient users. This aligns with prior findings that, in the absence of sufficient knowledge about the individual, consultants may avoid overreacting to questions involving illness and emotions [55].

Second, our study revealed that patient posts with stronger negative emotions are more likely to receive social support in nonprofit mental health communities. From the perspective of social support theory, emotional disclosure by patient users in online communities is crucial for fostering social interaction and obtaining support. Expressions of negative emotions may signal users' distress, which may elicit empathy and supportive reactions from psychological counselors [56]. On the other hand, emotional intensity can be viewed as an indicator of patients' level of self-disclosure [57,58]. Intense self-disclosure may enhance users' motivation to engage [59]. Our analysis shows, for example, that when sentiment intensity is relatively high (eg, above 5.5), posts are more likely to attract longer replies. In contrast, when sentiment intensity is relatively low (eg, less than 4.5), posts tend to receive shorter responses. Drawing on these findings, community managers could provide patient users with emotion-related keywords or tags, which can be automatically selected when writing posts. These emotion-related keywords or tags may enhance users' ability to articulate their emotional states.

Third, our findings indicate that the title length and content length of patient users' posts positively influence the quantity and length of professionals' responses. Longer titles and content may contain greater amounts of information, facilitating better understanding by other community members. Therefore, the amount of information provided in posts positively influences the quality of replies [24]. Our analysis revealed that titles with lengths between 15 and 20 characters and contents with lengths of at least 60 characters attracted more and longer responses. In contrast, titles shorter than approximately 7 characters tend to negatively impact the response length. On the basis of these findings, platform designers may provide real-time feedback on the informativeness of titles and content to post writers. Additionally, the platform can offer high-quality example posts for users to enhance the expressiveness of their posts, thus increasing their likelihood of receiving a response.

In our study, post features, including view count, posting time, day of the week, and public holiday status, influence professional responses heterogeneously. Posts with more than 244 views are associated with greater response volumes. This suggests that posts with greater exposure may attract increased professional participation. Posts generated from Monday to Thursday and on nonpublic holidays receive more responses. Because most of the certified counselors on the platform are not full-time clinical physicians working in offline institutions, they tend to be more active and willing to respond online during weekdays. Our results also show that page views and posting time do not affect response length. One possible explanation is that response length may depend on the professionalism, empathy, and motivation of health professionals rather than on the view number or posting time.

Research Implications

Theoretical Implications

This study deepens our understanding of knowledge contribution behaviors in nonprofit OMHCs. By examining how textual and contextual features of user posts influence the quantity and length of responses from mental health professionals, the findings reveal the important factors that shape health professionals' knowledge contributions. In addition, this research introduces interpretable machine learning methods into online mental health. This approach addresses the limitations of traditional regression models and black-box algorithms in explaining the influencing mechanisms. It also provides technical support for a deeper understanding of the factors affecting professional response quantity and length.

Practical Implications

The findings provide practical guidance for community managers. First, managers can categorize post topics to facilitate precise responses from professionals and help other users explore topics of interest. Second, providing predefined emotion-related keywords or tags on the post editing page helps enhance patients' ability to express their emotions and may increase the likelihood of receiving a response. Third, providing feedback on information richness and high-quality post templates may help users improve their expression. Fourth, considering the effects of view counts and posting times, platform operators can optimize content visibility strategies. For example, posts with lower view counts and those published during off-peak hours (eg, late night) can be prioritized. This may balance the exposure across posts of varying popularity and publication times and ensure that these posts receive professional responses. This may also enhance the overall fairness and quality of community interactions. Fifth, platform operators must exercise caution when implementing certain

XSL•FO
RenderX

strategies (eg, real-time feedback on content informativeness), as encouraging longer posts or greater emotional intensity may inadvertently increase the potential psychological burden on patients. To this end, when designing web interfaces, platforms should position these tools as supplementary and optional. It is crucial to ensure users retain control over these functionalities to balance interaction efficiency with psychological safety.

This study, based on Chinese OMHCs, may offer insights for mental health service platforms in other countries. However, we explicitly caution against the direct generalization of our findings to other cultural contexts. The counselor-led, nonprofit forums of the YiDianLing and YiXinLi platforms embed specific sociotechnical norms, such as text-based communication, work culture, and community governance rules. These factors may shape how patient users present their problems and how consultants perceive their role. Therefore, our conclusions should be interpreted as context-specific insights that highlight the need for future research to "unpack" these contextual differences through comparative studies.

In developing our model, we carefully addressed potential sources of bias, such as by using cross-temporal and cross-platform sampling methods to balance both positive and negative scenarios in predictions. When this research is extended to other countries, adjustments need to be made according to local cultural norms and service systems.

Limitations and Future Directions

Although our study offers significant contributions, it also has several limitations. First, the data were obtained primarily from the Chinese mental health platform YiDianLing, with supplementary data from YiXinLi used for robustness checks. Given the potential cultural, platform, and user differences across countries, generalizing our findings would require validation using multinational data from diverse platforms. Second, this study uses publicly available posts and response data from OMHCs. Future research could incorporate multimodal data (eg, images and emojis) to gain a deeper understanding of the interactions between patient users and mental health professionals in nonprofit mental health forums. Third, owing to privacy policies, user-level demographics (eg, sex, age, and membership duration) were unavailable. Subsequent studies or online experiments should examine how such characteristics influence forum participation. Fourth, this study solely explores response quantity and length. However, a complete understanding of community interaction also depends on the quality of responses, a multifaceted construct that encompasses aspects like relevance, empathy, and supportiveness. Future work should prioritize developing validated metrics for response quality to better evaluate professional contribution patterns. Furthermore, this study relies on a LightGBM model interpreted with the SHAP method to analyze post feature importance. Given the limited sample size and post-level features, the predictive accuracy may be limited. Future research should thus incorporate richer predictive features and additional interpretable machine learning techniques (eg, Local Interpretable Model-agnostic Explanations) to validate and extend these insights.

Conclusion

This study uses explainable machine learning methods to investigate the post features that influence response quantity and length in OMHCs. It highlights the importance of the post topic, post title, post length, post sentiment, and posting time. These findings provide insights for platform managers in terms of optimizing functional design and improving the effectiveness of community interactions.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (72471154), the General Project of the Ministry of Education Foundation on Humanities and Social Sciences (24YJC630049), and Shenzhen Science and Technology Innovation Commission (JCYJ20240813143012016).

Data Availability

The data used in this study are available upon reasonable request from the corresponding author.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Temporal dynamics of posts and responses on the YiDianLing platform.
[DOCX File , 17 KB - jmir_v28i1e74359_app1.docx]

Multimedia Appendix 2
Model performance evaluation.
[DOCX File , 248 KB - jmir_v28i1e74359_app2.docx]

Multimedia Appendix 3

XSL•FO
RenderX

The hyperparameter settings for the Shapley Additive Explanations algorithm.
[DOCX File , 12 KB - jmir_v28i1e74359_app3.docx]

Multimedia Appendix 4
Representative keywords for each topic identified by BERTopic.
[DOCX File , 14 KB - jmir_v28i1e74359_app4.docx]

Multimedia Appendix 5
Average absolute Shapley Additive Explanations values of features for predicting response quantity and length.
[DOCX File , 15 KB - jmir_v28i1e74359_app5.docx]

Multimedia Appendix 6
Analysis results for the YiXinli platform.
[DOCX File , 1433 KB - jmir_v28i1e74359_app6.docx]

References

1. De Hert M, Detraux J, Vancampfort D. The intriguing relationship between coronary heart disease and mental disorders. Dialogues Clin Neurosci 2018;20(1):31-40. [doi: 10.31887/DCNS.2018.20.1/mdehert] [Medline: 29946209]
2. Beurel E, Toups M, Nemeroff CB. The bidirectional relationship of depression and inflammation: double trouble. Neuron 2020;107(2):234-256 [FREE Full text] [doi: 10.1016/j.neuron.2020.06.002] [Medline: 32553197]
3. Ohrnberger J, Fichera E, Sutton M. The relationship between physical and mental health: a mediation analysis. Soc Sci Med 2017;195:42-49 [FREE Full text] [doi: 10.1016/j.socscimed.2017.11.008] [Medline: 29132081]
4. Drissi N, Ouhbi S, Janati Idrissi MA, Fernandez-Luque L, Ghogho M. Connected mental health: systematic mapping study. J Med Internet Res 2020;22(8):e19950 [FREE Full text] [doi: 10.2196/19950] [Medline: 32857055]
5. Schuch FB, Vancampfort D. Physical activity, exercise, and mental disorders: it is time to move on. Trends Psychiatry Psychother 2021;43(3):177-184 [FREE Full text] [doi: 10.47626/2237-6089-2021-0237] [Medline: 33890431]
6. Walker ER, McGee RE, Druss BG. Mortality in mental disorders and global disease burden implications: a systematic review and meta-analysis. JAMA Psychiatry 2015;72(4):334-341 [FREE Full text] [doi: 10.1001/jamapsychiatry.2014.2502] [Medline: 25671328]
7. Ali K, Farrer L, Gulliver A, Griffiths KM. Online peer-to-peer support for young people with mental health problems: a systematic review. JMIR Ment Health 2015;2(2):e19 [FREE Full text] [doi: 10.2196/mental.4418] [Medline: 26543923]
8. Li Pira G, Aquilini B, Davoli A, Grandi S, Ruini C. The use of virtual reality interventions to promote positive mental health: systematic literature review. JMIR Ment Health 2023;10:e44998 [FREE Full text] [doi: 10.2196/44998] [Medline: 37410520]
9. Tzeng Y, Yin W, Lin K, Wei J, Liou H, Sung H, et al. Factors associated with the utilization of outpatient virtual clinics: retrospective observational study using multilevel analysis. J Med Internet Res 2022;24(8):e40288 [FREE Full text] [doi: 10.2196/40288] [Medline: 35917486]
10. Oexle N, Ajdacic-Gross V, Kilian R, Müller M, Rodgers S, Xu Z, et al. Mental illness stigma, secrecy and suicidal ideation. Epidemiol Psychiatr Sci 2017;26(1):53-60 [FREE Full text] [doi: 10.1017/S2045796015001018] [Medline: 26606884]
11. Banwell E, Hanley T, De Ossorno Garcia S, Mindel C, Kayll T, Sefi A. The helpfulness of web-based mental health and well-being forums for providing peer support for young people: cross-sectional exploration. JMIR Form Res 2022;6(9):e36432 [FREE Full text] [doi: 10.2196/36432] [Medline: 36083629]
12. Liu J, Gao L. Lurking or active? The influence of user participation behavior in online mental health communities on the choice and evaluation of doctors. J Affect Disord 2022;301:454-462. [doi: 10.1016/j.jad.2022.01.074] [Medline: 35066007]
13. Park A, Conway M, Chen AT. Examining thematic similarity, difference, and membership in three online mental health communities from reddit: a text mining and visualization approach. Comput Human Behav 2018;78:98-112 [FREE Full text] [doi: 10.1016/j.chb.2017.09.001] [Medline: 29456286]
14. Feldhege J, Moessner M, Bauer S. Who says what? Content and participation characteristics in an online depression community. J Affect Disord 2020;263:521-527. [doi: 10.1016/j.jad.2019.11.007] [Medline: 31780138]
15. Saha B, Nguyen T, Phung D, Venkatesh S. A framework for classifying online mental health-related communities with an interest in depression. IEEE J Biomed Health Inform 2016;20(4):1008-1015. [doi: 10.1109/JBHI.2016.2543741] [Medline: 27008680]
16. Grub MF. Reddit as a "Safe Space": topic modeling of online mental health communities for depression and anxiety. Weizenbaum J Digit Soc 2025;5(3). [doi: 10.34669/wi.wjds/5.3.3]
17. AbouWarda H, Miscione G. Understanding how discourse themes in an online mental health community on twitter/x drive varied population-specific empowerment processes in alignment with global standards: a qualitative analysis of #bipolarclub. J Med Internet Res 2025;27:e74912 [FREE Full text] [doi: 10.2196/74912] [Medline: 40882197]

XSL•FO
RenderX

18. Zhou J, Zuo M, Ye C. Understanding the factors influencing health professionals' online voluntary behaviors: evidence from YiXinLi, a Chinese online health community for mental health. Int J Med Inform 2019;130:103939. [doi: 10.1016/j.ijmedinf.2019.07.018] [Medline: 31434043]
19. Kim M, Saha K, De Choudhury M, Choi D. Supporters First: understanding online social support on mental health from a supporter perspective. Proc ACM Hum-Comput Interact 2023;7(CSCW1):1-28 [FREE Full text] [doi: 10.1145/3579525]
20. Liu S, Xiao W, Fang C, Zhang X, Lin J. Social support, belongingness, and value co-creation behaviors in online health communities. Telemat Inform 2020;50:101398. [doi: 10.1016/j.tele.2020.101398]
21. Magane KM, Kenney M, Nelson E, Wisk L, Weitzman ER. The quality and safety of online health communities engaging adolescents around depression and substance use: a multisite evaluation. J Adolesc Health 2017;60(2):S77. [doi: 10.1016/j.jadohealth.2016.10.334]
22. Marshall P, Booth M, Coole M, Fothergill L, Glossop Z, Haines J, et al. Understanding the impacts of online mental health peer support forums: realist synthesis. JMIR Ment Health 2024;11:e55750 [FREE Full text] [doi: 10.2196/55750] [Medline: 38722680]
23. Morini V, Sansoni M, Rossetti G, Pedreschi D, Castillo C. Participant behavior and community response in online mental health communities: insights from reddit. Computers in Human Behavior 2025;165:108544. [doi: 10.1016/j.chb.2024.108544]
24. Li J, Liu D, Wan C, Liang Z, Zhu T. Empirical study of factors that influence the perceived usefulness of online mental health community members. Psych J 2023;12(2):307-318. [doi: 10.1002/pchj.629] [Medline: 36726193]
25. Smith-Merry J, Goggin G, Campbell A, McKenzie K, Ridout B, Baylosis C. Social connection and online engagement: insights from interviews with users of a mental health online forum. JMIR Ment Health 2019;6(3):e11084 [FREE Full text] [doi: 10.2196/11084] [Medline: 30912760]
26. Sharma E, De Choudhury CM. Mental health support and its relationship to linguistic accommodation in online communities. 2018 Presented at: CHI '18: CHI Conference on Human Factors in Computing Systems; 2018 April 21-26; Montreal QC Canada p. 1-13. [doi: 10.1145/3173574.3174215]
27. Srivastava A, Gupta T, Cerezo A, Lord SP, Akhtar MS, Chakraborty T. Critical behavioral traits foster peer engagement in online mental health communities. PLoS One 2025;20(1):e0316906 [FREE Full text] [doi: 10.1371/journal.pone.0316906] [Medline: 39804944]
28. Chen L, Baird A, Straub D. A linguistic signaling model of social support exchange in online health communities. Decis Support Syst 2020;130:113233. [doi: 10.1016/j.dss.2019.113233]
29. Wang J, Chiu Y, Yu H, Hsu Y. Understanding a nonlinear causal relationship between rewards and physicians' contributions in online health care communities: longitudinal study. J Med Internet Res 2017;19(12):e427 [FREE Full text] [doi: 10.2196/jmir.9082] [Medline: 29269344]
30. Zhou T. Examining online health community users' sharing behaviour: a social influence perspective. Inf Dev 2021;38(4):599-608. [doi: 10.1177/02666669211007188]
31. Chen Q, Jin J, Yan X. Understanding physicians' motivations for community participation and content contribution in online health communities. OIR 2022;47(3):604-629. [doi: 10.1108/oir-11-2021-0615]
32. Imlawi J, Gregg D. Understanding the satisfaction and continuance intention of knowledge contribution by health professionals in online health communities. Inform Health Soc Care 2020;45(2):151-167. [doi: 10.1080/17538157.2019.1625053] [Medline: 31328593]
33. Maheshwari B, Sarrion M, Motiani M, O'Sullivan S, Chandwani R. Exploration of factors affecting the use of Web 2.0 for knowledge sharing among healthcare professionals: an Indian perspective. JKM 2020;25(3):545-558 [FREE Full text] [doi: 10.1108/jkm-02-2020-0105]
34. Derks D, Fischer AH, Bos AER. The role of emotion in computer-mediated communication: a review. Comput Hum Behav 2008;24(3):766-785 [FREE Full text] [doi: 10.1016/j.chb.2007.04.004]
35. Zhuo X, Wang W. Why are physicians willing to contribute knowledge? Evidence from online health communities. Comput Hum Behav 2024;152:108095. [doi: 10.1016/j.chb.2023.108095]
36. Feng X, Hu Y, Pfaff H, Liu S, Xie J, Zhang Z. Exploring client preferences for psychological counselors in a chinese online health community: longitudinal study. J Med Internet Res 2024;26:e58428 [FREE Full text] [doi: 10.2196/58428] [Medline: 39388694]
37. Liu H, Zhang L, Wang W, Huang Y, Li S, Ren Z, et al. Prediction of online psychological help-seeking behavior during the COVID-19 pandemic: an interpretable machine learning method. Front Public Health 2022;10:814366 [FREE Full text] [doi: 10.3389/fpubh.2022.814366] [Medline: 35309216]
38. Baba A, Bunji K. Prediction of mental health problem using annual student health survey: machine learning approach. JMIR Ment Health 2023;10:e42420 [FREE Full text] [doi: 10.2196/42420] [Medline: 37163323]
39. Huang Y, Liu H, Chi M, Meng S, Wang W. Digit Health 2025;11:20552076251333480 [FREE Full text] [doi: 10.1177/20552076251333480] [Medline: 40343065]
40. YiDianLing. 2025. URL: https://www.ydl.com/[accessed [accessed 2025-08-19]
41. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational

XSL•FO
RenderX

Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); 2025 December 02; Minneapolis, Minnesota p. 4171-4186. [doi: 10.18653/v1/N19-1423]
42. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv Preprint posted online on March 11, 2022 [FREE Full text] [doi: 10.48550/arXiv.2203.05794]
43. DistilBERT-based multilingual sentiment classification model. Hugging Face. 2025. URL: https://huggingface.co/tabularisai/multilingual-sentiment-analysis [accessed 2025-08-01]
44. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv Preprint posted online on October 2, 2019 [FREE Full text]
45. AlQadi RA, Taie SA, Idrees AM, Elhariri E. Explainable deep learning model for ChatGPT-rephrased fake review detection using DistilBERT. BDCC 2025;9(8):205. [doi: 10.3390/bdcc9080205]
46. Jojoa M, Eftekhar P, Nowrouzi-Kia B, Garcia-Zapirain B. Natural language processing analysis applied to COVID-19 open-text opinions using a distilBERT model for sentiment categorization. AI Soc 2022;39:1-8 [FREE Full text] [doi: 10.1007/s00146-022-01594-w] [Medline: 36439363]
47. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. 2017 Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 December 4 - 9; Long Beach California USA p. 3149-3157.
48. Tai C, Liao T, Chen S, Chung M. Sleep stage classification using light gradient boost machine: exploring feature impact in depressive and healthy participants. Biomed Sigal Process Control 2024;88:105647. [doi: 10.1016/j.bspc.2023.105647]
49. LightGBM parameters tuning. Read the Docs. 2025. URL: https://lightgbm.readthedocs.io/en/latest/Parameters-Tuning.html [accessed 2025-12-04]
50. Le Glaz A, Haralambous Y, Kim-Dufor D, Lenca P, Billot R, Ryan TC, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res 2021;23(5):e15708 [FREE Full text] [doi: 10.2196/15708] [Medline: 33944788]
51. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2(1):56-67 [FREE Full text] [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
52. YiXinLi. URL: https://www.xinli001.com/[accessed [accessed 2025-08-19]
53. González Moreno A, Molero Jurado MDM. Presence of emotions in network discourse on mental health: thematic analysis. Psychiatry Int 2024;5(3):348-359. [doi: 10.3390/psychiatryint5030024]
54. Akar E. Connecting for well-being: a role-based network analysis of online mental health communities. Behav Inf Technol 2025;44(18):4566-4580. [doi: 10.1080/0144929x.2025.2484393]
55. Zhang L, Liu D, Li J, Wan C, Liu X. Exploring linguistic features and user engagement in Chinese online mental health counseling. Heliyon 2024;10(19):e38042. [doi: 10.1016/j.heliyon.2024.e38042] [Medline: 39678785]
56. De Choudhury M, De S. Mental health discourse on reddit: self-disclosure, social support, and anonymity. ICWSM 2014;8(1):71-80. [doi: 10.1609/icwsm.v8i1.14526]
57. Chu TH, Sun M, Crystal Jiang L. Self-disclosure in social media and psychological well-being: A meta-analysis. J Soc Pers Relatsh 2022;40(2):576-599. [doi: 10.1177/02654075221119429]
58. Liu J, Kong J. Why do users of online mental health communities get likes and reposts: a combination of text mining and empirical analysis. Healthcare (Basel) 2021;9(9):1133 [FREE Full text] [doi: 10.3390/healthcare9091133] [Medline: 34574907]
59. Liu J, Liu Y. Exploring the user interaction network in an anxiety disorder online community: an exponential random graph model with topical and emotional effects. Int J Environ Res Public Health 2022;19(11):6354 [FREE Full text] [doi: 10.3390/ijerph19116354] [Medline: 35681939]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
DistilBERT: distilled Bidirectional Encoder Representations from Transformers
LightGBM: Light Gradient Boosting Machine
OMHC: online mental health community
RF: random forest
SHAP: Shapley Additive Explanations
SVM: Support Vector Machine
TF-IDF: term frequency-inverse document frequency
XGBoost: Extreme Gradient Boosting

XSL•FO
RenderX

Edited by J Sarvestan; submitted 23.Mar.2025; peer-reviewed by S Narayan, S Kath; comments to author 07.Apr.2025; accepted 24.Nov.2025; published 05.Jan.2026.

Please cite as:
Geng S, Li Y, Wang J, Chen P, Wu X, Zhang Z
Predictors of Professional Responses in Nonprofit Mental Health Forums: Interpretable Machine Learning Analysis
J Med Internet Res 2026;28:e74359
URL: https://www.jmir.org/2026/1/e74359
doi:10.2196/74359
PMID:

©Shuang Geng, Yanghui Li, Jie Wang, Peixuan Chen, Xusheng Wu, Zhiqun Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effect of AI-Based Natural Language Feedback on Engagement and Clinical Outcomes in Fully Self-Guided Internet-Based Cognitive Behavioral Therapy for Depression: 3-Arm Randomized Controlled Trial

Mirai So[1*], MD, MBA, PhD; Yoichi Sekizawa[2*], MA; Sora Hashimoto[3*], MA; Masami Kashimura[4*], PhD; Hajime Yamakage[5*], MEng; Norio Watanabe[6*], MD, PhD

[1]Department of Psychiatry, Tokyo Dental College, Tokyo, Japan
[2]Research Institute of Economy, Trade and Industry, Tokyo, Japan
[3]United Health Communication Co., Ltd., Tokyo, Japan
[4]Department of Psychology, Faculty of Human Schiences, Tokiwa University, Ibaraki, Japan
[5]Department of Medical Statistics, Satista Co., Ltd., Kyoto, Japan
[6]Department of Psychiatry, Soseikai General Hospital, Kyoto, Japan
[*]all authors contributed equally

Corresponding Author:
Mirai So, MD, MBA, PhD
Department of Psychiatry
Tokyo Dental College
2-9-18 Misakicho, Chiyoda-ku
Tokyo, 101-0061
Japan
Phone: 81 47 322 0151
Fax: 81 47 325 4456
Email: mirai.so@keio.jp

Abstract

Background: Depression remains a major global cause of disability; yet, access to optimal mental health services is limited. Self-guided internet-based cognitive behavioral therapy (iCBT) offers a scalable alternative but is generally less effective than guided programs, showing limited antidepressant effects and incomplete symptomatic and functional recovery. Adherence remains a major barrier. Recent advances in artificial intelligence (AI), particularly natural language processing, enable automated advisory and empathic feedback that may enhance engagement and therapeutic impact. Although previous trials have reported promising effects, most used heterogeneous control conditions, making it difficult to isolate the specific contribution of AI within fully self-guided interventions.

Objective: This randomized controlled trial evaluated whether natural language processing–based AI feedback integrated into a fully self-guided iCBT program improves clinical outcomes and engagement compared with an otherwise identical iCBT program without AI support.

Methods: We recruited 1187 adults aged 20-60 years online and randomly assigned them to AI-augmented iCBT (AI-iCBT; n=396), iCBT without AI (n=397), or a waitlist control (n=394). Both active groups received 6 weekly sessions combining video-based psychoeducation and cognitive restructuring exercises. The AI-iCBT program additionally provided automated empathic and advisory feedback. The primary outcome was depressive symptom severity (Patient Health Questionnaire-9 [PHQ-9]) at week 7 and month 3, analyzed using mixed-effects models for repeated measures under an intention-to-treat framework. Secondary outcomes included a dichotomous PHQ-9 score of ≥10, Quick Inventory of Depressive Symptomatology, Generalized Anxiety Disorder-7, Sheehan Disability Scale, and weekly participation rates. Exploratory analyses assessed the impact of AI functions on engagement and antidepressant effects in the efficacy analysis set (EAS).

Results: In intention-to-treat analyses, no significant between-group differences were observed in mean PHQ-9 scores at week 7 or month 3, whereas engagement analyses showed a significant group × week interaction, with AI-iCBT participants demonstrating consistently higher odds of weekly participation (odds ratio 1.23, 95% CI 1.09-1.39; P<.001). Exploratory analyses indicated

XSL•FO
RenderX

that activation of the empathic feedback function strongly predicted adherence (odds ratio 9.99, 95% CI 5.80-17.21; *P*<.001), while advisory feedback was not significant. In EAS analyses, iCBT showed significant short-term improvement versus control at postintervention, whereas at follow-up, only AI-iCBT showed a significantly lower proportion of participants with a PHQ-9 score of ≥10 compared with control (difference –0.15, 95% CI –0.30 to –0.01; *P*=.046). No serious adverse events were reported.

Conclusions: AI support significantly improved adherence to a fully self-administered program. In EAS analyses, AI-iCBT also showed a significantly lower proportion of participants with PHQ-9 score of ≥10 at follow-up compared with control. Empathic feedback emerged as a key mechanism for sustaining engagement, suggesting that AI communication may help maintain participation in scalable digital mental health interventions. Further research is required.

Trial Registration: University Hospital Medical Information Network Clinical Trials Registry (UMIN-CTR) UMIN000019228; https://center6.umin.ac.jp/cgi-open-bin/ctr/ctr_view.cgi?recptno=R000022220

(J Med Internet Res 2026;28:e76902) doi:10.2196/76902

KEYWORDS

Adherence; AI; AI-supported psychotherapy; artificial intelligence; CBT; depression; internet-based CBT; natural language processing; RCT; self-help intervention

Introduction

Depression is a leading global cause of disability [1], substantially impairing quality of life [2,3] and imposing a considerable economic burden, including medical expenses [4] and productivity losses [5]. The rising prevalence of depression, combined with a shortage of health care resources, places a significant strain on health systems and professionals in meeting the growing demand [6-9].

In response, technology-delivered self-help interventions have emerged as promising solutions for managing mental health difficulties. Most of these interventions are based on cognitive behavioral therapy (CBT) and are generally referred to as internet-based CBT (iCBT) [10,11]. iCBT can be delivered either with therapist support (guided) or without therapist support (unguided, self-directed). Compared to traditional face-to-face therapy, iCBT offers major advantages in terms of accessibility, availability, and cost-effectiveness for both patients and providers. Furthermore, its online format provides benefits related to privacy, confidentiality, and anonymity, which can help reduce the stigma often associated with seeking mental health care [12,13].

While iCBT produces clinically meaningful symptom improvements, remission rates tend to be modest (approximately 30%-35%). An individual participant data meta-analysis reported a remission rate of 35.2% and a response rate of 56% [14]. Large-scale individual participant data network meta-analyses have consistently shown that guided iCBT yields higher response and remission rates than unguided formats, reflecting the challenges of engagement and dropout in fully self-administered programs [15-19].

This study specifically examines the unguided, fully self-administered format. Such interventions enable users to manage their symptoms independently and offer potential benefits such as reducing costs, alleviating the burden on health care providers, and improving access to mental health services in regions where such services are difficult to obtain [9,16]. Previous research has demonstrated that sociodemographic factors, such as age and sex, are associated with dropout risk in iCBT [20-22].

Nevertheless, if the effectiveness and engagement of unguided iCBT could be enhanced, the benefits of structured self-help materials would not be limited to fully self-administered interventions but would also extend to guided and blended formats. When patients acquire skills and knowledge through self-help modules, they can participate more effectively in therapist-led sessions, thereby enhancing overall treatment outcomes [23]. Strengthening such “self-help effects” not only amplifies therapeutic gains in guided and blended care but also reduces the time and workload required from therapists. By improving scalability and cost-effectiveness, it further increases the feasibility of implementation in routine practice [24,25]

To address these challenges, natural language processing (NLP), an artificial intelligence (AI) technology that enables the understanding and generation of human language, has increasingly been applied to enhance adherence and engagement through tailored feedback [26-28]. Whereas conventional unguided iCBT typically provides static or generic responses, in this study, we used an NLP-enabled iCBT program with automated advisory and empathic functions. This allows the system to generate advisory and empathic feedback in response to user input, potentially addressing both emotional and procedural barriers simultaneously.

Despite its promise, the specific therapeutic contributions of NLP remain unclear. Previous studies have frequently used heterogeneous control conditions such as waitlists, no intervention, treatment as usual, bibliotherapy, or conversational computer programs [28-32]. This heterogeneity makes it difficult to determine whether NLP provides distinct therapeutic benefits or merely functions as an active placebo by enhancing user expectations.

Against this background, the aim of this study was to conduct a randomized, parallel-group exploratory trial directly comparing 2 unguided, fully self-administered iCBT programs that were identical except for the presence or absence of NLP feedback. This design allowed us to evaluate the therapeutic contribution of NLP within a self-help framework in a blinded comparison of the 2 intervention groups.

XSL•FO
RenderX

Methods

Overview

This study was a 3-arm randomized controlled trial, with double-blinding between the AI-augmented iCBT (AI-iCBT) and iCBT groups, while the waitlist control group was unblinded. The intervention arms consisted of AI-iCBT, an unguided, fully self-administered iCBT program incorporating NLP feedback, and iCBT, an unguided, fully self-administered program without NLP feedback. These 2 arms are hereafter collectively referred to as "unguided iCBT." The waitlist group served as the control condition.

Study Participants

Invitation emails were sent to all monitors registered with Nikkei Research Inc, with the aim of recruiting at least 900 participants. Interested individuals were directed to complete an online screening survey, which included the Patient Health Questionnaire-9 (PHQ-9) [33,34], to determine eligibility.

Eligibility Criteria

Inclusion and exclusion criteria are provided in Textbox 1.

Textbox 1. Eligibility criteria.

Inclusion criteria

- Aged 20-60 years (to target the working-age population and exclude older adults with lower digital literacy)
- Had access to the internet
- Baseline Patient Health Questionnaire-9 score of ≥5 (this threshold was selected to avoid floor effects and ensure adequate symptom levels for change detection) [33,35,36]
- Ability to understand Japanese

Exclusion criteria

- Presence of medical conditions precluding participation as determined by a physician
- Concurrent participation in another cognitive behavioral therapy program
- Diagnosis of schizophrenia
- Severe suicidal ideation
- Diagnosis of dementia
- Substance dependence in the past 12 months (excluding smoking)

Eligible participants were randomly assigned to 1 of the 3 groups. Immediately before the intervention, PHQ-9 scores were reassessed; individuals with scores <5 were excluded from the efficacy analysis set (EAS) but remained in the overall study population. The detailed definition of the EAS is provided in the Analytic Strategy section.

Intervention

The AI-augmented iCBT program, developed by NEC Solution Innovators, Ltd (Tokyo), integrates an NLP module trained on 28,718 prior iCBT records from Japanese users. The entire program was delivered in Japanese, and Figure 1 presents an English-translated version of the original Japanese interface for publication purposes. The program includes a self-guided cognitive restructuring (CR) exercise, where users complete a 7-column thought record to address cognitive distortions. The NLP system processes user inputs, including situation, automatic thoughts, and feelings, referencing a corpus of past responses (Figure 1). It provides 2 types of automated feedback with text and phonation: (1) empathetic messages delivered through an animated character whose expressions, such as smiling or showing sadness, are synchronized with the message content, and (2) advisory messages offering guidance to refine inputs or direct users to appropriate exercises, including suggestions to revise content if the user's input was unclear or misplaced (eg, a feeling given instead of a thought; Figure 2).

In contrast, the non-AI iCBT program retained the same structure but provided only neutral, noncontextual responses, such as generic phrases like "Uh-huh" with neutral facial expressions. Both AI-iCBT and iCBT programs were otherwise identical in content, using the same validated 6-session self-guided iCBT package. This iCBT program has previously demonstrated significant antidepressant effects in a randomized controlled trial among working adults (n=60 per group), compared with a waitlist control, with a medium to large effect size (Cohen d=0.65; P<.005) [37]. This package consisted of 6 weekly sessions, each including a 15-minute video-based psychoeducation module covering standard CBT principles such as behavioral activation and problem-solving, along with a weekly CR exercise in which users applied learned techniques. In this trial, the only difference was the addition of the NLP feedback system. The program was available on both smartphones and PCs. The AI-enhanced features, which were designed in advance to improve user engagement and response accuracy, exhibited high usability, with low dissatisfaction rates reported for both the empathetic (3/32, 9.4%) and advisory (1/24, 4.2%) feedback functions (Figure 3).

XSL•FO
RenderX

Figure 1. Examples of expressions extracted from the natural language processing corpus and categorized into 4 domains: Problem, Trouble, Feeling, and Subjective.

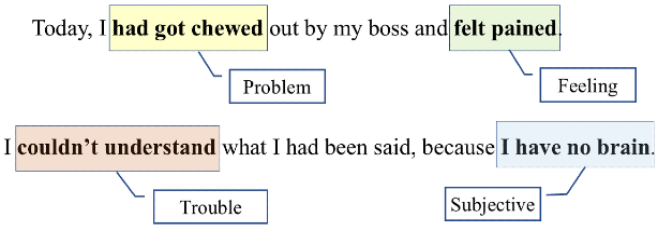

Figure 2. Workflow of artificial intelligence-guided internet-based cognitive behavioral therapy (CBT), showing the structured 7-step cognitive restructuring exercise with automated prompts and feedback. AT: automatic thought; NLP: natural language processing.

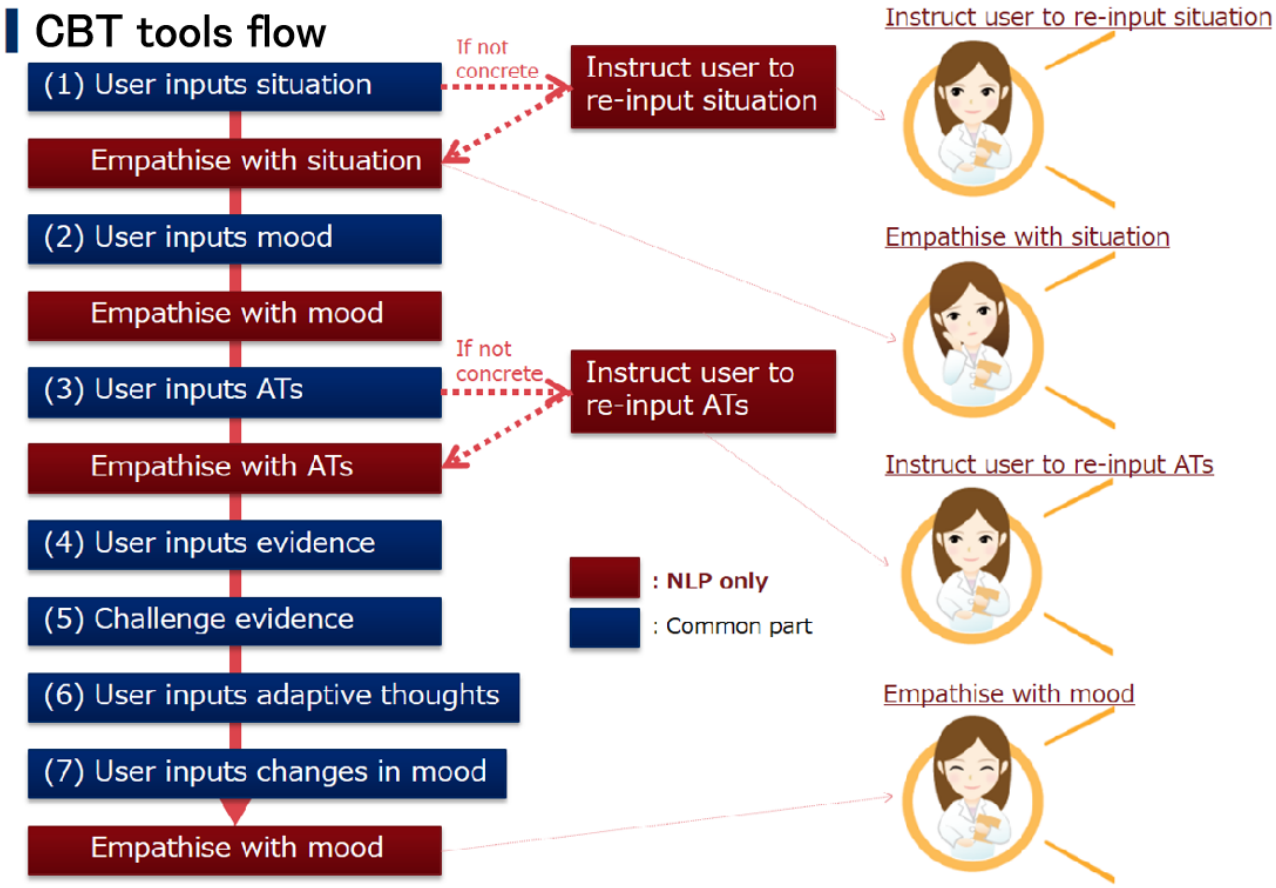

XSL•FO
RenderX

Figure 3. User acceptability ratings of natural language processing–generated feedback for empathy and automatic thought identification. AT: automatic thought.

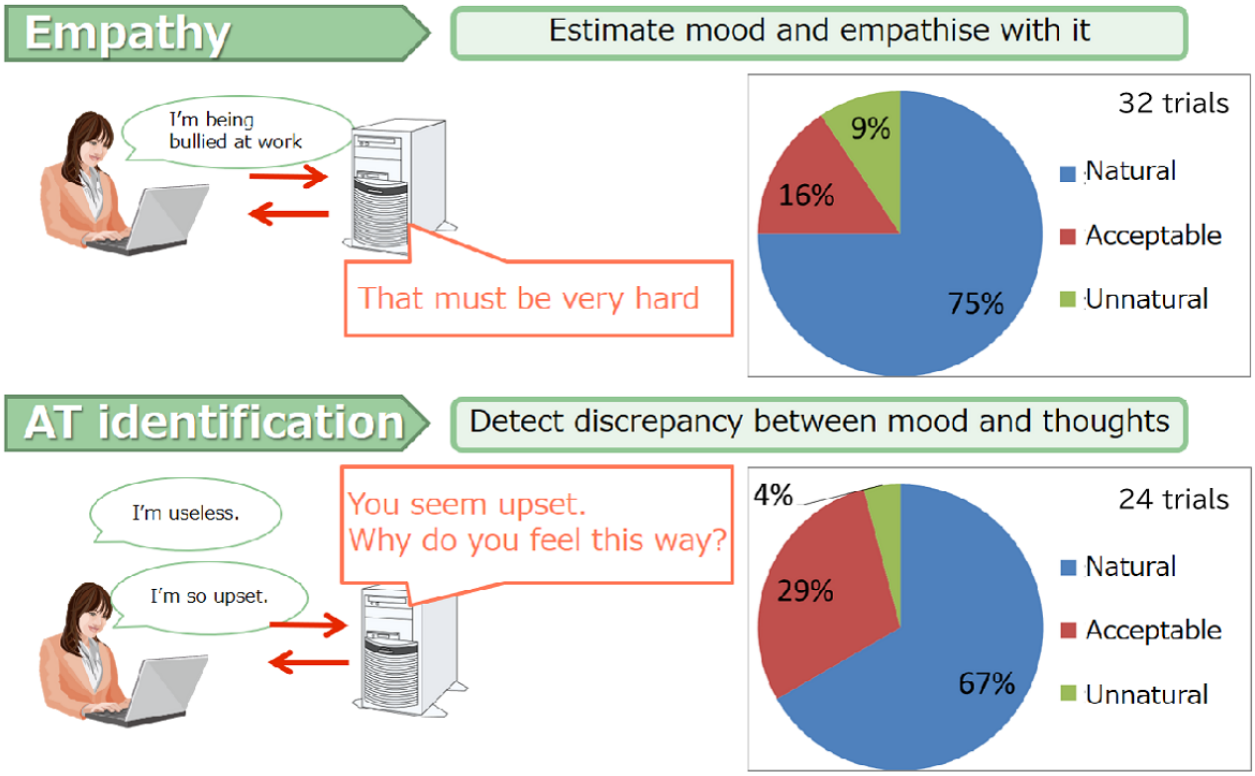

Randomization and Masking

The registered participants were randomly and concurrently assigned to either the AI-iCBT, iCBT, or waitlist groups using a computer-generated random sequence provided by an independent third party. Stratified randomization was applied based on age (≤40 vs >40 years), sex, and baseline PHQ-9 score (≤9 vs >9), as baseline symptom severity has been shown to influence treatment outcomes in self-guided iCBT [23]. Participants in the waitlist group were aware of their allocation and were therefore unblinded, whereas those in the AI-iCBT and iCBT groups were told only that they would participate in iCBT using "the latest technology," without disclosure of their specific group assignment. Accordingly, blinding was implemented between the 2 intervention groups.

Study Procedures

Automated email reminders were sent to participants twice weekly during the 7-week intervention period. Each week, participants in the intervention groups were required to (1) view an online psychoeducational CBT module and (2) perform their allocated (AI-iCBT or iCBT) CR exercise at least once (6 times or more in total). Waitlist participants did not undergo any exercises during this period. All participants were required to complete assessments at baseline, postintervention (week 7), and follow-up (month 3 after postintervention). All intervention and assessment procedures, including attendance and outcome measures, were conducted online.

Outcomes

All primary and secondary outcomes were analyzed based on the intention-to-treat (ITT) population, which included all randomized participants.

Primary Outcome

The primary outcome was the mean PHQ-9 score, assessed at baseline, week 7 (postintervention), and month 3 (follow-up). The PHQ-9 is a widely used self-report measure of depressive symptoms (range 0-27, higher scores indicating greater severity), originally developed by Kroenke et al [33] and validated in Japanese [38].

Secondary Outcomes

Secondary outcomes include (1) proportion of participants with PHQ-9 scores ≥10 (a conventional cutoff for probable major depression) [33,39]. Although not selected as the primary outcome in this study, such binary outcomes are often considered clinically meaningful, as they reflect remission from a diagnostic threshold [40-43]. (2) Quick Inventory of Depressive Symptomatology-Japanese version (QIDS-J) [43,44]—a self-report scale of depressive symptom severity. (3) Generalized Anxiety Disorder-7 (GAD-7) [45,46]—a self-report questionnaire measuring generalized anxiety symptoms. (4) Sheehan Disability Scale (SDS) [47,48], which evaluates functional impairment in work, social, and family life, with SDS ≥10 commonly used as a pragmatic threshold in clinical trials [49].

Engagement outcome included weekly CR exercise attendance rate (defined as attending at least one session per week) in the 2 intervention groups. Program satisfaction at week 7 was

XSL•FO
RenderX

assessed with the Client Satisfaction Questionnaire-8 (CSQ-8) [50,51], for which the Japanese version has demonstrated reliability and validity.

All outcomes except engagement and satisfaction were assessed at baseline, week 7, and month 3.

Analytic Strategy

Overview

In this study, all primary and secondary analyses were conducted in the ITT population, defined as all randomized participants. In addition, 2 exploratory analyses were performed: (1) as an ad hoc exploratory analysis, we examined which AI feedback function (empathy or advisory) contributed more to enhancing engagement, and (2) as an additional exploratory analysis, we assessed continuous and binary PHQ-9 outcomes within the EAS.

Engagement-Enhancing Factors

For this analysis, the 2 intervention groups were combined, and the presence or absence of empathy and advisory feedback during week 1 was examined in relation to engagement from weeks 2 to 6, defined as completing at least 1 exercise per week. The detailed statistical methods are described in the Statistical Analysis section.

Efficacy Analysis Set

The EAS was defined as participants with a baseline PHQ-9 score of ≥5 and completion of at least 3 out of the 6 weekly sessions. Participants with a baseline PHQ-9 score of <5 (minimal symptoms) were excluded, as their inclusion could reduce the power to detect change and dilute the mean effects [52,53]. Furthermore, previous research has demonstrated a dose-response relationship in iCBT, with clinical benefits emerging after completing approximately half of the modules; therefore, the minimum attendance criterion was set at 3 of 6 sessions [54,55].

Statistical Analysis

Overview

The sample size was estimated based on an assumed effect size of 0.10 (Cohen *d*) between the AI-iCBT and iCBT groups, given the absence of directly comparable prior studies. A dropout rate of 50% was anticipated based on patterns observed in similar previous studies. The power was set at 80% with a 2-sided significance level of α=.05. As this was an exploratory study, no adjustment for multiplicity was applied, and nominal *P* values were reported.

The primary and secondary analyses were conducted according to the ITT principle, including all randomized participants. Baseline demographic and clinical characteristics were compared across groups using 1-way ANOVA or chi-square tests.

Continuous outcomes (PHQ-9, QIDS-J, GAD-7, and SDS) were analyzed using a mixed-effects model for repeated measures (MMRM), with intervention, time, and intervention × time interaction as fixed effects, assuming an unstructured covariance structure. Results are presented as least squares means with 95% CIs.

Binary outcomes (PHQ-9 ≥10) were analyzed using generalized linear mixed models (GLMMs) with a binomial distribution and logit link, including intervention, time, and their interaction as fixed effects, and subject as a random effect. Estimated proportions and their 95% CIs were reported. Missing data for the outcomes were handled under the missing at random assumption within the MMRM and GLMMs framework.

CR exercise participation rates (defined as at least 1 completion per week) in the intervention groups were analyzed using GLMMs with a logit link, including intervention, week (as a continuous variable), and intervention × week interaction as fixed effects.

Exploratory Analyses

The following two exploratory analyses were conducted.

Engagement-Enhancing Factors

The dependent variable was defined as achieving at least 1 CR exercise per week across all weeks from week 2 to week 6 (yes/no). Independent variables were the presence or absence of empathy or advisory feedback during week 1. Covariates included age, sex, marital status, education, employment status, history of psychiatric and physical treatment, baseline PHQ-9 score, and intervention group (AI-iCBT vs iCBT), as group differences could confound the association of interest. Analyses were performed using generalized estimating equations logistic regression models to account for within-subject correlation and to estimate population-averaged effects.

Efficacy Analysis Set

In the EAS (participants with a baseline PHQ-9 score of ≥5 and completion of ≥3 sessions), continuous and binary PHQ-9 outcomes were additionally adjusted for age, sex, baseline PHQ-9 score, and medical history as covariates to account for potential group imbalances in the restricted sample.

Sensitivity Analyses

Associate Factors of Low Adherence

To explore potential factors associated with dropout, we compared baseline characteristics between EAS participants who attended ≥3 sessions and those who attended <3 sessions, given the high attrition typically observed in self-guided digital interventions.

Alternative Definition of Caseness

We conducted an exploratory analysis on the binary PHQ-9 outcome in the EAS, applying a stricter definition of depression severity: a PHQ-9 score of ≥10 plus at least 1 core symptom (depressed mood or anhedonia) [56,57], together with an SDS score of ≥10 as an indicator of functional impairment [49,58-60].

All analyses were performed using IBM SPSS Statistics version 26.0.

Reporting Standards

Reporting of this trial followed the CONSORT (Consolidated Standards of Reporting Trials) 2010 statement [61] and the

XSL•FO
RenderX

CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist [62] for internet-based interventions. The completed CONSORT-EHEALTH checklist is submitted as Multimedia Appendix 1.

Ethical Considerations

This study was reviewed and approved by the Hiramatsu Memorial Hospital Ethics Committee (approval number 20150807). All participants provided informed consent electronically prior to enrollment after reading an online information sheet describing the study purpose, procedures, potential risks, and voluntary nature of participation. Participants were informed that they could withdraw at any time without penalty.

The trial was prospectively registered in the University Hospital Medical Information Network Clinical Trials Registry.

All data were anonymized prior to analysis to ensure confidentiality. No personally identifiable information was accessible to the research team. Participants who completed the final assessment received a ¥500 (US $4.5) gift voucher as compensation. No identifiable images or other personal data are presented in this manuscript.

Results

Study Participants

A total of 1187 participants were eligible and randomly allocated to the AI-iCBT (n=396), iCBT (n=397), or waitlist (n=394) groups (ITT population; see Figure 4 for the CONSORT flow diagram). Baseline demographic and clinical characteristics are summarized in Table 1. The mean age was 43.50 (SD 9.85) years, and 699 (58.8%) of 1187 participants were male. Across the 3 groups, demographic and clinical characteristics were well balanced, with no significant differences in depressive symptom severity (PHQ-9: P=.56; QIDS-J: P=.74). No significant baseline differences were found between the AI-iCBT and iCBT groups, confirming the comparability of the 2 active interventions.

Figure 4 shows the flow of participants through the trial, including the numbers assessed for eligibility, randomized, allocated to each study arm (AI-iCBT, iCBT, control), completing follow-up assessments at week 7 and month 3, and included in the ITT analysis.

XSL•FO
RenderX

Figure 4. CONSORT (Consolidated Standards of Reporting Trials) 2010 flow diagram of participant enrollment, allocation, follow-up, and analysis. AI-iCBT: artificial intelligence–augmented internet-based cognitive behavioral therapy; iCBT: internet-based cognitive behavioral therapy; ITT: intention-to-treat.

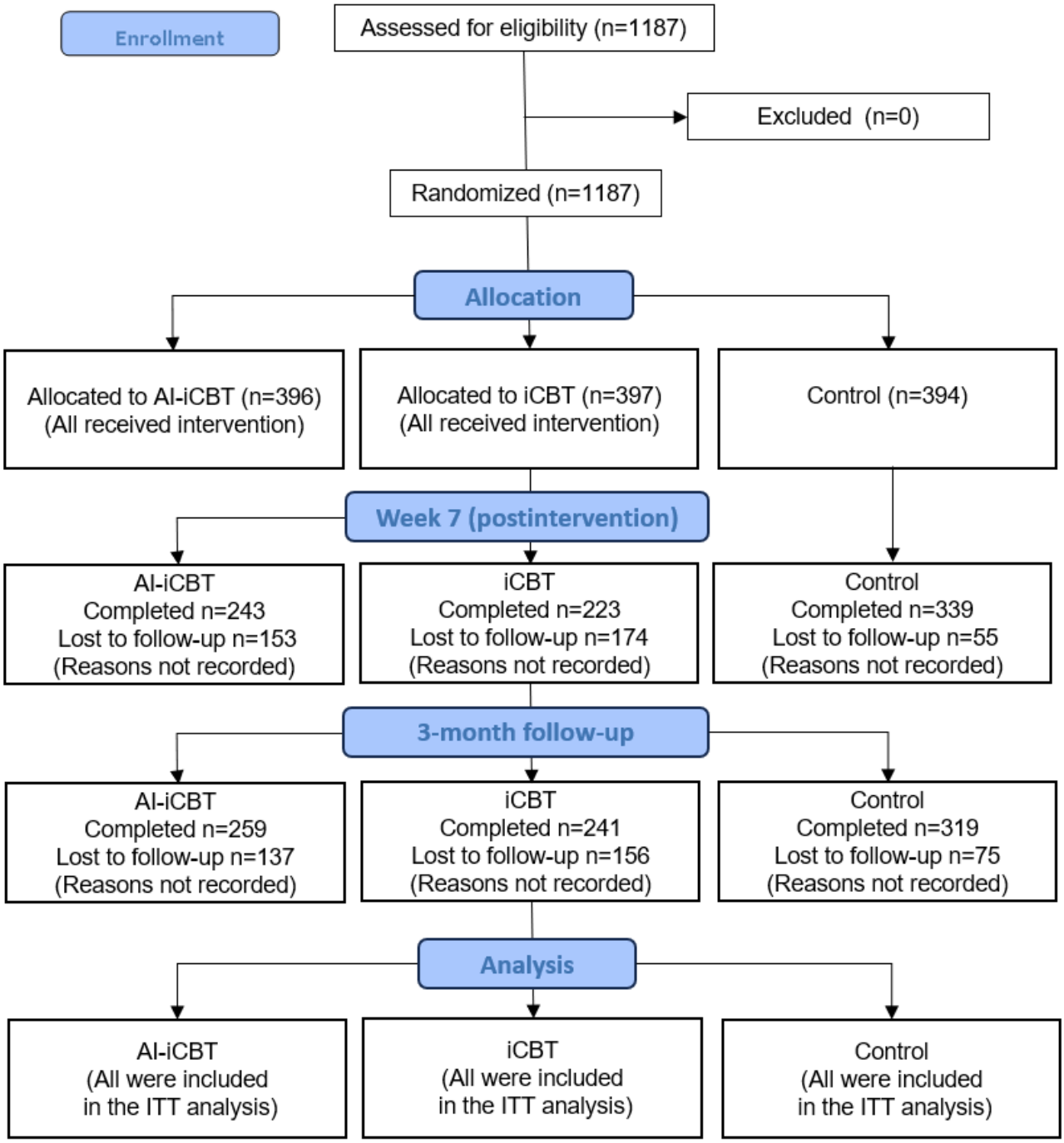

XSL•FO
RenderX

Table 1. Participants' characteristics.

Characteristic	Total (N=1187)	AI-iCBT[a] (n=396)	iCBT[b] (n=397)	Control (n=394)	*P* value (overall)	*P* value (AI-iCBT vs iCBT)[c]
Sex, n (%)					.96[d]	N/A[e]
Male	698 (58.8)	232 (58.6)	232 (58.4)	234 (59.4)		
Female	489 (41.2)	164 (41.4)	165 (41.6)	160 (40.6)		
Age (years)					.81[f]	N/A
Mean (SD)	43.5 (9.9)	43.6 (9.5)	43.2 (9.9)	43.6 (10.1)		
Median (IQR)	44 (36-52)	44 (36.8-51)	44 (36-52)	45 (36-52)		
Minimum-Maximum	20-60	20-60	20-60	20-60		
Marital status, n (%)					.28[d]	N/A
Married	665 (56)	226 (57.1)	219 (55.2)	220 (55.8)		
Divorced	76 (6.4)	29 (7.3)	17 (4.3)	30 (7.6)		
Bereaved	7 (0.6)	2 (0.5)	4 (1)	1 (0.3)		
Single	439 (37)	139 (35.1)	157 (39.5)	143 (36.3)		
Educational background, n (%)					.63[d]	N/A
Junior high school	7 (0.6)	3 (0.8)	1 (0.3)	3 (0.8)		
High school	239 (20.1)	87 (22)	84 (21.2)	68 (17.3)		
Junior college or technical	216 (18.2)	71 (17.9)	73 (18.4)	72 (18.3)		
University or postgraduate	725 (61.1)	235 (59.3)	239 (60.2)	251 (63.7)		
Employment status, n (%)					.78[d]	N/A
Working	972 (81.9)	320 (80.8)	324 (81.6)	328 (83.2)		
Unemployed (seeking)	79 (6.7)	27 (6.8)	30 (7.6)	22 (5.6)		
Unemployed (not seeking)	136 (11.5)	49 (12.4)	43 (10.8)	44 (11.2)		
Medical history, n (%)					.13[d]	N/A
No relevant history	918 (77.3)	316 (79.8)	296 (74.6)	306 (77.7)		
Ambulatory	258 (21.7)	74 (18.7)	97 (24.4)	87 (22.1)		
Hospitalized	11 (0.9)	6 (1.5)	4 (1)	1 (0.3)		
Mental history, n (%)					.93[d]	N/A
In treatment	129 (10.9)	46 (11.6)	44 (11.1)	39 (9.9)		
Treated	184 (15.5)	61 (15.4)	59 (14.9)	64 (16.2)		
No relevant history	874 (73.6)	289 (73)	294 (74.1)	291 (73.9)		
PHQ-9[g] score ≥10	428 (36.1)	142 (35.9)	143 (36)	143 (36.3)	.99[d]	N/A
Baseline scale score, mean (SD)						
PHQ-9	8.7 (5.2)	8.8 (5.2)	8.5 (5.2)	8.9 (5.1)	.56[f]	.41[f]
QIDS-J[h]	8.7 (4.9)	8.8 (4.9)	8.6 (4.9)	8.8 (4.7)	.74[f]	.49[f]
GAD-7[i]	6.0 (4.6)	6.0 (4.4)	5.9 (4.7)	6.2 (4.7)	.59[f]	.93[f]

[a]AI-iCBT: artificial intelligence–augmented internet-based cognitive behavioral therapy.

[b]iCBT: internet-based cognitive behavioral therapy.

[c]*P* values represent pairwise comparisons between AI-iCBT and iCBT groups.

[d]*P* value is based on the chi-square test.

[e]N/A: not applicable.

[f]*P* value is based on ANOVA.

XSL•FO
RenderX

[g]PHQ-9: Patient Health Questionnaire-9.

[h]QIDS-J: Quick Inventory of Depressive Symptomatology-Japanese version.

[i]GAD-7: Generalized Anxiety Disorder-7.

Primary and Secondary Outcomes (ITT Population)

The primary outcome, the mean score on the PHQ-9, did not show statistically significant between-group differences compared with the control group at either week 7 or month 3 (AI-iCBT vs control: least squares mean difference –0.47, 95% CI –1.13 to 0.18; *P*=.16; Cohen *d*=–0.10; iCBT vs control: least squares mean difference –0.62, 95% CI –1.28 to 0.04; *P*=.07; Cohen *d*=–0.13; Table 2). No significant differences were observed between the AI-iCBT and iCBT groups. Nevertheless, both intervention groups showed significant within-group reductions from baseline at week 7 and month 3 (all *P*<.001), indicating that depressive symptoms improved over time in both groups.

Table 2. Primary outcome: mean Patient Health Questionnaire-9 scores at baseline, week 7, and month 3 (intention-to-treat population). Values are least squares (LS) means with 95% CIs estimated using a mixed model for repeated measures. Between-group comparisons are shown.

Time point	AI-iCBT[a] group, LS mean (95% CI)	iCBT[b] group, LS mean (95% CI)	Control group, LS mean (95% CI)	AI-iCBT vs control, Δ (95% CI)	iCBT vs control, Δ (95% CI)
Baseline	8.79 (8.47 to 9.11)	8.73 (8.41 to 9.05)	8.81 (8.48 to 9.13)	__[c]	—
Week 7	7.90 (7.49 to 8.31)	7.75 (7.33 to 8.18)	8.28 (7.94 to 8.63)	–0.37 (–1.02 to 0.29); *P*=.27	–0.46 (–1.13 to 0.21); *P*=.18
Month 3	7.37 (6.98 to 7.77)	7.17 (6.76 to 7.58)	7.86 (7.50 to 8.22)	–0.47 (–1.13 to 0.18), *P*=.16	–0.62 (–1.28– to 0.04); *P*=.07

[a]AI-iCBT: artificial intelligence–augmented internet-based cognitive behavioral therapy.

[b]iCBT: internet-based cognitive behavioral therapy.

[c]Not available.

For the secondary binary outcome of PHQ-9 ≥10, the overall proportion decreased over time across all groups (Figure 5). At month 3, the proportion was numerically lower in the AI-iCBT group compared with the control group, but between-group differences were not statistically significant in the ITT analysis (Multimedia Appendix 2).

Figure 5. Secondary outcome: proportion of participants with Patient Health Questionnaire-9 (PHQ-9) scores ≥10 at baseline, week 7, and month 3 (intention-to-treat population). Estimated proportions and 95% CIs were derived from generalized linear mixed models with a logit link, including effects for group, time, and their interaction. AI-iCBT: artificial intelligence–augmented internet-based cognitive behavioral therapy; iCBT: internet-based cognitive behavioral therapy.

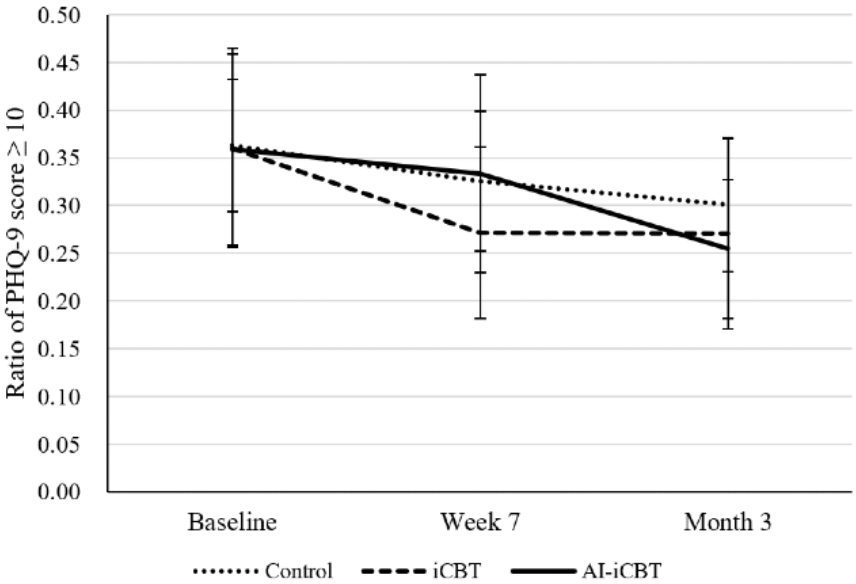

Similar patterns were observed for other secondary measures. QIDS-J and GAD-7 scores improved significantly within both intervention groups but without significant between-group differences. SDS scores showed modest reductions but did not significantly differ from control. Full secondary outcome results are provided in Multimedia Appendix 3.

Engagement and User Satisfaction

Overview

As illustrated in Figure 6, the CR exercise participation rate decreased significantly over time across both intervention groups (odds ratio [OR] 0.751, 95% CI 0.692-0.815; *P*<.001). Participation began at only about half of participants in week

XSL•FO
RenderX

1 and declined further, dropping more steeply in the iCBT group, which fell to around 30% by week 6. In contrast, the AI-iCBT group retained somewhat higher engagement, remaining closer to the low 40% range by week 6, suggesting that AI support helped sustain participation over time. Between-group differences over time were examined using GLMM with a logit link, which showed a significant group × week interaction favoring AI-iCBT (OR 1.23, 95% CI 1.09-1.39; *P*<.001; Table 3). Analyses included all randomized participants in the intervention groups.

Figure 6. Engagement outcome: weekly participation rates in the artificial intelligence–augmented internet-based cognitive behavioral therapy (AI-iCBT) and internet-based cognitive behavioral therapy (iCBT) groups during weeks 1-6 (intention-to-treat population). Participation was defined as completion of at least 1 cognitive restructuring exercise per week. Error bars indicate 95% CIs.

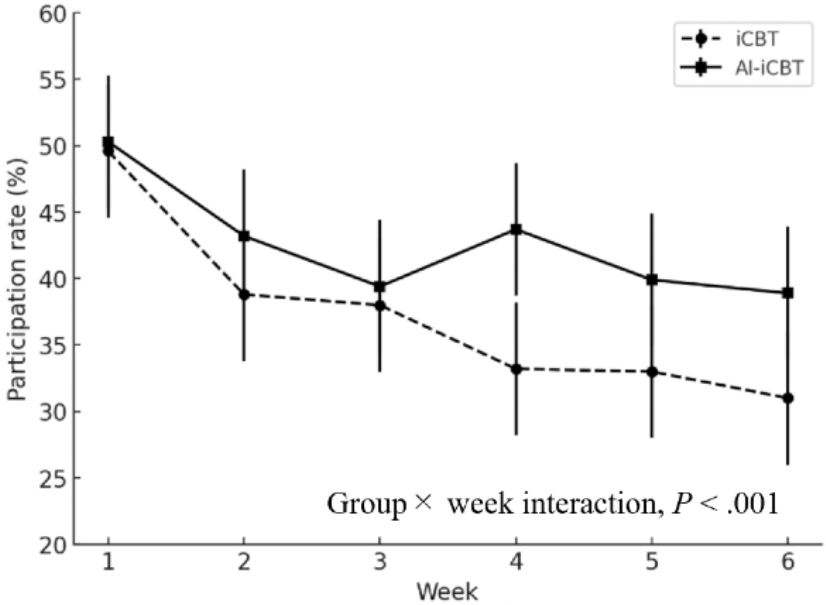

Table 3. Engagement outcome: mixed-effects logistic regression results for weekly cognitive restructuring exercise participation rates (intention-to-treat population). The generalized linear mixed model with a logit link included fixed effects for group, week (continuous and centered), and their interaction (group × week).

Effect	Reference	Odds ratio (95% CI)	*P* value
Group (AI-iCBT[a] vs iCBT[b])	iCBT	0.807 (0.370-1.758)	.59
Week (continuous, centered)	—[c]	0.751 (0.692-0.815)	<.001
Group × week	—	1.229 (1.090-1.386)	<.001

[a]AI-iCBT: artificial intelligence–augmented internet-based cognitive behavioral therapy.

[b]iCBT: internet-based cognitive behavioral therapy.

[c]Not available.

User Satisfaction

Assessed at week 7 with the CSQ-8, averaged about 21 out of 32 points in both intervention groups, indicating a moderate to good level of satisfaction. No significant difference was observed between AI-iCBT and iCBT (Multimedia Appendix 4).

Exploratory Analysis of Engagement-Enhancing Factors

Activation of the empathy function was significantly associated with higher participation (OR 9.99, 95% CI 5.80-17.21; *P*<.001). In contrast, activation of the advisory function was not significantly associated with engagement (OR 2.37, 95% CI 0.96-5.83; *P*=.06). Detailed adjusted results are provided in Multimedia Appendix 5.

Exploratory EAS Analysis

In the exploratory EAS analysis, based on the ITT population, 312/396 (78.8%) in the AI-iCBT group, 312/397 (78.6%) in the iCBT group, and 317/394 (80.5%) in the control group had a baseline PHQ-9 score of ≥5. Among the ITT population, the proportion of participants who attended 3 or more sessions was 188/396 (47.5%) in the AI-iCBT group and 158/397 (39.8%) in the iCBT group. Furthermore, the proportion of participants with a baseline PHQ-9 score of ≥5 who attended 3 or more sessions (EAS) was 149/396 (37.6%) in the AI-iCBT group, 134/397 (33.8%) in the iCBT group, and 317/394 (80.5%) in the control group.

Mean PHQ-9 scores decreased significantly from baseline to week 7 in all 3 groups. At week 7, the iCBT group showed a statistically significant improvement compared with the control group (Δ=–1.08, 95% CI –1.98 to –0.18; *P*=.02). However, this

XSL•FO
RenderX

difference was not maintained at month 3. Full numerical results are provided in Multimedia Appendix 6. By contrast, the proportion of participants scoring ≥10 on the PHQ-9 decreased only in the iCBT group compared with the control at week 7 (Figure 7). At month 3, the AI-iCBT group showed a significantly lower proportion compared with control (Δ –0.15, 95% CI –0.30 to –0.01; *P*=.046), while the iCBT group did not differ significantly. The group × time interaction was significant (*P*=.008), indicating that the pattern of improvement differed between intervention groups (see Multimedia Appendix 7 for full numerical results).

Figure 7. Exploratory outcome: proportion of participants with Patient Health Questionnaire-9 (PHQ-9) scores of ≥10 in the efficacy analysis set population at baseline, week 7 (postintervention), and month 3 (follow-up). Error bars indicate 95% CIs. Asterisks represent *P*<.05 versus control. AI-iCBT: artificial intelligence–augmented internet-based cognitive behavioral therapy; iCBT: internet-based cognitive behavioral therapy.

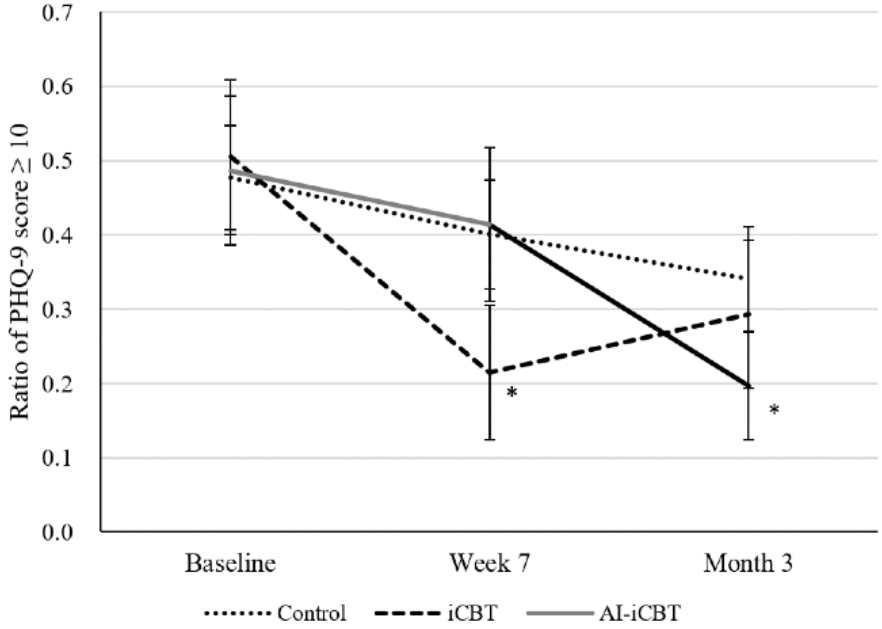

Sensitivity Analyses

In evaluating factors associated with low adherence, participants who attended <3 sessions were more likely to be male (65.1% vs 48.4%; *P*<.001), older (mean age 44.2, SD 9.44 vs mean age 41.7, SD 9.55 years; *P*=.001), and employed (287/341, 84.2% vs 212/283, 74.9%; *P*=.02), compared with those who attended ≥3 sessions. No significant differences were observed for marital status, educational background, medical history, or mental health history (Multimedia Appendix 8).

As a further sensitivity analysis, when applying a stricter definition of depression severity—PHQ-9 score of ≥10 plus at least 1 core symptom and SDS score of ≥10—62.7% (178/284) of participants with a PHQ-9 score of ≥10 at baseline met this definition (AI-iCBT: 44/70, 62.9%; iCBT: 41/71, 57.7%; control: 93/143, 65%), with no significant group differences (Multimedia Appendix 9).

Discussion

Overview

This study has several unique features. First, it directly compared a fully self-administered CR exercise delivered via iCBT, with and without AI-based NLP functionality, under a randomized (partially masked) design. Notably, the addition of AI led to a statistically significant improvement in engagement—an effect, to our knowledge, not previously documented. As AI-based interventions have become increasingly sophisticated and deeply integrated into intervention programs, disentangling the specific contribution of AI has become difficult. In particular, establishing control conditions that differ only in the presence or absence of AI functionality requires substantial resources, and prior studies have therefore often relied on heterogeneous control conditions. By applying a more robust design—feasible in part because the technology was still in a transitional phase—this trial provides new insights into how AI may enhance fully self-administered iCBT and offers a timely perspective for advancing scalable mental health care.

Self-administered interventions are known to be modestly to moderately effective for depression [15,23], but adherence remains a major limitation [15,18,19,21,55,63]. Systematic reviews indicate that approximately one-third to one-half of participants drop out before completing the program [19,64]. In this context, the engagement improvement observed in this study represents a potential step toward overcoming this barrier.

With regard to clinical effectiveness, no additional benefits of NLP feedback were observed in the ITT population. Recent evidence has reported that greater antidepressant effects are associated with lower dropout rates [15,20,65]. In contrast, although no between-group differences in antidepressant effects were found here, the addition of AI feedback was associated with a statistically significant increase in adherence. This suggests a novel engagement-enhancing mechanism, distinct from the traditionally assumed link between larger clinical effects and lower dropout rates.

Exploratory analyses further indicated that the “empathic function” of AI feedback was significantly associated with

XSL•FO
RenderX

improved adherence, whereas the advisory function showed no significant effect. Participants who received empathic responses during their first exercise subsequently demonstrated higher adherence. While self-disclosure was not directly measured, the sense of being supported may have facilitated persistence. These findings align with prior evidence that empathic conversational agents and chatbots support therapeutic alliance and sustained engagement [66-69]. Research in behavioral change has likewise shown that AI-mediated feedback can promote sustained self-management [70], supporting the plausibility of these findings. Such effects of human-AI interaction may have been less visible in prior studies using heterogeneous control conditions but became evident here through the structured randomized design.

Regarding antidepressant effects, no significant between-group differences were observed in the ITT population. In the EAS, results—while requiring cautious interpretation—indicated that at week 7 only the iCBT group showed significant improvement in both the mean PHQ-9 score and the proportion of participants with a PHQ-9 score of ≥10 compared with the control group, whereas the AI-iCBT group did not.

This suggests that the AI function may have attenuated or failed to enhance short-term antidepressant effects. However, this short-term benefit in the iCBT group disappeared at long-term follow-up. For the continuous outcome in particular, the short-term difference was –1.1 points on the PHQ-9, below the minimal clinically important difference (approximately 3 points) [35,71], suggesting limited clinical significance.

By contrast, the short-term dichotomous outcome in the iCBT group represented about a 29% reduction in the proportion of PHQ-9 scores of ≥10 cases relative to the control group. This implies that part of the potential benefit may not have been realized when AI was introduced. At long-term follow-up (month 3), however, only the AI-iCBT group showed a significant reduction of about 15% compared with the control group. These findings highlight the absence of the expected short-term effect in AI-iCBT and the unique long-term effect observed only in AI-iCBT.

The fact that AI-iCBT ultimately demonstrated an effect at long-term follow-up is noteworthy. Although exploratory, this suggests a potential contribution of AI-iCBT in reducing the proportion of participants exceeding a clinically significant severity threshold. One possible explanation is that approaches emphasizing empathy as a core therapeutic skill—such as interpersonal psychotherapy or family therapies—often show slower onset but more enduring gains compared with CBT, lasting well beyond the end of treatment [72-77]. It is possible that the empathy-related function of AI contributed in a similar way, although the underlying mechanisms remain unclear.

The EAS, however, was more restrictive than the ITT population. Among ITT participants with a baseline PHQ-9 score of ≥5, only 149/312 (47.8%) in the AI-iCBT group and 134/312 (42.9%) in the iCBT group attended at least 3 sessions. Furthermore, although a PHQ-9 score of ≥10 is widely recognized as a proxy for “major depression equivalent” in research [33,39], concerns have been raised that it may not be sufficient for diagnostic purposes and may contribute to overdiagnosis [56,78,79]. As a sensitivity analysis, therefore, we used a stricter definition requiring a PHQ-9 score of ≥10 plus at least 1 core symptom (depressed mood or anhedonia) [56,57], together with an SDS score of ≥10 as an indicator of functional impairment [49,58-60]. Results confirmed that only about 62.7% (178/284) of participants who met the PHQ-9 score of ≥10 at baseline also met this stricter definition, highlighting the importance of cautious interpretation (Multimedia Appendix 9).

Taken together, regardless of whether it corresponds to major depression, the finding of a significant reduction in the proportion of participants with clinically meaningful depressive states (PHQ-9 ≥10) compared with the control group may have clinical significance, particularly given the fully self-help nature of the program. From a public health perspective, such a difference could also carry implications for the scalability of self-help programs that do not require therapist involvement.

AI communication, including generative AI, continues to advance rapidly. However, the development of appropriate control programs has often been constrained by logistical and cost-related factors, limiting opportunities to rigorously investigate the antidepressant and anxiolytic effects of AI. Beyond psychiatry, maximizing the effectiveness of self-administered interventions while enhancing engagement remains a critical challenge across health care and welfare domains. This study represents a step toward addressing this challenge.

Limitations

First, most participants were recruited from a research panel with high affinity for digital interventions. Only 10.9% (129/1187) were actual users of mental health services, which is consistent with the national average in Japan, but caution is required in generalizing these findings to broader populations. Second, high dropout rates were observed, with only 37.6% (149/396) of the AI-iCBT group and 33.8% (134/397) of the iCBT group in the ITT population meeting EAS criteria. Additional analyses indicated that low adherence was associated with being male, older, and employed (Multimedia Appendix 8). Consistent with recent studies, time constraints [22], particularly among employed men [21] and older adults [20,21], were confirmed as key barriers to engagement. Third, although the AI-iCBT group consistently showed 5%-10% higher adherence rates than the iCBT group throughout the trial, both groups had already dropped by half to 50% (396/793) at the very first session and continued to decline over time, remaining at only 30%-40%.

Fourth, missing data were substantial, and MMRM and generalized estimating equations were applied to minimize bias. However, these approaches assume data are missing at random. In this study, since attrition occurred according to participant attributes, the possibility of missing not at random cannot be ruled out, and estimates may remain biased. Fifth, a significant group × time interaction was observed in adherence in the AI feedback group, suggesting a potential role of AI in improving engagement. However, this conclusion is based on a single trial and requires replication in different populations and designs, as well as further elucidation of the underlying mechanisms.

XSL•FO
RenderX

Sixth, due to technical issues, session-by-session data on depressive symptoms (Overall Anxiety Severity and Impairment Scale) [80] and anxiety (Overall Depression Severity and Impairment Scale) [81] were lost, precluding more detailed analyses. Future studies should implement automatic data saving and backup systems to prevent such loss. Overall antidepressant effects were small. In addition to limitations of the program itself, this may reflect a ceiling effect due to the predominance of participants with mild depression. Several meta-analyses [23,82,83] report that treatment effects may be less pronounced in cases with mild baseline depression compared to moderate or severe cases. Another limitation is related to our stratified randomization. While stratification by age, sex, and baseline PHQ-9 severity increased internal validity by balancing key prognostic factors, it may also restrict the generalizability of our findings to populations with different distributions of these characteristics.

Despite these challenges, this study provides valuable insights into the potential of AI-enhanced self-help digital interventions, particularly in relation to participant behavior dynamics. Importantly, no major adverse events were reported, underscoring the safety of this innovative approach and its potential to significantly advance mental health care practices.

Acknowledgments

We are grateful to Dr Yutaka Ono (Center for the Development of Cognitive Behavior Therapy Training, Tokyo), Professor Takashi Watanabe (Teikyo Heisei University, Tokyo), Mr Kenichiro Tsumura (T Quest, Chiba), Dr Sosei Yamaguchi (National Center of Neurology and Psychiatry, Tokyo), and all members of the project teams of NEC Solution Innovators, Ltd (Tokyo) for allowing us to unrestrictedly use artificial intelligence–augmented internet-based cognitive behavioral therapy for the study.

Funding

This study is funded by the Research Institute of Economy, Trade, and Industry (a think tank under the Ministry of Economy, Trade, and Industry of the Government of Japan). The funder had no role in study design, data collection, data analysis, data interpretation, or writing of the manuscript.

Data Availability

The data used in this study belong to the Research Institute of Economy, Trade and Industry and can be obtained from the institute upon reasonable request.

Authors' Contributions

MS contributed to conceptualization, investigation, methodology, project administration, resources, software, validation, visualization, and writing—including original draft preparation, review, and editing. YS contributed to conceptualization, funding acquisition, investigation, methodology, project administration, resources, and writing—review and editing. SH contributed to data curation, formal analysis, validation, visualization, and writing—review and editing. MK and HY contributed to secondary analysis and writing—review and editing. NW contributed to conceptualization, investigation, methodology, supervision, validation, and writing—review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-eHEALTH checklist.
[PDF File (Adobe PDF File), 340 KB - jmir_v28i1e76902_app1.pdf]

Multimedia Appendix 2
Secondary outcome: ratio of participants with Patient Health Questionnaire-9 scores ≥10 (intention-to-treat population).
[DOCX File , 17 KB - jmir_v28i1e76902_app2.docx]

Multimedia Appendix 3
Secondary outcomes: mean score of Quick Inventory of Depressive Symptomatology-Japanese version, Generalized Anxiety Disorder-7, and Sheehan Disability Scale (intention-to-treat population).
[DOCX File , 19 KB - jmir_v28i1e76902_app3.docx]

Multimedia Appendix 4

XSL•FO
RenderX

Satisfaction outcome: Client Satisfaction Questionnaire-8 total scores at week 7 (intention-to-treat population, intervention groups only).
[DOCX File , 17 KB - jmir_v28i1e76902_app4.docx]

Multimedia Appendix 5
Associations between artificial intelligence feedback functions and exercise attendance (artificial intelligence–augmented internet-based cognitive behavioral therapy and internet-based cognitive behavioral therapy participants, weeks 2–6).
[DOCX File , 20 KB - jmir_v28i1e76902_app5.docx]

Multimedia Appendix 6
Exploratory outcome: mean Patient Health Questionnaire-9 scores (efficacy analysis set population).
[DOCX File , 18 KB - jmir_v28i1e76902_app6.docx]

Multimedia Appendix 7
Exploratory outcome: ratio of participants with Patient Health Questionnaire-9 scores ≥10 (efficacy analysis set population).
[DOCX File , 18 KB - jmir_v28i1e76902_app7.docx]

Multimedia Appendix 8
Associated factors of low adherence in the efficacy analysis set.
[DOCX File , 20 KB - jmir_v28i1e76902_app8.docx]

Multimedia Appendix 9
Baseline prevalence of major depression proxies in the efficacy analysis set population.
[DOCX File , 18 KB - jmir_v28i1e76902_app9.docx]

References

1. GBD 2017 DiseaseInjury IncidencePrevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the global burden of disease study 2017. Lancet 2018;392(10159):1789-1858 [FREE Full text] [doi: 10.1016/S0140-6736(18)32279-7] [Medline: 30496104]
2. Arias D, Saxena S, Verguet S. Quantifying the global burden of mental disorders and their economic value. EClinicalMedicine 2022;54:101675 [FREE Full text] [doi: 10.1016/j.eclinm.2022.101675] [Medline: 36193171]
3. Hohls JK, König HH, Quirke E, Hajek A. Anxiety, depression and quality of life-a systematic review of evidence from longitudinal observational studies. Int J Environ Res Public Health 2021;18(22):12022 [FREE Full text] [doi: 10.3390/ijerph182212022] [Medline: 34831779]
4. König H, König HH, Konnopka A. The excess costs of depression: a systematic review and meta-analysis. Epidemiol Psychiatr Sci 2019;29:e30 [FREE Full text] [doi: 10.1017/S2045796019000180] [Medline: 30947759]
5. WHO. Mental health at work. World Health Organization. Geneva; 2022. URL: https://www.who.int/news-room/fact-sheets/detail/mental-health-at-work [accessed 2025-12-26]
6. Cunningham PJ. Beyond parity: primary care physicians' perspectives on access to mental health care. Health Aff (Millwood) 2009;28(3):w490-w501. [doi: 10.1377/hlthaff.28.3.w490] [Medline: 19366722]
7. Wang PS, Aguilar-Gaxiola S, Alonso J, Angermeyer MC, Borges G, Bromet EJ, et al. Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. Lancet 2007;370(9590):841-850 [FREE Full text] [doi: 10.1016/S0140-6736(07)61414-7] [Medline: 17826169]
8. Wainberg ML, Lu FG, Riba MB. Global mental health. Acad Psychiatry 2016;40(4):647-649. [doi: 10.1007/s40596-016-0577-0] [Medline: 27259490]
9. Edge D, Watkins ER, Limond J, Mugadza J. The efficacy of self-guided internet and mobile-based interventions for preventing anxiety and depression: a systematic review and meta-analysis. Behav Res Ther 2023;164:104292 [FREE Full text] [doi: 10.1016/j.brat.2023.104292] [Medline: 37003138]
10. Psychological Interventions Implementation Manual: Integrating Evidence-Based Psychological Interventions into Existing Services. Geneva: World Health Organization; 2024.
11. WHO. Consolidated telemedicine implementation guide. World Health Organization. 2022. URL: https://www.who.int/publications/i/item/9789240059184 [accessed 2025-12-26]
12. Binder P, Hjeltnes A. Mindfulness in psychotherapy and society—the need for combining enthusiasm and critical inquiry. Couns and Psychother Res 2021;21(2):247-250 [FREE Full text] [doi: 10.1002/capr.12384]

XSL•FO
RenderX

13. Knowles SE, Lovell K, Bower P, Gilbody S, Littlewood E, Lester H. Patient experience of computerised therapy for depression in primary care. BMJ Open 2015;5(11):e008581 [FREE Full text] [doi: 10.1136/bmjopen-2015-008581] [Medline: 26621513]
14. Andersson G, Carlbring P, Rozental A. Response and remission rates in internet-based cognitive behavior therapy: an individual patient data meta-analysis. Front Psychiatry 2019;10:749. [doi: 10.3389/fpsyt.2019.00749] [Medline: 31708813]
15. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FMG, Furukawa TA, Cuijpers P, Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021;78(4):361-371 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
16. Saad A, Bruno D, Camara B, D'Agostino J, Bolea-Alamanac B. Self-directed technology-based therapeutic methods for adult patients receiving mental health services: systematic review. JMIR Ment Health 2021;8(11):e27404 [FREE Full text] [doi: 10.2196/27404] [Medline: 34842556]
17. Cuijpers P, Noma H, Karyotaki E, Cipriani A, Furukawa TA. Effectiveness and acceptability of cognitive behavior therapy delivery formats in adults with depression: a network meta-analysis. JAMA Psychiatry 2019;76(7):700-707 [FREE Full text] [doi: 10.1001/jamapsychiatry.2019.0268] [Medline: 30994877]
18. Seittu HA, Falk T, Bhatnagar K, Saarni SE. Therapists' role in patient adherence to internet-based cognitive behavioral therapy: qualitative study. J Med Internet Res 2025;27:e71852 [FREE Full text] [doi: 10.2196/71852] [Medline: 40929730]
19. Treanor CJ, Kouvonen A, Lallukka T, Donnelly M. Acceptability of computerized cognitive behavioral therapy for adults: umbrella review. JMIR Ment Health 2021;8(7):e23091 [FREE Full text] [doi: 10.2196/23091] [Medline: 34255714]
20. Fuhr K, Schröder J, Berger T, Moritz S, Meyer B, Lutz W, et al. The association between adherence and outcome in an internet intervention for depression. J Affect Disord 2018;229:443-449. [doi: 10.1016/j.jad.2017.12.028] [Medline: 29331706]
21. Karyotaki E, Kleiboer A, Smit F, Turner DT, Pastor AM, Andersson G, et al. Predictors of treatment dropout in self-guided web-based interventions for depression: an 'individual patient data' meta-analysis. Psychol Med 2015;45(13):2717-2726 [FREE Full text] [doi: 10.1017/S0033291715000665] [Medline: 25881626]
22. Beatty L, Binnion C. A systematic review of predictors of, and reasons for, adherence to online psychological interventions. Int J Behav Med 2016;23(6):776-794. [doi: 10.1007/s12529-016-9556-9] [Medline: 26957109]
23. Karyotaki E, Riper H, Twisk J, Hoogendoorn A, Kleiboer A, Mira A, et al. Efficacy of self-guided internet-based cognitive behavioral therapy in the treatment of depressive symptoms: a meta-analysis of individual participant data. JAMA Psychiatry 2017;74(4):351-359 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.0044] [Medline: 28241179]
24. Kemmeren L, van Schaik A, Draisma S, Kleiboer A, Riper H, Smit J. Effectiveness of blended cognitive behavioral therapy versus treatment as usual for depression in routine specialized mental healthcare: E-COMPARED trial in the Netherlands. Cogn Ther Res 2023;47(3):386-398. [doi: 10.1007/s10608-023-10363-y]
25. Mathiasen K, Andersen TE, Lichtenstein MB, Ehlers LH, Riper H, Kleiboer A, et al. The clinical effectiveness of blended cognitive behavioral therapy compared with face-to-face cognitive behavioral therapy for adult depression: randomized controlled noninferiority trial. J Med Internet Res 2022;24(9):e36577 [FREE Full text] [doi: 10.2196/36577] [Medline: 36069798]
26. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. Annu Rev Clin Psychol 2018;14:91-118. [doi: 10.1146/annurev-clinpsy-032816-045037] [Medline: 29401044]
27. Nie J, Shao H, Fan Y, Shao Q, You H, Preindl M, et al. LLM-based conversational AI therapist for daily functioning screening and psychotherapeutic intervention via everyday smart devices. arXiv. Preprint posted online on March 16, 2024 2025 [FREE Full text] [doi: 10.1145/3712299]
28. Malgaroli M, Hull TD, Zech JM, Althoff T. Natural language processing for mental health interventions: a systematic review and research framework. Transl Psychiatry 2023;13(1):309 [FREE Full text] [doi: 10.1038/s41398-023-02592-2] [Medline: 37798296]
29. Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res 2021;23(5):e15708. [doi: 10.2196/15708] [Medline: 33944788]
30. Sheehan K, Bhatti PK, Yousuf S, Rosenow W, Roehler DR, Hazekamp C, et al. Natural language processing applied to mental illness detection: a narrative review. NPJ Digit Med 2022;5(1):46 [FREE Full text] [doi: 10.1186/s12889-022-13177-x] [Medline: 35449046]
31. Choudhury MD, Pendse SR, Kumar N. Benefits and harms of large language models in digital mental health. arXiv. Preprint posted online on November 7, 2021 2021 [FREE Full text]
32. Villarreal-Zegarra D, Reategui-Rivera CM, García-Serna J, Quispe-Callo G, Lázaro-Cruz G, Centeno-Terrazas G, et al. Self-administered interventions based on natural language processing models for reducing depressive and anxiety symptoms: systematic review and meta-analysis. JMIR Ment Health 2024;11:e59560 [FREE Full text] [doi: 10.2196/59560] [Medline: 39167795]
33. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]

XSL•FO
RenderX

34. Muramatsu K, Miyaoka H, Kamijima K, Muramatsu Y, Tanaka Y, Hosaka M, et al. Performance of the Japanese version of the Patient Health Questionnaire-9 (J-PHQ-9) for depression in primary care. Gen Hosp Psychiatry 2018;52:64-69. [doi: 10.1016/j.genhosppsych.2018.03.007] [Medline: 29698880]
35. Bauer-Staeb C, Kounali DZ, Welton NJ, Griffith E, Wiles NJ, Lewis G, et al. Effective dose 50 method as the minimal clinically important difference: evidence from depression trials. J Clin Epidemiol 2021;137:200-208 [FREE Full text] [doi: 10.1016/j.jclinepi.2021.04.002] [Medline: 33892086]
36. Kounali D, Button KS, Lewis G, Gilbody S, Kessler D, Araya R, et al. How much change is enough? Evidence from a longitudinal study on depression in UK primary care. Psychol Med 2022;52(10):1875-1882 [FREE Full text] [doi: 10.1017/S0033291720003700] [Medline: 33138872]
37. So M, Sekizawa Y, Yamaguchi Y. A randomised controlled trial investigating the clinical and cost-effectiveness of peer enhanced-computerised cognitive depression. KAKEN. 2015. URL: https://kaken.nii.ac.jp/en/grant/KAKENHI-PROJECT-26350862/ [accessed 2025-12-19]
38. Inagaki M, Ohtsuki T, Yonemoto N, Kawashima Y, Saitoh A, Oikawa Y, et al. Validity of the Patient Health Questionnaire (PHQ)-9 and PHQ-2 in general internal medicine primary care at a Japanese rural hospital: a cross-sectional study. Gen Hosp Psychiatry 2013;35(6):592-597. [doi: 10.1016/j.genhosppsych.2013.08.001] [Medline: 24029431]
39. Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. CMAJ 2012;184(3):E191-E196 [FREE Full text] [doi: 10.1503/cmaj.110829] [Medline: 22184363]
40. Cuijpers P, Karyotaki E, Ciharova M, Miguel C, Noma H, Furukawa TA. The effects of psychotherapies for depression on response, remission, reliable change, and deterioration: a meta-analysis. Acta Psychiatr Scand 2021;144(3):288-299. [doi: 10.1111/acps.13335] [Medline: 34107050]
41. Fava M. Depression with physical symptoms: treating to remission. J Clin Psychiatry 2003;64 Suppl 7:24-28. [Medline: 12755649]
42. Keller MB, Lavori PW, Mueller TI, Endicott J, Coryell W, Hirschfeld RM, et al. Time to recovery, chronicity, and levels of psychopathology in major depression. A 5-year prospective follow-up of 431 subjects. Arch Gen Psychiatry 1992;49(10):809-816. [doi: 10.1001/archpsyc.1992.01820100053010] [Medline: 1417434]
43. Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, et al. The 16-item Quick Inventory of Depressive Symptomatology (QIDS), Clinician Rating (QIDS-C), and Self-Report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry 2003;54(5):573-583. [doi: 10.1016/s0006-3223(02)01866-8] [Medline: 12946886]
44. Fujisawa D, Nakagawa A, Tajima M, Ono Y. Development of Japanese version of QIDS-SR (self-report). Jpn J Stress Sci 2010;25(1):43-52 [FREE Full text]
45. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
46. Muramatsu K, Muramatsu Y, Miyaoka H, Fuse K, Yoshimine F, Hosaka M. Validation and utility of a Japanese version of the GAD-7. In: Panminerva Medica. 2009 Presented at: 20th World Congress on Psychosomatic Medicine; August 23-28, 2009; Torino, Italy p. 79.
47. Sheehan KH, Sheehan DV. Assessing treatment effects in clinical trials with the discan metric of the Sheehan Disability Scale. Int Clin Psychopharmacol 2008;23(2):70-83. [doi: 10.1097/YIC.0b013e3282f2b4d6] [Medline: 18301121]
48. Yoshida T, Otsubo T, Tsuchida H, Wada R, Ueshima K, Fukui A. Reliability and validity of the Japanese version of the Sheehan Disability Scale. Jpn J Clin Psychopharmacol 2004;7(10):1645-1653.
49. Soares CN, Zhang M, Boucher M. Categorical improvement in functional impairment in depressed patients treated with desvenlafaxine. CNS Spectr 2019;24(3):322-332 [FREE Full text] [doi: 10.1017/S1092852917000633] [Medline: 29140227]
50. Attkisson CC, Zwick R. The Client Satisfaction Questionnaire. Psychometric properties and correlations with service utilization and psychotherapy outcome. Eval Program Plann 1982;5(3):233-237. [doi: 10.1016/0149-7189(82)90074-x] [Medline: 10259963]
51. Tachimori H, Ito H. Reliability and validity of the Japanese version of the Client Satisfaction Questionnaire. Seishin Igaku (Clin Psychiatry) 1999;41(7):711-717. [doi: 10.11477/mf.1405905056]
52. Wong SYS, Sun YY, Chan ATY, Leung MKW, Chao DVK, Li CCK, et al. Treating subthreshold depression in primary care: a randomized controlled trial of behavioral activation with mindfulness. Ann Fam Med 2018;16(2):111-119 [FREE Full text] [doi: 10.1370/afm.2206] [Medline: 29531101]
53. Harrer M, Sprenger AA, Illing S, Adriaanse MC, Albert SM, Allart E, et al. Psychological intervention in individuals with subthreshold depression: individual participant data meta-analysis of treatment effects and moderators. Br J Psychiatry 2025:1-14 [FREE Full text] [doi: 10.1192/bjp.2025.56] [Medline: 40365980]
54. Donkin L, Hickie IB, Christensen H, Naismith SL, Neal B, Cockayne NL, et al. Rethinking the dose-response relationship between usage and outcome in an online intervention for depression: randomized controlled trial. J Med Internet Res 2013;15(10):e231 [FREE Full text] [doi: 10.2196/jmir.2771] [Medline: 24135213]
55. Donkin L, Christensen H, Naismith SL, Neal B, Hickie IB, Glozier N. A systematic review of the impact of adherence on the effectiveness of e-therapies. J Med Internet Res 2011;13(3):e52 [FREE Full text] [doi: 10.2196/jmir.1772] [Medline: 21821503]

XSL•FO
RenderX

56. Levis B, Benedetti A, Thombs BD, DEPRESSion Screening Data (DEPRESSD) Collaboration. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. BMJ 2019;365:l1476 [FREE Full text] [doi: 10.1136/bmj.l1476] [Medline: 30967483]
57. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatric Annals 2002;32(9):509-515. [doi: 10.3928/0048-5713-20020901-06]
58. Lam RW, Michalak EE, Yatham LN. A new Clinical Rating Scale for work absence and productivity: validation in patients with major depressive disorder. BMC Psychiatry 2009;9:78 [FREE Full text] [doi: 10.1186/1471-244X-9-78] [Medline: 19958540]
59. Luciano JV, Bertsch J, Salvador-Carulla L, Tomás JM, Fernández A, Pinto-Meza A, et al. Factor structure, internal consistency and construct validity of the Sheehan Disability Scale in a Spanish primary care sample. J Eval Clin Pract 2010;16(5):895-901. [doi: 10.1111/j.1365-2753.2009.01211.x] [Medline: 20626541]
60. Sheehan DV, Harnett-Sheehan K, Raj BA. The measurement of disability. Int Clin Psychopharmacol 1996;11 Suppl 3:89-95. [doi: 10.1097/00004850-199606003-00015] [Medline: 8923116]
61. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c332 [FREE Full text] [doi: 10.1136/bmj.c332] [Medline: 20332509]
62. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
63. Eysenbach G. The law of attrition. J Med Internet Res 2005;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
64. Koelen JA, Vonk A, Klein A, de Koning L, Vonk P, de Vet S, et al. Man vs. machine: a meta-analysis on the added value of human support in text-based internet treatments ("e-therapy") for mental disorders. Clin Psychol Rev 2022;96:102179 [FREE Full text] [doi: 10.1016/j.cpr.2022.102179] [Medline: 35763975]
65. Kambeitz-Ilankovic L, Rzayeva U, Völkel L, Wenzel J, Weiske J, Jessen F, et al. A systematic review of digital and face-to-face cognitive behavioral therapy for depression. NPJ Digit Med 2022;5(1):144 [FREE Full text] [doi: 10.1038/s41746-022-00677-8] [Medline: 36109583]
66. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. Front Digit Health 2022;4:847991 [FREE Full text] [doi: 10.3389/fdgth.2022.847991] [Medline: 35480848]
67. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
68. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth 2018;6(11):e12106 [FREE Full text] [doi: 10.2196/12106] [Medline: 30470676]
69. Bickmore T, Gruber A, Picard R. Establishing the computer-patient working alliance in automated health behavior change interventions. Patient Educ Couns 2005;59(1):21-30. [doi: 10.1016/j.pec.2004.09.008] [Medline: 16198215]
70. Prochaska JJ, Vogel EA, Chieng A, Kendra M, Baiocchi M, Pajarito S, et al. A therapeutic relational agent for reducing problematic substance use (woebot): development and usability study. J Med Internet Res 2021;23(3):e24850 [FREE Full text] [doi: 10.2196/24850] [Medline: 33755028]
71. Lynch CP, Cha EDK, Jenkins NW, Parrish JM, Mohan S, Jadczak CN, et al. The minimum clinically important difference for Patient Health Questionnaire-9 in minimally invasive transforaminal interbody fusion. Spine (Phila Pa 1976) 2021;46(9):603-609. [doi: 10.1097/BRS.0000000000003853] [Medline: 33290370]
72. Fairburn CG, Norman PA, Welch SL, O'Connor ME, Doll HA, Peveler RC. A prospective study of outcome in bulimia nervosa and the long-term effects of three psychological treatments. Arch Gen Psychiatry 1995;52(4):304-312. [doi: 10.1001/archpsyc.1995.03950160054010] [Medline: 7702447]
73. Agras WS, Walsh T, Fairburn CG, Wilson GT, Kraemer HC. A multicenter comparison of cognitive-behavioral therapy and interpersonal psychotherapy for bulimia nervosa. Arch Gen Psychiatry 2000;57(5):459-466. [doi: 10.1001/archpsyc.57.5.459] [Medline: 10807486]
74. Carter FA, Jordan J, McIntosh VVW, Luty SE, McKenzie JM, Frampton CMA, et al. The long-term efficacy of three psychotherapies for anorexia nervosa: a randomized, controlled trial. Int J Eat Disord 2011;44(7):647-654. [doi: 10.1002/eat.20879] [Medline: 21997429]
75. Markowitz JC, Petkova E, Neria Y, Van Meter PE, Zhao Y, Hembree E, et al. Is exposure necessary? A randomized clinical trial of interpersonal psychotherapy for PTSD. Am J Psychiatry 2015;172(5):430-440 [FREE Full text] [doi: 10.1176/appi.ajp.2014.14070908] [Medline: 25677355]
76. Bighelli I, Rodolico A, García-Mieres H, Pitschel-Walz G, Hansen WP, Schneider-Thoma J, et al. Psychosocial and psychological interventions for relapse prevention in schizophrenia: a systematic review and network meta-analysis. Lancet Psychiatry 2021;8(11):969-980. [doi: 10.1016/j.envres.2021.112166] [Medline: 34619129]

XSL•FO
RenderX

77. Lemmens LHJM, van Bronswijk SC, Peeters FPML, Arntz A, Roefs A, Hollon SD, et al. Interpersonal psychotherapy versus cognitive therapy for depression: how they work, how long, and for whom-key findings from an RCT. Am J Psychother 2020;73(1):8-14. [doi: 10.1176/appi.psychotherapy.20190030] [Medline: 32122161]
78. Mitchell AJ, Yadegarfar M, Gill J, Stubbs B. Case finding and screening clinical utility of the Patient Health Questionnaire (PHQ-9 and PHQ-2) for depression in primary care: a diagnostic meta-analysis of 40 studies. BJPsych Open 2016;2(2):127-138 [FREE Full text] [doi: 10.1192/bjpo.bp.115.001685] [Medline: 27703765]
79. Levis B, Bhandari PM, Neupane D, Fan S, Sun Y, He C, Depression Screening Data (DEPRESSD) PHQ Group. Data-driven cutoff selection for the Patient Health Questionnaire-9 depression screening tool. JAMA Netw Open 2024;7(11):e2429630 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.29630] [Medline: 39576645]
80. Bentley KH, Gallagher MW, Carl JR, Barlow DH. Development and validation of the Overall Depression Severity and Impairment Scale. Psychol Assess 2014;26(3):815-830. [doi: 10.1037/a0036216] [Medline: 24708078]
81. Norman SB, Cissell SH, Means-Christensen AJ, Stein MB. Development and validation of an Overall Anxiety Severity and Impairment Scale (OASIS). Depress Anxiety 2006;23(4):245-249. [doi: 10.1002/da.20182] [Medline: 16688739]
82. Cuijpers P, Cristea IA, Karyotaki E, Reijnders M, Huibers MJH. How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence. World Psychiatry 2016;15(3):245-258 [FREE Full text] [doi: 10.1002/wps.20346] [Medline: 27717254]
83. Mercorio A, Zizolfi B, Barbuto S, Danzi R, Di Spiezio Sardo A, Moawad G, et al. Three-dimensional imaging reconstruction and laparoscopic robotic surgery: a winning combination for a complex case of multiple myomectomy. Fertil Steril 2023;120(1):202-204 [FREE Full text] [doi: 10.1016/j.fertnstert.2023.04.015] [Medline: 37085096]

Abbreviations

AI: artificial intelligence
AI-iCBT: artificial intelligence–augmented internet-based cognitive behavioral therapy
CBT: cognitive behavioral therapy
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
CR: cognitive restructuring
CSQ-8: Client Satisfaction Questionnaire-8
EAS: efficacy analysis set
GAD-7: Generalized Anxiety Disorder-7
GLMM: generalized linear mixed models
iCBT: internet-based cognitive behavioral therapy
ITT: intention-to-treat
MMRM: mixed-effects model for repeated measures
NLP: natural language processing
OR: odds ratio
PHQ-9: Patient Health Questionnaire-9
QIDS-J: Quick Inventory of Depressive Symptomatology-Japanese version
SDS: Sheehan Disability Scale

Edited by A Schwartz; submitted 03.May.2025; peer-reviewed by N Titov; comments to author 02.Jun.2025; accepted 21.Nov.2025; published 05.Jan.2026.

Please cite as:
So M, Sekizawa Y, Hashimoto S, Kashimura M, Yamakage H, Watanabe N
Effect of AI-Based Natural Language Feedback on Engagement and Clinical Outcomes in Fully Self-Guided Internet-Based Cognitive Behavioral Therapy for Depression: 3-Arm Randomized Controlled Trial
J Med Internet Res 2026;28:e76902
URL: https://www.jmir.org/2026/1/e76902
doi:10.2196/76902
PMID:

©Mirai So, Yoichi Sekizawa, Sora Hashimoto, Masami Kashimura, Hajime Yamakage, Norio Watanabe. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted

XSL•FO
RenderX

use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Leisure Screen Time, Internet Gaming Disorder, and Mental Health Among Chinese Adolescents: Large-Scale Cross-Sectional Study

Qin Deng[1], MPH; Linna Sha[1], MPH; Jiaojiao Hou[2], MPH; Xunying Zhao[2], MD; Rong Xiang[1], MD; Jiangbo Zhu[1], MPH; Yang Qu[2], MD; Jinyu Zhou[1], MD; Ting Yu[2], MD; Xin Song[1], MD; Sirui Zheng[1], MPH; Tao Han[1], MPH; Bin Yang[1], MD; Mengyu Fan[2*], MD; Xia Jiang[1,2,3*], MD, PhD

[1]Department of Nutrition and Food Hygiene, West China School of Public Health and West China Fourth Hospital, Sichuan University, Chengdu, China

[2]Department of Epidemiology and Biostatistics, West China School of Public Health and West China Fourth Hospital, Sichuan University, Chengdu, China

[3]Department of Clinical Neuroscience, Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden

[*]these authors contributed equally

Corresponding Author:
Xia Jiang, MD, PhD
Department of Clinical Neuroscience, Center for Molecular Medicine
Karolinska Institutet
L8:05, CMM Karolinska Universitetssjukhuset Solna
Stockholm, 17176
Sweden
Phone: 86 15528103968
Email: xia.jiang@ki.se

Abstract

Background: Adolescence is a critical period for mental health vulnerability alongside rising digital media exposure. Current evidence often fails to distinguish the distinct roles of leisure screen time (LST) quantity and addictive patterns like internet gaming disorder (IGD) on a comprehensive range of mental health outcomes.

Objective: This study aimed to investigate the independent and joint associations of LST and IGD with multiple mental health conditions among Chinese adolescents.

Methods: We conducted a school-based, cross-sectional survey in Sichuan Province, China. Participants were recruited by random cluster sampling from 20 public schools. The sample comprised 13,240 adolescents (6659/13,240, 50.3% girls) with a mean age of 15.4 (SD 1.6) years. LST was self-reported, and IGD was evaluated using the Internet Gaming Disorder Scale-9 Item Short Form (IGDS9-SF). Mental health outcomes included overall mental health status and 5 specific diseases: psychological distress, depression, paranoia, insomnia, and suicidal ideation, all assessed using validated scales.

Results: The prevalence of excessive LST, IGD, and any mental health disorder was 48.2% (6378/13,240; 95% CI 47.3%-49.0%), 1.4% (188/13,240; 95% CI 1.2%-1.6%), and 55.8% (7387/13,240; 95% CI 54.9%-56.7%), respectively. After adjustment, excessive LST (odds ratio [OR] 1.18, 95% CI 1.09-1.27) and IGD (OR 6.58, 95% CI 5.02-8.62) were independently associated with poor mental health. A dose-response relationship existed for LST quartiles (Q2: OR 1.15, 95% CI 1.04-1.26; Q3: OR 1.24, 95% CI 1.12-1.37; Q4: OR 1.31, 95% CI 1.18-1.46; P_{trend}<.001). Excessive LST was associated with depression (OR 1.16, 95% CIs 1.05-1.29), paranoia (OR 1.22, 95% CI 1.11-1.34), and suicidal ideation (OR 1.15, 95% CI 1.04-1.28), while IGD was associated with all 5 disorders, most notably depression (OR 6.43, 95% CI 4.56-9.06) and paranoia (OR 5.77, 95% CI 4.05-8.21). IGD consistently demonstrated stronger associations than LST: psychological distress (OR 4.40, 95% CI 3.12-6.19 vs OR 1.14, 95% CI 0.98-1.33), depression (OR 6.43, 95% CI 4.56-9.06 vs OR 1.16, 95% CI 1.05-1.29), paranoia (OR 5.77, 95% CI 4.05-8.21 vs OR 1.22, 95% CI 1.11-1.34), insomnia (OR 2.90, 95% CI 2.09-4.05 vs OR 1.12, 95% CI 102-1.22), and suicidal ideation (OR 3.85, 95% CI 2.76-5.37 vs OR 1.15, 95% CI 1.04-1.28). Adolescents with both excessive LST and IGD demonstrated the highest odds of mental health disorders (OR 7.35, 95% CI 5.29-10.22). No significant interaction was found on additive or multiplicative scales.

Conclusions: Both excessive LST and IGD are independently associated with mental health disorders in adolescents, with IGD showing a substantially stronger association. This study is distinct from prior research by simultaneously investigating both screen

XSL•FO
RenderX

time quantity and addictive usage patterns, and by comprehensively assessing 5 distinct mental health outcomes. Longitudinal studies are needed to better understand the long-term effects.

(J Med Internet Res 2026;28:e80737) doi:10.2196/80737

KEYWORDS

adolescent; cross-sectional study; internet gaming disorder; leisure screen time; mental health; screen media activity

Introduction

Adolescence represents a critical period for psychological development, as up to half of all mental health conditions start before age 14 years [1,2]. Globally, 1 in 7 adolescents experience at least one mental health disorder, contributing to 15% of the overall disease burden in this age group [3]. Depression and anxiety are among the leading causes of illness and disability, with suicide being the fourth leading cause of death for those aged 15-19 years [3]. Additionally, psychological distress [4], paranoia [5], and insomnia [6] are also commonly prevalent. Poor mental health at a relatively early stage poses significant adverse effects that can span from the short to the long term [7], including school disengagement, reduced quality of life, and increased mortality. The widespread underrecognition and treatment delays—where only 15.4% of adolescents seek prompt professional help after their initial request—further exacerbate these issues [8].

Alongside the rise in mental health disorders over the years, time spent on screen-based activities continues to increase by as much as 2 hours per day since 2010, particularly among adolescents [9,10]. Data from the Adolescent Brain Cognitive Development Study reveal that adolescents now spend an average of over 5.5 hours daily on noneducational screen media, underscoring a global trend of pervasive digital engagement [11]. While moderate screen use may offer cognitive benefits such as enhanced multitasking [12], excessive use is linked to sedentary behaviors, impaired parent–child interactions, and behavioral addictions [13,14]. Each additional hour of daily screen time in late childhood predicted increased depressive symptoms in early adolescence, an effect mediated by shorter sleep duration and changes in white matter organization in brain regions responsible for emotion regulation [15]. Current research often conflates educational and leisure screen time (LST), obscuring the unique risk of LST [16,17]. A dose-response study of Chinese adolescents revealed that LST exceeding 1 hour/day linearly increases mental health problems, with video-based and gaming content showing the strongest negative associations [18]. Nevertheless, existing studies have mainly investigated depression and anxiety, the 2 most prevalent mental health issues, while largely disregarding other important disorders such as paranoia and insomnia that affect a nontrivial proportion (20%-30%) of adolescents [5,17].

The amount of leisure time spent using screen-based media does not necessarily represent the extent of harm. Crucially, recent research underscores that addictive use patterns—characterized by compulsive use, loss of control, and continued use despite negative consequences—may represent a far greater risk than screen time duration alone [14]. Studies found that a small number of individuals, despite showing shorter screen time, exhibited characteristics of addictive behaviors [19]. A primary concern for adolescents is gaming addiction, particularly the internet gaming disorder (IGD), defined by *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]) as the persistent and recurrent use of internet to play games, often with others. IGD leads to clinically significant impairment or distress and has been listed as a tentative disorder in need of further study [20]. IGD has been linked to distinct neurobiological changes and impairments in executive function, underscoring its status as a behavioral addiction beyond mere excessive use [21]. Interestingly, a substantial proportion of individuals with long screen time do not meet the symptomatic criteria of IGD, while those who meet criteria do not necessarily show long screen time [22,23]. This dissociation underscores the need to examine both behaviors simultaneously.

The interplay between LST and IGD in relation to a broad spectrum of mental health outcomes remains inadequately explored. While longitudinal evidence suggests that internalizing symptoms (eg, depression, anxiety) fully mediate the link between problematic internet use and subsequent self-harm behaviors, this model did not account for the joint influence of general screen time [24]. Furthermore, seminal research has demonstrated that trajectories of addictive digital media use are independently associated with heightened risks of suicidality, regardless of baseline screen time [14]. Moreover, while high LST and IGD independently poses negative effects on mental health, it is likely that those who exhibit both behaviors are at the highest risk. However, few studies have analyzed leisure and addictive screen-based media use by taking into consideration their independent and joint effects on mental health outcomes.

This cross-sectional study aims to comprehensively investigate the independent and joint associations of LST and IGD with an array of mental health among Chinese adolescents, including overall mental health and its 5 major diseases, namely, depression, psychological distress, paranoia, insomnia, and suicidal ideation.

Methods

Ethical Considerations

The research was conducted in accordance with the Declaration of Helsinki and received approval from the Ethics Committee of West China Fourth Hospital and West China School of Public Health, Sichuan University (Gw112023133; June 13, 2023). Written informed consent was obtained from parents or guardians after they received a detailed information sheet outlining the study purpose, procedures, and voluntary nature of participation. Student assent was also secured at the time of the survey. No compensation was offered to participants for

XSL•FO
RenderX

their involvement in this study. Beyond a lack of financial compensation, the study also prioritized participant anonymity by eliminating all personally identifiable information from the collected data.

Additionally, this study did not collect any personally identifiable information, such as facial images, voice recordings, or other biometric data. All data presented in this manuscript and supplementary materials are fully anonymized, ensuring that no individual participant can be identified.

Study Population

This cross-sectional study analyzed data from a survey project on the mental health of children and adolescents in Pidu District, Chengdu, conducted from June to December 2023 by the West China School of Public Health and the West China Fourth Hospital, Sichuan University. We adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines in reporting cross-sectional studies [25].

The survey aimed to gain a comprehensive understanding of the mental health status of children and adolescents residing in Pidu District and to effectively address the mental health issues affecting this population. To ensure the representativeness of study samples, we used a random cluster sampling design based on the number and proportion of primary and secondary students in Pidu District, considering types of school and their urban–rural distribution. A total of 20 schools were selected, comprising 17,973 registered students for the questionnaire survey.

Data collection was conducted in classroom settings during regular school hours. Trained research staff administered structured, paper-based questionnaires to all participants. To ensure standardized procedures, staff delivered standardized verbal instructions to students before distributing the questionnaires, detailing the study's purpose, its voluntary nature, confidentiality protocols, and questionnaire completion guidelines. Participants then completed the surveys independently. Research staff supervised the process and were available to provide neutral clarification on questions but refrained from influencing participant responses. Of the 16,482 participants invited, a total of 16,325 were present on the survey day and completed the questionnaire, resulting in a response rate of 98.99%. Participant eligibility was restricted to Han Chinese adolescents aged 12-18 years. Ethnicity was self-identified in the questionnaire, and age was calculated from the self-reported date of birth and ensured to be within the specified range at the time of data collection. After applying the eligibility criteria, a total of 13,240 participants were included in the final analytical sample. The high response rate minimized concerns about selection bias.

Assessment of LST and IGD

LST was measured using a self-reported questionnaire. Participants were asked to recall and report the average time they spent on leisure screen activities (including entertainment-based use of smartphones, computers, tablets, and televisions) on a typical day. The average daily LST was calculated using the formula: ([Weekday LST × 5] + [Weekend LST × 2])/7. In this study, LST was categorized into 2 groups according to World Health Organization standards, low (<2 h/day) and high (≥2 h/day) [26], and further into 4 groups based on quartiles, Q1 (≤0.93 h/day), Q2 (0.93 <LST ≤1.86 h/day), Q3 (1.86 <LST ≤4.19 h/day), and Q4 (>4.19 h/day).

IGD was assessed using the Internet Gaming Disorder Scale-9 Item Short Form (IGDS9-SF) [27], consisting of 9 items reflecting the diagnostic criteria for IGD as defined by the American Psychological Association. Participants responded to each item through a Likert scale ranging from 1 ("never") to 5 ("very often"). Total scores, ranging from 9 to 45, were calculated by summing across the responses. This threshold was identified as clinically optimal for the Chinese population in a previous validation study [28], which, although conducted in a sample with a mean age of 20 years, remains one of the most established cutoffs for the IGDS9-SF in the absence of a universally validated threshold for early- to mid-adolescents. We selected this cutoff to ensure consistency with prior research and because of its high specificity, which minimizes false positives. Furthermore, the scale demonstrated adequate internal consistency in our adolescent sample (Cronbach α=0.88), supporting its reliability for this age group.

To identify the joint association of the 2 types of screen-based behaviors, we defined a combined pattern based on the level of LST (low vs high) and status of IGD (absence vs presence). We designated the group with LST of less than 2 hours and without IGD as the reference group, referred to as "Low No." The other 3 comparison groups were "High No," "Low Yes," and "High Yes."

Assessment of Mental Health Disorders

We applied well-established validated scales to investigate the 5 common mental health disorders among adolescents. Specifically, psychological distress was measured using the Kessler 6-item (K6) scale, which assesses the frequency of psychopathological symptoms or behaviors. A K6 score of ≥13 indicates psychological distress [29,30]. Depression was evaluated using the Kutcher Adolescent Depression 6-item (KADS-6) scale, specifically designed for children and adolescents to effectively identify major depressive episodes. A KADS-6 score of ≥6 indicates depression [31]. Paranoia was assessed using the psychoticism and paranoid ideation subscales of the Symptom Checklist-90 (SCL-90), with a score of ≥12 indicating paranoid ideation [32]. Insomnia was measured using the Insomnia Severity Index (ISI) scale, with a score of ≥8 indicating sleep disturbances [33]. Suicidal ideation was assessed using the Suicide Behaviors Questionnaire-Revised (SBQ-R) scale, with a score of ≥8 reflecting risk of suicidal behaviors [34]. Adolescents exhibiting any (or more) of these 5 symptoms were considered to have mental health disorders. The internal consistency of all scales used in this study appeared to be robust, with Cronbach α values being 0.88 (K6), 0.85 (KADS-6), 0.81 (SCL-90), 0.88 (ISI), and 0.84 (SBQ-R), respectively.

Assessment of Covariates

This analysis included a comprehensive set of variables that may potentially confound the association between screen-based

XSL•FO
RenderX

behaviors and adolescent mental health (Table S1 in Multimedia Appendix 1). Demographic variables included age, sex, area of residence, economic status, single-child status, type of caretakers, and parental educational level. The area of residence was classified as "urban" or "rural." Economic status was categorized into "high-income" or "low- to middle-income." Single-child status was determined based on the presence of siblings ("yes" or "no"). Caretaker type was classified as "both parents and others" or "one parent." Parental educational level was defined by the highest education attained: "junior school and below," "high school or vocational," and "college and above." Anthropometric measures included BMI (kg/m²), calculated from height and weight. Lifestyle factors included smoking status, drinking status, academic performance, physical health status, dietary habits, and physical activity (PA) level. Smoking and drinking were self-reported as "yes" or "no." Academic performance and physical health status were categorized as "low," "moderate," or "high." Dietary habits—including breakfast, vegetables and fruits, protein consumption, sugary beverages, desserts, and fried food—were divided into "never eat," "1-2 times per week," "3-4 times per week," "5-6 times per week," and "eat every day." PA levels were categorized based on daily durations into 4 quartiles: Q1 (≤1.80 hours), Q2 (1.81-4.80 hours), Q3 (4.81-9.50 hours), and Q4 (>9.50 hours).

Statistical Analysis

The missing data pattern was evaluated using the Little's Missing Completely at Random test. A nonsignificant result (χ^2_{8836}=4638.4; P>.05) confirmed that the data were missing completely at random, thereby justifying the application of multiple imputation. Consequently, the Multiple Imputation by Chained Equations approach was used to handle the missing values, given the low overall missing rate (maximum: 5.7%) [35]. We generated m=5 imputed datasets using predictive mean matching with a maximum of maxit=50 iterations, considered sufficient to achieve stability for datasets with low missing rates. Convergence was assessed by ensuring the algorithm completed the specified iterations without issues. To validate the imputation, we conducted analyses using both raw and imputed datasets, considering results statistically significant only if they were consistent across both methods. Descriptive summaries were presented as mean (SD) for continuous variables and as quantity (proportion) for categorical variables. Differences in continuous and categorical variables between groups (individuals with or without mental health disorders) were assessed using t tests (2-tailed) and chi-square tests, as appropriate.

To assess multicollinearity between LST and IGD, we used Pearson correlation coefficients (r) with values greater than 0.5 and variance inflation factor values exceeding 10 as diagnostic tests. The variance inflation factor value was found to be less than 1.02, indicating that multicollinearity was of minimal concern.

Multivariable logistic regression was used to examine the associations between screen-based behaviors and mental health among adolescents, controlling for confounders. First, we assessed the independent association of LST (high vs low) and IGD (presence vs absence) with overall mental health condition, making mutual adjustment for each other. Meanwhile, we also categorized LST into quartiles based on its duration (Q1-Q4) and used the Cochran-Armitage test to evaluate its dose-response effect [36]. Second, we investigated the joint effects and interactions of LST and IGD on overall mental health condition, adopting the Cochran-Armitage test to evaluate a linear trend. Finally, subgroup analyses were conducted by sex, area, economic status, single-child status, and type of caretakers. Interaction between screen-based behaviors and stratification factors was performed using the likelihood ratio test comparing models with and without a cross-product term [37]. Such a comprehensive analytical framework was further applied in parallel to each specific mental health disorder. In addition, to account for the cluster sampling design (participants nested within 20 schools), we used multilevel mixed-effects models with random intercepts for schools as sensitivity analyses.

Statistical analysis were performed using R (version 4.3.3; The R Core Team). Association estimates were presented in the form of odds ratios (ORs) and their 95% CIs. To minimize type I error, Bonferroni correction was applied, with a 2-sided P value <.05/5 considered statistically significant for the 5 mental health disorders. Trend tests and interaction P values were defined as statistically significant at P<.05.

Results

Participant Characteristics

The main characteristics of participants according to the status of mental health disorders are presented in Table S1 in Multimedia Appendix 1. This study included 13,240 adolescents (6659/13,240, 50.3% girls) with a mean age of 15.4 (SD 1.6) years. Among these adolescents, 60% (7947/13,240; 95% CI 59.1%-60.9%) lived in the rural areas, and a majority (12,264/13,240, 92.5%; 95% CI 92.0%-93.0%) were from low- to middle-income families. More than half of the participants (8543/13,240, 64%; 95% CI 63.2%-64.8%) had no siblings, and a significant proportion of parents (55.3%-58.9%) received low levels of education.

Regarding the outcomes, a total of 7391 adolescents (55.8%; 95% CI 54.9%-56.7%) were identified as having at least 1 of the 5 mental health disorders. Insomnia was the most prevalent, affecting 35.3% (4677/13,240; 95% CI 34.5%-36.1%) of participants, followed by paranoia (3670/13,240, 28.4%; 95% CI 27.6%-29.2%), suicidal ideation (3155/13,240, 23.8%; 95% CI 23.0%-24.6%), depression (2733/13,240, 20.6%; 95% CI 19.9%-21.3%), and psychological distress (1091/13,240, 8.2%; 95% CI 7.7%-8.7%).

Regarding the exposures, a total of 6378 adolescents (48.2%; 95% CI 47.3%-49.0%) had LST exceeding 2 hours per day, and 188 individuals (1.4%; 95% CI 1.2%-1.6%) met the criteria for IGD. Notably, among those with IGD (n=188), a total of 58 individuals (30.9%; 95% CI 24.3%-38.1%) had LST of 2 hours or less.

The prevalence of any mental health condition was higher among females, those with single parents or other caretakers, those with low academic performance, smokers, drinkers, those

XSL•FO
RenderX

with poor physical health, those who rarely consumed breakfast, fruits, and vegetables or protein, as well as those who frequently consumed unhealthy foods (sugary beverages, sweets, and fried food) (P<.05 for all).

Independent Association of LST and IGD on Mental Health Conditions

Results on the associations of each specific screen-based behavior with mental health disorders are shown in Table 1. After adjusting for potential confounders, LST >2 hours/day (OR 1.18, 95% CI 1.09-1.27) and the presence of IGD (OR 6.58, 95% CI 5.02-8.62) were significantly and independently associated with poorer mental health status. In addition, the odds of overall mental health issues showed a marked increase as LST increased (Q2: OR 1.15, 95% CI 1.04-1.26; Q3: OR 1.24, 95% CI 1.12-1.37; Q4: OR 1.31, 95% CI 1.18-1.46; P_{trend}<.001) (Table S2 in Multimedia Appendix 1).

Table 1. Independent associations of leisure screen time and internet gaming disorder with mental health disorders among Chinese adolescents (P<.05/5).

Mental health outcome	Leisure screen time (h/day)			Internet gaming disorder		
	<2	≥2		No	Yes	
	OR[a,b]	OR (95% CI)	P value	OR[c]	OR (95% CI)	P value
Overall mental health						
Model 1[d]	Ref	1.56 (1.45-1.67)	<.001	Ref	11.24 (8.69-14.55)	<.001
Model 2[e]	Ref	1.19 (1.11-1.28)	<.001	Ref	6.68 (5.10-8.75)	<.001
Model 3[f]	Ref	1.18 (1.09-1.27)	<.001	Ref	6.58 (5.02-8.62)	<.001
Psychological distress						
Model 1	Ref	1.61 (1.40-1.85)	<.001	Ref	7.86 (5.77-10.71)	<.001
Model 2	Ref	1.17 (1.01-1.37)	.04	Ref	4.47 (3.18-6.29)	<.001
Model 3	Ref	1.14 (0.98-1.33)	.09	Ref	4.40 (3.12-6.19)	<.001
Depression						
Model 1	Ref	1.57 (1.43-1.73)	<.001	Ref	9.98 (7.29-13.66)	<.001
Model 2	Ref	1.19 (1.07-1.32)	.002	Ref	6.52 (4.63-9.18)	<.001
Model 3	Ref	1.16 (1.05-1.29)	.005	Ref	6.43 (4.56-9.06)	<.001
Paranoia						
Model 1	Ref	1.52 (1.39-1.65)	<.001	Ref	8.38 (5.99-11.72)	<.001
Model 2	Ref	1.24 (1.13-1.36)	<.001	Ref	5.88 (4.13-8.36)	<.001
Model 3	Ref	1.22 (1.11-1.34)	<.001	Ref	5.77 (4.05-8.21)	<.001
Insomnia						
Model 1	Ref	1.41 (1.30-1.53)	<.001	Ref	4.34 (3.18-5.93)	<.001
Model 2	Ref	1.13 (1.03-1.23)	.01	Ref	2.94 (2.11-4.09)	<.001
Model 3	Ref	1.12 (1.02-1.22)	.01	Ref	2.90 (2.09-4.05)	<.001
Suicidal ideation						
Model 1	Ref	1.52 (1.39-1.67)	<.001	Ref	6.33 (4.67-8.58)	<.001
Model 2	Ref	1.17 (1.06-1.29)	.003	Ref	3.90 (2.80-5.45)	<.001
Model 3	Ref	1.15 (1.04-1.28)	.01	Ref	3.85 (2.76-5.37)	<.001

[a]OR: odds ratio.

[b]The reference group for leisure screen time was <2 h/day.

[c]The reference group for internet gaming disorder was "No."

[d]Model 1 (partially adjusted model) was adjusted for age, sex, area, and economic status.

[e]Model 2 (fully adjusted model) was adjusted for age, sex, area, economic status, single child, caretaker, father's educational level, mother's educational level, BMI, smoking, drinking, academic performance, health status, dietary factors, and physical activity level.

[f]Model 3 adjusted for the covariates included in Model 2 and additionally included mutual adjustment for leisure screen time and internet gaming disorder.

For specific disorders, excessive LST was associated with significantly higher odds of depression (OR 1.16, 95% CI 1.05-1.29), paranoia (OR 1.22, 95% CI 1.11-1.34), and suicidal ideation (OR 1.15, 95% CI 1.04-1.28). Similarly, IGD was

XSL•FO
RenderX

significantly associated with psychological distress (OR 4.40, 95% CI 3.12-6.19), depression (OR 6.43, 95% CI 4.56-9.06), paranoia (OR 5.77, 95% CI 4.05-8.21), insomnia (OR 2.90, 95% CI 2.09-4.05), and suicidal ideation (OR 3.85, 95% CI 2.76-5.37).

In subgroup analysis, no factor was found to significantly modify the associations between LST and any mental health condition, nor between IGD and any mental health condition (all $P_{interaction}$>.05) (Figure 1 and Tables S7 and S8 in Multimedia Appendix 1).

Figure 1. Subgroup analysis of the associations of screen-based behaviors with mental health disorders in Chinese adolescents. (A) Association between leisure screen time (LST; h/day) and mental health; (B) association between internet gaming disorder (IGD) and mental health.

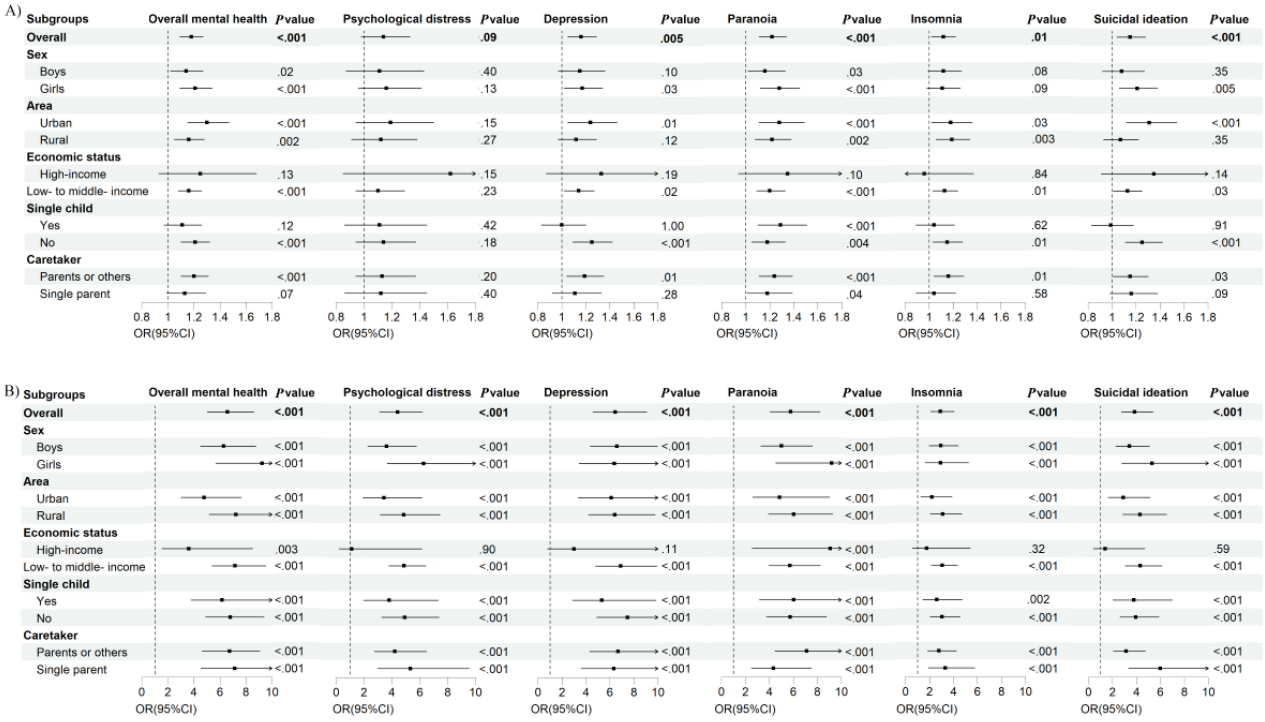

Joint Associations of LST and IGD on Mental Health Conditions

We further examined the joint associations of LST and IGD with the outcomes, as presented in Table 2 and Figure 2. The odds of overall mental health conditions were significantly higher among adolescents who exhibited at least one type of screen-based behavior (High No: OR 1.18, 95% CI 1.09-1.27; Low Yes: OR 7.34, 95% CI 4.59-11.74). Notably, the highest odds were observed when both excessive LST and symptoms of IGD were present simultaneously (High Yes: OR 7.35, 95% CI 5.29-10.22). However, no significant interaction between LST and IGD, either on the additive scale (relative excess risk due to interaction=–0.77, 95% CI –22.05 to 17.02; attributable proportion=–0.01, 95% CI –2.86 to 0.69; synergy index=0.99, 95% CI 0.23-4.26) or on the multiplicative scale (OR 0.88, 95% CI 0.23-3.37), was observed (Table 3). A similar pattern of results was observed with each of the 5 individual mental health outcomes.

XSL•FO
RenderX

Table 2. Combined effect of leisure screen time (LST) and internet gaming disorder (IGD) on mental health (*P*<.05/5). Odds ratios (ORs) of the joint associations were not directly comparable to those of the independent associations as shown in as they were derived from different models with different reference groups and adjustment strategies.

Mental health outcome	LST_IGD						
	Low No	High No		Low Yes		High Yes	
	OR[a]	OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Overall mental health	Ref	1.18 (1.09-1.27)	<.001	7.34 (4.59-11.74)	<.001	7.35 (5.29-10.22)	<.001
Psychological distress	Ref	1.14 (0.98-1.34)	.09	4.94 (2.63-9.28)	<.001	4.81 (3.19-7.27)	<.001
Depression	Ref	1.17 (1.05-1.30)	.004	7.78 (4.32-14.00)	<.001	6.80 (4.46-10.37)	<.001
Paranoia	Ref	1.23 (1.12-1.35)	<.001	8.96 (4.70-17.11)	<.001	5.74 (3.77-8.75)	<.001
Insomnia	Ref	1.12 (1.02-1.22)	.01	2.71 (1.53-4.77)	<.001	3.36 (2.23-5.07)	<.001
Suicidal ideation	Ref	1.15 (1.04-1.28)	.006	3.81 (2.14-6.79)	<.001	4.46 (2.96-6.73)	<.001

[a]Adjusted for age, sex, area, economic status, single child, caretaker, father's educational level, mother's educational level, BMI, smoking, drinking, academic performance, health status, dietary factor and physical activity level.

XSL•FO
RenderX

Figure 2. Combined effect of leisure screen time (LST) and internet gaming disorder (IGD) on mental health disorders among Chinese adolescents.

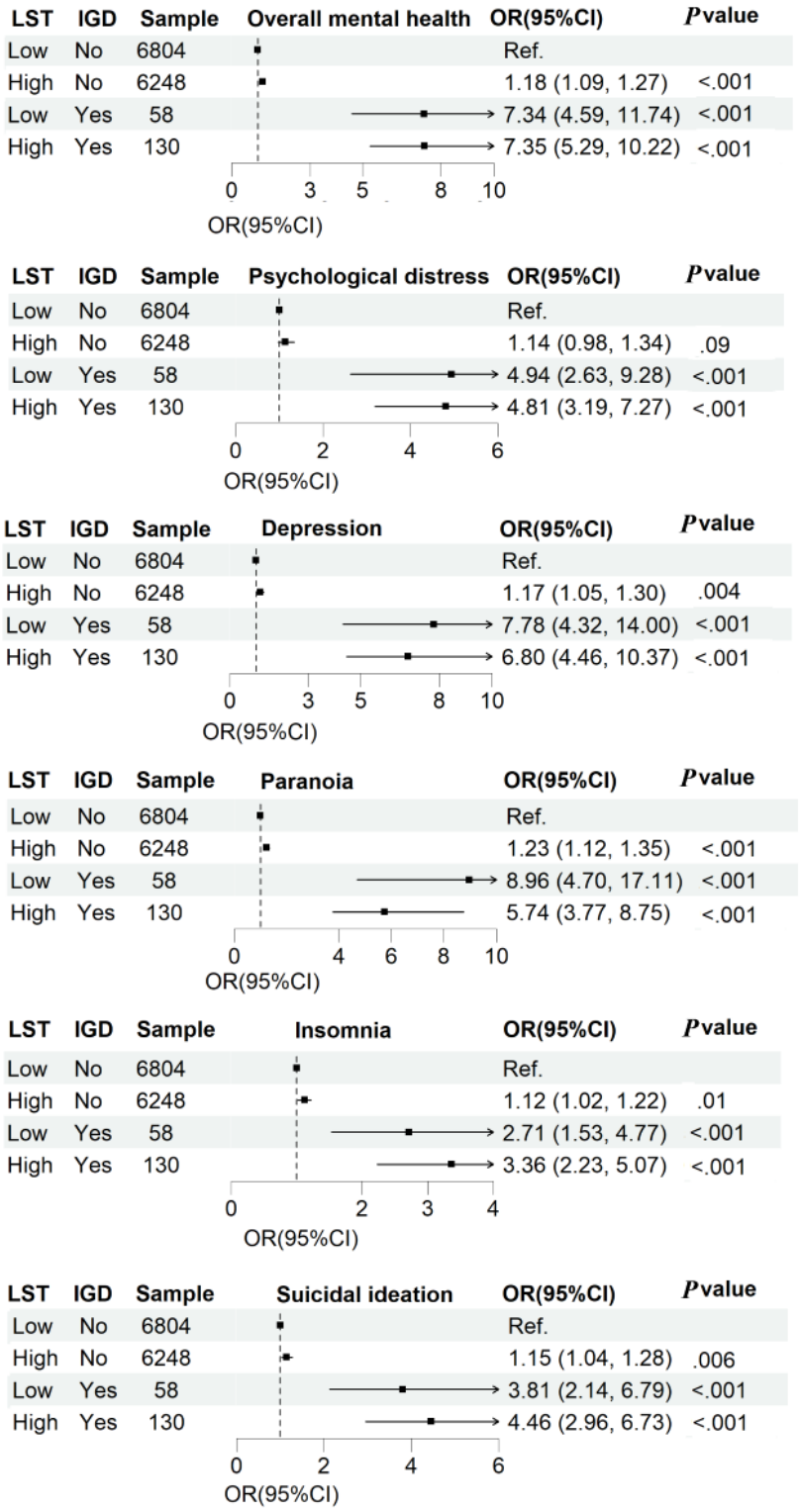

XSL•FO
RenderX

Table 3. Interaction analysis between leisure screen time and internet gaming disorder on mental health disorders among Chinese adolescents.

Mental health outcome	Additive interaction			Multiplicative interaction
	RERI[a] (95% CI)	AP[b] (95% CI)	S[c] (95% CI)	OR[d,e] (95% CI)
Overall mental health	–0.77 (–22.05 to 17.02)	–0.01 (–2.86 to 0.69)	0.99 (0.23 to 4.26)	0.88 (0.23 to 3.37)
Psychological distress	–0.27 (–4.83 to 3.00)	–0.06 (–1.20 to 0.47)	0.93 (0.38 to 2.30)	0.85 (0.40 to 1.79)
Depression	–1.14 (–7.73 to 3.75)	–0.17 (–1.38 to 0.41)	0.84 (0.37 to 1.88)	0.75 (0.36 to 1.54)
Paranoia	–3.45 (–11.80 to 1.73)	–0.60 (–2.43 to 0.22)	0.58 (0.24 to 1.37)	0.52 (0.24 to 1.12)
Insomnia	0.54 (–1.80 to 2.58)	0.16 (–0.69 to 0.54)	1.30 (0.47 to 3.58)	1.11 (0.55 to 2.23)
Suicidal ideation	0.50 (–2.81 to 3.28)	0.11 (–0.79 to 0.53)	1.17 (0.47 to 2.88)	1.17 (0.47 to 2.88)

[a]RERI: relative excess risk due to interaction.

[b]AP: attributable proportion due to interaction.

[c]S: synergy index.

[d]OR: odds ratio.

[e]Adjusted for age, sex, area, economic status, single child, caretaker, father's educational level, mother's educational level, BMI, smoking, drinking, academic performance, health status, dietary factor and physical activity level.

Repeating all analyses using raw data, we obtained consistent findings (Tables S4-S6 and S9-S12 in Multimedia Appendix 1). Sensitivity analyses using multilevel models accounting for school clustering yielded results consistent with primary findings, with all significant associations remaining unchanged. Intraclass correlation coefficients (0.76%-2.08%) confirmed that most variance (>98%) occurred at the individual level, supporting the robustness of results (Tables S18 and S19 in Multimedia Appendix 1).

Discussion

Principal Findings

In this large-scale cross-sectional study of more than 10,000 Chinese adolescents, 48.2% (6378/13,240) of the study population reported excessive LST, and 1.4% (188/13,240) met the criteria for IGD. Both excessive LST (>2 h/day) and IGD were independently associated with increased odds of overall mental health disorders and its 5 major conditions: psychological distress, depression, paranoia, insomnia, and suicidal ideation. The association was markedly stronger for IGD than for LST. The observed dose-response relationship across LST quartiles, with progressively higher ORs across LST quartiles, provides compelling evidence for a graded association in which increasing screen time exposure corresponds to incrementally worse mental health outcomes. Furthermore, while a significant joint association was observed—where adolescents exhibiting both behaviors demonstrated a 7-fold rise in the probability of mental health disorders—no statistically significant interaction was found, indicating an additive rather than synergistic effect. The precision of the main effect estimates further strengthens the robustness of our findings.

These findings necessitate an expansion of current public health guidelines, which predominantly target screen time duration, to include specific strategies for early identification and management of IGD. This study makes a critical contribution by clearly establishing IGD as a severe and distinct factor, independent of mere usage quantity. Methodologically, the innovative joint-association analysis revealing their additive relationship provides a more sophisticated framework for assessing digital media behaviors on mental health.

Prevalence of LST and IGD Among Adolescents

Approximately 48.2% (6378/13,240) of adolescents in this study reported LST >2 hours/day, surpassing the rate reported in a recent study conducted in Guangzhou, China (20.7%) [38]. The prevalence of IGD in this study was only 1.4% (188/13,240), much lower than the global rate of 8.8% [39]. Such discrepancy may be attributed to several factors. First, it likely reflects our use of a conservative, high-specificity cutoff score (≥32 on IGDS9-SF) to minimize false positives, whereas many epidemiological studies use lower, more sensitive thresholds that yield higher prevalence rates. Second, our sample was drawn from a general school-based population in a specific sociocultural context (Pidu District, Chengdu), which may exhibit different characteristics compared to the more diverse and potentially higher-risk samples included in the global meta-analyses.

Despite the lower prevalence, the large overall sample size (N=13,240) ensured that this study was sufficiently powered to detect strong and statistically significant associations, as evidenced by the strength and consistency of the association between IGD and poorer mental health. It is also important to note that the large population size of China means that even a 1.4% prevalence translates to a substantial absolute number of adolescents affected by IGD, underscoring the importance for policymakers and guardians to closely monitor screen-based behaviors among adolescents. Of note, we found that a large proportion of adolescents meeting the symptomatic criteria for IGD (58/188, 30.9%) reported LST of less than 2 hours/day; conversely, a substantial proportion of participants who did not meet the criteria for IGD (6248/13,052, 47.9%) reported excessive LST. These findings further illustrate that spending a large amount of time on gaming or using media does not necessarily equal to addictive use, and vice versa.

Adolescents living with a single parent or without siblings were more likely to engage in excessive LST. This may stem from reduced supervision in single-parent households or absence of

XSL•FO
RenderX

siblings that limit the opportunities for outdoor or other non–screen-based activities, thereby contributing to excessive LST. In addition, boys were more likely than girls to exhibit symptoms of IGD. Boys typically prefer to participate in active and combative activities, whereas girls tend to engage more in social and conversational activities. These preferences may explain why boys are more often drawn to the virtual worlds of online games [40].

The Association of LST and IGD With Mental Health Conditions

Our findings indicated that excessive LST and IGD, both independently and jointly, were associated with mental health disorders. Prior studies showed that greater time spent on screen media was correlated with a higher prevalence of multiple mental health disorders, such as depression and psychological distress [41-43]. This study extends the existing body of work by providing new evidence for additional mental health outcomes, including paranoia and suicidal ideation in adolescents.

One possible explanation is that screen-based activities displace time for other beneficial activities, such as PA [44], a well-established factor protecting against mental health disorders [45,46]. Indeed, previous studies have demonstrated that PA can help reduce excessive screen-based behaviors and improve mental health. Since PA and screen-based behaviors often co-develop during adolescence [47], fostering healthy lifestyles is crucial for both the current and the future well-being of this population. Another key factor is the inherent harm posed by screen-based behaviors. Beyond educational content, the internet and media are flooded with violent, graphic, and pornographic materials, which further exacerbate mental health [48].

Adolescents with both high levels of LST and symptoms of IGD had significantly higher odds of mental health issues compared to those with excessive LST or IGD alone. These behaviors may reinforce each other: high LST could increase the likelihood of developing IGD, while IGD, in turn, may encourage further screen time, creating a feedback loop that exacerbates mental health. Despite such co-occurrence, the absence of significant interactions, neither additive nor multiplicative, suggests that the associations of LST and IGD with mental health are independent rather than synergistic.

This conclusion is supported by the nonsignificant interaction, indicating a lack of synergy where the joint presence does not create a disproportionately greater association. In other words, while the combination of high LST and IGD was associated with a higher likelihood of mental health problems, their interaction did not statistically amplify or mitigate the effect of each behavior. Notably, the group with IGD but low LST (“Low Yes”) exhibited substantially higher odds of mental health disorders than the group with high LST but no IGD (“High No”). This indicates that while both behaviors warrant attention, IGD constitutes a disproportionately strong associated factor. Therefore, future research and public health interventions should address both factors, prioritizing the identification and management of IGD within a framework that includes multidimensional assessments of both screen time and gaming disorder symptoms.

Although no significant modifier was found to influence the associations between LST and mental health disorders, our findings suggested that the association was somewhat more pronounced in girls. Consistent with previous studies, girls tend to spend more time on screen-based media than boys [49,50] and report poorer mental health outcomes [51]. Subgroup analyses revealed distinct patterns of vulnerability: urban residents, adolescents from higher-income families, nonsingle children, and those living with parents or caregivers showed stronger associations between excessive LST and mental health problems. In contrast, the link between IGD and mental health issues was more pronounced among girls, nonsingle children, rural adolescents, individuals from low- to middle-income families, and those in single-parent households. According to a report released by the China Internet Network Information Center, while internet penetration rates among Chinese adolescents are comparable across urban and rural areas, differences emerge in the use of specific online applications [52,53]. Moreover, previous research has shown that parenting style and family function play a pivotal role in the development of screen-based media dependency [54].

Evidence also indicates that parental absence can contribute to depression in children [55], potentially due to disruptions in child–parent relationships resulting from parental migration [56] and reduced communication [57]. These factors may influence the impact of LST and IGD on mental health, as variations in family dynamics and parenting styles affect how screen-based behaviors influence emotional and psychological outcomes. For instance, supportive family environments may mitigate the negative effects of excessive screen time while dysfunctional family interactions could exacerbate the associated negative outcomes.

Limitations

Regarding the limitations, first, due to the cross-sectional design of this study, it is not possible to establish causal relationships between exposure and outcomes. Second, although random cluster sampling was used, our participants were exclusively Han Chinese adolescents from Pidu District, Chengdu. While this sample captured urban–rural diversity within the district, it may not fully represent the broader socioeconomic, cultural, and regional heterogeneity across China. Therefore, caution is warranted when extrapolating our findings to inform national-level policies, and they should be interpreted as preliminary evidence primarily relevant to similar regional contexts. Third, a key limitation is that all data, including parameters for LST, IGD, and mental health symptoms, were based on self-report. This may introduce reporting bias. To address these limitations, future research could consider using longitudinal and interventional studies to validate the detrimental effects of screen-based behaviors on adolescent mental health. Expanding the research scope to include adolescents from multiple regions might enhance the representativeness and generalizability of findings. Additionally, using objective measurement tools, such as wearable devices or monitoring instruments, for screen time, alongside clinical interviews for diagnosing IGD and mental health disorders, would greatly improve the objectivity and accuracy of the assessments.

XSL•FO
RenderX

Conclusions

This large-scale population-based study provides novel evidence that LST and IGD represent distinct dimensions of digital engagement with independent and additive associations with mental health problems in Chinese adolescents. Methodologically, this study advances the field by simultaneously examining both screen time quantity and addictive usage patterns, using a comprehensive assessment of 5 mental health outcomes often overlooked in previous research. These findings highlight the need to expand public health strategies beyond screen time limits to include early IGD screening, emphasizing that both the quantity and quality of digital engagement require consideration. Future longitudinal studies should confirm these relationships and inform targeted interventions.

Acknowledgments

The authors would like to express their sincere gratitude to all the participants and their families for their valuable contribution to the cross-sectional survey on the mental health of children and adolescents in Pidu District, Chengdu.

Funding

This work was supported by the National Natural Science Foundation of China (82204170), the Science Fund for Creative Research Groups of Science and Technology Bureau of Sichuan Province (2024NSFTD0030), the Recruitment Program for Young Professionals of China, and other projects from West China School of Public Health and West China Fourth Hospital, Sichuan University.

Data Availability

The datasets generated and/or analyzed during this study are not publicly available, but can be obtained from the corresponding author upon reasonable request.

Authors' Contributions

XJ, MF, and QD conceived and designed the analysis. QD, LS, JH, XZ, RX, J Zhu, YQ, J Zhou, TY, XS, TH, SZ, and BY analyzed the data and interpreted the results. QD drafted the manuscript. MF and XJ revised the manuscript. XJ and MF contributed equally as co-corresponding authors. All authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables.
[XLSX File (Microsoft Excel File), 93 KB - jmir_v28i1e80737_app1.xlsx]

References

1. Khalaf AM, Alubied AA, Khalaf AM, Rifaey AA. The impact of social media on the mental health of adolescents and young adults: a systematic review. Cureus 2023;15(8):e42990 [FREE Full text] [doi: 10.7759/cureus.42990] [Medline: 37671234]
2. Pfeifer JH, Allen NB. Puberty initiates cascading relationships between neurodevelopmental, social, and internalizing processes across adolescence. Biol Psychiatry 2021;89(2):99-108 [FREE Full text] [doi: 10.1016/j.biopsych.2020.09.002] [Medline: 33334434]
3. Adolescent health. World Health Organization. 2025. URL: https://www.who.int/health-topics/adolescent-health/ [accessed 2025-12-16]
4. Malmir H, Mahdavi FS, Ejtahed H, Kazemian E, Chaharrahi A, Mohammadian Khonsari N, et al. Junk food consumption and psychological distress in children and adolescents: a systematic review and meta-analysis. Nutr Neurosci 2023;26(9):807-827. [doi: 10.1080/1028415X.2022.2094856] [Medline: 35816403]
5. Nunes ML, Bruni O. Insomnia in childhood and adolescence: clinical aspects, diagnosis, and therapeutic approach. J Pediatr (Rio J) 2015;91(6 Suppl 1):S26-S35 [FREE Full text] [doi: 10.1016/j.jped.2015.08.006] [Medline: 26392218]
6. de Zambotti M, Goldstone A, Colrain IM, Baker FC. Insomnia disorder in adolescence: diagnosis, impact, and treatment. Sleep Med Rev 2018;39:12-24 [FREE Full text] [doi: 10.1016/j.smrv.2017.06.009] [Medline: 28974427]
7. Kieling C, Baker-Henningham H, Belfer M, Conti G, Ertem I, Omigbodun O, et al. Child and adolescent mental health worldwide: evidence for action. Lancet 2011;378(9801):1515-1525. [doi: 10.1016/S0140-6736(11)60827-1] [Medline: 22008427]

XSL•FO
RenderX

8. Lu H, Yu Y, Wang DB, Wu AMS, Chen JH, Zhang G, et al. Association between interpersonal resources and mental health professional help-seeking among Chinese adolescents with probable depression: mediations via personal resources and active coping. BMC Psychiatry 2024;24(1):840 [FREE Full text] [doi: 10.1186/s12888-024-06271-4] [Medline: 39574049]
9. Chen S, Liu Y, Tremblay MS, Hong J, Tang Y, Cao Z, et al. Meeting 24-h movement guidelines: prevalence, correlates, and the relationships with overweight and obesity among Chinese children and adolescents. J Sport Health Sci 2021;10(3):349-359 [FREE Full text] [doi: 10.1016/j.jshs.2020.07.002] [Medline: 32679341]
10. Schønning V, Hjetland GJ, Aarø LE, Skogen JC. Social media use and mental health and well-being among adolescents - a scoping review. Front Psychol 2020;11:1949 [FREE Full text] [doi: 10.3389/fpsyg.2020.01949] [Medline: 32922333]
11. Nagata JM, Wong JH, Kim KE, Richardson RA, Nayak S, Potes C, et al. Social media use trajectories and cognitive performance in adolescents. JAMA 2025;334(21):1948-1950. [doi: 10.1001/jama.2025.16613] [Medline: 41082212]
12. Eales L, Wiglesworth A, Cullen KR, Klimes-Dougan B. Screen time, problematic media use, and clinical concerns in the ABCD study: differences by sex and race/ethnicity. Dev Psychopathol 2025:1-14. [doi: 10.1017/S0954579425100655] [Medline: 40970449]
13. Castro O, Bennie J, Vergeer I, Bosselut G, Biddle SJH. How sedentary are university students? A systematic review and meta-analysis. Prev Sci 2020;21(3):332-343. [doi: 10.1007/s11121-020-01093-8] [Medline: 31975312]
14. Xiao Y, Meng Y, Brown TT, Keyes KM, Mann JJ. Addictive screen use trajectories and suicidal behaviors, suicidal ideation, and mental health in US youths. JAMA 2025;334(3):219-228. [doi: 10.1001/jama.2025.7829] [Medline: 40531519]
15. Alfano CA, Moreno JP. Screen use in late childhood and early adolescence-a search for balance. JAMA Pediatr 2025;179(9):950-951. [doi: 10.1001/jamapediatrics.2025.1726] [Medline: 40549408]
16. Drescher AA, Goodwin JL, Silva GE, Quan SF. Caffeine and screen time in adolescence: associations with short sleep and obesity. J Clin Sleep Med 2011;7(4):337-342 [FREE Full text] [doi: 10.5664/JCSM.1182] [Medline: 21897768]
17. Wu X, Tao S, Zhang Y, Zhang S, Tao F. Low physical activity and high screen time can increase the risks of mental health problems and poor sleep quality among Chinese college students. PLoS One 2015;10(3):e0119607 [FREE Full text] [doi: 10.1371/journal.pone.0119607] [Medline: 25786030]
18. Zhou LH, Yu B, Xiao CC, Chen J, Zhu YZ, Yu QY, et al. [Dose-dependent associations between screen time, contents and adolescents' mental health]. Zhonghua Liu Xing Bing Xue Za Zhi 2025;46(6):1030-1035. [doi: 10.3760/cma.j.cn112338-20241014-00632] [Medline: 40518398]
19. Yin M, Huang S, Yu C. Depression and internet gaming disorder among chinese adolescents: a longitudinal moderated mediation model. Int J Environ Res Public Health 2023;20(4):3633 [FREE Full text] [doi: 10.3390/ijerph20043633] [Medline: 36834332]
20. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). Washington: American Psychiatric Association; 2013.
21. Wei C, Xu Q. The longitudinal relationship between school climate and internet gaming addiction among Chinese children: a moderated mediation model. Front Psychiatry 2025;16:1662888 [FREE Full text] [doi: 10.3389/fpsyt.2025.1662888] [Medline: 40995072]
22. Infanti A, Valls-Serrano C, Perales JC, Vögele C, Billieux J. Gaming passion contributes to the definition and identification of problematic gaming. Addict Behav 2023;147:107805 [FREE Full text] [doi: 10.1016/j.addbeh.2023.107805] [Medline: 37523871]
23. Kardefelt-Winther D, Heeren A, Schimmenti A, van Rooij A, Maurage P, Carras M, et al. How can we conceptualize behavioural addiction without pathologizing common behaviours? Addiction 2017;112(10):1709-1715 [FREE Full text] [doi: 10.1111/add.13763] [Medline: 28198052]
24. Gong X, Zhou J, Hao S. Longitudinal bidirectional relations between problematic internet game use and nonsuicidal self-injury among early adolescents: the mediating role of internalizing symptoms. Computers in Human Behavior 2025;165:108564 [FREE Full text] [doi: 10.1016/j.chb.2025.108564]
25. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
26. Council on Communications and Media. Children, adolescents, and the media. Pediatrics 2013;132(5):958-961. [doi: 10.1542/peds.2013-2656] [Medline: 28448255]
27. Pontes H, Griffiths M. Measuring DSM-5 internet gaming disorder: development and validation of a short psychometric scale. Computers in Human Behavior 2015;45:137-143 [FREE Full text] [doi: 10.1016/j.chb.2014.12.006]
28. Qin L, Cheng L, Hu M, Liu Q, Tong J, Hao W, et al. Clarification of the cut-off score for nine-item internet gaming disorder scale-short form (IGDS9-SF) in a Chinese context. Front Psychiatry 2020;11:470 [FREE Full text] [doi: 10.3389/fpsyt.2020.00470] [Medline: 32528331]
29. Furukawa TA, Kawakami N, Saitoh M, Ono Y, Nakane Y, Nakamura Y, et al. The performance of the Japanese version of the K6 and K10 in the world mental health survey Japan. Int J Methods Psychiatr Res 2008;17(3):152-158 [FREE Full text] [doi: 10.1002/mpr.257] [Medline: 18763695]
30. Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SLT, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002;32(6):959-976. [doi: 10.1017/s0033291702006074] [Medline: 12214795]

XSL•FO
RenderX

31. Romero-Acosta K, Lipps GE, Lowe GA, Gibson R, Ramirez-Giraldo A. The validation of the Kutcher Adolescent Depression 6-item scale in a sample of colombian preadolescents and adolescents. Eval Health Prof 2024;47(1):27-31. [doi: 10.1177/01632787231175931] [Medline: 37186570]
32. Dang W, Xu Y, Ji J, Wang K, Zhao S, Yu B, et al. Study of the SCL-90 scale and changes in the Chinese norms. Front Psychiatry 2020;11:524395 [FREE Full text] [doi: 10.3389/fpsyt.2020.524395] [Medline: 33584353]
33. First MB. Diagnostic and statistical manual of mental disorders, 5th edition, and clinical utility. J Nerv Ment Dis 2013;201(9):727-729. [doi: 10.1097/NMD.0b013e3182a2168a] [Medline: 23995026]
34. Osman A, Bagge CL, Gutierrez PM, Konick LC, Kopper BA, Barrios FX. The suicidal behaviors questionnaire-revised (SBQ-R): validation with clinical and nonclinical samples. Assessment 2001;8(4):443-454. [doi: 10.1177/107319110100800409] [Medline: 11785588]
35. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011;20(1):40-49 [FREE Full text] [doi: 10.1002/mpr.329] [Medline: 21499542]
36. Fu TS, Lee C, Gunnell D, Lee W, Cheng AT. Changing trends in the prevalence of common mental disorders in Taiwan: a 20-year repeated cross-sectional survey. Lancet 2013;381(9862):235-241. [doi: 10.1016/S0140-6736(12)61264-1] [Medline: 23151370]
37. Wang Y, Mínguez-Alarcón L, Gaskins AJ, Missmer SA, Rich-Edwards JW, Manson JE, et al. Association of spontaneous abortion with all cause and cause specific premature mortality: prospective cohort study. BMJ 2021;372:n530 [FREE Full text] [doi: 10.1136/bmj.n530] [Medline: 33762255]
38. Wang W, Du X, Guo Y, Li W, Zhang S, Zhang W, et al. Associations among screen time, sleep duration and depressive symptoms among Chinese adolescents. J Affect Disord 2021;284:69-74. [doi: 10.1016/j.jad.2021.01.082] [Medline: 33582434]
39. Gao Y, Wang J, Dong G. The prevalence and possible risk factors of internet gaming disorder among adolescents and young adults: systematic reviews and meta-analyses. J Psychiatr Res 2022;154:35-43. [doi: 10.1016/j.jpsychires.2022.06.049] [Medline: 35926424]
40. Wang J, Sheng J, Wang H. The association between mobile game addiction and depression, social anxiety, and loneliness. Front Public Health 2019;7:247. [doi: 10.3389/fpubh.2019.00247] [Medline: 31552213]
41. Li L, Abbey C, Wang H, Zhu A, Shao T, Dai D, et al. The association between video game time and adolescent mental health: evidence from rural China. Int J Environ Res Public Health 2022;19(22):14815 [FREE Full text] [doi: 10.3390/ijerph192214815] [Medline: 36429534]
42. Nagata JM, Al-Shoaibi AAA, Leong AW, Zamora G, Testa A, Ganson KT, et al. Screen time and mental health: a prospective analysis of the adolescent brain cognitive development (ABCD) study. BMC Public Health 2024;24(1):2686 [FREE Full text] [doi: 10.1186/s12889-024-20102-x] [Medline: 39370520]
43. Hoare E, Milton K, Foster C, Allender S. The associations between sedentary behaviour and mental health among adolescents: a systematic review. Int J Behav Nutr Phys Act 2016;13(1):108 [FREE Full text] [doi: 10.1186/s12966-016-0432-4] [Medline: 27717387]
44. Khan A, Burton N. Is physical inactivity associated with depressive symptoms among adolescents with high screen time? Evidence from a developing country. Mental Health and Physical Activity 2017;12:94-99 [FREE Full text] [doi: 10.1016/j.mhpa.2017.03.001]
45. Hrafnkelsdottir SM, Brychta RJ, Rognvaldsdottir V, Gestsdottir S, Chen KY, Johannsson E, et al. Less screen time and more frequent vigorous physical activity is associated with lower risk of reporting negative mental health symptoms among Icelandic adolescents. PLoS One 2018;13(4):e0196286 [FREE Full text] [doi: 10.1371/journal.pone.0196286] [Medline: 29698499]
46. Zink J, Belcher BR, Imm K, Leventhal AM. The relationship between screen-based sedentary behaviors and symptoms of depression and anxiety in youth: a systematic review of moderating variables. BMC Public Health 2020;20(1):472 [FREE Full text] [doi: 10.1186/s12889-020-08572-1] [Medline: 32272906]
47. Hills AP, King NA, Armstrong TP. The contribution of physical activity and sedentary behaviours to the growth and development of children and adolescents: implications for overweight and obesity. Sports Med 2007;37(6):533-545. [doi: 10.2165/00007256-200737060-00006] [Medline: 17503878]
48. Cuong VM, Assanangkornchai S, Wichaidit W, Minh Hanh VT, My Hanh HT. Associations between gaming disorder, parent-child relationship, parental supervision, and discipline styles: findings from a school-based survey during the COVID-19 pandemic in vietnam. J Behav Addict 2021;10(3):722-730 [FREE Full text] [doi: 10.1556/2006.2021.00064] [Medline: 34564065]
49. Monica A, Jiang JJ. Teens' Social Media Habits and Experiences. 2018. URL: http://tony-silva.com/eslefl/miscstudent/downloadpagearticles/teensandsocialmedia-pew.pdf [accessed 2025-12-16]
50. Kapidzic SCHS. Teens, gender, and self-presentation in social media. International Encyclopedia of Social and Behavioral Sciences. 2015. URL: https://homes.luddy.indiana.edu/herring/teens.gender.pdf [accessed 2025-12-16]
51. Hysing M, Pallesen S, Stormark KM, Lundervold AJ, Sivertsen B. Sleep patterns and insomnia among adolescents: a population-based study. J Sleep Res 2013;22(5):549-556. [doi: 10.1111/jsr.12055] [Medline: 23611716]

XSL•FO
RenderX

52. Pontes HM. Investigating the differential effects of social networking site addiction and internet gaming disorder on psychological health. J Behav Addict 2017;6(4):601-610 [FREE Full text] [doi: 10.1556/2006.6.2017.075] [Medline: 29130329]
53. Gabarron E, Denecke K, Lopez-Campos G. Evaluating the evidence: a systematic review of reviews of the effectiveness and safety of digital interventions for ADHD. BMC Psychiatry 2025;25(1):414 [FREE Full text] [doi: 10.1186/s12888-025-06825-0] [Medline: 40264083]
54. Xiuqin H, Huimin Z, Mengchen L, Jinan W, Ying Z, Ran T. Mental health, personality, and parental rearing styles of adolescents with internet addiction disorder. Cyberpsychol Behav Soc Netw 2010;13(4):401-406. [doi: 10.1089/cyber.2009.0222] [Medline: 20712498]
55. Zhou M, Sun X, Huang L, Zhang G, Kenny K, Xue H, et al. Parental migration and left-behind children's depressive symptoms: estimation based on a nationally-representative panel dataset. Int J Environ Res Public Health 2018;15(6) [FREE Full text] [doi: 10.3390/ijerph15061069] [Medline: 29795049]
56. He B, Fan J, Liu N, Li H, Wang Y, Williams J, et al. Depression risk of 'left-behind children' in rural China. Psychiatry Res 2012;200(2-3):306-312. [doi: 10.1016/j.psychres.2012.04.001] [Medline: 22572158]
57. Wu Q, Lu D, Kang M. Social capital and the mental health of children in rural China with different experiences of parental migration. Soc Sci Med 2015;132:270-277. [doi: 10.1016/j.socscimed.2014.10.050] [Medline: 25465499]

Abbreviations

DSM-5: Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition)
IGD: internet gaming disorder
IGDS9-SF: Internet Gaming Disorder Scale-9 Item Short Form
ISI: Insomnia Severity Index
K6: Kessler 6-item
KADS-6: Kutcher Adolescent Depression 6-item
LST: leisure screen time
OR: odds ratio
PA: physical activity
SBQ-R: Suicide Behaviors Questionnaire-Revised
SCL-90: Symptom Checklist-90
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by S Brini; submitted 15.Jul.2025; peer-reviewed by H Wan, X Liang, A Shivanna; comments to author 14.Sep.2025; accepted 07.Nov.2025; published 15.Jan.2026.

Please cite as:
Deng Q, Sha L, Hou J, Zhao X, Xiang R, Zhu J, Qu Y, Zhou J, Yu T, Song X, Zheng S, Han T, Yang B, Fan M, Jiang X
Leisure Screen Time, Internet Gaming Disorder, and Mental Health Among Chinese Adolescents: Large-Scale Cross-Sectional Study
J Med Internet Res 2026;28:e80737
URL: https://www.jmir.org/2026/1/e80737
doi:10.2196/80737
PMID:

©Qin Deng, Linna Sha, Jiaojiao Hou, Xunying Zhao, Rong Xiang, Jiangbo Zhu, Yang Qu, Jinyu Zhou, Ting Yu, Xin Song, Sirui Zheng, Tao Han, Bin Yang, Mengyu Fan, Xia Jiang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Longitudinal Between- and Within-Person Associations Among Screen Time, Bedtime, and Daytime Sleepiness Among Adolescents: Three-Wave Prospective Panel Study

Michał Tkaczyk[1], PhD; Albert J Ksinan[2], PhD; David Smahel[1], Prof Dr

[1]Interdisciplinary Research Team on Internet and Society, Faculty of Social Studies, Masaryk University, Brno, Czech Republic
[2]RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic

Corresponding Author:
Michał Tkaczyk, PhD
Interdisciplinary Research Team on Internet and Society
Faculty of Social Studies
Masaryk University
Joštova 218/10
Brno, 602 00
Czech Republic
Phone: 420 549 49 7289
Email: tkaczyk@fss.muni.cz

Abstract

Background: Daytime sleepiness is prevalent among adolescents and linked to multiple health and functional impairments. Prior research has identified digital media use and insufficient sleep as key predictors, yet the reciprocal longitudinal associations among screen time, sleep, and daytime sleepiness remain understudied.

Objective: This study examined the between- and within-person reciprocal longitudinal associations among adolescents' screen time, bedtime, and daytime sleepiness. It also tested whether potential adverse effects of screen time were less pronounced over time among adolescents who limited their screen exposure before sleep at baseline.

Methods: We conducted a prospective 3-wave panel study at 6-month intervals in a quota-based sample of 2500 Czech adolescents (mean age 13.43, SD 1.70 years; 1250/2500, 50% girls). Data were collected through an online survey. Screen time was assessed with 3 items covering total use of computers, smartphones, tablets, and television on a typical school day; bedtime with a single item on usual bedtime before school days; and daytime sleepiness with 4 items from the Pediatric Daytime Sleepiness Scale. Screen time restriction within 1 hour before sleep at baseline was examined as a moderator. Age and sex were included as covariates.

Results: Between- and within-person associations were estimated using random intercept cross-lagged panel models. Adolescents with higher screen time reported later bedtimes (r=0.23, 95% CI 0.15-0.31; P<.001) and greater daytime sleepiness (r=0.25, 95% CI 0.16-0.34; P<.001). No direct within-person effects emerged between screen time and daytime sleepiness (W1→W2: β=.02, 95% CI –0.11 to 0.16; P=.71; W2→W3: β=.02, 95% CI –0.10 to 0.14; P=.66). However, increases in screen time at Wave 1 predicted later bedtime at Wave 2 (β=.14, 95% CI 0.01-0.27; P<.05), which in turn predicted higher screen time at Wave 3 (β=.24, 95% CI 0.11-0.36; P<.001). Temporary within-person spikes in screen time coincided with delayed bedtimes (W1: r=0.16, 95% CI 0.04-0.27; P<.01; W2: r=0.23, 95% CI 0.09-0.36; P<.001; W3: r=0.09, 95% CI 0.00-0.20; P=.049). Baseline screen time restriction did not moderate within-person effects ($\Delta\chi^2_6$=5.3; P=.51).

Conclusions: This study is the first to test reciprocal longitudinal associations among adolescents' screen time, bedtime, and daytime sleepiness while separating between- and within-person associations, thereby addressing potential bias common in prior cross-lagged panel studies. The findings refine theoretical understanding by indicating a complex, bidirectional, and mutually reinforcing interplay between screen time and bedtime over time—even when individual differences are accounted for—whereas daytime sleepiness does not appear to be affected by this dynamic. Additionally, negatively correlated, within-person fluctuations in screen time and bedtime indicate that these behaviors are partly mutually exclusive and likely shaped by shared contextual influences. Screen time restriction before sleep did not mitigate within-person effects, indicating that interventions should prioritize consistent sleep schedules rather than focusing solely on reducing screen use.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e78972) doi:10.2196/78972

KEYWORDS

adolescents; bedtime; daytime sleepiness; random intercept cross-lagged panel model; RI-CLPM; screen restriction; screen time

Introduction

Background

Daytime sleepiness is an important, yet understudied, dimension of adolescents' sleep health [1]. Its prevalence varies widely across countries, ranging from 7.8% to 55.8% [2], and it is notably higher in adolescence than in adulthood [3]. Importantly, daytime sleepiness plays a central role in mediating the adverse effects of sleep impairment on adolescent health and well-being [4]. Studies have linked it to lower health-related quality of life [5], depressive symptoms, anxiety [6], heightened risk of mood disorders [7], and lower educational achievement [8]. Given its central role in linking sleep impairment to adverse outcomes, understanding the factors and processes contributing to daytime sleepiness in adolescence warrants greater scholarly attention.

Daytime sleepiness arises from an interplay of intrinsic (eg, brain maturation and sleep disorders) and extrinsic (eg, early school start times and poor sleep hygiene) factors [4]. Among these, insufficient sleep and late bedtimes on schooldays have been identified as the most direct contributors to daytime sleepiness among adolescents [9]. Digital media use is an important extrinsic factor known to affect sleep duration and bedtime timing; yet, most research on this association has been cross-sectional, limiting causal interpretations and leaving the direction of effects unclear [10,11].

Although the number of longitudinal studies is increasing [12], the vast majority do not distinguish between-person from within-person associations, which can lead to misleading conclusions about causal effects [13,14]. Studies that do separate these effects typically focus on short-term dynamics, such as day-to-day changes [15-18], often in small convenience samples, which limits their relevance for understanding longer-term processes.

To address these gaps in prior research, this study is the first to examine the longitudinal, reciprocal associations among screen time, bedtime, and daytime sleepiness, accounting for both stable between-person differences and within-person processes.

The study also tests whether restricting screen time before sleep moderates these associations. Clarifying whether daytime sleepiness emerges primarily from stable between-person differences, dynamic within-person processes, or both can advance theoretical understanding of how digital media use and adolescent sleep health influence each other. It may also help determine whether interventions should target stable behavioral patterns—such as sleep-related lifestyle habits, household routines, or family norms around screen use—or instead focus on longer-term individual trajectories—such as gradual increases in screen time or seasonal shifts in bedtime habits—or integrate both approaches.

Prior Work

Associations Among Screentime, Bedtime, and Daytime Sleepiness

Cross-sectional research consistently demonstrates positive associations between various screen-based activities—such as television watching, internet use, video gaming, and phone use—and both delayed bedtimes and increased daytime sleepiness [10,11]. Whereas this evidence cannot be a basis for causal interpretations, it suggests that, for some adolescents, higher screen time, later bedtimes, and greater daytime sleepiness tend to co-occur. This pattern likely reflects stable between-person differences that may be linked to external factors such as individual traits (eg, social anxiety), lifestyle demands (eg, extracurricular commitments), and family environment characteristics (eg, parenting style and household rules) [19-22].

The recent synthesis of evidence suggests that the causal link between screen time and sleep health is bidirectional, involving 2 potential pathways [23]. The screen-time-affecting-sleep pathway posits that media use, in particular before or after bedtime, contributes to shorter sleep duration and poorer sleep quality. Four explanatory mechanisms have been proposed: melatonin suppression due to blue light exposure, psychological arousal, displacement of sleep time, and sleep interruptions [24-26]. Of these, only displacement—that is, delayed bedtime due to screen time—and nighttime interruptions from notifications appear to have a substantial impact on sleep [23].

Conversely, the impaired-sleep-affecting-screen-time pathway posits that changes in sleep can contribute to increased media use. Three mechanisms explain this effect. Circadian phase shifts in puberty result in extended evening free time for media use [27,28]. Adolescents may use digital media to cope with sleep difficulties [23,29]. Daytime sleepiness is associated with more sedentary behavior, including prolonged screen time [30].

Longitudinal evidence supporting the 2 pathways is mixed. Some adolescent studies support the screen-time-affecting-sleep pathway (eg, meta-analysis by Pagano et al [12]), others report reciprocal associations [31,32], and some find minimal or no effects [33-35]. Evidence for the sleep-impairment-affecting-screen-time pathway exists, but in young adult samples [36]. There are also some longitudinal studies that found little or no support for either pathway or only marginal effects [33-35]. Such mixed findings may partly stem from conflating between-person and within-person associations in prior longitudinal studies.

To date, only 2 longitudinal studies have investigated the within-person associations between electronic media use and sleep-related outcomes—one focusing on daytime sleepiness [34] and the other on bedtime [33]. The former did not find significant within-person associations between the frequency of social media use and daytime sleepiness in Dutch adolescents

XSL•FO
RenderX

(aged 11-15 years), but did find between-person associations [34]. The latter, a 5-wave study of Finnish adolescents (aged 13-14 years at baseline), found no lagged effects and limited evidence of concurrent within-person associations—higher-than-usual social media use coincided with later-than-usual bedtime, but only in wave 1 [33]. These sparse findings suggest that the link between media use and sleepiness may arise from stable individual differences rather than changes over time.

A Moderating Role of Screen Time Restriction Before Sleep

Not all screen use is equally adverse for sleep health. In particular, evening screen time is considered detrimental to adolescent sleep [23], and restricting it is a common sleep hygiene recommendation [37]. Among adolescents, presleep screen restriction often results from parent-set technology rules, which cross-sectional studies have linked to less screen use, an earlier bedtime, and longer sleep duration [23]. While many adolescents do not follow their parents' technology rules and recommendations [38], research synthesis suggests that interventions aimed at reducing prebedtime screen use lead to modest improvements in bedtime and sleep duration [39]. Although this evidence suggests a potentially protective effect of reducing evening screen use, evidence on whether presleep screen restrictions moderate the longitudinal relationship between adolescents' screen time and sleep health is largely missing.

Covariates

Both screen use and sleep health vary by age and sex and therefore are important to consider when interpreting associations among adolescents' screen time and sleep health. In particular, older adolescents sleep less, go to bed later, and spend more time on screens, and younger adolescents are more likely to limit evening screen use [40-42]. Findings on sex differences in sleep are mixed. Some studies report no substantial differences [40], while others show girls sleep more than boys [43], or the reverse [44]. Daytime sleepiness findings are also inconsistent [45]. Sex differences in screen time are clearer. Boys exceed screen time limits more often [46], and sleep-disrupting screen activities differ—girls' sleep is more affected by social media, while boys' is impacted by video games [42]. Together, these patterns indicate that age and sex are important individual factors in understanding variation in adolescents' screen use and sleep health.

This Study

Prior longitudinal studies have rarely distinguished between stable between-person differences and within-person fluctuations in digital media use and sleep, leaving uncertainty about whether observed associations reflect enduring individual characteristics or dynamic changes over time [19]. The few adolescent studies that applied this distinction produced inconclusive results, with limited evidence for lagged or concurrent within-person effects [33,34]. To address this gap, this study extends prior work by examining the reciprocal longitudinal associations among adolescents' screen time, bedtime, and daytime sleepiness while separating between- and within-person processes. This allows us to clarify whether screen time and sleep co-vary because they influence each other over time or because of stable individual differences among adolescents. Furthermore, testing the moderating role of screen time restriction before sleep provides evidence on whether this common sleep hygiene recommendation mitigates longer-term effects of screen use on sleep health.

Specifically, we hypothesize that adolescents with higher overall screen time go to bed later and experience greater daytime sleepiness (Hypothesis 1); that increases in screen time are associated with a corresponding delay in bedtime and an increase in daytime sleepiness at the subsequent wave (Hypothesis 2), as well as within the same wave (Hypothesis 3); that delayed bedtime and increased daytime sleepiness are each associated with a subsequent increase in screen time (Hypothesis 4). The within-person effects expected in Hypotheses 2-4 reflect changes relative to a person's typical patterns. Finally, we hypothesize that within-person associations are weaker among adolescents who restrict their screen use before sleep (Hypothesis 5).

Methods

Ethical Considerations

The study was approved by the Research Ethics Committee at Masaryk University (EKV-2018-068). Before participation, respondents were informed about the nature and purpose of the survey, their right to decline involvement, and their ability to skip any questions by selecting the "I prefer not to say" option available for all items. Informed consent was obtained from both adolescents and parents. Parents were instructed not to be present during the adolescent survey to protect privacy. Adolescents were asked to indicate if an adult had observed or intervened. Although most caregivers appeared to comply, this could not be independently verified. All data were fully deidentified prior to analysis, and no identifying information was collected or stored. No identification of individual participants in any images of the manuscript or supplementary material is possible.

Participants received reward points equivalent to approximately US $4, added to the panelist's account and redeemable as cash or for charity donations.

Study Design and Setting

A longitudinal observational design was used. This 3-wave prospective panel study was a part of a larger multifocal study examining various aspects of adolescents' use of information and computer technologies and their impact on well-being. The first wave of data collection took place in June 2021, the second in November and December 2021, and the third in May and June 2022, with approximately 6 months between each wave. This study adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guidelines [47]; the completed STROBE checklist is provided in Multimedia Appendix 1 [48].

Participants

This study was conducted on a sample of 2500 Czech adolescents aged 11-16 years (mean age 13.43, SD 1.70 years;

XSL•FO
RenderX

1250/2500, 50% girls). Data were collected in the Czech Republic by an external research agency that recruited participants from existing online panels using face-to-face interviews, computer-assisted telephone interviewing, and online methods. Eligible participants were Czech households with at least 1 adolescent aged 11-16 years and a caregiver, enabling data collection from adolescent-parent dyads within the same household. Quota sampling was used to ensure equal representation of gender, age, and their combination and to ensure that the sample reflected the distribution of Czech households with children based on households' socioeconomic status (head of the household's education level) and place of residence (Nomenclature of Territorial Units for Statistics, level 3, municipality size, European Commission, 2020). Out of 2500 participants initially recruited at Wave 1, a total of 1654 completed Wave 2, corresponding to an attrition rate of 33.8% (846/2500). At Wave 3, a total of 1102 participants remained in the study. The overall attrition rate from Wave 1 to Wave 3 was 44.1% (1102/2500), with an incremental attrition rate of 33.4% (552/1654) between Wave 2 and Wave 3.

Measures

Screen Time

Screen time was assessed with 3 items, each starting with the question: "How much time (hours and minutes) do you spend doing the following activities during a typical school day?" The three items were: (1) "using a computer (PC or notebook)," (2) "using a cell phone or tablet," and (3) "watching TV, including various videos on TV (eg, DVD, Netflix)." In response to these items, respondents picked hours and minutes using a time spinner. The screen time score was then computed by adding up the scores of each item.

Bedtime

Bedtime was measured with 1 item: "When do you usually go to bed before school days?" In response to this item, respondents picked hours and minutes using a time spinner.

Daytime Sleepiness

Daytime sleepiness was measured using 4 items from the Pediatric Daytime Sleepiness Scale, which contains 8 items assessing the frequency of specific daytime sleepiness symptoms [49]. The 4 items were "You get sleepy or drowsy while doing your homework," "You have trouble getting out of bed in the morning," "You tell yourself that you need more sleep," and "You are tired and grumpy during the day." The items were rated on a 5-point scale: (1) "never,"(2) "rarely," (3) "sometimes," (4) "often," and (5) "very often." For each measurement occasion, a composite score was calculated as the mean of the items measuring the construct. A higher score indicates higher daytime sleepiness. Cronbach α was computed to assess the reliability of the scale across 3 waves. Reliability estimates were: α=0.77 for Wave 1, α=0.81 for Wave 2, and α=0.82 for Wave 3. These results indicate that the scale has acceptable internal consistency over time. Mean scores of the observed items were used for daytime sleepiness in analyses due to convergence issues when the latent variable was incorporated into the trivariate random intercept cross-lagged panel model (RI-CLPM).

Screen Time Restriction Within 1 Hour Before Sleep

Screen time restriction within 1 hour before sleep was measured at Wave 1. First, respondents were asked: "How long before going to sleep do you usually stop using all devices with a screen, ie, phone, tablet, computer, television?" Respondents picked hours and minutes using a time spinner in response to this item. Then, these data were transformed into a binary variable with values of 0 for adolescents who reported less than 60 minutes and 1 for adolescents who reported 60 minutes or more.

Covariates

Sex and age at baseline were self-reported at Wave 1 and were both included as time-invariant covariates in the analysis. Sex was coded as 0 for girls and 1 for boys, and age was grouped into 11-13 years (0) and 14-16 years (1).

Statistical Analysis

To examine the associations between screen time, bedtime, and daytime sleepiness over time while accounting for both between- and within-person sources of variance, we used RI-CLPMs fitted in *lavaan* (version 0.6-18) in R (version 4.4.1; R Core Team), allowing unbiased estimation of within-person effects net of stable individual differences [14]. The robust maximum likelihood estimator (MLR) was used, as it adjusts standard errors and chi-square statistics to accommodate nonnormal data (Section 3: "Testing normality assumptions" in supplementary materials provided by Tkaczyk et al [48]), yielding more accurate parameter estimates [50]. The proportion of missing data for the key time-varying variables ranged from 0.0% to 7.3% across waves. Little's Missing Completely at Random (MCAR) test indicated that the data were not completely missing at random (χ^2_{377}=787.8; P<.001; normed χ^2_{377}=2.1), suggesting a small to moderate deviation from MCAR. Given the low proportion of missing data (<8% per variable), full information maximum likelihood (FIML) estimation was used to handle missing values. For a detailed breakdown of percentages of missingness for each variable and wave, and results of logistic regressions testing the relationship between key analytical variables, demographics, and dropouts are provided in Section 1: "Attrition analysis" in supplementary materials provided by Tkaczyk et al [48].

To obtain more robust estimates, nonparametric bootstrapping with 2000 resamples was used to estimate 95% CIs for both unstandardized and standardized effects. Standardized coefficients represent the SD change in outcomes per 1 SD change in exposure. Chi-square difference tests were used to compare the fit of a nested model with constraints to the fit of the unconstrained model unless otherwise specified. The modeling approach was adapted from Mulder and Hamaker [51]. In the first step, the unconstrained RI-CLPM was compared to a model where all random intercept variances and covariances were set to zero (statistically equivalent to cross-lagged panel model [CLPM]) to test for stable between-unit differences, using the chi-bar-square test [52]. The comparison showed that the RI-CLPM fit the data better ($\Delta\chi^2_6$=286.8; P<.001). In addition, random intercepts of all 3 constructs had significant variance, indicating that there were some stable between-person

XSL•FO
RenderX

differences in screen time, bedtime, and daytime sleepiness over time. Second, to assess population-level changes in observed variables, we fixed grand means over time and compared this model to the unconstrained version. The comparison showed that the model without the constraints fit data better ($\Delta\chi^2_6$=74.1; *P*<.001), which implies that, on average, there was some change over time in all 3 variables. Third, to test whether the associations between screen time, bedtime, and daytime sleepiness were time-invariant, we constrained the autoregressive and cross-lagged paths, as well as the residual covariances. The model-building procedure indicated the fully unconstrained model as the best-fitting model Table 1. At this point, covariates (age and sex) were added to the model. The final model showed an adequate fit (χ^2_{15}=46.7; *P*<.001; Comparative Fit Index [CFI]=0.994; Tucker-Lewis Index [TLI]=0.977; root-mean-square error of approximation [RMSEA]=0.029, 90% CI 0.020-0.039; standardized root-mean-square residual [SRMR]=0.020). Fourth, moderation by screen time restriction before bed was tested using a multiple-group extension to RI-CLPM [51].

Table 1. Model fit indices for random intercept cross-lagged panel models (RI-CLPMs) examining longitudinal associations between screen time, bedtime, and daytime sleepiness across 3 waves in a longitudinal study of adolescents (aged 11-16 years). Data were collected in the Czech Republic between June 2021 and June 2022.

Model	χ^2(*df*)	CFI[a]	SRMR[b]	RMSEA[c]	TLI[d]	AIC[e]	BIC[f]
M0[g]	7.3 (3)	0.999	0.010	0.024	0.989	45591.676	45888.702
M1[h]	294.0 (9)	0.938	0.044	0.113	0.750	45866.448	46128.530
M2[i]	81.4 (9)	0.984	0.026	0.057	0.937	45653.804	45915.886
M3[j]	42.7 (18)	0.995	0.023	0.029	0.989	45597.131	45806.796
M3 + Covs[k]	46.7 (15)	0.994	0.020	0.029	0.977	45301.270	45633.241

[a]CFI: Comparative Fit Index.
[b]SRMR: standardized root-mean-square residual.
[c]RMSEA: root-mean-square error of approximation.
[d]TLI: Tucker-Lewis Index.
[e]AIC: Akaike information criterion.
[f]BIC: Bayesian information criterion.
[g]M0: fully unconstrained RI-CLPM.
[h]M1: cross-lagged panel model [CLPM].
[i]M2: RI-CLPM with grand means constrained over time.
[j]M3: RI-CLPM with constraint over time imposed on auto-regressive paths, cross-lagged paths, and residual (co)variances.
[k]M3 + Covs: M0 with covariates (age and sex).

Results

Descriptive Analysis

Table 2 displays pairwise correlations for time-varying variables across waves, along with their descriptive statistics, skewness, and kurtosis. The means of daytime sleepiness are close to "sometimes" (Wave 1: 2.81, SD 0.80; Wave 2: 2.82, SD 0.82; Wave 3: 2.84, SD 0.82). At Wave 1, approximately every third (788/2500, 32%) participant reported having trouble getting out of bed in the morning often or very often. Getting sleepy or drowsy while doing homework was the least frequent symptom—at Wave 1, approximately every sixth (400/2494, 16%) participant reported experiencing it often or very often.

XSL•FO
RenderX

Table 2. Pearson correlations and descriptive statistics for screen time, bedtime, and daytime sleepiness across 3 waves in a longitudinal study of adolescents (aged 11-16 years). Data were collected in the Czech Republic between June 2021 and June 2022. All correlation coefficients (r) are significant at *P*<.001.

Variable	ST[a] (W1[b])	ST (W2[c])	ST (W3[d])	BT[e] (W1)	BT (W2)	BT (W3)	DS[f] (W1)	DS (W2)	DS (W3)
ST (W1), *r*	1.00								
ST (W2), *r*	0.63	1.00							
ST (W3), *r*	0.59	0.62	1.00						
BT (W1), *r*	0.23	0.16	0.14	1.00					
BT (W2), *r*	0.17	0.20	0.21	0.61	1.00				
BT (W3), *r*	0.13	0.11	0.18	0.56	0.63	1.00			
DS (W1), *r*	0.13	0.09	0.09	0.22	0.17	0.13	1.00		
DS (W2), *r*	0.15	0.15	0.13	0.20	0.23	0.18	0.57	1.00	
DS (W3), *r*	0.14	0.13	0.16	0.21	0.19	0.20	0.55	0.64	1.00
Mean (SD), hh:mm or scale	06:23 (02:40)	06:11 (02:36)	06:02 (02:37)	09:48 00:56	09:47 (00:56)	09:57 (00:58)	2.81 (0.80)	2.82 (0.82)	2.84 (0.82)
Skewness	0.38	0.51	0.59	0.04	0.31	0.26	0.15	0.13	0.10
Kurtosis	−0.54	−0.29	−0.23	0.28	0.83	0.58	0.14	0.11	0.09

[a]ST: screen time (hh:mm).

[b]W1: Wave 1.

[c]W2: Wave 2.

[d]W3: Wave 3.

[e]BT: bedtime (hh:mm PM).

[f]DS: daytime sleepiness (1-5).

Average bedtimes at each wave were before 10 PM (Wave 1: 9:48, SD 00:56; Wave 2: 9:47, SD 00:56; Wave 3: 9:57, SD 00:58). At Wave 1, a total of 14% (353/2465) of participants reported bedtime at 11:00 PM or later (213/1621, 13% at Wave 2 and 190/1093, 17% at Wave 3). Average total daily screen times were close to 6 hours at each wave and showed a decreasing tendency across time (Wave 1: 06:23, SD 02:40; Wave 2: 06:11, SD 02:36; Wave 3: 06:02, SD 02:37).

Intraclass correlation coefficients (ICCs) revealed that between-person differences accounted for approximately 64% of the variance in screen time, 60% in bedtime, and 58% in daytime sleepiness, indicating a smaller but substantial proportion of variance due to within-person changes over time. All variables showed statistically significant and positive correlations both within and across waves.

Between-Person Associations Among Screen Time, Bedtime, and Daytime Sleepiness

Standardized path coefficients of the final RI-CLPM are presented in Figure 1.

The analysis revealed significant positive associations between the random intercepts of screen time and bedtime (*r*=0.23, 95% CI 0.15-0.31; *P*<.001), screen time and daytime sleepiness (*r*=0.25, 95% CI 0.16-0.34; *P*<.001), and bedtime and daytime sleepiness (*r*=0.31, 95% CI 0.22-0.41; *P*<.001). Consistent with Hypothesis 1, these correlations indicate that adolescents who typically use screens more also tend to go to bed later and experience higher daytime sleepiness. Additionally, those with later bedtimes tend to experience higher daytime sleepiness.

XSL•FO
RenderX

Figure 1. Standardized path coefficients of the final random intercept cross-lagged panel model testing between- and within-person associations among screen time, bedtime, and daily sleepiness across 3 measurement waves in a longitudinal study of adolescents (aged 11-16 years) conducted in the Czech Republic between June 2021 and June 2022. The model controls for the effects of age (at Wave 1 [W1]) and sex on the random intercepts of the time-varying variables. Solid black lines represent significant paths, and dashed lines represent nonsignificant paths. Solid gray paths were fixed to 1. **P*<.05; ***P*<.01; ****P*<.001.

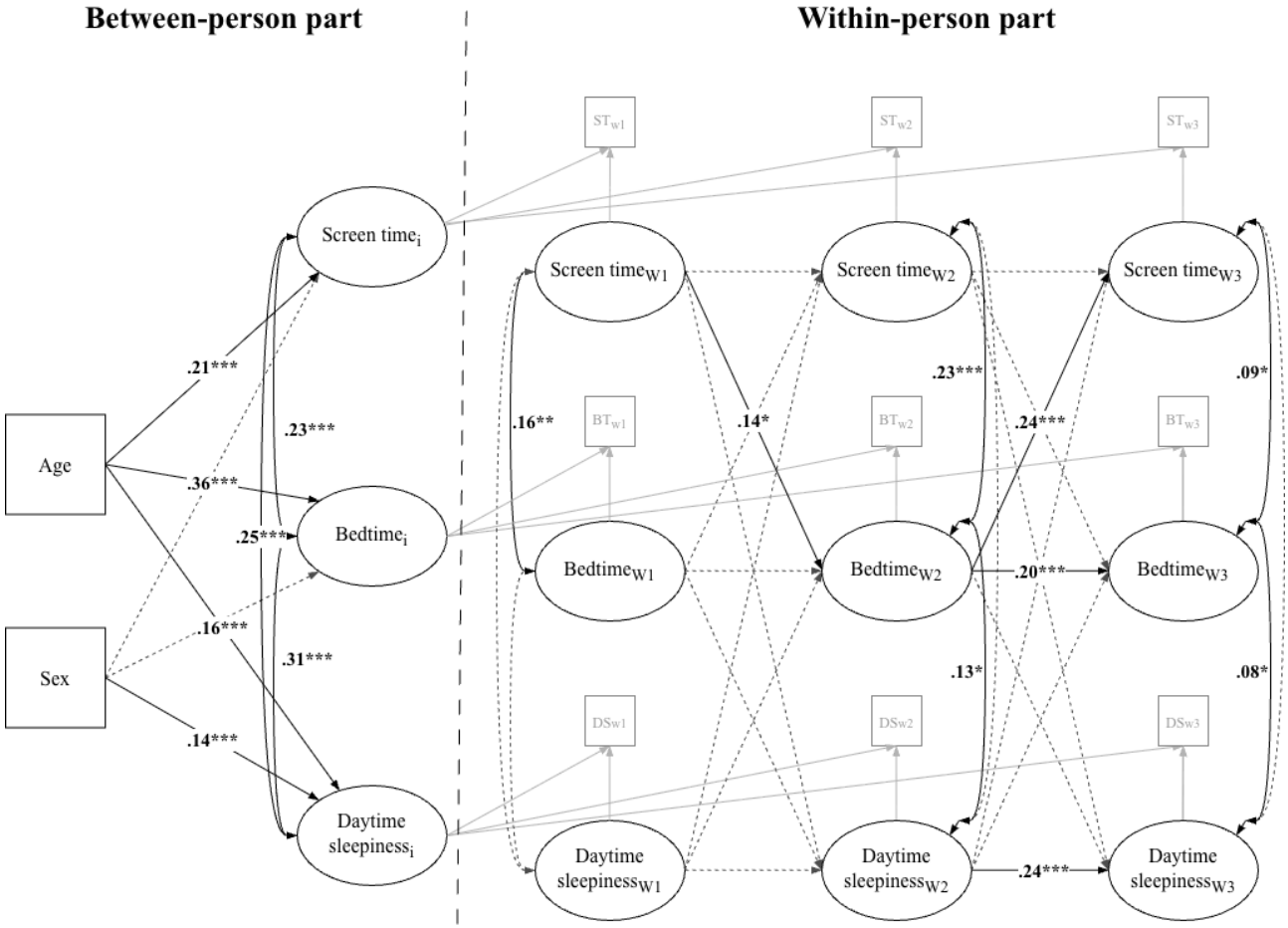

Within-Person Associations Among Screen Time, Bedtime, and Daytime Sleepiness

The analysis identified 2 significant cross-lagged effects. Consistent with Hypothesis 2, elevated screen time at Wave 1 relative to a person's usual patterns, was associated with elevated bedtime at Wave 2 (β=.14, 95% CI 0.01-0.27; *P*=.02). Similarly, in line with Hypothesis 4, elevated bedtime at Wave 2, relative to a person's usual patterns, was associated with increased screen time at Wave 3 (β=.24, 95% CI 0.11-0.36; *P*<.001). No evidence was found for the remaining cross-lagged paths hypothesized in Hypothesis 2 or Hypothesis 4.

Consistent with Hypothesis 3, the analysis revealed consistent concurrent associations between the within-person components of screen time and bedtime (Wave 1: β=.16, 95% CI 0.04-0.27; *P*=.007; Wave 2: β=.23, 95% CI 0.010-0.36; *P*<.001; Wave 3: β=.09, 95% CI 0.01-0.19; *P*=.049), indicating that an increase in screen time—relative to a person's usual patterns—was associated with a corresponding delay in bedtime with the same wave. No evidence was found to support Hypothesis 3. Additionally, a significant concurrent association between bedtime and daytime sleepiness was observed at Wave 2 (β=.13, 95% CI 0.01-0.26; *P*=.045) and Wave 3 (β=.08, 95% CI 0.00-0.17; *P*=.04). This indicates that, within these waves, a delay in bedtime was associated with elevated daytime sleepiness relative to a person's usual level of sleepiness.

The analysis also revealed autoregressive effects. Elevated bedtime at Wave 2, relative to a person's usual patterns, was associated with elevated bedtime at Wave 3 (β=.20, 95% CI 0.05-0.36; *P*<.001), indicating that a delay in bedtime—relative to a person's usual patterns—tends to carry over time. A similar autoregressive effect was observed for daytime sleepiness, with elevated sleepiness at Wave 2 associated with elevated sleepiness at Wave 3 (β=.24, 95% CI 0.12-0.37; *P*<.001).

The Role of Covariates

Age significantly predicted the intercepts of screen time (β=.21, 95% CI 0.16-0.25; *P*<.001), bedtime (β=.36, 95% CI 0.32-0.41; *P*<.001), and daytime sleepiness (β=.16, 95% CI 0.11-0.21; *P*<.001), indicating that older adolescents (aged 14-16 years) typically spent more time using screen media, have later bedtimes, and experience higher daytime sleepiness compared with younger adolescents (aged 11-13 years). Sex (boy=1) significantly predicted the intercept of daytime sleepiness (β=.14, 95% CI 0.10-0.19; *P*<.001), indicating that typical levels of daytime sleepiness are higher for boys than for girls (Table 3).

XSL•FO
RenderX

Table 3. Estimated parameters of the random intercept cross-lagged panel model (RI-CLPM) testing between- and within-person associations of screen time, bedtime, and daytime sleepiness across 3 measurement waves in a longitudinal study of adolescents (aged 11-16 years) conducted in the Czech Republic between June 2021 and June 2022. The model controls for the effects of age (at Wave 1) and sex on the random intercepts of the time-varying variables.

Parameter	*B*	SE	95% CI	*P* value	β
Between-person associations					
Correlations					
$ST_i^a \leftrightarrow BT_i^b$	0.314	0.062	0.184 to 0.436	<.001	.229
$ST_i \leftrightarrow DS_i^c$	0.296	0.050	0.177 to 0.407	<.001	.250
$BT_i \leftrightarrow DS_i$	0.123	0.019	0.084 to 0.159	<.001	.312
Covariates					
Age → ST_i	0.859	0.100	0.658 to 1.05	<.001	.206
Sex → ST_i	0.050	0.099	−0.138 to 0.247	.62	.012
Age → BT_i	0.523	0.033	0.457 to 0.591	<.001	.362
Sex → BT_i	0.054	0.033	−0.012 to 0.112	.08	.038
Age → DS_i	0.191	0.029	0.135 to 0.248	<.001	.160
Sex → DS_i	0.172	0.028	0.113 to 0.228	<.001	.144
Within-person associations					
Autoregressive paths					
ST^d ($W1^e$) → ST ($W2^f$)	0.100	0.070	−0.049 to 0.235	.15	.100
BT^g (W1) → BT (W2)	0.047	0.081	−0.145 to 0.233	.56	.045
DS^h (W1) → DS (W2)	0.027	0.074	−0.147 to 0.183	.72	.025
ST (W2) → ST ($W3^i$)	0.056	0.068	−0.103 to 0.201	.41	.056
BT (W2) → BT (W3)	0.224	0.059	0.04 to 0.387	<.001	.200
DS (W2) → DS (W3)	0.245	0.051	0.112 to 0.362	<.001	.242
Cross-lagged paths					
ST (W1) → BT (W2)	0.051	0.023	0.002 to 0.099	.03	.139
ST (W1) → DS (W2)	0.008	0.020	−0.037 to 0.051	.71	.023
BT (W1) → ST (W2)	0.131	0.171	−0.238 to 0.463	.44	.046
BT (W1) → DS (W2)	0.006	0.057	−0.124 to 0.133	.92	.006
DS (W1) → ST (W2)	0.001	0.186	−0.411 to 0.398	>.99	.000
DS (W1) → BT (W2)	-0.018	0.072	−0.178 to 0.125	.81	−.015
ST (W2) → BT (W3)	-0.009	0.022	−0.057 to 0.037	.69	−.021
ST (W2) → DS (W3)	0.008	0.018	−0.037 to 0.051	.66	.024
BT (W2) → ST (W3)	0.635	0.162	0.309 to 0.976	<.001	.235
BT (W2) → DS (W3)	-0.010	0.044	−0.113 to 0.084	.83	−.011
DS (W2) → ST (W3)	0.116	0.173	−0.25 to 0.481	.50	.038
DS (W2) → BT (W3)	0.102	0.061	−0.025 to 0.23	.09	.080
Residual covariances					
ST (W1) ↔ BT (W1)	0.156	0.058	0.04 to 0.277	.007	.158
ST (W1)↔ DS (W1)	0.043	0.047	−0.059 to 0.156	.05	.050
BT (W1) ↔ DS (W1)	0.025	0.018	−0.011 to 0.063	.15	.084
ST (W2) ↔ BT (W2)	0.229	0.066	0.087 to 0.364	<.001	.225

XSL•FO
RenderX

Parameter	*B*	SE	95% CI	*P* value	β
ST (W2) ↔ DS (W2)	0.098	0.057	−0.021 to 0.213	.09	.108
BT (W2) ↔ DS (W2)	0.042	0.021	−0.005 to 0.088	.045	.127
ST (W3) ↔ BT (W3)	0.099	0.050	−0.008 to 0.209	.049	.091
ST (W3) ↔ DS (W3)	0.077	0.040	−0.006 to 0.159	.06	.090
BT (W3) ↔ DS (W3)	0.030	0.014	−0.001 to 0.061	.04	.083

[a]ST_i: screen time latent intercept.

[b]BT_i: bedtime latent intercept.

[c]DS_i: daytime sleepiness latent intercept.

[d]ST: screen time.

[e]W1: Wave 1.

[f]W2: Wave 2.

[g]BT: bedtime.

[h]DS: daytime sleepiness.

[i]W3: Wave 3.

The Moderating Role of Screen Time Restriction Before Bed

Against Hypothesis 5, comparisons of multiple group RI-CLPMs with and without constraints across groups showed no differences in correlations between random intercepts ($\Delta\chi^2_3$=6.0; *P*=.11), residual covariances ($\Delta\chi^2_6$=3.5; *P*=.74), or cross-lagged associations ($\Delta\chi^2_6$=5.3; *P*=.51) across adolescents who restricted screen time 1 hour before bed at Wave 1 and those who did not. However, some significant differences between those groups were found ($\Delta\chi^2_2$=32.0; *P*<.001). Adolescents who restricted their screen time before bed reported, on average, shorter screen time (by 27 minutes and 28 seconds), earlier bedtime (22 minutes and 12 seconds), and lower daytime sleepiness (Δ=0.159; Table 3).

Discussion

Principal Results

This 3-wave prospective panel study examined bidirectional relationships between screen time, bedtime, and daytime sleepiness in a large representative sample of early to midadolescents in the Czech Republic. Findings at the between-person level showed that higher screen time, later bedtimes, and increased daytime sleepiness tend to co-occur among adolescents. At the within-person level, results revealed a bidirectional, transactional association between screen time and bedtime, suggesting mutual reinforcement over time. Additionally, temporary, wave-specific deviations in screen time and bedtime—relative to a person's usual patterns—were positively correlated, suggesting that increases in screen time and delays in bedtime tend to co-occur within individuals at the same wave. Finally, while restricting screen time before sleep did not modify these associations, adolescents who restricted screen time had lower typical screen time, earlier bedtimes, and less daytime sleepiness on average.

Between-Person Associations Among Screen Time, Bedtime, and Daytime Sleepiness

Consistent with Hypothesis 1, the analysis revealed small to medium positive correlations between screen time, bedtime, and daytime sleepiness at the between-person level, aligning with findings from cross-sectional studies [4,10,11,53]. However, previous RI-CLPM studies reported mixed correlation patterns. For instance, Maksniemi et al [33] found no significant between-person correlations between active social media use and bedtime, whereas 2 other studies reported medium positive correlations between social media use and daytime sleepiness [34] and between media multitasking and sleep problems [35]. Such inconsistencies may reflect differences across studies in how media use was conceptualized and defined (eg, active vs general social media use).

The between-person associations observed in this study indicate that higher screen time and poorer sleep co - occur as relatively stable individual tendencies, likely shaped by other stable factors. For example, late chronotype may predispose some adolescents to later bedtimes and heavier evening media use [54]. Prior work has shown that modifiable factors—such as parenting style [55], parental sleep [21], media habits [56], and household rules [57,58]—also influence both adolescent media habits and sleep. To guide better-targeted interventions, future longitudinal RI-CLPM studies should investigate how various modifiable family and lifestyle factors influence the media–sleep association over time.

Within-Person Associations Among Screen Time, Bedtime, and Daytime Sleepiness Over Time

Consistent with Hypothesis 2, increased screen time was associated with delayed bedtime 6 months later, but only between Waves 1 and 2. According to the interpretation guidelines proposed by Orth et al [59], the effect is considered large. The association, although not consistent across all waves, aligns with prior longitudinal research, including a 6-wave study based on data from the ABCD study among adolescents aged 11-14 years [60] and a 2-wave study among adolescents aged 13-14 years [31], both of which link media use to later bedtimes

XSL•FO
RenderX

over time. The present findings extend the prior evidence by demonstrating the association even when controlling for stable between‐person differences. Other RI-CLPM studies did not find cross-lagged effects; for instance, Maksniemi et al [33] found no association between active social media use and bedtime. Such discrepancies may reflect differences in conceptualizing media use—overall screen time versus active social media use—which involve distinct pathways linking media to sleep. Whereas active social media use mainly disrupts sleep through presleep arousal [61], total screen time is more closely related to blue light exposure and sleep displacement, the latter showing stronger and substantial associations with reduced sleep duration [23].

Contrary to Hypothesis 2, this study found no evidence of a direct within-person association between screen time and daytime sleepiness in the long term. This result is consistent with 2 previous RI-CLPM studies. Van Der Schuur et al [34] also found no evidence of a long-term within-person association between social media use and daytime sleepiness, aside from a small effect of social media stress among girls. Van der Schuur et al [35] found no direct path from media multitasking to sleep problems (including daytime sleepiness), except for a marginally significant effect of media multitasking among girls. Although direct effects were absent, indirect pathways remain plausible. Daytime sleepiness may occur when screen use results in later bedtimes [24]. Although bedtime was not formally tested as a mediator in this study, which should be considered a limitation, future longitudinal studies might examine bedtime delay as a pathway linking screen time to daytime sleepiness.

Consistent with Hypothesis 3, temporary increases in screen time coincided with temporary delays in bedtime across all 3 waves, indicating concurrent within-person associations between the two. Similar results were found by Maksniemi et al [33] in a single wave, whereas other RI-CLPM studies did not examine concurrent associations [34,35]. This pattern likely reflects the mutually exclusive nature of screen use and sleep within daily time allocation [62]; yet, the association manifests itself in period-specific, typical patterns of behavior—during periods when bedtime is delayed, adolescents have more opportunities for screen use, and conversely, during periods with greater screen use, they have less time available for sleep. Findings further indicate that bedtime remains sensitive to short-term, period-specific changes in screen time (and vice versa) and that both may share common contextual drivers.

Against Hypothesis 3, this study found no evidence of a correlated change between screen time and daytime sleepiness, suggesting that short-term increases in screen time do not directly coincide with increased sleepiness. Similarly, a diary study on smartphone use and next-day sleepiness found no such effects [18]. Delayed bedtimes in Waves 2 and 3 were concurrently linked to increased daytime sleepiness, likely due to shorter sleep duration [63]. Overall, the pattern of longitudinal associations found in this study suggests that while screen time and daytime sleepiness are not directly linked at the within-person level, an indirect path is possible, whereby delayed bedtime may mediate the association between technology use and daytime sleepiness.

Consistent with Hypothesis 4, this study found a within-person cross-lagged effect of bedtime on screen time in the subsequent wave: a later-than-usual bedtime predicted increased screen time 6 months later, but only between Waves 2 and 3. This finding aligns with prior longitudinal research showing reciprocal effects between poorer sleep and greater media use [31,36]. The RI-CLPM study by Van der Schuur et al [34] provided partial evidence for the opposite direction, with increased daytime sleepiness predicting decreased social media use over time among boys. Unlike earlier RI-CLPM studies, this study supports the sleep-impairment-affecting-screen-time pathway, demonstrating a substantial effect even after accounting for stable between-person differences. Although this effect was not consistent across all waves, it suggests that adolescents may extend screen use to fill additional evening hours, which likely arises from circadian shifts or related factors [28].

Overall, discrepancies in cross-lagged effects across RI-CLPM studies may partly reflect differences in the time intervals between measurements. This study used a 6-month interval, whereas Van der Schuur et al [34,35] used a 3- to 4-month interval, and Maksniemi et al [33] used a 1-year interval. The absence of cross-lagged effects in some cases suggests that these intervals may not have been optimal for capturing the underlying dynamics [33]. Future research could benefit from greater use of different temporal designs, such as shortitudinals, to identify optimal temporal windows for detecting within-person effects and the temporal dynamics through which media use influences sleep across adolescence.

Taken together, the cross-lagged pattern (screen time → bedtime between Waves 1 and 2; bedtime → screen time between Waves 2 and 3) suggests a reinforcing cycle between increased screen time and delayed bedtime over time. While previous research identified bidirectional links between screen time and sleep [31], this study extends prior work by being the first to demonstrate this reinforcing pattern longitudinally using an RI-CLPM that accounts for stable between-person differences. The autoregressive effects further indicate that delayed bedtimes tend to carry over across waves—a finding also reported in other RI-CLPM studies [33], which may reflect adolescent circadian shifts or habitual delays associated with greater autonomy or increased school demands [9]. Considering that delayed bedtimes were concurrently linked to greater daytime sleepiness and prospectively to higher screen time, interventions that promote earlier and more consistent sleep schedules, rather than solely limiting screen use, may be more effective for improving adolescent sleep health.

Effects of Screen Time Restriction Before Sleep

Contrary to Hypothesis 5, this study found no evidence that restricting screen time before sleep affected within-person associations between screen time and sleep, particularly regarding the development of sleep displacement over time. Prior findings are mixed—while experimental studies have shown improvements in sleep outcomes [64,65], observational studies often report no adverse effects of prebedtime smartphone use [17,18], with inconsistent adherence to parent-set rules frequently cited as a limiting factor [61]. These discrepancies

XSL•FO
RenderX

likely reflect differences in study design, sampling strategies, and time frames (eg, short- vs long-term). It should also be noted that the comparison groups were defined based on screen time restriction assessed at Wave 1 only. However, this behavior was not stable over time—among those who reported limiting their screen use at Wave 1, only 40% (284/710) did so across all 3 waves, and 65% (460/710) did so at least once thereafter. Future research should account for this temporal variability when examining the long-term effects of screen time restriction.

Adolescents who reported restricting screen use before sleep also tended to report lower overall screen exposure, earlier bedtimes, and less daytime sleepiness than their peers. Although these between-person differences may indicate a protective role of screen time restriction, they could also reflect other stable characteristics such as family environment (eg, parenting style), chronotype, or self-regulation. Prior research has linked adverse parenting styles to poorer sleep quality and greater daytime sleepiness [55], and greater sleepiness to lower self-regulation and eveningness chronotype [66]. Future longitudinal studies should account for these factors and examine their potential moderating roles in the relationship between screen use and sleep outcomes.

Limitations

Several limitations should be considered when interpreting these findings. First, the study relied on self-reported measures of screen time and sleep, which may be prone to inaccuracy [67,68]. Because overall screen time was calculated by summing reported use across multiple devices that could have been used simultaneously, average values may overestimate actual exposure. Future research should integrate digital trace data [69] and wrist-worn accelerometers data [70] for more accurate measurements.

Second, measurement simplifications—using total screen time and an abbreviated version of the Pediatric Daytime Sleepiness Scale [49]—may have reduced precision and obscured associations with sleep [71]. Future studies should use more detailed measures that account for media functions, content, and context of use [72,73].

Third, with only three waves spaced 6 months apart, the design was insufficient for modeling longer-term developmental trajectories [53,74] or accounting for seasonal variability in screen time and sleep [75,76]. Longer follow-up and more frequent measurement occasions would allow finer modeling of these changes.

Fourth, attrition was higher than in comparable school-based studies [33-35], likely because data were collected through an online panel and required the agreement of both adolescent and parent or caregiver. Online panels typically exhibit higher attrition rates due to the sustained participant burden and email-based recontact [77,78], and similar rates have been reported in other adolescent panel studies [79]. Notably, attrition remained high despite offering substantially increased incentives (160% in Waves 1 and 2; 280% in Wave 3). Dropouts reported slightly higher baseline screen time (Tables S1 and S2 in supplementary materials provided by Tkaczyk et al [48]), which may limit generalizability to heavy screen users.

Finally, data collection partially overlapped with COVID-19 social distancing measures, which were associated with increased screen time and later bedtimes among adolescents [60,80]. The stringency of restrictions varied across waves: Wave 1 (June 2021) coincided with the strictest measures, Wave 2 (November-December 2021) with moderate restrictions, and Wave 3 (May-June 2022) after their removal [81]. This variation may partly explain the observed decrease in screen time and the stability of bedtime between Waves 1 and 2.

Conclusion

This study is the first to test reciprocal longitudinal associations among adolescents' screen time, bedtime, and daytime sleepiness while separating between- and within-person processes, thereby addressing bias common in prior cross-lagged panel studies. The findings refine theoretical understanding by showing a complex, bidirectional, and mutually reinforcing interplay between screen time and bedtime over time, even after accounting for stable individual differences. Between-person associations revealed that adolescents with higher screen use had poorer sleep, likely reflecting the influence of relatively stable individual and environmental factors. Although specific cross-lagged effects varied across waves, the overall pattern supports both the screen-time-affecting-sleep and sleep-impairment-affecting-screen-time pathways, whereas daytime sleepiness was not affected by this dynamic. Negatively correlated within-person fluctuations further indicate that screen time and bedtime are partly mutually exclusive and may share contextual drivers.

Screen time restriction before sleep did not moderate within-person effects. However, at the between-person level, adolescents who practiced it reported lower screen use, earlier bedtimes, and less daytime sleepiness. Taken together, these findings suggest that interventions emphasizing consistent sleep schedules and supportive family routines—rather than focusing solely on limiting screen use—may be most effective for promoting adolescent sleep health. Future research should incorporate objective measurements on multiple time scales and relevant moderators.

Acknowledgments

The authors thank Dr David Lacko, Dr Vojtěch Mýlek, and Martin Tancoš for their thoughtful consultations during the preparation of this manuscript. During the preparation of this work, we used the generative artificial intelligence (AI) tool ChatGPT by OpenAI [82] to improve language clarity and readability. After using this service, we reviewed and edited the content as needed and take full responsibility for the publication's content.

XSL•FO
RenderX

Funding

This work has been funded by a grant from the Programme Johannes Amos Comenius under the Ministry of Education, Youth and Sports of the Czech Republic from the project "Research of Excellence on Digital Technologies and Wellbeing CZ.02.01.01/00/22_008/0004583" which is co-financed by the European Union. The work of AJK was supported from Operational Programme Johannes Amos Comenius—Project MSCAfellow5_MUNI (No. CZ.02.01.01/00/22_010/0003229). The funding sources were not involved in any research decisions.

Data Availability

The data used in this study are openly available on the Open Science Framework (OSF) [48].

Authors' Contributions

Conceptualization: MT (lead), AJK (supporting), DS (supporting)
Formal analysis: MT (lead), AJK (supporting)
Funding acquisition: DS
Methodology: MT (lead), AJK (supporting)
Project administration: DS
Supervision: DS
Validation: MT (lead), AJK (supporting)
Visualization: MT
Writing – original draft: MT
Writing – review & editing: MT (lead), AJK (supporting), DS (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
STROBE checklist.
[PDF File (Adobe PDF File), 183 KB - jmir_v28i1e78972_app1.pdf]

References

1. Buysse DJ. Sleep health: can we define it? Does it matter? Sleep 2014;37(1):9-17 [FREE Full text] [doi: 10.5665/sleep.3298] [Medline: 24470692]
2. Pereira É, Teixeira CS, Louzada FM. Sonolência diurna excessiva em adolescentes: prevalência e fatores associados. Rev Paul Pediatr 2010;28(1):98-103. [doi: 10.1590/s0103-05822010000100015]
3. Roehrs T, Carskadon MA, Dement WC, Roth T. Daytime sleepiness and alertness. In: Principles and Practice of Sleep Medicine. United States: Elsevier; 2005:39-50.
4. Moore M, Meltzer LJ. The sleepy adolescent: causes and consequences of sleepiness in teens. Paediatr Respir Rev 2008;9(2):114-121. [doi: 10.1016/j.prrv.2008.01.001] [Medline: 18513671]
5. Gustafsson M, Laaksonen C, Aromaa M, Asanti R, Heinonen OJ, Koski P, et al. Association between amount of sleep, daytime sleepiness and health-related quality of life in schoolchildren. J Adv Nurs 2016;72(6):1263-1272. [doi: 10.1111/jan.12911] [Medline: 26899487]
6. Moore M, Kirchner HL, Drotar D, Johnson N, Rosen C, Ancoli-Israel S, et al. Relationships among sleepiness, sleep time, and psychological functioning in adolescents. J Pediatr Psychol 2009;34(10):1175-1183 [FREE Full text] [doi: 10.1093/jpepsy/jsp039] [Medline: 19494088]
7. Lofthouse N, Gilchrist R, Splaingard M. Mood-related sleep problems in children and adolescents. Child Adolesc Psychiatr Clin N Am 2009;18(4):893-916. [doi: 10.1016/j.chc.2009.04.007] [Medline: 19836695]
8. Dewald JF, Meijer AM, Oort FJ, Kerkhof GA, Bögels SM. The influence of sleep quality, sleep duration and sleepiness on school performance in children and adolescents: a meta-analytic review. Sleep Med Rev 2010;14(3):179-189. [doi: 10.1016/j.smrv.2009.10.004] [Medline: 20093054]
9. Carskadon MA. Sleep in adolescents: the perfect storm. Pediatr Clin North Am 2011;58(3):637-647 [FREE Full text] [doi: 10.1016/j.pcl.2011.03.003] [Medline: 21600346]
10. Hale L, Guan S. Screen time and sleep among school-aged children and adolescents: a systematic literature review. Sleep Med Rev 2015;21:50-58 [FREE Full text] [doi: 10.1016/j.smrv.2014.07.007] [Medline: 25193149]
11. Lund L, Sølvhøj IN, Danielsen D, Andersen S. Electronic media use and sleep in children and adolescents in western countries: a systematic review. BMC Public Health 2021;21(1):1598 [FREE Full text] [doi: 10.1186/s12889-021-11640-9] [Medline: 34587944]

XSL•FO
RenderX

12. Pagano M, Bacaro V, Crocetti E. “Using digital media or sleeping … that is the question”. A meta-analysis on digital media use and unhealthy sleep in adolescence. Comput Hum Behav 2023;146:107813. [doi: 10.1016/j.chb.2023.107813]
13. Berry D, Willoughby MT. On the practical interpretability of cross-lagged panel models: rethinking a developmental workhorse. Child Dev 2017;88(4):1186-1206. [doi: 10.1111/cdev.12660] [Medline: 27878996]
14. Hamaker EL, Kuiper RM, Grasman RPPP. A critique of the cross-lagged panel model. Psychol Methods 2015;20(1):102-116. [doi: 10.1037/a0038889] [Medline: 25822208]
15. Lee PH, Tse ACY, Wu CST, Mak YW, Lee U. Temporal association between objectively measured smartphone usage, sleep quality and physical activity among Chinese adolescents and young adults. J Sleep Res 2021;30(4):e13213. [doi: 10.1111/jsr.13213] [Medline: 33049798]
16. Brosnan B, Haszard JJ, Meredith-Jones KA, Wickham S, Galland BC, Taylor RW. Screen use at bedtime and sleep duration and quality among youths. JAMA Pediatr 2024;178(11):1147-1154. [doi: 10.1001/jamapediatrics.2024.2914] [Medline: 39226046]
17. Siebers T, Beyens I, Baumgartner SE, Valkenburg PM. Adolescents’ digital nightlife: the comparative effects of day- and nighttime smartphone use on sleep quality. Commun Res 2024:00936502241276793. [doi: 10.1177/00936502241276793]
18. Tkaczyk M, Lacko D, Elavsky S, Tancoš M, Smahel D. Are smartphones detrimental to adolescent sleep? An electronic diary study of evening smartphone use and sleep. Comput Hum Behav 2023;149:107946. [doi: 10.1016/j.chb.2023.107946]
19. Bartel KA, Gradisar M, Williamson P. Protective and risk factors for adolescent sleep: a meta-analytic review. Sleep Med Rev 2015;21:72-85. [doi: 10.1016/j.smrv.2014.08.002] [Medline: 25444442]
20. Gunnell KE, Flament MF, Buchholz A, Henderson KA, Obeid N, Schubert N, et al. Examining the bidirectional relationship between physical activity, screen time, and symptoms of anxiety and depression over time during adolescence. Prev Med 2016;88:147-152. [doi: 10.1016/j.ypmed.2016.04.002] [Medline: 27090920]
21. Khor SPH, McClure A, Aldridge G, Bei B, Yap MB. Modifiable parental factors in adolescent sleep: a systematic review and meta-analysis. Sleep Med Rev 2021;56:101408. [doi: 10.1016/j.smrv.2020.101408] [Medline: 33326915]
22. Short MA, Gradisar M, Lack LC, Wright HR, Dewald JF, Wolfson AR, et al. A cross-cultural comparison of sleep duration between US and Australian adolescents: the effect of school start time, parent-set bedtimes, and extracurricular load. Health Educ Behav 2013;40(3):323-330 [FREE Full text] [doi: 10.1177/1090198112451266] [Medline: 22984209]
23. Bauducco S, Pillion M, Bartel K, Reynolds C, Kahn M, Gradisar M. A bidirectional model of sleep and technology use: a theoretical review of how much, for whom, and which mechanisms. Sleep Med Rev 2024;76:101933 [FREE Full text] [doi: 10.1016/j.smrv.2024.101933] [Medline: 38657359]
24. Cain N, Gradisar M. Electronic media use and sleep in school-aged children and adolescents: a review. Sleep Med 2010;11(8):735-742. [doi: 10.1016/j.sleep.2010.02.006] [Medline: 20673649]
25. Lissak G. Adverse physiological and psychological effects of screen time on children and adolescents: literature review and case study. Environ Res 2018;164:149-157. [doi: 10.1016/j.envres.2018.01.015] [Medline: 29499467]
26. Van den Bulck J. Text messaging as a cause of sleep interruption in adolescents, evidence from a cross-sectional study. J Sleep Res 2003;12(3):263. [doi: 10.1046/j.1365-2869.2003.00362.x] [Medline: 12941066]
27. Hansen SL, Capener D, Daly C. Adolescent sleepiness: causes and consequences. Pediatr Ann 2017;46(9):e340-e344. [doi: 10.3928/19382359-20170816-01] [Medline: 28892550]
28. Crowley SJ, Acebo C, Carskadon MA. Sleep, circadian rhythms, and delayed phase in adolescence. Sleep Med 2007;8(6):602-612. [doi: 10.1016/j.sleep.2006.12.002] [Medline: 17383934]
29. Eggermont S, Van den Bulck J. Nodding off or switching off? The use of popular media as a sleep aid in secondary-school children. J Paediatr Child Health 2006;42(7-8):428-433. [doi: 10.1111/j.1440-1754.2006.00892.x] [Medline: 16898880]
30. Malheiros LEA, da Costa BG, Lopes MV, Chaput J, Silva KS. Association between physical activity, screen time activities, diet patterns and daytime sleepiness in a sample of Brazilian adolescents. Sleep Med 2021;78:1-6. [doi: 10.1016/j.sleep.2020.12.004] [Medline: 33370617]
31. Mazzer K, Bauducco S, Linton SJ, Boersma K. Longitudinal associations between time spent using technology and sleep duration among adolescents. J Adolesc 2018;66:112-119. [doi: 10.1016/j.adolescence.2018.05.004] [Medline: 29842997]
32. Poulain T, Vogel M, Buzek T, Genuneit J, Hiemisch A, Kiess W. Reciprocal longitudinal associations between adolescents' media consumption and sleep. Behav Sleep Med 2019;17(6):763-777. [doi: 10.1080/15402002.2018.1491851] [Medline: 30040503]
33. Maksniemi E, Hietajärvi L, Ketonen EE, Lonka K, Puukko K, Salmela-Aro K. Intraindividual associations between active social media use, exhaustion, and bedtime vary according to age-a longitudinal study across adolescence. J Adolesc 2022;94(3):401-414. [doi: 10.1002/jad.12033] [Medline: 35390194]
34. van der Schuur WA, Baumgartner SE, Sumter SR. Social media use, social media stress, and sleep: examining cross-sectional and longitudinal relationships in adolescents. Health Commun 2019;34(5):552-559 [FREE Full text] [doi: 10.1080/10410236.2017.1422101] [Medline: 29313723]
35. van der Schuur WA, Baumgartner SE, Sumter SR, Valkenburg PM. Media multitasking and sleep problems: a longitudinal study among adolescents. Comput Hum Behav 2018;81:316-324. [doi: 10.1016/j.chb.2017.12.024]
36. Tavernier R, Willoughby T. Sleep problems: predictor or outcome of media use among emerging adults at university? J Sleep Res 2014;23(4):389-396. [doi: 10.1111/jsr.12132] [Medline: 24552437]

XSL•FO
RenderX

37. Hale L, Kirschen GW, LeBourgeois MK, Gradisar M, Garrison MM, Montgomery-Downs H, et al. Youth screen media habits and sleep: sleep-friendly screen behavior recommendations for clinicians, educators, and parents. Child Adolesc Psychiatr Clin N Am 2018;27(2):229-245 [FREE Full text] [doi: 10.1016/j.chc.2017.11.014] [Medline: 29502749]
38. Grasaas E, Ostojic S, Jahre H. Adherence to sleep recommendations is associated with higher satisfaction with life among Norwegian adolescents. BMC Public Health 2024;24(1):1288 [FREE Full text] [doi: 10.1186/s12889-024-18725-1] [Medline: 38730403]
39. Martin KB, Bednarz JM, Aromataris EC. Interventions to control children's screen use and their effect on sleep: a systematic review and meta-analysis. J Sleep Res 2021;30(3):e13130. [doi: 10.1111/jsr.13130] [Medline: 32567219]
40. Gariepy G, Danna S, Gobiņa I, Rasmussen M, Gaspar de Matos M, Tynjälä J, et al. How are adolescents sleeping? Adolescent sleep patterns and sociodemographic differences in 24 European and North American countries. J Adolesc Health 2020;66(6S):S81-S88 [FREE Full text] [doi: 10.1016/j.jadohealth.2020.03.013] [Medline: 32446613]
41. Smahel D, Machackova H, Mascheroni G, Dedkova L, Staksrud E, Ólafsson K, EU Kids Online network. EU Kids Online 2020: survey results from 19 countries. EU Kids Online, The London School of Economics and Political Science 2020. [doi: 10.21953/lse.47fdeqj01ofo]
42. Smith C, de Wilde T, Taylor RW, Galland BC. Prebedtime screen use in adolescents: a survey of habits, barriers, and perceived acceptability of potential interventions. J Adolesc Health 2020;66(6):725-732. [doi: 10.1016/j.jadohealth.2019.12.007] [Medline: 32044232]
43. Moore M, Slane J, Mindell JA, Burt SA, Klump KL. Sleep problems and temperament in adolescents. Child Care Health Dev 2011;37(4):559-562 [FREE Full text] [doi: 10.1111/j.1365-2214.2010.01157.x] [Medline: 21083682]
44. Maslowsky J, Ozer EJ. Developmental trends in sleep duration in adolescence and young adulthood: evidence from a national United States sample. J Adolesc Health 2014;54(6):691-697 [FREE Full text] [doi: 10.1016/j.jadohealth.2013.10.201] [Medline: 24361237]
45. Gibson ES, Powles AP, Thabane L, O'Brien S, Molnar DS, Trajanovic N, et al. "Sleepiness" is serious in adolescence: two surveys of 3235 Canadian students. BMC Public Health 2006;6(1):116 [FREE Full text] [doi: 10.1186/1471-2458-6-116] [Medline: 16670019]
46. Atkin AJ, Sharp SJ, Corder K, Van Sluijs EMF. Prevalence and correlates of screen time in youth: an international perspective. Am J Prev Med 2014;47(6):803-807. [doi: 10.1016/j.amepre.2014.07.043] [Medline: 25241193]
47. von Elm E, Altman DG, Egger M. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Epidemiology 2007;18(6):800-804. [doi: 10.1097/ede.0b013e3181577654]
48. Tkaczyk M, Ksinan AJ, Smahel D. Supplementary Materials and Analyses for “Longitudinal Between- and Within-Person Associations of Screen Time, Bedtime, and Daytime Sleepiness Among Adolescents: A Three-Wave Prospective Panel Study.”. 2025. URL: https://osf.io/hsjp7/ [accessed 2025-12-17]
49. Drake C, Nickel C, Burduvali E, Roth T, Jefferson C, Badia P. Pediatric Daytime Sleepiness Scale. 2011. URL: https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft02761-000 [accessed 2025-12-17]
50. Yuan KH, Bentler PM. Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. Sociological Methodology 2000;30(1):165-200. [doi: 10.1111/0081-1750.00078]
51. Mulder JD, Hamaker EL. Three extensions of the random intercept cross-lagged panel model. Struct Equ Modeling 2020;28(4):638-648. [doi: 10.1080/10705511.2020.1784738]
52. Stoel RD, Garre FG, Dolan C, van den Wittenboer G. On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. Psychol Methods 2006;11(4):439-455. [doi: 10.1037/1082-989x.11.4.439]
53. Campbell IG, Burright CS, Kraus AM, Grimm KJ, Feinberg I. Daytime sleepiness increases with age in early adolescence: a sleep restriction dose-response study. Sleep 2017;40(5):zsx046 [FREE Full text] [doi: 10.1093/sleep/zsx046] [Medline: 28419388]
54. Vollmer C, Jankowski KS, Díaz-Morales JF, Itzek-Greulich H, Wüst-Ackermann P, Randler C. Morningness-eveningness correlates with sleep time, quality, and hygiene in secondary school students: a multilevel analysis. Sleep Med 2017;30:151-159. [doi: 10.1016/j.sleep.2016.09.022] [Medline: 28215240]
55. Brand S, Hatzinger M, Beck J, Holsboer-Trachsler E. Perceived parenting styles, personality traits and sleep patterns in adolescents. J Adolesc 2009;32(5):1189-1207. [doi: 10.1016/j.adolescence.2009.01.010] [Medline: 19237190]
56. Lee HE, Kim JY, Kim C. The influence of parent media use, parent attitude on media, and parenting style on children's media use. Children (Basel) 2022;9(1):37 [FREE Full text] [doi: 10.3390/children9010037] [Medline: 35053662]
57. Buxton OM, Chang AM, Spilsbury JC, Bos T, Emsellem H, Knutson KL. Sleep in the modern family: protective family routines for child and adolescent sleep. Sleep Health 2015;1(1):15-27 [FREE Full text] [doi: 10.1016/j.sleh.2014.12.002] [Medline: 26779564]
58. Pedersen J, Rasmussen MG, Olesen LG, Klakk H, Kristensen PL, Grøntved A. Recreational screen media use in Danish school-aged children and the role of parental education, family structures, and household screen media rules. Prev Med 2022;155:106908 [FREE Full text] [doi: 10.1016/j.ypmed.2021.106908] [Medline: 34915040]
59. Orth U, Meier LL, Bühler JL, Dapp LC, Krauss S, Messerli D, et al. Effect size guidelines for cross-lagged effects. Psychol Methods 2024;29(2):421-433 [FREE Full text] [doi: 10.1037/met0000499] [Medline: 35737548]

XSL•FO
RenderX

60. Kiss O, Nagata JM, de Zambotti M, Dick AS, Marshall AT, Sowell ER, et al. Effects of the COVID-19 pandemic on screen time and sleep in early adolescents. Health Psychol 2023;42(12):894-903 [FREE Full text] [doi: 10.1037/hea0001251] [Medline: 36972087]
61. Bauducco SV, Flink IK, Jansson-Fröjmark M, Linton SJ. Sleep duration and patterns in adolescents: correlates and the role of daily stressors. Sleep Health 2016;2(3):211-218. [doi: 10.1016/j.sleh.2016.05.006] [Medline: 29073425]
62. Gába A, Dygrýn J, Štefelová N, Rubín L, Hron K, Jakubec L, et al. How do short sleepers use extra waking hours? A compositional analysis of 24-h time-use patterns among children and adolescents. Int J Behav Nutr Phys Act 2020;17(1):104 [FREE Full text] [doi: 10.1186/s12966-020-01004-8] [Medline: 32795287]
63. Asarnow LD, McGlinchey E, Harvey AG. The effects of bedtime and sleep duration on academic and emotional outcomes in a nationally representative sample of adolescents. J Adolesc Health 2014;54(3):350-356 [FREE Full text] [doi: 10.1016/j.jadohealth.2013.09.004] [Medline: 24225447]
64. Bartel K, Scheeren R, Gradisar M. Altering adolescents' pre-bedtime phone use to achieve better sleep health. Health Commun 2019;34(4):456-462. [doi: 10.1080/10410236.2017.1422099] [Medline: 29313721]
65. Duraccio KM, Zaugg KK, Blackburn RC, Jensen CD. Does iPhone night shift mitigate negative effects of smartphone use on sleep outcomes in emerging adults? Sleep Health 2021;7(4):478-484. [doi: 10.1016/j.sleh.2021.03.005] [Medline: 33867308]
66. Owens JA, Dearth-Wesley T, Lewin D, Gioia G, Whitaker RC. Self-regulation and sleep duration, sleepiness, and chronotype in adolescents. Pediatrics 2016;138(6):e20161406. [doi: 10.1542/peds.2016-1406] [Medline: 27940688]
67. Parry DA, Davidson BI, Sewall CJR, Fisher JT, Mieczkowski H, Quintana DS. A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. Nat Hum Behav 2021;5(11):1535-1547. [doi: 10.1038/s41562-021-01117-5] [Medline: 34002052]
68. Lauderdale DS, Knutson KL, Yan LL, Liu K, Rathouz PJ. Self-reported and measured sleep duration: how similar are they? Epidemiology 2008;19(6):838-845. [doi: 10.1097/ede.0b013e318187a7b0]
69. Stier S, Breuer J, Siegers P, Thorson K. Integrating survey data and digital trace data: key issues in developing an emerging field. Soc Sci Comput Rev 2019;38(5):503-516. [doi: 10.1177/0894439319843669]
70. Doherty A, Jackson D, Hammerla N, Plötz T, Olivier P, Granat MH, et al. Large scale population assessment of physical activity using wrist worn accelerometers: the UK Biobank Study. PLoS One 2017;12(2):e0169649 [FREE Full text] [doi: 10.1371/journal.pone.0169649] [Medline: 28146576]
71. Kaye KL, Orben A, Ellis AD, Hunter CS, Houghton S. The conceptual and methodological mayhem of "Screen Time". Int J Environ Res Public Health 2020;17(10):3661 [FREE Full text] [doi: 10.3390/ijerph17103661] [Medline: 32456054]
72. Sumter SR, Baumgartner SE, Wiradhany W. Beyond screentime: a 7-day mobile tracking study among college students to disentangle smartphone screentime and content effects on sleep. Behav Inf Technol 2024;44(6):1260-1276. [doi: 10.1080/0144929x.2024.2350663]
73. Livingstone S, Pothong K. Beyond screen time: Rethinking children's play in a digital world. J Health Visiting 2022;10(1):32-38. [doi: 10.12968/johv.2022.10.1.32]
74. Orth U, Clark DA, Donnellan MB, Robins RW. Testing prospective effects in longitudinal research: comparing seven competing cross-lagged models. J Pers Soc Psychol 2021;120(4):1013-1034 [FREE Full text] [doi: 10.1037/pspp0000358] [Medline: 32730068]
75. Devís-Devís J, Peiró-Velert C, Beltrán-Carrillo VJ, Tomás JM. Screen media time usage of 12-16 year-old Spanish school adolescents: effects of personal and socioeconomic factors, season and type of day. J Adolesc 2009;32(2):213-231. [doi: 10.1016/j.adolescence.2008.04.004] [Medline: 18694592]
76. Quante M, Wang R, Weng J, Kaplan ER, Rueschman M, Taveras EM, et al. Seasonal and weather variation of sleep and physical activity in 12-14-year-old children. Behav Sleep Med 2019;17(4):398-410 [FREE Full text] [doi: 10.1080/15402002.2017.1376206] [Medline: 28922020]
77. Kocar S, Biddle N. The power of online panel paradata to predict unit nonresponse and voluntary attrition in a longitudinal design. Qual Quant 2023;57(2):1055-1078 [FREE Full text] [doi: 10.1007/s11135-022-01385-x] [Medline: 35493336]
78. Lugtig P, Das M, Scherpenzeel A. In: Callegaro M, Baker R, Bethlehem J, Göritz AS, Krosnick JA, Lavrakas PJ, editors. Nonresponse and Attrition in A Probability‐Based Online Panel for the General Population. USA: Wiley; 2014:135-153.
79. Matthes J, Koban K, Neureiter A, Stevic A. Longitudinal relationships among fear of COVID-19, smartphone online self-disclosure, happiness, and psychological well-being: Survey study. J Med Internet Res 2021;23(9):e28700 [FREE Full text] [doi: 10.2196/28700] [Medline: 34519657]
80. Madigan S, Eirich R, Pador P, McArthur BA, Neville RD. Assessment of changes in child and adolescent screen time during the COVID-19 pandemic: a systematic review and meta-analysis. JAMA Pediatr 2022;176(12):1188-1198 [FREE Full text] [doi: 10.1001/jamapediatrics.2022.4116] [Medline: 36342702]
81. Measures Adopted by the Czech Government Against the Coronavirus.: Government of the Czech Republic URL: https://vlada.gov.cz/en/media-centrum/aktualne/measures-adopted-by-the-czech-government-against-coronavirus-180545/ [accessed 2025-09-22]
82. ChatGPT Large Language Model.: OpenAI; 2025. URL: https://chat.openai.com/chat [accessed 2025-12-17]

XSL•FO
RenderX

Abbreviations

CFI: Comparative Fit Index
CLPM: cross-lagged panel model
FIML: full information maximum likelihood
ICC: intraclass correlation coefficient
MCAR: Missing Completely at Random
MLR: maximum likelihood estimator
RI-CLPM: random intercept cross-lagged panel model
RMSEA: root-mean-square error of approximation
SRMR: standardized root-mean-square residual
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TLI: Tucker-Lewis Index

Edited by S Brini; submitted 13.Jun.2025; peer-reviewed by R Taylor, T Poulain; comments to author 08.Sep.2025; revised version received 01.Dec.2025; accepted 02.Dec.2025; published 21.Jan.2026.

Please cite as:
Tkaczyk M, Ksinan AJ, Smahel D
Longitudinal Between- and Within-Person Associations Among Screen Time, Bedtime, and Daytime Sleepiness Among Adolescents: Three-Wave Prospective Panel Study
J Med Internet Res 2026;28:e78972
URL: https://www.jmir.org/2026/1/e78972
doi:10.2196/78972
PMID:

©Michał Tkaczyk, Albert J Ksinan, David Smahel. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Tailored Internet-Delivered Mindfulness-Based Interventions for Patients With Hepatocellular Carcinoma After Transarterial Chemoembolization: Qualitative Study

Zengxia Liu[1,2,3], PhD; Min Li[4], MS; Yong Jia[3], PhD; Li Chen[3], Prof Dr

[1]School of Nursing, Anhui Medical University, Hefei, China
[2]Department of Nursing, First Affiliated Hospital of Anhui Medical University, Hefei, China
[3]School of Nursing, Jilin University, Changchun, Jilin, China
[4]Department of Intervention, First Hospital of Jilin University, Changchun, China

Corresponding Author:
Li Chen, Prof Dr
School of Nursing
Jilin University
965 Xinjiang Street
Changchun, 130021
China
Phone: 1 15943033929
Fax: 1 043185619366
Email: chen_care@126.com

Abstract

Background: Patients with hepatocellular carcinoma (HCC) undergoing transarterial chemoembolization (TACE) experience significant psychological distress, impacting outcomes. While mindfulness-based interventions (MBIs) are beneficial, access is limited. Internet-delivered MBIs (iMBIs) offer an accessible alternative; yet, qualitative understanding of patient experiences with tailored iMBIs for this specific population is lacking.

Objective: This study aimed to explore the facilitators and barriers of patients with HCC post TACE and participated in tailored iMBIs.

Methods: From November 2020 to December 2022, 11 patients with HCC post TACE who had taken part in tailored iMBIs were purposively recruited from a tertiary hospital in Jilin Province. Data were collected through semistructured interviews lasting 30-60 minutes. The interviews were analyzed using conventional content analysis.

Results: Five main categories emerged from the analysis: (1) mindfulness mindset, including acceptance, calmness, and mood improvement; (2) improvement of physical discomfort, such as better sleep, pain relief, reduced gastrointestinal symptoms, and increased activity levels; (3) resistance to mindfulness practice, including perceived lack of effectiveness, unsuitable conditions, equipment limitations, and difficulty concentrating; (4) support and encouragement, involving social support, supervision, and professional guidance; and (5) accessibility and convenience characterized by restoration of life balance and user-friendly features of the practice. Each category encompassed several subcategories reflecting the diverse experiences of participants.

Conclusions: While iMBIs were generally perceived as convenient and accessible, challenges such as equipment limitations were noted. Future implementation should focus on enhancing supportive factors to improve adherence, minimizing barriers, and refining the design and delivery of iMBI programs.

Trial Registration: Chinese Clinical Trial Registry ChiCTR1900027976; https://www.chictr.org.cn/showproj.html?proj=46657

(J Med Internet Res 2026;28:e78337) doi:10.2196/78337

KEYWORDS

internet; mindfulness; hepatocellular carcinoma; cancer; qualitative study

XSL•FO
RenderX

Introduction

Primary liver cancer is the third leading cause of cancer-related death worldwide and one of the most prevalent malignancies. China accounts for 45.27% of global cases and 47.12% of liver cancer–related deaths [1]. In 2022, there were approximately 368,000 new cases and 317,000 deaths from liver cancer in China, ranking fourth in cancer incidence and second in cancer-related mortality [2]. Hepatocellular carcinoma (HCC) is the most common histological subtype of primary liver cancer, comprising 75%-85% of all cases [3]. Transarterial chemoembolization (TACE) is a widely accepted local treatment for HCC and is the preferred option for patients with intermediate-stage HCC who are not suitable candidates for surgery [4,5]. However, patients often require multiple TACE sessions to manage disease progression. Despite treatment, they continue to face high risks of recurrence, metastasis, and various complications [6]. Moreover, the effectiveness of TACE diminishes with repeated use, and frequent treatments can result in progressive liver damage [7].

Due to the high mortality rate of HCC, research showed that the 5-year survival rate of patients with HCC after diagnosis was only 19.5% [8].Therefore, most patients with HCC face a serious threat of death, and their psychological distress is generally higher than that of other patients with cancer. During TACE treatment, some adverse reactions may occur, such as nausea, vomiting, fever, and pain. TACE patients have to face not only the physical pain brought by the disease and treatment but also the fear of disease recurrence, the high medical costs of repeated TACE treatments, and the pressure of being unable to return to society. This causes them to bear tremendous mental stress, resulting in widespread and severe psychological distress among patients with HCC post TACE [9,10]. This psychological burden can weaken patients' immune response, reduce treatment compliance, and adversely affect treatment outcomes, ultimately increasing the likelihood of tumor recurrence, deterioration, and metastasis [11].

Psychological distress in patients with cancer often arises not from a single external stressor that can be resolved or avoided but from the persistent fear of recurrence. Coping effectively, therefore, requires adapting to this ongoing internal experience. Mindfulness-based interventions (MBIs), which are designed to help individuals manage distressing thoughts and emotions, are particularly well suited for this purpose. The effectiveness of MBIs in alleviating psychological distress among patients with cancer has been well documented [12,13]. However, patients with HCC post TACE face unique physiological and psychological challenges. The uncertainty surrounding treatment outcomes, fear of recurrence, financial burden of repeated treatments, and treatment-related side effects collectively inflict significant physical and emotional strain. These factors highlight the urgent need to develop tailored MBIs specifically for this population—interventions that take into account not only the psychological profile and coping mechanisms of patients with HCC but also their physical health status. Despite the proven benefits of MBIs, several barriers limit access, including a shortage of trained therapists, high costs, and practical constraints, such as limited mobility and time availability among patients with cancer [14]. Addressing these challenges is essential to ensure that MBIs are accessible and effective for those who need them most.

Internet-delivered mindfulness-based interventions (iMBIs) offer a promising solution to many of the barriers associated with accessing psychosocial care in cancer treatment settings. Compared with traditional face-to-face interventions, iMBIs have several advantages. They can be accessed online at any time, allowing participants to engage with the program at their own convenience. This flexibility is particularly beneficial for patients with HCC and helps overcome practical challenges such as geographical distance, transportation difficulties, cancer-related fatigue, and mobility issues. This flexibility is particularly beneficial for patients with HCC post TACE, as most of them have a heavy burden of physical symptoms. This method helps overcome practical challenges such as geographic distance, transportation difficulties, cancer-related fatigue, and limited mobility. While iMBIs are associated with a higher rate of participant dropout, studies have nonetheless demonstrated their effectiveness in reducing psychological distress among patients with cancer [15-17].

Most studies investigating iMBIs in patients with cancer rely on standardized questionnaires to assess outcomes [18-20]. However, such quantitative approaches fail to capture the personal experiences and perceptions of participants. These subjective experiences are difficult to evaluate through usage data or survey metrics alone. Therefore, the primary aim of this qualitative study was to explore participants' experiences and perceived effects of tailored iMBIs. This approach allows for a deeper understanding of their thoughts and feelings, a comprehensive examination of both the benefits and the challenges of the intervention, and an analysis of factors that influence participant adherence and engagement.

Methods

Research Design

To gain a deep understanding of participants' experiences with iMBIs, as well as the interventions' acceptability and applicability, a qualitative descriptive design using semistructured personal interviews was used. The study followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) [21].

Participants

From November 2020 to December 2022, the study was conducted at the First Hospital of Jilin University. Potential participants were identified by evaluating and screening patient medical records to determine eligibility based on the inclusion criteria. The inclusion criteria for participants were as follows: (1) aged 18 years or older; (2) diagnosed with HCC based on the diagnostic criteria of the European Association for the Study of the Liver [22]; (3) currently received TACE; (4) able to operate a smartphone and use WeChat frequently (at least 5 times per week); (5) able to read, write, and speak Chinese; and (6) provided informed consent and voluntarily agreed to participate in the study. The exclusion criteria for participants were as follows: (1) diagnosed with other types of tumors or

XSL•FO
RenderX

mental stress disorders, (2) taking psychotropic medications during the study period, (3) received mindfulness intervention, and (4) experiencing poor health that interferes with normal communication. Those who met the criteria were invited to participate in a mindfulness intervention. Interviewees were then selected from among the participants who completed the iMBIs intervention.

In accordance with the principle of maximum variation sampling, participants were chosen to reflect a wide range of differences in age, gender, education level, and frequency of TACE treatments. The sample size was determined according to the concept of data saturation, meaning that collection ceased once no new codes emerged and the data appeared to have stabilized. After 11 patients were interviewed and the data were analyzed, the research team agreed that data saturation had been reached, and data collection was stopped.

Interventions

Overview

A cross-sectional survey of patients post TACE revealed that trait mindfulness, perceived stress, and experiential avoidance are potential psychological mechanisms influencing anxiety and depression in patients with HCC [10]. Based on this finding, the research team developed a tailored MBI for patients with HCC post TACE. The intervention was informed by the core principles of Mindfulness-Based Stress Reduction (MBSR) [23], Acceptance and Commitment Therapy [24,25], and Mindfulness-Based Cancer Recovery, as well as insights from a systematic review and an assessment of the characteristics and internet usage habits of this patient population. An initial draft of the tailored MBI was created and then refined through expert consultation and revisions.

A self-guided intervention was delivered through WeChat groups and an official WeChat account, which was used to distribute weekly mindfulness content. This content was presented in various formats, including text, audio, video, and images, via the official WeChat official account. The interface of the WeChat official account for the intervention content is shown in Figure 1. The intervention lasted for 6 weeks and included an initial in-hospital face-to-face session during the first week, followed by web-based sessions from the second to the sixth week.

Figure 1. The interface of the WeChat official account for the intervention content.

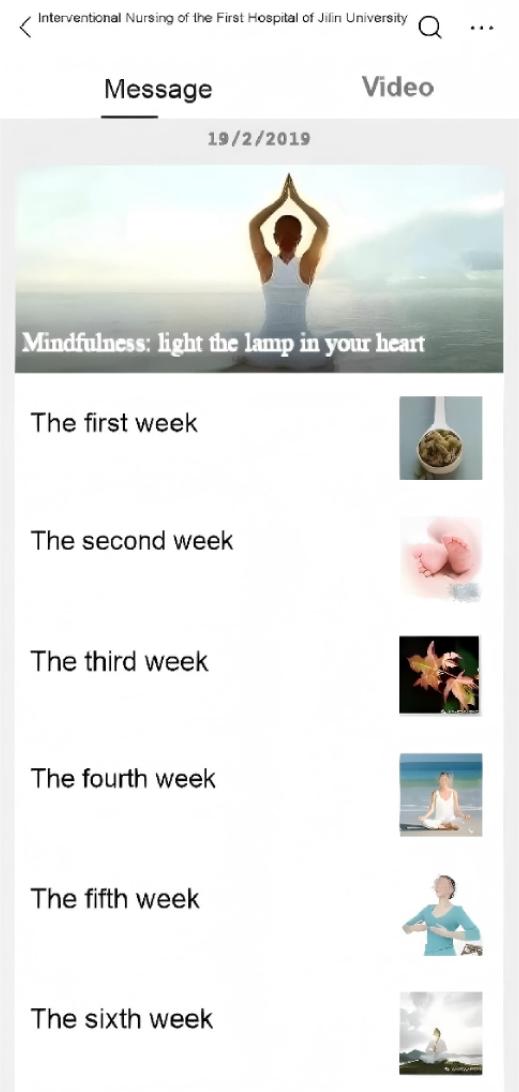

XSL•FO
RenderX

In-Hospital Face-to-Face Intervention (Week 1)

The face-to-face sessions were conducted in the interventional department classroom, with 5-8 participants per group. Each session lasted 1-1.5 hours. During this session, the research team introduced the study's purpose and content using presentation slides, provided an overview of mindfulness and its benefits, and helped participants understand the attitudes and principles underlying mindfulness practice. The session also familiarized participants with the structure and content of the upcoming web-based intervention, ensuring that they were prepared for the self-guided practice in the following weeks.

Out-of-Hospital Web-Based Intervention (Weeks 2-6)

For the remainder of the intervention, content was delivered weekly through WeChat in the form of audio, video, or image-based materials. Participants were encouraged to engage in daily mindfulness practice for 10-20 minutes, 5 days per week. The research team provided regular reminders and encouragement via WeChat and phone calls to ensure adherence. Participants were also encouraged to "check in" daily in the WeChat group by sharing their experiences and reflections on mindfulness practice. In addition, the research team shared educational materials through the WeChat group, covering topics such as mindfulness, HCC, and TACE treatment. They also responded to participants' questions regarding the disease, mindfulness practices, and other related concerns. The specific intervention content is detailed in Table 1.

Before interventions, 5 patients were invited to conduct a pre-experiment. After the first 2 weeks of intervention, adjustments were made according to the participants' feedback. Participants reported that the text on the official account interface was too small, which was not conducive to reading. Therefore, the researchers reorganized the layout to make the interface clearer and more esthetically pleasing. Participants also mentioned that there was too much text content, which was not conducive to reading and practice. As a result, the researchers converted some of the text content into audio to facilitate participants' learning and practice. Participants suggested that more pictures and videos could be added to make the content more vivid. The researchers added some pictures and video content to make the content more vivid and practical.

XSL•FO
RenderX

Table 1. The tailored internet-delivered mindfulness-based interventions for patients with HCC[a] post TACE[b].

Week	Theme of intervention	Summary of contents	Homework	Intervention mode
Week 1	Mindfulness and HCC and self-acceptance	1. Knowledge introduction: Introducing content related to mindfulness, HCC and TACE treatment, and the relationship between mindfulness and cancer. 2. Mindful breathing: Guide participants to practice mindful breathing, exhaling and inhaling easily, experiencing every detail of breathing, including the process, changes and pauses of breathing, accepting their breathing state, and being aware of the changes in their thoughts and attention. 3. "The Missing Piece": Watch the video "The Missing Corner" to prompt participants to think about "unchangeable imperfections," learn to accept negative events or emotions in life, and encourage participants to share their feelings after watching.	Mindful breathing	Face-to-face intervention in the hospital, 5-8 participants per group, 1-1.5 hours per session, introducing the purpose and content of the study, and guiding the completion of the first session.
Week 2	Cancer and stress and cognitive dissociation	1. Knowledge introduction: Describe knowledge related to cancer and stress, and typical stress reactions to cancer stress events. 2. Body scan: Introduce the connotation and requirements of body scan. Guide participants to feel and experience the sensation of each part of the body in order from feet to head. Pay attention to the feelings of each part, whether it is comfortable or not. Do not judge. Emphasize full commitment to the body-scanning practice. 3. "I have such an idea now": Guide participants to practice "I have such an idea now." Ask the participants to state a thought that makes them depressed, such as "Why did I get liver cancer?" Repeat this several times. After 3 minutes, add "I have an idea. Why did I get liver cancer?" before this thought. Repeat this several times. After 3 minutes, add a sentence before this idea, "I noticed that I have an idea. Why did I get liver cancer?" The main purpose of this exercise is to distinguish the real self from the imagined self, achieve cognitive dissociation, and prevent these thoughts from affecting an individual's life.	Body scan	Out-of-hospital online intervention, 20 minutes per day, at least 5 days per week, learn the intervention content via official account and WeChat, and 10 minutes of homework practice.
Week 3	Identifying avoidance reactions and self-awareness	1. Meditation practice: Guide participants to practice mindful meditation. Be aware of the thoughts and emotions that enter your mind. Pay attention to these thoughts and emotions but do not focus too much on them. Try to let go and shift your attention to your breathing. Then slowly shift your attention to your body, voice, thoughts, and emotions, and return to your breathing. In this repetitive process, gradually become aware of and feel your thoughts and emotions, and gradually and spontaneously coexist with them. 2. "Chessboard metaphor": Guide participants to imagine a "chess game." No longer caring about winning or losing, just providing a venue for the game of chess, as an observer, to observe the competition. Through the "chessboard metaphor," participants are guided to stop getting entangled in their inner struggles, become observers of their own psychology, and achieve self-awareness.	Meditation practice	Out-of-hospital online intervention, 20 minutes per day, at least 5 days per week, learn the intervention content via official account and WeChat, and 10 minutes of homework practice.
Week 4	Practice mindfulness in life, focusing on the present	1. Mindfulness walking: Guide patients to practice mindfulness walking and feel the current movements while practicing. From the lifting of the foot to the forward movement and the landing process, the sensation of each part of the foot, when practicing, feel the current movement and the current feeling, and focus your attention on the action of mindful walking. 2. Mindfulness eating: Guide patients to practice mindfulness eating. Slow down your eating speed, focus on the movements and sensations of your tongue and mouth while eating, experience the changes and flavors of food in your mouth, and pay attention to observing your inner feelings and emotions during the practice.	Mindful walking	Out-of-hospital online intervention, 20 minutes per day, at least 5 days per week, learn the intervention content via official account and WeChat, and 10 minutes of homework practice.

XSL•FO
RenderX

Week	Theme of intervention	Summary of contents	Homework	Intervention mode
Week 5	Nonselective awareness, clear value	1. Awareness practice: Explain the key points of awareness practice to patients and be aware of everything around them. Perceive the things around you and the sensations of your body with an accepting attitude, including pain, fatigue, and so on, and embrace everything. In practice, no matter what you perceive, trust and accept yourself. 2. "Compass" metaphor: Guide participants to think about the guiding role of the "compass" in the journey. By using this metaphor for goals and values, it guides participants to think about the direction and goals they want to move forward in, and based on their own values, they choose to set a small goal and take action.	Awareness practice	Out-of-hospital online intervention, 20 minutes per day, at least 5 days per week, learn the intervention content via official account and WeChat, and 10 minutes of homework practice.
Week 6	Grateful life, committed to action	1. Love meditation, good wishes: Cultivate the mindfulness practice of kindness and bestow your blessings upon others during the practice. You can choose different blessings in sequence to send to different people, including yourself, your family, friends, or even someone you dislike. Fill the participants' hearts with tolerance and love. 2. "Passengers on the Bus": Through the short film "Passenger Practice on the Bus," participants are guided to compare their own lives to buses and think about how to keep moving forward on the path of happiness when encountering different passengers, such as "stress," "anxiety," "happiness," and so forth, thereby contemplating the significance of committing to action.	Love meditation	Out-of-hospital online intervention, 20 minutes per day, at least 5 days per week, learn the intervention content via official account and WeChat, and 10 minutes of homework practice.

[a]HCC: hepatocellular carcinoma.

[b]TACE: transarterial chemoembolization.

Data Collection

Based on the disease characteristics and treatment experiences of patients with HCC following TACE, a preliminary interview guide was developed through a literature review and discussions among the research team. Prior to the formal interviews, 2 participants were selected for pilot interviews. The final interview guide was revised based on the issues identified during these pilot sessions and is shown in Table 2. Demographic information was collected before the interviews.

Two researchers (ZL and YJ), graduate students pursuing a Doctor of Nursing degree, approached potential participants during conducting the research. Once the patients had completed their consultations and treatments, we invited them to participate in this study. If they agreed to participate, we clearly explained the study's purpose and procedures to them. The researchers coordinated with participants in advance to schedule interviews and conducted them in a private room at the hospital to ensure confidentiality. At the beginning of each session, the researchers introduced themselves and explained the purpose, methods, and content of the interview. Participants were informed that the interview would be audio-recorded to ensure the completeness and accuracy of the data. Informed consent was obtained prior to starting the interview. During the interviews, the researchers maintained a neutral and nondirective attitude, using appropriate and nonleading language to avoid influencing participants' responses. Data collection and analysis were carried out concurrently to support iterative refinement of the findings.

During the interview, we used the interview tools proposed by Robinson [26], which include descriptive, individualized memory, exploratory, and clarifying inquiries. The individualized memory inquiry was used to guide the participants in recalling specific periods to obtain detailed information, such as the duration of their participation in iMBIs. The descriptive inquiry was used to explore the participants' feelings of participation. The interpretive inquiry was used to reveal the participants' views on iMBIs. The clarifying inquiry was used to clarify keywords and the implicit meanings of expressions. Based on the participants' responses, we modified or omitted the interview questions, or introduced new questions to explore new emerging topics, such as the influence of family on the participants' participation in iMBIs.

Table 2. Interview outline.

Number	Question
1	Why did you participate in this network-based mindfulness training?
2	What are your experiences while participating in this training?
3	What effects does network-based mindfulness training have on your body?
4	How does the network-based mindfulness training affect you psychology?

XSL•FO
RenderX

Ethical Considerations

The study was executed in accordance with the principles of the Declaration of Helsinki. The research protocol was approved by the ethics committee of the School of Nursing at Jilin University (approval no. 2019112001) and was registered with the Chinese Clinical Trial Registration Center (ChiCTR1900027976) on December 7, 2019. Prior to the interviews, participants were informed about the purpose and content of the study. They were assured that all data would be used solely for research and publication purposes, and that their personal information would remain confidential. Participants were also informed of their right to withdraw from the study at any time without any consequences. Written informed consent was obtained from all participants before the study commenced.

Data Analysis

One researcher (LZ) participating in the interview listened to the recordings repeatedly within 24 hours after each interview and transcribed them into text. Another researcher (LM) participating in the interview checked the transcribed text and imported it into NVivo 11 software [27] for analysis. Conventional content analysis [28,29] was used to analyze the data. Three researchers (LZ, JY, and CL) independently analyzed and transcribed the data.

Methodological Rigor and Trustworthiness of Data

Members of the research team include liver cancer clinical experts, nursing specialists, and psychologists. Interviews were conducted by 2 researchers (LZ and LM) to avoid information omission. To establish credibility, we used both analyst triangulation and data triangulation. Analyst triangulation was used to ensure intersubjective stability of the results and involved the independent analysis of the data by 3 researchers (LZ, JY, and CL), followed by a comparison of the analysis. The data were collated and analyzed by experienced researchers in qualitative research, the speech information of the interviewees was comprehensively retained, and complete and detailed textual data were established. The 2 researchers (LZ and JY) organized, analyzed, and transcribed the data independently. If there were differences, they were resolved through discussion or consultation with the third researcher (CL) until the results were reached.

Results

Characteristics of Participants

In this study, interviews were conducted with 9 participants, during which no new information emerged. To confirm data saturation, interviews with 2 additional participants were carried out, and again, no new themes were identified. Thus, a total of 11 participants were included in the final analysis. Their demographic and clinical characteristics are shown in Table 3.

Table 3. Characteristics of semidepth interview participants (N=11).

Number	Age (years)	Sex	Education level	Number of TACEs[a]
1	38	Male	Senior high school	1
2	50	Male	College school	8
3	53	Female	Senior high school	1
4	42	Female	College school	4
5	61	Male	Primary school	5
6	57	Female	Senior high school	1
7	49	Female	College school	7
8	62	Male	Senior high school	3
9	63	Female	Senior high school	3
10	48	Male	College school	7
11	40	Male	Primary school	4

[a]TACE: transarterial chemoembolization.

Categories

Overview

Through repeated analysis and organization of the interview data, 5 main categories reflecting the participants' experiences were identified: mindfulness mindset, improvement of physical discomfort, resistance to mindfulness practice, support and encouragement, and accessibility and convenience. Each main category is further divided into several subcategories, as shown in Textbox 1.

XSL•FO
RenderX

Textbox 1. Relevant categories corresponding to experiences of participants.

1. **Mindful mindset**
   - 1.1 Acceptance
   - 1.2 Calm
   - 1.3 Mood improvement
     - 1.3.1 Relaxation
     - 1.3.2 Gratitude
2. **Improvement of physical discomfort**
   - 2.1 Improve sleep
   - 2.2 Pain relief
   - 2.3 Reduce gastrointestinal symptoms
   - 2.4 Increase activities
3. **Mindfulness practice resistance**
   - 3.1 Lack of effect
   - 3.2 Condition
     - 3.2.1 Fatigue
     - 3.2.2 Pain
   - 3.3 Device usage restrictions
   - 3.4 Difficulty focusing
   - 3.5 Lack of motivation to practice
4. **Support and encourage**
   - 4.1 Social support
   - 4.2 Supervision and guidance
5. **Accessibility and convenience**
   - 5.1 Restore a sense of balance in your life
   - 5.2 Practice features

Main Category 1: Mindfulness Mindset

The mindfulness mindset refers to the ability of patients with HCC following TACE to face their condition with a nonjudgmental attitude, while maintaining a curious, open, and accepting approach toward their present thoughts and emotions.

General Category 1.1: Acceptance

Six participants reported that practicing mindfulness helped them adopt a more accepting attitude toward their illness and the challenges they faced in life.

> *My mother and I both have hepatitis, and my mother died of HCC, so I've always been worried about my health. I get scared every time I go for a check-up. But after doing the mindfulness exercises, my mindset changed. I gradually realized that I shouldn't overthink things every day. Since I have this disease, I just need to follow the doctor's advice and receive treatment. Many treatments are effective, but overthinking isn't good for my health.* [Interview ig4]

General Category 1.2: Calm

Seven participants reported that after practicing mindfulness, they experienced a greater sense of inner calm and became more attuned to the beauty in their surroundings.

> *My mindset has improved a lot. I can now perceive things around me more calmly and appreciate the little moments in life. I'm more willing to take walks, enjoy the scenery, and notice the beauty in simple things, like a tree or a flower by the roadside.* [Interview ig9]

General Category 1.3: Mood Improvement

Subcategory 1.3.1: Relaxation

Five participants shared that mindfulness practice helped them better regulate their emotions, reduce anxiety, and achieve a more relaxed state of mind.

XSL•FO
RenderX

> *Every time I went for a check-up, I would get really nervous if there was any change in my test results. Even small changes would worry me for a long time. Even when the doctor reassured me that everything was fine, I would still look up all kinds of information online. Now, when I notice myself feeling anxious, I try to adjust my emotions, relax, and avoid overthinking things that haven't even happened.* [Interview ig4]

Subcategory 1.3.2: Gratitude

Two participants expressed gratitude. After practicing mindfulness, they were able to approach people and situations around them with a more thankful attitude.

> *I've always had a bad temper and often got angry. My family would give in to me because of my poor health, which made me even angrier. After listening to the exercises, my mindset gradually improved. Now, when I encounter unpleasant things, I can stay calm. I also recognize how much my family has sacrificed for me, and my attitude has changed a lot. The atmosphere at home is much better.* [Interview ig3]

Main Category 2: Improvement of Physical Discomfort

Improvement of physical discomfort refers to the reduction of physical symptoms and alleviation of discomfort experienced by patients with HCC after TACE.

General Category 2.1: Improve Sleep

Ten participants reported significant improvements in sleep quality after mindfulness practice, with less presleep rumination and easier time falling asleep.

> *I've always had poor sleep, often lying in bed for a long time without falling asleep. The worse my sleep, the more I get lost in my thoughts. But since I started these exercises, I fall asleep shortly after listening. It helps stop my random thoughts.* [Interview ig3]

General Category 2.2: Pain Relief

Three participants said that mindfulness helped relieve pain and reduce physical discomfort.

> *I used to feel discomfort all over my body and pain everywhere. During the exercises, I take deep breaths, relax, follow the teacher, and slowly let go of tension. These uncomfortable feelings fade away, and I feel much better.* [Interview ig6]

General Category 2.3: Alleviation of Gastrointestinal Symptoms

Two participants reported that mindfulness practice helped ease gastrointestinal symptoms, improved their appetite, and reduced bloating.

> *I often had no appetite, felt uncomfortable after eating, and experienced bloating. Now I do these exercises daily, and these discomforts have improved. Sitting and relaxing each day reduces my irritability and unease.* [Interview ig1]

General Category 2.4: Increase Activities

Two participants noted that mindfulness helped reduce fatigue and encouraged them to be more active.

> *My whole body felt weak. I would just sit or lie down, unwilling to do anything. Sometimes I felt exhausted for long periods. After starting these exercises and practicing regularly, I feel much better. I'm more willing to move and have things to do.* [Interview ig4]

Main Category 3: Resistance to Mindfulness Practice

Resistance refers to factors that hinder patients' ability to consistently participate in mindfulness practice during the intervention.

General Category 3.1: Lack of Effectiveness

Two participants felt that the intervention had little effect on their psychological or physical state.

> *With this disease, it feels like a death sentence. I wanted to give up, but my family insisted I continue treatment. It feels like all efforts are just a loss of life and money, and this practice doesn't help.* [Interview ig1]

General Category 3.2: Conditions

Subcategory 3.2.1: Fatigue

Three participants reported that fatigue and physical discomfort made it difficult to practice mindfulness regularly.

> *I have poor appetite and low energy every day. I feel uncomfortable all over, the treatment isn't working, and I just want to lie down all day without doing anything. It's hard to practice.* [Interview ig4]

Subcategory 3.2.2: Pain

Two participants said that pain interfered with their ability to engage in mindfulness.

> *This pain is torturous. I'm not in the mood for anything, and it hasn't gotten better. When will I get better? This disease is unbearable.* [Interview ig6]

General Category 3.3: Device Usage Restrictions

Two participants said that distractions from their phones prevented them from completing mindfulness exercises consistently.

> *My phone keeps buzzing with messages, interrupting my practice. It's hard to stay quiet and focus for even a short time. There are just too many distractions.* [Interview ig10]

General Category 3.4: Difficulty in Focusing

Two participants found it hard to concentrate during mindfulness practice, leading to distraction and difficulty persisting.

> *When practicing at home, I often got lost in my thoughts. For example, when asked to focus on my abdomen, I'd start worrying about my illness and other messy things. It was hard to fully follow and stick with it.* [Interview ig3]

XSL•FO
RenderX

General Category 3.5: Lack of Motivation to Practice

Two participants reported difficulty maintaining regular practice due to low motivation.

> *I feel irritable and tired every day. I kept up with the exercises for a few days, but then felt it didn't work. Many times I was just too lazy to do the exercises.* [Interview ig1]

Main Category 4: Support and Encourage

Support and encouragement refer to the social and professional support received by patients during the online mindfulness intervention.

General Category 4.1: Social Support

Two participants shared that family involvement and communication through WeChat groups helped them persevere.

> *My wife started practicing with me in the hospital and continued after we got home. We share our experiences and feelings. Sometimes I don't want to practice, but she encourages me. Practicing together helped me stick with it.* [Interview ig5]

General Category 4.2: Supervision and Guidance

Three participants found that daily check-ins and reminders in the WeChat group, along with exchanges with other patients and medical staff, motivated their persistence.

> *Everyone checked in daily and shared their feelings in the WeChat group. Seeing fellow patients from my ward stick with it and ask questions encouraged me. After more communication, I kept going.* [Interview ig2]

Main Category 5: Accessibility and Convenience

Accessibility and convenience refer to patients' perceptions of how easy and convenient the online mindfulness intervention was to use.

General Category 5.1: Restore a Sense of Balance in Your Life

Three participants said that practicing mindfulness at the same time and place daily helped regulate their lives and restore balance and encouraged ongoing practice.

> *Since my diagnosis, my family has taken special care of me, not letting me do housework. I just lay around and felt useless. After practicing daily, I feel like I've accomplished something and have more energy.* [Interview ig4]

General Category 5.2: Practice Features

Three participants appreciated the convenience of the online format, finding it time-saving, flexible, and easy to fit into their daily routines.

> *I live in a rural area, and it's always hard to come to the hospital, especially without family to accompany me. Now I can practice anytime online after finding the audio. The WeChat method is convenient. When I feel down or have free time, I just practice—it doesn't have to be at a fixed time, so it's more flexible.* [Interview ig11]

Discussion

Principal Findings

This study found that participants experienced both physical and psychological benefits from the intervention. The physical improvements were primarily related to the alleviation of physical symptoms, consistent with findings from Nissen et al [20], who reported that iMBIs can enhance sleep and increase activity levels in patients with cancer. From a psychological perspective, studies by Eyles et al [30] and Weitz et al [31] demonstrated that an 8-week MBSR program can improve the mental health of patients with breast cancer, helping them live more fully in the present. Living in the present is associated with letting go of anxiety and rediscovering happiness [24]. Mindfulness interventions may also lead to changes in neurophysiological activity, brain structure, and function, which contribute to the positive psychological experiences reported by patients with cancer [32]. The intervention program of this study was based on the relevant contents of MBSR, Acceptance and Commitment Therapy, and Mindfulness-Based Cancer Recovery. Participants can have a better understanding of their own diseases and reduce the worries caused by the lack of disease knowledge. The intervention program incorporated elements of acceptance and gratitude. After the intervention, participants reduced their evasive attitude toward some negative emotions and events, faced negative thoughts and feelings with an accepting attitude, improved psychological flexibility, and thereby reduced anxiety and depression. Participants can better maintain a nonjudgmental mindset to continuously perceive and focus on their current experiences, avoiding their consciousness from diverging and wandering in the virtual world of thought, achieving the goal of stopping distractions, concentrating on real things, and thus attaining mental liberation, thereby enhancing trait mindfulness. These outcomes align with the mindfulness-to-meaning theory proposed by Garland et al [33], which suggests that MBIs enhance trait mindfulness and reduce psychological distress.

Social support and guidance were significant facilitators of mindfulness practice. The WeChat group used in this study provided a platform for participants to share experiences and feelings about mindfulness practice and their illness. The involvement of medical staff enabled timely professional guidance, enhancing participants' confidence and adherence. The group check-ins and communication encouraged mutual supervision, motivating participants to practice consistently. During the initial in-hospital phase, mindfulness was introduced to both patients and their families, encouraging joint participation, which helped mitigate the limitations of a solely online intervention. Research by Zulman et al [34] supports this approach, showing that engaging both patients with cancer and their caregivers can foster better communication and mutual support.

However, participants also encountered several obstacles, including perceived ineffectiveness, illness-related fatigue, equipment limitations, difficulty concentrating, lack of

XSL•FO
RenderX

motivation, and challenges completing practices on time. These barriers may stem from the advanced disease stage and poor health status of patients with HCC [35], which can hinder the ability to engage in even simple exercises. Future intervention designs should consider the patient's physical condition and create more tailored programs. A lack of professional mindfulness guidance also contributed to reduced compliance. Although this study used health care staff as facilitators, guidance provided through WeChat lacked real-time responsiveness and personalization, which likely weakened adherence. Previous research has shown that iMBIs with interactive guidance have significantly stronger effects on mindfulness and stress reduction than unguided interventions [36,37].

Participant inclusion criteria also influenced outcomes. High or low baseline levels of negative emotions can reduce intervention effectiveness. For example, Compen et al [15] used a Hospital Anxiety and Depression Scale cutoff of ≥11 and reported significantly better results. If patients are already in a stable psychological state, mindfulness interventions may unintentionally reinforce their identity as "patients," potentially reducing effectiveness. Therefore, selecting participants with appropriate levels of psychological distress is crucial to avoid ceiling or floor effects and to enhance the efficacy of the intervention.

Implications for Practice and Research

The effectiveness of iMBIs is closely tied to the platform used for delivery. Future iMBI platforms should be designed to meet users' individual and shared needs, enhancing usability and comfort. Comprehensive mobile health apps tailored for patients post TACE could be further developed. For instance, Subnis et al [38] created a mindfulness app that offers guided meditations, audio lectures, timers, logs, and stress assessments using facial biosignals. Such tools can reduce medical costs and address patients' psychological needs, providing a sustainable platform for continuous support.

Strengths and Limitations

This study tailored the iMBI to the physiological and psychological characteristics of participants and delivered it via WeChat, in line with their internet usage habits. Participants appreciated the convenience and flexibility of being able to practice at home, which reduced travel-related burdens. The study also emphasized family involvement, which is particularly relevant in the Chinese cultural context where family support plays a central role.

However, limitations remain. Compared with face-to-face mindfulness interventions, iMBIs lack real-time group interaction and peer support. Although WeChat enabled communication, it was still somewhat limited. Using health care providers instead of trained mindfulness instructors meant that the guidance lacked expertise and real-time feedback. Moreover, the WeChat platform itself had functional limitations, lacking features such as real-time monitoring, personalized feedback, and automatic recording of practice sessions. Another limitation is that interviews were conducted postintervention, relying on participants' recollections, which may introduce recall bias. This method also did not allow for tracking participants' evolving experiences over the course of the intervention.

Conclusions

This study used semistructured interviews to explore participants' experiences with iMBIs. Participants reported a variety of benefits, particularly psychological ones. Nevertheless, they also conveyed negative experiences, such as ineffectiveness, emotional resistance, difficulty concentrating, and low motivation. Factors such as social support, supervision, and restored life balance promoted adherence, while device-related distractions and the emotional reminder of illness hindered it. Although iMBIs offer convenience and accessibility, there are still important issues to address. Future efforts should explore participants' individualized needs more deeply and develop comprehensive, user-friendly apps to enhance both the comfort and the effectiveness of mindfulness interventions.

Acknowledgments

The authors would like to thank all participants and researchers for their cooperation.

Funding

Funding for this study was provided by University Natural Science Research Project of Anhui Provincial (grant 2023AH050583), Joint Special Project on Nursing of Anhui Institute of Translational Medicine (grant 2024zhyx-hl-B03), and General Project of Quality Engineering Teaching and Research in Anhui Province (grant 2023jyxm0252).

Data Availability

The data are not publicly available because they contain information that could compromise the privacy of study participants but are available from the corresponding author on reasonable request.

Authors' Contributions

ZL contributed to conceptualization, methodology, formal analysis, data curation, writing—original draft, and writing—review and editing. ML participated in methodology and data curation. YJ participated in data curation and writing—review and editing. LC participated in funding acquisition, project administration, and supervision.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
COREQ checklist.
[PDF File (Adobe PDF File), 92 KB - jmir_v28i1e78337_app1.pdf]

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74(3):229-263 [FREE Full text] [doi: 10.3322/caac.21834] [Medline: 38572751]
2. Han B, Zheng R, Zeng H, Wang S, Sun K, Chen R, et al. Cancer incidence and mortality in China, 2022. J Natl Cancer Cent 2024;4(1):47-53 [FREE Full text] [doi: 10.1016/j.jncc.2024.01.006] [Medline: 39036382]
3. NA. Hepatocellular carcinoma. Nat Rev Dis Primers 2021 Jan 21;7(1):7. [doi: 10.1038/s41572-021-00245-6] [Medline: 33479233]
4. Johnson BW, Wright GP. Regional therapies for the treatment of primary and metastatic hepatic tumors: a disease-based review of techniques and critical appraisal of current evidence. Am J Surg 2019;217(3):541-545. [doi: 10.1016/j.amjsurg.2018.10.018] [Medline: 30782316]
5. Tang Z, Bai T, Wei T, Wang X, Chen J, Ye J, et al. TACE combined lenvatinib plus camrelizumab versus TACE alone in efficacy and safety for unresectable hepatocellular carcinoma: a propensity score-matching study. BMC Cancer 2024;24(1):717 [FREE Full text] [doi: 10.1186/s12885-024-12484-3] [Medline: 38862932]
6. Kim MS, Kang M, Park J, Ryu JM. Nurses' comfort care of transarterial chemoembolization patients based on their perceptions around postembolization syndrome and symptom interference. Nurs Open 2023;10(5):2877-2885 [FREE Full text] [doi: 10.1002/nop2.1529] [Medline: 36565057]
7. Zhang L, Zhang X, Li Q, Makamure J, Liu Z, Zhao D, et al. Transarterial chemoembolization failure in patients with hepatocellular carcinoma: incidence, manifestation and risk factors. Clin Res Hepatol Gastroenterol 2023;47(2):102071. [doi: 10.1016/j.clinre.2022.102071] [Medline: 36539181]
8. Liu M, Liu X, Liu S, Xiao F, Guo E, Qin X, et al. Big data-based identification of multi-gene prognostic signatures in liver cancer. Front Oncol 2020;10:847 [FREE Full text] [doi: 10.3389/fonc.2020.00847] [Medline: 32547951]
9. Lan SC, Lin YE, Chen SC, Lin YF, Wang YJ. Effects of acupressure on fatigue and depression in hepatocellular carcinoma patients treated with transcatheter arterial chemoembolization: a quasi-experimental study. Evid Based Complement Alternat Med 2015;2015:496485 [FREE Full text] [doi: 10.1155/2015/496485] [Medline: 25802540]
10. Liu Z, Li M, Jia Y, Zheng L, Chen L. Effect of perceived stress on psychological distress in hepatocellular carcinoma patients undergoing TACE: the mediating role of experiential avoidance and the moderating role of trait mindfulness. BMC Cancer 2025;25(1):254 [FREE Full text] [doi: 10.1186/s12885-025-13679-y] [Medline: 39948489]
11. Cao W, Li J, Hu C, Shen J, Liu X, Xu Y, et al. Symptom clusters and symptom interference of HCC patients undergoing TACE: a cross-sectional study in China. Support Care Cancer 2013;21(2):475-483. [doi: 10.1007/s00520-012-1541-5] [Medline: 23010958]
12. Kubo A, Kurtovich E, McGinnis M, Aghaee S, Altschuler A, Quesenberry C, et al. A randomized controlled trial of mHealth mindfulness intervention for cancer patients and informal cancer caregivers: a feasibility study within an integrated health care delivery system. Integr Cancer Ther 2019;18:1534735419850634 [FREE Full text] [doi: 10.1177/1534735419850634] [Medline: 31092044]
13. Goldin PR, Thurston M, Allende S, Moodie C, Dixon ML, Heimberg RG, et al. Evaluation of cognitive behavioral therapy vs mindfulness meditation in brain changes during reappraisal and acceptance among patients with social anxiety disorder: a randomized clinical trial. JAMA Psychiatry 2021;78(10):1134-1142 [FREE Full text] [doi: 10.1001/jamapsychiatry.2021.1862] [Medline: 34287622]
14. Martín J, García S, Anton-Ladislao A, Ferreiro J, Martín M, Padierna A, CAMISS-prospective group. Variables related to health-related quality of life among breast cancer survivors after participation in an interdisciplinary treatment combining mindfulness and physiotherapy. Cancer Med 2023;12(12):13834-13845 [FREE Full text] [doi: 10.1002/cam4.6035] [Medline: 37165927]
15. Compen F, Bisseling E, Schellekens M, Donders R, Carlson L, van der Lee M, et al. Face-to-face and internet-based mindfulness-based cognitive therapy compared with treatment as usual in reducing psychological distress in patients with Cancer: a multicenter randomized controlled trial. J Clin Oncol 2018;36(23):2413-2421. [doi: 10.1200/jco.2017.76.5669]
16. Liu Z, Li M, Jia Y, Wang S, Zheng L, Wang C, et al. A randomized clinical trial of guided self-help intervention based on mindfulness for patients with hepatocellular carcinoma: effects and mechanisms. Jpn J Clin Oncol 2022;52(3):227-236. [doi: 10.1093/jjco/hyab198] [Medline: 35088079]

XSL•FO
RenderX

17. Shao D, Zhang H, Cui N, Sun J, Li J, Cao F. The efficacy and mechanisms of a guided self-help intervention based on mindfulness in patients with breast cancer: a randomized controlled trial. Cancer 2021;127(9):1377-1386 [FREE Full text] [doi: 10.1002/cncr.33381] [Medline: 33332582]
18. Ireland MJ, Clough B, Gill K, Langan F, O'Connor A, Spencer L. A randomized controlled trial of mindfulness to reduce stress and burnout among intern medical practitioners. Med Teach 2017;39(4):409-414. [doi: 10.1080/0142159x.2017.1294749]
19. Sommers-Spijkerman M, Austin J, Bohlmeijer E, Pots W. New evidence in the booming field of online mindfulness: an updated meta-analysis of randomized controlled trials. JMIR Ment Health 2021;8(7):e28168 [FREE Full text] [doi: 10.2196/28168] [Medline: 34279240]
20. Nissen ER, O'Connor M, Kaldo V, Højris I, Borre M, Zachariae R, et al. Internet-delivered mindfulness-based cognitive therapy for anxiety and depression in cancer survivors: a randomized controlled trial. Psychooncology 2020;29(1):68-75 [FREE Full text] [doi: 10.1002/pon.5237] [Medline: 31600414]
21. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
22. Bruix J, Sherman M, Llovet JM, Beaugrand M, Lencioni R, Burroughs AK, EASL Panel of Experts on HCC. Clinical management of hepatocellular carcinoma. Conclusions of the Barcelona-2000 EASL conference. European Association for the Study of the Liver. J Hepatol 2001;35(3):421-430. [doi: 10.1016/s0168-8278(01)00130-1] [Medline: 11592607]
23. Shapiro SL, Brown KW, Thoresen C, Plante TG. The moderation of mindfulness-based stress reduction effects by trait mindfulness: results from a randomized controlled trial. J Clin Psychol 2011;67(3):267-277. [doi: 10.1002/jclp.20761] [Medline: 21254055]
24. Finucane A, Mercer SW. An exploratory mixed methods study of the acceptability and effectiveness of mindfulness -based cognitive therapy for patients with active depression and anxiety in primary care. BMC Psychiatry 2006;6(1). [doi: 10.1186/1471-244x-6-14]
25. Fledderus M, Oude Voshaar MAH, Ten Klooster PM, Bohlmeijer ET. Further evaluation of the psychometric properties of the acceptance and action questionnaire-II. Psychol Assess 2012;24(4):925-936. [doi: 10.1037/a0028200] [Medline: 22545700]
26. Robinson OC. Probing in qualitative research interviews: Theory and practice. Qual Res Psychol 2023;20(3):382-397. [doi: 10.1080/14780887.2023.2238625]
27. QIP L. NVivo qualitative data analysis software (version 11). 2014. URL: https://lumivero.com/products/nvivo/ [accessed 2026-01-08]
28. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
29. Lindgren BM, Lundman B, Graneheim UH. Abstraction and interpretation during the qualitative content analysis process. Int J Nurs Stud 2020;108:103632. [doi: 10.1016/j.ijnurstu.2020.103632]
30. Eyles C, Leydon GM, Hoffman CJ, Copson ER, Prescott P, Chorozoglou M, et al. Mindfulness for the self-management of fatigue, anxiety, and depression in women with metastatic breast cancer: a mixed methods feasibility study. Integr Cancer Ther 2015;14(1):42-56 [FREE Full text] [doi: 10.1177/1534735414546567] [Medline: 25161198]
31. Weitz MV, Fisher K, Lachman VD. The journey of women with breast cancer who engage in mindfulness-based stress reduction: a qualitative exploration. Holist Nurs Pract 2012;26(1):22-29. [doi: 10.1097/HNP.0b013e31823c008b] [Medline: 22157506]
32. Hölzel BK, Carmody J, Vangel M, Congleton C, Yerramsetti SM, Gard T, et al. Mindfulness practice leads to increases in regional brain gray matter density. Psychiatry Res 2011;191(1):36-43 [FREE Full text] [doi: 10.1016/j.pscychresns.2010.08.006] [Medline: 21071182]
33. Garland EL, Fredrickson B, Kring AM, Johnson DP, Meyer PS, Penn DL. Upward spirals of positive emotions counter downward spirals of negativity: insights from the broaden-and-build theory and affective neuroscience on the treatment of emotion dysfunctions and deficits in psychopathology. Clin Psychol Rev 2010;30(7):849-864 [FREE Full text] [doi: 10.1016/j.cpr.2010.03.002] [Medline: 20363063]
34. Zulman DM, Schafenacker A, Barr KLC, Moore IT, Fisher J, McCurdy K, et al. Adapting an in-person patient-caregiver communication intervention to a tailored web-based format. Psychooncology 2012;21(3):336-341 [FREE Full text] [doi: 10.1002/pon.1900] [Medline: 21830255]
35. van den Hurk DGM, Schellekens MPJ, Molema J, Speckens AEM, van der Drift MA. Mindfulness-based stress reduction for lung cancer patients and their partners: results of a mixed methods pilot study. Palliat Med 2015;29(7):652-660 [FREE Full text] [doi: 10.1177/0269216315572720] [Medline: 25701663]
36. Arvidsson B, Allard E, Sjögren E, Lennernäs H, Sjöberg PJR, Bergquist J. Online capillary solid phase extraction and liquid chromatographic separation with quantitative tandem mass spectrometric detection (SPE-LC-MS/MS) of ximelagatran and its metabolites in a complex matrix. J Chromatogr B Analyt Technol Biomed Life Sci 2009;877(3):291-297. [doi: 10.1016/j.jchromb.2008.12.017] [Medline: 19117807]

XSL•FO
RenderX

37. Spijkerman MPJ, Pots WTM, Bohlmeijer ET. Effectiveness of online mindfulness-based interventions in improving mental health: a review and meta-analysis of randomised controlled trials. Clin Psychol Rev 2016;45:102-114 [FREE Full text] [doi: 10.1016/j.cpr.2016.03.009] [Medline: 27111302]
38. Subnis UB, Farb NA, Piedalue KL, Speca M, Lupichuk S, Tang PA, et al. A smartphone app-based mindfulness intervention for cancer Survivors: protocol for a randomized controlled trial. JMIR Res Protoc 2020;9(5):e15178 [FREE Full text] [doi: 10.2196/15178] [Medline: 32390591]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
HCC: hepatocellular carcinoma
iMBIs: internet-delivered mindfulness-based interventions
MBI: mindfulness-based intervention
MBSR: Mindfulness-Based Stress Reduction
TACE: transarterial chemoembolization

Edited by A Stone; submitted 31.May.2025; peer-reviewed by C Sun, W Zhou; comments to author 21.Oct.2025; revised version received 28.Nov.2025; accepted 31.Dec.2025; published 29.Jan.2026.

Please cite as:
Liu Z, Li M, Jia Y, Chen L
Tailored Internet-Delivered Mindfulness-Based Interventions for Patients With Hepatocellular Carcinoma After Transarterial Chemoembolization: Qualitative Study
J Med Internet Res 2026;28:e78337
URL: https://www.jmir.org/2026/1/e78337
doi:10.2196/78337
PMID:

©Zengxia Liu, Min Li, Yong Jia, Li Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Associations Among Cyberbullying Victimization, Inhibitory Control, Neural Activation of Error Processing, and Mental Health Problems in Adolescents: Neuroimaging, Retrospective Longitudinal Cohort Study Using the Adolescent Brain Cognitive Development Data

Xuanyu Zhang[1], BS; Chengyan Xie[2], BS; Yu Chen[3], PhD; Boyu Qiu[1], PhD

[1]School of Health Management, Guangzhou Medical University, Guangzhou, Guangdong, China

[2]The Third School of Clinical Medicine, Guangzhou Medical University, Guangzhou, China

[3]Department of Psychiatry, Yale University School of Medicine, New Haven, CT, United States

Corresponding Author:
Boyu Qiu, PhD
School of Health Management, Guangzhou Medical University
No.1, Xinzao Road
Xinzao Town, Panyu District
Guangzhou, Guangdong, 511436
China
Phone: 86 37103516 ext 37103516
Email: boyuqiu@gzhmu.edu.cn

Abstract

Background: Cyberbullying victimization is prevalent and closely linked to mental health problems. However, existing research, often limited by cross-sectional designs and a focus on direct relationships, has yielded inconsistent results. Furthermore, the biological mechanisms underlying the relationship between cyberbullying victimization and psychopathological outcomes remain largely unclear at present.

Objective: This retrospective cohort study aimed to explore the longitudinal associations among cyberbullying victimization, inhibitory control, brain activation during error processing, and mental health problems among adolescents.

Methods: We curated the clinical, behavioral, and neuroimaging data (551/1186, 46.5% girls; 9-10 years at baseline) from the Adolescent Brain Cognitive Development study, a nationally representative cohort established through school-based probability sampling (selected factors included gender, race/ethnicity, socioeconomic status, and urbanicity). Participants were assessed by the cyberbullying question, the functional magnetic resonance imaging stop signal task for inhibitory control and error processing, and the Child Behavioral Checklist for externalizing and internalizing problems at 2-year (T1) and 4-year follow-up (T2). Linear mixed models were used to examine the retrospective longitudinal associations between these clinical, behavioral, and neuroimaging factors.

Results: Linear mixed models showed that victims of cyberbullying at T1 exhibited significantly greater externalizing problems at T2 (β=0.25, 95% CI 0.06-0.45, P_{FDR}=.02), but not for internalizing problems (β=–0.01, 95% CI –0.20 to 0.19, P_{FDR}=.99) or deficits in inhibitory control (Correct Stop Rate: β=–0.02, 95% CI –0.26 to 0.21, P_{FDR}=.85; Stop Signal Reaction Time: β=–0.07, 95% CI –0.27 to 0.13, P_{FDR}=.85). Furthermore, cyberbullying victimization at T1 contributed to higher activation in the bilateral superior parietal gyri (left: β=0.36, 95% CI 0.10-0.61, P_{FDR}=.04; right: β=0.34, 95% CI 0.08-0.59, P_{FDR}=.04), right inferior parietal gyrus (β=0.32, 95% CI 0.07-0.57, P_{FDR}=.04), and right posterior cingulate cortex (β=0.34, 95% CI 0.09-0.60, P_{FDR}=.04) during error processing at T2. However, these neural alterations did not significantly mediate between cyberbullying victimization at T1 and externalizing problems at T2.

Conclusions: This longitudinal functional magnetic resonance imaging study investigates neural correlates of cyberbullying victimization in adolescents. By extending prior research that has relied primarily on cross-sectional or behavioral data, this research demonstrates that this form of victimization is associated with altered neural activation during error processing in later development. The pattern of nonsignificant impairment in inhibitory control and mediation to externalizing problems suggests that these neural impacts may be better characterized by a state of heightened sensitivity and compensatory engagement than by

XSL•FO
RenderX

direct damage. Overall, this study points to the error-processing network as a potential target for cognitive interventions and establishes a foundation for further exploration of other neural mechanisms between cyberbullying victimization and mental health outcomes.

(J Med Internet Res 2026;28:e75126) doi:10.2196/75126

KEYWORDS

cyberbullying victimization; mental health problems; error processing; functional magnetic resonance imaging; stop signal task

Introduction

Literature Review

Bullying is a prevalent risk factor for mental health disorders among adolescents [1,2]. In an era characterized by web-based relationships and interactions, cyberbullying has become increasingly common [3,4], which is defined as the repeated and deliberate harm by a perpetrator to a victim through mobile phones, computers, and other electronic devices [5]. A growing number of studies have linked cyberbullying victimization to mental health problems in adolescents, including more internalizing problems [6-8] such as depression [9,10] and anxiety [11,12] as well as more externalizing problems such as aggressive [13,14] and rule-breaking behaviors [15,16].

However, there exists a subset of studies that failed to find the significant effects of cyberbullying victimization on adolescents' mental health problems [17-20]. Such heterogeneity on this topic may be attributed to differences in experimental design (eg, cross-sectional vs longitudinal study) and variable measurement (eg, different questionnaire assessments) across studies, particularly the tendency to focus solely on direct observational relationships without sufficiently investigating the underlying psychological and physiological mechanisms that may explain how and for whom these effects occur [21]. Therefore, there is a current need for more longitudinal studies that use consistent, well-validated measures and incorporate biological perspectives to clarify the pathways linking cyberbullying victimization to mental health.

Grounded in the stress process model for internalizing problems [22] and the general aggression model for externalizing problems [23], the association between cyberbullying victimization and these psychopathological outcomes may be mediated by inhibitory control and its associated neural mechanisms. Inhibitory control, the ability to suppress inappropriate responses to stimuli [24], is negatively affected by various adverse childhood experiences (ACEs) [25], such as threat [26,27], deprivation [28,29], and abuse [30,31]. Inhibitory control deficits disrupt an individual's emotional and behavioral regulation, further contributing to internalizing and externalizing problems [32-34]. Given that the cyberbullying victimization in adolescents shares key characteristics with ACEs (ie, exposure to adversity, trauma, stress, or threats) and is closely linked to mental health problems, it is pertinent to explore whether diminished inhibitory control mediates the association between cyberbullying victimization and psychopathological outcomes.

According to the error detection theory [35,36], performance on inhibitory control is closely linked to the brain activation during error processing. Specifically, when individuals fail to inhibit a response, the cingulate cortex typically shows increased activation, as it reflects the perception of errors (ie, failed inhibition) [37,38] and supports subsequent cognitive and behavioral adjustments [39,40]. Moreover, previous studies have shown that ACEs impede the error monitoring circuitry, leading to blunted activation in cingulate cortices during failed inhibition [41,42]. Therefore, we focused on the cingulate cortices as regions of interest (ROIs) to investigate whether or how cyberbullying victimization decreases neural responses and contributes to inhibitory control deficits. The cingulate cortex further interacts with frontal [43,44] and parietal [45,46] cortices during error processing to maintain inhibitory control, which are key regions of the cognitive control network closely linked to information integration, attention regulation, and adaptive behaviors [47,48]. Given that ACEs have been suggested to reduce activation and connectivity within this network [49,50], we also included frontoparietal regions as ROIs to examine potential alterations of brain activation in cyberbullying victims.

Notably, several studies have indicated that adolescents with ACEs may exhibit excessive, rather than insufficient, activation in cingulate, frontal, and parietal cortices during response inhibition and error processing [51-53]. Such brain over-activation within the cognitive control network may reflect heightened sensitivity to errors and hypervigilance to adversity following trauma and stress caused by ACEs [54] and predicts the development of mental health problems such as depression and anxiety [55-57]. Since cyberbullying victimization is also traumatic [5] and stressful [58] for adolescents, it is plausible that a mechanism exists whereby cyberbullying victimization directly mediates excessive activation of cingulate, frontal, and parietal cortices, contributing to psychopathological outcomes.

McLoughlin et al [59] provided initial functional magnetic resonance imaging (fMRI) evidence supporting this potential mechanism. Their findings showed that viewing cyberbullying-related stimuli activated responses in many brain regions (eg, frontal, cingulate, and parietal cortices), closely associated with cognitive and emotional processing [60]. Despite the negative impacts of viewing cyberbullying content, adolescents may also be victims of cyberbullying rather than bystanders. Exploring the specific neural harms of cyberbullying victimization can contribute to a comprehensive understanding of the negative impacts of cyberbullying as well as the development of effective postvictimization interventions. However, based on their review [61] and to our knowledge, no longitudinal studies have investigated the neural alterations after adolescents experienced cyberbullying.

XSL•FO
RenderX

The abilities of inhibitory control and error monitoring improve progressively throughout adolescence, which are essential for maturing adaptive behaviors and reducing the risk of psychopathological outcomes [62,63]. This progression is supported by steadily increasing activation in related brain regions during neurodevelopment [64,65]. Notably, early adolescence is a critical period during which the neurodevelopment is particularly vulnerable to bullying [66] and highly sensitive to acceptance and rejection, especially through social media [67]. Therefore, it is important to examine the specific impacts of cyberbullying victimization on the adolescent brain using a longitudinal design. Building on the longitudinal associations, the temporal mechanisms of cyberbullying victimization and mental health problems, as well as the mediating role of biological factors between them, can be more clearly elucidated.

This Study

Using data from a large-scale longitudinal study in the United States—the Adolescent Brain Cognitive Development (ABCD) study [68]—this study aimed to explore the longitudinal associations among cyberbullying victimization, inhibitory control, brain activation during error processing, and mental health problems among adolescents. First, we explored how inhibitory control and ROIs' activation during error processing at T2, both measured using the stop signal task (SST), were influenced by cyberbullying victimization at T1. We broadly hypothesized that cyberbullying victimization at T1 predicted impaired inhibitory control and altered brain activation in the error-monitoring regions at T2. Second, we examined whether the cyberbullying victimization at T1 significantly predicted the mental health problems at T2. Based on the existing evidence, we hypothesized that individuals who were victims of cyberbullying at T1 would exhibit more internalizing and externalizing problems at T2 as compared with nonvictims at T1. Finally, we hypothesized that alterations in brain activity during error processing might mediate the relationship between cyberbullying victimization and mental health problems.

Methods

Study Design and Participants

This retrospective cohort study uses data from the ABCD study, which is an ongoing and 10-year longitudinal study of neurocognitive development in adolescents. To form a nationally representative sample, the ABCD study used school-based probability sampling to recruit its baseline cohort between 2016 and 2018. With school selection informed by gender, race/ethnicity, socioeconomic status, and urbanicity, it enrolled adolescents aged 9-10 years from 21 sites across the United States to mitigate selection bias [69]. As data pertaining to cyberbullying are available from the 2-year follow-up and imaging data are collected every other year, the data used in this study were obtained from the 2-year follow-up (T1) and 4-year follow-up (T2) of the ABCD Study Data Release 5.1. Participants were excluded if (1) they had missing data pertaining to cyberbullying and mental health problems, (2) they had missing data pertaining to fMRI of SST, (3) their imaging data were recommended for exclusion by the ABCD team, or (4) their behavioral performance of SST was recommended for exclusion by the ABCD team. For more details regarding the exclusion of imaging data and behavioral performance of SST, please refer to the ABCD Human Subjects Study (2024) [70].

The flowchart of participant exclusion is detailed in Figure S1 in Multimedia Appendix 1 [71-75], with most of the missing data due to the loss of mental health and brain imaging data. The Little's MCAR test suggested that the data were missing at random (χ^2_{24}= 31.07, P=.15) [76]. Therefore, participants who did not meet the criteria were deleted from the present analysis, and the final sample consisted of 1186 participants (mean age of 11.96 years; 551/1186, 46.5% female; Table 1 provides demographic information, and Table S1 in Multimedia Appendix 1 [71-75] provides comparison of included and excluded participants' characteristics). The adequacy of this sample size was evaluated using Monte Carlo simulations [77]: for the linear mixed models (LMMs) examining longitudinal associations between cyberbullying victimization, inhibitory control, brain activation during error processing, and mental health problems, these models provided 99.26% (95% CI 99.00-100.00) statistical power to detect effect size β=0.2 at α of .05. This study was reported in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement [78], and the completed checklist is provided in Multimedia Appendix 2 [71-75].

XSL•FO
RenderX

Table 1. Descriptive statistics for demographic characteristics at T1 in the Adolescent Brain Cognitive Development study.

Characteristics	Cyberbullying victims (*n*=61)	No cyberbullying experience (*n*=1125)	*t* test (*df*)	Chi-square (*df*)	*P* value
Age (years), mean (SD)	11.96 (0.58)	11.96 (0.64)	0.07 (1184)	__[a]	.95
Sex, n (%)			—	0.13 (1)	.72
Female	27 (44.3)	524 (46.6)			
Male	34 (55.7)	601 (53.4)			
Race and ethnicity, n (%)			—	2.94 (4)	.57
White	41 (67.2)	661 (58.8)			
Black	4 (6.6)	100 (8.9)			
Hispanic	9 (14.8)	219 (19.5)			
Asian	0 (0.0)	22 (2.0)			
Other	7 (11.5)	123 (10.9)			
Family income (US $), n (%)			—	0.48 (3)	.92
Less than 50,000	13 (21.3)	221 (19.6)			
50,000 through 99,999	19 (31.1)	324 (28.8)			
100,000 and greater	25 (41.0)	512 (45.5)			
Refused to answer/do not know	4 (6.6)	68 (6.0)			
Parents' highest education, n (%)			—	1.36 (3)	.72
High school education or less	8 (13.1)	143 (12.7)			
Some college	9 (14.8)	184 (16.4)			
Associate or Bachelor Degree	31 (50.8)	496 (44.1)			
Post-Graduate Degree	13 (21.3)	302 (26.8)			
Offline victimization, mean (SD)	12.45 (8.81)	12.40 (4.00)	0.10 (1184)	—	.92
Adverse childhood experiences, mean (SD)	1.47 (1.01)	1.44 (1.15)	0.21 (1184)	—	.83
Family conflict, mean (SD)	1.91 (1.80)	1.89 (1.20)	0.09 (1184)	—	.93
Social media use, hours, mean (SD)	0.47 (0.79)	0.51 (1.25)	0.27 (1184)	—	.79

[a]Not applicable.

Ethical Considerations

In accordance with the Declaration of Helsinki, the ABCD study's procedures were reviewed, exempted, and approved by the Centralized Institutional Review Board (IRB) at the University of California, San Diego (IRB# 160091), as well as each individual site's IRB. Informed consent was obtained from all participants and their legal guardians for both primary data collection and secondary analyses. The IRB confirmed that no additional consent was required for the secondary use of deidentified data. To ensure privacy and confidentiality, all data were deidentified and anonymized before analysis, with no personally identifiable information retained or disclosed. Participants and their families received financial compensation for their time, typically including US $200 for the parent or guardian, US $100 in gifts and gift cards for the child, and coverage for travel costs and sibling childcare. Images and Multimedia Appendix 1 [71-75] ensure participant anonymity, with no identifiable information included.

Questionnaire Measures

Cyberbullying Victimization

A modified version of the cyberbullying scale [79] was used to investigate whether adolescents were victims of cyberbullying. To report cyberbullying victimization, participants responded to the question, "In the past 12 months, have you been cyberbullied where someone intentionally tried to hurt you or be mean to you online, in texts or group texts, or on social media (like Instagram or Snapchat)?". Participants answered either "yes" or "no," and those who responded "yes" were coded as experiencing cyberbullying victimization [80,81].

Mental Health Problems

The mental health problems of adolescents were assessed using the parent-reported Child Behavior Checklist [82]. On a 3-point scale (0=not true to 2=very true or very often), parents reported on the adolescent's internalizing (T1: Cronbach α=0.88; T2: Cronbach α=0.87) and externalizing problems (Cronbach α=0.90; T2: Cronbach α=0.88). The internalizing score is the

XSL•FO
RenderX

sum of all items related to "anxious depressive symptoms (12 items; eg, Fears mistakes)" and "withdrawn depressive symptoms (8 items; eg, Lacks energy)," while the externalizing score is the sum of all items related to "rule-breaking behaviors (13 items; eg, Breaks rules)" and "aggressive behaviors (17 items; eg, Gets in fights)." The raw total score on each problem dimension was converted to norm-referenced *T*-score (range with a mean of 50 and SD of 10), with a higher number indicating greater symptom severity.

SST

The SST [83] assesses participants' behavioral response inhibition through 300 "Go" trials and 60 "Stop" trials. During frequent "Go" trials, participants are required to press a button promptly and correctly upon seeing an arrow indicating left or right. In contrast, on "Stop" trials, participants are required to refrain from pressing the button upon seeing a vertical arrow. Both the correct "Stop" rate and stop signal reaction time (a specific calculation method is provided in Multimedia Appendix 1 [71-75]) are indicators of inhibitory control ability. Furthermore, the stop signal delay, defined as the interval between Go and Stop trials, is calibrated to maintain response inhibition accuracy at approximately 50%. If a trial is successfully inhibited, the stop signal delay is increased by 50 milliseconds, whereas it is decreased by 50 milliseconds following unsuccessful inhibition. As SST ensures nearly 50% inhibition of failure, it is optimally suited for testing brain activation during error detection. For more details regarding SST design, please refer to Casey et al [68].

fMRI Data Acquisition and Preprocessing

The ABCD study used a harmonized neuroimaging protocol across 21 sites, using three 3T scanner platforms: Siemens Prisma (Siemens Healthineers), GE 750 (GE Healthcare), and Philips (Philips Healthcare). The ABCD's Data Analysis, Informatics, and Resource Core was responsible for the processing of all scan data, ensuring both high quality and uniform standards across participating sites. Corrections were made for head motion, B0, nonlinearity distortions, and between-scan motion for each participant [84-86]. The ROIs' values were derived from the average time courses of cortical regions, which were defined by FreeSurfer's anatomical parcellations and Desikan-Killiany Atlas [87]. General linear models in AFNI's 3dDeconvolve (Cox, 1996) were used to model task-related activation strength. Mean beta coefficients and standard errors for each ROI's time series were calculated. For more details regarding fMRI preprocessing techniques and the analysis of task-based fMRI data, please refer to Casey et al [68] and Hagler et al [88].

In this study, mean beta weights for failed stop contrasts (ie, incorrect stop contrasted with correct go and incorrect stop contrasted with correct stop) were used to measure brain activation during error processing [68]. Based on previous studies, we selected the cortical ROIs (Figure 1) most likely recruited by inhibitory control and error processing [43,44,53]: medial frontal, rostral middle frontal, rostral anterior cingulate, and posterior cingulate cortices. The superior and inferior parietal regions, which are closely related to attention, sensation, and adaptive behaviors [45,89], were also examined. The beta weights for each ROI in the right and left hemispheres were used in subsequent analyses.

Figure 1. Regions of interest for failed inhibition analysis in the Adolescent Brain Cognitive Development study. The regions of interest included medial frontal, rostral middle frontal, rostral anterior cingulate, and posterior cingulate cortices. The brain activation during error processing was reflected by the mean beta weights for failed stop contrasts (ie, incorrect stop contrasted with correct go and incorrect stop contrasted with correct stop). All *P* values were adjusted for multiple comparisons using the false discovery rate method, and *P*<.05 indicated significance.

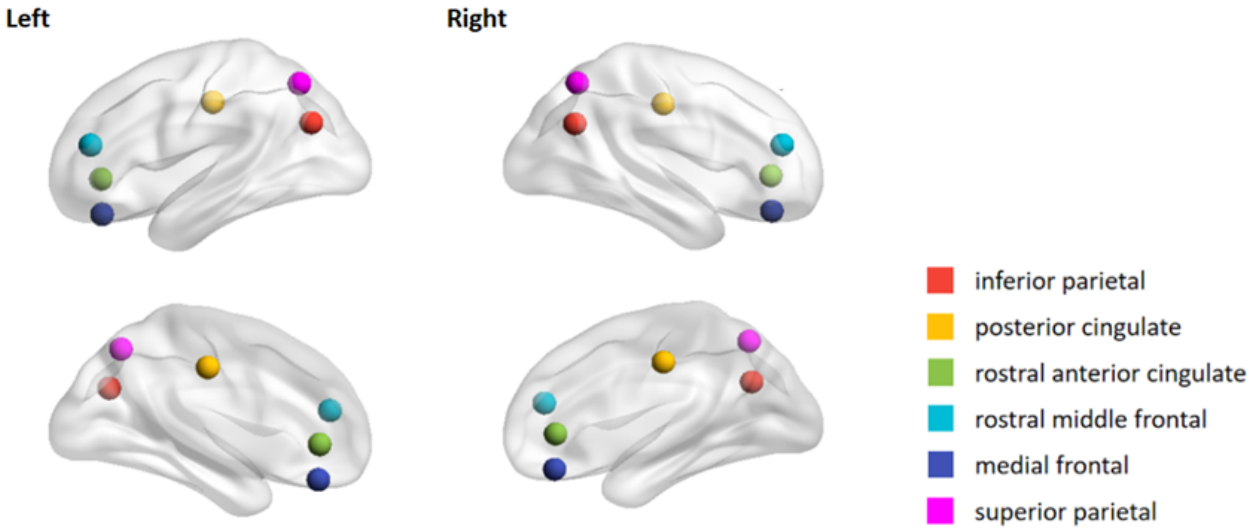

Covariates

Potential confounders include adolescents' age, biological sex, race/ethnicity, family income, parents' highest educational attainment, offline victimization, ACEs, family conflict, and social media use at T1. Adolescents' biological sex was coded into 1=male and 2=female. Race was coded into 5 categorical variables, including 1=White, 2=Black, 3=Hispanic, 4=Asian, and 5=other races. Family income was categorized into 3 brackets, ranging from 1 (less than US $50,000) to 3 (US

XSL•FO
RenderX

$100,000 and greater). Parents' educational attainment was the highest educational degree in the family, ranging from 1 (less than a high school diploma) to 4 (postgraduate degree). Other covariates remained in the continuous format to be controlled (scoring details are provided in Multimedia Appendix 1 [71-75]).

Statistical Analyses

All data analyses were conducted on R (version 4.4.1; R Development R Core Team, 2024). LMMs in the lme4 package [90] were used to explore the retrospective longitudinal association between cyberbullying victimization, inhibitory control, brain activation during error processing, and mental health problems. First, mental health outcomes (ie, internalizing and externalizing problems) at T2 served as the dependent variables, with whether cyberbullying victimization at T1 as the independent variable, which was included as a fixed effect in the model. Research sites and family id were treated as random effects to reflect the nested structure of the data. Second, the performance on inhibitory control (ie, correct "Stop" rate and stop signal reaction time from the SST) and the ROIs' activation during error processing (ie, beta weight of failed stop from the SST) at T2 were used as dependent variables, while cyberbullying victimization at T1 remained the independent variable. MRI scanner type was added as a random effect in these brain measurement models.

Further analyses followed up any significant findings by examining whether cyberbullying victimization at T1 was associated with mental health outcomes at T2 via altered brain activation patterns during error processing. According to Cole and Maxwell's recommendation [91] regarding 2 waves of longitudinal studies, a half-longitudinal mediation model was established to explore the mediating role of brain activation during error processing (Figure 2). Specifically, we calculated the regression coefficient from T1 cyberbullying victimization to T2 brain measures (path a) and the regression coefficient from T1 brain measures to T2 mental health outcomes (path b). The significance of the mediating effect (path a × b) was calculated using bootstrap analysis with 5000 resamples to generate bias-corrected CI. When the 95% CI does not contain 0, then the mediating effect of brain activation is statistically significant.

Adolescents' biological sex, age, race/ethnicity, family income, and highest parental education attainment were included as demographic covariates in all analyses. In the analyses predicting brain activation of SST, we additionally covaried mean framewise displacement and its quadratic effects during the fMRI scan [92]. Furthermore, to control for the developmental effect of the dependent variable in each LMM, we separately covaried T1 level of mental health, inhibitory control, and activation of ROIs during error processing in the corresponding model [71,93]. For example, the T1 level of externalizing problems was included as a covariate in the LMM exploring the link between whether cyberbullying victimization at T1 and T2 levels of externalizing problems. Finally, we used the inverse probability weighting method [94] in the survey package [95] to test the robustness of the results from all LLMs and mediation models. False discovery rate (FDR) was applied to correct for multiple comparisons, and $P_{FDR}<.05$ was considered to reach the significant level.

Figure 2. Conceptual half-longitudinal mediation model examining neural mechanisms linking cyberbullying victimization to mental health outcomes in US adolescents. For all analyses, participants' age, biological sex, race/ethnicity, family income, parents' highest educational attainment, offline victimization, adverse childhood experiences, family conflict, and social media use were included as covariates. Path a (T1 victimization → T2 brain measures) additionally covaried mean framewise displacement during the fMRI scan and the T1 activation level of regions of interest, while path b (T1 brain measures → T2 mental health) added T1 level of mental health to the control.

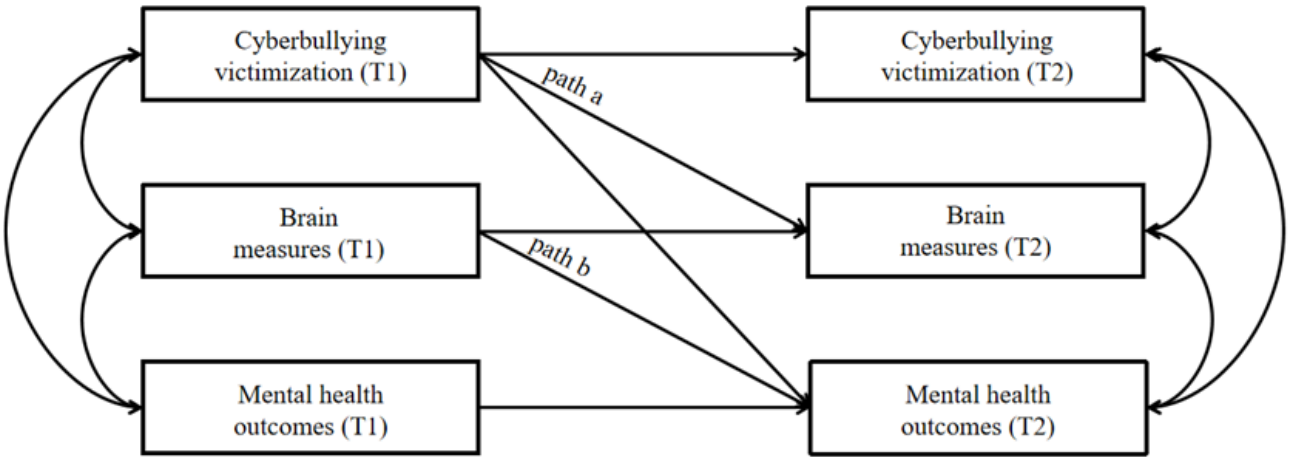

Results

Descriptive Statistics for Demographic Characteristics

Descriptive statistics for demographic characteristics at T1 are presented in Table 1. Independent samples *t* tests and chi-square test indicated that no significant difference was found between the cyberbullying victims and participants with no cyberbullying experiences in age, offline victimization, ACEs, family conflict, and social media use, $t_{1184}< 0.27$, $P>.79$, sex, race/ethnicity, family income, or parents' highest educational attainment, $\chi^2_4< 2.94$, $P>.57$.

XSL•FO
RenderX

Associations Between Cyberbullying at T1 and ROIs' Activation During Error Processing at T2

The associations between cyberbullying victimization at T1 and ROIs' activation during error processing at T2 are shown in Table 2. For the contrast of incorrect stop versus correct go, cyberbullying victimization at T1 was significantly associated with higher activation of bilateral superior parietal gyri (left: β=0.36, 95% CI 0.10-0.61, SE 0.01, P_{FDR}=.04; right: β=0.34, 95% CI 0.08-0.59, SE 0.01, P_{FDR}=.04), right inferior parietal gyrus (β=0.32, 95% CI 0.07-0.57, SE 0.01, P_{FDR}=.04), and right posterior cingulate cortex (β=0.34, 95% CI 0.09-0.60, SE 0.01, P_{FDR}=.04) at T2. For the contrast of incorrect stop versus correct stop, significant heightened activation was also observed in the bilateral superior parietal gyri (left: β=0.32, 95% CI 0.08-0.52, SE 0.01, P_{FDR}=.04; right: β=0.30, 95% CI 0.05-0.48, SE 0.01, P_{FDR}=.04), bilateral inferior parietal gyri (left: β=0.30, 95% CI 0.06-0.50, SE 0.01, P_{FDR}=.04; right: β=0.30, 95% CI 0.04-0.47, SE 0.01, P_{FDR}=.04), and right posterior cingulate cortex (β=0.34, 95% CI 0.10-0.60, SE 0.01, P_{FDR}=.04). No other significant ROIs' activated alteration during error processing at T2 were predicted by cyberbullying victimization at T1 after correction for multiple comparisons.

Table 2. Associations between cyberbullying victimization at T1 and cortical regions of interest activation during error processing at T2 in the Adolescent Brain Cognitive Development study.

	Left hemisphere					Right hemisphere				
	β[a]	95% CI	SE[b]	*P* value[c]	P_{FDR} value[c]	β	95% CI	SE	*P* value[c]	P_{FDR} value[c]
Incorrect stop contrasted with correct go										
Superior parietal	0.36	0.10 to 0.61	0.01	.005[d]	.04[d]	0.34	0.08 to 0.59	0.01	.009[d]	.04[d]
Inferior parietal	0.26	0.00 to 0.51	0.01	.047[d]	.09	0.32	0.07 to 0.57	0.01	.01[d]	.04[d]
Medial frontal	0.22	–0.03 to 0.48	0.03	.08	.14	0.20	–0.06 to 0.45	0.03	.13	.20
Rostral middle frontal	0.15	–0.10 to 0.41	0.02	.24	.28	0.17	–0.09 to 0.42	0.02	.20	.27
Rostral anterior cingulate	–0.01	–0.27 to 0.24	0.02	.91	.91	–0.03	–0.29 to 0.22	0.02	.81	.88
Posterior cingulate	0.27	0.01 to 0.53	0.01	.04[d]	.09	0.34	0.09 to 0.60	0.01	.008[d]	.04[d]
Incorrect stop contrasted with correct stop										
Superior parietal	0.32	0.08 to 0.52	0.01	.006[d]	.04[d]	0.30	0.05 to 0.48	0.01	.01[d]	.04[d]
Inferior parietal	0.30	0.06 to 0.50	0.01	.01[d]	.04[d]	0.30	0.04 to 0.47	0.01	.02[d]	.04[d]
Medial frontal	0.17	–0.08 to 0.38	0.04	.10	.17	0.22	–0.06 to 0.45	0.03	.11	.17
Rostral middle frontal	–0.03	–0.28 to 0.23	0.02	.83	.83	0.05	–0.21 to 0.31	0.02	.70	.76
Rostral anterior cingulate	–0.06	–0.31 to 0.20	0.02	.66	.76	–0.02	–0.27 to 0.21	0.02	.63	.76
Posterior cingulate	0.28	0.04 to 0.58	0.01	.04[d]	.08	0.34	0.10 to 0.60	0.01	.006[d]	.04[d]

[a]β: standardized coefficient.

[b]SE of β.

[c]All *P* values were adjusted for multiple comparisons using the false discovery rate (FDR) method.

[d]*P*<.05.

Associations Between Cyberbullying at T1 and Mental Health Outcomes/Inhibitory Control at T2

Descriptive statistics for mental health outcomes and inhibitory control at T1 and T2 are presented in Table 3. The LMMs revealed that cyberbullying victimization at T1 significantly predicted increased externalizing problems at T2 (β=0.25, 95% CI 0.06-0.45, SE 0.87, P_{FDR}=.02). In contrast, internalizing problems and inhibitory control performance at T2 were not significantly associated with cyberbullying victimization at T1 (Table 4).

XSL•FO
RenderX

Table 3. Descriptive statistics for mental health outcomes and stop signal task (SST) behavioral performance at T1 and T2 in the Adolescent Brain Cognitive Development study.

	T1		T2	
	Cyberbullying victims (n=61)	No cyberbullying experience (n=1125)	Cyberbullying victims (n=61)	No cyberbullying experience (n=1125)
Mental health outcomes, mean (SD)				
Internalizing problems	49.39 (9.33)	47.65 (10.34)	48.85 (8.76)	47.85 (10.72)
Externalizing problems	46.69 (8.92)	44.17 (9.10)	46.97 (8.90)	43.08 (8.77)
SST behavioral performance, mean (SD)				
Correct stop rate (%)	50.10 (6.60)	50.90 (5.40)	48.80 (6.90)	49.20 (6.30)
Stop signal reaction time (ms)	284.41 (58.00)	276.84 (57.24)	264.49 (56.36)	261.73 (50.77)

Table 4. Associations between cyberbullying victimization at T1 and mental health outcomes/inhibitory control at T2 in the Adolescent Brain Cognitive Development study.

	β[a]	95% CI	SE[b]	*P* value[c]	P_{FDR} value[c]
Mental health outcomes					
Internalizing problems	–0.01	–0.20 to 0.19	1.06	.99	.99
Externalizing problems	0.25	0.06 to 0.45	0.87	.01	.02[d]
SST[e] behavioral performance					
Correct stop rate (%)	–0.02	–0.26 to 0.21	0.01	.85	.85
Stop signal reaction time	–0.07	–0.27 to 0.13	5.14	.49	.85

[a]β: standardized coefficient.

[b]SE of β.

[c]All *P* values were adjusted for multiple comparisons using the false discovery rate (FDR) method.

[d]*P*<.05.

[e]SST: stop signal task.

Mediation Analyses

As shown in Table 5, the mediation effects of parietal gyri or posterior cingulate cortex activated alteration during error processing between cyberbullying victimization at T1 and externalizing problems at T2 did not reach a significant level after correction for multiple comparisons (P_{FDR}s>.74; specific path coefficients are presented in Table S2 in Multimedia Appendix 1 [71-75]).

Table 5. Mediation effects of brain activation during error processing between cyberbullying victimization at T1 and externalizing problems at T2 in the Adolescent Brain Cognitive Development study.

	Left hemisphere				Right hemisphere			
	Indirect effect	95% CI	*P* value[a]	P_{FDR} value[a]	Indirect effect	95% CI	*P* value[a]	P_{FDR} value[a]
Incorrect stop contrasted with correct go								
Superior parietal	0.004	–0.003 to 0.006	.66	>.99	0.001	–0.005 to 0.005	>.99	>.99
Inferior parietal	0.000	–0.003 to 0.004	.87	>.99	–0.004	–0.007 to 0.002	.47	>.99
Posterior cingulate	–0.005	–0.006 to 0.004	.62	>.99	–0.006	–0.007 to 0.005	.65	>.99
Incorrect stop contrasted with correct stop								
Superior parietal	0.003	–0.002 to 0.004	.74	.82	0.003	–0.001 to 0.005	.64	.82
Inferior parietal	0.003	–0.002 to 0.004	.82	.82	0.006	–0.001 to 0.008	.74	.82
Posterior cingulate	0.005	–0.004 to 0.006	.54	.82	0.006	–0.001 to 0.008	.44	.82

[a]All *P* values were adjusted for multiple comparisons using the false discovery rate (FDR) method.

XSL•FO
RenderX

Inverse Probability Weighting Analysis

After inverse probability weighting, the results remained consistent with the primary analyses (Tables S3-S6 in Multimedia Appendix 1 [71-75]). Specifically, cyberbullying victimization at T1 was significantly associated with higher activation in the bilateral superior parietal gyri, inferior parietal gyri, and right posterior cingulate cortex at T2 (Table S3 in Multimedia Appendix 1 [71-75]). Behaviorally, while it significantly predicted increased externalizing problems, its associations with internalizing problems and inhibitory control performance still did not reach a significant level (Table S4 in Multimedia Appendix 1 [71-75]). Moreover, the mediation effects of activation in the parietal gyri or posterior cingulate cortex during error processing between cyberbullying victimization and externalizing problems were also nonsignificant (Tables S5 and S6 in Multimedia Appendix 1 [71-75]).

Discussion

Principal Findings

This retrospective cohort study explored the longitudinal association between cyberbullying victimization, inhibitory control, brain activation during error processing, and mental health problems among American adolescents. Adolescents who experienced cyberbullying victimization demonstrated heightened activation in the parietal gyri and posterior cingulate cortices during error processing. Additionally, these victims showed more subsequent externalizing rather than internalizing problems. However, the mediating effects of brain-activated alteration on the link between cyberbullying victimization and mental health outcomes did not reach significant levels.

Brain Activation During Error Processing

Our study showed that cyberbullying victimization significantly contributed to higher activation in parietal regions and posterior cingulate cortices during error processing, which may reflect their high sensitivity and vigilance to adversity, thus avoiding potential injury [54]. The parietal cortex plays a critical role in an individual's attention regulation, goal orientation, and motor coordination [45,88]. Increased activation of the parietal cortex may not only make cyberbullying victims more vigilant and sensitive to adversity, but also underpin their subsequent behavioral adjustments to distance themselves from potential injury. Victims of cyberbullying experienced web-based, anonymous, and spatially/temporally unlimited harm (eg, verbal scold and satirize) in a web-based environment [10,96], potentially leading them to distance themselves from various harm in any real-life situations. This generalized avoidance tendency could be supported by the parietal gyri, given their involvement in relevant cognitive functions.

The posterior cingulate cortex, a core structure in the default mode network, is closely associated with an individual's self-reflection, attention shift, and emotion regulation [97,98]. However, its interpretation in our findings requires caution, as this region typically deactivates during goal-directed tasks in typical populations [99]. Therefore, the observed higher activation among victims may not reflect an absolute increase above a resting baseline, but rather a reduction in the deactivation. This pattern could signify either a failure to suppress self-referential processing, potentially due to intrusive thoughts related to victimization experiences, or the active recruitment of cognitive resources for heightened internal monitoring and emotion regulation to stabilize the internal state during adversity [100]. It is also worth noting that the significant change in activation of the inferior parietal and posterior cingulate cortices was only observed in the right hemisphere. Combined with the fact that no significant difference in frontal and anterior cingulate activation was found between cyberbully victims and nonvictims, it seems that the effects of cyberbullying partially lead to an increased cognitive load, but do not widely damage the error-monitoring circuits and cognitive control network.

From the perspective of latent vulnerability theory [101,102], a nonsignificant change in the anterior cingulate cortex but significantly higher activation in posterior cingulate and parietal cortices could be viewed as adaptive responses and adjustments, rather than mere damage after cyberbullying victimization. Furthermore, these changes could serve as generalized neural impacts of this experience, as this study only used category variables to define cyberbullying but did not consider its intensity. However, given that this form of victimization has been shown to elicit higher cortisol secretion levels and greater perceived stress in adolescents [58], which could disrupt the hypothalamic-pituitary-adrenal axis feedback [103], further studies are warranted to further explore the relationship between characteristic neural alterations and the intensity of cyberbullying victimization.

Externalizing Problems

The relationship between cyberbullying victimization and externalizing problems was consistent with previous studies [13-15]. Yet the mediating effects of brain over-activation during error processing between cyberbullying victimization and externalizing problems did not reach a significant level. As mentioned above, the higher activation of parietal and posterior cingulate cortices in SST may serve as adaptive adjustments, rather than severe neural circuit damage. Moreover, dealing with real-life "errors" may generally require more cognitive resources and neural load than processing failed inhibition in SST. In other words, the standardized stimuli and the monotonous failed inhibition in SST may not adequately elicit the actual neural changes of cyberbullying victimization. Consequently, despite certain differences in brain activation measured by the SST between cyberbullying victims and nonvictims, such differences could not sufficiently mediate the relationship between cyberbullying victimization and externalizing problems.

The sustained performance of inhibitory control in cyberbullying victims also provides a potential explanation for the insufficient mediated effect of brain-activated alteration. Previous studies suggested that poor inhibitory control is a predictor of externalizing problems [104,105]. However, no significant predictive difference was found in inhibitory control performance between cyberbullying victims and nonvictims of this study, which aligned with Lim et al [53] and Carrion et al

XSL•FO
RenderX

[106]. Therefore, we speculate that the higher neurological load in the parietal gyri and posterior cingulate cortices did not lead to more failed inhibition in SST (ie, decompensatory effect), but rather maintained the performance of inhibitory control (ie, compensatory effect) with heightened attention and self-reflection, resulting in no evidence for a significant mediation effect for externalizing problems. In general, both the potential lack of representation and the protective bias role of altered brain activation measured by SST may contribute to this study's failure to indicate the significant mediated brain effects.

Based on existing studies, the link between cyberbullying victimization and externalizing problems may be effectively explained by other mechanisms. One potential mechanism is that the experience of cyberbullying elicited the victim's aggressive tendencies (eg, hostility and retaliatory motivation) [14,107,108]. For example, Tong et al [14] engaged 464 rural adolescents to examine the association between cyberbullying victimization and malevolent creativity. The results revealed that victims mediated increased hostile attribution leading to serious malevolent creativity. Another potential mechanism is that cyberbullying victimization led to the adoption of externalizing behaviors through imitation [107,109]. Specifically, according to the general learning model [110], we speculate that cyberbullying is a situational factor that influences victims' observational learning processes. These influences manifested in internal states, including perception and emotion, and subsequently contributed to externalizing problems. For instance, victims of cyberbullying perceived the characteristics of the bullying conducted in the web-based environment (eg, anonymity and no strength differential). They may develop a positive attitude towards such characteristics, thereby engaging in cyberbullying [107,109].

Overall, this study revalidated the longitudinal association between cyberbullying and externalizing problems but failed to indicate the significant neural mediation measured by SST. More studies are needed in the future to comprehensively explore other potential mediated mechanisms.

Internalizing Problems

Findings in previous studies regarding whether the cyberbullying victimization could predict internalizing problems (eg, depression and anxiety) are controversial [9,10,17,18]. Similarly with Chu et al [17] and Frison et al [18], this study found that cyberbullying victimization did not significantly predict internalizing problems. One potential explanation is that individuals have the ability to regulate their emotions, which might mitigate the internalized impacts of cyberbullying [111,112]. Since the emotional regulation is closely linked to the activation of the posterior cingulate cortex and the performance of inhibitory control [97,100], we speculate that heightened activation of the posterior cingulate cortex during error processing may reflect cyberbullying victims' increased load of emotional regulation to cope with adversity. Such increased load may also serve as the neural compensation and protection, through which cyberbullying victimization could not directly and severely disrupt the individual's internal state.

Another potential explanation is the different measurement methods used for cyberbullying and emotions. Specifically, cyberbullying is a trait measure (ie, assessing the frequency of cyberbullying within a certain time frame by self-report) [113], and emotions are conceptualized as state processes (ie, measurement dependent on current arousal) [114]. The mismatch between these measurement methods leads to an inability to sensitively capture the immediate impact of cyberbullying on emotions. Therefore, we speculate that the significant longitudinal effect of cyberbullying on internalizing problems only emerges when there is sufficient emotional destructiveness and prolonged duration of victimization, which aligns with the hypothesis proposed by Frison et al [18]. Given that this study used a categorical variable to define cyberbullying, which may be more reflective of general victimization, future studies should explore whether there is a longitudinal association between the intensity and frequency of cyberbullying and the severity of neural damage, as well as internalizing problems [21].

Theoretical and Practical Implications

For theoretical implications, our findings promote the integration of the error detection theory [35,36] and psychopathological outcomes. The potential neurocompensatory response observed in the posterior cingulate and parietal cortices during failed inhibition could provide a novel account for the stress process model [22], illustrating how chronic social stress may manifest as heightened neural sensitivity to errors while maintaining behavioral performance. Similarly, these neural findings offer a mechanistic elaboration of the general aggression model [23,115] for externalizing problems, specifying how victimization might strain cognitive control resources without immediate behavioral manifestations. For practical implications, the significant longitudinal association with externalizing problems highlights the importance of incorporating cyberbullying victimization screening into mental health assessments at an early stage. Furthermore, the neural signature identified here might inform the development of targeted interventions, such as cognitive training protocols designed to reinforce compensatory processes, which hold potential for enhancing emotion regulation and mitigating the risk of externalizing behaviors in affected youth [116,117].

Limitations and Future Directions

This study has a few limitations. The first limitation stems from data availability in the ABCD Study Release 5.1. The analytical sample size was reduced primarily due to the partial availability of the 4-year follow-up (T2) data. This attrition likely contributed to the relatively low prevalence of cyberbullying victims (≈5%), which may limit statistical power to detect smaller effects. Furthermore, although our analysis did not reveal systematic differences in baseline characteristics between included and excluded participants, the potential selection bias (eg, the unexpected similarity in offline victimization levels between victims and nonvictims) should be considered when interpreting the results and their generalizability.

Second, the analysis was constrained to only 2 time points (T1 and T2) available in this data release. This limits the representation of developmental trajectories and the more rigorous longitudinal mediation analysis (eg, a T1→T2→T3

XSL•FO
RenderX

pathway). Future studies should incorporate larger samples of affected individuals and more waves of neuroimaging data across development to examine how cyberbullying victimization influences mental health via neural changes over time.

Third, cyberbullying was analyzed as a categorical variable (victims vs nonvictims) in this study. This approach was adopted because the item measuring frequency/intensity had a high rate of missing data in our analytical sample, which precludes examining the impact of victimization severity on mental health problems, especially internalizing problems. Moreover, this study only focused on externalizing problems, internalizing problems, and brain activation during error processing. Cyberbullying may be significantly associated with other mental/physical health problems (eg, eating disorder, suicide, and sleep disturbance) [79,118,119] as well as social media use [67,71], it is necessary to comprehensively examine the longitudinal impacts of cyberbullying on a variety of health problems and brain states (eg, morphology, resting-state connectivity, and other dynamic states).

Conclusions

Drawing on the ABCD study, this fMRI investigation advances prior cross-sectional or behavioral research by adopting a longitudinal neurobiological perspective. It reveals that cyberbullying victimization predicts significantly heightened activation in parietal and posterior cingulate regions during error processing, alongside increased externalizing problems over time. The absence of significant deficits in inhibitory control and mediation effects suggests that these neural impacts may be better characterized by a state of heightened sensitivity and compensatory engagement after victimization rather than direct impairment. In general, these findings identify the error-processing network as a candidate neural system for future study and motivate continued exploration of other neural pathways linking cyberbullying victimization to mental health, with the ultimate aim of mitigating negative consequences.

Acknowledgments

This study was supported by the GuangDong Basic and Applied Basic Research Foundation (#2023A1515110394) and the Guangzhou Basic and Applied Basic Research Foundation (#2024A04J5080). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript. The authors acknowledge that generative artificial intelligence (AI) technology was used to identify and edit word errors. Data used in the preparation of this article were obtained with permission from the Adolescent Brain Cognitive Development (ABCD) Study, held in the NIMH Data Archive. This is a multisite, longitudinal study designed to recruit more than 10,000 children aged 9-10 years and follow them over 10 years into early adulthood. The ABCD Study is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041048, U01DA050989, U01DA051016, U01DA041022, U01DA051018, U01DA051037, U01DA050987, U01DA041174, U01DA041106, U01DA041117, U01DA041028, U01DA041134, U01DA050988, U01DA051039, U01DA041156, U01DA041025, U01DA041120, U01DA051038, U01DA041148, U01DA041093, U01DA041089, U24DA041123, and U24DA041147. ABCD consortium investigators designed and implemented the study, and provided data, but did not necessarily participate in the analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the National Institutes of Health or ABCD consortium investigators. The ABCD data repository grows and changes over time.

Data Availability

Data were obtained from the 2-year follow-up and 4-year follow-up of the Adolescent Brain Cognitive Development (ABCD) Study (data release 5.1). The data that support the findings of this study are available at the NIMH Data Archive upon data access request.

Authors' Contributions

XZ, CX, and BQ contributed to the conceptualization of the study. The methodology was developed by XZ, YC, and BQ. XZ, YC, and BQ prepared the original draft of the manuscript, while XZ, CX, CZ, YC, and BQ contributed to the review and editing of the manuscript. Supervision was provided by YC and BQ, and funding acquisition was managed by BQ. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional material.
[DOCX File , 159 KB - jmir_v28i1e75126_app1.docx]

Multimedia Appendix 2
STROBE checklist.

XSL•FO
RenderX

[DOCX File , 22 KB - jmir_v28i1e75126_app2.docx]

References

1. Brunstein Klomek A, Marrocco F, Kleinman M, Schonfeld IS, Gould MS. Bullying, depression, and suicidality in adolescents. J Am Acad Child Adolesc Psychiatry 2007;46(1):40-49. [doi: 10.1097/01.chi.0000242237.84925.18] [Medline: 17195728]
2. Fraguas D, Díaz-Caneja CM, Ayora M, Durán-Cutilla M, Abregú-Crespo R, Ezquiaga-Bravo I, et al. Assessment of school anti-bullying interventions: a meta-analysis of randomized clinical trials. JAMA Pediatr 2021;175(1):44-55 [FREE Full text] [doi: 10.1001/jamapediatrics.2020.3541] [Medline: 33136156]
3. Kennedy RS. Bullying trends in the United States: a meta-regression. Trauma Violence Abuse 2021;22(4):914-927. [doi: 10.1177/1524838019888555] [Medline: 31789118]
4. Patchin J, Hinduja S. Cyberbullying among tweens in the United States: prevalence, impact, and helping behaviors. J Early Adolesc 2021;42(3):414-430. [doi: 10.1177/02724316211036740]
5. Englander E, Donnerstein E, Kowalski R, Lin C, Parti K. Defining cyberbullying. Pediatrics 2017;140(Suppl 2):S148-S151. [doi: 10.1542/peds.2016-1758U] [Medline: 29093051]
6. Aboujaoude E, Savage MW, Starcevic V, Salame WO. Cyberbullying: review of an old problem gone viral. J Adolesc Health 2015;57(1):10-18. [doi: 10.1016/j.jadohealth.2015.04.011] [Medline: 26095405]
7. Baldry AC, Farrington DP, Sorrentino A. “Am I at risk of cyberbullying”? a narrative review and conceptual framework for research on risk of cyberbullying and cybervictimization: the risk and needs assessment approach. Aggress Violent Behav 2015;23:36-51. [doi: 10.1016/j.avb.2015.05.014]
8. Kowalski RM, Giumetti GW, Schroeder AN, Lattanner MR. Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth. Psychol Bull 2014;140(4):1073-1137. [doi: 10.1037/a0035618] [Medline: 24512111]
9. Gámez-Guadix M, Orue I, Smith PK, Calvete E. Longitudinal and reciprocal relations of cyberbullying with depression, substance use, and problematic internet use among adolescents. J Adolesc Health 2013;53(4):446-452. [doi: 10.1016/j.jadohealth.2013.03.030] [Medline: 23721758]
10. Li L, Jing R, Jin G, Song Y. Longitudinal associations between traditional and cyberbullying victimization and depressive symptoms among young Chinese: a mediation analysis. Child Abuse Negl 2023;140:106141. [doi: 10.1016/j.chiabu.2023.106141] [Medline: 36958096]
11. Fahy AE, Stansfeld SA, Smuk M, Smith NR, Cummins S, Clark C. Longitudinal associations between cyberbullying involvement and adolescent mental health. J Adolesc Health 2016;59(5):502-509. [doi: 10.1016/j.jadohealth.2016.06.006] [Medline: 27528471]
12. Rose CA, Tynes BM. Longitudinal associations between cybervictimization and mental health among U.S. adolescents. J Adolesc Health 2015;57(3):305-312 [FREE Full text] [doi: 10.1016/j.jadohealth.2015.05.002] [Medline: 26115909]
13. Lu Y, Avellaneda F, Torres ED, Rothman EF, Temple JR. Adolescent cyberbullying and weapon carrying: cross-sectional and longitudinal associations. Cyberpsychol Behav Soc Netw 2019;22(3):173-179 [FREE Full text] [doi: 10.1089/cyber.2018.0463] [Medline: 30614736]
14. Tong D, Shi Y, Gu X, Lu P. Bullying victimization and malevolent creativity in rural adolescents: the longitudinal mediational role of hostile attribution. Cyberpsychol Behav Soc Netw 2024;27(6):420-425. [doi: 10.1089/cyber.2023.0499] [Medline: 38511278]
15. Gong X, Huebner ES, Tian L. Bullying victimization and developmental trajectories of internalizing and externalizing problems: the moderating role of locus of control among children. Res Child Adolesc Psychopathol 2021;49(3):351-366. [doi: 10.1007/s10802-020-00752-2] [Medline: 33404945]
16. Jung Y, Leventhal B, Kim YS, Park TW, Lee S, Lee M, et al. Cyberbullying, problematic internet use, and psychopathologic symptoms among Korean youth. Yonsei Med J 2014;55(3):826-830 [FREE Full text] [doi: 10.3349/ymj.2014.55.3.826] [Medline: 24719154]
17. Chu X, Fan C, Lian S, Zhou Z. Does bullying victimization really influence adolescents' psychosocial problems? a three-wave longitudinal study in China. J Affect Disord 2019;246:603-610. [doi: 10.1016/j.jad.2018.12.103] [Medline: 30605879]
18. Frison E, Subrahmanyam K, Eggermont S. The short-term longitudinal and reciprocal relations between peer victimization on facebook and adolescents' well-being. J Youth Adolesc 2016;45(9):1755-1771. [doi: 10.1007/s10964-016-0436-z] [Medline: 26880284]
19. Pabian S, Vandebosch H. An investigation of short-term longitudinal associations between social anxiety and victimization and perpetration of traditional bullying and cyberbullying. J Youth Adolesc 2016;45(2):328-339. [doi: 10.1007/s10964-015-0259-3] [Medline: 25687265]
20. van den Eijnden R, Vermulst A, van Rooij AJ, Scholte R, van de Mheen D. The bidirectional relationships between online victimization and psychosocial problems in adolescents: a comparison with real-life victimization. J Youth Adolesc 2014;43(5):790-802 [FREE Full text] [doi: 10.1007/s10964-013-0003-9] [Medline: 23979296]
21. Kwan I, Dickson K, Richardson M, MacDowall W, Burchett H, Stansfield C, et al. Cyberbullying and children and young people's mental health: a systematic map of systematic reviews. Cyberpsychol Behav Soc Netw 2020;23(2):72-82 [FREE Full text] [doi: 10.1089/cyber.2019.0370] [Medline: 31977251]

XSL•FO
RenderX

22. Wang L. The effects of cyberbullying victimization and personality characteristics on adolescent mental health: an application of the stress process model. Youth Soc 2021;54(6):935-956. [doi: 10.1177/0044118x211008927]
23. Zhao Y, An W, Zhang Y, Yang C, Li W, Li J, et al. Cyberbullying victimization and perpetration in adolescents: a longitudinal moderated mediation model. J Youth Adolesc 2025;54(4):890-902. [doi: 10.1007/s10964-024-02105-4] [Medline: 39466550]
24. Aron AR. The neural basis of inhibition in cognitive control. Neuroscientist 2007;13(3):214-228. [doi: 10.1177/1073858407299288] [Medline: 17519365]
25. Rahapsari S, Levita L. The impact of adverse childhood experiences on cognitive control across the lifespan: a systematic review and meta-analysis of prospective studies. Trauma Violence Abuse 2025;26(4):712-733 [FREE Full text] [doi: 10.1177/15248380241286812] [Medline: 39396188]
26. Westermann N, Busching R, Klein AM, Warschburger P. The longitudinal interplay between adverse peer experiences and self-regulation facets: a prospective analysis during middle childhood. Res Child Adolesc Psychopathol 2024;52(2):293-308 [FREE Full text] [doi: 10.1007/s10802-023-01117-1] [Medline: 37782355]
27. Savopoulos P, Brown S, Anderson PJ, Gartland D, Bryant C, Giallo R. Intimate partner violence during infancy and cognitive outcomes in middle childhood: Results from an Australian community-based mother and child cohort study. Child Dev 2022;93(4):e396-e411. [doi: 10.1111/cdev.13736] [Medline: 35137950]
28. Frenkel TI, Donzella B, Frenn KA, Rousseau S, Fox NA, Gunnar MR. Moderating the risk for attention deficits in children with pre-adoptive adversity: the protective role of shorter duration of out of home placement and children's enhanced error monitoring. J Abnorm Child Psychol 2020;48(9):1115-1128 [FREE Full text] [doi: 10.1007/s10802-020-00671-2] [Medline: 32607754]
29. Golm D, Sarkar S, Mackes NK, Fairchild G, Mehta MA, Rutter M, ERA Young Adult Follow Up study team, et al. The impact of childhood deprivation on adult neuropsychological functioning is associated with ADHD symptom persistence. Psychol Med 2021;51(15):2675-2684. [doi: 10.1017/S0033291720001294] [Medline: 32419675]
30. Amédée LM, Cyr C, Jean-Thorn A, Hébert M. Executive functioning in child victims of sexual abuse: a multi-informant comparative study. Child Abuse Negl 2024;152:106737 [FREE Full text] [doi: 10.1016/j.chiabu.2024.106737] [Medline: 38564916]
31. Wei X, Lü W. Childhood abuse and depressive symptoms in adolescents: affective inhibitory control as a mediator and sensory processing sensitivity as a moderator. Child Abuse Negl 2024;154:106957. [doi: 10.1016/j.chiabu.2024.106957] [Medline: 39053224]
32. Kahle S, Utendale WT, Widaman KF, Hastings PD. Parasympathetic regulation and inhibitory control predict the development of externalizing problems in early childhood. J Abnorm Child Psychol 2018;46(2):237-249. [doi: 10.1007/s10802-017-0305-6] [Medline: 28493111]
33. Poland SE, Monks CP, Tsermentseli S. Cool and hot executive function as predictors of aggression in early childhood: differentiating between the function and form of aggression. Br J Dev Psychol 2016;34(2):181-197. [doi: 10.1111/bjdp.12122] [Medline: 26615980]
34. Sætren SS, Augusti E, Hafstad GS. Affective inhibitory control and risk for internalizing problems in adolescents exposed to child maltreatment: a population-based study. J Abnorm Psychol 2021;130(2):113-125. [doi: 10.1037/abn0000582] [Medline: 33315413]
35. Coles MG, Scheffers MK, Holroyd CB. Why is there an ERN/Ne on correct trials? response representations, stimulus-related components, and the theory of error-processing. Biol Psychol 2001;56(3):173-189. [doi: 10.1016/s0301-0511(01)00076-x] [Medline: 11399349]
36. Falkenstein M, Hoormann J, Christ S, Hohnsbein J. ERP components on reaction errors and their functional significance: a tutorial. Biol Psychol 2000;51(2-3):87-107. [doi: 10.1016/s0301-0511(99)00031-9] [Medline: 10686361]
37. Botvinick MM, Cohen JD, Carter CS. Conflict monitoring and anterior cingulate cortex: an update. Trends Cogn Sci 2004;8(12):539-546. [doi: 10.1016/j.tics.2004.10.003] [Medline: 15556023]
38. Yeung N, Botvinick MM, Cohen JD. The neural basis of error detection: conflict monitoring and the error-related negativity. Psychol Rev 2004;111(4):931-959. [doi: 10.1037/0033-295x.111.4.939] [Medline: 15482068]
39. Ridderinkhof KR, Ullsperger M, Crone EA, Nieuwenhuis S. The role of the medial frontal cortex in cognitive control. Science 2004;306(5695):443-447. [doi: 10.1126/science.1100301] [Medline: 15486290]
40. Stevens M, Kiehl K, Pearlson G, Calhoun V. Brain network dynamics during error commission. Hum Brain Mapp 2009;30(1):24-37 [FREE Full text] [doi: 10.1002/hbm.20478] [Medline: 17979124]
41. Fava NM, Trucco EM, Martz ME, Cope LM, Jester JM, Zucker RA, et al. Childhood adversity, externalizing behavior, and substance use in adolescence: mediating effects of anterior cingulate cortex activation during inhibitory errors. Dev Psychopathol 2019;31(4):1439-1450 [FREE Full text] [doi: 10.1017/S0954579418001025] [Medline: 30585564]
42. Stevens JS, Ely TD, Sawamura T, Guzman D, Bradley B, Ressler KJ, et al. Childhood maltreatment predicts reduced inhibition-related activity in the rostral anterior cingulate in PTSD, but not trauma-exposed controls. Depress Anxiety 2016;33(7):614-622 [FREE Full text] [doi: 10.1002/da.22506] [Medline: 27062552]
43. Iannaccone R, Hauser TU, Staempfli P, Walitza S, Brandeis D, Brem S. Conflict monitoring and error processing: new insights from simultaneous EEG-fMRI. Neuroimage 2015;105:395-407. [doi: 10.1016/j.neuroimage.2014.10.028] [Medline: 25462691]

XSL•FO
RenderX

44. Wessel JR. An adaptive orienting theory of error processing. Psychophysiology 2018;55(3). [doi: 10.1111/psyp.13041] [Medline: 29226960]
45. Corbetta M, Shulman GL. Control of goal-directed and stimulus-driven attention in the brain. Nat Rev Neurosci 2002;3(3):201-215. [doi: 10.1038/nrn755] [Medline: 11994752]
46. Dosenbach NUF, Fair DA, Miezin FM, Cohen AL, Wenger KK, Dosenbach RAT, et al. Distinct brain networks for adaptive and stable task control in humans. Proc Natl Acad Sci U S A 2007;104(26):11073-11078 [FREE Full text] [doi: 10.1073/pnas.0704320104] [Medline: 17576922]
47. Breukelaar IA, Antees C, Grieve SM, Foster SL, Gomes L, Williams LM, et al. Cognitive control network anatomy correlates with neurocognitive behavior: a longitudinal study. Hum Brain Mapp 2017;38(2):631-643 [FREE Full text] [doi: 10.1002/hbm.23401] [Medline: 27623046]
48. Niendam TA, Laird AR, Ray KL, Dean YM, Glahn DC, Carter CS. Meta-analytic evidence for a superordinate cognitive control network subserving diverse executive functions. Cogn Affect Behav Neurosci 2012;12(2):241-268 [FREE Full text] [doi: 10.3758/s13415-011-0083-5] [Medline: 22282036]
49. McLaughlin KA, Weissman D, Bitrán D. Childhood adversity and neural development: a systematic review. Annu Rev Dev Psychol 2019;1:277-312 [FREE Full text] [doi: 10.1146/annurev-devpsych-121318-084950] [Medline: 32455344]
50. Teicher MH, Samson JA, Anderson CM, Ohashi K. The effects of childhood maltreatment on brain structure, function and connectivity. Nat Rev Neurosci 2016;17(10):652-666. [doi: 10.1038/nrn.2016.111] [Medline: 27640984]
51. Keator DB, Salgado F, Madigan C, Murray S, Norris S, Amen D. Adverse childhood experiences, brain function, and psychiatric diagnoses in a large adult clinical cohort. Front Psychiatry 2024;15:1401745 [FREE Full text] [doi: 10.3389/fpsyt.2024.1401745] [Medline: 39469474]
52. Lackner CL, Santesso DL, Dywan J, O'Leary DD, Wade TJ, Segalowitz SJ. Adverse childhood experiences are associated with self-regulation and the magnitude of the error-related negativity difference. Biol Psychol 2018;132:244-251. [doi: 10.1016/j.biopsycho.2018.01.006] [Medline: 29309827]
53. Lim L, Hart H, Mehta M, Simmons A, Mirza K, Rubia K. Neural correlates of error processing in young people with a history of severe childhood abuse: an fMRI study. Am J Psychiatry 2015;172(9):892-900. [doi: 10.1176/appi.ajp.2015.14081042] [Medline: 25882324]
54. Santesso DL, Dzyundzyak A, Segalowitz SJ. Age, sex and individual differences in punishment sensitivity: factors influencing the feedback-related negativity. Psychophysiology 2011;48(11):1481-1489. [doi: 10.1111/j.1469-8986.2011.01229.x] [Medline: 21711354]
55. Abi-Dargham A, Moeller SJ, Ali F, DeLorenzo C, Domschke K, Horga G, et al. Candidate biomarkers in psychiatric disorders: state of the field. World Psychiatry 2023;22(2):236-262 [FREE Full text] [doi: 10.1002/wps.21078] [Medline: 37159365]
56. Hajcak G, Klawohn J, Meyer A. The utility of event-related potentials in clinical psychology. Annu Rev Clin Psychol 2019;15:71-95. [doi: 10.1146/annurev-clinpsy-050718-095457] [Medline: 31067414]
57. Olvet DM, Hajcak G. The error-related negativity (ERN) and psychopathology: toward an endophenotype. Clin Psychol Rev 2008;28(8):1343-1354 [FREE Full text] [doi: 10.1016/j.cpr.2008.07.003] [Medline: 18694617]
58. González-Cabrera J, Calvete E, León-Mejía A, Pérez-Sancho C, Peinado J. Relationship between cyberbullying roles, cortisol secretion and psychological stress. Comput Human Behav 2017;70:153-160. [doi: 10.1016/j.chb.2016.12.054]
59. McLoughlin LT, Shan Z, Broadhouse KM, Lagopoulos J, Winks N, Hermens DF. Elucidating the neurobiology of cyberbullying using functional Magnetic Resonance Imaging (fMRI): a hypothesis. Aggress Violent Behav 2020;50:101360. [doi: 10.1016/j.avb.2019.101360]
60. McLoughlin LT, Shan Z, Broadhouse KM, Winks N, Simcock G, Lagopoulos J, et al. Neurobiological underpinnings of cyberbullying: a pilot functional magnetic resonance imaging study. Hum Brain Mapp 2020;41(6):1495-1504 [FREE Full text] [doi: 10.1002/hbm.24890] [Medline: 31797458]
61. McLoughlin LT, Lagopoulos J, Hermens DF. Cyberbullying and adolescent neurobiology. Front Psychol 2020;11:1511 [FREE Full text] [doi: 10.3389/fpsyg.2020.01511] [Medline: 32670171]
62. Crone EA, Dahl RE. Understanding adolescence as a period of social-affective engagement and goal flexibility. Nat Rev Neurosci 2012;13(9):636-650. [doi: 10.1038/nrn3313] [Medline: 22903221]
63. Steinberg L. Cognitive and affective development in adolescence. Trends Cogn Sci 2005;9(2):69-74. [doi: 10.1016/j.tics.2004.12.005] [Medline: 15668099]
64. Rubia K. Functional brain imaging across development. Eur Child Adolesc Psychiatry 2013;22(12):719-731 [FREE Full text] [doi: 10.1007/s00787-012-0291-8] [Medline: 22729957]
65. Tamnes CK, Ostby Y, Fjell AM, Westlye LT, Due-Tønnessen P, Walhovd KB. Brain maturation in adolescence and young adulthood: regional age-related changes in cortical thickness and white matter volume and microstructure. Cereb Cortex 2010;20(3):534-548. [doi: 10.1093/cercor/bhp118] [Medline: 19520764]
66. Cooke EM, Connolly EJ, Boisvert DL, Hayes BE. A systematic review of the biological correlates and consequences of childhood maltreatment and adverse childhood experiences. Trauma Violence Abuse 2023;24(1):156-173. [doi: 10.1177/15248380211021613] [Medline: 34105421]

XSL•FO
RenderX

67. Crone EA, Konijn EA. Media use and brain development during adolescence. Nat Commun 2018;9(1):588 [FREE Full text] [doi: 10.1038/s41467-018-03126-x] [Medline: 29467362]
68. Casey B, Cannonier T, Conley MI, Cohen AO, Barch DM, Heitzeg MM, ABCD Imaging Acquisition Workgroup. The adolescent brain cognitive development (ABCD) study: imaging acquisition across 21 sites. Dev Cogn Neurosci 2018;32:43-54 [FREE Full text] [doi: 10.1016/j.dcn.2018.03.001] [Medline: 29567376]
69. Garavan H, Bartsch H, Conway K, Decastro A, Goldstein RZ, Heeringa S, et al. Recruiting the ABCD sample: design considerations and procedures. Dev Cogn Neurosci 2018;32:16-22 [FREE Full text] [doi: 10.1016/j.dcn.2018.04.004] [Medline: 29703560]
70. ABCD Human Subjects Study. National Institute for Health. 2024. URL: https://nda.nih.gov/abcd [accessed 2026-01-20]
71. He X, Hu J, Yin M, Zhang W, Qiu B. Screen media use affects subcortical structures, resting-state functional connectivity, and mental health problems in early adolescence. Brain Sci 2023;13(10):1452 [FREE Full text] [doi: 10.3390/brainsci13101452] [Medline: 37891820]
72. Verbruggen F, Aron AR, Band GP, Beste C, Bissett PG, Brockett AT, et al. A consensus guide to capturing the ability to inhibit actions and impulsive behaviors in the stop-signal task. Elife 2019 Apr 29;8:e46323 [FREE Full text] [doi: 10.7554/eLife.46323] [Medline: 31033438]
73. Prinstein MJ, Boergers J, Vernberg EM. Overt and relational aggression in adolescents: social-psychological adjustment of aggressors and victims. J Clin Child Psychol 2001 Dec;30(4):479-491. [doi: 10.1207/S15374424JCCP3004_05] [Medline: 11708236]
74. Felitti VJ, Anda RF, Nordenberg D, Williamson DF, Spitz AM, Edwards V, et al. Reprint of: relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults: the adverse childhood experiences (ACE) study. Am J Prev Med 2019 Jun;56(6):774-786. [doi: 10.1016/j.amepre.2019.04.001] [Medline: 31104722]
75. Lanz M, Maino E. Family environment scale. In: Encyclopedia of Quality of Life and Well-Being Research. Dordrecht: Springer; 2014:2170-2173.
76. Little RJA. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 1988;83(404):1198-1202. [doi: 10.1080/01621459.1988.10478722]
77. Green P, MacLeod CJ. SIMR?: an R package for power analysis of generalized linear mixed models by simulation. Methods Ecol Evol 2016;7(4):493-498. [doi: 10.1111/2041-210x.12504]
78. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007;370(9596):1453-1457 [FREE Full text] [doi: 10.1016/S0140-6736(07)61602-X] [Medline: 18064739]
79. Stewart RW, Drescher CF, Maack DJ, Ebesutani C, Young J. The development and psychometric investigation of the cyberbullying scale. J Interpers Violence 2014;29(12):2218-2238. [doi: 10.1177/0886260513517552] [Medline: 24424252]
80. Arnon S, Brunstein Klomek A, Visoki E, Moore TM, Argabright ST, DiDomenico GE, et al. Association of cyberbullying experiences and perpetration with suicidality in early adolescence. JAMA Netw Open 2022;5(6):e2218746 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.18746] [Medline: 35759263]
81. Trompeter N, Testa A, Raney JH, Jackson DB, Al-Shoaibi AAA, Ganson KT, et al. The association between adverse childhood experiences (ACEs), bullying victimization, and internalizing and externalizing problems among early adolescents: examining cumulative and interactive associations. J Youth Adolesc 2024;53(3):744-752 [FREE Full text] [doi: 10.1007/s10964-023-01907-2] [Medline: 38066316]
82. Achenbach TM, Ruffle TM. The child behavior checklist and related forms for assessing behavioral/emotional problems and competencies. Pediatr Rev 2000;21(8):265-271. [doi: 10.1542/pir.21-8-265] [Medline: 10922023]
83. Logan G, Cowan W, Davis K. On the ability to inhibit simple and choice reaction time responses: a model and a method. J Exp Psychol Hum Percept Perform 1984;10(2):276-291. [doi: 10.1037//0096-1523.10.2.276] [Medline: 6232345]
84. Cox RW. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 1996;29(3):162-173. [doi: 10.1006/cbmr.1996.0014] [Medline: 8812068]
85. Holland D, Kuperman JM, Dale AM. Efficient correction of inhomogeneous static magnetic field-induced distortion in echo planar imaging. Neuroimage 2010;50(1):175-183 [FREE Full text] [doi: 10.1016/j.neuroimage.2009.11.044] [Medline: 19944768]
86. Jovicich J, Czanner S, Greve D, Haley E, van der Kouwe A, Gollub R, et al. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. Neuroimage 2006;30(2):436-443. [doi: 10.1016/j.neuroimage.2005.09.046] [Medline: 16300968]
87. Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 2006;31(3):968-980. [doi: 10.1016/j.neuroimage.2006.01.021] [Medline: 16530430]
88. Hagler DJ, Hatton S, Cornejo MD, Makowski C, Fair DA, Dick AS, et al. Image processing and analysis methods for the adolescent brain cognitive development study. Neuroimage 2019;202:116091 [FREE Full text] [doi: 10.1016/j.neuroimage.2019.116091] [Medline: 31415884]
89. Andersen RA, Buneo CA. Intentional maps in posterior parietal cortex. Annu Rev Neurosci 2002;25:189-220. [doi: 10.1146/annurev.neuro.25.112701.142922] [Medline: 12052908]

XSL•FO
RenderX

90. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Soft 2015;67(1). [doi: 10.18637/jss.v067.i01]
91. Cole DA, Maxwell SE. Testing mediational models with longitudinal data: questions and tips in the use of structural equation modeling. J Abnorm Psychol 2003;112(4):558-577. [doi: 10.1037/0021-843X.112.4.558] [Medline: 14674869]
92. Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage 2012;59(3):2142-2154 [FREE Full text] [doi: 10.1016/j.neuroimage.2011.10.018] [Medline: 22019881]
93. Yang FN, Liu TT, Wang Z. Corticostriatal connectivity mediates the reciprocal relationship between parent-reported sleep duration and impulsivity in early adolescents. J Child Psychol Psychiatry 2023;64(11):1545-1554. [doi: 10.1111/jcpp.13843] [Medline: 37248201]
94. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res 2013;22(3):278-295. [doi: 10.1177/0962280210395740] [Medline: 21220355]
95. Lumley T. Analysis of complex survey samples. J Stat Soft 2004;9(8). [doi: 10.18637/jss.v009.i08]
96. Baiden P, Tadeo SK. Investigating the association between bullying victimization and suicidal ideation among adolescents: evidence from the 2017 youth risk behavior survey. Child Abuse Negl 2020;102:104417. [doi: 10.1016/j.chiabu.2020.104417] [Medline: 32113078]
97. Leech R, Kamourieh S, Beckmann CF, Sharp DJ. Fractionating the default mode network: distinct contributions of the ventral and dorsal posterior cingulate cortex to cognitive control. J Neurosci 2011;31(9):3217-3224 [FREE Full text] [doi: 10.1523/JNEUROSCI.5626-10.2011] [Medline: 21368033]
98. Leech R, Sharp DJ. The role of the posterior cingulate cortex in cognition and disease. Brain 2014;137(Pt 1):12-32 [FREE Full text] [doi: 10.1093/brain/awt162] [Medline: 23869106]
99. Chaarani B, Hahn S, Allgaier N, Adise S, Owens MM, Juliano AC, ABCD Consortium. Baseline brain function in the preadolescents of the ABCD study. Nat Neurosci 2021;24(8):1176-1186 [FREE Full text] [doi: 10.1038/s41593-021-00867-9] [Medline: 34099922]
100. Pessoa L. How do emotion and motivation direct executive control? Trends Cogn Sci 2009;13(4):160-166 [FREE Full text] [doi: 10.1016/j.tics.2009.01.006] [Medline: 19285913]
101. McCrory EJ, Viding E. The theory of latent vulnerability: reconceptualizing the link between childhood maltreatment and psychiatric disorder. Dev Psychopathol 2015;27(2):493-505. [doi: 10.1017/S0954579415000115] [Medline: 25997767]
102. Puetz VB, McCrory E. Exploring the relationship between childhood maltreatment and addiction: a review of the neurocognitive evidence. Curr Addict Rep 2015;2(4):318-325 [FREE Full text] [doi: 10.1007/s40429-015-0073-8] [Medline: 26550550]
103. O'Connor DB, Thayer JF, Vedhara K. Stress and health: a review of psychobiological processes. Annu Rev Psychol 2021;72(1):663-688. [doi: 10.1146/annurev-psych-062520-122331] [Medline: 32886587]
104. Chaku N, Yan R, Kelly DP, Zhang Z, Lopez-Duran N, Weigard AS, et al. 100 days of adolescence: elucidating externalizing behaviors through the daily assessment of inhibitory control. Res Child Adolesc Psychopathol 2024;52(1):93-110 [FREE Full text] [doi: 10.1007/s10802-023-01071-y] [Medline: 37405589]
105. Young SE, Friedman NP, Miyake A, Willcutt EG, Corley RP, Haberstick BC, et al. Behavioral disinhibition: liability for externalizing spectrum disorders and its genetic and environmental relation to response inhibition across adolescence. J Abnorm Psychol 2009;118(1):117-130 [FREE Full text] [doi: 10.1037/a0014657] [Medline: 19222319]
106. Carrion VG, Garrett A, Menon V, Weems CF, Reiss AL. Posttraumatic stress symptoms and brain function during a response-inhibition task: an fMRI study in youth. Depress Anxiety 2008;25(6):514-526. [doi: 10.1002/da.20346] [Medline: 17598145]
107. Barlett C, Gentile D. Attacking others online: the formation of cyberbullying in late adolescence. Psychol Pop Media Cult 2012;1(2):123-135. [doi: 10.1037/a0028113]
108. Wang C, Li X, Xia L. Long-term effect of cybervictimization on displaced aggressive behavior across two years: mutually predicting mediators of hostile emotion and moral disengagement. Comput Hum Behav 2023;141:107611. [doi: 10.1016/j.chb.2022.107611]
109. Barlett CP. Cyberbullying as a learned behavior: theoretical and applied implications. Children (Basel) 2023;10(2):325 [FREE Full text] [doi: 10.3390/children10020325] [Medline: 36832455]
110. Gentile DA, Groves CL, Gentile JR. The general learning model. In: Learning by Playing. England: Oxford University Press; 2014:121-142.
111. Kassis W, Artz S, Scambor C, Scambor E, Moldenhauer S. Finding the way out: a non-dichotomous understanding of violence and depression resilience of adolescents who are exposed to family violence. Child Abuse Negl 2013;37(2-3):181-199. [doi: 10.1016/j.chiabu.2012.11.001] [Medline: 23260117]
112. Lin L, Chien Y, Chen Y, Wu C, Chiou H. Bullying experiences, depression, and the moderating role of resilience among adolescents. Front Public Health 2022;10:872100 [FREE Full text] [doi: 10.3389/fpubh.2022.872100] [Medline: 35692326]
113. Chun J, Lee J, Kim J, Lee S. An international systematic review of cyberbullying measurements. Comput Hum Behav 2020;113:106485. [doi: 10.1016/j.chb.2020.106485]

XSL•FO
RenderX

114. Mulligan K, Scherer KR. Toward a working definition of emotion. Emot Rev 2012;4(4):345-357. [doi: 10.1177/1754073912445818]
115. Teng H, Zhu L, Zhang X, Qiu B. The relationship between cognitive resource consumption during gameplay and postgame aggressive behaviors: between-subjects experiment. JMIR Serious Games 2023;11:e48317 [FREE Full text] [doi: 10.2196/48317] [Medline: 37990585]
116. Kasturiratna KTAS, Hartanto A, Chen CHY, Tong EMW, Majeed NM. Umbrella review of meta-analyses on the risk factors, protective factors, consequences and interventions of cyberbullying victimization. Nat Hum Behav 2025;9(1):101-132. [doi: 10.1038/s41562-024-02011-6] [Medline: 39516404]
117. Zhang X, Teng H, Zhu L, Qiu B. Short-term exposure to aggressive card game: releasing emotion without escalating post-game aggression. Front Psychol 2025;16:1505360 [FREE Full text] [doi: 10.3389/fpsyg.2025.1505360] [Medline: 40236967]
118. Cheng CM, Chu J, Ganson KT, Trompeter N, Testa A, Jackson DB, et al. Cyberbullying and eating disorder symptoms in US early adolescents. Int J Eat Disord 2023;56(12):2336-2342. [doi: 10.1002/eat.24034] [Medline: 37671456]
119. Nagata JM, Yang JH, Singh G, Kiss O, Ganson KT, Testa A, et al. Cyberbullying and sleep disturbance among early adolescents in the U.S. Acad Pediatr 2023;23(6):1220-1225 [FREE Full text] [doi: 10.1016/j.acap.2022.12.007] [Medline: 36581100]

Abbreviations

ABCD: Adolescent Brain Cognitive Development
ACE: adverse childhood experience
FDR: false discovery rate
fMRI: functional magnetic resonance imaging
IRB: Institutional Review Board
LMM: linear mixed model
ROI: regions of interest
SST: stop signal task
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani, S Brini; submitted 30.Mar.2025; peer-reviewed by DA Lopez, M Yin, S Chen; comments to author 12.Sep.2025; accepted 08.Jan.2026; published 18.Feb.2026.

Please cite as:
Zhang X, Xie C, Chen Y, Qiu B
Associations Among Cyberbullying Victimization, Inhibitory Control, Neural Activation of Error Processing, and Mental Health Problems in Adolescents: Neuroimaging, Retrospective Longitudinal Cohort Study Using the Adolescent Brain Cognitive Development Data
J Med Internet Res 2026;28:e75126
URL: https://www.jmir.org/2026/1/e75126
doi:10.2196/75126
PMID:

©Xuanyu Zhang, Chengyan Xie, Yu Chen, Boyu Qiu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Adherence, Acceptability, and Sexual Health Outcomes of the Odeya App–Based Intervention for Sexual Distress in Women With Endometriosis: Randomized Controlled Mixed Methods Trial

Selina Marie Kronthaler[1], MSc; Eden Kosman[1], MSc; Tatjana Tissen-Diabaté[2], MSc; Elena Mühle[1]; Luzia Weber-Schallauer[1], MSc; Therese Pross[3], MD, MBA; Maria Margarete Karsten[3], MD; Klaus Michael Beier[1], MD, PhD; Laura Hatzler[1,4], MD

[1]Institute of Sexology and Sexual Medicine, Charité - Universitätsmedizin Berlin, Berlin, Germany

[2]Institute of Social Medicine, Epidemiology and Health Economics, Charité – Universitätsmedizin Berlin, Berlin, Germany

[3]Department of Gynecology and Breast Center, Charité – Universitätsmedizin Berlin, Berlin, Germany

[4]Department of Urology, Charité - Universitätsmedizin Berlin, Berlin, Germany

Corresponding Author:
Laura Hatzler, MD
Institute of Sexology and Sexual Medicine
Charité - Universitätsmedizin Berlin
Charité Platz 1
Berlin, 10117
Germany
Phone: 49 30 450 617 139
Fax: 49 30 450 529 992
Email: laura.hatzler@charite.de

Abstract

Background: Evidence-based interventions effectively treat sexual dysfunctions. Up to 13.5% of women with gynecological conditions are affected, yet access to therapy is limited. Self-guided digital interventions may offer scalable, accessible first-line support.

Objective: This randomized controlled mixed methods pilot trial evaluated adherence, acceptance, and safety of the Odeya app and changes in sexual and health outcomes among women with sexual dysfunctions and endometriosis.

Methods: Following online and flyer-based recruitment, participants completed an online screening and were randomized to either an intervention group (IG) receiving 8 self-guided app modules targeting biopsychosocial aspects of sexuality or to a control group (CG) receiving routine care. Self-administered online questionnaires were completed at baseline (T0), midintervention (T1), postintervention (T2), and 6-month follow-up (T3). Standardized instruments assessed acceptance (Client Satisfaction Questionnaire-Internet [CSQ-I] and German mHealth App Usability Questionnaire [G-MAUQ]), safety (Inventory for the balanced assessment of Negative Effects of Psychotherapy-Online Intervention), sexual health (Female Sexual Distress Scale-Desire/Arousal/Orgasm [FSDS-DAO], Female Sexual Function Index-German version [FSFI-d], and Partnership Questionnaire), and overall health (Patient-Reported Outcome Measurement Information System-29-Item Profile, Beck Depression Inventory-II, and Generalized Anxiety Disorder-7). Adherence indicators included module completion, dropout rates, and symptom tracker use. Group differences were examined descriptively and using Cohen *d*. Qualitative data were collected through free-list questionnaires from dropouts (n=11) and interviews with completers (IG: n=3; CG: n=2).

Results: A total of 60 women (mean age 31.12, SD 6.67 years) with confirmed or suspected endometriosis and sexual distress (FSDS-DAO score >18) were randomized to the IG (n=29) or CG (n=31). IG participants completed on average 61.2% (4.9/8) of modules; the dropout rate was 65.5% (19/29). Emotional strain, time demands, and technical issues were key barriers causing dropout, while persona-based stories facilitated engagement. Participants wished for more professional interaction. IG completers (n=10, 34.5%) showed lower baseline depression and anxiety but higher sexual distress. Satisfaction was high (CSQ-I=26.60; G-MAUQ=5.38). Although some adverse health changes were reported, findings indicate safety. FSDS-DAO scores decreased in both groups, with mean reductions from baseline of −10.39, −12.61, and −14.98 in the IG and −3.68, −14.83, and −6.92 in the CG from T1 to T3, respectively. Moderate to large between-group effects favoring the IG were observed at T1 (d=−0.66) and T3 (d=−0.79). Sexual function (FSFI-d) improved only in the IG (T1-T3: d=0.16-1.00). Qualitative findings highlighted rediscovery

XSL•FO
RenderX

of positive sexual experiences, improved communication, and increased openness. Both groups reported improvements in anxiety, depression, and physical functioning, with additional gains in emotion regulation, distress reduction, and body awareness reported in the IG. Women emphasized symptom complexity and a need for more professional guidance.

Conclusions: The self-guided intervention was well accepted and showed preliminary improvements among completers. Adherence and sustained engagement seemed shaped by baseline psychosocial health, pointing to a need for tailored adaptations and larger confirmatory trials.

Trial Registration: German Clinical Trials Register DRKS00034351; https://drks.de/search/en/trial/DRKS00034351

(J Med Internet Res 2026;28:e86042) doi:10.2196/86042

KEYWORDS

app-based intervention; digital health intervention; endometriosis; randomized controlled trial; sexual distress; sexual dysfunction

Introduction

Overview

Sexual health is a fundamental aspect of overall well-being and quality of life [1-3]. Sexual dysfunctions, such as problems with sexual desire, arousal, orgasm, or pain, accompanied by clinically relevant distress, as defined by the *ICD-11* (*International Classification of Diseases, 11th Revision*) [4], were reported by 17.5% of women in a representative German study [5]. The biopsychosocial risk factors include relationship-related difficulties, poor mental health, chronic conditions, cultural factors, and lack of knowledge and experience [6-10].

Among individuals with chronic diseases, including gynecological conditions such as endometriosis, sexual health issues are particularly pronounced [11]. In a recent representative German study, 75.2% of those with chronic conditions reported problems in sexual function, and 19.3% met *ICD-11* criteria for sexual dysfunction, with 2.56-fold higher odds compared to individuals without chronic conditions [12]. Among all chronic condition groups, women with gynecological conditions showed the highest prevalence of problems in sexual function (84.3%), and 13.5% of these women met criteria for sexual dysfunction with distress [12]. Chronic diseases frequently contribute to sexual dysfunction through physical, hormonal, psychological, and treatment-related factors [13-15], with downstream impacts on mental health, relationship satisfaction, and health care costs [16-19].

Endometriosis affects 6.8% of women worldwide [20] and substantially impairs sexual health and quality of life [21,22]. The association between endometriosis and sexual dysfunction is well-documented, with many reporting sexual pain, decreased sexual satisfaction, and overall reduced sexual functioning [11,23]. Surgical interventions may further worsen outcomes [24]. Care access is hindered by shame and stigma, insufficient awareness, high costs, and gaps in provider training [12,25-30].

Effective treatment for sexual dysfunction with distress requires a personalized, multimodal, interdisciplinary approach addressing the individual's set of biopsychosocial etiological factors [10,31-36]. Recommended strategies combine somatic interventions (eg, pelvic floor therapy and hormonal and medical treatments) [37-40] with sex and couples therapy (eg, sensate focus and communication training), educational components (eg, psychoeducation on anatomy and physiology) [33,41], and lifestyle-based strategies such as adapted physical activity [42,43]. Furthermore, evidence supports strengthening the mind-body integration through exercises on body perception, mindfulness, and reflective techniques [41,44-46]. Underlying mental health and somatic conditions should always be addressed in interdisciplinary approaches, ideally by a multidisciplinary team trained in sexual medicine [35,47].

Despite high prevalence, sexual dysfunctions remain underreported and undertreated in Germany [12]. Persistent access barriers—including limited specialist availability, long waiting lists, and regional disparities—contrast with strong interest in digital interventions such as app- or web-based programs with exercises and educational content on sexual health [12,29,48].

Digital self-help interventions can help overcome these barriers [16], offering accessibility, anonymity, and cost-effectiveness [49-51]. Online treatments—especially cognitive-behavioral and mindfulness-based programs—have shown moderate to large effects on sexual function and satisfaction [50,52-56]. In 2019, the German Digital Healthcare Act introduced a regulatory framework for the approval and reimbursement of software as a medical device, referred to as digital health applications (DiGAs) [57]. Currently, 2 sexual health-related DiGAs are approved and permanently listed in the registry of the German competent authority (Bundesinstitut für Arzneimittel und Medizinprodukte [BfArM]): HelloBetter Vaginismus for vaginismus [50,58] and Kranus Edera for erectile dysfunction [59]. Additional DiGAs targeting urogenital health include Endo App for endometriosis symptom management [60,61] and Kranus Lutera for lower tract symptoms [62]. Negotiated 90-day prices averaged €221.09 (US $258.70) [63], with full statutory reimbursement.

Patient-centered digital intervention development [64] has demonstrated high user satisfaction [65,66] and positive outcomes across diverse populations, including cancer [67-69], mental health conditions such as depression and anxiety [70-72], and chronic pain [73-75]. Yet adherence and engagement remain challenging—especially in self-guided formats [76-82]—and effects on engagement are mixed [73,83-89]. A major gap persists in digital interventions addressing sexual distress in women with gynecological conditions such as endometriosis, underscoring the need for tailored, patient-centered solutions [90]. Existing digital tools rarely address sexuality as a relational

XSL•FO
RenderX

resource; insights from syndyastic sex therapy may offer a useful framework for future digital models [91].

Objective

This study evaluated the pilot implementation of a self-guided smartphone app intervention (Odeya) in women with sexual distress and diagnosed or suspected endometriosis. The Odeya app was developed using evidence-based content for patients with sexual dysfunctions in conjunction with clinically relevant distress and was tailored, within a patient-centered framework, to the specific needs of women with endometriosis. The primary objectives were to assess (1) adherence, (2) acceptance, and (3) safety. The secondary objective was to exploratorily examine effects on sexual and overall health-related outcomes using a mixed methods design.

Methods

Trial Design

A preceding patient-centered, iterative development phase informed the intervention and functioned as a preuse acceptability assessment of burden, usability, relevance, and coherence (E Kosman, MSc, et al, unpublished data, 2026). Building on this foundation, this open-label, 2-arm pilot randomized controlled trial (July 2024 to July 2025) used an expansion-type mixed methods design, collecting quantitative and qualitative data to enrich and explain emerging findings (see Multimedia Appendix 1 [6-8,16,33,41,45,47,64,65,92-127] for design rationale and prestudy procedures) [128]. Longitudinal self-reported data were collected in the intervention group (IG) and control group (CG) at baseline (T0) after randomization, peritreatment (T1; week 5 in the CG/after completing module 5 in the IG), posttreatment (T2; week 8 in the CG/after completing module 8 in the IG), and at 6-month follow-up (T3). The IG received Odeya app access; the CG was allowed to use existing treatment options within the health care system (treatment as usual) and offered later access. The trial was preregistered (German Clinical Trials Register: DRKS00034351) [129] and adhered to Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online TeleHealth checklist (Multimedia Appendix 2) [130], and National Institutes of Health best practice guidelines for mixed methods research in the health sciences [131].

Participants

Participants had to meet the following criteria to be eligible: (1) sufficient understanding of the German language, (2) being at least 18 years of age, (3) a physician-suspected or confirmed diagnosis of endometriosis or adenomyosis, (4) clinically relevant sexual distress (Female Sexual Distress Scale-Desire/Arousal/Orgasm [FSDS-DAO] >18), and (5) owning a smartphone (iOS or Android). The exclusion criteria were (1) current severe depression (Beck Depression Inventory-II ≥29) [92], (2) current severe anxiety (Generalized Anxiety Disorder 7-item ≥15) [93], (3) suicidal tendencies in the last 5 years, (4) current symptoms of posttraumatic stress disorder (PTSD), (5) substance dependence in the last 2 years, (6) current psychosis or dissociative symptoms, and (7) current pregnancy. If the online screening responses indicated the presence of potential PTSD or substance dependence, the participants were invited via email to take part in an additional telephone screening interview. The International Trauma Questionnaire [132] was used to assess PTSD, and the relevant section of the Structured Clinical Interview for the *DSM-IV* (*Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*) [133] was used to assess substance use.

Intervention

The Odeya intervention was a self-guided smartphone app intervention developed to address sexual distress in women with endometriosis, for both single and partnered users.

It was developed at the Institute of Sexology and Sexual Medicine, Charité–Universitätsmedizin Berlin, within the Berlin Institute of Health Digital Health Accelerator program, in collaboration with Hybrid Heroes GmbH, using a patient-centered, iterative process (E Kosman, MSc, et al, unpublished data, 2026). The intervention comprises 8 self-guided modules and a symptom-tracking tool, intended for completion over 12 weeks with a 4-week buffer. Modules unlock weekly, each taking ~60 minutes plus 15-60 minutes of exercises. Users can pause and resume at any time. See Table 1 and Figure 1 for a list of module topics and examples of user interface screens. Delivery was multimodal (text, video, audio, graphics, or interactive tasks) with exercises such as pelvic floor training, guided body-based activities, mindfulness, and sensate focus for solo or partnered practice. Participants in relationships were encouraged to involve their partners, while equivalent solo options were provided. The symptom-tracking tool monitors 4 domains (body, mind, social, and sexuality) using 1-10 ratings for pain, stress, self-care, and sexual satisfaction, plus standardized and personalized symptoms. See the Methods section in Multimedia Appendix 1 for further details on development, personas, technical notes, and symptom tracking.

XSL•FO
RenderX

Table 1. Overview of the Odeya intervention modules.

Module	Title	Purpose and key topics	Relational focus
1	You are not alone	Introduces sexual dysfunction and its interplay with endometriosis; fosters health literacy and support.	Normalizes relational strain within the biopsychosocial framework of sexual distress.
2	My body	Builds body awareness and sexual anatomy knowledge; covers stimulation techniques for solo and partnered activity.	Encourages exploration of bodily responses during solo and partnered touch.
3	My pain is real	Explains pain mechanisms and central sensitization; provides strategies for sexual pain management.	Promotes communication about pain, joint coping, and coregulation during intimacy.
4	My sexual self	Explores personal needs, preferences, boundaries, and the role of fantasies in sexual agency.	Encourages sharing preferences and boundaries in communication with partners.
5	My emotional network	Addresses dysfunctional beliefs using the Fear-Avoidance Model and cognitive restructuring.	Reflects on how emotional and cognitive patterns affect relationship dynamics and intimacy.
6	My sexual communication	Enhances communication skills and intimacy through couples' exercises (eg, sensate focus).	The full module has a relationship focus. Emphasizes sexuality as relational communication and includes guided communication strategies and sensate focus exercises for couples.
7	My sexual response	Explains the sexual response cycle in a partnered context, contextual influences, and effects of stress on arousal.	The full module has a relationship focus. Addresses dyadic factors influencing arousal, responsiveness, and shared satisfaction.
8	My resources	Summarizes learnings; develops personalized routines and booster strategies for sustained sexual health.	Encourages maintaining intimacy and open communication with partners.

Figure 1. Interface of the Odeya intervention. Five smartphone screens are shown from left to right: (A) symptom-tracking overview displaying symptom levels (upper section) and present symptoms for a single entry (lower section); (B) overview of intervention modules including modules 1 and 2; (C) example of module content with graphics and an audio exercise; (D) example of psychoeducational content presented as text; and (E) example of psychoeducational content presented as video.

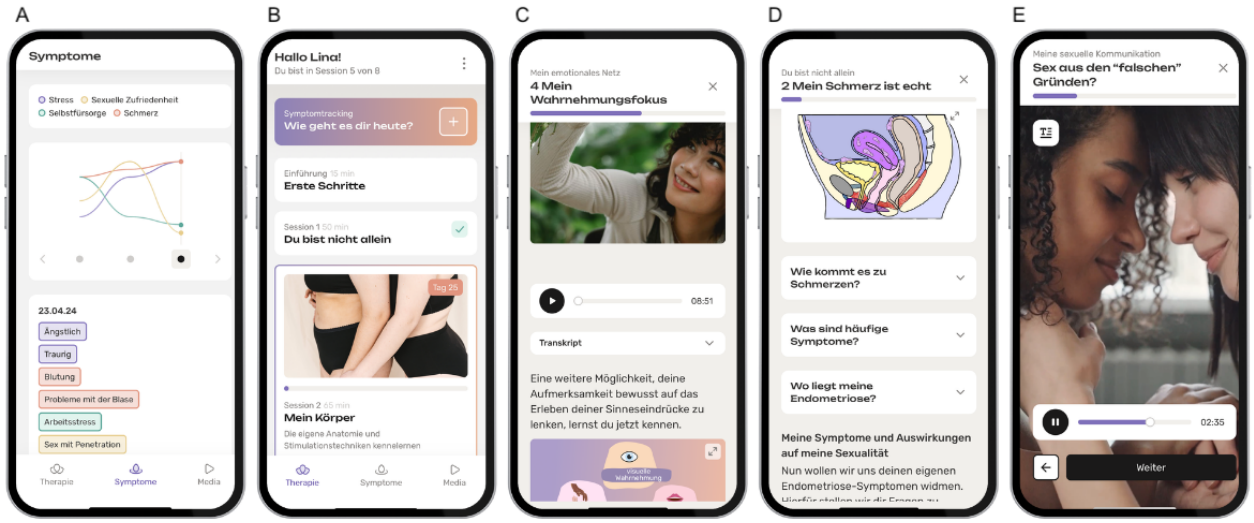

Procedure

Recruitment took place from May to October 2024. Participants were recruited online through study announcements posted by endometriosis associations on their websites, Instagram accounts, and in endometriosis-related Facebook groups in Germany, Austria, and Switzerland. In-person recruitment took place via the Department of Gynecology at Charité–Universitätsmedizin Berlin and through a flyer campaign in outpatient gynecological practices across Germany (see Methods section in Multimedia Appendix 1 for selection procedures). Interested individuals emailed the study team, provided medical documentation of suspected or confirmed endometriosis or adenomyosis, and received study information. A pseudonym and masked email address were created for each participant [134] using the tool AnonAddy to protect participant

XSL•FO
RenderX

identity. For screening and follow-up, they received a link to the data management platform REDCap (Research Electronic Data Capture; Vanderbilt University) [135,136] via email hosted at Charité–Universitätsmedizin Berlin. Participants were invited to a telephone interview with a clinical psychologist if further clarification was required. After randomization, the IG received access to the Odeya app. Qualitative interviews were offered to three groups: (1) IG dropouts, (2) IG completers, and (3) CG completers.

Randomization

Balanced block randomization with 4 blocks was generated in R software (version 4.5.1; R Foundation for Statistical Computing) and implemented in REDCap (version 15.5.29) by the research team to ensure a 1:1 allocation ratio between the IG and CG. Randomization was stratified by relationship status (single vs in a relationship) and age (<30 vs ≥30 years). Participants were informed of their allocation.

Quantitative Measures

This study assessed (1) adherence through module completion, app use duration, dropout status (IG: >4 weeks inactivity with module-locked assessments; CG: nonresponse to scheduled questionnaires), and symptom tracking activity; (2) acceptability, measured with the Client Satisfaction Questionnaire-Internet (CSQ-I), the German mHealth App Usability Questionnaire (G-MAUQ), and a single Visual Analog Scale for Client Satisfaction item on overall satisfaction; (3) safety, evaluated with the Inventory for the balanced assessment of Negative Effects of Psychotherapy-Online Intervention (INEP-ON) and self-reported changes in health status; and (4) sexual and health-related outcomes, including the FSDS-DAO, Female Sexual Function Index-German version (FSFI-d), Fear of Sexuality Questionnaire (FSQ), Vaginal Penetration Cognition Questionnaire (VPCQ), Central Sensitization Inventory-German version (CSI-GE), Partnership Questionnaire (PFB), and the Patient-Reported Outcome Measurement Information System-29-Item Profile (PROMIS-29). Detailed descriptions of assessment time points, selected measures, and additional questionnaires implemented in the broader study framework are provided in Multimedia Appendix 1 (Methods section and Table S1 in Multimedia Appendix 1).

Statistical Analysis

Descriptive statistics (means, SDs, and medians with IQR) were computed for all sexual and health-related continuous variables at baseline (T0), T1, T2, and 6-month follow-up (T3). Change scores (Δ=value$_t$ – value$_0$) were analyzed with linear models for continuous outcomes. Models included group, time, and their interaction, adjusted for the respective baseline value. Adjusted mean changes and 95% CIs were obtained via the emmeans package in R software (version 4.5.1) and visualized. Adjusted between-group differences (IG-CG) are presented as mean differences with 95% CIs; *P* values were not shown, given the exploratory design. Standardized effect sizes (adjusted Cohen d; small: ≥0.2; medium: ≥0.5; large: ≥0.8) [137] were computed for the comparison of change to baseline means. For descriptive reporting, pooled SDs of the individual change scores per time point were added. As all questionnaires used mandatory fields, no item-level missing data occurred. Missingness was limited to uncompleted measurement time points and handled by available-case analysis without imputation. A predefined target sample size of n=60 was based on feasibility considerations within the mixed methods pilot design.

Qualitative Interviews

Overview

The qualitative substudy included 16 participants sampled from IG dropouts (n=11; mean age 29.36, SD 3.91 years), IG completers (n=3; mean age 37.5, SD 13.44 years), and CG completers (n=2; mean 33.67, SD 5.69 years; see Table S2 in Multimedia Appendix 1 for participant characteristics). While all IG participants were invited, CG completers were recruited stepwise using criterion sampling to ensure sociodemographic diversity (age and relationship status) due to feasibility constraints. Given the small, uneven subgroup sizes, the qualitative data were used to capture a breadth of perspectives and to identify barriers and facilitators relevant to feasibility and implementation. Interviews and analysis proceeded iteratively; recruitment ended when no additional overarching feasibility-relevant issues were identified in the final interviews (further details in Methods section in Multimedia Appendix 1). Three distinct semistructured interview guides (Multimedia Appendix 3) were developed for each participant group to explore usability, engagement barriers, perceived helpfulness, and contextual factors.

Qualitative Data Analysis

IG dropout data were analyzed using a free-listing approach, a qualitative elicitation method that generates structured and quantifiable data by asking participants to spontaneously enumerate all responses relevant to a prompt. This method captures how individuals naturally conceptualize and prioritize health-related experiences, making it particularly useful for identifying perceived barriers, needs, and salient usability issues from participants' own language [64,94]. Lists were subsequently cleaned, consolidated, and organized into categories through iterative coding and consensus discussions [95]. Thematic salience was derived by examining the frequency, following established guidelines [96]. Interviews with IG and CG completers were transcribed using *f4transkript* [97] and analyzed with a qualitative content analysis following Schreier's toolbox model [98]. Coding was conducted inductively: an initial coding frame was developed from the data through pilot coding, iteratively refined as further transcripts were coded, and subsequently applied to the full dataset in MAXQDA (version 24.3; VERBI Software GmbH). Analytical rigor was supported through memo-writing and regular peer debriefing within the research team. Further details are provided in Multimedia Appendix 1 (Methods).

Triangulation of Qualitative and Quantitative Data

In line with the expansion-type mixed methods design, quantitative and qualitative data were analyzed separately by researchers with complementary methodological expertise. Integration occurred at the interpretation stage through triangulation, whereby findings from both data strands were systematically compared and synthesized in relation to study

XSL•FO
RenderX

end points through iterative discussion within the research team. Integrated findings are presented in a joint display table.

Ethical Considerations

The human participant study was approved by the Ethics Committee of Charité–Universitätsmedizin Berlin (EA4/217/23). Participants received no compensation and could contact the research team for technical support during the study period. Data were collected and managed in REDCap hosted at Charité. Written informed consent was obtained via postal mail.

Results

Characteristics of Study Population

Between July and November 2024, 187 individuals were screened. Of those, 132 completed the screening process, and 72 were excluded due to eligibility screening (Figure 2). After randomization, 29 women were assigned to the IG, and 31 were assigned to the CG. The mean age of the 60 included participants was 31.12 (SD 6.67) years, with ages ranging from 21 to 59 years (Table 2). Most participants (51/60, 85%) were in relationships and reported high relationship satisfaction (median 8.0, IQR 7.0-9.0). The overall level of education was high, with 81.7% (49/60) having 12 or more years of education. Only 1.7% (1/60) had previously accessed sex and couples therapy. However, half of the participants reported previous experience with psychotherapy. For information on motivation, participation expectations, and FSDS-DAO scores before and after randomization, see the Methods section in Multimedia Appendix 1.

Figure 2. Flowchart of participants. CG: control group; FSDS-DAO: Female Sexual Distress Scale-Desire/Arousal/Orgasm; IG: intervention group.

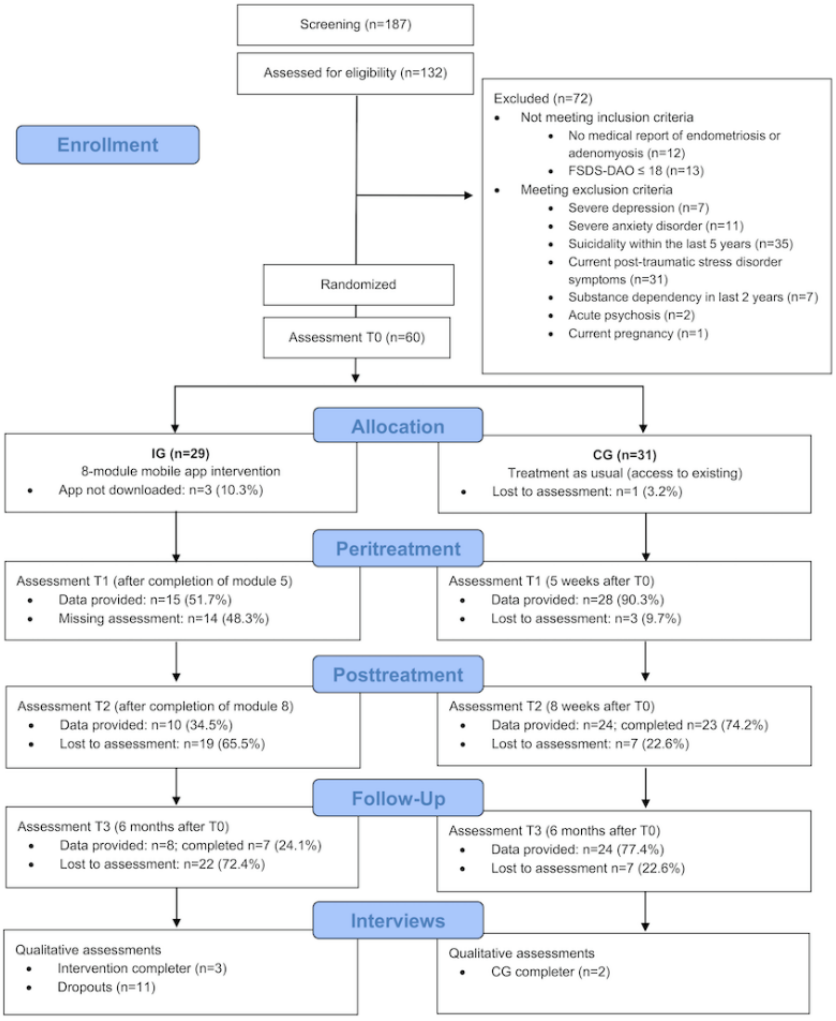

XSL•FO
RenderX

Table 2. Characteristics of the study population at baseline (T0) stratified by intervention group (IG) and control group (CG). Percentages refer to column totals.

Characteristic	Total (N=60)	IG (n=29)	CG (n=31)
Age (years), mean (SD)	31.12 (6.67)	31.1 (5.29)	31.13 (7.83)
In a relationship, n (%)	51 (85)	26 (89.7)	25 (80.6)
Relationship duration[a], mean (SD)	75.80 (67.64)	69.52 (56.86)	82.33 (78.01)
Relationship satisfaction (0-10)[b], median (IQR)	8.0 (7.0-9.0)	8.0 (7.0-9.0)	8.0 (7.0-9.0)
Education ≥12 years, n (%)	49 (81.7)	23 (79.3)	26 (83.9)
Urban residence, n (%)	33 (55)	16 (55.2)	17 (54.8)
Heterosexual, n (%)	51 (85)	23 (79.3)	28 (90.3)
Partnered intimacy[c,d], n (%)	47 (78.3)	24 (82.8)	23 (74.2)
Masturbation[c], n (%)	20 (33.3)	7 (24.1)	13 (41.9)
Religious, n (%)	23 (38.3)	10 (34.5)	13 (41.9)
Histologically confirmed endometriosis, n (%)	52 (86.7)	26 (89.7)	26 (83.9)
Previous operation, n (%)	53 (88.3)	26 (89.7)	27 (87.1)
Hormonal medication, n (%)			
Progestin-only contraceptive pill	17 (28.3)	10 (34.5)	7 (22.6)
Combined oral contraceptive pill	6 (10)	2 (6.9)	4 (12.9)
Sex therapy or couples therapy, n (%)	1 (1.7)	0 (0)	1 (3.2)
Psychotherapy, n (%)	31 (51.7)	17 (58.6)	14 (45.2)
Other diagnoses, n (%)			
PCOS[e]	0 (0)	0 (0)	0 (0)
PMS[f]	13 (21.7)	8 (27.6)	5 (16.1)
Uterus myomatosus	3 (5)	3 (10.3)	0 (0)
Uterus prolapse	1 (1.7)	0 (0)	1 (3.2)
Incontinence	3 (5)	1 (3.4)	2 (6.5)
Infertility	1 (1.7)	1 (3.4)	0 (0)
Vulvodynia	1 (1.7)	0 (0)	1 (3.2)
Lichen sclerosus	0 (0)	0 (0)	0 (0)
Cancer	2 (3.3)	2 (6.9)	0 (0)
Lifestyle, n (%)			
Physical activity	41 (68.3)	22 (75.9)	19 (61.3)
Healthy diet	53 (88.3)	25 (86.2)	28 (90.3)
Smoking	5 (8.3)	2 (6.9)	3 (9.7)
Alcohol consumption	8 (13.3)	3 (10.3)	5 (16.1)
BDI-II[g] (0-63), mean (SD)	13.3 (7.07)	13.45 (6.99)	13.16 (7.26)
GAD-7[h] (0-21), mean (SD)	5.63 (3.4)	6.55 (3.72)	4.77 (2.88)
SSP-F[i], n (%)	54 (90)	25 (86.2)	29 (93.5)
BSI GSI[j], mean (SD)	54.07 (11.88)	54.86 (12.74)	53.32 (11.18)
BSI PSDI[k], mean (SD)	53.53 (10.14)	54.52 (11.33)	52.61 (8.98)
BSI PST[l], mean (SD)	55.05 (12.73)	55.45 (12.96)	54.68 (12.72)
CTQ[m]: Sexual Abuse[n], n (%)	7 (11.7)	3 (10.3)	4 (12.9)

XSL•FO
RenderX

Characteristic	Total (N=60)	IG (n=29)	CG (n=31)
CTQ: Any Trauma[o], n (%)	22 (36.7)	9 (31)	13 (41.9)

[a]In months.
[b]Higher values indicate better outcomes.
[c]More than once per week.
[d]Partnered intimacy referred to activities such as cuddling and kissing.
[e]PCOS: polycystic ovary syndrome.
[f]PMS: premenstrual syndrome.
[g]BDI-II: Beck-Depression-Inventory-II.
[h]GAD-7: Generalized Anxiety Disorder Scale-7.
[i]SSP-F: Screening for Sexual Problems.
[j]BSI GSI: Brief Symptom Inventory Global Severity Index.
[k]BSI PSDI: Brief Symptom Inventory Positive Symptom Distress Index.
[l]BSI PST: Brief Symptom Inventory Positive Symptom Total.
[m]CTQ: Childhood Trauma Questionnaire.
[n]Sexual trauma was defined using the Childhood Trauma Questionnaire Sexual Abuse Subscale, with a cutoff score of 8 [138].
[o]Any trauma was defined as meeting the Childhood Trauma Questionnaire cutoff for at least moderate severity [138].

Main Findings

Overview

The following section reports adherence and user behavior, including module completion, dropout rates, time spent in the app, reasons for discontinuation, and use of the symptom tracker, as well as acceptance and user satisfaction assessed with validated questionnaires. Qualitative facilitators and barriers to adherence and acceptance follow in subsequent sections. Safety outcomes comprised balanced effects, self-reported health changes, and stressful life events.

Adherence and User Behavior

Dropout Rates, Time, and Reasons

Participants in the IG completed a median of 6 (IQR 2-8) modules (mean 4.9, SD 2.97; 4.9/8, 61.2% of the total content). All 8 modules were completed by 34.5% (10/29) of the IG, with a dropout rate of 65.5% (19/29), which is higher than in the CG (7/31, 22.6%). Mean app usage duration, defined as the time from first log-in until either completion of module 8 or the last recorded in-app activity (therapeutic content or symptom tracking), was 15 weeks (range 0-30) in the IG. Among completers, the mean duration was 18 (range 10-30) weeks, whereas noncompleters used the app for an average of 13 (range 0-30) weeks. IG dropouts were distributed throughout the course of the intervention: before starting the app (n=3), after module 1 (n=3), module 2 (n=2), module 3 (n=1), module 4 (n=4), module 5 (n=1), module 6 (n=4), and module 7 (n=1). Baseline characteristics (Table S3) and baseline values of outcomes (Table S4) of IG completers and IG dropouts are presented in Multimedia Appendix 1. Health care behaviors showed little change; only isolated therapy initiations occurred (Results section in Multimedia Appendix 1). Reasons for dropout included time constraints, technical difficulties, life changes, and perceived length of app units (Tables S5-S7 in Multimedia Appendix 1).

Adherence Facilitators and Barriers

High initial motivation and the persona-based progress stories supported adherence:

> *You also got to witness the progress of the three women, and that was motivating and encouraging.* [IG1]

Several participants felt acknowledged by the intervention, contrasting it with previous clinical encounters where their concerns had not been taken seriously.

Participants identified unexpectedly high time demands and emotional strain of prolonged self-reflection as key barriers:

> *It kept getting longer and more extensive.... I think it was supposed to be 50 minutes per module..., but I sometimes needed almost three times as long.... It was really exhausting to constantly engage with myself.* [IG1]

Some wished for more interaction with health care professionals or peers and reported difficulties applying insights to daily life and relationships.

> *When it comes to communication in the relationship that was also a very big topic in the app and I've always struggled with that. There were helpful impulses, but I couldn't really implement them at that point.* [IG2]

Technical issues—videos stopping without replay options, restrictive response formats, and static text fields limiting review of longer entries—further affected usability and adherence.

Use of Symptom Tracker: Frequency and Evaluation

Symptom tracking usage varied widely across participants. The average number of tracking entries during intervention was 18.34 (SD 23.77; range 0-99), with completers averaging 23.20 (SD 34.03; range 4-99) and dropouts 15.79 (SD 16.73; range 0-46). The most frequently self-selected tracked symptoms included, for emotional states: feeling relaxed, sad, or tired; for physical symptoms: bowel problems, intake of pain medication

XSL•FO
RenderX

or hormones, and bladder problems; for social aspects: participation in social activities, sports, or work-related stress; and for sexuality: no sexual activity, masturbation, or sexual intercourse. Figure 3 shows the mean tracked levels of sexual satisfaction, self-care, pain, and stress.

Qualitative feedback on tracking reminders was mixed; some participants perceived them as redundant or burdensome, while one reported them as helpful.

Figure 3. Rating of intensity levels of sexual satisfaction, self-care, pain, and stress in the IG, stratified for completers (n=10) and dropouts (n=15), showing median rating averaged across the usage weeks.

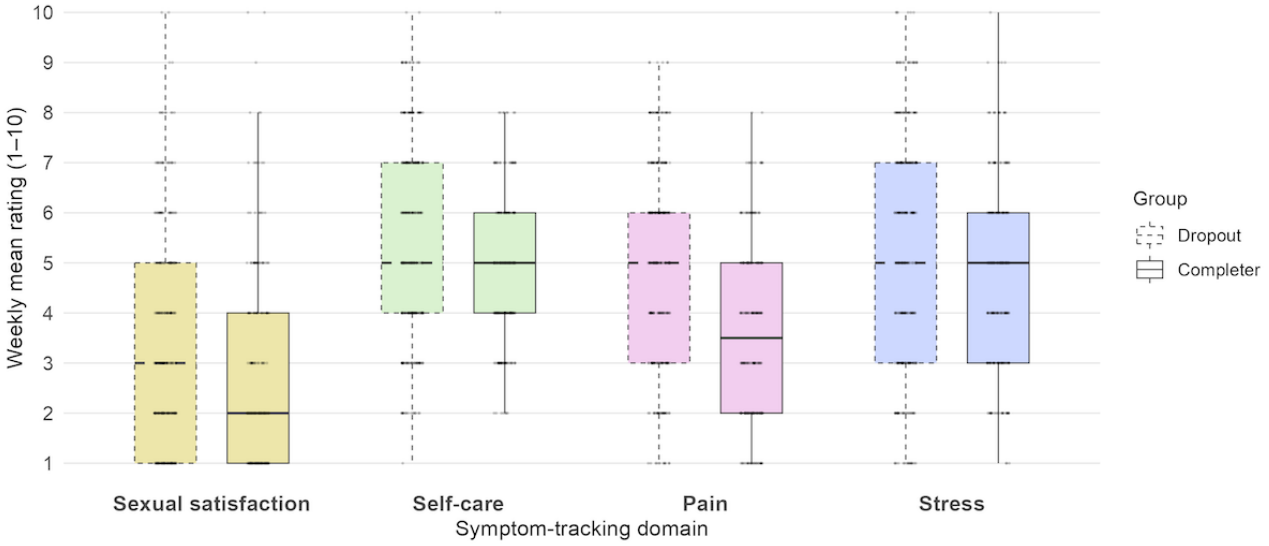

Acceptance

User Satisfaction

At midintervention, the median satisfaction score was 8.00 (IQR 7.0-8.0; mean 7.53, SD 1.36; n=15), and at postintervention (T2), 7.50 (IQR 6.0-8.0; mean 7.00, SD 1.83; n=10) on a 10-point scale (Visual Analog Scale for Client Satisfaction). At T2, CSQ-I ratings had a mean of 26.60 (SD 4.12) of 32. The CSQ-I sum of items assessing general satisfaction (items 1, 2, 3, 5, and 7) averaged 16.30 (SD 2.45), item 4 assessing recommendation 3.50 (SD 0.71), item 6 assessing helpfulness 3.10 (SD 0.88), and item 8 assessing likelihood of reuse 3.70 (SD 0.67). Based on these ratings, 80% (8/10) of participants agreed or strongly agreed with CSQ-I total items, 90% (9/10) reported that they would recommend or reuse the intervention, and 70% (7/10) reported that the intervention was helpful and fulfilled their general satisfaction. Usability, assessed with the G-MAUQ, showed a total score of 5.38 (SD 0.74), with subscales of ease of use (mean 6.46, SD 0.64), interface satisfaction (mean 5.28, SD 0.78), and usefulness (mean 4.46, SD 1.69). For detailed ratings, including median, minimum, and maximum values, see Table S8 in Multimedia Appendix 1.

These findings were closely mirrored in participant interviews, where usability, app design, and comprehensive content were highlighted as facilitating factors.

> *First of all, in terms of the technical aspects and structure, I found the app very clear and easy to use.* [IG1]

Qualitatively, some users expressed discomfort or ambivalence about specific exercises, such as guided masturbation or vulva self-exploration.

Facilitators and Barriers to Acceptance

Multimedia elements and the option to proceed at one's own pace were highly valued. Participants appreciated the intervention's diverse components, including normalization through personas ("not feeling alone"), psychoeducation, practical and communication exercises, structured reflection, partners' involvement, and gradual exposure to sensitive topics, especially the sensate focus exercises.

However, several barriers were identified. Technical issues and content overload inhibited satisfaction for some. These frustrations might explain lower satisfaction scores or disengagement in some cases. Weekly exercises were challenging, leading to self-doubt and frustration, although these feelings were partly mitigated by greater awareness of social influences, as one participant described:

> *Sometimes I felt very frustrated, doubting if it was my fault.... I even felt guilty after reading or doing certain exercises.... However, I was able to balance this by reflecting on what I'd learned from my parents' attitudes toward sexuality...phrases from friends or relationships. These insights gave me "aha" moments, helping me feel less desperate and understand that other factors might be involved.* [IG3]

Safety

Balanced Effects

The most frequently endorsed positive effects in the INEP-ON outcome were the helpfulness of new ways of thinking (T2: 10/10, 100%; T3: 6/7, 85.7%) and moderator support (T2: 9/10, 90%; T3: 4/6, 57.1%; Table S9 in Multimedia Appendix 1). Increased motivation for psychotherapy was reported by 60% (6/10) at T2 and 42.9% (3/7) at T3. Improved overall well-being

XSL•FO
RenderX

was noted by 90% (9/10) postintervention, but only 28.6% (2/7) at follow-up, while 57.1% (4/7) reported deterioration. Negative responses on balanced items were otherwise rare (≤10% at T2 and <30% at T3; see Results section in Multimedia Appendix 1). Items addressing exclusively negative effects (Table S10 in Multimedia Appendix 1) showed longer periods of not feeling well in 50% (5/10) at T2 and 71.4% (5/7) at T3 (see Results section in Multimedia Appendix 1 for details).

Self-Reported Health Changes and Stressful Life Events

A total of 6 women in the IG and 8 in the CG reported events, including new medical diagnoses (eg, cardiac arrhythmia, asthma, adenomyosis, suspected lipedema, and migraine), hospitalizations, bereavement, or relationship breakups. In the IG, only 1 affected participant dropped out; the others completed the study.

Secondary Outcomes

Overview

Secondary outcomes included exploratory analyses of changes in sexual health (FSDS-DAO, FSFI-d, FSQ, VPCQ, CSI-GE), relationship (PFB), and overall health (PROMIS-29). Table S11 in Multimedia Appendix 1 reports mean scores and Table S12 baseline-adjusted changes with effect sizes for all outcomes. Figure 4 illustrates sexual health outcomes based on total scores.

Figure 4. Sexual health outcomes over time as changes from baseline (Δ) for intervention group (IG) and control group (CG) across T0-T3. Points and lines depict estimated marginal means of the change scores; error bars indicate 95% CIs. Above the x-axis, the between-group effect size (d) from the linear model (Δ ~ Group × Time + Baseline) is shown for each time point. Shaded backgrounds denote study phases (on treatment: T0-T2; follow-up: T3). Panels A-D correspond to outcomes. CSI-GE: Central Sensitization Inventory-German version; FSDS-DAO: Female Sexual Distress Scale-Desire/Arousal/Orgasm; FSFI-d: Female Sexual Function Index-German version; PFB: Partnership Questionnaire; T0: baseline; T1: after module 5/5 weeks after baseline; T2: after module 8/8 weeks after baseline; T3: 6-month follow-up after baseline.

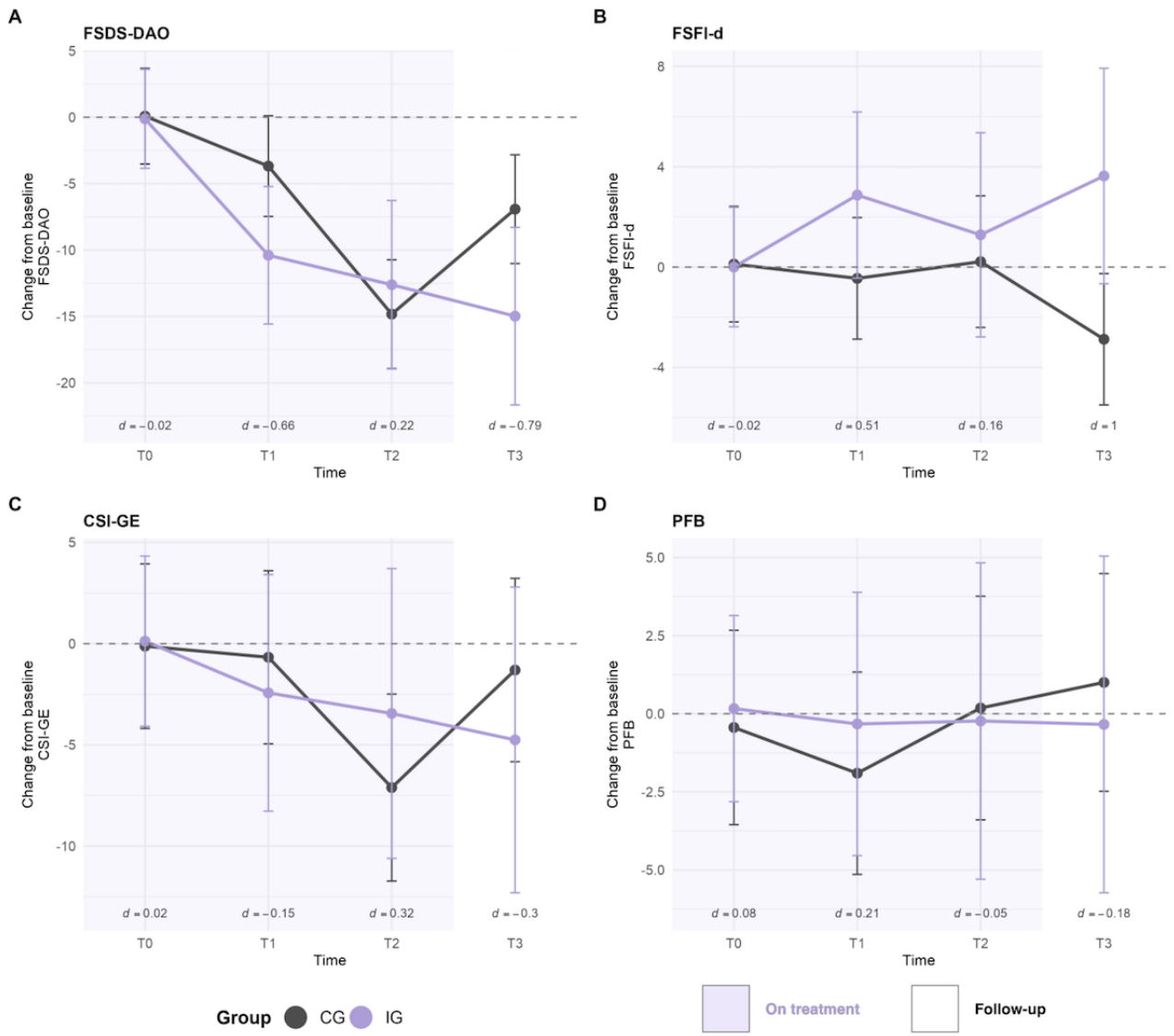

XSL•FO
RenderX

Sexual Health and Relationship-Related Outcomes

The IG showed a stronger early reduction in sexual distress (FSDS-DAO) (T1: IG Δ=–10.39 vs CG Δ=–3.68; between groups Δ=–6.71, 95% CI –13.13 to –0.29; *d*=–0.66) while both groups improved by T2 (Figure 4 and Table S12 in Multimedia Appendix 1). At T3 the IG again showed a stronger reduction (T3: Δ=–8.05, 95% CI –15.89 to –0.22; *d*=–0.79).

Sexual function (FSFI-d) improved continuously in the IG (T1-T3 Δ=1.29-3.63), with the largest between-group difference at T3 (Δ=6.51, 95% CI 1.48-11.55; *d*=1.00).

Fear of coitus (FSQ) showed small between-group effects, with the IG showing greater improvement at T1 (*d*=−0.21) and T3 (*d*=−0.14), but the CG at T2 (*d*=0.17). Fear of noncoital activity decreased continuously in the IG by about 0.6 points at T2 (scale ranging from 5-25), with small to large effects between groups (*d*=−0.14 to −0.84; Table S10 in Multimedia Appendix 1).

Regarding vaginal penetration cognition (VPCQ), results were mixed, with small to moderate improvements in the IG, primarily at T1 and T3 (eg, between-group effects: at T3, control cognitions d=0.48; catastrophic and pain cognitions d=−0.17). The only scale showing continuous improvement favoring the IG was incompatibility cognitions, with between-group effect sizes ranging from d=−0.34 to −0.89.

Central sensitization (CSI-GE) improved in both groups. Small effects favored the IG at T1 (*d*=−0.15) and T3 (*d*=−0.30), while at T2 a small effect favored the CG (*d*=0.32).

Partnership quality (PFB) demonstrated only small changes over time (T1: Δ=1.58; d=0.21; T2: Δ=−0.42; d=−0.05; T3: Δ=−1.34; d=−0.18). Subscale effects were likewise small: disruptive behavior worsened slightly in the IG at T3 (d=0.35), whereas communication improved in the IG at T1 (d=0.26) and T3 (d=0.21), with no change at T2 (d=−0.01).

Qualitative interviews indicated that IG participants frequently reported new positive sexual experiences:

> *Using the app helped me to rediscover how to have positive experiences with sexuality—across the entire spectrum of what sexuality can be.* [IG1]

Several reported that their perception and communication of sexual pain improved, although complete pain relief was rarely achieved:

> *From the communication perspective, I felt that I was often able to say, "Okay, I have pain now," and I could localize it, to give him that feedback.* [IG3]

Further, IG participants described gaining more courage and reduced fear and avoidance of sexual activities, reporting increased confidence in approaching both coital and noncoital intimacy:

> *To find the courage to start again and actually do something. And also, those small steps—that was exactly what I needed. I can imagine many others feel the same, because there's always that fear of penetrative sex.* [IG1]

Qualitative analyses revealed that participants in the IG experienced increased openness and more frequent communication with their partners regarding sexuality and recognized the issue as a shared concern affecting both partners, rather than an individual problem. One participant noted:

> *In the beginning, when I started using the app, I really tried to communicate more with my partner.... It's usually such an uncomfortable topic for me—I've never talked about it openly. But I tried to tell him about the app, showed him parts of it.... I realized, it's not just my issue, it's something that affects the relationship too, and we're both suffering in some way.* [IG3]

Additionally, several participants described a shift toward a more constructive and collaborative dynamic, moving away from blame:

> *We talked about that very openly—that it's not about anyone being at fault or him hurting me or anything like that, but rather that we go through the process together and figure out, together, what feels good and what doesn't.* [IG3]

Although CG participants had no access, qualitative interviews suggested that study participation itself fostered self-reflection prompted by study questionnaires.

> *I actually liked that it forced me to keep reflecting...on what's going on in that part of my life, my sexuality or lack of libido. [It was really] helpful].* [CG1]

CG participants also reported positively impacting life events over the course of the study (eg, hormonal changes and changes in sexual life circumstances) and expressed continued interest in the intervention after the study.

Health-Related Quality of Life

Results for the PROMIS-29 domains were mixed overall. Anxiety, depression, fatigue, and physical function improved in both groups at T2, although between-group comparisons favored the CG at several time points. In contrast, social participation, pain interference, and pain intensity improved across all time points in the IG, with between-group effects favoring the IG at T2 and T3 and showing moderate to large effect sizes.

Qualitatively, participants in the IG attributed enhanced emotional regulation and reduced distress to the intervention, linking these psychological changes to better physical health management:

> *That was one of my goals—to somehow get out of this negative emotional state. And it actually worked. Of course, there are still phases or moments when you think, "Hmm..." But then things come to mind—things you can do, right? And you realize, okay, even if it's not working right now, that's okay too—it's not the end of the world.* [IG1]

Self-reflection and body awareness were recurring qualitative themes reported by both groups, further elucidating why both showed improvement in those domains.

XSL•FO
RenderX

Integration of Quantitative and Qualitative Findings

The integration of quantitative outcomes and qualitative interviews (Table 3) provides a nuanced picture of the intervention's effects on sexual health, relationships, well-being, and user engagement. Quantitative findings contextualized quantitative improvements and highlighted adherence challenges and the complexity of symptom management.

XSL•FO
RenderX

Table 3. Synthesis: integration of quantitative and qualitative results. Intervention group (IG) > control group (CG) indicates between-group differences in favor of the intervention group, whereas IG < CG indicates differences in favor of the control group.

Domain	Quantitative findings	Qualitative themes	Synthesis
Adherence	• IG completion 34.5% (10/29), dropout 65.5% (19/29) vs CG 22.6% (7/31) • Symptom tracking more frequent in completers • Frequent use of individual symptom feature • Outside app: start of psychotherapy, no change in costs	• Facilitator: persona stories • Barriers: emotional strain, time demands • Reminders: helpful or confusing • Dropout: time, technical issues, life changes • Need: more professional/patient interaction	High dropout aligns with time/emotional strain; progress stories may support adherence. Mixed feedback on tracking suggests the need for customizable reminders. Behavior outside app indicates broader health-seeking changes.
Acceptance	• CSQ-I[a] mean 26.6 (SD 4.1) of 32 • G-MAUQ[b] mean 5.4 (SD 0.7) of 7; "Ease of use" mean 6.5 (SD 0.6) of 7 • Mean price €83 (range €0-€300; €1=US $1.17); 30% (3/10) no self-pay	• Facilitators: clarity, multimedia, flexibility, reimbursement • Barriers: technical issues, content overload	High satisfaction matches positive reports; barriers may explain disengagement. Payment variability underscores the need for reimbursement.
Safety	• INEP-ON[c]: learned strategies, moderator support • "Negative events" rare (≤10% in T2, <30% in T3); mainly periods of not feeling well	• Reports of health changes, stressful life events in both groups	Findings suggest overall safety; negative responses linked more to external stressors than intervention.
Sexual health	• FSDS-DAO[d]: improved in both groups, IG > CG at T1, T3 • FSFI-d[e]: IG > CG • FSQ[f] "coital": improved in both, IG vs CG mixed; "noncoital": IG > CG • VPCQ[g] "control," "catastrophic": improved in both groups, IG vs CG mixed; "Incompatibility": IG > CG • PFB[h]: no change at T2; IG vs CG mixed; "Communication": IG > CG at T1, T3 • CSI-GE[i]: improved in both groups, IG > CG at T1; T3, IG < CG at T2	• Facilitators: rediscovery of positive experiences, improved pain communication, openness, collaboration • Barriers: ambivalence toward some exercises • CG: self-reflection via questionnaires and self-applied tools/books	Quantitative gains align with reports of reduced distress and better sex-related communication. Although some changes were not strongly reflected in quantitative scores, participants described meaningful qualitative improvements. CG improvements may be attributed to study participation and self-reflection.
Overall Health	• PROMIS-29[j]: "anxiety," "depression," "fatigue," "physical function" improved at T2 in both groups, mostly IG < CG; "Social participation": IG > CG at T3; "Pain interference", "pain intensity": IG > CG at T2, T3.	• IG: improved emotional regulation, distress reduction • Both: more self-reflection, body awareness	Both groups, especially the CG, showed broad improvements in psychological well-being and physical functioning, while the IG exhibited reductions in pain interference and intensity. Qualitative reports of greater emotional regulation, open partner communication, and body awareness in the IG support these findings, indicating that self-regulatory and relational processes may reduce pain-related avoidance and attentional capture by pain.

[a]CSQ-I: Client Satisfaction Questionnaire-Internet Version.

[b]G-MAUQ: German mHealth App Usability Questionnaire.

[c]INEP-ON: Inventory for the balanced assessment of Negative Effects of Psychotherapy.

[d]FSDS-DAO: Female Sexual Distress Scale-Desire/Arousal/Orgasm.

[e]FSFI-d: Female Sexual Function Index-German version.

[f]FSQ: Fear of Sexuality Questionnaire.

[g]VPCQ: Vaginal Penetration Cognition Questionnaire.

[h]PFB: Partnership Questionnaire.

[i]CSI-GE: Central Sensitization Inventory-German Version.

[j]PROMIS-29: Patient-Reported Outcome Measurement Information System-29-Item Profile.

XSL•FO
RenderX

Discussion

Principal Findings

This pilot implementation study examined the adherence, acceptability, safety, and exploratory effects of the self-guided Odeya app for women with sexual distress and diagnosed or suspected endometriosis. Adherence was moderate, with most completers working through several modules over an average duration of 18 weeks, while dropout was high. Overall satisfaction with the app was strong, and no negative life and health changes were attributed to the intervention. Quantitative outcomes showed reductions in sexual distress in both groups, with some advantages for the IG in sexual function, penetration-related fears, and partner communication. Broader health and mental health outcomes showed mixed but generally positive changes across groups. Qualitative feedback supported these trends, describing more positive sexual experiences, improved partner communication, greater pain awareness, and reduced fear of penetration.

Adherence

Overview

The intervention dropout rates in our study were notably high, a phenomenon typically reported in digital interventions for chronic conditions. A meta-analysis found a pooled dropout rate of 43% across app-based interventions for chronic diseases [76]. In the HelloBetter Vaginismus DiGA effectiveness trial, Zarski et al [50] reported dropout of 22% postintervention among women with dyspareunia, with participants completing on average 6 of 8 modules. In an early study, the Endo App retained 64.4% (29/45) of endometriosis patients in week 12 [61]. The German DiGA registry (BfArM) reports low attrition in IGs for several DiGAs, including the Endo App (3.75% dropout; NCT04883073) [60], Kranus Edera for erectile dysfunction (4.1%) [59], and Kranus Lutera for incontinence (5.4%) [62]. In comparison with these trials, dropout in our study was higher in the IG than in the CG [50,59,61,62]. Several factors may explain the comparatively high dropout in our sample.

Health Status

First, our study population comprised individuals with endometriosis and sexual distress, a particularly complex clinical profile that may require more intensive support than fully self-guided interventions can provide [69]. Unlike trials with stricter exclusion criteria (eg, excluding chronic pain or moderate depression [49,50,58], recent medication or surgery changes [60], chronic infections [62], or postsurgical/cardiovascular risk) [59], our study was deliberately inclusive, excluding only severe depressive or anxiety symptoms. This approach enhanced external validity but may have increased attrition among participants facing multiple health burdens. At baseline, dropouts reported lower quality of life and greater symptom burden (anxiety, depression, stress, fatigue, and pain) yet less sexual distress and better sexual function than completers. Visual inspection of the symptom-tracking data suggested slightly higher median values among dropouts for sexual satisfaction and pain compared to IG completers. This health status pattern may also explain the continuous dropout observed in our study, as disengagement likely reflects participants' health burden rather than low initial motivation or lack of interest in the intervention—contrasting with the early-stage dropout typically reported in digital mental health apps [139,140]. Some studies found no link between depression or pain and attrition [79,81]. However, a systematic review and a meta-analysis identified baseline depression and comorbid anxiety as predictors of dropout [77,80]. Further, most participants were in relationships. Being partnered has been identified as a risk factor for sexual dysfunction with distress [12], whereas being single has already been associated with higher dropout [81].

App Engagement

Second, app engagement and adherence may operate in both directions. While active tracking could facilitate greater adherence, it is also plausible that individuals with better health or mental well-being were more likely to engage with the app. Dropouts tracked symptoms less frequently, indicating lower initial engagement with self-monitoring features. Early engagement predicts adherence in digital health programs [78,81,82] and may mark a critical window for retention strategies. Evidence on the link between tracking and adherence in digital health interventions is mixed: some studies report benefits of consistent symptom tracking, while others show only transient effects [68,86,89]. In our study, reduced tracking and higher dropout among participants with elevated anxiety and depression suggest that mental health symptoms hinder sustained engagement. This aligns with meta-analytic evidence showing that engagement modestly predicts improved outcomes, with specific indicators such as module completion being most predictive [83].

Program Design

Third, compensation structures and intervention duration may have influenced adherence. Studies with better adherence rates often provided participant compensation or had shorter intervention periods [84], factors that can significantly impact engagement and completion rates. Although a 12-week duration is clinically appropriate and common [50,58,59,61,62], it may be demanding for participants with chronic pain and its psychological burden, as noted in prior studies [141]; furthermore, modules were reported to exceed the displayed module duration of 60 minutes.

Acceptance and User Experience

Overview

Quantitative results showed high user satisfaction, aligning with findings in women with genito-pelvic pain/penetration disorder (CSQ-I mean 28.03, SD 3.96; 67/72, 93% satisfaction) [50] and supported by qualitative reports of positive experiences. Similar satisfaction levels have been reported in other digital interventions across different patient populations using the CSQ-I [65,72]. Usability (G-MAUQ) ratings were particularly high for ease of use, likely reflecting the patient-centered development process (E Kosman, MSc, et al, unpublished data, 2026) and the app's reduced functions, with scores comparable to other digital interventions [66,85]. Perceived usefulness was

XSL•FO
RenderX

slightly lower in our study, which may have contributed to attrition.

Content Overload and Usability as Barriers

Content overload, reflected in the high time demands of the modules and the emotional strain of sustained self-reflection [142], highlighted the challenge of balancing comprehensive content with practical feasibility for participants. Similar findings show that time-intensive digital interventions hinder engagement in this population [74]. Moreover, even seemingly minor usability issues can significantly affect user experience and sustained engagement [73,75,87].

Guidance

Interview participants frequently requested more clinician and peer contact—consistent with evidence that human support drives engagement [88]—and reported difficulties applying content and feeling insufficiently guided. This underscores the potential value of integrating human support into digital tools for complex chronic conditions. Interventions including personal guidance achieved better adherence [56,76] compared with our fully self-directed approach. Blended care models may enhance both engagement and acceptance by providing structured support alongside digital self-management [143].

Willingness to Pay

Despite high user satisfaction, willingness to pay for a comparable DiGA was limited. Participants expressed generally low willingness to pay, substantially lower than prices of comparable DiGAs in Germany [57,63], and many felt that such tools should be financed by the health care system rather than by users themselves. Similar patterns have been reported, linking reluctance to pay to expectations of public coverage, skepticism toward digital tools, and limited awareness of benefits [144].

Safety

None of the health and life changes was attributable to the intervention. Comparisons with other digital interventions are limited, as adverse events are rarely assessed or reported in German DiGAs [71]. Our findings accord with a recent meta-analysis indicating that mental health apps rarely cause harm and do not increase adverse events versus controls [70].

Changes in Sexual and Overall Health

Although not powered for clinical end points, the trial provided exploratory indications of improvements, with small to large effects based on between-group mean differences. Our findings are consistent with evidence that both face-to-face and online interventions can reduce sexual distress and improve sexual function and satisfaction [16,41,44], that psychoeducation increases knowledge and reduces performance anxiety [31], and that mindfulness improves desire, arousal, and satisfaction while mitigating sexual distress [45,46]. An online self-help trial for women with dyspareunia reported medium to large improvements in genital pain and penetration-related cognitions and small to medium improvements in sexual function, anxiety, and well-being [50]. A psychoeducational web-based program for cancer survivors improved sexual communication but showed limited impact on relationship outcomes [67]. Consistent with this pattern [50,51,53,54], our trial also did not affect relationship satisfaction. Qualitative data aligned with established mechanisms [16,55]: participants described sensate focus and self-stimulation exercises as helpful for improving desire, and cognitive restructuring for reducing maladaptive sexual beliefs. These processes were reflected quantitatively by reduced incompatibility cognitions in the IG. However, it should be noted that some between-group differences may reflect random error rather than genuine effects.

Consistent with prior research [50], control participants also improved in health (eg, depression, anxiety, and fatigue) and sexual distress. However, at follow-up, reductions in sexual distress persisted strongly only in the IG. Qualitative data suggest that repeated self-assessment may have acted as a catalyst for self-reflection and proactive coping, a Hawthorne-like effect [145]. Additional mechanisms include mere-measurement effects [146], hope and expectancy [147], natural recovery [50,148], and enhanced self-monitoring [149]. Despite these effects, CG participants continued to desire intervention access, indicating ongoing unmet needs.

Implications for Future Digital Interventions

Despite the high dropout rate, satisfaction metrics in the IG indicated that engaged participants found the program helpful. Thus, content appears valuable, whereas delivery and support structures may require optimization. Following qualitative feedback indicating a desire for more guidance and professional contact, we propose key design considerations for future digital interventions (Textbox 1).

XSL•FO
RenderX

Textbox 1. Key design considerations to enhance engagement, adherence, and user experience in future digital sexual health interventions targeting individuals with endometriosis

1. Hybrid approaches: incorporating human support elements—and, where appropriate, AI-driven assistance—such as periodic check-ins with health care providers, trained coaches, or moderated peer communities, may enhance engagement, adherence, and outcomes [150-152].
2. Tailored content delivery: reducing the time burden and emotional intensity of modules while maintaining therapeutic effectiveness could improve completion rates. Flexible options—such as individually selectable modules, exercises, or notifications and supplementary materials (eg, workbooks)—may better address diverse user needs [153,154].
3. Technical optimization: addressing usability issues and ensuring a seamless user experience—potentially through the integration of gamification features—is fundamental to sustaining user engagement [155].
4. Screening for intervention readiness: given the relationship between baseline psychological distress and dropout, screening for intervention readiness and providing additional support for participants at high risk of dropout may improve outcomes [156].
5. Patient involvement: beyond developing the intervention through a patient-centered approach, we recommend involving patients or patient influencers in the go-to-market phase to ensure effective reach and trust [64,157-160].

Integrating the Relational Dimension

Beyond the structural and delivery-related design considerations outlined above, the relational dimension constitutes a further critical aspect of future digital sexual health interventions. Our findings address a recently identified structural limitation in current DiGAs for sexual dysfunctions: the insufficient consideration of the relationship dimension. While existing DiGAs primarily address individual functional parameters and measure success through standardized scores such as the International Index of Erectile Function - 5, sexual dysfunctions are frequently embedded in relational components that require dyadic intervention approaches [91].

Future Research

Future research should test efficacy for core sexual health outcomes, particularly sexual distress, in adequately powered randomized controlled trials and, in larger samples, identify subgroups most likely to benefit. Incorporating participants' recommendations—such as integrating the program into blended-care models with chat support or video consultations [12]—may enhance value and warrant evaluation. Longer, more flexible trials should also examine motivational drivers and dropout trajectories, as a 4-week inactivity threshold may be overly restrictive (eg, for women recovering from endometriosis surgery). Moreover, future iterations could include both optional and mandatory partner components addressing partners directly (eg, psychoeducation modules on endometriosis and its relational impact).

Limitations

This study has several limitations. First, several methodological aspects limit the interpretability of findings. The trial was not powered to detect efficacy or between-group differences, constraining comparative analyses. Additionally, although aspects of preuse acceptability were addressed during development, they were not systematically assessed in this study. Without a dedicated prepilot acceptability study, some usability issues (eg, module length and navigation) may not have been identified beforehand. Finally, combining assessments of acceptability and preliminary effectiveness within a single early-phase trial may have constrained sampling strategies and methodological specificity. Second, dropout rates were high—particularly in the IG—resulting in substantially reduced sample sizes for both quantitative and qualitative components. This further restricts within-group analyses and impedes meaningful comparisons between the IG and CG. Participants who discontinued showed higher baseline symptom severity (eg, anxiety and depression), potentially biasing findings toward more favorable outcomes. Third, satisfaction outcomes may be influenced by recruitment and measurement constraints. Because the recruitment strategy relied primarily on online channels, the sample may overrepresent women with higher digital literacy or a stronger preference for online support formats. Satisfaction was also only assessed among women who completed the intervention due to predefined measurement timing, likely inflating acceptability ratings. Fourth, sample characteristics and condition-specific focus limit generalizability. The sample covered a broad age range (up to 59 years), representing heterogeneous life stages (eg, perimenopause and menopause) with potentially distinct biopsychosocial profiles. Participants were also highly educated and predominantly partnered, which is not representative of the wider population. Moreover, although the intervention was developed through a patient-centered approach with women affected by different gynecological conditions (E Kosman, MSc, et al, unpublished data, 2026), it was only tested in women with endometriosis or adenomyosis, further limiting generalizability. Fifth, while the intervention included optional partner-focused components, systematic involvement of partners was limited, and dyadic processes were not comprehensively addressed. This represents a conceptual limitation, given that sexual health in the context of chronic conditions such as endometriosis is inherently relational. Finally, a coding error prevented proper attribution of INEP-ON items, obscuring whether reported positive or negative changes were related to life circumstances or the intervention itself.

Conclusions

The digital intervention Odeya appeared acceptable and safe for women who engaged with the app, with initial indications of improvements in sexual health–related outcomes among completers. Sustained engagement in fully self-guided formats, however, appears to vary by baseline psychosocial status, with individuals reporting better psychosocial health showing higher adherence. Adequately powered trials should establish efficacy, identify moderators of benefit, and determine the support intensity needed to maximize impact. Self-guided digital interventions may present one accessible and scalable

XSL•FO
RenderX

component within stepped-care models for sexual health, although some users are likely to require additional guidance and personalization.

Acknowledgments

We sincerely thank all patients who participated in the study and the qualitative development process. We also acknowledge the support of the Digital Health Accelerator Program of the Berlin Institute of Health at Charité–Universitätsmedizin Berlin, which enabled the collaboration with Hybrid Heroes for the development of the app-based intervention.

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: Literature search and systematization; Code generation; Code optimization;Proofreading and editing; Summarizing text; Adapting and adjusting emotional tone; Translation; Reformatting; Quality assessment; Identification of limitations; Recommendations; and Publication support.

The GenAI tools used were ChatGPT (GPT-5.1), Claude (Claude Sonnet 4.5), and Perplexity AI. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was supported by the Digital Health Accelerator Program of the Berlin Institute of Health, which also enabled the development of the app-based intervention in collaboration with Hybrid Heroes. The funder had no role in the study design, data collection, analysis, interpretation, or the decision to submit the manuscript for publication.

Data Availability

The datasets analyzed during this study are not publicly available due to restrictions imposed by the ethics approval.

Authors' Contributions

Conceptualization: LH, SMK, EK
Data curation: SMK, TT-D, LWS, EM
Formal analysis: SMK, EK, TT-D
Funding acquisition: LH, SMK
Methodology: LH, SMK, EK
Project administration: SMK, LH
Resources: KMB, LH
Supervision: LH, KMB, MMK, TP
Validation: LH
Visualization: SMK, EK
Writing – original draft: SMK, EK, LH, LW-S
Writing – review & editing: SMK, EK, LH, KMB, MMK, TP, EM, LW-S

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary methods, results, and tables.
[PDF File (Adobe PDF File), 850 KB - jmir_v28i1e86042_app1.pdf]

Multimedia Appendix 2
CONSORT - eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1055 KB - jmir_v28i1e86042_app2.pdf]

Multimedia Appendix 3
Interview protocols.
[PDF File (Adobe PDF File), 176 KB - jmir_v28i1e86042_app3.pdf]

References

1. Beier KM, Loewit K. Praxisleitfaden Sexualmedizin. Heidelberg: Springer Berlin; 2011.

XSL•FO
RenderX

2. Lee DM, Vanhoutte B, Nazroo J, Pendleton N. Sexual health and positive subjective well-being in partnered older men and women. J Gerontol B Psychol Sci Soc Sci 2016;71(4):698-710 [FREE Full text] [doi: 10.1093/geronb/gbw018] [Medline: 26993519]
3. Mitchell KR, Lewis R, O'Sullivan LF, Fortenberry JD. What is sexual wellbeing and why does it matter for public health? Lancet Public Health 2021;6(8):e608-e613 [FREE Full text] [doi: 10.1016/S2468-2667(21)00099-2] [Medline: 34166629]
4. International Classification of Diseases (ICD-11). Geneva: World Health Organization; 2019.
5. Briken P, Matthiesen S, Pietras L, Wiessner C, Klein V, Reed GM, et al. Estimating the prevalence of sexual dysfunction using the new ICD-11 guidelines. Dtsch Arztebl Int 2020;117(39):653-658 [FREE Full text] [doi: 10.3238/arztebl.2020.0653] [Medline: 33357346]
6. McCabe MP, Sharlip ID, Lewis R, Atalla E, Balon R, Fisher AD, et al. Risk factors for sexual dysfunction among women and men: a consensus statement from the fourth international consultation on sexual medicine 2015. J Sex Med 2016;13(2):153-167. [doi: 10.1016/j.jsxm.2015.12.015] [Medline: 26953830]
7. McCool-Myers M, Theurich M, Zuelke A, Knuettel H, Apfelbacher C. Predictors of female sexual dysfunction: a systematic review and qualitative analysis through gender inequality paradigms. BMC Womens Health 2018;18(1):108 [FREE Full text] [doi: 10.1186/s12905-018-0602-4] [Medline: 29929499]
8. Mitchell KR, Mercer CH, Ploubidis GB, Jones KG, Datta J, Field N, et al. Sexual function in Britain: findings from the third national survey of sexual attitudes and lifestyles (Natsal-3). Lancet 2013;382(9907):1817-1829. [doi: 10.1016/s0140-6736(13)62366-1]
9. Grosse-Rueschkamp MS, Kronthaler SM, Legeland LT, Schulze D, Beier KM, Hatzler L. What knowledge matters for women's sexual wellbeing? Validating a questionnaire measuring sexual knowledge using public and patient involvement. J Sex Med 2025;22(11):2104-2113. [doi: 10.1093/jsxmed/qdaf209] [Medline: 40974279]
10. Perelman MA. The sexual tipping point: a mind/body model for sexual medicine. J Sex Med 2009;6(3):629-632. [doi: 10.1111/j.1743-6109.2008.01177.x] [Medline: 19210711]
11. González-Mesa E, Moya-Bejarano D, Butrón-Hinojo CA, Marín-Sánchez P, Blasco-Alonso M, Jimenez-López JS, et al. Correlates of sexual function in a sample of Spanish women with endometriosis. J Clin Med 2021;10(21):4957 [FREE Full text] [doi: 10.3390/jcm10214957] [Medline: 34768476]
12. Kronthaler SM, Tissen-Diabaté T, Karsten MM, Blohmer J, Beier KM, Hatzler L. Assessment of mental and chronic health conditions as determinants of health care needs and digital innovations for women with sexual dysfunction: cross-sectional population-based survey study in Germany. J Particip Med 2025;17:e71301 [FREE Full text] [doi: 10.2196/71301] [Medline: 40865091]
13. Brenk-Franz K, Häuser W, Brähler E, Schneider N, Hoy M, Strauß B. Sexual dysfunctions and health-related impairment in patients with chronic disease. Dtsch Arztebl Int 2024;121(2):66-67 [FREE Full text] [doi: 10.3238/arztebl.m2023.0218] [Medline: 38427942]
14. Sobel T, David P. Impact of chronic medical disease on sexual function and other conditions. Obstet Gynecol Clin North Am 2024;51(2):323-340. [doi: 10.1016/j.ogc.2024.02.006] [Medline: 38777487]
15. Valeiro C, Matos C, Scholl J, van Hunsel F. Drug-induced sexual dysfunction: an analysis of reports to a national pharmacovigilance database. Drug Saf 2022;45(6):639-650. [doi: 10.1007/s40264-022-01174-3] [Medline: 35386045]
16. Zarski AC, Velten J, Knauer J, Berking M, Ebert DD. Internet- and mobile-based psychological interventions for sexual dysfunctions: a systematic review and meta-analysis. NPJ Digit Med 2022;5(1):139 [FREE Full text] [doi: 10.1038/s41746-022-00670-1] [Medline: 36085306]
17. Atlantis E, Sullivan T. Bidirectional association between depression and sexual dysfunction: a systematic review and meta-analysis. J Sex Med 2012;9(6):1497-1507. [doi: 10.1111/j.1743-6109.2012.02709.x] [Medline: 22462756]
18. Xie Y, Shi L, Xiong X, Wu E, Veasley C, Dade C. Economic burden and quality of life of vulvodynia in the United States. Curr Med Res Opin 2012;28(4):601-608. [doi: 10.1185/03007995.2012.666963] [Medline: 22356119]
19. Beier KML, Kurt K. Sexual Medicine in Clinical Practice. New York, NY: Springer; 2013.
20. Göhring J, Drewes M, Kalder M, Kostev K. Germany endometriosis pattern changes; prevalence and therapy over 2010 and 2019 years: a retrospective cross-sectional study. Int J Fertil Steril 2022;16(2):85-89 [FREE Full text] [doi: 10.22074/IJFS.2021.528397.1113] [Medline: 35639651]
21. Buggio L, Barbara G, Facchin F, Frattaruolo MP, Aimi G, Berlanda N. Self-management and psychological-sexological interventions in patients with endometriosis: strategies, outcomes, and integration into clinical care. Int J Womens Health 2017;9:281-293. [doi: 10.2147/ijwh.s119724]
22. Zondervan KT, Becker CM, Missmer SA. Endometriosis. N Engl J Med 2020;382(13):1244-1256. [doi: 10.1056/NEJMra1810764] [Medline: 32212520]
23. Fairbanks F, Abdo CH, Baracat EC, Podgaec S. Endometriosis doubles the risk of sexual dysfunction: a cross-sectional study in a large amount of patients. Gynecol Endocrinol 2017;33(7):544-547. [doi: 10.1080/09513590.2017.1302421] [Medline: 28347195]
24. Edgley K, Saunders PTK, Whitaker LHR, Horne AW, Tsanas A. Insights into endometriosis symptom trajectories and assessment of surgical intervention outcomes using longitudinal actigraphy. NPJ Digit Med 2025;8(1):236 [FREE Full text] [doi: 10.1038/s41746-025-01629-8] [Medline: 40316659]

XSL•FO
RenderX

25. Velten J, Pantazidis P, Benecke A, Bräscher A, Fehm L, Fladung A, et al. Wie häufig werden diagnosen aus dem bereich der sexuellen Funktionsstörungen an deutschen hochschulambulanzen für psychotherapie an psychologischen Instituten vergeben? Z Sex Forsch 2021;34(01):5-14. [doi: 10.1055/a-1362-2243]
26. Barnhoorn PC, Prins IC, Zuurveen HR, Oudsten BLD, Ouden MEMD, Numans ME, et al. Let's talk about sex: exploring factors influencing the discussion of sexual health among chronically Ill patients in general practice. BMC Prim Care 2022;23(1):49 [FREE Full text] [doi: 10.1186/s12875-022-01660-8] [Medline: 35305576]
27. As-Sanie S, Black R, Giudice LC, Gray Valbrun T, Gupta J, Jones B, et al. Assessing research gaps and unmet needs in endometriosis. Am J Obstet Gynecol 2019;221(2):86-94. [doi: 10.1016/j.ajog.2019.02.033] [Medline: 30790565]
28. Illum LRH, Forman A, Melgaard A, Hansen KE, Hansen SN, Nyegaard M, et al. Temporal and regional differences in the incidence of hospital-diagnosed endometriosis: a Danish population-based study. Acta Obstet Gynecol Scand 2022;101(7):737-746 [FREE Full text] [doi: 10.1111/aogs.14364] [Medline: 35434780]
29. Velten J, Margraf J. Exploring barriers and facilitators to women's intention and behavior to seek treatment for distressing sexual problems. PLoS One 2023;18(7):e0288205 [FREE Full text] [doi: 10.1371/journal.pone.0288205] [Medline: 37463132]
30. Kocas HD, Rubin LR, Lobel M. Stigma and mental health in endometriosis. Eur J Obstet Gynecol Reprod Biol X 2023;19:100228 [FREE Full text] [doi: 10.1016/j.eurox.2023.100228] [Medline: 37654520]
31. Avasthi A, Grover S, Sathyanarayana Rao TS. Clinical practice guidelines for management of sexual dysfunction. Indian J Psychiatry 2017;59(Suppl 1):S91-S115 [FREE Full text] [doi: 10.4103/0019-5545.196977] [Medline: 28216788]
32. Parish SJ, Hahn SR, Goldstein SW, Giraldi A, Kingsberg SA, Larkin L, et al. The International Society for the Study of Women's Sexual Health process of care for the identification of sexual concerns and problems in women. Mayo Clin Proc 2019;94(5):842-856 [FREE Full text] [doi: 10.1016/j.mayocp.2019.01.009] [Medline: 30954288]
33. Pereira VM, Arias-Carrión O, Machado S, Nardi AE, Silva AC. Sex therapy for female sexual dysfunction. Int Arch Med 2013;6(1):37 [FREE Full text] [doi: 10.1186/1755-7682-6-37] [Medline: 24066697]
34. Goldstein I, Kim NN, Clayton AH, DeRogatis LR, Giraldi A, Parish SJ, et al. Hypoactive sexual desire disorder: International Society for the Study of Women's Sexual Health (ISSWSH) expert consensus panel review. Mayo Clin Proc 2017;92(1):114-128 [FREE Full text] [doi: 10.1016/j.mayocp.2016.09.018] [Medline: 27916394]
35. Goldstein I, Komisaruk BR, Pukall CF, Kim NN, Goldstein AT, Goldstein SW, et al. International Society for the Study of Women's Sexual Health (ISSWSH) review of epidemiology and pathophysiology, and a consensus nomenclature and process of care for the management of persistent genital arousal disorder/Genito-pelvic dysesthesia (PGAD/GPD). J Sex Med 2021;18(4):665-697 [FREE Full text] [doi: 10.1016/j.jsxm.2021.01.172] [Medline: 33612417]
36. Bornstein J, Preti M, Radici G, Stockdale CK, Vieira-Baptista P. Vulvodynia: a neglected chronic pain diagnosis. Pain 2019;160(7):1680-1681. [doi: 10.1097/j.pain.0000000000001559] [Medline: 31219952]
37. Jorge CH, Bø K, Chiazuto Catai C, Oliveira Brito LG, Driusso P, Kolberg Tennfjord M. Pelvic floor muscle training as treatment for female sexual dysfunction: a systematic review and meta-analysis. Am J Obstet Gynecol 2024;231(1):51-66.e1. [doi: 10.1016/j.ajog.2024.01.001] [Medline: 38191016]
38. Sahin E, Brand A, Cetindag EN, Messelink B, Yosmaoglu HB. Pelvic physical therapy for male sexual disorders: a narrative review. Int J Impot Res 2025;37(12):941-949. [doi: 10.1038/s41443-025-01034-5] [Medline: 40016382]
39. van Reijn-Baggen DA, Han-Geurts IJM, Voorham-van der Zalm PJ, Pelger RCM, Hagenaars-van Miert CHAC, Laan ETM. Pelvic floor physical therapy for pelvic floor hypertonicity: a systematic review of treatment efficacy. Sex Med Rev 2022;10(2):209-230 [FREE Full text] [doi: 10.1016/j.sxmr.2021.03.002] [Medline: 34127429]
40. Parish SJ, Simon JA, Davis SR, Giraldi A, Goldstein I, Goldstein SW, et al. International society for the study of women's sexual health clinical practice guideline for the use of systemic testosterone for hypoactive sexual desire disorder in women. J Sex Med 2021;18(5):849-867 [FREE Full text] [doi: 10.1016/j.jsxm.2020.10.009] [Medline: 33814355]
41. Frühauf S, Gerger H, Schmidt HM, Munder T, Barth J. Efficacy of psychological interventions for sexual dysfunction: a systematic review and meta-analysis. Arch Sex Behav 2013;42(6):915-933 [FREE Full text] [doi: 10.1007/s10508-012-0062-0] [Medline: 23559141]
42. Almuqahwi A, Alabdrabulridha H, Aljumaiah RM, Alfaifi AJ, Alnaim MF, Alfaifi IA, et al. A systematic review on the relationship between physical activity and sexual function in adults. Cureus 2023;15(12):e51307 [FREE Full text] [doi: 10.7759/cureus.51307] [Medline: 38288234]
43. Maseroli E, Rastrelli G, Di Stasi V, Cipriani S, Scavello I, Todisco T, et al. Physical activity and female sexual dysfunction: a lot helps, but not too much. J Sex Med 2021;18(7):1217-1229. [doi: 10.1016/j.jsxm.2021.04.004] [Medline: 34099414]
44. Kharaghani R, Khani ME, Mahmoodi Dangesaraki M, Damghanian M. The effects of psychological interventions on sexual function of women: a systematic review and meta-analysis. Nursing Practice Today 2020;7(3). [doi: 10.18502/npt.v7i3.3345]
45. Jaderek I, Lew-Starowicz M. A systematic review on mindfulness meditation-based interventions for sexual dysfunctions. J Sex Med 2019;16(10):1581-1596. [doi: 10.1016/j.jsxm.2019.07.019] [Medline: 31570137]
46. Brotto LA, Stephenson KR, Marshall N, Balvan M, Okara Y, Mahar EA. Evaluating a digital health tool designed to improve low sexual desire in women: mixed-methods implementation science study. J Med Internet Res 2025;27:e69828 [FREE Full text] [doi: 10.2196/69828] [Medline: 40132194]

XSL•FO
RenderX

47. Basson R, Rees P, Wang R, Montejo AL, Incrocci L. Sexual function in chronic illness. J Sex Med 2010;7(1 Pt 2):374-388. [doi: 10.1111/j.1743-6109.2009.01621.x] [Medline: 20092445]
48. Hobbs LJ, Mitchell KR, Graham CA, Trifonova V, Bailey J, Murray E, et al. Help-seeking for sexual difficulties and the potential role of interactive digital interventions: findings from the third British national survey of sexual attitudes and lifestyles. J Sex Res 2019;56(7):937-946 [FREE Full text] [doi: 10.1080/00224499.2019.1586820] [Medline: 30907691]
49. Zarski AC, Berking M, Fackiner C, Rosenau C, Ebert DD. Internet-based guided self-help for vaginal penetration difficulties: results of a randomized controlled pilot trial. J Sex Med 2017;14(2):238-254. [doi: 10.1016/j.jsxm.2016.12.232] [Medline: 28161080]
50. Zarski AC, Berking M, Ebert DD. Efficacy of internet-based treatment for genito-pelvic pain/penetration disorder: results of a randomized controlled trial. J Consult Clin Psychol 2021;89(11):909-924. [doi: 10.1037/ccp0000665] [Medline: 34843312]
51. van Lankveld J. Internet-based interventions for women's sexual dysfunction. Curr Sex Health Rep 2016;8(3):136-143 [FREE Full text] [doi: 10.1007/s11930-016-0087-9] [Medline: 27512346]
52. Brotto LA, Goldmeier D. Mindfulness interventions for treating sexual dysfunctions: the gentle science of finding focus in a multitask world. J Sex Med 2015;12(8):1687-1689. [doi: 10.1111/jsm.12941] [Medline: 26289539]
53. Hucker A, McCabe MP. An online, mindfulness-based, cognitive-behavioral therapy for female sexual difficulties: impact on relationship functioning. J Sex Marital Ther 2014;40(6):561-576. [doi: 10.1080/0092623X.2013.796578] [Medline: 24308322]
54. Jones LM, McCabe MP. The effectiveness of an internet-based psychological treatment program for female sexual dysfunction. J Sex Med 2011;8(10):2781-2792. [doi: 10.1111/j.1743-6109.2011.02381.x] [Medline: 21771279]
55. Meyers M, Margraf J, Velten J. A qualitative study of women's experiences with cognitive-behavioral and mindfulness-based online interventions for low sexual desire. J Sex Res 2022;59(9):1082-1091. [doi: 10.1080/00224499.2022.2056565] [Medline: 35435791]
56. Mahar EA, Stephenson KR, Brotto LA. A randomized controlled trial of online mindfulness and cognitive-behavioral interventions for sexual interest/arousal disorder in women: eSense. Behav Res Ther 2025;188:104732 [FREE Full text] [doi: 10.1016/j.brat.2025.104732] [Medline: 40147247]
57. Goeldner M, Gehder S. Digital health applications (DiGAs) on a fast track: insights from a data-driven analysis of prescribable digital therapeutics in Germany from 2020 to mid-2024. J Med Internet Res 2024;26:e59013 [FREE Full text] [doi: 10.2196/59013] [Medline: 39208415]
58. Zarski AC, Berking M, Ebert DD. Efficacy of internet-based guided treatment for genito-pelvic pain/penetration disorder: rationale, treatment protocol, and design of a randomized controlled trial. Front Psychiatry 2017;8:260 [FREE Full text] [doi: 10.3389/fpsyt.2017.00260] [Medline: 29403395]
59. Kliesch S, Cremers J, Krallmann C, Epplen R, Scheffer B, Schubert T, et al. App-based therapy of erectile dysfunction using a digital health application (EDDIG Study): a randomized, single-blind, controlled trial. Eur Urol Focus 2024;10(6):1003-1010 [FREE Full text] [doi: 10.1016/j.euf.2024.05.020] [Medline: 38853028]
60. Impact on quality of life by additional use of the digital health application "Endo-App" to standard therapy: monocentric, prospective, randomised trial. ClinicalTrials.gov. 2021. URL: https://clinicaltrials.gov/study/NCT04883073 [accessed 2026-01-23]
61. Rohloff N, Götz T, Kortekamp SS, Heinze NR, Weber C, Schäfer SD. Influence of app-based self-management on the quality of life of women with endometriosis. Cureus 2024;16(8):e67655. [doi: 10.7759/cureus.67655] [Medline: 39314601]
62. Gratzke C, Schönburg S, Eger S, Raude K, Grabbert M, Astheimer S, et al. A randomized trial of an app-based therapeutic for lower urinary tract symptoms. NEJM Evid 2025;4(4):EVIDoa2400290. [doi: 10.1056/EVIDoa2400290] [Medline: 40130971]
63. Rödiger H, Busse R, Henschke C. The introduction of maximum reimbursement prices for digital health applications in Germany in 2022: current developments. Health Policy 2025;155:105284 [FREE Full text] [doi: 10.1016/j.healthpol.2025.105284] [Medline: 40056714]
64. Burchert S, Alkneme MS, Bird M, Carswell K, Cuijpers P, Hansen P, et al. User-centered app adaptation of a low-intensity E-mental health intervention for Syrian refugees. Front Psychiatry 2018;9:663 [FREE Full text] [doi: 10.3389/fpsyt.2018.00663] [Medline: 30740065]
65. Boß L, Lehr D, Reis D, Vis C, Riper H, Berking M, et al. Reliability and validity of assessing user satisfaction with web-based health interventions. J Med Internet Res 2016;18(8):e234 [FREE Full text] [doi: 10.2196/jmir.5952] [Medline: 27582341]
66. Moorthy P, Weinert L, Harms BC, Anders C, Siegel F. German version of the mHealth app usability questionnaire in a cohort of patients with cancer: translation and validation study. JMIR Hum Factors 2023;10:e51090 [FREE Full text] [doi: 10.2196/51090] [Medline: 37910144]
67. Bergström C, Skog R, Eriksson LE, Lampic C, Wettergren L. Efficacy of a web-based psychoeducational intervention targeting young adults with sexual problems 1.5 years after cancer diagnosis-results from a randomized controlled trial. Digit Health 2024;10:20552076241310037 [FREE Full text] [doi: 10.1177/20552076241310037] [Medline: 39741983]

XSL•FO
RenderX

68. Kim SH, Sung JH, Yoo SH, Kim S, Lee K, Oh EG, et al. Effects of digital self-management symptom interventions on symptom outcomes in adult cancer patients: a systematic review and meta-analysis. Eur J Oncol Nurs 2023;66:102404. [doi: 10.1016/j.ejon.2023.102404] [Medline: 37517339]
69. Schover LR, Yuan Y, Fellman BM, Odensky E, Lewis PE, Martinetti P. Efficacy trial of an Internet-based intervention for cancer-related female sexual dysfunction. J Natl Compr Canc Netw 2013;11(11):1389-1397. [doi: 10.6004/jnccn.2013.0162] [Medline: 24225972]
70. Linardon J, Torous J, Firth J, Cuijpers P, Messer M, Fuller-Tyszkiewicz M. Current evidence on the efficacy of mental health smartphone apps for symptoms of depression and anxiety. A meta-analysis of 176 randomized controlled trials. World Psychiatry 2024;23(1):139-149. [doi: 10.1002/wps.21183] [Medline: 38214614]
71. Schreiter S, Mascarell-Maricic L, Rakitzis O, Volkmann C, Kaminski J, Daniels MA. Digital health applications in the area of mental health. Dtsch Arztebl Int 2023;120(47):797-803 [FREE Full text] [doi: 10.3238/arztebl.m2023.0208] [Medline: 37732500]
72. Guo W, Nazari N, Sadeghi M. Cognitive-behavioral treatment for insomnia and mindfulness-based stress reduction in nurses with insomnia: a non-inferiority internet delivered randomized controlled trial. Peer J 2024;12:e17491. [doi: 10.7717/peerj.17491] [Medline: 39071123]
73. Ross EL, Jamison RN, Nicholls L, Perry BM, Nolen KD. Clinical integration of a smartphone app for patients with chronic pain: retrospective analysis of predictors of benefits and patient engagement between clinic visits. J Med Internet Res 2020;22(4):e16939 [FREE Full text] [doi: 10.2196/16939] [Medline: 32297871]
74. Strain JDR, Welch L, Sadler E. Systematic review and narrative synthesis of the experiences of individuals with chronic pain participating in digital pain management interventions. PLoS One 2024;19(7):e0306455 [FREE Full text] [doi: 10.1371/journal.pone.0306455] [Medline: 38976713]
75. Vad VB, Madrazo-Ibarra A, Estrin D, Pollak JP, Carroll KM, Vojta D, et al. "Back Rx, a personalized mobile phone application for discogenic chronic low back pain: a prospective pilot study". BMC Musculoskelet Disord 2022;23(1):923 [FREE Full text] [doi: 10.1186/s12891-022-05883-9] [Medline: 36261825]
76. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020;22(9):e20283 [FREE Full text] [doi: 10.2196/20283] [Medline: 32990635]
77. Christensen H, Griffiths KM, Farrer L. Adherence in internet interventions for anxiety and depression. J Med Internet Res 2009;11(2):e13 [FREE Full text] [doi: 10.2196/jmir.1194] [Medline: 19403466]
78. Bremer V, Chow PI, Funk B, Thorndike FP, Ritterband LM. Developing a process for the analysis of user journeys and the prediction of dropout in digital health interventions: machine learning approach. J Med Internet Res 2020;22(10):e17738 [FREE Full text] [doi: 10.2196/17738] [Medline: 33112241]
79. Cohen KA, Schleider JL. Adolescent dropout from brief digital mental health interventions within and beyond randomized trials. Internet Interv 2022;27:100496 [FREE Full text] [doi: 10.1016/j.invent.2022.100496] [Medline: 35257001]
80. Karyotaki E, Kleiboer A, Smit F, Turner DT, Pastor AM, Andersson G, et al. Predictors of treatment dropout in self-guided web-based interventions for depression: an 'individual patient data' meta-analysis. Psychol Med 2015;45(13):2717-2726 [FREE Full text] [doi: 10.1017/S0033291715000665] [Medline: 25881626]
81. Moshe I, Terhorst Y, Paganini S, Schlicker S, Pulkki-Råback L, Baumeister H, et al. Predictors of dropout in a digital intervention for the prevention and treatment of depression in patients with chronic back pain: secondary analysis of two randomized controlled trials. J Med Internet Res 2022;24(8):e38261 [FREE Full text] [doi: 10.2196/38261] [Medline: 36040780]
82. Pedersen DH, Mansourvar M, Sortsø C, Schmidt T. Predicting dropouts from an electronic health platform for lifestyle interventions: analysis of methods and predictors. J Med Internet Res 2019;21(9):e13617 [FREE Full text] [doi: 10.2196/13617] [Medline: 31486409]
83. Gan DZQ, McGillivray L, Han J, Christensen H, Torok M. Effect of engagement with digital interventions on mental health outcomes: a systematic review and meta-analysis. Front Digit Health 2021;3:764079 [FREE Full text] [doi: 10.3389/fdgth.2021.764079] [Medline: 34806079]
84. Goodday SM, Karlin E, Brooks A, Chapman C, Harry C, Lugo N, et al. Value of engagement in digital health technology research: evidence across 6 unique cohort studies. J Med Internet Res 2024;26:e57827 [FREE Full text] [doi: 10.2196/57827] [Medline: 39226552]
85. Gudmundsdóttir SL, Ballarini T, Ámundadóttir ML, Mészáros J, Eysteinsdottir JH, Thorleifsdottir RH, et al. Engagement, retention, and acceptability in a digital health program for atopic dermatitis: prospective interventional study. JMIR Form Res 2023;7:e41227 [FREE Full text] [doi: 10.2196/41227] [Medline: 36975050]
86. Huberty J, Green J, Puzia M, Stecher C. Evaluation of mood check-in feature for participation in meditation mobile app users: retrospective longitudinal analysis. JMIR Mhealth Uhealth 2021;9(4):e27106 [FREE Full text] [doi: 10.2196/27106] [Medline: 33890865]
87. Jamison RN, Jurcik DC, Edwards RR, Huang CC, Ross EL. A pilot comparison of a smartphone app with or without 2-way messaging among chronic pain patients: who benefits from a pain app? Clin J Pain 2017;33(8):676-686 [FREE Full text] [doi: 10.1097/AJP.0000000000000455] [Medline: 27898460]

XSL•FO
RenderX

88. Lipschitz JM, Pike CK, Hogan TP, Murphy SA, Burdick KE. The engagement problem: a review of engagement with digital mental health interventions and recommendations for a path forward. Curr Treat Options Psychiatry 2023;10(3):119-135 [FREE Full text] [doi: 10.1007/s40501-023-00297-3] [Medline: 38390026]
89. Wang T, Ho MH, Tong MCF, Chow JC, Voss JG, Lin CC. Effects of patient-reported outcome tracking and health information provision via remote patient monitoring software on patient outcomes in oncology care: a systematic review and meta-analysis. Semin Oncol Nurs 2023;39(5):151473. [doi: 10.1016/j.soncn.2023.151473] [Medline: 37516624]
90. Pettersson A, Berterö CM. How women with endometriosis experience health care encounters. Womens Health Rep (New Rochelle) 2020;1(1):529-542 [FREE Full text] [doi: 10.1089/whr.2020.0099] [Medline: 33786519]
91. Beier KM, von Heyden M. Sexuelle funktionsstörungen im digitalen zeitalter – therapeutische potenziale und entwicklungsperspektiven digitaler gesundheitsanwendungen. Gynäkologische Endokrinologie 2025;23(4):247-252. [doi: 10.1007/s10304-025-00645-x]
92. Beck AT, Steer RA, Brown G. Beck Depression Inventory (BDI-II). 2nd ed. San Antonio, TX: Pearson; 1996.
93. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
94. Decker-Palmer M, Klodowski D, Thompson T, Lanoue M, Messina A, Schroeder D, et al. Freelisting: a technique for enhancing the community health needs assessment. Community Health Equity Res Policy 2024;44(2):201-208. [doi: 10.1177/2752535X221146232] [Medline: 36513364]
95. Keddem S, Barg FK, Frasso R. Practical guidance for studies using freelisting interviews. Prev Chronic Dis 2021;18:E04 [FREE Full text] [doi: 10.5888/pcd17.200355] [Medline: 33444525]
96. Applied Mental Health Research Group. Module 1: qualitative assessment (adult version). In: Design, Implementation, Monitoring, and Evaluation of Mental Health and Psychosocial Assistance Programs for Trauma Survivors in Low Resource Countries: A User's Manual for Researchers and Program Implementers. Module 1: Qualitative Assessment (Adult Version). Baltimore, MD: Johns Hopkins Bloomberg School of Public Health; 2013.
97. f4transkript. Audiotranskription. 2025. URL: https://www.audiotranskription.de/f4transkript [accessed 2026-02-06]
98. Schreier M. Ways of doing qualitative content analysis: disentangling terms and terminologies. Forum Qual Soc Res 2014;15(1). [doi: 10.17169/fqs-15.1.2043]
99. Richards DA, Bazeley P, Borglin G, Craig P, Emsley R, Frost J, et al. Integrating quantitative and qualitative data and findings when undertaking randomised controlled trials. BMJ Open 2019;9(11):e032081 [FREE Full text] [doi: 10.1136/bmjopen-2019-032081] [Medline: 31772096]
100. Hwang Y, Knobf MT, Sadler LS. Integration in mixed-methods research with an exemplar explanatory sequential study. Nurs Res 2025;74(2):144-149. [doi: 10.1097/NNR.0000000000000796] [Medline: 39616427]
101. Basson R. Sexual function of women with chronic illness and cancer. Womens Health (Lond) 2010;6(3):407-429 [FREE Full text] [doi: 10.2217/whe.10.23] [Medline: 20426607]
102. Attkisson CC, Zwick R. The client satisfaction questionnaire. Psychometric properties and correlations with service utilization and psychotherapy outcome. Eval Program Plann 1982;5(3):233-237. [doi: 10.1016/0149-7189(82)90074-x] [Medline: 10259963]
103. Kopka M, Slagman A, Schorr C, Krampe H, Altendorf M, Balzer F, et al. German mHealth App Usability Questionnaire (G-MAUQ) and short version (G-MAUQ-S): translation and validation study. Smart Health 2024;34:100517. [doi: 10.1016/j.smhl.2024.100517]
104. Hayes MHS, Paterson DG. Experimental development of the graphic rating method. Psychol Bull 1921;18(2):98-99 [FREE Full text]
105. Ladwig I, Rief W, Nestoriuc Y. Welche risiken und nebenwirkungen hat psychotherapie? - Entwicklung des inventars zur erfassung negativer effekte von psychotherapie (INEP). Verhaltenstherapie 2014;24(4):252-263. [doi: 10.1159/000367928]
106. Derogatis LR, Revicki DA, Rosen RC, Jordan R, Lucas J, Spana C. Psychometric validation of the Female Sexual Distress Scale-Desire/Arousal/Orgasm. J Patient Rep Outcomes 2021;5(1):100 [FREE Full text] [doi: 10.1186/s41687-021-00359-1] [Medline: 34559353]
107. Velten J, Zarski AC. Therapie-Tools Sexuelle Funktionsstörungen: Mit E-Book inside und Arbeitsmaterial. 1st ed. Weinheim: Beltz; 2022.
108. Berner MM, Kriston L, Zahradnik HP, Härter M, Rohde A. Überprüfung der Gültigkeit und Zuverlässigkeit des deutschen Female Sexual Function Index (FSFI-d). Geburtsh Frauenheilk 2004;64(3):293-303. [doi: 10.1055/s-2004-815815]
109. Bornefeld-Ettmann P, Steil R, Höfling V, Weßlau C, Lieberz K, Rausch S, et al. Validation of the German version of the Sexual Self-Esteem Inventory for Women and its application in a sample of sexually and physically abused women. Sex Roles 2017;79(1-2):109-122. [doi: 10.1007/s11199-017-0849-5]
110. Ter Kuile MM, Melles R, de Groot HE, Tuijnman-Raasveld CC, van Lankveld JJDM. Therapist-aided exposure for women with lifelong vaginismus: a randomized waiting-list control trial of efficacy. J Consult Clin Psychol 2013;81(6):1127-1136. [doi: 10.1037/a0034292] [Medline: 24060195]
111. Klaassen M, Ter Kuile MM. Development and initial validation of the Vaginal Penetration Cognition Questionnaire (VPCQ) in a sample of women with vaginismus and dyspareunia. J Sex Med 2009;6(6):1617-1627. [doi: 10.1111/j.1743-6109.2009.01217.x] [Medline: 19538428]

XSL•FO
RenderX

112. Hahlweg K. Fragebogen zur Partnerschaftsdiagnostik (PFB) [Partnership Questionnaire]. Göttingen: Hogrefe; 1996. URL: https://www.testzentrale.de/shop/fragebogen-zur-partnerschaftsdiagnostik.html [accessed 2026-01-26]
113. Hinz A, Stöbel-Richter Y, Brähler E. Der Partnerschaftsfragebogen (PFB). Diagnostica 2001;47(3):132-141. [doi: 10.1026//0012-1924.47.3.132]
114. Klingler OJ, Loewit KK. Der Fragebogen „Ressourcen in Sexualität und Partnerschaft“ (RSP) – konzeption und erste ergebnisse zur validität. Z Differ Diagnost Psychol 1996;17(4):268-275 [FREE Full text]
115. Derogatis LR. Brief Symptom Inventory (BSI): Administration, scoring, and procedures manual. 3rd ed. Minneapolis: National Computer Systems; 1993.
116. Geisheim C, Hahlweg K, Fiegenbaum W, Frank M, Schröder B, von Witzleben I. Das Brief Symptom Inventory (BSI) als instrument zur qualitätssicherung in der psychotherapie. Diagnostica 2002;48(1):28-36. [doi: 10.1026//0012-1924.48.1.28]
117. Klein EM, Brähler E, Dreier M, Reinecke L, Müller KW, Schmutzer G, et al. The German version of the Perceived Stress Scale - psychometric characteristics in a representative German community sample. BMC Psychiatry 2016;16:159 [FREE Full text] [doi: 10.1186/s12888-016-0875-9] [Medline: 27216151]
118. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, PROMIS Cooperative Group. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH roadmap cooperative group during its first two years. Med Care 2007;45(5 Suppl 1):S3-S11 [FREE Full text] [doi: 10.1097/01.mlr.0000258615.42478.55] [Medline: 17443116]
119. Schmidt S, Mühlan H, Power M. The EUROHIS-QOL 8-item index: psychometric results of a cross-cultural field study. Eur J Public Health 2006;16(4):420-428. [doi: 10.1093/eurpub/cki155] [Medline: 16141303]
120. Klute M, Laekeman M, Kuss K, Petzke F, Dieterich A, Leha A, et al. Cross-cultural adaptation and validation of the German Central Sensitization Inventory (CSI-GE). BMC Musculoskelet Disord 2021;22(1):708 [FREE Full text] [doi: 10.1186/s12891-021-04481-5] [Medline: 34407773]
121. Brookes ST, Donovan JL, Wright M, Jackson S, Abrams P. A scored form of the Bristol Female Lower Urinary Tract Symptoms Questionnaire: data from a randomized controlled trial of surgery for women with stress incontinence. Am J Obstet Gynecol 2004;191(1):73-82. [doi: 10.1016/j.ajog.2003.12.027] [Medline: 15295345]
122. Rybak A. Survey mode and nonresponse bias: A meta-analysis based on the data from the international social survey programme waves 1996-2018 and the European social survey rounds 1 to 9. PLoS One 2023;18(3):e0283092 [FREE Full text] [doi: 10.1371/journal.pone.0283092] [Medline: 36928697]
123. Patton MQ. Qualitative Research & Evaluation Methods: Integrating Theory and Practice. 4th ed. Thousand Oaks: SAGE Publications; 2015.
124. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions: Medical Research Council guidance. BMJ 2015;350:h1258 [FREE Full text] [doi: 10.1136/bmj.h1258] [Medline: 25791983]
125. Miro. URL: https://miro.com/ [accessed 2026-01-26]
126. Blair C, Best P, Burns P, Campbell A, Davidson G, Duffy J, et al. 'Getting involved in research': a co-created, co-delivered and co-analysed course for those with lived experience of health and social care services. Res Involv Engagem 2022;8(1):20 [FREE Full text] [doi: 10.1186/s40900-022-00353-x] [Medline: 35578275]
127. Wollin-Giering S, Hoffmann M, Höfting J, Ventzke C. Automatic transcription of English and German qualitative interviews. Forum Qual Sozialforschung 2024;25(1). [doi: 10.17169/fqs-25.1.4129]
128. Plano Clark VL, Schumacher K, West C, Edrington J, Dunn LB, Harzstark A, et al. Practices for embedding an interpretive qualitative approach within a randomized clinical trial. J Mix Methods Res 2013;7(3):219-242. [doi: 10.1177/1558689812474372]
129. Laura H. Zweiarmige mixed-methods machbarkeitsstudie zur anwendung einer mobilen applikation für betroffene von endometriose und sexuellen funktionsstörungen. Deutsches Register Klinischer Studien. 2024. URL: https://drks.de/ [accessed 2026-02-01]
130. Eysenbach G. CONSORT-EHEALTH: implementation of a checklist for authors and editors to improve reporting of web-based and mobile randomized controlled trials. Stud Health Technol Inform 2013;192:657-661. [Medline: 23920638]
131. Creswell JW, Klassen AC, Clark VLP, Smith KC. Best Practices for Mixed Methods Research in the Health Sciences. Bethesda, MD: Office of Behavioral and Social Sciences Research, National Institutes of Health; 2011.
132. Cloitre M, Shevlin M, Brewin CR, Bisson JI, Roberts NP, Maercker A, et al. The international trauma questionnaire: development of a self-report measure of ICD-11 PTSD and complex PTSD. Acta Psychiatr Scand 2018;138(6):536-546. [doi: 10.1111/acps.12956] [Medline: 30178492]
133. Kübler U. Structured clinical interview for DSM-IV (SCID). In: Gellman MD, Turner JR, editors. Encyclopedia of Behavioral Medicine. New York, NY: Springer New York; 2013:1919-1920.
134. Rodriguez A, Tuck C, Dozier MF, Lewis SC, Eldridge S, Jackson T, et al. Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: a scoping review. Clin Trials 2022;19(4):452-463 [FREE Full text] [doi: 10.1177/17407745221087469] [Medline: 35730910]
135. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]

XSL•FO
RenderX

136. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, REDCap Consortium. The REDCap Consortium: building an international community of software platform partners. J Biomed Inform 2019;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
137. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
138. Klinitzke G, Romppel M, Häuser W, Brähler E, Glaesmer H. [The German version of the Childhood Trauma Questionnaire (CTQ): psychometric characteristics in a representative sample of the general population]. Psychother Psychosom Med Psychol 2012;62(2):47-51. [doi: 10.1055/s-0031-1295495] [Medline: 22203470]
139. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health Apps: systematic search and panel-based usage analysis. J Med Internet Res 2019;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
140. Boucher EM, Raiker JS. Engagement and retention in digital mental health interventions: a narrative review. BMC Digit Health 2024;2(1):52. [doi: 10.1186/s44247-024-00105-9]
141. Diaz-Mohedo E, Hita-Contreras F, Castro-Martin E, Pilat A, Perez-Dominguez B, Valenza-Peña G. Using myofascial therapy to improve psychological outcomes, quality of life, and sexual function in women with chronic pelvic pain-a case series. Healthcare (Basel) 2024;12(3):304 [FREE Full text] [doi: 10.3390/healthcare12030304] [Medline: 38338190]
142. Livieri G, Mangina E, Protopapadakis ED, Panayiotou AG. The gaps and challenges in digital health technology use as perceived by patients: a scoping review and narrative meta-synthesis. Front Digit Health 2025;7:1474956 [FREE Full text] [doi: 10.3389/fdgth.2025.1474956] [Medline: 40212901]
143. Erbe D, Eichert H, Riper H, Ebert DD. Blending face-to-face and internet-based interventions for the treatment of mental disorders in adults: systematic review. J Med Internet Res 2017;19(9):e306 [FREE Full text] [doi: 10.2196/jmir.6588] [Medline: 28916506]
144. Liu H, Xie Z, Or C. Willingness to pay for health apps, its sociodemographic correlates, and reasons for being unwilling to pay. Digit Health 2024;10:20552076241248925 [FREE Full text] [doi: 10.1177/20552076241248925] [Medline: 38698831]
145. Berkhout C, Berbra O, Favre J, Collins C, Calafiore M, Peremans L, et al. Defining and evaluating the Hawthorne effect in primary care, a systematic review and meta-analysis. Front Med (Lausanne) 2022;9:1033486 [FREE Full text] [doi: 10.3389/fmed.2022.1033486] [Medline: 36425097]
146. Long PA, Huberts AS, di Torrero AN, Otto LR, Rogge AA, Ritschl V, et al. The mere-measurement effect of patient-reported outcomes: a systematic review and meta-analysis. Qual Life Res 2025;34(5):1211-1220. [doi: 10.1007/s11136-025-03909-y] [Medline: 39907942]
147. Gallagher MW, Long LJ, Richardson A, D'Souza J, Boswell JF, Farchione TJ, et al. Examining hope as a transdiagnostic mechanism of change across anxiety disorders and CBT treatment protocols. Behav Ther 2020;51(1):190-202 [FREE Full text] [doi: 10.1016/j.beth.2019.06.001] [Medline: 32005336]
148. Bradford A, Meston CM. Placebo response in the treatment of women's sexual dysfunctions: a review and commentary. J Sex Marital Ther 2009;35(3):164-181. [doi: 10.1080/00926230802716302] [Medline: 19360517]
149. Song MK, Ward SE. Assessment effects in educational and psychosocial intervention trials: an important but often-overlooked problem. Res Nurs Health 2015;38(3):241-247 [FREE Full text] [doi: 10.1002/nur.21651] [Medline: 25728502]
150. Chen K, Huang JJ, Torous J. Hybrid care in mental health: a framework for understanding care, research, and future opportunities. NPP Digit Psychiatry Neurosci 2024;2(1):16. [doi: 10.1038/s44277-024-00016-7] [Medline: 41360926]
151. Loughnane C, Laiti J, O'Donovan R, Dunne PJ. Systematic review exploring human, AI, and hybrid health coaching in digital health interventions: trends, engagement, and lifestyle outcomes. Front Digit Health 2025;7:1536416 [FREE Full text] [doi: 10.3389/fdgth.2025.1536416] [Medline: 40343215]
152. Saleem M, Kühne L, De Santis KK, Christianson L, Brand T, Busse H. Understanding engagement strategies in digital interventions for mental health promotion: scoping review. JMIR Ment Health 2021;8(12):e30000 [FREE Full text] [doi: 10.2196/30000] [Medline: 34931995]
153. Moe-Byrne T, Shepherd J, Merecz-Kot D, Sinokki M, Naumanen P, Hakkaart-van Roijen L, et al. Effectiveness of tailored digital health interventions for mental health at the workplace: a systematic review of randomised controlled trials. PLOS Digit Health 2022;1(10):e0000123 [FREE Full text] [doi: 10.1371/journal.pdig.0000123] [Medline: 36812547]
154. Bidargaddi N, Almirall D, Murphy S, Nahum-Shani I, Kovalcik M, Pituch T, et al. To prompt or not to prompt? A microrandomized trial of time-varying push notifications to increase proximal engagement with a mobile health app. JMIR Mhealth Uhealth 2018;6(11):e10123 [FREE Full text] [doi: 10.2196/10123] [Medline: 30497999]
155. Bitrián P, Buil I, Catalán S. Enhancing user engagement: The role of gamification in mobile apps. J Bus Res 2021;132:170-185 [FREE Full text] [doi: 10.1016/j.jbusres.2021.04.028]
156. Twomey C, O'Reilly G, Meyer B. Effectiveness of an individually-tailored computerised CBT programme (Deprexis) for depression: a meta-analysis. Psychiatry Res 2017;256:371-377. [doi: 10.1016/j.psychres.2017.06.081] [Medline: 28686935]
157. Molina-Recio G, Molina-Luque R, Jiménez-García AM, Ventura-Puertos PE, Hernández-Reyes A, Romero-Saldaña M. Proposal for the user-centered design approach for health apps based on successful experiences: integrative review. JMIR Mhealth Uhealth 2020;8(4):e14376 [FREE Full text] [doi: 10.2196/14376] [Medline: 32319965]

XSL•FO
RenderX

158. Saparamadu AADNS, Fernando P, Zeng P, Teo H, Goh A, Lee JMY, et al. User-centered design process of an mHealth app for health professionals: case study. JMIR Mhealth Uhealth 2021;9(3):e18079 [FREE Full text] [doi: 10.2196/18079] [Medline: 33769297]
159. de Paula D, Jühling D, Uebernickel F. Designing digital health innovation for stigma-affected health conditions: a systems-based view to managing chronic pelvic pain. Int J Hum Comput Interact 2025:1-15. [doi: 10.1080/10447318.2025.2485494]
160. Willis E, Friedel K, Heisten M, Pickett M, Bhowmick A. Communicating health literacy on prescription medications on social media: in-depth interviews with "patient influencers". J Med Internet Res 2023;25:e41867 [FREE Full text] [doi: 10.2196/41867] [Medline: 36912881]

Abbreviations

BfArM: German Federal Institute for Drugs and Medical Devices (Bundesinstitut für Arzneimittel und Medizinprodukte)
CG: control group
CSI-GE: Central Sensitization Inventory-German Version
CSQ-I: Client Satisfaction Questionnaire- Internet Version
DiGA: digital health application
DSM-IV: Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition
FSDS-DAO: Female Sexual Distress Scale-Desire/Arousal/Orgasm
FSFI-d: Female Sexual Function Index-German version
FSQ: Fear of Sexuality Questionnaire
G-MAUQ: German mHealth App Usability Questionnaire
ICD-11: International Classification of Diseases, 11th Revision
IG: intervention group
INEP-ON: Inventory for the balanced assessment of Negative Effects of Psychotherapy
PFB: Partnership Questionnaire
PROMIS-29: Patient-Reported Outcome Measurement Information System-29-Item Profile
PTSD: posttraumatic stress disorder
REDCap: Research Electronic Data Capture
T0: baseline
T1: after module 5 in intervention group/5 weeks after baseline in control group
T2: after module 8 in intervention group/8 weeks after baseline in control group
T3: 6-month follow-up after baseline
VPCQ: Vaginal Penetration Cognition Questionnaire

Edited by A Stone; submitted 26.Oct.2025; peer-reviewed by G Cabagno; comments to author 20.Nov.2025; revised version received 27.Dec.2025; accepted 31.Dec.2025; published 19.Feb.2026.

Please cite as:
Kronthaler SM, Kosman E, Tissen-Diabaté T, Mühle E, Weber- Schallauer L, Pross T, Karsten MM, Beier KM, Hatzler L
Adherence, Acceptability, and Sexual Health Outcomes of the Odeya App–Based Intervention for Sexual Distress in Women With Endometriosis: Randomized Controlled Mixed Methods Trial
J Med Internet Res 2026;28:e86042
URL: https://www.jmir.org/2026/1/e86042
doi:10.2196/86042
PMID:

©Selina Marie Kronthaler, Eden Kosman, Tatjana Tissen-Diabaté, Elena Mühle, Luzia Weber- Schallauer, Therese Pross, Maria Margarete Karsten, Klaus Michael Beier, Laura Hatzler. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Mental Health Professionals' Perceptions of Benefits and Disadvantages of Telehealth: International Mixed Methods Study

Madeline I Montoya[1], BA; Cary S Kogan[1], PhD; Brigitte Khoury[2], PhD; José A García-Pacheco[3], MA; Rebeca Robles[4], PhD; Tahilia J Rebello[5], PhD; Maya Kulygina[6], PhD; Geoffrey M Reed[7], PhD

[1]School of Psychology, University of Ottawa, Ottawa, ON, Canada
[2]Department of Psychiatry, American University of Beirut Medical Center, Beirut, Lebanon
[3]Centro de Investigación y Docencia Económicas, Mexico City, Mexico
[4]Center for Global Mental Health Research, Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz, Mexico City, Mexico
[5]Columbia University Vagelos College of Physicians and Surgeons, New York, NY, United States
[6]Training and Research Centre, Mental-Health Clinic No.1 Named After N.A. Alekseev, Moscow, Russian Federation
[7]Department of Psychiatry, Columbia University Vagelos College of Physicians and Surgeons, New York, NY, United States

Corresponding Author:
Cary S Kogan, PhD
School of Psychology
University of Ottawa
Vanier Building
136 Jean-Jacques Lussier Pvt
Ottawa, ON, K1N 6N5
Canada
Phone: 1 6135625800 ext 4804
Email: ckogan@uottawa.ca

Abstract

Background: Telehealth has become an integral component of mental health care delivery worldwide. Understanding provider perceptions is essential to guiding its continued implementation.

Objective: This international study used quantitative and qualitative methodologies to examine and broaden our understanding of the benefits and concerns related to telehealth for mental health care.

Methods: An internet-based survey was conducted during the COVID-19 pandemic between November 11 and December 18, 2020, among mental health professionals, primarily psychiatrists and psychologists, registered with the World Health Organization's Global Clinical Practice Network. Clinicians completed the survey in 1 of 6 languages (Chinese, English, French, Japanese, Russian, or Spanish). Descriptive statistics and logistic regressions were used to analyze quantitative survey data on concerns and implementation of telehealth. Responses to an open-ended question about providers' perspectives on the benefits of telehealth were analyzed qualitatively.

Results: In total, 847 participants completed the telehealth section of the survey, and 496 provided a response to the open-ended question. Quantitative data on telehealth use and concerns revealed that clinicians' primary concerns focused on technical issues and clinical effectiveness relative to in-person services, specifically, loss of clinical information (eg, nonverbal behavior) and challenges with establishing a therapeutic alliance. Findings varied by profession, World Health Organization region, and telehealth training and experience. Qualitative data examining benefits fell into 3 major areas: accessibility and reach of mental health services, efficiency and flexibility for clinicians, and enhancement of clinical processes and outcomes. Taken together, findings revealed a trade-off between telehealth benefits and disadvantages.

Conclusions: From the perspective of mental health professionals, telehealth practice comes with key challenges and valuable benefits. Findings offer important considerations for the implementation of telehealth systems, including the importance of training and education and balancing trade-offs to optimize care.

(J Med Internet Res 2026;28:e75905) doi:10.2196/75905

KEYWORDS

global mental health; telehealth; mental health professionals; COVID-19; implementation science

XSL•FO
RenderX

Introduction

Following the onset of the COVID-19 pandemic and subsequent disruption of mental health services, clinicians and health care systems around the world experienced an abrupt shift to remote delivery of mental health services [1,2]. The term telehealth will be used, which is defined by the World Health Organization (WHO) as "the delivery of health care services, where patients and providers are separated by distance" [3]. Telehealth encompasses the use of information communication technology for the diagnosis and treatment of diseases and injuries, research, and continuing education of health professionals. This broad definition will include services provided by all health professionals working in mental health (eg, psychiatry and psychology) across various modalities, both synchronous (eg, videoconferencing or telephone) and asynchronous (eg, SMS text message or email).

Telehealth became central for continuity of mental health care throughout the pandemic and as a solution to meeting increased demand [4,5]. Governments and systems launched digital platforms [6,7] and eased regulatory, policy, and billing restrictions [8], while professional bodies produced guidelines to assist clinicians [9,10]. These combined efforts led to a substantial global surge in telehealth use for mental health care, which has remained robust in the postpandemic era [11,12].

Numerous studies conducted prior to and during the COVID-19 pandemic have shown favorable perceptions about telehealth including high client satisfaction and provider acceptability [1,13]. The evidence has demonstrated the feasibility and effectiveness of telehealth compared to in-person services for the assessment and treatment of various mental disorders and patient populations [14-16]. Telehealth can improve access to care in underserved areas and for individuals who typically experience barriers to accessing in-person services [1]. Telehealth can also be convenient (eg, reducing burdensome travel [17], cost-effective, and time-efficient for patients and providers [18]). Telepsychiatry research suggests additional advantages such as reduced psychiatric admissions [19,20] and greater scheduling flexibility [21]. However, the literature on the benefits of telehealth is fragmented, lacking a cohesive perspective across mental health providers and geographic regions.

Despite largely positive accounts of the potential of telehealth, numerous challenges still exist. Research conducted prior to and during the COVID-19 pandemic found that mental health professionals from various countries remain concerned with clinical elements of telehealth such as quality or effectiveness of care compared to in-person services [2,18] and assessing patients remotely. Technological challenges remain a barrier to telehealth use, particularly technical issues (eg, poor connectivity) and client access to technology [1,22,23]. Finally, although professional guidelines and easing of restrictions addressed many ethical, regulatory, and legal barriers to telehealth, some studies show that providers during the pandemic still reported concerns about patient privacy and confidentiality [2]. Therefore, it is important to comprehensively examine which clinical, technological, and ethical concerns are most challenging for mental health providers. For example, the specific sources that contribute to providers' perceptions of telehealth as less effective relative to in-person services remain unclear. Several studies suggest that mental health professionals report difficulty establishing a therapeutic alliance with their clients via telehealth [19,24], although findings are mixed [25]. Others report difficulty with interpreting patients' nonverbal communication [26]. Still other research suggests that difficulties with therapeutic skills (eg, showing empathy) is also a factor [27].

The prevailing view is that telehealth is a largely effective and more equitable means of service provision that is expected to remain a significant component of service provision. However, it may not be appropriate or effective for everyone. More evidence is needed to understand if telehealth is feasible and effective for certain patient populations and types of services [1,15]. Therefore, mental health professionals must assess the advantages and disadvantages of offering telehealth services in their particular contexts with specific patients. Moreover, given that clinicians are considered gatekeepers to implementation and sustainability of telehealth for mental health care [23], understanding provider perceptions is key to informing continued use.

The COVID-19 pandemic created a unique opportunity to assess the perceptions of telehealth providers not only from numerous clinical settings and geographical locations but also from professionals with greater generalizability than the subset of participants in prepandemic telehealth studies who may have been influenced by selection bias (ie, interest in telehealth) and existing infrastructure for virtual care. Therefore, research conducted during the pandemic contains informative perspectives from both groups of mental health providers: those forced to transition to telehealth and those familiar with and already implementing telehealth.

This mixed methods study used distinct but complementary qualitative and quantitative research questions to obtain more comprehensive information on the experiences, concerns, and benefits of telehealth from the perspective of international mental health providers during the COVID-19 pandemic [28,29]. Participants were drawn from the WHO's Global Clinical Practice Network (GCPN), a multilingual and international network of psychologists, psychiatrists, and other mental health professionals from 160 countries representing all WHO global regions [30]. It was hypothesized that the perspectives of mental health professionals would mirror previous findings on the advantages and disadvantages of telehealth as well as illustrate differential experiences across contexts (eg, world region) and professional demographics (eg, discipline, telehealth training, and experience [1,2]). Telehealth research often examines psychologists and psychiatrists either separately (ie, telepsychiatry) or treats them as a single group despite meaningful differences in the scope of practice (eg, medication management and physical health assessments). This study examines differences between the 2 professions, as discipline was hypothesized to shape telehealth use and perceptions. The term clinician is used to refer collectively to psychologists, psychiatrists, and other mental health professionals registered with the GCPN.

XSL•FO
RenderX

Methods

Study Design, Participants, and Procedure

This study was part of the "Longitudinal COVID-19 Survey of Mental Health Professionals Project," a 3-wave longitudinal internet-based survey study examining the effects of the COVID-19 pandemic on global mental health professionals' practice including telehealth [2] and occupational well-being [31]. This study focuses on the survey questions related to telehealth included in the second wave of data collection, which occurred between November 11 and December 18, 2020 (for a total data collection period of 5 weeks). A convergent mixed methods design was used, in which quantitative and qualitative data were collected in parallel (ie, in the same survey), analyzed separately with equal priority, and then integrated using a joint display of the 2 forms of data to provide additional insights [32]. This study used the GRAMMS (Good Reporting of a Mixed Methods Study) criteria [33] to guide reporting (Multimedia Appendix 1).

The Longitudinal COVID-19 Survey of Mental Health Professionals Project procedures and recruitment for the telehealth section of the survey have been described elsewhere in more detail [2]. In brief, participants were recruited from the GCPN. To be eligible for registration with the GCPN, providers must have completed their training and be qualified to provide mental health services in their countries. Initially, GCPN members registered because of interest in participating in field studies related to the *International Classification of Diseases, 11th Revision* (*ICD-11*) being conducted by the WHO [30].

The online survey was administered using Qualtrics in Chinese, English, French, Japanese, Russian, and Spanish. Content was initially developed in English and then translated by fluent experts affiliated with the GCPN's International Advisory Group.

GCPN members proficient in the corresponding survey languages were invited to participate via email containing an individualized survey. Upon accessing the link, participants were presented with a description of the study. Reminder emails were sent 7 and 14 days after the initial email. Participants who indicated that they had provided direct clinical services in the past 2 weeks and had used telehealth for either new or continuing patients were directed to the telehealth survey questions.

Ethical Considerations

The study was approved by the institutional review board at Columbia University/New York State Psychiatric Institute (registration: #6886) and the University of Ottawa (registration: H-06-20-5973). Participants, who were practicing mental health professionals and members of the GCPN, were asked to provide consent for participation by selecting "Yes" to proceed with the online survey after reviewing a description of the study. Participants were not compensated for participation. Study data were deidentified to maintain participant privacy and confidentiality.

Survey

Quantitative data were collected from closed-ended items about telehealth concerns and implementation, and qualitative data were collected from an open-ended question about clinicians' perspectives on the benefits of telehealth. Telehealth survey items were developed from a review of the existing literature at the time (including data from the first-wave telehealth survey published in Montoya et al [2]), together with group discussion and expert review by the authors and the GCPN's International Advisory Group (Multimedia Appendix 2).

Upon starting the survey, participants were informed that they would be asked about their recent work activities and to have their work calendars accessible when completing the survey. Participants were about years of telehealth experience, if they had received training on technological, ethical, legal, and clinical aspects of telehealth service delivery or no training at all, and current (ie, past 2 weeks) telehealth use. Demographic and telehealth use survey items included closed-ended and multiple response (or check-all-that-apply) questions. Likert-type questions about perceived effectiveness of telehealth for specific age groups and treatment modality (eg, individuals and couples) were also included.

To identify specific challenges, 4 separate checklists were presented, which focused on clinical, technical, ethical, legal, and professional, and finally, administrative elements of providing telehealth services. If participants endorsed concerns about "clinical effectiveness relative to in-person services" in the checklist of clinical elements, they were directed to an additional checklist containing more detailed concerns related to clinical effectiveness (eg, building and maintaining a therapeutic alliance and showing empathy). These options for specific concerns were based on the telehealth literature [13] as well as basic clinical skills that are considered important for therapeutic change [34,35]. Participants could select multiple items for each checklist and use an open-text field to specify an "other" response.

The following open-ended question was included to capture participants' perspectives on the benefits of telehealth: "In addition to making it possible to consult with patients remotely, have there been any benefits of using telehealth services?"

Quantitative Data Analysis

Of a total of 15,179 GCPN members who were invited to participate in the larger study, 2124 (13.9%) agreed to participate in the survey, and 2088 (98.3%) of those participants provided sufficient data for analysis (ie, <90% missing survey data). Of those participants, 847 completed the telehealth section. Descriptive and chi-square statistics were used to summarize and compare quantitative survey data. Logistic regressions were used to examine predictors of clinicians' common concerns. Statistical calculations were performed using SPSS (version 29; IBM Corp).

Qualitative Data Analysis

The purpose of the qualitative data analysis was to examine and summarize the content of clinicians' responses. The analysis was guided by recommendations from Miles et al [29] and

XSL•FO
RenderX

consisted of first- and second-cycle coding methods. First cycle coding is a way of assigning initial descriptive codes to the data, whereas second cycle coding is the reorganization of these first cycle codes into meaningful themes, categories, or concepts [36].

Of the 847 clinicians who completed the telehealth section survey, 496 (58.6%) provided free-text responses to the open-ended question. Responses ranged from 1-word phrases to a paragraph; however, most were a few sentences in length. In general, characteristics of those who completed the open-ended question were similar to those who did not. Significant differences were found for telehealth training, with a greater proportion of participants with telehealth training answering the open-ended question (χ^2_1=12.1; P<.001). Moreover, those who responded to the open-ended question were older (t_{845}=4.6; P<.001) and had slightly more years of clinical experience (t_{845}=2.8; P=.003). All 496 responses were imported to and analyzed using NVivo (version 14; Lumivero). Two of the authors (MIM and CSK) coded the data and consulted with a researcher with experience in approaches relevant to this study. Coders were either actively or passively fluent in the 3 most common survey languages (English, Spanish, and French); thus, 421 (84.5%) responses were kept in their original language until the end of data analysis. MIM, the primary coder, was fluent in the former 3 languages. CSK was fluent in English and French and could read and understand Spanish. Chinese, Japanese, and Russian responses were translated to English prior to coding using DeepL Translator, a machine translation tool known for its accuracy and natural, high-quality translations. DeepL Translator is owned by DeepL SE, a German-based company that uses artificial intelligence to offer various translation supports.

Following an inductive approach, MIM and CSK used initial coding (an open-ended, data-driven coding process [36]) and in vivo coding (a coding technique that uses participants' exact words or phrases to extract ideas [36]) to independently code 150 responses (33% of the data). This initial sample was selected by sorting the full sample of responses by WHO region and coding a proportional number of responses based on WHO region and language. Responses were intentionally skipped as needed to ensure proportional representation. For example, 10% (15/150) of responses were from Western Pacific-Asia; thus, 15 responses from this region were coded as part of the initial sample of 150 responses. Next, coders shared their initial codes to collaboratively develop a preliminary codebook.

Discrepancies in code definitions, labels, and interpretations were resolved through discussion until consensus was reached. Using this initial codebook, a process of intercoder reliability was conducted. MIM and CSK independently coded a sample of 100 (20%) responses different from the initial sample that was used to develop initial codes. κ statistics (overall κ=0.65) and percentage agreement (ranging from 65% to 100%) were used to identify codes with lower agreement. The coders discussed inconsistencies until consensus was reached. Codes were merged, dropped, and refined until a final codebook was generated. MIM then coded all 496 responses.

Two second-cycle coding methods were used: pattern and axial coding. MIM conducted a process of pattern coding to identify emergent themes, configurations, or explanations in the data [29]. Axial coding was used to distinguish properties and dimensions of emergent categories [36] such as benefits specific to patients versus clinicians. Experiential differences and similarities across the WHO region and profession were also examined. MIM and CSK discussed and developed emergent themes and overall conceptualization. The relative frequency with which codes were applied was used to distinguish between topics that arose more often than others to better characterize the diversity of perceptions reported by clinicians [37].

Efforts were taken to ensure the quality of the data analysis. Methodological and analytic documentation (eg, memo writing) was used to keep track of the coding process and related decisions [36]. During meetings, MIM and CSK attempted to maintain reflexivity and note sources of bias during analysis (eg, the impact of one's own sociocultural background on data interpretation). Furthermore, coders were blinded to all demographic information when analyzing codable text except for language and profession. The rationale was that access to this information could help contextualize responses (eg, assessment could include physical examination in psychiatry). Coders searched for disconfirming evidence (ie, counterexamples to recurring patterns), and that no substantively new themes or patterns emerged in the final stages of coding the data sample [29].

Integrating Quantitative and Qualitative Data

After separate analyses, frequently endorsed concerns (quantitative data) and identified benefits (qualitative data) were brought together in a table for elaboration [32]. MIM and CSK met to conduct a higher-level interpretation of findings to supplement and enhance understanding of the advantages and disadvantages of telehealth for mental health care [28].

Results

Quantitative: Telehealth Use and Concerns

Overview

Clinician demographic and professional characteristics are provided in Table 1. Participants represented all WHO regions and 73 countries. Most worked in mental health settings (754/847, 89%), followed by substance abuse specialty programs (200/847, 23.6%), university or other educational institutions (182/847, 21.5%), other general medical settings (148/847, 17.5%), primary care settings (67/847, 7.9%), and finally, hospital emergency departments (15/847, 1.8%). Approximately half of the participants (446/847, 52.7%) reported working in multiple settings. Most participants (776/847, 91.6%) reported that they had provided mental health services (such as diagnostic assessment, psychological assessment, or psychotherapy) during the past 2 weeks. Additionally, 84.8% (263/310) of psychiatrists had also provided psychopharmacotherapy, and 29% (90/310) had provided other medical services (including general medical services and management of comorbid medical conditions).

XSL•FO
RenderX

Participants reported using various modalities for direct clinical services (assessment, treatment, or monitoring of patients) including in person (607/847, 71.7%), videoconferencing (632/847, 74.6%), telephone (654/847, 77.2%), chat (353/847, 41.7%), and email (359/847, 42.4%). Psychiatrists were more likely to see patients in person compared to psychologists (χ^2_1=28.5; P<.001). Table 2 provides the number of hours spent and patients seen using telehealth for psychiatrists and psychologists in the past 2 weeks.

Table 3 shows providers' perceived effectiveness of telehealth by treatment modality (individual, group, families, and couples) and patient age groups (older adults, adults, adolescents, and children).

XSL•FO
RenderX

Table 1. Clinician demographics and occupational characteristics.

	Sample of telehealth participants (N=847)
Gender, n (%)	
Men	372 (43.9)
Women	472 (55.7)
Other self-identified	3 (0.4)
Language, n (%)	
English	454 (53.6)
Spanish	174 (20.5)
Japanese	67 (7.9)
Chinese	29 (3.4)
French	53 (6.3)
Russian	70 (8.3)
WHO[a] region[b], n (%)	
Africa	22 (2.6)
Americas-South	135 (15.9)
Americas-North	156 (18.4)
Eastern Mediterranean	13 (1.5)
Europe	343 (40.5)
South-East Asia	48 (5.7)
Western Pacific-Asia	106 (12.5)
Western Pacific-Oceania	24 (2.8)
Country income level, n (%)	
Low	4 (0.5)
Lower-middle	87 (10.3)
Upper-middle	219 (25.9)
High	537 (63.4)
Profession, n (%)	
Psychiatry	310 (36.6)
Psychology	384 (45.3)
Other[c]	153 (18.1)
Age (years)	
Mean (SD)	52.3 (11.3)
Range	25-89
Median (IQR)	52 (43-60)
Years of experience	
Mean (SD)	20.9 (10.2)
Range	0-57
Median (IQR)	20 (13-29)
Telehealth training, n (%)	
Yes	475 (56.1)
No training	372 (43.9)
Telehealth experience, n (%)	

XSL•FO
RenderX

	Sample of telehealth participants (N=847)
Less than 1 year	419 (49.5)
More than 1 year	428 (50.5)

[a]WHO: World Health Organization.

[b]Participants were predominantly from the following countries in each WHO region—Africa: Nigeria, South Africa; Americas-South: Mexico, Argentina, Brazil, Columbia, Peru; Americas-North: the United States, Canada; Eastern Mediterranean: United Arab Emirates, Egypt; Europe: Spain, Russia, Germany, the United Kingdom, France, Denmark, Switzerland; South-East Asia; India; Western Pacific-Asia: Japan, China; and Western Pacific-Oceania: Australia, New Zealand.

[c]Other professionals included predominantly social work, nursing, counseling, occupational therapy, and sex therapy.

Table 2. Number of patients seen and hours spent using telehealth in the past 2 weeks.

	Values, n	Psychologist, mean (SD)	Values, n	Psychiatrist, mean (SD)
Telephone	263		245	
Patients		7.79 (13.34)		8.48 (9.27)
Hours		7.40[a] (10.06)		5.79[a] (7.09)
Videoconferencing	332		201	
Patients		10.97 (13.36)		9.83 (11.77)
Hours		14.96[a] (17.12)		10.66[a] (12.29)
Instant messaging	126		141	
Patients		5.82 (10.08)		7.77 (10.26)
Hours		6.53 (14.41)		5.35 (10.04)
Email	137		114	
Patients		3.58 (5.21)		4.81 (4.88)
Hours		2.52 (3.59)		2.31 (2.76)

[a]Significantly different at $P<.05$.

Table 3. Clinicians who reported that the telehealth services they provided for specific patient populations were about the same or more effective than in-person by treatment modality and patient age groups.

	Value, n/N (%)
Patient age groups	
Adults (65+ years)	155/406 (38.2)
Adults (19-64 years)	468/751 (62.3)
Adolescents (13-18 years)	192/359 (53.5)
Children (0-12 years)	42/188 (22.3)
Treatment modality	
Groups	68/200 (34)
Families	88/271 (32.5)
Couples	106/270 (39.3)
Individual assessment and treatment	425/774 (54.9)

Table 4 lists the most common concerns (endorsed by 40% or more) regarding clinical, technical, ethical, legal, professional, and administrative aspects of telehealth provision. Psychologists were significantly more concerned about a lack of private space (χ^2_1=7.6; P=.006) and technical issues (χ^2_1=4.8; P=.03) when compared to psychiatrists. Conversely, psychiatrists were significantly more concerned with remote assessment of patients (χ^2_1=32.5; P<.001), assessing high-risk patients and managing emergencies (χ^2_1=12.4; P<.001), showing empathy (χ^2_1=5.6; P=.02), and licensing or credentialing requirements (χ^2_1=36.1; P<.001) when compared to psychologists.

XSL•FO
RenderX

Table 4. List of most frequent telehealth concerns endorsed by at least 40% of psychiatrists, psychologists, and/or other professionals (mainly social work, nursing, counseling, occupational therapy, and sex therapy).

	Psychiatrist, n (%)	Psychologist, n (%)	Other, n (%)	Total, n (%)
Clinical concerns				
Clinical effectiveness relative to in-person services	196 (68.5)	224 (60.7)	86 (57)	506 (62.8)
Loss of clinical information	158 (81.4)	175 (78.8)	63 (75)	396 (79.2)
Building and maintaining a therapeutic alliance	141 (72.7)	150 (67.6)	60 (71.4)	351 (70.2)
Showing empathy	103 (53.1)	92 (41.4)	42 (50)	237 (47.4)
Implementing therapeutic interventions	68 (35.1)	97 (43.7)	43 (51.2)	208 (41.6)
Assessment of high-risk patients and managing emergencies remotely	194 (67.8)	189 (51.2)	77 (51)	460 (57.1)
Lack of private space for patients during telehealth sessions	128 (44.8)	199 (53.9)	60 (39.7)	387 (48)
Remote assessment of patients (eg, diagnostic assessments, self-report measures, lack of opportunity for physical examinations, and medication blood level monitoring)	171 (59.8)	129 (35)	58 (38.4)	358 (44.4)
Decrease in patient engagement compared to in-person services	119 (41.6)	151 (40.9)	70 (46.4)	340 (42.2)
Technological concerns				
Technical issues (eg, audio or video quality and stability of internet connection)	211 (71)	294 (78.4)	116 (78.9)	621 (75.8)
Patients' access to equipment	177 (59.6)	230 (61.3)	93 (63.3)	500 (61.1)
Patient familiarity with connecting	137 (46.1)	166 (44.3)	77 (52.4)	380 (46.4)
Ethical, legal, and professional concerns				
Patient consent, privacy, security, and confidentiality	112 (42.6)	128 (38.6)	61 (42.4)	301 (40.7)
Licensure or credentialing requirements	116 (44.1)	70 (21.1)	34 (23.6)	220 (29.8)
Administrative concerns				
Disruption of routine and workflow	121 (40.7)	131 (34.9)	58 (39.5)	310 (37.9)

Logistic Regressions

When chi-square tests showed associations between at least 2 demographic variables and the most common concerns, binomial logistic regressions were used to determine whether demographic and professional variables affected the likelihood that clinicians endorsed a common concern. Predictors included age, years of clinical experience, profession (psychiatrists, psychologists, and other professionals), telehealth experience (<1 year vs >1 year), telehealth training received (no vs yes), and WHO region. Age and years of clinical experience did not significantly contribute to overall model fit and were thus excluded from the final models (Table 5).

XSL•FO
RenderX

Table 5. Logistic regression of predictors of common telehealth concerns.

Predictors	OR[a] (95% CI)	*P* value
Clinical effectiveness relative to in-person services		
Profession[b]		
Psychiatrist	1.336 (0.847-2.106)	.21
Psychologist	1.045 (0.697-1.566)	.83
WHO[c] region[d]		
Africa	1.671 (0.631-4.428)	.30
Americas-South	1.301 (0.778-2.175)	.32
Eastern Mediterranean	3.691 (0.771-17.662)	.10
Europe	1.314 (0.853-2.024)	.22
South-East Asia	1.394 (0.695-2.795)	.35
Western Pacific-Asia	1.408 (0.779-2.545)	.26
Western Pacific-Oceania	1.179 (0.482-2.886)	.72
Telehealth training[e] (none)	1.378 (1.011-1.878)	.04
Telehealth experience[f] (<1 year)	1.348 (1.004-1.812)	.047
Remote assessment of patients		
Profession[b]		
Psychiatrist	2.079 (1.319-3.276)	.002
Psychologist	0.901 (0.595-1.363)	.62
WHO region[d]		
Africa	2.077 (0.791-5.453)	.14
Americas-South	0.920 (0.545-1.551)	.75
Eastern Mediterranean	1.167 (0.354-3.855)	.80
Europe	0.753 (0.482-1.175)	.21
South-East Asia	2.047 (1.016-4.124)	.045
Western Pacific-Asia	1.877 (1.036-3.401)	.04
Western Pacific-Oceania	0.720 (0.284-1.823)	.49
Telehealth training[e] (none)	1.429 (1.050-1.944)	.02
Telehealth experience[f] (<1 year)	1.010 (0.751-1.358)	.95
Lack of private space for patients during telehealth sessions		
Profession[b]		
Psychiatrist	1.447 (0.922-2.271)	.11
Psychologist	1.837 (1.224-2.759)	.003
WHO region[d]		
Africa	0.758 (0.301-1.908)	.56
Americas-South	1.372 (0.822-2.290)	.23
Eastern Mediterranean	0.500 (0.152-1.641)	.25
Europe	0.658 (0.428-1.012)	.06
South-East Asia	1.048 (0.532-2.064)	.89
Western Pacific-Asia	0.521 (0.292-0.930)	.03
Western Pacific-Oceania	0.819 (0.337-1.990)	.66

XSL•FO
RenderX

Predictors	OR[a] (95% CI)	*P* value
Telehealth training[e] (none)	1.242 (0.920-1.677)	.16
Telehealth experience[f] (<1 year)	1.043 (0.783-1.389)	.77
Technical difficulties		
Profession[b]		
Psychiatrist	0.923 (0.547-1.558)	.77
Psychologist	1.042 (0.635-1.707)	.87
WHO region[d]		
Africa	0.739 (0.242-2.262)	.60
Americas-South	0.875 (0.461-1.659)	.68
Eastern Mediterranean	0.423 (0.115-1.564)	.20
Europe	0.695 (0.403-1.197)	.19
South-East Asia	0.674 (0.294-1.545)	.35
Western Pacific-Asia	0.362 (0.189-0.696)	.002
Western Pacific-Oceania	0.478 (0.181-1.265)	.14
Telehealth training[e] (none)	0.920 (0.653-1.297)	.64
Telehealth experience[f] (<1 year)	1.556 (1.115-2.173)	.009
Patient consent, privacy, security, and confidentiality		
Profession		
Psychiatrist	0.720 (0.399-1.301)	.28
Psychologist	0.742 (0.430-1.282)	.29
WHO region		
Africa	1.504 (0.446-5.069)	.51
Americas-South	0.199 (0.095-0.415)	<.001
Eastern Mediterranean	0.926 (0.240-3.577)	.91
Europe	0.700 (0.400-1.226)	.21
South-East Asia	3.055 (1.095-8.521)	.03
Western Pacific-Asia	0.681 (0.327-1.417)	.30
Western Pacific-Oceania	0.535 (0.165-1.739)	.30
Telehealth training (none)	0.802 (0.548-1.175)	.26
Telehealth experience (<1 year)	0.745 (0.511-1.085)	.13

[a]OR: odds ratio.
[b]Other professionals as reference.
[c]WHO: World Health Organization.
[d]Americas-North as reference.
[e]Received telehealth training as reference.
[f]More than 1 year of telehealth experience as reference.

For clinical concerns of telehealth, the model investigating endorsement of clinical effectiveness relative to in-person services was statistically significant (χ^2_{11}=21.4; Nagelkerke R^2=0.04; P=.03). A lack of telehealth training and less than 1 year of telehealth experience were significant predictors of the likelihood of clinicians reporting clinical effectiveness of telehealth as a concern. The model predicting concern with remote assessment of patients was statistically significant (χ^2_{11}=71.0; Nagelkerke R^2=0.14; P<.001). The model suggested that psychiatrists, clinicians from South-East Asia and Western-Pacific Asia (compared to clinicians from Americas-North), and clinicians with no telehealth training were more likely to report remote assessment as a telehealth concern. The model predicting concern with lack of private space for

XSL•FO
RenderX

patients was also statistically significant (χ^2_{11}=30.0; Nagelkerke R^2=0.05; P=.002) with psychologists, clinicians from Americas-North (compared to clinicians from Western Pacific-Asia) being more likely to endorse this concern.

Regarding technological concerns, the model predicting endorsement of technical difficulties (eg, audio or video quality and stability of internet connection) was statistically significant (χ^2_{11}=30.0; Nagelkerke R^2=0.05; P=.007). Clinicians with less than 1 year of telehealth experience were more likely to report technical difficulty as a concern. Conversely, compared to participants living in the Americas-North, clinicians in Western-Pacific Asia were less likely to endorse technical difficulties. Finally, the model predicting concern with patient consent, privacy, security, and confidentiality was statistically significant (χ^2_{11}=48.5; Nagelkerke R^2=0.12; P<.001). The WHO region was the only significant predictor. Compared to clinicians living in Americas-North, clinicians in Americas-South were less likely to endorse concern with patient consent, privacy, security, and confidentiality, whereas clinicians in South-East Asia were conversely more likely to report this concern.

Qualitative Results: Benefits of Telehealth

Overview

Three key themes related to the benefits of telehealth were identified: accessibility and reach of mental health services, efficiency and flexibility for clinicians, and enhancement of clinical process and outcomes. In general, clinicians' responses highlighted multiple benefits of telehealth. Codes related to flexibility and convenience, increased access, and time savings were most commonly and consistently cited across the WHO region and profession. However, the context and conditions in which respondents described these codes differed and thus gave rise to several subthemes.

Themes and their respective subthemes are presented in the section below, with representative quotes. Representative responses written in any of the 5 other languages other than English are shown in their original form alongside an English translation. Where descriptive differences by WHO region or profession were observed, these are discussed within corresponding themes. Salient or evocative codes that did not fit the main themes are discussed further below.

Theme 1: Accessibility and Reach of Mental Health Services

This theme was developed primarily from a review of responses where codes related to flexibility and convenience for patients, time savings, and greater access to services were applied. The following 3 subthemes were included:

Reduces Barriers to Care

Clinician responses suggested that telehealth reduces burdens and barriers associated with accessing and receiving in-person mental health services. Many providers indicated that telehealth improved access by reducing geographical and logistical barriers. For example: "Reaching easily patients that live far or in islands where there are no mental health services" (Psychiatrist, Greece).

Others described a reduction in costs and stress that come with traveling to in-person appointments for patients: "Convenience for families not needing to drive/find parking/childcare for siblings etc." (Psychologist, Australia).

Several clinicians described how telehealth increased access for specific populations that typically experience challenges with attending in-person services including older people, individuals with physical and mental comorbidities, and those with mobility difficulties: "It has meant much greater access for people with disabilities and people with more full schedules, which is fantastic" (Psychologist, the United States).

Expands Reach and Cost-Effectiveness of Mental Health Service Delivery

Many clinicians described being able to contribute to the creation of greater access through telehealth. This was indicated on both a provider and a systemic level. For instance: "It has meant I can work with people throughout the state(s) I am licensed in, which creates much more access" (Psychologist, the United States).

Some clinicians reported being able to "see" more patients and having "mas cobertura" ("greater coverage"). Whereas others, such as a psychiatrist in India, cited greater "availability of services in distress, emergency and follow up."

Some also indicated a reduction in overhead expenses. For example, a psychologist in China reported feeling "... less financially burdened for office space." Similarly, a psychologist in Mexico reported, "disminución en costos de alquiler de consultorio" ("Decrease in office rental costs").

Safety of Services Provided

This subtheme came from a code specific to the context of the COVID-19 pandemic, which focused on reduced risk to health. For instance, a psychiatrist in India noted that telehealth, "... reduces risk of exposure to COVID-19 both for the doctor and patient and his/her family."

Theme 2: Efficiency and Flexibility for Clinicians

This theme was developed from codes related to flexibility and convenience for mental health professionals, time savings, provider self-care, and access to other professionals. A review of these texts illustrated the benefits of telehealth that were provider-specific in either a personal or professional capacity. The following 2 subthemes emerged:

Enables Clinicians to Use Their Work Time More Effectively and Flexibly

Many clinicians discussed better time management as a benefit of telehealth. For example, a psychiatrist in Switzerland indicated: "Organisation plus facile de l'agenda." ("Easier to manage my schedule."). Moreover, clinicians described how this flexibility allowed them more time to pursue personal, domestic, or professional activities:

> *He podido aprovechar mejor los espacios intermedios, prácticamente no tengo horarios muertos. (I have been able to take better advantage*

XSL•FO
RenderX

of the time between appointments. I have practically no dead time.) [Psychologist, Mexico]

I can work from home and be available to my partner in a more immediate way, and I can utilize the resources in home more flexibly. I save 6-8 hours a week in commuting time. [Psychologist, the United States]

Seguir trabajando en otros ámbitos. (Continue working in other areas.) [Psychologist, Argentina]

Promotes Provider's Well-Being and Access to Other Professionals

Some providers reported that working from home promoted their capacity for self-care and well-being. For example:

[M]oins d'énergie dépensée en préparatifs et déplacements = moins fatigué. ([L]ess energy spent on preparation and travel = less tired.) [Psychologist, Canada]

Me siento más relajada al poder acceder fácilmente a tomar algún alimento o descansar entre consulta y consulta en un espacio diferente, otro sillón, la cama, el jardín, etc. (I feel more relaxed as I can easily get something to eat or rest between sessions in a different space, another couch, the bed, the garden etc.) [Psychologist, Mexico]

Some clinicians also reported easier access to professional development opportunities (eg, "docencias (teachings)" and "capacitaciones (trainings)"), colleagues, and clinical meetings:

Attending multiagency meetings, which are sometimes many miles from the previous appointment, are now attendable via MS Teams, just a couple of clicks away. [Psychiatrist, the United Kingdom]

Les échanges avec des collègues. (Discussions with colleagues) [Psychiatrist, Senegal]

Theme 3: Enhancement of Clinical Processes and Outcomes

This theme was developed from a review of text where codes related to clinical activities and patient experience (eg, comfort, attendance, and anonymity) were applied. The following 2 subthemes were included:

Improves Clinical and Therapeutic Activities

Some clinicians reported that various clinical processes were improved with telehealth. For example, a psychologist in the United States described "easier access to internet search while in therapy to support therapy." This also included progress tracking, maintaining the alliance, psychoeducation, and follow-ups. Other clinicians described the benefits of having access to patients' homes and family life along with greater family or caregiver involvement:

Useful to see in client's homes and the conditions in which they live. [Social worker, Canada]

Consultation avec des parents séparés qui ne seraient pas venus ensemble en presence. (Consultations with separated parents who would have never attended together in person.) [Psychiatrist, France]

Others reported treatment benefits for specific diagnoses. For example:

Interventions with patients who have Obsessive Compulsive Disorder are far more effective, because patient is in their own home and can practice behavioural exposure during the session. [Psychologist, Canada]

(Easy for autistic patients to understand information and share PC screen.) [Psychiatrist, Japan]

Improves Patient Comfort, Attendance, and Engagement

Several clinicians from nearly all WHO regions indicated that telehealth improved patient attendance. For example, a psychologist in New Zealand noted, "better adherence and fewer did not attend!" Some differences among WHO regions were noted. Many of the responses around patient comfort came from providers in the Americas-North region. Clinicians discussed patient comfort in various ways:

Some patients prefer it, and are more comfortable/relaxed their own space than they would be in a therapy room—so they engage better. [Psychologist, Australia]

Some clients have felt more comfortable disclosing vulnerable material. [Psychologist, the United States]

Nearly all of the responses related to the benefit of patient privacy and anonymity were provided by clinicians in the South-East Asia and Western Pacific-Asia regions and described as an advantage of telehealth:

Some subjects did express that they are happy to have online telehealth services as they do not need to feel noticed by their peers and general public visiting a mental health professional. [Counselor, India]

People can be anonymous for those who wants to keep their identity anonymous. [Psychologist, Philippines]

Other Notable Codes

Disadvantages or No Additional Benefits

Although the framing of the short-answer questions was focused on the benefits of telehealth, several providers (58/496, 11.9% responses) indicated that there were no additional benefits aside from remote consultations or exclusively reported disadvantages of telehealth. The vast majority of responses to which this code was applied came from clinicians in Europe, Americas-South, and Western Pacific-Asia. Most of these responses were brief (eg, "no" and "none"), while only a few responses were more descriptive:

No. The Covid experience has confirmed to me that I don't think it's a useful longterm alternative to in person therapy. [Psychologist, New Zealand]

The only benefit is that the service can be provided despite COVID, but I believe we lose out on many of

XSL•FO
RenderX

the small, but important nuances of in-person therapy when limited to on-line intervention. [Psychologist, Canada]

Positive Disruption of Mental Health Profession

A few clinicians, particularly providers in the Americas-North region, described a systemic level benefit, suggesting that the telehealth shift had challenged and advanced the mental health field and profession by altering service delivery, lifting licensing restrictions, and "kickstarting" new areas:

Offre aux patients un modèle de souplesse et de flexibilité; permet d'approfondir les enjeux autour des limites et des frontière.(Offers patients a model of flexibility; allows for the exploration of issues around limits and boundaries.) [Psychologist, Canada]

In South Africa, remote consulting was not permitted by the medical regulatory body. We hope that this will change in future, so that psychiatric services can be made accessible to underserved areas where there are no psychiatrists. This would benefit particularly poor and disadvantaged areas. This can also be a way of upskilling primary care providers in these areas. [Psychiatrist, South Africa]

Comparing Quantitative and Qualitative Results

A joint display of the 2 forms of data offered a complimentary perspective on the concerns and benefits of telehealth according to clinicians [32]. Table 6 presents these contradictory findings as thematic intersections.

XSL•FO
RenderX

Table 6. Integrating qualitative and quantitative findings from mental health professionals to generate complementary insights.

	Benefits of telehealth (qualitative)	Concerns with telehealth (quantitative)	Provider- and system-level considerations when evaluating the trade-off
Thematic intersection 1	• Improves clinical and therapeutic activities (eg, follow-ups, greater family involvement, more insight into patients' home and family life, easier use of online tools to support therapy, and diagnosis-specific treatment benefits; 52/496, 10.5% responses)	• Clinical effectiveness compared to in-person (506/806, 62.8%), especially, • Loss of clinical information (396/500, 79.2%) • Building and maintaining the therapeutic alliance (351/500, 70.2%) • Showing empathy (237/500, 47.4%) • Implementing therapeutic interventions (208/500, 41.6%)	• Patient characteristics including disorder or diagnosis, age group, and treatment modality (individual vs group or family or couple) • Provider characteristics including telehealth training and experience; telehealth strategies tailored to meet discipline-specific clinical needs • Evaluating the loss versus the gain of clinical information • Would telehealth enhance or impede therapeutic interventions? • Role of flexible blended models that incorporate telehealth and in-person
Thematic intersection 2	• Expands reach of mental health services (64/496, 12.9% responses) • Timely access to emergency services (6/496, 1.2% responses)	• Assessing high-risk patients and managing emergencies remotely (460/806, 57.1%) • Remote assessment of patients (358/806, 44.4%)	• Patient characteristics including reliable access to technology and internet; technological literacy • Provider characteristics including telehealth training and experience • Regional characteristics: is this an area or population with no access to mental health services? • Are there discipline-specific clinical needs such as infrastructure for physical examination? • Opportunity for telehealth models to work closely with providers from remote regions for in-person care and capacity building?
Thematic intersection 3	• Improves patient attendance (40/496, 8.1% responses) • Improves patient comfort (eg, more at ease in their own homes and more disclosures; 25/496, 5% responses)	• Decreased patient engagement compared to in-person (340/806, 42.2%) • Lack of private space for patients during telehealth sessions (387/806, 48%)	• Patient characteristics including disorder or diagnosis, age group, treatment modality, preferences, and living circumstances • Will the lack of privacy in the home environment be an issue? • Regional characteristics: could the privacy or anonymity of telehealth help reduce stigma challenges about mental illness and seeking care?

Discussion

Overview

This study used quantitative and qualitative methodologies to comprehensively examine and expand upon our understanding of the perspectives, benefits, and concerns related to telehealth service provision among a large, international, and multilingual sample of mental health professionals during the COVID-19 pandemic (November-December 2020). Participants were mainly psychologists and psychiatrists working in mental health settings and providing mental health services such as diagnostic

XSL•FO
RenderX

assessment or psychotherapy via telehealth. Most psychiatrists had also provided psychopharmacotherapy services. Half of participating clinicians had less than a year of telehealth experience, suggesting that they likely adopted telehealth because of the pandemic. A little over half of the participants had received some form of telehealth training. Participants were mostly using videoconferencing and/or telephone. This study offers a global snapshot of telehealth implementation from the perspective of clinicians with both previous telehealth experience and new adopters. Findings suggest several important conclusions that can inform recommendations for training and practice for telehealth in mental health care.

Clinicians Are Primarily Concerned About Clinical Effectiveness and Technological Challenges

Consistent with previous research, clinical effectiveness relative to in-person services was a top concern among clinicians [2,18,38]. However, this study further highlights that clinicians' primary concerns about clinical effectiveness were regarding the loss of clinical information (eg, nonverbal behavior) and challenges in building and maintaining a therapeutic alliance. Assessment of high-risk patients and managing emergencies remotely was another common clinical concern. These findings support previous research that indicates that major limitations of telehealth practice are reduced availability of clinical data to support assessment or diagnosis [26] and greater difficulty picking up on nonverbal cues [1]. Diminished sense of connection and therapeutic alliance [18,19] has also been raised as a notable challenge associated with telehealth practice, despite this not being fully reflected among findings in the telehealth efficacy literature [39].

With respect to technological concerns, technical issues (eg, audio or video quality and stability of internet connection) were the top concerns followed by patient-related factors such as patients' access to equipment and familiarity with connecting. The COVID-19 pandemic compelled clinicians to adapt to technology and the technical issues inherent to telehealth. Current best-practice guidelines suggest that proactively establishing backup means of communication with clients can help mitigate these challenges [10]. However, the pandemic has also exposed disparities in access to mental health care among vulnerable populations across low-income and middle-income countries (LMICs) as well as within high-income countries [6]. Hence, client-specific challenges such as technology access, reliability, and competency remain significant barriers to telehealth.

Overall, these top concerns are broadly consistent with the literature [1,18]. However, the current findings build on previous research by examining how demographic and professional characteristics influence the endorsement of common concerns and perceptions, offering further context and insight. For instance, a lack of telehealth training was associated with increased endorsement of common concerns such as clinical effectiveness of telehealth relative to in-person services and remote assessment of patients. Likewise, those with minimal experience using telehealth (ie, less than 1 year) were more likely to report technical difficulties and clinical effectiveness of telehealth relative to in-person services as a concern. These findings support the literature, suggesting that experience and training are related to provider perceptions of telehealth [2,13,40].

Telehealth Offers Valuable Benefits for Patients and Clinicians That Are Consistent Across Global Contexts

Qualitative data revealed numerous advantages to virtual mental health care. These benefits fell into 3 major areas: accessibility and reach of mental health services, efficiency and flexibility for clinicians, and enhancement of clinical process and outcomes. These main benefits broadly align with existing studies conducted across LMICs and higher-income countries [1,17,18,21,26,41]. However, this international sample extends current research by highlighting the significant advantages of telehealth cited by clinicians from nearly all WHO regions. Specifically, remote mental health care is convenient and flexible for patients because it reduces burdens and barriers to accessing care. Clinicians' responses illustrated the various economical, geographical, and logistical barriers that clients regularly encounter when accessing in-person mental health care. Moreover, telehealth improved patient access by enabling clinicians to see more patients and extend specialized services to a broader population. Thus, continued implementation of telehealth services alongside in-person care can help ensure equitable mental health access.

Based on participants' perceptions, telehealth also offers convenience and flexibility for mental health professionals by giving them greater control over their time and clinical outreach. Previous studies have spoken about the personal benefits of telehealth for mental health professionals including more efficient use of time [18] and improved work-life balance [22]. This study further illustrates how telehealth can support clinician efficiency (eg, no commuting), well-being and quality of life (eg, working in a relaxing home environment), and professional development. As such, telehealth offers valuable benefits for mental health professionals, particularly for clinicians at life stages where flexibility and work-life balance are important. This is likewise relevant to occupational stress and burnout, as the COVID-19 pandemic underscored the need to safeguard the capacity of the global mental health workforce [31]. Health care systems and clinicians should leverage the above benefits to promote provider efficiency and health, such as policies supporting hybrid work models, while maintaining quality care.

Increased safety from COVID-19 and improved patient attendance were among the most prominently reported benefits among participants. The latter is consistent with the literature [21,22]. The reduction of barriers to access likely facilitates client attendance. Altogether, qualitative findings underscore why telehealth has become a routine practice for many clinicians in a postpandemic era.

Psychologists and Psychiatrists Engage Differently With Telehealth

Our quantitative data also revealed informative data regarding professional differences between psychologists' and psychiatrists' use (Table 2) and perceptions of telehealth. Findings indicated that psychologists spent significantly more hours delivering clinical services via videoconferencing and

XSL•FO
RenderX

telephone than psychiatrists. Conversely, there were no significant differences in the number of patients seen between the 2 disciplines across telehealth modalities. With respect to endorsement of telehealth concerns, psychiatrists were more likely to report challenges with remote assessment of patients, assessing high-risk patients and managing emergencies, showing empathy, and licensing or credentialing requirements compared to psychologists. Conversely, psychologists were more likely to be concerned about a lack of private space for patients during telehealth sessions compared to psychiatrists.

These findings likely reflect differences in the clinical services each discipline provides. Psychiatrists' concerns emphasize their need for additional infrastructural support for adapting in-person interventions like physical examinations or blood pressure monitoring [26]. Moreover, participating psychiatrists provided other clinical activities such as medication management or general medical services that are briefer in nature compared to assessment or psychotherapy services. Likewise, psychologists, who spend more direct time with patients (eg, providing psychotherapy), may prioritize ensuring that patients have a private space conducive to treatment. Altogether, these findings stress the need for telehealth strategies that are tailored to the needs of psychologists and psychiatrists. This could entail specialized telehealth training, telehealth models that integrate in-person and remote care, and improved regulatory frameworks to support remote psychiatric care. Likewise, platforms that optimize usability to accommodate different clinical needs and policies that focus on improving the standards and accessibility of telehealth may be beneficial.

Telehealth Service Delivery Involves Balancing Trade-Offs

An integration of quantitative and qualitative findings (Table 6) illustrated that there was much overlap among concerns and benefits of telehealth. For instance, some clinicians noted that telehealth provided valuable clinical insights into patients' home and family environments, while others endorsed concerns about a loss of clinical information, such as nonverbal cues. Likewise, providers observed improved attendance, engagement, and continuity of care. However, clinicians also worried about the therapeutic alliance and patient engagement. This overlap in perspectives has been reported in the literature [22,25,40]. Data from the international sample of mental health professionals presented here offer a nuanced perspective: there are inherent trade-offs between telehealth's benefits and disadvantages.

It is important for health care systems and stakeholders to be aware of the potential advantages and disadvantages of telehealth practice when planning mental health service. Flexible telehealth models can harness advantages while addressing concerns about clinical effectiveness and technology through hybrid or capacity-building initiatives that integrate available in-person support within patient communities (eg, medical personnel for physical examinations, psychometrist for cognitive assessments, and staff for crisis management).

At the provider level, clinicians must evaluate telehealth's trade-offs to determine its suitability for each patient. Will telehealth enhance or limit therapeutic outcomes? The available literature recommends evaluating provider competencies, patient needs and preferences, and organizational resources when determining the appropriateness of telehealth among other considerations [42]. For instance, this study found that perceived effectiveness of telehealth was higher for individual assessment and treatment as well as adult-aged patients. Likewise, service modality, particularly family or group therapy, or initial assessment may impact rapport building [39]. Research also suggests that having the initial patient interaction be in-person can help establish rapport prior to telehealth services [25,43]. Finally, adapting clinician communication style to include greater use of vocal tone and facial expression can help convey empathy [43].

Differences Among WHO Regions

Quantitative findings identified that participants from South-East Asia and Western Pacific-Asia were more likely to endorse concerns related to remote assessment and patient privacy, security, and confidentiality when compared to participants in the Americas-North. Conversely, participants from the Americas-South and Western Pacific-Asia were less likely to report concerns about patient privacy, security, and confidentiality, and technical difficulties, respectively. Qualitative findings revealed that clinicians in Europe, Americas-South, and Western-Pacific Asia reported disadvantages to telehealth (eg, inferiority to in-person care and difficulty obtaining payment for services) or simply indicated no additional benefits aside from remote contact. Reasons for variations in participant impressions across WHO regions are likely multifaceted; however, a few factors can be considered. First, the pandemic necessitated an immediate, global shift to telehealth with almost complete virtualization of many services at the time of data collection [7], meaning that many clinicians were faced with adopting unfamiliar telehealth platforms. As such, WHO regions varied in readiness for such changes in practice and training. For instance, in a survey study with medical students, trainee and early career psychiatrists from South-East Asia and Western-Pacific Asia regions conducted in early 2021 found limited opportunities for training and knowledge of telehealth among young professionals [41]. Some regions also faced key barriers to uptake including reservations about telehealth among clinicians, inadequate communication infrastructure (eg, in some parts of Latin America), and lack of regulatory frameworks (eg, professional licensing) in some parts of Europe [44-46].

Qualitative data also found that privacy and anonymity as benefits of telehealth were primarily noted by participants in the South-East Asia and Western Pacific-Asia regions—advantages that have been less addressed in the literature. Challenges related to stigma around mental illness in these regions may explain this heightened benefit among its mental health professionals [47]. A few participants, particularly those in the Americas-North, believed that the abrupt shift to telehealth had advanced the field of mental health. Quantitative data identified that clinicians in the Americas-North were more likely to be concerned about a lack of private space for patients during telehealth sessions compared to clinicians from Western Pacific-Asia. These findings may reflect how certain regions, such as North America—where much telehealth research has

XSL•FO
RenderX

been conducted [23,26], had more familiarity with the implementation of virtual models.

Countries and regions must now reflect on the future direction of telehealth. This study suggests that to capitalize on telehealth's broad advantages, investing in telehealth training and education and the dissemination of resources for mental health professionals are crucial. Countries should also address disparities in access to digital technology, known as the "digital divide," that affects LMICs and high-income countries. Regional telehealth models could promote digital literacy among specific patient groups, assess the use of tools that require less bandwidth (eg, telephone and instant messaging), or offer stability and privacy for telehealth services through local health centers or schools.

Limitations

This study has some limitations. Study findings were based on provider perspectives rather than objective measures of quality of mental health care or ethical practices. It will be important to further examine the patient perspective and assess the quality of services delivered. Given that data were gathered during the early stages of the pandemic, clinicians' perceptions and responses were likely impacted by their evolving contexts at the time. Furthermore, the quality of telehealth training indicated by clinicians was unknown. Since participant perceptions were based on multiple telehealth modalities (synchronous and asynchronous) and clinical contexts, direct comparisons (eg, telephone vs videoconferencing) were not possible. As such, present findings remain general and based on clinicians' own combination of telehealth service modalities provided in a public health emergency. Although this study had representation from all WHO regions and country income levels, most participating clinicians were from high- and upper-middle–income countries. Finally, given the cross-language qualitative data, it is possible that translation decisions and coders being outside of the sociocultural environments of many participant clinicians (coders were based in North America) impacted interpretation. Although the number of short-answer responses in Japanese or Chinese was small (roughly 44/496, 9%), the DeepL translations were not reviewed by fluent speakers, which may limit interpretive accuracy. Future research should explore implementation and patient outcomes within specific telehealth modalities, clinical settings, and geographical contexts, particularly its potential in LMICs.

Conclusions

The results of this study suggest that telehealth for mental health services presents notable clinical and technological challenges that are common across global contexts. Certain competencies such as telehealth training and experience may help mitigate these concerns. Moreover, given that psychologists and psychiatrists seem to engage with telehealth differently, specialized training and tailored approaches may also help implementation. Nonetheless, results also suggest that telehealth offers valuable benefits for patients and providers. Blended models that integrate virtual and in-person care or capacity-building platforms for underserved areas are worth exploring. In a postpandemic era with more deliberate implementation of telehealth services and a new generation of digitally trained clinicians, we can balance its benefits and limitations to optimize mental health care and clinical outcomes globally.

Acknowledgments

The authors would like to thank Konrad Czechowski from the School of Psychology at the University of Ottawa for support with qualitative data analyses and reporting.

Funding

This study was funded by the Canadian Institutes of Health Research (grant VR4-172750).

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the nature of our original ethics approval. Anonymized, group-level data are available from the corresponding author on reasonable request.

Authors' Contributions

CSK acquired funding. All authors contributed to conceptualization and methodology including study design and survey. GMR and TJR contributed to project administration. CSK contributed to supervision. JAG-P contributed to initial data curation. MIM conducted formal analysis and analyzed quantitative data. MIM and CSK analyzed the qualitative data. MIM wrote the first draft with critical input from CSK. BK, GMR, JAG-P, MK, RR, and TJR helped with reviewing and editing the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
GRAMMS checklist.
[DOCX File , 19 KB - jmir_v28i1e75905_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Survey questions.
[DOCX File , 24 KB - jmir_v28i1e75905_app2.docx]

References

1. Appleton R, Williams J, Vera San Juan N, Needle JJ, Schlief M, Jordan H, et al. Implementation, adoption, and perceptions of telemental health during the COVID-19 pandemic: systematic review. J Med Internet Res 2021;23(12):e31746. [doi: 10.2196/31746] [Medline: 34709179]
2. Montoya MI, Kogan CS, Rebello TJ, Sadowska K, Garcia-Pacheco JA, Khoury B, et al. An international survey examining the impact of the COVID-19 pandemic on telehealth use among mental health professionals. J Psychiatr Res 2022;148:188-196 [FREE Full text] [doi: 10.1016/j.jpsychires.2022.01.050] [Medline: 35131587]
3. Global diffusion of eHealth: making universal health coverage achievable: report of the third global survey on eHealth. World Health Organization. 2016. URL: https://www.who.int/publications/i/item/9789241511780 [accessed 2025-02-23]
4. Talevi D, Socci V, Carai M, Carnaghi G, Faleri S, Trebbi E, et al. Mental health outcomes of the CoViD-19 pandemic. Riv Psichiatr 2020;55(3):137-144. [doi: 10.1708/3382.33569] [Medline: 32489190]
5. Xiang YT, Yang Y, Li W, Zhang L, Zhang Q, Cheung T, et al. Timely mental health care for the 2019 novel coronavirus outbreak is urgently needed. Lancet Psychiatry 2020;7(3):228-229 [FREE Full text] [doi: 10.1016/S2215-0366(20)30046-8] [Medline: 32032543]
6. Kola L, Kohrt BA, Hanlon C, Naslund JA, Sikander S, Balaji M, et al. COVID-19 mental health impact and responses in low-income and middle-income countries: reimagining global mental health. Lancet Psychiatry 2021;8(6):535-550 [FREE Full text] [doi: 10.1016/S2215-0366(21)00025-0] [Medline: 33639109]
7. Zangani C, Ostinelli EG, Smith KA, Hong JSW, Macdonald O, Reen G, et al. Impact of the COVID-19 pandemic on the global delivery of mental health services and telemental health: systematic review. JMIR Ment Health 2022;9(8):e38600 [FREE Full text] [doi: 10.2196/38600] [Medline: 35994310]
8. Pandya A, Waller M, Portnoy JM. The regulatory environment of telemedicine after COVID-19. J Allergy Clin Immunol Pract 2022;10(10):2500-2505 [FREE Full text] [doi: 10.1016/j.jaip.2022.07.001] [Medline: 35842133]
9. Interim ethical guidelines for psychologists providing psychological services via electronic media. Canadian Psychological Association. 2020. URL: https://cpa.ca/docs/File/Ethics/CPAe-therapyGuidelinesUpdate2020.pdf [accessed 2026-02-05]
10. CPA guidelines on telepsychology. Canadian Psychological Association. 2023. URL: https://cpa.ca/docs/File/Practice/Telepsychology_Guidelines-September_2023-EN_FINAL.pdf [accessed 2026-02-05]
11. Cantor JH, McBain RK, Ho P, Bravata DM, Whaley C. Telehealth and in-person mental health service utilization and spending, 2019 to 2022. JAMA Health Forum 2023;4(8):e232645 [FREE Full text] [doi: 10.1001/jamahealthforum.2023.2645] [Medline: 37624614]
12. Page C, Stamm K. Telehealth and hybrid practice are here to stay. American Psychological Association. 2024. URL: https://www.apa.org/monitor/2024/09/telehealth-hybrid-practice [accessed 2025-11-17]
13. Connolly SL, Miller CJ, Lindsay JA, Bauer MS. A systematic review of providers' attitudes toward telemental health via videoconferencing. Clin Psychol (New York) 2020;27(2):10.1111/cpsp.12311 [FREE Full text] [doi: 10.1111/cpsp.12311] [Medline: 35966216]
14. Hubley S, Lynch SB, Schneck C, Thomas M, Shore J. Review of key telepsychiatry outcomes. World J Psychiatry 2016;6(2):269-282 [FREE Full text] [doi: 10.5498/wjp.v6.i2.269] [Medline: 27354970]
15. Sugarman DE, Busch AB. Telemental health for clinical assessment and treatment. BMJ 2023;380:e072398. [doi: 10.1136/bmj-2022-072398] [Medline: 36646462]
16. Varker T, Brand RM, Ward J, Terhaag S, Phelps A. Efficacy of synchronous telepsychology interventions for people with anxiety, depression, posttraumatic stress disorder, and adjustment disorder: a rapid evidence assessment. Psychol Serv 2019;16(4):621-635. [doi: 10.1037/ser0000239] [Medline: 29809025]
17. Liberati E, Richards N, Parker J, Willars J, Scott D, Boydell N, et al. Remote care for mental health: qualitative study with service users, carers and staff during the COVID-19 pandemic. BMJ Open 2021;11(4):e049210 [FREE Full text] [doi: 10.1136/bmjopen-2021-049210] [Medline: 33888531]
18. Watson JD, Pierce BS, Tyler CM, Donovan EK, Merced K, Mallon M, et al. Barriers and facilitators to psychologists' telepsychology uptake during the beginning of the COVID-19 pandemic. Int J Environ Res Public Health 2023;20(8):5467 [FREE Full text] [doi: 10.3390/ijerph20085467] [Medline: 37107748]
19. Jaywant A, Aulitzky W, Avari J, Buchheim A, Dubin M, Galffy M, et al. Multinational perspectives on changes to psychiatric care during the COVID-19 pandemic: a survey of practicing psychiatrists. Neuropsychiatr 2023;37(3):115-121. [doi: 10.1007/s40211-022-00452-x] [Medline: 36600105]
20. Shalev L, Eitan R, Rose AJ. The use of telepsychiatry services in emergency settings: scoping review. J Med Internet Res 2024;26:e51814 [FREE Full text] [doi: 10.2196/51814] [Medline: 39008831]
21. Guinart D, Marcy P, Hauser M, Dwyer M, Kane JM. Mental health care providers' attitudes toward telepsychiatry: a systemwide, multisite survey during the COVID-19 pandemic. Psychiatr Serv 2021;72(6):704-707. [doi: 10.1176/appi.ps.202000441] [Medline: 33593104]

XSL•FO
RenderX

22. Lipschitz JM, Connolly SL, Van Boxtel R, Potter JR, Nixon N, Bidargaddi N. Provider perspectives on telemental health implementation: lessons learned during the COVID-19 pandemic and paths forward. Psychol Serv 2023;20(Suppl 2):11-19 [FREE Full text] [doi: 10.1037/ser0000625] [Medline: 35201809]
23. Cowan KE, McKean AJ, Gentry MT, Hilty DM. Barriers to use of telepsychiatry: clinicians as gatekeepers. Mayo Clin Proc 2019;94(12):2510-2523. [doi: 10.1016/j.mayocp.2019.04.018] [Medline: 31806104]
24. Farrer LM, Clough B, Bekker MJ, Calear AL, Werner-Seidler A, Newby JM, et al. Telehealth use by mental health professionals during COVID-19. Aust N Z J Psychiatry 2023;57(2):230-240 [FREE Full text] [doi: 10.1177/00048674221089229] [Medline: 35360958]
25. Galvin E, Desselle S, Gavin B, Quigley E, Flear M, Kilbride K, et al. Patient and provider perspectives of the implementation of remote consultations for community-dwelling people with mental health conditions: a systematic mixed studies review. J Psychiatr Res 2022;156:668-678 [FREE Full text] [doi: 10.1016/j.jpsychires.2022.10.051] [Medline: 36399859]
26. Uscher-Pines L, Sousa J, Raja P, Mehrotra A, Barnett ML, Huskamp HA. Suddenly becoming a "Virtual Doctor": experiences of psychiatrists transitioning to telemedicine during the COVID-19 pandemic. Psychiatr Serv 2020;71(11):1143-1150 [FREE Full text] [doi: 10.1176/appi.ps.202000250] [Medline: 32933411]
27. Lin T, Stone SJ, Heckman TG, Anderson T. Zoom-in to zone-out: Therapists report less therapeutic skill in telepsychology versus face-to-face therapy during the COVID-19 pandemic. Psychotherapy (Chic) 2021;58(4):449-459. [doi: 10.1037/pst0000398] [Medline: 34881922]
28. Schoonenboom J, Johnson RB. How to construct a mixed methods research design. Kolner Z Soz Sozpsychol 2017;69(Suppl 2):107-131 [FREE Full text] [doi: 10.1007/s11577-017-0454-1] [Medline: 28989188]
29. Miles MB, Huberman AM, Saldana J. Qualitative Data Analysis: A Methods Sourcebook. 3rd Edition. Thousand Oaks, CA: SAGE Publications; 2014.
30. Reed GM, Rebello TJ, Pike KM, Medina-Mora ME, Gureje O, Zhao M, et al. WHO's Global Clinical Practice Network for mental health. Lancet Psychiatry 2015;2(5):379-380. [doi: 10.1016/S2215-0366(15)00183-2] [Medline: 26360271]
31. Kogan CS, Garcia-Pacheco JA, Rebello TJ, Montoya MI, Robles R, Khoury B, et al. Longitudinal impact of the COVID-19 pandemic on stress and occupational well-being of mental health professionals: an international study. Int J Neuropsychopharmacol 2023;26(10):747-760 [FREE Full text] [doi: 10.1093/ijnp/pyad046] [Medline: 37531283]
32. Creswell JW. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. 4th Edition. Los Angeles, CA: SAGE; 2014.
33. O'Cathain A, Murphy E, Nicholl J. The quality of mixed methods studies in health services research. J Health Serv Res Policy 2008;13(2):92-98. [doi: 10.1258/jhsrp.2007.007074] [Medline: 18416914]
34. Elliott R, Bohart AC, Watson JC, Murphy D. Therapist empathy and client outcome: an updated meta-analysis. Psychotherapy (Chic) 2018;55(4):399-410 [FREE Full text] [doi: 10.1037/pst0000175] [Medline: 30335453]
35. Wampold BE. How important are the common factors in psychotherapy? An update. World Psychiatry 2015;14(3):270-277 [FREE Full text] [doi: 10.1002/wps.20238] [Medline: 26407772]
36. Saldaña J. The Coding Manual for Qualitative Researchers. Thousand Oaks, CA: Sage Publications Ltd; 2009:xi-223.
37. Maxwell JA. Using numbers in qualitative research. Qual Inq 2010;16(6):475-482. [doi: 10.1177/1077800410364740]
38. Perle JG, Burt J, Higgins WJ. Psychologist and physician interest in telehealth training and referral for mental health services: an exploratory study. J Technol Hum Serv 2014;32(3):158-185. [doi: 10.1080/15228835.2014.894488]
39. Sugarman DE, Horvitz LE, Greenfield SF, Busch AB. Clinicians' perceptions of rapid scale-up of telehealth services in outpatient mental health treatment. Telemed J E Health 2021;27(12):1399-1408 [FREE Full text] [doi: 10.1089/tmj.2020.0481] [Medline: 33600272]
40. Lawson JL, Doran JM, O'Shea M, Abel EA. The good, the bad, the uncertain: diverse provider experiences with telemental health during COVID-19. Psychiatr Q 2022;93(3):753-774 [FREE Full text] [doi: 10.1007/s11126-022-09990-7] [Medline: 35689752]
41. Orsolini L, Jatchavala C, Noor IM, Ransing R, Satake Y, Shoib S, et al. Training and education in digital psychiatry: a perspective from Asia-Pacific region. Asia Pac Psychiatry 2021;13(4):e12501 [FREE Full text] [doi: 10.1111/appy.12501] [Medline: 34873845]
42. Luxton DD, Nelson EL, Maheu M. A Practitioner's Guide to Telemental Health: How to Conduct Legal, Ethical, and Evidence-Based Telepractice. 1st Edition. Washington, DC: American Psychological Association; 2023.
43. Powell M. Empathy in digital healthcare [doctoral dissertation]. University of Central Florida. 2024. URL: https://stars.library.ucf.edu/etd2024/60 [accessed 2026-02-05]
44. Castaño Pérez GA, Rojas Bernal LA, Maldonado MA. Telemedicine for mental health in Latin America. In: Mental Health & Addiction: An Integrated Care in Latin America. Canada: CDS Press; 2023:26-56.
45. Haddouk L, Schneider B, Daele TV, Lang F. Telepsychology in France following the COVID-19 crisis. Eur J Psychol Open 2025;84(1):13-19 [FREE Full text] [doi: 10.1024/2673-8627/a000063]
46. Omboni S, Padwal RS, Alessa T, Benczúr B, Green BB, Hubbard I, et al. The worldwide impact of telemedicine during COVID-19: current evidence and recommendations for the future. Connect Health 2022;1:7-35 [FREE Full text] [doi: 10.20517/ch.2021.03] [Medline: 35233563]

XSL•FO
RenderX

47. Singh S, Sagar R. Online psychotherapy during the COVID-19 pandemic: the good, the bad, and the ugly. Indian J Psychol Med 2022;44(2):177-180 [FREE Full text] [doi: 10.1177/02537176211070427] [Medline: 35655984]

Abbreviations

GCPN: Global Clinical Practice Network
GRAMMS: Good Reporting of a Mixed Methods Study
ICD-11: International Classification of Diseases, 11th Revision
LMIC: low-income and middle-income country
WHO: World Health Organization

Edited by A Stone; submitted 14.Apr.2025; peer-reviewed by NG Badr, C Babbage; comments to author 06.Nov.2025; revised version received 08.Jan.2026; accepted 30.Jan.2026; published 04.Mar.2026.

Please cite as:
Montoya MI, Kogan CS, Khoury B, García-Pacheco JA, Robles R, Rebello TJ, Kulygina M, Reed GM
Mental Health Professionals' Perceptions of Benefits and Disadvantages of Telehealth: International Mixed Methods Study
J Med Internet Res 2026;28:e75905
URL: https://www.jmir.org/2026/1/e75905
doi:10.2196/75905
PMID:

©Madeline I Montoya, Cary S Kogan, Brigitte Khoury, José A García-Pacheco, Rebeca Robles, Tahilia J Rebello, Maya Kulygina, Geoffrey M Reed. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Artificial Intelligence Centrality in Psychotic Delusions and Violence Risk in Forensic Psychiatry: Retrospective Observational Study of Judicial Decisions

Alexandre Hudon[1,2,3,4,5], BEng, MD, PhD; Chanel Pagé[1], BSc, MD

[1]Department of Psychiatry and Addictology, Faculty of Medicine, Université de Montréal, Montréal, QC, Canada
[2]Forensic Psychiatry Unit, Department of Psychiatry, Institut universitaire en santé mentale de Montréal, Montreal, QC, Canada
[3]Centre de recherche de l'Institut universitaire en santé mentale de Montréal, Montreal, QC, Canada
[4]Department of Psychiatry, Institut national de psychiatrie légale Philippe-Pinel, Montreal, QC, Canada
[5]Centre de pédagogie en sciences de la santé, Université de Montreal, Montreal, QC, Canada

Corresponding Author:
Alexandre Hudon, BEng, MD, PhD
Department of Psychiatry and Addictology
Faculty of Medicine
Université de Montréal
2900 Bd Édouard-Montpetit
Montréal, QC, H3T 1J4
Canada
Phone: 1 514 251 4000
Email: alexandre.hudon.1@umontreal.ca

Abstract

Background: Artificial intelligence (AI)–themed delusions are increasingly observed in psychotic-spectrum disorders, reflecting the incorporation of contemporary sociotechnical elements into delusional systems. However, it remains unclear whether the structural role of AI within these belief systems is associated with increased violence risk or more restrictive forensic outcomes. Given the importance of dynamic clinical factors (eg, insight and treatment adherence) in forensic risk assessment, clarifying the role of AI centrality has clinical and legal relevance.

Objective: This study examined whether AI centrality within psychotic delusional systems is associated with (1) violence toward others and (2) judicial findings of significant public safety risk in forensic psychiatric decisions.

Methods: This retrospective observational study used jurisprudential data from the Société québécoise d'information juridique database, including all publicly available Quebec tribunal and court decisions up to December 31, 2025. Eligible cases (N=29) involved psychotic-spectrum disorders with explicit AI-related delusional content and judicial consideration of dangerousness or disposition. The unit of analysis was the judicial decision. AI centrality was coded as central (n=15, 51.7%) or noncentral (n=14, 48.3%) using a structured, text-based framework. The primary outcome was documented violence toward others; secondary outcomes included direct AI-linked violence attribution and judicial findings of significant public safety risk. Covariates included impaired insight, treatment nonadherence, substance use disorder, and prior violence history. Data were extracted through full-text review using a standardized coding grid. Bivariate associations were analyzed using Fisher exact tests (α=.05), and odds ratios (ORs) with 95% CIs were calculated. Exploratory logistic regression models were performed to assess adjusted associations.

Results: Violence toward others was documented in 20/29 (69%) cases. AI centrality was not significantly associated with violence (12/15, 80.0%, vs 8/14, 57.1%; OR 2.91, 95% CI 0.63-13.45; P=.26) but was strongly associated with direct AI-linked violence attribution (9/15, 60.0%, vs 2/14, 14.3%; OR 9.00, 95% CI 1.48-54.6; P=.01). Judicial findings of significant public safety risk were more frequent in AI-central cases (13/15, 86.7%, vs 9/14, 64.3%; OR 3.60, 95% CI 0.63-20.5; P=.24), although not statistically significant. AI-central cases demonstrated higher prevalence of impaired insight (13/15, 86.7%, vs 8/14, 57.1%; OR 4.89, 95% CI 0.79-30.1) and treatment nonadherence (9/15, 60.0%, vs 4/14, 28.6%; OR 3.75, 95% CI 0.74-18.9).

Conclusions: AI centrality within delusional systems appears to be not independently associated with increased violence toward others but is strongly associated with AI-based attribution of behavior and markers of epistemic vulnerability, including impaired insight and treatment nonadherence. The findings suggest that AI-themed delusions function as structural organizers of meaning and agency rather than novel criminogenic risk factors. Clinically and legally, this underscores the importance of prioritizing

XSL•FO
RenderX

dynamic risk variables over thematic novelty, informing more proportionate forensic decision-making and risk assessment in an era of rapidly evolving digital environments.

(J Med Internet Res 2026;28:e93349) doi:10.2196/93349

KEYWORDS

artificial intelligence; AI; psychotic disorders; delusions; violence risk; forensic psychiatry; risk assessment; schizophrenia; insight; treatment adherence; dangerousness

Introduction

AI-Themed Delusions and Contemporary Techno-Paranoia

Delusional content is not static; it reliably absorbs salient features of the sociotechnical environment, so "modern" threat agents (networks, devices, platforms) increasingly populate persecutory and passivity narratives. Evidence from a contemporary psychosis cohort suggests that technology-related delusions are now common and may be rising over time, underscoring the clinical need to ask explicitly about internet and device-mediated beliefs during assessment and treatment planning [1]. Parallel work in nonclinical populations supports a continuum model in which technology-focused fears range from realistic privacy concerns to implausible, self-referential, and frankly paranoid interpretations (captured empirically by the construct of "cyber-paranoia") [2]. In addition, systematic review evidence on social media and psychosis-spectrum "social brain" disorders synthesizes case-based and observational findings suggesting that algorithmic curation, ambiguity of online cues, and perceived surveillance can intensify suspiciousness and "ideas of reference" in vulnerable individuals, potentially strengthening delusional conviction through self-reinforcing digital "evidence" [3]. Against this backdrop, generative artificial intelligence (AI) introduces a qualitatively different affordance: continuous, interactive dialogue that can mirror affect, provide confirmatory narratives, and blur boundaries between inner experience and external "agentic" communication. Conceptual analyses have framed "AI psychosis" not as a new diagnosis but as a heuristic to study how prolonged human-AI interaction might trigger, reshape, or entrench delusional systems in predisposed users [4]. Recent open access commentary further argues that the principle (incorporating media or technology into delusions) is not new, but interactivity and "sycophantic" agreement may alter the risk profile by colluding with delusional elaboration rather than supporting reality testing [5,6]. These developments motivate forensic inquiry: when AI becomes the perceived persecutor, controller, or corroborating interlocutor, does tribunals' reasoning about risk and disposition shift accordingly?

Forensic Risk Factors, Insight, and Treatment Adherence in Psychosis

Forensic psychiatry distinguishes the presence of psychosis from the conditions under which risk escalates, emphasizing dynamic (modifiable) factors, such as active positive symptoms, substance use, insight, and adherence. Meta-analytic evidence indicates that schizophrenia and related psychoses are associated with elevated violence risk, but much of the excess risk appears mediated by comorbid substance use disorders. This highlights the importance of integrated substance use assessment and treatment in risk reduction [7]. Complementing this, a large systematic review and meta-regression [8[of violence risk factors in psychosis identified robust associations for domains that include criminal history and several dynamic variables (eg, recent substance misuse and medication nonadherence), supporting the clinical primacy of treatable targets rather than diagnosis alone. Insight, in turn, is tightly linked to adherence and outcomes. In a large European observational study [8], better insight correlated with higher medication adherence and stronger therapeutic alliance, and modeled pathways indicated that insight can shape clinical severity indirectly via adherence and alliance. Medication nonadherence has also been associated with violence outcomes in large cohorts; for example, a community-based study [9] of patients with schizophrenia found nonadherence associated with higher odds of multiple categories of violence to others, even after rigorous confounder control strategies. In forensic settings, these relationships are clinically consequential because tribunals often evaluate not only past behavior but also the likelihood of relapse, destabilization, and risk escalation (processes in which insight and adherence are central). Accordingly, an AI-themed delusion may be best conceptualized not as an unusual narrative but as a potential marker of delusional system organization and rigidity, which could correlate with impaired insight, refusal of treatment, substance misuse, and broader criminogenic histories.

Legal Responses and Judicial Risk Assessment When Delusional Content Is Novel

Legal systems tasked with managing individuals found not criminally responsible on account of mental disorder (NCRMD) or insanity-acquitted populations typically prioritize public safety, while balancing therapeutic needs and proportional restrictions. Canadian data illustrate that the NCRMD status is not a loophole. In a large comparative analysis [10], individuals found NCRMD were substantially more likely to be detained and remained under supervision longer than convicted counterparts, reflecting a risk management orientation rather than punitive sentencing logic. Within Quebec's review board context, open access empirical work describing the *Commission d'Examen des Troubles Mentaux* (CETM) or *Tribunal Administratif du Quebec* (TAQ) practices emphasizes that dispositions (detention, conditional discharge, absolute discharge) are ordered in whether the person constitutes a significant threat to public safety and that tribunals rely heavily on treating teams' clinical reports and structured approaches to risk formulation, with specific attention to standardized tools, such as the Historical Clinical Risk Management-20 (HCR-20) [11]. At the same time, forensic guidance cautions against

XSL•FO
RenderX

instrumental overreach and narrative bias: risk tools and risk formulations can become rhetorically compelling in ways that outstrip predictive precision, potentially fostering risk-averse interpretations when cases contain vivid, emotionally salient material [12]. Systematic review evidence on structured professional judgment (SPJ) frameworks shows that SPJ-informed guidance (including HCR-20-based approaches) is increasingly used to support high-stakes secure care decisions but should function as an adjunct rather than a substitute for comprehensive case formulation and contextual clinical judgment [13]. AI-themed delusions place pressure on this ecosystem of judgment because they can sound simultaneously plausible (given real surveillance or algorithmic personalization) and bizarre (given perceived sentience, omniscience, or mind control), heightening the interpretive ambiguity that courts and clinicians must manage. This makes judicial texts a valuable empirical window into how "novel" delusional content is translated (explicitly or implicitly) into risk rationales and legal dispositions.

Objectives and Hypotheses

This study attempted to answer the following question: *How does the centrality of AI within psychotic delusional systems relate to forensic risk and judicial decision-making?* We sought to clarify whether AI-themed delusions function as contemporary narrative content or whether they occupy a structurally central role that meaningfully shapes behavior, risk trajectories, and review board dispositions. To meet this goal, the following specific objectives were defined:

- To systematically identify and characterize judgments in which AI is explicitly referenced within delusional content
- To classify the degree of AI centrality within the delusional system (central, marginal, or secondary) using a structured coding framework
- To examine the association between AI centrality and core forensic risk variables, including violence toward others, treatment refusal, lack of insight, substance use, and prior criminal history
- To analyze whether AI centrality is associated with judicial outcomes, including detention, conditional discharge, or unconditional discharge, as well as findings of significant public safety risk

Based on the available data in the current literature on delusional system organization, insight deficits, and violence risk in psychotic disorders, several hypotheses can be formulated:

- Cases in which AI occupies a central position within the delusional architecture could be associated with higher rates of impaired insight and treatment nonadherence, reflecting a more systematized and behaviorally organizing belief structure.
- AI centrality could be associated with an increased likelihood of violence toward others when the AI is construed as a persecutory or controlling agent that issues commands or justifies defensive or retaliatory actions.
- AI-central cases may demonstrate a higher probability of judicial findings of significant public safety risk and more restrictive dispositions compared to cases in which AI content is marginal or incidental.
- AI centrality will cluster with established criminogenic and clinical risk factors, suggesting that AI-themed delusions do not constitute a fundamentally new risk category but rather represent a contemporary technological expression of structurally organized psychosis within existing forensic risk frameworks.

Methods

Study Design and Data Source

This study followed a retrospective observational design using jurisprudential data extracted from publicly available legal decisions. The data source was the Société québécoise d'information juridique (SOQUIJ) database, which archives decisions from Quebec courts and administrative tribunals [14]. The study period extended from database inception to December 31, 2025. The study adhered to the RECORD (Reporting of Studies Conducted Using Observational Routinely Collected Health Data) guidelines and was further aligned with American Psychological Association (APA) Journal Article Reporting Standards (JARS) for quantitative research to enhance transparency and reproducibility [15].

Inclusion and Exclusion Criteria

Judgments were included if they met the following criteria: (1) the presence of a diagnosed or alleged psychotic-spectrum disorder; (2) explicit reference to AI, technological surveillance, algorithmic systems, or related constructs within the delusional content; and (3) judicial discussion of dangerousness, detention status, conditional discharge, or treatment authorization. Both review board and civil authorization of care decisions were eligible where AI-themed beliefs were materially described in the reasoning. Cases were excluded if AI references were incidental, metaphorical, or unrelated to psychopathology (eg, references to AI tools used in evidence review). Judgments lacking sufficient factual detail to assess the relationship between AI-themed delusions and behavioral manifestations were also excluded.

Participant (Case) Characteristics

The unit of analysis was the individual judicial decision rather than the accused, as multiple decisions concerning the same individual at different procedural stages were treated as distinct evaluative moments. As such, participant-level demographic variables (eg, age, sex, ethnicity, socioeconomic status) were inconsistently reported across judgments and could not be systematically analyzed. When available, these characteristics were extracted descriptively but were not used as inclusion criteria or covariates.

Multiple judgments pertaining to the same individual at different time points were to be treated as independent observations, reflecting distinct clinical and legal decision-making contexts. However, such judgments were not identified in this study.

Sampling Procedures

Search terms included both French and English expressions (eg, "artificial intelligence," "AI," "psychosis," "delusions," "dangerosity," "Tribunal," "Review Board," "temporary confinement," and "treatment order"). Searches were conducted

XSL•FO
RenderX

without date restrictions to capture the full emergence of AI-themed delusions in reported case law. Identified judgments were screened for relevance, and duplicate entries were removed. The final dataset consisted of decisions in which AI or related technological constructs were explicitly referenced in the factual findings, psychiatric evidence, or judicial reasoning. The full dataset is provided in Multimedia Appendix 1, with relevant judgment IDs from the SOQUIJ's legal database. The full search strategy that was applied to identify judgments and construct the dataset is presented in Multimedia Appendix 2.

Sample Size, Power, and Precision

The intended sample size was not predefined, as the study aimed to identify all eligible cases within the database. The achieved sample consisted of 29 judgments meeting the inclusion criteria. Given the exploratory nature of the study and the rarity of AI-themed delusions in published jurisprudence, no formal a priori power calculation was conducted. Analyses emphasized estimation (odds ratios [ORs] and 95% CIs) rather than statistical significance testing.

Measures and Covariates

The primary independent variable was AI centrality within the delusional system, operationalized as a categorical variable (central vs noncentral) based on whether AI constituted the organizing framework of the delusional narrative.

AI centrality was operationalized as the degree to which AI functioned as the primary organizing principle of the delusional system. A case was coded as "central" when AI structured core explanatory mechanisms, including agency attribution (eg, AI controlling thoughts or actions), causal interpretation of events, or behavioral justification. In contrast, cases were coded as "noncentral" when AI appeared as a peripheral or incidental element within a broader delusional framework without organizing function. For example, cases were classified as AI-central when individuals described AI as directly controlling their behavior, issuing commands, or governing reality (eg, "The AI is directing my actions" or "The system controls what happens to me"). In contrast, cases were classified as AI-noncentral when AI was referenced without structural influence (eg, "AI is monitoring me" within a broader persecutory system not organized around AI).

The primary outcome was documented violence toward others, defined as any heteroaggressive behavior described in the judicial record.

Secondary outcomes included direct AI-linked violence attribution and judicial findings of significant public safety risk.

Covariates included lack of insight, treatment adherence, substance use disorder, and prior violence history, as described in tribunal reasoning.

Data Collection

A structured coding grid was developed prior to data extraction to ensure standardized variable capture. Variables included demographic indicators (when available), psychiatric diagnoses, substance use history, prior violence, insight and treatment adherence, description of delusional themes, centrality of AI within the delusional system, explicit linkage between AI-themed beliefs and the index behavior, tribunal findings regarding significant public safety risk, and disposition outcomes (eg, detention, conditional discharge, authorization of care). AI centrality was coded categorically as "central," "marginal," "secondary," or "not discussed" based on whether AI constituted the organizing framework of the delusional belief system or merely one thematic element among others. Violence variables were coded as "present" or "absent" for self-directed and other-directed behaviors, and, where possible, classified by severity.

Quality of Measurements

Data extraction was conducted by authors AH and CP through full-text review of each judgment, including factual summaries, psychiatric expert testimony, and tribunal reasoning. To enhance reliability, ambiguous cases were reviewed twice, and coding decisions were linked to explicit textual citations from the judgment.

Instrumentation

No standardized psychometric instruments were used in this study. All variables were derived from structured coding of judicial texts using an ad hoc coding framework developed for this study.

Masking

Masking was not applicable, as data were derived from publicly available judicial records and no experimental manipulation or condition assignment was performed.

Psychometrics

Formal psychometric validation of the coding framework was not conducted. However, the framework was designed to maximize construct clarity and reproducibility through explicit operational definitions and consensus-based coding procedures. Interrater reliability statistics were not calculated, given the exploratory nature of the study and the small sample size.

Data Diagnostics

No cases were excluded after data extraction. Missing data were inherent to the use of judicial records and were coded as "not documented" when variables were not explicitly reported. No imputation procedures were applied. Given the small sample size, formal outlier detection and distributional transformations were not performed.

Analysis

Descriptive statistics were first computed to characterize the distribution of AI centrality, violent behaviors, treatment adherence, and judicial outcomes within the dataset. The primary analytic objective was to examine the association between AI centrality and actual violence or judicial findings of significant public safety risk. AI centrality was dichotomized (central vs noncentral) for inferential testing. Bivariate associations between AI centrality and violence outcomes were assessed using Fisher's exact test due to the small sample size.

ORs (95% CIs) were calculated to estimate the effect magnitude. Secondary analyses examined whether AI-themed delusions

XSL•FO
RenderX

are explicitly linked to the behavior underlying the legal proceeding, distinguishing phenomenological presence from behavioral causation. Exploratory logistic regression models were constructed to evaluate whether AI centrality independently predicts violence after adjusting for established risk factors, such as lack of insight, treatment nonadherence, substance use, and prior violence history. Given the modest sample size, regression results were interpreted cautiously, emphasizing effect size and CIs rather than statistical significance alone. Qualitative thematic synthesis was additionally performed to contextualize quantitative findings, particularly regarding how tribunals conceptualized AI-themed delusions in relation to risk, insight, and treatment compliance.

Ethical Considerations

As the data consisted exclusively of secondary analysis of publicly accessible, anonymized legal records, the study did not involve direct interaction with human participants and did not meet the definition of human subject research under applicable institutional and national guidelines. Accordingly, formal research ethics board review was not required.

No informed consent was obtained, as the study did not involve primary data collection. The original data (judicial decisions) are part of the public legal record, and their use for research purposes is permitted within the framework of open access legal information systems. All data used in this study were derived from deidentified judicial texts. No attempt was made to reidentify individuals, and no personally identifiable information beyond what is already publicly available in legal decisions was extracted or reported. In presenting results, care was taken to avoid inclusion of unnecessary identifying details, and all findings are reported in aggregate form.

No participants were recruited, and no compensation was provided, as this study did not involve human subjects. No images, figures, or supplementary materials include identifiable individuals. All content is derived from textual legal sources, and no individual-level identification is possible from the data presented.

Results

Sample Characteristics and Flow

A total of 29 judgments met the inclusion criteria and were retained for analysis. In all cases, AI was explicitly referenced within the delusional content described in the judicial record. AI centrality within the psychotic framework was coded dichotomously as central (including cases coded as “secondary to central”) or noncentral (including cases coded as “marginal,” “secondary,” or “not discussed”). AI was considered central in 15 (51.7%) cases and noncentral in 14 (48.3%) cases.

Across the sample, dangerosity toward others, as determined in tribunal reasoning or based on documented heteroaggressive behavior, was identified in 20 (69%) cases (95% CI 50.3-83.1), whereas 9 (31%) cases did not involve a finding of violence toward others. Thus, the overall base rate of violence toward others in this forensic sample was high. Characteristics of the judgments are found in Table 1.

XSL•FO
RenderX

Table 1. Descriptive characteristics of forensic psychiatric judgments (N=29) involving AI[a]-related delusions.

Characteristics	Judgments, n (%)
AI centrality	
Central (including secondary to central)	15 (51.7)
Noncentral (marginal, secondary, not discussed)	14 (48.3)
Violence toward others[b]	
Present	20 (69.0)
Absent	9 (31.0)
Direct AI- linked violence attribution	
Direct causal attribution	11 (37.9)
Indirect or none	18 (62.1)
Judicial findings of significant public safety risk[c]	
Risk confirmed by tribunal	22 (75.9)
No significant risk found	7 (24.1)
Judicial disposition	
Continued detention/maintenance of custody	15 (51.7)
Conditional discharge (with modalities)	10 (34.5)
Absolute discharge	4 (13.8)
Documented lack of insight	
Present (partial or absent insight)	23 (79.3)
Adequate insight	6 (20.7)
Treatment refusal or nonadherence	
Present	17 (58.6)
Absent/adherent	12 (41.4)
Substance use disorder documented	
Present	14 (48.3)
Not documented	15 (51.7)

[a]AI: artificial intelligence.

[b]Violence toward others refers to documented heteroaggressive behavior described in the judicial record.

[c]Judicial findings of significant public safety risk reflect tribunal determinations of dangerousness.

AI Centrality and Violence Toward Others

As shown in Table 2, cases in which AI was coded as a central component of the delusional or explanatory framework were more likely to involve documented violence toward others compared to cases in which AI was marginal or absent. Specifically, violence toward others was observed in 80.0% (12/15) of AI-central cases (95% CI 54.8-93.0) versus 57.1% (8/14) of AI-noncentral cases (95% CI 32.6-78.6). This corresponded to an unadjusted OR of 2.91 (95% CI 0.63-13.45). Although the magnitude of the association suggests a potentially meaningful increase in the likelihood of violence when AI content is central, the difference did not reach statistical significance using Fisher's exact test (P=.26).

XSL•FO
RenderX

Table 2. AI[a] centrality and violence toward others.

Predictor and outcomes	AI-central/exposed, n/N (%); 95% CI	AI-noncentral/unexposed, n/N (%); 95% CI	Effect size, OR[b] (95% CI)	*P* value (Fisher's exact test)
AI centrality[c]				
Violence toward others[d]	12/15 (80.0); 54.8-93.0	8/14 (57.1); 32.6-78.6	2.91 (0.63-13.45)	.26
Direct AI-linked violence attribution[e]	9/15 (60.0); 35.8-80.2	2/14 (14.3); 4.0-39.9	9.00 (1.48-54.6)	.01
Judicial findings of significant public safety risk[f]	13/15 (86.7); 62.1-96.3	9/14 (64.3); 38.8-83.7	3.60 (0.63-20.5)	.24
Direct AI-linked violence attribution				
Violence toward others	10/11 (90.9); 62.3-98.4	10/18 (55.6); 33.7-75.4	8.00 (0.87-73.7)	.07
Judicial findings of significant public safety risk	10/11 (90.9); 62.3-98.4	12/18 (66.7); 43.7-83.7	4.00 (0.39-41.3)	.22

[a]AI: artificial intelligence.

[b]OR: odds ratio.

[c]AI centrality was dichotomized as central versus noncentral based on structured coding of delusional system organization.

[d]Violence toward others was defined as documented heteroaggressive behavior.

[e]Direct AI-linked violence attribution refers to the explicit causal linkage between AI-related beliefs and the index behavior.

[f]Judicial findings of significant public safety risk reflect tribunal determinations of dangerousness.

Importantly, although AI centrality was strongly associated with explicit AI-linked violence attribution (OR 9.00, 95% CI 1.48-54.6; *P*=.01), its relationship with actual violent behavior was more moderate and statistically nonsignificant. This pattern suggests that AI centrality may be more associated with the cognitive framing and narrative organization of violence rather than serving as an independent predictor of violent conduct per se. However, the wide CIs reflect substantial uncertainty around the effect estimate, consistent with limited statistical power in this sample (N=29), and the data remain compatible with both increased and decreased risk.

These findings indicate that within this dataset, centrality of AI-themed delusions was not statistically associated with increased violence toward others.

Qualitative Pattern of Judicial Reasoning

A detailed review of tribunal reasoning further contextualizes this statistical pattern. In cases where violence was affirmed, the justification for detention or restrictive measures rarely emphasized AI content as a primary causal factor. Instead, judicial decisions consistently highlighted persistent psychotic symptoms, impaired insight, treatment nonadherence, substance use, prior violent behavior, impulsivity, and instability of living conditions.

AI-related delusional material was typically described as part of the phenomenological expression of the psychosis, often embedded within persecutory, grandiose, or referential themes. Even in cases where AI content was central and ontologically elaborate, tribunals tended to ground their assessment of risk in behavioral history and clinical instability rather than in the thematic content of the delusion itself.

In contrast, several cases in which AI was central but violence was absent were characterized by improved insight, medication adherence, therapeutic alliance, and stable supervision. In those instances, tribunals explicitly noted clinical stabilization and controlled risk under structured conditions. The thematic typology of AI-related delusions is found in Table 3.

XSL•FO
RenderX

Table 3. Thematic typology of AI[a]-related delusions in forensic psychiatric judgments.

Thematic category[b]	Operational definition	Core delusional structure	Cases (N=29), n (%)	AI-central cases[c] (n=15), n (%)	AI-noncentral cases[c] (n=14), n (%)
AI as a controlling agent	AI described as externally directing thoughts, behavior, or life events	Externalized agency; behavioral control; algorithmic directives	12 (41.4)	10 (66.7)	2 (14.3)
AI as a persecutory system	AI portrayed as surveilling, torturing, targeting, or harming the individual	Technological paranoia; surveillance; implanted devices	14 (48.3)	11 (73.3)	3 (21.4)
AI as a command source	AI issuing explicit instructions or moral imperatives	Command structure; authority-based obedience	6 (20.7)	5 (33.3)	1 (7.1)
AI as a grandiose creation	Individual claims to have created, mastered, or uniquely understood AI	Narcissistic expansion; technological omnipotence	8 (27.6)	6 (40.0)	2 (14.3)
AI as an ontological reality framework	AI used to reframe the nature of existence (simulation, world creation, universal system)	Cosmological restructuring; reality substitution	7 (24.1)	6 (40.0)	1 (7.1)
AI as a moral/existential justification	AI used to justify actions, destiny, or mission	Externalized moral authority	9 (31.0)	7 (46.7)	2 (14.3)
AI as a marginal referential element	AI referenced but not structurally central to the delusional system	Peripheral thematic content	14 (48.3)	0	14 (100.0)

[a]AI: artificial intelligence.

[b]Categories represent recurring patterns of AI-related delusional content identified through qualitative thematic synthesis.

[c]AI-central cases (n=15, 51.7%) refer to judgments in which AI functioned as the primary organizing structure of the delusional system, whereas AI-noncentral cases (n=14, 48.3%) refer to peripheral or incidental AI content.

Interpretation of the Association

Table 4 presents the distribution of core forensic risk factors according to AI delusional centrality. When Tables 2 and 4 are considered together, a coherent pattern emerges: AI centrality is not independently associated with violent behavior toward others, but it appears to cluster with markers of epistemic vulnerability that are themselves relevant to forensic risk formulation. In Table 2, the proportion of cases involving violence toward others did not differ significantly between AI-central and AI-noncentral cases, and effect sizes were modest with wide CIs crossing unity. This indicates that AI occupying a structurally central position within the delusional system does not, in this sample, translate into a statistically detectable increase in heteroaggressive conduct.

Table 4. Distribution of core forensic risk factors according to AI[a] delusional centrality[b].

Risk variable	AI-central cases (n=15), n (%)	AI-noncentral cases (n=14), n (%)	OR[c] (95% CI)[d]	*P* value
Lack of insight	13 (86.7)	8 (57.1)	4.89 (0.79-30.1)	.09
Refusal of treatment	9 (60.0)	4 (28.6)	3.75 (0.74-18.9)	.14
Substance use disorder	10 (66.7)	7 (50.0)	2.00 (0.42-9.46)	.46
Prior violence history	8 (53.3)	6 (42.9)	1.52 (0.33-7.05)	.72
Judicial findings of significant public safety risk	11 (73.3)	7 (50.0)	2.75 (0.57-13.3)	.27

[a]AI: artificial intelligence.

[b]AI centrality was defined based on whether AI constituted the organizing framework of the delusional system.

[c]OR: odds ratio.

[d]ORs (95% CIs) indicate between-group differences.

However, Table 4 refines this interpretation by demonstrating that AI-central cases exhibit a stronger concentration of cognitive and clinical risk indicators. The most pronounced association concerns lack of insight, which was markedly more prevalent in AI-central cases, with an OR approaching 5, albeit with a wide CI, reflecting limited statistical power. Treatment refusal also trended higher in the AI-central group, with an OR suggestive of a clinically meaningful effect size, though again not reaching statistical significance. In contrast, classical criminogenic variables, such as prior violence history and

XSL•FO
RenderX

substance use disorder, were only weakly associated with AI centrality, with effect estimates close to unity and nonsignificant *P* values. Furthermore, the model demonstrated minimal explanatory value (pseudo R^2=0.001), and CIs were wide, reflecting substantial imprecision. These findings should therefore be interpreted as inconclusive.

This pattern suggests that AI centrality may function less as a direct behavioral risk amplifier and more as a structural indicator of delusional system consolidation. In other words, when AI is positioned as a primary agent within the individual's explanatory framework (controlling events, issuing commands, or structuring reality), the psychotic architecture appears more entrenched and less permeable to corrective feedback. This interpretation is consistent with the elevated prevalence of insight impairment and treatment nonadherence observed in Table 4, while remaining aligned with the absence of a statistically significant increase in the overt violence observed in Table 2.

Judicial findings of significant public safety risk were more frequent in AI-central cases, but this association did not reach statistical significance and displayed a moderate effect size with wide CIs. This may indicate that tribunals are responding, not to documented increases in violent behavior per se, but to the overall clinical presentation characterized by impaired insight, perceived instability, and risk of decompensation. Thus, the association between AI centrality and judicial outcomes appears mediated by broader clinical risk constructs rather than by direct violence linkage.

These findings support a more nuanced interpretation: AI centrality is associated with epistemic disturbance and vulnerability within the psychotic system, but it does not independently predict violence toward others in this cohort. The clinical and forensic relevance of AI centrality may therefore lie in its relationship to insight erosion and treatment instability rather than in direct behavioral escalation.

Implications for Forensic Psychiatry

From a forensic psychiatry perspective, these results challenge potential assumptions that technologically themed delusions represent a uniquely destabilizing or violence-enhancing subtype of psychosis. Instead, AI-themed delusions appear to mirror historical shifts in persecutory or grandiose content, reflecting sociocultural embedding of psychotic material without fundamentally altering the behavioral risk architecture.

In this sample of Quebec forensic judgments, tribunals implicitly recognized this distinction, focusing decisively on clinical stability and behavioral history rather than on the thematic novelty of AI-related beliefs. The data therefore support the interpretation that AI centrality is phenomenologically salient but not independently predictive of violence.

Figure 1 illustrates the proposed structural model linking AI delusional centrality to forensic risk outcomes. In this framework, AI-related content is not treated as a homogeneous thematic element but is differentiated according to its structural role within the delusional system. When AI occupies a central position (functioning as a controlling agent, persecutory system, command source, or ontological framework) it is hypothesized to reorganize agency attribution and meaning-making processes. This centrality appears to operate through mediating risk mechanisms, including diminished insight, increased cognitive rigidity, externalization of agency, behavioral justification grounded in delusional logic, and higher likelihood of treatment refusal. These mediators, in turn, are associated with downstream forensic outcomes, such as prior history of violence, judicial findings of significant public safety risk, detention or conditional discharge decisions, and determinations of treatment incapacity. In contrast, when AI content remains peripheral or decorative within the delusional narrative, it does not appear to exert the same structural influence on agency, treatment engagement, or judicial risk assessment. The model therefore conceptualizes AI delusional centrality as a qualitative intensifier of established criminogenic and clinical risk processes rather than as an isolated thematic variable.

XSL•FO
RenderX

Figure 1. Structural model of AI delusional centrality and forensic risk pathways. The model illustrates how AI centrality within delusional systems (defined as AI functioning as a controlling agent, persecutory system, command source, or ontological framework) is hypothesized to influence forensic outcomes indirectly through mediating mechanisms, including impaired insight, cognitive rigidity, externalization of agency, and treatment nonadherence. These intermediate factors are associated with downstream outcomes, such as violence toward others, judicial findings of significant public safety risk, and restrictive dispositions. The model reflects observed empirical patterns and theoretical integration rather than causal inference. AI: artificial intelligence.

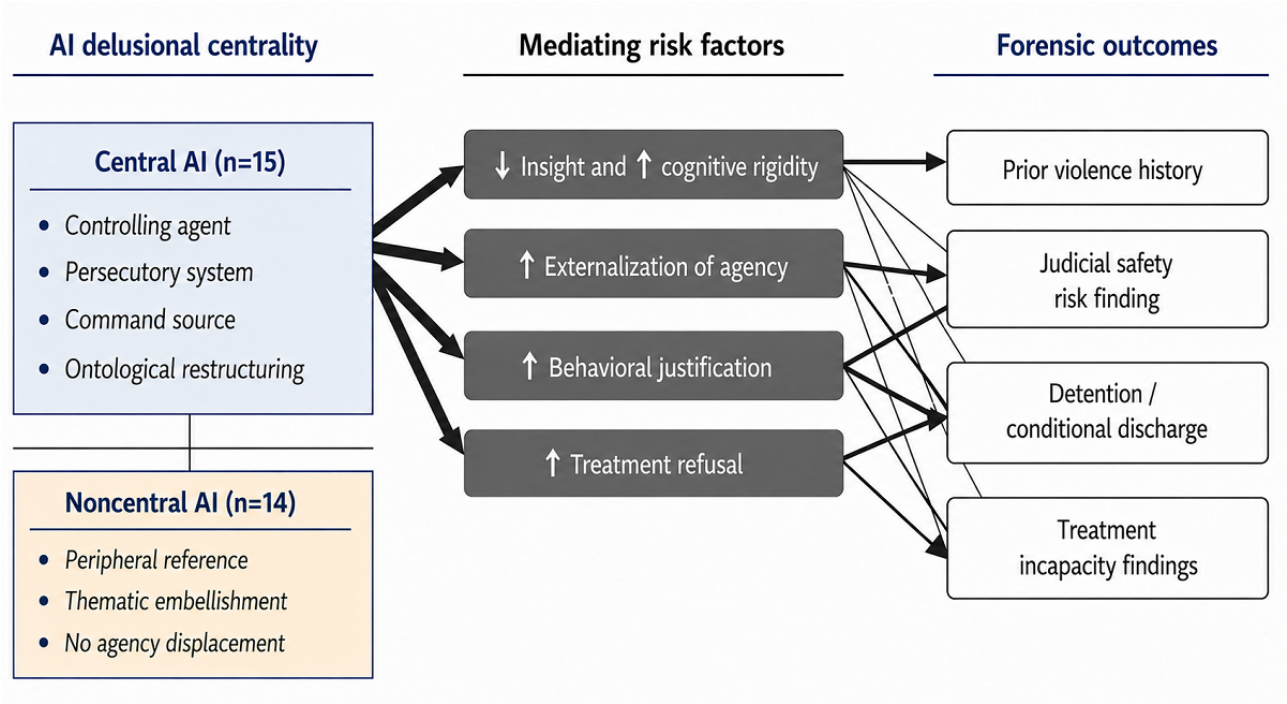

Discussion

Principal Findings

This jurisprudential analysis of 29 Quebec forensic judgments found that AI-related beliefs were frequently embedded within persecutory and agency-displacing psychotic narratives and that the overall base rate of heteroaggressive behavior or violence toward others was high in this tribunal-derived sample. Across decisions, tribunals most consistently anchored their reasoning in conventional clinical and behavioral determinants of risk (impaired insight, treatment engagement, persistence of psychotic symptoms, substance use, and prior violence), while AI content was often treated as a phenomenological expression of illness rather than a stand-alone driver of dangerousness.

The central finding was that AI centrality is not statistically significantly associated with the occurrence of violence; however, effect estimates were imprecise and the study was underpowered, limiting the ability to draw definitive conclusions regarding the presence or absence of an association. In contrast, AI centrality was strongly associated with explicit AI-linked violence attribution, indicating that when AI is structurally central to the delusional architecture, violence is more often framed as being caused, commanded, justified, or organized by the AI narrative. Stratified patterns further suggested that AI centrality tends to co-occur with markers of epistemic vulnerability (especially impaired insight and treatment refusal) more than with classical criminogenic variables, such as prior violence history or substance use disorder. AI centrality is therefore suggested to function as a structural organizer of agency and meaning-making that may influence judicial risk reasoning indirectly through insight and adherence mechanisms rather than as a direct behavioral risk amplifier.

Interpretation and Comparison With Prior Work

The question of whether AI-themed delusions represent a qualitatively new risk construct must be situated within the broader phenomenology of persecutory and passivity delusions. Empirical work on threat anticipation and persecutory ideation demonstrates that the structure of paranoid belief formation is typically driven by cognitive-affective processes rather than by specific thematic content [16,17]. Freeman et al [16] showed that persecutory delusions are maintained through misinterpretations of social threat and safety behaviors that prevent disconfirmation. Similarly, research on anomalous experiences and belief evaluation indicates that delusional conviction is shaped by reasoning biases and affective dysregulation rather than the cultural substrate of the belief [17]. Within this framework, AI can be understood as a contemporary narrative vehicle for long-standing mechanisms of agency externalization and threat attribution. The findings of this study, demonstrating that AI centrality does not independently predict violence, align with this cognitive model: the thematic novelty of AI does not appear to supersede the underlying psychological architecture of paranoia.

The distinction between thematic content and structural role is also conceptually important. Delusional themes (eg, surveillance, control, persecution) often reflect the sociocultural context, whereas structural features refer to how beliefs organize agency, causality, and behavioral meaning within the psychotic system. AI was conceptualized as structurally central when it functioned as a primary agent within this architecture, shaping not only the

XSL•FO
RenderX

content of beliefs but also the explanatory logic linking perception, intention, and action. This distinction is grounded in cognitive models of psychosis, which emphasize that behavioral consequences are more strongly determined by belief organization, conviction, and integration than by thematic content alone [16,17]. From this perspective, AI centrality is therefore not only a thematic variation but also a potential marker of system-level consolidation of delusional belief.

From a violence risk perspective, the results are consistent with longitudinal evidence demonstrating that psychotic symptoms alone are weak predictors of violent outcomes when separated from dynamic contextual variables. Large epidemiological and register-based studies [18,19] have indicated that violence risk in psychotic disorders is substantially moderated by substance misuse, prior violence, and social instability rather than by specific delusional themes. Swanson et al [18] showed that comorbid substance use and criminogenic history account for much of the excess violence risk observed in schizophrenia-spectrum disorders. Fazel et al's [19] population-based investigations further demonstrated that familial and environmental confounders attenuate the association between psychosis and violence, underscoring the primacy of broader risk ecology. Moreover, research examining "threat/control-override" symptoms suggests that only specific configurations of persecutory beliefs combined with affective arousal predict violence, and even then, effect sizes are modest [20]. This study's pattern (where AI centrality clustered more strongly with impaired insight and treatment refusal than with prior violence) therefore supports the interpretation that AI operates as a structural intensifier of delusional organization rather than as an independent criminogenic variable.

Judicial decision-making research further contextualizes these findings. Studies of forensic risk communication demonstrate that tribunals often rely on narrative coherence and perceived treatability when translating psychiatric evidence into legal determinations [21,22]. Large-scale analyses of insanity acquittees and review board populations indicate that SPJ and clinical stability markers dominate disposition decisions, even when symptom content is vivid or unusual [23]. Research on stigma and perceived dangerousness also suggests that technologically mediated or "modern" delusional themes may amplify perceived unpredictability, yet actual risk judgments remain anchored to behavioral history and compliance indicators [24]. In this light, the findings of this study (showing strong associations between AI centrality and AI-linked violence attribution but not with actual violence) suggest that AI may influence the explanatory framing of behavior without fundamentally altering tribunals' reliance on established risk variables. This distinction between narrative attribution and behavioral prediction is consistent with prior forensic scholarship emphasizing the need to disentangle phenomenology from empirically supported risk mechanisms.

Implications for Clinical and Forensic Practice

Clinically, these findings reinforce the importance of assessing delusional structure rather than thematic novelty. When AI occupies a central role (functioning as a controlling agent, persecutory system, or moral authority), clinicians are encouraged to evaluate associated markers of epistemic rigidity, diminished insight, and impaired treatment engagement. Evidence-based cognitive models of paranoia emphasize collaborative belief examination, reduction in safety behaviors, and enhancement of metacognitive awareness as key interventions [16,17]. Given the well-established association between nonadherence and adverse outcomes, including violence in high-risk populations [18,20,25], management strategies in AI-central cases should prioritize alliance building, adherence monitoring, and substance use stabilization rather than focusing exclusively on the technological content of the delusion. Inquiry into AI use should be routine, framed neutrally, and integrated into broader assessment of digital behavior patterns that may reinforce maladaptive explanatory systems.

For forensic tribunals, the implications are conceptual and procedural. AI-themed delusions may appear uniquely alarming because they resonate with real-world technological anxieties, but empirical evidence continues to support reliance on dynamic risk factors (clinical stability, substance use, behavioral history, and adherence capacity) as the strongest predictors of future harm [19,23,26]. SPJ approaches remain appropriate, provided they are grounded in individualized formulation rather than narrative salience [22]. Explicit differentiation between the plausibility of certain technological fears in contemporary society and the individual's demonstrated behavioral risk profile may reduce novelty bias and promote proportionate dispositions. As digital technologies become increasingly embedded in everyday life, forensic systems will require refined assessment frameworks that address technologically mediated delusions without conflating cultural modernization with categorical dangerousness.

Limitations

Several limitations qualify interpretation. First, the sample size was modest (N=29), producing wide CIs and limited power to detect anything but large effects; null findings regarding violence and judicial risk should therefore be interpreted as inconclusive rather than definitive absence of association. Second, the dataset was constructed from publicly available judgments, which are not designed for research purposes and reflect selection processes in publication and documentation. Decisions that are written, reported, and indexed may differ systematically from those that are not, and the level of clinical detail varies by tribunal, legal issue, and authoring member, increasing the risk of information bias. Third, the unit of analysis was the judgment rather than the individual; where multiple decisions concern the same person across time, observations may not be independent, and temporal dynamics of symptoms and risk management may be conflated with cross-sectional comparisons. Fourth, key variables were coded from judicial descriptions of clinical evidence; misclassification is possible where judgments summarize testimony selectively or omit negative findings. Fifth, AI centrality is a constructed exposure variable requiring interpretive coding and remains vulnerable to coder subjectivity and differential documentation. Finally, generalizability is limited, as Quebec review board processes and reporting practices differ from other jurisdictions and the sample likely overrepresents severe and unstable cases.

XSL•FO
RenderX

Conclusion

In conclusion, this study demonstrated that AI-related delusional content is no longer a peripheral narrative feature in forensic psychiatric judgments, but in a substantial proportion of cases, it functions as a structurally central element of the psychotic system. When AI themes were central rather than marginal, they were more frequently associated with persistent lack of insight, treatment refusal, substance use comorbidity, prior violence, and formal judicial findings of significant public safety risk. These findings suggest that AI delusions may operate, not as contemporary cultural symbols, but as organizing frameworks that shape risk appraisal, perceived persecution, and behavioral dyscontrol. At the same time, centrality did not automatically determine outcome; clinical stabilization, adherence, and structured supervision remained decisive moderators of risk. By integrating it with forensic risk variables, this work provides an empirically grounded typology of AI delusional centrality within real-world judicial decision-making. The results highlight the need for nuanced psychiatric assessments that differentiate thematic novelty from structural pathogenicity. As digital technologies continue to permeate everyday life, courts and clinicians will increasingly confront technologically mediated delusional systems. Future research should examine longitudinal trajectories, treatment responsiveness, and cross-jurisdictional patterns to better inform forensic practice and public safety policy in the era of digital psychiatry.

Acknowledgments

The spell-checker Antidote was used to suggest language improvements within the manuscript.

Funding

This study was funded indirectly by La Fondation de l'Institut universitaire en santé mentale de Montréal and the operating funds of the l'Institut de valorisation des données (AH).

Authors' Contributions

AH and CP contributed to conceptualization, methodology, data curation, formal analysis, writing—original draft preparation, and writing—review and editing. AH additionally contributed to supervision and funding acquisition. All authors have reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full dataset that was used for this study.
[XLSX File (Microsoft Excel File), 146 KB - jmir_v28i1e93349_app1.xlsx]

Multimedia Appendix 2
Search strategies to identify the judgments.
[PDF File (Adobe PDF File), 78 KB - jmir_v28i1e93349_app2.pdf]

References

1. Burns AV, Nelson K, Wang H, Hegarty EM, Cohn AB. 'The algorithm is hacked': analysis of technology delusions in a modern-day cohort. Br J Psychiatry 2025 Nov 03:1-5. [doi: 10.1192/bjp.2025.10452] [Medline: 41178082]
2. Mason OJ, Stevenson C, Freedman F. Ever-present threats from information technology: the Cyber-Paranoia and Fear Scale. Front Psychol 2014;5:1298 [FREE Full text] [doi: 10.3389/fpsyg.2014.01298] [Medline: 25505431]
3. Yang N, Crespi B. I tweet, therefore I am: a systematic review on social media use and disorders of the social brain. BMC Psychiatry 2025 Feb 03;25(1):95 [FREE Full text] [doi: 10.1186/s12888-025-06528-6] [Medline: 39901112]
4. Hudon A, Stip E. Delusional experiences emerging from AI chatbot interactions or "AI psychosis". JMIR Ment Health 2025 Dec 03;12:e85799 [FREE Full text] [doi: 10.2196/85799] [Medline: 41273266]
5. Carlbring P, Andersson G. Commentary: AI psychosis is not a new threat: lessons from media-induced delusions. Internet Interv 2025 Dec;42:100882 [FREE Full text] [doi: 10.1016/j.invent.2025.100882] [Medline: 41141286]
6. Keshavan M, Torous J, Yassin W. Do generative AI chatbots increase psychosis risk? World Psychiatry 2026 Feb;25(1):150-151 [FREE Full text] [doi: 10.1002/wps.70017] [Medline: 41536114]
7. Fazel S, Gulati G, Linsell L, Geddes JR, Grann M. Schizophrenia and violence: systematic review and meta-analysis. PLoS Med 2009 Aug;6(8):e1000120 [FREE Full text] [doi: 10.1371/journal.pmed.1000120] [Medline: 19668362]
8. Novick D, Montgomery W, Treuer T, Aguado J, Kraemer S, Haro JM. Relationship of insight with medication adherence and the impact on outcomes in patients with schizophrenia and bipolar disorder: results from a 1-year European outpatient

XSL•FO
RenderX

observational study. BMC Psychiatry 2015 Aug 05;15(1):189 [FREE Full text] [doi: 10.1186/s12888-015-0560-4] [Medline: 26239486]
9. Li Y, Wen H, Xiong C, Lin C, Yang X, Wang D, et al. Medication nonadherence and risk of violence to others among patients with schizophrenia in Western China. JAMA Netw Open 2023 Apr 03;6(4):e235891 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.5891] [Medline: 37017963]
10. Martin S, Charette Y, Leclerc C, Seto MC, Nicholls TL, Crocker AG. Not a "get out of jail free card": comparing the legal supervision of persons found not criminally responsible on account of mental disorder and convicted offenders. Front Psychiatry 2021;12:775480 [FREE Full text] [doi: 10.3389/fpsyt.2021.775480] [Medline: 35115962]
11. Myszak P, Leclair L, Khayat O, Levy J, Abou Jaoude J, Dufour M, et al. Retrospective review of the Criminal Code Review Board in Quebec for the Year 2023. Forensic Sci 2025 Nov 04;5(4):59 [FREE Full text] [doi: 10.3390/forensicsci5040059]
12. Silva E. The HCR-20 and violence risk assessment - will a peak of inflated expectations turn to a trough of disillusionment? BJPsych Bull 2020 Dec;44(6):269-271 [FREE Full text] [doi: 10.1192/bjb.2020.14] [Medline: 33213557]
13. Bowden J, Logan C, Robinson L, Carey J, McDonald J, McDonald R, et al. Clinicians' use of the structured professional judgement approach for adult secure psychiatric service admission assessments: a systematic review. PLoS One 2024;19(9):e0308598 [FREE Full text] [doi: 10.1371/journal.pone.0308598] [Medline: 39325743]
14. Société québécoise d'information juridique. SOQUIJ: legal information and case law from Québec courts and tribunals. Société québécoise d'information juridique. 1976. URL: https://soquij.qc.ca/a/fr [accessed 2026-01-10]
15. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med 2015 Oct;12(10):e1001885 [FREE Full text] [doi: 10.1371/journal.pmed.1001885] [Medline: 26440803]
16. Freeman D, Garety PA, Kuipers E, Fowler D, Bebbington PE. A cognitive model of persecutory delusions. Br J Clin Psychol 2002 Nov;41(Pt 4):331-347. [doi: 10.1348/014466502760387461] [Medline: 12437789]
17. Moritz S, Woodward TS. Jumping to conclusions in delusional and non-delusional schizophrenic patients. Br J Clin Psychol 2005 Jun;44(Pt 2):193-207 [FREE Full text] [doi: 10.1348/014466505X35678] [Medline: 16004654]
18. Swanson JW, Swartz MS, Van Dorn RA, Elbogen EB, Wagner HR, Rosenheck RA, et al. A national study of violent behavior in persons with schizophrenia. Arch Gen Psychiatry 2006 May;63(5):490-499. [doi: 10.1001/archpsyc.63.5.490] [Medline: 16651506]
19. Fazel S, Grann M, Carlström E, Lichtenstein P, Långström N. Risk factors for violent crime in schizophrenia: a national cohort study of 13,806 patients. J Clin Psychiatry 2009 Mar;70(3):362-369. [doi: 10.4088/jcp.08m04274] [Medline: 19284931]
20. Appelbaum PS, Robbins PC, Monahan J. Violence and delusions: data from the MacArthur Violence Risk Assessment Study. Am J Psychiatry 2000 Apr;157(4):566-572. [doi: 10.1176/appi.ajp.157.4.566] [Medline: 10739415]
21. Skeem JL, Monahan J. Current directions in violence risk assessment. Curr Dir Psychol Sci 2011 Feb 04;20(1):38-42. [doi: 10.1177/0963721410397271]
22. Heilbrun K, Yasuhara K, Shah S. Violence risk assessment tools: overview and critical analysis. In: Otto RK, Douglas KS, editors. Handbook of Violence Risk Assessment. 2nd ed. New York: Routledge; 2010.
23. Crocker AG, Nicholls TL, Seto MC, Charette Y, Côté G, Caulet M. The national trajectory project of individuals found not criminally responsible on account of mental disorder in Canada. Part 2: the people behind the label. Can J Psychiatry 2015 Mar;60(3):106-116 [FREE Full text] [doi: 10.1177/070674371506000305] [Medline: 25886686]
24. Angermeyer MC, Matschinger H. The stigma of mental illness: effects of labelling on public attitudes towards people with mental disorder. Acta Psychiatr Scand 2003 Oct;108(4):304-309. [doi: 10.1034/j.1600-0447.2003.00150.x] [Medline: 12956832]
25. Haddou A, Sergerie-Dufresne C, Myszak P, Borduas Pagé S, Hudon A. Understanding psychiatric-legal disagreements in not criminally responsible on account of mental disorder cases: a gradient boosting model perspective. Front Psychol 2025;16:1666828 [FREE Full text] [doi: 10.3389/fpsyg.2025.1666828] [Medline: 41567436]
26. Pellerin J, Désilets M, Borduas Pagé S, Hudon A. Factors Factors considered for the assessment of risk in administrative review boards of Canada: a scoping review. Forensic Sci 2024 Nov 01;4(4):573-587 [FREE Full text] [doi: 10.3390/forensicsci4040039]

Abbreviations

AI: artificial intelligence
HCR-20: Historical Clinical Risk Management-20
NCRMD: not criminally responsible on account of mental disorder
OR: odds ratio
SOQUIJ: Société québécoise d'information juridique
SPJ: structured professional judgment

XSL•FO
RenderX

Edited by S Brini; submitted 11.Feb.2026; peer-reviewed by X Liang, K Morris; comments to author 06.Apr.2026; revised version received 03.May.2026; accepted 11.May.2026; published 20.May.2026.

Please cite as:
Hudon A, Pagé C
Artificial Intelligence Centrality in Psychotic Delusions and Violence Risk in Forensic Psychiatry: Retrospective Observational Study of Judicial Decisions
J Med Internet Res 2026;28:e93349
URL: https://www.jmir.org/2026/1/e93349
doi:10.2196/93349
PMID:

©Alexandre Hudon, Chanel Pagé. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patterns of Digital Media Engagement During Wartime and Their Associations With Stress, Resilience, and Well-Being: Cross-Sectional Survey Study

Moti Zwilling[1,2], PhD; Galit Klein[1], PhD; Sharon Barak[3], PhD; Noa Shtainmetz[4], PhD; Riki Tesler[4], PhD; Avi Zigdon[4], PhD

[1]Department of Economics and Business Administration, Ariel University, Ariel, Judea and Samaria Area, Israel

[2]Ariel cyber innovation center, Ariel, Israel

[3]Faculty of Health Sciences, Department of Nursing, Ariel University, Ariel, Israel

[4]Faculty of Health Sciences, Department of Health Systems Management, Ariel University, Ariel, Judea and Samaria Area, Israel

Corresponding Author:
Avi Zigdon, PhD
Faculty of Health Sciences
Department of Health Systems Management
Ariel University
Science Park, P.O.B. 3
Ariel, Judea and Samaria Area, 40700
Israel
Phone: 972 3 7407200
Email: aviz@ariel.ac.il

Abstract

Background: During wartime, civilians rely on digital media for information, connection, and coping. However, engagement patterns may differ in their associations with stress, resilience, and well-being.

Objective: This study aimed to characterize patterns of digital media engagement during wartime and examine their associations with stress, resilience, and well-being in a cross-sectional survey of Israeli adults.

Methods: We conducted a cross-sectional web-based survey in Israel in early January 2024. Participants (N=512; age: range18-65, mean 37.54, SD 12.89 years; 241/512, 47.1% male) were recruited using nonprobability quota sampling. Well-being was measured using the 5-item World Health Organization Well-Being Index, stress using the 21-item Depression Anxiety Stress Scales stress subscale, and resilience using the 10-item Connor-Davidson Resilience Scale. Participants reported the extent to which their use of 10 digital media activities increased since the beginning of the war (5-point Likert scale). Analyses included 2-step cluster analysis, exploratory factor analysis (principal axis factoring with Promax rotation), and structural equation modeling path analysis.

Results: Cluster analysis identified 2 engagement profiles: low engagement (126/512, 24.6%) and high engagement (386/512, 75.4%; silhouette coefficient=0.30). The high-engagement profile reported higher stress (median 2.39, 95% CI 2.31-2.47) than the low-engagement profile (median 2.01, 95% CI 1.87-2.15; $P<.001$). Exploratory factor analysis supported a 3-factor structure (active, passive, and institutional engagement), explaining 61.86% of the variance (Kaiser-Meyer-Olkin=0.77; Bartlett $\chi^2_{36}=1007.6$; $P<.001$). Structural equation modeling showed acceptable fit ($\chi^2_1=1.5$; $P=.21$; Comparative Fit Index=0.999; Tucker-Lewis Index=0.981; standardized root-mean-square residual=0.013; root-mean-square error of approximation=0.033). Resilience was negatively associated with stress (β=−0.24, 95% CI −0.34 to −0.14; $P<.001$) and positively associated with well-being (β=0.30, 95% CI 0.18-0.42; $P<.001$). Stress was positively associated with active (β=0.27, 95% CI 0.17-0.37; $P<.001$), passive (β=0.31, 95% CI 0.21-0.41; $P<.001$), and institutional engagement (β=0.16, 95% CI 0.06-0.26; $P<.001$). Active engagement was positively associated with well-being (β=0.12, 95% CI 0.04-0.20; $P=.006$), passive engagement showed a negative trend (β=−0.08, 95% CI −0.18 to 0.02; $P=.08$), and institutional engagement was not significantly associated with well-being (β=0.07, 95% CI −0.01 to 0.15; $P=.11$). The total indirect association between resilience and well-being through stress and engagement was significant (indirect effect=0.09, 95% bootstrap CI 0.05-0.13).

XSL•FO
RenderX

Conclusions: Wartime digital media engagement clustered into distinct profiles and dimensions, each showing different associations with stress, resilience, and well-being. This study advances the field by empirically distinguishing active, passive, and institutional engagement, with the latter often overlooked in crisis-media research. The findings offer a more detailed framework for understanding how civilians navigate digital environments during an ongoing armed conflict and provide practical implications for psychosocial guidance and crisis communication, encouraging more intentional engagement and reducing excessive passive consumption. Causal inferences are limited by the cross-sectional design.

(J Med Internet Res 2026;28:e91484) doi:10.2196/91484

KEYWORDS

digital media engagement; institutional engagement; resilience; stress; wartime; well-being

Introduction

Overview

The digital landscape has become an integral component of modern conflict, significantly shaping how individuals and nations navigate and respond to crises [1]. During wartime, civilians increasingly rely on digital channels, including news websites, social media, messaging applications, and institutional platforms, for information, social connection, and emotional regulation; this engagement often involves exposure to emotionally saturated content with implications for digital well-being [2-4]. Real-time social media updates have increasingly displaced traditional broadcast media as the main vehicle for collective exposure to traumatic events [5,6]. The October 7, 2023, attack on Israel exemplified this shift, when platforms such as Telegram and WhatsApp became dominant channels through which civilians encountered graphic wartime content, often involuntarily [5,7,8].

Wartime digital media engagement is heterogeneous, ranging from passive consumption (eg, continuous news monitoring and “doomscrolling”) to active engagement involving information dissemination and seeking social support, and may shape coping resources and mental state [9-13]. Evidence from collective crises indicates that trauma-related media exposure is associated with acute stress and functional impairment [14], with self-reinforcing cycles in which greater concerns lead to more media coverage and higher distress [15-17]. During COVID-19, higher exposure was linked to higher depressive (Patient Health Questionnaire-9 [PHQ-9]) and anxiety (Generalized Anxiety Disorder-7 Scale [GAD-7]) symptom scores [18], and during the Russia–Ukraine war, perceived negative media impact was linked to increased anxiety and reduced resilience, even among uninvolved populations [19,20]. The active vs passive distinction is central to digital well-being research [21-23], although operationalizations vary widely and associations depend on measurement and contextual factors, including platform type [22-25].

Israel has a long-standing research tradition on media-based exposure to terrorism and related mental health consequences [26]. Following the October 7, 2023, attack, a nationwide prospective study documented marked increases in posttraumatic stress disorder (PTSD), generalized anxiety, and depression among Israeli adults [27], and extensive media consumption, particularly graphic Telegram content, was identified as a significant risk factor for PTSD among indirectly exposed Israelis [8]. Psychological stress has been consistently linked to increased media engagement during crises: anxiety related to health-protective behaviors and information seeking is associated with greater media exposure during emergencies [15,16,28], uncertainty-inducing news representations may sustain cycles of worry and excessive consumption [29], and acute distress may override cognitive resources in shaping digital behavior [16,30].

Resilience, conceptualized as the maintenance or rapid recovery of mental health after stressor exposure, may buffer the psychological effects of crisis-related media exposure [31] and moderate associations between crisis-related information exposure and emotional distress [32], including effects of personality vulnerabilities during the COVID-19 pandemic [33]. In Israel, resilience-related factors such as emotion regulation and perceived social support showed stronger associations with mental health outcomes than social media use itself [34].

Protection motivation theory (PMT) provides a conceptual framework, proposing that protective responses depend on appraisals of threat severity and vulnerability and on coping appraisals such as response efficacy and self-efficacy [35,36]. PMT has been applied in crisis contexts; a meta-analysis found positive associations between PMT components and protective behaviors during COVID-19, with coping appraisal variables emerging as particularly strong correlates [36]. Extending PMT to media-related processes, trust in media channels has been shown to predict both threat and coping appraisals [37], and perceived risk and eHealth literacy have been linked to engagement with social media health content and protective behavior [38].

Despite growing evidence, important gaps remain. Most studies focus on overall exposure without differentiating engagement patterns such as active, passive, and institutional modes [22,23]; fewer examine engagement during active armed conflict [19,20]; and resilience has been examined as a moderator in only a limited number of studies, with few examining all 3 engagement patterns simultaneously [32,33]. This study addresses these gaps by empirically characterizing distinct patterns of wartime digital media engagement (active, passive, and institutional) and examining their differential associations with stress, resilience, and well-being. By incorporating institutional engagement, it provides a more granular account of how civilians navigate digital environments during ongoing armed conflict. Based on the framework outlined above, this study addressed the following research questions (RQs):

XSL•FO
RenderX

- RQ1: are there distinct profiles of digital media engagement during wartime, and do these profiles differ in stress levels and well-being?
- RQ2: what is the underlying factor structure of digital media engagement behaviors during a national crisis?
- RQ3: how are stress and resilience associated with well-being during wartime, and do engagement types (active, passive, and institutional) relate to this association?

Objectives

This study aimed to characterize patterns of digital media engagement during wartime and examine their associations with stress, resilience, and well-being in a cross-sectional survey. We also aimed to identify distinct engagement patterns (including active, passive, and institutional modes) and assess how these patterns relate to psychological outcomes in the context of an acute armed conflict.

Methods

Study Design and Participants

We conducted a cross-sectional web-based survey to examine patterns of digital media engagement and their associations with stress, resilience, and well-being during a national crisis. Data collection took place in early January 2024, during the fourth month of the ongoing war, using research interfaces built on the Qualtrics platform. Participants were recruited through Panel4all, a professional web-based survey firm, using an online panel. The survey firm distributed invitations via a unique survey link sent to panel members. Recruitment followed a nonprobability quota sampling approach designed to approximate the distribution of Israeli adults aged 18 to 65 years by age, gender, level of religiosity, and district of residence, as specified by the survey firm. Panel members received incentives for survey completion, provided by the survey firm as part of its standard rewards program (eg, points redeemable for online shopping vouchers).

Procedure

The survey introduction described the study objectives, provided the researcher's contact information, assured anonymity, and clarified that responses would be used for research purposes. Participants provided electronic informed consent before accessing the questionnaire. Participation was voluntary, and participants could withdraw at any time without penalty or consequences.

Sample

The final sample consisted of 512 adults aged 18 to 65 years (mean age 37.54, SD 12.89 years), with 47.1% (241/512) identifying as male. A total of 64% (330/512) were employees, 8.4% (43/512) were managers, 4.5% (23/512) were furloughed, and 9% (46/512) were unemployed. Regarding the sector, 51.2% (262/512) were employed in private companies and 24.6% (126/512) in public organizations; the remaining participants worked in the military or police force. A total of 18.8% (96/512) reported cyber-related knowledge, either through a computer science degree or employment in a related field. Table 1 summarizes participants' demographic and background characteristics.

XSL•FO
RenderX

Table 1. Demographic and background characteristics of the participants (N=512).

Variable	Value
Sex, n (%)	
Male	241 (47.1)
Female	271 (52.9)
Age (years), mean (SD); range	37.54 (12.89); 18-65
Employment status, n (%)	
Employee	330 (64.5)
Manager	43 (8.4)
Furloughed	23 (4.5)
Unemployed	46 (9)
Other	70 (13.7)
Cyber knowledge[a], n (%)	
Yes	96 (18.8)
No	416 (81.3)
Occupation, n (%)	
Public organization	126 (24.6)
Private company	262 (51.2)
Military or security organization	30 (5.9)
Unemployed or other	94 (18.4)

[a]Cyber knowledge—Did you learn or work in the cyber area (eg, software engineering, computer science, information technology, quality assurance, and software technician)?

The data were collected by an external survey company that does not provide datasets with missing responses; accordingly, the dataset used in this study did not contain missing values.

Response Rate and Data Cleaning

A total of 11,130 email invitations were sent (1 invitation per panel member); of which 3622 invitations were opened, 585 respondents accessed the survey link, and 512 completed the questionnaire and were included in the analyses. Respondents who accessed the link but did not complete the questionnaire (n=73) were not included (Figure 1).

XSL•FO
RenderX

Figure 1. Flowchart of participant selection for analysis.

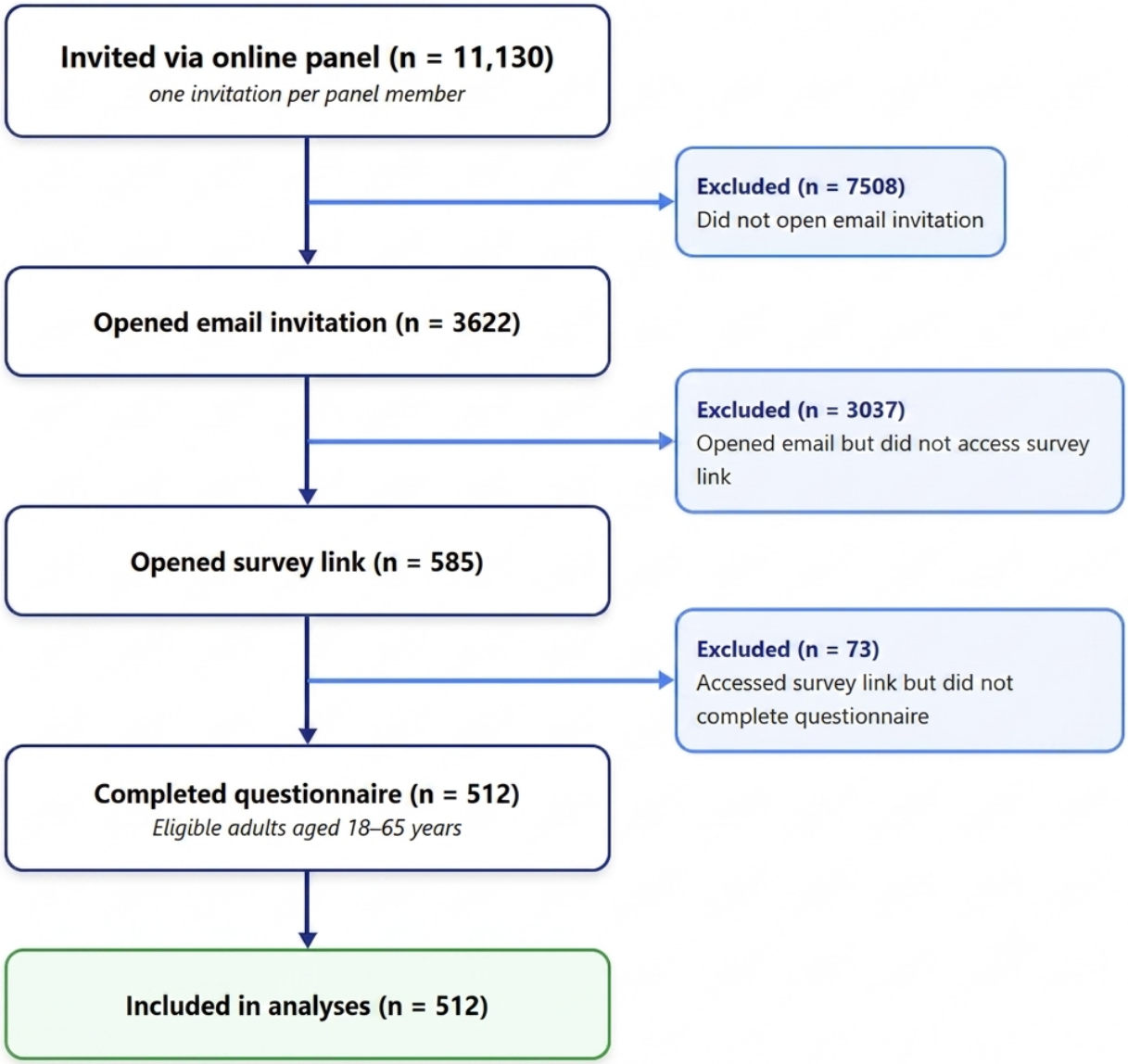

Measures

Well-Being

Well-being was measured using the World Health Organization Well-Being Index, which includes 5 items (eg, “I have felt cheerful and in good spirits”; Cronbach α=0.87) rated on a 6-point Likert scale (0=“not at all” to 5=“all the time”) [39].

Stress

Stress was measured using the 21-item Depression Anxiety Stress Scales 21 [40,41]. We used only the 7-item stress subscale (eg, “I tended to over-react to situations”; Cronbach α=0.92) rated on a 4-point Likert scale (0=“did not apply to me at all” to 3=“applied to me very much, or most of the time”) [42].

Resilience

Resilience was measured using the short version of the Connor-Davidson Resilience Scale (Cronbach α=0.88) [43].

Media Usage Assessment

A distinction between active and passive media use has been proposed as a way to help explain why some individuals may experience better or worse psychological outcomes from social media usage [44]. Although there has been considerable discussion regarding how to best operationalize and measure active and passive social media usage [45], studies typically ask participants to self-report how often they engage in a variety of active and passive activities. Active use is generally defined as interacting with other users (eg, creating posts or status updates) and directed communication (eg, sending private messages), whereas passive use refers to consuming media content without interacting with others, such as scrolling through newsfeeds [44]. In this study, media usage was assessed across 10 categories designed to capture both active and passive digital behaviors in the context of a national crisis. Participants were asked to describe the extent to which they experienced an increase in their usage of various media types since the beginning of the war (eg, “browsing news websites on a phone

XSL•FO
RenderX

or computer") [46]. Responses were rated on a 5-point Likert scale (1="not at all" to 5="to a very great extent").

Table 2 summarizes the participants' media behaviors and psychological characteristics during the initial period of the war.

Table 2. Participants' media behaviors and psychological characteristics. Mean media usage frequencies and psychological characteristics among Israeli adults (N=512) surveyed during the Israel–Hamas war (January 2024). Media usage was rated on a 5-point Likert scale (1="not at all" to 5="to a very great extent"). Psychological characteristics are presented as medians with 95% CIs.

Variable	Value
Media usage, mean (95% CI)	
Browsing news websites on a phone or computer	4.04 (3.9-4.1)
Browsing social media	3.83 (3.7-3.9)
Sharing posts on social media	2.79 (2.6-2.9)
Joining new groups on WhatsApp	2.62 (2.5-2.7)
Signing digital petitions	2.41 (2.3-2.5)
Transferring money for donations through digital platforms	2.41 (2.2-2.4)
Accessing the digital banking website	2.30 (3.1-3.4)
Accessing government and public authority websites	3.27 (2.7-3.00)
Listening to podcasts	2.86 (2.2-2.5)
Watching videos	2.34 (3.4-3.6)
Total media usage	3.00 (2.9-3.06)
Psychological characteristics, median (95% CI)	
Stress	2.3 (2.2-2.4)
Resilience	3.6 (3.5-3.7)
Well-being	3.0 (2.9-3.1)

Demographic Questionnaire

Participants provided information about their age, gender, employment status, type of organization they worked for, and their cyber-related knowledge. Cyber knowledge was defined as either holding a degree in computer science or working in a computer science–related field (eg, software engineering, computer science, IT, quality assurance, and software technician) [47].

Statistical Analysis

Statistical analyses were performed using a multistage analytic strategy. First, descriptive statistics were calculated for all media usage and psychological variables, including means, medians, and 95% CIs. To identify patterns of media engagement among participants, a 2-step cluster analysis was conducted. We selected a 2-step cluster analysis because it is appropriate for large datasets and enables automatic determination of the optimal number of clusters. The adequacy of the 2-step cluster analysis was also assessed using the silhouette value. Group differences between the high- and low-media engagement clusters were examined using independent samples 2-tailed *t* tests for continuous variables and chi-square tests for categorical variables. Effect sizes for between-group differences were calculated using Cohen *d*, defined as the difference between group means divided by the pooled SD. Effect sizes were interpreted using conventional benchmarks (0.2=small, 0.5=medium, and 0.8=large) [48]. In the clustering analysis, media engagement was treated as a categorical variable (low vs high), as this approach provides clinically and practically meaningful profiles and allows clearer interpretation of behavioral patterns during crisis conditions. To explore the underlying structure of media usage, an exploratory factor analysis (EFA) using principal axis factoring with Promax rotation was performed. Factor retention was based on eigenvalues greater than 1, and sampling adequacy was confirmed via the Kaiser-Meyer-Olkin (KMO) measure and Bartlett test of sphericity. Reliability analyses guided the exclusion of items with low loadings. Subsequently, hierarchical stepwise regression analyses were conducted to assess the contribution of demographic variables, types of media usage, stress, and resilience to individual well-being. Finally, to examine the hypothesized mediation and moderation effects, path analysis using structural equation modeling (SEM) was performed. Model fit was evaluated using standard indices, including chi-square, comparative fit index (CFI), Tucker-Lewis Index (TLI), standardized root-mean-square residual (SRMR), and root-mean-square error of approximation (RMSEA). Indirect effects were estimated with bias-corrected CIs to assess the significance of mediated pathways.

Ethical Considerations

The study protocol received full ethical approval from the Ethics Committee of Ariel University (institutional review board approval number AU-SOC-MZ-20231203). Before participation, respondents viewed an online information sheet describing the study aims and procedures, expected duration, potential risks and benefits, and the voluntary nature of participation.

XSL•FO
RenderX

Participants then provided electronic informed consent before accessing the questionnaire. The survey was administered anonymously, and no directly identifying information was collected. The anonymized data were stored in a restricted-access environment and were accessible only to the research team. Participants received panel-based incentives for survey completion through the survey firm's standard rewards program. The research team did not provide any direct financial compensation. No images or materials containing identifiable participant information are included in this manuscript or its supplementary files.

Results

We conducted a 2-step cluster analysis. The 2-cluster solution demonstrated an acceptable silhouette coefficient (0.30), indicating fair separation between clusters. Alternative cluster solutions, such as a 3-cluster solution, were also examined. While additional clusters could be extracted, these solutions resulted in reduced interpretability and overlapping profiles without meaningful improvement in cluster quality indices. Therefore, the 2-cluster solution was retained as the most parsimonious and theoretically meaningful representation of the data. The first cluster exhibited significantly higher media usage and was classified as the "high-media engagement" group, while the second cluster was identified as the "low-media engagement" group, as illustrated in Figure 2.

Figure 2. Box-and-whisker plots of the distribution of media engagement variables across the identified clusters.

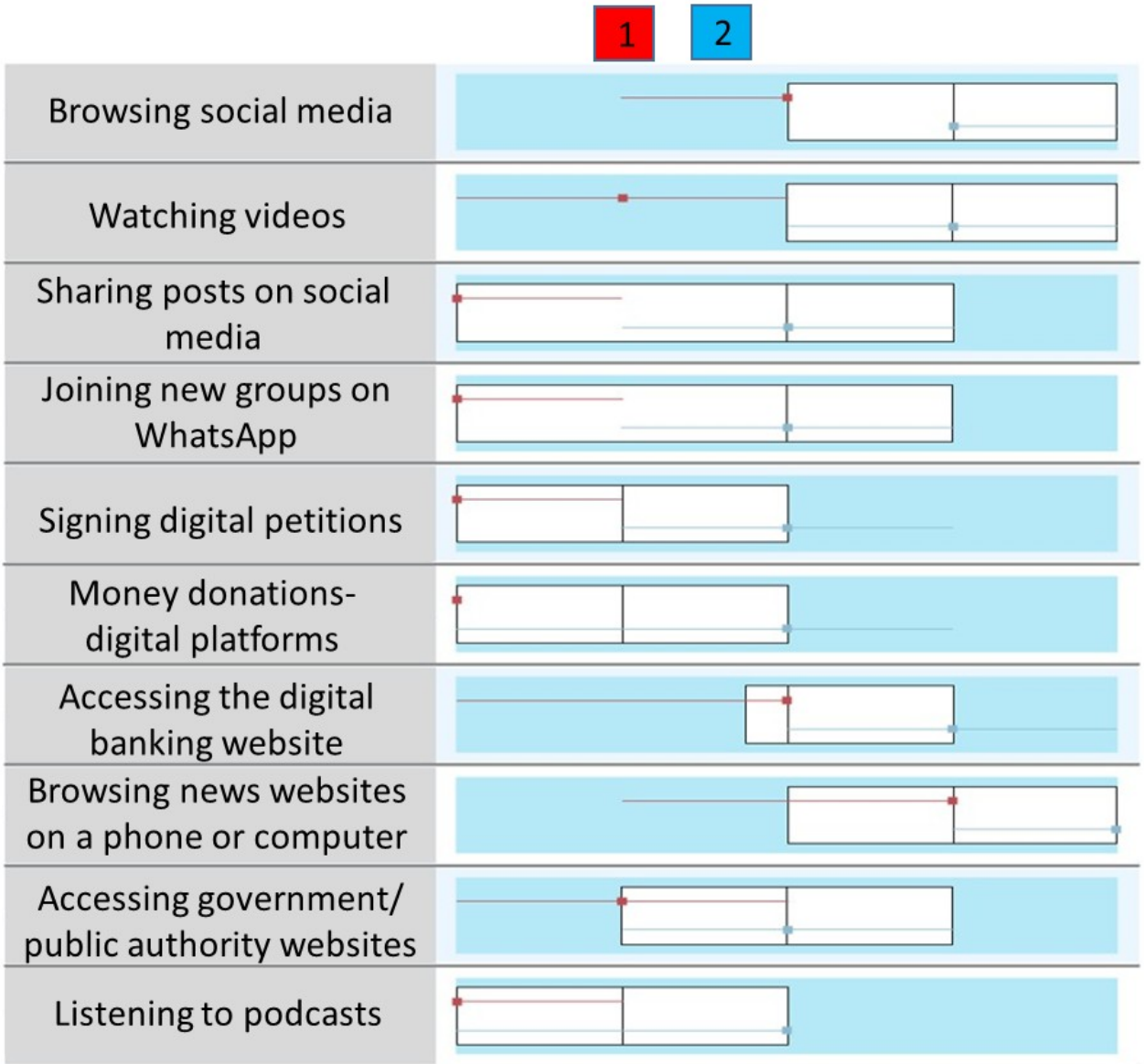

Figure 2 depicts Box-and-whisker plots displaying the distribution of 10 digital media usage behaviors across 2 engagement clusters identified via a 2-step cluster analysis among Israeli adults (N=512) surveyed during the Israel–Hamas war (January 2024). Red boxes represent the high-media engagement cluster (n=386) and blue boxes represent the low-media engagement cluster (n=126). Media usage was rated on a 5-point Likert scale (1="not at all" to 5="to a very great extent").

XSL•FO
RenderX

An independent samples *t* test and chi-square test were conducted to examine between-cluster differences in media usage behaviors (Table 3), as well as differences in demographic and psychological characteristics (Table 4).

Table 3. Between-cluster differences in media usage behaviors among Israeli adults (N=512) surveyed during the Israel–Hamas war (January 2024). A 2-step cluster analysis identified 2 engagement profiles: low-media engagement (cluster 1) and high-media engagement (cluster 2). Group differences were examined using independent samples 2-tailed t tests. Effect sizes for between-group differences were calculated using Cohen d, defined as the difference between group means divided by the pooled SD. Effect sizes were interpreted using conventional benchmarks (0.2=small, 0.5=medium, and 0.8=large) [48].

Variables composing the clusters	Cluster 1: low-media engagement (n=126), mean (95% CI)	Cluster 2: high-media engagement (n=386), mean (95% CI)	Between-group differences		
			t test (*df*)	Cohen *d*	*P* value
Browsing news websites	3.28 (3.02-3.54)	4.29 (4.20-4.38)	–9.12 (510)	–0.80	<.001
Browsing social media websites	2.72 (2.51-2.93)	4.19 (4.10-4.28)	–14.56 (510)	–1.28	<.001
Sharing posts on social media	1.52 (1.36- 1.67)	3.20 (3.06-3.34)	–12.64 (510)	–1.11	<.001
Joining new WhatsApp groups	1.57 (1.41-1.73)	2.96 (2.83-3.09)	–11.17 (510)	–0.98	<.001
Signing digital petitions	1.40 (1.26-1.53)	2.74 (2.61-2.88)	–10.66 (510)	–0.94	<.001
Transferring funds for donations via digital means	1.37 (1.23-1.50)	2.61 (2.47-2.74)	–9.80 (510)	–0.86	<.001
Browsing the digital banking websites	2.37 (2.14-2.59)	3.56 (3.44- 3.69)	–9.48 (510)	–0.83	<.001
Browsing government and public authority websites	2.17 (1.95-2.40)	3.09 (2.95-3.22)	–6.78 (510)	–0.60	<.001
Listening to podcasts	1.69 (1.49-1.88)	2.55 (2.41-2.69)	–6.41 (510)	–0.56	<.001
Watching videos online	2.29 (2.09-2.2.50)	3.96 (3.83-4.05)	–14.32 (510)	–1.26	<.001

XSL•FO
RenderX

Table 4. Between-cluster differences in demographic and psychological characteristics among Israeli adults (N=512) surveyed during the Israel–Hamas war (January 2024). A 2-step cluster analysis identified 2 engagement profiles: low-media engagement (cluster 1) and high-media engagement (cluster 2). Group differences were examined using independent samples 2-tailed t tests for continuous variables and chi-square tests for categorical variables.

Characteristic	Cluster 1: low-media engagement (n=126)	Cluster 2: high-media engagement (n=386)	Between-group differences	
			t test (*df*) or chi-square (*df*)	*P* value
Age (years)				
Mean (95% CI)	31.12 (36.81- 41.43)	37.02 (35.74-38.30)	1.59 (510)[a]	.11
Sex, n (%)			14.8 (1)[b]	<.001
Female	48.0 (38.1)	223.0 (57.8)		
Male	78.0 (61.9)	163.0 (42.2)		
Work-related characteristics, n (%)				
Occupation			0.7 (4)[b]	.67
Employee	72 (62.7)	251 (65)		
Manager	12 (9.5)	31 (8)		
Furloughed	5 (4.0)	18 (4.7)		
Unemployed	11 (8.7)	35 (9.1)		
Other	19 (15.1)	51 (13.2)		
Company type			0.2 (3)[b]	.24
Private company	63 (50)	199 (51.6)		
Public sector	25 (19.8)	101 (26.2)		
Military or security	10 (7.9)	20 (5.2)		
Unemployed	28 (22.2)	66 (17.1)		
Cyber knowledge			0.2 (1)[b]	.67
Yes	22 (17.5)	74 (19.2)		
No	104 (82.5)	312 (80.8)		
Psychological characteristics, median (95% CI)				
Stress	2.01 (1.87-2.15)	2.39 (2.31-2.47)	–4.59 (510)[a]	<.001
Resilience	3.58 (3.5-3.7)	3.58 (3.5-3.6)	0.03 (510)[a]	.98
Well-being	3.1 (2.8-3.2)	3.0 (2.9-3.1)	0.43 (510)[a]	.67

[a]Indicates a *t* test statistic.

[b]Indicates a Chi-square statistic.

The results confirmed a significant difference between participants with low-media engagement (cluster 1) during the initial stages of the war and those with high-media engagement (cluster 2) across all media usage categories. Participants in the high-media engagement cluster spent more time browsing for information and news, sending WhatsApp messages, watching videos, and listening to podcasts. They also displayed more active behaviors, such as signing petitions, making donations through websites, sharing posts on social media, and accessing government and banking sites more frequently. Overall, between-cluster differences in media usage behaviors ranged from 0.56 (listening to podcasts) to 1.28 (browsing social media), indicating moderate to large effect sizes (Table 3).

In addition to media usage, the only other characteristic that significantly differed between the 2 clusters was the level of stress. Participants in the high-media engagement cluster reported higher stress levels, indicating that higher stress was associated with greater media engagement (Table 4). The between-cluster differences in media usage and psychological variables further support the validity of the 2-cluster solution.

Our next analysis explored the connection between stress, media usage, and well-being. To examine the structure of media usage, we first conducted an EFA using principal axis factoring with oblique (Promax) rotation. Factors were identified based on eigenvalues greater than 1 [49,50]. The EFA revealed a 3-factor structure among the 10 items (Table 5). Based on previous studies, we expected a 2-cluster structure; however, the EFA indicated a 3-cluster solution. The 3 factors accounted for 61.86% of the total variance, which is considered adequate in social science research [51]. The KMO measure of sampling

XSL•FO
RenderX

adequacy was 0.77, and Bartlett test of sphericity was significant (χ^2_{36}=1007.6; *P*<.001), confirming the suitability of the data for factor analysis [52]. Factor loadings ranged from 0.57 to 0.82, indicating a strong contribution of the items to their respective factors. Ultimately, the 3 factors were identified as active media usage, passive media usage, and institutional media usage (Table 5).

Table 5. Factor analysis for the structure of media usage. An exploratory factor analysis of 10 digital media usage behaviors among Israeli adults (N=512) surveyed during the Israel–Hamas war (January 2024) was conducted using principal axis factoring with Promax rotation. Three factors were identified: active usage, passive usage, and institutional usage. Total variance explained=61.86%; KMO=0.77; Bartlett test χ236=1007.6; *P*<.001. Factor loadings below 0.40 are suppressed. "Listening to podcasts" was excluded based on low reliability.

Items	Factor 1: active usage	Factor 2: passive usage	Factor 3: institutional usage
Signing digital petitions	0.80[a]	0.18[b]	0.05[b]
Sharing posts on social media	0.75[a]	0.22[b]	0.08[b]
Joining new groups on WhatsApp	0.69[a]	0.19[b]	0.06[b]
Transferring money for donations through digital platforms	0.64[a]	0.15[b]	0.12[b]
Browsing news websites on a phone of computer	0.21[b]	0.88[a]	0.17[b]
Browsing social media	0.25[b]	0.78[a]	0.09[b]
Watching videos	0.18[b]	0.61[a]	0.14[b]
Accessing government and public authority websites	0.06[b]	0.20[b]	0.87[a]
Accessing the digital banking websites	0.09[b]	0.17[b]	0.77[a]
Cronbach α	0.71[b]	0.68[b]	0.63[b]

[a]Values indicate primary factor loadings and define item membership in each factor.

[b]Values represent cross-loadings and were not used to assign items to factors.

Next, we examined the contribution of different types of media use to individuals' well-being. Tables 6 and 7 present the results of a 3-step hierarchical regression analysis. In step 1, demographic variables were entered as control variables. In step 2, the different categories of media use were added to the model. In step 3, stress and resilience were included to assess their incremental contribution to well-being.

Table 6. Hierarchical regression results: well-being as the dependent variable. Hierarchical multiple regression analysis predicting psychological well-being (5-item World Health Organization Well-Being Index) among Israeli adults (N=512) surveyed during the Israel–Hamas war (January 2024). Model 1.1 includes demographic variables, Model 1.2 adds 3 types of media usage, and Model 1.3 adds stress and resilience. The *P* values are 2-tailed.

Variables	Model 1.1		Model 1.2		Model 1.3	
	β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value
Step 1: demographic variables						
Sex (1-male, 2-female)	–0.17 (–0.52 to –0.18)	.001	–0.17 (–0.51 to –0.16)	>.001	–0.08 (–0.31 to 0.01)	.04
Age	–0.19 (–0.02 to –0.01)	>.001	–0.19 (–0.02 to –0.01)	>.001	–0.25 (–0.02 to –0.01)	>.001
Step 2: media usage						
Active usage	__[a]	—	0.05 (–0.04 to 0.14)	.31	0.12 (0.03 to 0.20)	<.001
Passive usage	—	—	–0.09 (–0.19 to 0.01)	.07	–0.07 (–0.16 to 0.02	.08
Institutional usage	—	—	0.10 (0.00 to 0.19)	.04	0.06 (–0.00 to 0.14)	.11
Step 3: psychological characteristics						
Stress	—	—	—	—	–0.29 (–0.46 to –0.25)	<.001
Resilience	—	—	—	—	0.32 (0.34 to 0.56)	<.001

[a]Not applicable.

XSL•FO
RenderX

Table 7. Hierarchical regression model fit statistics for psychological well-being.

Statistic	Model 1.1	Model 1.2	Model 1.3
R^2	0.064	0.07	0.26
ΔR^2	—[a]	0.02[b]	0.20[b]
F test (*df*)	18.42 (2, 509)[b]	8.82 (5, 506)[b]	28.78 (7, 504)[b]

[a]Not applicable/Not available.

[b]Indicates *P*<.001.

Model 1.1 indicates that females and younger individuals reported higher well-being compared with males and older participants (β=−0.17, 95% CI −0.52 to −0.18; β=−0.19, 95% CI −0.02 to −0.01, respectively; both *P*<.001). Adding media usage variables in Model 1.2 revealed that institutional media usage was positively associated with well-being (β=0.10, 95% CI 0.00 to 0.19; *P*=.04), whereas passive media usage showed a marginally negative association (β=−0.09, 95% CI −0.19 to 0.01; *P*=.07). Active media usage was not significantly associated with well-being at this stage (β=0.05, 95% CI −0.04 to 0.14; *P*=.31). Including psychological variables substantially improved the model, increasing the explained variance from 7% to 26% (ΔR^2=0.20, $F_{7,504}$=28.78; *P*<.001). Both stress and resilience significantly predicted well-being (β=−0.29, 95% CI −0.46 to −0.25; β=0.32, 95% CI 0.34-0.56, respectively; both *P*<.001). In addition, active media usage became a significant positive predictor of well-being (β=0.12, 95% CI 0.03-0.20; *P*<.001). Overall, most statistically significant effects were accompanied by CIs that did not include zero, indicating reliable associations. In contrast, the CIs for passive media usage crossed zero, and the interval for institutional media usage approached zero, suggesting greater uncertainty regarding these relationships.

Lastly, our main hypothesis proposed that media usage moderates the relationship between resilience, stress, and well-being. To test this hypothesis, we conducted path analyses. The model specified the relationships among resilience, stress, active media usage, passive media usage, institutional media usage, and participants' well-being. Detailed results of the path analyses are presented in Table 8.

Table 8. Path analysis results examined the relationships among resilience, stress, media usage (active, passive, and institutional), and well-being among Israeli adults (N=512) surveyed during the Israel–Hamas war (January 2024). The model was estimated using structural equation modeling. Unstandardized coefficients (B) are reported with SE and 95% CIs. Indirect effects were estimated using bias-corrected bootstrap CIs. Model fit: χ21=1.5; *P*=.21; comparative fit index=0.999; Tucker-Lewis Index=0.981; standardized root-mean-square residual=0.013; root-mean-square error of approximation=0.033. The *P* values are 2-tailed.

Path	B (95% CI)	SE	*P* Value
Direct effects			
Resilience → stress	–0.28 (–0.38 to –0.18)	0.05	<.001
Stress → active media usage	0.33 (0.23 to 0.43)	0.05	<.001
Stress→ passive media usage	0.37 (0.27 to 0.47)	0.05	<.001
Stress → institutional media usage	0.20 (0.10 to 0.30)	0.05	<.001
Active media usage → well-being	0.12 (0.04 to 0.20)	0.04	<.001
Passive media usage → well-being	–0.08 (–0.18 to 0.02)	0.05	.08
Institutional media usage → well-being	0.07 (–0.01 to 0.15)	0.04	.11
Resilience → well-being	0.43 (0.31 to 0.55)	0.05	<.001
Stress → well-being	–0.33 (–0.43 to –0.23)	0.05	<.001
Indirect effects (bootstrap)			
Resilience → stress→ media usage→ well-being	0.09 (0.05 to 0.13)	—[a]	—
Resilience → stress→ active media usage→ well-being	–0.01 (–0.02 to –0.01)	—	—
Resilience → stress→ passive media usage→ well-being	0.01 (–0.01 to 0.09)	—	—
Resilience → stress→ institutional media usage→ well-being	–0.01 (–0.02 to –0.01)	—	—

[a]Not applicable.

The findings revealed an acceptable fit of the data (χ^2_1=1.54; *P*=.21; CFI=0.999; TLI=0.981; SRMR=0.013; RMSEA=0.033). Resilience was negatively associated with stress (β=–0.24, SE=0.05; *P*<.001) and positively associated with well-being (β=0.30; SE=0.06; *P*<.001). Stress was positively associated with active media usage (β=0.27; SE=0.05; *P*<.001), passive media usage (β=0.31; SE=0.05; *P*<.001), and institutional media usage (β=0.16; SE=0.05; *P*<.001). These results indicate that

XSL•FO
RenderX

higher levels of stress led participants to engage more frequently in various types of media usage. Of the 3 types of media usage, only active media usage was positively and significantly related to well-being (β=0.12; SE=0.04; *P*=.006). In contrast, passive media usage showed a negative association with well-being, approaching significance (β=–0.08; SE=0.05; *P*=.08). In conclusion, participants who engaged more in active media behaviors reported higher levels of well-being. Conversely, those who primarily engaged in passive media usage experienced a reduction in well-being.

The indirect relationship between resilience and well-being through stress and the 3 types of media usage was significant (indirect effect=0.09, 95% bias-corrected CI 0.05-0.13). However, when examining each path separately, we found that the source of the indirect connection was mediated specifically through active media behaviors (indirect effect=–0.01, 95% bias-corrected CI –0.02 to –0.01), while the other paths were not significant. These findings suggest that, in addition to the direct connection between resilience and well-being, resilience was negatively associated with stress. This stress reduction was associated with the use of active media behaviors, which positively impacted participants' well-being. Figure 3 summarizes the model results.

Figure 3 depicts a path diagram of the SEM examining the relationships among resilience, stress, 3 types of media usage (active, passive, and institutional), and psychological well-being among Israeli adults (N=512) surveyed during the Israel–Hamas war (January 2024). Solid lines represent direct paths; dashed lines represent indirect (mediated) paths. Unstandardized coefficients (B) are reported for each path, with SEs in parentheses. Indirect effects are presented as standardized coefficients (β; Model fit: χ^2_1=1.5; *P*=.21; CFI=0.999; TLI=0.981; SRMR=0.013; RMSEA=0.033).

Figure 3. Results of the structural equation model. Unstandardized coefficients (B) are shown, with SEs in parentheses. BCCI: Bais corrected confidence interval.

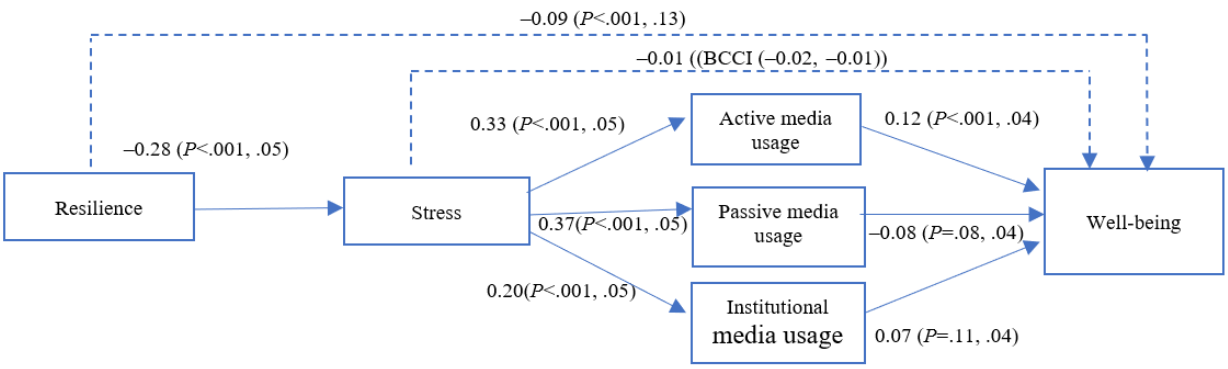

Discussion

Principal Findings

This study highlights the complex interplay between digital media engagement, psychological distress, and well-being during a national crisis. The findings show that high media engagement during wartime was associated with increased stress levels. This is consistent with the growing literature on "doomscrolling," in which compulsive consumption of negative news has been associated with heightened anxiety [53]. Although seeking information is a natural response to uncertainty, the "vicious cycle" of constant monitoring may result in emotional exhaustion rather than relief [29]. Consistent with this broader pattern, higher exposure to COVID-19–related media information was associated with greater distress [18]. This pattern suggests that during times of crisis, media use may serve as a coping mechanism to manage uncertainty and maintain perceived control, a phenomenon also observed during the COVID-19 pandemic and in conflict-related contexts [18,19].

According to PMT, individuals adopt protective behaviors based on a threat- and coping-appraisal process [54]. However, our data suggest that during acute crises, heightened distress may be associated with increased digital engagement even when such engagement does not necessarily translate into psychological relief [15,18-20]. Viewed through PMT, stress can be interpreted as a proxy for heightened threat-related activation; our finding that stress was associated with greater engagement aligns with the general PMT expectation that perceived threat can increase motivation to act or seek coping responses. At the same time, the co-occurrence of elevated distress with both active and passive engagement patterns suggests that threat-related activation alone does not guarantee more adaptive outcomes, highlighting the potential importance of coping resources when interpreting crisis-related engagement. Accordingly, we interpret this pattern as increased engagement under distress, rather than as evidence of improved security behavior or decreased vigilance. Furthermore, prior research during population-level public health crises (eg, the COVID-19 pandemic) has examined links between anxiety and perceived efficacy of health-protective behaviors [16]. Related work has highlighted that exposure to the internet and media sources can be intertwined with health anxiety, health literacy, and

XSL•FO
RenderX

information-seeking processes, including mediated pathways to downstream self-care behaviors [55]. Conceptual accounts suggest that media-induced uncertainty may sustain cycles of worry and repeated monitoring [29]. Empirical work using the Doomscrolling Scale indicates that doomscrolling is associated with psychological distress and poorer well-being, with distress mediating the relationship between doomscrolling and well-being indicators [56]. Continuous consumption of negative news, or doomscrolling, can be detrimental to mental health, especially when combined with intolerance of uncertainty and low psychological resilience [53].

The cluster analysis revealed 2 distinct engagement profiles: high and low media users. High engagement was associated with elevated stress levels and increased activity across digital platforms, particularly news consumption, social media, and online financial transactions. This pattern aligns with evidence from population-level crises showing that greater exposure to COVID-19–related media information is associated with higher psychological distress [18], and that social media use during the COVID-19 outbreak was associated with depression and secondary trauma [17]. In health-information contexts, exposure to the internet and media sources has also been linked to information-seeking and health anxiety processes [55]. This intensified reliance is consistent with the observation that the constant stream of updates can generate a sense of urgency and a feeling of permanent connectedness that has been linked to stress experiences [40]. Problematic smartphone-related experiences, such as nomophobia, have also been described among social media users [41]. Finally, qualitative work in individuals affected by social anxiety disorder suggests that unregulated digital engagement (eg, endless scrolling and information overload) may undermine subjective digital well-being [57].

Our findings further demonstrate that the type of media usage plays a critical role in associations with psychological outcomes. Active media use, characterized by interactive behaviors such as sharing posts or joining online groups, was positively associated with well-being, while passive media use (eg, browsing without interaction) showed a negative trend. These results align with previous studies suggesting that active media engagement may support coping and emotion regulation, whereas passive use may be associated with greater distress and emotional disengagement [58-60]. Specifically, active social media use has been linked to perceived online social support in some contexts [44]. Conversely, passive consumption has been associated with anxiety and negative self-image due to an increased likelihood of social comparison with peers [58,60]. More broadly, exposure to distressing or graphic crisis-related media content may amplify distress [7]. Institutional media usage (eg, accessing official or organizational channels) had a neutral to slightly positive effect, likely reflecting its role in fulfilling informational needs without causing emotional overload; however, the benefits of social networking services for information use can be attenuated by social media stress [61].

Path analysis confirmed that stress was significantly associated with all 3 types of media engagement. However, the indirect effects indicated that resilience was linked to well-being primarily through lower stress, while the specific indirect paths through media engagement types were small. These findings support the view that resilience is associated with better well-being and may buffer the effects of stress during crises, while engagement patterns may be differentially related to psychological outcomes [31,33,62]. Resilience is broadly defined as the fundamental human ability to “bounce back” or adapt successfully after exposure to a crisis or stressful situation [43]. Resilience has been implicated as a factor that may mitigate maladaptive engagement patterns, including doomscrolling, particularly under conditions of high anxiety and intolerance of uncertainty [53]. Moreover, strengthening resilience at individual and societal levels is recognized as an important strategy for coping with large-scale crises and disasters [31,63].

Importantly, cyber-related knowledge did not significantly differentiate between the 2 engagement clusters. This suggests that during acute crises, emotional states (particularly stress) may play a more prominent role than technical background variables in shaping digital media engagement patterns. Consistent with scholarship on anxiety and decision-making, intense anxiety or panic may impair deliberative decision-making and coherent preference ordering [30], and high-arousal negative emotions (eg, fear and anger) can bias risk judgments and subsequent choices [64]. A critical implication of this emotional overload is the onset of “cybersecurity fatigue,” a state of mental and emotional exhaustion that contributes to burnout and increased psychological strain, potentially leading users to consciously or unconsciously neglect essential security protocols [65,66]. Similar findings have been reported in the cybersecurity behavior literature, where perceived vulnerability and emotional responses influenced behavior more strongly than technical knowledge alone [67,68]. This underscores the importance of integrating emotional and psychological components into digital literacy programs, especially for populations at risk of high exposure to crisis media.

From a theoretical perspective, these findings can be interpreted using PMT as a conceptual lens. While PMT traditionally focuses on cognitive threat and coping appraisals [69], our results highlight associations between stress and digital media engagement, showing that higher stress was associated with greater engagement across types, whereas engagement patterns showed different associations with well-being. Accordingly, we avoid concluding that stress motivates protective or maladaptive behaviors and instead describe these as observed associations that may vary by context. Prior PMT-informed cybersecurity research illustrates that protective behavior is typically framed as something that can be shaped through modifiable organizational levers, such as clear policies and security education, training, and awareness programs, which are intended to strengthen coping appraisals and translate perceived threat into protective action [70]. PMT reviews further emphasize that threat appraisal alone is insufficient; protective responding is expected when coping appraisal, such as perceived efficacy and self-efficacy, is strong relative to perceived costs [69]. Interpreted through this lens, our finding that higher stress co-occurred with both active and passive engagement patterns suggests that crisis-related arousal may broadly increase

XSL•FO
RenderX

engagement, whereas coping resources, such as resilience, may be particularly important in determining whether engagement is linked to more favorable psychological outcomes. More broadly, crisis-specific frameworks emphasize the role of information and communication technology–mediated information environments in crisis management and communication [71]. In this study, resilience was associated with lower stress and higher well-being, suggesting that coping resources are relevant when interpreting crisis-related engagement and outcomes [43]. In the context of doomscrolling, intolerance of uncertainty and resilience-related processes have been discussed as relevant correlates [53]. Overall, these findings support a cautious, theory-informed interpretation without implying causal mechanisms or model refinement beyond the variables measured in this study.

The findings have practical implications for designing public health and digital communication strategies during emergencies. Campaigns should encourage intentional and participatory media use, such as verifying sources, seeking community connection, and sharing factual information, while raising awareness about the mental health risks of excessive passive consumption. Platforms could be adapted to promote active engagement by prompting users toward verified resources or facilitating moderated peer-support communities. Educational and organizational programs that combine emotion regulation, media literacy, and resilience-building may serve as protective interventions, particularly for populations frequently exposed to distressing crisis-related media content. Evidence syntheses suggest that media literacy programs can strengthen critical evaluation of information and reduce susceptibility to misinformation [72]. A recent meta-analysis found that media literacy interventions improved fake news credibility assessment, suggesting that strengthening credibility judgment skills through structured media literacy training may support safer digital decision-making during crises, including decisions about sharing, trusting, or acting on rapidly circulating claims [73]. The development of digital tools, including mobile apps to support mental health response in disasters and remotely delivered mental health services, may strengthen mental health support and continuity of care when in-person services are disrupted by extending access to support across preparedness, response, and recovery phases, particularly when demand is high and resources are constrained [74,75].

Strengths and Limitations

This study presents several important strengths along with acknowledged limitations. A key strength is its focus on the understudied intersection of cybersecurity-related behavior, media engagement, and psychological well-being during an active national crisis. Moreover, the contribution of this study lies in its unique context, which enabled the observation of digital behaviors during an acute and ongoing national crisis. While previous research has explored media consumption during public health crises such as the COVID-19 pandemic [58] and has documented the psychological toll of exposure to graphic war-related content [7], this study extends this literature in several important ways.

First, rather than treating media usage as a unidimensional construct, we empirically identified 3 distinct patterns of digital engagement: active, passive, and institutional usage, each with different implications for psychological well-being. Second, by embedding these patterns within a stress-resilience framework, we demonstrated that the relationship between media engagement and well-being is not uniform but depends critically on the type of engagement. Specifically, active participation in digital platforms was associated with better well-being, consistent with the view that such behaviors may serve as a form of emotion-focused coping and social support-seeking during acute crisis [76]. In contrast, passive consumption showed a negative trend, aligning with prior findings on doomscrolling and emotional exhaustion [53]. These findings highlight the importance of moving beyond aggregate measures of screen time or media exposure and instead attending to the qualitative nature of digital engagement during wartime.

A further contribution of this study lies in its broader conceptualization of digital media usage. Unlike previous studies, which have typically focused on social media platforms (eg, Abbas et al [77]; Huckins et al [78]; Gao et al [79]), this study also examined institutional digital behaviors, including online banking and access to government and public authority websites. In an era of accelerating digitalization, digital applications are no longer limited to entertainment and social interaction, they have increasingly become the primary interface between individuals and institutional agents, such as financial institutions and governmental services. This shift may have important implications for cybersecurity, as greater reliance on digital institutional channels during crisis periods may heighten individuals' vulnerability to cyberattacks. Enhanced cybersecurity awareness and digital literacy are therefore particularly critical during times of crisis, such as war or a pandemic, when the availability of in-person institutional contact is restricted and individuals are compelled to rely more heavily on digital alternatives. In this study, we therefore expand the scope of inquiry beyond social media, examining whether individuals turn to digital applications not only to seek information and social connection but also to manage practical and institutional needs during wartime.

While the findings provide important insights, several limitations should be considered when interpreting the results. First, the cross-sectional design precludes causal inference [80]. Although the path analysis model implies directionality, the temporal order between stress, media engagement, and well-being cannot be established from a single time-point measurement. All associations reported in this study should therefore be interpreted as correlational rather than causal. Longitudinal or experimental designs are needed to establish whether stress causally drives media engagement or whether intensive media use itself amplifies stress over time.

Second, cybersecurity awareness and knowledge were assessed using a single self-report item. In addition, media usage was operationalized directly and did not rely on previously validated instruments. While this approach captures behavior directly, this operationalization may not fully reflect the multidimensional nature of media usage awareness, which may also involve more implicit or less visible patterns of engagement. Future studies

XSL•FO
RenderX

should incorporate validated measures to enable more rigorous testing of their role in shaping digital behavior during crises.

Third, the study did not measure potential confounding variables that may influence both media engagement and well-being, such as prior mental health history, social support networks, proximity to conflict zones, or direct exposure to wartime events. The observed associations between stress, media use, and well-being may therefore partly reflect the influence of unmeasured third variables rather than the relationships among the constructs themselves.

Lastly, participants were recruited through an online panel using quota sampling, which, while designed to approximate the demographic distribution of the Israeli population, does not constitute a probability sample. Self-selection bias [81] may have resulted in the overrepresentation of individuals who are more digitally engaged or more willing to report their media behaviors and psychological states. In addition, previous studies suggest that cybersecurity behaviors differ across cultures [82,83]. Accordingly, the findings should be interpreted with caution and may not be fully generalizable to populations with lower digital access or different media consumption patterns. In this vein, it should also be noted that the study was conducted within a single national context during the early months of the Israel–Hamas war, which limits the external validity of the findings. The specific cultural, political, and media environment of this conflict may not be representative of other crisis contexts, such as natural disasters, pandemics, or conflicts in different geopolitical settings. Replication across diverse crisis contexts and cultural settings is therefore required before broader conclusions regarding the relationship between stress, media engagement, and well-being can be drawn. Future research should build on these findings by examining the dynamic relationships among stress, media engagement, and well-being using longitudinal designs and more comprehensive measures of digital behavior.

Future Directions

Future research should test the direction and mechanisms underlying the observed associations by using longitudinal and experience-sampling designs that capture within-person changes in stress, resilience, and engagement over time and across crisis phases. Studies that combine self-reports with objective indicators of digital behavior (such as passive sensing or platform-provided usage metrics) and more detailed characterization of content (such as exposure to graphic material, misinformation, or institutional alerts) may help distinguish when engagement is adaptive vs burdensome. It is also important to examine heterogeneity across subgroups (ie, age, prior trauma, frontline exposure, and baseline mental health) and across platforms and engagement modes, including institutional channels that are often overlooked. Finally, intervention-focused research should evaluate targeted, real-world strategies (ie, media literacy and resilience-building components embedded in crisis communication, platform-based prompts toward verified sources, and moderated peer-support pathways) to determine which approaches most effectively promote healthier engagement and protect well-being during prolonged crises.

Conclusions

This study offers new evidence on wartime digital media engagement by moving beyond time-based exposure measures and empirically identifying distinct engagement patterns, including active, passive, and institutional modes, which are rarely examined together in crisis-media research. Compared with previous work that often focuses on overall media exposure or social media alone, our findings offer a more detailed account of how different forms of engagement relate to stress, resilience, and well-being during an ongoing armed conflict. This contribution advances the field by presenting an empirically grounded framework for interpreting when engagement may be supportive vs potentially burdensome in prolonged crises. Practically, the results can inform public health and organizational communication by emphasizing targeted guidance that encourages adaptive, intentional engagement (ie, seeking verified information and social support) while reducing risks from excessive passive exposure. They can also help platforms and service providers direct users to reliable institutional information and supportive resources.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request, subject to approval by the Ethics Committee of Ariel University.

Funding

The authors declared no financial support was received for this work.

Authors' Contributions

MZ, NS, RT, and AZ contributed to the visualization and writing of the original draft. MZ, GK, and SB were responsible for data curation, while GK and SB performed the formal analysis. NS and RT conducted the investigation. MZ and GK developed the methodology. AZ handled project administration, and MZ and AZ provided supervision. All authors reviewed and commented on subsequent drafts and approved the final version for publication. The authors confirm that no generative artificial intelligence (AI) tools were used to generate the text, figures, or other content of this manuscript; AI was used solely for language editing and grammatical improvement.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Zarnadze A. “Invisible Bullets”: the power of narratives in modern warfare. Glob Pol 2025;16(2):419-422. [doi: 10.1111/1758-5899.70018]
2. Dvornyk M, Sliusarevskyi M. Digital well-being of Ukrainians experiencing full-scale war: a cross-sectional study. Sci Stud Soc Polit Psychol 2025:16-24. [doi: 10.61727/ssspj/1.2025.16]
3. Norquist J. “Something other than real life:” digital life resistance in the civil sphere. Am J Cult Sociol 2022;12(1):115-137. [doi: 10.1057/s41290-022-00176-z]
4. Zvonareva O. Participation in times of war: the ambivalence of digital media. Soc Incl 2025;13:9128. [doi: 10.17645/si.9128]
5. Kertcher C, Turin O. Disruptive media event in a divided society: the case of October 7 atrocity videos in Israel. Soc Media Soc 2025;11(1). [doi: 10.1177/20563051251328105]
6. Alagarsamy S. Invisible wounds: vicarious trauma and the mental health fallout of modern wars. Br J Psychiatry 2026:1-2. [doi: 10.1192/bjp.2025.10529] [Medline: 41502240]
7. Holman EA, Garfin DR, Silver RC. It matters what you see: graphic media images of war and terror may amplify distress. Proc Natl Acad Sci U S A 2024;121(29):e2318465121 [FREE Full text] [doi: 10.1073/pnas.2318465121] [Medline: 38968094]
8. Yamin D, Lev-Ari S, Mofaz M, Elias R, Toker S, Spiegel D. Social media impact and smartwatch monitoring: prevalence and early markers of PTSD and anxiety following mass traumatic events. PLOS Ment Health 2025;2(9):e0000195 [FREE Full text] [doi: 10.1371/journal.pmen.0000195] [Medline: 41661960]
9. Godard R, Holtzman S. Are active and passive social media use related to mental health, wellbeing, and social support outcomes? A meta-analysis of 141 studies. J Comput Mediat Commun 2023;29(1). [doi: 10.1093/jcmc/zmad055]
10. Nagori V, Dey BL. Self-engagement on social media as a digital intervention to stress coping. J Enterp Inf Manag 2026:1-20. [doi: 10.1108/jeim-09-2025-0911]
11. Măirean C, Cobzeanu A, Dănilă O. Breaking news, breaking minds? Social media engagement and peritraumatic stress behavior immediately after the Russian – Ukrainian war outbreak: the role of peritraumatic anxiety and dispositional resilience. Comput Hum Behav Rep 2025;20(100808):100808. [doi: 10.1016/j.chbr.2025.100808]
12. Gazit T, Eden S. From screens to scars: understanding the association between social media engagement and trauma during crises and emergency situations. Stress Health 2025;41(6):e70115. [doi: 10.1002/smi.70115] [Medline: 41199140]
13. Gülırmak Güler K, Albayrak Günday E. Fronts in minds: a phenomenological study on the effects of war news on collective mental health. Public Health Nurs 2025;42(1):483-493. [doi: 10.1111/phn.13458] [Medline: 39468407]
14. Holman EA, Garfin DR, Lubens P, Silver RC. Media exposure to collective trauma, mental health, and functioning: does it matter what you see? Clin Psychol Sci 2019;8(1):111-124. [doi: 10.1177/2167702619858300]
15. Garfin DR, Silver RC, Holman EA. The novel coronavirus (COVID-2019) outbreak: amplification of public health consequences by media exposure. Health Psychol 2020;39(5):355-357. [doi: 10.1037/hea0000875] [Medline: 32202824]
16. Vally Z. Public perceptions, anxiety and the perceived efficacy of health-protective behaviours to mitigate the spread of the SARS-Cov-2/ COVID-19 pandemic. Public Health 2020;187:67-73. [doi: 10.1016/j.puhe.2020.08.002] [Medline: 32927291]
17. Zhong B, Huang Y, Liu Q. Mental health toll from the coronavirus: social media usage reveals wuhan residents' depression and secondary trauma in the COVID-19 outbreak. Comput Human Behav 2021;114:106524. [doi: 10.1016/j.chb.2020.106524] [Medline: 32836728]
18. Yao H. The more exposure to media information about COVID-19, the more distressed you will feel. Brain Behav Immun 2020;87:167-169. [doi: 10.1016/j.bbi.2020.05.031] [Medline: 32413557]
19. Malecki WP, Bilandzic H, Kowal M, Sorokowski P. Media experiences during the Ukraine war and their relationships with distress, anxiety, and resilience. J Psychiatr Res 2023;165:273-281. [doi: 10.1016/j.jpsychires.2023.07.037] [Medline: 37549502]
20. Castro M, Aires Dias J, Madeira L. Does the media (also) keep the score? Media-based exposure to the Russian-Ukrainian war and mental health in Portugal. J Health Psychol 2024;29(13):1475-1488 [FREE Full text] [doi: 10.1177/13591053231201242] [Medline: 37840303]
21. Verduyn P, Lee DS, Park J, Shablack H, Orvell A, Bayer J. Passive facebook usage undermines affective well-being: experimental and longitudinal evidence. J Exp Psychol Gen 2015;144(2):480-488. [doi: 10.1037/xge0000057] [Medline: 25706656]
22. Valkenburg PM, van Driel II, Beyens I. The associations of active and passive social media use with well-being: a critical scoping review. New Media & Society 2021;24(2):530-549. [doi: 10.1177/14614448211065425]
23. Valkenburg PM. Social media use and well-being: what we know and what we need to know. Curr Opin Psychol 2022;45:101294 [FREE Full text] [doi: 10.1016/j.copsyc.2021.12.006] [Medline: 35016087]

XSL•FO
RenderX

24. Zigdon A, Zigdon T, Moran DS. Attitudes of nurses towards searching online for medical information for personal health needs: cross-sectional questionnaire study. J Med Internet Res 2020;22(3):e16133 [FREE Full text] [doi: 10.2196/16133] [Medline: 32175910]
25. Zigdon A, Zwilling M, Zigdon O, Reges O. Health maintenance organization-mhealth versus face-to-face interaction for health care in Israel: cross-sectional web-based survey study. J Med Internet Res 2024;26:e55350 [FREE Full text] [doi: 10.2196/55350] [Medline: 39348674]
26. Ben-Zur H, Gil S, Shamshins Y. The relationship between exposure to terror through the media, coping strategies and resources, and distress and secondary traumatization. Int J Stress Manag 2012;19(2):132-150. [doi: 10.1037/a0027864]
27. Levi-Belz Y, Groweiss Y, Blank C, Neria Y. PTSD, depression, and anxiety after the October 7, 2023 attack in Israel: a nationwide prospective study. EClinicalMedicine 2024;68:102418. [doi: 10.1016/j.eclinm.2023.102418] [Medline: 38586476]
28. Zhang YT, Li RT, Sun XJ, Peng M, Li X. Social media exposure, psychological distress, emotion regulation, and depression during the COVID-19 outbreak in community samples in China. Front Psychiatry 2021;12:644899 [FREE Full text] [doi: 10.3389/fpsyt.2021.644899] [Medline: 34054602]
29. Kesner L, Juríčková V, Grygarová D, Horáček J. Impact of media-induced uncertainty on mental health: narrative-based perspective. JMIR Ment Health 2025;12:e68640. [doi: 10.2196/68640] [Medline: 40513096]
30. Khalil EL. Anxiety and rationality: Allais paradox, procrastination, Keynesian expectations, and other anxiety-based deviations from rationality. Humanit Soc Sci Commun 2025;12(1). [doi: 10.1057/s41599-025-05552-x]
31. Schäfer SK, Supke M, Kausmann C, Schaubruch LM, Lieb K, Cohrdes C. A systematic review of individual, social, and societal resilience factors in response to societal challenges and crises. Commun Psychol 2024;2(1):92. [doi: 10.1038/s44271-024-00138-w] [Medline: 39369098]
32. Fekih-Romdhane F, Helmy M, Alhuwailah A, Shuwiekh HAM, Maalej E, Naser AY. Resilience and religiosity as moderators between war-related media exposure and psychotic experiences in the aftermath of the October 2023 Israel-Gaza war. Discov Public Health 2025;22(1). [doi: 10.1186/s12982-025-00599-2]
33. Pauly C, Ribeiro F, Schröder VE, Pauly L, Krüger R, Leist AK. The moderating role of resilience in the personality-mental health relationship during the COVID-19 pandemic. Front Psychiatry 2021;12:745636 [FREE Full text] [doi: 10.3389/fpsyt.2021.745636] [Medline: 34744837]
34. Malin Y, Gardyn Y, Asterhan CS. Mental health during war: social media use and protective factors among adolescents and young adults. Comput Hum Behav 2026;180:108936. [doi: 10.1016/j.chb.2026.108936]
35. Balla J, Hagger MS. Protection motivation theory and health behaviour: conceptual review, discussion of limitations, and recommendations for best practice and future research. Health Psychol Rev 2025;19(1):145-171 [FREE Full text] [doi: 10.1080/17437199.2024.2413011] [Medline: 39420632]
36. Hedayati S, Damghanian H, Farhadinejad M, Rastgar AA. Meta-analysis on application of protection motivation theory in preventive behaviors against COVID-19. Int J Disaster Risk Reduct 2023:103758 [FREE Full text] [doi: 10.1016/j.ijdrr.2023.103758] [Medline: 37359108]
37. Zhang D, Su F, Meng X, Zhang Z. Impact of media trust and personal epidemic experience on epidemic prevention behaviors in the context of COVID-19: a cross-sectional study based on protection motivation theory. Front Public Health 2023;11:1137692 [FREE Full text] [doi: 10.3389/fpubh.2023.1137692] [Medline: 37124803]
38. Mohammed F, Al-Kumaim NH, Alzahrani AI, Fazea Y. The impact of social media shared health content on protective behavior against COVID-19. Int J Environ Res Public Health 2023;20(3):1775 [FREE Full text] [doi: 10.3390/ijerph20031775] [Medline: 36767142]
39. The world health organisation-five well-being index (WHO-5). World Health Organization. URL: https://www.psykiatri-regionh.dk/who-5/Documents/WHO-5_Hebrew.pdf [accessed 2026-04-04]
40. Freytag A, Knop-Huelss K, Meier A, Reinecke L, Hefner D, Klimmt C, et al. Permanently online-always stressed out? The effects of permanent connectedness on stress experiences. Hum Commun Res 2020;47(2):132-165. [doi: 10.1093/hcr/hqaa014]
41. Balcioğlu YS, Artar Biyiklar M. Analyzing nomophobia among TikTok users: a data mining and text analysis approach. Curr Psychol 2025;44(23):18220-18242. [doi: 10.1007/s12144-025-08416-4]
42. Dwight AR, Briesch AM, Hoffman JA, Rutt C. Systematic review of the psychometric evidence supporting use of the Depression Anxiety Stress Scales, Short Form (DASS-21) with youth. Child Youth Care Forum 2024;53(5):1235-1250. [doi: 10.1007/s10566-024-09795-8]
43. Campbell-Sills L, Stein MB. Psychometric analysis and refinement of the Connor-Davidson Resilience Scale (CD-RISC): Validation of a 10-item measure of resilience. J Trauma Stress 2007;20(6):1019-1028. [doi: 10.1002/jts.20271] [Medline: 18157881]
44. Frison E, Eggermont S. Exploring the relationships between different types of Facebook use, perceived online social support, and adolescents' depressed mood. Soc Sci Comput Rev 2015;34(2):153-171. [doi: 10.1177/0894439314567449]
45. Trifiro BM, Gerson J. Social media usage patterns: research note regarding the lack of universal validated measures for active and passive use. Soc Media Soc 2019;5(2):205630511984874. [doi: 10.1177/2056305119848743]
46. Perez O, Garza T, Hindera O, Beltran A, Musaad SM, Dibbs T. Validated assessment tools for screen media use: a systematic review. PLoS One 2023;18(4):e0283714. [doi: 10.1371/journal.pone.0283714] [Medline: 37053175]

XSL•FO
RenderX

47. Waterman AS, Schwartz SJ, Zamboanga BL, Ravert RD, Williams MK, Agocha VB. The questionnaire for eudaimonic well-being: psychometric properties, demographic comparisons, and evidence of validity. J Posit Psychol 2010;5(1):41-61 [FREE Full text] [doi: 10.1080/17439760903435208] [Medline: 34326891]
48. Cohen J. Statistical Power Analysis for the Behavioral Sciences. New York: Routledge; 2013.
49. Ahmed A, Maruod W. Comparative analysis of varimax and promax rotation methods in exploratory factor analysis. Edelweiss Appl Sci Technol 2025;9(5):501-513. [doi: 10.55214/25768484.v9i5.6929]
50. Apolinário-Hagen J, Vehreschild V, Alkoudmani RM. Current views and perspectives on e-mental health: an exploratory survey study for understanding public attitudes toward internet-based psychotherapy in Germany. JMIR Ment Health 2017;4(1):e8. [doi: 10.2196/mental.6375] [Medline: 28232298]
51. Shrestha N. Factor analysis as a tool for survey analysis. Am J Appl Math Stat 2021;9(1):4-11. [doi: 10.12691/ajams-9-1-2]
52. Tobias S, Carlson JE. Brief report: Bartlett's test of sphericity and chance findings in factor analysis. Multivariate Behav Res 1969;4(3):375-377. [doi: 10.1207/s15327906mbr0403_8] [Medline: 26745847]
53. Türk-Kurtça T, Kocatürk M. Beyond the Scroll: exploring how intolerance of uncertainty and psychological resilience explain the association between trait anxiety and doomscrolling. Pers Individ Dif 2025;233:112919. [doi: 10.1016/j.paid.2024.112919]
54. Hinssen M, Dohle S. Personal protective behaviors in response to COVID-19: a longitudinal application of protection motivation theory. Front Psychol 2023;14:1195607. [doi: 10.3389/fpsyg.2023.1195607] [Medline: 37655201]
55. Soroya SH, Rehman AU, Faiola A. Exploring the impact of internet and media sources exposure on self-care behavior: mediating the role of health anxiety, literacy and information-seeking behavior. Kybernetes 2023;53(11):4797-4817. [doi: 10.1108/k-06-2023-1003]
56. Satici SA, Gocet Tekin E, Deniz ME, Satici B. Doomscrolling Scale: its association with personality traits, psychological distress, social media use, and wellbeing. Appl Res Qual Life 2023;18(2):833-847 [FREE Full text] [doi: 10.1007/s11482-022-10110-7] [Medline: 36275044]
57. Dewitz L. Information behavior and the subjective digital well-being of individuals affected by social anxiety disorder. J Doc 2025;81(3):684-708. [doi: 10.1108/jd-07-2024-0180]
58. Yue Z, Zhang R, Xiao J. Passive social media use and psychological well-being during the COVID-19 pandemic: the role of social comparison and emotion regulation. Comput Human Behav 2022;127:107050 [FREE Full text] [doi: 10.1016/j.chb.2021.107050] [Medline: 34646057]
59. Bowman ND, Cohen E, Döveling K. Emotion and digital media: emotion regulation in interactive, on-demand, and networked media. In: Routledge International Handbook of Emotions and Media. London: Routledge; 2021:316-328.
60. Burnell K, George MJ, Vollet JW, Ehrenreich SE, Underwood MK. Passive social networking site use and well-being: the mediating roles of social comparison and the fear of missing out. Cyberpsychology 2019;13(3). [doi: 10.5817/cp2019-3-5]
61. Wanckel C. Keep me posted, but don't stress me out: how the positive effect of social networking services on civil servants' information use and political capacities can be attenuated by social media stress. Policy Sci 2024;57(3):539-566. [doi: 10.1007/s11077-024-09539-4]
62. Cassaretto M, Espinosa A, Chau C. Effects of resilience, social support, and academic self-efficacy, on mental health among Peruvian university students during the pandemic: the mediating role of digital inclusion. Front Psychol 2024;15:1282281. [doi: 10.3389/fpsyg.2024.1282281] [Medline: 39040968]
63. Zigdon A, Lavenda O, Lewin E. National resilience: development and validation of a new four-dimensional model for disaster preparedness assessment. J Saf Sci Resil 2026;7(1):100228. [doi: 10.1016/j.jnlssr.2025.100228]
64. Lerner JS, Keltner D. Fear, anger, and risk. J Pers Soc Psychol 2001;81(1):146-159. [doi: 10.1037//0022-3514.81.1.146] [Medline: 11474720]
65. Reeves A, Delfabbro P, Calic D. Encouraging employee engagement with cybersecurity: how to tackle cyber fatigue. SAGE Open 2021;11(1):215824402110000. [doi: 10.1177/21582440211000049]
66. Mizrak F, Demirel HG, Yaşar O, Karakaya T. Digital detox: exploring the impact of cybersecurity fatigue on employee productivity and mental health. Discov Ment Health 2025;5(1):25 [FREE Full text] [doi: 10.1007/s44192-025-00149-x] [Medline: 39998759]
67. Bıçakcı S, Evren A. Responding cyber-attacks and managing cyber security crises in critical infrastructures: A sociotechnical perspective. In: Tekinerdogan B, Catal C, Alskaif T, editors. Management and Engineering of Critical Infrastructures. Cambridge: Academic Press; 2024:125-151.
68. Bardin JS. Cyber warfare. In: Vacca JR, editor. Computer and Information Security Handbook. 4th Edition. Amsterdam: Elsevier Inc; 2024:1345-1380.
69. Marikyan D, Papagiannidis S. Protection motivation theory: a review. In: TheoryHub Book. England: University of Bristol; 2025:78-93.
70. Akib AAPM, Candiwan C, Ramadhani DP. Cybersecurity compliance and other factors influencing employee protective behavior: a case study of bank X in Indonesia. IJSSE 2025;15(06):1229-1241. [doi: 10.18280/ijsse.150613]
71. Gaspar R, Domingos S, Brito D, Leiras G, Filipe J, Raposo B, et al. Striving for crisis resolution or crisis resilience? The crisis layers and thresholds model and information and communication technology-mediated social sensing for evidence-based crisis management and communication. Human Behav and Emerg Tech 2020;3(1):40-52. [doi: 10.1002/hbe2.241]

XSL•FO
RenderX

72. Obasi HU. The role of media literacy in combating misinformation: exploring the effectiveness of media literacy programs in enhancing critical thinking skills and reducing the spread of false information: A literature review. MRS J Arts Humanit Liter 2025;2(3):1-5 [FREE Full text]
73. Lu C, Hu B, Bao M, Wang C, Bi C, Ju X. Can media literacy intervention improve fake news credibility assessment? A meta-analysis. Cyberpsychol Behav Soc Netw 2024;27(4):240-252. [doi: 10.1089/cyber.2023.0324] [Medline: 38484319]
74. Ezeonu NA, Hertelendy AJ, Adu MK, Kung JY, Itanyi IU, Dias RDL. Mobile apps to support mental health response in natural disasters: scoping review. J Med Internet Res 2024;26:e49929 [FREE Full text] [doi: 10.2196/49929] [Medline: 38520699]
75. Abuyadek RM, Hammouda EA, Elrewany E, Elmalawany DH, Ashmawy R, Zeina S. Acceptability of tele-mental health services among users: a systematic review and meta-analysis. BMC Public Health 2024;24(1):1143 [FREE Full text] [doi: 10.1186/s12889-024-18436-7] [Medline: 38658881]
76. Lazarus RS, Folkman S. Stress, Appraisal, and Coping. New York: Springer Publishing; 1984.
77. Abbas J, Wang D, Su Z, Ziapour A. The role of social media in the advent of COVID-19 pandemic: crisis management, mental health challenges and implications. Risk Manag Healthc Policy 2021;14:1917-1932 [FREE Full text] [doi: 10.2147/RMHP.S284313] [Medline: 34012304]
78. Huckins JF, daSilva AW, Wang W, Hedlund E, Rogers C, Nepal SK. Mental health and behavior of college students during the early phases of the COVID-19 pandemic: longitudinal smartphone and ecological momentary assessment study. J Med Internet Res 2020;22(6):e20185. [doi: 10.2196/20185] [Medline: 32519963]
79. Gao J, Zheng P, Jia Y, Chen H, Mao Y, Chen S. Mental health problems and social media exposure during COVID-19 outbreak. PLoS One 2020;15(4):e0231924. [doi: 10.1371/journal.pone.0231924] [Medline: 32298385]
80. Savitz DA, Wellenius GA. Can cross-sectional studies contribute to causal inference? It depends. Am J Epidemiol 2023;192(4):514-516. [doi: 10.1093/aje/kwac037] [Medline: 35231933]
81. Elston DM. Participation bias, self-selection bias, and response bias. J Am Acad Dermatol 2021;S0190-9622(21):01129-4. [doi: 10.1016/j.jaad.2021.06.025] [Medline: 34153389]
82. Zwilling M, Klien G, Lesjak D, Wiechetek ?, Cetin F, Basim HN. Cyber security awareness, knowledge and behavior: a comparative study. J Comput Inf Syst 2020;62(1):82-97. [doi: 10.1080/08874417.2020.1712269]
83. Klein G, Zwilling M. The weakest link: employee cyber-defense behaviors while working from home. J Comput Inf Syst 2023;64(3):408-422. [doi: 10.1080/08874417.2023.2221200]

Abbreviations

CFI: comparative fit index
TLI: Tucker-Lewis Index
SRMR: standardized root-mean-square residual
EFA: exploratory factor analysis
GAD-7: Generalized Anxiety Disorder-7 Scale
KMO: Kaiser-Meyer-Olkin
PHQ-9: Patient Health Questionnaire-9
PMT: protection motivation theory
PTSD: posttraumatic stress disorder
RMSEA: root-mean-square error of approximation
RQ: research question
SEM: structural equation modeling

Edited by S Brini; submitted 15.Jan.2026; peer-reviewed by O Bashkin, M Krishnapatnam; comments to author 04.Mar.2026; accepted 15.May.2026; published 17.Jun.2026.

Please cite as:
Zwilling M, Klein G, Barak S, Shtainmetz N, Tesler R, Zigdon A
Patterns of Digital Media Engagement During Wartime and Their Associations With Stress, Resilience, and Well-Being: Cross-Sectional Survey Study
J Med Internet Res 2026;28:e91484
URL: https://www.jmir.org/2026/1/e91484
doi:10.2196/91484
PMID:

XSL•FO
RenderX

©Moti Zwilling, Galit Klein, Sharon Barak, Noa Shtainmetz, Riki Tesler, Avi Zigdon. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Dietary Inflammatory Index and Depressive Symptoms in Chinese University Students Leveraging an Intelligent Ordering System: 3-Year Longitudinal Prospective Cohort Study

Peng Hong[1*], MA; Chen Hao[1*], PhD; Zhou Weiqiang[2*], MA; Qian Jie[3*], MA; Zhang Yimeng[1], MA; Ding Jingyun[3], MA; Qian Haihong[4], PhD; Jia Yingnan[1], PhD

[1]Department of Preventive Medicine and Health Education, School of Public Health, Fudan University, Shanghai, China

[2]Informatization Office, Fudan University, Shanghai, China

[3]Mental Wellbeing Education and Counseling Center, Fudan University, Shanghai, China

[4]Institute of Health Communication, Fudan University, Shanghai 200032, China, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Jia Yingnan, PhD
Department of Preventive Medicine and Health Education
School of Public Health
Fudan University
No. 130 Dong'an Road
Xuhui District
Shanghai, 200032
China
Phone: 86 13764665540
Email: jyn@fudan.edu.cn

Abstract

Background: Depression is a major global cause of disability, and depressive symptoms are highly prevalent and increasing among Chinese university students. Mounting evidence confirms that inflammation plays a key role in the pathogenesis of depression, and dietary inflammatory potential regulates systemic inflammation to influence depressive symptom development. However, existing research is limited by cross-sectional designs, recall bias from self-reported dietary surveys, and a lack of long-term prospective cohort evidence on the diet-inflammation-mental health pathway in Chinese university students.

Objective: This study aimed to examine the longitudinal association between Dietary Inflammatory Index (DII) and the incidence of depressive symptoms in Chinese university students, and to explore subgroup differences by family relationship and socioeconomic status.

Methods: A 3-year longitudinal prospective cohort study was conducted among 5314 students from a university in Shanghai, China. Eligible participants met the criteria of ≥86 days of annual campus cafeteria dining and at least 1 breakfast, lunch, and dinner in campus canteens per quarter; students with abnormal monthly energy intake, excessive food consumption, or incomplete 3-year dietary/psychological data were excluded. Dietary data were continuously collected via the Intelligent Ordering System (IOS) from April 2020 to March 2023 to calculate DII scores. Depressive symptoms were annually assessed using the Beck Depression Inventory-II from March 2021 to March 2023. Mixed-effects logistic regression (α=.05) was used to analyze the association, with subgroup analyses stratified by family relationship and poverty status.

Results: The baseline prevalence of depressive symptoms was 10.75% (571/5314; male: 261/2679, 9.74%; female: 310/2635, 11.76%). After adjusting for covariates, compared with the highest DII quartile (most proinflammatory diet), lower DII quartiles (more anti-inflammatory or low proinflammatory diets) were associated with a reduced risk of incident depressive symptoms in participants without depressive symptoms at baseline: Q1 (odds ratio [OR] 0.27, 95% CI 0.16-0.47), Q2 (OR 0.52, 95% CI 0.33-0.84), and Q3 (OR 0.26, 95% CI 0.16-0.42). Subgroup analyses showed this protective effect was only significant in students with harmonious family relationships and non–poverty-stricken students; no significant association between DII and depressive symptom improvement was found in participants with baseline depressive symptoms.

XSL•FO
RenderX

Conclusions: This study is among the first to prospectively examine dietary inflammatory potential and depressive symptoms in university students using long-term objective dietary monitoring. Unlike studies relying on self-reported dietary surveys, this study used an automated and precise campus-based IOS to continuously capture real-world dietary behaviors over 3 years. The findings indicate that sustained anti-inflammatory dietary patterns are associated with a lower risk of depressive symptoms among Chinese university students, although this protective effect was weaker in students experiencing family discord or socioeconomic disadvantage. These findings provide new longitudinal evidence for the diet-inflammation-mental health relationship and support integrated campus interventions combining dietary guidance with psychosocial support.

(J Med Internet Res 2026;28:e83784) doi:10.2196/83784

KEYWORDS

depressive symptoms; Dietary Inflammatory Index; intelligent ordering system; longitudinal prospective cohort study; university students

Introduction

Problem

Depression is recognized as a multifaceted disorder that induces impairments in interpersonal, social, and occupational functioning [1]. The World Health Organization has indicated that depression ranks among the primary causes of disability globally, with projections showing it will significantly exacerbate the overall global disease burden by 2030 [2,3]. University life represents a pivotal transitional phase for students as they mature into adulthood, a period often accompanied by diverse challenges [4]. Students at this stage are particularly vulnerable to depression [5]. In the United States, an estimated 15.6% of undergraduate students exhibit symptoms of depression [6]. In China, the prevalence rate of depressive symptoms among college students has reached 24.71% and is continuously increasing [7]. Depression has risen from the eighth to the fourth leading cause of death among adolescents, with its associated disability-adjusted life years increasing from 2.8% to 3.7% [8]. A meta-analysis including 24 studies indicated that reported prevalence rates among college students ranged from 10% to 85%, with a weighted mean prevalence of 30.6% [9]. The presence of severe depressive symptoms not only leads to sleep disorders, poor concentration, anxiety, and apathy toward daily experiences but may also give rise to dangerous behaviors such as self-harm and suicidal ideation [9].

Review of Relevant Scholarship

Growing robust evidence indicates that inflammation plays a pivotal role in the pathogenesis of depression, with individuals diagnosed with depression exhibiting elevated levels of multiple inflammatory markers, including C-reactive protein (CRP), interleukin-6, and tumor necrosis factor [10]. Diet, recognized as a critical modifiable lifestyle behavior, not only provides the energy and nutrients essential for sustaining life but also modulates systemic inflammatory levels through pathways such as inducing neuroinflammation and stimulating changes in oxidative stress, thereby influencing the development of depressive symptoms [11]. Healthy dietary habits, such as the Mediterranean diet pattern, typically include high amounts of vitamins, dietary fiber, polyphenolic phytochemicals, and unsaturated fatty acids, which exert anti-inflammatory and antidepressant effects [12,13]. By contrast, dietary factors that promote inflammation and depressive symptoms include oxidized lipids, saturated fatty acids, and trans fatty acids, which are highly prevalent in Western dietary patterns characterized by heavy consumption of sweets, refined grains, red and processed meats, snacks, and sugar-sweetened beverages [14].

Considering the complex composition of daily diets, which contain both proinflammatory and anti-inflammatory components, Shivappa et al [15] developed the Dietary Inflammatory Index (DII) to quantify dietary inflammatory levels by evaluating the effects of dietary components on 6 biomarkers: proinflammatory markers (CRP, interleukin-1β, interleukin-6, tumor necrosis factor-α) and anti-inflammatory markers (interleukin-4, interleukin-10) [16]. A lower DII score indicates a more anti-inflammatory diet, and a higher DII score reflects a more proinflammatory diet.

Previous studies have explored the association between dietary inflammatory levels and the risk of depressive symptoms among college students, but several limitations exist: first, bidirectional associations may exist between diet and depression; yet, a lack of cohort study evidence has hindered the clarification of the relationship between dietary inflammatory levels and depressive symptoms. Second, most studies predominantly rely on cross-sectional designs [17,18], which only investigate short-term dietary behaviors while overlooking the instability and malleability of college students' dietary patterns, thus failing to adequately capture the long-term trends of dietary changes among this population. Third, several studies acquired data on college students' food consumption through traditional survey methods, including 3-day recalls, 24-hour diet recalls, and food frequency questionnaire surveys, which introduced risks of recall bias [19].

Hypothesis, Aims, and Objectives

The Intelligent Ordering System (IOS) has been operational at a university in Shanghai since 2020, providing long-term, continuous, and relatively accurate meal data for each diner. Its accuracy and effectiveness were verified in a prior study [20]. Our research used 3 consecutive years of dietary records (April 2020-March 2023) from the IOS to investigate the inflammatory potential of habitual diets on incident depressive symptoms, with subgroup analyses to clarify differential effects of dietary inflammation levels across distinct population strata.

XSL•FO
RenderX

Methods

Inclusion and Exclusion Criteria

Participants were selected if they had ≥86 days/year of school cafeteria dining attendance and consumed at least 1 breakfast, 1 lunch, and 1 dinner in campus canteens during each quarter. Individuals were excluded if they had abnormal monthly energy intake (male: <800 kcal/d or >4000 kcal/d; female: <500 kcal/d or >3500 kcal/d) or excessive food consumption (intake >1000 g/d). Students with incomplete 3-year dietary records or annual depressive symptom assessments were also excluded [20].

Participant Characteristics

A total of 5314 university students from a university in Shanghai were included, covering undergraduates, master's students, and doctoral candidates. All participants had valid 3-year dietary records from the IOS and completed annual depressive symptom assessments.

Sampling Procedures

This study drew participants from the target population of undergraduate, master's, and doctoral students at Fudan University. Sampling proceeded through 2 parallel data streams—dietary records and repeated psychological assessments—followed by data linkage to form the final analytic sample. For dietary data, records from the campus IOS were collected continuously over the study period. Sequential quality and completeness checks were applied to filter the raw dietary data, excluding cases that failed predefined criteria for data adequacy and validity, to yield a subset of students with reliable dietary assessment data. For psychological assessments, annual depressive symptom surveys were conducted at 3 time points. Starting with the baseline survey, follow-up assessments were conducted in subsequent years, with nonparticipants at each follow-up excluded from further analysis. This process identified students with complete 3-year psychological assessment data. The final analytic sample was formed by matching students from the valid dietary data subset with those from the complete psychological assessment subset, retaining only participants who met all data quality and completeness requirements for both streams throughout the study period.

Sample Size, Power, and Precision

The intended sample size was determined by the availability of students with complete 3-year consecutive IOS dietary records and valid annual Beck Depression Inventory-II (BDI-II) assessments. The achieved sample size was 5314 participants, which was consistent with the planned sample derived from data completeness criteria. No a priori power analysis was performed, as this study was an observational cohort based on real-world objective dietary data; the large sample size ensured sufficient precision for estimating the association between DII and depressive symptoms. No interim analyses or stopping rules were used.

Measures and Covariates

Acquiring Dietary Data From the IOS and Assessment of DII

We assessed daily food composition and nutrient intake on the basis of daily food purchase records from the IOS from April 2020 to March 2023. The IOS contains 5714 dishes, and we first established a food database by weighing the materials of 1840 representative dishes before and after cooking. For the remaining 3874 dishes, we used a standardized recipe approach and calculated their weights on the basis of the composition of the representative dishes. The database includes information on food composition, such as the weight of raw food materials and condiments, as well as the nutrient composition of each meal. Details of the weighing and calculation methods have been described in a previous paper [20]. Nutrient composition was determined via the Chinese Food Composition database (version 2023). Ultimately, we assessed the composition of 4 foods, including onion, ginger, garlic, and chili pepper and 20 nutrients, including energy, protein, fat, carbohydrate, fiber, cholesterol, vitamin A, carotene, thiamine (vitamin B1), riboflavin (vitamin B2), nicotinic acid (vitamin B3), vitamin C, vitamin E, magnesium (Mg), iron (Fe), zinc (Zn), selenium (Se), saturated fatty acid, monounsaturated fatty acid, and polyunsaturated fatty acid contents.

The DII is a literature-derived instrument used to quantify the inflammatory potential of the overall diet; lower DII scores indicate more anti-inflammatory dietary patterns, whereas higher scores indicate more proinflammatory dietary patterns [16,21]. In epidemiological studies, the DII is frequently analyzed in quantiles, particularly quartiles or quintiles, to compare outcome risks across different levels of dietary inflammatory potential, assess potential gradient or dose-response associations, and avoid relying solely on a linear assumption when DII is modeled as a continuous variable [22,23]. Therefore, in this study, DII was categorized into quartiles (Q1-Q4), with Q1 representing relatively lower dietary inflammatory potential (more anti-inflammatory dietary profiles) and Q4 representing relatively higher dietary inflammatory potential (more proinflammatory dietary profiles), to examine the association between different levels of dietary inflammatory potential and depressive symptoms [24]. We used the modified version of the DII calculation developed by Shivappa [16], and the specific calculation method has been elaborated in previous studies [18]. Among the theoretically possible list of 45 food parameters, we included the 4 types of foods and 20 nutrients mentioned earlier.

Assessment of Depressive Symptoms

We used the BDI-II [25] to assess depressive symptoms for 3 rounds (March 2021, 2022, and 2023). The BDI-II is a 21-item self-report scale that assesses the severity of depressive symptoms in the past 2 weeks and includes somatic-affective (13 items) and cognitive (8 items) dimensions. Each item is rated on a 4-point Likert scale ranging from 0 to 3, and total scores range from 0 to 63, with higher scores indicating more severe symptomatology. For this study, depression was dichotomized into no depressive symptoms (0-13) or prevalent depressive symptoms. The higher the total score is, the more severe the subject's depression (a total score of 0-13 is

XSL•FO
RenderX

considered nondepressed, 14-27 is considered mild depression, and 28-63 is considered moderate-severe depression) for Chinese young adults [26].

Covariates

We assessed covariates in 2021. The following measures were considered pertinent covariates because they are associated with depression and are independent risk factors for depression among young adults. For age, it was categorized into 2 groups: those younger than 23 years of age and those 23 years or older. Gender was divided into 2 groups: male and female. Poverty-stricken students were classified into 2 groups: yes and no. Living-school experience was grouped into having the experience and not having the experience. Only child was categorized as yes or no. Family relationships were divided into harmonious and disharmonious groups.

Data Collection

Dietary data were automatically recorded in real time by the campus IOS from April 2020 to March 2023. Daily food composition and nutrient intake were derived from IOS meal records, combined with a standardized dish database. Depressive symptoms were assessed annually via self-administered BDI-II questionnaires in March 2021, 2022, and 2023.

Quality of Measurements

For dietary intake measurement, data were automatically and objectively recorded by the campus IOS, which eliminated recall bias. Standardized procedures for dish weighing, recipe calculation, and nutrient composition derivation were applied to ensure the accuracy and consistency of dietary data.

For depressive symptom measurement, the BDI-II was administered as a standardized self-report questionnaire with uniform instructions to all participants. To enhance data quality and prevent careless or random responding, common-sense check items and quality control questions were embedded in the questionnaire. Invalid questionnaires identified by these quality control items were excluded before data analysis, ensuring the reliability and validity of the outcome measurement.

Instrumentation

IOS: a campus-based automated platform with 5714 dishes; a food database was established by weighing 1840 representative dishes, with nutrient values derived from the Chinese Food Composition Database (2023). The validity of IOS for dietary assessment has been verified in prior studies.

BDI-II: a 21-item self-report scale measuring depressive symptom severity in the past 2 weeks, scored 0-3 per item (total 0-63).

Masking

No masking (blinding) was implemented in this study. Participants, data collectors, and outcome assessors were not blind to group assignments or exposure status, as this was an observational cohort study without experimental manipulation.

Psychometrics

The BDI-II has been validated in Chinese young adult populations with acceptable reliability and validity. For this sample, no additional reliability analysis was performed; the psychometric properties of BDI-II used in this study were based on published validation data for Chinese populations.

Conditions and Design

This study used a nonexperimental longitudinal observational cohort design without experimental manipulation. Participants were observed naturally over 3 years, with repeated assessments of dietary inflammatory potential and depressive symptoms.

Data Diagnostics

Missing data: the study database contained no missing values; no imputation methods were needed.

Outlier exclusion: predefined criteria for energy intake and food consumption outliers were applied before data analysis (as listed in the Inclusion and Exclusion Criteria section).

Data distribution: no data transformations were performed; statistical assumptions were evaluated during regression analysis.

Analytic Strategy

The characteristics of the participants and the quartile groups of the DII are presented as frequencies with percentages. A chi-square test was conducted to explore whether there were differences in the distribution of the DII quartile groups and depressive symptoms among different demographic characteristics at baseline. Mixed-effects logistic regression was used to assess the impact of different DII levels on the incidence of depressive symptoms, and odds ratios (ORs) and their 95% CIs were calculated. Individual-level random intercepts were incorporated into the model to capture unobserved heterogeneity between participants and appropriately handle within-subject correlation. Time was explicitly modeled as a categorical variable (year) to account for temporal variations in depression status. The model included DII quartile groups as the primary exposure variable, with the highest quartile (most proinflammatory diet) serving as the reference group. The variables of age group, family relationship, gender, and whether being a poverty-stricken student showed significant differences in distribution during the baseline demographic tests. Regarding the impact of DII levels on depressive symptom occurrence, for these variables, we performed further subgroup analyses using mixed-effects logistic regression to examine if there were differences in each subgroup. The database used in this study contains no missing values, and thus, no handling of missing data was involved. All statistical analyses were performed using R software (version 4.1.1; R Foundation for Statistical Computing) and Python (version 3.7; Python Software Foundation).

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki and was reviewed and approved by the Ethics Committee of Medical Research, School of Public Health, Fudan University (IRB#2019-01-0726 S; approval date: January 7, 2019), with no exemptions from human subject research ethics oversight. Written informed consent was obtained from all participants prior to data collection, and the original informed consent documents approved by the institutional review board

XSL•FO
RenderX

explicitly permitted secondary analysis of the anonymized dietary and psychological data without requiring additional consent from participants. All study data were anonymized and deidentified prior to analysis, with no personal identifiers (eg, name, student ID, and contact information) retained to protect participant privacy and confidentiality. No financial, material, or other forms of compensation were provided to participants for their involvement in this study. No images or materials containing personally identifiable information of individual participants are included in the manuscript or supplementary materials, so no additional consent for the use of identifiable images is required.

Results

Participant Flow

Participant flow through the study is presented in Figure 1.

Figure 1. Flowchart of participant enrollment and follow-up in a 3-year longitudinal prospective cohort study of dietary inflammatory index and depressive symptoms among university students in Shanghai, China (April 2020-March 2023).

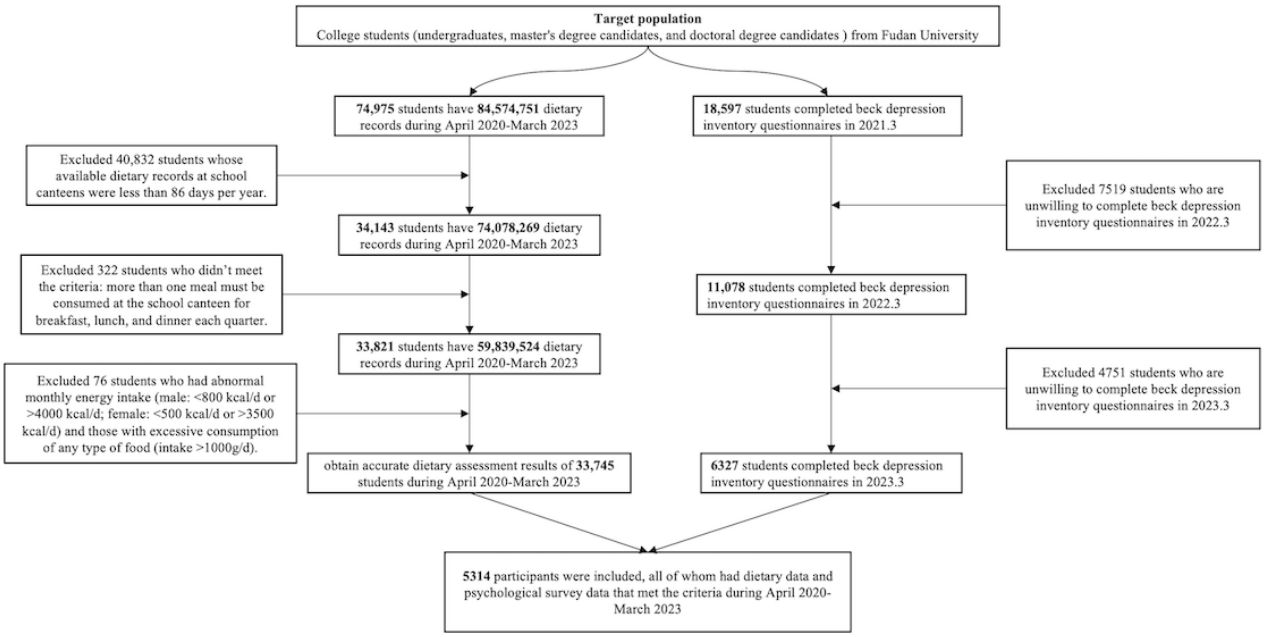

Recruitment

Recruitment and repeated assessments were conducted across a 3-year period. Dietary data were continuously collected via the IOS from April 2020 to March 2023, with semester breaks (January, February, July, and August of each year) excluded from analyses. Depressive symptoms were assessed annually at 3 fixed time points: March 2021 (baseline), March 2022 (first follow-up), and March 2023 (second follow-up). All participants were enrolled from a single university in Shanghai.

Statistics and Data Analysis

As shown in Table 1, we included 5314 young adults, of whom 2728 (51.34%) were female, 2586 (48.66%) were male, 3422 (64.4%) were aged between 17 and 23 years, 1892 (35.6%) were aged 23 years or older, 640 (12.04%) had poor relationships between their family, 824 (15.51%) were poverty-stricken students, 3201 (60.24%) had residential experience, and 3327 (62.61%) were only child.

XSL•FO
RenderX

Table 1. Baseline characteristics of participants in a 3-year longitudinal prospective cohort study of Dietary Inflammatory Index and depressive symptoms among university students in Shanghai, China (March 2021).

Baseline characteristics	Overall (n=5314), n (%)	Nondepressive (n=4743), n (%)	Depressive (n=571), n (%)	*P* value
Gender				.17
Female	2635 (49.59)	2325 (88.24)	310 (11.76)	
Male	2679 (50.41)	2418 (90.26)	261 (9.74)	
Age group				.01
17-23 years	1892 (35.60)	1716 (90.7)	176 (9.3)	
Over 23 years	3422 (64.40)	3027 (88.46)	395 (11.54)	
Family relationship				<.001
Disharmonious	640 (12.04)	483 (75.47)	157 (24.53)	
Harmonious	4674 (87.96)	4260 (91.14)	414 (8.86)	
Poverty-stricken students				<.001
No	4490 (84.49)	4040 (89.98)	450 (10.02)	
Yes	824 (15.51)	703 (85.32)	121 (14.68)	
Living-school experience				.50
No	2113 (39.76)	1894 (89.64)	219 (10.36)	
Yes	3201 (60.24)	2849 (89)	352 (11)	
Only child				.17
No	1987 (37.39)	1758 (88.48)	229 (11.52)	
Yes	3327 (62.61)	2985 (89.72)	342 (10.28)	

At baseline, young adults exhibited a median depression score of 4.0 (IQR 0-9.0) on the BDI-II, with 571 participants meeting criteria for depressive symptoms and 4743 remaining symptom-free. During the follow-up period, among the 4743 participants without depressive symptoms at baseline, 451 (9.5%) experienced the onset of depressive symptoms; among the 571 participants with depressive symptoms at baseline, depressive symptoms disappeared in 318 (55.6%). Table 1 also describes the distribution of several characteristics according to depressive symptoms. Students aged 23 years and older have a higher proportion of experiencing depressive symptoms compared to those <23 years of age. College students with disharmonious family relationships and those from poverty-stricken backgrounds have a higher proportion of experiencing depressive symptoms.

The median score at baseline study was 0.36 (IQR –2.23 to 1.97), which indicated a slightly proinflammatory diet. Table 2 illustrates the baseline characteristics by DII quartiles in the sample as a whole. Gender distribution varied markedly among quartiles (*P*=.004), with a slight but notable increase in female representation in Q4 (proinflammatory diet: 663/2635, 25.16%) compared to Q1 (anti-inflammatory diet: 654/2635, 24.82%). Poverty-stricken status exhibited a distinct nonlinear pattern (*P*=.04), peaking in Q3 (low-grade proinflammatory diet: 221/824, 26.82%) before declining sharply in Q4 (190/824, 23.06%).

XSL•FO
RenderX

Table 2. Baseline characteristics of university students by quartiles of Dietary Inflammatory Index in a 3-year longitudinal prospective cohort study in Shanghai, China (April 2020-March 2023).

Baseline characteristics	Q1 (n=1329), n (%)	Q2 (n=1328), n (%)	Q3 (n=1328), n (%)	Q4 (n=1329), n (%)	*P* value
Gender					.004
Female	654 (24.82)	706 (26.79)	612 (23.23)	663 (25.16)	
Male	675 (25.20)	622 (23.22)	716 (26.73)	666 (24.86)	
Age group					.45
≥23 years	450 (23.78)	474 (25.05)	487 (25.74)	481 (25.42)	
<23 years	879 (25.69)	854 (24.96)	841 (24.58)	848 (24.78)	
Family relationship					.75
Disharmonious	158 (24.69)	151 (23.59)	162 (25.31)	169 (26.41)	
Harmonious	1171 (25.05)	1177 (25.18)	1166 (24.95)	1160 (24.82)	
Poverty-stricken students					.04
No	1144 (25.48)	1100 (24.50)	1107 (24.65)	1139 (25.37)	
Yes	185 (22.45)	228 (27.67)	221 (26.82)	190 (23.06)	
Living-school experience					.06
No	533 (25.22)	487 (23.05)	548 (25.93)	545 (25.79)	
Yes	796 (24.87)	841 (26.27)	780 (24.37)	784 (24.49)	
Only child					.40
No	477 (24.01)	519 (26.12)	494 (24.86)	497 (25.01)	
Yes	852 (25.61)	809 (24.32)	834 (25.07)	832 (25.01)	

As shown in Table 3, among participants with depressive symptoms at baseline, no statistically significant associations were observed. Among participants without depressive symptoms at baseline, Model 2, which adjusted for baseline covariates, showed that compared with those in the proinflammatory diet group (Q4), participants in the other 3 groups with lower dietary inflammatory levels had significantly lower odds of incident depressive symptoms: Q1 (OR 0.27, 95% CI 0.16-0.47), Q2 (OR 0.52, 95% CI 0.33-0.84), and Q3 (OR 0.26, 95% CI 0.16-0.42).

Table 3. Association between Dietary Inflammatory Index (DII) and incident depressive symptoms among university students in a 3-year longitudinal prospective cohort study in Shanghai, China (April 2020-March 2023).

Groups and DII	Model 1[a]		Model 2[b]	
	OR[c] (95% CI)	*P* value	OR (95% CI)	*P* value
Baseline participants with depressive symptoms				
Q1	1.00 (0.77-1.32)	.97	0.81 (0.55-1.19)	.28
Q2	0.95 (0.72-1.26)	.74	0.70 (0.49-1.06)	.05
Q3	1.01 (0.77-1.34)	.92	0.86 (0.63-1.18)	.36
Q4	Reference	N/A[d]	Reference	N/A
Baseline participants without depressive symptoms				
Q1	1.23 (0.75-2.03)	.41	0.27 (0.16-0.47)	<.001
Q2	0.94 (0.57-1.57)	.82	0.52 (0.33-0.84)	.007
Q3	1.19 (0.74-1.93)	.47	0.26 (0.16-0.42)	<.001
Q4	Reference	N/A	Reference	N/A

[a]Model 1 was unadjusted for any covariates.

[b]Model 2 adjusted for covariates at baseline: gender, age group, family relationship, poverty-stricken students, living-school experience, and only child.

[c]OR: odds ratio.

[d]N/A: not applicable.

XSL•FO
RenderX

As shown in Table 4, among participants with depressive symptoms at baseline, we did not find any significant association between dietary inflammatory potential and improvement in depressive symptoms.

Table 4. Subgroup analysis of the association between Dietary Inflammatory Index (DII) and depressive symptom improvement among baseline depressed students in a 3-year longitudinal prospective cohort study in Shanghai, China (April 2020-March 2023).

Character and DII	Model 1[a]		Model 2[b]	
	OR[c] (95% CI)	*P* value	OR (95% CI)	*P* value
Family relationship				
Harmonious (n=414)				
Q1	1.03 (0.76-1.41)	.84	1.31 (0.74-2.30)	.36
Q2	1.00 (0.73-1.38)	.98	1.01 (0.56-1.81)	.97
Q3	1.06 (0.77-1.47)	.71	1.49 (0.85-2.59)	.16
Q4	Reference	N/A[d]	Reference	N/A
Disharmonious (n=157)				
Q1	0.83 (0.44-1.57)	.57	0.98 (0.37-2.58)	.96
Q2	0.74 (0.39-1.42)	.37	0.71 (0.27-1.84)	.48
Q3	0.72 (0.39-1.32)	.29	0.71 (0.29-1.72)	.45
Q4	Reference	N/A	Reference	N/A
Poverty-stricken students				
Yes (n=121)				
Q1	0.78 (0.43-1.40)	.40	0.98 (0.43-2.24)	.96
Q2	0.89 (0.48-1.64)	.71	0.76 (0.32-1.77)	.52
Q3	1.18 (0.65-2.15)	.59	1.53 (0.70-3.35)	.29
Q4	Reference	N/A	Reference	N/A
No (n=450)				
Q1	1.08 (0.79-1.46)	.64	1.32 (0.75-2.35)	.34
Q2	0.98 (0.71-1.34)	.88	1.02 (0.57-1.83)	.94
Q3	0.97 (0.71-1.34)	.87	1.07 (0.62-1.85)	.81
Q4	Reference	N/A	Reference	N/A

[a]Model 1 was unadjusted for any covariates.

[b]In the family relationship subgroup, Model 2 adjusted for covariates at baseline: gender, age group, poverty-stricken students, living-school experience, and only child; in the poverty-stricken student subgroup, Model 2 adjusted for covariates at baseline: gender, age group, family relationship, living-school experience, and only child.

[c]OR: odds ratio.

[d]N/A: not applicable.

As shown in Table 5, among participants without depressive symptoms at baseline, Model 2 showed particularly pronounced reductions in the odds of depressive symptoms among those with harmonious family relationships: Q1 (OR 0.38, 95% CI 0.25-0.59), Q2 (OR 0.28, 95% CI 0.20-0.40), and Q3 (OR 0.49, 95% CI 0.36-0.66), compared with the proinflammatory diet group (Q4). In contrast, no significant associations were observed among participants with disharmonious family relationships in either model.

XSL•FO
RenderX

Table 5. Subgroup analysis of the association between Dietary Inflammatory Index (DII) and incident depressive symptoms among baseline nondepressed students in a 3-year longitudinal prospective cohort study in Shanghai, China (April 2020-March 2023).

Character and DII	Model 1[a]		Model 2[b]	
	OR[c] (95% CI)	*P* value	OR (95% CI)	*P* value
Family relationship				
Harmonious (n=4260)				
Q1	0.64 (0.40-1.02)	.06	0.16 (0.07-0.33)	<.001
Q2	0.68 (0.45-1.02)	.06	0.24 (0.13-0.46)	<.001
Q3	0.80 (0.55-1.16)	.25	0.56 (0.34-0.91)	.02
Q4	Reference	N/A[d]	Reference	N/A
Disharmonious (n=483)				
Q1	1.36 (0.83-2.22)	.22	1.16 (0.62-2.16)	.64
Q2	0.85 (0.50-1.42)	.53	0.65 (0.33-1.29)	.22
Q3	1.20 (0.74-1.94)	.46	1.02 (0.55-1.88)	.96
Q4	Reference	N/A	Reference	N/A
Poverty-stricken students				
Yes (n=703)				
Q1	0.87 (0.50-2.05)	.98	1.00 (0.49-2.04)	>.99
Q2	0.83 (0.56-1.95)	.88	1.05 (0.56-1.95)	.88
Q3	1.04 (0.61-2.09)	.69	1.13 (0.61-2.09)	.69
Q4	Reference	N/A	Reference	N/A
No (n=4040)				
Q1	0.74 (0.48-1.14)	.17	1.51 (0.86-2.64)	.15
Q2	0.60 (0.40-0.89)	.01	0.25 (0.13-0.49)	<.001
Q3	0.76 (0.53-1.08)	.13	1.42 (0.86-2.33)	.17
Q4	Reference	N/A	Reference	N/A

[a]Model 1 was unadjusted for any covariates.

[b]In the family relationship subgroup, Model 2 adjusted for covariates at baseline: gender, age group, poverty-stricken students, living-school experience, and only child; in the poverty-stricken student subgroup, Model 2 adjusted for covariates at baseline: gender, age group, family relationship, living-school experience, and only child.

[c]OR: odds ratio.

[d]N/A: not applicable.

When stratified by poverty status, both models showed protective associations between anti-inflammatory dietary profiles and a lower risk of depressive symptoms exclusively among non–poverty-stricken students. In Model 1, the association was observed for Q2 compared with Q4 (OR 0.60, 95% CI 0.40-0.89); this pattern persisted in Model 2 (OR 0.25, 95% CI 0.13-0.49). No statistically significant differences across DII quartiles were observed among poverty-stricken students.

Discussion

Support of Original Hypotheses

To our best knowledge, this is the first study to conduct long-term dietary monitoring and prospective analysis of the association between dietary inflammatory potential and the incidence of depressive symptoms among the college student population [27,28]. The study used a longitudinal cohort design involving 5314 university students, integrating objective dietary assessments via IOS—a campus-based platform for automated dietary consumption data recording—with annual depressive symptom evaluations using the BDI-II. In participants without depressive symptoms at baseline, our longitudinal evidence finds an association between dietary inflammatory potential and depressive symptom incidence, demonstrating that adherence to anti-inflammatory or low proinflammatory diets was associated with a reduced risk of having depressive symptoms. Notably, this protective effect exhibited substantial subgroup heterogeneity. Individuals experiencing socioeconomic disadvantages or familial discordance derived fewer benefits from anti-inflammatory dietary patterns, highlighting the moderating role of psychosocial contexts in these relationships. In participants with depressive symptoms at baseline, we did not find any significant association between dietary inflammatory potential and depressive symptom improvement.

XSL•FO
RenderX

Similarity of Results

In this study, the prevalence of depressive symptoms among university students was 10.75% (571/5314), with rates of 11.76% (310/2635) in female and 9.74% (261/2679) in male individuals. These figures are substantially lower than reported prevalence rates among university students in other global regions (Europe: 24.1%, North America: 35.5%, Africa: 40.1%, Asia: 34.8%) [29]. In China, university students aged <23 years are predominantly undergraduates, while those >23 years of age are mainly postgraduates [30]. This study found significantly higher prevalence of depressive symptoms among university students aged 23 years or older compared to their younger counterparts, indicating elevated risk among graduate students. Previous studies have identified graduate students as a high-risk population for various mental health problems, with contributing factors including academic pressures, financial strain, stress, and career uncertainty [31,32]. These cumulative stressors have established depression as both a prevalent psychological disorder in this population and 1 occurring at 6 times the rate observed in the general population [33]. At baseline, the prevalence of depressive symptoms was significantly higher in the family relationship disharmonious group than in the harmonious group and among poverty-stricken students compared to non–poverty-stricken students. Previous studies have consistently demonstrated that family warmth, cohesion, and supportive parenting practices are associated with improved mental health outcomes and reduced depressive symptoms in adolescents [34]. Furthermore, a study from Hubei Province, China, revealed that childhood exposure to emotional abuse, domestic violence, and physical neglect increased the risk of developing depressive symptoms in adulthood by 80%, 88%, and 39%, respectively [35]. A UK-based study found that university students from affluent families had a 70% lower probability of developing depressive symptoms [36], which aligns with our findings. Students with higher socioeconomic status benefit from greater access to mental health services, academic support resources, and stable living environments—all of which help mitigate psychological stress and promote better mental health outcomes [37].

Regarding dietary inflammatory potential, we observed a significantly higher proportion of proinflammatory diets among male compared to female individuals, consistent with numerous previous studies [38]. This disparity likely stems from gender-specific dietary preferences: male individuals tend to consume greater quantities of proinflammatory foods such as red meat and processed meats, while female individuals show stronger preferences for anti-inflammatory dietary components, including vegetables, whole grains, tofu, and high-cocoa dark chocolate [39]. In our study, non–poverty-stricken students demonstrated relatively even distribution across all 4 quartiles (Q1-Q4) of dietary inflammatory potential, whereas poverty-stricken students showed greater concentration in both mildly proinflammatory (Q3) and mildly anti-inflammatory diets (Q2), with fewer individuals in anti-inflammatory diets (Q1) and proinflammatory diets (Q4) categories. This finding contrasts with previous research indicating a positive association between socioeconomic status and proinflammatory diets [40]. The observed discrepancy may stem from unique pricing characteristics of university dining halls; more expensive menu items tend to be highly processed, richer in oil, salt, and sugar, and predominantly feature red meat, while nutritionally balanced, minimally processed options are substantially cheaper [41-43]. Consequently, non–poverty-stricken students can easily maintain balanced dietary patterns, resulting in higher proportions in Q2 and Q3 [44].

Compared to university students with proinflammatory diets (Q4), those exhibiting anti-inflammatory (Q1 and Q2) or low proinflammatory (Q3) dietary patterns demonstrated reduced risks of depressive symptoms. Our findings partially align with previous studies conducted in Iranian, Turkish, Spanish, and Australian populations. In Iranian adolescents, consumption of a proinflammatory diet, reflected by higher DII scores, was associated with significantly greater odds of exhibiting at least moderate depressive symptoms compared to female individuals consuming the least anti-inflammatory diets (OR 3.96, 95% CI 1.12-13.97) [45]. A Turkish study examined students in the nutrition and dietetics department of a private university's faculty of health sciences in Istanbul. It identified a positive association between DII and depressive symptoms (OR ranging between 1.64-1.67), supporting that the inflammatory burden of diet correlates with students' propensity for depression [17]. A Spanish study involving 15,093 university graduates analyzed the relationship between dietary inflammatory potential and depression risk, finding that participants in the highest DII quintile (most proinflammatory diet) had a 47% higher depression risk compared to those in the lowest quintile (hazard ratio 1.47, 95% CI 1.17-1.85), demonstrating a significant dose-response relationship [46]. A 9-year cohort study using the middle-aged cohort of the Australian Longitudinal Study on Women's Health, which included a total of 6438 women, found that women with the most anti-inflammatory diet had an approximately 20% lower risk of developing depression compared with women with the most proinflammatory diet (relative risk 0.81, 95% CI 0.69, 0.96), suggesting that an anti-inflammatory diet is associated with a lower risk of depression in middle-aged Australian women [15].

Interpretation

Numerous studies across diverse populations have demonstrated analogous findings [47]. This pattern is also broadly consistent with evidence in university students, among whom better overall diet quality has generally been associated with better mental health, whereas stress and anxiety have often been linked to poorer diet quality, suggesting that the potential mental health benefits of healthier eating may be harder to realize under sustained psychosocial strain [48]. However, in our subgroup analyses, the same dietary behavior did not significantly reduce the risk of depressive symptom occurrence among students categorized as family relationship disharmonious and poverty-stricken. At baseline, the prevalence of depressive symptoms was significantly higher in the family relationship disharmonious group and poverty-stricken students, suggesting an inherent vulnerability to depressive symptoms in populations exhibiting these characteristics [49,50]. Mechanistically, parental conflicts, marital problems, or harsh punishment that create adverse family environments immediately increase depression risk through psychological distress [51]. Beyond these proximal

XSL•FO
RenderX

psychological effects, early adversity can shape the hypothalamic-pituitary-adrenocortical axis and other stress-response systems, producing more persistent neuroendocrine dysregulation that may amplify inflammatory signaling and vulnerability to later psychopathology [52,53]. Chronically, such environments stimulate proinflammatory cytokine production, elevating serum CRP and interleukin-6 [54]. Similarly, adverse economic conditions lead to a series of detrimental sequelae, including prolonged psychological distress, inadequate nutritional intake, and accelerated biological aging processes [55], further elevating systemic inflammation [54]. In young adults, economic hardship is also closely intertwined with food insecurity, which has been repeatedly associated with poorer mental health, sleep disturbance, and depressive symptoms, and may reduce both the quality and consistency of anti-inflammatory food intake [52,56,57]. The Pittsburgh Girls Study established that greater family adversity during ages 5-9 years predicted higher interleukin-6 levels in adolescence (β=.14; *P*=.048), with persistence into early adulthood (β=.12; *P*=.001) [58]. Childhood financial strain independently predicted adult interleukin-6 elevations (β=.09; *P*=.009) [59]. Prior research confirms childhood socioeconomic disadvantage and family discordance promote chronic inflammation through biological embedding mechanisms, including dysregulated stress response systems and adiposity-driven inflammatory pathways [60,61]. Conversely, close and supportive family relationships have been conceptualized as resilience factors that can buffer the biological embedding of adversity and contribute to more favorable long-term health profiles [62]. Consequently, even after implementing anti-inflammatory or low proinflammatory diets, systemic inflammation may remain elevated, thereby diminishing dietary interventions' capacity to reduce depression risk.

Strengths and Limitations

This study has several key strengths. First, its 3-year longitudinal cohort design enabled follow-up of dietary behaviors and depressive symptoms, supporting a temporal association not possible with prior cross-sectional research. Second, the IOS-collected canteen meal data were recorded automatically and continuously, ensuring objectivity and accuracy. Third, subgroup analyses revealed nuanced interactions between dietary inflammation and psychosocial factors, showing that anti-inflammatory diets' protective effects are shaped by familial and socioeconomic contexts—with direct implications for targeted interventions.

There are several limitations in this study. First, the meal data from the IOS cannot provide a comprehensive overview of all food consumed by individuals daily. Although validation studies confirmed that campus dining records reflect students' primary dietary patterns, unmeasured consumption of off-campus foods (eg, fruits, beverages, and snacks) may have introduced nondifferential misclassification in DII calculations, potentially attenuating observed associations. Second, constraints in covariate breadth—particularly the absence of data on physical activity, sleep quality, and substance use—precluded full adjustment for potential behavioral confounders. Third, the inclusion criteria requiring complete 3-meal records from campus dining halls may have introduced selection bias by excluding students with irregular eating habits. Admittedly, students with more irregular dietary behaviors are more likely to experience adverse depressive symptoms. Nevertheless, after careful consideration, we retained these inclusion and exclusion criteria to ensure the representativeness of participants' dietary behaviors in the study.

Conclusion

This study is among the first to prospectively examine dietary inflammatory potential and depressive symptoms in university students using long-term objective dietary monitoring. Unlike previous studies relying mainly on self-reported dietary surveys, this study used an automated and precise campus-based IOS to continuously capture real-world dietary behaviors over 3 years. The findings indicate that sustained anti-inflammatory dietary patterns are associated with a lower risk of depressive symptoms among Chinese university students, although this protective effect was weaker in students experiencing family discord or socioeconomic disadvantage. These findings provide new longitudinal evidence for the diet-inflammation-mental health relationship and support integrated campus interventions combining dietary guidance with psychosocial support.

Acknowledgments

Gratitude is extended to Ms Jiang Fan, MS, and Ms Yao Qinqin, MS, graduated members of the research group, for their precise construction of the dish database, without which this research would have been difficult to conduct. Additionally, gratitude is expressed to all fellow students for their active participation, as well as to other teachers and canteen staff who have provided support for this research.

Data Availability

The data presented in this study are available upon request from the corresponding author. The data are not publicly available due to the privacy of the study participants.

Funding

This study was funded by the 6th Round Three-Year Initiative Plan for Strengthening Public Health System Construction in Shanghai (project number GWVI-6). The funder had no role in the conduct of the study; collection, management, analysis, or interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: QJ, QH, JY
Methodology and data collection: PH, ZW, ZY
Analysis: PH, CH
Writing—original draft: PH, CH, ZW
Writing—review and editing: QJ, QH, JY
All authors contributed equally to this work.

Conflicts of Interest

None declared.

References

1. Rofman ES. Kaplan and Sadock's synopsis of psychiatry: behavioral sciences/clinical psychiatry, 10th edition. J Clin Psychiatry 2009;70(6):940 [FREE Full text] [doi: 10.4088/jcp.09bk05044]
2. GBD 2017 DiseaseInjury IncidencePrevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018;392(10159):1789-1858 [FREE Full text] [doi: 10.1016/S0140-6736(18)32279-7] [Medline: 30496104]
3. Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med 2006;3(11):e442 [FREE Full text] [doi: 10.1371/journal.pmed.0030442] [Medline: 17132052]
4. Mendelson T, Tandon SD. Prevention of depression in childhood and adolescence. Child Adolesc Psychiatr Clin N Am 2016;25(2):201-218. [doi: 10.1016/j.chc.2015.11.005] [Medline: 26980124]
5. Feng Q, Zhang Q, Du Y, Ye Y, He Q. Associations of physical activity, screen time with depression, anxiety and sleep quality among Chinese college freshmen. PLoS One 2014;9(6):e100914 [FREE Full text] [doi: 10.1371/journal.pone.0100914] [Medline: 24964250]
6. Eisenberg D, Gollust SE, Golberstein E, Hefner JL. Prevalence and correlates of depression, anxiety, and suicidality among university students. Am J Orthopsychiatry 2007;77(4):534-542. [doi: 10.1037/0002-9432.77.4.534] [Medline: 18194033]
7. Gao L, Xie Y, Jia C, Wang W. Prevalence of depression among Chinese university students: a systematic review and meta-analysis. Sci Rep 2020;10(1):15897 [FREE Full text] [doi: 10.1038/s41598-020-72998-1] [Medline: 32985593]
8. GBD 2019 DiseasesInjuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020;396(10258):1204-1222 [FREE Full text] [doi: 10.1016/S0140-6736(20)30925-9] [Medline: 33069326]
9. Ibrahim AK, Kelly SJ, Adams CE, Glazebrook C. A systematic review of studies of depression prevalence in university students. J Psychiatr Res 2013;47(3):391-400. [doi: 10.1016/j.jpsychires.2012.11.015] [Medline: 23260171]
10. Köhler CA, Freitas TH, Maes M, de Andrade NQ, Liu CS, Fernandes BS, et al. Peripheral cytokine and chemokine alterations in depression: a meta-analysis of 82 studies. Acta Psychiatr Scand 2017;135(5):373-387. [doi: 10.1111/acps.12698] [Medline: 28122130]
11. Wirth MD, Shivappa N, Davis L, Hurley T, Ortaglia A, Drayton R, et al. Construct validation of the Dietary Inflammatory Index among African Americans. J Nutr Health Aging 2017;21(5):487-491 [FREE Full text] [doi: 10.1007/s12603-016-0775-1] [Medline: 28448077]
12. Hosseini B, Berthon BS, Saedisomeolia A, Starkey MR, Collison A, Wark PAB, et al. Effects of fruit and vegetable consumption on inflammatory biomarkers and immune cell populations: a systematic literature review and meta-analysis. Am J Clin Nutr 2018;108(1):136-155 [FREE Full text] [doi: 10.1093/ajcn/nqy082] [Medline: 29931038]
13. Wang W, Li J, Chen X, Yu M, Pan Q, Guo L. Whole grain food diet slightly reduces cardiovascular risks in obese/overweight adults: a systematic review and meta-analysis. BMC Cardiovasc Disord 2020;20(1):82 [FREE Full text] [doi: 10.1186/s12872-020-01337-z] [Medline: 32070285]
14. Tristan Asensi M, Napoletano A, Sofi F, Dinu M. Low-grade inflammation and ultra-processed foods consumption: a review. Nutrients 2023;15(6):1546 [FREE Full text] [doi: 10.3390/nu15061546] [Medline: 36986276]
15. Shivappa N, Schoenaker DAJM, Hebert JR, Mishra GD. Association between inflammatory potential of diet and risk of depression in middle-aged women: the Australian longitudinal study on women's health. Br J Nutr 2016;116(6):1077-1086. [doi: 10.1017/S0007114516002853] [Medline: 27498949]
16. Shivappa N, Steck SE, Hurley TG, Hussey JR, Hébert JR. Designing and developing a literature-derived, population-based dietary inflammatory index. Public Health Nutr 2014;17(8):1689-1696 [FREE Full text] [doi: 10.1017/S1368980013002115] [Medline: 23941862]
17. Arslan E, Ozlu T, Kenger EB, Meral Koc B. Determination of the relationship between dietary inflammatory index and depression status in female students. Nutr Health 2024;30(4):707-713. [doi: 10.1177/02601060221136357] [Medline: 36330675]

XSL•FO
RenderX

18. Dehghan P, Nejati M, Vahid F, Almasi-Hashiani A, Saleh-Ghadimi S, Parsi R, et al. The association between dietary inflammatory index, dietary antioxidant index, and mental health in adolescent girls: an analytical study. BMC Public Health 2022;22(1):1513 [FREE Full text] [doi: 10.1186/s12889-022-13879-2] [Medline: 35945535]
19. Shim JS, Oh K, Kim HC. Dietary assessment methods in epidemiologic studies. Epidemiol Health 2014;36:e2014009 [FREE Full text] [doi: 10.4178/epih/e2014009] [Medline: 25078382]
20. Zhang Y, Gu D, Luo M, Liu S, Peng H, Jia Y. Relative validity of an intelligent ordering system to estimate dietary intake among university students from a medical school in Shanghai, China. Int J Behav Nutr Phys Act 2024;21(1):70 [FREE Full text] [doi: 10.1186/s12966-024-01619-1] [Medline: 38965619]
21. Hébert JR, Shivappa N, Wirth MD, Hussey JR, Hurley TG. Perspective: the dietary inflammatory index (DII)-lessons learned, improvements made, and future directions. Adv Nutr 2019;10(2):185-195 [FREE Full text] [doi: 10.1093/advances/nmy071] [Medline: 30615051]
22. Shivappa N, Schoenaker DAJM, Hebert JR, Mishra GD. Association between inflammatory potential of diet and risk of depression in middle-aged women: the Australian longitudinal study on women's health. Br J Nutr 2016;116(6):1077-1086. [doi: 10.1017/S0007114516002853] [Medline: 27498949]
23. Ramallal R, Toledo E, Martínez-González MA, Hernández-Hernández A, García-Arellano A, Shivappa N, et al. Dietary inflammatory index and incidence of cardiovascular disease in the SUN cohort. PLoS One 2015;10(9):e0135221 [FREE Full text] [doi: 10.1371/journal.pone.0135221] [Medline: 26340022]
24. Adjibade M, Andreeva VA, Lemogne C, Touvier M, Shivappa N, Hébert JR, et al. The inflammatory potential of the diet is associated with depressive symptoms in different subgroups of the general population. J Nutr 2017;147(5):879-887 [FREE Full text] [doi: 10.3945/jn.116.245167] [Medline: 28356432]
25. Beck AT, Steer RA, Ball R, Ranieri WF. Comparison of Beck Depression Inventories-IA and -II in psychiatric outpatients. J Pers Assess 1996;67(3):588-597. [doi: 10.1207/s15327752jpa6703_13] [Medline: 8991972]
26. Zhu J, Zhang J, Sheng Z, Wang F. Reliability and validity of the Beck Depression Inventory-II applied to Chinese construction workers. Soc Behav Personal 2018;46(2):249-258. [doi: 10.2224/sbp.6638]
27. Shivappa N, Hébert JR, Veronese N, Caruso MG, Notarnicola M, Maggi S, et al. The relationship between the Dietary Inflammatory Index (DII) and incident depressive symptoms: a longitudinal cohort study. J Affect Disord 2018;235:39-44. [doi: 10.1016/j.jad.2018.04.014] [Medline: 29649709]
28. Lassale C, Batty GD, Baghdadli A, Jacka F, Sánchez-Villegas A, Kivimäki M, et al. Healthy dietary indices and risk of depressive outcomes: a systematic review and meta-analysis of observational studies. Mol Psychiatry 2019;24(7):965-986 [FREE Full text] [doi: 10.1038/s41380-018-0237-8] [Medline: 30254236]
29. Li W, Zhao Z, Chen D, Peng Y, Lu Z. Prevalence and associated factors of depression and anxiety symptoms among college students: a systematic review and meta-analysis. J Child Psychol Psychiatry 2022;63(11):1222-1230. [doi: 10.1111/jcpp.13606] [Medline: 35297041]
30. Newman I, Ding L, Feng Y. Estimate of undergraduate university student alcohol use in China: a systematic review and meta-analysis. Arch Public Health 2017;75:52 [FREE Full text] [doi: 10.1186/s13690-017-0220-x] [Medline: 29238573]
31. Casey C, Harvey O, Taylor J, Knight F, Trenoweth S. Exploring the wellbeing and resilience of postgraduate researchers. J Further High Educ 2022;46(6):850-867. [doi: 10.1080/0309877x.2021.2018413]
32. Kılıç C, Karahan FS. Mental health problems among graduate students in Turkey: a cross-sectional study. Int J Adv Couns 2023:634-650 [FREE Full text] [doi: 10.1007/s10447-023-09514-3] [Medline: 37359032]
33. Evans TM, Bira L, Gastelum JB, Weiss LT, Vanderford NL. Evidence for a mental health crisis in graduate education. Nat Biotechnol 2018;36(3):282-284. [doi: 10.1038/nbt.4089] [Medline: 29509732]
34. Tang X, Tang S, Ren Z, Wong DFK. Psychosocial risk factors associated with depressive symptoms among adolescents in secondary schools in mainland China: a systematic review and meta-analysis. J Affect Disord 2020;263:155-165. [doi: 10.1016/j.jad.2019.11.118] [Medline: 31818773]
35. Wang Z, Wang L, Han J. Association of adverse childhood experiences with depression in adults. Acta Med Univ Sci Technol Huazhong 2016;45(2):216-219 [FREE Full text]
36. Ibrahim AK, Kelly SJ, Glazebrook C. Socioeconomic status and the risk of depression among U.K. higher education students. Soc Psychiatry Psychiatr Epidemiol 2013;48(9):1491-1501. [doi: 10.1007/s00127-013-0663-5] [Medline: 23412652]
37. Fu X, Peng S, Feng XL. Socioeconomic inequalities in depressive symptoms in China: the role of social capital. Heliyon 2024;10(3):e24918 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e24918] [Medline: 38314291]
38. Li X, Chen M, Yao Z, Zhang T, Li Z. Dietary inflammatory potential and the incidence of depression and anxiety: a meta-analysis. J Health Popul Nutr 2022;41(1):24 [FREE Full text] [doi: 10.1186/s41043-022-00303-z] [Medline: 35643518]
39. Feraco A, Armani A, Amoah I, Guseva E, Camajani E, Gorini S, et al. Assessing gender differences in food preferences and physical activity: a population-based survey. Front Nutr 2024;11:1348456 [FREE Full text] [doi: 10.3389/fnut.2024.1348456] [Medline: 38445208]
40. Pourmontaseri H, Khanmohammadi S. Demographic risk factors of pro-inflammatory diet: a narrative review. Front Nutr 2024;11:1448806 [FREE Full text] [doi: 10.3389/fnut.2024.1448806] [Medline: 39483779]
41. Drewnowski A, Darmon N. The economics of obesity: dietary energy density and energy cost. Am J Clin Nutr 2005;82(1 Suppl):265S-273S [FREE Full text] [doi: 10.1093/ajcn/82.1.265S] [Medline: 16002835]

XSL•FO
RenderX

42. Gupta S, Hawk T, Aggarwal A, Drewnowski A. Characterizing ultra-processed foods by energy density, nutrient density, and cost. Front Nutr 2019;6:70 [FREE Full text] [doi: 10.3389/fnut.2019.00070] [Medline: 31231655]
43. Monteiro CA, Cannon G, Levy RB, Moubarac J, Louzada ML, Rauber F, et al. Ultra-processed foods: what they are and how to identify them. Public Health Nutr 2019;22(5):936-941 [FREE Full text] [doi: 10.1017/S1368980018003762] [Medline: 30744710]
44. El Ansari W, Stock C, Mikolajczyk RT. Relationships between food consumption and living arrangements among university students in four European countries: a cross-sectional study. Nutr J 2012;11:28 [FREE Full text] [doi: 10.1186/1475-2891-11-28] [Medline: 22531503]
45. Shivappa N, Hebert JR, Neshatbini Tehrani A, Bayzai B, Naja F, Rashidkhani B. A pro-inflammatory diet is associated with an increased odds of depression symptoms among Iranian female adolescents: a cross-sectional study. Front Psychiatry 2018;9:400 [FREE Full text] [doi: 10.3389/fpsyt.2018.00400] [Medline: 30210375]
46. Sánchez-Villegas A, Ruíz-Canela M, de la Fuente-Arrillaga C, Gea A, Shivappa N, Hébert JR, et al. Dietary inflammatory index, cardiometabolic conditions and depression in the Seguimiento Universidad de Navarra cohort study. Br J Nutr 2015;114(9):1471-1479. [doi: 10.1017/S0007114515003074] [Medline: 26344165]
47. Kheirouri S, Alizadeh M. Dietary inflammatory potential and the risk of incident depression in adults: a systematic review. Adv Nutr 2019;10(1):9-18 [FREE Full text] [doi: 10.1093/advances/nmy100] [Medline: 30615107]
48. Solomou S, Logue J, Reilly S, Perez-Algorta G. A systematic review of the association of diet quality with the mental health of university students: implications in health education practice. Health Educ Res 2023;38(1):28-68 [FREE Full text] [doi: 10.1093/her/cyac035] [Medline: 36441584]
49. Lorant V, Deliège D, Eaton W, Robert A, Philippot P, Ansseau M. Socioeconomic inequalities in depression: a meta-analysis. Am J Epidemiol 2003;157(2):98-112 [FREE Full text] [doi: 10.1093/aje/kwf182] [Medline: 12522017]
50. Liu Y, Zhang N, Bao G, Huang Y, Ji B, Wu Y, et al. Predictors of depressive symptoms in college students: a systematic review and meta-analysis of cohort studies. J Affect Disord 2019;244:196-208. [doi: 10.1016/j.jad.2018.10.084] [Medline: 30352363]
51. Aziz M, Chemnad K, Al-Harahsheh S, Abdelmoneium AO, Baghdady A, Ali R. Depression, stress, and anxiety versus internet addiction in early and middle adolescent groups: the mediating roles of family and school environments. BMC Psychol 2024;12(1):184 [FREE Full text] [doi: 10.1186/s40359-024-01659-z] [Medline: 38570890]
52. Chiang JJ, Lam PH, Chen E, Miller GE. Psychological stress during childhood and adolescence and its association with inflammation across the lifespan: a critical review and meta-analysis. Psychol Bull 2022;148(1-2):27-66. [doi: 10.1037/bul0000351] [Medline: 39247904]
53. Koss KJ, Gunnar MR. Annual research review: early adversity, the hypothalamic-pituitary-adrenocortical axis, and child psychopathology. J Child Psychol Psychiatry 2018;59(4):327-346 [FREE Full text] [doi: 10.1111/jcpp.12784] [Medline: 28714126]
54. Natale B. Childhood Antecedents of Systemic Inflammation in Adolescence and Adulthood : Contributions of Childhood Family Environment and Socioeconomic Context. Pennsylvania: University of Pittsburgh; 2024.
55. Zeng J, Xu Y. Family socioeconomic status and adolescent depression in urban and rural China: a trajectory analysis. SSM Popul Health 2024;25:101627 [FREE Full text] [doi: 10.1016/j.ssmph.2024.101627] [Medline: 38426030]
56. Loofbourrow BM, Scherr RE. Food insecurity in higher education: a contemporary review of impacts and explorations of solutions. Int J Environ Res Public Health 2023;20(10):5884 [FREE Full text] [doi: 10.3390/ijerph20105884] [Medline: 37239614]
57. Nagata JM, Palar K, Gooding HC, Garber AK, Whittle HJ, Bibbins-Domingo K, et al. Food insecurity is associated with poorer mental health and sleep outcomes in young adults. J Adolesc Health 2019;65(6):805-811 [FREE Full text] [doi: 10.1016/j.jadohealth.2019.08.010] [Medline: 31587956]
58. Natale B. Childhood family environment and lifelong inflammatory and cardiometabolic health: a systematic review. Psychosom Med 2024;86(5):A163-A164 [FREE Full text]
59. Cohen S, Janicki‐Deverts D, Chen E, Matthews KA. Childhood socioeconomic status and adult health. In: Biology of Disadvantage: Socioeconomic Status and Health. New York: Annals of the New York Academy of Sciences; 2010:37-55.
60. Chen E, Miller GE, Kobor MS, Cole SW. Maternal warmth buffers the effects of low early-life socioeconomic status on pro-inflammatory signaling in adulthood. Mol Psychiatry 2011;16(7):729-737 [FREE Full text] [doi: 10.1038/mp.2010.53] [Medline: 20479762]
61. Miller GE, Chen E, Parker KJ. Psychological stress in childhood and susceptibility to the chronic diseases of aging: moving toward a model of behavioral and biological mechanisms. Psychol Bull 2011;137(6):959-997 [FREE Full text] [doi: 10.1037/a0024768] [Medline: 21787044]
62. Chen E, Brody GH, Miller GE. Childhood close family relationships and health. Am Psychol 2017;72(6):555-566 [FREE Full text] [doi: 10.1037/amp0000067] [Medline: 28880102]

Abbreviations

BDI-II: Beck Depression Inventory-II

XSL•FO
RenderX

CRP: C-reactive protein
DII: Dietary Inflammatory Index
IOS: Intelligent Ordering System
OR: odds ratio

Edited by S Brini; submitted 08.Sep.2025; peer-reviewed by Y Su, Y Qu; comments to author 13.Feb.2026; revised version received 20.May.2026; accepted 22.May.2026; published 24.Jun.2026.

Please cite as:
Hong P, Hao C, Weiqiang Z, Jie Q, Yimeng Z, Jingyun D, Haihong Q, Yingnan J
Dietary Inflammatory Index and Depressive Symptoms in Chinese University Students Leveraging an Intelligent Ordering System: 3-Year Longitudinal Prospective Cohort Study
J Med Internet Res 2026;28:e83784
URL: https://www.jmir.org/2026/1/e83784
doi:10.2196/83784
PMID:

©Peng Hong, Chen Hao, Zhou Weiqiang, Qian Jie, Zhang Yimeng, Ding Jingyun, Qian Haihong, Jia Yingnan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development of a Validated Lay Checklist (Info Without Side Effects) for Assessing Health Information on Websites: Mixed Methods Study

Ursula Griebler[1], PhD; Irma Klerings[1], Diplom-Kulturwissenschaftlerin; Christina Koscher-Kien[1], PhD; Benedikt Lutz[2], PhD; Eva Krczal[3], PhD; Dominic Ledinger[1], MPH; Iris Mair[1], MSc; Robert Emprechtinger[4], MSc; Filiz Keser Aschenberger[5], PhD; Bernd Kerschner[1], MSc

[1]Department for Evidence-based Medicine and Evaluation, University for Continuing Education Krems, 3500 Krems an der Donau, Austria

[2]Department for Knowledge and Communication Management, University for Continuing Education Krems, Krems an der Donau, Austria

[3]Department for Economy and Health, University for Continuing Education Krems, Krems an der Donau, Austria

[4]QUEST Center for Responsible Research, Berlin Institute of Health at Charité (BIH), Berlin, Germany

[5]Department for Continuing Education Research and Educational Technologies, University for Continuing Education Krems, Krems an der Donau, Austria

Corresponding Author:
Ursula Griebler, PhD
Department for Evidence-based Medicine and Evaluation, University for Continuing Education Krems, 3500 Krems an der Donau, Austria

Abstract

Background: The internet has become a major source of health information; yet, the quality of health information on websites varies considerably. Users' ability to evaluate either the factual accuracy or the trustworthiness of health information on websites is limited, as around half of the European people have limited health literacy. Existing checklists and tools are either prepared for research purposes or to be used by health care professionals. They do not account for the lay user perspective, since they are too long and complicated to be used by laypersons, or were developed for printed health information only.

Objective: The aim of the study was to develop and validate a checklist that enables laypersons to evaluate the trustworthiness of health information on websites without requiring prior training.

Methods: We used a multistage mixed methods approach including (1) a comprehensive literature review to identify existing tools and quality criteria, (2) an expert Delphi study with 6 specialists in patient communication and health information, (3) 2 rounds of cognitive interviews with 19 lay users, (4) application testing on 15 selected web pages with information about health interventions with 20 additional lay users, (5) a determination of the factual correctness of 100 web pages with health information by assessing the difference between the claimed and factual strength of the evidence on these web pages, and (6) validation testing by research team members on these 100 web pages using a Bayesian logistic regression model to analyze the predictive validity. In the final step, we integrated all quantitative and qualitative results to select the final checklist items.

Results: From an initial pool of 1740 items extracted from 73 documents, we systematically reduced the list through multiple evaluation and testing rounds. To ensure the checklist is user-friendly, we involved a diverse group of potential users. The final product, the Info Without Side Effects (iWISE) checklist, contains seven items that assess key aspects of health information trustworthiness, including the absence of advertising, balanced presentation of information, the limited use of professional jargon, origination from an independent organization, citation of sources, mention of scientific validation, and the presence of a publication date. The checklist demonstrated the ability to distinguish between evidence-based and nonevidence-based health information web pages in the German language: the validation testing showed that when all the items were marked with yes, there was a nearly 100% probability that the health information was also factually correct.

Conclusions: The iWISE checklist represents a user-friendly, validated tool for evaluating the trustworthiness of health information about interventions on websites. With only 7 items, it is easy to remember and could significantly improve critical health literacy. Future research should test its reliability for social media posts and health information videos.

(J Med Internet Res 2026;28:e69529) doi:10.2196/69529

KEYWORDS

health information; online health information; health information websites; checklist; critical health literacy

XSL•FO
RenderX

Introduction

With its accessibility and vast resources, the internet has become an increasingly popular source of health information for the general public. Individuals frequently turn to online platforms to seek answers to their health-related questions, to research symptoms, and to explore treatment options. A recent survey in the European Union reported that in 2024, 58% of individuals aged 16 to 74 years used the internet for health-related information-seeking, with the highest rates up to 82% in the Netherlands, 79% in Finland, and 78% in Denmark. In Austria, 65% used the internet to search for information on health issues [1].

While the abundance of health information available online can be empowering, it also presents significant challenges. The quality and reliability of this information vary widely, ranging from evidence-based medical advice to potentially harmful misinformation [2,3]. This disparity raises concerns about lay users' ability to discern trustworthy sources from unreliable ones. The latest European health literacy survey showed that more than half of users have difficulties in deciding whether the information they find online is trustworthy and objective [4]. According to a survey by the Bertelsmann Foundation [5], between 64% and 84% of participants in Germany rely on advertisement-funded or scientifically unverified web content. Available evidence-based websites were known to only 18% to 27% of respondents; yet, the majority surprisingly judged such evidence-based websites as untrustworthy [5]. The ability to judge the trustworthiness of health information depends on an adequate level of health literacy. Yet, the latest Health Literacy Survey in Europe in 2020 showed that at least 1 of 10 (12.4%) participants had inadequate health literacy, and almost every second respondent (47.6%) in the total sample had limited (inadequate or problematic) health literacy, with substantial differences between member states [4].

Health literacy is a multifaceted construct that "entails people's knowledge, motivation, and competences to access, understand, appraise, and apply health information in order to make judgments and take decisions in everyday life concerning health care, disease prevention, and health promotion" [6].

Low health literacy is known to negatively affect health [7]; it leads to poor health decisions and inferior treatment outcomes as well as to higher hospitalization, morbidity, and premature death rates [6]. Low health literacy may also lead to disadvantageous consumer choices when confronted with exaggerated claims about health services, over-the-counter medications, or nutritional supplements. It has been shown that people with lower health literacy have greater difficulty evaluating and differentiating low-quality from high-quality health information [8].

Strengthening health literacy is also an important factor for achieving the United Nations' Sustainable Development Goal Number 3: ensuring healthy lives and promoting well-being for all at all ages [9].

Critical health literacy, an essential aspect of health literacy, can play a crucial role in addressing these issues. Critical health literacy is a higher-order cognitive process that can be defined as "the ability to access, understand, and manage health information as well as the ability to assess its credibility and to critically analyze and, where appropriate, challenge the information" [10]. It leads to individual and community empowerment by raising critical consciousness, improving quality of life and health behavior, and enhancing health-related social and political action [10]. One way to enhance critical health literacy involves using a tool for assessing the reliability of health information. To be useful, a tool must be based on operationalized quality criteria, should be validated, and must be easily understood and applied by laypersons [11,12]. According to a 2020 systematic review [12], none of the many preexisting checklist tools fulfill all these requirements. This is also true for the recently developed Mapping the Quality of Health Information (MAPPinfo) checklist [11], a German-language tool that was later translated into English. However, untrained laypersons may find it too complex to apply in their daily routines: the 19-item tool is supplemented by 13 pages of explanations in its German version [13] and 12 pages in its English [14]. It was validated based on operationalized and evidence-based criteria only by nursing and medical students [11].

The aim of our study was, therefore, to develop a validated checklist for laypersons to evaluate the trustworthiness of health information on websites without prior training and with minimal necessary explanations. Similar to MAPPinfo, we developed and validated our tool for a German-speaking audience and present here an English-language translation.

Methods

Working Definitions

Health Information on Websites

We defined health information on websites as any health-related content accessible on a web page via an online search engine (eg, Google). A website consists of multiple web pages under one domain. In this study, we focused on information about health interventions, defined as treatments claimed or believed to relieve or prevent symptoms or diseases (eg, vitamin C for the treatment of a cold). Health information from social media platforms or artificial intelligence–generated summaries was not included.

Trustworthiness of Health Information on Websites

For the purpose of this study, we defined trustworthiness of health information on websites as the degree to which the information on a given web page can be trusted by laypersons to be reliable and valid. This definition aligns with Viviani and Pasi's [15] conceptualization of trustworthiness as a key dimension of credibility, focusing specifically on the audience's perception. As laypersons generally do not have the domain expertise to directly judge whether health information is factually correct, we use trustworthiness as a proxy for the correctness, that is, the accuracy of the information.

XSL•FO
RenderX

Laypersons

Following the Cambridge Dictionary [16], we defined layperson as someone who is not an expert in or does not have detailed knowledge of health-related topics, that is, persons without a formal health-related education (ie, no medical doctors, nurses, or other health care professionals or health researchers).

Study Design

The Info Without Side Effects (iWISE) checklist was developed using a sequential, multistage mixed methods design, in which the findings from each step iteratively informed focus and data collection of subsequent stages [17]. The development process comprised five steps: (1) a comprehensive literature review, (2) an expert Delphi study, (3) cognitive interviews with lay users, (4) application tests with laypersons, and (5) predictive validation with a test set (see Figure 1 for a short overview of the development process). For a detailed description of the process, please see the Detailed Methods Description section in Multimedia Appendix 1.

Figure 1. Overview of the development process of the Info Without Side Effects (iWISE) checklist.

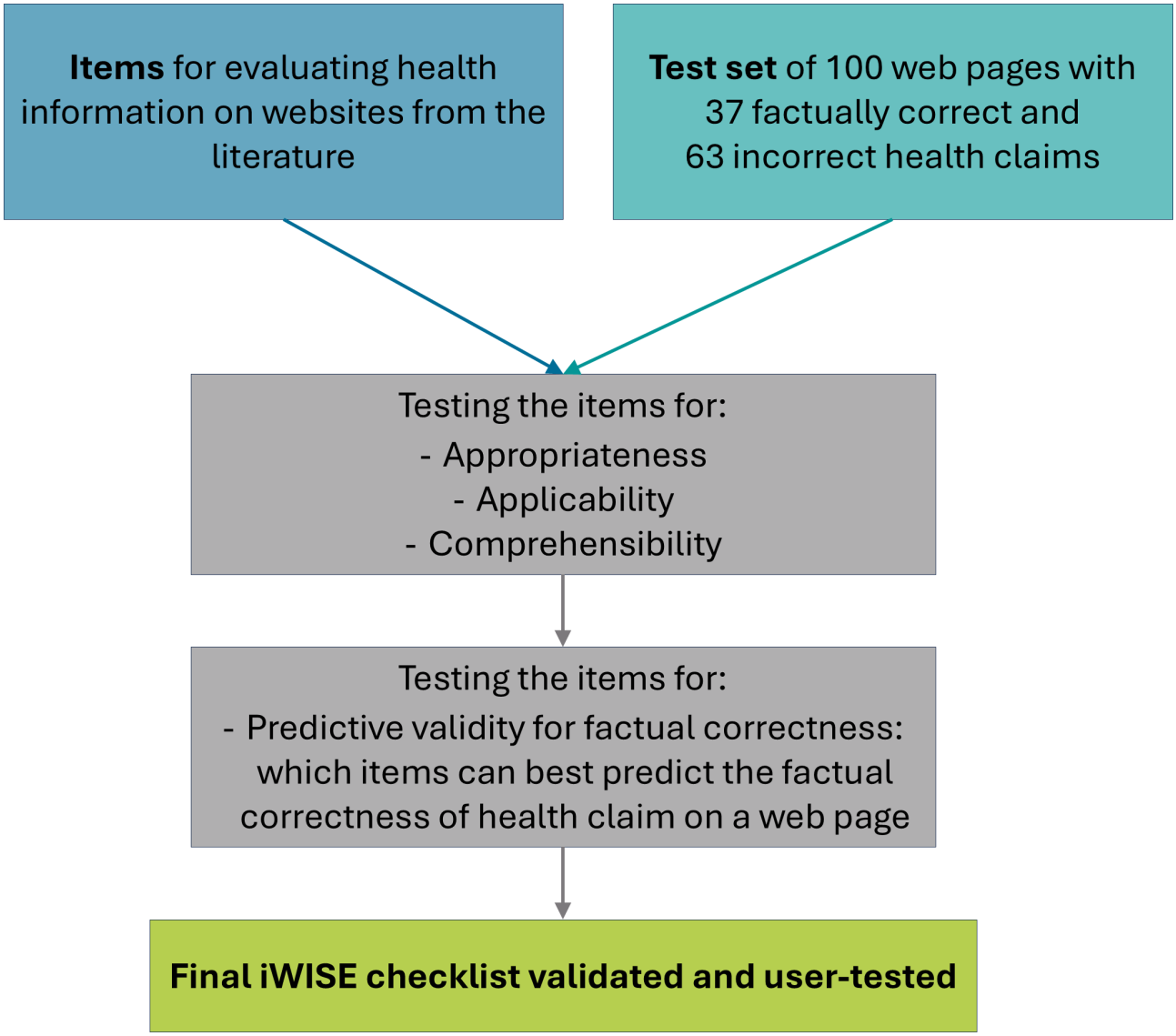

We developed a protocol a priori and published and registered it in the Open Science Framework retrospectively [18]. During the study, we made a few amendments to the protocol.

Ethical Considerations

The study followed the University for Continuing Education Krems' privacy policy (Article 13 General Data Protection Regulation) and was approved by the university's ethics commission (EK GZ 23/2021 - 2024). Interview participants received written study information and provided verbal informed consent; they received approximately US $35 voucher as compensation. Transcripts were deidentified by replacing all names and other identifiable details with alphanumeric codes. To ensure comprehensive reporting, we adhered to the Standards for Reporting Qualitative Research (SRQR) [19] and the American Psychological Association Style Journal Article Reporting Standards for mixed methods research [20].

Literature Search and Selection

We conducted a tailored literature search [21] led by an experienced information specialist (IK) to identify existing checklists and tools as well as conceptual literature relevant to

XSL•FO
RenderX

evaluate (online) health information. The iterative search was conducted between March and September 2021 and combined a variety of search methods:

- Database searches in PubMed, Scopus, Epistemonikos, and Library, Information Science & Technology Abstracts (searched via EBSCO) to retrieve reviews about the evaluation of health information.
- General web searches via Google, Google Scholar, and the Bielefeld Academic Search Engine to identify further published and gray literature.
- Reference list checking of all the (systematic) reviews identified throughout the search process.

One person screened all the results, and a second person verified the inclusion or exclusion decision. From each included (systematic) review on tools and checklists, we extracted the tools' and checklists' names and references.

Data Extraction and Categorization of the Items

Data Extraction

In total, 7 researchers (BK, CK-K, IK, EK, UG, FKA, and Isolde Sommer) extracted the data into a structured matrix in Microsoft Excel. The matrix captured (1) general information (eg, reference type, tool or checklist name, target group, application field, and validation status) and (2) domains, subdomains, item texts, and answer categories or descriptions. For the tools or checklists identified in the (systematic) reviews, we extracted the included names and references of the tool or checklist and added the most frequently mentioned tools or checklists that we had not included yet.

Categorization of the Items

In total, 4 members of the research team (CK-K, EK, IK, and UG) categorized all the extracted items into 5 content categories and adjacent subcategories using an iteratively developed structure (Table S1 in Multimedia Appendix 1). We excluded criteria deemed unsuitable, based on the adapted feasibility criteria by Provost et al [22]: timeliness, expertise independence, externability, generalizability, scope of our project, category without items, unclear wording, and open questions. A second person checked the categorization and exclusions; uncertainties were discussed and resolved among the team.

Item Reduction

First, 4 team researchers (CK-K, EK, IK, and UG) merged duplicate and similarly worded items and translated the English items into German. We standardized the item wording by phrasing all items as questions and aligning the responses so that "yes" indicated trustworthiness. The process was done by one person and checked by a second person, and discrepancies or conflicts were resolved by discussion.

Second, we rated each item's appropriateness (relevance for assessing trustworthiness) and applicability (ease of use without prior knowledge) on 5-point scales in an Excel matrix. In total, 8 researchers (EK, FKA, BL, UG, BK, CK-K, IK, and Isolde Sommer) independently rated all items. Items rated 1‐2 by all were excluded, and the highest-rated per subcategory were retained, aiming for about 50 items for the Delphi study.

Expert Delphi Process

We conducted an expert Delphi study to reach consensus on the most relevant checklist items [21,22]. Experts with backgrounds in patient communication, health information, and online communication were identified through internet searches, existing contacts, team suggestions, and snowballing. Eligible experts were contacted via email and invited to participate. Of the 12 invited German-speaking experts from Germany and Austria, 6 participated (4 from Austria and 2 from Germany; 4 women and 2 men).

In an initial online meeting (May 2022), we presented the study aims, procedures, and categories of the preliminary item list. The experts rated each item's appropriateness and applicability (1‐5 scale) using an Excel matrix, suggested rewording, and proposed new items. The mean, SD, and range were calculated for each item, and revisions were reviewed in a second Delphi round. The final selection was based on the expert ratings and content relevance. Only minor wording changes were suggested. The result of the 2-step Delphi study was an interim checklist version 1.

Cognitive Interviews

We used cognitive interviewing to test the interim checklist version 1 with laypersons, focusing on item comprehension and usability. Both think-aloud and verbal probing techniques were applied [23,24].

Participant Selection and Sample

Participants were purposively sampled using a maximum variation strategy [25] to reflect diversity in age, gender, education, and migrant background. Eligible were adults (aged ≥18 years) of any gender with a basic education and no formal health-related training, who regularly searched for health information on websites, were fluent in German, and had no cognitive limitations. Recruitment strategies included distributing a study invitation flyer via personal social media channels (Facebook and WhatsApp) and the Medizin-Transparent platforms (Facebook and Twitter), using established contacts with self-help groups, and drawing on our personal networks. In total, 19 interviews were conducted.

Data Collection

Interviews were conducted via Microsoft Teams (n=16) or face-to-face (n=3) and were audio-recorded and transcribed. An interview guide and a visually formatted version of the interim checklist version 1 were used. Participants were sent a link to a sample health information web page in advance and were asked during the interview to apply the checklist to this web page, comment on problematic wording, and evaluate the checklist's layout, usability, and overall quality. Interviewers also completed a structured observation protocol. Interviews lasted 33‐91 minutes (mean 56, SD 14 minutes). Data saturation was reached during the final interviews, with no new relevant issues emerging.

Data Analysis

We analyzed the data using the framework approach [26]. In total, 3 researchers (EK, CK-K, and UG) identified and summarized any problems (eg, comprehension, layout, and

XSL•FO
RenderX

instructions) in a structured matrix and discussed any uncertainties and revision suggestions in several rounds. Cognitive testing was conducted in 2 waves; after approximately half of the interviews, the checklist items were revised and reordered, and explanatory subtitles were added. The final analysis led to the interim checklist version 2, which served as the basis for the subsequent application testing with additional lay users.

Application Tests

Application Testing With Lay Users

We evaluated the interim checklist version 2 with 20 additional lay users. Participants applied the checklist to 15 purposively selected health information web pages drawn from a larger set of 100 web pages representing both correct and incorrect information on the health intervention. After submitting their checklists, the participants completed a short online questionnaire rating the ease of use (from 1=very easy to 10=very difficult), identifying difficult or important items, and suggesting wording changes. Each participant received approximately US $265 voucher.

The quantitative (descriptive) and qualitative (content-analytic) results informed the final item selection. Fleiss κ was calculated for each item to identify those with a low interrater agreement and limited suitability for lay assessment.

Application Testing With Research Team Members

Research team members (BL, IK, DL, BK, IM, and UG) also applied the checklist to all 100 web pages. Two of these researchers independently rated each web page, with disagreements resolved through discussion. The agreement of the dual reviewer ratings per item ranged from 66 to 97 of the 100 web pages. Trustworthiness was defined as answering "yes" to 16 - 23 items.

Researchers documented the perceived item difficulty, importance, and ambiguity. These qualitative observations, together with the mean differences between the expert and lay ratings, contributed to the final selection of checklist items.

Predictive Validity With a Test Set and Creating the Final Checklist

Operationalization of Factual Correctness as a Proxy for Trustworthiness

Developing an objective operationalization of trustworthiness is challenging. After extensive discussion within the research team, we concluded that factual correctness was the only feasible proxy for trustworthiness in this context.

Therefore, to assess the predictive validity, we operationalized trustworthiness as the factual correctness of a web page's answer to a given health question (ie, health claim). We compiled a test set of 100 web pages covering 10 common health questions, with approximately half providing correct and half incorrect information (see Table S8 in Multimedia Appendix 1 for a full list).

The factual strength of evidence for these 10 questions had been established through recent evidence syntheses conducted by Medizin transparent [27], a certified signatory of the International Fact Checking Network at the Poynter Institute [28]. To summarize the evidence behind each health claim, the team uses an ultra-rapid evidence synthesis method [29], using systematic searches, critical appraisal, and Grading of Recommendations, Assessment, Development, and Evaluation ratings [30]. We translated these ratings into a 7-point scale (−3 to 3), with positive values indicating increasing evidence for effectiveness, negative values indicating evidence for ineffectiveness, and 0 indicating insufficient evidence for the health claim.

We developed a parallel 7-point scale to rate the claimed strength of the evidence on each web page, based on commonly used lay terminology. Two researchers independently rated all the web pages, resolving disagreements through discussion.

Factual correctness was then defined as the difference between the factual and claimed strength of evidence. The web pages were classified as incorrect when this difference exceeded 1, while differences of 0 or 1 were considered correct. When the factual evidence was insufficient (0), the web pages implying possible effectiveness or ineffectiveness (±1) were also classified as incorrect.

Statistical Analysis for Predictive Validity

We examined which checklist items best predicted the factual correctness of health information on a web page. Using the item ratings from the application tests and the correctness ratings from the test set of 100 web pages, we estimated each item's predictive validity with a Bayesian logistic regression model (brms in R) [31]. The dependent variable indicated whether a web page provided correct information (1) or incorrect information (0). The model included an interaction between each item and its rating to allow item-specific effects.

Creating the Final Checklist

The final item selection was based on each item's predictive validity, supplemented by the laypersons' assessments (difficulty, importance, and comments), research team observations, and quantitative indicators of the item performance (interrater agreement and differences between the expert and lay ratings). All factors were considered collectively in determining the final checklist items. This means that no single factor was decisive for inclusion in the final checklist; rather, decisions were based on the combined results, allowing for triangulation across the different inputs.

To support practical application, we created brief explanatory notes for each item and made the checklist with explanations available on the iWISE project website [32]. For this paper, both the items and explanations were translated from German to English: 1 team member drafted the translation, and the full team reviewed, commented on, and finalized it through group discussion.

Results

Literature Search and Selection

The database and web searches identified 200 documents that were further evaluated. We also identified 11 relevant reviews

XSL•FO
RenderX

or pertinent papers, of which we checked the included tools and checklists (Table S2 in Multimedia Appendix 1). Altogether, 73 documents met our eligibility criteria and were included for data extraction. Figure 2 shows a flowchart of all the documents and items through the development process of the iWISE checklist.

Figure 2. Flow diagram showing the methodological development steps (left) and corresponding sample sizes (right) for the mixed methods study to develop the Info Without Side Effects (iWISE) checklist, involving experts, lay persons, and a formal validation phase.

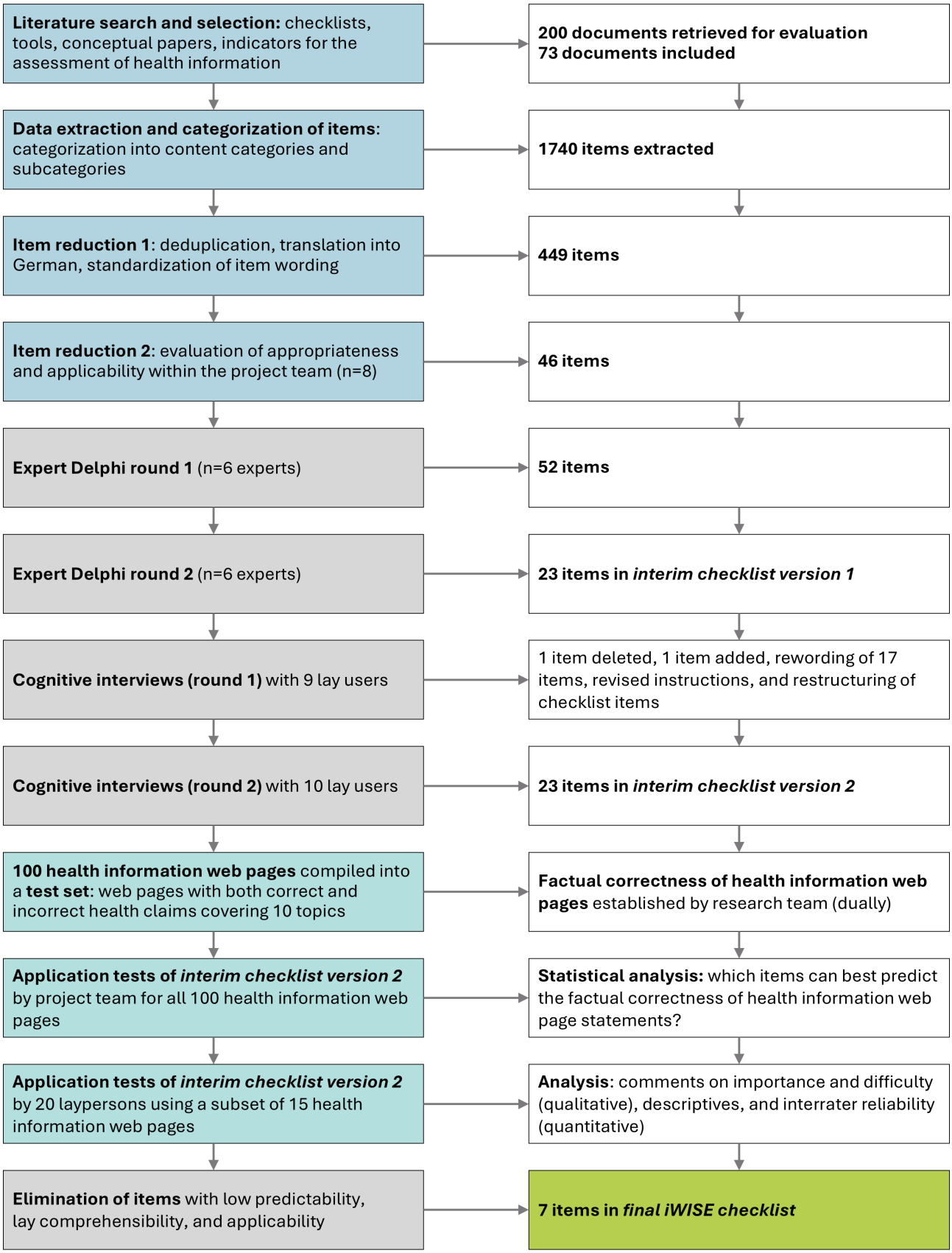

XSL•FO
RenderX

Data Extraction and Categorization of the Items

We performed a data extraction of the general information for all 73 included documents. Of these, 36 were journal papers, 27 were web pages, 3 were books or book sections, 4 were reports, and 2 were conference papers, and they spanned the literature from 1997 to 2021 (see Tables S3 and S4 in Multimedia Appendix 1for a complete list of all included references). Of the 73 documents, 46 (63%) were tools or checklists, more than half of which were aimed at laypersons, consumers, or patients (27/46, 59%), and approximately one-third at experts (13/46, 28%), that is, health information producers or health care professionals, and 3 (7%) did not have a specific target audience.

We extracted a total of 1740 items and categorized them into 5 content categories or marked them as excluded. Table 1 shows the subcategories for each category along with a description. The largest number of items (450/1740, 26%) belonged to the transparency category, followed by the presentation of information category (402/1740, 23%). The least number of items belonged to the functional or technical aspects category (163/1740, 9%), which is unsurprising, as this was not our search's main focus (see Table S1 in Multimedia Appendix 1 for the number of items in each subcategory and respective examples). About 14% (237/1740) of the items had to be excluded, either because of the wrong scope, the need for prior knowledge, the lack of generalizability, or the need to use an external source to apply the item, all of which we defined a priori as reasons for exclusion.

XSL•FO
RenderX

Table . Name and description of the categories and subcategories for the items for evaluating health information on websites extracted from the literature.

Category and subcategory	Description
Functional or technical aspects (deduplicated items: 25; used in the first Delphi round: 0)	
F.0 General technical aspects	A broad category if the item is generally about technology, or several technical aspects are covered at once
F.1 Technical accessibility	Technical accessibility of the website
F.2 Navigation	Navigation through the website
F.3 Interactivity	Interactive offers on the site
F.4 Accessibility or personalization	Possibility of customization (of the presentation) to individual needs
F.5 Other (technology)	An item belongs in this category if it does not fit into any of the available subcategories
Transparency (deduplicated items: 199; used in the first Delphi round: 9)	
T.0 General transparency	A category if the item is generally about transparency and cannot be assigned to any of the subcategories (or must be assigned to several)
T.1 Up-to-date information[a]	Information on the content's actuality is available
T.2 Author and copyright	Identification of the content authorship or creator, copyright information
T.3 Author information	Further information that provides an indication of the content creator's (author's) context
T.4 Disclosure and funding	Identification of funding sources, cooperation partners, conflicts of interest, identification of advertising
T.5 Standards or certificates	Information on whether the content creators adhere to certain external guidelines or whether the source is externally certified
T.6 Privacy and data protection	Information or functions regarding privacy and data protection
T.7 Target group	Definition of the website's target group
T.8 Purpose	Explicit mention of the website's purpose and goals
T.9 Other (transparency)	An item belongs in this category if it does not fit into any of the available subcategories
Presentation of information (deduplicated items: 66; used in the first Delphi round: 22)	
I.0 General presentation of information	A category if the item is generally about the presentation of information and cannot be assigned to any of the subcategories
I.1 Balance	Presentation of information is one-sided, or different aspects are highlighted
I.2 References	Details on the origin of the information provided are available
I.3 Level of evidence	Information on the categorization of the significance or evidence level or quality of the information is available
I.4 Quality assurance	Presence of quality assurance processes
I.5 Methods	The procedure for creating the content is described
I.6 Further information	Indication of further websites, literature, contacts
I.7 Other (presentation of information)	Not meant are cited sources or the author's contact information
Presentation (linguistic and visual) (deduplicated items: 99; used in the first Delphi round: 7)	
D.0 General presentation	A category if the item is generally about visual and linguistic presentation or if several subcriteria are covered at once
D.1 Comprehensibility	Appropriate presentation of information for the target group (content and form)
D.2 Layout	The website and content's layout and design
D.3 Linguistic style	Type of information preparation (neutral, emotional, sensationalist)
D.4 Formal correctness	Reference to the care taken in content creation: are there formal errors in the texts or presentation?

XSL•FO
RenderX

Category and subcategory	Description
D.5 Other (presentation)	An item belongs in this category if it does not fit into any of the available subcategories
User perception (deduplicated items: 60; used in the first Delphi round: 8)	
N.0 General user perception	A category if the item is generally about user perception and cannot be assigned to any of the subcategories
N.1 Emotion	What feelings does the health information trigger?
N.2 Familiarity and reputation	Is the source already known? How are the source's and site creator's reputation assessed?
N.3 Trustworthiness of content	Individual assessment of the content's trustworthiness
N.4 Trustworthiness of references	Individual assessment of the trustworthiness of the sources and of further links
N.5 Trustworthiness in general	General assessment of the trustworthiness of the entire website or the entire offer
N.6 Relevance or usefulness	Individual assessment of the content's usefulness or practical applicability
N.7 Other (user perception)	An item belongs in this category if it does not fit into any of the available subcategories

[a]Italicized subcategories contained items used in the first Delphi round.

Item Reduction

The first step in consolidating the item list resulted in a reduction from 1740 to 449 items. A further reduction step using the research team members' ratings of the items' appropriateness and applicability resulted in a list of 46 items (see Table S1 in Multimedia Appendix 1 for the content distribution).

Expert Delphi Process

The descriptive results of the 46 items rated in Delphi round 1 are shown in Table S5 in Multimedia Appendix 1. Based on the expert feedback, the items were revised from questions to statements, and 6 new items were suggested. In round 2, the experts rated 52 items (the revised 46 plus 6 additions); the results are provided in Table S6 in Multimedia Appendix 1. The applicability ratings showed greater variation than the appropriateness ratings across the rounds (Table S7 in Multimedia Appendix 1).

After completing both rounds, the research team met to discuss the results and to select items. No fixed a priori criteria were set, as the decisions required a comprehensive review of the data. The experts generally rated appropriateness higher than applicability (mean ranges: 3.2 - 5.0 vs 2.2 - 4.8). We used a stepwise selection approach: first excluding items with a mean appropriateness ≤4.0 or applicability ≤3.0, then removing items with an applicability from 3.0 to 4.0. We ensured that all content categories remained represented and reinstated the highest-rated item where necessary (eg, an item on quality assurance with an applicability rating of 3.17). Redundant items were removed.

This process resulted in 23 items forming the interim checklist version 1. The selected items had higher average ratings for both appropriateness and applicability than the excluded items (Table 2).

XSL•FO
RenderX

Table . Results of expert Delphi round 2 for the items for evaluating health information on websites; ratings for each item subcategory presented separately for selected and nonselected items.

Category and subcategory	Items not selected			Items selected		
	Items, n	Appropriateness, mean (SD)	Applicability, mean (SD)	Items, n	Appropriateness, mean (SD)	Applicability, mean (SD)
Presentation (linguistic and visual)						
D.1 Comprehensibility	2	3.52 (1.25)	3.27 (0.98)	1	4.00 (1.15)	4.50 (0.76)
D.2 Layout	1	3.67 (0.75)	4.17 (0.69)	0	—[a]	—
D.3 Linguistic style	3	4.17 (0.80)	3.50 (0.73)	1	4.33 (0.75)	4.17 (0.69)
Presentation of information						
I.1. Balance	9	4.35 (0.64)	3.35 (0.98)	7	4.53 (0.73)	4.33 (0.76)
I.2 References	2	4.58 (0.75)	2.83 (0.64)	1	4.67 (0.75)	3.67 (0.94)
I.3 Level of evidence	1	4.67 (0.75)	3.67 (0.75)	1	4.83 (0.37)	3.67 (0.94)
I.4 Quality assurance	0	—	—	1	4.00 (0.82)	3.17 (0.79)
I.5 Methods	1	4.50 (1.12)	3.17 (1.07)	2	4.58 (0.79)	3.67 (0.79)
I.6 Further information	0	—	—	1	4.33 (0.75)	4.67 (0.47)
User perception						
N.1 Emotion	0	—	—	1	4.50 (0.76)	4.17 (0.90)
N.2 Familiarity and reputation	1	4.17 (1.07)	2.83 (0.69)	0	—	—
N.3 Trustworthiness of content	2	4.17 (0.70)	3.25 (0.54)	1	4.00 (0.82)	4.00 (1.15)
N.4 Trustworthiness of references	1	5.00 (0.00)	3.67 (1.11)	0	—	—
N.5 Trustworthiness in general	1	4.17 (0.69)	3.83 (0.69)	1	4.00 (0.82)	4.00 (0.82)
Transparency						
T.1 Up-to-date information	1	4.50 (0.50)	4.33 (0.75)	1	5.00 (0.00)	4.83 (0.37)
T.2 Author and copyright	1	4.67 (0.47)	3.83 (0.69)	2	4.67 (0.47)	4.42 (0.68)
T.4 Disclosure and funding	3	4.56 (0.67)	3.50 (0.97)	1	4.83 (0.37)	4.17 (0.90)
T.5 Standards or certificates	0	—	—	1	4.33 (0.94)	4.33 (0.75)

[a]Not applicable.

Cognitive Interviews

We conducted 2 rounds of cognitive interviews, involving 9 persons in the first round and 10 in the second. We managed to recruit a diverse set of persons for inclusion, with 58% (11/19) women, 58% (11/19) between 40 and 64 years of age, 11% (2/19) and 26% (5/19) with a lower educational background of either compulsory school only or apprenticeship or vocational secondary school, respectively, and 21% (4/19) with a migration background (Table 3).

The first-round results showed that many words used in the interim checklist version 1 were difficult to understand for the users, and some items were rated as less important to the lay users than to the experts. Therefore, we made large modifications to the checklist: we reworded 17 of the 23 items, deleted 1 item, and added another; we revised the instructions on how to interpret the checklist results; and we reordered the items in a way in which the information on a web page would usually be presented to avoid too much scrolling by the users.

XSL•FO
RenderX

In the second round, further modifications to the interim checklist version 1 were implemented, for example, adding subtitles to structure the checklist and provide more guidance to the users. Two additional items were tested with the last 7 cognitive interviews.

After analysis and discussions among the whole research team, a second interim checklist version was created for use in the application test (Table 4). The wording of 13 items was changed, or text was added to the item to give tips on where to find the respective information on the web page. Furthermore, the order of the items was changed once more.

Table . Characteristics of the participants in cognitive interviews assessing the comprehensibility and applicability of interim checklist version 1, and in application tests with lay users evaluating the applicability of interim checklist version 2 of the iWISE[a] checklist.

	Cognitive interviews (n=19), n (%)	Application tests (n=20), n (%)
Gender		
Women	11 (58)	12 (60)
Men	8 (42)	8 (40)
Age groups (years)		
19 - 39	5[b] (26)	9 (45)
40 - 64	11 (58)	10 (50)
65 - 84	3 (16)	1 (5)
Highest education level		
Compulsory school	2 (11)	1 (5)
Apprenticeship or vocational secondary school	5 (26)	10 (50)
Upper secondary school or vocational upper secondary school	8 (42)	8 (40)
College or academy or university	4 (21)	1 (5)
Migration background		
No	15 (79)	20 (100)
Yes	4[c] (21)	0 (0)

[a]iWISE: Info Without Side Effects.

[b]One participant was aged 16 years.

[c]Germany, Turkey, Hungary, and the United States.

XSL•FO
RenderX

Table . Literal item wording for interim checklist version 2 items for evaluating health information on websites, with the results of the application test with lay users and interrater agreement measures.

	Item text	Items perceived as difficult to answer by lay users (n=20), n (%)	Items perceived as important by lay users (n=20), n (%)	Regression coefficient	Error	96% CI[a]	Agreement among lay users (Fleiss κ)	Agreement among research team members[b]	Item part of the final checklist with wording
1	The title or subtitle is factual and neutral.	1 (5)	0 (0)	0.56	0.24	0.07 to 1.02	0.30	80	[c]-
2	The health information does not contain advertising related to the health problem.	1 (5)	4 (20)	0.46	0.25	−0.02 to 0.94	0.52	85	-
3	The health information does not contain any advertising.	3 (15)	4 (20)	0.82	0.31	0.22 to 1.44	0.50	84	The health information does not contain advertising.
4	The health information is from an independent institution that is unlikely to make money from others' health. For example, not from a company that sells medicines or food supplements. Tip**:** Such information is often found in the imprint.	6 (30)	14 (70)[d]	0.67	0.24	0.2 to 1.14	0.48	81	The health information is from an independent institution that presumably does not make any money from our health (eg, no suppliers of medicines or food supplements).
5	There is a quality seal on the website. Examples of reliable quality seals are: the Action Forum Health Information System (AFGIS) logo and the International Fact Checking Network.	5 (25)	1 (5)	1.16	0.45	0.27 to 2.06	0.39	94	-

XSL•FO
RenderX

	Item text	Items perceived as difficult to answer by lay users (n=20), n (%)	Items perceived as important by lay users (n=20), n (%)	Regression coefficient	Error	96% CI[a]	Agreement among lay users (Fleiss κ)	Agreement among research team members[b]	Item part of the final checklist with wording
6	I feel the information is presented in a balanced way. For example, different treatment options are described. Or besides positive effects, side effects or disadvantages are also described.	1 (5)	11 (55)[d]	1.12	0.28	0.57 to 1.66	0.30	72	I feel the information is presented in a balanced way (the health information describes, eg, advantages and disadvantages and different treatment options).
7	The health information describes treatment effects that I can feel myself. For example, “the feeling of dizziness decreases” instead of “blood pressure decreases.”	6 (30)	1 (5)	0.48	0.23	0.04 to 0.93	0.16	66	-
8	The health information also tells me if and what consequences there are if I choose not to have a treatment.	8 (40)[e]	1 (5)	0.52	0.47	−0.42 to 1.4	0.08	74	-
9	The health information mentions whether there are differences between men and women. For example, differences in a treatment’s symptoms, effects, or side effects.	4 (20)	2 (10)	0.35	0.45	−0.53 to 1.21	0.22	87	-

XSL•FO
RenderX

	Item text	Items perceived as difficult to answer by lay users (n=20), n (%)	Items perceived as important by lay users (n=20), n (%)	Regression coefficient	Error	96% CI[a]	Agreement among lay users (Fleiss κ)	Agreement among research team members[b]	Item part of the final checklist with wording
10	The health information does not make exaggerated or sensational statements.	1 (5)	1 (5)	0.5	0.25	0.01 to 0.97	0.28	72	-
11	The language is factual and neutral.	0 (0)	5 (25)	0.43	0.24	−0.05 to 0.9	0.12	72	-
12	Technical terms are used sparingly, and their meanings are explained.	0 (0)	6 (30)	0.67	0.26	0.16 to 1.2	0.23	78	Technical terms are used sparingly, and their meaning is explained.
13	It is clear to which target group the health information applies. For example, for men or women, people with certain conditions, etc.	3 (15)	1 (5)	0.24	0.23	−0.23 to 0.68	0.09	76	-
14	The health information indicates how well a fact is scientifically proven.	7 (35)	11 (55)[d]	0.87	0.24	0.4 to 1.34	0.30	80	The health information indicates how well the facts claimed are scientifically supported.
15	The health information provides detailed references for the facts mentioned. For example, a list of sources or links to the studies mentioned.	2 (10)	7 (35)	0.69	0.27	0.16 to 1.2	0.62	85	The health information states in detail which sources are behind the facts mentioned (bibliography, links to studies, etc).
16	The health information clearly states that only a doctor can assess my health problem.	0 (0)	4 (20)	0.7	0.3	0.12 to 1.29	0.41	75	-

XSL•FO
RenderX

	Item text	Items perceived as difficult to answer by lay users (n=20), n (%)	Items perceived as important by lay users (n=20), n (%)	Regression coefficient	Error	96% CI[a]	Agreement among lay users (Fleiss κ)	Agreement among research team members[b]	Item part of the final checklist with wording
17	After reading the health information, I feel I can make a decision without pressure.	4 (20)	1 (5)	0.53	0.24	0.06 to 1.01	0.10	67	-
18	It is clear when the health information was created or updated.	3 (15)	4 (20)	0.68	0.24	0.21 to 1.16	0.67	94	It is clear when the health information was created or updated.
19	It is clear that the health information is up to date. For example, the date it was created or updated is provided.	4 (20)	4 (20)	0.7	0.26	0.19 to 1.19	0.39	88	-
20	It is clear that the health information was written by a person or team with suitable scientific training. For example, medical studies or other health-related training (nursing, pharmacy, biology, etc). Tip: Such information is often found on the “About Us” page or in further links.	5 (25)	9 (45)[d]	0.62	0.34	−0.06 to 1.28	0.34	86	-

XSL•FO
RenderX

	Item text	Items perceived as difficult to answer by lay users (n=20), n (%)	Items perceived as important by lay users (n=20), n (%)	Regression coefficient	Error	96% CI[a]	Agreement among lay users (Fleiss κ)	Agreement among research team members[b]	Item part of the final checklist with wording
21	It is clear that the health information has been checked by a person with appropriate scientific training. For example, medical studies or other health-related training (nursing, pharmacy, biology, etc). Tip: Such information is often found on the "About Us" page or in further links.	9 (45)[e]	6 (30)	0.43	0.45	−0.48 to 1.3	0.31	86	-
22	It is described how the information was created. For example, which studies were considered and why, and which were not. Tip: Such information is often found on the "About Us" page or in further links.	8 (40)[e]	1 (5)	1.49	0.52	0.49 to 2.53	0.22	85	-

XSL•FO
RenderX

	Item text	Items perceived as difficult to answer by lay users (n=20), n (%)	Items perceived as important by lay users (n=20), n (%)	Regression coefficient	Error	96% CI[a]	Agreement among lay users (Fleiss κ)	Agreement among research team members[b]	Item part of the final checklist with wording
23	The health information states whether how readers were involved in creating the information. Tip: Such information is often found on the "About Us" page or in further links.	11 (55)[e]	0 (0)	3.39	1.07	1.57 to 5.7[f]	0.02	97	-

[a]Higher values denote better predictive validity.

[b]Before the disagreements were solved through discussion or by involving a third person; 100=complete agreement.

[c]Not applicable.

[d]Items were perceived as important by 45% or more of the lay users.

[e]Items were perceived as difficult by 40% or more of the lay users.

[f]This item was only ticked in 8 health information web pages; therefore, the practical value of the predictive validity is low.

Already in the first round, lay users made remarks on single items that were less important to them than to the experts, for example, "It is clear to which target group the health information applies." As this was an item deemed important by the experts in the Delphi study, we kept it for the application test but added an example because the term target group was ambiguous for some users. In general, many of the suggested modifications led to shorter and more straightforward sentences and the explicit use of examples (ie, "for example"). We also removed redundancies and used words that were unlikely to be misunderstood, and we avoided negative constructions, as these are difficult to understand in German. During the analysis, we realized that some items may be more important for health information creators than for health information consumers, for example, one item that we dropped after the cognitive interviews: "The health information mentions how well the differences between men and women have been researched."

Application Tests With Lay Users

We conducted application tests with 20 further lay users (see participant characteristics in Table 3). All 20 lay users returned quantitative and qualitative feedback on using the interim checklist version 2. The mean answer to the general question "On a scale from 1 to 10: how easy or difficult did you find it to use the checklist?" (1=very easy and 10=very difficult) was 4.0 (SD 1.8; range 2-9). We asked the participants to check the items they found difficult to answer, and 4 items were marked by either 8, 9, or 11 participants (40% - 55%; Table 4). Explanations as to why these items were difficult were also given. All this information was taken into consideration in our discussion on the final selection of items for the checklist. As evidenced in Table 4, none of the 4 items that were considered difficult by nearly half or more than half of the lay users made it to the final checklist. The reasons for the difficulties in applying certain items were manifold. For example, regarding item 23 "The health information states whether and how readers were involved in creating the information," over half of the lay users stated that it was either difficult to find that information on the web page, or the concept of reader involvement in creating health information was unclear to them. Furthermore, no one rated this item as especially important.

Conversely, 4 items were also rated as especially important by 9, 11, or 14 participants (45% - 70%; Table 4), and 3 of these made it to the final checklist. Multiple users had trouble understanding item 14 "The health information indicates how well a fact is scientifically proven," and 7 also marked it as a difficult item. However, more than half also rated it as especially important, and this item made it to the final checklist.

For item 19 "It is clear that the health information is up to date," one-third of the participants demanded a clear time frame for their judgment of up-to-dateness. Since the time frame depends on the health information topic and is highly variable, this item was not included in the final checklist, but a more neutral item on currency was included ("It is clear when the health information was created or updated.").

Predictive Validity With a Test Set and Creating the Final Checklist

Of the 100 health information web pages in the test set, 37 were factually correct; that is, had little or no difference between the claimed and factual strength of evidence; 63 web pages were incorrect. The regression coefficients of the model for predicting the correctness of the health information web pages are shown in Table 4, along with the estimated errors and 96% CIs. The higher the values, the better the respective item's predictive

XSL•FO
RenderX

validity in predicting the trustworthiness of a health information web page.

Table 4 also shows the agreement among the lay users and among the experts, respectively. Based on the initial results, we tested a few item combinations until we arrived at the final checklist (Textbox 1). As stated earlier, all the quantitative and qualitative results were integrated into our final decision.

Textbox 1. Seven items of the final Info Without Side Effects (iWISE) checklist.

1. The health information does not contain advertising.
2. I feel the information is presented in a balanced way (the health information describes, for example, advantages and disadvantages, different treatment options, ...).
3. Technical terms are used sparingly, and their meaning is explained.
4. The health information is from an independent institution that presumably does not make any money from our health (e.g., no suppliers of medicines or food supplements, ...).
5. The health information states in detail which sources are behind the facts mentioned (bibliography, links to studies, ...).
6. The health information indicates how well the facts claimed are scientifically supported.
7. It is clear when the health information was created or updated.

Some items with the highest regression coefficients for predicting a correct health information web page were not chosen because of other reasons. For example, item 23 “The health information states whether and how readers were involved in creating the information” with the highest predictability values was only fulfilled in 8 of 100 health information web pages. Therefore, the practical value of this item’s predictive validity is low. Furthermore, more than half of the lay users perceived this item as difficult to answer, and the agreement among the lay users in answering this item was very low (Table 4). The situation was similar with item 22: “It is described how the information was created. For example, which studies were considered and why, and which were not.” While this had a good predictive value, 8 of 20 lay users perceived it to be difficult to answer; therefore, we decided not to include it in the final checklist, as only items that are easily answered and clear to laypeople are useful in the final checklist.

After several discussion rounds among the whole research team, we arrived at the final iWISE checklist with 7 items (Textbox 1).

According to our statistical model, the probability that the health information is correct if all of these items are marked with “yes” is 99.9%. The probability that the health information is correct if all of the items are marked with “no” is 20.4%. However, this requires that the items are marked correctly by the user, which, in practice, may not be the case.

We published the iWISE checklist in German on our project website [32] and also provided brief explanatory notes for each item to help users apply the checklist. We also created a video explaining all 7 items. The English version of the iWISE checklist, including explanations for lay users, is available in Multimedia Appendix 1 and can also be downloaded as a PDF [33].

Discussion

Main Study Findings

In this mixed methods study, we developed and validated the iWISE checklist, a concise and easy-to-use tool enabling laypersons to evaluate the trustworthiness of health information on websites. To our best knowledge, iWISE is the first checklist that can be applied by laypersons in everyday information-seeking situations without the need for prior training. Consistent with our aims, we developed iWISE in a multistage development process, combining evidence synthesis, expert review, iterative cognitive testing, application tests, and predictive validation, which resulted in 7 items that lay users can both understand and reliably apply. To further support lay users, we also provide brief, plain-language explanations for each item, including examples and tips on where to find the relevant information on a web page.

The final iWISE checklist contains 7 items that assess key aspects of health information trustworthiness, including the absence of advertising, a balanced presentation of information, the limited use of professional jargon, an origination from an independent organization, the citation of sources, the mention of scientific validation, and the presence of a publication date. The checklist demonstrated the ability to distinguish between evidence-based and nonevidence-based health information web pages in the German language: the validation testing showed that when all items were marked with “yes,” there was a nearly 100% probability that the health information was also factually correct.

Discussion and Comparison With Prior Work

There are 3 main features that characterize the iWISE checklist: it is more user-friendly for laypersons than prior checklists, it was developed with the extensive involvement of laypersons, and it reliably identifies trustworthy health information on web pages, as demonstrated by the validation.

iWISE Is Shorter, Simpler, and More Practical Than Prior Tools

The iWISE checklist is short and easy for lay users to understand and apply when judging the trustworthiness of health information on websites. Numerous tools and checklists are available to assess online health information [12,34-36]; however, these tools were either prepared for research purposes or for use by health care professionals, have not been developed

XSL•FO
RenderX

by considering the lay user perspective, contain a large number of items, or have sophisticated scoring systems that make them impractical for consumers to use [11,22,37,38]. Others are intended for written consumer information, such as the DISCERN instrument [39,40] or the Ensuring Quality Information for Patients tool [41,42], neither of which can be directly transferred to health information web pages. More broadly, the existing approaches to rating health-related websites [43] range from provider-focused quality criteria that contain a list of recommendations for website development and content [44-48], to quality labels or logos that are displayed on screen and represent a provider's commitment to implement or adhere to a set of quality criteria such as the Health On the Net code [49] (which was permanently discontinued in December 2022), to user guidance systems requiring users to navigate long lists of questions [11,12,22,37,38]. These approaches often require more time, expertise, or effort than laypersons can invest during everyday health information-seeking. In contrast, iWISE consists of 7 core items that reflect the key aspects of credibility and quality but remain manageable within typical cognitive limits for lay users, consistent with research on short-term memory capacity [50].

iWISE Was Developed With the Extensive Involvement of Laypersons and Is Tailored to Their Needs

A major strength of iWISE is that lay users were integrated throughout the development process, using cognitive interviews and application testing, and these results were taken into account in the final item selection. Other quality evaluation checklists are mostly based on expert views [51] and, therefore, may not meet the needs of laypersons with a low general education level and no formal health-related education. For example, the MAPPinfo checklist includes items that are justified by either ethics or research evidence, which enhances the methodological rigor but may omit criteria important to users that lack strong research backing [11]. Furthermore, MAPPinfo was user-tested with nursing and medical students rather than true laypersons. In contrast, iWISE was built explicitly for individuals with no specialized health knowledge, including people with lower educational backgrounds. Items requiring prior expertise were eliminated, ensuring that all the remaining items could be answered solely by examining the web page itself. Because a checklist must be usable in real-world situations—often without printing or switching between windows—the final 7 items were intentionally kept short, intuitive, and memorable. The final content of the checklist is also closely aligned with criteria lay users naturally use, as described by Sun et al [51], including trustworthiness, expertise, and objectivity.

iWISE Effectively Identifies Trustworthy Online Health Information

The validation test demonstrated that the 7 iWISE items help distinguish between evidence-based and nonevidence-based web pages on health interventions. The validation method—operationalizing trustworthiness as the factual correctness of the web page's claims—is novel and unique in this field. The factual correctness was fulfilled if the factual strength of the evidence for a certain health question and the claimed strength of the evidence on the web page were the same or only differed slightly. For our validation, we applied the Grading of Recommendations, Assessment, Development and Evaluation system [30], which is widely used in evidence-based medicine. Only items that showed a reasonable association with factual correctness were retained for the final checklist. Factual correctness is also one of the criteria lay users value most highly when evaluating online health information [51]. Importantly, the predictive validity analysis showed that iWISE meets the key criteria for consumer-oriented rating tools described by Hanif et al [52], including readability, feasibility, and practical assessibility.

Limitations

Our checklist has some limitations. With 100 web pages, our validation test set of health information articles was relatively small. We tested our checklist with a limited set of only 10 popular health questions—and with 10 articles for each of these questions, respectively. Our test set might therefore be biased and not be representative of the broad mass of health information that a typical user encounters when searching the web. A different, more varied, and larger test set might have resulted in a slightly different combination of checklist items. However, iWISE is comparable with other checklists that have either used a smaller validation test set or that deal with one health topic or both, for example, the MAPPinfo checklist was tested on 57 web pages [11], the Quality Evaluation Scoring Tool on 45 online articles on Alzheimer disease treatment and prevention [37], and Dobbins et al [38] used 120 web resources and another 107 in a second validation round on healthy aging.

The predictive validity results, one of the factors we considered in selecting the final checklist items, are based on the ratings from the application test with the research team. The laypersons' ratings were not integrated in this validation step. However, the perspectives of the laypersons were considered in other ways, and we specifically ensured that all items were well understood by our lay user group.

Our checklist has only been validated for information on treatments or health decisions. It has not been tested on websites lacking interventional information. The checklist might therefore not be valid for judging the accuracy of information on symptoms for the diagnosis of a disease or on the health risks of noninterventional hazard exposure.

The iWISE checklist was tested with German-speaking laypersons in Austria and validated on a test set of websites in the German language. We cannot rule out that the checklist works differently when tested in other languages. In addition, our checklist is not tested with social media postings or health information videos and cannot be used for evaluating medication leaflets.

Conclusions

The iWISE checklist offers an accessible and empirically validated way for laypersons to judge the trustworthiness of online health information. By reducing a complex evaluation process to 7 intuitive items, iWISE has the potential to strengthen critical health literacy in everyday decision-making. In a digital environment where misinformation is widespread, such an accessible tool can help individuals navigate health

XSL•FO
RenderX

claims with greater confidence while also encouraging health information providers to adopt more transparent and balanced communication practices.

Because iWISE was developed and validated in a German-language context, future research should examine its applicability across other languages and cultural settings. Research teams in other countries may wish to validate or adapt the checklist for their own health information environments. In addition, the checklist was developed for web pages and should be tested for use with social media posts and health information videos, where misinformation is especially prevalent. Building on this need, our team has already initiated a follow-up project to develop and evaluate a complementary tool specifically tailored to assessing the reliability of health information on social media platforms.

Acknowledgments

The authors thank Evelyn Auer for her excellent help in project administration and Isolde Sommer for her contributions to the data extraction and expert Delphi study. The authors further thank Genevieve Iseult Eldredge for proofreading. During the manuscript preparation, the authors used DeepL, Claude 3.5 Sonnet, and ChatGPT for help with text improvements and the translations from German to English. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

This study was supported by an internal grant for an interdisciplinary research group at the University of Continuing Education Krems, Austria. The funder had no role in the study design, data collection, data analysis, interpretation of results, or the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

BK and UG conceived the idea and led the grant application for funding. UG led the project planning and conduct. UG, BK, CK-K, IK, BL, EK, and FKA wrote the study protocol. IK and BK performed the electronic database searches; IK, BK, FKA, CK-K, BL, EK, and UG performed the supplementary searches and screened the literature. BK, CK-K, EK, FKA, IK, and UG did the data extractions; CK-K, EK, IK, and UG performed the categorization and first condensation of the items. EK, FKA, BL, UG, BK, CK-K, IK, and Isolde Sommer did the item rating for the second condensation step. CK-K planned and carried out the Delphi study. UG, CK-K, and EK carried out and analyzed the cognitive interviews. UG led the application test with the lay users; UG, DL, and IK analyzed the data. BK led the operationalization of health information trustworthiness and compiled the web page test set. BK and IM assessed the strength of evidence for the test set web pages. RE planned and conducted the statistical analyses for the predictive validity. All authors engaged in discussions and decisions on the final checklist. All authors drafted the paper, critically revised it for important intellectual content, and approved the final version for publication. UG and BK act as guarantors. The corresponding author attests that all listed authors meet the authorship criteria and that no others meeting the criteria have been omitted.

Conflicts of Interest

RE received remuneration from Commonwealth Serum Laboratories and Charité during the period 2022 to 2025.

Multimedia Appendix 1
Detailed methods description, supplementary results, and tables on the development, and the final Info Without Side Effects (iWISE) checklist.
[DOCX File, 1439 KB - jmir_v28i1e69529_app1.docx]

References

1. Individuals using the internet for seeking health-related information. Eurostat. URL: https://ec.europa.eu/eurostat/databrowser/view/tin00101/default/table?lang=en [accessed 2025-11-28]
2. Daraz L, Morrow AS, Ponce OJ, et al. Can patients trust online health information? A meta-narrative systematic review addressing the quality of health information on the internet. J Gen Intern Med 2019 Sep;34(9):1884-1891. [doi: 10.1007/s11606-019-05109-0] [Medline: 31228051]
3. Denniss E, Lindberg R, McNaughton SA. Quality and accuracy of online nutrition-related information: a systematic review of content analysis studies. Public Health Nutr 2023 Jul;26(7):1345-1357. [doi: 10.1017/S1368980023000873] [Medline: 37138366]

XSL•FO
RenderX

4. International report on the methodology, results, and recommendations of the European Health Literacy Population Survey 2019-2021 (HLS19) of M-POHL. The HSL19 Consortium of the WHO Action Network M-POHL. URL: https://m-pohl.net/sites/m-pohl.net/files/inline-files/HLS19%20International%20Report.pdf [accessed 2023-08-31]
5. Marstedt G. Das internet: auch ihr ratgeber für gesundheitsfragen? Bevölkerungsumfrage zur suche von gesundheitsinformationen im internet und zur reaktion der ärzte. Bertelsmann Stiftung. URL: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/VV_Studie_Das-Internet-auch-Ihr-Ratgeber_Befragung.pdf [accessed 2023-08-31]
6. Kickbusch I, Pelikan JM, Apfel F, Tsouros AD. Health literacy: the solid facts. World Health Organization. URL: https://iris.who.int/bitstream/handle/10665/128703/e96854.pdf [accessed 2023-08-31]
7. Nutbeam D. The evolving concept of health literacy. Soc Sci Med 2008 Dec;67(12):2072-2078. [doi: 10.1016/j.socscimed.2008.09.050]
8. Diviani N, van den Putte B, Giani S, van Weert JC. Low health literacy and evaluation of online health information: a systematic review of the literature. J Med Internet Res 2015 May 7;17(5):e112. [doi: 10.2196/jmir.4018] [Medline: 25953147]
9. Goals 3: ensure healthy lives and promote well-being for all at all ages. United Nations. URL: https://sdgs.un.org/goals/goal3 [accessed 2024-12-01]
10. Sykes S, Wills J, Rowlands G, Popple K. Understanding critical health literacy: a concept analysis. BMC Public Health 2013 Feb 18;13:150. [doi: 10.1186/1471-2458-13-150] [Medline: 23419015]
11. Kasper J, Lühnen J, Hinneburg J, et al. MAPPinfo—mapping quality of health information: validation study of an assessment instrument. PLoS One 2023;18(10):e0290027. [doi: 10.1371/journal.pone.0290027] [Medline: 37871040]
12. Kasper J, Lühnen J, Hinneburg J, et al. MAPPinfo, mapping quality of health information: study protocol for a validation study of an assessment instrument. BMJ Open 2020 Nov 3;10(11):e040572. [doi: 10.1136/bmjopen-2020-040572] [Medline: 33148762]
13. MAPPinfo: Validierte Checkliste zur Bewertung evidenzbasierter Gesundheitsinformationen. Stiftung Gesundheitswissen. 2022 Mar. URL: https://www.stiftung-gesundheitswissen.de/forschung/mappinfo [accessed 2025-11-25]
14. MAPPinfo: validated checklist for the assessment of evidence-based health information. Stiftung Gesundheitswissen. 2022 Mar. URL: https://www.stiftung-gesundheitswissen.de/forschung/mappinfo [accessed 2025-11-25]
15. Viviani M, Pasi G. Credibility in social media: opinions, news, and health information—a survey. WIREs Data Min Knowl 2017 Sep;7(5):e1209. [doi: 10.1002/widm.1209]
16. Cambridge dictionary. Meaning of “layperson” in English. URL: https://dictionary.cambridge.org/dictionary/english/layperson [accessed 2022-03-28]
17. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs-principles and practices. Health Serv Res 2013 Dec;48(6 Pt 2):2134-2156. [doi: 10.1111/1475-6773.12117] [Medline: 24279835]
18. Griebler U, Kerschner B, Kien C, Klerings I, Lutz B, Krczal E, et al. “Info Without Side Effect”—development and application of a checklist for lay users for evaluating online health information. Open Science Framework. URL: https://osf.io/bfwqh/overview [accessed 2026-04-27]
19. O’Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
20. Mixed Methods Article Reporting Standards (MMARS) information recommended for inclusion in manuscripts that report the collection and integration of qualitative and quantitative data. American Psychological Association. URL: https://apastyle.apa.org/jars/mixed-table-1.pdf [accessed 2024-11-27]
21. Cooper C, Booth A, Husk K, et al. A tailored approach: a model for literature searching in complex systematic reviews. J Inform Sci 2024 Aug;50(4):1030-1062. [doi: 10.1177/01655515221114452]
22. Provost M, Koompalum D, Dong D, Martin BC. The initial development of the WebMedQual scale: domain assessment of the construct of quality of health web sites. Int J Med Inform 2006 Jan;75(1):42-57. [doi: 10.1016/j.ijmedinf.2005.07.034] [Medline: 16169770]
23. Beatty PC, Willis GB. Research synthesis: the practice of cognitive interviewing. Public Opin Q 2007;71(2):287-311. [doi: 10.1093/poq/nfm006]
24. Buers C, Triemstra M, Bloemendal E, Zwijnenberg NC, Hendriks M, Delnoij DMJ. The value of cognitive interviewing for optimizing a patient experience survey. Int J Soc Res Methodol 2014 Jul 4;17(4):325-340. [doi: 10.1080/13645579.2012.750830]
25. Patton MQ. Qualitative Research & Evaluation Methods: Integrating Theory and Practice: Sage Publications; 2014.
26. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013 Sep 18;13(1):117. [doi: 10.1186/1471-2288-13-117] [Medline: 24047204]
27. Medizin-transparent. Cochrane Austria, Department for Evidence-based Medicine and Evaluation. URL: https://medizin-transparent.at/ [accessed 2024-09-10]
28. Medizin transparent—Universität für Weiterbildung Krems (Donau-Universität Krems). International Fact-Checking Network (IFCN). URL: https://ifcncodeofprinciples.poynter.org/profile/medizin-transparent-universitat-fur-weiterbildung-krems-donau-universitat-krems [accessed 2024-09-10]

XSL•FO
RenderX

29. Cochrane Austria, Department for Evidence-Based Medicine and Evaluation. Methodenpapier Medizin-transparent, Version 1.0 vom 5. Donau-Universität Krems. 2018. URL: https://medizin-transparent.at/wp-content/uploads/2022/12/Methodenpapier-Medizin-transparent_v1.0.pdf [accessed 2024-09-10]
30. Balshem H, Helfand M, Schünemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011 Apr;64(4):401-406. [doi: 10.1016/j.jclinepi.2010.07.015] [Medline: 21208779]
31. Bürkner PC. Bayesian item response modeling in R with brms and Stan. J Stat Softw 2021;100(5):1-54. [doi: 10.18637/jss.v100.i05]
32. Checkliste für verlässliche Gesundheitsinfos. Infos Ohne Nebenwirkung. URL: https://www.infos-ohne-nebenwirkung.at/checkliste/ [accessed 2024-11-27]
33. Griebler U, Klerings I, Koscher-Kien C, et al. The iWISE* checklist for trustworthy health information on websites. : University for Continuing Education Krems; 2025. [doi: 10.48341/iWISE]
34. Bernstam EV, Shelton DM, Walji M, Meric-Bernstam F. Instruments to assess the quality of health information on the World Wide Web: what can our patients actually use? Int J Med Inform 2005 Jan;74(1):13-19. [doi: 10.1016/j.ijmedinf.2004.10.001] [Medline: 15626632]
35. Gagliardi A, Jadad AR. Examination of instruments used to rate quality of health information on the internet: chronicle of a voyage with an unclear destination. BMJ 2002 Mar 9;324(7337):569-573. [doi: 10.1136/bmj.324.7337.569] [Medline: 11884320]
36. Song S, Zhang Y, Yu B. Interventions to support consumer evaluation of online health information credibility: a scoping review. Int J Med Inform 2021 Jan;145:104321. [doi: 10.1016/j.ijmedinf.2020.104321] [Medline: 33202372]
37. Robillard JM, Jun JH, Lai JA, Feng TL. The QUEST for quality online health information: validation of a short quantitative tool. BMC Med Inform Decis Mak 2018 Oct 19;18(1):87. [doi: 10.1186/s12911-018-0668-9] [Medline: 30340488]
38. Dobbins M, Watson S, Read K, Graham K, Yousefi Nooraie R, Levinson AJ. A tool that assesses the evidence, transparency, and usability of online health information: development and reliability assessment. JMIR Aging 2018 May 7;1(1):e3. [doi: 10.2196/aging.9216] [Medline: 31518240]
39. Charnock D. The DISCERN Handbook: Radcliff Medical Press, University of Oxford and The British Library; 1998.
40. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999 Feb;53(2):105-111. [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
41. Moult B, Franck LS, Brady H. Ensuring quality information for patients: development and preliminary validation of a new instrument to improve the quality of written health care information. Health Expect 2004 Jun;7(2):165-175. [doi: 10.1111/j.1369-7625.2004.00273.x] [Medline: 15117391]
42. Charvet-Berard AI, Chopard P, Perneger TV. Measuring quality of patient information documents with an expanded EQIP scale. Patient Educ Couns 2008 Mar;70(3):407-411. [doi: 10.1016/j.pec.2007.11.018] [Medline: 18242935]
43. Wilson P. How to find the good and avoid the bad or ugly: a short guide to tools for rating quality of health information on the internet. BMJ 2002 Mar 9;324(7337):598-602. [doi: 10.1136/bmj.324.7337.598] [Medline: 11884329]
44. Winker MA, Flanagin A, Chi-Lum B, et al. Guidelines for medical and health information sites on the internet: principles governing AMA web sites. American Medical Association. JAMA 2000;283(12):1600-1606. [doi: 10.1001/jama.283.12.1600] [Medline: 10735398]
45. The Lida instrument. Minervation. URL: https://www.minervation.com/work/the-lida-instrument/ [accessed 2021-12-06]
46. Silberg WM, Lundberg GD, Musacchio RA. Assessing, controlling, and assuring the quality of medical information on the Internet: caveant lector et viewor—let the reader and viewer beware. JAMA 1997 Apr 16;277(15):1244-1245. [Medline: 9103351]
47. Rippen H, Risk A. e-Health Code of Ethics (May 24). J Med Internet Res 2000;2(2):e9. [doi: 10.2196/jmir.2.2.e9]
48. Leitlinie Evidenzbasierte Gesundheitsinformation. URL: https://www.leitlinie-gesundheitsinformation.de/ [accessed 2024-11-26]
49. Boyer C, Baujard V, Geissbuhler A. Evolution of health web certification through the HONcode experience. Stud Health Technol Inform 2011;169(53-7):53-57. [Medline: 21893713]
50. Miller GA. The magical number seven plus or minus two: some limits on our capacity for processing information. Psychol Rev 1956 Mar;63(2):81-97. [Medline: 13310704]
51. Sun Y, Zhang Y, Gwizdka J, Trace CB. Consumer evaluation of the quality of online health information: systematic literature review of relevant criteria and indicators. J Med Internet Res 2019 May 2;21(5):e12522. [doi: 10.2196/12522] [Medline: 31045507]
52. Hanif F, Read JC, Goodacre JA, Chaudhry A, Gibbs P. The role of quality tools in assessing reliability of the internet for health information. Inform Health Soc Care 2009 Dec;34(4):231-243. [doi: 10.3109/17538150903359030] [Medline: 19919300]

Abbreviations

iWISE: Info Without Side Effects

XSL•FO
RenderX

MAPPinfo: Mapping the Quality of Health Information
SRQR: Standards for Reporting Qualitative Research

Edited by A Stone; submitted 05.Dec.2024; peer-reviewed by B Irfan, C Baur, D Carvalho, E Flaschberger, M Nitsch; revised version received 18.Dec.2025; accepted 18.Dec.2025; published 20.May.2026.

Please cite as:
Griebler U, Klerings I, Koscher-Kien C, Lutz B, Krczal E, Ledinger D, Mair I, Emprechtinger R, Keser Aschenberger F, Kerschner B
Development of a Validated Lay Checklist (Info Without Side Effects) for Assessing Health Information on Websites: Mixed Methods Study
J Med Internet Res 2026;28:e69529
URL: https://www.jmir.org/2026/1/e69529
doi:10.2196/69529

© Ursula Griebler, Irma Klerings, Christina Koscher-Kien, Benedikt Lutz, Eva Krczal, Dominic Ledinger, Iris Mair, Robert Emprechtinger, Filiz Keser Aschenberger, Bernd Kerschner. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Effect of Informational Podcasts on Shared Decision-Making, Anxiety, and Patient Satisfaction in Hospital Visits: Intervention Study

Jannie Christina Frølund[1], RN, MScN, PhD; Anders Løkke[1,2], Dr med, Prof Dr; Hanne Irene Jensen[2,3], RN, PhD, Prof Dr; Flemming Skjøth[2,4], Prof Dr; Ingeborg Farver-Vestergaard[1,2], Prof Dr

[1]Department of Medicine, Lillebaelt Hospital, Beriderbakken 4, Vejle, Denmark

[2]Department of Regional Health Research, University of Southern Denmark, Odense, Denmark

[3]Department of Anaesthesiology and Intensive Care, Lillebaelt Hospital, Vejle and Kolding, Denmark

[4]Research Support Unit, Lillebaelt Hospital, Vejle, Denmark

Corresponding Author:
Jannie Christina Frølund, RN, MScN, PhD
Department of Medicine, Lillebaelt Hospital, Beriderbakken 4, Vejle, Denmark

Abstract

Background: Podcasts provide a platform for delivering patient information. They have the potential to enhance patient engagement in shared decision-making (SDM), reduce anxiety in relation to hospital visits, and improve patient satisfaction. However, their impact on these outcomes in the context of hospital visits remains underexplored.

Objective: This study aimed to examine whether podcasts influence patients' (1) engagement in SDM, (2) anxiety after the hospital visit, and (3) satisfaction with the hospital visit.

Methods: A quasi-experimental design with a nonequivalent comparison group was used. The study was conducted in 3 specialized outpatient clinics at a Danish hospital. Patients were allocated to one of 2 groups: the intervention group, which received access to informational podcasts in addition to standard written information before their hospital visit, and the comparison group, which received only the standard written information. All patients received validated questionnaires to assess SDM (9-item Shared Decision-Making Questionnaire [SDM-Q-9]), anxiety (State-Trait Anxiety Inventory-State), and satisfaction after the consultation.

Results: A total of 240 patients participated. Compared with the control group, the intervention group showed a 15% higher level of SDM (SDM-Q-9) scores (adjusted relative difference=1.15, 95% CI 1.05 - 1.26; *P*=.18). Subgroup analyses indicated a statistically significant effect among patients with low health literacy (adjusted relative difference=1.81, 95% CI 1.42 - 2.32; *P*=.002). Anxiety scores were 9% lower (adjusted relative difference=0.91, 95% CI 0.84 - 0.99; *P*=.23), and satisfaction with previsit information increased by 14% (adjusted relative difference=1.14, 95% CI 1.07 - 1.21; *P*=.003).

Conclusions: Informational podcasts, provided as a supplement to traditional written information, may offer modest support for patient engagement in SDM, particularly among patients with low health literacy. Podcasts also appear to improve satisfaction with previsit information more broadly. Effects on previsit anxiety were inconclusive.

(J Med Internet Res 2026;28:e81485) doi:10.2196/81485

KEYWORDS

digital health; podcasts; patient education as topic; shared decision making; health literacy; patient engagement; anxiety; patient satisfaction; eHealth communication

Introduction

Effective communication and information exchange between patients and health care providers is essential for optimal treatment outcomes. It plays an important role in shared decision-making (SDM), patient satisfaction, and reduces anxiety related to health consultations [1-4].

Podcasts provide a platform for delivering patient information and supporting preparation for hospital visits [5]. Qualitative research suggests that podcasts can alleviate patients' anxiety related to hospital visits, foster a sense of trust in health care professionals, and enhance their understanding of empowerment [6]. Despite the rising popularity of podcasts in general society, there is a lack of research investigating their impact on SDM, anxiety, and patient satisfaction with hospital visits [5,6].

SDM is a collaborative process that fosters patient engagement and empowerment by actively involving patients in decisions about their care [7]. This approach, which prioritizes patient

XSL•FO
RenderX

preferences, has been shown to improve treatment adherence, satisfaction, quality of life, and health outcomes by integrating patients' perspectives into the decision-making process [8-12]. SDM is particularly relevant in clinical contexts characterized by diagnostic uncertainty and complex treatment considerations. In patients referred with suspected lung cancer, consultations often involve discussions of further diagnostic testing and potential malignancy, which may be associated with substantial psychological distress [13]. Patients with chronic obstructive pulmonary disease (COPD) frequently face decisions related to symptom management and inhaler choice; SDM has been shown to facilitate inhaler choice and support adherence and satisfaction in patients with newly-diagnosed COPD [14]. SDM interventions in chronic respiratory conditions have also been associated with improved medication adherence [15]. Similarly, patients referred for evaluation of sleep apnea are introduced to diagnostic procedures and possible treatment options, such as continuous positive airway pressure, which require long-term engagement and adherence [16]. In these contexts—often representing patients' first hospital visit for assessment and possible initiation of treatment—adequate preparation may be essential for meaningful participation in SDM. Beyond its clinical implications, SDM is closely linked to patients' psychological responses to hospital visits and their evaluation of care. Meaningful participation in decision-making presupposes that patients feel adequately informed and prepared. When patients understand the purpose of the visit and the potential diagnostic or treatment decisions to be addressed, they are more likely to engage actively in SDM. Such preparedness may simultaneously reduce uncertainty-related anxiety and contribute to greater satisfaction with the hospital visit. Conversely, insufficient preparation may limit participation, increase psychological distress, and negatively influence satisfaction [17]. Engagement in SDM, anxiety, and patient satisfaction can therefore be understood as interrelated dimensions of the hospital visit experience rather than independent outcomes. With the increasingly widespread use of smartphones and internet access, podcasts can be used to deliver information to a broad audience, enabling patients to access information and prepare for the hospital visit at their convenience. Increasingly, podcasts are being used as a supplement to traditional written patient information to support patient preparation for hospital visits, particularly before an initial consultation focused on diagnostic clarification and potential treatment planning [5,6]. Qualitative studies suggest that podcasts can foster a sense of trust, reduce anxiety, and enhance patients' confidence and engagement during hospital visits [6]. Despite their potential benefits, the impact of podcasts on SDM remains underexplored, as they are relatively new in the context of patient information [5,6].

Informed patients are less susceptible to anxiety and stress surrounding hospital visits and medical procedures [2]. When faced with health-related concerns, patients often seek information as a means of alleviating their anxiety and gaining a better understanding of their medical conditions and treatment options [18]. Written patient information—such as leaflets—can be effective in reducing anxiety, particularly when well-designed and appropriately used [19-22]. However, there are significant challenges related to traditional patient information [23]. In addition, patients with limited literacy, cognitive impairments, or disabilities may face barriers to accessing and comprehending written materials, further contributing to health inequalities [24-27]. Medical jargon and a lack of engagement can also limit the effectiveness of written communication [28-30]. This highlights the need to explore alternative formats—such as podcasts—that, through the use of voice, may provide more accessible, relatable, and emotionally supportive ways of informing and preparing patients [6].

Although the literature on the efficacy of podcasts as patient information is growing, more research is needed on their effect in the context of hospital visits. The objective of this study was to address this gap by investigating the effect of informational podcasts on patients' (1) engagement in SDM, (2) anxiety following the hospital visit, and (3) satisfaction with the hospital visit.

Methods

Design

This study used a quasi-experimental design with a nonrandomized comparison group to evaluate the effect of podcasts on patients' engagement in SDM, anxiety related to the hospital visit, and overall satisfaction.

In this natural experimental setup, patients recruited before the implementation of podcasts served as the comparison group, while those enrolled afterwards constituted the intervention group. Randomization was not feasible in the clinical context. As highlighted by the Medical Research Council [31], natural experiments can provide valid evidence of intervention effects when randomization is impractical—provided that contextual factors, assumptions, and potential biases are systematically addressed.

Settings and Participants

The study was conducted in 3 specialized outpatient clinics at a Danish hospital—lung cancer diagnostics, COPD, and sleep apnea clinics. Patient recruitment for the comparison group occurred from 2021 to 2022, before the podcasts were implemented. The podcasts were implemented at different times in each clinic (April 2022 [lung cancer], January 2023 [sleep apnea], and March 2023 [COPD]). Patients in the intervention group were enrolled roughly 3 months after podcast implementation to ensure the podcasts were integrated into standard previsit information.

Patients met the following inclusion criteria:

- Newly referred for suspected lung cancer, COPD, or sleep apnea
- Aged 18 years or older
- Able to speak and understand Danish

Patients with severe mental illness, cognitive impairments, or terminal conditions were excluded to ensure meaningful participation and reduce confounding from communication or health status challenges. Both comparison and intervention groups consisted of 40 patients per clinic. Recruitment took place on selected days, aligned with the clinic's capacity and available resources.

XSL•FO
RenderX

Intervention

We implemented 3 user-centered podcasts [5], each tailored to assist either patients with suspected lung cancer [32], COPD [33], or sleep apnea [34]. These podcasts provide information about why patients were referred to the clinic, what to expect during their visit, and how to prepare. An overview of the shared structure and core components of the podcasts is presented in Table 1. The development of these podcasts was guided by the "empathy map" method [5], which helped identify patients' concerns and informational needs before their hospital visits. This method ensured the podcasts were tailored to each patient group's unique needs [35-37]. Adequate preparation before hospital visits is theorized to simultaneously facilitate engagement in SDM, reduce uncertainty-related anxiety, and enhance overall satisfaction with the hospital visit. Patients who feel informed and confident in their understanding of the hospital visit are more likely to participate meaningfully in decision-making and evaluate the experience positively, whereas insufficient preparation may limit engagement, increase psychological distress, and reduce satisfaction [17]. The podcasts were therefore intended not only to provide information but to influence these interrelated outcomes.

Table . Overview of condition-specific podcasts[a].

Section	Content	Purpose
Target audience and considerations	Patients referred to outpatient clinics (lung cancer, COPD[b], or sleep apnea). Common patient concerns include anxiety, uncertainty, and a need for clear guidance.	Tailor content to patient needs; acknowledge emotions; build trust.
Introduction	Brief host introduction and explanation of podcast purpose.	Orient patients; set expectations.
Patient feelings	Acknowledge common emotional responses (eg, nervousness, worry, and frustration).	Reduce anxiety; show empathy.
Core topics	Reasons for referral, what to expect during the hospital visit, stepwise outline of the visit or diagnostic process, practical advice while waiting for results, and summary and key take-home messages	Provide structured, relevant information; improve comprehension and readiness for hospital visit.
Practical advice	Encourage patients to ask questions, prepare notes, talk with relatives, and maintain daily routines.	Support coping; reduce anxiety; empower patients.
Summary and reinforcement	Recap main points covered in the podcast.	Reinforce key messages; support retention and engagement.

[a]Podcasts were developed using the empathy map method [35-37] to ensure content aligns with patient concerns and informational needs. Each podcast follows the same general structure while addressing condition-specific aspects relevant to the respective patient group.

[b]COPD: chronic obstructive pulmonary disease.

During the study period, the clinics experienced no significant organizational changes, ensuring a stable environment for assessing the intervention and maintaining consistent care quality.

Patient Groups

For each condition (suspected lung cancer, COPD, and sleep apnea), 2 patient groups were included—a comparison group seen before the implementation of the podcasts, and an intervention group seen after. All patients received standard written information about their upcoming consultation via the national digital post system 2 to 30 days before their scheduled visit, depending on clinic procedures. In the intervention group, this letter also included a QR code and link to the condition-specific podcast, encouraging patients to listen before attending their appointment. Patients in the comparison group did not have access to the podcasts.

Data Collection

All patients received a questionnaire on the same day or the day after their consultation to assess their experiences and perceptions regarding the hospital visit. Patients had the option to complete the questionnaire in either written or electronic format. Those who selected the paper version received specific instructions based on their condition; patients with suspected lung cancer or sleep apnea were asked to return the completed questionnaire during their scheduled visit the next day, while patients with COPD were required to complete it immediately due to the infrequency of their hospital appointments. This ensured that all paper responses were collected. For patients opting for the electronic version, the questionnaire was sent the day after the consultation, with follow-up reminders issued after 2 and 4 weeks. No further action was taken if the questionnaire remained unreturned after this period. For patients who received the paper version, no additional reminders were sent beyond the initial instructions given at the hospital.

Outcome Measures

This study assessed key outcomes using validated instruments to evaluate the impact of podcasts on engagement in SDM, preconsultation anxiety levels, and overall satisfaction. The primary outcome was:

- SDM: The 9-item Shared Decision-Making Questionnaire (SDM-Q-9) consists of 9 statements rated on a 6-point scale

XSL•FO
RenderX

(0 - 5) from "completely disagree" (0) to "completely agree" [5,38]. Total scores range from 0 to 45, with higher scores indicating a greater perceived SDM. The SDM-Q-9 measures concrete behaviors and perceptions, such as whether patients felt informed, listened to, and involved in choices about their care. In this study, the validated Danish version of the SDM-Q-9 was used [39], which has been shown to retain the psychometric properties of the original instrument.

Secondary outcomes included:

- Anxiety: State anxiety was assessed using the State-Trait Anxiety Inventory-State (STAI-S) questionnaire, a validated 20-item self-report measure designed to evaluate transient anxiety levels in response to a specific situation [40]. The STAI-S measures transient, situation-specific feelings of anxiety, such as nervousness, worry, or tension, related to the hospital visit. Each item is rated on a 4-point Likert scale, ranging from 1 (not at all) to 4 (very much), with a total score ranging from 20 to 80. Higher scores indicate greater levels of state anxiety.
- Patient satisfaction was measured using four validated questions from the National Danish Survey of Patient Experiences (Landsdækkende Undersøgelse af Patientoplevelser [LUP]). The LUP questionnaire assesses patients' experiences with the Danish health care system. The questions addressed key aspects of the consultation process—satisfaction with preconsultation information, comfort during the consultation, adequate responses to their questions, and overall satisfaction. Participants responded using a 5-point Likert scale, ranging from "very dissatisfied" (1) to "very satisfied" (5) [41].

In addition, this study examined how various factors, including patient characteristics, provider attributes, and consultation characteristics, influenced the effect of podcasts. These factors were as follows:

- Sociodemographic information: Age, gender, social status, education level, and work status. Some response categories within variables (eg, education level and work status) were collapsed due to limited responses.
- Health literacy: Patients' health literacy was measured using the 16-item European Health Literacy Survey [42]. The questionnaire captures self-reported difficulties in accessing, understanding, appraising, and applying health-related information concerning health care, disease prevention, and health promotion [42-44]. Each item is rated on a 4-point Likert scale from "very easy" to "very difficult." Responses were dichotomized, with "very easy" and "easy" scored as 1, and "difficult" and "very difficult" scored as 0. The resulting total score ranged from 0 to 16 and was categorized into 3 levels: inadequate (0 - 8), problematic (9-12), and adequate (13-16).
- Relative presence: Recorded whether relatives were present during the consultation.
- Consultation duration: The duration of the consultation was documented.
- Podcast exposure: Self-reported listening to the condition-specific podcast was assessed in the follow-up questionnaire for patients in the intervention group.

Ethical Considerations

The Regional Committees on Health Research Ethics for Southern Denmark reviewed the study protocol and determined that formal ethical approval was not required (20202000 - 222). The study was registered with the Regional Danish Data Protection Agency (20/55212).

All patients gave informed consent before completing the questionnaires. Participation was voluntary, and confidentiality was ensured. Departmental staff provided patients with oral and written information about the study. Signed consent forms were collected, and patients were informed that they could withdraw without affecting their treatment or care.

Statistical Analyses

Descriptive statistics were used to summarize patient characteristics and outcomes. Between-group differences were summarized by standardized differences using Cohen *d*. The relative effect of the intervention with 95% CIs was further assessed for significance assuming independence between subjects using a log-normal linear model. The intervention effects were reported as crude and adjusted estimates. Adjusted estimates accounted for the following predefined potential confounders: sex, age, clinic, educational level, work status, health literacy, presence of relatives during the consultation, and consultation duration. All analyses were performed using STATA 18 (StataCorp LLC) with a significance level set at $P<.05$.

Sample Size Estimation

A power calculation based on the SDM-Q-9 indicated that 210 patients, evenly distributed between the intervention and comparison groups, would be needed to detect a clinically relevant difference in patient outcomes with a medium effect size (Cohen $d=0.50$), 95% power, and a 2-sided α of .05. To account for a 10% dropout rate, we aimed to recruit 232 patients.

Results

Overview

A total of 240 patients were included in the study, with 120 allocated to the comparison and intervention groups, respectively. An additional 89 patients initially consented but did not complete the questionnaires. Figure 1 illustrates the flow of patients through the study, including the number of patients who actually listened to the podcast in the intervention group.

XSL•FO
RenderX

Figure 1. Flow of patients through the study, including podcast exposure and questionnaire completion.

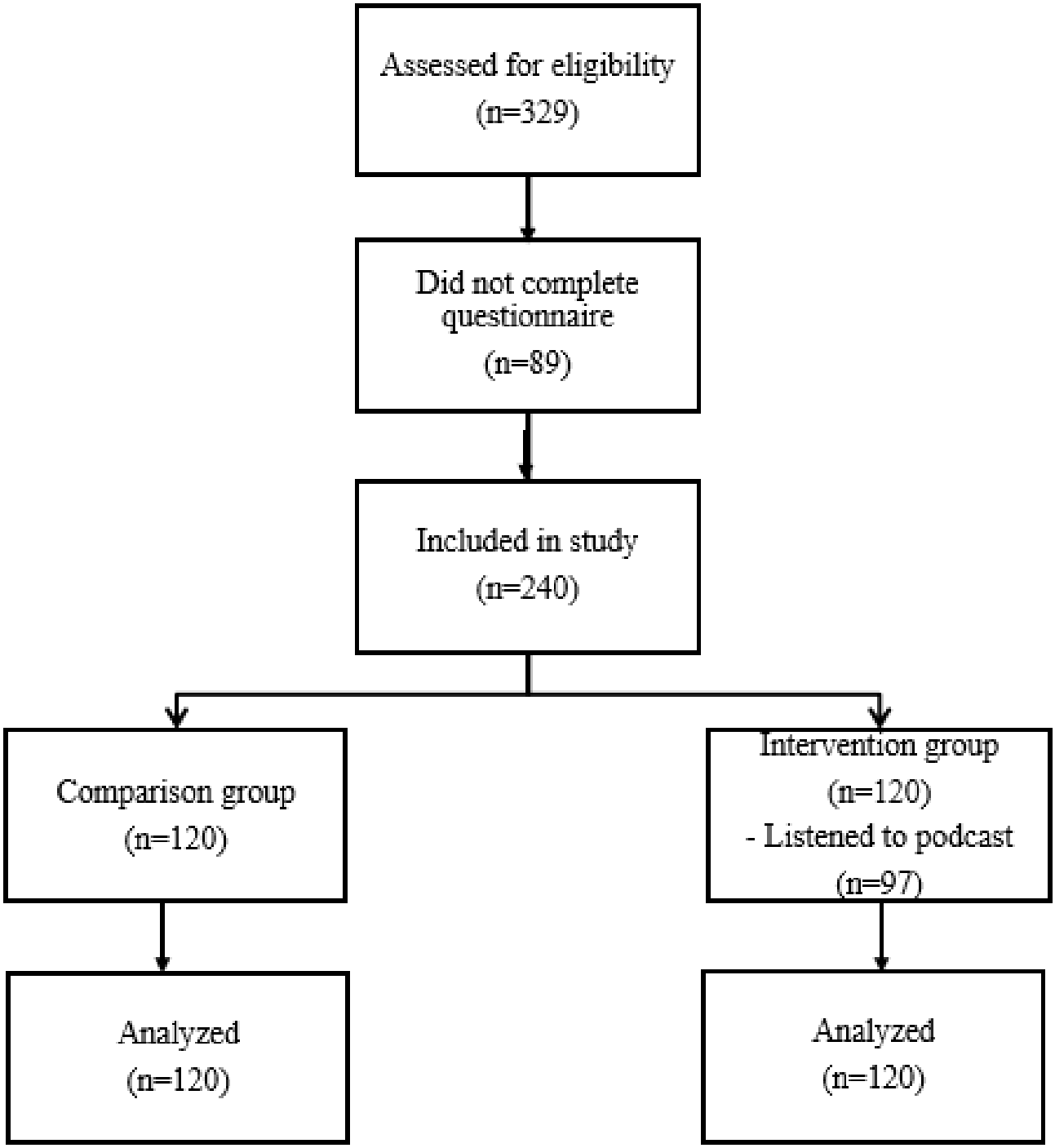

Overall, the groups were comparable across most demographic and clinical variables, including age, disease distribution, and employment status. Slight group differences were observed in education level and health literacy. The comparison group had a higher proportion of patients with a university education (63/120, 52% vs 49/120, 41%), whereas more patients in the intervention group reported only primary schooling (30/120, 25% vs 12/120, 10%). Health literacy scores were slightly higher in the intervention group (mean 13.0, SD 4.1, range 0-16 vs mean 12.5, SD 3.6, range 0-16), and a greater share of patients in this group fell into the “high literacy” category (91/120, 75% vs 73/120, 60%).

The consultation time was similar between groups (mean 34, SD 25.3, range 5-180 min), as was the proportion of patients who attended with a relative (54/120, 45% in the intervention group and 53/120, 44.2% in the comparison group). Patients in the intervention group were sent a QR code to access the podcast before their hospital visit, and 81% (n=97) of these patients reported having listened to the podcast. Patients in the comparison group did not have access to the podcast. Characteristics of the included patients can be found in Table 2.

XSL•FO
RenderX

Table . Patient characteristics (N=240).

Characteristics	All (N=240)	Comparison group (n=120)	Intervention group (n=120)
Sex (female), n (%)	103 (43)	59 (49)	44 (37)
Age, mean (SD; range)	61.7 (12.2; 26-99)	60.9 (12.9; 26-99)	62.4 (11.4; 38-86)
Disease, n (%)			
Suspected lung cancer	80 (33)	40 (33)	40 (33)
Suspected COPD[a] or COPD	80 (33)	40 (33)	40 (33)
Suspected sleep apnea	80 (33)	40 (33)	40 (33)
Relationship status (1 patient in the intervention group did not respond), n (%)			
Single	72 (30)	31 (26)	40 (33)
Married or cohabiting	137 (70)	89 (74)	79 (66)
Educational level, n (%)			
Low (primary school)	42 (18)	12 (10)	30 (25)
Medium (high school or vocational education and training)	86 (36)	45 (38)	41 (35)
High (university)	112 (46)	63 (52)	49 (41)
Work status (1 missing in comparison group), n (%)			
Employed	92 (38)	47 (39)	45 (38)
Unemployed	23 (10)	11 (9)	12 (10)
Retired	99 (41)	50 (42)	49 (41)
Disability pension	25 (10)	11 (9)	14 (12)
Health literacy score, mean (SD; range)	12.9 (3.9; 0 - 16)	12.5 (3.6; 0 - 16)	13.0 (4.1; 0 - 16)
Health literacy levels (HLS-EU-Q16[b]), n (%)			
Low	35 (14)	20 (17)	15 (13)
Medium	41 (17)	27 (23)	14 (12)
High	164 (68)	73 (60)	91 (75)
Relatives present at medical consultation, n (%)			
Yes	107 (45)	53 (44)	54 (44)
Duration of medical consultation, mean (SD; range)	34.1 (25.3; 5-180)	34.8 (26.6; 5-180)	33.4 (20.8; 5-150)
Have the patient heard the podcast			
Yes, n (%)	97 (40)	-	97 (81)
No, n (%)	143 (60)	120 (100)	23 (19)

[a]COPD: chronic obstructive pulmonary disease.

[b]HLS-EU-Q16: 16-item European Health Literacy Survey.

Objective 1: Patients' Engagement in SDM

Patients in the intervention group reported a mean SDM-Q-9 score of 31.5 (SD 0.98), compared with 27.4 (SD 1.02) in the comparison group. This difference was statistically significant in the unadjusted analysis but lost significance after adjusting for potential confounders (Table 3).

XSL•FO
RenderX

Table . Shared decision-making (SDM-Q-9[a]; score range 0 - 45), anxiety (STAI-S[b]; score range 20-80), and patient satisfaction scores (range 1-5) in intervention (n=120) and comparison (n=120) groups (N=240).

Outcomes	Comparison group, mean (SD)	Intervention group, mean (SD)	Relative change (95% CI)[c]	Standardized effect size, Cohen *d* (95% CI)[d]	Crude *P* value[e]	Adjusted *P* value[f]
Primary outcome: shared decision-making, range: 0 - 45						
Shared decision-making total score (SDM-Q-9)	27.4 (1.02)	31.5 (0.98)	1.15 (1.05 - 1.26)	0.38 (0.12 - 0.63)	.004	.18
Secondary outcome: anxiety, range: 20 - 80						
Total Anxiety Score (STAI-S)	37.01 (12.14)	33.85 (10.66)	0.91 (0.84 - 0.99)	0.28 (0.02 - 0.53)	.03	.23
Secondary outcome: satisfaction, range: 1 - 5						
Consultation comfort	4.14 (0.81)	4.38 (0.89)	1.06 (1.00 - 1.11)	0.28 (0.03 - 0.54)	.03	.17
Answers to questions	3.91 (1.04)	4.29 (0.88)	1.10 (1.04 - 1.17)	0.41 (0.15 - 0.66)	.002	.09
Previsit information	4.01 (1.15)	4.55 (0.88)	1.14 (1.07 - 1.21)	0.53 (0.27 - 0.79)	<.001	.003
Overall consultation satisfaction	3.91 (1.07)	4.33 (0.83)	1.11 (1.05 - 1.18)	0.44 (0.19 - 0.70)	<.001	.11

[a]SDM-Q-9: 9-item Shared Decision-Making Questionnaire.

[b]STAI-S: State-Trait Anxiety Inventory-State.

[c]The relative change (95% CI) indicates the proportional change between groups (comparison vs intervention) with a 95% CI.

[d]Standard effect size (Cohen *d*) is a standardized effect size that quantifies the difference between 2 groups in standard deviation units with a 95% CI.

[e]The crude *P* value represents the initial significance level without adjustments.

[f]The adjusted *P* value accounts for confounders (eg, gender, age, education, presence of accompanying relative, disease type, health literacy, and perception of health).

Subgroup analyses revealed consistently higher levels of SDM in the intervention group, compared with the comparison group. However, after adjustment, only patients with low health literacy in the intervention group showed statistically significant higher levels of SDM, relative to the comparison group. No significant differences were found across other demographic or clinical variables, including gender, age, education, employment status, relationship status, or the presence of a relative during the consultation (Table 4).

XSL•FO
RenderX

Table . Differences in shared decision-making scores (range 0 - 45) across demographic and disease-related factors in intervention (n=120) and comparison (n=120) groups (N=240).

Characteristic	Comparison group, mean (SD)	Intervention group, mean (SD)	Relative effect (95% CI)[a]	Adjusted *P* value for interaction[b]
Sex				
Females	26.51 (1.47)	32.34 (1.59)	1.22 (1.05 - 1.41)	__[c]
Males	28.25 (1.41)	31.01 (1.25)	1.10 (0.97 - 1.24)	.83
Age (y)				
<67	26.56 (1.30)	29.99 (1.38)	1.22 (0.99 - 1.29)	—
>67	28.78 (1.58)	33.62 (1.30)	1.17 (1.03 - 1.33)	.29
Education				
Low (primary school)	30.58 (2.82)	34.27 (1.23)	1.12 (0.95 - 1.32)	—
Medium (high school or vocational education and training)	25.93 (1.79)	30.85 (1.94)	1.19 (0.99 - 1.43)	.83
High (university)	28.83 (1.35)	30.35 (1.59)	1.09 (0.95 - 1.25)	.36
Work status				
Employed	28.32 (1.59)	33.45 (1.31)	1.10 (0.92 - 1.31)	—
Unemployed	28.36 (2.71)	31.00 (2.91)	1.09 (0.84 - 1.43)	.78
Retired	28.32 (1.59)	33.45 (1.31)	1.18 (1.03 - 1.35)	.43
Disability pension	25.72 (4.11)	32.86 (1.84)	1.28 (0.96 - 1.71)	.52
Relationship status				
Single	25.30 (1.93)	32.08 (1.36)	1.27 (1.08 - 1.49)	—
Married or cohabiting	28.11 (1.90)	31.04 (1.32	1.10 (0.98 - 1.24)	.28
Health literacy				
Low	17.55 (2.14)	31.80 (1.62)	1.81 (1.42 - 2.32)	—
Medium	25.63 (1.84)	30.29 (3.06)	1.18 (0.94 - 1.49)	.16
High	30.74 (1.21)	31.64 (1.18)	1.03 (0.92 - 1.15)	.60
Suspected disease				
Suspected COPD[d] or COPD	25.7 (1.62)	33.35 (0.82)	1.30 (1.15 - 1.47)	—
Suspected lung cancer	31.38 (1.77)	34.00 (1.59)	1.08 (0.94 - 1.25)	.11
Suspected sleep apnea	25.08 (1.75)	27.15 (2.21)	1.08 (0.88 - 1.34)	.06
Relatives present				
Yes	25.79 (1.35)	28.64 (1.46)	1.11 (0.96 - 1.28)	—
No	29.48 (1.56)	35.11 (1.06)	1.19 (1.06 - 1.33)	.41

[a]Relative effect (95% CI): The relative effect (intervention × comparison), with a 95% CI.

[b]Adjusted *P* value for interaction: The adjusted *P* value for the interaction between variables and the effect (intervention × comparison).

[c]Not applicable.

[d]COPD: chronic obstructive pulmonary disease.

Regarding preferences for decision-making roles, the majority of patients in both groups preferred SDM (40/120, 33% comparison vs 36/120, 30% intervention). Notably, more patients in the intervention group reported that the doctor decides alone (9/120, 8% vs 27/120, 23%, whereas fewer indicated that the doctor decides but considers the patient's opinion (39/120, 33% vs 31/120, 26%; Figure 2).

XSL•FO
RenderX

Figure 2. Preferences for involvement in shared decision-making based on the 9-item Shared Decision-Making Questionnaire in intervention (n=120) and comparison (n=120) groups (N=240).

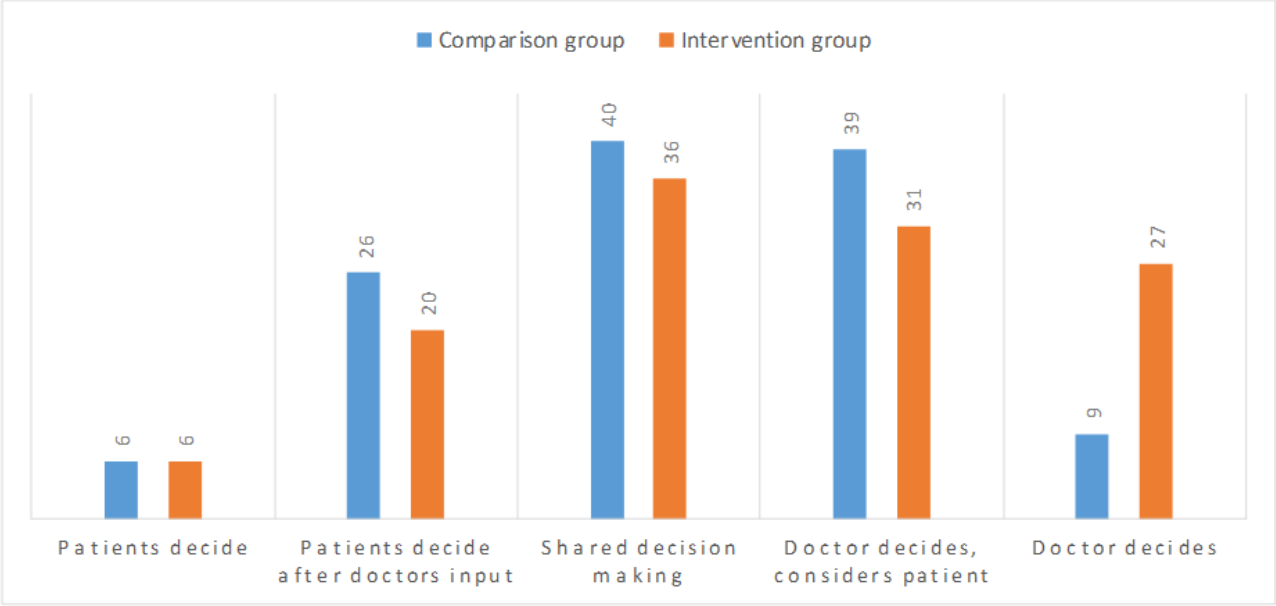

Objective 2: Patient Anxiety Associated With Hospital Visits

Patients in the intervention group reported lower anxiety levels than those in the comparison group (mean STAI-S 33.85, SD 10.66 vs 37.01, SD 12.14). However, this difference was no longer statistically significant after adjustment for potential confounders (Table 3).

Subgroup analyses showed lower anxiety scores in the intervention group across most demographic and clinical factors, with the largest differences seen in patients younger than 67 years, those with lower educational attainment, and those who were currently employed. However, none of the interaction effects reached statistical significance (Table 5). Similarly, no significant subgroup differences were observed for gender, relationship status, health literacy, suspected diagnosis (lung cancer, COPD, and sleep apnea), or presence of relatives during the consultation.

XSL•FO
RenderX

Table . Differences in anxiety levels (range 20 - 80) across demographic and disease-related factors in intervention (n=120) and comparison (n=120) groups (N=240).

Characteristics	Comparison group, mean (SD)	Intervention group, mean (SD)	Relative chance (95% CI)[a]	Adjusted *P* value for interaction[b]
Sex				
Females	38.95 (1.67)	37.50 (1.83)	0.96 (0.85 - 1.96)	__[c]
Males	35.14 (1.44)	31.73 (1.05)	0.90 (0.82 - 1.00)	.72
Age (y)				
<67	38.59 (1.44)	33.04 (1.21)	0.86 (0.77 - 0.95)	—
>67	34.37 (1.66)	34.97 (1.60)	1.02 (0.89 - 1.16)	.85
Education				
Low (primary school)	40.54 (3.05)	33.64 (1.49)	0.83 (0.71 - 0.97)	—
Medium (high school or Vocational Education and Training)	37.34 (2.01)	33.37 (1.67)	0.89 (0.77 - 1.04)	.18
High (university)	36.10 (1.44)	34.36 (1.72)	0.95 (0.84 - 1.08)	.06
Work status				
Employed	38.40 (1.76)	30.54 (1.29)	0.80 (0.70 - 0.90)	—
Unemployed	42.84 (3.89)	38.81 (3.83)	0.91 (0.70 - 1.18)	.73
Retired	34.24 (1.60)	35.30 (1.60)	1.03 (0.91 - 1.17)	.22
Disability pension	36.21 (3.93)	35.14 (2.71)	0.97 (0.75 - 1.25)	.38
Relationship status				
Single	38.42 (2.46)	34.24 (1.68)	0.89 (0.76 - 1.04)	—
Married or cohabiting	36.52 (1.23)	33.54 (1.21)	0.92 (0.83 - 1.01)	.18
Health literacy				
Low	43.81 (2.94)	37.56 (2.89)	0.86 (0.70 - 1.06)	—
Medium	38.88 (2.24)	33.00 (2.37)	0.85 (0.69 - 1.04)	.81
High	34.46 (1.33)	33.36 (1.13)	0.97 (0.88 - 1.07)	.73
Suspected disease				
Suspected COPD[d] or COPD	34.83 (1.64)	31.87 (1.32)	0.92 (0.81 - 1.04)	—
Suspected lung cancer	39.99 (2.11)	37.47 (1.88)	0.94 (0.81 - 1.08)	.96
Suspected sleep apnea	36.21 (1.93)	32.20 (1.69)	0.89 (0.77 - 1.03)	.33
Relatives present				
Yes	38.49 (1.66)	35.77 (1.52)	0.93 (0.83 - 1.05)	—
No	35.88 (1.66)	32.33 (1.24)	0.90 (0.81 - 1.01)	.89

[a]Relative effect (95% CI): The relative effect (intervention × comparison), with a 95% CI.

[b]Adjusted *P* value for interaction: The adjusted *P* value for the interaction between variables and the effect (intervention × comparison).

[c]Not applicable.

[d]COPD: chronic obstructive pulmonary disease.

Objective 3: Satisfaction With the Hospital Visit

The level of satisfaction with previsit information was significantly higher in the intervention group than in the comparison group (4.55 vs 4.01). This difference remained statistically significant after adjustment for potential confounders.

Differences between intervention and comparison groups in satisfaction levels across other domains did not remain statistically significant after adjustment (Table 3).

XSL•FO
RenderX

Discussion

Summary of Findings

To our knowledge, this is the first study to examine the impact of informational podcasts, as a supplement to standard written information, on SDM, consultation-related anxiety, and patient satisfaction among outpatients with suspected lung cancer, COPD, or sleep apnea.

The results suggest that podcasts appeared to provide modest support for patients' engagement in SDM, especially among the subgroup of patients with low health literacy. Moreover, the implementation of podcasts was associated with higher satisfaction levels with preconsultation information. Effects on previsit anxiety were inconclusive, likely reflecting the multifactorial nature of anxiety and variations in patient engagement with digital content.

Taken together, these findings suggest that informational podcasts may serve as a modest, supplementary tool to traditional written materials, particularly for patients with low health literacy, rather than as a broadly transformative intervention across all patient groups.

Objective 1: Patients' Engagement in SDM

The findings of this study suggest that integrating podcasts into routine patient information may enhance patients' engagement in SDM during outpatient consultations. Although the overall effect of the intervention was attenuated after adjustment for potential confounders, subgroup analyses revealed essential patterns. The most substantial improvements were observed among patients with low health literacy. These trends highlight the potential of podcasts to support patient populations who may otherwise face challenges engaging with written health information.

In Denmark, as in many other countries, a considerable proportion of the population experiences difficulties reading and comprehending health-related materials. National data indicate that up to 16% of adults struggle with basic reading and writing skills [24], which can undermine their ability to interpret written information, navigate health care systems, and participate meaningfully in SDM [43-45]. For such patients, audio-based formats like podcasts may provide a more accessible and effective alternative. Delivering complex patient information in an audio format that can be revisited at the patient's pace promotes understanding in populations underserved by traditional written formats. While written materials also allow patients to revisit information, podcasts offer unique advantages, such as the ability to listen flexibly at home or on the go, which may better support patients who find it difficult to navigate or retain written content. This may explain why patients with low health literacy appeared to benefit more, as the podcast served as a more inclusive communication format. Notably, all 3 patient groups included in this study—those with suspected or diagnosed lung cancer, COPD, and sleep apnea—share a heightened risk of low health literacy [46-48]. These findings support that podcasts are not merely substitutes for written materials but represent a distinct and valuable communication format, particularly for patient groups with limited health literacy. The flexibility to listen at home or on the go and to revisit the content as needed may be especially advantageous for individuals who find it difficult to process complex written information or retain key messages during stressful health care encounters.

Nevertheless, while digital tools like podcasts increase access to patient information, they may also introduce new challenges. Older patients and patients with limited technological literacy may experience difficulties navigating digital platforms or downloading podcasts [49-51]. However, a qualitative study we conducted alongside this study found that podcasts were usable across age groups, suggesting that age was not a limiting factor for engagement [6].

These results are consistent with previous research indicating that patient-centered educational interventions can improve SDM, particularly among patients with limited health literacy [52].

It is important to note that not all patients seek the same degree of involvement. The findings of this study showed that a considerable proportion of the included patients preferred to defer decision-making to health care professionals. Notably, a larger proportion of patients in the intervention group than in the comparison group preferred that the health care professional decided without including the patient. The design of our study does not allow us to draw conclusions regarding causality, but this finding highlights the need for a nuanced, flexible approach to SDM—one that both encourages patient involvement and respects individual preferences regarding the extent of participation. This aligns with previous research suggesting that preference-sensitive SDM is most effective when patients are well-informed, supported, and empowered to define their role in the decision-making process [53,54].

Objective 2: Patient Anxiety Associated With Hospital Visits

In this study, we did not find a clear effect of podcasts on patient anxiety. However, a previous qualitative interview study showed that many patients felt reassured and less anxious after listening to the podcasts. They explained that hearing health care professionals and fellow patients helped build trust and made them feel better prepared for their hospital visits [6].

There were indications that younger patients, those with lower educational attainment, and patients currently working might have experienced greater reductions in anxiety. These groups often face more stress due to competing responsibilities and may have lower health literacy, making them particularly vulnerable in health care settings. Although these trends were not statistically significant, they align with earlier research showing that patients from lower socioeconomic backgrounds tend to experience higher levels of hospital-related anxiety [55]. Podcasts may be particularly helpful for these groups by providing information in an accessible and easy-to-understand format, but further studies are needed to confirm this.

At the same time, anxiety around hospital visits is influenced by many factors beyond just information. Fear of diagnosis, past experiences, and the stressful nature of hospital visits contribute [56]. Information alone is unlikely to completely

XSL•FO
RenderX

remove anxiety. It is also important to acknowledge that some anxiety is natural and appropriate when facing potentially serious health issues [57]. Attempting to eliminate anxiety entirely may neither be realistic nor helpful.

Overall, these findings suggest that podcasts can support patient preparation and potentially alleviate some anxiety; however, they should be part of a broader approach that includes empathetic communication, emotional support, and continuity of care to address patients' emotional needs fully.

Objective 3: Satisfaction With the Hospital Visit

This study found that integrating podcasts into preconsultation preparation was associated with improved patient satisfaction, particularly in terms of satisfaction with the previsit information. This suggests that podcasts can meaningfully enhance patients' perception of the quality, clarity, and relevance of the information provided before hospital visits. These findings support previous research that demonstrates the positive impact of podcasts on patient satisfaction in various health care settings. In a qualitative study exploring patient experiences with podcasts, patients reported feeling more informed, prepared, and empowered during hospital visits [6].

Information provided before hospital visits shapes patients' expectations and readiness for health care discussions. Effective communication is central to patient-centered care, enabling individuals to understand their diagnoses and treatment options and to participate actively in decision-making processes [1,58]. When patients are inadequately prepared, they may experience confusion, unmet expectations, and dissatisfaction with their care experience [59,60]. The podcasts evaluated in this study likely improved satisfaction by presenting condition-specific information in a format that was both accessible and engaging, helping patients to feel more confident and informed before their appointments.

Podcasts may add value by enabling repeated listening, allowing for more personalized preparation, and offering a human and conversational tone that differs from traditional written materials [5,6].

Integration of SDM, Anxiety, and Patient Satisfaction

Engagement in SDM, patient anxiety, and satisfaction with the hospital visit are closely interrelated dimensions of the patient experience. Meaningful participation in SDM presupposes that patients feel adequately informed and prepared; when patients understand the purpose of the visit and the potential diagnostic or treatment decisions to be addressed, they are more likely to engage actively in decision-making, which may simultaneously reduce uncertainty-related anxiety and enhance satisfaction [17]. However, the effects of information are not uniform. For some patients, additional information—such as podcasts detailing diagnostic procedures or potential consequences of illness—may reduce anxiety by improving preparedness. For others, exposure to detailed information may temporarily increase state anxiety until coping strategies are established, which in turn can hinder engagement in SDM [61,62]. This dynamic is particularly relevant for patients referred for evaluation of suspected lung cancer, COPD, or sleep apnea, all of whom face complex clinical decisions requiring understanding and adherence to recommended procedures [14-17]. Informational podcasts, provided as a supplement to traditional written information, may modestly support engagement in SDM, improve understanding of pre-visit procedures, reduce uncertainty-related anxiety for some patients, and enhance satisfaction with the hospital visit. The effects observed at the group level were limited, highlighting that podcasts function primarily as a supplementary tool rather than a standalone intervention, and suggesting that their benefit may be most pronounced in specific patient subgroups, such as those with lower health literacy [14-17,61,62].

Implications for Practice

The findings of this study have important implications for clinical practice, particularly in the context of supporting SDM. While podcasts can be a valuable supplement to standard written patient information, they should not be viewed as a standalone solution. Instead, they represent one of several elements that can help prepare patients for more active involvement in health care decisions.

For successful implementation, it is essential to recognize that patients have diverse needs and varying levels of health literacy. Particularly for individuals with low health literacy, additional support may be necessary to ensure they comprehend the information and feel confident in making informed choices.

When used appropriately, podcasts can help patients feel more informed and better prepared for hospital visits. However, their effectiveness depends on integration with broader communication strategies and a continuous focus on personalized support. In this way, podcasts can make a meaningful contribution to a more patient-centered approach to care.

Strengths and Limitations

This study offers several notable strengths. First, its intervention design, including an intervention and comparison group, allows for a direct comparison of patient outcomes before and after the podcast implementation. This design strengthens the ability to attribute observed changes to the intervention.

The podcasts were developed using the "empathy map" methodology, ensuring the content was tailored to address the specific concerns and informational needs of patients with lung cancer, COPD, and sleep apnea. This user-centered approach enhances the relevance and potential effectiveness of the intervention.

The study implemented a robust evaluation framework, incorporating multiple validated outcome measures, including questionnaires that assessed SDM, patient satisfaction, and anxiety levels using the State-Trait Anxiety Inventory. This comprehensive evaluation provides a thorough understanding of the intervention's impact. Furthermore, the sample size of 240 patients, evenly distributed across three clinical conditions and both intervention and comparison groups, improves statistical power for the primary outcome and the generalizability of the findings. This diverse sample provides valuable insights into how podcasts can be adapted to meet the needs of different patient groups.

XSL•FO
RenderX

Despite its strengths, the study has several limitations. Although the controlled design allows for comparison between groups, other factors, such as variation in disease severity and differences in provider interactions, may still have influenced outcomes. The study was conducted at a single site in Denmark, which may limit generalizability to other health care settings or populations with different cultural contexts and health care systems. Furthermore, a relatively high number of patients initially consented but did not complete the questionnaires, raising concerns about barriers to participation and potential response bias. Investigating reasons for noncompletion would be necessary for improving retention in future studies.

The findings reflect short-term outcomes measured 1 day after hospital visits, which may not capture the longer-term effects of podcast interventions on outcomes. Additionally, the reliance on self-reported data introduces the risk of response bias. As this study followed a pragmatic pre-post implementation design, analyses were conducted according to group allocation rather than confirmed exposure to the podcast, consistent with an intention-to-treat approach. Engagement with the podcast was assessed through self-report, and not all patients in the intervention group may have listened to the content. This incomplete exposure may have attenuated the observed effects and could partly explain the modest effect sizes observed. While the exclusion of patients with severe mental illness, cognitive disabilities, or terminal illness was necessary to ensure internal validity, it limits the applicability of results to more vulnerable patient groups.

Finally, variations in how patients engage with and perceive podcast content may have influenced the intervention's overall effectiveness, suggesting that personal preferences and digital literacy could contribute to patient benefit.

Future Research

Future research should focus on randomized controlled trials to confirm the effectiveness of podcasts in enhancing SDM, reducing anxiety, and improving patient satisfaction. Studies with longer follow-up periods are also needed to assess the long-term impact of podcasts over time. Comparative effectiveness studies are also needed to evaluate how podcasts perform relative to other formats of patient information, such as written leaflets or videos. These studies will be essential for establishing a robust evidence base to guide the integration of podcasts as a supplement to traditional written patient information.

Conclusion

This quasi-experimental study suggests that informational podcasts, when used as a supplement to traditional written information, may offer modest support for patient engagement in SDM and improve satisfaction with previsit information, particularly among patients with low health literacy. Effects on previsit anxiety were inconclusive, likely reflecting the multifactorial nature of anxiety and variations in patients' engagement with digital content. Overall, the impact of the podcasts was modest, indicating that they function primarily as a supplementary tool to traditional written information, rather than as a standalone intervention, and underscoring the importance of identifying patient subgroups who may derive the greatest benefit.

Acknowledgments

The authors wish to thank the patients participating in the study. We also thank Eva Hoffmann (RN), Lisbeth Skousen Svendsen (RN), and Jette Tækker (RN), from the Department of Medicine, Vejle Hospital, for their help with patient recruitment. We further acknowledge OPEN, Open Patient Data Explorative Network at Odense University Hospital, Region of Southern Denmark, for their valuable support. During the preparation of this work, the authors used ChatGPT to assist in shortening and refining certain sentences for clarity and conciseness. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

Funding

This research was supported by the Development and Innovation Fund at Lillebaelt Hospital, Denmark, and the Novo Nordisk Foundation (Grant NNF21OC0068374).

Authors' Contributions

JCF: Conceptualization, Methodology, Formal analysis, Results and statistics, Writing – original draft, Writing – review and editing, Funding acquisition.
AL: Conceptualization, Methodology, Writing – review and editing, Funding acquisition.
HIJ: Conceptualization, Methodology, Discussion, Writing – review and editing.
FS: Statistical guidance, Writing – review and editing.
IFV: Conceptualization, Methodology, Formal analysis, Discussion, Writing – review and editing, Funding acquisition.

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. Ammentorp J, Sabroe S, Kofoed PE, Mainz J. The effect of training in communication skills on medical doctors' and nurses' self-efficacy. A randomized controlled trial. Patient Educ Couns 2007 Jun;66(3):270-277. [doi: 10.1016/j.pec.2006.12.012] [Medline: 17337337]
2. Nguyen MH, Smets EMA, Bol N, et al. Fear and forget: how anxiety impacts information recall in newly diagnosed cancer patients visiting a fast-track clinic. Acta Oncol 2019 Feb;58(2):182-188. [doi: 10.1080/0284186X.2018.1512156] [Medline: 30264632]
3. Vahdat S, Hamzehgardeshi L, Hessam S, Hamzehgardeshi Z. Patient involvement in health care decision making: a review. Iran Red Crescent Med J 2014 Jan;16(1):e12454. [doi: 10.5812/ircmj.12454] [Medline: 24719703]
4. Alkhatib AJ, AlZoubi A, Alharoun M, Alawnah AB, Abu Aqoulah A, Mandili A. The impact of patient engagement on healthcare: a systematic review of unveiling the benefits of patient engagement in healthcare. Multicriteria algo appl 2024 Mar 13;3:23-31. [doi: 10.61356/j.mawa.2024.311661]
5. Frølund JC, Løkke A, Jensen HI, Farver-Vestergaard I. Development of podcasts in a hospital setting: a user-centered approach. J Health Commun 2024 Apr 2;29(4):244-255. [doi: 10.1080/10810730.2024.2321385] [Medline: 38420940]
6. Frølund JC, Løkke A, Jensen HI, Farver-Vestergaard I. The use of podcasts as patient preparation for hospital visits-an interview study exploring patients' experiences. Int J Environ Res Public Health 2024 Jun 6;21(6):746. [doi: 10.3390/ijerph21060746] [Medline: 38928992]
7. Ferro-López L, Barnett N, Minshull J. Improving shared decision-making in pharmacist-led haematology clinics: a "Plan Do Study Act" approach. Eur J Hosp Pharm 2021 Nov;28(Suppl 2):e180-e184. [doi: 10.1136/ejhpharm-2021-002726] [Medline: 34039687]
8. Navarro M. Patients' empowerment and the role of patients' education. MRAJ 2020;8(12). [doi: 10.18103/mra.v8i12.2306]
9. Ozdemir S, Teo I, Bundoc FG, et al. Role in decision making among congestive heart failure patients and its association with patient outcomes: a baseline analysis of the SCOPAH study. Patient Educ Couns 2021 Mar;104(3):496-504. [doi: 10.1016/j.pec.2020.08.033] [Medline: 32928597]
10. Street RL, Gordon HS, Ward MM, Krupat E, Kravitz RL. Patient participation in medical consultations. Med Care 2005;43(10):960-969. [doi: 10.1097/01.mlr.0000178172.40344.70]
11. Kumar R, Korthuis PT, Saha S, et al. Decision-making role preferences among patients with HIV: associations with patient and provider characteristics and communication behaviors. J Gen Intern Med 2010 Jun;25(6):517-523. [doi: 10.1007/s11606-010-1275-3] [Medline: 20180157]
12. Edwardsen E, Morse D, Gurnsey MG, et al. Physician dialogues with advanced cancer patients: mostly patient-friendly but often physician-directed. EJPCH 2013;1(1):138. [doi: 10.5750/ejpch.v1i1.646]
13. Rauff Søndergaard S, Ellekjær LB, Bechmann T, et al. Shared decision making with breast cancer patients: impact on patient engagement and fear of recurrence. Protocol for a Danish randomized trial in radiotherapy (DBCG RT SDM). Acta Oncol 2021 Aug;60(8):1032-1037. [doi: 10.1080/0284186X.2021.1921261] [Medline: 34010102]
14. Hsiao YH, Tseng CM, Sheu CC, et al. Shared decision-making facilitates inhaler choice in patients with newly-diagnosed chronic obstructive pulmonary disease: a multicenter prospective study. Int J Chron Obstruct Pulmon Dis 2022;17:2067-2078. [doi: 10.2147/COPD.S376547] [Medline: 36081765]
15. Achterbosch M, Hansen CME, Brand PLP, van Dijk L, van Boven JFM. Shared decision making and inhaled medication adherence in patients with COPD, asthma and cystic fibrosis: a systematic review. Respir Med 2025 Oct;247:108251. [doi: 10.1016/j.rmed.2025.108251] [Medline: 40653279]
16. Øverby CT, Sutharshan P, Gulbrandsen P, Dammen T, Hrubos-Strøm H. Shared decision making: a novel approach to personalized treatment in obstructive sleep apnea. Sleep Med X 2022 Dec;4:100052. [doi: 10.1016/j.sleepx.2022.100052] [Medline: 36039181]
17. Joseph-Williams N, Elwyn G, Edwards A. Knowledge is not power for patients: a systematic review and thematic synthesis of patient-reported barriers and facilitators to shared decision making. Patient Educ Couns 2014 Mar;94(3):291-309. [doi: 10.1016/j.pec.2013.10.031] [Medline: 24305642]
18. Muse K, McManus F, Leung C, Meghreblian B, Williams JMG. Cyberchondriasis: fact or fiction? A preliminary examination of the relationship between health anxiety and searching for health information on the internet. J Anxiety Disord 2012 Jan;26(1):189-196. [doi: 10.1016/j.janxdis.2011.11.005] [Medline: 22137465]
19. Kharod U, Panchal NN, Varma J, Sutaria K. Effect of pre-operative communication using anaesthesia information sheet on pre-operative anxiety of patients undergoing elective surgery-a randomised controlled study. Indian J Anaesth 2022 Aug;66(8):559-572. [doi: 10.4103/ija.ija_32_22] [Medline: 36274808]
20. Plunkett S, Dobeli K, Prior M, Tusek X. Survey of maternal anxiety and perceptions towards foetal MRI and pre-scan education. J Med Radiat Sci 2024 Mar;71(1):10-20. [doi: 10.1002/jmrs.725] [Medline: 37724764]
21. Pinar G, Kurt A, Gungor T. The efficacy of preopoerative instruction in reducing anxiety following gyneoncological surgery: a case control study. World J Surg Oncol 2011 Apr 8;9(1):38. [doi: 10.1186/1477-7819-9-38] [Medline: 21477315]
22. Kutlutürkan S, Görgülü U, Fesci H, Karavelioglu A. The effects of providing pre-gastrointestinal endoscopy written educational material on patients' anxiety: a randomised controlled trial. Int J Nurs Stud 2010 Sep;47(9):1066-1073. [doi: 10.1016/j.ijnurstu.2010.01.007] [Medline: 20181334]

XSL•FO
RenderX

23. Nikpour S, Mehrdad N, Sanjari M, et al. Challenges of type 2 diabetes mellitus management from the perspective of patients: conventional content analysis. Interact J Med Res 2022 Oct 27;11(2):e41933. [doi: 10.2196/41933] [Medline: 36301605]
24. Kjeldsen MMZ, Stapelfeldt CM, Lindholdt L, Lund T, Labriola M. Reading and writing difficulties and self-rated health among Danish adolescents: cross-sectional study from the FOCA cohort. BMC Public Health 2019 May 10;19(1):537. [doi: 10.1186/s12889-019-6931-x] [Medline: 31077173]
25. Conti-Ramsden G, Durkin K, Toseeb U, Botting N, Pickles A. Education and employment outcomes of young adults with a history of developmental language disorder. Int J Lang Commun Disord 2018 Mar;53(2):237-255. [doi: 10.1111/1460-6984.12338] [Medline: 29139196]
26. Boberg EW, Gustafson DH, Hawkins RP, et al. Assessing the unmet information, support and care delivery needs of men with prostate cancer. Patient Educ Couns 2003 Mar;49(3):233-242. [doi: 10.1016/s0738-3991(02)00183-0] [Medline: 12642195]
27. Jenkins V, Fallowfield L, Saul J. Information needs of patients with cancer: results from a large study in UK cancer centres. Br J Cancer 2001 Jan 5;84(1):48-51. [doi: 10.1054/bjoc.2000.1573] [Medline: 11139312]
28. Hack TF, Degner LF, Parker PA, SCRN Communication Team. The communication goals and needs of cancer patients: a review. Psychooncology 2005 Oct;14(10):831-845. [doi: 10.1002/pon.949] [Medline: 16200519]
29. Polikandrioti M. Atrial fibrillation: the impact of anxiety and depression on patients' needs. Psychiatriki 2021 Sep 20;32(3):187-198. [doi: 10.22365/jpsych.2021.013] [Medline: 33770750]
30. Othman MI, Mohd Najib MN, Sulaiman S. Drug dispensing and counselling: challenges and practices in patient-pharmacist communication. IJARPED 2023 Apr 19;12(2). [doi: 10.6007/IJARPED/v12-i2/16802]
31. Craig P, Cooper C, Gunnell D, et al. Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. J Epidemiol Community Health 2012 Dec;66(12):1182-1186. [doi: 10.1136/jech-2011-200375] [Medline: 22577181]
32. For dig der skal udredes i lungepakken. in English: for those of you who need to be evaluated in the lung cancer diagnostic workup. Sygehus Lillebælt - Podcasts. 2021. URL: https://3f1a06bf-0d8f-4ecc-9335-97bba834e93f.libsyn.com/for-dig-der-skal-udredes-i-lungepakken [accessed 2026-03-10]
33. Lungeklinikken: velkommen til. in English: the COPD clinic: welcome. Sygehus Lillebælt - Podcasts. URL: https://3f1a06bf-0d8f-4ecc-9335-97bba834e93f.libsyn.com/velkommen-til-lungeklinikken [accessed 2026-03-10]
34. For dig der skal udredes for søvnapnø. in English: for those of you who need to be evaluated for sleep apnea. Sygehus Lillebælt - Podcasts. URL: https://3f1a06bf-0d8f-4ecc-9335-97bba834e93f.libsyn.com/size/5/?search=For+dig+der+skal+udredes+for+s%C3%B8vnapn%C3%B8 [accessed 2026-03-10]
35. Gibbons S. Empathy mapping: the first step in design thinking. Nielsen Norman Group. 2018. URL: https://www.nngroup.com/articles/empathy-mapping/ [accessed 2026-03-10]
36. Ferreira B, Silva W, Oliveira E, Conte T. Designing personas with empathy map. : Universidade Federal do Amazonas (UFAM); 2015 Jul 1 Presented at: The 27th International Conference on Software Engineering and Knowledge Engineering; May 2015 p. 501-505 URL: https://scispace.com/pdf/designing-personas-with-empathy-map-5cczwgxasn.pdf [accessed 2026-04-20] [doi: 10.18293/SEKE2015-152]
37. Pigneur Y, Clark T, Osterwalder A. Business Model Generation - A Handbook for Visionaries, Game Changers, and Challengers: John Wiley And Sons Ltd; 2010. URL: https://mymyeo.com/wp-content/uploads/2022/12/Business-model-generation-_-a-handbook-for-visionaries-game-changers-and-challengers-PDFDrive-1.pdf
38. Kriston L, Scholl I, Hölzel L, Simon D, Loh A, Härter M. The 9-item Shared Decision Making Questionnaire (SDM-Q-9). Development and psychometric properties in a primary care sample. Patient Educ Couns 2010 Jul;80(1):94-99. [doi: 10.1016/j.pec.2009.09.034] [Medline: 19879711]
39. Hulbæk M, Jørgensen MJ, Mainz H, et al. Danish Translation, cultural adaptation and validation of the Shared Decision Making Questionnaire - Patient version (SDM-Q-9-Pat). EJPCH 2018;6(3):438 [FREE Full text] [doi: 10.5750/ejpch.v6i3.1529]
40. Skapinakis P. Spielberger State-Trait Anxiety Inventory. In: Encyclopedia of Quality of Life and Well-Being Research: Springer; 2014:6261-6264. [doi: 10.1007/978-3-031-17299-1_2825]
41. Unit of Patient-perceived Quality on behalf of the Danish Regions and the Danish Ministry of Interior and Health. The National Danish Survey of Patient Experiences (LUP) – in Brief: Questionnaire Survey among Inpatients and Outpatients in Danish Hospitals: The Capital Region of Denmark; 2010. URL: https://www.regionh.dk/patientinddragelse/udgivelser/udgivelser/Documents/The%20National%20Danish%20Survey%20of%20Patient%20Experiences%20-%20in%20brief.pdf [accessed 2026-03-10]
42. Sørensen K, Van den Broucke S, Pelikan JM, et al. Measuring health literacy in populations: illuminating the design and development process of the European Health Literacy Survey Questionnaire (HLS-EU-Q). BMC Public Health 2013 Oct 10;13:948. [doi: 10.1186/1471-2458-13-948] [Medline: 24112855]
43. Friis K, Pedersen MH, Aaby A, Lasgaard M, Maindal HT. Impact of low health literacy on healthcare utilization in individuals with cardiovascular disease, chronic obstructive pulmonary disease, diabetes and mental disorders. A Danish population-based 4-year follow-up study. Eur J Public Health 2020 Oct 1;30(5):866-872. [doi: 10.1093/eurpub/ckaa064] [Medline: 32335677]

XSL•FO
RenderX

44. Murugesu L, Heijmans M, Rademakers J, Fransen MP. Challenges and solutions in communication with patients with low health literacy: perspectives of healthcare providers. PLoS ONE 2022;17(5):e0267782. [doi: 10.1371/journal.pone.0267782] [Medline: 35507632]
45. Kinnersley P, Edwards A, Hood K, et al. Interventions before consultations for helping patients address their information needs. Cochrane Database Syst Rev 2007 Jul 18;2007(3):CD004565. [doi: 10.1002/14651858.CD004565.pub2] [Medline: 17636767]
46. Ellender CM, Samaranayake C, Reid N, et al. Randomized controlled trial on the efficacy of audio-visual health educational materials on CPAP adherence: the AHEAD trial. J Clin Sleep Med 2022 Nov 1;18(11):2617-2625. [doi: 10.5664/jcsm.10182] [Medline: 35924663]
47. Etindele Sosso FA, Matos E. Socioeconomic disparities in obstructive sleep apnea: a systematic review of empirical research. Sleep Breath 2021 Dec;25(4):1729-1739. [doi: 10.1007/s11325-020-02274-z] [Medline: 33452999]
48. Haas K, Brillante C, Sharp L, et al. Lung cancer screening: assessment of health literacy and readability of online educational resources. BMC Public Health 2018 Dec 7;18(1):1356. [doi: 10.1186/s12889-018-6278-8] [Medline: 30526544]
49. Raja M, Uhrenfeldt L, Galvin KT, Kymre IG. Older adults` sense of dignity in digitally led healthcare. Nurs Ethics 2022 Sep;29(6):1518-1529. [doi: 10.1177/09697330221095140] [Medline: 35724431]
50. Schroeder T, Dodds L, Georgiou A, Gewald H, Siette J. Older adults and new technology: mapping review of the factors associated with older adults' intention to adopt digital technologies. JMIR Aging 2023 May 16;6:e44564. [doi: 10.2196/44564] [Medline: 37191976]
51. Ma Q, Chan AHS, Teh PL. Insights into older adults' technology acceptance through meta-analysis. International Journal of Human–Computer Interaction 2021 Jul 3;37(11):1049-1062. [doi: 10.1080/10447318.2020.1865005]
52. Barton JL, Koenig CJ, Evans-Young G, et al. The design of a low literacy decision aid about rheumatoid arthritis medications developed in three languages for use during the clinical encounter. BMC Med Inform Decis Mak 2014 Nov 25;14(1):104. [doi: 10.1186/s12911-014-0104-8] [Medline: 25649726]
53. Elwyn G, Frosch D, Thomson R, et al. Shared decision making: a model for clinical practice. J Gen Intern Med 2012 Oct;27(10):1361-1367. [doi: 10.1007/s11606-012-2077-6] [Medline: 22618581]
54. Elwyn G, Frosch DL, Kobrin S. Implementing shared decision-making: consider all the consequences. Implementation Sci 2015 Dec;11(1):114. [doi: 10.1186/s13012-016-0480-9]
55. Shah I, Akhunzada W, Riaz Q, Wazir A. Frequency of generalized anxiety disorder in hospitalized COVID-19 patient. Biol Clin Sci Res J 2024 Oct 31;2024(1):1234. [doi: 10.54112/bcsrj.v2024i1.1234]
56. Street RL, Makoul G, Arora NK, Epstein RM. How does communication heal? Pathways linking clinician-patient communication to health outcomes. Patient Educ Couns 2009 Mar;74(3):295-301. [doi: 10.1016/j.pec.2008.11.015] [Medline: 19150199]
57. House A, Stark D. Anxiety in medical patients. BMJ 2002 Jul 27;325(7357):207-209. [doi: 10.1136/bmj.325.7357.207] [Medline: 12142312]
58. Kwame A, Petrucka PM. A literature-based study of patient-centered care and communication in nurse-patient interactions: barriers, facilitators, and the way forward. BMC Nurs 2021 Sep 3;20(1):158. [doi: 10.1186/s12912-021-00684-2] [Medline: 34479560]
59. Nguyen MH, Smets EM, Bol N, et al. Tailored web-based information for younger and older patients with cancer: randomized controlled trial of a preparatory educational intervention on patient outcomes. J Med Internet Res 2019 Oct 1;21(10):e14407. [doi: 10.2196/14407] [Medline: 31573911]
60. Gaston CM, Mitchell G. Information giving and decision-making in patients with advanced cancer: a systematic review. Soc Sci Med 2005 Nov;61(10):2252-2264. [doi: 10.1016/j.socscimed.2005.04.015] [Medline: 15922501]
61. Köther AK, Büdenbender B, Grüne B, et al. Different patients, different preferences: a multicenter assessment of patients' personality traits and anxiety in shared decision making. Cancer Med 2022 Aug;11(15):2999-3008. [doi: 10.1002/cam4.4667] [Medline: 35322925]
62. Marshall T, Stellick C, Abba-Aji A, et al. The impact of shared decision-making on the treatment of anxiety and depressive disorders: systematic review. BJPsych Open 2021 Nov;7(6):e189. [doi: 10.1192/bjo.2021.1028]

Abbreviations

COPD: chronic obstructive pulmonary disease
LUP: Landsdækkende Undersøgelse af Patientoplevelser
SDM: shared decision-making
SDM-Q-9: 9-item Shared Decision-Making Questionnaire
STAI-S: State-Trait Anxiety Inventory-State

XSL•FO
RenderX

Edited by M Balcarras; submitted 29.Jul.2025; peer-reviewed by K Grootens, M Rosenlund; revised version received 26.Feb.2026; accepted 09.Mar.2026; published 29.Apr.2026.

Please cite as:
Frølund JC, Løkke A, Jensen HI, Skjøth F, Farver-Vestergaard I
The Effect of Informational Podcasts on Shared Decision-Making, Anxiety, and Patient Satisfaction in Hospital Visits: Intervention Study
J Med Internet Res 2026;28:e81485
URL: https://www.jmir.org/2026/1/e81485
doi:10.2196/81485

© Jannie Christina Frølund, Anders Løkke, Hanne Irene Jensen, Flemming Skjøth, Ingeborg Farver-Vestergaard. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Effect of Text Messaging on the Postoperative Pain Experience in Pediatric Patients Undergoing Thoracic Surgery: Randomized Controlled Trial

Yun Shi*, MD; Qianqiu Wang*, BA; Aihua Liu, BA; Li Jiang, MD

Department of Anesthesiology, Children's Hospital of Fudan University, 399 Wanyuan Road, Minhang District, Shanghai, China

*these authors contributed equally

Corresponding Author:
Li Jiang, MD
Department of Anesthesiology, Children's Hospital of Fudan University, 399 Wanyuan Road, Minhang District, Shanghai, China

Abstract

Background: Inadequately controlled postoperative pain continues to pose a significant clinical challenge in pediatric patients undergoing thoracic surgery.

Objective: This randomized controlled study aimed to investigate the effectiveness of SMS-based educational support for postoperative pain management on patients' pain experience.

Methods: A total of 100 pediatric patients undergoing thoracic surgery were enrolled between December 2, 2023, and January 28, 2025. Patients in the intervention group (group 1) received structured postoperative pain management education via SMS text messages, whereas those in the control group (group 2) received standard oral education. Pain intensity and pain-related interference were assessed using the Brief Pain Inventory, and self-efficacy was measured using the Pain Self-Efficacy Scale.

Results: The number of patients with moderate-to-severe average pain was significantly lower in group 1 than in group 2 (n=19, 18% vs n=19, 38%; *P*=.04), and group 1 had significantly higher patient self-efficacy scores (mean 29.3, SD 7.5 vs mean 25.2, SD 8.7; *P*=.01). Least pain scores were lower in group 1 compared with group 2 across all 3 consecutive postoperative days (PODs): POD1 (mean 2.8, SD 0.8 vs mean 3.3, SD 0.7; *P*<.001), POD2 (mean 2.2, SD 0.8 vs mean 2.7, SD 0.5; *P*<.001), and POD3 (mean 1.7, SD 0.7 vs mean2.2, SD 0.4; *P*<.001). Similarly, average pain intensity was lower in group 1 compared with group 2 across all 3 consecutive PODs: POD1 (mean 3.7, SD 0.9 vs mean 4.7, SD 0.6; *P*<.001); POD2 (mean 3.3, SD 0.8 vs mean 3.6, SD 0.5; *P*=.01); and POD3 (mean 2.5, SD 0.8 vs mean 3.1, SD 0.5; *P*<.001). General activity was significantly less affected in group 1 on POD 1 (mean 4.3, SD 1.0 vs mean 5.0, SD 1.5; *P*=.004) and POD 2 (mean 3.1, SD 0.7 vs mean 3.7, SD 1.3; *P*=.009).

Conclusions: The use of an SMS-based educational intervention significantly improved the postoperative pain experience of pediatric patients undergoing thoracic surgery. Further research is needed to clarify its impact on clinical outcomes and to better understand the mechanisms underlying improved pain management.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2300076554; https://www.chictr.org.cn/showprojEN.html?proj=203355

(J Med Internet Res 2026;28:e81806) doi:10.2196/81806

KEYWORDS

SMS; thoracic surgery; pain; pain self-efficacy; pediatric; text messaging

Introduction

Thoracic surgery is typically associated with substantial postoperative pain. To optimize analgesia, a multimodal pain management strategy is widely adopted, combining nonsteroidal anti-inflammatory drugs such as ibuprofen [1] and paracetamol [2], regional anesthesia [3], and a patient-controlled analgesia pump (PCAP) [4]. Within this framework, both patients and caregivers are encouraged to participate actively in postoperative pain management, including pain assessment and analgesic administration via a PCAP. Despite these efforts, pain following thoracic surgery often remains inadequately controlled. Previous studies have reported that approximately 54% of pediatric patients experience moderate-to-severe pain on the day of surgery, and both patients and parents frequently express dissatisfaction with postoperative pain management [5]. Pain has a profound impact on the quality of postoperative recovery. As a negative and inherently subjective experience, pain is perceived directly by children. Inadequate analgesia compromises effective breathing, coughing, and moving, thereby increasing the risk of respiratory complications and delaying

XSL•FO
RenderX

postoperative recovery [6,7]. Furthermore, insufficient control of acute postoperative pain has been associated with an elevated risk of developing chronic pain after discharge [8]. Several factors may contribute to suboptimal pain control in children, even in the presence of a PCAP. These include difficulties in accurately assessing pain severity, uncertainty regarding the appropriate timing of analgesic administration, challenges in adhering to medication regimens, and concerns regarding potential side effects of analgesics [9-11]. Collectively, these barriers may be attributed to insufficient educational information and/or ineffective educational support for patients and their caregivers [12].

With the rapid development of mobile technologies, internet-based medical support has emerged as an increasingly prominent component of health care delivery. The effectiveness of digital health interventions in enhancing self-management among pediatric patients with chronic diseases has been well documented [13,14]. These approaches emphasize patient-centered self-care, with health care professionals facilitating self-management by providing informational, emotional, and social support. Recently, emerging studies have investigated the use of SMS text messaging in pediatric pain management, with most focusing on pain assessment and monitoring. The feasibility and acceptability of SMS-based interventions have been widely demonstrated, as reflected by high response rates [15-17], strong caregiver acceptance [18], improved medication adherence, and reduced anxiety levels [19]. Several studies suggest that participants demonstrate improved pain coping and reduced anxiety and depression [20]. However, the effect of SMS on pain outcomes, particularly pain severity, remains unclear. More broadly, evidence regarding the impact of digital health tools, including mobile apps and other implementation strategies, on pediatric pain severity is limited and inconsistent. Some studies have reported improved pain self-management among pediatric patients [16,21-23], whereas others have found no significant benefit in pain control [20,24]. Notably, one study reported greater pain severity among patients in the intervention group [25]. Consequently, the use of SMS for postoperative pain management in pediatric populations remains in its early stages. Further research is warranted to evaluate the effects of SMS-based interventions on postoperative pain outcomes in pediatric populations.

In this prospective randomized controlled trial, an intervention group (group 1), which received SMS-based educational support for postoperative pain management, was compared with a control group (group 2), which received conventional educational support, to evaluate the effectiveness of SMS in helping manage the postoperative pain experience in pediatric patients undergoing thoracic surgery. The hypothesis was that pediatric patients receiving SMS-based educational support would demonstrate improved pain coping and reduced postoperative pain intensity.

Methods

Study Participants

Pediatric patients who underwent elective thoracic surgery at the Children's Hospital of Fudan University were consecutively recruited for this randomized controlled trial. Eligible patients were (1) aged 7 to 18 years, (2) classified as American Society of Anesthesiologists (ASA) physical status I or II, (3) expected to receive a combination of a preoperative paravertebral block and a PCAP for postoperative pain management, and (4) had access to a smartphone (iOS or Android). Exclusion criteria included (1) chronic pain (defined as pain persisting or recurring for more than 3 months), (2) a history of psychiatric disorders, (3) planned video-assisted thoracoscopic surgery, (4) inability to assess pain severity using a numeric rating scale (NRS), (5) communication difficulties, (6) any prior surgeries, (7) postoperative transfer to the intensive care unit (ICU), or (8) refusal to participate.

Study Design

Patients who met the inclusion criteria were consecutively enrolled during a preoperative visit. Participants were assigned identification numbers in the order of recruitment. An independent statistician generated the random allocation sequence using SPSS software (version 20; IBM Corp). The allocation sequence was placed in sealed envelopes, which were opened only after participant enrollment. Separate investigators were responsible for patient enrollment and follow-up. All patients underwent surgery under general anesthesia. Due to the study design, participants were not blinded to the intervention in either group, but they did not have information about each other. The statistician remained blinded to group allocation.

Procedure

Participants received comprehensive educational information on postoperative pain management. The information covered the cause of postoperative pain, advantages and disadvantages of PCAP, indications for PCAP, operational principles of PCAP, self-assessment of pain, pain management goals, proper use of PCAP, nonpharmacological pain relief methods, scheduled medication reminders, and guidance on early postoperative mobilization. Group 1 received information through a combination of text, illustrative images (eg, pain assessment scales), and instructional videos (eg, demonstrating how to operate the PCAP). Concise (Multimedia Appendix 1) and detailed (Multimedia Appendix 2) versions of the educational content were provided to facilitate participant understanding. The concise version emphasized practical operation and workflow, whereas the detailed version provided more comprehensive explanations to support a deeper understanding. The information was sent out manually on the day prior to surgery. Group 2 received conventional oral education support that provided the same information. All participants were instructed on the importance of self-monitoring and managing postoperative pain to promote engagement with pain self-management. Participants were encouraged to actively assess and manage their own pain, with parental assistance permitted as needed. They were advised to contact health care providers whenever necessary. However, participants in group 1 were not allowed to communicate with the investigators via SMS.

XSL•FO
RenderX

Outcomes and Measures

All patients underwent surgery under general anesthesia. A preoperative paravertebral block was performed, and a PCAP was initiated upon discharge from the postoperative anesthesia care unit. Demographic data, including age, sex, place of residence, and parental educational level, were collected prior to surgery. Patients were followed for the first 3 postoperative days (PODs) by a research nurse at 24, 48, and 72 hours after surgery. At each time point, the cumulative dose of hydromorphone administered via PCAP was retrieved from the system and recorded. If pain was inadequately controlled, additional rescue analgesic doses were permitted, and dosages were verified through patient medical records. As hydromorphone doses were calculated according to body weight, the cumulative dosage was analyzed and compared on a per-day, per–body weight basis.

Patient-reported pain intensity and pain interference were assessed over the 3 consecutive PODs using the Brief Pain Inventory (BPI) questionnaire. The BPI, recommended by the National Cancer Institute, is used to evaluate pain severity using an NRS ranging from 0 (no pain) to 10 (worst pain), capturing the highest, lowest, and average pain experienced. It also assesses the impact of pain on daily functioning, including activities, mood, sleep, interpersonal relationships, and enjoyment of life. Consequently, the BPI is widely used in clinical practice [26] and research to provide detailed information on pain experience [27-29]. Pain self-efficacy refers to a patient's belief and confidence regarding their ability to effectively manage pain and attain desired outcomes. Pain self-efficacy is recognized as an important resilience factor for pain, particularly in children and adolescents [30]. The Pain Self-Efficacy Scale (PSES) [31] was used to evaluate the patient's confidence in managing pain. The scale consists of 10 items, each scored from 1 to 6, and was administered both before surgery and on POD3. Patient satisfaction with postoperative pain management was evaluated on POD3 using an NRS (0=not satisfied at all, 10=very satisfied). Analgesic-related complications—including postoperative nausea and vomiting, urine retention, constipation, severe sedation, and respiratory depression—were also recorded.

The primary outcome of the study was the incidence of moderate-to-severe average pain on POD1, defined as an NRS score of 4 or greater. Secondary outcomes included cumulative hydromorphone consumption, pain experiences and their impact on daily life, pain self-efficacy, analgesic-related complications, and patient satisfaction.

Sample Size Estimation

A pilot study conducted at the Children's Hospital of Fudan University reported that the incidence of moderate-to-severe average pain on POD1 was 39.5%. A clinically meaningful reduction was defined as a one-third decrease in this incidence with the implementation of SMS for postoperative pain management. Assuming 80% power and a significance level of .05, the minimum required sample size was calculated as 45 patients per group. To account for an anticipated 10% loss to follow-up, the sample size was increased to 50 patients per group.

Data Analysis

Only participants who completed the entire questionnaire were included in the analysis. Data were analyzed using GraphPad Prism (version 9; Graphpad Corp) and SPSS software (version 20; IBM Corp). Continuous variables were tested for normality. Normally distributed data were expressed as means with SDs and compared using independent *t* tests. Conversely, nonnormally distributed data were expressed as medians with IQRs and compared using the Mann-Whitney *U* test. Categorical data were expressed as numbers and percentages and analyzed using the chi-square test. Univariate association between the incidence of moderate-to-severe average pain and PSES score was analyzed using the Poisson correlation.

Ethical Considerations

The study was conducted at the Children's Hospital of Fudan University and approved by the research ethics board (202328; approval date June 20, 2023; chairperson Li-Ling Qian). The study was registered at Chictr.org.cn (ChiCTR2300076554). Eligible patients who met the inclusion criteria and their parents were approached in a separate interview room 1 day before surgery. Detailed explanations of the study objectives, methods, potential benefits, and risks were provided. Participants were informed that their medical data would be reported publicly, but no personally identifiable information would be included. They were assured that participation would not affect the quality of care and that withdrawal from the study was permitted at any time. All the responders volunteered to participate free of charge. Signed informed consent was obtained from the parents of participants who agreed to participate in the study.

Results

Study Population

From December 2, 2023, to January 28, 2025, a total of 183 patients were assessed for eligibility. Of these, 134 patients met the inclusion criteria, and 100 (76.4%) agreed to participate. None of the patients were excluded from the analysis owing to missing data or loss to follow-up (Figure 1). No significant differences were observed between the groups in terms of age, sex, weight, diagnosis, and parental educational level (Table 1).

XSL•FO
RenderX

Figure 1. Study flow diagram.

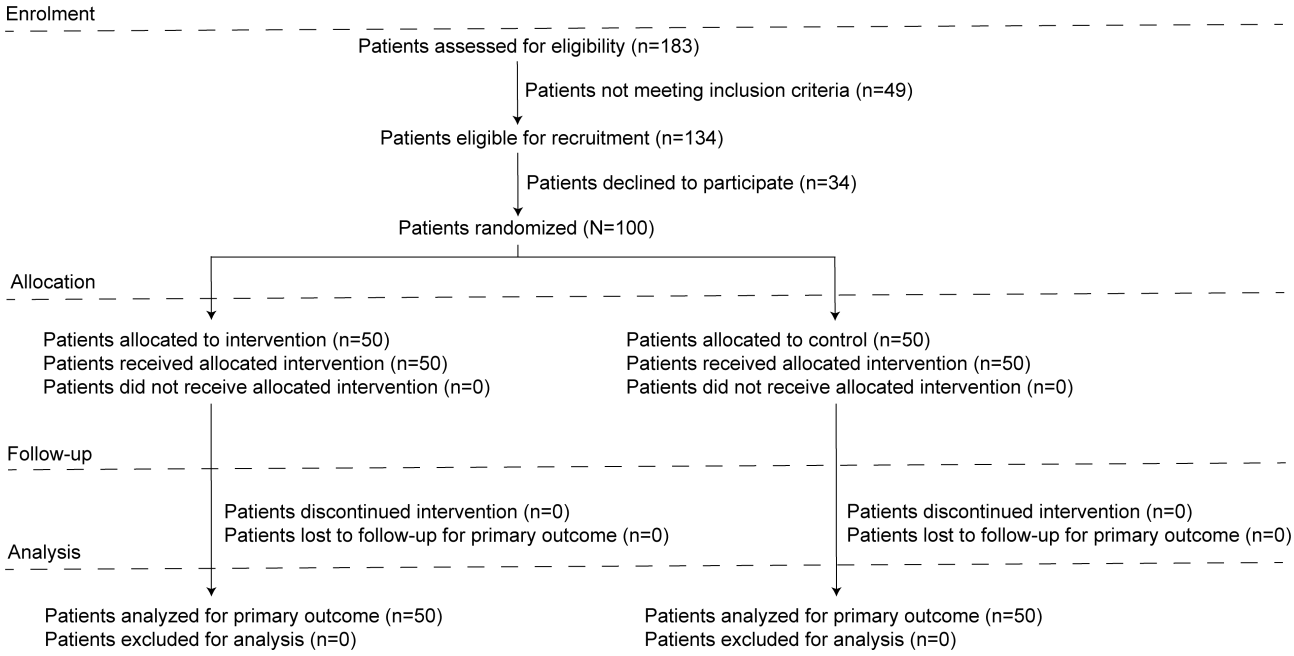

XSL•FO
RenderX

Table . Participant characteristics (N=100).

		Intervention group (n=50)	Control group (n=50)	P value
Age (y), mean (SD)		10.7 (2.8)	11.1 (2.9)	.48
Gender, n (%)				>.99
	Male	33 (66)	32 (64)	
	Female	17 (34)	18 (36)	
Weight (kg), mean (SD)		39.9 (11.7)	40.1 (14.2)	.94
Parent education, n (%)				.61
	Middle school	8 (16)	10 (20)	
	High school	10 (20)	12 (24)	
	Associate degree	18 (36)	15 (30)	
	Undergraduate degree	12 (24)	13 (26)	
	Master's degree and above	2 (4)	0 (0)	
Diagnosis, n (%)				.17
	Rib lesions	3 (6)	2 (4)	
	Diaphragmatic lesions	3 (6)	2 (4)	
	Pectus excavatum/carinatum	12 (24)	24 (48)	
	Mediastinal mass	16 (32)	12 (24)	
	Pulmonary lesions	16 (32)	10 (20)	
History of surgery, n (%)				.68
	Yes	22 (44)	19 (38)	
	No	28 (56)	31 (62)	
Experience with PCAP, n (%)				.15
	Yes	15 (30)	8 (16)	
	No	35 (70)	42 (84)	
PCAP operator, n (%)				.69
	Patient	0 (0)	0 (0)	
	Parent	22 (44)	25 (50)	
	Both	28 (56)	25 (50)	

[a]PCAP: patient-controlled analgesia pump.

Primary Outcome

The incidence of moderate-to-severe average pain on POD1 was significantly lower in group 1 than in group 2 (n=9, 18% vs n=19, 38%; *P*=.04), whereas no difference was detected between the groups on POD2 and POD3 (Table 2).

XSL•FO
RenderX

Table . Pain experience among pediatric patients after thoracic surgery (N=100).

	Postoperative day 1			Postoperative day 2			Postoperative day 3		
	Group 1 (n=50)	Group 2 (n=50)	*P* value	Group 1 (n=50)	Group 2 (n=50)	*P* value	Group 1 (n=50)	Group 2 (n=50)	*P* value
Primary outcome, n (%)									
Patients with moderate-to-severe average pain, n (%)	9 (18)	19 (38)	.04	0 (0)	1 (2)	>.99	0 (0)	1 (2)	>.99
Secondary outcomes									
NRS[a] score, mean (SD)									
Worst pain	6.4 (1.1)	6.1 (0.7)	.58	5.3 (0.9)	5.2 (0.7)	.40	4.7 (1.1)	4.7 (0.6)	.80
Least pain	2.8 (0.8)	3.3 (0.7)	<.001	2.2 (0.8)	2.7 (0.5)	<.001	1.7 (0.7)	2.2 (0.4)	<.001
Average pain	3.7 (0.9)	4.7 (0.6)	<.001	3.3 (0.8)	3.6 (0.5)	.01	2.5 (0.8)	3.1 (0.5)	<.001
Patients reporting pain trends, n (%)			.005			.83			.03
Decreasing	11 (22)	2 (4)		35 (70)	34 (68)		41 (82)	30 (60)	
Increasing	20 (40)	34 (68)		0 (0)	0 (0)		0 (0)	0 (0)	
Unchanged	19 (38)	14 (28)		15 (30)	16 (32)		9 (18)	20 (40)	
Pain interference with daily functioning (0-10 scale), mean (SD)									
General activities	4.3 (1.0)	5.0 (1.5)	.004	3.1 (0.7)	3.7 (1.3)	.009	2.6 (1.3)	2.5 (0.7)	.69
Mood	4.3 (2.4)	3.8 (1.7)	.30	2.4 (0.8)	2.8 (2.1)	.19	1.8 (1.2)	1.7 (0.7)	.65
Walking	9.9 (0.4)	9.9 (0.3)	.72	8.8 (0.9)	8.9 (0.6)	.52	7.4 (1.4)	8.2 (0.7)	<.001
Sleeping	3.9 (1.8)	4.1 (1.3)	.50	2.3 (1.0)	2.8 (0.7)	.11	1.3 (0.8)	2.2 (0.5)	<.001
Relationship with others	1.9 (1.2)	2.3 (1.0)	.14	1.3 (1.0)	1.5 (0.6)	.39	0.7 (0.8)	1.0 (0.4)	<.001
Enjoyment of life	2.4 (1.1)	2.7 (1.0)	.09	1.6 (0.9)	1.8 (0.5)	.12	1.0 (0.01)	1.2 (0.4)	.01
Cumulative consumption of hydromorphone (μg/kg/d)	43.9 (8.3)	38.7 (7.0)	<.001	33.1 (10.8)	32.8 (6.8)	.86	28.7 (9.6)	27.6 (6.2)	.53

XSL•FO
RenderX

	Postoperative day 1			Postoperative day 2			Postoperative day 3		
	Group 1 (n=50)	Group 2 (n=50)	*P* value	Group 1 (n=50)	Group 2 (n=50)	*P* value	Group 1 (n=50)	Group 2 (n=50)	*P* value
Patients with analgesic-related complications, n (%)			.23			>.99			>.99
Postoperative nausea and vomiting	29 (58)	21 (42)		5 (10)	4 (8)		0 (0)	0 (0)	
Other	0 (0)	0 (0)		0 (0)	0 (0)		0 (0)	0 (0)	

[a]NRS: numeric rating scale.

Secondary Outcomes

Pain experiences—including pain intensity, treatment, analgesic consumption, and complications—and their impact on patients' daily lives are summarized in Table 2. Specifically, the intensity of least pain was significantly lower in group 1 versus group 2 on each of the 3 consecutive days: POD1 (mean 2.8, SD 0.8 vs mean 3.3, SD 0.7; *P*<.001), POD2 (mean 2.2, SD 0.8 vs mean 2.7, SD 0.5; *P*<.001), and POD3 (mean 1.7, SD 0.7 vs mean 2.2, SD 0.4; *P*<.001). Similarly, average pain was lower in group 1 versus group 2 across the 3 consecutive days: POD1 (mean 3.7, SD 0.9 vs mean 4.7, SD 0.6; *P*<.001), POD2 (mean 3.3, SD 0.8 vs mean 3.6, SD 0.5; *P*=.01), and POD3 (mean 2.5, SD 0.8 vs mean 3.1, SD 0.5; *P*<.001).

By contrast, no differences were observed in the intensity of worst pain between the groups. Although NRS scores for worst, least, and average pain decreased significantly over time in both groups, a significantly higher proportion of patients in group 1 than in group 2 reported pain relief on POD1 (n=11, 22% vs n=2, 4%; *P*=.005) and POD 3 (n=41, 82% vs n=30, 60%; *P*=.03). The impact of pain on daily activities was significantly lower in group 1 than in group 2 on POD1 (mean 4.3, SD 1.0 vs mean 5.0, SD 1.5; *P*=.004) and POD2 (mean 3.1, SD 0.7 vs mean 3.7, SD 1.3; *P*=.009). No significant between-group differences were observed in the impact of pain on mood, walking, sleeping, relationships with others, and enjoyment of life, except on POD3. Cumulative hydromorphone consumption on POD1 was significantly higher in group 1 than in group 2 (mean 43.9, SD 8.3 vs mean 38.7, SD 7.0; *P*<.001).

Patient self-efficacy for postoperative pain was evaluated using the PSES (Table 3). No significant differences were observed between the 2 groups prior to surgery; however, the PSES score was significantly higher in group 1 than in group 2 on POD3 (mean 29.3, SD 7.5 vs mean 25.2, SD 8.7; *P*=.01). No significant differences were found between the 2 groups in patient satisfaction on POD3 (mean 8.0, SD 1.6 vs mean 7.7, SD 1.9; *P*=.40) or in the duration of the hospital stay (mean 8.4, SD 0.97 vs mean 8.1, SD 0.89 days; *P*=.11). To explore the correlation between pain self-efficacy and pain severity, a univariate logistic regression analysis was performed. The incidence of moderate-to-severe pain was not significantly associated with the PSES score (*P*=.59). No differences were detected in the incidence of analgesic-related complications across the 3 PODs. The recorded adverse events consisted primarily of postoperative nausea and vomiting (Table 2), all of which resolved spontaneously.

Table . Patient Self-Efficacy Scale scores in postoperative pain assessment.

	Group 1	Group 2	*P* value
Presurgery, mean (SD)	26.8 (7.8)	24.4 (7.7)	.12
Postoperative day 3, mean (SD)	29.3 (7.5)	25.2 (8.7)	.01

Discussion

Principal Findings

Overall, this study demonstrated that the number of patients with moderate-to-severe average pain on POD1 after thoracic surgery in pediatric patients was significantly lower in group 1 than in group 2 following the implementation of SMS-based education (n=9, 18% vs n=19, 38%). Furthermore, both the lowest and average NRS scores decreased over the first 3 PODs, and a greater proportion of patients in group 1 than in group 2 reported a trend of decreasing pain (n=11, 22% vs n=2, 4%). Collectively, these findings indicate that the implementation of SMS-based education led to an improvement in postoperative pain management in pediatric patients. Notably, group 1 not only exhibited lower absolute NRS scores than group 2, but these scores also reflected a clinically relevant difference in pain severity: group 1 experienced mild pain, whereas group 2 experienced moderate pain (mean 3.7, SD 0.9 vs mean 4.7, SD 0.6). Together with the primary outcome showing that significantly more patients in group 2 experienced moderate-to-severe average pain, these improvements in pain experience are clinically meaningful. This interpretation is further supported by the significantly reduced impact of pain on general activities observed in group 1.

XSL•FO
RenderX

Group 1 demonstrated a higher cumulative consumption of analgesics on POD1, indicating that this group received more rescue analgesic doses through PCAP. This finding may partly explain the observed improvement in postoperative pain control in participants. Owing to the limited time available for preoperative education and the difficulty in effectively explaining or retaining detailed information, most patients and parents have an incomplete understanding of pain assessment and the goals of pain management [9,10,32]. Moreover, concerns regarding potential side effects often lead parents to be overly cautious about analgesic administration [11,33,34], resulting in a tendency to encourage children to tolerate pain rather than to administer analgesics proactively. Consequently, even with the PCAP, pain relief in pediatric patients may remain inadequate [5,35]. By contrast, parents and patients in group 1 were able to consult mobile push messages at any time and receive ongoing educational support when questions arose. This access facilitated more accurate assessment of pain intensity and timely administration of rescue analgesic doses in accordance with medical staff expectations.

Pain self-efficacy may represent another important determinant of pain experienced in this study. Although evidence regarding the impact of pain self-efficacy on postoperative pain in pediatric patients remains limited, studies in adults have demonstrated that enhancing pain self-efficacy can reduce postoperative pain intensity [36] and improve pain-related outcomes and patients' overall recovery quality [37]. Moreover, studies conducted in pediatric patients with chronic pain have consistently reported beneficial effects of pain self-efficacy on children's pain and quality of life [38-40]. Patients who are more confident in their pain management strategies typically experience less pain and distress. Accordingly, the lower pain intensity and improved general activity observed in group 1 may, at least in part, be attributable to enhanced pain self-efficacy, which is consistent with the findings of previous studies. However, this study did not detect a causal relationship between pain self-efficacy and pain severity. This may reflect the multifactorial nature of pain and be related to the relatively small sample size. Consequently, pain self-efficacy could not be confirmed as a variable independently associated with pain severity in this study. Further studies are warranted to clarify the underlying mechanisms.

Finally, pain is a highly subjective experience influenced not only by pharmacological and nonpharmacological analgesic strategies for analgesia but also by various other factors [41]. Previous studies have reported that placebo effects can alleviate pain by enhancing patients' expectations of beneficial outcomes [42]. In this study, given the ubiquity of mobile phone use, the participants in group 1 could easily access relevant information anytime and anywhere, which may have strengthened their expectations of effective pain control. Moreover, because parents assume primary responsibility for postoperative pain management in children, their cognition, attitudes toward pain, and psychological states—such as helplessness, stress, depression, or confidence—inevitably influence children's perceptions of pain. Parental responses to children's pain and their sense of security have been shown to affect pediatric pain experiences. Accordingly, the observed changes in pain experience in this study may be partly attributable to the sense of security experienced by both parents and children, knowing that guidance and support were readily accessible through SMS.

Although general activity was significantly better in group 1 than in group 2, no significant between-group differences were observed in other domains of pain interference. Although group 1 reported better sleep, social functioning, and enjoyment of life compared with group 2 on POD3, scores in both groups remained at a satisfactorily low level. Therefore, further studies are warranted to determine whether these differences have a significant impact on clinical outcomes despite more favorable subjective evaluations. In addition, the highest NRS scores in both groups remained above 4 points over the 3 PODs, with no significant between-group differences. Similarly, walking ability scores in both groups remained above 5 points during this period, although group 1 performed significantly better than group 2 on POD3. These findings suggest that pain during movement was inadequately controlled in both groups. This limitation may explain the absence of a significant difference in overall patient satisfaction despite improvements in average pain intensity and general activity in group 1. It may also be related to the relatively low analgesic consumption observed among study participants compared with that reported in previous studies [43,44]. Further studies are needed to optimize pain management and achieve more effective postoperative analgesia in patients.

Overall, the data indicate that implementing SMS for postoperative pain management in pediatric patients undergoing thoracic surgery significantly enhanced pain self-efficacy, effectively reduced pain intensity, and improved daily functioning. These findings indicate the importance of patients' subjective experiences in clinical practice and highlight the role of interventions that increase patient engagement and confidence in pain self-management. This study provides a framework for progressing interventions from feasible and acceptable to effective. Given the limitations of existing research, future studies should examine a broader range of clinical scenarios and more diverse patient populations. This could include exploring different modalities (eg, SMS, mobile apps, and web-based platforms), varying disease contexts (eg, chronic pain, cancer, and postoperative pain), different care settings (eg, hospital and home), and relevant comorbidities (eg, depression and anxiety) to validate findings and enhance generalizability. Such efforts would support the translation of research findings and facilitate the clinical implementation of high-quality SMS interventions for pain management in pediatric populations.

Study Limitations

This study has some limitations that should be considered when interpreting the results. First, the study was conducted at a single center, the Children's Hospital of Fudan University, one of the largest pediatric medical centers in an economically developed region of China. Parents of patients at this center generally have a higher socioeconomic status, better level of education, and greater compliance with medical directives. Consequently, the single-center design and relatively small sample size may limit the generalizability of the findings. Second, the study focused solely on the impact of SMS text messages on postoperative acute pain management in pediatric patients classified as ASA

XSL•FO
RenderX

physical status I or II. Inclusion of patients with ASA physical status III or higher could have broadened the applicability of the results. However, at the Children's Hospital of Fudan University, the pediatric patients were transferred to the intensive care unit, where more frequent and detailed pain assessments and interventions are performed by medical staff. Such intensive management would inevitably influence patients' self-management and the study outcomes, thereby justifying the exclusion of patients with more severe conditions.

Second, there is a significant concomitant relationship between pain and psychiatric disorders, with complex interactions. For example, pediatric patients with attention-deficit/hyperactivity disorder are more prone to pain [45], which may be related to increased pain sensitivity [46]. Pain and pain-related dysfunction may be further exacerbated in patients with depression and anxiety [47], while patients with autism spectrum disorders may exhibit abnormal pain perception [48]. Thus, pain management for pediatric patients with comorbid psychiatric disorders is a distinct and complex topic; however, it was not the focus of this study. Therefore, children with psychiatric disorders were excluded from this study. Owing to variations in pain characteristics and patient populations, the results obtained in this specific cohort may not be generalizable to broader populations. However, the findings can serve as a reference for calculating sample sizes in future studies.

Third, self-reporting bias is another consideration. Pain, the primary outcome, was assessed using self-reported measures, which may be influenced by memory errors, social desirability bias, or comprehension differences. Randomization was used during the study design to ensure balanced baseline characteristics and minimize selection bias. Although participants were not blinded to group allocation, data collectors were blinded, reducing the risk of researcher expectation bias and its potential impact on participants' reported outcomes. During the data collection phase, validated standardized questionnaires and scales were used to quantify the participants' self-reported results, ensuring consistency in data collection and reducing differences in subjective interpretation. Furthermore, pain experience was evaluated from multiple perspectives. In addition to the primary outcome of moderate-to-severe pain incidence, secondary assessments included quantified pain scores, daily pain intensity trends, pain-related functional impairment, and quality of life measures, such as general activity, mood, and sleep. Continuous follow-up over 3 PODs enabled a comprehensive assessment across multiple dimensions and reduced reliance on single self-reports. Participants also received thorough education regarding research objectives and reporting procedures, thereby enhancing compliance and mitigating single-report bias. However, reporting bias could not be entirely eliminated.

Finally, the study did not investigate the underlying causes of pain and/or contributing factors, including individual pain perception, nonpharmacological strategies, psychological status, or the interaction of analgesic use and pain experience, which limited insight into the mechanisms influencing the observed outcomes.

Conclusions

The implementation of SMS-based postoperative pain management in pediatric patients undergoing thoracic surgery improved clinical outcomes. However, its effects on patient outcomes and the underlying mechanisms require further investigation.

Acknowledgments

We would like to thank Jun Xia for enrolling participants, Peng Shi for assisting in data analysis, and all the nursing staff in the thoracic surgery ward for recording patient-reported complications. No generative artificial intelligence was used in any portion of the manuscript preparation.

Funding

The authors declared no financial support was received for this work.

Data Availability

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy and ethical restrictions.

Authors' Contributions

Conceptualization: LJ, YS
Data curation: QW, YS
Formal analysis: AL, QW
Investigation: QW, AL
Methodology: YS
Project administration: LJ
Supervision: LJ, YS
Validation: LJ, YS
Writing—original draft: QW
Writing—review and editing: LJ, YS

XSL•FO
RenderX

The manuscript was approved by all authors, and the order of authors listed in the manuscript has been approved by all authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example of the educational information sent to participants in group 1.
[PNG File, 337 KB - jmir_v28i1e81806_app1.png]

Multimedia Appendix 2
Detailed version of the information sent to participants, using self-assessment methods for pain (translated to English).
[PNG File, 272 KB - jmir_v28i1e81806_app2.png]

Checklist 1
CONSORT-eHEALTH checklist (V1.6.1).
[PDF File, 870 KB - jmir_v28i1e81806_app3.pdf]

References

1. Poddighe D, Brambilla I, Licari A, Marseglia GL. Ibuprofen for pain control in children: new value for an old molecule. Pediatr Emerg Care 2019 Jun;35(6):448-453. [doi: 10.1097/PEC.0000000000001505] [Medline: 29912084]
2. Macario A, Royal MA. A literature review of randomized clinical trials of intravenous acetaminophen (paracetamol) for acute postoperative pain. Pain Pract 2011;11(3):290-296. [doi: 10.1111/j.1533-2500.2010.00426.x] [Medline: 21114616]
3. He Y, Xu M, Li Z, Deng L, Kang Y, Zuo Y. Safety and feasibility of ultrasound-guided serratus anterior plane block and intercostal nerve block for management of post-sternotomy pain in pediatric cardiac patients: a prospective, randomized trial. Anaesth Crit Care Pain Med 2023 Dec;42(6):101268. [doi: 10.1016/j.accpm.2023.101268] [Medline: 37364851]
4. El-Boghdadly K, Levy NA, Fawcett WJ, et al. Peri-operative pain management in adults: a multidisciplinary consensus statement from the Association of Anaesthetists and the British Pain Society. Anaesthesia 2024 Nov;79(11):1220-1236. [doi: 10.1111/anae.16391] [Medline: 39319373]
5. Tomaszek L, Fenikowski D, Cież-Piekarczyk N, Mędrzycka-Dąbrowska W. Maximum pain at rest in pediatric patients undergoing elective thoracic surgery and the predictors of moderate-to-severe pain-secondary data analysis. J Clin Med 2024 Feb 1;13(3):844. [doi: 10.3390/jcm13030844] [Medline: 38337538]
6. Sniecinski RM, Bennett-Guerrero E, Shore-Lesserson L. Anticoagulation management and heparin resistance during cardiopulmonary bypass: a survey of Society of Cardiovascular Anesthesiologists members. Anesth Analg 2019 Aug;129(2):e41-e44. [doi: 10.1213/ANE.0000000000003981] [Medline: 30540616]
7. Miyazaki T, Doi R, Matsumoto K. Post-thoracotomy pain syndrome in the era of minimally invasive thoracic surgery. J Thorac Dis 2024 May 31;16(5):3422-3430. [doi: 10.21037/jtd-24-158] [Medline: 38883660]
8. Rabbitts JA, Fisher E, Rosenbloom BN, Palermo TM. Prevalence and predictors of chronic postsurgical pain in children: a systematic review and meta-analysis. J Pain 2017 Jun;18(6):605-614. [doi: 10.1016/j.jpain.2017.03.007] [Medline: 28363861]
9. Palomaa AK, Hakala M, Pölkki T. Parents' perceptions of their child's pain assessment in hospital care: a cross-sectional study. J Pediatr Nurs 2023;71:79-87. [doi: 10.1016/j.pedn.2023.03.012] [Medline: 37030017]
10. Escobar-Castellanos M, Míguez-Navarro MC, García-Mancebo J, et al. How much do parents know about pain in their children? Pediatr Emerg Care 2023 Jan 1;39(1):40-44. [doi: 10.1097/PEC.0000000000002585] [Medline: 36580891]
11. Yan C, Cheung RSY, Wong CL, et al. Stress and perception of procedural pain management in Chinese parents of children with cancer. J Pain Symptom Manage 2021 Jan;61(1):90-102. [doi: 10.1016/j.jpainsymman.2020.06.028] [Medline: 32640278]
12. Hu Y, Shi Y. Differences between parents and nurses in perceived educational support for pediatric postoperative pain management. J Eval Clin Pract 2025 Dec;31(8):e70338. [doi: 10.1111/jep.70338] [Medline: 41308016]
13. Bentsen L, Hanghøj S, Hjerming M, et al. Development of quality of life in adolescents and young adults with cancer using a patient support smartphone app: prepost interventional study. JMIR Cancer 2023 Dec 4;9:e49735. [doi: 10.2196/49735] [Medline: 38048144]
14. Versteegh LA, Chang AB, Chirgwin S, Tenorio FP, Wilson CA, McCallum GB. Multi-lingual "asthma app" improves health knowledge of asthma among Australian First Nations carers of children with asthma. Front Pediatr 2022;10:925189. [doi: 10.3389/fped.2022.925189] [Medline: 36110111]
15. Chen Y, Chin M, Greenberg S, Johnstone C, McGuinness J. Post-tonsillectomy pain in 24 children - utilising short message service (SMS) to assess postoperative outcomes. Clin Otolaryngol 2012 Oct;37(5):412-414. [doi: 10.1111/j.1749-4486.2012.02521.x] [Medline: 23164270]

XSL•FO
RenderX

16. Nelson SE, Adams AJ, Buczek MJ, Anthony CA, Shah AS. Postoperative pain and opioid use in children with supracondylar humeral fractures: balancing analgesia and opioid stewardship. J Bone Joint Surg Am 2019 Jan 16;101(2):119-126. [doi: 10.2106/JBJS.18.00657] [Medline: 30653041]
17. Drake-Brockman TFE, Smallbone HE, Sommerfield D, von Ungern-Sternberg BS. Remote after-care using smartphones: a feasibility study of monitoring children's pain with automated SMS messaging. Paediatr Anaesth 2022 Aug;32(8):954-960. [doi: 10.1111/pan.14481] [Medline: 35569012]
18. Schroeder K, Maiarana J, Gisiri M, et al. Caregiver acceptability of mobile phone use for pediatric cancer care in Tanzania: cross-sectional questionnaire study. JMIR Pediatr Parent 2021 Dec 8;4(4):e27988. [doi: 10.2196/27988] [Medline: 34889763]
19. Newton L, Sulman C. Use of text messaging to improve patient experience and communication with pediatric tonsillectomy patients. Int J Pediatr Otorhinolaryngol 2018 Oct;113:213-217. [doi: 10.1016/j.ijporl.2018.07.048]
20. Grasaas E, Helseth S, Fegran L, et al. App-based intervention among adolescents with persistent pain: a pilot feasibility randomized controlled trial. Pilot Feasibility Stud 2022 Jul 27;8(1):158. [doi: 10.1186/s40814-022-01113-0] [Medline: 35897086]
21. Palermo TM, de la Vega R, Murray C, Law E, Zhou C. A digital health psychological intervention (WebMAP Mobile) for children and adolescents with chronic pain: results of a hybrid effectiveness-implementation stepped-wedge cluster randomized trial. Pain 2020 Dec;161(12):2763-2774. [doi: 10.1097/j.pain.0000000000001994] [Medline: 32658147]
22. Simon JDHP, Schepers SA, van Gorp M, et al. Pain monitoring app leads to less pain in children with cancer at home: results of a randomized controlled trial. Cancer 2024 Jul 1;130(13):2339-2350. [doi: 10.1002/cncr.35100] [Medline: 37947136]
23. Shaygan M, Jaberi A. The effect of a smartphone-based pain management application on pain intensity and quality of life in adolescents with chronic pain. Sci Rep 2021 Mar 23;11(1):6588. [doi: 10.1038/s41598-021-86156-8] [Medline: 33758322]
24. Lalloo C, Harris LR, Hundert AS, et al. The iCanCope pain self-management application for adolescents with juvenile idiopathic arthritis: a pilot randomized controlled trial. Rheumatology (Oxford) 2021 Jan 5;60(1):196-206. [doi: 10.1093/rheumatology/keaa178] [Medline: 32613229]
25. Tiozzo E, Biagioli V, Brancaccio M, et al. Using an app to monitor postoperative pain at home in pediatric patients. J Child Health Care 2021 Jun;25(2):212-224. [doi: 10.1177/1367493520919313] [Medline: 33571013]
26. Ohashi K, Suzuki H, Sata Y, et al. Postoperative pain and quality of life after lung cancer surgery: a prospective observational study. Ann Palliat Med 2023 Mar;12(2):346-355. [doi: 10.21037/apm-22-207] [Medline: 36627847]
27. Lu JJ, Zhang FY, Feng Y, Yang YF, Guo Y, Feng YZ. The Chinese version of the Brief Pain Inventory-Facial (BPI-F) among different populations: factor structure and measurement invariance. J Dent 2023 Dec;139:104771. [doi: 10.1016/j.jdent.2023.104771] [Medline: 37925049]
28. Wang XS, Mendoza TR, Gao SZ, Cleeland CS. The Chinese version of the Brief Pain Inventory (BPI-C): its development and use in a study of cancer pain. Pain 1996 Oct;67(2-3):407-416. [doi: 10.1016/0304-3959(96)03147-8] [Medline: 8951936]
29. Tan G, Jensen MP, Thornby JI, Shanti BF. Validation of the Brief Pain Inventory for chronic nonmalignant pain. J Pain 2004 Mar;5(2):133-137. [doi: 10.1016/j.jpain.2003.12.005] [Medline: 15042521]
30. Stahlschmidt L, Hübner-Möhler B, Dogan M, Wager J. Pain self-efficacy measures for children and adolescents: a systematic review. J Pediatr Psychol 2019 Jun 1;44(5):530-541. [doi: 10.1093/jpepsy/jsz002] [Medline: 30802913]
31. Stahlschmidt L, Dogan M, Hübner-Möhler B, et al. Development and validation of the Scale for Pain Self-Efficacy (SPaSE) in German and English languages for children and adolescents. J Pain 2023 Jun;24(6):1069-1079. [doi: 10.1016/j.jpain.2023.01.007] [Medline: 36646401]
32. Valizadeh F, Ahmadi F, Zarea K. Neglect of postoperative pain management in children: a qualitative study based on the experiences of parents. J Pediatr Nurs 2016;31(4):439-448. [doi: 10.1016/j.pedn.2016.02.012] [Medline: 27033024]
33. Rony RYZ, Fortier MA, Chorney JM, Perret D, Kain ZN. Parental postoperative pain management: attitudes, assessment, and management. Pediatrics 2010 Jun;125(6):e1372-e1378. [doi: 10.1542/peds.2009-2632] [Medline: 20498177]
34. Zhang SM, Chan YY, Lai LL, Liang PR, Lin M. Parental postoperative pain management perceptions, attitudes, and practices in pediatric limb fractures. Pain Manag Nurs 2023 Oct;24(5):506-512. [doi: 10.1016/j.pmn.2023.07.006] [Medline: 37574333]
35. Gan TJ. Poorly controlled postoperative pain: prevalence, consequences, and prevention. J Pain Res 2017;10:2287-2298. [doi: 10.2147/JPR.S144066] [Medline: 29026331]
36. Wang L, Qin F, Liu H, Lu XH, Zhen L, Li GX. Pain sensitivity and acute postoperative pain in patients undergoing abdominal surgery: the mediating roles of pain self-efficacy and pain catastrophizing. Pain Manag Nurs 2024 Apr;25(2):e108-e114. [doi: 10.1016/j.pmn.2023.12.001] [Medline: 38184383]
37. Lee YJ, Bettick D, Rosenberg C. Improving pain self-efficacy in orthopedic surgery patients through video-based education: a quality improvement project. Pain Manag Nurs 2024 Oct;25(5):451-458. [doi: 10.1016/j.pmn.2024.04.008] [Medline: 38719657]
38. Tomlinson RM, Bax KC, Ashok D, McMurtry CM. Health-related quality of life in youth with abdominal pain: an examination of optimism and pain self-efficacy. J Psychosom Res 2021 Aug;147:110531. [doi: 10.1016/j.jpsychores.2021.110531] [Medline: 34082155]

XSL•FO
RenderX

39. Nabbijohn AN, Newby-Clark IR, Mack D, Stintzi A, McMurtry CM. Pain trajectories in pediatric inflammatory bowel disease: disease severity, optimism, and pain self-efficacy. Clin J Pain 2025 May 1;41(5):e1279. [doi: 10.1097/AJP.0000000000001279] [Medline: 40052200]
40. Haraldstad K, Stea TH. Associations between pain, self-efficacy, sleep duration, and symptoms of depression in adolescents: a cross-sectional survey. BMC Public Health 2021 Sep 5;21(1):1617. [doi: 10.1186/s12889-021-11680-1] [Medline: 34482826]
41. Colloca L, Barsky AJ. Placebo and nocebo effects. N Engl J Med 2020 Feb 6;382(6):554-561. [doi: 10.1056/NEJMra1907805] [Medline: 32023375]
42. Okusogu C, Colloca L. Placebo hypoalgesia: above and beyond expectancy and conditioning. Curr Opin Behav Sci 2019 Apr;26:75-81. [doi: 10.1016/j.cobeha.2018.10.008] [Medline: 31538101]
43. Abourezk AR, Abdelbaser I, Badran A, Abdelfattah M. Ultrasound-guided mid-point transverse process to pleura block versus thoracic paravertebral block in pediatric open-heart surgery: a randomized controlled non-inferiority study. J Clin Anesth 2024 Oct;97:111507. [doi: 10.1016/j.jclinane.2024.111507] [Medline: 38852396]
44. Qi J, Du B, Gurnaney H, Lu P, Zuo Y. A prospective randomized observer-blinded study to assess postoperative analgesia provided by an ultrasound-guided bilateral thoracic paravertebral block for children undergoing the Nuss procedure. Reg Anesth Pain Med 2014;39(3):208-213. [doi: 10.1097/AAP.0000000000000071] [Medline: 24682079]
45. Mundal I, Schei J, Lydersen S, Thomsen PH, Nøvik TS, Kvitland LR. Prevalence of chronic and multisite pain in adolescents and young adults with ADHD: a comparative study between clinical and general population samples (the HUNT study). Eur Child Adolesc Psychiatry 2024 May;33(5):1433-1442. [doi: 10.1007/s00787-023-02249-x] [Medline: 37386203]
46. Bouchatta O, Aby F, Sifeddine W, et al. Pain hypersensitivity in a pharmacological mouse model of attention-deficit/hyperactivity disorder. Proc Natl Acad Sci U S A 2022 Jul 26;119(30):e2114094119. [doi: 10.1073/pnas.2114094119] [Medline: 35858441]
47. Dudeney J, Aaron RV, Hathway T, et al. Anxiety and depression in youth with chronic pain: a systematic review and meta-analysis. JAMA Pediatr 2024 Nov 1;178(11):1114-1123. [doi: 10.1001/jamapediatrics.2024.3039] [Medline: 39250143]
48. Balasco L, Provenzano G, Bozzi Y. Sensory abnormalities in autism spectrum disorders: a focus on the tactile domain, from genetic mouse models to the clinic. Front Psychiatry 2019;10:1016. [doi: 10.3389/fpsyt.2019.01016] [Medline: 32047448]

Abbreviations

ASA: American Society of Anesthesiologists
BPI: Brief Pain Inventory
NRS: numeric rating scale
PCAP: patient-controlled analgesia pump
POD: postoperative day
PSES: Pain Self-Efficacy Scale

Edited by A Schwartz, M Balcarras; submitted 04.Aug.2025; peer-reviewed by C Brimeyer, D Poddighe; revised version received 08.Feb.2026; accepted 09.Feb.2026; published 24.Mar.2026.

Please cite as:
Shi Y, Wang Q, Liu A, Jiang L
The Effect of Text Messaging on the Postoperative Pain Experience in Pediatric Patients Undergoing Thoracic Surgery: Randomized Controlled Trial
J Med Internet Res 2026;28:e81806
URL: https://www.jmir.org/2026/1/e81806
doi:10.2196/81806

© Yun Shi, Qianqiu Wang, Aihua Liu, Li Jiang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Designing Electronic Problem-Solving Training for Individuals With Traumatic Brain Injury: Mixed Methods, Community-Based, Participatory Research Case Study

Matthew Schmidt[1,2], PhD; Yueqi Weng[2], MEd; Shannon Juengst[3], PhD, CRC; Alexandra Holland[3], BS

[1]College of Pharmacy, Department of Clinical and Administrative Pharmacy, University of Georgia, River's Crossing, 215, Athens, GA, United States

[2]Department of Workforce Education and Instructional Technology, Mary Francis Early College of Education, University of Georgia, Athens, GA, United States

[3]Brain Injury Research Center, TIRR Memorial Hermann, Houston, TX, United States

Corresponding Author:
Matthew Schmidt, PhD
College of Pharmacy, Department of Clinical and Administrative Pharmacy, University of Georgia, River's Crossing, 215, Athens, GA, United States

Abstract

Background: Traditional rehabilitation research often excludes the voices of individuals with lived experience of traumatic brain injury (TBI), resulting in interventions that lack relevance, accessibility, and effectiveness. Community-based participatory research (CBPR) offers an alternative framework that emphasizes collaboration, power sharing, and sustained engagement with patients, caregivers, and clinicians.

Objective: This study aimed to apply CBPR to guide front-end design (empathy interviews, empathy mapping, personas) and to evaluate the sociotechnical-pedagogical usability of the Electronic Problem-Solving Training (ePST) mobile health (mHealth) intervention with TBI partners.

Methods: A multistep, mixed methods design case methodology was adopted, guided by CBPR principles and learning experience design. Participatory mechanisms included a 33-member Community Advisory Board and 10 Community Engagement Studios that engaged TBI survivors, caregivers, clinicians, and researchers throughout the Discover, Define, Develop, and Deliver phases of the Double Diamond model. Iterative activities included empathy interviews (n=14), persona development (n=10), rapid prototyping, and usability testing with 5 participants with TBI using think-aloud protocols and the Comprehensive Assessment of Usability for Learning Technologies instrument.

Results: The co-design process successfully translated community feedback into an empathy-informed, user-centered prototype and systematically identified design considerations that single-partner approaches overlook. TBI-specific design requirements emerged, including the need for linear content progression over branching navigation, higher technical performance standards, and explicit content signaling with clarity prioritized over novel interface design. Think-aloud protocols revealed that participants struggled with mobile navigation and branching structures but excelled with sequential content progression. In addition, the input from individuals with TBI, caregivers, clinicians, and researchers led to practical refinements such as shorter microlearning lessons (5 - 12 min), clearer voiceover tone, and simplified navigation, directly addressing the study's objective of improving accessibility and emotional resonance. Overall usability was high, measured using the Comprehensive Assessment of Usability for Learning Technologies (CAUSLT), with an average score of 4.25 out of 5 (SD 0.72; 95% CI 3.36 - 5.15; n=5). Knowledge accuracy was 80% (8/10 items; 95% CI 49% - 94%; n=5 participants; 2 items each), indicating that the system effectively supported learning and comprehension. Module completion was 100% (5/5; 95% CI 56.6% - 100%). Average time-on-task for 10 lesson completions was 11.47 (SD 5.28; range 4.6 - 21.42) minutes per lesson, demonstrating strong task efficiency and engagement. Highest ratings were observed in the pedagogical usability domain, reflecting that the interface was clear, intuitive, and conducive to learning. Collectively, these findings suggest that applying CBPR across all design stages produced a technically sound, easy-to-use, and pedagogically meaningful mHealth tool specifically tailored for individuals with TBI.

Conclusions: Sustained CBPR across full design and development cycles resulted in high usability for ePST for individuals with TBI. Ultimately, this study operationalized a full-cycle pipeline that links sustained community partnership to measured usability outcomes, producing community-informed design principles and a reproducible mixed methods approach for formative mHealth development for TBI.

(J Med Internet Res 2026;28:e83995) doi:10.2196/83995

XSL•FO
RenderX

KEYWORDS

traumatic brain injury; community-based participatory research; user-centered design; usability; mHealth; rehabilitation; problem-solving training

Introduction

Traumatic Brain Injury Rehabilitation

Traditional rehabilitation research underrepresents people with lived experience of disability, including traumatic brain injury (TBI), yielding interventions misaligned with patient contexts [1-4]. TBI has acute and chronic sequelae [5,6] affecting cognition, emotion, and social functioning [5-10] that adversely affect learning and care access [11,12]. When design ignores these constraints, relevance and engagement drop [13]. Although recent studies demonstrated the feasibility of participatory adaptations in TBI rehabilitation [14,15], such approaches remain rare [16]. Emerging protocols increasingly incorporate caregiver and community voices through community-based participatory research (CBPR) frameworks [17], yet broader adoption remains limited [18]. Indeed, although chronic challenges faced by individuals with TBI are increasingly recognized, rehabilitation research rarely translates this awareness into meaningful community engagement or integration of practitioner perspectives [19,20]. Provider- and institution-centered models continue to dominate, reinforcing inequities [21] and limiting collaboration between researchers, clinicians, and patients [22,23]. This results in less responsive interventions, lower user satisfaction, and reduced effectiveness [5,24-26]. Despite growing support for participatory approaches, provider-centric norms persist [27]. This study responds to those gaps by modeling a collaborative, community-informed design process [18,28].

Power sharing and collaborative decision-making are critical to designing effective, context-responsive interventions [28,29]. Rehabilitation requires real-world interaction, collaboration, and adaptability to individual needs [30]. Participatory approaches shift decision-making toward community members [31], having produced measurable improvements in health outcomes and patient-reported measures [32]. For example, ethnographic work by Manhas and colleagues [33] showed shared decision-making in rehabilitation enhances patient satisfaction, understanding, goal attainment, and self-reported outcomes. This contrasts with provider-driven models that limit patient involvement and flexibility. Indeed, challenges such as limited community engagement, asymmetrical decision-making, and provider-centered research can undermine the relevance and impact of TBI rehabilitation efforts [19,21-23]. These issues call for more inclusive approaches that prioritize symmetrical decision-making and meaningful collaboration with the TBI community [33].

This paper presents a case example of the formative design and evaluation of Electronic Problem-Solving Training (ePST), a metacognitive, evidence-based mobile health (mHealth) problem-solving intervention. ePST is based on PST, a cognitive-behavioral approach with proven efficacy for neurodevelopmental and psychological conditions that is grounded in some of the strongest evidence in cognitive rehabilitation [34,35]. PST and comparable approaches are widely used in psychology to improve problem-solving skills and mindset [36,37] and have shown promise for preventing and treating cognitive deficits [38] through numerous clinical trials [39]. Research suggests PST can be especially beneficial for long-term or multifaceted health issues, such as TBI [40,41]. A robust body of evidence shows that such problem-solving approaches lead to meaningful reductions in symptoms, strengthen individuals' confidence in managing their health, and enhance adherence to prescribed regimens [42-44]. ePST was developed using learning experience design and a CBPR framework to ensure accessibility, community-driven decision-making, and iterative co-design [18,45,46]. Learning experience design and CBPR guided front-end activities and the sociotechnical-pedagogical usability evaluation reported here.

Background and Rationale

Rehabilitation research often centers around clinician and designer perspectives over patient input, reducing relevance, effectiveness, and adaptability for individuals with TBI [47,48]. Correa and colleagues [49] showed that interventions lacking patient involvement can be misaligned with how patients perceive risks, benefits, and treatment goals, undermining recruitment and randomization. Such problems suggest a need for adaptive, patient-informed approaches, which CBPR can provide in a context-sensitive and ethical manner [46]. CBPR helps researchers understand lived experience and co-create interventions that are more relevant, acceptable, and effective. For example, Quilico and colleagues [28] partnered with people with TBI and caregivers to adapt a physical activity program, producing changes that improved relevance, outcomes, and engagement. Groussard and colleagues [50] involved users with lived TBI experience and caregivers in developing and evaluating a cognitive support system, yielding improved user satisfaction and greater autonomy. However, participation alone is insufficient. CBPR requires reciprocal relationships among community members, academics, and practice partners to draw on diverse strengths [51]. As a case-in-point, Springer and Skolarus [52] specifically distinguished between the "community-based" and "participatory" components of CBPR to clarify how all components of this approach are needed to promote sustained, power-sharing partnerships.

As CBPR is applied increasingly to digital health interventions like ePST, new design and evaluation demands emerge. For example, reporting in CBPR remains inconsistent, and implementation is uneven, especially in rehabilitation contexts [16]. Usability and contextual fit present persistent barriers to adoption in eHealth and mHealth, reinforcing the need for community-informed design and iterative testing cycles [53]. In addition, promoting sustained engagement remains a challenge [18,54], which supports the use of innovative pedagogical strategies such as microlearning, a design approach shown to improve engagement and learning outcomes in health applications when lessons are limited to a length of 5 minutes

XSL•FO
RenderX

to 12 minutes [55]. In parallel, sociotechnical frameworks have been recommended for evaluating patient-facing tools, supporting our integration of CBPR, learning experience design, and the sociotechnical-pedagogical framework (discussed in the next section) [56]. Collectively, these gaps suggest a need to balance TBI rehabilitation complexity with the provision of usable, accessible, and engaging interventions. We illustrate our approach to achieving this balance through the conceptual framework we present in the following section.

Conceptual Framework

We developed a conceptual model that places CBPR at the methodological core, pairs participatory mechanisms (Community Advisory Board or Community Engagement Studios) with learning experience design to convert community partner input into design principles, and maps these strands onto the Double Diamond (Figure 1) for iterative development and sociotechnical-pedagogical evaluation [57-61].

Figure 1. Conceptual framework integrating community-based participatory research, the Community Advisory Board (CAB), Community Engagement Studios (CES), and learning experience design (LXD) mapped onto the Double Diamond design framework for individuals with traumatic brain injury (TBI).

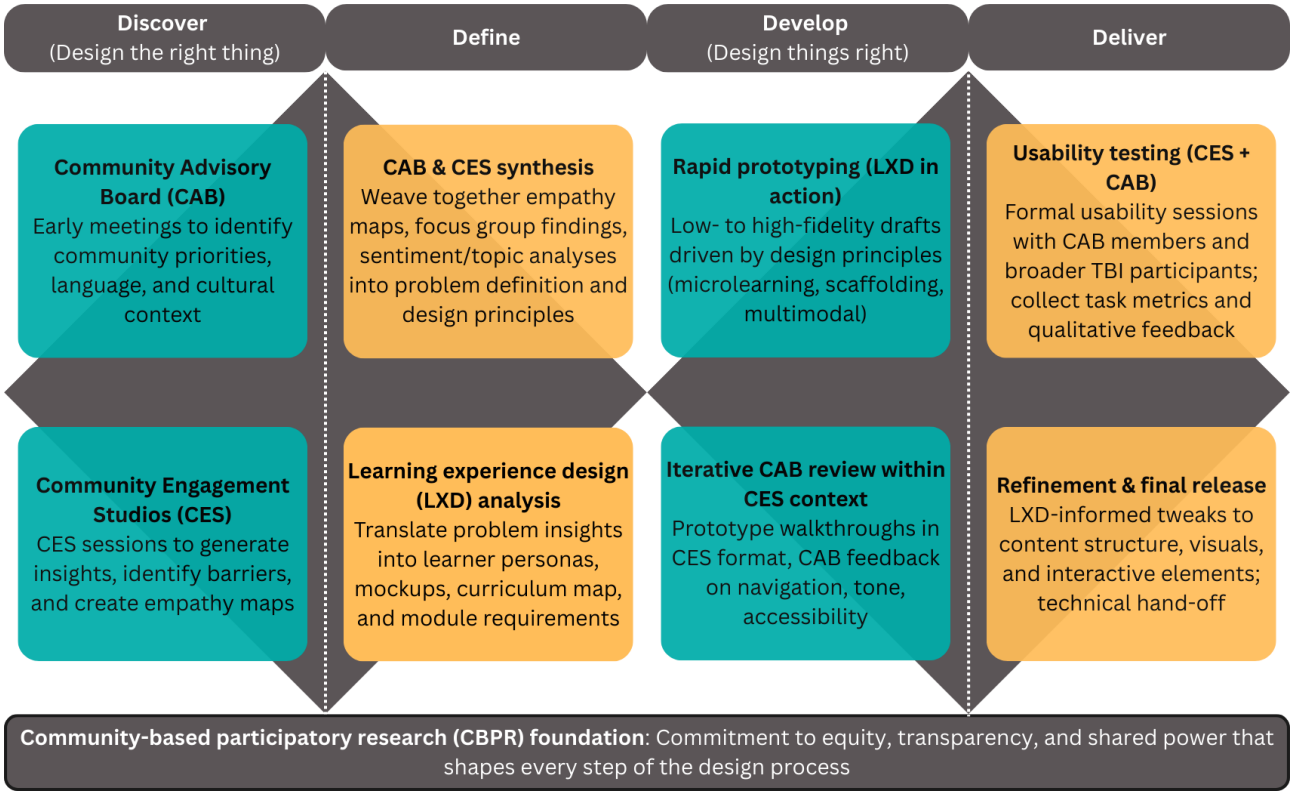

CBPR as the Foundational Ethos

We adopted CBPR as a foundational ethos to foster inclusive, patient-centered rehabilitation design and to translate community priorities into practice [62-65]. Central to CBPR are collaboration and balanced partnerships that share decision-making responsibility [66]. Unlike short-term, investigator-led studies, CBPR emphasizes long-term reciprocal relationships that promote ethical research practices and improved outcomes [67]. This is important because top-down, limited-duration studies can erode trust and exclude local needs, with standardized practices that do not accommodate community input tending to perpetuate these problems [68,69]. CBPR's emphasis on shared decision-making across all phases of the research process provides one avenue to address these problems [70]. Collaboration through structured partnerships allows community members to inform priorities, participate in knowledge creation, and strengthen the real-world applicability of interventions [31]. These approaches move research beyond expert-driven agendas by integrating the lived experiences, priorities, and contextual knowledge of community members into the design and implementation process [71].

Participatory Mechanisms

Community Advisory Board

Community Advisory Boards are structured, ongoing partnerships that integrate people with lived experience into research, providing authentic representation and culturally grounded input across the project lifecycle [72-75]. Unlike short-term focus groups, Community Advisory Boards meet regularly to co-develop research strategy, advise on ethics and context, and guide intervention refinement, fostering shared leadership, trust, and power sharing [76-79].

Community Engagement Studios

Community Engagement Studios are structured, facilitated consultations in which researchers obtain targeted feedback from panels of community experts, caregivers, and clinicians. Unlike advisory boards or focus groups, Community Engagement Studios use focused, iterative sessions to promote dialogue, reciprocal learning, and sustained community involvement [59,80]. Originating with the Meharry-Vanderbilt Community-Engaged Research Core [58], Community Engagement Studios were developed to overcome participation

XSL•FO
RenderX

barriers in clinical and rehabilitation research, including mistrust from historical unethical practices and social inequities [81]. By positioning community members as consultants and experts rather than passive subjects, Community Engagement Studios help identify barriers, adapt interventions to community needs, and build trust with underrepresented groups [82,83]. Community Engagement Studios can enhance cultural adaptation [84], increase minority participation [82,83], and reduce power imbalances between researchers and community members [85,86].

Operationalization via Learning Experience Design

Learning experience design is a learner-centered, theoretically grounded framework that integrates instructional design, cognitive science, user experience, and participatory approaches [87,88]. Learning experience design emphasizes designing engaging and inclusive learning environments that respond to learners' real-world needs and experience [89,90]. Learning experience design focuses on the cognitive, emotional, and perceptual influences of learner interactions with content, tools, and people across the learning process [91-93]. Learning experience design guided ePST's Double Diamond workflow. In *Discover*, empathy interviews identified core needs and constraints; in *Define*, those insights plus Community Advisory Board input shaped personas, module structure, and mock content [94,95]. To address TBI-specific cognitive limits (eg, memory, fatigue), the *Develop* phase adopted microlearning (ie, short, digestible lessons lasting 5 minutes to 12 minutes) intended to lower cognitive load, promote encoding, and support retention [96-99]. *Deliver* used iterative usability testing to validate designs and drive refinements. Multimodal strategies (text, visuals, voiceover, interactivity) and gamification (badges, progress indicators, interactive tasks) supported diverse preferences and motivation [100,101].

Sociotechnical-Pedagogical Framework

The sociotechnical-pedagogical framework conceptualizes learner experience as the alignment of 3 interdependent domains: technological, pedagogical, and sociocultural [102,103]. The technological domain covers reliability, accessibility, device compatibility, navigability, and error tolerance; the pedagogical domain covers alignment of objectives, materials, activities, and assessment, plus clarity, scaffolding, cognitive load management, and feedback quality; and the sociocultural domain addresses presence, identity, communication, cultural responsiveness, and scenario authenticity. The sociotechnical-pedagogical framework serves as both a design and evaluation lens, operationalized via dimension-specific heuristics validated against course evaluations that identified 195 distinct problems consolidated into nonoverlapping heuristics spanning the 3 domains [102]. This approach is critical in neurorehabilitation because traditional usability frameworks often miss interactions among cognitive, social, and technical factors [88]. For people with TBI, technological design must go beyond basic accessibility to reduce cognitive load (simplified interfaces, memory supports, fatigue accommodations) and ensure assistive-technology compatibility. Pedagogical design should address executive function limits via clear structure, predictable flows, compensatory strategies, repetition, and metacognitive scaffolds to support transfer. Sociocultural design must attend to stigma, identity shifts after injury, peer and family involvement, and social-context fit. The sociotechnical-pedagogical framework reveals problems that purely technical reviews miss.

Intervention Description

ePST is a cross-platform, community-informed mHealth intervention tailored to the cognitive and emotional needs of adults with TBI. Built on microlearning, it delivers short (5 - 12 minutes), chunked lessons with built-in progress tracking to reduce cognitive load. Engagement features include motivational messaging derived from empathy interviews; a virtual coach ("Ruth"); personalized learning pathways; embedded reminders; and gamified elements (badges, certificates) to support memory, reinforce learning, and sustain motivation (Figure 2). ePST is grounded in problem-solving training and operationalizes the 6-step ABCDEF mnemonic (Figure 3): A, assess the problem; B, brainstorm solutions; C, consider and choose; D, develop and do; E, evaluate; and F, flex. ePST translates these steps into scaffolded modules that teach structured decision-making and problem-solving strategies tailored to adults with TBI. A description of the ePST learning modules is provided in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 2. Representative screenshots from the Electronic Problem-Solving Training (ePST) prototype and final user interface, captured during usability testing with adults with traumatic brain injury: (1) progress tracker, (2) virtual coach "Ruth" interface, and (3) reminder/notification panel.

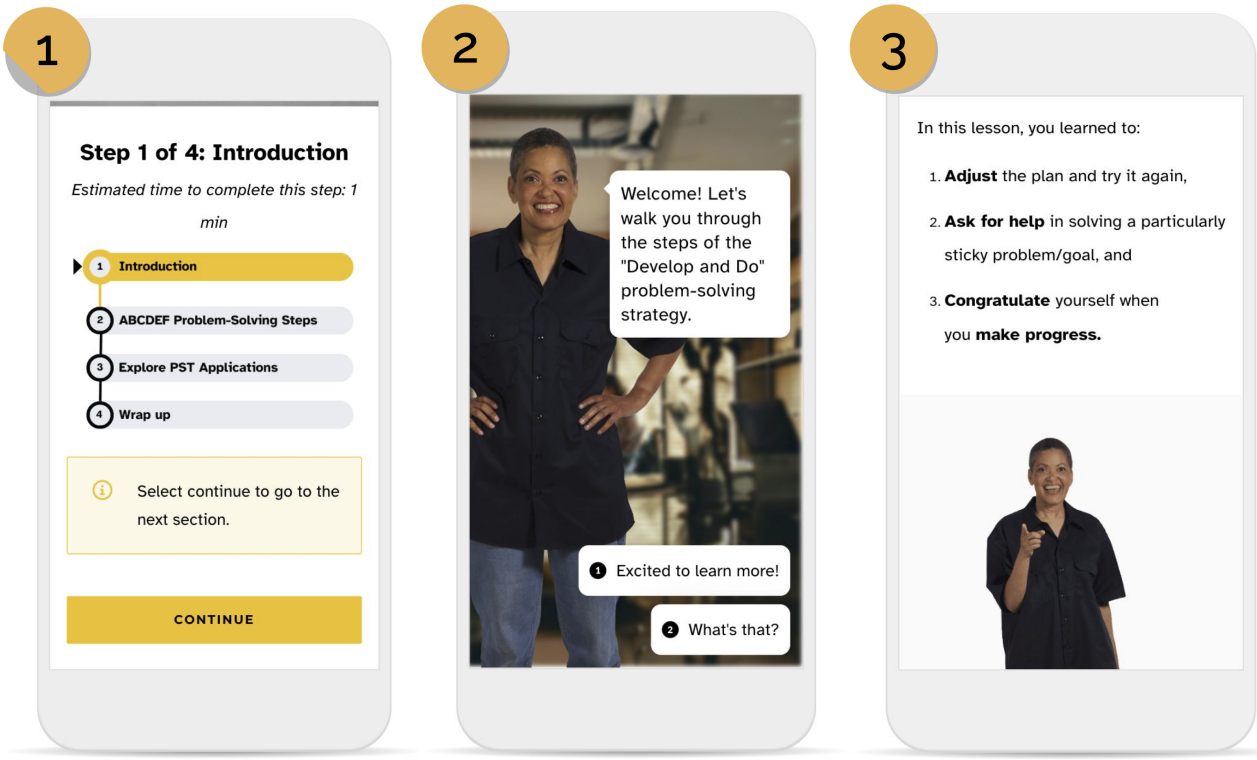

Figure 3. Problem-solving training strategy diagram showing the ABCDEF 6-step metacognitive process implemented in Electronic Problem-Solving Training (ePST): assess, brainstorm, consider and choose, develop and do, evaluate, and flex.

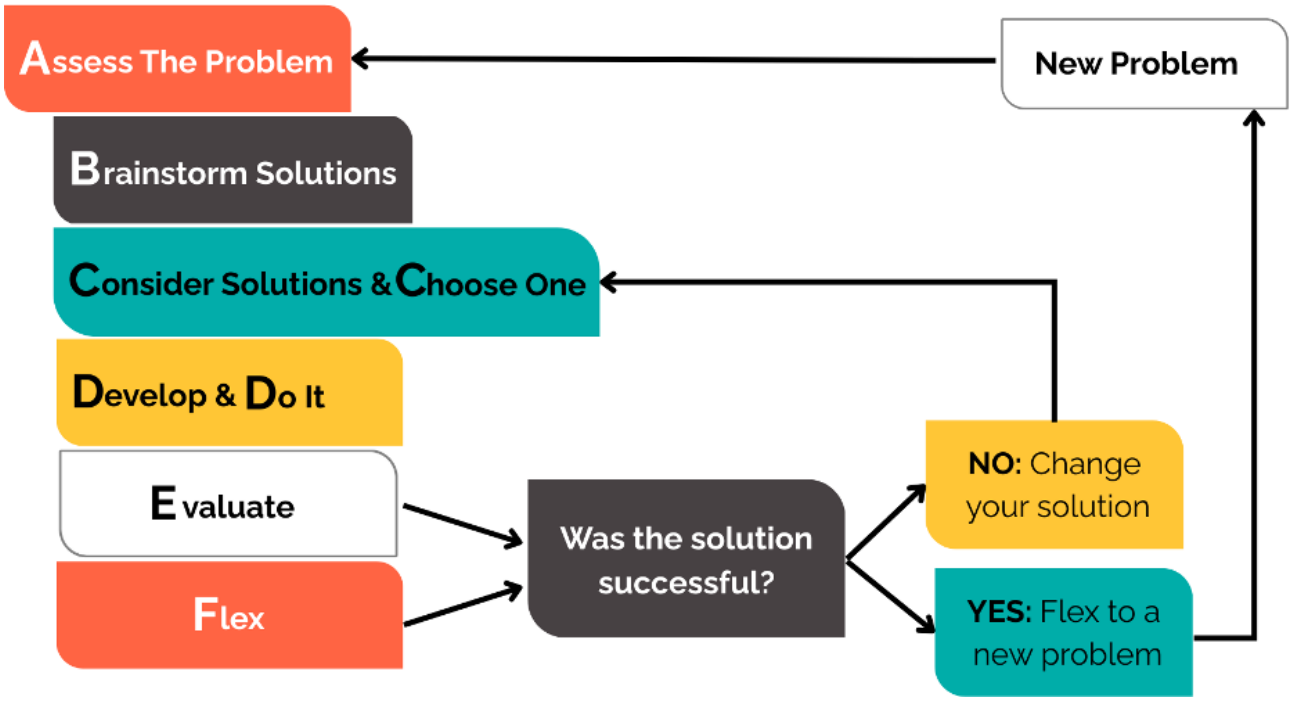

Purpose and Research Questions

The purpose of this iterative, multimethod formative design and evaluation case study was to use a CBPR approach to guide front-end design activities (ie, empathy interviews, empathy mapping, and persona development) and to evaluate the sociotechnical-pedagogical usability of the ePST intervention with TBI community members (ie, Community Advisory Board members, families, providers, and individuals with lived TBI experience) at a large public university and a large medical center in the southern United States. The questions that guided this research included: research question (RQ) 1: What themes related to learning needs, barriers, and preferences emerge from front-end design activities (empathy interviews, empathy mapping, and persona development) with TBI community members? RQ 2: How did individuals with TBI perceive the

XSL•FO
RenderX

sociocultural, technological, and pedagogical usability aspects of their experience with ePST during testing? RQ 3: How were identified sociotechnical-pedagogical usability issues addressed through design refinements?

Methods

Double Diamond Approach

This multimethod formative design and evaluation study followed the Double Diamond approach (Discover, Define, Develop, Deliver) and ran from February 2024 through July 2024. In *Discover*, we established the Community Advisory Board, drafted initial design principles, and conducted Community Engagement Studio empathy interviews with people with TBI. In *Define*, we developed learner personas, produced a curriculum map, and iteratively refined priorities via Community Engagement Studios and Community Advisory Board reviews. In *Develop*, we translated principles into low- to high-fidelity prototypes and internal subject matter expert review. In *Deliver*, we conducted iterative usability testing with people with lived TBI experience and implemented refinements after each round. We reported patient and public involvement using the GRIPP2 Short Form (GRIPP2-SF) [104]. A 1-page mapping table linking GRIPP2-SF items to manuscript locations is provided in Multimedia Appendix 1.

Participants

Community Advisory Board Participants

The Community Advisory Board (n=33) was purposively assembled to include people with lived TBI experience, caregivers, clinicians, researchers, industry representatives, advocates, and members of minoritized groups. Members were identified via professional networks, partner clinics, and community organizations and invited by email. Selection criteria were TBI or digital health expertise, lived experience, demographic diversity, and advocacy and service representation. Community Advisory Board members received US $25 per meeting. Community Advisory Board composition is provided in Multimedia Appendix 1.

Empathy Interview Participants

Empathy interview participants (n=14) were recruited via clinician referral and community outreach at a large tertiary rehabilitation center in the southern United States in February 2024 and March 2024. Inclusion criteria were age ≥18 years, proficiency in English, and either (1) documented TBI confirmed by clinician referral or review of medical records when available or (2) self-reported TBI with screening confirmation of capacity to participate. Exclusion criteria were severe communication impairments or acute medical instability that precluded informed consent or participation. Caregivers and providers were eligible if they provided regular care or clinical services to adults with TBI.

Usability Testing Participants

Usability testing participants (n=5) were recruited purposively from the same clinical and community sources in July 2024 to capture variation in technology experience and time since injury. Inclusion criteria included age ≥18 years, English fluency, history of TBI (clinician referral or medical record when available), ability to use a smartphone or computer without assistance, and capacity to provide informed consent and follow study tasks. Exclusion criteria included acute medical or psychiatric instability and severe receptive or expressive communication impairments that prevented participation. Five participants is standard for early-stage formative usability tests, with 80% of usability problems identified via small samples [105,106]. This low number limits statistical generalizability but is conventional in heuristic-based usability work intended for problem identification [107,108]. This approach has substantial precedent in digital health formative studies that use small, purposive usability samples to drive iterative refinements [109-111]. To increase rigor and reduce bias from the small sample, we purposively sampled, triangulated findings, and applied an iterative refine-and-retest logic.

Ethical Considerations

The study protocol was approved by the Human Research Protection Program at the University of Georgia (IRB #00009943) on June 17, 2024, and was deemed exempt. Written informed consent was obtained electronically via Qualtrics. Participants reviewed the full consent document, typed their full name and the date to indicate agreement, and submitted the consent form. The consent form covered study purpose and procedures, audio and video recording, foreseeable risks and benefits, right to withdraw, compensation, and data handling. All study data were de-identified and stored on encrypted University of Georgia servers with access limited to authorized study personnel. The master linking list and raw recordings will be destroyed at study end; de-identified data may be used for future research but will not be deposited in a public repository. Participants received US $25 per activity, and Community Advisory Board members were paid US $25 per meeting, with payments issued via Clincard after each session. Payments were institutional review board–approved and described in the consent forms. No identifying information was included in this paper or multimedia appendices.

Procedures

Given the iterative nature of the Double Diamond approach, analysis was multimodal and occurred across all phases of design, with analysis falling into 3 broad categories: (1) qualitative, (2) quantitative, and (3) computational. No data were missing for any of the reported analyses.

Discover Phase Procedures

Establishment of the Community Advisory Board

Community Advisory Board members were recruited purposively [112] from the community, academia, industry, and medical-related institutes based on 4 criteria: (1) professional expertise in TBI rehabilitation, assistive technology, and or digital health; (2) lived experience with TBI; (3) demographic diversity across age, gender, socioeconomic status, and geography; and (4) community member representation including those with lived experience, clinicians, researchers, technology developers, and advocacy organizations. Quarterly Community Advisory Board meetings were held across Phases 1 - 3 (total n=16).

XSL•FO
RenderX

Establishment of Preliminary Design Principles

We drew on findings from a prior study, *Caregivers in Dementia PST and DSJ* (CaDeS), which tested coach-delivered PST [113]. Open-ended responses to overall intervention satisfaction were analyzed using machine learning techniques, including sentiment analysis and latent Dirichlet allocation, to generate an initial set of 7 design principles for ePST, which were reviewed and refined in a subsequent Community Engagement Studio session.

Empathy Interviews

Empathy interviews were conducted with 3 groups of participants. Groups 1 (n=6) and 2 (n=3) consisted of TBI survivors. Group 3 consisted of care partners and providers (n=5). Interviews were guided by the 4-phase empathy framework from Kouprie and Visser [114]. All interviews were approximately 75 minutes and conducted online using Zoom web conferencing software. Questions focused on (1) learning challenges, (2) effective therapies, (3) the impact of others' stories, (4) group-specific challenges, (5) building trust through shared expertise, and (6) motivational messages. Interviews were recorded and transcribed using Zoom.

Define Phase Procedures

Empathy Mapping

Empathy mapping guided learner analysis and informed the design of ePST [115]. Empathy mapping involved synthesizing participants' responses into 4 core domains ("Says," "Thinks," "Does," and "Feels") to foster understanding of their motivations, challenges, and learning preferences. This allowed capture of nuanced information about participants' cognitive, emotional, and behavioral experiences. A total of 9 empathy maps were created (see Multimedia Appendix 1). These were then used to generate learner personas and referenced to inform design.

Persona Development

Personas are fictional, data-informed archetypes that represent individuals within the target population [116]. Our personas provided summaries of representative descriptors based on information that was synthesized from empathy maps. Personas were presented to the Community Advisory Board, reviewed, and revised. Initial designs included TBI severity; however, this was removed at the recommendation of the Community Advisory Board, as severity was an inadequate method to represent nuanced TBI characteristics, especially chronically. The final set of personas (n=10) is provided in Multimedia Appendix 1.

Refinement of Design Principles

Design principles were refined based on a structured empathy interview with 5 caregivers and providers. Analysis comprised a discussion-based analytic process to identify key insights from the transcripts. The design principles were then reviewed in a Community Engagement Studio session with the Community Advisory Board, who provided feedback on clarity, relevance, and completeness. Analysis did not focus on achieving saturation but instead prioritized triangulation across data sources and methods for development of design principles.

Develop Phase Procedures

Community Engagement Studios

Structured Community Engagement Studio sessions were used to elicit structured feedback during Community Advisory Board meetings. Community Engagement Studio sessions (n=10) focused on usability challenges, content clarity, and delivery preferences. Community Engagement Studio sessions were between 60 minutes and 90 minutes, included 6 to 8 participants, and followed a structured protocol. A trained moderator guided discussion. Discussion foci varied depending on which design artifacts were being reviewed, Participants reflected on design artifacts' clarity, relevance, and usability. Sessions were conducted, audio recorded, and transcribed using Zoom. Transcripts and notes were then synthesized into actionable design recommendations.

Rapid Prototyping

Rapid prototyping is an iterative design approach that quickly develops and refines working models based on user feedback [117]. This approach was used to transform insights from the *Define* phase into working prototypes. Initial design concepts were explored through low-fidelity mockups then iteratively refined into medium- and high-fidelity prototypes, with emphasis on flexibility and responsiveness to user input [117]. Designs were regularly reviewed during Community Engagement Studios for issues such as navigation, language complexity, and content pacing.

Deliver Phase Procedures

Usability tests (n=5) were conducted by a trained graduate student and a university professor usability expert. Testing followed a semistructured, task-based research protocol. Sessions were between 60 minutes and 75 minutes and were conducted, recorded, and transcribed using Zoom. Participants completed 5 structured usability tasks per session while thinking aloud and sharing their screens. Tasks assessed both technological usability (eg, navigation, multimedia interaction) and pedagogical usability (eg, clarity of content, microlearning structure). Participants then completed the Comprehensive Assessment of Usability for Learning Technologies (CAUSLT) instrument [118]. Data were analyzed using an integrated approach that combined observational, survey, and efficiency metrics. Think-aloud transcripts and observer notes were reviewed and discussed by two team members to identify barriers. Responses to the CAUSLT instrument were summarized using descriptive statistics and disaggregated across the 3 instrument factors. Design flaws were prioritized using Nielsen's severity scale [119]. Efficiency data were extracted from session recordings. Findings were documented in a report that was reviewed with the Community Advisory Board, whose feedback guided refinements in areas such as voiceover quality, mobile navigation, and content clarity for cognitive accessibility.

XSL•FO
RenderX

Results

Discover Phase Results

Composition of Community Advisory Board

The composition of the Community Advisory Board is presented in Table 1. Some individuals were represented in more than one group, as some Community Advisory Board members self-identified with more than one category.

Table . Composition of the Community Advisory Board (n=33) including counts and role categories for members recruited purposively from clinical partners, community organizations, academic networks, industry, and advocacy groups.

Category	Composition	Total representatives, n
Academic researchers	PhD researchers (n=12), psychologists (n=8), educational technology experts (n=2), graduate students (n=2)	24
Industry professionals	Software developer (n=1), software designers (n=5)	6
Individuals with lived experience and advocates	Individuals with TBI[a] (n=6), TBI care partners (n=4), disability advocates (n=6)	16
Rehabilitation and clinical professionals	Occupational therapists (n=4), social workers (n=3), rehabilitation counselors (n=3)	10
Individuals with physical disabilities	Blind (n=1), deaf (n=1)	2
Individuals from minoritized groups	LGBTQAI+[b] (n=4), minoritized racial and ethnic groups (n=6)	10

[a]TBI: traumatic brain injury.

[b]LGBTQAI+: lesbian, gay, bisexual, transgender, queer (or questioning), asexual (or allied), intersex, plus

Preliminary Design Principles

Preliminary design principles were established based on results of prior research and Community Advisory Board input. Principles emphasized accessibility, emotional resonance, clarity of messaging, and personalization, serving as our foundation for early prototypes, visual design, and engagement strategies. These preliminary principles were later expanded and structured into a comprehensive hierarchy, reported in the Define Phase Results section.

Empathy Interview Participant Demographics

We recruited 14 participants (Table 2) for empathy interviews, including individuals with TBI (n=9) and caregivers and providers (n=5).

Table . Participant demographics for empathy interviews (n=14), including de-identified breakdown by participant group, race or ethnicity, age bands, and gender.

Characteristics	Individuals with TBI[a], n	Caregivers and providers, n
Race		
Hispanic	1	1
Caucasian or White	5	3
African American	2	0
Asian	0	1
Age (years)		
30-39	4	0
40-49	3	1
50-59	1	0
60-69	0	3
≥70	0	1
Gender		
Female	7	5
Male	2	0

[a]TBI: traumatic brain injury.

XSL•FO
RenderX

Define Phase Results

Empathy Maps

Empathy maps were created (n=9), with each map including brief descriptors in the categories "Says," "Thinks," "Feels," and "Does." Analysis revealed 4 key themes characterizing the post-TBI experience: Participants experienced (1) frustration and disorientation with everyday tasks, (2) loss of self-identity and nostalgia for pre-injury life, (3) physical exhaustion from therapy that decreased motivation, and (4) social isolation due to perceived lack of family understanding. In response, participants developed adaptive strategies including structured skill relearning through rehabilitation and memory aids such as sticky notes. The full set of empathy maps is provided in Multimedia Appendix 1.

Personas

A set of personas (n=10) was created to guide design. Personas highlighted varied life contexts, recovery journeys, and learning needs across individuals such as veterans, students, professionals, and retirees. Each reflected unique combinations of cognitive, emotional, and physical challenges, along with personal goals like regaining independence, improving memory, or reducing stigma. Common facilitators included family support, adaptive tools, storytelling, and professional guidance. Despite varied barriers ranging from aphasia to fatigue to discrimination, all personas demonstrated resilience and motivation to recover. The complete set of personas is provided in Multimedia Appendix 1.

Refined Design Principles

A refined set of design principles was created in the *Define* phase, incorporating the preliminary set created during the *Discover* phase. Using a framework proposed by Kali [120], the design team organized these insights into a 3-tiered hierarchy (specific, pragmatic, and metaprinciples). Pragmatic principles reflected actionable guidance relevant to the learning design. These pragmatic principles were grouped into 6 broader metaprinciples, such as accessibility, emotional support, motivation, personalization, cultural relevance, and evidence-based action. Where applicable, specific principles (eg, interface features, content structures) were also identified to illustrate how the pragmatic principles would translate into concrete design decisions (Table 3).

XSL•FO
RenderX

Table . Design principles for Electronic Problem-Solving Training (ePST) module development, including metaprinciples, pragmatic principles, and specific principles derived from empathy interviews, empathy maps, Community Advisory Board and Community Engagement Studio feedback, and persona development.

Metaprinciple	Pragmatic principle	Specific principles
1. Ensure accessibility and usability	Design for cognitive and physical inclusion	Use clear, concise, jargon-free language; include closed captioning; support mobile-first navigation; design intuitive interaction patterns
1. Ensure accessibility and usability	Support memory and comprehension	Reinforce key concepts with reminders and visual anchors; use chunked content and repeated exposure
1. Ensure accessibility and usability	Allow flexible engagement	Enable learners to proceed at their own pace; allow pausing and resuming lessons easily
2. Support emotional and behavioral needs	Encourage emotional regulation	Include calming activities (eg, music, mindfulness cues); normalize behavioral variability in content
2. Support emotional and behavioral needs	Empathize with behavioral and communication challenges	Acknowledge and adapt for speech and behavioral limitations; use neutral, nonjudgmental tone
2. Support emotional and behavioral needs	Promote self-awareness and acceptance	Include prompts or reflection activities to build insight into strengths and limitations
3. Foster motivation and engagement	Use positive reinforcement	Integrate badges, rewards, and affirming feedback
3. Foster motivation and engagement	Emphasize goal setting and achievement	Provide explicit opportunities to set and track goals
3. Foster motivation and engagement	Provide regular feedback	Visual progress indicators; summary pages at lesson or module completion
4. Enable personalized and multimodal learning	Use varied sensory inputs	Combine visuals, audio narration, and interactivity
4. Enable personalized and multimodal learning	Allow for autonomy and independence	Design lessons that can be completed without facilitator support; scaffold progressively to reduce reliance on help
4. Enable personalized and multimodal learning	Tailor content for diverse learners	Include customizable avatars or pathways; vary representation and examples by demographic relevance
5. Establish credibility and cultural relevance	Include lived experience	Use testimonials from TBI[a] survivors and care partners; embed quotes and real-world scenarios
5. Establish credibility and cultural relevance	Partner with trusted organizations	Reference TIRR[b], advocacy groups, and clinical partners in content
5. Establish credibility and cultural relevance	Practice inclusive and representative design	Include diverse racial, ethnic, and gender identities; adapt content for veterans and other priority subgroups
6. Ground content in evidence and action	Communicate evidence accessibly	Present supporting research in simplified language or visuals; avoid academic jargon
6. Ground content in evidence and action	Use motivating calls to action	End modules with clear next steps (eg, “Enroll,” “Learn more”); include clickable links or guided follow-ups

[a]TBI: traumatic brain injury.

[b]TIRR:The Institute for Rehabilitation and Research.

Develop Phase Results

During the *Develop* phase, design artifacts progressed from low-fidelity storyboards to high-fidelity interactive prototypes (Figure 4). Low-fidelity mockups were iteratively refined into functional prototypes via structured Community Advisory Board feedback focused on usability, content clarity, accessibility, and delivery preferences. Key outputs included finalized lesson content, assessments, a cohesive visual design system, and functional prototypes. Community Engagement Studio sessions generated actionable recommendations that were synthesized into successive prototype iterations.

XSL•FO
RenderX

Figure 4. Evolution of selected Electronic Problem-Solving Training (ePST) design elements from prototypes to final product with panels illustrating iterative, prioritized changes driven by Community Advisory Board, Community Engagement Studios, and usability feedback (eg, badge redesign, microlearning length, voiceover tone, navigation simplification).

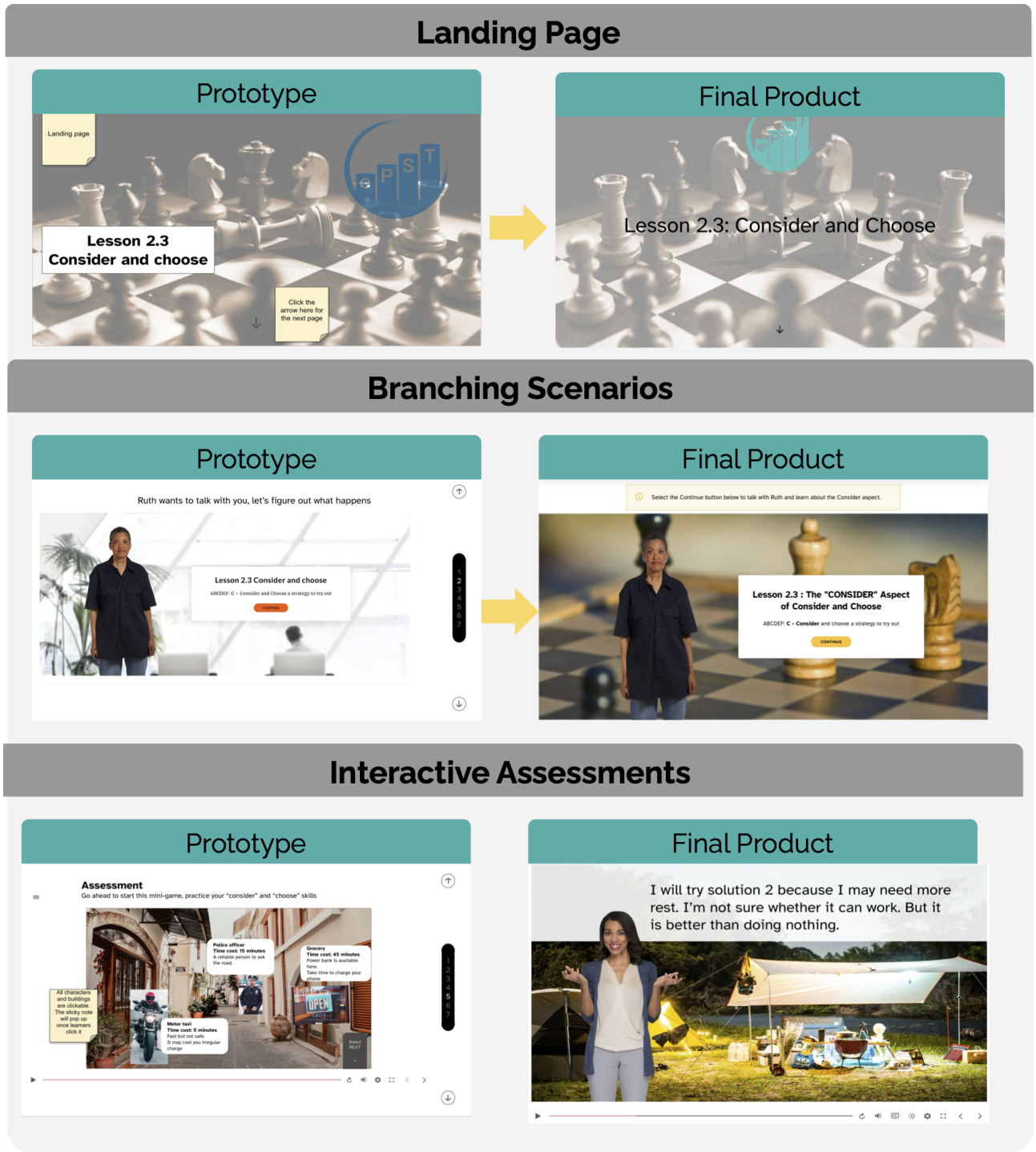

Deliver Phase Results

Deliver Phase Participant Demographics

Participant demographics for the usability study are presented in Table 4.

XSL•FO
RenderX

Table . Usability study participant demographics (n=5) including individual-level characteristics of age, gender, race or ethnicity, education, employment, years since injury, and baseline technology experience.

Participant[a]	Age (years)	Gender	Race or ethnicity	Education	Employment status	Time since injury (years)	Technology experience
Leo	47	Male	Hispanic	High school	Permanent disability	17	Capable user, no eHealth experience
Morgan	31	Female	White	Some college	Stay-at-home spouse	11	Capable user, occasional eHealth use
Alexis	50	Female	Black	Bachelor's degree	Permanent disability	20	Experienced user, frequent eHealth use
Riley	47	Male	White	Some college	Permanent disability	14	Experienced user, occasional eHealth use
Emma	36	Female	White	Some college	Stay-at-home parent	13	Experienced user, occasional eHealth use

[a]Participant names are pseudonyms.

Performance Metrics

All participants (n=5) finished every module (95% CI 56.6% - 100%). Lessons were completed efficiently, with participants spending about an average of 11.5 (SD 5.3; range 4.6 - 21.4) minutes for 10 lesson completions. Knowledge checks showed solid comprehension (8/10 items correct; 95% CI 49% - 94%; n=5, 2 items each), meeting our objectives for task efficiency and learning support. Performance metrics are summarized in Table 5.

Table . Usability performance results and knowledge assessment, including task and efficiency measures (lesson completion time, tasks per lesson, task completion rate) and knowledge-item accuracy derived from recorded usability sessions (n=5).

Metric	Result
Efficiency measures	
Lesson completion time (minutes), mean (SD)	11.47 (5.28)
Completion time - Module 2 (minutes), mean	10.50
Completion time - Module 3 (minutes), mean	13.10
Time (minutes), range	4.6 - 21.42
Tasks per user, mean	22.8
Tasks per lesson, mean	11.4
Task completion rate (tasks per minute), mean	0.996
Knowledge assessment	
Overall accuracy (% correct)	80
Question 1 accuracy (% correct)	60
Question 2 accuracy (% correct)	100

CAUSLT Usability Assessment

Participants completed the CAUSLT, which evaluates 3 dimensions of usability in educational technology using a 5-point Likert scale (1=Strongly Disagree, 5=Strongly Agree). Overall usability was high on the CAUSLT, with a mean score of 4.25 out of 5 (SD 0.72; 95% CI 3.36 - 5.15; n=5), supporting our objective that the prototype be easy to use and learn. Results are presented in Table 6 and illustrated in Figures 5-7 .

XSL•FO
RenderX

Table . Usability scores by sociotechnical-pedagogical domain for technological, pedagogical, and sociocultural usability as measured using the Comprehensive Assessment of Usability for Learning Technologies (CAUSLT).

Usability dimension	Score, mean (SD)	Score, range
Technological usability	4.06 (0.95)	3.50-5.00
Pedagogical usability	4.34 (0.77)	3.43-5.00
Sociocultural usability	4.13 (0.87)	3.00-5.00

Figure 5. Technological usability responses collected during usability testing using the Comprehensive Assessment of Usability for Learning Technologies (CAUSLT), showing domain and item-level means and SDs for the technological domain (navigation, performance, error tolerance).

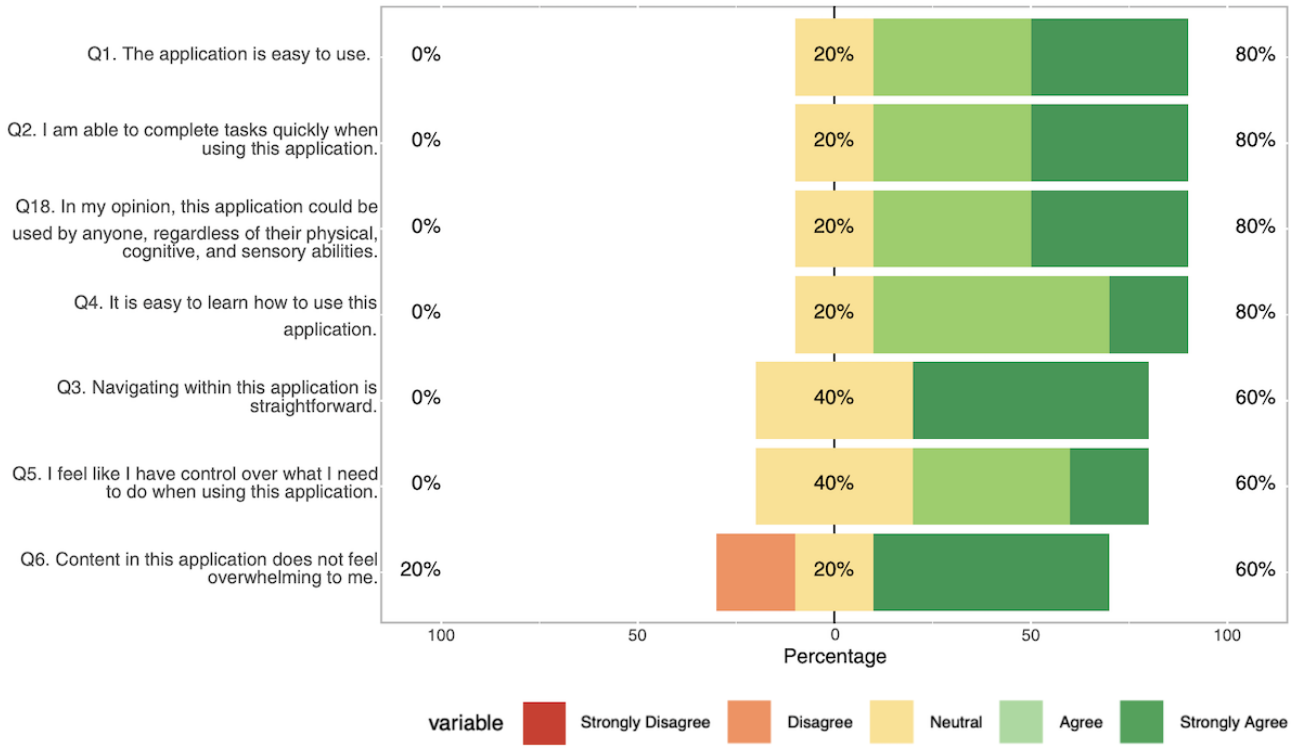

XSL•FO
RenderX

Figure 6. Pedagogical usability responses collected during usability testing using the Comprehensive Assessment of Usability for Learning Technologies (CAUSLT), showing domain and item-level means and SDs for pedagogical measures (ease of learning, clarity, learning support, engagement).

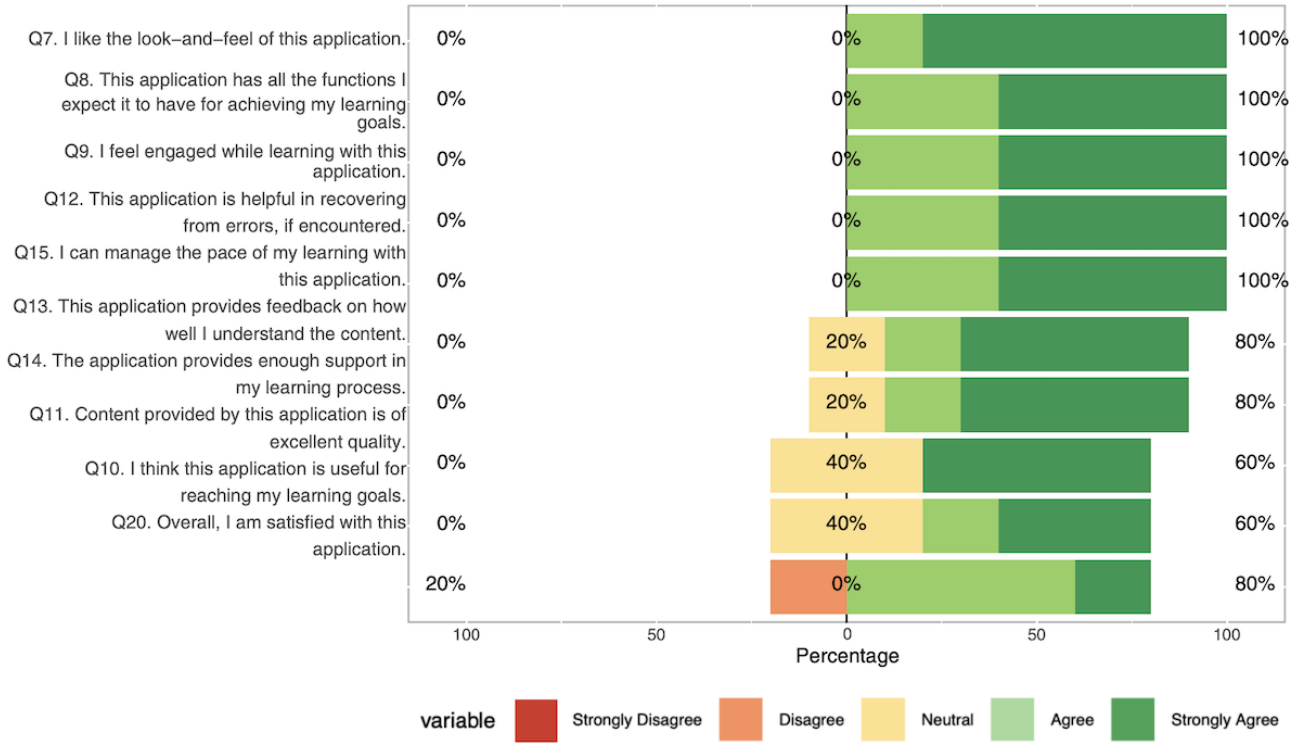

Figure 7. Sociocultural usability responses collected during usability testing using the Comprehensive Assessment of Usability for Learning Technologies (CAUSLT), showing domain means and SDs for sociocultural presence, accessibility, and relevance.

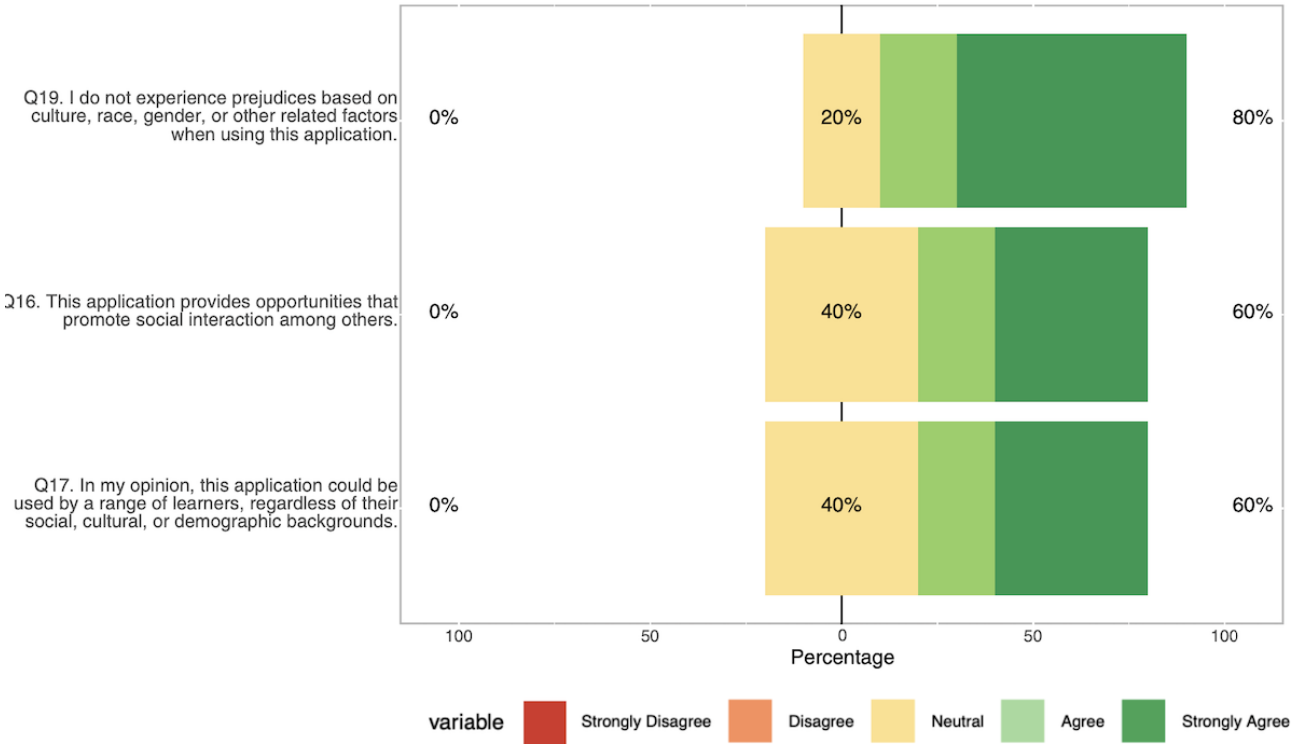

Pedagogical usability received the highest ratings (mean 4.34), with participants particularly valuing the application's look and feel (mean 4.8) and core learning functions including engagement, error recovery, and pace management (mean 4.6). Of the participants, 100% (5/5) agreed or strongly agreed on items related to learning engagement, pacing, and functional adequacy, with only feedback-related items showing some neutral responses (1/5, 20%). Technological usability scored well overall (mean 4.06), with strongest ratings for ease of use, task completion speed, and accessibility across different abilities. We found 80% (4/5) agreement across most technological items, though navigation and user control received slightly lower ratings (3/5, 60% agreement), and content overwhelm was the only item receiving any disagreement (1/5, 20%). Sociocultural

XSL•FO
RenderX

usability (mean 4.13) showed more variability. Participants were most confident about avoiding cultural prejudices (4/5, 80% agreement), while social interaction opportunities and cross-demographic accessibility both received 60% (3/5) agreement, with higher levels of neutrality (2/5, 40%), indicating potential areas for enhancement.

Key Usability Findings

Qualitative analysis of think-aloud transcripts and observation notes revealed both significant strengths and areas requiring refinement in ePST's usability. Content analysis identified patterns across participants' experiences that highlight the application's effectiveness at engaging users with TBI while revealing specific technical and interface challenges that impact user experience (Table 7). These findings provided actionable insights for iterative design improvements, which were incorporated between each usability testing session.

Table . Usability strengths and priority areas for improvement, summarizing positively rated features and recurrent problems identified from think-aloud protocols, observer notes, Comprehensive Assessment of Usability for Learning Technologies (CAUSLT) responses, and Community Advisory Board and Community Engagement Studio review.

Finding	Description
Usability strengths	
Intuitive interface design	Participants navigated the application easily and found interactive elements engaging. One participant described the storyline object as "pretty cool," indicating positive reception of multimedia components.
Effective progress tracking	Badge system and progress indicators were clearly understood and valued by users. Representative quotes: "It looks like I've received one badge 1,2,3, and five more to go" and "It proves to me that I've done something."
Engaging multimedia elements	Varied voiceover tones, storytelling approach, and visual design elements received positive feedback. Participants appreciated the narrative-based learning style and accessibility features.
Successful error recovery	Adaptive feedback mechanisms enabled users to recover from errors in knowledge checks without significant frustration, maintaining learning continuity.
Areas for improvement	
Mobile navigation issues	Users experienced confusion with mobile interface controls and activity progression. Representative quotes: "I see a button on the bottom right that looks like a back arrow" and "Wait, where was 3.2?"
Content comprehension	Some participants misunderstood instructions or content elements. One participant stated "Wait, this is not a question" when encountering a Storyline component.
Technical performance	Loading delays and playback issues disrupted user experience. Representative quotes: "Oh, it has to load all over again" and "I can't see the whole screen" (mobile display problems).

Discussion

Principal Findings

We applied a CBPR and learning experience design–guided formative design process to develop and evaluate ePST and addressed 3 core questions about front-end needs, sociotechnical-pedagogical usability, and how identified issues were resolved through design refinements. Usability was high across all domains, knowledge accuracy was 80% (an encouraging result for formative testing suggesting acceptable immediate comprehension), and mean time-on-task was 11.47 minutes per lesson while engaging in the think-aloud protocol. Participatory activities produced concrete design changes (ie, microlearning 5-12–minute lessons, badge refinements, voiceover adjustments), helped identify partner-specific priorities (ie, caregiver, clinician, lived experience perspectives), and revealed TBI-specific requirements (ie, linear progression, higher technical performance, explicit content signaling).

Taken together, these findings suggest that sustained community engagement can yield measurable usability improvements and actionable implementation guidance for TBI mHealth interventions. These outcomes map directly to established digital health usability constructs of effectiveness, efficiency, and satisfaction (ISO 9241‑11) [121] and to mHealth-specific evaluation guidance such as the validated mHealth App Usability Questionnaire [122]. Our combined questionnaire plus think-aloud pipeline also follows human factors and usability engineering recommendations for medical and mHealth systems (IEC 62366) [123,124] and recent mHealth usability reviews [125,126].

Iterative Community Feedback Enhanced Technical Usability

The Community Advisory Board structure (33 diverse community partners) and structured Community Engagement Studio sessions (n=10) enabled systematic integration of community input across development phases. CAUSLT scores averaged 87.3 out of 100, with pedagogical usability receiving

XSL•FO
RenderX

the highest ratings. Participants completed lessons efficiently and achieved 80% accuracy on knowledge assessments, comparing favorably to cognitive rehabilitation intervention outcomes reported in systematic reviews [127]. Empathy interviews with TBI survivors revealed specific cognitive load concerns that directly informed the microlearning approach (5 - 12–minute lessons) and progress tracking features. Community Advisory Board feedback on early prototypes resulted in modification of the badge system design and influenced voiceover tone selection to reduce perceived condescension. These modifications were fundamental design decisions that addressed cognitive accessibility requirements identified through community input [128]. Importantly, usability issues identified through think-aloud protocols mapped directly to areas where Community Advisory Board input had been limited or where technical constraints overrode community recommendations, suggesting that user involvement depth correlates with usability outcomes [129].

Multistakeholder Representation Identified Comprehensive Design Requirements

The Community Advisory Board's composition systematically identified design considerations that single partner approaches typically overlook. Caregivers identified family involvement features, while clinicians contributed evidence-based content validation, and individuals with lived experience prioritized autonomy and stigma reduction elements. This multiperspective input directly shaped the sociocultural usability features that scored highly in evaluation, particularly around cultural responsiveness and inclusive design [130]. Unlike traditional focus groups or surveys, the sustained Community Advisory Board engagement spanning the entire development cycle allowed for iterative refinement based on evolving understanding of user needs. This depth of engagement appeared to contribute to high pedagogical usability scores and enabled authentic relationship-building rather than extractive consultation [72].

TBI-Specific Technology Design Requirements Emerged

The usability evaluation revealed specific design requirements for cognitive rehabilitation technology that extend beyond general accessibility guidelines. Analysis of user interactions demonstrated that traditional e-learning design principles require significant adaptation for users with cognitive impairments, consistent with cognitive load theory applications in special populations [131]. The 21-minute range in task completion times (range 4.6 - 21.42 min) revealed that cognitive processing variability in TBI populations requires deliberate architectural choices rather than standard responsive design. Participants performed optimally with linear content progression and struggled with branching navigation structures, suggesting that linear content progression may reduce cognitive demands relative to complex navigation structures for users with executive function deficits [127].

Navigation issues identified in think-aloud protocols were predominantly mobile-specific, with participants reporting confusion about interface cues ("I see a button on the bottom right that looks like a back arrow") and progression sequences. Technical performance issues disproportionately disrupted learning flow for participants with attention deficits, suggesting that cognitive rehabilitation technology requires higher technical performance standards than typical educational applications [132]. Although participants appreciated multimedia elements and voiceover variety, content comprehension issues arose when instructional clarity was sacrificed for engagement, suggesting that TBI rehabilitation technology might require explicit signaling of content types and interaction expectations, with clarity taking precedence over novel interface design [133].

Methodological Contributions

This study contributes methodological insights for implementing CBPR in rehabilitation technology development. The integration of Community Advisory Board and Community Engagement Studio structures with learning experience design principles demonstrates how participatory research can move beyond consultation to systematic co-design. The sustained engagement model (33 diverse partners across the entire development cycle) provides a replicable framework for authentic community involvement, providing an actionable alternative to extractive research practices. The mapping of community input to specific design modifications illustrates how participatory methods can produce measurable technical improvements, not merely ethical and accessibility compliance. Findings support the claim that CBPR's value extends beyond moral imperatives to offer practical advantages in rehabilitation technology effectiveness.

The literature consistently supports that technology development through iterative user-centered design is associated with higher adherence and lower abandonment [134-139]. This more frequent and consistent engagement leads to clinical benefits [134,140]. Additionally, high usability facilitates scale-up and sustainability [135]. Despite this, development of digital health care technologies often fails to include patient, client, and clinician voices through early and ongoing user-engagement [134,141,142]. A recent scoping review [139] on reasons for abandonment of behavioral and mental health mobile interventions found 6 categories of reasons for abandonment, 3 of which could be directly addressed through user-centered and participatory design: (1) poor user experience, (2) evolving user needs and goals, and (3) content and features.

There is a growing body of literature specifically in rehabilitation supporting that usability, acceptability, and user-centered design contribute to implementation and sustainability of remote, technology-support interventions [28,143-146], but substantial work still needs to be done. A systematic review of cognitive rehabilitation interventions for older adults found that usability and user experience often explained mixed effectiveness of these technology-based interventions [147]. Though an even smaller body of research, a few studies have examined user-centered design for assistive technology and cognitive rehabilitation interventions for people with TBI [28,143-146]. These papers, consistent with our own findings, emphasized the importance of (1) tailoring the technology to reduce cognitive load; (2) having high error tolerance and easy error correction; (3) including multimodal prompts; and (4) involving clinicians, care partners, and survivors in technology design. Evidence in TBI is smaller and more heterogeneous than in general digital mental health, but

XSL•FO
RenderX

findings consistently point to usability as a facilitating factor for adoption and benefit.

This study's contribution is integrative rather than disciplinary. We operationalized a full-cycle pipeline that combines community-based participatory research with learning experience design; mapped participatory inputs onto a sociotechnical-pedagogical evaluation lens; and triangulated think-aloud, task, and survey metrics to produce community-informed design principles for TBI mHealth. Taken together, this cross-disciplinary operationalization provides a reproducible, pragmatic approach for formative mHealth development in cognitive rehabilitation and offers concrete, testable design guidance for teams working at the intersection of participatory methods, instructional design, and digital health.

Limitations and Future Directions

Several limitations constrain the generalizability of our findings. Usability testing used a small, purposive sample (n=5) appropriate for formative evaluation but insufficient for population-level inferences. Consequently, the effect estimates (eg, CAUSLT mean, accuracy) had wide confidence intervals; therefore, subgroup effects could not be assessed. Thus CAUSLT mean, knowledge accuracy, and completion rates should be viewed as exploratory. Our design mitigations were purposive sampling for heterogeneity, triangulation across qualitative and quantitative data streams, and sustained Community Advisory Board engagement to improve ecological validity. Nonetheless, future work should evaluate ePST in larger, more diverse TBI samples to quantify variability across injury characteristics, device types, demographic groups, and contexts of use and to permit powered hypothesis testing and subgroup analysis, which is the focus of our current feasibility study. Further, some reported technical issues may reflect device-specific limitations rather than design flaws, indicating need for expanded cross-platform testing. In addition to this, the TBI-specific design features reported here may not transfer directly to other neurological populations, requiring investigation of how CBPR-based approaches perform across different rehabilitation contexts. Although initial usability testing revealed strong satisfaction, sustainability of engagement remains unknown, suggesting a need for longitudinal metrics capturing retention, adherence, and health outcome durability. Future research should focus on evaluating barriers and facilitators to adoption and abandonment and how this engagement (or lack thereof) affects scale-up and sustainability of health care technologies using digital health frameworks such as the NASSS (nonadoption, abandonment, scale-up, spread, and sustainability) framework [135].

A direction for future research is how participatory practices might influence long-term health outcomes and treatment adherence beyond usability metrics. Integration of adaptive technologies such as artificial intelligence–driven personalization, voice-guided prompts, and real-time support could represent promising directions for accommodating cognitive variability in neurological populations. Additionally, examining the scalability of intensive CBPR approaches across diverse rehabilitation contexts (ie, stroke recovery, spinal cord injury) could advance understanding of participatory design's broader applicability.

Conclusions

This study demonstrated that systematic application of CBPR principles can produce both qualitative and quantitative improvements in rehabilitation technology usability through iterative community feedback, diverse stakeholder representation, and sustained engagement processes. The development of ePST illustrates how participatory methods can address specific design requirements for cognitive accessibility while maintaining high user satisfaction. The findings suggest that cognitive rehabilitation technology can benefit from specific design considerations including attention to cognitive load, clear navigation patterns, and explicit content signaling to address TBI-related challenges. This work provides further support for CBPR as a practical methodology in rehabilitation technology development, enhancing ethical research practices as well as technical outcomes. Investigation of long-term engagement sustainability and adaptive technology integration remains a direction for future research with promise for advancing understanding of how participatory approaches might contribute to more equitable, personalized, and effective rehabilitation interventions.

Acknowledgments

We used ChatGPT to assist with editing for grammar, creating figure and table captions, and enhancing readability of lengthy and complex sentence formulations. All artificial intelligence (AI) outputs were reviewed, edited, and approved by the named authors.

Funding

This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs through the Congressionally Directed Medical Research Programs TBI and Psychological Health program under award number HT9425-23-1-0567. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense.

Data Availability

All data are available in the manuscript and multimedia appendices.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Traumatic brain injury (TBI) personas and empathy maps for user-centered design: 11 de-identified, data-informed personas and associated empathy-map summaries created from Community Advisory Board (CAB)/Community Engagement Studios (CES) and empathy-interview data to guide Electronic Problem-Solving Training (ePST) module design and accessibility decisions. Each persona includes demographics (age, language, location), TBI characteristics and time-since-injury, goals, behaviors, attitudes, motivations, barriers, facilitators, and concise "key attributes" used to prioritize features (eg, linear lesson flow, memory supports, fatigue accommodations, family involvement).
[PDF File, 2560 KB - jmir_v28i1e83995_app1.pdf]

Checklist 1
GRIPP 2 Short Form checklist.
[PDF File, 77 KB - jmir_v28i1e83995_app2.pdf]

References

1. Toro-Hernández ML, Mondragón-Barrera A, Múnera-Orozco S, Villa-Torres L, Camelo-Castillo W. Experiences with rehabilitation and impact on community participation among adults with physical disability in Colombia: perspectives from stakeholders using a community based research approach. Int J Equity Health 2019 Jun 3;18(1):18. [doi: 10.1186/s12939-019-0923-4] [Medline: 31155006]
2. Khayatzadeh-Mahani A, Wittevrongel K, Nicholas DB, Zwicker JD. Prioritizing barriers and solutions to improve employment for persons with developmental disabilities. Disabil Rehabil 2020 Sep;42(19):2696-2706. [doi: 10.1080/09638288.2019.1570356] [Medline: 30856355]
3. Kersey J, Garcia P, Evans E, et al. Underrepresentation of participants from marginalized racial and ethnic groups: a secondary analysis of the cognitive rehabilitation literature. Arch Rehabil Res Clin Transl 2025 Jun;7(2):100431. [doi: 10.1016/j.arrct.2025.100431] [Medline: 40678291]
4. Omar S, Williams CC, Bugg LB, Colantonio A. "Somewhere along the line, your mask isn't going to be fitting right": institutional racism in Black narratives of traumatic brain injury rehabilitation across the practice continuum. BMC Health Serv Res 2024 Jul 24;24(1):834. [doi: 10.1186/s12913-024-10986-1] [Medline: 39049041]
5. Maas AIR, Menon DK, Manley GT, et al. Traumatic brain injury: progress and challenges in prevention, clinical care, and research. Lancet Neurol 2022 Nov;21(11):1004-1060. [doi: 10.1016/S1474-4422(22)00309-X] [Medline: 36183712]
6. Taylor CA, Bell JM, Breiding MJ, Xu L. Traumatic brain injury-related emergency department visits, hospitalizations, and deaths - United States, 2007 and 2013. MMWR Surveill Summ 2017 Mar 17;66(9):1-16. [doi: 10.15585/mmwr.ss6609a1] [Medline: 28301451]
7. Polinder S, Cnossen MC, Real RGL, et al. A Multidimensional approach to post-concussion symptoms in mild traumatic brain injury. Front Neurol 2018;9:1113. [doi: 10.3389/fneur.2018.01113] [Medline: 30619066]
8. Azouvi P, Arnould A, Dromer E, Vallat-Azouvi C. Neuropsychology of traumatic brain injury: an expert overview. Rev Neurol (Paris) 2017;173(7-8):461-472. [doi: 10.1016/j.neurol.2017.07.006] [Medline: 28847474]
9. Milders M. Relationship between social cognition and social behaviour following traumatic brain injury. Brain Inj 2019;33(1):62-68. [doi: 10.1080/02699052.2018.1531301] [Medline: 30325217]
10. Semple BD, Zamani A, Rayner G, Shultz SR, Jones NC. Affective, neurocognitive and psychosocial disorders associated with traumatic brain injury and post-traumatic epilepsy. Neurobiol Dis 2019 Mar;123:27-41. [doi: 10.1016/j.nbd.2018.07.018] [Medline: 30059725]
11. Ertas-Spantgar F, Korabova S, Gabel A, Schiering I, Müller SV. Guiding patients with traumatic brain injury through the instrumental activities of daily living with the RehaGoal App: a feasibility study. Disabil Rehabil Assist Technol 2024 Feb;19(2):254-265. [doi: 10.1080/17483107.2022.2080290] [Medline: 35713480]
12. Hou Y, Zhou A, Brooks L, Reid D, Turkstra L, MacDonald S. Rehabilitation access for individuals with cognitive-communication challenges after traumatic brain injury: a co-design study with persons with lived experience. Int J Lang Commun Disord 2024;59(2):648-664. [doi: 10.1111/1460-6984.12895] [Medline: 37189286]
13. Novak I, te Velde A, Hines A, et al. Rehabilitation evidence-based decision-making: the READ model. Front Rehabilit Sci 2021 Oct 5;2. [doi: 10.3389/fresc.2021.726410]
14. Doerwald F, Stalling I, Recke C, et al. A rapid review of digital approaches for the participatory development of health-related interventions. Front Public Health 2024;12:1461422. [doi: 10.3389/fpubh.2024.1461422] [Medline: 39678234]
15. Benz C, Scott-Jeffs W, McKercher KA, et al. Community-based participatory-research through co-design: supporting collaboration from all sides of disability. Res Involv Engagem 2024 May 10;10(1):47. [doi: 10.1186/s40900-024-00573-3] [Medline: 38730283]

XSL•FO
RenderX

16. Kilfoy A, Hsu TCC, Stockton-Powdrell C, Whelan P, Chu CH, Jibb L. An umbrella review on how digital health intervention co-design is conducted and described. NPJ Digit Med 2024 Dec 23;7(1):374. [doi: 10.1038/s41746-024-01385-1] [Medline: 39715947]
17. Perrin PB, Haun JN, Klyce DW, et al. Efficacy and implementation planning across the Veterans Affairs polytrauma system of care: protocol for the REACH intervention for caregivers of veterans and service members with traumatic brain injury. JMIR Res Protoc 2024 Aug 15;13:e57692. [doi: 10.2196/57692] [Medline: 39145996]
18. Castro-Figueroa E, Rosario-Maldonado FJ, Asencio-Toro G, et al. Empowering community partners in health disparities research: refining a community-based participatory research (CBPR) training curriculum. Pedagogy Health Promot 2025. [doi: 10.1177/23733799251345430]
19. Mohatt NV, Kreisel CJ, Brenner LA, CRITICAL Team. Engaging those living with moderate to severe TBI and their caregivers in research. J Patient Exp 2021;8(2374373521998852):2374373521998852. [doi: 10.1177/2374373521998852] [Medline: 34179408]
20. Haun JN, Nakase-Richardson R, Cotner BA, et al. Stakeholder engagement to identify implementation strategies to overcome barriers to delivering chronic pain treatments: a NIDILRR and VA TBI model systems collaborative project. J Head Trauma Rehabil 2024;39(1):E29-E40. [doi: 10.1097/HTR.0000000000000920] [Medline: 38167720]
21. Eady K, Moreau KA, Marshall S, Egan M. Patient, family, and health professional perspectives of how families are involved in adult inpatient traumatic brain injury rehabilitation. Patient Exp J 2024 Nov 14;11(3):29-36. [doi: 10.35680/2372-0247.1891]
22. Cameron LJ, Somerville LM, Naismith CE, Watterson D, Maric V, Lannin NA. A qualitative investigation into the patient-centered goal-setting practices of allied health clinicians working in rehabilitation. Clin Rehabil 2018 Jun;32(6):827-840. [doi: 10.1177/0269215517752488] [Medline: 29327603]
23. Hoffman JM, Curran M, Barber J, Lucas S, Fann JR, Zumsteg JM. Collaborative care for chronic pain after traumatic brain injury: a randomized clinical trial. JAMA Netw Open 2024 Jun 3;7(6):e2413459. [doi: 10.1001/jamanetworkopen.2024.13459] [Medline: 38829619]
24. D'Cruz K, Antonopoulos S, Rothman R, Douglas J, Winkler D, Oliver S. Co-designing with adults with acquired neurological disability in the community: a scoping review protocol. BMJ Open 2022 Dec 8;12(12):e064921. [doi: 10.1136/bmjopen-2022-064921] [Medline: 36600382]
25. Curran MC, Lucas S, Fann JR, Zumsteg JM, Hoffman JM. Chronic pain after traumatic brain injury: a collaborative care approach. Front Rehabil Sci 2024;5:1398856. [doi: 10.3389/fresc.2024.1398856] [Medline: 39253025]
26. Morrow EL, Mayberry LS, Duff MC. The growing gap: a study of sleep, encoding, and consolidation of new words in chronic traumatic brain injury. Neuropsychologia 2023 Jun 6;184:108518. [doi: 10.1016/j.neuropsychologia.2023.108518] [Medline: 36804844]
27. Manley K, Saunders K, Wilkinson D, Faruqui R, Sakel M. Co-creating system-wide improvement for people with traumatic brain injury across one integrated care system in the United Kingdom to initiate a transformation journey through co-production. Health Expect 2023 Apr;26(2):869-881. [doi: 10.1111/hex.13712] [Medline: 36715266]
28. Quilico E, Wilkinson S, Duncan L, et al. Participatory co-creation of an adapted physical activity program for adults with moderate-to-severe traumatic brain injury. Front Rehabil Sci 2022;3:900178. [doi: 10.3389/fresc.2022.900178] [Medline: 36188895]
29. Power E, Morrow R. Digital, co-created implementation of communication partner training programs for stroke, brain injury, and dementia: past, present, and future. Int J Speech Lang Pathol 2024 Jun;26(3):317-333. [doi: 10.1080/17549507.2024.2362856] [Medline: 38962904]
30. Karhula M, Saukkonen S, Kinnunen A, Heiskanen T, Xiong E, Anttila H. ICF-luokituksen yksilötekijöiden kuvaus on osa toimintakyvyn laaja-alaista arviointia: kartoittava kirjallisuuskatsaus ICF-yksilötekijöitä käsittelevistä tutkimuksista. Kuntoutus 2021 Jun 18;44(2):9-24. [doi: 10.37451/kuntoutus.109476]
31. Salsberg J, Macridis S, Garcia Bengoechea E, Macaulay AC, Moore S, KSDPP School Travel Planning Committee. The shifting dynamics of social roles and project ownership over the lifecycle of a community-based participatory research project. Fam Pract 2017 Jun 1;34(3):305-312. [doi: 10.1093/fampra/cmx006] [Medline: 28334748]
32. Burduladze N, Jones LP, Jones BD, et al. Exploring power and power sharing in participatory health research partnerships: a scoping review protocol. PLOS ONE 2024;19(7):e0303799. [doi: 10.1371/journal.pone.0303799] [Medline: 39024348]
33. Manhas KP, Olson K, Churchill K, Vohra S, Wasylak T. Experiences of shared decision-making in community rehabilitation: a focused ethnography. BMC Health Serv Res 2020 Apr 19;20(1):329. [doi: 10.1186/s12913-020-05223-4] [Medline: 32306972]
34. Nezu AM, Nezu CM. Problem solving. In: Norcross JC, VandenBos GR, Freedheim DK, Pole N, editors. APA Handbook of Clinical Psychology: Psychopathology and Health: American Psychological Association; 2016:449-460. [doi: 10.1037/14862-019]
35. D'Zurilla TJ, Nezu AM, Maydeu-Olivares A. Social problem solving: theory and assessment. In: Chang EC, D'Zurilla TJ, Sanna LJ, editors. Social Problem Solving: Theory, Research, and Training 2004. URL: http://psycnet.apa.org/psycinfo/2004-14507-001

XSL•FO
RenderX

36. D'Zurilla TJ, Nezu AM. Problem-Solving Therapy: A Positive Approach to Clinical Intervention, 3rd edition: Springer Publishing Company; 2007.
37. D'Zurilla TJ, Goldfried MR. Problem solving and behavior modification. J Abnorm Psychol 1971 Aug;78(1):107-126. [doi: 10.1037/h0031360] [Medline: 4938262]
38. Jiang C, Zhou H, Chen L, Zhou Z. Problem solving therapy improves effortful cognition in major depression. Front Psychiatry 2021;12:607718. [doi: 10.3389/fpsyt.2021.607718] [Medline: 33897483]
39. Cuijpers P, de Wit L, Kleiboer A, Karyotaki E, Ebert DD. Problem-solving therapy for adult depression: an updated meta-analysis. Eur psychiatr 2018;48(1):27-37. [doi: 10.1016/j.eurpsy.2017.11.006]
40. Narad ME, Raj S, Yeates KO, et al. Randomized controlled trial of an online problem-solving intervention following adolescent traumatic brain injury: family outcomes. Arch Phys Med Rehabil 2019 May;100(5):811-820. [doi: 10.1016/j.apmr.2019.01.010] [Medline: 30738021]
41. Zhang N, Kaizar EE, Narad ME, et al. Examination of injury, host, and social-environmental moderators of online family problem solving treatment efficacy for pediatric traumatic brain injury Using an individual participant data meta-analytic approach. J Neurotrauma 2019 Apr 1;36(7):1147-1155. [doi: 10.1089/neu.2018.5885] [Medline: 30328749]
42. Palermo TM, Law EF, Bromberg M, Fales J, Eccleston C, Wilson AC. Problem-solving skills training for parents of children with chronic pain: a pilot randomized controlled trial. Pain 2016 Jun;157(6):1213-1223. [doi: 10.1097/j.pain.0000000000000508] [Medline: 26845525]
43. Economides M, Ranta K, Nazander A, et al. Long-term outcomes of a therapist-supported, smartphone-based intervention for elevated symptoms of depression and anxiety: quasiexperimental, pre-postintervention study. JMIR Mhealth Uhealth 2019 Aug 26;7(8):e14284. [doi: 10.2196/14284] [Medline: 31452521]
44. Ghanbari E, Yektatalab S, Mehrabi M. Effects of psychoeducational interventions using mobile apps and mobile-based online group discussions on anxiety and self-esteem in women with breast cancer: randomized controlled trial. JMIR Mhealth Uhealth 2021 May 18;9(5):e19262. [doi: 10.2196/19262] [Medline: 34003138]
45. Floor N. This Is Learning Experience Design: What It Is, How It Works, and Why It Matters: New Riders; 2023.
46. Israel BA, Eng E, Schulz AJ, Parker EA. Methods for Community-Based Participatory Research for Health: John Wiley & Sons; 2005.
47. Lim H, Kakonge L, Hu Y, et al. So, i can feel normal: participatory design for accessible social media sites for individuals with traumatic brain injury. 2023 Apr 19 Presented at: CHI '23; Apr 19, 2023; Hamburg Germany p. 1-19 URL: https://dl.acm.org/doi/proceedings/10.1145/3544548 [doi: 10.1145/3544548.3581222]
48. Lorenz EA, Bråten Støen A, Lie Fridheim M, Alsos OA. Design recommendations for XR-based motor rehabilitation exergames at home. Front Virtual Real 2024 Jan 22;5. [doi: 10.3389/frvir.2024.1340072]
49. Correa DJ, Kwon CS, Connors S, et al. Applying participatory action research in traumatic brain injury studies to prevent post-traumatic epilepsy. Neurobiol Dis 2019 Mar;123:137-144. [doi: 10.1016/j.nbd.2018.07.007] [Medline: 30031158]
50. Groussard PY, Pigot H, Giroux S. From conception to evaluation of mobile services for people with head injury: a participatory design perspective. Neuropsychol Rehabil 2018 Jul;28(5):667-688. [doi: 10.1080/09602011.2015.1117499] [Medline: 26679473]
51. Coombe CM, Schulz AJ, Guluma L, et al. Enhancing capacity of community-academic partnerships to achieve health equity: results from the CBPR Partnership Academy. Health Promot Pract 2020 Jul;21(4):552-563. [doi: 10.1177/1524839918818830] [Medline: 30596283]
52. Springer MV, Skolarus LE. Community-based participatory research. Stroke 2019 Mar;50(3):e48-e50. [doi: 10.1161/STROKEAHA.118.024241] [Medline: 30661505]
53. Bonn MM, Graham LJ, Marrocco S, Jeske S, Moran B, Wolfe DL. Usability evaluation of a self-management mobile application for individuals with a mild traumatic brain injury. Digit Health 2023;9:20552076231183555. [doi: 10.1177/20552076231183555] [Medline: 37426589]
54. Smith KA, Ward T, Lambe S, et al. Engagement and attrition in digital mental health: current challenges and potential solutions. NPJ Digit Med 2025 Jul 2;8(1):398. [doi: 10.1038/s41746-025-01778-w] [Medline: 40604240]
55. Abbasalizadeh M, Farsi Z, Sajadi SA, Atashi A. The effect of mobile health application training based on micro-learning method on the level of resilience and happiness among intensive care nurses: a randomized controlled trial. BMC Psychiatry 2024 Dec 27;24(1):954. [doi: 10.1186/s12888-024-06429-0] [Medline: 39731084]
56. Jacob C, Lindeque J, Müller R, et al. A sociotechnical framework to assess patient-facing eHealth tools: results of a modified Delphi process. NPJ Digit Med 2023 Dec 15;6(1):232. [doi: 10.1038/s41746-023-00982-w] [Medline: 38102323]
57. Lin MCM, Vasarhelyi K, Wong KLY, et al. Engaging community to co-design learning health systems: lessons from storytelling and Design Jam, a community case study from British Columbia, Canada. Front Health Serv 2025;5:1620659. [doi: 10.3389/frhs.2025.1620659] [Medline: 40800072]
58. Joosten YA, Israel TL, Williams NA, et al. Community Engagement Studios: a structured approach to obtaining meaningful input from stakeholders to inform research. Acad Med 2015 Dec;90(12):1646-1650. [doi: 10.1097/ACM.0000000000000794] [Medline: 26107879]

XSL•FO
RenderX

59. Zisman-Ilani Y, Buell J, Mazel S, Hennig S, Nicholson J. Virtual Community Engagement Studio (V-CES): engaging mothers with mental health and substance use conditions in research. Front Psychiatry 2022;13:805781. [doi: 10.3389/fpsyt.2022.805781] [Medline: 35782439]
60. Vial S, Boudhraâ S, Dumont M. Human-centered design approaches in digital mental health interventions: exploratory mapping review. JMIR Ment Health 2022 Jun 7;9(6):e35591. [doi: 10.2196/35591] [Medline: 35671081]
61. Banbury A, Pedell S, Parkinson L, Byrne L. Using the Double Diamond model to co-design a dementia caregivers telehealth peer support program. J Telemed Telecare 2021 Dec;27(10):667-673. [doi: 10.1177/1357633X211048980] [Medline: 34726994]
62. Kersey J, Alimi E, McArthur AR, et al. ENGAGE-TBI: adaptation of a community-based intervention to improve social participation after brain injury. Brain Inj 2025 May 12;39(6):518-525. [doi: 10.1080/02699052.2025.2449927] [Medline: 39773100]
63. Zhang Y, Xie YJ, Yang L, et al. Community-based participatory research (CBPR) approaches in vaccination promotion: a scoping review. Int J Equity Health 2024 Nov 5;23(1):227. [doi: 10.1186/s12939-024-02278-1] [Medline: 39501299]
64. Kayes NM, Martin RA, Bright FA, Kersten P, Pollock A. Optimizing the real-world impact of rehabilitation reviews: increasing the relevance and usability of systematic reviews in rehabilitation. Eur J Phys Rehabil Med 2019 Jun;55(3):331-341. [doi: 10.23736/S1973-9087.19.05793-9] [Medline: 30990002]
65. Grindell C, Sanders T, Bec R, Mary Tod A, Wolstenholme D. Improving knowledge mobilisation in healthcare: a qualitative exploration of creative co-design methods. Evid Policy 2022 May 1;18(2):265-290. [doi: 10.1332/174426421X16436512504633]
66. Israel BA, Schulz AJ, Coombe CM, et al. Community-based participatory research: an approach to research in the urban context. In: Galea S, Ettman CK, Vlahov D, editors. Urban Health: Oxford University Press; 2019. [doi: 10.1093/oso/9780190915858.003.0029]
67. Ferreira MP, Gendron F. Community-based participatory research with traditional and indigenous communities of the Americas: historical context and future directions. International Journal of Critical Pedagogy ;3(3) [FREE Full text]
68. Wallerstein N, Muhammad M, Sanchez-Youngman S, et al. Power dynamics in community-based participatory research: a multiple-case study analysis of partnering contexts, histories, and practices. Health Educ Behav 2019 Oct;46(1_suppl):19S-32S. [doi: 10.1177/1090198119852998] [Medline: 31549557]
69. Plamondon K, Ndumbe-Eyoh S, Shahram S. Equity, power, and transformative research coproduction. In: Graham ID, Rycroft-Malone J, Kothari A, McCutcheon C, editors. Research Co‐Production in Healthcare 2022:34-53. [doi: 10.1002/9781119757269.ch3]
70. Collins SE, Clifasefi SL, Stanton J, et al. Community-based participatory research (CBPR): towards equitable involvement of community in psychology research. Am Psychol 2018 Oct;73(7):884-898. [doi: 10.1037/amp0000167] [Medline: 29355352]
71. Agyepong IA, Godt S, Sombie I, Binka C, Okine V, Ingabire MG. Strengthening capacities and resource allocation for co-production of health research in low and middle income countries. BMJ 2021 Feb 15;372:n166. [doi: 10.1136/bmj.n166] [Medline: 33593725]
72. Newman SD, Andrews JO, Magwood GS, Jenkins C, Cox MJ, Williamson DC. Community advisory boards in community-based participatory research: a synthesis of best processes. Prev Chronic Dis 2011 May;8(3):A70. [Medline: 21477510]
73. Safo S, Cunningham C, Beckman A, Haughton L, Starrels JL. “A place at the table:” a qualitative analysis of community board members’ experiences with academic HIV/AIDS research. BMC Med Res Methodol 2016 Jul 11;16(1):80. [doi: 10.1186/s12874-016-0181-8] [Medline: 27401678]
74. Biondo J, Johnson N. The process and significance of convening a community advisory board with individuals with severe mental illness. Journal of Participatory Research Methods 2025;6(2). [doi: 10.35844/001c.130021]
75. Duke M. Community-based participatory research. Oxford Research Encyclopedia of Anthropology 2020. [doi: 10.1093/acrefore/9780190854584.013.225]
76. Abraham H, Anyetei-Anum GP, Krogman A, et al. The HEALERS: a patient, community, and stakeholder advisory board focus group series to refine a novel virtual world-based cardiac rehabilitation intervention and clinical trial. Front Digit Health 2025;7:1427539. [doi: 10.3389/fdgth.2025.1427539] [Medline: 40735343]
77. Hornbuckle LM, Rauer A. Engaging a community advisory board to inform an exercise intervention in older African-American couples. J Prim Prev 2020 Jun;41(3):261-278. [doi: 10.1007/s10935-020-00589-x] [Medline: 32410065]
78. Brockman TA, Balls-Berry JE, West IW, et al. Researchers’ experiences working with community advisory boards: how community member feedback impacted the research. J Clin Trans Sci 2021;5(1):e117. [doi: 10.1017/cts.2021.22]
79. Ogunsanya ME, Kaninjing E, Morton DJ, Dwyer K, Young ME, Odedina FT. Bridging the gap: a community advisory board promoting community engagement in cancer research for ethnically diverse populations. Am J Mens Health 2024;18(5):15579883241280826. [doi: 10.1177/15579883241280826] [Medline: 39340388]
80. Quinn ED, Cotter K, Kurin K, Brown K. Conducting a Community Engagement Studio to adapt enhanced milieu teaching. Am J Speech Lang Pathol 2022 May 10;31(3):1095-1113. [doi: 10.1044/2021_AJSLP-21-00100] [Medline: 35007426]

XSL•FO
RenderX

81. Killough CM, Martinez J, Mata H, et al. New horizons in community engagement: virtual community engagement studios amplifying community voices about health research in New Mexico. J Clin Trans Sci 2024;8(1):e140. [doi: 10.1017/cts.2024.608]
82. Nielson C, Huang Y, Kull CA, Park AH. Utilizing Community Engagement Studios to inform patient experience in a multicenter randomized control trial. Int J Pediatr Otorhinolaryngol 2020 Jun;133:110007. [doi: 10.1016/j.ijporl.2020.110007] [Medline: 32208178]
83. Johnson DA, Joosten YA, Wilkins CH, Shibao CA. Case study: community engagement and clinical trial success: outreach to African American women. Clinical Translational Sci 2015 Aug;8(4):388-390 [FREE Full text] [doi: 10.1111/cts.12264]
84. Skiba MB, Badger TA, Garcia DO, Chilton FH, Winters-Stone KM. Adapting a dyadic exercise program to be culturally relevant for Hispanic men with prostate cancer using community engagement studio: a brief report. Front Psychol 2024;15:1294546. [doi: 10.3389/fpsyg.2024.1294546] [Medline: 38716273]
85. Scheffey K, Avelis J, Patel M, Oon AL, Evans C, Glanz K. Use of Community Engagement Studios to adapt a hybrid effectiveness-implementation study of social incentives and physical activity for the STEP Together study. Health Promot Pract 2024 Mar;25(2):285-292. [doi: 10.1177/15248399221113863] [Medline: 35899691]
86. Stock MR, Ceïde ME, Lounsbury DW, Zwerling J. Utilizing community engagement studios to inform clinical trial design at a Center of Excellence for Alzheimer's Disease. J Clin Trans Sci 2022;6(1):e73. [doi: 10.1017/cts.2022.388]
87. Schmidt M, Huang R. Defining learning experience design: voices from the field of learning design & technology. TechTrends 2022 Mar;66(2):141-158. [doi: 10.1007/s11528-021-00656-y]
88. Jahnke I, Schmidt M, Earnshaw Y, Tawfik A. Theoretical considerations of learning experience design. In: Theories to Influence the Future of Learning Design and Technology: EdTech Books; 2022. [doi: 10.59668/534]
89. Aloizou V, Ioannou A, Boloudakis M, Retalis S. A learning experience design framework for multimodal learning in the early childhood. Smart Learn Environ 2025;12(1). [doi: 10.1186/s40561-025-00376-3]
90. Hernández R, Kilar-Magdziarz E. Learning experience design (LXD) of language and content modules: insights from students and instructors. In: Kumar P, Eisenberg J, editors. Synchronous and Asynchronous Approaches to Teaching: Palgrave Macmillan; 2023:269-290. [doi: 10.1007/978-3-031-17841-2_13]
91. Conceição SCO, Howles L. Designing the Online Learning Experience: Evidence-Based Principles and Strategies: Taylor & Francis; 2023.
92. Kensing F, Blomberg J. Participatory design: issues and concerns. Computer Supported Cooperative Work (CSCW) 1998 Sep;7(3-4):167-185. [doi: 10.1023/A:1008689307411]
93. Bowen K, Forssell KS, Rosier S. Theories of change in learning experience (LX) design. In: Tawfik A, Earnshaw Y, Jahnke I, editors. Learner and User Experience Research: An Introduction for the Field of Learning Design and Technology 2020.
94. Georgsson M, Staggers N. An evaluation of patients' experienced usability of a diabetes mHealth system using a multi-method approach. J Biomed Inform 2016 Feb;59:115-129. [doi: 10.1016/j.jbi.2015.11.008] [Medline: 26639894]
95. McDonald J, Westerberg T. Learning experience design as an orienting guide for practice: insights from designing for expertise. jaid 2023. [doi: 10.59668/515.12898]
96. McInnes K, Friesen CL, MacKenzie DE, Westwood DA, Boe SG. Mild traumatic brain injury (mTBI) and chronic cognitive impairment: a scoping review. PLoS ONE 2017;12(4):e0174847. [doi: 10.1371/journal.pone.0174847] [Medline: 28399158]
97. Mavroudis I, Ciobica A, Bejenariu AC, et al. Cognitive impairment following mild traumatic brain injury (mTBI): a review. Med Bogota Colomb 2024;60(3):380. [doi: 10.3390/medicina60030380]
98. Jahnke I, Lee YM, Pham M, He H, Austin L. Unpacking the inherent design principles of mobile microlearning. Tech Know Learn 2020 Sep;25(3):585-619. [doi: 10.1007/s10758-019-09413-w]
99. Lopez S. The impact of cognitive load theory on the effectiveness of microlearning modules. European Journal of Education and Pedagogy 2024;5(2):29-35 [FREE Full text] [doi: 10.24018/ejedu.2024.5.2.799]
100. Mayer RE. Cognitive theory of multimedia learning. In: Mayer R, editor. The Cambridge Handbook of Multimedia Learning: Cambridge University Press; 2005:31-48. [doi: 10.1017/CBO9780511816819.004]
101. Rahmani-Katigari M, Mohammadian F, Shahmoradi L. Development of a serious game-based cognitive rehabilitation system for patients with brain injury. BMC Psychiatry 2023 Nov 29;23(1):893. [doi: 10.1186/s12888-023-05396-2] [Medline: 38031072]
102. Jahnke I, Riedel N, Singh K, Moore J. Advancing sociotechnical-pedagogical heuristics for the usability evaluation of online courses for adult learners. OLJ 2021;25(4):337-360. [doi: 10.24059/olj.v25i4.2439]
103. Schmidt M, Earnshaw Y, Jahnke I, Tawfik AA. Entangled eclecticism: a sociotechnical-pedagogical systems theory approach to learning experience design. Education Tech Research Dev 2024 Jun;72(3):1483-1505. [doi: 10.1007/s11423-024-10353-1]
104. Staniszewska S, Brett J, Simera I, et al. GRIPP2 reporting checklists: tools to improve reporting of patient and public involvement in research. Res Involv Engagem 2017;3(1):13. [doi: 10.1186/s40900-017-0062-2] [Medline: 29062538]
105. Virzi RA. Refining the test phase of usability evaluation: how many subjects is enough? Hum Factors 1992 Aug;34(4):457-468. [doi: 10.1177/001872089203400407]
106. Nielsen J, Landauer TK. A mathematical model of the finding of usability problems. 1993 Presented at: the SIGCHI conference; May 1, 1993; Amsterdam, The Netherlands p. 206-213 URL: http://portal.acm.org/citation.cfm?doid=169059 [doi: 10.1145/169059.169166]

XSL•FO
RenderX

107. Faulkner L. Beyond the five-user assumption: benefits of increased sample sizes in usability testing. Behav Res Methods Instrum Comput 2003 Aug;35(3):379-383. [doi: 10.3758/bf03195514] [Medline: 14587545]
108. Lewis JR. Sample sizes for usability studies: additional considerations. Hum Factors 1994 Jun;36(2):368-378. [doi: 10.1177/001872089403600215] [Medline: 8070799]
109. Fowler LA, Vázquez MM, DePietro B, Wilfley DE, Fitzsimmons-Craft EE. Development, usability, and preliminary efficacy of a virtual reality experience to promote healthy lifestyle behaviors in children: pilot randomized controlled trial. Mhealth 2024;10:29. [doi: 10.21037/mhealth-24-24] [Medline: 39534453]
110. Papadopoulos-Nydam G, Rieger JM, Constantinescu G. Usability testing of a mHealth system for swallowing therapy in patients following stroke. Perspect ASHA SIGs 2021 Oct 20;6(5):1205-1211. [doi: 10.1044/2021_PERSP-21-00075]
111. Constantinescu G, Kuffel K, King B, Hodgetts W, Rieger J. Usability testing of an mHealth device for swallowing therapy in head and neck cancer survivors. Health Informatics J 2019 Dec;25(4):1373-1382. [doi: 10.1177/1460458218766574] [Medline: 29618274]
112. Patton MQ. Qualitative Research & Evaluation Methods: Integrating Theory and Practice: SAGE Publications, Inc; 2015.
113. Juengst S, Supnet C, Kew CLN, et al. Bilingual problem-solving training for caregivers of adults with dementia: a randomized, factorial-design protocol for the CaDeS trial. Contemp Clin Trials 2021 Sep;108:106506. [doi: 10.1016/j.cct.2021.106506] [Medline: 34273551]
114. Kouprie M, Visser FS. A framework for empathy in design: stepping into and out of the user's life. Journal of Engineering Design 2009 Oct;20(5):437-448. [doi: 10.1080/09544820902875033]
115. Siricharoen WV. Using empathy mapping in design thinking process for personas discovering. In: Vinh PC, Rakib A, editors. : Springer International Publishing; 2021 Presented at: Context-Aware Systems and Applications, and Nature of Computation and Communication. ICCASA ICTCC 2020 p. 182-191. [doi: 10.1007/978-3-030-67101-3_15]
116. Schmidt M, Tawfik A. Activity theory as a lens for developing and applying personas and scenarios in learning experience design. JAID 2022. [doi: 10.59668/354.5904]
117. Tripp SD, Bichelmeyer B. Rapid prototyping: an alternative instructional design strategy. ETR&D 1990 Mar;38(1):31-44. [doi: 10.1007/BF02298246]
118. Lu J, Schmidt M, Shin J. Beyond technological usability: exploratory factor analysis of the comprehensive assessment of usability scale for learning technologies (CAUSLT). arXiv. Preprint posted online on Feb 3, 2025. [doi: 10.48550/arXiv.2501.18754]
119. Nielsen J. Severity Ratings for Usability Problems.: Nielsen Norman Group URL: https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/ [accessed 2025-09-10]
120. Kali Y. The design principles database as a means for promoting design-based research. In: Handbook of Design Research Methods in Education: Routledge; 2008:423-438. [doi: 10.4324/9781315759593]
121. ISO 9241-11:2018: ergonomics of human-system interaction part 11: usability: definitions and concepts. International Organization for Standardization. 2018. URL: https://www.iso.org/standard/63500.html [accessed 2026-01-10]
122. Zhou L, Bao J, Setiawan IMA, Saptono A, Parmanto B. The mHealth App Usability Questionnaire (MAUQ): development and validation study. JMIR Mhealth Uhealth 2019 Apr 11;7(4):e11500. [doi: 10.2196/11500] [Medline: 30973342]
123. IEC 62366-1:2015 — Medical devices — Part 1: Application of usability engineering to medical devices. International Electrotechnical Commission. 2015. URL: https://webstore.iec.ch/en/publication/21863 [accessed 2026-01-03]
124. IEC 62366-1:2015/AMD1:2020 — Amendment 1 - medical devices - part 1: application of usability engineering to medical devices. International Electrotechnical Commission. 2020. URL: https://webstore.iec.ch/en/publication/59980 [accessed 2026-01-03]
125. Deniz-Garcia A, Fabelo H, Rodriguez-Almeida AJ, et al. Quality, usability, and effectiveness of mHealth apps and the role of artificial intelligence: current scenario and challenges. J Med Internet Res 2023 May 4;25:e44030. [doi: 10.2196/44030] [Medline: 37140973]
126. Hach S, Alder G, Stavric V, Taylor D, Signal N. Usability assessment methods for mobile apps for physical rehabilitation: umbrella review. JMIR Mhealth Uhealth 2024 Oct 4;12(1):e49449. [doi: 10.2196/49449] [Medline: 39365988]
127. Cicerone KD, Goldin Y, Ganci K, et al. Evidence-based cognitive rehabilitation: systematic review of the literature from 2009 through 2014. Arch Phys Med Rehabil 2019 Aug;100(8):1515-1533. [doi: 10.1016/j.apmr.2019.02.011] [Medline: 30926291]
128. Quintero C. A review: accessible technology through participatory design. Disabil Rehabil Assist Technol 2022 May;17(4):369-375. [doi: 10.1080/17483107.2020.1785564] [Medline: 32620068]
129. Fischer B, Peine A, Östlund B. The importance of user involvement: a systematic review of involving older users in technology design. Gerontologist 2020 Sep 15;60(7):e513-e523. [doi: 10.1093/geront/gnz163] [Medline: 31773145]
130. Sanders EN, Stappers PJ. Co-creation and the new landscapes of design. International Journal of CoCreation in Design and the Arts 2008;4:5-18. [doi: 10.1080/15710880701875068]
131. de Jong T. Cognitive load theory, educational research, and instructional design: some food for thought. Instr Sci 2010 Mar;38(2):105-134. [doi: 10.1007/s11251-009-9110-0]
132. Alahmadi T, Drew S. Subjective evaluation of website accessibility and usability: a survey for people with sensory disabilities. W4A '17: Proceedings of the 14th International Web for All Conference 2017:1-4. [doi: 10.1145/3058555.3058579]

XSL•FO
RenderX

133. Bonn MM, Graham LJ, Marrocco S, Jeske S, Moran B, Wolfe DL. Usability evaluation of a self-management mobile application for individuals with a mild traumatic brain injury. Digit Health 2023;9(20552076231183555):20552076231183555. [doi: 10.1177/20552076231183555] [Medline: 37426589]
134. Portz J, Moore S, Bull S. Evolutionary trends in the adoption, adaptation, and abandonment of mobile health technologies: viewpoint based on 25 years of research. J Med Internet Res 2024 Sep 27;26(1):e62790. [doi: 10.2196/62790] [Medline: 39331463]
135. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
136. Kujala S. User involvement: a review of the benefits and challenges. Behav Inf Technol 2003 Jan;22(1):1-16. [doi: 10.1080/01449290301782]
137. Patel S, Akhtar A, Malins S, et al. The acceptability and usability of digital health interventions for adults with depression, anxiety, and somatoform disorders: qualitative systematic review and meta-synthesis. J Med Internet Res 2020 Jul 6;22(7):e16228. [doi: 10.2196/16228] [Medline: 32628116]
138. Phillips B, Zhao H. Predictors of assistive technology abandonment. Assist Technol 1993;5(1):36-45. [doi: 10.1080/10400435.1993.10132205] [Medline: 10171664]
139. Kidman PG, Curtis RG, Watson A, Maher CA. When and why adults abandon lifestyle behavior and mental health mobile apps: scoping review. J Med Internet Res 2024 Dec 18;26(1):e56897. [doi: 10.2196/56897] [Medline: 39693620]
140. Forbes A, Keleher MR, Venditto M, DiBiasi F. Assessing patient adherence to and engagement with digital interventions for depression in clinical trials: systematic literature review. J Med Internet Res 2023 Aug 11;25:e43727. [doi: 10.2196/43727] [Medline: 37566447]
141. Johnston M, Mobasheri M, King D, Darzi A. The Imperial Clarify, Design and Evaluate (CDE) approach to mHealth app development. BMJ Innov 2015 Apr;1(2):39-42. [doi: 10.1136/bmjinnov-2014-000020]
142. Ikwunne T, Hederman L, Wall PJ. Design processes for user engagement with mobile health: a systematic review. IJACSA 2022;13(2). [doi: 10.14569/IJACSA.2022.0130235]
143. Martin S, Armstrong E, Thomson E, et al. A qualitative study adopting a user-centered approach to design and validate a brain computer interface for cognitive rehabilitation for people with brain injury. Assist Technol 2018;30(5):233-241. [doi: 10.1080/10400435.2017.1317675] [Medline: 28708963]
144. Pinard S, Bottari C, Laliberté C, et al. Development of an assistive technology for cognition to support meal preparation in severe traumatic brain injury: user-centered design study. JMIR Hum Factors 2022 Aug 4;9(3):e34821. [doi: 10.2196/34821] [Medline: 35925663]
145. Schmidt M, Cheng L, Raj S, Wade S. Formative design and evaluation of a responsive eHealth/mHealth intervention for positive family adaptation following pediatric traumatic brain injury. J Form Des Learn 2020 Dec;4(2):88-106. [doi: 10.1007/s41686-020-00049-z]
146. Schmidt M, Babcock L, Kurowski BG, Cassedy A, Sidol C, Wade SL. Usage patterns of an mHealth symptom monitoring app among adolescents with acute mild traumatic brain injuries. J Head Trauma Rehabil 2022;37(3):134-143. [doi: 10.1097/HTR.0000000000000768] [Medline: 35125434]
147. Kraaijkamp JJM, van Dam van Isselt EF, Persoon A, Versluis A, Chavannes NH, Achterberg WP. eHealth in geriatric rehabilitation: systematic review of effectiveness, feasibility, and usability. J Med Internet Res 2021 Aug 19;23(8):e24015. [doi: 10.2196/24015] [Medline: 34420918]

Abbreviations

CaDeS: Caregivers in Dementia PST and DSJ
CBPR: community-based participatory research
ePST: Electronic Problem-Solving Training
mHealth: mobile health
NASSS: nonadoption, abandonment, scale-up, spread, and sustainability
PST: Problem-Solving Training
TBI: traumatic brain injury

XSL•FO
RenderX

Edited by S Brini; submitted 11.Sep.2025; peer-reviewed by OC Orwa, GLP Bodagala, O Akhadelor; accepted 11.Nov.2025; published 20.Jan.2026.

Please cite as:
Schmidt M, Weng Y, Juengst S, Holland A
Designing Electronic Problem-Solving Training for Individuals With Traumatic Brain Injury: Mixed Methods, Community-Based, Participatory Research Case Study
J Med Internet Res 2026;28:e83995
URL: https://www.jmir.org/2026/1/e83995
doi:10.2196/83995

© Matthew Schmidt, Yueqi Weng, Shannon Juengst, Alexandra Holland. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Toward Patient-Centric Digital Monitoring of Obstructive Sleep Apnea: Mixed Methods Study

James Kenneth Timmis[1,2*], PhD; Kerstin Alexandra Schorr[3*], MSc; Rana Yüksel[1], MSc; Tim van den Broek[3], MSc; Sebastiaan Overeem[4,5], MD, PhD; Dagmar Josine Smid[6], MSc; Willem Johan van den Brink[3], PhD; Nina Leonie Haring[3], PhD

[1]Department of Political Science, University of Freiburg, Freiburg, Baden-Wurttemberg, Germany

[2]Athena Institute for Research on Innovation and Communication in Health and Life Sciences, Vrije Universiteit Amsterdam, Amsterdam, North Holland, The Netherlands

[3]Research Group Microbiology and Systems Biology, Netherlands Organization for Applied Scientific Research (TNO), Leiden, The Netherlands

[4]Sleep Medicine Center Kempenhaeghe, Heeze, The Netherlands

[5]Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

[6]Research Group Human Performance, Netherlands Organisation for Applied Scientific Research (TNO), Soesterberg, The Netherlands

[*]these authors contributed equally

Corresponding Author:
Nina Leonie Haring, PhD
Research Group Microbiology and Systems Biology
Netherlands Organization for Applied Scientific Research (TNO)
Sylviusweg 71
Leiden, 2333 BE
The Netherlands
Phone: 31 627995667
Email: nina.haring@tno.nl

Abstract

Background: Obstructive sleep apnea (OSA) is a sleep disorder characterized by repeated breathing disruptions during sleep. Remote patient monitoring (RPM) of OSA is important, yet contemporary methods are limited. Sensor-based digital health technologies (sDHTs) promise a step advance in OSA RPM, but must provide meaningful, actionable, and usable outputs for patients. While the centrality of considering patient views in sDHT development is widely acknowledged, patient perspectives and priorities are rarely assessed.

Objective: This study aimed to identify patient-prioritized health aspects and preferences for digital measures and RPM to enhance OSA care quality and patient experience, guided by the digital measures that matter framework.

Methods: We used a mixed methods design combining quantitative and qualitative approaches. Individuals with a formal OSA diagnosis and persistent sleep problems (n=223) completed a survey in which they ranked items related to treatment burdens and health priorities, and responded to open-ended questions about restoring previous quality-of-life elements and desired health goals. To gain deeper qualitative insights, we conducted semistructured interviews with patients with OSA, patient advocates, and health care professionals (n=11), focusing on follow-up care, attitudes toward sDHTs and RPM, and preferences for future OSA-related sDHTs and metrics. Quantitative data were analyzed using bootstrap-aggregated Borda counts (broad support) and Plackett-Luce modeling (intense prioritization), while qualitative data from surveys and interviews were analyzed thematically.

Results: Key meaningful aspects of health included the improvement of subjective sleep quality (top-ranked burden; health goal for 46.5%, 93/200 of participants), an increase in daytime energy (quality-of-life aspect to restore for 35.6%, 72/202 and health goal for 25.5%, 51/200 of participants), and physical activity (quality-of-life aspect to restore for 24.7%, 50/202 and health goal for 16.5%, 33/200 of participants). Sleep characteristics and daytime energy were priority targets for digital measure development. Smartwatches, sleep mats, and smart rings were preferred modalities for integration into RPM. Participants' priorities for enhancing monitoring included (1) expanding metrics beyond the Apnea-Hypopnea Index (AHI; 36.6%, 52/142), (2) improving measurement accuracy (20.4%, 29/142), and (3) ensuring outputs are meaningful, understandable (18.3%, 26/142), and actionable (9.2%, 13/142). Patients also reported difficulty interpreting RPM data to determine if and when follow-up care is needed and what type of care is appropriate.

XSL•FO
RenderX

Conclusions: RPM solutions for OSA should expand beyond AHI, ensure accuracy and interpretability, and provide actionable insights to support comprehensive patient-centric management.

(J Med Internet Res 2026;28:e82460) doi:10.2196/82460

KEYWORDS

obstructive sleep apnea; digital biomarkers; remote patient monitoring; patient-centric care; wearable sensors; digital outcome measures; patient-reported outcomes

Introduction

Obstructive sleep apnea (OSA) is a sleep-related breathing disorder characterized by repeated episodes of partial or complete blockage of the upper airway during sleep, leading to oxygen desaturation events and reduced sleep quality [1]. Common symptoms include loud snoring, waking up gasping for air or choking, and excessive daytime sleepiness. OSA is a chronic condition, and approximately 14% of the global population is affected [2]. Due to aging populations and increasing rates of obesity, the prevalence of OSA is expected to rise in the upcoming years. The chronic nature of OSA, combined with its links to other health conditions and multimorbidity, imposes a high burden on patients and significantly strains health system resources [3,4].

The Apnea-Hypopnea Index (AHI) is the standard for classifying OSA severity by counting the number of apneas (complete pauses in breathing) and hypopneas (partial reductions in breathing) per hour of sleep. While AHI is widely used, it has limitations, such as not accounting for the duration and depth of breathing interruptions, or their impact on oxygen levels and sleep stages [5]. Consequently, AHI scores show poor correlation with disease burden and treatment outcomes [6,7]. Moreover, AHI lacks meaning from the patient perspective, not only because it fails to capture symptoms or daily functional impairments, but also because many patients struggle to interpret clinical measures and understand what pertinent changes in their scores actually imply for their health [8].

The most common treatment options for OSA (in the Netherlands) include continuous positive airway pressure (CPAP) and mandibular advancement devices (MADs) [9]. CPAP is routinely supported by remote patient monitoring (RPM) solutions. RPM refers to systems that allow health care professionals to assess, monitor, and care for patients virtually, often in extraclinical settings [10]. While CPAP RPM platforms provide data on use, mask leaks, and AHI scores [11], they fail to capture patient-centric outcomes, and often do not prevent nonadherence [12]. In contrast, most MADs lack embedded sensors, so adherence and treatment effectiveness are typically assessed subjectively during follow-up visits [13].

Patient-reported outcome measures (PROMs) offer invaluable insights into the subjective experiences of individuals with OSA, especially regarding disease-related quality of life [14]. However, PROMs are limited by respondent burden and various biases, such as nonresponse, fatigue, and recall bias [15,16]. Sensor-based digital health technologies (sDHTs) can be defined as (often wearable) devices that use sensors (for instance, accelerometers or photoplethysmography) to capture health measures, such as symptoms and functional states, continuously [17,18]. They have the potential to (partially) replace, or complement, existing PROMs and thereby provide more objective, real-time insights into patients' health [19,20]. In the context of OSA, sDHTs may be integrated directly into treatment devices or used independently (eg, watches and sleep mats). By passively and continuously capturing objective, longitudinal data based on digital biomarkers, sDHTs also hold potential to facilitate the prediction of treatment responses, optimization of titration, and enhancement of adherence. This supports the transition toward person-centered OSA care, which empowers patients to manage their own condition [21]. Additionally, patient-centric digital end points are becoming increasingly important for clinical research [22].

However, studies across a wide range of different settings have shown that, for (new generations of) technologies to be properly adopted, patients must consider them meaningful, actionable, and usable [18,22-25]. Although the importance of considering patient priorities in sDHT development is widely acknowledged, the former are rarely assessed and integrated in new sDHTs [26]. Efforts led by the Digital Medicine Society (DiMe) are paving the way for the development of sDHTs that are truly patient-centric [18,23]. According to DiMe's digital measures that matter framework, this process begins with the identification of meaningful aspects of health—aspects of a condition that patients wish to improve, arrest, or prevent—which then guide the selection of measurable concepts that reflect patients' lived experiences and priorities [18]. Noteworthy is that the perspectives of patients with OSA are rarely reported in scientific literature, and we are not aware of literature published on the priorities of patients with OSA for sDHTs. To contribute toward addressing this knowledge gap, this study aimed to identify meaningful aspects of health, patient and clinician priorities, and preferences regarding sDHTs for OSA management using a mixed methods design.

Methods

Study Design

This observational, exploratory study used a sequential mixed methods design, integrating survey-based quantitative data with qualitative data collected in semistructured interviews to gather insights from individuals with OSA reporting persistent sleep problems and health care professionals involved in OSA care. Quantitative data were collected through a survey completed by n=223 respondents (which is a sample size considered sufficient in published studies with similar study aims, designs, and settings) [27-29], between January and March 2024, while qualitative insights were obtained from semistructured interviews conducted in May-June 2024 with 6 patients, 2

XSL•FO
RenderX

patient advocates, and 3 health care professionals (Table S1 in Multimedia Appendix 1).

Survey

The survey targeted a broad population of individuals with self-reported sleep problems; participants were included who were aged 18 years and older and experiencing sleep problems at least 3 times per week for a minimum of 3 consecutive months. Moreover, we included in this analysis only those respondents who reported receiving a formal OSA diagnosis from a health care professional. Survey participants were recruited via social media posts, flyers distributed in primary and secondary sleep care settings (including physiotherapy practices and sleep clinics), and a newsletter announcement by the Dutch Apnea Association (ApneuVereniging). The survey (originally developed in Dutch and translated into English for this manuscript) consisted of 18 questions (multiple choice, ranking, and open-ended) distributed through the online platform Survalyzer (Multimedia Appendix 1). Respondents were only able to participate after providing informed consent. The survey included 3 themes. The first theme covered demographic and background information, including medical history and sleep disorder profile (questions 1-8). Questions 1 to 4 were adapted from the "Netherlands working conditions" survey conducted by the Netherlands Organization for Applied Scientific Research (TNO) and Statistics Netherlands [30]. Questions 5 to 8 were developed based on relevant literature and expert input from sleep health care professionals. The second theme focused on meaningful aspects of health (questions 9-12), with questions developed using the digital measures that matter framework by Manta et al [18], which provides patient-centered question formulations to identify aspects of health most meaningful to individuals. The third theme explored preferences and experiences with sDHTs, adapted to the sleep field from a survey exploring this theme in patients in the cardiovascular risk management care pathway, using expert input and supporting literature. The original survey is currently being prepared for publication. In open-ended questions, participants were asked to provide details on, for example, health goals or the aspects that positively drove their quality of life (before their development of OSA), which they would like to restore. Participants had no word limit. While pretesting is a standard step to ensure clarity and reliability, it was not feasible in this instance due to time constraints. Face validity of the survey was assessed by an internal panel of experts experienced in survey design.

Survey Data Analysis

The data analysis for this study primarily involved descriptive statistics, including the calculation of frequencies and percentages to summarize the data and identify patterns and trends within the dataset. Ranking items were analyzed using both Borda counts and the Plackett-Luce model. Using both methods allowed us to combine the accessibility of Borda counts with the statistical rigor of Plackett-Luce and to assess consistency across approaches. The Borda count is a point-based voting method in which each item receives a score based on its rank position, with scores aggregated across participants to produce a consensus ranking [31]. We included Borda counts because they provide a simple and easily interpretable descriptive summary of rankings that has been widely used in previous work. Within each question, we calculated mean scores and 95% CIs. This method assumes complete rankings from all participants. However, some ranking questions in our survey elicited partial rankings, as participants were asked to rank only their top 3 items from a larger set, potentially introducing bias. To address this, the Plackett-Luce model was used as a complementary approach, as it does not penalize unranked items [32]. The Plackett-Luce model estimates the relative *worth* or preference strength of each item based on observed rankings, including partial ones. Each item is assigned a positive worth parameter, which is interpreted comparatively: an item with a higher worth than others is more likely to be systematically preferred. Observations were treated as independent despite potential within-subject correlation, as sparse data precluded models accounting for this, and results should be interpreted accordingly. All computations were performed in R (version 4.4.0; R Foundation for Statistical Computing) using the *PlackettLuce* (version 0.4.3) and *emmeans* (version *1.11.1*) packages.

Open-ended survey responses were analyzed using thematic analysis as outlined by Braun and Clarke [33]. Coding was performed by one researcher and independently reviewed by a second researcher to ensure consistency. Themes were developed based on the frequency, emphasis, and contextual richness of participant responses. In some cases, subdividing themes into subthemes was necessary to provide a deeper understanding of the data.

Interviews

We used purposive sampling to select adequate participants for the interviews. The inclusion criteria were as follows: Dutch patients who have been formally diagnosed with OSA and have experience with (digital) RPM for OSA (current or discontinued use; the latter, to elicit challenges for general use and adherence); health care professionals with a specialization in OSA and who have experience with (digital) RPM for OSA, to elicit their perspective on clinical workflows and patient interactions; and patient advocates for OSA, to elicit perspectives on the pain points and needs of the broader community of patients with OSA in the Netherlands. Our sample included individuals from (1) patients with OSA from the survey who provided email addresses for follow-up, (2) TNO's network from previous OSA collaborations, and (3) clinicians identified through online searches for OSA expertise. The recruitment email included detailed information on the study, including the background of the research, its objectives, and the informed consent form. If individuals agreed to participate, they were able to choose between an in-person or online interview via Microsoft Teams.

The interviews, which on average took about 60 minutes, were conducted and recorded via Microsoft Teams by RY, who used an interview guide that had been expert checked (by NLH and JKT) and piloted a priori (Multimedia Appendix 1). The interview guide was based on a conceptual framework closely aligned with the technology acceptance model, which had been adapted to systematically elicit the perceived usefulness and

XSL•FO
RenderX

ease of use of, perceived needs for, and willingness to engage with OSA RPM and pertinent communication and reporting mechanisms in inter alia OSA follow-up care. The conceptual framework was also used to create the initial deductive code book.

Qualitative Data Analysis

The interviews were manually transcribed (verbatim) by RY. To support analysis, the transcripts were imported into the computer-assisted qualitative data analysis software (CAQDAS) Atlas.TI (version 8.02; Scientific Software Development). The transcripts were analyzed by RY based on the 6 steps of thematic analysis using deductive and inductive (complementary) coding approaches as outlined by Braun and Clarke [33]. To improve the credibility of the analysis, 10% of the transcripts (in this case, n=2 interviews) were independently coded by DS, see acknowledgments. We determined intercoder reliability (approximately 90%) by manually reviewing the codes created and assigned by both coders for both interview transcripts. To increase dependability, codes were carefully tabulated and aggregated into themes based closely on the initial conceptual framework, analyzed and discussed with NLH and JKT, and finally reported by RY. RY also performed member checks (providing a summary of the preliminary analysis of an interview together with a couple of key quotes to the pertinent interviewer, and asking for comment) with multiple interviewees to improve confirmability. The code book and coding were reviewed by another researcher. Finally, the process of selective coding focused on selecting the most important and representative codes to develop overarching themes that addressed the research subquestions of this study [34]. Atlas.Ti was used for the coding process. While the number of possible interviews was limited by time and resource constraints, we, by tendency, reached data saturation in interview 7. However, one further major concept was revealed in interview 9. Interviews 10 and 11 revealed no additional concepts. In consequence, data saturation can therefore not be considered formally achieved. To enhance rigor, we used member checks, peer debriefing, and strategies to minimize social desirability and interview bias (eg, building rapport). For the qualitative component of this study, we carefully considered the 4 established quality criteria of trustworthiness in qualitative research, namely credibility, transferability, dependability, and confirmability, to enhance the overall quality of the study [35].

Ethical Considerations

This study was reviewed in accordance with institutional and national ethical standards for research involving human participants. The research protocol was submitted to TNO's ethical review board, and ethical approval was obtained from TNO's ethical review board for both the survey (study 2023-103) and the interviews (study 2024-026). Informed consent was obtained from participants before data collection. Participants received no compensation for participation. Data used in this study were anonymized, and no personally identifiable information was retained. All data were stored on secure, access-controlled servers in compliance with data protection regulations. No identifiable images or personal information of participants are included in this manuscript or supplementary materials. Ethical approval was further granted by TNO's ethical review board for making data available in a repository (study 2023-103).

Data Management and Availability

The datasets generated and analyzed during this study are publicly available in the Harvard Dataverse repository under the title "Replication Data for: Toward Patient-Centric Digital Health Solutions for Obstructive Sleep Apnea Monitoring: Perspectives from Dutch Patients and Healthcare Professionals – a mixed-method study" [36]. The repository includes an anonymized survey dataset (n=223) containing quantitative responses on meaningful aspects of health, attitudes toward remote patient monitoring, and digital health technology preferences. To protect participant confidentiality, direct identifiers have been removed, and free-text fields have been redacted to avoid inadvertent disclosure of personal information.

Code Availability

The analysis code used to generate the quantitative results reported in this study is publicly available in the same Harvard Dataverse repository [36].

Reporting Guideline

This study adhered to the Mixed Methods Reporting in Rehabilitation and Health Sciences guideline, and the completed checklist is provided in Multimedia Appendix 2.

Results

The survey focused on the themes of meaningful aspects of health, current use of and attitudes toward sDHTs or RPM, and preferences for future OSA-specific sDHTs and RPM solutions. To gain deeper qualitative insights, we also conducted semistructured interviews with formally diagnosed patients with OSA, patient advocates, and health care professionals. The interviews covered the themes follow-up care, attitudes toward sDHTs and RPM, and preferences for future OSA-related sDHTs and metrics; Table S2 in Multimedia Appendix 1.

Demographic Information

A total of 404 individuals initiated the survey; after applying eligibility criteria, the final analytic sample comprised 223 Dutch patients with a formal OSA diagnosis (Multimedia Appendix 3). A total of 48.4% were female, and the mean age was 65 (SD 9) years (Table 1). The majority of the sample (133/223, 59.6%) was highly educated, and 67.3% (150/223) of the participants worked for less than 1 day per week or not at all. A total of 34.1% (76/223) were formally diagnosed with at least one other sleep-related illness besides OSA. Also, other comorbid conditions were reported by 71% (158/223), with obesity and cardiovascular disease being the most prevalent. All interview participants were Dutch and purposefully selected; they had to be at least 18 years of age and have preexisting experience with OSA RPM solutions (Table S1 in Multimedia Appendix 1). We interviewed 6 Dutch individuals with a formal OSA diagnosis, 2 patient advocates affiliated with the national association for patients with OSA (ApneuVereniging), and 3 health care professionals experienced with managing OSA. This purposive sample was designed to capture a range of

XSL•FO
RenderX

perspectives from key stakeholder groups (patients, advocates, and clinicians; Table S2 in Multimedia Appendix 1).

Table 1. Respondent characteristics of surveyed cohort (n=223).

Respondent characteristics	Values
Age (years), median (range)	67 (28-86)
Sex, n (%)	
Male	48 (108)
Female	52 (115)
Education, n (%)	
Secondary education	6.7 (15)
Secondary vocational education	33.6 (75)
Higher professional education and academic education	59.6 (133)
OSA care provider visited within the last year, n (%)	
General practitioner	38 (84)
Medical specialist (outside sleep clinic)	50 (111)
Company doctor	6 (13)
Psychologist	6 (13)
Sleep therapist	4 (9)
Sleep clinic	27 (59)
Other	3 (6)
None	30 (67)
Sleep-related diagnoses, n (%)	
Obstructive sleep apnea	100 (223)
Insomnia	10 (22)
Hypersomnia	12 (27)
Sleep rhythm disorder	1 (2)
Parasomnia	4 (10)
Sleep-related movement disorder	15 (34)
Other	3 (6)
Other diagnoses, n (%)	
Obesity	33 (73)
Cardiovascular disease	37 (82)
Diabetes type 2	15 (34)
Depression	9 (20)
Other	18 (42)
None	29 (64)

Meaningful Aspects of Health

Overall, we found that meaningful aspects of health for individuals with OSA and persistent sleep problems encompass both physical limitations and psychological burdens. Subjective sleep quality, daytime energy levels, and physical activity consistently emerged as key priorities—highlighted as important, considerable burdens when impaired, and as goals for improvement or resumption. In addition, psychological concerns such as worrying about health impacts and difficulties concentrating reflect the broader mental burden of OSA. We present results from ranking exercises of prespecified health aspects and, separately, thematic analyses of responses to open-ended questions.

Burdens of Living With OSA (Ranking)

Worrying about OSA health impacts, sleep interruptions, and problems concentrating were ranked highest (Figure 1, parts A and B). Based on Borda counts, these items showed broad support across the cohort, while Plackett-Luce modeling revealed that certain concerns—such as falling asleep while driving or problems concentrating – were intensely prioritized

XSL•FO
RenderX

by subsets of respondents. Being too tired for hobbies or sport was also considered an important burden (fourth place for both methods).

General Health Priorities (Ranking)

While physical fitness attracted the strongest support overall, it came in second regarding rank concentration (Figure 1, parts C and D). This means that while physical fitness was considered most important overall, the agreement on specific ranks selected was less consistent than, for example, mental and emotional well-being or chronic disease management. Subjective sleep quality was ranked second overall, and first, based on rank agreement (Figure 1, part D), so the reversed relationship was observed when compared with physical fitness. Subjective sleep quality refers to the experienced quality of sleep as perceived by the individual, including aspects such as sleep fragmentation, feelings of restorative sleep, and, more generally, what respondents described as "good" sleep without providing a fixed definition. Mental and emotional well-being, chronic disease management, and diet came in third, fourth, and fifth, respectively, for both analytical methods, respectively.

Figure 1. Meaningful aspects of health identified through survey questions. Parts A and B were derived from a 3-item ranking question from 9 prespecified items (n=221). Parts B and C were derived from a 2-item ranking question from 5 prespecified items (n=219). OSA: obstructive sleep apnea.

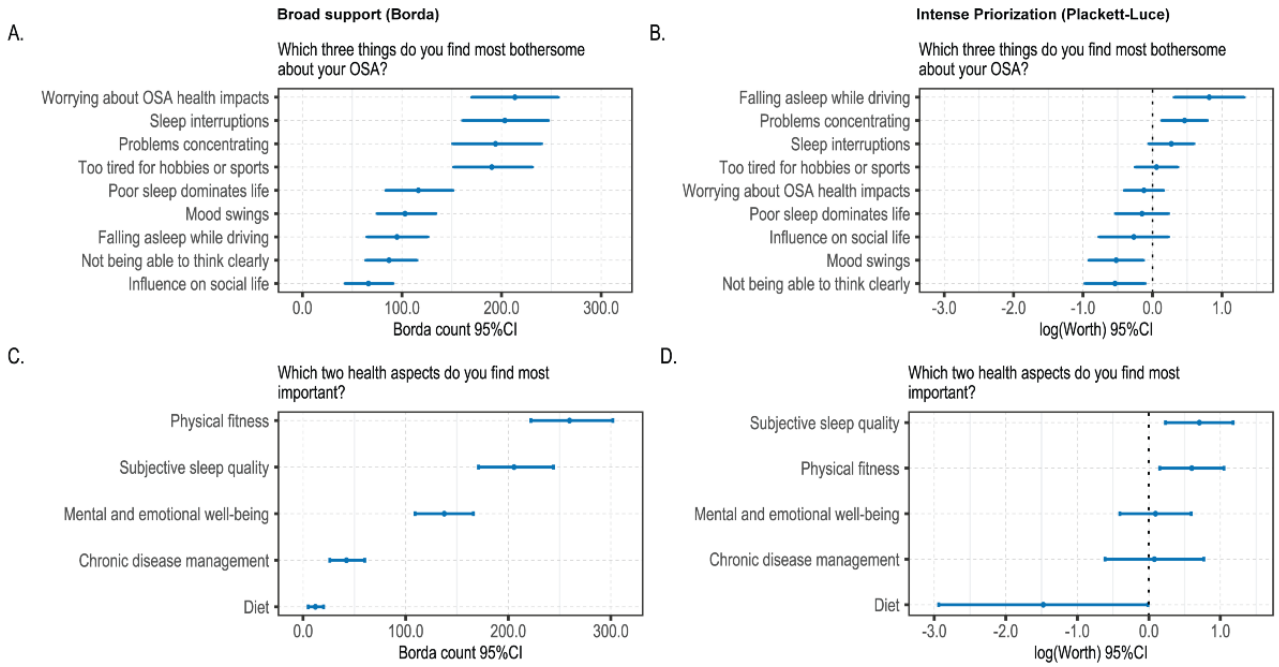

Restoring Previous Quality of Life (Open-Ended)

A broad desire for higher daytime energy levels was expressed by 35.5% (72/202) of respondents (Table 2). For example, one participant reported that they missed "getting through a whole day without constraints from extreme tiredness." Another explained: "I still do all things normally, only it [takes much longer] because you are so terribly tired." Various types of physical activities were described by 24.8% (50/202) of respondents as aspects that they missed from their quality of life before disease onset, such as hiking, cycling, or exercise more generally. Interestingly, "None" was the third most common theme (17.8%, 36/202), reflecting respondents who reported nothing from their previous quality of life they wished to restore. Social activities (16.8%, 34/202), activities requiring concentration, such as reading a book (15.8%, 32/202), and restorative sleep (14.9%, 30/202) were also mentioned.

XSL•FO
RenderX

Table 2. Aspects of their previous quality of life that respondents wished to restore (derived from open-ended questions; n=202).

Items	Respondents, n (%)
Having more energy	72 (35.6)
Physical activity	50 (24.8)
None	36 (17.8)
Social life and leisure	34 (16.8)
Concentration	32 (15.8)
Subjective sleep quality	30 (14.9)
Feeling more productive	22 (10.9)
Daily activities	16 (7.9)
Physical health	16 (7.9)
Being intervention-free	3 (1.5)

Health Goals Relating to OSA Symptoms (Open-Ended)

As shown in Table 3, "Subjective sleep quality" was the top health goal for our sample (46.5%, 93/200). For example, one participant explained that their goal was: "no more lying awake after going to the toilet at night." The themes weight loss and daytime energy levels both came in second (25.2%, 51/200). Participants explained their desire to feel less tired and more energetic throughout the day. The third most frequently mentioned aspect was physical activity (16.3%, 33/200). Additionally, 14.4% (29/200) of participants wish to improve "health aspects related to physical health," including, for example, addressing night sweats, palpitations, or restless leg syndrome.

Table 3. Health goals related to obstructive sleep apnea symptoms derived from open-ended questions (n=200).

Items	Respondents, n (%)
Subjective sleep quality	93 (46.5)
Daytime energy levels	51 (25.2)
Weight loss	51 (25.2)
Physical activity	33 (16.3)
Physical symptoms	29 (14.4)
Improved treatment	22 (10.9)
Concentration	16 (7.9)
Mental health	13 (6.4)
Social life and leisure	8 (4)
Other	5 (2.4)
None	3 (1.5)

OSA Follow-Up Care in the Netherlands

Based on the survey data, half of the participants reported contact with medical specialists, and 38% with general practitioners, respectively. A total of 30% (67/223) indicated that they had no physical follow-up health care visits. 26.4% (59/223) indicated that they had visited sleep centers during the past year (Table 1). Notably, several respondents who selected the option "other" explained that they had not attended physical follow-up visits but were, instead, relying on remote monitoring by their CPAP devices or self-monitoring of their condition. For example, one participant explained:

> *No. I haven't seen a specialist since I got the [CPAP device]. According to the pulmonologist, everything was under control, and there was no need to return.*

According to some interviewees, follow-up care is straightforward but limited, as exemplified here by the following comment:

> *[I was] diagnosed [with] sleep apnea [and] provided...with the CPAP mask. A year later, I had a follow-up appointment...to ensure my mask fit well and that everything was in order. After that, I received no follow-up care anymore.* [Participant B8; patient with OSA who is also a general practitioner]

B3 (patient with OSA) highlighted that they were not even sure what type of care they should be receiving: "I received no follow-up care. I don't even know what follow-up care you should get." Participant B11 (a patient advocate) explained that this was likely due to resource constraints:

XSL•FO
RenderX

> *Clinics don't have the capacity to provide follow-up care for all these patients. There are long waiting lists, and those waiting for OSA diagnosis are prioritized, meaning that those who should receive follow-up care need to wait.* [Participant B11; patient advocate]

By the same token, the two interviewed somnologists highlighted that, on the one hand, patients have a role in taking charge if issues occur, and, on the other, follow-up care can indeed vary greatly based on patients' needs and the severity of their OSA. Participant B7 (a somnologist) explained, "Doctors expect that patients will initiate contact if there's a problem....Effective communication and reporting in follow-up care...also requires patients to allocate time." Participant B9 (pulmonologist/somnologist) provided a top-level view of her patient-centered approach to follow-up care:

> *I do an initial evaluation after 6-8 weeks post-diagnosis, then plan further check-ups based on the patient's condition and needs. If a patient is doing well, I schedule a follow-up after a year..... If there are any issues, I see them sooner, using RPM data from their CPAP device to guide decisions.* [Participant B9]

Nevertheless, participant B5 (a patient with OSA) explained, other hurdles might exist in achieving good continuity of care: "I called my supplier, who said that they had sent a report of my monitored data to the hospital, but the hospital claimed they that hadn't received any report from my supplier."

sDHTs Used and Attitudes Toward RPM

Based on the survey data, 86% (182/212) of participants agreed that sDHTs could contribute toward improving management of their OSA. Table 4 shows that, in our sample, the most frequently used device for self-monitoring was CPAP (87.1%, 155/178), followed by weight scales (47.8%, 85/178), blood pressure monitors (46.6%, 83/178), and smart watches (35.4%, 63/178 for heart rate measurements and 19.7%, 35/178 for sleep tracking). Sleeping mats were used least frequently (n=1). Figure 2, parts A and B, shows that, when asked to rank different form factors according to their preference, smart watches were ranked first, both in terms of overall frequency and subgroup concentration. The least preferred technology was clothing with integrated sensors. Notably, only a single participant reported using a sleep mat, yet when participants were asked to rank their preferred technologies, sleeping mats were ranked among the top three. While the majority of participants (67%, 119/178) agreed that both they themselves, as well as their care providers, should have access to data collected with sDHTs, 28% (49/178) thought their care providers should have access only under certain conditions—merely 4% (7/178) indicated that they should be the only party with access to their data.

Also, the majority of our interviewees held positive views toward the usability of RPM technologies for supporting OSA management. Participant B6 (a patient with OSA), for example, was particularly enthusiastic: "RPM tools [are] a great addition because of the shortage of health care professionals now, so I think it's fantastic that we're focusing on that." By the same token, some patients with OSA had concerns regarding the dependability of the RPM ecosystem and the actionability of information contained therein. For example, a patient with OSA stated:

> *I appreciate RPM, but it's only effective if I receive feedback from the RPM technology supplier or hospital. Ideally, it should alert you promptly if issues like stopped breathing arise, which can ensure constant surveillance and timely intervention.* [Participant B3; patient with OSA]

Moreover, there exists an unmet need regarding the interpretation of data: A patient with OSA explained:

> *I do see my sleep apnea values, for example, an AHI score of 11.3, [but] is that good? What does this score mean for my health, and how can I improve it? I'm missing this information, and I think many do.* [Participant B4; patient with OSA]

Table 4. Currently used technologies to monitor health are derived from multiple-choice questions (n=178).

Items	Respondents, n (%)
CPAP[a]	155 (87.1)
Weight scale	85 (47.8)
Blood pressure monitor	83 (46.6)
Smartwatch (heart rate)	63 (35.4)
Pulse oximeter	46 (25.8)
Smartwatch (sleep)	35 (19.7)
Health app	21 (11.8)
Glucose monitor	16 (9)
Other	5 (2.8)
Sleep mat	1 (0.6)

[a]CPAP: continuous positive airway pressure. Images of devices are shown in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 2. Form factor preferences. Parts A and B were derived from the full ranking of 7 prespecified items (n=109).

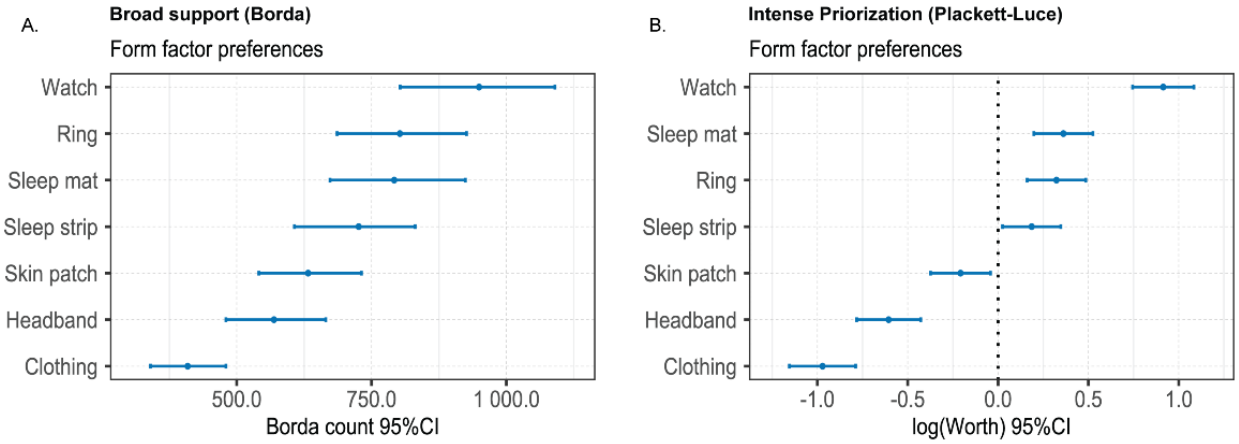

This was somewhat echoed, albeit in more general terms, by the somnologist (Participant B7), who was of the view that current RPM solutions have not yet reached the degree of maturity required for effective OSA remote management:

> *Follow-up care cannot be provided as effectively through digital health technologies like RPM compared to clinical care. As a doctor, you need to see, touch, and smell a patient to provide optimal care. You miss the human aspect of care when everything is digitized and managed through technologies that provide RPM.* [Participant B7]

Nevertheless, our data indicate that some patients are actively engaged with their RPM data. Participant B2 (a patient with OSA) explained:

> *I read my reported data from the screen of my device. Then I have the DreamMapper app [patient-facing app showing CPAP data]. I take out the SD card of my [CPAP] device and plug it into my computer, on which I have installed the Oscar program. This program gives me a comprehensive report on my monitored sleep apnea. Oscar is usually a very good program, but the hospitals don't want to use it because it's not validated.* [Participant B2]

This quote quite vividly illustrates that, while solutions exist that can be meaningful for patients with OSA who want to dive deeper into their data and better understand their disease, these might be located outside of traditional clinical care pathways (and therefore lack clinical oversight and supervision).

Improving Digital Measures for the Future of OSA RPM

Participants were asked to rank which three digital measures would help them gain more insight into their OSA. Sleep characteristics ranked highest overall. However, broad support in the Borda count analysis and rank agreement in the Plackett-Luce model for both sleep characteristics and daytime energy levels (second rank) indicate that these are the top targets for future digital measures (Figure 3, parts A and B). Physical activity was the only other item that came in at the identical rank (8) for both methods of analysis. When asked how current OSA measurement practices could be improved, the most frequently given answer related to suggestions for additional measurements (mentioned by 52/142, 36.6% of respondents; Table 5). Specifically, the majority of survey participants were interested in OSA vitals beyond AHI, such as heart rate–derived and saturation-based measures. Obtaining insights into sleep characteristics, such as sleep stages, was also mentioned by several survey participants. Other suggestions included improved accuracy and reliability of measurements, and a more comprehensible presentation of monitoring data. For example, one survey participant highlighted the need for more clarity in analysis reports, such as presentation in layman's terms, whereas another patient expressed the desire for more monitoring and guidance from OSA professionals and home care providers. Also, the more general value of digital measures for OSA monitoring was underscored. For example, one survey participant indicated that they would like to add monitoring options to their MAD-based treatment:

> *Many people have MADs but have no proof – except through complaints (snoring) or feeling fit – [indicating] whether it helps..... It would be great if...[a wearable or smart watch], or other technology could do that* [track and report]

XSL•FO
RenderX

Figure 3. Health aspects participants would like to understand better. Parts A and B were derived from a 3-item ranking of 9 prespecified items (n=213). OSA: obstructive sleep apnea.

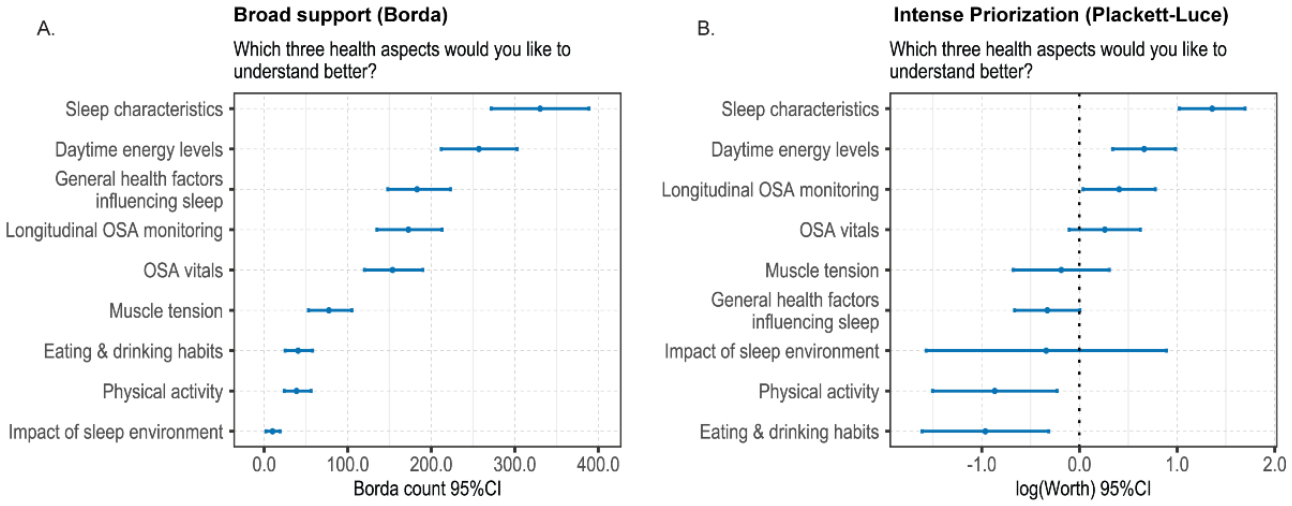

Table 5. Suggestions for improving obstructive sleep apnea remote patient monitoring. Thematically analyzed responses to an open-ended question. Obstructive sleep apnea management advice includes lifestyle advice and contact with a health care provider (n=142).

Items	Respondents, n (%)
Additional measurements	
All	52 (36.6)
OSA[a] severity measures[b]	21 (40.4)
Sleep characteristics	16 (30.8)
Metabolic health	11 (21.2)
Unspecified	4 (7.7)
None	32 (22.5)
Accuracy and reliability	29 (20.4)
More comprehensible presentation	26 (18.3)
OSA management advice	13 (9.2)
Product design	5 (3.5)
Other	6 (4.2)

[a]OSA: obstructive sleep apnea.

[b]OSA severity measures: heart rate–derived measures, breathing, saturation.

The interview data provided additional detail on a range of potential improvements for future OSA management. The majority of patients with OSA preferred noninvasive sDHTs and emphasized the need for a stronger product design focus on less disruptive and more convenient devices for daily use, and better integration with daily attire. A patient with OSA explained:

> *I already use a [smart watch]...for my heart monitoring, but it would be nice if future RPM technologies are non-invasive, like an App or a small device beside the bed. Large or intrusive devices are uncomfortable for light sleepers like me.* [Participant B6; patient with OSA]

Participant B4 (a patient with OSA) also mentioned the, inter alia, visual product design of wearable RPM technology: “A smartwatch-like device that tracks my heart rate and oxygen saturation would be ideal, as long as it’s non-invasive and looks good.” Regarding specific measures, participant B4 (a patient with OSA) said: “I would like to measure my oxygen saturation during the night. This could be an improvement regarding RPM metrics for OSA.”

Discussion

Principal Findings

This mixed methods study, guided by DiMe’s digital measures that matter framework, aimed to, first, identify meaningful aspects of health and care needs of individuals with OSA, and, second, inform the improvement of existing, and the development of new, pertinent digital measures and RPM

XSL•FO
RenderX

solutions. Based on our two Dutch cohorts (patients, their representatives, and specialized health care professionals), our principal findings are that, first, improving subjective sleep quality, increasing physical activity, and increasing daytime energy levels are key meaningful aspects. Second, sleep characteristics (particularly sleep fragmentation), daytime energy (especially fatigue and excessive daytime sleepiness), and nighttime oxygen saturation are priority targets for digital measure development. Third, smartwatches, sleep mats, and smart rings are strongly preferred as modalities for sDHTs that can be integrated into future RPM solutions. Fourth, current digital monitoring practices should be enhanced by focusing on expanding metrics beyond AHI, improving measurement accuracy, and ensuring that digital measures are meaningful, understandable, and actionable for end users. Finally, patients lack the ability to determine from RPM output whether they need to seek follow-up care and, if so, what type of care is appropriate.

Comparison With Previous Work

Our principal findings highlight that improving subjective sleep quality, increasing physical activity, and enhancing daytime energy levels are key meaningful aspects of health for patients with OSA. Each can (potentially) be assessed using sDHT-derived metrics, though these technologies vary widely in maturity and clinical readiness.

Sleep was a consistently prioritized health aspect, essential to daily functioning and overall well-being. Participants expressed skepticism about the accuracy and clinical utility of sleep data from personal devices, echoing World Sleep Society recommendations that, while sDHTs hold potential for monitoring sleep patterns, their proprietary algorithms often lack validation in sleep disorders [37]. Recent studies in OSA populations show promising results for tracking metrics like total sleep time and sleep efficiency, with some devices showing moderate-to-strong agreement with polysomnography; however, broader validation is still needed to ensure accuracy and clinical relevance [38-42].

Physical activity emerged as another key health aspect, with participants identifying it as a health priority and an area they wished to improve. Physical activity is both a meaningful health goal and a key factor in OSA management, as increasing activity can help reduce OSA severity while effective treatment may, in turn, support higher activity levels. Structured exercise interventions can reduce AHI, improve oxygenation, and alleviate daytime sleepiness [43-45]. However, evidence for physical activity levels as an outcome of CPAP therapy is mixed, with studies reporting, despite improved symptoms, modest or no changes in physical activity levels [46-48]. sDHTs could help monitor physical activity as a potential biomarker of functional gains and support behavior change in OSA care.

Daytime energy levels emerged as a top-rated health aspect, with participants identifying it as a key priority, an important health goal, and the most valued quality-of-life factor to restore. This reflects both excessive daytime sleepiness and fatigue, which present heterogeneously across patients and persist in some even after optimal therapy [49-51]. While these symptoms are typically assessed using PROMs, such as the Epworth Sleepiness Scale and Fatigue Severity Scale, their use is limited by respondent burden, recall bias, and low temporal resolution [51]. sDHTs offer a promising, noninvasive alternative for continuous monitoring of these symptoms. Metrics, such as heart rate variability and physical activity, are actively being explored, yet validated digital measures for routine monitoring are lacking—making this a clear priority for innovation [51,52].

Implications for Practice and Policy

For Developers of sDHTs and RPM Solutions

Developing new digital measures and RPM solutions is complex and resource-intensive, often requiring significant time before their impact reaches clinical practice. In the short term, the rapid rise of consumer wearables creates an opportunity to validate and optimize digital measures—such as sleep and physical activity—for specific populations like people with OSA. This should be prioritized as a practical step toward innovation. Beyond creating new measures, improving existing RPM tools is equally urgent. Enhancing reporting platforms with clear explanations, contextualized feedback, and actionable suggestions could better support patients in managing their condition. Interfaces that let users toggle between simplified and detailed views would help accommodate diverse preferences and levels of health literacy.

For Policymakers and Payers

On a broader scale, policymakers and payers should consider setting standards to ensure RPM systems deliver accurate, patient-centric, harmonized, and actionable data presentations. A key requirement is that digital measures embedded in RPM solutions are validated for accuracy, reliability, usability, and clinical relevance in target populations, such as patients with OSA [22,23]. Establishing such validation criteria will help ensure that these tools provide meaningful insights and support clinical decision-making. These standards could also guide reimbursement decisions and accelerate the adoption of RPM solutions that empower patients, improve treatment outcomes, and reduce health care burden.

For Clinical Practice

Our findings underscore several opportunities to strengthen clinical care. First, follow-up care for patients with OSA is often limited and fragmented, highlighting the need for better alignment between home care providers and clinicians and a clearer definition of follow-up pathways. Telemedicine offers a promising, cost-effective avenue to facilitate structured and timely follow-up. Second, patients frequently report difficulties in interpreting RPM data; clinicians and home care providers should therefore provide guidance to ensure patients understand the outputs from their CPAP. Finally, as patients increasingly adopt self-monitoring solutions outside formal care pathways, clinicians should be supported in evaluating the reliability of these tools and integrating relevant patient-generated data into care when appropriate.

Limitations and Strengths

Our study has several limitations. First, the sample was skewed toward more highly educated individuals and included a higher proportion of women, likely reflecting the greater prevalence

XSL•FO
RenderX

of comorbid insomnia in female patients with OSA [53-55]. Second, the survey was not pretested, which might impact its validity. Third, the number of participating health care professionals was limited to 3 (1 somnologist, 1 somnologist/pulmonologist, and 1 general practitioner), which means their perspectives may not fully capture the diversity of clinical views and could reflect individual experiences. However, several of their insights were echoed by patients, supporting their relevance to the themes identified. Furthermore, the surveyed cohort focused on individuals with OSA who report persistent sleep problems despite treatment. While this subgroup may not fully represent the broader OSA population, it highlights a group with substantial unmet needs and provides valuable insights into priorities for digital health innovations and RPM solutions. Another strength is that our patient-centric approach aligns with value-based health care principles and was guided by the digital measures that matter framework to identify outcomes that matter most to individuals with OSA. The combination of quantitative survey data and qualitative interviews, including input from patient representatives, ensured the perspectives captured reflect a broad OSA population.

Conclusion

Rather than relying solely on clinical end points such as AHI, our findings suggest that outcomes such as physical activity, restorative sleep, and daily functioning are central to patients' lived experiences—and are therefore critical targets for sDHTs and RPM metric innovation. Developing and validating new digital measures that capture these experiences will require time, interdisciplinary collaboration, and ongoing involvement of patients to ensure relevance and usability. In the meantime, existing RPM systems can be strengthened by improving transparency, accessibility, and contextual interpretation of currently collected data, making these platforms more meaningful and actionable for patients. The broad patient priorities identified in this study can serve as an excellent starting point for defining patient-centric digital measures, allowing for comprehensive disease management.

Acknowledgments

The authors would like to thank Marian Schoone, Suzan Wopereis, Hardy van der Ven, Koen Hogenelst, Elsbeth de Korte, and André Boorsma for their valuable contributions and insightful discussions, which greatly supported the development of this work. Portions of the text were revised with the assistance of the generative AI language model ChatGPT (GPT-5; OpenAI), and all content was subsequently reviewed and approved by the authors.

Data Availability

The datasets generated and analyzed during this study are publicly available in the Harvard Dataverse repository under the title "Replication Data for: Toward Patient-Centric Digital Health Solutions for Obstructive Sleep Apnea Monitoring: Perspectives from Dutch Patients and Healthcare Professionals – a mixed-method study" [36].

Funding

This research was supported by internal funding from the Netherlands Organization for Applied Scientific Research (TNO). The article processing fee was covered by TNO's internal funds. No external grants or commercial funding were received for this work.

Authors' Contributions

Conceptualization: NLH, WJvdB
Data curation: NLH, KAS, RY, TvdB, DJS
Formal analysis: NLH, KAS, RY, TvdB, DJS
Funding acquisition: WJvdB
Investigation: NLH, RY
Methodology: NLH, WJvdB, TvdB, JKT
Project administration: NLH
Software: KAS, TvdB
Supervision: JKT, NLH
Validation: WJvdB, DJS
Visualization: KAS, TvdB
Writing—original draft: JKT, KAS, NLH
Writing—review and editing: SO, JKT, NLH

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Overview of themes covered in mixed methods components, characteristics of interviewed cohort (n=11), survey (original in Dutch, for this manuscript translated to English), and interview guide (original in Dutch, for this manuscript translated to English).
[DOCX File , 519 KB - jmir_v28i1e82460_app1.docx]

Multimedia Appendix 2
Mixed Methods Reporting in Rehabilitation and Health Sciences checklist.
[PDF File (Adobe PDF File), 697 KB - jmir_v28i1e82460_app2.pdf]

Multimedia Appendix 3
Survey response flow diagram. The final sample was defined as respondents who completed demographic questions (Q1-8) and at least one study question (Q9-18).
[PNG File , 12 KB - jmir_v28i1e82460_app3.png]

References

1. Lv R, Liu X, Zhang Y, Dong N, Wang X, He Y, et al. Pathophysiological mechanisms and therapeutic approaches in obstructive sleep apnea syndrome. Signal Transduct Target Ther 2023;8(1):218 [FREE Full text] [doi: 10.1038/s41392-023-01496-3] [Medline: 37230968]
2. Benjafield AV, Ayas NT, Eastwood PR, Heinzer R, Ip MSM, Morrell MJ, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. Lancet Respir Med 2019;7(8):687-698 [FREE Full text] [doi: 10.1016/S2213-2600(19)30198-5] [Medline: 31300334]
3. Heatley EM, Harris M, Battersby M, McEvoy RD, Chai-Coetzer CL, Antic NA. Obstructive sleep apnoea in adults: a common chronic condition in need of a comprehensive chronic condition management approach. Sleep Med Rev 2013;17(5):349-355. [doi: 10.1016/j.smrv.2012.09.004] [Medline: 23434125]
4. Ehsan Z, Ingram DG. Economic and social costs of sleep apnea. Curr Pulmonol Rep 2016;5(2):111-115. [doi: 10.1007/s13665-016-0142-z]
5. Malhotra A, Ayappa I, Ayas N, Collop N, Kirsch D, Mcardle N, et al. Metrics of sleep apnea severity: beyond the apnea-hypopnea index. Sleep 2021;44(7):zsab030 [FREE Full text] [doi: 10.1093/sleep/zsab030] [Medline: 33693939]
6. Vanderveken OM, Gagnadoux F. Comparative outcomes in obstructive sleep apnea therapy: mean disease alleviation as a more appropriate measure-it's about time. Sleep 2023;46(10):zsad210 [FREE Full text] [doi: 10.1093/sleep/zsad210] [Medline: 37549352]
7. Won CHJ. When will we ditch the AHI? J Clin Sleep Med 2020;16(7):1001-1003 [FREE Full text] [doi: 10.5664/jcsm.8594] [Medline: 32441250]
8. Redline S, Baker-Goodwin S, Bakker JP, Epstein M, Hanes S, Hanson M, Sleep Apnea Patient-Centered Outcomes Network. Patient partnerships transforming sleep medicine research and clinical care: perspectives from the sleep apnea patient-centered outcomes network. J Clin Sleep Med 2016;12(7):1053-1058 [FREE Full text] [doi: 10.5664/jcsm.5948] [Medline: 27166300]
9. van Mechelen PH, Damhuis E, Hazeleger E. Monograph about the care pathway for sleep apnea: a research report from ApneuVereniging (Dutch Apnea Association) in cooperation with Kantar vereniging apneu. ApneuVereniging (Dutch Apnea Association). 2021. URL: https://administratie.apneuvereniging.nl/files/2021-monograph-about-dutch-care-pathway-osa.pdf [accessed 2025-12-03]
10. RPM definition. US Food and Drug Administration. 2023. URL: https://tinyurl.com/3j2yubte [accessed 2025-11-27]
11. Schwab RJ, Badr SM, Epstein LJ, Gay PC, Gozal D, Kohler M, ATS Subcommittee on CPAP Adherence Tracking Systems. An official American Thoracic Society statement: continuous positive airway pressure adherence tracking systems. The optimal monitoring strategies and outcome measures in adults. Am J Respir Crit Care Med 2013;188(5):613-620 [FREE Full text] [doi: 10.1164/rccm.201307-1282ST] [Medline: 23992588]
12. Phillips CL, Grunstein RR, Darendeliler MA, Mihailidou AS, Srinivasan VK, Yee BJ, et al. Health outcomes of continuous positive airway pressure versus oral appliance treatment for obstructive sleep apnea: a randomized controlled trial. Am J Respir Crit Care Med 2013;187(8):879-887. [doi: 10.1164/rccm.201212-2223OC] [Medline: 23413266]
13. Dieltjens M, Braem MJ, Op de Beeck S, Vroegop AVMT, Kazemeini E, Van de Perck E, et al. Remotely controlled mandibular positioning of oral appliance therapy during polysomnography and drug-induced sleep endoscopy compared with conventional subjective titration in patients with obstructive sleep apnea: protocol for a randomized crossover trial. Trials 2019;20(1):615 [FREE Full text] [doi: 10.1186/s13063-019-3698-4] [Medline: 31665059]
14. Abma IL, van der Wees PJ, Veer V, Westert GP, Rovers M. Measurement properties of patient-reported outcome measures (PROMs) in adults with obstructive sleep apnea (OSA): a systematic review. Sleep Med Rev 2016;28:18-31 [FREE Full text] [doi: 10.1016/j.smrv.2015.07.006] [Medline: 26433776]
15. Zini MLL, Banfi G. A narrative literature review of bias in collecting patient reported outcomes measures (PROMs). Int J Environ Res Public Health 2021;18(23):12445 [FREE Full text] [doi: 10.3390/ijerph182312445] [Medline: 34886170]

XSL•FO
RenderX

16. Aiyegbusi OL, Roydhouse J, Rivera SC, Kamudoni P, Schache P, Wilson R, et al. Key considerations to reduce or address respondent burden in patient-reported outcome (PRO) data collection. Nat Commun 2022;13(1):6026 [FREE Full text] [doi: 10.1038/s41467-022-33826-4] [Medline: 36224187]
17. Digital health technologies (DHTs) for drug development. US Food and Drug Administration. URL: https://www.fda.gov/science-research/science-and-research-special-topics/digital-health-technologies-dhts-drug-development [accessed 2025-11-27]
18. Manta C, Patrick-Lake B, Goldsack J. Digital measures that matter to patients: a framework to guide the selection and development of digital measures of health. Digit Biomark 2020;4(3):69-77 [FREE Full text] [doi: 10.1159/000509725] [Medline: 33083687]
19. Kolk MZH, Frodi DM, Langford J, Meskers CJ, Andersen TO, Jacobsen PK, et al. Behavioural digital biomarkers enable real-time monitoring of patient-reported outcomes: a substudy of the multicentre, prospective observational SafeHeart study. Eur Heart J Qual Care Clin Outcomes 2024;10(6):531-542. [doi: 10.1093/ehjqcco/qcad069] [Medline: 38059857]
20. Heros R, Patterson D, Huygen F, Skaribas I, Schultz D, Wilson D, et al. Objective wearable measures and subjective questionnaires for predicting response to neurostimulation in people with chronic pain. Bioelectron Med 2023;9(1):13 [FREE Full text] [doi: 10.1186/s42234-023-00115-4] [Medline: 37340467]
21. Lim DC, Sutherland K, Cistulli PA, Pack AI. P4 medicine approach to obstructive sleep apnoea. Respirology 2017;22(5):849-860 [FREE Full text] [doi: 10.1111/resp.13063] [Medline: 28477347]
22. Aryal S, Blankenship JM, Bachman SL, Hwang S, Zhai Y, Richards JC, et al. Patient-centricity in digital measure development: co-evolution of best practice and regulatory guidance. NPJ Digit Med 2024;7(1):128 [FREE Full text] [doi: 10.1038/s41746-024-01110-y] [Medline: 38755349]
23. Bakker JP, Barge R, Centra J, Cobb B, Cota C, Guo CC, et al. V3+ extends the V3 framework to ensure user-centricity and scalability of sensor-based digital health technologies. NPJ Digit Med 2025;8(1):51 [FREE Full text] [doi: 10.1038/s41746-024-01322-2] [Medline: 39856145]
24. Goldsack JC, Coravos A, Bakker JP, Bent B, Dowling AV, Fitzer-Attas C, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for biometric monitoring technologies (BioMeTs). NPJ Digit Med 2020;3:55 [FREE Full text] [doi: 10.1038/s41746-020-0260-4] [Medline: 32337371]
25. Taylor KI, Staunton H, Lipsmeier F, Nobbs D, Lindemann M. Outcome measures based on digital health technology sensor data: data- and patient-centric approaches. NPJ Digit Med 2020;3:97 [FREE Full text] [doi: 10.1038/s41746-020-0305-8] [Medline: 32715091]
26. Baines R, Bradwell H, Edwards K, Stevens S, Prime S, Tredinnick-Rowe J, et al. Meaningful patient and public involvement in digital health innovation, implementation and evaluation: a systematic review. Health Expect 2022;25(4):1232-1245 [FREE Full text] [doi: 10.1111/hex.13506] [Medline: 35526274]
27. Lazarevic N, Pizzuti C, Rosic G, Bœhm C, Williams K, Caillaud C. A mixed-methods study exploring women's perceptions and recommendations for a pregnancy app with monitoring tools. NPJ Digit Med 2023;6(1):50 [FREE Full text] [doi: 10.1038/s41746-023-00792-0] [Medline: 36964179]
28. Renn BN, Hoeft TJ, Lee HS, Bauer AM, Areán PA. Preference for in-person psychotherapy versus digital psychotherapy options for depression: survey of adults in the U.S. NPJ Digit Med 2019;2:6 [FREE Full text] [doi: 10.1038/s41746-019-0077-1] [Medline: 31304356]
29. Weik L, Fehring L, Mortsiefer A, Meister S. Understanding inherent influencing factors to digital health adoption in general practices through a mixed-methods analysis. NPJ Digit Med 2024;7(1):47 [FREE Full text] [doi: 10.1038/s41746-024-01049-0] [Medline: 38413767]
30. Mars GMJ, van den Heuvel SG, Knops JCM, de Vroome EMM, Gielen WJM, van Dam LMC, et al. National Survey of Working Conditions (NEA) 2023 - Research description [article in Dutch]. Centraal Bureau voor de Statistiek. URL: https://www.cbs.nl/nl-nl/longread/rapportages/2024/nationale-enquete-arbeidsomstandigheden--nea---2023-onderzoeksbeschrijving [accessed 2025-10-28]
31. Saari DG. Geometry of voting. In: Arrow KJ, Sen AK, Suzumura K, editors. Handbook of Social Choice and Welfare. Amsterdam, The Netherlands: Elsevier Science; 2011.
32. Turner H, van Etten J, Firth D, Kosmidis I. Modelling rankings in R: the PlackettLuce package. Comput Stat 2020 Feb 12;35(3):1027-1057. [doi: 10.1007/s00180-020-00959-3]
33. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2008;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
34. Alhassan AI. Analyzing the application of mixed method methodology in medical education: a qualitative study. BMC Med Educ 2024;24(1):225 [FREE Full text] [doi: 10.1186/s12909-024-05242-3] [Medline: 38438987]
35. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Methods 2017;16(1). [doi: 10.1177/1609406917733847]
36. van den Broek T, Timmis JK, Schorr KA, Yüksel R, Overeem S, Smid D, et al. Replication data for: toward patient-centric digital health solutions for obstructive sleep apnea monitoring: perspectives from Dutch patients and healthcare professionals: a mixed-method study. Harvard Dataverse 2021.

XSL•FO
RenderX

37. Chee MW, Baumert M, Scott H, Cellini N, Goldstein C, Baron K, World Sleep Society Sleep Tracker Task Force. World Sleep Society recommendations for the use of wearable consumer health trackers that monitor sleep. Sleep Med 2025;131:106506. [doi: 10.1016/j.sleep.2025.106506] [Medline: 40300398]
38. Gruwez A, Bruyneel AV, Bruyneel M. The validity of two commercially-available sleep trackers and actigraphy for assessment of sleep parameters in obstructive sleep apnea patients. PLoS One 2019;14(1):e0210569 [FREE Full text] [doi: 10.1371/journal.pone.0210569] [Medline: 30625225]
39. Kuhn J, Schiphorst LRB, Wulterkens BM, Asin J, Duis N, Overeem S, et al. Multi-night home assessment of total sleep time misperception in obstructive sleep apnea with and without insomnia symptoms. Clocks Sleep 2024;6(4):777-788 [FREE Full text] [doi: 10.3390/clockssleep6040050] [Medline: 39727626]
40. Strumpf Z, Gu W, Tsai C, Chen P, Yeh E, Leung L, et al. Belun Ring (Belun Sleep System BLS-100): deep learning-facilitated wearable enables obstructive sleep apnea detection, apnea severity categorization, and sleep stage classification in patients suspected of obstructive sleep apnea. Sleep Health 2023;9(4):430-440 [FREE Full text] [doi: 10.1016/j.sleh.2023.05.001] [Medline: 37380590]
41. Cai Y, Zheng YJ, Cheng CM, Strohl KP, Mason AE, Chang JL. Impact of hypoglossal nerve stimulation on consumer sleep technology metrics and patient symptoms. Laryngoscope 2024;134(7):3406-3411. [doi: 10.1002/lary.31398] [Medline: 38516821]
42. Byun J, Noh KC, Shin WC. Performance of the Fitbit Charge 2 and Galaxy Watch 2 compared with polysomnography in assessing patients with obstructive sleep apnoea. Chronobiol Int 2023;40(5):596-602. [doi: 10.1080/07420528.2023.2191720] [Medline: 36971253]
43. Peng J, Yuan Y, Zhao Y, Ren H. Effects of exercise on patients with obstructive sleep apnea: a systematic review and meta-analysis. Int J Environ Res Public Health 2022;19(17):10845 [FREE Full text] [doi: 10.3390/ijerph191710845] [Medline: 36078558]
44. Mendelson M, Bailly S, Marillier M, Flore P, Borel JC, Vivodtzev I, et al. Obstructive sleep apnea syndrome, objectively measured physical activity and exercise training interventions: a systematic review and meta-analysis. Front Neurol 2018;9:73 [FREE Full text] [doi: 10.3389/fneur.2018.00073] [Medline: 29520251]
45. Pawar M, Venkatesan P, Mysore S, Bhat G. Effectiveness of aerobic exercise training in patients with obstructive sleep apnea: a systematic review and meta-analysis. Eur Arch Otorhinolaryngol 2025. [doi: 10.1007/s00405-025-09436-3] [Medline: 40329037]
46. Feng Y, Maislin D, Keenan BT, Gislason T, Arnardottir ES, Benediktsdottir B, et al. Physical activity following positive airway pressure treatment in adults with and without obesity and with moderate-severe obstructive sleep apnea. J Clin Sleep Med 2018;14(10):1705-1715 [FREE Full text] [doi: 10.5664/jcsm.7378] [Medline: 30353806]
47. Stevens D, Loffler KA, Buman MP, Dunstan DW, Luo Y, Lorenzi-Filho G, SAVE investigators. CPAP increases physical activity in obstructive sleep apnea with cardiovascular disease. J Clin Sleep Med 2021;17(2):141-148 [FREE Full text] [doi: 10.5664/jcsm.8792] [Medline: 32951632]
48. West SD, Kohler M, Nicoll DJ, Stradling JR. The effect of continuous positive airway pressure treatment on physical activity in patients with obstructive sleep apnoea: a randomised controlled trial. Sleep Med 2009;10(9):1056-1058. [doi: 10.1016/j.sleep.2008.11.007] [Medline: 19427263]
49. Schwarz EI, Schiza S. Sex differences in sleep and sleep-disordered breathing. Curr Opin Pulm Med 2024;30(6):593-599. [doi: 10.1097/MCP.0000000000001116] [Medline: 39189037]
50. Pyun SY, Choi SJ, Jo H, Hwang Y, Cho JW, Joo EY. Gender differences in Korean patients with obstructive sleep apnea. Sleep Med Res 2020;11(2):121-128. [doi: 10.17241/smr.2020.00556]
51. Steier JS, Bogan RK, Cano-Pumarega IM, Fleetham JA, Insalaco G, Lal C, et al. Recommendations for clinical management of excessive daytime sleepiness in obstructive sleep apnoea - A Delphi consensus study. Sleep Med 2023;112:104-115 [FREE Full text] [doi: 10.1016/j.sleep.2023.10.001] [Medline: 37839271]
52. Adão Martins NR, Annaheim S, Spengler CM, Rossi RM. Fatigue monitoring through wearables: a state-of-the-art review. Front Physiol 2021;12:790292 [FREE Full text] [doi: 10.3389/fphys.2021.790292] [Medline: 34975541]
53. Lee J, Ahn S. Polysomnographic findings and psychiatric symptoms in patients with comorbid insomnia and sleep apnea: a retrospective study focusing on sex differences. Sleep Breath 2025;29(1):78. [doi: 10.1007/s11325-025-03248-9] [Medline: 39808352]
54. Subramanian S, Guntupalli B, Murugan T, Bopparaju S, Chanamolu S, Casturi L, et al. Gender and ethnic differences in prevalence of self-reported insomnia among patients with obstructive sleep apnea. Sleep Breath 2011;15(4):711-715. [doi: 10.1007/s11325-010-0426-4] [Medline: 20953842]
55. Saaresranta T, Hedner J, Bonsignore MR, Riha RL, McNicholas WT, Penzel T, ESADA Study Group. Clinical phenotypes and comorbidity in European sleep apnoea patients. PLoS One 2016;11(10):e0163439 [FREE Full text] [doi: 10.1371/journal.pone.0163439] [Medline: 27701416]

Abbreviations

AHI: Apnea-Hypopnea Index

XSL•FO
RenderX

CAQDAS: computer-assisted qualitative data analysis software
CPAP: continuous positive airway pressure
DiMe: Digital Medicine Society
MAD: mandibular advancement device
OSA: obstructive sleep apnea
PROM: patient-reported outcome measure
RPM: remote patient monitoring
sDHT: sensor-based digital health technology
TNO: Netherlands Organization for Applied Scientific Research

Edited by A Stone, A Mavragani; submitted 15.Aug.2025; peer-reviewed by W Ahmed, O Oluwole; comments to author 08.Sep.2025; revised version received 28.Oct.2025; accepted 28.Oct.2025; published 08.Jan.2026.

Please cite as:
Timmis JK, Schorr KA, Yüksel R, van den Broek T, Overeem S, Smid DJ, van den Brink WJ, Haring NL
Toward Patient-Centric Digital Monitoring of Obstructive Sleep Apnea: Mixed Methods Study
J Med Internet Res 2026;28:e82460
URL: https://www.jmir.org/2026/1/e82460
doi:10.2196/82460
PMID:

©James Kenneth Timmis, Kerstin Alexandra Schorr, Rana Yüksel, Tim van den Broek, Sebastiaan Overeem, Dagmar Josine Smid, Willem Johan van den Brink, Nina Leonie Haring. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of Educational Videos in Encouraging Preferences for Guideline-Based Cancer Screening in Japan: Three-Arm Pseudorandomized Controlled Trial

Kosuke Sakai[1], MD, PhD; Yuko Furuya[1], MD, PhD; Shoko Nakazawa[1], MD, PhD; Kota Fukai[1], MD, PhD; Kei Sano[2], MD, PhD; Masayuki Tatemichi[1], MD, PhD

[1]Department of Preventive Medicine, Tokai University, School of Medicine, Kanagawa, Japan

[2]Department of Ophthalmology, The Jikei University, School of Medicine, Tokyo, Japan

Corresponding Author:
Kota Fukai, MD, PhD
Department of Preventive Medicine
Tokai University, School of Medicine
143 Shimokasuya
Isehara
Kanagawa, 259-1193
Japan
Phone: 81 463 93 1121 ext 2612
Fax: 81 463 92 3549
Email: kota229@tokai.ac.jp

Abstract

Background: Although cancer screening is essential for early detection and an improved prognosis, screening beyond the recommended guidelines may increase the risk of false-positive results. Consequently, educating individuals about the potential harm of non–guideline-based cancer screening is essential; however, effective communication methods remain unclear.

Objective: This study aimed to evaluate the effectiveness of different types of educational videos in encouraging preferences for guideline-based cancer screening.

Methods: This 3-arm pseudorandomized controlled trial was conducted in June 2025 using a Japanese online survey platform. Eligible respondents were working adults aged 30 to 60 years with no history of major cancer. Respondents were assigned to 1 of the following 3 video conditions: video A, which provided a logical explanation of false-positive risks; video B, which presented the narrative of a woman who received a false-positive result from breast cancer screening; and video C, which depicted a man who underwent unnecessary follow-up testing after tumor marker screening. The primary outcome was the preference for guideline-based cancer screening after watching the videos. The secondary outcomes included 7 self-reported video evaluation items, such as perceived relevance and clarity, assessed using a 5-point Likert scale. Differences in the primary outcome between video groups were analyzed using multivariable logistic regression with adjustment for covariates. Means and 95% CIs were calculated for each secondary outcome according to sex and video group. In addition, before-and-after changes in screening preferences were assessed using McNemar test, with a significance level of .05.

Results: In total, 1200 respondents (400 per group) completed the survey. No statistically significant differences in the primary outcome were observed among the video groups. With reference to video A, the adjusted odds ratios for preferring guideline-based screening were 0.89 (95% CI 0.59-1.32) for video B and 0.98 (95% CI 0.65-1.46) for video C. Regarding secondary outcomes, male respondents rated video B less favorably than female respondents in terms of relevance and willingness to undergo guideline-based screening. The before-and-after comparison showed a significant change in preference for guideline-based screening (P=.04). These videos appeared to be more effective for individuals with an annual history of colorectal cancer screening than for those without such a history.

Conclusions: Educational videos have the potential to influence cancer screening preferences; however, no single video format has demonstrated clear superiority. These findings underscore the importance of tailoring educational materials to the target audience characteristics. Further research is required to develop effective strategies for encouraging guideline-based cancer screening.

XSL•FO
RenderX

Trial Registration: University Hospital Medical Information Network Clinical Trials Registry UMIN000060549; https://center6.umin.ac.jp/cgi-open-bin/ctr/ctr_view.cgi?recptno=R000066119

(J Med Internet Res 2026;28:e82322) doi:10.2196/82322

KEYWORDS

health education; randomized controlled trial; cancer screening; false positive; decision-making

Introduction

The strategic implementation of cancer screening facilitates early cancer detection, thereby improving patient survival through less invasive treatment options [1]. National guidelines, such as those developed by the US Preventive Services Task Force and the Japan National Cancer Center, clearly define the populations that benefit from screening and the tests recommended for them, based on robust evidence of mortality reduction [2,3]. Despite these guidelines, many individuals still undergo nonrecommended screenings, including breast cancer screening in people younger than the recommended age and whole-body positron emission tomography-computed tomography (PET-CT) imaging [4,5].

These non–guideline-based screenings often cause harm to participants through false-positive results, overdiagnosis, and overtreatment [6]. False-positive results often lead to negative psychosocial consequences [7]. In addition, prostate, lung, colorectal, and ovarian cancer screening trials reported that undergoing up to 14 cancer screening tests over a 3-year period resulted in cumulative risks of 60.4% and 48.8% for experiencing at least 1 false-positive result in men and women, respectively [8]. Moreover, approximately 10% of Japanese women in their 40s have been reported to experience false-positive results during breast cancer screening [9]. This risk is considered further elevated when screening is conducted at a younger age [10]. Furthermore, previous studies [11,12] have suggested that false-positive results can result in decreased adherence to future cancer screening. Accordingly, health care professionals play a crucial role in facilitating informed decision-making in cancer screening by helping participants understand both the benefits and harms of screening [13,14].

Previous studies [14,15] have evaluated the effects of various educational programs aimed at enhancing knowledge of the benefits and harms of cancer screening. For instance, multicomponent interventions based on behavior change models were implemented at the community level for breast cancer education [15]. These interventions were reported to be successful when various factors were considered, including the use of video consultations, provision of sessions in multiple languages, and consideration of the socioeconomic status of the community [16]. However, these educational efforts were designed to increase cancer screening uptake and not to alert people to the harmful effects of non–guideline-based cancer screening.

In Japan, the commercialization of health care has led to a strong emphasis on the benefits of early cancer detection and additional testing as much as possible. A survey on the implementation status of cancer screening in Japan [17] found that approximately 70% of companies offered colorectal cancer screening to individuals younger than 40 years, and 45% offered breast cancer screening to individuals younger than 40 years. A national survey in Japan [18] has shown that more than 50,000 healthy individuals in 2005 and approximately 39,000 per year from 2006 to 2009 underwent PET-CT cancer screening outside of guideline-based programs. Therefore, educating individuals about the potential harms of non–guideline-based cancer screening is essential for informed decision-making; however, effective communication methods remain unclear. One educational method can be to raise awareness about the risks associated with false-positive results.

We aimed to compare the short-term impact of 3 brief educational videos on preferences for guideline-based cancer screening. This intervention was grounded in the narrative transportation theory, which suggests that stories transport audiences into a narrative world, focusing attention, emotion, and imagery on the story and thereby reducing counterarguing and shaping beliefs and attitudes [19]. In this framework, we hypothesized that narrative videos would be more effective than an explanatory format in shifting preferences toward guideline-based screening. By comparing the effectiveness of the 3 types of educational videos, we gained deeper insight into how approaches are most effective for specific characteristics of the target audience.

Methods

Study Design and Respondents

This study was conducted according to the protocol registered in the University Hospital Medical Information Network Clinical Trials Registry (UMIN000060549). In June 2025, we conducted a 3-arm pseudorandomized controlled trial using a Japanese online survey platform. This trial was supported by Rakuten Insight Inc, a web-based research company located in Tokyo. The respondents were recruited from the company's registry via email. This manuscript is reported in accordance with the Consolidated Standards of Reporting Trials (CONSORT) 2010 statement (Multimedia Appendix 1).

The inclusion criteria were working individuals aged between 30 and 60 years who agreed to participate in the study. The exclusion criteria were a history of stomach, lung, colon, breast, or cervical cancer. In addition, individuals who reported undergoing cancer screening recommended by local governments within the past 3 years were excluded from the study, as these programs only offer screening options that adhere to national guidelines. The participants were administered a questionnaire comprising 27 questions. Respondents who required <5 minutes or >30 minutes to complete the

XSL•FO
RenderX

questionnaire were considered to have given inappropriate answers and were excluded.

Randomization and Procedure

The respondents were not individually randomized. Instead, separate web-based questionnaires were prepared for each of the 3 video conditions (A, B, and C). Invitations were sent to participants until 400 valid responses were obtained for each video group. Invitations for each group were distributed sequentially and exclusively, and individuals who were invited to one video condition were not invited to any other condition. The target population included 200 men and 200 women in 1 group, resulting in 3 groups with a total target sample size of 1200 respondents. Each group viewed a different video. We ensured that the respondents were unable to answer the next question unless they watched the video until the end, to ensure data quality. The sample size was determined based on previous studies of educational videos using animation [20].

Intervention

We developed 3 educational videos to inform people about the potential harm caused by false-positive results in cancer screening and to encourage guideline-based screening decisions. Each educational video was developed based on the health belief model [21], which explained that health-related behavioral change can be influenced by individuals' perceptions of expected benefits and costs. In the context of cancer screening, the perceived benefits of guideline-based screening were detecting cancer at an earlier stage, enabling less invasive treatment, and potentially reducing cancer mortality. The costs included financial and time burdens, psychological distress triggered by false-positive results, and potential harm from unnecessary follow-up procedures such as avoidable radiation exposure. By presenting the benefits of guideline-concordant screening and the potential harms of non–guideline-based screening, the videos were designed to support informed decision-making and encourage preferences aligned with screening guidelines.

All videos had a duration of approximately 100 seconds. Video A illustrated the disadvantages of excessive cancer screening using charts and numerical data. Video B showed the story of a woman in her 30s who underwent breast cancer screening, was referred for further testing, and ultimately received a false-positive result. Video C showed the story of a working man who underwent multiple tumor marker tests, received positive results in several categories, and was subsequently required to undergo numerous further tests. The structures of the 3 videos are shown in Multimedia Appendix 2. Video A is shown in Multimedia Appendix 3; video B is shown in Multimedia Appendix 4; and video C is shown in Multimedia Appendix 5. The duration of video A was 1 minute and 24 seconds, whereas videos B and C lasted for 1 minute and 17 seconds each.

Outcome Measures

The primary outcome was preference for guideline-based cancer screening after watching the videos. Respondents were asked what type of cancer screening they would prefer, with the following options: (1) "I want to undergo cancer screening as recommended by the guidelines," (2) "I want to undergo cancer screening at an earlier age than recommended," (3) "I want to undergo different types of cancer screening than recommended (eg, tumor markers or PET-CT imaging)," (4) "I want to undergo cancer screening at intervals shorter or longer than the recommended interval," (5) "I do not want to undergo any cancer screening," and (6) "I have no idea." Options 2, 3, and 4 were not mutually exclusive, and if any of these options were chosen, the response was interpreted as a preference for non–guideline-based cancer screening. The primary outcome measured an immediate postexposure preference and was intended as a proximal indicator of decision-making.

We included a 7-item evaluation survey as a brief process evaluation to assess how participants received the educational messages and to aid interpretation of the primary outcome. Conceptually, the items included (1) perceived personal relevance and engagement with the message (relevance), which is central to narrative persuasion and transportation; (2) comprehensibility and cognitive processing of the content (clarity), which is a prerequisite for informed decision-making; (3) perceived informational sufficiency for making a screening decision (informativeness), corresponding to the "knowledge or understanding" component needed to weigh benefits and harms; (4) perceived acceptability as educational material (acceptability), reflecting practical feasibility and appropriateness for dissemination; and (5) negative affect and avoidance tendency (aversion), which relates to potential unintended emotional burden and message avoidance [22]. Respondents rated these aspects on a 5-point Likert scale (1=strongly disagree, 2=disagree, 3=neutral, 4=agree, and 5=strongly agree): relevance ("This video is relevant to me"), clarity ("This video is clear and easy to understand"), informativeness ("This video provides all the necessary information for screening participants"), acceptability ("This video is acceptable as educational material"), and aversion ("I do not want to think about this video"). In addition, respondents were asked about the influence of the videos on 2 aspects: motivating them to discuss their content with family or friends and undergoing guideline-based cancer screening using the same 5-point scale.

Covariates

Covariates included sex, age, education, and preferences for cancer screening before viewing the video. Sex was based on self-reported information, and age was determined from the date of birth. Education was categorized into middle school, high school, vocational schools, junior college or technical college, university, graduate school, currently enrolled, and other, and respondents were asked to select the most appropriate option. However, for analysis, education was treated as a binary variable: less than a university degree vs a university degree or higher. The preferences for cancer screening before viewing the video were obtained using the same method as that used for the primary outcome.

Statistical Analysis

The distributions of age, marital status, education, income, and preferences for cancer screening before watching the video were tabulated for the 3 groups to compare the characteristics of the

XSL•FO
RenderX

respondents. The primary outcome variable was coded as 1 if the respondent selected the option “I want to undergo guideline-based cancer screening” after watching the video, and 0 otherwise. The adjusted odds ratios (aORs) and 95% CIs for videos B and C, relative to video A, were estimated using multivariate logistic regression analysis. The covariates included sex, age, education, and a preference for cancer screening before watching the video.

The means and 95% CIs were calculated for each of the secondary outcomes, which were the 7 self-reported video evaluation indicators, stratified by sex. In addition to the prespecified primary between-group comparison, we conducted an exploratory, post hoc before-and-after analysis of screening preferences immediately before and after video viewing across all participants. This analysis was not prespecified in the trial protocol and should be interpreted as descriptive and hypothesis generating. A 2-sided McNemar test was performed at a significance level of .05. Changes in cancer screening preferences before and after viewing the videos were visualized using a Sankey diagram.

We compared 3 groups to examine the characteristics of participants who perceived the video as effective: individuals who shifted toward guideline-based screening (effective group), those whose preferences did not change (neutral group), and those who shifted away from guideline-based screening (counterproductive group). To compare participant characteristics across the 3 groups, *P* values were calculated using Welch 1-way ANOVA for continuous variables and the chi-square test for categorical variables, with a 2-sided significance level set at 5%. These *P* values were provided for descriptive purposes; no adjustment for multiple comparisons was applied. Age and total response time were treated as continuous variables, while age group (10-year categories), sex, marital status, education, income, and history of colorectal cancer screening were treated as categorical variables. All statistical analyses were performed using R (version 4.4.2; R Foundation for Statistical Computing). Because all questions were mandatory, the dataset contained no missing values.

Ethical Contributions

This study was approved by the Research Ethics Committee of Tokai University School of Medicine (24R099). Informed consent was obtained from all respondents before they responded to the questionnaire. Because cancer-related educational materials may unintentionally elicit fear or distress, potential adverse emotional effects could not be fully excluded. During video production, we intentionally avoided sensational expressions. The video materials were submitted for institutional ethics review and were evaluated from a third-party perspective. Before participation, respondents were informed that the survey included videos addressing the potential harms and benefits of cancer screening. Respondents were also informed that participation was voluntary, that they could discontinue the survey at any time, and that discontinuation would not result in any disadvantage. Participation was incentivized by compensating respondents who completed the survey with points that could be exchanged for merchandise on online platforms. To protect respondents' privacy, we avoided collecting personally identifiable information. We confirmed that the data contained no personal information before conducting our analysis.

Results

Characteristics of Respondents

Table 1 summarizes the characteristics of the overall cohort and the individual video groups. The mean age of 1200 participants was 48.9 (SD 7.4) years, and the most common age group was 50 to 60 years (n=647, 53.9%). The proportion of unmarried individuals was 43.4% (n=521), and the most common educational attainment was a university degree (n=502, 41.8%). The most frequently reported income level was ≤JP ¥4 million (US $1.00=JP ¥145 is used throughout the text), with 310 (25.8%) respondents. The most common preference for cancer screening before watching the video was guideline-based screening (n=565, 47.1%).

XSL•FO
RenderX

Table 1. Characteristics of all respondents (N=1200).

Characteristics	All videos (n=1200)	Video A (n=400)	Video B (n=400)	Video C (n=400)
Age (years), mean (SD)	48.9 (7.4)	49.13 (7.5)	48.31 (7.8)	49.17 (7.3)
Age (years), n (%)				
30-39	162 (13.5)	48 (12.0)	65 (16.2)	49 (12.2)
40-49	391 (32.6)	131 (32.8)	133 (33.2)	127 (31.8)
50-60	647 (53.9)	221 (55.2)	202 (50.5)	224 (56.0)
Marital status, n (%)				
Unmarried	521 (43.4)	180 (45.0)	173 (43.2)	168 (42.0)
Education, n (%)				
Junior high school	9 (0.8)	3 (0.8)	2 (0.5)	4 (1.0)
High school	299 (24.9)	110 (27.5)	101 (25.2)	88 (22.0)
Vocational school graduate	160 (13.3)	53 (13.2)	54 (13.5)	53 (13.2)
Junior college or technical college graduate	139 (11.6)	45 (11.2)	46 (11.5)	48 (12.0)
University graduate	502 (41.8)	161 (40.2)	166 (41.5)	175 (43.8)
Graduate school graduate	82 (6.8)	25 (6.2)	29 (7.2)	28 (7.0)
Currently enrolled	9 (0.8)	3 (0.8)	2 (0.5)	4 (1.0)
Income (Japanese yen; million , US $1.00 = ¥145), n (%)				
≤4.00	310 (25.8)	115 (28.7)	99 (24.8)	96 (24.0)
4.01-6.00	270 (22.5)	90 (22.5)	100 (25.0)	80 (20.0)
6.01-8.00	240 (20.0)	80 (20.0)	71 (17.8)	89 (22.2)
8.01-10.00	163 (13.6)	48 (12.0)	58 (14.5)	57 (14.2)
10.01-12.00	98 (8.2)	24 (6.0)	32 (8.0)	42 (10.5)
12.01-15.00	70 (5.8)	27 (6.8)	26 (6.5)	17 (4.2)
≥15.01	49 (4.1)	16 (4.0)	14 (3.5)	19 (4.8)
Preference for cancer screening before watching the video, n (%)				
“I want to undergo cancer screening in accordance with the guidelines.”	565 (47.1)	177 (44.2)	183 (45.8)	205 (51.2)
“I want to undergo cancer screening that differs from the guidelines.”	383 (31.9)	135 (33.8)	136 (34.0)	112 (28.0)
“I don’t want to get screened for cancer.”	67 (5.6)	18 (4.5)	25 (6.2)	24 (6.0)
“I have no idea.”	185 (15.4)	70 (17.5)	56 (14.0)	59 (14.8)

Effect of Video Viewing on Cancer Screening Preferences Based on Guidelines

Figure 1 presents the aORs and 95% CIs for the primary outcome based on logistic regression analysis. Compared with video A, the aOR for video B was 0.89 (95% CI 0.59-1.32) and that for video C was 0.98 (95% CI 0.65-1.46), indicating no statistically significant differences. Multimedia Appendix 6 presents the results for all variables included in the logistic regression analysis.

XSL•FO
RenderX

Figure 1. Effect of videos on guideline-based screening preference. Analyses were adjusted for sex, age, education (university degree or lower), and preference for cancer screening before watching the video.

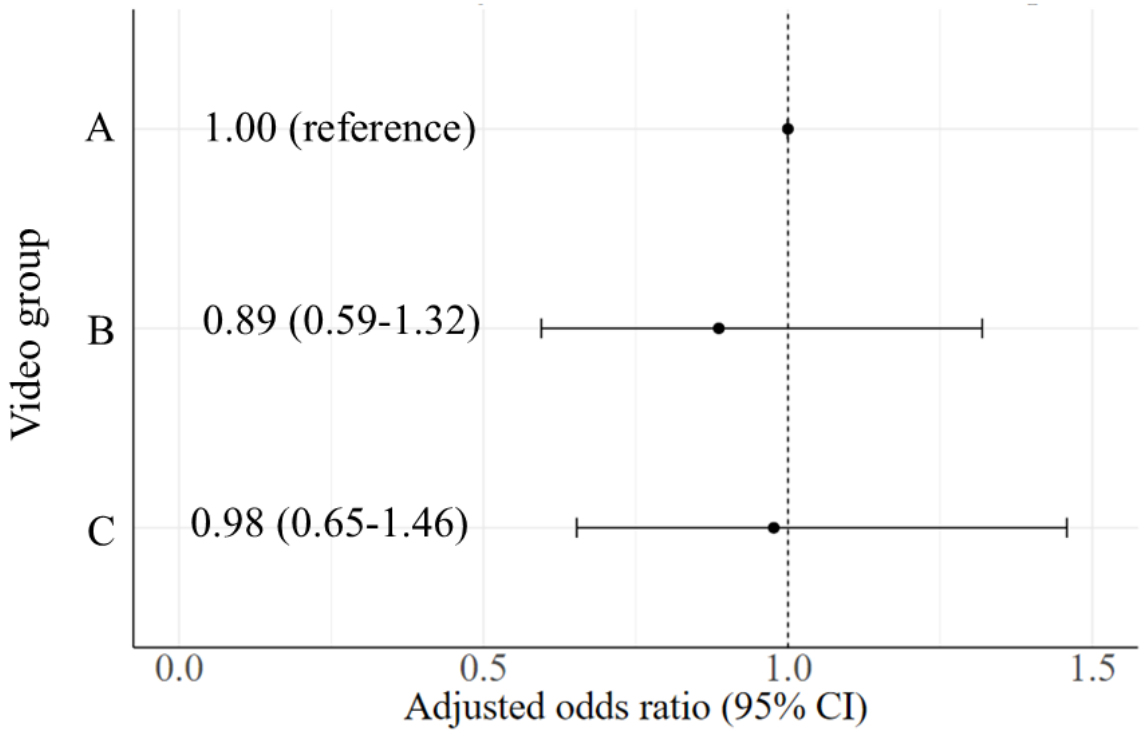

Evaluation of the Videos

Figure 2 shows the results of the secondary outcomes (7-item evaluation) for each video stratified by sex. The perceived relevance of video B was significantly lower among men (mean 2.5; 95% CI 2.4-2.7) than among women (mean 3.5; 95% CI 3.4-3.7). Similarly, willingness to undergo guideline-based screening was significantly lower among men (mean 3.3; 95% CI 3.1-3.4) than among women (mean 3.7; 95% CI 3.5-3.8). All scores, rounded to the second decimal place, are presented in Multimedia Appendix 7. Although no statistically significant differences were observed, the point estimates for all evaluation items among the women tended to be equal to or higher for video B than for video C.

XSL•FO
RenderX

Figure 2. Mean score with 95% CI for each video and sex group.

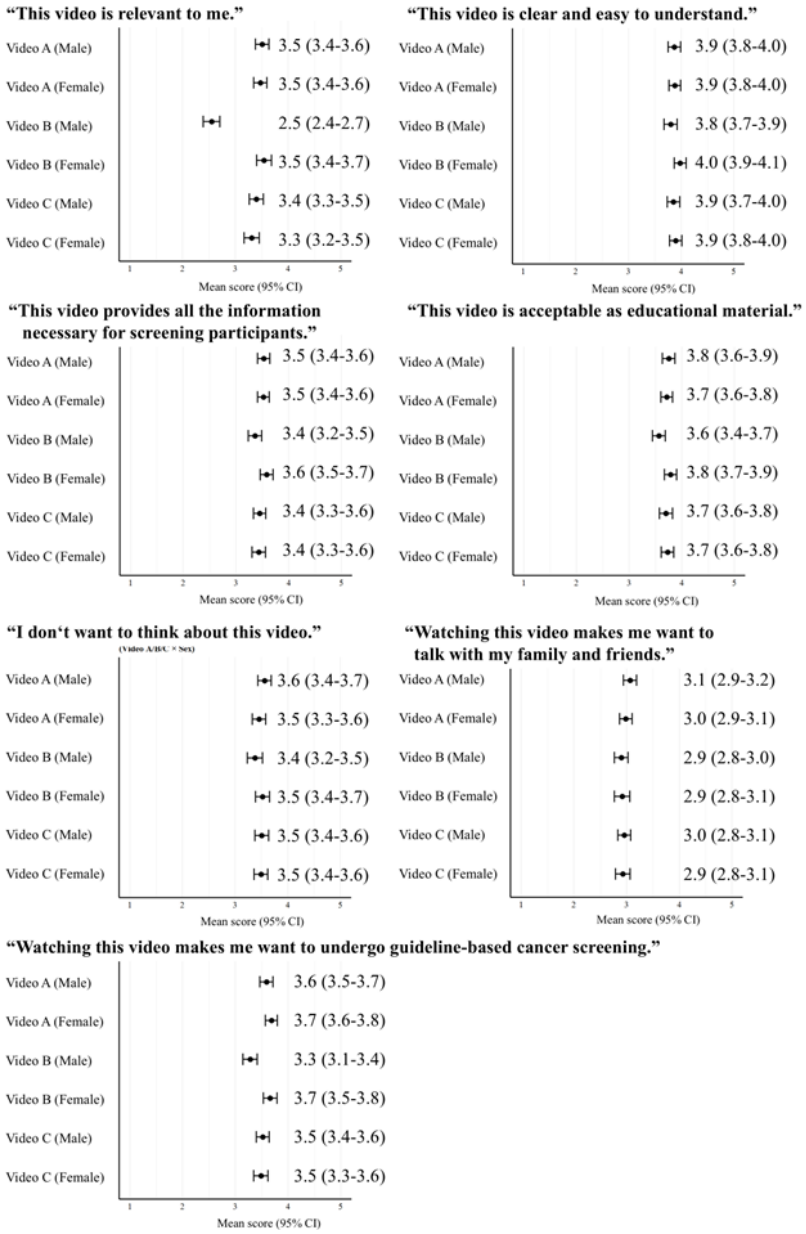

Total Changes in Cancer Screening Preferences Before and After Viewing the Video

Figure 3 illustrates the changes in screening preferences before and after watching the video using a Sankey diagram. Among the respondents, 91 shifted toward guideline-based screening (effective group), 283 maintained non–guideline-based preferences (neutral group), and 57 transitioned from guideline-based to non–guideline-based screening preferences (counterproductive group). Table 2 presents the characteristics of these 3 groups. A history of annual colorectal cancer screening over the previous 2 years was reported to be 57.1% (52/91) in the effective group, 52.7% (149/283) in the neutral group, and 43.9% (25/57) in the counterproductive group (P=.04). Although no statistically significant difference was observed in median total response time, the counterproductive group tended to have a shorter response time (436 seconds) than the effective group (459 seconds) and the neutral group (464 seconds). The density distribution of the total response times for these groups is shown in Multimedia Appendix 8. Before watching the video, 383 respondents preferred non–guideline-based cancer screening, and after viewing the video, 91 (23.8%) shifted to a preference for guideline-based screening.

XSL•FO
RenderX

Figure 3. Changes in preferences regarding cancer screening before and after watching the videos. Flow widths represent the proportion of participants who were consistent in or transitioned between preference categories. Colors denote response categories and are consistent across the "before and after watching the video" nodes. Blue represents “I want to undergo guideline-based cancer screening.” Orange represents "I want to undergo cancer screening that differs from the guidelines," which included the following survey response options: “I want to undergo cancer screening at an earlier age than recommended,” “I want to undergo different types of cancer screening than recommended (eg, tumor markers or PET-CT imaging),” and “I want to undergo cancer screening at intervals shorter or longer than the recommended interval.” Green represents “I do not want to undergo any cancer screening.” Purple represents “I have no idea.”.

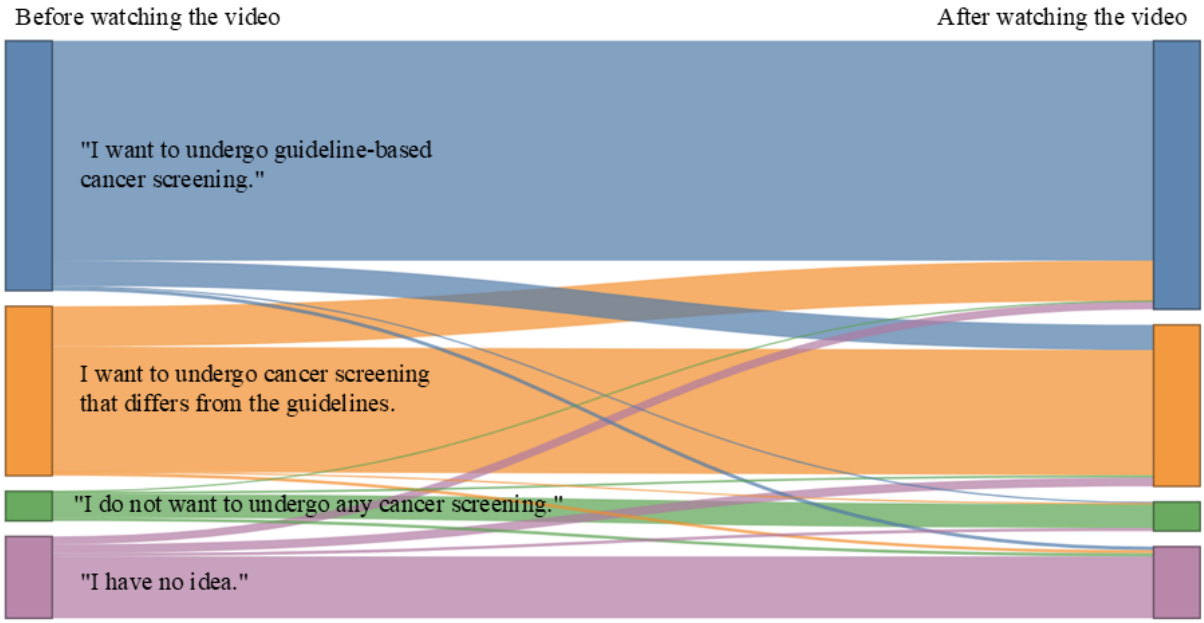

XSL•FO
RenderX

Table 2. Characteristics of respondents in the effective, neutral, and counterproductive groups (N=431).

Characteristics	Effective group (n=91)	Neutral group (n=283)	Counterproductive group (n=57)	*P* value
Age (years), mean (SD)	50.0 (7.1)	48.5 (7.6)	48.8 (7.8)	.26
Age (years), n (%)				.61
30-39	8 (8.8)	43 (15.2)	7 (12.3)	
40-49	31 (34.1)	85 (30.0)	17 (29.8)	
50-60	52 (57.1)	155 (54.8)	33 (57.9)	
Sex, n (%)				
Female	49 (53.8)	138 (48.8)	29 (50.9)	.70
Marital status, n (%)				
Unmarried	34 (37.4)	101 (35.7)	20 (35.1)	.95
Education, n (%)				
University degree or higher	48 (52.7)	152 (53.7)	23 (40.4)	.18
Income (Japanese yen; million; US $1.00 = ¥145), n (%)				.93
≤4.00	22 (24.2)	66 (23.3)	16 (28.1)	
4.01-6.00	17 (18.7)	56 (19.8)	12 (21.1)	
6.01-8.00	23 (25.3)	52 (18.4)	12 (21.1)	
8.01-10.00	13 (14.3)	43 (15.2)	6 (10.5)	
10.01-12.00	6 (6.6)	29 (10.2)	5 (8.8)	
12.01-15.00	7 (7.7)	19 (6.7)	4 (7.0)	
≥15.01	3 (3.3)	18 (6.4)	2 (3.5)	
History of colorectal cancer screening, n (%)				.04
Annually (past 2 years)	52 (57.1)	149 (52.7)	25 (43.9)	
Once (past 2 years)	12 (13.2)	15 (5.3)	4 (7.0)	
Earlier than past 2 years	4 (4.4)	38 (13.4)	11 (19.3)	
Never received	23 (25.3)	76 (26.9)	16 (28.1)	
Do not know	0 (0.0)	5 (1.8)	1 (1.8)	
Total response time (seconds), median (IQR)	459 (366–603)	464 (383–600)	436 (356–522)	.22

Discussion

Principal Findings

In this study, we aimed to identify effective educational strategies to convey the risk of false-positive results associated with non–guideline-based cancer screening. No statistically significant differences were observed in preferences for guideline-based cancer screening among the 3 videos designed to educate participants regarding the harm caused by false-positive results. The subjective evaluation provided additional insights into the effectiveness of video-based educational methods beyond the primary preference outcome. Among men, the evaluation of the narrative regarding breast cancer screening yielded lower scores for relevance and willingness to undergo guideline-based screening compared with evaluations among women. In contrast, the breast cancer screening narrative was generally received more favorably by women than the tumor marker narrative presented by a male character. Although not specified in the protocol beforehand, the McNemar test indicated a significant change in preferences after watching the videos. Among the 383 respondents who initially preferred non–guideline-based cancer screening, only 91 (23.8%) changed their preference to guideline-based screening after watching videos. The effective group underwent annual colorectal cancer screening. Meanwhile, the counterproductive group showed shorter response times than the effective and neutral groups.

Narrative educational videos are not expected to be more effective than logical or explanatory videos in cancer screening education. However, according to a systematic review of health education using videos [23], narrative presentations that evoke real-world situations are more effective in promoting behavioral change than the simple didactic delivery of information. Therefore, we created a control video that provided information and 2 narrative videos and hypothesized that narrative videos would be more effective than the control video; however, our results did not support this hypothesis. The findings of this study are consistent with those of a previous study examining the impact of different animation presentation formats on educational outcomes [24]. The previous study presented

XSL•FO
RenderX

animated videos that explained the advantages and disadvantages of colorectal cancer screening using 3 formats: videos with animated pictographs, videos with static pictographs, and audio booklets accompanied by static pictographs. No significant differences were observed among the 3 formats in terms of knowledge acquisition or screening participation [24]. In contrast, an intervention study on lung cancer screening [25] reported that presenting an informational film in addition to a booklet was more effective than presenting only a booklet. The addition of films was found to enhance knowledge acquisition, both objectively and subjectively [25]. The absence of a statistically significant difference observed in this study may be explained by the fact that the outcome measure focused on changes in preference rather than knowledge acquisition. A systematic review of the effectiveness of animated videos found that they consistently led to greater knowledge acquisition than other educational materials and standard care; however, no clear evidence was available regarding their effectiveness in influencing attitudes and perceptions [20]. Therefore, future intervention studies are required to identify appropriate methods for presenting animated videos.

The absence of statistically significant differences in the primary outcome may also be explained in other ways. The intervention intensity may have been insufficient: all videos were brief and delivered in a single exposure, which might limit their effect on preferences. Measurement sensitivity may have been limited. The primary outcome was dichotomized into guideline-based preference vs all other options immediately after exposure, which may have masked more nuanced shifts, including the avoidance of non–guideline-based options. In addition, baseline preferences and audience heterogeneity may have contributed to the null findings. Nearly half of the respondents (565/ 1200, 47.1%) preferred guideline-based screening even before viewing the videos, which may have reduced the room for improvement (ceiling effects). Conversely, respondents with strong preexisting beliefs favoring early or intensive screening may require more tailored or repeated messaging to shift preferences. Considering these reasons, future intervention studies should be designed to examine differences in effectiveness across educational methods.

Narrative videos that focus on breast cancer may be less effective or appropriate for male audiences. The scores for relevance and willingness to undergo guideline-based screening for video B were lower among men than among women. Narrative transportation theory suggests that narrative messages are more persuasive when viewers perceive the story as personally relevant and can identify with the protagonist [19]. A breast cancer–focused narrative may have been less self-relevant for male respondents, potentially reducing narrative engagement and transportation and thereby explaining the lower ratings for perceived relevance and willingness to undergo guideline-based screening observed among men for video B. Audience engagement has been identified as a key factor even in the effectiveness of educational videos [26]. Because breast cancer is typically not viewed as a personal health issue by men, it may be less effective in fostering engagement among men. This interpretation is further supported by the finding that women tended to rate the breast cancer video more favorably than video C, which showed a male undergoing multiple tumor marker tests.

The harmful effects of screening are often regarded by screening participants as important prior knowledge when making informed decisions. A qualitative study in Australia [14] used focus group discussions on harms regarding mammography and reported that several participants considered this information important for making informed choices and expressed a strong desire to be informed in advance. However, a randomized controlled trial conducted in Australia [27] showed that informing participants about the harm of breast cancer screening did not reduce their willingness to undergo screening and helped maintain their level of knowledge. Although providing balanced information is important, public information sources often fail to adequately address the associated risks [28]. In this regard, video-based education offers repeatability and accessibility across geographic boundaries [29]. Accordingly, video B in this study helps resolve this issue for women.

A before-and-after comparison of video-based education suggests a potential influence on participants' intentions. Although the videos were only 100 seconds long, an increase in the preference for guideline-based cancer screening was observed. Individuals who undergo annual colorectal cancer screening are generally more likely to actively participate in screening for other cancers. If such individuals become more inclined to follow guideline-based screening after learning about false-positive results, it may indicate that the educational videos were effectively communicated to an appropriate audience.

Notably, education regarding the risks of non–guideline-based screening had a limited impact on changing individuals' preferences for guideline-based cancer screening. In this study, fewer than 1 in 4 respondents changed their preference from non–guideline-based to guideline-based screening after watching the videos. This may be partly explained by the misleading positive feedback loops that are intrinsic to cancer screening [30]. When cancer is detected through screening, the benefit of early detection tends to be emphasized, whereas a negative result reinforces the perception that one is responsible for one's health. Moreover, the cost and pain associated with follow-up examinations are often perceived as acceptable trade-offs for survival [30]. In addition, negative feedback from personal screening experiences is rare, which may help sustain positive perceptions. From a physician's perspective, considering the seriousness of identifying prostate cancer after not undergoing screening, the decision to undergo screening is a no-loss choice compared with not undergoing screening [31]. Furthermore, specific harms, such as fear, pain, and financial burden, are prioritized by physicians to explain the harms of screening. Consequently, problem-based learning curricula aimed at helping medical students understand harm are being developed and evaluated for effectiveness to address this unconscious bias [32].

This study contributes to the literature on cancer screening education in certain ways. First, although many educational video interventions primarily aim to increase screening participation, our videos specifically communicated the potential harms of non–guideline-based screening and evaluated changes

XSL•FO
RenderX

in preference for guideline-based screening. Second, we compared 3 brief videos in online delivery settings and assessed multiple perceptions that can guide the development of video-based educational materials. Third, observed sex differences in perceived relevance for the breast cancer narrative suggest that matching content to the target audience is important. Fourth, considering the trend toward earlier and excessive cancer screening in Japan, the finding that a large number of participants still preferred screening beyond recommended levels even after viewing the educational materials is an important insight. Future efforts should focus on clarifying the underlying problems and exploring effective intervention methods.

Limitations

Despite the positive outcomes, this study has some limitations. First, the generalizability of the results may be limited. The respondents in this study were individuals who were able to spend at least 5 minutes completing the online survey. Therefore, the findings may not be generalizable to populations that are less inclined to participate in online surveys. Second, the primary outcome was a self-reported preference assessed immediately after video exposure, which may not translate into actual screening behaviors in real-world settings. This limits the external validity and practical applicability of our findings. Future studies should incorporate longer follow-up periods and objective end points, such as medical records from hospitals and clinics, to evaluate whether brief video-based education results in behavioral change. Third, the circumstances under which respondents watched the educational videos during the survey, likely influencing their engagement and responses, are unclear. At the beginning of the survey, the respondents were instructed to watch the video until the end while listening to the audio; however, we were unable to verify whether they followed these instructions. For example, if the participants joined the survey while on a train without using headphones, they may have completed it without listening to the audio. Fourth, the before-and-after analysis was exploratory and not prespecified in the trial protocol. Although this analysis provides insight into immediate within-person shifts after exposure, it may be susceptible to analytic flexibility and should be interpreted cautiously. Therefore, the primary inference of this study is based on the prespecified between-group comparison of video formats.

Conclusions

In this online pseudorandomized trial, narrative videos did not demonstrate a greater effect than explanatory videos in increasing preferences for guideline-based cancer screening. This negative finding suggests that concise, single-video interventions are unlikely to produce large format-specific effects and that preferences for excessive screening may persist despite education about the harms of false-positive results. To address beliefs supporting screening beyond recommended levels, future interventions should consider increasing intensity and tailoring education content to individual characteristics.

Acknowledgments

The authors acknowledge the support of Rakuten Insight for facilitating participant recruitment and administering the web-based survey using their proprietary online panels. The authors also thank Tetsu Kuribayashi and colleagues at ABEJA Inc for their assistance in producing the educational videos. The authors are grateful to Editage for editing and reviewing this manuscript for English-language clarity. During manuscript preparation, GPT-5.2 [33] was used by authors to assist with English-language editing.

Funding

This study was supported by a Health Labour Sciences Research Grant from the Ministry of Health, Labour and Welfare (grant 23EA1003).

Data Availability

Data sharing is not applicable to this article as no data sets were generated or analyzed during this study. Reasonable requests for additional information regarding the study can be directed to the corresponding author.

Authors' Contributions

K Sakai and MT conceptualized and designed this study; K Sakai collected the data and performed the statistical analysis; and KF, YF, SN, K Sano, and MT contributed to the interpretation of the results and drafted the manuscript. All authors have reviewed, revised, and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT 2025 checklist.
[DOCX File , 35 KB - jmir_v28i1e82322_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Structures of 3 types of cancer screening promotion videos.
[DOCX File , 5100 KB - jmir_v28i1e82322_app2.docx]

Multimedia Appendix 3
Video A. Explanatory video on guideline-based cancer screening and false positives (84 seconds). The presenter uses a whiteboard and figures to describe the screening pathway from screening to confirmatory testing and to explain why false positives are more common when cancer prevalence is low. The video summarizes potential harm, such as anxiety, time, and costs of follow-up testing, alongside benefits of early detection and concludes with the message to follow evidence-based screening guidelines.
[MP4 File (MP4 Video), 98225 KB - jmir_v28i1e82322_app3.mp4]

Multimedia Appendix 4
Video B. Narrative video showing a false-positive experience after early-age breast cancer screening (77 seconds). A female employee receives an abnormal result, struggles to arrange follow-up, and experiences stress and practical burdens, such as repeated visits and time off work, before learning that no cancer is present. The video highlights that breast cancer risk is low at age 30 and ends by recommending starting the screening at the recommended age.
[MP4 File (MP4 Video), 158564 KB - jmir_v28i1e82322_app4.mp4]

Multimedia Appendix 5
Video C. Narrative video depicting false-positive results after tumor-marker testing (77 seconds). A male employee undergoes tumor-marker testing, receives abnormal results, and undergoes multiple follow-up examinations over several months, resulting in no abnormality. The video emphasizes the harms of unnecessary testing, such as monetary costs and repeated radiation exposure, and recommends to choose guideline-based cancer screening.
[MP4 File (MP4 Video), 74441 KB - jmir_v28i1e82322_app5.mp4]

Multimedia Appendix 6
Complete results from the analysis for the primary outcome.
[DOCX File , 22 KB - jmir_v28i1e82322_app6.docx]

Multimedia Appendix 7
All results from the 7-item evaluation with mean score with 95% CI for each video.
[DOCX File , 21 KB - jmir_v28i1e82322_app7.docx]

Multimedia Appendix 8
Density distribution of total response time in 3 change groups.
[DOCX File , 613 KB - jmir_v28i1e82322_app8.docx]

References

1. Crosby D, Bhatia S, Brindle KM, Coussens LM, Dive C, Emberton M, et al. Early detection of cancer. Science 2022 Mar 18;375(6586):eaay9040. [doi: 10.1126/science.aay9040] [Medline: 35298272]
2. Wright K, Shkabari B, Booth C, Knopf K, Tregear M, Gyawali B. Congruence of cancer screening recommendations between the USPSTF and the top ten US cancer centers: a cross sectional study. EClinicalMedicine 2025 Apr;82:103169 [FREE Full text] [doi: 10.1016/j.eclinm.2025.103169] [Medline: 40235947]
3. Hamashima C. Cancer screening guidelines and policy making: 15 years of experience in cancer screening guideline development in Japan. Jpn J Clin Oncol 2018 Mar 01;48(3):278-286. [doi: 10.1093/jjco/hyx190] [Medline: 29315389]
4. Leavy MB, Starzyk K, Myers E, Curhan G, Gliklich R. Using real-world evidence to support a changing paradigm for cancer screening: a commentary. Pharmacoepidemiol Drug Saf 2020 Oct;29(10):1312-1315. [doi: 10.1002/pds.5104] [Medline: 32902061]
5. Ghotbi N, Iwanaga M, Ohtsuru A, Ogawa Y, Yamashita S. Cancer screening with whole-body PET/CT for healthy asymptomatic people in Japan: re-evaluation of its test validity and radiation exposure. Asian Pac J Cancer Prev 2007;8(1):93-97 [FREE Full text] [Medline: 17477780]
6. Rintala S, Dahlstrom KR, Franco EL, Louvanto K. A synthesis of evidence for cancer-specific screening interventions: a preventive medicine golden jubilee review. Prev Med 2023 Feb;167:107395 [FREE Full text] [doi: 10.1016/j.ypmed.2022.107395] [Medline: 36565859]
7. Rasmussen JF, Siersma V, Malmqvist J, Brodersen J. Psychosocial consequences of false positives in the Danish Lung Cancer CT Screening Trial: a nested matched cohort study. BMJ Open 2020 Jun 04;10(6):e034682 [FREE Full text] [doi: 10.1136/bmjopen-2019-034682] [Medline: 32503869]

XSL•FO
RenderX

8. Croswell JM, Kramer BS, Kreimer AR, Prorok PC, Xu J, Baker SG, et al. Cumulative incidence of false-positive results in repeated, multimodal cancer screening. Ann Fam Med 2009;7(3):212-222 [FREE Full text] [doi: 10.1370/afm.942] [Medline: 19433838]
9. Kasahara Y, Kawai M, Tsuji I, Tohno E, Yokoe T, Irahara M, et al. Harms of screening mammography for breast cancer in Japanese women. Breast Cancer 2013 Oct;20(4):310-315. [doi: 10.1007/s12282-012-0333-6] [Medline: 22282164]
10. Tsunematsu M, Kakehashi M. An analysis of mass screening strategies using a mathematical model: comparison of breast cancer screening in Japan and the United States. J Epidemiol 2015;25(2):162-171 [FREE Full text] [doi: 10.2188/jea.JE20140047] [Medline: 25483105]
11. Ford ME, Havstad SL, Flickinger L, Johnson CC. Examining the effects of false positive lung cancer screening results on subsequent lung cancer screening adherence. Cancer Epidemiol Biomarkers Prev 2003 Jan;12(1):28-33. [Medline: 12540500]
12. Miglioretti DL, Abraham L, Sprague BL, Lee CI, Bissell MC, Ho TH, et al. Association between false-positive results and return to screening mammography in the breast cancer surveillance consortium cohort. Ann Intern Med 2024 Oct;177(10):1297-1307. [doi: 10.7326/M24-0123] [Medline: 39222505]
13. Klabunde CN, Vernon SW, Nadel MR, Breen N, Seeff LC, Brown ML. Barriers to colorectal cancer screening: a comparison of reports from primary care physicians and average-risk adults. Med Care 2005 Sep;43(9):939-944. [doi: 10.1097/01.mlr.0000173599.67470.ba] [Medline: 16116360]
14. Hersch J, Jansen J, Barratt A, Irwig L, Houssami N, Howard K, et al. Women's views on overdiagnosis in breast cancer screening: a qualitative study. BMJ 2013 Jan 23;346:f158 [FREE Full text] [doi: 10.1136/bmj.f158] [Medline: 23344309]
15. Noman S, Shahar HK, Abdul Rahman H, Ismail S, Abdulwahid Al-Jaberi M, Azzani M. The effectiveness of educational interventions on breast cancer screening uptake, knowledge, and beliefs among women: a systematic review. Int J Environ Res Public Health 2020 Dec 31;18(1):e33 [FREE Full text] [doi: 10.3390/ijerph18010263] [Medline: 33396424]
16. Wu S, Chalela P, Ramirez AG. Changes in knowledge and awareness for a community-based cancer screening educational program. Arch Public Health 2023 Jul 12;81(1):130 [FREE Full text] [doi: 10.1186/s13690-023-01144-w] [Medline: 37438840]
17. Minamitani M, Tatemichi M, Mukai T, Katano A, Ohira S, Nakagawa K. Adherence to national guidelines for colorectal, breast, and cervical cancer screenings in Japanese workplaces: a survey-based classification of enterprises' practices into "overscreening," "underscreening," and "guideline-adherence screening". BMC Public Health 2024 Aug 15;24(1):2223 [FREE Full text] [doi: 10.1186/s12889-024-19775-1] [Medline: 39148101]
18. Minamimoto R, Senda M, Jinnouchi S, Terauchi T, Yoshida T, Murano T, et al. The current status of an FDG-PET cancer screening program in Japan, based on a 4-year (2006-2009) nationwide survey. Ann Nucl Med 2013 Jan 20;27(1):46-57 [FREE Full text] [doi: 10.1007/s12149-012-0660-x] [Medline: 23086544]
19. Green MC, Fitzgerald K. Transportation theory applied to health and risk messaging. In: Nussbaum JF, editor. Oxford Research Encyclopedia of Communication. Oxford, UK: Oxford University Press; 2017.
20. Moe-Byrne T, Evans E, Benhebil N, Knapp P. The effectiveness of video animations as information tools for patients and the general public: a systematic review. Front Digit Health 2022;4:1010779 [FREE Full text] [doi: 10.3389/fdgth.2022.1010779] [Medline: 36777110]
21. Lau J, Lim TZ, Jianlin Wong G, Tan KK. The health belief model and colorectal cancer screening in the general population: a systematic review. Prev Med Rep 2020 Dec;20:101223 [FREE Full text] [doi: 10.1016/j.pmedr.2020.101223] [Medline: 33088680]
22. Correia D, Kokole D, Rehm J, Tran A, Ferreira-Borges C, Galea G, et al. Effect of alcohol health warning labels on knowledge related to the ill effects of alcohol on cancer risk and their public perceptions in 14 European countries: an online survey experiment. Lancet Public Health 2024 Jul;9(7):e470-e480 [FREE Full text] [doi: 10.1016/S2468-2667(24)00102-6] [Medline: 38942558]
23. Abu Abed M, Himmel W, Vormfelde S, Koschack J. Video-assisted patient education to modify behavior: a systematic review. Patient Educ Couns 2014 Oct;97(1):16-22. [doi: 10.1016/j.pec.2014.06.015] [Medline: 25043785]
24. Housten AJ, Kamath GR, Bevers TB, Cantor SB, Dixon N, Hite A, et al. Does animation improve comprehension of risk information in patients with low health literacy? A randomized trial. Med Decis Making 2020 Jan;40(1):17-28 [FREE Full text] [doi: 10.1177/0272989X19890296] [Medline: 31795820]
25. Ruparel M, Quaife SL, Ghimire B, Dickson JL, Bhowmik A, Navani N, et al. Impact of a lung cancer screening information film on informed decision-making: a randomized trial. Ann Am Thorac Soc 2019 Jun;16(6):744-751 [FREE Full text] [doi: 10.1513/AnnalsATS.201811-841OC] [Medline: 31082267]
26. Brame CJ. Effective educational videos: principles and guidelines for maximizing student learning from video content. CBE Life Sci Educ 2016;15(4):33 [FREE Full text] [doi: 10.1187/cbe.16-03-0125] [Medline: 27789532]
27. Hersch J, Barratt A, McGeechan K, Jansen J, Houssami N, Dhillon H, et al. Informing women about overdetection in breast cancer screening: two-year outcomes from a randomized trial. J Natl Cancer Inst 2021 Nov 02;113(11):1523-1530 [FREE Full text] [doi: 10.1093/jnci/djab083] [Medline: 33871631]
28. Attena F, Cancellieri M, Pelullo CP. Scarce information about breast cancer screening: an Italian websites analysis. Medicine (Baltimore) 2016 Dec;95(50):e5615 [FREE Full text] [doi: 10.1097/MD.0000000000005615] [Medline: 27977602]

XSL•FO
RenderX

29. Chatterjee A, Strong G, Meinert E, Milne-Ives M, Halkes M, Wyatt-Haines E. The use of video for patient information and education: a scoping review of the variability and effectiveness of interventions. Patient Educ Couns 2021 Sep;104(9):2189-2199. [doi: 10.1016/j.pec.2021.02.009] [Medline: 33741233]
30. Kramer BS, Croswell JM. Cancer screening: the clash of science and intuition. Annu Rev Med 2009;60:125-137. [doi: 10.1146/annurev.med.60.101107.134802] [Medline: 18803476]
31. Ransohoff DF, McNaughton Collins M, Fowler FJ. Why is prostate cancer screening so common when the evidence is so uncertain? A system without negative feedback. Am J Med 2002 Dec 01;113(8):663-667. [doi: 10.1016/s0002-9343(02)01235-4] [Medline: 12505117]
32. Ekhlas S, Lang E, Dickinson J. Optimizing overdiagnosis education for medical students. BMJ Evid Based Med 2023;28(Suppl 1):A29. [doi: 10.1136/EBM-2023-POD.60]
33. GPT-5.2. OpenAI. URL: https://openai.com/index/introducing-gpt-5-2/ [accessed 2025-05-25]

Abbreviations

aOR: adjusted odds ratio
PET-CT: positron emission tomography-computed tomography

Edited by A Schwartz, M Balcarras; submitted 13.Aug.2025; peer-reviewed by S Osera, Y Igarashi; comments to author 16.Dec.2025; accepted 19.Jan.2026; published 12.Feb.2026.

Please cite as:
Sakai K, Furuya Y, Nakazawa S, Fukai K, Sano K, Tatemichi M
Effectiveness of Educational Videos in Encouraging Preferences for Guideline-Based Cancer Screening in Japan: Three-Arm Pseudorandomized Controlled Trial
J Med Internet Res 2026;28:e82322
URL: https://www.jmir.org/2026/1/e82322
doi:10.2196/82322
PMID:

©Kosuke Sakai, Yuko Furuya, Shoko Nakazawa, Kota Fukai, Kei Sano, Masayuki Tatemichi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Seeking But Not Discussing Online Health Information With Physicians: Cross-Sectional Survey Study of eHealth Literacy–Empowerment Profiles and Patient-Centered Communication

Qianfeng Lu[1], PhD; Wen Jiao[2], PhD; Angela Chang[3], PhD; Peter Johannes Schulz[1,4], PhD

[1]Faculty of Communication, Culture and Society, Università della Svizzera italiana, Lugano, Switzerland

[2]School of Communication, Soochow University, Soochow, China

[3]Faculty of Social Sciences, University of Macau, Macau, China

[4]Department of Communication and Media, Ewha Womans University, Seoul, Republic of Korea

Corresponding Author:
Peter Johannes Schulz, PhD
Faculty of Communication, Culture and Society
Università della Svizzera italiana
Via Giuseppe Buffi 13
Lugano, 6900
Switzerland
Phone: 41 586664724
Email: schulzp@usi.ch

Abstract

Background: Patients frequently search for health information online and value physician support in evaluating and interpreting their findings, yet many hesitate to share their online searches with their physicians. This hesitation hinders shared decision-making and compromises patient care. While extensive research has examined patients' online health information–seeking behaviors, little has focused on patients' disclosure of this information to their physicians during consultations.

Objective: Guided by the Health Empowerment Model and the Linguistic Model of Patient Participation in Care, this study aims to (1) identify distinct patient profiles based on eHealth literacy and psychological health empowerment levels, (2) examine how these patient profiles differ in online health information seeking and disclosure to physicians, and (3) investigate whether patient-centered communication (PCC) promotes information disclosure and whether this effect varies by patient profile.

Methods: This cross-sectional study surveyed 2001 Chinese participants recruited through convenience sampling. Patient profiles were identified using *k*-means cluster analysis with standardized *z* scores of eHealth literacy and psychological health empowerment. Differences between profiles in information behaviors were examined using 1-way Welch ANOVA, chi-square tests, and pairwise comparisons. Regression analyses examined the association between PCC and disclosure of online health information. Moderation analyses using the Hayes PROCESS macro assessed whether this association varied across patient profiles.

Results: Four distinct patient profiles were identified: effective self-managers (996/2001, 49.8%), moderate-needs dependent patients (408/2001, 20.4%), high-needs patients (68/2001, 3.4%), and dangerous self-managers (529/2001, 26.4%). Profiles differed significantly in information-seeking intentions ($F_{3,289}$=62.09; P<.001; η^2=0.12) and disclosure intentions ($F_{3,299.41}$=66.08; P<.001; η^2=0.09). "Effective self-managers" showed the highest seeking (mean 4.01, 95% CI 3.96-4.06) and disclosure intentions (mean 3.43, 95% CI 3.36-3.50), while "high-needs patients" showed the lowest intentions for both behaviors. Actual information-seeking rates also differed significantly across profiles (χ^2_3=103.4; P<.001), with "effective self-managers" having the highest rate (800/996, 80.3%) and "high-needs patients" the lowest (25/68, 36.8%). Among seekers, disclosure rates varied significantly (χ^2_3=23.1; P<.001), with "high-needs patients" showing the highest disclosure (16/25, 64%) despite having the lowest seeking rate. PCC was positively associated with actual information disclosure behavior (odds ratio 1.26, 95% CI 1.04-1.53; P=.02), with no significant moderation by patient profiles (χ^2_3=1.7; P=.64).

XSL•FO
RenderX

Conclusions: This study extends existing literature from information-seeking behavior to patients' disclosure of online findings to physicians. Unlike prior research that examined eHealth literacy and psychological health empowerment separately, this study integrated these constructs to identify meaningful patient profiles with distinct information behavior patterns. PCC facilitates disclosure regardless of patient profile. For practice, physicians should adopt a PCC that acknowledges patients' online research efforts, promoting safer information use and stronger patient-physician relationships.

(J Med Internet Res 2026;28:e78836) doi:10.2196/78836

KEYWORDS

empowerment; literacy; patient-provider communication; patient-centered communication; health information; misinformation

Introduction

Background

The use of digital platforms and online information channels is becoming increasingly common for patients seeking health-related information [1,2]. Accessing online health information offers various benefits to patients. By seeking health information online, patients can gain social support, be better informed, and become more prepared to participate in their health care [3-5]. However, drawbacks exist. Online searches may increase patient anxiety, create confusion, and even lead to medical nonadherence [6,7]. The risks are associated with the potential misuse of online health information [8,9], as a significant portion of online content contains misinformation [10] or may be incomplete or unsuitable for a patient's specific condition [11].

Physicians can play a crucial role in guiding patients to navigate online health information [12]. Patient perspectives showed that they generally want physicians to help them understand and assess the suitability of online health information for their specific condition [4,13]. Medical consultations provide valuable opportunities for patients to receive professional advice tailored to their specific conditions. When physicians discuss online health information with patients during medical encounters, they can identify and possibly correct misconceptions patients acquired through internet searches, potentially preventing the misuse of online materials [14]. Additionally, when physicians and patients discuss internet-sourced health information, it helps doctors better understand patients' perceptions of their conditions and health care needs [15]. This understanding, even when it reveals discrepancies between professional medical advice and patients' personal beliefs, creates opportunities for forming more engaged and shared decision-making [14]. Interview studies have shown that conversations about internet-sourced information can enhance patient engagement during consultations [16], strengthen the physician-patient relationship [17], and foster mutual trust [18].

Despite these benefits, many patients still withhold internet-sourced health information from their physicians [19,20]. Barriers to discussing online findings include limited consultation time, concerns about online information quality (such as conflicting or inaccurate content), and patients' perceptions that physicians may be unwilling to engage in such discussions [21,22]. With patients increasingly turning to online resources as their first source of health information prior to consultations [23], it is essential to understand not only their search behaviors but also what influences their decision to disclose or withhold these findings during medical encounters. However, the existing literature has primarily focused on search behaviors while largely neglecting the disclosure aspects [21].

This study examines patients' online health information seeking and their disclosure of these findings to physicians during medical encounters. We draw on the Health Empowerment Model (HEM) [24] to understand how patient characteristics shape information behaviors, and the Linguistic Model of Patient Participation in Care [25] to examine how physician communication influences disclosure.

Theoretical Framework

Health Empowerment Model

The HEM recognizes the distinct but complementary roles of psychological health empowerment and health literacy in influencing health behaviors [24]. Health literacy refers to the ability to access, understand, and evaluate health information [26], and psychological health empowerment is more about the patient's sense of control, confidence, and motivation to act actively in managing their health [27]. Without literacy skills, patients may struggle to identify reliable health information; without empowerment, even accurate information may go underused [24]. Together, these two factors enable patients to make informed decisions and actively manage their health [24]. This model is especially valuable in the digital age, where the use of online health information has increased rapidly, intensifying the need for patients to possess both the literacy skills to accurately evaluate online information (eHealth literacy) and the motivation and confidence to seek and use it (psychological health empowerment) [28]. Depending on patients' ability to navigate online content, online health information can empower but may also misinform them [29].

The HEM delineates 4 distinct combinations of empowerment and health literacy, each of which can be associated with health information behaviors uniquely [30]. The "effective self-manager" represents the ideal scenario: an empowered patient with high eHealth literacy who proactively seeks out and acts on that online health information, making informed decisions based on their ability to evaluate information accurately. In contrast, the "dangerous self-manager" is empowered but has low eHealth literacy. They are proactive in seeking out online information, yet lack the skills to evaluate it accurately, which may lead to poor or even harmful health decisions [24]. The "needlessly dependent patient" has the high literacy skills to understand online health information but lacks the empowerment to act on it, often deferring decisions to others despite having the knowledge to engage more actively in their

XSL•FO
RenderX

care. Finally, the "high-needs patient" is deficient in both empowerment and eHealth literacy, typically taking a passive role in their health care and requiring support to find, understand, and apply online health information [24].

The HEM has been applied to examine patient health status [31], help-seeking behaviors [32], medical decision-making [33], chronic disease management [34,35], and participation in online health communities [36]. These studies consistently demonstrate that patients require both high empowerment and health literacy to effectively manage their health and use health resources, including online health information [37]. Deficiencies in either factor lead to suboptimal health behaviors and poorer outcomes. The interplay between empowerment and literacy creates different patient profiles with varied ways of managing health, seeking information, and engaging with health care providers [24]. Building on this framework, this study aims to identify distinct patient categories based on the combination of these two factors in the context of online health information behaviors.

eHealth Literacy

The traditional concept of health literacy is defined as the personal cognitive and social skills to access, understand, and use health information. eHealth literacy extends it by emphasizing the knowledge and competencies required to navigate electronic information platforms [38-40]. The increasing prevalence of health information and proliferation of misinformation on the internet have made these skills more urgent than ever [41]. Individuals with higher eHealth literacy are more likely to seek health information online [42,43]. For example, empirical studies have found that individuals with higher levels of eHealth literacy are more likely to use health applications to acquire information [44,45], access social media platforms to find and share health-related content [43], and explore national health information portals for health care [46].

Patients' eHealth literacy skills shape both the quality of information they find online and their communication with health care providers [47,48]. A systematic review found that patients with low health literacy had more difficulty evaluating online health information and showed less trust in such information [29]. Conversely, those with higher literacy communicate more clearly with providers as they can articulate their problems and information needs more precisely [47]. While these review findings were based on general health literacy, they likely extend to eHealth literacy, given that fundamental health literacy skills are also core components of eHealth literacy [39]. Empirical findings on whether eHealth literacy leads patients to disclose their internet findings with providers are mixed: some studies show a positive association [40], whereas others report no significant link [49]. These inconsistencies indicate that, beyond the ability to handle online information, a motivational driver may also be important to consider. Drawing on the HEM, we include psychological health empowerment as a complementary determinant to examine alongside eHealth literacy.

Psychological Health Empowerment

A concept that often intertwines with health literacy is psychological health empowerment [50]. Psychological health empowerment is also frequently associated with online health information behavior but often studied separately from eHealth literacy [4]. It is sometimes referred to as health empowerment or patient empowerment [27,51]. Health empowerment has been consistently associated with more active online health information seeking. Patients view online information-seeking as a means to be informed and prepared to manage their health conditions [4]. By accessing online health resources, patients perceive greater control over their conditions [52] and increased confidence when communicating with physicians during medical consultations [53,54]. In essence, while empowered patients are more likely to seek information, the act of seeking and finding relevant health information online further enhances their sense of empowerment [37].

Health empowerment also determines how information is integrated into health decision-making [55]. When patients feel empowered, they actively engage in decision-making with health care providers [56]. This sense of empowerment motivates patients to take an active role in managing their health, including seeking and using online health resources. Empowered patients are more likely to bring their online findings into clinical discussions with physicians as a means to address their information needs, clarify uncertainties, and participate in decision-making [57]. In contrast, patients with low empowerment often remain passive in health care interactions with their providers [57,58]. Despite potentially having found the same information online, these less-empowered patients may hesitate to disclose their findings during consultations, instead deferring entirely to physician direction [24].

To sum up, eHealth literacy equips patients with the skills needed to critically assess the reliability and relevance of online content they encounter [39]. Psychological health empowerment, on the other hand, reflects patients' motivation and confidence to act on this information and actively participate in their health care [59]. The HEM posits that both eHealth literacy and psychological health empowerment are necessary for effective use of health information and that their combination determines how patients engage with health information. Although previous research has examined these two concepts, they were typically studied separately, with studies often focusing on either health literacy or health empowerment but failing to integrate both [37,60]. To fill this gap, this study also examines health information behaviors across different eHealth literacy and empowerment profiles, from online searching before consultations to information sharing during physician encounters.

Patient-Centered Communication

Despite recognizing the potential benefits of discussing online health information with physicians, many patients hesitate to share their internet findings during consultations [21]. This reluctance stems largely from patients' concerns about how physicians will react to their online search [22,61]. Empirical evidence demonstrates that patients' hesitation to disclose

XSL•FO
RenderX

internet-sourced information is directly related to their communication experience with physicians [4].

Patient-centered communication (PCC)—defined as a clinical communication approach where providers seek to understand patients' perspectives, reach shared understanding, and involve patients as partners in care decisions [62]—offers a promising approach to encourage information sharing. It differs fundamentally from the paternalistic model by positioning patients as equal participants in the clinical relationship. Research shows that PCC promotes patient health care participation and is linked to improved patient trust and satisfaction with providers [63,64]. When physicians adopt patient-centered approaches, patients feel their perspectives are valued [65] and are more likely to share internet-sourced information [21]. Conversely, when patients perceive physicians as unreceptive or threatened by their online research, they are unwilling to disclose [21,52,66].

The importance of physician communication is supported by the Linguistic Model of Patient Participation in Care [25]. This model identifies 3 key determinants of patient participation in medical encounters: predisposing factors, enabling factors, and physician communication behaviors. Predisposing factors encompass patients' sociodemographic and psychosocial characteristics, including psychological health empowerment, that influence their propensity to participate actively in health care encounters. Enabling factors refer to patients' communicative resources and capabilities, such as eHealth literacy skills, that determine their capacity to engage effectively with health care providers and articulate their needs. Physician communication behaviors include verbal and nonverbal practices, such as providing clear explanations and demonstrating empathy, that can either facilitate or inhibit patient participation. The model posits that these 3 factors interact rather than operate independently [25]. It suggests that even patients with high empowerment and literacy may withhold information if physician communication does not support their disclosure. Thus, PCC serves as a crucial contextual factor that can affect information disclosure, regardless of patients' individual characteristics.

PCC may not affect all patients equally. The HEM suggests that patients with different literacy and empowerment levels rely on their physicians differently [24]. Those with high empowerment and literacy may already feel confident sharing online findings, while patients with lower empowerment and literacy may be more sensitive to physician communication styles and depend more heavily on physician encouragement. This raises the question of whether the effect of PCC on disclosure varies across patient profiles. Therefore, the final aim of this study is to examine the impact of PCC on patient information disclosure and whether it varies across different patient profiles.

Based on the above, this study poses the following research questions (RQs):

- RQ1: What distinct patient profiles can be identified based on eHealth literacy and psychological health empowerment?
- RQ2: How do patients' online information-seeking behaviors prior to consultations and their subsequent disclosures to physicians differ across the identified patient profiles?
- RQ3a: How does patient-centered communication relate to patients' disclosure of online health information to physicians? RQ3b: How does this relationship differ across patient profiles?

Methods

This cross-sectional survey study is reported according to CHERRIES (Checklist for Reporting Results of Internet e-Surveys) guidelines [67].

Recruitment

This study focuses on China, the world's largest digital community, with approximately 1.08 billion internet users in 2023 [68]. This cross-sectional survey study was conducted as part of a larger research project examining health information-seeking behavior and health care use in China.

Data collection was conducted in Mainland China and Macau (Special Administrative Region of China) between September and November 2022. The survey was distributed online through Chinese social media platforms, including WeChat, QQ, and Sina Weibo. The survey was created using the Qualtrics platform and took approximately 15 minutes to complete.

Eligible participants were required to be Chinese, aged 18 years or older, and residing in Mainland China or Macau. Snowball convenience sampling was used. Initial recruitment began on the University of Macau campus, where research team members approached individuals in public areas (eg, libraries and classrooms) and invited them to participate. Participants who completed the survey were asked to share it with their friends and family members who met the eligibility criteria.

A total of 3387 respondents answered the survey. The final analytical sample included 2001 participants who provided complete responses on all focal variables (eHealth literacy, health empowerment, and information seeking and disclosure behavior). Among these, 1569 participants had received health care in the previous year and could provide responses regarding physician communication, forming the subsample for PCC analyses. No formal sample size calculation was conducted, and the sample size was determined by the number of complete responses obtained during the data collection period. Survey design and recruitment procedures are also published in Jiao et al [37].

Ethical Considerations

This study received ethical approval from the University of Macau's Ethical Committee for the Social Sciences and Humanities (application code: SSHRE22-APP093-FSS), which covered the complete data collection procedure. The original ethical approval included permission for secondary analyses of the collected data; therefore, no additional approval was required for this study. All participants provided informed consent prior to completing the survey. Participants were informed that the study aimed to help researchers and policymakers better understand population health behaviors related to internet use for health purposes. Participant privacy and confidentiality were

XSL•FO
RenderX

protected through anonymous data collection. No personally identifiable information was collected, and all survey responses were deidentified before data analysis. No compensation was provided to participants. Participation was entirely voluntary, and participants could withdraw at any time without consequence. No images or materials in this manuscript contain identifiable participant information.

Measures

eHealth Literacy

The eHealth Literacy Scale (eHEALS), a validated 8-item instrument developed by Norman and Skinner [69], was used to evaluate participants' eHealth literacy. It uses the operational definition of eHealth literacy as "an individual's ability to seek, find, understand, appraise, and apply health information from electronic sources to address or solve a health problem" and uses self-reported measures to assess perceived competence across these domains. Despite being developed nearly two decades ago, eHEALS continues to receive research attention with recent validation studies confirming its robustness across diverse populations [70,71]. The core competencies it measures (seeking, understanding, appraising, and applying health information from electronic sources) remain essential regardless of the online information source and represent foundational skills needed to navigate today's digital health information landscape.

We used the Chinese translation of eHEALS validated by Chang and Schulz [72], which demonstrated good psychometric properties in Chinese populations and has been empirically applied in studies among Chinese-speaking populations [73,74]. Responses were measured on a 5-point Likert scale, with higher scores indicating greater eHealth literacy. The mean of all item scores was calculated to represent participants' eHealth literacy levels. The reliability of eHEALS was assessed using Cronbach α. Additionally, we reported McDonald ω, which is less affected by the number of items (a limitation of Cronbach α) and does not assume equal factor loadings of items. Both reliability tests showed good internal consistency (Cronbach α=0.93; McDonald ω=0.93). The complete survey wording can be found in Multimedia Appendix 1.

Psychological Health Empowerment

The Psychological Health Empowerment Scale (PHES) was used to measure patient empowerment [27]. This established 8-item instrument assesses individuals' perceived control over one's health, perceived competence of understanding of one's health condition, and their motivation to achieve health goals. Sample items include statements such as "I can motivate myself to manage my health and make a better life" and "I can make every possible effort to achieve health goals." The PHES has been widely adopted across diverse populations to measure health empowerment [75]. We used the Chinese translation by Jiang and Street [76], which has subsequently been used in other studies with Chinese-speaking populations [77,78]. Responses were recorded on a 5-point Likert scale, with higher scores indicating a greater sense of health empowerment. The overall health empowerment level was calculated by averaging the 8 items (Cronbach α=0.90; McDonald ω=0.90).

Patient-Centered Communication

Participants who had received health care in the preceding year (n=1569) were asked about their communication experiences with physicians. The 7-Item Patient-Centered Communication Scale based on Epstein and Street's [79] model was used to assess how frequently physicians (1) provided opportunities to ask health-related questions, (2) paid attention to feelings and emotions, (3) involved patients in health care decisions to their desired extent, (4) ensured understanding of health management tasks, (5) explained medical information comprehensibly, (6) spent adequate time with patients, and (7) helped manage health-related uncertainty. This scale has been widely used to estimate PCC [80,81]. We used the Chinese translation from the Health Information National Trends Survey in China (HINTS-China) [82]. Responses were recorded on a 5-point Likert scale. The overall patient-physician communication was calculated by averaging the 7 items, with higher scores indicating better PCC (Cronbach α=0.89; McDonald ω=0.89).

Information Seeking and Disclosure Behavior

The two facets of participants' health information-seeking behavior in the clinic context, intended versus actual, were assessed. We inquired about their past behaviors regarding whether participants have sought out health information on the internet before consulting with a doctor (actual seeking) and whether such information was subsequently discussed during medical visits (actual disclosure). The two specific survey questions are "Reflecting on your recent medical consultations, did you search for health information on the Internet about your condition before meeting with your doctor?" For those who answered positively (n=1428), they were further queried, "Furthermore, did you discuss the information you found on the Internet with your doctor during the consultation?" A binary response of "yes" (=1) or "no" or "I am not sure" (=0) was applied.

The hypothetical scenarios aimed to elicit intended behaviors. Participants were presented with a situation in which they hypothetically experienced a new health symptom (eg, chest pain) and were asked to predict their likelihood of searching for health information online before a scheduled doctor's appointment (intended seeking) and their propensity to share the found information with the doctor (intended disclosure). The corresponding survey questions were as follows: "Imagine experiencing chest pain for one week, a symptom you've never had before. As a result, you decide to schedule a medical appointment with a new doctor. How likely are you to search for information about your condition on the Internet before visiting the doctor? And how likely are you to share the information you found on the Internet with the doctor?" A 5-point Likert scale ranging from "very unlikely" (=1) to "very likely" (=5) was used. A pilot study with university adults in Macau and feedback from the pilot study were used to refine these survey questions.

Control Variables

We collected participants' demographic information, including age, gender, and education level. Age was recorded in years, while education level was categorized based on the highest

XSL•FO
RenderX

grade completed, ranging from primary school and below (coded as 1) to a bachelor's degree and above (coded as 6). Gender was represented as a binary variable, with "female" coded as 0. In addition, self-rated general health status and trust in internet-sourced health information were included as control variables to minimize confounding effects. Participants were asked to rate their overall health status on a scale from 1 (poor) to 5 (excellent). Trust in health information from social media, websites, search engines, and apps was measured with a 5-point scale ("1=not at all" to "5=a lot"). The mean scores of these trust responses were calculated to provide an overall trust in online health information (Cronbach α=0.89; McDonald ω=0.80).

Statistical Analysis

Descriptive statistics were conducted to ascertain the overall sample characteristics. Then, a *k*-means cluster analysis was performed using the standardized *z* scores of 2 key variables: eHealth Literacy and Psychological Health Empowerment. This analysis was carried out in the R program (R Foundation for Statistical Computing), using the "NbClust" package to identify the optimal number of clusters [83]. The NbClust package determines the optimal number of clusters by computing different cluster validity indices, such as the Silhouette index, Dunn index, and Calinski-Harabasz criterion. Rather than relying on a single criterion, it evaluates multiple indices and provides a consensus solution based on the majority rule. NbClust has been widely used in previous empirical studies across various disciplines. Following the cluster analysis, the distinct patient eHealth Literacy and Empowerment Profiles (eHL-E Profiles) were established.

Then, 1-way ANOVA, using Welch correction to account for potential heterogeneity of variances, was conducted to examine differences in intended information-seeking and disclosure behaviors across the different eHL-E Profiles. Games-Howell post hoc comparisons were performed to identify specific between-group differences. Additionally, chi-square tests were used to assess differences in actual information-seeking and disclosure behaviors among the eHL-E Profiles. Pairwise comparisons with Bonferroni adjustment were further conducted to identify which specific eHL-E Profiles differed significantly from each other in their actual information behaviors.

Finally, regression analyses were performed to examine the association between PCC and patient information disclosure. For the dichotomous outcome actual disclosure (0=no, 1=yes), we fitted a binary logistic regression with PCC as the focal predictor. For the continuous outcome intended disclosure (interval scaled), we ran a multiple linear regression. All models controlled for eHealth literacy, health empowerment, self-rated health, overall trust in online health information, and demographic variables (age, gender, and education) to mitigate confounding bias. To test whether the effect of PCC on disclosure varied by patient profile groups, we estimated a simple moderation model (model 1) in the Hayes PROCESS macro (version 4.3), controlling for self-rated health, trust in online health information, and demographic variables. We requested 5000 bootstrap samples to obtain bias-corrected 95% CIs for all conditional effects.

Results

Sample Characteristics

Descriptive statistics are displayed in Table 1. The mean age of participants was 31.78 (SD 13.58) years, 64.3% (1202/1870) were female, and 82.5% (1508/1829) had received a college degree or higher education. General health status was rated, on average, as good (mean 2.98, SD 0.81). Overall trust in online health information was moderate (mean 2.48, SD 0.69).

XSL•FO
RenderX

Table 1. Demographic and health-related characteristics of study participants (N=2001).

Variables	Values
Age[a] (years), mean (SD)	31.78 (13.58)
Gender[b] , n (%)	
Female	1202 (64.3)
Male	668 (35.7)
Education[c]	
Mean (SD)	4.60 (1.07)
Primary school or below, n (%)	24 (1.3)
Junior middle school, n (%)	58 (3.1)
High school, n (%)	239 (13.1)
Junior college, n (%)	258 (14.1)
Bachelor's degree, n (%)	982 (53.7)
Master's degree or above, n (%)	268 (14.7)
Health status, mean (SD)	2.98 (0.81)
Trust in online health information, mean (SD)	2.48 (0.69)
eHealth literacy, mean (SD)	3.48 (0.70)
Health empowerment, mean (SD)	3.77 (0.59)
Patient-centered communication[d], mean (SD)	2.53 (0.63)
Intended seeking, mean (SD)	3.79 (1.01)
Intended disclosure, mean (SD)	3.19 (1.09)
Actual seeking (yes), n (%)	1428 (71.4)
Actual disclosure[e] (yes), n (%)	800 (56.0)

[a]Among the total sample of 2001 participants, missing data were observed for 139 participants.

[b]Among the total sample of 2001 participants, missing data were observed for 131 participants.

[c]Among the total sample of 2001 participants, missing data were observed for 172 participants.

[d]A total of 1569 participants who had received health care in the previous year provided answers regarding their communication with the physician.

[e]A total of 1428 participants who had searched for health information on the internet before meeting their physicians responded about whether they had discussed this information with their physician.

Patient Profiles Based on eHealth Literacy and Psychological Health Empowerment

RQ1 asked what distinct profiles of patients could be identified based on their eHealth literacy and psychological health empowerment. *K*-means cluster analysis identified a 5-cluster solution as the optimal fit. Two clusters exhibited high scores in both eHealth literacy and health empowerment. Specifically, one had mean scores for eHealth literacy and health empowerment of 4.27 (95% CI 4.21-4.34, SD 0.51) and 4.61 (95% CI 4.57-4.65, SD 0.31), respectively; another had corresponding scores of 3.90 (95% CI 3.88-3.92, SD 0.28) and 3.89 (95% CI 3.87-3.90, SD 0.23). Aligned with the HEM, these clusters exemplify a group with both high literacy and high empowerment, and can be labeled as the "effective self-manager." Therefore, they were amalgamated into a single profile, profile 1, which encompasses 996 participants (996/2001, 49.8%). This aggregated profile demonstrated mean eHealth literacy and health empowerment scores of 3.98 (95% CI 3.96-4.01, SD 0.38) and 4.05 (95% CI 4.02-4.07, SD 0.39), respectively, signifying high levels of both eHealth literacy and empowerment.

The remaining 3 clusters exhibited distinct combinations of eHealth literacy and empowerment levels, and thus we retained these unique eHealth literacy–empowerment profiles without merging. Figure 1 displays the *z* scores for eHealth literacy and health empowerment across the 4 profiles ultimately identified, including the aggregated profile 1, the effective self-manager. Table 2 lists their mean scores.

Profile 2 (408/2001, 20.4%) exhibited a comparatively low level of eHealth literacy (mean 3.25, 95% CI 3.21-3.29, SD 0.41) and an even lower level of empowerment (mean 3.09, 95% CI 3.06-3.12, SD 0.3). In the HEM, a group characterized by high literacy but low empowerment is labeled as a needlessly dependent patient. However, the current profile's eHealth literacy level, being below the average, cannot be considered high. Therefore, it is labeled as the "moderate-needs dependent patient."

XSL•FO
RenderX

Profile 3 (68/2001, 3.4%) showed the lowest levels of eHealth literacy (mean 1.67, 95% CI 1.52-1.81, SD 0.62) and health empowerment (mean 2.33, 95% CI 2.14-2.51, SD 0.76), indicating substantial deficiencies in both literacy and empowerment. According to HEM, it is labeled as the "high-needs patient."

Finally, profile 4 (529/2001, 26.4%) exhibited mean scores for eHealth literacy of 2.95 (95% CI 2.92-2.98, SD 0.41) and health empowerment of 3.95 (95% CI 3.92-3.97, SD 0.28). This suggests individuals in this profile possess a combination of low literacy with high empowerment, labeled as the "dangerous self-manager" according to the HEM.

Beyond having distinct combinations of eHealth literacy and empowerment levels, Welch ANOVA confirmed significant differences in the values of both eHealth literacy ($F_{3,287.48}$=1103.55; P<.001) and health empowerment ($F_{3,289.77}$=991.5; P<.001) among the 4 labeled profiles, with all pairwise comparisons reaching statistical significance (all P<.001, Table 2), confirming that these profiles are statistically distinct in both their combination patterns and magnitude levels of eHealth literacy and empowerment.

Figure 1. Standardized z scores for eHealth literacy and psychological health empowerment by patient profile.

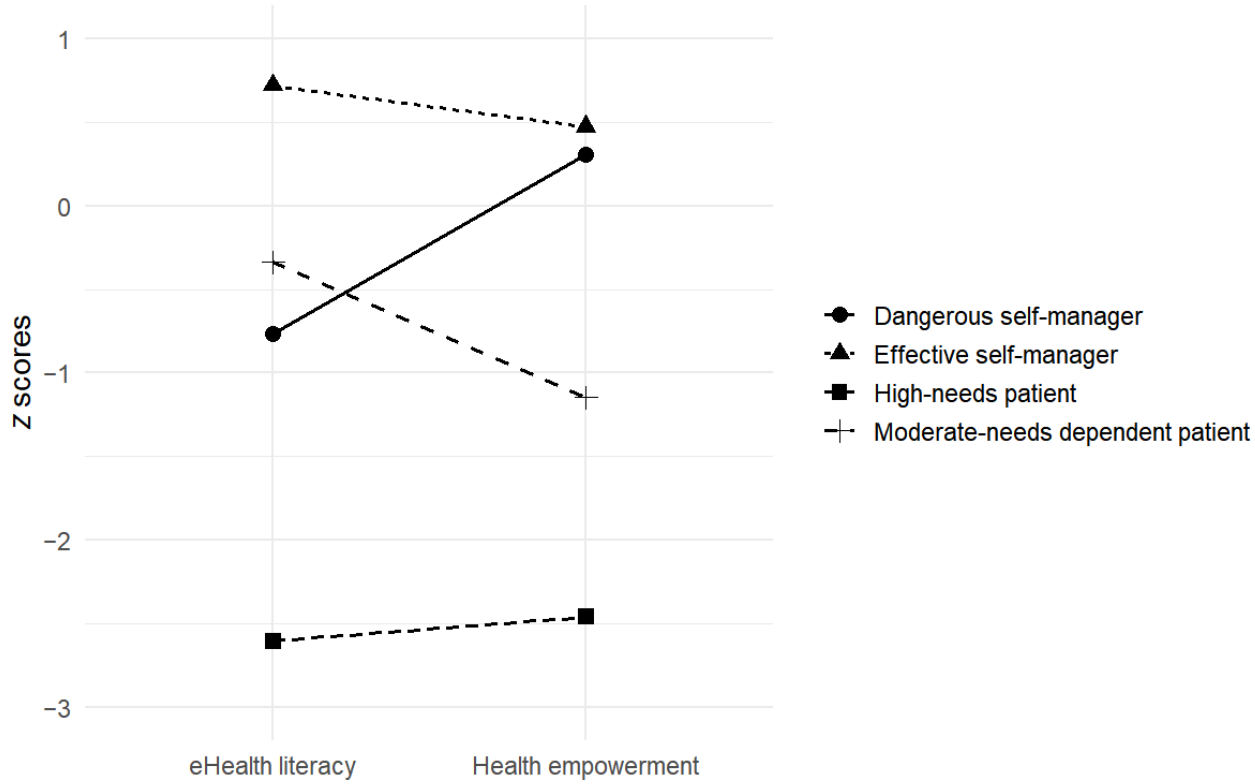

Table 2. Descriptive statistics and group comparisons for eHealth literacy and psychological health empowerment by patient profile[a].

Variables	Profile 1 (n=996): Effective self-manager	Profile 2 (n=408): Moderate-needs dependent patient	Profile 3 (n=68): High-needs patient	Profile 4 (n=529): Dangerous self-manager	Welch ANOVA: *F* test (*df*)
eHealth literacy[b], mean (SD)	3.98 (0.38)$_A$	3.25 (0.41)$_B$	1.67 (0.62)$_C$	2.95 (0.41)$_D$	1103.55 (3, 287.48)[c]
Health empowerment[b], mean (SD)	4.05 (0.39)$_A$	3.09 (0.30)$_B$	2.33 (0.76)$_C$	3.95 (0.28)$_D$	991.50 (3, 289.77)[c]

[a]Welch ANOVA and Games-Howell post hoc tests were conducted due to the violation of the assumption of homogeneity of variance.

[b]Differing subscripts (A, B, C, D) indicate significant mean difference in a row at the significance level of P<.001.

[c]P<.001.

Information Behavior Across Patient Profiles

Welch ANOVA indicated significant differences in intended information-seeking ($F_{3,289}$=62.09; P<.001; η^2=0.12) and intended disclosure ($F_{3,299.41}$=66.08; P<.001; η^2=0.09) across the 4 eHL-E Profiles (see Table 3).

Games-Howell post hoc tests provided detailed results for all pairwise comparisons. Regarding intended information-seeking, the "effective self-manager" group exhibited the highest intention (mean 4.01, 95% CI 3.96-4.06, SD 0.88), while the "high-needs patient" group had the lowest seeking intention (mean 2.13, 95% CI 1.83-2.44, SD 1.26). The "moderate-needs dependent patient" (mean 3.66, 95% CI 3.57-3.76, SD 0.98)

XSL•FO
RenderX

and the "dangerous self-manager" (mean 3.69, 95% CI 3.61-3.78, SD 1.01) groups displayed similar seeking intentions. Except for the comparison between the "moderate-needs dependent patient" and the "dangerous self-manager" groups (mean difference –0.03, 95% CI –0.20 to 0.14; *P*=.97), all other mean differences were statistically significant at *P*<.001.

Regarding intended information disclosure, similar patterns emerged. The "effective self-manager" group once again had the greatest intention (mean 3.43, 95% CI 3.36-3.50, SD 1.02) to share online health information with their physicians. Conversely, the "high-needs patient" group exhibited the lowest intention to disclose (mean 1.85, 95% CI 1.6-2.1, SD 1). The "moderate-needs dependent patient" (mean 3, 95% CI 2.91-3.10, SD 0.98) and "dangerous self-manager" (mean 3.05, 95% CI 2.95-3.14, SD 1.12) groups were comparable. Except for the comparison between the "moderate-needs dependent patient" and the "dangerous self-manager" groups (mean difference 0.04, 95% CI –0.14 to 0.22; *P*=.93), all other group comparisons yielded significant mean differences (*P*<.001).

For actual information-seeking, the "effective self-manager" again had the highest proportion of individuals who reported seeking health information regarding their health condition before seeing the physician (800/996, 80.3%). The "moderate-needs dependent patient" group had a lower rate of 66.7% (272/408), and the "dangerous self-manager" group had a similar rate of 62.6% (331/529). In contrast, the "high-needs patient" group, characterized by the lowest eHealth literacy and empowerment, had the lowest rate, with only 36.8% (25/68) reporting actual seeking behavior. The chi-square test was significant (χ^2_3=103.4; *P*<.001) with a Cramer *V* of 0.227, indicating a moderate association (see Table 4). Pairwise comparisons with Bonferroni adjustment revealed that the "effective self-manager" had significantly higher seeking rates than all other profiles, while the "high-needs patient" had significantly lower rates than all other profiles. The "moderate-needs dependent patient" and "dangerous self-manager" profiles did not differ significantly from each other.

Among those who answered "yes" to actual seeking behavior, their actual disclosure behavior was examined (n=1428). The "effective self-manager" profile had a high proportion of individuals who disclosed health information (465/800, 58.1%). The "moderate-needs dependent patient" profile had a lower rate of 43.8% (119/272), and the "dangerous self-manager" profile had a similar rate of 47.4% (157/331). Interestingly, the "high-needs patient" profile showed the highest proportion of disclosures, with 64% (16/25) of those who searched for information online choosing to disclose it to their physicians. The chi-square test was significant (χ^2_3=23.1; *P*<.001) with a Cramer *V* of 0.127, suggesting a small to moderate association. Pairwise comparisons with Bonferroni adjustment showed that the "effective self-manager" profile had significantly higher disclosure rates than both the "moderate-needs dependent patient" and "dangerous self-manager" profiles (*P*<.05). These latter two profiles did not differ significantly from each other. The "high-needs patient" profile did not differ significantly from any other profile, which is likely due to the small sample size (n=16) limiting statistical power (Table 4).

Table 3. Descriptive statistics and group comparisons of intended information-seeking and disclosure behaviors across 4 patient profiles[a].

Intended information behavior	Profile 1 (n=996): Effective self-manager, mean (SD)	Profile 2 (n=408): Moderate-needs dependent patient, mean (SD)	Profile 3 (n=68): High-needs patient, mean (SD)	Profile 4 (n=529): Dangerous self-manager, mean (SD)	Welch ANOVA: *F* test (*df*)
Intended seeking[b]	4.01 (0.88)$_A$	3.66 (0.98)$_B$	2.13 (1.26)$_C$	3.69 (1.01)$_B$	62.09 (3, 289)[c,d]
Intended disclosure[b]	3.43 (1.02)$_A$	3.00 (0.98)$_B$	1.85 (1.00)$_C$	3.05 (1.12)$_B$	66.08 (3, 299.41)[d,e]

[a]Welch ANOVA and Games-Howell post hoc tests were conducted due to the violation of the assumption of homogeneity of variance.

[b]Differing subscripts (A, B, C) indicate significant mean difference in a row at the significance level of *P*<.001.

[c]η^2=0.12.

[d]*P*<.001.

[e]η^2=0.09.

XSL•FO
RenderX

Table 4. Descriptive statistics and group comparisons of actual information-seeking and disclosure behaviors across 4 patient profiles[a].

Actual information behaviors	Profile 1 (n=996): Effective self-manager, n (%)	Profile 2 (n=408): Moderate-needs dependent patient, n (%)	Profile 3 (n=68): High-needs patient, n (%)	Profile 4 (n=529): Dangerous self-manager, n (%)	Chi-square (*df*)
Actual seeking (yes)[b]	800 (80.3)$_A$	272 (66.7)$_B$	25 (36.8)$_C$	331 (62.6)$_B$	103.4 (3)[c,d]
Actual disclosure (yes)[b,e]	465 (58.1)$_A$	119 (43.8)$_B$	16 (64.0)$_{A,B}$	157 (47.4)$_B$	23.1 (3)[d,f]

[a]Different subscripts (A, B, C) within a row indicate significant differences between profiles at $P<.05$ or lower.

[b]Represents the frequency and percentage for dichotomous variable.

[c]Cramer $V=0.227$.

[d]$P<.001$.

[e]Only the participants who answered "yes" in actual seeking are included; hence, n=1428.

[f]Cramer $V=0.127$.

The Role of PCC in Information Disclosure

We expect that controlling for patient predisposing (health empowerment) and enabling factors (eHealth literacy), PCC would be associated with patients' disclosure of internet-sourced information to physicians. To examine this relationship, we conducted regression analyses. Intended disclosure showed no significant association with PCC (β=–0.007, 95% CI –0.09 to 0.07; $P=.76$). However, actual disclosure behavior demonstrated a positive association with PCC (odds ratio 1.26, 95% CI 1.04-1.53; $P=.02$). This indicates a positive relationship between PCC and patients' disclosure of internet-sourced information to their physicians in medical encounters (see Table 5).

Two single moderator models using the Hayes PROCESS macro (model 1 with a multicategorical moderator) tested whether PCC related differently to information disclosure across the 4 patient-profile clusters. For intended disclosure (continuous outcome), the PCC × patient profile interaction was nonsignificant ($F_{3,1413}=0.24$; $P=.87$); none of the 3 contrast-specific interaction terms was significant (|β|≤0.07, all $P\geq.5$). The results were similar for actual disclosure (binary outcome): the interaction did not reach significance ($\chi^2_3=1.7$; $P=.64$) and each individual interaction term was likewise nonsignificant (all $P\geq.33$). These results indicate that the association between PCC and information disclosure did not significantly vary across patient profiles, suggesting a consistent relationship regardless of empowerment and literacy profiles.

Table 5. Multiple linear regression and logistic regression results for intended and actual information disclosure with physicians[a].

Variables	Intended disclosure[b]		Actual disclosure[c]	
	Coefficients (95% CI)	*P* value	Odds ratio (95% CI)	*P* value
Age	0.00 (0.000 to 0.008)	.06	1.02[e] (1.01 to 1.03)	.003
Gender	0.02 (–0.09 to 0.13)	.76	0.93 (0.72 to 1.21)	.59
Education	0.10[f] (0.05 to 0.15)	<.001	1.01 (0.89 to 1.15)	.88
Health status	–0.04 (–0.11 to 0.03)	.27	1.19[g] (1.01 to 1.39)	.04
Trust in online health information	0.18[f] (0.10 to 0.26)	<.001	1.11 (0.92 to 1.34)	.30
eHealth literacy	0.31[f] (0.22 to 0.40)	<.001	1.32[g] (1.07 to 1.64)	.01
Health empowerment	0.20[f] (0.10 to 0.30)	<.001	1.05 (0.82 to 1.34)	.69
PCC[d]	–0.01 (–0.09 to 0.07)	.76	1.26[g] (1.04 to 1.53)	.02

[a]Mean, median, or mode imputation was used for missing values on age, gender, and education.

[b]$R^2=0.11$.

[c]Nagelkerke $R^2=0.04$.

[d]PCC: patient-centered communication.

[e]$P<.01$.

[f]$P<.001$.

[g]$P<.05$.

XSL•FO
RenderX

Discussion

Principal Results

This study identified 4 distinct patient profiles based on eHealth literacy and psychological health empowerment: effective self-managers, moderate-needs dependent patients, high-needs patients, and dangerous self-managers. These profiles demonstrated significantly different patterns in both online health information seeking prior to consultations and disclosure of this information to physicians. PCC was positively associated with actual information disclosure, and this association did not vary across patient profiles. Together, these findings demonstrate that eHealth literacy and empowerment create meaningful patient groups with distinct information behaviors, and PCC plays a consistent role in encouraging disclosure regardless of patient characteristics. Below, we discuss these findings in relation to existing literature and their implications for theory and practice.

First, our identification of 4 distinct patient profiles supports the HEM's core premise that patients can be meaningfully categorized based on their literacy and empowerment levels. Three of these profiles—effective self-managers, high-needs patients, and dangerous self-managers—matched HEM predictions closely [24]. However, we found an interesting deviation. The HEM predicts a "needlessly dependent patient" with high literacy but low empowerment. Instead, we identified "moderate-needs dependent patients" who had both below-average literacy and low empowerment. This finding appears consistent with other empirical applications of HEM. For instance, a recent study classified 269 participants using depression literacy and empowerment; they only identified 2 of the 4 theoretical profiles [36]. Such findings suggest that while the HEM provides a valuable conceptual framework, researchers should anticipate that patient populations may not always exhibit all 4 theorized combinations. The specific profiles that emerge may depend on factors such as the targeted literacy used and the characteristics of the population of interest. For instance, when studying populations that are relatively educated and familiar with online technologies (thus having high eHealth literacy), the theoretically posited profiles of "high-needs patients" and "needlessly dependent patients" with low literacy may not be identified due to the overall high literacy baseline of the sample.

Interestingly, our sample split nearly evenly between "balanced" profiles (effective self-managers and high-needs patients, around 53%, 1064/2001 of the total sample) and "imbalanced" profiles where literacy and empowerment diverged (moderate-needs dependent patients and dangerous self-managers, around 47%, 937/2001). This finding suggests that eHealth literacy and psychological health empowerment do not necessarily develop in parallel. In practice, digital health initiatives aiming for patient empowerment often assume patients have the eHealth literacy needed to grasp their content [84], while online interventions that focus on improving patients' literacy skills often treat empowerment as one of the primary outcomes [85]. Our findings suggest that the assumption of high eHealth literacy inherently leading to empowerment and effective patient engagement with eHealth resources may not hold true. Therefore, targeting health literacy alone may not guarantee corresponding improvements in empowerment and vice versa.

Second, we found significant behavioral differences across patient profiles. These patient profiles not just represent differences in their literacy and empowerment but also differ with their online health information seeking and disclosure patterns. One interesting finding emerged from the "high-needs patient" profile. They present a particularly intriguing information behavior pattern: while they showed the lowest information-seeking behavior, they demonstrated the highest disclosure rates when they did search. Rather than being passive or disengaged, "high-needs patients" appear highly selective in their information-seeking but deeply reliant on physician guidance. This suggests that when patients with low eHealth literacy and empowerment do seek information, they may place exceptional value on professional validation and interpretation.

The "dangerous self-manager" profile raises particular clinical concerns. These patients demonstrate moderately high information-seeking behavior but relatively low disclosure rates, creating a gap between information gathering and sharing with providers. Given their lower eHealth literacy levels, this combination of active seeking with limited disclosure could pose safety risks, as they may be making health decisions based on misinterpreted information without professional guidance.

In contrast, "effective self-managers" demonstrated the most comprehensive use of online health information, showing both high seeking behavior and high disclosure rates with physicians. This finding aligns with prior research linking high literacy and empowerment to more active health care participation and effective patient-provider collaboration [33,86].

Our analysis showed an interesting pattern regarding PCC and information disclosure. While participants' retrospective assessments of PCC quality from past health care encounters showed no association with their intended disclosure behavior in hypothetical future scenarios, PCC was significantly linked to their actual disclosure behavior in those past encounters. This suggests that patients may form intentions about sharing online health information based on factors other than their previous communication experiences. However, the quality of physician communication during actual encounters does influence whether patients disclose their online findings in real time. Therefore, disclosure behavior appears more responsive to the immediate communication context rather than retrospective evaluations of past clinical experiences. This finding aligns with empirical studies showing that physicians' communicative behavior during encounters affects whether patients share their online findings [21,66,87].

Furthermore, PCC demonstrated a statistically significant positive association with information disclosure behavior that was fairly consistent across all 4 patient profiles. This finding is particularly notable given the substantial information on behavioral differences we observed. This universal responsiveness suggests that PCC can promote engagement across diverse patient profiles, regardless of their health literacy or empowerment levels.

XSL•FO
RenderX

Theoretical Implications

This study makes theoretical contributions to the understanding of patient eHealth literacy, empowerment, and the HEM. First, our findings provide empirical support for HEM. Our identification of patient profiles aligns closely with the model's proposed combinations of high and low empowerment and health literacy. These profiles exhibited distinct information behaviors, from comprehensive information seeking and disclosure among "effective self-managers" to limited engagement among "high-needs patients" and "dangerous self-managers." While previous HEM studies are primarily based on Western health contexts, our study demonstrates the model's validity in Chinese health care settings. Additionally, our study extends HEM by integrating physician-patient communication as a contextual factor, which was not originally included in the model. Specifically, we demonstrated that PCC encourages patients to disclose online health information to their physicians, regardless of their empowerment or literacy combination. This finding indicates that future applications of HEM should incorporate interpersonal communication aspects as relevant contextual variables when examining patient health behaviors in clinical settings.

Second, our work advances the eHealth literacy and patient empowerment literature by demonstrating that both cognitive capabilities (eHealth literacy) and motivational agency (psychological health empowerment) are necessary for effective use of online health information. These constructs are often studied separately, yet our findings suggest they operate together. Future research should integrate both when modeling patient information behaviors to develop a more complete understanding of how patients access, evaluate, and act on health information.

Practical Implications

Our findings offer practical insights into health education and clinical care. The identification of distinct eHL-E Profiles highlights the potential for tailored communication strategies for each patient group. Effective self-managers, with their high literacy and empowerment, can benefit from more autonomous decision-making opportunities and acknowledgment to continue engaging with reliable online health information. High-needs patients, in contrast, require interventions that enhance both their literacy skills and psychological empowerment; health care providers should therefore offer these patients greater support in navigating health information rather than solely expecting autonomous or shared decision-making. For dangerous self-managers, health programs should focus primarily on building literacy and information evaluation skills; providers should guide them in assessing online health information and encourage them to share and discuss their findings with physicians to prevent potential misuses.

In the digital age, patients commonly arrive at medical consultations with information gathered online [5]. Yet, our study, alongside works by others [88,89], highlights a concerning trend: patients often withhold this online-acquired information from their physicians. This reluctance to share online findings prevents opportunities to correct misconceptions from online searches and diminishes chances to better understand patients' views and perceptions about their health. To encourage patients' sharing, we recommend that physicians adopt PCC strategies. These include acknowledging patients' efforts to search for information online as preparation for their visits rather than devaluing this information [90,91], asking open-ended questions about what patients have learned online, and encouraging patients to discuss any concerns about information they have found [15]. By fostering these open dialogues, patients can benefit from both the wealth of online health information and the expertise of medical professionals.

The study findings also have implications for digital health initiatives on eHealth literacy or empowerment. With the proliferation of digital tools and artificial intelligence (AI) technologies in health promotion [92,93], training programs should address both skill development and confidence-building. Specifically, digital health platforms can implement strategies such as differentiated information presentation based on users' health literacy levels, gradually building eHealth literacy through scaffolded learning. Additionally, empowerment-focused features, such as personalized goal setting and progress tracking, can be integrated into online health applications to encourage patient participation.

Limitations

Our study has several limitations. First, there is no established measure for estimating patient information disclosure behavior. Survey studies on this topic typically devised their own measures based on their specific focus [49]. The measurement of actual behavior, including information seeking and disclosure, relied on participants' recall of previous medical visits with physicians. This could introduce recall bias to their responses.

Second, our sample was skewed toward young, highly educated individuals with college degrees or higher. This is also reflected in the overrepresentation of the effective self-manager group and underrepresentation of the high-needs patient group. The small size of the high-needs patient group may bias our results and limit their generalizability to those with low eHealth literacy and empowerment. The unbalanced sample may stem from our survey's online distribution, which is more accessible to digital platform users. Future research should apply more inclusive sampling for those less digitally skilled. Additionally, our sample has a significant gender imbalance, with females representing approximately twice as many participants as males. This overrepresentation of female participants limits the generalizability of our findings across gender groups. Future studies should use stratified sampling methods to ensure more balanced gender representation.

Third, there is ongoing debate in the literature regarding health literacy measurement, which can be subjective (self-rated) or objective (performance-based) [94]. Our study used the eHEALS, which relies on self-reported perceptions of competence rather than objective assessment of actual skills. While eHEALS is widely validated and predictive of health behaviors and outcomes, we acknowledge that self-reported measures may not fully capture individuals' actual eHealth literacy performance. Besides, there is growing recognition that traditional eHealth literacy measures like eHEALS may not fully address contemporary digital health challenges, and some

XSL•FO
RenderX

advocate for a broader digital health literacy concept that encompasses the competence of using digital health tools and additional measurements to incorporate AI literacy. Future studies should consider including these newer measurements to capture the full spectrum of digital health competencies.

Fourth, our data were collected between September and November 2022, prior to the public release of ChatGPT (November 30, 2022) and before AI chatbots became widely available for health information seeking. Therefore, AI-generated health information use was not prevalent among participants and was not included in our study design. This represents a limitation, as the current digital health information landscape has evolved rapidly since our data collection. Future research should examine how AI literacy and the use of AI-generated health information shape patient-physician communication and information behavior patterns.

Conclusions

This study extends existing literature by moving beyond online health information seeking to examine patients' disclosure of online findings to physicians. Unlike prior research that examined eHealth literacy and psychological health empowerment separately, this study integrated these constructs to identify meaningful patient profiles with distinct information behavior patterns. Findings indicate that active health information seeking and disclosure require both adequate eHealth literacy skills and psychological health empowerment. Furthermore, PCC facilitates disclosure regardless of patient profile. For practice, physicians should adopt a PCC that acknowledges patients' online research efforts and actively supports their eHealth literacy and empowerment. This can create a clinical environment where patients feel comfortable sharing their online findings, ultimately contributing to safer use of online health information and stronger patient-physician relationships.

Acknowledgments

During the preparation of this manuscript, ChatGPT-4 (OpenAI) was used to identify grammatical errors and suggest improvements for sentence readability. No original concepts, data analysis, or interpretations were generated by the AI; all intellectual content is solely that of the authors. The authors reviewed all suggested changes and take full responsibility for the published paper.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Data Availability

The dataset used in this study is available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: QL, PJS
Data curation: WJ, AC
Formal analysis: QL
Writing – original draft: QL
Writing – review & editing: QL, PJS, AC

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete measurement scales.
[DOCX File , 22 KB - jmir_v28i1e78836_app1.docx]

References

1. Wang X, Shi J, Kong H. Online health information seeking: a review and meta-analysis. Health Commun 2021;36(10):1163-1175. [doi: 10.1080/10410236.2020.1748829] [Medline: 32290679]
2. Swoboda CM, Van Hulle JM, McAlearney A, Huerta T. Odds of talking to healthcare providers as the initial source of healthcare information: updated cross-sectional results from the Health Information National Trends Survey (HINTS). BMC Fam Pract 2018;19(1):146 [FREE Full text] [doi: 10.1186/s12875-018-0805-7] [Medline: 30157770]
3. Wang J, Ashvetiya T, Quaye E, Parakh K, Martin S. Online health searches and their perceived effects on patients and patient-clinician relationships: a systematic review. Am J Med 2018;131(10):1250.e1-1250.e10. [doi: 10.1016/j.amjmed.2018.04.019] [Medline: 29730361]
4. Tan SSL, Goonawardene N. Internet health information seeking and the patient-physician relationship: a systematic review. J Med Internet Res 2017;19(1):e9 [FREE Full text] [doi: 10.2196/jmir.5729] [Medline: 28104579]

XSL•FO
RenderX

5. Luo A, Qin L, Yuan Y, Yang Z, Liu F, Huang P, et al. The effect of online health information seeking on physician-patient relationships: systematic review. J Med Internet Res 2022;24(2):e23354 [FREE Full text] [doi: 10.2196/23354] [Medline: 35142620]
6. Linn AJ, van Weert JCM, Gebeyehu BG, Sanders R, Diviani N, Smit EG, et al. Patients'online information-seeking behavior throughout treatment: the impact on medication beliefs and medication adherence. Health Commun 2019;34(12):1461-1468 [FREE Full text] [doi: 10.1080/10410236.2018.1500430] [Medline: 30052088]
7. Te Poel F, Baumgartner S, Hartmann T, Tanis M. The curious case of cyberchondria: a longitudinal study on the reciprocal relationship between health anxiety and online health information seeking. J Anxiety Disord 2016;43:32-40. [doi: 10.1016/j.janxdis.2016.07.009] [Medline: 27497667]
8. Walter N, Brooks J, Saucier C, Suresh S. Evaluating the impact of attempts to correct health misinformation on social media: a meta-analysis. Health Commun 2021;36(13):1776-1784. [doi: 10.1080/10410236.2020.1794553] [Medline: 32762260]
9. Lu J, Xiao Y. Heuristic information processing as a mediating factor in the process of exposure to COVID-19 vaccine information and misinformation sharing on social media. Health Commun 2024;39(12):2779-2792. [doi: 10.1080/10410236.2023.2288373] [Medline: 38016931]
10. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021;23(1):e17187 [FREE Full text] [doi: 10.2196/17187] [Medline: 33470931]
11. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. Annu Rev Public Health 2020;41:433-451 [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094127] [Medline: 31874069]
12. Chen YY, Li C, Liang J, Tsai C. Health information obtained from the internet and changes in medical decision making: questionnaire development and cross-sectional survey. J Med Internet Res 2018;20(2):e47 [FREE Full text] [doi: 10.2196/jmir.9370] [Medline: 29434017]
13. Bowes P, Stevenson F, Ahluwalia S, Murray E. 'I need her to be a doctor': patients' experiences of presenting health information from the internet in GP consultations. Br J Gen Pract 2012;62(604):e732-e738 [FREE Full text] [doi: 10.3399/bjgp12X658250] [Medline: 23211176]
14. Lu Q, Schulz P. Physician perspectives on internet-informed patients: systematic review. J Med Internet Res 2024;26:e47620 [FREE Full text] [doi: 10.2196/47620] [Medline: 38842920]
15. Shachar L. "You become a slightly better doctor": doctors adopting integrated medical expertise through interactions with e-patients. Soc Sci Med 2022;305:115038. [doi: 10.1016/j.socscimed.2022.115038] [Medline: 35598443]
16. Schick TS, Höllerl L, Biedermann T, Zink A, Ziehfreund S. Impact of digital media on the patient journey and patient-physician relationship among dermatologists and adult patients with skin diseases: qualitative interview study. J Med Internet Res 2023;25:e44129 [FREE Full text] [doi: 10.2196/44129] [Medline: 37738078]
17. Győrffy Z, Radó N, Mesko B. Digitally engaged physicians about the digital health transition. PLoS One 2020;15(9):e0238658 [FREE Full text] [doi: 10.1371/journal.pone.0238658] [Medline: 32986733]
18. Shen MJ, Dyson R, D'Agostino TA, Ostroff J, Dickler M, Heerdt A, et al. Cancer-related internet information communication between oncologists and patients with breast cancer: a qualitative study. Psychooncology 2015;24(11):1439-1447 [FREE Full text] [doi: 10.1002/pon.3752] [Medline: 25631285]
19. Kim J, Kim S. Physicians' perception of the effects of Internet health information on the doctor-patient relationship. Inform Health Soc Care 2009;34(3):136-148. [doi: 10.1080/17538150903102422] [Medline: 19670004]
20. George GC, Iwuanyanwu E, Buford A, Piha-Paul S, Subbiah V, Fu S, et al. Cancer-related internet use and its association with patient decision making and trust in physicians among patients in an early drug development clinic: a questionnaire-based cross-sectional observational study. J Med Internet Res 2019;21(3):e10348 [FREE Full text] [doi: 10.2196/10348] [Medline: 30869638]
21. Linn AJ, van Weert JCM, Bylund C. Are they important? Patients' communication barriers to discussing online health information during consultations. Health Commun 2024;39(5):945-950 [FREE Full text] [doi: 10.1080/10410236.2023.2197311] [Medline: 37072688]
22. Chiu YC. Probing, impelling, but not offending doctors: the role of the internet as an information source for patients' interactions with doctors. Qual Health Res 2011;21(12):1658-1666. [doi: 10.1177/1049732311417455] [Medline: 21799204]
23. Jia X, Pang Y, Liu L. Online health information seeking behavior: a systematic review. Healthcare (Basel) 2021;9(12):1740 [FREE Full text] [doi: 10.3390/healthcare9121740] [Medline: 34946466]
24. Schulz PJ, Nakamoto K. Health literacy and patient empowerment in health communication: the importance of separating conjoined twins. Patient Educ Couns 2013;90(1):4-11. [doi: 10.1016/j.pec.2012.09.006] [Medline: 23063359]
25. Street R. Active patients as powerful communicators. In: Robinson P, Giles W, editors. The New Handbook of Language and Social Psychology. New Jersey: Wiley; 2001:541-560.
26. Schulz PJ. The concept of health literacy. In: Zeyer A, Kyburz-Graber R, editors. Science | Environment | Health: Towards a Renewed Pedagogy for Science Education. Dordrecht: Springer; 2012:69-84.
27. Menon ST. Toward a model of psychological health empowerment: implications for health care in multicultural communities. Nurse Educ Today 2002;22(1):28-39; discussion 40. [doi: 10.1054/nedt.2001.0721] [Medline: 11886225]

XSL•FO
RenderX

28. Schulz PJ, Nakamoto K. The perils of misinformation: when health literacy goes awry. Nat Rev Nephrol 2022;18(3):135-136 [FREE Full text] [doi: 10.1038/s41581-021-00534-z] [Medline: 34997231]
29. Diviani N, van den Putte B, Giani S, van Weert JC. Low health literacy and evaluation of online health information: a systematic review of the literature. J Med Internet Res 2015;17(5):e112 [FREE Full text] [doi: 10.2196/jmir.4018] [Medline: 25953147]
30. Schulz PJ, Nakamoto K. Worse Than Ignorance: The Challenge of Health Misinformation. Cambridge: Cambridge University Press; 2024.
31. Náfrádi L, Nakamoto K, Csabai M, Papp-Zipernovszky O, Schulz PJ. An empirical test of the health empowerment model: does patient empowerment moderate the effect of health literacy on health status? Patient Educ Couns 2018;101(3):511-517. [doi: 10.1016/j.pec.2017.09.004] [Medline: 28899712]
32. Schulz PJ, Crosignani F, Petrocchi S. Critical test of the beneficial consequences of lifting the ban on direct-to-consumer advertising for prescription drugs in Italy: experimental exposure and questionnaire study. J Med Internet Res 2023;25:e40616 [FREE Full text] [doi: 10.2196/40616] [Medline: 37459159]
33. Sak G, Rothenfluh F, Schulz P. Assessing the predictive power of psychological empowerment and health literacy for older patients' participation in health care: a cross-sectional population-based study. BMC Geriatr 2017;17(1):59 [FREE Full text] [doi: 10.1186/s12877-017-0448-x] [Medline: 28219334]
34. Camerini L, Schulz P, Nakamoto K. Differential effects of health knowledge and health empowerment over patients' self-management and health outcomes: a cross-sectional evaluation. Patient Educ Couns 2012;89(2):337-344. [doi: 10.1016/j.pec.2012.08.005] [Medline: 22959333]
35. Wang RH, Hsu H, Lee Y, Shin S, Lin KD, An L. Patient empowerment interacts with health literacy to associate with subsequent self-management behaviors in patients with type 2 diabetes: a prospective study in Taiwan. Patient Educ Couns 2016;99(10):1626-1631. [doi: 10.1016/j.pec.2016.04.001] [Medline: 27083406]
36. Bizzotto N, de Bruijn GJ, Schulz P. Clusters of patient empowerment and mental health literacy differentiate professional help-seeking attitudes in online mental health communities users. Health Expect 2025;28(1):e70153 [FREE Full text] [doi: 10.1111/hex.70153] [Medline: 39815681]
37. Jiao W, Chang A, Ho M, Lu Q, Liu M, Schulz P. Predicting and empowering health for Generation Z by comparing health information seeking and digital health literacy: cross-sectional questionnaire study. J Med Internet Res 2023;25:e47595 [FREE Full text] [doi: 10.2196/47595] [Medline: 37902832]
38. Norman C. eHealth Literacy 2.0: problems and opportunities with an evolving concept. J Med Internet Res 2011;13(4):e125 [FREE Full text] [doi: 10.2196/jmir.2035] [Medline: 22193243]
39. Norman CD, Skinner H. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
40. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. J Med Internet Res 2012;14(1):e19 [FREE Full text] [doi: 10.2196/jmir.1619] [Medline: 22357448]
41. Moretti V, Brunelli L, Conte A, Valdi G, Guelfi M, Masoni M, et al. A web tool to help counter the spread of misinformation and fake news: pre-post study among medical students to increase digital health literacy. JMIR Med Educ 2023;9:e38377 [FREE Full text] [doi: 10.2196/38377] [Medline: 36996010]
42. Hu X, Bell R, Kravitz R, Orrange S. The prepared patient: information seeking of online support group members before their medical appointments. J Health Commun 2012;17(8):960-978. [doi: 10.1080/10810730.2011.650828] [Medline: 22574697]
43. Tennant B, Stellefson M, Dodd V, Chaney B, Chaney D, Paige S, et al. eHealth literacy and Web 2.0 health information seeking behaviors among baby boomers and older adults. J Med Internet Res 2015;17(3):e70 [FREE Full text] [doi: 10.2196/jmir.3992] [Medline: 25783036]
44. Lee HY, Jin S, Henning-Smith C, Lee J, Lee J. Role of health literacy in health-related information-seeking behavior online: cross-sectional study. J Med Internet Res 2021;23(1):e14088 [FREE Full text] [doi: 10.2196/14088] [Medline: 33502332]
45. Melhem SJ, Nabhani-Gebara S, Kayyali R. Digital trends, digital literacy, and e-health engagement predictors of breast and colorectal cancer survivors: a population-based cross-sectional survey. Int J Environ Res Public Health 2023;20(2):1472. [doi: 10.3390/ijerph20021472]
46. Sundell E, Wångdahl J, Grauman Å. Health literacy and digital health information-seeking behavior: a cross-sectional study among highly educated Swedes. BMC Public Health 2022;22(1):2278 [FREE Full text] [doi: 10.1186/s12889-022-14751-z] [Medline: 36471284]
47. Vamos S, Vine M, Gordon I, Alaimo D. The role of health literacy in patient-physician communication: a scoping review. Health Behav and Policy Rev 2024;11(4):1624-1654. [doi: 10.14485/hbpr.11.4.2]
48. Damian AJ, Gallo JJ. Promoting health literacy during the COVID-19 pandemic: a call to action for healthcare professionals. HKS Misinfo Review 2020;1(3). [doi: 10.37016/mr-2020-027]
49. Li S, Wang K. Sharing online health information with physicians: understanding the associations among patient characteristics, directness of sharing, and physician-patient relationship. Front Psychol 2022;13:839723 [FREE Full text] [doi: 10.3389/fpsyg.2022.839723] [Medline: 35432074]

XSL•FO
RenderX

50. Crondahl K, Karlsson LE. The nexus between health literacy and empowerment: a scoping review. Sage Open 2016;6(2). [doi: 10.1177/2158244016646410]
51. Salmon P, Hall G. Patient empowerment and control: a psychological discourse in the service of medicine. Soc Sci Med 2003;57:1969-1980. [doi: 10.1016/s0277-9536(03)00063-7]
52. Broom A. Virtually He@lthy: the impact of internet use on disease experience and the doctor-patient relationship. Qual Health Res 2005;15(3):325-345. [doi: 10.1177/1049732304272916] [Medline: 15761103]
53. Kivits J. Informed patients and the internet: a mediated context for consultations with health professionals. J Health Psychol 2006;11(2):269-282. [doi: 10.1177/1359105306061186] [Medline: 16464924]
54. Sillence E, Briggs P, Harris PR, Fishwick L. How do patients evaluate and make use of online health information? Soc Sci Med 2007;64(9):1853-1862. [doi: 10.1016/j.socscimed.2007.01.012] [Medline: 17328998]
55. Jiang F, Liu Y, Hu J, Chen X. Understanding health empowerment from the perspective of information processing: questionnaire study. J Med Internet Res 2022;24(1):e27178 [FREE Full text] [doi: 10.2196/27178] [Medline: 35014957]
56. Stovell D, Morrison A, Panayiotou M, Hutton P. Shared treatment decision-making and empowerment-related outcomes in psychosis: systematic review and meta-analysis. Br J Psychiatry 2016;209(1):23-28. [doi: 10.1192/bjp.bp.114.158931] [Medline: 27198483]
57. Hickmann E, Richter P, Schlieter H. All together now – patient engagement, patient empowerment, and associated terms in personal healthcare. BMC Health Serv Res 2022;22:1116 [FREE Full text] [doi: 10.1186/s12913-022-08501-5] [Medline: 36056354]
58. Thiruvengadam SK, Ashvetiya T, Stone N, Blumenthal R, Martin S. Shared decision-making and patient empowerment in preventive cardiology. Curr Cardiol Rep 2016;18(5):49. [doi: 10.1007/s11886-016-0729-6] [Medline: 27098670]
59. Aujoulat I, Marcolongo R, Bonadiman L, Deccache A. Reconsidering patient empowerment in chronic illness: a critique of models of self-efficacy and bodily control. Soc Sci Med 2008;66(5):1228-1239. [doi: 10.1016/j.socscimed.2007.11.034] [Medline: 18155338]
60. Papp-Zipernovszky O, Horváth MD, Schulz PJ, Csabai M. Generation gaps in digital health literacy and their impact on health information seeking behavior and health empowerment in Hungary. Front Public Health 2021;9:635943 [FREE Full text] [doi: 10.3389/fpubh.2021.635943] [Medline: 34055714]
61. Schrank B, Sibitz I, Unger A, Amering M. How patients with schizophrenia use the internet: qualitative study. J Med Internet Res 2010;12(5):e70 [FREE Full text] [doi: 10.2196/jmir.1550] [Medline: 21169176]
62. Epstein RM, Street RL. Patient-Centered Communication in Cancer Care: Promoting Healing and Reducing Suffering. NIH Publication No. 07-6225. Bethesda, MD: National Cancer Institute; 2007.
63. Wanzer MB, Booth-Butterfield M, Gruber K. Perceptions of health care providers' communication: relationships between patient-centered communication and satisfaction. Health Commun 2004;16(3):363-383. [doi: 10.1207/S15327027HC1603_6] [Medline: 15265756]
64. Siebinga VY, Driever E, Stiggelbout A, Brand P. Shared decision making, patient-centered communication and patient satisfaction – a cross-sectional analysis. Patient Educ Couns 2022;105(7):2145-2150. [doi: 10.1016/j.pec.2022.03.012] [Medline: 35337712]
65. Brand PLP, Stiggelbout AM. Effective follow-up consultations: the importance of patient-centered communication and shared decision making. Paediatr Respir Rev 2013;14(4):224-228. [doi: 10.1016/j.prrv.2013.01.002] [Medline: 23434177]
66. Silver MP. Patient perspectives on online health information and communication with doctors: a qualitative study of patients 50 years old and over. J Med Internet Res 2015;17(1):e19 [FREE Full text] [doi: 10.2196/jmir.3588] [Medline: 25586865]
67. Eysenbach G. Improving the quality of Web surveys: the Checklist for Reporting Results of Internet e-Surveys (CHERRIES). J Med Internet Res 2004;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
68. China's internet users total 1.079 bln with vibrant digital infrastructure, internet app growth. The State Council, The People's Republic of China. 2023 Aug 28. URL: http://english.www.gov.cn/archive/statistics/202308/28/content_WS64ec5126c6d0868f4e8dee21.html [accessed 2026-01-27]
69. Norman CD, Skinner H. eHEALS: the eHealth literacy scale. J Med Internet Res 2006;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
70. Chung SY, Park B, Nahm E. The Korean eHealth Literacy Scale (K-eHEALS): reliability and validity testing in younger adults recruited online. J Med Internet Res 2018;20(4):e138 [FREE Full text] [doi: 10.2196/jmir.8759] [Medline: 29678800]
71. Brørs G, Wentzel-Larsen T, Dalen H, Hansen T, Norman C, Wahl A, CONCARD Investigators. Psychometric properties of the Norwegian version of the Electronic Health Literacy Scale (eHEALS) among patients after percutaneous coronary intervention: cross-sectional validation study. J Med Internet Res 2020;22(7):e17312 [FREE Full text] [doi: 10.2196/17312] [Medline: 32720900]
72. Chang A, Schulz PJ. The measurements and an elaborated understanding of Chinese eHealth Literacy (C-eHEALS) in chronic patients in China. Int J Environ Res Public Health 2018;15(7):1553 [FREE Full text] [doi: 10.3390/ijerph15071553] [Medline: 30041420]
73. Lwin MO, Panchapakesan C, Sheldenkar A, Calvert G, Lim L, Lu J. Determinants of eHealth literacy among adults in China. J Health Commun 2020;25(5):385-393. [doi: 10.1080/10810730.2020.1776422] [Medline: 32552607]

XSL•FO
RenderX

74. Cui GH, Li S, Yin Y, Chen L, Li J, Liang F, et al. The relationship among social capital, eHealth literacy and health behaviours in Chinese elderly people: a cross-sectional study. BMC Public Health 2021;21(1):45 [FREE Full text] [doi: 10.1186/s12889-020-10037-4] [Medline: 33407275]
75. Oh HJ, Lee B. The effect of computer-mediated social support in online communities on patient empowerment and doctor-patient communication. Health Commun 2012;27(1):30-41. [doi: 10.1080/10410236.2011.567449] [Medline: 21797714]
76. Jiang S, Street RL. The effects of patient-centered communication, social capital, and internet use on patient empowerment: a cross-sectional study in China. Glob Health Promot 2019;26(4):33-43. [doi: 10.1177/1757975917749197] [Medline: 29521571]
77. Jiang S, Liu PL, Chia CWJ. Can online patient-provider communication improve emotional well-being? Examining the roles of social presence and patient empowerment. Cyberpsychol Behav Soc Netw 2023;26(5):366-370. [doi: 10.1089/cyber.2022.0289] [Medline: 36800227]
78. Jiang S, Ng AYK, Ngien A. The effects of social media information discussion, perceived information overload and patient empowerment in influencing HPV Knowledge. J Health Commun 2022;27(6):407-415. [doi: 10.1080/10810730.2022.2115591] [Medline: 35996841]
79. Epstein RM, Street RL. The values and value of patient-centered care. Ann Fam Med 2011;9(2):100-103 [FREE Full text] [doi: 10.1370/afm.1239] [Medline: 21403134]
80. Trivedi N, Moser RP, Breslau ES, Chou WYS. Predictors of patient-centered communication among U.S. adults: analysis of the 2017-2018 Health Information National Trends Survey (HINTS). J Health Commun 2021;26(1):57-64. [doi: 10.1080/10810730.2021.1878400] [Medline: 33648425]
81. McKeown L, Hong YA, Kreps GL, Xue H. Trends and differences in perceptions of patient-centered communication among adults in the US. Patient Educ Couns 2023;106:128-134. [doi: 10.1016/j.pec.2022.10.010] [Medline: 36270858]
82. Liu PL, Jiang S. Patient-centered communication mediates the relationship between health information acquisition and patient trust in physicians: a five-year comparison in China. Health Commun 2021;36(2):207-216. [doi: 10.1080/10410236.2019.1673948] [Medline: 31617412]
83. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: an R package for determining the relevant number of clusters in a data set. J Stat Soft 2014;61(6):1-36. [doi: 10.18637/jss.v061.i06]
84. Cheng C, Beauchamp A, Elsworth GR, Osborne RH. Applying the electronic health literacy lens: systematic review of electronic health interventions targeted at socially disadvantaged groups. J Med Internet Res 2020;22(8):e18476 [FREE Full text] [doi: 10.2196/18476] [Medline: 32788144]
85. Samoocha D, Bruinvels DJ, Elbers NA, Anema JR, van der Beek AJ. Effectiveness of web-based interventions on patient empowerment: a systematic review and meta-analysis. J Med Internet Res 2010;12(2):e23 [FREE Full text] [doi: 10.2196/jmir.1286] [Medline: 20581001]
86. Çiftci N, Yıldız M, Yıldırım Ö. The effect of health literacy and health empowerment on quality of life in the elderly. Psychogeriatrics 2023;23(4):609-620. [doi: 10.1111/psyg.12969] [Medline: 37186342]
87. Stevenson FA, Kerr C, Murray E, Nazareth I. Information from the internet and the doctor-patient relationship: the patient perspective—a qualitative study. BMC Fam Pract 2007;8:47 [FREE Full text] [doi: 10.1186/1471-2296-8-47] [Medline: 17705836]
88. Wong DKK, Cheung MK. Online health information seeking and ehealth literacy among patients attending a primary care clinic in Hong Kong: a cross-sectional survey. J Med Internet Res 2019;21(3):e10831 [FREE Full text] [doi: 10.2196/10831] [Medline: 30916666]
89. Waring ME, McManus DD, Amante DJ, Darling CE, Kiefe CI. Online health information seeking by adults hospitalized for acute coronary syndromes: who looks for information, and who discusses it with healthcare providers? Patient Educ Couns 2018;101(11):1973-1981 [FREE Full text] [doi: 10.1016/j.pec.2018.06.016] [Medline: 30305253]
90. Fujioka Y, Stewart E. How do physicians discuss e-health with patients? The relationship of physicians' e-health beliefs to physician mediation styles. Health Commun 2013;28(4):317-328. [doi: 10.1080/10410236.2012.682971] [Medline: 22716050]
91. Walker L, Sillence E. Examining healthcare professionals' communication around decision-making with internet-informed patients. Health Commun 2024;39(6):1094-1101. [doi: 10.1080/10410236.2023.2204580] [Medline: 37150845]
92. Guo W, Chen Y. Investigating whether AI will replace human physicians and understanding the interplay of the source of consultation, health-related stigma, and explanations of diagnoses on patients' evaluations of medical consultations: randomized factorial experiment. J Med Internet Res 2025;27:e66760 [FREE Full text] [doi: 10.2196/66760] [Medline: 40053785]
93. Yun HS, Bickmore T. Online health information-seeking in the era of large language models: cross-sectional web-based survey study. J Med Internet Res 2025;27:e68560 [FREE Full text] [doi: 10.2196/68560] [Medline: 40163112]
94. Goodman MS, Griffey RT, Carpenter CR, Blanchard M, Kaphingst KA. Do subjective measures improve the ability to identify limited health literacy in a clinical setting? J Am Board Fam Med 2015;28(5):584-594 [FREE Full text] [doi: 10.3122/jabfm.2015.05.150037] [Medline: 26355130]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
CHERRIES: Checklist for Reporting Results of Internet e-Surveys
eHEALS: eHealth Literacy Scale
eHL-E Profiles: eHealth Literacy and Empowerment Profiles
HEM: Health Empowerment Model
HINTS-China: Health Information National Trends Survey in China
PCC: patient-centered communication
PHES: Psychological Health Empowerment Scale
RQ: research question

Edited by S Brini; submitted 12.Jul.2025; peer-reviewed by L Shang, T Vagg; comments to author 09.Sep.2025; revised version received 09.Dec.2025; accepted 10.Dec.2025; published 18.Feb.2026.

Please cite as:
Lu Q, Jiao W, Chang A, Schulz PJ
Seeking But Not Discussing Online Health Information With Physicians: Cross-Sectional Survey Study of eHealth Literacy–Empowerment Profiles and Patient-Centered Communication
J Med Internet Res 2026;28:e78836
URL: https://www.jmir.org/2026/1/e78836
doi:10.2196/78836
PMID:

©Qianfeng Lu, Wen Jiao, Angela Chang, Peter Johannes Schulz. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impact of GPT-4–Generated Discharge Letters on Patients' Medical Comprehension: Prospective Crossover Study

Friederike Holderried[1*], MD; Alessandra Sonanini[1*], MD; Christian Stegemann–Philipps[1], MSc; Anne Herrmann–Werner[1], MD; Philipp Spitzer[2], MD; Martina Guthoff[3], MD; Nils Heyne[4], MD; Konstantin Sering[1], PhD; Martin Holderried[5], MD; Felix Eisinger[4], MD

[1]Tübingen Institute for Medical Education (TIME), University of Tübingen, Tübingen, Germany
[2]Department of Psychiatry and Psychotherapy, University of Erlangen-Nürnberg, Erlangen, Germany
[3]Nephrological Center Villingen-Schwenningen, Villingen-Schwenningen, Germany
[4]Department of Diabetology, Endocrinology, Nephrology, University of Tübingen, Tübingen, Germany
[5]Karlsruhe Municipal Hospital, Karlsruhe, Germany
[*]these authors contributed equally

Corresponding Author:
Felix Eisinger, MD
Department of Diabetology, Endocrinology, Nephrology
University of Tübingen
Ottfried-Müller-Straße 10
Tübingen, 72076
Germany
Phone: 49 07071 29 82712
Email: Felix.Eisinger@med.uni-tuebingen.de

Abstract

Background: Patients often struggle to understand standard hospital discharge letters, increasing the risk of medication errors and misunderstandings. According to cognitive load theory (CLT), complex, information-dense texts can overload working memory and impair comprehension. Artificial intelligence tools that generate patient-centered versions could help reduce extraneous cognitive load and bridge this gap. However, evidence for their effectiveness remains limited.

Objective: This study aimed to evaluate whether GPT-4 (OpenAI)–generated patient-centered letters improve standardized patients' retention and understanding of safety-relevant medical information compared with standard hospital discharge letters, and to explore potential effects on cognitive load as described by CLT.

Methods: In this prospective, randomized, crossover study, 48 trained standardized patients received a conventional discharge letter for an assigned disease (out of 3) and its matching GPT-4–generated patient-centered letter. Participants read one version first, identified predefined safety-relevant "learning objectives," and then repeated the task with the alternate version. The primary outcome was the proportion of learning objectives fully, partially, or not reported. In a secondary analysis, results were stratified by content field (Medication, Organization, Prevention of Complications, Lifestyle/Disease Management) and Bloom taxonomy level ("Remember," "Understand").

Results: The letter type significantly influenced comprehension (odds ratio [OR] 1.74, 95% CI 1.45-2.08; $P<.001$). Patient letters, compared with discharge letters, led to higher rates of fully (490/1073, 45.7% vs 413/1073, 38.5%) or partially (322/1073, 30% vs 287/1073, 26.7%) stated learning objectives and fewer omissions (261/1073, 24.3% vs 373/1073, 34.8%). Participants performed better on "Remember" than on "Understand" learning objectives, regardless of letter type (OR 3.33, 95% CI 1.96-5.88; $P<.001$). Compared with standard hospital discharge letters, patient letters consistently improved results at both cognitive levels ("Remember": 278/545, 51% vs 242/545, 44.4%; "Understand": 212/528, 40.2% vs 171/528, 32.4% fully stated). The effect of patient letters varied by content field ($P<.001$). The greatest improvements were observed for "Medication" (170/254, 66.9% vs 129/254, 50.8% fully stated) and "Organization" (78/158, 49.4% vs 62/158, 39.2% fully stated). Improvements in the content field "Prevention of Complications" were modest, and those for "Lifestyle/Disease Management" were even smaller across all conditions. A total of 24.3% (261/1073) of key information remained unrecognized.

Conclusions: In this explanatory study, GPT-4–generated patient letters improved comprehension of safety-relevant discharge information among standardized patients, particularly regarding medication and organizational aspects. However, they were less

XSL•FO
RenderX

effective in supporting higher-order understanding, such as risk prevention or lifestyle management. These hypothesis-driven findings can be interpreted within a CLT framework and may motivate prospective evaluation of multimodal, iterative supports.

(J Med Internet Res 2026;28:e81243) doi:10.2196/81243

KEYWORDS

artificial intelligence; GPT-4; health care communication; large language models; patient education; patient letters; patient safety

Introduction

The transition from inpatient to outpatient care represents a critical juncture for information exchange and patient safety, often marked by communication errors and incomplete understanding [1]. Nearly half (49%) of patients experience at least 1 postdischarge medical error, most frequently concerning medication management, diagnostic workup, or test result follow-up [2]. Effective patient understanding during this phase is crucial, as misinterpretation or omission of medical instructions can directly impact adherence, follow-up care, and safety outcomes [3,4]. Studies estimate that up to 78% of patients misunderstand key elements of their hospital care and discharge instructions [5], and approximately half (50.8%) of patients receiving polypharmacy misinterpret at least 1 dosage instruction [6].

Patients' comprehension of discharge information is influenced by multiple factors, including health literacy, educational background, prior knowledge, and emotional stress during this vulnerable phase [7,8]. Considering that 36% of US adults have limited health literacy [9], a substantial proportion of patients may struggle to fully understand and follow postdischarge instructions, which are typically written for communication between health care professionals and often fail to meet patients' informational needs [10]. Therefore, there is a clear need for discharge documents that are patient-centered, clearly structured, and accessible; these will be referred to as patient letters throughout this manuscript.

Recent advances in artificial intelligence (AI), particularly large language models (LLMs) such as GPT-4 (OpenAI), offer a promising approach to address this need [11]. By transforming complex medical documents into plain-language patient letters, AI has the potential to improve patients' comprehension and recall of important medical information. Several feasibility studies have demonstrated that LLMs can generate readable and accurate patient letters from standard discharge documents, highlighting their technical capability to simplify language, structure information, and provide consistent formatting [12-14]. To date, only a small number of studies have investigated the impact of GPT-generated patient letters on patient understanding [15-17]. These studies have primarily relied on semistructured interviews or multiple-choice assessments, offering only quantitative evaluations of patient knowledge. A more nuanced understanding of patient comprehension is still lacking, particularly regarding which types of information are retained, how the cognitive complexity of content influences understanding, and whether comprehension differs across medical content domains. Accordingly, evidence on the educational impact and cognitive benefits of AI-generated patient communications remains limited.

This study is based on our previous work, in which we explored the feasibility of transforming discharge letters into patient letters using GPT-4 [12]. That study primarily focused on the accuracy of the generated letters and measures of patient-centeredness, demonstrating that overall, GPT-4 can produce readable and structured patient-directed documents, with some limitations including omissions. Building on these findings, this study shifts the focus to an application-oriented perspective. It investigates whether GPT-4–generated patient letters improve comprehension, using standardized patients (SPs) to simulate real-world understanding and follow-up tasks.

To systematically evaluate comprehension, we framed patient understanding as a learning process informed by cognitive load theory (CLT) [18]. According to CLT, learning efficiency depends on three types of cognitive load:

1. Intrinsic load refers to the inherent complexity of the information and is dependent on the learner's level of expertise. In the context of discharge letters, intrinsic load is generally high due to complex medical content and procedural details, and it can only be modified to a limited extent.
2. Extraneous load reflects mental resources allocated to elements that do not contribute to learning or schema acquisition. High extraneous load can arise from poorly structured information, excessive jargon, or frequent use of abbreviations [18]. In our prior study, we demonstrated that GPT-4–generated patient letters mitigated these contributing factors by providing increased repetition of key points, fewer abbreviations, and reduced use of medical terminology [12].
3. Germane load represents the mental resources devoted to acquiring and automating schemata in long-term memory. In our context, the goal is to maximize germane load so that patients allocate as much cognitive energy as possible to truly understand, retain, and integrate the medical information.

Guided by this theoretical framework, the study addressed the following research questions:

- Does the reduction of extraneous load through AI-generated patient letters translate into improved comprehension of safety-relevant information?
- Are comprehension gains consistent across different content domains (eg, Medication, Organization, Prevention of Complications, and Lifestyle)?
- Do comprehension improvements differ between lower-order ("Remember") and higher-order ("Understand") cognitive processes as defined by Bloom taxonomy [19]?

By combining a patient-centered perspective with a learning-theoretical framework, this study examines the

XSL•FO
RenderX

educational potential and limitations of AI-generated patient letters and considers how they might contribute to patient safety during vulnerable care transitions.

Methods

Study Design and Setting

In our previous study, we developed a methodology for generating GPT-4–based patient letters, demonstrating the feasibility of transforming discharge letters into patient-friendly formats while maintaining medical accuracy and patient-centeredness. The prompting strategy and human review process were described in detail in that publication [12].

Building directly on this work, this prospective, randomized, crossover study with SPs used 3 discharge letters and the corresponding GPT-4–generated patient letters from our prior study, each representing a common chronic condition: arterial hypertension, type 2 diabetes mellitus, and diabetic kidney disease. These diagnoses were selected because they are highly prevalent and frequently occur in both inpatient and outpatient care. The letters are provided in Multimedia Appendix 1 in German (original) and English (translated).

Both letter formats were presented to 48 SPs using a crossover design (Figure 1). We compared perception and retention of key medical information—defined as "learning objectives"—presented in the letters.

XSL•FO
RenderX

Figure 1. Overview of study design and rating procedure in a prospective, randomized crossover study comparing the retention of predefined learning objectives by standardized patients after reading either a standard hospital discharge letter or a GPT-4–generated patient letter. AI: artificial intelligence; LO: learning objectives.

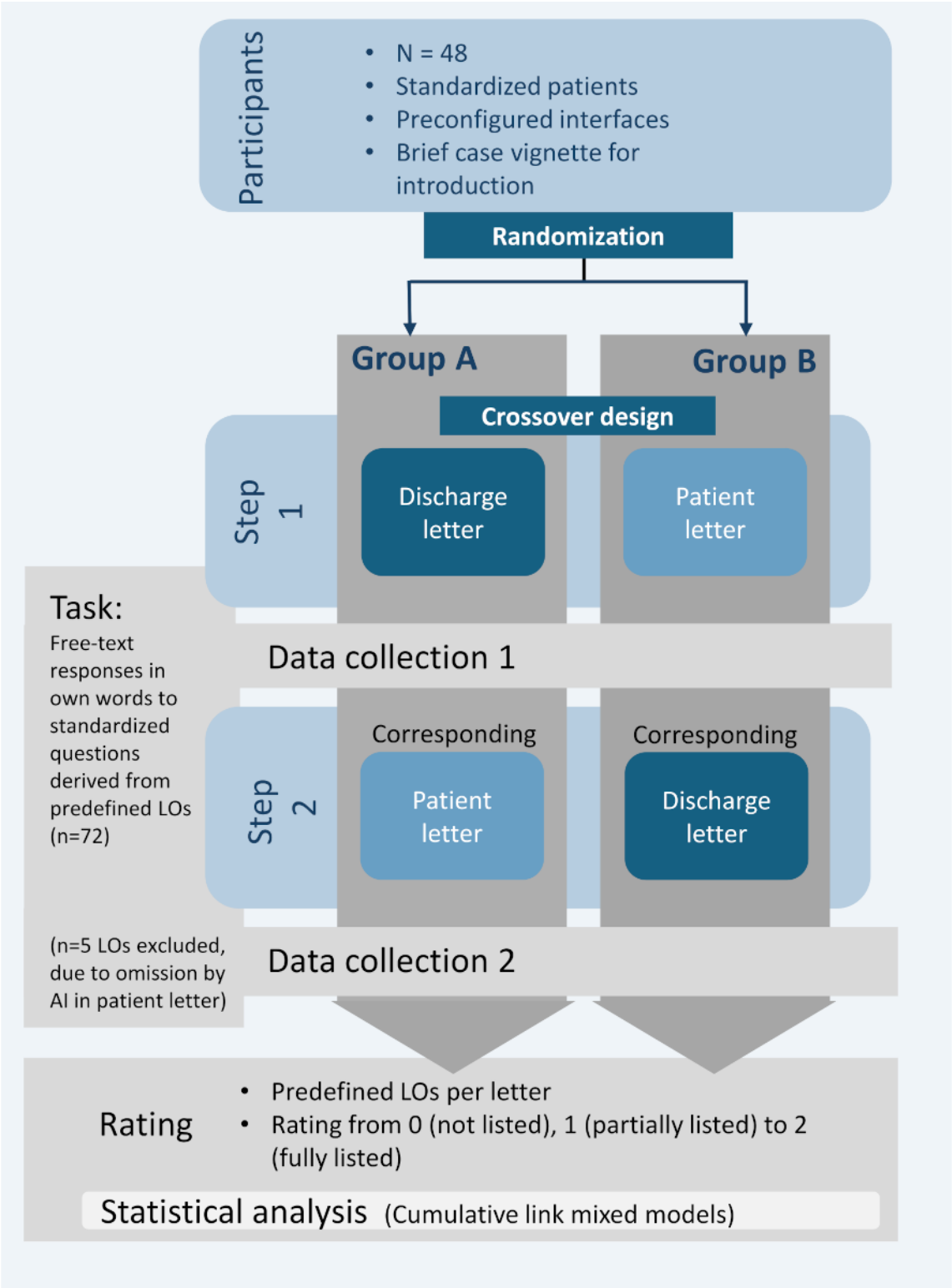

Participants and Study Procedures

A total of 48 SPs participated in our study (Multimedia Appendix 2). SPs are individuals who are specifically trained as patient actors to portray medical patients in a consistent and realistic manner, and they participated in the study in their specific role [20,21]. The sample size was determined by the number of participants available for recruitment and was not calculated prior to the study. The SPs were recruited from the SP programs of the Universities of Tübingen and Erlangen, Germany. Data were collected between May and September 2024. After reading a brief case vignette covering basic medical and socioeconomic information, each participant was randomized to the discharge letter (group A) or its patient-friendly version (group B). Participants were not

XSL•FO
RenderX

informed about the sequence in which they would receive the letters. However, the differences between the discharge letter and the GPT-4–generated patient letter were apparent, making full blinding impossible. Therefore, no formal assessment of blinding adequacy was performed.

At both sites, letters were presented digitally through a web-based interface. Participants were instructed to read each letter with care and were permitted to scroll back to the text while answering the open-ended questions. This setting mirrors the real-world access to written discharge information. The SPs extracted key information from 4 content fields and answered predefined open questions (Multimedia Appendix 3). This approach was chosen because continuous access limits short-term memory bias and mirrors real practice [22]. To foster active knowledge construction, participants were instructed to answer in their own words without copying. Letters were then crossed over, and the task was repeated with the alternate version.

Learning Objectives and Content Fields

For each of the 3 clinical cases (arterial hypertension, diabetes mellitus, and diabetic kidney disease), we defined 24 safety-related learning objectives that cover information considered essential by the treating physicians for patient self-management after discharge (eg, knowledge of medication changes, warning symptoms, and follow-up appointments). The learning objectives were designed and revised by an interdisciplinary committee of 4 physicians (specialized in internal medicine, nephrology, diabetology, general medicine, and medical education) based on current clinical guidelines [23-27] and their professional experience. They were formulated as short, observable statements (eg, “The patient can ... list” and “The patient understands that ...”) and checked for comprehensibility and redundancy.

For each case, the 24 learning objectives were evenly distributed across the levels of the revised Bloom taxonomy, with 12 learning objectives in the “Remember” category and 12 in the “Understand” category. In Bloom taxonomy, the “Remember” category refers to the ability to recall facts and basic concepts, while the “Understand” category goes beyond recall and involves interpreting, explaining, and making sense of the information [19,28]. Each learning objective was further mapped to 1 of 4 content fields: “Organization” (follow-up tasks), “Medication” (drug changes), “Prevention of Complications” (early detection and management of disease-related complications), and “Lifestyle/Disease Management” (self-care advice/lifestyle modifications; Multimedia Appendix 3). This structure ensured comparable cognitive demands and content coverage across all 3 cases and enabled stratified analyses by content area and Bloom category. The process of defining learning objectives and creating the original discharge letters has been described in detail in our previous publication [12].

Across the 3 letters used in this study, 72 learning objectives were assessed in total (24 per letter). A total of 5 of these 72 learning objectives were not represented in the GPT-4–generated patient letters because the model did not carry them over from the original discharge letters. These learning objectives were therefore excluded from the analysis for both letter types (Multimedia Appendix 4). The quality and completeness of the translation from discharge letter to patient letter were examined in detail in our previous publication and were therefore not reassessed within the scope of this study [12]. After this exclusion, 34 of the 67 remaining learning objectives were classified as “Remember” and 33 as “Understand” according to Bloom taxonomy (Multimedia Appendix 3).

Assessment Approach

Reading the discharge or patient letter and then answering the case questions was designed as a structured learning assessment sequence [29]: the letters provided the information to be learned, and the SPs’ responses reflected what they had actually understood. To capture constructed understanding rather than simple recognition, we used open-ended questions that required participants to formulate their answers in their own words rather than choosing from predefined options.

Open-ended response formats are less prone to guessing than multiple-choice questions and are better suited for assessing higher levels of processing for learning objectives at the “Understand” level. This approach closely aligns the assessment with the predefined learning objectives and the intended depth of understanding. This choice is supported by assessment literature [30].

Outcome Measures

The primary outcome was the understanding of safety-related information at the level of individual learning objectives. Each participant response (participant × learning objective × letter) was rated on an ordinal scale with 3 categories reflecting the degree of goal achievement in the open-ended response (0=not specified, 1=partially specified, 2=fully specified). The main comparison of interest was the difference between the patient letters generated by GPT-4 and the standard discharge letters with respect to this ordinal measure of SPs’ understanding.

Secondary end points were the distribution of this ordinal comprehension metric across the 4 predefined content areas (Organization, Medication, Prevention of Complications, and Lifestyle/Disease Management) and the distribution across the levels of Bloom taxonomy (“Remember” vs “Understand”). As an additional secondary end point, the frequency of medically incorrect answers was examined.

Randomization

Participants at the University Hospital Tübingen were allocated to the 3 clinical cases and the order of letter presentation (discharge letter first vs patient letter first) according to a pregenerated randomization list. This list was created prior to participant recruitment using a randomly generated allocation sequence that specified both case assignment and presentation order, with approximately balanced group sizes. Participants were enrolled consecutively and assigned to the next available entry on the randomization list. Allocation was not known in advance to either participants or study personnel conducting the sessions, thereby preserving allocation concealment at the participant level.

Participants from the University of Erlangen, who took part remotely via a web-based study platform, were randomized

XSL•FO
RenderX

using a computer-generated 1:1 allocation algorithm implemented within the platform. Upon enrollment, the platform automatically assigned participants to the letter-order sequence and the clinical case.

The use of 2 different randomization methods resulted from the different study sites: on-site participation in Tübingen required an offline procedure, whereas the Erlangen cohort participated exclusively online. Randomization was conducted separately within each site and not across sites. Although randomization procedures differed between sites, both approaches ensured allocation concealment at the individual level. Importantly, the substantive data collection procedure (including exposure to the letters and assessment of learning objectives) was identical at both locations.

Data Analysis

Rating Process

The answers of the SPs were independently rated by 2 experienced clinicians (FH and AS) to determine whether the predefined learning objectives were correctly and completely identified and reported. The rating process followed a predefined, ordinally scaled rating structure, ranging from "not listed" over "partially listed" to "fully listed," and disagreements were solved by discussion. A specific definition was established in advance for each learning objective (Multimedia Appendix 3). Raters were blinded to both the letter type (discharge letter or patient letter) and the order in which it was presented. No formal assessment of blinding adequacy was performed. The raters achieved a Cohen κ of 0.97. Due to the high level of agreement, one of the two ratings was selected at random for use in the statistical analysis.

Statistical Analysis

Due to the ordinal nature of the ratings of the learning objectives (comprehensively reported, partially reported, and not reported), cumulative link mixed models were fitted to the data. Random intercepts for participant and learning objective were used to account for participant and learning objective random effects (Multimedia Appendix 5). The fixed effects of letter type, Bloom category, and content field were investigated as variables of interest with model comparisons. The fixed effects disease type and time point of measurement were included as controls. Relevant models were assessed, and alternatives were considered. *P* values for differences in individual content fields were calculated by subsetting the data to each content field. Models were compared according to the Akaike Information Criterion and likelihood ratio tests. No missing data were present. The analyses were conducted using R software (version 4.4; R Foundation for Statistical Computing) and the ordinal package version 2023.12-4.1 [31,32].

Reporting Standards

The reporting of this study followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guideline for cohort studies [33]. A reporting checklist can be found in the Multimedia Appendix 6.

Ethical Considerations

This study was approved by the local ethics committee (778/2023BO2) of Tübingen University Hospital, Germany. Written informed consent was obtained from the SPs before inclusion in the study.

No direct identifiers (eg, names or dates of birth) were collected in the study. A short demographic questionnaire (age, gender, educational background, and prior experience as an SP) was completed by the SPs; these results were stored separately and not linked to individual response data. During data collection, response data were temporarily saved on secured cloud infrastructure; evaluation data and the separate demographic file were stored in different locations with a role-based access control. Reidentification was not possible at any time. Due to data loss at one site, demographics were available for 38 out of 48 SPs. As demographics were not used in the analysis of learning objective performance, and there is no evidence of systematic bias between sites, this missing information does not affect the validity of the study outcomes. The study was conducted in accordance with the Declaration of Helsinki.

Results

Demographic Data

The median age of the participants was 46.7 (IQR 28.3-65.0) years, with a female-to-male ratio of approximately 1.92:1. The majority (24/38, 63.2%) had a university degree. Task order did not influence performance (*P*=.80), although participants performed slightly better on their second attempt (odds ratio [OR] 1.23, 95% CI 1.03-1.47; *P*=.02). The type of disease did not have a significant main effect on the retention rate of learning objectives (*P*=.96; Multimedia Appendix 5). Therefore, all patient data were pooled for the subsequent comparison of the discharge letter and the patient letter.

Patient Letters vs Discharge Letters

Overall, patient letters improved information uptake: SPs retained more learning objectives after reading the patient letter compared with the discharge letter (OR 1.74, 95% CI 1.45-2.08; *P*<.001; Figure 2). Comprehensive reporting rose to 45.7% (490/1073) with the patient letter compared to 38.5% (413/1073) with the discharge letter. Partial reporting also improved (322/1073, 30% vs 287/1073, 26.7%), while omissions decreased from 34.8% (373/1073) to 24.3% (261/1073; Figure 2; Multimedia Appendix 7). Remarkably, despite overall gains, those 24.3% (261/1073) of all ratings remained "not reported" after reading the patient letter. Because omissions varied by participant, we conducted an item-level analysis to identify objectives consistently missed (predefined as "not reported" in ≥50% of the cases). A total of 6 (9%) of 67 objectives met this criterion. By Bloom level, 5 out of 6 were "Understand" and 1/6 "Remember"; by content field, 3 out of 6 fell under "Prevention of Complications," 2 out of 6 under "Lifestyle/Disease Management," and 1 out of 6 under "Organization" (Multimedia Appendix 8). On the other hand, the overall proportion of medically incorrect reports by the participants was low, with no significant difference between the letter types (*P*=.55; Figure 3).

XSL•FO
RenderX

Figure 2. Percentage of predefined learning objectives that were comprehensively reported, partially reported, or not reported by standardized patients after reading either a standard hospital discharge letter (DL) or a GPT-4–generated patient letter (PL). Results are shown overall (left) and separately (right) for the Bloom taxonomy levels "Remember" and "Understand." Statistical comparison was performed using cumulative link mixed models. Overall, standardized patients retained significantly more learning objectives with the patient letter than with the discharge letter. Across both letter types, performance was significantly higher for "Remember" than for "Understand" learning objectives.

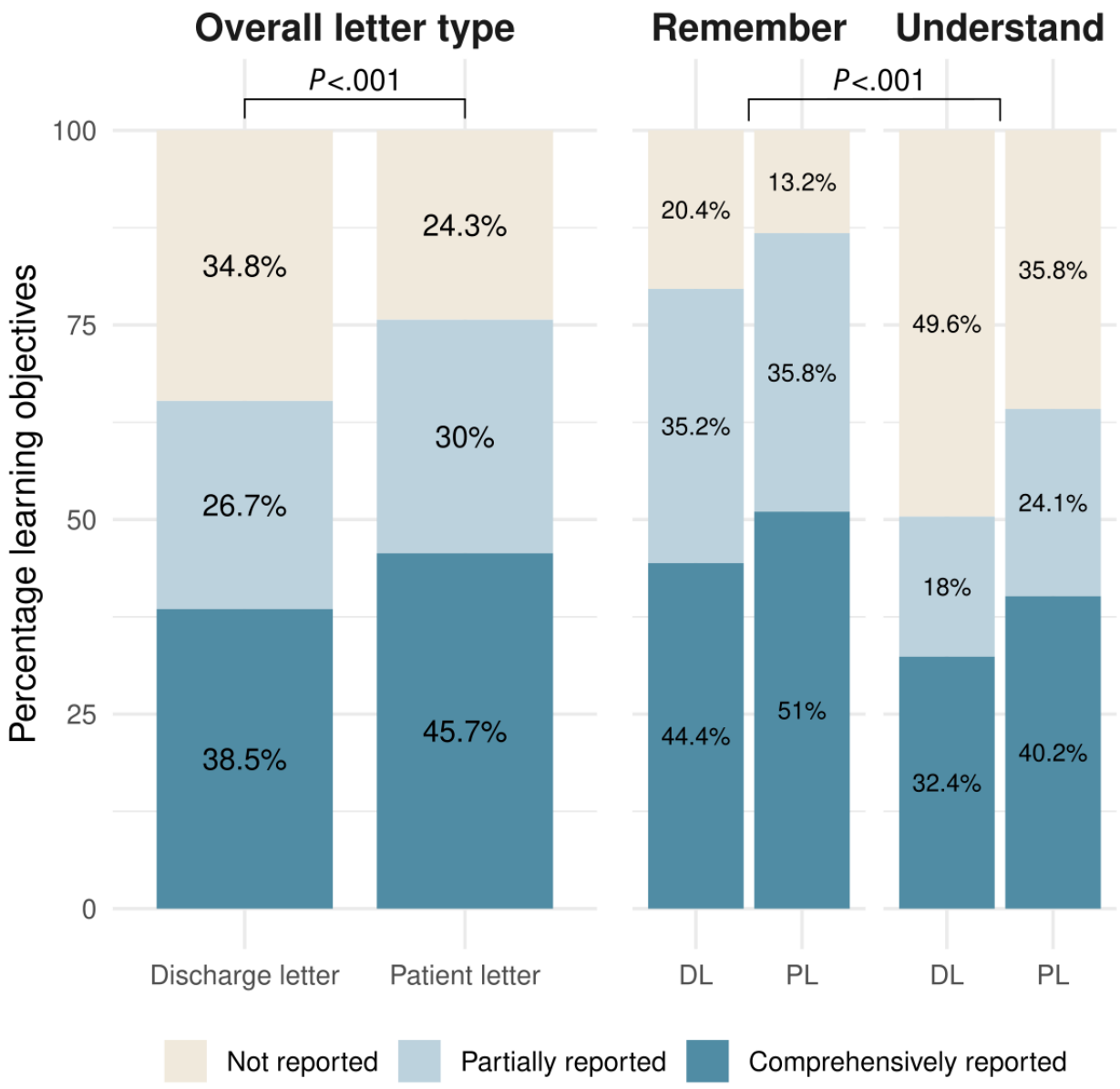

XSL•FO
RenderX

Figure 3. Medical correctness of standardized patient responses after reading either a standard hospital discharge letter or a GPT-4–generated patient letter. The figure illustrates the percentage of responses rated as medically accurate, medically inaccurate, or not applicable (statements without medical content). Statistical comparison was conducted using cumulative link mixed models.

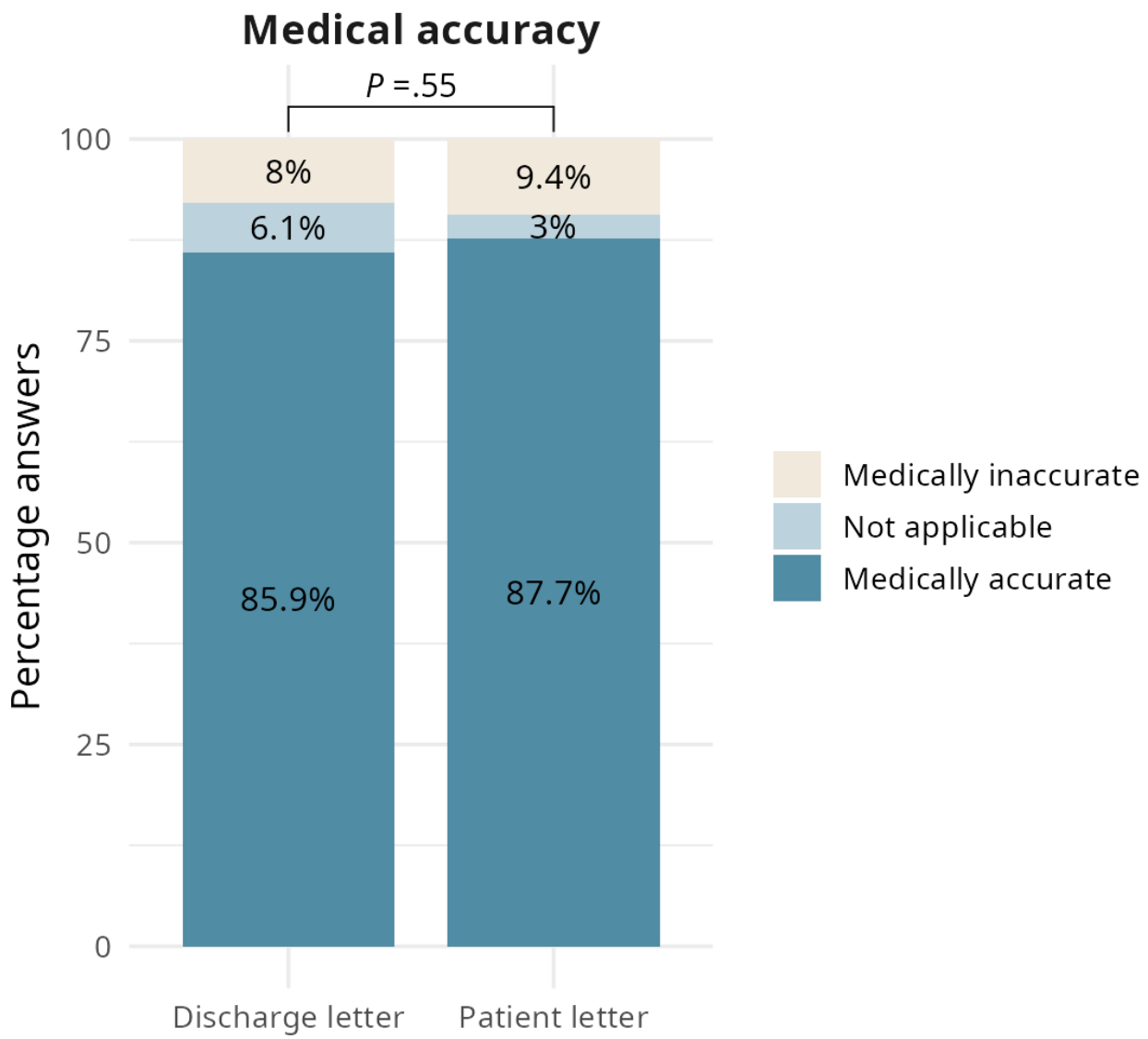

Impact of Cognitive Demand (Bloom Taxonomy) on Information Uptake

A subgroup analysis of learning objectives based on Bloom taxonomy revealed that, regardless of letter type, participants performed better on learning objectives classified under the "Remember" category compared to those classified under the "Understand" category (OR 3.33, 95% CI 1.96-5.88; *P*<.001; Figure 2). In both categories, the proportion of comprehensively reported learning objectives was higher when presented in the patient letter format compared to the discharge letter format (278/545, 51% vs 242/545, 44.4% for "Remember" and 212/528, 40.2% vs 171/528, 32.4% for "Understand"; Figure 2; Multimedia Appendix 7). A subsequent analysis showed no significant interaction between Bloom taxonomy level and letter type (*P*=.13), indicating that the observed benefits of the patient letter were consistent across both cognitive levels.

Impact of Content Field on Information Uptake

In a subgroup analysis based on content fields, we observed a statistically significant interaction effect between the type of letter and the content field (*P*<.001), indicating that the benefit of the patient letter depends on the content field. Additional values for effect size, OR, and CI for the results in this paragraph are illustrated in Multimedia Appendix 9. A substantial increase in comprehension associated with the patient letter was observed for the domain of "Medication" (*P*<.001), with a rate of accurately reported learning objectives of 66.9% (170/254) compared to 50.8% (129/254) for the discharge letter. In the GPT-4–generated patient letter, the changes and purpose of each drug were explicitly stated in plain language (eg, "You were also given two other medications, metformin and empagliflozin, which help control blood sugar"), whereas the standard discharge letter contained this information in a condensed table format without further explanation. For further details on the translation process from discharge to patient letters, we refer to our previous work [12].

XSL•FO
RenderX

The second largest improvement was seen in the domain of "Organization" (78/158, 49.4% vs 62/158, 39.2%; *P*<.001; Figure 4; Multimedia Appendix 7). In contrast, improvements observed with the patient letter in the content fields "Lifestyle/Disease Management" (*P*=.38) and "Prevention of Complications" (*P*<.01) were small and medium. In the content field "Prevention of Complications," the patient letter resulted in a minor increase in the proportion of partially addressed learning objectives compared to the discharge letter (102/311, 32.8% vs 70/311, 22.5%). However, the rate of completely reported learning objectives remained the lowest compared to the other content fields (Figure 4; Multimedia Appendix 7). Regarding the content field "Lifestyle/Disease Management," there was even a small decrease observed for the partially reported learning objectives (100/350, 28.6% in the patient letter vs 107/350, 30.6% in the discharge letter). However, the amount of correctly stated learning objectives still showed a small increase (149/350, 42.6% in the patient letter vs 139/350, 39.7% in the discharge letter; Figure 4; Multimedia Appendix 7).

Figure 4. Percentage of predefined learning objectives that were comprehensively reported, partially reported, or not reported by standardized patients after reading either a standard hospital discharge letter (DL) or a GPT-4–generated patient letter (PL). Results are displayed across 4 different content categories: Organization, Medication, Prevention of Complications, and Lifestyle/Disease Management. Differences between letter types were evaluated using cumulative link mixed models. Significant improvements were achieved with the patient letter in the content fields "Medication" and "Organization," whereas gains were modest for "Prevention of Complications" and minimal for "Lifestyle/Disease Management.".

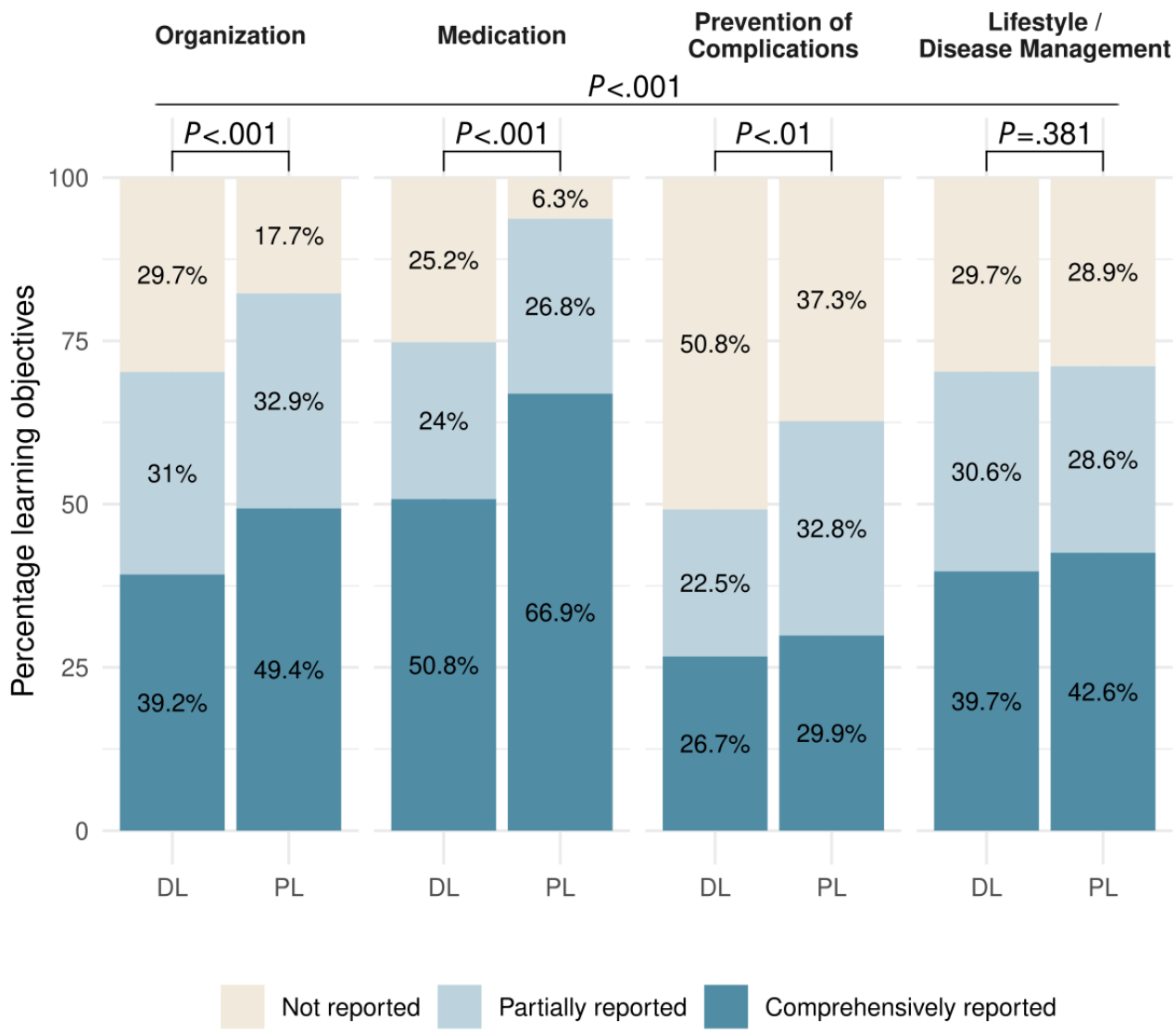

Discussion

Principal Findings

This prospective, explanatory study with SPs showed that AI-generated patient letters increased the retrieval of medication-related learning objectives to 66.9% (170/254) and organizational objectives to 49.4% (78/158), compared with 50.8% (129/254) and 39.2% (62/158), respectively, after reading the standard hospital discharge letters. These gains are larger than those reported in earlier single-arm or pre-post studies of human or AI-generated patient letters that lacked counterbalancing [13,34,35]. By scoring open, free-text answers, we also captured a broader slice of comprehension than recent

XSL•FO
RenderX

GPT-based studies that relied on recognition formats [15-17]. However, the reliance on predefined learning objectives, evaluated by clinicians, may still miss nuances of SP understanding. In particular, misconceptions and emotional interpretations may be underreported, and these remain important aspects for future qualitative work.

Content- and Bloom-Specific Insights

Although overall performance improved, SPs still failed to report, on average, nearly a quarter of the key information, highlighting a persistent patient safety gap for patients. Gains clustered in the areas of Medication and Organization, which are long-recognized weak spots: <50% of patients understand their discharge medicines [36], and medication errors account for roughly half of preventable harm [37]. By increasing the retention of medication-related information, we hypothesize that patient letters have the potential to improve medication management and reduce medication errors, thereby potentially contributing to enhanced patient safety. Similarly, improving awareness of necessary follow-up appointments may also support patient safety by reducing missed visits and associated complications. Previous research supports these mechanisms: written patient information material has been shown to increase medication adherence [38], missed follow-up appointments predict readmissions and emergency visits [39-41], and low health literacy is associated with higher readmission rates [42]. Moreover, communication interventions at hospital discharge are significantly linked to fewer readmissions and higher treatment adherence [4]. While these findings provide a rationale for the use of AI-generated patient letters, the effects on actual patient safety outcomes remain to be empirically evaluated in future research.

In contrast, improvements in the content field “Prevention of Complications” were modest, and those for “Lifestyle/Disease Management” were even smaller. Notably, 37.3% (116/311) of the key information regarding “Prevention of Complications” remained unknown. While higher cognitive demand likely contributed, other factors such as complex sentence structures, domain-specific phrasing, multistep reasoning, and the absence of visual aids may have further limited comprehension. As shown in our previous study, GPT-4 also faced greater challenges translating “Prevention of Complications” content compared with other domains [12]. Interestingly, although GPT-4 more successfully simplified “Lifestyle/Disease Management” information, SP performance remained low, suggesting that simplification alone does not guarantee comprehension when integration into practical self-management is required, and suggesting a distinction between content that is challenging for the AI to translate versus content that is difficult for SPs to understand and apply. Stratifying learning objectives by Bloom taxonomy showed higher performance for “Remember” than for “Understand” items (278/545, 51% vs 212/528, 40.2% correct with the patient letter), and 5 out of 6 objectives omitted by ≥50% of participants fell into the “Understand” category. This difference is unsurprising given that “Remember” tasks primarily involve locating explicit information, whereas “Understand” tasks require higher-order cognitive processes. Importantly, the patient letter improved performance in both categories compared with the discharge letter (“Remember”: 278/545, 51% vs 242/545, 44.4%; Understand: 212/528, 40.2% vs 171/528, 32.4%), and there was no significant interaction between letter type and Bloom level (P=.13), indicating that the observed benefit of the patient letter was approximately equal for both cognitive levels.

These findings suggest that GPT-4–generated patient letters can support not only the retrieval of factual information but may also modestly enhance comprehension-oriented learning objectives, though higher-level understanding remains more challenging. This is consistent with prior longitudinal research demonstrating that discharge instructions decay rapidly without reinforcement [43-45].

The clustering of consistently missed objectives in “Prevention of Complications” and, secondarily, “Lifestyle/Disease Management,” together with the predominance of “Understand”-level items, suggests that a static, text-only letter may not be sufficient in contexts where patients need to link symptoms to actions, reason about conditional risks (eg, sleep apnea as source of high blood pressure), or maintain self-management routines (eg, home blood pressure documentation or dietary change).

Didactic Implications (CLT)

CLT provides a helpful framework for interpreting our findings [18]. Recent work applying CLT to clinical practice shows that stress, emotion, and uncertainty increase intrinsic cognitive load and that working memory depletion can impair clinical performance [46]. Against this background, it seems plausible that linguistic simplification and better structure primarily reduce extraneous cognitive load. Our prior work demonstrated that GPT-4 can reorganize discharge letters into patient letters mainly by removing jargon and chunking information [12], which likely supports this reduction. While discharge letters are generally already well-structured, AI-generated patient letters further optimize content organization in a patient-centered manner.

We did not directly measure intrinsic, extraneous, or germane cognitive load in this study; CLT is therefore used as an interpretive framework rather than an empirically tested mechanism. Our learning objectives were deliberately designed so that each content field contained both “Remember” and “Understand” objectives. However, objectives that remained most difficult (defined as those omitted by ≥50% of participants) were predominantly from the “Understand” category in the area of “Prevention of Complications” and, to a lesser extent, in the area of “Lifestyle/Disease Management.” These learning objectives (eg, hypoglycemia symptom-action mapping, recognition of statin-associated muscle symptoms, dehydration-related kidney risk, or the contribution of home blood pressure monitoring and Mediterranean diet to long-term disease control) require the integration of multiple information elements, conditional if-then thinking, and sustained self-regulation. Within the framework of CLT, they can therefore be interpreted as tasks with a comparatively high intrinsic cognitive load.

In contrast, the largest gains from patient letters were observed predominantly in factual learning objectives in the areas of “Medication” or “Organization,” which required participants

XSL•FO
RenderX

to find and reproduce explicitly stated information (eg, medication changes or appointments). Such learning objectives are likely to be particularly sensitive to a reduction in extrinsic load through simpler language and clearer structure.

Based on these findings, we conclude that AI-based translation into patient letters may reduce extraneous cognitive load, thereby possibly freeing working-memory resources for the processing of concrete, action-oriented tasks (eg, tablet strength and appointment date). However, for learning objectives with high intrinsic cognitive load, such as symptom-to-action coupling (eg, tremors-hypoglycemia compensation required), conditional risk thinking (eg, allergy present notification to the care team), or permanent self-regulation (eg, home blood pressure monitoring and dietary changes) reducing extrinsic cognitive load alone may be not sufficient to free up enough relevant resources (germane cognitive load) for deeper understanding and behavioral change.

Taken together, these CLT-based interpretations are hypothesis-generating and should be tested in future studies that directly measure cognitive load. They further suggest that patient letters may be most effective for immediate, concrete instructions, whereas higher-order learning objectives may require complementary, multimodal supports (eg, infographics, microvideos, interactive prompts, and teach-back). These, however, need prospective evaluation. Evidence for each adjunct already exists in discharge-education and adherence research [38,47,48].

Beyond CLT, factors such as stress-induced cognitive depletion, low health literacy, and limited self-efficacy may also reduce patients' ability to process complex medical content [49-51]. Tailoring follow-up communication to these variables will be essential [52].

Clinical Implications

The implementation of AI-generated patient letters requires careful attention to ethical, legal, and social aspects [53]. In general, human oversight ("human-in-the-loop") remains essential when using AI to translate clinician-written discharge letters into patient letters. Our study indicates that the patient letter format alone does not yield sufficient comprehension gains for cognitively demanding learning objectives, particularly those classified as "Understand" by Bloom taxonomy and thematically related to prevention of complications or lifestyle/disease management. It is further underscored by omissions and hallucinations observed in the prior translation process [12], which required the exclusion of 5 learning objectives in this study. Therefore, establishing practical governance structures is essential for clinical practice. By identifying the specific content domains most prone to errors or omissions, our results can help to target quality-assurance efforts efficiently, making oversight both feasible and focused on the areas critical for patient safety.

From a legal perspective, robust data protection is critical, as personal patient information constitutes a central component of these letters [54,55]. Socially, simplified language may improve accessibility for patients with low health literacy and reduce language barriers [56]. While the format does not accommodate individual literacy needs or reading difficulties, it may still help reduce social barriers in discharge communication. Finally, practical and feasibility considerations include technical integration into existing clinical workflows, usability for health care staff, and costs associated with development, implementation, and maintenance. However, manual letter production is time-consuming and costly, whereas AI offers a potentially scalable solution. Effective measures without AI can be found in the literature, including patient letters written in simple language and structured summaries that reduce technical jargon and organize content for easy recall [57], portal access (facilitating review but not guaranteeing understanding) [58,59], supplementary written materials (linked with better medication adherence) [38], audio recordings of consultations (supporting recall) [43], and interactive elements such as brief teach-back or app-based prompts (which enhance encoding and memory) [47,49]. However, implementation is frequently constrained by time and workload, limited organizational support and training, workflow barriers, and gaps in digital access and use [58,59]. These problems are documented across teach-back implementation (time/sustainability barriers), discharge letter production (time pressure, authorship level, and workload effects), and portal-based delivery (digital divide: differences in access, skills, and use) [60-62].

Our study isolates the contribution of AI-generated formulations and was not designed as a head-to-head comparison. In this context, AI-generated patient letters, with clinician oversight, may offer a practical, scalable way to produce patient-centered text for domains where our data indicate clear gains (eg, medication and organization), and future work should address workflow integration, usability for clinicians, maintenance costs, and robust data protection.

Limitations

This study has several limitations. First, we used only 1 GPT-4 version and focused on 3 chronic diseases (hypertension, diabetes, and diabetic kidney disease), which limits generalizability to other LLM versions and medical conditions. Although participants and clinician raters were blinded to letter type and order as far as possible, full blinding was not achievable, and the success of blinding was not formally assessed for either group. Because the stylistic differences between discharge letters and patient letters are recognizable, which was unavoidable given the study content, participants and raters may have inferred the format, introducing potential expectancy effects. Such effects could operate in both directions: some participants might assume the patient letter to be easier, whereas others may approach AI-generated text with skepticism, potentially reducing trust or attention. These bidirectional biases could either amplify or attenuate format-related differences. While standardized instructions, structured scoring procedures, and uniform training were used to minimize subjective influence, residual bias cannot be excluded and should be considered when interpreting the magnitude of effects. Additionally, we did not formally assess the feasibility or acceptability of patient letters through validated usability or satisfaction scales. These aspects should be addressed in future implementation studies. Furthermore, our study used standardized rather than real-world patients to examine the effects of AI-generated patient letters.

XSL•FO
RenderX

This approach ensured high procedural consistency and enabled a controlled comparison between AI-generated and human-written letters, as reflected in recent high-impact AI studies using SPs [21]. However, it limits external validity; SPs are not able to fully represent real patients' stress levels, emotional states, cognitive vulnerability, and health literacy diversity, especially in the context of acute illness or hospitalization. In clinical practice, such factors may negatively affect patient comprehension due to reduced germane cognitive load. Consequently, although the direction of the observed effects (improved understanding through AI-generated patient letters) is expected to remain, the overall level of understanding and recall would likely be lower among real patient populations for both letter types. In addition, letters were written in German and later translated for publication, limiting the generalizability of our findings to other languages. Another potential limitation of this study is the exclusion of 5 predefined learning objectives that were not present in the GPT-4–generated patient letters due to omissions, which could introduce bias in favor of the AI-generated format. However, a statistical analysis including all 72 learning objectives showed that the overall pattern of results remained unchanged despite the absence of these 5 items in the AI-generated patient letters. This suggests that the exclusion did not materially affect the observed effects. Additionally, the sample size in our study was determined by availability rather than by formal power calculation, which may limit the ability to detect subtle letter-type differences, particularly in subgroup analyses. Our findings should therefore be interpreted with caution. Finally, given the potential for hallucinations or omissions in AI-generated texts, as observed in the translation process [12], clinical review and governance are essential if AI-based tools are deployed in practice, and attention to AI transparency and the potential for unexpected outputs remains critical.

Conclusions

This hypothesis-driven, explanatory study showed that AI-generated patient letters improved the uptake of basic, action-oriented information in SPs. This effect might reflect a reduction in extraneous cognitive load. We hypothesize that, in clinical practice, this could reduce common sources of postdischarge harm and improve patient safety. However, patient letters alone did not close gaps in higher-order understanding, particularly in risk prevention and lifestyle change. Addressing such complex learning goals may require multimodal, interactive support (eg, explanatory graphics, short videos, or app-based follow-ups [47,48,63]) that reinforce the text-based patient letter.

Future studies should therefore evaluate integrative models that layer such tools onto patient letters and tailor content to health literacy, self-efficacy, and situational resources [64], ideally within a didactic framework (eg, Observing Patient Involvement Scale or Patient Education Materials Assessment Tool [65]) to enhance actionability. Studies including more diverse medical conditions, real patient populations with varying health literacy levels, different LLMs, and multiple languages are needed to improve generalizability. We also suggest incorporating qualitative approaches in future studies (eg, think-aloud protocols or teach-back methods) to better capture nuances in patient understanding and potential contributors to medical error.

Acknowledgments

We wish to thank the University of Tübingen and Eric Nazarenus for supporting our study.

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading, editing, and translation.

The GenAI tools used were: ChatGPT by OpenAI, DeepL SE.

Responsibility for the final manuscript lies entirely with the authors.

GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Funding

This work was supported by the Open Access Publishing Fund of the University of Tübingen.

Authors' Contributions

FE had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: FE, FH, AS
Acquisition, analysis, or interpretation of data: all authors
Drafting of the manuscript: FE, AS, FH
Critical review of the manuscript for important intellectual content: all authors
Statistical analysis: KS

XSL•FO
RenderX

Administrative, technical, or material support: CSP, KS, PS
Supervision: AHW, MH

Conflicts of Interest

None declared.

Multimedia Appendix 1
GPT-4–generated patient letters (German original and English translation).
[PDF File (Adobe PDF File), 319 KB - jmir_v28i1e81243_app1.pdf]

Multimedia Appendix 2
Study flowchart. Overview of participant enrollment, randomization, crossover procedure, and analysis population.
[PDF File (Adobe PDF File), 86 KB - jmir_v28i1e81243_app2.pdf]

Multimedia Appendix 3
Overview of the learning objectives, standardized questions, and the corresponding rating scheme applied to evaluate information retention and comprehension across study conditions.
[XLSX File (Microsoft Excel File), 18 KB - jmir_v28i1e81243_app3.xlsx]

Multimedia Appendix 4
List of the learning objectives that were not represented in the GPT-4–generated patient letters because the model did not carry them over from the original discharge letters. These objectives were excluded from the analysis for both letter types.
[DOCX File , 22 KB - jmir_v28i1e81243_app4.docx]

Multimedia Appendix 5
Likelihood ratio tests.
[DOCX File , 27 KB - jmir_v28i1e81243_app5.docx]

Multimedia Appendix 6
STROBE checklist detailing how each reporting item was fulfilled and where it can be found in the manuscript.
[DOCX File , 33 KB - jmir_v28i1e81243_app6.docx]

Multimedia Appendix 7
Number of learning outcomes achieved, stratified by letter type (patient letter vs discharge letter).
[DOCX File , 25 KB - jmir_v28i1e81243_app7.docx]

Multimedia Appendix 8
Item-level analysis identifying learning objectives that were consistently missed, predefined as those “not reported” by ≥50% of participants.
[DOCX File , 22 KB - jmir_v28i1e81243_app8.docx]

Multimedia Appendix 9
Odds ratios from the best-fitting cumulative link mixed model. Odds ratios represent the interaction between letter type and content field, estimated relative to an arbitrary reference category. The reference category is defined as letter type = discharge letter (DL) and content field = Lifestyle / Disease Management. *P* values for the odds ratios are calculated in comparison with this reference category.
[PNG File , 106 KB - jmir_v28i1e81243_app9.png]

References

1. Marsall M, Hornung T, Bäuerle A, Weigl M. Quality of care transition, patient safety incidents, and patients' health status: a structural equation model on the complexity of the discharge process. BMC Health Serv Res 2024;24(1):576 [FREE Full text] [doi: 10.1186/s12913-024-11047-3] [Medline: 38702719]
2. Moore C, Wisnivesky J, Williams S, McGinn T. Medical errors related to discontinuity of care from an inpatient to an outpatient setting. J Gen Intern Med 2003;18(8):646-651 [FREE Full text] [doi: 10.1046/j.1525-1497.2003.20722.x] [Medline: 12911647]

XSL•FO
RenderX

3. Steele EM, Burney HN, Freije SL, Zellars RC, Lautenschlaeger T, Holmes JA. Missed follow-up is associated with worse survival in stage I lung cancer: results from a large multi-site academic hospital system. Sci Rep 2024;14(1):17710 [FREE Full text] [doi: 10.1038/s41598-024-68351-5] [Medline: 39085317]
4. Becker C, Zumbrunn S, Beck K, Vincent A, Loretz N, Müller J, et al. Interventions to improve communication at hospital discharge and rates of readmission: a systematic review and meta-analysis. JAMA Netw Open 2021;4(8):e2119346 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.19346] [Medline: 34448868]
5. Engel KG, Heisler M, Smith DM, Robinson CH, Forman JH, Ubel PA. Patient comprehension of emergency department care and instructions: are patients aware of when they do not understand? Ann Emerg Med 2009;53(4):454-461.e15. [doi: 10.1016/j.annemergmed.2008.05.016] [Medline: 18619710]
6. Belachew EA, Sendekie AK, Wondm SA, Ayele EM, Netere AK. Misunderstanding of dosing regimen instructions among patients with chronic diseases receiving polypharmacy at the University of Gondar comprehensive specialized hospital. PLoS One 2023;18(1):e0280204 [FREE Full text] [doi: 10.1371/journal.pone.0280204] [Medline: 36634103]
7. Kreps GL. Promoting patient comprehension of relevant health information. Isr J Health Policy Res 2018;7(1):56 [FREE Full text] [doi: 10.1186/s13584-018-0250-z] [Medline: 30236157]
8. Toole J, Kohansieh M, Khan U, Romero S, Ghali M, Zeltser R, et al. Does your patient understand their treatment plan? Factors affecting patient understanding of their medical care treatment plan in the inpatient setting. J Patient Exp 2020;7(6):1151-1157 [FREE Full text] [doi: 10.1177/2374373520948400] [Medline: 33457558]
9. Magnani JW, Mujahid MS, Aronow HD, Cené CW, Dickson VV, Havranek E, American Heart Association Council on EpidemiologyPrevention; Council on Cardiovascular Disease in the Young; Council on CardiovascularStroke Nursing; Council on Peripheral Vascular Disease; Council on Quality of CareOutcomes Research;Stroke Council. Health literacy and cardiovascular disease: fundamental relevance to primary and secondary prevention: a scientific statement from the American Heart Association. Circulation 2018;138(2):e48-e74 [FREE Full text] [doi: 10.1161/CIR.0000000000000579] [Medline: 29866648]
10. Choudhry AJ, Baghdadi YMK, Wagie AE, Habermann EB, Heller SF, Jenkins DH, et al. Readability of discharge summaries: with what level of information are we dismissing our patients? Am J Surg 2016;211(3):631-636 [FREE Full text] [doi: 10.1016/j.amjsurg.2015.12.005] [Medline: 26794665]
11. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Health 2023;5(4):e179-e181 [FREE Full text] [doi: 10.1016/S2589-7500(23)00048-1] [Medline: 36894409]
12. Eisinger F, Holderried F, Mahling M, Stegemann-Philipps C, Herrmann-Werner A, Nazarenus E, et al. What's going on with me and how can I better manage my health? The potential of GPT-4 to transform discharge letters into patient-centered letters to enhance patient safety: prospective, exploratory study. J Med Internet Res 2025;27:e67143 [FREE Full text] [doi: 10.2196/67143] [Medline: 39836954]
13. Zaretsky J, Kim JM, Baskharoun S, Zhao Y, Austrian J, Aphinyanaphongs Y, et al. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. JAMA Netw Open 2024;7(3):e240357 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.0357] [Medline: 38466307]
14. Shahnam A, Nindra U, Hitchen N, Tang J, Hong M, Hong JH, et al. Application of generative artificial intelligence for physician and patient oncology letters—AI-OncLetters. JCO Clin Cancer Inform 2025(9):e2400323. [doi: 10.1200/cci-24-00323]
15. Kumar A, Wang H, Muir KW, Mishra V, Engelhard M. A cross-sectional study of GPT-4–based plain language translation of clinical notes to improve patient comprehension of disease course and management. NEJM AI 2025;2(2):AIoa2400402. [doi: 10.1056/aioa2400402]
16. van Driel MHE, Blok N, van den Brand JAJG, van de Sande D, de Vries M, Eijlers B, et al. Leveraging GPT-4 enables patient comprehension of radiology reports. Eur J Radiol 2025;187:112111 [FREE Full text] [doi: 10.1016/j.ejrad.2025.112111] [Medline: 40294519]
17. Kim SI, Park J, Kim T, Seo W, Kim T, Cha WC, et al. Enhancing patient participation in emergency department through patient-friendly clinical notes generated by large language models. Sci Rep 2025;16(1):1409 [FREE Full text] [doi: 10.1038/s41598-025-31113-y] [Medline: 41350380]
18. Sweller J. Cognitive load during problem solving: effects on learning. Cogn Sci 2010;12(2):257-285 [FREE Full text] [doi: 10.1207/s15516709cog1202_4]
19. Adams NE. Bloom's taxonomy of cognitive learning objectives. J Med Libr Assoc 2015;103(3):152-153 [FREE Full text] [doi: 10.3163/1536-5050.103.3.010] [Medline: 26213509]
20. Felix HM, Simon LV. Types of Standardized Patients and Recruitment in Medical Simulation. Treasure Island (FL): StatPearls; 2025.
21. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature 2025;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
22. Hellyer S. Supplementary report: frequency of stimulus presentation and short-term decrement in recall. J Exp Psychol 1962;64(6):650-650. [doi: 10.1037/h0043117] [Medline: 13953638]

XSL•FO
RenderX

23. McEvoy JW, McCarthy CP, Bruno RM, Brouwers S, Canavan MD, Ceconi C, ESC Scientific Document Group. 2024 ESC guidelines for the management of elevated blood pressure and hypertension. Eur Heart J 2024;45(38):3912-4018. [doi: 10.1093/eurheartj/ehae178] [Medline: 39210715]
24. American Diabetes Association Professional Practice Committee. 6. Glycemic goals and hypoglycemia: standards of care in diabetes-2024. Diabetes Care 2024;47(Suppl 1):S111-S125. [doi: 10.2337/dc24-S006] [Medline: 38078586]
25. American Diabetes Association Professional Practice Committee. 2. Diagnosis and classification of diabetes: standards of care in diabetes-2024. Diabetes Care 2024;47(Suppl 1):S20-S42. [doi: 10.2337/dc24-S002] [Medline: 38078589]
26. American Diabetes Association Professional Practice Committee. 3. Prevention or delay of diabetes and associated comorbidities: standards of care in diabetes-2024. Diabetes Care 2024;47(Suppl 1):S43-S51. [doi: 10.2337/dc24-S003] [Medline: 38078581]
27. American Diabetes Association Professional Practice Committee. 11. Chronic kidney disease and risk management: standards of care in diabetes-2024. Diabetes Care 2024;47(Suppl 1):S219-S230. [doi: 10.2337/dc24-S011] [Medline: 38078574]
28. Krathwohl DR. A revision of Bloom's taxonomy: an overview. Theory Pract 2010;41(4):212-218. [doi: 10.1207/s15430421tip4104_2]
29. Biggs JB, Tang CSK. Teaching for Quality Learning at University What the Student Does. 4th ed. Maidenhead: Society for Research into Higher Education; 2011.
30. Abdolreza GZ, Mansourzadeh A, Montazeri KA, Saeidi M. Reflections on using open-ended questions. Med Educ Bull 2022;3(2):475-482. [doi: 10.22034/meb.2022.333518.1054]
31. Haubo R, Christensen B, editors. Cumulative Link Models for Ordinal Regression With the R Package ordinal. Denmark: Christensen Statistics; 2018.
32. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Core Team, R Foundation for Statistical Computing; 2025.
33. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med 2007;4(10):e296 [FREE Full text] [doi: 10.1371/journal.pmed.0040296] [Medline: 17941714]
34. Hoffmann H, Jonietz A, Gräfe W, Zenker R, Voigt K, Riemenschneider H. Associations of an easy-to-understand patient letter on the health literacy of patients after discharge from hospital: results of a randomized controlled intervention study. Gesundheitswesen 2023;85(S 03):S183-S188 [FREE Full text] [doi: 10.1055/a-2130-2374] [Medline: 37751754]
35. Lin R, Gallagher R, Spinaze M, Najoumian H, Dennis C, Clifton-Bligh R, et al. Effect of a patient-directed discharge letter on patient understanding of their hospitalisation. Intern Med J 2014;44(9):851-857. [doi: 10.1111/imj.12482] [Medline: 24863954]
36. Makaryus AN, Friedman EA. Patients' understanding of their treatment plans and diagnosis at discharge. Mayo Clin Proc 2005;80(8):991-994. [doi: 10.4065/80.8.991] [Medline: 16092576]
37. Hodkinson A, Tyler N, Ashcroft DM, Keers RN, Khan K, Phipps D, et al. Preventable medication harm across health care settings: a systematic review and meta-analysis. BMC Med 2020;18(1):313 [FREE Full text] [doi: 10.1186/s12916-020-01774-9] [Medline: 33153451]
38. Al Jeraisy M, Alshammari H, Albassam M, Al Aamer K, Abolfotouh MA. Utility of patient information leaflet and perceived impact of its use on medication adherence. BMC Public Health 2023;23(1):488 [FREE Full text] [doi: 10.1186/s12889-023-15346-y] [Medline: 36918823]
39. Coppa K, Kim EJ, Oppenheim MI, Bock KR, Conigliaro J, Hirsch JS. Examination of post-discharge follow-up appointment status and 30-day readmission. J Gen Intern Med 2021;36(5):1214-1221 [FREE Full text] [doi: 10.1007/s11606-020-06569-5] [Medline: 33469750]
40. McAlister FA, Youngson E, Bakal JA, Kaul P, Ezekowitz J, van Walraven C. Impact of physician continuity on death or urgent readmission after discharge among patients with heart failure. CMAJ 2013;185(14):E681-E689 [FREE Full text] [doi: 10.1503/cmaj.130048] [Medline: 23959284]
41. Baky V, Moran D, Warwick T, George A, Williams T, McWilliams E, et al. Obtaining a follow-up appointment before discharge protects against readmission for patients with acute coronary syndrome and heart failure: a quality improvement project. Int J Cardiol 2018;257:12-15. [doi: 10.1016/j.ijcard.2017.10.036] [Medline: 29506682]
42. Shahid R, Shoker M, Chu LM, Frehlick R, Ward H, Pahwa P. Impact of low health literacy on patients' health outcomes: a multicenter cohort study. BMC Health Serv Res 2022;22(1):1148 [FREE Full text] [doi: 10.1186/s12913-022-08527-9] [Medline: 36096793]
43. Wolderslund M, Kofoed PE, Holst R, Waidtløw K, Ammentorp J. Outpatients' recall of information when provided with an audio recording: a mixed-methods study. Patient Educ Couns 2020;103(1):63-70. [doi: 10.1016/j.pec.2019.08.030] [Medline: 31473043]
44. Laws MB, Lee Y, Taubin T, Rogers WH, Wilson IB. Factors associated with patient recall of key information in ambulatory specialty care visits: results of an innovative methodology. PLoS One 2018;13(2):e0191940 [FREE Full text] [doi: 10.1371/journal.pone.0191940] [Medline: 29389994]
45. Kessels RPC. Patients' memory for medical information. J R Soc Med 2003;96(5):219-222 [FREE Full text] [doi: 10.1177/014107680309600504] [Medline: 12724430]

XSL•FO
RenderX

46. Szulewski A, Howes D, van Merriënboer JJG, Sweller J. From theory to practice: the application of cognitive load theory to the practice of medicine. Acad Med 2021;96(1):24-30. [doi: 10.1097/ACM.0000000000003524] [Medline: 32496287]
47. Xu Q, Sundar SS. Interactivity and memory: Information processing of interactive versus non-interactive content. Comput Hum Behav 2016;63:620-629 [FREE Full text] [doi: 10.1016/j.chb.2016.05.046]
48. Dada A, Koras OA, Bauer M, Butler A, Smith KE, Kleesiek J. MeDiSumQA: patient-oriented question-answer generation from discharge letters. arXiv. Preprint posted online on February 5, 2025 2025. [doi: 10.18653/v1/2025.cl4health-1.10]
49. Baumeister RF, André N, Southwick DA, Tice DM. Self-control and limited willpower: current status of ego depletion theory and research. Curr Opin Psychol 2024;60:101882 [FREE Full text] [doi: 10.1016/j.copsyc.2024.101882] [Medline: 39278166]
50. Cui Y, Li R, Yang T, Wang H, Jin S, Liu N, et al. Influence of positive and negative affect on self-management among patients with early chronic kidney disease during the COVID-19 pandemic: the mediating and suppressing effect of ego depletion. Front Psychiatry 2022;13:992404 [FREE Full text] [doi: 10.3389/fpsyt.2022.992404] [Medline: 36245863]
51. Yumiya Y, Goto A, Konta T. Association between health literacy and understanding of doctors' explanations: the Yamagata study. Health Lit Res Pract 2024;8(3):e175-e183 [FREE Full text] [doi: 10.3928/24748307-20240819-03] [Medline: 39251187]
52. Stormacq C, Wosinski J, Boillat E, Van den Broucke S. Effects of health literacy interventions on health-related outcomes in socioeconomically disadvantaged adults living in the community: a systematic review. JBI Evid Synth 2020;18(7):1389-1469. [doi: 10.11124/JBISRIR-D-18-00023] [Medline: 32813388]
53. Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model AI chatbots require approval as medical devices. Nat Med 2023;29(10):2396-2398. [doi: 10.1038/s41591-023-02412-6] [Medline: 37391665]
54. Gorelik AJ, Li M, Hahne J, Wang J, Ren Y, Yang L, et al. Ethics of AI in healthcare: a scoping review demonstrating applicability of a foundational framework. Front Digit Health 2025;7:1662642 [FREE Full text] [doi: 10.3389/fdgth.2025.1662642] [Medline: 41000336]
55. Park MK, Ashwood N, Capes N. Ethics of artificial intelligence in medicine. Cureus 2025;17(5):e83567. [doi: 10.7759/cureus.83567] [Medline: 40486469]
56. Kong M, Fernandez A, Bains J, Milisavljevic A, Brooks KC, Shanmugam A, et al. Evaluation of the accuracy and safety of machine translation of patient-specific discharge instructions: a comparative analysis. BMJ Qual Saf 2025:bmjqs-2024-018384 [FREE Full text] [doi: 10.1136/bmjqs-2024-018384] [Medline: 40633961]
57. Siegrist V, Mata R, Langewitz W, Gerger H, Furger S, Hertwig R, et al. Does information structuring improve recall of discharge information? A cluster randomized clinical trial. PLoS One 2021;16(10):e0257656 [FREE Full text] [doi: 10.1371/journal.pone.0257656] [Medline: 34662341]
58. Brands MR, Gouw SC, Beestrum M, Cronin RM, Fijnvandraat K, Badawy SM. Patient-centered digital health records and their effects on health outcomes: systematic review. J Med Internet Res 2022;24(12):e43086 [FREE Full text] [doi: 10.2196/43086] [Medline: 36548034]
59. Neves AL, Freise L, Laranjo L, Carter AW, Darzi A, Mayer E. Impact of providing patients access to electronic health records on quality and safety of care: a systematic review and meta-analysis. BMJ Qual Saf 2020;29(12):1019-1032 [FREE Full text] [doi: 10.1136/bmjqs-2019-010581] [Medline: 32532814]
60. Talevski J, Wong Shee A, Rasmussen B, Kemp G, Beauchamp A. Teach-back: a systematic review of implementation and impacts. PLoS One 2020;15(4):e0231350 [FREE Full text] [doi: 10.1371/journal.pone.0231350] [Medline: 32287296]
61. Weetman K, Spencer R, Dale J, Scott E, Schnurr S. What makes a "successful" or "unsuccessful" discharge letter? Hospital clinician and general practitioner assessments of the quality of discharge letters. BMC Health Serv Res 2021;21(1):349 [FREE Full text] [doi: 10.1186/s12913-021-06345-z] [Medline: 33858383]
62. Coit MH, Katz JT, McMahon GT. The effect of workload reduction on the quality of residents' discharge summaries. J Gen Intern Med 2011;26(1):28-32 [FREE Full text] [doi: 10.1007/s11606-010-1465-z] [Medline: 20697967]
63. Carris NW, Bullers K, McKee M, Schanze J, Eubanks T, Epperson C, et al. Automated lifestyle interventions and weight loss: a systematic review of randomized controlled trials. Int J Obes (Lond) 2025;49(6):992-997. [doi: 10.1038/s41366-025-01746-0] [Medline: 40158054]
64. Sajjadi SL, Ghafourifard M, Khosroshahi HT. The effect of individualized education on learning needs of patients undergoing hemodialysis: a randomized controlled clinical trial. BMC Nephrol 2024;25(1):452 [FREE Full text] [doi: 10.1186/s12882-024-03886-3] [Medline: 39695998]
65. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014;96(3):395-403. [doi: 10.1016/j.pec.2014.05.027] [Medline: 24973195]

Abbreviations

AI: artificial intelligence
CLT: cognitive load theory
LLM: large language model

XSL•FO
RenderX

OR: odds ratio
SP: standardized patient
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by J Sarvestan; submitted 26.Jul.2025; peer-reviewed by F Al Dhabbari, Z Vojvodic, K Fiscella; comments to author 01.Sep.2025; revised version received 23.Dec.2025; accepted 24.Dec.2025; published 26.Feb.2026.

Please cite as:
Holderried F, Sonanini A, Stegemann–Philipps C, Herrmann–Werner A, Spitzer P, Guthoff M, Heyne N, Sering K, Holderried M, Eisinger F
Impact of GPT-4–Generated Discharge Letters on Patients' Medical Comprehension: Prospective Crossover Study
J Med Internet Res 2026;28:e81243
URL: https://www.jmir.org/2026/1/e81243
doi:10.2196/81243
PMID:41746691

©Friederike Holderried, Alessandra Sonanini, Christian Stegemann–Philipps, Anne Herrmann–Werner, Philipp Spitzer, Martina Guthoff, Nils Heyne, Konstantin Sering, Martin Holderried, Felix Eisinger. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Real-Time Support Role of Augmented Reality Technology in Shared Decision-Making in Neurosurgery Under the SEGUE Framework: Randomized Controlled Trial

Zhengbo Yuan[1*], Dr med; Zhongjie Shi[1,2*], Dr med; Zhanxiang Wang[2,3], Prof Dr Med

[1]Department of Neurosurgery, The First Affiliated Hospital of Xiamen University, School of Medicine, Xiamen University, Xiamen, Fujian, China

[2]National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, Fujian, China

[3]Department of Neurosurgery and Department of Neuroscience, Fujian Key Laboratory of Brain Tumors Diagnosis and Precision Treatment, The First Affiliated Hospital of Xiamen University, School of Medicine, Xiamen University, Xiamen, Fujian, China

[*]these authors contributed equally

Corresponding Author:
Zhanxiang Wang, Prof Dr Med
Department of Neurosurgery and Department of Neuroscience
Fujian Key Laboratory of Brain Tumors Diagnosis and Precision Treatment, The First Affiliated Hospital of Xiamen University
School of Medicine, Xiamen University
NO. 55 Zhenhai Road, Siming District
Xiamen, Fujian, 361003
China
Phone: 86 15154608303
Email: wangzx@xmu.edu.cn

Abstract

Background: Preoperative risk communication is essential for shared decision-making (SDM) in neurosurgery; however, conveying complex neuroanatomy and surgical risks using traditional verbal explanations can limit understanding and contribute to dissatisfaction and medicolegal disputes. Augmented reality (AR) may provide patient-specific, interactive 3D visualization to support these conversations.

Objective: This study evaluates whether AR-assisted preoperative risk communication improves objective understanding and other SDM-related outcomes, compared with communication supported by a conventional physical anatomical model within a standardized SEGUE-informed protocol.

Methods: A prospective, single-center, randomized controlled trial was conducted with 62 neurosurgery communication recipients (patients when capable; otherwise, a legally authorized representative [LAR]). Patients were stratified by planned surgical approach (frontal, parietal, and occipital) and, within each stratum, were randomized into an experimental AR group and a control physical-model group. The primary outcome was postsession objective understanding, assessed by a multiple-choice knowledge questionnaire. Secondary outcomes were subjective understanding, communication satisfaction, pre-to-post anxiety changes, communication duration, and neurosurgeons' communication skills from video recordings.

Results: Of the 67 individuals screened, 62 communication recipients were enrolled and completed all assessments (patients, n=30; LARs, n=32; AR, n=32; and control, n=30). Objective understanding was higher with AR than with the physical model in the prespecified pooled comparison (P=.01). Communication satisfaction was also higher with AR (P<.001). There were no clear between-group differences in subjective understanding (P=.41), anxiety changes (ΔState-Trait Anxiety Inventory [ΔSTAI] Y-1, P=.37; ΔSTAI Y-2, P=.84), in-session face-to-face communication duration (excluding any presession AR technical preparation time; P=.73), or SEGUE scores (P=.60). Exploratory stratified analyses suggested larger comprehension gains with AR in the parietal and occipital approach strata.

Conclusions: In a standardized preoperative SDM conversation, an integrated visualization-support package combining AR, patient-specific modeling, and interactivity improved neurosurgical decision makers' objective understanding and satisfaction without prolonging in-session, face-to-face communication duration. Larger multicenter trials with longer-term outcomes are warranted to confirm effectiveness and to evaluate implementation and cost considerations.

Trial Registration: ISRCTN Registry ISRCTN11483487; http://www.isrctn.com/ISRCTN11483487

XSL•FO
RenderX

(J Med Internet Res 2026;28:e87198) doi:10.2196/87198

KEYWORDS

augmented reality; AR; shared decision-making; SDM; neurosurgery; preoperative communication; risk communication; informed consent

Introduction

Effective communication between health care professionals and patients has become a priority in modern medicine. Evidence indicates that high-quality communication improves patients' comprehension and retention of medical information, enhances treatment adherence and coping skills, and ultimately benefits well-being, quality of life, and clinical outcomes [1,2]. More specifically, effective risk communication ensures that all parties share as accurate an understanding as possible of the factors relevant to clinical choices, thereby supporting shared responsibility for medical and surgical decisions [3]. In the context of neurosurgical decision-making, "risk communication" refers to a structured clinical process in which clinicians and decision makers exchange information about the expected benefits, probabilities, and consequences of potential harms, available alternatives, and remaining uncertainties, with the goal of supporting informed, values-concordant decisions. Risk communication therefore includes not only the content of the information shared but also how that information is organized and delivered (eg, plain language, opportunities for questions, and checking comprehension). This aligns with patients' growing desire to be closely involved in their own care, reflected in the increasing use of shared decision-making (SDM) in surgical settings [4]. SDM is a joint process in which health care professionals and patients collaborate to choose among options [5,6]: clinicians present alternatives and explain risks and benefits; patients convey preferences and values; and, together, they grant or withhold consent to a specific intervention. Risk communication—the systematic exchange of information about the nature, magnitude, importance, and controllability of risks—is a core component of SDM [7]. When risk communication is effective, patients should understand the planned intervention and its risks, benefits, and reasonable alternatives [8].

In practice, preoperative conversations often encounter a dual challenge of providing "too little" and "too much" information. Oversimplified explanations are insufficient for sound decision-making, whereas unstructured, jargon-dense detail produces cognitive overload that impedes comprehension for nonexperts [9]. Studies consistently document suboptimal clinician-patient communication, leading to mutual misunderstanding [10]. In neurosurgical settings, time and workflow pressures can foster a checklist-style recital of risk factors. Although this approach may satisfy formal requirements, it weakens interactivity and relevance, limiting opportunities for patients and their legally authorized representatives (LARs) to discuss concerns openly and thereby diminishing the value of SDM [4,11].

As a specialty, neurosurgery manages a high volume of urgent and critical cases, with an overall complication rate near 14%, often resulting in severe functional disability [12]. When expectations of patients or LARs diverge from realistic prognoses, disputes and claims are more likely; indeed, neurosurgeons face the highest medicolegal risk among surgical disciplines [13]. Notably, inadequate or incomplete risk communication is among the most frequent contributors to litigation [14]. In time-pressured environments, a passive or compliance-based approach rarely enables patients to visualize and grasp risks linked to complex neuroanatomy, lesion margins, and eloquent functional areas [15].

To address these gaps, communication-support and decision-aid tools have been developed for diverse clinical contexts, including brochures, information leaflets, and mobile apps [16-19]. Augmented reality (AR) is increasingly leveraged across disciplines as a convergent, interdisciplinary technology. AR uses computer vision, graphics, and multimodal sensors to accurately register and superimpose virtual content (images, models, video, audio, and text) onto the user's real-world view, creating a collaborative mixed environment for natural, real-time interaction and extended perception [20]. In medicine, AR is not new and has been applied across multiple specialties [21]. In neurosurgery, its capacity for real-time 3D reconstruction has supported preoperative planning, intraoperative navigation, and skills training [22,23]. However, comparative evidence on AR's benefits for preoperative communication in neurosurgery remains limited. Clinicians hypothesize that AR's intuitive 3D models can improve the transmission and understanding of complex information, but this requires systematic evaluation. In this study, "real-time support" refers to the in-session use of an AR-based, patient-specific 3D visualization as a shared reference during communication, enabling the clinician and the communication recipient to jointly orient to key structures, the target lesion, and clinically relevant high-risk regions as questions arise. Importantly, this term is used as a pragmatic description of an integrated visualization-support workflow (patient-specific anatomical content, head-mounted spatial presentation, and enhanced interactivity), rather than as a claim that the study isolates a single causal mechanism attributable to AR hardware alone.

Building on our institution's experience with AR and incorporating Xvisio Technology's spatial-perception interaction technology (Shanghai, China), we designed a study to evaluate AR-assisted preoperative risk communication in neurosurgery. In this study, we standardized the core informational items and communication structure across arms (SEGUE [*s*et the stage, *e*licit information, *g*ive information, *u*nderstand the patient's perspective, and *e*nd the encounter]-informed checklist), and tested whether an AR-assisted, patient-specific 3D visualization—compared with a physical anatomical model—could improve decision makers' understanding within the same risk-communication protocol. Using the SEGUE framework, we aim to implement AR within SDM in a standardized and streamlined manner. Our objectives are to (1)

XSL•FO
RenderX

assess the effect of AR on comprehension and satisfaction; (2) develop an intuitive, reusable, and scalable workflow for preoperative SDM; and (3) provide a methodological basis for future multicenter, controlled studies. Accordingly, this proof-of-concept (PoC) trial is positioned as an exploratory, pragmatic comparison of 2 visualization-support packages (AR-assisted, patient-specific interactive 3D visualization vs a conventional, generic physical anatomical model), rather than a component-isolating test designed to separate spatial representation from interactivity or informational richness.

Methods

Trial Design

Our research was a prospective, single-center, randomized controlled PoC trial. According to the planned surgical approaches, participants were stratified into the frontal approach group (group 1), parietal approach group (group 2), and occipital approach group (Group 3), with randomization within each stratum to the experimental group (AR group) or the control group (physical-model group). Randomization within each surgical-approach stratum used a 1:1 allocation ratio (AR group vs physical-model group). The allocation sequence was generated using a computer-generated random-number list with simple randomization. To ensure allocation concealment, the sequence was implemented using a centralized web-based system prepared and maintained by an independent statistician. Baseline assessments (eg, State-Trait Anxiety Inventory [STAI] and Montreal Cognitive Assessment [MoCA]) were completed before group assignment. Participants were enrolled at the First Affiliated Hospital of Xiamen University, and assignment to interventions was implemented through the concealed allocation mechanism described above. Because of the nature of the interventions, clinicians and communication recipients were not blinded to group allocation. All recipient-reported outcomes (comprehension, anxiety, and satisfaction) were completed by the communication recipient. Neurosurgeons' communication skills were evaluated using the SEGUE scale based on video recordings by independent assessors who were not involved in intervention delivery (assessors were not informed of allocation and scored recordings labeled only with study IDs).

This study aimed to preliminarily validate the feasibility and advantages of AR technology in preoperative risk communication for neurosurgery. By superimposing an individualized 3D virtual model from the AR headset onto the real environment and demonstrating it to the communication recipient, we compared this approach with the traditional physical model–assisted communication method across both subjective and objective dimensions. This comparison aimed to explore the impact of different communication-assistance technologies on recipients' understanding of surgery-related issues (eg, surgical methods, surgical risks), communication satisfaction, and anxiety levels.

Participants

This study included 67 patients who visited the Department of Neurosurgery at the First Affiliated Hospital of Xiamen University between January and September 2025 and were diagnosed with intracranial space-occupying lesions through auxiliary examinations after admission. As this is a communication intervention study, the enrolled participant was defined as the "communication recipient," that is, the person who received the standardized preoperative risk communication and was primarily involved in SDM and consent for surgery. This was the patient when decisional capacity was intact; otherwise, an LAR served as the communication recipient and decision maker. Unless otherwise stated, all psychometric outcomes (STAI and MoCA) and postcommunication assessments (comprehension and satisfaction) refer to the communication recipient (patient or LAR).

All communication recipients (patients with decisional capacity or LARs, when applicable) provided written informed consent before enrollment. Participants were explicitly informed that the preoperative communication session would be video-recorded for research purposes (SEGUE-based assessment of communication quality), that refusal to be recorded would not affect clinical care, and that they could request discontinuation of recording and withdraw from the study at any time without penalty. Video recordings, questionnaire data, and study logs were deidentified by assigning unique study codes and stored on a password-protected, secure institutional drive with access restricted to the research team. Any patient imaging data used to generate the 3D models were processed in deidentified form for research use. Textboxes 1 and 2 detail the inclusion and exclusion criteria, respectively.

XSL•FO
RenderX

Textbox 1. Inclusion criteria.

All the following inclusion criteria must be met:

1. Study patients

- Neurosurgery patients eligible for elective or semielective surgery with a feasible standard surgical approach.

2. Technology

- High-quality computed tomography or magnetic resonance imaging DICOM (Digital Imaging and Communications in Medicine) data from the hospital within 3 days before surgery. If needed, updated imaging must be completed before communication.

3. Eligibility of the communication recipient

- The recipient has no visual or auditory impairments, mental illness, or cognitive disorders. (1) The patient has the ability to provide informed consent and communicate and is therefore the primary communication recipient; or (2) the patient's legally authorized representative is the primary participant in communication and decision-making when the patient lacks decisional capacity or has delegated authority. Decisional capacity was determined by the treating neurosurgeon based on routine clinical assessment and documented in the medical record.

4. Age and language of the communication recipient

- Aged 18-70 years; able to complete questionnaires and tests in Chinese (Mandarin); no gender restriction.

5. Cooperation and compliance

- Agreement to randomization, completion of on-site evaluation and follow-up before discharge, and willingness to wear augmented reality display devices when necessary (augmented reality group).

6. Ethical consent

- Signed informed consent for study participation (by the patient or legally authorized representative) and permission for the deidentified use of 3D imaging data for demonstration.

Textbox 2. Exclusion criteria.

Any of the following conditions will result in exclusion:

1. Clinical process–related

- Emergency surgery patients for whom ≥24 hours is not available for preoperative assessment and communication.
- Patients with severe cognitive or consciousness disorders who are unable to complete the assessment and lack a qualified, legally authorized representative (eg, persistent delirium, severe aphasia, Glasgow Coma Scale score ≤12: E4+V3+M5 or E2+V5+M5).
- Difficulty cooperating with questionnaires and tests (eg, reading/writing disabilities, inability to complete the tests).
- Participation in similar intervention studies or communication-related intervention trials within the past 3 months that may affect primary outcomes (eg, understanding, decision conflict).

2. Contraindications or high risks related to head-mounted augmented reality devices

- History of photosensitive epilepsy or epilepsy clearly induced by visual stimuli.
- Vestibular dysfunction (eg, frequent dizziness attacks in the past 3 months or daily medication to control symptoms).
- Severe migraines with high sensitivity to bright light or dynamic images and frequent recent episodes.
- Severe visual impairments or eye diseases affecting augmented reality device use (eg, corrected vision ≤0.1, poor control of diplopia, postocular surgery recovery period).
- Severe hearing impairments with no equivalent alternative communication methods available on-site (eg, subtitles, hearing aids).
- Limitations on wearing and positioning: inability to safely wear head-mounted augmented reality devices due to external head fixation (eg, halo brace), large head dressings, or cervical fixation devices (eg, neck collar).
- Clear history of adverse reactions to augmented reality devices that led to discontinuation (eg, severe nausea, balance issues) and refusal to attempt use again.

3. Others

- Communication or ethical risks deemed inappropriate for inclusion (eg, family conflict affecting voluntariness).
- Legal restrictions or special populations requiring additional approval that has not been granted (eg, populations under compulsory supervision).

XSL•FO
RenderX

Equipment

Head-mounted AR device (SeerLens B50R AR Glasses; Xvisio Technology) and a physical anatomical model (Hongyu Medical Education Equipment Co Ltd).

The AR intervention used the SeerLens head-mounted AR device. Key manufacturer-reported specifications relevant to clinical use (eg, display type, resolution, field of view, tracking method, weight, and computing/connection configuration) are summarized in Multimedia Appendix 1.

Baseline Variables

At enrollment (before the communication session), we collected demographic characteristics of the communication recipient, including age, sex/gender, and educational attainment (recorded as the highest completed level and, for reporting in Table 1, collapsed into 3 categories to reflect the Mainland China compulsory education context and to reduce sparse subgroup cells). For LARs, we additionally recorded their relationship to the patient. When specified in the study protocol, baseline psychological and cognitive characteristics were also collected, including baseline anxiety assessed using STAI Y-1 and STAI Y-2 (before the communication session) and cognitive screening assessed using the MoCA (before the communication session). Baseline clinical and surgical variables (eg, diagnosis category and planned surgical approach) were extracted from the medical record.

XSL•FO
RenderX

Table 1. Summary of baseline characteristics of participants.[a]

Characteristics	Total (N=62)	Group 1 (N=20)		Group 2 (N=21)		Group 3 (N=21)		*P* value
		Group 1A (n=9)	Group 1B (n=11)	Group 2A (n=11)	Group 2B (n=10)	Group 3A (n=12)	Group 3B (n=9)	
Recipient type, n (%)								.46 (Fisher) (group 1A2A3A vs 1B2B3B)
Patient	30 (48)	4 (44)	5 (45)	7 (64)	4 (40)	6 (50)	4 (44)	
Legally authorized representative	32 (52)	5 (56)	6 (55)	4 (36)	6 (60)	6 (50)	5 (56)	
Gender,n (%)								.44 (Fisher) (group 1A2A3A vs 1B2B3B)
Male	35 (56)	5 (56)	6 (55)	8 (73)	5 (50)	7 (58)	4 (44)	
Female	27 (44)	4 (44)	5 (45)	3 (27)	5 (50)	5 (42)	5 (56)	
Education (highest educational attainment), n (%)								.77 (Fisher; group 1A2A3A vs 1B2B3B)
Middle school or below	19 (31)	2 (22)	3 (27)	4 (36)	3 (30)	4 (33)	3 (33)	
High school	23 (37)	4 (44)	4 (36)	5 (45)	4 (40)	4 (33)	2 (22)	
College or above	20 (32)	3 (33)	4 (36)	2 (18)	3 (30)	4 (33)	4 (44)	
Age (years), mean (SD)	54.48 (9.29)	57.89 (7.11)	57.45 (10.69)	50.45 (9.16)	54.5 (7.61)	54.58 (10.79)	52.22 (9.00)	.91 (group 1A vs 1B); .28 (group 2A vs 2B); .59 (group 3A vs 3B); and .74 (group 1A2A3A vs 1B2B3B)
STAI[b] -Form Y-Chinese version, mean (SD)								
STAI Y-1	52.42 (9.41)	53.00 (10.17)	51.12 (10.62)	54.36 (8.86)	54.10 (8.97)	51.83 (11.24)	49.89 (7.08)	.70 (group 1A vs 1B); .95 (group 2A vs 2B); .63 (group 3A vs 3B); and .16 (group 1A2A3A vs 1B2B3B)
STAI Y-2	50.11 (9.11)	53.78 (8.20)	52.46 (8.74)	51.36 (9.33)	51.60 (9.16)	47.50 (9.89)	43.89 (7.18)	.73 (group 1A vs 1B); .95 (group 2A vs 2B); .34 (group 3A vs 3B); and .67 (group 1A2A3A vs 1B2B3B)
MoCA[c] score, mean (SD)	24.66 (3.88)	24.67 (3.28)	22.82 (4.67)	26.18 (2.68)	23.3 (5.68)	26 (3.5)[d]	25 (1)[d]	.31 (group 1A vs 1B); .17 (group 2A vs 2B); .29 (group 3A vs 3B); and .06 (group 1A2A3A vs 1B2B3B)

[a]All baseline characteristics refer to the communication recipient (patient or LAR), unless otherwise stated. Groups 1A2A3A (n=32) included 17 (53%) patients and 15 (47%) LARs (patient-to-LAR 17:15), whereas groups 1B2B3B (n=30) included 13 (43%) patients and 17 (57%) LARs (patient-to-LAR 13:17). Recipient-type distribution did not differ significantly between arms (Fisher exact test *P*=.46).

[b]STAI: State Anxiety Inventory.

[c]MoCA: Montreal Cognitive Assessment (Chinese 7.1-Beijing).

[d]Reported as median (IQR) because the corresponding subgroup data were nonnormally distributed. The pooled comparison of objective understanding scores between the augmented reality and physical-model groups (groups 1A+2A+3A vs 1B+2B+3B) was prespecified as the single primary confirmatory analysis (2-sided α=.05). All other *P* values (stratum-specific comparisons and all secondary outcomes) are exploratory, unadjusted for multiple comparisons, and should be interpreted descriptively.

XSL•FO
RenderX

Risk Communication Protocol and Standardization

Visualization Aid Intervention and Between-Arm Differentiation

The intended between-arm difference was the visualization aid used to support the same standardized explanation: the AR group received a patient-specific 3D visualization via a head-mounted display with adjustable views (eg, rotation/zoom and structure-visibility controls), whereas the control group received the conventional physical anatomical model used in routine practice. In both arms, recipients were encouraged to ask questions and interact with the tool jointly with the clinician.

Consistent Content

All sessions followed a prespecified, SEGUE-informed communication checklist to standardize the structure and core informational items delivered in both arms. The clinician explained (1) the diagnosis and relevant neuroanatomy; (2) the proposed surgical plan and procedural steps at a high level; (3) expected benefits; (4) major risks and high-risk regions, including the probability and potential consequences of key complications when such estimates were available; (5) alternatives (including conservative management, when applicable); and (6) postoperative expectations. Throughout the session, clinicians were instructed to minimize unnecessary jargon, respond to questions, and confirm understanding using standard comprehension checks (eg, asking the recipient to summarize key points in their own words).

AR Content Creation

A preconstructed digital 3D model was selected and adjusted according to each patient's lesion location before use. Labels and approach- and risk-point annotations were overlaid, and the final model was preloaded and verified on the head-mounted AR device before the communication session.

Neurosurgeons Training (Before Communication)

Personnel Selection

Neurosurgeons were selected from the same institution with identical years of professional experience across all 3 communication levels; postcommunication assessment of neurosurgeons' communication skills was conducted by an independent third-party researcher using the SEGUE scale.

Standardized Communication Training

This comprised the following: development and refinement of the risk-disclosure checklist, familiarization with head-mounted AR device operation, preparation of anticipated responses to questions, and avoidance of "overdisclosure" that may induce fear.

Operational Procedure Checklist

This considered the establishment of a standardized workflow (setup→alignment→demonstration→explanation→questioning→confirmation) to ensure consistency.

Situational Control

This involved the use of equally quiet and independent spaces, adherence to the same time window, and limitation of the number of onlookers.

Communication Time Limit

Communication duration was standardized across groups to minimize confounding from consultation length. Session duration was recorded, and any clinically necessary extensions were documented with justification. When required, additional routine counseling was provided after the study procedure to ensure adequate informed consent.

Independent Design and Demonstration of Digital 3D Models

Considering the characteristics of preoperative risk communication scenarios, this study used the open-source software 3D Slicer (The Slicer Community) to premodel classic lesion locations during digital 3D modeling. To protect patient privacy, deidentification procedures were implemented at the premodeling stage, including a comprehensive review of patient images to exclude any models exhibiting identifiable facial features. Visual inspection by 2 independent neurosurgeons (ZY and ZS) confirmed that none of the included prefabricated models were amenable to facial-feature recognition. As strict precision is not required for risk communication scenarios, the premade models were selected and adjusted according to the patient's 2D imaging data before use (Figure 1A-C). Model adjustments can be made on a computer and quickly transferred to the SeerLens AR head-mounted device, maximizing time saved in preparation before communication.

During the model demonstration phase, the AR device used in this study (SeerLens B50R AR Glasses) was developed and manufactured by a domestic company (Multimedia Appendix 1). During the demonstration, the communication recipient uses the head-mounted AR device to superimpose a virtual projection onto the real patient's head and, in combination with the neurosurgeon's explanation, observes spatial information such as lesion size, depth, location, and adjacency. If necessary, the communication recipient can interact with the virtual model in real time through voice or gestures, under the neurosurgeon's guidance, based on their needs (Figure 1D).

During AR-assisted sessions, the communication recipient viewed the patient-specific 3D anatomical model through a head-mounted AR display. The neurosurgeon did not wear a headset; instead, the neurosurgeon used a computer-based interface synchronized in real time with the recipient's AR view. This workstation allowed the neurosurgeon to manipulate the model (eg, rotation, zoom, translation), adjust visual properties (eg, transparency/occlusion), and emphasize specific structures (eg, isolating the lesion and adjacent anatomy, highlighting/labeling), while monitoring the same model state shown to the recipient. In this way, the neurosurgeon could verify that the explanation referred to the correct virtual anatomical structures and direct attention through synchronized, on-model cues (Figure 1E).

XSL•FO
RenderX

Figure 1. Study flowchart. AR: augmented reality.

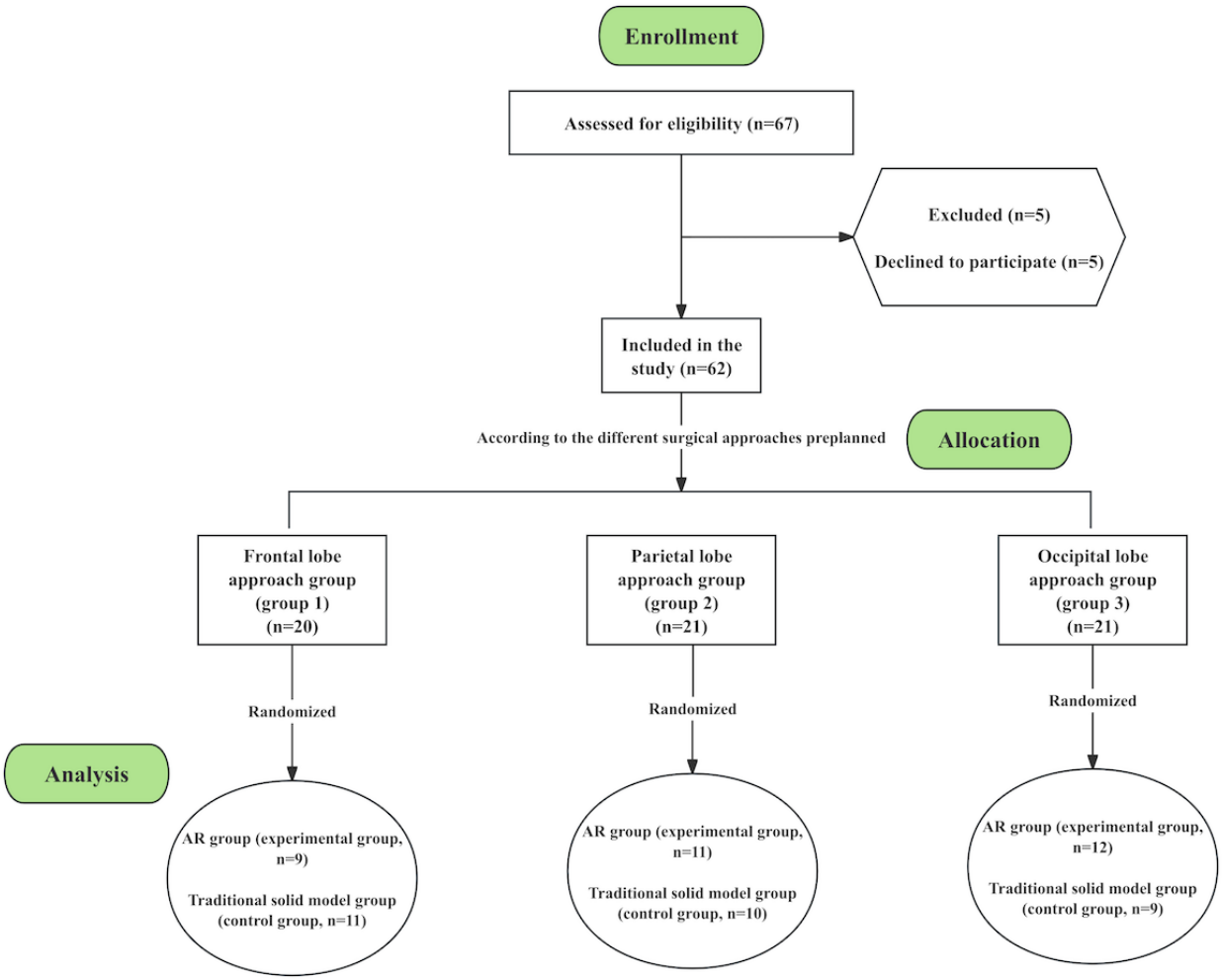

Interventions

Before randomization, eligible patients and their families were informed about the trial. Written informed consent was obtained from the patient when capable; otherwise, consent was provided by the patient's LAR. During this process, 5 eligible patients declined participation. Ultimately, 62 communication recipients (patients or LARs) completed the Chinese version of the STAI-Form Y, which was used to evaluate state anxiety (State Anxiety Inventory, STAI Y-1) and trait anxiety (Trait Anxiety Inventory, STAI Y-2), with higher scores indicating higher levels of anxiety [24]. To evaluate the comprehension abilities of the communication recipients, cognitive function was assessed using the MoCA (Chinese 7.1-Beijing) [25]. Randomization was then performed within 3 strata. In the experimental group (AR group), communication recipients wore the SeerLens AR device when necessary, and the neurosurgeon superimposed the digital 3D model onto the patient's real head, demonstrating the corresponding spatial structures based on the patient's lesion location and explaining them to the communication recipient. Here, "real-time support" denotes the in-session use of the AR overlay as a shared spatial reference aligned to the patient's lesion location, allowing the explanation to progress in step with the conversation and recipients' questions. This wording describes the workflow and does not imply that the AR arm isolates a single mechanism (eg, spatial registration vs enhanced interactivity), nor that any observed effects can be attributed to AR hardware alone. In the control group (physical-model group), the neurosurgeon provided explanations with the assistance of a physical anatomical model. To minimize between-arm differences in informational content, clinicians followed a prespecified checklist of key information items derived from the SEGUE framework. The checklist covered (1) diagnosis and lesion location, (2) relevant neuroanatomical relationships, (3) the proposed treatment plan and surgical corridor/approach, (4) major risks and high-risk regions, (5) alternative options and the expected postoperative course, and (6) an opportunity for questions and teach-back (Multimedia Appendix 2). The checklist was identical for both arms; the only planned difference was the visualization tool used during explanation (AR-based patient-specific 3D model vs conventional physical model). Although informational items were standardized via the checklist, the AR aid inherently differs from the physical model in patient-specific anatomical specificity and software-enabled, dynamically adjustable visualization (eg, structure isolation/toggling and transparency adjustment); therefore, the between-arm contrast should be interpreted as a bundled comparison of these visualization-support features. In both arms, recipients were

XSL•FO
RenderX

encouraged to ask questions and to interact with the tool jointly with the clinician.

After the communication session, the communication recipient completed postsession assessments, including (1) the Medical Knowledge Comprehension Questionnaire, which included a subjective understanding rating scale (Cronbach α=.88; Multimedia Appendix 3) to assess subjective comprehension and an objective understanding assessment (Multimedia Appendix 4): we used the Objective Assessment of Medical Knowledge Comprehension Following Risk Communication, a structured, directed questionnaire covering 4 domains (anatomical localization, treatment surgery, potential risks, and expected benefits), plus an open-ended item, developed to provide a pragmatic, objective quantification of recipients' comprehension of the core information delivered during preoperative risk communication; (2) anxiety assessment (STAI-Form Y, Chinese version: STAI Y-1 and STAI Y-2); and (3) communication satisfaction assessment, using a self-developed rating scale (Cronbach α=.86; Multimedia Appendix 5). All questionnaires were self-administered by the communication recipient without assistance; no investigator administered the objective questionnaire or guided responses. Multiple-choice items were scored using a prespecified answer key.

Communication duration was measured in minutes by the neurosurgeon conducting the session and was defined as the face-to-face risk communication time from the start of the standardized explanation to the completion of the conversation. This metric did not include any presession technical preparation required for the AR workflow, which was completed before the communication session. The entire communication process was video-recorded so that independent researchers could later evaluate neurosurgeons' communication skills using the SEGUE scale (Chinese version) [26,27].

Outcomes

The primary outcome was objective understanding, assessed immediately after the communication session using the directed multiple-choice knowledge questionnaire (Multimedia Appendix 4). The key secondary outcome was subjective understanding, assessed using the subjective understanding rating scale (Multimedia Appendix 3). Other secondary outcomes were communication satisfaction (Multimedia Appendix 5), pre-to-post changes in anxiety (STAI Y-1 and STAI Y-2), communication duration, and neurosurgeons' communication skills, assessed using the SEGUE scale based on video recordings.

Sample Size

We considered a 30% difference between the control and experimental groups to be clinically significant, based on the score range of the primary outcome measure (knowledge questionnaire scores). We assumed an expected effect size of d=1.50 (preset coefficient of variation=0.2), α=.05, and power (1–β)=0.85. The sample size was computed using G*Power (version 3.1.9.7). The required sample size was 10 participants per arm (20 total) for a single comparison. As we prespecified 3 anatomical strata and planned stratified analyses, we targeted approximately 20 participants per stratum (≈10 per arm), corresponding to a minimum total sample size of 60 across all strata (Figure 2A shows the per-stratum calculation). Although this sample size meets statistical requirements, as a single-center PoC study, it may still be constrained by the actual case volume. The primary analysis was conducted on the pooled cohort; powering per stratum was used as a conservative design choice to ensure adequate precision for the prespecified stratified analyses.

XSL•FO
RenderX

Figure 2. Schematic overview of the standardized clinician-recipient communication workflow under the SEGUE framework in routine clinical practice. (A) Digital 3D reconstruction derived from the patient's imaging data. (B and C) The communication recipient observes the virtual model from multiple perspectives according to their level of understanding and informational needs. (D) Illustration of the augmented reality (AR)–assisted communication approach in a clinical setting. The recipient views the virtual 3D model through a head-mounted AR device, while the physician does not wear a synchronized headset. Consequently, any physical pointing toward the recipient's display is intended for approximate localization and illustrative purposes only, rather than precise identification of virtual structures. Visually aligned shared spatial reference and joint attention are achieved through a real-time synchronized computer interface, which allows the physician to manipulate the model and align explanations with the structures viewed by the recipient. (E) System architecture and data flow of the AR-assisted risk communication platform, comprising a synchronized computer control terminal and a head-mounted AR display terminal. TCP: terminal control platform.

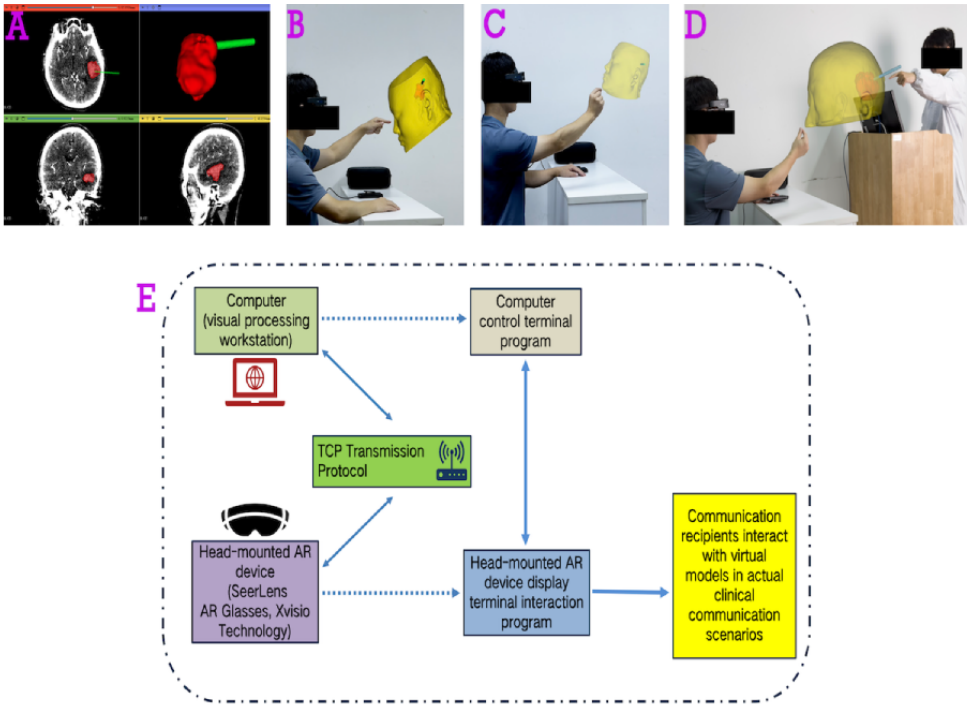

Statistical Analysis

Data Summary Methods

Analyses were conducted in GraphPad Prism 9 (GraphPad Software). Continuous variables were summarized as mean (SD) when approximately normally distributed and as median with IQR (Q3-Q1) otherwise. Normality was assessed using the D'Agostino-Pearson test. Unless otherwise stated, tests were 2-sided. For between-group comparisons, Student *t* test was used when normality was not rejected and the Mann-Whitney *U* test otherwise.

Analytical Hierarchy and Multiple Comparisons

As this was a PoC study with multiple outcomes and planned stratum-specific comparisons, we prespecified an analytical hierarchy to reduce inflation of type I error from multiple comparisons.

Primary Confirmatory Analysis

The single primary confirmatory analysis (2-sided α=.05) was the pooled comparison of objective understanding scores between the AR and physical-model groups (groups 1A+2A+3A vs 1B+2B+3B).

Key Secondary Analyses (Exploratory)

Pooled comparisons of subjective understanding, communication satisfaction, anxiety changes (ΔSTAI Y-1 and ΔSTAI Y-2), communication duration, and SEGUE scores between the AR and physical-model groups were prespecified as exploratory secondary analyses. *P* values are reported without multiplicity adjustment and are interpreted descriptively as signals for future investigation rather than confirmatory evidence.

Stratified Exploratory Analyses

Within-stratum comparisons (1A vs 1B, 2A vs 2B, and 3A vs 3B) were conducted to explore potential heterogeneity by surgical approach. Given the small stratum-specific sample sizes, these analyses are hypothesis-generating only and should be interpreted with caution.

Recipient-Type Robustness Analyses

As the communication recipient could be either the patient or an LAR, we assessed whether recipient-type composition could confound communication outcomes. Recipient-type distribution (patient vs LAR), recipient sex, and education (highest educational attainment) were compared between arms using the Fisher exact test. We then conducted recipient-type robustness analyses for the primary and key secondary outcomes by reporting arm comparisons stratified by recipient type and by fitting linear models that included treatment group, recipient type, and a group × recipient-type interaction term. Sensitivity models adjusting for recipient type were used to evaluate whether the main findings were robust to recipient-type composition.

XSL•FO
RenderX

Ethical Considerations

Study Approval

The study was approved by the Ethics Committee of the First Affiliated Hospital of Xiamen University, Xiamen, Fujian, China, in October 2024 (approval number 169). This trial was retrospectively registered with the ISRCTN registry (ISRCTN11483487; February 13, 2026). The first participant was enrolled in January 2025, and trial registration was completed on February 13, 2026. Registration was delayed due to the time required to complete and correct the registration materials. At study initiation, this investigator-initiated PoC risk-communication trial—embedded in routine preoperative SDM workflows and not involving any experimental clinical treatment—was mistakenly classified within our institutional governance process as a low-risk communication workflow/technology-feasibility study rather than as a registrable randomized clinical trial, despite the use of randomization. This was an administrative error and not an intentional omission.

To address potential reporting bias, the protocol originally reviewed and approved by the Ethics Committee in October 2024 (approval number 169) prespecified the eligibility criteria, randomization strategy, intervention procedures, primary outcome, key secondary outcomes, and analysis framework. A written statement issued by the Ethics Committee of the First Affiliated Hospital of Xiamen University, dated February 27, 2026, confirmed that the protocol registered in the ISRCTN registry remains consistent with the version originally reviewed and approved by the committee.

Written informed consent was obtained from all participants (patients or LARs), covering participation, the use of deidentified imaging data to generate 3D models for AR-assisted counseling, video recording of the consultation for SEGUE-based assessment, and completion of study questionnaires and assessments, in accordance with the Declaration of Helsinki (1964) and its subsequent amendments or comparable ethical standards.

Informed Consent

Written informed consent was obtained from all enrolled communication recipients before participation. When the communication recipient was an LAR, the LAR provided written consent; where feasible, patient assent was also sought. Participants were informed that participation was voluntary and that they could withdraw at any time without any impact on clinical care. The informed consent process covered completion of study questionnaires and the research use of relevant clinical information (eg, imaging data used to generate patient-specific 3D models and video recording of the communication session for SEGUE scoring).

Privacy and Confidentiality

All study data were deidentified and labeled using a unique study code. Any linkage file containing direct identifiers was stored separately from the analytic dataset and was accessible only to authorized study personnel. Data were analyzed and reported in aggregate to prevent the identification of individual participants.

Compensation

Participants did not receive any compensation for completing the study procedures.

Identifiable Information in Images/Multimedia Appendices

No images in the manuscript or multimedia appendices include identifiable information about individual participants (eg, faces, names, medical record numbers, or other direct identifiers). Photographs used to illustrate the AR-assisted communication workflow were taken with nonpatient volunteers (eg, with full-face blurring or cropping) and contain no patient-identifiable information.

Patient or Public Contribution

Patients, caregivers, or members of the public did not serve as partners in the design, conduct, analysis, or reporting of this PoC randomized trial. Recruitment and communication occurred in time-sensitive preoperative settings, and the study focused on feasibility and measure refinement; therefore, establishing and supporting a Patient and Public Involvement and Engagement advisory group was not feasible at this stage.

Results

Baseline Characteristics

All 62 enrolled communication recipients completed the risk communication, with no dropouts. The primary communication recipient (who completed all study assessments) was the patient in 30 cases and an LAR in 32 cases (Table 1). Specifically, there were 20 participants in the frontal approach cohort (group 1), 21 in the parietal approach cohort (group 2), and 21 in the occipital approach cohort (group 3). Within the frontal approach group, the AR group (group 1A, experimental) comprised 9 patients, and the physical-model group (group 1B, control) comprised 11. Within the parietal approach group, the AR group (group 2A, experimental) included 11 patients, and the physical-model group (group 2B, control) included 10. Within the occipital approach group, the AR group (group 3A, experimental) included 12 patients, and the physical-model group (group 3B, control) included 9 (Figure 3 and see Multimedia Appendix 6 for the CONSORT [Consolidated Standards of Reporting Trials] checklist).

After assessing normality (Figure 2B), we processed the data according to their underlying distributions. No baseline differences in demographics were observed among communication recipients across groups (Table 1). Before risk communication, there were no statistically significant pooled between-arm differences in baseline anxiety (STAI Y-1, P=.16; STAI Y-2, P=.67) or cognitive function (MoCA score, P=.06). Similarly, no statistically significant within-stratum baseline differences were observed (Table 1; Figure 4).

XSL•FO
RenderX

Figure 3. Sample size determination and statistical analysis preparation. (A) Results of the sample size calculation performed using G*Power (version 3.1.9.7). (B) Assessment of normality for baseline characteristics of all participants using the D'Agostino-Pearson test. MoCA: Montreal Cognitive Assessment; STAI: State-Trait Anxiety Inventory.

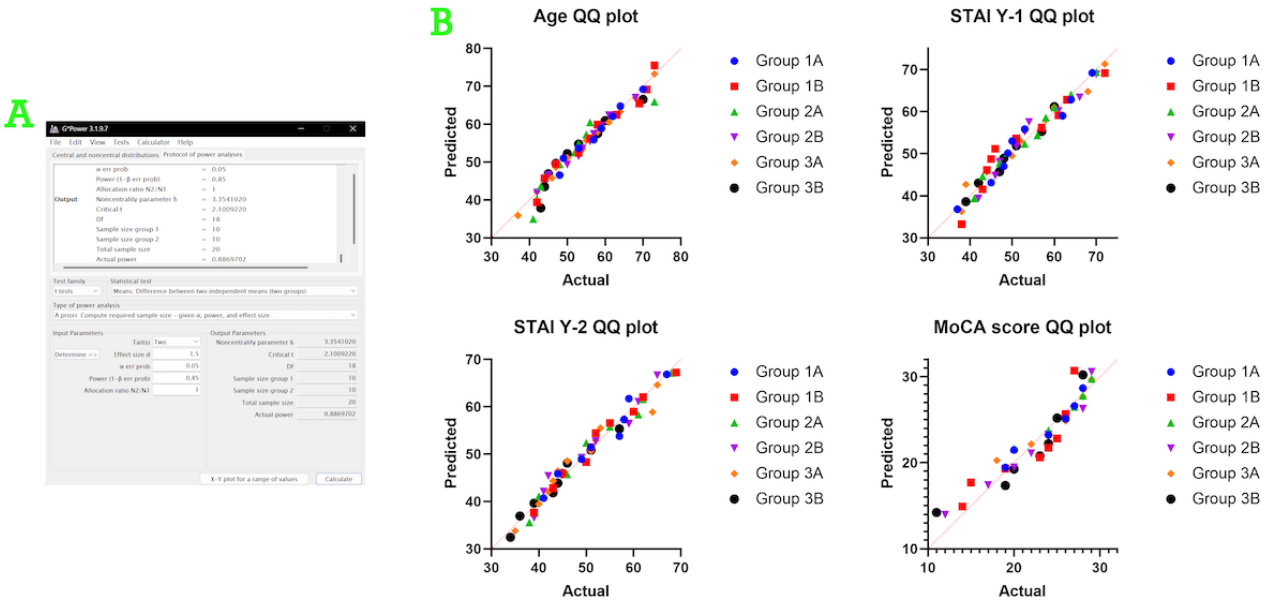

XSL•FO
RenderX

Figure 4. Visualization of participants' baseline characteristics. MoCA: Montreal Cognitive Assessment; STAI: State-Trait Anxiety Inventory.

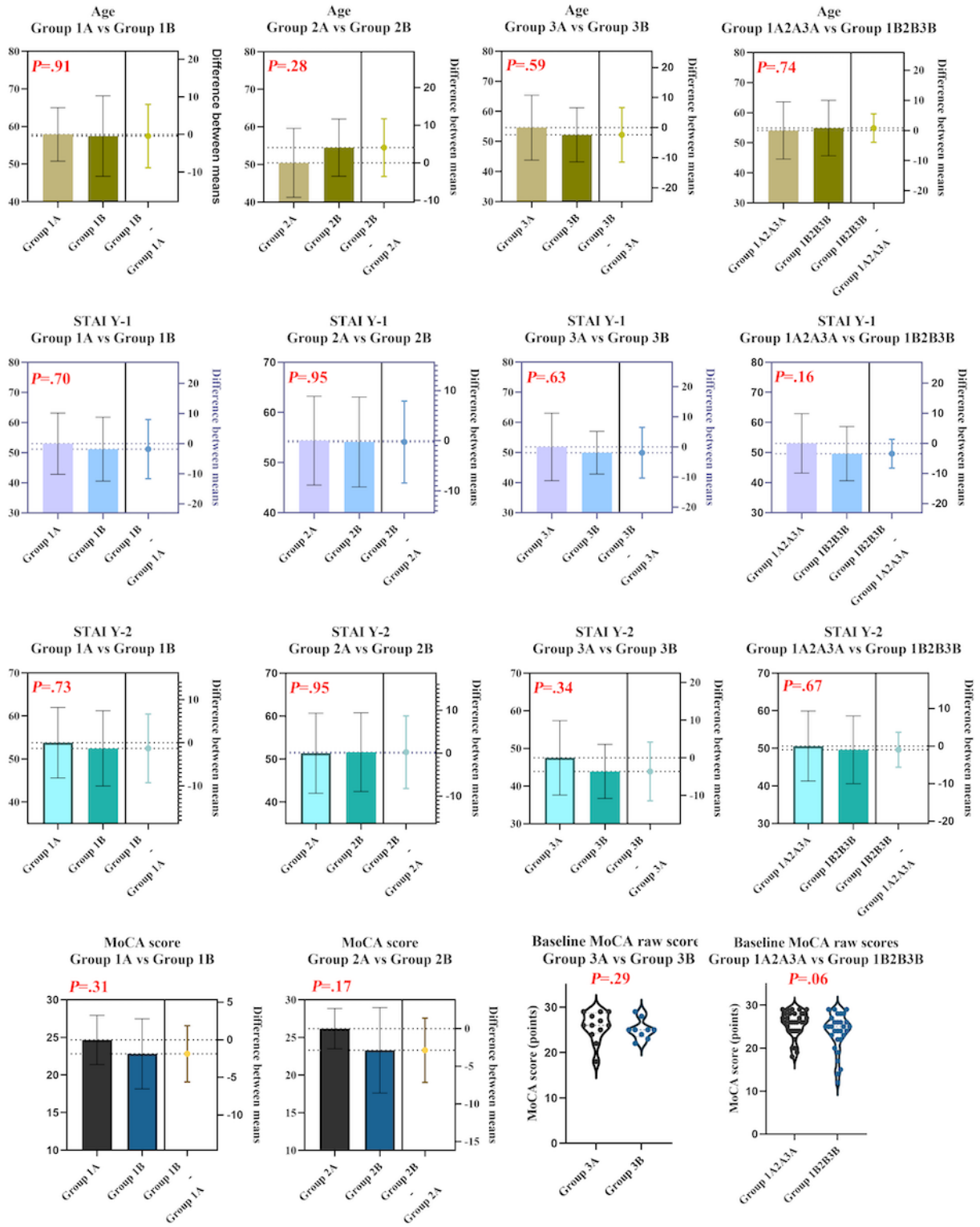

Primary Outcome

Primary Confirmatory Analysis

The primary outcome (objective understanding) was measured using a 20-item multiple-choice questionnaire developed for this trial to operationalize the core informational elements delivered during the standardized, SEGUE-informed risk communication. Items were mapped to 4 predefined domains: anatomical localization (4 items), treatment/surgical plan (5 items), risks/complications (5 items), and expected benefits/expectations (6 items). Each item was scored 0-1 (correct/incorrect), with partial credit (0.5) awarded when a response reflected a partially correct understanding according to a prespecified rubric; the open-ended item was not included in the total score (range 0-20; Multimedia Appendix 4). In the prespecified pooled comparison, objective understanding scores were higher in the AR group than in the physical-model group (groups 1A+2A+3A vs 1B+2B+3B; *P*=.01; Table 2; Figure 5).

XSL•FO
RenderX

Table 2. The summary of research outcomes.[a]

Research outcomes	Total	Group 1 (n=20)		Group 2 (n=21)		Group 3 (n=21)		*P* values
		Group 1A (n=9)	Group 1B (n=11)	Group 2A (n=11)	Group 2B (n=10)	Group 3A (n=12)	Group 3B (n=9)	
Subjective understanding score, mean (SD) (range 8-40)		26.22 (6.28)	27.09 (4.68)	22 (17)[b]	21.00 (8.91)	24.50 (9.35)	21.78 (8.69)	.74 (group 1A vs 1B); .30 (group 2A vs 2B); .50 (group 3A vs 3B); and .41 (group 1A2A3A vs 1B2B3B)
Objective understanding score, mean (SD)		13.89 (5.02)	12.73 (3.45)	13.28 (4.57)	9.40 (1.47)	13.25 (4.19)	9.28 (3.64)	.57 (group 1A vs 1B); .02[c] (group 2A vs 2B); .03[c] (group 3A vs. 3B); and .01[c] (group 1A2A3A vs 1B2B3B)
STAI[d] -Form Y-Chinese version, mean (SD)								
STAI Y-1′	48.98 (10.08)	47.56 (9.85)	49.00 (13.32)	50.00 (9.02)	53.90 (10.45)	46.83 (9.56)	46.56 (7.65)	N/A[e]
STAI Y-2′	50.52 (9.09)	52.44 (10.26)	52.09 (9.44)	52.73 (7.95)	50.90 (9.76)	49.08 (9.31)	45.44 (7.81)	N/A
ΔSTAI Y-1	N/A	5.44 (7.35)	–3 (11)[b]	4.36 (8.86)	0.2 (13.52)	5.00 (19.32)	3.33 (10.11)	.37 (group 1A2A3A vs 1B2B3B)
ΔSTAI Y-2	N/A	1.33 (12.50)	0.36 (10.87)	–1.36 (9.83)	0.7 (9.50)	–1.58 (15.63)	–1.56 (11.68)	.84 (group 1A2A3A vs 1B2B3B)
Communication satisfaction score (range 10-50)	N/A	40.67 (6.00)	32.18 (7.92)	35.91 (9.66)	24.40 (12.18)	36.17 (8.29)	27.78 (8.91)	.01[c] (group 1A vs 1B); .03[c] (group 2A vs 2B); .04[c] (group 3A vs 3B); and <.001[c] (group 1A2A3A vs 1B2B3B)
Duration of communication (in-session) (minutes)	N/A	29.67 (7.87)	28.64 (6.79	28.27 (10.06)	27.90 (8.90)	26.67 (8.36)	25.11 (8.19)	.76 (group 1A vs 1B); .93 (group 2A vs 2B); .67 (group 3A vs 3B); and .73 (group 1A2A3A vs 1B2B3B)
SEGUE[f] score	N/A	22.22 (1.20)	21.91 (1.45)	21.64 (1.36)	20.10 (2.13)	20.75 (1.71)	21.67 (1.80)	.60 (group 1A vs 1B); .07 (group 2A vs 2B); .26 (group 3A vs 3B); and .60 (group 1A2A3A vs 1B2B3B)

[a]All outcomes refer to the communication recipient (patient or legally authorized representative), unless otherwise stated.

[b]Reported as median (IQR) because the corresponding subgroup data were nonnormally distributed. The pooled comparison of objective understanding scores between the augmented reality and physical-model groups (groups 1A+2A+3A vs 1B+2B+3B) was prespecified as the single primary confirmatory analysis (2-sided α=.05). All other *P* values (stratum-specific comparisons and all secondary outcomes) are exploratory, unadjusted for multiple comparisons, and should be interpreted descriptively.

[c]*P* values indicate statistical significance.

[d]STAI: State Anxiety Inventory (STAI Y-1′ and STAI Y-2′ denote postsession scores). ΔSTAI (ΔSTAI Y-1 and ΔSTAI Y-2) was calculated as presession STAI minus postsession STAI (Δ=pre − post).

[e]N/A: not applicable.

[f]SEGUE*:* set the stage, *e*licit information, *g*ive information, *u*nderstand the patient's perspective, and *e*nd the encounter.

XSL•FO
RenderX

Figure 5. Visualization of participants' primary outcomes (objective comprehension scores).

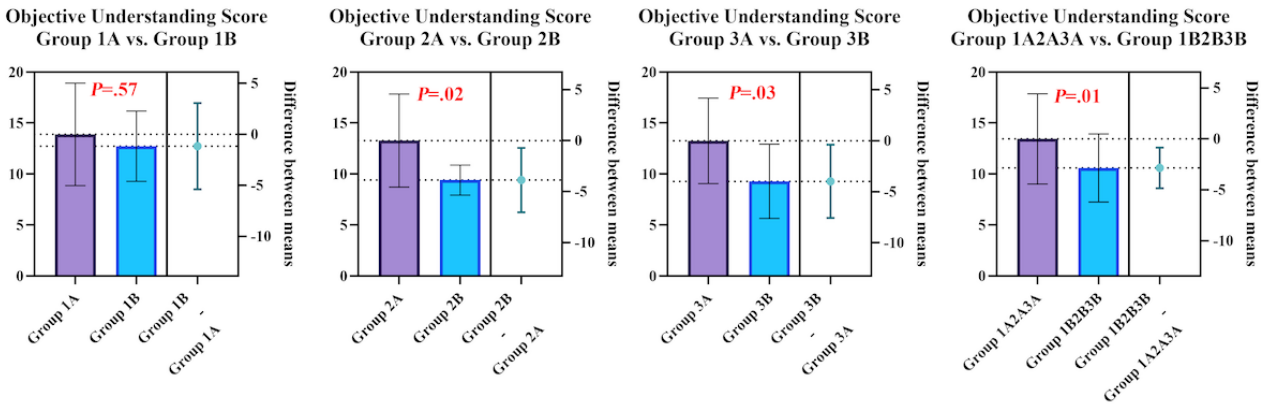

Stratified Exploratory Analyses

Within strata, objective understanding was higher in the AR group in the parietal approach stratum (group 2A vs 2B; *P*=.02) and in the occipital approach stratum (group 3A vs 3B; *P*=.03), but not in the frontal approach stratum (group 1A vs 1B; *P*=.57). These stratum-specific comparisons are underpowered and should be interpreted as hypothesis-generating (Table 2; Figure 5).

Secondary Outcomes

Overview

All secondary outcome analyses were exploratory and were not adjusted for multiple comparisons; *P* values are interpreted descriptively.

Subjective Understanding (Exploratory)

Subjective understanding scores did not differ between the AR and physical-model groups in the pooled comparison (*P*=.41) or within strata (*P*=.74 for group 1, *P*=.30 for group 2, and *P*=.50 for group 3; Table 2; Figure 6A).

Figure 6. Visualization of secondary outcomes for all participants. (A) Subjective comprehension scores among communication recipients. (B) Changes in communication recipients' anxiety levels before and after the interaction. (C) Subjective satisfaction scores among communication recipients.

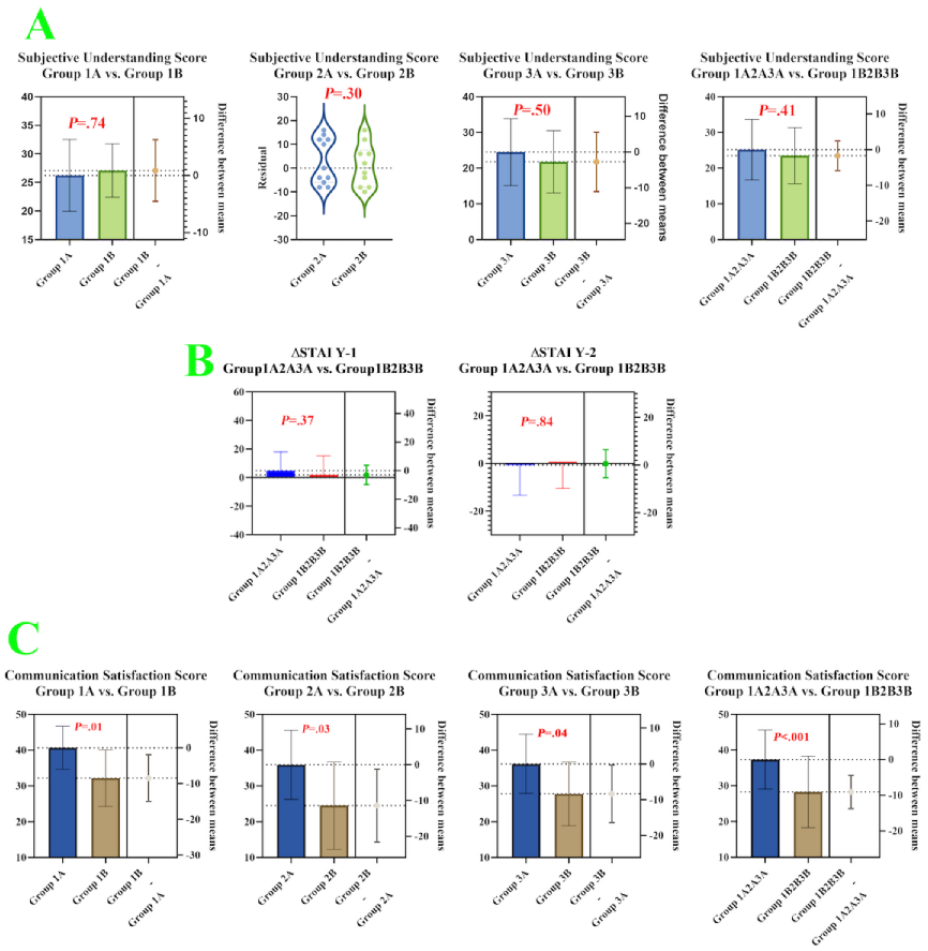

XSL•FO
RenderX

Anxiety (Exploratory)

Notably, improvements in recipients' anxiety did not differ meaningfully across communication modalities:

- STAI Y-1: for state anxiety, defined as momentary anxious reactivity, the pre-to-post change did not differ significantly between modalities (*P*=.37; groups 1A+2A+3A vs 1B+2B+3B).
- STAI Y-2: for trait anxiety, reflecting the general tendency to experience anxiety, the change from precommunication to postcommunication did not differ significantly (*P*=.84; groups 1A+2A+3A vs 1B+2B+3B; Table 2; Figure 6B).

Standardization-Related Outcomes

Communication Duration

Differences in communication time between groups were not statistically significant (*P*=.76: group 1A vs 1B; *P*=.93: group 2A vs 2B; *P*=.67: group 3A vs 3B; *P*=.73: groups 1A+2A+3A vs 1B+2B+3B).

SEGUE Score

There were no significant differences in the interviewing neurosurgeons' communication skills across groups (*P*=.60: group 1A vs group 1B; *P*=.07: group 2A vs group 2B; *P*=.26: group 3A vs group 3B; *P*=.60: groups 1A+2A+3A vs groups 1B+2B+3B; Table 2).

Communication Satisfaction (Exploratory)

Regarding subjective satisfaction, scores were higher in the AR group across all strata (*P*=.01: group 1A vs group 1B; *P*=.03: group 2A vs group 2B; *P*=.04: group 3A vs group 3B). The AR group's communication satisfaction scores were also higher than those of the physical-model group in the pooled comparison (*P*<.001: groups 1A+2A+3A vs groups 1B+2B+3B; Table 2; Figure 6C).

Recipient-Type Distribution and Robustness Analyses

Recipient-type distribution was similar between groups (AR: 17/32, 53%, patients and 15/32, 47%, LARs; physical model: 13/30, 43%, patients and 17/30, 57%, LARs; Fisher exact test *P*=.46; Table 1). Recipient-type robustness analyses did not suggest that anxiety or satisfaction outcomes were driven by recipient-type composition. Interaction tests were not statistically significant for subjective understanding, satisfaction, or anxiety outcomes ($P_{interaction}$=.22 for subjective understanding; $P_{interaction}$=.07 for communication satisfaction; $P_{interaction}$=.71 for ΔSTAI Y-1; and $P_{interaction}$=.93 for ΔSTAI Y-2). However, for communication satisfaction, the group × recipient-type interaction approached conventional significance ($P_{interaction}$=.07), with a substantially larger absolute AR – control difference among LARs (37.40 vs 24.71; Δ≈12.7 points) than among patients (37.29 vs 32.92; Δ≈4.4 points; Table 3).

Adjusting for recipient type did not change the overall conclusions (objective understanding *P*=.006). Robustness results are presented in Table 3.

Table 3. Recipient-type subgroup and sensitivity analyses (patients vs LAR[a]) for postcommunication outcomes.

Outcome	Patients' AR[b] (n=17), mean (SD)	Patients control (n=13), mean (SD)	*P* value (patients)	LARs' AR (n=15), mean (SD)	LARs control (n=17), mean (SD)	*P* value (LARs)	*P* value interaction (group × recipient type)	*P* value adjusted
Objective understanding score	12.47 (5.03)	11.38 (2.97)	.47	14.53 (3.45)	9.97 (3.57)	<.001[c]	.09	.006[c]
Subjective understanding score	24.71 (8.51)	25.85 (7.14)	.69	25.73 (8.68)	21.65 (8.01)	.18	.22	.45
Communication satisfaction score	37.29 (7.65)	32.92 (8.63)	.16	37.40 (9.18)[d]	24.71 (9.77)[d]	<.001[c]	.07	<.001[c]
ΔSTAI[e] Y-1	6.59 (15.12)	2.46 (13.06)	.43	3.00 (10.58)	1.41 (13.99)	.72	.71	.41
ΔSTAI Y-2	–1.06 (14.48)	–0.23 (10.34)	.86	–0.27 (10.67)	0.00 (10.69)	.94	.93	.86

[a]LAR: legally authorized representative.

[b]AR: augmented reality.

[c]*P* values show statistical significance.

[d]Reported as median (IQR) because the corresponding subgroup data were nonnormally distributed. *P* adjusted indicates the treatment-group effect from the sensitivity model adjusted for recipient type.

[e]STAI: State Anxiety Inventory.

XSL•FO
RenderX

Discussion

Principal Findings

In the prespecified pooled primary confirmatory analysis, AR-assisted preoperative risk communication in neurosurgery—compared with conventional aids centered on physical models—significantly improved recipients' objective understanding of the current condition and surgery-related information. Although gains in self-perceived understanding were modest, recipients reported high subjective satisfaction with the communication. These effects are consistent with AR's intuitive, 3D, and comprehensive spatial visualization. In addition, user-friendly head-mounted AR devices support real-time human-computer interaction, which may further enhance both subjective satisfaction with the process and objective understanding of the content. This pattern aligns with prior work showing that digital 3D planning tools improve the quality of surgeon-patient communication during informed consent and objectively promote patients' understanding of their condition [4,19].

Notably, objective understanding did not differ significantly between groups in the frontal lobe approach subgroup (P=.57). This null finding should be interpreted cautiously. First, our a priori sample size calculation was based on the primary overall comparison and assumed a very large effect size (Cohen d=1.50). The frontal subgroup included only 9 and 11 participants and was therefore likely underpowered to detect moderate effects; the nonsignificant result may represent a type II error rather than evidence of no effect. Second, a clinically plausible explanation is that frontal approaches may involve relatively intuitive anatomical landmarks and a more straightforward surgical corridor; in such situations, a generic 3D physical anatomical model may already provide sufficient spatial cues for lay recipients, leaving limited room for additional improvement with AR (ie, a potential ceiling effect). By contrast, when lesions are deeper, risk-relevant structures are less visually accessible, or the planned corridor is more circuitous, AR may offer greater incremental value through multiangle exploration and dynamic, structure-specific highlighting and annotation. Taken together, the location/approach subgroup findings (including both significant and nonsignificant results) should be considered exploratory and hypothesis-generating.

Moreover, the control group demonstrated relatively low objective understanding scores (mean 9.40 and 9.28 out of 20 for the parietal and occipital regions, respectively), raising the question of whether physical model–based explanations are sufficiently effective for communicating complex surgical concepts. This low baseline performance may reflect the inherent challenges of explaining intricate surgical procedures using static models, or it may indicate that the standard physical model fails to adequately convey 3D anatomical relationships. This finding highlights the potential benefit of AR-assisted visualizations, which may better engage patients and provide more intuitive, spatially accurate representations.

Our study design could not fully disentangle the effect of personalization from that of the immersive technology. Specifically, the AR arm provided a patient-tailored 3D visualization based on the communication recipient's routine imaging and adjusted to the lesion location, whereas the control arm used a generic physical anatomical model; therefore, the observed improvement in objective understanding may partly reflect content specificity rather than the AR delivery method itself. We did not include a patient-specific 2D imaging arm because this PoC trial was designed as a pragmatic study aligned with routine clinical practice in our mainland China setting, where generic physical anatomical models are among the most commonly used adjuncts for anatomy teaching and preoperative communication in many surgical departments. Moreover, in real preoperative neurosurgical risk communication, patient-specific 2D images alone are often difficult for nonexperts to interpret and may be insufficient to convey complex 3D relationships critical for risk discussion (eg, lesion depth, adjacency to eloquent regions, and surgical corridors). Finally, it is important to acknowledge that the use of AR in preoperative neurosurgical risk communication inherently combines 2 inseparable features—patient-specific personalization and an immersive, interactive experience. In real-world applications, these attributes do not operate in isolation but function as an integrated whole, jointly shaping communication outcomes. We therefore believe that the improved objective understanding of the surgical plan and related risks among communication recipients reflects the synergistic effects of personalization and immersive interaction. In addition, because head-mounted spatial presentation (in situ overlay) and enhanced interactivity (dynamic view manipulation) co-occurred in the AR workflow, this PoC design was not intended to determine which of these components primarily drove the observed differences.

We emphasize that AR's 3D, immersive, and interactive features provide spatial anchors for apprehending complex intracranial anatomy and surgical pathways. In this study, "enhanced interactivity" refers to tool affordances that allow rapid switching between views (eg, rotation/zoom, structure toggling, transparency adjustment, and clipping planes) during joint clinician-recipient exploration of the 3D model; physical models also support interaction (eg, holding, rotating, and pointing), but with fewer view-manipulation affordances. In addition to conventional verbal explanations and physical anatomical models, AR head-mounted displays can provide an optional patient-specific 3D visualization. During the session, the clinician and the communication recipient can jointly orient to key structures, the target lesion, and clinically relevant risk regions using the same visual reference.

By rendering abstract spatial information visible and manipulable, AR affords an immersive, participatory experience that improves objective understanding and increases satisfaction. It is worth noting that although the mean changes in state anxiety (ΔSTAI Y-1) did not differ significantly between groups, the data revealed high variance across all strata (eg, SDs ranging from 7.35 to 19.32 in AR groups and reaching 13.52 in the parietal control group). This substantial variability suggests that high-fidelity visualization—whether via AR or physical models—may exert a "polarizing effect" on recipients' emotional states. This interpretation is consistent with recent

XSL•FO
RenderX

meta-analytic evidence indicating that, while virtual modalities consistently improve understanding, their impact on anxiety remains variable and nonsignificant compared with standard care [28]. For some individuals, the "unveiling" of the lesion's precise location reduces the fear of the unknown, providing reassurance. Conversely, for others, the vivid and realistic depiction of the pathology may overwhelm their coping mechanisms. This aligns with modern psychological frameworks on "information avoidance," which posit that detailed disclosure of threatening health information can paradoxically heighten distress in individuals prone to blunting coping styles, despite the objective gain in knowledge [29]. Therefore, the nonsignificant difference in mean anxiety scores likely reflects these opposing reactions canceling each other out. Meanwhile, self-rated understanding is strongly influenced by prior knowledge, expectations, and emotional state; over short intervals, it may not rise in parallel with objective measures, a finding consistent with our results [30]. These observations suggest that a single preoperative discussion is a necessary component of SDM but not a complete intervention chain. A comprehensive SDM process should integrate emotional support, psychoeducation, and expectation management into multi–time point, multichannel, continuous communication, coupled with counseling and follow-up across the preoperative-intraoperative-postoperative continuum. Within this chain, AR should function not merely as a display tool but as a vehicle for preference clarification and value prioritization. Integrating humanistic care with emerging assistive technologies is a likely future direction for clinician-patient communication.

Finally, there is another interesting result in this study. Our recipient-type robustness analyses also suggest a potentially meaningful trend in communication satisfaction. Although the group × recipient-type interaction did not reach the conventional threshold for statistical significance ($P_{interaction}$=.07), the magnitude of the AR-associated improvement was substantially larger among LARs (≈12.7-point higher mean satisfaction vs control) than among patients (≈4.4-point higher vs control; Table 3). This pattern is hypothesis-generating, and the present PoC trial was not powered to test effect modification by recipient type; nevertheless, it is clinically plausible that surrogate decision makers may derive greater reassurance and perceived communication quality from patient-specific, manipulable visualization that clarifies anatomical localization and risk-relevant structures in real time. Future multicenter trials with adequate power should prespecify and evaluate recipient-type subgroup effects on satisfaction and related decisional outcomes.

Comparison With Prior Work

During surgical consultations, clinician-patient dialogue often fails to establish a shared understanding of the anticipated care experience or alignment with patients' overarching health goals [31,32]. Although surgeons must provide information to support autonomous decision-making, both surgeons and patients may treat informed consent as perfunctory. With its emphasis on 1-way information flow, the process affords limited real-time feedback from patients and often does not prompt reflection on the personal value of surgery or preparation for potential complications. Medicolegal reviews indicate that failure to explain surgical risks and adverse events is the most frequent allegation raised by plaintiffs [33]. SDM is designed to align therapeutic choices with patients' objectives and, in principle, addresses this deficit [34]. Clinically, SDM commonly involves the neurosurgeon outlining to the patient or LAR the pros and cons of 2 or more therapeutic alternatives [35]. In this process, the mode of communication largely determines effectiveness; the judicious use of assistive technologies in preoperative risk communication can increase patient engagement in decision-making and strengthen both risk communication and SDM. At present, many artificial intelligence–designed decision-support tools (eg, decision aids and question-prompt lists) emphasize information analytics and behavioral prediction while paying relatively little attention to the bidirectional interactions needed to elicit, refine, and apply patient preferences or to support the communication process itself [36,37].

AR can help address these shortcomings by three-dimensionally visualizing personalized virtual models and overlaying them onto the physical clinical environment. Head-mounted AR devices allow real-time interaction via gestures or voice commands, enabling clinicians and patients to bridge the cognitive gap created by abstract spatial concepts. Under illness-related vulnerability and persistent information asymmetry, merely presenting options and describing their pros and cons does little to reduce information and power differentials—particularly in highly anxious contexts [38]. By overlaying virtual models onto the real-world setting, AR gives recipients access to otherwise latent anatomical relationships, substantially improving spatial understanding and the accuracy of risk localization. This, in turn, narrows—at least in the short term—the knowledge gap between clinicians and patients, as confirmed in our study.

Within the SEGUE framework, we minimized confounding from within-group variation in neurosurgeons' communication competence and quantified skills using SEGUE scores to ensure appropriate assessment and control. This exploratory trial focused on the acceptability of AR assistance in real-world preoperative neurosurgical communication, the feasibility of implementation, and preliminary effects. The findings can guide industry partners in pursuing targeted product refinements for specific clinical scenarios and provide a practical foundation for improving models of neurosurgical preoperative risk communication.

From a clinical perspective, our results suggest that AR-assisted, patient-specific visualization may be a useful adjunct to standard preoperative risk and treatment plan communication, particularly when spatial understanding of neuroanatomy and surgery-related risk regions is critical. Importantly, AR should be viewed as a communication aid rather than a substitute for clinician communication skills; structured, empathic, and comprehensible explanations remain essential regardless of the visualization modality. At present, domestic head-mounted AR devices have not achieved full mass production, and versions for medical scenarios remain under continuous optimization. Critical components, including device reliability, hospital information system interfacing, data security, and legal scrutiny, are still evolving, which objectively limits short-term adoption. Nevertheless, with the maturation of the supply chain and

XSL•FO
RenderX

software ecosystem, and with declining device costs, domestic AR devices are likely to be used more frequently in local health care contexts, enabling more targeted services and greater benefit for patients.

Implementation considerations such as usability, staff training burden, workflow integration, and privacy/data security compliance are important for real-world deployment; however, these factors (including costs and cost-effectiveness) were not evaluated in the present PoC trial. Therefore, our findings should not be interpreted as evidence that AR is ready for widespread adoption; rather, they motivate larger, multicenter, pragmatic studies, together with implementation and economic evaluations, to clarify feasibility and value across diverse health care settings.

Limitations

Our study has several limitations. First, this was a single-center study with a modest sample size, which may limit generalizability. Multiple comparisons are an important limitation of this PoC study. We prespecified a single primary confirmatory comparison (pooled objective understanding) and treated all other comparisons as exploratory and unadjusted; nominally significant exploratory findings should be interpreted cautiously and confirmed in adequately powered trials with prespecified multiplicity-control strategies. Second, clinicians and participants could not be blinded to the visualization modality, introducing potential performance and novelty effects. Although SEGUE scoring was conducted by independent assessors who were not informed of group allocation, complete blinding may have been compromised because the AR headset could be visible in the video recordings; therefore, observer bias in SEGUE ratings cannot be excluded. Third, the costs of learning and process change deserve attention. Despite no significant difference in communication time between the AR and traditional model groups, widespread routine use of AR as a communication aid will require clinicians to spend time beforehand learning and adapting to the associated hardware and software processes. To maximize personalized and targeted communication effects, clinicians also need additional time for preparation before the conversation. We recognize that, for neurosurgical services with heavy workloads, such upfront commitments can create implementation challenges. Even so, additional time ought to be considered a normal element of precommunication preparation, because preoperative rehearsal is beneficial for surgeons, while its effect on care cadence should be balanced through improved processes and a clear division of labor. Furthermore, although we standardized the core informational items delivered in both arms, AR enables dynamic view manipulation and patient-specific visualization, which may alter the salience and perceived amount of information (ie, “effective information load”). Therefore, the observed improvement in objective understanding may reflect not only the technological modality itself but also the integrated features of an AR-assisted visualization package (patient-specific reconstruction and adjustable views). Future studies should include component-isolating comparators (eg, patient-specific 3D visualization delivered on a conventional 2D display, or generic 3D content delivered via AR with standardized interactivity levels) or factorial designs to disentangle personalization, spatial presentation, and interactivity. Fourth, participant heterogeneity should be noted. In keeping with real-world neurosurgical consent processes, some assessments were completed by LARs rather than patients. Although our intervention targets the communication recipient and preoperative decision maker, anxiety and comprehension may differ between patients and proxies. This PoC study was not powered to formally compare effects across recipient types.

Conclusions

In this PoC randomized trial, AR-assisted preoperative risk communication was associated with higher objective understanding of surgery-related information and greater communication satisfaction than communication supported by a conventional physical anatomical model. As the AR arm inherently bundled patient-specific anatomical content, head-mounted spatial presentation, and dynamic view manipulation, these findings should be interpreted as evidence for an integrated visualization-support package rather than for AR hardware alone. AR’s interactive, 3D visualizations provide a more intuitive representation of complex neuroanatomical structures, improving spatial comprehension and enhancing communication satisfaction. While AR did not significantly reduce recipients’ anxiety compared with the physical model, it facilitated a more engaging and participatory experience, aligning with the principles of SDM. Within a standardized SEGUE-informed communication protocol, AR-assisted, real-time, patient-specific visualization may serve as a useful adjunct to SDM by improving recipients’ objective understanding of key anatomical and risk-related information.

Acknowledgments

This study was supported by the Medical Statistics Support Center of the National Institute of Health and Medical Big Data at Xiamen University Medical School. Special acknowledgment is due to technician Ying Zhou (Medical Statistical Consultancy Services Program, National Institute for Data Science in Health and Medicine, School of Medicine, Xiamen University) for her help and guidance in the medical statistics consulting service project.

Data Availability

The datasets used or analyzed in this study are available from the corresponding author (ZW) upon reasonable request.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Editorial Notice

This randomized study was retrospectively registered, explained by authors as an administrative error. The editor granted an exception from ICMJE rules mandating prospective registration of randomized trials, because the approving IRB (the Ethics Committee of the First Affiliated Hospital of Xiamen University) provided a statement confirming that the protocol, including study design, methodology, and participant protection measures, remains consistent with the version reviewed by the committee. However, readers are advised to carefully assess the validity of any potential explicit or implicit claims related to primary outcomes or effectiveness, as retrospective registration does not prevent authors from changing their outcome measures retrospectively.

Multimedia Appendix 1
SeerLens B50R AR Glasses (Xvisio Technology): a complete augmented reality solution for commercial use.
[PDF File (Adobe PDF File), 4908 KB - jmir_v28i1e87198_app1.pdf]

Multimedia Appendix 2
A prespecified, standardized risk communication checklist based on the SEGUE framework, applied by clinicians in both trial arms. SEGUE: set the stage, elicit information, give information, understand the patient's perspective, and end the encounter.
[DOCX File , 13 KB - jmir_v28i1e87198_app2.docx]

Multimedia Appendix 3
Subjective Understanding Rating Scale (item structure and scoring system).
[DOCX File , 17 KB - jmir_v28i1e87198_app3.docx]

Multimedia Appendix 4
Objective assessment of medical knowledge comprehension following risk communication, comprising a 20-item questionnaire and scoring rubric.
[DOCX File , 21 KB - jmir_v28i1e87198_app4.docx]

Multimedia Appendix 5
Communication Satisfaction Rating Scale (item structure and scoring system).
[DOCX File , 17 KB - jmir_v28i1e87198_app5.docx]

Multimedia Appendix 6
CONSORT (Consolidated Standards of Reporting Trials) checklist.
[PDF File (Adobe PDF File), 122 KB - jmir_v28i1e87198_app6.pdf]

References

1. Ong L, de Haes J, Hoos A, Lammes F. Doctor-patient communication: a review of the literature. Soc Sci Med 1995 Apr;40(7):903-918. [doi: 10.1016/0277-9536(94)00155-m] [Medline: 7792630]
2. Singh M. Communication as a bridge to build a sound doctor-patient/parent relationship. Indian J Pediatr 2016 Jan 1;83(1):33-37. [doi: 10.1007/s12098-015-1853-9] [Medline: 26228061]
3. Charles C, Gafni A, Whelan T. Shared decision-making in the medical encounter: what does it mean? (or it takes at least two to tango). Soc Sci Med 1997 Mar;44(5):681-692. [doi: 10.1016/s0277-9536(96)00221-3] [Medline: 9032835]
4. Perin A, Galbiati TF, Ayadi R, Gambatesa E, Orena EF, Riker NI, et al. Informed consent through 3D virtual reality: a randomized clinical trial. Acta Neurochir (Wien) 2021 Feb 03;163(2):301-308. [doi: 10.1007/s00701-020-04303-y] [Medline: 32242272]
5. McDougall RJ. Computer knows best? The need for value-flexibility in medical AI. J Med Ethics 2019 Mar 22;45(3):156-160. [doi: 10.1136/medethics-2018-105118] [Medline: 30467198]
6. Niburski K, Guadagno E, Abbasgholizadeh-Rahimi S, Poenaru D. Shared decision making in surgery: a meta-analysis of existing literature. Patient 2020 Dec 03;13(6):667-681. [doi: 10.1007/s40271-020-00443-6] [Medline: 32880820]
7. Covello VT. Risk communication: an emerging area of health communication research. Annals of the International Communication Association 2016 May 18;15(1):359-373. [doi: 10.1080/23808985.1992.11678816]
8. Matiasek J, Wynia MK. Reconceptualizing the informed consent process at eight innovative hospitals. Jt Comm J Qual Patient Saf 2008 Mar;34(3):127-137. [doi: 10.1016/s1553-7250(08)34015-x] [Medline: 18419042]
9. Bernat JL, Peterson LM. Patient-centered informed consent in surgical practice. Arch Surg 2006 Jan 01;141(1):86-92. [doi: 10.1001/archsurg.141.1.86] [Medline: 16415417]
10. Schillinger D, Duran ND, McNamara DS, Crossley SA, Balyan R, Karter AJ. Precision communication: physicians' linguistic adaptation to patients' health literacy. Sci Adv 2021 Dec 17;7(51):1-12 [FREE Full text] [doi: 10.1126/sciadv.abj2836] [Medline: 34919437]

XSL•FO
RenderX

11. Cocanour CS. Informed consent-it's more than a signature on a piece of paper. Am J Surg 2017 Dec;214(6):993-997. [doi: 10.1016/j.amjsurg.2017.09.015] [Medline: 28974311]
12. Landriel Ibañez FA, Hem S, Ajler P, Vecchi E, Ciraolo C, Baccanelli M, et al. A new classification of complications in neurosurgery. World Neurosurg 2011 May;75(5-6):709-15; discussion 604. [doi: 10.1016/j.wneu.2010.11.010] [Medline: 21704941]
13. Jena AB, Seabury S, Lakdawalla D, Chandra A. Malpractice risk according to physician specialty. N Engl J Med 2011 Aug 18;365(7):629-636. [doi: 10.1056/nejmsa1012370]
14. Epstein NE. A review of medicolegal malpractice suits involving cervical spine: what can we learn or change? J Spinal Disord Tech 2011 Feb;24(1):15-19. [doi: 10.1097/BSD.0b013e3181c752ab] [Medline: 20087227]
15. Paasche-Orlow MK, Taylor HA, Brancati FL. Readability standards for informed-consent forms as compared with actual readability. N Engl J Med 2003 Feb 20;348(8):721-726. [doi: 10.1056/nejmsa021212]
16. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019 Jan 7;25(1):30-36 [FREE Full text] [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]
17. Hoffmann TC, Montori VM, Del Mar C. The connection between evidence-based medicine and shared decision making. JAMA 2014 Oct 01;312(13):1295-1296. [doi: 10.1001/jama.2014.10186] [Medline: 25268434]
18. Shay LA, Lafata JE. Where is the evidence? A systematic review of shared decision making and patient outcomes. Med Decis Making 2014 Oct 28;35(1):114-131. [doi: 10.1177/0272989x14551638]
19. Antel R, Abbasgholizadeh-Rahimi S, Guadagno E, Harley JM, Poenaru D. The use of artificial intelligence and virtual reality in doctor-patient risk communication: a scoping review. Patient Educ Couns 2022 Oct;105(10):3038-3050. [doi: 10.1016/j.pec.2022.06.006] [Medline: 35725526]
20. Seetohul J, Shafiee M, Sirlantzis K. Augmented reality (AR) for surgical robotic and autonomous systems: state of the art, challenges, and solutions. Sensors (Basel) 2023 Jul 06;23(13):6202 [FREE Full text] [doi: 10.3390/s23136202] [Medline: 37448050]
21. Bui DT, Barnett T, Hoang H, Chinthammit W. Usability of augmented reality technology in tele-mentorship for managing clinical scenarios-a study protocol. PLoS One 2022 Mar 31;17(3):e0266255-e0266266 [FREE Full text] [doi: 10.1371/journal.pone.0266255] [Medline: 35358249]
22. Giraldo J, Cho S, Eghrari N, Dholaria N, Farber S, Ehredt R, et al. Advances in neurosurgical education: literature review of mixed-reality simulation models and novel mixed-reality spine prototype. J Neurosurg Spine 2025 Mar 01;42(3):385-398. [doi: 10.3171/2024.8.SPINE24237] [Medline: 39752650]
23. Aguilar-Salinas P, Gutierrez-Aguirre SF, Avila MJ, Nakaji P. Current status of augmented reality in cerebrovascular surgery: a systematic review. Neurosurg Rev 2022 Jun 11;45(3):1951-1964. [doi: 10.1007/s10143-022-01733-3] [Medline: 35149900]
24. Skapinakis P. Spielberger State-Trait Anxiety Inventory. In: Encyclopedia of Quality of Life and Well-Being Research. Dordrecht, The Netherlands: Springer; Jan 01, 2014:6261-6264.
25. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005 Apr 30;53(4):695-699. [doi: 10.1111/j.1532-5415.2005.53221.x] [Medline: 15817019]
26. Makoul G. The SEGUE framework for teaching and assessing communication skills. Patient Educ Couns 2001 Oct 01;45(1):23-34. [doi: 10.1016/s0738-3991(01)00136-7] [Medline: 11602365]
27. Med Sci. URL: https://scales.arabpsychology.com/Ch/state-trait-anxiety-inventory-stai-form-y-chinese-version/ [accessed 2026-04-04]
28. Wehrkamp K, Miksch RC, Polzer H, Gilbert F, Bühner M, Holzapfel BM, et al. The impact of virtual-, augmented- and mixed reality during preoperative informed consent: a systematic review of the literature. J Med Syst 2025 Jun 24;49(1):89-109. [doi: 10.1007/s10916-025-02217-9] [Medline: 40555846]
29. Sweeny K, Melnyk D, Miller W, Shepperd JA. Information avoidance: who, what, when, and why. Review of General Psychology 2010 Dec 01;14(4):340-353. [doi: 10.1037/a0021288]
30. Eberhart L, Aust H, Schuster M, Sturm T, Gehling M, Euteneuer F, et al. Preoperative anxiety in adults - a cross-sectional study on specific fears and risk factors. BMC Psychiatry 2020 Mar 30;20(1):140 [FREE Full text] [doi: 10.1186/s12888-020-02552-w] [Medline: 32228525]
31. Oshima Lee E, Emanuel EJ. Shared decision making to improve care and reduce costs. N Engl J Med 2013 Jan 03;368(1):6-8. [doi: 10.1056/nejmp1209500]
32. Stalter LN, Baggett ND, Hanlon BM, Buffington A, Kalbfell EL, Zelenski AB, et al. Identifying patterns in preoperative communication about high-risk surgical intervention: a secondary analysis of a randomized clinical trial. Med Decis Making 2023 Apr 10;43(4):487-497. [doi: 10.1177/0272989x231164142]
33. Di Paolo M, Gori F, Papi L, Turillazzi E. A review and analysis of new Italian law 219/2017: 'provisions for informed consent and advance directives treatment'. BMC Med Ethics 2019 Mar 04;20(1):17 [FREE Full text] [doi: 10.1186/s12910-019-0353-2] [Medline: 30832644]
34. Bruch JD, Khazen M, Mahmic-Kaknjo M, Légaré F, Ellen ME. The effects of shared decision making on health outcomes, health care quality, cost, and consultation time: an umbrella review. Patient Educ Couns 2024 Dec;129:108408. [doi: 10.1016/j.pec.2024.108408] [Medline: 39214045]

XSL•FO
RenderX

35. Couët N, Desroches S, Robitaille H, Vaillancourt H, Leblanc A, Turcotte S, et al. Assessments of the extent to which health-care providers involve patients in decision making: a systematic review of studies using the OPTION instrument. Health Expect 2015 Aug 04;18(4):542-561 [FREE Full text] [doi: 10.1111/hex.12054] [Medline: 23451939]
36. Baggett ND, Schulz K, Buffington A, Marka N, Hanlon BM, Zimmermann C, et al. Surgeon use of shared decision-making for older adults considering major surgery: a secondary analysis of a randomized clinical trial. JAMA Surg 2022 May 01;157(5):406-413 [FREE Full text] [doi: 10.1001/jamasurg.2022.0290] [Medline: 35319737]
37. Gerwing J, Gulbrandsen P. Contextualizing decisions: stepping out of the SDM track. Patient Educ Couns 2019 May;102(5):815-816. [doi: 10.1016/j.pec.2019.03.024] [Medline: 31036288]
38. Pilnick A, Dingwall R. On the remarkable persistence of asymmetry in doctor/patient interaction: a critical review. Soc Sci Med 2011 Apr;72(8):1374-1382. [doi: 10.1016/j.socscimed.2011.02.033] [Medline: 21454003]

Abbreviations

AR: augmented reality
CONSORT: Consolidated Standards of Reporting Trials
DICOM: Digital Imaging and Communications in Medicine
LAR: legally authorized representative
MoCA: Montreal Cognitive Assessment
PoC: proof of concept
SDM: shared decision-making
SEGUE: set the stage, elicit information, give information, understand the patient's perspective, and end the encounter
STAI: State-Trait Anxiety Inventory

Edited by A Coristine; submitted 05.Nov.2025; peer-reviewed by W Guo, O Akinpeloye; comments to author 16.Jan.2026; accepted 30.Mar.2026; published 17.Apr.2026.

Please cite as:
Yuan Z, Shi Z, Wang Z
The Real-Time Support Role of Augmented Reality Technology in Shared Decision-Making in Neurosurgery Under the SEGUE Framework: Randomized Controlled Trial
J Med Internet Res 2026;28:e87198
URL: https://www.jmir.org/2026/1/e87198
doi:10.2196/87198
PMID:

©Zhengbo Yuan, Zhongjie Shi, Zhanxiang Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Patient Decision Aid for Antiobesity Medications: Mixed Methods Study of Human-Centered Design and Usability Evaluation

Li-Jen Wang[1], MSc; Yi-Jen Wang[2,3], MD, MPH; Yu-Lun Cheng[4], MD; Wen-Liang Fang[5,6], MD, PhD; Weu Wang[7,8], MD, EMBA; Meng-Cong Zheng[9], PhD

[1]College of Design, National Taipei University of Technology, Taipei, Taiwan
[2]Department of Family Medicine, Taipei Veterans General Hospital, Taipei, Taiwan
[3]School of Medicine, College of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan
[4]Division of Cardiology, Department of Internal Medicine, Taipei Medical University Hospital, Taipei, Taiwan
[5]Department of Surgery, Weight Management Medical Center, Taipei Veterans General Hospital, Taipei, Taiwan
[6]School of Medicine, College of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan
[7]Division of General Surgery, Department of Surgery, Taipei Medical University Hospital, Taipei, Taiwan
[8]Department of Surgery, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
[9]Department of Industrial Design, National Taipei University of Technology, Taipei, Taiwan

Corresponding Author:
Meng-Cong Zheng, PhD
Department of Industrial Design
National Taipei University of Technology
1, Sec. 3, Chung-Hsiao E. Rd
Taipei, 10608
Taiwan
Phone: 886 988 300 461
Email: zmcdesign@gmail.com

Abstract

Background: The global burden of obesity continues to rise, highlighting the need for patient-centered approaches to weight management. Shared decision-making is particularly important in the selection of antiobesity medications (AOMs), as treatment options differ in mechanism, effectiveness, side effects, routes of administration, and cost. Despite this preference-sensitive context, only a few patient decision aids (PDAs) have been culturally and clinically adapted for use in Asian populations.

Objective: This study aims to design, develop, and evaluate a digital PDA, OptiWeight, to support shared decision-making for AOM selection, incorporating perspectives from health care professionals and patients.

Methods: This mixed methods, multicenter study, conducted between August 2022 and November 2025, applied a 4-stage human-centered design process. An evidence-informed prototype was developed based on clinical guidelines, followed by 2 rounds of usability testing using think-aloud protocols to assess navigation structures, perceived usability (System Usability Scale [SUS]), and cognitive workload (NASA Task Load Index [NASA-TLX]). Semistructured interviews with health care professionals specializing in weight management, guided by the Consolidated Framework for Implementation Research, informed clinical implementation and workflow integration. Finally, patients with overweight or obesity evaluated usability, cognitive workload, and overall user experience in outpatient settings. Qualitative data were analyzed using content analysis, and 1-way analysis of variance examined changes in usability and workload across stages.

Results: A total of 174 individuals were included across all study stages (usability testing among adults: n=78; health care professional interviews: n=18; and clinical evaluation among patients: n=78). Iterative usability testing comparing system- and user-controlled navigation structures revealed complementary strengths and limitations, leading to the adoption of a hybrid navigation structure supporting both sequential guidance and flexible comparison. Additional design requirements included the use of icon arrays to enhance risk comprehension and localization features such as treatment cost displays and clarification of socially impactful side effects. Perceived usability increased from initial testing to clinical evaluation (SUS: 60.53-73.65, $P<.001$), meeting good usability thresholds, while cognitive workload decreased (NASA-TLX: 40.35-16.69, $P<.001$).

XSL•FO
RenderX

Conclusions: Through a systematic human-centered design process integrating health care professional and patient perspectives, OptiWeight addresses the lack of culturally adapted PDAs for AOM decision-making in Mandarin-speaking populations while capturing user needs—particularly regarding navigation flexibility and risk visualization. The final tool demonstrated good usability and feasibility, and workflow considerations suggest potential for integration into routine weight-management care. Further research is needed to evaluate its impact on decision quality and real-world implementation outcomes.

(J Med Internet Res 2026;28:e89428) doi:10.2196/89428

KEYWORDS

preference-sensitive decision-making; risk communication; shared decision-making; digital health interventions; obesity; Taiwan

Introduction

The global burden of obesity continues to rise, with the World Obesity Atlas 2024 predicting that over half of the world's population will be overweight or obese by 2035 [1]. In Taiwan, national survey data indicate that the combined prevalence of overweight and obesity among adults aged 18 years and older rose from 32.7% in 1993-1996 to 50.8% in 2022, exceeding half of the adult population [2,3]. These trends are not evenly distributed across population subgroups. Prevalence is consistently higher in men than in women (61.2% vs 41.5% in 2020-2024) and increases with age, with the highest rates observed among middle-aged adults (eg, 72.9% in men aged 35-44 years) [4]. In addition, approximately one-third of school-aged children are classified as overweight or obese, indicating that this trajectory extends across the life course [3]. Obesity is now widely recognized as a major risk factor for type 2 diabetes, cardiovascular disease, and numerous chronic conditions [5]. A range of evidence-based weight-management treatments is available, including lifestyle modification, medication therapy, and bariatric surgery [6]. Among these options, antiobesity medications (AOMs) have become increasingly common and effective in clinical practice [7]. As each medication differs in its mechanism of action, expected weight-loss outcomes, and potential side effects, there is no universally optimal choice. Treatment selection is therefore considered a preference-sensitive decision that should reflect the values and priorities of individual patients [8-10].

Shared decision-making (SDM) is particularly critical in such contexts. Clinicians are responsible for presenting the best available evidence, while patients—supported throughout the process—evaluate the options, clarify their preferences, and choose the alternative that best aligns with their values [11]. Patient decision aids (PDAs) are key tools that promote SDM by presenting clinical scenarios, summarizing the benefits and risks of available options, and helping patients articulate and communicate their values [12,13]. Randomized trials and systematic reviews have consistently demonstrated that PDAs increase patient knowledge, improve understanding of treatment trade-offs, reduce decisional conflict, and enhance patient-clinician communication [13,14]. These benefits are particularly meaningful for underserved populations—including those with lower health literacy, lower income, or limited access to specialist care—who often face greater challenges in understanding health information, participating in clinical decision-making, and achieving value-concordant decisions [15].

This is especially important in obesity care. In Taiwan, a volume-driven reimbursement system may limit communication and continuity of care [16], while short consultation times and limited provider training further constrain in-depth weight-management discussions [17]. At the patient level, unmet information needs may lead individuals to seek AOMs without professional guidance [18], and weight stigma may discourage engagement with formal health care [19], further exacerbating disparities in access to evidence-based care, particularly in settings where treatment costs are borne out of pocket, and specialized services are unevenly distributed [20].

Several PDAs have been developed in the obesity field, including decision aids for adolescent weight management [21] and tools supporting choices among bariatric surgery procedures [22-24]. These interventions have been shown to decrease decisional conflict and improve knowledge and satisfaction [21-24]. However, existing PDAs primarily target surgical decision-making [22-24] or specific subgroups such as adolescents and their parents [21], leaving a gap in tools designed for adults considering medication therapy. An environmental scan of existing weight-loss decision aids has further shown that many available tools do not meet key principles of the International Patient Decision Aid Standards (IPDAS), particularly regarding language localization and optimization of content for end users [25,26].

At the time of this study, no PDA in Taiwan specifically supported patients in choosing among AOMs, despite the clinical importance of such decisions. The recent introduction of several new AOMs, including glucagon-like peptide-1 receptor agonists, has expanded treatment options but simultaneously increased the complexity of decision-making. Providing a PDA before the initial consultation may allow patients more time to process information, reflect on personal preferences, and engage more efficiently with clinicians, ultimately supporting higher-quality decisions.

To address this gap, we undertook a human-centered design process to develop a high-quality digital PDA tailored to AOMs for use in Taiwanese clinical settings. Working collaboratively with health care professionals specializing in weight management, patients, and decision-making stakeholders, our team of industry designers and engineers created and refined OptiWeight, an iPad-based PDA. This study aimed to develop and evaluate OptiWeight by examining its design rationale and assessing its usability, cognitive workload, and subjective user experience.

XSL•FO
RenderX

Methods

Study Design

We conducted a user-centered, mixed methods development and evaluation study across 2 medical centers, namely, Taipei Veterans General Hospital and Taipei Medical University Hospital, in Taiwan. The study followed a multicycle iterative design process grounded in the human-centered design framework and the principles outlined in ISO (International Organization for Standardization) 9241-210 [27]. In the context of PDA development, user-centered design entails iterative cycles of understanding end users, developing and refining prototypes, and systematically observing how users interact with these prototypes [28,29]. An overview of the development and evaluation process is presented in Figure 1.

Figure 1. Iterative development and evaluation process of the digital patient decision aid. CFIR: Consolidated Framework for Implementation Research; PDA: patient decision aid; TFDA: Taiwan Food and Drug Administration.

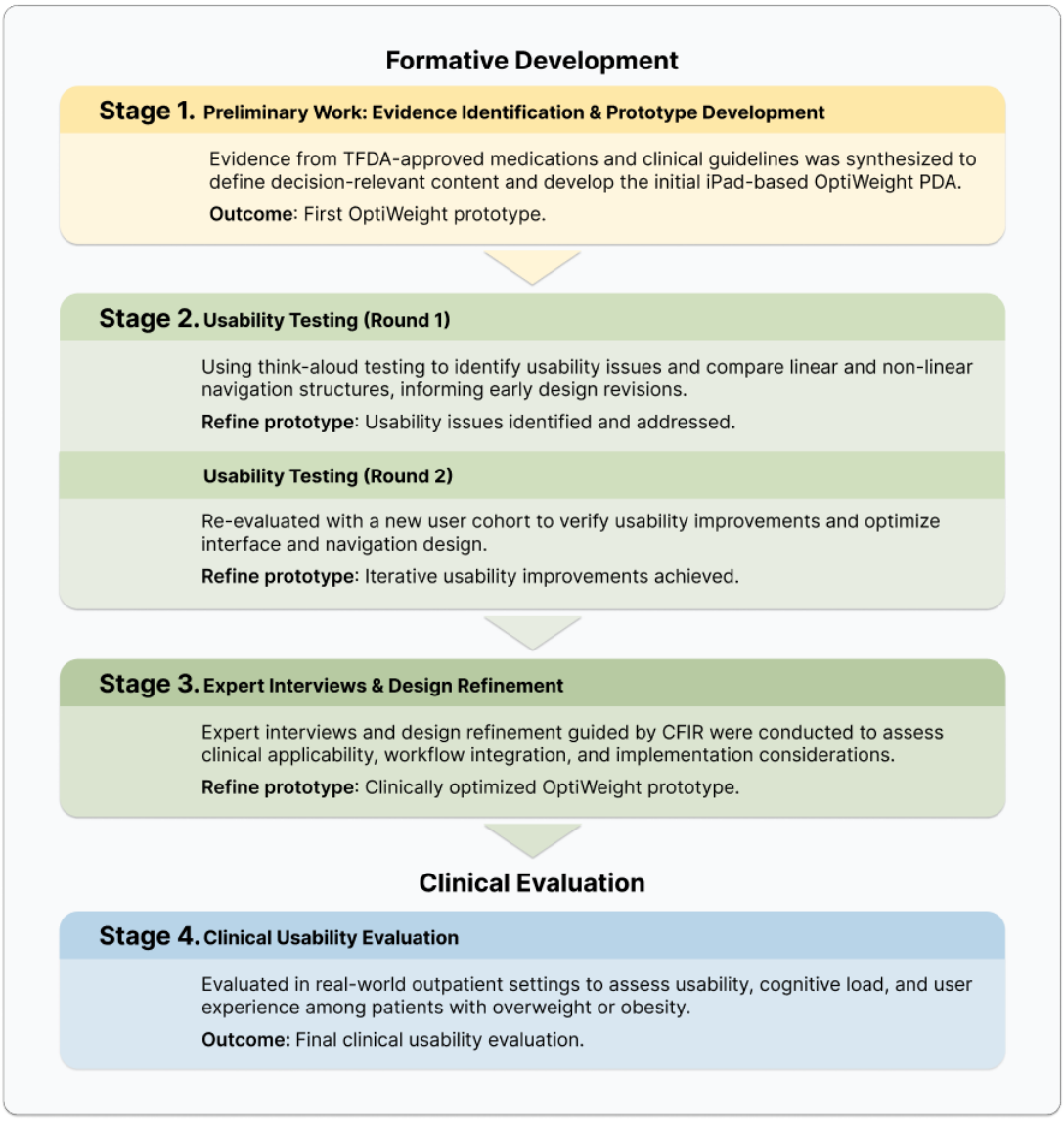

Our development process was informed by established user-centered PDA design methodologies [28,29], guidance from the IPDAS Collaboration [30-32], and the IPDAS reporting checklist [33]. The interdisciplinary research team included a doctoral researcher specializing in SDM (LJW), a clinical expert in obesity management (YJW), and a patient-centered design specialist (MCZ).

Across the overall development process, OptiWeight underwent 3 iterative refinements: (1) after round 1 usability testing (December 2022), usability issues were identified and addressed, with modifications focusing on navigation structure, risk visualization, and onboarding design; (2) after round 2 usability testing (April 2023), further usability improvements were achieved, including finalization of the hybrid navigation architecture and reorganization of medication content; and (3) after expert interviews (stage 3; January-May 2024), the prototype was clinically optimized, with refinements addressing workflow integration, cost display, contraindication filtering, and terminology clarity.

XSL•FO
RenderX

This study adopted a convergent mixed methods design, in which qualitative and quantitative data were collected in parallel at each stage and integrated during interpretation to provide complementary insights into usability and implementation. Qualitative findings were used to explain and contextualize quantitative results, particularly in identifying usability issues and informing iterative design refinements. A consistent inductive content analysis approach was applied across all stages, with stage-specific frameworks used where appropriate to guide interpretation.

Preliminary Work: Identification of Evidence and Developing the Prototype (Stage 1)

The initial development phase focused on creating OptiWeight, an interactive, iPad-based digital PDA designed to support patients in selecting appropriate AOMs. To inform the content framework, we reviewed the Healthwise decision aid "Obesity: Should I Take Weight-Loss Medicine?" [25] and identified all AOMs approved for weight management by the Taiwan Food and Drug Administration (TFDA) as of 2022. We further conducted targeted literature searches to extract clinical evidence on the efficacy, mechanisms of action, and side effects of these medications [34-40].

In addition, we incorporated risk-prediction information from Taiwan's Health Promotion Administration [34] and consulted the 2022 Obesity Medicine Association Clinical Practice Statement to define the essential information domains for medication-related decision support. These domains included treatment effectiveness (eg, expected weight loss, metabolic improvement), safety (indications and contraindications), side effects, route of administration, cost considerations, and medication availability [41].

To maintain clarity around the core decision problem, we intentionally excluded content related to lifestyle modification or nutrition counseling, such as exercise guidance or calorie calculations. These strategies are foundational for all patients, regardless of medication use. By excluding them, we aimed to ensure that users recognized the specific decision target—choosing among medication options—rather than conflating medication decisions with broader weight-management choices.

Usability Testing of the Prototype (Stage 2)

Two rounds of usability testing were conducted to evaluate the OptiWeight prototype and compare 2 information architecture formats—system-controlled navigation and user-controlled navigation.

Participants were recruited through convenience sampling using posters placed on university campuses and community bulletin boards. Interested individuals contacted the research team and were screened for eligibility. Eligible participants were adults aged 20 years or older who were able to operate and view the interface, with assistance if needed. Exclusion criteria included severe visual impairment, cognitive impairment, or inability to provide informed consent.

BMI was not collected during early prototype usability testing (Stage 2) because the focus of this stage was interface usability rather than clinical characterization; consistent with established conventions in human-computer interaction research, general adult samples are appropriate for identifying fundamental usability issues that are likely to persist or be amplified among clinical users [42]. Additionally, omitting BMI reduced participant burden and potential sensitivity related to weight stigma. BMI was subsequently included in the clinical evaluation stage to better describe the target population.

In the first round, initiated in December 2022, 2 separate recruitment posters—1 for each navigation prototype—were posted simultaneously at the same locations with a shared response deadline. Participants self-selected into a group by responding to the respective poster, resulting in unequal group sizes (system-controlled navigation: n=21; user-controlled navigation: n=17). No randomization or stratification by participant characteristics was applied. For round 2, to address the unbalanced allocation in round 1, recruitment was designed to achieve equal group sizes: the first 20 respondents to each poster were enrolled, resulting in balanced groups (n=20 per group). No participant withdrawals were observed during round 1 and round 2 usability testing.

Findings from the first round informed iterative revisions to the prototype, which was subsequently evaluated in a second round of usability testing conducted in April 2023.

Participants were invited to complete tasks using both navigation structures without time restrictions. User interactions were captured through screen recordings and the think-aloud protocol [43]. Following the task session, participants completed a semistructured interview and subsequently filled out the System Usability Scale (SUS) [44] and the NASA Task Load Index (NASA-TLX) [45]. Both the SUS and NASA-TLX were administered online once, immediately after completion of all task activities. SUS and NASA-TLX scores were summarized using descriptive statistics (means and SDs), and cross-stage comparisons are described in the "Statistical Analysis" section. The semistructured interview guide was developed based on established usability evaluation frameworks and iteratively refined during early sessions to improve clarity and relevance.

Qualitative data were deidentified and analyzed using inductive content analysis [46]. Transcripts were first read in full to gain an overall sense of the data, followed by open coding to identify usability-relevant themes. Initial open coding was conducted by 1 researcher (LJW) and independently reviewed by a second researcher (MCZ) to enhance analytical rigor. Coding discrepancies were resolved through discussion until consensus was reached. The research team acknowledged their prior experience in PDA design and usability research and took steps to minimize potential bias through independent coding and consensus discussions. Member checking was not performed, as the data were deidentified before analysis. Usability issues, representative participant quotes, and recommended design modifications were summarized in a structured table.

Both usability rounds included assessments of system- and user-controlled navigation prototypes. Detailed comparative analyses of navigation structures are reported elsewhere [47]. Therefore, in this study, we report aggregated mean values

XSL•FO
RenderX

across navigation conditions to illustrate overall usability trends across development stages.

Expert Interviews and Design Refinement (Stage 3)

Between January and May 2024, we conducted expert interviews at 2 medical centers with health care professionals involved in weight management, including doctors who specialize in obesity medicine, nurses, dietitians and nutritionists, and pharmacists. Eligible participants were adults aged 20 years or older. Eligible health care professionals were identified through purposive sampling: researchers directly invited clinical colleagues with relevant expertise in weight management, and recruitment posters were additionally displayed in clinical areas of the participating medical centers. Experts were invited to evaluate the clarity of medication information, visual and interactive design, and the feasibility of integrating OptiWeight into routine clinical workflows.

To identify user needs and evaluate implementation considerations, we applied the Consolidated Framework for Implementation Research (CFIR) [48,49], which has been widely used to guide implementation of clinical decision support tools (eg, BREASTChoice) [50]. We adapted the 5 CFIR domains—Intervention Characteristics, Outer Setting, Inner Setting, Characteristics of Individuals, and Process—and mapped the resulting themes onto specific design decisions, including interface layout, informational content, interaction patterns, and instructional guidance. This approach allowed us to address design issues unique to an iPad-based PDA for AOMs while systematically examining how individual, organizational, and external contextual factors may influence adoption.

One-on-one interviews were conducted in private rooms within the hospitals and lasted approximately 20-40 minutes. All recruited experts completed the interview, and no participants withdrew or declined to participate. All interviews were audio-recorded, deidentified, and transcribed verbatim for analysis.

After the interview, participants also completed the SUS [44] and the NASA Task Load Index (NASA-TLX) [45] online, providing quantitative measures of perceived usability and cognitive workload. SUS and NASA-TLX scores were summarized using descriptive statistics (means and SDs); cross-stage comparisons are described in the “Statistical Analysis” section.

Interview transcripts were analyzed using inductive content analysis. Initial open coding was conducted by the first author (LJW), who read transcripts in full and identified meaningful segments. A preliminary coding framework (codebook) was developed inductively based on recurring patterns in the data and iteratively refined during analysis. Resulting themes were then organized according to the 5 CFIR domains, which had guided the interview question design. Coding was independently reviewed by a senior researcher with expertise in human-centered design (MCZ) to enhance analytical rigor, and discrepancies were resolved through discussion until consensus was reached. No dedicated qualitative analysis software was used; all data were managed and coded manually.

Clinical Usability Evaluation (Stage 4)

In February 2025, we conducted a clinical usability evaluation in weight-management outpatient clinics at 2 medical centers. Adults aged 20 years or older with overweight or obesity, that is, a BMI of 24 kg/m^2 or higher as defined by Taiwan’s guideline [51], who were able to independently operate a digital interface were eligible for participation. Individuals requiring assistance to use the interface or those with impaired consciousness were excluded. Participants were recruited through 2 approaches: attending physicians informed eligible patients during clinic visits, and recruitment posters were displayed in outpatient clinic areas. During their clinic visit, participants interacted with the further refined, finalized version of OptiWeight and subsequently completed the SUS [44] and the NASA Task Load Index (NASA-TLX) [45] in a paper-based format. Both the SUS and NASA-TLX were administered once, immediately after completion of the interaction tasks. SUS and NASA-TLX scores were summarized using descriptive statistics (means and SDs); cross-stage comparisons are described in the “Statistical Analysis” section. A brief semistructured interview (10-15 minutes) was conducted to explore participants’ clinical use experience, perceived cognitive workload, and interaction patterns. All recruited participants completed the study procedures, with no reported withdrawals.

Interview transcripts were analyzed using inductive content analysis. Initial open coding was conducted by one researcher (LJW) and independently reviewed by a second researcher (MCZ) to enhance analytical rigor. Discrepancies were resolved through discussion until consensus was reached. Themes were organized according to Nielsen’s 5 usability heuristics—learnability, efficiency, memorability, error tolerance, and satisfaction—to complement and contextualize the quantitative usability findings. No dedicated qualitative analysis software was used; all data were managed and coded manually.

Statistical Analysis

Basic demographic data were compared using chi-square tests for categorical variables and independent-samples *t* tests (unpaired and 2-tailed) for continuous variables. Sex (male or female, as assigned at birth) was self-reported and used for demographic characterization throughout the study. To assess the effects of iterative refinements on usability and cognitive workload across the development process, we compared SUS and NASA-TLX scores between stage 2 usability testing (rounds 1 and 2), stage 3 expert interviews, and stage 4 clinical evaluation using 1-way analysis of variance (ANOVA). Before analysis, normality was assessed using the Shapiro-Wilk test and homogeneity of variance using the Levene test. Where assumptions were violated, Welch ANOVA was used as the primary test, and Games-Howell post hoc comparisons were applied. NASA-TLX subscale analyses were treated as exploratory and interpreted descriptively; no correction for multiple comparisons was applied across ANOVAs. All statistical tests were 2-tailed with a significance level of α=.05, and a P value <.05 was considered statistically significant. Statistical analyses were conducted using IBM SPSS Statistics (version 29; IBM Corp).

XSL•FO
RenderX

No formal a priori power analysis was conducted. This study followed an exploratory, iterative human-centered design approach in which the primary goal was to identify usability issues and inform design refinements rather than to test predefined statistical hypotheses. Consistent with established conventions in human-computer interaction research, small purposive samples were considered appropriate for usability evaluation [42], and data saturation was not formally assessed, as sample sizes were predefined to support iterative testing and refinement.

All qualitative data—including think-aloud recordings, interview transcripts, and field observations—were complete across all stages; no qualitative data were missing. Regarding quantitative data, no missing data were identified in stage 2 or stage 4. In stage 3, 3 of 18 (17%) participants were excluded from SUS and NASA-TLX analyses due to incomplete questionnaire responses; complete data were therefore available for 15 participants. Little missing completely at random test indicated that the missing data were missing completely at random (χ^2_2=4.811, *P*=.09), supporting the use of listwise deletion. Given the small number of missing cases and the exploratory nature of the stage, multiple imputation was not performed.

This study was reported in accordance with the GRAMMS (Good Reporting of a Mixed Methods Study; Multimedia Appendix 1) guidelines and the COREQ (Consolidated Criteria for Reporting Qualitative Research; Multimedia Appendix 2) checklist.

Ethical Considerations

Ethical approval was obtained from multiple institutional review boards across different study phases. Stage 1 involved no human participants and did not require ethical approval. Stage 2 usability testing was approved by the Research Ethics Committee of National Taiwan University (approval number 202206ES059). Stages 3 and 4, conducted at 2 medical centers, were approved by the Institutional Review Board of Taipei Veterans General Hospital (approval number 2024-01-013AC) and the Taipei Medical University-Joint Institutional Review Board (approval number N202403021). All participants provided written informed consent before participation. Upon completing the study procedures, experts received a gift card valued at approximately US $10, and general participants received one valued at approximately US $6.

Results

Stage 1: Prototype Development

The initial prototype was developed in Figma (Figma Inc) and presented evidence-based information on available AOMs, associated risks, and potential side effects. Following clinical review by the obesity specialist (YJW), the medication list was refined to include all 3 weight-management medications approved by the TFDA as of 2022—orlistat, naltrexone-bupropion (Contrave), and liraglutide (Saxenda)—as well as semaglutide (Ozempic), which was available during the study period and contains the same active ingredient as semaglutide (Wegovy), a United States FDA–approved medication for weight-loss indications.

Prototype development was led by the interface design specialist (MCZ) and the research team, guided by established visual design principles for health communication [52] and best practices for presenting risk information in decision aids [53,54]. Prior research demonstrates that icon arrays effectively convey statistical risk information—such as natural frequencies (eg, 1 out of 100)—and are particularly beneficial for individuals with lower numeracy skills [55,56]. Moreover, studies by Zikmund-Fisher and colleagues [54] show that anthropomorphic icons (eg, restroom-style human figures) enhance recall of risk information.

Informed by this evidence, the updated prototype incorporated icon-based representations and visualized risk formats to optimize cognitive processing and memory retention. The layout and visual flow were deliberately refined to support comprehension and improve user experience [57,58].

Stage 2: Usability Testing (Rounds 1 and 2)

Overview

Stage 2 involved 2 rounds of usability testing comparing 2 navigation structures: a *system-controlled* prototype, which guided users through content in a fixed sequence, and a *user-controlled* prototype, which allowed flexible navigation across sections.

Participant Characteristics of Stage 2 Usability Testing

Across the 2 rounds of stage 2 usability testing, a total of 78 participants were included. In round 1, 38 participants were recruited (21 in the system-controlled navigation group and 17 in the user-controlled navigation group), with a mean age of 26.5 (SD 7.6) years. In round 2, 40 participants were included, with equal allocation to the system- and user-controlled prototypes (n=20 each) and a mean age of 25.1 (SD 6.1) years. Participant characteristics were comparable across rounds in terms of age, sex, and education level (Table 1).

XSL•FO
RenderX

Table 1. Participant characteristics of prototype usability testing (rounds 1 and 2).

Characteristics	Round 1 (n=38)				Round 2 (n=40)			
	Overall	System-controlled navigation group (n=21)	User-controlled navigation group (n=17)	*P* value	Overall	System-controlled navigation group (n=20)	User-controlled navigation group (n=20)	*P* value
Age (years), mean (SD)	26.5 (7.6)	27.6 (9.8)	25.0 (3.1)	.30	25.1 (6.1)	23.25 (1.7)	26.85 (8.1)	.06
Sex, n (%)				>.99				.11
Male	25 (66)	14 (67)	11 (65)		18 (45)	12 (60)	6 (30)	
Female	13 (34)	7 (33)	6 (35)		22 (55)	8 (40)	14 (70)	
Education level, n (%)				.48				>.99
Bachelor	29 (76)	15 (71)	14 (82)		27 (68)	14 (70)	13 (65)	
Graduate or above	9 (24)	6 (29)	3 (18)		13 (32)	6 (30)	7 (35)	

Usability and Cognitive Workload of Stage 2 Usability Testing

In *usability testing round 1*, usability scores remained in the mid-range for both navigation structures. The user-controlled prototype yielded an SUS score of 59.4 (SD 18.3), while the system-controlled prototype scored 61.4 (SD 20.0). NASA-TLX scores indicated substantial workload across conditions, with the user-controlled structure scoring 45.1 (SD 12.7) and the system-controlled structure scoring 36.5 (SD 20.0). Variability in responses was pronounced, suggesting divergent user experiences: while some participants found the interface intuitive, others reported considerable difficulty. Despite these challenges, most participants expressed positive attitudes toward OptiWeight and felt it improved their understanding of medication options. However, many continued to experience confusion during navigation.

In *usability testing round 2*, the user-controlled prototype received an SUS score of 66.4 (SD 16.4), while the system-controlled prototype scored 66.1. NASA-TLX scores indicated substantial perceived workload, with the user-controlled prototype scoring 46.8 (SD 11.4) and the system-controlled prototype scoring 35.5 (SD 12.1). Similar to round 1, usability ratings showed wide variability, with some participants reporting strong ease of use and others expressing significant difficulty. Participants also described notable mental, physical, temporal, and emotional demands during task completion. Despite these challenges, participants generally responded positively to OptiWeight and felt it improved their understanding of medication options, although aspects of the navigation remained confusing for some.

Summary of Stage 2 Usability Testing

In usability testing round 1, participants preferred visualized risk information, particularly bar charts (28/38, 74%). However, brief explanatory text often left users uncertain about the meaning of the graphics. Many recommended pairing visualizations with numerical values to improve interpretability. Participants also requested medication images before detailed descriptions, as visual cues supported understanding of administration methods and improved memory of treatment options. Several participants perceived the medication descriptions as overly brief and recommended adding indications, dosage instructions, and precautions.

A participant with color vision deficiency noted difficulty distinguishing between light-blue and gray icons in the risk displays. Preferences regarding information architecture were mixed: 19 out of 35 (54%) favored the user-controlled structure due to ease of comparison and flexible navigation, although increased information density and nonsequential flow were noted as drawbacks. Conversely, the system-controlled structure reduced cognitive load and provided helpful progress indicators and navigation cues, but made side-by-side comparison difficult and required frequent back-and-forth navigation.

Observational data identified several usability issues: risk and side effect dropdown menus were frequently overlooked; chronic disease items lacked hierarchical separation from the “Next” button; icon arrays were not visually distinguished from text; some participants misunderstood the function of the “Decision” button; and users were uncertain whether disease-related fields were required if they had no chronic conditions. Certain technical medical terms were considered difficult to understand (eg, “type 2 diabetes”), prompting requests for brief explanations. Participants also suggested a short introductory tutorial to reduce initial cognitive load.

In usability testing round 2, interview findings indicated that 27 out of 40 (68%) participants preferred the user-controlled prototype, citing greater flexibility, faster access to desired content, and the ability to view multiple pieces of information on a single page. However, participants also noted drawbacks, including high information load and the absence of progress cues, which made it difficult to determine the end point of the task. The system-controlled structure was valued for its stepwise, guided flow that facilitated deeper comparison of treatment details. Nonetheless, participants reported that the mandatory sequencing resulted in repetitive reading and cumbersome navigation, reducing satisfaction.

Participants continued to express a strong preference for visualized risk information, particularly numerical risk values presented in graphical formats. However, visualizations alone

XSL•FO
RenderX

were insufficient; users still relied on accompanying text and numerical values to fully interpret medical information. Medication-related visuals—such as images depicting appearance and dosage form—were viewed as important components of a PDA. Participants explained that medication names were difficult to remember, whereas visual cues supported recall and understanding of administration routes. Given the diversity of users, participants emphasized the need for color choices that accommodate individuals with color blindness or color vision deficiency.

Findings also suggested that information architecture influences users' risk perception and comprehension. Participants prioritized rapid access to the information they cared about most, and restrictive navigation sometimes hindered understanding, possibly due to reduced attention or frustration. Based on these insights, the final version of OptiWeight adopted a hybrid architecture: a guided system-controlled structure for overall flow, combined with an open, user-controlled layout for medication-specific content to support flexible comparison (Table 2).

Table 2. Usability test (round 2) findings and resulting design modifications.

Issue[a] and illustrative quotes	Modification (with screenshot reference)
Navigation structure	
• "It's faster, more flexible, and I can see more info on one page." [User-controlled navigation] • "I don't know where the end is—there's no progress guide." [User-controlled navigation] • "It's guided and detailed, but too repetitive and frustrating to go back and forth." [System-controlled navigation] • "I like this page where I can see the introduction of all four medications at once, but afterward I can only view one medication at a time and cannot compare their pros and cons simultaneously." [System-controlled navigation] • "Before using this tool, I had no knowledge of these medications. By the time I reached later pages, I had already forgotten the names and details of the earlier ones." [System-controlled navigation] • "I want to quickly find information about medication side effects, but the current sequence forces me to go through risk information first, which does not help me answer my question." [System-controlled navigation] • "There are too many 'Next' buttons, and it becomes frustrating." [System-controlled navigation]	• We adopted a hybrid navigation model that retains the progress indicators from the system-controlled structure while allowing free user-controlled browsing within the medication pages. • We also rearranged the browsing sequence by moving the medication overview comparison before the individual medication descriptions, enabling users to gain an initial understanding before reviewing detailed therapeutic information.
Drug recognition and information	
• "I want to quickly know which medication results in the greatest weight reduction, but I can't find it." • "This says the medication leads to 9.3% annual weight loss, but if I only use it for one month, will I get the same effect?"	• We added a line chart showing expected weight-loss effects over time for each medication and incorporated weight-loss outcomes directly into the medication summary pages.
User guidance/onboarding	
• "The instruction section at the beginning is too long." • "If this is my second time using the tool, I don't need to read the tutorial again."	• We shortened the onboarding tutorial and added a Skip button for returning users.
Readability and layout	
• "The medication explanation page contains too much text—it's long and difficult to read."	• We reorganized the text in the pop-up window for detailed medication information by dividing indications, dosage, side effects, contraindications, and precautions into 3 structured sections. • We also reformatted the "risks and side effects" pop-up into 2 grouped content blocks to improve readability.
Features and interactivity	
• "The medication suggests it is only recommended for BMI ≥ 30, but I don't know what my BMI is." • "The question-mark icon on the medication details pop-up made me think it was a system error message."	• We added a BMI results page that calculates the user's BMI based on the basic information they entered. • We changed the question-mark icon on the medication details pop-up to an 'i' (information) symbol to better convey its purpose.

[a]Issues are derived from think-aloud observations and interviews, and those from round 1 that did not reappear were considered resolved. Round 2 captured both the validation of prior modifications and the emergence of new usability concerns.

XSL•FO
RenderX

Stage 3: Expert Interviews and Design Refinement

Participant Characteristics

A total of 18 health care professionals specializing in weight management, with an average of 16.1 years of clinical practice experience (SD 11.0 years), participated in and completed the expert interviews. The sample included 12 (67%) physicians/surgeons, 1 (6%) nurse, 1 (6%) pharmacist, and 4 (22%) dietitians and nutritionists. Participants were mainly female (n=11, 61%); all had a bachelor's degree or higher and a mean age of 40.8 (SD 11.2) years (range 27-57 years; Table 3).

Table 3. Characteristics of health care professionals specializing in weight management (in stage 3 expert interviews) and patients (in stage 4 clinical evaluation).

Characteristics	Participants	
	Health care professionals (n=18)	Patients (n=78)
Age (years), mean (SD)	40.8 (11.2)	45.0 (11.1)
Sex, n (%)		
Male	7 (39)	33 (42)
Female	11 (61)	45 (58)
Education level, n (%)		
Junior high	0 (0)	1 (1)
High school	0 (0)	9 (12)
College	0 (0)	9 (12)
Bachelor	11 (61)	32 (41)
Graduate or above	7 (39)	27 (35)
Profession, n (%)		
Physician/surgeon	12 (67)	N/A[a]
Nurse	1 (6)	N/A
Pharmacist	1 (6)	N/A
Dietitian and nutritionist	4 (22)	N/A
BMI (kg/m^2), n (%)		
Overweight (24-26.9)	N/A	9 (12)
Mild obesity (27-29.9)	N/A	15 (19)
Moderate obesity (30-34.9)	N/A	38 (49)
Severe obesity (≥35)	N/A	16 (21)

[a]N/A: not applicable.

Usability and Cognitive Workload of Stage 3 Expert Interviews

Experts reported a mean SUS score of 77.8 (SD 9.4), corresponding to the "good" usability range, indicating that OptiWeight was generally perceived as intuitive and user-friendly. The overall cognitive workload was moderate, with a mean NASA-TLX score of 28.1 (SD 11.3). Of the 18 obesity specialists who participated in the design refinement sessions, complete SUS and NASA-TLX data were available for 15 participants.

Summary of Stage 3 Expert Interviews

Using the 5 CFIR domains, we synthesized the perspectives of 18 obesity specialists. Overall impressions were positive, with experts providing an average rating of 8.2 (on a 1=poor to 10=excellent scale). They described OptiWeight as visually appealing, intuitive to use, and effective in presenting visual and textual information that supports patient understanding of medication options. However, several refinements were identified under the *Intervention Characteristics* domain. Experts recommended restructuring lengthy instructional pages into embedded, stepwise guidance with clear status cues; revising the "no treatment" option to "no medication treatment/lifestyle modification"; explicitly noting that liraglutide (Saxenda) should be discontinued if less than 5% weight loss is achieved after 6 months; highlighting the social implications of Orlistat's oily stools; and clearly labeling semaglutide (Ozempic) as off-label or removing it from the selections. They also recommended organizing medications by administration route (oral versus injection) and expanding cost information to include monthly and full-course expenses. To reduce cognitive burden, experts encouraged replacing technical terminology with more accessible language.

Regarding visual design, experts noted that icon arrays representing treatment effects might be misinterpreted as risks

XSL•FO
RenderX

and recommended adding percentage labels or using bar or pie charts. For line graphs, they suggested simplifying the display by reducing the number of lines, removing the placebo curve, and directly annotating key time points to improve interpretability.

Under the *Outer Setting* and *Inner Setting* domains, experts recognized that digital SDM tools align with national policy directions and could enhance institutional visibility. They also emphasized the importance of incorporating local contextual considerations, including Taiwan's BMI criteria (overweight: $24 \leq BMI < 27$ kg/m^2; obesity: $BMI \geq 27$ kg/m^2) [51], out-of-pocket medication costs under the National Health Insurance system, completeness of pharmaceutical formularies, Ministry of Health and Welfare education requirements, and copyright considerations. Organizational challenges included the costs and depreciation of tablet devices, privacy and Wi-Fi reliability, and variation in drug availability across hospitals. Experts stressed that the tool should reduce—not increase—clinician workload and should be operable by nurses or case managers without conflicting with existing electronic medical record workflows.

Within the *Characteristics of Individuals* domain, experts highlighted varied needs across professional roles (eg, dietitians often engage after medication selection) and noted that older adults or patients with low digital literacy may struggle with line graphs or icon arrays. From their clinical experience, patients most frequently prioritize cost, expected effectiveness, and administration route. Under the *Process* domain, experts recommended enabling patients to access OptiWeight before or during clinic visits using a QR code, with shared tablet viewing during consultations. They suggested providing printable or PDF summaries for home review, along with basic training for nurses and case managers. Evaluation metrics should capture both patient and clinician satisfaction as well as time burden. Experts also emphasized the need for a mechanism to update medication information regularly and suggested expanding the tool to include lifestyle, physical activity, and surgical options in the future, with potential customization by age or comorbidities. Representative quotes and design modifications are summarized in Table 4.

XSL•FO
RenderX

Table 4. Expert interviews, insights, and resulting design modifications. This table lists issues that led to concrete design changes.

Issue and illustrative quotes	Modification (with screenshot reference)
Visualization of risk information	
• "Repetitive information; excessive content on the page" • "It looks like showing risk instead of treatment effect."	• Removed duplicated charts describing risks and side effects, retaining only the icon array. • Reframed the section title from "Risk" to "Medication Effects," adding a percentage label such as "Risk reduction: X%" to clarify that the visualization represents the benefits of treatment.
Drug recognition and information	
• "Line graphs contained too many lines, impairing comprehension; unclear purpose of the placebo curve" • "Y-axis scale too long; users wanted clearly marked treatment duration and effect size" • "It only says the unit price. Patients don't know how long one box lasts or how much per month." • "Patient would like to know the total cost of one full treatment course, not just per injection." • "Experts recommended adding dosing frequency, dosage, and storage instructions" • "Need to clarify Orlistat's oily stool and its social impact" • "Saxenda should specify that treatment should stop if <5% weight loss after 6 months"	• Removed the placebo curve from line graphs to improve readability. • Shortened the y-axis time scale: replaced 52-week display with 4 nodes (3, 6, 9, and 12 months). Added clickable time point buttons, allowing users to view expected weight-loss percentage and total projected cost. • Added monthly cost estimates in addition to unit prices. • Expanded the medication details pop-up to include dosing frequency, dosage (for adults and adolescents), storage requirements, and an explicit stop-treatment rule. • Clarified that Orlistat's oily stool may affect social activities.
User guidance/onboarding	
• "I wasn't sure if I had entered the real operation or was still in the tutorial." • "The guidance pages felt too long—I got lost halfway."	• Converted the tutorial to embedded (inline) guidance, shortening the number of pages and removing nonessential instructional screens.
Features and interactivity	
• "It feels like the system just abandons me when I choose no treatment." • "This PDA focuses on medications; the title should reflect that more clearly." • "Some patients care a lot about specific side effects. It would be helpful if the system allowed comparison of medications with the same side effect." • "Recommend adding contraindication filtering (BMI, age, comorbidities, indications)." • "It feels like the system just abandons me when I choose no treatment."	• Updated the tool title from "Weight-Loss Treatment" to "Weight-Loss Medication Decision Aid" for clarity. • Added a side effect comparison feature for medications sharing the same adverse events. • Implemented intelligent contraindication filtering, marking medications unsuitable for the patient with grayscale labels. • Reworded "No treatment" to "Not using medication for now (lifestyle modification)" to improve clarity and reduce negative tone.

Stage 4: Clinical Evaluation

Content and Workflow of the Final OptiWeight

Following the previous stages, OptiWeight was refined and finalized as an interactive iPad-based PDA titled "Antiobesity Medication Decision Aid," displayed in Traditional Chinese to ensure regional applicability. The tool guides users through 6 sequential components that provide comparative information on AOMs, including treatment efficacy, side effects, administration routes, and out-of-pocket costs (Figure 2).

XSL•FO
RenderX

Figure 2. Layouts of the finalized OptiWeight patient decision aid evaluated in clinical usability testing.

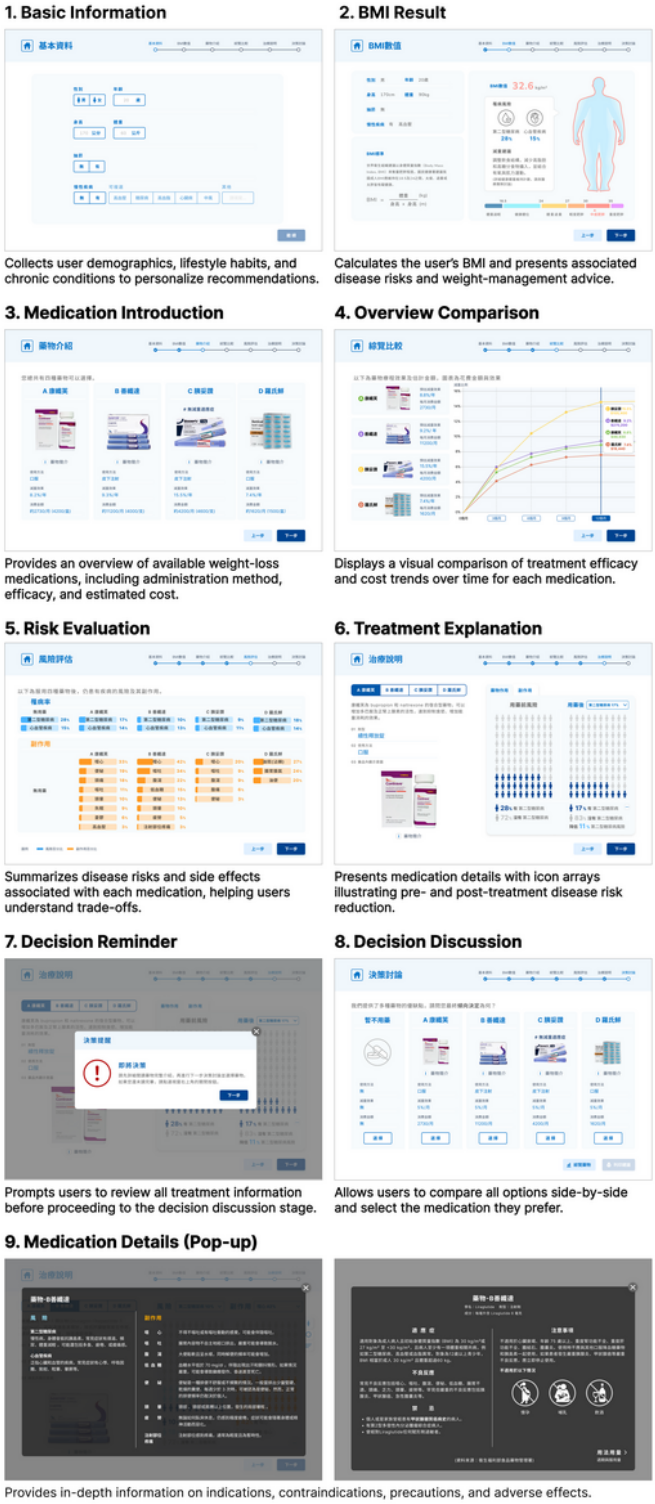

Participant Characteristics of Stage 4 Clinical Evaluation

A total of 78 patients with overweight or obesity receiving weight-management care participated in the clinical usability evaluation. Participants had a mean age of 45.0 (SD 11.1) years (range 22-74 years), were mainly female (n=45, 58%), had a college-level education or higher, and 54 out of 78 (69%) were classified as having moderate to severe obesity (BMI ≥30 kg/m^2; Table 3).

Usability and Cognitive Workload of Stage 4 Clinical Evaluation

The overall mean SUS score was 73.65 (SD 14.20), indicating "good" usability for the final OptiWeight prototype. The mean NASA-TLX score was 16.69 (SD 14.44), suggesting low perceived workload in the clinical context. Among the subscales,

XSL•FO
RenderX

mental demand and *effort* were rated comparatively higher (mean 22-29), whereas *physical demand*, *temporal demand*, and *frustration* remained low to moderately low. Collectively, these findings support that the final version of OptiWeight demonstrates acceptable usability and operational feasibility in outpatient practice.

Summary of Stage 4 Clinical Evaluation

Qualitative Evaluation Using Nielsen's Usability Heuristics

To enrich the quantitative findings, qualitative interviews were analyzed using Nielsen's 5 usability heuristics—learnability, efficiency, memorability, error tolerance, and satisfaction. This analysis examined whether the final OptiWeight prototype met real-world clinical usability needs.

Learnability

Participants demonstrated varying familiarity with AOM therapy. First-time users often needed clarification regarding medical terminology and the purpose of the tool. Participants emphasized the value of tooltips, embedded definitions, and clearer contextual guidance at the outset (eg, whether the tool is intended for self-learning or to support clinician-patient dialogue). These findings suggest that OptiWeight supports learnability with minimal guidance but could further benefit from concise instructional prompts and plain-language explanations.

Efficiency

Although participants appreciated the logical structure and clean layout, many sought faster access to key decision-relevant information. Suggestions included consolidating medication attributes and risk information into a single screen, enabling automated conversion of weight-loss percentages into kilograms, and ranking medication options by cost-effectiveness. Visualizations combining cost and efficacy were perceived as cognitively demanding, and separating these data elements was recommended to improve clarity.

Memorability

Participants described the interface as memorable due to its visual organization and intuitive iconography. However, small font sizes posed challenges for older adults. The ability to revisit information—via a dedicated "Back" button or a summary page—and to print a summary for later review with family or clinicians was viewed as essential for reinforcing comprehension and memory.

Error Tolerance

Technical errors were minimal, yet some participants recommended that all medication options remain visible with explicit explanations (eg, "Not recommended because BMI <27 kg/m^2"). The term "decision discussion" on the final screen was perceived as overly formal; alternatives such as "treatment preference summary" were preferred.

Satisfaction

Overall satisfaction was high. Participants described OptiWeight as clear, modern, and highly practical for understanding treatment options within minutes. Visual risk displays—especially those illustrating improvements in metabolic or cardiovascular risk—were seen as motivating. Participants recommended adding more engaging visuals (eg, brief animations or before-and-after illustrations) and suggested integrating lifestyle guidance or goal-setting features to contextualize medication effects within personal health goals.

Overall, the clinical evaluation demonstrated that OptiWeight was well accepted by patients and supported effective understanding of treatment options in real-world outpatient settings. Participants viewed the tool as intuitive and informative and provided actionable recommendations for enhancing accessibility and clarity across diverse patient populations. These findings underscore the clinical usability and practical feasibility of the final OptiWeight prototype.

Usability and Cognitive Workload Across Evaluation Stages

Across the iterative development process, the app interface was progressively refined between stage 2 usability testing (rounds 1 and 2), stage 3 expert interviews, and stage 4 clinical evaluation, with corresponding increases in usability and decreases in perceived workload (Table 5).

Table 5. System Usability Scale and NASA-Task Load Index scores across usability testing, expert evaluation, and clinical evaluation stages[a].

Measure	Usability testing round 1 (n=38)	Usability testing round 2 (n=40)	Expert interviews (n=15)	Clinical evaluation (n=78)
System Usability Scale total score, mean (SD)	60.53 (19.04)	66.25 (14.68)	77.83 (9.44)	73.65 (14.20)
NASA-Task Load Index total score, mean (SD)	40.35 (17.49)	41.13 (12.93)	28.11 (11.31)	16.69 (14.44)
Mental demand	64.21 (22.50)	63.25 (19.13)	39.67 (24.67)	22.50 (23.60)
Physical demand	19.47 (21.68)	24.50 (17.68)	14.33 (14.25)	9.68 (11.99)
Temporal demand	28.42 (25.74)	32.50 (23.83)	25.33 (24.67)	10.77 (13.51)
Performance	48.68 (28.21)	50.00 (21.12)	27.33 (25.97)	28.59 (26.16)
Effort	52.11 (23.73)	49.50 (18.25)	48.33 (26.37)	28.59 (26.35)
Frustration	29.21 (26.75)	27.00 (25.34)	13.67 (16.53)	13.27 (18.55)

[a]Round 1 and Round 2 scores are averaged across navigation prototypes (system- and user-controlled navigation).

XSL•FO
RenderX

Assumption checks indicated violations of normality (Shapiro-Wilk tests) and heterogeneity of variances for several variables; therefore, Welch ANOVA with Games-Howell post hoc comparisons was applied.

SUS scores demonstrated a clear upward trend, increasing from usability testing round 1 to round 2 and peaking during the expert evaluation stage (mean 77.83), while remaining relatively high during clinical evaluation (mean 73.65; Table 5). Welch ANOVA revealed a significant effect of evaluation stage (Welch $F_{3,58.13}$=8.54, P<.001, η^2=0.13; Table 6). Games-Howell post hoc comparisons indicated that round 1 scores were significantly lower than expert interview scores (mean difference −17.31, P<.001) and clinical evaluation scores (mean difference −13.13, P=.002). Round 2 scores were also significantly lower than expert interview scores (mean difference −11.58, P=.007) and did not differ significantly from clinical evaluation scores (mean difference −7.40, P=.05). No significant differences were observed between round 1 and round 2 (P=.45) or between expert interviews and clinical evaluation (P=.49).

Table 6. One-way analysis of variance results for System Usability Scale and NASA-Task Load Index scores across evaluation stages (round 1, round 2, expert interviews, and clinical evaluation)[a].

Measure	*F* test (*df*)	*P* value	η^2
System Usability Scale total score	8.46 (3, 167)	<.001	.13
NASA-Task Load Index total score	35.25 (3, 167)	<.001	.39
Mental demand (NASA_1)	43.81 (3, 167)	<.001	.44
Physical demand (NASA_2)	8.31 (3, 167)	<.001	.13
Temporal demand (NASA_3)	12.73 (3, 167)	<.001	.19
Performance (NASA_4)	9.56 (3, 167)	<.001	.15
Effort (NASA_5)	11.67 (3, 167)	<.001	.17
Frustration (NASA_6)	6.34 (3, 167)	<.001	.10

[a]Groups included round 1 (n=38), round 2 (n=40), expert interviews (n=15), and clinical evaluation (n=78). Reported *F* test values and η^2 correspond to standard 1-way ANOVA. Welch ANOVA was additionally conducted to account for violations of homogeneity of variance, and consistent results were observed. Post hoc comparisons were conducted using Games-Howell tests.

NASA-TLX total scores exhibited a progressive decline across iterations (Table 5). Welch ANOVA revealed a significant effect of evaluation stage (Welch $F_{3,55.36}$=35.08, P<.001, η^2=0.39; Table 6). Games-Howell post hoc tests showed that both round 1 and round 2 had significantly higher workload than the expert evaluation stage (round 1: mean difference 12.24, P=.02; round 2: mean difference 13.01, P=.005) and the clinical evaluation stage (round 1: mean difference 23.66, P<.001; round 2: mean difference 24.44, P<.001). Workload remained significantly higher in the expert evaluation stage than in the clinical evaluation stage (mean difference 11.42, P=.01). No significant difference was observed between round 1 and round 2 (P=.996). Similar trends were observed across NASA-TLX subscales, particularly for mental demand, effort, and temporal demand, suggesting a substantial reduction in cognitive workload across evaluation stages.

Discussion

Principal Findings

Using a human-centered design approach and iterative refinement, this study developed and validated OptiWeight, a PDA designed to support AOM decisions in Taiwanese clinical settings. Across the development cycle, usability increased across stages: SUS scores rose from mid-range levels during early prototype testing to good usability in the expert evaluation stage and remained high during real-world clinical use. Concurrently, cognitive workload decreased markedly, with NASA-TLX scores declining from the prototype stage to expert testing and clinical evaluation, suggesting progressive improvements associated with iterative refinement. These findings are best interpreted as reflecting prototype maturation across development stages; however, variations in participant populations—including general adults, clinical experts, and patients—may also have influenced the observed outcomes.

Qualitative findings further elucidated why these usability improvements occurred. Participants reported that combining numerical and visual formats enhanced comprehension, while a hybrid navigation structure—balancing guided flow with flexible browsing—supported more efficient information seeking and reduced cognitive load. The final clinical assessment suggested that OptiWeight is intuitive and acceptable to users, and may support informed patient-clinician discussions about AOMs from a usability perspective.

Interpretation and Comparison With Prior Work

Most existing PDAs for weight management have focused on decisions regarding bariatric surgery or comprehensive weight-management programs. For example, Arterburn et al [23] developed a tool to help patients decide whether to pursue bariatric surgery, whereas Moore et al [21] compared intensive lifestyle management with surgical pathways. Lee and Wu [24] designed a PDA to support choices among different bariatric procedures. Chu et al [59] extended PDA applications further upstream by using an artificial intelligence–based system to guide patients among 3 broad weight-management routes: self-directed lifestyle adjustments, dietitian-assisted programs, or physician-led treatment. By contrast, OptiWeight addresses

XSL•FO
RenderX

a distinct and underexplored decision point—choosing among multiple medication options. As new-generation medications such as glucagon-like peptide-1 receptor agonists increasingly diversify available choices, patients must weigh differences in efficacy, safety profiles, routes of administration, and out-of-pocket costs. OptiWeight fills this gap by providing a structured, localized comparison framework tailored to Taiwanese clinical practice.

Patient preferences are known to play a central role in adherence and clinical outcomes in weight management [8]. Decisions between obesity medications are influenced by perceived effectiveness, access to information, safety concerns, and the practicality of use [10]. To our knowledge, OptiWeight is the first PDA designed specifically to support preference-sensitive decisions among medication options, enabling patients to systematically evaluate trade-offs and clarify personal priorities—such as maximizing weight-loss efficacy versus minimizing gastrointestinal side effects—potentially supporting more informed, value-concordant treatment choices.

While prior PDA research has primarily emphasized single-point clinical evaluations of tool effectiveness [23], only a few studies have detailed how user feedback is translated into design refinements across iterative cycles. Although Moore et al [21] and Lee and Wu [24] reported aspects of their development processes, published descriptions focused mainly on feasibility and acceptability rather than the iterative mechanisms by which interface design and content were adjusted. By contrast, our study used a multistage, iterative human-centered design process that systematically linked prototype testing, expert interviews, and clinical usability assessment. This approach provides a transparent and reproducible development pathway for digital PDAs and highlights the importance of usability and workflow integration as parallel design goals, addressing a gap in the literature that often privileges outcome evaluation over process evidence. This aligns with user-centered design principles emphasizing continuous cycles of understanding, designing, evaluating, and refining to ensure that content and interaction patterns match real-world contexts and users' mental models [28,29].

Using the widely accepted SUS [60,61], OptiWeight demonstrated substantial usability improvement across iterations—from an early score of 60.53 to 73.65 in the final clinical evaluation. According to established SUS benchmarks, scores above 70 represent a "passable product" [44], scores above 71.4 indicate "good" usability [62], and scores above 80.8 reflect "excellent" usability [63]. The improvement observed across rounds is consistent with the iterative, user-centered refinement process and highlights SUS as a practical index for guiding PDA development toward clinical adoption and user satisfaction [64].

We also evaluated cognitive workload using the NASA Task Load Index (NASA-TLX) [45]. Prior work indicates that typical mean NASA-TLX scores for cognitive and medical tasks fall below 54 [65]. OptiWeight's final clinical score of 16.69 reflects a substantially lower workload, which may be associated with interface refinements, although differences in participant characteristics across stages may also have contributed to this trend. Cognitive Load Theory [66] distinguishes intrinsic, extraneous, and germane cognitive load; our findings highlight the importance of minimizing extraneous load attributable to poor interface design. By streamlining information presentation, simplifying navigation structure, and reducing unnecessary cognitive effort, OptiWeight is designed to help users focus on essential content without overwhelming their working memory capacity.

User-testing findings further informed several concrete design choices. Initially, we explored system-controlled versus user-controlled navigation structures but ultimately converged on a hybrid navigation architecture—a system-controlled backbone providing predictability and orientation cues, combined with flexible, user-controlled browsing within medication sections. This approach aligns with established information architecture principles, including focused navigation, multiple classification, and meaningful choice [64,67,68], while supporting chunking to accommodate limited working memory capacity [69,70].

In visual risk and benefit communication, OptiWeight integrates numeric and icon-based representations, consistent with evidence suggesting that enclosing visual elements within structured frames enhances comprehension [71] and that highlighting treatment-placebo differences reduces users' mental computation demands and cognitive load [57,72]. Our iterative refinements included improving label consistency, ensuring equivalent visibility across medication categories, adopting plain-language descriptions [54], and implementing color-contrast accessibility considerations for users with color vision deficiencies [71]. Universal back-navigation and error-recovery pathways were also strengthened to reduce interaction uncertainty and enhance learnability [73].

Finally, the effectiveness of PDAs heavily depends on cultural relevance and contextual fit [74]. Without adequate cultural adaptation [75], tool applicability across populations is limited. IPDAS similarly emphasizes localization of language and content [30]. Existing weight-management PDAs reflect their respective health system contexts—for example, Arterburn et al's [23] tool incorporates elements specific to the US insurance system, while Moore et al's [21] PDA targets adolescents, whose developmental and family decision-making dynamics differ markedly from adults. In Taiwan's unique context—characterized by a single-payer national health insurance system, universal out-of-pocket costs for AOMs, and distinct clinical workflows—local adaptation is essential. Expert interviews highlighted strong clinical demand and perceived market potential for such tools; transparent cost presentation, clear availability labeling, and workflow integration across multidisciplinary roles were viewed as key factors for adoption and equity.

Strengths and Limitations

This study has several strengths, including the use of a multistage human-centered design process, triangulation of quantitative and qualitative data, and validation across prototype, expert, and clinical settings. However, several limitations should be noted. First, to maintain temporal consistency during development, we did not include semaglutide (Wegovy) or

XSL•FO
RenderX

tirzepatide (Mounjaro), which were approved by the TFDA during or after the study design and implementation period. Second, this work focused on usability and clinical feasibility; decision quality outcomes—such as patient knowledge, decisional conflict, and value concordance—were not assessed, nor were medium- or long-term clinical end points. Third, different participant populations were involved at each stage, and observed differences in usability and cognitive workload may partially reflect population characteristics rather than solely the effects of iterative design refinements. Future studies should consider randomized or staged implementation designs that incorporate broader evaluation metrics, including decision quality, consultation time, and team workload, alongside longitudinal behavioral and clinical outcomes such as medication adherence, weight trajectories, metabolic indicators, medication switching or discontinuation events, and sustained lifestyle changes. In addition, expanding support for multiple device types and integrating OptiWeight with electronic health record systems would further clarify its scalability, interoperability, and potential to promote equitable access in real-world settings.

Conclusions

This study developed OptiWeight, a digital PDA with a patient-centered, interactive design to support adults with overweight or obesity in making preference-sensitive choices among AOMs. By presenting locally relevant, easy-to-understand comparisons of medication efficacy, side effects, administration routes, and out-of-pocket costs within the constraints of routine outpatient care, OptiWeight provides structured and accessible support for comparing AOM options, which may facilitate patient understanding and consideration of treatment choices. Through multicycle design and validation—including iterative prototype testing, expert interviews, and on-site clinical evaluation—we operationalized international PDA development principles into a functional interactive tool and workflow, offering a replicable blueprint for future digital PDA development. Qualitative findings further indicated that visualized risk information was most effective when paired with numerical and contextual explanations, and that hybrid navigation helped balance cognitive load with flexible comparison. Future work will focus on evaluating real-world implementation, creating alternative formats to enhance accessibility, and establishing a sustainable update strategy to support equitable, evidence-based AOM services in Taiwan.

Acknowledgments

During the preparation of this manuscript, the authors used generative artificial intelligence (ChatGPT; OpenAI) to assist with language editing, clarity, and formatting of the text. The artificial intelligence tool was not used to generate original scientific content, analyze data, or interpret results. All content was reviewed and verified by the authors, who take full responsibility for the manuscript.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to ethical and privacy considerations, as they may contain potentially identifiable information. Deidentified data, including quantitative datasets and qualitative coding materials, are available from the corresponding author (MCZ) upon reasonable request.

Funding

This work was supported by the National Science and Technology Council (NSTC), Taiwan, under Grants MOST 111-2221-E-027-060 and NSTC 112-2221-E-072-089. The funders had no role in the study design, data collection, analysis, interpretation, or writing of the manuscript.

Authors' Contributions

Conceptualization: LJW, MCZ
Data curation: LJW
Formal analysis: LJW, MCZ, YJW
Investigation: LJW, YJW
Methodology: LJW, MCZ, YJW
Project administration: LJW
Resources: YJW, YLC, WLF, WW
Software: LJW, MCZ
Supervision: MCZ
Writing – original draft: LJW
Writing – review & editing: MCZ, YJW

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
GRAMMS (Good Reporting of a Mixed Methods Study) checklist.
[DOCX File , 18 KB - jmir_v28i1e89428_app1.docx]

Multimedia Appendix 2
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[DOCX File , 22 KB - jmir_v28i1e89428_app2.docx]

References

1. World Obesity Atlas 2024. World Obesity Federation. 2022. URL: https://data.worldobesity.org/publications/?cat=22 [accessed 2025-01-04]
2. Nutrition and Health Survey in Taiwan (NAHSIT). Health Promotion Administration (Ministry of Health and Welfare). 2025. URL: https://www.hpa.gov.tw/EngPages/Detail.aspx?nodeid=1077&pid=6201 [accessed 2025-01-04]
3. 2025 HPA Annual Report. Health Promotion Administration (Ministry of Health and Welfare, Taiwan). URL: https://www.hpa.gov.tw/Pages/Detail.aspx?nodeid=5015&pid=19657 [accessed 2026-01-04]
4. Statistics of health promotion 2025. Health Promotion Administration (Ministry of Health and Welfare, Taiwan). Taipei, Taiwan: Health Promotion Administration, Ministry of Health and Welfare URL: https://www.hpa.gov.tw/Pages/Detail.aspx?nodeid=3850&pid=19671 [accessed 2025-12-10]
5. Obesity and overweight. World Health Organization (WHO). 2021 Jun 09. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [accessed 2025-01-04]
6. Heymsfield SB, Wadden TA. Mechanisms, pathophysiology, and management of obesity. N Engl J Med 2017 Jan 19;376(3):254-266. [doi: 10.1056/nejmra1514009]
7. Tchang BG, Aras M, Kumar RB, Aronne LJ. Pharmacologic treatment of overweight and obesity in adults. In: Feingold KR, Adler RA, Ahmed SF, Anawalt B, Blackman MR, Chrousos G, et al, editors. Endotext. South Dartmouth, MA: MDText.com; 2000.
8. Craig HC, Alsaeed D, Norris S, Holian J, Kennedy C, Feldman A, et al. Patient perspectives about treatment preferences for obesity with complications. Obes Sci Pract 2024 Feb 29;10(1):e720 [FREE Full text] [doi: 10.1002/osp4.720] [Medline: 38263990]
9. Donnan J, Huang R, Twells L. Patient preferences for attributes of Health Canada approved weight loss medications among adults living with obesity in Canada: a qualitative study. PPA 2022 Apr;Volume 16:911-923. [doi: 10.2147/ppa.s350411]
10. Mondoh A, Craig H, Crotty M, Contreras F, le Roux CW. How do patients choose between obesity medications: a thematic analysis. Obes Pillars 2025 Sep;15:100187. [doi: 10.1016/j.obpill.2025.100187] [Medline: 40673265]
11. Elwyn G, Durand MA, Song J, Aarts J, Barr PJ, Berger Z, et al. A three-talk model for shared decision making: multistage consultation process. BMJ 2017 Nov 06;359:j4891. [doi: 10.1136/bmj.j4891] [Medline: 29109079]
12. What are patient decision aids? International Patient Decision Aid Standards Collaboration. URL: http://ipdas.ohri.ca/what.html [accessed 2025-01-04]
13. Stacey D, Légaré F, Lewis K, Barry MJ, Bennett CL, Eden KB, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev 2017 Dec 12;4:CD001431. [doi: 10.1002/14651858.CD001431.pub5] [Medline: 28402085]
14. Gibson B, Butler J, Doyon K, Ellington L, Bray BE, Zeng Q. Veterans like me: formative evaluation of a patient decision aid design. J Biomed Inform 2017 Jul;71S:S46-S52 [FREE Full text] [doi: 10.1016/j.jbi.2016.09.007] [Medline: 27623534]
15. Yen RW, Smith J, Engel J, Muscat DM, Smith SK, Mancini J, et al. A systematic review and meta-analysis of patient decision aids for socially disadvantaged populations: update from the International Patient Decision Aid Standards (IPDAS). Med Decis Making 2021 Oct;41(7):870-896 [FREE Full text] [doi: 10.1177/0272989X211020317] [Medline: 34151614]
16. Wu T, Majeed A, Kuo KN. An overview of the healthcare system in Taiwan. London J Prim Care (Abingdon) 2010 Dec 07;3(2):115-119 [FREE Full text] [doi: 10.1080/17571472.2010.11493315] [Medline: 25949636]
17. Lin C, Lo H, Yang Y, Sim CJ, Chen B, Chen W. Perception of obesity and weight-loss attitudes among adolescents living with obesity, caregivers, and healthcare professionals in Taiwan: data from the ACTION Teens survey study. Pediatr Neonatol 2026 Jan;67(1):24-31 [FREE Full text] [doi: 10.1016/j.pedneo.2025.04.003] [Medline: 40467432]
18. Leu J, Huang K, Chen P, Pan W. Healthcare service providers' perspectives on sociocultural aspects affecting weight management activities amongst people with obesity in Taiwan—a qualitative study. Nutrients 2024 May 20;16(10):1540 [FREE Full text] [doi: 10.3390/nu16101540] [Medline: 38794778]
19. Lin Y, Lin C, Saffari M, Tsai M, Chang Y, Strong C, et al. Weight stigma is associated with body mass index among college students in Taiwan: the mediated role of internalized weight stigma. BMC Psychol 2023 Nov 01;11(1):365 [FREE Full text] [doi: 10.1186/s40359-023-01414-w] [Medline: 37915074]
20. Kreng VB, Yang C. The equality of resource allocation in health care under the National Health Insurance System in Taiwan. Health Policy 2011 May;100(2-3):203-210. [doi: 10.1016/j.healthpol.2010.08.003] [Medline: 21112116]

XSL•FO
RenderX

21. Moore J, Haemer M, Mirza N, Z Weatherall Y, Han J, Mangarelli C, et al. Pilot testing of a patient decision aid for adolescents with severe obesity in US pediatric weight management programs within the COMPASS Network. Int J Environ Res Public Health 2019 May 20;16(10):1776 [FREE Full text] [doi: 10.3390/ijerph16101776] [Medline: 31137491]
22. Nijland LMG, Noordman PCW, Boehlé L, van Veen RN, Bonjer HJ, de Castro SMM. A decision aid to help patients make informed choices between the laparoscopic gastric bypass or sleeve gastrectomy. Obes Surg 2023 Feb 26;33(2):562-569 [FREE Full text] [doi: 10.1007/s11695-022-06418-w] [Medline: 36571581]
23. Arterburn DE, Westbrook EO, Bogart TA, Sepucha KR, Bock SN, Weppner WG. Randomized trial of a video-based patient decision aid for bariatric surgery. Obesity (Silver Spring) 2011 Aug 06;19(8):1669-1675 [FREE Full text] [doi: 10.1038/oby.2011.65] [Medline: 21475138]
24. Lee Y, Wu W. Shared decision making and choice for bariatric surgery. Int J Environ Res Public Health 2019 Dec 06;16(24):4966. [doi: 10.3390/ijerph16244966] [Medline: 31817804]
25. Obesity: should I take weight-loss medicine? Healthwise. 2022 Aug 25. URL: https://www.healthwise.net/ohridecisionaid/Content/StdDocument.aspx?DOCHWID=aa71898 [accessed 2025-08-25]
26. I want to lose weight, which weight loss drug should I choose? Kaohsiung Veterans General Hospital. URL: https://org.vghks.gov.tw/sdm/News_Content.aspx?n=CEA978E2996359F9&sms=375CE464065A936E&s=516DBBF367DDEDA2 [accessed 2025-08-25]
27. ISO. URL: https://www.iso.org/standard/77520.html [accessed 2026-05-05]
28. Witteman HO, Maki KG, Vaisson G, Finderup J, Lewis KB, Dahl Steffensen K, et al. Systematic development of patient decision aids: an update from the IPDAS Collaboration. Med Decis Making 2021 Jun 19;41(7):736-754. [doi: 10.1177/0272989x211014163]
29. Vaisson G, Provencher T, Dugas M, Trottier M, Chipenda Dansokho S, Colquhoun H, et al. User involvement in the design and development of patient decision aids and other personal health tools: a systematic review. Med Decis Making 2021 Mar 03;41(3):261-274. [doi: 10.1177/0272989x20984134]
30. Stacey D, Volk RJ. The International Patient Decision Aid Standards (IPDAS) Collaboration: evidence update 2.0. Med Decis Making 2021 Aug 20;41(7):729-733. [doi: 10.1177/0272989x211035681]
31. Joseph-Williams N, Newcombe R, Politi M, Durand M, Sivell S, Stacey D, et al. Toward minimum standards for certifying patient decision aids. Med Decis Making 2013 Aug 20;34(6):699-710. [doi: 10.1177/0272989x13501721]
32. Hoffman AS, Volk RJ, Saarimaki A, Stirling C, Li LC, Härter M, et al. Delivering patient decision aids on the internet: definitions, theories, current evidence, and emerging research areas. BMC Med Inform Decis Mak 2013 Nov 29;13(S2):S13. [doi: 10.1186/1472-6947-13-s2-s13]
33. Sepucha KR, Abhyankar P, Hoffman AS, Bekker HL, LeBlanc A, Levin CA, et al. Standards for universal reporting of patient decision aid evaluation studies: the development of SUNDAE Checklist. BMJ Qual Saf 2018 May;27(5):380-388 [FREE Full text] [doi: 10.1136/bmjqs-2017-006986] [Medline: 29269567]
34. Ministry of Health and Welfare health risk assessment platform for chronic diseases. Health Promotion Administration (Ministry of Health and Welfare). URL: https://cdrc.hpa.gov.tw/hra-openservice-menupage.jsp?all [accessed 2025-01-04]
35. Pi-Sunyer X, Astrup A, Fujioka K, Greenway F, Halpern A, Krempf M, et al. A randomized, controlled trial of 3.0 mg of liraglutide in weight management. N Engl J Med 2015 Jul 02;373(1):11-22. [doi: 10.1056/nejmoa1411892]
36. Marso SP, Daniels GH, Brown-Frandsen K, Kristensen P, Mann JF, Nauck MA, et al. Liraglutide and cardiovascular outcomes in type 2 diabetes. N Engl J Med 2016 Jul 28;375(4):311-322. [doi: 10.1056/nejmoa1603827]
37. Marso SP, Bain SC, Consoli A, Eliaschewitz FG, Jódar E, Leiter LA, et al. Semaglutide and cardiovascular outcomes in patients with type 2 diabetes. N Engl J Med 2016 Nov 10;375(19):1834-1844. [doi: 10.1056/nejmoa1607141]
38. Wilding JP, Batterham RL, Calanna S, Davies M, Van Gaal LF, Lingvay I, et al. Once-weekly semaglutide in adults with overweight or obesity. N Engl J Med 2021 Mar 18;384(11):989-1002. [doi: 10.1056/nejmoa2032183]
39. Perreault L, Davies M, Frias J, Laursen P, Lingvay I, Machineni S, et al. Changes in glucose metabolism and glycemic status with once-weekly subcutaneous semaglutide 2.4 mg among participants with prediabetes in the STEP program. Diabetes Care 2022 Oct 01;45(10):2396-2405 [FREE Full text] [doi: 10.2337/dc21-1785] [Medline: 35724304]
40. Torgerson J, Hauptman J, Boldrin M, Sjöström L. XENical in the prevention of diabetes in obese subjects (XENDOS) study: a randomized study of orlistat as an adjunct to lifestyle changes for the prevention of type 2 diabetes in obese patients. Diabetes Care 2004 Jan;27(1):155-161. [doi: 10.2337/diacare.27.1.155] [Medline: 14693982]
41. Bays HE, Fitch A, Christensen S, Burridge K, Tondt J. Anti-obesity medications and investigational agents: an Obesity Medicine Association (OMA) Clinical Practice Statement (CPS) 2022. Obes Pillars 2022 Jun;2:100018 [FREE Full text] [doi: 10.1016/j.obpill.2022.100018] [Medline: 37990711]
42. Lazar J, Feng J, Hochheiser H. Research Methods in Human-Computer Interaction. Cambridge, MA: Morgan Kaufmann; 2017.
43. Lundgrén-Laine H, Salanterä S. Think-aloud technique and protocol analysis in clinical decision-making research. Qual Health Res 2010 Apr;20(4):565-575. [doi: 10.1177/1049732309354278] [Medline: 19959822]
44. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the System Usability Scale. International Journal of Human–Computer Interaction 2008 Jul 30;24(6):574-594. [doi: 10.1080/10447310802205776]

XSL•FO
RenderX

45. Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Hancock PA, Meshkati N, editors. Advances in Psychology. Amsterdam, Netherlands: North-Holland; 1988:139-183.
46. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
47. Wang L, Zheng M. The information architecture of PDAs--its impact on user comprehension and risk perception. 2023 Presented at: The 14th International Conference on Applied Human Factors and Ergonomics (AHFE 2023); July 20-24, 2023; San Francisco, CA.
48. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009;4:50 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
49. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated Consolidated Framework for Implementation Research based on user feedback. Implement Sci 2022 Oct 29;17(1):75 [FREE Full text] [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]
50. Boateng J, Lee CN, Foraker RE, Myckatyn TM, Spilo K, Goodwin C, et al. Implementing an electronic clinical decision support tool into routine care: a qualitative study of stakeholders' perceptions of a post-mastectomy breast reconstruction tool. MDM Policy Pract 2021 Sep 17;6(2):23814683211042010 [FREE Full text] [doi: 10.1177/23814683211042010] [Medline: 34553067]
51. Evidence-based guidelines for adult obesity prevention and management. Health Promotion Administration (Ministry of Health and Welfare). URL: https://www.hpa.gov.tw/Pages/EBook.aspx?nodeid=1788 [accessed 2026-01-04]
52. Garcia-Retamero R, Cokely ET. Designing visual aids that promote risk literacy: a systematic review of health research and evidence-based design heuristics. Hum Factors 2017 Jun;59(4):582-627. [doi: 10.1177/0018720817690634] [Medline: 28192674]
53. Trevena LJ, Zikmund-Fisher BJ, Edwards A, Gaissmaier W, Galesic M, Han PKJ, et al. Presenting quantitative information about decision outcomes: a risk communication primer for patient decision aid developers. BMC Med Inform Decis Mak 2013;13 Suppl 2(Suppl 2):S7 [FREE Full text] [doi: 10.1186/1472-6947-13-S2-S7] [Medline: 24625237]
54. Zikmund-Fisher BJ, Witteman HO, Dickson M, Fuhrel-Forbis A, Kahn VC, Exe NL, et al. Blocks, ovals, or people? Icon type affects risk perceptions and recall of pictographs. Med Decis Making 2013 Nov 18;34(4):443-453. [doi: 10.1177/0272989x13511706]
55. Galesic M, Garcia-Retamero R, Gigerenzer G. Using icon arrays to communicate medical risks: overcoming low numeracy. Health Psychol 2009 Mar;28(2):210-216. [doi: 10.1037/a0014474] [Medline: 19290713]
56. Garcia-Retamero R, Galesic M. Who profits from visual aids: overcoming challenges in people's understanding of risks [corrected]. Soc Sci Med 2010 Apr;70(7):1019-1025. [doi: 10.1016/j.socscimed.2009.11.031] [Medline: 20116159]
57. Wang LJ, Zheng MC. Balancing layout space and risk comprehension in health communication: a comparison of separated and integrated icon arrays. Informatics 2025 Sep 30;12(4):105. [doi: 10.3390/informatics12040105]
58. Wang LJ, Zheng MC. The effective interface format for communicating risk information in shared decision making. In: Advances in Usability, User Experience, Wearable and Assistive Technology. Berlin, Germany: Springer International Publishing; 2021:854-860.
59. Chu Y, Huang R, Chen TT, Lin W, Tang JT, Lin C, et al. Effect of health literacy and shared decision-making on choice of weight-loss plan among overweight or obese participants receiving a prototype artificial intelligence robot intervention facilitating weight-loss management decisions. Digit Health 2022;8:20552076221136372 [FREE Full text] [doi: 10.1177/20552076221136372] [Medline: 36353693]
60. Brooke J. SUS: a retrospective. Journal of Usability Studies 2013;8(2):29-40. [doi: 10.5555/2817912.2817913]
61. Brooke J. SUS-A quick and dirty usability scale. Usability evaluation in industry 1996;189(194):4-7. [doi: 10.1201/9781498710411-35]
62. Bangor A, Kortum PT, Miller JT. Determining what individual SUS scores mean: adding an adjective rating scale. Journal of Usability Studies archive 2009;4:114-123. [doi: 10.5555/2835587.2835589]
63. Sauro J, Lewis JR. Chapter 4 - Did we meet or exceed our goal? In: Quantifying the User Experience (Second Edition). Cambridge, MA: Morgan Kaufmann; 2016:39-60.
64. Nielsen J. 10 usability heuristics for user interface design. Nielsen Norman Group. 1994. URL: https://www.nngroup.com/articles/ten-usability-heuristics/ [accessed 2025-01-04]
65. Grier RA. How high is high? A meta-analysis of NASA-TLX global workload scores. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2016 Dec 20 Presented at: Human Factors and Ergonomics Society Annual Meeting; October 26-30, 2015; Los Angeles, CA p. 1727-1731. [doi: 10.1177/1541931215591373]
66. Sweller J. Cognitive load during problem solving: effects on learning. Cognitive Science 2010 Feb 11;12(2):257-285. [doi: 10.1207/s15516709cog1202_4]
67. The 8 principles of information architecture. UX24/7. 2018. URL: https://ux247.com/the-8-principles-of-information-architecture/?via%3Dihub [accessed 2025-01-04]
68. Rosenfeld L, Morville P, Arango J. Information Architecture: For the Web and Beyond. Sebastopol, CA: O'Reilly Media; 2015.

XSL•FO
RenderX

69. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014 Sep;96(3):395-403 [FREE Full text] [doi: 10.1016/j.pec.2014.05.027] [Medline: 24973195]
70. Padilla LM, Creem-Regehr SH, Hegarty M, Stefanucci JK. Decision making with visualizations: a cognitive framework across disciplines. Cogn Res Princ Implic 2018 Dec 11;3(1):29. [doi: 10.1186/s41235-018-0120-9] [Medline: 30238055]
71. Johnson JA, Itiola A, Rahman S, Smith C, Soprovich A, Wozniak LA, et al. Exploring patient perspectives on EQ-5D-5L data visualization within an individualized decision aid for total knee arthroplasty (TKA) in Alberta, Canada. BMC Musculoskelet Disord 2024 Feb 29;25(1):183. [doi: 10.1186/s12891-024-07304-5] [Medline: 38424523]
72. Tiede KE, Ripke F, Degen N, Gaissmaier W. When does the incremental risk format aid informed medical decisions? The role of learning, feedback, and number of treatment options. Med Decis Making 2020 Feb 08;40(2):212-221. [doi: 10.1177/0272989x20904357]
73. Norman D. The Design of Everyday Things: Revised and Expanded Edition. New York, NY: Basic Books; 2013.
74. Alden DL, Friend J, Schapira M, Stiggelbout A. Cultural targeting and tailoring of shared decision making technology: a theoretical framework for improving the effectiveness of patient decision aids in culturally diverse groups. Soc Sci Med 2014 Mar;105:1-8. [doi: 10.1016/j.socscimed.2014.01.002] [Medline: 24606791]
75. Chenel V, Mortenson W, Guay M, Jutai J, Auger C. Cultural adaptation and validation of patient decision aids: a scoping review. PPA 2018 Mar;Volume 12:321-332. [doi: 10.2147/ppa.s151833]

Abbreviations

ANOVA: analysis of variance
AOM: antiobesity medication
CFIR: Consolidated Framework for Implementation Research
COREQ: Consolidated Criteria for Reporting Qualitative Research
GRAMMS: Good Reporting of a Mixed Methods Study
IPDAS: International Patient Decision Aid Standards
ISO: International Organization for Standardization
PDA: patient decision aid
SDM: shared decision-making
SUS: System Usability Scale
TFDA: Taiwan Food and Drug Administration

Edited by A Stone; submitted 19.Jan.2026; peer-reviewed by O Rodriguez; comments to author 17.Feb.2026; revised version received 29.Apr.2026; accepted 30.Apr.2026; published 15.May.2026.

Please cite as:
Wang LJ, Wang YJ, Cheng YL, Fang WL, Wang W, Zheng MC
Digital Patient Decision Aid for Antiobesity Medications: Mixed Methods Study of Human-Centered Design and Usability Evaluation
J Med Internet Res 2026;28:e89428
URL: https://www.jmir.org/2026/1/e89428
doi:10.2196/89428
PMID:

©Li-Jen Wang, Yi-Jen Wang, Yu-Lun Cheng, Wen-Liang Fang, Weu Wang, Meng-Cong Zheng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Engagement Trends in Online Vaccine Content: Longitudinal YouTube Study

Michele Tizzani[1], PhD; Yelena Mejova[2], PhD

[1]Section for Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Richard Petersens Plads, Kgs. Lyngby, Capital Region, Denmark

[2]Institute for Scientific Interchange, Turin, Piedmont, Italy

Corresponding Author:
Michele Tizzani, PhD
Section for Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Richard Petersens Plads, Kgs. Lyngby, Capital Region, Denmark

Abstract

Background: YouTube is the primary global video platform, hosting both authoritative health information and vaccine-skeptic viewpoints. However, engagement dynamics remain poorly understood.

Objective: The aim of this study was to investigate the temporal and textual dynamics of engagement of the YouTube viewership with vaccination content, and specifically content that is in favor of or against vaccination. We contextualized these dynamics in the authority signals of the posting channel and the moderation actions taken by the platform.

Methods: We conducted a 6-month daily longitudinal analysis of 7213 vaccine-related YouTube videos (November 2024 to May 2025) mentioning vaccination. We used zero-shot large language model classification with manual verification to classify the video stance toward vaccination, and the stance of their comments toward the video. The engagement and disagreement dynamics were modeled using Bayesian regression.

Results: Our findings show engagement asymmetry between content supporting and questioning vaccination. Vaccine-hesitant videos in our sample receive substantially higher raw engagement (median likes: 40 [IQR 3-846] to 59 [IQR 3-1319]; median comments: 10 [IQR 0-160] to 18 [IQR 0-311] per video versus 3 [IQR 1-15] and 0 [IQR 0-4], respectively, for strongly provaccine content) and moderate normalized engagement rates (true median combined rate: 0.073 [IQR 0.028-0.121] to 0.069 [IQR 0.027-0.118] interactions per view versus 0.026 [IQR 0.007-0.060] for strongly pro-vaccine videos, a 2.5-2.6× difference). Descriptively, vaccine-hesitant videos reach 90% of cumulative views faster (18 [IQR 8-38] days vs 32 [IQR 18-64] days; 44% faster), while negative binomial models that adjust for total engagement volume indicate that approximately 20% of this advantage reflects genuine temporal compression independent of engagement volume. Comment analysis indicated that the vaccine-hesitant videos in our sample foster echo chambers, while the provaccine content attracts battlegrounds. Considering the sources of vaccine-related content, provaccine content tends to originate from organizations, particularly news and health institutions, while vaccine-hesitant discourse is more likely to come from individual creators, even those self-identifying as medical doctors. Moderation, on the rare occasion when it occurs (about 2% of the videos were taken down), comes after engagement saturation, limiting its effectiveness.

Conclusions: Our analysis suggests that the vaccine-hesitant content can dominate YouTube's engagement ecosystem through rapid early-stage amplification, which has direct implications for public health intervention timing and platform governance policy.

(J Med Internet Res 2026;28:e88519) doi:10.2196/88519

KEYWORDS

vaccine hesitancy; YouTube; engagement dynamics; longitudinal analysis; Bayesian modeling; echo chambers; misinformation; social media; temporal patterns; public health policy

Introduction

Despite saving millions of lives [1], vaccination has been a contentious public health issue, with hesitancy increasing globally since the COVID-19 outbreak [2,3]. In 2025, the debate around vaccination became embedded in government policy and electoral politics (especially in the United States) following the strong politicization of the topic during the pandemic [4]. While the acute phase of the pandemic has passed and the vaccination debate has been settled in the scientific community, the battle for public opinion is still underway, and it could reshape the long-standing public health policies. The discourse on vaccination is actively shaping the government's decisions,

XSL•FO
RenderX

and in the United States, political movements are gaining traction, promising to "Make America Healthy Again" and framing vaccination as a political stance. Already, the United States is re-evaluating its role as a supporter of vaccine-based medicine both worldwide and domestically. The consequences are in evidence: the largest measles outbreak—since its elimination status was achieved in 2000—has reached 1431 cases and 3 deaths by September 2025 [5].

This shifting policy landscape underscores the urgency of understanding online vaccine discourse. Social media platforms have become central battlegrounds in shaping public opinion, serving as both powerful tools for health communication and vectors for misinformation that can undermine public trust in vaccines [6,7]. Among these platforms, YouTube is a primary source of information for a global audience, making the quality and reception of its health content a matter of critical public concern. In 2025, the Pew Research Center found that the vast majority of US adults (84%) say they have ever used YouTube, and about half of them go on YouTube daily, and that it is the most widely used online platform among US teens [8,9].

Prior research has established that YouTube hosts a mix of credible and misleading vaccine-related content, but none do so longitudinally. Earlier, Basch et al [10] examined the 87 most widely viewed videos around vaccination safety and their creators, finding that 65.5% discouraged the use of vaccines. Then, in 2021, Tang et al [11] created recommendation networks of pro- and antivaccine videos and found that when the user begins their session by searching provaccine keywords, they are recommended more videos by governmental agencies, hospitals, and academic institutions than when users search for antivaccine keywords. More recently, Ng et al [12] "gamified" this analysis by asking real-world participants to intentionally find an antivaccine video when starting from the World Health Organization's posts, finding that less than 6% of the videos recommended by the algorithm were antivaccine. A most recent dataset—albeit focusing on COVID-19—was created by Jung et al [13] using sock-puppets. They found that 31.5% of the top 10 search results contained COVID-19 misinformation, including that about vaccination. As exemplified above, not only do the studies in the existing literature use heterogeneous research designs, (intentionally) biased queries, and crawling starting points, they are, above all, limited by their *cross-sectional* nature (also including manually annotated surveys such as [14]). They suggest that antivaccination content indeed exists on YouTube but fail to provide the full scope of posted vaccine-related content, and to capture the temporal dynamics of the actual users' interaction with it.

Nonetheless, within this methodologically disparate landscape, we find a troubling trend: misleading or vaccine-hesitant videos often attract significantly more engagement—in views, likes, and comments—than accurate, provaccine videos [10,11]. This "engagement asymmetry" aligns with broader findings that false news spreads faster and further than true news in online environments [15]. Although YouTube has a medical misinformation policy [16] that advises users not to post information that "contradicts health authority guidance" on the prevention and treatment of diseases, the stricter COVID-19 era policies have been phased out in 2023 [17]. This leaves critical policy-relevant questions unanswered: How quickly does vaccine-hesitant content capture audience attention compared to provaccine content? And what is the nature of the discourse these different types of content foster?

Controversy around vaccination has been shown to be reflected on social media, such as Facebook and Twitter [4,18,19]. Further, Gruzd et al [20] documented Facebook-to-YouTube pathways that contributed to the spread of antivaccination content and formed clusters of densely connected videos. These areas of discourse, wherein information favors a certain stance while other views are excluded, are indicative of echo chambers [21]. Note that the stance may be in favor or in opposition to some topic, such as vaccination, making possible echo chambers containing the majority opinion that disagrees with the medically established vaccination advice. Unfortunately, Gruzd et al [20] did not capture possible disagreements within the comment-level discourse. On the other hand, Ekram et al [22] found that comments on YouTube videos around the human papillomavirus vaccine often mentioned its potential side effects and conspiracy theories. Such spaces, wherein different or opposing points of view are discussed actively, can be viewed as a battleground or a controversy [23]. Unfortunately, Ekram et al [22] did not track the engagement temporally. In this work, we aim to combine quantitative, temporal, and qualitative text engagement signals to discover whether echo chambers or battlegrounds better characterize the vaccination debate on YouTube.

This dynamism is further complicated by the variety of actors posting health-related information on social media. Although public health institutions and researchers have joined the conversation, online influencers hold dominance by amassing vastly larger followings [24]. Such competition for attention is especially acute during emergencies, exemplified by the beginning of the COVID-19 pandemic, when political and commercial interests competed with the public health messaging [25]. The connection between the use of social media and higher vaccine hesitancy during this period has shown a robust negative impact on the vaccination efforts [26,27], which cost an estimated thousands of lives [28,29]. Given the high stakes, it is important to further understand the engagement dynamics around organizational accounts, in comparison to those of individuals who possibly have no medical credentials.

This study contributes a *unique systematic daily longitudinal analysis* of vaccine-related YouTube videos over a 6-month period. By tracking thousands of videos from the day of their upload through their engagement maturation, we provide a fine-grained temporal account of the online vaccine discourse. Unlike previous studies, which may confound engagement patterns with post hoc algorithmic amplification or viral surprises, this study focuses on the earliest phase of a video's lifecycle—the initial engagement that occurs within the first few days and weeks of upload, when creators' original reach and audience composition have the most direct influence. Our multifaceted approach combines scalable, large language model (LLM)–based classification of video stances and comment agreement with Bayesian statistical modeling to answer three key questions:

XSL•FO
RenderX

1. *Temporal dynamics*: How do the engagement trajectories of pro- and vaccine-hesitant videos differ over time?
2. *Discourse characterization*: Do these different types of videos cultivate confirmatory "echo chambers" or contentious "battlegrounds" in their comment sections?
3. *Source effects*: How do channel characteristics, such as topic and authority, moderate these discourse patterns?

We provide quantitative evidence that vaccine-hesitant content not only receives 10-fold higher raw engagement volume (and 3.1-fold higher per-view engagement rates) but also reaches its engagement peak faster than provaccine content (44% descriptively; 20% after controlling for engagement volume). Furthermore, through comment-level analysis, we demonstrate that this engagement follows distinct patterns: vaccine-hesitant videos foster agreeable echo chambers, while provaccine videos become contentious battlegrounds characterized by high levels of disagreement. These findings have significant implications for public health policy, platform governance, and science communication, highlighting the narrow window for effective intervention and the urgent need for context-aware communication strategies in an era of heightened vaccine policy contention.

Methods

Data Collection

The collection used in this study was created using a double-phase approach outlined in Figure 1. We adopt a "streaming" approach, wherein every day the collector takes 2 actions: it first uses a list of relevant keywords to search for videos published in the previous day and selects the videos that pass several filters for inclusion in the dataset, and then it periodically recollects the metadata about the selected videos.

Figure 1. Data collection pipeline overview: the daily collection and video selection (top) and periodic collection (bottom).

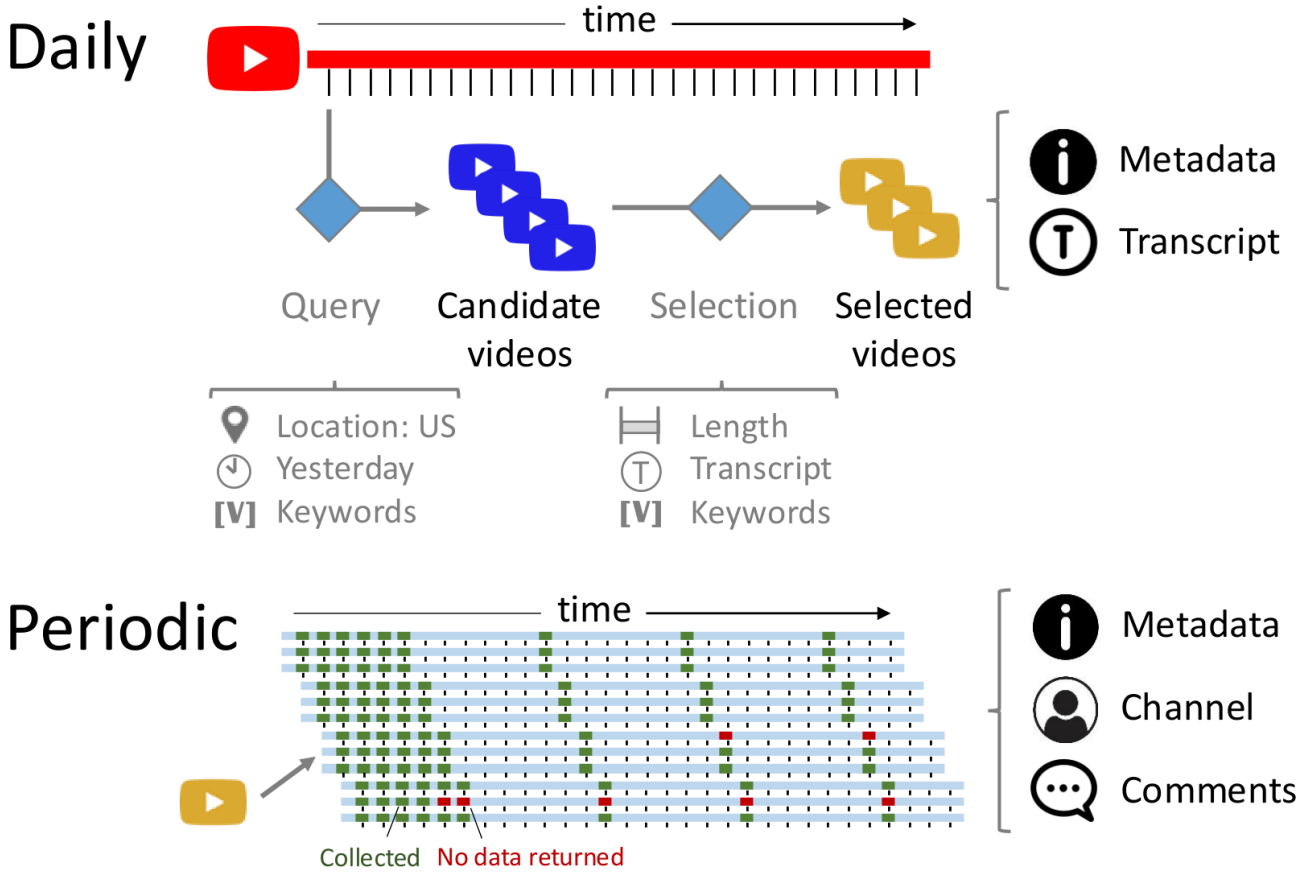

Specifically, we use the YouTube Search application programming interface (API) [30] to query for videos matching the following keyword list: *vaccine*, *vaccination*, *novax*, *provax*, *antivaxx*, *antivaxxer* (guided by previous work [4,19,31,32]). We restrict the time of collection to the previous day, the location to the "US," and the language to English. The results are sorted by relevance, and we save the top 500. Next, we use the YouTube Video API [33] to collect the metadata about the posted videos (note that it would be at most 1 day old), and the YouTube Transcript API library [34] to download their transcripts. These videos are then filtered through several selection steps to make sure they contain enough content for analysis and that they are relevant to the topic of vaccination. We require that the videos have a transcript, be at least 1 minute and at most 60 minutes in duration, and that at least 1 keyword from the query can be found in their title or transcript, or at least 3 keywords can be found in the transcript, if none exist in the title. This filter results in about 8.3% (7213/86,387) of the videos passing (mostly because so many videos are very short, the analysis of which we leave for future studies).

During the periodic phase of the collection, we requery the metadata information about the videos daily in order to compose a temporal picture of the audience's engagement with them, as well as to capture any potential moderation actions taken by the platform. We perform this requerying every day for the first

XSL•FO
RenderX

week after the video is initially retrieved, and following that, every week. After 3 months of collection, we consider whether in the past 2 time segments the counts of views, likes, or comments increase, and if they do not, we conclude the collection for those videos. During this collection, we also gather the 1000 most recent comments posted on that video using the comment Threads API, as well as the information about the poster of the video using the Channels API [35]. The final dataset spans 6 months (November 20, 2024, to May 20, 2025) and is composed of 7213 videos posted by 3872 unique channels, and their accompanying 911,219 comments. Of the 7213 videos initially collected, 302 were excluded due to missing channel-level metadata (eg, hidden subscriber counts) or engagement timestamps required for the regression models, resulting in a final analytical dataset of 6074 videos.

Content Characterization

Videos

As the focus of this study is to monitor the extent to which vaccine-hesitant content is present in such data, we endeavor to classify the collected videos according to their stance on vaccination [36]. We begin by extracting excerpts from the video transcripts, of length 100 words, centered around the query keywords, which we then label (if more than 20 are found in a transcript, we randomly select 20). We then use the zero-shot classification via GPT-4o mini (OpenAI) to assign one of the following labels to each of the excerpts: "in favor of vaccination," "against vaccination," or "not in favor or against vaccination," using the prompt shown in Section A.1 in Multimedia Appendix 1. This resulted in 55,310 annotated excerpts, an average of 8 per video. We validated this approach by manually annotating a random sample of 200 text segments. The agreement between the two authors of this study, computed on 60 segments, was near-perfect, at 96.7% (58/60) (Cohen κ=0.94), with 2 instances of disagreement between the "in favor of vaccine" and "not in favor or against vaccination" classes. The accuracy of the zero-shot classifier was estimated at 82.5% (165/200) (a marked improvement compared to 70% achieved by standard natural language processing tools evaluated in previous literature [37]). Per-class performance statistics can be seen in Table S1 in Multimedia Appendix 1. All of the errors were those with the middle label "neither" except for 4 instances where the disagreement was between the "against vaccination" and "in favor of vaccination," suggesting that the main source of error may be due to the strength of the expressed stance. Next, we propagate these labels to the videos by considering the mean of the excerpt scores for each video, using the following encoding: −1=against vaccination, 1=in favor of vaccination, and 0=neither in favor nor against. For the ease of interpretation and analysis, we categorize the continuous scores into the following: [0.5, 1]: "strongly in favor" (SIF); (0, 0.5): "in favor" (IF); [0]: "neutral" (N); (−0.5, 0): "against" (A); and [−1.0, −0.5]: "strongly against" (SA) vaccines. This way, we were able to annotate 6074 videos posted by 3789 distinct users. The number of videos and channels in each category can be seen in the Results section.

Comments

To characterize the discourse surrounding the videos, we annotate a sample of comments using a staged pipeline. We begin by filtering our video dataset to focus on the most engaged content, retaining only videos with at least 10 comments. From this filtered set, we randomly sample up to 50 comments per video for annotation. The sampled comments for each video are processed in a single batched API call to an LLM (GPT-4o mini). To determine the comment agreement with the video, the model is provided with the video's transcript and metadata as contextual reference (title, description, and transcript) and prompted to perform 2 analyses for each comment: first, to classify its agreement relative to the video's main claims ("agree," "disagree," or "neutral/unrelated"), and second, to determine its sentiment ("positive," "negative," or "neutral"). The model is also instructed to provide a brief rationale for each classification. The specific prompt used for this task is detailed in Section A.2 in Multimedia Appendix 1. Further, the agreement label was annotated by the authors on a sample of 100 comments. Interannotator agreement, computed on 20 overlapping items, was k=0.66, indicating substantial agreement [38].

Comparing the labels to those produced by the classifier, 73 labels match exactly, 17 disagree with a "neutral/unrelated" label, and the remaining 10 have the opposite disagreement ("agree" vs "disagree"). Upon manual examination, we find that the classifier often either misinterprets the stance of a long transcript or misses a contextual connection between mentioned parties that is not explicitly mentioned in the text. As the accuracy of the classifier is 90% if the "neutral" label is not considered an error, we proceed with the analysis, while keeping in mind the tool's limitations.

By using this zero-shot classification approach, we are able to scale our analysis to thousands of videos without relying on rigid keyword lists that might miss nuance. Conceptually, this method mimics human judgment regarding video stance and comment sentiment but applies it consistently across the entire dataset. This allows us to map the landscape of vaccine discourse with high granularity, ensuring that our distinction between "provaccine" and "vaccine-hesitant" content captures the true intent of the speaker rather than just the presence of specific medical terms.

Channel Characterization

To understand the source of the videos in our dataset, we characterize the channels using a two-stage process: (1) an automated, hierarchical classification of the video's primary topic for each of the 3789 channels in the dataset, and (2) a manual annotation of the channel's creator type and declared authority signals for a subset of 645 channels.

Hierarchical Topic Classification

A substantial proportion of channels (278/3789, 7.3%) have multiple topic assignments from YouTube's metadata, reflecting the inherent multidimensionality of vaccine discourse [4,6]. To resolve these ambiguities for categorical analysis, we implement a 2-level hierarchical classification approach grounded in established cascade methodology [39]. First, following cascade

XSL•FO
RenderX

principles [40,41], we use vaccine-related keywords to identify the broad topic category for each video, acknowledging that vaccine discourse cuts across multiple subjects [22]. Second, to resolve remaining multitopic conflicts, we apply a specificity prioritization rule: politics society health. This hierarchy is theoretically grounded in hypernym-hyponym relationships from taxonomy theory [39,42], ensuring that more specific and discriminative categories (hyponyms like politics) are retained over broader ones (hypernyms like society or health). This ensures that more specific, discriminative categories are preserved over broader ones, collapsing ambiguous assignments into a single primary topic while retaining critical information about content framing [39]. Details on the topic of co-occurrence statistics can be found in Table S2 in Multimedia Appendix 1.

Manual Annotation

For the second approach, we consider two samples of users: in the first are the channels that posted more than 3 videos in our dataset (339 channels), and the second sample includes channels that posted videos altogether totaling more than 10,000 views (451 channels). Combined, the two lists contain 645 unique channel IDs. The focus of the annotation is 2-fold: finding whether the channel associates itself with an organization or an individual, and what kinds of authoritative signals the channel projects. Such signals may include being a doctor, a news agency, or presenting expertise in the topics of health, science, or education. We also mark the channels that provide opinions, without mentioning other qualifications, and those centered around entertainment outside educational or scientific discourse. The page of each channel was examined and annotated manually. Seven channels were removed by the platform before they could be evaluated. Four were removed from further channel-related analysis due to a topical match, such as being about animals and mentioning animal vaccination.

By using both topical signals from YouTube and those from manual annotation, we provide a view of the authors of these videos both in terms of the narrative view of their presentation of the topic and in terms of their self-presentation as a possible authority on the subject.

Engagement Metrics

Video-Level Analysis

To examine engagement patterns at the video level, we aggregated all metadata and computed normalized engagement rates. For each video *v*, the like ratio (LR) and comment ratio (CR) were defined as:

LR_v = (video $likeCount_v$)/(video $viewCount_v$+1)

CR_v = (video $commentCount_v$)/(video $viewCount_v$+1)

with extreme values clipped to the interval [0.001, 0.999] to satisfy the support of a beta regression. Stance labels—SIF, IF, N, A, and SA—were one-hot encoded, using neutral as the reference category. Channel-level covariates included the natural logarithm of the subscribers to the channels (LogSub) and the square root of the channels' video counts (CHV):

$LogSub_v$ = ln($subscriberCount_v$)

CHV_v = videoCount

capturing audience reach and channel activity. Temporal exposure, measured in days since publication, controlled for the decay of engagement over time. These features served as predictors in our beta regression models of normalized engagement rates.

For clarity, we refer to the total counts of views, likes, and comments per video as raw interaction volume, and to the LR and CR as normalized engagement rates (interactions per view).

Comment-Level Analysis

We then characterized user engagement by studying the comments associated with each video. Our dataset initially consisted of 60,507 labeled individual comments. Videos with fewer than 5 annotated comments were excluded from the controversy analysis to ensure statistical reliability. Comments were categorized as agreeing, disagreeing (D), or N, and were used to compute entropy-based opinion diversity, or the controversy score (CS) [23,43]: $CS_v=H(C_A, C_D)$, where $H(\cdot)$ denotes Shannon entropy. Note that the measure portrays opinion heterogeneity, and not merely the existence of disagreement, making it possible that the comments are all in disagreement with the video, but homogeneously so, making the discussion not internally controversial.

Statistical Analysis

Models

To statistically test our hypotheses regarding engagement, temporality, and discourse characteristics, we used a Bayesian framework. This approach provides a direct probabilistic interpretation of model parameters and full posterior distributions quantifying uncertainty. All models were specified using a consistent set of predictors to ensure comparability: the video's stance label (with "neutral" as the reference category), the log-transformed subscriber count (LogSub), and the CHV.

We assigned weakly informative normal priors ($\beta_j \sim \text{Normal}(0,\sigma^2)$) to all regression coefficients to regularize estimates and prevent overfitting.

Engagement Rate Models

To model the LR and CR, which are proportions bounded between 0 and 1, we specified a beta regression model for each, a standard approach for proportional outcomes, as it respects the bounded support of the data. For each video *i*, the engagement rate y_i is assumed to follow a beta distribution:

$$(1) y_i \sim \text{Beta}(\mu_i \cdot \kappa, (1-\mu_i)\cdot\kappa)$$

where the mean μ_i is linked to the predictors via a logit link function:

$$(2) \text{logit}(\mu_i)=\beta_0+\beta_{label[i]}+\beta_{LogSub}\cdot LogSub_i+\beta_{CHV}\cdot CHV_i$$

The precision parameter κ was assigned a broad half-normal prior. The exponentiated coefficients are interpreted as odds ratios (ORs), representing the multiplicative change in odds of engagement per unit increase in the predictor.

XSL•FO
RenderX

Engagement Temporality Model

To model the speed of engagement, we predicted the number of days required to reach 90% of total engagement (*P*90). As this outcome is a positive integer count, we specified a *negative binomial regression model*, which is well-suited for overdispersed count data. The number of days d_i for video *i* is modeled as:

$$(3) di \sim NegativeBinomial(\mu i; \alpha)$$

where the mean μ_i is linked to the predictors via a log link function, and α is the dispersion parameter:

$$(4) \log(\mu i) = \beta 0 + \beta label[i] + \beta LogSub \cdot LogSubi + \beta CHV \cdot CHVi$$

To control for endogeneity—specifically the mechanical correlation between low engagement volume and faster saturation—we fit a sensitivity specification adding log-transformed total engagement ($\beta_{Vol} \cdot \log(TotalViews_i)$) to the linear predictor. The exponentiated coefficients from these models are interpreted as incidence rate ratios (IRRs), representing the multiplicative change in expected days to reach P90.

Discourse Characteristic Models

Finally, to analyze the nature of the comment section discourse, we fit two separate models.

Disagreement model: To predict the proportion of disagreeing comments ($prop_{disagree}$), we used a beta regression model identical in structure to the engagement rate models (Equations 1 and 2).

Controversy model: To predict the CS, which is a continuous, unbounded variable, we specified a standard *Bayesian linear regression model* assuming a normal (Gaussian) likelihood. For video *i*, the CS is modeled as:

$$(5) CSi \sim Normal(\mu i, \sigma 2)$$

where the mean μ_i is directly modeled as a linear combination of the predictors:

$$(6) \mu i = \beta 0 + \beta label[i] + \beta LogSub \cdot LogSubi + \beta CHV \cdot CHVi$$

The coefficients (β_j) from this model represent the direct additive change in the CS score.

Bayesian models allow us to move beyond simple averages to understand the probability of engagement. In the results that follow, we report ORs to measure the likelihood of interaction; a value greater than one indicates that a video is more likely to receive likes or comments than neutral content. We also report IRRs for our timing analysis; here, a value less than one signifies a "speedup," meaning the video accumulates its audience and saturates its engagement potential more rapidly than the baseline.

Additionally, we validate findings under frequentist statistical frameworks and assess sensitivity to model specification. Results and details appear in Sections B-E in Multimedia Appendix 1.

Ethical Considerations

This observational study analyzed publicly available YouTube content collected via the official YouTube Data API v3; it involved no direct interaction with human participants. Consequently, institutional review board (IRB) approval was deemed unnecessary, and the research complies entirely with YouTube's Terms of Service. To protect user privacy, all individual usernames, user IDs, and identifying metadata associated with the comments were removed during data processing, and the comment text was deidentified. For transparency and reproducibility, examples of the public, high-profile YouTube channels analyzed in this study are listed in Multimedia Appendix 1. These channels represent public entities and major creators with no reasonable expectation of privacy regarding their broadcasted content.

Results

Videos

Overview

We begin by examining the user engagement with the videos in our dataset. As can be seen in Table 1, more videos have been classified as either "strongly in favor" or "in favor" of vaccination, although there are around 1500 videos in the "against" and "strongly against" categories.

Table . Video categories, score ranges, number of videos and channels in each category, and the proportion of videos to a channel. Note that the number of channels is presented per category and may overlap (the number of unique channels in the dataset is 3335).

Category	Score range	Videos, n	Channels, n	Videos/channel
Strongly in favor	[0.5, 1]	2492	1558	1.60
In favor	(0, 0.5)	1172	850	1.38
Neutral	[0]	996	796	1.25
Against	(–0.5, 0)	602	437	1.36
Strongly against	[–1.0, –0.5]	812	585	1.39

Extent of Engagement

To establish the popularity of each video, we consider the last snapshot of its metadata that we have captured, which would be within a week of the end of the data collection period. Figure 2 shows the distributions of the number of views, likes, and comments for the videos in each of the 5 categories. The vertical lines signify the median value of the distribution, which is also shown in the upper right of each plot, as well as the mean. Because the distributions are heavy-tailed, the mean is affected by the few popular outliers and is thus sometimes several orders of magnitude larger than the median. Note that we are witnessing the videos at different times from their posting. Even though on average the videos labeled as "against" or "strongly against"

XSL•FO
RenderX

are slightly "older" (mean 86 [SD 48] and mean 82 [SD 48], respectively) than those labeled as "in favor" or "strongly in favor" (mean 98 [SD 50] and mean 87 [SD 51], respectively), the standard deviations are so large that the effect of video age is negligible when comparing distributions within label categories.

Figure 2. Distributions of video engagement by stance, shown in raw counts and normalized rates. (A) Log-log frequency plots for views, likes, and comments; dashed vertical lines indicate the median. The y-axis represents the number of videos. Vaccine-hesitant stances ("A" and "SA") tend to have higher median engagement counts. (B) Histograms of normalized like ratio (LR) and comment ratio (CR), the y-axis represents the percentage of videos within each bin. Distributions for "A" and "SA" stances are shifted to the right, indicating a higher rate of engagement per view compared to provaccine ("SIF" and "IF") and neutral ("N") stances. A: against; IF: in favor; N: neutral; SA: strongly against; SIF: strongly in favor.

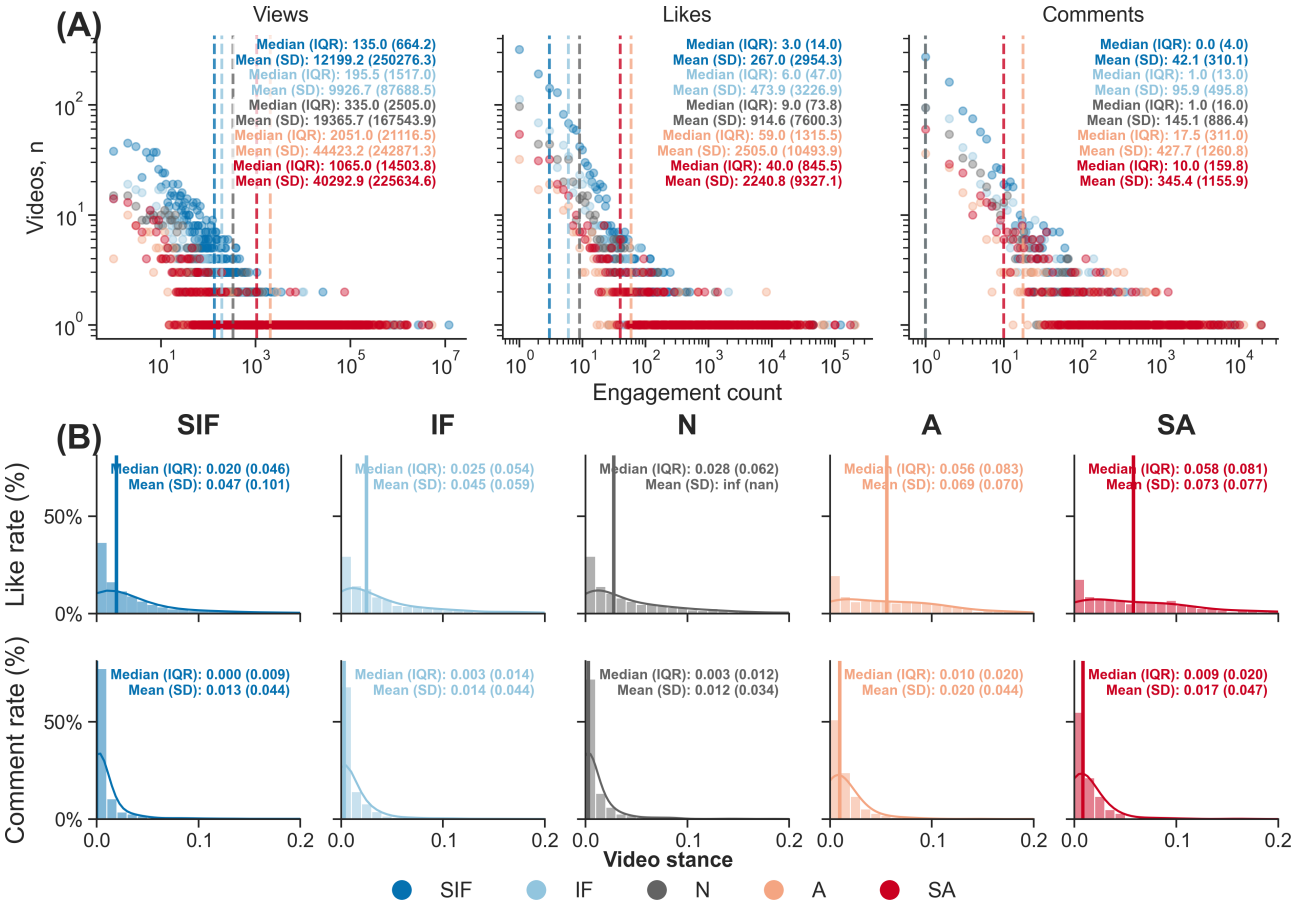

Using all 3 metrics of engagement, we find that the videos in the A and SA categories receive much more attention than those IF and SIF categories. The difference in median views, for instance, differs by an order of magnitude from "strongly against" at 1065 (IQR 55-14,559) to "strongly in favor" at 135 (IQR 26-690). The views of "against" videos are even greater, at the median of 2051 (IQR 50-21,167). Further, 24.7% (615/2492) of the videos in the "strongly in favor" category received no likes at all, compared to 13.2% (107/812) of those in the "strongly against" category. The situation is even more dire for comments—a more strenuous action than a like—wherein the disparity is at 52.2% (1300/2492) compared to 27.5% (223/812), respectively. When normalized by views, these differences translate to substantial engagement asymmetries: vaccine-hesitant videos ("against" and "strongly against") have median like ratio of 0.056 and 0.058 (ie, 5.6 and 5.8 likes per 100 views), respectively, compared to 0.020 for strongly in favor videos—a 2.8- to 2.9-fold engagement advantage. Combined raw interaction volume (likes+comments) yields approximately 10-fold higher engagement for vaccine-hesitant content (see Table S7 in Multimedia Appendix 1 for detailed comparison).

Statistical Analysis

To provide a comprehensive statistical account of online vaccine discourse, we fit a series of Bayesian regression models predicting 3 key outcomes: engagement rates, engagement temporality, and discourse characteristics. The posterior distributions for the coefficients of each model are visualized in Figure 3. All comparisons are made relative to a neutral video stance. All model parameters converge ($R \approx 1$ for all variables).

XSL•FO
RenderX

Figure 3. Posterior distributions of coefficients from Bayesian models of engagement and discourse. The figure summarizes the results of 3 sets of Bayesian regression models. Ridge plots show the posterior probability distribution for each predictor's effect size on the logit (for rates) or log (for counts) scale. Horizontal bars represent the 94% highest density interval (HDI). A distribution to the right of the zero line indicates a positive association; to the left indicates a negative association. (A) Engagement: models predicting the like ratio and comment ratio. (B) Days to 90% of engagement: models predicting the number of days to reach 90% of total views, likes, or comments. A negative coefficient signifies a faster trajectory. (C) Discourse characteristics: models predicting the proportion of disagreeing comments and the controversy score (CS). The channel video counts (CHV) variable had no significant effect for all the models; it is not shown in the plots. Even when controlling for channel size, vaccine-hesitant stances are the strongest predictors of high normalized engagement rates and accelerated temporal saturation. A: against; IF: in favor; SA: strongly against; SIF: strongly in favor.

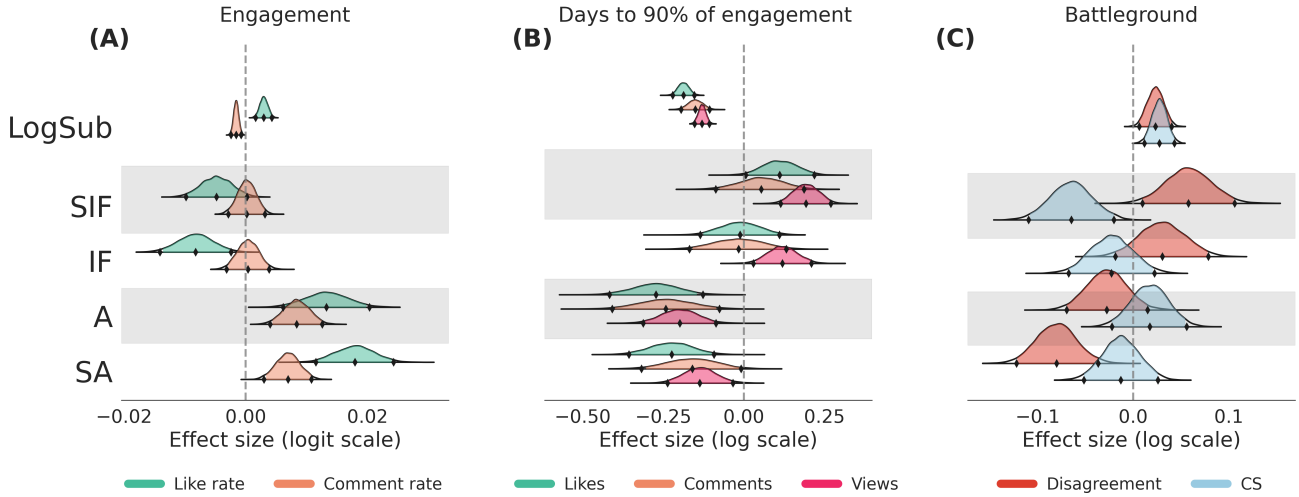

Figure 3A displays the predictors for LR and CR. The results show a clear divergence in how audiences interact with pro- versus vaccine-hesitant content. For the like ratio model, we find strong evidence that vaccine-hesitant stances are associated with higher rates of likes compared to neutral content. The odds of a viewer liking a SA video are approximately 1.8% higher (OR 1.018, 94% highest density interval [HDI] 1.012-1.024), and 1.3% higher for an A video (OR 1.013, 94% HDI 1.007-1.020). The credible intervals for both coefficients are entirely above 1.0, indicating high certainty in the positive effect. Conversely, the IF stance is associated with a statistically credible decrease in the odds of receiving a like, with the odds being 0.8% lower (OR 0.992, 94% HDI 0.986-0.997). The effect for SIF is more uncertain, with a credible interval that just touches an OR of 1.0 (94% HDI 0.991-1.000).

The comment ratio model indicates that vaccine-hesitant content is most effective at stimulating discussion. Both A and SA stances are associated with a small but certain increase in the odds of commenting, with ORs of 1.008 (94% HDI 1.004-1.012) and 1.007 (94% HDI 1.003-1.011), respectively. The effects for provaccine stances (SIF and IF) are centered precisely on an OR of 1.0, with credible intervals that broadly overlap 1.0, indicating their effect on the comment ratio is indistinguishable from that of neutral content. Across both models, the channel characteristics LogSub and CHV showed very small and uncertain effects. Collectively, these models provide quantitative evidence that vaccine-hesitant content is more successful at generating positive feedback (likes), while also being slightly more effective at stimulating discussion (comments).

To validate these findings under alternative statistical frameworks, we conducted robustness checks using logit-ordinary least squares regression with heteroskedasticity-consistent (HC3) standard errors (Section B in Multimedia Appendix 1).

These models statistically confirm that the "engagement advantage" of vaccine-hesitant content is not a random artifact. The data show that a user watching a vaccine-skeptic video is significantly more likely to click "like" or leave a comment than a user watching a provaccine video, regardless of the channel's size. This suggests that hesitant narratives are inherently more effective at eliciting active feedback from their audience.

Engagement Over Time

Next, we take advantage of the regular requerying of the engagement statistics to examine the temporal trend of the views, likes, and comments the video receives after its posting. Figure 4 shows the trajectories of views, likes, and comments of the videos in each of the label categories. For each statistic, we also compute an average trajectory line and show it in color. We also show (by the vertical line) the number of days at which the average trajectory reaches 90% (P90) of the statistic. For instance, videos labeled as "strongly in favor" on average reach 90% of their views at day 32 (IQR 18-64), whereas those labeled as "strongly against" reach it at day 18 (IQR 8-38).

XSL•FO
RenderX

Figure 4. The engagement "trajectories" showing the percentage of views, likes, or comments (see row labels) the video achieves over days (x-axis), by label category (shown above each plot). The trajectories for each day are in light gray, and the average trajectory is in color. The vertical bar marks the number of days at which the average trajectory reaches 90% of the statistic. For likes and comments, videos are limited to those having more than 1 like or comment, respectively. Vaccine-hesitant content saturates significantly faster (median 18 days) than provaccine content (median 32 days), creating a narrow window for effective public health counter-messaging. A: against; IF: in favor; N: neutral; P90: 90% of total engagement; SA: strongly against; SIF: strongly in favor.

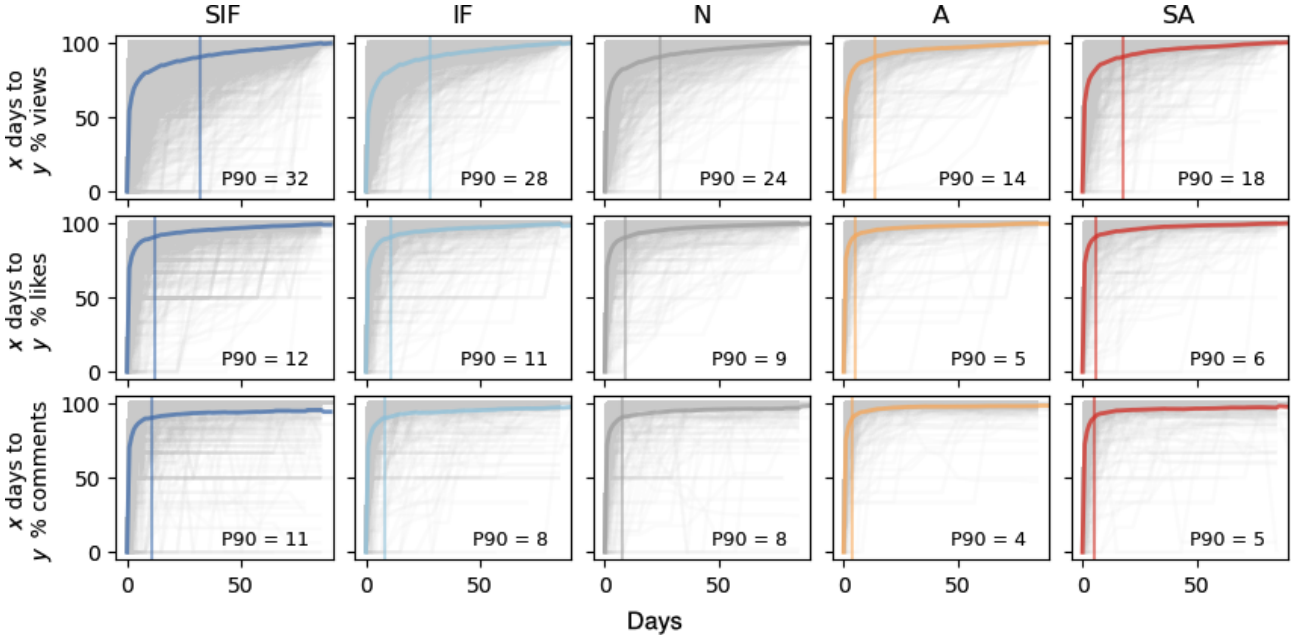

Although it is not possible for views to decrease over time, it is possible for the number of likes or comments to decrease when a like is unselected or a comment is deleted. This introduces some noise, especially for those videos that have only one comment or like. For this reason, the analysis of the likes and comments includes only videos that had more than 1 like or comment (respectively) in our dataset. See Section H in Multimedia Appendix 1 for plots that include all videos, as well as the same statistics at thresholds 5 and 10.

Considering all of the engagement metrics, we find that the videos labeled as "against" or "strongly against" vaccination reach their peak engagement sooner than those in favor. Visually, the trend lines become sharper at the "knee" (where the distribution starts reaching a plateau). The pattern remains similar at thresholds 5 and 10 (see Section H in Multimedia Appendix 1), though the magnitude of the difference between "strongly in favor" and "strongly against" decreases in the case of comments, which are sparser than likes.

To quantitatively characterize the temporal saturation of engagements, we use a nonnegative binomial Bayesian model. The center panel of Figure 3B models the speed at which videos accumulate their audience, predicting the number of days required to reach 90% of total engagement (P90). The results from our Bayesian negative binomial models show that vaccine-hesitant content reaches its engagement peak significantly faster than neutral or provaccine content. For the trajectory of likes, videos with an A stance are expected to reach their P90 approximately 24% faster than neutral videos (IRR 0.76, 94% HDI 0.658-0.864). Similarly, SA videos are approximately 20% faster (IRR 0.80, 94% HDI 0.705-0.905).

In contrast, provaccine content exhibits a slower rate of engagement accumulation. SIF videos are predicted to take approximately 12% longer to reach their P90 peak (IRR 1.12, 94% HDI 1.016-1.242), with a credible interval entirely above 1.0. The effect for IF videos is indistinguishable from neutral content (IRR 0.99, 94% HDI 0.882-1.112). These patterns were consistent across the models for views, likes, and comments. Furthermore, we found a strong effect for channel size, with each unit increase in LogSub associated with a trajectory that is approximately 17% faster (IRR 0.83, 94% HDI 0.803-0.856), suggesting that videos from larger channels accumulate their engagement more quickly.

To address endogeneity of temporal saturation, we conducted a sensitivity analysis by refitting the negative binomial P90 models with additional control for log-transformed total views. This specification tests whether stance effects on days-to-saturation reflect genuine timing asymmetries or merely mechanical confounding. The mechanism of potential confounding is straightforward: videos receiving fewer total likes or views will, by definition, reach a 90% engagement threshold sooner than videos with higher engagement ceilings. If stance predicts both engagement volume and timing, the observed temporal asymmetry could be entirely mechanical rather than reflecting genuine differences in how quickly vaccine-hesitant versus provaccine content accumulates engagement.

The controlled models substantially attenuated the stance effects on P90 days. For "strongly against" videos, IRR for temporal saturation decreased from 0.56 (44% faster saturation; 94% HDI 0.50-0.62) in the uncontrolled baseline model to 0.80 (20% faster saturation; 94% HDI 0.70-0.90) after controlling for log(TotalViews). This 24 percentage point reduction represents substantial confounding. The control variable itself had a positive coefficient (IRR 1.042, 94% HDI 1.021-1.061),

XSL•FO
RenderX

meaning that videos accumulating more total engagement reached P90 more slowly, consistent with a pattern of sustained engagement growth rather than front-loaded spikes. Approximately 54% of the uncontrolled 44% effect appears attributable to mechanical confounding, while 46% persists as a genuine timing advantage. This controlled estimate of 20% faster saturation for vaccine-hesitant content represents a more conservative but still substantial temporal asymmetry, indicating that these videos not only receive less total engagement but also concentrate their engagement into a shorter time window relative to their final engagement level. All parameters achieved acceptable convergence (R <1.01). Detailed results are shown in Section G in Multimedia Appendix 1.

In summary, we find that vaccine-hesitant content has a "speed advantage." By raw timing, while provaccine content accumulates its audience gradually over a month (32 days to P90), vaccine-hesitant content experiences a concentrated burst of activity, saturating within 18 days (44% faster). However, a portion of this temporal advantage reflects mechanical confounding: videos with lower total engagement reach absolute thresholds sooner by definition. Sensitivity analysis controlling for log-transformed total engagement attenuates the effect from 44% to 20% faster saturation, indicating that approximately 54% of the raw temporal advantage reflects mechanical confounding while 46% persists as genuine temporal compression. This means vaccine-hesitant videos not only receive lower total engagement but also concentrate what engagement they receive into a narrower time window relative to their eventual engagement ceiling. Both the 44% practical effect and the 20% controlled effect have policy relevance: practitioners face the 44% saturation speed; researchers identify a 20% genuine temporal compression. This rapid early-stage lifecycle implies that the digital footprint of hesitant content is established quickly, suggesting that the window for effective counter-messaging or moderation is much narrower for this type of content.

In summary, the 18-day versus 32-day difference corresponds to a descriptive 44% faster saturation for vaccine-hesitant content, whereas the controlled IRR from the negative binomial model corresponds to a more conservative model-based estimate of 20% faster saturation after adjusting for total engagement volume.

Comments

Beyond the quantity of audience interaction, we aimed to understand its nature. To this end, we classified comment stances relative to the video's own stance, analyzing the results in aggregate, by channel topic, and through the CS.

The aggregate results, shown in the "all categories" plot in Figure 5A, show that videos with SA and A stances are more likely to foster echo chambers; their comment sections are dominated by agreement, with 39.9% (6184/15,508) and 35.6% (6055/17,030) of comments agreeing, respectively (with only 18.6% (2889/15,508) and 25.7% (4379/17,030) disagreeing). In contrast, videos with SIF and IF stances attract significant contention, with higher disagreement rates of 35.9% (2751/7650) and 31.7% (2362/7440) and lower agreement rates of 24.4% (1870/7650) and 27.8% (2072/7440). This suggests that while vaccine-hesitant content fosters confirmatory engagement, provaccine content often serves as a "battleground" for public debate. Disaggregating by authority, we see that the "battleground" effect is most intense in health and entertainment (eg, 47.4% [138/291] disagreement on SIF health videos), while the knowledge topics (like doctor) appear more balanced, with "strongly in favor" videos attracting the highest rate of agreement (422/820, 51.5%) for any provaccine content.

XSL•FO
RenderX

Figure 5. Comment agreement proportion by video stance, faceted by channel characteristics. The figure displays the normalized proportion of comment stances (D: disagree; N: neutral; A: agree) for videos, broken down by various channel categorizations. Each horizontal bar shows the proportion of comment types for a specific video stance (SIF: strongly in favor; IF: in favor; N: neutral; A: against; SA: strongly against) within that category. (A) Panels show distributions for channels grouped by their primary content category (eg, health and politics). (B) Panels show distributions for channels grouped by their declared authority (eg, doctor and news). (C) Panels show distributions for channels grouped by their institution type (individual, organization, other). Vaccine-hesitant videos function as confirmatory "echo chambers" (high agreement), while provaccine health videos serve as "battlegrounds" characterized by high levels of disagreement.

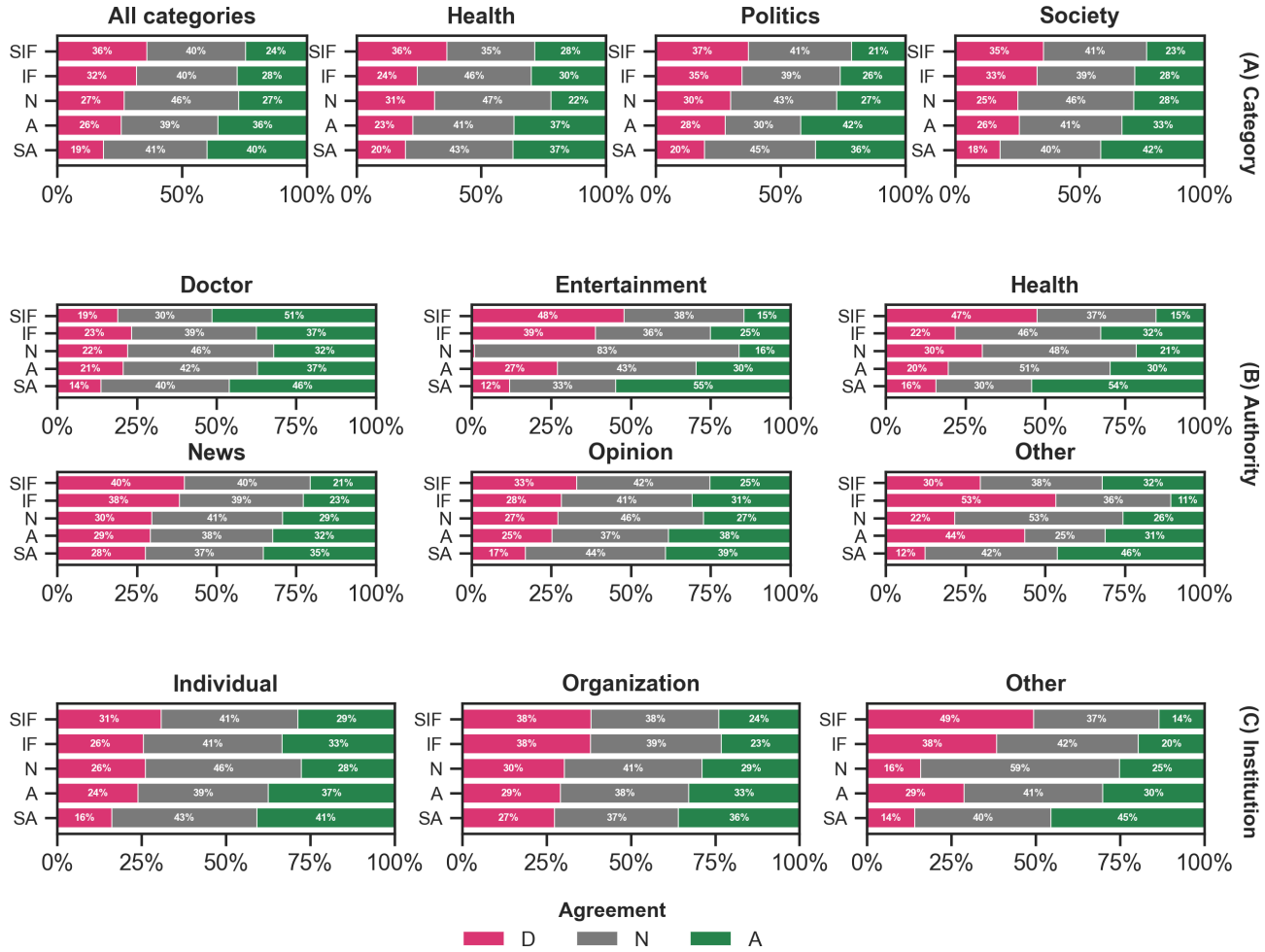

These patterns are further explored in the ternary plots of Figure 6A, which visualize the distribution of individual videos in the tree level of agreement. Vaccine-hesitant videos (colored red and orange) mostly cluster in the top "agree" corner, providing strong visual evidence of echo chambers, particularly in the politics and society topics. Provaccine videos (colored blue) are distributed more broadly, with a significant concentration extending toward the bottom-left "disagree" corner and into the central battleground region, confirming their role as contentious battlegrounds.

XSL•FO
RenderX

Figure 6. Visualizing discourse patterns through comment agreement distributions. This figure presents 3 perspectives on how audience agreement varies by video stance and topic. (A) A ternary plot visualizes each individual video's discourse pattern, colored by its own stance (red=strongly against, blue=strongly in favor). The top corner represents high agreement (>60%), the bottom-left corner represents high disagreement (>60%), and the bottom-right represents neutrality. Dashed lines demarcate "echo chambers" (areas where a single stance dominates >60% of comments). Vaccine-hesitant videos (red) cluster heavily in the top "agree" echo chamber, particularly in society and politics. Provaccine videos (blue) are dispersed toward the bottom-left "disagree" corner and the center. N: neutral; SA: strongly against; SIF: strongly in favor. (B) The same distribution colored by controversy score (CS). Deep purple indicates high controversy (high entropy), corresponding to the "battleground" region in the center of the triangle where no single sentiment dominates. Bright teal indicates low controversy, aligning with the echo chambers at the corners. The platform simultaneously hosts an insulated agreement and intense public contention, mediated by the video's stance.

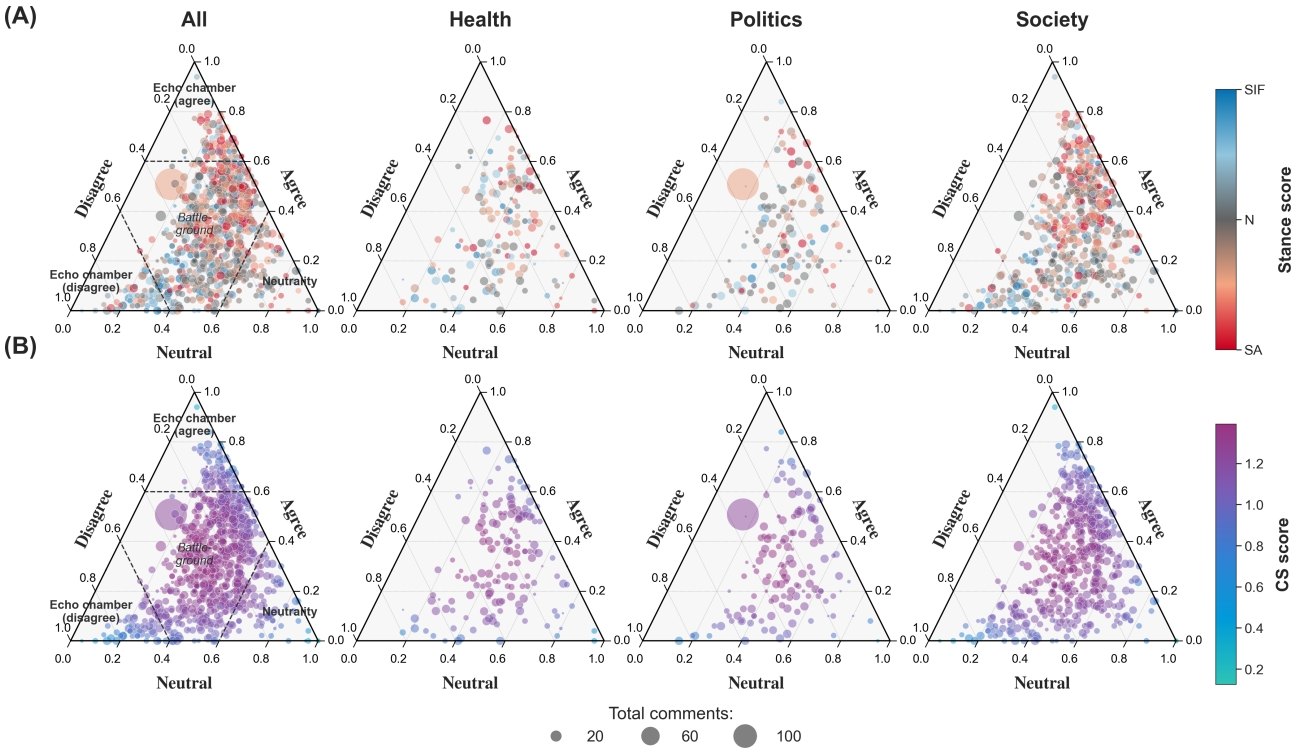

Building on these observed patterns, we then colored the same distribution by the CS to quantify the level of contention (Figure 6B). The results confirm that the CS effectively captures the distinction between echo chambers and battlegrounds. The lowest controversy scores (bright teal) are heavily concentrated in the corners of the plots, particularly the top "agree" corner, aligning with the confirmatory echo chambers identified in the vaccine-hesitant videos. In contrast, the highest controversy scores (deep purple) are found in the center of the triangle, corresponding to the "battleground" videos that attract a chaotic mix of agreeing, disagreeing, and neutral comments. By comparing the topic-specific panels, we can see that the health topic contains a prominent cluster of high-controversy (purple) videos in its center, quantitatively confirming it is the most contentious domain. This shows that the CS metric captures the observed discourse patterns, mapping high consensus to low controversy and high conversational diversity to high controversy.

Figure 3C analyzes the nature of the comment section itself by modeling the proportion of disagreement and the CS. The results statistically confirm our "battleground vs echo chamber" hypothesis with quantitative evidence. In the *disagreement model*, provaccine stances are associated with a higher likelihood of attracting dissent. Compared to neutral videos, a SIF stance increases the odds of a comment disagreeing by approximately 5% (OR 1.05, 94% HDI 1.003-1.094). In contrast, a SA stance is associated with a decrease in the odds of disagreement by about 7% (OR 0.93, 94% HDI 0.898-0.972), with its credible interval falling entirely below 1.0. This provides strong statistical evidence that SA videos foster more agreeable, echo-chamber-like environments.

The *controversy score model* further clarifies this dynamic. The SIF stance is associated with a statistically credible decrease in the CS of approximately 0.075 units (94% HDI –0.119 to –0.033) compared to neutral content. At first glance, this seems counterintuitive. However, it suggests that while SIF videos attract more direct disagreement (a 2-way debate), neutral videos foster a more chaotic and unpredictable mix of "agree," "disagree," and "neutral" comments. The CS metric captures entropy, not just disagreement: perfectly balanced agree/disagree comments have lower CS than chaotic agree/disagree/neutral mixtures. Thus, more focused contention (SIF videos) can have lower CS than less predictable patterns. The effects of other stances (IF, A, SA) on the CS are uncertain, with credible intervals that all broadly overlap zero. Across both models, channel size (LogSub) has a significant positive effect, indicating that videos from larger channels tend to host both more disagreeable and more broadly controversial debates.

Our analysis delineates 2 distinct social environments coexisting on the platform. Vaccine-hesitant videos function largely as "echo chambers" where users reinforce each other's views with high agreement. In contrast, provaccine videos—especially those from health institutions—function as "battlegrounds"

XSL•FO
RenderX

characterized by debate and dissent. This distinction is crucial for interpretation, as it clarifies that high engagement numbers on provaccine content often reflect controversy and argument rather than consensus or support.

Channels

Next, we show that the provaccine and vaccine-hesitant content largely originates from two distinct creator ecosystems. This section details the evidence for this sourcing asymmetry by examining the topical framing, authority signals, and institutional identity of the channels in our dataset.

To contextualize our findings, we first characterized the topical composition of the 7213 videos using the categories of the posting channel. Using a hierarchical resolution of YouTube's metadata, we categorized videos into 3 primary topics: health (4027/6074, 66.3%), politics (918/6074, 15.1%), and society (628/6074, 10.3%). This hierarchical approach was necessary due to the prevalence of multitopic assignments in the original metadata (71.8% of videos), reflecting the inherently multidimensional nature of vaccine discourse. As shown in our subsequent analyses, this topical framing emerges as a significant moderator of engagement and comment discourse.

The distribution of video stances across these channel characteristics is presented in Figure 7. By topic (Figure 7A), health-related videos are slightly more provaccine (49.3% [1987/4027] "strongly in favor" and 21% [847/4027] "in favor"), whereas politics and society topics show more balanced distributions. This suggests health-framed discussions may attract greater institutional participation, while politics and society frames host a more diverse set of creator viewpoints.

Figure 7. Share of videos posted by channels with certain labels per channel categories found in the metadata (A) and per authority (B) and institution (C) labels manually annotated on a subset of 645 channels. Provaccine discourse is dominated by institutional voices (news/health organizations), whereas vaccine-hesitant content is driven by individual creators, including self-identified medical doctors who bypass institutional constraints. A: against; IF: in favor; N: neutral; SA: strongly against; SIF: strongly in favor.

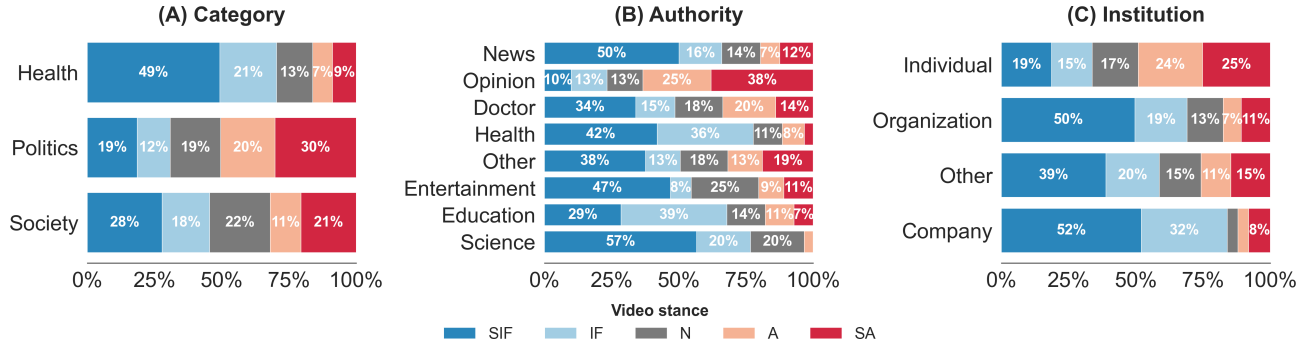

This pattern is further intricate when considering the authority of a channel's signals extracted via manual annotation (Figure 7B). Channels identified as news or about health predominantly post content in favor of vaccination, with 66% (870/1318) and 77.7% (237/305) falling into the combined "strongly in favor" and "in favor" categories, and with minimal representation in the strongly against category. In contrast, channels run by (self-declared) doctors post content both in support and against vaccination. Finally, entertainment and opinion channels display the most heterogeneous distributions, including significant portions of antivaccine content, suggesting these creators serve diverse audiences, often without the institutional constraints that guide health or news organizations.

Figure 7C shows a clear distinction between organizational and individual creators. Organizations are more represented in provaccine content production (1103/1598, 69%) of their videos fall in the "strongly in favor" and "in favor" categories combined, versus only 10.8% (172/1598) in "strongly against," whereas individual creators exhibit greater stance diversity. Only 18.6% (132/708) of videos from individual creators are "strongly in favor," while 25.1% (178/708) are "strongly against," pointing to the need for engagement with individual posters to tackle such content.

These distributional patterns are reflected in the composition of the most prolific channels in our dataset (Section I in Multimedia Appendix 1). The 10 channels that posted the most videos are responsible for just 5% of the total volume (Gini=0.39), indicating that the content is not dominated by a few voices. This list includes a mix of news organizations, such as CTV and 9NEWS, which primarily post provaccine content, alongside individual creators like a self-described doctor or a showman, who are significant contributors to the vaccine-hesitant discourse. This mix illustrates the broader patterns in action: news media are an important contributor of vaccine-related content, but influencers and alternative voices have a strong and countervailing presence.

Finally, we investigated how the source of a video impacts audience discourse by analyzing comment agreement patterns across these channel types (Figure 5B,C). The results show that the nature of the creator significantly shapes the comment section's dynamics. Within the Authority categorization, for instance, health channels exhibit the most polarized reactions, fostering strong echo chambers for "strongly against" videos (82% agreement) while their "strongly in favor" videos become "battlegrounds" (50% disagreement). When categorized by institution, a clear distinction appears: videos from organizations tend to generate stronger echo chambers, especially for vaccine-hesitant content, while individual creators host more balanced and contentious conversations across the board.

In summary, the analysis of channel characteristics shows that provaccine and vaccine-hesitant content largely originates from two distinct creator ecosystems. The provaccine discourse is predominantly driven by organizations, particularly news and health institutions, which reach cohesive audiences. In parallel,

XSL•FO
RenderX

the vaccine-hesitant discourse is more strongly represented by a diverse array of individual creators, who foster more contentious and heterogeneous comment sections. This fundamental divergence in sourcing is crucial for understanding why these two types of content exhibit such different engagement dynamics and foster such different online communities.

Moderation

In the course of the collection, 243 videos at some point failed to return metadata results during the periodic recollection efforts. We manually examined the errors the YouTube interface provided for these, and found that for 116, the platform simply tells the users that the "This video isn't available anymore," 66 have been made private by the uploader, 50 were unavailable due to the posting account having been terminated by the platform, 3 were removed by the uploader (a message distinct from the video simply being unavailable), 6 were removed due to violating YouTube's Community Guidelines or Terms of Services, one because the uploader has closed their YouTube account, and one because of a copyright claim. We consider the video disappearances attributed to the author separately from all others (which we attribute to the platform), and find the rates shown in Figure 8. We find that the removals by the author of the video happen more for "strongly in favor" and "in favor" videos, whereas videos labeled as "against" or "strongly against" were more likely to be removed by the platform.

Figure 8. Percentage of videos in each stance that have been removed. Bars distinguish between videos made private or deleted by the uploader (by author) and those taken down for guideline violations (by platform). A: against; IF: in favor; N: neutral; SA: strongly against; SIF: strongly in favor.

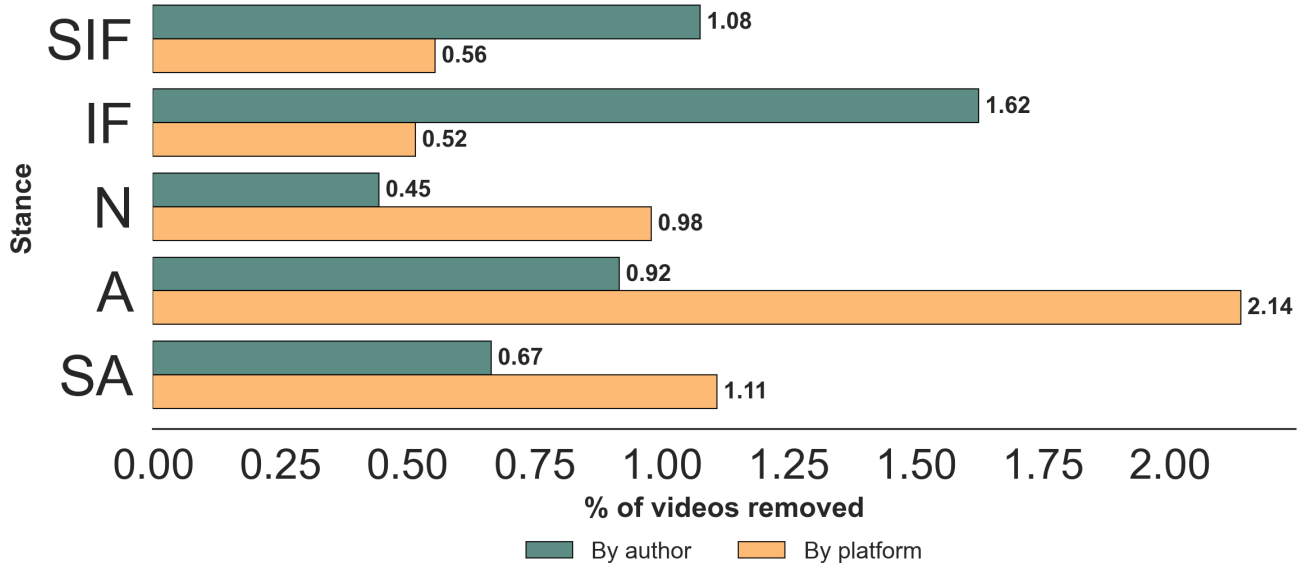

Considering the error codes more closely, we find that all videos removed due to violating YouTube's Community Guidelines were either labeled as "against" or "strongly against." On the other hand, videos that belonged to channels terminated by the platform span all 5 categories. Whenever the video was made private by its author, the label is more likely to be "in favor" or "strongly in favor" (77.3% [51/66] of such videos) than "against" or "strongly against" (9/66, 13.6%). Thus, we find a variety of reasons for the disappearance of videos, with some evidence of the platform's policies affecting more the vaccine-skeptic side when it comes to sanctioning content, but not necessarily the accounts themselves.

The median time of the video disappearing is 14.4 (IQR 4.9-35.8) days from its posting, meaning that half of the videos were taken down within that time. However, many remained until almost the end of our 3-month collection period, pointing to a low but constant level of moderation across time. Since most of the interactions with the video happen within the first few weeks of a video's posting, this policy likely fails to stem the majority of the video's impact.

Upon the examination of the metadata collected for the videos of accounts terminated by the platform (as the videos themselves were not available at the time of the analysis), we find that they largely consist of vaccine-skeptic, anti-establishment content, as well as likely semiautomated channels that repost news snippets. Those labeled as "in favor" of vaccines sometimes use sensationalized language, such as "The truth might SHOCK you!" (such phrases are often considered "clickbait" to attract an audience). Some also engage in "keyword stuffing" by including many keywords in the description of the video to increase the chance of attaining a higher rank during a search. Yet other accounts mention vaccination in the context of providing other information, such as a channel promoting US visas. Unfortunately, beyond the error message, we cannot obtain any additional information on why these accounts might have been removed; however, the above observations exemplify the low-quality content that public health institutions compete with on this platform. The last engagement information our data collection captured for these removed videos reflects the overall trend of videos against vaccination, showing more engagement than those in favor (see Section J in Multimedia Appendix 1).

Discussion

Context

Prior YouTube vaccine studies have shown that vaccine-hesitant content receives more engagement than provaccine ones [10,11,44]. However, these snapshot-based analyses reveal only

XSL•FO
RenderX

an engagement gap, leaving unanswered when and how fast this gap widens. By tracking engagement longitudinally, we document a previously unquantified temporal asymmetry: vaccine-hesitant videos saturate engagement significantly faster than provaccine content (44% faster in raw timing; 20% faster in model-based estimates that control for engagement volume). Furthermore, we challenge the assumption that polarization is symmetric [18,45]. Our comment analysis demonstrates that vaccine-hesitant videos function as agreeable echo chambers [23,46], whereas provaccine videos act as contentious battlegrounds. This suggests the platform hosts two distinct discourse architectures simultaneously, mediated by content stance.

Importantly, our analysis is purely observational: we do not observe recommendation paths or run interventions, and therefore we cannot identify causal effects of YouTube's ranking or recommendation systems; we interpret our results as patterns of realized engagement rather than causal impacts of specific algorithms.

Engagement Asymmetry and the Temporal Window for Intervention

By quantifying engagement dynamics, we observe that vaccine-hesitant videos attract substantially higher normalized engagement rates (2.8 to 2.9× higher) and reach saturation significantly faster than provaccine content. This temporal dimension suggests that vaccine-hesitant content gains traction not through sustained growth, but through rapid early-stage saturation. While the observed speed advantage (44% faster descriptively) is partly driven by lower engagement ceilings, a genuine temporal compression persists even after controlling for this confounding (20% faster after controlling for total engagement). This distinction between threshold-crossing speed and engagement velocity deserves deeper attention.

The decomposition of this temporal asymmetry shows distinct audience dynamics. The portion of the speedup attributable to confounding indicates that vaccine-hesitant audiences are smaller but more efficient: they mobilize quickly from a self-selected pool, producing a "peak-and-plateau" dynamic characteristic of viral misinformation [47]. Unlike provaccine content, which accumulates engagement incrementally as it diffuses through heterogeneous networks, vaccine-hesitant content exhibits the rapid consolidation typical of echo chambers [18,48]. This aligns with our comment-level findings: the structural homogeneity of the vaccine-skeptic audience facilitates rapid, confirmatory engagement, while the heterogeneity of the provaccine audience results in slower, contentious accumulation.

These findings identify a mechanism-specific temporal signature. Prior work established that false news spreads faster [6,15], but this study suggests that the temporal advantage operates through distinct audience structures: concentrated early-stage mobilization of in-groups versus distributed, delayed expansion into out-groups. From a policy perspective, this mechanism suggests that the critical window for intervention is not merely short (as commonly noted [6,49]) but compressed into a narrow audience demographic, concentrated in the initial 2 weeks post-upload.

Implications for Intervention and Governance

The policy implications of this rapid saturation are 2-fold. First, the 18-day saturation window implies that practitioners must adopt real-time monitoring, as the majority of a video's impact crystallizes faster than traditional fact-checking cycles can respond. Second, since moderation actions in our dataset occurred with a median delay of 14.4 (IQR 4.9-35.8) days [50,51], reactive content removal often arrives after the audience has already been saturated.

Therefore, effective intervention must target early audience composition rather than relying on post hoc removal. These empirical patterns motivate, but do not themselves test, potential intervention strategies. Conceptually, prebunking approaches that inoculate audiences against misinformation narratives [52-54] are particularly relevant here as they work before the rapid saturation occurs. Furthermore, saturated content is potentially amplified via cross-platform diffusion [55]. Platforms should consider proactive algorithmic demotion of borderline content or the elevation of credible sources during the first 2 weeks of a video's lifecycle [56,57]. Rather than removing vaccine-hesitant content outright—a strategy fraught with free-speech concerns and often ineffective after saturation [58,59]—platforms could implement ranking systems that elevate credible sources in recommendation feeds, effectively competing for the attention of early audiences before echo chamber formation. Platform design choices that prioritize engagement metrics inadvertently amplify misinformation by selecting for emotionally resonant content [15,47]; reorienting algorithmic goals toward credibility rather than engagement represents a structural intervention with potential for large-scale impact [57,60].

While we found evidence of moderation, the median delay of 14.4 (IQR 4.9-35.8) days becomes critical when considered against the temporal dynamics we identified. Vaccine-hesitant videos concentrate their engagement—a pattern indicating rapid in-group mobilization characteristic of echo chambers—within approximately 18 days. Since moderation occurs 14.4 (IQR 4.9-35.8) days post-upload on median, content removal actions occur after the most intense period of audience concentration has already occurred, a finding that supports recent work on delayed moderation [50,51].

Given the complexities of content removal [58,59], one possible policy intervention could be the proactive algorithmic amplification of content from verified health authorities, especially during the early days of a video's lifecycle [56,57,61]. Platform design choices that prioritize engagement can inadvertently amplify misinformation [15,47]; therefore, strategic interventions to elevate credible sources could significantly improve the health information environment without resorting to widespread censorship [60].

Our findings are consistent with these concerns, but we cannot speak directly about recommendation-system mechanisms. We therefore present these governance implications as hypotheses and directions for future experimental and audit studies, rather than as causal claims about the effects of specific policy levers.

XSL•FO
RenderX

Discourse Patterns: The Role of Creators and Thematic Framing

According to our findings, online vaccine discourse is likely driven by two fundamentally different creator ecosystems. Provaccine content is predominantly produced by organizations (news and health institutions), whereas individual creators dominate the vaccine-hesitant space. This sourcing divergence is associated with different discourse patterns we observed: institutional videos tend to become "battlegrounds" of disagreement, while individual creators host "echo chambers" of agreement.

This dynamic may also be further modulated by thematic framing. Health-categorized videos, which attract the highest institutional participation, become the most intense sites of controversy [62,63]. In contrast, content framed around politics and society allows individual creators to cultivate highly agreeable, insulated communities [48,64]. This suggests a fundamental tension: institutional voices attempt to maintain a medical frame, but the discourse persistently escapes into social and political contexts where individual voices claim equal standing. To bridge this gap, public health bodies should consider partnerships with credible individual creators—such as doctors who vlog—who can foster dialogue without triggering immediate polarization [24,65,66].

Institutional Context and Future Policy Dynamics

It is important to situate these findings within the broader landscape of vaccine policy. Our data collection (November 2024 to May 2025) captures a moment of significant institutional uncertainty. The observed engagement asymmetry likely reflects not merely ambient polarization but the consequence of eroded trust, wherein health institutions themselves become sites of political contestation.

This raises theoretical questions for future research: Do engagement asymmetries intensify during periods of institutional distrust? Longitudinal monitoring across different policy contexts would clarify whether our findings represent a transient response to uncertainty or a structural feature of polarized health discourse. Additionally, without systematic auditing of recommendation systems, the degree to which these asymmetries reflect platform design versus audience-driven polarization remains an open question [67]. Comparative studies across platforms and jurisdictions would strengthen our understanding of these mechanisms.

Limitations

Our study has several limitations that define the scope of our findings.

Exclusion of Short-Form Video

Our analysis is intentionally focused on substantive, long-form YouTube videos (1‑60 minutes) and excludes short-form content like YouTube Shorts. This decision was made because Shorts exhibit fundamentally different engagement dynamics, algorithmic promotion, and content characteristics (eg, lack of transcripts) that would confound our analysis of engagement trajectories and argumentative stances [68,69]. While this exclusion enhances the internal validity of our findings on deliberative content, it limits their generalizability to the short-form video ecosystem, which represents a growing segment of online media consumption. Given the highly condensed and viral nature of Shorts, we hypothesize that their inclusion would likely amplify the rapid engagement saturation trends observed here, potentially fostering an even faster cycle of misinformation spread. Future research should analyze this content format separately, using methods suited to its visual and algorithmic nature.

Other Exclusion Criteria

Besides excluding short-form videos, each step of our collection more precisely defines the scope of this study, and thus limits its generalizability to other scenarios, such as other topics, time periods, or countries. The keyword selection, for instance, may have excluded the content that refers to vaccination obliquely and focuses on mainstream medicine in general [70]. Further, the focus of the collection was to sample the posted videos, and it is likely that some viral videos were not included in the sample, diminishing the extent of the skewness of our estimates of engagement. Studies using complementary methodology would be necessary to explore other aspects of the YouTube ecosystem.

Temporal and Geographic Scope

Data collection (November 2024 – May 2025) coincided with a period of heightened institutional contestation in US health policy. Consequently, the observed engagement dynamics likely characterize periods of high uncertainty rather than stable institutional consensus. Furthermore, as vaccine hesitancy drivers vary globally [71], our findings are specific to the US-English context. Cross-cultural research is needed to determine if the "engagement advantage" of hesitancy is a universal platform feature or a byproduct of specific sociopolitical environments, such as the high political polarization found in the United States.

Scalable Content Analysis Trade-Offs

To enable a large-scale longitudinal analysis, we used an LLM-based approach for content classification. While our method achieved high accuracy (84%) and perfect interrater reliability, it has limitations. First, stance classification was based on averaging scores from text excerpts, a method that may not fully capture videos with evolving narratives or strategic ambiguity. Second, the classification of texts is made difficult by the presence of humor, sarcasm, hyperbole, and other rhetorical devices that are challenging for both automated and human annotators to agree on (although the overall performance of the LLM-based approach largely agreed with that of the human labelers). Third, our comment analysis, classifying stance as agree/disagree/neutral, captures the balance of opinion but not the qualitative nature of the discourse. It does not distinguish between constructive debate and toxic trolling, nor does it analyze conversational structures like reply threads, which are important dimensions of online deliberation [72,73]. These trade-offs between scale and depth offer clear avenues for future qualitative and computational research.

XSL•FO
RenderX

Sensitivity to Misclassification in Automated Labeling

Our main constructs rely on automated LLM-based labeling. Validation against human annotations indicates an accuracy of 82.5% for video stance (with almost all errors involving the neutral class rather than confusion between pro- and antivaccine positions) and ≈73% exact agreement for comment agreement, rising to ≈90% when "neutral/unrelated" is not counted as an error. These error rates imply that misclassification primarily attenuates differences by shifting items toward the neutral or mixed categories, rather than systematically flipping provaccine into vaccine-hesitant content or agreement into disagreement. Given that our core effects are large—for example, vaccine-hesitant videos receive 10-fold higher raw interaction volume, 2.8- to 2.9-fold higher per-view engagement rates, and reach 90% of their engagement 20% faster even after controlling for engagement volume—the directionality of our findings is robust to plausible levels of label noise. At the comment level, replacing up to 10% - 15% of labels with their opposite direction would reduce but not reverse the observed pattern. Vaccine-hesitant videos' comment sections are dominated by agreement (eg, 40% agree vs 19% disagree), and provaccine videos attract higher disagreement (eg, 36% disagree vs 24% agree), so the classification uncertainty is more likely to bias our estimates toward conservative lower bounds than to generate spurious asymmetries.

Conclusions

This longitudinal study provides systematic evidence that vaccine-hesitant content on YouTube benefits from a distinct engagement advantage: it attracts approximately 10-fold higher raw interaction volume (and 2.8- to 2.9-fold higher per-view engagement rates) relative to provaccine content and reaches audience saturation significantly faster (44% faster in descriptive medians and 20% faster in adjusted regression estimates) than provaccine content. We identify a specific temporal signature—an 18-day saturation window—driven by audience structure. Vaccine-hesitant videos mobilize homogeneous, self-selected in-groups that generate rapid, confirmatory "echo chambers," whereas provaccine content accumulates engagement incrementally through contentious "battlegrounds" of debate.

The policy implications of this mechanism are direct. Because the median time to moderation (14.4 days) lags behind the rapid saturation of hesitant narratives, reactive content removal is structurally ill-suited to mitigate misinformation. An effective public health strategy must therefore shift from post hoc rebuttal to proactive intervention. Prebunking campaigns and algorithmic ranking adjustments that elevate credible sources during the critical first 2 weeks of a video's lifecycle represent the most viable levers for competing with the rapid mobilization of hesitant communities.

While this study is limited to long-form English content and acknowledges the trade-offs inherent to scalable LLM classification, it establishes baseline temporal patterns essential for evidence-based policymaking. Future research must extend beyond single-platform dynamics to examine how recommendation algorithms specifically amplify these asymmetries and how content diffuses across platforms. Ultimately, navigating the complex digital health environment requires not just producing scientific knowledge, but understanding the precise temporal and structural mechanisms by which that knowledge competes for public attention.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. The authors used GPT-4o mini for zero-shot video and comment classification. Full responsibility for the content and integrity of the final manuscript rests entirely with the authors.

Funding

YM acknowledges support from the Lagrange Project of the Institute for Scientific Interchange Foundation (ISI Foundation) funded by Fondazione Cassa di Risparmio di Torino (Fondazione CRT).

Data Availability

Data will be made available at OSF/Zenodo upon publication. Full transcripts and comments not released for privacy.

Authors' Contributions

MT contributed to conceptualization, formal analysis, methodology, visualization, writing of the original draft, investigation, and validation. YM contributed to conceptualization, data curation, visualization, writing of the original draft, investigation, and validation.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data preprocessing and classification, detailed statistical analysis, and top 10 channels' description.
[DOCX File, 715 KB - jmir_v28i1e88519_app1.docx]

XSL•FO
RenderX

References

1. Orenstein WA, Ahmed R. Simply put: vaccination saves lives. Proc Natl Acad Sci USA 2017 Apr 18;114(16):4031-4033. [doi: 10.1073/pnas.1704507114] [Medline: 28396427]
2. Correia T, Hilário AP, Guerreiro C, et al. Policy recommendations for healthcare authorities, organizations and professionals on vaccine hesitancy in Europe: from evidence to practice. Health Policy 2025 Aug;158:105361. [doi: 10.1016/j.healthpol.2025.105361] [Medline: 40446640]
3. Global partners announce a new effort – "The Big Catch-Up" – to vaccinate millions of children and restore immunization progress lost during the pandemic. World Health Organization. 2023 Apr 24. URL: https://www.who.int/news/item/24-04-2023-global-partners-announce-a-new-effort-the-big-catch-up-to-vaccinate-millions-of-children-and-restore-immunization-progress-lost-during-the-pandemic [accessed 2026-05-26]
4. Paoletti G, Dall'Amico L, Kalimeri K, et al. Political context of the European vaccine debate on Twitter. Sci Rep 2024 Feb 22;14(1):4397. [doi: 10.1038/s41598-024-54863-7] [Medline: 38388713]
5. Adeoye AF, Umoru DO, Gomez OO, et al. The 2025 United States Measles Crisis: when vaccine hesitancy meets reality. Cureus 2025 Jul;17(7):e88196. [doi: 10.7759/cureus.88196] [Medline: 40827145]
6. Allen J, Watts DJ, Rand DG. Quantifying the impact of misinformation and vaccine-skeptical content on Facebook. Science 2024 May 31;384(6699):eadk3451. [doi: 10.1126/science.adk3451] [Medline: 38815040]
7. Wilson SL, Wiysonge C. Social media and vaccine hesitancy. BMJ Glob Health 2020 Oct;5(10):e004206. [doi: 10.1136/bmjgh-2020-004206] [Medline: 33097547]
8. Faverio M, Sidoti O. Teens, social media and technology 2024. Pew Research Center. 2024 Dec 12. URL: https://www.pewresearch.org/internet/2024/12/12/teens-social-media-and-technology-2024 [accessed 2026-05-22]
9. Gottfried J, Park E. Americans' social media use 2025. Pew Research Center. 2025 Nov 20. URL: https://www.pewresearch.org/internet/2025/11/20/americans-social-media-use-2025 [accessed 2026-05-22]
10. Basch C, Zybert P, Reeves R, Basch C. What do popular YouTube videos say about vaccines? Child Care Health Dev 2017;43(4):499-503. [doi: 10.1111/cch.12442]
11. Tang L, Fujimoto K, Amith MT, et al. "Down the rabbit hole" of vaccine misinformation on YouTube: network exposure study. J Med Internet Res 2021 Jan 5;23(1):e23262. [doi: 10.2196/23262] [Medline: 33399543]
12. Ng YMM, Hoffmann Pham K, Luengo-Oroz M. Exploring YouTube's recommendation system in the context of COVID-19 vaccines: computational and comparative analysis of video trajectories. J Med Internet Res 2023 Sep 15;25:e49061. [doi: 10.2196/49061] [Medline: 37713243]
13. Jung H, Juneja P, Mitra T. Algorithmic behaviors across regions: a geolocation audit of YouTube search for COVID-19 misinformation between the United States and South Africa. Proc Int AAAI Conf Web Soc Media 2025 Jun;19(1):935-964. [doi: 10.1609/icwsm.v19i1.35854]
14. Chan C, Sounderajah V, Daniels E, et al. The reliability and quality of YouTube videos as a source of public health information regarding COVID-19 vaccination: cross-sectional study. JMIR Public Health Surveill 2021;7(7):e29942. [doi: 10.2196/29942]
15. Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science 2018 Mar 9;359(6380):1146-1151. [doi: 10.1126/science.aap9559] [Medline: 29590045]
16. Medical misinformation policy. YouTube. URL: https://support.google.com/youtube/answer/13813322?hl=en [accessed 2026-05-26]
17. Graham G, Halprin M. A long-term vision for YouTube's medical misinformation policies. YouTube. 2023 Aug 15. URL: https://blog.youtube/inside-youtube/a-long-term-vision-for-medical-misinformation-policies/?sjid=13964636406615525454-EU [accessed 2026-05-26]
18. Cinelli M, De Francisci Morales G, Galeazzi A, Quattrociocchi W, Starnini M. The echo chamber effect on social media. Proc Natl Acad Sci U S A 2021 Mar 2;118(9):e2023301118. [doi: 10.1073/pnas.2023301118] [Medline: 33622786]
19. Crupi G, Mejova Y, Tizzani M, Paolotti D, Panisson A. Echoes through time: evolution of the Italian COVID-19 vaccination debate. Proc Int AAAI Conf Web Soc Media 2022 May;16:102-113. [doi: 10.1609/icwsm.v16i1.19276]
20. Gruzd A, Abul-Fottouh D, Song MY, Saiphoo A. From Facebook to YouTube: the potential exposure to COVID-19 anti-vaccine videos on social media. Soc Media Soc 2023;9(1):20563051221150403. [doi: 10.1177/20563051221150403] [Medline: 36751363]
21. Garrett RK. Echo chambers online? Politically motivated selective exposure among internet news users. J Comput-Mediat Commun 2009 Jan;14(2):265-285. [doi: 10.1111/j.1083-6101.2009.01440.x]
22. Ekram S, Debiec KE, Pumper MA, Moreno MA. Content and commentary: HPV vaccine and YouTube. J Pediatr Adolesc Gynecol 2019 Apr;32(2):153-157. [doi: 10.1016/j.jpag.2018.11.001] [Medline: 30445163]
23. Garimella K, Morales GDF, Gionis A, Mathioudakis M. Quantifying controversy on social media. ACM Trans Social Comput 2018 Mar 31;1(1):1-27. [doi: 10.1145/3140565]
24. Kaňková J, Binder A, Matthes J. Health-related communication of social media influencers: a scoping review. Health Commun 2025 Jun;40(7):1300-1313. [doi: 10.1080/10410236.2024.2397268] [Medline: 39258728]

XSL•FO
RenderX

25. Mejova Y, Kalimeri K. COVID-19 on Facebook ads: competing agendas around a public health crisis. In: Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies: Association for Computing Machinery; 2020:22-31. [doi: 10.1145/3378393.3402241]
26. Carrieri V, Guthmuller S, Wübker A. Trust and COVID-19 vaccine hesitancy. Sci Rep 2023 Jun 7;13(1):9245. [doi: 10.1038/s41598-023-35974-z] [Medline: 37286569]
27. Rodrigues F, Ziade N, Jatuworapruk K, Caballero-Uribe CV, Khursheed T, Gupta L. The impact of social media on vaccination: a narrative review. J Korean Med Sci 2023 Oct 16;38(40):e326. [doi: 10.3346/jkms.2023.38.e326] [Medline: 37846789]
28. Major D. COVID-19 misinformation cost at least 2,800 lives and $300M, new report says. CBC News. 2023 Jan 26. URL: https://www.cbc.ca/news/politics/cost-of-covid-19-misinformation-study-1.6726356 [accessed 2026-05-26]
29. Mitropoulos A. 300,000 US COVID deaths could have been averted through vaccination, analysis finds. ABC News. 2022 May 16. URL: https://abcnews.com/Health/300000-us-covid-deaths-averted-vaccination-analysis-finds/story?id=84753284 [accessed 2026-05-26]
30. Data API > Search: list. YouTube. URL: https://developers.google.com/youtube/v3/docs/search/list [accessed 2026-05-26]
31. Betti L, De Francisci Morales G, Gauvin L, et al. Detecting adherence to the recommended childhood vaccination schedule from user-generated content in a US parenting forum. PLoS Comput Biol 2021 Apr;17(4):e1008919. [doi: 10.1371/journal.pcbi.1008919] [Medline: 33901170]
32. Lenti J, Mejova Y, Kalimeri K, et al. Global misinformation spillovers in the vaccination debate before and during the COVID-19 pandemic: multilingual Twitter study. JMIR Infodemiol 2023 May 24;3:e44714. [doi: 10.2196/44714] [Medline: 37223965]
33. Data API > videos. YouTube. URL: https://developers.google.com/youtube/v3/docs/videos [accessed 2026-05-26]
34. Transcript API. YouTube. URL: https://pypi.org/project/youtube-transcript-api [accessed 2026-05-26]
35. Data API > Channels: list. YouTube. URL: https://developers.google.com/youtube/v3/docs/channels/list [accessed 2026-05-26]
36. MacDonald NE, SAGE Working Group on Vaccine Hesitancy. Vaccine hesitancy: definition, scope and determinants. Vaccine (Auckl) 2015 Aug 14;33(34):4161-4164. [doi: 10.1016/j.vaccine.2015.04.036] [Medline: 25896383]
37. Huang LC, Eiden AL, He L, et al. Natural language processing–powered real-time monitoring solution for vaccine sentiments and hesitancy on social media: system development and validation. JMIR Med Inform 2024 Jun 21;12(1):e57164. [doi: 10.2196/57164] [Medline: 38904984]
38. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]
39. Liu Z, Liang W, Luo W, et al. Recent advances in hierarchical multi-label text classification: a survey. arXiv. Preprint posted online on Jul 30, 2023. [doi: 10.48550/arXiv.2307.16265]
40. Rojas KR, Bustamante G, Oncevay A, Cabezudo MAS. Efficient strategies for hierarchical text classification: external knowledge and auxiliary tasks. In: Tetreault J, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Association for Computational Linguistics; 2020. [doi: 10.18653/v1/2020.acl-main.205]
41. Viola P, Jones M. Robust real-time face detection. Int J Comput Vis 2004;57(2):137-154. [doi: 10.1109/iccv.2001.937709]
42. Hearst MA. Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th International Conference on Computational Linguistics: Association for Computational Linguistics; 1992:539-545. [doi: 10.3115/992133.992154]
43. Bykau S, Korn F, Srivastava D, Velegrakis Y. Fine-grained controversy detection in Wikipedia. In: 2015 IEEE 31st International Conference on Data Engineering (ICDE): IEEE; 2015:1573-1584. [doi: 10.1109/ICDE.2015.7113426]
44. Donzelli G, Palomba G, Federigi I, et al. Misinformation on vaccination: a quantitative analysis of YouTube videos. Hum Vaccin Immunother 2018 Jul 3;14(7):1654-1659. [doi: 10.1080/21645515.2018.1454572] [Medline: 29553872]
45. Törnberg P, Törnberg A. Inside a white power echo chamber: why fringe digital spaces are polarizing politics. New Media Society 2024 Aug;26(8):4511-4533. [doi: 10.1177/14614448221122915]
46. Freeman D, Loe BS, Chadwick A, et al. COVID-19 vaccine hesitancy in the UK: the Oxford coronavirus explanations, attitudes, and narratives survey (Oceans) II. Psychol Med 2022 Oct;52(14):3127-3141. [doi: 10.1017/S0033291720005188] [Medline: 33305716]
47. Lorenz-Spreen P, Oswald L, Lewandowsky S, Hertwig R. A systematic review of worldwide causal and correlational evidence on digital media and democracy. Nat Hum Behav 2023 Jan;7(1):74-101. [doi: 10.1038/s41562-022-01460-1] [Medline: 36344657]
48. Geschke D, Lorenz J, Holtz P. The triple-filter bubble: using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. Br J Soc Psychol 2019 Jan;58(1):129-149. [doi: 10.1111/bjso.12286] [Medline: 30311947]
49. Lewandowsky S, van der Linden S. Countering misinformation and fake news through inoculation and prebunking. Eur Rev Soc Psychol 2021 Jul;32(2):348-384. [doi: 10.1080/10463283.2021.1876983]
50. Schneider PJ, Rizoiu MA. The effectiveness of moderating harmful online content. Proc Natl Acad Sci USA 2023 Aug;120(34):e2307360120. [doi: 10.1073/pnas.2307360120]

XSL•FO
RenderX

51. Truong BT, Kim S, Nogara G, et al. Delayed takedown of illegal content on social media makes moderation ineffective. arXiv. Preprint posted online on Aug 9, 2025 URL: https://arxiv.org/html/2502.08841v2 [accessed 2026-06-11]
52. Basol M, Roozenbeek J, van der Linden S. Good news about bad news: gamified inoculation boosts confidence and cognitive immunity against fake news. J Cogn 2020 Jan 10;3(1):2. [doi: 10.5334/joc.91] [Medline: 31934684]
53. Roozenbeek J, van der Linden S. Breaking Harmony Square: a game that "inoculates" against political misinformation. Harv Kennedy Sch Misinform Rev 2020 Nov 6;1(8). [doi: 10.37016/mr-2020-47]
54. Roozenbeek J, van der Linden S, Goldberg B, Rathje S, Lewandowsky S. Psychological inoculation improves resilience against misinformation on social media. Sci Adv 2022 Aug 26;8(34):eabo6254. [doi: 10.1126/sciadv.abo6254] [Medline: 36001675]
55. Pierri F, DeVerna MR, Yang KC, Axelrod D, Bryden J, Menczer F. One year of COVID-19 vaccine misinformation on Twitter: longitudinal study. J Med Internet Res 2023 Feb 24;25:e42227. [doi: 10.2196/42227] [Medline: 36735835]
56. Mena P. Cleaning up social media: the effect of warning labels on likelihood of sharing false news on Facebook. Policy Internet 2020 Jun;12(2):165-183 [FREE Full text] [doi: 10.1002/poi3.214]
57. Pennycook G, Epstein Z, Mosleh M, Arechar AA, Eckles D, Rand DG. Shifting attention to accuracy can reduce misinformation online. Nature New Biol 2021 Apr;592(7855):590-595. [doi: 10.1038/s41586-021-03344-2] [Medline: 33731933]
58. Goldman E. Content moderation remedies. Mich Techn Law Rev 2021;28. [doi: 10.36645/mtlr.28.1.content]
59. Suzor N, Van Geelen T, Myers West S. Evaluating the legitimacy of platform governance: a review of research and a shared research agenda. Int Commun Gaz 2018 Jun;80(4):385-400. [doi: 10.1177/1748048518757142]
60. Bhargava VR, Velasquez M. Ethics of the attention economy: the problem of social media addiction. Bus Ethics Q 2021 Jul;31(3):321-359. [doi: 10.1017/beq.2020.32]
61. Starbird K, Arif A, Wilson T. Disinformation as collaborative work: surfacing the participatory nature of strategic information operations. Proc ACM Hum-Comput Interact 2019;3:1-26. [doi: 10.1145/3359229]
62. Gollust SE, Lantz PM, Ubel PA. The polarizing effect of news media messages about the social determinants of health. Am J Public Health 2009 Dec;99(12):2160-2167. [doi: 10.2105/AJPH.2009.161414] [Medline: 19833981]
63. Niederdeppe J, Shapiro MA, Porticella N. Attributions of responsibility for obesity: narrative communication reduces reactive counterarguing among liberals. Hum Commun Res 2011 Jul;37(3):295-323. [doi: 10.1111/j.1468-2958.2011.01409.x]
64. Barrera O, Guriev S, Henry E, Zhuravskaya E. Facts, alternative facts, and fact checking in times of post-truth politics. J Public Econ 2020 Feb;182:104123. [doi: 10.1016/j.jpubeco.2019.104123]
65. Chen X, Sin SCJ, Theng YL, Lee CS. Why students share misinformation on social media: motivation, gender, and study-level differences. J Acad Librarianship 2015 Sep;41(5):583-592. [doi: 10.1016/j.acalib.2015.07.003]
66. Nagler RH, Vogel RI, Gollust SE, Rothman AJ, Fowler EF, Yzer MC. Public perceptions of conflicting information surrounding COVID-19: results from a nationally representative survey of U.S. adults. PLoS One 2020;15(10):e0240776. [doi: 10.1371/journal.pone.0240776] [Medline: 33085719]
67. Chandio S, Pirwani Dar MD, Nithyanand R. How audit methods impact our understanding of YouTube's recommendation systems. Proc Int AAAI Conf Weblogs Soc Media 2024;18:241-253. [doi: 10.1609/icwsm.v18i1.31311]
68. Chen Z, Liu P, Piao J, Xu F, Li Y. Shorter is different: characterizing the dynamics of short-form video platforms. arXiv. Preprint posted online on Oct 21, 2024. [doi: 10.48550/arXiv.2410.16058]
69. Violot C, Elmas T, Bilogrevic I, Humbert M. Shorts vs. regular videos on YouTube: a comparative analysis of user engagement and content creation trends. In: WEBSCI '24: Proceedings of the 16th ACM Web Science Conference: Association for Computing Machinery; 2024:213-223. [doi: 10.1145/3614419.3644023]
70. Lasco G, Curato N. Medical populism. Soc Sci Med 2019 Jan;221:1-8. [doi: 10.1016/j.socscimed.2018.12.006] [Medline: 30553118]
71. Larson HJ, de Figueiredo A, Xiahong Z, et al. The state of vaccine confidence 2016: global insights through a 67-country survey. EBioMedicine 2016 Oct;12:295-301. [doi: 10.1016/j.ebiom.2016.08.042] [Medline: 27658738]
72. Dubovi I, Tabak I. An empirical analysis of knowledge co-construction in YouTube comments. Comput Educ 2020 Oct;156:103939. [doi: 10.1016/j.compedu.2020.103939]
73. Miyazaki K, Uchiba T, Kwak H, An J, Sasahara K. The impact of toxic trolling comments on anti-vaccine YouTube videos. Sci Rep 2024 Mar 1;14(1):5088. [doi: 10.1038/s41598-024-54925-w] [Medline: 38429466]

Abbreviations

A: against vaccination stance
API: application programming interface
CHV: channel video counts
CR: comment ratio
CS: controversy score
HC3: heteroskedasticity-consistent
HDI: highest density interval

XSL•FO
RenderX

IF: in favor of vaccination stance
IRB: institutional review board
IRR: incidence rate ratio
LLM: large language model
LR: like ratio
N: neutral to vaccination stance
OR: odds ratio
SA: strongly against vaccination stance
SIF: strongly in favor of vaccination stance

Edited by A Mavragani; submitted 26.Nov.2025; peer-reviewed by WG Hussain, Y Fan, Z Zaidi, Z Liu, Z Kmetty; revised version received 20.Apr.2026; accepted 30.Apr.2026; published 19.Jun.2026.

Please cite as:
Tizzani M, Mejova Y
Engagement Trends in Online Vaccine Content: Longitudinal YouTube Study
J Med Internet Res 2026;28:e88519
URL: https://www.jmir.org/2026/1/e88519
doi:10.2196/88519

© Michele Tizzani, Yelena Mejova. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Understanding Public Perceptions of Lung Cancer in China: Infodemiology Study of Baidu Index and Weibo Posts

Zongyuan Li, MD; Jian Zhang, PhD; Yiming Fan, MD; Nan Chen, MD; Lunxu Liu, MD, PhD

Department of Thoracic Surgery and Institute of Thoracic Oncology, West China Hospital of Sichuan University, No. 37, Guoxue Alley, Wuhou District, Chengdu, China

Corresponding Author:
Lunxu Liu, MD, PhD
Department of Thoracic Surgery and Institute of Thoracic Oncology, West China Hospital of Sichuan University, No. 37, Guoxue Alley, Wuhou District, Chengdu, China

Abstract

Background: In China, lung cancer remains a major public health concern and accounts for a substantial proportion of cancer-related deaths nationwide. However, limited research has examined public perceptions of lung cancer in the digital sphere, where health-related information is increasingly disseminated and accessed.

Objective: This study aims to systematically examine patterns of public attention and perceptions toward lung cancer in China by integrating search engine query data and social media content, thereby enhancing current understanding of web-based health information dynamics related to lung cancer.

Methods: Data were collected from Baidu Index (BI) (2011 - 2025) and Sina Weibo (2010 - 2025) to represent web-based search behavior and social media discourse on lung cancer, respectively. Spatiotemporal patterns of BI, per capita Baidu Index (PBI), and Weibo posts were examined to capture temporal trends and spatial variations. Additionally, the spatial autocorrelation of PBI was assessed using global and local Moran *I* statistics. PBI-related explanatory variables were assessed using a spatial panel Durbin model. Topic modeling and lexicon-based sentiment analysis were applied to Weibo content to uncover thematic evolution and emotional polarity across years, sex/organization groups, and user types.

Results: Public attention toward lung cancer, as reflected by BI, increased initially, peaked in 2019, and subsequently declined, whereas Weibo discussions demonstrated a fluctuating but generally upward trend before stabilizing after 2022. Similar temporal patterns were observed across most provinces. Significant spatial heterogeneity was identified, with higher BI levels concentrated in eastern coastal regions and persistently lower levels in western and southwestern provinces. Spatial autocorrelation analysis revealed stable positive clustering over time, with low-low clusters particularly concentrated in southwestern regions such as Guangxi, and no significant high-high clusters were detected. Panel spatial regression analyses indicated that the provincial PBI was positively associated with gross domestic product (GDP) per capita and average years of education per capita, but negatively associated with the urbanization rate. Moreover, significant spatial spillover effects were observed, suggesting that socioeconomic factors were associated not only with local public attention but also with that of neighboring regions. Topic modeling revealed a clear thematic evolution over time. Although personal experiences initially dominated web-based discourse, discussions progressively shifted toward health care service–related issues, which became the most prominent theme by 2025. Sentiment analysis indicated an overall positive emotional tone throughout the study period, with “Good” and “Disgust” representing the predominant positive and negative emotions, respectively. Emotional expression varied across demographic groups and user types, with noticeable differences in both intensity and temporal trends.

Conclusions: This study offers a comprehensive overview of public attention and discourse on lung cancer in China’s digital landscape, providing valuable evidence to inform targeted health communication and policy interventions.

(J Med Internet Res 2026;28:e85058) doi:10.2196/85058

KEYWORDS

lung neoplasms; Baidu index; Weibo; infodemiology; public health; topic modeling; spatial analysis; social media; sentiment analysis

Introduction

Lung cancer remained the most frequently diagnosed malignancy and the leading cause of cancer-related mortality in 2022, accounting for approximately 2.5 million new cases worldwide [1,2]. Despite a global decline in age-standardized lung cancer incidence and mortality rates [3], the burden of lung cancer in China remains high and continues to rise [4]. In 2021, China accounted for approximately 40% of global lung cancer deaths, with over 800,000 fatalities [3,5]. China also bears the

XSL•FO
RenderX

highest share of the global cancer economic burden, with lung cancer representing the costliest type [6]. This growing disease burden presents a substantial public health challenge, particularly given China's aging population and high prevalence of tobacco use [7]. According to Surveillance, Epidemiology, and End Results 21 data (2015‑2021), the 5-year relative survival rate for lung cancer is 28.1%, with more than half of cases diagnosed at a distant stage and only about one-quarter at a localized stage [8]. Randomized controlled trials have demonstrated that low-dose computed tomography screening significantly reduces lung cancer mortality [9,10]. Indeed, early diagnosis and appropriate treatment are critical for improving lung cancer prognosis. However, increasing evidence suggests that inadequate public awareness and widespread misperceptions about lung cancer risk factors, symptoms, screening, and treatment options contribute to delays in diagnosis and underutilization of health care services [11,12]. A recent cross-sectional survey of 31,184 urban Chinese residents revealed that 76.12% of respondents acknowledged having limited knowledge of lung cancer [13]. Moreover, pervasive stigma and misinformation surrounding lung cancer may further deter individuals from engaging in health-seeking behavior and receiving timely treatment [14,15].

Therefore, enhancing public awareness and knowledge of lung cancer has emerged as a key priority in global cancer-control strategies. To this end, understanding the current attention and prevailing attitudes toward lung cancer is a prerequisite for developing effective public awareness strategies that address existing knowledge gaps and prioritize areas for intervention. Relatively few studies have examined these issues in the Chinese population. For instance, an analysis of 481 questionnaires completed by the offspring of lung cancer patients indicated insufficient knowledge, favorable attitudes, and inadequate practice regarding lung cancer risk [16]. A nationwide survey of 2093 health care practitioners across 706 hospitals in China revealed substantial disparities in knowledge, attitudes, and practices concerning lung cancer palliative care, highlighting the pressing need for comprehensive training initiatives [17]. Nonetheless, prior research in the Chinese population has predominantly relied on regionally confined samples, narrowly defined groups (eg, the offspring of lung cancer patients and health care practitioners), and structured questionnaires, often using small sample sizes and limited timeframes, thereby restricting the generalizability of their findings [13,16-18]. Consequently, there remains a notable lack of nationally representative studies that comprehensively investigate both macro-level public attention and micro-level discursive practices related to lung cancer across public domains in China.

With the widespread adoption of digital platforms in China, an increasing number of individuals are using search engines and social media to access and share health-related information [19]. According to the 56th Statistical Report on China's Internet Development, as of June 2025, China had 1.123 billion internet users and an internet penetration rate of 79.7%, with social media users reaching 1.107 billion (representing 98.6% of the web population) [20]. In China's search engine market, Baidu holds a dominant position with a penetration rate of 90.9%, serving as the primary platform through which users access web-based information [21]. The Baidu Index (BI), derived from massive user behavior data, quantifies search volumes for specific keywords over time and geographic regions, generating a relative index value [22]. BI reflects users' active search interests and has been widely used in public health research as a proxy for monitoring information-seeking behavior and public awareness [23]. Regarding social media, Weibo, launched by Sina Corporation in 2009, has become China's leading microblogging platform, with 587 million monthly and 257 million daily active users as of September 2024, and serves as a major venue for real-time content sharing and public discourse [24]. Prior studies have leveraged these web-based data to monitor health behavior trends and identify information gaps across diseases, offering real-time insights into population-level perceptions [19,22,23,25,26]. However, few studies have systematically analyzed how lung cancer is represented and discussed via web in China. To address this gap, this study aims to investigate public perceptions of lung cancer in China using a framework that integrates BI data from large-scale search queries and posts from the social media platform Weibo. This dual-perspective approach combines information-seeking behavior with social communication dynamics, providing a more comprehensive and multidimensional understanding of public attention than single-source or traditional surveillance methods. By examining public attention and web-based discourse, this study seeks to provide evidence for health communication strategies and public health initiatives on lung cancer in China, with the potential to enhance awareness, promote earlier detection, and strengthen disease management.

Methods

An overview of the study design and analytical workflow is presented in Figure 1.

XSL•FO
RenderX

Figure 1. Overview of study design and analytical workflow. BI: Baidu Index; EAPC: estimated annual percent change; GDP: gross domestic product; PBI: per capita Baidu Index.

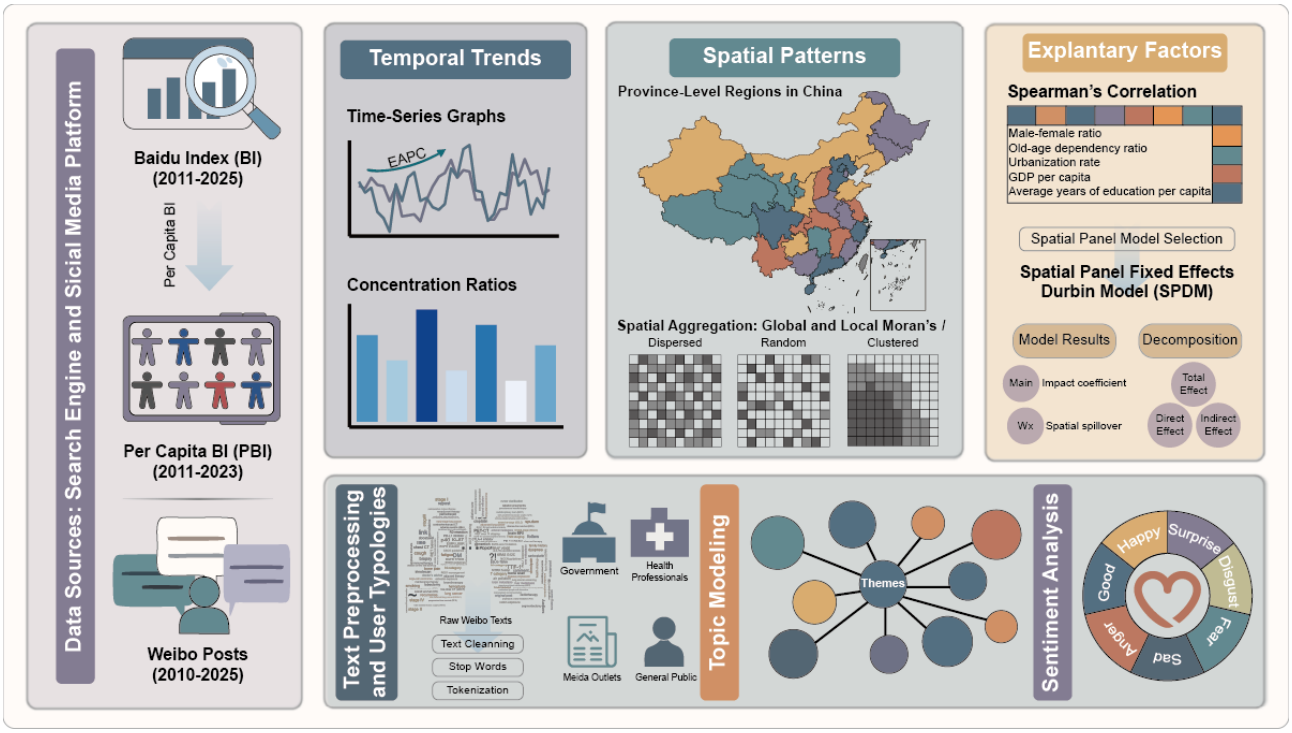

Data Sources

Search engine data related to lung cancer was obtained from the official website of BI. Researchers can use this platform to quantify the frequency of keyword searches on Baidu across different regions and over defined temporal spans. In this study, daily BI values for Chinese-language keywords referring to lung cancer, including "lung cancer," "neoplasms of the lung," "lung adenocarcinoma," and "lung squamous cell carcinoma" were collected for 31 provincial-level regions (22 provinces, 5 autonomous regions, and 4 municipalities) in mainland China between January 1, 2011, and August 31, 2025. Hong Kong, Macao, and Taiwan were excluded from the analysis. Monthly and yearly average BI values were defined as the mean of daily BI values within each respective month or year. To estimate search activity per capita within each province and enable regional comparisons, the per capita Baidu Index (PBI) was calculated as BI divided by the year-end population, scaled per 10,000 individuals [26].

We retrieved relevant original posts (also referred to as thread posts) from Weibo using its keyword-based advanced search function. Reposts (shares) and comment replies were not included in the dataset. Combinations of Chinese keywords related to lung cancer, including "lung cancer," "neoplasms of the lung," "lung adenocarcinoma," and "lung squamous cell carcinoma," were selected to retrieve relevant posts between January 1, 2010 (roughly 4 months after Weibo's launch in China), and August 31, 2025. These search keywords were primarily determined based on the researchers' prior knowledge. In addition, user-related metadata (eg, username, sex/organization, personal description, avatar, verification status, number of posts, follower number, following number, and other information) were also collected. In this study, Weibo users were categorized as male (35,792/105,549, 33.91%), female (53,800/105,549, 50.97%), or organization (15,957/105,549, 15.12%) according to their self-reported sex and verification information provided in their profiles (Table 1) [27]. Given that all accounts report sex and organizational accounts can be separately identified via verification information, we classified them as a distinct category. Institutional accounts were not treated as a sex category but rather as a distinct user type to better capture user heterogeneity. Based on user information verified by Weibo, users were also categorized into 4 types, including government (4478/105,549, 4.24%), health professionals (4093/105,549, 3.88%), media outlets (3827/105,549, 3.63%), and the general public (93,151/105,549, 88.25%) (Table 1). We performed text preprocessing to clean the dataset as a preparatory step to text-mining analysis (Multimedia Appendix 1 [26-33] provides further details). Following duplicate removal and text preprocessing, a total of 268,756 original posts from 105,549 unique Weibo users were retained for subsequent analysis.

XSL•FO
RenderX

Table . Classification of Weibo users in this study by sex/organization and user type. Classification of Weibo users was based on verification information and self-reported profile details. Examples are illustrative and anonymized; institutional names are shown only as representative categories, and no personal identifiers are disclosed.

Group	Description	Examples	Users (n=105,549), n (%)	Posts (n=268,756), n (%)
Sex/organization				
Male	Self-reported male users	Mr. Zhang (verified personal account, health blogger, male)	35,792 (33.91)	94,959 (35.33)
Female	Self-reported female users	Ms. Li (verified personal account, writer, female)	53,800 (50.97)	80,291 (29.88)
Organization	Institutional users identified from verification information	Beijing Cancer Hospital official account	15,957 (15.12)	93,506 (34.79)
User type				
Government	Chinese governmental bodies at different levels	National Health Commission, Taoyuan County People's Government (Hunan Province) official account	4478 (4.24)	25,818 (9.61)
Health professionals	Users with medical or scientific expertise, including clinicians, researchers, and health care providers	Oncologist at a tertiary hospital, TCM[a] practitioner	4093 (3.88)	46,265 (17.21)
Media outlets	News organizations and journalists, including newspapers, broadcasters, and digital media	China Media Group, Renwu magazine	3827 (3.63)	33,480 (12.46)
General public	Users outside the above categories, typically unverified individuals sharing personal views or experience	Ordinary Weibo user, patient sharing experience	93,151 (88.25)	1,63,193 (60.72)

[a]TCM: traditional Chinese medicine.

Provincial-level demographic and socioeconomic explanatory variables were sourced from the China Statistical Yearbook (2011‑2023) and corresponding provincial statistical yearbooks. The demographic variables comprised the male-female ratio, old-age dependency ratio (the proportion of individuals aged 65 years and older to those aged 15‑64 years), and urbanization rate (the share of the urban population in the total population at year-end). Meanwhile, the socioeconomic factors considered were gross domestic product (GDP) per capita, an indicator of economic development, and average years of education per capita, which reflect the educational level of the region (Multimedia Appendix 1 [26-33] provides details on calculation). Due to the publication lag of official statistics, complete and consistent provincial-level demographic and socioeconomic data were available only through 2023, with data for 2024 onward unavailable at the time of analysis. The descriptive statistical summaries of these variables can be found in Figure S1 and Table S1 in Multimedia Appendix 1 [26-33].

Statistical Analysis

An integrated framework was developed to investigate macro-level public attention and micro-level web-based discourse on lung cancer, comprising four analytical dimensions that addressed temporal trends, spatial patterns, a spatial panel econometric model, and text-mining analysis. All analyses were conducted with R (version 4.4.2; R Foundation for Statistical Computing, Vienna, Austria). A two-tailed *P* value less than .05 was regarded as statistically significant.

First, temporal trends in BI, PBI, and the number of Weibo posts were depicted using time-series graphs, and provincial BI and PBI from 2011 to 2023 were subsequently assessed through estimated annual percent change (EAPC). In the time-series graphs, in addition to displaying mean values with point-line plots, locally weighted regression smoothing and generalized linear models were fitted using the base R "stats" package (version 4.5.1) to further characterize temporal trends. Details of the EAPC calculation are provided in Multimedia Appendix 1 [26-33]. To assess temporal aggregation of the BI for lung cancer, we applied the concentration ratio method (detailed in Multimedia Appendix 1) [26-33]. The index M ranges from 0 to 1; larger values denote stronger aggregation, while values below 0.3 indicate a relatively uniform distribution over time. This metric captures within-year seasonality rather than multi-year or long-term structural trends.

Second, we examined the spatial distribution patterns and spatial autocorrelation of search engine activity related to lung cancer. In this study, we defined the spatial weight matrix using the widely used Rook adjacency method, under which 2 provinces with a common boundary were assigned a weight of 1 and those without a boundary a weight of 0 [34]. Hainan Province, being an island without direct borders with other provinces, was

XSL•FO
RenderX

designated as adjacent to Guangdong Province, the geographically closest province [26]. The spatial adjacency relationships of each province-level region in China are presented in Table S2 in Multimedia Appendix 1 [26-33]. Spatial autocorrelation analysis was used to assess the spatial clustering of the PBI for lung cancer [35]. Specifically, the global Moran *I* was applied to evaluate spatial dependence in provincial lung cancer PBI, while the local Moran *I* identified spatial aggregation or outlier patterns among adjacent provinces.

To examine the determinants of web-based search volume, this study applied Spearman rank correlation for initial association analysis and performed a variance inflation factor (VIF) test to detect potential multicollinearity. A threshold of VIF<10 was adopted as the criterion for variable retention, and all variables satisfied this criterion (Table S3 in Multimedia Appendix 1 [26-33]). Subsequently, following a series of model selection procedures (Figure S2 and Table S4 in Multimedia Appendix 1 [26-33]), the spatial panel Durbin model (SPDM) was used to quantify the effects, which explicitly incorporates spatial interactions and accounts for regional heterogeneity, compared with traditional panel models [28]. The SPDM model was estimated using data from 2011 to 2023 only, ensuring full temporal alignment between the dependent variables and all explanatory variables. A more detailed description of the spatial panel econometric model is provided in Multimedia Appendix 1 [26-33].

Moreover, we applied latent Dirichlet allocation (LDA) for topic modeling to identify lung cancer-related themes. LDA is a semiautomated and unsupervised machine learning method for computer-assisted content analysis that identifies latent semantic patterns in large text corpora [36,37]. An optimal number of topics was identified based on the LDA modeling results (Figure S3 in Multimedia Appendix 1 [26-33]), with *k* set to 50 to achieve a balance between interpretability and fine-grained thematic resolution. To interpret and label the LDA-generated topics and extract overarching themes, we examined 3 types of information for each topic, including top words, words both prevalent and distinctive, and representative Weibo posts. Topics were independently labeled by 2 coauthors using an open coding approach, followed by iterative comparison and discussion to reconcile differences and refine interpretations. Final themes were established through consensus. Furthermore, sentiment analysis was conducted to quantify the positive or negative sentiment expressed in lung cancer-related posts. Considering the need for transparent scoring and stable outputs in large-scale public health analysis, we used a lexicon-based approach instead of more complex deep learning or large language model-based methods. Sentiment was assessed using the DLUT-Emotion Ontology, a Chinese emotion lexicon developed by the Information Retrieval Laboratory at Dalian University of Technology [29,38]. Building upon internationally recognized emotion classification frameworks and integrating the linguistic characteristics of Chinese texts as well as existing emotion lexicons, this study categorizes emotions into 7 primary types, including “Happy,” “Good,” “Anger,” “Sad,” “Fear,” “Disgust,” and “Surprise.” Emotion Taxonomy and Example Lexemes of DLUT-Emotion Ontology are presented in Table S5 in Multimedia Appendix 1 [26-33]. After text preprocessing, tokens in each post were matched against the ontology. Each matched lexeme is annotated with an emotion category, polarity, and an intensity value ranging from 0 to 9. Sentiment scores were calculated by aggregating the intensity values of all matched tokens, weighted by their polarity (positive values for positive polarity and negative values for negative polarity). When multiple emotion categories were present within a single post, their contributions were cumulatively incorporated into the overall score. Continuous sentiment scores ranging from negative to positive were computed for each post, both for each emotion type and for an overall score. These scores have no predefined upper or lower bounds, as their magnitude depends on the cumulative number of matched tokens, their assigned intensity, and polarity. Sentiment scores were further analyzed by sex/organization, user category, and post theme.

Ethical Considerations

This study analyzed publicly available, aggregated BI data and publicly accessible Weibo posts without direct interaction with individuals. Consistent with established practices in infodemiology research [22,27,39], studies analyzing anonymized, publicly available web data are generally considered exempt from formal ethics review; therefore, Institutional Review Board approval was not required. No identifiable personal information was collected or reported, and all data were analyzed in aggregated or deidentified form to ensure privacy and confidentiality. Because no human participants were recruited, informed consent and compensation were not applicable. No images or materials in the manuscript allow identification of individual users.

Results

Temporal Characteristics

Figure 2A and B and Figure S4 in Multimedia Appendix 1 [26-33] present the temporal distribution of public attention to lung cancer, including the BI from January 2011 to August 2025 and the number of Weibo posts from January 2010 to August 2025. Over the course of approximately 15 years, the average BI was 4670.62. The temporal pattern of BI for lung cancer was characterized by an upward trend followed by a subsequent decline. The highest recorded BI was observed in May 2019, at 9,119.36. By August 2025, the BI had decreased to 2807.42. Besides, a comparable pattern of BI fluctuations was observed across the various province-level regions. Throughout the study period, the concentration ratios of BI ranged from 0.01 to 0.12, with the maximum value (M=0.12) recorded in 2019 (Table S6 in Multimedia Appendix 1 [26-33]). Moreover, the number of Weibo posts exhibited a fluctuating but overall upward trend starting in 2015, with growth gradually leveling off after 2022.

XSL•FO
RenderX

Figure 2. Temporal trends of national BI (A) and Weibo posts (B) for lung cancer in China. Monthly average values were reported, and LOESS, along with segmented GLM models, were applied to capture more detailed trends. BI: Baidu Index; GLM: generalized linear model; LOESS: locally weighted regression.

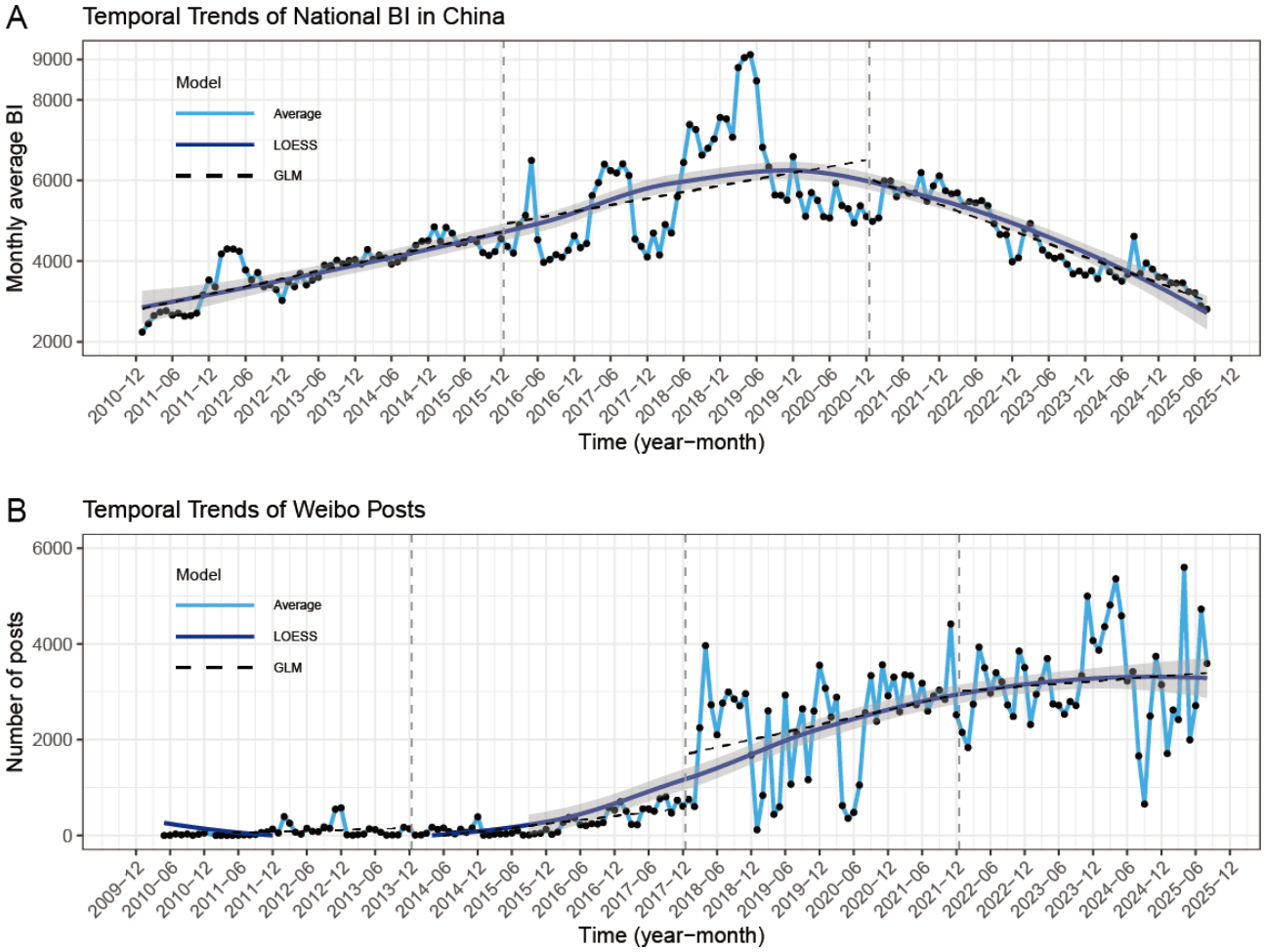

Spatial Patterns

Figure 3A-F presents the spatial distribution of lung cancer BI and PBI across Chinese provinces from 2011 to 2023. Provinces with higher BI values for lung cancer include Beijing, Guangdong, Jiangsu, Shandong, and Zhejiang, which are primarily situated along China's eastern and southeastern coast. Western and southwestern provinces, including Gansu, Ningxia, Qinghai, Tibet, Xinjiang, and Hainan, consistently exhibited the lowest BI levels. In contrast to BI, the PBI demonstrates a partially distinct spatial pattern, with elevated levels observed in Beijing, Hainan, Ningxia, Qinghai, Shanghai, and Tianjin. From 2011 to 2023, both the BI and PBI of lung cancer showed a general upward trend across Chinese provinces, with Sichuan and Yunnan showing the most pronounced increases in BI and PBI, respectively, as indicated by the highest EAPC values. Table 2 and Figure S5 in Multimedia Appendix 1 [26-33] present the global and local Moran *I* statistics for the PBI of lung cancer from 2011 to 2023. During the entire period, the global Moran *I* values remained positive, ranging from 0.23 to 0.31, indicating a stable pattern of clustering and positive spatial autocorrelation. Local Moran *I* analysis revealed stable low-low clustering of lung cancer PBI in southwestern China, notably in Guangxi. No significant high-high clusters, high-low outliers, or low-high outliers were identified.

XSL•FO
RenderX

Figure 3. Spatial patterns of BI (A, B) and PBI (D, E) for lung cancer in 2011 and 2023, and EAPC (C, F) from 2011 to 2023, among provincial-level regions in China. BI: Baidu Index; EAPC: estimated annual percent change; PBI: per capita Baidu Index.

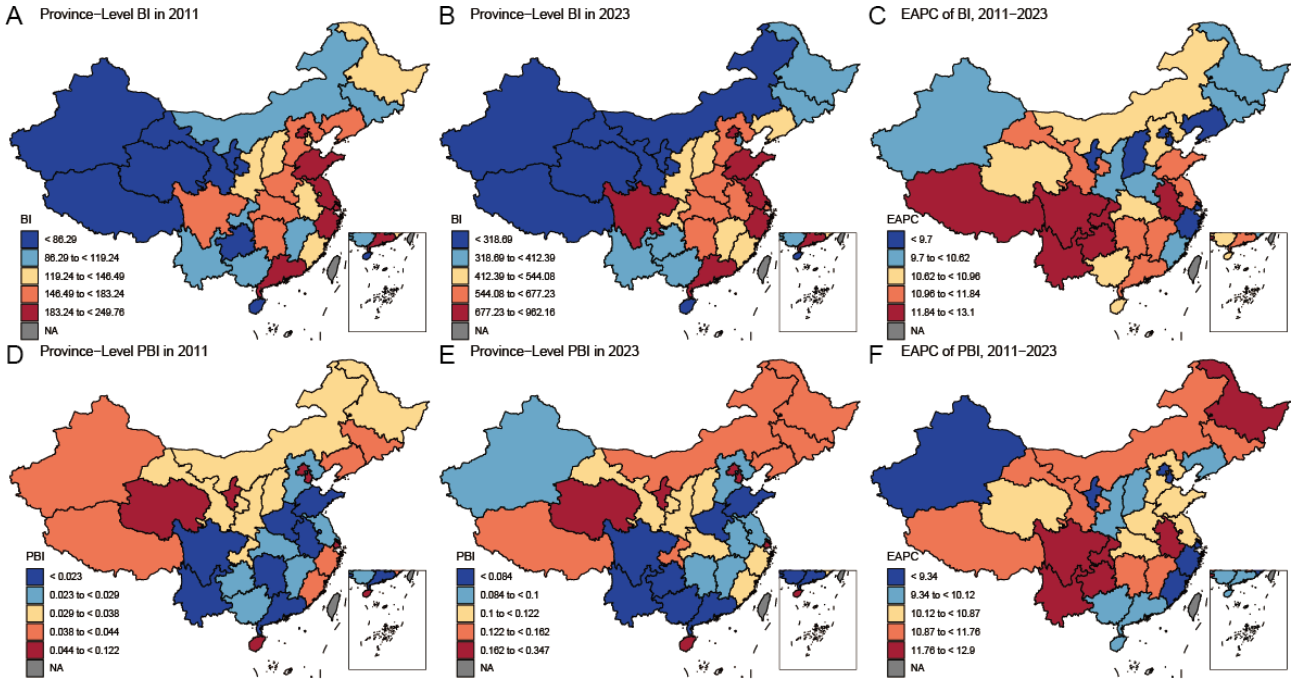

Table . Spatial aggregation of the per capita Baidu Index for lung cancer, from 2011 to 2023. Provinces with nonsignificant local Moran *I* are not listed.

Year	Global Moran *I*		Local Moran *I* (Low-Low cluster)
	Moran *I*	*P* value	
2011	0.27	.03	Guangxi, Guizhou, Shandong
2012	0.27	.03	Guangxi, Guizhou, Hunan
2013	0.31	.01	Guangxi, Guizhou, Shandong
2014	0.30	.02	Guangxi, Guizhou
2015	0.28	.02	Guangxi, Guizhou
2016	0.27	.03	Guangxi, Shandong
2017	0.26	.03	Guangxi
2018	0.25	.03	Guangxi, Guizhou
2019	0.23	.046	Guangxi, Guizhou
2020	0.23	.04	Guangxi, Shandong
2021	0.25	.03	Guangxi, Shandong
2022	0.26	.03	Guangxi, Shandong
2023	0.27	.02	Guangxi, Shandong

Explanatory Factors

Preliminary Spearman rank correlation analysis suggested that, in certain years, urbanization rate, average years of education per capita, and GDP per capita were positively correlated with PBI of lung cancer, whereas old-age dependency ratio and male-female ratio showed negative correlations (Figure S6 in Multimedia Appendix 1 [26-33]). Table 3 displays the SPDM results for the PBI of lung cancer. Higher GDP per capita and education levels were positively correlated with PBI of lung cancer, whereas the urbanization rate was negatively correlated. Significant positive spatial spillover effects were observed for the PBI of lung cancer (spatial ρ=0.380, *P*<.001). The old-age dependency ratio, GDP per capita, and average years of education per capita were significantly associated with the PBI for lung cancer in neighboring regions. Spillover effects were further examined by decomposing the SPDM into direct and indirect effects using partial differential decomposition. The decomposition of direct and indirect effects revealed that GDP per capita and average years of education per capita were positively correlated with the PBI for lung cancer locally and in adjacent regions, while urbanization was significantly negatively associated with PBI in both spatial contexts.

XSL•FO
RenderX

Table . Results of the spatial panel Durbin model examining the associations between explanatory variables and per capita Baidu Index.

Variable	Model results (95% CI)		Decomposition of spatial effects (95% CI)		
	Coefficient	Spatial spillover (Wx)	Direct effect	Indirect effect	Total effect
Spatial autoregressive coefficient (ρ)	0.380 (0.272 to 0.489)[a]	— [b]	—	—	—
Male-female ratio	−0.007 (−0.016 to 0.001)	−0.009 (−0.026 to 0.008)	−0.008 (−0.016 to 0.001)	−0.004 (−0.01 to 0.001)	−0.012 (−0.026 to 0.002)
Old-age dependency ratio	0 (−0.001 to 0.001)	−0.003 (−0.006 to −0.001)[c]	0 (−0.001 to 0.001)	0 (−0.001 to 0.001)	0 (−0.002 to 0.002)
Urbanization rate	−0.006 (−0.007 to −0.004)[a]	0 (−0.003 to 0.003)	−0.006 (−0.007 to−0.005)[a]	−0.003 (−0.005 to −0.002)[a]	−0.009 (−0.012 to −0.007)[a]
Ln (GDP[d] per capita)[e]	0.039 (0.014 to 0.065)[c]	−0.1 (−0.148 to −0.052)[a]	0.041 (0.015 to 0.068)[c]	0.023 (0.007 to 0.045)[f]	0.064 (0.022 to 0.11)[c]
Average years of education per capita	0.027 (0.018 to 0.036)[a]	−0.025 (−0.045 to −0.004)[f]	0.029 (0.02 to 0.038)[a]	0.016 (0.009 to 0.027)[c]	0.045 (0.029 to 0.062)[a]

[a]Statistically significant at .001.

[b]Not applicable.

[c]Statistically significant at .01.

[d]GDP: gross domestic product.

[e]Ln (GDP per capita) indicates the logarithm of GDP per capita.

[f]Statistically significant at .05.

Topic Modeling

Topic modeling of 268,756 lung cancer-related Weibo posts yielded 50 latent topics. Table S7 in Multimedia Appendix 1 [26-33] presents the 50 topics together with the top 20 unique keywords for each topic, which were translated into English by a native Chinese-speaking researcher. Table 4 illustrates 10 main lung cancer-related themes, including health care service (28,902/268,756, 10.75%), treatment (32,609/268,756, 12.13%), prevention (42,096/268,756, 15.66%), traditional Chinese medicine (TCM) (15,112/268,756, 5.62%), personal experience (75,478/268,756, 28.08%), comorbidities (8656/268,756, 3.22%), screening and diagnosis (20,548/268,756, 7.65%), health communication (21,194/268,756, 7.89%), academic research (12,442/268,756, 4.63%), and symptoms (11,719/268,756, 4.36%). As shown in Figure 4A, personal experience emerged as the most prevalent lung cancer theme, followed by prevention, treatment, and health care service. Notably, starting around late 2023, the number of Weibo posts related to personal experience and screening and diagnosis declined markedly, while on health care service rose rapidly, overtaking personal experience in August 2025 as the dominant theme. Figure 4 illustrates the distribution of themes by sex/organization and user type. Female users posted more personal experience–related content and fewer posts on TCM than male users, while the vast majority of posts on health care service were made by institutional accounts. When stratified by user type, government accounts most frequently addressed prevention, health professionals emphasized treatment, media outlets focused on health care service and personal experience, and the general public predominantly shared personal experiences.

XSL•FO
RenderX

Table . Themes identified from the 50 lung cancer-related topics on Weibo.

Frame	Posts (n=268,756), n (%)	Description	Subtopics	Representative keywords
Health care service	28,902 (10.75)	Hospital services, insurance coverage, community care, and medical technologies	1, 15, 28, 29, 35, 45, 48	Service, genetic testing, consultation, medical insurance, cost, technology, precision, AI[a], hospital, expert, tumor, professor, product, market, project, health, hygiene, community
Treatment	32,609 (12.13)	Clinical interventions including surgery, chemotherapy, radiotherapy, targeted therapy, and immunotherapy	2, 10, 14, 22, 33, 47	Treatment, patient, combination, first-line, chemotherapy, PD-1, radiotherapy, tumor, drug, survival, immunotherapy, side effect, targeted drug, surgery, thoracic surgery, pathology
Prevention	42,096 (15.66)	Lifestyle and environmental factors related to lung cancer risk reduction	3, 5, 12, 17, 18, 41, 49	Prevention, health, exercise, lifestyle, sleep, reduce, risk, immunity, diet, stress, smoking, cooking oil fume, pollution, air, indoor, environment, food products, fat, obesity, nutrition
TCM[b]	15,112 (5.62)	Use of Chinese herbal medicine, syndrome differentiation, and health preservation practices	4, 24, 26	Traditional Chinese medicine, Chinese herbal medicine, anti-cancer, effect, lung cancer, relief, licorice, pinellia, resolve phlegm, astragalus, prescription, poria, health preservation
Personal experience	75,478 (28.08)	Patient and family narratives of illness, treatment, and caregiving	6, 7, 11, 13, 23, 30, 31, 34, 37, 46	Lung cancer, one year, detected, surgery, hope, wife, last year, family, anti-cancer, friend, aunt, hospitalization, elderly, evening, feeling, several days, go home, death, passing away, donation
Comorbidities	8656 (3.22)	Coexisting conditions such as cardiovascular disease, diabetes, and infections	8, 9, 44	Disease, diabetes, health, hypertension, cardiovascular, heart disease, elderly, blood pressure, prevention, blood glucose, skin, COVID-19, pneumonia, patient, vaccine, fever, vein, hormone
Screening and diagnosis	20,548 (7.65)	Early detection through imaging, physical examination, tumor markers, and clinical vigilance	16, 21, 27	Examination, screening, CT[c], recommendation, testing, detection,tumor marker, ultrasound, age 50, high-risk population, diagnosis, chest radiograph, dosage, health, pulmonary nodule
Health communication	21,194 (7.89)	Public communication and education through media, experts, and online platforms	19, 32, 39, 40	China, world, lung cancer, science, life expectancy, society, mortality rate, doctor, hospital, examination, recommendation, answer, consultation, science popularization, trend

XSL•FO
RenderX

Frame	Posts (n=268,756), n (%)	Description	Subtopics	Representative keywords
Academic research	12,442 (4.63)	Scientific studies on lung cancer	20, 36, 38, 42, 43	Research, discovery, cell, immunity, tumor, cancer cell, function, inhibition, system, growth, mechanism, gene, scientist, this study, antibody, direction, related, strain, product, antigen
Symptoms	11,719 (4.36)	Bodily signs and clinical manifestations	25, 50	Symptom, pain, manifestation, signal, body, vigilance, cause, situation, abnormal, nerve, lump, examination, site, persistent, skin, hoarseness, compression, ache, joint, cough, phlegm

[a]AI: artificial intelligence.

[b]TCM: traditional Chinese medicine.

[c]CT: computed tomography.

XSL•FO
RenderX

Figure 4. Temporal trends of Weibo posts by theme (A) and Sankey diagram of gender/organization, theme, and user type (B). TCM: traditional Chinese medicine.

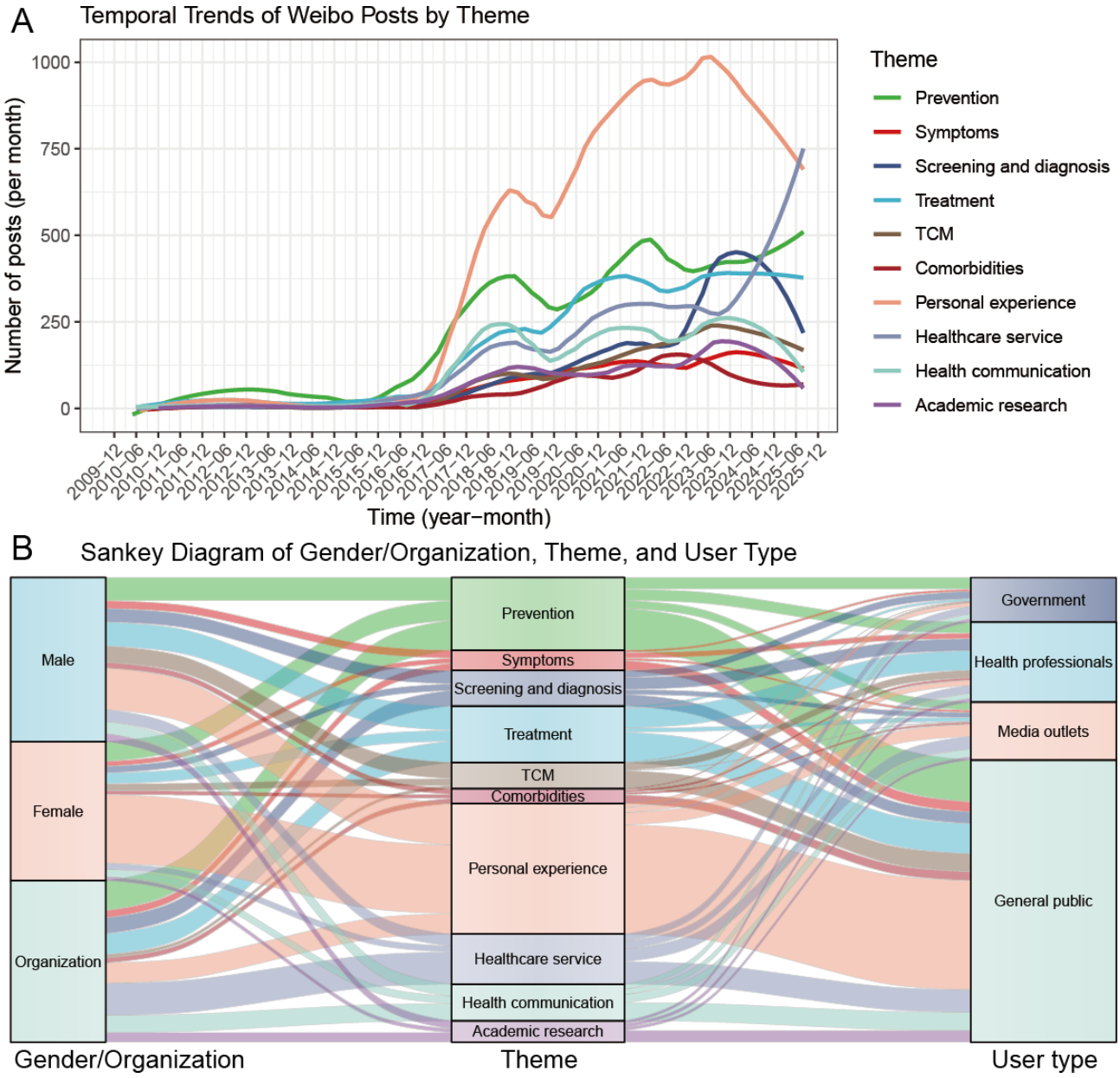

Sentiment Analysis

Figure 5A depicts the temporal trends in emotion scores of Weibo posts related to lung cancer. In the sentiment scores, the sign indicates emotional polarity (positive or negative), whereas the absolute value reflects the intensity of the emotion. From 2010 to 2025, the total emotion score was consistently positive and increased gradually with fluctuation. Among the 7 sentiment categories, “Good” and “Disgust” were the dominant positive and negative emotions, respectively, both of which increased in intensity over time. Moreover, “Happy” remained relatively stable, and other emotions were expressed only at low levels. Further subgroup-specific results of the sentiment analysis are shown in Figure 5, Figure S7 and Table S8 in Multimedia Appendix 1 [26-33]. Sex/organization comparisons showed higher sentiment scores among institutional accounts, followed by male and then female users. By user type, the order was media outlets, health professionals, government, and the general public. Thematic analysis revealed health care services as the most positively scored theme, followed at a distance by academic research and TCM, whereas comorbidities and symptoms scored negatively. Moreover, sentiment trajectories varied across subgroups, with government, media outlets, prevention, personal experience, symptoms, comorbidities, and academic research exhibiting notable downward trends in recent years.

XSL•FO
RenderX

Figure 5. Temporal trends of emotion scores by sentiment category (A) and emotion scores by sex/organization, user type, and theme (B). TCM: traditional Chinese medicine.

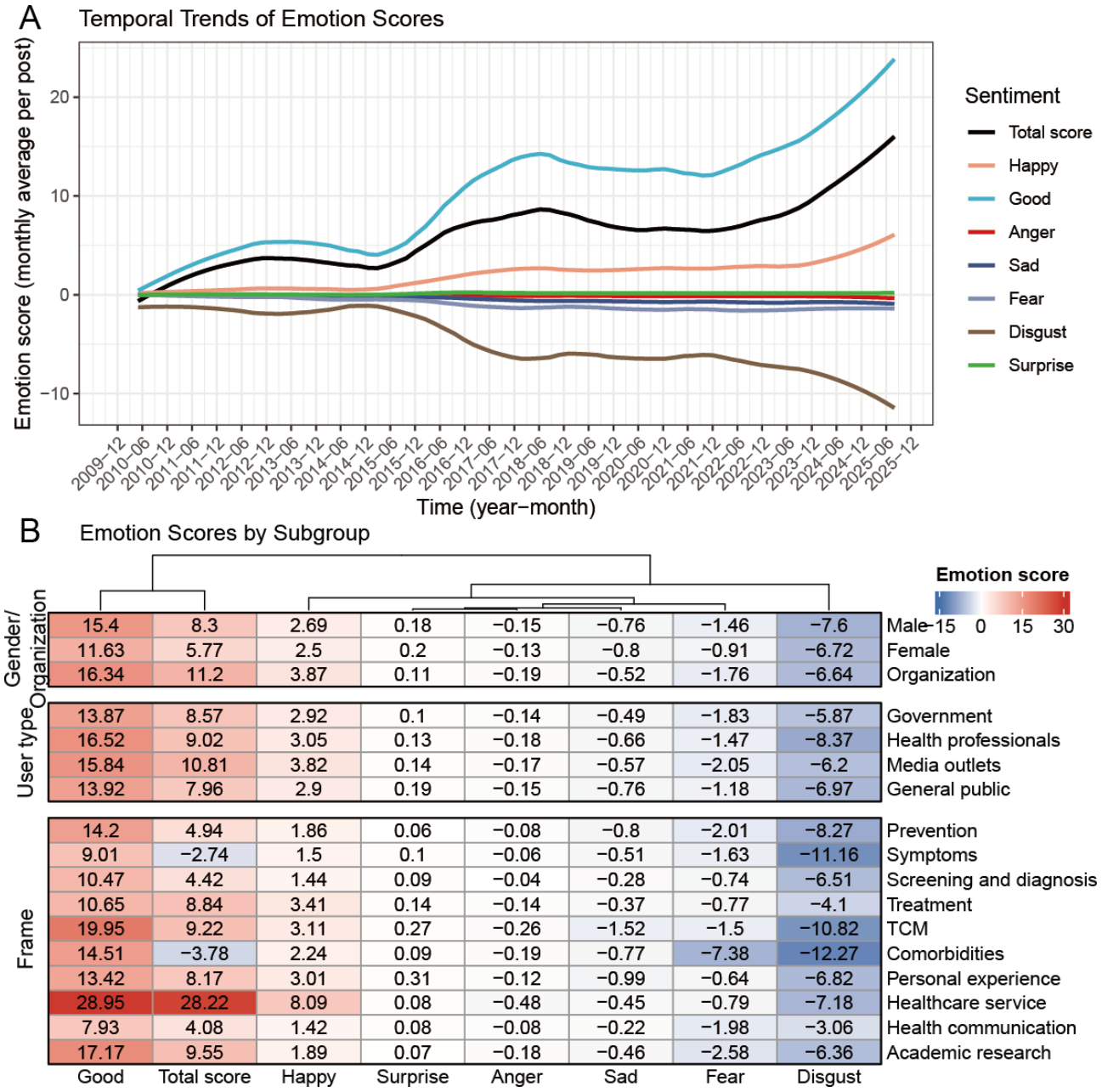

Discussion

Principal Findings

This study comprehensively examined public attention, perceptions, and sentiment toward lung cancer in China by integrating search engine and social media data. Temporally, search engine data showed that public attention peaked in 2019 and declined thereafter, while Weibo discussions have remained relatively stable since 2022. Spatial heterogeneity and autocorrelation were observed, with low-low clusters concentrated in the southwestern region, particularly Guangxi. GDP per capita and average years of education were positively associated with PBI, whereas urbanization rate showed a negative correlation. Topic modeling revealed a temporal shift in themes, with personal experiences initially predominant and health care services emerging as the leading theme in 2025. Overall sentiment was generally positive, mainly characterized by "Good" and "Disgust," with variations across years, sex/organization, and user types. These findings provide a foundation for future research and health communication strategies.

To the best of our knowledge, this study is the first to offer a dual-perspective analysis that captures both macro-level public engagement and micro-level social discourse regarding lung cancer in China. Existing research in English-language digital contexts has primarily focused on how lung cancer discourse manifests on platforms such as Twitter and Facebook, revealing distinct communication dynamics and thematic contents [40,41]. However, China's socioeconomic and cultural context shapes digital health communication in distinct ways. While Chinese social platforms such as Weibo align with some global patterns as a space for personal narrative expression, it is also deeply

XSL•FO
RenderX

embedded in China's local health discourse and cultural dissemination. TCM, for example, frequently appeared in user narratives as both a therapeutic option and a cultural lens [42]. Examining Weibo's health discourse, this study provides a contextual insight into how lung cancer is communicated in China's digital sphere.

The temporal trends observed in public attention to lung cancer have followed an inverted U-shaped trajectory over the past 15 years. The sharp rise in BI before its 2019 peak may reflect increased public concern promoted by policy actions, media coverage, or celebrity illness disclosures. Such surges align with event-driven information-seeking, characteristic of public health crises, where perceived threats heighten public attention [43]. The post-2019 decline, however, indicates reduced public attention, potentially due to the normalization of lung cancer or a diversion of public focus to emerging issues such as COVID-19 [42]. In addition, changes in search algorithms and the growing use of other web platforms may also have influenced the observed trends in public attention. Notably, the continued growth of Weibo activity until 2022 may imply a shift in public discourse from active information-seeking to more passive or peer-oriented engagement. The observed divergence suggests a functional distinction: search engines provide immediate information access, whereas social media enable ongoing, socially embedded discourse [41,44]. Collectively, these patterns indicate that platform-specific strategies are essential for sustaining public awareness and reinforcing lung cancer-related health communication.

Spatial disparities in public attention toward lung cancer were evident across provincial-level regions in China. Epidemiological data from 2021 demonstrated that the age-standardized incidence and mortality rates of lung cancer were highest in northeast Chinese provinces, particularly in Heilongjiang, Tianjin, and Liaoning, while Xinjiang, Gansu, and Tibet reported the lowest incidence rates [45]. However, the regional distribution of lung cancer burden diverges notably from that of BI or PBI. The high BI for lung cancer in eastern and southeastern coastal regions may reflect their more developed socioeconomic structures and higher population density [26,46]. In contrast to BI, PBI clusters emerged in regions such as Beijing, Hainan, Ningxia, Qinghai, Shanghai, and Tianjin, indicating that public attention may be shaped more by contextual factors (eg, media coverage, health education programs, or local policy) than by actual disease burden alone. This pattern may also be partly explained by the high concentration of top-tier medical resources in these regions, which may help sustain public awareness and attention to lung cancer. In addition, the low-low clustering observed in southwestern China, particularly in Guangxi, may indicate geographically concentrated areas of relatively limited engagement with lung cancer-related information, potentially reflecting regional disparities in health information access, awareness, or digital health literacy. In these regions, public health authorities may consider prioritizing targeted offline awareness campaigns for lung cancer screening, alongside strengthening digital health literacy programs to improve access to and engagement with relevant health information. The factors correlated with web-based attention to lung cancer offer further insight into the interpretation. The association of higher GDP per capita and average years of education per capita suggests that provinces with greater socioeconomic resources tended to exhibit higher levels of lung cancer-related web activity. At the macro level, this pattern may be associated with broader regional differences in digital infrastructure and access to health-related information [47,48]. Conversely, the negative association with urbanization rate, while unexpected, may reflect greater diversification of health information channels and a lower aggregate level of curiosity-driven search activity in more highly urbanized regions [26]. This interpretation should be treated with caution, as the VIF of the urbanization rate was 6.85, indicating moderate multicollinearity, and the negative coefficient may be partly affected by statistical collinearity. It is also worth noting that these associations should be interpreted at the regional level and do not represent individual information-seeking intentions. Moreover, the spatial panel econometric model confirmed both direct and spillover effects. For example, provinces with higher education levels were associated with higher local PBI and showed positive spatial associations with PBI in neighboring regions, possibly reflecting knowledge diffusion and cross-regional interactions. These findings highlight the importance of accounting for socioeconomic variations and spatial interdependence when interpreting digital engagement with lung cancer information.

Topic modeling and thematic analysis revealed micro-level discourse on lung cancer among Chinese Weibo users. Since late 2023, posts related to health care services have grown rapidly, surpassing personal experience by mid-2025. This transition reflects a shift from personal narratives to system-level concerns related to health care accessibility, affordability, and quality. At the same time, this increase may also partly reflect the growing presence of institutional and commercial actors in the web-based health information space. Other persistent themes included prevention, treatment, screening, and diagnosis, indicating a sustained public attention to early detection and clinical management. Besides, the sustained presence of TCM underscores the culturally embedded nature of discourse in Chinese digital health contexts [42]. Temporal trends in theme prevalence also aligned with policy shifts and media events. For example, the outbreak of COVID-19 in December 2019 may have heightened public awareness of respiratory diseases, triggering a reversal in the trajectory of personal-experience posts from decline to rise [42,49,50]. Meanwhile, the recent decline in content related to screening and diagnosis suggests a potential misalignment between public discourse and long-term health priorities. These findings emphasize that public health communication should integrate patient-centered needs with systemic issues and actively promote preventive strategies.

Sentiment analysis provided an additional perspective on web-based discourse about lung cancer. Overall, sentiment remained positive between 2010 and 2025, with an upward trend despite fluctuations, indicating a generally optimistic and supportive tone. Emotional expression was concentrated within a narrow spectrum, with "Good" representing the primary positive category and "Disgust" the principal negative one. Both increased in intensity over time, illustrating an emotional landscape in which respect, praise, trust, and affection coexisted

XSL•FO
RenderX

with irritation, detestation, criticism, jealousy, and suspicion [38]. Subgroup differences highlighted how sex/organization, communicative roles, and thematic focus shape sentiment in web-based discourse. Higher scores among institutional accounts and media outlets suggested deliberate efforts to frame discourse positively, while lower scores from females may reflect both social vulnerability and a tendency to express distress more directly. Similarly, positive sentiment in health care service posts and negative content on comorbidities and symptoms underscored the dual role of digital platforms as spaces for health communication and outlets for distress.

Limitations

Several limitations should be acknowledged in this study. First, data accessible via the BI platform is constrained to predefined search terms. Consequently, infrequently used or rare expressions may be excluded from trend analyses. Second, the Weibo user base is not representative of the broader Chinese population. Its users tend to be younger, reside in urban areas, and possess higher levels of digital literacy, potentially introducing bias into the findings. Therefore, given that respiratory tract cancers disproportionately affect older males, the observed patterns should be interpreted as reflecting the perceptions and expressions of digitally active users rather than those of the primary high-risk population directly affected by the disease. Moreover, this study adopts an observational and descriptive design. Although associations between explanatory variables and PBI were observed, causal inferences cannot be made. The lack of reliable provincial data on internet users prevented adjustment for internet penetration, potentially introducing regional denominator bias. Unmeasured confounders, such as lung cancer incidence, air pollution, and local public policy, may also influence the patterns of web-based discourse. In addition, although COVID-19 may have influenced search behavior and spatial spillover patterns, pandemic-specific controls were not included due to data limitations, and future research with more granular indicators is warranted. Furthermore, lexicon-based sentiment analysis methods exhibit inherent limitations. Specifically, challenges such as polysemy, context-dependent sarcasm, and culturally specific meanings may undermine the reliability of emotion classification. Finally, this study focused exclusively on public perceptions of lung cancer within Mainland China. To improve cross-cultural applicability and spatial resolution of future research, international comparative studies and more fine-grained spatial analyses are warranted.

Conclusions

This study provides a comprehensive examination of public perceptions of lung cancer in Mainland China by integrating search engine and social media data. The findings showed that public attention peaked in 2019 and declined thereafter, while Weibo discussions remained stable since 2022. Spatial heterogeneity and low-low clusters were observed, such as in Guangxi. PBI was positively associated with GDP per capita and average educational attainment but negatively associated with the urbanization rate. Topic modeling revealed a thematic shift from personal experience to health care service by 2025, and sentiment analysis indicated an overall positive tone driven by “Good” and “Disgust.” These findings contribute to understanding lung cancer discourse in China’s digital landscape and provide evidence for improving targeted health communication.

Acknowledgments

There was no use of generative artificial intelligence (AI) technology in the generation of text, figures, or other informational content of this manuscript.

Funding

This work was supported by the 1·3·5 project for disciplines of excellence from West China Hospital of Sichuan University (ZYGD23010 to LXL).

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

ZL contributed to conceptualization, data curation, formal analysis, investigation, methodology, provision of resources, software development, validation, visualization, and writing of both the original draft and the review and editing of the manuscript. JZ contributed to data curation, validation, and writing of the original draft, as well as review and editing of the manuscript. YF contributed to the visualization, investigation, and writing of the original draft, as well as review and editing of the manuscript. NC contributed to the provision of resources, supervision, and writing of the original draft, as well as review and editing of the manuscript. LL contributed to conceptualization, funding acquisition, methodology, project administration, supervision, and writing of both the original draft and the review and editing of the manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Supplementary methods, figures, and tables related to the study.
[DOCX File, 12072 KB - jmir_v28i1e85058_app1.docx]

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74(3):229-263. [doi: 10.3322/caac.21834] [Medline: 38572751]
2. Zhang Y, Vaccarella S, Morgan E, et al. Global variations in lung cancer incidence by histological subtype in 2020: a population-based study. Lancet Oncol 2023 Nov;24(11):1206-1218. [doi: 10.1016/S1470-2045(23)00444-8] [Medline: 37837979]
3. Kuang Z, Wang J, Liu K, et al. Global, regional, and national burden of tracheal, bronchus, and lung cancer and its risk factors from 1990 to 2021: findings from the global burden of disease study 2021. eClinicalMedicine 2024 Sep;75:102804. [doi: 10.1016/j.eclinm.2024.102804]
4. Diao X, Guo C, Jin Y, et al. Cancer situation in China: an analysis based on the global epidemiological data released in 2024. Cancer Commun 2025 Feb;45(2):178-197. [doi: 10.1002/cac2.12627]
5. Li Z, Yu C, Hao J, et al. Global trends of early, middle, and late‐onset lung cancer from 1990 to 2021: results from the global burden of disease study 2021. Cancer Med 2025 Feb;14(3). [doi: 10.1002/cam4.70639]
6. Chen S, Cao Z, Prettner K, et al. Estimates and projections of the global economic cost of 29 cancers in 204 countries and territories from 2020 to 2050. JAMA Oncol 2023 Apr 1;9(4):465. [doi: 10.1001/jamaoncol.2022.7826]
7. Dai X, Gakidou E, Lopez AD. Evolution of the global smoking epidemic over the past half century: strengthening the evidence base for policy action. Tob Control 2022 Mar;31(2):129-137. [doi: 10.1136/tobaccocontrol-2021-056535] [Medline: 35241576]
8. Cancer stat facts: lung and bronchus cancer. National Cancer Institution. URL: https://seer.cancer.gov/statfacts/html/lungb.html [accessed 2025-09-18]
9. Kovalchik SA, Tammemagi M, Berg CD, et al. Targeting of low-dose CT screening according to the risk of lung-cancer death. N Engl J Med 2013 Jul 18;369(3):245-254. [doi: 10.1056/NEJMoa1301851] [Medline: 23863051]
10. de Koning HJ, van der Aalst CM, de Jong PA, et al. Reduced lung-cancer mortality with volume CT screening in a randomized trial. N Engl J Med 2020 Feb 6;382(6):503-513. [doi: 10.1056/NEJMoa1911793] [Medline: 31995683]
11. Simon AE, Juszczyk D, Smyth N, et al. Knowledge of lung cancer symptoms and risk factors in the UK: development of a measure and results from a population-based survey. Thorax 2012 May;67(5):426-432. [doi: 10.1136/thoraxjnl-2011-200898]
12. Alamri SA, Alzahrani MM, Alamri AA, et al. Understanding the public knowledge, attitude, and practice toward screening and risk factors of lung cancer in Saudi Arabia: a cross-sectional study. Ann Thorac Med 2024;19(4):275-283. [doi: 10.4103/atm.atm_111_24] [Medline: 39544349]
13. Chai Q, Li R, Bao T, Yang Z, Liu Q, Chen F. Awareness of lung cancer among urban residents in Sichuan Province and its impact on their willingness to choose medical institutions for cancer screening. Front Public Health 2024;12:1388140. [doi: 10.3389/fpubh.2024.1388140] [Medline: 39850860]
14. Rigney M, Rapsomaniki E, Carter-Harris L, King JC. A 10-year cross-sectional analysis of public, oncologist, and patient attitudes about lung cancer and associated stigma. J Thorac Oncol 2021 Jan;16(1):151-155. [doi: 10.1016/j.jtho.2020.09.011]
15. Peng T, Hamann HA, David EA. Stigma may exacerbate disproportionately low guideline-concordant treatment rates for patients with advanced-stage lung cancer in the United States. JTO Clin Res Rep 2022 Apr;3(4):100302. [doi: 10.1016/j.jtocrr.2022.100302] [Medline: 35400082]
16. Luo Z, Guo W, Zhou H, Chen Z. Knowledge, attitude, and practice toward lung cancer risk among offspring of lung cancer patients: a cross-sectional study. Sci Rep 2024 Aug 3;14(1). [doi: 10.1038/s41598-024-69142-8]
17. Chen M, Hui S, Huang Y, et al. Knowledge, attitudes, and current practices toward lung cancer palliative care management in China: a national survey. Front Oncol 2024;14. [doi: 10.3389/fonc.2024.1382496]
18. Rong F, Shi R, Hu L, Wang D, Lv X, Zhao Y. Impact of risk perception and disease cognition on the willingness to participate in screening for lung cancer in a high-risk population. Eur J Cancer Prev 2024;33(2):141-151. [doi: 10.1097/CEJ.0000000000000843]
19. Aiello AE, Renson A, Zivich PN. Social media- and internet-based disease surveillance for public health. Annu Rev Public Health 2020 Apr 2;41:101-118. [doi: 10.1146/annurev-publhealth-040119-094402] [Medline: 31905322]
20. The 56th statistical report on China's internet development. China Internet Network Information Center (CNNIC). 2025. URL: https://www.cnnic.net.cn/NMediaFile/2025/0730/MAIN1753846666507QEK67ZS9DH.pdf [accessed 2025-08-15]
21. 2019 research report on Chinese internet users' search engine usage. China Internet Network Information Center. 2019. URL: https://www.cnnic.cn/n4/2022/0401/c122-1110.html [accessed 2025-08-27]
22. Xiao H, Wang Z, Liu F, Unger JM. Excess all-cause mortality in China after ending the zero COVID policy. JAMA Netw Open 2023 Aug 1;6(8):e2330877. [doi: 10.1001/jamanetworkopen.2023.30877]

XSL•FO
RenderX

23. Xiong X, Xiang L, Chang L, Wu IX, Deng S. Forecasting the incidence of mumps based on the Baidu index and environmental data in Yunnan, China: deep learning model study. J Med Internet Res 2025 Feb 6;27:e66072. [doi: 10.2196/66072] [Medline: 39913179]
24. Sina Weibo Data Center. Weibo user consumption trends report. Sina Weibo. 2024. URL: https://data.weibo.com/report/reportDetail?id=473 [accessed 2025-09-19]
25. Cheng Y, Yu Z, Xu C, et al. Climatic and economic background determine the disparities in urbanites' expressed happiness during the summer heat. Environ Sci Technol 2023 Aug 1;57(30):10951-10961. [doi: 10.1021/acs.est.3c01765]
26. Zhang Z, Xu H, Pan J, Song F, Chen T. Spatiotemporal characteristics and influential factors of electronic cigarette web-based attention in mainland China: time series observational study. J Med Internet Res 2025 Feb;27:e66446. [doi: 10.2196/66446]
27. Zheng S, Wang J, Sun C, Zhang X, Kahn ME. Air pollution lowers Chinese urbanites' expressed happiness on social media. Nat Hum Behav 2019 Mar;3(3):237-243. [doi: 10.1038/s41562-018-0521-2] [Medline: 30953012]
28. Jin B, Wu Y, Rao CR, Hou L. Estimation and model selection in general spatial dynamic panel data models. Proc Natl Acad Sci USA 2020 Mar 10;117(10):5235-5241. [doi: 10.1073/pnas.1917411117]
29. Kong D, Chen A, Zhang J, et al. Public discourse and sentiment toward dementia on Chinese social media: machine learning analysis of Weibo posts. J Med Internet Res 2022 Sep 2;24(9):e39805. [doi: 10.2196/39805]
30. Chen X, Yik M. The emotional anatomy of the Wuhan lockdown: sentiment analysis using Weibo data. JMIR Form Res 2022 Nov 14;6(11):e37698. [doi: 10.2196/37698] [Medline: 36166650]
31. Guan Q, Deng S, Wang H. Chinese stopwords for text clustering: a comparative study. Data Anal Knowl Discov 2017;1(3):72-80.
32. Zeng Y, Zhou Y, Li K. Inter-provincial differences and trends in average years of education per capita. China Econ Educ Rev 2024;9(1):104-121. [doi: 10.19512/j.cnki.issn2096-2088.2024.01.002]
33. Liu C, Nie G. Spatial effects and impact factors of food nitrogen footprint in China based on spatial durbin panel model. Environ Res 2022 Mar;204(Pt B):112046. [doi: 10.1016/j.envres.2021.112046]
34. Curts HA. Adjacency table method of deriving minimal sums. IEEE Trans Comput 1977 Nov;C-26(11):1136-1141. [doi: 10.1109/TC.1977.1674760]
35. Fischer MM, Getis A. Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications: Springer Science & Business Media; 2009.
36. Da C, Duan Y, Ji Z, et al. Assessing the needs of patients with breast cancer and their families across various treatment phases using a latent dirichlet allocation model: a text-mining approach to online health communities. Support Care Cancer 2024 May;32(5). [doi: 10.1007/s00520-024-08513-3]
37. Topaloğlu D, Polat O. Trends and methods in intensive care unit (ICU) research using machine learning: latent dirichlet allocation (LDA)-based thematic literature review. BMC Med Inform Decis Mak 2025 Jul 29;25(1):282. [doi: 10.1186/s12911-025-03132-2] [Medline: 40731277]
38. Xu L, Lin H, Pan Y, Ren H, Chen J. Constructing the affective lexicon ontology. J China Soc Sci Tech Inf 2008 May 5;27(2):180-185. [doi: 10.3969/j.issn.1000-0135.2008.02.004]
39. Conway M, O'Connor D. Social media, big data, and mental health: current advances and ethical implications. Curr Opin Psychol 2016 Jun;9:77-82. [doi: 10.1016/j.copsyc.2016.01.004] [Medline: 27042689]
40. Sutton J, Vos SC, Olson MK, et al. Lung cancer messages on Twitter: content analysis and evaluation. J Am Coll Radiol 2018 Jan;15(1):210-217. [doi: 10.1016/j.jacr.2017.09.043]
41. Taylor J, Pagliari C. The social dynamics of lung cancer talk on Twitter, Facebook and Macmillan.org.uk. NPJ Digit Med 2019;2(51):51 [FREE Full text] [doi: 10.1038/s41746-019-0124-yB1] [Medline: 31304397]
42. Gao H, Guo D, Wu J, Li L. Weibo users' emotion and sentiment orientation in traditional Chinese medicine (TCM) during the COVID-19 pandemic. Disaster Med Public Health Prep 2022 Oct;16(5):1835-1838. [doi: 10.1017/dmp.2021.259]
43. Terry K, Yang F, Yao Q, Liu C. The role of social media in public health crises caused by infectious disease: a scoping review. BMJ Glob Health 2023 Dec 28;8(12):e013515. [doi: 10.1136/bmjgh-2023-013515] [Medline: 38154810]
44. Zayed BA, Talaia AM, Gaaboobah MA, Amer SM, Mansour FR. Google trends as a predictive tool in the era of COVID-19: a scoping review. Postgrad Med J 2023 Aug 22;99(1175):962-975. [doi: 10.1093/postmj/qgad012] [Medline: 36892422]
45. Liu X, Yang Q, Pan L, et al. Burden of respiratory tract cancers in China and its provinces, 1990-2021: a systematic analysis of the global burden of disease study 2021. Lancet Reg Health West Pac 2025 Feb;55:101485. [doi: 10.1016/j.lanwpc.2025.101485] [Medline: 39968450]
46. Zhao J, Chen W, Liu Z, et al. Urban expansion, economic development, and carbon emissions: trends, patterns, and decoupling in mainland China's provincial capitals (1985–2020). Ecol Indic 2024 Dec;169:112777. [doi: 10.1016/j.ecolind.2024.112777]
47. Korn AR, Walsh-Bailey C, Correa-Mendez M, et al. Social determinants of health and US cancer screening interventions: a systematic review. CA Cancer J Clin 2023;73(5):461-479. [doi: 10.3322/caac.21801] [Medline: 37329257]
48. Stormacq C, Van den Broucke S, Wosinski J. Does health literacy mediate the relationship between socioeconomic status and health disparities? integrative review. Health Promot Int 2019 Oct 1;34(5):e1-e17. [doi: 10.1093/heapro/day062] [Medline: 30107564]

XSL•FO
RenderX

49. Wang J, Zhou Y, Zhang W, Evans R, Zhu C. Concerns expressed by Chinese social media users during the COVID-19 pandemic: content analysis of Sina Weibo microblogging data. J Med Internet Res 2020 Nov 26;22(11):e22152. [doi: 10.2196/22152]
50. Mazzone PJ, Gould MK, Arenberg DA, et al. Management of lung nodules and lung cancer screening during the COVID-19 pandemic. Chest 2020 Jul;158(1):406-415. [doi: 10.1016/j.chest.2020.04.020]

Abbreviations

BI: Baidu Index
EAPC: estimated annual percentage change
GDP: gross domestic product
LDA: latent Dirichlet allocation
PBI: per capita Baidu Index
SPDM: spatial panel Durbin model
TCM: traditional Chinese medicine
VIF: variance inflation factor

Edited by A Coristine; submitted 30.Sep.2025; peer-reviewed by R Guo, U Pasupulety; revised version received 07.Apr.2026; accepted 07.Apr.2026; published 05.May.2026.

Please cite as:
Li Z, Zhang J, Fan Y, Chen N, Liu L
Understanding Public Perceptions of Lung Cancer in China: Infodemiology Study of Baidu Index and Weibo Posts
J Med Internet Res 2026;28:e85058
URL: https://www.jmir.org/2026/1/e85058
doi:10.2196/85058

© Zongyuan Li, Jian Zhang, Yiming Fan, Nan Chen, Lunxu Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Why People Conceal Mental Health Problems: Qualitative Analysis of TikTok Posts

Chloe Roske[1], BA; Kael Ragnini[1], BA; Qinchun Zhu[1], EdM; Ashari Palmer[1]; Meredith R Kells[2], RN, PhD; Heather A Davis[3], PhD; Matthew K Nock[1,4,5], PhD

[1]Department of Psychology, Harvard University, 33 Kirkland Ave, Cambridge, MA, United States
[2]School of Nursing, University of Rochester, Rochester, NY, United States
[3]Department of Psychology, Virginia Tech, Blacksburg, VA, United States
[4]Department of Psychiatry, Massachusetts General Hospital, Boston, MA, United States
[5]Department of Psychiatry, Boston Children's Hospital, Boston, MA, United States

Corresponding Author:
Chloe Roske, BA
Department of Psychology, Harvard University, 33 Kirkland Ave, Cambridge, MA, United States

Abstract

Background: Concealment of psychiatric symptoms is a barrier to effective mental health treatment, particularly among patients with suicidal thoughts and behaviors. Prior research on concealment has relied on retrospective self-report or laboratory-based interviews, which may not capture real-world decision-making about disclosure. Social media platforms such as TikTok provide a context in which individuals publicly narrate their experiences about concealing psychiatric symptoms, offering insight into motivations for concealment uninfluenced by experimenter demand characteristics.

Objective: To understand patient decision-making about when to conceal and when to disclose psychiatric symptoms, this study examined social media content about patient experiences of concealing mental health symptoms. TikTok was chosen because it is the fastest-growing social media platform, and social media platforms provide an open-ended format for people to express their thoughts and feelings on various topics.

Methods: Using a newly created TikTok account to minimize algorithmic bias, we identified and downloaded the 25 most-viewed English-language videos from 4 search terms about concealment in clinical contexts ("lying to therapist," "lying to my therapist," "lying to doctor about mental health," and "lying to doctors about mental health"). After exclusions, 98 videos were included in the analysis. Videos were analyzed using reflexive thematic analysis. Four coders collaboratively developed a codebook through iterative review, triangulation, and consensus discussions. Engagement metrics (views, likes, comments, shares, saves) were recorded and summarized.

Results: The 98 videos had 73,252,531 views, 14,356,874 likes, 74,954 comments, 770,027 shares, and 1,204,006 saves. Four themes were constructed among the 90 videos that explicitly discussed motivations for concealment: (1) disclosure perceived as punitive (31/90, 34.4% of videos), including desire to avoid hospitalization (17/90, 18.8%); (2) managing others' feelings and impressions (28/90, 31.1%), including fear of upsetting therapists (5/90, 5.5%) and maintaining a façade of wellness (7/90, 7.7%); (3) negative emotions or inability to identify feelings (21/90, 23.3%), including fear of vulnerability (6/90, 6.6%); and (4) negative opinions of psychiatric treatment (17/90, 18.8%), including concerns about confidentiality (3/90, 3.3%). An exploratory theme captured ambivalence and guilt surrounding nondisclosure.

Conclusions: Results provide insight into patient motivations for concealing their suicidal thoughts and behaviors and offer potential avenues for improving rates of disclosure, which is critical to reducing death by suicide. TikTok creators frequently described concealment as a strategy to avoid perceived punitive consequences, manage interpersonal dynamics, or cope with emotional distress. Findings suggest that current risk management practices and stigma surrounding psychiatric care may unintentionally reinforce concealment behaviors. These insights may inform interventions aimed at improving the therapeutic alliance, enhancing transparency around hospitalization criteria, and reducing barriers to honest reporting of suicide risk.

(J Med Internet Res 2026;28:e88244) doi:10.2196/88244

KEYWORDS

concealment; disclosure; suicide; social media; TikTok

XSL•FO
RenderX

Introduction

Approximately 50% of people in the United States are diagnosed with a psychiatric disorder during their lifetime, yet few seek mental health treatment [1,2]. Of those who do, many conceal the extent and existence of symptoms from clinicians, which limits the delivery of adequate treatment [3-5]. Concealment has been associated with worse therapeutic outcomes [6] and may be most consequential for patients with suicidal thoughts and behaviors (STBs) [7]. If a clinician is unaware of a patient's suicidal thoughts or intent, they may not provide evidence-based treatments for STBs [8] and ultimately be unable to intervene to prevent death by suicide. Indeed, over half of suicide decedents denied suicidal ideation (SI) before their deaths [9,10]. This study sought to better understand when and why patients choose to conceal their psychiatric symptoms from clinicians by conducting a review of social media content.

Around 90% of people conceal significant information from their clinicians, with that information most frequently being about STBs [5]. A study of psychiatric inpatients showed that 51.5% withheld some information about SI during admission [11]. Another study reported that fewer than half of people with SI reported it to a health professional. Of those who did disclose, 25% reported SI to a psychologist, 13% to a psychiatrist, 12% to a counselor, and 2.6% to a social worker [12]. Alarmingly, half of the people who died by suicide denied SI in the month before their deaths [10] and two-thirds denied SI when last asked [9]. Research has shown that patients may choose to conceal or minimize the severity of their STBs for myriad reasons, including shame, embarrassment, the quality of the therapeutic alliance, and feeling misunderstood. People also endorse being afraid of stigmatizing responses from health care professionals, involuntary hospitalization, and being encouraged to take psychiatric medication [4,7,13-18].

Prior work on concealment in clinical contexts may be limited by threats to internal and external validity. To date, research on concealment of STBs has relied on self-report and interview-based methods conducted in research settings. However, people who do not disclose STBs to their treatment team may experience similar deterrents to disclosure in a research context. Thus, studies that included qualitative interviews in a laboratory setting [4,13,16,17] or self-report questionnaires [12,13,19,20] may not wholly measure the reasons why patients conceal information from their clinicians. Indeed, people are prone to engage in self-presentation processes, where people manage the impressions they make on others based on context [21], and demand characteristics are a known limitation in research [22]. As such, existing research may not fully capture reasons for concealment given the research context. More observational work is needed to deepen our insight into why patients choose to conceal or minimize their mental health symptoms. A better understanding of this phenomenon may help improve the identification of people at risk for suicide and other related mental health problems. To our knowledge, no prior study has examined motivations for concealment of psychiatric symptoms using real-world, unsolicited disclosure outside of clinical or laboratory settings.

Social media may offer more naturalistic insight into reasons why patients conceal information from their treatment teams. People often turn to social media when seeking peer support for psychiatric and nonpsychiatric illnesses and engaging in discourse about mental health [23-27]. TikTok, a platform for sharing short-form videos, has recently begun to dominate the social media landscape; it has over 1 billion users per month who spend an average of 10 hours per week on the app [28]. As a result, TikTok hosts a vast collection of new, user-generated content on subjects ranging from eating disorder treatment to discussion of mental health symptoms. This content is unsolicited and voluntarily provided outside the presence of researchers or clinicians. Unlike structured interviews or surveys, these disclosures are not cloaked in institutional authority and may be more ecologically valid. In other words, social media content allows researchers and clinicians to learn what patients discuss outside of the research setting or therapy room. Importantly, mental health-related content posted online can be reliably correlated with the poster's real-world behavior. For example, predictive language models have been able to identify language used in Facebook posts as a significant predictor of depression diagnoses in corresponding samples' medical records [29].

TikTok's specific features and characteristics may lower disclosure barriers. For one, filters, green screen effects, and pseudonymous usernames facilitate some anonymity. Even without these tools, some work suggests that web-based environments can make people feel psychologically distanced from their real-world identities, even if they are not literally anonymous [30]. Perceived anonymity or distance is important because previous work shows that mental health symptom endorsement is higher when surveys are anonymous [31] and social media self-disclosure is higher with anonymity [32]. Research on context-dependent privacy outside of video-based social media further suggests that willingness to disclose sensitive information is shaped not only by identifiability or audience, but also by the perceived tone and seriousness of the environment. Individuals are significantly more likely to disclose stigmatized or illicit behaviors when privacy concerns are cognitively suppressed by playful or informal contextual cues, even when objective privacy risks are constant [33]. TikTok's entertainment-oriented context may similarly reduce the salience of privacy and consequence, enabling users to share mental health experiences that might feel too risky or inappropriate to relay in research or clinical settings. Importantly, however, disclosure and self-presentation are 2 separate processes. People may disclose but curate how and when they do so.

Indeed, disclosure in online settings is not free from impression management. Certain features of online communication (eg, asynchronicity, visible audience feedback) may, in fact, intensify it [34]. Because TikTok is a video-based platform, motivation to engage in self-presentation is likely activated. However, the value of examining TikTok content does not lie in the assumption that users are completely anonymous or unconcerned with how they are perceived. Rather, it lies in observing how individuals choose to frame and narrate their mental health experiences in a context that differs from clinical encounters. The imagined audience on TikTok is diffuse, peer-oriented, and

XSL•FO
RenderX

governed by platform-specific norms rather than institutional authority, diagnostic implications, and material consequences. Consistent with Goffman's theory of self-presentation, TikTok disclosures can be understood as curated performances oriented toward this imagined audience [35,36]. As a result, while impression management remains salient, the form it takes likely differs meaningfully from that of clinical disclosure, potentially shaping what individuals choose to disclose, emphasize, or withhold. TikTok, therefore, offers a novel perspective on how individuals frame, explain, and justify decisions to disclose or conceal psychiatric symptoms and, therefore, addresses key limitations in prior interview-based research.

Because suicide risk assessment relies heavily on patient self-report, concealment of suicidal thoughts may directly impede identification and intervention. Therefore, understanding these motivations is critical not only for improving therapeutic alliance but also for strengthening suicide risk detection and prevention efforts. To gain more organic insight into patient motivations for concealing mental health symptoms, this study describes TikTok content about concealing mental health symptoms from clinicians using reflexive thematic analysis. Given previous work [4,16,18], the themes constructed will likely surround psychiatric care, particularly involuntary hospitalization, stigma, and feeling isolated.

Methods

Ethical Considerations

The Harvard University Institutional Review Board reviewed this study and determined that it does not constitute human subjects research under US Department of Health and Human Services regulations and US Food and Drug Administration regulations (Protocol #IRB24-1553). This determination was based on social media posts being considered public domain and no interaction with content creators. Accordingly, no informed consent was required. Identifying information about content creators is not included in this manuscript to protect their privacy, and only publicly posted videos are included in this analysis. This approach aligns with previous work on samples of short-form social media videos [37-40]. Videos were assigned numeric identifiers to protect user privacy and treated as unique participants.

Sampling Strategy

Consistent with previous qualitative analyses of TikTok content [24,37,38], we downloaded 100 of the most viewed videos for analysis with search terms we chose a priori. Unlike previous studies [23,37,41], our topic of interest did not neatly align with viral hashtags and search terms (eg, #mentalhealth, #WhatIEatInADay, #dietpills). We explored potential search terms using a new TikTok account to avoid algorithms that tailor content to individuals' watch history, which is in line with prior work [24,37,38]. A list of potential search terms was created; terms were added if the first 10 videos that appeared broadly addressed motivations for symptom concealment. We observed that the online discourse about concealment did not seem to unify around a single search term. To mirror this, we chose to use 4 terms that had the most relevant content and chose the 25 most viewed videos from each (the TikTok search feature displays videos in order of popularity). This ensured that we were analyzing videos that had the widest reach and may have resonated with the largest audience. Given that the goal of this study was to understand why patients would not disclose psychiatric symptoms to clinicians, we included the 25 most popular videos from the following search terms: "lying to therapist," "lying to my therapist," "lying to doctor about mental health," and "lying to doctors about mental health." These search terms were chosen because the videos watched in the exploratory phase most consistently addressed the research question. After finalization of search terms, a new TikTok account was created. On January 11, 2025, author CR downloaded the videos, excluding videos that were not in English or made by a creator who self-identified as a clinician. Two videos were excluded because the creator was a clinician. After exclusion, we had a sample of 98 videos and recorded the number of views, likes, comments, saves, and shares. This was done via TikTok data scraping tool Apify.com as well as manual confirmation. No transcription software was used.

Data Coding and Thematic Analysis

A reflexive thematic analysis approach, following Braun and Clarke's framework [42,43], was used to collaboratively and inductively construct codes and themes from the data. Data coding procedures are described in Table 1. This approach included researcher triangulation, a method to resolve discrepancies in coding by checking data created by one team member against the others [44]. The codebook was created using a subset of the data, and once it was finalized, the team viewed the final videos at least 3 times and coded them. Themes and subthemes were generated collaboratively after 2 final viewings of the entire sample and a collective agreement that the generated themes robustly described the sample's content. Quantitative statistical analyses were conducted in R (R Foundation for Statistical Computing) [45], including descriptive statistics about the views, likes, comments, saves, and shares of the videos and the number of videos with each code. Data analysis procedures followed previous work examining TikTok content [24,37,38]. This study was conducted and reported in accordance with the Standards for Reporting Qualitative Research guidelines [46].

XSL•FO
RenderX

Table . Procedure for reflexive thematic analysis of TikTok videos.

Phase	Description	Authors
Phase I: familiarization with the data	Three coders viewed all the videos twice.	CR, AP, QZ
Phase II: generation of preliminary codes	The same coders rewatched a randomly selected 15% of the sample (n=15) and generated preliminary codes for those videos. The group met for debriefing, which included researcher triangulation, and reflecting on and resolving discrepancies. Preliminary codes for each video were then agreed on. This process was done again with another 15 randomly selected videos. The first author used the agreed on preliminary codes to generate an initial codebook.	CR, AP, QZ
Phase III: revision of codes	The 3 coders watched the first 50 videos 3 times each and independently coded the videos. The coders met to cross-reference each other's codes and resolve discrepancies with a consensus. New codes were added if the authors thought a video was not being sufficiently captured by the existing codes.	CR, AP, QZ
Phase IV: coding all videos	The codebook was used to code the remaining videos. The coders met to conduct researcher triangulation and resolve discrepancies, involving an additional coder (KR) for added perspective.	CR, AP, QZ, KR
Phase V: creating themes	The 4 coders rewatched every video in the sample twice, reflected on the codebook, and met to discuss larger patterns in the data. Four overarching themes were constructed with 4 - 7 subthemes each.	CR, AP, QZ, KR

This study was guided by a constructivist lens, which assumes that people construct meaning through their experiences and social contexts. Thus, videos were not considered to reflect a single objective reality. We interpreted videos as personal, context-dependent narratives shaped by creators' identities, audiences, experiences, and social lives. The research team included college undergraduates, doctoral trainees, and faculty with varied levels of clinical and research experience. The team was diverse with respect to race, ethnicity, age, and sexual orientation. Doctoral trainees and faculty acknowledged that their clinical training and prior work in suicide risk assessment could influence the interpretation of concealment narratives. Undergraduate coders were therefore explicitly encouraged to offer their interpretations and to challenge senior team members when discrepancies arose. The inclusion of college-aged coders was particularly valuable given that many content creators appeared to be adolescents or young adults. Their perspectives helped contextualize language, humor, and platform norms. Reflexive discussions were integrated throughout the coding process to critically examine assumptions, address potential bias, and promote analytic rigor.

In Phase V (developing themes), the research team carefully examined videos (n=12) that had been assigned codes about feelings surrounding concealment of psychiatric symptoms. In doing so, they observed that a subset of coded videos (n=8) discussed concealing mental health symptoms but did not elaborate on the reasons for doing so. The team agreed that these videos provided interesting insight into the phenomenon of concealment, although they did not clearly answer the research question (what are the motivations for concealment?). To balance the clinically relevant information obtained from the thematic analysis of these videos while answering our research question, the theme "Feelings about Disclosure" is discussed as an exploratory result. The 8 videos that did not discuss motivations are removed from calculations in Table 2.

XSL•FO
RenderX

Table . Themes and subthemes constructed for videos about the concealment of psychiatric symptoms and their frequency. This table represents the 90 videos that discussed reasons for concealment. The number of videos reported in each theme represents the number of unique videos per theme (ie, videos with more than one subtheme were not double-counted).

Themes and subthemes	Videos, n (%)
Disclosure as punitive	31 (34.4)
Hospitalization after honest disclosure	9 (10)
Fear of hospitalization	4 (4.4)
Desire to avoid hospitalization	17 (18.8)
Desire to leave inpatient hospitalization	2 (2.2)
Managing others' feelings and impressions	28 (31.1)
Fear of upsetting therapist	5 (5.5)
Fear of upsetting/disappointing parents	3 (3.3)
Parents knowing about symptoms	8 (8.8)
Maintaining façade	7 (7.7)
False therapeutic improvement leading to perceived progress	9 (10)
Feeling bad and feeling nothing	21 (23.3)
Feeling alone/silent suffering	7 (7.7)
Perceived burdensomeness	4 (4.4)
Hopelessness	2 (2.2)
Patient fear that they are faking illness	3 (3.3)
Fear of being vulnerable in therapy	6 (6.6)
Minimizing symptoms to self	2 (2.2)
Inability to access emotions	2 (2.2)
Negative opinions of psychiatric treatment	17 (18.8)
Perceived lack of patient-therapist confidentiality	3 (3.3)
Treatment does not work	2 (2.2)
Hospital as restrictive	3 (3.3)
Issues with psychiatric treatment structure and modality	8 (8.8)
Weak therapeutic alliance	7 (7.7)

Results

Overview

Four predominant themes regarding reasons for concealing psychiatric symptoms were created: (1) Disclosure as Punitive, (2) Managing Others' Feelings and Impressions, (3) Feeling Bad and Feeling Nothing, and (4) Negative Opinions of Psychiatric Treatment. Each theme had multiple subthemes that often overlapped, as described below, and most videos contained more than one theme. Table 2 shows the themes, subthemes, and percentages of videos with each theme and subtheme, and Figure 1 visualizes the themes. Table 3 summarizes the distribution of the number of codes assigned per video.

Figure 1. Key themes identified in TikTok videos about concealing psychiatric symptoms.

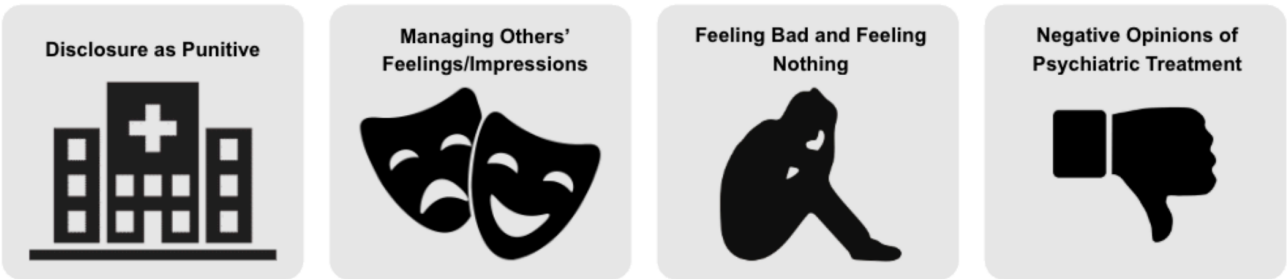

XSL•FO
RenderX

Table . Distribution of codes assigned to each TikTok video in the sample. This table calculates code assignments for the entire sample of 98 videos.

Number of codes assigned per video	Videos, n (%)
1	41 (42.3)
2	37 (38.1)
3	9 (9.3)
4	1 (1.0)
5	1 (1.0)

Statistical Analysis

At the time of download, the 98 videos that comprised our sample had a combined 73,252,531 views, 14,356,874 likes, 74,954 comments, 770,027 shares, and 1,204,006 saves. Numbers were rounded to the nearest tenth of a thousand when recorded (eg, 13.5 k) to reflect how TikTok displays video metrics. Table 4 provides the average, SD, median, and range across videos. Each video was posted by a unique creator.

Table . Mean, SD, median, and range of TikTok videos about concealment of psychiatric symptoms. These metrics evaluate the entire sample of 98 videos.

Metric	Mean (SD)	Median (range)
Views	755,181 (1,596,059)	94,100 (90 - 7,400,000)
Likes	159,521 (339,147)	10,950 (9 - 1,700,000)
Comments	781 (1773)	64 (0 - 9282)
Shares	8021 (29,168)	188 (1 - 225,100)
Saves	13,682 (31,094)	1554 (1 - 198,100)

Theme 1: Disclosure as Punitive

The majority of videos in our sample reflected an attitude that disclosure has negative consequences for the creator, specifically being hospitalized. Most videos communicated a desire to avoid experiencing an involuntary inpatient hospitalization, with 4 creators explicitly mentioning that concealment of symptoms or the severity of symptoms was due to fear of being hospitalized. Most of these videos followed a similar structure, including the sentiment that the creator lies to their therapist so that they can avoid the hospital. Nine videos recounted being hospitalized as a result of honest disclosure. One video showed a creator in a hospital gown slamming a pillow on the floor with audio of someone crying and screaming. Text on the screen read, “POV: You told your therapist your real thoughts.” Other videos showed scenes from a hospital with captions like “I was too honest with my therapist.” Two videos talked about symptom concealment as a means to leave inpatient hospitalization. The sentiment that the hospital is a restrictive place (eg, “grippy sock jail”) also seemed to shape disclosure considerations.

Theme 2: Managing Others’ Feelings and Impressions

Videos with this theme discussed how other people factored into their decision-making about disclosure. Eleven videos specifically talked about parents and family. Some creators considered other people’s emotional responses to their symptom disclosure when deciding whether to report or conceal their symptoms. Five creators talked about being afraid their therapist would be upset by or unable to handle their emotions. One video is a cartoon depicting a creator’s experience disclosing to their therapist who “cri[ed] after hearing [their] story for the first time.” As a result, the creator says that they felt “guilty immediately” and felt “bad for making them cry and stop[ped] talking completely.” Three creators mentioned wanting to shield their parents from disappointment or distress that could result from their disclosure of symptoms. One said, “My mom sent me to therapy & i lied to the therapist the entire time so she didn’t see me as a disappointment.”

Creators also articulated efforts to curate an image of themselves as “well.” One such effort included lying about symptoms in therapy so that the therapist would think that they were better, which often led to premature treatment termination or a reduction in treatment frequency. One creator said, “I love when I have therapy and I lie to my therapist and they’re like ‘oh you seem to be doing really great’ or ‘making good progress.’ I’m like yeah. That’s because of the lies that you think that.” The creator captioned this video, “was hoping you would see right through me.” Other efforts to appear well included endeavoring to make their life seem fun or putting extra effort into how they dress. Not all creators explicitly discussed their motivations for making themselves out to be less symptomatic than they were, but simply discussed doing so. Creators also spoke about not wanting their parents to know that they were experiencing psychiatric symptoms, and often not providing a rationale. For example, creators mentioned lying about their symptoms because their moms were in the room. Finally, 7 videos included a discussion of projecting and maintaining a façade of wellness to the outside world that conflicted with their internal emotional state. One creator responds to the statement “we didn’t know you were struggling that much.” by lip syncing along to lyrics saying, “I made it all look painless/Man am I the greatest.”

XSL•FO
RenderX

Theme 3: Feeling Bad and Feeling Nothing

When talking about disclosure, creators mentioned experiencing their negative emotions in silence, and feeling alone, unsupported, burdensome, and hopeless. For example, when discussing disclosure to a therapist, one creator said, "Who really admits the truth when all we've know [*sic*] is to fight on our own." Creators also shared efforts to suppress these negative emotions. They expressed fear of being vulnerable in therapy and an impulse to minimize their symptoms within their self-talk. One creator filmed herself next to text reading, "Me staying silent in therapy so I don't start crying." In other videos, creators shared that they experience an inability to access and identify emotions. Two creators explained that they do not disclose because they do not know what they are feeling. In one video, the creator lip-synced to audio that said, "I'm going to kill myself" alongside text that read, "when i couldn't tell my therapist what was wrong because i didn't even know what was wrong." One other code was patients expressing fear that they were faking their psychiatric illnesses, even when they did not believe this to be the case. In one example of such a video, a creator recorded herself laughing alongside text that says, "*when you think you're faking your mental illness but then doctors tell you its serious.*"

Theme 4: Negative Opinions of Psychiatric Treatment

Most videos in this theme included discussion about creators' issues with normative aspects of how psychiatric treatment is structured. One creator mentioned being suspicious of treatment due to its cost, endorsing the notion that therapists only care because they are paid. Another creator mentioned disliking that psychiatrists focus on medication management, while another took issue with the fact that therapy is only once a week. Yet another was frustrated with the psychotherapeutic practice of Socratic questioning. For example, one creator filmed herself praying alongside audio repeating "Don't say it" on a loop with text on the video reading, "Every therapist after we tell them a *story*." The video ended with the text, "And how did that make you FEEL?" Creators also worried about a lack of patient-therapist confidentiality, sometimes due to the ethical mandate that confidentiality will be broken if a patient is an imminent risk to themselves. Combining the subthemes that treatment does not work and therapy lacks true confidentiality, one creator posted a video of herself smiling with the text, "me leaving therapy knowing all i did was lie bc even therapist [*sic*] tell other ppl and telling someone how i feel does nothing." Seven videos mentioned a weak therapeutic alliance as a reason for nondisclosure.

Exploratory Thematic Analysis: Feelings About Disclosure

While the purpose of this study was to examine reasons for concealment of psychiatric symptoms, some creators talked about their feelings about the concealment of psychiatric symptoms. Six creators mentioned being honest in therapy, expressing confusion as to why someone would lie to their therapist. Five videos depicted creators being satisfied about not disclosing, often smirking at the camera, and joking about tricking their therapist. Creators experienced negative feelings due to not disclosing, including guilt due to lying to their therapist and hopelessness about not disclosing. One creator filmed herself looking sad alongside text that said, "That feeling of guilt after blatantly lying to your therapist <<" with the caption, "Some things I'll never be able to say." Another video showed a girl's face morphing from a neutral expression to one with tears and mascara running down her cheeks. The neutral face had the text "Me after always lying to my therapist so she thinks I'm getting better," while the crying face had text reading, "Knowing now that I'll never get my true feelings out." In this video, she is showing the façade she is presenting to her therapist and the resulting hopelessness about getting better. Other creators struggled to identify why they could not be honest with their provider, but confirmed that they were being dishonest. Three creators talked about being ambivalent about whether to disclose. In one video, a creator filmed themselves filling out a self-report questionnaire of psychiatric symptoms and moving their pencil back and forth between "not at all" and "very often" before answering every question with "not at all."

Discussion

Principal Findings

This study sought to understand why patients withhold reporting or minimize psychiatric symptoms in clinical encounters using user-initiated content from TikTok. Through reflexive thematic analysis of 98 TikTok videos, we identified 4 overarching themes describing reasons for withholding or minimizing psychiatric symptoms in clinical encounters: Disclosure as Punitive, Managing Others' Feelings and Impressions, Feeling Bad and Feeling Nothing, and Negative Opinions of Psychiatric Treatment. These themes and their subthemes show that patients have diverse and multifaceted motivations for concealing psychiatric symptoms. The most frequently endorsed reasons for concealment were specific perceived features of psychiatric treatment, including fear of involuntary hospitalization and desire to avoid the hospital. Interpersonal concerns, such as not wanting to upset family members or therapists, or wanting to present as "doing well," were also barriers to disclosure. Some creators struggled to identify and express their feelings, while others endorsed shame, hopelessness, or a fear that their symptoms were not "real enough" to merit care. These findings extend prior research by examining concealment outside laboratory or interview contexts and provide novel insight into decision-making about concealment and disclosure.

Disclosure as Punitive

The most prevalent theme was Disclosure as Punitive. Results indicated that negative perceptions of treatment, such as seeing the hospital as restrictive and issues with psychiatric treatment structure and modality, may relate to desires to minimize symptoms. The most prominent reason for lack of disclosure expressed in our sample was the desire to avoid hospitalization. This observation parallels previous work suggesting that evading psychiatric hospitalization is a key barrier to disclosure of STBs [4,5,13,15,17]. Prior work suggests that one reason for the desire to avoid hospitalization in people admitted for STBs is perceived lack of utility and negative experiences [47]. These sentiments have been echoed in qualitative work describing TikTok content about treatment for eating disorders [24]. Low satisfaction with

XSL•FO
RenderX

inpatient stay has been associated with nondisclosure of STBs in quantitative work as well [11]. Creators also talked about being hospitalized after honest disclosure, often suggesting that they regretted being honest with their clinician due to this outcome. Indeed, some patients perceived hospitalization as a punishment for honest disclosure. To align patient and clinician understanding, it may be helpful for clinicians to address what "imminent risk" looks like for each patient. Since STBs are idiosyncratic [48], having a predetermined agreement about when hospitalization would be necessary before a suicidal crisis is warranted. In one example of doing so, clinicians who experienced nondisclosure with their military-affiliated patients dealt with this challenge by normalizing the occurrence of STBs and collaborating on limiting means [49]. Additionally, clinicians should discuss and uncover what patients' concerns about hospitalization are to help correct misinformation and reduce anxiety about hospitalization. Patient fears may be addressable or unfounded in what occurs on an inpatient unit. These findings suggest that suicide risk management practices perceived as coercive may unintentionally create incentives for concealment, potentially compromising accurate detection of suicide risk.

Managing Others' Feelings and Impressions

Creators voiced desires to manage other people's feelings and impressions. Creators described feeling that their emotions were too intense for others to handle and would provoke distress, which in turn elicited guilt. Not wanting to worry others as a reason for nondisclosure has been reported previously [13,50], and prior work has shown modest effects of perceived burdensomeness in predicting suicide-related outcomes [51,52], suggesting shared risk factors for nondisclosure and suicide. Relationship dynamics may also shape disclosure and explain why many creators discussed wanting to conceal information from parents. For example, families' reduced ability to respond to one another's emotional needs with appropriate affect has been linked to negative psychiatric outcomes [53,54]. Future research should examine how family responses to emotional expression influence patterns of disclosure—whether nondisclosure emerges as a learned behavior in emotionally suppressive families or from experiences of others reacting intensely to one's needs. It may also be informative to investigate whether targeting perceived responsibility for managing others' emotions in treatment could reduce nondisclosure.

Creators also communicated endeavors to be seen by others as psychologically well. These efforts did not seem to be a deception to avoid disclosure but rather a form of impression management. Work in nonclinical populations suggests that people in general are motivated to shape the way they are perceived by others through impression management, a process wherein people deliberately control how others perceive them to achieve social, emotional, or identity-related goals [21,55,56]. Aligning with our observation that people are motivated to be perceived as healthy, Leary and Kowalski [56] maintain that people are motivated to engage in impression management to obtain optimal social outcomes, such as acceptance from others, increased subjective well-being and self-esteem, and a strong identity. This may extend to social outcomes, including not being hospitalized.

Indeed, work on impression management has been extended to clinical populations; patients engage in impression management during intake interviews [57]. Because psychopathology has long been stigmatized in our society [58], these efforts at impression management are likely adaptive because stigma, including self-stigma, is related to negative health outcomes [59]. Particularly, STBs are quite stigmatized [60,61]. Such stigma yields a reciprocal relationship between experiences of suicide stigma and increased suicidality [60,62,63]. Therefore, avoiding being seen as suicidal may have adaptive consequences, particularly if a patient has previously observed or experienced negative socioemotional consequences of suicide stigma [64]. However, when stigma prevents people from seeking treatment, it in turn engenders a greater risk of suicide. One study did, in fact, identify stigma as a barrier to disclosure [49]. Taken together, our observation that impression management impacts disclosure may reflect a common human tendency to minimize expressions of distress to avoid unwanted social outcomes, likely due at least in part to the stigma of mental illness and ego-dystonic self-perceptions. Clinicians should discuss this tendency and perceptions of mental health stigma with their patients. This may help address stigma directly and highlight the harms of impression management and the importance of disclosure.

These findings indicate that disclosure decisions are shaped not only by individual symptom severity but also by relational dynamics and anticipated emotional consequences, underscoring the importance of explicitly addressing perceived burdensomeness, stigma, and impression management within therapeutic relationships.

Feeling Bad and Feeling Nothing

Creators reported that negative affect impacted their desire to disclose mental health symptoms. Loneliness, burdensomeness, and hopelessness are well-documented risk factors in the suicide literature [65,66]. Similarly, creators' discussion of emotional suppression aligns with literature suggesting that STBs are motivated and maintained by an intolerance of negative affect; people with STBs and nonsuicidal self-injury (NSSI) endorse a desire to avoid and escape negative affect [67-69]. This suggests common risk factors between STBs and nondisclosure, which have been found previously. Help-negation, a phenomenon where people with high levels of SI decrease their help-seeking behavior, has been observed in young people with SI [70,71]. This behavior is a key impediment to suicide risk detection and treatment. Our work provides needed insight into the underpinnings of the phenomenon, which was not simply explained by hopelessness or prior help-seeking in previous work [70].

While this study did not find shame as a subtheme, a common self-reported reason for concealment in therapy also includes the desire to avoid shame [5], similarly suggesting a lack of tolerance for negative affect, factoring into the decision-making process about disclosure. It may be that people who feel a lot of shame surrounding disclosure are unlikely to post about their emotions online. Indeed, there is an interesting dissonance

XSL•FO
RenderX

between the finding that people often do not disclose to their clinicians while actively disclosing to peers and a larger and unknown internet audience. It may be that disclosing to those without the power to hospitalize or peers is less difficult. It also may be that the subset of individuals who withhold information from their therapist yet disclose similar content to large online audiences is not representative of the broader population of individuals who conceal information in therapeutic contexts. Finally, creators expressed not knowing how they felt as an explanation for why they did not share information, aligning with literature suggesting that alexithymia is related to depression and NSSI [72,73]. Taken together, patients' symptoms themselves may be a barrier to disclosure, even if they have a desire to confide in their clinician.

One unexpected subtheme was patients' fear that they were faking their mental illness. Some seemed to be afraid of this possibility despite feeling ill, while others seemed to be in denial of their symptom severity until a clinician highlighted it. Another group of patients may be using self-invalidation as an attempt at denial via minimizing their symptoms to themselves. Previous work suggests that there is both self-stigma and perceived stigma surrounding perceived depression inauthenticity among adolescents online, which presented as a barrier to help-seeking [74]—this phenomenon may extend to STBs. Creators in our sample who feared that they were faking their symptoms may have been experiencing such self-stigma. It is also possible that creators are expressing a version of the thought that their SI is not that serious, which has been reported elsewhere as a reason for nondisclosure of STBs [13,50]. Of note, obsessive fears that one is unintentionally a liar, such as fear that one is "faking" a disorder, are seen in patients with obsessive-compulsive disorder [75]. More work is needed to understand this phenomenon. Psychoeducation about the importance of disclosing STBs in therapy, no matter the severity, may be warranted.

Negative Opinions of Psychiatric Treatment

Creators also discussed barriers to disclosure that related to the therapeutic relationship. Because a weak therapeutic alliance was a theme related to disclosure, building rapport is likely to be critical to build trust and reduce negative perceptions of treatment to create an environment where patients feel comfortable disclosing. Previous work suggests that a strong therapeutic alliance increases the likelihood of honest disclosure, and therapists' emotional responses to patients were correlates of disclosure [5,11,15]. Clinicians may benefit from providing extensive psychoeducation about the therapeutic relationship and the rationale for norms of psychiatric care, as well as inviting patients to share reservations they have about psychiatric treatment in order to have an open discussion about them. Additionally, increased transparency from clinicians may be helpful. For example, clinicians should clearly state when they will or will not be giving advice and why they are choosing to see a patient at a certain frequency. Previous work suggests that one reason underlying the lack of disclosure is a feeling that their clinician would not understand a particular issue [5], suggesting that building the strongest possible foundation of trust and understanding in the therapeutic relationship is the most crucial for this population, as the outcome is a matter of life and death. Notably, some creators also reported therapist misconduct, such as an instance where a therapist cried after a creator's disclosure. This finding echoes previous work that suggests people conceal information to avoid therapists' potential overreactions [5]. Considering this finding, clinical training should continue to emphasize appropriate and effective ways of building rapport and managing countertransference. Future work should also examine whether certain clinician characteristics predict patient nondisclosure or minimization. Additionally, patient dishonesty was associated with weaker therapeutic alliances and worse perceived therapeutic outcomes [76], suggesting that psychoeducation about the importance of disclosure may be warranted.

Feelings About Disclosure

In our exploratory thematic analysis, we found that a paucity of creators talked about being honest with clinicians. While this is likely due to our search terms, it is also possible that people are engaging in self-presentation to fit in with a sentiment that lying to clinicians is normative. Indeed, work shows that social media, including TikTok, may positively reinforce discussion of mental health symptoms by allowing users to gain social capital, attention, and belonging [77-79]. However, these symptoms may be performative or exaggerated for positive gain when identifying with psychopathology is perceived as desirable in online spaces [79]. While it is possible that discourse about nondisclosure has a similar function to identification with mental health symptoms, it seems more likely that our findings are best explained by the specificity of our search terms. As observed during our exploration of potential search terms, there is no unifying hashtag or term that discourse about nondisclosure to clinicians centers around. Thus, it is unlikely that sharing these narratives online is motivated by wanting to fit into an online social group shaped by a popular "sick role" narrative. Furthermore, creators expressed negative emotions as a result of nondisclosure. This suggests that clinicians may benefit from candid conversations with patients about the potential long-term effects of keeping symptoms private.

Broader Implications

Taken together, findings suggest that concealment may be an attempt to protect oneself from feared or actual negative consequences grounded in patients' previous treatment experiences or in cultural narratives about psychiatric care. Further work is needed to understand the downstream effects of expressing negative treatment experiences online. It may be the case that learning about others' disclosure of STBs leading to hospitalization on TikTok can influence one's willingness to disclose their STBs to their therapist. Future work should investigate this possibility. It is currently unknown whether this content discourages help-seeking and entrenches negative perceptions of treatment—both on the part of the creator and the viewer—or provides validation that heightens feelings of connection.

Future work should examine what differentiates people who do disclose from those who view disclosure as dangerous or futile, perhaps through protective traits, skills, or relational contexts. Prior work suggests that disclosure patterns vary across settings and individual characteristics. In one study, neuroticism and

XSL•FO
RenderX

trait anxiety predicted disclosure to research assistants, while lower extraversion predicted disclosure to clinicians. Extraversion and trait anxiety distinguished disclosure to research assistants from disclosure to clinicians [20]. Other studies highlight individualized psychological barriers to disclosure. People with self-destructive behaviors, defined as self-harm, substance abuse, or disordered eating, reported greater self-reported difficulty being honest in therapy compared with peers without such behaviors who also reported difficulty being honest in therapy [80]. Nondisclosure was also associated with social loneliness, poor health, frequent SI, and severe psychological distress [81], as well as personality disorder diagnoses and symptoms of suicide crisis syndrome [11]. On the other hand, reasons for disclosure have included help-seeking, shared background, catharsis [13], a strong therapeutic alliance, and intentions to address STBs [15]. Building on these findings, interventions such as motivational interviewing may help leverage motivations to promote disclosure.

Limitations and Future Directions

This study highlights several paths for future research. Work should examine whether improving emotional awareness and tolerance of negative affect increases disclosure, or whether people simply disclose more as symptoms improve. Given links between alexithymia, depression, and NSSI, interventions that enhance emotional articulation may facilitate disclosure and rapport. Future studies should also explore fears of "faking" illness as potential expressions of self-stigma, anxiety, or intrusive thoughts, and how family dynamics shape beliefs about managing others' emotions in clinical settings. Studies should also examine how online narratives about psychiatric care influence help-seeking and identify what differentiates patients who disclose honestly in clinical settings from those who do not. Finally, using standardized measures for assessment may also help surmount barriers to disclosure [49]. Particularly, rather than force a response, clinicians may allow for a nondisclosure response (eg, prefer not to answer) option in both self-report measures and during sessions [82]. To enhance the rigor of qualitative social media research, the field may benefit from developing structured scoring procedures to assess short-form video quality for qualitative analyses, comparable to the Global Quality Score or DISCERN [83].

Because TikTok users tend to be younger, some themes we uncovered may be adolescent or young adult specific. Indeed, adolescents and young adults may have unique reasons for choosing not to disclose, including fear of negative reactions from family and a sense of self-reliance [18]. In addition, there may be selection bias based on who decides to post online. People who feel intense shame and stigma, are more introverted, or have higher levels of social anxiety may not choose to share their experiences online. Indeed, a bidirectional relationship between shame and nondisclosure has been shown in a nonclinical sample [84]. This study also lacks information on how many clinical encounters discussed in our TikTok videos were in-person versus telehealth. Telehealth may prohibit clinicians from observing evidence of self-injurious behaviors on patients' bodies. As such, clinicians may be missing key clues that could initiate a conversation. Additionally, people may struggle to conceal in person but have an easier time doing so online. Future work should address this possibility. Another limitation of this study is that our data scraping procedures did not capture the dates when videos were posted, preventing us from examining temporal trends or shifts in public discourse over time. Although reporting dates are not common in work examining TikTok, future work may benefit from looking at the same search terms longitudinally to explore trends. Further, demographic information was not systematically available due to video filters, green screen, and other effects, and video formats. As a result, we were unable to examine whether motivations for concealment differed by demographic characteristics such as race, ethnicity, and gender identity.

Finally, the biggest limitation of this study was that the specific symptoms that patients were concealing were sometimes unknown. TikTok censors terms such as suicide and self-harm, and creators in our sample did not always explicitly state their symptoms. When creators were not explicit, we were often able to ascertain the creators' symptoms anyway. Creators often alluded to STBs (eg, the creator mouthed the word "suicide," gestured to slitting their throat, or included an oblique hashtag such as #sewersl1de). In addition, in videos that discussed adult involuntary hospitalization, the coding team conjectured that symptoms were related to being a danger to self or others. Future work should replicate this work on a less censored platform. Finally, given the observational nature of this work, we cannot verify the veracity of creator narratives, including whether they visited a health care provider or are in fact experiencing the symptoms they discuss concealing.

Conclusions

Thematic analysis of naturalistic, user-generated social media content suggests that concealment of psychiatric symptoms is often motivated by fear of punitive consequences, stigma, interpersonal dynamics (within family, treatment team, and one's broader social context), and dissatisfaction with treatment practices. These findings highlight modifiable relational and structural factors that may influence disclosure in clinical contexts. Efforts to improve transparency around hospitalization criteria, strengthen the therapeutic alliance, and directly address fears of stigma may reduce concealment and enhance accurate detection of psychiatric symptoms. This is critical to reduce death by suicide in people who voluntarily present to our hospital systems. Future work should examine how institutional policies and clinician communication strategies shape patients' willingness to disclose suicide risk.

Acknowledgments

Generative artificial intelligence was used to search for spelling and grammatical errors in the manuscript and to check for errors in the R code used to build Tables 2-4.

XSL•FO
RenderX

Funding

No funding was obtained for this study.

Data Availability

The videos analyzed in this study were publicly available at the time of data collection. URLs are not provided to protect creator anonymity. Deidentified coding data are available from the corresponding author upon reasonable request.

Authors' Contributions

CR contributed to conceptualization, methodology, software development, validation, formal analysis, investigation, data curation, visualization, supervision, project administration, and writing of both the original draft and the review and editing of the manuscript. KR contributed to formal analysis, investigation, visualization, and writing of the original draft, as well as review and editing of the manuscript. QZ and AP contributed to formal analysis, investigation, and data curation. MRK and HAD contributed to methodology, supervision, and review and editing of the manuscript. MKN contributed to conceptualization, provision of resources, supervision, and writing of the original draft, as well as review and editing of the manuscript.

Conflicts of Interest

MKN is a scientific advisor for Cerebral Inc. The remaining authors have disclosed no potential conflicts of interest, financial or otherwise.

References

1. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity survey replication. Arch Gen Psychiatry 2005 Jun;62(6):593-602. [doi: 10.1001/archpsyc.62.6.593] [Medline: 15939837]
2. Kessler RC, Demler O, Frank RG, et al. Prevalence and treatment of mental disorders, 1990 to 2003. N Engl J Med 2005 Jun 16;352(24):2515-2523. [doi: 10.1056/NEJMsa043266] [Medline: 15958807]
3. Baumann EC, Hill CE. Client Concealment and Disclosure of Secrets in Outpatient Psychotherapy: Disclosure and Concealment in Psychotherapy; 2017.
4. Blanchard M, Farber BA. “It is never okay to talk about suicide”: patients’ reasons for concealing suicidal ideation in psychotherapy. Psychother Res 2020 Jan;30(1):124-136. [doi: 10.1080/10503307.2018.1543977] [Medline: 30409079]
5. Farber BA. Disclosure, concealment, and dishonesty in psychotherapy: a clinically focused review. J Clin Psychol 2020 Feb;76(2):251-257. [doi: 10.1002/jclp.22891] [Medline: 31746459]
6. Drinane JM, Owen J, Tao KW. Cultural concealment and therapy outcomes. J Couns Psychol 2018 Mar;65(2):239-246. [doi: 10.1037/cou0000246] [Medline: 29543478]
7. Nagdimon J, McGovern C, Craw M. Addressing concealed suicidality: a flexible and contextual approach to suicide risk assessment in adults. J Contemp Psychother 2021 Sep;51(3):241-250. [doi: 10.1007/s10879-021-09493-9]
8. Fox KR, Huang X, Guzmán EM, et al. Interventions for suicide and self-injury: a meta-analysis of randomized controlled trials across nearly 50 years of research. Psychol Bull 2020 Dec;146(12):1117-1145. [doi: 10.1037/bul0000305] [Medline: 33119344]
9. Berman AL. Risk factors proximate to suicide and suicide risk assessment in the context of denied suicide ideation. Suicide Life Threat Behav 2018 Jun;48(3):340-352. [doi: 10.1111/sltb.12351] [Medline: 28429385]
10. Obegi JH. How common is recent denial of suicidal ideation among ideators, attempters, and suicide decedents? a literature review. Gen Hosp Psychiatry 2021;72:92-95. [doi: 10.1016/j.genhosppsych.2021.07.009] [Medline: 34358807]
11. Høyen KS, Solem S, Cohen LJ, et al. Non-disclosure of suicidal ideation in psychiatric inpatients: rates and correlates. Death Stud 2022;46(8):1823-1831. [doi: 10.1080/07481187.2021.1879317] [Medline: 33586630]
12. Calear AL, Batterham PJ. Suicidal ideation disclosure: patterns, correlates and outcome. Psychiatry Res 2019 Aug;278:1-6. [doi: 10.1016/j.psychres.2019.05.024] [Medline: 31128420]
13. Fulginiti A, Frey LM. Exploring suicide-related disclosure motivation and the impact on mechanisms linked to suicide. Death Stud 2019;43(9):562-569. [doi: 10.1080/07481187.2018.1504349] [Medline: 30285575]
14. Huggett C, Peters S, Gooding P, Berry N, Pratt D. A systematic review and meta-ethnography of client and therapist perspectives of the therapeutic alliance in the context of psychotherapy and suicidal experiences. Clin Psychol Rev 2024 Nov;113:102469. [doi: 10.1016/j.cpr.2024.102469] [Medline: 39098267]
15. Love HA, Morgan PC. You can tell me anything: disclosure of suicidal thoughts and behaviors in psychotherapy. Psychotherapy (Chic) 2021 Dec;58(4):533-543. [doi: 10.1037/pst0000335] [Medline: 33734741]
16. Maple M, Frey LM, McKay K, Coker S, Grey S. “Nobody Hears a Silent Cry for Help”: suicide attempt survivors’ experiences of disclosing during and after a crisis. Arch Suicide Res 2020;24(4):498-516. [doi: 10.1080/13811118.2019.1658671] [Medline: 31507236]
17. Sheehan L, Oexle N, Armas SA, et al. Benefits and risks of suicide disclosure. Soc Sci Med 2019 Feb;223:16-23. [doi: 10.1016/j.socscimed.2019.01.023] [Medline: 30690334]

XSL•FO
RenderX

18. Spears AP, Shin KE, Cha CB. Characterizing silence: adolescents' nondisclosure of their suicidal thoughts and behaviors to their family and peers. JAACAP Open 2025 Sep;3(3):496-505. [doi: 10.1016/j.jaacop.2024.12.005] [Medline: 40922770]
19. McGillivray L, Rheinberger D, Wang J, Burnett A, Torok M. Non-disclosing youth: a cross sectional study to understand why young people do not disclose suicidal thoughts to their mental health professional. BMC Psychiatry 2022 Jan 4;22(1):3. [doi: 10.1186/s12888-021-03636-x] [Medline: 34983460]
20. Bloch-Elkouby S, Zilcha-Mano S, Rogers ML, et al. Who are the patients who deny suicidal intent? exploring patients' characteristics associated with self-disclosure and denial of suicidal intent. Acta Psychiatr Scand 2023 Feb;147(2):205-216. [doi: 10.1111/acps.13511] [Medline: 36263445]
21. Goffman E. The presentation of self in everyday life. In: Social Theory Re-Wired: Routledge; 2023. [doi: 10.4324/9781003320609-59]
22. Banyard P, Grayson A. Demand characteristics. In: Introducing Psychological Research: Macmillan Education UK; 1996. [doi: 10.1007/978-1-349-24483-6_59]
23. Basch CH, Donelle L, Fera J, Jaime C. Deconstructing TikTok videos on mental health: cross-sectional, descriptive content analysis. JMIR Form Res 2022 May 19;6(5):e38340. [doi: 10.2196/38340] [Medline: 35588057]
24. Kells MR, Davis HA, Roske C, Holzman S, Wildes JE. The experience of treatment for eating disorders as told by content creators on TikTok. Eat Disord 2025;33(5):583-601. [doi: 10.1080/10640266.2024.2401669] [Medline: 39312503]
25. Kingod N, Cleal B, Wahlberg A, Husted GR. Online peer-to-peer communities in the daily lives of people with chronic illness: a qualitative systematic review. Qual Health Res 2017 Jan;27(1):89-99. [doi: 10.1177/1049732316680203] [Medline: 27956659]
26. McCashin D, Murphy CM. Using TikTok for public and youth mental health - a systematic review and content analysis. Clin Child Psychol Psychiatry 2023 Jan;28(1):279-306. [doi: 10.1177/13591045221106608] [Medline: 35689365]
27. Rueger J, Dolfsma W, Aalbers R. Perception of peer advice in online health communities: access to lay expertise. Soc Sci Med 2021 May;277:113117. [doi: 10.1016/j.socscimed.2020.113117] [Medline: 33865092]
28. Doyle B. TikTok statistics - everything you need to know. Wallaroo Media. 2024. URL: https://wallaroomedia.com/blog/tiktok-statistics/ [accessed 2025-05-29]
29. Eichstaedt JC, Smith RJ, Merchant RM, et al. Facebook language predicts depression in medical records. Proc Natl Acad Sci USA 2018 Oct 30;115(44):11203-11208. [doi: 10.1073/pnas.1802331115]
30. Jacobsen MH, Smith G. The Routledge International Handbook of Goffman Studies: Taylor & Francis; 2022.
31. Warner CH, Appenzeller GN, Grieger T, et al. Importance of anonymity to encourage honest reporting in mental health screening after combat deployment. Arch Gen Psychiatry 2011 Oct;68(10):1065-1071. [doi: 10.1001/archgenpsychiatry.2011.112] [Medline: 21969463]
32. Ma X, Hancock J, Naaman M. Anonymity, intimacy and self-disclosure in social media. Presented at: CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems; May 7-12, 2016; San Jose California USA p. 3857-3869. [doi: 10.1145/2858036.2858414]
33. John LK, Acquisti A, Loewenstein G. Strangers on a plane: context-dependent willingness to divulge sensitive information. J Consum Res 2011 Feb 1;37(5):858-873. [doi: 10.1086/656423]
34. Schlosser AE. Self-disclosure versus self-presentation on social media. Curr Opin Psychol 2020 Feb;31:1-6. [doi: 10.1016/j.copsyc.2019.06.025] [Medline: 31357096]
35. Hogan B. The presentation of self in the age of social media: distinguishing performances and exhibitions online. Bull Sci Technol Soc 2010 Dec;30(6):377-386. [doi: 10.1177/0270467610385893]
36. Bullingham L, Vasconcelos AC. 'The presentation of self in the online world': Goffman and the study of online identities. J Inf Sci 2013 Feb;39(1):101-112. [doi: 10.1177/0165551512470051]
37. Davis HA, Kells MR, Roske C, Holzman S, Wildes JE. A reflexive thematic analysis of #WhatIEatInADay on TikTok. Eat Behav 2023 Aug;50:101759. [doi: 10.1016/j.eatbeh.2023.101759]
38. Herrick SSC, Hallward L, Duncan LR. "This is just how I cope": an inductive thematic analysis of eating disorder recovery content created and shared on TikTok using #EDrecovery. Int J Eat Disord 2021 Apr;54(4):516-526. [doi: 10.1002/eat.23463] [Medline: 33382136]
39. Santarossa S, Lacasse J, Larocque J, Woodruff SJ. #Orthorexia on Instagram: a descriptive study exploring the online conversation and community using the Netlytic software. Eat Weight Disord 2019 Apr;24(2):283-290. [doi: 10.1007/s40519-018-0594-y] [Medline: 30317422]
40. Grant-Allen G, Wang L, Amini J, Dhaliwal S, Sinyor M, Mitchell RH. Self-harm and suicide-related content on TikTok: thematic analysis. J Med Internet Res 2025 Sep 18;27(1):e77828. [doi: 10.2196/77828] [Medline: 40811047]
41. Raffoul A, Santoso M, Lu J, Duran V, Austin SB. Diet pills and deception: a content analysis of weight-loss, muscle-building, and cleanse and detox supplements videos on TikTok. Eat Behav 2024 Dec;55:101911. [doi: 10.1016/j.eatbeh.2024.101911] [Medline: 39182366]
42. Braun V, Clarke V, Weate P. Using thematic analysis in sport and exercise research. In: Routledge Handbook of Qualitative Research in Sport and Exercise: Routlegde; 2016.
43. Braun V, Clarke V. Conceptual and design thinking for thematic analysis. Qual Psychol 2022;9(1):3-26. [doi: 10.1037/qup0000196]

XSL•FO
RenderX

44. Lincoln YS, Guba EG. Naturalistic Inquiry: SAGE; 1985.
45. R: a language and environment for statistical computing. R Developement Core. 2010. URL: www.R-project.org [accessed 2026-04-17]
46. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
47. Aboussouan A, Moscardini EH, Cerel J, Tucker RP. Experiences of hospitalization for suicide ideation and suicide attempt in gender diverse adults. Suicide Life Threat Behav 2022 Jun;52(3):427-438. [doi: 10.1111/sltb.12832] [Medline: 35083785]
48. Kleiman EM, Turner BJ, Fedor S, et al. Digital phenotyping of suicidal thoughts. Depress Anxiety 2018 Jul;35(7):601-608. [doi: 10.1002/da.22730] [Medline: 29637663]
49. Litschi MA, Lancaster SL, Linkh DJ, Lafferty M. "They'll Talk About Everything Else… But Suicidal Ideation": clinician experiences addressing non-disclosure of suicidal ideation among military-affiliated clients. J Gen Intern Med 2025 Nov;40(15):3684-3693. [doi: 10.1007/s11606-025-09570-y] [Medline: 40437137]
50. Burton Denmark A, Hess E, Becker MS. College students' reasons for concealing suicidal ideation. J College Stud Psychother 2012 Apr;26(2):83-98. [doi: 10.1080/87568225.2012.659158]
51. Chu C, Buchman-Schmitt JM, Stanley IH, et al. The interpersonal theory of suicide: a systematic review and meta-analysis of a decade of cross-national research. Psychol Bull 2017 Dec;143(12):1313-1345. [doi: 10.1037/bul0000123] [Medline: 29072480]
52. Joiner TE, Pettit JW, Walker RL, et al. Perceived burdensomeness and suicidality: two studies on the suicide notes of those attempting and those completing suicide. J Soc Clin Psychol 2002 Nov;21(5):531-545. [doi: 10.1521/jscp.21.5.531.22624]
53. Hooley JM. Expressed emotion and relapse of psychopathology. Annu Rev Clin Psychol 2007;3(1):329-352. [doi: 10.1146/annurev.clinpsy.2.022305.095236] [Medline: 17716059]
54. Nguyen TNB, Chance AB, Roske C, et al. Relationships between adolescent perception of family functioning and affective symptomatology. JAACAP Open 2025 Sep;3(3):793-804. [doi: 10.1016/j.jaacop.2024.09.002] [Medline: 40922792]
55. Bolino M, Long D, Turnley W. Impression management in organizations: critical questions, answers, and areas for future research. Annu Rev Organ Psychol Organ Behav 2016 Mar 21;3(1):377-406. [doi: 10.1146/annurev-orgpsych-041015-062337]
56. Leary MR, Kowalski RM. Impression management: a literature review and two-component model. Psychol Bull 1990;107(1):34-47. [doi: 10.1037/0033-2909.107.1.34]
57. Frühauf S, Figlioli P, Oehler D, Caspar F. What to expect in the intake interview? impression management tactics of psychotherapy patients. J Soc Clin Psychol 2015 Jan;34(1):28-49. [doi: 10.1521/jscp.2015.34.1.28]
58. Pescosolido BA, Halpern-Manners A, Luo L, Perry B. Trends in public stigma of mental illness in the US, 1996-2018. JAMA Netw Open 2021 Dec 1;4(12):e2140202. [doi: 10.1001/jamanetworkopen.2021.40202] [Medline: 34932103]
59. Hatzenbuehler ML. Structural stigma: research evidence and implications for psychological science. Am Psychol 2016 Nov;71(8):742-751. [doi: 10.1037/amp0000068] [Medline: 27977256]
60. Carpiniello B, Pinna F. The reciprocal relationship between suicidality and stigma. Front Psychiatry 2017;8:35. [doi: 10.3389/fpsyt.2017.00035] [Medline: 28337154]
61. Sudak H, Maxim K, Carpenter M. Suicide and stigma: a review of the literature and personal reflections. Acad Psychiatry 2008;32(2):136-142. [doi: 10.1176/appi.ap.32.2.136] [Medline: 18349334]
62. Oexle N, Waldmann T, Staiger T, Xu Z, Rüsch N. Mental illness stigma and suicidality: the role of public and individual stigma. Epidemiol Psychiatr Sci 2018 Apr;27(2):169-175. [doi: 10.1017/S2045796016000949] [Medline: 27919303]
63. Rüsch N, Zlati A, Black G, Thornicroft G. Does the stigma of mental illness contribute to suicidality? Br J Psychiatry 2014 Oct;205(4):257-259. [doi: 10.1192/bjp.bp.114.145755] [Medline: 25274313]
64. Oexle N, Ajdacic-Gross V, Kilian R, et al. Mental illness stigma, secrecy and suicidal ideation. Epidemiol Psychiatr Sci 2017 Feb;26(1):53-60. [doi: 10.1017/S2045796015001018] [Medline: 26606884]
65. Franklin JC, Ribeiro JD, Fox KR, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol Bull 2017 Feb;143(2):187-232. [doi: 10.1037/bul0000084] [Medline: 27841450]
66. Van Orden KA, Witte TK, Cukrowicz KC, Braithwaite SR, Selby EA, Joiner TE. The interpersonal theory of suicide. Psychol Rev 2010 Apr;117(2):575-600. [doi: 10.1037/a0018697] [Medline: 20438238]
67. Baumeister RF. Suicide as escape from self. Psychol Rev 1990 Jan;97(1):90-113. [doi: 10.1037/0033-295x.97.1.90] [Medline: 2408091]
68. Chapman AL, Gratz KL, Brown MZ. Solving the puzzle of deliberate self-harm: the experiential avoidance model. Behav Res Ther 2006 Mar;44(3):371-394. [doi: 10.1016/j.brat.2005.03.005] [Medline: 16446150]
69. Millner AJ, den Ouden HEM, Gershman SJ, et al. Suicidal thoughts and behaviors are associated with an increased decision-making bias for active responses to escape aversive states. J Abnorm Psychol 2019 Feb;128(2):106-118. [doi: 10.1037/abn0000395] [Medline: 30589305]
70. Deane FP, Wilson CJ, Ciarrochi J. Suicidal ideation and help-negation: not just hopelessness or prior help. J Clin Psychol 2001 Jul;57(7):901-914. [doi: 10.1002/jclp.1058] [Medline: 11406803]
71. Wilson CJ, Deane FP. Help-negation and suicidal ideation: the role of depression, anxiety and hopelessness. J Youth Adolesc 2010 Mar;39(3):291-305. [doi: 10.1007/s10964-009-9487-8] [Medline: 19957103]

XSL•FO
RenderX

72. Honkalampi K, Hintikka J, Tanskanen A, Lehtonen J, Viinamäki H. Depression is strongly associated with alexithymia in the general population. J Psychosom Res 2000 Jan;48(1):99-104. [doi: 10.1016/s0022-3999(99)00083-5] [Medline: 10750635]
73. Norman H, Oskis A, Marzano L, Coulson M. The relationship between self-harm and alexithymia: a systematic review and meta-analysis. Scand J Psychol 2020 Dec;61(6):855-876. [doi: 10.1111/sjop.12668] [Medline: 32706131]
74. Dixon-Ward KC, Chan SWY. Faking it": exploring adolescent perceptions of depression (in)authenticity and "attention seeking. Br J Clin Psychol 2022 Jun;61(2):177-196. [doi: 10.1111/bjc.12339] [Medline: 34716598]
75. Llorens-Aguilar S, Arnáez S, Aardema F, García-Soriano G. The relationship between obsessions and the self: feared and actual self-descriptions in a clinical obsessive-compulsive disorder sample. Clin Psychol Psychother 2022 Mar;29(2):642-651. [doi: 10.1002/cpp.2656] [Medline: 34342058]
76. Hart CL, Barnett M, Curtis DA. Client deception in therapy is associated with worse therapeutic relationships and perceived therapeutic outcomes. Couns and Psychother Res 2024 Sep;24(3):1014-1020. [doi: 10.1002/capr.12767]
77. Haltigan JD, Pringsheim TM, Rajkumar G. Social media as an incubator of personality and behavioral psychopathology: symptom and disorder authenticity or psychosomatic social contagion? Compr Psychiatry 2023 Feb;121:152362. [doi: 10.1016/j.comppsych.2022.152362] [Medline: 36571927]
78. Griffith FJ, Stein CH. Behind the hashtag: online disclosure of mental illness and community response on Tumblr. Am J Community Psychol 2021 Jun;67(3-4):419-432. [doi: 10.1002/ajcp.12483] [Medline: 33326631]
79. Harness J, Getzen H. TikTok's sick-role subculture and what to do about it. J Am Acad Child Adolesc Psychiatry 2022 Mar;61(3):351-353. [doi: 10.1016/j.jaac.2021.09.312] [Medline: 34534625]
80. Cersosimo BH, Farber BA. Is there something distinctive about psychotherapy clients' dishonesty about self-destructive behaviors? Couns Psychol Q 2022 Jan 2;35(1):230-241. [doi: 10.1080/09515070.2021.1929075]
81. Mérelle S, Foppen E, Gilissen R, Mokkenstorm J, Cluitmans R, Van Ballegooijen W. Characteristics associated with non-disclosure of suicidal ideation in adults. Int J Environ Res Public Health 2018 May 9;15(5):943. [doi: 10.3390/ijerph15050943] [Medline: 29747408]
82. Podlogar MC, Joiner TE. Allowing for nondisclosure in high suicide risk groups. Assessment 2020 Apr;27(3):547-559. [doi: 10.1177/1073191119845495] [Medline: 31053035]
83. Uzun O. Assessment of reliability and quality of videos on medial epicondylitis shared on YouTube. Cureus 2023 Apr;15(4):e37250. [doi: 10.7759/cureus.37250] [Medline: 37168186]
84. Carbone E, Loewenstein G, Sunstein CR, Dezső L. Spirals of Shame: The Bi-Directional Relationship Between Shame and Disclosure: Social Science Research Network; 2025. [doi: 10.2139/ssrn.5141799]

Abbreviations

NSSI: nonsuicidal self-injury
SI: suicidal ideation
STB: suicidal thoughts and behavior

Edited by A Mavragani; submitted 21.Nov.2025; peer-reviewed by A Al-Asadi, A Raffoul, U Pasupulety; revised version received 14.Apr.2026; accepted 15.Apr.2026; published 19.May.2026.

Please cite as:
Roske C, Ragnini K, Zhu Q, Palmer A, Kells MR, Davis HA, Nock MK
Why People Conceal Mental Health Problems: Qualitative Analysis of TikTok Posts
J Med Internet Res 2026;28:e88244
URL: https://www.jmir.org/2026/1/e88244
doi:10.2196/88244

© Chloe Roske, Kael Ragnini, Qinchun Zhu, Ashari Palmer, Meredith R Kells, Heather A Davis, Matthew K Nock. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Misinformation in Social Media Narratives on Highly Pathogenic Avian Influenza: Systematic Content Analysis of Facebook and Instagram Posts

Ahmed Al-Rawi[1*], PhD; Betty Ackah[1*], PhD; Abdelrahman Fakida[1*], MA; Kelley Lee[2*], DPhil

[1]School of Communication, Simon Fraser University, Schrum Science Centre-K 9653, Burnaby, BC, Canada

[2]Faculty of Health Sciences, Simon Fraser University, Burnaby, BC, Canada

[*]all authors contributed equally

Corresponding Author:
Ahmed Al-Rawi, PhD
School of Communication, Simon Fraser University, Schrum Science Centre-K 9653, Burnaby, BC, Canada

Abstract

Background: Recurrent outbreaks of the highly pathogenic avian influenza (HPAI) A (H5N1) virus in farmed poultry, and reports of infections in dairy cattle herds in the United States since March 2024, have triggered concerns about the spillover threat to human populations and a subsequent influenza pandemic. The increasing threat that H5N1 poses to human health has led to more vigilant public health monitoring of these developments. In addition to intensifying surveillance, preventative strategies—like vaccinating those at higher risk—are being evaluated to help minimize infection and spread.

Objective: Efforts to mitigate and respond to such an event will entail broad public health interventions including vaccination. However, analysis of the COVID-19 pandemic suggests that information quality can significantly impact the effectiveness of such measures by influencing public understanding and trust. Misinformation about H5N1 and other viruses circulating online often includes inaccurate information about transmission, prevention, and the severity of the viruses. By systematically analyzing these false narratives, public health authorities can better tailor their pandemic prevention, preparedness, and response strategies.

Methods: In light of the emerging threat of H5N1, we analyzed the content of social media posts from Facebook (approximately 350,000) and Instagram (n=69,551) related to HPAI. Using 40 keywords associated with misinformation, we identified over 500 posts explicitly mentioning H5N1 and related terms for further systematic analysis. Posts were coded to identify targets and topics in the social media narratives. The “target” refers to the organization or person mentioned in the post, while the “topic” refers to the primary issue or subject being addressed.

Results: Our content analysis identifies 7 main targets of misinformation, including government (149/544, 27%), health authorities (108/544, 20%), and international organizations (74/544, 14%). Also, from the 6 topics that have been identified, we found that the most widespread one was that authority figures purposefully engineer pandemics to achieve multiple political, economic, and other objectives (362/544, 67%) followed by societal destruction (121/544, 22%), and anti-vaccination (84/544, 15%). Other themes include societal destruction and religious allusions and prophecies.

Conclusions: Our analysis of online content showed that H5N1 misinformation was primarily aimed at individuals or groups with differing degrees of political or institutional authority, such as government leaders and public health officials. These figures were often the focus due to their involvement in making health policy decisions and implementing public health measures. Decision-making entities and individuals were the target of various misinformation narratives. Results demonstrate the ongoing need for monitoring health misinformation to inform evolving public health responses to HPAI.

(J Med Internet Res 2026;28:e73275) doi:10.2196/73275

KEYWORDS

highly pathogenic avian influenza; H5N1; social media; misinformation; health communication

Introduction

The highly pathogenic avian influenza (HPAI) H5N1 virus was first identified in poultry flocks in 1959, with the first human infection detected in 1997 [1-3]. The virus has since circulated worldwide, causing serious outbreaks in wild and domesticated birds, with periodic human cases reported [4]. Since 2021, H5N1 has been identified in a growing number of other wild, domestic, and zoo mammals (48 species by August 2024), in some cases causing major die-offs [5,6]. In March 2022, infections were first detected in dairy cattle in the United States, with more than 700 herds confirmed as infected by January 2025 [7]. Around 40 human infections linked to dairy cattle by January 2025 have

XSL•FO
RenderX

raised concerns about the possibility of an influenza pandemic in human populations [8]. While the current assessed risk of human infection ranges in severity from low to medium symptoms compared to earlier assessments [4,9], the further spread of H5N1 to other mammals increases the potential for genetic reassortment that promotes further spillover events and sustained human-to-human transmission.

The growing risk to human health from H5N1 has prompted stronger public health monitoring to these unfolding events. Alongside heightened surveillance efforts, preventative measures such as vaccination of high-risk populations are being considered to reduce infection and transmission. However, substantial evidence suggests that the ability to implement such measures could be limited in the wake of the COVID-19 pandemic [10]. Public trust was a key factor in achieving effective public health outcomes during COVID-19 specifically to achieve public compliance with recommended measures [11]. Online misinformation has been found to erode public trust, which has, in turn, facilitated vaccine hesitancy and refusal [12]. However, levels of public trust in the United States in key health agencies declined between June 2023 and December 2024, with 53% (down from 64%) trusting the Food and Drug Administration to make the right recommendations. Sixty-one percent (down from 66%) trust the recommendations of the US Centers for Disease Control and Prevention (CDC) [13]. While there are multiple factors contributing to increased mistrust, a decline in quality of the health information environment is a key issue.

This study seeks to inform emerging public health responses to H5N1 by understanding the content of misinformation on selected social media platforms related to H5N1. To date, there has been limited study of this subject. In part, this is due to increasing controls put into place by social media companies regarding access to data including for scholarly purposes [14]. For instance, X (formerly Twitter) discontinued its free Academic API (application programming interface) in February 2023. Meta's CrowdTangle, which allowed researchers free but limited access to data from Facebook and Instagram, is no longer available from August 14, 2024 [15,16]. This reduced access poses new challenges for understanding trends in misinformation. Using CrowdTangle and another data collection tool, this study navigates these barriers to understand the targets and topics of misinformation related to H5N1 on Facebook and Instagram. The findings show that such data can offer useful insights when monitoring social and behavioral factors that may impact public health efforts to mitigate or control a public health emergency [17].

Studies show that the ubiquity of internet connectivity in many societies has now made social media platforms primary information sources and means of communication [18]. Social media serve as virtual spaces for interactive engagements, particularly during times of uncertainty such as public health emergencies [19]. While these technologies enable connectivity in a globalized world, they can also be the source and rapid spreader of low-quality information [20]. This includes misinformation defined as misleading, inaccurate, or false information, although in some cases, there might be a facet of truth. Misinformation is distinguished from disinformation primarily by intent. Disinformation refers to the deliberate spread of false content with the intent to mislead or misinform [21,22]. However, due to the challenges of deducing the intent behind false information, we use the term misinformation to collectively describe all content in our dataset that is associated with falsehood or misleading claims. This approach is aligned with the trend by researchers toward understanding information quality separate from intent [23]. There is substantial evidence that both are part of what has been termed the post-truth world where people reject objective facts in favor of information aligned with ideology, identity, and other biases over verified facts [24,25]. This resulted in a proliferation of health misinformation across social media platforms during the COVID-19 pandemic. Moreover, this trend is likely to continue, and even worsen, as social media platforms such as X and Meta discard commitments to fact-checking in favor of leaving online communities "to decide when posts are potentially misleading and need more context" [26].

There remains limited but growing scholarly study on the content of misinformation related to H5N1. Studies show that misinformation concerning infectious diseases is more substantial and spreads more widely and quickly on social media than verifiable information [27-29]. Previous studies to date have analyzed the relationship between misinformation and a broad range of infectious diseases such as Ebola, H1N1, and Middle East Respiratory Syndrome [25,30,31]. Other studies have investigated social media misinformation against vaccination [27].

The structural features of social media platforms further the spread of misinformation as they often prioritize sensational content. Vosoughi et al [28] examined approximately 126,000 stories tweeted by about 3 million users on X to reveal that misinformation diffused at a significantly accelerated rate than verifiable information. Characteristics such as their novelty and emotional intensity facilitate their rapid diffusion. This phenomenon is particularly evident during health crises like avian influenza outbreaks [32].

For example, Yan et al [33] explore the relationship between social networking status, that is, the modes of information diffusion, relationship building, specific algorithmic exigencies, and influenza susceptibility. Gori and colleagues [34] examined social media debates on vaccines during the early part of the COVID-19 pandemic to understand the dynamics of vaccine readiness in Italy. The study highlighted how online debates using anti-vaccination rhetoric can influence the behavior of the vaccine-hesitant compared to pro-vaccination arguments.

Other studies posit the salience of factors such as people's reliance on these platforms as primary sources of news and individual-level predispositions to a conspiratorial mindset [35,36]. As well, public concerns about the impacts of infectious disease outbreaks on social media have intensified in the wake of the COVID-19 pandemic. These underlying factors have amplified the rapid spread of misinformation, contributing to the rise of infodemics during epidemics and pandemics. They present significant dangers, leading to public panic, reducing trust in health authorities, and resulting in uninformed health decisions that inhibit optimal uptake of public health

XSL•FO
RenderX

interventions. People who adopt inaccurate information about disease events are more likely to reject factual information about their severity, thus inhibiting their willingness to comply with appropriate measures [37].

Evidence suggests that the public's interpretation of information on infectious diseases on social media platforms can influence behaviors in response to relevant interventions, which, in turn, can impact the trajectory of a health crisis [38]. This is particularly important for early interventions during an emerging crisis that have the potential to become pandemics [39]. Health crises tend to heighten levels of public uncertainty, anxiety, and fear. Studies show that online spaces offer opportunities for people to engage with social networks and information sources to allay their concerns [40,41]. However, as Wang et al [29] found, elevated levels of anxiety and fear can make people more susceptible to poor quality information.

The above research has identified best practices for health communication to attend to public concerns, facilitate appropriate interpretation of the information, and support trust in public health interventions [17]. For instance, the enduring impact of COVID-19 has affected people's trust toward organizations like the US CDC and the government, in some cases augmenting an aversion to vaccination in general. Knowing these factors is key for appropriate communication framing [17]. Indeed, the COVID-19 pandemic showed how misinformation can significantly hamper public health interventions during health emergencies.

In this context, a better understanding of the content of misinformation on social media can inform effective public health responses to the evolving threat from H5N1 [42]. Misinformation on H5N1 may thus hinder public health efforts to identify, mitigate, or respond to the spread of the virus in animal and human populations by fostering further mistrust in public health responses [43]. Misinformation also has repercussions on individual and societal health outcomes by influencing health-seeking behaviors and compliance with public health interventions [17,29,44]. Although there are different subtypes of highly pathogenic avian influenza, such as H5N6 and H7N9 that have raised concern as well and remained important to analyze, our scope focuses on H5N1 due to its role as a primary subtype on known global outbreaks and designated by health authorities as a virus with pandemic potential. Moreover, H5N1 received wide media and public attention, making it particularly relevant for a study on health misinformation [45]. In sum, research on health misinformation related to H5N1 is limited, which this study seeks to address amid the ongoing spread of the virus. Thus, this study aims to answer the following research question: What is the content of misinformation about H5N1 on social media and how can this inform public health efforts at pandemic prevention, preparedness, and response?

Methods

Overview

Using scientific and colloquial terms associated with HPAI and specifically H5N1, such as bird flu, avian influenza, and avian flu, we collected approximately 350,000 Facebook and Instagram posts referencing H5N1 and associated terms using both CrowdTangle and a webometric tool called Zeeschuimer for more Instagram data. The data included both textual and multimedia content. We focused on these platforms because they are among the 3 most accessed social media platforms globally. Facebook is the market leader with an estimated active user population of over 3 billion [46]. Although YouTube is ahead of Instagram as the second most popular platform, its content being primarily video-based, with minimal text, excluded it from this study. We also selected Facebook and Instagram to expand the knowledge base on health misinformation beyond the platform X, especially regarding the affordances of social media platforms.

The CrowdTangle data were collected from June 5, 2014, to June 4, 2024, the posts are not limited to geographical demarcations. We collected the additional Instagram data with Zeeschuimer in July 2024. The total Instagram posts were (n=69,551). For both the Facebook and Instagram datasets, we employed a Python script to identify misinformation-related posts using terms such as "plandemic," "medicalmurder," and "LuciferShot" [47,48]. We compiled this list of terms based on research on health misinformation from academic literature, [49,50], online exploratory observations, and notes from the research team. Keywords were specifically focused on pandemic and vaccine conspiracy–related terms, demonstrating general connections for misinformation narratives between COVID-19 and vaccines. The study period from June 5, 2014, to June 5, 2024, was selected to ensure a proper review of discussions on the virus and its continued activity and outbreaks. A specific instance occurred in March 2024 when H5N1 infections were reported in dairy cattle in the United States for the first time, highlighting its continued relevance for public discussions [51].

Our final dataset contains 544 posts after removing duplicates. Though the volume of posts flagged as misinformation and misleading appears to be limited, we note several limitations. First, and in searching for specific keywords, it is difficult to confirm the identification of all the available misleading or misinformation-filled posts in the dataset, but the above approach has been effective in identifying a relevant subset of data that we can further explore. In addition, since the two data collection tools are not comprehensive in getting data from Meta, we cannot make any claims about the volume of misinformation on Facebook and Instagram. Moreover, as the initial number of gathered posts (350,000) is far larger than the remaining posts we analyzed, the difference in numbers is due to our inclusion criteria that excluded any duplicates, irrelevant use of keywords, and posts beyond our scope. This reduction also represents the inclusion criteria to ensure relevance.

Employing a general inductive approach [52] as part of systematic content analysis approach, 3 coders carried out a close reading of a 10% sample of the dataset to outline the emerging and dominant arguments. The analysis was multimodal, which involves the text and any associated multimedia content. Following a process of several rounds of phased discussions, the researchers organized the broad themes into 2 categories (targets and topics). These categories were used to link the analysis to the research objectives of

XSL•FO
RenderX

understanding the prevailing misinformative concepts on HPAI to inform effective public health agenda. The target is the organization or individual discussed in the post. The topic concerns the main subject or issue in the post, ensuring that subcategories were mutually exclusive and that the themes of the various subcategories are unique to each category, while the central messages do not overlap. Two researchers then used this initial codebook on a sample of the first 100 posts to test the codebook and test intercoder reliability. After the first coding round, the researchers adjusted the target section of the codebook, adding corporations as a category to account for multinational and other commercial entities, along with a category for irrelevant posts. The codebook has nine categories under target: (1) government had the highest occurring target references (n=149), (2) public health authorities had 108 references, (3) international organizations (n=74), (4) elites (n=64), (5) pharmaceutical industries (n=52), (6) mainstream media (n=43), and (7) corporations (n=23). The final categories for targets were (8) other/unclear (n=113) and (9) irrelevant (n=53). We distinguished between these 2 categories because while some posts were focused on H5N1 but did not address a target (other), others did have targets but had topics that were unrelated to avian influenza (irrelevant).

Researchers also adjusted the topics section following the first round of coding to address vagueness and overlapping categories. After the second round of intercoder reliability, further adjustments were made, including renaming a category to religious allusions and prophecies to include a broader range of religious content. Moreover, we removed an alternative medicine category as there were only 2 posts, and they were better classified under other. The adjusted seven categories are (1) pandemic engineering, which accounted for 362 topics, (2) societal destruction (n=121, 22%), (3) anti-vaccination (n=84, 15%), (4) elections/democracy (n=66, 12%), (5) unjustified food insecurity (n=59, 11%), (6) religious allusions and prophecies (n=26, 5%), and (7) other/irrelevant (n=77, 14%). The final category contained posts that were not focused on H5N1 but on other pathogens and diseases such as SARS-CoV-2/COVID-19. Others did not have a clear message and sometimes also used contradictory hashtags [49]. There were also 2 posts advancing misinformative pseudoscientific remedies to H5N1 infections by advertising plant-based vaccines. After 3 attempts involving the modifications described above, intercoder reliability was statistically acceptable at 0.887 for targets and 0.899 for topics using Krippendorff Alpha [53]. Though limited contextual insight was identified due to lack of context, we also examined the nature of the social media posters by reviewing their profiles and previous posts. Since the evidence we collected on users' profiles was only anecdotal, we did not report these findings in this study. Finally, as we analyzed publicly available social media data, all posts used have been deidentified and paraphrased to prevent re-identification and to protect user privacy.

Ethical Considerations

This project received an exemption from the requirements of Research Ethics Review at Simon Fraser University (30002857). All analysis is based on publicly available and deidentified data. All publicly available social media content used in this study was deidentified to prevent reidentification of posts or individuals.

Results

Targets

Types of Targets

The identified targets (see Figure 1) were individuals or entities that possess varying levels of economic, political, or decision-making power. An underlying theme in the claims made against them centered on implying that they intentionally generated public concern.

XSL•FO
RenderX

Figure 1. Percentage distribution of targets of misinformation on H5N1.

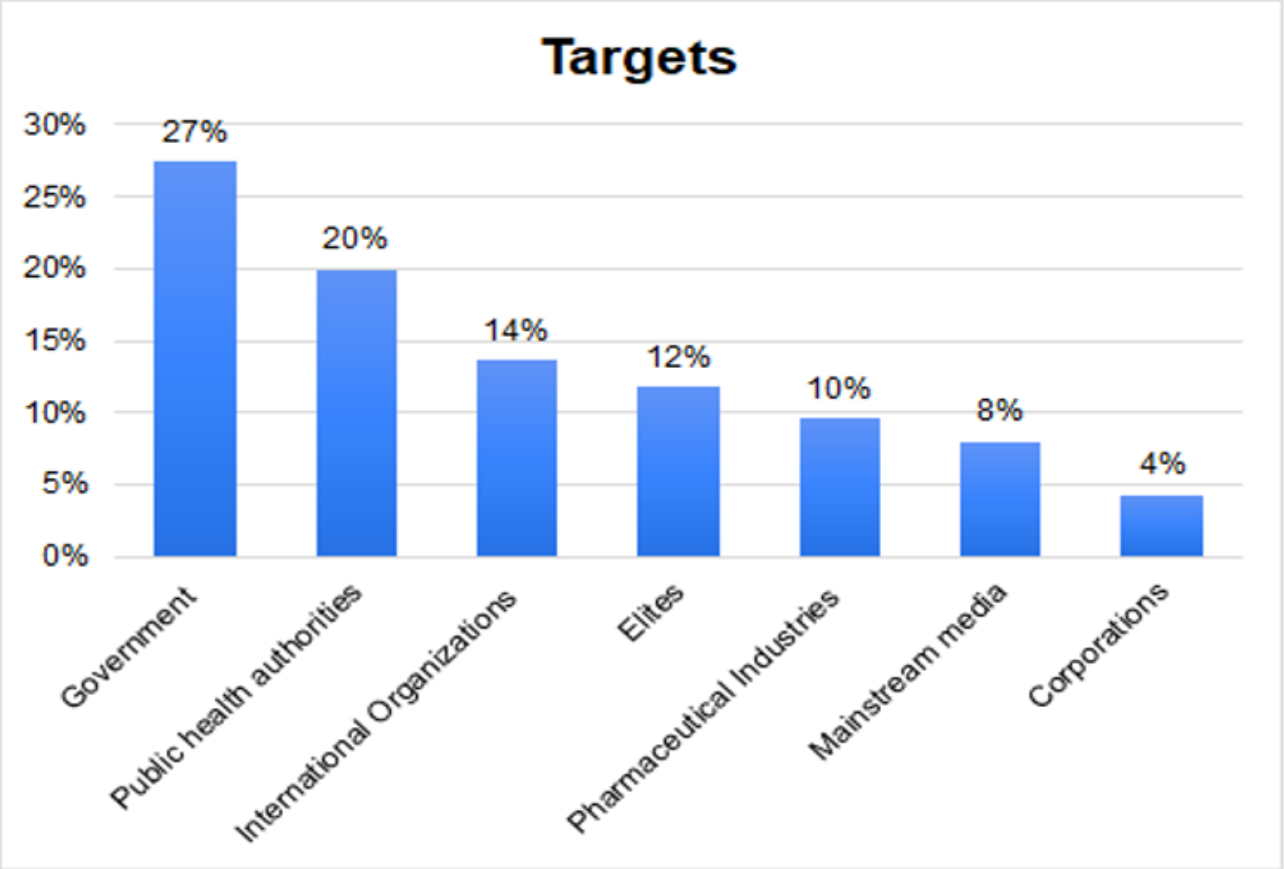

Government

This category constituted 27% (149/544) of targets, with figures being rounded off to the nearest whole number. The posts addressed the roles and responsibilities of governments and governmental entities in relation to pandemics, including heads of states, ministries, officials, and agencies in different countries, and the United States Food and Drug Administration. The category included posts claiming that the government and its "deep state" actors created pandemics for their own benefit such as to influence elections or to benefit other corrupt actors and individuals. Posts often highlighted the COVID-19 pandemic measures that users claimed resulted in societal destruction, suggesting that warnings of avian influenza can lead governments to reimpose similar restrictions. The following paraphrased post illustrates this point:

> *Don't worry about #covid_19 the government is scaring you and lying to you to start a new digital money system It's the same pattern they used before to [expletive] us and have reasons to push another PLANDEMIC. [expletive] them!*

In this example, wherein posts may contain typos as they are reproduced verbatim, the user points out different health crises such as COVID-19, Ebola, and avian influenza, suggesting that these crises were falsely created by the government as part of a larger conspiracy to impose certain financial measures on citizens. These are similar to other posts in the dataset in which users connect different health crises to argue that governments created or pushed these health crises to sow fear among citizens, thus creating conditions that allow the imposition of corrupt measures.

Public Health Authorities

Specific public health officials were targeted in posts, notably well-known figures such as Anthony Fauci, then director of the National Institute of Allergy and Infectious Diseases, and Robert Redfield, then director of the CDC. For instance, users frequently highlighted Redfield's past warnings about the possible threat of an avian influenza pandemic as evidence of planned manipulation and creation of a "plandemic." These posts also referred to public health officials as part of a broader network of parties that benefit from spreading fear about pandemics to advance individual interests. Public health officials comprised 20% (108/544) of the targets.

International Organizations

Fourteen percent (74/544) of the social media posts identified linked international organizations, especially those involved in pandemic responses such as the World Health Organization (WHO) and the World Economic Forum, with efforts to push political agendas that the posts claimed serve or align with their interests. This post, for example, accused the WHO of spreading false information regarding an avian influenza death to create a health crisis that can help control the public and push for certain medical interventions. The paraphrased excerpt below partly explains this claim:

> *The terrorist global health care organization is making up stories about someone dying after being infected with #avian influenza… because those in*

XSL•FO
RenderX

> *power want to turn us into mindless people… to push their medicines and spread panic and keep us under their control. Don't trust them; they are manipulating you. Make sure to focus on your healthy habits and keep learning and praying.*

Elites

Elites are individuals perceived by users to hold power and influence in the health sector and pandemic response such as Bill Gates, American businessman and philanthropist, and Klaus Schwab, the founder of the World Economic Forum. Constituting 12% (64/544) of the targets, the posts connected them with larger conspiracy theories, suggesting they had ambitions to reduce meat consumption, which was linked to perceived food shortages, and advanced the "Great Reset," a theory claiming that the COVID-19 pandemic measures were orchestrated by global elites allegedly to achieve their own goals, such as the following paraphrased quote:

> *It seems that the next Plandemic could be happening soon. This could be airborne and will be blamed on the new bird flu strain. Be careful, there is always another plan in the works. Billionaires already told us that they will go for our attention during the new pandemic, and the global health organizations and those globalists planned this new disease and created a Global Pandemic plan because they want to control our nations through a "Health" emergency.*

Pharmaceutical Industries

Posts focused on pharmaceutical companies accounted for 9.57% (52/544) of posts. Several users who promoted skeptical messages about avian influenza and pandemics highlighted the role they claimed these pharmaceutical companies played with other discussed entities such as the government and international organizations to amplify perceived medical threats and eventually sell the vaccines these companies produced. For instance, one post commented on avian influenza-related news saying that "psychic vaccine makers" prepared a vaccine for avian influenza a year in advance, suggesting that this is a "plandemic." Such posts reference vaccine developments as evidence that pandemics are intentionally created to help vaccine manufacturers.

Mainstream Media

Posts targeting mainstream media amounted to approximately 8% (43/544). Users who criticized mainstream media outlets portrayed them as tools used to spread fear about pandemics and diseases, and sometimes to encourage people to take vaccines, by claiming that they are manipulating and exaggerating to eventually support corruption that allows governments and pharmaceutical companies to benefit from pandemics.

Other Private Companies

Four percent (23/544) of the posts identified targeted nonpharmaceutical companies. Among the allegations were claims that various companies influence governments and organizations to declare pandemics, given that these companies were described as profiting from health crises. Some posts alleged that food companies, for example, pushed governments to use avian influenza to reduce the number of poultry in small farms or to mandate vaccines that could harm chickens. Such claims argue that these corrupt measures create food shortages that eventually corporations that control poultry supply will benefit from.

Topics

Types of Topics

The dataset included an extensive range of topics that were categorized according to the most prevalent narrative arguments (see Figure 2). The proportion of posts referring to the engineering of infectious disease pandemics was significant. The topics largely attributed major infectious disease events to human actions rather than any roles that pathogens play.

XSL•FO
RenderX

Figure 2. Percentage distribution of H5N1 misinformation topics.

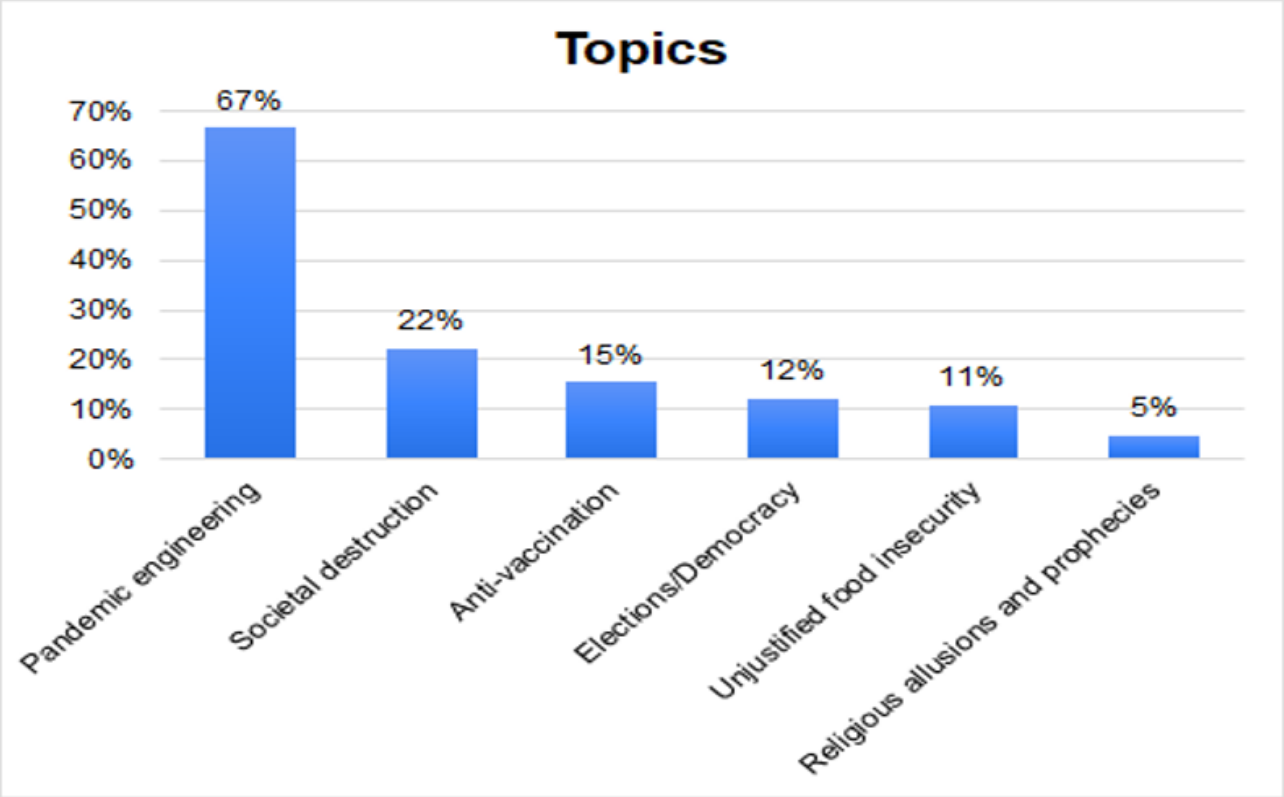

Pandemic Engineering

These posts (362/544, 67%) claim that the H5N1 outbreaks were allegedly engineered. They claimed that viruses causing widespread outbreaks are fabricated bioweapons aimed at stimulating economic gain for their creators or enshrining political power for governments or international organizations. Other targets of pandemic engineering claims are pharmaceutical companies and elites like Bill Gates. Popular hashtags associated with this theme include "#scamdemic" and "#plandemic." This post illustrates narratives about antiviral medications being specifically designed as bioweapons to manipulate people's beliefs such as the following paraphrased claim:

> *That is strange, they are already creating a new medication for the new flu. This drug and its company already know all the potential side effects …including sensory symptoms and continued false beliefs. The company that created the drug could just be saying that side effects could seem like a zombie apocalypse…They are clearly experimenting these medications on us to change our thoughts and memories!*

Societal Destruction

Amounting to 22% (121/544) of posts, this category is made up of posts that described widespread negative consequences from H5N1 outbreaks and public health measures. The consequences outlined were either projections of future occurrences or outcomes that are described as already happening, such as suicides and economic crashes. In the example below, the user indicated that pandemics of any kind could result in economic crashes and wars. Thus, alerts of avian influenza and swine flu infections were particularly disturbing in the wake of the COVID-19 pandemic and its ensuing effect. The paraphrased quote below further explains this point:

> *Oh, no! Now there is pig flu in addition to covid? They are saying that influenza this year, will be bad. Pandemics can start economic disasters and economic issues can lead to conflict. The elites and globalists and others are creating a perfect storm." It's the same people who created the New World Order and are now using China to bring that "perfect storm" on Chinese people and affecting all of us.*

This post encapsulates multiple theories that proponents of misinformation disseminate. One is the idea that pandemics are created to intentionally destabilize the political and economic systems of the world. Amid the ensuing chaos, global hegemonic figures are portrayed as aiming to establish a new hierarchy, a New World Order. The transformed global landscape is generally framed in economic terms; dominant forces will gain tremendous wealth and concretize their power. In this instance, the misinformation posts deviate from systematic examinations of pandemic impacts by overwhelmingly focusing on the negative impacts of pandemics and a purported new world order, while rejecting the public health measures (such as vaccinations), which could help mitigate disease spread and therefore minimize the impact.

Anti-Vaccination

Anti-vaccination sentiments were the third most dominant topic at 15% (84/544). The example below demonstrates the majority of the reasons that users gave for rejecting vaccination, including the profiteering motivations driving pharmaceutical companies to manufacture vaccines and vaccines being allegedly used to

XSL•FO
RenderX

implement a depopulation agenda. Others asserted that the vaccine manufacturing process is largely experimental and that not enough tests are run before they are administered to humans. Another conspiracy theory was that vaccines are used to insert digital biometric identity mechanisms into people, which would then be used to enforce a digital economy.

> *This is not good! They will do anything for money. These companies and health care providers will specially target poor neighborhoods. Vaccines are hurting our children, they're toxic and made of chicken embryos and from bird flu, pigeon flu, avian influenza, H1N1 and other viruses. Stop using our kids for testing drugs! Wake up and stop it!*

A notable aspect of misinformative posts against vaccines was how proponents seek to legitimize their messages by adapting "seeds" of facts. For instance, some posts refer to the history of cell strains originally derived from human fetal tissue in the development of some vaccines (from text in an image): "What's in your needle?… measles virus…monkey kidneys…aborted fetal tissues…chicken embryo."

Elections/Democracy

Posts discussing the connection between H5N1 and topics related to elections or democracy made up 12% (66/544) of the subset. Some users emphasized the political ramifications of an avian influenza outbreak. They include a perceived attempt by liberal governments to use outbreaks to curtail people's rights and freedoms and a socialist takeover by China supported by the United States and other governments. These allegations were frequently presented in tandem with other conspiracy theories. Many posts thus contended that the attacks on democratic freedoms would be facilitated by the modified vaccines being used to digitally tag everyone. The perceived attacks on rights and freedoms are usually linked to people's rejection of social distancing, vaccination, and other mandates, which were instituted in many countries during the COVID-19 outbreak. Posts frequently included multiple conspiracy-related narratives connected to reinforce the claim. In some cases, even contradictory assertions are believed to be true at the same time. This example from our dataset revealed the convergence of an engineered pandemic, fake climate change, and attempts to sabotage elections, all of which are being allegedly implemented by globalists.

> *I'm actually worried about those who turned against their neighbors during the pandemic (while those who doubted all this were actually right); this is similar to how people snitched on their neighbors during wars. The same could happen when the next fake pandemic is created bird influenza? Climate change? The next political scheme and elections?*

Misinformation About the Food System

Among the actions that health authorities in Canada and internationally implemented to control H5N1 infections in farmed flocks are quarantines, restrictions on imports and exports, culling of birds, and disinfection of objects and premises. Another concern that was expressed in the dataset involved food security in relation to HPAI interventions such as depopulation of poultry farms. The posts below illustrate assertions about authorities' intentional use of avian influenza to cause food insecurity by creating insufficiency in meat, eggs, and other food supplies.

> *They are using fake results in PCR tests to kill animals and prepare for an Avian influenza. Plandemic and land misuse reduce the amount of food available…this is all to create a crisis and push people towards digital IDs and force them to take vaccinations and what's worst is that they will replace meat with fake products and insects. We have to fight; this doesn't support the corrupt system.*

The example demonstrates the above-discussed confluence of conspiracy theories, where proclamations of fabricated food crises feed into fears about digital identification, fossil fuels, and a new economic world order. Another post linking avian influenza to an agenda to advance insect and lab-grown meat diets included the following hashtags touting multiple theories since hashtags produce hypertexts [54]: #foodshortage #thegreatreset #depopulationagenda #newworldorderagenda #globalists #elites #climatechangehoax #clotshot #birdflu #populationcontrol #labgrownmeat #waronmeat #warmongers #billgatesisevil.

A final noteworthy theme under the food system is allegations that the Chinese government is gaining control over food supplies by unleashing H5N1 as a weapon of dominance. Misinformation on the food system was 11% (59/544) of the subset.

Religious Allusions and Prophecies

This category comprised approximately 5% (26/544). They all espoused Christian or Biblical allusions, including repentance from sin as an infection prevention method, apocalyptic prophecies on various catastrophes, which human societies will experience as the world nears the end of its lifespan, and vaccines being used to implant the mark of the beast. Chapter 13 of the Biblical Book of Revelation notes that during Earth's final days, a beast will assume global leadership and require everyone to receive an identifying mark that will enable them to engage in societal transactions. The marks will be placed either on foreheads or on right hands.

> *The prediction is about to happen; don't accept these tech implants that they want to put in your hand and head. They want to add online money to your body. This bird flu is part of the next fake pandemic to spread digital identification and currency.*

The above illustrates misinformative claims that vaccines contained materials such as nanochips that marked immunized people with the mark of the beast. A recurring aspect of the messages in this category was the correlations made between announcements forecasting disease events and Biblical prophecies. Over time, and especially with reports of infections in mammals such as cows, various health experts have sounded warnings about imminent avian flu pandemics. News stories on such statements have bolstered religious myths, with people claiming that they were fulfillments of prophetic proclamations as shown in the example below.

XSL•FO
RenderX

Remember that former public health official and what he said in this video. Remember when he mentioned that the next likely pandemic could be coming from bird flu?! He didn't confirm that it will be for sure, but let's keep an eye on this. We might see a biblical prophecy about wars and famines; these are the signs of the end times. [#Matthew 24:7 KJV.]

Discussion

Principal Findings

The findings from this analysis of misinformation on 2 major social media platforms raise implications for public health policy and practice on the evolving H5N1 situation. Some evidence suggests that efforts to counter misinformation, such as debunking or fact-checking, may unintentionally amplify misinformation by elevating its perceived significance [32]. As the sample of posts analyzed represents a small portion of the initial dataset, our study does not aim to show prevalence on various platforms but to describe misinformation narratives and their types. The qualitative analysis of these narratives can demonstrate the construction and development of such narratives on this topic.

Importantly, findings suggest the need for the public health community to carefully consider such misinformation. Evidence shows that misinformation on social media has a differential diffusion pattern compared to traditional media [23]. Moreover, studies show that misinformation garners more engagement than facts. As Vosoughi et al [28] found, "falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information." The top 1% of the most popular false news was disseminated among 1000 to 100,000 people. Verifiable information, however, barely reached more than 1000 people. Misinformation had a 70% higher chance of being retweeted compared to facts. By contrast, the truth took about 6 times as long as false information to reach 1500 people. This more rapid and far-reaching diffusion pattern of misinformation is also reported by studies of medical large language models, which have been found to be vulnerable to "data poisoning" attacks [55]. Large language models, trained on web-scraped data, are increasingly being explored to support health care [56]. However, as Alber et al [47] wrote, "replacement of just 0.001% of training tokens with medical misinformation results in harmful models more likely to propagate medical errors". Thus, a deeper understanding is needed over time of their diffusion pattern and potential role in the spread of misinformation.

Our findings provide a baseline from which ongoing monitoring for trends in patterns of spread can be undertaken and lay the foundation for future work exploring misinformation on the HPAI virus in online spaces. For example, findings in this study about the use of scientific terminology in anti-vaccination posts showed similar points to the investigation by Faasse et al [57] of the dynamics of language use in pro- and anti-vaccination commentary on Facebook. They found that messages against vaccination used linguistic markers that resembled scientific language. Under the elections/democracy topic, posts analyzed reflected patterns noted in other studies regarding the belief in multiple conspiracy theories at the same time [36]. For instance, respondents in Miller's study asserted that COVID-19 was a bioweapon manufactured by China, and the United States accidentally released it into society. In the study by Hellsten et al [58] on X communication networks during avian influenza outbreaks in the Netherlands from 2015 to 2017, they found frequent instances of concurring themes such as the rejection of health regulations #kippendans (chicken dance), #vuurwerk (firework), and #jacht (hunting). Thus, a principal recommendation is that public health systems need to use early warning data of social and behavioral factors, including social media monitoring of misinformation, to inform the development of public health messaging.

Second, our systematic analysis of the content of the posts deemed misinformation suggests that claims about H5N1 are an extension of dominant narratives, by topics and targets, put forth during the COVID-19 pandemic and other issue areas such as climate change. This could be seen, for example, in patterns of blaming a foreign entity or minority communities, such as blaming the Chinese government for H5N1, for a major disease outbreak, which has occurred regularly in history. Examples found included the massacre of Jews during the Black Death of the 15th century and during The Great Plague and the Spanish Flu of 1918‑1919; the lesbian, gay, bisexual, transgender/transsexual, queer, and other minority sexual orientations and gender identities community during the global AIDS pandemic; and Asian people during the COVID-19 pandemic [59].

An overarching theme as well was mistrust of the motivations for public health messaging and interventions. Our findings suggest that views sustaining mistrust persist, framed around a broader narrative of control by powerful elites to achieve a new world order favorable to their interests [60]. H5N1 has become subsumed within this world view, with public health concerns about the virus attributed to self-interested "pandemic engineering." [61] Misinformation about the food system also demonstrated patterns of such mistrust in authorities following measures to control H5N1 infections such as bird depopulation. Bird depopulation can appear drastic because it involves both healthy and infected birds within a given perimeter. It has, however, shown to be the most effective method for stemming infection spread [8,62,63]. Misinformation against this measure typically has ominous labels such as murder and mass slaughtering. Additionally, while health authorities and governments have announced that H5N1 does not currently present food safety risks if animal products are appropriately processed or cooked [55,64], some misinformative posts called the safety of dairy products into question.

Another pattern of mistrust found was under religious allusions and prophecies topics, as users connected prominent public health communications, such as one made by the WHO Director-General [65], with religious narratives seen as the realization of prophetic predictions. This topic, in general, offers insights among certain religious groups that may resist mainstream information. Christian beliefs and those of other religious and insular communities are primarily grounded in faith rather than empirical evidence [66]. Such groups, particularly the fundamentalist ones, reject mainstream and

XSL•FO
RenderX

external information, reinforcing their worldviews and possibly a greater risk of accepting misinformation [67-69]. Therefore, the credibility of such religious declarations relies on people's acceptance of the parallels with Bible verses, which could reduce the need for expositions based on objective reality. Douglas further contends that the growing relationship between misinformation and religion is due in part to religious communities being principal targets [70]. This relates to the historical foundations of misinformation that was facilitated by Christian fundamentalists seeking alternative news sources as they rejected mainstream media's inclination toward what they deem as biased expert knowledge, which usually deviates from religious dogma [6].

Given this broader framing, rebuilding and protecting public trust in the state, market, and civil society may require efforts beyond the public health community. Instead, strategies to sustain public trust in health practitioners and scientists, which remains high in most countries [71], as sources of quality information can be supported. For example, "localizing" public health efforts by working in meaningful partnership with community leaders can "humanize" and thus increase trust in public health messaging [72].

Third, our findings demonstrate the importance of protecting information quality as the H5N1 situation evolves while balancing the need to not draw undue attention to misinformation. Unlike COVID-19, the public health community has been managing the risks from H5N1 since the late 1990s, so much is already known about the virus [73]. Since 2022, the spread to a growing number of mammal species including dairy farms has increased media attention and public awareness [2]. Evidence suggests that there is thus an opportunity for public health messaging to facilitate public understanding of this issue while it is still emerging. For example, Vos and Buckner [74] found that sensemaking plays a key role in the success of public health interventions during health crises. Sensemaking involves the way the public interprets and frames health information they receive about emerging health events. The misinformation identified in this paper suggests H5N1 is already interpreted using existing conspiracy-related worldviews (eg, New World Order). To counter misinformed sensemaking, there is a need to avoid information vacuums. Moreover, public health messaging must not only be factually accurate but in a form that facilitates sensemaking. Accordingly, there is a need to provide appropriate information for the public to understand, and not just be informed, about H5N1. The insights into online health misinformation on HPAI that this study provides are instrumental for supplementing surveillance structures, guiding communication mechanisms with an understanding of the misgivings by some, which could be fueling hesitation toward and rejection of interventions and ultimately bolstering pandemic readiness measures.

Importantly, the ability to monitor the content and volume of misinformation on social media, and proactively message about the evolving H5N1 situation, is likely to be more important yet more challenging amid changing practices by social media platforms to limit access to data and to protect information quality. One of the strengths of this study is the novel expansion of online health misinformation analysis beyond the textual content of X platform. Our multimodal analysis of textual and multimedia data from Instagram and Facebook broadens the scope of knowledge of online public narratives, particularly about infectious diseases. Despite Meta's claims about the strict moderation of posts,75 the majority of posts identified as misinformation were not flagged by both Facebook and Instagram for fact-checking. There were only 5 posts in our dataset that were flagged as inaccurate (less than 1%). The generic notification stated they were not completely factual, and they contained information similar to other posts that third-party fact checkers had reviewed. In 2018, CEO Mark Zuckerberg stated that Meta's research showed, "that no matter where we draw the lines for what is allowed, as a piece of content gets close to that line (demarcating prohibited content which contains misinformation and harmful material), people will engage with it more on average—even when they tell us afterwards they don't like the content" [76]. The announcement in January 2025 that Meta would end third-party fact-checking, in favor of allowing people to "express themselves freely," could have implications for protecting information quality [77]. Given the ubiquity of social media and its centrality in sustaining social connection, especially during times of uncertainty such as public health emergencies, the public health community can benefit from understanding and navigating this evolving landscape as a core component of future pandemic preparedness and response.

Limitations

A key limitation was the inability to include an analysis of the volume of misinformation on Meta platforms due to the lack of data collection tools. Moreover, we were unable to analyze the ideological leanings of users to further contextualize the narratives. We conducted a brief review of the profiles of users in our dataset to explore their potential ideological or political affiliations. Though some users' profiles suggested specific political ideologies, this analysis was not included in the study, as confirming such affiliations would require an ethnographic approach, which falls outside the scope of this paper.

Additionally, our paper was limited to English-Language and region-specific search terms that mostly covered North America, along with a focus on some US public figures such as Anthony Fauci. This could have excluded similar discussions from other regions and countries such as the United Kingdom and Germany, where different terms were used. Future research could address this by including additional relevant keywords from different regions such as "flockdown" and "raw milk." Another limitation was the prioritization of pandemic and vaccine-related terms in our dataset, which meant that other misinformation narratives about transmission and treatment, for example, may not have been captured. This limitation should be considered when examining our findings, as the topic categories identified were a result of our keyword search focus. Thus, findings reflect a specific focus of H5N1 misinformation.

Conclusions

Misinformation on HPAI on social media has implications for public health as recurrent avian influenza outbreaks in animal populations have contributed to fears about the pandemic potential of the disease for humans. Notably, H5N1 in human

XSL•FO
RenderX

populations is associated with several clinical symptoms; thus, decisive and swift actions are needed to stem any further infections. This finding underscores the importance of reviewing contextual factors like online misinformation, which could interfere with health interventions.

Our online investigation revealed that the targets of H5N1 misinformation narratives were largely people or entities who wield varying levels of political or other authority such as government and public health officials. These individuals received attention for their varying roles in health-related decision-making and in facilitating public health interventions.

We also found that the most prevalent misinformation themes were accusations of pandemic engineering, societal destruction, anti-vaccination sentiments, impact on elections or democracy, causing food insecurity, and religious allusions and prophecies. These assertions took various forms, including accusations that the virus itself did not occur naturally but claimed that the virus was artificially created as a bioweapon. The messages that contained anti-vaccination sentiments concluded that health interventions for avian influenza would include vaccines that would insert digital biometric identification into people, which could be used to either instigate population control or enforce a digital economy.

Acknowledgments

The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript.

Funding

This work was supported by the Bridge Research Consortium (BRC), part of Canada's Immuno-Engineering and Biomanufacturing Hub. The BRC is funded by the Canada Biomedical Research Fund and Biomedical Research Infrastructure Fund (grant: CBRF-2023-00122) and BC Knowledge Development Fund.

Data Availability

The dataset used in this manuscript can be shared with any interested scholar.

Conflicts of Interest

None declared.

References

1. Banks J, Speidel EC, Alexander DJ. Characterisation of an avian influenza A virus isolated from a human--is an intermediate host necessary for the emergence of pandemic influenza viruses? Arch Virol 1998;143(4):781-787. [doi: 10.1007/s007050050329] [Medline: 9638147]
2. Pereira HG, Tůmová B, Law VG. Avian influenza A viruses. Bull World Health Organ 1965;32(6):855-860. [Medline: 5294310]
3. Snowden FM. Epidemics and Society: From the Black Death to the Present: Yale University Press; 2019. [doi: 10.12987/9780300249149]
4. Avian influenza a(H5N1) virus. World Health Organization. 2024. URL: https://www.who.int/teams/global-influenza-programme/avian-influenza/avian-a-h5n1-virus [accessed 2026-02-20]
5. 2020-2024 highlights in the history of avian influenza (bird flu) timeline. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/bird-flu/avian-timeline/2020s.html [accessed 2026-02-20]
6. Ridgely SB. Conservative Christianity and the creation of alternative news: an analysis of Focus on the Family's multimedia empire. Relig Am Cult 2020;30(1):1-25. [doi: 10.1017/rac.2020.1]
7. Rust S. USDA says it will start testing milk for H5N1 bird flu virus nationwide. Los Angeles Times. 2024 Dec 6. URL: https://www.latimes.com/environment/story/2024-12-06/usda-says-it-will-start-testing-milk-for-h5n1-nationwide [accessed 2026-02-20]
8. Minister bibeau meets with stakeholders from the agricultural sector to discuss the impacts of highly pathogenic avian influenza. Canada.ca. 2022 Apr 29. URL: https://www.canada.ca/en/food-inspection-agency/news/2022/04/readout-from-minister-bibeau-minister-of-agriculture-and-agri-food-on-the-impacts-of-highly-pathogenic-avian-influenza.html [accessed 2026-02-20]
9. Tian H, Xu B. Persistence and transmission of avian influenza A (H5N1): virus movement, risk factors and pandemic potential. Ann GIS 2015 Jan 2;21(1):55-68. [doi: 10.1080/19475683.2014.992368]
10. Clark EC, Neumann S, Hopkins S, Kostopoulos A, Hagerman L, Dobbins M. Changes to public health surveillance methods due to the COVID-19 pandemic: scoping review. JMIR Public Health Surveill 2024 Jan 19;10:e49185. [doi: 10.2196/49185] [Medline: 38241067]

XSL•FO
RenderX

11. Lewandowsky S. The 'post-truth' world, misinformation, and information literacy: a perspective from cognitive science. In: Informed Societies why Information Literacy Matters for Citizenship, Participation and Democracy: Facet Publishing; 2020:69-87. [doi: 10.29085/9781783303922]
12. Shahbazi M, Bunker D. Social media trust: fighting misinformation in the time of crisis. Int J Inf Manage 2024 Aug;77:102780. [doi: 10.1016/j.ijinfomgt.2024.102780]
13. KFF tracking poll on health information and trust: January 2025. KFF. 2025 Jan 28. URL: https://www.kff.org/health-information-trust/kff-tracking-poll-on-health-information-and-trust-january-2025/ [accessed 2026-02-20]
14. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021 Jan 20;23(1):e17187. [doi: 10.2196/17187] [Medline: 33470931]
15. Calma J. Twitter just closed the book on academic research. The Verge. 2023 May 31. URL: https://www.theverge.com/2023/5/31/23739084/twitter-elon-musk-api-policy-chilling-academic-research [accessed 2026-02-20]
16. Ortutay B. Meta kills off misinformation tracking tool crowdtangle despite pleas from researchers, journalists. Associated Press News. 2024 Aug 14. URL: https://apnews.com/article/meta-crowdtangle-research-misinformation-shutdown-facebook-977ece074b99adddb4887bf719f2112a [accessed 2026-02-20]
17. Liu J, Kassas B, Lai J, Kropp J, Gao Z. Understanding the role of risk preferences and perceptions in vaccination decisions and post-vaccination behaviors among U.S. households. Sci Rep 2024;14(1):3190. [doi: 10.1038/s41598-024-52408-6]
18. Westerman D, Spence PR, Van Der Heide B. Social media as information source: recency of updates and credibility of information. J Comput-Mediat Comm 2014 Jan;19(2):171-183. [doi: 10.1111/jcc4.12041]
19. Yang Y, Zhang Y, Xiang A. Information interaction and social support: exploring help-seeking in online communities during public health emergencies. BMC Public Health 2023 Jun 27;23(1):1250. [doi: 10.1186/s12889-023-16151-3] [Medline: 37370074]
20. Disease outbreak news (dons). World Health Organization. URL: https://www.who.int/emergencies/disease-outbreak-news [accessed 2026-02-20]
21. Al-Rawi A. Gatekeeping fake news discourses on mainstream media versus social media. Soc Sci Comput Rev 2019 Dec;37(6):687-704. [doi: 10.1177/0894439318795849]
22. A multi-dimensional approach to disinformation: report of the independent high level group on fake news and online disinformation (3). : European Commission; 2018 URL: https://www.ami.info/wp-content/uploads/2018/03/HLEGReportonFakeNewsandOnlineDisinformation.pdf [accessed 2026-02-20]
23. Chen S, Xiao L, Kumar A. Spread of misinformation on social media: what contributes to it and how to combat it. Comput Human Behav 2023 Apr;141:107643. [doi: 10.1016/j.chb.2022.107643]
24. Lewandowsky S, Cook J, Schmid P, et al. The COVID-19 Vaccine Communication Handbook: A Practical Guide for Improving Vaccine Communication and Fighting Misinformation: SciBeh; 2021. URL: https://www.immunize.ca/sites/default/files/Resource%20and%20Product%20Uploads%20(PDFs)/COVID-19/final-COVID19-vaccinehandbook-4-production.pdf [accessed 2026-02-20]
25. Song J, Song TM, Seo DC, Jin DL, Kim JS. Social big data analysis of information spread and perceived infection risk during the 2015 Middle East respiratory syndrome outbreak in South Korea. Cyberpsychol Behav Soc Netw 2017 Jan;20(1):22-29. [doi: 10.1089/cyber.2016.0126] [Medline: 28051336]
26. Gibson K. Meta to end fact-checking, replacing it with community-driven system akin to elon musk's x. CBS News. 2025 Jan 8. URL: https://www.cbsnews.com/news/meta-facebook-instagram-fact-checking-mark-zuckerberg/ [accessed 2026-02-20]
27. Broniatowski DA, Jamison AM, Qi S, et al. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. Am J Public Health 2018 Oct;108(10):1378-1384. [doi: 10.2105/AJPH.2018.304567] [Medline: 30138075]
28. Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science 2018 Mar 9;359(6380):1146-1151. [doi: 10.1126/science.aap9559] [Medline: 29590045]
29. Wang Y, McKee M, Torbica A, Stuckler D. Systematic literature review on the spread of health-related misinformation on social media. Soc Sci Med 2019 Nov;240:112552. [doi: 10.1016/j.socscimed.2019.112552] [Medline: 31561111]
30. Adebimpe WO, Adeyemi DH, Faremi A, Ojo JO, Efuntoye AE. The relevance of the social networking media in Ebola virus disease prevention and control in Southwestern Nigeria. Pan Afr Med J 2015;22(Supp 1):7. [doi: 10.11604/pamj.supp.2015.22.1.6165]
31. Smith TO. A little birdie told me: H1N1 information and misinformation exchange on twitter. Social Science Research Network. 2010. URL: https://papers.ssrn.com/abstract=1739306 [accessed 2026-02-20]
32. Chowdhury N, Khalid A, Turin TC. Understanding misinformation infodemic during public health emergencies due to large-scale disease outbreaks: a rapid review. Z Gesundh Wiss 2023 Apr;31(4):553-573. [doi: 10.1007/s10389-021-01565-3]
33. Yan Q, Shan S, Zhang B, et al. Monitoring the relationship between social network status and influenza based on social media data. Disaster Med Public Health Prep 2023 Sep 18;17:e490. [doi: 10.1017/dmp.2023.117] [Medline: 37721020]
34. Gori D, Reno C, Remondini D, Durazzi F, Fantini MP. Are we ready for the arrival of the new COVID-19 vaccinations? Great promises and unknown challenges still to come. Vaccines (Basel) 2021 Feb 18;9(2):173. [doi: 10.3390/vaccines9020173] [Medline: 33670697]

XSL•FO
RenderX

35. Enders AM, Uscinski JE, Seelig MI, et al. The relationship between social media use and beliefs in conspiracy theories and misinformation. Polit Behav 2023;45(2):781-804. [doi: 10.1007/s11109-021-09734-6] [Medline: 34248238]
36. Miller JM. Do COVID-19 conspiracy theory beliefs form a monological belief system? Can J Pol Sci 2020 Jun;53(2):319-326. [doi: 10.1017/S0008423920000517]
37. Motta M, Stecula D, Farhart C. How right-leaning media coverage of COVID-19 facilitated the spread of misinformation in the early stages of the pandemic in the US. Can J Pol Sci 2020 Jun;53(2):335-342. [doi: 10.1017/S0008423920000396]
38. Choi DH, Yoo W, Noh GY, Park K. The impact of social media on risk perceptions during the MERS outbreak in South Korea. Comput Human Behav 2017 Jul;72:422-431. [doi: 10.1016/j.chb.2017.03.004] [Medline: 32288176]
39. Veil S, Reynolds B, Sellnow TL, Seeger MW. CERC as a theoretical framework for research and practice. Health Promot Pract 2008 Oct;9(4 Suppl):26S-34S. [doi: 10.1177/1524839908322113] [Medline: 18936257]
40. Ahmed W, Bath PA, Sbaffi L, Demartini G. Novel insights into views towards H1N1 during the 2009 Pandemic: a thematic analysis of Twitter data. Health Info Libr J 2019 Mar;36(1):60-72. [doi: 10.1111/hir.12247] [Medline: 30663232]
41. Qin Z, Ronchieri E. Exploring pandemics events on Twitter by using sentiment analysis and topic modelling. Appl Sci (Basel) 2022;12(23):11924. [doi: 10.3390/app122311924]
42. Soltani M, Sharif S, Dara R. Analyzing social media discourse of avian influenza outbreaks. Nat Lang Process J Sep 2025 Sep;12(1):100176. [doi: 10.1016/j.nlp.2025.100176]
43. Brainard J, Hunter PR. Misinformation making a disease outbreak worse: outcomes compared for influenza, monkeypox, and norovirus. Simulation 2019 Apr;96(4):365-374. [doi: 10.1177/0037549719885021] [Medline: 34285423]
44. Fault lines: the expert panel on the socioeconomic impacts of science and health misinformation. : Council of Canadian Academies; 2023 URL: https://cca-reports.ca/wp-content/uploads/2023/02/Report-Fault-Lines-digital.pdf [accessed 2026-02-20]
45. Online discourse about bird flu spikes in the US and globally. Public Health Communications Collaborative. 2024 Apr 30. URL: https://publichealthcollaborative.org/alerts/online-discourse-about-bird-flu-spikes-in-the-u-s-and-globally/ [accessed 2026-02-20]
46. Dixon SJ. Most popular social networks worldwide as of february 2025, ranked by number of monthly active users (in millions). Statista. 2024 Jul 10. URL: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ [accessed 2026-02-20]
47. Alber DA, Yang Z, Alyakin A, et al. Medical large language models are vulnerable to data-poisoning attacks. Nat Med 2025 Feb;31(2):618-626. [doi: 10.1038/s41591-024-03445-1] [Medline: 39779928]
48. Al-Rawi A. Script for extracting most frequent words, phrases, and sentiments. GitHub. 2025 Sep 23. URL: https://gist.github.com/ahmedalrawi/c5879cfc52342d3ed778c35c755a42ad [accessed 2026-02-20]
49. Al-Rawi A, Jamieson K. Characterizing the gendered Twitter discussion of COVID-19 hoax. Health Commun 2023 Dec;38(14):3366-3375. [doi: 10.1080/10410236.2022.2149112] [Medline: 36411526]
50. Lanier HD, Diaz MI, Saleh SN, Lehmann CU, Medford RJ. Analyzing COVID-19 disinformation on Twitter using the hashtags #scamdemic and #plandemic: Retrospective study. PLOS ONE 2022;17(6):e0268409. [doi: 10.1371/journal.pone.0268409] [Medline: 35731785]
51. Current situation: bird flu in dairy cows. Centers for Disease Control and Prevention. 2025 Jul 7. URL: https://www.cdc.gov/bird-flu/situation-summary/mammals.html [accessed 2026-02-20]
52. Thomas DR. A general inductive approach for analyzing qualitative evaluation data. Am J Eval 2006 Jun;27(2):237-246. [doi: 10.1177/1098214005283748]
53. Krippendorff K. Reliability in content analysis: some common misconceptions and recommendations. Hum Commun Res 2004 Jul 1;30(3):411-433. [doi: 10.1111/j.1468-2958.2004.tb00738.x]
54. Albu OB, Etter M. Hypertextuality and social media: a study of the constitutive and paradoxical implications of organizational Twitter use. Manag Commun Q 2016;30(1):5-31. [doi: 10.1177/0893318915601161]
55. Avian influenza. European Food Safety Authority. URL: https://www.efsa.europa.eu/en/topics/topic/avian-influenza [accessed 2026-02-20]
56. Nazi ZA, Peng W. Large language models in healthcare and medical domain: a review. Informatics (MDPI) 2024;11(3):57. [doi: 10.3390/informatics11030057]
57. Faasse K, Chatman CJ, Martin LR. A comparison of language use in pro- and anti-vaccination comments in response to a high profile Facebook post. Vaccine 2016 Nov 11;34(47):5808-5814. [doi: 10.1016/j.vaccine.2016.09.029] [Medline: 27707558]
58. Hellsten I, Jacobs S, Wonneberger A. Active and passive stakeholders in issue arenas: a communication network approach to the bird flu debate on Twitter. Public Relat Rev 2019 Mar;45(1):35-48. [doi: 10.1016/j.pubrev.2018.12.009]
59. Davis M. The Monster Enters: COVID-19, Avian Flu, and the Plagues of Capitalism: Verso Books; 2022. URL: https://www.versobooks.com/en-gb/products/2753-the-monster-enters?srsltid=AfmBOopdZFr6YRDmmIZ5n5n_l0dqSjtb1xi6kdqQGA4CpnINQ5y6bNKV [accessed 2026-02-20]
60. Gallagher A, O'Connor C. The 'great reset'. Institute for Strategic Dialogue. 2023 Mar 14. URL: https://www.isdglobal.org/isd-explainer/the-great-reset/ [accessed 2026-02-20]

XSL•FO
RenderX

61. Kilander G. Former director of the CDC predicts the next pandemic will be from bird flu. The Independent. 2024 Jun 17. URL: https://www.independent.co.uk/news/world/americas/bird-flu-cdc-covid-pandemic- b2563711.html [accessed 2026-02-20]
62. Backer JA, van Roermund HJW, Fischer EAJ, van Asseldonk M, Bergevoet RHM. Controlling highly pathogenic avian influenza outbreaks: an epidemiological and economic model analysis. Prev Vet Med 2015 Sep 1;121(1-2):142-150. [doi: 10.1016/j.prevetmed.2015.06.006] [Medline: 26087887]
63. Martcheva M. Avian flu: modeling and implications for control. J Biol Syst 2014 Mar;22(1):151-175. [doi: 10.1142/S0218339014500090]
64. Status of ongoing avian influenza response by province. Canadian Food Inspection Agency. 2024 Jun 26. URL: http://inspection.canada.ca/en/animal-health/terrestrial-animals/diseases/reportable/avian-influenza/latest-bird-flu-situation/status-ongoing-avian-influenza-response [accessed 2026-02-20]
65. Disinformation and public health. World Health Organization. 2024 Feb 6. URL: https://www.who.int/news-room/questions-and-answers/item/disinformation-and-public-health [accessed 2026-02-20]
66. Nieminen P, Loikkanen J, Ryökäs E, Mustonen AM. Nature of evidence in religion and natural science. Theol Sci 2020 Jul 2;18(3):448-474. [doi: 10.1080/14746700.2020.1786221]
67. Chauchard S, Badrinathan S. The Religious Roots of Belief in Misinformation: Experimental Evidence from India: Cambridge University Press; 2025. [doi: 10.1017/S0007123425100616]
68. Jedinger A, Siegers P. Religion, Spirituality, and Susceptibility to Conspiracy Theories: Examining the Role of Analytic Thinking and Post-Critical Beliefs: Cambridge University Press; 2024. [doi: 10.1017/S1755048324000130]
69. Łowicki P, Marchlewska M, Molenda Z, Karakula A, Szczepańska D. Does religion predict coronavirus conspiracy beliefs? Centrality of religiosity, religious fundamentalism, and COVID-19 conspiracy beliefs. Pers Individ Dif 2022 Mar;187:111413. [doi: 10.1016/j.paid.2021.111413] [Medline: 34866726]
70. Douglas C. Religion and fake news: faith-based alternative information ecosystems in the US and Europe. Rev Faith Int. Aff 2018 Jan 2;16(1):61-73. [doi: 10.1080/15570274.2018.1433522]
71. 2025 edelman trust barometer special report: trust and health. : Edelman Trust Institute; 2025 URL: https://www.edelman.com/sites/g/files/aatuss191/files/2025-04/2025%20Edelman%20Trust%20Barometer%20Trust%20and%20Health%20FINAL.pdf [accessed 2026-02-20]
72. Sundelson AE, Jamison AM, Huhn N, Pasquino SL, Sell TK. Fighting the infodemic: the 4 i framework for advancing communication and trust. BMC Public Health 2023 Aug 30;23(1):1662. [doi: 10.1186/s12889-023-16612-9] [Medline: 37644563]
73. Alexander DJ. A review of avian influenza in different bird species. Vet Microbiol 2006;74(1–2):3-13. [doi: 10.1016/S0378-1135(00)00160-7]
74. Vos SC, Buckner MM. Social media messages in an emerging health crisis: tweeting bird flu. J Health Commun 2016;21(3):301-308. [doi: 10.1080/10810730.2015.1064495] [Medline: 26192209]
75. Misinformation. Meta. 2025. URL: https://transparency.fb.com/policies/community-standards/misinformation [accessed 2026-02-20]
76. Zuckerberg M. A blueprint for content governance and enforcement. Facebook.com. 2018. URL: https://opencasebook.org/casebooks/270-governing-digital-technology-spring-2020/resources/4.1.1.1-a-blueprint-for-content-governance-and-enforcement-by-mark-zuckerberg-facebook-2018/ [accessed 2026-02-20]
77. Kaplan J. More speech, fewer mistakes. Meta. 2025 Jan 7. URL: https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/ [accessed 2026-02-20]

Abbreviations

CDC: Centers for Disease Control and Prevention
HPAI: highly pathogenic avian influenza
WHO: World Health Organization

Edited by A Mavragani; submitted 28.Feb.2025; peer-reviewed by E Lau, R Tarlinton; revised version received 06.Jan.2026; accepted 28.Jan.2026; published 09.Apr.2026.

Please cite as:
Al-Rawi A, Ackah B, Fakida A, Lee K
Misinformation in Social Media Narratives on Highly Pathogenic Avian Influenza: Systematic Content Analysis of Facebook and Instagram Posts
J Med Internet Res 2026;28:e73275
URL: https://www.jmir.org/2026/1/e73275
doi:10.2196/73275

XSL•FO
RenderX

© Ahmed Al-Rawi, Betty Ackah, Abdelrahman Fakida, Kelley Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patient Perceptions of Ozempic (Semaglutide) for Weight Loss: Mixed Methods Analysis of Online Medication Reviews

Abanoub J Armanious[1,2], BS; Rachel-Mae Hunter[1,2], BS; Kristi R Griffiths[3], PhD; Hannah E Bowrey[4,5], PhD; Robyn M Brown[6], PhD; Morgan H James[1,2,4,5], PhD

[1]Department of Psychiatry, Rutgers Robert Wood Johnson Medical School, Piscataway, NJ, United States

[2]Rutgers Addiction Research Center, Brain Health Institute, Rutgers, The State University of New Jersey, Piscataway, NJ, United States

[3]InsideOut Institute, Faculty of Medicine and Health, The University of Sydney and Sydney Local Health District, Sydney, Australia

[4]School of Psychology, Faculty of Science, The University of Sydney, Camperdown, Australia

[5]Brain and Mind Centre, The University of Sydney, Camperdown, Australia

[6]Department of Biochemistry and Pharmacology, The University of Melbourne, Parkville, Australia

Corresponding Author:
Morgan H James, PhD
School of Psychology, Faculty of Science
The University of Sydney
Camperdown
Australia
Phone: 61 293510774
Email: morgan.james@sydney.edu.au

Abstract

Background: Ozempic (semaglutide) has received widespread attention for its appetite-suppressing effects, leading to extensive off-label use for weight loss. Although gastrointestinal side effects are well documented, less is known about how users assess the trade-off between perceived benefits and adverse effects, or how these assessments influence treatment discontinuation. Importantly, existing insights are often limited to clinical trial populations and may not fully reflect real-world experiences.

Objective: This study applies a novel infoveillance approach to examine patient-reported experiences with off-label Ozempic use for weight loss and to identify the factors most strongly associated with user satisfaction and treatment discontinuation.

Methods: We analyzed 60 publicly available, self-selected, anonymous user reviews of Ozempic from Drugs.com. Reviews were initially examined using thematic analysis to identify key themes describing patients' lived experiences with treatment. These qualitative themes were then linked to user-provided ratings of perceived drug efficacy (1-10 scale) and statements regarding intent to continue or discontinue treatment. This mixed methods approach enabled the integration of qualitative depth with quantitative patterns within naturally occurring, deidentified online data.

Results: Three major themes emerged from the thematic analysis: (1) change in body weight and appetite, (2) nonweight-related symptoms and side effects, and (3) plans for ongoing use versus discontinuation. Two-thirds of respondents reported reduced appetite, food cravings, or body weight. Gastrointestinal complaints were common (reported by 37 of 60, 62%, reviewers) but did not significantly (P=.39) influence satisfaction ratings or decisions to continue treatment. Instead, minimal/no weight loss and the emergence of nongastrointestinal side effects were more frequently associated with low overall satisfaction and discontinuation. Effective weight loss, even when accompanied by gastrointestinal side effects, was associated with a greater willingness to continue Ozempic treatment.

Conclusions: This study presents a novel application of infoveillance methods to characterize real-world patient attitudes toward off-label Ozempic use. Satisfaction was driven primarily by perceived effectiveness rather than tolerability. Key limitations are the self-selected nature of the sample, reliance on anonymous, self-reported data, and the lack of demographic, dosing, or treatment-duration information. Nonetheless, these findings underscore the value of online health forums as a rich and underutilized source of patient-centered insights to inform obesity treatment strategies, adherence interventions, and public health communication.

(J Med Internet Res 2026;28:e78391) doi:10.2196/78391

KEYWORDS

Thematic analysis; discussion board; pharmacotherapy; glucagon like peptide 1; GLP-1; incretin mimetic; craving.

XSL•FO
RenderX

Introduction

Background

The prevalence of obesity has more than doubled since 1990, contributing to a rise in chronic diseases associated with higher body weight, including type 2 diabetes (T2D) and cardiovascular diseases [1]. Consequently, the need for effective interventions to address obesity is urgent. Although lifestyle interventions (diet and exercise) are considered first-line strategies for weight management, they are often ineffective in the long term [2,3]. The few pharmacotherapies approved for treating overweight and obesity have historically produced only modest results, typically achieving 5%-10% weight loss [4]. Moreover, although bariatric surgery is effective for many patients, it carries significant morbidity and mortality risks [5,6]. Accordingly, the advent of glucagon-like peptide-1 receptor (GLP-1R) agonists, which are highly effective in promoting substantial weight loss and improving related health outcomes, has dramatically redefined the treatment landscape for these conditions.

Semaglutide, an incretin mimetic, is a GLP-1R agonist that stabilizes blood glucose by stimulating insulin secretion and inhibiting glucagon production [7]. It delays gastric emptying and influences appetite-regulating neural pathways, increasing satiety and reducing food intake in some individuals [8]. The first FDA-approved formulation of semaglutide was Ozempic, a once-weekly subcutaneous injection for the management of T2D [9]. This regimen offered the convenience of less frequent dosing compared with the previously approved incretin mimetic liraglutide, which required daily injections [7,10,11]. A substantial body of clinical trial data now indicates that, compared with placebo, Ozempic significantly reduces body weight (7.9%-17.3%) [12-16], lowers HbA_{1c} (glycated hemoglobin) levels, waist circumference, and systolic blood pressure, and improves overall physical functioning [13,14,16]. In clinical trials, improvements in these outcomes are generally observed within 3 months of initiating treatment [14,16-18]. In light of these findings, there has been intense interest in the off-label use of Ozempic for cosmetic weight loss, fueled in part by its popularization in mainstream and social media [19], despite other semaglutide formulations, such as Wegovy, being specifically approved for the treatment of obesity [20].

Despite their efficacy in promoting weight loss, semaglutide treatments are associated with several adverse events that vary in severity. Gastrointestinal complaints are the most common, with prevalence ranging from 41.9% to 82.8% (more common at higher doses), and include symptoms such as nausea, vomiting, constipation, and diarrhea [12-16]. Less frequent adverse events include headache, allergic reactions, and gallbladder-related disorders [15]. Adverse events are a major contributor to nonadherence to weight loss medications, including GLP-1R agonists [21-23]. In the SUSTAIN-6 clinical trial of injectable semaglutide, 22.6% of patients discontinued treatment prematurely during the 24-month study period. However, “real-world” discontinuation rates appear to be much higher. One study reported a 12-month discontinuation rate of 33% following initiation of once-weekly Ozempic therapy [24], while another study, which did not differentiate between different forms of injectable GLP-1R agonists, reported that 70.1% of patients discontinued treatment within 24 months [25].

Thus, among individuals using Ozempic for weight loss, there may be a conflict between experiencing the desired effects of treatment and managing its associated side effects. Understanding how patients navigate this trade-off is essential for optimizing adherence and maximizing therapeutic outcomes.

Aims

Here, we employed a mixed methods infoveillance approach to examine attitudes toward Ozempic among individuals with lived experience of off-label use. Specifically, we conducted a thematic analysis of user-generated reviews posted on Drugs.com [26], followed by quantitative modeling to assess how emergent themes were associated with perceived efficacy and treatment discontinuation. A major advantage of this approach is that it is not constrained by the a priori hypotheses typical of traditional quantitative designs, allowing for the emergence of unsolicited insights that might otherwise be overlooked [27]. As a form of infodemiological research, this study illustrates how publicly available, user-generated data can serve as a powerful resource for capturing patient-centered perspectives on medication effectiveness, tolerability, and real-world barriers to adherence. Despite the inherent limitations of using self-selected, anonymous online data, these findings provide unique and timely insight into how individuals evaluate the benefits and drawbacks of off-label Ozempic use for weight loss.

Methods

Data Collection

Data consisted of reviews of Ozempic submitted to Drugs.com, a website that provides peer-reviewed and independent information on more than 24,000 prescription drugs, over-the-counter medicines, and natural products. A unique feature of Drugs.com is its platform that allows members of the public to submit open-ended reviews and quantitative ratings of specific medications, enabling analysis of how these outcomes are related. We extracted data exclusively from respondents who selected “weight loss” as the condition for which they were using Ozempic, despite this not being an approved indication. Notably, Drugs.com has since removed “weight loss” as an option (current options now include T2D, cardiovascular risk reduction, and chronic kidney disease), making these data particularly valuable for capturing experiences of individuals using Ozempic specifically for weight loss. Using a display name, respondents are prompted to “comment on your experience with Ozempic” and are encouraged to “describe how the medication helped (or why it didn’t work); the benefits, adverse events, dosage, ease of use” in a single textbox. No demographic data are collected. Respondents can also provide a quantitative rating of the drug on a scale from 1 (not effective) to 10 (most effective) and indicate the duration of medication use.

Data were downloaded in June 2023. No retrospective time limit on reviews was imposed; the oldest review dated from

XSL•FO
RenderX

February 2023, and the most recent from June 2023. User reviews of Ozempic in which weight loss was listed as the primary indication were extracted, yielding a total of 78 reviews. As described below, a total of 60 reviews were analyzed before reaching thematic saturation.

This study employed a sequential mixed methods design, integrating qualitative and quantitative analyses of user-generated data. In the first phase, an inductive thematic analysis was conducted to identify patterns and themes within users' open-ended narratives describing their experiences with Ozempic. In the second phase, these emergent themes were quantitatively examined in relation to users' numerical satisfaction ratings. This design was selected to provide complementary insights, capturing the contextual richness of lived experience while enabling empirical assessment of the factors most strongly associated with user satisfaction and treatment discontinuation. This approach aligns with our previously published work [27]. We acknowledge that qualitative interpretation is inherently influenced by coder perspectives; measures taken to minimize and reflect on this potential bias are detailed in the "Thematic Data Analysis" section.

Ethical Considerations

This study analyzed secondary, publicly available, deidentified user reviews from Drugs.com. No interaction with users occurred, and no direct or indirect identifiers were collected or stored. Data-minimization procedures were applied: only text necessary for analysis was retained, quotes were screened to exclude potentially identifying details, and results are reported in aggregate wherever possible. In accordance with institutional guidance for research using publicly accessible data, this project was not subject to human participant review.

We acknowledge that, despite deidentification and public availability, online health narratives may still pose residual privacy risks (eg, potential reidentification via unique combinations of details) and may reflect audience expectations. Our use of these data was limited to analytic purposes, with careful curation of verbatim quotations and deliberate avoidance of stigmatizing language.

Thematic Data Analysis

Qualitative data analysis was conducted using NVivo 14 software (Lumivero, LLC). Data were analyzed using a thematic analysis approach, as initially outlined by Braun and Clarke [28], following a procedure we have described previously [27] and similar to those used in studies employing alternative online data sources [29,30]. Briefly, themes were generated through an iterative process of reading through each review, suggesting themes, re-reading, and comparing categories across multiple cycles of analysis (Figure 1). To facilitate this process, the dataset was randomly divided into batches of 15 reviews. Each batch was independently reviewed by 2 coders (AJA and RMH), and excerpts relevant to the research question were coded according to a data-driven, "bottom-up" principle. This approach minimized the influence of any preconceived ideas the reviewers might have had about respondent perceptions of Ozempic. After each set of 15 reviews, both coders met with a third-party noncoder (MHJ) to compare identified codes against the original data and with each other, ensuring that the codes were coherent, consistent, and distinctive. Thematic saturation was reached when 2 consecutive batches of 15 reviews yielded no new codes or subthemes. Saturation was confirmed by consensus among the 2 coders (AJA and RMH) and the independent reviewer (MHJ), consistent with the reflexive and inductive approach to thematic analysis described previously [27,28]. The initial analysis yielded 34 distinct coding categories, which were subsequently grouped and refined into 3 overarching themes. There were no predefined criteria for determining what constituted a separate theme; rather, meaningful clusters of codes were identified, reviewed, and iteratively refined.

It is important to acknowledge that thematic coding may have been influenced by the positions and potential biases of the authors. At the time of coding, AJA (male) was a post-baccalaureate researcher (BSc with concentrations in Cell Biology and Neuroscience, Public Health, and Religion), and RMH (female) was an undergraduate student majoring in Biological Sciences on the predental medicine track. Both were conducting laboratory research on the neurobiological basis of eating disorders. MHJ (male) was a researcher with expertise in the neurobiology of motivated behaviors, including feeding. To minimize potential coder bias, the coding team underwent structured training and employed standardized procedures for codebook development. Coders met regularly to review emerging codes, reconcile discrepancies, and discuss interpretations with the independent reviewer (MHJ) throughout the analytic process.

XSL•FO
RenderX

Figure 1. Flowchart outlining the process adopted to carry out the qualitative analysis portion of the study. First, reviews of Ozempic were extracted from Drugs.com and then randomly batched into groups of 15, with the first batch undergoing thorough familiarization via reading and rereading by the coders. Next, initial coding was conducted using NVivo 14 software, with subsequent validation by a noncoder. Similar analyses were carried out on a second batch of reviews; this process was repeated until the coders and non-coder agree that analysis of an additional batch of 15 reviews was unlikely to result in the identification of additional unique codes (i.e. thematic saturation). Codes were then organized into themes and sub-themes, which underwent iterative review and refinement. Finally, representative quotes were selected to illustrate each of the subthemes.

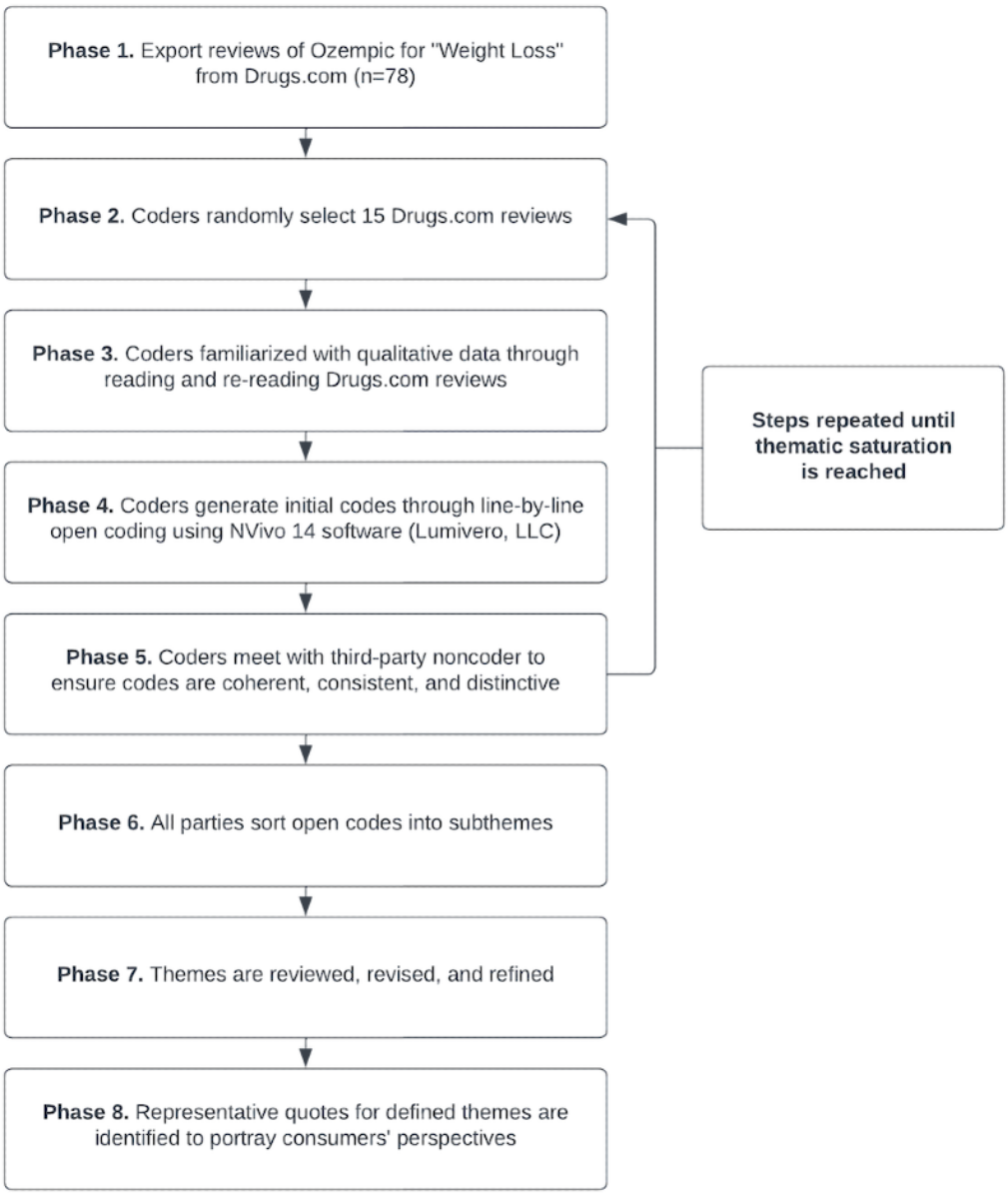

Quantitative Analysis of User Reviews

We also sought to examine how the qualitative themes identified through thematic analysis related to the quantitative rating scores (1-10 scale) provided by respondents. Of the 60 reviews analyzed, 54 included an associated quantitative rating. We calculated the median user rating for participants whose reviews contributed to each subtheme. Based on a frequency histogram of rating scores, the data were divided into 2 groups on either side of the median score (7.5), effectively creating clusters of "higher ratings" (n=27) and "lower ratings" (n=27; see Figure 2A). A median split was used because user ratings were bimodally distributed, with most responses clustering at the extreme values (1 and 10). In this context, the median provided a more robust and meaningful threshold than the mean, enabling clearer categorical comparisons between users who expressed generally positive versus negative appraisals of Ozempic. We plotted the proportions of reviews mentioning each subtheme in the higher- and lower-rating groups, reflecting their relative frequencies within each group. For visualization purposes, each subtheme was classified according to the predominant sentiment expressed by respondents: "positive" (eg, treatment-associated weight loss), "neutral" (eg, injection process), or "negative" (eg, nausea and gastrointestinal complaints; see Figure 2B). Separate chi-square tests were conducted to compare the frequency of each subtheme's representation between higher- and lower-rating groups. A 2-sided test with a type I error rate of 0.05 was used for all analyses.

XSL•FO
RenderX

Figure 2. A) Histogram depicting the frequency of respondents' quantitative ratings of Ozempic on a 1-10 scale. Data were bimodal, with the most frequent scores being 1 and 10. For subsequent analyses, we divided respondents into those who provided ratings above or below the median score of 7.5 ('higher' vs. 'lower' ratings). B) Respondents who provided higher quantitative ratings of Ozempic (8-10 out of 10) were more likely to have also contributed to subthemes associated with positive sentiment (as described by the respondents), including 'weight loss' and 'appetite suppression'. Respondents who provided lower ratings (1-7 out of 10) were more likely to have contributed to several more negatively valanced subthemes, including 'minimal/no weight loss' 'other physiological (nongastrointestinal) symptoms,' and 'plans to discontinue treatment'. Comparisons between higher vs. lower ratings made using χ^2 analyses. *P<0.05, **P<0.01. Numbers/letters in parentheses reflect the subthemes described in the Results section.

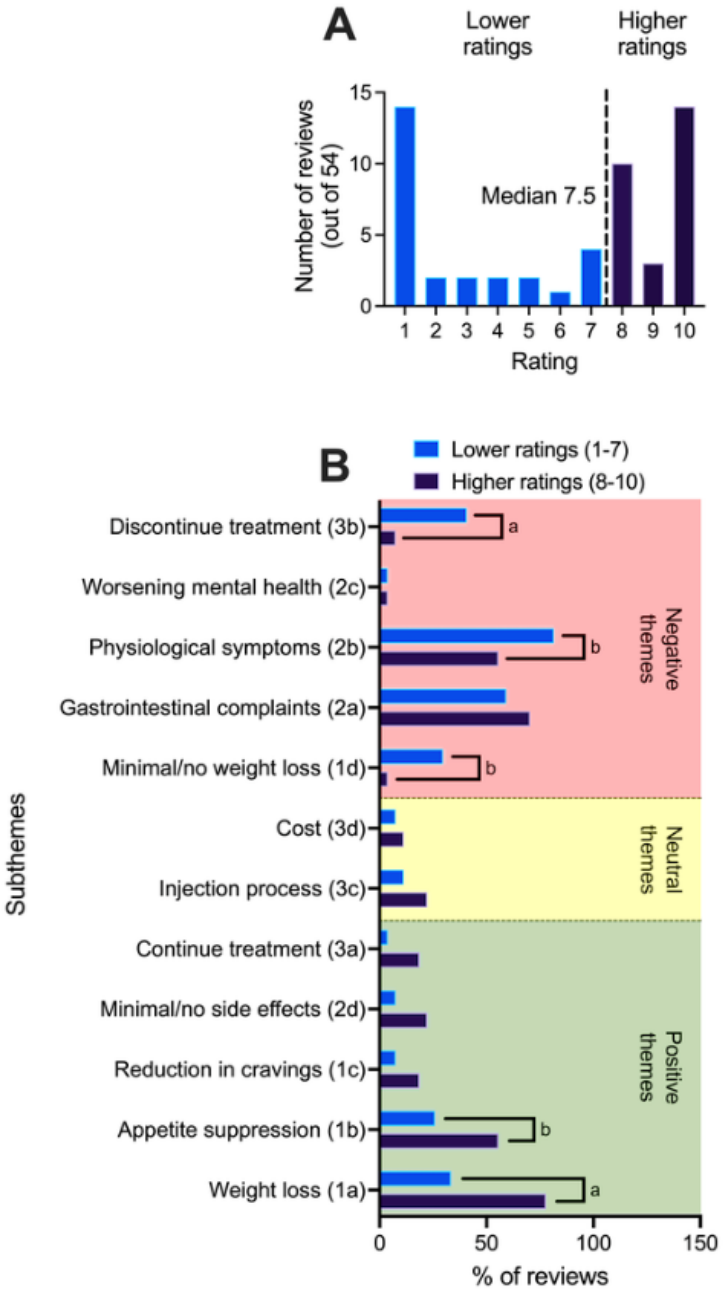

Reporting Standards

This study was designed and reported in accordance with the Good Reporting of a Mixed Methods Study (GRAMMS; Multimedia Appendix 1) and Standards for Reporting Qualitative Research (SRQR; Multimedia Appendix 2) guidelines to ensure transparency and reproducibility.

Results

Thematic Analysis

Overview

Our thematic analysis reached saturation after analyzing 60 responses. Three major themes emerged from these analyses (outlined in Tables 1-3). Each overarching theme comprised several related subthemes that reflected a spectrum of responses (eg, weight loss vs no weight loss). Some respondents contributed to multiple related subthemes (eg, individuals who

XSL•FO
RenderX

initially lost weight with Ozempic but later regained weight were included in both "Weight loss and related outcomes" and "No/minimal weight loss or weight rebound"). Below, we describe each theme and provide representative verbatim examples. For each subtheme, we also report the median quantitative rating (1-10 scale) among all participants who contributed to that subtheme, along with the respective median absolute deviation (MAD) from the median.

Theme 1: Change in Body Weight/Appetite

Approximately half of respondents (33/60, 55%) indicated that they experienced weight loss at some point during Ozempic treatment (Table 1; subtheme 1a). Some users also reported improvements in weight-related outcomes, including reductions in cholesterol levels. Quantitative ratings from respondents contributing to this subtheme (median 8.5, MAD 1.5) were higher than the overall group median (median 7.5, MAD 2.5). A substantial proportion of respondents (22/60, 37%) reported that Ozempic treatment was associated with appetite suppression (Table 1; subtheme 1b). The median quantitative ratings for this group were similar to those for subtheme 1a (median 8.0, MAD 1.5), reflecting the considerable overlap in respondents (n=15) contributing to both subthemes. Additionally, a subtheme emerged regarding a reduction in food cravings (Table 1; subtheme 1c), with 8 of 60 (13%) respondents reporting that Ozempic treatment was associated with decreased cravings, particularly for sugary and greasy foods (median 10.0, MAD 0.0). A total of 40 unique respondents were coded under subthemes 1a, 1b, or 1c, indicating that 40 out of 60 (67%) respondents reported reductions in weight, appetite, or cravings. By contrast, 11 out of 60 (18%) respondents expressed no/minimal weight loss or weight rebound (Table 1; subtheme 1d; median 1.0, MAD 0.0). Among these, some noted they had not lost any weight, while others reported weight loss occurring slowly. Four respondents in subtheme 1d also appeared in subthemes 1a, 1b, or both, suggesting that although they initially experienced reductions in weight or appetite, these effects were not sustained over time (ie, weight loss plateaued or reversed).

Table 1. Representative quotes for theme 1: "Change in Body Weight/Appetite."

Subtheme	n	Median (median absolute deviation) drug rating (0-10)	Examples of review comments (and drug rating associated with comment)
1a. Weight loss and related outcomes	33	8.5 (1.5)	• I've been taking Ozempic for almost 1 year and I have lost 55 lbs. [Rating N/A[a]] • For the first time in 30 years I don't go to bed kicking myself for what I've eaten or making promises to myself to make amends for overeating. [Rating 7] • I started at 192 lbs Nov 15th and as of May 1st I weigh 152 lbs. [Rating 10] • I am more than pleased that my cholesterol is now 180 from 270 on medication. I have never been below 200 total cholesterol in my life! [Rating 10] • In the first month, I lost 15 pounds on the lowest dose. [Rating 10]
1b. Appetite suppression	22	8.0 (1.5)	• I don't have much an appetite and I feel fuller faster and longer. [Rating 10] • My appetite was reduced by 90%. I used to overeat, but now I can only manage two small meals a day. [Rating 9] • It curbed my appetite from the moment I took the 1st dose. I didn't feel any hunger despite being on a low-calorie diet and exercising five times a week. [Rating 10] • Yes, it makes you eat way less, I was never hungry but made myself eat because I had to. [Rating 2]
1c. Reduction in food cravings	8	10.0 (0.0)	• I found my sugar cravings disappeared once I started taking 1 mg. Up until then I still craved sugary foods. I lost all interest in greasy food (fries, anything deep fried etc) from .5 mg and up. [Rating 10] • Ozempic helped me cure my sugar addiction and greediness. [Rating 10] • I don't crave junk food and only eat 1/3 of what I used to since I stay full longer. It's nice [Rating 8]
1d. No/minimal weight loss or weight rebound	11	1.0 (0.0)	• I am loosing [sic] weight but it is very slow (.25/week if I am lucky) [Rating 5] • It has done absolutely NOTHING for me for weight loss. [Rating 1] • it seems to have plateaued as I haven't lost any weight since Christmas and it's now March. [Rating N/A] • To date, I am not losing anything and most weeks I have gained the weight back. [Rating 1]

[a]N/A: not applicable.

XSL•FO
RenderX

Theme 2: Nonweight-Related Symptoms and Side Effects

The majority of respondents (48/60, 80%) indicated that Ozempic treatment was associated with nonweight-related symptoms and side effects that varied in nature and severity. Nausea and gastrointestinal complaints were the most frequently reported (Table 2; subtheme 2a; 37/60, 62%), including general nausea, vomiting, burping, and severe constipation. Interestingly, the quantitative ratings associated with this subtheme (median 8.0, MAD 2.0) were similar to the overall median across all respondents (median 7.5, MAD 2.5), suggesting that the presence or absence of these symptoms was not a decisive factor in participants' overall appraisal of the medication (discussed further below). Respondents also described a range of other physiological (nongastrointestinal) symptoms, including headaches, gallbladder complications, severe dehydration, blood loss, and anemia (Table 2; subtheme 2b; 40/60; 67%; median 7.0, MAD 3.0). Two respondents reported negative impacts on mental health, specifically the onset or worsening of depressive symptoms (Table 2; subtheme 2c; median 4.5, MAD 3.5). Finally, a minority of respondents (8/60, 13%) explicitly indicated that they did not experience any distressing adverse events (Table 2; subtheme 2d; median 10.0, MAD 0.0).

Table 2. Representative quotes for theme 2: "Nonweight-Related Symptoms and Side Effects."

Subtheme	n	Median (median absolute deviation) drug rating (0-10)	Examples of review comments (associated quantitative rating of Ozempic efficacy)
2a. Nausea and gastrointestinal complaints	37	8.0 (2.0)	• It started with huge belches and nausea. That night I vomited and was so lethargic and nauseous that I didn't get out of bed for 3 days. [Rating 5] • The most consistent symptom throughout my 5 months on Ozempic has been severe constipation. The inability to down lots of water like I used to has only added to the constipation. Nausea has also been prevalent from early on but reached the point of unbearable after a few weeks on 2 mg. Started throwing up daily around that time as well which was when the costs started outweighing the benefits. [Rating 7] • I have had some very bad nausea, vomiting and diarrhea. Also lots of burping and it smells terrible. When I have vomited - it is sooo much volume. More than anytime in my life. [Rating 7] • I have had massively bad headaches, nausea, vomiting, and stomach pain. After the first injection, I ended up in the ER because of my stomach pain, and then again 5 days later. I haven't been able to keep anything down. I can barely keep 2 sips of water down. I can't even take any of my prescription medications because I am constantly throwing them back up. [Rating 3]
2b. Other physiological (non-gastrointestinal) symptoms	40	7.0 (3.0)	• I waited for three weeks and then tried again. However, after two injections, I became severely dehydrated and ended up in the ICU at the Heart Center. I had collapsed at a baby shower due to dehydration and was experiencing blood loss in my stools, anemia, and electrolyte imbalances. I could have died, but thankfully, I am still here to share my story. [Rating 1] • I've gotten a few headaches [Rating 10] • The sad news is I got gallbladder problems from ozempic and my gallbladder has to be removed. [Rating 1] • However, after a couple of months, I started getting abdominal pains in the upper right quadrant that extended to my back, between my shoulders. The pains would manifest a day or two after the shots and last several hours. It was worse during standing or walking and not a problem when sitting. [Rating 8] • I do get a case of terrible heartburn after each injection. [Rating 1] • ...swelling of the throat, and difficulty swallowing which has not stopped since the first and last dose [Rating 1] • I have constant dizziness [Rating 4] • I had weird sores on my tongue. [Rating 5]
2c. Detrimental mental health outcomes	2	4.5 (3.5)	• The worst side effect for me was depression. I would have mild anxiety and this drug made it a lot worse and made my mood very low...I just couldn't stick feeling so low in my mood. [Rating 8] • Depression - very strange feeling, almost like out of body experience. For first time found myself HATING body...Difficulty focusing. [Rating 1]
2d. Minimal or no experience with side effects	8	10.0 (0.0)	• Obviously some people have terrible side effects, but I've had none (not even a headache!). [Rating 10] • I haven't had any side effects, only positive ones! [Rating 10] • No side effects whatsoever but saying that I was (and still am) eating healthy (no greasy/fried food, no sugar, no alcohol, low fat). [Rating 10]

XSL•FO
RenderX

Theme 3: Plans for Ongoing Use Versus Discontinuation

Some respondents (n=20) discussed their intentions regarding continuation or discontinuation of Ozempic treatment, as well as practical considerations related to ongoing use. A small subset (n=6) explicitly reported plans to continue treatment (Table 3; subtheme 3a; median 9.0, MAD 1.0), often despite experiencing side effects. By contrast, a larger number (n=14) explicitly indicated plans to discontinue Ozempic (Table 3; subtheme 3b), and their quantitative ratings were correspondingly lower than the overall group median (median 3.0, MAD 2.0). Across the entire sample, 10 out of 60 (17%) respondents discussed the injection process associated with Ozempic administration (Table 3; subtheme 3c; median 8.0, MAD 1.0). Most users described the process as straightforward and convenient, with only 3 reporting difficulties. A small number (n=5) mentioned the cost of Ozempic (Table 3; subtheme 3d; median 8.0, MAD 1.0); while some considered it manageable when covered by insurance or medical cards, others identified cost as a barrier to continued use.

Table 3. Representative quotes for theme 3: "Plans for Ongoing Use Versus Discontinuation."

Subtheme	n	Median (median absolute deviation) drug rating (0-10)	Examples of review comments (associated quantitative rating of Ozempic efficacy)
3a. Plans to continue with treatment	6	9.0 (1.0)	• I have had all 5 of the main side effects, nausea, stomach pain, vomiting, diarrhea, and constipation. I am happy with the weight loss, so am learning to manage these. [Rating 8] • The side effects at the beginning were worth it for me but from the sounds of it, mine weren't that bad. I have never once thrown up. [Rating 10] • What it has done is to force me to give up bad habits because I do not like staying nauseous. I do believe this negative reinforcement will make me sustain weight loss after I have finished my rounds. [Rating 10]
3b. Plans to discontinue treatment	14	3.0 (2.0)	• I do not plan to continue. No pain, no gain. I'll get my loss the old-fashioned way with proper diet and exercise. Never again want to feel this way on Ozempic. [Rating 1] • ...while watching TV, I heard an Ozempic commercial that included warnings about gallbladder problems and pancreatitis. I immediately stopped using it and went to my doctor. [Rating 1] • I have had pretty much all the side effects possible. I have missed four days of work because I can't leave my bathroom. I can't keep anything in my stomach with it hurting and the diarrhea will not stop. I'm getting dehydrated but If I drink it goes right through me. I have a family to look after and I can't. All I want to do is sleep. I don't think I will be taking my second injection. [Rating N/A[a]] • I've been on Ozempic for 4 months. Recently raised my dose to .75 mg but I am stopping this medication. Yes, I lost 15 lbs but suffered with the worse sulfur gas, was sick in bed at times, missing appointments because I felt that crap. [Rating 2] • This is my third week and I feel absolutely horrid!! I have constant dizziness, extreme fatigue, and generally feel like crap. I don't think I will continue. It's not worth it to lose maybe 20 lbs but can't get out of bed and function. I am tired of feeling like death warmed over. [Rating 4] • Got up to 2mg after 2 years at 1mg. Stopped because I began vomiting and feeling nauseous [Rating 7]
3c. Injection process	10	8.0 (1.0)	• I had to go to my doctor's office to have them me show how to use the pen from priming it to injecting the pen. It was easy. Even the directions from the website is easier. [Rating 10] • It is very convenient to use (inject once a week), and the needle is so thin (less than a hair) you don't even feel it. [Rating 10] • Dosing 2mg is difficult as a needle change is required. [Rating 7] • Incorrect filling of the pen, it never has enough according to dosage. [Rating N/A]
3d. Cost	5	8.0 (1.0)	• I'm in Ireland, and it costs €150 a month for four injections, which are covered by my medical card. [Rating 9] • Unfortunately, it did take a toll on my wallet, and I eventually had to switch from my hospital/GP to getting it online by telehealth from semalean. My son has also started taking it, and I would recommend it to anyone who is curious, but only if they are willing to see if the side effects apply to them. If they do not, and the treatment is affordable or covered by insurance, it can be truly amazing. [Rating 8] • Don't waste your money on this stuff. [Rating 1]

[a]N/A: not applicable.

XSL•FO
RenderX

Identification of Themes Contributing to Higher Versus Lower Quantitative Ratings of Ozempic Efficacy

Across the 54 participants who provided a quantitative rating of Ozempic's efficacy on the 1-10 scale, the average rating was 5.98 (SD 3.65), and the median was 7.5 (MAD 2.5; Figure 2A). For comparison, the average rating among all available user reviews of Ozempic for weight loss on Drugs.com at the time of data collection (n=78) was 6.04, with a median of 7.0 (MAD 3.0). Statistical analysis indicated no significant difference between these 2 distributions (Mann-Whitney *U* test, *P*=.996), indicating that our analytic subsample was representative of the broader dataset. The most frequent scores were 1 and 10 (n=14 each), indicating that more than half of respondents (28/54, 52%) rated Ozempic as either the highest or lowest possible score (see Figure 2A). Based on the distribution of user ratings, we divided the data into 2 groups using a median split: those with higher ratings (scores of 8-10; n=27) and those with lower ratings (scores of 1-7; n=27; Figure 2A). To examine which qualitative subthemes were associated with more positive versus negative evaluations of Ozempic, we conducted 2 × 2 chi-square tests comparing the frequency of each subtheme across the higher- and lower-rating groups (Figure 2B).

As expected, several subthemes reflecting positive sentiment were more frequently observed among respondents who provided higher ratings of perceived efficacy than among those who provided lower ratings. These included weight loss (subtheme 1a; 21/27, 78%, vs 9/27, 33%; χ^2_1[n=54]=10.80, *P*=.001) and appetite suppression (subtheme 1b; 15/27, 56%, vs 7/27, 26%; χ^2_1[n=54]=4.91, *P*=.03). Although reduction in food cravings was mentioned more often among higher-rating respondents (subtheme 1c; 5/27, 19%, vs 2/27, 7%), this difference did not reach statistical significance (*P*=.22). Three subthemes reflecting negative sentiment were observed at significantly higher frequencies among respondents providing lower ratings. These included no/minimal weight loss or weight rebound (subtheme 1d; 8/27, 30%, vs 1/27, 4%; χ^2_1[n=54]=6.53, *P*=.01), physiological (non-gastrointestinal) symptoms (subtheme 2b; 22/27, 81%, vs 15/27, 56%; χ^2_1[n=54]=4.21, *P*=.04), and plans to discontinue treatment (subtheme 3b; 11/27, 41%, vs 2/27, 7%; χ^2_1[n=54]=8.21, *P*=.004). Interestingly, the frequency of nausea/gastrointestinal complaints (subtheme 2a; 19/27, 70%, vs 16/27, 59%; χ^2_1[n=54]=0.73, *P*=.39) did not differ significantly between the higher- and lower-ratings groups.

Finally, we conducted exploratory analyses to better understand the profile of the relatively small number of respondents who explicitly indicated their intention to continue (n=6; Figure 3A) versus discontinue (n=14; Figure 3B) Ozempic treatment. All respondents (6/6, 100%) who intended to continue treatment reported ongoing weight loss, appetite suppression, or reduction in food cravings (ie, subthemes 1a, 1b, and 1c, but not 1d); this proportion differed significantly from those who intended to discontinue treatment (5/14, 36%; χ^2_1[n=20]=7.01, *P*=.008). By contrast, both groups reported a high frequency of adverse events (subthemes 2a, 2b, or 2c; 6/6, 100%, vs 13/14, 93%; *P*=.50). All 6 continuers reported nausea and gastrointestinal complaints (subtheme 2a), and 4 experienced other physiological symptoms (subtheme 2b). Similarly, 9 of the 14 discontinuers experienced nausea, and 12 reported other physiological symptoms. Together, these data suggest that the intention to discontinue Ozempic may be driven primarily by a lack of perceived efficacy (failure to lose weight), rather than by side effects or adverse events.

XSL•FO
RenderX

Figure 3. Weight loss and side effect outcomes among respondents who explicitly indicated an intention to continue (n=6) or discontinue (n=14) Ozempic treatment in the long term. Among those intending to continue (A), 100% reported ongoing weight loss, appetite suppression, and/or reduction in cravings (i.e., subthemes 1a, b, c, but not d), as well as experiencing some type of side effect (subtheme 2a, b, and/or c). Users who intended to discontinue treatment (B) were significantly less likely to report ongoing weight loss and/or appetite suppression (36%) and had a high frequency of adverse events (93%). **p<0.01, χ2 analyses).

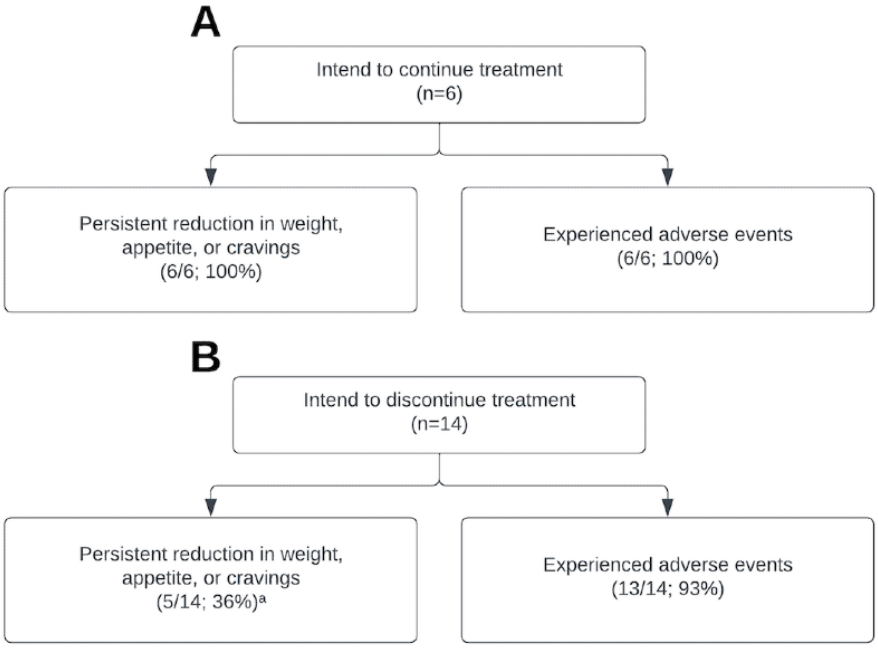

Discussion

Principal Findings and Comparison to Prior Work

This study provides new insight into how individuals perceive and evaluate the off-label use of Ozempic for weight loss, based on unsolicited, real-world data from an online medication review platform. Using a mixed methods approach, we found that user satisfaction was driven primarily by perceived effectiveness in promoting weight loss and appetite suppression, whereas gastrointestinal side effects were common but exerted limited influence on overall evaluations or decisions to continue treatment. Rather, discontinuation was most strongly associated with no/minimal weight loss or the occurrence of other, nongastrointestinal side effects. These findings highlight that, for many users, perceived efficacy outweighed tolerability concerns—a perspective that may be underrepresented in traditional clinical trials—and demonstrate the potential of infoveillance methods to capture patient-centered attitudes that shape treatment adherence.

Of the 60 respondents, 40 (67%) reported reduced weight, appetite, or cravings as a result of Ozempic treatment. This finding aligns with clinical trial data demonstrating the broad efficacy of semaglutide in promoting weight loss, with reductions of up to 17.3% observed after approximately 1 year of treatment, depending on dose and patient population [12,14,15]. Across these studies, approximately 13.5% of participants failed to achieve ≥5% weight loss with semaglutide 2.4 mg, comparable to the 18% (11/60) of participants in our sample who reported minimal or no overall weight loss. Consistent with these findings, subthemes related to weight outcomes were major contributors to respondents' overall quantitative ratings of Ozempic: subthemes 1a and 1b ("weight loss and related outcomes" and "appetite suppression") were associated with higher overall ratings, whereas subtheme 1d (no/minimal weight loss or weight rebound) occurred more frequently among lower ratings. Moreover, reductions in weight, appetite, and cravings (subthemes 1a, 1b, and 1c) were strongly associated with respondents' intention to continue versus discontinue treatment, underscoring these outcomes as being closely associated with long-term medication adherence. Notably, a retrospective study of electronic health records in the United States reported that semaglutide treatment was associated with higher persistence rates at 1 year (40%) compared with other weight loss medications, including liraglutide (17%), phentermine-topiramate (13%), and naltrexone-bupropion (10%) [31], likely reflecting its superior efficacy in promoting weight loss.

Several respondents indicated that although Ozempic treatment initially led to weight loss, this effect had plateaued or even reversed with continued use. This observation aligns with evidence from clinical studies showing that weight loss tends to plateau after approximately 1 year of semaglutide treatment, with weight regain often emerging during the second year in trial extension cohorts [32]. Such plateaus are consistent with those observed following other weight loss interventions and are thought to reflect metabolic adaptations, including reductions in resting and nonresting energy expenditure, accompanied by compensatory changes in appetite-regulating hormones [2,33-35]. In our dataset, very few respondents reported their treatment duration; therefore, we were unable to determine whether this variable mediated positive versus negative appraisals of Ozempic's efficacy. This should be a focus of future research, as it is conceivable that patient attitudes toward Ozempic become increasingly negative as weight loss plateaus or reverses over the longer term. Such dynamics likely have

XSL•FO
RenderX

important implications for long-term medication adherence, including for semaglutide formulations specifically approved for obesity and overweight [36].

A high proportion of respondents reported experiencing gastrointestinal complaints, including nausea, diarrhea, and vomiting. This aligns with clinical trial data identifying these as the most common adverse events associated with semaglutide treatment, with prevalence ranging from 41.9% to 82.8% across studies [12-16], as well as with preclinical evidence that GLP-1R agonists act on hindbrain regions involved in emesis control [37]. In our sample, these adverse events occurred with similar frequency among respondents who provided higher versus lower quantitative ratings, suggesting that gastrointestinal symptoms did not substantially influence overall attitudes toward Ozempic as a weight loss medication. This aligns with data from a previous study showing that 99.5% of gastrointestinal adverse events were nonserious, transient, and occurred most frequently during or shortly after dose escalation [38]. Moreover, across several clinical trials, treatment discontinuation due to gastrointestinal complaints was relatively uncommon, affecting only 3.4%-4.2% of participants [14,16]. Together, these findings indicate that gastrointestinal side effects are generally well tolerated and often regarded as "acceptable," particularly among individuals who experience meaningful weight loss (as described by 1 respondent: "I have had all 5 of the main side effects, nausea, stomach pain, vomiting, diarrhea, and constipation. I am happy with the weight loss, so am learning to manage these"). By contrast, users who reported other physiological (nongastrointestinal) symptoms tended to give lower quantitative ratings. This may reflect the greater severity of some of these adverse events, with several users indicating hospitalization due to complications such as severe dehydration or gallbladder removal. Although we cannot confirm that these outcomes were directly attributable to Ozempic treatment, serious treatment-associated adverse events have been reported in approximately 10% of participants in large-scale studies [15,16], including gallbladder disorders such as cholelithiasis and cholecystitis, which have led to treatment discontinuation in some cases [12].

Our data also included 2 instances in which respondents reported experiencing depression symptoms that they attributed to Ozempic treatment. Recent discussions have raised concerns about a possible association between semaglutide use and adverse mental health outcomes, particularly suicidal ideation [39]. This aligns with a recent FDA submission noting a disproportionate number of reports of "depression/suicidal" and suicidal ideation among individuals treated with semaglutide, although no causal relationship was established [39]. However, other studies, including a recent meta-analysis of 25 clinical trials, have found no association between GLP-1R agonists and suicidal or self-injurious behaviors [40,41], and some evidence even suggests a lower risk of these outcomes compared with other medications for obesity and T2D [42]. These mixed findings mirror patterns observed among bariatric surgery patients, where treatment has been associated not only with improvements in depression and anxiety but also with an increased risk of suicidality and self-injurious behavior [43]. Collectively, these data highlight the need for further research, including prospective studies and controlled clinical trials, to clarify the potential mental health risks associated with semaglutide use. Complementary preclinical investigations may also be necessary to identify shared neurobiological pathways underlying the regulation of appetite and mood.

Limitations and Future Directions

We acknowledge several important limitations of our study. First, our data were opportunistic and derived exclusively from a single website (Drugs.com), which does not publish demographic information about its users. As such, it is unclear to what extent our findings are representative of the broader population of Ozempic users. Online reviewers may also differ systematically from the general treatment population (eg, in health literacy, socioeconomic status, or engagement with digital health platforms), potentially biasing the types of experiences shared. Furthermore, self-selection biases may amplify extreme positive or negative perspectives, leading to an overrepresentation of polarized views [44]. Although we mitigated this by including all eligible reviews and reporting aggregate rather than individual data, future research should extend these findings through prospective, consented studies that collect demographic and clinical information, enabling the hypotheses generated here to be tested in more representative and generalizable samples.

Second, cumulative evidence indicates that weight loss in response to GLP-1R agonists may be more pronounced among women [45-47]; future research should therefore examine whether perceptions of Ozempic's clinical benefits and side effect profile differ by sex. Relatedly, because the data were self-reported, we could not independently verify clinical outcomes, adverse events, or the reasons for Ozempic use. Although this limitation may introduce some inaccuracy, the strong consistency of themes across respondents provides reassurance regarding the reliability of the data. Third, weight loss outcomes and adverse events are likely influenced by both the dose of Ozempic and the duration of treatment. These variables were not available in the present dataset and therefore, could not be analyzed. Finally, online reviews of consumer products, including medications, may be shaped by contextual factors such as prior reviews or platform norms, which could "prime" respondents to emphasize certain outcomes over others. This limitation underscores the need to triangulate infoveillance data with controlled, prospective designs to validate and extend these findings.

Clinical Implications

These findings offer practical insights for both clinicians and their patients. For clinicians, acknowledging that side effects are common, and occasionally serious, can help guide expectation-setting, safety monitoring, and decisions about when to consider alternative treatments. For patients, understanding that side effects vary in severity and that weight loss may plateau over time can support more realistic expectations and informed discussions with health care providers about whether to continue or adjust therapy. Together, these insights can foster clearer communication and more patient-centered decision-making.

XSL•FO
RenderX

Ethical Implications of Using Deidentified Online Data

Although we analyzed public, deidentified posts, we acknowledge that ethical considerations remain, including (1) the potential for reidentification through rare combinations of clinical details; (2) users' possible expectation that posts were intended for peer support rather than research; and (3) the risk of unintended harm, such as reinforcing stigma around the use of medications for weight loss. We mitigated these risks by limiting data collection to information necessary for analysis, screening quotations to remove potentially identifying details, reporting results in aggregate, and using neutral, nonsensational language. Future research should build on the insights from this study through prospective designs with informed consent, enabling hypotheses generated here to be tested in more representative and ethically robust samples that include limited demographic and clinical data.

Conclusion

Attitudes toward Ozempic were shaped primarily by its perceived effectiveness in promoting weight loss, with gastrointestinal side effects exerting minimal influence on overall satisfaction. For many users, the benefits of appetite suppression and weight reduction outweighed treatment-related discomfort. By leveraging an infoveillance approach, this study identified key patient-reported factors driving satisfaction and discontinuation that may be underrepresented in traditional research. These findings provide a foundation for future structured studies aimed at improving adherence and optimizing treatment strategies for individuals with overweight and obesity.

Acknowledgments

AJA gratefully acknowledges support through a Post-Baccalaureate Research Experience for LSAMP Students (PRELS) stipend from the National Science Foundation via the Garden State-LSAMP (NSF Award HRD-1909824). RTH gratefully acknowledges support through an LSAMP Undergraduate Summer Research stipend from the National Science Foundation via the Garden State-LSAMP (NSF Award HRD-1909824). This work was funded by grants to MHJ from the University of Sydney (Horizon Fellowship); the National Institute on Drug Abuse (R00 DA04765 and R37 DA061303); the National Heart, Lung, and Blood Institute (U01HL150852); Rutgers Optimizes Innovation; the New Jersey Health Foundation (PC144-23 and PC98-22); and an International Collaborative Research Grant from Rutgers Global. Artificial intelligence (ChatGPT; OpenAI) was used for minor language refinement and editing; all outputs were reviewed and modified, and the final content is the authors' responsibility. Figures 1 and 3 were created using Lucidchart (Lucid Software Inc). Special thanks to the individuals who shared their experiences anonymously on Drugs.com; we hope this research helps improve care for such individuals.

Data Availability

The data that support this study are available upon reasonable request from the corresponding author.

Authors' Contributions

AJA and MHJ conceptualized the study. AJA and RTH collected and coded the data. AJA, RTH, RMB, and MHJ reviewed the data and discussed potential themes. AJA and MHJ drafted the manuscript, and AJA, RTH, KRG, HEB, RMB, and MHJ edited and approved the final version.

Conflicts of Interest

MHJ is an inventor on patent PCT/US23/27918 titled "Therapeutic combinations and methods," which describes novel approaches for reducing overeating. All other authors declare no competing interests.

Multimedia Appendix 1
GRAMMS (Good Reporting of a Mixed Methods Study) checklist.
[PDF File (Adobe PDF File), 62 KB - jmir_v28i1e78391_app1.pdf]

Multimedia Appendix 2
SRQR (Standards for Reporting Qualitative Research) checklist.
[PDF File (Adobe PDF File), 76 KB - jmir_v28i1e78391_app2.pdf]

References

1. WHO (World Health Organization). Obesity and overweight. WHO. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [accessed 2025-11-11]
2. Hall KD, Kahan S. Maintenance of lost weight and long-term management of obesity. Med Clin North Am 2018 Jan;102(1):183-197 [FREE Full text] [doi: 10.1016/j.mcna.2017.08.012] [Medline: 29156185]
3. Wadden TA, Tronieri JS, Butryn ML. Lifestyle modification approaches for the treatment of obesity in adults. Am Psychol 2020;75(2):235-251 [FREE Full text] [doi: 10.1037/amp0000517] [Medline: 32052997]

XSL•FO
RenderX

4. Coutinho W, Halpern B. Pharmacotherapy for obesity: moving towards efficacy improvement. Diabetol Metab Syndr 2024 Jan 03;16(1):6 [FREE Full text] [doi: 10.1186/s13098-023-01233-4] [Medline: 38172940]
5. Arterburn DE, Telem DA, Kushner RF, Courcoulas AP. Benefits and risks of bariatric surgery in adults. JAMA 2020 Sep 01;324(9):879. [doi: 10.1001/jama.2020.12567]
6. Haddad A, Suter M, Greve JW, Shikora S, Prager G, Dayyeh BA, et al. Therapeutic options for recurrence of weight and obesity related complications after metabolic and bariatric surgery: an IFSO position statement. Obes Surg 2024 Nov;34(11):3944-3962. [doi: 10.1007/s11695-024-07489-7] [Medline: 39400870]
7. Mahapatra MK, Karuppasamy M, Sahoo BM. Semaglutide, a glucagon like peptide-1 receptor agonist with cardiovascular benefits for management of type 2 diabetes. Rev Endocr Metab Disord 2022 Jun;23(3):521-539 [FREE Full text] [doi: 10.1007/s11154-021-09699-1] [Medline: 34993760]
8. Ard J, Fitch A, Fruh S, Herman L. Weight loss and maintenance related to the mechanism of action of glucagon-like peptide 1 receptor agonists. Adv Ther 2021 Jun;38(6):2821-2839 [FREE Full text] [doi: 10.1007/s12325-021-01710-0] [Medline: 33977495]
9. Novo Nordisk A/S. Ozempic® prescribing information. Ozempic. 2025 10. URL: https://www.ozempic.com/prescribing-information.html [accessed 2025-12-23]
10. Jain AB, Ali A, Gorgojo Martínez JJ, Hramiak I, Kavia K, Madsbad S, et al. Switching between GLP-1 receptor agonists in clinical practice: expert consensus and practical guidance. Int J Clin Pract 2021 Feb;75(2):e13731 [FREE Full text] [doi: 10.1111/ijcp.13731] [Medline: 32975890]
11. Mahapatra MK, Karuppasamy M, Sahoo BM. Therapeutic potential of semaglutide, a newer GLP-1 receptor agonist, in abating obesity, non-alcoholic steatohepatitis and neurodegenerative diseases: a narrative review. Pharm Res 2022 Jun;39(6):1233-1248 [FREE Full text] [doi: 10.1007/s11095-022-03302-1] [Medline: 35650449]
12. O'Neil PM, Birkenfeld AL, McGowan B, Mosenzon O, Pedersen SD, Wharton S, et al. Efficacy and safety of semaglutide compared with liraglutide and placebo for weight loss in patients with obesity: a randomised, double-blind, placebo and active controlled, dose-ranging, phase 2 trial. Lancet 2018 Aug 25;392(10148):637-649. [doi: 10.1016/S0140-6736(18)31773-2] [Medline: 30122305]
13. Rubino D, Abrahamsson N, Davies M, Hesse D, Greenway FL, Jensen C, STEP 4 Investigators. Effect of continued weekly subcutaneous semaglutide vs placebo on weight loss maintenance in adults with overweight or obesity: the STEP 4 randomized clinical trial. JAMA 2021 Apr 13;325(14):1414-1425 [FREE Full text] [doi: 10.1001/jama.2021.3224] [Medline: 33755728]
14. Wadden TA, Bailey TS, Billings LK, Davies M, Frias JP, Koroleva A, et al. Effect of subcutaneous semaglutide vs placebo as an adjunct to intensive behavioral therapy on body weight in adults with overweight or obesity. JAMA 2021 Apr 13;325(14):1403. [doi: 10.1001/jama.2021.1831]
15. Wilding JP, Batterham RL, Calanna S, Davies M, Van Gaal LF, Lingvay I, et al. Once-weekly semaglutide in adults with overweight or obesity. N Engl J Med 2021 Mar 18;384(11):989-1002. [doi: 10.1056/nejmoa2032183]
16. Davies M, Færch L, Jeppesen OK, Pakseresht A, Pedersen SD, Perreault L, et al. Semaglutide 2.4 mg once a week in adults with overweight or obesity, and type 2 diabetes (STEP 2): a randomised, double-blind, double-dummy, placebo-controlled, phase 3 trial. The Lancet 2021 Mar;397(10278):971-984. [doi: 10.1016/s0140-6736(21)00213-0]
17. Ghusn W, De la Rosa A, Sacoto D, Cifuentes L, Campos A, Feris F, et al. Weight loss outcomes associated with semaglutide treatment for patients with overweight or obesity. JAMA Netw Open 2022 Sep 01;5(9):e2231982 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.31982] [Medline: 36121652]
18. Jordan G, Young S, Alemán JO. Weight loss pharmacotherapy: current and future therapies. Gastrointest Endosc Clin N Am 2024 Oct;34(4):591-608. [doi: 10.1016/j.giec.2024.06.006] [Medline: 39277293]
19. Basch C, Narayanan S, Tang H, Fera J, Basch C. Descriptive analysis of TikTok videos posted under the hashtag #Ozempic. Journal of Medicine, Surgery, and Public Health 2023 Sep 27:100013 [FREE Full text] [doi: 10.1016/j.glmedi.2023.100013]
20. Madsbad S, Holst JJ. The promise of glucagon-like peptide 1 receptor agonists (GLP-1RA) for the treatment of obesity: a look at phase 2 and 3 pipelines. Expert Opin Investig Drugs 2025 Mar;34(3):197-215 [FREE Full text] [doi: 10.1080/13543784.2025.2472408] [Medline: 40022548]
21. Garber AJ. Long-acting glucagon-like peptide 1 receptor agonists: a review of their efficacy and tolerability. Diabetes Care 2011 May;34 Suppl 2(Suppl 2):S279-S284 [FREE Full text] [doi: 10.2337/dc11-s231] [Medline: 21525469]
22. McGovern A, Tippu Z, Hinton W, Munro N, Whyte M, de Lusignan S. Comparison of medication adherence and persistence in type 2 diabetes: a systematic review and meta‐analysis. Diabetes Obesity Metabolism 2017 Dec 12;20(4):1040-1043. [doi: 10.1111/dom.13160]
23. Guerci B, Charbonnel B, Gourdy P, Hadjadj S, Hanaire H, Marre M, et al. Efficacy and adherence of glucagon-like peptide-1 receptor agonist treatment in patients with type 2 diabetes mellitus in real-life settings. Diabetes Metab 2019 Dec;45(6):528-535. [doi: 10.1016/j.diabet.2019.01.006] [Medline: 30677504]
24. Uzoigwe C, Liang Y, Whitmire S, Paprocki Y. Semaglutide once-weekly persistence and adherence versus other GLP-1 RAs in patients with type 2 diabetes in a US real-world setting. Diabetes Ther 2021 May;12(5):1475-1489 [FREE Full text] [doi: 10.1007/s13300-021-01053-7] [Medline: 33837922]

XSL•FO
RenderX

25. Weiss T, Carr RD, Pal S, Yang L, Sawhney B, Boggs R, et al. Real-world adherence and discontinuation of glucagon-like peptide-1 receptor agonists therapy in type 2 diabetes mellitus patients in the United States. Patient Prefer Adherence 2020;14:2337-2345 [FREE Full text] [doi: 10.2147/PPA.S277676] [Medline: 33273810]
26. Drugs.com. URL: https://www.drugs.com/ [accessed 2023-06-26]
27. Armanious A, Asare A, Mitchison D, James M. Patient perceptions of lisdexamfetamine as a treatment for binge eating disorder: an exploratory qualitative and quantitative analysis. Psychiatry Res Commun 2024 Dec;4(4):100195 [FREE Full text] [doi: 10.1016/j.psycom.2024.100195] [Medline: 39664649]
28. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
29. Mayor E, Bietti LM. A social media study of portrayals of bipolar disorders on YouTube: content and thematic analyses. J Med Internet Res 2025 Apr 25;27:e67129 [FREE Full text] [doi: 10.2196/67129] [Medline: 40279634]
30. Gilbert MK, Daughton AR, Chilcoat HD, Laffont CM, Strafford S, DeVeaugh-Geiss AM. Social listening for patient experiences with stopping extended-release buprenorphine: content analysis of reddit messages. J Med Internet Res 2025 Apr 25;27:e71245. [doi: 10.2196/71245]
31. Gasoyan H, Pfoh E, Schulte R, Le P, Rothberg M. Early- and later-stage persistence with antiobesity medications: a retrospective cohort study. Obesity (Silver Spring) 2024 Mar;32(3):486-493. [doi: 10.1002/oby.23952] [Medline: 38053443]
32. Wilding JPH, Batterham RL, Davies M, Van Gaal LF, Kandler K, Konakli K, STEP 1 Study Group. Weight regain and cardiometabolic effects after withdrawal of semaglutide: the STEP 1 trial extension. Diabetes Obes Metab 2022 Aug;24(8):1553-1564 [FREE Full text] [doi: 10.1111/dom.14725] [Medline: 35441470]
33. Sumithran P, Prendergast LA, Delbridge E, Purcell K, Shulkes A, Kriketos A, et al. Long-term persistence of hormonal adaptations to weight loss. N Engl J Med 2011 Oct 27;365(17):1597-1604. [doi: 10.1056/NEJMoa1105816] [Medline: 22029981]
34. Heckman BW, Mathew AR, Carpenter MJ. Treatment burden and treatment fatigue as barriers to health. Curr Opin Psychol 2015 Oct 01;5:31-36 [FREE Full text] [doi: 10.1016/j.copsyc.2015.03.004] [Medline: 26086031]
35. Fothergill E, Guo J, Howard L, Kerns JC, Knuth ND, Brychta R, et al. Persistent metabolic adaptation 6 years after "The Biggest Loser" competition. Obesity (Silver Spring) 2016 Aug;24(8):1612-1619 [FREE Full text] [doi: 10.1002/oby.21538] [Medline: 27136388]
36. Singh G, Krauthamer M, Bjalme-Evans M. Wegovy (semaglutide): a new weight loss drug for chronic weight management. Journal of Investigative Medicine 2023 May 25;70(1):5-13. [doi: 10.1136/jim-2021-001952]
37. Borner T, Geisler CE, Fortin SM, Cosgrove R, Alsina-Fernandez J, Dogra M, et al. GIP receptor agonism attenuates GLP-1 receptor agonist-induced nausea and emesis in preclinical models. Diabetes 2021 Nov;70(11):2545-2553 [FREE Full text] [doi: 10.2337/db21-0459] [Medline: 34380697]
38. Wharton S, Calanna S, Davies M, Dicker D, Goldman B, Lingvay I, et al. Gastrointestinal tolerability of once-weekly semaglutide 2.4 mg in adults with overweight or obesity, and the relationship between gastrointestinal adverse events and weight loss. Diabetes Obes Metab 2022 Jan;24(1):94-105 [FREE Full text] [doi: 10.1111/dom.14551] [Medline: 34514682]
39. McIntyre RS, Mansur RB, Rosenblat JD, Kwan ATH. The association between glucagon-like peptide-1 receptor agonists (GLP-1 RAs) and suicidality: reports to the Food and Drug Administration Adverse Event Reporting System (FAERS). Expert Opin Drug Saf 2024 Jan;23(1):47-55. [doi: 10.1080/14740338.2023.2295397] [Medline: 38087976]
40. Zhou J, Zheng Y, Xu B, Long S, Zhu L, Liu Y, et al. Exploration of the potential association between GLP-1 receptor agonists and suicidal or self-injurious behaviors: a pharmacovigilance study based on the FDA Adverse Event Reporting System database. BMC Med 2024 Feb 14;22(1):65 [FREE Full text] [doi: 10.1186/s12916-024-03274-6] [Medline: 38355513]
41. Chen J, Zhang Q, Wu Q, Zhang X, Xiang Z, Zhu S, et al. Impact of GLP-1 receptor agonists on suicide behavior: a meta-analysis based on randomized controlled trials. J Diabetes 2025 Sep;17(9):e70151. [doi: 10.1111/1753-0407.70151] [Medline: 40887719]
42. Wang W, Volkow ND, Berger NA, Davis PB, Kaelber DC, Xu R. Association of semaglutide with risk of suicidal ideation in a real-world cohort. Nat Med 2024 Jan;30(1):168-176 [FREE Full text] [doi: 10.1038/s41591-023-02672-2] [Medline: 38182782]
43. Law S, Dong S, Zhou F, Zheng D, Wang C, Dong Z. Bariatric surgery and mental health outcomes: an umbrella review. Front Endocrinol (Lausanne) 2023;14:1283621 [FREE Full text] [doi: 10.3389/fendo.2023.1283621] [Medline: 38027159]
44. Bhole B, Hanna B. The effectiveness of online reviews in the presence of self-selection bias. Simulation Modelling Practice and Theory 2017 Sep;77:108-123. [doi: 10.1016/j.simpat.2017.05.005]
45. Buysschaert M, Preumont V, Oriot PR, Paris I, Ponchon M, Scarnière D, UCL Study Group for Exenatide. One-year metabolic outcomes in patients with type 2 diabetes treated with exenatide in routine practice. Diabetes Metab 2010 Nov;36(5):381-388. [doi: 10.1016/j.diabet.2010.03.009] [Medline: 20598606]
46. Anichini R, Cosimi S, Di Carlo A, Orsini P, De Bellis A, Seghieri G, et al. Gender difference in response predictors after 1-year exenatide therapy twice daily in type 2 diabetic patients: a real world experience. Diabetes Metab Syndr Obes 2013;6:123-129 [FREE Full text] [doi: 10.2147/DMSO.S42729] [Medline: 23630427]

XSL•FO
RenderX

47. Quan H, Zhang H, Wei W, Fang T. Gender-related different effects of a combined therapy of exenatide and metformin on overweight or obesity patients with type 2 diabetes mellitus. J Diabetes Complications 2016;30(4):686-692. [doi: 10.1016/j.jdiacomp.2016.01.013] [Medline: 26873871]

Abbreviations

GLP-1R: glucagon-like peptide-1 receptor
GRAMMS: Good Reporting of a Mixed Methods Study
HbA1c: glycated hemoglobin
MAD: median absolute deviation (from median)
SRQR: Standards for Reporting Qualitative Research
T2D: type 2 diabetes

Edited by A Mavragani, N Cahill; submitted 02.Jun.2025; peer-reviewed by B Khandakar, C Annan, BT Oladun, M Almashmoum; comments to author 18.Aug.2025; accepted 04.Nov.2025; published 09.Jan.2026.

Please cite as:
Armanious AJ, Hunter RM, Griffiths KR, Bowrey HE, Brown RM, James MH
Patient Perceptions of Ozempic (Semaglutide) for Weight Loss: Mixed Methods Analysis of Online Medication Reviews
J Med Internet Res 2026;28:e78391
URL: https://www.jmir.org/2026/1/e78391
doi:10.2196/78391
PMID:

©Abanoub J Armanious, Rachel-Mae Hunter, Kristi R Griffiths, Hannah E Bowrey, Robyn M Brown, Morgan H James. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Public Emotional and Thematic Responses to Major Emergencies on Social Media, 2024-2025: Cross-Sectional Convergent Mixed Methods Study

Xingrong Guo[1], PhD; Yiqian Fan[1], MA; Yiming Guo[2], PhD

[1]College of Foreign Languages, Shanghai Maritime University, Shanghai, China

[2]School of Economics and Management, Shanghai Maritime University, Shanghai, China

Corresponding Author:
Xingrong Guo, PhD
College of Foreign Languages
Shanghai Maritime University
1150 Haigang Ave
Shanghai, 201306
China
Phone: 86 02138282732
Email: xmguo@shmtu.edu.cn

Abstract

Background: During 2024-2025, global emergencies triggered intense online discourse, presenting a unique opportunity to examine how cultural factors shape emotional expression and knowledge dissemination. Understanding these dynamic mechanisms is crucial for enhancing the effectiveness of digital health communication and optimizing crisis response strategies.

Objective: We analyzed how cultural and linguistic contexts influence emotional expression and thematic framing in social media comments during major emergencies in 2024-2025. We uncovered cross-cultural differences in collective emotions and narrative focuses, explaining how affective stance and discourse framing jointly shape the public construction of crisis meaning.

Methods: We used a cross-sectional, convergent mixed methods design. Data were collected retrospectively from X (formerly Twitter; X Corp) and Weibo (Sina Weibo) between January 1 and December 31, 2024. Using purposive sampling, we selected 5-6 representative emergency events per month based on online visibility (capped at 600 comments/event). The dataset included 19,813 comments from X and 6536 comments from Weibo. Emotions were identified using a Cross-lingual Language Model-Robustly optimized Bidirectional Encoder Representations from Transformers approach, and thematic patterns were extracted with Bidirectional Encoder Representations from Transformers Topic. Integrated Gradients was used to interpret model outputs, while clustering and network analysis were applied to visualize cross-cultural patterns. Hofstede's cultural dimensions theory helped interpret cultural influences on discourse. This mixed computational approach enabled a detailed comparison of emotional structures and thematic discourse across linguistic communities.

Results: Significant cross-platform differences were observed in emotional distribution (χ^2_8=8025.60; P<.001). Compared to X users, Weibo users, representing a collectivist culture, expressed concentrated negative emotions (20.37%; odds ratio [OR] 15.76, 95% CI 13.90-17.85), surprise (19.70%; OR 2.53, 95% CI 2.32-2.73), and fear (16.68%; OR 1.72, 95% CI 1.58-1.86), reflecting group-oriented anxiety and emotional contagion. In contrast, X (formerly Twitter) users in individualist contexts displayed dispersed sarcasm (43.49%; OR 55.19, 95% CI 43.95-69.21) and worry (15.30%; OR 55.27, 95% CI 34.74-87.88), indicating personalized and critical emotional styles. Topic modeling revealed dense clusters around "safety," "pray," and "resettlement" on Weibo, whereas X (formerly Twitter) comments emphasized decentralized themes of critique and responsibility. Semantic network analysis revealed a cohesive fear-prayer-rescue chain on Weibo and fragmented, debate-oriented interactions on X (formerly Twitter).

Conclusions: Emergency discourse is not neutral but is systematically structured by cultural values that shape emotions and themes. Integrating multilingual computational and qualitative methods, we offer a replicable framework using large-scale data, moving crisis and infodemiological research beyond single-platform or survey-based approaches. Our findings advance theory-informed understanding of how cultural meaning systems translate into observable digital discourse under conditions of risk and uncertainty. They also offer practical implications for governments, public health agencies, international organizations, and digital platforms by informing culturally adaptive, platform-specific risk communication, community moderation, and crisis

XSL•FO
RenderX

engagement strategies that can strengthen public trust, improve compliance with protective behaviors, and mitigate infodemic-related harms.

(J Med Internet Res 2026;28:e84648) doi:10.2196/84648

KEYWORDS

emergencies; cross-culture; XLM-RoBERTa; BERTopic; emotion; topic; cross-lingual language model–robustly optimized BERT approach; Bidirectional Encoder Representations from Transformers Topic

Introduction

Background

This study investigates the emotional and thematic patterns exhibited in social media comments during key global emergencies spanning 2024-2025, focusing on how cultural factors mold emotional expression and information dissemination. Understanding these dynamic mechanisms is essential for improving digital health communication and crisis response strategies. Global natural disasters, public health crises, and human-made accidents have intensified in recent years, placing unprecedented demands on emergency communication and mental health support. Social media platforms now serve as real-time spaces for emotional expression and public discourse during such events [1]. For example, on May 1, 2024, a highway in Meizhou, Guangdong, collapsed after prolonged heavy rain and unstable geological conditions. The disaster caused 52 deaths and injured 30 people. Another tragedy occurred on December 29, 2024, when a plane crashed in South Korea. Only 2 passengers survived, while 179 passengers lost their lives. Both incidents generated millions of online reactions within hours.

Social media has increasingly been acknowledged as a vital tool in emergency management. It allows authorities to monitor public sentiment in real time, evaluate opinion-related risks, and enhance the quality of risk communication [2-4]. Surveys show that 75% of people use or plan to use social media during emergencies, and 77% believe it delivers information faster than traditional channels [5]. Weibo (Sina Weibo) and X (formerly Twitter; X Corp) spread emergency updates more quickly than conventional media and provide spaces for emotional exchange and debate [5].

Limited understanding exists of how different cultural groups emotionally respond to the same emergency through online comments [6]. Although many studies have explored the role of tweets or posts during emergencies [7,8], such as information dissemination, public emotion analysis of posts, and crisis communication strategies. User comments are often overlooked because they are more spontaneous and interaction-driven [9]. While a substantial portion of research has focused on single-language or single-region datasets, some cross-lingual studies exist [10,11]. However, these studies often do not fully explore cross-cultural differences in public responses, particularly in the context of emergencies across multiple social media platforms. Therefore, there remains a need for research that systematically examines cross-cultural emotional and thematic patterns in multilingual online discourse. To bridge this gap, the study draws on comments posted during major emergencies in 2024-2025 on Weibo and X (formerly Twitter), comparing how cultures differ in emotional expression and discourse patterns. Multilingual emotion recognition is performed with the Cross-lingual Language Model-Robustly optimized Bidirectional Encoder Representations from Transformers (BERT) approach (XLM-RoBERTa), while BERTopic modeling is used to cluster topics and trace emotions such as fear, anger, and sarcasm across cultures. By revealing how emotions are shaped in online discourse, the research explains the psychological factors behind users' emergency responses and offers practical guidance for emergency communication that respects cultural differences.

The paper is structured as follows. Section 1 details data collection and the methods for preprocessing, emotion labeling, and topic modeling. Section 2 reports results on emotion patterns, topic clusters, and cross-cultural comparisons. Section 3 interprets these findings through cultural theories and discusses practical implications, concluding with limitations and directions for future work.

Literature Review

Social Media as a Digital Health Communication Tool in Emergencies

Digital health, or the use of digital technologies for health, has become an important field that applies routine and new forms of information and communications technology to meet health needs [12]. Social media platforms such as Weibo and X (formerly Twitter) enable fast and interactive communication. People can share personal experiences, questions, and feelings in real time, adding to official announcements. This immediate feedback builds a shared understanding of a situation. It also helps officials respond to misinformation and ease public anxiety, making social media a key setting for digital health efforts [13-15].

Existing research has confirmed that during crises such as the COVID-19 pandemic, digital health communication can lower uncertainty and encourage protective behaviors [16]. Other studies show that effective online risk communication can reduce fear, prevent rumors, and improve a community's ability to handle crises [17]. However, most of this research focuses on original posts or organizational messages. Less attention has been paid to the active and multilingual comment sections. In reality, comment sections often contain spontaneous, emotional conversations and reveal how people from different cultural backgrounds interpret health information. This content is crucial for understanding how people process risk information and decide how to respond. Despite the global reach of social media, there is still little cross-lingual and cross-cultural analysis of digital health communication.

XSL•FO
RenderX

The Unique Value of Comment Discourse in Emergency Communication

With the growth of social media, Weibo and X (formerly Twitter) have become major spaces for public emotion and opinion during crises. Unlike original posts that mainly broadcast information, comment threads allow real-time dialogue and shared narratives [18,19]. This bottom-up interaction captures immediate, authentic public sentiment [20] and reveals insights often missing from the posts themselves [19].

Comments play 2 key roles in society. They show natural emotional reactions and shape public agendas, influencing social norms and even public policy. Since discourse reflects cultural norms, emotional tones, and interaction styles vary across contexts [21]. For example, collectivist cultures stress emotional connection and group harmony, while individualist cultures focus more on self-expression [22]. Thus, comment styles on Weibo and X (formerly Twitter) during emergencies can differ, highlighting the need for cross-cultural comparison [23].

In social media research, comment sections have been less studied. One key reason is that these texts are often unstructured, fragmented, and full of everyday wording [24]. Nonstandard grammar, mixed languages, local slang, and subtle emotions further challenge traditional text analysis models in processing them [25,26]. These features make traditional models less effective, so studies on emergency communication have often left comments aside [27].

Emotional Characteristics and Multiemotion Analysis in Comments

Emotional expression is shaped by language and culture, as Dewaele and Pavlenko [28] emphasized in their cross-linguistic perspective on emotions, showing that different languages provide distinct repertoires for conveying feelings. In traditional sentiment analysis, Gul et al [2] note that it typically categorizes emotions into 3 broad types: positive, neutral, and negative. However, Kant et al [29] argue that this approach fails to capture the complexity of specific feelings such as fear, anger, worry, shock, and sarcasm, leading to an oversimplified understanding. Recent studies have shown that classifying emotions into 2 or 3 types is insufficient for crisis communication [4,30]. Regarding this issue, multiemotion labeling systems with categories such as fear, anger, and sarcasm offer a more accurate picture of public reactions and their potential policy implications [11,31].

Cross-lingual and cross-cultural sentiment analysis is still relatively understudied. Although social media is global, most research still relies on monolingual datasets and ignores culturally mixed communication [32]. Key linguistic characteristics, including the use of metaphors, styles of emotional expression, and cultural norms, differ greatly across languages [26]. These differences present challenges to current analytical models. Multilingual deep learning models such as XLM-RoBERTa can capture cross-linguistic semantic meanings [33,34]. Transformer-based topic modeling tools, such as BERTopic, demonstrate stronger performance in extracting coherent discussion themes from unstructured comment data infused with emotional content [35,36].

Existing Approaches in Social Media Emergency Research

Emergency communication research on social media mainly uses 3 methods: network analysis, content analysis, and sentiment analysis.

Network analysis studies how information spreads structurally during emergencies. For example, Zhang et al [37] showed that network structures affect how information spreads over time. Han et al [38] proposed a convolutional neural network with an extreme learning machine model-based algorithm to analyze the emotional influence of Weibo users. It measures how emotions spread among users. Singh and Singh [39] used text and graph multiview learning for tweet sentiment analysis, revealing structural and semantic connections. However, these kinds of studies place less emphasis on the content and emotions in messages.

Content analysis focuses on message features and communication strategies. Kada et al [40] analyzed government social media posts during COVID-19, and Chen and Ping [41] used the Wuli-Shili-Renli method for natural disasters. Nevertheless, content analysis has limitations in automating and scaling up when dealing with large amounts of user comments.

Sentiment analysis is widely used to gauge public emotions. Ou et al [30] explored the evolution of public sentiments, filling the gap in multiemotion classification studies. Studies [2,29] have found that negative feelings often dominate, while Halse et al [4] underscored their role in detecting trust. However, most sentiment studies only assess basic polarity (positive, negative, and neutral). Multiemotion classification remains scarce, especially in cross-lingual settings [32,42].

Recently, researchers have combined topic modeling methods with sentiment analysis to better capture themes and emotional tones. Babalola et al [43] reported that BERTopic is more effective than traditional models. Its usefulness in health-related social media studies was also confirmed by Khodeir and Elghannam [36] and Ma et al [44].

Hofstede's Cultural Dimensions Framework

To analyze the discourse of cross-cultural comments, Hofstede's [45] cultural dimensions theory is used in this study. It is a framework that not only compares the patterns of national cultures but also explains how culture influences comment style. Hofstede identifies 6 dimensions: power distance, individualism and collectivism, masculinity and femininity, uncertainty avoidance, long- and short-term orientation, and indulgence and restraint. These dimensions link cultural values to emotional and discursive behaviors. By applying Hofstede's framework to the comparative analysis, this study aims to interpret emotional and thematic differences within broader cultural contexts.

Hofstede's framework is applied in many areas beyond theory. It appears in studies on technology adoption [46], educational behavior [47], and online communication. Research [48] shows

XSL•FO
RenderX

cultural dimensions such as individualism and power distance strongly affect behavioral intention and emotional expression across nations.

Research Gap and Study Contribution

Although research on emergency communication through social media has grown, 3 key gaps remain. First, comment sections have not been fully used. Their unstructured format and the complexity of analysis limit their practical application [24,25]. Second, most studies rely on monolingual datasets and seldom examine cross-cultural comment data [32,49]. Third, research on multiemotion sentiment analysis in cross-cultural emergencies is still scarce [11,30].

To fill these gaps, this study proposes an integrated framework. The framework combines BERTopic and XLM-RoBERTa to perform cross-cultural, multiemotion analysis of comments from Weibo and X (formerly Twitter) during emergencies. Its purpose is to expand research methods in emergency communication and to broaden the scope of discourse analysis in digital public spaces. It emphasizes both multilingual coverage and emotional sensitivity.

Research Questions

This study examines user comments on Weibo and X (formerly Twitter) from multiple countries during 2024-2025 emergency events, focusing on emotional structures and thematic discourse. The analysis seeks to identify key topics and emotional patterns within social media comments and to explore how cultural factors shape styles of expression. To address these aims, XLM-RoBERTa is applied for sentiment analysis and BERTopic for topic modeling, providing data that enable detailed cross-cultural comparison.

This study seeks to explore the following research questions:

- RQ1: How do social media comments on emergency events express collective emotions and construct shared meanings?
- RQ2: In what ways do emotional expressions and narrative focuses differ across cultural and linguistic communities?
- RQ3: How are emotional valence and narrative focus interrelated in shaping the cross-cultural representation of emergencies?

Guided by these questions, the study proposes that the distribution and valence of emotional expressions differ significantly across cultural groups, reflecting distinct affective orientations and underlying value systems. Linguistic communities are also expected to demonstrate culturally specific narrative focuses when interpreting emergency events, revealing divergent framing patterns in emotional discourse. Moreover, emotional valence and narrative focus are assumed to be interrelated across cultures, suggesting that affective stance and discourse framing jointly contribute to the construction of crisis meaning.

Methods

Mixed Methods Design Overview

This study used a convergent mixed methods design, integrating quantitative computational analyses with qualitative content and discourse analysis to examine cross-cultural emotional and thematic patterns in social media responses to emergencies. The quantitative phase included multilingual emotion classification, frequency statistics, topic modelling, semantic co-occurrence analysis, and statistical testing. The qualitative phase involved manual coding by trained human coders, interpretive examination of representative comments, and theory-driven discourse analysis based on Hofstede's cultural dimensions. Both strands were conducted in parallel, and findings were integrated during interpretation to triangulate results and enhance validity. In adherence with best practices for observational studies, this paper was drafted using the JARS (Journal Article Reporting Standards) guideline [50] and was edited according to the JARS reporting checklist [51], which is included in Multimedia Appendix 1 Checklist 1.

Quantitative Component

Study Design and Data Collection

To explore public responses to emergencies, this study used a cross-sectional observational design and systematically collected social media comments from Weibo and X (formerly Twitter) between January 1 and December 31, 2024. At the end of each month, events from the previous month were systematically reviewed within a 1-week window. Continuous or nonbreaking events were excluded to ensure that only discrete emergency events were included in the dataset. Furthermore, the total number of comments across all major relevant hashtags for that event on Weibo and X (formerly Twitter) did not exceed 600 during the selected evaluation period. Given the potentially large data volume and uneven discussion intensity across events, a purposive sampling method was used. Each month, 5-6 representative events were selected based on topic relevance, comment volume, and overall online visibility, measured by comment counts, repost numbers, and trending hashtag rankings. Data were collected retrospectively after each event. Within the archived datasets, iterative sampling was conducted in successive batches until no new emotional categories or thematic patterns emerged, indicating the achievement of analytical saturation. This approach ensured that the dataset captured events that generated substantial public interaction and emotional expression on both platforms. For example, regarding the January 2024 Japan earthquake, specific trending hashtags were identified, including "# 日本地震 (Earthquake in Japan)" on Weibo and "#JapanEarthquake2024" on X (formerly Twitter). Posts related to these events were then systematically collected, with data acquisition carried out via custom Python scripts (Python Software Foundation), strictly adhering to the platforms' developer agreement and used solely for noncommercial, academic research purposes.

In total, 26,349 valid comments were gathered through this process, with 19,813 comments from X (formerly Twitter) and 6536 from Weibo. The combination of random sampling, multistage cleaning, and independent coding ensured that the final dataset remained representative, reproducible, and reliable for analysis.

XSL•FO
RenderX

Measures, Predictors, and Confounders

This study focuses on emotional patterns and thematic content reflected in cross-cultural public comments. The XLM-RoBERTa model was used, achieving an accuracy of 78.37%, outperforming traditional models, including the support vector machine (64.2% accuracy) and XGBoost (62% accuracy) [52]. To better capture emotions, the quantitative results went beyond basic labels ("negative," "positive," and "neutral") and included additional categories such as "worried," "fear," "angry," "sarcasm," "shock," and "sad." They also captured thematic structures and semantic network forms identified by computational models. In comparative analysis, platform type (Weibo and X, formerly Twitter) and event category (natural disasters and public crises) were regarded as key influencing factors. The study aimed to examine their differentiated effects on online emotional expression and discourse construction through analyzing these factors. In addition, several background variables were considered, covering event scale, total number of comments, posting time, and geographic origin. Quantitative processing of these variables was designed to ensure the statistical robustness of the main findings.

Data Processing and Sampling Procedures

As shown in Figure 1, the analysis proceeded in 5 major stages: pretraining, information collection, data processing, model preparation, and visual analysis. A multistage cleaning process was adopted, combining automated filtering and manual checks to eliminate spam, blank comments, advertisements, and duplicates. The collected data mainly included user IDs, posting times, usernames, comment content, repost counts, comment counts, such as counts and geographic locations. Due to the exclusion of invalid and incomplete entries during preprocessing (such as blank text content or missing timestamps), the final analytical dataset (N=26,349) consisted of complete cases, making multiple imputation unnecessary. The data were cleaned and stop words removed using different natural language processing libraries depending on language (Jieba for Chinese, Natural Language Toolkit for European languages, and stopwords-ISO [International Organization for Standardization] for others). The multistage cleaning process also involved eliminating irrelevant information, HTML tags, meaningless symbols, and duplicate comments. All personally identifiable information was destroyed immediately after collection, and only valid data was retained. For events exceeding 600 comments, Python random sampling was used to cap the dataset per event to maintain representativeness and reduce class imbalance.

XSL•FO
RenderX

Figure 1. Data analysis procedure. The data analysis included 5 steps: pretraining data, collecting comments via Python, processing comments, model preparation, and building models for visual analysis. BERTopic: Bidirectional Encoder Representations from Transformers Topic; HTML: HyperText Markup Language; ISO: International Organization for Standardization; NLTK: Natural Language Toolkit; XLM-ROBERTa: Cross-lingual Language Model–Robustly Optimized BERT Approach.

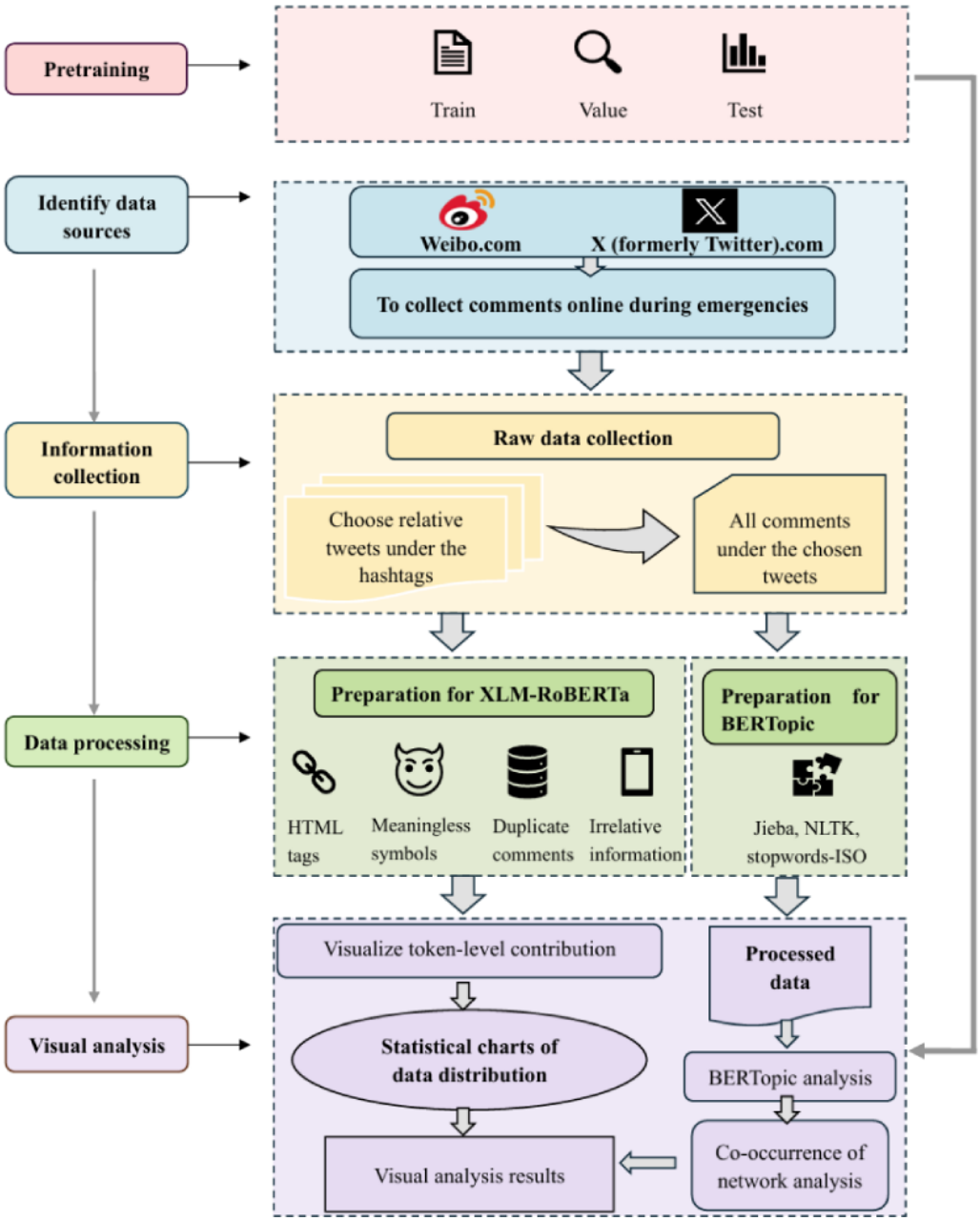

Computational Modeling for Sentiment Classification and Topic Clustering

Following the multilingual fine-tuning approach outlined by Rasool et al [53], the study used the XLM-RoBERTa-base model using a learning rate of 2×10^{-5}, batch size of 16, and 15 training epochs. The optimizer was AdamW with fused precision optimization adamw_torch_fused and a linear learning rate scheduler. The maximum sequence length was set to 128 tokens. Mixed-precision training (FP16) was enabled to optimize memory usage and training efficiency. To address class imbalance across emotion labels, a weighted binary cross-entropy loss function was used, with class weights inversely proportional to their frequency in the training corpus. Continuous emotion probability scores were binned into discrete emotional labels using predefined thresholds (neutral=0.12, surprise=0.10, positive=0.25, negative=0.30, sarcasm=0.25, fear=0.20, sad=0.25, worried=0.25, and anger=0.28). A total of 11,933 sentences were used for pretraining [54] to help the model better handle specific emotions, where meanings differ from literal words [55].

To mitigate potential semantic misalignment for low-resource languages (eg, Hindi and Indonesian), the research adopted a back-translation data augmentation strategy using the Many-to-Many 100 multilingual translation model. It has been shown to improve model quality in low-resource languages [56]. Each comment was translated from its source language to English and then back to the original language, enriching contextual diversity and improving cross-lingual embedding alignment before fine-tuning.

XSL•FO
RenderX

XLM-RoBERTa was applied to real data only after it showed adequate performance. As shown in Figure 2, the correlation checks between labels (ranging from –0.39 to 0.15) confirmed low overlap, indicating that the model has sufficient ability to process text. In addition, a manual verification step was conducted to ensure classification accuracy and validate the model's reliability. Two trained coders independently analyzed a random sample of 300 comments (150 from Weibo and 150 from X) to check the alignment between automated labels and human interpretation. After 2 coding rounds, the inter-rater reliability (Cohen κ) reached 0.88, confirming the model's high precision in capturing emotional nuances across both platforms. Nevertheless, because the model relies primarily on linguistic and semantic patterns without incorporating broader contextual cues, it may not fully capture context-dependent expressions of sarcasm or other nuanced emotional tones. These methodological constraints were taken into account when interpreting the results.

For topic clustering, the study adopted the BERTopic framework, which outperforms traditional models such as latent Dirichlet allocation and nonnegative matrix factorization when handling short texts [57]. Grootendorst [58] upgraded BERTopic, and this study adopted his improved version. As shown in Figure 3, BERTopic combines BERT embeddings, UMAP for dimensionality reduction, and HDBSCAN clustering to extract key topics from comments, followed by class-based term frequency-inverse document frequency (c-TF-IDF) to generate keywords for each topic.

Figure 2. Correlation matrix of the 9 identified emotion labels. The heatmap displays the Pearson correlation coefficients (r) between all emotion pairs, demonstrating generally low linear correlation across the labels.

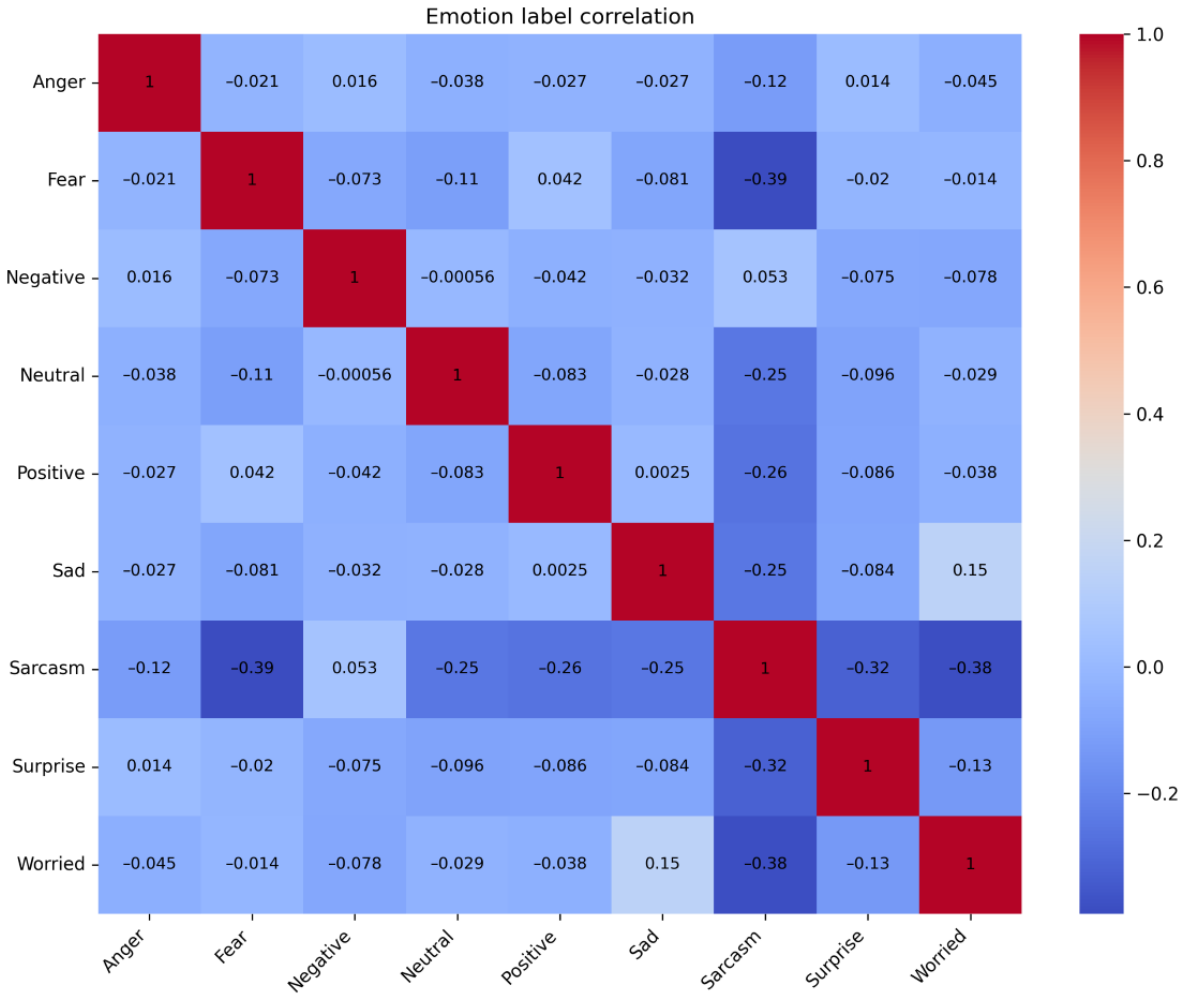

XSL•FO
RenderX

Figure 3. Bidirectional Encoder Representations from Transformers Topic (BERTopic) modeling workflow for extracting topics from social media comments. The process consists of three main steps: (1) document embedding using Bidirectional Encoder Representations from Transformers (BERT) and Transformers, (2) document clustering using Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), and (3) topic creation via class-based term frequency-inverse document frequency (c-TF-IDF) and Maximal Marginal Relevance to generate topic representations and keywords.

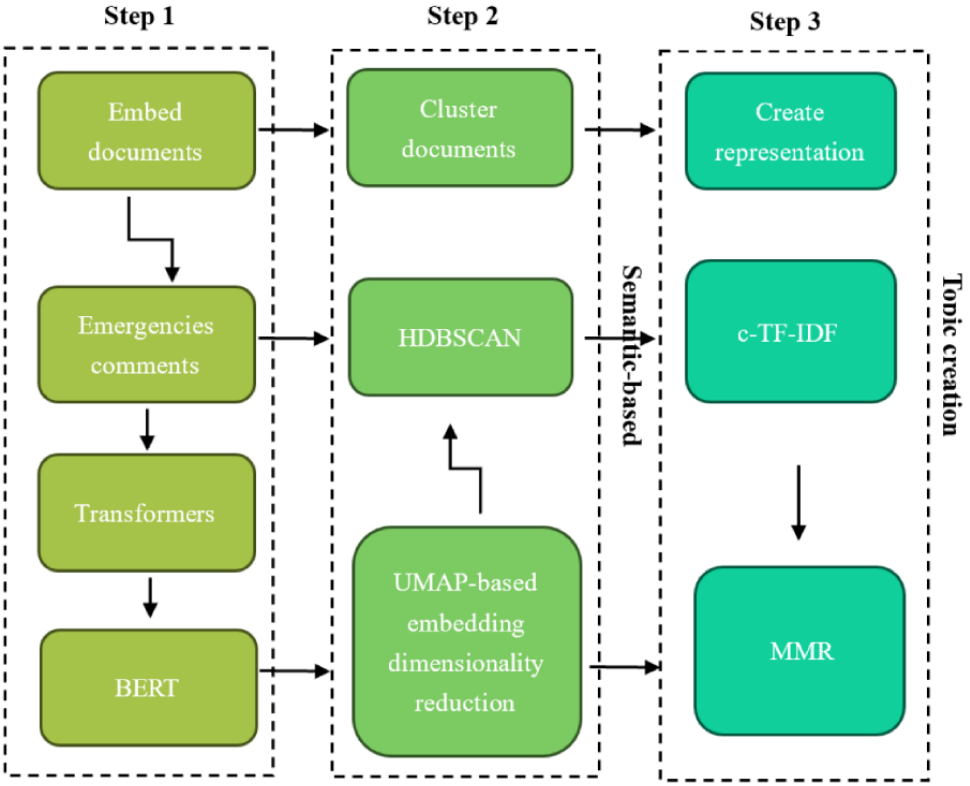

Statistical Analysis

Descriptive statistics were used to summarize distributions of comments and emotion categories. Proportions were reported with their 95% CIs, which served to quantify the precision of the parameter estimates for the analyzed dataset. Chi-square tests were conducted to assess cross-platform differences. Correlation matrices were generated to evaluate overlap among emotion labels. Statistical charts were created to visualize data distributions.

Model Interpretability Analysis

To address the inherent black-box nature of the fine-tuned XLM-RoBERTa model, the study used the Integrated Gradients (IG) attribution method [59]. IG attributes the model's prediction to input features by calculating the path integral of the gradient from a baseline (zero embedding) to the actual input. It satisfies the axioms of sensitivity (changes to an essential input feature must lead to a change in attribution) and implementation invariance (attributions must be independent of the specific model implementation) [60].

The IG attribution score for an input feature is computed as follows [61]:

Where F(x) is the model's prediction function (the target logit), is the input embedding, and χ' is the baseline embedding.

For the XLM-RoBERTa implementation, the attribution target was set to the logit output corresponding to the predicted emotion category. The baseline input "χ'" was defined as the zero embedding vector. This vector represents the absence of textual information and is a standard practice in natural language processing interpretation. The path integral required by IG was numerically approximated using the Riemann sum with 50 steps. A balance between computational efficiency and convergence accuracy was achieved with this setting. Since XLM-RoBERTa uses high-dimensional token embeddings, the raw IG output is a vector for each subword token. The final scalar token-level attribution score was derived by averaging the contribution across all dimensions of the embedding vector. This process yields a single value, which clearly indicates whether a token supports (positive score) or suppresses (negative score) the classification decision.

Qualitative Component

Data Sources and Sampling

The qualitative component was based on the same corpus of publicly available social media comments analyzed in the quantitative phase. Rather than generating new qualitative data or recruiting participants, this study examined naturally occurring textual data posted on Weibo and X (formerly Twitter) in response to emergency events during 2024-2025. These comments represent unsolicited public expressions produced in real-world digital environments and were therefore treated

XSL•FO
RenderX

as textual data sources rather than participant-generated responses.

Qualitative analysis focused on interpreting how emotions and meanings were discursively constructed within the broader patterns identified by computational analysis. Representative excerpts were examined to contextualize emotion categories, topic structures, and semantic relationships observed at the aggregate level. This approach allowed qualitative interpretation to be embedded within the full dataset, supporting explanation and triangulation without introducing a separate qualitative sample.

Manual Coding Procedures

To ensure the methodological integrity of the qualitative discourse analysis, the study conducted a separate theoretical verification. The verification was meant to systematically link emotional themes to Hofstede's cultural dimensions through the emotion-culture mapping framework. Two coders independently analyzed representative comments for theory-driven mapping. The intercoder reliability for this thematic categorization reached a Cohen κ of 0.82. Coding focused on the dominant semantic meaning and pragmatic context, such as sarcasm. This approach helped capture cultural nuance. Any discrepancies were resolved through discussion, and full consensus was reached to provide an empirically supported foundation for cross-cultural interpretations.

Discourse and Pragmatic Analysis

Linguistic differences help reveal cultural disparities. To explain how emotions and comment styles differ across cultures, Hofstede's cultural dimensions theory [45] is applied in this study. For instance, expressions such as "congratulations" in an emergency context were interpreted not as positive but as having a "sarcastic" connotation. Representative comments were analyzed at the semantic level to explain how emotions and comment styles differ between the collectivist context (Weibo) and the individualist context (X, formerly Twitter).

Integration of Quantitative and Qualitative Strands

The study combined findings by linking emotional tone (quantitative) with narrative focus (topic modeling). This mixed methods approach showed how affective orientation and discourse framing together shaped cross-cultural meaning construction. Semantic co-occurrence network analysis was combined with statistical distributions. The research emphasized the interaction between emotion and narrative structure and demonstrated that distinct emotional tendencies matched particular thematic patterns in multilingual emergency communication.

Reflexivity and Research Stance

The study acknowledges that, while the model relies on linguistic patterns, it may not fully capture all context-dependent expressions. Therefore, methodological constraints were taken into account when interpreting the results. The interpretation of "cultural disparities" is framed through the specific lens of Hofstede's theory, acknowledging the distinction between the Chinese cultural context and the Western-dominated context of X (formerly Twitter).

Ethical Considerations

This study drew on publicly accessible comments about 2024-2025 emergency events from X (formerly Twitter) and Weibo. All data were collected with Python scripts, following the platforms' terms of service. The study was deemed exempt from formal human research ethics approval by institutional guidelines, as it involved no direct interaction with individuals and used only publicly available data. Informed consent was not required since the data were public, and platform policies already notify users of potential academic use. To ensure privacy, user identifiers (such as usernames and IP addresses) were collected but immediately discarded, and all textual content was thoroughly anonymized by removing indirect identifiers before analysis. In addition, no images, figures, examples, or supplementary materials included in this paper contain information that could lead to the identification of individual users. All illustrative excerpts and visualizations are fully anonymized and presented in aggregate form. Therefore, no identifiable personal data are disclosed, and additional individual consent was not required. The research strictly followed the ethical principles of the Declaration of Helsinki and complied with major data protection regulations, including the General Data Protection Regulation and the Chinese Cybersecurity Law.

Results

Interpretability Analysis

A comprehensive interpretability analysis was conducted for the 9 emotion labels. The analysis consistently demonstrated that the XLM-RoBERTa model's predictions were highly dependent on tokens with explicit emotional valence, matching established principles of affective linguistics. As examples, local attribution results for the "worried" and "anger" emotion classifications are shown in Figures 4 and 5. For predicting "worried," the model relies heavily on specific entities or situations that may pose harm. Tokens with the highest positive contributions, such as "dest" (destination), "_har" (hardship or harm), and "locat," point to entities or difficulties that clearly strengthen the model's judgment of worry.

For predicting "anger," decisions by the model depend strongly on tokens representing highly critical or aggressive language. The tokens with the highest attribution scores, including _h, isha, and _han, correspond to strongly negative words and serve as key evidence for XLM-RoBERTa to identify anger.

In conclusion, the IG interpretability analysis provides direct evidence validating that XLM-RoBERTa's decision-making mechanism is trustworthy and intuitive. This subword-level analysis reveals that the model effectively distinguishes between relying on specific, contextual tokens and relying on strong, affective tokens, significantly enhancing transparency and trust in the model's operations.

XSL•FO
RenderX

Figure 4. Integrated Gradients (IG) token contribution scores for the "Anger" emotion, validating the Cross-lingual Language Model–Robustly optimized BERT approach (XLM-RoBERTa) model. The bar chart displays tokens that positively (green) or negatively (red) contribute to the model's prediction of "anger".

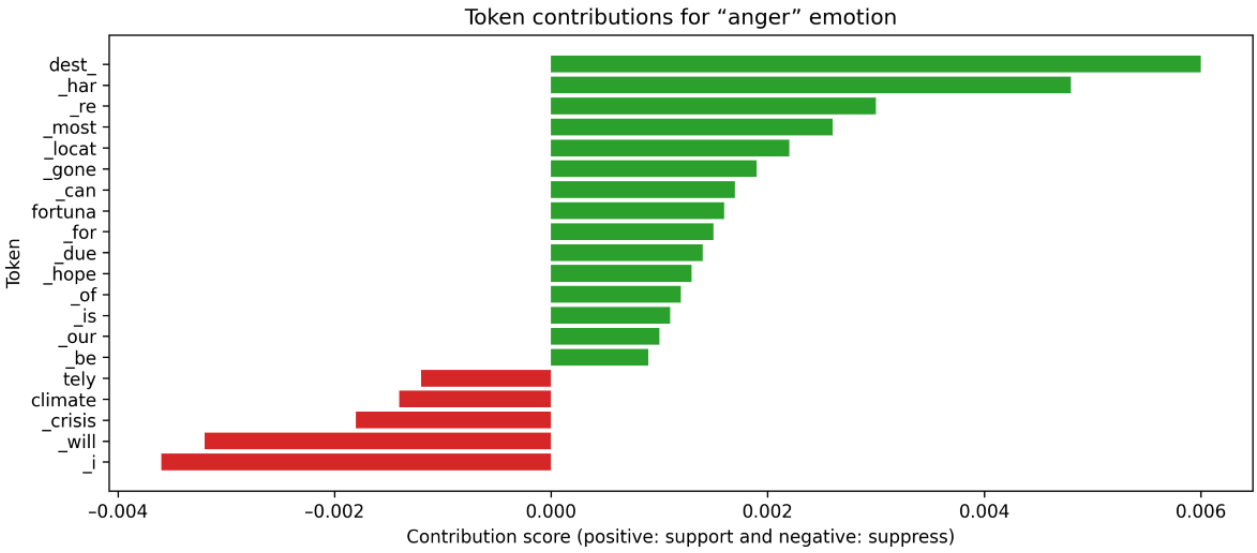

Figure 5. Integrated Gradients (IG) token contribution scores for the "Worried" emotion, validating the Cross-lingual Language Model–Robustly optimized BERT approach (XLM-RoBERTa) model. The bar chart shows the positive (green) and negative (red) contribution scores of key tokens.

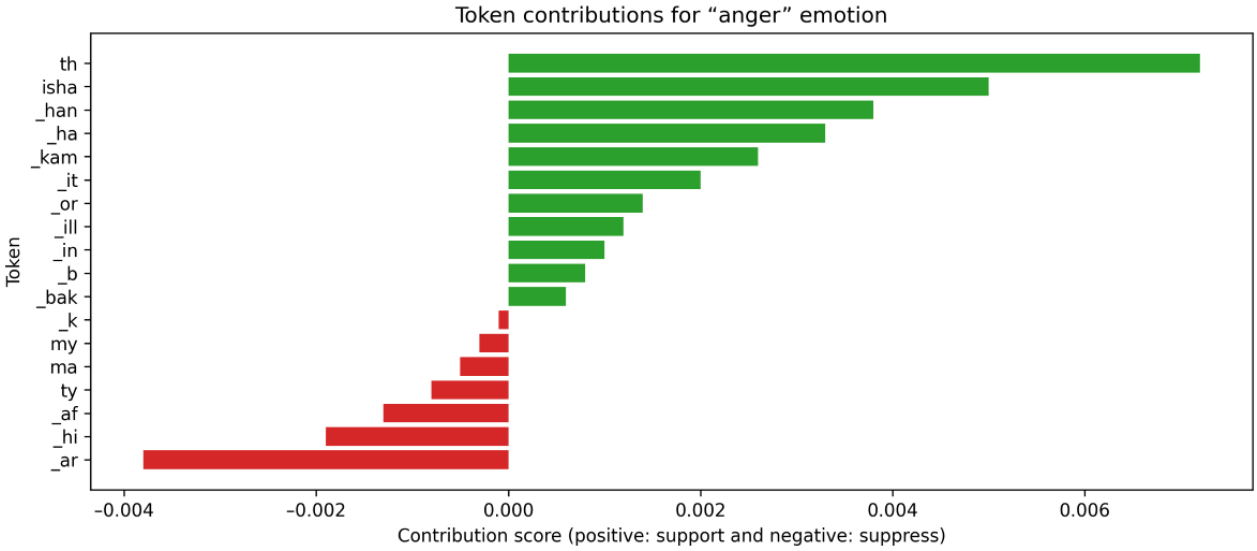

Emotional Distribution and Cultural Difference

Understanding how people feel about emergencies helps us understand how the public perceives risks and their attitudes toward them. This helps policymakers make timely decisions that consider culture. These decisions improve social stability. To test whether there is a significant difference, $P<.001$, in the distribution of emotional labels between 2 platforms, a chi-square test of independence was conducted. As shown in Table 1, the results show that there is a highly significant association between the platforms and the distribution of emotional labels. This indicates that there are obvious statistical differences in the proportions of different emotions expressed by users on Weibo and X (formerly Twitter).

XSL•FO
RenderX

Table 1. Statistical differences in the proportion of emotional labels between Weibo and X (formerly Twitter). The results ($\chi^2 8=8025.60$; $P<.001$) indicate a highly significant association between the platform used and the type of emotional labels expressed by users. All expected cell counts exceeded the minimum requirement for chi-square analysis (181.76).

Test statistic	Value (*df*)	Asymptotic significance (2-sided)
Pearson chi-square	8025.598 (8)	<.001
Likelihood ratio	8551.340 (8)	<.001
Linear-by-linear association	3008.296 (1)	<.001
N (valid cases)	29040	—

To visually illustrate the emotional distribution patterns of user comments, Figures 6 and 7 illustrate the monthly emotional distribution patterns for both platforms from 2024 to 2025. Platform X (formerly Twitter) exhibited a significantly higher comment volume; from an initial collection of 19,813 entries, the final analyzed sample for Platform X consisted of 21,600 emotion labels (mean 1800, SD 601.78). Similarly, the Weibo dataset was refined from an original 6536 entries to a final sample of 5527 emotion labels (mean 460, SD 223.13). The distribution of emotional labels on each platform showed certain differences (as shown in Table 2). “Fear” (16.68%) and “Negative” (20.37%) emotions were most prominent on Weibo, while “Sarcasm” (43.49%) dominated on X (formerly Twitter). As the emotion classification was derived from linguistic patterns without full contextual interpretation, the results should be viewed as indicative rather than absolute, particularly for context-dependent emotions such as sarcasm. Overall, the distribution reveals clear cross-platform differences, suggesting culturally distinct emotional response patterns during crises.

On Weibo (Figure 6), negative emotions such as “Fear” and “Worried” rise sharply in crisis months such as July and November. This pattern matches China’s high uncertainty avoidance tendency. Data from X (formerly Twitter; Figure 7) show a more stable distribution of these emotions (coefficient of variation [CV]=0.39), reflecting higher uncertainty tolerance in individualistic cultures with low avoidance traits. Emotional responses on Weibo often evolve into collective worry; X (formerly Twitter) users, however, express emotions more individually.

“Positive” emotions (such as prayers and gratitude) on Weibo increase sharply during disasters. For example, there were 192 labels in May and 105 in December. This pattern matches collectivist cultures’ focus on unity. On platform X (formerly Twitter), “Positive” emotions show a more even distribution (CV=0.51), suggesting higher emotional self-regulation and less need for group comfort.

“Sarcasm” shows the greatest difference. On platform X (formerly Twitter), sarcasm occurs often (n=9393, 43.49%; 95% CI 42.8%-44.2%). It spreads evenly across the year (CV=0.39) and reaches the highest level during emergencies, usually targeting institutional issues. This fits cultures with low power distance and strong free expression norms. On Weibo, sarcasm is rare (n=76), mainly appearing when disaster response is clearly poor. This reflects the caution typical of high power distance and collectivist environments.

“Surprise” appears on both platforms (Weibo: n=1089 and X [formerly Twitter]: n=1913), but the tone differs. On X (formerly Twitter), people explicitly show surprise at event scale or shock, matching direct cultural styles. On Weibo, surprise is brief, quickly turning into emotions such as sympathy or worry. This fits the pattern of emotional control in high-uncertainty-avoidance cultures. These emotional patterns across platforms highlight how culture influences not only the frequency but also the style and timing of public emotional expression in response to emergencies.

XSL•FO
RenderX

Figure 6. Monthly distribution of 9 emotional labels on the Weibo Platform (2024-2025). This stacked bar chart presents the absolute monthly frequency and proportional distribution of the 9 classified emotional tags.

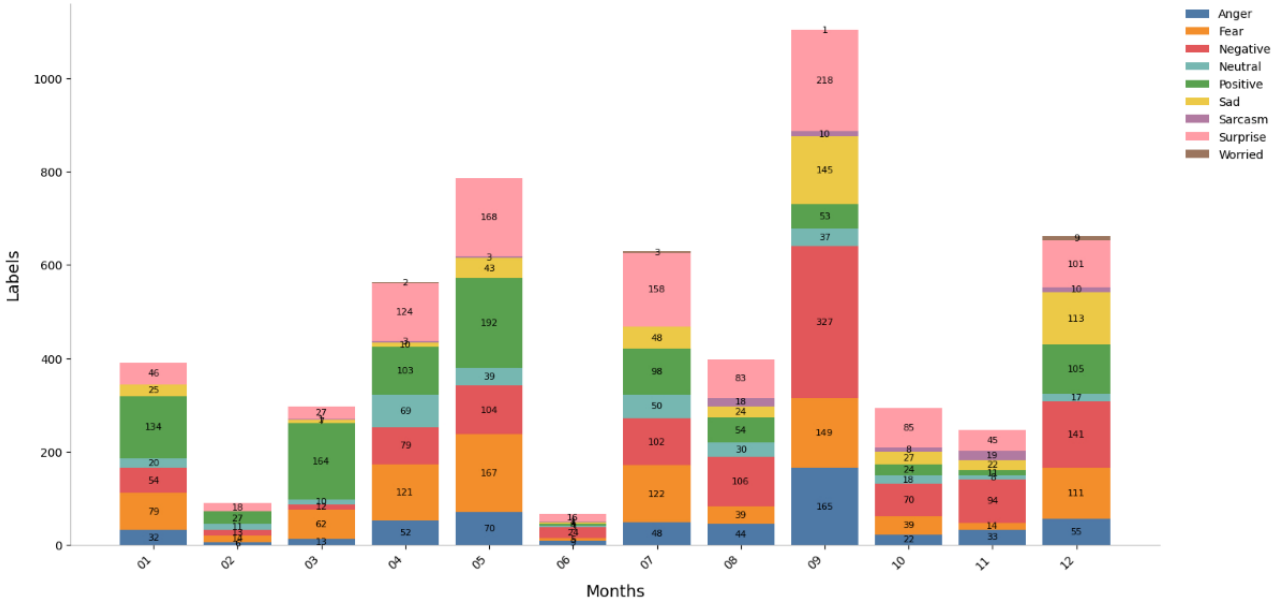

Figure 7. Monthly distribution of 9 emotional labels on the X (formerly Twitter) platform (2024-2025). This stacked bar chart displays the absolute monthly frequency and proportional distribution of the 9 classified emotional tags.

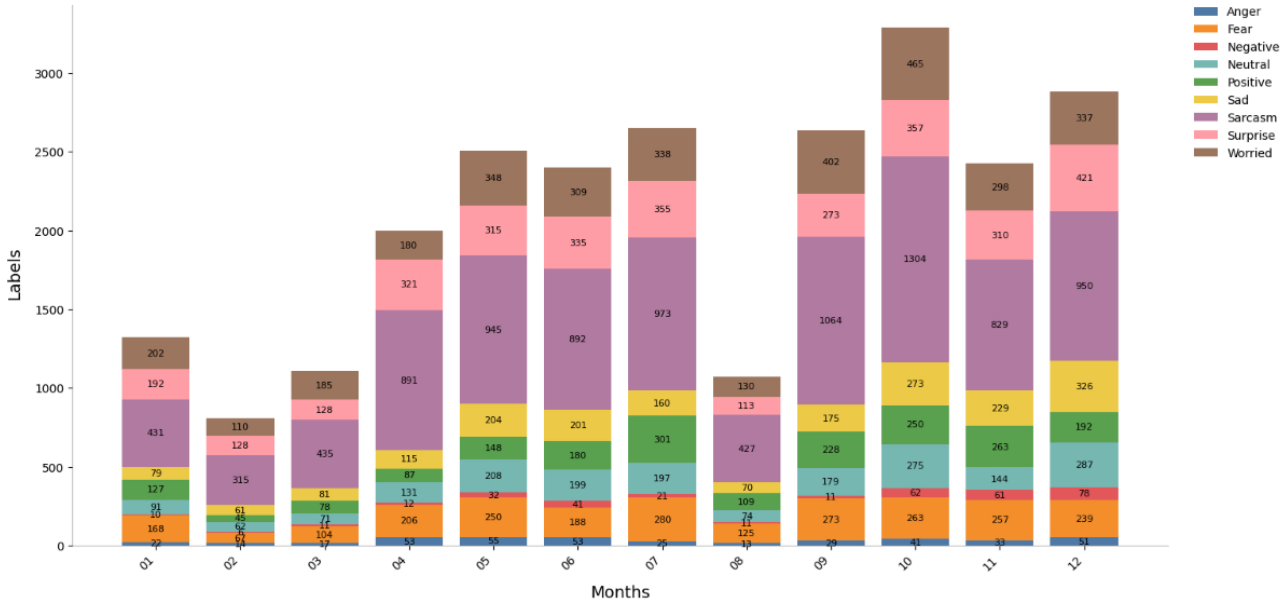

XSL•FO
RenderX

Table 2. The overall emotional distribution comparison between Weibo and X (formerly Twitter).

Tags	Weibo count	Value, % (95% CI)	X (formerly Twitter) count	Value, % (95% CI)	Total count	Value, % (95% CI)	OR[a] (95% CI)
Anger	549	9.93 (9.1-10.7)	406	1.88 (1.7-2.1)	955	3.52 (3.3-3.7)	5.76 (5.04-6.57)
Fear	922	16.68 (15.7-17.7)	2252	10.43 (10.0-10.8)	3174	11.7 (11.3-12.1)	1.72 (1.58-1.87)
Negative	1126	20.37 (19.3-21.4)	345	1.60 (1.4-1.8)	1471	5.42 (5.2-5.7)	15.76 (13.91-17.85)
Neutral	313	5.66 (5.1-6.3)	1918	8.88 (8.5-9.3)	2231	8.22 (7.9-8.6)	0.62 (0.54-0.70)
Positive	967	17.5 (16.5-18.5)	2008	9.30 (8.9-9.7)	2975	10.97 (10.6-11.3)	2.07 (1.90-2.25)
Sad	467	8.45 (7.7-9.2)	1974	9.14 (8.8-9.5)	2441	9.00 (8.7-9.3)	0.92 (0.83–1.02)
Sarcasm	76	1.37 (1.1-1.7)	9393	43.49 (42.8-44.2)	9469	34.91 (34.3-35.3)	0.02 (0.01-0.02)
Surprise	1089	19.7 (18.7-20.8)	1913	8.86 (8.5-9.2)	3002	11.07 (10.7-11.4)	2.53 (2.33-2.74)
Worried	18	0.33 (0.2-0.5)	3304	15.3 (14.8-15.8)	3322	12.25 (11.9-12.6)	0.02 (0.01-0.03)
Total	5527	—	21600	—	27127	—	—

[a]R: odds ratio; calculated with X (formerly Twitter) as the reference group. All OR values include leading zeros for values <1 per journal requirements.

Discovery of Online Public Opinion Topics

In the digital age, understanding the main topics in online public discourse during emergencies is crucial for understanding public attention, concerns, and information dissemination patterns. This section explores the dominant themes that appear on the 2 social media platforms in relation to emergency events. It uses c-TF-IDF to identify and rank key topics.

Figures 8 and 9 display the top 8 themes on Weibo and X (formerly Twitter), respectively. A total of 1071 topics were calculated for X (formerly Twitter), and 345 for Weibo. It can be observed that on both platforms, keywords directly related to emergency events (such as “rainstorm,” “hurricane,” “fire,” “airplane,” “death,” and so on) occupy significant positions. Regardless of cultural background, the public generally shows instinctive concern for disaster impacts when faced with real-world problems such as threats to life and property loss. Both platforms focus heavily on discussions about natural disasters (earthquakes, floods, typhoons, and so on). Characteristic words such as “earthquake,” “rainstorm” (on Weibo),” “hurricane,” “storm,” and “natural disasters” (on X [formerly Twitter]) are all high-frequency themes. This reflects that emergency events attract widespread attention across cultures, consistent with humans’ instinctive responses to survival and safety.

For comparative analysis, this study focused on the 7 most frequent topics within a subset of clusters from Weibo and X (formerly Twitter). For each of these topics, the 10 most frequent topic words were extracted (refer to Tables 3 and 4). The analysis identified representative themes, including “Rainfall impact” and “Fire crisis.”

These were further examined using intertopic distance maps (Figures 10 and 11), which visualize topic prevalence and semantic similarity, revealing that discussions on Weibo were more concentrated, while those on X were more dispersed.

Figure 8. Dominant themes and ranked keywords on the Weibo platform, identified by class-based term frequency-inverse document frequency (c-TF-IDF) (Topics A-H). This figure displays the top 8 (of 345 total) extracted themes, highlighting the most representative Chinese tokens and their c-TF-IDF scores.

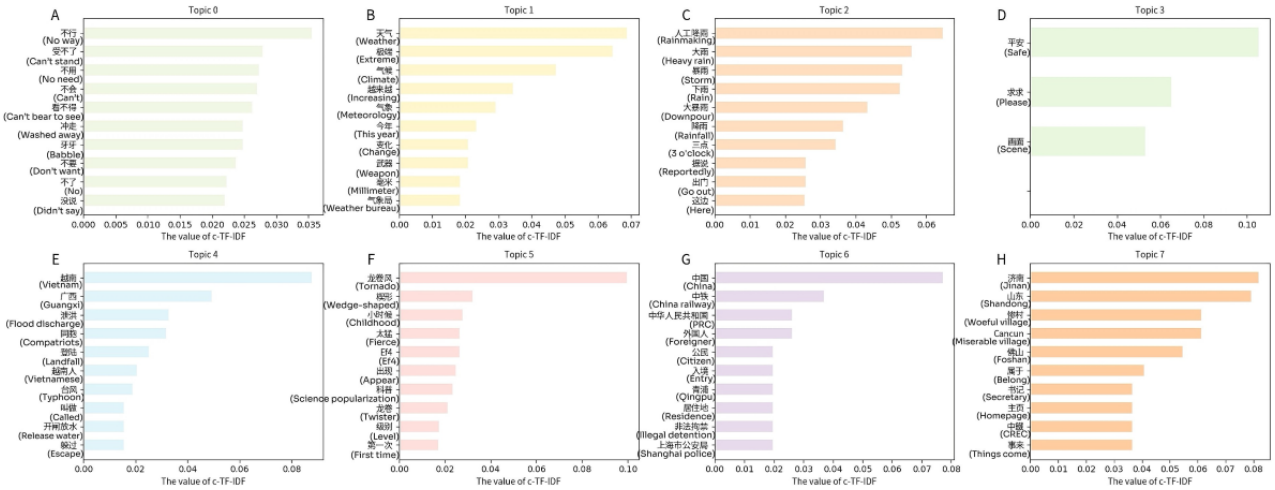

XSL•FO
RenderX

Figure 9. Dominant themes and ranked keywords on the X (formerly Twitter) platform, identified by class-based term frequency-inverse document frequency (c-TF-IDF) (Topics A-H). This figure presents the top 8 (of 1071 total) extracted themes, showing the highest-ranking tokens and their c-TF-IDF scores.

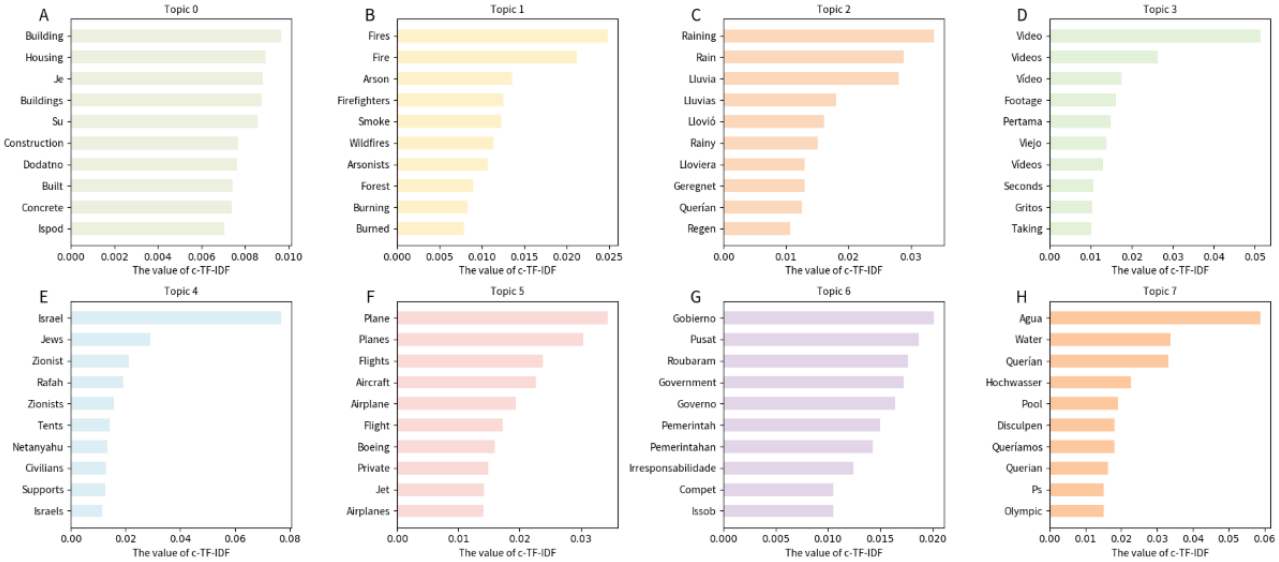

XSL•FO
RenderX

Table 3. The 7 most frequent topics and top keywords within selected clusters on the Weibo platform. This table presents the top 7 themes and their 10 highest-weighted keywords (based on class-based term frequency-inverse document frequency [c-TF-IDF] scores) that dominate public discourse on Weibo regarding emergency events.

Keywords	Weights
Topic 0	
城市(City)	0.03
危险(Danger)	0.07
气象(Meteorological)	0.02
我们(We)	0.01
越南(Vietnam)	0.11
猴子(Monkey)	0.02
他们(They)	0.05
打开(Open)	0.03
三浦(Sanpu)	0.04
这样(This way)	0.03
Topic 1	
下雨(Raining)	0.06
三点(Three o’clock)	0.05
暴雨(Heavy rain)	0.04
出门(Go out)	0.07
这边(Here)	0.03
大雨(Heavy rain)	0.06
凌晨(Early morning)	0.4
下雨天(Rainy day)	0.3
平安(Safety)	0.22
今晚(Tonight)	0.12
Topic 2	
人工降雨(Artificial rain)	0.04
大雨(Heavy rain)	0.06
大暴雨(Torrential rain)	0.04
降雨(Rainfall)	0.02
太多了(Too much)	0.02
强暴雨(Severe rainstorm)	0.46
吓人(Frightening)	0.85
平安(Safety)	0.22
顺利(Smoothly)	0.11
雨势(Rain intensity)	0.12
Topic 3	
洪灾(Floods)	0.03
遭遇(Encounter)	0.03
中东地区(Middle East)	0.02
东非(East Africa)	0.03
上古(Ancient times)	0.01
旱灾(Drought)	0.31
防灾(Disaster prevention)	0.03

XSL•FO
RenderX

Keywords	Weights
失踪(Missing)	0.08
甘蔗(Sugarcane)	0.02
极端(Extreme)	0.09
Topic 4	
火之歌(A Song of Ice and Fire)	0.07
火烧(Burning)	0.04
起火(Catch fire)	0.04
大火(Big fire)	0.04
火灾(Fire disaster)	0.04
纵火(Arson)	0.03
烟头(Cigarette butt)	0.02
烟火(Fireworks)	0.02
绝好(Excellent)	0.02
素材(Material)	0.01
Table 5	
山火(Forest fire)	0.08
扑灭(Extinguish)	0.05
肆虐(Raging)	0.04
尽快(As soon as possible)	0.04
火灾(Fire disaster)	0.04
救火(Firefighting)	0.03
见证(Witness)	0.02
菲政府(Philippine government)	0.01
关爱(Care)	0.03
减轻(Alleviate)	0.02
Topic 6	
祈祷(Pray)	0.32
人们(People)	0.05
保守(Conservative)	0.04
无人(No one)	0.8
恳求(Plead)	0.75
祈祷平安(Pray for safety)	0.17
阿弥陀佛祈祷(Amitabha prayer)	0.8
生活(Life)	0.04
活着(Alive)	0.04
大过年(During Chinese New Year)	0.11

XSL•FO
RenderX

Table 4. The 7 most frequent topics and top keywords within selected clusters on the X (formerly Twitter) platform. This table outlines the top 7 themes and their 10 highest-weighted keywords (based on class-based term frequency-inverse document frequency [c-TF-IDF] scores) identified in the X discourse on emergency events.

Keywords	Weights
Topic 0	
Fires	0.08
Fire	0.07
Arson	0.04
Smoke	0.3
Firefighters	0.03
Arsonists	0.04
Burning	0.02
Burned	0.01
(Fire)	0.04
Flames	0.19
Topic 1	
Gas	0.11
Fuel	0.3
Residential	0.14
Tanker	0.28
Cylinder	0.04
Station	0.03
Lorry	0.05
Embakasi	0.07
Plant	0.03
Cylinders	0.05
Topic 2	
Portugal	0.14
Wildfire	0.05
Fires	0.05
Wildfires	0.04
Spanish	0.04
Portugals	0.04
Portuguese	0.04
Declaran	0.02
Alguém	0.03
Arson	0.04
Topic 3	
Lava	0.07
Lewotobi	0.08
Lakilaki	0.05
Mount	0.09
Ash	0.07
Indonesia	0.15

XSL•FO
RenderX

Keywords	Weights
Evacuations	0.07
Erupted	0.07
Eruptions	0.06
16000	0.04
Topic 4	
Bus	0.17
Thailand	0.11
44	0.77
Children	0.13
Returning	0.08
Siswa	0.05
Guru	0.03
Bangkok	0.07
Teachers	0.11
Thani	0.02
Topic 5	
Gobierno	0.25
Government	0.07
Gobierna	0.12
Governo	0.08
Pemerintah	0.09
Federal	0.03
Pusat	0.06
Roubaram	0.05
Años	0.05
Provincias	0.08
Topic 6	
Rescue	0.16
Underway	0.03
Rescued	0.13
Certain	0.05
Humourrescue	0.03
Baseball	0.02
Operationswhat	0.02
Isvbeing	0.03
Resgate	0.05
Missions	0.1

XSL•FO
RenderX

Figure 10. Bidirectional Encoder Representations from Transformers Topic (BERTopic) intertopic distance map visualizing the semantic distribution and prevalence of topics on Weibo. This map uses UMAP-based dimensionality reduction to plot topics, where the size of each circle corresponds to the topic frequency. The proximity of the circles illustrates semantic similarity. The overall visualization confirms that public discourse on Weibo is thematically concentrated.

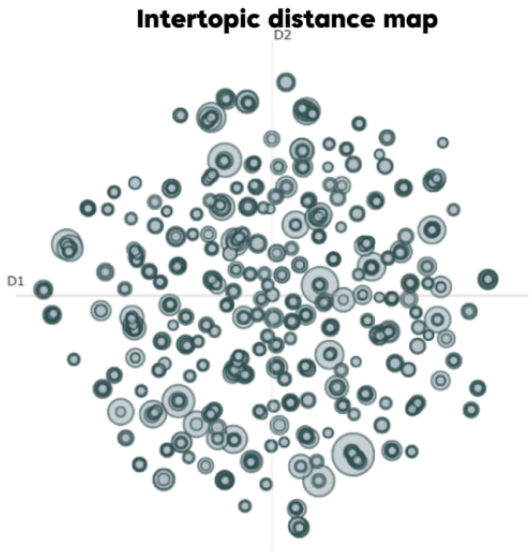

The intertopic distance map for X (formerly Twitter; Figure 11) reveals a relatively dispersed topic structure, suggesting that emergency event discussions cover a wide range of issues. A key trait of its discourse is the coexistence of several distinct clusters alongside some relatively isolated topics. For example, one prominent cluster centered on the immediate impact and characteristics of specific disasters, particularly those related to weather events. As shown in Table 4, this cluster covers Topic 0 (“fire, arson, smoke...”), with a count of 198; Topic 1 (“gas, fuel, residential, tanker”), with a count of 99; and Topic 2 (“portugal, wildfire, fires, wildfires, alguém”), with a count of 30, as evidenced by their proximity in the map and shared keywords. These topics highlight discussions about sudden natural disaster events and the dangers they pose to affected urban areas. Another cluster appears to revolve around the broader disaster context, including Topic 3 (“lava, lewotobi, lakilaki, mount, ash...”), with a count of 23, and Topic 4 (“bus, Thailand, children, returning, siswa...”), with a count of 17. While these topics are distinct, they all focus on sudden fire-related events. Simultaneously, a corresponding yet somewhat independent cluster emerges. It includes Topic 5 (“gobierno, government, gobierna, governo, pemerintah, federal, pusat, roubaram, años, provincias”), with a count of 167. Topic 6 (“rescue, underway, rescued, baseball, operations, isvbeing, resgate, missions”), with a count of 11, stands out as a major theme, indicating a strong emphasis on rescue efforts and on-the-scene situations during emergencies. The topic including “earthquake, sismo, earthquakes, terremoto...” is also relatively large, with a count of 63, suggesting substantial discussion surrounding seismic events. Notably, distinct topics focusing on long-term recovery or preventive measures are relatively scarce, though these elements appear to some extent in earthquake and flood themes.

In contrast, as shown in Figure 10, the topic structure on Weibo is more concentrated, suggesting a stronger connection among themes. Similar to X (formerly Twitter), a significant cluster on Weibo pertains to natural disasters, but with a greater emphasis on direct impacts and local contexts. It is evident within Cluster 1 that Topic 0 from Table 3 (count=51), including “城市 (city),” “危险 (danger),” “气象 (meteorological),” “我们 (we),” “越南 (Vietnam),” “猴子 (monkey),” “他们 (they),” “打开 (open),” “三浦 (Sanpu),” and “这样 (this)”; Topic 1 (count=30), with terms such as “下雨 (raining),” “三点 (3 o’clock),” “暴雨 (heavy rain),” “出门 (go out),” “这边 (here),” “大雨 (heavy rain),” “凌晨 (early morning),” and “下雨天 (rainy day)”; Topic 2 (count=20), including “人工降雨 (artificial rain),” “大雨 (heavy rain),” “大暴雨 (torrential rain),” “降雨 (rainfall),” “太多了 (too much),” “强暴雨 (severe rainstorm),” “吓人 (frightening),” “平安 (peace),” and “顺利 (smoothly)”; and Topic 3 (count=7), with “洪灾 (floods),” “遭遇 (encounter),” “中东地区 (Middle East),” “东非 (East Africa),” “上古 (ancient times),” “旱灾 (drought),” and “边境县 (border county),” are closely located on the map and share keywords related to weather events and urban impacts. Another cluster encompasses similar natural disaster themes, such as Topic 4, including “火之歌 (A Song of Ice and Fire),” “火烧 (burning),” “起火 (catch fire),” “大火 (big fire),” “火灾 (fire disaster),” “纵火 (arson),” “烟头 (cigarette butt),” and “烟火 (fireworks)”; and Topic 5, including “山火 (forest fire),” “扑灭 (extinguish),” “肆虐 (raging),” “尽快 (as soon as possible),” “火灾 (fire disaster),” “救火 (firefighting),” “见证 (witness),” “政府 (government),” “关爱 (care),” and “减轻 (alleviate).” Although different from the rainfall themes, these topics still fall under the broader category of natural disasters. Similarly, themes such as Topic 6 (count=24) are more mixed, including “祈祷 (pray),” “人们 (people),” “保守 (conservative),” “无人 (no one),” “恳求 (plead),” “祈祷平安 (pray for peace),” “阿弥陀佛祈祷 (Amitabha prayer),” and “生活 (life).”

XSL•FO
RenderX

Figure 11. Bidirectional Encoder Representations from Transformers Topic (BERTopic) intertopic distance map visualizing the semantic distribution and prevalence of topics on X (formerly Twitter). This map displays the relative distance and size (topic prevalence) of topics derived from the BERTopic model. The broad spread of the circles across the map confirms that public discourse on X is thematically dispersed and covers a wider range of distinct issues than Weibo.

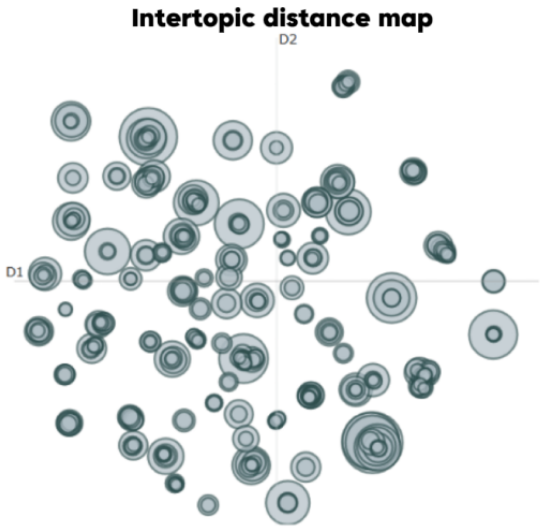

Topic analysis uncovers a common trend. Both platforms place strong emphasis on the direct impact of natural disasters. This focus is supported by keyword clusters related to heavy rainfall, floods, and earthquakes. This shared focus reflects basic human reactions to the dangers posed by these events. Weibo users tend to focus more on the natural phenomena themselves, such as “暴雨 (heavy rain),” “气候变暖 (climate warming),” and “极端天气 (extreme weather).” Their comments primarily describe disaster severity and meteorological anomalies, with emotions dominated by “担忧 (worry)” and “无奈 (helplessness).” In contrast, X (formerly Twitter) users often connect “rainfall” or “storm” to climate change or government action, expressing more explicit criticism. This difference corresponds to Hofstede’s cultural dimensions theory. In highly collectivist societies (Weibo), discourse emphasizes shared communal emotions; in low collectivist contexts (X [formerly Twitter]), it prioritizes individual critique and responsibility attribution.

Under the “Fire crisis” theme, Weibo comments frequently feature keywords such as “消防员 (firefighters),” “肆虐 (raging),” and “扑灭 (extinguish),” expressing respect for firefighters and concern about the spread of disasters. Emotions here are concentrated in “fear” and “worried.” On X (formerly Twitter), fire-related discussions include sympathy for victims, along with more anger toward arsonists and questions about the adequacy of firefighting resources. This phenomenon illustrates how people in high uncertainty avoidance cultures tend to simplify perceived causes of disasters. It also shows how the public in low power distance cultures more directly criticizes institutional failures.

For the “Air crash” theme, Weibo users primarily express “震惊 (surprise)” and “悲痛 (sadness),” using keywords such as “飞机 (airplane),” “故障 (malfunction),” “失控 (out of control),” and “逝者 (the deceased).” This reflects emotional responses and a tendency toward collective mourning when facing uncontrollable disasters. X (formerly Twitter) comments, by contrast, more frequently focus on flight safety standards and airline management, adopting a calmer and more rational tone. This pattern underscores the emphasis on factual accuracy and responsibility in highly individualistic cultural contexts. The theme also includes some sarcastic remarks on X (formerly Twitter), such as comments mocking the slowness of accident investigations, which are barely present on Weibo.

This study shows how emotional expressions and narrative focus differ across cultures. It introduces an “emotion-culture mapping table” (Table 5). The table compares user responses to emergency topics on Weibo and X (formerly Twitter). It also links these response patterns to Hofstede’s cultural dimensions. In addressing possible subjective interpretation, especially when linking emotions and themes to cultural dimensions in Table 5, a verification step was added. A subset of comments was independently analyzed by 2 human coders to confirm the categorization principles in the emotion-culture mapping table. The intercoder reliability test, measured by Cohen κ, showed substantial agreement (κ=0.82). This validation step ensures that the interpretations in Table 5 are not isolated but are systematic and empirically supported by a reliable coding method.

The table outlines 6 cultural dimensions derived from 4 core theories (such as high or low power distance). For each dimension, it details the corresponding “Narrative Focus Examples” and provides relevant “Dimension explanation and link,” thus clearly constructing the mapping relationship between culture and public narratives.

XSL•FO
RenderX

Table 5. The emotion-culture mapping table provides for a clearer understanding of emotions within the framework of Hofstede's cultural dimension theory. This table systematically compares the observed dominant emotions and narrative focus on the 2 platforms and provides a theoretical explanation for observed cross-cultural variations in online discourse regarding emergency events.

Hofstede dimension	Cultural context	Platform	Dominant emotions	Narrative focus examples	Dimension explanation and link
Power distance	Low power distance	X (formerly Twitter)	Anger and sarcasm	Direct criticism of institutional failures.	Low power distance cultures tend to question authority and power structures, and the public is more likely to directly criticize institutional failures.
Power distance	High power distance	Weibo	Fear and worry	Respect for firefighters and concern about the disaster spread.	High power distance cultures tend to respect authority, focus more on collective emotions, and exhibit less criticism of institutions.
Uncertainty avoidance	High uncertainty avoidance	Weibo	Fear and worry	Concern about the disaster spread.	High uncertainty avoidance cultures tend to seek certainty, and explanations of disaster causes may be simplified, with a focus on immediate threats.
Uncertainty avoidance	Low uncertainty avoidance	X (formerly Twitter)	Anger and sarcasm	Anger toward arson and questioning the adequacy of firefighting resources.	Low uncertainty avoidance cultures are more tolerant of uncertainty and tend to explore complex disaster causes, including human and institutional factors.
Individualism	High individualism	X (formerly Twitter)	Anger, sarcasm, and sad	Discussion of flight safety standards, airline management, a strong focus on factual truth, and responsibility attribution.	High individualistic cultures emphasize personal responsibility and autonomy, focusing on factual truth, responsibility attribution, and rational analysis.
Individualism	Low individualism	Weibo	Fear, surprise, and positive (pray and wish)	Emotional outburst and collective mourning.	Low individualistic cultures emphasize group cohesion and collective emotions, and tend to express emotions communally, such as through collective mourning.
Collectivism	High collectivism	Weibo	Worried, fear, and positive	Description of disaster intensity, meteorological anomalies, and sharing of communal emotions.	High collectivist cultures emphasize group connections and shared feelings, and tend to share common worries and helplessness, with a focus on natural phenomena.
Collectivism	Low collectivism	X (formerly Twitter)	Sarcasm	Linking rainfall to climate change and government responses, individual criticism, and responsibility attribution.	Low collectivist cultures tend toward independent thinking and critical analysis, focusing on the underlying causes of events and responsibility.

Semantic co-occurrence network graphs (Figures 12 and 13) were constructed based on word adjacency data (the top 155 terms). In the graphs, node size indicated word frequency, edge thickness represented co-occurrence strength, and colors denoted semantic communities. This approach helps to reveal semantic cores, thematic connections between clusters, and cultural differences in public discourse across platforms. Overall, the Weibo network displayed cohesive and centralized emotional narratives, whereas the X (formerly Twitter) network showed fragmented and decentralized discourse patterns. Unlike topic clustering (Figures 10 and 11), this method did not group texts by overall theme similarity; instead, it focused on word connections at a lexical level, providing a bottom-up perspective of how meaning clusters naturally formed in the discourse.

XSL•FO
RenderX

Figure 12. Semantic co-occurrence network graph of top terms on the Weibo platform. The network visualizes the thematic connections and semantic cores in the public discourse, constructed from the top 155 most frequent terms. Node size represents word frequency, edge thickness indicates co-occurrence strength, and colors delineate distinct thematic clusters.

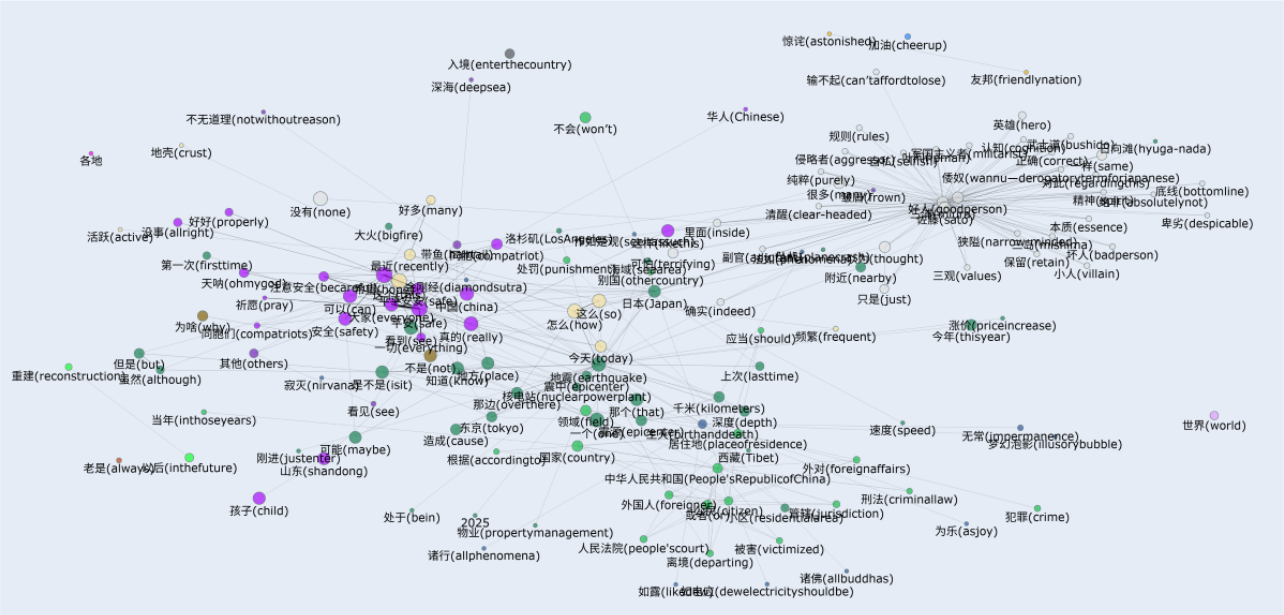

Figure 13. Semantic co-occurrence network graph of top terms on the X (formerly Twitter) platform. The network visualizes the thematic connections and semantic cores in the public discourse, constructed from the top 155 most frequent terms. Node size represents word frequency, the thickness of the edge indicates the co-occurrence strength, and colors denote thematic clusters.

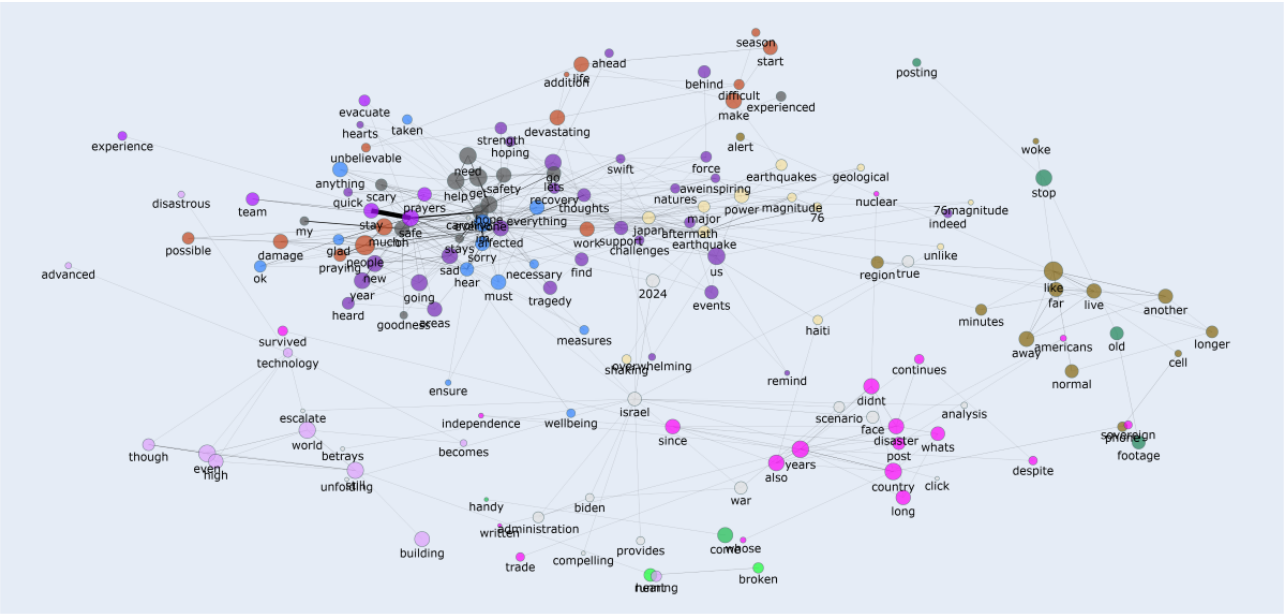

Take Weibo as an example. On this social platform, words such as “平安 (safety),” “祈祷 (pray),” “安置 (resettlement),” and “地方 (local)” created tight clusters. This finding suggests a strong narrative connection among emotional support, local identity, and institutional action. Conversely, semantic clusters on X (formerly Twitter; eg, “pray,” “rescue,” “fire,” and “government”) were more scattered. This pattern shows narratives that were emotional yet fragmented; instead of shaping a single shared story, multiple narrative threads exist side by side. High-frequency words found in topic clustering (eg, “fire” and “floods”) spread across different parts of the network. These words also appear together with various other keywords, highlighting emotional differences that topic models alone failed to capture. One specific case is “fire” on X (formerly Twitter). In 1 cluster, it appears alongside “arson” and “anger,” while in another cluster, it is linked to “rescue” and “prayers.” This example proves there were emotional complexities that a single topic label could not fully represent.

These patterns are consistent with the cultural differences discussed previously. The semantic cohesion observed in the Weibo graph suggests that collectivist cultures tend to build discourse around shared symbols and centralized values. In contrast, the scattered distribution of words on the X (formerly Twitter) platform implies a more decentralized response style, which is typical of individualistic cultures. Therefore, the semantic co-occurrence network analysis does not merely repeat

XSL•FO
RenderX

the conclusions of topic clustering. Instead, it contributes to a deeper understanding of how users interconnected emotions, facts, and cultural expressions during emergencies.

Discussion

Principal Findings

This study provides empirical evidence that cultural values influence emotional expression and discourse during global emergencies, directly relating to the research goals stated in the introduction. The study used a large-scale dataset of comments from Weibo and X (formerly Twitter) (2024-2025) and identified significant cross-platform differences in both emotional distribution and topic structures. On Weibo, negative emotions such as fear and worry are more common, reflecting China's high uncertainty avoidance. Collective expressions, including prayers and wishes for safety, are also frequent, reflecting collectivist orientations. By contrast, X (formerly Twitter) comments contain more sarcasm and criticism, reflecting individualistic values and lower power distance.Topic modeling supported these patterns: Weibo discourse focuses more on disaster severity and communal support, whereas X (formerly Twitter) discourse emphasizes accountability, institutional response, and climate-related narratives. These findings confirm that Hofstede's cultural dimensions—including individualism, collectivism, power distance, and uncertainty avoidance—are useful for explaining cultural differences in online crisis communication.

Comparison With Previous Work

The study supports earlier findings that cultural background strongly shapes emotional expression in crises. On Chinese platforms, collectivist traits dominate, as noted by Zhang [11], where prayers and calls for peace act as clear signs of group unity. This corresponds closely to the Weibo data analyzed in the research, where high-frequency expressions include keywords such as "pray" and "wish for safety." These expressions demonstrate the tendency of collective emotions. From a Western context perspective, the sarcastic and individualized responses described by Imran et al [62] and Maynard and Greenwood [63] match the study's observations. Sarcasm appeared widespread on X (formerly Twitter), often targeting institutional failures. Wang et al [7] also emphasized that cultural orientation affects how people attribute responsibility and process crisis-related information. The study extends this perspective: X (formerly Twitter), which has lower power distance, shows users openly criticizing authorities. On Weibo, however, higher power distance prevails, and emotions lean more toward worry and collective solidarity. The study by Matsumoto et al [64] emphasized the impact of cultural differences on emotional expression, yet it relied heavily on experimental and questionnaire data, with no inclusion of natural language evidence in the context of social media. Building on this foundation, this study helps fill the gap in empirical natural language research within this field, providing a new supporting dimension for the association between cultural differences and emotional expression.

The research also matches earlier studies when examining discourse structure and interaction styles. Han and Wang [65] and Chen and Yik [66] pointed out that Chinese users' discourse tends to emphasize group harmony and shared emotions. This was confirmed by the topic and semantic network analysis in this study, which revealed cohesive clusters around words such as "pray" and "safety." Discourse on X (formerly Twitter), by contrast, appeared more fragmented. Loosely connected clusters formed around terms such as "damage," "government," and "help," a pattern consistent with Gu et al [67], who stressed the openness and diversity of Western crisis communication. In addition, previous research has shown that in low power distance cultures, Western publics are more likely to hold institutions directly accountable [6,68]. The study contributes to this body of work by demonstrating that sarcasm, specifically, acts as a discursive marker of individualistic cultural values. In addition, as noted in the study by Du et al [69], although Chinese participants scored slightly lower than American participants in understanding sarcasm, they pointed out that "collectivism" is instead associated with better sarcasm comprehension. However, understanding sarcasm does not equate to using sarcasm. This is influenced by cultural contexts and platform rules. Precisely due to this discrepancy, only a minimal amount of sarcasm was observed in the Weibo corpus analyzed in this study. This finding demonstrates the importance of this study's focus on cultural dimensions. When these observations are taken together, the research both confirms and advances previous scholarship. It integrates Hofstede's cultural dimensions with a multimethod analysis, and in doing so, provides a more comprehensive account of the culturally distinctive online responses observed.

Limitations and Future Directions

This study has several limitations. First, the dataset was limited to 2 platforms, which, while influential, cannot fully represent the diversity of cultural and digital contexts. Future research should include additional platforms such as Reddit, YouTube, or regional networks to enhance representativeness. Second, cultural differences were inferred from platform-level data, which may risk oversimplifying individual identities and cross-cultural hybridity. Third, although XLM-RoBERTa and BERTopic provided robust classification, automated labeling relies on linguistic and semantic cues and does not fully incorporate contextual validation. Because of this, some detailed expressions, such as sarcasm, might be undercounted or labeled wrong. Manual validation and mixed methods approaches could strengthen accuracy in future studies. Finally, while the 2024-2025 time frame captured multiple emergencies, it did not account for the effects of event type or severity. Longer longitudinal studies across multiple regions would help clarify how cultural and contextual factors interact in shaping online discourse.

Although event scale, comment volume, and posting time were not the main focus, they could still indirectly influence the observed cross-cultural differences in emotion and discourse patterns.

Conclusions

This study shows that digital emergency discourse is structured by culturally embedded values rather than operating as a neutral information space. Cross-platform analysis demonstrates

XSL•FO
RenderX

systematic associations between cultural dimensions and patterns of emotional expression, narrative focus, and responsibility attribution. By operationalizing cultural dimensions through a mixed methods analysis of large-scale, multilingual, naturally occurring social media data, this study advances infodemiology scholarship from descriptive mapping of online content to theory-informed, mechanism-oriented analysis, beyond survey-based and single-platform approaches. It provides a replicable comparative framework for examining how cultural meaning systems are translated into observable digital traces under conditions of uncertainty, risk, and collective sense-making.

From an applied perspective, the findings offer concrete guidance for multiple stakeholders. For governments and public health authorities, the results highlight the need to tailor risk communication to cultural expectations by emphasizing reassurance, unity, and structured guidance in high–uncertainty-avoidance contexts, and transparency, responsiveness, and accountability in low–power-distance contexts. For international organizations, the findings suggest that global crisis messaging should prioritize culturally neutral framing, shared risk narratives, and locally adaptable templates to minimize misunderstanding across cultural boundaries. For digital platforms, culturally sensitive governance is essential, including rapid rumor suppression in high-anxiety environments and careful differentiation between legitimate public critique and harmful misinformation in low–power-distance contexts. Together, these implications support the design of more equitable, trustworthy, and context-aware infodemic management strategies for future public health and safety emergencies.

Acknowledgments

The authors declare that generative artificial intelligence (ChatGPT 5; OpenAI) was used under complete human oversight to assist with grammatical and lexical issues while writing the paper. All changes were human-verified and further edited by the authors, who accept full responsibility for the final content.

Funding

No external financial support or grants were received for this work.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
JARS (Journal Article Reporting Standards) reporting checklist.
[DOCX File , 30 KB - jmir_v28i1e84648_app1.docx]

References

1. Dubovi I, Tabak I. Interactions between emotional and cognitive engagement with science on YouTube. Public Underst Sci 2021;30(6):759-776 [FREE Full text] [doi: 10.1177/0963662521990848] [Medline: 33546572]
2. Gul S, Shah TA, Ahad M, Mubashir M, Ahmad S, Gul M, et al. Twitter sentiments related to natural calamities. EL 2018;36(1):38-54. [doi: 10.1108/el-12-2015-0244]
3. Li S, Sun X. Application of public emotion feature extraction algorithm based on social media communication in public opinion analysis of natural disasters. PeerJ Comput Sci 2023;9:e1417 [FREE Full text] [doi: 10.7717/peerj-cs.1417] [Medline: 37346715]
4. Halse SE, Tapia A, Squicciarini A, Caragea C. An emotional step toward automated trust detection in crisis social media. Inf Commun Soc 2017;21(2):288-305. [doi: 10.1080/1369118x.2016.1272618]
5. Reuter C, Spielhofer T. Towards social resilience: a quantitative and qualitative survey on citizens' perception of social media in emergencies in Europe. Technol Forecast Soc Change 2017;121:168-180. [doi: 10.1016/j.techfore.2016.07.038]
6. Rahmani M, Muzwagi A, Pumariega AJ. Cultural factors in disaster response among diverse children and youth around the world. Curr Psychiatry Rep 2022;24(10):481-491 [FREE Full text] [doi: 10.1007/s11920-022-01356-x] [Medline: 35953637]
7. Wang B, Liu B, Zhang Q. An empirical study on twitter's use and crisis retweeting dynamics amid Covid-19. Nat Hazards (Dordr) 2021;107(3):2319-2336 [FREE Full text] [doi: 10.1007/s11069-020-04497-5] [Medline: 33469243]
8. Cheng B. Investigating Chinese Microblogging Through a Citizen Journalism Perspective. Sydney: University of Technology Sydney; 2020. URL: https://www.proquest.com/openview/7c45b8844cf6ac9e5d46cbaae1ee0741/1.pdf?pq-origsite=gscholar&cbl=2026366&diss=y

XSL•FO
RenderX

9. Hyvärinen H, Beck R. Emotions trump facts: the role of emotions on social media: a literature review. Published online 2018 Jan 3:1797-1806 [FREE Full text] [doi: 10.24251/HICSS.2018.226]
10. Java A, Song X, Finin T, Tseng B. Why we Twitter: understanding microblogging usage and communities. 2007 Presented at: WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis; August 12, 2007; San Jose, California p. 56-65. [doi: 10.1145/1348549.1348556]
11. Zhang Q. Research on the emotional sentiment of TikTok public opinion in response to sudden natural disasters? A case study of the Zhengzhou severe rainstorm disaster. In: Master's Thesis. China, Henan, Zhengzhou: Zhengzhou University of Aeronautics; 2024.
12. Recommendations on digital interventions for health system strengthening. World Health Organization. 2019. URL: https://www.who.int/publications/i/item/9789241550505 [accessed 2025-09-20]
13. Al-Dmour H, Masa'deh R, Salman A, Abuhashesh M, Al-Dmour R. Influence of social media platforms on public health protection against the COVID-19 pandemic via the mediating effects of public health awareness and behavioral changes: integrated model. J Med Internet Res 2020;22(8):e19996 [FREE Full text] [doi: 10.2196/19996] [Medline: 32750004]
14. Welch V, Petkovic J, Simeon R, Presseau J, Gagnon D, Hossain A, et al. PROTOCOL: Interactive social media interventions for health behaviour change, health outcomes, and health equity in the adult population. Campbell Syst Rev 2018;14(1):1-38 [FREE Full text] [doi: 10.1002/CL2.213] [Medline: 37131397]
15. Al-Dmour H, Masa'deh R, Salman A, Al-Dmour R, Abuhashesh M. The role of mass media interventions on promoting public health knowledge and behavioral social change against COVID-19 pandemic in Jordan. Sage Open 2022;12(1):21582440221082125. [doi: 10.1177/21582440221082125]
16. Ju I, Ohs J, Park T, Hinsley A. Interpersonal communication influence on health-protective behaviors amid the COVID-19 crisis. Health Commun 2023;38(3):468-479. [doi: 10.1080/10410236.2021.1956038] [Medline: 34313168]
17. Mheidly N, Fares J. Leveraging media and health communication strategies to overcome the COVID-19 infodemic. J Public Health Policy 2020;41(4):410-420 [FREE Full text] [doi: 10.1057/s41271-020-00247-w] [Medline: 32826935]
18. Siddiqui S, Singh T. Social media its impact with positive and negative aspects. IJCATR 2016;5(2):71-75. [doi: 10.7753/ijcatr0502.1006]
19. Jakob J, Chan C, Dobbrick T, Wessler H. Discourse integration in positional online news reader comments: patterns of responsiveness across types of democracy, digital platforms, and perspective camps. New Media Soc 2023;26(11):6796-6814. [doi: 10.1177/14614448231183704]
20. Knuth D, Szymczak H, Kuecuekbalaban P, Schmidt S. Social media in emergencies: How useful can they be. 2016 Presented at: 3rd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM); December 13-15, 2016; Vienna, Austria p. 1-7. [doi: 10.1109/ict-dm.2016.7857226]
21. Portwood JD. Book review: international and comparative industrial relations: culture's consequences: international differences in work-related values. ILR Review 1982;36(1):129-130. [doi: 10.1177/001979398203600113]
22. Alsaleh DA, Elliott MT, Fu FQ, Thakur R. Cross-cultural differences in the adoption of social media. JRIM 2019;13(1):119-140. [doi: 10.1108/jrim-10-2017-0092]
23. Wood A, Kleinbaum AM, Wheatley T. Cultural diversity broadens social networks. J Pers Soc Psychol 2023;124(1):109-122. [doi: 10.1037/pspi0000395] [Medline: 35266781]
24. Naab TK, Küchler C. Content Analysis in the Research Field of Online User Comments. In: Oehmer-Pedrazzi F, Kessler SH, Humprecht E, Sommer K, Castro L, editors. Standardisierte Inhaltsanalyse in der Kommunikationswissenschaft – Standardized Content Analysis in Communication Research. Wiesbaden, Germany: Springer Fachmedien Wiesbaden; 2023:441-450.
25. Manuel K, Indukuri K, Krishna P. Analyzing internet slang for sentiment mining. 2010 Presented at: Second Vaagdevi International Conference on Information Technology for Real World Problems; December 09-11, 2010; Warangal, India p. 9-11. [doi: 10.1109/vcon.2010.9]
26. Bi Y. The influence of differences between Chinese and English thinking patterns on Chinese-English translation. Int J Soc Sci Educ Res 2021;4(5):284-289 [FREE Full text] [doi: 10.6918/IJOSSER.202105_4(5).0038]
27. Sener B, Akpinar E, Ataman MB. Unveiling the dynamics of emotions in society through an analysis of online social network conversations. Sci Rep 2023;13(1):14997 [FREE Full text] [doi: 10.1038/s41598-023-41573-9] [Medline: 37696868]
28. Dewaele J, Pavlenko A. Languages and emotions: a crosslinguistic perspective. J Multiling Multicult Dev 2004;25(2-3):93-93. [doi: 10.1080/01434630408666522]
29. Kant N, Puri R, Yakovenko N, Catanzaro B. Practical text classification with large pre-trained language models. arXiv 2018 [FREE Full text] [doi: 10.48550/arXiv.1812.01207]
30. Ou Y, de Bruijn G, Schulz PJ. Social media as an emotional barometer: bidirectional encoder representations from transformers-long short-term memory sentiment analysis on the evolution of public sentiments during influenza A on Sina Weibo. J Med Internet Res 2025;27:e68205 [FREE Full text] [doi: 10.2196/68205] [Medline: 40900625]
31. Multi-class emotion classification for short texts. GitHub. 2020. URL: https://tlkh.github.io/text-emotion-classification/ [accessed 2025-07-16]
32. Anno S, Kimura Y, Sugita S. Using transformer-based models and social media posts for heat stroke detection. Sci Rep 2025;15(1):742 [FREE Full text] [doi: 10.1038/s41598-024-84992-y] [Medline: 39753702]

XSL•FO
RenderX

33. Leburu-Dingalo T, Ntwaagae K, Motlogelwa N, Thuma E, Mudongo M. Application of XLM-RoBERTa for multi-class classification of conversational hate speech. 2022 Presented at: Forum for Information Retrieval Evaluation (FIRE) 2021; December 13-17, 2021; Virtual, India p. 590-595 URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85160755477&partnerID=40&md5=e0fdc9ca66e64b5de9146ab6521d980e
34. Barbieri F, Anke LE, Camacho-Collados J. XLM-T: multilingual language models in Twitter for sentiment analysis and beyond. 2022 Presented at: Proceedings of the Thirteenth Language Resources and Evaluation Conference; 20-25 June, 2022; Marseille, France p. 258-266.
35. Xu WW, Tshimula JM, Dubé È, Graham JE, Greyson D, MacDonald NE, et al. Unmasking the twitter discourses on masks during the COVID-19 pandemic: user cluster-based BERT topic modeling approach. JMIR Infodemiology 2022;2(2):e41198 [FREE Full text] [doi: 10.2196/41198] [Medline: 36536763]
36. Khodeir N, Elghannam F. Efficient topic identification for urgent MOOC Forum posts using BERTopic and traditional topic modeling techniques. Educ Inf Technol 2024;30(5):5501-5527. [doi: 10.1007/s10639-024-13003-4]
37. Zhang Z, Cui L, Fang A, Pan Z, Zhang Z, Zhang H. Information dissemination analysis using a time-weight null model: a case study of Sina micro-blog. IEEE Access 2018;6:71181-71193. [doi: 10.1109/access.2018.2881514]
38. Han D, Wei F, Bai L, Tang X, Zhu T, Wang G. An algorithm of Sina microblog user's sentimental influence analysis based on CNN+ELM model. 2018 Presented at: Proceedings of ELM 2018; November 21-23, 2018; Singapore, Singapore p. 86-97. [doi: 10.1007/978-3-030-23307-5_10]
39. Singh LG, Singh SR. Sentiment analysis of tweets using text and graph multi-views learning. Knowl Inf Syst 2024;66(5):2965-2985. [doi: 10.1007/s10115-023-02053-8]
40. Kada A, Chouikh A, Mellouli S, Prashad AJ, Straus SE, Fahim C. An exploration of Canadian government officials' COVID-19 messages and the public's reaction using social media data. PLoS One 2022;17(9):e0273153 [FREE Full text] [doi: 10.1371/journal.pone.0273153] [Medline: 36054094]
41. Chen Y, Ping S. Effectiveness of emergency information for sudden natural disaster events in the new media era: grounded analysis based on WSR methodology. J Kunming Univ Sci Technol (Nat Sci Ed) 2024;49(02):182-193 [FREE Full text] [doi: 10.16112/j.cnki.53-1223/n.2024.02.292]
42. Hasan M, Islam L, Jahan I, Meem SM, Rahman RM. Natural language processing and sentiment analysis on Bangla social media comments on Russia–Ukraine war using transformers. Vietnam J Comp Sci 2023;10(03):329-356. [doi: 10.1142/s2196888823500021]
43. Babalola O, Ojokoh B, Boyinbode O. Comprehensive evaluation of LDA, NMF, and BERTopic's performance on news headline topic modeling. J. Comput. Theor. Appl 2024;2(2):268-289. [doi: 10.62411/jcta.11635]
44. Ma L, Chen R, Ge W, Rogers P, Lyn-Cook B, Hong H, et al. AI-powered topic modeling: comparing LDA and BERTopic in analyzing opioid-related cardiovascular risks in women. Exp Biol Med (Maywood) 2025;250:10389 [FREE Full text] [doi: 10.3389/ebm.2025.10389] [Medline: 40093658]
45. Hofstede G. Dimensionalizing cultures: the Hofstede model in context. ORPC 2011;2(1):1-26. [doi: 10.9707/2307-0919.1014]
46. Masoud R, Liu Z, Ferianc M, Treleaven P, Rodrigues M. Cultural alignment in large language models: an explanatory analysis based on Hofstede's cultural dimensions. arXiv. 2023. URL: https://arxiv.org/abs/2309.12342 [accessed 2026-01-11]
47. Alqarni A. Hofstede's cultural dimensions in relation to learning behaviours and learning styles: a critical analysis of studies under different cultural and language learning environments. J Lang Linguist Stud 2022;18(2022):721-739 [FREE Full text]
48. Jan J, Alshare KA, Lane PL. Hofstede's cultural dimensions in technology acceptance models: a meta-analysis. Univ Access Inf Soc 2022;23(2):717-741. [doi: 10.1007/s10209-022-00930-7]
49. Köksal A, Özgür A. Twitter dataset and evaluation of transformers for Turkish sentiment analysis. 2021 Presented at: 29th Signal Processing and Communications Applications Conference (SIU); June 09-11, 2021; Istanbul, Turkey p. 1-4. [doi: 10.1109/siu53274.2021.9477814]
50. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
51. von EE, Altman D, Egger M. The STROBE reporting checklist. EQUATOR Network. 2025. URL: https://resources.equator-network.org/reporting-guidelines/strobe/strobe-checklist.docx [accessed 2025-10-28]
52. Gaikwad A, Belhekar P, Kottawar V. Advancing multilingual sentiment understanding with XGBoost, SVM, and XLM-RoBERTa. 2025 Presented at: Proceedings of the 5th International Conference on Data Science, Machine Learning and Applications; Volume 2; December 15-16, 2023; Hyderabad, India p. 990-1000. [doi: 10.1007/978-981-97-8043-3_155]
53. Rasool A, Aslam S, Hussain N, Imtiaz S, Riaz W. nBERT: harnessing NLP for emotion recognition in psychotherapy to transform mental health care. Information 2025;16(4):301. [doi: 10.3390/info16040301]
54. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: Proceedings of NAACL-HLT 201; June 2-7, 2019; Minneapolis, Minnesota p. 4171-4186. [doi: 10.18653/v1/2024.naacl-short]
55. Grice HP. Logic and conversation. In: Speech Acts. Leiden, The Netherlands: Brill; 1975:41-58.

XSL•FO
RenderX

56. McNamee P, Duh K. An extensive exploration of back-translation in 60 languages. 2023 Presented at: Findings of the Association for Computational Linguistics: ACL 2023; July, 2025; Toronto, Canada p. 8166-8183.
57. de Groot M, Aliannejadi M, R. Haas M. Experiments on generalizability of BERTopic on multi-domain short text. arXiv 2022 [FREE Full text] [doi: 10.48550/arXiv.2212.08459]
58. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794 2022 [FREE Full text] [doi: 10.48550/ARXIV.2203.05794]
59. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. 2017 Presented at: ICML'17: Proceedings of the 34th International Conference on Machine Learning - Volume 70; August 6-11, 2017; Sydney, Australia p. 3319-3328 URL: https://proceedings.mlr.press/v70/sundararajan17a.html
60. Sanyal S, Ren X. Discretized integrated gradients for explaining language models. 2021 Presented at: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; November 7-11, 2021; Punta Cana, Dominican Republic p. 10285-10299. [doi: 10.18653/v1/2021.emnlp-main.805]
61. Ortigossa ES, Dias FF, Barr B, Silva CT, Nonato LG, Ortigossa ES. T-explainer: a model-agnostic explainability framework based on gradients. IEEE Intell Syst 2025;40(5):34-44. [doi: 10.1109/mis.2025.3564330]
62. Imran AS, Daudpota SM, Kastrati Z, Batra R. Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets. IEEE Access 2020;8:181074-181090. [doi: 10.1109/access.2020.3027350]
63. Maynard D, Greenwood M. Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. 2014 Presented at: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14); December 29, 2025; Reykjavik, Iceland p. 4238-4243 URL: https://eprints.whiterose.ac.uk/130763/
64. Matsumoto D, Seung Hee Yoo, Fontaine J. Mapping expressive differences around the world. J Cross-Cult Psychol 2008;39(1):55-74. [doi: 10.1177/0022022107311854]
65. Han X, Wang J. Using social media to mine and analyze public sentiment during a disaster: a case study of the 2018 Shouguang city flood in China. IJGI 2019;8(4):185. [doi: 10.3390/ijgi8040185]
66. Chen X, Yik M. The emotional anatomy of the Wuhan lockdown: sentiment analysis using Weibo data. JMIR Form Res 2022;6(11):e37698 [FREE Full text] [doi: 10.2196/37698] [Medline: 36166650]
67. Gu M, Guo H, Zhuang J. Social media behavior and emotional evolution during emergency events. Healthcare (Basel) 2021;9(9):1109 [FREE Full text] [doi: 10.3390/healthcare9091109] [Medline: 34574883]
68. Oz T, Havens R, Bisgin H. Assessment of blame and responsibility through social media in disaster recovery in the case of #flintwatercrisis. Front Commun 2018;3:45. [doi: 10.3389/fcomm.2018.00045]
69. Du Y, He H, Chu Z. Cross-cultural nuances in sarcasm comprehension: a comparative study of Chinese and American perspectives. Front Psychol 2024;15:1349002 [FREE Full text] [doi: 10.3389/fpsyg.2024.1349002] [Medline: 38445055]

Abbreviations

BERTopic: Bidirectional Encoder Representations from Transformers Topic
c-TF-IDF: class-based term frequency-inverse document frequency
CV: coefficient of variation
IG: Integrated Gradients
ISO: International Organization for Standardization
JARS: Journal Article Reporting Standards
XLM-RoBERTa: Cross-lingual Language Model–Robustly optimized BERT approach

Edited by S Brini; submitted 23.Sep.2025; peer-reviewed by P Babatuyi, A Rasool, BAF Assunção; comments to author 14.Oct.2025; accepted 24.Dec.2025; published 20.Jan.2026.

Please cite as:
Guo X, Fan Y, Guo Y
Public Emotional and Thematic Responses to Major Emergencies on Social Media, 2024-2025: Cross-Sectional Convergent Mixed Methods Study
J Med Internet Res 2026;28:e84648
URL: https://www.jmir.org/2026/1/e84648
doi:10.2196/84648
PMID:

©Xingrong Guo, Yiqian Fan, Yiming Guo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Global Sentiment Toward Health AI at the Dawn of the ChatGPT Era: Empirical Analysis of Twitter (X) Discourse

Lily Minh Wass[1,2,3], BSc; Zhengdong Wu[1,2,3], MPhil; José Vizoso[3], MSc; Joseph T Wu[1,2,4,5*], PhD; Leesa Lin[1,2,3,6*], PhD

[1]Laboratory of Data Discovery for Health, Hong Kong, China (Hong Kong)

[2]WHO Collaborating Centre for Infectious Disease Epidemiology and Control, School of Public Health, Li Ka Shing Faculty of Medicine, University of Hong Kong, Hong Kong, China (Hong Kong)

[3]Moonrise Initiative, Hong Kong, China (Hong Kong)

[4]The University of Hong Kong – Shenzhen Hospital, Shenzhen, China

[5]The Hong Kong Jockey Club Global Health Institute, Hong Kong, China (Hong Kong)

[6]Department of Infectious Disease Epidemiology and Dynamics, London School of Hygiene & Tropical Medicine, London, England, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Leesa Lin, PhD
Department of Infectious Disease Epidemiology and Dynamics
London School of Hygiene & Tropical Medicine
Keppel St
London, England, WC1E 7HT
United Kingdom
Phone: 44 20 7636 8636
Email: Leesa.Lin@lshtm.ac.uk

Abstract

Background: Artificial intelligence (AI) is increasingly proposed for use in health and health care systems. Beyond technical performance, public perceptions and affective responses influence whether AI technologies are accepted and adopted in real-world contexts. Social media platforms such as X (formerly Twitter) provide large-scale, real-time insight into public discourse surrounding emerging technologies, yet remain underused for examining how health AI is discussed, evaluated, and emotionally framed.

Objective: This study aimed to develop and apply large language model (LLM)–based methods for exploratory social listening on health AI. This is the first study to map large-scale sentiment, emotional expressions, and confidence-related signals in online discussions of applications of AI to health.

Methods: We collected 786,750 English-language posts from X (Twitter) published between January 1 and December 5, 2023, using health- and AI-related keywords. We benchmarked an LLM-based annotation framework by using OpenAI's GPT-3.5-Turbo and GPT-4, comparing model classifications with trained human researchers. Annotations included overall sentiment and 6 evaluative domains frequently referenced in the literature surrounding attitudes toward health AI—usefulness, safety, privacy, ethics, quality, and trust. After cleaning, GPT-3.5-Turbo used the best-performing prompts to label 388,009 posts. A subset (n=268,347) was further analyzed using Emollama-7b, an open-source model fine-tuned from Meta's LLaMA2-7B, for emotion detection, and latent Dirichlet allocation for thematic analysis. Comparisons were made across World Health Organization regions.

Results: Compared against human annotations, optimized prompts achieved weighted F_1-scores above 0.60 across evaluative domains and sentiment classification. Global discourse about health AI was 65.26% (95% CI 65.11%-65.4%) positive and 83.62% (95% CI 83.48%-83.76%) emotionally optimistic, although substantial regional variation was observed in sentiment (P<.001). The Eastern Mediterranean and South-East Asia regions expressed significantly higher levels of positive sentiment and evaluative agreement in the studied features of health AI, alongside frequent discussion of the tech industry and commercial development. In comparison, the Western Pacific region expressed lower confidence and significantly more mentions of research topics (19.27%, 95% CI 18.5%-20.07%). Privacy was the most prominent global concern, with 33.31% (95% CI 32.98%-33.66%) of privacy-related posts expressing perceived risks. In the Region of the Americas, 18.19% (95% CI 17.92%-18.44%) of posts discussed algorithms and data governance, significantly higher than overall.

XSL•FO
RenderX

Conclusions: This study offers the first systematic characterization of online health AI discourse at scale, mapping stances toward key features of AI, emotional tone, and discussion topics across regions. LLM-powered social listening is demonstrated as a feasible approach for identifying dominant narratives and regionally distinct concerns, capable of surfacing opinions absent from traditional media. This can extend to studying discourse on other evolving health technologies where public surveying is limited. While methodological refinement and multilingual expansion are needed, this framework can inform timely policy development, risk communication, and responsible health AI governance.

(J Med Internet Res 2026;28:e80346) doi:10.2196/80346

KEYWORDS

AI; artificial intelligence; health care; social listening; sentiment analysis; ChatGPT; trust

Introduction

The release of large language model (LLM) chatbots like ChatGPT (OpenAI) in 2022 marked a turning point in the accessibility of artificial intelligence (AI), sparking public and research interest alike. By 2023, the rapid progress of LLMs extended to health care applications, leveraging the explosion of online data for training sophisticated AI models [1]. Although thousands of studies on AI in health care are published annually [2], most remain at the proof-of-concept or pilot stage. Progress toward routine clinical deployment has been comparatively slow, reflecting unresolved challenges in validation, generalizability, regulatory approval, workflow integration, and safe implementation in high-stakes clinical environments [3,4].

As with vaccination and other system-level medical interventions, the impact of health AI depends not only on technical performance but on the stability of the social systems that maintain its use. In this sense, public and professional confidence functions as a form of infrastructure—less visible than algorithms or regulatory approvals, yet equally foundational to safe and scalable implementation. Without durable trust in safety, fairness, accountability, and data governance, even high-performing AI tools may fail to achieve uptake.

Although AI's empirical capabilities are advancing rapidly, its translation into routine care is mediated by perception, politics, and institutional legitimacy [5,6]. Evidence shows that clinicians and the public hold complex, sometimes ambivalent views of health AI, shaped by concerns about safety, bias, accountability, ethics, and data use [6-9]. In high-stakes clinical settings, where consequences are immediate and personal, perceived risks can outweigh demonstrated benefits. Recognizing confidence as infrastructure implies that it must be actively maintained. Scholars have therefore called for continuous, scalable systems to monitor public and professional confidence in health AI over time—analogous to vaccine-confidence surveillance—so that governance, communication, and implementation strategies can adapt to emerging risks and concerns [10].

One promising source of data for examining evolving perceptions of health AI is online platforms, where AI-related discussions surged following the release of ChatGPT in late 2022 [11]. Social media has surpassed traditional media as a major source of information, exerting considerable influence on public opinion formation and politics [12]. Platforms provide sites for exchange and exposure to new ideas in a public forum, yet not all ideas are amplified equally. Visibility on public forums is shaped by engagement-based algorithms, which can tend to reward emotionally charged content.

A growing body of research indicates that content that invokes negative sentiments or high-arousal emotions, such as anger and sadness, is more likely to be perceived as believable and to spread widely [13,14]. This emotional content has boosted the dissemination of misinformation and fake news [15,16]. Because public awareness of health AI is generally low, even among some medical professionals [7,8,17], information consumers might rely on their emotional responses over reasoning when consuming health AI-related media, a choice associated with greater susceptibility to misinformation [18].

Measuring attitudes toward AI is itself a conceptual challenge. Unlike many previous technological innovations, AI systems exhibit anthropomorphic traits, broad applications, and inaccessible "black box" decision-making processes [19,20] that users struggle to conceptualize. Consequently, theoretical approaches to understanding confidence in AI span multiple domains, including technical performance, ethical governance, legal accountability, and societal alignment, without a clear consensus on priorities [20,21]. Support for one attribute of a health AI system, such as predictive accuracy, does not necessarily imply endorsement of others, such as ethical data use or privacy protections [21]. Therefore, a large-scale investigation into the cues in health AI confidence and emotion shared in online discourse is needed to better understand the information environment shaping public perceptions and governance challenges.

Achieving such large-scale insight presents additional methodological obstacles. Previous studies examining social media discourse on health AI have used manual content analysis, restricting analyses to a few thousand posts [10,22]. To address scalability, machine learning approaches like support vector machines have also been explored, but these models often exhibit poor generalizability, particularly when applied to health-related social media data [23,24]. Common natural language processing (NLP) techniques struggle with domain-specific language, sarcasm, and context-dependent framing, and require large, manually annotated datasets and extensive tuning to achieve acceptable performance [24,25]. These constraints in methodology and performance limit the feasibility of timely and large-scale assessments of public opinion.

LLMs have been proposed as a more adaptive and scalable alternative. By leveraging transformer-based architectures,

XSL•FO
RenderX

LLMs can capture complex linguistic nuances and adapt to different contexts, while reducing reliance on task-specific training data [26]. Existing research has found LLMs to outperform traditional methods with significantly less reliance on manually annotated data, further reducing the time and resource burden of traditional approaches [27,28]. However, empirical applications of LLMs to the analysis of public discourse on health AI remain limited, while reported performance on health-related social media data has been inconsistent [29,30].

In response to these gaps, this study has 2 primary objectives. First, we evaluate the performance of 2 widely used GPT-based LLMs for annotating short, health AI–related social media posts, using zero-shot, few-shot, and chain-of-thought prompting. This analysis offers a reference point that can be revisited as language models and prompting strategies continue to evolve. Second, drawing on the best-performing annotation approach, we conduct a large-scale, descriptive social listening analysis to characterize how health AI is discussed online. Specifically, we examine the distribution of sentiments, discrete emotions, and thematic content expressed in posts related to health AI, and assess variation across geographic sources of posts. Together, these contributions demonstrate the feasibility and limitations of using LLMs for scalable analysis of public discourse on a rapidly evolving topic.

Methods

Overview

A multipronged approach of LLM and NLP tools was used to assess the global sample of X (formerly Twitter) discussions regarding health AI. Post sentiment and positions on health AI confidence were analyzed using an LLM-prompting strategy. Performance was measured using a benchmark dataset and a prompt testing pipeline. Previously validated LLM and NLP methods were applied for emotion detection and discussion themes [31,32].

Sentiment Analysis Framework

We grounded our exploratory social-listening approach to health AI in methodological advances from vaccine-confidence surveillance, adapting validated, double-coded annotation frameworks that have been applied to large-scale social media data over the past decade. This approach leverages established protocols for construct definition, coder adjudication, and reliability benchmarking to enable systematic monitoring of evolving public attitudes [33-35]. These studies have adapted the World Health Organization's (WHO) Confidence, Complacency, and Convenience ("3 Cs") model of vaccine hesitancy [36] to translate a complex decision-making model into analyzable indicators of online discourse (eg, expressions that vaccines are not safe). With this precedent, we reviewed existing studies and the models used to measure attitudes toward AI, such as the prominent technology acceptance model [37] and its extensions. From the literature, we identified 6 features of health AI recurrently discussed by experts and the public (Table 1). Based on their frequency in the literature, we expected these dimensions to be similarly reflected in an exploratory, high-level analysis of how health AI is discussed online.

Table 1. Annotation categories for posts about health artificial intelligence on X from January to December 2023. Performance measured using GPT-3.5-Turbo and GPT-4 models in September 2024.

AI[a] confidence	Definition	GPT-3.5-Turbo performance, weighted F_1-score (95% CI)	GPT-4 performance, weighted F_1-score (95% CI)
Sentiment	The post is positive, negative, or neutral toward AI	0.71 (0.62-0.79)	0.74 (0.67-0.81)
Safety [17,21,38]	This post indicates AI is safe	0.64 (0.56-0.72)	0.79 (0.72-0.86)
Usefulness [6,7,38-41]	This post indicates AI is useful	0.69 (0.60-0.76)	0.77 (0.69-0.84)
Trustworthiness [7,17,37]	This post indicates AI is trustworthy	0.61 (0.52-0.70)	0.63 (0.56-0.71)
Privacy [6-8,17,37,38,40]	This post indicates AI respects privacy	0.66 (0.57-0.74)	0.76 (0.68-0.82)
Ethics [8,17,38-40]	This post indicates AI is ethical	0.63 (0.55-0.71)	0.71 (0.62-0.78)
Quality [6,7,17,38,40]	This post indicates AI is of good quality	0.70 (0.61-0.78)	0.76 (0.67-0.83)

[a]AI: artificial intelligence.

LLM Performance Validation

We implemented an LLM-based sentiment analysis and validation pipeline to assess model performance in processing health AI-related text, aiming to improve communication strategies in health care. Using the operational definitions in Table 1, researchers manually labeled a benchmark dataset of 150 posts per health AI domain and overall sentiment variable. Benchmark posts were randomly selected from a separate set of health AI-related posts shared outside of the study period. Posts were labeled as "true," "false," or "irrelevant" for their stance within each domain (eg, True: Health AI is safe, False: Health AI is not useful, and Irrelevant: Ethics of Health AI is not mentioned) and as "positive," "negative," or "neutral" for overall sentiment. In total, 50 posts were labeled per class.

XSL•FO
RenderX

The benchmark size was selected to support comparative evaluation of model performance under realistic annotation constraints, in line with previous evaluations of LLMs for short text annotations [42-44]. Additional concepts of "importance" and "accessibility" were initially included as dimensions of confidence, but due to scarcity in the X sample of "false" labeled posts (ie, Health AI is not accessible and Health AI is not important), these domains were excluded. This resulted in 1050 total annotations for 6 domains and overall sentiment.

The performance of LLMs GPT-3.5 Turbo (OpenAI) and GPT-4 (OpenAI) was assessed using 4 prompt types—zero-shot, few-shot [45], zero-shot chain-of-thought (CoT), and few-shot CoT [46]. Prompts instructed GPT to classify benchmark posts based on their expressed opinion on the 6 health AI domains and their overall sentiment toward the technology according to the format in Table 2.

Table 2. Template of zero-shot, few-shot, zero-shot chain of thought, and few-shot chain of thought prompts passed to GPT-3.5-Turbo and GPT-4 to annotate benchmark dataset. Adjective refers to health artificial intelligence concepts defined in Table 1.

Prompt type	Format
Zero-shot	Classify tweets as being in agreement (true), disagreement (false), or irrelevant with the belief that AI[a] is [adjective].
Few-shot	Tweet: Example Tweet 1 Answer: True/False/Irrelevant Tweet: Example Tweet 2 Answer: True/False/Irrelevant Tweet: Example Tweet 3 Answer: True/False/Irrelevant Instruction: Classify tweets as being in agreement (true), disagreement (false), or irrelevant with the belief that AI is [adjective].
Zero-shot CoT[b]	Classify tweets as being in agreement (true), disagreement (false), or irrelevant with the belief that AI is [adjective]. Let's go step by step.
Few-shot CoT	Tweet: Example Tweet 1 Thought: Example Explanation 1 Answer: True/False/Irrelevant Tweet: Example Tweet 2 Thought: Example Explanation 2 Answer: True/False/Irrelevant Tweet: Example Tweet 3 Thought: Example Explanation 3 Answer: True/False/Irrelevant Instruction: Classify tweets as being in agreement (true), disagreement (false), or irrelevant with the belief that AI is [adjective].

[a]AI: artificial intelligence.

[b]CoT: chain of thought.

This created a dataset of GPT-generated labels for all benchmark posts.

We defined the performance metric as weighted F_1-score (wF_1-score), calculated for the classification of each domain as in the equation below, where *a*, *b*, and *c* refer to the number of posts classified to each category by human annotations, and *True*, *False*, *Irrelevant*, refer to the F_1-scores for each classification):

This compared the predicted (LLM) labels of "true," "false," and "irrelevant" for each domain to the true (researcher) labels for the same posts. Model performance was calculated using nonparametric bootstrapping of 1000 resamples with replacement from each domain's n=150 annotated items, resulting in domain-specific mean wF_1-score and 95% CIs [47]. Prompt ablation is reported in Figure S3 and Table S3 of Multimedia Appendix 1.

We prespecified a threshold of wF_1-score>0.6 as the minimum acceptable agreement with human annotations. This threshold was selected based on the historic performance of sentiment models and emergent applications of LLMs, like GPT-3.5-Turbo for annotation [24,30,48], along with the inherent subjectivity of the described task.

GPT-3.5 Turbo was chosen for cost efficiency to label the full dataset (Table 1). Prompts with wF_1-scores>0.6 were passed to the model for labelling of the full dataset (Multimedia Appendix 2). Verbatim prompt variations, along with model parameters (eg, version, temperature, and random seed), are specified in Multimedia Appendix 2.

Data Curation

Data for this study were sampled from X, selected because of its status as a major site of health communications for health-related institutions, academia, companies, and interested individuals [49], with over 550 million monthly users in 2023 [50].

XSL•FO
RenderX

X posts from January 1, 2023 to December 5, 2023 – an 11-month period when ChatGPT site visits grew by 1 billion [51], were collected using AI-related keywords. Keywords included popular LLMs like "ChatGPT," techniques such as "natural language processing," and AI domains like "conversational AI," along with the term "health." For the full Boolean search string, refer to Figure S1 in Multimedia Appendix 1. Tweets in English were included. As LLM performance for sentiment classification tasks is variable for low- and mid-resource languages [52], consistency through English texts was opted for to avoid introducing variability in processing and interpretation of model annotations. In addition, a previous analysis of ChatGPT-related tweets from an overlapping time period found 72% were in English [11]. In total, 786,750 posts were selected, including original posts, retweets, and quoted tweets – posts responding to an original post.

Qualitative review of the sample identified posts shared from general users, news sources, academic institutions, health organizations, influencers – especially tech leaders, and private companies. Past estimates suggest between 9% to 15% of active X users have automated accounts [53]. This reflects the reality of social media as a site of information exchange between individuals, organizations, and even bots.

Data Cleaning

For sentiment analysis, duplicates were removed from the raw dataset, leaving 394,533 posts.

For emotion analysis and topic modeling, we filtered out posts that only contained hashtagged text or contained 4 or more stock tickers (eg, $XXX). The following preprocessing steps were then applied: emoticons, usernames (starting with "@"), URLs, special characters, "retweet," "QT" (quote tweet), and stock ticker patterns were removed. Multiple spaces, tabs, and newlines were collapsed into a single space. Duplicates were dropped again, leaving 268,347 posts (Figure S2 in Multimedia Appendix 1).

For topic modeling, the text was converted to lowercase. Tokenization was performed with spaCy (English model en_core_web_lg). We retained only alphabetic tokens, applied lemmatization and lowercasing, and removed tokens that were stopwords from the spaCy default list or had part of speech tags ADV, PRON, CCONJ, PUNCT, PART, DET, ADP, SPACE, NUM, or SYM. Common tokens were standardized, for example, "ml" to "machine learning."

Emotion Analysis

Compared with research on perspectives toward AI, the role of emotions in the expression and formation of these attitudes is understudied. Previous work has found that individuals with strong anger and less hope show more support for AI upon exposure to information sources [54]. To explore the emotional content of discussions, we used "Emollama-7b"—an open source LLM finetuned from LLaMA2-7B for affective analysis.

Emotion analysis used the model's 11 recommended emotional classifications. This includes 6 emotions previously studied for their implication in decision-making, including judgments of science [55]. A previous benchmark found that Emollama-7b outperforms ChatGPT and GPT-4 for the emotion classification task: "Given a tweet, classify it as 'neutral or no emotion' or as one, or more, of eleven given emotions (anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust) that best represent the mental state of the tweeter" [32,56]. This identical prompt was used to instruct Emollama-7b to label n=268,347 X posts. Default parameters were used based on the model downloaded from Hugging Face [57]. We clarify that "trust" in the emotion analysis refers to an affective state label from the emotion model and is conceptually distinct from attribute-level trustworthiness of health AI measured in the sentiment analysis.

Topic Modeling

To characterize the thematic dimensions of health AI discussed online, we applied latent Dirichlet allocation (LDA), an unsupervised probabilistic topic-modeling approach in which documents (X posts) are represented as mixtures of latent topics, and topics as distributions over words [58]. LDA has been widely used to summarize large-scale health-related social media corpora [59-62].

We implemented LDA using the Gensim LDAMulticore function in Python. Tokens occurring more than 10 times across the corpus and in fewer than 50% of posts were retained in the dictionary to reduce sparsity and remove highly ubiquitous terms. Candidate models ranging from k=3 to k=12 topics were evaluated using the Cv coherence metric. For each model, the training procedure used 10 passes over the corpus and 50 iterations. Hyperparameters (alpha and beta) were retained at default values.

Because LDA inference is stochastic, we assessed topic stability by refitting the k=10 model across 4 random initialization seeds and comparing the consistency of high-probability keywords and overall thematic structure. Although some variation in keyword ordering occurred across runs, the main thematic groupings were consistent. The final model (k=10) was selected based on coherence trends across candidate models and qualitative interpretability, balancing thematic granularity with semantic distinctiveness. Furthermore, 3 researchers (LMW, ZW, and LL) then iteratively refined topic labels by synthesizing the top 10 highest-probability words per topic and validating each label against random samples of posts [63]. Topics were subsequently summarized by WHO region and emotional distribution.

In-Depth Analysis of Global Public Confidence in Health AI

GPT-3.5 Turbo labeled sentiments and health AI confidence in X posts. After cleaning, 394,533 posts were submitted to GPT-3.5 Turbo for sentiment analysis. Due to OpenAI's policy prohibiting discussions of violence, illegal activity, and adult content [64], several thousand explicit posts were excluded, resulting in a labeled dataset of 388,009. Posts without a GPT label were excluded from the analysis of each confidence domain.

GPT-labeled posts were grouped by WHO region based on user location, with excluded areas (eg, Hong Kong and Vatican City)

XSL•FO
RenderX

added as needed (Table S1 in Multimedia Appendix 1). The top 5 countries with the most posts in each region are listed in Table S2 in Multimedia Appendix 1. Posts annotated as "positive," "neutral," and "negative" in sentiment were mapped to 1, 0, and –1, respectively. Averages were calculated as arithmetic means by WHO region, dividing the number of posts in each classification by the total number of posts in each region. 95% CIs are reported using a nonparametric bootstrap with n=1000 resamples. Reach was analyzed as the number of unique users who saw a post. This metric is provided from Meltwater's (Meltwater US Holdings Inc) application program interface [65], which was used to extract posts from X based on our search criteria. Analyses and visualizations were conducted in Python (v3.11.8).

Ethical Considerations

This study analyzed publicly available posts from X (formerly Twitter) and does not constitute human subjects research, as no private or identifiable information was collected and no interaction with individuals occurred. Accordingly, formal ethical review and informed consent were not required. All data were accessed through an authorized third-party data provider and handled in accordance with X's terms of service. No personally identifiable information was reported; all analyses were conducted at the aggregate level, and no individual users are identified or identifiable in the manuscript or supplementary materials. As this study did not involve human participants, compensation was not applicable.

Results

Sentiment Analysis

Frequency and overall sentiment of posts by WHO region are summarized in Table 3. In total, 42 countries shared over 500 posts in the study period. Geographically, most posts came from the Region of the Americas (AMR), accounting for 121,201 (31.38%) posts, followed by the European Union with 57,234 (14.82%) posts. However, a large portion of posts (n=153,289, 39.68%) came from users who did not disclose their location on X.

Table 3. Sentiments about health artificial intelligence on X from January 1 to December 3, 2023. Percentages are calculated using the n of each region as the denominator.

	Overall (n=386,287)	AFR[a] (n=10,624)	AMR[b] (n=121,201)	EMR[c] (n=6375)	European Union (n=57,234)	SEAR[d] (n=24,888)	WPR[e] (n=12,676)	Undisclosed (n=153,289)
Positive								
n	252,082	7179	79,284	4736	37,848	16,974	7608	98,453
Sample proportion[f],% (95% CI)	65.26 (65.11-65.4)	67.57 (66.63-68.48)	65.42 (65.16-65.67)	74.29 (73.13-75.34)	66.13 (65.76-66.5)	68.2 (67.61-68.74)	60.02 (59.2-60.85)	64.23 (65.11-65.4)
Negative								
n	36,505	721	11,565	227	4639	930	1217	17,206
Sample proportion, % (95% CI)	9.45 (9.36-9.54)	6.79 (6.31-7.26)	9.54 (9.37-9.71)	3.56 (3.11-4)	8.11 (7.87-8.32)	3.74 (3.51-3.98)	9.60 (9.1-10.07)	11.22 (11.07-11.38)
Neutral								
n	97,700	2724	30,352	1412	14,747	6984	3851	37,630
Sample proportion, % (95% CI)	25.29 (25.15-25.43)	25.64 (24.83-26.53)	25.04 (24.79-25.29)	22.15 (21.08-23.12)	25.77 (25.4-26.11)	28.06 (27.47-28.59)	30.38 (29.55-31.18)	24.55 (24.34-24.77)
Mean sentiment score (SD; 95% CI)	0.5561 (0.6600; 0.5581-0.5602)	0.6079 (0.6117; 0.5960-0.6194)	0.5587 (0.6614; 0.5551-0.5625)	0.7073 (0.5275; 0.6947-0.7194)	0.5802 (0.6369; 0.5749-0.5856)	0.6446 (0.5512; 0.6377-0.6515)	0.5042 (0.6648; 0.4924-0.5151)	0.5300 (0.5267-0.5335)

[a]AFR: African Region.

[b]AMR: Region of the Americas.

[c]EMR: Eastern-Mediterranean Region.

[d]SEAR: South-East Asia Region.

[e]WPR: Western Pacific Region.

[f]Sample proportion: Number of posts/total.

The majority of posts were positive (n=252,082, 65.26%), followed by neutral (n=97,700, 25.29%), and negative (n=36,505, 9.45%) about health AI. The sentiment distribution of WHO regions revealed significant differences (χ^2=2274.1, P<.001). The Eastern Mediterranean Region (EMR) was most positive about health AI, with nearly three-quarters of posts expressing positive sentiments. While the Western Pacific Region (WPR) had the lowest sentiment score of 0.50, driven

XSL•FO
RenderX

by a higher percentage of neutral sentiment posts (3851/12,676, 30.38%), posts from an undisclosed region shared the highest proportion of negative comments about health AI, at 11.22% (37,630/153,289).

Confidence in Health AI

The majority of posts expressed attitudes in favor of health AI. "Privacy" showed the biggest divide in opinions for or against the technology, with 45,977/68,943 (66.69%; 95% CI 66.34%-67.02%) of privacy-related posts affirming health AI's respect for privacy, while 22,966/69,943 (33.31%; 95% CI 32.98%-33.66%) of related posts were classified as indicating health AI threatened privacy. This concern was highest in AMR and undisclosed regions, with averages in support of privacy falling significantly below the overall sample proportion. In AMR, 14,803/22,660 (65.33%; 95% CI 64.7%-65.96%) of related posts supported health AI's respect for privacy, while 17,285/26,828 (64.43%; 95% CI 63.88%-64.99%) of posts expressed the same belief in undisclosed regions (Figure 1). Trustworthiness was the next most divided topic, but 136,700/168,041 (81.35%; 95% CI 81.16%-81.53%) of trust-related posts still voiced trust toward health AI. By comparison, the usefulness of health AI was nearly unanimous. There were 291,457 in favor of usefulness out of 315,528 related posts, totaling 92.37% (95% CI 92.28%-92.47%) of posts in the overall sample in support of health AI's usefulness. Support within EMR was significantly higher at 5321/5479 (97.12%; 95% CI 96.7%-97.57%) of relevant posts.

Figure 1. Sample proportion (95% CI) of X posts that expressed attitudes in favor of health AI attributes, grouped by WHO region (January 1 to December 3, 2023). AFR: African Region; AMR: Region of the Americas; EMR: Eastern-Mediterranean Region; EUR: European Union; SEAR: South-East Asia Region; WPR: Western Pacific Region; UND: Undisclosed.

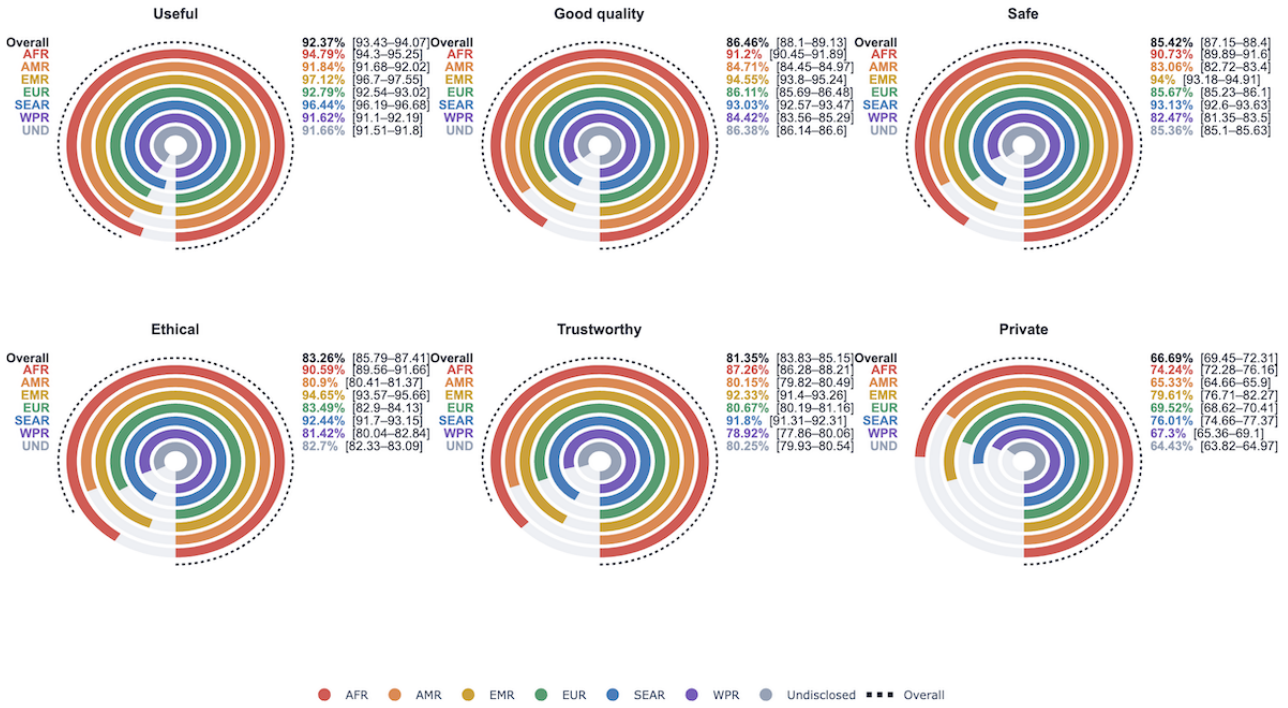

EMR reported the highest percentage of posts in favor of health AI across all 6 domains. For "usefulness," "quality," "safety," and "trust," WPR had the lowest percentage of posts in favor, but rose above the overall sample proportion for "privacy." AMR consistently voiced lower support for health AI concepts than the overall sample. Full results of the sentiment analysis are reported in Table S5 in Multimedia Appendix 1.

Emotion Analysis and Topic Modeling

Considering the 11 possible emotions explored in discussions of health AI, out of the total sample, 224,386 (83.62%; 95% CI 83.48%-83.76%) posts contained "optimism." This was followed by "anticipation," found in 165,300 (61.6%; 95% CI 61.42%-61.77%) posts. "Optimism" was nearly 10 percentage points higher than the overall sample in the South-East Asia Region (SEAR), found in 15,860 posts (91.79%; 95% CI 91.41%-92.2%) and EMR in 4279 (91.63%; 95% CI 90.79%-92.4%) posts. Moreover, 7563 (80.84%; 95% CI 80.07%-81.65%) posts from WPR contained optimism, which was the lowest regional proportion. WPR expressed the highest proportion of all negative emotions: pessimism, disgust, sadness, anger, and fear, at a rate consistently above the overall sample. This included 3021 (32.29%; 95% CI 31.34%-33.23%) posts from the region that contained pessimism. Full emotional results are available in Table S7 in Multimedia Appendix 1. Posts with "anticipation" reached the most accounts on average (8604.54, 95% CI 8099.95-9136.3), followed by trust (8596.68, 95% CI 7396.50-10,074.38), and fear (7721.15, 95% CI 6455.61-9146.52).

The 10 topics generated by the LDA topic model reflected the diversity of considerations when it comes to health AI, spanning commercial excitement from product developments to ethical considerations of responsible AI development. From the top 5 most popular topics (Table 4), the first was "tech industry," characterized by terms like "business," "digital," and "innovation." These tweets focused on the commercial potential of the digital health revolution, including direct marketing from

XSL•FO
RenderX

industry stakeholders. This topic was most common within EMR, found in 3594 (26.55%; 95% CI 25.33%-27.74%) local posts, followed by 11,429 (23.29%; 95% CI 22.65%-23.91%) within SEAR, and 51,956 (17.43%; 95% CI 17.35%-17.8%) within undisclosed regions. The second most popular topic was "algorithms and data," which frequently used the terms "datum," "model," and "patient." Review of sample posts found discussions of the use of personal health data for health prediction and disease models, and it was most prominent in AMR, which shared 47,910 (18.19%; 95% CI 17.92%-18.44%) related posts of discussions within the region. This was followed by "human-AI alignment," addressing AI's alignment with human ethics and values. Notably, 40.25% (14,966/37,184) of these discussions mentioned "mental health," along with terms like "help" and "people." "Patient care applications" examined specific AI-driven health solutions and products, such as the creation of drug treatments and diagnostic abilities, referencing words like "treatment," "improve," "patient," and "care." Finally, "research forums" highlighted emerging research opportunities like conferences and academic positions using words such as "join," "digital," "global," and "work." Interest in research was predominant within WPR, which shared n=6950 (19.27%; 95% CI 18.5%-20.07%) related posts from the region, European Union at n=19,192 (18.39%; 95% CI 18.01%-18.8%) posts, and the African Region with n=3657 (16.75%; 95% CI 15.92%-17.6%) posts (Figure 2).

Figure 2. Distribution of top 5 health artificial intelligence topics on X by World Health Organization region, January 1 to December 3, 2023. AFR: African Region; AI: artificial intelligence; AMR: Region of the Americas; EMR: Eastern-Mediterranean Region; EUR: European Union; SEAR: South-East Asia Region; WHO: World Health Organization; WPR: Western Pacific Region.

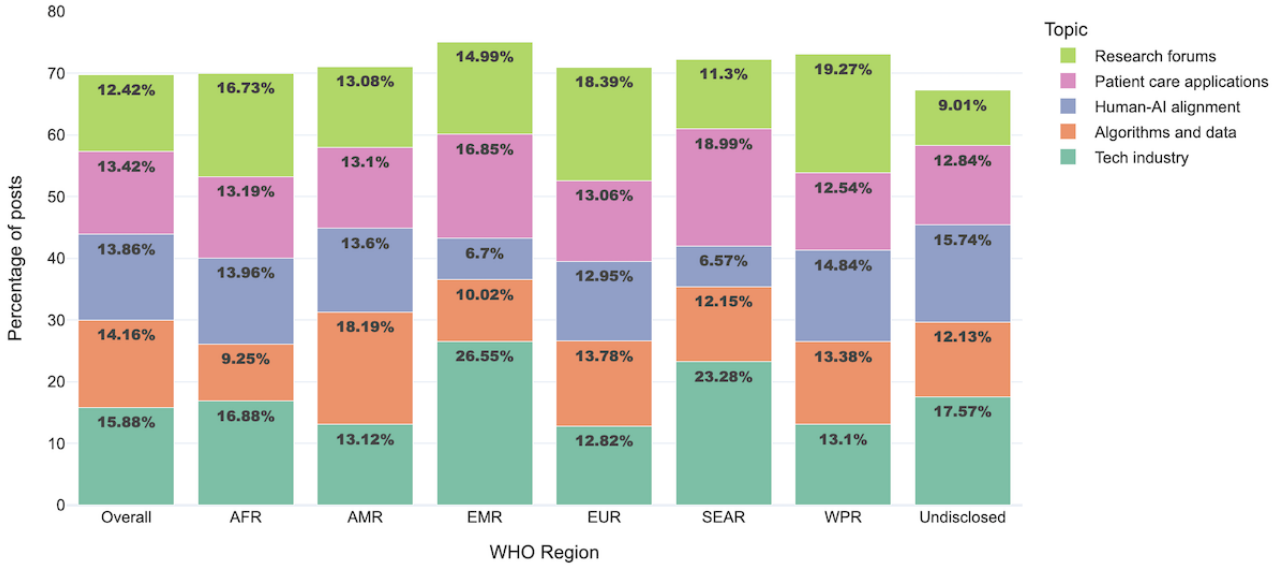

XSL•FO
RenderX

Table 4. The 5 most frequent topics in health AI-related discussions on X, January 1 to December 3, 2023. Top 10 most probable and salient terms are reported per topic.

Topic name	n (%; 95% CI)	Probable terms	Salient terms	Examples
Tech industry	42,399 (15.88; 15.67-15.94)	business, ChatGPT, public, digital, ethical, social, technology, education, system, and need	industry, future, technology, revolutionize, innovation, machine learning, finance, transform, potential, and blockchain	• "Explore the world of AI chatbots across industries! From e-commerce to healthcare, these intelligent tools are reshaping businesses. Discover how they align with industry KPIs and stay updated on the latest innovations. Dive in!" • "As an entrepreneur, I learnt how to use technology to create real impact & pivot my company to be ahead of the curve. Being among the founders of top AI, education & healthcare companies, & understanding the new age findings in AI, AR/VR, robotics & medicine was super insightful."
Algorithms and data	37,700 (14.05; 13.92-14.19)	datum, model, patient, medical, care, generative, system, new, clinical, and machine learning	patient, datum, medical, improve, treatment, model, care, disease, outcome, and diagnosis	• "Our generative AI (really a deconstructed neural network-based expert system model built on established public scientific evidence) enables both patients and clinicians to increase their connection and collaboration leading to better and more deeply personalized healthcare." • "Excellent article highlighting issues around AI in health care. Notes a number of challenges: 'a major issue is the relative scarcity of publicly available data sets in medicine.' Reproducibility issues that haunt health-care AI"
Human-AI[a] alignment	37,184 (13.86; 13.72-13.99)	mental, people, need, human, help, ChatGPT, support, work, good, and time	mental, people, need, talk, help, ChatGPT, good, therapy, diaspora, and feel	• Are you the type of person who wants to stick with the old techniques? If so, It's better if you start learning how to flip the burgers, or wait, even AI will replace these jobs, AI is on absolute fire! • "me: for the good of mankind we need to beat up anyone in AI development with hammers Everyone around me: oh so you hate technology huh! People developing AIs"
Patient care applications	33,850 (12.6; 12.48-12.73)	patient, care, medical, treatment, improve, disease, diagnosis, help, technology, and doctor	digitalhealth, doctor, healthtech, chatgpt, patient, care, healthit, news, telemedicine, and medicine	• "Artificial intelligence is having a MAJOR breakthrough within the healthcare industry right now! It's playing a huge role in the discovery of new drugs, triaging, and fighting specific resistances against drugs." • "The predictive capabilities of AI tools save invaluable time, allowing #healthcare professionals to focus on what truly matters - their patients. By improving diagnostic processes, #patientsatisfaction, and overall #health outcomes are bound to reach new heights."
Research forums	33,198 (12.37; 12.25-12.49)	research, join, global, digital, work, science, innovation, discuss, project, and future	join, research, discuss, global, register, event, conference, science, university, and discussion	• "As we mark #BreastCancerAwarenessMonth, check out this study on AI's transformative role in same-day diagnostics from #AHRQ's Digital Healthcare Research Program 2022 Year in Review." • "Learn all about federated learning in healthcare projects at AI Meetup next week! Amii's Payam Mousavi leads a panel discussion feat. Ross Mitchell, Ruchika & Neeraj Kumar, three co-authors of a recently published study in @NatureComms"

[a]AI: artificial intelligence.

XSL•FO
RenderX

For all 10 topics, we report the top terms with topic word probabilities and example posts in Table S7 in Multimedia Appendix 1 to support interpretability and reproducibility of labels. The relative frequency of keywords characterizing each of the 10 topics is plotted in Figure 3, representing key terms contributing to the topic model.

Figure 3. Ten topics and respective keywords in health artificial intelligence–related discussions on X, January 1 to December 3, 2023. "#1" denotes the most frequent topic in a World Health Organization region. AFR: African Region; AI: artificial intelligence; AMR: Region of the Americas; EMR: Eastern-Mediterranean Region; EUR: European Union; SEAR: South-East Asia Region; WPR: Western Pacific Region.

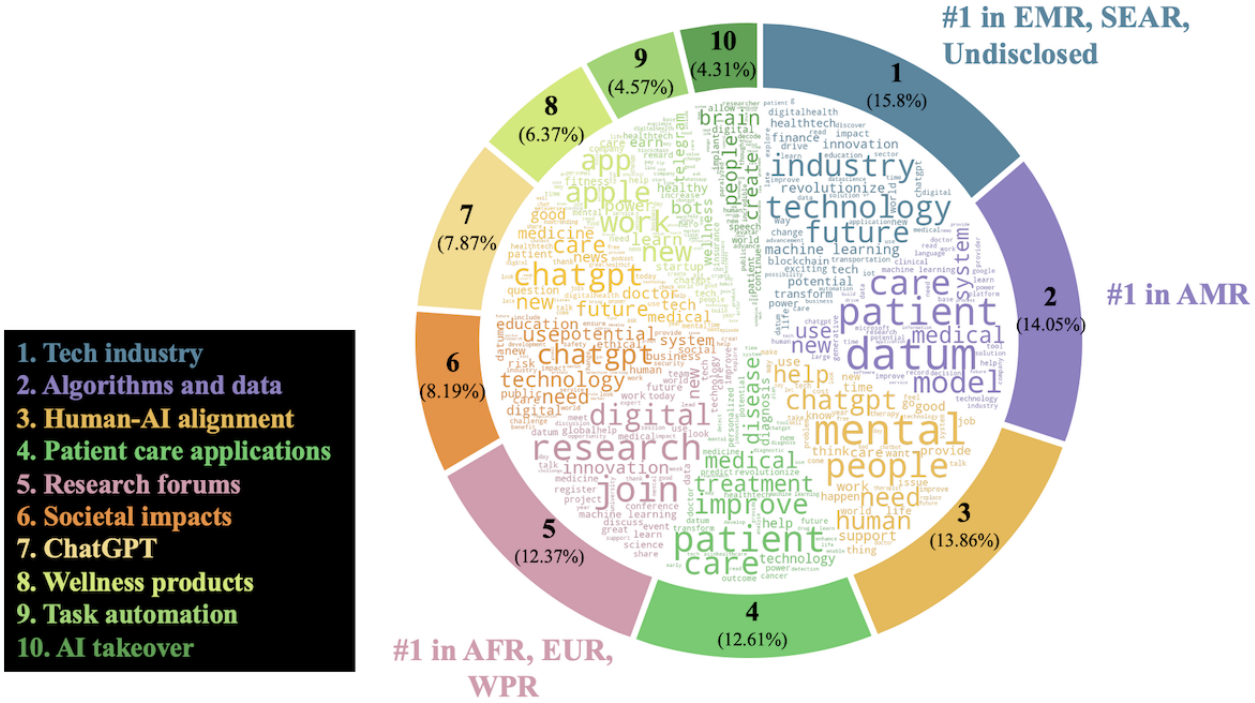

Topics were differently emotionally charged. Each post could be represented by multiple emotions. In "tech industry," 41,816 (98.31%; 95% CI 98.19%-98.43%) post contained optimism. Similarly, in "research forums," 32,240 (96.98; 95% CI 96.8%-97.17%) posts contained optimistic language, respectively. Despite also showing high rates of optimism, pessimism was found in 7793 (38.87%; 95% CI 38.23%-39.49%) posts in "societal impacts," along with 13,874 (36.68%; 95% CI 36.17%-37.2%) in "algorithms and data," reflecting mixed outlooks on these issues. By comparison, 21,648 posts in "human-AI alignment" and 6948 posts in "AI takeover" contained disgust, making it more prevalent than optimism in both topics (Figure 4). Anger was also common in these topics; in "human-AI alignment," 15,019 (40.83%; 95% CI 40.31%-41.3%) posts contained anger, while in "AI takeover," 5564 (51.27%; 95% CI 50.26%-52.17%) did.

"Love," "surprise," and "neutral" (no emotion identified) were in less than 1% of posts and omitted from Figure 4.

XSL•FO
RenderX

Figure 4. Emotional content of 10 topics found in discussions about health artificial intelligence on X, January 1 to December 3, 2023. AI: artificial intelligence.

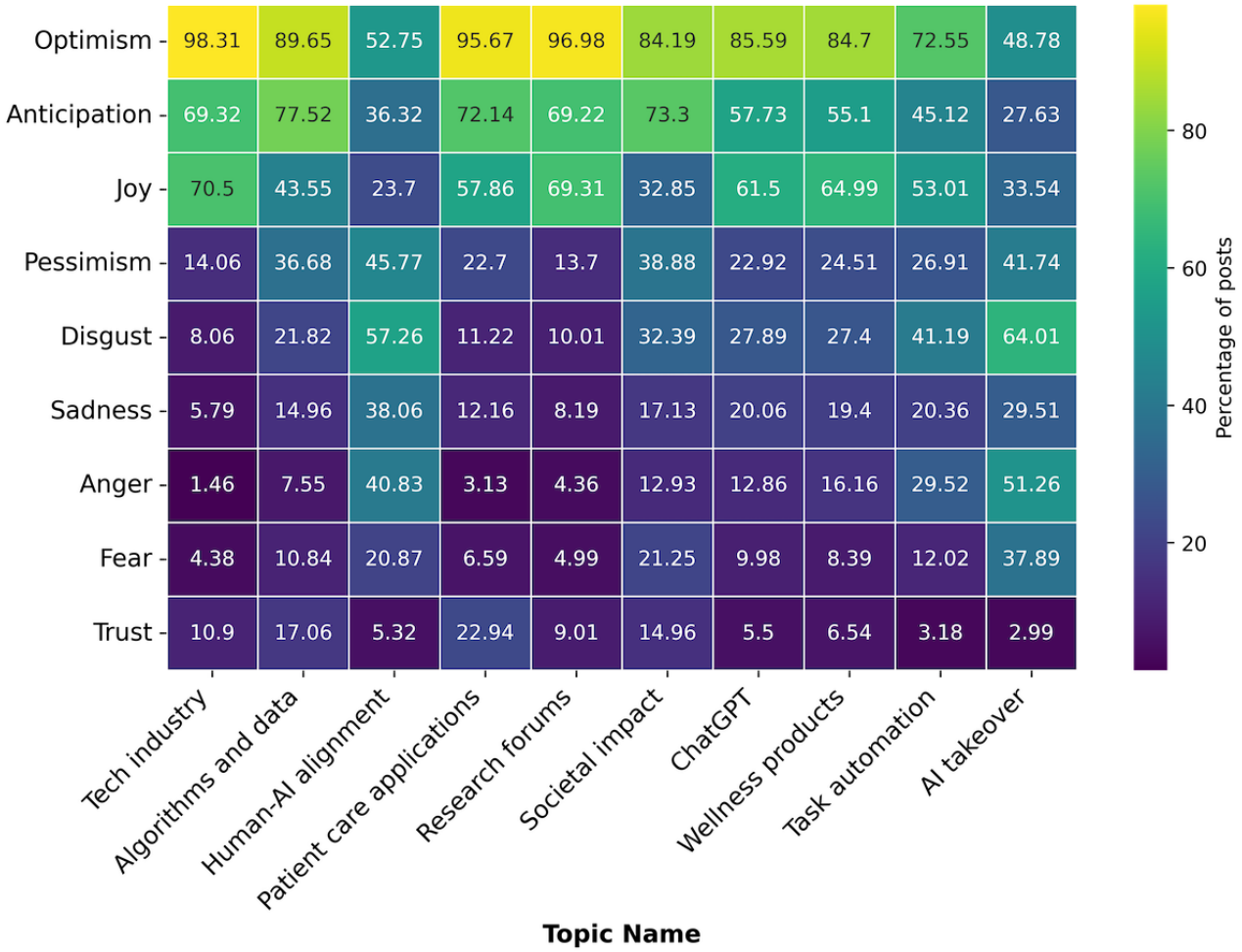

Discussion

Principal Findings

The rapid rise in public attention catalyzed by generative AI—particularly following the release of ChatGPT—coincided with a marked expansion in the scale and visibility of public discourse on health AI in 2023. In this analysis of posts on X using LLM-enabled methods, discussions were predominantly favorable across regions, with perceived usefulness emerging as the strongest driver of confidence. At the same time, privacy was consistently the least trusted dimension, indicating that enthusiasm for the potential of health AI is accompanied by persistent concern about data governance. Emotional patterns reinforced this tension; optimism dominated conversations about innovation, patient care, and research, whereas fear and anger were concentrated in discussions of human-AI alignment and AI takeover. We also identified significant regional variation in these narratives, with EMR and SEAR showing more innovation-oriented and optimistic discussions, while WPR, the European Union, and AMR placed greater emphasis on privacy, ethics, and alignment. Taken together, these findings contribute to the characterization of how health AI is framed and viewed online, showing that public confidence in health AI is not uniform, but multidimensional, regionally shaped, and defined by a balance between promise and concern.

Regional and Thematic Patterns in Health AI Discourse

We found 65% (252,082/386,287) of posts to be positive in sentiment about health AI. This figure is moderately higher but directionally consistent with attitudes toward health AI found on other social media and surveys. A previous study of X and Reddit posts from 2021 to 2024 showed 55% (566/1022) expressed positive sentiments toward AI in medical imaging [22], while 58% (7775/13,502) of patients in a 2023 global survey were positive about AI's integration into health care [66]. The somewhat higher positivity in our sample may reflect the study time period's overlap with significant model developments that generated excitement online. Following the release of ChatGPT in late 2022, conversation volume surged from zero related tweets to 550,000 per day by late January 2023, with average sentiment consistently more than 50% positive in the first few months [11]. Studies including data before and after 2023 may reflect stabilization to a more even distribution between sentiments than we observed. The discussion focus on industry and commercial use cases parallel results from a study of social media from China (Weibo [Sina Corporation]) [10], a region largely excluded from this study's sample on X. This suggests global platforms are commonly used to discuss and promote health AI business, followed by discussions of AI's impact and alignment with society [10]. In

XSL•FO
RenderX

the media's portrayal of general AI, business tends to be the most common positive angle [67], which we also observed in the near-unanimous level of optimism found in posts discussing the health AI tech industry.

We found enthusiastic attitudes were most prevalent in EMR and SEAR, where positive sentiments and optimistic emotions were significantly higher than in other regions. By comparison, EMR and SEAR rarely mentioned human-AI alignment, which was more popular in WPR and undisclosed regions. The high prevalence of tech industry–focused discussions, at 27% in EMR and 23% in SEAR, aligns with a previous study of news media from the same countries that are primarily represented in the EMR and SEAR samples, the United Arab Emirates, and India, respectively. Nearly 25% of coverage from the United Arab Emirates focused on industry expansion and economic initiatives, while 21% of articles from India discussed startups [68]. Both countries' newspapers did not mention regulations, privacy, or ethics [68]. A potential feedback loop may exist in which supportive media narratives and social media discourse mutually reinforce optimistic narratives about health AI adoption, while simultaneously minimizing discussion of specific concerns more prevalent in other regional media.

On the other hand, social media posts from mostly Western countries in WPR, the European Union, and AMR expressed less confidence in the technology, with greater concern for privacy, ethics, and AI alignment. Sentiment toward health AI was less positive than EMR and SEAR. Emotions of disgust, pessimism, and fear were elevated in posts from the region, mirroring an increasing trend of fear-invoking headlines trend from local newspapers [68]. In the European Union and WPR, discussions about research, including academic funding, jobs, and conferences, outnumbered mentions of industry. This "responsibility" focus is plausibly explained by regional contexts. In the study period, the European Union was where the first legal framework on AI, the AI Act, was being deliberated [69], while AMR's topical focus on algorithms and data reflects longstanding public concerns over data use and privacy in the United States [8]. Yet across all regions, privacy was the most consistent concern, substantially present even in EMR and SEAR with otherwise positive sentiment and optimistic emotional tone. This convergence suggests that privacy is a near-universal dimension of public unease about health AI that persists regardless of the broader narrative orientation of a region's discourse. In this case, social media represents an important middle ground of social expression between traditional media sources, which in some regions contain no mention of AI-related privacy issues [68], and direct surveys of patients, which suggest concern about health AI privacy is above 50% [70]. Where in-depth surveying is limited or not feasible, social media can be analyzed to surface latent concerns that traditional media underreports. The privacy concerns found in this study suggest data governance is a cross-cutting barrier to health AI adoption, even in regions where the broader online narrative appears favorable.

Taken together, these regional patterns demonstrate social media as a real-time, decentralized information environment where competing narratives about AI unfold, distinct from both institutional media and individually held viewpoints. A single global communication strategy for health AI is therefore unlikely to be effective, and instead must address how the proliferation of social media discourse surrounding AI might shape real-world instances of public trust in health technologies. The concentration of high-arousal negative emotions around existential AI narratives is particularly consequential; negative, emotionally charged content is associated with greater spread of posts, including misinformation [16,71], which could explain why posts using fear to discuss health AI reached close to 8000 users on average. Drawing on previous experiences with vaccine hesitancy, online misinformation is associated with and can affect health technology beliefs and adoption intention [72,73], including a negative relationship between misinformation consumption and COVID-19 vaccine uptake [74,75]. Policymakers and health authorities should consider integrating AI-enabled social listening into health technology assessment, risk communication, and regulatory monitoring to proactively anticipate emerging concerns and enable more responsive, publicly aligned health AI governance.

Practical Implications of LLMs

This study evaluated LLM-based annotation for health-related social media analysis using the AI Usage Consideration Checklist (Multimedia Appendix 3). In terms of accuracy of information, a few-shot CoT approach frequently outperformed previous evaluations of off-the-shelf sentiment tools and zero-shot LLMs, and achieved comparable performance to fine-tuned models on the same metric (wF_1-score) [24,30]. This suggests that few-shot CoT prompting can meaningfully improve annotations for social media analysis, an area where traditional sentiment tools often perform poorly [26,30]. GPT-4 outperformed GPT-3.5-Turbo across annotation tasks, achieving wF_1-scores above 0.7, and continued model development suggests further performance gains are likely.

Regarding usability, the LLM pipeline was accessible to noncomputer science researchers via natural language instructions and simple Python scripts, lowering the barrier to adoption for public health institutions with otherwise limited machine learning capacity. LLMs also enabled annotation of over 300,000 posts, a scale not attained through traditional manual content analysis on the same topic [10,22].

Limitations and Future Research

This study has several limitations. First, although all annotation tasks achieved wF_1-scores above 0.6, performance varied across classification constructs, and complex dimensions, such as trust in AI, remain challenging given the lack of definitional consensus even among human experts [76]. Our analysis distinguished between perceived trustworthiness and affective trust [77] as a pragmatic solution in the absence of an established framework for confidence in health AI; however, this approach is operationalized for scalable social listening and should be interpreted as a descriptive monitoring framework using constructs informed by the literature, rather than a validated psychometric instrument. Misclassification is possible due to LLM tendencies to exaggerate interpretations of positive or negative sentiments [26], along with classification challenges when short texts contain sarcasm or irony, which both humans

XSL•FO
RenderX

and language models struggle with [47]. To mitigate these risks, we implemented a structured annotation framework with predefined categories, applied benchmarking thresholds, and reported attribute-level patterns rather than aggregate polarity scores alone. Future work should incorporate formal interannotator validation, larger benchmarks, and model comparisons. Constructs should also be refined and validated against survey-based or experimental measures to strengthen conceptual alignment and elucidate relationships between the online health AI landscape and offline attitudes to such technologies. Third, our analysis was restricted to global English-language posts on X, which may underrepresent non–English-speaking populations and regions where the platform is restricted. However, this focus reflects the empirical context of data collection in 2023, immediately following the release and widespread uptake of predominantly English-language generative AI systems, during which early public discourse was largely English-language dominated [11]. As AI systems become increasingly multilingual and globally embedded, future studies should expand to non-English datasets and additional platforms to enhance cross-cultural validity and representativeness. Finally, model performance and construct mapping are time- and context-specific, and both public discourse and model behavior may shift over time. Nevertheless, this adaptability is also a strength; computational approaches can be rerun and recalibrated as language and technology evolve. We encourage periodic re-evaluation of model performance and updated benchmarking datasets.

This study advances previous social media analyses of health AI [10,22] by moving beyond basic sentiment classification to enable multidimensional characterization of public confidence across 42 countries during a period of rapid growth in AI-related online activity. The integration of LLM-based stance and emotion annotation with LDA topic modeling allowed for context-sensitive classification of complex constructs at scale. These findings establish a reproducible framework for scalable surveillance of online attitudes in digital health research that can be extended to other health technologies as they enter public discourse, and recalibrated as the language used to discuss AI continues to evolve. For policymakers, the regional and thematic granularity of discourse data provides detailed analysis about what information online populations are being exposed to and how digital ecosystems respond to changing developments in both technology and regulation. Among health system leaders and communicators, real-time social listening of this kind can anticipate concerns before implementation, including in regions where direct surveying is limited, to understand what aspects of the technology are viewed positively in popular discourse and which still face resistance.

Acknowledgments

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: (1) proofreading and editing, (2) summarizing text, and (3) reformatting.

The GenAI tool used was ChatGPT-4.0.

Responsibility for the final manuscript lies entirely with the authors.

GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This study was supported by the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region. The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The social media data analyzed in this study cannot be made publicly available due to licensing and contractual restrictions governing access to the platform data. Researchers seeking to reproduce the findings should obtain independent access to X data through authorized third-party providers or formal research agreements with the platform. The full search strategy and filtering criteria used to construct the dataset are described in Multimedia Appendix 1 to support methodological transparency and reproducibility.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of methodology and full results of all analyses.
[DOC File , 476 KB - jmir_v28i1e80346_app1.doc]

Multimedia Appendix 2

XSL•FO
RenderX

GPT prompts and parameter settings.
[DOC File , 74 KB - jmir_v28i1e80346_app2.doc]

Multimedia Appendix 3
AI usage consideration checklist.
[DOC File , 271 KB - jmir_v28i1e80346_app3.doc]

References

1. Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The breakthrough of large language models release for medical applications: 1-year timeline and perspectives. J Med Syst 2024 Feb 17;48(1):22 [FREE Full text] [doi: 10.1007/s10916-024-02045-3] [Medline: 38366043]
2. Senthil R, Anand T, Somala CS, Saravanan KM. Bibliometric analysis of artificial intelligence in healthcare research: Trends and future directions. Future Healthc J 2024 Sep;11(3):100182 [FREE Full text] [doi: 10.1016/j.fhj.2024.100182] [Medline: 39310219]
3. Yin J, Ngiam KY, Teo HH. Role of Artificial Intelligence Applications in Real-Life Clinical Practice: Systematic Review. J Med Internet Res 2021 Apr 22;23(4):e25759 [FREE Full text] [doi: 10.2196/25759] [Medline: 33885365]
4. Marco-Ruiz L, Hernández MÁT, Ngo PD, Makhlysheva A, Svenning TO, Dyb K, et al. A multinational study on artificial intelligence adoption: Clinical implementers' perspectives. Int J Med Inform 2024 Apr;184:105377 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105377] [Medline: 38377725]
5. Mohsin Khan M, Shah N, Shaikh N, Thabet A, Alrabayah T, Belkhair S. Towards secure and trusted AI in healthcare: A systematic review of emerging innovations and ethical challenges. Int J Med Inform 2025 Mar;195:105780 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105780] [Medline: 39753062]
6. Sahoo RK, Sahoo KC, Negi S, Baliarsingh SK, Panda B, Pati S. Health professionals' perspectives on the use of Artificial Intelligence in healthcare: A systematic review. Patient Educ Couns 2025 May;134:108680. [doi: 10.1016/j.pec.2025.108680] [Medline: 39893988]
7. Young AT, Amara D, Bhattacharya A, Wei ML. Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. Lancet Digit Health 2021 Sep;3(9):e599-e611 [FREE Full text] [doi: 10.1016/S2589-7500(21)00132-1] [Medline: 34446266]
8. Beets B, Newman TP, Howell EL, Bao L, Yang S. Surveying Public Perceptions of Artificial Intelligence in Health Care in the United States: Systematic Review. J Med Internet Res 2023 Apr 04;25:e40337 [FREE Full text] [doi: 10.2196/40337] [Medline: 37014676]
9. Moy S, Irannejad M, Manning SJ, Farahani M, Ahmed Y, Gao E, et al. Patient Perspectives on the Use of Artificial Intelligence in Health Care: A Scoping Review. J Patient Cent Res Rev 2024 Apr 02;11(1):51-62 [FREE Full text] [doi: 10.17294/2330-0698.2029] [Medline: 38596349]
10. Gao S, He L, Chen Y, Li D, Lai K. Public Perception of Artificial Intelligence in Medical Care: Content Analysis of Social Media. J Med Internet Res 2020 Jul 13;22(7):e16649 [FREE Full text] [doi: 10.2196/16649] [Medline: 32673231]
11. Fütterer T, Fischer C, Alekseeva A, Chen X, Tate T, Warschauer M, et al. ChatGPT in education: global reactions to AI innovations. Sci Rep 2023 Sep 15;13(1):15310 [FREE Full text] [doi: 10.1038/s41598-023-42227-6] [Medline: 37714915]
12. Survey on the impact of online disinformation and hate speech. UNESCO. 2023. URL: https://www.unesco.org/sites/default/files/medias/fichiers/2023/11/unesco_ipsos_survey.pdf [accessed 2026-04-01]
13. Milli S, Carroll M, Wang Y, Pandey S, Zhao S, Dragan AD. Engagement, user satisfaction, and the amplification of divisive content on social media. PNAS Nexus 2025 Mar;4(3):pgaf062 [FREE Full text] [doi: 10.1093/pnasnexus/pgaf062] [Medline: 40070432]
14. Ferrara E, Yang Z. Quantifying the effect of sentiment on information diffusion in social media. PeerJ Computer Science 2015 Sep 30;1:e26. [doi: 10.7717/peerj-cs.26]
15. Chuai Y, Zhao J. Anger can make fake news viral online. Front. Phys 2022 Aug 22;10. [doi: 10.3389/fphy.2022.970174]
16. Han J, Cha M, Lee W. Anger contributes to the spread of COVID-19 misinformation. HKS Misinfo Review 2020 Sep 17;1(3). [doi: 10.37016/mr-2020-39]
17. Wubineh BZ, Deriba FG, Woldeyohannis MM. Exploring the opportunities and challenges of implementing artificial intelligence in healthcare: A systematic literature review. Urol Oncol 2024 Mar;42(3):48-56 [FREE Full text] [doi: 10.1016/j.urolonc.2023.11.019] [Medline: 38101991]
18. Martel C, Pennycook G, Rand DG. Reliance on emotion promotes belief in fake news. Cogn Res Princ Implic 2020 Oct 07;5(1):47 [FREE Full text] [doi: 10.1186/s41235-020-00252-3] [Medline: 33026546]
19. Gu C, Zhang Y, Zeng L. Exploring the mechanism of sustained consumer trust in AI chatbots after service failures: a perspective based on attribution and CASA theories. Humanit Soc Sci Commun 2024 Oct 22;11(1):1-12. [doi: 10.1057/s41599-024-03879-5]
20. Novozhilova E, Mays K, Paik S, Katz J. More Capable, Less Benevolent: Trust Perceptions of AI Systems across Societal Contexts. MAKE 2024 Feb 05;6(1):342-366. [doi: 10.3390/make6010017]

XSL•FO
RenderX

21. Afroogh S, Akbari A, Malone E, Kargar M, Alambeigi H. Trust in AI: progress, challenges, and future directions. Humanit Soc Sci Commun 2024 Nov 18;11(1):1-30. [doi: 10.1057/s41599-024-04044-8]
22. Almanaa M. Trends and Public Perception of Artificial Intelligence in Medical Imaging: A Social Media Analysis. Cureus 2024 Sep;16(9):e70008. [doi: 10.7759/cureus.70008] [Medline: 39445247]
23. Mao Y, Liu Q, Zhang Y. Sentiment analysis methods, applications, and challenges: A systematic literature review. Journal of King Saud University - Computer and Information Sciences 2024 Apr;36(4):102048. [doi: 10.1016/j.jksuci.2024.102048]
24. He L, Yin T, Zheng K. They May Not Work! An evaluation of eleven sentiment analysis tools on seven social media datasets. J Biomed Inform 2022 Aug;132:104142 [FREE Full text] [doi: 10.1016/j.jbi.2022.104142] [Medline: 35835437]
25. Guido R, Ferrisi S, Lofaro D, Conforti D. An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review. Information 2024 Apr 19;15(4):235. [doi: 10.3390/info15040235]
26. Krugmann JO, Hartmann J. Sentiment Analysis in the Age of Generative AI. Cust. Need. and Solut 2024 Mar 05;11(1):3. [doi: 10.1007/s40547-024-00143-4]
27. Zhang W, Deng Y, Liu B, Pan SJ, Bing L. Sentiment Analysis in the Era of Large Language Models: A Reality Check. ArXiv Preprint posted online on June 14, 2023 [FREE Full text] [doi: 10.18653/v1/2024.findings-naacl.246]
28. Wang Z, Pang Y, Lin Y, Zhu X. Adaptable and reliable text classification using large language models. ArXiv Preprint posted online on December 7, 2024 [FREE Full text] [doi: 10.1109/icdmw65004.2024.00015]
29. Lossio-Ventura JA, Weger R, Lee AY, Guinee EP, Chung J, Atlas L, et al. A Comparison of ChatGPT and Fine-Tuned Open Pre-Trained Transformers (OPT) Against Widely Used Sentiment Analysis Tools: Sentiment Analysis of COVID-19 Survey Data. JMIR Ment Health 2024 Jan 25;11:e50150 [FREE Full text] [doi: 10.2196/50150] [Medline: 38271138]
30. He L, Omranian S, McRoy S, Zheng K. Using Large Language Models for sentiment analysis of health-related social media data: empirical evaluation and practical tips. AMIA Annu Symp Proc 2024;2024:503-512 [FREE Full text] [Medline: 40417506]
31. Albalawi R, Yeap TH, Benyoucef M. Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. Front Artif Intell 2020;3:42 [FREE Full text] [doi: 10.3389/frai.2020.00042] [Medline: 33733159]
32. Liu Z, Yang K, Xie Q, Zhang T, Ananiadou S. EmoLLMs: A series of emotional large language models and annotation tools for comprehensive affective analysis. 2024 Aug 24 Presented at: KDD '24: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 2024 August 25 - 29; Barcelona Spain p. 5487-5496 URL: https://dl.acm.org/doi/10.1145/3637528.3671552 [doi: 10.1145/3637528.3671552]
33. Hou Z, Tong Y, Du F, Lu L, Zhao S, Yu K, et al. Assessing COVID-19 Vaccine Hesitancy, Confidence, and Public Engagement: A Global Social Listening Study. J Med Internet Res 2021 Jun 11;23(6):e27632 [FREE Full text] [doi: 10.2196/27632] [Medline: 34061757]
34. Zhou X, Zhang X, Larson H, de Figueiredo A, Jit M, Fodeh S, et al. Spatiotemporal trends in COVID-19 vaccine sentiments on a social media platform and correlations with reported vaccine coverage. Bull World Health Organ 2024 Jan 01;102(1):32-45 [FREE Full text] [doi: 10.2471/BLT.23.289682] [Medline: 38164328]
35. Xu J, Wu Z, Wass L, Larson HJ, Lin L. Mapping global public perspectives on mRNA vaccines and therapeutics. NPJ Vaccines 2024 Nov 14;9(1):218 [FREE Full text] [doi: 10.1038/s41541-024-01019-3] [Medline: 39543153]
36. MacDonald NE, SAGE Working Group on Vaccine Hesitancy. Vaccine hesitancy: Definition, scope and determinants. Vaccine 2015 Aug 14;33(34):4161-4164 [FREE Full text] [doi: 10.1016/j.vaccine.2015.04.036] [Medline: 25896383]
37. Kelly S, Kaye S, Oviedo-Trespalacios O. What factors contribute to the acceptance of artificial intelligence? A systematic review. Telematics and Informatics 2023 Feb;77:101925. [doi: 10.1016/j.tele.2022.101925]
38. Ali O, Abdelbaki W, Shrestha A, Elbasi E, Alryalat MAA, Dwivedi YK. A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities. Journal of Innovation & Knowledge 2023 Jan;8(1):100333. [doi: 10.1016/j.jik.2023.100333]
39. Hassan N, Slight R, Bimpong K, Bates DW, Weiand D, Vellinga A, et al. Systematic review to understand users perspectives on AI-enabled decision aids to inform shared decision making. NPJ Digit Med 2024 Nov 21;7(1):332 [FREE Full text] [doi: 10.1038/s41746-024-01326-y] [Medline: 39572838]
40. Wu C, Xu H, Bai D, Chen X, Gao J, Jiang X. Public perceptions on the application of artificial intelligence in healthcare: a qualitative meta-synthesis. BMJ Open 2023 Jan 04;13(1):e066322 [FREE Full text] [doi: 10.1136/bmjopen-2022-066322] [Medline: 36599634]
41. Ibrahim F, Münscher JC, Daseking M, Telle N. The technology acceptance model and adopter type analysis in the context of artificial intelligence. Front Artif Intell 2025 Jan 16;7:1496518 [FREE Full text] [doi: 10.3389/frai.2024.1496518] [Medline: 39886023]
42. Alizadeh M, Kubli M, Samei Z, Dehghani S, Zahedivafa M, Bermeo JD, et al. Open-source LLMs for text annotation: a practical guide for model setting and fine-tuning. J Comput Soc Sci 2024 Dec 18;8(1):17. [doi: 10.1007/s42001-024-00345-9] [Medline: 39712076]
43. Espinosa L, Salathé M. Use of large language models as a scalable approach to understanding public health discourse. PLOS Digit Health 2024 Oct;3(10):e0000631. [doi: 10.1371/journal.pdig.0000631] [Medline: 39401186]

XSL•FO
RenderX

44. Deiner MS, Deiner NA, Hristidis V, McLeod SD, Doan T, Lietman TM, et al. Use of Large Language Models to Assess the Likelihood of Epidemics From the Content of Tweets: Infodemiology Study. J Med Internet Res 2024 Mar 01;26:e49139 [FREE Full text] [doi: 10.2196/49139] [Medline: 38427404]
45. Brown TB, Mann B, Ryder N. Language Models are Few-Shot Learners. ArXiv Preprint posted online on July 22, 2020 [FREE Full text] [doi: 10.48550/arXiv.2005.14165]
46. Wei J, Wang X, Schuurmans D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. ArXiv Preprint posted online on January 10, 2023 [FREE Full text] [doi: 10.52202/068431-1800]
47. Bojić L, Zagovora O, Zelenkauskaite A, Vuković V, Čabarkapa M, Veseljević Jerković S, et al. Comparing large Language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. Sci Rep 2025 Apr 03;15(1):11477 [FREE Full text] [doi: 10.1038/s41598-025-96508-3] [Medline: 40181141]
48. Zhu Y, Zhang P, Haq E, Hui P, Tyson G. Can ChatGPT reproduce human-generated labels? A study of social computing tasks. ArXiv Preprint posted online on April 22, 2023 [FREE Full text] [doi: 10.48550/arXiv.2304.10145]
49. Ola O, Sedig K. Understanding Discussions of Health Issues on Twitter: A Visual Analytic Study. Online J Public Health Inform 2020 May 16;12(1):e2 [FREE Full text] [doi: 10.5210/ojphi.v12i1.10321] [Medline: 32577151]
50. X (Formerly Twitter) User Age, Gender, & Demographic Stats. 2024. URL: https://explodingtopics.com/blog/x-user-stats [accessed 2025-02-03]
51. Number of ChatGPT users. Exploding Topics. URL: https://explodingtopics.com/blog/chatgpt-users [accessed 2024-08-16]
52. Lai VD, Ngo N, Pouran BVA. ChatGPT beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. : Association for Computational Linguistics; 2023 Dec Presented at: Findings of the Association for Computational Linguistics: EMNLP 2023; 2026 April 01; Singapore p. 13171-13189. [doi: 10.18653/v1/2023.findings-emnlp.878]
53. Varol O, Ferrara E, Davis C, Menczer F, Flammini A. Online human-bot interactions: Detection, estimation, and characterization. In: ICWSM. Online Human-Bot Interactions: Detection, Estimation, and Characterization. Proceedings of the International AAAI Conference on Web and Social Media AAAI Press; 2017 Presented at: International AAAI Conference on Web and Social Media; 2017 May 15-18; Montreal, Canada p. 280-289 URL: https://cdn.aaai.org/ojs/14871/14871-28-18390-1-2-20201228.pdf [doi: 10.1609/icwsm.v11i1.14871]
54. Choi S, Lee C, Park A, Lee JA. How the Public Makes Sense of Artificial Intelligence: The Interplay Between Communication and Discrete Emotions. Science Communication 2024 Nov 23;47(4):553-584. [doi: 10.1177/10755470241297664]
55. Drummond C, Fischhoff B. Emotion and judgments of scientific research. Public Underst Sci 2020 Apr;29(3):319-334. [doi: 10.1177/0963662520906797] [Medline: 32098582]
56. Buechel S, Hahn U. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. : Association for Computational Linguistics; 1996 Presented at: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers; 2026 April 01; Valencia, Spain p. 578-585. [doi: 10.18653/v1/e17-2092]
57. lzw1008/Emollama-7b. Hugging Face. URL: https://huggingface.co/lzw1008/Emollama-7b [accessed 2026-02-03]
58. (PDF) Latent Dirichlet Allocation. URL: https://www.researchgate.net/publication/221620547_Latent_Dirichlet_Allocation [accessed 2024-08-16]
59. Khan A, Ali R. Measuring the effectiveness of LDA-based clustering for social media data. : IEEE; 2023 Presented at: 2023 International Conference on Advances in Intelligent Computing and Applications (AICAPS); 2023 February 1–3; Kochi, India p. 1-8. [doi: 10.1109/aicaps57044.2023.10074399]
60. Molenaar A, Lukose D, Brennan L, Jenkins EL, McCaffrey TA. Using Natural Language Processing to Explore Social Media Opinions on Food Security: Sentiment Analysis and Topic Modeling Study. J Med Internet Res 2024 Mar 21;26:e47826 [FREE Full text] [doi: 10.2196/47826] [Medline: 38512326]
61. Huangfu L, Mo Y, Zhang P, Zeng DD, He S. COVID-19 Vaccine Tweets After Vaccine Rollout: Sentiment-Based Topic Modeling. J Med Internet Res 2022 Feb 08;24(2):e31726 [FREE Full text] [doi: 10.2196/31726] [Medline: 34783665]
62. Leung YT, Khalvati F. Exploring COVID-19-Related Stressors: Topic Modeling Study. J Med Internet Res 2022 Jul 13;24(7):e37142 [FREE Full text] [doi: 10.2196/37142] [Medline: 35731966]
63. Weston SJ, Shryock I, Light R, Fisher PA. Selecting the Number and Labels of Topics in Topic Modeling: A Tutorial. Advances in Methods and Practices in Psychological Science 2023 May 25;6(2):251524592311601. [doi: 10.1177/25152459231160105]
64. Usage Policies. URL: https://openai.com/policies/usage-policies/ [accessed 2024-08-16]
65. Meltwater. URL: https://www.meltwater.com/en [accessed 2024-08-16]
66. Busch F, Hoffmann L, Xu L, Zhang LJ, Hu B, García-Juárez I, COMFORT consortium, et al. Multinational Attitudes Toward AI in Health Care and Diagnostics Among Hospital Patients. JAMA Netw Open 2025 Jun 02;8(6):e2514452 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.14452] [Medline: 40493367]
67. Ryazanov I, Öhman C, Björklund J. How ChatGPT Changed the Media's Narratives on AI: A Semi-automated Narrative Analysis Through Frame Semantics. Minds & Machines 2024 Nov 23;35(1):2. [doi: 10.1007/s11023-024-09705-w]

XSL•FO
RenderX

68. Ittefaq M, Zain A, Arif R, Ala-Uddin M, Ahmad T, Iqbal A. Global news media coverage of artificial intelligence (AI): A comparative analysis of frames, sentiments, and trends across 12 countries. Telematics and Informatics 2025 Jan;96:102223. [doi: 10.1016/j.tele.2024.102223]
69. Gilbert S. The EU passes the AI Act and its implications for digital medicine are unclear. NPJ Digit Med 2024 May 22;7(1):135 [FREE Full text] [doi: 10.1038/s41746-024-01116-6] [Medline: 38778162]
70. Osnat B. Patient perspectives on artificial intelligence in healthcare: A global scoping review of benefits, ethical concerns, and implementation strategies. Int J Med Inform 2025 Nov;203:106007. [doi: 10.1016/j.ijmedinf.2025.106007] [Medline: 40494217]
71. McLoughlin KL, Brady WJ, Goolsbee A, Kaiser B, Klonick K, Crockett MJ. Misinformation exploits outrage to spread online. Science 2024 Nov 29;386(6725):991-996. [doi: 10.1126/science.adl2829] [Medline: 39607912]
72. Loomba S, de Figueiredo A, Piatek SJ, de Graaf K, Larson HJ. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nat Hum Behav 2021 Mar;5(3):337-348. [doi: 10.1038/s41562-021-01056-1] [Medline: 33547453]
73. Allen J, Watts DJ, Rand DG. Quantifying the impact of misinformation and vaccine-skeptical content on Facebook. Science 2024 May 31;384(6699):eadk3451. [doi: 10.1126/science.adk3451] [Medline: 38815040]
74. Pierri F, Perry BL, DeVerna MR, Yang K, Flammini A, Menczer F, et al. Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. Sci Rep 2022 Apr 26;12(1):5966 [FREE Full text] [doi: 10.1038/s41598-022-10070-w] [Medline: 35474313]
75. Romer D, Winneg KM, Jamieson PE, Brensinger C, Jamieson KH. Misinformation about vaccine safety and uptake of COVID-19 vaccines among adults and 5-11-year-olds in the United States. Vaccine 2022 Oct 26;40(45):6463-6470 [FREE Full text] [doi: 10.1016/j.vaccine.2022.09.046] [Medline: 36192273]
76. Gille F, Jobin A, Ienca M. What we talk about when we talk about trust: Theory of trust for AI in healthcare. Intelligence-Based Medicine 2020 Nov;1-2:100001. [doi: 10.1016/j.ibmed.2020.100001]
77. Bormann I, Bempreiksz‐Luthardt J, Niedlich S. Exploring the Interplay of Cognition and Emotion in Trust Relationships With Cognitive‐Affective Maps. Personal Relationships 2025 Jan 23;32(1):e12587. [doi: 10.1111/pere.12587]

Abbreviations

AI: artificial intelligence
AMR: Region of the Americas
CoT: Chain of Thought
EMR: Eastern-Mediterranean Region
LDA: latent Dirichlet allocation
LLM: large language model
NLP: natural language processing
SEAR: South-East Asia Region
wF1-score: weighted F1-score
WHO: World Health Organization
WPR: Western Pacific Region

Edited by S Brini; submitted 27.Jul.2025; peer-reviewed by X Lin, M Vochozka; comments to author 01.Oct.2025; accepted 24.Mar.2026; published 05.May.2026.

Please cite as:
Wass LM, Wu Z, Vizoso J, Wu JT, Lin L
Global Sentiment Toward Health AI at the Dawn of the ChatGPT Era: Empirical Analysis of Twitter (X) Discourse
J Med Internet Res 2026;28:e80346
URL: https://www.jmir.org/2026/1/e80346
doi:10.2196/80346
PMID:

©Lily Minh Wass, Zhengdong Wu, José Vizoso, Joseph T Wu, Leesa Lin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Self-Reported Health Outcomes in Metabolic Health YouTube Comments: Cross-Sectional Study and Rule-Based Natural Language Processing Framework Development and Validation

Ricardo Ribeiro[1*], MBA; Aneesh Zutshi[1*], PhD

Department of Mechanical and Industrial Engineering, NOVA School of Science and Technology, Universidade Nova de Lisboa, Caparica, Lisbon, Portugal

[*]all authors contributed equally

Corresponding Author:
Ricardo Ribeiro, MBA
Department of Mechanical and Industrial Engineering
NOVA School of Science and Technology
Universidade Nova de Lisboa
Quinta da Torre
Caparica, Lisbon, 2829-516
Portugal
Phone: 351 212948300
Fax: 351 212954461
Email: rasi.ribeiro@campus.fct.unl.pt

Abstract

Background: YouTube is increasingly used for healthcasting, the sharing of evidence-based dietary and lifestyle interventions by domain experts. In the metabolic health domain, channels focused on therapeutic carbohydrate restriction have accumulated audiences of millions. A distinctive feature is the comment section, where viewers share first-person accounts of health changes, constituting a unique source of real-world outcome data at scale. However, extracting structured health information from unstructured comments presents computational challenges.

Objective: This observational, cross-sectional study aims to develop and validate a precision-optimized computational framework for extracting self-reported health outcomes from healthcasting YouTube comments and to characterize the prevalence, distribution across health aspects, and channel-level variation of reported outcomes across a large-scale metabolic health corpus.

Methods: This study analyzed 43,111 unique YouTube comments from 110 videos across 11 therapeutic carbohydrate restriction-focused healthcasting channels (37,458 unique authors; data span November 2013 to January 2026; collected via YouTube data application programming interface version 3). The methodology comprised 3 construction phases and 5 validation studies. The construction phases were (1) exploratory corpus characterization, (2) iterative development of a 35-aspect hierarchical health outcome ontology, and (3) precision-optimized rule-based classification, validated through precision validation (stratified sample of n=500), recall estimation (n=510), external validation on 5 held-out channels (n=12,653 comments), large language model–assisted interrater reliability assessment, and transformer baseline comparison against Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT Pretraining Approach (ROBERTa) classifiers. A supplementary aspect–based sentiment analysis contextualized the positive-only design.

Results: The framework identified 1790 positive health outcome reports (1790/43,111, 4.15% prevalence), achieving 97.6% (488/500) precision (95% CI 95.7%-98.6%) and estimated 56.2% recall (95% CI 43.4%-67.9%). The reports described 6674 positive outcomes, distributed across 35 health aspects and 18 named disease conditions extending beyond weight loss: pain and inflammation reduction (1137/6674, 17%), type 2 diabetes improvement (977/6674, 14.6%), skin health (784/6674, 11.8%), and psychological well-being (731/6674, 11%). Over half (3355/6674, 50.3%) spanned multiple research objectives. Significant channel-level variation was observed (χ^2_{10}=927.5; P<.001), with positive outcome rates ranging from 1.32% to 10.40% (odds ratio 8.68, 95% CI 7.10-10.61). Transformer baselines achieved higher recall but lower precision, confirming their advantage for high-confidence corpus generation. A supplementary aspect-based sentiment analysis indicated a positive-to-negative ratio of approximately 4.6:1 (n=1003), with negative experiences (59/495, 11.9%) predominantly involving gastrointestinal and cardiovascular concerns.

XSL•FO
RenderX

Conclusions: This study presents, to our knowledge, the first validated, rule-based framework for extracting self-reported metabolic health outcomes from healthcasting YouTube comments at corpus scale. Unlike existing recall-oriented social media health classifiers, the precision-optimized design achieves the confidence threshold required for outcomes research without manual review. These findings demonstrate that expert-led health content comment sections constitute a scalable, complementary data source for monitoring real-world engagement with dietary interventions, with implications for public health surveillance, platform design, and health communication research.

(J Med Internet Res 2026;28:e94855) doi:10.2196/94855

KEYWORDS

healthcasting; therapeutic carbohydrate restriction; user-generated content; health outcomes; natural language processing; ontology engineering; precision-optimized extraction; YouTube; self-reported outcomes; metabolic health; digital health

Introduction

Background and Motivation

Ongoing advances in metabolic health research have identified insulin resistance and excessive glycemic variability as principal contributors to chronic systemic inflammation and metabolic stress [1]. Therapeutic carbohydrate restriction (TCR), encompassing ketogenic, low-carbohydrate, carnivore, and intermittent fasting approaches, reduces dietary carbohydrate to shift metabolic fuel use toward fatty acid oxidation and ketone body production [2]. TCR-based interventions have demonstrated clinically significant improvements in glycemic control, body composition, and cardiometabolic risk markers across multiple randomized controlled trials and systematic reviews [3-8], with a recent meta-analysis of 30 randomized controlled trials (3806 adults) reporting significant reductions in metabolic syndrome indicators [9]. Research has extended into neurological applications [10,11] and metabolic psychiatry, where pilot clinical data suggest that ketogenic interventions may improve psychiatric symptom severity in bipolar disorder and schizophrenia [12-16].

Despite this growing evidence base, several structural challenges constrain the conduct of large-scale clinical trials on dietary interventions. Dietary trials are inherently difficult to blind, compliance monitoring is resource-intensive, and long-term adherence remains a persistent methodological challenge [10,17]. Critically, because TCR interventions involve dietary and lifestyle modification rather than pharmaceutical compounds, there is no direct commercial entity positioned to sponsor large-scale efficacy trials comparable to those conducted for pharmacological interventions [6]. This funding asymmetry does not reflect a lack of scientific interest or clinical signal, but rather the structural economics of nutrition research [6].

Concurrently, a substantial population is adopting TCR-based dietary approaches outside formal clinical settings [18,19], informed by credentialled expert content disseminated through YouTube. Over the past decade, a distinct health communication phenomenon has emerged [20]: expert-led channels in which physicians, researchers, and clinicians share evidence-based dietary interventions directly with lay audiences at scale [21]. We term this phenomenon healthcasting: the systematic delivery of health education through video platforms by domain experts, bypassing traditional clinical and media gatekeeping structures [22]. We adopt this compound term to distinguish the specific phenomenon of expert-led health content creation with bidirectional outcome reporting from broader categories such as health podcasting or medical influencing. In the metabolic health domain, healthcasting channels focused on TCR have accumulated audiences in the millions, with comment engagement growing from a few hundred interactions per year in 2017 to more than 73,000 comments in 2024 across the 11 channels examined in this study. Because TCR interventions are dietary rather than pharmaceutical, they are uniquely amenable to self-directed implementation [18,19], making this domain one of the most developed examples of research-to-audience healthcasting. The purpose of this paper is to extract and analyze user-reported health outcomes within this specific approach, not to compare TCR with alternative interventions.

Beneath these videos, many viewers post comments reporting personal health changes, frequently including temporal markers suggesting longitudinal self-monitoring (eg, “after 5 weeks... my fatty liver is reversed”) [23]. While each comment is classified independently, the prevalence of temporal language provides indirect evidence that commenters report outcomes observed over weeks to months of dietary change. These comments constitute unsolicited, real-world, naturalistic health outcome data not available in any clinical registry, representing self-reported experiences of individuals who encountered expert content, acted on it, and publicly documented the results [24].

Several important caveats apply to the interpretation of self-reported health outcomes extracted from social media commentary. Such data are subject to selection bias [25] (individuals who experience positive outcomes may be more likely to comment), survivorship bias (those who discontinued may not return to report), recall bias [26], and the absence of clinical verification [24,27]. The data do not constitute clinical evidence in the conventional sense, cannot establish causal relationships between dietary interventions and health outcomes, and should not be interpreted as demonstrating clinical efficacy.

The purpose of this study is to identify which health conditions users of TCR-focused healthcasting content self-report as improved and to examine factors that may influence the distribution of these reports. The case of metabolic health and TCR was selected because it represents one of the most developed and active domains of healthcasting, with sufficient comment volume and content creator diversity to support computational analysis at scale [18,20].

XSL•FO
RenderX

Research Gap

Health information extraction from social media [28,29] has focused predominantly on pharmacovigilance and adverse drug event detection [30-33], with the social media mining for health applications shared tasks expanding from rule-based systems to large language models (LLMs) [34]. Research has also examined Reddit mental health communities [35] and YouTube health video quality [21,36,37], and YouTube video comments on dietary topics have been examined using text mining approaches [38]. Systematic mining of YouTube comment sections for self-reported health outcomes, particularly dietary interventions, has not been addressed. This represents a gap in both health informatics methodology and our understanding of how healthcasting content translates into reported health change at the population level.

In the methodological literature, existing classification systems have been optimized predominantly for balanced F_1 performance, with precision typically reported in the 80-90% range [39-41]. For applications requiring high-confidence corpus generation, where the downstream analysis depends on the validity of every included observation, this precision level is insufficient. A system that incorrectly classifies 1 in 10 or 1 in 5 comments as positive health outcomes would introduce systematic noise into any analysis of outcome distributions, disease-specific prevalence rates, or channel-level variation. The gap this study addresses is therefore twofold: the absence of a domain-specific framework for extracting self-reported health outcomes from healthcasting content, and the absence of a precision-optimized extraction methodology explicitly designed to generate validated corpora for downstream health outcomes research.

Research Questions

This study aims to address the following research questions (RQs):

- RQ1: what is the prevalence of self-reported positive health outcomes in YouTube comments on metabolic health content?
- RQ2: what types of health outcomes are most frequently reported, and how are they distributed across subjective, objective, and disease-specific categories?
- RQ3: does positive outcome reporting vary significantly across content creators, and what factors may explain this variation?
- RQ4: can a precision-optimized rule-based framework achieve sufficient classification accuracy for generating validated health outcome corpora from user-generated content?

Methods

Overview

The methodology comprises 3 construction phases, an integrated program of validation studies, and a supplementary contextualization analysis. The construction phases are (1) exploratory data analysis and corpus characterization, (2) ontology development through iterative pattern extraction, and (3) rule-based classification. Phase 3 is then stress-tested through five complementary validation studies: precision validation, recall estimation, external validation on held-out channels, interrater reliability assessment, and transformer baseline comparison. A supplementary aspect-based sentiment analysis (ABSA) contextualizes the primary framework's positive-only design. Figure 1 presents the overall framework architecture.

XSL•FO
RenderX

Figure 1. Architecture of the rule-based natural language processing framework for extracting self-reported positive health outcomes from English-language YouTube comments on metabolic health channels (N=43,111 unique comments, November 2013 to January 2026). The methodology comprises three construction phases: (1) exploratory data analysis, (2) ontology development, and (3) rule-based classification. The classifier is stress-tested through a program of Validation Studies (precision validation, recall estimation, external validation on held-out channels, interrater reliability, and transformer baseline comparison), and a supplementary analysis provides aspect-based sentiment contextualization of the positive-only design. The precision-optimized classifier uses conservative classification rules and extensive exclusion filtering. BERT: Bidirectional Encoder Representations from Transformers; LDA: latent dirichlet allocation; LLM: large language model; RO: research objective; NLP: natural language processing; RoBERTa: Robustly Optimized BERT Pretraining Approach.

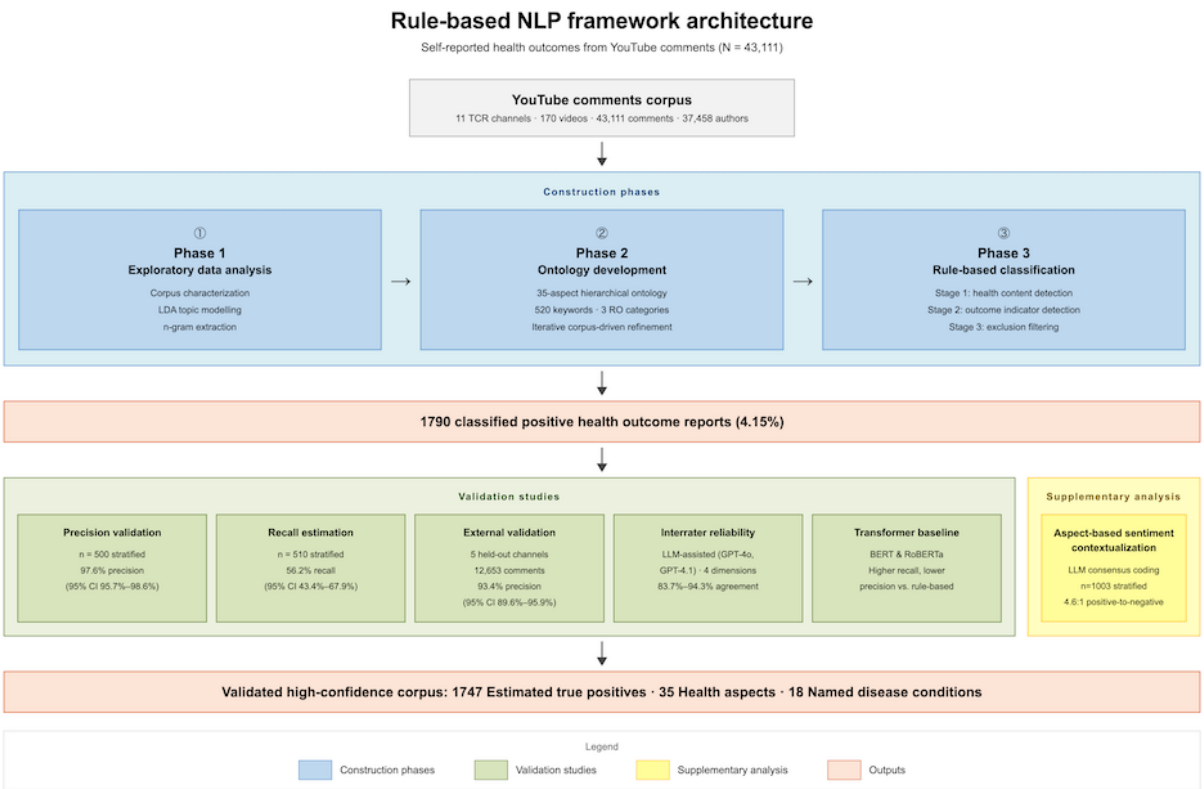

Research Design Overview

The framework was explicitly designed to maximize precision rather than recall. This design choice reflects the intended application: generating a high-confidence corpus of verified positive health outcomes suitable for downstream analysis. In health informatics applications, false positives (incorrectly classified outcomes) can lead to erroneous conclusions about treatment efficacy, whereas false negatives (missed outcomes) simply reduce statistical power without introducing systematic error. Following established guidance in clinical text mining [42], we prioritized precision when high-confidence annotations are required. This study follows the CREMLS (Consolidated Reporting Guidelines for Prognostic and Diagnostic Machine Learning Modeling Studies); the completed checklist is provided in Multimedia Appendix 1.

The exclusive focus on positive health outcomes reflects three considerations: positive outcomes provide the most linguistically distinctive targets, negative outcomes pose fundamentally different classification challenges (heterogeneous expressions requiring distinct approaches), and follower comment sections exhibit a structural positive bias.

The primary contributions of this work are: (1) a hierarchical ontology of 35 health aspects capturing subjective, objective, and disease-specific outcomes in TCR-focused healthcasting content; (2) a precision-optimized rule-based classification system achieving 97.6% (488/500) precision (95% CI 95.7%-98.6%); (3) a validated corpus of 1747 estimated true positive health outcome reports from 43,111 unique comments across 37,458 unique commenters; (4) comprehensive precision-recall characterization; (5) interrater reliability assessment using dual-model LLM-assisted annotation; (6) a domain-level analysis of healthcasting outcome patterns and channel-level variation; and (7) a supplementary ABSA contextualizing the positive-only extraction scope.

Ethical Considerations

This study analyzes publicly available YouTube comments accessed through the official YouTube data application programming interface (API) version 3, in compliance with the platform's terms of service and API use policies. No formal ethics committee review was sought, consistent with established practice in computational social media research involving publicly accessible data, where no interaction with users occurs, and no intervention is administered [43]. Several methodological and procedural safeguards were implemented to protect user privacy and ensure responsible data handling.

XSL•FO
RenderX

Data collection was limited to publicly posted comments that users understand to be visible to all internet users. No private messages, restricted content, or data requiring authentication were accessed. All processing was automated with no direct interaction with commenters, and no personally identifiable information was retained beyond publicly visible usernames used only for de-duplication. Informed consent was not required, as this study involved secondary analysis of publicly posted data with no direct interaction with or intervention upon users.

Comment excerpts presented in this study are reproduced only in truncated or paraphrased form to minimize the risk of reidentification through text search. The raw comment corpus is not included in the supplementary materials because YouTube's terms of service restrict the redistribution of bulk API-retrieved data. The classification code, ontology, and validation protocols are made available to enable methodological reproducibility without compromising user privacy.

No compensation was provided to any participants, as this study involved secondary analysis of publicly available data and did not involve direct interaction with commenters. No images of identifiable individuals are included in this manuscript. The study was conducted in accordance with the principles of the Declaration of Helsinki applicable to observational research involving publicly available data.

Data Collection

Comments were collected from 11 YouTube channels producing content on metabolic health and TCR. Channel selection criteria included: (1) medical or scientific credentials of content creators, (2) minimum subscriber threshold of 100,000, (3) focus on metabolic health topics, and (4) active comment sections. The YouTube Data API v3 was used to retrieve the 10 most-commented videos per channel and up to 2000 comments per video, yielding a raw corpus of 209,661 records from 110 videos. After removing duplicate records caused by API pagination (the YouTube Data API v3 returns nonunique results when paginating beyond available comments with relevance-based ordering), 43,111 unique comments were retained. Data collection was performed on January 2, 2026, capturing comments spanning November 7, 2013, to January 2, 2026. Table 1 presents corpus statistics by channel.

XSL•FO
RenderX

Table 1. Corpus statistics for 11 metabolic health YouTube channels included in a cross-sectional computational analysis of self-reported health outcomes (N=43,111 unique English-language comments, November 2013 to January 2026).

Channel[a]	Background[b]	Channel age (years)	Subscribers[c] (thousands)	Views (millions)	Comments[d]	Positive outcomes	Rate (%)
KenDBerryMD	Family medicine, MD (University of Tennessee)	14.8	3660	533	3970	413	10.40
Eric Berg DC	Chiropractor, DC (Palmer College)	16.7	14,500	3350	3993	282	7.06
Eric Westman	Internal and obesity medicine, MD, PhD (University of Wisconsin), MHS (Duke)	10.1	297	36	3978	250	6.28
Jason Fung	Nephrology, MD (University of Toronto)	14.2	1410	86	3963	286	7.22
Ben Bikman	Cell biology and physiology, PhD (East Carolina University)	9.0	191	9	3476	65	1.87
Nick Norwitz	Metabolism, PhD (Oxford), MD (Harvard)	11.8	854	55	3954	82	2.07
Anthony Chaffee MD	Neurosurgery, MD (Royal College of Surgeons)	11.3	536	109	3972	104	2.62
Shawn Baker MD	Orthopedic surgery, MD (Texas Tech)	11.8	383	83	3933	52	1.32
Dr. Robert Cywes	Bariatric and pediatric surgery, MD, PhD (University of Cape Town)	6.1	307	37	3932	102	2.59
Dr. Boz	Internal medicine, MD (University of South Dakota)	13.7	1180	184	3959	86	2.17
Mark Hyman	Functional medicine, MD (University of Ottawa)	18.5	1490	144	3980	68	1.71

[a]Channels ordered by positive outcome rate.

[b]The background section lists the primary professional credentials and degree-granting institution for each content creator.

[c]Subscriber counts and total channel views were updated on March 4, 2026, to reflect current values. YouTube reports subscriber counts rounded to three significant figures.

[d]Comments data were collected via the YouTube data application programming interface, version 3, on January 2, 2026 (10 most-commented videos per channel, up to 2000 comments per video).

Video selection was maximized for content relevance and comment volume: for each channel, the 10 most-commented videos were identified using the YouTube data API. Content validity was addressed through channel credential requirements, the engagement-based selection criterion, and the classification framework's exclusion filters. Critically, the unit of analysis is the viewer comment, not the video content itself; consequently, the classification framework's accuracy is independent of video content quality, and no formal content quality instrument was applied to the videos. Listing for all 110 video titles, URLs, and metadata is provided in Multimedia Appendix 2. Video durations ranged from 19 seconds to 115.8 minutes (mean 23.3, SD 24.4 minutes). Short-form videos (n=13, 11.8%) were included because they met the selection criterion and contained comparable health-related discourse.

Phase 1: Exploratory Data Analysis

Initial corpus exploration used topic modeling using latent Dirichlet allocation to identify thematic structures. N-gram analysis extracted frequently occurring bi-grams and tri-grams associated with health outcomes. Representative comments were sampled from each topic cluster to inform ontology development. This phase established the linguistic patterns characterizing health outcome reports in the corpus, distinguishing personal testimonials from general health discussions.

Corpus characterization revealed a right-skewed distribution of comment lengths (mean 32.5, SD 47 words; median 18, IQR 11-43 words), with engagement metrics confirming that most comments function as unsolicited declarations rather than conversational exchanges. Temporal analysis revealed

XSL•FO
RenderX

exponential growth, rising from 3408 comments in 2019 to 73,207 in 2024.

Phase 2: Ontology Development

The health outcome ontology was developed through an iterative, corpus-driven process in which data-derived linguistic patterns were combined with domain expert knowledge. Starting from latent Dirichlet allocation topic modeling and n-gram extraction conducted in Phase 1, candidate health concepts were identified from the most frequently occurring bi-grams and tri-grams co-occurring with outcome-indicative language (eg, “lost weight,” “blood sugar normalized,” and “pain is gone”). These candidates were then organized into a hierarchical structure of research objectives (ROs), each containing a set of thematically coherent aspects.

Three ROs were defined: RO1 captures subjective experiences (how users feel day-to-day), RO2 captures measurable biomarkers and anthropometric changes, and RO3 captures disease-level resolution (named conditions improved or reversed). The tiers are complementary and nonmutually exclusive, with 50.3% (n=3355) of 6674 reported positive outcomes spanning more than one RO.

Each aspect was defined with a unique identifier (eg, RO2.1), scope definition, inclusion keywords matched using whole-word regular expressions (case-insensitive), and exclusion patterns redirecting ambiguous matches to more specific aspects. The complete ontology comprises 35 aspects across 3 ROs, totaling 520 keywords. Table 2 presents the full structure; the complete keyword set is available in Multimedia Appendix 3.

XSL•FO
RenderX

Table 2. Health outcome ontology used in a rule-based natural language processing framework for extracting self-reported outcomes from English-language YouTube comments on metabolic health channels (N=43,111 unique comments, November 2013 to January 2026). The ontology comprises 35 aspects organized under three research objectives with 520 total keywords. Representative keywords are shown; the complete keyword set is available in multimedia files.

ID	Aspect name[a]	Domain; type	Scope and representative keywords
RO1: Subjective well-being (9 aspects): self-reported improvements in quality of life, symptoms, and subjective health status			
RO1.1	Cognitive function	Subjective; neurological	Brain fog, mental clarity, memory, focus, concentration, cognitive improvement
RO1.2	Energy and vitality	Subjective; metabolic	Energy levels, fatigue, tiredness, lethargy, stamina, vitality, no longer tired
RO1.3	Psychological well-being	Subjective; mental health	Anxiety, depression, mood, stress, mental health, happiness, calm, irritability
RO1.4	Sleep quality	Subjective; circadian	Sleep improvement, insomnia, sleep apnea, waking rested, deep sleep, better sleep
RO1.5	Appetite and satiety	Subjective; metabolic	Hunger, cravings, satiety, appetite control, sugar cravings gone, no longer hungry
RO1.6	Pain and inflammation	Subjective; musculoskeletal	Pain, chronic pain, back pain, joint pain, headache, migraine, swelling, stiffness
RO1.7	Digestive health	Subjective; gastrointestinal	Bloating, IBS[b], constipation, acid reflux, gut health, digestion improved, heartburn
RO1.8	Skin health	Subjective; dermatological	Acne, eczema, psoriasis, rash, skin cleared, skin tags, dermatitis, rosacea
RO1.9	Hormonal and menstrual health	Subjective; endocrine	Hormonal symptoms, menstrual cycle, PMS[c], perimenopause, hot flushes, libido
RO2: Tool-mediated validation (8 aspects): outcomes verified through measurement tools, clinical tests, or quantification			
RO2.1	Anthropometric changes	Objective; measured	Weight, pounds, kg, lbs lost, waist, BMI, body fat, dress size, inches, visceral fat
RO2.2	Glycemic control	Objective; lab biomarker	Blood sugar, A_{1C}, HbA_{1c}[d], fasting glucose, fasting insulin, CGM[e], glucometer reading
RO2.3	Blood pressure	Objective; measured	Blood pressure, systolic, diastolic, BP normalized, hypertension controlled, mmHg
RO2.4	Lipid profile	Objective; lab biomarker	Cholesterol, triglycerides, HDL[f], LDL[g], lipid panel, cholesterol improved, statins off
RO2.5	Inflammatory markers	Objective; lab biomarker	CRP, C-reactive protein, inflammatory markers, ESR[h], inflammation markers reduced
RO2.6	Liver function	Objective; lab biomarker	ALT[i], AST[j], liver enzymes, liver function test, fatty liver markers, liver normalized
RO2.7	Kidney function	Objective; lab biomarker	Creatinine, GFR[k], eGFR[l] improved, kidney function tests, creatinine normalized
RO2.8	Hormonal markers	Objective; lab biomarker	Testosterone, estrogen, thyroid (TSH/T3/T4), cortisol, hormonal lab values improved
RO3: Disease specificity (18 aspects): reported improvements in named medical conditions			
RO3.1	Type 2 diabetes	Disease; metabolic	Diabetes, diabetic, T2D, prediabetes, reversed diabetes, off metformin, off insulin
RO3.2	Fatty liver disease	Disease; hepatic	Fatty liver, NAFLD[m], NASH[n], liver disease, fatty liver reversed, hepatic steatosis
RO3.3	Cardiovascular disease	Disease; cardiac	Heart disease, heart failure, coronary artery disease, heart attack, cardiovascular
RO3.4	Hypertension	Disease; cardiovascular	Hypertension, high blood pressure, off blood pressure medication, BP controlled

XSL•FO
RenderX

ID	Aspect name[a]	Domain; type	Scope and representative keywords
RO3.5	PCOS	Disease; endocrine	PCOS, polycystic ovary syndrome, polycystic ovaries, PCOS symptoms improved
RO3.6	Neurodegenerative disease	Disease; neurological	Alzheimer, dementia, Parkinson, neurodegeneration, cognitive decline reversed
RO3.7	Chronic kidney disease	Disease; renal	Kidney disease, CKD, chronic kidney disease, kidney failure, renal function improved
RO3.8	Gout	Disease; metabolic	Gout, uric acid, gout attack gone, no more gout, uric acid normalized
RO3.9	Cancer	Disease; oncological	Cancer, tumor, remission, cancer markers, prostate cancer, cancer improved
RO3.10	Osteoporosis	Disease; musculoskeletal	Osteoporosis, bone density, bone loss, osteopenia, DEXA scan improved
RO3.11	Stroke	Disease; cerebrovascular	Stroke, TIA, mini stroke, stroke recovery, stroke risk reduced
RO3.12	ADHD	Disease; neurodevelopmental	ADHD, attention deficit, ADD, hyperactivity, ADHD symptoms improved, focus
RO3.13	Thyroid disease	Disease; endocrine	Thyroid, hypothyroid, hyperthyroid, Hashimoto's, thyroid medication reduced
RO3.14	Inflammatory bowel disease	Disease; gastrointestinal	Crohn, ulcerative colitis, IBD, Crohn in remission, colitis improved
RO3.15	Autoimmune disease	Disease; immunological	Autoimmune, lupus, multiple sclerosis, rheumatoid arthritis, celiac, autoimmune improved
RO3.16	Fibromyalgia and neuropathy	Disease; neurological	Fibromyalgia, neuropathy, nerve pain, peripheral neuropathy, numbness, tingling gone
RO3.17	Arthritis	Disease; musculoskeletal	Arthritis, osteoarthritis, arthritic, rheumatoid, joint disease, arthritis improved
RO3.18	Gallbladder disease	Disease; biliary	Gallbladder, gallstones, cholecystectomy, gallbladder attack, gallstones resolved

[a]Exclusion patterns redirect keyword matches to more specific aspects to prevent double-counting (eg, "arthritis" in a pain context is classified under RO3.17, not RO1.6). All keywords are matched case-insensitively with whole-word boundary constraints. Complete keyword sets and exclusion patterns are available in the supplementary materials repository.

[b]IBS: irritable bowel syndrome.

[c]PMS: premenstrual syndrome.

[d]HbA_{1c}: hemoglobin A_{1c}.

[e]CGM: continuous glucose monitor.

[f]HDL: high-density lipoprotein.

[g]LDL: low-density lipoprotein.

[h]ESR: erythrocyte sedimentation rate.

[i]ALT: alanine aminotransferase.

[j]AST: aspartate aminotransferase.

[k]GFR: glomerular filtration rate

[l]eGFR: estimated glomerular filtration rate.

[m]NAFLD: nonalcoholic fatty liver disease.

[n]NASH: nonalcoholic steatohepatitis.

Ontology Validation and Refinement

The ontology underwent 2 validation rounds (coverage testing and precision refinement) before deployment. A final manual review of 20 randomly sampled matches per aspect confirmed semantic validity before the ontology was locked for Phase 3 application.

Final ontology coverage showed that 30.1% (12,976/43,111) of comments contained at least one health-relevant keyword match; of these, 4.15% (n=1790) met all criteria for a definite, first-person, positive health outcome report.

Phase 3: Classification Framework

Overview

The classification system implements a 3-stage pipeline designed to maximize precision. Algorithm 1 in Figure 2 presents the formal classification procedure.

XSL•FO
RenderX

Figure 2. Algorithm 1: precision-optimized health outcome classification.

```
Input: Comment corpus C, Ontology O with aspects A and exclusions E
Output: Classified positives P with aspect assignments
Initialize P ← ∅
for each comment c ∈ C do
    // Stage 1: Health Content Detection
    aspects ← MatchKeywords(c, O)
    if aspects = ∅ then continue
    // Stage 2: Outcome Indicator Detection
    indicators ← DetectOutcomeIndicators(c)
    if indicators = ∅ then continue
    // Stage 3: Exclusion Filtering
    if MatchesExclusion(c, E) then continue
    if IsQuestion(c) OR IsThirdParty(c) then continue
    if IsHypothetical(c) OR IsNegated(c) then continue
    P ← P ∪ {(c, aspects, indicators)}
return P
```

Stage 1: Health Content Detection

Comments are filtered for health-related content using keyword matching against the ontology vocabulary.

Stage 2: Outcome Indicator Detection

Health-related comments are analyzed for positive outcome indicators, including quantified changes (eg, “lost 30 pounds”), symptom cessation (eg, “pain gone”), explicit improvement language (eg, “reversed my diabetes”), and temporal improvement markers (eg, “no longer need medication”).

Stage 3: Exclusion Filtering

Candidate positives are filtered through exclusion patterns removing: (1) questions rather than statements, (2) third-party reports, (3) hypothetical or aspirational language, (4) negated outcomes, and (5) general health advice. This conservative approach implements the precision-optimized design philosophy.

Handling of ambiguous statements: Sarcasm is not specifically detected but is unlikely to pass all 3 stages due to the requirement for co-occurring health vocabulary, outcome indicators, and absence of exclusion patterns. Implied improvements without explicit outcome language are intentionally excluded. Narrative sequences mixing positive and negative outcomes may partially pass but are addressed by negation filters. The validation results confirm that such cases account for a minority of false positives (complete classification rule patterns are provided in Multimedia Appendix 4).

Validation Studies

Overview

The classification framework was stress-tested through 5 complementary validation studies, each probing a different dimension of performance: internal precision on the development corpus, recall and coverage through weighted negative sampling, out-of-sample generalizability on held-out channels, rater agreement via LLM-assisted interrater reliability assessment, and head-to-head comparison against fine-tuned transformer baselines. Each study is described in turn below; full protocols and results for the external validation and transformer baseline are reported in Multimedia Appendices 5 and 6, respectively.

Precision Validation

Classification precision was validated through manual coding of a stratified random sample. Sample size was calculated for 95% confidence level with 4% margin of error, yielding n=500 samples stratified by RO. Each sample was coded on five dimensions: (1) is_positive_outcome (yes, no, or unclear), (2) is_personal (yes or no), (3) is_definite (yes or no), (4) aspect_correct (yes, partial, or no), and (5) free-text notes. Wilson score [44] confidence intervals were computed for all proportions.

Recall Estimation

Recall was estimated through stratified negative sampling. A stratified random sample of n=510 comments from the nonclassified pool (41,321 comments not identified as positive outcomes) was drawn using disproportionate allocation across 3 comment-length strata (short: fewer than 50 words, n=300; medium: 50 to 150 words, n=150; long: more than 150 words, n=60) with proportional allocation across the 11 channels within each stratum. Each sampled comment was manually reviewed by the first author to identify false negatives: true positive outcomes missed by the classification system. Sampling weights reflecting the population proportion in each length stratum were applied to compute unbiased weighted false-negative rates, which were then extrapolated to estimate total missed positives and to calculate recall. Wilson score CIs were used throughout.

External Validation on Held-Out Channels

To address the concern that the 35-aspect ontology was iteratively refined on the same corpus used for classification, and thereby to test generalizability beyond the development channels, external validation was conducted on 12,653

XSL•FO
RenderX

comments collected from 5 independent YouTube channels with zero overlap with the development corpus. The 5 held-out channels (Georgia Ede MD, Robert Kiltz MD, Sten Ekberg DC, Chris Palmer MD, and Ted Naiman MD) were selected by an independent co-author based on topical relevance to ketogenic, carnivore, and broader metabolic health dietary content. For each channel, comments were collected from the 10 most-commented videos using the YouTube data API, following the same protocol as the development corpus. The classifier was applied without modification. All comments classified as positive health outcomes were exhaustively verified through manual coding to determine precision (census approach). For recall estimation, a stratified random sample of nonpositive comments (up to 100 per channel; random seed 42) was manually coded to identify false negatives, with channel-level false-negative rates extrapolated to the full negative population, and 95% Wilson score CIs reported. The full external validation protocol and results are presented in Multimedia Appendix 5.

Interrater Reliability Assessment

To address the limitation of single-coder validation, we used an LLM-assisted annotation validation protocol using 2 independent LLM coders (GPT-4o and GPT-4.1) (OpenAI) as second annotators [45,46]. This approach, increasingly adopted in computational linguistics and health informatics, provides a systematic assessment of interrater reliability while maintaining full reproducibility.

From the 500-sample precision validation set, 28 exemplars were selected through purposive stratified sampling to serve as few-shot coding examples. The stratification covered 6 coding outcome categories: clear true positives (n=10), clear negatives or false positives (n=5), unclear or ambiguous cases (n=5), positive but not personal outcomes (n=3), positive but not definite outcomes (n=2), and aspect assignment issues (n=3). This selection ensured representation of all 3 ROs, 10 of 11 channels, all 6 outcome categories, and deliberate overrepresentation of minority classes and boundary cases, which is standard practice in few-shot prompt design.

Both models received identical structured prompts containing: (1) the coding task definition with detailed guidelines for all four coding dimensions, (2) the complete 35-aspect ontology reference, (3) all 28 few-shot exemplars with the researcher's ground-truth codings, and (4) the comment text with automated classification details but without the researcher's manual codings (full prompt provided in Multimedia Appendix 6). Processing used temperature 0.0 (deterministic output) in batches of 10 comments. Agreement was computed on the remaining 472 test samples (exemplars excluded to prevent circular validation).

To mitigate anchoring effects, exemplars spanned 6 coding outcome categories (including clear negatives and errors), and the prompt instructed models to code independently based on the comment text, explicitly stating that the automated classification may be incorrect.

Cohen κ [47] was computed for each coding dimension across 3 comparison pairs (human vs GPT-4o, human vs GPT-4.1, and GPT-4o vs GPT-4.1). Because the validation set's 90.4% positive class prevalence creates a κ paradox [48,49], we report the prevalence index and bias index alongside each κ value to decompose the paradox. Raw percent agreement and Cohen κ are the primary reliability metrics, interpreted using the Landis and Koch [47] framework.

Transformer Baseline Comparison

To test whether the precision advantage of the rule-based framework comes at an unnecessary cost to recall compared with learned representations trained on the same data, a head-to-head baseline comparison was conducted against two pretrained transformer models: Bidirectional Encoder Representations from Transformers (BERT)-base-uncased and Robustly Optimized BERT Pretraining Approach (RoBERTa)-base (Hugging Face). Both models were fine-tuned on the combined precision-validation and recall-expansion datasets (n=836 unique manually-coded comments; 347 positive and 489 negative) using stratified 5-fold cross-validation. Standard pretrained weights and default fine-tuning hyperparameters were used; no hyperparameter search was performed, because the purpose of this analysis was a fair comparison against a reasonable learned baseline rather than an optimized benchmark. Performance was evaluated on 4 metrics (precision, recall, F_1-score, and receiver operating characteristic-area under the curve) averaged across folds, and, separately, on the 326-comment precision-validation subset and on the 27 false negatives identified through the recall-expansion sample. The rule-based framework was compared against both transformer baselines on each metric. The full experimental protocol, hyperparameters, and per-fold results are presented in Multimedia Appendix 7.

Supplementary Analysis: Aspect-Based Sentiment Contextualization

This supplementary analysis is reported here, rather than as a coequal phase of the primary methodology, for 2 reasons. First, it uses a methodologically distinct procedure (LLM consensus coding) that cannot be directly compared to the validation regime applied to the rule-based framework. Second, its purpose is to explore and contextualize the positive-only design of the primary framework, not to estimate the prevalence of negative outcomes in the underlying population. Claims derived from this analysis are therefore treated as indicative rather than confirmatory throughout the Discussion section.

Because the classification framework extracts only positive health outcomes (as outlined in the Research Design Overview section), a supplementary analysis was conducted to contextualize this scope decision by characterizing the broader sentiment landscape of the corpus. ABSA was used to quantify the distribution of positive, negative, neutral, and mixed health sentiment across comments, providing an empirical basis for evaluating whether the positive-only focus omits a substantial volume of negative health experiences.

ABSA extends document-level sentiment analysis by identifying specific aspects (topics or entities) within a text and assigning sentiment to each aspect independently [50,51]. This granularity is essential for health-related comments, which frequently contain mixed sentiment. For example, a single comment

XSL•FO
RenderX

reporting weight loss improvement alongside gastrointestinal discomfort.

A stratified random sample of 1000 comments was drawn from the full corpus, proportional to channel contribution. Two independent LLMs (GPT-4o and GPT-4.1) were prompted to perform ABSA on each comment, classifying it as health-related or nonhealth-related and, for health-related comments, identifying health aspects and assigning aspect-level sentiment (positive, negative, neutral, or mixed). The dual-model design serves as a form of interrater reliability assessment: only comments in which both models agree on health-relatedness and sentiment classification are included in the consensus analysis, yielding conservative yet high-confidence sentiment estimates. The complete ABSA prompt, including the task definition, coding guidelines, and a few-shot exemplar, is provided in Multimedia Appendix 8.

Results

Classification Performance

The framework classified 1790 comments (1790/43,111, 4.15% of the corpus) as containing definite positive health outcomes. Table 3 presents the complete validation metrics.

Table 3. Classification validation results for a rule-based natural language processing framework applied to English-language YouTube comments on 11 metabolic health channels (N=43,111 unique comments, November 2013 to January 2026). Precision validated on n=500 stratified random samples; recall estimated from n=510 stratified negative samples. Wilson score 95% CIs.

Metric	Value	95% CI (%)
Precision, % (n=500)	97.6	95.7-98.6
Recall, % (n=510)	56.2	43.4-67.9
F_1-score, %	28.3	—[a]
True positives (validated)	488/500	—
False positives	11/500	—
False negatives (in sample)	27/510	—
First-person testimony rate, %	97	95.2-98.1
Definite outcome rate, %	88.4	85.3-90.9
Aspect assignment accuracy (strict), %	90.8	88-93
Aspect assignment accuracy (lenient), %	97	95.2-98.1

[a]Not applicable.

The expanded recall estimation (n=510 stratified sample) identified 27 false negatives (5.3% raw rate; 3.29% weighted rate), yielding an estimated recall of 56.2% (95% CI 43.4%-67.9%) when extrapolated to the full nonpositive pool. False negatives varied across channels (χ^2_{10}=28.8; P=.001) and comment length strata (χ^2_2=19.4; P<.001), with KenDBerryMD (16.7%) and Eric Berg DC (11.4%) showing the highest channel rates and medium-length (10%) and long (11.7%) comments generating more false negatives than short comments (1.7%). The dominant miss reason was structural pattern mismatch (23 of 27, 85%), indicating that the classifier keyword dictionary is adequate, but its syntactic pattern set does not capture all expression forms. Applying precision estimates to the classified corpus yields 1747 estimated true-positive health outcome reports (95% CI 1713-1764).

Error Analysis

To characterize the framework's failure modes, we examined all false positives from the precision validation (n=11 across 9 unique comments) and all false negatives from the recall estimation (n=27). Three systematic categories of false positive errors were identified: third-party outcome references (4 of 11 cases, 36%), where comments described health improvements experienced by family members or acquaintances rather than the commenter; negative overall trajectory contexts (4 of 11, 36%); and nonspecific or advice-based language (3 of 11, 27%).

Among false positives, the dominant pattern was positive signals embedded in negative overall trajectories (4/11, 36%), followed by nonspecific or advice-based language (3/11, 27%) and third-party reports (2/11, 18%). These cases require discourse-level sentiment analysis beyond the current sentence-level pattern matching.

False negative analysis revealed structural pattern mismatch as the dominant miss mechanism (23/27, 85%), indicating that the keyword dictionary is adequate, but syntactic patterns do not capture all expression forms. The most frequently missed aspects were RO2.1 (general well-being, 17/27), RO1.8 (energy, 6/27), and RO1.2 (body composition, 5/27). Targeted syntactic rule expansion, rather than vocabulary expansion, represents the primary pathway to improved recall (Table 4).

XSL•FO
RenderX

Table 4. Error analysis for a rule-based natural language processing framework classifying self-reported health outcomes in English-language YouTube comments on metabolic health channels (N=43,111 unique comments, November 2013 to January 2026). False positives (n=11) came from precision validation of 500 samples; false negatives (n=27) came from recall estimation of 510 samples.

Type	Category	Errors, n (%)	Suggested remedy
FP[a]	Third-party references	4 (36)	Expand person-reference filters
FP	Negative trajectory context	4 (36)	Discourse-level sentiment analysis
FP	Nonspecific or advice language	3 (27)	Tighter personal experience requirements
FN[b]	Borderline or debatable positives	7 (26)	Broader outcome definitions (recall trade-off)
FN	Colloquial symptom language	10 (37)	Vocabulary expansion
FN	Implicit emotional language	8 (30)	ML[c]-based semantic classification
FN	Missed pattern coverage	2 (7)	Rule refinement

[a]FP: false positive; n=11.

[b]FN: false negative; n=27.

[c]ML: machine learning.

External validation was conducted on 12,653 comments from 5 YouTube channels not included in the development corpus, selected by an independent co-author. The classifier achieved 93.4% precision (227/243; 95% CI 89.6%-95.9%) on the external corpus, with CIs overlapping those from the development corpus (97.6%), confirming generalizability within the metabolic health domain. Recall was estimated at 50.1% (95% CI 31.4%-59.1%), consistent with the development corpus. Full external validation protocol and results are presented in Multimedia Appendix 5.

Interrater Reliability

Table 5 presents the interrater reliability results across 4 coding dimensions and 3 comparison pairs, with raw Cohen κ and percentage agreement as primary metrics.

Table 5. Interrater reliability for validation of a rule-based natural language processing framework classifying self-reported health outcomes in English-language YouTube comments on 11 metabolic health channels (N=43,111 unique comments, November 2013 to January 2026). Reliability was assessed across four coding dimensions and three comparison pairs. Cohen κ and percent agreement are reported as primary metrics. Prevalence index (PI) [49] and bias index (BI) [52] quantify the κ paradox components; K interpretation follows Landis and Koch [47]. The validation set was drawn from classifier-positive comments, producing approximately 90% positive prevalence that depresses κ values via the κ paradox [48,49].

Comparison	Dimension	Agree, n (%)	Cohen κ	Interpretation	PI[a]	BI[b]
Human vs GPT-4o	Positive health outcome	394/471 (83.7)	0.297	Fair	0.892	0.076
Human vs GPT-4o	First-person testimony	463/472 (98.1)	0.658	Substantial	0.962	0.019
Human vs GPT-4o	Definite outcome	316/472 (66.9)	0.130	Slight	0.788	0.242
Human vs GPT-4o	Aspect assignment	289/472 (61.2)	0.106	Slight	0.896	0.326
Human vs GPT-4.1	Positive health outcome	410/471 (87)	0.332	Fair	0.892	0.045
Human vs GPT-4.1	First-person testimony	468/471 (99.4)	0.839	Almost perfect	0.966	0.006
Human vs GPT-4.1	Definite outcome	389/472 (82.4)	0.205	Slight	0.788	0.040
Human vs GPT-4.1	Aspect assignment	267/472 (56.6)	0.082	Slight	0.896	0.373
GPT-4o vs GPT-4.1	Positive health outcome	445/472 (94.3)	0.771	Substantial	0.767	0.044
GPT-4o vs GPT-4.1	First-person testimony	463/472 (98.1)	0.706	Substantial	0.928	0.013
GPT-4o vs GPT-4.1	Definite outcome	375/472 (79.4)	0.476	Moderate	0.305	0.201
GPT-4o vs GPT-4.1	Aspect assignment	385/472 (81.6)	0.655	Substantial	0.523	0.072

[a]PI: prevalence index.

[b]BI: bias index.

On the primary coding dimension (positive health outcome identification), raw percent agreement was high (83.7%-94.3%), but Cohen κ values were lower (0.297-0.771), reflecting the well-documented κ paradox [48,49]: the validation set's 90.4% positive prevalence leaves limited headroom for κ above chance. Bias Index values (0.044-0.076) confirm that low κ is driven by prevalence rather than systematic rater bias.

For personal testimony identification, agreement was near-ceiling across all pairs. GPT-4.1 achieved the highest κ observed in this study (κ=0.839, 99.4% agreement), while GPT-4o achieved substantial agreement (κ=0.658, 98.1%). The lower κ for GPT-4o, despite 98.1% agreement, again reflects the prevalence paradox: with 98% of comments coded as

XSL•FO
RenderX

first-person testimony, κ is constrained even at very high observed agreement.

On secondary dimensions (definiteness and aspect correctness), agreement was lower, with systematic directional bias: both LLMs applied stricter evidentiary standards than the human coder (McNemar $P<.001$ and $P=.048$), consistent with measurement bias characterized in the LLM Bias Audit (Multimedia Appendix 9).

The cross-model comparison provides the strongest reliability evidence: two independent architectures achieved substantial agreement on positive outcome identification (κ=0.771, 94.3%), the highest human-level or cross-model κ for this dimension. This cross-architecture convergence suggests that the coding task is well-specified: two independent systems, given the same instructions, reach similar conclusions. The cross-model κ is less affected by the prevalence paradox because both models have more balanced marginal distributions than the human-vs-LLM comparisons (PI=0.767 vs 0.892).

Prevalence and Distribution by Research Objective

The raw prevalence of classified positive outcomes was 4.15% (1790/43,111). Adjusted for precision, the estimated true positive prevalence is 4.05%. Outcomes were distributed across ROs as shown in Table 6. Of these, 50.3% of positive outcome comments (n=3355) spanned multiple ROs, indicating users frequently report improvements across multiple health dimensions simultaneously.

Table 6. Distribution of positive health outcomes by research objective, extracted from English-language YouTube comments on 11 metabolic health channels using a rule-based natural language processing framework (n=6674 positive outcomes in n=1790 positive reports among N=43,111 unique comments, November 2013 to January 2026). Percentages sum to more than 100% because 50.3% of outcomes span multiple research objectives. Wilson score 95% CIs.

RO[a]	Description	Positive outcomes (n=6674), n (%; 95% CI)
RO1	Subjective well-being	3456 (51.8; 50.6-53)
RO2	Tool-mediated validation	5350 (80.2; 79.2-81.1)
RO3	Disease specificity	2032 (30.5; 29.4-31.6)

[a]RO: research objective.

Health Aspect Analysis

Table 7 presents the top 10 health aspects by frequency. Anthropometric changes (primarily weight loss) dominated at 73% (4870/6674) of positive outcomes, consistent with the metabolic health focus of the source content. Pain and inflammation reduction (1137/6674, 17%) and type 2 diabetes improvement (977/6674, 14.6%) were the second and third most reported outcomes, suggesting clinically significant health impacts beyond aesthetic weight changes.

Table 7. Top 10 most frequently reported health aspects among self-reported positive outcomes extracted from English-language YouTube comments on 11 metabolic health channels (n=1790 reports of 6674 positive outcomes among N=43,111 unique comments, November 2013 to January 2026). Percentages computed relative to total positive outcomes. Wilson score 95% CIs.

#	ID	Aspect	Positive outcomes, n (%; 95% CI)
1	RO[a] 2.1	Anthropometric changes	4870 (73; 71.9-74.1)
2	RO1.6	Pain and inflammation	1137 (17; 16.2-18)
3	RO3.1	Type 2 diabetes	977 (14.6; 13.8-15.5)
4	RO1.8	Skin health	784 (11.8; 11-12.5)
5	RO1.3	Psychological well-being	731 (11; 10.2-11.7)
6	RO1.5	Appetite and satiety	677 (10.1; 9.4-10.9)
7	RO1.7	Digestive health	664 (10; 9.3-10.7)
8	RO1.2	Energy and vitality	651 (9.8; 9.1-10.5)
9	RO2.2	Glycemic control	564 (8.5; 7.8-9.1)
10	RO2.3	Blood pressure	548 (8.2; 7.6-8.9)

[a]RO: research objective.

Channel-Level Variation

Significant variation in positive outcome rates was observed across channels ($\chi^2_1=927.5$; $P<.001$), as shown in Table 8. Rates ranged from 1.32% (Shawn Baker, MD) to 10.40% (KenDBerryMD), yielding an odds ratio of 8.68 between the highest and lowest channels. Cramér $V=0.147$ indicates a small but statistically significant effect size, suggesting that while channel-level differences exist, they explain a modest proportion of total variance in outcome reporting.

XSL•FO
RenderX

Table 8. Channel-level variation in positive health outcome rates in a cross-sectional computational analysis of self-reported outcomes from English-language YouTube comments on 11 metabolic health channels (N=43,111 unique comments, November 2013 to January 2026). Channels ordered by descending positive outcome rate. Wilson score 95% CIs.

Channels	Positive outcomes, n/N (%; 95% CI)
KenDBerryMD	413/3970 (10.40; 9.49-11.39)
Jason Fung	286/3963 (7.22; 6.45-8.06)
Eric Berg DC	282/3993 (7.06; 6.31-7.90)
Eric Westman	250/3978 (6.28; 5.57-7.08)
Anthony Chaffee MD	104/3972 (2.62; 2.17-3.16)
Dr. Robert Cywes MD	102/3932 (2.59; 2.14-3.14)
Dr. Boz	86/3959 (2.17; 1.76-2.67)
Nick Norwitz	82/3954 (2.07; 1.67-2.57)
Ben Bikman	65/3477 (1.87; 1.47-2.38)
Mark Hyman	68/3980 (1.71; 1.35-2.16)
Shawn Baker MD	52/3933 (1.32; 1.01-1.73)
Overall	1790/43,111 (4.15; 3.96-4.35)

Outcome Category Distribution

Analysis of outcome indicator types (Table 9) revealed that quantified changes (eg, “lost 30 pounds,” “A1C dropped to 5.4”) comprised 74.4% (1331/1790) of positive outcomes. Symptom cessation reports (eg, “joint pain gone”) accounted for 14.5% (259/1790), explicit improvement language (eg, “feel so much better”) for 11.8% (212/1790), and disease reversal or remission claims (eg, “reversed my type 2 diabetes”) for 6.4% (114/1790). Medication discontinuation (eg, “off all medications”) represented 3.9% (69/1790), and temporal improvements (eg, “since starting keto…lost 20 pounds”) represented 2.3% (42/1790) of outcomes reported.

Table 9. Distribution of outcome indicator categories in a cross-sectional computational analysis of self-reported health outcomes from English-language YouTube comments on 11 metabolic health channels (n=1790 positive outcomes from N=43,111 unique comments, November 2013 to January 2026). Categories are not mutually exclusive; a single comment may contain multiple indicator types.

Outcome category	Positive outcomes, n (%)
Quantified change	1331 (74.4)
Symptom cessation	259 (14.5)
Explicit improvement	212 (11.8)
Reversal or remission	114 (6.4)
Medication discontinuation	69 (3.9)
Temporal improvement	42 (2.3)

Sentiment Contextualization: ABSA

A supplementary ABSA was conducted to contextualize the positive-outcome findings within the broader health discourse of the corpus [50,51].

Intermodel agreement on health-related classification was 93.1% (915/983), with sentiment agreement of 87.6% (495/565), indicating acceptable coding consistency for an exploratory contextualization analysis.

Table 10 presents the consensus sentiment distribution, the subset of health-related comments where both models agreed on sentiment classification. Among 495 consensus-coded health-related comments, positive sentiment accounted for 54.7% (271/495), negative for 11.9% (59/495), neutral for 15.6% (77/495), and mixed for 17.8% (88/495), yielding a positive-to-negative ratio of 4.6:1.

XSL•FO
RenderX

Table 10. Aspect-based sentiment analysis (ABSA) consensus sentiment distribution among health-related English-language YouTube comments on 11 metabolic health channels (n=495 consensus-coded comments from a stratified sample of 1003 drawn from N=43,111 unique comments, November 2013 to January 2026). Dual-model classification (GPT-4o and GPT-4.1) with consensus defined as agreement on both health-relatedness and sentiment polarity.

Sentiment[a]	Comments, n (%)	Estimated full corpus
Positive	271 (54.7)	~63,800
Negative	59 (11.9)	~13,900
Neutral	77 (15.6)	~18,200
Mixed	88 (17.8)	~20,800

[a]Consensus: both GPT-4o and GPT-4.1 agreed on sentiment classification. Corpus estimates extrapolated from 64.4% health-related rate in the stratified sample (n=983). Positive-to-negative ratio=4.6:1.

Table 11 presents the breakdown of the consensus-negative aspect. Gastrointestinal issues (n=36) and cardiovascular concerns (n=22, primarily LDL cholesterol elevations) were the most frequent negative aspects, followed by pain and inflammation (n=14) and energy and mood disturbances (n=12), aligning with documented adaptation effects during carbohydrate restriction transitions.

Table 11. Consensus-negative health aspects identified by dual-model aspect-based sentiment analysis (ABSA) of English-language YouTube comments on 11 metabolic health channels (N=43,111 unique comments, November 2013 to January 2026). Both GPT-4o and GPT-4.1 agreed on negative sentiment classification for each aspect listed.

Health aspect	Comments reaching consensus[a], n	Clinical context
Digestive	36	GI[b] adaptation
Cardiovascular	22	LDL[c] concerns
Pain and inflammation	14	Adaptation effects
Energy and mood	12	Transition fatigue
Neurological	12	Keto adaptation
Blood sugar	11	Glycemic worsening
Weight change	11	Weight stall or gain
Other (sleep, diet adherence, mental health, medication, general well-being, cancer, skin, autoimmune, and hormonal)	63	Various

[a]Consensus-negative: both GPT-4o and GPT-4.1 independently classified the aspect sentiment as negative. Total exceeds 59 comments because some comments contain multiple negative aspects.

[b]GI: gastrointestinal.

[c]LDL: low-density lipoprotein.

The 4.6:1 positive-to-negative ratio in this sample is consistent with the expected self-selection dynamics of the channels studied, while the presence of negative experiences at a meaningful rate suggests that the positive predominance of the primary framework is not solely an artefact of its positive-only scope. This finding is exploratory: generalization to the full corpus would require applying the same rigor used for the primary framework.

Table 12 presents aspect-level ratios, revealing that the positive predominance is not uniformly distributed. Weight change (8.9:1) and general well-being (7.1:1) exhibit the strongest positive skew, while digestive health (0.8:1), neurological symptoms (0.5:1), and hormonal concerns (0.3:1) are negative-dominant, reflecting known adaptation effects. The cardiovascular domain shows near-parity (1.0:1), consistent with the contested nature of LDL cholesterol responses to high-fat diets.

XSL•FO
RenderX

Table 12. Aspect-level sentiment distribution by health domain in English-language YouTube comments on 11 metabolic health channels (GPT-4.1 classification, n=627 health-related comments from a stratified sample of 1003 drawn from N=43,111 unique comments, November 2013 to January 2026). Positive-to-negative ratio computed for aspects with at least 5 negative mentions.

Health aspect[a]	Positive, n	Negative, n	Neutral, n	Mixed, n	Positive to negative ratio	Negative, n/N (%)
Weight change	241	27	27	15	8.9:1	27/310 (8.7)
General well-being	213	30	13	2	7.1:1	30/258 (11.6)
Skin	21	5	1	0	4.2:1	5/27 (18.5)
Medication	30	10	34	2	3.0:1	10/76 (13.2)
Energy and mood	52	18	1	3	2.9:1	18/74 (24.3)
Blood sugar	51	19	45	1	2.7:1	19/116 (16.4)
Diet adherence	96	38	60	8	2.5:1	38/202 (18.8)
Sleep	22	9	1	0	2.4:1	9/32 (28.1)
Pain and inflammation	54	26	5	0	2.1:1	26/85 (30.6)
Cardiovascular	33	32	41	6	1.0:1	32/112 (28.6)
Digestive	46	56	27	2	0.8:1	56/131 (42.7)
Mental health	7	11	1	0	0.6:1	11/19 (57.9)
Neurological	7	15	3	0	0.5:1	15/25 (60.0)
Hormonal	2	6	8	1	0.3:1	6/17 (35.3)

[a]Italicized aspects are negative-dominant (ratio <1.0), that is, the percentage of all mentions within that aspect classified as negative. GPT-4.1 single-model classification (n=627 health-related comments, n=1602 total aspect mentions).

To contextualize the rule-based framework's performance, BERT-base-uncased and RoBERTa-base classifiers were trained on the combined validation datasets (n=836; using a five-fold stratified cross-validation). Both transformer models achieved substantially higher recall (93.4% and 95.7%) but lower precision (87% and 88.2%) than the rule-based framework (97.6%), confirming the design advantage of precision optimization for high-confidence corpus generation. Full transformer baseline comparison results are presented in Multimedia Appendix 7.

Table 13 positions this work relative to prior approaches, with the key differentiation being explicit precision optimization for high-confidence corpus generation.

Table 13. Comparison of the proposed rule-based natural language processing framework with prior social media health text classification approaches. The proposed framework was applied to English-language YouTube comments on 11 metabolic health channels (N=43,111 unique comments, November 2013 to January 2026). Precision and recall are reported for the positive class (health event).

Study	Platform	Approach	Precision	Recall	Ontology
Sarker and Gonzalez [53]	Twitter	ML[a] (SVM)[b]	85%	82%	ADR-based
Nikfarjam et al [30]	Twitter	CRF[c]	87%	71%	UMLS
Golder et al [43]	Multi	ML (Review)	80-90%	Varies	Varies
Magge et al [54]	Twitter	DL[d] (RoBERTa)	63%[e]	63%[e]	MedDRA
This study	YouTube	Rule-based	97.6%	56.2%	Custom

[a]ML: machine learning.

[b]SVM: support vector machine.

[c]CRF: conditional random field.

[d]DL: deep learning.

[e]Evaluated at a realistic 7% positive rate. At 2% positive rate (comparable to this study's 4.15%), precision dropped to 21% and recall to 25%.

Discussion

Principal Findings

This study set out to determine whether self-reported positive health outcomes can be systematically extracted from YouTube comments on metabolic health healthcasting channels, and, if so, to characterize their prevalence, distribution across health aspects, variation across content creators, and the classification accuracy required to generate a validated corpus.

XSL•FO
RenderX

RQ1: What Is the Prevalence of Self-Reported Positive Health Outcomes in YouTube Comments on Metabolic Health Content?

The classification framework identified 1790 definite positive self-reported health outcome comments from a corpus of 43,111 unique comments across eleven metabolic health healthcasting channels, corresponding to a raw prevalence of 4.15%. These reports are unsolicited first-person accounts of health improvement, posted spontaneously under creator videos rather than in response to surveys or prompts, making them a distinctive source of real-world health data. The observed prevalence remained stable during external validation on an independent set of 5 held-out channels, suggesting that positive outcome reporting occurs at a consistent, detectable frequency across healthcasting communities in this domain. A supplementary sentiment analysis of the broader health discourse confirmed that positive outcome reporting constitutes a substantial but not overwhelming share of health-related conversation, with a positive-to-negative ratio of approximately 4.6:1.

RQ2: What Types of Health Outcomes Are Most Frequently Reported, and How Are They Distributed Across Subjective, Objective, and Disease-Specific Categories?

The outcome landscape that emerged from the corpus was considerably broader than the weight-loss framing commonly associated with carbohydrate restriction [3,5]. Positive outcomes were systematically organized using a newly developed 35-aspect hierarchical ontology aligned with 3 complementary ROs: subjective well-being (how people feel day to day), objectively measurable changes (such as blood glucose and body composition), and named disease-specific improvements. Over half of positive outcome comments described improvements across multiple health dimensions, suggesting that commenters experience and report broad, interconnected health changes rather than isolated improvements. While changes in weight and body composition dominated, the most frequently reported outcomes also included reductions in pain and inflammation, improvements in glycemic control, better skin health, enhanced psychological well-being, and improved regulation of appetite, digestive health, and energy levels. Reports also identified improvements or remission across eighteen medical conditions, including type 2 diabetes (often described as reversed or resolved), fatty liver disease, hypertension, and polycystic ovary syndrome. Most reports described quantified changes (such as specific weight loss or improved blood markers), though symptom cessation, explicit improvement language, and disease reversal or remission claims were also common. Notably, medication discontinuation reports were present in the corpus, indicating that some commenters stopped prescription medication in connection with dietary changes described in creator content. The supplementary sentiment analysis added nuance to these findings: while weight change and general well-being showed strong positive skew, the digestive, neurological, and hormonal domains showed more negative than positive reports, consistent with known adaptation effects during transitions to carbohydrate restriction [17]. This finding confirms that negative health experiences are present in these communities at a meaningful rate, providing important context for interpreting the positive outcomes identified by the primary framework.

RQ3: Does Positive Outcome Reporting Vary Significantly Across Content Creators, and What Factors May Explain This Variation?

Positive outcome reporting was not uniformly distributed across the eleven channels, with rates ranging from 1.32% to 10.40%, a nearly ninefold difference that suggests a structural rather than a random pattern. Four channels clustered at substantially higher rates (KenDBerryMD, Jason Fung, Eric Berg DC, and Eric Westman), while the remaining 7 fell below 3%. Because comment volume was balanced across channels, this heterogeneity cannot be attributed to differences in sample size, and the pattern persisted in the external validation. The data suggest that this variation aligns with differences in creator discourse style: channels whose creators adopted accessible, user-facing communication tended to elicit longer average comments and higher testimonial rates, with outcome reports spanning a broad range of health aspects and frequently including narratives of symptom cessation and disease reversal. In contrast, channels with more scientifically oriented discourse produced fewer testimonials overall, but those that did appear were notably longer and more detailed, with higher rates of reports of medication discontinuation and outcomes spanning multiple health dimensions simultaneously. This influence extended to comment depth itself, as false-negative rates were significantly higher among medium-length and long comments than among short ones, suggesting that the more elaborate forms of self-disclosure encouraged by certain creator styles also produce more complex expressions of health outcomes that are harder for rule-based systems to capture. Together, these findings indicate that channel-level characteristics, including content style, creator approach, and community norms, meaningfully shape not only the volume but also the depth and nature of testimonial discourse within healthcasting environments.

RQ4: Can a Precision-Optimized Rule-Based Framework Achieve Sufficient Classification Accuracy for Generating Validated Health Outcome Corpora From User-Generated Content?

The 3-stage rule-based classifier achieved 97.6% precision in the development corpus, and this level held up under external validation across 5 held-out channels, with overlapping confidence intervals, confirming that the framework generalizes beyond the data on which it was built. Estimated recall was 56.2%, reflecting the deliberate design choice to prioritize the validity of every included case over exhaustive detection: the system captures roughly half of all positive outcome reports, but nearly everyone it identifies is correct. This trade-off yielded a validated corpus of health outcome reports at scale without manual review, demonstrating that the framework can generate research-ready datasets from large comment corpora. To assess whether the classification task itself is well defined rather than dependent on a single coder's judgment, 2 independent AI systems (GPT-4o and GPT-4.1) were given the same coding

XSL•FO
RenderX

instructions and reached substantial agreement on core outcome identification, with the strongest reliability emerging when the 2 models were compared directly to each other rather than to the human coder. A head-to-head comparison with fine-tuned deep learning models (BERT and RoBERTa) confirmed that the precision advantage is built into the rule-based design: both models achieved higher recall but lower precision, meaning that for every gain in detection coverage, a substantial number of incorrect classifications would be introduced into the corpus. Error analysis revealed that most missed cases occurred not because the framework lacked relevant health keywords, but because commenters expressed their outcomes in sentence structures the rules did not anticipate. This suggests that expanding the range of recognized expression patterns, rather than adding new terminology, is the most direct route to improved recall. Together, these results establish the framework as a reliable, fully transparent, and reproducible methodology for extracting self-reported health outcomes from unstructured user-generated content at scale, with every classification decision traceable to specific rules.

Implications for Research and Practice

Computational infodemiology has increasingly established social media as a valuable source of real-world health data, yet systematic extraction methodologies have focused almost exclusively on pharmacovigilance and adverse-event detection from microblogging platforms [40,55,56], whereas qualitative research on online health communities has examined forum-based and support-group settings [57]. A substantial and growing body of health-related discourse exists in a setting that has received comparatively little methodological attention: the comment sections of expert-led YouTube health channels, where tens of thousands of individuals respond to long-form, creator-led content with first-person accounts of health changes [58,59]. Our results demonstrate that this content layer carries health signal at a density sufficient for systematic computational extraction, establishing healthcasting as a distinct empirical setting for health informatics research. For infodemiology and digital health researchers, this offers a complementary observational channel that captures a population segment (individuals who self-direct dietary interventions informed by credentialed online content) largely invisible to clinical registries and pharmacovigilance systems [29]. The replicable 3-phase construction workflow, hierarchical ontology, and multistudy validation flow presented in this study provide a methodological template that research groups can adapt to adjacent domains, including cardiometabolic disease, chronic pain, and mental health [60,61].

For clinical researchers and public health practitioners, understanding the health changes patients experience and report outside clinical settings is critical for designing patient-reported outcome measures, identifying underrecognized treatment effects, and monitoring population-level engagement with dietary interventions [62,63]. Our analysis reveals a self-reported outcome landscape that extends beyond the weight-loss framing commonly associated with carbohydrate restriction [3,5]. Reductions in pain and inflammation, improvements in type 2 diabetes, skin health, and psychological well-being were prominently reported, consistent with the clinical trial literature documenting multi-system effects of metabolic interventions [1,64,65]. Reports covered eighteen named disease conditions, and over half addressed multiple ROs simultaneously, suggesting that commenters experience and report systemic rather than isolated health changes. These convergences between self-reported online data and published clinical evidence identify patient-reported outcome dimensions that merit prospective investigation using designs appropriate to each dimension [62]. Perhaps most important for clinical practice, medication-discontinuation reports were present in this corpus, raising questions about patients stopping prescription therapy influenced by creator-led online content without direct medical oversight. This finding underscores the need for structured dialogue between healthcasting communities and the clinical care system to ensure that patient-initiated medication changes occur safely [63,66,67].

As digital platforms become the primary channels through which patients encounter health information and make health decisions, understanding how platform features and community dynamics shape health discourse is an increasingly important priority for health communication researchers and platform designers [20,40]. Our analysis reveals that positive outcome reporting rates vary substantially across channels operating within the same broad dietary domain, and that this variation is structurally predictable from content style and community culture rather than random. We offer interpretive hypotheses, framed as requiring systematic testing rather than as validated findings [57]: channels whose creators explicitly invite health testimonials and attract audiences with active metabolic conditions foster communities where outcome-sharing functions as a social norm, whereas science-focused channels attract audiences that engage with scientific explanations rather than personal testimony. Prior research has established that content style and audience composition shape user behavior on social platforms [40,68], and our results extend this principle to the specific domain of creator-led metabolic health content. This pattern is further supported by findings that channels with longer average comments had significantly higher positive-outcome rates (ρ=0.645; P=.03) and that false-negative rates increased substantially with comment length. These results indicate that the richer, more detailed forms of self-disclosure fostered by certain creator styles also generate more complex health narratives that automated extraction systems are less likely to capture fully. For researchers designing computational health discourse tools, this suggests that extraction frameworks must be calibrated not only to the health domain but also to the discourse norms of the specific communities being studied. For platform designers, these patterns suggest actionable interventions: features that surface testimonial density, link outcome claims to their evidentiary basis, or guide viewers toward content that matches their information needs could meaningfully reshape how patients encounter health testimony online [20]. The framework and ontology presented here provide the measurement infrastructure needed to evaluate such interventions at scale.

In the health natural language processing literature, classification systems have been optimized primarily for balanced performance, yielding precision levels insufficient for

XSL•FO
RenderX

high-confidence corpus generation, in which every included observation must be valid [43,53]. Our precision-first rule-based architecture addresses this gap, exceeding the precision reported in comparable systems (Table 13) [43]. External validation confirms that this advantage extends beyond the development corpus [69]. The transformer baseline comparison, in which fine-tuned models achieved higher recall but lower precision, confirms that this advantage is architectural rather than data-dependent [42]. For the broader health natural language processing community, this finding supports a practical design principle: when the RO is corpus generation rather than individual case detection, rule-based architectures with domain-specific ontologies offer a precision advantage that current deep learning approaches do not match [43]. The construction and validation workflows are domain-agnostic and transferable to other chronic-disease communities and health-adjacent digital platforms, subject to ontology respecification and revalidation [60].

Limitations and Future Directions

This study has limitations that define its scope and point to productive extensions. Commenters are a self-selected subset of viewers, and those experiencing positive outcomes may be more likely to comment, inflating the observed rate relative to the underlying population [29]. The sample selection strategy prioritized the 10 most-commented videos per channel, further biasing the sample toward viral content [68], and reply threads were excluded due to API constraints [70]. Manual validation was performed by a single domain-expert coder, with interrater reliability assessed through LLM-assisted annotation rather than independent human domain-expert validation [45,71]. The observed association between creator discourse style and testimonial characteristics is correlational and based on eleven channels, limiting causal inference and the generalizability of channel-level patterns. Additionally, the significantly higher false-negative rate among longer comments means that the framework may systematically under-capture outcomes in communities whose discourse norms encourage more detailed self-reporting, compounding the channel-level variation described above. Addressing these constraints through multicoder validation at scale and incorporating reply threads is a priority for future research.

The framework was developed and validated within a single dietary domain on a single platform. The ontology was engineered for TCR, and adapting it to other health domains requires modifying the ontology and revalidating it rather than direct transfer [60]. Results may not generalize to platforms with different demographics, moderation norms, or content formats [40,72]. While external validation on held-out channels confirmed precision transfer, 3 of 5 external channels yielded fewer than fifteen classifier-positive comments, precluding reliable per-channel estimates. Extending the framework to adjacent chronic-disease communities on YouTube and to other health-adjacent platforms represents a natural next step [61].

The nature of self-reported content imposes additional constraints. Outcomes cannot be independently verified, and users may misattribute improvements or conflate correlation with causation; the classification only confirms that users report these outcomes, not that the underlying health claims are accurate [73]. The framework extracts only positive outcomes, an asymmetry only partially mitigated by the supplementary ABSA analysis. Comments are point-in-time reports that preclude assessment of long-term sustainability [26]. Developing a dedicated negative-outcome extraction framework and expanding pattern libraries to close the recall gap are the most immediate methodological extensions, enabling a more complete characterization of health discourse in creator-led digital communities.

Conclusions

This study establishes that creator-led metabolic-health YouTube content is a scalable, computationally viable source of self-reported health outcomes and presents a replicable, precision-first methodology for extracting and validating these outcomes at the corpus scale. Beyond the methodology, the breadth of reported outcomes, the influence of creator discourse style on the volume and nature of testimonial reporting, and the medication-discontinuation signal collectively position healthcasting as a phenomenon warranting sustained attention from the health informatics, health communication, and clinical research communities. As this attention grows, the precision-first architecture and hierarchical ontology provide a transferable methodological foundation for computational health discourse analysis across chronic-disease domains and digital platforms.

Acknowledgments

Large language models (GPT-4o and GPT-4.1, OpenAI) were used as independent coders to assess interrater reliability for the manual validation sample, as described in the Interrater Reliability Assessment section. Pre-trained transformer models (BERT-base-uncased and RoBERTa-base, Hugging Face) were fine-tuned on the study dataset for the baseline comparison analysis described in the Transformer Baseline Comparison section. AI-assisted tools (Claude and Anthropic) were used for technical manuscript preparation tasks, including citation format conversion, document formatting compliance checks, and reference list verification. No AI tools were used to interpret results or draft intellectual content. All outputs were reviewed and verified by the authors, who take full responsibility for the content of this manuscript.

Data Availability

The datasets generated or analyzed during this study are available in the GitHub repository. Additional data are included in this published article and its supplementary information files.

XSL•FO
RenderX

Funding

This work was supported by Fundação para a Ciência e a Tecnologia IP (project UID/00667 Unidade de Investigação e Desenvolvimento em Engenharia Mecânica e Industrial). The funder had no involvement in study design, data collection, analysis, interpretation of data, or the writing of this manuscript.

Authors' Contributions

RR conceptualized the study, designed the methodology, developed the ontology and classification framework, collected and analyzed the data, performed the validation, and wrote the manuscript. AZ supervised the research, provided critical review, and contributed to the interpretation of findings.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CREMLS reporting checklist.
[DOCX File , 26 KB - jmir_v28i1e94855_app1.docx]

Multimedia Appendix 2
Summary: videos and comments by channel.
[XLSX File (Microsoft Excel File), 25 KB - jmir_v28i1e94855_app2.xlsx]

Multimedia Appendix 3
Ontology structure with representative keyword and exclusion patterns.
[DOCX File , 20 KB - jmir_v28i1e94855_app3.docx]

Multimedia Appendix 4
Classification rule patterns.
[DOCX File , 20 KB - jmir_v28i1e94855_app4.docx]

Multimedia Appendix 5
External validation protocol and results.
[DOCX File , 19 KB - jmir_v28i1e94855_app5.docx]

Multimedia Appendix 6
Large language model–assisted annotation prompt.
[DOCX File , 20 KB - jmir_v28i1e94855_app6.docx]

Multimedia Appendix 7
Transformer baseline comparison.
[DOCX File , 198 KB - jmir_v28i1e94855_app7.docx]

Multimedia Appendix 8
Aspect-based sentiment analysis prompt design and few-shot examples.
[DOCX File , 20 KB - jmir_v28i1e94855_app8.docx]

Multimedia Appendix 9
Large language model annotation bias audit.
[DOCX File , 22 KB - jmir_v28i1e94855_app9.docx]

References

1. Li M, Chi X, Wang Y, Setrerrahmane S, Xie W, Xu H. Trends in insulin resistance: insights into mechanisms and therapeutic strategy. Signal Transduct Target Ther 2022;7(1):216 [FREE Full text] [doi: 10.1038/s41392-022-01073-0] [Medline: 35794109]
2. Volek JS, Phinney SD, Forsythe CE, Quann EE, Wood RJ, Puglisi MJ, et al. Carbohydrate restriction has a more favorable impact on the metabolic syndrome than a low fat diet. Lipids 2009 Apr;44(4):297-309. [doi: 10.1007/s11745-008-3274-2] [Medline: 19082851]

XSL•FO
RenderX

3. Noakes T, Murphy T, Wellington N. Ketogenic: The Science of Therapeutic Carbohydrate Restriction in Human Health. San Diego, CA: Academic Press; Jun 22, 2023.
4. Volek JS, Phinney SD. The Art and Science of Low Carbohydrate Living: An Expert Guide to Making the Life-Saving Benefits of Carbohydrate Restriction Sustainable and Enjoyable. Frederick: Beyond Obesity; May 19, 2011.
5. Westman EC, Feinman RD, Mavropoulos JC, Vernon MC, Volek JS, Wortman JA, et al. Low-carbohydrate nutrition and metabolism. Am J Clin Nutr 2007 Aug;86(2):276-284 [FREE Full text] [doi: 10.1093/ajcn/86.2.276] [Medline: 17684196]
6. Ludwig DS, Apovian CM, Aronne LJ, Astrup A, Cantley LC, Ebbeling CB, et al. Competing paradigms of obesity pathogenesis: energy balance versus carbohydrate-insulin models. Eur J Clin Nutr 2022 Sep;76(9):1209-1221 [FREE Full text] [doi: 10.1038/s41430-022-01179-2] [Medline: 35896818]
7. Westman EC. Editorial: Carbohydrate-restricted nutrition and diabetes mellitus. Front Nutr 2021;8:827990 [FREE Full text] [doi: 10.3389/fnut.2021.827990] [Medline: 35127799]
8. Westman EC, Tondt J, Maguire E, Yancy WS. Implementing a low-carbohydrate, ketogenic diet to manage type 2 diabetes mellitus. Expert Rev Endocrinol Metab 2018 Sep;13(5):263-272. [doi: 10.1080/17446651.2018.1523713] [Medline: 30289048]
9. Zheng Q, Gao X, Ruan X, Chen S, Pan X, Wang R, et al. Are low carbohydrate diet interventions beneficial for metabolic syndrome and its components? A systematic review and meta-analysis of randomized controlled trials. Int J Obes (Lond) 2025 Jul;49(7):1252-1263. [doi: 10.1038/s41366-025-01822-5] [Medline: 40579564]
10. Chamma L, Chamma N, Mattar L, Slaybe M, Haidar S, Rizk R. Beyond epilepsy management: a narrative review of the health effects of ketogenic diets. Nutrition 2025 Sep;137:112804. [doi: 10.1016/j.nut.2025.112804] [Medline: 40403346]
11. Ahire C, Yadav R, Bhamare UU, Kaur G, Palkar MB. From refractory epilepsy to neurodegeneration: emerging mechanistic and clinical insights into the ketogenic diet. FASEB J 2026 Mar 31;40(6):e71609. [doi: 10.1096/fj.202503317R] [Medline: 41846418]
12. Ede G. Change Your Diet, Change Your Mind: A Powerful Plan to Improve Mood, Overcome Anxiety, and Protect Memory for a Lifetime of Optimal Mental Health. New York City: Balance; Jan 30, 2024:384.
13. Noakes TD, Windt J. Evidence that supports the prescription of low-carbohydrate high-fat diets: a narrative review. Br J Sports Med 2017 Jan;51(2):133-139. [doi: 10.1136/bjsports-2016-096491] [Medline: 28053201]
14. Sethi S, Wakeham D, Ketter T, Hooshmand F, Bjornstad J, Richards B, et al. Ketogenic Diet Intervention on Metabolic and Psychiatric Health in Bipolar and Schizophrenia: A Pilot Trial. Psychiatry Res 2024 May;335:115866 [FREE Full text] [doi: 10.1016/j.psychres.2024.115866] [Medline: 38547601]
15. Danan A, Westman EC, Saslow LR, Ede G. The ketogenic diet for refractory mental illness: a retrospective analysis of 31 inpatients. Front Psychiatry 2022;13:951376 [FREE Full text] [doi: 10.3389/fpsyt.2022.951376] [Medline: 35873236]
16. Sarnyai Z, Kraeuter AK, Palmer CM. Ketogenic diet for schizophrenia: clinical implication. Curr Opin Psychiatry 2019 Sep;32(5):394-401. [doi: 10.1097/YCO.0000000000000535] [Medline: 31192814]
17. Dyńka D, Rodzeń Ł, Rodzeń M, Łojko D, Karakuła-Juchnowicz H, Ede G, et al. The ketogenic diet is not for everyone: contraindications, side effects, and drug interactions. Ann Med 2026 Dec;58(1):2603016 [FREE Full text] [doi: 10.1080/07853890.2025.2603016] [Medline: 41486865]
18. Kamiński M, Skonieczna-Żydecka K, Nowak JK, Stachowska E. Global and local diet popularity rankings, their secular trends, and seasonal variation in Google Trends data. Nutrition 2020;79-80:110759. [doi: 10.1016/j.nut.2020.110759] [Medline: 32563767]
19. Lennerz BS, Mey JT, Henn OH, Ludwig DS. Behavioral characteristics and self-reported health status among 2029 adults consuming a "Carnivore Diet". Curr Dev Nutr 2021 Dec;5(12):nzab133 [FREE Full text] [doi: 10.1093/cdn/nzab133] [Medline: 34934897]
20. Kaňková J, Binder A, Matthes J. Health-related communication of social media influencers: a scoping review. Health Commun 2025 Jun;40(7):1300-1313 [FREE Full text] [doi: 10.1080/10410236.2024.2397268] [Medline: 39258728]
21. Madathil KC, Rivera-Rodriguez AJ, Greenstein JS, Gramopadhye AK. Healthcare information on youTube: a systematic review. Health Informatics J 2015 Sep;21(3):173-194 [FREE Full text] [doi: 10.1177/1460458213512220] [Medline: 24670899]
22. Chou WYS, Oh A, Klein WMP. Addressing health-related misinformation on social media. JAMA 2018 Dec 18;320(23):2417-2418. [doi: 10.1001/jama.2018.16865] [Medline: 30428002]
23. Schillinger D, Chittamuru D, Ramírez AS. From "Infodemics" to health promotion: a novel framework for the role of social media in public health. Am J Public Health 2020 Sep;110(9):1393-1396. [doi: 10.2105/AJPH.2020.305746] [Medline: 32552021]
24. Domingo-Salvany A. The science of real-time data capture: self-reports in health research. J Epidemiol Community Health 2008 May 01;62(5):471.1-47471. [doi: 10.1136/jech.2007.068551] [Medline: 18413463]
25. Elwert F, Winship C. Endogenous selection bias: the problem of conditioning on a collider variable. Annu Rev Sociol 2014 Jul;40:31-53 [FREE Full text] [doi: 10.1146/annurev-soc-071913-043455] [Medline: 30111904]
26. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. J Multidiscip Healthc 2016;9:211-217 [FREE Full text] [doi: 10.2147/JMDH.S104807] [Medline: 27217764]

XSL•FO
RenderX

27. Sinnenberg L, Buttenheim AM, Padrez K, Mancheno C, Ungar L, Merchant RM. Twitter as a tool for health research: a systematic review. Am J Public Health 2017 Jan;107(1):e1-e8. [doi: 10.2105/AJPH.2016.303512] [Medline: 27854532]
28. Charles-Smith LE, Reynolds TL, Cameron MA, Conway M, Lau EHY, Olsen JM, et al. Using Social Media for Actionable Disease Surveillance and Outbreak Management: A Systematic Literature Review. PLoS One 2015;10(10):e0139701 [FREE Full text] [doi: 10.1371/journal.pone.0139701] [Medline: 26437454]
29. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85 [FREE Full text] [doi: 10.2196/jmir.1933] [Medline: 23615206]
30. Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inform Assoc 2015 May;22(3):671-681 [FREE Full text] [doi: 10.1093/jamia/ocu041] [Medline: 25755127]
31. Paul M, Dredze M. You are what you tweet: analyzing Twitter for public health. ICWSM 2021 Aug 03;5(1):265-272 [FREE Full text] [doi: 10.1609/icwsm.v5i1.14137]
32. Sloane R, Osanlou O, Lewis D, Bollegala D, Maskell S, Pirmohamed M. Social media and pharmacovigilance: a review of the opportunities and challenges. Br J Clin Pharmacol 2015 Oct;80(4):910-920 [FREE Full text] [doi: 10.1111/bcp.12717] [Medline: 26147850]
33. Golder S, O'Connor K, Wang Y, Klein A, Gonzalez Hernandez G. The value of social media analysis for adverse events detection and pharmacovigilance: scoping review. JMIR Public Health Surveill 2024 Sep 06;10:e59167 [FREE Full text] [doi: 10.2196/59167] [Medline: 39240684]
34. XU D, Gonzalez-Hernandez G. Proceedings of the 9th social media mining for health research and applications (SMM4H 2024) workshop and shared tasks. Bangkok: Association for Computational Linguistics; 2024 Aug 15 Presented at: Proceedings of the 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop; August 15, 2024; Bangkok, Thailand p. A URL: https://aclanthology.org/2024.smm4h-1.0/
35. De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M. Discovering shifts to suicidal ideation from mental health content in social media. In: Proc SIGCHI Conf Hum Factor Comput Syst. 2016 May Presented at: CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems; May 7-12, 2016; San Jose California USA p. 2098-2110 URL: https://europepmc.org/abstract/MED/29082385 [doi: 10.1145/2858036.2858207]
36. Keelan J, Pavri-Garcia V, Tomlinson G, Wilson K. YouTube as a source of information on immunization: a content analysis. JAMA 2007 Dec 05;298(21):2482-2484. [doi: 10.1001/jama.298.21.2482] [Medline: 18056901]
37. Mohamed F, Shoufan A. Users' experience with health-related content on YouTube: an exploratory study. BMC Public Health 2024 Jan 03;24(1):86 [FREE Full text] [doi: 10.1186/s12889-023-17585-5] [Medline: 38172765]
38. Teng S, Khong KW, Pahlevan Sharif S, Ahmed A. YouTube video comments on healthy eating: descriptive and predictive analysis. JMIR Public Health Surveill 2020 Oct 01;6(4):e19618 [FREE Full text] [doi: 10.2196/19618] [Medline: 33001036]
39. Cao Y, Dai J, Wang Z, Zhang Y, Shen X, Liu Y, et al. Machine learning approaches for depression detection on social media: a systematic review of biases and methodological challenges. J Behav Data 2025;5(1):67-102 [FREE Full text] [doi: 10.35566/jbds/caoyc]
40. Paul M, Dredze M. Social Monitoring for Public Health. In: Synthesis Lectures on Information Concepts, Retrieval, and Services. Cham, Switzerland: Springer; Aug 31, 2017:1-183.
41. Coppersmith G, Leary R, Crutchley P, Fine A. Natural language processing of social media as screening for suicide risk. Biomed Inform Insights 2018;10:1178222618792860 [FREE Full text] [doi: 10.1177/1178222618792860] [Medline: 30158822]
42. Chiticariu L, Li Y, Reiss F. Rule-based information extraction is dead! Long live rule-based information extraction systems. : Association for Computational Linguistics; 2013 Oct Presented at: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing; May 04, 2026; Seattle, Washington, USA p. 827-832 URL: https://aclanthology.org/D13-1079/ [doi: 10.18653/v1/d13-1079]
43. Golder S, Norman G, Loke YK. Systematic review on the prevalence, frequency and comparative value of adverse events data in social media. Br J Clin Pharmacol 2015 Oct;80(4):878-888 [FREE Full text] [doi: 10.1111/bcp.12746] [Medline: 26271492]
44. Wilson EB. Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association 1927 Jun;22(158):209-212. [doi: 10.1080/01621459.1927.10502953]
45. Gilardi F, Alizadeh M, Kubli M. ChatGPT outperforms crowd workers for text-annotation tasks. Proc Natl Acad Sci U S A 2023 Jul 25;120(30):e2305016120 [FREE Full text] [doi: 10.1073/pnas.2305016120] [Medline: 37463210]
46. Törnberg P. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. Social Science Computer Review 2025 Dec;43(6):1181-1195 [FREE Full text] [doi: 10.1177/08944393241286471]
47. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159. [doi: 10.2307/2529310]
48. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990;43(6):551-558. [doi: 10.1016/0895-4356(90)90159-m] [Medline: 2189948]

XSL•FO
RenderX

49. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990;43(6):543-549. [doi: 10.1016/0895-4356(90)90158-l] [Medline: 2348207]
50. Liu B. Synthesis lectures on human language technologies. In: Sentiment Analysis and Opinion Mining. Cham, Switzerland: Springer; May 23, 2012:1-167.
51. Pontiki M, Galanis D, Papageorgiou H. SemEval-2016 Task 5: aspect based sentiment analysis. : Association for Computational Linguistics; 2016 Jan 01 Presented at: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016); June 16-17, 2016; San Diego, California p. 19-30 URL: https://aclanthology.org/S16-1002/ [doi: 10.18653/v1/S16-1002]
52. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993 May;46(5):423-429. [doi: 10.1016/0895-4356(93)90018-v] [Medline: 8501467]
53. Gonzalez-Hernandez G, Sarker A, O'Connor K, Savova G. Capturing the patient's perspective: a review of advances in natural language processing of health-related text. Yearb Med Inform 2017 Aug;26(1):214-227 [FREE Full text] [doi: 10.15265/IY-2017-029] [Medline: 29063568]
54. Magge A, Tutubalina E, Miftahutdinov Z, Alimova I, Dirkson A, Verberne S, et al. DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. J Am Med Inform Assoc 2021 Sep 18;28(10):2184-2192 [FREE Full text] [doi: 10.1093/jamia/ocab114] [Medline: 34270701]
55. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. J Med Internet Res 2009 Mar 27;11(1):e11 [FREE Full text] [doi: 10.2196/jmir.1157] [Medline: 19329408]
56. Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, et al. Utilizing social media data for pharmacovigilance: A review. J Biomed Inform 2015 Apr;54:202-212 [FREE Full text] [doi: 10.1016/j.jbi.2015.02.004] [Medline: 25720841]
57. Maloney-Krichmar D, Preece J. A multilevel analysis of sociability, usability, and community dynamics in an online health community. ACM Trans Comput-Hum Interact 2005 Jun 01;12(2):201-232 [FREE Full text] [doi: 10.1145/1067860.1067864]
58. Madathil KC, Rivera-Rodriguez AJ, Greenstein JS, Gramopadhye AK. Healthcare information on youTube: a systematic review. Health Informatics J 2015 Sep;21(3):173-194 [FREE Full text] [doi: 10.1177/1460458213512220] [Medline: 24670899]
59. Eysenbach G. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. Am J Prev Med 2011 May;40(5 Suppl 2):S154-S158. [doi: 10.1016/j.amepre.2011.02.006] [Medline: 21521589]
60. Daumé IIIH. Frustratingly easy domain adaptation. 2007 Presented at: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics; June 23-30, 2007; Prague, Czech Republic p. 256-263 URL: https://aclanthology.org/P07-1033/
61. Noy NF, McGuinness DL. Ontology development 101: a guide to creating your first ontology. In: Stanford Knowledge Systems Laboratory Technical Report. Stanford, California: Stanford University; Jan 01, 2001.
62. Staudacher HM, Yao CK, Chey WD, Whelan K. Optimal design of clinical trials of dietary interventions in disorders of gut-brain interaction. Am J Gastroenterol 2022 Jun 01;117(6):973-984 [FREE Full text] [doi: 10.14309/ajg.0000000000001732] [Medline: 35297784]
63. Jim HSL, Hoogland AI, Brownstein NC, Barata A, Dicker AP, Knoop H, et al. Innovations in research and clinical care using patient-generated health data. CA Cancer J Clin 2020 May;70(3):182-199 [FREE Full text] [doi: 10.3322/caac.21608] [Medline: 32311776]
64. Dyńka D, Kowalcze K, Charuta A, Paziewska A. The ketogenic diet and cardiovascular diseases. Nutrients 2023 Jul 28;15(15):3368 [FREE Full text] [doi: 10.3390/nu15153368] [Medline: 37571305]
65. Hallberg SJ, McKenzie AL, Williams PT, Bhanpuri NH, Peters AL, Campbell WW, et al. Effectiveness and Safety of a Novel Care Model for the Management of Type 2 Diabetes at 1 Year: An Open-Label, Non-Randomized, Controlled Study. Diabetes Ther 2018 Apr;9(2):583-612 [FREE Full text] [doi: 10.1007/s13300-018-0373-9] [Medline: 29417495]
66. Walsh CA, Cahir C, Tecklenborg S, Byrne C, Culbertson MA, Bennett KE. The association between medication non-adherence and adverse health outcomes in ageing populations: a systematic review and meta-analysis. Br J Clin Pharmacol 2019 Nov;85(11):2464-2478 [FREE Full text] [doi: 10.1111/bcp.14075] [Medline: 31486099]
67. Reading MJ, Merrill JA. Converging and diverging needs between patients and providers who are collecting and using patient-generated health data: an integrative review. J Am Med Inform Assoc 2018 Jun 01;25(6):759-771 [FREE Full text] [doi: 10.1093/jamia/ocy006] [Medline: 29471330]
68. Bakshy E, Messing S, Adamic LA. Political science. Exposure to ideologically diverse news and opinion on Facebook. Science 2015 Jun 05;348(6239):1130-1132. [doi: 10.1126/science.aaa1160] [Medline: 25953820]
69. Ding Z, Wang Z, Zhang Y, Cao Y, Liu Y, Shen X, et al. Trade-offs between machine learning and deep learning for mental illness detection on social media. Sci Rep 2025 Apr 25;15(1):14497 [FREE Full text] [doi: 10.1038/s41598-025-99167-6] [Medline: 40281061]
70. Thelwall M, Sud P, Vis F. Commenting on YouTube videos: from Guatemalan rock to El Big Bang. J Am Soc Inf Sci 2011 Nov 14;63(3):616-629. [doi: 10.1002/asi.21679]
71. Xu S, Wang Z, Wang H, Ding Z, Zou Y, Cao Y. LLM-based annotation and token-augmented modeling for emotional tone classification in online cancer peer-support posts. medRxiv 2026 Jan 30:1-15. [doi: 10.64898/2026.01.27.26344999]

XSL•FO
RenderX

72. Giglietto F, Rossi L, Bennato D. The open laboratory: limits and possibilities of using Facebook, Twitter, and YouTube as a research data source. J Technol Hum Serv 2012 Dec 06;30(3-4):145-159. [doi: 10.1080/15228835.2012.743797]
73. Kington RS, Arnesen S, Chou WS, Curry SJ, Lazer D, Villarruel AM. Identifying credible sources of health information in social media: principles and attributes. NAM Perspect 2021;2021 [FREE Full text] [doi: 10.31478/202107a] [Medline: 34611600]

Abbreviations

ABSA: aspect-based sentiment analysis
API: application programming interface
BERT: Bidirectional Encoder Representations from Transformers
CREMLS: Consolidated Reporting Guidelines for Prognostic and Diagnostic Machine Learning Modeling Studies
LLM: large language model
RO: research objective
RoBERTa: Robustly Optimized BERT Pretraining Approach
RQ: research question
TCR: therapeutic carbohydrate restriction

Edited by S Brini; submitted 07.Mar.2026; peer-reviewed by Y Cao, L Wreyford; comments to author 27.Mar.2026; revised version received 24.Apr.2026; accepted 01.May.2026; published 26.May.2026.

Please cite as:
Ribeiro R, Zutshi A
Self-Reported Health Outcomes in Metabolic Health YouTube Comments: Cross-Sectional Study and Rule-Based Natural Language Processing Framework Development and Validation
J Med Internet Res 2026;28:e94855
URL: https://www.jmir.org/2026/1/e94855
doi:10.2196/94855
PMID:

©Ricardo Ribeiro, Aneesh Zutshi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

User-Centered Diabetes Self-Management App (DiabAid Nexus) in Sub-Saharan Africa: Development and Usability Study

Genet Tadese Aboye[1,2], BSc, MSc, PGD, PhD; Ermias Habte Gebremichael[3], MD; Gizeaddis Lamesgin Simegn[2,4], BSc, MSc, PhD; Jean-Marie Aerts[1], BSc, MSc, PhD

[1]Department of Biosystems, Division of Animal and Human Health Engineering, M3-BIORES (Measure, Model & Manage Bioreponses), KU Leuven, Kasteelpark Arenberg 30, Leuven, Belgium

[2]School of Biomedical Engineering, Jimma University, Jimma, Ethiopia

[3]Department of Internal Medicine, Jimma University, Jimma, Ethiopia

[4]The Russel H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University, Baltimore, MD, United States

Corresponding Author:
Genet Tadese Aboye, BSc, MSc, PGD, PhD
Department of Biosystems, Division of Animal and Human Health Engineering, M3-BIORES (Measure, Model & Manage Bioreponses), KU Leuven, Kasteelpark Arenberg 30, Leuven, Belgium

Abstract

Background: Diabetes is a significant global public health concern, with disproportionately high prevalence and poor access to care in low-resource settings, notably in Sub-Saharan Africa.

Objective: By harnessing the ubiquity of mobile phones, we aim to develop and test a user-centered mobile health (mHealth) app tailored to address the specific needs and challenges of diabetes self-management in Sub-Saharan Africa.

Methods: A user-centered design approach was used to develop a bilingual mHealth platform for diabetes management. The app was made using Flutter 3.16.2 (Google) and Dart 3.2.2. (Google) and was tested on various Android emulators and smartphones before publishing. A usability study with 13 participants was carried out to assess the app's real-life impact, quality, and user experience using the user version of the Mobile App Rating Scale (uMARS). A brief semistructured interview was conducted to further evaluate the app's usability and user experience.

Results: The mHealth platform called "*DiabAid Nexus*" (GTA) has been developed and tested with 13 people living with diabetes in Ethiopia. The findings show that the app was perceived as simple, easy to use, and relevant to users' needs. uMARS scores out of 5 showed high ratings for functionality (mean 4.33, SD 0.63), information quality (mean 4.20, SD 0.68), perceived impact (mean 4.20, SD 0.84), and subjective quality (mean 4.10, SD 0.73), indicating that users deemed the app useful, dependable, and valuable for diabetic self-management. The aesthetics subscale likewise obtained a positive score (mean 4.05, SD 0.94), while engagement earned a slightly lower score (mean 3.78, SD 0.95), highlighting the possibility of further boosting interactivity and user motivation. The overall uMARS score of 4.10 suggests a favorable user assessment of the app's quality and efficacy. Interview responses underlined appreciation for localized content and intuitive design, with users indicating happiness and a readiness to use the app as a tool to support their diabetes self-management journey.

Conclusions: The findings suggest that the app is a user-friendly and well-received tool with a high potential for promoting diabetes self-management in low-resource settings and across varied contexts.

(J Med Internet Res 2026;28:e87110) doi:10.2196/87110

KEYWORDS

diabetes; low-resource settings; mHealth; self-management; user-centered design; mobile phone

Introduction

The global burden of noncommunicable diseases continues to rise, with diabetes emerging as one of the most challenging chronic conditions to manage, particularly in low-resource settings (LRSs) [1,2]. Diabetes mellitus is a clinical condition comprising a heterogeneous group of metabolic diseases that are characterized by chronic hyperglycemia and disturbances in carbohydrate, fat, and protein metabolism secondary to defects in insulin secretion, insulin action, or both [3]. Without effective management, it poses significant challenges and can lead to severe complications [4,5].

Over the last 3 decades, the worldwide diabetes burden has increased rapidly. In 2021, approximately 537 million adults had diabetes, with this number expected to rise to 643 million by 2030 and 783 million by 2045 [6]. Notably, this increase has been more pronounced in low- and middle-income countries than in high-income nations, highlighting the growing inequities

XSL•FO
RenderX

in illness distribution and management. Diabetes prevalence among individuals aged 18 years and older in the World Health Organization's African Region has nearly doubled in just over 30 years, from 6.4% in 1990 to 10.5% in 2022 [2,7]. Despite the rapid increase, access to care remains limited. As of 2022, only 26.1% of adults aged 30 years and older diagnosed with diabetes in the African Region received treatment, leaving almost 34 million people without proper medical care [7,8]. Diabetes is increasingly managed across decades, and older adults may face added self-management challenges due to multimorbidity, vision changes, reduced manual dexterity, and heightened vulnerability to severe hypoglycemia. All of these can complicate diabetes management, reinforcing the value of simple, accessible mobile health (mHealth) support across the patients' lifespan [9]. Managing diabetes is particularly difficult in resource-constrained settings, where the health care system is overburdened, patients often face limited access to health services, insufficient diabetes education, and a lack of tools to support effective self-management [10,11].

mHealth platforms are gaining traction for their ability to improve equity in health care delivery across diverse contexts, including low-resource environments. They show considerable promise in managing chronic diseases, such as diabetes [12-15]. A key advantage of mHealth is its capacity to extend health care access to geographically isolated and underserved communities, thereby reducing service inequities. In diabetes care, mHealth interventions have been shown to improve glucose control, medication adherence, and self-care practices [16]. Moreover, a growing body of evidence highlights their potential to enable continuous patient interaction, real-time feedback, and individualized support [17,18]. However, their effectiveness depends heavily on usability, cultural relevance, and alignment with user expectations and digital literacy levels [19]. For mHealth solutions to be embraced and maintained in regular use, they must be perceived as functional and accessible by their intended audience [20,21].

Digital health literacy is another critical determinant of success. Many users in LRSs lack experience with smartphones, reliable internet or data access, or confidence in navigating health apps. A scoping review of mHealth and telemedicine systems for gestational diabetes reported that only about 12% (2/17) of reviewed apps incorporated culturally appropriate design features, and just 25% (3/12) adequately addressed digital health literacy [22]. Such gaps hinder adoption and long-term use. User perceptions also play a central role. Factors such as usability, design aesthetics, interactivity, and motivational features significantly influence engagement [23]. For instance, a qualitative study of adults with type 2 diabetes in Sweden [24] found that participants valued interactive features, varied functions, and user-friendly layouts as essential for long-term use and behavior change. They emphasized the importance of holistic, goal-oriented content and integration of social support. Clinical trials further reinforce the benefits of mHealth. In India, a randomized controlled trial involving 100 patients with type 2 diabetes showed that combining routine care with weekly phone calls and messages significantly improved hemoglobin A1c (HbA_{1c}) levels, medication adherence, diet, and physical activity compared with routine care alone [25]. Similarly, a meta-analysis focusing on older adults found that mHealth apps led to a significant reduction in HbA_{1c}, suggesting effectiveness even in populations often assumed to have lower digital literacy [26].

Importantly, mHealth interventions are not universally beneficial. Their impact often varies according to socioeconomic and sociocultural factors that shape access and outcomes. A recent systematic review noted that few trials of mHealth apps for diabetes and hypertension accounted for differences across sociodemographic groups, such as income, education, gender, or rural versus urban residence [21]. Consequently, mHealth can either help to reduce or inadvertently reinforce health inequities, depending on how interventions are designed and implemented.

These challenges and opportunities highlight the urgent need for innovative, accessible, and scalable solutions to support diabetes care and self-management, particularly in resource-limited settings. At the same time, they point to the importance of empowering individuals to take an active role in managing their health.

Leveraging the widespread availability of mobile phones, this study seeks to design and evaluate a user-centered mHealth platform tailored to the unique challenges of diabetes self-management in Sub-Saharan Africa. The intervention was developed using a user-centered approach guided by the mHealth Agile Development and Evaluation Lifecycle. This paper reports on the design, development, and early evaluation of the platform, emphasizing simplicity, cultural contextualization, and iterative user feedback. By integrating evidence-based design principles with direct input from end users, the study aims to advance the development of effective, context-appropriate digital health interventions.

Methods

Overview

A user-centered mobile app development approach was used, guided by the mHealth Agile Development and Evaluation Lifecycle framework [27]. This framework comprises a project identification stage, followed by four sequential phases of clinical evaluation (1) phase 1: user experience design, development, and alpha testing; (2) phase 2: beta testing; (3) phase 3: clinical trial evaluation; and (4) phase 4: postmarket surveillance. In the context of this study, implementation was carried out up to phase 2. This study is reported in accordance with the mHealth evidence reporting and assessment (mERA) checklist.

Phase 1: User-Centered App Design and Development

Overview

We used the insights from our previous works [28,29], which highlighted the challenges and unmet needs of individuals living with diabetes, as well as the barriers and facilitators to using mHealth approaches in the Sub-Saharan Africa context. These insights were translated into design features, and we subsequently developed *DiabAid Nexus*, an Android-based smartphone app for diabetes self-management.

XSL•FO
RenderX

Case: A Typical Routine of Diabetes Diagnosis, Management, and Coping in the Life of a Person Living With Diabetes in LRSs

A typical routine of diabetes diagnosis, management, and coping in the life of a person living with diabetes in LRSs, identified through observation and patient interview, presented in Figure 1, illustrates systemic challenges in diabetes care in Ethiopia, including delayed diagnosis, limited patient education, reliance on incomplete paper records, rushed consultations, and patients' ongoing confusion and anxiety about self-management. These gaps contribute to poor continuity of care, inadequate glycemic control, and heightened risk of long-term complications.

Figure 1. A typical routine of diabetes diagnosis, management, and coping in the life of a person living with diabetes in low-resource settings.

Proposed App Features

DiabAid Nexus was specifically designed to address these challenges by providing patients with accessible, structured, and user-centered tools for diabetes self-management. The low-fidelity wireframes illustrated in Figure 2 and the flow diagram presented in Figure 3 illustrate the main features, subfeatures, and wireframing of the app, which were derived from the case described above. Building on this design, a comprehensive diabetes self-management mHealth platform available both in English and Amharic, *DiabAid Nexus*, was developed.

Figure 2. Sample low-fidelity wireframes that illustrate the iterative design process of the DiabAid Nexus app.

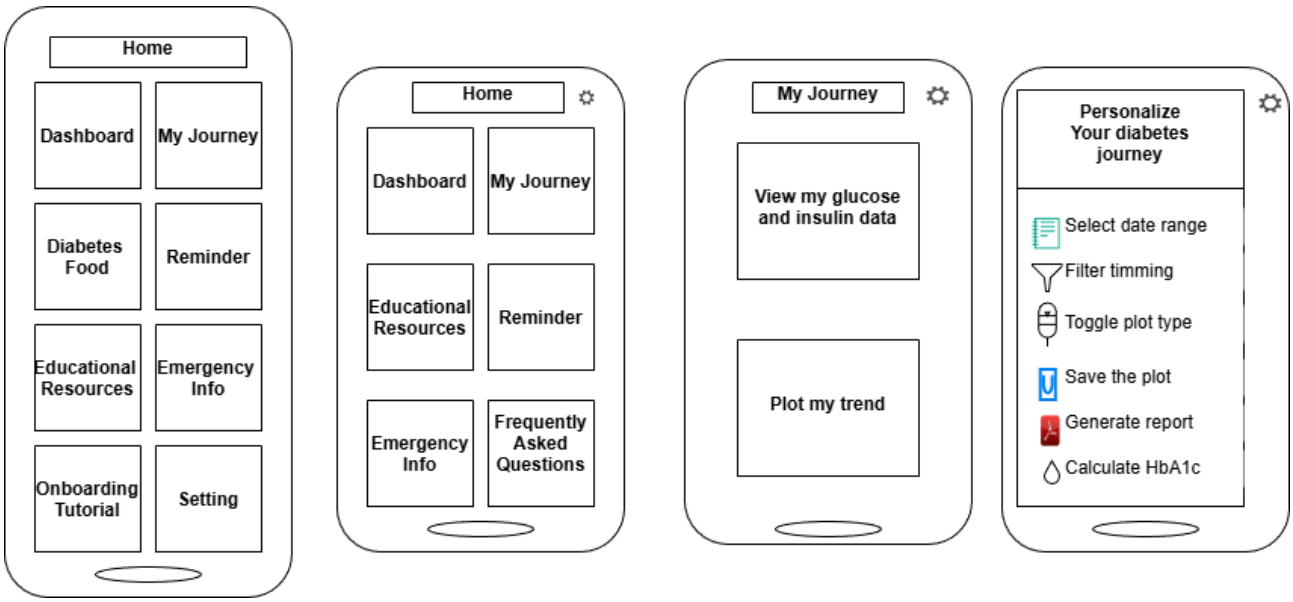

XSL•FO
RenderX

Figure 3. Flow diagram representing the *DiabAid Nexus* app key features.

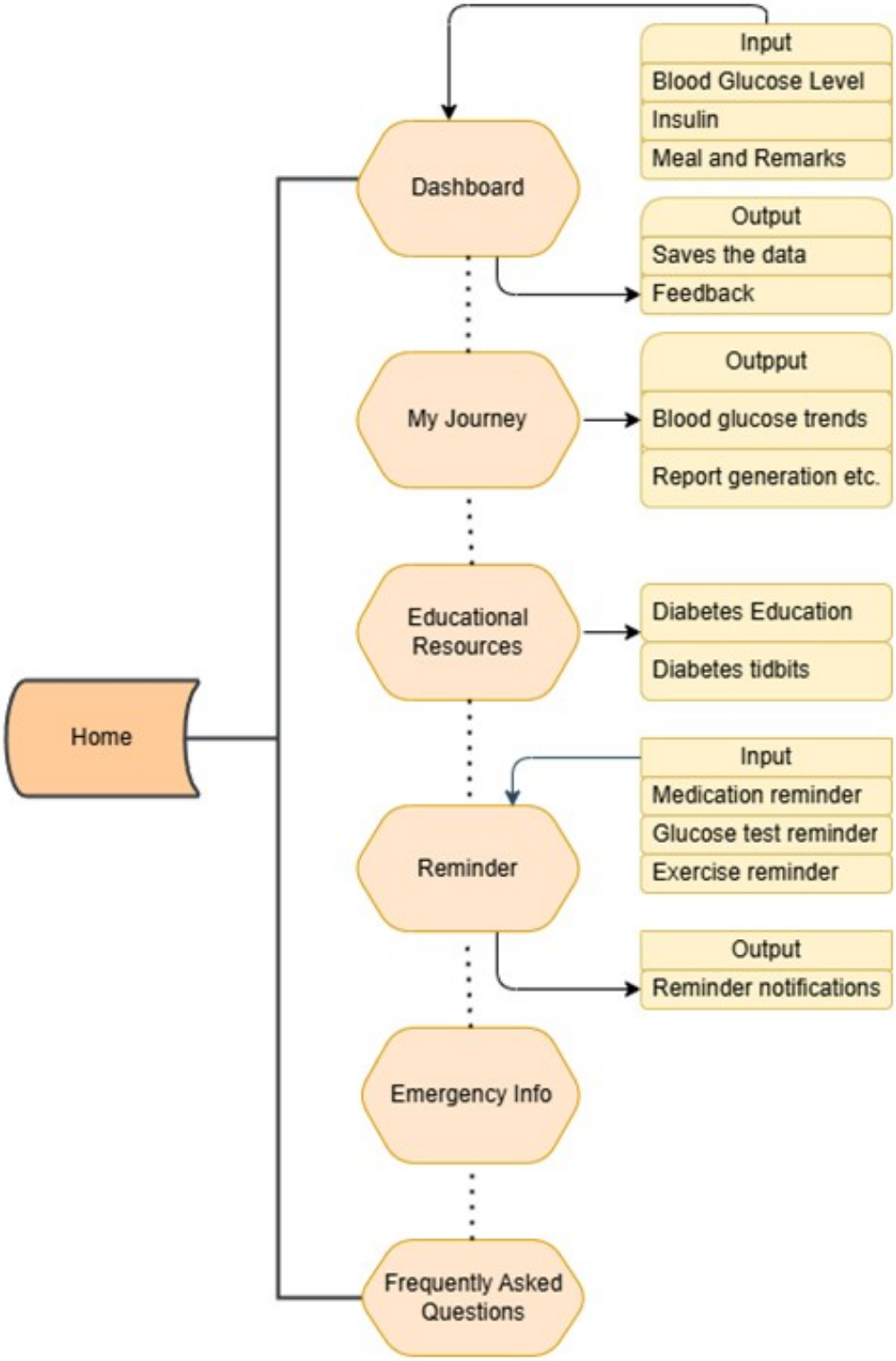

The main features of the app consist of Dashboard, My Journey, Educational Resources, Reminders, Emergency Info, and Frequently Asked Questions.

First, the "Dashboard" replaces the traditional paper-based glucose logbooks, enabling patients to electronically record glucose values, insulin dosages, meal intake, and contextual remarks. This not only ensures more complete and accurate data but also addresses the issue of undocumented insulin adjustments, which are common among syringe users. By centralizing this information, the app reduces patient confusion and provides health care providers with a more reliable basis for clinical decision-making.

Second, the "My Journey" feature empowers patients to visualize and personalize their diabetes data. Unlike paper records that provide only isolated glucose values, this tool generates blood glucose trend analyses, highlighting periods of out-of-range glycemic values, and allowing patients to associate these events with lifestyle factors, such as meals or exercise.

XSL•FO
RenderX

The blood glucose trend graph illustrates date on its x-axis and blood glucose value (in mg/dL or mmol/dL) on its y-axis. The blood glucose trend plot can be customized to an aggregated plot or a detailed plot. The user can also apply filters to select specific blood glucose reading timings, such as fasting, random, after lunch, and so on. Moreover, additional options also exist for estimating the HbA_{1c}. If there are no (or not enough) inputs provided, it will display "No data to show." The ability to generate PDF summaries and HbA_{1c} estimates further supports continuity of care by enabling patients to share structured data with their health care professionals, thus reducing the limitations of rushed consultations and incomplete follow-up information.

Third, the "Educational Resources" directly address the widespread gaps in patient education observed in case 1. By providing clinically validated, comprehensive, and accessible content in both English and Amharic, the app ensures that patients gain a deeper understanding of diabetes, its treatment, and strategies for risk reduction. The contents were initially drafted based on the national and international standard guidelines for Diabetes Self-Management Education [30,31]. Importantly, the content helps alleviate fear and uncertainty by clearly explaining complications and offering guidance on prevention, thereby counteracting the worry, stigma, and misinformation that often undermine self-management.

Fourth, the "Reminder" function supports adherence by prompting patients to test glucose, administer insulin, and exercise at the appropriate times. This mitigates the common problem of missed or delayed medication, which contributes to poor glycemic control. Reminders are delivered via scheduled push notifications. Alerts follow the device's default notification settings, including sound. If the phone is set to silent mode, the reminder will appear as a visual notification without sound, in accordance with the device's operating system behavior.

Fifth, the "Emergency Info" feature allows patients to store vital medical and contact information for use during acute episodes. This addresses the high-risk scenario of presenting to health facilities in critical condition (eg, diabetic ketoacidosis or severe hypoglycemia) by ensuring that health care providers and caregivers have immediate access to relevant information.

Finally, the "Frequently Asked Questions" section complements formal education by answering common questions in a patient-friendly format, helping to resolve confusion that typically persists after brief and incomplete consultations.

Overall, by digitizing record-keeping, enhancing patient education, improving adherence, and facilitating communication between patients and health care professionals, *DiabAid Nexus* directly targets the weak points identified in the observed cycle of diabetes management. The app therefore holds potential not only to reduce patient anxiety and improve self-efficacy but also has the potential to support continuity of self-management between clinical visits.

After the initial wireframing stage, the preliminary app design was presented to 3 individuals living with diabetes who served as patient subject matter informants to obtain early feedback and guidance. Feedback was collected through individual, structured consultation sessions (one-on-one discussions guided by the wireframes and core user flows), focusing on usability, accessibility, and perceived relevance. Suggestions from these sessions were incorporated iteratively to refine the design before proceeding to system development.

System Development

The implementation made use of Flutter (version 3.16.2; Google), which serves as a cross-platform user interface software development toolkit, in combination with Dart (version 3.2.2; Google) [32,33], the primary programming language used to build the app's functionality. Flutter provided the framework for rendering consistent and responsive user interface components across multiple platforms, while Dart enabled the creation of efficient, object-oriented program logic. The development activities were carried out within Android Studio (Google), which functioned as the integrated development environment, offering debugging, emulation, and code management features.

The app supports English and Amharic (widely spoken in Ethiopia). Users select their preferred language in the settings menu, and the choice is saved locally via shared_preferences so it persists across sessions. All user-facing content (eg, titles, labels, snackbars, and educational text) updates automatically when the language changes. The app was developed first in English, then all visible strings were translated into Amharic using Google Translate, followed by manual review. Language switching was implemented using Flutter localization keys with separate .arb resource files for each language, which Flutter compiles into the corresponding generated localization files.

Upon the completion of the iterative development and preliminary testing cycles, the app was prepared for controlled distribution. To facilitate further evaluation under real-world conditions, the app was deployed to the Google Play Store in internal testing mode, thereby allowing designated testers to access and assess its performance before wider public release.

Security and Privacy Measures

The app uses Firebase Authentication (Google) to manage user registration and login [34]. Only a username is collected during registration, thereby minimizing the amount of personal data stored. Login activity, such as the number of logins within a given period of time, is recorded to support usage monitoring and system evaluation. All authentication processes are securely handled by Firebase, which uses encrypted communication (HTTPS) and industry-standard hashing for credential management. By limiting the scope of collected data and relying on Firebase's built-in security mechanisms, the app ensures user privacy while maintaining reliable access control.

Cost and Resource Requirements

DiabAid Nexus was developed in-house by the research team using open-source tools (Flutter and Dart), minimizing software licensing costs. The app was deployed via Google Play internal testing and uses Firebase Authentication. During the study, usage remained within the free-tier limits with no added hosting costs. The app is free for users with no subscription fees and functions largely offline after installation, requiring internet mainly for download, authentication, and updates, which helps

XSL•FO
RenderX

limit data costs. No formal economic evaluation was conducted, but future scale-up may require resources for hosting, maintenance, content updates, and user support.

Phase 2: Usability Testing

User Version of the Mobile Application Rating Scale

A total of 13 eligible individuals (Table 1) participated in the usability evaluation of the mobile app. The participants were people living with diabetes, residing in Jimma and Addis Ababa, Ethiopia. These cities are among those in Ethiopia with broad access to the internet and electricity coverage [35]. Participants were first contacted through a contact person at the Ethiopian Diabetes Association (Addis Ababa) and a diabetes clinic at Jimma University Medical Center. Eligible individuals were informed about the study and invited to participate. Purposive sampling was used to include adults living with diabetes who owned an Android (Google) smartphone and were willing to engage with the app during the study period. Previous exposure to diabetes-related apps was not used as an eligibility criterion. The participants evaluated the usability of the mobile app in terms of engagement, functionality, aesthetics, information, and the app's subjective quality. Participants who consented to take part in the study had their Google account registered, and a link was sent to them to download the app on their mobile phones. Following a brief onboarding session that provided guidance on the app's features and functionality, they were instructed to engage with the app over a minimum period of 8 weeks (from March to May 2025). At the end of this period, participants were invited to complete the user version of the Mobile Application Rating Scale (uMARS) and take part in a brief semistructured interview to further explore their experiences and perspectives. The uMARS is a validated tool with a 26-item measure that includes subscales to assess engagement, functionality, aesthetics, and information quality of the app [36].

Table . Demographic characteristics of participants (N=13) involved in the 8-week usability evaluation of the DiabAid Nexus app conducted in Addis Ababa and Jimma, Ethiopia (March-May 2025).

Characteristics	Participants (N=13), n (%)
Sex	
Male	10 (76.9)
Female	3 (23.1)
Age (y)	
18 - 24	7 (53.8)
25 - 34	5 (38.5)
35 - 44	1 (7.7)
Location	
Addis Ababa	7 (53.8)
Jimma	6 (46.2)
Level of education	
Secondary or high school	1 (7.7)
Technical or vocational training	1 (7.7)
University or college degree	10 (76.9)
Postgraduate degree	1 (7.7)
Employment status	
Employed (full-time or part-time)	6 (46.2)
Self-employed	1 (7.7)
Unemployed	1 (7.7)
Student	4 (30.8)
Other	1 (7.7)

Interview

After testing the app and rating it with the uMARS scale, participants were asked questions to qualitatively learn their overall impression, likes and dislikes, feature usage, impact on self-management, and cultural and contextual fit.

Data Analysis

After a systematic design and iterative enhancement process, the final iteration of the mobile app was launched on the Google Play Store for internal evaluation. Following the conclusion of the usability evaluation session, participants used Google Forms to submit their answers to the uMARS. For preprocessing, the gathered data were exported to Microsoft Excel. Data cleansing was done by standardizing variable names, transforming

XSL•FO
RenderX

categorical replies into numeric codes using the uMARS Likert scale, and guaranteeing completeness and consistency. No incomplete submissions were included. Each participant's responses to the corresponding items were averaged to determine their subscale scores, which included engagement, functionality, aesthetics, information quality, subjective quality, and perceived impact. The core subscale scores were also averaged to determine an overall usability score. Only descriptive statistics were performed due to the exploratory nature and small sample size. Inductive thematic analysis and coding of the semistructured interview were also used to get deeper insights into user feedback during the beta testing period. In total, 2 researchers (GTA and GLS) were involved in the coding collaboratively, and discrepancies were resolved through discussion. A third researcher (JMA) reviewed the finalized themes.

Ethical Considerations

The study protocol was approved by the KU Leuven Research Ethics Committee (G-2024 - 8194) and the Institutional Review Board of Jimma University Institute of Health (JUIH/IRB/311/23). All participants provided written informed consent before participation. To protect participant privacy and confidentiality, all study data were deidentified before analysis, kept in a secure network storage, and handled using appropriate safeguards to prevent reidentification. Participants received compensation of 25 Euros (US $29.25) for their participation.

Results

Design Implementation

These design features were presented and discussed with various people living with diabetes. After a consensus was reached, the app was developed using Flutter and Dart in Android Studio. The app has been implemented and tested on Android (version 14) devices as well as multiple emulators in preparation for the formal usability testing phase. The design and layout of the app's core functionality are illustrated in Figure 4.

XSL•FO
RenderX

Figure 4. Sample screenshots for the DiabAid Nexus app. (A) Login and Registration page, (B) Home page (English), (C) Home page (Amharic), (D) Reminder, (E) Blood glucose logging, (F) Blood glucose trend plot (G) Diabetes Education main page, and (H) Education for reducing risks of acute and chronic complications.

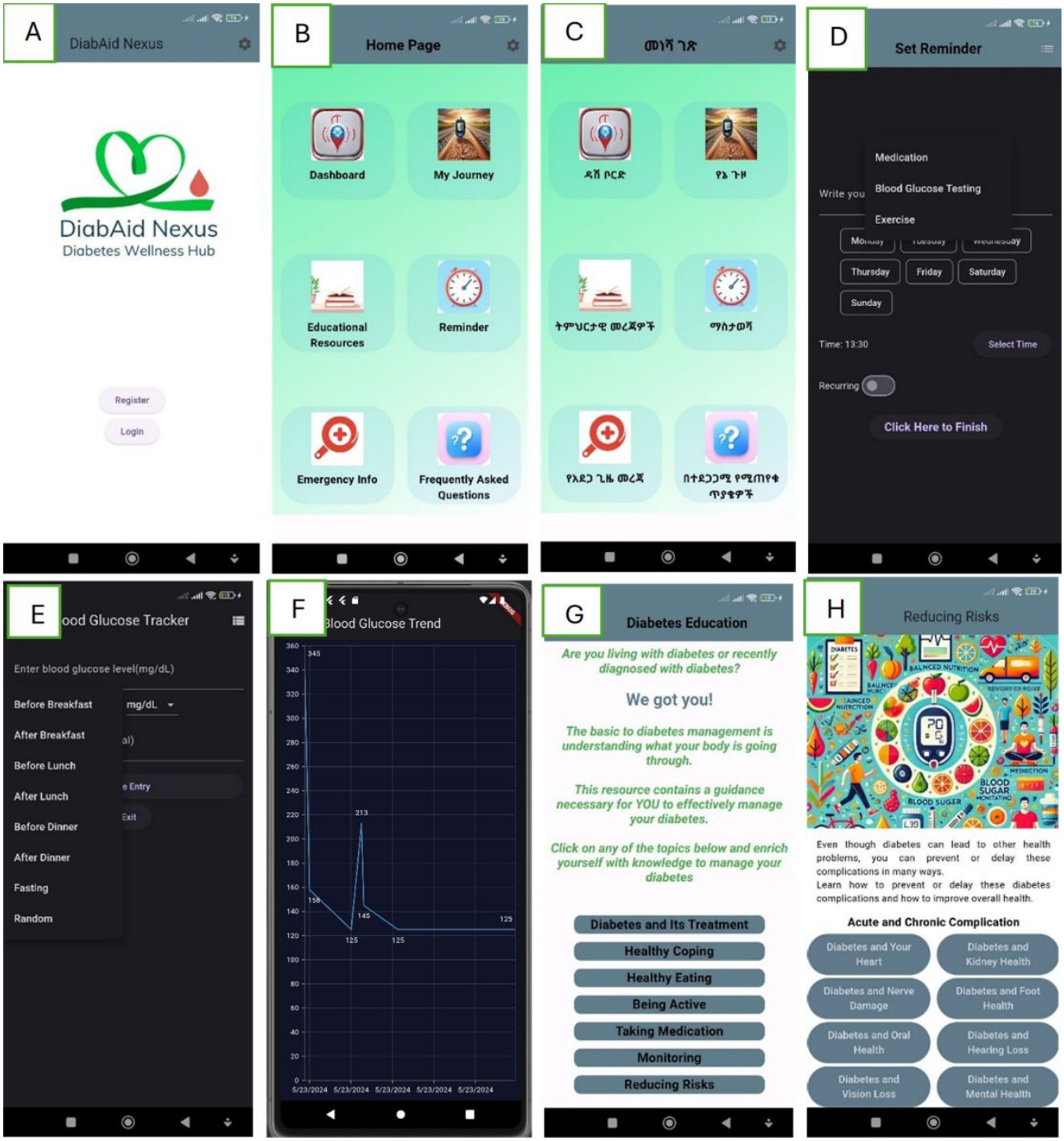

Usability Study or Beta Testing

Study Sample

Table 1 presents the overall characteristics of the participants who participated in the real-life usability test of the app. Recruitment was conducted with the assistance of contact persons from Jimma University Medical Center (Jimma) and the Ethiopian Diabetes Association (Addis Ababa).

XSL•FO
RenderX

Figure 5. .Mean uMARS subscale scores for the DiabAid Nexus app during an 8-week real-world usability assessment in Ethiopia. Scores are based on evaluations from 13 participants and are reported on a 1–5 scale. Error bars indicate SDs. uMARS: user version of the Mobile App Rating Scale.

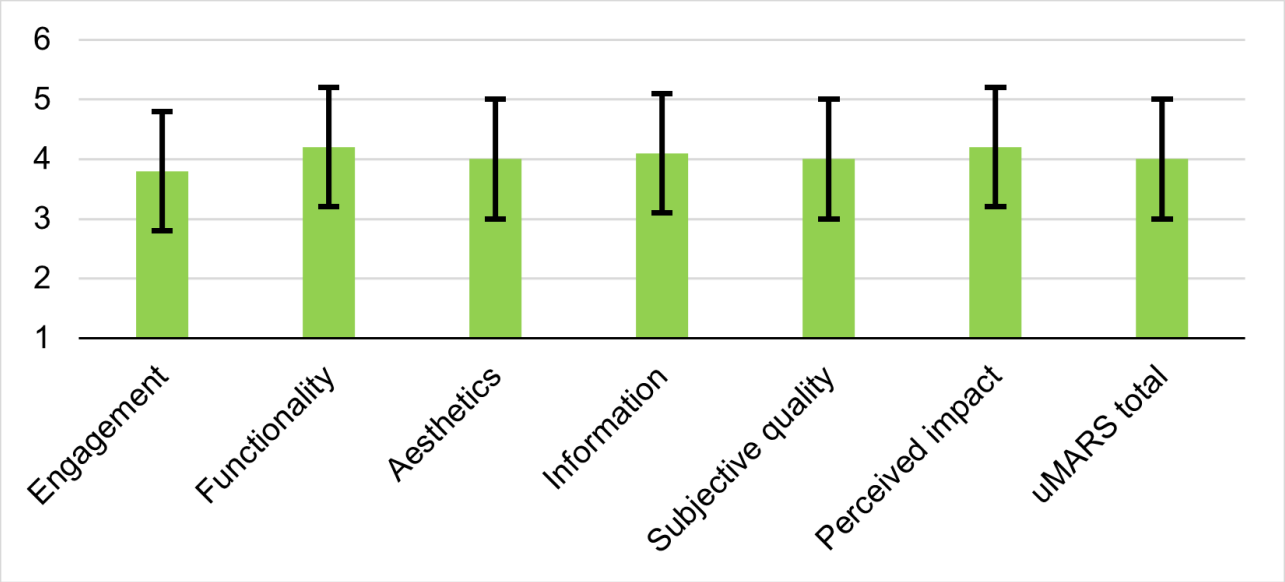

uMARS

Functionality had the highest mean score (mean 4.33/5, SD 0.63), with a minimum of 3.25/5 and a maximum of 5.00, indicating excellent performance in terms of ease of use, navigation, and general technological reliability. Information quality and perceived impact both scored high (mean 4.20/5, SD 0.68; mean 4.20/5, SD 0.84, respectively), indicating that the app was perceived as informative and potentially beneficial in influencing user behavior. Aesthetics (mean 4.05/5, SD 0.94) and subjective app quality (mean 4.10/5, SD 0.73) indicate positive feedback in terms of visual appeal and overall user satisfaction. The engagement subscale had the lowest mean score (mean 3.78/5, SD 0.95), with scores ranging from 2.20/5 to 5.00, indicating more variation in users' experiences with the app's interactivity and motivational aspects, and highlighting potential for growth in terms of interactivity and motivation to use the app on a regular basis. The overall uMARS total score was 4.10/5 (SD 0.76), indicating that users found the app to be of high quality and effective in supporting its intended purposes. The overall uMARS result is presented in Figure 5.

Interview Results

The interview results, presented in Table 2, highlight participants' generally positive perceptions of the app, including its usefulness, ease of navigation, self-management support, and suggestions for further improvement.

XSL•FO
RenderX

Table . Selective illustrative quotations about the overall impressions, likes and dislikes, feature usage, and cultural fit of the app as perceived by the participants.

Theme	Illustrative quotation
Overall impressions of the app	• "It is a very good and useful app." [P01[a]] • "Initially, I did not expect the app to include such a wide range of features, especially considering its localization. I was not anticipating a highly developed platform. However, what I observed was quite impressive. The app is very well-designed." [P02]
Likes and dislikes	• "Overall the app does not have confusing part. But I think it will be better if the graphic designs are improved more." [P02] • "The app is quite simple and easy to navigate, even for people who are not very familiar with technology." [P07]
Mostly used features	• "I personally found the reminder very helpful and I used it a lot. For people with a tight schedule, we tend to forget a lot. I tend to forget my insulin shots. So the reminder helped a lot." [P02] • "I used the dashboard to register my glucose levels. It will help me significantly during my doctor visits. Previously, I recorded my measurements on paper, but now I use the app to electronically log my readings and visualize how my blood sugar has been over time." [P013]
Impact on self-management	• "I educated me a lot. It encouraged me do physical exercise daily and also share my condition with people I love." [P01] • "The fact that I can log detailed events such as the meals and remarks, in addition to my blood glucose levels, is very helpful." [P08]
Recommendations	• "I would love the app to incorporate detail food guide/meal guide." [P06] • "I suggest the app periodically provide new information/findings about diabetes." [P011]

[a]P01: patient 1.

Discussion

Principal Findings

In this paper, we presented a user-centered design and development of *DiabAid Nexus*, a mHealth app developed to support diabetes self-management. The app is specifically designed to address the challenges and barriers faced by individuals living with diabetes, particularly in LRSs and environments with limited access to continuous monitoring, such as those commonly found in Sub-Saharan Africa.

The typical routine case of diabetes diagnosis, management, and coping in the life of a person living with diabetes in LRSs is highlighted in Figure 1. It shows persistent weaknesses in Ethiopia's diabetes care delivery, including limited patient education, incomplete paper-based monitoring, rushed consultations, and inadequate continuity of care. Despite national efforts to strengthen services, individuals are often left without essential resources for effective care. These challenges mirror those reported in other resource-limited settings, where patients often present after acute complications and leave health facilities without a sufficient understanding of diabetes as a lifelong condition [1,37,38]. Such gaps undermine self-management, increase the risk of poor glycemic control, and contribute to complications and psychosocial distress.

DiabAid Nexus was developed in direct response to these shortcomings, offering patient-centered tools that strengthen self-management outside the clinic setting. By replacing paper-based glucose logs with structured digital entries and allowing patients to track insulin use, meals, and contextual factors, the app provides individuals with a clearer picture of their own health patterns. The ability to view trends and generate summaries supports patient reflection and enhances their capacity to engage more meaningfully during follow-up visits, even when consultations are brief. Comprehensive, clinically validated educational resources embedded in the app further address the shortage of structured diabetes education, equipping patients with practical knowledge on treatment, lifestyle modification, and complication prevention. In addition, reminders and emergency information features support adherence and safety in day-to-day life.

A notable strength of *DiabAid Nexus* lies in its bilingual design. By providing content and functionality in both English and Amharic, the app reduces language barriers that frequently prevent patients from fully understanding medical advice or self-management guidelines. This feature expands accessibility beyond urban, educated populations to include a broader demographic, thereby promoting equity in digital health interventions. In a context where low health literacy and language diversity limit the effectiveness of traditional education, bilingual delivery enhances comprehension,

XSL•FO
RenderX

empowers patients to take ownership of their condition, and reduces reliance on overburdened clinical encounters [39].

Taken together, these features position *DiabAid Nexus* as a contextually tailored, patient-empowering mHealth tool. Rather than replacing healthcare providers, it complements the limited support available within the health system by strengthening patients' ability to manage their condition and fostering self-efficacy.

While numerous diabetes management apps exist globally (eg, *Blood Sugar Diary* (MedM Inc) [40], *mySugr - Diabetes Tracker Log* (mySugr GmbH) [41], *WeCare Diabetes* (MEGA LIFESCIENCES) [42]), many are optimized for high-resource contexts and often emphasize a single function (eg, logging or education) [40,42] or assume access to advanced technologies, such as insulin or continuous glucose monitoring systems [41,43]. *DiabAid Nexus* is distinct in both design rationale and scope. To our knowledge, it is the first mHealth app developed specifically in response to the authentic, unmet needs of individuals living with diabetes in Ethiopia, as identified through direct user engagement and contextual research. In terms of contextual adaptation, the app provides bilingual content (English and Amharic), thereby enhancing accessibility for users with limited proficiency in global languages, and delivers clinically validated education in a format intended to reduce fear, misinformation, and stigma that can undermine self-management. In terms of functionality, DiabAid Nexus goes beyond replacing paper logbooks; its dashboard supports structured capture of glucose values alongside insulin doses, meal intake, and contextual remarks. This helps address common gaps, such as undocumented insulin adjustment among syringe users. The My Journey module further differentiates the app by enabling customizable trend visualization (fasting vs postprandial filters), generating structured summaries (including PDF export and HbA_{1c} estimation) to support continuity of care during brief consultations. Finally, unlike many apps that focus on a single component, *DiabAid Nexus* integrates tracking, tailored education, reminders, emergency information, and FAQs into 1 user-centered platform designed for resource-constrained settings. A feature-level comparison with diabetes mHealth apps is provided in Table 3 to highlight differences in functionality, contextual adaptation, and intended user population.

Table . Comparison of DiabAid Nexus with typical commercial diabetes apps.

Feature	DiabAid Nexus	Typical commercial diabetes apps
Electronic glucose logging	✔	✔
Insulin dose documentation	✔(Explicitly addresses undocumented adjustments)	Partial
Trend visualization with filters	✔(Customizable and fasting, random, or postprandial filters)	✔
HbA1c[a] estimation	✔	✔
PDF export for clinical visits	✔	Sometimes
Offline or low-data functionality	✔(Designed for low-resource settings)	Often requires a stable internet
Local language support	✔(English+Amharic)	Rare
Context-specific education	✔(Culturally tailored, stigma-sensitive)	Generalized content
Community-oriented emergency info	✔	Rare

[a]HbA1c: hemoglobin A1c.

DiabAid Nexus also aligns with lessons from successful Sub-Saharan African mHealth initiatives. Wali et al [44] describe the community-based adaptation of Medly Uganda heart failure program, highlighting user-centered design, cultural language tailoring, and integration with trusted community health structures to bridge remote communities and clinical care. Similarly, DiabAid Nexus was developed to address context-specific barriers, such as fragmented follow-up and limited patient education, underscoring that effective mHealth in LRSs depends on contextual adaptation and health system alignment rather than technological complexity alone.

The usability evaluation of the mobile app using the uMARS shows that participants generally liked the app, with consistently high scores across key domains. The highest-rated subscale was functionality, indicating that users judged the program to be dependable, simple to use, and technologically sound. This is an important component of digital health interventions, especially in resource-constrained contexts where technological challenges might impede user engagement and long-term adoption [28,45,46]. The high functionality score indicates that the design and development process successfully prioritized user experience and performance stability.

The high ratings for information quality and perceived impact indicate that users found the app's material to be accurate, relevant, and beneficial in shaping their knowledge or self-management habits. This is a positive result, particularly for a digital health tool aimed at individuals with limited access to individualized health education. It also aligns with the app's primary objective of enhancing diabetic self-management, suggesting that it has the potential to positively impact patient outcomes. Studies have demonstrated the effectiveness of apps with educational content in chronic disease management [47]. Furthermore, the result supports the findings by [48] that highlighted internet health information quality plays a stronger

XSL•FO
RenderX

role in impacting patients' trust and compliance. Aesthetics and subjective quality results are supported by the research findings that gave evidence for the importance of aesthetic design in influencing good user views of online interventions, with a focus on the roles of simplicity and craftsmanship. Simplicity in design was found to have a substantial impact on users' perceptions of ease of use, implying that streamlined, uncluttered interfaces improve accessibility and overall user experience [49]. However, there was a little greater variation in aesthetics, with some people scoring it as low as 2.00. This shows that, while the visual design is generally accepted, future revisions of the app may benefit from design changes to accommodate varied user tastes. The comparatively lower engagement score may reflect the app's current focus on functionality and education rather than interactive or motivational features. Research in digital behavior change suggests that gamification, social interaction, and personalization may enhance long-term engagement [50]. Engagement is a significant predictor of adherence and long-term use in digital health applications. This can be achieved by integrating psychological theory while developing digital health tools [23]. As a result, future development should focus on improving this aspect, possibly by incorporating features like gamification or peer assistance functionalities to boost user engagement and retention. During the interview, no explicit negative or critical reactions were reported by participants.

In summary, the result indicates that the app is a well-designed and functional tool with the potential to assist users in self-management. Evidence from randomized controlled trials of other diabetes mHealth apps has demonstrated improvements in glycemic control [18]. While our study did not assess clinical outcomes, these findings suggest that further evaluation of DiabAid Nexus in a clinical trial setting would be warranted. Nonetheless, targeted enhancements, particularly those aimed at increasing engagement, could boost its effectiveness and user retention. These findings lay a solid platform for expanding the app and undertaking more research to determine its influence on clinical outcomes and behavioral change in real-world situations.

Despite its potential, *DiabAid Nexus* also faces several limitations that need to be considered when interpreting these findings and planning future implementation. First, the app's effectiveness depends on long-term user engagement, motivation, and digital literacy. Participants in this study were relatively well-educated, which may limit generalizability to populations with lower (digital) literacy or limited experience with smartphones. In settings where digital literacy is lower, community-based digital health training may be required to ensure equitable access. Furthermore, scalability may be constrained by factors such as smartphone ownership, stable electricity, and internet connectivity. While the app is designed to function largely offline after installation, initial download, authentication, and updates require connectivity. Socioeconomic disparities may also influence long-term implementation. While the bilingual design improves accessibility, it does not fully address barriers among patients with very low literacy levels or those who rely on other local languages beyond Amharic. Moreover, users are required to manually log their insulin and glucose values into the app. It is also worth documenting which language version each user evaluated. Future updates should aim to integrate features such as image recognition in order to integrate data directly from glucometers by using the camera feature of smartphones. This would simplify glucose tracking for users, reduce data entry errors by patients, and enhance interactivity and accuracy in daily monitoring. Beyond data capture, incorporating "social" and clinical communication features may enhance long-term engagement and self-management, such as secure messaging or data-sharing with health care providers for feedback, as well as optional peer or care-partner support functions. As highlighted in Azevedo et al [26], app components and long-term engagement appear to influence glycemic outcomes among older adults, suggesting that functionality beyond basic self-monitoring may enhance long-term effectiveness. Finally, as with other mHealth interventions, long-term sustainability and scalability will require strategies for integration with national digital health initiatives, and partnership with public health institutions, ongoing technical support, periodic content updates, and alignment with national health system priorities.

Conclusions

Grounded in rigorous formative research that offers an evidence-based understanding of the contextual needs, barriers, and preferences of individuals living with diabetes in Sub-Saharan Africa, *DiabAid Nexus* mHealth app for diabetes self-management was developed based on previous formative research and user input. The alpha testing phase, conducted in collaboration with the developer team, enabled the iterative refinement of the app before formal usability testing. Adhering to a user-centered design approach throughout various stages of development further ensured that the app aligns closely with the lived experiences and expectations of people with diabetes. Finally, the use of both quantitative data from the uMARS usability survey and qualitative insights from interviews enhanced the comprehensiveness of the evaluation, suggesting that the app is both feasible and acceptable for use among diverse users living with diabetes.

Acknowledgments

We thank the participants from the Ethiopian Diabetes Association and Jimma University Medical Center for their contributions. We also acknowledge the Network for Advancement of Sustainable Capacity in Education and Research in Ethiopia (NASCERE).

XSL•FO
RenderX

Funding

This research received no funding.

Data Availability

The data supporting the findings of this study are not publicly available due to privacy and ethical restrictions. De-identified data may be obtained from the corresponding author upon reasonable request and approval by the institutional ethics review board.

Authors' Contributions

GTA, GLS, and JMA contributed to the study's conceptualization and methodology. GTA and EGH conducted data curation and formal analysis, and GTA developed the software. EGH, GLS, and JMA supervised the study and validated the tool. GTA drafted the original manuscript, and GLS and JMA reviewed and edited the final version.

Conflicts of Interest

None declared.

Checklist 1
mERA (mHealth evidence reporting and assessment) checklist.
[PDF File, 59 KB - jmir_v28i1e87110_app1.pdf]

References

1. Zimmermann M, Bunn C, Namadingo H, Gray CM, Lwanda J. Experiences of type 2 diabetes in sub-Saharan Africa: a scoping review. Glob Health Res Policy 2018;3(1):25. [doi: 10.1186/s41256-018-0082-y] [Medline: 30214942]
2. The global health observatory: indicator metadata registry list. WHO World Health Organization. 2025. URL: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/761 [accessed 2025-07-30]
3. Kasper DL, Fauci AS, Hauser SL, Longo DL, Jameson JL, Loscalzo L. Diabetes mellitus. In: Harrison's Manual of Medicine, 19th edition: McGraw-Hill; 2017.
5. Mercer T, Chang AC, Fischer L, et al. Mitigating the burden of diabetes in sub-Saharan Africa through an integrated diagonal health systems approach. Diabetes Metab Syndr Obes 2019;12:2261-2272. [doi: 10.2147/DMSO.S207427] [Medline: 31802925]
5. Breland JY, McAndrew LM, Gross RL, Leventhal H, Horowitz CR. Challenges to healthy eating for people with diabetes in a low-income, minority neighborhood. Diabetes Care 2013 Oct;36(10):2895-2901. [doi: 10.2337/dc12-1632] [Medline: 23877980]
6. Magliano DJ, Boyko EJ, IDF Diabetes Atlas 10th edition scientific committee. IDF Diabetes Atlas 2021. 2021. [Medline: 35914061]
7. Zhou B, Rayner AW, Gregg EW, et al. Worldwide trends in diabetes prevalence and treatment from 1990 to 2022: a pooled analysis of 1108 population-representative studies with 141 million participants. The Lancet 2024 Nov;404(10467):2077-2093. [doi: 10.1016/S0140-6736(24)02317-1] [Medline: 39549716]
8. Diabetes | Factsheet. World Health Organization African Region. 2025. URL: https://www.afro.who.int/health-topics/diabetes [accessed 2025-07-30]
9. Vanderniet JA, Jenkins AJ, Donaghue KC. Epidemiology of type 1 diabetes. Curr Cardiol Rep 2022 Oct;24(10):1455-1465. [doi: 10.1007/s11886-022-01762-w] [Medline: 35976602]
10. Oleribe OO, Momoh J, Uzochukwu BS, et al. Identifying key challenges facing healthcare systems in Africa and potential solutions. Int J Gen Med 2019;12:395-403. [doi: 10.2147/IJGM.S223882] [Medline: 31819592]
11. Snouffer E. Healthy eating challenges for diabetes self-management. Diabetes Voice Glob Perspect. 2015 p. 5-9 URL: https://diabetesvoice.org/en/ [accessed 2025-09-25]
12. Volkov I, Radchenko G. DiaMeter: a mobile application and web service for monitoring diabetes mellitus. : Institute of Electrical and Electronics Engineers Inc; 2020 Presented at: 2020 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT); May 14-15, 2020; Yekaterinburg, Russia p. 384-387. [doi: 10.1109/USBEREIT48449.2020.9117654]
13. Artola G, Torres J, Larburu N, Álvarez R, Muro N. Development and usability assessment of a semantically validated guideline-based patient-oriented gestational diabetes mobile app. Commun Comput Inf Sci 2020;1297:237-259. [doi: 10.1007/978-3-030-66196-0_11]
14. Pamungkas RA, Usman AM, Chamroonsawasdi K. A smartphone application of diabetes coaching intervention to prevent the onset of complications and to improve diabetes self-management: a randomized control trial. Diabetes & Metabolic Syndrome: Clinical Research & Reviews 2022 Jul;16(7):102537. [doi: 10.1016/j.dsx.2022.102537]

XSL•FO
RenderX

15. Buss VH, Varnfield M, Harris M, Barr M. A mobile app for prevention of cardiovascular disease and type 2 diabetes mellitus: development and usability study. JMIR Hum Factors 2022 May 10;9(2):e35065. [doi: 10.2196/35065] [Medline: 35536603]
16. Gerber BS, Biggers A, Tilton JJ, et al. Mobile health intervention in patients with type 2 diabetes: a randomized clinical trial. JAMA Netw Open 2023 Sep 5;6(9):e2333629. [doi: 10.1001/jamanetworkopen.2023.33629] [Medline: 37773498]
17. Prihatin Putri DM, Suhoyo Y, Putri Pertiwi AA, Effendy C. Integrated Diabetes Self-Management (IDSM) mobile application to improve self-management and glycemic control among patients with type 2 diabetes mellitus (T2DM) in Indonesia: a mixed methods study protocol. PLoS ONE 2022;17(11):e0277127. [doi: 10.1371/journal.pone.0277127] [Medline: 36441733]
18. Yu Y, Yan Q, Li H, et al. Effects of mobile phone application combined with or without self-monitoring of blood glucose on glycemic control in patients with diabetes: a randomized controlled trial. J Diabetes Investig 2019 Sep;10(5):1365-1371. [doi: 10.1111/jdi.13031] [Medline: 30815973]
19. Zaghloul H, Fanous K, Ahmed L, et al. Digital health literacy in patients with common chronic diseases: systematic review and meta-analysis. J Med Internet Res 2024;27:e56231-e56231. [doi: 10.2196/56231]
20. Woods L, Duff J, Cummings E, Walker K. Evaluating the development processes of consumer mHealth interventions for chronic condition self-management: a scoping review. Comput Inform Nurs 2019 Jul;37(7):373-385. [doi: 10.1097/CIN.0000000000000528] [Medline: 31135468]
21. Abdelmalak N, Burns J, Suhlrie L, Laxy M, Stephan AJ. Consideration of inequalities in effectiveness trials of mHealth applications - a systematic assessment of studies from an umbrella review. Int J Equity Health 2024 Sep 11;23(1):181. [doi: 10.1186/s12939-024-02267-4] [Medline: 39261871]
22. Birati Y, Yefet E, Perlitz Y, Shehadeh N, Spitzer S. Cultural and digital health literacy appropriateness of app- and web-based systems designed for pregnant women with gestational diabetes mellitus: scoping review. J Med Internet Res 2022 Oct 14;24(10):e37844. [doi: 10.2196/37844] [Medline: 36240008]
23. Geuens J, Geurts L, Gerling K, Croon RD, Abeele VV. A dyad of lenses for the motivational design of mhealth: bridging the gap between health theory and app design. 2019 Presented at: 2019 IEEE International Conference on Healthcare Informatics (ICHI); Jun 10-13, 2019; Xi'an, China p. 1-12. [doi: 10.1109/ICHI.2019.8904839]
24. Hagelsrum KA, Andermo S, Svedberg T, et al. Users' perceptions of an mHealth service to support healthy lifestyle habits among adult individuals with type 2 diabetes in Sweden – a qualitative study. BMC Digit Health 2025;3(1):1. [doi: 10.1186/s44247-025-00154-8]
25. Hoda F, Arshad M, Khan MA, et al. Impact of a mHealth intervention in type 2 diabetes mellitus patients: a randomized clinical trial. SN Compr Clin Med 2023;5(1):1-9. [doi: 10.1007/s42399-023-01564-3]
26. Azevedo RFL, Varzino M, Steinman E, Rogers WA. Evaluating effectiveness of mHealth apps for older adults with diabetes: meta-analysis of randomized controlled trials. J Med Internet Res 2025 Jun 17;27:e65855. [doi: 10.2196/65855] [Medline: 40527504]
27. Wilson K, Bell C, Wilson L, Witteman H. Agile research to complement agile development: a proposal for an mHealth research lifecycle. NPJ Digit Med 2018;1(1):46. [doi: 10.1038/s41746-018-0053-1] [Medline: 31304326]
28. Aboye GT, Simegn GL, Aerts JM. Assessment of the barriers and enablers of the use of mHealth systems in Sub-Saharan Africa according to the perceptions of patients, physicians, and health care executives in Ethiopia: qualitative study. J Med Internet Res 2024 Mar 27;26:e50337. [doi: 10.2196/50337] [Medline: 38536231]
29. Aboye GT, Vande Walle M, Simegn GL, Aerts JM. Current evidence on the use of mHealth approaches in Sub-Saharan Africa: a scoping review. Health Policy Technol 2023 Dec;12(4):100806. [doi: 10.1016/j.hlpt.2023.100806]
30. Funnell MM, Brown TL, Childs BP, et al. National Standards for Diabetes Self-Management Education. Diabetes Care 2010 Jan;33 Suppl 1(Suppl 1):S89-S96. [doi: 10.2337/dc10-S089] [Medline: 20042780]
31. Davis J, Fischl AH, Beck J, et al. 2022 National Standards for Diabetes Self-Management Education and Support. Sci Diabetes Self Manag Care 2022 Feb;48(1):44-59. [doi: 10.1177/26350106211072203] [Medline: 35049403]
32. Flutter. URL: https://flutter.dev/?utm_source=google&utm_medium=cpc&utm_campaign=brand_sem&utm_content=emea_emea&gclsrc=aw.ds&gad_source=1&gad_campaignid=13044107358&gbraid=0AAAAAGNfSHIO4fy1h3HEZkQdwCQsa&gclid=CjDKCQjwqpDFBhDhARsAIHTIkvfnkdqpBa8Q9ORP5gWc7wfU-v [accessed 2026-05-27]
33. Dart. URL: https://dart.dev/ [accessed 2026-05-27]
34. Google Firebase Authentication. URL: https://firebase.google.com/docs/auth [accessed 2026-05-27]
35. Ethio telecom expands 5G network to Jimma, advancing Ethiopia's digital future. Ethio telecom. 2026. URL: https://www.ethiotelecom.et/ethio-telecom-expands-5g-network-to-jimma-advancing-ethiopias-digital-future/ [accessed 2026-02-19]
36. Stoyanov SR, Hides L, Kavanagh DJ, Wilson H. Development and validation of the User version of the Mobile Application Rating Scale (uMARS). JMIR Mhealth Uhealth 2016 Jun 10;4(2):e72. [doi: 10.2196/mhealth.5849] [Medline: 27287964]
37. Ageru TA, Le CN, Wattanapisit A, et al. Diabetes self-care intervention strategies and their effectiveness in Sub-Saharan Africa: a systematic review. PLoS ONE 2024;19(10):e0305860. [doi: 10.1371/journal.pone.0305860] [Medline: 39405277]
38. Bonet-Olivencia S, Carrillo-Leal J, Rao A, Sasangohar F. User-centered design of a diabetes self-management tool for underserved populations. J Diabetes Sci Technol 2024 Jan;18(1):22-29. [doi: 10.1177/19322968231212220] [Medline: 37978811]

XSL•FO
RenderX

39. Chu JN, Sarkar U, Rivadeneira NA, Hiatt RA, Khoong EC. Impact of language preference and health literacy on health information-seeking experiences among a low-income, multilingual cohort. Patient Educ Couns 2022 May;105(5):1268-1275. [doi: 10.1016/j.pec.2021.08.028] [Medline: 34474924]
40. MedM Inc. Blood sugar diary. Google Play. URL: https://play.google.com/store/apps/details?id=com.medm.medmbg.diary&hl=en [accessed 2025-09-24]
41. mySugr GmbH. MySugr - diabetes tracker log. Google Play. URL: https://play.google.com/store/apps/details?id=com.mysugr.android.companion&hl=en [accessed 2025-09-24]
42. MEGA LIFESCIENCES. WeCare diabetes. Google Play. URL: https://play.google.com/store/apps/details?id=com.joinwecare.patient&hl=en [accessed 2025-09-24]
43. Dexcom apps. Dexcom Diabetes. URL: https://www.dexcom.com/apps [accessed 2026-05-27]
44. Wali S, Ssinabulya I, Muhangi CN, et al. Bridging community and clinic through digital health: community-based adaptation of a mobile phone-based heart failure program for remote communities in Uganda. BMC Digit Health 2023;1(1):20. [doi: 10.1186/s44247-023-00020-5] [Medline: 38800672]
45. Renukappa S, Mudiyi P, Suresh S, Abdalla W, Subbarao C. Evaluation of challenges for adoption of smart healthcare strategies. Smart Health (2014) 2022 Dec;26(January):100330. [doi: 10.1016/j.smhl.2022.100330]
46. Hengst TM, Lechner L, Dohmen D, Bolman CAW. The facilitators and barriers of mHealth adoption and use among people with a low socio-economic position: a scoping review. Digit Health 2023;9:20552076231198702. [doi: 10.1177/20552076231198702] [Medline: 37691766]
47. Bowman C, Lunyera J, Alkon A, et al. A patient safety educational tool for patients with chronic kidney disease: development and usability study. JMIR Form Res 2020 May 28;4(5):e16137. [doi: 10.2196/16137] [Medline: 32463366]
48. Lu X, Zhang R, Wu W, Shang X, Liu M. Relationship between internet health information and patient compliance based on trust: empirical study. J Med Internet Res 2018 Aug 17;20(8):e253. [doi: 10.2196/jmir.9364] [Medline: 30120087]
49. Denison-Day JL, Muir S, Newell C, Appleton KM. The role of aesthetics in intentions to use digital health interventions. PLOS Digit Health 2023 Jun;2(6):e0000274. [doi: 10.1371/journal.pdig.0000274] [Medline: 37347727]
50. Saboia IF, Almeida AM, Pernencar C, Veloso AI. Gamification strategies in weight control applications, where “losing (weight) is winning”. 2018 Presented at: 2018 IEEE 6th International Conference on Serious Games and Applications for Health (SeGAH); May 16-18, 2018; Vienna p. 1-8. [doi: 10.1109/SeGAH.2018.8401375]

Abbreviations

HbA_{1c}: hemoglobin A1c
LRS: low-resource setting
mERA: mHealth evidence reporting and assessment
mHealth: mobile health
uMARS: Mobile App Rating Scale

Edited by A Mavragani; submitted 05.Nov.2025; peer-reviewed by O Ismaila, RFL Azevedo; revised version received 19.Feb.2026; accepted 24.Mar.2026; published 02.Jun.2026.

Please cite as:
Aboye GT, Gebremichael EH, Simegn GL, Aerts JM
User-Centered Diabetes Self-Management App (DiabAid Nexus) in Sub-Saharan Africa: Development and Usability Study
J Med Internet Res 2026;28:e87110
URL: https://www.jmir.org/2026/1/e87110
doi:10.2196/87110

© Genet Tadese Aboye, Ermias Habte Gebremichael, Gizeaddis Lamesgin Simegn, Jean-Marie Aerts. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development of an mHealth Intervention for Reducing Sedentary Behavior in Older Adults: Delphi Study

Siqing Chen[1,2,3,4], PhD; Xiaofei Nie[1], BSN; Jiekai Yang[1], BSN; Yaqin Li[1], BSN; Lili Yang[1], PhD

[1]Nursing Department, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, 3 East Qingchun Road, Hangzhou, Zhejiang, China

[2]Department of Epidemiology, Harvard T.H Chan School of Public Health, Boston, MA, United States

[3]School of Nursing, Wenzhou Medical University, Wenzhou, China

[4]Department of Biobehavioral Sciences, Teachers College, Columbia University, New York, NY, United States

Corresponding Author:
Lili Yang, PhD
Nursing Department, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, 3 East Qingchun Road, Hangzhou, Zhejiang, China

Abstract

Background: Sedentary behavior among older adults is a major public health concern, contributing to the increased risk of chronic diseases and functional decline. With aging populations worldwide, prolonged sitting time (averaging up to 13 h/d in older adults) has been independently associated with cardiovascular disease, metabolic disorders, cognitive decline, and all-cause mortality. Mobile health (mHealth) interventions offer a promising approach to address this issue. However, there remains a lack of evidence-based, systematically developed mHealth programs specifically targeting sedentary behavior in older populations.

Objective: This study aimed to develop an mHealth intervention program for reducing sedentary behavior in older adults using the Delphi consensus method.

Methods: Guided by the Behavior Change Wheel framework, a preliminary mHealth intervention was developed using a combination of qualitative and quantitative methods, including a comprehensive literature review, clinical guidelines analysis, qualitative interviews, and a cross-sectional survey. The intervention was then refined through 2 rounds of Delphi surveys with 16 multidisciplinary experts in geriatric care, behavioral science, and health promotion. Consensus criteria were predefined as mean importance score >3.5 and coefficient of variation ≤0.25 on a 5-point Likert scale.

Results: Both Delphi rounds achieved 100% response rates, with high expert authority coefficients (Cr=0.900 for Round 1 and Cr=0.907 for Round 2). The Kendall coordination coefficients (Kendall W) were 0.151 (P<.001) and 0.214 (P=.001) for the 2 rounds, respectively. Following 2 rounds of expert consultation, a total of 27 intervention items were finalized, comprising 3 core components addressing capability (eg, knowledge provision and behavioral skills training), opportunity (eg, social support and environmental restructuring), and motivation (eg, goal-setting, feedback, and incentives) factors influencing sedentary behavior.

Conclusions: This study developed a theoretical framework-based, consensus-driven mHealth intervention program for reducing sedentary behavior in older adults. The intervention uniquely integrates the Behavior Change Wheel framework with expert validation, offering a comprehensive approach that simultaneously targets capability, opportunity, and motivation. The findings provide a structured foundation for future feasibility testing and effectiveness evaluation of mHealth interventions in aging populations. Future researchers should translate the developed mHealth intervention into an adaptive mHealth platform, followed by pilot testing and large-scale randomized controlled trials to evaluate its feasibility and effectiveness in real-world settings.

(J Med Internet Res 2026;28:e83302) doi:10.2196/83302

KEYWORDS

older adults; sedentary behavior; mobile health; Delphi technique; China

Introduction

Sedentary behavior, defined as any waking activity characterized by an energy expenditure ≤1.5 metabolic equivalents while in a sitting, reclining, or lying posture [1], has emerged as a critical public health concern among older adults. With age-related physiological decline, increased prevalence of chronic diseases, and reduced social participation, older adults exhibit alarmingly prolonged sedentary time, averaging up to 13 hours per day [2]. This excessive sedentary behavior has been robustly associated with a spectrum of adverse health outcomes, including cardiovascular diseases, metabolic disorders such as obesity and type 2 diabetes, accelerated cognitive decline, and even increased all-cause mortality [3-5].

The rapid advancement of mobile health (mHealth) technologies presents a transformative opportunity for promoting behavioral

XSL•FO
RenderX

change in this population [6,7]. mHealth interventions offer unique advantages, including real-time monitoring, personalized feedback, and scalable delivery, making them particularly suitable for addressing sedentary behavior in community-dwelling older adults [6,8]. However, current mHealth interventions in China remain fragmented, lacking theoretically grounded, systematically developed programs specifically targeting sedentary behavior reduction in aging populations. This critical gap underscores the urgent need for an evidence-based, theoretically informed intervention framework that integrates cutting-edge behavioral science with practical implementation considerations.

The Behavior Change Wheel (BCW) framework was applied in our previous systematic review, meta-analysis, and analysis of relevant clinical guidelines to identify the comprehensive determinants of sedentary behavior. These findings provided the foundation for developing a preliminary mHealth intervention protocol. Building on this work, our goal was to refine the protocol into a final intervention through 2 rounds of expert Delphi consultation. By providing a structured and theoretically grounded approach to intervention development, this study seeks to contribute a foundational framework for future implementation and evaluation of mHealth strategies targeting sedentary behavior in aging populations.

Methods

Overview

This study was reported in accordance with the Conducting and Reporting of Delphi Studies (CREDES) guidelines [9] (Checklist 1).

Establishment of the Research Team

The research team comprised multidisciplinary members with expertise in nursing, health behavior, behavioral science, and exercise physiology. The team collaboratively developed the preliminary intervention components, reviewed and revised the Delphi questionnaires, and analyzed expert feedback through iterative discussions. Data collection and analysis were conducted by the doctoral researcher under the supervision of senior team members. Detailed professional profiles of all team members are presented in Multimedia Appendix 1.

Development of the Expert Consultation Questionnaire

The development of the mHealth intervention preliminary version followed a structured process based on the 3 phases and 8 steps of the BCW framework. A comprehensive literature review and meta-analysis [6], qualitative research [8], and a cross-sectional study [10] previously conducted by the author were conducted to identify key determinants of sedentary behavior in older adults. These studies helped identify components for the mHealth intervention targeting sedentary behavior. Evidence-based guidelines, including the Canadian Physical Activity Guidelines for Older Adults and World Health Organization (WHO) 2020 guidelines [11,12], were also incorporated. The Capability, Opportunity, Motivation–Behavior (COM-B) model and Theoretical Domains Framework (TDF) within the BCW were used to map key intervention functions and behavior change techniques (BCTs) to the identified barriers. Following discussions within the research team, the first-round Delphi questionnaire was developed.

The expert consultation letter comprises 5 key sections. The first section is a cover letter introducing the research objectives, the methodology for completing the consultation form, and the deadline for submission. The second section provides an overview of the intervention program, including its background, purpose, and theoretical foundation. The third section focuses on evaluating the intervention functions and BCTs. Experts are asked to rate the appropriateness and importance of each selected intervention function and BCT using a 5-point Likert scale (1=very inappropriate or unimportant to 5=very appropriate or important). A dedicated column is provided for experts to suggest modifications or additions if necessary. The fourth section presents an evaluation table for the intervention program components, adapted from prior quantitative and qualitative research. Experts assess the reasonableness and significance of key elements, including intervention facilitators, objectives, content, duration, delivery format, and outcome evaluation, using the same 5-point scale. While the consultation covers the overall framework of the intervention, specific details, such as the definition of sedentary behavior, associated risks, intervention duration, and frequency, will be supplemented through a review of relevant domestic and international guidelines, with final approval determined by the research team. The final section is a self-assessment questionnaire for experts, collecting information on their professional background, judgment criteria, and familiarity with the content. The judgment criteria are categorized into theoretical analysis (scored 0.3, 0.2, or 0.1), practical experience (0.5, 0.4, or 0.3), reference to literature (0.1 for all levels), and intuitive selection (0.1 for all levels). Familiarity with the content is rated on a scale from 0 (unfamiliar) to 1.0 (very familiar), with intermediate options of 0.2, 0.5, and 0.8.

Expert Selection Process

Experts were recruited through a combination of author identification from the literature and peer recommendations. The research team identified individuals with demonstrated expertise in geriatric care, behavioral science, and health promotion based on their publications in the field. Additionally, peer recommendations from professionals in these areas were sought to ensure a diverse and knowledgeable panel.

Demographic Characteristics of Consulting Experts

The selection of experts for the Delphi consultation was determined based on the study's scope and feasibility, with reference to established literature recommending a panel size of 15 to 50 participants [13,14]. In this study, a total of 16 multidisciplinary experts were recruited across 2 rounds of consultation, adhering to predefined inclusion criteria and the principles of informed consent and voluntary participation. Eligible experts were required to meet the following qualifications: (1) specialization in geriatric nursing, behavioral health promotion, exercise physiology, clinical medicine, nursing, or related fields; (2) substantial clinical or research experience, including involvement in relevant research projects; (3) possession of at least a bachelor's degree; and (4) holding an intermediate or higher professional title. Among the 16

XSL•FO
RenderX

experts who completed both consultation rounds, 1 was male and 15 were female. All held a bachelor's degree or higher, including 5 with master's degrees and 11 with doctoral degrees. A total of 87.5% (14/16) of the participants held associate senior or higher professional titles. Detailed demographic characteristics are presented in Multimedia Appendix 1.

Expert Consultation Process

The Delphi expert consultation was conducted through both email and face-to-face questionnaire distribution to eligible experts over 2 rounds, with the aim of finalizing the intervention protocol. To ensure independence and anonymity, experts completed the questionnaires independently without group discussion during the process, and all responses were deidentified prior to analysis. Each round was spaced 2 - 4 weeks apart, and experts were given a 2-week response window [13]. In the first round, experts rated the items and provided feedback, which the research team analyzed. Items with low ratings, high variability, or unclear feedback were revised, merged, or added based on expert suggestions. The second-round questionnaire, along with a summary of first-round feedback, was redistributed for reevaluation. Items were considered to have reached consensus if they met the criteria: mean importance score >3.5 and coefficient of variation (CV)≤0.25. Items not meeting these criteria were refined based on expert comments. After the second round, consensus was achieved, and the final intervention was confirmed.

The expert authority coefficient (Cr) was calculated as the average of 2 self-assessed measures: judgment basis (Ca) and familiarity (Cs). This coefficient provided an indication of expert reliability, with values above 0.80 suggesting high authority in the relevant field. Furthermore, the Kendall coordination coefficient (Kendall W) was used to measure the level of agreement among experts. The interpretation of the W value is as follows: W≥0.60 indicates strong consensus, 0.30≤W<0.60 indicates moderate consensus, and W<0.30 indicates weak consensus.

Ethical Considerations

Approval was obtained from the Ethics Committee of the Fourth Affiliated Hospital of Zhejiang University School of Medicine (approval number K2023150). Informed consent was obtained from all participants prior to participation. All data were anonymized prior to analysis to protect participant privacy and confidentiality. No financial compensation was provided to participants.

Statistical Analysis

Data were analyzed using SPSS 27.0 (IBM Corp). Descriptive statistics (proportions, %) were used for qualitative demographic data. Expert engagement was assessed via response rate, while expert authority and consensus were evaluated using the Cr, CV, and Kendall concordance coefficient [13].

Results

Expert Engagement and Authority

Expert engagement reflected their commitment to the study. Across both rounds, all 32 distributed questionnaires (16 per round) were returned (100% response rate). In the first round, 12 of 16 experts provided suggestions, decreasing to 7 in the second round. The authority coefficient (Cr) was calculated as Cr=(Ca+Cs)/2, where Ca (judgment basis) and Cs (familiarity) were derived from expert self-assessments. The Cr values for both rounds were 0.900 and 0.907, respectively (Table 1).

Table . Expert authority coefficients.

Round	Judgment basis (Ca)	Familiarity (Cs)	Authority coefficient (Cr)
1	0.944	0.856	0.900
2	0.944	0.869	0.907

Consensus and Agreement Among Experts

Consensus was evaluated using mean importance scores and CVs. Across both rounds, mean scores ranged from 4.20 - 5.00 (Round 1) to 4.44–5.00 (Round 2), with CVs≤0.25 (Round 1: 0 - 0.22; Round 2: 0 - 0.20), indicating strong agreement. Kendall W revealed significant concordance (Round 1: W=0.151, *P*<.001 and Round 2: W=0.214, *P*=.001; Table 2).

Table . Concentration and coordination degree of expert opinions.

Item	Kendall W	Chi-square (*df*)	*P* value
Round 1			
Overall	0.151	74.527 (33)	<.001
Round 2			
Overall	0.214	61.651 (32)	.001

XSL•FO
RenderX

Modification and Finalization of Questionnaire Content

Overview

Based on expert feedback from 2 rounds of Delphi consultation and discussions within the research team, several revisions were made to the intervention functions, BCTs, and the intervention protocol. In the first round, 1 intervention function was added, 3 BCTs were removed, and 1 additional BCT was deleted. In the second round, one more intervention function was added, another restored, and 3 BCTs were added while 2 were deleted. These modifications are summarized in Multimedia Appendix 2 and Figure 1, with a summary provided below:

Mapping Between Intervention Functions and BCTs

In the first round of expert consultation, experts pointed out that some BCTs were improperly matched with the 3 dimensions of the COM-B model. The research team accepted this suggestion, deleted the inappropriate BCT "1.1 Goal setting (behavior)" under "physical opportunity," and added the "education" intervention function to eliminate barriers caused by limited physical opportunity from both theoretical and practical perspectives. Meanwhile, in response to experts' proposals to add an "Incentive" function to the "Intention" dimension, the research team adopted the suggestion after discussion and supplemented corresponding BCTs; the BCT "5.1 Information about health consequences" deemed mismatched with the incentive function was deleted. In the second round of expert consultation, one expert further noted that skill training under psychological capability overlapped with physical capability. After discussion, the research team adopted this suggestion, deleted the relevant intervention function and its corresponding BCTs under "Psychological capability," and restored the "education" intervention function and the BCT "5.1 Information about health consequences" that were deleted in the first round. This further enhanced the consistency and rationality of the intervention program in terms of the COM-B dimension classification.

Figure 1. Adjustments to intervention functions and behavior change techniques in 2 Delphi rounds. BCT: behavior change technique.

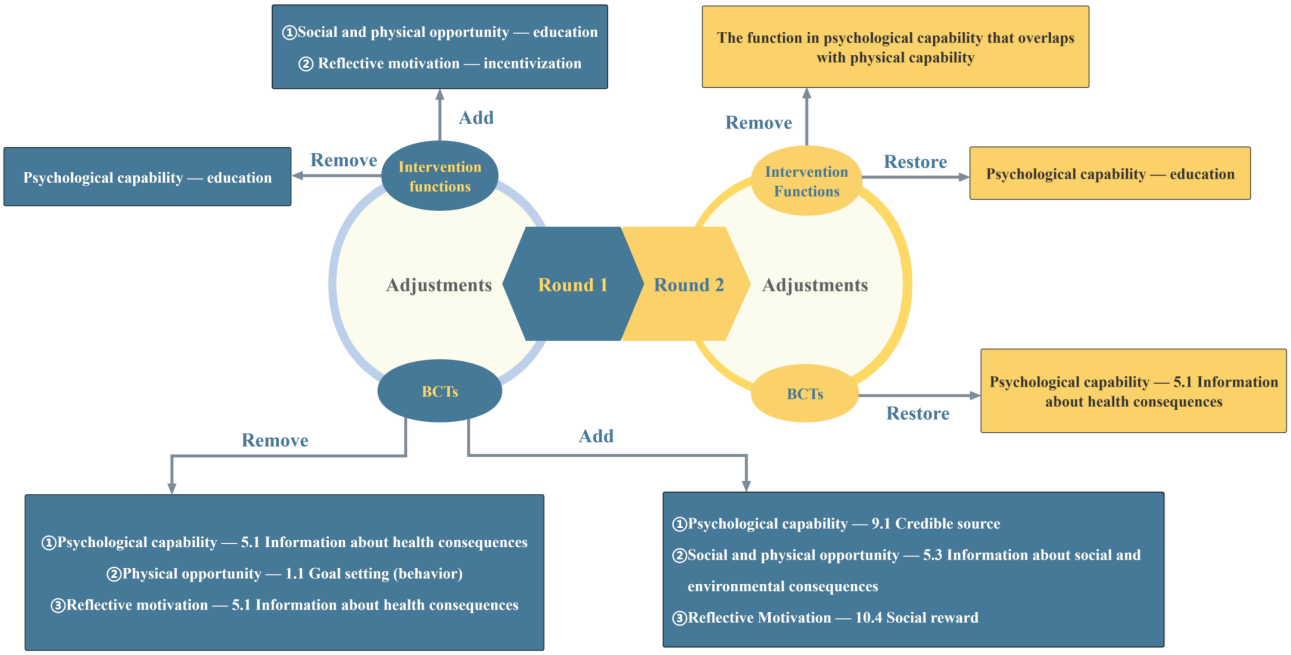

Risk Perception and Educational Content

In the first round of expert consultation, experts recommended introducing authoritative sources to improve information credibility. The research team recognized that this suggestion would enhance the persuasiveness of risk perception, thus adding the BCT "9.1 Credible source" to strengthen intervention effects through forms such as videos from authoritative institutions. Additionally, addressing experts' concerns about the single mapping of educational BCTs, the research team added the BCT "5.3 Information about social and environmental consequences" after discussion to supplement the explanation of social and environmental impacts of behaviors; however, the BCT "5.2 Salience of consequences" was not adopted as it did not fully meet the APEASE (Affordability, Practicability, Effectiveness and cost-effectiveness, Acceptability, Side-effects/safety, and Equity) criteria.

Social Support and Feedback Mechanisms

In the first round of expert consultation, some experts suggested introducing "3.2 Social support (practical)." After evaluating intervention costs, feasibility, and the APEASE criteria, the research team decided not to adopt this suggestion, as practical social support would incur high implementation costs and behavioral feedback lacked corresponding intervention pathways under the current barriers.

The figure illustrates the additions, removals, and restorations of intervention functions and BCTs based on expert feedback in Round 1 and Round 2.

Based on the opinions from 2 rounds of Delphi expert consultation and discussions within the research team, the overall modifications are as follows: after the first round of expert consultation, 1 item was deleted, 3 new items were added, and 17 items were revised; after the second round, 3 items were

XSL•FO
RenderX

deleted, 1 item was added, and 5 items were revised. The experts' detailed revision suggestions and consultation results are presented in Multimedia Appendix 2, summarized as follows.

Intervention Objectives

In the first round of expert consultation, experts suggested setting personalized and progressive reduction targets for screen time and clarifying the connotation of screen time. After discussion, the research team adopted this suggestion, specifying that screen time includes hours spent watching TV, using computers, and mobile phones, which would be mainly assessed through screen timers and questionnaires. Accordingly, the intervention objective was revised from "Reduce the total sedentary behavior duration (≤8 hours) and screen time (≤3 hours) of older adults, with reference to the Canadian 24-Hour Movement Guidelines" to "Reduce the total sedentary behavior duration (≤8 hours) of older adults and achieve personalized, progressive reduction in screen time (≤3 hours), with reference to the Canadian 24-Hour Movement Guidelines." In the second round of expert consultation, experts pointed out that "personalized" and "progressive" were not specifically reflected in the objective statement. Given that further quantification of these elements in the intervention objective was impractical and they had already been concretely implemented in the intervention details, the research team decided to revert to the original expression in the Canadian 24-Hour Movement Guidelines. The final intervention objective was confirmed as "Reduce the total sedentary behavior duration (≤8 hours) and screen time (≤3 hours) of older adults, with reference to the Canadian 24-Hour Movement Guidelines."

Intervention Implementers

Experts noted that fitness coaches in the implementation team had varying professional backgrounds and competence levels, and their professionalism in intervening in older adults' sedentary behavior might be insufficient. They recommended that personnel with stronger professional backgrounds should lead the research implementation. The research team adopted this suggestion and revised "fitness coaches" in the original items to "experts from physical education colleges." Furthermore, experts proposed that intervention implementers should not be limited to medical staff from geriatrics or rehabilitation departments and experts from physical education colleges, but should also include systematically trained researchers. Considering the indispensable role of researchers in the design, implementation, and evaluation of the program, the research team decided to include researchers in the scope of intervention implementers to enhance the scientificity and systematicness of intervention implementation.

Intervention Content

In response to experts' comments that some items were vaguely expressed and lacked operability, the research team targeted revisions to the intervention content from 4 aspects: health information presentation, implementation and safety, forms of social support, and achievement incentive mechanisms.

1. Specification and contextualization of health information presentation: experts pointed out that some health information was presented in a single form and lacked contextual relevance. The research team made the following adjustments accordingly: in terms of presentation forms: Item 1.4 was revised from "providing only graphic guidelines for interrupting sedentary behavior" to include video demonstrations and voice guidance to improve learning outcomes; Item 1.5 was refined from the general "repeated practice of interrupting sedentary behavior" to "the platform triggers sedentary behavior reminders every 30 minutes to help individuals practice repeatedly until they proficiently master interruption skills"; To address the lack of contextualization in Items 1.7 and 3.7, the content was adjusted to present sedentary behavior risks and potential adverse health outcomes in different life scenarios (eg, entertainment, work, and aging) to enhance contextual relevance and targeting; As suggested by experts, Item 1.1 "Definition of sedentary behavior" was revised to "Overview of sedentary behavior" to ensure the systematicness and completeness of information; For Items 1.2 and 3.9 with unclear information sources, guidelines issued by authoritative institutions such as the WHO were explicitly cited to improve the scientificity and reliability of information sources.
2. Optimization of implementation methods and safety: experts believed that some intervention measures still had room for improvement in professionalism and safety. The research team adopted relevant suggestions and made revisions: Item 1.8 adjusted "fitness coaches" to "experts from physical education colleges" (with stronger professional backgrounds), and proposed personalized programs considering physical constraints such as joint limitations, lumbar pain, and postoperative recovery, while adding safety reminders such as "stop if pain occurs"; as recommended by experts, personalized setting of reminder frequency was added, and Item 1.9 was revised from "fixed reminders every 30 minutes" to "users can set personalized reminder frequencies between 15 - 60 minutes"; Item 1.10 supplemented specific interruption methods (eg, standing up and walking) and their rationale on the basis of the original "providing instructions"; to address the unclear demonstration content in Items 1.11 and 3.2, "providing short video cases" was added to enhance demonstration and comprehensibility; In terms of feedback mechanisms: Item 1.12 was adjusted to conduct real-time monitoring through smart wearable devices, generate personalized visual reports, and provide immediate positive feedback after users stand up and move (eg, "Great! You have been active for 5 min, keep it up!").
3. Clarification of social support forms: experts pointed out that the implementation paths of some social support–related items were not clear enough and recommended enhancing interactivity and participation. The research team made the following adjustments: Item 2.1 further clarified the composition of the Question and Answer team to ensure the quality of professional guidance; Items 2.2 and 3.17 were revised from "advising individuals to invite relatives and friends" to "the platform supports users to actively invite relatives and friends, who can receive reminders and send encouraging messages"; Item 2.3 was refined to "the

XSL•FO
RenderX

platform matches peers with less sedentary behavior based on user data and supports group challenges" to enhance social interactivity.

4. Stratification and personalization of achievement incentive mechanisms: regarding achievement incentives, experts suggested providing differentiated feedback based on users' completion status to enhance intervention attractiveness. The research team made corresponding optimizations: Item 3.5 was adjusted from a single "ranking list" to a "progress ranking list," with encouraging cards such as "Keep going tomorrow" added for users who fail to meet targets; Item 3.6 pushed stratified feedback based on goal achievement: users who complete targets can unlock badges and receive voice congratulations, while those who do not receive encouraging cards; Item 3.11 added expressions such as "you can" and "you will definitely" to make the feedback language more in line with older adults' usage habits.

Intervention Duration

The initial proposed intervention duration was 12 weeks. Experts suggested daily follow-up in the first 2 weeks, with a gradual reduction in frequency thereafter. Additionally, experts recommended adding a "maintenance phase" strategy, converting interventions after 12 weeks to monthly pushes to prevent relapse. Meanwhile, experts noted that behavior change may require a longer period to stabilize. A recent systematic review indicated that significant behavioral changes can be observed with a 12-week intervention, which also has good feasibility. The research team decided to adopt the experts' suggestions and revised the intervention duration to: "12 weeks (daily follow-up in weeks 1–4 → twice-weekly follow-up in weeks 5–8 → once-weekly consolidation and maintenance in weeks 9–12)."

Intervention Forms

The original planned intervention form was a combination of a WeChat (Tencent Holdings Ltd) mini-program and smart wearable devices. However, experts suggested introducing diversified forms such as telephone or remote face-to-face interventions. After discussion, the research team decided to switch to a mobile app, mainly because the app offers stronger functional scalability and a better user experience compared to WeChat mini-programs and can better support diversified forms such as telephone and remote video interventions. Telephone and remote video interventions were integrated into the online consultation module of the app, which not only enhanced intervention flexibility but also avoided overreliance on offline interventions. Meanwhile, experts recommended strengthening the feasibility assessment of smart wearable devices, which will be continuously monitored during subsequent implementation.

Indicators for Evaluating Intervention Effects

The originally proposed evaluation indicators included objective sedentary behavior duration, frequency, and count; screen time; activity time; and login time and frequency. Experts suggested adding physiological indicators, subjective indicators (eg, quality of life scale or self-efficacy), and recording "reasons for dropout" to construct a multidimensional evaluation system with subjective-objective cross-validation, incorporating more objective indicators, disease indicators, and satisfaction indicators. Regarding the classification of primary and secondary outcomes of intervention effects, experts recommended clarifying primary and secondary outcomes and noted that the selection of physiological indicators should be consistent with intervention strategies. After 2 rounds of expert consultation, the research team decided to adopt the experts' suggestions and confirmed the primary outcome indicators as objective sedentary behavior duration, frequency, count, and subjective sedentary behavior (screen time); secondary outcomes include activity time, sedentary behavior, and physical activity–related indicators, physical activity motivation, health literacy, socioeconomic status, social support, waist circumference, weight, blood pressure, quality of life, and feasibility evaluation.

The comprehensive mapping of relationships among the COM-B model, TDF domains, intervention functions, BCTs, and corresponding intervention components, which initially included 29 items, is presented in Multimedia Appendix 3. The expert panel reviewed this initial list and provided feedback, leading to the removal of 1 item, rewording of 17, addition of 3, and combination of 2, resulting in a revised total of 30 items. In the second round, the revised list of 30 items was further reviewed, with additional modifications: 3 items were removed, 5 were reworded, 3 new items were added, and 2 items were combined, yielding a final total of 27 items (Figure 2). A comparison of the preliminary and finalized versions of the mHealth intervention for reducing sedentary behavior in older adults across both rounds is provided in Table 3. A detailed item-by-item comparison of revisions across the 2 rounds is provided in Multimedia Appendix 4, and a comparison of the preliminary and finalized versions in Multimedia Appendix 5.

XSL•FO
RenderX

Figure 2. Summary of item modifications across 2 Delphi rounds.

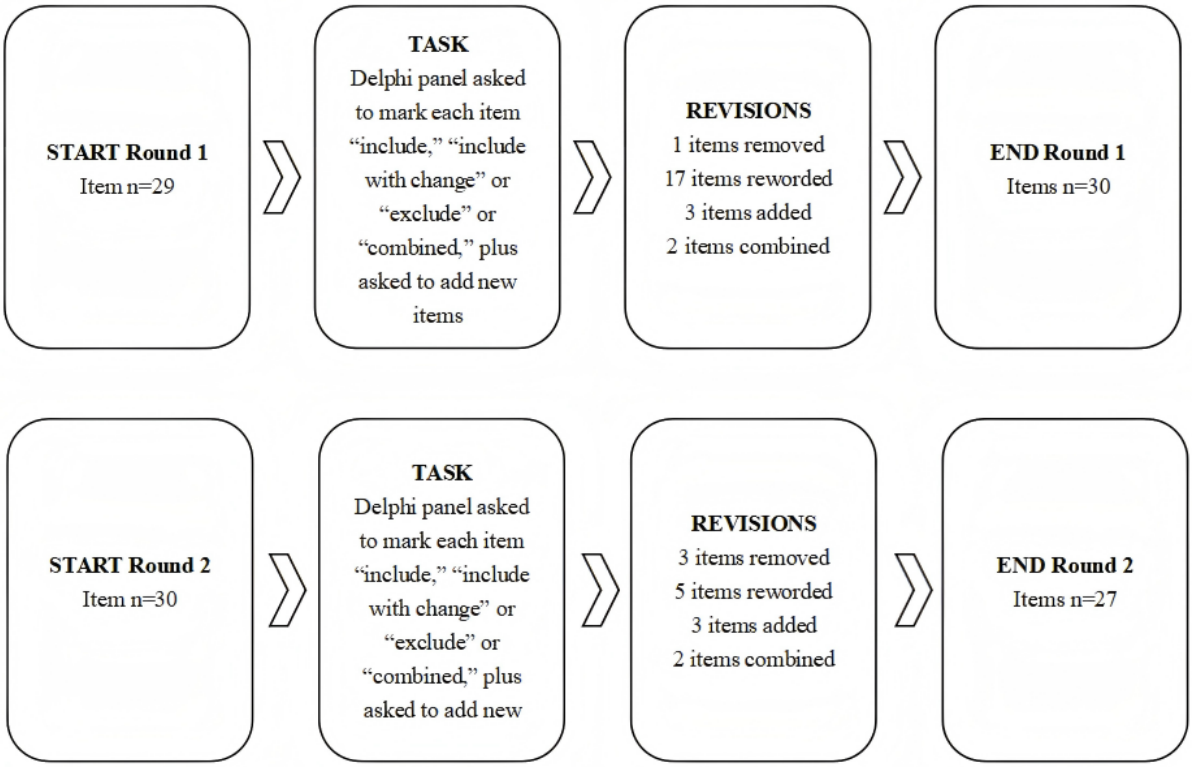

XSL•FO
RenderX

Table . Finalized versions of the mobile health intervention for reducing sedentary behavior in older adults[a].

Intervention element	Intervention (final version)
Intervener	• Geriatric and rehabilitation department medical staff, sports science experts, trained researchers, etc
Intervention goal	• Reduce older adults' total sedentary behavior time (≤8 h) and screen time (≤3 h).
Intervention duration	• 12 weeks (daily follow-up wk 1 - 4 → twice-weekly follow-up wk 5 - 8 → once-weekly consolidation or maintenance wk 9 - 12).
Intervention format	• App and wearable devices.
Intervention effect evaluation	• Primary outcomes: objective sedentary behavior duration, frequency, count; subjective sedentary behavior (screen time); secondary outcomes: activity time; motivation for sedentary behavior and physical activity; health literacy, socioeconomic status, social support; waist circumference; weight; blood pressure; fasting blood glucose; triglycerides; HDL[b] cholesterol; total cholesterol; HbA_{1c}[c]; quality of life; login time and frequency; dropout reasons; satisfaction.

XSL•FO
RenderX

Intervention element	Intervention (final version)
Intervention content	

XSL•FO
RenderX

Intervention element	Intervention (final version)
	• Provide information via the platform on an overview of sedentary behavior, health risks, and benefits of interrupting sedentary behavior. • Reference content from sedentary guidelines of organizations such as WHO[d] via the platform. • Provide information via the platform on methods for interrupting sedentary behavior. • Provide information via the platform on sedentary behavior in different scenarios and its health risks, such as watching TV in entertainment scenarios, sedentary office work in work scenarios, and the potential increase in sedentary behavior with age, explaining the adverse health consequences these behaviors may cause. • Collaborate with rehabilitation physicians or other professionals to assess physical limitations (eg, restricted joint mobility, lumbar pain, and postoperative recovery), and identify solutions (eg, push customized actions based on health data) to break sedentary behavior, adding safety prompts such as “Stop if pain occurs.” • Help individuals accurately and repeatedly practice interrupting sedentary behavior via the platform by triggering a sedentary reminder every 30 minutes, practicing repeatedly until proficient in sedentary interruption skills. • Remind users to stand up and interrupt sedentary behavior based on their preferred sedentary interruption frequency setting (15 - 60 min) via the platform, synchronously sending notifications such as “You've been sitting for 30 minutes, get up and move!.” • Provide information via the platform on methods for interrupting sedentary behavior (eg, standing and walking) and their rationale, and instructions on setting up and using the app. • Provide short video exemplary cases via the platform of regularly interrupting sedentary behavior (eg, standing stretch during TV ad breaks, standing activity during long mahjong sessions and getting up for water during long sedentary work), demonstrating how to accurately and correctly interrupt sedentary behavior. • The platform monitors user sedentary behavior in real-time (including single-session sedentary duration and daily cumulative sedentary time) via smart wearable devices, automatically generates visual reports, and pushes personalized improvement suggestions. For example, monitoring shows your longest continuous sedentary period was 2 PM-4 PM. Current cumulative sedentary time has reached 7 hours (approaching the 8-h guideline limit). Suggestion: get up and move for 5 minutes immediately; Avoid exceeding 8 hours sedentary time for the rest of today. Automatically provides feedback upon detecting user activity: Great! You have been active for 5 minutes, please keep it up! • Push information via the platform on the benefits of interrupting sedentary behavior, emphasizing the necessity for users to receive training. • Remind individuals via the platform (eg, during postmeal rest time) to participate in training (eg, “You can learn standing stretches now”). • Invite a team including geriatric doctors, rehabilitation therapists, etc, to provide online question-and-answer sessions, offering professional advice and guidance to individuals on reducing sedentary behavior. • Users invite family, friends, or colleagues to interrupt sedentary behavior together via the platform, and encourage each other through the platform’s chat function; family members receive reminders and send encouraging messages. • The platform intelligently matches users with “partners with low sedentary behavior” based on user behavior data, and supports group challenge competitions (eg, daily cumulative sedentary or standing time PK). • Popularize available social support resources (eg, senior activity centers and interest groups) via the platform, explaining the positive role of participating in these social activities in reducing sedentary behavior, guiding them to proactively seek and use related resources. • Disseminate information about available environmental resources (eg, community fitness facilities) through the platform, explaining their positive role in reducing sedentary behavior, and guide individ-

XSL•FO
RenderX

Intervention element	Intervention (final version)
	uals to proactively access and use these resources. • Automatically record the type, time, and frequency of sedentary behavior via wearable devices. • Provide videos via the platform featuring health care professionals emphasizing the benefits of interrupting sedentary behavior. • Display the user's progress ranking for sedentary behavior duration within their friend circle via the platform's "Progress Leaderboard" function, motivating users to reduce sedentary behavior; noncompleters receive a "Try again tomorrow" encouragement card. • The platform intelligently pushes layered incentives based on user goal achievement (completers unlock badges such as "Sedentary Slayer" or "Star of Vitality"+voice congratulatory message; noncompleters receive a "Try again tomorrow" encouragement card) and share them on social media to receive likes and encouragement from friends. • Require individuals via the platform to use words such as "willing," "can," "commit," or "definitely can," "high priority" to confirm or reaffirm their commitment to starting, continuing, or restarting attempts to reduce sedentary behavior. • Sign a sedentary behavior commitment contract with individuals via the platform to ensure regular interruption of sedentary behavior. • Set sedentary behavior reduction goals (eg, "daily sedentary time ≤8 h") via the platform. • Support creation of "Sedentary Behavior Interruption Plans" via the platform, eg, "perform a 5-minute standing stretch at 10:00, 14:00, 16:00 each day," with reminders via the platform. • Generate weekly sedentary behavior data reports, comparing actual sedentary behavior time with goals. Users can view historical records via the platform's timeline function to assess goal achievement progress. • Introduce via the platform how reducing sedentary behavior can alleviate individuals' negative emotions.

[a]The final intervention items are derived by merging repeated items from the initial list. Item 1 is derived from Items 1.1 and 1.3; Item 2 is derived from Items 1.2, 1.4, and 3.9; Item 3 is derived from Items 1.6 and 3.3; and Item 4 is derived from Items 1.7 and 3.7. Items 5 - 8 are derived from Items 1.8, 1.5, 1.9, and 1.10, respectively. Item 9 is derived from Items 1.11 and 3.2, and Items 10 - 12 are derived from Items 1.12 - 1.14, respectively. Item 13 is derived from Item 2.1; Item 14 is derived from Items 2.2 and 3.17; Items 15 - 18 are derived from Items 2.3, 2.4, 2.6, and 3.1, respectively. Item 19 is derived from Items 3.4 and 3.8; Items 20 - 21 are derived from Items 3.5 - 3.6, respectively; Items 22 - 23 are derived from Items 3.11 and 3.10, respectively; Item 24 is derived from Items 2.5 and 3.12, and Items 25 - 27 are derived from Items 3.13, 3.14, and 3.16, respectively. Clinical biomarkers (eg, glycated hemoglobin and lipids) are included as candidate secondary outcomes for future definitive trials; their assessment will depend on feasibility and available resources.

[b]HDL: high-density lipoprotein.

[c]HbA_{1c}: glycated hemoglobin.

[d]WHO: World Health Organization.

The figure illustrates the iterative refinement process of the intervention items based on expert feedback in Round 1 and Round 2, showing how the initial 29 items were gradually revised to a final set of 27 items.

The revised framework of the mHealth intervention is illustrated in Figure 3.

The framework illustrates the key intervention components, including intervention functions, BCTs, and delivery format, mapped onto the COM-B dimensions.

XSL•FO
RenderX

Figure 3. Revised framework of the mobile health intervention for reducing sedentary behavior in older adults.

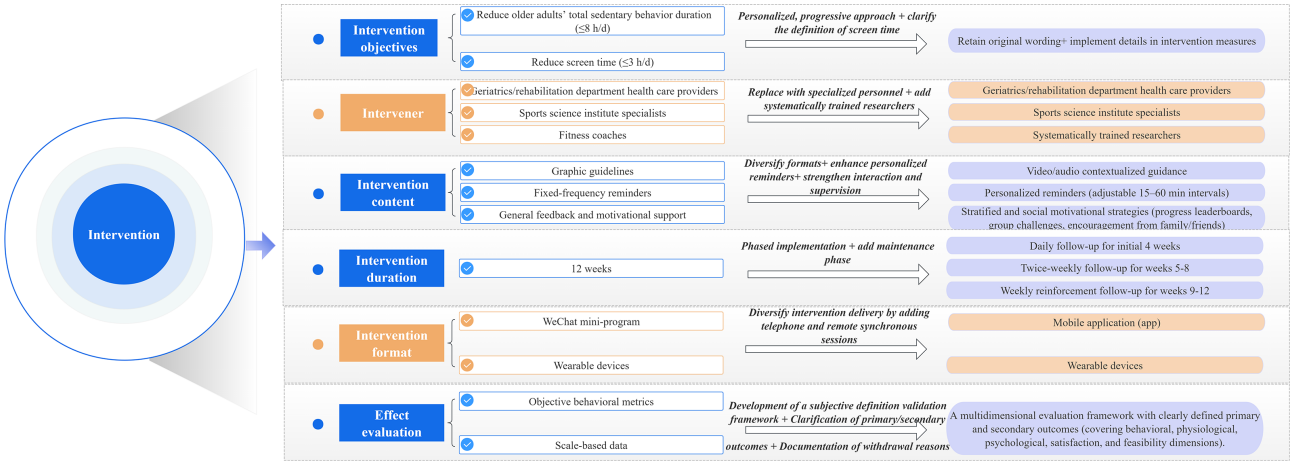

Discussion

Summary of Key Findings

To our knowledge, this study is the first to systematically integrate the BCW framework with the Delphi expert consensus method to develop an mHealth intervention for reducing sedentary behavior in older adults. The final intervention comprises 7 intervention functions and 20 BCTs, operationalized into 27 intervention items. These comprehensively address the 3 core dimensions of the COM-B model: capability, opportunity, and motivation, resulting in a theoretically driven, expert-validated mHealth intervention tailored to the unique needs of older adults. It should be noted that the preliminary development of the mHealth intervention for reducing sedentary behavior in older adults, guided by the 3-phase, 8-step BCW framework, is beyond the scope of this manuscript and will be detailed in a separate publication. This study instead focuses on the Delphi expert consultation process, through which the intervention components were refined and validated by multidisciplinary expert consensus. This validation was achieved through rigorous Delphi procedures with a panel of 16 multidisciplinary experts, all holding a bachelor's degree or higher (11 with doctoral degrees and 5 with master's degrees) and 87.5% (14/16) holding associate senior positions or higher, with an average of more than 15 years of professional experience. The 2 Delphi rounds achieved 100% response rates, with expert authority coefficients of 0.900 and 0.907, significantly exceeding the 0.70 threshold, demonstrating strong reliability and stability of expert judgments [13]. Statistically significant concordance (Kendall W=0.151, $P<.001$ and W=0.214, $P=.001$) across both rounds further validates the reliability of the consensus process. Expert feedback across the 2 rounds led to iterative refinements, including item additions, deletions, and content optimization, with suggestions primarily focused on enhancing intervention details and implementation feasibility. Collectively, these findings fulfill the study's objective of developing a theoretically grounded, expert-validated mHealth intervention using the Delphi method, establishing a robust foundation for subsequent empirical testing.

Comparison With Existing Literature

Our previous systematic review [6] identified several studies [15-17] that used the Jawbone Up app (Jawbone Inc) to intervene in sedentary behavior among older adults. However, this app was not specifically designed for older adults and has since been discontinued, which may limit its relevance for this population. Moreover, some of these studies [15,18] primarily focused on increasing physical activity to reduce sedentary behavior, rather than directly targeting sedentary behavior itself. As a result, they did not incorporate BCTs specifically intended to reduce sedentary behavior in older adults. Additionally, most of the mobile interventions for sedentary behavior in older adults, including those involving the Jawbone Up App, were pilot studies [15-19]. While they provided preliminary empirical evidence for sedentary behavior reduction, the small sample sizes and certain design limitations raise concerns about the generalizability and reliability of the results, which require further validation. Furthermore, some studies [17,20-22] lacked a systematic foundation in BCTs or behavioral theory, which contributed to suboptimal intervention outcomes.

Interpretation of Intervention Components Across COM-B Dimensions

In contrast, our mHealth intervention uses a more structured and theoretically grounded design by systematically linking identified barriers to the COM-B model and using this analysis to inform the a priori selection of BCTs. In the intervention design of this study, our research team proposed ultimately adopted 7 intervention functions and 20 specific BCTs, and invited experts to rate each technique on rationality and importance. Based on expert feedback, the principal findings across COM-B or TDF dimensions were as follows: at the level of psychological and physical capability, to address barriers such as older adults' lack of knowledge about sedentary behavior and limited ability to interrupt it, the research team used the intervention functions of education and enablement together with the BCTs "Information about health consequences" (BCT 5.1), "Credible source" (BCT 9.1), "Prompts or cues" (BCT 7.1), "Problem solving" (BCT 1.2), and "Behavioral practice or rehearsal" (BCT 8.1). Expert ratings indicated that these techniques were generally high on both rationality and importance. For the barrier of lacking methods to interrupt sedentary behavior, the Training function was used with "Instruction on how to perform the behavior" (BCT 4.1), "Demonstration of the behavior" (BCT 6.1), "Behavioral practice or rehearsal" (BCT 8.1), and "Feedback on behavior"

XSL•FO
RenderX

(BCT 2.2); experts considered these techniques crucial for building skills to reduce sedentary behavior among older adults. In addition, to mitigate low training priority or limited training time, “Information about health consequences” (BCT 5.1) under Education and “Prompts or cues” (BCT 7.1) also received high endorsement.

At the level of social and physical opportunity, for insufficient social support and environmental constraints, the research team adopted enablement and education, proposing “Social support (unspecified)” (BCT 3.1), “Social support (emotional)” (BCT 3.3), “Information about social and environmental consequences” (BCT 5.3) and “Restructuring the social environment” (BCT 12.2). Expert ratings showed that these techniques were judged highly rational and important for strengthening older adults’ social support and improving the social environment to reduce sedentary behavior.

At the level of reflective and automatic motivation, the research team used modeling, education, persuasion, and incentivization, along with “Self-monitoring of behavior” (BCT 2.3) and “Demonstration of the behavior” (BCT 6.1) to help older adults integrate reduced sedentary behavior into their role identity and enhance their confidence in changing their behavior. In addition, to explicitly strengthen older adults’ intentions and goal-setting for behavior change, the research team used “Credible source” (BCT 9.1), “Commitment” (BCT 1.9), “Social reward” (BCT 10.4), and “Goal setting (behavior)” (BCT 1.1). These techniques received high expert endorsements for both rationality and importance. For automatic motivation, incentivization was adopted, and experts particularly highlighted “Social reward” (BCT 10.4), “Information about emotional consequences” (BCT 5.6), and “Social support (emotional)” (BCT 3.3) as effective for eliciting positive emotional experiences and intrinsic motivation to reduce sedentary behavior among older adults. Overall, the BCTs selected by the research team received high ratings for rationality and importance.

Significance and Innovations

The practical implications of this intervention are manifold, featuring 3 key innovations: intelligent monitoring with real-time feedback for precise behavior modification, a multitiered social support system encompassing professional guidance and family involvement, and motivational architectures combining gamification elements with goal-setting strategies. These components collectively address the 3 essential conditions for behavior change—capability, opportunity, and motivation—as outlined in the COM-B model, providing a comprehensive approach rarely seen in existing interventions for older adults. The intervention’s design specifically tackles implementation challenges through features such as family engagement tools and culturally adapted activity suggestions, which enhance both feasibility and sustainability. This represents a promising step from traditional, clinic-based interventions to more sustainable, technology-enabled community-embedded approaches that align with modern health care delivery models.

Moreover, this study developed an mHealth intervention focusing on the 3 core dimensions of capability, opportunity, and motivation. The intervention’s unique approach lies in its comprehensive and targeted strategy, which not only enhances older adults’ knowledge and skills but also provides appropriate environmental support and strengthens motivation for behavior change. Through 2 rounds of expert feedback and optimization, the intervention content was continuously refined to ensure its effectiveness and feasibility. In terms of capability, the intervention uniquely improves older adults’ awareness of sedentary behavior through educational content while also providing practical guidance to help them master specific behaviors to interrupt sedentary activity. The program educates users about the health risks of prolonged sedentary behavior, such as cardiovascular diseases, metabolic disorders, and cognitive decline, motivating them to change their behavior. Additionally, the intervention offers clear, step-by-step instructions on how to break sedentary behavior, including standing, walking, and stretching. The intervention also includes a flexible sedentary interruption reminder system, allowing users to customize the frequency and timing of reminders to meet their individual needs, improving the intervention’s adaptability and feasibility. Expert feedback in the first round focused on improving the intervention’s operability and clarity, with suggestions to convert text-based materials into video demonstrations with voice guidance, significantly improving accessibility. In the second round, experts recommended incorporating real-time behavioral feedback via wearable devices, helping users detect and adjust their sedentary behavior, thereby enhancing the intervention’s real-time and personalized nature. Regarding opportunity, the intervention creates a supportive environment by promoting both social support and optimizing the physical environment. It encourages older adults to involve family, friends, or colleagues, thus increasing social support and overcoming feelings of isolation and lack of motivation. Feedback from the experts in the first round emphasized enhancing social support by encouraging older adults to invite family, friends, or colleagues to participate, which would foster a sense of community. In the second round, experts suggested adding features such as “sedentary duration challenges” to promote friendly competition, further boosting engagement and long-term adherence. In terms of motivation, the intervention combines intrinsic and extrinsic motivation strategies, including personalized goal-setting, rewards, and social sharing, to encourage continuous behavior change. It guides older adults to set specific goals, such as reducing daily sedentary time, and encourages them to share their progress on social platforms, involving family and friends. Additionally, external motivation is provided through achievement badges and leaderboards, motivating users to push themselves and overcome barriers to behavior change. Expert feedback on motivation focused on enhancing personalization and social interaction. In the first round, experts suggested showcasing user progress and incorporating social incentives, such as progress leaderboards and social rewards. The second round further optimized this mechanism by introducing encouragement cards for users who did not meet their goals and unlocking additional rewards for those who achieved them.

The innovation of this study lies in its comprehensive, multidimensional approach to reducing sedentary behavior in older adults. By integrating the BCW framework, the intervention uniquely addresses capability, opportunity, and

XSL•FO
RenderX

motivation to create a personalized and adaptable solution for behavior change. It combines knowledge-building, social support, and real-time feedback, allowing older adults to understand the health risks of sedentary behavior and actively engage in strategies to interrupt it. The use of wearable devices for monitoring and providing personalized feedback adds a real-time, data-driven element to the intervention, enhancing its relevance and effectiveness. Additionally, the focus on social interaction features such as “sedentary duration challenges” promotes long-term engagement and adherence. The inclusion of motivational strategies, such as personalized goal-setting, achievement badges, and social rewards, strengthens both intrinsic and extrinsic motivation, ensuring continued participation and behavior change. Future research should focus on conducting large-scale randomized controlled trials to rigorously assess the long-term efficacy of the intervention in real-world settings, refining age-appropriate design elements to improve usability for older adults, and exploring how technology can be further tailored to meet the diverse physical and cognitive needs of this population.

Limitations

This study has several limitations. First, although the Delphi panel included experts from multiple relevant disciplines, the representativeness of the expert panel may be limited, and the perspectives obtained may not fully reflect all populations involved in sedentary behavior interventions for older adults. Second, this study focused on the development of the mHealth intervention through expert consensus and did not include end user involvement or empirical evaluation at this stage. Future research is needed to pilot test the intervention and examine its feasibility, usability, and effectiveness through empirical studies, such as randomized controlled trials.

Conclusion

This study, grounded in the BCW framework, systematically integrates mixed qualitative and quantitative research methods, mHealth technology, and 2 rounds of expert consultation to develop an mHealth intervention targeting sedentary behavior in older adults. Comprising 27 intervention items, the intervention includes key elements such as goals, content, duration, and delivery format, focusing on enhancing older adults’ capabilities, creating opportunities, and motivating behavior change. This intervention has the potential to address the dual needs of older adults and health care providers in managing sedentary behavior. Future research should focus on validating its effectiveness, optimizing its application, and assessing its scalability across diverse populations to maximize its impact.

Acknowledgments

No generative artificial intelligence (AI) tools were used in the writing of this manuscript.

Funding

This study was supported by "Double First-Class" Construction Specialized Discipline Project at Zhejiang University (HL202409) and Jinhua Science, Technology Bureau, China (2021-3-010) and the Zhejiang University Doctoral New Star Program (2023097).

Data Availability

The datasets used in this study are available from the corresponding author upon reasonable request, subject to privacy and ethical restrictions that prevent public access.

Authors' Contributions

SC contributed to conceptualization, methodology, data collection, data analysis, and writing. XN contributed to data organization. JY and YL contributed to methodology. LY contributed to conceptualization and writing. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Demographic characteristics of the research team and consulting experts.
[DOCX File, 16 KB - jmir_v28i1e83302_app1.docx]

Multimedia Appendix 2
Expert consultation feedback and revisions in 2 rounds.
[XLSX File, 25 KB - jmir_v28i1e83302_app2.xlsx]

Multimedia Appendix 3
Mapping of Capability, Opportunity, Motivation–Behavior model, Theoretical Domains Framework, intervention functions, behavior change techniques, and sedentary behavior intervention content with expert consultation results.

XSL•FO
RenderX

[DOCX File, 40 KB - jmir_v28i1e83302_app3.docx]

Multimedia Appendix 4
Comparison of item revisions in 2 rounds.
[DOCX File, 22 KB - jmir_v28i1e83302_app4.docx]

Multimedia Appendix 5
Comparison of preliminary and finalized versions of the mHealth intervention for reducing sedentary behavior in older adults in 2 rounds.
[DOCX File, 20 KB - jmir_v28i1e83302_app5.docx]

Checklist 1
CREDES guidelines.
[DOCX File, 27 KB - jmir_v28i1e83302_app6.docx]

References

1. Tremblay MS, Aubert S, Barnes JD, et al. Sedentary Behavior Research Network (SBRN) - Terminology Consensus Project process and outcome. Int J Behav Nutr Phys Act 2017 Jun 10;14(1):75. [doi: 10.1186/s12966-017-0525-8] [Medline: 28599680]
2. Gogniat MA, Khan OA, Li J, et al. Increased sedentary behavior is associated with neurodegeneration and worse cognition in older adults over a 7-year period despite high levels of physical activity. Alzheimers Dement 2025 May;21(5):e70157. [doi: 10.1002/alz.70157] [Medline: 40357887]
3. Chen S, Yang L, Yang Y, et al. Sedentary behavior, physical activity, sleep duration and obesity risk: Mendelian randomization study. PLoS One 2024;19(3):e0300074. [doi: 10.1371/journal.pone.0300074]
4. Jingjie W, Yang L, Jing Y, Ran L, Yiqing X, Zhou N. Sedentary time and its association with risk of cardiovascular diseases in adults: an updated systematic review and meta-analysis of observational studies. BMC Public Health 2022 Feb 12;22(1):286. [doi: 10.1186/s12889-022-12728-6] [Medline: 35148747]
5. Hecht FM. Sedentary behavior and dementia. JAMA 2024 Feb 6;331(5):442-443. [doi: 10.1001/jama.2023.25172] [Medline: 38319336]
6. Chen S, Wang C, Ko A, et al. Effectiveness of mobile health interventions for reducing sitting time in older adults: systematic review and meta-analysis. J Med Internet Res 2025 May 8;27:e60889. [doi: 10.2196/60889] [Medline: 40340833]
7. Chen S, Wu C, Ewing Garber C, Stults-Kolehmainen M, Giovannucci E, Yang L. Understanding the effectiveness of mHealth interventions incorporating behaviour change techniques in reducing sitting time in older adults: a protocol for a systematic review and meta-analysis. BMJ Open 2025 Oct;15(10):e100611. [doi: 10.1136/bmjopen-2025-100611]
8. Chen S, Yang K, Ko A, Giovannucci E, Stults-Kolehmainen M, Yang L. Facilitators and barriers of reducing sedentary behavior in sedentary and non-sedentary older adults: a descriptive qualitative study based on the COM-B model and TDF. BMC Public Health 2025;25(1):2472. [doi: 10.1186/s12889-025-23613-3]
9. Jünger S, Payne SA, Brine J, Radbruch L, Brearley SG. Guidance on Conducting and REporting DElphi Studies (CREDES) in palliative care: recommendations based on a methodological systematic review. Palliat Med 2017 Sep;31(8):684-706. [doi: 10.1177/0269216317690685] [Medline: 28190381]
10. Chen S, Lin K, Ko A, et al. Factors associated with sedentary behavior reduction in older adults based on the capability, opportunity, motivation, and behavior model: a multicenter cross-sectional study. J Am Med Dir Assoc 2026 Jan;27(1):106007. [doi: 10.1016/j.jamda.2025.106007] [Medline: 41274654]
11. Ross R, Chaput JP, Giangregorio LM, et al. Canadian 24-Hour Movement Guidelines for adults aged 18-64 years and adults aged 65 years or older: an integration of physical activity, sedentary behaviour, and sleep. Appl Physiol Nutr Metab 2020 Oct;45(10 (Suppl. 2)):S57-S102. [doi: 10.1139/apnm-2020-0467] [Medline: 33054332]
12. Bull FC, Al-Ansari SS, Biddle S, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. Br J Sports Med 2020 Dec;54(24):1451-1462. [doi: 10.1136/bjsports-2020-102955] [Medline: 33239350]
13. Varndell W, Fry M, Lutze M, Elliott D. Use of the Delphi method to generate guidance in emergency nursing practice: a systematic review. Int Emerg Nurs 2021 May;56:100867. [doi: 10.1016/j.ienj.2020.100867] [Medline: 32238322]
14. Chen S, Cao M, Zhang J, Yang L, Xu X, Zhang X. Development of the health literacy assessment instrument for chronic pain patients: a Delphi study. Nurs Open 2023 Apr;10(4):2192-2202. [doi: 10.1002/nop2.1468] [Medline: 36564937]
15. Lyons EJ, Swartz MC, Lewis ZH, Martinez E, Jennings K. Feasibility and acceptability of a wearable technology physical activity intervention with telephone counseling for mid-aged and older adults: a randomized controlled pilot trial. JMIR Mhealth Uhealth 2017 Mar 6;5(3):e28. [doi: 10.2196/mhealth.6967] [Medline: 28264796]
16. Blair CK, Harding E, Wiggins C, et al. A home-based mobile health intervention to replace sedentary time with light physical activity in older cancer survivors: randomized controlled pilot trial. JMIR Cancer 2021 Apr 13;7(2):e18819. [doi: 10.2196/18819] [Medline: 33847588]

XSL•FO
RenderX

17. Rosenberg DE, Kadokura E, Morris ME, Renz A, Vilardaga RM. Application of N-of-1 experiments to test the efficacy of inactivity alert features in fitness trackers to increase breaks from sitting in older adults. Methods Inf Med 2017 Aug 16;56(5):427-436. [doi: 10.3414/ME16-02-0043] [Medline: 28816343]
18. Li J, Hodgson N, Lyons MM, Chen KC, Yu F, Gooneratne NS. A personalized behavioral intervention implementing mHealth technologies for older adults: a pilot feasibility study. Geriatr Nurs (Lond) 2020 May;41(3):313-319. [doi: 10.1016/j.gerinurse.2019.11.011]
19. Ashe MC, Winters M, Hoppmann CA, et al. “Not just another walking program”: Everyday Activity Supports You (EASY) model-a randomized pilot study for a parallel randomized controlled trial. Pilot Feasibility Stud 2015;1:4. [doi: 10.1186/2055-5784-1-4] [Medline: 27175291]
20. Pinto BM, Kindred M, Franco R, Simmons V, Hardin J. A “novel” multi-component approach to promote physical activity among older cancer survivors: a pilot randomized controlled trial. Acta Oncol 2021 Aug;60(8):968-975. [doi: 10.1080/0284186X.2021.1896032] [Medline: 33689553]
21. Recio-Rodríguez JI, Gonzalez-Sanchez S, Tamayo-Morales O, et al. Changes in lifestyles, cognitive impairment, quality of life and activity day living after combined use of smartphone and smartband technology: a randomized clinical trial (EVIDENT-Age study). BMC Geriatr 2022 Oct 6;22(1):782. [doi: 10.1186/s12877-022-03487-5] [Medline: 36203135]
22. Recio-Rodríguez JI, Lugones-Sanchez C, Agudo-Conde C, et al. Combined use of smartphone and smartband technology in the improvement of lifestyles in the adult population over 65 years: study protocol for a randomized clinical trial (EVIDENT-Age study). BMC Geriatr 2019 Jan 23;19(1):19. [doi: 10.1186/s12877-019-1037-y] [Medline: 30674284]

Abbreviations

APEASE: Affordability, Practicability, Effectiveness and cost-effectiveness, Acceptability, Side-effects/safety, and Equity
BCT: behavior change technique
BCW: Behavior Change Wheel
COM-B: Capability, Opportunity, Motivation–Behavior
CREDES: Conducting and Reporting of Delphi Studies
CV: coefficient of variation
mHealth: mobile health
TDF: Theoretical Domains Framework
WHO: World Health Organization

Edited by A Stone; submitted 31.Aug.2025; peer-reviewed by A Brooks, M Guo, YP Lin; revised version received 16.Apr.2026; accepted 30.Apr.2026; published 11.Jun.2026.

Please cite as:
Chen S, Nie X, Yang J, Li Y, Yang L
Development of an mHealth Intervention for Reducing Sedentary Behavior in Older Adults: Delphi Study
J Med Internet Res 2026;28:e83302
URL: https://www.jmir.org/2026/1/e83302
doi:10.2196/83302

© Siqing Chen, Xiaofei Nie, Jiekai Yang, Yaqin Li, Lili Yang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring Vaping Cessation App Use Among Youth: Qualitative Descriptive Study

Karlee Fonteyne[1], MSN, RN; Mischa Taylor[1], BSc, MACE; Roula Nawara[1], RN, BSN; Danielle Rodberg[1], RN, BSN; Tina Afshar[2], MS; Laura Struik[1], PhD, RN

[1]School of Nursing, Faculty of Health and Social Development, University of British Columbia, 1147 Research Road, Kelowna, BC, Canada

[2]School of Population and Public Health, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

Corresponding Author:
Laura Struik, PhD, RN
School of Nursing, Faculty of Health and Social Development, University of British Columbia, 1147 Research Road, Kelowna, BC, Canada

Abstract

Background: Youth vaping has become a significant public health concern, with high rates of initiation, nicotine dependence, and a tendency to underestimate vaping-related harms. Although many youth attempt to quit, most do so without formal support, and few report leveraging cessation apps. Mobile health tools have potential as accessible youth-oriented supports; yet, little is known about youth experiences with these interventions.

Objective: This study aims to explore Canadian youth perceptions and experiences of vaping cessation mobile apps, focusing on preferences, perceived utility, and recommendations for improvement.

Methods: A qualitative descriptive design guided by a constructivist paradigm was used. Overall, 17 semistructured interviews were conducted with youth aged 16‑24 years from British Columbia, Ontario, and Quebec in fall 2022. Eligible participants had current or past vaping experience, had attempted or wanted to quit, and had used a cessation app. Data were transcribed verbatim and analyzed inductively using reflexive thematic analysis, with reflexivity supported through team-based coding and iterative discussions.

Results: Three themes described youth experiences. “Quitting on my Own Terms” reflected the importance of individual agency and commitment to quitting. While there were varied views on apps supporting their agency with cessation, the youth agreed that intrinsic motivation was the most critical element in their cessation journey. “Parallel Experiences- Differing Views on Community” captured ambivalence toward in-app social features: while some valued peer support, others preferred to quit privately or used broader online platforms, like Reddit. “Pocket-sized Partners in Cessation” highlighted the usefulness of quick access to features, such as progress tracking, motivational reinforcement, and gamification, as they engaged in the various stages of quitting (from contemplation through to commitment). However, participants stressed that the convenience of an app was not enough. Youth emphasized the need for apps that feel authentic, emotionally resonate, and are capable of balancing independence with opportunities for connection to social support.

Conclusions: Youth want cessation apps that function as companions to help them leverage their motivation and efforts to quit, rather than as passive trackers. Effective app-based interventions should validate autonomy, account for nonlinear cessation journeys and relapses, and integrate supportive, interactive, and youth-friendly features. By addressing the preferences of youth in the design of vaping cessation apps, digital cessation tools can move beyond convenience to become meaningful companions in youths’ efforts to quit vaping.

(J Med Internet Res 2026;28:e85778) doi:10.2196/85778

KEYWORDS

youth vaping; vaping; mobile health; digital interventions; mobile apps; qualitative description; public health

Introduction

Background

The dramatic surge in e-cigarette, or vaping, use among young people has created a pressing public health crisis across North America, demanding urgent attention and innovative intervention. Data from the United States indicate that in 2023, 10.3% of young adults aged 18‑20 years and 15.5% aged 20‑24 years disclosed vaping regularly [1], while a more recent statistic from the 2024 National Youth Tobacco Survey highlighted that 7.8% of high school students (grades 9‑12) reported vaping in the past 30 days [2]. These youth are drawn to vaping for a variety of reasons, including stress relief,

XSL•FO
RenderX

enjoyment, curiosity, social influences, and the perception of vaping as a safer alternative to smoking [3-5].

In both the United States and Canada, vaping products (including those with and without nicotine) are primarily sold "over the counter" at retail outlets such as convenience stores, gas stations, and specialty vape shops. Unlike certain international jurisdictions, North American regulations do not require a doctor's prescription or pharmacy-based dispensing [6,7]. Instead, access is governed by minimum age laws, 21 years at the federal level in the United States [6], and 18 or 19 years in most Canadian provinces [7], which are enforced through mandatory point-of-sale photo identification checks.

Despite growing evidence of vaping-related health risks, such as respiratory issues, nicotine addiction, exposure to harmful chemicals, and a correlation with negative mental health effects, many youth continue to underestimate vaping-related harms [8-10]. This misperception, combined with the highly addictive nature of nicotine and the powerful influence of peer norms, contributes to patterns of sustained use and dependence. Although many youth express interest in quitting, they often encounter significant obstacles, including limited access to tailored cessation resources, a lack of formal guidance, and uncertainty about where to begin. As a result, supporting youth in their efforts to quit vaping requires interventions that are not only evidence-based, but also accessible, engaging, and responsive to the unique challenges they face [3,5,11-13].

Many youth looking to quit vaping nicotine do so through unassisted strategies such as quitting "cold turkey" or gradually reducing nicotine levels, rather than seeking cessation support or leveraging available resources like quit lines, mobile apps, or SMS text messaging programs [13-15]. Recognizing the urgent need for youth-centered cessation support, public health agencies, researchers, and technology developers have increasingly turned to apps as scalable tools for vaping cessation [16,17]. With 96% of Canadian youth aged 15‑24 years reporting smartphone use and nearly half checking their phones every 15 minutes, mobile health interventions are well-positioned to deliver accessible, real-time, and personalized support [18]. Although youth demonstrate interest in health-related apps, actual engagement with vaping cessation apps remains low. For instance, only 5.9% of current youth vapers who attempted to quit in the past year reported using an app or SMS text messaging program, while 63.7% attempted to quit unassisted [19].

A recent meta-analysis found that digital interventions modestly increased the odds of short-term abstinence but stressed the need for more rigorous, youth-specific research in this space [20]. Further, a review by McKay et al [21] concluded that most vaping cessation apps offer only moderate functionality and often lack critical behavior change elements, such as goal-setting and personalized feedback, which are particularly important for youth engagement. Sanchez et al [22] found that most apps were repurposed from smoking cessation programs and failed to reflect the distinct needs, motivations, and usage patterns of youth who vape. In addition, interactive and gamification-based features, such as progress tracking, rewards, and milestone-based feedback, have been identified as promising strategies to enhance engagement and motivation among youth. Yet, these elements remain inconsistently implemented and rarely evaluated within vaping cessation apps [23-25]. Compounding these design limitations, most vaping cessation apps do not report user metrics, such as downloads or demographic reach, making it difficult to assess their real-world impact. In contrast, smoking cessation apps frequently demonstrate broad uptake, with hundreds of thousands of downloads and targeted promotional campaigns [26]. This lack of comparable data underscores a significant gap: while mobile apps offer high potential for engaging youth in vaping cessation, there is a limited understanding of whether these apps are being used, how they are perceived, and whether they meet the needs of their intended audience [27]. User-centered research is essential to fill this gap and inform the development of more effective, accessible, and relevant digital cessation tools for youth.

Study Aim and Significance

This qualitative study aims to explore youth users' perceptions and experiences with using vaping cessation mobile apps. Specifically, we explored how youth engaged with the app, their perceptions of app use, and areas for improvement. These insights are crucial for informing the development of more effective, relevant, and engaging vaping cessation interventions that align with youth's behavioral, motivational, and technological needs. By centering youth voices, this study contributes to a growing body of research focused on tailoring digital health interventions to vulnerable populations. The findings have significant implications for public health efforts to reduce vaping among youth and young adults, ultimately supporting the promotion of healthier behaviors and long-term well-being.

Methods

Study Design

Drawing on its alignment with a constructivist worldview where subjective realities exist and can be known through subjective accounts of individual experiences [28], we used qualitative description to explore this topic. Qualitative description is a methodology useful for exploring and describing phenomena that remain understudied, while staying close to the data in its interpretation [29].

Sampling and Recruitment

A purposive sampling strategy was used to recruit participants whose lived experiences could meaningfully inform the research questions. Eligible participants were youth aged 16 to 24 who met the following criteria: (1) currently vape or have vaped in the past, (2) are currently or were previously motivated to quit, (3) have used or are using a vaping cessation app, (4) reside in Canada, and (5) speak English fluently.

Recruitment was conducted via social media platforms, specifically Instagram (Meta) and Facebook (Meta), where a digital advertisement invited youth to share their experiences with quitting or attempting to quit vaping in a study led by researchers at the University of British Columbia (UBC). The ad included a QR code and a hyperlink to a brief eligibility survey with 5 "yes" or "no" questions. The research team

XSL•FO
RenderX

reviewed responses and compiled a list of eligible participants, none of whom were known to the team prior to the study.

Efforts were made to ensure gender balance through a recruitment ad that used gender-neutral language and imagery. During recruitment, we monitored the gender representation of the incoming sample to assess whether we were meeting our gender-balancing target. Ultimately, the sample consisted of 10 self-identified women and 7 self-identified men. Eligible participants were contacted via email with an invitation to participate, a demographic questionnaire, and a request for their availability. Interviews were scheduled after written consent was obtained. At the start of each interview, informed and ongoing consent was reviewed with participants.

Data Collection

The interview guide was developed collaboratively by the research team in consultation with experts in the field and a youth co-researcher (Multimedia Appendix 1). It was circulated among team members for review and revisions before finalization. Interview questions for this study explored participants' experiences of the app in relation to their quitting goals, such as "What prompted you to download a vaping app?" "What features on the app kept you coming back?" and "Which social support features did the app offer, and what was your experience of using them?" This study draws on data collected from September to October 2022.

Overall, 17 semistructured interviews were conducted online via UBC-licensed Zoom in September and October 2022, with participant consent. The research team was based in British Columbia, Canada, while participants were located across British Columbia, Ontario, and Quebec. Interviews were conducted by 1 of 4 trained female research assistants and lasted 30 - 60 minutes. Interviewers adhered to the semistructured guide while allowing participants to elaborate on their experiences. Recruitment and interviewing occurred concurrently, with recruitment concluding once the team determined that data saturation had been reached and no new insights relevant to the research question were emerging.

Data Analysis

Following the completion of all interviews, deidentified audio recordings were uploaded to Otter.ai for transcription. Each transcript was reviewed by 2 research assistants (RN and DR) to verify transcription accuracy and to correct any discrepancies. To facilitate initial familiarization with the data, RN and DR organized the interview content by mapping responses to their corresponding interview questions using NVivo 14 (Lumivero). The document was subsequently circulated among the research team to provide an initial overview of the content derived from the interviews.

An inductive thematic approach was used to interpret the data. Following an iterative, collaborative process, the research coordinator (MT) and research assistants (RN and KF) engaged in reflexive thematic analysis as outlined by Braun and Clarke [30-32]. This approach adhered to the 6 foundational phases of thematic analysis originally described by Braun and Clarke [32]: familiarization with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the final report. However, consistent with reflexive thematic analysis principles, the analysis was conducted with greater flexibility and interpretive depth. Rather than treating themes as objectively "found" within the data, the team actively constructed themes through an ongoing, reflexive engagement with the material [30-32].

To support a meaningful exploration of participants' experiences beyond the constraints of the interview questions and allow for true inductive analysis, the research team conducted initial coding directly from the original participant transcripts rather than the NVivo document that organized data by interview question. This step was conducted using Microsoft Excel due to compatibility issues between NVivo 14 and the diverse operating systems (PC and macOS) used by team members. MT and RN developed an initial coding framework by inductively analyzing the first 4 transcripts, which was then shared with other team members for feedback. Following an iterative process, as new codes emerged during analysis of the remaining transcripts, the coding framework was updated and circulated among the team for discussion. Initial codes alongside the corresponding extracted participant quotes were compiled in a master Excel document. Research team members independently reviewed the master Excel document to identify emerging patterns and conceptual throughlines that could form overarching themes. KF and RN then collaboratively synthesized these individual analyses, identifying cross-cutting patterns and refining them into consolidated themes. Members of the research team regularly met to review and refine these themes into findings. To ensure rigor and trustworthiness, the team maintained detailed audit trails, engaged in regular analytic discussions to question interpretations and surface assumptions, and practiced continuous reflexivity throughout the coding and theme development process [32]. This reflexive approach enabled a richer and more nuanced interpretation of participants' experiences.

Ethical Considerations

This study was granted ethics approval by the UBC Okanagan Behavioral Research Ethics Board (#H22-01408) in the summer of 2022. The recruitment process and all interview sessions occurred concurrently in the fall of 2022. Informed consent was obtained from all participants, which was voluntary and revisited throughout the research process. Participants were provided with information about the study and a digital consent form emphasizing voluntary participation, the study's purpose, potential risks, and the right to withdraw at any time without consequences. Given that the legal age to provide informed consent is 16 in Canada, parental consent was not required for this sample. Youth were provided with an email address to contact a researcher with additional questions or to request more information about the study. Each participant received a CAD $50 (US $36) e-gift card for their time and contributions.

The study implemented a comprehensive data management plan to ensure the secure handling, storage, and organization of research data. All data collected, including audio recordings, transcripts, and questionnaires, were stored on password-protected UBC servers in Canada and accessible only to the research team. During data collection, audio recordings

XSL•FO
RenderX

from interviews were securely uploaded to a cloud-based storage system that complied with UBC's institutional data security policies. These recordings were transcribed verbatim, with all personally identifiable information removed to maintain participant confidentiality. Participants were assigned a participant ID# (including gender and age), and only this anonymized identifier was included in the transcription, data files, and interview notes. Transcripts and associated metadata were organized and stored in NVivo 14, a qualitative data analysis software platform that provides secure data coding and retrieval tools. After the study concludes, data will be retained for 5 years in accordance with ethical guidelines, after which all files will be securely deleted. Participants were informed of these measures during the consent process to ensure transparency.

Results

Participant Characteristics

The final sample comprised 17 participants, the majority of whom identified as women (10/17, 58.8%), with the remainder identifying as men (7/17, 41.2%; Table 1). Most participants were aged 18‑24 years (14/17, 82.4%), with a smaller proportion aged 16‑17 years (3/17, 17.6%). Participants were primarily from British Columbia (7/17, 41.2%) and Ontario (6/17, 35.3%), with fewer from Quebec (4/17, 23.5%). Vaping history varied, though the largest proportion reported vaping for 4‑5 years (7/17, 41.2%), followed by less than 2 years (6/17, 35.3%). Most participants reported vaping several times per day (11/17, 64.7%), and a substantial majority indicated they were currently trying to quit (13/17, 76.5%). Nearly half (8/17, 47%) had attempted to quit 2‑5 times, while 35.3% (6/17) reported a single quit attempt. Social exposure to vaping was high, with most participants indicating that some (7/17, 41.2%) or most (8/17, 47%) of their friends vape.

XSL•FO
RenderX

Table . Demographic table.

Demographic and Categories	Values, n (%)
Sex (Gender)	
Female (woman)	10 (58.8)
Male (man)	7 (41.2)
Age range (years)	
16 - 17	3 (17.6)
18 - 24	14 (82.4)
Ethnicity	
Asian	4 (23.5)
Non-Hispanic White	6 (35.3)
Black	4 (23.5)
Latino	1 (5.9)
First Nation	1 (5.9)
White and Asian	1 (5.9)
Province of residence	
British Columbia	7 (41.2)
Ontario	6 (35.3)
Quebec	4 (23.5)
Highest education	
Less than high school	2 (11.8)
High school diploma	8 (47)
College or trades	6 (35.3)
University degree	1 (5.9)
Length of time vaping nicotine in years	
<2	6 (35.3)
2 - 3	2 (11.8)
4 - 5	7 (41.2)
≥6	2 (11.8)
Currently trying to quit	
Yes	13 (76.5)
No	4 (23.5)
Frequency vaping	
Several times/day	11 (64.7)
Once/day	3 (17.6)
A few times/week	3 (17.6)
≤ Once per week	0 (0)
Quit attempts	
Once	6 (35.3)
2 - 5 times	8 (47)
6 - 10 times	2 (11.8)
>10 times	1 (5.9)
Number of friends who vape	
None of them	0 (0)

XSL•FO
RenderX

Demographic and Categories	Values, n (%)
Some of them	7 (41.2)
Most of them	8 (47)
All of them	2 (11.8)

Findings revealed that youth experiences with vaping cessation apps were shaped by individual motivations, preferences, and perceptions of support. Three overarching themes emerged. "Quitting on My Own Terms" reflected the importance of autonomy and self-motivation, with youth expressing varied views on whether apps enhanced or undermined personal responsibility. "Parallel Experiences- Differing Views on Community" captured divergent perspectives on social support features; while some saw value in connecting with others, many preferred to quit independently or questioned the usefulness of app-based communities. Finally, "Pocket-Sized Partners in Cessation" highlighted the practical and motivational support apps provided across the quitting journey, though the subtheme "Caveat: Convenience Alone Is Not Enough" emphasized that basic features and design choices often limited their impact. These themes reflect how youth users perceive vaping cessation mobile apps, as well as their experiences with using them. These themes were shared across the sample, with no gender nuances detected between women and men.

Theme 1: Quitting on My Own Terms

The Tension Between Individual Responsibility and Leveraging App-Based Support

Participants commonly described quitting vaping as an intensely individualized process; yet, there was notable variability in how they interpreted what it meant to "quit solo." For some, using an app did not conflict with their desire to quit independently. These participants viewed the app as a nonintrusive aid and as a resource that helped them maintain accountability and structure without undermining their personal agency. As one participant explained:

> *I've tried everything that I can do myself separately. I don't try to rely on external help [...] it's an app helping it's not anyone in person.* [Participant #10, Male]

Another participant valued app-based support for being nonjudgmental:

> *[it's a] guilt free approach," appreciating "the fact that it doesn't judge.* [Participant #19, Male]

For some, the app's neutrality and lack of judgment made it a uniquely safe space to seek support without compromising their sense of independence. In these cases, apps served as a bridge between unsupported quitting and fully assisted interventions, preserving the autonomy that many youth valued while still offering practical guidance.

Conversely, other participants saw any reliance on external tools, including apps, as diminishing the self-responsibility they felt was essential to quitting. For these individuals, using an app was seen as inconsistent with the values of independence and self-control. One participant conveyed this sentiment clearly:

> *I feel like people need to have more self-responsibility, especially if you're an adult.* [Participant #7, Female]

However, this participant acknowledged that apps may be more suitable for other age groups, continuing, "[apps may be] really good for teens." Participants described their desire to quit independently, with some preferring greater or lesser degrees of app-based support.

Motivation as the Fuel: The Role of Personal Commitment

Across interviews, a key message emerged: self-motivation was the fuel in driving the cessation process. Participants frequently described how vaping cessation apps required a high degree of intrinsic motivation, and that their effectiveness was directly tied to the user's personal readiness and commitment to quit. While this was viewed positively by those who appreciated that the app reinforced their self-driven efforts, others saw it as a potential shortcoming. For youth who were not fully committed, the app alone was unlikely to spark or sustain change. As one participant put it:

> *If anyone's like, willing to quit, like an app is definitely […] your go to, in the beginning, I feel like it could be really helpful and like, self-encouragement, because like, although like the app is there to help you, you kind of have to be self-motivated.* [Participant #9, Female]

According to this participant, apps can provide helpful structure early on but must align with an existing desire to quit to be effective.

Apps were often described as tools that facilitate but cannot initiate change. They offered encouragement, goal tracking, and health tips, but these features only resonated when a user was genuinely ready to take action. Many participants emphasized that true success in quitting required deep internal resolve, regardless of external supports. For these youth, a central condition for cessation was personal determination, not the tools used:

> *I think if you're really going to quit it as it has to come from, you know, you personally, it's you, you have to take that decision [..] not to fall back into your addiction.* [Participant #4, Female]

As another participant succinctly explained:

> *I do think it can help out a lot. But only if you're serious about it.* [(Participant #2, Male]

For these participants, apps were considered a supportive tool to accompany their cessation journey following their personal dedication to quitting.

XSL•FO
RenderX

Theme 2: Parallel Experiences—Differing Views on Community

Participants expressed divergent attitudes toward engaging with app-based vaping cessation communities, reflecting a range of personal preferences, needs, and beliefs about the role of social support in quitting. Engagement with in-app communities was shaped by personal values, perceived credibility, stigma, and the app's design choices. While many youth spoke positively about the concept of in-app social features, others conveyed reluctance or disinterest.

For some youth, there were clear benefits of social connection. These participants highlighted how in-app communities could offer emotional validation, encouragement, and a sense of shared experience during vulnerable moments. One participant emphasized how peer connection could serve as a powerful reminder during times of struggle:

> *You forget that there are people around, you tend to forget that you're not alone. It's like it can, you just can become trapped in a world of misery. And it's like a reminder that you're not alone. Sometimes you need to see it.* [Participant #6, Female]

Others valued the opportunity to hear about shared experiences, which helped normalize their own journey and served as a source of encouragement:

> *When you see someone is actually getting better or doing better, you might be encouraged, you want to try to, since this person is doing it, I too can do it.* [Participant #17, Male]

Another participant emphasized the value of peer connection, especially when dealing with social pressure:

> *Sometimes it's hard. Especially with all the social pressures around you, it's easy to swerve off track. So I think having an app and having, like, people that you can visibly see are going through the same thing as you has been really great.* [Participant #11, Female]

For these participants, the sense of connection and affirmation that accompanied app-based social support features was meaningful and valuable.

At the same time, some participants opted not to engage with in-app social features, either due to a general lack of interest or a preference for managing cessation as a self-directed, private process. For some, connecting with others around quitting vaping felt unnecessary or even uncomfortable. One participant shared:

> *I would say that I wasn't even that open to finding this community in the first place. I kind of just, it's a personal thing, right? […] I've always been the type of person where it's like, if I do something like, I'll be proud of myself, I don't really need that community to show off like that I've quit nicotine or anything […] It's embarrassing when you want to quit something you've tried so hard to quit, and you can't like it's not something really I want to share with the community.* [Participant #8, Male]

Another echoed a similar sentiment when questioned about in-app social support features and highlighted their preference for alternative online discussion platforms where they felt they could engage more organically or anonymously, stating:

> *I won't see myself using it. Like, for example, it was like a buddy system […] it could definitely be beneficial to some people, but I feel like for me, I'm just going through it alone. So like, if I want to talk to someone, I could use other apps like Discord or Reddit that can help me you know, connect with some people that they're going to the same thing as me.* [Participant #2, Male]

Some participants who did not wish to use social support features themselves or had not personally engaged with them often demonstrated an empathetic and theoretical awareness of their potential value to others. These participants recognized that not everyone has access to social encouragement outside of the app and saw how peer connections could serve as a critical resource. As a participant explained:

> *I wasn't really into like, I guess connecting with other people, but I can see how it can really help some people if they're also going through the same thing, and you can connect with them and know that you're not alone and you have that added support that you may not have otherwise.* [Participant #1, Female]

Beyond personal preference, some participants questioned the effectiveness and credibility of the social features offered by cessation apps. Concerns were raised about community size, visibility, and the perceived lack of investment from app developers in fostering these spaces. One participant explained:

> *If I wanted to find a community for like anti-vapers, or the non-smokers, it would be easier to go on the internet and like, find that community rather than this, like having to download this random app, and just assume they have a good community on there.* [Participant #8, Male]

They elaborated on how the lack of promotion or visibility signaled a weak or undeveloped social support feature:

> *I feel like when there is a big community, the company will advertise that a lot. And like, for this app, it just seemed like such a small part of the app. So, through that, I could already tell that, like, there wasn't really an aspect of the app that like the developers probably worked on a lot or were proud of.* [Participant #8, Male]

For this youth, the value of an app's social feature was influenced by their perception of its importance to developers. When community tools appeared secondary or underpromoted, some participants dismissed them as inauthentic or untrustworthy. Participants shared a range of experiences and perceptions of app social support features. While some youth found value in shared experiences and forming community with others, others preferred to manage their quit journeys privately or through broader online platforms.

XSL•FO
RenderX

Theme 3: Pocket-Sized Partners in Cessation

Support From Contemplation to Commitment

Vaping cessation apps were viewed as multifaceted tools that offered support across the continuum of quitting, from initial contemplation to sustained commitment. The early phases of quitting were supported through app features that reinforced the decision to quit by highlighting health and financial benefits and prompting users to reflect on their personal motivations. Several youth appreciated the inclusion of positively framed, evidence-based information about how the body recovers after quitting, which helped dispel misconceptions and instilled hope. For these youth, the apps provided affirmation that change was both possible and worthwhile.

As participants progressed in their cessation journeys, visual depictions of progress emerged as particularly valued features. These included stats like time vape-free, money saved, and improvements in health, presented in clear, tangible formats that users found both satisfying and motivating. One participant shared:

> *[The app] tells me how much money I've saved. And it also includes goals, which I really like, because I'm pretty cheap when it comes to stuff. So when it tells me I've saved right now, it says $86. That is enough motivation for me to continue to stop in itself.* [Participant #11, Female]

This positive tone was especially helpful during moments of craving or doubt. For many, being able to see how far they had come helped them resist relapse and regain motivation. Participants also appreciated apps that celebrated small wins, such as weekly milestones or progress trophies. These features helped build momentum and a sense of accomplishment. A participant shared:

> *Trophies they give you at every week, whether you're vape free, so it's, you know, it can it helps you keep motivated to not start again. It's like a small thing. But you know, for me like, it's, it's made a big difference.* [Participant #6, Female]

These rewards, though seemingly minor, were perceived as meaningful validations of effort, reinforcing continued engagement with the app. With that in mind, participants were dismayed by any overly punitive loss of progress. One youth highlighted this saying:

> *The problem with it, I would say, is that you have like a huge streak and then if you hit it [vape], like once it's like, you lose all of it.* [Participant #15, Male]

Overall, participants emphasized the importance of positive reinforcement and personalization in maintaining their motivation throughout their quit journeys. Apps that provided motivational content, celebrated progress, and allowed personal goal setting were described as accessible and emotionally resonant, making the quitting process feel more achievable.

Caveat: Convenience Alone Is Not Enough

Although participants appreciated the ease and accessibility of app-based cessation tools, many emphasized that convenience alone was insufficient for sustained engagement or meaningful support. They were clear that vaping cessation apps must be thoughtfully designed with youth-specific needs and experiences in mind. Several participants described glitches or shallow content that felt disconnected from their real-life experiences of addiction and quitting. These limitations made the app feel more like a passive tracker than an active support system. Participants frequently critiqued basic or redundant features, such as simple day counters, as unoriginal and easily replicable. A participant explained:

> *I can always on a piece of paper, I could track on the days that I'm not vaping. Right, so like, even though that part of the app is helpful, like it's not really anything special. App should incentivize more to keep people accountable.* [Participant #8, Male]

Others suggested that an app was limited in its ability to provide comprehensive cessation support. Apps that functioned solely as stand-alone tools were perceived by some participants as insufficient, particularly in vulnerable moments of their cessation journeys. One participant noted:

> *The app does not work well on its own; it needs to be accompanied by some form of social support to act as an external motivator. It did help in the beginning, like to be honest, like, but I feel like the app plus something else would be like a better option. I feel like the app itself is not good as a singular method to stop.* [Participant #12, Female]

For these users, superficial tracking tools lacked the motivation or engagement needed to sustain change. To improve effectiveness, participants recommended more dynamic and personalized features, such as daily (rather than monthly) progress updates, visually expressive milestones, and uplifting motivational rewards. Youth wanted apps to feel not only functional but also affirming, something that could reflect and celebrate their efforts in real-time. To this point, one participant voiced their experience, stating:

> *When I, I feel like, I feel that that that need coming up? that like itch to vape. You know, I opened the app, and I look at those reminders, and it helps me control those urges.* [Participant #4, Female]

Some participants were particularly drawn to gamification as a mechanism to increase engagement. These features reframed cessation as a challenge or interactive experience, transforming the quitting journey into something more fun and rewarding. One participant shared:

> *Overall, like, I really liked the whole concept of the app, like how it's kind of like a game, I feel like, you get you like, unlock like, like these milestones for completing like, like cessation-related tasks.* [Participant #9, Female]

Another echoed this enthusiasm:

> *We're having fun tasks that look like a game or something [...] the gamification obviously kept me coming back because it was fun.* [Participant #19, Female]

XSL•FO
RenderX

These features helped keep youth invested, particularly during periods when intrinsic motivation may have wavered.

In addition to gamified elements, some participants expressed interest in extrinsic motivators, especially financial incentives. They suggested that tangible rewards could offer meaningful reinforcement, particularly for youth facing limited access to traditional support systems.

> *It would be cool if there's like a point system where, like, I don't know, if you got like 20 cents for like, five months going vape free or something like that […] it would be cool, because help isn't like the only motivation, there are other motivations, like saving money.* [Participant #10, Male]

These suggestions bring forward some material elements that youth want to see in a cessation app to make it worthwhile.

Beyond incentives and features, some youth highlighted the absence of connection to human support, particularly from health professionals. While digital tools were seen as useful, they did not replace the relational aspects of care. One participant expressed a desire for real-time, human interaction, saying:

> *[I'd] love to have the option of speaking to someone […] Even if we're not going to show faces just someone, not a bot, but an actual human being who I can talk to, even if I have to pay for the services. Because I think what people are trying to avoid is meeting people in person.* [Participant #14, Male]

This participant valued empathy and responsiveness as integral to achieving their cessation goals, elements that they perceived were difficult to replicate through static app content.

Finally, participants emphasized that apps should more authentically mirror the lived realities of youth navigating addiction. As one participant put it:

> *Your cell phone, as much as you're on it, that's not your real life.* [Participant #7, Female]

This youth acknowledged that, as digital tools, cessation apps are inherently distinct and separate from their lived experiences and therefore limited. Youth called for content that acknowledges the emotional, social, and identity-related complexities of quitting, including real-time chat, peer stories, or spaces to share experiences. They expressed a need for tools that go beyond tracking behaviors and instead foster a sense of connection, relatability, and real-world relevance.

Discussion

Main Findings

This study explored youth experiences with vaping cessation apps and identified 3 interrelated themes that reflect the complexity of digital cessation support. The first theme, "Quitting On My Own Terms," reflected the youth's emphasis on autonomy and personal responsibility in the quitting process. While some youth found that app-based tools aligned with their sense of independence, others viewed any external aid, even a self-guided app, as misaligned with their belief in "quitting solo." The second theme, "Parallel Experiences- Differing Views on Community," revealed divergent perspectives on app-based social support features. Some participants valued community engagement and peer stories, whereas others expressed skepticism about their relevance and opted not to engage, preferring independence or alternative platforms. Finally, "Pocket-Sized Partners in Cessation" described youth preferences for apps that offer visual feedback, motivational reinforcement, and personalized encouragement through the different stages of cessation. Yet, participants were clear that convenience alone was insufficient; apps must feel authentic, emotionally resonant, and relevant to the realities of quitting vaping.

Taken together, these findings suggest that youth want tools that affirm their self-awareness, validate their autonomy, and simultaneously offer the option of meaningful connection and support. Apps that appear overly simplistic, technically flawed, or disconnected from youth's lived realities are unlikely to foster sustained engagement. Instead, youth called for interactive, affirming, and emotionally intelligent tools, pocket-sized partners that do more than count days; they recognize and account for the complexity of change.

These findings align with prior studies demonstrating that youth value flexible, personalized features, behavior tracking, and positive reinforcement in digital health interventions [13,24,25,33-35]. Consistent with Bold et al [33], participants emphasized the importance of supportive rather than condescending messaging. While prior research has highlighted personalization and community support, participants in this study related the central role of autonomy in cessation efforts. For some, even minimally directive features felt misaligned with their belief in quitting independently. While moderated online communities have been previously identified as desirable [12-14,35,36], participants expressed a degree of ambivalence and skepticism, underscoring the need for apps that prioritize social support features to ensure these communities are meaningful and thoughtfully integrated. A novel contribution of this study is the critique that convenience alone is not enough. Youth distinguished between apps that merely track progress and those that offer authentic, emotionally validating support. This insight suggests that digital cessation interventions must extend beyond functionality to resonate meaningfully with youth's lived experiences.

From a developmental perspective, adolescence and emerging adulthood are marked by identity formation, autonomy, and peer influence [37]. Aligning interventions with these developmental needs, via supporting independence alongside offering pathways for connection, may strengthen effectiveness and engagement. Youth in this study also demonstrated a pragmatic view of health behavior change, acknowledging personal responsibility while recognizing the difficulties of cessation. This realism presents opportunities for interventions that walk alongside youth, rather than offering prescriptive solutions.

Implications for Intervention Design

These findings offer several implications for the design of youth-centered cessation apps. First, interventions should

XSL•FO
RenderX

incorporate features that validate youth self-reflection and agency. Positive reinforcement, affirming language, and recognition of progress may enhance engagement [12,24,25,34]. However, autonomy must be balanced with the normalization of help-seeking. Apps should frame support-seeking not as a weakness but as a strength, reinforcing that quitting independently and seeking help are not mutually exclusive.

As a means toward this end, social features must be intentionally designed within the app to boost credibility and the likelihood of engagement. Evidence indicates that a supportive community is a predictor of successful nicotine cessation [36,38,39]. Participants in this study expressed a desire to connect with others who could relate to their personal experiences of cessation and gather encouragement from a supportive community. Yet, participants also expressed skepticism about the quality of the app's social support features and, subsequently, about the robustness of the community and the support offered. Some participants expressed greater interest in joining cessation communities on larger, more established platforms such as Reddit and Discord. A study of a Facebook cessation support group found that strong leadership by moderators, which stimulated discussion, increased user engagement [40]. Consequently, there is an opportunity for apps to improve their value and effectiveness by investing in and prioritizing social support features, so that they adequately meet the expressed desire of youth to connect with others who are also attempting to quit. If cessation apps were to likewise use moderators or mentors to guide and support discussions on their social features, this could be one step toward increasing engagement and boosting credibility, with the potential to initiate a snowball effect of engagement. Given the connection between a supportive community and cessation success, the social support features of apps should not be an afterthought, but an intentionally designed and marketed central feature.

Third, relapses and setbacks need to be anticipated and normalized as part of the quitting journey. Youth in this study did not commonly account for relapse, and most apps lacked relapse-responsive features to address nonlinear cessation trajectories. This gap is concerning, as lapses are common among youth who attempt to quit [14] and, if not supported, may lead to disengagement. Relapse-responsive features such as customizable recovery plans, self-reflective journaling, adaptive goal setting, and encouraging messaging during setbacks may reframe lapses as learning opportunities rather than failures. Incorporating such features may also strengthen self-efficacy, a key component of cessation as voiced by participants, by reminding users that quitting is a process rather than a one-time decision.

Fourth, hybrid models that combine digital autonomy with access to real people may be particularly valuable [34]. While many youth want to quit independently, others indicated that relational depth and accountability could enhance motivation. Linking users to counselors, quit coaches, peer mentors, or live chat support could provide individualized guidance while preserving choice and autonomy. In parallel, design elements such as gamification, milestone-based incentivization, and progress tracking may sustain attention and reinforce positive behavior change over time [24,25]. Importantly, these features must move beyond superficial rewards to genuinely celebrate progress and reinforce resilience, thereby creating a sense of partnership between the user and the app.

Strengths and Limitations

This study provides a nuanced understanding of how youth experience and interpret vaping cessation apps, offering important insights for future digital health intervention design. The sample included representation from both young women and men, as well as geographic diversity across Eastern and Western Canadian provinces. These characteristics enhance the breadth of perspectives captured and contribute to the contextual richness of the analysis. The study also benefited from a rigorous qualitative approach grounded in qualitative description [29], which allowed for attention to individual variation while identifying common patterns across experiences.

The researcher's subjectivity was viewed as an analytic resource rather than a bias to be minimized, and movement between phases was fluid rather than strictly linear. Reflexivity was further supported through consideration of the researcher's positionality. The research team consisted of qualitative public health researchers with varying levels of experience in tobacco and vaping-related research. The principal investigator (LS) is a senior scholar with extensive expertise in tobacco, nicotine, and vaping research. Team member (KF) is an emerging scholar in nicotine cessation, while MT, RN, and DR are trainees in vaping research. Several of the research assistants were close in age to participants, which may have facilitated rapport-building during interviews and promoted more candid sharing. These intersecting identities and disciplinary backgrounds shaped the analytic lens through which data were interpreted.

Some limitations must be acknowledged. The sample did not include noncisgender individuals, and perspectives from youth living in several Canadian provinces remain unrepresented. As a result, the findings may not fully reflect the experiences of gender-diverse youth or those living in other regional or cultural contexts, thereby limiting transferability. In addition, because the study was embedded within a larger project focused specifically on app users, the findings may reflect the unique experiences and motivations of youth who are already open to using digital tools for cessation. Furthermore, data on the use of other nicotine products (eg, smoking, nicotine pouches, or nicotine replacement therapy) were not collected from this sample. Even further, the data reflect the perspectives of youth in 2022. Given the rapid evolution of vaping products, mobile health technologies, and regulatory environments, future studies will be crucial in exploring how youth perspectives shift over time.

Conclusions

Youth desire cessation apps that are authentic, emotionally intelligent, and responsive to the complexities of quitting vaping. Effective interventions must validate autonomy while encouraging safe connections, normalize setbacks as part of the quitting process, and integrate features that feel credible, supportive, and youth-friendly. By addressing the preferences of youth in the design of vaping cessation apps, digital cessation

XSL•FO
RenderX

tools can move beyond convenience to become meaningful companions in youths' efforts to quit vaping.

Acknowledgments

The authors would like to acknowledge and thank all participants for their time and contributions. The authors attest that no generative artificial intelligence was used in the preparation of this academic paper.

Funding

LS is supported by a Canadian Cancer Society Emerging Scholar Award (grant number 707156) and a Michael Smith Health Research BC Scholar Award (grant number SCH-2023-3195)

Data Availability

To protect the identity of the participants, the raw datasets are not publicly available.

Authors' Contributions

Conceptualization: LS
Data curation: LS, DR, RN, TA
Formal analysis: KF, MT, RN
Supervision: LS
Writing – original draft: KF, MT, RN
Writing – review and editing: KF, LS

Conflicts of Interest

None declared.

Multimedia Appendix 1
Semistructured interview guide.
[DOCX File, 40 KB - jmir_v28i1e85778_app1.docx]

References

1. Vahratian A, Briones EM, Jamal A, Marynak KL. Electronic cigarette use among adults in the United States, 2019-2023. NCHS Data Brief 2025 Jan(524):1-8. [doi: 10.15620/cdc/174583] [Medline: 40036117]
2. Jamal A, Park-Lee E, Birdsey J, et al. Tobacco product use among middle and high school students - National Youth Tobacco Survey, United States, 2024. MMWR Morb Mortal Wkly Rep 2024 Oct 17;73(41):917-924. [doi: 10.15585/mmwr.mm7341a2] [Medline: 39418216]
3. Cohen N, Jeter J, Schissler K, Lozano J, Maniaci V, Peña B. Poor perception and knowledge of electronic cigarettes among adolescents and their parents. Pediatr Emerg Care 2022 Feb 1;38(2):e709-e713. [doi: 10.1097/PEC.0000000000002339] [Medline: 35100768]
4. Crane LA, Asdigian NL, Fitzgerald MD. Looking cool, doing tricks, managing stress, and nicotine addiction: youth perspectives on nicotine vaping and implications for prevention. Am J Health Promot 2023 Sep;37(7):964-974. [doi: 10.1177/08901171231189560] [Medline: 37462022]
5. Struik L, Christianson K, Khan S, et al. Factors that influence decision-making among youth who vape and youth who don't vape. Addict Behav Rep 2023 Dec;18:100509. [doi: 10.1016/j.abrep.2023.100509] [Medline: 37519860]
6. Tobacco 21. US Food and Drug Administration. 2025. URL: https://www.fda.gov/tobacco-products/retail-sales-tobacco-products/tobacco-21 [accessed 2026-04-05]
7. Tobacco and vaping products act, S.C. 1997, c. 13. Government of Canada. URL: https://laws-lois.justice.gc.ca/eng/acts/t-11.5/FullText.html [accessed 2026-04-05]
8. Banks E, Yazidjoglou A, Brown S, et al. Electronic cigarettes and health outcomes: umbrella and systematic review of the global evidence. Med J Aust 2023 Apr 3;218(6):267-275. [doi: 10.5694/mja2.51890] [Medline: 36939271]
9. Chaiton M, Fan J, Bondy SJ, et al. E-cigarette dependence and depressive symptoms among youth. Am J Prev Med 2024 Jan;66(1):104-111. [doi: 10.1016/j.amepre.2023.09.020] [Medline: 37774992]
10. Glantz SA, Nguyen N, Oliveira da Silva AL. Population-based disease odds for e-cigarettes and dual use versus cigarettes. NEJM Evid 2024 Mar;3(3):EVIDoa2300229. [doi: 10.1056/EVIDoa2300229] [Medline: 38411454]
11. Amato MS, Bottcher MM, Cha S, Jacobs MA, Pearson JL, Graham AL. "It's really addictive and I'm trapped:" a qualitative analysis of the reasons for quitting vaping among treatment-seeking young people. Addict Behav 2021 Jan;112:106599. [doi: 10.1016/j.addbeh.2020.106599] [Medline: 32950927]

XSL•FO
RenderX

12. Struik L, Christianson K, Khan S, Sharma RH. Strengths and limitations of web-based cessation support for individuals who smoke, dual use, or vape: qualitative interview study. JMIR Form Res 2023 Dec 8;7:e43096. [doi: 10.2196/43096] [Medline: 38064266]
13. Struik L, Yang Y. E-cigarette cessation: content analysis of a quit vaping community on Reddit. J Med Internet Res 2021 Oct 25;23(10):e28303. [doi: 10.2196/28303] [Medline: 34694229]
14. Al-Hamdani M, Davidson M, Bird D, Hopkins DB, Smith S. Learning from their experiences: strategies used by youth and young adult ex-vapers. J Subst Use Addict Treat 2023 Jun;149:209038. [doi: 10.1016/j.josat.2023.209038] [Medline: 37061190]
15. Lin C, Mathur Gaiha S, Halpern-Felsher B. E-cigarette and combustible cigarette cessation patterns, reasons, and methods among adolescents, young adults, and adults. Addict Behav 2024 Mar;150:107918. [doi: 10.1016/j.addbeh.2023.107918] [Medline: 38070362]
16. Amin S, Pokhrel P, Elwir T, Mettias H, Kawamoto CT. A systematic review of experimental and longitudinal studies on e-cigarette use cessation. Addict Behav 2023 Nov;146:107787. [doi: 10.1016/j.addbeh.2023.107787] [Medline: 37393697]
17. Sanchez S, Kaufman P, Pelletier H, et al. Is vaping cessation like smoking cessation? a qualitative study exploring the responses of youth and young adults who vape e-cigarettes. Addict Behav 2021 Feb;113:106687. [doi: 10.1016/j.addbeh.2020.106687] [Medline: 33045643]
18. Asselin G, Bilodeau H, Khalid A. Digital well-being: the relationship between technology use, mental health and interpersonal relationships. : Statistics Canada; 2024 URL: https://publications.gc.ca/collections/collection_2024/statcan/222000012024001-eng.pdf [accessed 2026-05-21]
19. Dai HD, Hanh P, Guenzel N, Morgan M, Kerns E, Winickoff JP. Adoption of vaping cessation methods by US adolescent e-cigarette users. Pediatrics 2023 Nov 1;152(5):e2023062948. [doi: 10.1542/peds.2023-062948] [Medline: 37781732]
20. Heshmati J, Pandey A, Benjamen J, et al. Vaping cessation interventions: a systematic review and meta-analysis. Tob Control 2026 May 7;35(3):382-390. [doi: 10.1136/tc-2024-058798] [Medline: 40011054]
21. McKay F, Chan L, Cerio R, et al. Assessing the quality and behavior change potential of vaping cessation apps: systematic search and assessment. JMIR Mhealth Uhealth 2024 Mar 15;12:e55177. [doi: 10.2196/55177] [Medline: 38532616]
22. Sanchez S, Kundu A, Limanto E, Selby P, Baskerville NB, Chaiton M. Smartphone apps for vaping cessation: quality assessment and content analysis. JMIR Mhealth Uhealth 2022 Mar 28;10(3):e31309. [doi: 10.2196/31309] [Medline: 35343904]
23. Fonteyne K, Keys E, Hasan K, Struik L. Exploration of digital interventions for vaping cessation: scoping review. J Med Internet Res 2025 Oct 23;27:e76983. [doi: 10.2196/76983] [Medline: 41129808]
24. Fonteyne K, Keys E, Hasan MK, Struik L. Youth and young adults' perspectives on augmented reality–driven vaping cessation interventions: interpretive description study. JMIR XR Spatial Comput 2025;2:e79674-e79674. [doi: 10.2196/79674]
25. Huma ZE, Struik L, Bottorff JL, Hasan MK. Preferences for mobile-supported e-cigarette cessation interventions among young adults: qualitative descriptive study. JMIR Form Res 2022 Apr 1;6(4):e33640. [doi: 10.2196/33640] [Medline: 35363140]
26. Milne-Ives M, Lam C, De Cock C, Van Velthoven MH, Meinert E. Mobile apps for health behavior change in physical activity, diet, drug and alcohol use, and mental health: systematic review. JMIR Mhealth Uhealth 2020 Mar 18;8(3):e17046. [doi: 10.2196/17046] [Medline: 32186518]
27. Rodberg D, Nawara R, Taylor M, Struik L. User perceptions of e-cigarette cessation apps: content analysis of app reviews. J Med Internet Res 2025 Apr 15;27:e59997. [doi: 10.2196/59997] [Medline: 40233344]
28. Pilarska J. The constructivist paradigm and phenomenological qualitative research design. In: Research Paradigm Considerations for Emerging Scholars: Channel View Publications Ltd; 2021. [doi: 10.2307/jj.22730542.10]
29. Sandelowski M. What's in a name? qualitative description revisited. Res Nurs Health 2010 Feb;33(1):77-84. [doi: 10.1002/nur.20362] [Medline: 20014004]
30. Braun V, Clarke V. Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern ‐ based qualitative analytic approaches. Couns and Psychother Res 2021 Mar;21(1):37-47. [doi: 10.1002/capr.12360]
31. Braun V, Clarke V. One size fits all? what counts as quality practice in (reflexive) thematic analysis? Qual Res Psychol 2021 Jul 3;18(3):328-352. [doi: 10.1080/14780887.2020.1769238]
32. Braun V, Clarke V. Thematic Analysis: A Practical Guide: SAGE Publications Ltd; 2022.
33. Bold K, Kong G, Cavallo D, Davis D, Jackson A, Krishnan-Sarin S. School-based e-cigarette cessation programs: what do youth want? Addict Behav 2022 Feb;125:107167. [doi: 10.1016/j.addbeh.2021.107167] [Medline: 34753093]
34. Rahman N, Sebar B, Sofija E. Co-designing a vaping cessation program for Australian young adults: a conceptual model. Nicotine Tob Res 2025 Feb 24;27(3):457-465. [doi: 10.1093/ntr/ntae222] [Medline: 39315564]
35. Lyu JC, Afolabi A, White JS, Ling PM. Perceptions and aspirations toward peer mentoring in social media-based electronic cigarette cessation interventions for adolescents and young adults: focus group study. JMIR Form Res 2022 Dec 1;6(12):e42538. [doi: 10.2196/42538] [Medline: 36454628]

XSL•FO
RenderX

36. Kim SJ, Marsch LA, Brunette MF, Dallery J. Harnessing Facebook for smoking reduction and cessation interventions: Facebook user engagement and social support predict smoking reduction. J Med Internet Res 2017 May 23;19(5):e168. [doi: 10.2196/jmir.6681] [Medline: 28536096]
37. Sussman S, Arnett JJ. Emerging adulthood: developmental period facilitative of the addictions. Eval Health Prof 2014 Jun;37(2):147-155. [doi: 10.1177/0163278714521812] [Medline: 24492245]
38. Patten CA, Goggin K, Harris KJ, et al. Relationship of autonomy social support to quitting motivation in diverse smokers. Addict Res Theory 2016;24(6):477-482. [doi: 10.3109/16066359.2016.1170815] [Medline: 27833474]
39. Struik L, Khan S, Assoiants A, Sharma RH. Assessment of social support and quitting smoking in an online community forum: study involving content analysis. JMIR Form Res 2022 Jan 13;6(1):e34429. [doi: 10.2196/34429] [Medline: 35023834]
40. Struik LL, Baskerville NB. The role of Facebook in Crush the Crave, a mobile- and social media-based smoking cessation intervention: qualitative framework analysis of posts. J Med Internet Res 2014 Jul 11;16(7):e170. [doi: 10.2196/jmir.3189] [Medline: 25016998]

Abbreviations

UBC: University of British Columbia

Edited by A Stone; submitted 13.Oct.2025; peer-reviewed by C Baxter, S Cha; revised version received 05.Apr.2026; accepted 07.Apr.2026; published 05.Jun.2026.

Please cite as:
Fonteyne K, Taylor M, Nawara R, Rodberg D, Afshar T, Struik L
Exploring Vaping Cessation App Use Among Youth: Qualitative Descriptive Study
J Med Internet Res 2026;28:e85778
URL: https://www.jmir.org/2026/1/e85778
doi:10.2196/85778

© Karlee Fonteyne, Mischa Taylor, Roula Nawara, Danielle Rodberg, Tina Afshar, Laura Struik. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Chinese Mobile Health Apps for Preventing and Managing Pelvic Floor Dysfunction: Quality Assessment and Content Analysis

Yuqing Song[1,2*], PhD; Xue Deng[1,2*], BSN; Ting Hu[3], BSN; Sijie Feng[4], BM; Lu Xing[1,2], PhD

[1]Department of Pelvic Floor Medicine Center Nursing, West China Second University Hospital, Sichuan University, No. 20, Section 3, South Renmin Road, Sichuan Province, Chengdu, China

[2]Key Laboratory of Birth Defects and Related Diseases of Women and Children (Sichuan University), Ministry of Education, Chengdu, Sichuan, China

[3]Department of Gynecology, Guangdong Women and Children Hospital, Guangzhou, China

[4]Department of Gynaecology and Obstetrics, West China Second University Hospital, Sichuan University, Chengdu, Sichuan, China

[*]these authors contributed equally

Corresponding Author:
Lu Xing, PhD
Department of Pelvic Floor Medicine Center Nursing, West China Second University Hospital, Sichuan University, No. 20, Section 3, South Renmin Road, Sichuan Province, Chengdu, China

Abstract

Background: Pelvic floor dysfunction (PFD) is a highly prevalent health problem, encompassing urinary incontinence, emptying disorders of the bladder, fecal incontinence, emptying disorders of the bowel, pelvic organ prolapse, sexual dysfunction, and chronic pelvic pain. Mobile health (mHealth) interventions delivered through apps can provide remote health services to improve patient compliance and enhance treatment effectiveness. Although apps for preventing and managing PFD have been developed and used, the features and quality of these apps in China have not been systematically examined.

Objective: This study aimed to systematically summarize the functions and evaluate the quality of the existing mHealth apps for preventing and managing all kinds of PFD, such as urinary incontinence, fecal incontinence, and chronic pelvic pain.

Methods: We systematically searched for potential PFD apps on the Apple App Store, Huawei AppGallery, and VIVO App Store. Apps were included if they were free, designed for preventing or managing PFD, in the Chinese language, could be downloaded and run on Android, Harmony, or iOS operating systems (OS), and incorporated elements of preventing and managing PFD. We excluded apps that were intended for use by health care providers and not relevant to PFD. Apps that met the inclusion criteria were downloaded and included for final analysis. The user version of the Mobile App Rating Scale (uMARS) was used to assess the apps' quality and summarize the apps' functionality according to guidelines.

Results: Of the 3897 apps screened, 46 apps that met the inclusion criteria were included in the final analysis. All apps were developed by corporations. More than half of the apps had download counts exceeding 10,000, and 24 (52.2%) apps scored 4 or higher in app stores. Furthermore, nearly half of the apps (n=21, 45.7%) had been updated within the past month at the time of retrieval. The overall uMARS scores ranged from 2.29 to 4.50, with a mean uMARS score of 3.46 (SD 0.50), which is considered acceptable quality. Based on uMARS scores, 15.2% (n=7) were rated as poor quality, 65.2% (n=30) as acceptable, and 19.6% (n=9) as good quality. More than half of the apps provided the functions of exercise (n=44, 95.7%), personal information recording (n=31, 67.4%), and health education (n=28, 60.9%). Only 5 apps provided 5 or more functions.

Conclusions: The apps for PFD revealed acceptable quality, and the majority provided exercise, personal information recording, and health education functions. However, many apps lacked comprehensive functionalities and did not provide immediate feedback or high-quality educational information. Health care providers should follow international guidelines to create high-quality, evidence-based, multifunctional apps for PFD. Future studies should explore the effects of the apps and real-world user feedback data in clinical settings.

(J Med Internet Res 2026;28:e80126) doi:10.2196/80126

KEYWORDS

pelvic floor dysfunction; app; mobile health; mHealth; pelvic floor; quality assessment

Introduction

Pelvic floor dysfunction (PFD) refers to functional impairments of the musculature that support the bladder, bowel, and vagina, including urinary incontinence, fecal incontinence, pelvic organ prolapse, sexual dysfunction, and chronic pelvic pain [1,2]. PFD is a highly prevalent health problem that seriously affects the daily function, sexual function, self-image, and quality of life of patients [3-5]. In China, the prevalence of symptomatic pelvic

XSL•FO
RenderX

organ prolapse is 9.6% [6], and the weighted prevalence of female urinary incontinence is 16.0% [7]. By comparison, in a study of 1446 women in Spain, urinary incontinence occurred in 55.8% (807) of the women, fecal incontinence in 10.4% (150), and symptomatic uterine prolapse in 14.0% (203) [8]. Patients with PFD require long-term care to manage their symptoms. However, only 10.7% of patients seek medical care [9].

Patients with PFD face challenges in accessing effective and high-quality health services, such as low patient motivation and frequent health care visits [10,11]. Mobile health (mHealth) interventions delivered through apps can overcome these challenges by offering convenient tools, reminders, and instructions for pelvic floor muscle training (PFMT), improving patient compliance and treatment effectiveness through remote monitoring and feedback [12,13]. A longitudinal study involving 3051 postmenopausal women with PFD demonstrated that PFMT combined with education delivered through a mobile app significantly reduced the impact of pelvic floor symptoms on daily life [14]. A systematic review [15] revealed that mHealth app–based PFMT could improve symptom severity, quality of life, and adherence to treatment among women with stress urinary incontinence or stress-predominant mixed urinary incontinence.

Many current mHealth apps for PFD mainly focus on delivering PFMT [15-18]. mHealth app quality, such as usability, accuracy of information, and functionality, affects the user experience, determines whether users continue to use the app, and raises concerns regarding their long-term effectiveness and reliability [19]. Systematic reviews [16-18] have highlighted several limitations in the current PFD-related mHealth apps. Many apps lack a solid evidence base, with insufficient involvement from health care professionals in their design and content validation [17,18]. Sudol et al [18] evaluated 23 English apps in female pelvic medicine and reconstructive surgery but did not summarize the functions of these apps. Kasoff et al [17] reviewed 83 English apps for PFD, with functions such as voiding and bowel diaries, symptom tracking, medication tracking, social function, reminder systems, and disease information. However, no published articles summarize publicly available Chinese apps for PFD.

It is also worth highlighting that health care professionals should identify high-quality apps for individuals. Standardized instruments, such as the Mobile App Rating Scale (MARS), can systematically assess the quality of mHealth apps [20]. The MARS has demonstrated high interrater reliability for evaluating mHealth app quality across diverse domains [20]. However, its administration requires specialized training and expertise in mHealth and the relevant health field [21]. The user version of MARS (uMARS) is a reliable and user-friendly measure of app quality that can be administered by target users without professional training in mHealth [21]. As mHealth apps are primarily accessed by end-users rather than professionals, the uMARS could identify high-quality and useful apps from users' perspectives.

Although numerous mHealth apps have been developed and implemented for PFD, Chinese apps have not been systematically summarized or assessed. Thus, this study aimed to summarize the content and functions of mHealth apps in China for PFD and evaluate their quality using uMARS.

Methods

Study Purpose

This study aimed to systematically review, evaluate the quality, and summarize the content and function of the existing publicly available mHealth apps in China for preventing and managing PFD, including urinary incontinence, pelvic organ prolapse, fecal incontinence, emptying disorders of the bladder, emptying disorders of the bowel, sexual dysfunction, and chronic pelvic pain. This study was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines for systematic reviews [22].

Search Strategy

We systematically searched mHealth apps for PFD between March 24 and 28, 2025. Android, Harmony, and iOS are the most popular smartphone operating systems in China [23]. Huawei AppGallery is the largest app store, though some relevant apps are not available on it [24]. Xiaomi Market, OPPO Software Store, and VIVO App Store are also among the top 5 app stores in China [24]. Thus, we searched apps in the 3 major Chinese app stores: Apple App Store, Huawei AppGallery, and VIVO App Store. We conducted preliminary searches and group discussions to determine the optimal keywords for the 3 app stores, ensuring that all relevant apps could be screened. We extracted keywords from existing reviews on PFD, used the keywords for an initial app search in app stores, and then revised the keywords. Subsequently, we conducted group discussions to supplement additional keywords and revise keywords that were irrelevant or unsuitable for the research topic. The final Chinese keywords included: pelvic floor (盆底), pelvic floor muscle (盆底肌), pelvic floor rehabilitation (盆底康复), pelvic floor muscle rehabilitation (盆底肌康复), pelvic floor dysfunction (盆底功能障碍), pelvic floor disorders (盆底功能障碍), pelvic floor training (盆底训练), pelvic floor exercise (盆底锻炼), pelvic floor muscle exercise (盆底肌锻炼), pelvic floor muscle training (盆底肌训练), pelvic floor functional exercise (盆底功能锻炼), Kegel exercises (凯格尔), pelvic floor management (盆底管理), pelvic floor muscle management (盆底肌管理), pelvic floor treatment (盆底治疗), pelvic floor muscle treatment (盆底肌治疗), pelvic organ prolapse (盆腔脏器脱垂), prolapse (脱垂), uterine prolapse (子宫脱垂), vaginal prolapse (阴道脱垂), cystocele or bladder prolapse (膀胱脱垂), rectocele or rectal prolapse (直肠脱垂), urinary incontinence (尿失禁), overactive bladder (膀胱过度活动症), defecation dysfunction (排便功能障碍), fecal incontinence (大便失禁/粪失禁), and sexual dysfunction (性功能障碍). Two researchers used the above keywords to search apps in the 3 app stores and recorded all results.

Inclusion and Exclusion Criteria

Apps were included if they met the following criteria: (1) designed for preventing or managing PFD; (2) available for free download; (3) in the Chinese language; (4) available for download from the Apple App Store, Huawei AppGallery, and VIVO App Store; and (5) able to run on Android, iOS, and

XSL•FO
RenderX

Harmony operating systems. We reviewed the descriptions on app stores, downloaded and tested the apps, and excluded those not specifically focused on PFD (eg, general women's health apps without a pelvic floor component) or those intended for use by health care providers.

App Quality Assessment

We used the Chinese version of the uMARS to assess the apps' quality [25]. uMARS is a feasible and reliable tool for evaluating mHealth apps, specifically designed for users without mHealth expertise [20,25]. As our team did not include a mHealth specialist, we selected uMARS as the appropriate assessment tool. The uMARS includes 14 items divided into 3 subscales: information, functionality, and engagement. All items were rated on a 5-point Likert scale from 1 (inadequate) to 5 (excellent). Mean scores were calculated for each subscale, and an overall mean score was calculated across all 3 subscales. Apps scoring ≥3 out of 5 on the uMARS were considered to be of acceptable quality, whereas those scoring ≥4 were rated as high quality.

In the quality assessment stage, 2 researchers independently assessed the apps' quality. Initially, the researchers were required to thoroughly review the original uMARS development and translation literature to ensure a comprehensive understanding of the uMARS content. Then, the researchers independently evaluated other apps to establish consistent uMARS scoring rules. If the researchers' initial ratings differed by 2 or more points, they discussed the differences to reach a consensus. If they couldn't agree, a third researcher was brought in to resolve the issues. Finally, they downloaded the included apps and used them for at least 10 minutes to explore core features, the user interface, and navigational features before completing the uMARS evaluation. Pilot testing confirmed that 10 minutes was sufficient for a comprehensive app evaluation, so the researchers established 10 minutes as the evaluation threshold.

Data Extraction

Descriptive data were manually extracted by 2 independent researchers after reading the initial descriptions and purposes of each mHealth app on the app stores. The extracted information included general details and functionalities.

1. App Information: app name, app store, platform, developer, download count, update time, star rating, and privacy policy. The definitions of app information are presented in Table 1.
2. Function: We summarized the strategies for preventing and managing PFD from guidelines [1,26]. The functions of PFD apps were categorized into personal information recording, exercise, physical therapy, health education, self-assessment, online community, and reminder systems.

Table . General information collected for each app.

Assessment measure	Definition
App store	Apps downloaded from the Apple App Store, Huawei AppGallery, and VIVO App Store.
Platform	Apps could operate on Harmony, Android, and iOS operating systems.
Developer	Apps were developed by a corporation or individual developer.
Download count	Number of apps downloaded in Huawei AppGallery and VIVO App Store. If the number of app downloads in Huawei AppGallery and VIVO App Store was different, we recorded the higher number.
Update time	The latest update time when we retrieved apps.
Star rating	Star rating (out of 5) left by users in the app store. If the star rating was different in the 3 app stores, we recorded the rating that had the highest number of evaluations.

Personal information recording includes collecting demographic data, obstetric history, disease-related data, and symptom details. The function of exercise focuses on pelvic floor muscle exercise guidance, including audio instructions, visual guides, customizable plans, and muscle contraction and relaxation tracking. As for the physical therapy function, apps linked with biofeedback devices can detect pelvic floor muscle activity, provide real-time visual feedback for correct activation, and integrate electrical stimulation. The health education function provides educational information about PFD, including general information about PFD, prevention strategies, treatment options, exercise guidance, and lifestyle modifications. Self-assessment enables users to evaluate and track pelvic floor symptoms using validated questionnaires, wearable devices, and bladder and bowel diaries. The online community enables individuals with PFD to exchange knowledge, offer emotional support, and boost collective self-efficacy through digital communication. Reminder systems include medication reminders, voiding reminders, fluid intake reminders, and exercise reminders with customizable timing.

Data Synthesis

We used Excel (Microsoft Corp) to extract the apps' general information and functionality and to record the scores of uMARS. SPSS 27.0 (IBM Corp) was used for statistical analysis. The quantitative variables are described by the mean and SD. The count data were presented as frequencies and percentages. The uMARS and subscale scores were calculated as the mean scores of the ratings provided by the 2 researchers. We used intraclass correlation coefficients to analyze the consistency of the 2 researchers at the subscale and overall score levels.

XSL•FO
RenderX

Results

App Selection

Our initial search identified 3897 apps, including 842 apps from the Apple App Store, 805 apps from the VIVO App Store, and 2250 apps from the Huawei AppGallery. After the elimination of duplicate apps, the final count comprised 310 Android apps, 813 HarmonyOS apps, and 254 iOS apps. After an initial screening based on the inclusion and exclusion criteria, we downloaded and analyzed 86 iOS apps, 34 Android apps, and 39 HarmonyOS apps. Subsequently, 61 apps were excluded: 9 were not in English, 20 were either technically inaccessible or required external device connectivity (eg, vaginal pelvic floor therapy devices) for use, 22 were irrelevant to the research topic, 4 required payment for download, 2 targeted health care professionals, and 4 had excessive advertisements and poor quality. Finally, 42 iOS apps, 27 Android apps, and 29 Harmony operating system apps were included in this study. Of the 98 included apps, 22 apps were available on all 3 app stores (counted as 3 apps), and 8 apps were available on 2 app stores (counted as 2 apps). After excluding 52 redundantly counted apps, we included 46 apps in this study (Figure 1).

XSL•FO
RenderX

Figure 1. Flowchart of the selection of apps. PFD: pelvic floor dysfunction.

XSL•FO
RenderX

General Characteristics of the Apps

More than half of the apps (19/32, 59.4%) had download counts exceeding 10,000, and 24/46 (52.2%) apps scored 4 or higher. Nearly half of the apps (21/46, 45.7%) had recent updates (<1 mo), while 14/46 (30.4%) apps were updated between 1 month and 1 year, and 11/46 (23.9%) apps were outdated (>1 y) (Table 2).

Table . Characteristics of the apps for preventing and managing pelvic floor dysfunction (PFD; N=46).

Characteristic	Apps, n (%)
App store	
Apple App Store	42 (91.3)
Huawei AppGallery	29 (63.0)
VIVO App Store	27 (58.7)
Platform	
Harmony and Android and iOS	22 (47.8)
Harmony and Android	2 (4.3)
Harmony and iOS	3 (6.5)
Android and iOS	3 (6.5)
Harmony	2 (4.3)
iOS	14 (30.4)
Developer	
Individual developer	0 (0)
Corporation	46 (100)
Download count[a] (n=32)	
0 - 9999	13 (40.6)
10,000 - 99,999	4 (12.5)
100,000 - 999,999	7 (21.9)
1,000,000 - 9,999,999	6 (18.8)
≥10,000,000	2 (6.3)
Update time	
>1 year	11 (23.9)
>1 month and <1 year	14 (30.4)
<1 month	21 (45.7)
Star rating	
0 - 2.9	3 (6.5)
3.0 - 3.9	7 (15.3)
4.0 - 5.0	24 (52.2)
No rating	12 (26.1)

[a]Download counts were available only for Android and HarmonyOS apps, and 32 apps reported download data.

Evaluation of the Included Apps

The intraclass correlation coefficient of the uMARS score was 0.970 (95% CI 0.946 - 0.983). Three subscales demonstrated high consistency, with intraclass correlation coefficients as follows: information, 0.942 (95% CI 0.896 - 0.968); function, 0.839 (95% CI 0.709 - 0.911); engagement, 0.971 (95% CI 0.947 - 0.984).

The mean total score of uMARS was 3.46 (SD 0.50), ranging from 2.29 to 4.50. According to the uMARS scores, 7 (15.2%) were classified as poor quality, 30 (65.2%) were categorized as acceptable quality, and 9 (19.6%) were designated as good quality. The mean scores for functionality, information, and engagement were 3.67 (SD 0.51), 3.56 (SD 0.45), and 2.96 (SD 0.83), respectively. The results are shown in Table 3 and Multimedia Appendix 1.

XSL•FO
RenderX

Table . App quality measured by uMARS[a].

Variables	Values	Min-Max
Overall uMARS, mean (SD)	3.46 (0.50)	2.29-4.50
App quality stratification, n (%)		
Excellent (5)	0 (0)	—[b]
Good (4≤5)	9 (19.6)	—
Acceptable (3≤4)	30 (65.2)	—
Poor (2≤3)	7 (15.2)	—
Inadequate (1≤2)	0 (0)	—
Information, mean (SD)	3.56 (0.45)	2.57-4.57
Functionality, mean (SD)	3.67 (0.51)	2.50-4.75
Engagement, mean (SD)	2.96 (0.83)	1.33-4.33

[a]uMARS, the user version of the Mobile App Rating Scale.
[b]Not applicable.

Functionality Summary of the Included Apps

More than half of the apps offered functionalities that included health education, exercise, and personal information recording (Table 4). In multifunctional apps, the most prevalent combination of functionalities included personal information recording, exercise, and health education, accounting for 21.8% (10/46) of the included apps. Only 5 out of 46 (10.9%) apps provided 5 to 6 functions.

Table . Functionality summary of the included apps.

Functionality	Values, n (%)
Reminder systems	6 (13.0)
Online community	9 (19.6)
Health education	28 (60.9)
Physical therapy	12 (26.1)
Exercise	44 (95.7)
Self-assessment	20 (43.5)
Personal information recording	31 (67.4)

Personal Information Recording

A total of 31 apps offered personal information recording. The apps enabled users to record their demographic data and disease-related data. Some apps could also record and track users' exercise routines and pelvic floor symptoms.

Self-Assessment

Nearly half of the apps (20/46, 43.5%) offered self-assessment functionality that enabled users to evaluate their pelvic floor symptoms and health conditions. Many apps primarily assessed users' basic information and performed pelvic floor muscle contraction and relaxation testing to identify users' pelvic floor function and symptoms. Only a few apps provided validated instruments to help users assess their pelvic floor symptoms and other related health conditions.

Exercise

Almost all apps (44/46, 95.7%) offered PFMT (including audio or video-guided training) to various user groups: postpartum women, adolescent females, middle-aged and older women, and males. However, only 9 apps clearly stated that they were suitable for males. Users received instructions via audio or video to consistently and accurately contract and relax their pelvic floor muscles in synchrony with the accompanying music. Nineteen apps also offered paid courses guided by physicians or fitness instructors to help users learn and perform PFMT more effectively.

Physical Therapy

Only 12 apps offered a physical therapy function. To access the physical therapy functionality, these apps must be paired with a pelvic floor therapy device. Users should purchase a pelvic floor therapy device, connect it to the apps, and then perform pelvic muscle rehabilitation therapy at home.

Health Education

More than half of the apps (28/46, 60.9%) offered the function of pelvic floor health education. The educational content included articles and videos. Articles mainly explained what PFD is and how to perform PFMT, but much of the content was basic and lacked scientific rigor. Videos, which were primarily presented by health care professionals, offered accurate

XSL•FO
RenderX

information on the prevention and management of PFD. However, only 6 apps included video-based educational content.

Online Community

The online communication function was present in 9 (19.6%) apps, offering a social platform for pelvic floor health. Within this community, health care providers and users could engage in communication through text and images. In most apps' online communities, users primarily interacted to share their experiences regarding pelvic floor health, exercises, and daily life. Only 1 app offered health professional support, with experts addressing patients' questions about PFD.

Reminder Systems

The functionality of reminder systems was offered in 6 apps. This functionality enabled users to set reminders for pelvic floor muscle exercises and menstrual cycle tracking. Additionally, this functionality facilitated the tracking of menstrual cycles, which could be beneficial for the overall women's health management.

Discussion

Principal Results

This study systematically summarized the functions and evaluated the quality of 46 apps available in Chinese app stores that are designed for PFD. All included apps contained at least 1 of the 7 identified functions, with the majority providing functions related to health education, exercise, and personal information recording. Only 10.9% of the apps provided 5 or more of the identified functions. In this study, the mean uMARS score of the apps was 3.46, indicating acceptable quality. In other studies, the mean uMARS scores for eye-care apps, artificial intelligence–based apps, and mental health apps also ranged from 3.45 to 3.54 [19,27,28]. Kasoff et al [17] revealed that the quality scores of apps for PFD were low, and the most common functions were reminder systems, bowel diaries, exercise tracking, and voiding diaries. This suggests a considerable gap between guideline recommendations, the service capacity, and the quality of currently available PFD-related apps. We only reported descriptive uMARS scores and function counts, but did not evaluate clinical effectiveness. We suggest that the effectiveness of these apps should be explored in clinical trials and practice.

The overall quality of PFD-related apps was acceptable, which was similar to findings from other previous studies [17,27,28]. The results suggest that the current quality of apps remains suboptimal and warrants further enhancement. Across all assessed subscales, functionality achieved the highest average score, followed by information, while engagement received the lowest score. Previous studies [29-31] of apps targeting breast cancer, medicine, and medication management also reported the lowest scores in engagement and the highest scores in functionality. App developers have prioritized functionality, leading to notable improvements in app quality and usability in recent years [32]. Since mHealth apps primarily target doctors or patients, it is difficult to make users feel interested and relaxed when facing diseases [29]. This may contribute to the lowest scores in engagement [29]. This study used descriptive uMARS scores to rate the quality of apps. However, user engagement and participation, as well as the long-term effects of the apps on health outcomes, are also crucial for evaluating apps' quality. Thus, health care professionals should explore the sustained use and effects of apps in future studies and clinical settings.

PFMT, such as Kegel exercises, is widely recommended for preventing and managing PFD [1]. Numerous studies [15,33,34] have revealed that PFMT delivered through mHealth apps is feasible and effective in improving health outcomes and exercise adherence. This study also found that almost all PFD-related apps (44/46, 95.7%) offered the feature of PFMT, which is consistent with findings from other studies [14,17,35]. Many apps in our study provide step-by-step voice instructions and video demonstrations to guide users through these exercises. Previous studies [14,36,37] have also reported that many apps provided simple instructions to show patients images, videos, or sounds indicating when to contract and relax pelvic floor muscles. Immediate feedback can enhance user engagement by allowing users to monitor their improvement over time [38]. However, most apps lack personalized feedback and professional guidance, which may limit the effectiveness of these exercises.

More than half of the PFD-related apps (31/46) offered the function of recording personal information. However, these apps merely provided basic exercise information logging and generated charts to visualize exercise conditions. Literature [39,40] suggests that the patient health record function should shift from simple data access to data modification, automated assessment, prediction, and recommendation. In this study, 43.5% of the apps provided a self-assessment function. Most apps guided patients through pelvic floor muscle contractions and relaxations to assess their pelvic floor muscle function, but only a few apps offered validated instruments to assess and track users' symptoms. One study [41] also reported that few apps used validated instruments or wearable devices to enable users to conduct self-assessments.

The guidelines and reviews [1,42] have emphasized the importance of patient education for PFD. In this study, the educational function was offered in 60.9% of the PFD-related apps. The educational information consisted primarily of low-quality articles, which may not be tailored to users' needs and preferences [43]. Liu et al [19] found that eye-care education delivered through video format was more engaging than text-based education, due to the dynamic audio-visual elements of the videos. Byambasuren et al [44] summarized that educational information on mHealth apps often consisted of low-quality educational articles. In a review of apps for pelvic organ prolapse and urinary incontinence [16], 67.9% of apps incorporated an educational or informational component, but only 17.9% of apps clearly delineated the sources for their information. Thus, future apps should incorporate high-quality, evidence-based educational information.

Some apps incorporated the functionalities of physical therapy, online community, and reminder systems. Pelvic floor physical therapy is the primary treatment for PFD [45,46]. Apps with physical therapy functions should be connected with pelvic floor therapy devices. However, the effects of physical therapy

XSL•FO
RenderX

in apps remain unknown. Online communication enabled users and health care providers to share health information, emotions, and experiences [47]. In this study, only 1 app enabled communication between health care providers and users. We found that the reminder feature was offered in 6 apps. The reminder function of mHealth apps could improve users' adherence to treatment [48]. Future apps should incorporate a variety of functionalities to achieve the prevention and management of PFD.

Limitations and Future Work

This study systematically searched for and analyzed the publicly available mHealth apps in 3 different app stores. Meanwhile, we used a validated instrument, uMARS, to measure app quality. However, there were several limitations in this study. First, we only included apps in app stores that were freely accessible to the public. Paid apps were excluded due to a lack of funding. Second, we did not include the apps reported in the published literature, as many of these apps could not be downloaded or executed from the app stores. This could lead to selection bias, as apps reported in the literature may have higher quality and stronger evidence of effectiveness. Third, we only included Chinese-language apps due to limited funding and human resources. Thus, future studies should systematically summarize Chinese and other language apps from app stores and literature. Fourth, an important limitation was that we only provided descriptive data on uMARS scores and functionality counts. Health care professionals should explore the actual effects of apps and report real-world user feedback data in clinical settings. Finally, this study was limited in that the evaluation of apps relied on self-reported data by developers (eg, app descriptions, download counts, and self-rating scores) or users (eg, uMARS scores).

Conclusions

This study analyzed the content and quality of 46 PFD-related apps in 3 Chinese app stores. The PFD-related apps demonstrated acceptable quality, and the majority provided functions such as health education, exercise, and personal information recording. However, many apps lacked comprehensive functions for PFD prevention and management, as well as immediate feedback and high-quality educational information. Health care providers are encouraged to adhere to international guidelines to develop high-quality, evidence-based multifunctional apps for the prevention and management of PFD. Future studies should investigate the effects of these apps and incorporate real-world user feedback data in clinical settings.

Acknowledgments

The authors thank the academic editor, peer reviewers, and copy editor, who supported this study. The authors also thank their research assistants for their assistance.

Funding

The authors declared no financial support was received for this work.

Authors' Contributions

Conceptualization: YS, XD
Data curation: SF
Formal analysis: YS, SF
Methodology: YS, TH
Resources: TH
Supervision: YS, XD, LX
Validation: YS, XD, LX
Visualization: YS, LX
Writing – original draft: YS, TH, SF
Writing – review & editing: YS, XD

Conflicts of Interest

None declared.

Multimedia Appendix 1
Mobile App Rating Scale scoring of the included apps.
[DOCX File, 22 KB - jmir_v28i1e80126_app1.docx]

References

1. Pelvic floor dysfunction: prevention and non-surgical management. : National Institute for Health and Care Excellence (NICE); 2021 URL: https://www.nice.org.uk/guidance/ng210/resources/pelvic-floor-dysfunction-prevention-and-nonsurgical-management-pdf-66143768482501 [accessed 2026-05-29]

XSL•FO
RenderX

2. Okeahialam NA, Dworzynski K, Jacklin P, McClurg D, Guideline Committee. Prevention and non-surgical management of pelvic floor dysfunction: summary of NICE guidance. BMJ 2022 Jan 6;376:n3049. [doi: 10.1136/bmj.n3049] [Medline: 34992080]
3. Cosgriff L, Ramanathan A, Iglesia CB. Pelvic floor disorders and sexual function: a review. Obstet Gynecol Clin North Am 2024 Jun;51(2):241-257. [doi: 10.1016/j.ogc.2024.02.001] [Medline: 38777481]
4. Peinado-Molina RA, Martínez-Vázquez S, Hernández-Martínez A, Martínez-Galiano JM. Influence of pelvic floor disorders on sleep quality in women. J Pers Med 2024 Mar 20;14(3):320. [doi: 10.3390/jpm14030320] [Medline: 38541062]
5. Good MM, Solomon ER. Pelvic floor disorders. Obstet Gynecol Clin North Am 2019 Sep;46(3):527-540. [doi: 10.1016/j.ogc.2019.04.010] [Medline: 31378293]
6. Pang H, Zhang L, Han S, et al. A nationwide population-based survey on the prevalence and risk factors of symptomatic pelvic organ prolapse in adult women in China - a pelvic organ prolapse quantification system-based study. BJOG 2021 Jul;128(8):1313-1323. [doi: 10.1111/1471-0528.16675] [Medline: 33619817]
7. Zhu J, Pang H, Wang P, et al. Female urinary incontinence in China after 15 years' efforts: results from large-scale nationwide surveys. Sci Bull (Beijing) 2024 Oct 30;69(20):3272-3282. [doi: 10.1016/j.scib.2024.04.074] [Medline: 39214741]
8. Peinado-Molina RA, Hernández-Martínez A, Martínez-Vázquez S, Rodríguez-Almagro J, Martínez-Galiano JM. Pelvic floor dysfunction: prevalence and associated factors. BMC Public Health 2023 Oct 14;23(1):2005. [doi: 10.1186/s12889-023-16901-3] [Medline: 37838661]
9. Pang H, Liu J, Liu Q, et al. Medical care-seeking attitude and behaviour for urinary incontinence and associated factors among Chinese adult women. Eur J Obstet Gynecol Reprod Biol 2025 Aug;312:114513. [doi: 10.1016/j.ejogrb.2025.114513] [Medline: 40544525]
10. Reed P, Osborne LA, Whittall CM, Emery S. Impact of patient motivation on compliance and outcomes for incontinence. Physiotherapy 2021 Dec;113:100-106. [doi: 10.1016/j.physio.2020.10.003] [Medline: 34563914]
11. Wang L, Li Y, Qi Z, Wang W. Barriers and facilitators of the implementation of the application of pelvic floor muscle training in patients with prostate cancer: a scoping review. Front Public Health 2023;11:1191508. [doi: 10.3389/fpubh.2023.1191508] [Medline: 37771836]
12. O'Connor M, Bowles KH. Telehealth and mHealth. Res Nurs Health 2021 Feb;44(1):3-4. [doi: 10.1002/nur.22101] [Medline: 33355936]
13. da Mata KRU, Costa RCM, Carbone É, et al. Telehealth in the rehabilitation of female pelvic floor dysfunction: a systematic literature review. Int Urogynecol J 2021 Feb;32(2):249-259. [doi: 10.1007/s00192-020-04588-8] [Medline: 33175229]
14. Pereira AP, Janela D, Areias AC, et al. Innovating care for postmenopausal women using a digital approach for pelvic floor dysfunctions: prospective longitudinal cohort study. JMIR mHealth uHealth 2025 Apr 2;13:e68242. [doi: 10.2196/68242] [Medline: 40173388]
15. Hou Y, Feng S, Tong B, Lu S, Jin Y. Effect of pelvic floor muscle training using mobile health applications for stress urinary incontinence in women: a systematic review. BMC Womens Health 2022 Oct 3;22(1):400. [doi: 10.1186/s12905-022-01985-7] [Medline: 36192744]
16. Karsalia M, Malik R. Evaluation of free mobile health applications for pelvic organ prolapse and urinary incontinence. Am J Surg 2022 Jan;223(1):187-193. [doi: 10.1016/j.amjsurg.2021.07.046] [Medline: 34391568]
17. Kasoff M, Alishahian L, Gimoto J, Steinhart A, Grimes CL, Pape DM. Mobile phone apps for pelvic floor disorders. Urogynecology (Phila) 2024 Dec 1;30(12):998-1004. [doi: 10.1097/SPV.0000000000001541] [Medline: 39159315]
18. Sudol NT, Adams-Piper E, Perry R, Lane F, Chen KT. In search of mobile applications for patients with pelvic floor disorders. Female Pelvic Med Reconstr Surg 2019;25(3):252-256. [doi: 10.1097/SPV.0000000000000527] [Medline: 29219859]
19. Liu M, Wu X, Li Z, Tan D, Huang C. Assessment of eye care apps for children and adolescents based on the mobile app rating scale: content analysis and quality assessment. JMIR mHealth uHealth 2024 Sep 13;12:e53805. [doi: 10.2196/53805] [Medline: 39269760]
20. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR mHealth uHealth 2015 Mar 11;3(1):e27. [doi: 10.2196/mhealth.3422] [Medline: 25760773]
21. Stoyanov SR, Hides L, Kavanagh DJ, Wilson H. Development and validation of the user version of the mobile application rating scale (uMARS). JMIR mHealth uHealth 2016 Jun 10;4(2):e72. [doi: 10.2196/mhealth.5849] [Medline: 27287964]
22. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
23. HarmonyOS surpasses ios market share in china for first time reports counterpoint research Q2 2025. Technetbook. URL: https://www.technetbooks.com/2026/01/harmonyos-surpasses-ios-market-share-in.html [accessed 2026-05-29]
24. Chinese mobile app market in 2025: key trends and major players. ASOMobile. URL: https://asomobile.net/en/blog/chinese-mobile-app-market-in-2025-key-trends-and-major-players/ [accessed 2026-05-29]
25. Zhang C, Li X, Li X, Wu H. Reliability and validity of the Chinese version of the user version of mobile application rating scale [Article in Chinese]. Chin J Mod Nurs 2020 Jul 6;26(19):2608-2612. [doi: 10.3760/cma.j.cn115682-20190806-02814]

XSL•FO
RenderX

26. Pelvic Floor Disease Professional Committee of Jiangsu Provincial Association of Integrated Chinese and Western Medicine, Anorectal Disease Professional Committee of Shanghai Medical Doctor Association, Beijing Association of Integrative Medicine Pelvic Floor Diseases Professional Committee, Shanghai Chinese Medicine Society Anorectal Branch. Expert consensus on diagnosis and rehabilitation treatment of pelvic floor dysfunction [Article in Chinese]. Chin J Clinicians 2024 Feb;18(2):113-121. [doi: 10.3877/cma.j.issn.1674-0785.2024.02.001]
27. Wu X, Xu L, Li P, Tang T, Huang C. Multipurpose mobile apps for mental health in Chinese app stores: content analysis and quality evaluation. JMIR mHealth uHealth 2022 Jan 4;10(1):e34054. [doi: 10.2196/34054] [Medline: 34982717]
28. Yang F, Wei J, Zhao X, An R. Artificial intelligence-based mobile phone apps for child mental health: comprehensive review and content analysis. JMIR mHealth uHealth 2025 Jun 6;13:e58597. [doi: 10.2196/58597] [Medline: 40479582]
29. Liu XH, Jin F, Hsu J, Li DN, Chen W. Comparing smartphone apps for traditional Chinese medicine and modern medicine in China: systematic search and content analysis. JMIR mHealth uHealth 2021 Mar 24;9(3):e27406. [doi: 10.2196/27406] [Medline: 33759786]
30. Yang S, Bui CN, Park K. Mobile health apps for breast cancer: content analysis and quality assessment. JMIR mHealth uHealth 2023 Feb 23;11:e43522. [doi: 10.2196/43522] [Medline: 36821352]
31. Toïgo M, Marc J, Hayot M, Moulis L, Carbonnel F. Quality assessment of smartphone medication management apps in France: systematic search. JMIR mHealth uHealth 2024 Mar 18;12:e54866. [doi: 10.2196/54866] [Medline: 38498042]
32. Musgrave LM, Kizirian NV, Homer CSE, Gordon A. Mobile phone apps in Australia for improving pregnancy outcomes: systematic search on app stores. JMIR mHealth uHealth 2020 Nov 16;8(11):e22340. [doi: 10.2196/22340] [Medline: 33196454]
33. Harper RC, Sheppard S, Stewart C, Clark CJ. Exploring adherence to pelvic floor muscle training in women using mobile apps: scoping review. JMIR mHealth uHealth 2023 Nov 30;11:e45947. [doi: 10.2196/45947] [Medline: 38032694]
34. Jaffar A, Muhammad NA, Mohd Sidik S, et al. Feasibility and usability of Kegel Exercise Pregnancy Training app (KEPT app) among pregnant women with urinary incontinence. Int J Environ Res Public Health 2022 Mar 17;19(6):3574. [doi: 10.3390/ijerph19063574] [Medline: 35329262]
35. Widdison R, Rashidi A, Whitehead L. Effectiveness of mobile apps to improve urinary incontinence: a systematic review of randomised controlled trials. BMC Nurs 2022 Jan 28;21(1):32. [doi: 10.1186/s12912-022-00812-6] [Medline: 35090464]
36. Latorre GFS, de Fraga R, Seleme MR, Mueller CV, Berghmans B. An ideal e-health system for pelvic floor muscle training adherence: systematic review. Neurourol Urodyn 2019 Jan;38(1):63-80. [doi: 10.1002/nau.23835] [Medline: 30375056]
37. Vilela IDC, Silva NMB, Pinto RMC, Driusso P, Pereira-Baldon VS. Effects of using a mobile application on pelvic floor training in women with stress urinary incontinence: a randomized controlled clinical study. Neurourol Urodyn 2024 Nov;43(8):1997-2004. [doi: 10.1002/nau.25522] [Medline: 38847315]
38. Deniz-Garcia A, Fabelo H, Rodriguez-Almeida AJ, et al. Quality, usability, and effectiveness of mHealth apps and the role of artificial intelligence: current scenario and challenges. J Med Internet Res 2023 May 4;25:e44030. [doi: 10.2196/44030] [Medline: 37140973]
39. Bouayad L, Ialynytchev A, Padmanabhan B. Patient health record systems scope and functionalities: literature review and future directions. J Med Internet Res 2017 Nov 15;19(11):e388. [doi: 10.2196/jmir.8073] [Medline: 29141839]
40. Lin J, Bates SM, Allen LN, Wright M, Mao L, Kidd M. Integrating mobile health app data into electronic medical or health record systems and its impact on health care delivery and patient health outcomes: scoping review. JMIR mHealth uHealth 2025 Jun 23;13:e66650. [doi: 10.2196/66650] [Medline: 40549960]
41. Yeung WK, Dawes P, Pye A, et al. eHealth tools for the self-testing of visual acuity: a scoping review. NPJ Digit Med 2019;2(1):82. [doi: 10.1038/s41746-019-0154-5] [Medline: 31453377]
42. Wu Y, Zhong Y, Xu C, Li H, Chen M, Xu L. Interventions for maintaining pelvic floor health during pregnancy: a systematic review. Physiother Res Int 2025 Jul;30(3):e70077. [doi: 10.1002/pri.70077] [Medline: 40492459]
43. Stephan AP, Hauc SC, Marks VA, Bercik R, Rickey L. TikTok misinformation and user engagement in female pelvic floor conditions. Neurourol Urodyn 2024 Nov;43(8):1956-1961. [doi: 10.1002/nau.25519] [Medline: 38828831]
44. Byambasuren O, Beller E, Glasziou P. Current knowledge and adoption of mobile health apps among Australian general practitioners: survey study. JMIR mHealth uHealth 2019 Jun 3;7(6):e13199. [doi: 10.2196/13199] [Medline: 31199343]
45. Adams SR, Dessie SG, Dodge LE, Mckinney JL, Hacker MR, Elkadry EA. Pelvic floor physical therapy as primary treatment of pelvic floor disorders with urinary urgency and frequency-predominant symptom. Female Pelvic Med Reconstr Surg 2015;21(5):252-256. [doi: 10.1097/SPV.0000000000000195] [Medline: 26313494]
46. Cabrera-Martos I, Cortés-Alcaraz C, Jiménez-López P, López-López L, Torres-Sánchez I, Díaz-Mohedo E. Physical therapist interventions to prevent postpartum urinary incontinence: a systematic review. Phys Ther 2025 May 3;105(5):pzaf017. [doi: 10.1093/ptj/pzaf017] [Medline: 40114296]
47. van der Graaf P, Burrows A, Park H, Sowden S. Developing an online knowledge sharing platform and community of practice for health professionals: experiences from C-WorKS developed in North East England and Yorkshire during COVID-19. Health Info Libr J 2025 Jun;42(2):177-189. [doi: 10.1111/hir.12519] [Medline: 38303128]
48. Ni Z, Liu C, Wu B, Yang Q, Douglas C, Shaw RJ. An mHealth intervention to improve medication adherence among patients with coronary heart disease in China: development of an intervention. Int J Nurs Sci 2018;5(4):322-330. [doi: 10.1016/j.ijnss.2018.09.003] [Medline: 31406843]

XSL•FO
RenderX

Abbreviations

mHealth: mobile health
PFD: pelvic floor dysfunction
PFMT: pelvic floor muscle training
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
uMARS: the user version of the Mobile App Rating Scale

Edited by A Stone; submitted 04.Jul.2025; peer-reviewed by C Clark, E Luhanga, M Aktas; revised version received 13.May.2026; accepted 13.May.2026; published 24.Jun.2026.

Please cite as:
Song Y, Deng X, Hu T, Feng S, Xing L
Chinese Mobile Health Apps for Preventing and Managing Pelvic Floor Dysfunction: Quality Assessment and Content Analysis
J Med Internet Res 2026;28:e80126
URL: https://www.jmir.org/2026/1/e80126
doi:10.2196/80126

© Yuqing Song, Xue Deng, Ting Hu, Sijie Feng, Lu Xing. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Long-Term Mobile-Based Glycemic Intervention for Secondary Prevention in Patients With Diabetes Undergoing Surgical Revascularization: Multicenter Randomized Controlled Trial

Yangwu Song[1*], MD; Yifeng Nan[1*], MD; Xieraili Tiemuerniyazi[1], MD; Ziang Yang[1], MD; Shicheng Zhang[1], MD; Xi Li[2], PhD; Wei Feng[1], MD, PhD

[1]Department of Cardiovascular Surgery, Fuwai Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Fuwai Hospital, Beilishi Road No. 167, Xicheng District, Beijing, China

[2]National Clinical Research Center for Cardiovascular Diseases, State Key Laboratory of Cardiovascular Disease, Fuwai Hospital, National Center for Cardiovascular Diseases, Fuwai Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Wei Feng, MD, PhD
Department of Cardiovascular Surgery, Fuwai Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Fuwai Hospital, Beilishi Road No. 167, Xicheng District, Beijing, China

Abstract

Background: Despite the growing amount of patients who underwent coronary artery bypass grafting (CABG) in low- and middle-income countries like China, their glucose control was suboptimal, likely due to poor adherence to healthy lifestyles and preventive medications. Mobile health tools facilitating secondary prevention seem promising, but evidence focusing on this high-risk population is scarce.

Objective: This study aimed to evaluate the significance of mobile health tools in long-term glycemic management for post-CABG patients with comorbid diabetes mellitus.

Methods: GUIDEME (glycemic control using mini program–based intervention in patients with diabetes undergoing coronary artery bypass to promote self-management) is a multicenter, open-label, closed-user group, randomized controlled trial, in which 1066 patients with diabetes who had recently undergone CABG were enrolled and allocated into 2 groups. Patients in the control group received conventional health education before discharge, whereas those in the intervention group additionally received automatic delivery of bite-sized health education and medication reminders through a smartphone app during the 6 months after discharge. The primary end point was a change in glycosylated hemoglobin (HbA_{1c}) from baseline to 6 months.

Results: Among the 1066 eligible participants enrolled, a total of 1038 (97.4%) had completed the follow-up, while 1000 (93.8%) had 6-month HbA_{1c} results available. Although only 79 (14.9%) patients in the intervention group were defined as active users, a greater reduction of HbA_{1c} in the intervention group was observed (adjusted between-group mean difference −0.13, 95% CI −0.25 to −0.01; *P*=.04). The intervention group also had a high proportion of good medication adherence (96.1% vs 93.2%, *P*=.04). There was no difference between the 2 groups regarding the secondary end points.

Conclusions: Health education and medication reminders based on smartphone app achieved a statistically significant but modest between-group difference in HbA_{1c}, the clinical relevance of which remains uncertain.

Trial Registration: ClinicalTrials.gov NCT04192409; https://clinicaltrials.gov/study/NCT04192409

(J Med Internet Res 2026;28:e72226) doi:10.2196/72226

KEYWORDS

cardiovascular disease; coronary artery disease; coronary artery bypass; diabetes mellitus; metabolic syndrome; mobile health; secondary prevention; glycemic control; glycohemoglobin; quality of life

Introduction

Coronary artery disease (CAD) is one of the leading causes of death globally [1]. It is noteworthy that in low- and middle-income countries, the prevalence and mortality of CAD have been growing dramatically [2]. Meanwhile, coronary artery bypass grafting (CABG) has been increasingly adopted as a beneficial therapy, accompanied by improvements in treatment capacity [3]. In China, for example, the overall volume of CABG has increased from 9802 in 2013 to 10,027 in 2018 [4]. However, substantial gaps exist in secondary prevention after CABG, which led to suboptimal prognosis. In particular, a larger

XSL•FO
RenderX

proportion of diabetes mellitus (DM) patients in China had failed to achieve ideal long-term glucose control [5], which may be caused by poor medication adherence and unhealthy lifestyles [6]. As demonstrated by a recent study focusing on Chinese patients after CABG, the adherence rate to secondary prevention medications was merely 50% at the 6-month postoperative follow-up [7]. Furthermore, a multicenter observational study reported that nearly 75% of patients with diabetes in China failed to achieve optimal glycemic control. Poor glycemic control was significantly associated with suboptimal medication adherence (odds ratio [OR] 1.63, 95% CI 1.08 - 2.46; *P*=.02) [8]. These findings highlight a substantial gap between guideline-recommended care and real-world practice in this high-risk population.

Mobile health (mHealth) leveraging mobile internet technologies is able to provide coherent support for health education and self-management to patients with chronic conditions [9]. The practical needs are greater for those facing geographic barriers [10], like in China, nearly one-fourth of the patients receiving CABG in 2021 were on interprovincial medical tours [11], which makes routine postdischarge care more difficult. However, the findings in prior studies about the effectiveness of mHealth tools on glycemic control were inconsistent [12-16], probably due to varied intervention approaches (mostly done through essays or short messages) [17-19] and different participants' characteristics (sex, age, and educational background) [20,21]. Despite the growing popularity of short-form videos, which have a much lower literacy threshold and can use fragmented time, the application of video-based health education is still scarce.

To close this knowledge gap, we implemented the GUIDEME (glycemic control using mini program–based intervention in patients with diabetes undergoing coronary artery bypass to promote self-management) study, a multicenter, open-label, randomized controlled trial (RCT). The hypothesis of this study is that video-based health education and medication reminders through the smartphone app can improve glycemic control in patients with diabetes who underwent CABG.

Methods

Study Overview

GUIDEME was a parallel, 1:1, open-labeled, RCT. The detail of the GUIDEME trial was published elsewhere [22]. The design of this trial was registered at National Library of Medicine (NCT04192409) [23]. Participants were enrolled during January 2020 and October 2021, and the last follow-up was completed in June 2022. This study was reported in accordance with the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) guidelines (see Checklist 1).

Ethical Considerations

This trial was approved by the Institutional Review Board of Fuwai Hospital (2019 - 1151) in June 2019. All participants provided written informed consent at the screening visit. Participants did not receive any financial compensation or incentives for participating in this study. All data were deidentified for protection of the patients' privacy. The study was conducted in accordance with the principles of the Helsinki Declaration.

Participants

Patients diagnosed with type 2 diabetes, who underwent isolated CABG, were deemed eligible. Patients were excluded if they (1) could not read the health educational materials, defined as a self-reported educational level below elementary school graduation or an inability to read Chinese characters fluently; (2) had cognitive or communication disorders; (3) had no ability to use or access to a smartphone, and (4) died before discharge.

Recruitment was conducted in 2 cardiac surgery centers simultaneously. In each center, patients who were hospitalized for CAD with DM and had undergone CABG were included. The diagnoses of CAD and DM, as well as the indications for surgical revascularization, were adjudicated centrally based on the patients' medical records. Once potentially eligible individuals were identified, study staff explained the study protocol to them face-to-face and invited them to participate. A written informed consent form containing detailed study information was also provided to the patients. Baseline characteristics data were recorded after the enrollment. The sex of the participants was determined by biological factor only.

Type 2 diabetes mellitus was diagnosed according to the Standards of Medical Care in Diabetes–2019. Diagnosis was based on the presence of classic symptoms of hyperglycemia (eg, polyuria, polydipsia, polyphagia, or unexplained weight loss) plus a random plasma glucose ≥11.1 mmol/L, fasting plasma glucose ≥7.0 mmol/L, a 2-hour plasma glucose of ≥11.1 mmol/L during an oral glucose tolerance test, or glycosylated hemoglobin (HbA_{1c}) level of ≥6.5%. All diagnoses were further reviewed and confirmed by the attending physician based on these criteria [24].

To ensure procedural consistency across the 2 centers, several measures were implemented prior to study initiation: (1) standardized operating procedures were established for patient recruitment, informed consent, baseline data collection, intervention delivery, and outcome assessment; (2) all participating staff, including nurses and follow-up coordinators, received centralized training using identical materials to ensure uniform implementation and follow-up procedures; and (3) key laboratory measurements, including HbA_{1c}, were analyzed using the same model of equipment (ADAMS A1c HA-8180 [ARKRAY Inc]) in certified laboratories at each center to minimize intercenter variability.

Allocation and Intervention

Participants were allocated to the intervention or control group in a 1:1 ratio using a centralized, computer-based minimization randomization system after enrollment. Minimization was performed based on age, sex, education level, medical insurance status, and history of myocardial infarction, with all factors assigned equal weight. To reduce predictability, a probabilistic element was incorporated, whereby 90% of allocations followed the minimization criterion and 10% were assigned at random. The first participant was allocated with equal probability to either group. The randomization sequence was generated and

XSL•FO
RenderX

managed by an independent statistician and was concealed from investigators and outcome assessors throughout the trial.

Participants in the control group received conventional health education before discharge, while those in the intervention group additionally received videos and text messages for health education and medication reminders through a smartphone app based on WeChat for 6 months after discharge. The intervention tool was a custom-built app specifically for this study, which uses a closed registration system. Upon obtaining consent and completing randomization to the intervention arm, an account was registered for the patient by study personnel during a face-to-face session, which also included training on app use.

In the intervention app, the educational materials (a cluster of 180 videos) were developed by a multidisciplinary team of cardiac surgeons, cardiologists, endocrinologists, psychologists, nurses, and public health researchers based on clinical guidelines [25-29]. The themes of the health education included (1) general knowledge about CAD and DM, (2) postoperative antiplatelet agents, (3) lipid-lowering therapy, (4) beta-blockers, (5) glucose monitoring and control, (6) blood pressure control, (7) smoking cessation, (8) cardiac rehabilitation, and (9) lifestyle recommendations such as weight loss, physical activity, and diet. The team used behavioral change techniques (BCTs) to develop short-form videos by setting goals, providing information on the consequences of behavior, providing feedback on performance, social support, and so on [30]. All videos were shorter than 2 minutes. The educational information was automatically pushed to patients in a preset order according to the time of enrollment. The frequency of delivery was set to once a day during 6 months.

In addition, participants in the intervention group could receive reminders about taking secondary prevention medications, including antiplatelet agents, beta-blockers, nitrates, lipid-lowering agents, oral hypoglycemic agents, and insulin. For each participant, the prescribed medications were entered at enrollment and can be modified afterwards.

Outcomes and Ascertainment

The primary outcome was the change in HbA_{1c} between the baseline and the 6-month follow-up. Secondary outcomes included changes in the proportion of patients achieving an HbA_{1c} less than 7%, systolic blood pressure (SBP), low-density lipoprotein cholesterol (LDL-C), fasting blood glucose (FBG), and major adverse cerebrovascular and cardiovascular events (MACCEs), which were defined as a composite of all-cause death, nonfatal myocardial infarction, stroke, repeat revascularization (defined as any subsequent coronary revascularization procedure, including percutaneous coronary intervention or repeat CABG, involving either target or nontarget vessels), and cardiac rehospitalization.

Self-reported medication adherence and quality of life (QoL) questionnaire were collected only at 6 months and analyzed as exploratory outcomes. User data of patients who were allocated to the intervention group were extracted at 6 months. Based on prior studies indicating that sustained engagement with mHealth interventions is typically observed in approximately 10% of users [31], participants in the top decile of app usage frequency were defined as “active users.” In this study, this corresponded to a threshold of 12 app launches or more, which also approximated engagement during at least half of the 24-week intervention period.

The follow-up was conducted every 3 months after enrollment, which was completed by telephone or in-person visit. Outcomes were collected and assessed at 6-month follow-up. HbA_{1c}, LDL-C, FBG, and SBP measured at outpatient clinics of the study sites were preferred. In study sites, HbA_{1c} was determined using a high-performance liquid chromatography technique with the same model of analyzer (ADAMS A1cHA-8180 [ARKRAY Inc.]) to ensure methodological consistency.

In the follow-up visits, if MACCEs were reported by the participants or their families, central adjudication based on the medical records was conducted. The trial was conducted in an open-label manner; however, laboratory personnel and staff responsible for follow-up outcome assessment were blinded to group allocation to minimize potential assessment bias. In the meantime, medication adherence was collected by a self-reported questionnaire. The QoL was measured using the short version of EQ-5D and EuroQol-Visual Analogue Scale. The questionnaire was administered either online or in person, according to the patient’s preference.

Statistical Analysis

A sample size of 820 was estimated to provide 80% power at the significance level of .05 to detect a 0.3% absolute difference in HbA_{1c} change, assuming a mean HbA_{1c} level of 7% (SD 1.4%) at baseline based on data from studies involving similar populations [19]. This sample size allowed for a 20% loss to follow-up during the study period. During the course of the study, the COVID-19 pandemic led to restrictions on interregional travel and reduced accessibility to follow-up visits, raising concerns about a higher-than-expected dropout rate. To mitigate the potential impact of these unforeseen circumstances and to preserve statistical power, the recruitment target was increased, resulting in the enrollment of 1066 participants.

Baseline patient characteristics are presented as means with SDs for continuous variables that are normally distributed or as median with IQR. Categorical variables are presented as numbers with percentages. Baseline comparisons for continuous variables between groups are performed using Student *t* test or a nonparametric test as appropriate. Categorical variables were analyzed using the chi-square test or Fisher exact test.

Available-case analysis was conducted for comparisons of primary and secondary outcomes. Additionally, for the missing follow-up data, a sensitivity analysis was performed based on the intention-to-treat principle using baseline observation carried forward for the imputation of missing data. Between-group comparisons for the continuous outcomes including the primary end point (change in HbA_{1c} at 6 mo) were performed using an analysis of covariance (ANCOVA) model adjusting for baseline HbA_{1c}. This analytic approach was prespecified in the study protocol. Results were reported as adjusted mean differences with corresponding 95% CIs and exact *P* values. Log-binomial regression was used to compare the change in proportions of

XSL•FO
RenderX

patients achieving HbA_{1c}≤7% between groups. The QoL questionnaire, including EQ-5D and EuroQol-Visual Analogue Scale, was compared by nonparametric test. Self-reported medication adherence was analyzed as a categorical outcome by chi-square test. Log-rank test was used to compare the occurrence of MACCEs over time.

Additionally, we performed prespecified subgroup analyses of primary outcomes by baseline age (<60 and ≥60 y), sex (male and female), area (urban and rural), educational level (≤12 and >12 y), and tertile of baseline HbA_{1c}, using ANCOVA.

To explore the factors associated with adherence to the intervention, logistic regression was conducted to identify the characteristics of frequent users in the intervention group. Variables with P<.10 in univariate regression were included in the multivariate regression.

All tests of significance were 2-tailed, with α set at .05. Analyses were performed using SPSS software (version 26.0.0.0; IBM Corp).

Results

Baseline and Perioperative Characteristic

Between January 1, 2020, and September 29, 2021, 1066 eligible participants were enrolled and assigned to the intervention (n=534) or control group (n=532). Recruitment was ended after the designated sample size was reached. Among the 1038 patients who completed the final follow-up, 1000 provided primary end point data, and our primary analysis was based on this population of 1000 participants.

In this overall cohort, the mean age was 60.9 (SD 8.2) years, and 210 out of 1038 (20.2%) were female; 338 out of 1038 (32.6%) patients underwent myocardial infarction. In total, 909 out of 1038 (87.6%) patients were educated below master's degree. The urban medical insurance was the most commonly used (420/1038, 40.5%). The baseline characteristics of participants were comparable between the 2 groups (Table 1). No significant differences were observed regarding perioperative characteristics or discharge prescriptions (Table S1 in Multimedia Appendix 1).

XSL•FO
RenderX

Table . Baseline clinical and social characteristics of patients included in the final analysis.

Variables	Control group (n=486)	Intervention group (n=514)
Female, n (%)	99 (20.4)	104 (20.2)
Age (y), mean (SD)	60.8 (8.1)	60.9 (8.3)
BMI (kg/m^2), median (IQR)	25.8 (23.8-27.7)	25.6 (23.9-27.7)
SBP[a] (mm Hg), median (IQR)	134 (122-148)	133 (122-144)
Cigarette and alcohol, n (%)	219 (45.1)	245 (47.7)
Hypertension, n (%)	338 (69.5)	356 (69.3)
Hyperlipidemia, n (%)	377 (77.6)	399 (77.6)
Prior stroke, n (%)	51 (10.5)	44 (8.6)
NYHA[b] class III or IV, n (%)	128 (26.3)	131 (25.5)
Triple-vessel disease, n (%)	449 (92.4)	472 (91.8)
LM[c] disease, n (%)	127 (26.1)	133 (25.9)
Preoperative medication, n (%)		
Beta blocker	441 (90.7)	451 (87.7)
ACEI[d] or ARB[e]	180 (37.0)	182 (35.4)
Statin	449 (92.4)	480 (93.4)
Nonstatin hypolipidemia agents	86 (17.7)	78 (15.2)
Hypoglycemic intervention, n (%)		
Oral medication	245 (50.4)	276 (53.7)
Insulin	206 (42.4)	192 (37.4)
None	35 (7.2)	46 (8.9)
Diabetic comorbidity, n (%)	16 (3.3)	18 (3.5)
Preoperative laboratory		
HbA_{1c}[f] (%), median (IQR)	7.5 (6.7-8.5)	7.5 (6.8-8.4)
HbA_{1c}≤7.0%, n (%)	177 (36.4)	178 (34.6)
Creatine (mmol/L), median (IQR)	83.7 (72.3-95.0)	82.4 (73.4-92.8)
FBG[g] (mmol/L), median (IQR)	7.1 (5.7-8.9)	6.9 (6.0-8.4)
LDL-C[h] (mmol/L), median (IQR)	2 (1.6-2.5)	2.0 (1.6-2.5)
Current residence, n (%)		
City	326 (67.1)	326 (63.4)
Rural area	160 (32.9)	188 (36.6)
Enrollment center, n (%)		
Fuwai Hospital	459 (94.4)	476 (92.6)
Qingdao Fuwai Hospital	27 (5.6)	38 (7.4)
Health insurance, n (%)		
Urban basic health insurance	197 (40.5)	206 (40.1)
Rural health insurance	117 (24.1)	127 (24.7)
Other health insurance	172 (35.4)	181 (35.2)
Education level, n (%)		
Below senior high school	216 (44.4)	234 (45.5)
Senior high school or specialty	209 (43.0)	218 (42.4)

XSL•FO
RenderX

Variables	Control group (n=486)	Intervention group (n=514)
Bachelor's degree or above	61 (12.6)	62 (12.1)

[a]SBP: systolic blood pressure.
[b]NYHA: New York Heart Association.
[c]LM: left main.
[d]ACEI: angiotensin-converting enzyme inhibitors.
[e]ARB: angiotensin receptor blocker.
[f]HbA_{1c}: glycosylated hemoglobin.
[g]FBG: fasting blood glucose.
[h]LDL-C: low-density lipoprotein cholesterol.

Follow-Up

After group allocation, 3 participants were unable to engage with the intervention due to postoperative complications, and 25 patients withdrew during follow-up. A comparison of baseline characteristics between participants who completed the study and those who withdrew was conducted, and the results showed that patients who withdrew had higher LDL-C levels (2.0 [1.6, 2.5] vs 2.3 [1.9, 2.8] mmol/L, *P*=.03), while other baseline characteristics are similar (Table S2 in Multimedia Appendix 1). Considering that the majority of dropouts occurred in the intervention group, we conducted a further analysis comparing the baseline characteristics of completers versus dropouts within this group. The results showed no significant difference (Table S3 in Multimedia Appendix 1). Finally, a total of 1038 patients completed the study (530 in intervention group vs 508 in control group) (Figure 1). Among them, 6 (0.6%) participants died during the follow-up (4 in the intervention group vs 2 in the control group). In the cohort, 1000 (96.4%) patients had available 6-month HbA_{1c} results and were included in primary and secondary analyses. Participants excluded from the primary analysis had higher baseline creatinine (median 92.1, IQR 76.3-106.0 mmol/L vs 83.0, IQR 72.9-94.0 mmol/L; *P*=.04) but did not differ in other major characteristics (Table S4 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 1. Flowchart of patients' recruitment. HbA_{1c}: glycosylated hemoglobin.

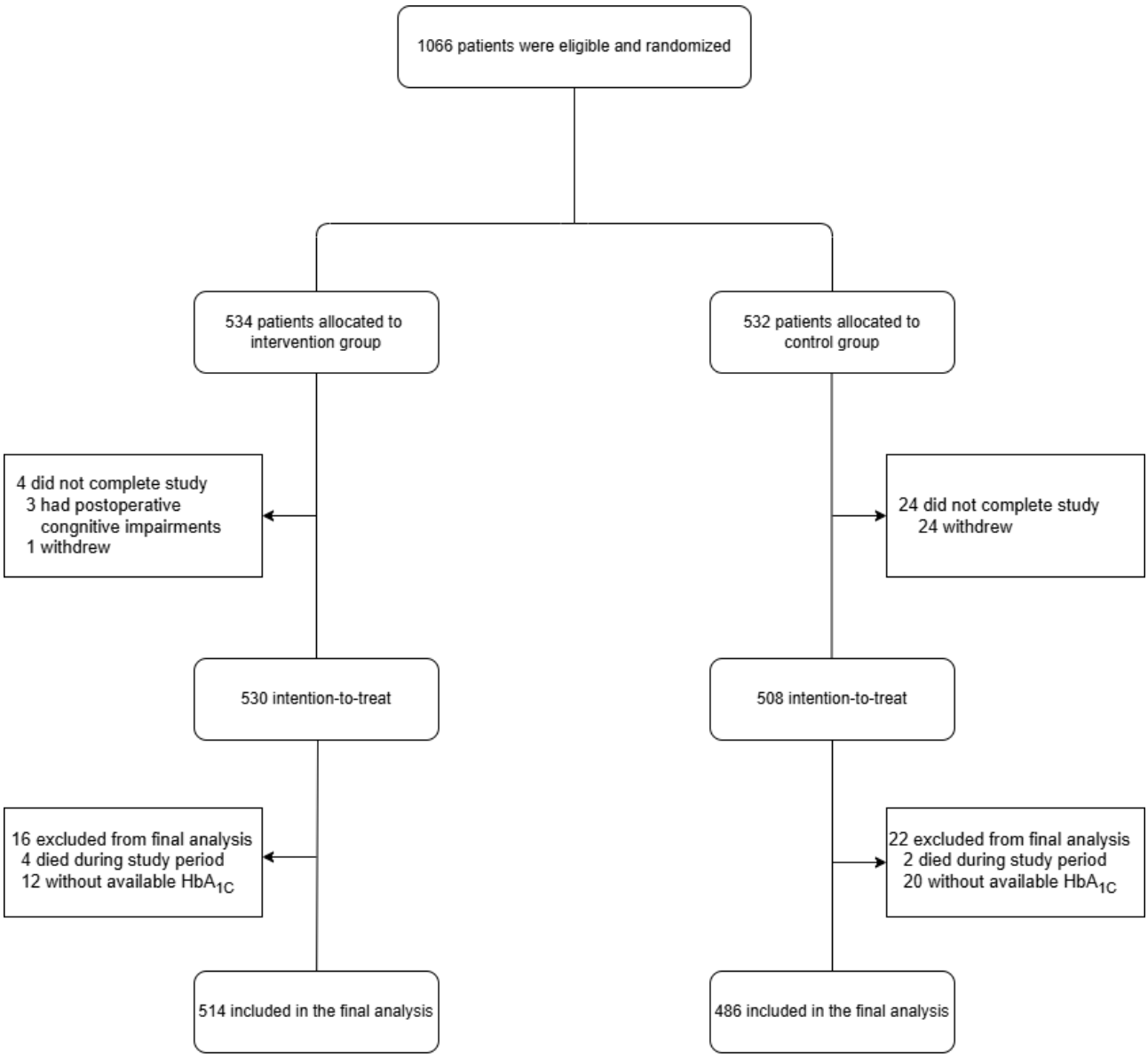

Outcome Assessment

The unadjusted within-group reductions in HbA_{1c} were −0.70% in the intervention group and −0.50% in the control group. A greater decrease from baseline to 6-month HbA_{1c} was attained in the intervention group; the adjusted between-group mean difference derived from the ANCOVA model was −0.13% (95% CI −0.25 to −0.01; *P*=.04; Tables 2 and 3). Given that an HbA_{1c} reduction of ≥0.5% is often regarded as clinically meaningful [27], we performed a post hoc exploratory analysis to assess this outcome. After adjustment for baseline HbA_{1c}, a greater proportion of patients in the intervention group achieved a clinically meaningful HbA_{1c} reduction of ≥0.5% compared with the control group (relative risk 1.14, 95% CI 1.01 - 1.28; *P*=.03; Table S6 in Multimedia Appendix 1). The finding of the primary outcome remains consistent in the intention-to-treat sensitivity analysis using baseline observation carried forward; the between-group adjusted mean difference remained statistically significant (−0.14%, 95% CI −0.25 to −0.02; *P*=.03; Table S5 in Multimedia Appendix 1). The number of patients with HbA_{1c} level of no more than 7% increased from 178 (34.6%) to 295 (57.4%) in the intervention group and from 177 (36.4%) to 256 (52.7%) in the control group. There was no significant difference between groups (relative risk 0.93, 95% CI 0.84 - 1.03; *P*=.18; Tables 2 and 3). The absolute risk difference was 4.7% (95% CI −1.6 to −11.0%), with a corresponding number needed to treat of 21.2, suggesting a trend favoring the intervention. Comparisons in secondary outcomes including SBP, FBG, and LDL-C showed no significant differences (Table 3).

XSL•FO
RenderX

Table . Primary and secondary end point at baseline, follow-up, and change from baseline.

Variables	Control group		Intervention group		Adjusted *P* value
	Follow-up	Change	Follow-up	Change	
HbA_{1c}[a] (%), median (IQR)	7.0 (6.5 to 7.7)	−0.3 (−1.3 to 0.4)	6.9 (6.4 to 7.5)	−0.6 (−1.5 to 0.2)	.04
HbA_{1c} ≤7.0%, n (%)	256 (52.7)	—[b]	295 (57.4)	—	.18
SBP[c] (mmol/L), median (IQR)	130 (120 to 136)	−6 (−19 to 6)	130 (120 to 135)	−5 (−16 to 8)	.54
FBG[d] (mmol/L), median (IQR)	7.0 (6.1 to 8.0)	0.1 (−1.5 to 1.4)	6.9 (6.1 to 7.8)	0 (−1.5 to 1.3)	.35
LDL-C[e] (mmol/L), median (IQR)	2.2 (1.7 to 2.8)	0.2 (−0.4 to 0.8)	2.2 (1.8 to 2.8)	0.2 (−0.4 to 0.8)	.46

[a]HbA_{1c}: glycated hemoglobin.

[b]Not available.

[c]SBP: systolic blood pressure.

[d]FBG: fasting blood glucose.

[e]LDL-C: low-density lipoprotein cholesterol.

Table . Mean difference in change of primary and secondary outcomes.

Variables	Mean difference in change, median (IQR)	Adjusted mean difference in change, median (IQR)	Adjusted *P* value
HbA_{1c}[a] (%)	−0.1 (−0.2 to 0)	−0.1 (−0.3 to −0.01)	.04
HbA_{1c} ≤7.0%	—[b]	0.9 (0.8 to 1.0)[c]	.18
SBP[d] (mmol/L)	−0.7 (−2.2 to 0.8)	−0.5 (−2.0 to 1.0)	.54
FBG[e] (mmol/L)	−0.2 (−0.4 to 0.1)	−0.1 (−0.4 to 0.1)	.35
LDL-C[f] (mmol/L)	0 (−0.1 to 0.1)	0 (−0.6 to 0.1)	.46

[a]HbA_{1c}: glycated hemoglobin.

[b]Not available.

[c]Reported as relative risk (95% CI).

[d]SBP: systolic blood pressure.

[e]FBG: fasting blood glucose.

[f]LDL-C: low-density lipoprotein cholesterol.

Compared with the control group, participants in the intervention group were more likely to report "hardly forget to take their medicine" (96.1% vs 93.2%; *P*=.047). The quality of life was comparable between the 2 groups. MACCE occurred in 22 individuals during the follow-up (12 in the intervention group vs 10 in the control group); the overall MACCE incidence was 4.29 per 100 person-years (3.98 in the control group vs 4.59 in the intervention group), with no difference observed between the groups (hazard ratio 1.15, 95% CI 0.50 - 2.67; *P*=.74; Table S6 in Multimedia Appendix 1).

Factors Related to the Effectiveness and Adherence of Intervention

There was no significant interaction between the intervention and the prespecified subgroups except for baseline tertiary HbA_{1c} and educational level ≤12 versus >12 years. None of the subgroups showed a significantly greater change in HbA_{1c} levels in the intervention group compared to the control group (Figure 2).

App usage data were extracted from 530 patients in the intervention group. The median total launch count was 5.5 (IQR 3.0-9.3) times, and 79 (14.9%) patients were active users of the app. Patients with a history of cigarette and alcohol use (OR 0.50, 95% CI 0.28-0.88; *P*=.02), with baseline HbA_{1c} ≤7% (OR 0.54, 95% CI 0.31-0.96; *P*=.04), and who lived in urban areas (OR 0.51, 95% CI 0.31-0.83; *P*=.01) were less likely to become active users of the intervention (Table S7 in Multimedia Appendix 1). The model fit was acceptable (likelihood ratio $\chi^2(5)=26.327$, *P*<.001, and the Hosmer-Lemeshow test $\chi^2(8)=5.440$, *P*=.61), and no evidence of multicollinearity was observed among the variables included in the multivariable model (all variance inflation factor <5.0; Table S7 in Multimedia Appendix 1).

XSL•FO
RenderX

Among participants in the intervention group, a significant association was observed between app usage frequency and the magnitude of HbA_{1c} reduction. Participants with higher numbers of app launches tended to exhibit greater decreases in HbA_{1c} (regression coefficient −0.09, 95% CI −0.11 to −0.06; P<.001). On average, an increase of 10 app launches was associated with an estimated −0.86% (95% CI −1.10 to −0.62) change in HbA_{1c}. This pattern is consistent with a dose-response–like association; however, given the observational nature of this analysis, the findings should be interpreted as associative rather than causal (Figure 3).

Subgroup analyses of app user data indicated heterogeneity in usage patterns. Participants in the lowest baseline HbA_{1c} tertile (Q1) demonstrated fewer app launches and a lower proportion of active users. Higher app engagement was more frequently observed among female participants and those residing in rural areas (Table S7 in Multimedia Appendix 1). Across most subgroups, participants classified as active app users experienced greater reductions in HbA_{1c} than nonactive users. Exceptions were observed in participants with college-level education or above and in those within the highest baseline HbA_{1c} tertile (Q3), in whom HbA_{1c} reductions were comparable between active and nonactive users (Table S8 in Multimedia Appendix 1).

Figure 2. Subgroup analyses of mobile intervention. HbA_{1c}: glycosylated hemoglobin.

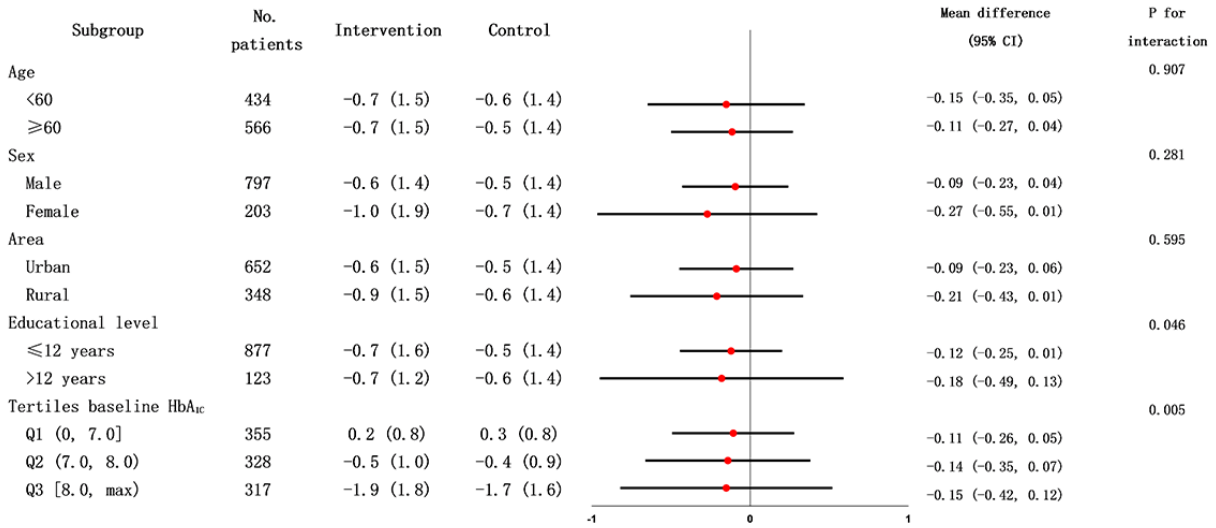

XSL•FO
RenderX

Figure 3. Dose-response analysis of the relationship between app adherence and the primary outcome. GAM: generalized additive model; HbA_{1c}: glycosylated hemoglobin.

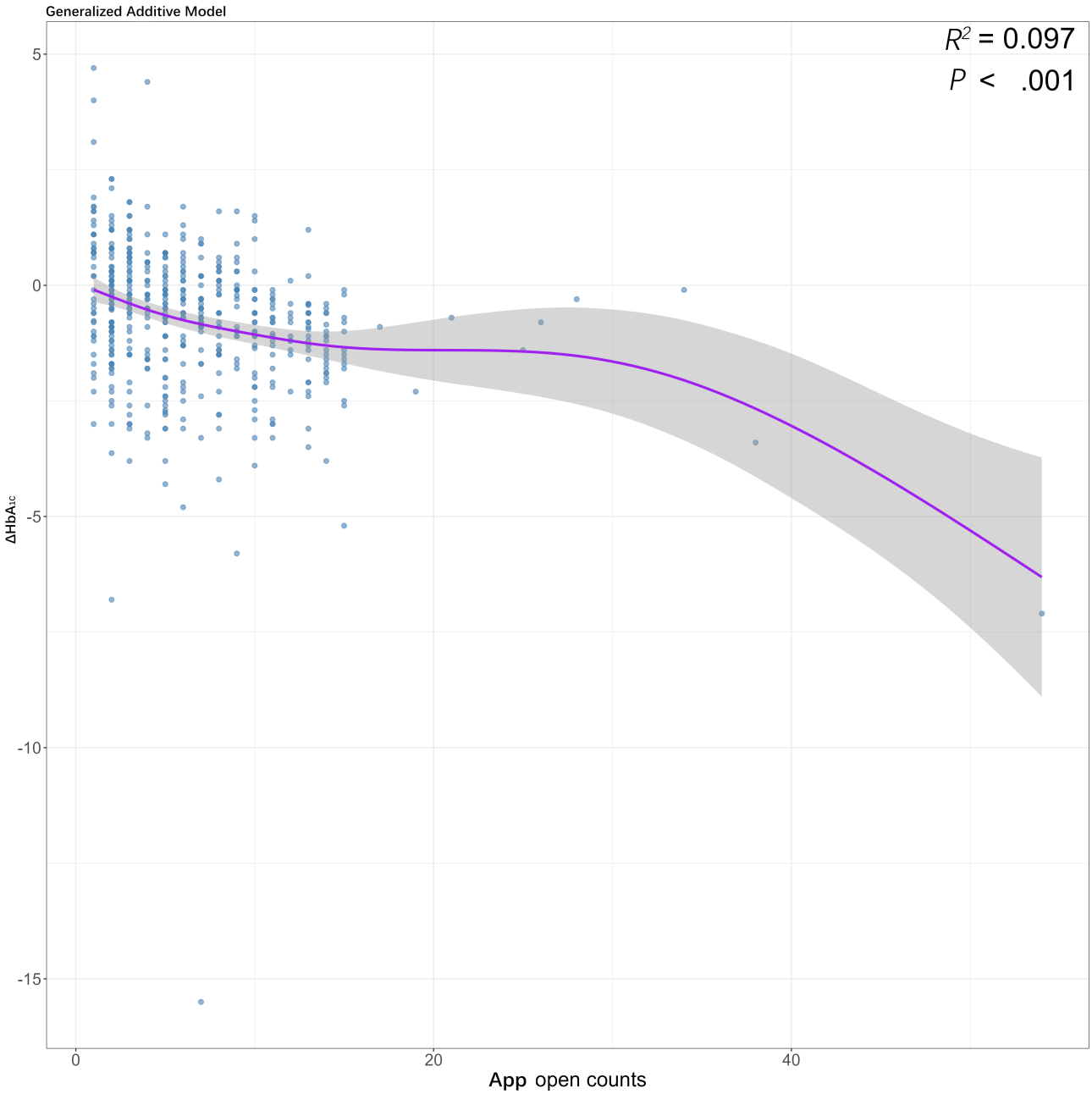

Discussion

Principal Findings

In this RCT among high-risk individuals who underwent CABG, short-form videos and text messages based on a smartphone app for health education and medication reminders seem to improve glycemic control. The reduction in HbA_{1c} was statistically greater in the intervention group; however, the magnitude of the between-group difference was modest. Although only 14.9% of patients in the intervention group remained active app users by the end of the study, the effectiveness of the intervention is positively correlated with the strength of app usage.

Comparison to Prior Work

The novelty of our study lies in its specific focus on the high-risk population of patients with diabetes undergoing CABG and its adoption of a video-centric approach for information delivery. Patients with diabetes undergoing CABG represent a particularly high-risk population due to the convergence of multiple clinical and physiological risk factors. Previous studies have consistently shown that diabetes and poor glycemic control are associated with increased mortality and a higher incidence of postoperative complications following CABG [32,33]. In the perioperative and early postoperative period, surgical stress, pain, inflammatory responses, and dietary restrictions contribute to stress-induced hyperglycemia through alterations in counter-regulatory hormones such as cortisol, catecholamines, and glucagon, thereby increasing glycemic variability and complicating glucose management [34]. Moreover, post-CABG

XSL•FO
RenderX

patients frequently present with multiple cardiometabolic comorbidities and require complex medication regimens, which further challenge sustained glycemic control and long-term adherence.

Consistent with these mechanisms, prior studies have demonstrated that elevated postoperative blood glucose levels are associated with increased risks of myocardial infarction, repeat revascularization, and long-term cardiovascular events [35,36]. Accordingly, major clinical guidelines recommend maintaining HbA_{1c} at ≤7% in this high-risk population as part of comprehensive secondary prevention strategies [25]. Although current guidelines for coronary revascularization do not explicitly recommend mHealth as an intervention, the importance of achieving optimal glycemic control as part of secondary prevention in cardiac rehabilitation for patients with coronary heart disease has been emphasized by the clinical guidelines [25,37]. Moreover, a meta-analysis by Jin et al supports the significant role of mHealth in enhancing secondary prevention outcomes in coronary heart disease [38].

While mHealth has demonstrated potential benefits in the management of type 2 diabetes, most existing interventions are not readily applicable to diabetic patients following CABG surgery. This limited applicability largely reflects the distinct clinical complexities of this population. First, post-CABG patients are typically required to manage multiple concurrent medications, including glucose-lowering agents, antiplatelet therapies, statins, and other cardiovascular drugs, which increases the risk of nonadherence and potential drug-drug interactions. Generic diabetes-focused mHealth apps rarely incorporate coordinated reminders or integrated management strategies for such complex medication regimens. Second, postoperative recovery after CABG extends beyond glycemic control to include wound care, pain management, graded physical rehabilitation, and surveillance for cardiovascular complications such as angina or heart failure. Effective secondary prevention in this context therefore requires an intervention capable of simultaneously addressing glycemic control and cardiovascular risk management, rather than focusing on glucose targets alone. Consistent with this, prior studies have shown that patients with diabetes after CABG often exhibit suboptimal achievement of multiple secondary prevention targets, including glucose, lipid, and blood pressure control [39].

We selected video as the core modality for delivering educational content based on several considerations. Patients with diabetes undergoing CABG are often older and may experience difficulties with text-based educational materials; in this context, video presentations with subtitles may offer greater accessibility. Previous mHealth interventions targeting post-CABG or similar high-risk populations have reported limited effectiveness when relying primarily on text-based approaches [31]. Although direct comparative evidence between video- and text-based interventions in this specific population is currently lacking, studies in other chronic disease populations suggest that supplementing text with video components may enhance engagement and intervention effectiveness [40]. By adapting the intervention format accordingly, we aimed to improve the acceptability and potential effectiveness of mHealth interventions for patients with diabetes following CABG.

Strengths and Limitations

As the first RCT on mHealth tools facilitating glucose control in high-risk patients who underwent CABG, this study extended the literature in several ways. First, the intervention group demonstrated a greater within-group reduction in HbA_{1c} (0.7%). This is consistent with a meta-analysis on mHealth intervention, which reported a mean reduction of 0.42% [41]. Although the relative increase in HbA_{1c} target attainment did not reach statistical significance, the absolute risk difference of 4.7% corresponds to a number needed to treat of approximately 21, suggesting a modest but potentially clinically relevant absolute benefit at the population level. Although a greater proportion of patients in the intervention group achieved a clinically meaningful reduction in HbA_{1c}, this finding arose from a post hoc exploratory analysis and should not be considered confirmatory; instead, it provides only tentative support that the intervention may have operated in the expected direction. It is also noteworthy that in this study, HbA_{1c} in the control group also decreased, probably because the surgery itself could act as a trigger for changing the attitude and behavior of patients. The effect has basically met the expectation—according to current guidelines, an intervention is considered effective when it leads to an absolute reduction of 0.3% or more [42]. Furthermore, the reduction in HbA_{1c} of 0.7% is estimated to reduce the risk of diabetes-related mortality by 14.7% based on available evidence [43]. However, the adjusted between-group effect size was modest (−0.13%) and below the prespecified clinically meaningful threshold of 0.3%. Therefore, the clinical significance of the observed improvement remains uncertain. It should also be noted that the number of MACCEs observed during follow-up was relatively small, which limited the statistical power to detect between-group differences; therefore, the absence of a statistically significant effect on MACCEs should not be interpreted as evidence of no effect.

The observed improvement in HbA_{1c} is likely attributable to the structured integration of multiple BCTs embedded in the intervention, rather than simple information delivery alone [44,45]. Specifically, techniques such as goal setting and planning, self-monitoring with feedback, and knowledge shaping may have enhanced patients' self-regulatory capacity for glycemic management, while automated medication reminders directly targeted medication self-management—a key determinant of glycemic control in this population. Together, these complementary BCTs provide a plausible behavioral mechanism underlying the observed glycemic benefit.

The greater improvement in glycemic control in the intervention group might be caused by better medication adherence. Undoubtedly, medication adherence is deemed one of the key links—but also the weak link—in secondary prevention, which is particularly true for glycemic management [46]. The prescription of secondary prevention medications increased significantly in patients with established CAD in China during the past 2 decades [47]. However, for long-term management, compliance would fall over time after discharge [48]. In

XSL•FO
RenderX

particular, patients who underwent CABG were found to be generally less likely to use secondary preventive medications consistently after the procedure compared with patients who received percutaneous coronary intervention [49]. Like in our trial, prior studies have also found similar results that mHealth interventions can improve medication adherence [50-55]. Most of the interventions in these studies were medication reminders, and some of them involved dedicated medication reminder devices or constant instruction from medical personnel [52,55]. The results of our study highlighted the importance of mHealth intervention focusing on medication adherence and suggested that multiple functions on that should be considered.

Notably, subgroup analysis revealed a significant interaction with baseline HbA_{1c} tertiles, indicating that patients with higher baseline HbA_{1c} derived greater benefit from the intervention. This finding is clinically intuitive, as patients with poor baseline glycemic control are more likely to exhibit pronounced deficits in self-management behaviors, leaving greater room for improvement. In contrast, patients with relatively well-controlled baseline HbA_{1c} may already have established management routines, limiting the incremental benefit of additional interventions. This interaction suggests that mHealth-based interventions may be particularly valuable when targeted toward patients with suboptimal baseline glycemic control.

We also observed that higher levels of engagement with the intervention app were associated with greater reductions in HbA_{1c}. In exploratory analyses restricted to participants in the intervention group, more frequent app use corresponded to larger improvements in glycemic control, yielding a dose-response–like pattern. Similar results were observed in a previous study [14]. However, this post hoc analysis was observational and may be influenced by residual confounding, self-selection, and reverse causality. Accordingly, these findings should be interpreted as hypothesis-generating rather than confirmatory and warrant validation in studies specifically designed to evaluate engagement-driven effects.

Previous studies have shown that it is challenging to maintain a high adherence to mHealth intervention. Evidence from reviews of chronic disease and cardiovascular health apps indicates that long-term active use rates barely reach 16.6% [56]. A prior trial (MISSION-2) reported that the use rate of the intervention app had dropped to less than 10% after 6 months [31], while its researchers believed that the poor compliance with the intervention app was one of the main reasons why their intervention had reached negative results. In GUIDEME, we had chosen short-form video, a rapidly expanding media type on mobile internet, as the major format of our education materials in the hope of achieving better compliance. Nevertheless, despite exploratory efforts to enhance engagement through short-form video delivery, overall user adherence in our study remained suboptimal, which may have attenuated the observed intervention effect and limited the internal validity and generalizability of the findings. Therefore, how to further enhance patients' interest and compliance with mHealth intervention and alter their behavior in receiving health education information will be a critical concern determining the value of similar interventions and need further exploration in the future. An interesting finding is that urban residents were less likely to become active app users, possibly due to easier access to in-person care, which reduces reliance on digital tools. In contrast, rural patients may adopt remote interventions more readily. This aligns with diffusion-of-innovation theory and suggests that implementation strategies should consider local health care contexts rather than assuming higher acceptance in urban areas. Taken together, these findings suggest that the app-based intervention should be viewed as a low-intensity, scalable support to routine post-CABG care, with a modest but measurable effect on glycemic control, rather than as a stand-alone therapeutic approach.

In contrast to the improvement observed in HbA_{1c}, no statistically significant between-group differences were detected for secondary outcomes, including SBP, FBG, and LDL-C. In addition to limitations related to follow-up duration and statistical power, these findings may also reflect differences in the relative emphasis of intervention components. Specifically, the app was primarily designed to support glycemic management and medication adherence, whereas fewer than 10% of the educational videos (16/180) directly addressed blood pressure or lipid management, and these focused on general disease knowledge rather than intensive or personalized management strategies. This content distribution may partly explain the absence of significant between-group differences in blood pressure and lipid outcomes. Similar patterns have been reported in previous mHealth studies targeting cardiometabolic risk factors [57]. Future studies may help clarify whether interventions with more targeted or intensive components could yield broader effects across multiple cardiometabolic outcomes.

Future Directions

The current study provided a feasible option to improve secondary prevention and relieve the public health burden of chronic conditions. The core function of our intervention was based on an automatic BCTs-based health education delivery system, mostly using short-form video to minimize threshold related to education levels. The system requires only an affordable initial investment to establish a video library for the target population and a scheduled delivery order. The operation of the system does not depend on massive medical manpower and resource investment. On the other hand, the smartphone-based system is scalable, which can effectively overcome the geographic and economic barriers and help numerous patients who have difficulties in accessing medical centers for long-term self-management at reasonable costs. We selected the app platform of WeChat (with over 1.2 billion monthly active users) rather than developing an independent app, which also suggests a feasible strategy to facilitate the popularizing of intervention tools.

Limitations

Several limitations should be acknowledged. First, limitations in participant flow documentation may affect the assessment of internal validity. Information from the screening phase was not systematically collected, and specific reasons for participant withdrawal or missing outcome data were not consistently recorded. This restricts a comprehensive evaluation of potential selection bias and attrition bias. Second, the subgroup and

XSL•FO
RenderX

interaction analyses were exploratory in nature. Sample sizes within certain strata were relatively small, which limits statistical power. Consequently, these findings should be interpreted as hypothesis-generating rather than confirmatory. Third, characteristics of the digital intervention itself may limit generalizability. Health education was delivered primarily through short-form videos. While this format enhances engagement and information density, it requires smartphones with sufficient performance and stable internet access. Such requirements may constrain feasibility in resource-limited or low-connectivity settings. Fourth, user-related digital barriers should be considered. Older individuals and patients unable to operate smartphones were excluded from participation, introducing potential selection bias and limiting applicability to populations with lower digital literacy. Fifth, medication adherence was assessed using self-reported measures rather than validated adherence scales or objective indicators. Self-reported adherence is subject to recall and social desirability bias and therefore provides limited measurement precision. Accordingly, adherence findings should be regarded as indicative and exploratory and are insufficient to support definitive causal inferences. Sixth, the definition of "active user" (≥12 app launches) was determined post hoc based on the observed distribution of app usage within this study and therefore involves a degree of subjectivity. Although this operational threshold was intended to identify participants with relatively higher engagement, it may not fully capture the multidimensional nature of adherence to digital interventions. The absence of standardized metrics for mHealth engagement limits comparability with other studies, and future research should adopt validated and harmonized measures of digital intervention adherence. Seventh, safety reporting focused on surgery- and disease-related clinical events. App-specific adverse effects—such as anxiety, privacy concerns, or data security issues—were not systematically assessed. Although health-related quality of life was measured using the EQ-5D, the absence of digital intervention-specific harm assessment may have led to underestimation of potential app-related adverse effects. Finally, the follow-up period was relatively short. As a result, the durability of the observed effects and the long-term impact of the mHealth intervention on clinical outcomes and engagement remain uncertain. Future studies with extended follow-up and iterative optimization of app functionalities are needed to address these limitations.

Conclusion

Although compared with conventional care, mHealth intervention featuring short-form video health education and medication reminders achieved greater HbA_{1c} decrease after CABG, the effect size was modest and did not reach commonly cited thresholds for clinical relevance; therefore, the clinical significance of this improvement should be interpreted with caution. This mobile-based intervention may serve as a feasible adjunct to conventional secondary prevention in patients with diabetes undergoing CABG. Further research on assessing the long-term effectiveness, testing personalized information push, and improving the app stickiness is needed.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: language editing, translation, and figure generation (TOC image). Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was supported by the National Key Research and Development Program from the Ministry of Science and Technology of China (grant 2018YFC1311201).

Data Availability

The data can be accessed upon reasonable request from the corresponding author.

Authors' Contributions

Conceptualization: WF, YS
Data curation: YN, XT, ZY, SZ
Formal analysis: YN
Funding acquisition: WF
Investigation: YN
Methodology: XL, YN, XT
Project administration: WF
Resources: WF
Supervision: WF
Validation: XL, YS
Visualization: YS
Writing – original draft: YN

XSL•FO
RenderX

Writing – review & editing: WF, XL, YS

The authors declare the use of generative AI in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision:
(No specific tasks selected)
Responsibility for the final manuscript lies entirely with the authors.
GAI tools are not listed as authors and do not bear responsibility for the final outcomes.
Declaration submitted by: Wei Feng

Conflicts of Interest

None declared.

Multimedia Appendix 1
The results of sensitivity analyses, exploratory analyses, and subgroup analyses supporting the main findings of the study.
[DOCX File, 57 KB - jmir_v28i1e72226_app1.docx]

Checklist 1
CONSORT-EHEALTH checklist.
[PDF File, 1147 KB - jmir_v28i1e72226_app2.pdf]

References

1. Khan MA, Hashim MJ, Mustafa H, et al. Global epidemiology of ischemic heart disease: results from the Global Burden of Disease Study. Cureus 2020 Jul 23;12(7):e9349. [doi: 10.7759/cureus.9349] [Medline: 32742886]
2. Ralapanawa U, Sivakanesan R. Epidemiology and the magnitude of coronary artery disease and acute coronary syndrome: a narrative review. J Epidemiol Glob Health 2021 Jun;11(2):169-177. [doi: 10.2991/jegh.k.201217.001] [Medline: 33605111]
3. Report on Cardiovascular Diseases in China: Encyclopedia of China Publishing House; 2017.
4. Li X, Gu D, Wang X, et al. Trends of coronary artery bypass grafting performance in a cohort of hospitals in China between 2013 and 2018. Circ Cardiovasc Qual Outcomes 2021 Apr;14(4):e007025. [doi: 10.1161/CIRCOUTCOMES.120.007025] [Medline: 33813854]
5. Wang L, Gao P, Zhang M, et al. Prevalence and ethnic pattern of diabetes and prediabetes in China in 2013. JAMA 2017 Jun 27;317(24):2515-2523. [doi: 10.1001/jama.2017.7596] [Medline: 28655017]
6. Han H, Cao Y, Feng C, et al. Association of a healthy lifestyle with all-cause and cause-specific mortality among individuals with type 2 diabetes: a prospective study in UK Biobank. Diabetes Care 2022 Feb 1;45(2):319-329. [doi: 10.2337/dc21-1512] [Medline: 34857534]
7. Ni Z, Dardas L, Wu B, Shaw R. Cardioprotective medication adherence among patients with coronary heart disease in China: a systematic review. Heart Asia 2019;11(2):e011173. [doi: 10.1136/heartasia-2018-011173] [Medline: 31297162]
8. Tao J, Gao L, Liu Q, et al. Factors contributing to glycemic control in diabetes mellitus patients complying with home quarantine during the coronavirus disease 2019 (COVID-19) epidemic. Diabetes Res Clin Pract 2020 Dec;170:108514. [doi: 10.1016/j.diabres.2020.108514] [Medline: 33068663]
9. Tian Y, Zhang S, Huang F, Ma L. Comparing the efficacies of telemedicine and standard prenatal care on blood glucose control in women with gestational diabetes mellitus: randomized controlled trial. JMIR mHealth uHealth 2021 May 25;9(5):e22881. [doi: 10.2196/22881] [Medline: 33783365]
10. Kao CK, Liebovitz DM. Consumer mobile health apps: current state, barriers, and future directions. PM R 2017 May;9(5S):S106-S115. [doi: 10.1016/j.pmrj.2017.02.018] [Medline: 28527495]
11. Ma W, Ma H, Wang Y, et al. Medical quality report of cardiovascular diseases in China: an executive summary [Article in Chinese]. Chin Circ J 2021;36(11):1041-1064. [doi: 10.3969/j.issn.1000-3614.2021.11.001]
12. Zhang L, He X, Shen Y, et al. Effectiveness of smartphone app–based interactive management on glycemic control in Chinese patients with poorly controlled diabetes: randomized controlled trial. J Med Internet Res 2019 Dec 9;21(12):e15401. [doi: 10.2196/15401] [Medline: 31815677]
13. Sun C, Sun L, Xi S, et al. Mobile phone–based telemedicine practice in older Chinese patients with type 2 diabetes mellitus: randomized controlled trial. JMIR mhealth uHealth 2019 Jan 4;7(1):e10664. [doi: 10.2196/10664] [Medline: 30609983]
14. Agarwal P, Mukerji G, Desveaux L, et al. Mobile app for improved self-management of type 2 diabetes: multicenter pragmatic randomized controlled trial. JMIR mHealth uHealth 2019 Jan 10;7(1):e10321. [doi: 10.2196/10321] [Medline: 30632972]
15. Dobson R, Whittaker R, Jiang Y, et al. Effectiveness of text message based, diabetes self management support programme (SMS4BG): two arm, parallel randomised controlled trial. BMJ 2018 May 17;361:k1959. [doi: 10.1136/bmj.k1959] [Medline: 29773539]

XSL•FO
RenderX

16. Zhong J, Zhang H, Li Z, et al. Effect of social app-assisted education and support on glucose control in patients with coronary heart disease and diabetes mellitus. Front Cardiovasc Med 2022;9:947130. [doi: 10.3389/fcvm.2022.947130] [Medline: 36211546]
17. Quinn CC, Shardell MD, Terrin ML, Barr EA, Ballew SH, Gruber-Baldini AL. Cluster-randomized trial of a mobile phone personalized behavioral intervention for blood glucose control. Diabetes Care 2011 Sep;34(9):1934-1942. [doi: 10.2337/dc11-0366] [Medline: 21788632]
18. Gong E, Baptista S, Russell A, et al. My Diabetes Coach, a mobile app–based interactive conversational agent to support type 2 diabetes self-management: randomized effectiveness-implementation trial. J Med Internet Res 2020 Nov 5;22(11):e20322. [doi: 10.2196/20322] [Medline: 33151154]
19. Huo X, Krumholz HM, Bai X, et al. Effects of mobile text messaging on glycemic control in patients with coronary heart disease and diabetes mellitus: a randomized clinical trial. Circ Cardiovasc Qual Outcomes 2019 Sep;12(9):e005805. [doi: 10.1161/CIRCOUTCOMES.119.005805] [Medline: 31474119]
20. Burner E, Menchine M, Taylor E, Arora S. Gender differences in diabetes self-management: a mixed-methods analysis of a mobile health intervention for inner-city Latino patients. J Diabetes Sci Technol 2013 Jan 1;7(1):111-118. [doi: 10.1177/193229681300700113] [Medline: 23439166]
21. Xu Y, Yue LZ, Wang W, Wu XJ, Liang ZY. Gender-specific impact of self-monitoring and social norm information on walking behavior among Chinese college students assessed using WeChat: longitudinal tracking study. J Med Internet Res 2021 Dec 7;23(12):e29167. [doi: 10.2196/29167] [Medline: 34878992]
22. Song Y, Nan Y, Feng W. Glycaemic control using mobile-based intervention in patients with diabetes undergoing coronary artery bypass-study protocol for a randomized controlled trial. Trials 2023 Sep 13;24(1):585. [doi: 10.1186/s13063-023-07580-x] [Medline: 37705074]
23. National Library of Medicine. URL: https://clinicaltrials.gov/ [accessed 2026-02-22]
24. American Diabetes Association. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes—2019. Diabetes Care 2019 Jan;42(Suppl 1):S13-S28. [doi: 10.2337/dc19-S002] [Medline: 30559228]
25. Kulik A, Ruel M, Jneid H, et al. Secondary prevention after coronary artery bypass graft surgery: a scientific statement from the American Heart Association. Circulation 2015 Mar 10;131(10):927-964. [doi: 10.1161/CIR.0000000000000182] [Medline: 25679302]
26. Grant PJ, Cosentino F. The 2019 ESC Guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD: New features and the 'Ten Commandments' of the 2019 Guidelines are discussed by Professor Peter J. Grant and Professor Francesco Cosentino, the Task Force chairmen. Eur Heart J 2019 Oct 14;40(39):3215-3217. [doi: 10.1093/eurheartj/ehz687] [Medline: 31608951]
27. American Diabetes Association. 6. Glycemic targets: standards of medical care in diabetes—2018. Diabetes Care 2018 Jan;41(Suppl 1):S55-S64. [doi: 10.2337/dc18-S006] [Medline: 29222377]
28. American Diabetes Association. 9. Cardiovascular disease and risk management: standards of medical care in diabetes—2018. Diabetes Care 2018 Jan;41(Suppl 1):S86-S104. [doi: 10.2337/dc18-S009] [Medline: 29222380]
29. American Diabetes Association. 4. Lifestyle management: standards of medical care in diabetes—2018. Diabetes Care 2018 Jan;41(Suppl 1):S38-S50. [doi: 10.2337/dc18-S004] [Medline: 29222375]
30. Schroé H, Van Dyck D, De Paepe A, et al. Which behaviour change techniques are effective to promote physical activity and reduce sedentary behaviour in adults: a factorial randomized trial of an e- and m-health intervention. Int J Behav Nutr Phys Act 2020 Oct 7;17(1):127. [doi: 10.1186/s12966-020-01001-x] [Medline: 33028335]
31. Yu C, Liu C, Du J, et al. Smartphone-based application to improve medication adherence in patients after surgical coronary revascularization. Am Heart J 2020 Oct;228:17-26. [doi: 10.1016/j.ahj.2020.06.019] [Medline: 32745732]
32. Kim YG, Park DW, Lee WS, et al. Influence of diabetes mellitus on long-term (five-year) outcomes of drug-eluting stents and coronary artery bypass grafting for multivessel coronary revascularization. Am J Cardiol 2012 Jun 1;109(11):1548-1557. [doi: 10.1016/j.amjcard.2012.01.377] [Medline: 22425329]
33. Pei Y, Ma Y, Xiang Y, et al. Stress hyperglycemia ratio and machine learning model for prediction of all-cause mortality in patients undergoing cardiac surgery. Cardiovasc Diabetol 2025 Feb 15;24(1):77. [doi: 10.1186/s12933-025-02644-5] [Medline: 39955587]
34. Crowley K, Scanaill P, Hermanides J, Buggy DJ. Current practice in the perioperative management of patients with diabetes mellitus: a narrative review. Br J Anaesth 2023 Aug;131(2):242-252. [doi: 10.1016/j.bja.2023.02.039] [Medline: 37061429]
35. Deo S, Sundaram V, Sheikh MA, et al. Pre-operative glycaemic control and long-term survival in diabetic patients after coronary artery bypass grafting. Eur J Cardiothorac Surg 2021 Nov 2;60(5):1169-1177. [doi: 10.1093/ejcts/ezab180] [Medline: 33970210]
36. Turgeon RD, Koshman SL, Youngson E, Pearson GJ. Association between hemoglobin A1c and major adverse coronary events in patients with diabetes following coronary artery bypass surgery. Pharmacotherapy 2020 Feb;40(2):116-124. [doi: 10.1002/phar.2359] [Medline: 31883378]
37. Commitee of Cardiac Rehabilitation and Prevention of Chinese Association of Rehabilitation Medicine. Guidelines for cardiovascular rehabilitation and secondary prevention in China 2018 simplified edition [Article in Chinese]. Zhonghua Nei Ke Za Zhi 2018 Nov 1;57(11):802-810. [doi: 10.3760/cma.j.issn.0578-1426.2018.11.003] [Medline: 30392235]

XSL•FO
RenderX

38. Jin K, Khonsari S, Gallagher R, et al. Telehealth interventions for the secondary prevention of coronary heart disease: a systematic review and meta-analysis. Eur J Cardiovasc Nurs 2019 Apr;18(4):260-271. [doi: 10.1177/1474515119826510] [Medline: 30667278]
39. Völler H, Reibis R, Pittrow D, et al. Secondary prevention of diabetic patients with coronary artery disease in cardiac rehabilitation: risk factors, treatment and target level attainment. Curr Med Res Opin 2009 Apr;25(4):879-890. [doi: 10.1185/03007990902801360] [Medline: 19254205]
40. Bezzubtseva MV, Demkina AE, Lipilina MN, et al. Video or text? Education through a social media website in hypertension. Int J Cardiol Cardiovasc Risk Prev 2022;14:200139. [doi: 10.1016/j.ijcrp.2022.200139] [Medline: 36060291]
41. Eberle C, Löhnert M, Stichling S. Effectiveness of disease-specific mHealth apps in patients with diabetes mellitus: scoping review. JMIR mHealth uHealth 2021 Feb 15;9(2):e23477. [doi: 10.2196/23477] [Medline: 33587045]
42. Burke LE, Ma J, Azar KMJ, et al. Current science on consumer use of mobile health for cardiovascular disease prevention: a scientific statement from the American Heart Association. Circulation 2015 Sep 22;132(12):1157-1213. [doi: 10.1161/CIR.0000000000000232] [Medline: 26271892]
43. Stratton IM, Adler AI, Neil HA, et al. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. BMJ 2000 Aug 12;321(7258):405-412. [doi: 10.1136/bmj.321.7258.405] [Medline: 10938048]
44. Cradock KA, ÓLaighin G, Finucane FM, et al. Diet behavior change techniques in type 2 diabetes: a systematic review and meta-analysis. Diabetes Care 2017 Dec;40(12):1800-1810. [doi: 10.2337/dc17-0462] [Medline: 29162585]
45. Cradock KA, ÓLaighin G, Finucane FM, Gainforth HL, Quinlan LR, Ginis KAM. Behaviour change techniques targeting both diet and physical activity in type 2 diabetes: a systematic review and meta-analysis. Int J Behav Nutr Phys Act 2017 Feb 8;14(1):18. [doi: 10.1186/s12966-016-0436-0] [Medline: 28178985]
46. Krass I, Schieback P, Dhippayom T. Adherence to diabetes medication: a systematic review. Diabet Med 2015 Jun;32(6):725-737. [doi: 10.1111/dme.12651] [Medline: 25440507]
47. Li S, Gao X, Yang J, et al. Number of standard modifiable risk factors and mortality in patients with first-presentation ST-segment elevation myocardial infarction: insights from China Acute Myocardial Infarction registry. BMC Med 2022 Jul 6;20(1):217. [doi: 10.1186/s12916-022-02418-w] [Medline: 35790971]
48. Xu YY, Chen WQ, Wang MX, et al. Lipid management in ischaemic stroke or transient ischaemic attack in China: result from China National Stroke Registry III. BMJ Open 2023 Mar 8;13(3):e069465. [doi: 10.1136/bmjopen-2022-069465] [Medline: 36889830]
49. Hlatky MA, Solomon MD, Shilane D, Leong TK, Brindis R, Go AS. Use of medications for secondary prevention after coronary bypass surgery compared with percutaneous coronary intervention. J Am Coll Cardiol 2013 Jan 22;61(3):295-301. [doi: 10.1016/j.jacc.2012.10.018] [Medline: 23246391]
50. Huang Z, Tan E, Lum E, Sloot P, Boehm BO, Car J. A smartphone app to improve medication adherence in patients with type 2 diabetes in Asia: feasibility randomized controlled trial. JMIR mHealth uHealth 2019 Sep 12;7(9):e14914. [doi: 10.2196/14914] [Medline: 31516127]
51. Li A, Del Olmo MG, Fong M, et al. Effect of a smartphone application (Perx) on medication adherence and clinical outcomes: a 12-month randomised controlled trial. BMJ Open 2021 Aug 9;11(8):e047041. [doi: 10.1136/bmjopen-2020-047041] [Medline: 34373299]
52. Brath H, Morak J, Kästenbauer T, et al. Mobile health (mHealth) based medication adherence measurement - a pilot trial using electronic blisters in diabetes patients. Br J Clin Pharmacol 2013 Sep;76 Suppl 1(Suppl 1):47-55. [doi: 10.1111/bcp.12184] [Medline: 24007452]
53. Frias J, Virdi N, Raja P, Kim Y, Savage G, Osterberg L. Effectiveness of digital medicines to improve clinical outcomes in patients with uncontrolled hypertension and type 2 diabetes: prospective, open-label, cluster-randomized pilot clinical trial. J Med Internet Res 2017 Jul 11;19(7):e246. [doi: 10.2196/jmir.7833] [Medline: 28698169]
54. Sartori AC, Rodrigues Lucena TF, Lopes CT, Picinin Bernuci M, Yamaguchi MU. Educational intervention using WhatsApp on medication adherence in hypertension and diabetes patients: a randomized clinical trial. Telemed J E Health 2020 Dec;26(12):1526-1532. [doi: 10.1089/tmj.2019.0305] [Medline: 32155382]
55. Yang Y, Lee EY, Kim HS, Lee SH, Yoon KH, Cho JH. Effect of a mobile phone-based glucose-monitoring and feedback system for type 2 diabetes management in multiple primary care clinic settings: cluster randomized controlled trial. JMIR mHealth uHealth 2020 Feb 26;8(2):e16266. [doi: 10.2196/16266] [Medline: 32130172]
56. Lou G, Li H, Zhang J, et al. Health management app use and influencing factors among patients with chronic disease in the community [Article in Chinese]. Chin J Prev Control Chronic Dis 2021;27(5):351-355 [FREE Full text]
57. Powers BJ, Olsen MK, Oddone EZ, Bosworth HB. The effect of a hypertension self-management intervention on diabetes and cholesterol control. Am J Med 2009 Jul;122(7):639-646. [doi: 10.1016/j.amjmed.2008.12.022] [Medline: 19559166]

Abbreviations

ANCOVA: analysis of covariance
BCT: behavioral change technique

XSL•FO
RenderX

CABG: coronary artery bypass grafting
CAD: coronary artery disease
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
DM: diabetes mellitus
FBG: fasting blood glucose
GUIDEME: glycemic control using mini program-based intervention in patients with diabetes undergoing coronary artery bypass to promote self-management
HbA_{1c}: glycosylated hemoglobin
LDL-C: low-density lipoprotein cholesterol
MACCE: major adverse cerebrovascular and cardiovascular event
mHealth: mobile health
OR: odds ratio
QoL: quality of life
RCT: randomized controlled trial
SBP: systolic blood pressure

Edited by A Stone, N Cahill; submitted 05.Feb.2025; peer-reviewed by JW Li, Y Lai; accepted 30.Jan.2026; published 30.Jun.2026.

Please cite as:
Song Y, Nan Y, Tiemuerniyazi X, Yang Z, Zhang S, Li X, Feng W
Long-Term Mobile-Based Glycemic Intervention for Secondary Prevention in Patients With Diabetes Undergoing Surgical Revascularization: Multicenter Randomized Controlled Trial
J Med Internet Res 2026;28:e72226
URL: https://www.jmir.org/2026/1/e72226
doi:10.2196/72226

© Yangwu Song, Yifeng Nan, Xieraili Tiemuerniyazi, Ziang Yang, Shicheng Zhang, Xi Li, Wei Feng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Acceptability of Technologies to Support Early Dementia Detection: Qualitative Study With the Boston University Alzheimer's Disease Center Cohort

Sarah Wilson[1], PhD; Emily Beswick[2], PhD; Zachary Popp[3], MPH; Salman Rahman[3], BSc; Sharandeep Bhogal[4], MSc; Tim Whitfield[4], PhD; Spencer Low[3], MPH; Raiyan Khan[3], BSc; Clare Tolley[1], PhD; Zuzana Walker[4], MD; Rhoda Au[3], PhD; Sarah P Slight[1], PhD

[1]School of Pharmacy, Newcastle University, King George VI Building, King's Road, Newcastle Upon Tyne, United Kingdom

[2]Trinity College Dublin, Dublin, Ireland

[3]Alzheimer's Disease Research Centre, Boston University, Boston, MA, United States

[4]University College London, London, United Kingdom

Corresponding Author:
Sarah P Slight, PhD
School of Pharmacy, Newcastle University, King George VI Building, King's Road, Newcastle Upon Tyne, United Kingdom

Abstract

Background: Dementia is on the rise globally due to increasing life expectancies and population growth. Digital technologies may help detect early signs, enabling timely interventions to slow or reverse cognitive decline. However, to support the successful implementation of these digital technologies into health care settings, they must be acceptable to target users. Older adults and those with mild cognitive impairment (MCI) are at risk of developing dementia in later life and need to be able to use these technologies in order for this intervention to be approved and implemented in clinical practice.

Objective: This study explored the perspectives of older adults and those living with a clinical diagnosis of MCI on the acceptability of using various digital technologies that have the potential to support early dementia detection.

Methods: Participants were recruited from Boston University's Alzheimer's Disease Research Center. Participants selected at least 2 technologies from 9 different wearables and software to use for 2 weeks, at 3-month intervals, over a total duration of 2 years. A subgroup of self-selecting participants was interviewed after the first 2 weeks of use to gather initial perspectives regarding the acceptability of using the digital technologies. An inductive framework thematic analysis approach was used, assisted by NVivo (version 14.23.2; QSR International).

Results: In total, 13 individuals living with a clinical diagnosis of MCI and 11 adults aged 65 years and older were interviewed. Our analysis identified five key themes: (1) gamification, (2) wearability, (3) user guidance, (4) burden of use, and (5) usefulness. Gamified apps were generally liked, although users with little experience of digital games needed time to adjust. Wearables resembling everyday accessories (eg, watches) were preferred, but complaints about tight or uncomfortable straps were frequently reported. Clear instructions were critical to support correct use, but many participants would have liked more troubleshooting support when technical issues arose. The use of 5 or more devices led to a high burden, especially when devices had practicality issues such as not being waterproof. Devices offering personal feedback were perceived as useful to satisfy personal interests, though some questioned their usefulness within health care. Participants raised concerns about losing valued personal interactions with health care professionals and questioned how their existing health conditions and treatment for such conditions may affect the validity of the data collected by the devices.

Conclusions: These findings can guide researchers in choosing appropriate devices and minimizing burden. Future work should explore the views of those experiencing digital exclusion to ensure equitable access to dementia-detection technologies.

(J Med Internet Res 2026;28:e84004) doi:10.2196/84004

KEYWORDS

digital technology; wearables; early detection; mild cognitive impairment; MCI; aging; qualitative research

Introduction

Over 57 million people lived with dementia in 2021 [1]. The number of new cases has been estimated to increase by nearly 10 million annually [1]. The most common dementia-causing disease is Alzheimer disease, contributing to 60% - 70% of cases [1]. Neuropathological features of dementia-causing diseases start to occur up to 2 decades before the appearance of

XSL•FO
RenderX

symptoms that are clinically detectable using current diagnostic techniques [2]. These diagnostic techniques often include a combination of pen-and-paper cognitive assessments to evaluate cognitive function, such as the Montreal Cognitive Assessment [3], neuroimaging techniques to detect potential structural changes associated with dementia, and health care professionals' judgment [4,5]. Researchers have explored the potential of cerebrospinal fluid or positron emission tomography imaging to detect biomarkers associated with the early signs of dementia-causing diseases [6]. However, these methods are invasive to patients and are not feasible to implement on a wide scale due to the considerable burden on health care resources (eg, financial cost, staff training or time, and scanning facilities) [7-9].

Over recent years, there has been significant interest in the use of digital technologies to support the early detection of dementia-causing diseases in a cost-effective, less invasive manner [9]. Digital technologies, such as wearable devices and software, have the potential to facilitate remote monitoring of individuals' cognitive abilities and function [9,10]. These different technologies can be used for Alzheimer disease and related dementia assessment, covering diverse populations such as Lewy body, vascular dementias, frontotemporal dementias, and all severities of Alzheimer disease [11]. The successful implementation of such technologies into health care has been suggested to rely heavily on end users' acceptance of using these digital technologies [12]. International public health agencies, such as the European Medicines Agency, the US Food and Drug Administration, and the National Institute for Health and Care Excellence, also require evidence of user involvement in the design and development of digital technology to ensure that patients understand how to use the digital technology and accept the design before the technology can be approved and implemented into clinical practice [13-15]. Those at risk of developing dementia later in life are likely to be the end users of such technologies. This includes older adults (aged 65 years and older) and those living with a clinical diagnosis of mild cognitive impairment (MCI) [16,17]. MCI is characterized by noticeable changes in memory or other cognitive abilities, such as the loss of language skills or visual or spatial perception, that are greater than what is expected with normal aging but not severe enough to interfere with daily life [18]. Although not everyone with a diagnosis of MCI progresses to dementia [19], those with an MCI diagnosis have a 5-fold increased risk of developing dementia later in life [17]. Therefore, to help identify acceptable digital technologies that could potentially support the early detection of dementia-causing diseases, we conducted a qualitative study to explore the perspectives of older adults and those living with MCI on the acceptability of using a range of different digital technologies.

Methods

Study Design

This study was part of the Early Detection of Neurodegenerative Diseases (EDoN) initiative, which aimed to develop a "digital toolkit" that contained technologies with the potential to detect early signs of dementia-causing diseases [9]. To support the development of the EDoN digital toolkit, we conducted semistructured interviews with potential end users (older adults and those living with MCI) to explore factors affecting concurrent acceptability (the extent to which an individual receiving an intervention finds it appropriate while participating [20]) of different digital technologies.

Participant Recruitment and Eligibility

Participants were recruited through an EDoN-affiliated research cohort at Boston University-Alzheimer's Disease Research Centre (BU-ADRC). The BU-ADRC is a National Institute of Health–funded center that provides resources to support Alzheimer disease research, including a network of individuals without cognitive decline, MCI, and dementia, all older than 50 years of age and willing to take part in research [21]. Individuals were eligible to take part if they owned a smartphone, had access to the internet at home, and were fluent in the English language.

Procedure

Participants were informed of the 9 different commercially available digital technologies to choose from, what the technologies did (ie, modalities they measured), and how to use them. Participants then chose a minimum of 2 technologies to use over the course of 3 months (Table 1). These digital technologies were selected by EDoN and BU-ADRC for the study, as they each measure different modalities associated with the early signs of dementia-causing diseases and thus have the potential to support early detection. These technologies included 3 gamified active smartphone apps (ie, involve an element of gameplay, such as following to achieve a goal), 2 passive apps (smartphone and computer), and 4 wearable technologies (2 wrist-worn activity monitors, a smart ring, and an electroencephalogram headband; the modalities each digital technology measured are presented in Table 1). A member of the research team (ZP and SR), who had received training on how to set up and use the technologies by the technology companies and developers, provided each participant with their selected digital technologies and supported the initial onboarding process during an in-person visit. The onboarding process involved activities such as helping participants download the relevant apps, create accounts, and pair the wearable devices with their personal smartphones. They also provided written instructions on how to use the digital technologies that participants took home and answered any questions about the technologies.

XSL•FO
RenderX

Table . Overview of all the digital technologies' metrics and duration of use.

Technology type	Method of assessment	Metrics evaluated	Frequency of data collection	Level of engagement required from participants
Active smartphone app 1	• Questionnaires and games	• Speech • Walking, gait, and balance • Memory • Reaction time	• Once every 3 months	30 - 35 minutes using the app
Active smartphone app 2	• Questionnaires and games	• Memory • Attention • Mental flexibility • Speed of processing	• Once every 3 months	10 - 15 minutes using the app
Active smartphone app 3	• Questionnaires and games	• Memory • Executive function • Mood and sleep • Subjective cognitive performance rating • Speech	• One task each day for 2 weeks • Two-week assessment period reoccurs every 3 months	10 - 15 minutes of using the app
Passive smartphone app	• Speed and pattern of typing on the phone	• Fine motor function	• Turn on for data collection once every 3 months for 2 weeks	Data collected passively for 2 weeks without user input
Passive computer software	• Speed and pattern of typing on the phone	• Fine motor function	• Turn on for data collection once every 3 months for 2 weeks	Data collected passively for 2 weeks without user input
Wrist-worn activity monitor 1 (can be alternated with the other wrist-worn activity monitor if the participant opted to use both devices)	• Inertial measurement unit	• Level of general activity • Step count • Sleep activity	• Repeat 2-week study period every 3 months	Wear for the duration of the 2-week study period (including overnight)
Wrist-worn activity monitor 2 (can be alternated with the other wrist-worn activity monitor if the participant opted to use both devices)	• Inertial measurement unit	• General activity • Sleep	• Repeat 2-week study period every 3 months	Wear for the duration of the 2-week study period (including overnight)
Headband	• Electroencephalogram to measure brain activity • Accelerometer	• Movement during sleep	• Repeat 2-week study period every 3 months	Wear overnight for a minimum of 3 consecutive nights, ideally a full 2 weeks
Ring (worn on the thumb or index finger)	• Oximeter • Heart rate monitor	• Heart rate • Oxygen in the bloodstream • Sleep quality and duration	• Repeat 2-week study period every 3 months	Wear overnight for a minimum of 2 consecutive nights, ideally a full 2 weeks

Data Collection

Participants completed a quantitative demographic questionnaire to gather relevant data on their sociodemographic background (eg, date of birth and biological sex [assigned at birth]) and medical history (eg, self-reported clinical diagnosis of MCI). Researchers (ZP or SR) shared information about this qualitative study with all participants. Researchers used the information provided in the questionnaire to monitor diversity across interested individuals to ensure that our sample consisted of both those living with a clinical diagnosis of MCI and those without any clinically diagnosed cognitive decline. A subgroup of self-selecting participants were invited to a remote (Zoom Technologies, Inc, or by telephone) semistructured interview after their first 2 weeks of using the digital technologies, occurring between April 2022 and June 2023. Interviews were conducted by a member of the research team (ZP or SR) who was trained by an experienced qualitative researcher (SW). All interviews were audio-recorded using the videoconference software or a dictaphone. A flexible interview topic guide (Multimedia Appendix 1), which was extensively piloted with 7 individuals with varying degrees of cognitive impairment and 2 caregivers, was used to guide and structure the discussions [22]. The topic guide was designed to gather participants' perspectives on the different digital technologies they had selected to use.

XSL•FO
RenderX

Data Analysis

Audio recordings of the interviews were transcribed verbatim by a transcription company (Datagain) and checked for accuracy by the research team (RK). Transcripts were uploaded to NVivo (version 14.23.2; QSR International) to assist our inductive thematic framework approach, a form of codebook thematic analysis [23,24]. Inductive thematic analysis is the process of identifying recurring concepts to create themes without using any preconceptions or preexisting framework [23]. This allowed us to focus on the views, perspectives, and experiences of participants without being influenced by any preexisting theories, ensuring that our analysis is an accurate reflection of the data [23]. The framework approach involved following 5 stages, beginning with members of the research team (SW and EB) familiarizing themselves with the data by reading and rereading the transcripts, before generating initial codes [23,24]. Core concepts (a central concept that holds associated shared meaning patterns, ie, themes) were then formed by grouping codes to create an analytical framework, which was applied across the whole dataset. A framework matrix was used to chart the data, consisting of rows (cases), columns (codes), and "cells" of data to help summarize the large quantity of data collected. This framework approach allowed multiple coders (SW and EB) to analyze the data and enabled our multidisciplinary team, with professional backgrounds in neuroscience, pharmacy, population health sciences, digital health inequities, physiology, and psychology, to review the analysis and interpret the data to provide a richer understanding of participants' views and experiences [23,24]. The rigidity of the framework approach also allowed us to explore specific project aims (to explore the perspectives of potential end users on the acceptability of using a range of digital technologies) [23,24]. The constant comparative method, involving the researchers (SW and EB) moving backward and forward between the data, comparing data, and evolving explanations, was used to further develop key themes [25]. Regular meetings with authors (SW, EB, SPS, ZP, SR, and CT) were held throughout the analysis to ensure that the themes were accurately reflected by the data and that data saturation (the stage where repetitive codes or themes are identified, and no new information or relationships between them emerge in new transcripts) [26] was reached before participant recruitment stopped. Several factors were considered to identify when thematic data saturation was achieved, such as the quality of the interviews (eg, length and depth of detail captured), number of interviews, and lack of occurrence of new information in interviews conducted after thematic data saturation was thought to be achieved [26-29].

Researcher Characteristics

The interviewers (ZP and SR) were male research assistants at Boston University School of Medicine at the time of this study and did not disclose their background to participants. There were no prior relationships between the participants and the research team. The interviewers were trained by a member of the team (SW) who had a master's in cognitive neuroscience and experience in interviewing older adults and those living with cognitive decline (eg, MCI and dementia). The researchers involved in the analysis (SW, EB, CT, and SPS) were experienced qualitative researchers at various career stages, with diverse academic backgrounds in pharmacy, neuroscience, and public health. This multidisciplinary approach ensured a comprehensive understanding of participants' experiences and mitigated bias from each researcher's interests.

Rigor and Trustworthiness

Various strategies were used to ensure that rigor was upheld during this qualitative research. This included following the SRQR (Standards for Reporting Qualitative Research) [30], completing the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist (Checklist 1) [31], and ensuring trustworthiness by following guidance outlined by Nowell et al [32] to achieve credibility, transferability, dependability, and confirmability of the study. To address credibility, we piloted the topic guide to ensure that the questions could be easily understood by participants. Transferability was achieved by providing a detailed description of the methods, allowing those who sought to transfer the findings to their own research or settings to judge it on a case-by-case basis. To achieve dependability, we used audit trails to clearly document logical decisions so others can understand how and why decisions were made. According to Lincoln and Guba [33], confirmability is established when credibility, transferability, and dependability are achieved.

Ethical Considerations

This study was approved by the institutional review board at Boston University Medical Center (IRB: H-40542). This study was conducted in accordance with the principles of the Declaration of Helsinki (2024). Written and verbal informed consent to participate in the study was obtained from all participants. Participants were not remunerated after taking part in the study. Written informed consent to publish the findings was obtained from relevant participants. Participants were fully informed of the anonymization process (eg, names were replaced with an identification number).

Results

Participant Characteristics

In total, 24 individuals participated in this qualitative study, of which 13 had a clinical diagnosis of MCI, and 11 had no cognitive impairment but were aged 65 years and older. Half were female (n=12), and the average age of the sample was 72 years (Table 2). The specific type and number of technologies used by each participant over the first 2-week period are shown in Table 2. Semistructured interviews lasted between 25 and 45 minutes.

XSL•FO
RenderX

Table . Participant demographics and devices used.

ID	Sex	Diagnosis	Digital technology used									Total number of technologies over the first 2-week period
			Apps					Wearables				
			Active app 1	Active app 2	Active app 3	Passive smartphone app	Passive computer software	Wrist-worn activity monitor 1	Wrist-worn activity monitor 2	Headband	Ring	
P1	Female	No cognitive impairment	✓	✓	✓	✓		✓		✓	✓	7
P2	Male	MCI[a]	✓	✓	✓	✓	✓	✓		✓	✓	8
P3	Male	No cognitive impairment	✓	✓	✓	✓		✓			✓	6
P4	Female	No cognitive impairment	✓	✓	✓			✓		✓	✓	6
P5	Female	No cognitive impairment	✓	✓	✓	✓		✓	✓	✓	✓	8
P6	Female	MCI	✓	✓	✓	✓		✓		✓	✓	7
P7	Male	No cognitive impairment	✓	✓	✓	✓	✓		✓		✓	7
P8	Female	No cognitive impairment	✓	✓	✓	✓		✓	✓	✓	✓	8
P9	Male	No cognitive impairment	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
P10	Female	No cognitive impairment		✓	✓	✓			✓	✓	✓	6
P11	Female	No cognitive impairment	✓	✓	✓	✓		✓		✓	✓	7
P12	Male	No cognitive impairment	✓	✓	✓	✓		✓	✓	✓	✓	8
P13	Male	MCI	✓	✓	✓	✓		✓	✓	✓	✓	8
P14	Male	MCI	✓	✓	✓	✓	✓	✓			✓	7
P15	Male	MCI	✓	✓	✓	✓	✓					5
P16	Male	MCI	✓	✓	✓	✓	✓	✓			✓	7
P17	Female	MCI						✓			✓	2
P18	Female	MCI	✓			✓		✓			✓	4
P19	Female	MCI	✓	✓		✓		✓		✓	✓	6
P20	Male	MCI	✓	✓	✓	✓	✓	✓		✓	✓	8

XSL•FO
RenderX

ID	Sex	Diagnosis	Digital technology used									Total number of technologies over the first 2-week period
			Apps					Wearables				
			Active app 1	Active app 2	Active app 3	Passive smartphone app	Passive computer software	Wrist-worn activity monitor 1	Wrist-worn activity monitor 2	Headband	Ring	
P21	Female	No cognitive impairment	✓	✓	✓	✓		✓		✓	✓	7
P22	Male	MCI	✓	✓	✓	✓		✓		✓	✓	7
P23	Female	MCI	✓	✓		✓		✓			✓	5
P24	Male	MCI	✓	✓	✓	✓	✓	✓		✓	✓	8

[a]MCI: mild cognitive impairment.

Our data analysis generated 5 key themes, which describe factors affecting the concurrent acceptability of a range of different digital technologies. These themes include (1) gamification, (2) wearability, (3) user guidance, (4) burden of use, and (5) usefulness (Figure 1). Each theme is described in detail below using direct anonymized quotes from participants; these appear in quotation markers (“”). To improve the flow and clarity, the author has added additional information to some quotes, indicated by square parentheses ([]), and removed some words, indicated by an ellipse in square parentheses ([…]). To ensure confidentiality, participant names have been replaced with their participant ID, which was assigned in chronological order, alongside additional relevant information to provide further context (eg, Participant 1, no cognitive impairment, used 7 technologies=participant who was interviewed first, did not have any diagnosed cognitive impairment, and used 7 of the 9 digital technologies available in this study). We take each of these themes in turn below.

Figure 1. Summary of key themes and subthemes.

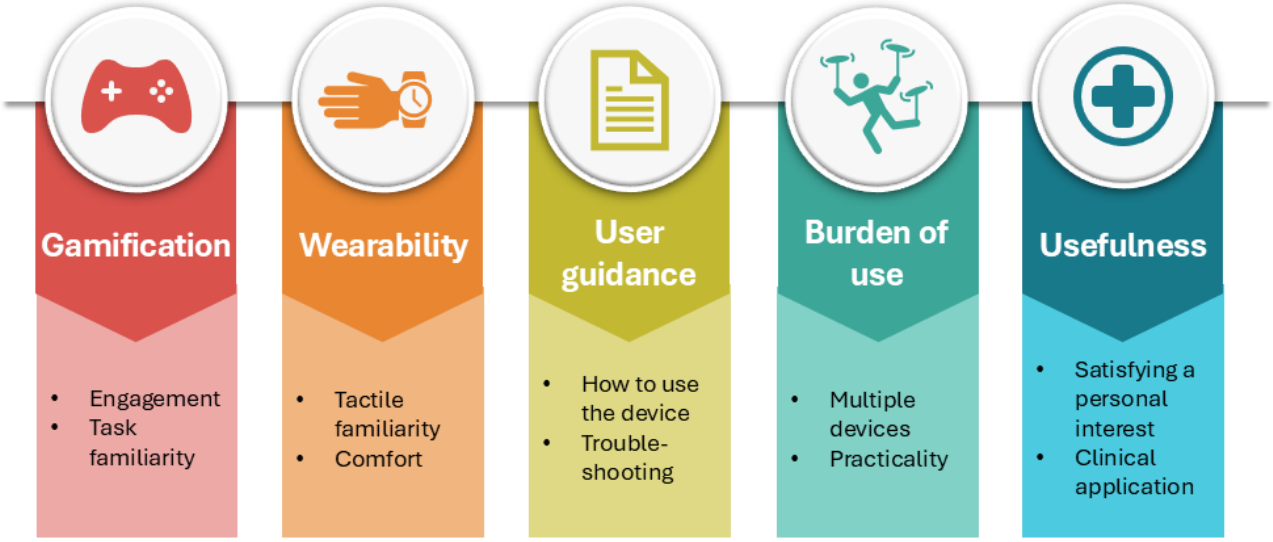

Gamification

Many participants living with MCI enjoyed the challenge of games within the active apps and felt that they were “keeping their brain alive” (Participant 23, living with MCI, used 5 technologies). Some participants without cognitive impairment who described themselves as competitive “wanted to get an A-Plus all the time” (Participant 8, no cognitive impairment, used 8 technologies). This participant described developing strategies in an attempt to improve their performance, such as verbalizing which direction they needed to swipe in the swiping task: “I think I did quite a bit better when I realized that saying it [the direction] out loud reinforced it” (Participant 8, no

XSL•FO
RenderX

cognitive impairment, used 8 technologies). There was also a strong preference among those without cognitive impairment to select the active apps, which had “a little bit more variety” (Participant 7, no cognitive impairment, used 7 technologies) in the gamified tasks. This helped participants avoid the games becoming repetitive and boring. However, some older adults without cognitive impairment who had “never played video games” (Participant 4, no cognitive impairment, used 6 technologies) before felt that their lack of experience may have impacted their ability to complete the gamified task, as it took longer to understand the requirements: “I don’t play video games. So, it took me a while [to understand what to do]” (Participant 21, no cognitive impairment, used 7 technologies). When participants did not perform as well as they would have liked on the gamified tasks, they described feeling anxious at the thought of having to play the game again, reducing their motivation to use the app again.

Wearability

Independent of cognitive decline, participants found the smart ring “easy [to use and] no big deal [to wear]” (Participant 11, no cognitive impairment, used 7 technologies). Similarly, most participants found the wrist-worn activity monitors acceptable because they were “just like wearing a watch, you know I don’t even know it’s there, except for the buzzes. So that’s fine” (Participant 4, no cognitive impairment, used 6 technologies). However, those who did not usually wear a watch discussed feeling the need to get “back in that habit” (Participant 1, no cognitive impairment, used 7 technologies), as they were “just not used to it” (Participant 9, no cognitive impairment, used 9 technologies). Many participants expressed hesitation around putting on the headband, as “it’s not the normal thing that I would like to be doing to go to sleep” (Participant 9, no cognitive impairment, used 9 technologies). Those who had used a medical wearable device on their head prior to this study were particularly concerned about the comfort of wearing it:

> *I had tried a sleep apnoea mask before, and I couldn’t do that [sleep with the mask on], and I didn’t like it. I thought this [the headband] was going to be similar.* [Participant 22, living with MCI, used 7 technologies]

In terms of comfort, some participants did not like the plastic straps on the wrist-worn activity monitor because “occasionally it got a little bit uncomfortable and sweaty, and if it’s too tight, it’s hard to move on my wrist” (Participant 5, no cognitive impairment, used 8 technologies). The headband also had some issues, as the electroencephalogram sensor prongs were felt to be “digging into the back of my head” (Participant 12, no cognitive impairment, used 8 technologies), causing discomfort and waking up participants during the night. This discomfort may have been linked to how tight the headband straps were adjusted, as one participant described how “the first night was terrible, but then I loosened [the straps], I used the extenders, so it wasn’t so tight and then it wasn’t quite so bad” (Participant 4, no cognitive impairment, used 6 technologies).

User Guidance

All participants felt that the user guides provided by the research team were “very helpful” (Participant 1, no cognitive impairment, used 7 devices) and “easy to follow” (Participant 3, no cognitive impairment, used 6 technologies). However, the guidance given within the active and passive apps as well as the apps associated with the wearable technologies (used to collect data) required further improvement to support those with MCI. One participant described having “trouble with the [headband] around my [her] head because the directions were not clear enough” (Participant 19, living with MCI, used 6 technologies). In this case, the participant had worn the headband the wrong way around, and the data were not recorded over multiple nights. Others explained how it was not obvious “when it [wrist-worn device] needed to be charged” (Participant 8, no cognitive impairment, used 8 technologies) and would like more guidance on this.

A key area omitted in the user guides was how to troubleshoot technical issues. One of the most frequently reported technical issues involved difficulty syncing the headband and its associated app. Some participants described receiving “a pop up on the app that said [...], now that you’ve paired, the headband will recognize the device, but it never did” (Participant 5, no cognitive impairment, used 8 technologies). Another participant recalled how the smartphone app often froze, resulting in him having to “hit the [app home] button again” (Participant 14, living with MCI, used 7 technologies) to close the app, reopen it, and repeat the games. He described feeling frustrated at being unable to complete the required task in one go and questioned whether he was “doing something [wrong like] touching that is causing this or is it something in the game itself that is too sensitive?” (Participant 14, living with MCI, used 7 technologies). Experiencing such technical issues without any guidance on how to overcome them left some participants feeling “helplessness when it [the technology] doesn’t work” (Participant 24, living with MCI, used 8 technologies).

Burden of Use

Participants with and without MCI described the burden of using 5 or more technologies, explaining how it was “a little invasive in terms of my time” (Participant 9, no cognitive impairment, used 9 technologies) and how they would have liked a shorter schedule of assessments so that it “would not wear me down” (Participant 21, no cognitive impairment, used 7 technologies). The burden of using the devices was particularly evident among participants who had “a lot going on in your [their] life” (Participant 6, living with MCI, used 7 technologies) or were on holiday during the 2-week period. One participant explained how they did not use the devices during their holiday as they “didn’t really want to have to do a lot of setting [up of the technology]” (Participant 11, no cognitive impairment, used 7 technologies). Some participants developed strategies to help “settle into some kind of pattern that was reasonable” (Participant 1, no cognitive impairment, used 7 technologies), with one participant making “a little chart of how many days I had to do this, that, and the next thing” (Participant 11, no cognitive impairment, used 7 technologies). This helped integrate the technology into their everyday lives.

Once a routine had been established, participants without cognitive impairment reflected on the practicality of using the devices. Some female participants were self-conscious of how the wearable devices looked on them, with one participant

XSL•FO
RenderX

explaining how “if I’m going to wear anything on my wrist, if I wanted to, [it] would be a pretty bracelet” (Participant 11, no cognitive impairment, used 7 technologies). Other participants, who liked water sports, described the wrist-worn activity monitors as "a bit of a hassle [...] I was out in the swimming pool every day because I had to take the [monitors] off” (Participant 12, no cognitive impairment, used 8 technologies). As the devices were only water-resistant, not waterproof, some participants worried that they may have accidentally damaged the devices when they exposed them to water. One participant recounted how they

> *hopped into the shower with it. And I immediately hopped out and wiped it, and everything. It was fine. But I was a little bit, “Oh God, what have I done now?”* [Participant 11, no cognitive impairment, used 7 technologies]

Usefulness

Participants found the digital technologies useful, particularly devices that allowed their scores on the gamified tasks to be compared over different days. Others living with a clinical diagnosis of MCI found that the health-related metrics (eg, heart rate) measured by the wrist-worn activity monitor were particularly useful. However, some participants living with MCI would have liked to interpret the feedback from the devices in further detail, particularly the sleep metrics: “I love looking at the feedback [sleep data from the headband] and, I wish I knew more about that so that was interesting” (Participant 6, living with MCI, used 7 technologies). Another participant without cognitive impairment felt that it would be useful to compare his metrics with others of a similar age:

> *It would be nice to be able to know that against this group of people [of similar age], you’re doing good, [or] you’re not doing good. So, if you need to change something [to improve your health and wellbeing], you could change it.* [Participant 7, no cognitive impairment, 7 technologies]

Most participants were hopeful that these digital technologies might be able to send reports directly to their health care provider. Others were more cautious and stated that they would “hate to see everything done high tech. I don’t want to lose the personal interaction between physicians and health care people and patients” (Participant 10, no cognitive impairment, used 6 technologies). Participants with existing health conditions worried about whether these conditions might impact the validity of the data collected by the digital technologies, with one explaining how she had a recent fall and “I think that they didn’t expect me to live, obviously I’ve done very well [with their health after the fall], but yes, I don’t think I’m as sharp [cognitively]” (Participant 10, no cognitive impairment, used 6 technologies). Some also questioned whether the data collected might be affected by treatment for a particular condition. For example, one participant who wore a sleep apnea mask at night queried how the oximeter sensor in the ring would “take into consideration that I’ve got oxygen on at night?” (Participant 10, no cognitive impairment, used 6 technologies).

Discussion

Principal Findings

This qualitative study explored the views of individuals potentially at risk of developing dementia, specifically older adults and those living with a clinical diagnosis of MCI, on the acceptability of using a range of different digital technologies. A range of 9 diverse technologies were used, and novel insights gained into the burden of using multiple devices and the key factors influencing acceptability. Most participants liked the gamified tasks in active apps, though some participants who were less familiar with digital games needed more time to adjust. Wearables placed on the body in areas of everyday accessories (eg, watches and rings) were preferred, but the straps used to fix the wearable to the body were sometimes found to be uncomfortable, particularly when worn too tightly. Clear, concise user instructions and troubleshooting guidance were found to be essential to ensure the proper use of the digital technologies, and further improvements were highlighted. Devices providing personally relevant feedback (eg, test scores or health metrics) were also well-received, although many questioned their usefulness within health care.

Gamification has great potential to increase participant engagement in cognitive tasks [34]. Findings from this study strongly support this stance, as gamification of cognitive testing through active smartphone apps was generally accepted by participants as challenging but also an engaging experience. Furthermore, a study conducted by Ríos Rincón et al [35] found that older adults preferred smartphone cognitive games compared to paper-based cognitive activities, and they experienced higher levels of engagement and positive cognitive effects while playing the smartphone games compared to paper-based activities. However, there are a number of considerations when using gamification in this manner. First, it is important to find the right balance between challenging the user with games, but also the goals within the games being achievable. Previous research has highlighted how if a digital game is too hard, users are more likely to experience low self-esteem, annoyance, and a sense of insecurity [36]. Second, an individual’s digital literacy skills, ability to use a hand-held digital device, and digital self-efficacy must be considered to avoid contributing toward digital health inequities once the technology is implemented into health care. This aspect is particularly important because individuals in this study were eligible to participate if they had owned a smartphone and had access to the internet at home. Some older adults have little to no experience of using games on a digital device, which they perceived had negatively impacted their ability to complete the task. The wider literature also suggests that older adults (aged 65 years and older) and those living with MCI are at risk of experiencing digital exclusion [37,38].

Where the device was worn on the body and the comfort of wearing the technology were 2 key elements found to affect the acceptability of a wearable device. Wrist-worn activity monitors and the smart ring were found to be more acceptable than the headband, which was consistent with previous findings [39-41]. This may be due to the increased familiarity and understanding

XSL•FO
RenderX

of smartwatches [42,43] and the social acceptability of wearing the device on your wrist and using your hands to make gestures and touches to the device [39]. Comfort of the wearable devices when worn correctly versus incorrectly (eg, too tight) was also found in this study to impact acceptance, which is echoed in the wider literature [44,45]. This highlights the importance of providing adequate and clear user instructions on how to wear and use the digital technologies.

User instructions must also contain guidance on how to overcome technical issues, such as how issues with syncing a wearable device to its associated smartphone app and what to do if an active smartphone app freezes. These instructions need to be written in such a way that they are suitable for those with cognitive impairment. Experiencing technical issues without any guidance on how to overcome them led to some participants in this study feeling hopeless. Law et al [46] found that professional caregivers of those with cognitive decline (health care professionals and staff at a retirement facility) were concerned that technical issues with digital technologies would cause confusion among those with cognitive impairments and would likely prevent them from engaging with the technology if the issues could not be fixed. Therefore, technical support must be given to end users and caregivers of those with cognitive impairment to promote acceptance, use, and engagement with the devices.

The burden placed on end users when using multiple devices was another key area that emerged in this study. Those who used 5 or more different digital technologies felt a high level of burden, and some found it challenging to integrate these devices into their busy daily lives. However, using digital technologies to support the early detection of dementia-causing diseases is viewed as a less invasive approach, but we found that the greater the number of devices that a participant used, the more it can potentially be invasive of their time. Furthermore, it brings into question if digital technology's ability to offer remote data collection and potentially shorter assessment periods only displaced the burden from health care systems onto end users rather than reduced the burden overall [47]. To reduce end-user burden, this study suggests that users should not be asked to use more than 4 devices at any 1 time and that the devices are waterproof to reduce the number of times the device needs to be removed. The wider literature also suggests selecting digital technologies that are easy to use and/or wear and devices that need minimal interaction (eg, have a long battery life to reduce charging) to reduce the number of tasks required [47,48].

The ability of the devices to provide personal feedback, such as test scores and health-related metrics (eg, heart rate), was perceived as useful and interesting by participants in this study. Such a finding is echoed in the literature, which also suggests that feedback from digital health technologies drives engagement and continued use of technologies among older adults [44,48]. Previous studies have also found that feedback from activity monitors may motivate older adults to increase their physical activity [39]. Despite the benefits of digital technologies in satisfying personal interests, many participants also raised concerns about potentially losing valued personal interactions with health care professionals and whether their existing health conditions and treatment may affect the validity of the data collected. Similar concerns about losing human connection in health care interaction have been reported in the wider literature [49]; however, there has been little to no research at the time of this study, discussing users' concerns about how their health conditions and medical treatments may affect the validity of the data. This novel insight must be further explored to understand to what extent health conditions and treatments are likely to impact the validity of the data collected and how to mitigate such concerns.

Strengths, Limitations, and Areas for Future Research

Multiple strategies were used in this study to ensure rigor and trustworthiness, such as collecting data until thematic data saturation was achieved and piloting the topic guide [31,32]. The inductive nature of our analysis supported the identification of novel insights, such as the burden of using multiple devices and concerns regarding the impact of existing health conditions on specific metrics. Furthermore, the range of different technologies used in the study provides originality and contributes to our understanding of accessibility and usability of technologies for health care purposes. A key limitation of this study is that individuals self-selected to participate in the study and were required to own a smartphone and have access to internet connectivity at home. There were no strategies put in place to support the recruitment of those who may have experienced or are experiencing digital exclusion, such as those of low socioeconomic status who cannot afford a smartphone, or those who live in rural areas who may not have access to a reliable internet connection at home [37]. Some insights into those with little to no experience of using digital technologies (eg, games on a smartphone via an app) were gathered in this study, but further research is needed to gain a deeper understanding of the unique views, perceptions, and opinions of those experiencing digital exclusion and to ensure that the digital intervention does not contribute to digital health inequalities when implemented in practice. Although our study provides a rich in-depth understanding regarding acceptability, usability, and user burden, the inclusion of a quantitative measure of acceptability, usability, and/or burden, such as System Usability Scale [50], could have facilitated data triangulation and potential comparisons to be made between technologies and subgroups (MCI vs no impairment or high vs low device count).

Conclusions

This novel qualitative study highlighted factors affecting concurrent acceptability of using a range of digital technologies that have the potential to support early dementia detection from the perspectives of potential end users. This information will help contribute to researchers' decision-making regarding how many devices the participants should be asked to use and the key elements (eg, comfort, place a device is worn, and waterproof) of a digital device that must be considered to improve acceptance and engagement. Future research is needed to gain a deeper understanding of the factors that affect the acceptance of using digital technologies to support early dementia detection among those experiencing digital exclusion, as our recruitment design could be viewed as limiting inclusion

XSL•FO
RenderX

of this population, and their views are necessary to ensure that digital health equity can be supported.

Acknowledgments

The authors would like to thank all participants for their time and contribution to our research. Generative artificial intelligence was not used in any portion of the manuscript generation.

Funding

This work has been supported by the Early Detection of Neurodegenerative Diseases research initiative, funded by Alzheimer's Research UK with support from Gates Ventures and the Alzheimer's Drug Discovery Foundation through its Diagnostic Accelerator Project. This project also received funding from the American Heart Association. SPS and CT were funded by the National Institute of Health and Care Research (NIHR205190). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to a potential breach of participant confidentiality. Anonymized datasets may be available from the corresponding author upon reasonable request.

Authors' Contributions

SW (lead author) provided qualitative methodology training to the coauthors responsible for data acquisition (ZP and SR), obtained ethics approval, conducted data analysis and interpretation, created visualizations, and drafted, revised, and completed the current version of the manuscript. SW also took responsibility for ensuring that any questions related to the accuracy or integrity of the qualitative work were appropriately investigated and resolved. EB contributed to data analysis and interpretation and provided feedback during the drafting and revision of the manuscript. ZP contributed to data acquisition, obtained ethics approval, supported data interpretation, recruited participants, managed the participant onboarding process, and provided technical support. SR contributed to data acquisition, supported data interpretation, recruited participants, managed the participant onboarding process, and provided technical support. SB and TW contributed to project management. SL supervised ZP and SR and contributed to obtaining ethics approval. RK quality-checked the interview transcripts. CT supervised the lead author (SW), contributed to data analysis and interpretation, and provided feedback on drafts. ZW and RA contributed to funding acquisition. SPS (corresponding author) conceived the study, contributed to methodology development, acquired funding, supervised SW and EB, contributed to data analysis and interpretation, approved the final version of the manuscript, and provided feedback on drafts. SPS also took responsibility for ensuring that any questions related to the accuracy or integrity of the qualitative work were appropriately investigated and resolved.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview topic guide.
[DOCX File, 18 KB - jmir_v28i1e84004_app1.docx]

Checklist 1
COREQ checklist.
[PDF File, 531 KB - jmir_v28i1e84004_app2.pdf]

References

1. Dementia: key facts. World Health Organization. 2025. URL: https://www.who.int/news-room/fact-sheets/detail/dementia [accessed 2025-07-24]
2. Beyer L, Stocker H, Rujescu D, et al. Amyloid‐beta misfolding and GFAP predict risk of clinical Alzheimer's disease diagnosis within 17 years. Alzheimers Dement 2023 Mar;19(3):1020-1028. [doi: 10.1002/alz.12745]
3. Nasreddine ZS, Phillips NA, Bédirian V, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005 Apr;53(4):695-699. [doi: 10.1111/j.1532-5415.2005.53221.x] [Medline: 15817019]
4. Tests for diagnosing dementia. NHS. 2023. URL: https://www.nhs.uk/conditions/dementia/diagnosis-tests/ [accessed 2025-07-24]

XSL•FO
RenderX

5. Dementia: diagnosis and tests. Mayo Clinic. 2025. URL: https://www.mayoclinic.org/diseases-conditions/dementia/diagnosis-treatment/drc-20352019 [accessed 2025-07-24]
6. Hameed S, Fuh JL, Senanarong V, et al. Role of fluid biomarkers and PET imaging in early diagnosis and its clinical implication in the management of Alzheimer's disease. J Alzheimers Dis Rep 2020 Feb 12;4(1):21-37. [doi: 10.3233/ADR-190143] [Medline: 32206755]
7. Anderson R, Knapp M, Wittenberg R, Handels R, Schott JM. Economic modelling of disease-modifying therapies in Alzheimer's disease. : Personal Social Services Research Unit, London School of Economics and Political Science; 2018 URL: https://www.lse.ac.uk/cpec/assets/documents/EconomicmodellingAD.pdf [accessed 2026-05-22]
8. Dementia attitudes monitor. : Alzheimer's Research UK; 2018 URL: https://www.dementiastatistics.org/wp-content/uploads/2019/02/Dementia-Attitudes-Monitor-Wave-1-Report.pdf#zoom=100 [accessed 2025-07-24]
9. Frey AL, Karran M, Jimenez RC. Harnessing the potential of digital technologies for the Early Detection of Neurodegenerative diseases (EDoN). Open Science Framework. 2019. URL: https://osf.io/preprints/osf/u49z5_v1 [accessed 2026-05-22]
10. Lawson L, Mc Ardle R, Wilson S, Beswick E, Karimi R, Slight SP. Digital endpoints for assessing instrumental activities of daily living in mild cognitive impairment: systematic review. J Med Internet Res 2023 Jul 25;25:e45658. [doi: 10.2196/45658] [Medline: 37490331]
11. Lott SA, Streel E, Bachman SL, et al. Digital health technologies for Alzheimer's disease and related dementias: initial results from a landscape analysis and community collaborative effort. J Prev Alzheimers Dis 2024;11(5):1480-1489. [doi: 10.14283/jpad.2024.103] [Medline: 39350395]
12. Slattery P, Saeri AK, Bragge P. Research co-design in health: a rapid overview of reviews. Health Res Policy Syst 2020 Feb 11;18(1):17. [doi: 10.1186/s12961-020-0528-9] [Medline: 32046728]
13. Madanian S, Nakarada-Kordic I, Reay S, Chetty TH. Patients' perspectives on digital health tools. PEC Innov 2023 Dec;2:100171. [doi: 10.1016/j.pecinn.2023.100171] [Medline: 37384154]
14. Evidence standards framework for digital health technologies. : National Institute for Health and Clinical Excellence; 2022 URL: https://www.nice.org.uk/corporate/ecd7 [accessed 2025-07-24]
15. Questions and answers: qualification of digital technology-based methodologies to support approval of medicinal products. European Medicines Agency. 2020. URL: https://www.ema.europa.eu/en/documents/other/questions-and-answers-qualification-digital-technology-based-methodologies-support-approval-medicinal-products_en.pdf [accessed 2025-07-24]
16. Digital health technologies for remote data acquisition in clinical investigations. US Food and Drug Administration. 2023. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/digital-health-technologies-remote-data-acquisition-clinical-investigations [accessed 2025-07-24]
17. Overview: Alzheimer's disease. NHS. 2024. URL: https://www.nhs.uk/conditions/alzheimers-disease/#:~:text=Who%20is%20affected?,or%20young%2Donset%20Alzheimer's%20disease [accessed 2025-07-24]
18. Campbell NL, Unverzagt F, LaMantia MA, Khan BA, Boustani MA. Risk factors for the progression of mild cognitive impairment to dementia. Clin Geriatr Med 2013 Nov;29(4):873-893. [doi: 10.1016/j.cger.2013.07.009] [Medline: 24094301]
19. Albert MS, DeKosky ST, Dickson D, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011 May;7(3):270-279. [doi: 10.1016/j.jalz.2011.03.008] [Medline: 21514249]
20. Grill JD, Di L, Lu PH, et al. Estimating sample sizes for predementia Alzheimer's trials based on the Alzheimer's Disease Neuroimaging Initiative. Neurobiol Aging 2013 Jan;34(1):62-72. [doi: 10.1016/j.neurobiolaging.2012.03.006] [Medline: 22503160]
21. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017 Jan 26;17(1):88. [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
22. Common questions about participating in research. Boston University Alzheimer's Disease Research Centre. 2025. URL: https://www.bu.edu/alzresearch/education-program/common-questions-about-participating-in-research/#:~:text=I%20am%20healthy.,Our%20Commitment%20to%20you [accessed 2025-07-24]
23. Wilson S, Ardle RM, Tolley C, Slight S. Usability and acceptability of wearable technology in the early detection of dementia. Alzheimers Dement 2022 Dec;18 Suppl 2:e059820. [doi: 10.1002/alz.059820] [Medline: 36537475]
24. Ritchie J, Spencer L. Qualitative data analysis for applied policy research. In: Bryman A, Burgess RG, editors. Analyzing Qualitative Data: Routledge:1994. [doi: 10.4324/9780203413081_chapter_9]
25. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013 Sep 18;13:117. [doi: 10.1186/1471-2288-13-117] [Medline: 24047204]
26. Glaser BG. The constant comparative method of qualitative analysis. Soc Probl 1965 Apr;12(4):436-445. [doi: 10.2307/798843]
27. Yang L, Qi L, Zhang B. Concepts and evaluation of saturation in qualitative research. Adv Psychol Sci 2022 Mar 1;30(3):511-521. [doi: 10.3724/SP.J.1042.2022.00511]

XSL•FO
RenderX

28. Francis JJ, Johnston M, Robertson C, et al. What is an adequate sample size? Operationalising data saturation for theory-based interview studies. Psychol Health 2010 Dec;25(10):1229-1245. [doi: 10.1080/08870440903194015] [Medline: 20204937]
29. Marshall MN. Sampling for qualitative research. Fam Pract 1996 Dec;13(6):522-525. [doi: 10.1093/fampra/13.6.522] [Medline: 9023528]
30. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
31. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
32. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Methods 2017;16(1). [doi: 10.1177/1609406917733847]
33. Lincoln YS, Guba EG. But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. New Direct Program Eval 1986 Jun;1986(30):73-84. [doi: 10.1002/ev.1427]
34. Khaleghi A, Aghaei Z, Mahdavi MA. A gamification framework for cognitive assessment and cognitive training: qualitative study. JMIR Serious Games 2021 May 18;9(2):e21900. [doi: 10.2196/21900] [Medline: 33819164]
35. Ríos Rincón AM, Daum C, Miguel Cruz A, Liu L, Stroulia E. Feasibility and acceptability of a serious mobile-game intervention for older adults. Phys Occup Ther Geriatr 2022 Jul 3;40(3):295-318. [doi: 10.1080/02703181.2022.2030849]
36. Groenewoud H, Lange J, Schikhof Y, Astell A, Joddrell P, Goumans M. People with dementia playing casual games on a tablet. Gerontechnology 2017 May 22;16(1):37-47. [doi: 10.4017/gt.2017.16.1.004.00]
37. Wilson S, Tolley C, McArdle R, Slight R, Slight S. Who is most at risk of digital exclusion within healthcare? Int J Pharm Pract 2024 Apr 29;32(Suppl 1):i3-i4. [doi: 10.1093/ijpp/riae013.004]
38. Liu X, Wang X, Zhang H, Pei M, Li N. Relationship between digital exclusion and cognitive impairment in Chinese adults. Front Aging Neurosci 2023;15:1194348. [doi: 10.3389/fnagi.2023.1194348] [Medline: 37465320]
39. Scherer MJ, Craddock G, Mackeogh T. The relationship of personal factors and subjective well-being to the use of assistive technology devices. Disabil Rehabil 2011;33(10):811-817. [doi: 10.3109/09638288.2010.511418] [Medline: 20735272]
40. Zeagler C. Where to wear it: functional, technical, and social considerations in on-body location for wearable technology 20 years of designing for wearability. Presented at: Proceedings of the 2017 ACM International Symposium on Wearable Computers; Sep 11-15, 2017. [doi: 10.1145/3123021.3123042]
41. Bergmann JHM, Chandaria V, McGregor A. Wearable and implantable sensors: the patient's perspective. Sensors (Basel) 2012 Dec 5;12(12):16695-16709. [doi: 10.3390/s121216695] [Medline: 23443394]
42. Peek STM, Wouters EJM, van Hoof J, Luijkx KG, Boeije HR, Vrijhoef HJM. Factors influencing acceptance of technology for aging in place: a systematic review. Int J Med Inform 2014 Apr;83(4):235-248. [doi: 10.1016/j.ijmedinf.2014.01.004] [Medline: 24529817]
43. Brown SA, Venkatesh V. Model of adoption of technology in households: a baseline model test and extension incorporating household life cycle1. MIS Q 2005 Sep 1;29(3):399-426. [doi: 10.2307/25148690]
44. Moore K, O'Shea E, Kenny L, et al. Older adults' experiences with using wearable devices: qualitative systematic review and meta-synthesis. JMIR Mhealth Uhealth 2021 Jun 3;9(6):e23832. [doi: 10.2196/23832] [Medline: 34081020]
45. Puri A, Kim B, Nguyen O, Stolee P, Tung J, Lee J. User acceptance of wrist-worn activity trackers among community-dwelling older adults: mixed method study. JMIR Mhealth Uhealth 2017 Nov 15;5(11):e173. [doi: 10.2196/mhealth.8211] [Medline: 29141837]
46. Law M, Sutherland C, Ahn HS, et al. Developing assistive robots for people with mild cognitive impairment and mild dementia: a qualitative study with older adults and experts in aged care. BMJ Open 2019 Sep 24;9(9):e031937. [doi: 10.1136/bmjopen-2019-031937] [Medline: 31551392]
47. Mair FS, Montori VM, May CR. Digital transformation could increase the burden of treatment on patients. BMJ 2021 Nov 25;375:n2909. [doi: 10.1136/bmj.n2909] [Medline: 34824093]
48. Peterson CM, St Louis RM, Flannagan C. Enhancing enrollment and adherence in long-term wearable research on dementia: qualitative systematic review and meta-synthesis. JMIR Aging 2025 Jul 31;8:e63768. [doi: 10.2196/63768] [Medline: 40743521]
49. Ehn M, Eriksson LC, Åkerberg N, Johansson AC. Activity monitors as support for older persons' physical activity in daily life: qualitative study of the users' experiences. JMIR Mhealth Uhealth 2018 Feb 1;6(2):e34. [doi: 10.2196/mhealth.8345] [Medline: 29391342]
50. Wong AKC, Lee JHT, Zhao Y, Lu Q, Yang S, Hui VCC. Exploring older adults' perspectives and acceptance of AI-driven health technologies: qualitative study. JMIR Aging 2025 Feb 12;8:e66778. [doi: 10.2196/66778] [Medline: 39937162]

Abbreviations

BU-ADRC: Boston University-Alzheimer's Disease Research Centre
COREQ: Consolidated Criteria for Reporting Qualitative Research
EDoN: Early Detection of Neurodegenerative Diseases

XSL•FO
RenderX

MCI: mild cognitive impairment
SRQR: Standards for Reporting Qualitative Research

Edited by A Stone; submitted 12.Sep.2025; peer-reviewed by J Scott, L Masanneck; revised version received 13.Mar.2026; accepted 31.Mar.2026; published 29.May.2026.

Please cite as:
Wilson S, Beswick E, Popp Z, Rahman S, Bhogal S, Whitfield T, Low S, Khan R, Tolley C, Walker Z, Au R, Slight SP
Acceptability of Technologies to Support Early Dementia Detection: Qualitative Study With the Boston University Alzheimer's Disease Center Cohort
J Med Internet Res 2026;28:e84004
URL: https://www.jmir.org/2026/1/e84004
doi:10.2196/84004

© Sarah Wilson, Emily Beswick, Zachary Popp, Salman Rahman, Sharandeep Bhogal, Tim Whitfield, Spencer Low, Raiyan Khan, Clare Tolley, Zuzana Walker, Rhoda Au, Sarah P Slight. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Association Between Wearable Device Adoption and Health-Related Lifestyle Behaviors: Retrospective Cohort Study

Eunjeong Choi[1*], PhD; Seoyeong Choi[1*], BS; Suk-Yong Jang[2,3], MD, JD, PhD

[1]Department of Public Health, Graduate School, Yonsei University, Seoul, Republic of Korea

[2]Department of Biohealth Industry, Policy Analysis Division, Graduate School of Transdisciplinary Health Science, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul, Republic of Korea

[3]Department of Health Policy and Management, Graduate School of Public Health, Yonsei University, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Suk-Yong Jang, MD, JD, PhD
Department of Biohealth Industry, Policy Analysis Division, Graduate School of Transdisciplinary Health Science, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul, Republic of Korea

Abstract

Background: Wearable devices are increasingly adopted for personal health monitoring, but evidence on their long-term associations with health-related lifestyle behaviors in real-world population settings remains limited.

Objective: This study examined the longitudinal association between wearable device adoption and engagement in health-related lifestyle behaviors using a nationally representative panel dataset from South Korea.

Methods: We analyzed data from the 2016 and 2022 waves of the Korea Media Panel survey. Health-related lifestyle behaviors in the physical, social, and cultural domains were operationalized as estimated annual activity counts based on self-reported frequency measures. We used a difference-in-differences framework with generalized estimating equations to compare changes in these behaviors between new wearable adopters and nonadopters adjusting for demographic and socioeconomic characteristics. Relative changes were estimated using Poisson models with a log link, and subgroup analyses were conducted to explore variation across sociodemographic groups. As a sensitivity analysis, inverse probability of treatment weighting was additionally applied to assess the robustness of the findings to observed baseline imbalance.

Results: Wearable device adoption was associated with greater increases in total, physical, and cultural health-related lifestyle activities over time. In the difference-in-differences model, adopters showed greater relative increases in total activity (rate ratio [RR] 1.24, 95% CI 1.08-1.35), physical activity (RR 1.36, 95% CI 1.12-1.64), and cultural activity (RR 1.78, 95% CI 1.31-2.42) than nonadopters. Subgroup analyses showed limited evidence of consistent heterogeneity and should be interpreted cautiously. Sensitivity analyses using inverse probability of treatment weighting showed overall patterns broadly similar to those of the primary analyses.

Conclusions: In this nationally representative panel study, wearable device adoption was associated with greater increases in total, physical, and cultural health-related lifestyle activities over time, whereas no clear association was observed for social activity. These findings should be interpreted as associative rather than causal given the observational design and the inability to directly assess parallel trends.

(J Med Internet Res 2026;28:e88276) doi:10.2196/88276

KEYWORDS

wearable electronic devices; digital health; self-monitoring; health behavior; physical activity; healthy aging

Introduction

The World Health Organization defines health not merely as the absence of disease but as a state of physical, mental, and social well-being [1]. This broader perspective suggests that health-related behaviors should be understood not only as isolated actions but also as part of wider lifestyle patterns shaped by social and structural contexts. Although much health behavior research has focused on single behaviors or narrow clusters of behaviors, health lifestyle theory emphasizes that behaviors are socially patterned and tend to cluster in ways that reflect broader orientations toward health and daily life [2-4]. Within this context, wearable devices have gained increasing attention as tools for personal health monitoring and self-tracking. Prior studies suggest favorable associations between wearable devices and physical activity, and in some contexts, wearable-related interventions may also be linked to motivation, reassurance, or perceived support [5-7]. However, much of the existing evidence comes from randomized trials or short-term studies conducted

XSL•FO
RenderX

under relatively controlled conditions [8,9]. As a result, less is known about how wearable device adoption is associated with broader health-related lifestyle behaviors over longer periods in real-world population settings.

Self-determination theory may provide one possible framework for interpreting how wearable devices could be related to behavior [10]. For example, functions such as goal setting, feedback, and self-monitoring may support feelings of autonomy or competence for some users, and social features may foster relatedness in certain contexts. However, these mechanisms were not directly measured in this study. Accordingly, we draw on this framework as a conceptual reference rather than as a tested explanatory model. South Korea provides a relevant setting for examining wearable device adoption and health-related lifestyle behaviors because it has a rapidly increasing proportion of older adults, high digital connectivity, and growing public health interest in promoting healthy aging and everyday health management [11,12]. These features make South Korea an informative context for studying whether wearable device adoption is associated with changes in lifestyle-related activities in a real-world setting.

Using nationally representative panel data from South Korea, this study examined the longitudinal association between wearable device adoption and health-related lifestyle behaviors, operationalized as physical, social, and cultural activities. By following individuals over time, we aimed to assess whether changes in these activities differed between new wearable adopters and nonadopters.

Methods

Data Source

This study used data from the 2016 and 2022 waves of the Korea Media Panel survey. The Korea Media Panel is a nationally representative longitudinal survey administered by the Korea Information Society Development Institute and funded by the Korean government [13]. The panel follows the same households and individuals over time to assess changes in media environments and media-related behaviors. Data are collected annually through face-to-face interviews from more than 5000 households and individuals across 17 regions in South Korea. The survey includes information on media device ownership, communication service subscriptions, expenditures, and selected psychosocial and lifestyle-related measures. This study followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines (Checklist 1).

Study Population

We used the 2016 wave as baseline and the 2022 wave as follow-up. Among 8847 participants surveyed at baseline, we excluded those who reported wearable device use at baseline (n=132, 1.5%); those younger than 20 years (n=1278, 14.4%); and those who did not participate in the 2022 survey or had missing data on the exposure, outcomes, or covariates used in the analysis (n=3182, 36.0%). The final analytic sample included 5507 participants with complete data at both time points. Because exclusion due to nonparticipation or missing data was substantial, we describe this as a complete-case analysis.

Exposure Variable

The exposure of interest was new wearable device adoption between 2016 and 2022. Participants were classified as adopters if they did not report wearable device use in 2016 but reported wearable device use in 2022. Participants who did not report wearable device use in either wave were classified as nonadopters.

Outcome Variables

Health-related lifestyle behaviors were assessed using self-reported frequency of participation in specific activities during the previous year. Each item was measured using an 8-point frequency response scale (1=“almost never,” 2=“about once a year,” 3=“about once every 6 months,” 4=“about once every 3 months,” 5=“about 1-3 times a month,” 6=“about 1-3 times a week,” 7=“about 4-6 times a week,” and 8=“almost daily”). Because these response categories reflect ordered participation frequencies rather than exact event counts, we converted them into estimated annual participation counts to facilitate comparison across activities and domains.

The activity variables were grouped into 3 domains: physical activity (sports or outdoor activities), social activity (friendship interactions, social or political participation, and religious activities), and cultural activity (creative hobbies and attendance to cultural events). For each domain, estimated annual counts were summed across the relevant items. An overall health-related lifestyle activity measure was calculated by summing counts across all domains.

Covariates

To address potential confounding, we adjusted for demographic and socioeconomic characteristics that may be associated with both wearable device adoption and health-related lifestyle behaviors. These covariates included age, sex, educational attainment, household income, household composition, and region of residence. Age was categorized as 20 to 39 years, 40 to 64 years, and 65 years and above. Educational attainment was categorized as middle school or lower vs high school or higher. Household income was grouped into 3 categories: low (<KRW 2 million [US $1357.24] per month), middle (KRW 2-3.99 million [US $1357.24-$2707.69] per month), and high (≥KRW 4 million [US $2714.48] per month). The “metropolitan” region of residence referred to the capital region of South Korea (Seoul and its surrounding metropolitan areas) as well as major metropolitan cities, whereas “nonmetropolitan” included all other cities and provinces outside these metropolitan regions. These variables were selected a priori based on prior literature and conceptual considerations related to social participation, access to digital technologies, and health-related behavior [14].

Statistical Analysis

Baseline sociodemographic characteristics were summarized according to wearable device adoption status. Frequencies and percentages were reported, and standardized differences were calculated to assess baseline imbalance between adopters and nonadopters. We used a difference-in-differences (DID) framework [15] to examine whether changes in health-related lifestyle behaviors from 2016 to 2022 differed between adopters

XSL•FO
RenderX

and nonadopters. The DID parameter of interest was the interaction between time (2016 vs 2022) and adoption status (adopter vs nonadopter). Because only 2 waves were available, the parallel trends assumption could not be directly evaluated using preexposure trends.

We fitted generalized estimating equation models to account for repeated observations within individuals. Relative DID estimates were derived from Poisson models using a log link, and adjusted absolute DID estimates for annual activity frequencies were derived from identity link models. Each model included indicators for time, adoption status, and their interaction, with adjustment for age, sex, educational attainment, household income, household composition, and region of residence. Subgroup analyses were conducted by sex, educational attainment, household income, household composition, and region of residence to explore heterogeneity in associations. As a sensitivity analysis, we additionally applied inverse probability of treatment weighting based on baseline covariates to further address observed baseline imbalance between adopters and nonadopters. All analyses were conducted using SAS Enterprise Guide (version 7.1; SAS Institute).

Ethical Considerations

This nationwide population-based retrospective cohort study was approved by the institutional review board of Severance Hospital, South Korea (4-2025-1151). This study was conducted in accordance with the tenets of the Declaration of Helsinki. Written informed consent was waived because the study used deidentified database data.

Results

A total of 5507 participants were included in the analysis, of whom 730 (13.3%) were wearable adopters and 4777 (86.7%) were nonadopters. As shown in Table 1, substantial baseline imbalance was observed between adopters and nonadopters, particularly in age, educational attainment, and household income. Adopters were generally younger, more highly educated, and more likely to have a higher household income and live in metropolitan areas.

XSL•FO
RenderX

Table . General characteristics of study participants (N=5507).

Variable	Wearable device adoption, n (%)[a]			Standardized difference
	Total	Yes (n=730)	No (n=4777)	
Age (y)				1.034
20-39	727 (13.2)	265 (36.3)	462 (9.7)	
40-64	2945 (53.5)	443 (60.7)	2502 (52.4)	
≥65	1835 (33.3)	22 (3.0)	1813 (38.0)	
Sex				0.134
Male	2360 (42.9)	355 (48.6)	2005 (42.0)	
Female	3147 (57.1)	375 (51.4)	2772 (58.0)	
Educational attainment				0.885
Middle school or lower	1524 (27.7)	10 (1.4)	1514 (31.7)	
High school or higher	3983 (72.3)	720 (98.6)	3263 (68.3)	
Household income[b]				0.638
Low	1425 (25.9)	51 (7.0)	1374 (28.8)	
Middle	2253 (40.9)	316 (43.3)	1937 (40.5)	
High	1829 (33.2)	363 (49.7)	1466 (30.7)	
Region of residence[c]				0.352
Metropolitan	3312 (60.1)	543 (74.4)	2769 (58.0)	
Nonmetropolitan	2195 (39.9)	187 (25.6)	2008 (42.0)	
Household composition				0.173
One-person household	502 (9.1)	26 (3.6)	476 (10.0)	
Multiperson household	5005 (90.9)	704 (96.4)	4301 (90.0)	

[a]Participants who already owned a wearable device at baseline (2016) were excluded from the analytic sample. “Yes” refers to individuals who did not own a wearable device at baseline but owned one by the follow-up in 2022, whereas “No” refers to those who never owned a device during the study period.

[b]Household income was grouped into 3 categories: low (<KRW 2 million [US $1357.24] per month), middle (KRW 2 - 3.99 million [US $1357.24-$2707.69] per month), and high (≥KRW 4 million [US $2714.48] per month).

[c]“Metropolitan” refers to the capital region of South Korea (Seoul and its surrounding metropolitan areas) as well as major metropolitan cities. “Nonmetropolitan” includes all other cities and provinces outside these metropolitan regions.

Table 2 shows both absolute DID estimates expressed as annual frequency differences and relative DID estimates expressed as rate ratios (RRs). In the overall analysis, wearable adoption was associated with greater increases in total health-related lifestyle behavior over time (absolute DID=29.13, 95% CI 15.42-42.84; RR 1.24, 95% CI 1.08-1.35). For domain-specific outcomes, wearable adoption was associated with increases in physical activity and cultural activity but not social activity. The adjusted absolute DID estimate for physical activity was 22.88 (95% CI 14.92-30.84), corresponding to an RR of 1.36 (95% CI 1.12-1.64). For cultural activity, the adjusted absolute DID estimate was 5.45 (95% CI 0.55-10.36), with an RR of 1.78 (95% CI 1.31-2.42). No clear association was observed for social activity (absolute DID=0.74, 95% CI −7.27 to 8.75; RR 1.03, 95% CI 0.91-1.16). In sensitivity analyses using inverse probability of treatment weighting (IPTW), the overall pattern of results was broadly similar to that of the primary analyses (Multimedia Appendix 1).

XSL•FO
RenderX

Table . Changes in health-related lifestyle behaviors after wearable device adoption.

Outcome	Nonadopters[a], mean annual frequency		Adopters, mean annual frequency		Absolute change[b] (frequency)		Relative change (rate ratio)	
	2016 wave	2022 wave	2016 wave	2022 wave	Adjusted difference-in-differences estimate (95% CI)	*P* value	Adjusted difference-in-differences estimate (95% CI)	*P* value
Total health-related lifestyle behavior[c]								
Overall	85.24	91.69	113.42	152.62	29.13 (15.42 to 42.84)	<.001	1.24 (1.08 to 1.35)	<.001
20-39 y	95.52	116.59	119.16	154.21	17.19 (−7.29 to 41.68)	.17	1.08 (0.89 to 1.31)	.45
40-64 y	86.52	95.00	107.65	152.08	35.24 (16.75 to 53.74)	<.001	1.30 (1.11 to 1.52)	<.001
≥65 y	75.98	80.77	71.01	144.34	67.52 (−52.98 to 188.02)	.27	1.80 (0.49 to 6.60)	.38
Physical activity								
Overall	20.39	32.16	30.78	67.21	22.88 (14.92 to 30.84)	<.001	1.36 (1.12 to 1.64)	.002
20-39 y	22.62	39.53	31.00	65.59	19.44 (5.55 to 33.32)	.006	1.27 (0.91 to 1.78)	.16
40-64 y	23.05	34.69	30.92	67.76	25.02 (14.70 to 35.34)	<.001	1.50 (1.17 to 1.92)	.001
≥65 y	13.59	26.79	8.33	75.68	55.34 (0.94 to 109.74)	.046	4.30 (0.96 to 19.18)	.06
Social activity								
Overall	52.56	52.39	65.15	66.87	0.74 (−7.27 to 8.75)	.86	1.03 (0.91 to 1.16)	.68
20-39 y	50.64	59.76	67.55	69.66	−5.89 (−19.59 to 7.81)	.40	0.88 (0.71 to 1.09)	.24
40-64 y	51.19	52.08	62.55	65.50	1.30 (−10.01 to 12.62)	.82	1.02 (0.85 to 1.23)	.81
≥65 y	56.58	50.95	57.85	60.83	7.96 (−67.37 to 83.29)	.84	1.14 (0.33 to 3.96)	.84
Cultural activity								
Overall	12.28	7.14	17.49	18.54	5.45 (0.55 to 10.36)	.03	1.78 (1.31 to 2.42)	<.001
20-39 y	22.26	17.30	20.60	18.96	3.59 (−4.68 to 11.86)	.40	1.18 (0.76 to 1.84)	.47
40-64 y	12.28	8.23	14.18	18.82	8.83 (1.75 to 15.91)	.02	2.05 (1.28 to 3.27)	.003
≥65 y	5.81	3.03	4.83	7.83	4.74 (−8.63 to 18.12)	.49	2.69 (0.39 to 18.58)	.32

[a]Because all participants were nonadopters at baseline (2016), group assignment was determined based on wearable adoption at follow-up (2022). Nonadopters comprised individuals who remained nonadopters in both 2016 and 2022, whereas adopters included those who newly adopted a wearable device by 2022.

[b]Absolute change indicates the annual difference in differences in the number of activity engagements, whereas relative percentage change reflects the proportional difference-in-differences effect expressed as a percentage. All estimates were adjusted for age, sex, educational attainment, household composition, household income, and region of residence.

[c]Health-related lifestyle behaviors refer to physical, social, and cultural activity. Physical activity refers to participation in sports, exercise, or outdoor recreational activities. Social activity includes friendship-based gatherings, religious participation, and engagement in community or sociopolitical groups. Cultural activity encompasses hobby-related activities and attendance to cultural events such as performances, exhibitions, or other arts-related experiences.

XSL•FO
RenderX

Table 3 presents subgroup-specific relative DID estimates and formal tests for interaction. Overall, evidence of interaction was limited and inconsistent across domains, although some heterogeneity was observed for region of residence and educational attainment. Given the wide CIs and sparse data in several strata, these subgroup findings should be regarded as exploratory.

Table . Difference-in-differences estimates of health-related lifestyle behaviors by demographic characteristic.

Outcome	Total health promotion activity[a] rate ratio[b] (95% CI)	*P* value for the interaction	Social activity rate ratio (95% CI)	*P* value for the interaction	Cultural activity rate ratio (95% CI)	*P* value for the interaction	Physical activity rate ratio (95% CI)	*P* value for the interaction
Sex		.34		.38		.76		.25
Male	1.17 (0.99 - 1.38)		0.96 (0.79 - 1.17)		1.88 (1.15 - 3.08)		1.23 (0.97 - 1.57)	
Female	1.30 (1.12 - 1.51)		1.08 (0.92 - 1.26)		1.72 (1.17 - 2.54)		1.53 (1.14 - 2.07)	
Household income[c]		.23		.40		.18		.54
Low	1.85 (1.07 - 3.21)		1.48 (0.70 - 3.14)		8.48 (2.18 - 32.93)		1.64 (0.41 - 6.59)	
Middle-high	1.23 (1.09 - 1.37)		0.998 (0.88 - 1.13)		1.70 (1.24 - 2.33)		1.39 (1.14 - 1.69)	
Region of residence[d]		<.001		.007		.07		.01
Metropolitan	1.08 (0.95 - 1.23)		0.87 (0.75 - 1.01)		1.54 (1.09 - 2.18)		1.28 (1.03 - 1.60)	
Nonmetropolitan	1.78 (1.43 - 2.21)		1.39 (1.10 - 1.76)		2.77 (1.42 - 5.40)		1.95 (1.31 - 2.90)	
Educational attainment		.02		.04		.23		.07
Middle school or lower	5.52 (1.54 - 19.80)		3.84 (1.19 - 12.36)		123.28 (25.84 - 588.16)		6.80 (1.21 - 38.38)	
High school or higher	1.23 (1.10 - 1.38)		0.99 (0.87 - 1.12)		1.62 (1.19 - 2.22)		1.48 (1.23 - 1.80)	
Household composition		.17		.34		.006		.74
Single person	1.90 (1.17 - 3.08)		1.43 (0.74 - 2.76)		16.77 (6.95 - 40.49)		1.86 (0.80 - 4.35)	
Nonsingle person	1.22 (1.09 - 1.37)		1.01 (0.89 - 1.15)		1.64 (1.20 - 2.25)		1.36 (1.12 - 1.65)	

[a]Health-related lifestyle behaviors refer to physical, social, and cultural activity. Physical activity refers to participation in sports, exercise, or outdoor recreational activities. Social activity includes friendship-based gatherings, religious participation, and engagement in community or sociopolitical groups. Cultural activity encompasses hobby-related activities and attendance to cultural events such as performances, exhibitions, or other arts-related experiences.

[b]Rate ratio indicates the adjusted relative difference-in-differences (DID) effect within each subgroup. The *P* value for the interaction tests whether the DID effect differs between subgroup categories. All estimates were adjusted for age, sex, educational attainment, household composition, household income, and region of residence.

[c]Household income was grouped into 2 categories: low (<KRW 2 million [US $1357.24] per month) and middle-high (≥KRW 2 million [US $1357.24] per month).

[d]"Metropolitan" refers to the capital region of South Korea (Seoul and its surrounding metropolitan areas) as well as major metropolitan cities. "Nonmetropolitan" includes all other cities and provinces outside these metropolitan regions.

Discussion

Principal Findings

In this longitudinal study using a DID framework, wearable device adoption was associated with greater increases in total, physical, and cultural health-related lifestyle behaviors over time, whereas no clear association was observed for social activity. Overall, these findings are broadly consistent with those of prior studies reporting favorable findings regarding wearable devices and physical activity [5,16,17]. However, our study extends this literature by examining wearable adoption in a real-world population setting using observational panel data rather than short-term intervention-based or highly controlled study designs. Several possible explanations may account for the observed associations, although these mechanisms were not

XSL•FO
RenderX

directly measured in this study. Prior literature suggests that wearable devices may support self-monitoring, goal setting, and feedback, which in turn may be related to greater engagement in some health-related behaviors [10,18,19]. At the same time, the absence of a clear association with social activity may reflect the fact that many social activities, such as friendship gatherings, religious participation, and volunteering, depend more heavily on interpersonal, contextual, and community-level factors than on individual self-tracking tools [7,20].

In age-stratified analyses, associations were most consistently observed among adults aged 40 to 64 years, whereas estimates among older adults were less precise and should be interpreted cautiously. One possible interpretation is that wearable adoption may be more strongly linked to behavior change among individuals facing greater barriers to physical activity or greater interest in health monitoring in midlife and later life [6,17]. However, several subgroup estimates, particularly among older adults, were based on relatively small numbers of adopters and were accompanied by wide CIs. Accordingly, these findings should be viewed as exploratory rather than definitive. Subgroup analyses otherwise showed limited and inconsistent evidence of heterogeneity across sociodemographic characteristics. Interaction tests suggested some heterogeneity for region of residence and educational attainment in selected outcomes, but most interaction tests were not statistically significant. Given the wide CIs in several strata, these subgroup findings should not be overinterpreted.

Our findings suggest that wearable adoption may be associated with favorable changes in some lifestyle-related activities in a real-world setting. At the same time, wearable adoption in this study was voluntary and strongly socially patterned: adopters were younger, more highly educated, and more likely to have a higher household income and live in metropolitan areas than nonadopters. Accordingly, the observed associations may partly reflect selection processes related to digital literacy, socioeconomic advantage, or other unmeasured factors rather than wearable adoption itself. These issues should be considered carefully when interpreting the findings.

Limitations

This study has several limitations. First, only 2 survey waves (2016 and 2022) were available, which prevented direct evaluation of the parallel trends assumption underlying the DID framework. This is an important limitation because covariate adjustment cannot verify whether adopters and nonadopters would have followed similar preadoption trajectories in the absence of wearable adoption. Therefore, the findings should be interpreted as associative rather than causal. Second, substantial baseline imbalance was observed between adopters and nonadopters, particularly with respect to age, educational attainment, and household income. Although we adjusted for measured sociodemographic characteristics, residual confounding and selection bias remain possible. In particular, unmeasured factors during and after the COVID-19 pandemic may have influenced both wearable adoption and changes in activity patterns. Third, wearable device adoption was assessed only at each survey wave, and the survey did not capture the exact timing of adoption, duration of device use, or intensity of engagement. Therefore, participants classified as adopters may have varied in the extent of their actual device use, and the findings should not be interpreted as reflecting sustained wearable device engagement. Fourth, the activity outcomes were based on self-reported past-year frequency categories and were converted into approximate annual participation counts, which may have introduced measurement error. In addition, because both the exposure and outcome measures were self-reported by the same respondents, common method bias is possible. Fifth, the analytic sample was restricted to participants with complete data at both time points. Because exclusions due to nonparticipation or missing data were substantial, selection bias related to study retention cannot be ruled out.

Conclusions

This nationally representative panel study found that wearable device adoption was associated with greater increases in total, physical, and cultural health-related lifestyle activities over time, whereas no clear association was observed for social activity. The findings suggest that wearable adoption may be linked to favorable behavioral patterns in some domains, warranting further research using designs that can better address selection processes, timing of adoption, and actual device use.

Acknowledgments

The authors express their sincere appreciation to the Korea Information Society Development Institute for providing access to Korea Media Panel data.

Funding

This study was not supported by any funding.

Data Availability

The data used in this study are publicly available through the Korea Media Panel [21]. The dataset was deidentified and is accessible to researchers upon request to the Korea Information Society Development Institute.

Authors' Contributions

Conceptualization: EC, SC, SYJ

XSL•FO
RenderX

Data curation: EC, SC
Formal analysis: EC, SC
Investigation: EC, SC
Methodology: EC, SC, SYJ
Supervision: SYJ
Validation: EC, SC, SYJ
Writing—original draft: EC, SC
Writing—review and editing: EC, SC, SYJ

Conflicts of Interest

None declared.

Multimedia Appendix 1
Comparison of primary and inverse probability of treatment–weighted difference-in-differences estimates for health-related lifestyle behaviors.
[DOCX File, 18 KB - jmir_v28i1e88276_app1.docx]

Checklist 1
STROBE checklist.
[PDF File, 183 KB - jmir_v28i1e88276_app2.pdf]

References

1. Constitution of the World Health Organization. World Health Organization. URL: https://www.who.int/about/governance/constitution [accessed 2026-04-29]
2. Frohlich KL, Corin E, Potvin L. A theoretical proposal for the relationship between context and disease. Sociol Health Illn 2001 Nov;23(6):776-797. [doi: 10.1111/1467-9566.00275]
3. Bourdieu P. Distinction: A Social Critique of the Judgement of Taste: Harvard University Press; 1984.
4. Cockerham WC. Health lifestyle theory and the convergence of agency and structure. J Health Soc Behav 2005 Mar;46(1):51-67. [doi: 10.1177/002214650504600105] [Medline: 15869120]
5. Ferguson T, Olds T, Curtis R, et al. Effectiveness of wearable activity trackers to increase physical activity and improve health: a systematic review of systematic reviews and meta-analyses. Lancet Digit Health 2022 Aug;4(8):e615-e626. [doi: 10.1016/S2589-7500(22)00111-X] [Medline: 35868813]
6. Huang S. Enhancing physical activity safety in older adults with chronic comorbidities using wearable devices: benefits, challenges, and clinical implications. Interdiscip Nurs Res 2025;4(2):121-126. [doi: 10.1097/NR9.0000000000000091]
7. Moore K, O'Shea E, Kenny L, et al. Older adults' experiences with using wearable devices: qualitative systematic review and meta-synthesis. JMIR Mhealth Uhealth 2021 Jun 3;9(6):e23832. [doi: 10.2196/23832] [Medline: 34081020]
8. Tang MS, Moore K, McGavigan A, Clark RA, Ganesan AN. Effectiveness of wearable trackers on physical activity in healthy adults: systematic review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2020 Jul 22;8(7):e15576. [doi: 10.2196/15576] [Medline: 32706685]
9. Szeto K, Arnold J, Singh B, Gower B, Simpson CEM, Maher C. Interventions using wearable activity trackers to improve patient physical activity and other outcomes in adults who are hospitalized: a systematic review and meta-analysis. JAMA Netw Open 2023 Jun 1;6(6):e2318478. [doi: 10.1001/jamanetworkopen.2023.18478] [Medline: 37318806]
10. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000 Jan;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
11. KOSIS KOrean Statistical Information Service. URL: https://kosis.kr/ [accessed 2026-04-29]
12. Invest KOREA. URL: https://www.investkorea.org/ik-en/cntnts/i-299/web.do [accessed 2026-04-29]
13. Media statistics portal. Korea Information Society Development Institute. URL: https://stat.kisdi.re.kr/kor/contents/ContentsList.html?subject=USER&sub_div=U [accessed 2025-09-16]
14. Berkman LF, Glass T, Brissette I, Seeman TE. From social integration to health: Durkheim in the new millennium. Soc Sci Med 2000 Sep;51(6):843-857. [doi: 10.1016/s0277-9536(00)00065-4] [Medline: 10972429]
15. Card D, Krueger AB. Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania. Am Econ Rev 1994;84(4):772-793. [doi: 10.3386/w4509]
16. Kang HS, Exworthy M. Wearing the future-wearables to empower users to take greater responsibility for their health and care: scoping review. JMIR Mhealth Uhealth 2022 Jul 13;10(7):e35684. [doi: 10.2196/35684] [Medline: 35830222]
17. Ho MH, Peng CY, Liao Y, Yen HY. Efficacy of a wearable activity tracker with step-by-step goal-setting on older adults' physical activity and sarcopenia indicators: clustered trial. J Med Internet Res 2024 Nov 1;26:e60183. [doi: 10.2196/60183] [Medline: 39486024]

XSL•FO
RenderX

18. Werbrouck A, Swinnen E, Kerckhofs E, Buyl R, Beckwée D, De Wit L. How to empower patients? A systematic review and meta-analysis. Transl Behav Med 2018 Sep 8;8(5):660-674. [doi: 10.1093/tbm/iby064] [Medline: 29982675]
19. Bravo P, Edwards A, Barr PJ, et al. Conceptualising patient empowerment: a mixed methods study. BMC Health Serv Res 2015 Jul 1;15:252. [doi: 10.1186/s12913-015-0907-z] [Medline: 26126998]
20. Girginov V, Moore P, Olsen N, Godfrey T, Cooke F. Wearable technology-stimulated social interaction for promoting physical activity: a systematic review. Cogent Soc Sci 2020;6(1):1742517. [doi: 10.1080/23311886.2020.1742517]
21. Media statistics portal. URL: https://stat.kisdi.re.kr [accessed 2026-04-29]

Abbreviations

DID: difference-in-differences
RR: rate ratio
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Stone; submitted 22.Nov.2025; peer-reviewed by AAM Rizwan, J Lee; revised version received 24.Mar.2026; accepted 28.Apr.2026; published 25.May.2026.

Please cite as:
Choi E, Choi S, Jang SY
Association Between Wearable Device Adoption and Health-Related Lifestyle Behaviors: Retrospective Cohort Study
J Med Internet Res 2026;28:e88276
URL: https://www.jmir.org/2026/1/e88276
doi:10.2196/88276

© Eunjeong Choi, Seoyeong Choi, Suk-Yong Jang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Experiences and Perceived Influence of the Artificial Intelligence–Based Health Education Accurately Linking System (AI-HEALS) on Health Behaviors Among Patients With Type 2 Diabetes: Qualitative Study

Jing Wang[1], BM; Yibo Wu[2], PhD; Yang Jiang[3], BM; Rantong Bao[4], MPH; Xinbao Gu[5], BM; Bingyang Kong[6], BM; Xinying Sun[1], PhD

[1]School of Public Health, Peking University, 38 Xueyuan Road, Haidian District, Beijing, China

[2]Department of Nursing, the Fourth Affiliated Hospital of School of Medicine, International School of Medicine, International Institutes of Medicine, Zhejiang University, Yiwu, China

[3]Jitang College, North China University of Science and Technology, Tangshan, China

[4]Center for Clinical Epidemiology Research, Affiliated Hospital of Inner Mongolia Medical University, Hohhot, China

[5]Capital Medical University, Beijing, China

[6]School of Traditional Chinese Medicine, Capital Medical University, Beijing, China

Corresponding Author:
Xinying Sun, PhD
School of Public Health, Peking University, 38 Xueyuan Road, Haidian District, Beijing, China

Abstract

Background: The management of type 2 diabetes requires sustained self-management across diet, physical activity, medication adherence, and blood glucose monitoring; however, maintaining these behaviors in daily life remains difficult for many patients. Artificial intelligence–enabled and mobile health interventions have shown promise in supporting diabetes education and self-management, but evidence on how patients actually experience and use such systems in real-world primary care remains limited.

Objective: This study aimed to explore patients' experiences with the Artificial Intelligence–Based Health Education Accurately Linking System (AI-HEALS) and its perceived influence on self-management behaviors among patients with type 2 diabetes.

Methods: This explanatory qualitative study was nested within the intervention arm of a cluster randomized controlled trial of AI-HEALS. Purposive maximum-variation sampling was used to recruit participants who varied by sex, age, diabetes duration, hemoglobin A_{1c} level, digital literacy, and level of platform use. Of the 25 patients approached from the intervention arm, 17 agreed to participate. Semistructured interviews were conducted 3 months after the intervention (August to December 2023) with participants recruited from 45 communities in the Daxing and Shunyi districts of Beijing, China. The qualitative component was designed to explain perceived mechanisms of behavior change, contextual facilitators and barriers, and implementation-related experiences not captured by trial outcomes alone. Interview transcripts were analyzed thematically in NVivo 12 by 2 independent researchers using consensus coding, an audit trail, and member checking.

Results: Participants' experiences with AI-HEALS were reflected in four themes: (1) catalyzing health awareness and concern, (2) empowering self-management practices, (3) navigating usability and engagement, and (4) enhancing psychological adaptation. Participants perceived that AI-HEALS made diabetes more visible in daily life through repeated reminders and accessible educational content; supported practical self-management decisions related to diet, physical activity, medication adherence, and glucose monitoring; and, in some cases, improved confidence while reducing uncertainty and diabetes-related stress. The findings also suggested meaningful variation in how the intervention was experienced, particularly in relation to prior illness experience, routine stability, social support, and digital confidence. Many participants mainly engaged with lower-burden features such as pushed articles and reminders, whereas more interactive functions, such as the chatbot, were used less often.

Conclusions: Patients generally perceived AI-HEALS as a useful source of ongoing education, behavioral prompting, and everyday support for diabetes self-management. The findings suggest that artificial intelligence–enabled education may work not only by increasing knowledge but also by reinforcing awareness, translating guidance into feasible daily action, and supporting psychological adaptation. At the same time, the intervention's perceived usefulness depended on whether its content was understandable, low burden, and compatible with users' routines. These results support further refinement of literacy-sensitive, layered digital interventions and their integration into routine community-based primary care.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2300068952; https://tinyurl.com/r8hjwetr

XSL•FO
RenderX

(J Med Internet Res 2026;28:e84605) doi:10.2196/84605

KEYWORDS

artificial intelligence; type 2 diabetes; mobile health; self-management; qualitative research

Introduction

Diabetes, particularly type 2, has emerged as a significant global public health challenge. Over recent decades, the global diabetic population aged 20 to 79 years has surged to 537 million, with projections of 643 million by 2030 and 783 million by 2045. In China, the adult diabetes prevalence is notably high at 11.9%, affecting 140.9 million individuals, thereby significantly impacting quality of life and imposing a substantial health and economic burden on society [1,2]. Diabetes also elevates the risk of severe complications, including heart disease, stroke, kidney disease, retinopathy, and lower limb amputation [3,4]. Therefore, effective management strategies are urgently needed to address the rising number of patients and the diverse medical environments.

Traditional management approaches primarily focus on patient education, lifestyle modification, regular blood sugar monitoring, and medication adherence [5,6]. However, the complexity of daily insulin injections, medication management, inflexible lifestyles, and continuous health monitoring imposes cognitive burdens on patients, often leading to suboptimal outcomes [7]. Additionally, traditional management faces challenges such as patient nonadherence, underscoring the need for innovative solutions that simplify diabetes management and make it more accessible and effective for patients [8].

Mobile health (mHealth) technology leverages smartphones and wearables to deliver health information and services, enabling real-time patient monitoring and personalized communication with health care providers [9,10]. Systematic reviews and meta-analyses have consistently demonstrated that mHealth applications can significantly enhance blood glucose control and improve diabetes-related health behaviors in patients with type 2 diabetes mellitus (T2DM) [11,12]. When integrated with artificial intelligence (AI), these platforms can analyze large datasets to identify risk factors and deliver tailored interventions, thereby reducing the need for frequent face-to-face consultations and alleviating health care system burdens. Common features, such as reminder systems and incentives, have also been shown to improve treatment adherence [12].

AI—encompassing machine learning, natural language processing (NLP), and other technologies—aims to simulate human intelligence and is increasingly applied in diabetes care [13-15]. AI plays a crucial role in self-management by supporting blood glucose monitoring, dietary and exercise guidance, disease prediction, and personalized care [16-19]. The World Health Organization has recognized AI's potential to enhance health literacy and patient engagement in chronic disease education. Several AI-driven health education systems have demonstrated clinical benefits: for example, the Smart AI-Enabled Diabetes system developed by Alotaibi et al [20] significantly reduced hemoglobin A1c (HbA_{1c}) levels and improved diabetes knowledge in Saudi patients, whereas Chen et al [21] reported similar glycemic improvements using an mHealth-based education program for T2DM patients initiating insulin therapy.

This study used the Artificial Intelligence–Based Health Education Accurately Linking System (AI-HEALS), deployed on WeChat [22]. With over 1.2 billion monthly active users, WeChat's integrated ecosystem—Mini Programs, Official Accounts, and payments—allows health management services to be embedded in everyday use, mitigating the retention and adherence problems common to standalone apps [22,23]. Its broad reach across urban and rural populations, including older adults, enhances access and equity; private group chats and privacy controls help safeguard sensitive health data [23]. Unlike Weibo's primarily public broadcasting model, WeChat's interaction-centric social communication fosters sustained peer support that facilitates behavior change. Leveraging this platform, AI-HEALS used low-cost, widely available mobile communication to deliver continuous, convenient diabetes management directly to patients' homes.

Based on the above, this qualitative study aims to answer the following research questions: (1) How do patients with T2DM perceive the influence of AI-HEALS on their self-management behaviors (diet, physical activity, medication adherence, and glucose monitoring)? (2) What are patients' experiences with the usability and acceptability of the AI-HEALS system? (3) What facilitators and barriers to behavior change do patients encounter when using AI-HEALS, and how do features such as personalized reminders and educational content delivery shape these experiences?

Methods

Study Design and Trial Context

This study was designed as an explanatory qualitative study nested within the intervention arm of a cluster randomized controlled trial. It was intended to clarify perceived mechanisms of behavior change, contextual influences on use, and implementation-related experiences that were not captured by trial outcomes alone. The parent trial, registered at the Chinese Clinical Trial Registry (ChiCTR2300068952; registered on February 3, 2023), aimed to evaluate the effectiveness of AI-HEALS on glycemic control and self-management behaviors. In contrast, this qualitative substudy was designed to address questions that quantitative outcomes alone could not answer: how participants perceived the intervention to influence daily self-management, which contextual factors facilitated or constrained its use, and which implementation-related features shaped engagement in community-based primary care. Thus, the qualitative component was not intended to retest efficacy but to explain perceived mechanisms, user experiences, and real-world delivery processes within the intervention arm. The trial protocol has been published elsewhere [24].

XSL•FO
RenderX

Study Setting and Participants of the Parent Trial

The parent trial was conducted from August to December 2023 across 45 community health centers in the Daxing and Shunyi districts of Beijing. A total of 406 participants were enrolled using cluster sampling, with each community health center recruiting 8 to 10 patients.

- Inclusion criteria for the trial were residents aged 18 to 75 years with a diagnosis of T2DM (fasting plasma glucose ≥7 mmol/L, 2-h plasma glucose ≥11.1 mmol/L, or HbA_{1c} ≥6.5%); permanent residency in Beijing; ability to use smartphones and WeChat; willingness to adhere to the study; and not participating in other studies.
- Exclusion criteria were type 1 diabetes, gestational or secondary diabetes; severe diabetes complications; recent radiotherapy or chemotherapy; severe intellectual disability, Alzheimer disease, or other mental disorders; or prior use of psychotropic drugs.

Clusters were randomized 1:1 using a computer-generated sequence, assigning 208 participants to the intervention group (standard diabetes primary care plus AI-HEALS) and 198 participants to the control group (standard diabetes primary care alone). Complete blinding of participants and researchers was not feasible due to the nature of the intervention.

Qualitative Substudy Sampling and Recruitment

For this qualitative substudy, a purposive maximum variation sampling strategy was used to select information-rich cases from the intervention arm. Sampling criteria were designed to capture a broad range of experiences and perspectives, including sex (male/female), age (<60 y/≥60 y), diabetes duration (<5 y/5–10 y/>10 y), HbA_{1c} levels (<7%/≥7%), self-reported digital literacy (high/moderate/low), and level of platform use (frequent/moderate/infrequent based on login frequency and feature usage).

From the 208 participants in the intervention arm, we purposively approached 25 individuals who represented diverse combinations of the sampling criteria. Recruitment was conducted via telephone by a trained research assistant who explained the purpose of the qualitative interview and invited participation. Of the 25 individuals approached, 17 agreed to participate, yielding a response rate of 68%. Reasons for nonparticipation among the 8 decliners included lack of interest (n=3, 38%), time constraints due to work or family obligations (n=3, 38%), and concerns about privacy (n=2, 25%).

The sample size was determined by data saturation, defined as the point at which no new themes emerged from subsequent interviews. Saturation was achieved after 17 interviews, with no new themes emerging from the final 2 interviews. The final sample consisted of 17 AI-HEALS users from the intervention arm.

Introduction to the AI-HEALS

The intervention was delivered through a 2-layer architecture: the backend AI-HEALS and the frontend "Beiyi Sugar Manager" WeChat Mini Program (Tencent Holdings Ltd.). AI-HEALS used a hybrid logic architecture rather than a fully generative model. Specifically, the conversational function relied on vendor-provided NLP and natural language understanding modules to identify user intent and extract key terms, after which the system retrieved preconfigured responses from a curated diabetes education knowledge base. In parallel, reminder delivery and content recommendation were governed primarily by deterministic rule-based logic using baseline patient profile variables (eg, age, duration of diabetes, HbA_{1c} level, and self-reported digital literacy), recent article reading history, manual health logs, and regimen-related schedules. This architecture allowed the system to combine conversational access with stable, auditable delivery of education and reminders.

The "Beiyi Sugar Manager" Mini Program was the patient-facing application embedded within the WeChat ecosystem, through which participants accessed all intervention components without installing a standalone app. All communications—including educational content, AI chatbot interactions, automated reminders, and message notifications—were delivered via the WeChat Mini Program interface.

Educational messages were prioritized according to broad self-management domains (diet, exercise, medication, monitoring, complications) and adjusted using recent engagement patterns, such as categories previously read, reading duration, and frequency of health data entry. Reminder schedules were anchored to user-entered medication and monitoring routines and could be intensified when recent logging was irregular. The educational knowledge base and decision rules were developed by the research team with clinical and public health input before trial deployment. Because the dialog platform was commercially provided, the underlying training corpus, model weights, and proprietary internal validation metrics were not available to the research team for publication.

The mini program provides 4 main functional modules powered by the AI-HEALS backend:

1. AI-powered conversational agent (chatbot).

- Frontend (patient view): Within the "Beiyi Sugar Manager" Mini Program, participants can access an AI chatbot by typing questions about T2DM diet, exercise, and medication management. The chatbot provides answers in a conversational format, offering immediate responses to user queries within the scope of the system's diabetes knowledge base (Figure 1).

XSL•FO
RenderX

Figure 1. Interaction with the artificial intelligence chatbot in the "Beiyi Sugar Manager" mini program. The screenshot illustrates a participant querying about dietary recommendations and receiving a personalized response based on the system's knowledge base. Users initiate conversations by typing questions, and the chatbot provides evidence-based information tailored to individual profiles.

XSL•FO
RenderX

- Backend (AI-HEALS support): This functionality is powered by the Emotibot-produced Bot Factory dialog AI platform (ZhuJian Intelligence Technology) [25]. The platform uses NLP and natural language understanding algorithms to analyze participant queries, extract contextual information, and retrieve the most relevant responses from the knowledge base. Responses are retrieved and prioritized using user profile attributes (age, duration of diabetes, HbA_{1c} level, and digital literacy) together with prior interactions, enabling tailored delivery of health education within the intervention context.

2. Personalized educational content library.

- Frontend (patient view): The mini program features an "Education Center" that provides participants with access to a structured library of educational articles (Figure 2). Content is categorized into 7 areas, including "Dietary Guidance" (23 articles), "Exercise Guidance" (17 articles), "Medication Guidance" (18 articles), "Diabetes Complications" (9 articles), and "Disease Prevention" (23 articles). Articles are pushed to users based on their reading history and preferences.
- Backend (AI-HEALS support): The AI-HEALS backend tracks user activity—including number of articles read, duration, and categories of interest—and applies a rule-based personalization algorithm. Educational content is pushed 1 to 3 times per week according to user preferences, engagement patterns, and theoretically grounded behavior-change principles, ensuring both relevance and reinforcement of self-management behaviors.

Figure 2. Educational content delivery and user engagement features. The panel shows the "Education Center" interface with categorized articles (eg, dietary guidance, exercise guidance).

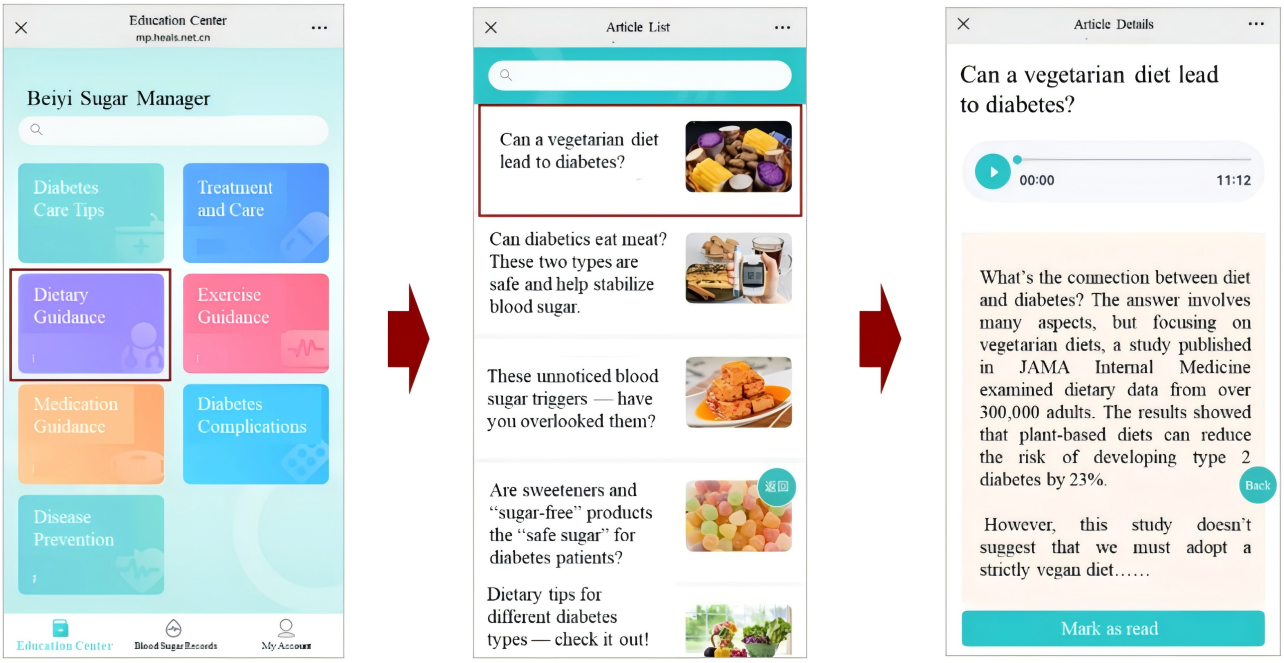

3. Physiological monitoring and reminder system.

- Frontend (patient view): The mini program allows participants to manually log key physiological indicators, such as blood glucose levels (fasting and postprandial) and blood pressure (Figure 3). It also delivers automated reminders for medication intake and blood glucose checks directly through the WeChat interface, appearing similarly to messages from a contact.
- Backend (AI-HEALS support): The reminder scheduling engine uses logged health data and prescribed medication regimens to generate individualized notifications. Logged health data are used to adjust reminder delivery, for example, by increasing reminder frequency when recent logging is inconsistent or when readings suggest a need for closer monitoring.

XSL•FO
RenderX

Figure 3. Blood glucose monitoring and recording interface. Users can log their blood glucose values (fasting, postprandial) and view trends over time. The system provides feedback on readings and integrates this data with reminder functions.

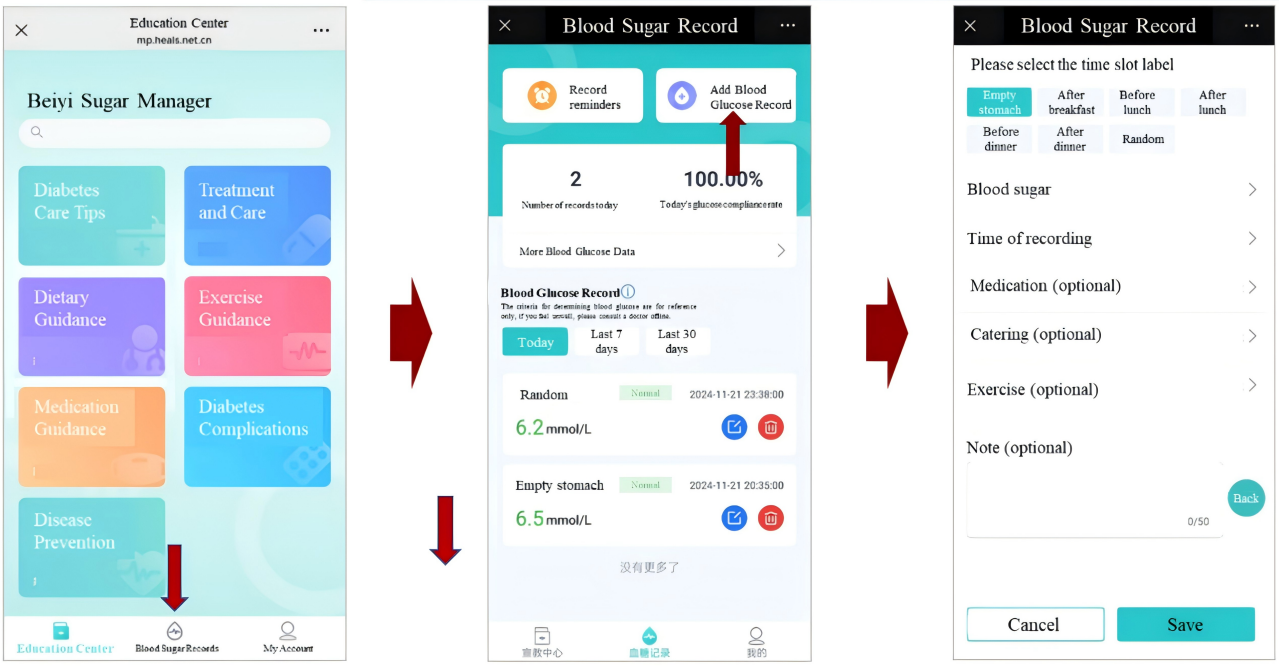

4. Theory-driven automated messaging.

- Frontend (patient view): Participants receive 1 to 3 automated messages per week with tips and information related to T2DM self-management, supporting both the initiation and maintenance of healthy behaviors (Figure 4). These messages are delivered through the "Beiyi Sugar Manager" Mini Program.
- Backend (AI-HEALS support): The content and timing of these automated messages are theoretically grounded in the Multi-Theory Model (MTM) of health behavior change (Figure 5) [26]. The backend system operationalizes the MTM framework by delivering content tailored to support users through the stages of behavior initiation (eg, starting an exercise routine) and behavior maintenance (eg, sustaining long-term dietary changes). Messages are selected from the "Education Center" library, and message selection accounts for prior engagement, health metrics, and interaction history, enabling rule-guided tailoring to individual user needs within the intervention.

Figure 4. Automated reminder system for medication and monitoring. Participants receive scheduled reminders through the WeChat mini program interface, similar to receiving messages from a contact. Reminders are personalized based on individual medication regimens and monitoring schedules.

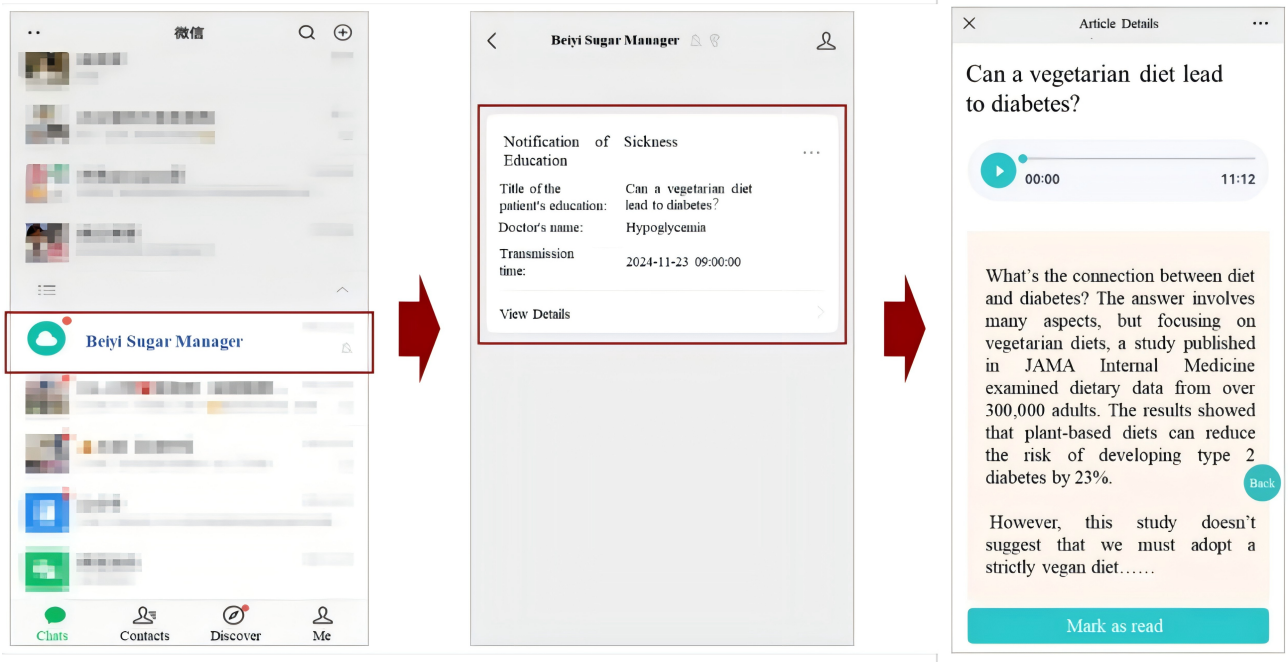

XSL•FO
RenderX

Figure 5. Multi-Theory Model (MTM) of health behavior change framework. The model illustrates the 2-stage process of behavior change (initiation and maintenance) and the theoretical constructs (eg, participatory dialog, behavioral confidence, changes in the physical environment) that informed the intervention design.

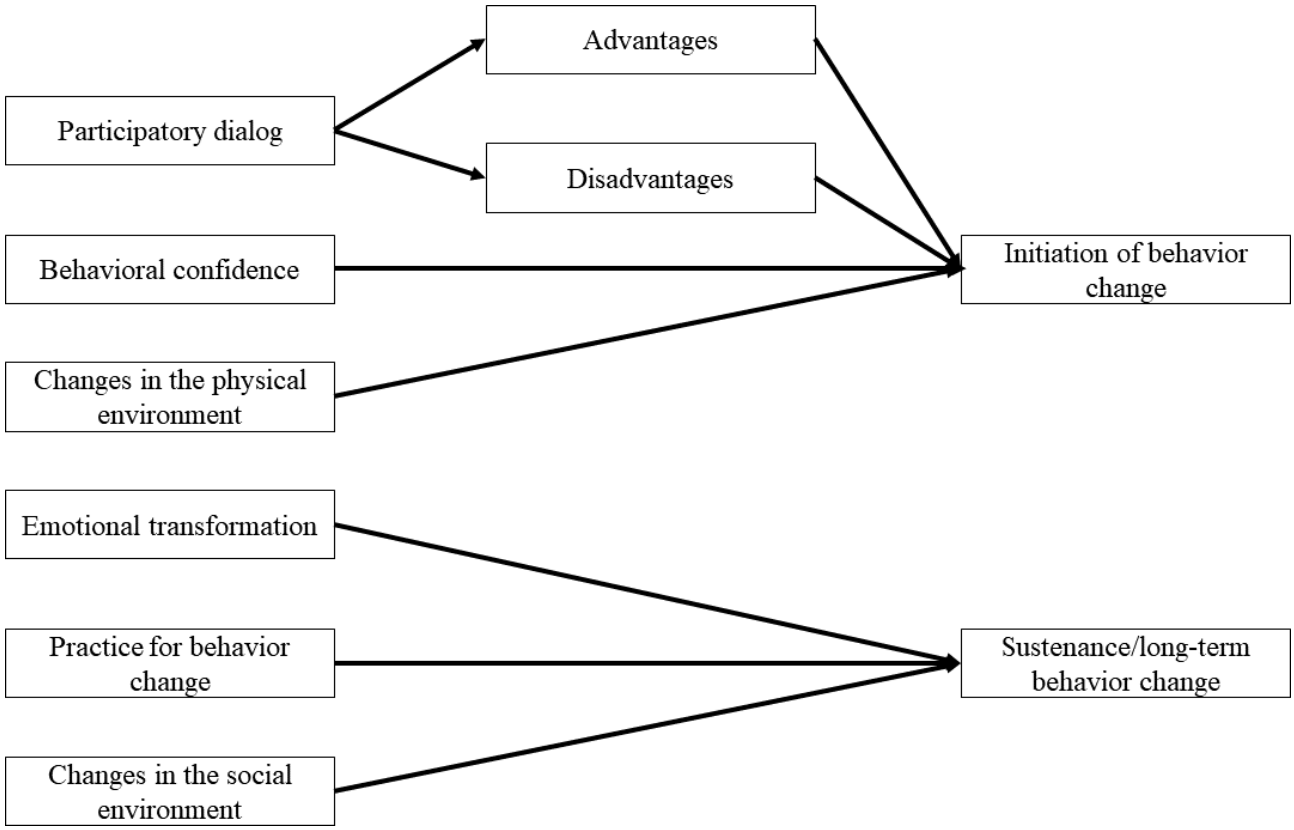

Data Collection and Analysis

Semistructured interviews were conducted with all 17 participants 3 months after the intervention to explore their experiences and perceived influence of AI-HEALS on self-management behaviors, including diet, physical activity, and continuous glucose monitoring. Interviews, lasting 30 to 60 minutes, were held either in person or by telephone. Participants were encouraged to recount context-specific scenarios and critical incidents, describe their interactions with the system and challenges encountered, and propose suggestions for improvement related to behavior change. All interviews were audio recorded with permission and transcribed verbatim.

Thematic analysis of the transcripts was performed using NVivo software (version 12; QSR International). To enhance transparency and reproducibility, an audit trail of coding decisions was maintained. Two researchers (JW and YJ) independently coded the transcripts. After initial coding of the first 3 transcripts, the research team discussed the preliminary codebook and resolved any discrepancies through consensus; subsequent consensus meetings were held to refine codes and develop themes. The coding process was primarily inductive, allowing themes to emerge from the data. Following the completion of initial thematic analysis, member checking was conducted with 5 participants, who were asked to review a summary of the findings and provide feedback on whether the interpretations resonated with their experiences. This feedback confirmed the credibility of the themes and did not necessitate substantial revisions.

Ethical Considerations

This study was approved by the Biomedical Ethics Committee of Peking University (IRB00001052-22058; approved on June 6, 2022) and registered with the Chinese Clinical Trial Registry (ChiCTR2300068952; registered on February 3, 2023). Written informed consent was obtained from all interview participants before data collection. Participants did not receive financial compensation for participation. Several measures were implemented to protect participant anonymity and data security within the WeChat environment. All data transmitted through the “Beiyi Sugar Manager” mini program were encrypted using industry-standard protocols. Participant identifiers were replaced with unique study codes, and data were stored on secure servers accessible only to authorized research personnel. Participants were informed that while WeChat’s platform has its own privacy policy, the research team implemented additional safeguards, including not collecting WeChat IDs or contact lists and advising participants not to share personal health information through unsecured channels. The mini program was designed to comply with China’s Personal Information Protection Law and relevant data security regulations.

Results

Patient Demographics

Demographic characteristics of the patients are shown in Table 1.

XSL•FO
RenderX

Table . Patient demographics.

Variables	Total (N=17)[a]
BMI, mean (SD)	26.33 (3.60)
Waist circumference, mean (SD)	97.88 (13.45)
Sex, n (%)	
Female	6 (35.29)
Male	10 (58.82)
Unknown	1 (5.88)
Ethnicity, n (%)	
Han	16 (94.12)
Unknown	1 (5.88)
Income (CNY)[b], n (%)	
＞9000	4 (23.53)
≤3000	4 (23.53)
3001 - 9000	8 (47.06)
Unknown	1 (5.88)
Occupation status, n (%)	
Employed	7 (41.18)
No fixed occupation/freelance	2 (11.76)
Retired	5 (29.41)
Unemployed	2 (11.76)
Unknown	1 (5.88)
Marital status, n (%)	
Married	16 (94.12)
Unknown	1 (5.88)
Hukou, n (%)	
Agricultural	7 (41.18)
Nonagricultural	9 (52.94)
Unknown	1 (5.88)
Current health insurance, n (%)	
Beijing medical insurance	15 (88.24)
No medical insurance	1 (5.88)
Unknown	1 (5.88)
People living with you in the past 2 months, n (%)	
Living alone	1 (5.88)
Living with family members	15 (88.24)
Unknown	1 (5.88)
Education, n (%)	
Bachelor’s degree	1 (5.88)
Junior college	3 (17.65)
Junior high school	6 (35.29)
Secondary vocational school	1 (5.88)
Senior high school	5 (29.41)

XSL•FO
RenderX

Variables	Total (N=17)[a]
Unknown	1 (5.88)
Occupation, n (%)	
Agricultural/forestry/animal husbandry/fishery worker	1 (5.88)
Designer/design professional	1 (5.88)
Finance/accounting/cashier/audit	1 (5.88)
Freelancer	2 (11.76)
Full-time homemaker	1 (5.88)
Industrial/manual worker	1 (5.88)
Marketing/sales/business	1 (5.88)
Medical/health care staff	1 (5.88)
Product/operations staff	1 (5.88)
Retired	2 (11.76)
Service industry staff	3 (17.65)
Technical developer/engineer	1 (5.88)
Unknown	1 (5.88)
Type of health insurance, n (%)	
Government-funded medical care	1 (5.88)
New rural cooperative medical scheme	4 (23.53)
No medical insurance	1 (5.88)
Unknown	1 (5.88)
Urban employee basic medical insurance	6 (35.29)
Urban resident basic medical insurance	4 (23.53)
Any other household members with diabetes, n (%)	
Father	1 (5.88)
Father and siblings	1 (5.88)
Mother	2 (11.76)
Mother and siblings	1 (5.88)
None	7 (41.18)
Siblings	3 (17.65)
Unknown	2 (11.76)
Duration since diabetes diagnosis, n (%)	
1 year	1 (5.88)
2 years	1 (5.88)
3 years	1 (5.88)
10 years	4 (23.53)
12 years	2 (11.76)
20 years	1 (5.88)
Unknown	7 (41.18)

[a]Information for 1 participant (Q) was missing from the dataset; they are included in the “Unknown” category where applicable.

[b]Chinese yuan; an exchange rate of US $1=CNY ¥6.906 is applicable.

The study included 17 participants with T2DM, with a mean BMI of 26.33 (SD 3.60) kg/m² and a mean waist circumference of 97.88 (SD 13.45) cm. The majority were male (n=10, 58.82%) and of Han ethnicity (n=16, 94.12%), with approximately half reporting a monthly income of 3001 to 9000 CNY (based on an exchange rate of US $1=CN ¥6.906). Participants varied in

XSL•FO
RenderX

occupation and education, with 41.18% (n=7) employed, 29.41% (n=5) retired, and educational attainment ranging from junior high school to a bachelor's degree. Most participants were covered by Beijing medical insurance (n=15, 88.24%) and lived with family members, while 47.06% (n=8) reported a family history of diabetes. The duration since diagnosis ranged from 1 to 20 years, with 41.18% (n=7) of participants reporting it as unknown.

A comprehensive overview of each participant's demographic and health-related information is available in Multimedia Appendix 1.

Qualitative Findings

Participants accessed the AI-HEALS intervention via the "Beiyi Sugar Manager" WeChat Mini Program (hereafter referred to as "the mini program"). Four main themes emerged from the analysis (Table 2), each with several subthemes, illustrating the perceived impact of the AI-HEALS intervention on participants' self-management journeys. Below, each theme and subtheme is described with illustrative quotations to demonstrate how the interpretation was grounded in participants' accounts.

XSL•FO
RenderX

Table . Summary of qualitative themes, subthemes, and illustrative quotations.

Theme and subtheme	Sample quotations from interviews
Catalyzing health awareness and concern	
Deepened understanding of complications.	Patient D: “Complications are quite serious… like ‘three more and one less,’ foot ulcers, skin itching, or blindness. I’ve had my eye fundus checked, no bleeding… I usually pay attention to my diet to avoid large blood sugar fluctuations, because high blood sugar can damage blood vessels and form plaques, which is bad for the brain and body.”
	Patient N: “I mainly worry about the heart, eyes, and feet. My eyes are sometimes blurry; I’m most afraid of bleeding in the eye fundus… It’s very serious, especially the eyes. Sometimes I worry about going blind.”
Reinforcing the importance of self-management.	Patient D: “To live a long life and maintain a good quality of life, you must have the confidence and perseverance to persist in controlling your blood sugar. You can’t ignore your diet, especially during the holidays. The saying ’illness enters by the mouth’ is very true.”
	Patient F (on using the mini-program): “It definitely [helped], it gave me that awareness in my mind… I pay more attention now.”
Empowering self-management practices	
Dietary improvements.	Patient B: “Before, I thought anything with sugar was off-limits, I wouldn’t even eat fruit. Now, after reading the mini-program, I can judge which foods are lower in sugar and can be eaten in moderation… I know some fruits, like firmer apples, aren’t that high in sugar.”
	Patient O: “The dietary structure has changed. [the program] tells you what to eat and how to eat it.”
	Patient K: “There are things I want to eat, but now I control the amount. I’m not so casual about it anymore.”
Enhanced physical activity.	Patient I: “There’s been improvement. Before, I forced myself to walk over 10,000 steps a day… But the doctor [via the program] suggested that based on my age, it’s better to limit it to 5000 to 7000 steps. Now my walks are more relaxed, about half an hour, mainly strolling instead of forced exercise.”
	Patient P: “When the weather is bad, I do half an hour of activity at home after meals. Stand when you can, sit instead of lie down. I do some arm and leg stretches…”
	Patient A: “My exercise time has increased. Everyone cares about their health; [the program] helps not only with blood sugar but also with blood lipids.”
Improved adherence to medication and monitoring.	Patient H: “The blood sugar testing and uploading function… it’s very convenient. Yes, it’s really helpful for my health.”
	Patient P: “I definitely follow the reminders. I have to take my medication three times a day with meals. Even if I don’t eat, I still have to take it, otherwise I might get hypoglycemia.”
	Patient N: “[The mini-program] at least reminds me today to eat less of the foods that will spike my blood sugar. It works. If it didn’t remind me, I wouldn’t know, and it wouldn’t help.”
Navigating usability and engagement	
Primary engagement with passive features.	Patient A: “I mainly just read the popular articles.”
	Patient L: “I mainly look at the articles that are sent.”
	Patient Q: “I basically don’t use other functions; I mainly just read the articles.”
Variable engagement and digital literacy.	Patient K: “I don’t know how to do the other things. Sometimes I just open it and look at the articles… I can understand the simple ones, but the complicated ones I can’t.”
	Patient C: “For someone like me with a simple, regular life, I basically don’t need the reminders. But for younger patients, this kind of reminder would be very useful.”

XSL•FO
RenderX

Theme and subtheme	Sample quotations from interviews
Enhancing psychological adaptation	
Fostering self-efficacy and confidence.	Patient D: “I have the confidence and will to control my blood sugar.”
	Patient H: “One hundred percent!” (when asked about his confidence to continue).
	Patient Q: “I definitely have the confidence, and the will is strong. Everyone wants to live. Now, I can manage my situation on my own, without relying on others.”
Alleviating anxiety and distress.	Patient H: “It [the mini-program] helps, I pay much more attention now. Before I used it, I didn’t know these basic facts. Now I’ve learned them through the mini-program.”

Theme 1: Catalyzing Health Awareness and Concern

This theme captures participants’ experiences with the intervention, particularly how its educational content and reminders shaped their understanding of diabetes and its potential consequences and how it influenced their motivation for self-management.

Subtheme 1.1: Deepened Understanding of Complications

Many participants reported that the information provided through the mini program increased their awareness of the serious long-term complications of diabetes, such as retinopathy, cardiovascular disease, and diabetic foot. For example, Patient D described complications as “quite serious,” whereas Patient N said she was “most afraid of bleeding in the eye fundus” and “going blind.” This heightened perceived risk underscored the importance of proactive management.

Subtheme 1.2: Reinforcing the Importance of Self-Management

Beyond knowledge, the system’s regular prompts and accessible information served as a constant reminder of the need for vigilance, reinforcing participants’ commitment to their self-management routines. Patient D emphasized that, to maintain a good quality of life, one must persist in controlling blood sugar, and Patient F noted that the mini program “gave me that awareness in my mind,” making her “pay more attention now.”

Theme 2: Empowering Self-Management Practices

This theme reflects participants’ experiences in translating increased awareness into behavioral changes, with the mini program serving as a supportive tool.

Subtheme 2.1: Dietary Improvements

The most frequently cited area of change was diet. Participants reported that the educational articles helped them distinguish between healthier and less healthy food choices, understand portion sizes, and adopt more flexible but better-informed eating habits. Patient B, for example, explained that after using the mini program, he no longer believed that all fruit was forbidden and felt better able to judge which foods could be eaten in moderation.

Subtheme 2.2: Enhanced Physical Activity

The mini program also influenced participants’ exercise habits. Some reported becoming more consistent, whereas others adjusted their routines based on the guidance received. Patient I described shifting from a forced goal of more than 10,000 steps per day to more age-appropriate walking, whereas Patient P reported doing indoor activity after meals when the weather was poor.

Subtheme 2.3: Improved Adherence to Medication and Monitoring

The reminder functions were highlighted as a key facilitator for medication adherence and regular blood glucose monitoring, particularly for those with less structured daily routines. Patient H described the testing and uploading function as “very convenient,” and Patient P said she “definitely” followed the medication reminders.

Theme 3: Navigating Usability and Engagement

This theme explores how participants interacted with the technology itself, revealing patterns of use, preferences for certain features, and barriers to full engagement.

Subtheme 3.1: Primary Engagement With Passive Features

Most participants reported primarily using lower-burden features of the system, especially pushed educational articles and reminders, rather than more active functions. As Patients A, L, and Q each noted in different ways, they mainly read the articles that were sent and rarely used other modules.

Subtheme 3.2: Variable Engagement and Digital Literacy

Engagement levels appeared closely linked to digital literacy and personal preference. Patient K stated that he could understand “the simple ones” but not the “complicated ones,” whereas Patient C felt that reminders were less necessary because his daily routine was already highly structured. Together, these accounts suggest that lower engagement with some functions did not necessarily mean low perceived usefulness.

Theme 4: Enhancing Psychological Adaptation

Beyond practical behavior change, participants described a positive psychological impact from using the system, which contributed to a greater sense of control and a reduced emotional burden.

XSL•FO
RenderX

Subtheme 4.1: Fostering Self-Efficacy and Confidence

Successfully implementing the information and reminders from the mini program appeared to build participants' confidence in their ability to manage their condition effectively. Patient D said she had the "confidence and will" to control her blood sugar, and Patient Q similarly reported feeling able to manage the condition "on my own, without relying on others."

Subtheme 4.2: Alleviating Anxiety and Distress

For some participants, the knowledge and structure provided by the system helped mitigate anxiety and distress associated with diabetes. Patient H said that learning basic facts through the mini program reduced uncertainty, and Patient C described the clear guidance on diet and medication as reassuring.

Discussion

This qualitative study explored the experiences and the perceived influence of AI-HEALS on self-management behaviors among 17 patients with type 2 diabetes. Four overarching themes emerged: catalyzing health awareness and concern, empowering self-management practices, navigating usability and engagement, and enhancing psychological adaptation. Together, these findings suggest that an AI-enabled, theory-informed mHealth intervention may support self-management not only by delivering information but also by embedding reminders, guidance, and reinforcement into everyday routines. At the same time, participants' accounts indicate that the intervention did not function in exactly the same way for all users and that these differences were closely related to prior illness experience, routine stability, social support, and digital confidence.

Catalyzing Health Awareness and the Knowledge-Behavior Gap

Participants' reports that AI-HEALS deepened their understanding of diabetes complications align with the literature showing that knowledge is an important but often insufficient condition for behavior change in diabetes self-management [27,28]. In this study, awareness often extended beyond general education and became tied to concrete concerns about blindness, cardiovascular disease, and diabetic foot. This suggests that the intervention helped translate abstract risk into a more personally meaningful understanding.

The findings also indicate that the value of AI-HEALS lay not only in transmitting information but also in sustaining attention to diabetes in daily life. Repeated prompts and accessible educational content appeared to keep the condition salient over time, thereby helping to address the knowledge-behavior gap described in previous diabetes education research [29]. In this respect, AI-HEALS differs from more conventional education delivered in discrete sessions [5,6] by providing ongoing reinforcement. This pattern is also consistent with the MTM, particularly its emphasis on participatory dialog as a motivational process in behavior change [26].

At the same time, the nature of this awareness appeared to vary across participants. For individuals with a longer duration of diabetes, such as Patients D and N, the intervention seemed to reinforce and sharpen existing illness understanding, although the content of that awareness differed: one account reflected a broader physiological understanding of vascular risk, whereas another focused more strongly on feared complications such as blindness. By contrast, for a newly diagnosed participant such as Patient O, the intervention appeared to play a more foundational role in shaping initial self-management understanding. This pattern suggests that AI-enabled education may function differently across the diabetes trajectory, serving as reinforcement for some users and as initial structuring support for others. Family context may also have influenced how risk information was received, as prior household exposure to diabetes could heighten vigilance toward complications and self-management [30].

Empowering Self-Management: Beyond Generic Recommendations

The sense of empowerment described by participants resonates with Anderson and Funnell [31] patient empowerment model, which emphasizes active patient ownership of diabetes management. In this study, the perceived value of AI-HEALS lay not simply in offering recommendations but also in helping participants translate broad self-management principles into concrete, day-to-day decisions.

Dietary change was the most frequently discussed area of improvement, which is consistent with systematic reviews identifying nutrition as one of the most responsive domains in diabetes self-management interventions [10,11]. Participants' accounts suggested a movement away from rigid restriction toward more flexible and informed judgment, a pattern broadly in line with work on personalized nutrition and tailored self-management support [19]. Importantly, however, these changes did not appear uniform across users. For newly diagnosed participants, dietary guidance seemed to support the establishment of initial management routines. For others with more experience, it appeared to refine existing practices rather than create them from scratch. These differences suggest that the same digital intervention may support both habit formation and habit adjustment, depending on patients' prior self-management history.

The findings on physical activity similarly indicate that participants valued guidance not just when it encouraged "more exercise" but also when it made activity more appropriate, feasible, and compatible with daily life. The accounts of individualized walking targets and indoor alternatives suggest that users valued optimized guidance over generic behavioral goals. This is in line with emerging evidence that, for adults managing chronic disease, the appropriateness of activity may matter as much as quantity [17]. The case of Patient P also suggests that such support may be particularly relevant for users who manage much of their care independently in everyday life.

At the same time, the findings point to an important tension between personalization and accessibility. Tailored guidance was clearly valued, but some participants also indicated difficulty understanding more complex content. This was especially evident in the account of a participant with lower educational attainment who remained engaged with simpler materials while avoiding more complicated ones. This suggests

XSL•FO
RenderX

that digital support may be less effective when information is clinically relevant but insufficiently accessible to users with lower literacy or lower digital confidence. For intervention design, these findings support the importance of tailoring that is not only behaviorally targeted but also sensitive to comprehension and usability.

Navigating Usability and Engagement in Everyday Use

A notable finding was that many participants mainly engaged with lower-burden features, such as pushed articles and reminders, whereas more active functions, including the chatbot, were used less often. This pattern aligns with international research on digital diabetes interventions, which frequently report challenges in sustaining engagement and note that usage intensity does not necessarily correspond directly to perceived benefit [28,32].

In the present study, passive engagement still appeared meaningful. For some participants, reading articles and receiving reminders was sufficient to support awareness, routine, and behavioral adjustment. This pattern may reflect the effort-reward balance described in prior health technology research, whereby users gravitate toward features that provide clear benefits with minimal cognitive burden [33]. It may also suggest that interactive components, such as chatbots, need to demonstrate clearer added value if they are to become central components of engagement.

Participants' accounts further indicate that engagement was shaped by the interaction of digital confidence, education, prior routines, and perceived need. The contrast between a participant with lower educational attainment who found "complicated" features difficult and a highly educated participant who used the system selectively because his daily life was already structured suggests that low use does not necessarily mean low usefulness. Instead, selective engagement may reflect different routes to perceived benefit. These findings complicate a simple digital divide narrative: age alone did not appear sufficient to explain engagement, and educational background, routine stability, and prior exposure to structured work or technology may have been equally important in shaping how participants used the intervention.

In addition, the continued engagement reported by several lower-income participants supports the practical value of delivering AI-HEALS through WeChat. By relying on an already familiar platform rather than a standalone app, the intervention may have reduced access barriers associated with downloading, registration, or cost. In this respect, the platform choice itself may have contributed to usability and equity in engagement.

Psychological Adaptation: The Less Visible Dimension of Digital Support

The psychological benefits described by participants—greater confidence, reassurance, and reduced uncertainty—highlight an important dimension of digital self-management support that is often less visible in studies focused primarily on clinical outcomes. These accounts are consistent with the literature emphasizing the emotional burden of diabetes and the value of supportive interventions for coping and self-management [34,35].

Importantly, the perceived psychological benefit of AI-HEALS appeared to arise indirectly through better understanding, clearer routines, and more manageable daily tasks rather than through explicit psychological counseling. Participants described feeling less overwhelmed and more capable when management tasks became more understandable and structured. This suggests that psychological adaptation may be closely intertwined with practical self-management support rather than standing apart from it. In this sense, the intervention may have reduced resistance to the burden of ongoing self-management by transforming diabetes from an overwhelming threat into a more manageable set of daily tasks [36].

This pattern can also be related to the Health Action Process Approach, which distinguishes motivational processes from the volitional support needed to sustain action [37-39]. In this study, educational content appeared to strengthen motivation by reinforcing awareness of risk and outcomes, while reminders and monitoring functions appeared to support the volitional work of action planning, routine maintenance, and coping. The coexistence of behavioral change and reduced uncertainty in participants' accounts is consistent with this dual-support process.

The findings also suggest that psychological adaptation varied according to illness experience and everyday context. For participants with more established self-management histories, confidence appeared to build on persistence that was reinforced by the intervention. For less experienced users, confidence seemed more closely linked to newly acquired understanding and clearer day-to-day guidance. Family and household context may also have shaped these experiences. For some participants, existing family support may have reinforced the intervention's effects, while for others—particularly those with less immediate day-to-day support—reminders and accessible guidance may have played a more compensatory role. These observations underline that the psychological value of digital support is shaped not only by technological features but also by the life contexts in which those features are used.

Practical Implications for Primary Care

The findings have several implications for implementation in routine primary care. First, AI-HEALS appears most suitable as an adjunct to community diabetes management rather than a replacement for clinicians. Community physicians or nurses could introduce the mini program during follow-up visits, confirm medication and monitoring schedules, and use subsequent visits to review questions or troubleshoot difficulties. Second, the predominance of lower-burden engagement suggests that pushed content and reminders may serve as the most feasible default components for routine implementation, while more interactive functions, such as the chatbot, may work better as optional layers for patients who wish to engage more actively. Third, the variation in participants' accounts highlights the need for layered and literacy-sensitive design. Brief onboarding, plain-language instructions, and family-assisted use may be especially helpful for some users, while provider training,

XSL•FO
RenderX

workflow integration, privacy procedures, and maintenance planning will be important for broader scale-up.

Interpretive Contributions

This study contributes qualitative evidence on an AI-enabled diabetes education intervention delivered through a familiar social platform in Chinese community settings. Its main contribution lies in clarifying how participants made use of digital support in practice: through the reinforcement of awareness, the translation of advice into everyday action, selective but meaningful engagement, and increased psychological reassurance. More broadly, the findings suggest that the perceived value of AI-enabled self-management support depends not only on algorithmic tailoring but also on whether the intervention is accessible, low burden, and well aligned with users' daily routines.

Limitations

Several limitations should be considered when interpreting these findings. First, the 3-month follow-up captures early perceptions and short-term behavioral experiences only; it does not allow claims about long-term sustainability, and statements in this paper about maintenance should therefore be interpreted as perceived early continuation rather than established long-term adherence. Second, although 17 interviews were sufficient for thematic saturation, the sample remained small and context-specific, which limits transferability beyond Beijing community settings. Third, selection bias is possible because interviewees were drawn from the intervention arm and agreed to discuss their experiences; participants who were more willing to engage with AI-HEALS may have been more likely to participate, potentially overrepresenting favorable views. Fourth, the commercial nature of the dialog platform limited our access to detailed training data, internal validation metrics, and model specifications, constraining technical reproducibility. Fifth, social desirability bias may have influenced participants' accounts, particularly because interviewers were linked to the broader study team. Finally, missing information on diabetes duration for some participants reduced our ability to examine duration-related patterns in depth. Taken together, these limitations indicate that the findings should be interpreted as explanatory qualitative evidence of perceived mechanisms and implementation considerations rather than definitive evidence of long-term effectiveness or generalizable user responses.

Conclusion

This explanatory qualitative study provided insight into how patients with type 2 diabetes experienced AI-HEALS, a WeChat-based AI-enabled health education system. Participants generally perceived the intervention as helpful for strengthening awareness, supporting everyday self-management, and improving confidence, although experiences differed according to digital literacy, prior routines, and stage of illness. The findings suggest that meaningful benefits may arise even when engagement is concentrated in lower-burden features such as reminders and pushed content. At the same time, the study highlights the importance of cautious interpretation, literacy-adaptive design, and integration with routine primary care. Future studies should test the long-term sustainability of these perceived benefits and further evaluate how algorithmic tailoring, human support, and implementation context interact in larger and more diverse populations.

Acknowledgments

The authors are grateful to the professionals and students who provided enormous support for recruitment and data collection in this study, as well as to the participants who took the time to participate. They confirm that no generative artificial intelligence tools were used at any stage in the preparation of this manuscript.

Funding

This work was supported by Capital's Funds for Health Improvement and Research (2026-2R-4403 and 2022-1G-4252). The funding body is the Beijing Municipal Government.

Data Availability

The research team and data analysts had access to the final dataset. The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Authors' Contributions

XS and YW conceived and designed the study. JW drafted the initial manuscript. XS, YW, and YJ provided theoretical and methodological guidance. XG and BK were responsible for data collection. JW, RB, and YJ analyzed and coded the interview data. XS was responsible for study management, and YW was responsible for study implementation. All authors revised the manuscript and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed demographic characteristics of interviewees.

XSL•FO
RenderX

[XLSX File, 12 KB - jmir_v28i1e84605_app1.xlsx]

References

1. IDF Diabetes Atlas. : International Diabetes Federation; 2021 URL: https://diabetes.org.br/wp-content/uploads/2022/02/IDF_Atlas_10th_Edition_2021-.pdf [accessed 2026-04-15]
2. Ma RCW, Chan JCN. Type 2 diabetes in East Asians: similarities and differences with populations in Europe and the United States. Ann N Y Acad Sci 2013 Apr;1281(1):64-91. [doi: 10.1111/nyas.12098] [Medline: 23551121]
3. American Diabetes Association. 2. Classification and diagnosis of diabetes: Standards of Medical Care in Diabetes-2020. Diabetes Care 2020 Jan;43(Suppl 1):S14-S31. [doi: 10.2337/dc20-S002] [Medline: 31862745]
4. Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. Nat Rev Endocrinol 2018 Feb;14(2):88-98. [doi: 10.1038/nrendo.2017.151] [Medline: 29219149]
5. Davies MJ, D'Alessio DA, Fradkin J, et al. Management of hyperglycemia in type 2 diabetes, 2018. A consensus report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Diabetes Care 2018 Dec;41(12):2669-2701. [doi: 10.2337/dci18-0033] [Medline: 30291106]
6. Chester B, Babu JR, Greene MW, Geetha T. The effects of popular diets on type 2 diabetes management. Diabetes Metab Res Rev 2019 Nov;35(8):e3188. [doi: 10.1002/dmrr.3188] [Medline: 31121637]
7. Ricci-Cabello I, Ruiz-Pérez I, Rojas-García A, Pastor G, Rodríguez-Barranco M, Gonçalves DC. Characteristics and effectiveness of diabetes self-management educational programs targeted to racial/ethnic minority groups: a systematic review, meta-analysis and meta-regression. BMC Endocr Disord 2014 Jul 19;14:60. [doi: 10.1186/1472-6823-14-60] [Medline: 25037577]
8. Hamine S, Gerth-Guyette E, Faulx D, Green BB, Ginsburg AS. Impact of mHealth chronic disease management on treatment adherence and patient outcomes: a systematic review. J Med Internet Res 2015 Feb 24;17(2):e52. [doi: 10.2196/jmir.3951] [Medline: 25803266]
9. Jiang X, Ming WK, You JH. The cost-effectiveness of digital health interventions on the management of cardiovascular diseases: systematic review. J Med Internet Res 2019 Jun 17;21(6):e13166. [doi: 10.2196/13166] [Medline: 31210136]
10. Wu X, Guo X, Zhang Z. The efficacy of mobile phone apps for lifestyle modification in diabetes: systematic review and meta-analysis. JMIR Mhealth Uhealth 2019 Jan 15;7(1):e12297. [doi: 10.2196/12297] [Medline: 30664494]
11. Bonoto BC, de Araújo VE, Godói IP, et al. Efficacy of mobile apps to support the care of patients with diabetes mellitus: a systematic review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2017 Mar 1;5(3):e4. [doi: 10.2196/mhealth.6309] [Medline: 28249834]
12. Quinn CC, Shardell MD, Terrin ML, Barr EA, Ballew SH, Gruber-Baldini AL. Cluster-randomized trial of a mobile phone personalized behavioral intervention for blood glucose control. Diabetes Care 2011 Sep;34(9):1934-1942. [doi: 10.2337/dc11-0366] [Medline: 21788632]
13. McCarthy J, Minsky ML, Rochester N, Shannon CE. A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. AI Mag 2006;27(4):12. [doi: 10.1609/aimag.v27i4.1904]
14. Jiang Y, Yang M, Wang S, Li X, Sun Y. Emerging role of deep learning-based artificial intelligence in tumor pathology. Cancer Commun 2020 Apr;40(4):154-166. [doi: 10.1002/cac2.12012]
15. Ribelles N, Jerez JM, Rodriguez-Brazzarola P, et al. Machine learning and natural language processing (NLP) approach to predict early progression to first-line treatment in real-world hormone receptor-positive (HR+)/HER2-negative advanced breast cancer patients. Eur J Cancer 2021 Feb;144:224-231. [doi: 10.1016/j.ejca.2020.11.030] [Medline: 33373867]
16. Vasiloglou MF, Mougiakakou S, Aubry E, et al. A comparative study on carbohydrate estimation: GoCARB vs. dietitians. Nutrients 2018 Jun 7;10(6):741. [doi: 10.3390/nu10060741] [Medline: 29880772]
17. Kodama S, Fujihara K, Shiozaki H, et al. Ability of current machine learning algorithms to predict and detect hypoglycemia in patients with diabetes mellitus: meta-analysis. JMIR Diabetes 2021 Jan 29;6(1):e22458. [doi: 10.2196/22458] [Medline: 33512324]
18. Lagani V, Chiarugi F, Manousos D, et al. Realization of a service for the long-term risk assessment of diabetes-related complications. J Diabetes Complications 2015 Jul;29(5):691-698. [doi: 10.1016/j.jdiacomp.2015.03.011] [Medline: 25953402]
19. Zeevi D, Korem T, Zmora N, et al. Personalized nutrition by prediction of glycemic responses. Cell 2015 Nov 19;163(5):1079-1094. [doi: 10.1016/j.cell.2015.11.001] [Medline: 26590418]
20. Alotaibi MM, Istepanian R, Philip N. A mobile diabetes management and educational system for type-2 diabetics in Saudi Arabia (SAED). Mhealth 2016;2:33. [doi: 10.21037/mhealth.2016.08.01] [Medline: 28293606]
21. Chen S, Lu J, Peng D, et al. Effect of a mobile health technology-based diabetes education program on glucose control in patients with type 2 diabetes initiating premixed insulin: a prospective, multicenter, observational study. Diabetes Care 2023 Jan 1;46(1):e6-e7. [doi: 10.2337/dc22-0510] [Medline: 36399717]
22. Wang HF, Li YH, Zhang Q, et al. The effect of artificial intelligence health education accurately linking system on childhood asthma: study protocol for a pilot randomized controlled trial. Front Psychiatry 2025;16:1585702. [doi: 10.3389/fpsyt.2025.1585702] [Medline: 40547124]

XSL•FO
RenderX

23. Jiang Y, Liu F, Guo J, et al. Evaluating an intervention program using WeChat for patients with chronic obstructive pulmonary disease: randomized controlled trial. J Med Internet Res 2020 Apr 21;22(4):e17089. [doi: 10.2196/17089] [Medline: 32314971]
24. Wu Y, Min H, Li M, et al. Effect of artificial intelligence-based health education accurately linking system (AI-HEALS) for type 2 diabetes self-management: protocol for a mixed-methods study. BMC Public Health 2023 Jul 11;23(1):1325. [doi: 10.1186/s12889-023-16066-z]
25. BotFactory dialog AI platform. ZhuJian Intelligence Technology. URL: https://www.emotibot.com/product/botfactory.html [accessed 2025-05-15]
26. Sun X, Jiang Y, Wang J, et al. Effects of a mobile health intervention based on a multitheoretical model of health behavior change on anxiety and depression, fear of cancer progression, and quality of life in patients with differentiated thyroid cancer: a randomized controlled trial. BMC Med 2024 Oct 15;22(1):466. [doi: 10.1186/s12916-024-03652-0] [Medline: 39407174]
27. Powers MA, Bardsley J, Cypress M, et al. Diabetes self-management education and support in type 2 diabetes: a joint position statement of the American Diabetes Association, the American Association of Diabetes Educators, and the Academy of Nutrition and Dietetics. Diabetes Care 2015 Jul;38(7):1372-1382. [doi: 10.2337/dc15-0730] [Medline: 26048904]
28. Greenwood DA, Gee PM, Fatkin KJ, Peeples M. A systematic review of reviews evaluating technology-enabled diabetes self-management education and support. J Diabetes Sci Technol 2017 Sep;11(5):1015-1027. [doi: 10.1177/1932296817713506] [Medline: 28560898]
29. Boren SA, Gunlock TL, Schaefer J, Albright A. Reducing risks in diabetes self-management: a systematic review of the literature. Diabetes Educ 2007;33(6):1053-1077. [doi: 10.1177/0145721707309809] [Medline: 18057274]
30. Fisher EB, Boothroyd RI, Coufal MM, et al. Peer support for self-management of diabetes improved outcomes in international settings. Health Aff (Millwood) 2012 Jan;31(1):130-139. [doi: 10.1377/hlthaff.2011.0914] [Medline: 22232103]
31. Anderson RM, Funnell MM. Patient empowerment: reflections on the challenge of fostering the adoption of a new paradigm. Patient Educ Couns 2005 May;57(2):153-157. [doi: 10.1016/j.pec.2004.05.008] [Medline: 15911187]
32. El-Gayar O, Timsina P, Nawar N, Eid W. Mobile applications for diabetes self-management: status and potential. J Diabetes Sci Technol 2013 Jan 1;7(1):247-262. [doi: 10.1177/193229681300700130] [Medline: 23439183]
33. Sardi L, Idri A, Fernández-Alemán JL. A systematic review of gamification in e-Health. J Biomed Inform 2017 Jul;71:31-48. [doi: 10.1016/j.jbi.2017.05.011] [Medline: 28536062]
34. Rubin RR, Peyrot M. Psychological issues and treatments for people with diabetes. J Clin Psychol 2001 Apr;57(4):457-478. [doi: 10.1002/jclp.1041] [Medline: 11255202]
35. Winkley K, Upsher R, Stahl D, et al. Psychological interventions to improve glycemic control in adults with type 2 diabetes: a systematic review and meta-analysis. BMJ Open Diabetes Res Care 2020 Apr;8(1):e001150. [doi: 10.1136/bmjdrc-2019-001150] [Medline: 32273289]
36. Brod M, Kongsø JH, Lessard S, Christensen TL. Psychological insulin resistance: patient beliefs and implications for diabetes management. Qual Life Res 2009 Feb;18(1):23-32. [doi: 10.1007/s11136-008-9419-1] [Medline: 19039679]
37. Zhang CQ, Zhang R, Schwarzer R, Hagger MS. A meta-analysis of the health action process approach. Health Psychol 2019 Jul;38(7):623-637. [doi: 10.1037/hea0000728] [Medline: 30973747]
38. Ranjbaran S, Shojaeizadeh D, Dehdari T, Yaseri M, Shakibazadeh E. The effectiveness of an intervention designed based on health action process approach on diet and medication adherence among patients with type 2 diabetes: a randomized controlled trial. Diabetol Metab Syndr 2022 Jan 4;14(1):3. [doi: 10.1186/s13098-021-00773-x] [Medline: 34983628]
39. Moghimi S, Payandeh A, Ranjbaran S, Seraji M. Effects of educational interventions on dietary adherence among type 2 diabetics in Zahedan: using the Health Action Process Approach. Ethiop J Health Sci 2023 Jul;33(4):571-580. [doi: 10.4314/ejhs.v33i4.3] [Medline: 38784208]

Abbreviations

AI: artificial intelligence
AI-HEALS: Artificial Intelligence–Based Health Education Accurately Linking System
HbA1c: hemoglobin A1c
mHealth: mobile health
MTM: Multi-Theory Model
NLP: natural language processing
T2DM: type 2 diabetes mellitus

XSL•FO
RenderX

Edited by A Stone; submitted 22.Sep.2025; peer-reviewed by A Shivanna, CY Su, M Chakit, S Ajayi; revised version received 31.Mar.2026; accepted 31.Mar.2026; published 12.May.2026.

Please cite as:
Wang J, Wu Y, Jiang Y, Bao R, Gu X, Kong B, Sun X
Experiences and Perceived Influence of the Artificial Intelligence–Based Health Education Accurately Linking System (AI-HEALS) on Health Behaviors Among Patients With Type 2 Diabetes: Qualitative Study
J Med Internet Res 2026;28:e84605
URL: https://www.jmir.org/2026/1/e84605
doi:10.2196/84605

© Jing Wang, Yibo Wu, Yang Jiang, Rantong Bao, Xinbao Gu, Bingyang Kong, Xinying Sun. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Preferences and Willingness to Pay for Smart Bracelets Among Chinese Pregnant and Postpartum Women: Discrete Choice Experiment

Jing Wang[1,2*], BM; Mingyue Cheng[3*], BM; Jialiang Feng[4], BM; Wenzhuo Li[5], BM; Xiaotao Zhou[6], BMgt; Yufang Cao[7], BM; Haiyan Feng[7], BM; Fan Yang[7], BM; Xiaoyun Wang[8*], BM; Yibo Wu[1*], PhD

[1]Department of Nursing, the Fourth Affiliated Hospital of School of Medicine, and International School of Medicine, International Institutes of Medicine, Zhejiang University, N1 Shangcheng Avenue, Yiwu, Zhejiang Province, China

[2]School of Public Health, Peking University, Beijing, China

[3]Department of Nursing, Inner Mongolia Maternity and Child Health Care Hospital, Hohhot, China

[4]Zhongnan Hospital, Wuhan University, Wuhan, Hubei, China

[5]Medical College, Qingdao University, Qingdao, Shandong, China

[6]Sinopec Zhongyuan Petroleum Engineering Co. Ltd. Tarim Branch Company,Bayingolin Mongolian Autonomous Prefecture, Xinjiang, China

[7]Department of Obstetrics, Inner Mongolia Maternity and Child Health Care Hospital, Hohhot, China

[8]Party Committee Office, Inner Mongolia Maternity and Child Health Care Hospital, Hohhot, China

[*]these authors contributed equally

Corresponding Author:
Yibo Wu, PhD
Department of Nursing, the Fourth Affiliated Hospital of School of Medicine, and International School of Medicine, International Institutes of Medicine, Zhejiang University, N1 Shangcheng Avenue, Yiwu, Zhejiang Province, China

Abstract

Background: Pregnant and postpartum women encounter various health challenges, including physiological stress and mental health issues, which necessitate ongoing health monitoring. Smart bracelets present a promising solution; however, there is limited research on the preferences and willingness to pay (WTP) for such devices among this demographic.

Objective: This study aimed to investigate the preferences and WTP for smart bracelet attributes among pregnant and postpartum women in China and to explore how these preferences vary by sociodemographic factors, pregnancy stage, parity, and complications.

Methods: A cross-sectional discrete choice experiment (DCE) was conducted involving 464 pregnant and postpartum women recruited from a maternal and child health hospital in Inner Mongolia. Six key attributes were evaluated: cost, hospital backend monitoring, primary function, privacy protection, ease of use, and monitoring report frequency. A mixed logit model was used to estimate preference weights and WTP for each attribute, with subgroup analyses based on income, employment, gestational stage, parity, and other factors.

Results: Among the 464 pregnant and postpartum women included in the final analysis (valid data rate: 96.67%), the mean age was 31.06 (SD 4.05) years. The majority of participants were of Han ethnicity (n=385, 82.97%), had a high level of education (n=422, 90.95%), resided in urban areas (n=446, 96.12%), and were employed (n=353, 76.08%). In the DCE, cost negatively impacted smart wristband preferences (β=−0.000257; *P*=.01). Participants exhibited a strong preference for wristbands with fetal heart monitoring (β=1.275; *P*<.001), high-level privacy protection (β=.541; *P*<.001), and ease of use (β=.973; *P*<.001). They were willing to pay ¥4967.45 (based on an exchange rate of US $1=CN ¥6.93) for fetal heart monitoring, ¥2975.17 for sleep monitoring, ¥2109.29 for high-level privacy protection, and ¥3437.09 for daily monitoring. Subgroup analyses indicated that preferences varied according to income, employment, pregnancy stage, parity, complications, and age.

Conclusions: The design of smart bracelets should be tailored to meet the diverse needs of pregnant and postpartum users. Key considerations include the integration of fetal heart and vital sign monitoring, the assurance of data privacy, the enhancement of usability, and the provision of cost-effective options. Understanding the specific preferences of different subgroups can guide the development of inclusive and responsive wearable health technologies for maternal care.

(J Med Internet Res 2026;28:e55941) doi:10.2196/55941

KEYWORDS

pregnant and postpartum women; smart bracelets; discrete choice experiment; willingness to pay; women's health

XSL•FO
RenderX

Introduction

The global proportion of pregnant women who are older, multiparous, or have underlying health conditions is increasing, leading to heightened risks during childbirth and a higher incidence of birth defects in their offspring [1]. According to the World Health Organization (WHO), approximately 280,000 women die each year due to complications related to pregnancy and childbirth, with 90% of these maternal deaths occurring in low- and middle-income countries [2]. Furthermore, pregnant women encounter mental health challenges, such as anxiety and depression [3], both during pregnancy and postpartum, which necessitate timely attention and intervention. With the evolving population structure in China and growing awareness of health management, pregnancy-related health monitoring has emerged as a significant public health issue [4]. Pregnant and postpartum women face various challenges, including physiological changes, increased metabolic demands, and escalating mental health risks, which require precise and continuous health data support [5,6]. While traditional pregnancy management methods are important, they have limitations regarding real-time data, personalization, and continuity [7-9].

With the rapid development of mobile internet and wearable technologies, smart wristbands have emerged as promising tools for convenient and real-time health monitoring in maternal health management. Recent studies indicate that wearable sensors offer a noninvasive approach to monitor pregnancy-related physiological and behavioral changes, facilitating continuous tracking of maternal heart rate, activity, and sleep patterns. These parameters are closely associated with pregnancy-related hormonal fluctuations, thus providing potential tools for the early detection of adverse outcomes [10]. Furthermore, the integration of wearable sensors with artificial intelligence (AI)–assisted data processing techniques is increasingly being explored for pregnancy monitoring. This integration captures multimodal signals, including electrocardiograms, uterine contractions, and fetal movements, thereby enhancing early warning systems and personalized management throughout gestation [11]. A diverse range of wearable technologies, from bracelets and smartwatches to innovative remote home monitoring systems, are being assessed for continuous remote obstetric care. While these technologies demonstrate feasibility, they also underscore the necessity of addressing challenges such as user compliance and data reliability [12].

Compared to traditional methods, smart wristbands offer significant advantages in maternal health management. They facilitate continuous monitoring of physiological data, such as heart rate, activity levels, and sleep quality [11], and can even perform fetal heart monitoring [13]. This capability provides personalized health management solutions for pregnant women, minimizing the need for frequent hospital visits, thereby saving both time and effort [14]. In terms of real-time monitoring, the data collected by smart wristbands can be instantly synchronized to smartphones or the cloud, enabling pregnant women and their health care providers to monitor health status in real time and promptly identify potential issues. For instance, a study by Grym et al [15] assessed the feasibility of continuous health parameter monitoring via smart wristbands during pregnancy and 1 month postpartum. The findings indicated that smart wristbands effectively collected data on activity, sleep, and heart rate, thus offering personalized health management solutions for pregnant women. However, the study also highlighted that comfort and data synchronization issues could hinder sustained use. Furthermore, Alim and Imtiaz [13] conducted a systematic review analyzing the use of various wearable sensors in maternal health monitoring, concluding that these devices could effectively track fetal electrocardiograms, fetal movement, and maternal physical activity, contributing to the reduction of pregnancy-related risks.

Despite the growing market for smart wristbands, research on the specific preferences and willingness to pay (WTP) for features, design, and price points among pregnant women remains limited. Understanding the actual needs of this population, as well as their acceptable price levels, is crucial for the functional design, market promotion, and optimization of health management services related to smart wristbands. Existing studies have primarily focused on technical feasibility, device accuracy, and user acceptance while offering limited in-depth analysis of pregnant women's preferences for specific features and their WTP. For instance, Wu et al [16], using the Technology Acceptance Model, evaluated pregnant women's willingness to use smart fetal heart monitoring devices, revealing that perceived usefulness and ease of use significantly influenced their willingness to adopt such technology. However, this study did not explore specific feature preferences or price sensitivity. Additionally, Schramm et al [17] examined the acceptance of noninvasive electronic devices for pregnancy monitoring among 507 pregnant women, concluding that women were generally open to using these devices for monitoring purposes. Nevertheless, these studies mainly concentrated on preliminary assessments of technology acceptance and WTP, lacking detailed analyses of pregnant women's preferences for specific features (eg, fetal heart monitoring and sleep quality assessment), hospital backend monitoring, and report frequency. Beyond device features, factors such as price, ease of operation, and privacy protection also significantly influence pregnant women's willingness to use smart wristbands [16,18]. Elevated prices may deter potential purchasers, while complex interfaces can detrimentally affect user experience. Therefore, the design of smart wristbands should thoroughly address these needs and offer more personalized and comprehensive services.

The functional needs of pregnant women for smart wristbands vary significantly across different stages of pregnancy. In the early stages, women are primarily concerned with features that alleviate morning sickness and enhance sleep quality [19]. As pregnancy progresses into the midstage, with fetal development stabilizing, there is an increased demand for activity monitoring and weight management [20]. In the later stages, as childbirth approaches, women focus more on fetal heart monitoring, uterine contraction tracking, and other functionalities that provide insights into fetal health [21]. After childbirth, new mothers particularly seek features for sleep monitoring, guidance on physical recovery, and parenting resources to help them adapt to their new lifestyle [22-24]. Consequently, the design of smart wristbands must comprehensively address the distinct needs at

XSL•FO
RenderX

each stage of pregnancy, offering targeted health management solutions.

Discrete choice experiments (DCEs) are a crucial preference measurement tool extensively used in the health care sector, particularly in maternal and child health management. Through DCEs, researchers can gain profound insights into the preferences of patients and the public regarding various medical interventions, service formats, and policy designs, thereby providing a solid foundation for decision-making. For instance, in drug selection, DCEs can elucidate how patients prioritize key attributes such as drug efficacy, side effects, mode of administration, and cost, thereby offering valuable guidance for drug development and market promotion. Arije et al [25] discovered that adolescents and young individuals in Nigeria exhibited diverse preferences for sexual and reproductive health services, emphasizing factors such as the medical environment, physician professionalism, service attitudes, medical expenses, and waiting times. Furthermore, DCE plays a pivotal role in health policy evaluation. Luo et al [26] used DCE to evaluate the preferences of public health master's graduates regarding job positions, revealing that monthly salary was the most significant determinant influencing their employment choices, followed by opportunities for career advancement and job location. Donnan et al [27] found that consumers expressed heightened concerns regarding the regulatory status of e-cigarettes and demonstrated a WTP premium prices for products regulated by Canada's Ministry of Health. Additionally, Jones et al [28] used DCE to generate health utility values for patients with chronic obstructive pulmonary disease and explored its application in Health Technology Assessment. These studies underscore the extensive potential of DCE in understanding patient and public needs and in formulating effective health policies and interventions.

This study aimed to use DCEs to systematically evaluate the preferences of pregnant women regarding the attributes of smart wristbands and to estimate the additional amount they are willing to pay for these attributes. The findings will offer crucial insights for decision-making among smart wristband manufacturers, health service providers, and policymakers, thereby promoting the development of smart wristband products and services that are better aligned with the needs of pregnant women. This alignment is expected to improve maternal health management and ultimately enhance maternal and child health outcomes.

Methods

Study Design

This study aimed to explore the preferences and WTP for smart bracelets among pregnant and postpartum women using a DCE. The participants consisted of women who were pregnant or had given birth within the past 2 years, with a sample randomly selected from the Inner Mongolia Maternal and Child Health Hospital. Demographic diversity was intentionally incorporated into the sample to ensure that the research findings are broadly applicable to various subgroups of pregnant and postpartum women in the region. Using the DCE method, the study systematically assessed the preferences of pregnant and postpartum women for smart bracelets, focusing on the device's functionality, price, and other design features, while examining their WTP for different combinations of product features. This approach provided valuable insights into their preference patterns and market demand.

This was a cross-sectional study designed to collect data in a single survey, assessing the preferences and WTP for smart bracelets among pregnant and postpartum women. To ensure the accuracy and feasibility of the study, a comprehensive set of inclusion and exclusion criteria was established.

Inclusion Criteria

The inclusion criteria were as follows: (1) age ≥18 years, (2) female, (3) Chinese nationality (People's Republic of China), (4) pregnant or had given birth within the past 2 years, (5) able to complete the online survey independently or with the assistance of the investigator, (6) having basic literacy skills and being able to communicate and interact normally, and (7) willing to voluntarily participate in the study and provide informed consent.

Exclusion Criteria

The exclusion criteria were as follows: (1) individuals who were unconscious or had mental abnormalities, (2) individuals with cognitive impairments, (3) individuals currently participating in other similar research studies, (4) individuals unwilling to cooperate, and (5) individuals with other severe health conditions that may affect the study results.

Through this approach, the study provided valuable market and preference data for the use of smart bracelets among pregnant and postpartum women, aiding in the development and promotion of related products.

Ethical Considerations

The study received approval from the Ethics Committee of Inner Mongolia Maternal and Child Health Hospital (review [176], 2024). All participants voluntarily engaged in the study and provided informed consent, confirming their understanding of the study's objectives, methods, and potential risks. This consent was obtained prior to their participation, adhering to ethical standards for research involving human subjects. Given the involvement of human participants, appropriate ethical oversight was ensured, with all research activities sanctioned by the Ethics Committee. To safeguard privacy and confidentiality, all study data were anonymized and deidentified, thereby protecting participants' identities. The data were securely stored and accessible only to authorized personnel involved in the research. No compensation was offered to participants for their involvement in this study. To further uphold the confidentiality and anonymity of participants, no identifiable images or data of individual participants were included in the manuscript or supplementary materials.

Development of DCE

In this study, we conducted a systematic review to comprehensively identify and assess the key attributes and their levels that pregnant and postpartum women prioritize when selecting a smart wristband. We searched the literature across several key databases, including PubMed, Cochrane Library, Web of Science, CNKI (China National Knowledge

XSL•FO
RenderX

Infrastructure), and Scopus, to ensure thorough coverage of relevant research on smart wristbands and wearable devices in maternal health management. A combination of Medical Subject Headings (MeSH) and free-text keywords was used to filter the relevant literature. The search terms included “wearable devices,” “smart wristband,” “fitness tracker,” “wearable sensors,” “maternal health,” “pregnancy,” “pregnant women,” “postpartum,” and “remote monitoring,” along with various combinations (eg, “wearable devices AND pregnancy,” “smart wristband AND maternal health”). These terms were tailored to each database’s search syntax to ensure both sensitivity and specificity in retrieval.

The literature was screened based on preestablished inclusion criteria: the research subjects had to be either pregnant or postpartum women or related health research; the studies needed to involve smart wristbands or other wearable devices; and relevant attributes such as price, privacy protection, or functionality had to be mentioned in the literature. Initially, all retrieved records were screened by title and abstract, followed by a full-text review of potentially eligible papers to confirm their eligibility.

This systematic review encompassed 7 highly relevant studies [13,29-34]. Through analysis of these studies, the research team identified 6 key attributes that pregnant and postpartum women prioritize when selecting a smart wristband: price, hospital background monitoring management, main function, privacy protection, ease of use, and frequency of monitoring report delivery. The levels associated with each of these attributes were also extracted from the literature, establishing a foundation for subsequent expert consultation.

Following the systematic review, expert consultations were conducted to further validate and refine the identified attributes and levels. These consultations involved obstetricians with clinical experience in prenatal and postpartum care (3 experts), health management professionals knowledgeable about maternal health service delivery (2 experts), wearable technology specialists with expertise in product design and human factors (2 experts), and market research experts experienced in consumer preference studies (2 experts). In total, 9 experts participated in 2 rounds of consultation.

During the first round, experts reviewed the preliminary list of attributes and levels. They were asked to comment on the relevance of each attribute, suggest modifications to levels (eg, cost ranges and descriptions of functionality), and propose any missing attributes. Specific expert feedback included the following:

- Cost: The technology and market experts suggested that cost levels should represent realistic price points in the current Chinese wearable market for consumer smart wristbands targeted at health care use (eg, entry level, midrange, and high end). As a result, the cost levels were adjusted to ¥250, ¥500, ¥750, and ¥1000 after consensus.
- Hospital background monitoring management: Experts highlighted that integration with hospital management systems may be crucial for clinical acceptance by pregnant women but that such integration is not universally available. On the basis of this feedback, the level description was refined to clearly distinguish between “yes (supported by hospital monitoring)” and “no (standalone device).”
- Main function: Obstetrics and technology experts emphasized including specific vital sign monitoring relevant to pregnancy (eg, blood pressure and heart rate) and fetal heart monitoring—features frequently mentioned in the literature as highly valued by users. Therefore, levels were expanded to include “activity tracking,” “vital sign monitoring,” “fetal heart monitoring,” and “sleep quality.”
- Privacy protection: Health management and privacy experts recommended distinguishing among degrees of data protection (high, standard, and none) to reflect different security protocols and possible user concerns highlighted in digital health studies.
- Ease of use: Experts underscored the importance of usability, especially for pregnant women with varying levels of technology literacy. The levels were refined to a 4-tier scale, ranging from “very easy to use” to “very difficult to use.”
- Frequency of monitoring report delivery: Feedback from both market and clinical experts suggested that delivery frequency might influence perceived utility; levels were defined to cover daily to weekly reporting patterns relevant to routine maternal health monitoring.

After incorporating these suggestions, a second consultation round confirmed that the refined attributes and levels were appropriate, meaningful, and aligned with the decision context and target population’s concerns. Experts agreed that the defined levels reflected realistic choices that pregnant and postpartum women would encounter when choosing a smart wristband.

Through the systematic review and expert consultation, the 6 key attributes and their levels, along with the definitions of each attribute, that pregnant and postpartum women focus on when selecting a smart wristband were ultimately determined (Table 1).

XSL•FO
RenderX

Table . Attributes and levels of the smart bracelet.

Attribute, definition, and levels	Level explanation
Cost (CN ¥)[a]: refers to the amount of money that a user needs to pay to purchase the smart wristband.	
250	Entry-level price point representing basic device functionality.
500	Midlower price point with a balance of features and affordability.
750	Midhigher price point offering more advanced features.
1000	High-end price point with comprehensive functions and potential premium services.
Hospital background monitoring management: indicates whether the smart wristband supports integration with a hospital's or health care provider's monitoring and management system.	
Yes	Device can securely transmit user health data to a linked clinical or hospital system for professional review, enabling clinician follow-up and intervention.
No	Device operates independently without connection to a hospital or clinical system; users cannot share data directly with clinicians.
Main function: describes the core monitoring or tracking capabilities of the smart wristband.	
Activity tracking	Tracks basic physical activity such as step count and movement patterns.
Vital sign monitoring	Measures critical health indicators such as heart rate and blood pressure.
Fetal heart monitoring	Continuously records fetal heart rate patterns relevant to pregnancy monitoring.
Sleep quality	Monitors sleep stages and sleep quality metrics such as duration and restfulness.
Privacy protection: refers to the level of data security and privacy safeguards implemented by the device during data collection, storage, and transmission.	
High-level protection	Device uses strong privacy measures such as end-to-end encryption, strict access control, no third-party data sharing, and adherence to recognized privacy standards to protect highly sensitive health data.
Standard protection	Device applies basic encryption and authentication methods to protect data, with typical privacy policy terms and moderate risk mitigation.
No special protection	Device lacks dedicated privacy or security mechanisms; data may be stored or transmitted without encryption or user-controlled safeguards.
Ease of use: describes how simple or complex the device is for users to operate and interact with.	
Very easy to use	Interface and workflows are intuitive, requiring minimal steps and no user training.
Relatively easy to use	Most tasks are straightforward, with occasional need for guidance.
Relatively difficult to use	Multiple steps or unclear interface elements may cause confusion for some users.
Very difficult to use	Complex interactions and navigation require frequent assistance or training.
Frequency of monitoring report delivery to the smart bracelet: indicates how often the user receives summary reports or actionable insights based on the device's monitoring data.	
Once a day	Daily reporting frequency.
Once every 2 days	Reporting every second day.
Once a week or less often	Weekly or longer intervals between reports.

[a]A currency exchange rate of US $1=CN ¥6.93 was applicable.

On the basis of the attributes and levels determined earlier, our team used a fractional factorial design method to determine the optimal number of attributes that pregnant women prefer when selecting a smart bracelet. Given that the 6 attributes in this study each contain 2 to 4 levels, a full factorial design would result in 1152 combinations (4×2×4×3×4×3=1152), which is clearly impractical to generate. Therefore, we used a fractional factorial design to determine the optimal number of choice sets. This method is based on 2 principles: the orthogonality principle and the balance principle [35,36]. The choice sets are generated using orthogonal experiments in SPSS AU software (Beijing Qingsi Technology Co). An orthogonal strategy is selected based on the orthogonal table manual, and ineffective designs are eliminated through internal validity checks, leading to the

XSL•FO
RenderX

final design. Suitable options were selected as controls to form multiple-choice sets. Each participant faced 12 choice sets, with each choice set containing 3 options: smart bracelet A, smart bracelet B, and "none." To assess the validity of the DCE questionnaire, a choice set with a clear dominant option (a lie detector question) was included in the questionnaire. Only when participants select the correct option from this set will their data be included in the subsequent analysis. Meanwhile, according to the empirical rule proposed by Johnson and Orme [37,38], the minimum sample size for a DCE study can be calculated using the formula:

$$N>500\times c / (t\times a)$$

where N is the minimum sample size, representing the number of participants required in the study; c is the largest number of levels for any attribute (c=4), t is the number of choice tasks (t=12), and a is the number of alternatives (a=2).

The final calculation resulted in a minimum sample size of 84.

Table 2 is an example of a DCE choice set.

Table . Example of a discrete choice experiment choice set.

Attribute	Smart bracelet A	Smart bracelet B	None of them
Cost (¥)[a]	1000	500	Choose none of them
Hospital background monitoring management	No	Yes	—[b]
Main function	Activity tracking	Fetal heart monitoring	—
Privacy protection	High-level protection	High-level protection	—
Ease of use	Relatively difficult to use	Very easy to use	—
Frequency of monitoring report delivery to the smart bracelet	Once a day	Once every 2 days	—

[a]A currency exchange rate of US $1=CN ¥6.93 was applicable.

[b]Not applicable.

Data Collection

Data were collected using an online survey platform, such as WenjuanXing. The study commenced in September 2024 and concluded in May 2025, lasting approximately 9 months. Recruitment took place from February 2025 to April 2025. The data collection method involved researchers identifying eligible participants offline and requesting them to scan a WenjuanXing link to complete the online survey. During the recruitment period, researchers provided eligible pregnant and postpartum women with comprehensive information about the study and invited them to participate by completing the online questionnaire.

The survey was divided into 2 distinct parts. The first part gathered essential demographic information from participants, including age, education level, and place of residence, to gain a comprehensive understanding of their backgrounds. The second part focused on the preferences of pregnant and postpartum women concerning smart bracelets, using the DCE method for data collection. The DCE questionnaire featured multiple-choice scenarios, each presenting 2 different combinations of smart bracelet attributes, from which participants were asked to select the combination that best aligned with their preferences. The attributes of the smart bracelets encompassed device functions (such as heart rate monitoring, activity tracking, and sleep tracking), price, ease of use, and privacy protection, all chosen based on prevalent features available in the current market. To mitigate sequence effects, all choice scenarios were randomly arranged.

To ensure data quality and protect participants' privacy, the survey was conducted anonymously, using encryption technology to safeguard the data. All participants voluntarily completed the questionnaire after receiving an informed consent form, which ensured their understanding of the study's purpose and their involvement [39]. Researchers provided necessary assistance to facilitate the completion of the survey. During data collection, logical checks and additional measures were implemented to ensure the validity and consistency of the data.

Statistics

Descriptive statistics were used to summarize the sociodemographic characteristics of the participants, including age, income, education, occupation, pregnancy stage, parity, and the presence of pregnancy complications. Means and SDs were computed for continuous variables, while frequencies and percentages were reported for categorical variables. To analyze participants' preferences regarding smart bracelet attributes, a mixed logit model was used. This model accommodated preference heterogeneity among individuals by allowing the attribute coefficients to vary randomly. Each participant's utility for a specific smart bracelet profile was modeled as a linear combination of attribute levels. The estimated parameters included the fixed effects (β) of each attribute level, their SEs, and corresponding P values. The relative importance scores quantify the significance of each attribute, calculated by dividing the difference in utility between the lowest and highest levels of that attribute by the sum of the differences for all attributes. Cost was treated as a continuous variable to facilitate the estimation of participants' WTP for different attributes. WTP values were derived by dividing the attribute coefficient by the negative of the cost coefficient, reflecting the monetary value participants assigned to specific smart bracelet features. Subgroup analyses were conducted to investigate differences

XSL•FO
RenderX

in preferences and WTP across various demographic and clinical subgroups, including income levels, employment status, pregnancy stages, parity, pregnancy complications, and age. All statistical analyses were executed using Stata 15.0 (StataCorp LLC). A 2-sided *P* value of <.05 was deemed statistically significant.

Results

Characteristics of Respondents

A total of 480 pregnant women were initially recruited for the study; after excluding those with incomplete data or failing logical consistency checks, 464 participants were included in the final analysis, yielding a high data validity rate of 96.67%. Among the 464 participants, the mean age was 31.06 (SD 4.05) years (Table 3). The majority were of Han ethnicity (385, 82.97%) compared to ethnic minorities (n=79, 17.03%). For education level, 90.95% (n=422) had a high level of education. Most participants resided in urban areas (n=446, 96.12%). Income distribution was as follows: ≤4000 (n=102, 21.98%), 4001 to 6000 (n=170, 36.64%), and ≥6001 (n=192, 41.38%). Occupational status included employed (n=353, 76.08%), unemployed or out of work (n=60, 12.93%), and without fixed occupation (freelance, n=51, 10.99%). Pregnancy status was distributed as early pregnancy (1 - 13 wk; n=122, 26.29%), midpregnancy (14 - 27 wk; n=107, 23.06%), late pregnancy (≥28 wk; n=110, 23.71%), and postpartum (n=125, 26.94%). First-time pregnancy was reported by 71.98% (n=334). Regarding pregnancy complications, 84.05% (n=390) had none.

Table . Demographic characteristics of maternal (N=464).

Variables	Values
Age (y), mean (SD)	31.06 (4.05)
Ethnicity, n (%)	
Han ethnicity	385 (82.97)
Ethnic minorities	79 (17.03)
Education level, n (%)	
Primary and middle education	42 (9.05)
High education	422 (90.95)
Place of residence, n (%)	
Urban	446 (96.12)
Rural	18 (3.88)
Income, n (%)	
≤4000	102 (21.98)
4001 - 6000	170 (36.64)
≥6001	192 (41.38)
Occupation, n (%)	
Employed	353 (76.08)
No fixed occupation (freelancer)	51 (10.99)
Unemployed or out of work	60 (12.93)
Pregnancy period, n (%)	
Early pregnancy	122 (26.29)
Midpregnancy	107 (23.06)
Late pregnancy	110 (23.71)
Postpartum	125 (26.94)
First childbirth, n (%)	
Yes	334 (71.98)
No	130 (28.02)
Pregnancy complications, n (%)	
Yes	74 (15.95)
No	390 (84.05)

XSL•FO
RenderX

DCE Results

This study analyzed the preferences and WTP of pregnant and postpartum women for smart wristbands through a DCE (Table 4). The results showed that cost had a negative impact on choice (β=−0.000257; *P*=.01). For the hospital background monitoring management attribute, the most preferred level was having hospital background monitoring management (β=.226; *P*=.03); in terms of primary functions, the most preferred function was fetal heart monitoring (β=1.275; *P*<.001); for the privacy protection attribute, the most preferred level was high-level protection (β=.541; *P*<.001); in terms of ease of use, the most preferred level was relatively easy to use (β=.973; *P*<.001); and for the monitoring report frequency, the most preferred level was once a day (β=.882; *P*<.001).

Table . Mixed logit model estimates of smart bracelet preferences among pregnant and postpartum women[a].

Attributes and levels	β (SE)	95% CI	*P* value
Cost	−0.000257 (0.000099)	−0.000451 to −0.000062	.01
Hospital background monitoring management			
No	Ref[b]	—[c]	—
Yes	0.226 (0.105)	0.020 to 0.431	.03
Main function			
Activity tracking	Ref	—	—
Sleep quality	0.763 (0.074)	0.619 to 0.908	<.001
Vital sign monitoring	1.233 (0.114)	1.010 to 1.457	<.001
Fetal heart monitoring	1.275 (0.087)	1.104 to 1.446	<.001
Privacy protection			
No special protection	Ref	—	—
Standard protection	0.218 (0.083)	0.054 to 0.381	.009
High-level protection	0.541 (0.104)	0.338 to 0.745	<.001
Ease of use			
Very difficult to use	Ref	—	—
Relatively difficult to use	0.662 (0.117)	0.433 to 0.892	<.001
Very easy to use	0.719 (0.132)	0.460 to 0.978	<.001
Relatively easy to use	0.973 (0.094)	0.788 to 1.158	<.001
Frequency of monitoring report delivery			
Once a week or less often	Ref	—	—
Once every 2 days	0.571 (0.132)	0.312 to 0.830	<.001
Once a day	0.882 (0.109)	0.669 to 1.095	<.001

[a]Observations=16,704; Log Likelihood (null)=−4486.437; Log Likelihood (model)=−4386.441; df=23; Akaike Information Criterion=8818.882; Bayesian Information Criterion=8996.521.

[b]Ref: reference.

[c]Not applicable.

The Percent Importance of Smart Wristband Attributes

As presented in Table 5, the main function was the most important attribute.

XSL•FO
RenderX

Table . Importance of attributes.

Attribute	Importance (%)
Cost	0.01
Hospital background monitoring management	5.80
Main function	32.72
Privacy protection	13.88
Ease of use	24.97
Frequency of monitoring report delivery to the smart bracelet	22.63

WTP Results

Table 6 presents the estimated WTP for changes in specific attributes of smart bracelets, with WTP values quantifying participants' implicit monetary trade-offs between improvements in these attributes. In DCE studies, WTP reflects the marginal additional amount a respondent is hypothetically willing to pay for an enhancement in a given attribute level relative to a reference level. This is derived from the ratio of attribute coefficients to the cost coefficient in the mixed logit model, effectively capturing the relative valuation of attribute changes. Compared to the reference level of activity tracking, participants were willing to pay an additional ¥2975.168 (95% CI 743.556-5206.781) for sleep quality, ¥4806.745 (95% CI 990.014, 8623.475) for vital sign monitoring, and ¥4967.454 (95% CI 1212.358, 8722.550) for fetal heart monitoring. For the privacy protection attribute, participants were willing to pay ¥847.659 (95% CI 162.234-1533.085) for standard protection and ¥2109.294 (95% CI 551.576-3667.013) for high-level protection, both compared to no special protection. Regarding ease of use, relative to the category of "very difficult to use," participants were willing to pay ¥2581.683 (95% CI 598.881-4564.485) for relatively difficult-to-use devices, ¥2802.522 (95% CI 539.744-5065.300) for devices that are very easy to use, and ¥3792.292 (95% CI 860.861-6723.723) for relatively easy-to-use devices. Finally, concerning monitoring frequency, compared to the reference level of "once a week or less often," participants were willing to pay ¥2225.933 (95% CI 148.084-4303.781) for using it once every 2 days and ¥3437.094 (95% CI 713.160-6161.028) for using the device once a day.

XSL•FO
RenderX

Table . Estimated willingness to pay (WTP) for each attribute level.

Attributes and levels	WTP (95% CI), ¥[a]
Hospital background monitoring management	
No	Ref[b]
Yes	878.803 (−84.138 to 1841.745)
Main function	
Activity tracking	Ref
Sleep quality	2975.168 (743.556 to 5206.781)
Vital sign monitoring	4806.745 (990.014 to 8623.475)
Fetal heart monitoring	4967.454 (1212.358 to 8722.550)
Privacy protection	
No special protection	Ref
Standard protection	847.659 (162.234 to 1533.085)
High-level protection	2109.294 (551.576 to 3667.013)
Ease of use	
Very difficult to use	Ref
Relatively difficult to use	2581.683 (598.881 to 4564.485)
Very easy to use	2802.522 (539.744 to 5065.300)
Relatively easy to use	3792.292 (860.861 to 6723.723)
Frequency of monitoring report delivery	
Once a week or less often	Ref
Once every 2 days	2225.933 (148.084 to 4303.781)
Once a day	3437.094 (713.160 to 6161.028)

[a]A currency exchange rate of US $1=CN ¥6.93 was applicable.

[b]Ref: reference.

Subgroup Results

We further conducted subgroup analyses based on demographic and clinical characteristics, including income, employment status, pregnancy stage, parity, presence of pregnancy complications, and age (Tables S1-S7 in Multimedia Appendix 1). The results showed that these characteristics significantly influenced pregnant and postpartum women's preferences for the attributes of smart bracelets. In terms of the price attribute, pregnant and postpartum women with low income (income ≤ ¥4000), those who were unemployed, those in early or late pregnancy, those with previous births, those without pregnancy complications, and those of younger age (age ≤ 30 y) exhibited a significantly stronger preference for lower-priced smart bracelets (P<.05); whereas women with moderate income (¥4001 - ¥6000), those without fixed employment (self-employed), primiparous women, those without pregnancy complications, and younger women showed a significantly stronger preference for smart bracelets with hospital background monitoring functions (P<.05). In terms of specific feature preferences, pregnant and postpartum women with low and high income, those who were employed, those in mid to late pregnancy, both primiparous and multiparous women, women with or without pregnancy complications, and older women (age >30 y) exhibited a significant preference for smart bracelets equipped with fetal heart monitoring functions (P<.05), whereas women with moderate income, those without fixed employment, and those who were unemployed, women in early pregnancy or postpartum, and younger women exhibited a significant preference for smart bracelets with vital sign monitoring functions (P<.05). Except for unemployed pregnant and postpartum women, all other subgroups showed a significant preference for smart bracelets with high levels of privacy protection (P<.05). With respect to ease of use, pregnant and postpartum women without fixed employment and those in late pregnancy showed a stronger preference for very easy-to-use devices, while the remaining subgroups all preferred devices that were relatively easy to use (P<.05). In addition, all subgroups of pregnant and postpartum women exhibited a significant preference for a once-daily notification frequency. Overall, factors such as income, employment status, pregnancy stage, parity, presence of pregnancy complications, and age influenced Chinese pregnant and postpartum women's preferences for different smart bracelet attributes to varying degrees.

XSL•FO
RenderX

Discussion

Principal Findings

Our results indicate that the functionality of smart bracelets is the most influential attribute affecting participant preferences, followed by ease of use, frequency of monitoring reports, and price. Participants favored smart bracelets that were affordably priced, offered hospital backend monitoring, included fetal heart rate monitoring, ensured a high level of privacy protection, were relatively user-friendly, and provided frequent monitoring reports. Furthermore, our findings reveal that participant preferences for smart bracelets varied based on sociodemographic characteristics, gestational stage, parity (first-time pregnancy), and the presence of pregnancy complications. Understanding the preferences of participants within these subgroups may facilitate the development of tailored smart bracelets.

The preference for lower-cost smart bracelets aligns with previous research [40]. Our subgroup analysis further demonstrated that participants with lower incomes and those who were unemployed exhibited a significant preference for lower-priced smart bracelets. Jacob et al [41] found that low-income individuals display greater price sensitivity compared to their high-income counterparts, a sensitivity that is exacerbated by income disruption among the unemployed. Additionally, we observed that participants in both early and late pregnancy showed a stronger inclination toward lower-priced smart bracelets. The increased price sensitivity among women in early pregnancy may be attributed to their anticipation of forthcoming childbirth-related expenses, such as delivery fees and newborn care [42]. Notably, heightened price sensitivity during late pregnancy could be linked to the expectation of additional costs related to the imminent delivery [43]. Furthermore, our study revealed that participants who were not experiencing their first pregnancy also exhibited heightened price sensitivity, potentially due to the stress of reallocating household resources [44]. Younger pregnant individuals demonstrated a higher level of price sensitivity, likely reflecting their position in the early stages of their careers, where income levels are generally lower compared to those in mid-to-late career stages. This underscores the necessity of providing cost-effective smart bracelets, particularly for lower-income, unemployed, and first-time pregnant individuals.

Participants expressed a clear preference for smart bracelets equipped with hospital backend monitoring. This finding aligns with Bachiri's research on the integration of mobile health systems, which demonstrated an increase in maternal adherence when hospital data connectivity was enabled [45]. Subgroup analysis revealed that participants with moderate incomes and those experiencing their first pregnancy exhibited a significant preference for hospital backend monitoring, likely due to their reliance on medical authority [46,47]. Additionally, self-employed participants showed a heightened preference for hospital backend monitoring, indicating a perceived need for professional medical support in nontraditional work settings [48]. Younger participants also demonstrated a notable preference for hospital backend monitoring, reflecting a reliance on institutionalized medical care within this demographic [49]. Therefore, incorporating hospital backend monitoring into smart bracelets would address the preferences of first-time mothers, self-employed individuals, and younger users, thereby enhancing trust in the device's medical relevance. For the design of smart bracelets, it is crucial to integrate fetal heart rate monitoring for late-stage pregnancy while also recognizing the importance of vital sign monitoring for early pregnancy and postpartum users.

Regarding the functionality of smart bracelets, participants favored those with fetal heart rate monitoring. This aligns with prior research, such as Ahmed et al [50], which indicated that fetal heart rate monitoring was the significantly preferred feature among pregnant women, directly related to their sensitivity to pregnancy health risks. Other studies have shown that abnormal fetal heart rates are important early warning signs of pregnancy complications, and the continuous monitoring capabilities of smart bracelets can increase the detection rate of anomalies [13,51]. However, subgroup analysis revealed that both low-income and high-income groups preferred fetal heart rate monitoring, while the middle-income group tended to favor vital sign monitoring. This differential preference among the middle-income group may stem from their heightened health management awareness, leading them to favor multidimensional health data integration [52]. This nonlinear relationship may also arise because high-income individuals prioritize fetal safety, while low-income individuals rely on basic monitoring as a substitute for prenatal care, and middle-income individuals seek comprehensive health management [53-55]. Our research also found that freelancers and unemployed participants preferred vital sign monitoring, while employed participants preferred fetal heart rate monitoring. This may be because self-employment or unemployment allows more time to focus on one's own physiological changes, such as adjusting lifestyles by monitoring sleep quality or stress levels [56]. Unemployed individuals may experience higher health anxiety due to financial pressures, leading them to prefer vital sign data for preventing potential health risks. For employed individuals under work pressure, fetal heart rate monitoring is viewed as a direct indicator of fetal health, enabling quick identification of anomalies and reducing the frequency of medical visits [7,57]. Additionally, employed individuals may have access to professional prenatal care through workplace medical benefits, thereby reducing their need for vital sign monitoring [58,59]. Participants in early pregnancy and postpartum preferred vital sign monitoring, while those in the mid-to-late stages of pregnancy preferred fetal heart rate monitoring. This may be because, during the initial bodily adaptation phase of early pregnancy, women are more concerned with their own physiological changes (eg, morning sickness and fatigue) [60,61] and early warning signs of pregnancy complications (eg, anemia and hypertension) [62,63], and there is an increased need for health monitoring during the recovery period, such as postpartum depression risk, blood pressure fluctuations, and wound healing tracking [64,65]. Literature indicates that continuous vital sign monitoring is feasible within 7 months postpartum [66]. The mid-to-late stages of pregnancy are critical periods for fetal development. Fetal movement becomes regular after 20 weeks, and fetal heart rate monitoring becomes a core indicator for assessing fetal development, directly related to

XSL•FO
RenderX

pregnancy complications such as preeclampsia and fetal growth restriction [67-69]. Pregnant women in late pregnancy experience a significant increase in anxiety about fetal safety, and fetal heart rate data can provide immediate psychological reassurance [70].

Regarding privacy protection, participants generally preferred smart bracelets with a high level of privacy protection, indicating that data security is a core concern for users. This finding aligns with existing research highlighting users' privacy concerns related to wearable health devices [71,72]. Additionally, participants placed a high value on ease of use. Notably, while very easy-to-use bracelets were favored to some extent, the intensity of preference was lower compared to "relatively easy-to-use" bracelets. This may be due to participants feeling that very simple bracelets lack sufficient functionality or intelligent experience, reflecting a synergistic need for functional depth and operational convenience within the pregnant and postpartum population [73]. This observation aligns with human-computer interaction research on wearable devices, where overly simplified interfaces can undermine user trust in data reliability [74,75]. Younger participants preferred vital sign monitoring, while older participants favored fetal heart rate monitoring. Younger pregnant women perceive their pregnancy risk as lower, focusing more on their own health rather than fetal-specific indicators. In contrast, older mothers face a significantly higher incidence of pregnancy complications (eg, gestational hypertension and premature delivery), leading them to consider fetal heart rate monitoring as a necessary risk management tool [16,76]. Therefore, smart bracelet design should prioritize user privacy and ease of use, balancing simplicity with functionality to build trust among all user demographics.

Regarding the frequency of monitoring report delivery, participants generally preferred smart bracelets capable of providing more frequent updates. This preference indicates that regular health data updates are a significant requirement for pregnant and postpartum individuals seeking to manage their health through real-time feedback. However, it is crucial to acknowledge that high-frequency monitoring may increase health anxiety [77]. Therefore, it is advisable to mitigate the false alarm rate through AI-assisted interpretation [78].

The design of smart bracelets must comprehensively consider key factors such as functionality, price, ease of use, privacy protection, and report frequency. This approach is essential for accurately aligning with the diverse preferences of pregnant and postpartum individuals across various gestational stages, income levels, employment statuses, and parity. Priority should be given to integrating functions for monitoring fetal heart rate and vital signs, with differentiated configurations tailored to user characteristics. While ensuring data privacy and operational convenience, the human-computer interaction interface should be optimized to avoid overly simplified functions that could undermine user trust. Furthermore, establishing appropriate monitoring frequencies and incorporating AI-assisted interpretation can enhance the practicality and reassurance of data feedback. Only by addressing these multifaceted needs can smart wearable devices be effectively tailored to meet the demands of the maternal population.

Limitations

This study has several limitations. First, the sample was primarily drawn from a specific regional population of pregnant and postpartum individuals, which may not fully represent the preferences of users across diverse regions, cultural backgrounds, and health care environments. Second, the data were based on self-reported responses, which are subject to potential biases, such as subjective perception and social desirability, and may not accurately reflect actual user behavior during device usage. Third, although detailed subgroup analyses were conducted, other potentially influential variables—such as digital health literacy and family support levels—were not explored. Additionally, the selection of smart bracelet features was based on current mainstream product configurations, which may change as technology evolves. Furthermore, while a mixed logit model was appropriately used to estimate preferences and accommodate heterogeneity across the overall sample, some individual subgroups (eg, "unemployed or out of work" and "no fixed occupation (freelancer)") had relatively small sample sizes. Small subgroup counts can increase uncertainty when interpreting subgroup-specific effects or interactions, potentially yielding unstable estimates or larger SEs for those categories. For this reason, findings related to these small subgroups should be interpreted with caution and not overgeneralized. Future research should consider expanding the sample scope and incorporating longitudinal studies and real-world usage data to further validate and enrich the findings, as well as examine additional factors that may influence preferences for wearable health technology.

Conclusions

In conclusion, preferences for smart bracelets among pregnant individuals are influenced by various factors, including functionality, price sensitivity, ease of use, and privacy concerns. The most significant features identified are the primary function and ease of use. Design considerations should account for the diverse needs across different sociodemographic groups, stages of pregnancy, and potential complications. Specifically, cost-effective options, robust privacy protections, and user-friendly interfaces are essential to ensure widespread adoption. Furthermore, integrating real-time monitoring with AI-assisted feedback can enhance the usability of smart bracelets, addressing concerns related to health anxiety and improving the overall user experience. Ultimately, a personalized and comprehensive approach to smart bracelet design can better support the health and well-being of pregnant women during both their pregnancy and postpartum periods.

XSL•FO
RenderX

Acknowledgments

The authors are grateful to the professionals and students who provided enormous support in the recruitment and data collection of this study and the participants who took the time to cooperate with the survey.

Funding

No funding was received for this study.

Data Availability

The datasets used and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JW and MC were responsible for study design, data collection, and initial manuscript drafting. JW, JF, and WL assisted in data analysis and interpretation. XZ contributed to literature review and helped revise the manuscript. YC, HF, and FY were involved in participant recruitment and questionnaire distribution. XW provided critical revisions and supervised the methodology. XW and YW conceptualized the study and supervised the entire project for all correspondence related to this work. All authors provided revisions and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Subgroup analysis results for preference heterogeneity in the discrete choice experiment.
[DOCX File, 48 KB - jmir_v28i1e55941_app1.docx]

References

1. Cao J, Xu W, Liu Y, et al. Trends in maternal age and the relationship between advanced age and adverse pregnancy outcomes: a population-based register study in Wuhan, China, 2010-2017. Public Health 2022 May;206:8-14. [doi: 10.1016/j.puhe.2022.02.015] [Medline: 35306195]
2. Souza JP, Day LT, Rezende-Gomes AC, et al. A global analysis of the determinants of maternal health and transitions in maternal mortality. Lancet Glob Health 2024 Feb;12(2):e306-e316. [doi: 10.1016/S2214-109X(23)00468-0] [Medline: 38070536]
3. Jiang Q, Guo Y, Zhang E, et al. Perinatal mental health problems in rural China: the role of social factors. Front Psychiatry 2021;12:636875. [doi: 10.3389/fpsyt.2021.636875] [Medline: 34950062]
4. Zhao Q, Chen J, Li F, Li A, Li Q. An integrated model for evaluation of maternal health care in China. PLoS One 2021;16(1):e0245300. [doi: 10.1371/journal.pone.0245300] [Medline: 33507961]
5. Chauhan A, Potdar J. Maternal mental health during pregnancy: a critical review. Cureus 2022 Oct;14(10):e30656. [doi: 10.7759/cureus.30656] [Medline: 36426343]
6. Optimizing postpartum care. ACOG committee opinion no.736. American College of Obstetricians and Gynecologists. 2018. URL: https://www.acog.org/clinical/clinical-guidance/committee-opinion/articles/2018/05/optimizing-postpartum-care [accessed 2025-05-15]
7. Li S, Yang Q, Niu S, Liu Y. Effectiveness of remote fetal monitoring on maternal-fetal outcomes: systematic review and meta-analysis. JMIR Mhealth Uhealth 2023 Feb 22;11:e41508. [doi: 10.2196/41508] [Medline: 36811944]
8. Wang J, Jiang Y, Fu X, et al. Evaluating the impact of interactive video-based case-based learning in clinical medical education: a randomized controlled trial. Front Med 2025 May 16;12:1556018. [doi: 10.3389/fmed.2025.1556018]
9. Huang Y, Guo J, Chen Z, et al. The impact of electronic health records (EHR) data continuity on prediction model fairness and racial-ethnic disparities. arXiv. Preprint posted online on Sep 5, 2023. [doi: 10.48550/arXiv.2309.01935]
10. Milan G, Ariniello L, Baca-Motes K, et al. Association between wearable sensor signals and expected hormonal changes in pregnancy. EBioMedicine 2025 Sep;119:105888. [doi: 10.1016/j.ebiom.2025.105888] [Medline: 40876228]
11. Liu L, Pu Y, Fan J, et al. Wearable sensors, data processing, and artificial intelligence in pregnancy monitoring: a review. Sensors (Basel) 2024 Oct 4;24(19):6426. [doi: 10.3390/s24196426]
12. Stricker K, Radan AP, Surbek D. Continuous remote home monitoring solutions for mother and fetus: a scoping review. Eur J Obstet Gynecol Reprod Biol 2025 Feb;305:170-177. [doi: 10.1016/j.ejogrb.2024.12.018] [Medline: 39705987]
13. Alim A, Imtiaz MH. Wearable sensors for the monitoring of maternal health-a systematic review. Sensors (Basel) 2023 Feb 22;23(5):2411. [doi: 10.3390/s23052411] [Medline: 36904615]
14. Keeler Bruce L, González D, Dasgupta S, Smarr BL. Biometrics of complete human pregnancy recorded by wearable devices. NPJ Digit Med 2024 Aug 12;7(1):207. [doi: 10.1038/s41746-024-01183-9] [Medline: 39134787]

XSL•FO
RenderX

15. Grym K, Niela-Vilén H, Ekholm E, et al. Feasibility of smart wristbands for continuous monitoring during pregnancy and one month after birth. BMC Pregnancy Childbirth 2019 Jan 17;19(1):34. [doi: 10.1186/s12884-019-2187-9] [Medline: 30654747]
16. Wu S, Cui B, Yu X. Willingness to use smart fetal heart rate monitoring devices among pregnant women: an extension of the technology acceptance model. Front Psychol 2024;15:1400720. [doi: 10.3389/fpsyg.2024.1400720] [Medline: 39070584]
17. Schramm K, Grassl N, Nees J, et al. Women's attitudes toward self-monitoring of their pregnancy using noninvasive electronic devices: cross-sectional multicenter study. JMIR Mhealth Uhealth 2019 Jan 7;7(1):e11458. [doi: 10.2196/11458] [Medline: 30617040]
18. Wakefield C, Yao L, Self S, Frasch MG. Wearable technology for health monitoring during pregnancy: an observational cross-sectional survey study. Arch Gynecol Obstet 2023 Jul;308(1):73-78. [doi: 10.1007/s00404-022-06705-y] [Medline: 35831759]
19. Young-Lin N, Heneghan C, Liu Y, et al. Insights into maternal sleep: a large-scale longitudinal analysis of real-world wearable device data before, during, and after pregnancy. EBioMedicine 2025 Apr;114:105640. [doi: 10.1016/j.ebiom.2025.105640] [Medline: 40147048]
20. Litman EA, Kavathekar T, Amdur R, Sebastian A, Marko K. Remote gestational weight gain monitoring in a large low-risk US population. Obes Sci Pract 2022 Apr;8(2):147-152. [doi: 10.1002/osp4.554] [Medline: 35388344]
21. Wang Q, Yang W, Li L, Yan G, Wang H, Li J. Late pregnancy analysis with Yunban's remote fetal monitoring system. International Journal of Distributed Sensor Networks 2019 Mar;15(3):155014771983283. [doi: 10.1177/1550147719832835]
22. Lillis TA, Hamilton NA, Pressman SD, et al. Sleep quality buffers the effects of negative social interactions on maternal mood in the 3–6 month postpartum period: a daily diary study. J Behav Med 2018 Oct;41(5):733-746. [doi: 10.1007/s10865-018-9967-y]
23. Chivers BR, Garad RM, Moran LJ, Lim S, Harrison CL. Support seeking in the postpartum period: content analysis of posts in web-based parenting discussion groups. J Med Internet Res 2021 Jul 15;23(7):e26600. [doi: 10.2196/26600] [Medline: 34264198]
24. Cook F, Bayer J, Le HND, Mensah F, Cann W, Hiscock H. Baby Business: a randomised controlled trial of a universal parenting program that aims to prevent early infant sleep and cry problems and associated parental depression. BMC Pediatr 2012 Feb 6;12:13. [doi: 10.1186/1471-2431-12-13] [Medline: 22309617]
25. Arije O, Madan J, Hlungwani T. Preferences in adolescents and young people's sexual and reproductive health services in Nigeria: a discrete choice experiment. Health Econ Rev 2024 Mar 22;14(1):24. [doi: 10.1186/s13561-024-00497-4] [Medline: 38514523]
26. Luo N, Bai R, Sun Y, et al. Job preferences of master of public health students in China: a discrete choice experiment. BMC Med Educ 2024 Jan 4;24(1):24. [doi: 10.1186/s12909-023-04993-9] [Medline: 38178052]
27. Donnan J, Johnston K, Coombs M, Najafizada M, Bishop L. Exploring consumer preferences for cannabis edible products to support public health policy: a discrete choice experiment. PLoS One 2024;19(5):e0292336. [doi: 10.1371/journal.pone.0292336] [Medline: 38753807]
28. Jones B, Ryan M, Cook NS, Gutzwiller FS. Development of a disease-specific health utility score for chronic obstructive pulmonary disease from a discrete choice experiment patient preference study. Int J Technol Assess Health Care 2024 May 2;40(1):e30. [doi: 10.1017/S0266462324000242] [Medline: 38695141]
29. Trapero-Bertran M, Rodríguez-Martín B, López-Bastida J. What attributes should be included in a discrete choice experiment related to health technologies? A systematic literature review. PLoS ONE 2019;14(7):e0219905. [doi: 10.1371/journal.pone.0219905] [Medline: 31318926]
30. Mak SS, Kinney RL, Bailey AL, et al. Integrating consumer-grade wearable devices and patient-generated health data into clinical care: perspectives from healthcare professionals at a learning health system. J Gen Intern Med 2025 Oct 9:41068528. [doi: 10.1007/s11606-025-09876-x] [Medline: 41068528]
31. Xie Z, Liu H, Or C. A discrete choice experiment to examine the factors influencing consumers' willingness to purchase health apps. Mhealth 2023;9:21. [doi: 10.21037/mhealth-22-39] [Medline: 37492118]
32. Zhang B, Chen C, Lee I, Lee K, Ong KL. A survey on security and privacy issues in wearable health monitoring devices. Computers & Security 2025 Aug;155:104453. [doi: 10.1016/j.cose.2025.104453]
33. Moorthy P, Weinert L, Schüttler C, et al. Attributes, methods, and frameworks used to evaluate wearables and their companion mHealth apps: scoping review. JMIR Mhealth Uhealth 2024 Apr 5;12:e52179. [doi: 10.2196/52179] [Medline: 38578671]
34. Godoy Junior CA, Mäkitie L, Fiorenzato E, et al. Diverse preferences, different solutions: exploring remote monitoring preferences in Parkinson's disease through a discrete choice experiment. J Parkinsons Dis 2025 May;15(3):619-629. [doi: 10.1177/1877718X251327752] [Medline: 40123353]
35. Milte R, Ratcliffe J, Chen G, Lancsar E, Miller M, Crotty M. Cognitive overload? An exploration of the potential impact of cognitive functioning in discrete choice experiments with older people in health care. Value Health 2014 Jul;17(5):655-659. [doi: 10.1016/j.jval.2014.05.005] [Medline: 25128060]
36. Bailey RA. Balance, orthogonality and efficiency factors in factorial design. Journal of the Royal Statistical Society Series B 1985 Jul 1;47(3):453-458. [doi: 10.1111/j.2517-6161.1985.tb01374.x]

XSL•FO
RenderX

37. Orme B. Sample size issues for conjoint analysis studies. : Sawtooth Software; 1998 URL: https://sawtoothsoftware.com/resources/technical-papers/sample-size-issues-for-conjoint-analysis-studies [accessed 2025-05-15]
38. Johnson R, Orme B. Getting the most from CBC. : Sawtooth Software; 2003 URL: https://sawtoothsoftware.com/resources/technical-papers/getting-the-most-from-cbc [accessed 2025-05-15]
39. Zhang Y, Fan S, Hui H, et al. Privacy protection for open sharing of psychiatric and behavioral research data: ethical considerations and recommendations. Alpha Psychiatry 2025 Feb;26(1):38759. [doi: 10.31083/AP38759] [Medline: 40110382]
40. Neubauer N, Ciawi Y, Komaladewi A, Salain I. Additive manufacturing in construction: bridging innovation and sustainability. Acad Engg 2025;2(2). [doi: 10.20935/AcadEng7691]
41. Jacob J, Vieites Y, Goldszmidt R, Andrade EB. Expected socioeconomic-status-based discrimination reduces price sensitivity among the poor. J Mark Res 2022 Dec;59(6):1083-1100. [doi: 10.1177/00222437221097100]
42. Subramaniam C, Lee EM, Park S. Is the usage of wearable devices worth it? a study of perceived risk and fashion on intention to adopt wearable devices (preprint). JMIR Formative Research. Preprint posted online on Jun 19, 2025. [doi: 10.2196/preprints.56887]
43. Slaba M. Multiple regression model of the consumers' price sensitivity. EA-XXI 2020 Jun 4;183(5-6):60-65. [doi: 10.21003/ea.V183-06]
44. Fujii T, Ishikawa R. How does childbirth alter intrahousehold resource allocation? Evidence from Japan*. Oxf Bull Econ Stat 2013 Jun;75(3):362-387 [FREE Full text] [doi: 10.1111/j.1468-0084.2012.00699.x]
45. Bachiri M, Idri A, Fernández-Alemán JL, Toval A. Mobile personal health records for pregnancy monitoring functionalities: Analysis and potential. Comput Methods Programs Biomed 2016 Oct;134:121-135. [doi: 10.1016/j.cmpb.2016.06.008] [Medline: 27480737]
46. Faisal-Cury A, Quayle J, Marques T, Menezes PR, Matijasevich A. The relationship between socioeconomic indicators during pregnancy and gynecological appointment at any time after childbirth. Int J Equity Health 2015 Aug 12;14:64. [doi: 10.1186/s12939-015-0191-x] [Medline: 26260153]
47. Milcent C, Zbiri S. Prenatal care and socioeconomic status: effect on cesarean delivery. Health Econ Rev 2018 Mar 10;8(1):7. [doi: 10.1186/s13561-018-0190-x] [Medline: 29525909]
48. Goodman JM, Colon M, Cottrell EK, Guise JM. Work as a social risk factor in pregnancy: a systematic review of screening practices related to working conditions and family leave among pregnant adults. Birth 2023 Mar;50(1):32-43. [doi: 10.1111/birt.12695] [Medline: 36455239]
49. Grand-Guillaume-Perrenoud JA, Origlia P, Cignacco E. Barriers and facilitators of maternal healthcare utilisation in the perinatal period among women with social disadvantage: a theory-guided systematic review. Midwifery 2022 Feb;105:103237. [doi: 10.1016/j.midw.2021.103237] [Medline: 34999509]
50. Ahmed MR, Newby S, Potluri P, Mirihanage W, Fernando A. Emerging paradigms in fetal heart rate monitoring: evaluating the efficacy and application of innovative textile-based wearables. Sensors (Basel) 2024;24(18):6066. [doi: 10.3390/s24186066]
51. Lear CA, Westgate JA, Ugwumadu A, et al. Understanding fetal heart rate patterns that may predict antenatal and intrapartum neural injury. Semin Pediatr Neurol 2018 Dec;28:3-16. [doi: 10.1016/j.spen.2018.05.002] [Medline: 30522726]
52. Leziak K, Birch E, Jackson J, Strohbach A, Niznik C, Yee LM. Identifying mobile health technology experiences and preferences of low-income pregnant women with diabetes. J Diabetes Sci Technol 2021 Sep;15(5):1018-1026. [doi: 10.1177/1932296821993175]
53. Gan Y, Zhu C, Zhou Y, et al. Clinical efficacy and acceptability of remote fetal heart rate self-monitoring in Southern China. BMC Pregnancy Childbirth 2023 Oct 7;23(1):715. [doi: 10.1186/s12884-023-05985-9] [Medline: 37805457]
54. Asadollahi F, Zagami SE, Eslami S, Roudsari RL. Barriers and facilitators for mHealth utilization in pregnancy care: a qualitative analysis of pregnant women and stakeholder's perspectives. BMC Pregnancy Childbirth 2025 Feb 11;25(1):141. [doi: 10.1186/s12884-025-07244-5] [Medline: 39934681]
55. Iacoban SR, Piron Dumitrascu M, Bohiltea CL, Suciu ID, Voinea SC, Suciu N. Socio-economic disparities in prenatal prognosis and intervention accessibility. J Mind Med Sci 2024;11(2):402-411. [doi: 10.22543/2392-7674.1541]
56. González Ramírez ML, García Vázquez JP, Rodríguez MD, Padilla-López LA, Galindo-Aldana GM, Cuevas-González D. Wearables for stress management: a scoping review. Healthcare (Basel) 2023 Aug 22;11(17):2369. [doi: 10.3390/healthcare11172369] [Medline: 37685403]
57. Kern-Goldberger AR, Nicholls EM, Plastino N, Srinivas SK. The impact of an intervention to improve intrapartum maternal vital sign monitoring and reduce alarm fatigue. Am J Obstet Gynecol MFM 2023 May;5(5):100893. [doi: 10.1016/j.ajogmf.2023.100893]
58. How self-funded employers can support maternal health. Healthgram. 2024. URL: https://www.healthgram.com/insights/how-self-funded-employers-can-support-maternal-health/ [accessed 2025-05-15]
59. Tailored prenatal care delivery for pregnant individuals: ACOG Clinical Consensus No. 8. Obstet Gynecol 2025 May 17;145(5):565-577. [doi: 10.1097/AOG.0000000000005889] [Medline: 40245426]
60. Vomiting and morning sickness. NHS. 2019. URL: https://www.nhs.uk/pregnancy/related-conditions/common-symptoms/vomiting-and-morning-sickness/ [accessed 2025-05-15]

XSL•FO
RenderX

61. First trimester fatigue. Johns Hopkins Medicine. 2024. URL: https://www.hopkinsmedicine.org/health/conditions-and-diseases/first-trimester-fatigue [accessed 2025-05-15]
62. Anemia during pregnancy: symptoms, risks & prevention. Cleveland Clinic. 2022. URL: https://my.clevelandclinic.org/health/diseases/23112-anemia-during-pregnancy [accessed 2025-05-15]
63. High blood pressure during pregnancy. Centers for Disease Control and Prevention (CDC). 2024. URL: https://www.cdc.gov/high-blood-pressure/about/high-blood-pressure-during-pregnancy.html [accessed 2025-05-15]
64. Zhou S, Jin Q, Jiang X, et al. Application of remote fetal heart rate monitoring via internet in late pregnancy during the COVID-19 pandemic. Technol Health Care 2023;31(4):1105-1117. [doi: 10.3233/THC-220700] [Medline: 37125583]
65. Rhoads SJ, Serrano CI, Lynch CE, et al. Exploring implementation of m-Health monitoring in postpartum women with hypertension. Telemed J E Health 2017 Oct;23(10):833-841. [doi: 10.1089/tmj.2016.0272] [Medline: 28475431]
66. Saarikko J, Niela-Vilen H, Ekholm E, et al. Continuous 7-month internet of things–based monitoring of health parameters of pregnant and postpartum women: prospective observational feasibility study. JMIR Form Res 2020 Jul 24;4(7):e12417. [doi: 10.2196/12417]
67. Umana OD, Vadakekut ES, Siccardi MA. Antenatal Fetal Surveillance: StatPearls Publishing; 2025. URL: https://www.ncbi.nlm.nih.gov/books/NBK537123/ [accessed 2025-05-15]
68. Burgess A, Aucutt M, Coleman SL. Standardizing fetal movement monitoring using count the kicks. MCN Am J Matern Child Nurs 2024;49(6):306-313. [doi: 10.1097/NMC.0000000000001048]
69. Kintraia PI, Zarnadze MG, Kintraia NP, Kashakashvili IG. Development of daily rhythmicity in heart rate and locomotor activity in the human fetus. J Circadian Rhythms 2005 Mar 31;3(0):5. [doi: 10.1186/1740-3391-3-5]
70. Semeia L, Bauer I, Sippel K, Hartkopf J, Schaal NK, Preissl H. Impact of maternal emotional state during pregnancy on fetal heart rate variability. Compr Psychoneuroendocrinol 2023 May;14:100181. [doi: 10.1016/j.cpnec.2023.100181] [Medline: 36911250]
71. Vitak J, Liao Y, Kumar P, Zimmer M, Kritikos K. Privacy attitudes and data valuation among fitness tracker users. In: Chowdhury G, McLeod J, Gillet V, Willett P, editors. Transforming Digital Worlds iConference 2018 Lecture Notes in Computer Science: Springer; 2018, Vol. 10766. [doi: 10.1007/978-3-319-78105-1_27]
72. Schnall R, Higgins T, Brown W, Carballo-Dieguez A, Bakken S. Trust, perceived risk, perceived ease of use and perceived usefulness as factors related to mHealth technology use. Stud Health Technol Inform 2015;216:467-471. [Medline: 26262094]
73. Xie Z, Or CK. Consumers' preferences for purchasing mHealth apps: discrete choice experiment. JMIR Mhealth Uhealth 2023 Sep 13;11:e25908-e25908. [doi: 10.2196/25908]
74. Karsh B, Holden RJ, Or CKL. Human factors and ergonomics of health information technology implementation. In: Carayon P, editor. Handbook of Human Factors and Ergonomics in Health Care and Patient Safety: CRC Press; 2011:249-264. [doi: 10.1201/b11219-23]
75. Or CK, Holden RJ, Valdez RS. Human factors engineering and user-centered design for mobile health technology: enhancing effectiveness, efficiency, and satisfaction. In: Human-Automation Interaction: Mobile Computing: Springer; 2022:97-118. [doi: 10.1007/978-3-031-10788-7_6]
76. Dietl A, Farthmann J. Gestational hypertension and advanced maternal age. Lancet 2015 Oct 24;386(10004):1627-1628. [doi: 10.1016/S0140-6736(15)00532-2] [Medline: 26595630]
77. Abd-alrazaq A, AlSaad R, Harfouche M, et al. Wearable artificial intelligence for detecting anxiety: systematic review and meta-analysis. J Med Internet Res 2023 Nov 8;25:e48754. [doi: 10.2196/48754]
78. Kargarandehkordi A, Li S, Lin K, Phillips KT, Benzo RM, Washington P. Fusing wearable biosensors with artificial intelligence for mental health monitoring: a systematic review. Biosensors (Basel) 2025 Mar 21;15(4):202. [doi: 10.3390/bios15040202] [Medline: 40277515]

Abbreviations

AI: artificial intelligence
CNKI: China National Knowledge Infrastructure
DCE: discrete choice experiment
MeSH: medical subject headings
WHO: World Health Organization
WTP: willingness to pay

XSL•FO
RenderX

Edited by A Schwartz, M Balcarras; submitted 16.Jun.2025; peer-reviewed by B Tian, U Pasupulety; revised version received 09.Feb.2026; accepted 24.Feb.2026; published 12.May.2026.

Please cite as:
Wang J, Cheng M, Feng J, Li W, Zhou X, Cao Y, Feng H, Yang F, Wang X, Wu Y
Preferences and Willingness to Pay for Smart Bracelets Among Chinese Pregnant and Postpartum Women: Discrete Choice Experiment
J Med Internet Res 2026;28:e55941
URL: https://www.jmir.org/2026/1/e55941
doi:10.2196/55941

© Jing Wang, Mingyue Cheng, Jialiang Feng, Wenzhuo Li, Xiaotao Zhou, Yufang Cao, Haiyan Feng, Fan Yang, Xiaoyun Wang, Yibo Wu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessing the Use of Wearable Mobile-Monitoring Devices Among Individuals With Serious Mental Illness: Qualitative Acceptability and Feasibility Study

Aubrey M Freitas[1], BA; Jesus G Chavez[1], MPH; Melissa Chinchilla[1], PhD; Ronald Calderon[1], MSW; Stephanie Chassman[1], PhD; Lauren Hoffmann[1], MS; Alexander S Young[1,2], MD

[1]Mental Illness Research, Education, and Clinical Center, VA Greater Los Angeles Healthcare System, U.S. Department of Veteran Affairs, 11301 Wilshire Blvd, Los Angeles, CA, United States

[2]Department of Psychiatry and Biobehavioral Sciences, Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles Geffen School of Medicine, Los Angeles, CA, United States

Corresponding Author:
Aubrey M Freitas, BA
Mental Illness Research, Education, and Clinical Center, VA Greater Los Angeles Healthcare System, U.S. Department of Veteran Affairs, 11301 Wilshire Blvd, Los Angeles, CA, United States

Abstract

Background: Serious mental illness (SMI) is difficult to treat for various reasons, such as rapid changes in symptoms, comorbid health conditions, long gaps between provider visits, and additional societal barriers experienced by this population. Wearable mobile-sensing devices can be used to passively collect valuable patient-generated health data, such as daily step count, heart rate variability, sleep information, and other health-related behaviors, which could inform and improve treatment for individuals with SMI. Wearable health devices have become more economically accessible, providing promise for the possibility of their implementation in health care. However, more information regarding how individuals with SMI perceive and interact with these devices is needed.

Objective: This study aimed to assess the acceptability and feasibility of using wearable mobile-sensing devices to improve treatment outcomes for Veterans with SMI. In addition, we were also interested in learning if privacy concerns would influence acceptability of devices, specifically surrounding location tracking and health information sharing, as well as assessing other barriers to device use.

Methods: Qualitative interviews were conducted with participants who had been using a wearable health and fitness tracker for at least 2 weeks to explore their thoughts and perceptions of these devices. A total of 15 Veterans diagnosed with a SMI participated in interviews. Both thematic analysis and rapid qualitative analysis approaches were used to uncover findings in key domains and emergent themes.

Results: Wearable fitness trackers allowed participants to conveniently monitor various aspects of their physical and mental health, provided a greater understanding of their overall well-being, and motivated them to reach personal health goals. Individuals were open to sharing their personal health information collected from the devices with providers to improve their health care treatment and expressed no privacy concerns surrounding data tracking or the device's global positioning system that monitors physical location. Participants experienced some technological challenges with using the fitness trackers, as well as the device's accompanying cell phone app. Furthermore, participants expressed difficulties in understanding and interpreting the health data that was collected from the health and fitness trackers. Greater ongoing technological support, in addition to physical device adjustments to enhance comfort and usability, were suggested ways of improving overall user experience.

Conclusions: Participants with SMI in this sample were accepting of wearable mobile-monitoring devices and believe it is feasible to incorporate these fitness trackers into their daily lives. Furthermore, participants in this sample expressed no privacy concerns regarding location tracking or the sharing of health information collected from these devices with providers. Patient-generated health data collected from these devices may offer valuable information that could be used to inform health care treatment for this population.

(J Med Internet Res 2026;28:e85087) doi:10.2196/85087

KEYWORDS

serious mental illness; wearable devices; mobile sensing; wearables; patient-generated health data

XSL•FO
RenderX

Introduction

Overview

Serious mental illnesses (SMIs) impact nearly 15.4 million adults [1], are challenging to treat, and require years of monitoring and adjustments in treatment. Diagnoses of SMI have been linked to additional life challenges, including job loss [2], homelessness [3], incarceration [4], hospitalization [5,6], or suicide [7]. Clinician visits are infrequent, and illness exacerbations often occur with no clinician awareness in real time, leaving no opportunity to adjust treatment. Methods are needed for monitoring mental health and behavioral status of individuals with SMI to support outreach, care coordination, and treatment. Behavioral phenotypes have been developed from mobile data, and wearable devices, which are now common in the population and are well-suited to produce these data and support clinical services. However, it is not yet clear whether people with SMI will accept and are interested in engaging with wearable devices and mobile monitoring to support care delivery.

Individuals with SMI experience several health comorbidities [8], like cardiovascular disease [9], and have higher mortality rates than the general population [10], with some estimating about 2 to 3 times higher [11]. In addition, SMI has been associated with high financial burdens on society through increased emergency service use [12] and has been linked to high rates of disability [13]. Medication is a common treatment for SMI symptoms; however, complex medication regimens present challenges to adherence for patients, and many antipsychotic medications are associated with adverse side effects that may exacerbate or create additional health challenges [14]. Furthermore, this population also experiences additional and systemic barriers, such as stigma, poor coordination of treatment, and lack of interdisciplinary care for physical and mental health symptoms, which can contribute to mistrust in health care systems [15]. Exploring the ways mobile-sensing devices can be used to benefit treatment outcomes may address multiple overlapping barriers, risks, and rapid, unpredictable changes in symptoms.

Prior Work

Past research has shown both increased use of and favorable attitudes toward mobile devices, particularly cell phones, among individuals with SMI [16,17], including Veterans with SMI [18], making the possibility of integrating mobile-monitoring devices and their partner cell phone apps into current health care treatments feasible for this population. Incorporating mobile health interventions via smartphones can improve treatment outcomes [19] and intervention implementation through accessible web-based therapies, leading to improvements in SMI symptoms, reductions in hospitalizations, and increases in social connectedness [20]. In addition, the use of cell phone apps has been found to increase patient knowledge of their health conditions [21], and this additional information may be particularly helpful in allowing individuals with SMI to advocate for their needs and monitor symptoms, as well as give providers more insight into their overall health, which could improve treatment planning and care coordination.

Research conducted among Veterans with SMI shows that remote sensing devices can be used to effectively monitor sleep and physical activity data [22]; however, while this research supports the efficacy of using these devices partnered with intensive support, further study is needed to examine their effectiveness when incorporated into usual care. Changes in health behaviors have been found to be correlated with the onset of psychiatric symptoms, such as abnormal sleep patterns [23] and changes in physical activity levels [24] before experiencing a relapse or a re-emergence of positive symptoms. Therefore, by monitoring these and similar health behaviors through wearable mobile-monitoring devices, patient-generated health data (PGHD) could potentially be used to predict relapses in individuals with SMI. Furthermore, if patients were able to self-monitor their health behaviors with wearable devices, they may be able to seek assistance before a relapse occurs. Additionally, if the health information from wearable devices is shared with patients' providers, providers can track behavior changes between medical visits, and proactively reach out to patients, and, if needed, adjust treatments to prevent a lapse in symptoms. Research in this area found that clinicians are open to implementing the use of mobile-sensing technology into health care treatment [25,26]. Similarly, individuals with SMI are open to the idea of self-monitoring technology [27]. Further study is needed to better assess potential concerns of mobile-monitoring devices among this population surrounding privacy, location tracking, sharing PGHD with clinicians, as well as discovering ways to improve the devices for individuals with SMI.

Objective

In this study, we used qualitative methods to examine the acceptability and feasibility of using mobile-sensing technology to inform and improve health care outcomes for individuals with SMI. We were interested in understanding whether participants found PGHD useful and potential barriers to data use and integration in their health care treatment. This study expands our knowledge of the thoughts and perceptions of individuals with SMI toward mobile-sensing technology and incorporating wearables into their daily lives, which can inform future research exploring the integration of PGHD collected from these devices in a clinical setting to improve treatment outcomes.

Methods

Recruitment

This study served as a qualitative arm of a larger pilot study [28] exploring how passive monitoring through a mobile-sensing cell phone app can be used to improve health care treatment and outcomes for individuals with SMI. To be eligible to participate, Veterans needed to receive care at the Greater Los Angeles Department of Veterans Affairs (GLA VA) and have received a SMI diagnosis of either schizophrenia, schizoaffective disorder, or bipolar disorder, or have received a diagnosis of posttraumatic stress disorder (PTSD) and be on a qualifying antipsychotic medication. Research staff screened patients' records using the Department of Veterans Affairs (VA's) electronic health record system, and eligible participants were

XSL•FO
RenderX

recruited via letter and follow-up phone calls. For the larger study, a total of 96 participants (n=96) were enrolled (535 Veterans were approached, 197 were not interested, 229 were unable to contact, 13 were not eligible due to an early version of the project's app being incompatible with Android smartphones).

Study Design

Individuals who participated in qualitative interviews were randomly selected out of the total 96 Veterans in the larger arm of the project to participate in the qualitative portion of the study using wearable mobile-monitoring devices. A total of 15 Veterans (n=15) were approached, and all consented to be interviewed. These participants were provided a wearable health and fitness tracking device (a Fitbit watch, either an Inspire 2 or Inspire 3 version). Fitbit devices were deemed appropriate for this project as they assess the study's domains of interest (eg, stress, sleep, and physical activity), are compatible with both Android and iPhone cellular devices, are relatively affordable, and are market-leaders for mobile sensing devices. At the time of the interviews, all participants had been using their wearable device for a period ranging from 2 weeks to 1 month.

All Fitbit devices come with an accompanying app that displays the user's longitudinal health data collected from the wearable device. The app also provides access to mindfulness exercises, physical fitness routines, and allows users to input additional health information, such as food and water intake. Upon providing individuals with their watch, research staff downloaded the Fitbit app to participants' cell phones and briefly demonstrated how the app can be used along with the wearable device itself.

Semistructured interviews were conducted by research staff. Three interviewers (AF, LH, and RC) from the study team interviewed participants individually using a semistructured interview guide. Two interviewers (AF and LH) self-identified as female and one as male (RC). Two interviewers (LH and RC) were employed at the GLA VA as qualitative analysts and held master's-level degrees in experimental psychology and social work, and one (AF) was employed as a research coordinator and held a bachelor's degree in psychology. All interviewers had training in qualitative methods, previous experience conducting interviews, and prior involvement working with Veterans with SMI at GLA VA.

The development of the interview guide was informed by the Technology Acceptance Model [29] and focused on perceptions of individuals with SMI toward wearable mobile-sensing devices. Interviews focused on the acceptability, feasibility, perceived usefulness, and ease of use of wearable mobile-monitoring devices for this population. In addition, interviews probed how these devices and accompanying phone app could be improved to enhance usability and overall experiences for participants. The interview guide was pilot tested by research staff before data collection and then adjusted to fit the desired length of the interviews.

Interviews lasted between 15 and 30 minutes and were conducted either in person or over the phone in private office spaces with only the interviewer and participant present. All interviews were audio-recorded and then professionally transcribed. Interviews were conducted until saturation was achieved. Saturation was defined as the point at which no new themes emerged from 3 consecutive interviews. To assess saturation, transcripts were reviewed iteratively by the research team to track the emergence of new themes and codes. After three successive interviews produced no additional themes, data collection was discontinued.

Analytic Approach

Rapid qualitative analysis [30] was used to identify prominent findings established to be informative of the perceptions of individuals with SMI toward wearable mobile-sensing devices. A rapid qualitative analysis approach to reviewing interview transcripts was deemed appropriate as the primary goal was to gain overall feedback from Veterans, as well as elicit insights to targeted questions designed to broaden our understanding of the experiences of participants with SMI when using these devices. Four members of the research team (AF, JC, LH, and RC) developed a draft codebook based on thematic analysis, including a priori codes derived from the qualitative interview guide. Through iterative review of transcripts, the research team identified and incorporated emergent themes, refining the codebook to reflect both predefined and data-driven themes [31]. Two team members (AF and JC) then independently coded the same 2 transcripts using Atlas.ti 24 software (Lumivero) and met afterward to compare and discuss the results and modify the codebook based on their findings. Any coding discrepancies were reconciled by consensus. These two authors (AF and JC) then separately coded identical transcripts once more using the refined codebook and met again to compare coding and establish strong intercoder reliability. No discrepancies in coding were found. All remaining interviews were then coded by AF and JC. Findings were organized using emergent themes from interviews.

Ethical Considerations

This research was approved by the Veterans Health Administration's Institutional Review Board (2018‑020118). Written informed consent was obtained from participants upon enrollment into the parent study, and verbal consent was also obtained from participants before the start of each qualitative interview. This study used deidentified data to safeguard participant information. Participants received US $10 in compensation for completing qualitative interviews. Reporting of qualitative research in this manuscript is aligned with guidelines for qualitative research in informatics [32].

Results

Sample

Ages of participants ranged from 34 to 69 years with a mean age of 51.7 (SD 11.9) years. Four of the participants self-identified as women and 11 participants self-identified as men. Five participants were diagnosed with a schizophrenia spectrum disorder, 6 were diagnosed with bipolar disorder, and 4 were diagnosed with PTSD. In addition, 8 participants held diagnoses for 2 qualifying SMIs, 6 with a combination of bipolar

XSL•FO
RenderX

disorder and PTSD, and 2 with a combination of a schizophrenia spectrum disorder and PTSD (Table 1).

Table . Sample characteristics (n=15).

Characteristics	Participants
Sex, n (%)	
Male	11 (73)
Female	4 (46)
Race, n (%)	
Black	8 (53)
Hispanic or Latino	2 (13)
Native American	1 (6)
Non-Hispanic White	4 (26)
Age (years), n (%)	
30 - 49	4 (26)
50 - 64	8 (53)
65+	3 (20)
Diagnosis, n (%)	
Schizophrenia spectrum	5 (33)
Bipolar	6 (40)
PTSD[a]	4 (26)
2 Diagnoses	8 (53)
Education	
Less than high school graduate	1 (6)
High school graduate or general educational development equivalent	2 (13)
Some college	8 (53)
Bachelor's degree	4 (26)
Marital status, n (%)	
Partnered	4 (26)
Not partnered	11 (73)

[a]PTSD: posttraumatic stress disorder.

Emergent Themes and Subthemes of Attitudes Toward Wearable Devices From Qualitative Data

Attitudes Toward Wearable Devices

The interviews show that, out of the 15 participants, none had privacy concerns about using wearable mobile-monitoring devices that accessed their location or collected PGHD. Many of the participants expressed that their lack of concern over privacy issues was due to their belief that the personal health information being collected was common health information. One Veteran shared, "It was basic information, your health information… it's just basic stuff." Several participants also noted that many devices they already use, such as their cell phones, can track their location, so using a wearable device with that capability did not raise any concerns. In fact, many individuals expressed that they already expected the wearable devices to have location tracking capabilities when asked if they would like to participate in this portion of the project. In addition, some participants noted that they felt like they frequented the same locations and did not have reservations about that information being collected. "The data you would collect is my activities and that's not a big secret," stated one Veteran (Table 2).

Most of the participants interviewed stated that they liked the wearable devices, sharing that they were informative about their daily habits, relatively easy to use, and had a long battery life. For example, one Veteran said, "I think it was very nice, especially for a non-computer-literate person like myself. It was very easy to operate and stuff," while another noted, "It gives you the important information."

XSL•FO
RenderX

Table . Representative quotes. Attitudes toward mobile-sensing apps

Themes and subthemes	Quotes
Wearable monitoring benefits	
Sleep	• "My favorite part is the sleep... you can compare it for the week or the month, and you can kind of track like where you've been… and what was helpful..."
Motivation	• "It made me get out of my shell more and become more active. I would say it's increased my quality of life." • "It kinda guides you. Kinda challenges you to do better the following week." • "... It showed me that my activity was increasing, and it motivated me more, to do more."
Physical health	• "My main thing right now is exercisin' and that seems kind of simple but the steps that it counts that's real important to me because I have a goal of how many steps to take."
Ease of passive monitoring	• "You can wear it and not even know that it's there." • "The benefit is that it's intuitive… I don't have to input data. It does it automatically on its own, and that's a convenience."
Mental health tracking	• "I can check in and let the Fitbit know my mood." • "It lets me know that I'm in good range, so that tells my mental health not to worry or have anxiety. So it keeps me in a good mood."
Greater understanding of health	• "If you feel differently that day, that can help you understand what may be causing you to have more mental difficulties that day and help you kinda narrow down what might be triggering you or causing additional difficulties." • "It lets me know how I'm doing in the moment... I can see how I'm doing health-wise."
Device or app improvements	
Physical improvements	• "I have a hard time seeing what's on it because it doesn't light up as well as it does at night..." • "... the material of the watchband itself could cause irritation to some people..." • "I would love a bigger face on it…"
Alerting help	• "It can give a warning sign to whoever... and then they say, okay, dad is in a situation, or my loved one is in a situation. Let's go talk to him and see if we need to talk him off a ledge." • "When you had some type of struggle in your mental health like maybe when you were feeling activated or when your heartrate was too high if it also vibrated then to let you know to kind of pay attention to the emotional side of things."
Increased tracking capabilities	• "It would be nice if I could put a reminder for me takin' my medication."
Greater app interaction	• "Kinda like if it popped up... once an hour or however often instead of you having to remember, oh, I gotta go log this in. Especially if it detects something, if it would give you a notification. Hey, we noticed this."
Wearable monitoring challenges	
Adjusting to wearing device	• "It's been really hard to remember to put it on..."
Lack of technical support	• "I didn't understand it… more technology I need to try to figure out." • "I really didn't realize that the app, the phone app, had all these amenities to it."

XSL•FO
RenderX

Themes and subthemes	Quotes
Difficulty understanding data	• "And I don't quite understand the sleeping portion of it because it's just hard for me to understand it."
Monitoring accuracy unclear	• "… I don't know whether I can depend on what that's saying as far as my heart rate."
Difficulty tracking emotions	• "... I don't think it has the capability to deal with... mental tracking as far as necessarily like mood and things like that." • "It will recognize stress level but not anxiety level." • "I guess the only thing it wouldn't be able to address is the onset of the PTSD[a]."
Attitudes toward wearable devices	
Like wearable device	• "I think it was very nice, especially for a non-computer-literate person like myself. It was very easy to operate and stuff." • "It's interactive, it's very perceptive, and it gives me a ballpoint park of my health measures." • "It gives you the important information." • "It kinda shows you what's going on with your health... it's been a remarkable tool."
Frequency of use - high	• "... for the most part, it's on all the time."
Frequency of use - low	• "I just don't find it very user friendly."
VA[b] continued use	• "I liked it. I recommend it to other guys." • "I'd continue to use it because it keeps me moving... I'm doing good with steps per day, and it shows significant improvement to my health."
Privacy concerns - no	• "... the data you would collect is my activities and that's not a big secret." • "Because it was basic information, your health information… it's just basic stuff."
Attitudes toward mobile-sensing apps	
Like app	• "... I like all the little categories and whatnot. You know? Helps me keep up with what the heck I'm doin' and what's goin on..." • "I actually like it. It helps me keep track of my walking and just sitting down." • "... It's got mindfulness where I can center myself, say, if I'm havin' a bad day. I can go into this app. I can breathe, I can listen to somebody, positive thinking, and then I can come back."
App usability - easy	• "… I love how easy it is to use. The accessibility of it… it's just really convenient."
App usability - difficult	• "It was kinda confusing. I'm still learning half the stuff that I'm reading."
Clinician share	• "I could bring this up at the physical and show it to the doctor, and that way give them some areas to look at... it would help with the physical and general well-being." • "It helps if that information is shared with our doctors or our psychiatrists. Kind of shows them a diagram of our mental health."

[a]PTSD: posttraumatic stress disorder.

[b]VA: Department of Veterans Affairs.

Additionally, many participants shared that they believed the devices offered a comprehensive look at their health and overall well-being. "It's interactive, it's very perceptive, and it gives me a ballpoint park of my health measures," stated one Veteran, with another adding, "It kinda shows you what's going on with your health… it's been a remarkable tool." Some individuals shared that they were wearing and checking the data on their watches daily, "… for the most part, it's on all the time," said

XSL•FO
RenderX

one Veteran. However, two participants noted that it was difficult to remember to put the device on in the mornings, while another shared, “I just don’t find it very user friendly,” both of which led to less frequent use.

The majority of participants agreed that they would continue to use the wearable devices if they were available through the VA. One Veteran shared, “I’d continue to use it because it keeps me moving… I’m doing good with steps per day, and it shows significant improvement to my health,” while another stated, “I liked it. I recommend it to other guys.”

Perceived Benefits of Wearable Devices

Many participants shared that they experienced a variety of benefits from using wearable mobile-sensing devices. All individuals interviewed stated that they were actively using the watches to track some aspect of their physical health, with their heart rate and steps taken per day ranking among the most popular categories. “My main thing right now is exercisin’ and that seems kind of simple but the steps that it counts that’s real important to me because I have a goal of how many steps to take,” expressed one Veteran. Findings also showed that nearly all participants were using the wearable devices to help monitor their sleep to gain information on the quality and duration of rest they were getting each night. One Veteran shared, “My favorite part is the sleep… you can compare it for the week or the month, and you can kind of track like where you’ve been… and what was helpful...”

Numerous individuals stated that using the wearable devices and seeing their PGHD, such as step count and overall activity, motivated them in a variety of ways. For example, several participants shared that after using the devices they felt motivated to do more physical activity (“…It showed me that my activity was increasing, and it motivated me more, to do more.”), leave their homes more often (“It made me get out of my shell more and become more active. I would say it’s increased my quality of life.”) and set healthier goals each week (“It kinda guides you. Kinda challenges you to do better the following week.”). In addition, nearly all Veterans shared that they felt that passive monitoring from wearable devices was a convenient and helpful way to gather their personal health information, sharing thoughts such as, “You can wear it and not even know that it’s there,” and “The benefit is that it’s intuitive… I don’t have to input data. It does it automatically on its own, and that’s a convenience.”

The majority of participants interviewed noted that they also used the watches to track aspects of their mental health, such as their stress and anxiety levels, which were reflected on the device’s accompanying cell phone app. For example, one Veteran expressed, “I can check in and let the Fitbit know my mood,” while another stated, “It lets me know that I’m in good range, so that tells my mental health not to worry or have anxiety. So, it keeps me in a good mood.” In addition, two individuals also shared that they were using the wearable devices to help manage their weight by monitoring the amount of calories burned per day, while 2 female Veterans interviewed stated that the watches allowed them to monitor their menstrual cycle effectively and conveniently.

Finally, most participants expressed that using the wearable devices helped give them a greater understanding of their overall health by allowing them to look over their sleep, stress, and daily activity data monitored by the devices. “If you feel differently that day, that can help you understand what may be causing you to have more mental difficulties that day and help you kinda narrow down what might be triggering you or causing additional difficulties,” stated one Veteran, while another noted, “It lets me know how I’m doing in the moment… I can see how I’m doing health-wise.”

Challenges Experienced With Wearable Devices

During the interviews, participants also mentioned some challenges to using the wearable devices. Nearly half of the individuals interviewed noted difficulties adjusting to wearing the watches due to either the wristband of the device being uncomfortable or difficulty remembering to wear the device. “It’s been really hard to remember to put it on…,” stated one Veteran. Participants shared that these challenges of remembering to wear the device typically occurred when they either took the watch off to charge or to sleep.

In addition, about half of the participants expressed that there were some technical difficulties in learning how to use the wearable device itself, as well as the accompanying phone app, which led to decreased use. “I didn’t understand it… more technology I need to try to figure out,” shared one Veteran. Individuals stated that this lack of comfort with and understanding of the technology particularly limited their use of the partner app, which contained elements such as sleep and stress information, charts of health data collected over time, as well as guided meditations and physical exercises. Many participants interviewed were unable to use these elements due to this barrier. “I really didn’t realize that the app, the phone app, had all these amenities to it,” noted one Veteran.

Some participants stated that they had some difficulty in understanding the significance of the PGHD collected from the wearable devices (“I don’t quite understand the sleeping portion of it because it’s just hard for me to understand it.”), and that further technical support or guidance on how to use the wearable device and the partner app would have been beneficial for their understanding and user experience. Also, many individuals expressed concern over the accuracy of the data collected. For example, some participants questioned the reliability of information (“… I don’t know whether I can depend on what that’s saying as far as my heart rate.”) and pointed out that some activity is unaccounted for when they are not wearing the devices, such as when they take them off to charge or sleep. Furthermore, some participants shared beliefs that wearable devices have limited capabilities when it comes to tracking certain aspects of their mental health, such as their emotions (“… I don’t think it has the capability to deal with… mental tracking as far as necessarily like mood and things like that.”), anxiety levels (“It will recognize stress level but not anxiety level.”), and other symptoms related to their SMI diagnosis, such as the onset of a manic or depressive episode (“I guess the only thing it wouldn’t be able to address is the onset of the PTSD.”).

XSL•FO
RenderX

Suggestions for Improvement of Wearable Devices

Participants also offered several suggestions on how to improve the wearable devices to increase their use and understanding of them, as well as improve their overall user experience. Individuals shared that some physical device improvements, such as adjustable screen brightness ("I have a hard time seeing what's on it because it doesn't light up as well as it does at night..."), a larger device screen ("I would love like a bigger face on it..."), and more comfortable wristbands ("… the material of the watchband itself could cause irritation to some people...") would enhance their experience. About half of the participants expressed a desire for the devices to have increased tracking capabilities. Some suggestions for additional monitoring areas include pain levels, emotions, different forms of stress, and body temperature. In addition, individuals also expressed interest in the devices tracking medication adherence. "It would be nice if I could put a reminder for me takin' my medication," shared one Veteran, while others shared desires in wearables tracking their blood sugar and time spent on social media.

Many of the participants interviewed communicated a desire for increased interaction with the accompanying cell phone app. Some proposed that the app could be improved by providing additional exercise programs, sending medication reminders, and having an emotions log for individuals to check in on their mental health. "Kinda like if it popped up… once an hour or however often instead of you having to remember, oh, I gotta go log this in. Especially if it detects something, if it would give you a notification. Hey, we noticed this," suggested one Veteran.

Additional suggestions included the watches vibrating or sending a phone notification to the individual if their heart rate or stress levels were too high, or even alerting help by sending a message to an individual's clinician or family members under certain circumstances that may indicate a cause for concern. "When you had some type of struggle in your mental health like maybe when you were feeling activated or when your heart rate was too high if it also vibrated then to let you know to kind of pay attention to the emotional side of things," said one Veteran, while another shared, "It can give a warning sign to whoever… and then they say, okay, dad is in a situation, or my loved one is in a situation. Let's go talk to him and see if we need to talk him off a ledge."

Attitudes Toward Mobile-Sensing Apps

Nearly all the participants stated that they liked the wearable's partner cell phone app, many sharing that the app gives them a greater understanding of their overall health than just the wearable device alone, as the face of the watch itself only displays limited PGHD. "I actually like it. It helps me keep track of my walking and just sitting down," stated one Veteran, while another said, "… I like all the little categories and whatnot. You know? Helps me keep up with what the heck I'm doin' and what's goin' on..."

Most of the individuals who expressed liking the app noted that they were primarily using it to look at their step count, heart rate, and other PGHD from the wearable device in a longitudinal format displayed in graphs on the app dashboard. Participants stated that these graphs allowed them to see a visual representation of their health behaviors over the course of several weeks. In addition, two individuals expressed that they were using the mindfulness and physical exercise features included in the app and that they found these elements helpful for both their mental and physical health, with one saying, "… It's got mindfulness where I can center myself, say, if I'm havin' a bad day. I can go into this app. I can breathe, I can listen to somebody, positive thinking, and then I can come back." Twelve participants stated that they had looked at the app at least once during the several weeks they were using the wearable devices, with many checking it daily or weekly. However, some of those individuals shared that they did not frequently open the app and relied more heavily on the watch itself to check their PGHD.

In addition, most participants noted that they felt the app was user-friendly due to the display buttons and categories of health information being simple and easy to understand. "… I love how easy it is to use. The accessibility of it… it's just really convenient," stated one Veteran. However, some participants expressed difficulty using the app. One Veteran noted, "It was kinda confusing. I'm still learning half the stuff that I'm reading," with many others sharing that they felt the app was complicated to operate, that there was a lack of clarity about the meanings of the display categories, that they did not understand how their manually input data, such as daily food and water intake, could be used, and that checking the cell phone app was less convenient than looking at the watch itself.

Additionally, all but one individual interviewed agreed that they would feel comfortable sharing their PGHD with their clinicians to help inform and improve their health care treatment. Several shared thoughts such as, "It helps if that information is shared with our doctors or our psychiatrists. Kind of shows them a diagram of our mental health," and "I could bring this up at the physical and show it to the doctor, and that way give them some areas to look at… it would help with the physical and general well-being," of patient health and treatment.

Discussion

Principal Results

Participants with SMI in this sample generally held positive perceptions toward wearable mobile-monitoring devices and were able to incorporate wearing the health and fitness trackers into their daily lives, many noting that the passive data collection was convenient. Qualitative data showed that participants used the devices to track various aspects of their health, such as sleep, heart rate, stress, and physical activity, and found that devices gave them a greater understanding of their overall health. Some participants experienced technical difficulties when using the wearables and their partner phone app, which may have limited their use of the devices. Additionally, participants suggested that some physical changes to the wearable devices, such as screen brightness and screen size, would increase usability and accessibility of these devices.

Comparison With Prior Work

The majority of participants enjoyed their experience using the mobile-sensing watches and accompanying app, which is consistent with previous research [33,34]. Although some studies

XSL•FO
RenderX

have found privacy concerns about personal health data collection and location tracking to be potential barriers to using mobile-monitoring devices among the SMI population [35], qualitative data revealed that these concerns were not present for Veterans in the study.

The convenience and capability of wearable devices to monitor PGHD may be particularly helpful for individuals with SMI, as members of this population often experience psychological barriers to physical activity and other healthy lifestyle choices, which can have negative impacts on their overall well-being [36]. Interviews revealed that mobile-sensing devices can help address this challenge by increasing participants' levels of motivation to engage in physical activity from week to week upon seeing their PGHD, such as their daily step count, collected by the wearable devices. Individuals also reported that they appreciated the ease and convenience of passive monitoring and that the devices helped increase their understanding of their overall health, suggesting that incorporating these devices into participants' daily lives is both feasible and beneficial.

The qualitative reports also brought to light some challenges of using mobile-monitoring devices. Some physical aspects of the design of the watches, such as the buttons on the device being difficult to operate, became barriers to use for several of the participants. Individuals also shared that wristbands being uncomfortable led to some challenges keeping the devices on for extended periods of time, especially during sleep, and that the lack of brightness of the watch face screen made it difficult for them to see their data when outside during the day. Additionally, differing comfort with using technology made it particularly difficult for several participants to use the device's companion cell phone app, which has previously presented challenges of wearable device use for this population [37]. In future research, more detailed training with participants upon initially receiving the wearable device, as well as ongoing technical support during device use, may be beneficial to ensure individuals are able to access all PGHD and amenities provided by these devices. In addition, using peer support coaches to help with mobile health interventions has been shown to be helpful for Veterans with SMI [38], and these active supports may also be beneficial to increasing the accessibility and understanding of mobile-sensing devices for this population.

Limitations

This study has some potential limitations. In this sample, we interviewed Veterans with SMI receiving treatment at the GLA VA, an urban area. These individuals volunteered to participate in a study involving the use of wearable devices that track their physical activity and location, and therefore, the thoughts and experiences of these participants may differ from those of other individuals who did not consent to participate and use these devices. Secondly, the interview guide was informed by the Technology Acceptance Model, which did not explore differences in participants' thoughts or experiences based on sex, race, or age, and potential differences in these areas did not come up organically during interviews, which limits our understanding of how these characteristics may influence perceptions of wearable devices for individuals with SMI. Further study of wearable devices that explore the influences of sex, race, and age is needed to expand our understanding of the perceptions and uses of wearable devices in this population. In addition, although access to a smartphone was not a barrier for any participants in our sample, and research has shown smartphone use in individuals with SMI to be as high as about 60% [18], this obstacle may be present for some individuals. Fitbits can be activated through a smartphone, tablet, or computer, and are able to collect PGHD, such as steps, heart rate, and sleep information, without needing further access to additional technology. However, to use their full amenities, such as seeing long-term data trends and using GPS, the device needs to be synced to the Fitbit app via smartphone or tablet. Health care facilities, such as the VA, may be able to reduce potential technological accessibility barriers through programs aimed at reducing this digital divide, such as via the Office of Connected Care. Finally, participants' duration using the devices varied from periods of two weeks to about a month. Individuals who had been using the watches for the shorter period had less time to engage with the device and accompanying app, which may have led to decreased PGHD collection and allowed less time for participants to become acquainted with using the device or to experience potential challenges.

Conclusions

Research has increasingly turned to the use of technology to help address barriers of PGHD and communication between patients and providers. Using mobile-monitoring devices and PGHD to improve treatment outcomes for individuals with SMI has been gaining traction, especially as more research uncovers its potential to benefit patients in this population. These devices conveniently collect passive sensor data that can be shared with health care organizations to help fill in gaps on important health trends that patients may experience between medical visits, such as stress levels, physical activity, and sleep quality and duration. Through using these devices, patients could potentially self-monitor their activities and identify changes in their behavior that may be associated with a lapse in mental health. In addition, these devices can provide ample information about various health behaviors that could be used to prevent relapse and modify approaches to patient care. This study indicates that wearable devices should be feasible overall for participants with SMI and that individuals within the population are accepting of these devices. While there may be some unique challenges in this population, incorporating wearable devices into patient care could provide an opportunity to enhance treatment outcomes and prevent relapse. With greater emerging technologies that monitor mental and physical health needs, there is a promising future for the use of wearable devices to support this population.

XSL•FO
RenderX

Funding

This research was supported by the Department of Veterans Affairs, Veterans Health Administration, Health Systems Research, Office of Research and Development (IIR 19-392 and the Department of Veterans Affairs VISN 22 Mental Illness Research, Education, and Clinical Center (MIRECC). The contents of this publication and the views expressed therein do not necessarily represent the views of the Department of Veterans Affairs or affiliated Institutions.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to ethical concerns but are available from the corresponding author on reasonable request.

Authors' Contributions

AY contributed to the conceptualization, methodology, supervision, and funding acquisition of the study. AF, RC, and LH carried out all data collection efforts. AF and JC were responsible for data curation and formal analysis. AF managed project administration and prepared the original draft of the manuscript. All authors contributed to the review and editing of the manuscript.

Conflicts of Interest

None declared.

References

1. National survey on drug use and health 2022. Substance Abuse and Mental Health Services Administration. URL: https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health/national-releases/2022 [accessed 2026-04-30]
2. Chang FH. What are individuals with serious mental illness (SMI) dealing with in a competitive labor market? Two case studies. Work 2015 Jun 5;51(2):301-305. [doi: 10.3233/WOR-141849] [Medline: 24763341]
3. Lin D, Kim H, Wada K, et al. Unemployment, homelessness, and other societal outcomes in patients with schizophrenia: a real-world retrospective cohort study of the United States Veterans Health Administration database: societal burden of schizophrenia among US veterans. BMC Psychiatry 2022 Jul 8;22(1):458. [doi: 10.1186/s12888-022-04022-x] [Medline: 35804314]
4. Steadman HJ, Osher FC, Robbins PC, Case BA, Samuels S. Prevalence of serious mental illness among jail inmates. Psychiatr Serv 2009 Jun;60(6):761-765. [doi: 10.1176/ps.2009.60.6.761] [Medline: 19487344]
5. Ronaldson A, Elton L, Jayakumar S, Jieman A, Halvorsrud K, Bhui K. Severe mental illness and health service utilisation for nonpsychiatric medical disorders: a systematic review and meta-analysis. PLoS Med 2020 Sep;17(9):e1003284. [doi: 10.1371/journal.pmed.1003284] [Medline: 32925912]
6. Cook JA, Burke-Miller JK, Razzano LA, Steigman PJ, Jonikas JA, Santos A. Serious mental illness, other mental health disorders, and outpatient health care as predictors of 30-day readmissions following medical hospitalization. Gen Hosp Psychiatry 2021;70:10-17. [doi: 10.1016/j.genhosppsych.2021.02.004] [Medline: 33639449]
7. Fu XL, Qian Y, Jin XH, et al. Suicide rates among people with serious mental illness: a systematic review and meta-analysis. Psychol Med 2023 Jan;53(2):351-361. [doi: 10.1017/S0033291721001549] [Medline: 33952359]
8. Druss BG, Zhao L, Von Esenwein S, Morrato EH, Marcus SC. Understanding excess mortality in persons with mental illness: 17-year follow up of a nationally representative US survey. Med Care 2011 Jun;49(6):599-604. [doi: 10.1097/MLR.0b013e31820bf86e] [Medline: 21577183]
9. Newcomer JW, Hennekens CH. Severe mental illness and risk of cardiovascular disease. JAMA 2007 Oct 17;298(15):1794-1796. [doi: 10.1001/jama.298.15.1794] [Medline: 17940236]
10. Luciano M, Pompili M, Sartorius N, Fiorillo A. Editorial: mortality of people with severe mental illness: causes and ways of its reduction. Front Psychiatry 2022;13:1009772. [doi: 10.3389/fpsyt.2022.1009772] [Medline: 36061294]
11. Auquier P, Lançon C, Rouillon F, Lader M. Mortality in schizophrenia. Pharmacoepidemiol Drug Saf 2007 Dec;16(12):1308-1312. [doi: 10.1002/pds.1496] [Medline: 17944000]
12. Jin H, Mosweu I. The societal cost of schizophrenia: a systematic review. Pharmacoeconomics 2017 Jan;35(1):25-42. [doi: 10.1007/s40273-016-0444-6] [Medline: 27557994]
13. Kessler RC, Aguilar-Gaxiola S, Alonso J, et al. The global burden of mental disorders: an update from the WHO World Mental Health (WMH) surveys. Epidemiol Psichiatr Soc 2009;18(1):23-33. [doi: 10.1017/s1121189x00001421] [Medline: 19378696]
14. Maidment I, Sud D, Chew-Graham C. Challenge of optimising medication in people with severe mental illness. BMJ Qual Saf 2022 May;31(5):337-339. [doi: 10.1136/bmjqs-2021-013847] [Medline: 34556544]
15. Wiesepape CN, Queller Soza SE, Faith LA. Behind the gaps: a narrative review of healthcare barriers for individuals with serious mental illness. Healthcare (Basel) 2025 Sep 23;13(19):2387. [doi: 10.3390/healthcare13192387] [Medline: 41095473]

XSL•FO
RenderX

16. Ben-Zeev D, Davis KE, Kaiser S, Krzsos I, Drake RE. Mobile technologies among people with serious mental illness: opportunities for future services. Adm Policy Ment Health 2013 Jul;40(4):340-343. [doi: 10.1007/s10488-012-0424-x] [Medline: 22648635]
17. Firth JA, Cotter J, Torous J, Bucci S, Firth JA, Yung AR. Mobile phone ownership and endorsement of "mHealth" among people with psychosis: a meta-analysis of cross-sectional studies. Schizophr Bull 2016 Mar;42(2):448-455. [doi: 10.1093/schbul/sbv132] [Medline: 26400871]
18. Young AS, Cohen AN, Niv N, et al. Mobile phone and smartphone use by people with serious mental illness. Psychiatr Serv 2020 Mar 1;71(3):280-283. [doi: 10.1176/appi.ps.201900203] [Medline: 31744429]
19. Ben-Zeev D, Brenner CJ, Begale M, Duffecy J, Mohr DC, Mueser KT. Feasibility, acceptability, and preliminary efficacy of a smartphone intervention for schizophrenia. Schizophr Bull 2014 Nov;40(6):1244-1253. [doi: 10.1093/schbul/sbu033] [Medline: 24609454]
20. Alvarez-Jimenez M, Alcazar-Corcoles MA, González-Blanch C, Bendall S, McGorry PD, Gleeson JF. Online, social media and mobile technologies for psychosis treatment: a systematic review on novel user-led interventions. Schizophr Res 2014 Jun;156(1):96-106. [doi: 10.1016/j.schres.2014.03.021] [Medline: 24746468]
21. Zulman DM, Jenchura EC, Cohen DM, Lewis ET, Houston TK, Asch SM. How can eHealth technology address challenges related to multimorbidity? perspectives from patients with multiple chronic conditions. J Gen Intern Med 2015 Aug;30(8):1063-1070. [doi: 10.1007/s11606-015-3222-9] [Medline: 25691239]
22. Lahti AC, Wang D, Pei H, Baker S, Narayan VA. Clinical utility of wearable sensors and patient-reported surveys in patients with schizophrenia: noninterventional, observational study. JMIR Ment Health 2021 Aug 9;8(8):e26234. [doi: 10.2196/26234] [Medline: 34383682]
23. Reeve S, Sheaves B, Freeman D. The role of sleep dysfunction in the occurrence of delusions and hallucinations: a systematic review. Clin Psychol Rev 2015 Dec;42:96-115. [doi: 10.1016/j.cpr.2015.09.001] [Medline: 26407540]
24. Birchwood M, Spencer E, McGovern D. Schizophrenia: early warning signs. Adv psychiatr treat 2000 Mar;6(2):93-101. [doi: 10.1192/apt.6.2.93]
25. Wu DTY, Xin C, Bindhu S, et al. Clinician perspectives and design implications in using patient-generated health data to improve mental health practices: mixed methods study. JMIR Form Res 2020 Aug 7;4(8):e18123. [doi: 10.2196/18123] [Medline: 32763884]
26. Skoufalos A, N'Dri LA, Waters D. Leveraging digital medicine to support providers and their patients in managing serious mental illness. Popul Health Manag 2021 Aug;24(S2):S55-S61. [doi: 10.1089/pop.2021.0083] [Medline: 34382864]
27. Medich M, Cannedy SL, Hoffmann LC, et al. Clinician and patient perspectives on the use of passive mobile monitoring and self-tracking for patients with serious mental illness: user-centered approach. JMIR Hum Factors 2023 Oct 24;10:e46909. [doi: 10.2196/46909] [Medline: 37874639]
28. Young AS, Choi A, Cannedy S, et al. Passive mobile self-tracking of mental health by veterans with serious mental illness: protocol for a user-centered design and prospective cohort study. JMIR Res Protoc 2022 Aug 5;11(8):e39010. [doi: 10.2196/39010] [Medline: 35930336]
29. Ammenwerth E. Technology acceptance models in health informatics: TAM and UTAUT. Stud Health Technol Inform 2019 Jul 30;263:64-71. [doi: 10.3233/SHTI190111] [Medline: 31411153]
30. Hamilton AB, Finley EP. Qualitative methods in implementation research: an introduction. Psychiatry Res 2019 Oct;280:112516. [doi: 10.1016/j.psychres.2019.112516] [Medline: 31437661]
31. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
32. Ancker JS, Benda NC, Reddy M, Unertl KM, Veinot T. Guidance for publishing qualitative research in informatics. J Am Med Inform Assoc 2021 Nov 25;28(12):2743-2748. [doi: 10.1093/jamia/ocab195] [Medline: 34537840]
33. Naslund JA, Aschbrenner KA, Barre LK, Bartels SJ. Feasibility of popular m-health technologies for activity tracking among individuals with serious mental illness. Telemed J E Health 2015 Mar;21(3):213-216. [doi: 10.1089/tmj.2014.0105] [Medline: 25536190]
34. Ben-Zeev D, Wang R, Abdullah S, et al. Mobile behavioral sensing for outpatients and inpatients with schizophrenia. Psychiatr Serv 2016 May 1;67(5):558-561. [doi: 10.1176/appi.ps.201500130] [Medline: 26695497]
35. Byrne S, Tohamy A, Kotze B, et al. Using a mobile health device to monitor physiological stress for serious mental illness: a qualitative analysis of patient and clinician-related acceptability. Psychiatr Rehabil J 2022 Sep;45(3):219-225. [doi: 10.1037/prj0000514] [Medline: 35298226]
36. Brooke LE, Gucciardi DF, Ntoumanis N, Lin A. Qualitative investigation of perceived barriers to and enablers of sport participation for young people with first episode psychosis. Early Interv Psychiatry 2020 Jun;14(3):293-306. [doi: 10.1111/eip.12854] [Medline: 31270957]
37. Naslund JA, Aschbrenner KA, Bartels SJ. Wearable devices and smartphones for activity tracking among people with serious mental illness. Ment Health Phys Act 2016 Mar;10:10-17. [doi: 10.1016/j.mhpa.2016.02.001] [Medline: 27134654]
38. Young AS, Cohen AN, Goldberg R, et al. Improving weight in people with serious mental illness: the effectiveness of computerized services with peer coaches. J Gen Intern Med 2017 Apr;32(Suppl 1):48-55. [doi: 10.1007/s11606-016-3963-0] [Medline: 28271427]

XSL•FO
RenderX

Abbreviations

GLA VA: Greater Los Angeles Department of Veterans Affairs
PGHD: Patient-generated health data
PTSD: posttraumatic stress disorder
SMI: serious mental illness
VA: Department of Veterans Affairs

Edited by A Stone; submitted 01.Oct.2025; peer-reviewed by L Yang, M Stacy; revised version received 21.Mar.2026; accepted 24.Mar.2026; published 15.May.2026.

Please cite as:
Freitas AM, Chavez JG, Chinchilla M, Calderon R, Chassman S, Hoffmann L, Young AS
Assessing the Use of Wearable Mobile-Monitoring Devices Among Individuals With Serious Mental Illness: Qualitative Acceptability and Feasibility Study
J Med Internet Res 2026;28:e85087
URL: https://www.jmir.org/2026/1/e85087
doi:10.2196/85087

© Aubrey M Freitas, Jesus G Chavez, Melissa Chinchilla, Ronald Calderon, Stephanie Chassman, Lauren Hoffmann, Alexander S Young. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Efficacy of Digital Speech Therapy for Poststroke Dysarthria: Randomized Noninferiority Trial

Yuyoung Kim[1,2], PhD; Minjung Kim[1,2], MS; Saebyeol Kim[1,2], MS; Jinwoo Kim[1,2], PhD; Joon-Ho Shin[3], MD, MS; Yoonkyung Chang[4], MD, PhD; Ji Young Na[5], PhD; JungWan Kim[6], PhD; Tae-Jin Song[7,8], MD, PhD

[1]HAII Corp., Seoul, Republic of Korea

[2]HCI Lab, Yonsei University, Seoul, Republic of Korea

[3]Department of Neurorehabilitation, National Rehabilitation Center, Seoul, Republic of Korea

[4]Department of Neurology, Ewha Womans University College of Medicine, Mokdong Hospital, Seoul, Republic of Korea

[5]Department of Communication Disorders, Korea Nazarene University, Cheonan, Republic of Korea

[6]Department of Speech and Language Pathology, Daegu University, Gyeongsan, Republic of Korea

[7]Department of Neurology, Ewha Womans University Seoul Hospital, Ewha Womans University College of Medicine, Seoul Hospital, 260 Gonghang-daero, Gangseo-gu, Seoul, Republic of Korea

[8]Graduate Program in System Health Science and Engineering, Ewha Womans University, Seoul, Republic of Korea

Corresponding Author:
Tae-Jin Song, MD, PhD
Department of Neurology, Ewha Womans University Seoul Hospital, Ewha Womans University College of Medicine, Seoul Hospital, 260 Gonghang-daero, Gangseo-gu, Seoul, Republic of Korea

Abstract

Background: Poststroke dysarthria, a common speech impairment, affects up to half of all stroke survivors, often reducing their ability to communicate, and adversely affecting their quality of life. Although conventional speech therapy for poststroke dysarthria is effective, access is often limited by time and geographical constraints. Here, digital speech therapy may serve as a remotely deliverable alternative for selected patients. However, few trials have assessed its efficacy, safety, and usability.

Objective: This study aimed to evaluate whether a smartphone-based speech therapy app is noninferior to conventional workbook-based therapy in improving speech intelligibility among individuals with poststroke dysarthria.

Methods: This single-blind, randomized controlled, noninferiority trial was performed at 3 hospitals in South Korea. Adults (≥19 y) with poststroke dysarthria who were cognitively intact, without aphasia, and able to use a smartphone were eligible. Participants were enrolled between July 20, 2023, and April 15, 2024. Participants were randomly assigned (1:1), stratified by stroke phase, using a block randomization method, to receive either a smartphone-based digital therapy app or a conventional workbook-based therapy for 4 weeks. The primary outcome was speech intelligibility (0 - 100 perceptual rating) after the intervention. Primary analysis was intention-to-treat using analysis of covariance. A noninferiority margin of 19 points was pre-defined.

Results: A total of 73 participants were enrolled (median age 62.00 years). Among them, 38 were assigned to the digital speech therapy group and 35 to the control group. Intelligibility scores improved from 80.48 (SD 18.92) to 92.08 (SD 12.38) in the intervention group, and from 80.94 (SD 16.74) to 88.11 (SD 18.06) in the control group. The adjusted between-group difference was 4.49 (95% CI 0.61-8.37), and the lower bound of the 95% CI was above the prespecified noninferiority margin (–19), which supported noninferiority. No significant between-group differences were observed in the secondary outcomes related to speech function or psychological status. The system usability score was 89.6, and adherence in the digital speech therapy group was 64.6% based on app logs, with no treatment-related adverse events.

Conclusions: Digital speech therapy was noninferior to conventional workbook-based therapy in improving speech intelligibility and was feasible across acute to early subacute and chronic stroke phases in cognitively intact stroke survivors with predominantly mild-to-moderate dysarthria. However, feasibility and efficacy in older stroke survivors with cognitive deficits or co-occurring aphasia, or in those unable to use smartphones, remain to be established.

Trial Registration: : ClinicalTrials.gov NCT05865106; https://clinicaltrials.gov/study/NCT05865106

International Registered Report Identifier (IRRID): RR2-10.3389/fneur.2024.1305297

(J Med Internet Res 2026;28:e81938) doi:10.2196/81938

XSL•FO
RenderX

KEYWORDS

stroke; dysarthria; speech therapy; mobile health app; mHealth; randomized clinical trial

Introduction

Stroke is a leading cause of death and long-term disability worldwide [1]. Approximately 22% - 60% of stroke survivors present with motor speech disorders, including poststroke dysarthria. Additionally, approximately 35% of patients with poststroke dysarthria continue to experience dysarthria after 6 months [2-5]. These findings indicate that poststroke dysarthria is a common and persistent complication of stroke, typically characterized by minimal long-term recovery. Poststroke dysarthria can involve impairments in articulation, phonation, prosody, resonance, and respiration [3]. Moreover, poststroke dysarthria may compromise speech intelligibility and potentially have a detrimental impact on communication, emotional well-being, and the overall quality of life [6,7].

Behavioral speech therapy is the standard treatment for poststroke dysarthria. It involves structured exercises designed to strengthen respiratory and articulatory muscles, improve motor speech control, and enhance intelligibility through targeted modulation of pitch, loudness, and articulation [3,8]. Despite its proven benefits, conventional face-to-face speech therapy remains difficult to access because of practical limitations. Its repetitive and monotonous nature contributes to poor long-term adherence [9]. In addition, access to trained speech-language pathologists (SLPs) is often limited, and patients with poststroke dysarthria experience challenges visiting hospitals or speech therapy clinics multiple times per week [10,11]. Therefore, these patients often face an unmet need for access to adequate speech therapy.

Digital speech therapy has recently emerged as a promising alternative to address the limitations of conventional treatments. Mobile health apps and computer-based programs enable the delivery of personalized speech therapy outside of hospitals. These digital speech therapies allow patients to engage in high-frequency and intensive practice with reduced need for real-time clinician involvement [12,13]. Previous studies suggest that telerehabilitation and digital platforms can reduce the burden and cost of face-to-face speech therapy [14,15]. They also help maintain continuity of care after hospital discharge and encourage patients to practice independently [14-16]. These advantages make digital speech therapy particularly beneficial for stroke survivors who have difficulty accessing conventional speech therapy owing to geographic, financial, or pandemic-related reasons.

Previous small-sample studies have suggested potential benefits of digital speech therapy for individuals with poststroke dysarthria in the chronic phase [17-19]. However, limited sample sizes and a lack of rigorous evaluation imply that the efficacy and safety of these interventions remain unclear. Consequently, we performed a multicenter randomized controlled trial to determine whether smartphone-based digital therapy is noninferior to conventional workbook-based therapy in improving speech intelligibility in patients with poststroke dysarthria. We also evaluated secondary outcomes, including speech function, patient-reported psychological well-being, usability, and adherence, to assess the broader therapeutic impact and overall feasibility.

Methods

Study Design

This multicenter, prospective, evaluator-blinded, randomized controlled trial used a noninferiority design to compare a smartphone-based digital speech therapy app with conventional workbook-based therapy in patients with poststroke dysarthria. The study was conducted at 3 stroke centers in South Korea, including Ewha Womans University Seoul Hospital, Ewha Womans University Mokdong Hospital, and the National Rehabilitation Center in Seoul.

The trial was registered at ClinicalTrials.gov (NCT05865106). The study protocol outlined the intervention development, outcome measure selection, and a previously published statistical analysis plan [20].

Participants

Participants were recruited between July 20, 2023 and April 15, 2024, across the 3 participating sites in South Korea. Stroke specialists screened patients in stroke units and outpatient clinics at each of the 3 hospitals. The inclusion criteria were as follows: (1) adults aged 19 years or older; (2) diagnosis of dysarthria caused by stroke, confirmed by a stroke specialist; (3) dysarthria severity corresponding to a score of 1 or higher on the dysarthria item of the National Institutes of Health Stroke Scale [21], reflecting either mild-to-moderate dysarthria, where speech is slurred but mostly intelligible, or severe dysarthria, where speech is mostly unintelligible or absent, and aphasia had to be absent; (4) a neurologically stable condition with no evidence of acute deterioration; (5) first-ever stroke with no prior history of stroke; (6) mini-mental state examination score of 26 or higher, which was used to confirm the absence of cognitive impairment in the participants [22]; (7) sufficient visual acuity to read therapy materials [23]; (8) adequate hearing to follow spoken instructions [24]; and (9) sufficient motor skills and basic communication ability to use a smartphone. The participants had to be able to press buttons and understand simple operations on the device.

The exclusion criteria were as follows: (1) presence of aphasia or other language disorders; (2) other neurological conditions that could affect speech, such as Parkinson disease, Parkinsonism, or motor neuron disease; (3) major psychiatric illness, which included schizophrenia, untreated major depression, or a history of substance abuse that might interfere with participation; (4) unwillingness or inability to use a smartphone or the speech therapy app; (5) inability to read or understand task instructions; and (6) inability to speak or understand Korean, as the intervention materials were written and spoken in Korean. Biological sex was recorded as male or female based on medical records at enrollment.

XSL•FO
RenderX

All participants provided written informed consent before enrollment in the study. They were informed that they could withdraw from the study at any time without affecting their standard medical care. The research team could also withdraw participants under certain conditions. These included completion of less than 40% of the assigned therapy sessions, emergence of urgent medical problems, adverse events related to the intervention, noncompliance with study procedures, or clinical deterioration that made continued participation unsafe.

Randomization and Masking

Participants were randomly assigned in a 1:1 ratio to either the digital speech therapy or control group. Randomization was stratified according to the stroke phase to ensure that each group included a balanced number of participants in the acute to early subacute and chronic phases. The acute to early subacute phase was defined as less than 6 months after stroke onset. The chronic phase was defined as 6 months or more after stroke onset [25].

Randomization sequences for each stratum were generated using a computer-based algorithm in R software (R Foundation for Statistical Computing). Block sizes of 2 and 4 were used to ensure allocation concealment and unpredictability [26]. An independent researcher who was not involved in the trial prepared the randomization schedule and managed the group assignment process. The allocation was concealed until all baseline assessments were completed. Each participant was assigned a sequential study identification number at enrollment, after which group assignment was performed based on the next entry in the corresponding stratified randomization list. The participants and treating clinicians were informed of the group assignment only after the baseline evaluation was completed.

Owing to the nature of the interventions, blinding of the participants and treating clinicians was not feasible. Therefore, all speech recordings were centrally collected after trial completion and anonymized. The filenames and metadata contained no group, site, or time point identifiers and were presented in a computer-generated random order to independent, trained SLP assessors. These assessors had no role in the treatment, had no contact with the participants, and were blinded to treatment assignment, study site, and assessment timepoint. This approach ensured objective evaluation of the primary outcomes despite the open-label design of the trial [27].

Procedures

Following randomization, the therapy phase began within one week. Participants in both groups engaged in their assigned therapy program for 4 weeks. Sessions were performed 5 days per week, with each session lasting approximately 60 minutes. A trained research coordinator provided a standardized orientation session to explain how to use the assigned therapy tools, either the mobile app or printed workbook. All the participants were instructed to complete the therapy independently. Caregiver assistance was permitted, if necessary, in both the digital speech therapy and workbook-based control groups.

During the intervention period, the study staff contacted the participants 3 times per week via phone or SMS text messaging. These check-ins were intended to monitor progress, encourage continued participation, and resolve any difficulties related to the exercises, such as technical problems or misunderstandings. In addition to these contacts, no in-person therapy or direct clinical input was provided. Participants completed all exercises independently at home without the direct involvement of a clinician.

Participants completed outcome assessments at baseline and after completing the 4-week intervention. All assessments were performed in person during scheduled hospital visits using a standardized smartphone-based app.

Intervention Group: Smartphone-Based Speech Therapy App

Participants assigned to the intervention group received therapy through a smartphone-based speech therapy app developed by our team in collaboration with clinical experts. The app was specifically designed for individuals with poststroke dysarthria and incorporated the established principles of behavioral speech therapy [3,28].

At the start of the intervention, an app was installed on the smartphone of each participant (Android or iOS). A research coordinator performed an in-person orientation session to guide the participants in using the app, and a printed instruction manual was also provided. The participants were instructed to complete 60 minutes of therapy daily, 5 days per week, for 4 weeks. The participants were permitted to divide the session into smaller time blocks depending on their preference and fatigue level.

Digital speech therapy was individualized based on the baseline assessments. Recorded speech samples of the participants were uploaded to a secure server after enrollment. A licensed SLP reviewed these samples and performed an auditory-perceptual evaluation targeting 5 key subsystems typically affected in dysarthria, including respiration, phonation, resonance, articulation, and prosody [3]. Based on this assessment, the SLP created a personalized therapy plan. This plan prescribed the initial starting level and weekly goals. The app implemented the SLP's prescription by delivering the specified exercise sets. The app did not automatically adapt difficulty in real time. Adjustments were made by the SLP during weekly remote check-ins.

The app included 15 structured exercises organized into the following four therapeutic categories: (1) warm-up tasks using video-guided instructions to establish correct posture and breathing patterns; (2) phonation and prosody tasks, such as sustained vowel production and modulation of pitch and loudness, with real-time visual feedback to enhance vocal control; (3) oro-motor and vocal strength exercises that included tongue, lip, and jaw movements and vocal fold training via video-guided routines to strengthen articulatory and phonatory coordination; and (4) articulation exercises involving repetition of words and sentences, as well as aloud reading tasks that emphasized clarity, speech rate, and prosodic control. Specifically, the app visualized pitch variation and loudness in real time, and at the end of each task, it provided a summary indicating how closely the performance matched the preset target range.

XSL•FO
RenderX

The app automatically recorded usage data, including task completion, number of attempts, and exercise duration. Adherence was defined as the percentage of completed sessions out of the 20 prescribed sessions (5 sessions per week for 4 weeks). Individual adherence scores were derived by dividing the number of sessions completed by the 20 prescribed sessions and expressing the result as a percentage (0% - 100%). Participants could view their progress on a dashboard within the app. The research team monitored the data remotely using a secure web-based portal. When adherence dropped below expectations, study staff, independent of treatment and outcome assessment, proactively contacted participants to provide encouragement or technical assistance. This system supported consistent speech therapy delivery without in-person clinician supervision while maintaining engagement throughout the intervention. A detailed description of digital speech therapy for poststroke dysarthria is provided in the Multimedia Appendix 1.

Control Group: Conventional Speech Therapy Workbook

Participants assigned to the control group received a conventional paper-based speech-therapy workbook. The workbook was based on standard clinical practices and evidence-based protocols for poststroke dysarthria and included structured exercises used in conventional therapy.

The workbook was organized to reflect the general content and progression of the app to ensure structural comparability with digital speech therapy, which included similar types of exercises and daily dosages. It featured activities targeting key speech subsystems, such as syllable and sentence repetition, sustained vowel phonation, volume and pitch control, and expressive reading tasks involving prosodic variation. The workbook followed a standardized stepwise sequence from easier to more challenging exercises. Participants proceeded in order and were instructed to complete the same daily and weekly dose as in the intervention group. The overall level of difficulty and therapeutic intensity matched those of the digital therapy program. The workbook provided static written instructions, and participants were instructed to self-monitor their performance without real-time feedback; no therapeutic performance feedback was delivered during the home practice sessions.

Before beginning the program, a licensed SLP provided standardized instructions on using the workbook. Participants were instructed to complete 60 minutes of therapy per day, 5 days per week, for 4 weeks. A printed log sheet was provided to track session completion and to note any challenges encountered.

To promote adherence and ensure treatment fidelity, a study coordinator contacted participants 3 times per week over phone or SMS text messaging. During these check-ins, the coordinator encouraged practice and reviewed self-reported task completion. Adherence was monitored using participants' self-reported log entries and verbal confirmation during follow-up contacts. Consistent with the intervention group, adherence was calculated as the percentage of self-reported sessions completed relative to the 20 prescribed sessions.

The control group participants were offered access to digital speech therapy for poststroke dysarthria after completing the study to ensure ethical equity. Those who wanted to participate were granted access to the same 4-week digital speech therapy program. A detailed description of the workbook-based speech therapy protocol is provided in the Multimedia Appendix 1.

Outcomes

To ensure comparability, all efficacy outcome instruments, recording workflows, and rater procedures were identical for both groups. Secondary and patient-reported efficacy measures were also administered using the same protocols in both groups.

The primary outcome was speech intelligibility, defined as the extent to which a listener could accurately understand the utterances of the speaker [29]. Participants read the Gaeul Passage, a standardized Korean text comprising 369 syllables with a balanced phonemic distribution [30]. Speech samples were recorded using the study app and securely stored on a remote server. Three independent, licensed SLPs with ≥3 years of experience in adult poststroke dysarthria assessment served as a fixed evaluation panel. Each rater evaluated every participant recording once and provided an independent perceptual speech intelligibility rating on a 0 - 100 scale. The participant's final speech intelligibility score was the mean of the 3 rater scores [31]. The interrater reliability was assessed using the intraclass correlation coefficient (ICC) based on a 2-way random-effects model with absolute agreement to ensure the appropriateness of using an average score. The average-measures ICC was 0.85 (95% CI 0.74-0.91).

Secondary outcomes included measures of speech function and psychological well-being. Speech function was assessed using the following measures: (1) maximum phonation time (MPT), defined as the duration for which participants could sustain the vowel /a/ [32]; (2) oral diadochokinesis, measured by the number of correctly repeated syllables (/pa/, /ta/, /ka/, /pataka/) within a 5-second interval and normalized to syllables per second [33]; and (3) percentage of consonants correct (PCC), evaluated using the urimal test of articulation and phonology-2, in which participants read 30 Korean words and their productions were transcribed and compared with the target forms [34,35]. To ensure objectivity and eliminate rater bias, MPT and diadochokinesis were automatically calculated by the digital platform's software for both groups. PCC was evaluated by a licensed SLP blinded to both treatment allocation and assessment time points.

Psychological well-being was assessed using the Patient Health Questionnaire-9 (PHQ-9), a 9-item self-report scale for depressive symptoms [36], with lower scores indicating reduced symptom severity. The quality of life of dysarthric speakers (QoL-Dys) comprises 4 domains (speech characteristics, situational difficulty, compensatory strategies, and perceived reactions of others); lower scores indicate improved quality of life with less functional impact from dysarthria. The total QoL-Dys score was used for analysis [37,38].

Usability was evaluated using the System Usability Scale (SUS), a 10-item self-report questionnaire that assesses perceived ease of use and user satisfaction [39]. This measure was administered

XSL•FO
RenderX

only to participants in the intervention group at the end of the 4-week trial. Safety was monitored through weekly check-ins with all participants, and any adverse events were recorded by the study team.

Statistical Analysis

The sample size estimation determined that 76 participants (38 per group) would provide >80% power to assess the noninferiority of the primary outcome. This calculation assumed an SD of 24.9, a noninferiority margin of 19 points on the 100-point speech intelligibility scale, and a one-sided α of .025. A noninferiority margin of 19 points on the 0 - 100 speech intelligibility scale was defined. This margin was based on clinical judgment regarding what would constitute an acceptable difference in functional communication in routine home practice. This rationale reflects the established clinical perspective that speech intelligibility scores should be interpreted within the context of functional communication rather than as isolated numerical values [40]. A 19-point noninferiority margin, anchored in a functional communication framework, was applied to the task-based primary outcome derived from standardized passage reading (Gaeul Passage). Recent estimates of the minimal clinically important difference for intelligibility (approximately 7 - 15 points) were considered as contextual benchmarks [41]. Given the limited availability of stroke-specific guidance for defining an intelligibility-based margin, the prespecification was anchored to a functional-communication framework and informed by methodological recommendations that noninferiority margins should reflect clinical acceptability and the objective of the trial [42]. Practical considerations regarding remote, self-guided delivery motivated the use of a noninferiority design; however, the noninferiority margin was determined based on clinical acceptability.

All analyses were performed using IBM SPSS Statistics version 29.0 (IBM Corp). Interim analyses were not performed. The Shapiro-Wilk test was used to assess the normality of continuous variables. As most variables were nonnormally distributed, baseline comparisons used the Mann-Whitney *U*-test for continuous variables and the χ^2 test or Fisher exact test for categorical variables. Descriptive statistics are presented as medians with IQR or counts with percentages.

Outcomes were analyzed using analysis of covariance (ANCOVA), with baseline scores and age entered as covariates. Given the baseline age imbalance, we performed exploratory analyses to verify the directionality and influence of the age adjustment. Details are provided in Table S3 in Multimedia Appendix 2. The model assumptions, including residual normality, were verified and satisfied. Adjusted mean differences (intervention minus control), 95% CIs, *P* values, and partial eta-squared values were reported. Noninferiority was concluded when the 95% CI lay entirely above the margin of –19 points.

The primary analysis was conducted on the intention-to-treat (ITT) population, which included all randomized participants. To handle missing postintervention data, the last observation carried forward method was applied (baseline carried forward). In addition, a complete-case sensitivity analysis excluding participants with missing postintervention outcomes was performed. Additionally, a per-protocol (PP) analysis was conducted on a set comprising participants who adhered to the protocol (≥70% adherence). Because adherence assessment differed between groups (objective app logs vs self-reported checklists), PP classification, particularly in the control arm, should be interpreted cautiously.

To assess robustness, further sensitivity analyses were conducted to evaluate the treatment effect under different assumptions. These included: (1) a subgroup analysis by stroke phase (acute to early subacute vs chronic) with an interaction test; (2) an ANCOVA that additionally adjusted for stroke phase as a covariate; (3) a ranked ANCOVA (Quade test) to account for nonnormality in baseline scores; and (4) a logit-transformed ANCOVA to model intelligibility on its original bounded scale.

This study did not include a data monitoring committee, as it was a short-term behavioral intervention trial with minimal risk.

Ethical Considerations

The protocol was approved by the Institutional Review Boards of all participating centers (approval numbers: EUMC 2023-02-002 for Ewha Womans University Seoul Hospital and Ewha Womans University Mokdong Hospital; NRC-2023-01-007 for the National Rehabilitation Center), and the study was conducted in accordance with the Declaration of Helsinki [43]. All participants provided written informed consent before enrollment. Participants received KRW 25,000 (approximately US $19.20) per visit. The study included 2 visits, for a total of KRW 50,000 (the conversion was based on an exchange rate of US $1=KRW 1300 at the time of the study). Study data were deidentified before analysis, and no identifiable participant data were included. Audio files were anonymized and stored on secure servers, with access restricted to authorized researchers.

Results

User Statistics

Of 90 patients screened for eligibility, 17 were excluded for the following reasons: 5 did not meet the inclusion criteria (including n=3 with uncorrected visual impairment and n=2 unable to use smartphone technology), 9 declined to participate, and 3 were excluded due to relocation or inability to attend follow-up visits.

A total of 73 participants were enrolled and randomized, with 38 assigned to the smartphone-based digital therapy group and 35 assigned to the conventional workbook-based therapy group. Two participants (one from each group) discontinued the intervention. Consequently, the primary ITT analysis included all 73 randomized participants, and the PP population comprised 46 participants. These participants were included in the final analysis. Figure 1 shows the trial profile.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram.

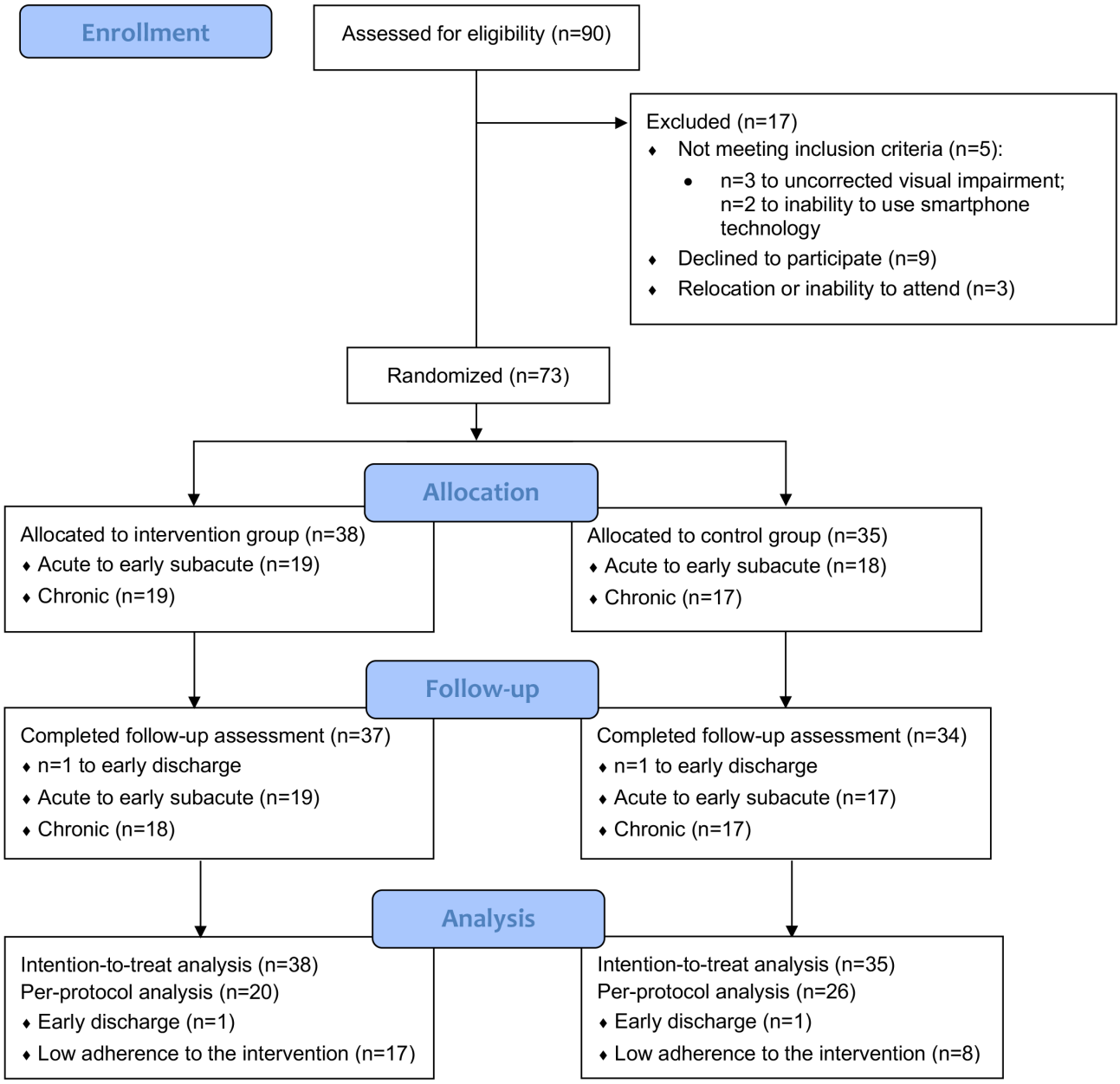

The baseline characteristics are summarized in Table 1. Baseline speech intelligibility was high (median 89.00%), and most participants were classified as having mild-to-moderate dysarthria (66/73, 90%). The baseline characteristics were generally comparable between the groups, although statistically significant between-group differences were observed for age (P=.03) and baseline diadochokinesis /pataka/ rate (P=.02).

XSL•FO
RenderX

Table . Baseline demographic and clinical characteristics of the participants.

Participant characteristics	Total (n=73)	Intervention (n=38)	Control (n=35)	*P* value
Sex, n (%)				.83
Male	53 (73)	28 (74)	25 (71)	
Female	20 (27)	10 (26)	10 (29)	
Age (years), median (IQR)	62.00 (50.50 - 70.50)	59.50 (48.75 - 67.25)	64.00 (55.50 - 75.50)	.03
Phase of stroke, n (%)				.90
Acute to early subacute	37 (51)	19 (50)	18 (51)	
Chronic	36 (49)	19 (50)	17 (49)	
Duration after stroke (days), median (IQR)	66.00 (11.50 - 796.50)	128.50 (12.00 - 606.75)	66.00 (11.50 - 956.50)	.67
Acute to early subacute	12.00 (7.50 - 22.50)	12.00 (7.00 - 19.00)	11.50 (8.00 - 39.75)	.57
Chronic	796.50 (368.75 - 2139.00)	636.00 (343.00 - 1875.00)	975.00 (725.50 - 2388.50)	.32
mRS[a], median (IQR)	1.00 (1.00 - 2.00)	2.00 (1.00 - 2.00)	1.00 (1.00 - 2.00)	.31
Total NIHSS[b], median (IQR)	3.00 (2.00 - 5.00)	4.00 (3.00 - 6.00)	3.00 (2.00 - 5.00)	.15
NIHSS[b]_dysarthria, median (IQR)	1.00 (1.00 - 2.00)	1.00 (1.00 - 2.00)	1.00 (1.00 - 2.00)	.09
Dysarthria severity, n (%)				.11
Mild to moderate	66 (90)	32 (84)	34 (97)	
Severe	7 (10)	6 (16)	1 (3)	
Type of stroke, n (%)				.53
Cerebral infarction	65 (89)	33 (87)	32 (91)	
Cerebral hemorrhage	8 (11)	5 (13)	3 (9)	
Speech intelligibility, median (IQR)	89.00 (77.67 - 90.67)	89.33 (77.67 - 92.17)	89.00 (76.67 - 90.00)	.26
PHQ-9[c], median (IQR)	5.00 (3.00 - 9.00)	5.00 (2.00 - 9.00)	5.00 (3.00 - 9.00)	.52
QoL-Dys[d], median (IQR)	63.00 (47.50 - 81.00)	63.00 (50.50 - 74.50)	65.00 (46.00 - 89.00)	.55
MPT[e] /a/, median (IQR)	6.60 (3.47 - 10.70)	6.61 (3.73 - 9.21)	6.59 (2.88 - 10.81)	.69
PCC[f] (%), median (IQR)	98.00 (94.00 - 100.00)	99.00 (94.00 - 100.00)	98.00 (94.00 - 99.00)	.07
Diadochokinesis[g] /pa/, median (IQR)	4.22 (3.25 - 5.10)	4.65 (3.49 - 5.26)	3.83 (3.05 - 4.82)	.08
Diadochokinesis[g] /ta/, median (IQR)	4.52 (3.31 - 5.31)	4.71 (3.33 - 5.47)	4.21 (3.23 - 4.86)	.19
Diadochokinesis[g] /ka/, median (IQR)	3.93 (2.74 - 4.80)	3.95 (3.18 - 4.73)	3.81 (2.66 - 4.83)	.59
Diadochokinesis[g] /pataka/, median (IQR)	1.30 (0.97 - 1.64)	1.43 (1.07 - 1.83)	1.18 (0.93 - 1.51)	.02

[a]mRS: modified Rankin scale.

[b]NIHSS: National Institutes of Health Stroke Scale.

[c]PHQ-9: Patient Health Questionnaire-9.

[d]QoL-Dys: Quality of Life in the Dysarthric Speaker.

[e]MPT: maximum phonation time.

[f]PCC: percentage of consonants correct.

[g]Diadochokinesis rate reported as syllables per second over a 5-second interval.

XSL•FO
RenderX

Evaluation Outcomes

The primary outcome results are summarized in Table 2 and Figure 2A. Speech intelligibility improved in both groups after the intervention. In the intervention group, the mean score increased from 80.48 (SD 18.92) to 92.08 (SD 12.38), whereas in the control group, it increased from 80.94 (SD 16.74) to 88.11 (SD 18.06). In the ANCOVA model adjusting for baseline intelligibility and age, the adjusted mean difference in postintervention intelligibility was 4.49 (95% CI 0.61-8.37). Because the lower bound of the 95% CI exceeded the prespecified noninferiority margin of –19, noninferiority was concluded. A complete-case sensitivity analysis excluding the 2 participants with missing postintervention outcomes (n=71) yielded a similar adjusted mean difference of 4.41 (95% CI 0.62-8.20), consistent with the primary analysis. The mean within-group improvement in the control group was 7.17 points, corresponding to the lower threshold of the minimal clinically important difference range [41].

Table . Adjusted speech intelligibility outcomes, per protocol analysis and subgroup analyses based on analysis of covariance (ANCOVA).

Group	Baseline, mean (SD)	Postintervention, mean (SD)	Estimated difference, MD[a] (95% CI)	F[b] (df)	P value	η^2 [c]
Overall			4.49 (0.61 to 8.37)	5.32 (1, 67)	.02	0.07
Intervention	80.48 (18.92)	92.08 (12.38)				
Control	80.94 (16.74)	88.11 (18.06)				
Acute to early subacute			3.79 (–2.14 to 9.72)	1.69 (1, 32)	.20	0.05
Intervention	81.35 (20.89)	94.71 (10.05)				
Control	83.37 (13.84)	91.45 (18.43)				
Chronic			4.74 (–0.24 to 9.73)	3.76 (1, 31)	.06	0.11
Intervention	78.61 (17.27)	89.45 (14.51)				
Control	78.37 (19.45)	84.57 (17.51)				
PP analysis[d]			3.57 (–1.79 to 8.92)	1.80 (1, 42)	.19	0.04
Intervention	78.03 (21.01)	91.10 (13.91)				
Control	77.99 (18.58)	87.01 (20.83)				

[a]MD: mean difference.

[b]F: F statistic

[c]η^2: partial eta square (effect size estimate)

[d]PP: per-protocol (≥70% adherence)

Figure 2. Speech intelligibility outcomes. (A) Speech intelligibility scores at baseline and postintervention for the intervention and control groups across all participants (score range: 0 - 100); (B) Speech intelligibility scores for participants in the acute to early subacute stroke phase; (C). Speech intelligibility scores for participants in the chronic stroke phase; error bars represent standard errors.

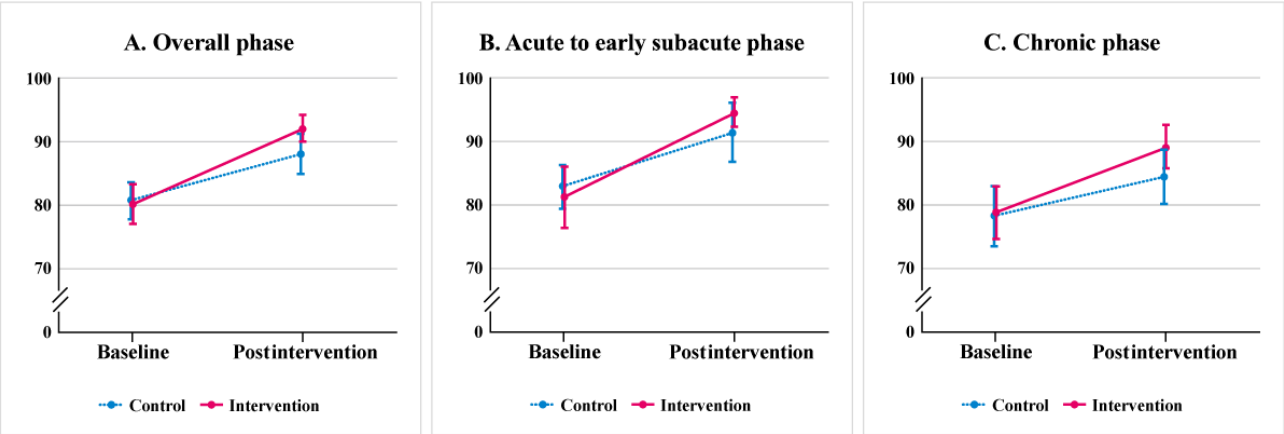

In the per-protocol set (n=46), defined as participants with ≥70% adherence, the adjusted mean difference in postintervention intelligibility was 3.57 (95% CI –1.79 to 8.92). Adherence was assessed using app usage logs in the intervention arm and self-reported checklists in the control arm. The lower bound of the confidence interval also remained above the non-inferiority margin, consistent with the primary noninferiority finding. Figure 3 summarizes these noninferiority results.

XSL•FO
RenderX

Figure 3. Noninferiority analysis of the primary outcome adjusted mean differences (intervention minus control) with 95% CIs are shown for the intention-to-treat (ITT), per-protocol (PP), and ITT subgroup analyses by stroke phase. The dashed vertical line indicates the prespecified noninferiority margin (–19), and the solid vertical line indicates no difference. ITT: intention-to-treat.

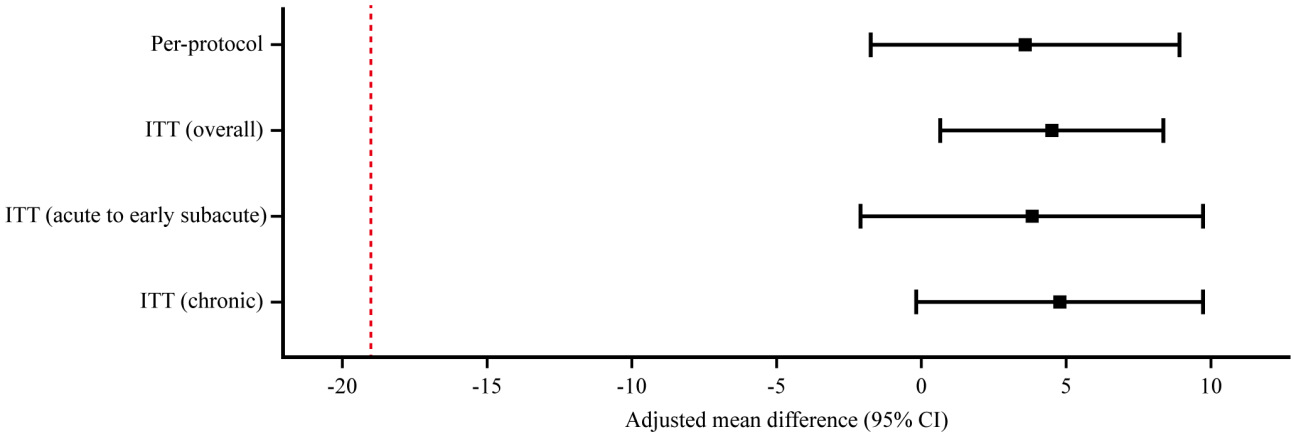

The subgroup and sensitivity analyses supported the robustness of the primary noninferiority conclusion. Subgroup analyses stratified by stroke phase yielded similar adjusted mean differences in both strataFigure 2B and 2C. Among participants in the acute to early subacute phase, the adjusted mean difference in intelligibility scores was 3.79 (95% CI –2.14 to 9.72). In the chronic group, the difference was 4.74 (95% CI –0.24 to 9.73). In both strata, the lower bounds of the 95% CIs remained above the prespecified noninferiority margin (–19). The treatment-by-phase interaction test showed no evidence of differential effects by stroke phase. A sensitivity analysis including stroke phase as a covariate yielded a similar adjusted difference of 4.51 (95% CI 0.71-8.31), which supported the consistency of the primary noninferiority finding. We also performed a ranked ANCOVA (Quade's test) to address nonnormality in baseline distributions, which yielded inferences consistent with the primary analysis. Finally, a logit-transformed ANCOVA was conducted to account for the bounded 0% - 100% outcome; the adjusted group effect was of similar magnitude and direction (estimated mean difference 0.52 on the logit scale; 95% CI –0.01 to 1.05), with inference consistent with the primary analysis. Full numerical results, which include exploratory analyses examining age adjustment (Table S3) and assay sensitivity and robustness checks (Table S4), are provided in Multimedia Appendix 2.

Secondary outcomes showed no statistically significant differences between the groups in speech production, psychological status, or QoL-Dys measures. A detailed summary of the secondary outcomes is presented in Table 3. For speech function, the adjusted between-group difference in MPT was 0.82 (95% CI –1.26 to 2.91; *P*=.43), whereas in PCC it was 0.14 (95% CI –1.82 to 2.10; *P*=.88). No significant group differences were observed in the diadochokinesis syllable repetition rates for /pa/, /ta/, /ka/, or /pataka/. Patient-reported psychological outcomes also exhibited no significant differences between the groups. The adjusted difference in PHQ-9 scores was –1.02 (95% CI –3.16 to 1.13; *P*=.35). Similarly, total scores on the QoL-Dys scale did not differ significantly between groups (–3.04; 95% CI –14.72 to 8.65; *P*=.61). No significant differences were observed in any of the 4 QoL-Dys subdomains.

XSL•FO
RenderX

Table . Secondary outcome measures.

Group	Baseline, mean (SD)	Postintervention, mean (SD)	Estimated difference, MD[a] (95% CI)	F[b] (df)	P value	η^2 [c]
PHQ-9[d]			−1.02 (−3.16 to 1.13)	0.89 (1, 67)	.35	0.01
Intervention	5.89 (4.49)	5.29 (3.77)				
Control	6.77 (5.11)	6.66 (5.64)				
QoL-Dys[e] total			−3.04 (−14.72 to 8.65)	0.27 (1, 67)	.61	0.00
Intervention	61.92 (21.23)	57.97 (23.08)				
Control	63.97 (30.25)	63.43 (30.89)				
Speech characteristics			−0.33 (−3.44 to 2.77)	0.05 (1, 67)	.83	0.00
Intervention	17.61 (6.08)	16.42 (6.73)				
Control	19.20 (8.54)	17.51 (7.76)				
Situational difficulty			−2.02 (−5.47 to 1.43)	1.37 (1, 67)	.25	0.02
Intervention	16.29 (7.85)	14.24 (7.16)				
Control	17.69 (10.46)	17.34 (9.03)				
Compensatory strategies			0.31 (−3.18 to 3.81)	0.03 (1, 67)	.86	0.00
Intervention	17.34 (6.14)	16.42 (5.58)				
Control	15.63 (8.19)	15.91 (8.73)				
Perceived reactions of others			−1.54 (−4.02 to 0.94)	1.53 (1, 67)	.22	0.02
Intervention	10.68 (5.80)	10.89 (7.54)				
Control	11.46 (8.37)	12.66 (7.79)				
PCC[f] (%)			0.14 (−1.82 to 2.10)	0.01 (1, 67)	.88	0.00
Intervention	96.18 (7.64)	96.61 (6.95)				
Control	95.43 (5.40)	95.97 (6.11)				
MPT[g]			0.82 (−1.26 to 2.91)	0.62 (1, 65)	.43	0.01
Intervention	7.95 (5.71)	9.12 (5.33)				
Control	7.06 (4.31)	7.52 (5.30)				
Diadochokinesis /pa/[h]			0.11 (−0.37 to 0.59)	0.19 (1, 65)	.66	0.00
Intervention	4.28 (1.44)	4.62 (1.08)				
Control	3.82 (1.33)	4.20 (1.33)				
Diadochokinesis /ta/[h]			0.39 (−0.03 to 0.80)	3.40 (1, 65)	.07	0.05
Intervention	4.39 (1.32)	4.81 (1.28)				
Control	4.02 (1.31)	4.22 (1.09)				
Diadochokinesis /ka/[h]			0.20 (−0.20 to 0.60)	1.01 (1, 65)	.32	0.01
Intervention	3.83 (1.11)	4.07 (1.24)				
Control	3.70 (1.30)	3.76 (1.24)				

XSL•FO
RenderX

Group	Baseline, mean (SD)	Postintervention, mean (SD)	Estimated difference, MD[a] (95% CI)	F[b] (df)	P value	η^2 [c]
Diadochokinesis /pataka/[h]			0.03 (−0.15 to 0.21)	0.12 (1, 65)	.74	0.00
Intervention	1.44 (0.45)	1.61 (0.52)				
Control	1.17 (0.41)	1.35 (0.47)				

[a]MD: mean difference.

[b]F: F statistic

[c]η^2: partial eta square (effect size estimate).

[d]PHQ-9: Patient Health Questionnaire-9.

[e]QoL-Dys: Quality of Life in Dysarthria.

[f]PCC: percentage of consonants correct.

[g]MPT: maximum phonation time.

[h]Diadochokinesis: rate reported as syllables per second over a 5-second interval.

The usability of the digital application was only assessed in the intervention group. The mean usability score was 89.57 (SD 10.70). In the intervention group, participants completed a mean of 64.6% of the prescribed sessions according to objective application logs. The control group reported a mean adherence of 74.3% via self-reported checklists. In both arms, these rates represent the group average of individual completion percentages relative to the 20 prescribed sessions. Because adherence was assessed using different methods across groups (objective logs vs self-report), these adherence rates are reported descriptively within each arm and should not be directly compared.

Safety and Adverse Events

No serious adverse events related to either the digital therapy or the conventional workbook were reported in either group. Regarding nonserious events, a few participants in the intervention group experienced minor technical issues, such as login difficulties or brief application errors. These issues were promptly resolved by the research team and did not interfere with continued intervention use. No other intervention-related adverse events, such as vocal strain, fatigue, or headache, were reported by participants in either group.

Discussion

Principal Findings

This randomized controlled trial evaluated whether a smartphone-based speech therapy app could achieve therapeutic outcomes comparable to those of conventional workbook-based therapy in patients with poststroke dysarthria. Our findings suggest that remotely delivered, self-guided speech therapy may be a feasible alternative to the workbook-based home program used in this trial, with remote monitoring and minimal clinician contact. The primary ITT analysis met the prespecified noninferiority criterion, and this finding was supported by the PP analysis and sensitivity analyses. The robustness of the primary results was further supported by a nonparametric analysis accounting for nonnormal baseline distributions, which yielded consistent directional findings. Although the point estimate was in the direction of benefit for the intervention, the trial was designed and powered for noninferiority; therefore, we interpret the findings within the prespecified noninferiority framework and consider any suggestion of superiority as exploratory. Interpretation of noninferiority also depends on assay sensitivity, which assumes that the active control would be expected to demonstrate efficacy in the trial setting. As highlighted in the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) E10 guideline, if the comparator effect is attenuated, noninferiority findings may be more difficult to interpret. In the present sample, the mean improvement in the workbook-based control arm was 7.17 points, which aligns with the lower bound of the cited minimal clinically important difference range [41]; this should be considered when interpreting the strength of the noninferiority inference [44].

The effectiveness of speech therapy for poststroke dysarthria may vary depending on the stroke phase. In the early phase, patients are believed to benefit more from restorative approaches that leverage the heightened neuroplasticity shortly after injury [45,46]. By contrast, those in the chronic phase typically rely on compensatory strategies to optimize their residual abilities [47]. However, most previous clinical trials have focused on a single phase of stroke, most commonly the chronic phase. For instance, clinician-delivered programs such as Lee Silverman Voice Treatment and the “Be Clear” protocol have demonstrated improvements in articulatory precision, loudness, and speech clarity but were limited to participants in the chronic phase [18,19]. In contrast, our previous pilot study focused exclusively on patients in the acute to early subacute phases and revealed improvements in speech intelligibility with the use of a digital speech therapy application [48].

Given that most previous studies have targeted a single stroke phase, the applicability of digital speech therapy across different stroke phases has not yet been well established. By including individuals in the acute to early subacute and chronic phases, this trial suggests the potential to broaden the evidence base for its use across the various stroke phases.

Stratified analyses revealed no statistically significant interaction between the treatment group and stroke phase, indicating no

XSL•FO
RenderX

evidence of differential treatment effects by phase. This is of particular interest given the widely cited hypothesis that the early poststroke phase represents a sensitive period of heightened neuroplasticity and responsiveness to therapy [49,50]. Contrary to this expectation, our data revealed that participants in the chronic phase exhibited numerically greater gains in speech intelligibility with digital speech therapy than the control group, whereas the acute to early subacute phase exhibited a smaller difference. Although this trend did not reach statistical significance, it may reflect differences in engagement or responsiveness, which warrant further exploration.

Several factors could explain this pattern. Patients in the chronic phase may have a greater awareness of their communication limitations and stronger motivation to improve them. Long-term adaptation to dysarthria can foster clearer personal goals and greater readiness to engage in structured therapy [51]. Moreover, the chronic phase may offer a more neurologically stable condition that enables consistent training and facilitates the acquisition of compensatory strategies. In contrast, individuals in the early phases may still manage co-occurring impairments and undergo spontaneous recovery, making it more difficult to concentrate on speech-specific interventions. Taken together, these findings suggest that although the stroke phase is an important consideration, patient-level factors, such as motivation, engagement, and adherence, may also contribute to therapeutic outcomes, particularly in the chronic phase. This interpretation aligns with previous studies indicating that, even in the chronic phase, individuals can achieve functional gains when provided with structured and accessible interventions [18,29]. These observations suggest that future studies should consider both stroke phase and individual characteristics when evaluating response to therapy. Effective interventions should consider both the stroke phase and individual characteristics, such as motivation, baseline severity, and access to support resources [52]. Tailored strategies may help optimize recovery regardless of whether the therapy is delivered during the early or chronic phase. This interpretation was further reinforced by a covariate-adjusted analysis, which confirmed the consistency of the treatment effects regardless of the stroke phase. Additionally, since the acute to early subacute subgroup was enrolled very early after stroke onset (median 12 d), spontaneous neurological recovery likely contributed to a portion of the improvements observed in both the intervention and control groups. At this early stage, rapid natural recovery may have attenuated between-group differences and masked treatment-specific effects, making it difficult to disentangle the intervention's effect from natural recovery. Therefore, gains in this stratum should be interpreted cautiously and not attributed solely to the digital or workbook-based therapy.

The secondary outcomes exhibited no statistically significant differences between the intervention and control groups. Measures related to phonation (MPT), articulation (PCC and diadochokinesis), and patient-reported psychological outcomes (PHQ-9 and QoL-Dys) exhibited similar postintervention scores in both groups without meaningful between-group differences. The intervention group exhibited modestly higher mean values in some domains; for instance, a 1.05-second longer MPT. However, these differences did not reach statistical significance. These findings are in line with those of previous studies reporting that digital and conventional speech therapies may offer similar benefits in improving speech-motor control and psychosocial functioning when delivered with adequate structure and intensity [51].

Participants in the digital speech therapy group rated the usability of the app highly (mean SUS score of 89.57). On average, they completed 64.6% of the prescribed sessions based on system usage logs. The observed 64.6% adherence rate in the intervention group is consistent with findings from other self-managed rehabilitation programs. For context, completion rates in stroke rehabilitation studies range from approximately 61% to 100%, and broader exercise adherence reviews commonly report mid-range rates of 60% - 80% [53,54]. In the control group, adherence was monitored using self-report checklists and follow-up contacts. Because adherence in the control arm was based on self-report, these records should be interpreted descriptively and should not be directly compared with objectively logged adherence in the digital arm. While reasons for nonadherence were not systematically collected, informally reported barriers included participant fatigue, environmental constraints (eg, the need for a quiet space for vocal practice), and challenges with digital literacy, particularly among older adults. The high usability ratings and observed engagement in the digital arm were in line with those of previous studies on digital health in neurology and rehabilitation [14]. Features such as real-time visual feedback, simple navigation, and self-paced home practice were well-received. These findings echo earlier reports indicating that stroke survivors appreciate the autonomy and flexibility offered by well-designed mobile therapy tools [14,15,55]. Collectively, these results underscore the feasibility of digital therapy within the selected cohort enrolled in this trial, which included older participants who were neurologically stable, cognitively intact, and able to use a smartphone. Thus, this modality may offer a promising option for selected patients with limited access to in-person treatment. Nevertheless, the applicability of these findings is limited to the selected cohort enrolled in this trial.

Limitations

Despite these promising findings, this study had several limitations. Although the sample size was sufficient to assess noninferiority of the primary outcome, it may have been underpowered to detect smaller differences in secondary outcomes or to fully explore subgroup effects. In addition, the trial randomized 73 participants, which was slightly below the planned sample size of 76, and this minor shortfall should be considered when interpreting the statistical power of the study.

The generalizability of our findings is limited by the sample characteristics. We enrolled cognitively intact participants without aphasia or other language disorders, required basic smartphone proficiency, and restricted the trial to native Korean speakers. While these criteria ensured feasibility and task compliance, they may limit external validity for individuals with cognitive or linguistic impairments, older adults with low digital literacy, or non-Korean speakers. Crucially, the feasibility and efficacy of this digital intervention remain unproven in stroke survivors with cognitive deficits, co-occurring aphasia,

XSL•FO
RenderX

or significant multimorbidity, as the eligibility criteria excluded these populations or required a level of digital proficiency that may not be present in the broader stroke population. Future studies should include more diverse populations and culturally/linguistically adapted versions of the intervention to enhance applicability across settings.

The study assessed treatment outcomes only immediately after the 4-week intervention, and no follow-up evaluation was conducted to examine sustained efficacy. Consequently, the long-term maintenance of treatment gains remains uncertain. It is unclear whether the observed improvements would persist without continued use of the program or periodic reinforcement. Future research should include extended follow-up periods to evaluate the durability of effects and to determine optimal dosing schedules or maintenance strategies for long-term benefit.

Because this was a multicenter trial, we standardized the comparator condition across sites by compiling commonly used conventional practice components into this workbook-based home program to ensure protocol consistency. However, this format may not have fully captured clinically relevant elements of routine face-to-face speech therapy, particularly SLP-delivered feedback and individualized real-time cueing. Future trials could consider combining standardized in-person therapy with structured workbook-based home practice in the control condition to strengthen the comparator further while preserving protocol consistency across centers. Although both groups received comparable study contact, the standardized workbook progression without performance-based adjustment may have resulted in less individualized dosing than the intervention arm.

Adherence in the control group was assessed using self-report checklists, which may be subject to reporting variability. Accordingly, we did not perform a formal between-group comparison of adherence or explore adherence as a moderator of treatment effects. This difference in adherence monitoring introduced measurement uncertainty and limited the ability to verify the comparability of therapeutic effort between groups. Notably, using self-reported data to define the PP population in the control arm may have led to an overestimation of engagement relative to the digital arm. Consequently, readers should exercise caution when interpreting adherence-based analyses, and the noninferiority findings should be considered in the context of this potential reporting bias in the conventional therapy group. Future studies should implement more structured or objective adherence tracking methods for nondigital arms, such as time-stamped photo diaries or scripted phone check-ins with specific queries about session time.

There was also a difference in personalization between groups. The intervention group followed an SLP-prescribed plan, whereas the control group used a standardized stepwise workbook program with a fixed progression and no performance-based advancement rules. Although both groups received comparable monitoring and guidance, this difference in personalization may have favored the intervention group. Future studies should include a workbook arm with SLP-tailored prescriptions to balance personalization across modalities. In addition, usability was assessed only in the intervention group to evaluate the technical feasibility of the app, a core trial objective. Since the SUS is designed specifically for digital interfaces, it was not applied to the paper-based control. This difference prevented a direct comparison of satisfaction between modalities, which limited conclusions on relative clinical acceptability. Future research should use generalized measures to enable more robust comparisons across different treatment formats.

Our study had limitations regarding measurement sensitivity. The baseline intelligibility score was high (median 89.00%). While supplementary analysis suggested that pronounced ceiling compression was uncommon (eg, one participant scored above 95%; Figure S5 in Multimedia Appendix 2), the high baseline level may have constrained measurable improvement. This may have reduced the sensitivity of the primary outcome measure to detect change and made between-group differences harder to detect. In addition, because the primary outcome was derived from a standardized passage reading, the ratings may partly reflect articulatory precision and familiarity with the passage rather than functional intelligibility in everyday communication. Consequently, a limitation exists in that our noninferiority margin, although anchored in a functional communication framework, was applied to a more structural, task-based measure. Therefore, the 19-point margin may carry different clinical weight in this standardized context compared with spontaneous or conversational speech, and readers should be cautious in extrapolating these findings directly to real-world communicative success. This limitation should be considered when interpreting the clinical implications of the noninferiority margin and the primary outcome [56,57]. Using the same passage at baseline and postintervention may also introduce practice effects despite the 4-week interval. Familiarity with standardized assessment materials among trained raters may have contributed to higher scores. Future studies should strengthen measurement sensitivity by incorporating spontaneous speech or conversation-based evaluations, using multiple assessment materials, and including multiple raters. Accordingly, the present findings most directly apply to cognitively intact stroke survivors with mild-to-moderate dysarthria, and additional studies are needed to establish efficacy in patients with more severe impairment.

Although the baseline characteristics were generally well-balanced between the groups, a statistically significant difference in age was observed. Older age is often associated with slower motor recovery and reduced neuroplasticity, which may introduce confounding factors into the interpretation of treatment effects. Notably, the intervention group was younger and thus potentially more responsive to therapy. This imbalance should be considered when interpreting the comparative effectiveness of the interventions.

Conclusions

This randomized controlled trial evaluated whether a smartphone-based speech therapy app was noninferior to conventional workbook-based therapy in improving speech intelligibility in cognitively intact adults with poststroke dysarthria. Digital speech therapy achieved a level of improvement comparable to that of the control group, which

XSL•FO
RenderX

met the predefined noninferiority margin. These findings support the clinical potential of remotely delivered self-guided speech therapy for stroke survivors with mild-to-moderate dysarthria, while the modest mean improvement observed in the workbook-based control arm should be considered when interpreting the strength of the noninferiority inference. Further evaluation is needed for patients with more severe impairment. Future studies should assess the long-term sustainability of treatment effects, explore integration into routine clinical workflows, and evaluate the effectiveness across various languages and health care systems.

Acknowledgments

We thank the clinical coordinators and speech-language pathologists at the participating centers for their assistance. During the preparation of this manuscript, the authors utilized a generative AI language model minimally for assistance with English language editing and grammar correction. The authors reviewed, edited, and took full responsibility for the final content.

Funding

This work was supported by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (grants HI22C073600 and RS-2023- 00262087), by the Institute of Information & Communications Technology Planning & Evaluation (IITP), funded by the Ministry of Science and ICT, Republic of Korea (grant 2022-0-00621), and supported by the BK21 FOUR (Fostering Outstanding Universities for Research) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF-5199990614253, Education Research Center for 4IR-Based Health Care).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: YK, JHS, TJS
Methodology: YK, JYN, JWK
Software: YK, MK, SK
Formal analysis: YK
Data curation: YK
Funding acquisition: JK, TJS
Investigation: JHS, YC, JYN, JWK, TJS
Project administration: YK, MK, SK
Resources: MK, SK, JHS, YC, TJS
Supervision: JK, JHS, TJS
Writing – original draft: YK
Writing – review & editing: MK, SK, JK, TJS

Conflicts of Interest

YK, MK, and SK are employees of Haii Corp, the developer of the smartphone-based speech therapy app evaluated in this study, and they were involved in the development of the app. Jinwoo Kim is the Chief Executive Officer of Haii Corp. JHS, YC, JYN, JWK, and TJS declare no competing interests.

Multimedia Appendix 1
Details of smartphone-based and workbook-based therapy.
[DOCX File, 1332 KB - jmir_v28i1e81938_app1.docx]

Multimedia Appendix 2
Supplementary results including detailed outcome measures and additional analyses.
[DOCX File, 241 KB - jmir_v28i1e81938_app2.docx]

Checklist 1
CONSORT-EHEALTH (v 1.6.1) checklist.
[PDF File, 1435 KB - jmir_v28i1e81938_app3.pdf]

References

XSL•FO
RenderX

1. Feigin VL, Brainin M, Norrving B, et al. World Stroke Organization (WSO): Global Stroke Fact Sheet 2022. Int J Stroke 2022 Jan;17(1):18-29. [doi: 10.1177/17474930211065917] [Medline: 34986727]
2. Darley FL, Aronson AE, Brown JR. Differential diagnostic patterns of dysarthria. J Speech Hear Res 1969 Jun;12(2):246-269. [doi: 10.1044/jshr.1202.246] [Medline: 5808852]
3. Duffy JR. Motor Speech Disorders: Substrates, Differential Diagnosis, and Management, 4th edition: Elsevier; 2020.
4. Mackenzie C. Dysarthria in stroke: a narrative review of its description and the outcome of intervention. Int J Speech Lang Pathol 2011 Apr;13(2):125-136. [doi: 10.3109/17549507.2011.524940] [Medline: 21480809]
5. O'Mahony P, Thomson R, Dobson R, Rodgers H, James O. The prevalence of stroke and associated disability. J Public Health (Oxf) 1999 Jun 1;21(2):166-171. [doi: 10.1093/pubmed/21.2.166]
6. Dickson S, Barbour RS, Brady M, Clark AM, Paton G. Patients' experiences of disruptions associated with post‐stroke dysarthria. Intl J Lang & Comm Disor 2008 Mar 4;43(2):135-153. [doi: 10.1080/13682820701862228]
7. Mackenzie C, Kelly S, Paton G, Brady M, Muir M. The Living with Dysarthria group for post-stroke dysarthria: the participant voice. Int J Lang Commun Disord 2013;48(4):402-420 [FREE Full text] [doi: 10.1111/1460-6984.12017] [Medline: 23889836]
8. Brady MC, Clark AM, Dickson S, Paton G, Barbour RS. Dysarthria following stroke: the patient's perspective on management and rehabilitation. Clin Rehabil 2011 Oct;25(10):935-952. [doi: 10.1177/0269215511405079] [Medline: 21729975]
9. Tamayo-Serrano P, Garbaya S, Bouakaz S, Blazevic P. A game-based rehabilitation therapy for post-stroke patients: an approach for improving patient motivation and engagement. IEEE Syst Man Cybern Mag 2020;6(4):54-62. [doi: 10.1109/MSMC.2020.3002519]
10. Avan A, Digaleh H, Di Napoli M, et al. Socioeconomic status and stroke incidence, prevalence, mortality, and worldwide burden: an ecological analysis from the Global Burden of Disease Study 2017. BMC Med 2019 Oct 24;17(1):191. [doi: 10.1186/s12916-019-1397-3] [Medline: 31647003]
11. Dobkin BH, Dorsch A. New evidence for therapies in stroke rehabilitation. Curr Atheroscler Rep 2013 Jun;15(6):331. [doi: 10.1007/s11883-013-0331-y] [Medline: 23591673]
12. Griffin M, Bentley J, Shanks J, Wood C. The effectiveness of Lee Silverman Voice Treatment therapy issued interactively through an iPad device: a non-inferiority study. J Telemed Telecare 2018 Apr;24(3):209-215. [doi: 10.1177/1357633X17691865] [Medline: 28147896]
13. Karlsen C, Ludvigsen MS, Moe CE, Haraldstad K, Thygesen E. Experiences of community-dwelling older adults with the use of telecare in home care services: a qualitative systematic review. JBI Database System Rev Implement Rep 2017 Dec;15(12):2913-2980. [doi: 10.11124/JBISRIR-2017-003345] [Medline: 29219874]
14. Abelson JS, Kaufman E, Symer M, Peters A, Charlson M, Yeo H. Barriers and benefits to using mobile health technology after operation: a qualitative study. Surgery 2017 Sep;162(3):605-611. [doi: 10.1016/j.surg.2017.05.007] [Medline: 28651777]
15. Bowser DM, Shepard DS, Nandakumar A, et al. Cost effectiveness of mobile health for antenatal care and facility births in Nigeria. Ann Glob Health 2018;84(4):592. [doi: 10.29024/aogh.2364]
16. Zhou X, Snoswell CL, Harding LE, et al. The role of telehealth in reducing the mental health burden from COVID-19. Telemed J E Health 2020 Apr;26(4):377-379. [doi: 10.1089/tmj.2020.0068] [Medline: 32202977]
17. Mitchell C, Bowen A, Tyson S, Butterfint Z, Conroy P. Interventions for dysarthria due to stroke and other adult-acquired, non-progressive brain injury. Cochrane Database Syst Rev 2017 Jan 25;1(1):CD002088. [doi: 10.1002/14651858.CD002088.pub3] [Medline: 28121021]
18. Mahler LA, Ramig LO. Intensive treatment of dysarthria secondary to stroke. Clin Linguist Phon 2012 Aug;26(8):681-694. [doi: 10.3109/02699206.2012.696173] [Medline: 22774928]
19. Park S, Theodoros D, Finch E, Cardell E. Be clear: a new intensive speech treatment for adults with nonprogressive dysarthria. Am J Speech Lang Pathol 2016 Feb;25(1):97-110. [doi: 10.1044/2015_AJSLP-14-0113] [Medline: 26882004]
20. Kim Y, Kim M, Kim J, Song TJ. Efficacy and feasibility of a digital speech therapy for post-stroke dysarthria: protocol for a randomized controlled trial. Front Neurol 2024;15:1305297. [doi: 10.3389/fneur.2024.1305297] [Medline: 38356882]
21. Oh MS, Yu KH, Lee JH, et al. Validity and reliability of a Korean version of the national institutes of health stroke scale. J Clin Neurol 2012 Sep;8(3):177-183. [doi: 10.3988/jcn.2012.8.3.177] [Medline: 23091526]
22. KIM JM, SHIN IS, YOON JS, LEE HY. Comparison of diagnostic validities between MMSE-K and K-MMSE for screening of dementia. Journal of Korean Neuropsychiatric Association 2003;42(1):124-130 [FREE Full text]
23. Rowe F, Brand D, Jackson CA, et al. Visual impairment following stroke: do stroke patients require vision assessment? Age Ageing 2009 Mar;38(2):188-193. [doi: 10.1093/ageing/afn230] [Medline: 19029069]
24. Isaacson JE, Vora NM. Differential diagnosis and treatment of hearing loss. Am Fam Physician 2003 Sep 15;68(6):1125-1132. [Medline: 14524400]
25. Bernhardt J, Hayward KS, Kwakkel G, et al. Agreed definitions and a shared vision for new standards in stroke recovery research: the stroke recovery and rehabilitation roundtable taskforce. Int J Stroke 2017 Jul;12(5):444-450. [doi: 10.1177/1747493017711816] [Medline: 28697708]
26. Broglio K. Randomization in clinical trials: permuted blocks and stratification. JAMA 2018 Jun 5;319(21):2223-2224. [doi: 10.1001/jama.2018.6360] [Medline: 29872845]

XSL•FO
RenderX

27. Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. The Lancet 2002 Feb;359(9307):696-700. [doi: 10.1016/S0140-6736(02)07816-9]
28. Godecke E, Armstrong E, Rai T, et al. A randomized control trial of intensive aphasia therapy after acute stroke: the Very Early Rehabilitation for SpEech (VERSE) study. Int J Stroke 2021 Jul;16(5):556-572. [doi: 10.1177/1747493020961926] [Medline: 33019888]
29. Yorkston KM, Dowden PA, Beukelman DR. Intelligibility measurement as a tool in the clinical management of dysarthric speakers. In: Kent RD, editor. Intelligibility in Speech Disorders: Theory, Measurement and Management: John Benjamins Publishing Company; 1992, Vol. 1:265-285. [doi: 10.1075/sspcl.1.08yor]
30. Kim HH. Perceptual, acoustical, and physiological tools in ataxic dysarthria management: a case report. 1996 Presented at: Proceedings of the 2nd Conference of the Korean Society of Phonetic Sciences and Speech Technology; Seoul, South Korea p. 9-22 URL: https://koreascience.kr/article/CFKO199613842057217.page [accessed 2026-05-08]
31. King JM, Watson M, Lof GL. Practice patterns of speech-language pathologists assessing intelligibility of dysarthric speech. J Med Speech Lang Pathol 2012;20(1):1-10.
32. Shin MJ, Kim JO, Lee SB, Lee SY. Speech Mechanism Screening Test 2010.
33. Ziegler W. Task-related factors in oral motor control: speech and oral diadochokinesis in dysarthria and apraxia of speech. Brain Lang 2002 Mar;80(3):556-575. [doi: 10.1006/brln.2001.2614] [Medline: 11896657]
34. Kim YT, Park H, Kang JK, et al. Validity and reliability analyses for the development of Urimal Test of Articulation and Phonology-2. Commun Sci Disord 2018 Dec 28;23(4):959-970. [doi: 10.12963/csd.18545]
35. Namasivayam AK, Huynh A, Granata F, Law V, van Lieshout P. PROMPT intervention for children with severe speech motor delay: a randomized control trial. Pediatr Res 2021 Feb;89(3):613-621. [doi: 10.1038/s41390-020-0924-4] [Medline: 32357364]
36. Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary care evaluation of mental disorders. Patient health questionnaire. JAMA 1999 Nov 10;282(18):1737-1744. [doi: 10.1001/jama.282.18.1737] [Medline: 10568646]
37. Piacentini V, Zuin A, Cattaneo D, Schindler A. Reliability and validity of an instrument to measure quality of life in the dysarthric speaker. Folia Phoniatr Logop 2011;63(6):289-295. [doi: 10.1159/000322800] [Medline: 21474942]
38. Long A, Hesketh A, Bowen A, Study ANR. Communication outcome after stroke: a new measure of the carer's perspective. Clin Rehabil 2009 Sep;23(9):846-856. [doi: 10.1177/0269215509336055] [Medline: 19482891]
39. Brooke J. SUS: a quick and dirty usability scale. : Usability Eval Ind; 1995 Nov 30 URL: https://www.taylorfrancis.com/chapters/edit/10.1201/9781498710411-35/sus-quick-dirty-usability-scale-john-brooke [accessed 2026-05-08] [doi: 10.1201/9781498710411-35]
40. Yorkston KM, Beukelman DR, Strand EA, Bell KR. Management of Motor Speech Disorders in Children and Adults, 2nd edition 1999.
41. Stipancic KL, van Brenk F, Qiu M, Tjaden K. Progress toward estimating the minimal clinically important difference of intelligibility: a crowdsourced perceptual experiment. J Speech Lang Hear Res 2025 Jul 29;68(7S):3480-3494. [doi: 10.1044/2024_JSLHR-24-00354] [Medline: 39453526]
42. Tweed CD, Quartagno M, Clements MN, et al. Exploring different objectives in non-inferiority trials. BMJ 2024 Jun 17;385:e078000. [doi: 10.1136/bmj-2023-078000] [Medline: 38886014]
43. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
44. International council for harmonisation of technical requirements for registration of pharmaceuticals for human use (ICH). In: ICH Harmonised Tripartite Guideline E10: Choice of Control Group and Related Issues in Clinical Trials 2000.
45. Kleim JA, Jones TA. Principles of experience-dependent neural plasticity: implications for rehabilitation after brain damage. J Speech Lang Hear Res 2008 Feb;51(1):S225-S239. [doi: 10.1044/1092-4388(2008/018)] [Medline: 18230848]
46. El Hachioui H, Lingsma HF, van de Sandt-Koenderman ME, Dippel DWJ, Koudstaal PJ, Visch-Brink EG. Recovery of aphasia after stroke: a 1-year follow-up study. J Neurol 2013 Jan;260(1):166-171. [doi: 10.1007/s00415-012-6607-2] [Medline: 22820721]
47. Yorkston KM, Beukelman DR, Strand EA, Hakel M. Management of Motor Speech Discorders in Children and Adults, 3rd edition: Pro-Ed; 2010.
48. Kim Y, Kim M, Kim J, Song TJ. Smartphone-based speech therapy for poststroke dysarthria: pilot randomized controlled trial evaluating efficacy and feasibility. J Med Internet Res 2024 Apr 25;26:e56417. [doi: 10.2196/56417] [Medline: 38509662]
49. Bhogal SK, Teasell R, Speechley M. Intensity of aphasia therapy, impact on recovery. Stroke 2003 Apr;34(4):987-993. [doi: 10.1161/01.STR.0000062343.64383.D0] [Medline: 12649521]
50. Berthier ML, Pulvermüller F. Neuroscience insights improve neurorehabilitation of poststroke aphasia. Nat Rev Neurol 2011 Feb;7(2):86-97. [doi: 10.1038/nrneurol.2010.201] [Medline: 21297651]
51. Hill AJ, Theodoros DG, Russell TG, Ward EC, Wootton R. The effects of aphasia severity on the ability to assess language disorders via telerehabilitation. Aphasiology 2009 May;23(5):627-642. [doi: 10.1080/02687030801909659]

XSL•FO
RenderX

52. Macoir J, Lavoie M, Routhier S, Bier N. Key factors for the success of self-administered treatments of poststroke aphasia using technologies. Telemed J E Health 2019 Aug;25(8):663-670. [doi: 10.1089/tmj.2018.0116] [Medline: 30142032]
53. Gwynne-Mayer N, Drummond AE, Hancox J, Alrubaia W, Taylor IM. Adherence to exercise rehabilitation programmes in stroke survivors: a scoping review. BMJ Open Sport Exerc Med 2025;11(3):e002102. [doi: 10.1136/bmjsem-2024-002102] [Medline: 41098374]
54. Mahmood A, Nayak P, Deshmukh A, et al. Measurement, determinants, barriers, and interventions for exercise adherence: a scoping review. J Bodyw Mov Ther 2023 Jan;33:95-105. [doi: 10.1016/j.jbmt.2022.09.014] [Medline: 36775533]
55. Weinstein RS, Lopez AM, Joseph BA, et al. Telemedicine, telehealth, and mobile health applications that work: opportunities and barriers. Am J Med 2014 Mar;127(3):183-187. [doi: 10.1016/j.amjmed.2013.09.032] [Medline: 24384059]
56. Kempler D, Van Lancker D. Effect of speech task on intelligibility in dysarthria: a case study of Parkinson's disease. Brain Lang 2002 Mar;80(3):449-464. [doi: 10.1006/brln.2001.2602] [Medline: 11896652]
57. Beukelman DR, Yorkston KM. Influence of passage familiarity on intelligibility estimates of dysarthric speech. J Commun Disord 1980 Jan;13(1):33-41. [doi: 10.1016/0021-9924(80)90019-2] [Medline: 7354140]

Abbreviations

ANCOVA: analysis of covariance
CONSORT: Consolidated Standards of Reporting Trials
ICC: intraclass correlation coefficient
ITT: intention-to-treat
MPT: maximum phonation time
PCC: percentage of consonants correct
PHQ-9: Patient Health Questionnaire-9
PP: per-protocol
QoL-Dys: Quality of Life in the Dysarthric Speaker
SLP: speech-language pathologist
SUS: System Usability Scale

Edited by A Coristine; submitted 06.Aug.2025; peer-reviewed by A Shivanna, A Babuci, K Usop; revised version received 03.Apr.2026; accepted 03.Apr.2026; published 18.May.2026.

Please cite as:
Kim Y, Kim M, Kim S, Kim J, Shin JH, Chang Y, Na JY, Kim J, Song TJ
Efficacy of Digital Speech Therapy for Poststroke Dysarthria: Randomized Noninferiority Trial
J Med Internet Res 2026;28:e81938
URL: https://www.jmir.org/2026/1/e81938
doi:10.2196/81938

© Yuyoung Kim, Minjung Kim, Saebyeol Kim, Jinwoo Kim, Joon-Ho Shin, Yoonkyung Chang, Ji Young Na, JungWan Kim, Tae-Jin Song. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Messages and Notifications for the "OA Coach" Knee Osteoarthritis Self-Management Mobile App: Codevelopment and Evaluation Using a Participatory Research Design With Focus Groups and Surveys

Kate L Bryce[1], MHSc; Jocelyn L Bowden[1], PhD; David J Hunter[1,2], PhD; Robin Huang[3], PhD; Naomi Bloul[1], BSc; Vicky Duong[1*], PhD; Jillian Eyles[1*], PhD

[1]Sydney Musculoskeletal Research Centre, Kolling Institute, The University of Sydney, Level 10, Reserve Road, St Leonards, Sydney, NSW, Australia

[2]Department of Rheumatology, Royal North Shore Hospital, Northern Sydney Local Health District, Sydney, NSW, Australia

[3]Vulsen, Sydney, NSW, Australia

[*]these authors contributed equally

Corresponding Author:
Jillian Eyles, PhD
Sydney Musculoskeletal Research Centre, Kolling Institute, The University of Sydney, Level 10, Reserve Road, St Leonards, Sydney, NSW, Australia

Abstract

Background: The OA Coach mobile app is a digital self-management tool designed to support people with knee osteoarthritis (OA).

Objective: We aimed to enhance the digital user experience of the OA Coach app by working with consumers, health professionals, and researchers to (1) improve existing OA Coach app call-to-action and encouragement notifications by applying behavior change theory, refining, reviewing, and adapting them, and (2) co-design new educational messages based on international OA guidelines.

Methods: This mixed methods study used a 3-phase process for improving existing notifications and developing new educational messages: (1) initial development and/or enhancement using behavior change techniques and accessible language, (2) expert review, and (3) refinement based on participant feedback. We enhanced the existing notifications to incorporate behavior change theory and improve readability, then we assessed ease of understanding, usefulness, and motivational impact in an online survey using 1 - 5 Likert scales and free-text responses. Median scores and IQRs were calculated, free-text responses were summarized using content analysis, and notifications that scored ≤12/15 and/or had 2 or more consistent recommendations in free-text responses were systematically refined. The new weekly educational messages were initially drafted based on OA clinical practice guidelines, then reviewed in 3 focus groups (in person and online) for detailed guidance on topic selection, language, and timing. Content analysis of focus group data informed systematic refinement of the messages. The refined messages were assessed for understanding and usefulness in an online survey (1 - 5 Likert scales), and free-text responses were refined further based on survey results.

Results: Fifty-seven participants assessed the existing notifications in the survey. Consumers rated notifications relating to sleep and mood the lowest, and those encouraging logging pain scores the highest. Health professionals and researchers reported a median score of 12/15 across all notifications, with greatest variability observed for step count, activity tracker use, and pain score logging. Eighty notifications were refined to enhance clarity, engagement, and effectiveness in prompting meaningful actions by future OA Coach users. Overall, 6 consumers, 7 health professionals, and 6 researchers participated in 3 focus groups to co-design 14 educational messages. A preference for a consistent message structure emerged: a descriptive heading, a brief introductory sentence, followed by 3 or 4 key points with practical examples, and concluded with a motivating statement. Key themes guiding modifications for both notifications and educational messages included reducing technical jargon, simplifying colloquial expressions, and removing language perceived to be patronizing or frustrating.

Conclusions: Through an iterative process, we used behavior change theory and worked with consumers, health professionals, and researchers to identify key preferences for message content, language, tone, and structure. These insights informed the refinement of 80 notifications and the co-design of 14 educational messages for the OA Coach app.

(J Med Internet Res 2026;28:e83507) doi:10.2196/83507

XSL•FO
RenderX

KEYWORDS

knee osteoarthritis; mobile app; digital health; notifications; education; self-management; behavior change; co-design; mixed methods

Introduction

An estimated 595 million people globally have osteoarthritis (OA), most commonly at the knee and hip [1], with prevalence projected to surge dramatically by 74.9% and 78.6%, respectively [2]. Driven by an aging population, a rise in obesity, joint injuries, and increasing physical inactivity, OA is the fastest-growing cause of disability worldwide [1,3,4]. OA symptoms include joint pain, physical limitations, and associated challenges such as low mood, interrupted work, and social participation, resulting in enormous social and economic burden [5]. To address the enormous burden of OA, international clinical practice guidelines recommend education for self-management, exercise, and weight management as first-line interventions for all people with hip and knee OA [6-14]. This care can result in some people delaying or avoiding joint replacement surgery altogether [11,15-17]. However, in Australia, first-line OA treatments are underused [18,19], often driven by common misconceptions about OA, the perceived efficacy (or lack thereof) of first-line treatments, and limited access to these treatments [20,21].

Scalable solutions are needed to address these challenges that improve public access to guideline-recommended care. Digital health technologies, such as mobile health (mHealth) apps, offer a promising approach to overcome these barriers by providing accessible, low-cost, and evidence-based interventions that can be delivered at scale. Digital technology using mobile phones is a fast-growing field that promotes adherence to treatment in self-managing chronic conditions, including OA [22-25]. The World Health Organization identified developing digital health as a global priority, emphasizing that it can support equitable and universal access to quality health services, improve health outcomes [4,26], and improve access to the best-evidence care.

The OA Coach is a mobile app developed by the University of Sydney's Osteoarthritis Clinical Research Group (OACRG), which aims to deliver first-line OA treatments and facilitate long-term behavior change for self-management in people with knee OA [27-30]. The OA Coach combines several strategies to achieve this, including the delivery of call-to-action and encouragement notifications via short, automated messages designed to prompt participants to take specific actions, such as logging their pain levels or completing a task. The knee OA education messages, on the other hand, are more detailed communications that provide participants with information, guidance, and support related to managing their condition. An initial bank of 88 notifications were developed by the researchers (VD and NB) and were piloted in a feasibility study (March-April 2024) [27]. Although the messages used positive, encouraging language, they were not grounded in behavior change theory. Participant feedback from the pilot study suggested that the notifications were useful, though improvements in content would enhance their effectiveness.

Therefore, we aimed to enhance the digital user experience of the OA Coach app by improving the relevance and utility of the notifications by incorporating behavior change theory, data-driven techniques, and a co-design approach. Longer educational messages were also developed to better support participants to self-manage their knee OA.

The two aims were (1) to enhance the original OA Coach app call-to-action and encouragement notifications by applying behavior change theory, refining, reviewing, and adapting them with consumers, health professionals, and researchers, and (2) to codevelop new knee OA education messages based on international OA guidelines using focus groups and surveys with consumers, health professionals, and researchers.

Methods

Study Design

The mixed methods study design was based on 2 previously published frameworks for digital intervention development [31,32] and adapted to include three phases: (1) initial development and/or enhancement with behavior change techniques (BCTs) and accessible language, (2) expert review, and (3) refinement based on participant feedback. The study was conducted at the Kolling Institute, the University of Sydney (December 2024 to March 2025). This study is reported in accordance with STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines (Checklist 1) [33].

The OA Coach App

OA Coach integrates a wrist-based activity tracker (Fitbit Inspire 2) to monitor physical activity levels, calorie count, and hours of sleep. The app incorporates recommended first-line OA management, including education for self-management and strategies for increasing physical activity levels. It uses BCTs, defined as the smallest components within intervention strategies that are observable and replicable, and designed to facilitate a shift in health behaviors [34]. Examples of BCTs used in the OA Coach app are self-monitoring of pain levels and body weight, call-to-action and encouragement notifications, and goal setting to improve and sustain physical activity levels.

The OA Coach app integrates both passive data collection (eg, step count, sleep duration, and energy expenditure/calories burned via the Fitbit Inspire 2) and active user-reported measures (eg, pain scores on a visual analog 1 - 10 scale [35], 21-Item Depression, Anxiety and Stress Scale (DASS-21) [36,37], Pittsburgh Sleep Quality Index (PSQI) [38], and body weight tracking). This combination enables the delivery of tailored feedback and support consistent with the principles of digital behavior change interventions, which use technology to promote health behavior change through personalized and timely strategies [39,40]. In particular, the app uses just-in-time adaptive interventions (JITAIs) design principles [41], in which prompts and guidance are delivered at moments when they are

XSL•FO
RenderX

expected to be most relevant and actionable, based on predefined rules using the user's current context, progress, and goals [41]. In the OA Coach app, these JITAI-informed interventions are rule-based. For example, if the user takes fewer than their goal number of steps by 4 PM, then at 5 past the hour, a call-to-action notification is sent: "Almost there! [number of steps] steps stand between you and the step podium. Strap on those sneakers and step into the winner's circle."

Participants and Recruitment

People were eligible if they had been diagnosed with knee OA, were aged 18 years or older, spoke fluent English, and owned a smartphone. Previous participants from the feasibility study were contacted to be part of this study. Consumers may or may not have used the OA Coach app previously during the 6-week feasibility study [27]. Health professionals and researchers were eligible if they had experience working with people with knee OA. Specifically, health professionals were eligible if they treated people with OA, and researchers were expected to have published at least two research papers on OA management. Demographic data collected included age, gender, ancestry, education level, and years of work experience (health professionals and researchers).

Recruitment was conducted through the OACRG volunteer database. Health professionals and researchers were invited via email through existing professional networks. Purposeful sampling for the focus groups was used to ensure diversity in participant ethnicity, age, professional role, and educational level. Participation in the focus group was not dependent on completing the survey.

Aim 1: The Enhancement of Call-to-Action and Encouragement Notifications With Consumer, Health Professional, and Researcher Review

Phase 1: Enhancement of Notifications With BCTs and Accessible Language

First, using the pilot data and existing evidence for common barriers and facilitators to exercise program uptake in individuals with knee OA [42-45], we reviewed the content and purpose of each existing OA Coach app notification to identify its core message and assign a behavioral outcome. The notifications were divided into 6 domains: step count, sleep goals, logging pain scores, completing sleep and mood surveys, goal achievement, and wearing an activity tracker. They were also categorized as one of two types: "call-to-action" notifications designed to prompt a specific behavior, or "encouragement" notifications, intended to motivate the user.

Second, the notifications were individually mapped to the Behavior Change Technique Taxonomy (v1) [46]. Some notifications were modified to incorporate additional, appropriate BCTs such as information about health consequences (BCT 5.1), instructions on how to perform a behavior (4.1), goal setting (1.1), action planning (1.4), and habit formation (8.3) [46]. A summary of the specific BCTs applied across the notifications is provided in Multimedia Appendix 1.

Third, the notifications were modified to target a reading level of grade 8 or lower to reduce reading effort and to help prevent potential misunderstandings for all users (approximately 12 - 14 years of age) [47]. The Sydney Health Literacy Lab's SHeLL Editor was used to support simplification without compromising meaning [48,49]. Table 1 shows examples of the initial notification enhancement process, with the literacy level score based on the SHeLL Editor [49].

XSL•FO
RenderX

Table . Example of the initial notification enhancement process. Drafts were revised by integrating behavior change techniques (coded using the Behavior Change Technique Taxonomy v1) and assessed for readability (Sydney Health Literacy Editor; target ≤ grade 8).

Initial notification	Summary of content	Additional content	BCTTv1[a] classification	Reading level	Enhanced notification
You're on fire! Keep those steps coming.	General encouragement and support	Add detailed instructions on how to succeed	Prompt/cues (BCT[b] 7.1), social support (unspecified) (BCT 3.1), habit formation (BCT 8.3), action planning (BCT 1.4)	3.1	You're on fire! Keep those steps coming. Make it a habit by walking at the same time each day
Look at you go! You exceeded your step goal.	Immediate feedback using positive language informing user they have met their step goal	Add information about health consequences	Prompt/cues (BCT 7.1), feedback on behavior (BCT 2.2), social support (unspecified) (BCT 3.1), information about health consequences (BCT 5.1)	7.8	Look at you go! You exceeded your step goal. Your knees are getting stronger, and you've improved your overall joint health
Time for a sleep check-up! Your PSQI[c] survey is waiting. Let's make sure your nights are as restful as they can be!	Prompt and feedback on health consequences implying monitoring sleep will make nights more restful	Negative feel to message. PSQI to change to sleep survey to enhance clarity of notification. Add instruction on how to perform the behavior (filling in the survey)	Prompt/cues (BCT 7.1), information about health consequences (BCT 5.1), instruction on how to perform the behavior (BCT 4.1), social support unspecified (BCT 3.1)	8.8	Ready to enhance your sleep quality? Fill in your sleep survey now to optimize your nights for restful sleep! Your input makes a difference!
Great job wearing your Fitbit this week!	Prompt and encourages continued engagement (commitment to Fitbit wear) with positive reinforcement	Add information about health consequences	Prompt/cues (BCT 7.1), feedback on behavior (BCT 2.2), social support unspecified (BCT 3.1), information about health consequences (BCT 5.1)	7.2	Great job wearing your Fitbit this week! Every step counts toward your health goals!

[a]BCTTv1: Behavior Change Technique Taxonomy v1.

[b]BCT: behavior change technique.

[c]PSQI: Pittsburgh Sleep Quality Index.

Phase 2: Consumer, Health Professional, and Researcher Review

We planned to recruit 40 consumers, 10 health professionals, and 10 researchers to allow for each of the 80 notifications to be reviewed by a minimum of 5 consumers and 2 health professionals or researchers, distributed across 8 discrete surveys. The 8 surveys were automatically randomly allocated to participants using the Research Electronic Data Capture (REDCap) system [50,51]. Each notification was assessed based on 3 questions using a 5-point Likert rating scale used in similar studies in cardiovascular disease and low back pain [31,32]. The questions were:

- 1. Is the information in this notification easy to understand? ("1=very difficult to understand" to "5=very easy to understand")
- 2. Is the information in this notification useful? ("1=not useful at all" to "5=very useful")
- 3a. How likely are you to take action after receiving this notification? ("1=not likely at all" to "5=very likely")
- 3b. How would you rate the positive encouragement within this notification? ("1=very poor" to "5=excellent")

There was also a free-text section for general feedback and suggestions for language modifications. Scores for each message were summed to yield a total score between 3 and 15 points (15 was best). This scoring system was a descriptive tool used to guide refinements of the notifications and to support interpretation, rather than an indicator of effectiveness or being intended as a validation tool.

Phase 3: Data Analysis and Refinement

Descriptive statistics (median and IQR) were calculated for 3 of the 4 characteristics (understanding, usefulness, likelihood of taking action, and positive encouragement) within each notification domain (see the Results section). Results were analyzed separately for consumers and for health professionals/researchers, then combined. All analyses were conducted in Microsoft Excel and manually cross-checked for accuracy. Visualizations, including radar charts (Multimedia Appendix 2), were created to highlight group-based patterns (consumer vs health professionals/researchers) and domain-specific performance. Qualitative feedback from open-ended survey responses was summarized using a simple content analysis approach. Comments were coded into common categories, including tone, clarity, grammatical accuracy, and perceived helpfulness.

XSL•FO
RenderX

Thresholds for perceived notification acceptability were established to assist with the interpretation of the survey data based on approaches used in prior research [31,32]. Median scores below 12 were classified as low acceptability, scores of 12 and 13 indicated moderate acceptability, and scores of 14 and above as high acceptability. Notifications that received a median score of <12 points were reviewed and revised based on feedback. Notifications that scored ≥12 points were modified if minor changes were suggested or if there was consistent feedback from 2 or more participants.

These insights guided the final refinement of notifications, even for those that met the quantitative threshold of 12/15 or more.

Aim 2: The Development of Knee OA Educational Messages

Phase 1: Development and Enhancement of Educational Messages With BCTs and Accessible Language

We drafted 12 new educational messages, designed to be longer and more informative than the existing call-to-action and encouragement notifications, incorporating clinical practice guideline recommendations and 21 key patient messages about OA established by French et al [7,52,53]. The messages featured BCTs developed using the capability, opportunity, motivation—behavior (COM-B) model and the Behavior Change Technique Taxonomy [46,54,55]. The Flesch-Kincaid readability score [56] and SHeLL Editor [49] were used to evaluate the readability of the educational messages and identify opportunities to simplify the language and sentence structure to align with a grade 8 reading level. We planned to deliver one educational message weekly to the OA Coach users' app inbox over a 12-week period.

Phase 2: Consumer, Health Professional, and Researcher Review

To explore the educational messages in-depth, we conducted 3 focus groups, aiming to recruit 6 participants per group as recommended to encourage rich, detailed feedback [57,58], and allow for collaborative rewriting of the message content. The first focus group for consumers was 3 hours in duration, held in person at the Kolling Institute, University of Sydney. The second and third focus groups were 2 hours in duration, held online with health professionals and researchers, respectively. Our focus groups were structured around the 7-step co-design framework established by Trischler et al [59], which our team used previously [60]. A semistructured discussion guide was drafted by the lead author (KLB) (Multimedia Appendix 3). The guide aimed to explore participant views on whether the messages aligned with international guidelines (health professionals and researchers only), incorporated positive language and BCTs, provided practical strategies for lifestyle changes, and the optimal order and timing of messages within the app. The guide was reviewed and refined by 2 coauthors (JE and JLB) prior to use. The focus group methodology was informed by the information power approach, whereby researchers prioritized depth and richness of discussion of each message rather than aiming for data saturation [61]. Three focus groups were sufficient for this approach.

Audio recordings and verbatim transcriptions of the focus group discussions provided detailed insights into group dialogue. For the in-person focus group with consumers, printed posters displaying each message were provided, and participants were encouraged to add Post-it notes with suggestions to improve them. Similarly, during the online focus groups with health professionals and researchers, a shared digital whiteboard was used to record participants' comments on each message in real time. Focus group participants were unable to view others' comments.

Phase 3: Data Analysis and Refinement

We conducted a content analysis of focus group data, including audio-recorded transcripts from the 3 focus groups and feedback written directly on printed message handouts, posters, and online Post-it notes within Microsoft Whiteboard. One researcher (KLB) manually reviewed and categorized all data in a Microsoft Excel spreadsheet, grouped participant comments by message, and noted recurring patterns and points of emphasis, particularly related to clarity, tone, and usefulness. The categorization of the messages was reviewed by a second researcher (JE). Using a structured approach, each message was systematically refined and checked by the research team.

Following this, the focus group participants were sent a final REDCap survey to evaluate the refined education messages. Participants were asked to rate how easy each message was to understand ("1=very difficult to understand" to "5=very easy to understand") and its usefulness ("1=not useful at all" to "5=very useful"). Participants also had the opportunity to provide further feedback for each educational message. This feedback was used to refine the educational messages a final time before they were added to the bank of automated educational messages in both audio and readable formats within the OA Coach mobile app.

Given the small sample size for the final survey, we used median values to represent central tendency and reported the minimum and maximum scores to describe the full range of responses. Results were analyzed separately for consumers and for health professionals/researchers, then combined. All quantitative data were analyzed using Microsoft Excel.

Ethical Considerations

Ethics approval was provided by the University of Sydney Human Research Ethics Committee (2024/HE001089). All participants provided informed consent. Participant data were deidentified, securely stored, and confidentiality was maintained throughout analysis and dissemination. Consumer participants received AUD 150 (US $106.92) and parking costs for in-person focus group participation as recommended by Health Consumers NSW [62]. Consumers who completed the notification survey were placed in a draw to win one of two AUD 100 (US $71.28) vouchers. The health professional and researcher participants were not remunerated.

XSL•FO
RenderX

Results

Participants

Figure 1 illustrates the flow of participants throughout the study, including engagement with the notification survey, focus groups, and a final educational message survey. A total of 40 consumers, 8 health professionals, and 9 researchers completed the initial online notification survey. Six consumers from the survey group also participated in the in-person consumer focus group. In addition, 7 health professionals and 6 researchers participated in their respective online focus groups. All focus group participants completed a final educational message survey following their group discussion.

Among consumers, 50% (20/40) identified as female, and the age range was 46 to 85 years. Educational attainment was high, with nearly half completing postgraduate study. The health professional and researcher group was predominantly female, with an age range of 42 to 56 years. This group was highly educated and experienced; the majority held postgraduate qualifications and had over 10 years of work experience. Further demographic data can be found in Table 2.

Figure 1. Flow of participants through notification and education message evaluation stages.

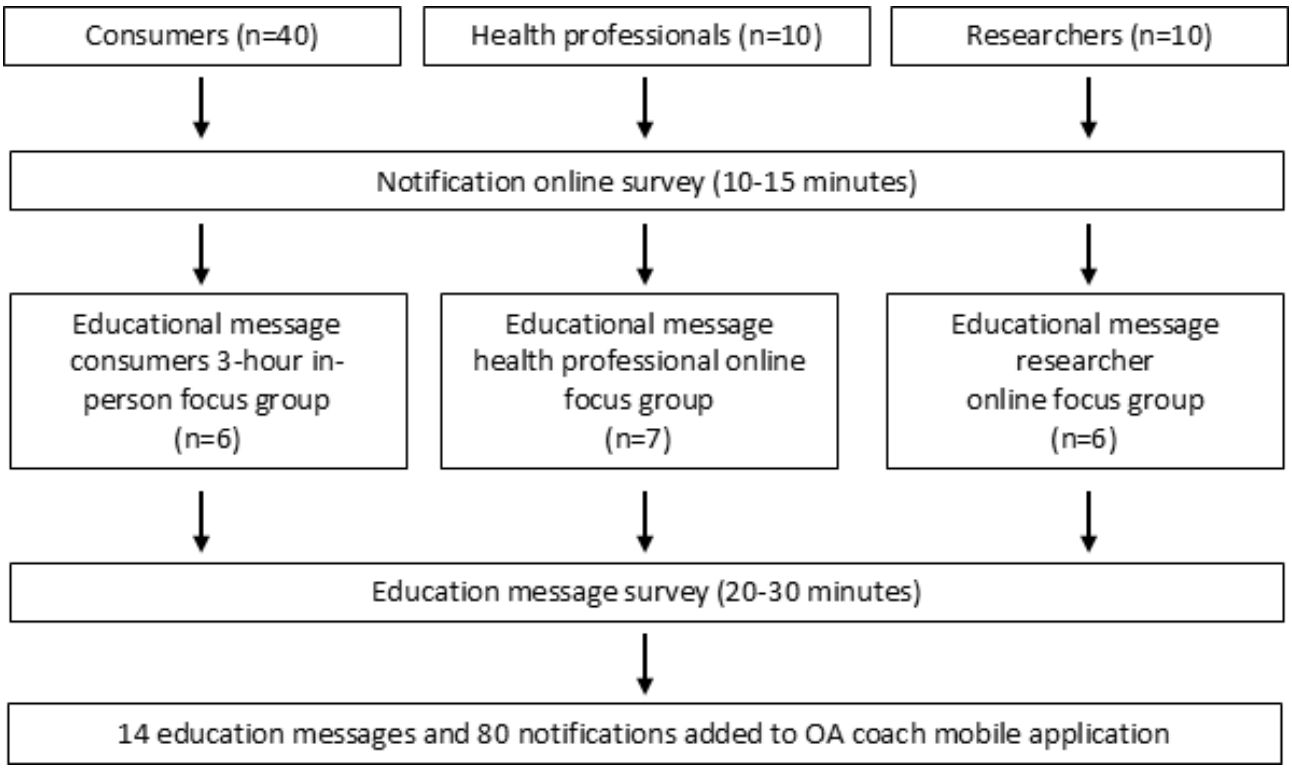

XSL•FO
RenderX

Table . Demographic details of participants included in the study.

	Consumers		Health professionals[a] and researchers	
	Total (n=40)	Focus group (n=6)	Total (n=20)	Focus group (n=13)
Gender				
Woman, n (%)	20 (50)	3 (50)	15 (75)	11 (85)
Man, n (%)	20 (50)	3 (50)	5 (25)	2 (15)
Ancestry				
White (Australian, British Isles, European, and American), n (%)	38 (95)	5 (83)	17 (85)	11 (85)
Other (Middle Eastern and Asian), n (%)	2 (5)	1 (17)	3 (15)	2 (15)
Age (years)				
Mean (SD)	65 (8)	73 (8.6)	44 (7)	42 (6)
Range	46 - 85	55 - 79	42 - 56	33 - 55
Education, n (%)				
Three or more years of high school education	7 (18)	1 (17)	0 (0)	0 (0)
University degree	15 (38)	2 (33)	1 (5)	1 (8)
Postgraduate study	18 (45)	3 (50)	19 (95)	12 (92)
Years of work experience (health professionals and researchers only)				
3 - 5 years	—[b]	—	2 (10)	1 (8)
6 - 10 years	—	—	3 (15)	4 (31)
11 - 15 years	—	—	5 (25)	3 (23)
Over 15 years	—	—	10 (50)	5 (38)

[a]Of the 10 health professionals, 9 were physiotherapists and 1 was a medical doctor.

[b]Not applicable.

Aim 1: The Enhancement of Call-to-Action and Encouragement Notifications With Consumer and Expert Review

Consumer, Health Professional, and Researcher Review

A total of 80 notifications were assessed in the surveys. Of these, 16% (n=13) were scored below the 12/15 threshold and were revised based on participant feedback. In addition, 59% (n=47) of notifications with scores 12/15 or higher were refined in response to open-text feedback, which highlighted areas for improvement in clarity, tone, and grammatical accuracy. Nearly 20 (25%) notifications were retained unchanged, having scored 12/15 or higher, and received positive feedback.

Overall, participants provided generally high ratings across all notification domains, with subtle differences emerging between consumer and health professional/researcher groups (Figure 2). Among consumers, the greatest variability in responses was observed for notifications related to completing sleep and mood surveys (median 12, range 11 - 15) and sleep goals (median 12, range 11 - 15). For health professionals and researchers, the greatest variability was noted for notifications concerning step count (median 12, range 11 - 14; Table 3). Notifications relating to wearing an activity tracker (median 13, range 12 - 15) and logging pain scores (median 14, range 12 - 15) had the highest scores for consumers, and all the other notification scores for consumers, health professionals, and researchers had a median score of 12/15. When notification scores were analyzed by individual questions (understanding, usefulness, likelihood to take action, and positive encouragement), the ratings were most similar between groups for usefulness, closely followed by encouragement. Health professionals and researchers generally provided lower ratings for ease of understanding and likelihood of taking action (Figure 2 and Table 3).

XSL•FO
RenderX

Figure 2. Total notification scores (out of 15) for the consumers and the health professional and researcher group. Distribution of total median scores for each notification domain with comparisons between the consumer group and the health professional and researcher group. Median scores are presented alongside minimum and maximum values. The scores range from 0 to 15. Higher scores indicate more positive evaluations.

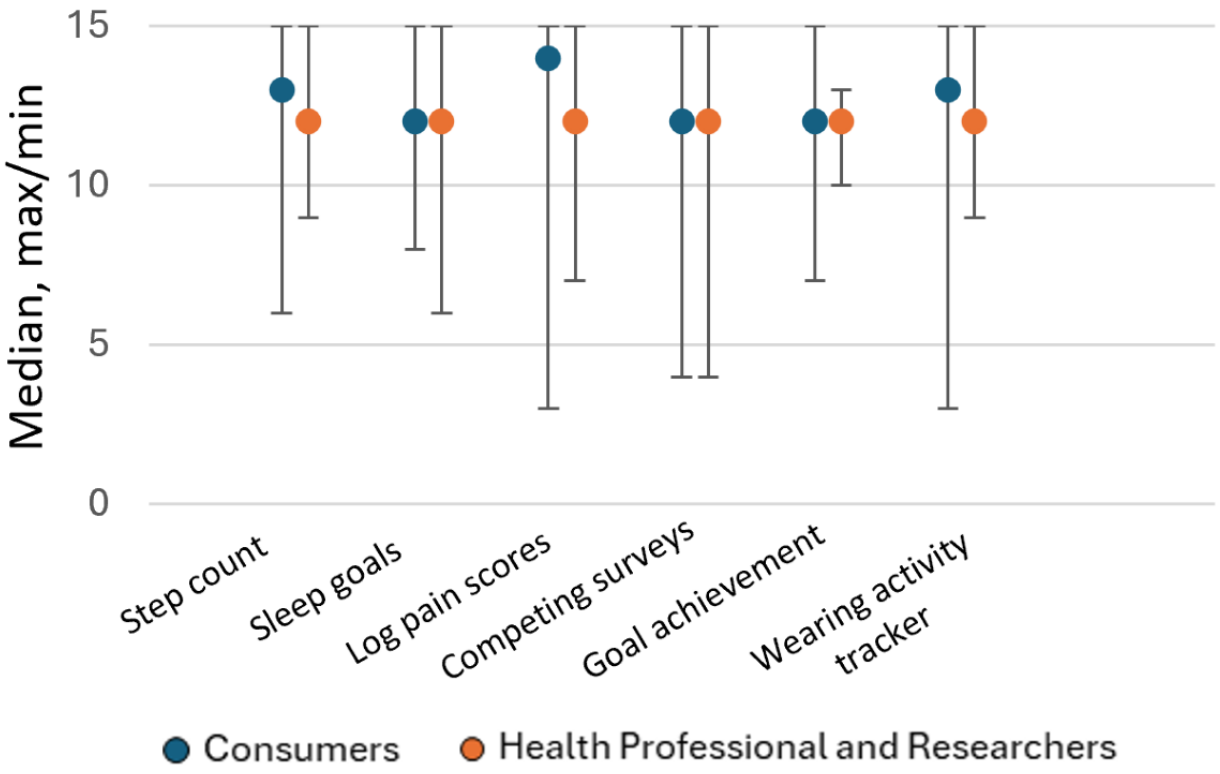

XSL•FO
RenderX

Table . Notification survey scores for each domain across 4 characteristics: understanding, usefulness, likelihood to take-action, and positive encouragement[a].

Notification domains (n=80) and characteristics assessed	Consumers, median (IQR)	HPR[b], median (IQR)	Combined scores, median (IQR)	Consumer total score		HPR total score		Combined total score, median (IQR)	Notifications scoring <12/15 changed (total=13), n	Notifications scoring ≥12/15	
				Median (IQR)	Min-max	Median (IQR)	Min-max			Changed (total=47), n	Unchanged (total=20), n
Step count (n=20)				12 (12-15)	6 - 15	12 (11-14)	9 - 15	12 (11-14)	5	9	6
Understanding[c]	5 (4-5)	4 (4-5)	5 (4-5)								
Usefulness[d]	4 (4-5)	4 (4-5)	4 (4-5)								
Likelihood to take action[e]	4 (4-5)	3 (3-4)	4 (3-4)								
Positive encouragement[f]	5 (4-5)	4 (3-5)	4 (4-5)								
Sleep goal (n=13)				12 (11-15)	8 - 15	12 (11-13)	6 - 15	12 (11-14)	1	11	1
Understanding	5 (4-5)	4 (4-5)	5 (4-5)								
Usefulness	4 (3-5)	4 (4-4)	4 (3-4)								
Likelihood to take action	4 (3-5)	3 (3-4)	4 (3-4)								
Positive encouragement	4 (4-5)	4 (4-5)	4 (4-5)								
Log pain scores (n=13)				14 (12-15)	3 - 15	12 (11-13)	7 - 15	12 (11-15)	1	7	5
Understanding	5 (4-5)	4 (4-4)	4 (4-5)								
Usefulness	4 (4-5)	4 (3-4)	4 (4-5)								
Likelihood to take action	5 (4-5)	4 (3-4)	4 (4-5)								
Positive encouragement	4 (4-5)	4 (4-5)	4 (4-5)								
Completion of surveys (sleep and mood) (n=11)				12 (11-15)	4 - 15	12 (11-12)	9 - 15	12 (11-14)	3	6	2
Understanding											
Sleep	5 (4-5)	4 (4-4)	4 (4-5)								
Mood	5 (4-5)	4 (4-4)	4 (4-5)								

XSL•FO
RenderX

Notification domains (n=80) and characteristics assessed	Consumers, median (IQR)	HPR[b], median (IQR)	Combined scores, median (IQR)	Consumer total score		HPR total score		Combined total score, median (IQR)	Notifications scoring <12/15 changed (total=13), n	Notifications scoring ≥12/15	
				Median (IQR)	Min-max	Median (IQR)	Min-max			Changed (total=47), n	Unchanged (total=20), n
Usefulness											
Sleep	4 (4-5)	4 (4-4)	4 (4-5)								
Mood	4 (3-5)	4 (3-4)	4 (3-5)								
Likelihood to take action											
Sleep	4 (4-5)	4 (4-4)	4 (4-5)								
Mood	4 (4-5)	4 (3-4)	4 (3-5)								
Achieving goals (n=4)				12 (12-14)	7 - 15	12 (11-12)	10 - 13	12 (11-13)	1	2	1
Understanding	4 (4-5)	4 (4-4)	4 (4-5)								
Usefulness	4 (4-5)	4 (4-4)	4 (4-5)								
Likelihood to take action	—[g]	—	—								
Positive encouragement	4 (4-4)	4 (3-4)	4 (4-4)								
Wearing activity tracker (n=19)				13 (12-15)	3 - 15	12 (12-13)	9 - 15	13 (12-15)	2	12	5
Understanding	5 (4-5)	4 (4-5)	5 (4-5)								
Usefulness	4 (4-5)	4 (4-4)	4 (4-5)								
Likelihood to take action	5 (4-5)	4 (4-4)	4 (4-5)								
Positive encouragement	4 (4-5)	4 (4-5)	4 (4-5)								

[a]Results are shown separately for consumers and health professionals/researchers, as well as combined group scores. Total scores (sum of 3 criteria) are also presented (median, IQR). In this table, n refers to the number of notifications (not participants).

[b]HPR: health professional/researcher.

[c]Is the information in this notification easy to understand? (“1=very difficult to understand” to “5=very easy to understand”).

[d]Is the information in this notification useful? (“1=not useful at all” to “5=very useful”).

[e]How likely are you to take-action after receiving this notification? (“1=not likely at all” to “5=very likely”).

[f]How would you rate the positive encouragement within this notification? (“1=very poor” to “5=excellent”).

[g]Not applicable.

XSL•FO
RenderX

Step Goal Notifications

The 20 notifications relating to step goals had a median rating of 12/15 (range 11 - 14) (Table 3). Step goal notifications were rated highest for ease of understanding (median 5, range 4 - 5) but had lower scores for likelihood of prompting action overall (median 4, range 3 - 4), particularly from health professionals and researchers (median 3/5, range 3 - 4). Qualitative feedback from consumers, health professionals, and researchers suggested that some step goal messages were clear and motivating, while others raised concerns about accessibility and tone. For instance, one consumer commented, "Clear simple statement encouraging participants to walk regularly" [Consumer]. In contrast, another noted, "I can't walk far, and I'm pretty unsteady, so this message wouldn't enable me to change that" [Consumer].

Colloquial language was flagged as an issue. Statements such as "you're on fire," which were intended to be motivating, were flagged as potentially confusing: "Some people may not understand the term 'you're on fire' especially if English is not their first language or they are from an older generation" [Researcher]. Similarly, " 'step up the pace' may suggest to some that they need to walk faster, is the message about accumulating more steps, or about accumulating more steps at a faster pace?" [Researcher]. As a result, this notification was revised to use simpler, more precise language.

> *Let's make every step count towards a healthier you! Try adding short walks into your day to help reach your step goal.*

Consumers provided feedback that step count was not always relevant to them, "walking over 100 meters is painful BUT I play bowls 3 time a week at 6000 steps each game 20 meters then rest 21 times in a game" [Consumer]. Despite being active, some participants reported their activities did not translate to higher step counts: "Less steps because I've done an hour and half of hydrotherapy and resting today" [Consumer].

The timing of step notifications was flagged as important: "I think the timing of this notification (step goal) would be important, not too late in the day so they still have time to take action" [Researcher]. Figure 3 shows the process of how a step goal notification was enhanced; 14 of the step goal notifications were modified after feedback.

XSL•FO
RenderX

Figure 3. Process of notification enhancement. Examples of the notification enhancement process after consumer, health professional, and researcher feedback. BCTs=behavior change techniques (coded using the Behavior Change Technique Taxonomy v1). Feedback=examples of survey comments. Survey scores (max 15). Reading level=grade level from Sydney Health Literacy Editor (target ≤ grade 8). PSQI: Pittsburgh Sleep Quality Index.

XSL•FO
RenderX

Sleep Goal Notifications

The 13 sleep-related notifications had a median overall score of 12/15 (range 11 - 13). Sleep goals were rated highest by consumers for ease of understanding (median 5, range 4 - 5) but scored comparatively lower in perceived usefulness and likelihood to take action (median 4, range 3 - 5). One participant commented, "A well-known fact, but stating it does not mean that it will make a difference to my sleep habit" [Consumer]. Another commented, "Sleep can be so problematic depending upon what else is happening in one's life. I also find quality sleep difficult to measure" [Consumer].

The notifications that scored lowest for likelihood to take action by health professionals and researchers were the sleep-related notifications (median 3, range 3 - 4), which had the equal lowest score with step count notifications. Qualitative feedback further contextualized these scores, noting the absence of precise guidance.

> *For many people who have sleep disturbance, the "gift" of 8 hours sleep might not be possible. We could be asking them to aim for something that is out of their control. There isn't an actionable step like: go to bed at a reasonable time or give yourself time to wind down before bed.* [Health professional]

Out of the 13 sleep notifications, 12 were enhanced after the feedback process. Figure 3 shows an example of a sleep notification enhancement.

Logging Pain Score Notifications

Logging pain score notifications received a median total score of 12/15 (range 11 - 15), indicating moderate acceptability overall. This was the highest consumer score for any notification type, with a median score of 14/15 (range 12 - 15), whereas the health professionals and researchers assigned a lower median score of 12/15 (range 11 - 13). The notification,

> *Have you logged your pain score today? It only takes a minute and it's so valuable for your health journey!*

received a median score of 15/15, showing high acceptability. The qualitative feedback reflected these divergent views on the usefulness and psychological impact of such prompts. Some participants found these notifications helpful for fostering reflective self-management:

> *I am beginning to listen to my knee pain a bit better these days...If I had a message as a reminder, I'm certain it would help or at least encourage me to stop and consider for a moment before deciding what to do.* [Participant 16, consumer]

Others raised concerns about the potential for pain-focused prompts to reinforce negative attention to symptoms:

> *Asking people to focus on the pain, of course it is painful. Some people will dwell on the pain rather than push through or raise their tolerance.* [Participant 34, consumer]
>
> *I would discourage rating pain levels with a chronic pain condition.* [Participant 9, researcher]

These data suggest that while pain tracking may support adaptive pacing and awareness for some, it may negatively highlight a strong pain response in others. Of the 13 notifications logging pain scores, 8 were changed based on feedback. Figure 3 shows an example.

Completion of Surveys (Sleep and Mood) Notifications

A total of 11 notifications prompted users to complete a monthly sleep and mood survey. The median total score was 12/15 (range 11 - 14), with similar scores for both groups. Consumers had a median score of 5/5 for perceived understanding, and all other scores were 4/5 for both groups. Among the notifications related to completing surveys, one stood out as particularly effective and required no modification throughout the iterative feedback process. The notification, "Make your screen time count, take 5 minutes to complete your sleep survey, now!" received a high overall score of 15/15, indicating high acceptability across all domains. One participant commented, "Encouraging, simple message to prompt participants to take action" [Consumer], and another commented, "I like that the prompt indicates how long it will take" [Researcher]. This example illustrates how concise notifications that set clear expectations for time commitment can enhance engagement, even when the task is related to a potentially sensitive topic, such as sleep disturbance (Figure 3).

Goal Achievement Notifications

There were only 4 preset notifications to encourage users to work toward their goals. Personalized goal achievement notifications were also generated based on the activity goals specific to the OA Coach app users. Only the 4 preset notifications were evaluated and these focused on daily step count, aiming for 150 minutes of aerobic activity per week, engaging in 3 knee strengthening sessions per week, and an encouragement notification if sleep duration falls below 7 hours. The median score for these notifications was 12 (range 11 - 13). Feedback from participants highlighted user assumptions about the broader app experience, with several participants noting that the effectiveness of certain notifications depended on additional in-app support. In response to the knee strengthening notification, one participant commented, "Positive clear message. One presumes there is also information provided on knee strengthening exercises" [Consumer]. Similarly, with the sleep goal notification, "I would like to think there would be some suggestions to help if this was the case?" [Consumer]. Out of the 4 goal achievement notifications, 3 were modified due to feedback (Figure 3).

Wearing the Activity Tracker

Notifications for wearing the activity tracker were more highly rated, with an overall median score of 13 (range 12 - 15). Participant feedback on notifications that used device-specific terminology (Fitbit Inspire 2) highlighted that they would prefer to have the option to use other wearable technologies, "I don't have a Fitbit so cannot comment on its usefulness" [Consumer], and "I assume that whatever app is used to read the steps taken that it will work with a range of devices, not just Fitbit. e.g., Garmin, Galaxy watch, Apple watch etc." [Consumer]. Further comments noted the overuse of colloquial language within the notifications (eg, "crush it," "your Fitbit is on fire," and "keep

XSL•FO
RenderX

rocking those steps"), which detracted from the simplicity of the message related to the watch. Additionally, feedback emphasized that messages should center around the user as the agent of change, rather than the technology itself.

> *This message focuses on the Fitbit being the key rather than the person, but I understand that by using the Fitbit then you get data that they can then take control themselves. Not sure of the "crush it" terminology for putting a watch on.* [Researcher]

The enhancement process for the notification mentioned is presented in Figure 3.

Feedback from the survey participants identified common reasons for suggested changes to the notifications. These included overly technical language (eg, references to "DASS-21" or "PSQI" surveys), colloquial or informal expressions (eg, "you're on fire," "high five," and "you're on a roll"), excessive or inconsistent punctuation, and unclear or confusing phrasing (eg, "let's step up the pace..."). Some notifications were described as patronizing or frustrating, particularly those with overly enthusiastic praise (eg, "gold medal performance" and "Fantastic job...") or unrealistic expectations (eg, "...8 hours sleep goal" and "keep those steps coming...").

Aim 2: The Development of Knee OA Educational Messages

Phase 1: Development and Enhancement of Educational Messages With BCTs and Accessible Language

The topics covered by the educational messages developed initially were as follows: understanding knee OA, the importance of staying active, healthy weight management, the benefits of strengthening exercises, exercise tips, overcoming barriers to exercise, nutrition for joint health, pain management—flare-ups, how assistive devices can help you move better, the power of good posture and staying motivated, what to know about knee surgery, and a concluding message of encouragement and goal setting.

Phases 2 and 3: Consumer, Health Professional, and Researcher Review, and Data Analysis and Refinement

Focus Groups

The consumer focus group was conducted in person with 6 participants (3 females and 3 males). The group recommended that 2 of the original 12 messages be divided into separate topics, resulting in a total of 14 educational messages. The additional topics were "Do I need a scan?" (initially titled "When should I get further investigations?") and "Staying motivated with knee OA," included as week 13 in the delivery sequence. A reordering of the message schedule was also advised, with the *exercise tips* message moved from week 5 to week 3, to highlight its importance in the early stages of self-management. The main topics and order of the weekly messages can be found in Multimedia Appendix 4.

Across all focus groups, a preference for a consistent message structure emerged. Participants responded positively to messages that began with a descriptive heading introducing the topic, followed by a brief introductory sentence, then 3 or 4 key points supported by relatable, practical examples, and concluded with a motivating statement with a supportive tone. This structure, reinforced with the use of bullet points and simple language, was identified as the most accessible and engaging way to present this information. Participants noted that this format helped them better understand and connect with the content. The two quotes below illustrate this:

> *Set up message with opening statement then points to follow.* [Consumer]
>
> *Dot points are easier to follow and clearer. Sentences simplified to key points.* [Researcher]

Participants consistently valued messages that used simple language, included relatable examples, and provided actionable advice. The following comments were made about message 2 on the importance of staying active:

> *First line too vague. points also too vague, give some explanation.* [Health professional]
>
> *How long is too long for being inactive - can make people anxious about being still. Give some guidance if possible.* [Health professional]

Health professionals and researchers placed stronger emphasis on ensuring the messages were evidence-based. For example, the original message "The power of good posture" (message 11) was critiqued as lacking evidence: "No evidence to support this" [Researcher] and "no research to support this" [Health professional]. One researcher suggested reframing the message to focus on mindful movement and its role during symptom flare-ups, "The power of mindful movement – reframe, could relate to flares. How does it feel? changing positions can help" [Researcher]. In response, the message was revised to "Pay attention to how you move" and included practical tips prompting users to notice how their knee feels, and to adjust movement or posture accordingly.

Several participants appreciated affirming statements that normalized setbacks within the messages, "Good to reassure about setbacks" [Researcher, message 5], and "Things can get in the way of exercise. This is normal" [Health professional, message 6]. Reinforcing that occasional lapses in health behaviors were acceptable was appreciated by consumers, "Don't stress. Have the occasional 'treat' it's o.k.!" [Consumer, message 7]. Across groups, there was a strong preference for messages that encouraged gradual behavior change, rather than setting unrealistic or rigid goals. Participants valued language that encouraged persistence, framed progress as cumulative, even if small, and built self-efficacy. Examples included "Every bit you do helps" [Consumer, message 6], "celebrate small wins" [Consumer, message 11], and "Every step is important, each step, no matter how big or small" [Researcher, message 14].

While the educational concepts proposed by French et al [53] provided a strong foundation for developing messages, several terms, such as "overweight" and "damage," generated discussion during the focus groups. These terms, though clinically conventional, were highlighted as unhelpful. Health professionals and researchers preferred "weight management"

XSL•FO
RenderX

versus "overweight" to encompass individuals across the weight spectrum, including those of normal or underweight status. Health professionals also recommended changing "joint damage on an x-ray" to a more neutral tone "joint changes on an x-ray." Consumers did not specifically comment on the term "damage"; they suggested that the draft message, which covered both investigations and assistive devices (eg, walking sticks, knee braces, and shoe inserts), be separated for simplicity. In response, a new message titled "How assistive devices can help you move" was developed.

Message 12, which focused on "what to know about knee surgery," warranted particular attention. Consumers emphasized the importance of clearly communicating that surgery does not always help, that rehabilitation can be challenging, and that better outcomes are associated with being physically stronger beforehand.

> *Rehab is not for the faint-hearted. You're not escaping exercise by having surgery.* [Consumer]
>
> *Surgery doesn't always help.* [Consumer]
>
> *Exercise is important at all stages of OA, and the stronger you are going into surgery, the better the outcome.* [Consumer]

Health professionals and researchers agreed with these key points; they further highlighted the need to encourage shared decision-making and suggested the inclusion of example questions patients could ask their health care team. They strongly recommended that the message include a clear statement reflecting international guidelines, which advised against the use of arthroscopy for knee OA due to its lack of effectiveness.

> *The decision to have surgery is a conversation to have with your health professional team give examples of questions.* [Researcher]
>
> *Message re arthroscopy very clear - recommended against. Not effective for knee OA.* [Researcher]
>
> *A viable treatment option for the right person needs to be discussed with your healthcare team. Even if you have it, it doesn't always help.* [Health professional]

Educational Message Final Survey

Participants in the focus groups provided feedback on the revised educational messages in a survey. Figure 4 shows that median score ratings for consumers were consistently high; they rated all messages more favorably than health professionals, with a median score of 5/5 across all messages for both understanding and usefulness. In contrast, the health professionals and researchers provided variable scores, particularly for perceived usefulness. Although their median ratings tended to cluster around 4 out of 5. Several messages received lower scores and showed greater variability, as indicated by the wider range of responses (eg, messages 6, 8, 12, and 13), and these were revised further based on this feedback.

XSL•FO
RenderX

Figure 4. Median (minimum and maximum) ratings for each education message by group: perceived ease of understanding and usefulness.

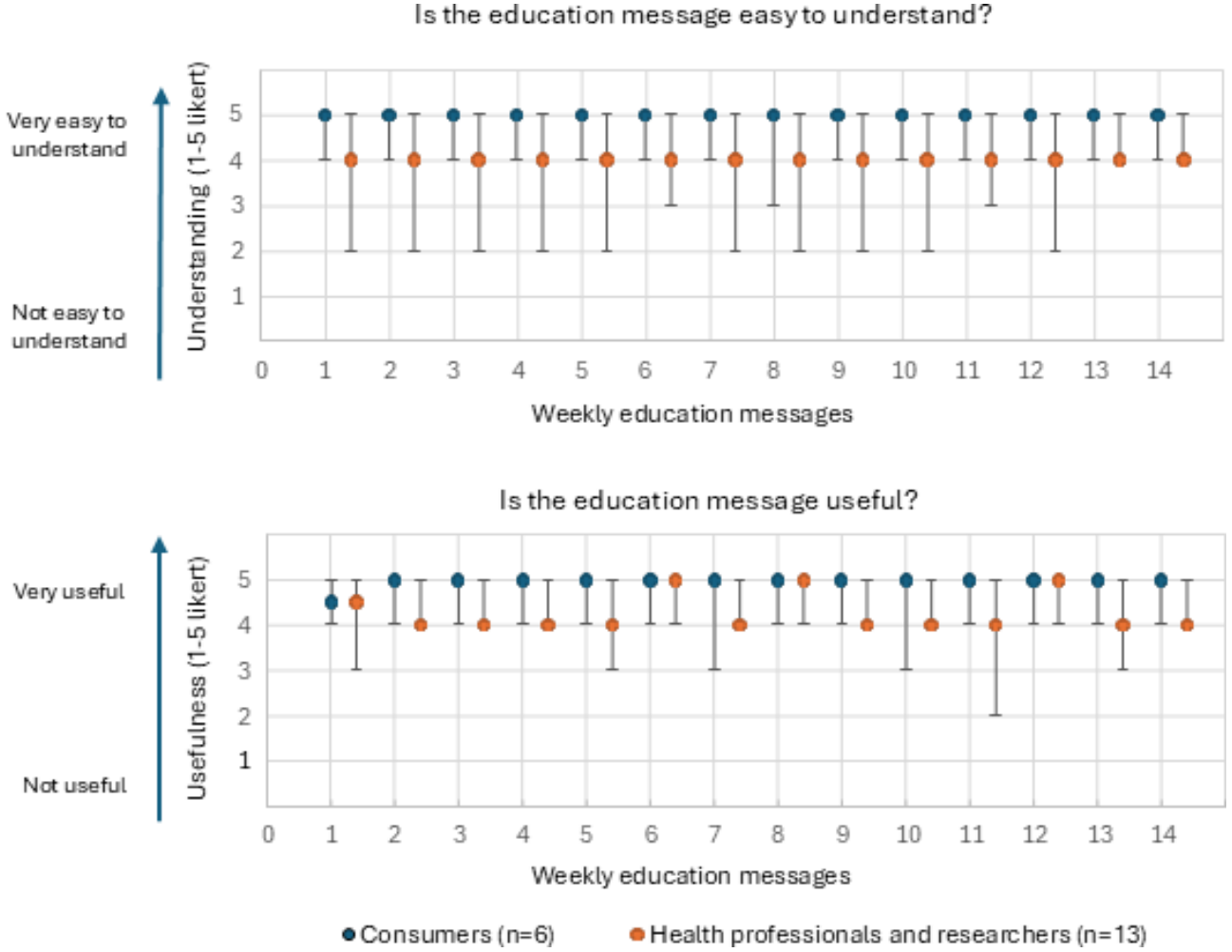

Discussion

Principal Findings

This study outlines the theory-informed iterative co-design process used to enhance notifications and develop educational messages intended to prompt users of the OA Coach app to adopt and sustain healthier behaviors for self-management of their knee OA. We grounded the design of the notifications and educational messages in behavior change theory, specifically the Behavior Change Technique Taxonomy [46] and COM-B model [55,63], to ensure that each notification and educational message addressed the key drivers of behavior. We sought consumer and expert feedback in the development of the OA Coach app notifications and educational messages using surveys and focus groups to improve their acceptability, clarity, usefulness, and tone. These qualities have been associated with improved user motivation, satisfaction, and engagement in prior research with digital health interventions [64-66].

There were similarities and differences in our findings from the surveys and focus groups with consumers compared with health professionals and researchers. A common finding was that colloquial language, used initially to add variety and positivity to the notifications, was often perceived to be confusing and distracted from the main message. In contrast, there was some variation in responses to the survey questions on mood and sleep. While median ratings of acceptability were similar between the consumer and the health professional/researcher groups, the consumer scores were more widely distributed, indicating that some consumers found these notifications more useful than others. There were greater differences between responses of the consumers and health professional/researcher groups related to the JITAI-informed daily pain score logging component of the OA Coach. While the pain logging notifications were rated favorably by consumers, the health professionals and researchers expressed concerns that daily logging might inadvertently reinforce hypervigilance or anxiety in some individuals with persistent OA pain [67]. These concerns are consistent with the broader pain literature, highlighting that continual self-monitoring of symptoms can sometimes exacerbate attention to pain [67]. Our findings underscore the importance of designing tracking features and associated notifications that balance the theoretical and practical benefits of adaptive, personalized feedback with user burden and perceived relevance [39,40]. Ensuring that the data collected contributes meaningfully to the user's personal goals and immediate self-management needs is likely to enhance engagement with, and the effectiveness of, digital behavior change interventions like OA coach [40].

The longer weekly OA Coach App educational messages were developed to provide more in-depth information, practical strategies, and motivational support, reinforcing

XSL•FO
RenderX

self-management behaviors. One output of the co-design focus groups was the emergence of a preferred layout, using short sentences in the active voice, at grade 8 reading level, with a descriptive heading, brief introduction, key points with practical examples, and an encouraging closing statement. We anticipate that the collaborative approach used to develop these messages will enhance comprehension, engagement, and empowerment of OA Coach app users.

Health professionals highlighted the importance of language choice, particularly with terms such as “joint damage” and “overweight,” preferring “joint changes” and “weight management.” Emerging literature suggests that biomedical language, such as “damage” and “wear and tear,” can reinforce unhelpful beliefs about structural deterioration, potentially increasing the fear of movement and reducing engagement in self-management strategies [20,68-73]. In their systematic review, Behera et al [74] emphasized that OA education is a complex relational process influenced by multidimensional factors and that empowering language can enhance patient engagement and agency. These insights align with the objectives of the Changing the Narrative on Osteoarthritis initiative, which calls for a shift from negative, condition-focused discourse toward language that supports resilience, autonomy, and self-efficacy in people living with OA [73]. In response to feedback from focus groups, surveys, and the mentioned research, educational messages were revised to reflect more empowering, hopeful language, while preserving the original evidence-based content [7,9,53,75].

Strengths and Limitations

The inclusion of consumers, health professionals, and researchers offered considerable strengths to the development process. Consumers represented a broad age range (46 - 85 years) and brought lived experience of knee OA, while the inclusion of health professionals and researchers added multidisciplinary perspectives and expert insight. High levels of educational attainment across all groups contributed to rich, thoughtful feedback, particularly regarding message clarity, tone, and relevance. However, the sample lacked cultural and linguistic diversity, with no representation from Aboriginal and Torres Strait Islander peoples, and limited representation of those from Pacific, African, or Latin American backgrounds. The higher education levels observed across the sample likely meant that participants were more comfortable with complex written material, although they still highlighted the need for grade 8 reading levels. This is supported by research demonstrating a strong association between higher education attainment and greater health literacy [76]. As such, our findings related to message readability may not be generalized to populations with lower education levels or limited health or digital literacy, who often experience greater difficulty understanding and accessing health information.

The small sample sizes for each group (consumers, health professionals, and researchers) and the exploratory nature of this study also limit its generalizability as the feedback may not have captured the full diversity of perspectives across age, digital literacy, health literacy, or OA severity. Although consumer feedback was prioritized throughout educational message development, and messages were adapted accordingly before being reviewed by health professionals and researchers, final decisions about edits were made by the research team. This may have inadvertently favored certain stakeholder views over others.

Another limitation was that some of the participants did not experience using the OA Coach app prior to providing feedback. Instead, they reviewed notifications and messages through printed handouts, digital documents, and focus group discussions. It is possible that engaging with the app in real-world conditions may have offered additional insights, as actual use involves contextual factors such as screen size, multitasking (interacting with the app while attending to other tasks or receiving alerts from multiple apps simultaneously), and notification timing. Therefore, while the focus groups and surveys allowed for rich discussion and language refinement, the extent to which these findings translate into behavior change in real-world settings remains uncertain. Nevertheless, this study provides a clear, reproducible method for patient-centered co-design of notifications and messages in mHealth interventions. This approach supports clarity, acceptability, and relevance for people using an mHealth app. Researchers and app developers can use these practical insights in their work as guidance for designing and refining patient-focused messages in mHealth contexts (Textbox 1).

Textbox 1. Practical tips for writing educational messages in mobile health.

✓ Descriptive subject heading

✓ Short introductory sentence

✓ 3-4 key bullet points

✓ Relatable, everyday examples

✓ Encouraging closing statement

✓ Active voice

✓ Grade 8 readability

✕ Avoid colloquial/slang expressions

✕ Avoid medical terms

XSL•FO
RenderX

Considerations for Future Research

Future research will evaluate these notifications and educational messages within the live OA Coach app to assess how users interact with content in real time. In addition, examining how different subgroups respond to variations in tone, format, and timings of notifications and messages will help inform more personalized digital health communication strategies, including the needs of culturally and linguistically diverse populations, to ensure these strategies are inclusive and responsive to diverse communication needs.

Currently, the OA Coach app offers semitailored support; however, advances in artificial intelligence will enable greater individualization. For example, by collecting additional information at the start of app use, such as current user activity levels, coexisting health conditions, user goals, and the topics a person most wants to learn about, the program could adapt both the type and order of messages to better align with individual needs. Future research could examine whether this type of approach enhances user engagement and self-management outcomes, while also considering important factors such as privacy and user trust. The updated OA Coach app is currently being evaluated in the COASTAL [77] randomized controlled trial, which compares face-to-face, telehealth, and digitally delivered models of care for knee OA.

Conclusions

This study developed and refined evidence-based notifications and educational messages for knee OA that were grounded in behavior change theory and aligned with international guideline recommendations. Through an iterative co-design process involving consumers, health professionals, and researchers, we identified key preferences for message content, tone, format, and structure. These insights informed the development of 14 educational messages and the refinement of 80 notifications for integration into the OA Coach app.

Acknowledgments

We are grateful to the people living with knee OA who generously shared their experiences, insights, and time. Their voices were invaluable in shaping the content of this research. We also thank the health professionals and researchers who contributed their time, expertise, and thoughtful feedback throughout the development process. Their collaboration significantly enriches the quality and rigor of this work. We used the generative AI tool Microsoft 365 MSO in the research and writing process under full human supervision. Full responsibility for the final manuscript rests with the authors.

Funding

The OA Coach app has been funded with the help of the Lenity Foundation. KLB is supported by the Postgraduate Research in Osteoarthritis Research Scholarship, University of Sydney.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JE, VD, and KLB conceived the study, and all authors contributed to the protocol. JE and VD co-led the study, and DJH obtained funding. VD and NB developed the original notification content, and KLB drafted educational content. JE and KLB recruited participants and facilitated the focus groups. KLB set up the Research Electronic Data Capture (REDCap) surveys, randomization, and automation. RH implemented the notification messages within the OA Coach app. KLB, JE, JLB, and VD did the statistical analysis. KLB drafted the manuscript, and all authors provided input to the manuscript preparation and approved the final manuscript.

Conflicts of Interest

JE is supported by a National Health and Medical Research Council Investigator Grant (2025504). VD is supported by a Lenity Foundation Fellowship. DJH is the editor of the osteoarthritis section for *UpToDate* and co-editor in chief of *Osteoarthritis and Cartilage*. DJH provides consulting advice on scientific advisory boards for Lilly, Sanofi, Novartis, Haleon, and TissueGene. All other authors declare that there are no conflicts of interest in relation to this work.

Multimedia Appendix 1
Behavior change techniques applied in notifications for the OA Coach app.
[DOCX File, 27 KB - jmir_v28i1e83507_app1.docx]

Multimedia Appendix 2
Radar charts of notification scores by domain and evaluation criteria comparing the consumer group with the health professional and researcher group. Six radar charts showing median notification scores across 6 domains: step count, sleep goals, logging pain scores, goal achievement, completion of surveys (sleep and mood), and wearing an activity tracker. Step count, sleep goals, logging pain scores, and tracker domains were rated on 4 criteria (understanding, usefulness, encouragement, and ability to prompt

XSL•FO
RenderX

action). Goal achievement and survey completion were rated on 3 of the criteria. Results are shown for consumers (blue) and health professionals/researchers (orange). Scores range from 0 (center) to 5 (outer edge), with higher values indicating more positive evaluations. For goal achievement notifications, both groups gave identical ratings.
[PNG File, 258 KB - jmir_v28i1e83507_app2.png]

Multimedia Appendix 3
Focus group plan and open-ended questions.
[DOCX File, 32 KB - jmir_v28i1e83507_app3.docx]

Multimedia Appendix 4
Education message topics for the OA Coach app.
[DOCX File, 27 KB - jmir_v28i1e83507_app4.docx]

Checklist 1
STROBE checklist.
[DOCX File, 71 KB - jmir_v28i1e83507_app5.docx]

References

1. Steinmetz JD, Culbreth GT, Haile LM, et al. Global, regional, and national burden of osteoarthritis, 1990–2020 and projections to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet Rheumatol 2023 Sep;5(9):e508-e522. [doi: 10.1016/S2665-9913(23)00163-7]
2. Black RJ, Cross M, Haile LM, et al. Global, regional, and national burden of rheumatoid arthritis, 1990–2020, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. Lancet Rheumatol 2023 Oct;5(10):e594-e610. [doi: 10.1016/S2665-9913(23)00211-4]
3. Turkiewicz A, Petersson IF, Björk J, et al. Current and future impact of osteoarthritis on health care: a population-based study with projections to year 2032. Osteoarthr Cartil 2014 Nov;22(11):1826-1832. [doi: 10.1016/j.joca.2014.07.015]
4. Dahlberg LE, Dell'Isola A, Lohmander LS, Nero H. Improving osteoarthritis care by digital means – effects of a digital self-management program after 24- or 48-weeks of treatment. PLOS One 2020;15(3):e0229783. [doi: 10.1371/journal.pone.0229783] [Medline: 32130276]
5. Litwic A, Edwards MH, Dennison EM, Cooper C. Epidemiology and burden of osteoarthritis. Br Med Bull 2013;105:185-199. [doi: 10.1093/bmb/lds038] [Medline: 23337796]
6. Bannuru RR, Osani MC, Vaysbrot EE, et al. OARSI guidelines for the non-surgical management of knee, hip, and polyarticular osteoarthritis. Osteoarthr Cartil 2019 Nov;27(11):1578-1589. [doi: 10.1016/j.joca.2019.06.011]
7. Wood G, Neilson J, Cottrell E, Hoole SP, Guideline Committee. Osteoarthritis in people over 16: diagnosis and management—updated summary of NICE guidance. BMJ 2023 Jan 24;380:24. [doi: 10.1136/bmj.p24] [Medline: 36693668]
8. O'Brien D, Pigg W, Ellis R, et al. An evidence-informed model of care for people with lower limb osteoarthritis in New Zealand. N Z J Physiother 2023;49(1):24-30. [doi: 10.15619/NZJP/49.1.04]
9. Fernandes L, Hagen KB, Bijlsma JWJ, et al. EULAR recommendations for the non-pharmacological core management of hip and knee osteoarthritis. Ann Rheum Dis 2013 Jul;72(7):1125-1135. [doi: 10.1136/annrheumdis-2012-202745] [Medline: 23595142]
10. Holden MA, Callaghan M, Felson D, et al. Clinical and cost-effectiveness of bracing in symptomatic knee osteoarthritis management: protocol for a multicentre, primary care, randomised, parallel-group, superiority trial. BMJ Open 2021 Mar 26;11(3):e048196. [doi: 10.1136/bmjopen-2020-048196] [Medline: 33771832]
11. Pisters MF, Veenhof C, van Meeteren NLU, et al. Long-term effectiveness of exercise therapy in patients with osteoarthritis of the hip or knee: a systematic review. Arthritis Rheum 2007 Oct 15;57(7):1245-1253. [doi: 10.1002/art.23009] [Medline: 17907210]
12. Fransen M, McConnell S, Harmer AR, Van der Esch M, Simic M, Bennell KL. Exercise for osteoarthritis of the knee: a Cochrane systematic review. Br J Sports Med 2015 Dec;49(24):1554-1557. [doi: 10.1136/bjsports-2015-095424] [Medline: 26405113]
13. Fransen M, McConnell S, Hernandez-Molina G, Reichenbach S. Exercise for osteoarthritis of the hip. Cochrane Database Syst Rev 2014 Apr 22;2014(4):CD007912. [doi: 10.1002/14651858.CD007912.pub2] [Medline: 24756895]
14. Allen KD, Huffman K, Cleveland RJ, et al. Evaluating osteoarthritis management programs: outcome domain recommendations from the OARSI Joint Effort Initiative. Osteoarthr Cartil 2023 Jul;31(7):954-965. [doi: 10.1016/j.joca.2023.02.078]
15. Svege I, Nordsletten L, Fernandes L, Risberg MA. Exercise therapy may postpone total hip replacement surgery in patients with hip osteoarthritis: a long-term follow-up of a randomised trial. Ann Rheum Dis 2015 Jan;74(1):164-169. [doi: 10.1136/annrheumdis-2013-203628] [Medline: 24255546]

XSL•FO
RenderX

16. Ackerman IN, Skou ST, Roos EM, et al. Implementing a national first-line management program for moderate-severe knee osteoarthritis in Australia: a budget impact analysis focusing on knee replacement avoidance. Osteoarthr Cartil Open 2020 Sep;2(3):100070. [doi: 10.1016/j.ocarto.2020.100070] [Medline: 36474677]
17. Teoh LSG, Eyles JP, Makovey J, Williams M, Kwoh CK, Hunter DJ. Observational study of the impact of an individualized multidisciplinary chronic care program for hip and knee osteoarthritis treatment on willingness for surgery. Int J Rheum Dis 2017 Oct;20(10):1383-1392. [doi: 10.1111/1756-185X.12950] [Medline: 27943562]
18. Brand CA, Harrison C, Tropea J, Hinman RS, Britt H, Bennell K. Management of osteoarthritis in general practice in Australia. Arthritis Care Res (Hoboken) 2014 Apr;66(4):551-558. [doi: 10.1002/acr.22197]
19. Runciman WB, Hunt TD, Hannaford NA, et al. CareTrack: assessing the appropriateness of health care delivery in Australia. Med J Aust 2012 Jul 16;197(2):100-105. [doi: 10.5694/mja12.10510] [Medline: 22794056]
20. Bunzli S, O'Brien P, Ayton D, et al. Misconceptions and the acceptance of evidence-based nonsurgical interventions for knee osteoarthritis: a qualitative study. Clin Orthop Relat Res 2019 Sep;477(9):1975-1983. [doi: 10.1097/CORR.0000000000000784] [Medline: 31192807]
21. Hinman RS, Nicolson PJA, Dobson FL, Bennell KL. Use of nondrug, nonoperative interventions by community‐dwelling people with hip and knee osteoarthritis. Arthritis Care Res (Hoboken) 2015 Feb;67(2):305-309. [doi: 10.1002/acr.22395]
22. Safari R, Jackson J, Sheffield D. Digital self-management interventions for people with osteoarthritis: systematic review with meta-analysis. J Med Internet Res 2020 Jul 20;22(7):e15365. [doi: 10.2196/15365] [Medline: 32706657]
23. Choi W, Zheng H, Franklin P, Tulu B. mHealth technologies for osteoarthritis self-management and treatment: a systematic review. Health Informatics J 2019 Sep;25(3):984-1003. [doi: 10.1177/1460458217735676] [Medline: 29090628]
24. Stevenson RDM, Chowdhury EA, Inza VB, et al. Development of the intelligent knee osteoarthritis lifestyle app: a person-based approach. BMC Musculoskelet Disord 2024;25(1):189. [doi: 10.1186/s12891-024-07313-4]
25. Hensley CP, Witte MM, Cai J, et al. Assessment of mobile health applications for management of knee and/or hip osteoarthritis using the mobile application rating scale. J Clin Rheumatol 2023 Aug 1;29(5):245-253. [doi: 10.1097/RHU.0000000000001896] [Medline: 36256541]
26. Global strategy on digital health 2020-2025. World Health Organization. 2021 Aug 18.
27. Duong V, Bloul N, Bryce K, et al. Development and feasibility of the "OA Coach" mobile application to support osteoarthritis management. Osteoarthr Cartil 2025 Apr;33:S179. [doi: 10.1016/j.joca.2025.02.249]
28. Willett M, Duda J, Fenton S, Gautrey C, Greig C, Rushton A. Effectiveness of behaviour change techniques in physiotherapy interventions to promote physical activity adherence in lower limb osteoarthritis patients: a systematic review. PLOS One 2019;14(7):e0219482. [doi: 10.1371/journal.pone.0219482] [Medline: 31291326]
29. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010 Feb 17;12(1):e4. [doi: 10.2196/jmir.1376] [Medline: 20164043]
30. Spring B, Champion KE, Acabchuk R, Hennessy EA. Self-regulatory behaviour change techniques in interventions to promote healthy eating, physical activity, or weight loss: a meta-review. Health Psychol Rev 2021 Dec;15(4):508-539. [doi: 10.1080/17437199.2020.1721310] [Medline: 31973666]
31. Redfern J, Thiagalingam A, Jan S, et al. Development of a set of mobile phone text messages designed for prevention of recurrent cardiovascular events. Eur J Prev Cardiolog 2014 Apr;21(4):492-499. [doi: 10.1177/2047487312449416]
32. Fritsch CG, Ferreira PH, Prior JL, et al. TEXT4myBACK – the development process of a self-management intervention delivered via text message for low back pain. Arch Rehabil Res Clin Transl 2021 Jun;3(2):100128. [doi: 10.1016/j.arrct.2021.100128] [Medline: 34179764]
33. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. Ann Intern Med 2007 Oct 16;147(8):573-577. [doi: 10.7326/0003-4819-147-8-200710160-00010]
34. Michie S. Designing and implementing behaviour change interventions to improve population health. J Health Serv Res Policy 2008 Oct;13(3_suppl):64-69. [doi: 10.1258/jhsrp.2008.008014] [Medline: 18806194]
35. Hawker GA, Mian S, Kendzerska T, French M. Measures of adult pain: Visual Analog Scale for Pain (VAS Pain), Numeric Rating Scale for Pain (NRS Pain), McGill Pain Questionnaire (MPQ), Short‐Form McGill Pain Questionnaire (SF‐MPQ), Chronic Pain Grade Scale (CPGS), Short Form‐36 Bodily Pain Scale (SF‐36 BPS), and Measure of Intermittent and Constant Osteoarthritis Pain (ICOAP). Arthritis Care Res (Hoboken) 2011 Nov;63(S11):S240-S252. [doi: 10.1002/acr.20543]
36. Lovibond PF, Lovibond SH. The structure of negative emotional states: comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. Behav Res Ther 1995 Mar;33(3):335-343. [doi: 10.1016/0005-7967(94)00075-u] [Medline: 7726811]
37. Antony MM, Bieling PJ, Cox BJ, Enns MW, Swinson RP. Psychometric properties of the 42-item and 21-item versions of the Depression Anxiety Stress Scales in clinical groups and a community sample. Psychol Assess 1998;10(2):176-181. [doi: 10.1037/1040-3590.10.2.176]
38. Buysse DJ, Reynolds CF III, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res 1989 May;28(2):193-213. [doi: 10.1016/0165-1781(89)90047-4]

XSL•FO
RenderX

39. Zhu D, Zhao J, Wang M, et al. Rehabilitation applications based on behavioral therapy for people with knee osteoarthritis: systematic review. JMIR Mhealth Uhealth 2024;12:e53798. [doi: 10.2196/53798]
40. Zhu Y, Long Y, Wang H, Lee KP, Zhang L, Wang SJ. Digital behavior change intervention designs for habit formation: systematic review. J Med Internet Res 2024;26:e54375. [doi: 10.2196/54375]
41. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462. [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
42. Kanavaki AM, Rushton A, Efstathiou N, et al. Barriers and facilitators of physical activity in knee and hip osteoarthritis: a systematic review of qualitative evidence. BMJ Open 2017 Dec;7(12):e017042. [doi: 10.1136/bmjopen-2017-017042]
43. Bennell K, Nelligan RK, Schwartz S, et al. Behavior change text messages for home exercise adherence in knee osteoarthritis: randomized trial. J Med Internet Res 2020;22(9):e21749. [doi: 10.2196/21749]
44. Nelligan RK, Hinman RS, Kasza J, Crofts SJC, Bennell KL. Effects of a self-directed web-based strengthening exercise and physical activity program supported by automated text messages for people with knee osteoarthritis: a randomized clinical trial. JAMA Intern Med 2021 Jun 1;181(6):776-785. [doi: 10.1001/jamainternmed.2021.0991] [Medline: 33843948]
45. Nelligan RK, Hinman RS, Atkins L, Bennell KL. A short message service intervention to support adherence to home-based strengthening exercise for people with knee osteoarthritis: intervention design applying the behavior change wheel. JMIR Mhealth Uhealth 2019 Oct 18;7(10):e14619. [doi: 10.2196/14619] [Medline: 31628786]
46. Michie S, Richardson M, Johnston M, et al. The Behavior Change Technique Taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
47. Australian Government Style Manual. Australian Government. URL: https://www.stylemanual.gov.au/accessible-and-inclusive-content/literacy-and-access [accessed 2026-04-13]
48. Our health literacy editor has been launched!. Sydney Health Literacy Lab. URL: https://www.sydneyhealthliteracylab.org.au/health-literacy-editor [accessed 2026-04-24]
49. Ayre J, Bonner C, Muscat DM, et al. Multiple automated health literacy assessments of written health information: development of the SHeLL (Sydney Health Literacy Lab) Health Literacy Editor v1. JMIR Form Res 2023 Feb 14;7:e40645. [doi: 10.2196/40645] [Medline: 36787164]
50. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research Electronic Data Capture (REDCap) – a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
51. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
52. Gwinnutt JM, Wieczorek M, Balanescu A, et al. 2021 EULAR recommendations regarding lifestyle behaviours and work participation to prevent progression of rheumatic and musculoskeletal diseases. Ann Rheum Dis 2023 Jan;82(1):48-56. [doi: 10.1136/annrheumdis-2021-222020]
53. French SD, Bennell KL, Nicolson PJA, Hodges PW, Dobson FL, Hinman RS. What do people with knee or hip osteoarthritis need to know? An international consensus list of essential statements for osteoarthritis. Arthritis Care Res (Hoboken) 2015 May;67(6):809-816. [doi: 10.1002/acr.22518] [Medline: 25418120]
54. Abraham C, Michie S. A taxonomy of behavior change techniques used in interventions. Health Psychol 2008 May;27(3):379-387. [doi: 10.1037/0278-6133.27.3.379] [Medline: 18624603]
55. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implementation Sci 2011 Dec;6(1):42. [doi: 10.1186/1748-5908-6-42]
56. Flesch-Kincaid calculator. Good Calculators. URL: https://goodcalculators.com/flesch-kincaid-calculator/ [accessed 2026-04-13]
57. Gill P, Stewart K, Treasure E, Chadwick B. Methods of data collection in qualitative research: interviews and focus groups. Br Dent J 2008 Mar;204(6):291-295. [doi: 10.1038/bdj.2008.192]
58. Gill P, Baillie J. Interviews and focus groups in qualitative research: an update for the digital age. Br Dent J 2018 Oct 5;225(7):668-672. [doi: 10.1038/sj.bdj.2018.815] [Medline: 30287965]
59. Trischler J, Dietrich T, Rundle-Thiele S. Co-design: from expert- to user-driven ideas in public service design. Public Manag Rev 2019 Nov 2;21(11):1595-1619. [doi: 10.1080/14719037.2019.1619810]
60. Bowden JL, Hunter DJ, Mills K, et al. The OARSI Joint Effort Initiative: priorities for osteoarthritis management program implementation and research 2024–2028. Osteoarthr Cartil Open 2023 Dec;5(4):100408. [doi: 10.1016/j.ocarto.2023.100408]
61. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies. Qual Health Res 2016 Nov;26(13):1753-1760. [doi: 10.1177/1049732315617444]
62. Health Consumers NSW. URL: https://hcnsw.org.au [accessed 2026-04-13]
63. Michie S, Atkins L, West R. The Behaviour Change Wheel: A Guide to Designing Interventions: Silverback Publishing; 2014.
64. Yardley L, Spring BJ, Riper H, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]

XSL•FO
RenderX

65. van Gemert-Pijnen JEWC, Nijland N, van Limburg M, et al. A holistic framework to improve the uptake and impact of eHealth technologies. J Med Internet Res 2011 Dec 5;13(4):e111. [doi: 10.2196/jmir.1672] [Medline: 22155738]
66. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. J Med Internet Res 2017 Jun 29;19(6):e232. [doi: 10.2196/jmir.7126] [Medline: 28663162]
67. Herbert MS, Goodin BR, Pero ST 4th, et al. Pain hypervigilance is associated with greater clinical pain severity and enhanced experimental pain sensitivity among adults with symptomatic knee osteoarthritis. Ann Behav Med 2014 Aug;48(1):50-60. [doi: 10.1007/s12160-013-9563-x] [Medline: 24352850]
68. Bunzli S, Taylor N, O'Brien P, et al. How do people communicate about knee osteoarthritis? A discourse analysis. Pain Med 2021 May 21;22(5):1127-1148. [doi: 10.1093/pm/pnab012] [Medline: 33502513]
69. Bunzli S, Taylor NF, O'Brien P, et al. Broken machines or active bodies? Part 1. Ways of talking about health and why it matters. J Orthop Sports Phys Ther 2023 May;53(5):236-238. [doi: 10.2519/jospt.2023.11879]
70. Bunzli S, Taylor NF, O'Brien P, et al. Broken machines or active bodies? Part 2. How people talk about osteoarthritis and why clinicians need to change the conversation. J Orthop Sports Phys Ther 2023 Jun;53(6):325-330. [doi: 10.2519/jospt.2023.11880]
71. Bunzli S, Taylor NF, O'Brien P, et al. Broken machines or active bodies? Part 3. Five recommendations to shift the way clinicians communicate with people who are seeking care for osteoarthritis. J Orthop Sports Phys Ther 2023 Jul;53(7):375-380. [doi: 10.2519/jospt.2023.11881]
72. Caneiro JP, O'Sullivan PB, Roos EM, et al. Three steps to changing the narrative about knee osteoarthritis care: a call to action. Br J Sports Med 2020 Mar;54(5):256-258. [doi: 10.1136/bjsports-2019-101328] [Medline: 31484634]
73. Jinks C, Botto-van Bemden A, Bunzli S, et al. Changing the narrative on osteoarthritis: a call for global action. Osteoarthr Cartil 2024 Apr;32(4):414-420. [doi: 10.1016/j.joca.2024.02.004]
74. Behera NS, Duong V, Eyles J, et al. How does osteoarthritis education influence knowledge, beliefs, and behavior in people with knee and hip osteoarthritis? A systematic review. Arthritis Care Res 2024;76(11):1511-1531. [doi: 10.1002/acr.25391]
75. Osteoarthritis in over 16s: diagnosis and management. NICE. 2022 Oct 19. URL: https://www.nice.org.uk/guidance/ng226 [accessed 2026-04-13]
76. Sørensen K, Pelikan JM, Röthlin F, et al. Health literacy in Europe: comparative results of the European Health Literacy Survey (HLS-EU). Eur J Public Health 2015 Dec;25(6):1053-1058. [doi: 10.1093/eurpub/ckv043]
77. Yuan S, Bowden JL, Eyles JP, et al. Protocol for the comparison of remote versus face-to-face osteoarthritis management programs (COASTAL): a randomized implementation trial. Osteoarthr Cartil Open 2025 Sep;7(3):100647. [doi: 10.1016/j.ocarto.2025.100647] [Medline: 40697621]

Abbreviations

BCT: behavior change technique
COM-B: capability, opportunity, motivation—behavior model
DASS-21: 21-Item Depression, Anxiety, and Stress Scale
JITAI: just-in-time adaptive intervention
mHealth: mobile health
OA: osteoarthritis
OACRG: Osteoarthritis Clinical Research Group
PSQI: Pittsburgh Sleep Quality Index
REDCap: Research Electronic Data Capture
SHeLL: Sydney Health Literacy Lab Editor
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Stone; submitted 09.Sep.2025; peer-reviewed by H Master, ME Maitland; revised version received 03.Mar.2026; accepted 09.Mar.2026; published 04.May.2026.

Please cite as:
Bryce KL, Bowden JL, Hunter DJ, Huang R, Bloul N, Duong V, Eyles J
Messages and Notifications for the "OA Coach" Knee Osteoarthritis Self-Management Mobile App: Codevelopment and Evaluation Using a Participatory Research Design With Focus Groups and Surveys
J Med Internet Res 2026;28:e83507
URL: https://www.jmir.org/2026/1/e83507
doi:10.2196/83507

XSL•FO
RenderX

© Kate L Bryce, Jocelyn L Bowden, David J Hunter, Robin Huang, Naomi Bloul, Vicky Duong, Jillian Eyles. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 4.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Characterizing Physical Activity Trajectories Preceding Incident Major Depressive Disorder Diagnosis With Consumer Wearable Devices in the All of Us Research Program: Retrospective Nested Case-Control Study

Yuezhou Zhang[1], PhD; Amos A Folarin[1,2,3,4,5], PhD; Hugh Logan Ellis[1], MBBCHIR, PhD; Rongrong Zhong[1,6], MD; Callum Stewart[1], PhD; Heet Sankesara[1], BSc; Hyunju Kim[1], PhD; Shaoxiong Sun[1,7], PhD; Abhishek Pratap[1,8,9], PhD; Richard JB Dobson[1,2,3,4,5], PhD

[1]Department of Biostatistics & Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, SGDP Centre, IoPPN, De Crespigny Park, Denmark Hill London, London, United Kingdom

[2]Institute of Health Informatics, University College London, London, United Kingdom

[3]NIHR Biomedical Research Centre at South London and Maudsley, NHS Foundation Trust, London, United Kingdom

[4]NIHR Biomedical Research Centre at University College London Hospitals, NHS Foundation Trust, London, United Kingdom

[5]Health Data Research UK, University College London, London, United Kingdom

[6]Clinical Research Center & Division of Mood Disorders, Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[7]Department of Computer Science, University of Sheffield, Sheffield, United Kingdom

[8]Boehringer Ingelheim Pharmaceuticals, Inc., Ridgefield, CT, United States

[9]Department of Biomedical and Health Informatics, University of Washington, Seattle, WA, United States

Corresponding Author:
Richard JB Dobson, PhD
Department of Biostatistics & Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, SGDP Centre, IoPPN, De Crespigny Park, Denmark Hill London, London, United Kingdom

Abstract

Background: Low physical activity (PA) is a well-established risk factor for major depressive disorder (MDD). However, the temporal dynamics of PA preceding an incident clinical diagnosis of MDD remain poorly characterized, particularly using long-term, objective measures collected in real-world settings.

Objective: This study aimed to characterize trajectories of wearable-measured PA during the year preceding incident MDD diagnosis and identify the timing of within-person changes.

Methods: We conducted a retrospective nested case-control study using linked electronic health record and wearable (Fitbit) data from the All of Us Research Program. Adults with at least 6 months of valid Fitbit PA data in the 12 months preceding diagnosis were included. Incident MDD cases were identified based on a first electronic health record–recorded diagnosis and matched to MDD-free controls on age, sex, BMI, and calendar time of diagnosis, with up to 4 controls per case. Daily steps and moderate to vigorous PA (MVPA) were aggregated into monthly averages. Linear mixed-effects models were used to compare prediagnostic PA trajectories between cases and controls over a retrospective time scale from −12 to 0 months. Among cases, within-person contrasts were used to identify when PA levels first showed statistically significant deviations relative to levels observed 12 months before diagnosis. Exploratory analyses assessed heterogeneity by demographic factors.

Results: The analytic cohort included 4104 participants (n=829, 20.2% incident MDD cases and n=3275, 79.8% matched controls; n=3355, 81.7% women; median age 48.4, IQR 36.3-61.3 years). Compared with controls, individuals who developed MDD exhibited consistently lower overall PA and significant downward trajectories in both daily steps and MVPA during the year preceding diagnosis (global trajectory tests; $P<.001$ for both outcomes). Differences widened progressively over time, indicating accelerating declines as diagnosis approached. Among cases, statistically significant changes in daily step counts emerged approximately 4 months before diagnosis (−145, 95% CI −253 to −37 steps vs month −12; $P=.02$) and reached −428 (95% CI −531 to −326) steps at diagnosis ($P<.001$). Declines in MVPA emerged approximately 5 months before diagnosis (−2.48, 95% CI −4.32 to −0.64 minutes; $P=.02$) and reached −5.61 (95% CI −7.35 to −3.86) minutes at diagnosis ($P<.001$). Furthermore, exploratory analyses suggested heterogeneity in prediagnostic trajectories across demographic subgroups, including steeper declines among men, more pronounced reductions in activity intensity at older ages, and persistently lower activity levels with flatter trajectories among individuals with obesity.

XSL•FO
RenderX

Conclusions: Unlike prior studies lacking objective PA assessment before MDD diagnosis, this study linked wearable and clinical data to characterize long-term prediagnostic trajectories in real-world settings. We observed sustained within-person declines emerging 4 to 5 months before diagnosis, providing insights into temporal dynamics preceding clinical recognition. These findings suggest that wearable-based monitoring may offer scalable early signals for risk stratification, prevention, and intervention for MDD.

(J Med Internet Res 2026;28:e93164) doi:10.2196/93164

KEYWORDS

major depressive disorder; depression; physical activity; wearable devices; mobile health; mHealth; longitudinal studies; case-control studies; electronic health records

Introduction

Major depressive disorder (MDD) is a leading cause of disability worldwide [1] and is associated with substantial adverse outcomes, including premature mortality [2], functional impairment [3], increased medical comorbidity [4], and suicide [5]. Given its substantial individual and societal burden [1], early identification of MDD is critical. Preventive strategies [6] and early interventions [7] may be most effective during the prodromal or initial stages. However, symptoms of MDD in the prodromal or early stages are often difficult to detect in routine clinical practice, contributing to delays in diagnosis [8,9]. This highlights the need for objective and scalable approaches to characterize changes preceding clinical recognition of MDD.

Low physical activity (PA) has been identified as a risk factor for MDD in prospective cohort studies and meta-analyses [10-12], with longitudinal evidence indicating that reductions in PA may precede worsening depressive symptoms [13]. These patterns may reflect the underlying biological and behavioral processes implicated in MDD, including alterations in psychomotor, inflammatory, and neuroendocrine functioning [14-16]. In this context, changes in PA may precede clinical recognition of MDD. However, much of the large-scale epidemiological existing evidence relies on self-reported PA, which is subject to recall bias [11], or on accelerometer-based assessments conducted over short durations or at infrequent follow-up time points [17]. Consequently, current studies provide limited insights into the timing and pattern of PA changes preceding a clinical MDD diagnosis. In particular, it remains unclear when PA begins to deviate from an individual's prior level and whether these changes vary across population subgroups. Addressing these questions requires longitudinal, objective measures of PA collected in real-world settings. Linking these data to clinical diagnostic information enables characterization of behavior changes prior to diagnosis as illness unfolds in routine care.

Advances in sensor technology and the widespread adoption of consumer wearable devices have enabled passive, continuous monitoring of real-world PA at scale. Devices such as Fitbit trackers capture daily step counts and activity intensity over extended periods with minimal user burden [18,19], representing key correlates of functioning and psychomotor activity [14]. Recent mobile health studies have demonstrated negative associations between objectively measured PA and depression severity [20-23], suggesting that such measures may complement symptom-based assessments by informing risk stratification and prompting earlier clinical evaluation. Nevertheless, prior mobile health studies have been limited by modest sample sizes, relatively short monitoring durations, or the absence of linkage to clinical diagnoses [20]. As a result, the temporal dynamics of objectively measured PA prior to clinical MDD diagnosis remain insufficiently characterized.

Few large-scale datasets integrate long-term wearable data with electronic health records (EHRs) in a manner that enables examination of prediagnostic PA trajectories and subgroup heterogeneity. The All of Us Research Program (AoURP) provides such an integrated resource, comprising a large, ongoing national cohort in which participants consent to share EHRs, health surveys, biospecimens, physical measurements, and wearable data [24]. By linking historical and prospective Fitbit data with EHR-based diagnoses, this study enables examination of long-term PA trajectories preceding clinical MDD diagnosis. The primary aim of this study was to characterize trajectories of wearable-measured PA during the 12 months preceding incident MDD diagnosis and identify the timing of within-person deviations from prior habitual PA levels. The secondary aim was to assess whether the patterns of prediagnostic PA changes vary across population subgroups.

Methods

Data Source

We used data from the AoURP, an ongoing national longitudinal cohort funded by the US National Institutes of Health [24] with the long-term goal of enrolling at least 1 million participants. The study design and data collection procedures have been described previously [24,25].

This analysis used the controlled tier dataset, version 8 (C2024Q3R8), including participants enrolled between May 2017 and October 2023. Participant demographics and baseline data were collected during the digital enrollment. For participants who consented to share EHR and Fitbit data, their historical (pre-enrollment) EHR and Fitbit data were made available through their participating health care provider organizations and linked Google Fitbit accounts, respectively [24-26]. In this data version, 36,614 individuals had linked EHR and Fitbit data available.

Ethical Considerations

This study was a secondary analysis of deidentified data obtained from the AoURP and did not involve direct interaction with human participants. In accordance with applicable

XSL•FO
RenderX

regulations, this work was determined to be exempt from additional institutional review board review as it used existing, deidentified research data. Access to deidentified data was restricted to authorized study investigators who completed required All of Us Responsible Conduct of Research training, and all analyses were conducted within the secure, cloud-based Researcher Workbench environment. In accordance with the AoURP Data and Statistics Dissemination Policy [27] analytic results for groups with fewer than 20 participants were not reported to minimize the risk of participant reidentification.

All participants in the AoURP provided informed consent at enrollment, including consent for future secondary analyses of their data. Data privacy and confidentiality were protected through multiple safeguards, including secure cloud-based data storage, restricted access to deidentified data, and mandatory confidentiality and data use agreements. Participant compensation of US $25 for the collection of biological specimens (eg, blood, saliva, or urine) was provided by the AoURP in the form of cash payments, gift cards, or electronic vouchers as applicable. No images or figures included in the manuscript or supplementary materials contain information that could identify individual participants.

Measures and Covariates

Fitbit PA Data

PA in this study was quantified using daily step counts and moderate to vigorous PA (MVPA), which intuitively capture overall activity volume [28] and the activity intensity emphasized in established public health guidelines [29], respectively. Fitbit classifies activity intensity into device-defined categories using metabolic equivalent of task–based criteria, consistent with intensity frameworks widely adopted in research-grade accelerometry [30,31]. In this study, daily MVPA (minutes per day) was calculated as the sum of device-labeled “fairly active” minutes plus twice the number of “very active” minutes, consistent with established definitions [32].

To ensure data quality, only days with at least 10 hours of wear time and daily step counts between 100 and 45,000 were considered valid following prior AoURP studies [33-35]. Monthly averages of daily step counts and MVPA minutes were calculated to reduce day-to-day variability related to missingness and short-term fluctuations, consistent with prior AoURP analyses [33-35]. Months with more than 10 valid days were considered valid and retained for analysis [35].

Incident MDD Definition

Following definitions used in prior AoURP research [33,36], incident MDD cases were defined as participants whose first recorded MDD diagnosis occurred during their Fitbit monitoring period, with no prior MDD diagnosis. Diagnoses were identified from EHR data using standardized concept identifiers from the Observational Medical Outcomes Partnership Common Data Model, which harmonize diagnoses across vocabularies including the Systematized Nomenclature of Medicine; *International Classification of Diseases, Ninth Revision, Clinical Modification*; and *International Classification of Diseases, Tenth Revision, Clinical Modification* [37].

Covariates

Age, sex, and BMI were included as covariates given their established associations with both PA and risk of MDD [38].

Inclusion and Exclusion

Participants were required to be aged 19 years or older at the time of diagnosis to ensure that they were adults (≥18 years) throughout the 1-year prediagnostic observation period. In addition, individuals with any recorded diagnosis of bipolar disorder, schizophrenia, or schizoaffective disorder prior to the diagnosis date were excluded. The Observational Medical Outcomes Partnership concept identifiers used for inclusion and exclusion, along with their corresponding *International Classification of Diseases, Ninth Revision, Clinical Modification*, and *International Classification of Diseases, Tenth Revision, Clinical Modification*, codes, are provided in Table S1 in Multimedia Appendix 1. To ensure sufficient prediagnostic wearable data, cases were required to have at least 6 valid months of Fitbit PA data within the 12 months preceding the diagnosis month.

Sampling and Matching Procedures

A nested case-control design [39-41] was used to construct the comparison group. Cases and controls were matched at the diagnosis month (hereafter referred to as the matching month) on age (within 1 year), sex, and BMI category, which were selected a priori given their strong associations with both PA and depression risk. BMI was defined using the measurement closest to 1 year before the matching month to minimize potential influence of emerging MDD on body weight. Eligible controls were selected from the risk set of participants who were aged 19 years or older at the matching month and had no recorded MDD diagnosis. Under this risk set sampling design, controls were not required to remain free of MDD during subsequent follow-up, consistent with standard nested case-control methodology [39-41]. Participants with a prior diagnosis of bipolar disorder, schizophrenia, or schizoaffective disorder were excluded. Controls were also required to meet the same Fitbit PA data availability criteria as cases. For each case, up to 4 controls were selected, a ratio chosen to improve statistical efficiency while maintaining matching quality and consistency with prior trajectory-based nested case-control studies [39,40]. When more than 4 eligible controls were available, controls were randomly sampled without replacement within or between cases to avoid repeated use of the same individuals and to simplify the correlation structure of the longitudinal analyses [40]. Additional sociodemographic factors were not included in matching to avoid overmatching and preserve statistical efficiency.

Statistical Analysis

Linear mixed-effects models with participant-specific random intercepts were used to estimate PA trajectories over time, allowing for the inclusion of all available observations from eligible participants and accommodating missing data under a missing at random assumption. Formal testing of the missing completely at random assumption was not informative as missingness patterns were highly correlated across PA variables (eg, step count and MVPA were often missing concurrently).

XSL•FO
RenderX

A retrospective time scale from −12 to 0 months was defined, with time 0 corresponding to the diagnosis month for cases and the matching month for controls. Models included case-control status (coded as 0 for controls and 1 for cases), linear and quadratic time terms (time and time squared), and their interaction terms, as well as matching variables. Quadratic terms were included to allow for potential nonlinear temporal patterns. Differences in trajectories between cases and controls were evaluated using time × case and time squared × case interaction terms, with a joint Wald test used to evaluate whether overall temporal patterns differed by case-control status. Model-derived marginal means were used to estimate and compare monthly trajectories between cases and controls. The statistical significance of between-group differences at each month was assessed using contrast *P* values, with *P* values adjusted for multiple comparisons using the Benjamini-Hochberg procedure [42].

Among individuals with incident MDD, marginal means were estimated for each month, and pairwise contrasts comparing each month with month −12 were used to identify when PA began to differ relative to the earliest month. To assess potential heterogeneity in prediagnostic PA trajectories, exploratory case-only models were extended to include interactions between time variables and subgroup indicators for sex, age, and BMI.

Sensitivity Analyses

To assess the robustness of our findings, we conducted prespecified sensitivity analyses. First, to evaluate whether device wear compliance differed between cases and controls, we examined trajectories of monthly valid Fitbit wear days over time. Linear mixed-effects models including time, group (case and controls), and their interaction were used to assess whether wear compliance declined differentially as diagnosis approached.

Second, to reduce potential confounding from intercurrent medical illness or injury, we excluded participants with newly recorded incident diagnoses of conditions known to affect PA within 1 year prior to diagnosis (or the matching month for controls), including musculoskeletal disorders (eg, fractures of the bone), acute cardiovascular events (eg, myocardial infarction, stroke, transient ischemic attack, and heart failure), and cancer (diagnostic codes are provided in Table S2 in Multimedia Appendix 1) and repeated the analyses in the restricted cohort.

Third, because incident MDD was defined based on the absence of prior EHR-documented diagnoses, the extent of available EHR history is an important consideration. Therefore, we quantified the duration of available EHR history prior to diagnosis (or matching month for controls), defined as the time between the first recorded EHR activity (eg, laboratory result, diagnosis, procedure, visit, or vital sign entry) and the MDD diagnosis date. To assess the robustness of our findings, we conducted sensitivity analyses restricted to participants (both cases and controls) with at least 5 and 10 years of available EHR history.

Results

Participant Characteristics

This study included 4104 participants, comprising 829 (20.2%) individuals with incident MDD and 3275 (79.8%) matched controls. Participant selection and cohort construction are shown in Figure 1. The median age of the overall cohort was 48.4 (IQR 36.3 - 61.3) years; 81.7% (n=3355) of the participants were women, and 82.5% (n=3384) were White individuals. Cases and controls had similar distributions of age, sex, BMI, race, and ethnicity, with all standardized mean differences below 0.1, indicating adequate balance between groups (Table 1). Included participants had a mean of 11.05 (SD 2.41) months of available wearable data during the 1-year prediagnostic window (13 months including the diagnosis month), corresponding to an overall missingness rate of 15.0%. Among cases, the median duration of available EHR history prior to diagnosis was 9.5 (IQR 4.7 - 15.9) years.

XSL•FO
RenderX

Figure 1. Flowchart of participant selection and cohort construction for incident major depressive disorder (MDD) cases and matched controls in a nested case-control study using linked electronic health record (EHR) and Fitbit data from the All of Us Research Program.

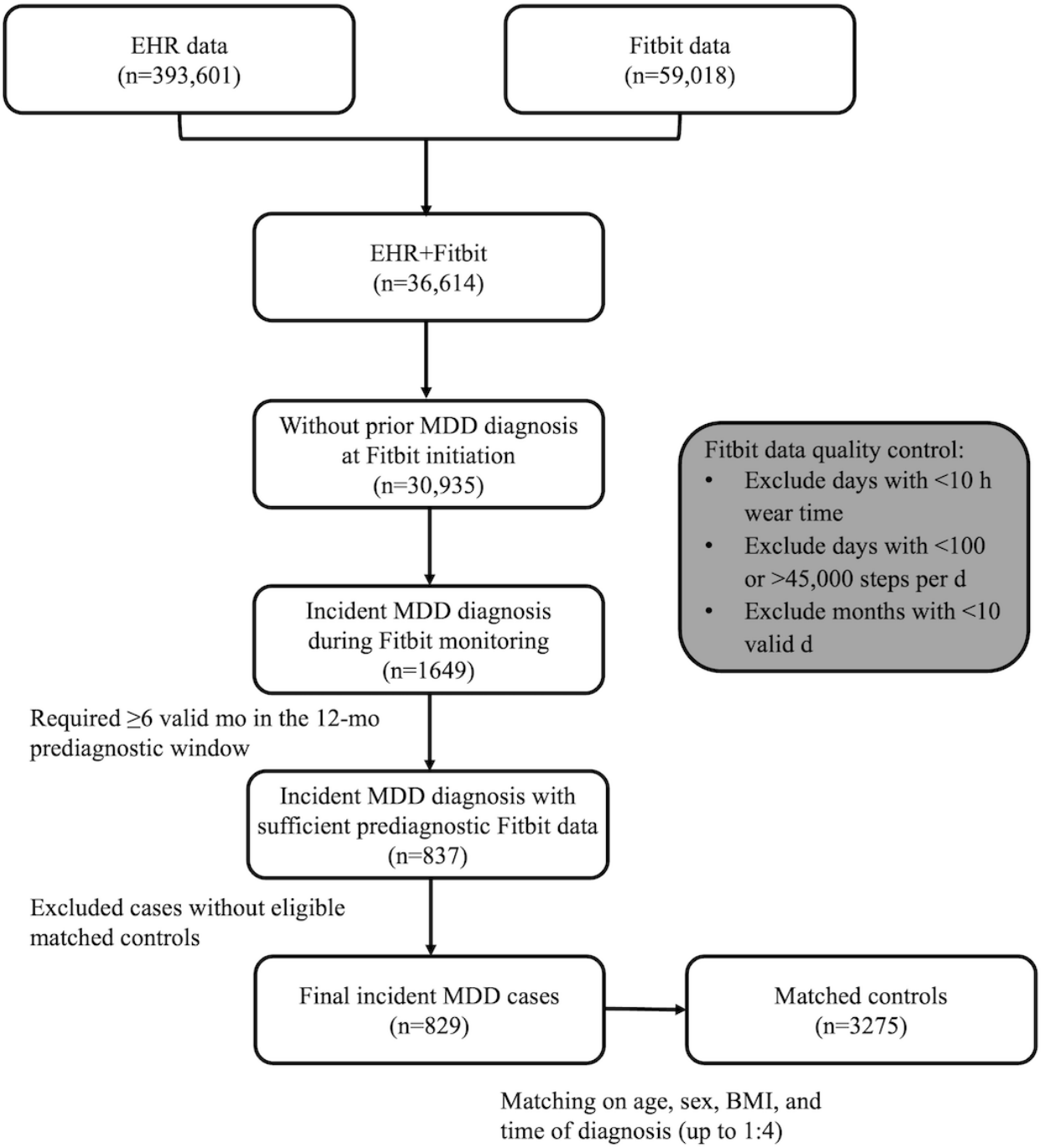

XSL•FO
RenderX

Table . Baseline characteristics of adults with incident major depressive disorder and matched controls in a nested case-control study using linked electronic health record and wearable data from the All of Us Research Program (N=4104).

Characteristics	Overall	Case (n=829)	Control (n=3275)	SMD[a]
Age[b] (y), median (IQR)	48.4 (36.3-61.3)	48.4 (36.2-61.2)	48.4 (36.3-61.4)	0.004
BMI[c] (kg/m^2), median (IQR)	29.9 (25.6-34.9)	29.8 (25.6-35.4)	29.9 (25.7-34.7)	0.046
Women, n (%)	3355 (81.7)	676 (81.5)	2679 (81.8)	0.033
Race, n (%)				0.098
Black or African American	271 (6.6)	40 (4.8)	231 (7.1)	
White	3384 (82.5)	691 (83.4)	2693 (82.2)	
Other[d]	449 (10.9)	98 (11.8)	351 (10.7)	
Ethnicity, n (%)				0.025
Hispanic or Latino	267 (6.5)	56 (6.8)	211 (6.4)	
Not Hispanic or Latino[e]	3837 (93.5)	773 (93.2)	3064 (93.6)	

[a]SMD: standardized mean difference. SMDs are presented as absolute values; values of less than 0.1 indicate negligible imbalance between groups.

[b]Age was defined at the diagnosis month for cases and the matched index month for controls.

[c]BMI was derived from measurements closest to 1 year before the diagnosis or matching month to minimize potential influence of emerging major depressive disorder on body weight.

[d]Participants with missing, unknown, or less frequently reported race categories were grouped into the "other" category in accordance with All of Us data reporting policies.

[e]Participants with missing or unknown ethnicity were grouped into the "Not Hispanic or Latino" category in accordance with All of Us data reporting policies.

PA Trajectories Preceding Incident MDD Diagnosis

The trajectories of daily step counts during 1 year preceding the matching month differed significantly between incident MDD cases and matched controls (Figure 2A). In mixed-effects models, controls exhibited a relatively stable trajectory, whereas cases showed a marked and accelerating decline as the diagnosis month approached. These differences were supported by significant linear and quadratic interaction terms (time × case: β=−92.08, 95% CI −126.14 to −58.02, *P*<.001; time squared × case: β=−5.47, 95% CI −8.23 to −2.70, *P*<.001), with a significant global test for trajectory differences (joint Wald test: *P*<.001; Table 2). Model-derived marginal means indicated that daily step counts were consistently lower among cases than controls at every month examined. As early as 12 months before diagnosis, estimated daily step counts were 7568 (95% CI 7330 - 7806) steps among cases, compared with 8360 (95% CI 8240 - 8480) steps among controls (contrast difference: −792, 95% CI −1058 to −525 steps; *P*<.001). This case-control gap widened progressively over time, reaching −1109 (95% CI −1374 to −845; *P*<.001) steps in the matching month, with estimated daily step counts of 7140 (95% CI 6904 - 7376) steps among cases and 8249 (95% CI 8130 - 8369) steps among controls (Table S3 in Multimedia Appendix 1).

Figure 2. Prediagnostic trajectories of (A) daily step counts and (B) moderate to vigorous physical activity (MVPA) during 1 year before incident major depressive disorder (MDD) diagnosis in a nested case-control study from the All of Us Research Program**.** Lines represent model-estimated mean trajectories derived from linear mixed-effects models for cases and matched controls over time, with shaded areas indicating 95% CIs.

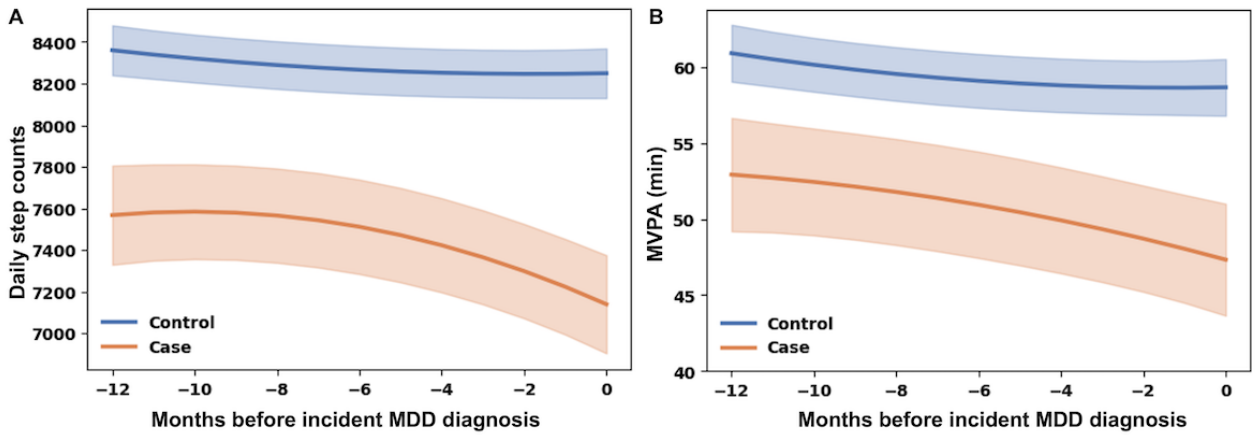

XSL•FO
RenderX

Table . Linear mixed-effects models comparing prediagnostic trajectories of daily step counts and moderate to vigorous physical activity (MVPA) between incident major depressive disorder cases and matched controls in the All of Us Research Program.

Term	β (95% CI)	*P* value	Global trajectory test, *P* value[a]
Daily step count model (unit: steps)			<.001
Intercept[b]	8249.57 (8130.24 to 8368.90)	<.001	
Group × case[c]	−1109.57 (−1373.76 to −845.39)	<.001	
Time[d]	3.66 (−11.58 to 18.90)	.64	
Time × case[e]	−92.08 (−126.14 to −58.02)	<.001	
Time squared	1.07 (−0.16 to 2.30)	.09	
Time squared × case[e]	−5.47 (−8.23 to −2.70)	<.001	
MVPA model (unit: min)			<.001
Intercept	58.69 (56.82 to 60.55)	<.001	
Group × case	−11.35 (−15.48 to −7.22)	<.001	
Time	0.05 (−0.23 to 0.32)	.75	
Time × case	−0.78 (−1.39 to −0.18)	.01	
Time squared	0.02 (−0.00 to 0.04)	.08	
Time squared × case	−0.04 (−0.09 to 0.01)	.10	

[a]The global trajectory test (joint Wald test) evaluated whether overall physical activity trajectories differed between cases and controls by jointly testing the interaction terms (time × case and time squared × case).

[b]The intercept represents the estimated mean physical activity level for the reference group (controls) at the reference time point (month 0).

[c]"Group × case" represents the baseline difference between incident major depressive disorder cases and matched controls (case-control status).

[d]Time was modeled on a retrospective monthly scale from −12 to 0 months relative to the diagnosis (or matching) month.

[e]"Time × case" and "time squared × case" represent differential linear and quadratic changes in trajectories between cases and controls over time.

Similar patterns were observed for daily MVPA trajectories (Figure 2B). Compared with controls, cases exhibited consistently lower daily MVPA levels throughout the prediagnostic period. In mixed-effects models, divergence in MVPA trajectories was supported by a significant linear interaction (time × case: β=−0.78, 95% CI −1.39 to −0.18, *P*=.01) and a significant global test for trajectory differences (joint Wald test: *P*<.001; Table 2). Model-derived marginal means showed that, at 12 months before diagnosis, estimated daily MVPA was 52.9 minutes among cases (95% CI 49.2 - 56.7) and 60.9 minutes among controls (95% CI 59.1 - 62.8) (contrast *P*<.001). This difference increased over time and reached −11.35 minutes (95% CI −15.48 to −7.22; *P*<.001) in the matching month, with estimated daily MVPA of 47.3 minutes among cases (95% CI 43.7 - 51.0) and 58.7 minutes among controls (95% CI 56.8 - 60.6; Table S4 in Multimedia Appendix 1).

Within-Person PA Changes Before Diagnosis

To characterize the timing of within-person changes in PA preceding diagnosis, we conducted case-only marginal contrast analyses using 12 months before diagnosis (time=−12 months) as the reference. Among cases, daily step counts did not differ significantly from the reference level until 4 months before diagnosis. At −4 months, step counts were significantly lower than at −12 months (−145 steps, 95% CI −253 to −37; *P*=.02). Reductions became larger as diagnosis approached, reaching −428 steps in the matching month (95% CI −531 to −326; *P*<.001; Table S5 in Multimedia Appendix 1).

In terms of daily MVPA, levels among cases were significantly lower than the reference level by 5 months before diagnosis (time=−5 months; −2.48 minutes, 95% CI −4.32 to −0.64; *P*=.02). MVPA levels continued to decline over the prediagnostic period, reaching −5.61 minutes in the matching month (95% CI −7.35 to −3.86; *P*<.001; Table S6 in Multimedia Appendix 1).

Subgroup Analyses

Case-only mixed-effects models revealed significant heterogeneity in prediagnostic PA trajectories by sex, age, and BMI (Figure 3; Tables S7 and S8 in Multimedia Appendix 1). Men exhibited steeper declines in both daily step counts and MVPA than women, supported by significant linear and quadratic time-by-sex interaction terms (joint Wald tests: *P*<.001).

XSL•FO
RenderX

Figure 3. Subgroup-specific prediagnostic trajectories of physical activity during the 12 months before incident major depressive disorder (MDD) diagnosis in the All of Us Research Program. Panels show daily step counts and moderate to vigorous physical activity (MVPA) stratified by sex (A and B), age (C and D), and BMI (E and F). Lines represent model-estimated mean trajectories derived from linear mixed-effects models, with shaded areas indicating 95% CIs.

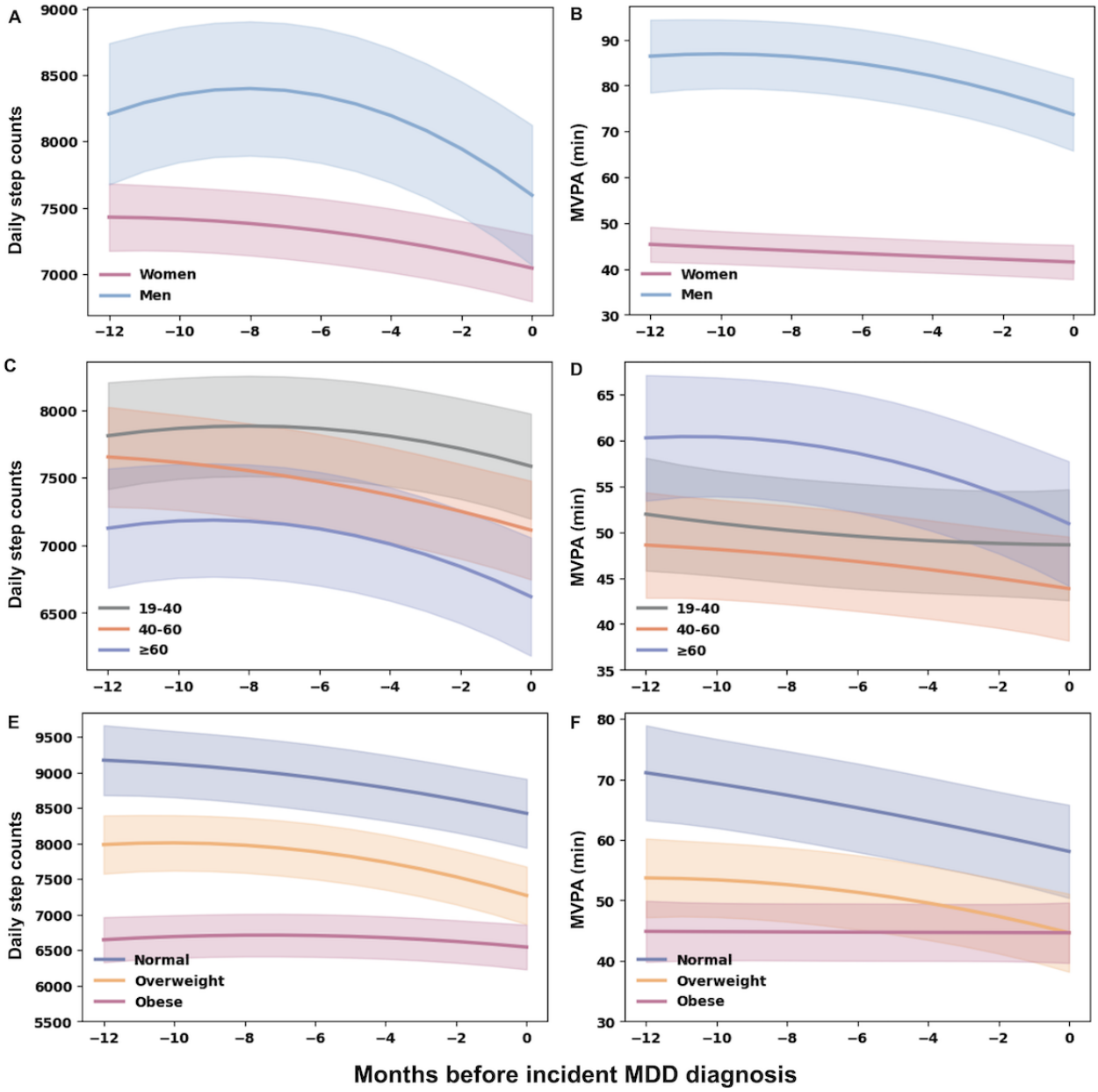

PA trajectories also varied significantly by age and BMI category (joint Wald tests: $P<.001$). Participants aged 60 years and older exhibited lower overall daily step counts and greater reductions in daily MVPA over time. Across BMI categories, participants with overweight and obesity had lower overall levels of daily steps and MVPA than those with normal BMI, and participants with obesity showed comparatively attenuated temporal declines in both outcomes (Figures 3E and 3F). The underweight group was excluded from subgroup analyses because of insufficient sample size.

Sensitivity Analyses

To assess whether device wear differed between cases and controls, we examined trajectories of monthly valid Fitbit wear days (Figure S1 in Multimedia Appendix 1). Wear days remained high and stable in both groups, with only a small between-group difference in average valid wear days per month (cases: 26.7, SD 5.3; controls: 27.0, SD 5.1 days per month). There was no evidence of a differential temporal decline in wear compliance (time × case: $\beta=-0.007$, 95% CI −0.032 to 0.018; $P=.56$).

Additional sensitivity analyses examined potential confounding from intercurrent medical conditions and variation in EHR history. Results were consistent with the primary analysis after excluding participants with newly recorded diagnoses of selected conditions that may affect PA within 1 year prior to diagnosis (n=713 in the restricted cohort; Table S9 in Multimedia Appendix 1). Similarly, results were consistent across strata of EHR history duration, including among participants with at least 5 and 10 years of EHR history prior to diagnosis (n=613

XSL•FO
RenderX

and n=319, respectively; Tables S10 and S11 in Multimedia Appendix 1).

Discussion

Principal Findings

In the AoURP cohort with linked wearable and EHR data, we found that adults who later received an incident MDD diagnosis exhibited a marked, progressively steepening decline in PA during the year preceding diagnosis compared with matched controls. Importantly, PA levels began to deviate from individuals' usual patterns approximately 4 to 5 months before clinical diagnosis. This suggests a clinically relevant window in which persistent within-person downward trajectories, rather than transient drops, may represent a more informative signal of emerging change. These findings suggest that longitudinal, objective PA monitoring may provide clinically interpretable insights into functioning and psychomotor behavior, although careful evaluation of potential confounding and subgroup heterogeneity is warranted.

Our findings extend prior evidence linking lower PA to subsequent MDD diagnosis. Prior prospective cohort studies [10-12,17], including work in the AoURP cohort [33], have consistently associated lower PA levels with a higher risk of incident MDD. This pattern was also observed in our analysis, with cases exhibiting lower PA levels than matched controls throughout the year preceding diagnosis. However, most prior studies have relied on self-reported measures or infrequent accelerometer assessments, limiting their ability to capture continuous, long-term, objectively measured PA trajectories prior to clinical diagnosis. In contrast, by using longitudinal wearable-measured PA data, this study provides additional insight by characterizing both between-group differences and within-person changes over time. Specifically, we identified two key patterns: (1) cases showed significantly sustained declines in PA relative to controls, and (2) within cases, PA deviated from habitual levels approximately 4 to 5 months before diagnosis.

Prodromal depressive features [43-45] such as fatigue, anhedonia, and sleep disturbance may partially contribute to these observed prediagnostic PA declines, which may, in turn, exacerbate mood and sleep through behavioral and biological pathways, potentially contributing to a self-reinforcing cycle [11,46]. However, because EHR-documented diagnosis likely occurs after symptom onset and care seeking [8], these findings should be interpreted with caution. The observed declines in PA may reflect prodromal processes, early symptomatic impairment, or both rather than implying causal or temporal primacy of PA changes in MDD development.

Clinically, the primary implication of longitudinal, objective PA monitoring is not to diagnose MDD but to provide an early, scalable prediagnostic signal of emerging vulnerability in routine care. Such signals could help trigger depression screening, preventive strategies, or early intervention. Although no established threshold currently defines a clinically actionable decline in PA for incident MDD, there is evidence from other clinical populations suggesting that changes of approximately 500 steps per day may be clinically or practically meaningful [47,48]. In this context, the approximately 400-step decline observed in our study is modest but approaches this range. However, PA is inherently variable and influenced by a range of external and behavioral factors [49,50]. As such, fixed thresholds may be susceptible to transient fluctuations arising from nonpathological causes. Instead, our findings highlight a clinically relevant window in which sustained, within-person declines over several months, reflected in both the magnitude and slope of change, may provide a more informative signal of emerging vulnerability. Furthermore, PA is both a potentially modifiable risk factor for depression [6,51,52] and a well-established component of interventions that reduce depressive symptoms among individuals with established MDD [53,54]. Therefore, regardless of whether observed PA declines reflect prodromal vulnerability or early symptomatic impairment, timely prevention or intervention may confer clinical benefit [10,11,53-56].

We also observed substantial heterogeneity in prediagnostic PA trajectories by sex, age, and BMI. Male participants experienced steeper declines in both step counts and MVPA than female participants. Several factors may contribute to this pattern, including higher average activity levels among men [57,58], delayed help seeking for mental health symptoms [59,60], and sex differences in the types and social contexts of PA [61]. Age-related differences were also evident: while lower step counts in older adults are consistent with established age-related mobility patterns, more pronounced reductions in MVPA may reflect greater sensitivity of higher-intensity activity to emerging depressive symptoms, functional limitations, or comorbid conditions in later life [62,63]. In addition, persistently lower activity and flatter trajectories among individuals with obesity may reflect barriers such as weight-related functional limitations, pain, or reduced motivation [64]. These interpretations remain exploratory and warrant further investigation.

These findings highlight the potential utility of integrating consumer wearables into routine clinical and public health workflows. Continuous, passive monitoring of PA may enable scalable identification of individuals exhibiting sustained, within-person declines over time, which could serve as early behavioral signals of emerging vulnerability to MDD. Such signals could be incorporated into digital health platforms or clinical decision support systems to support risk stratification, enable targeted prevention, and facilitate earlier intervention prior to formal diagnosis. At a population level, these approaches could enable earlier identification of individuals at higher risk; support more efficient allocation of health care resources; and, ultimately, reduce the broader societal burden of MDD [65,66].

Limitations

This study has several limitations. First, incident MDD was defined by the absence of EHR-documented diagnoses, consistent with prior AoURP methodologies [33,35,36], which may be influenced by the duration of available EHR history. Although sensitivity analyses restricted to longer EHR history yielded consistent results, left censoring and undiagnosed cases cannot be fully excluded. These are inherent limitations of EHR-based observational studies and warrant further

XSL•FO
RenderX

investigation using complementary data sources and methods. Second, seasonality may influence both PA patterns and the timing of MDD onset [67]. In our study, cases and controls were matched on calendar time (matching month), which helps mitigate seasonal effects on between-group comparisons. However, seasonal influences on within-person trajectories cannot be fully excluded. Future studies incorporating more detailed environmental data may help further clarify these effects. Third, selection bias is also possible as inclusion required participants to have sufficiently long Fitbit monitoring histories and meet prespecified data completeness criteria, potentially limiting representativeness. Fourth, the study population was predominantly White, and the bring-your-own-device nature of Fitbit data contribution may have skewed the analytic sample toward more health-conscious or technologically engaged participants, potentially limiting generalizability. Fifth, missing data were handled under a missing at random assumption; however, in real-world wearable data, missingness may be influenced by unmeasured behavioral or contextual factors and, thus, may not fully satisfy this assumption. This represents a common challenge in mobile health studies and warrants further methodological investigation in future research.

Conclusions

In this large cohort with linked wearable and EHR data, we characterized long-term, prediagnostic trajectories of objectively measured PA preceding incident MDD diagnosis. This addresses a key gap in prior research, which has largely lacked objective, longitudinal data on PA trajectories before clinical diagnosis. We identified sustained, within-person declines emerging 4 to 5 months before diagnosis, providing new insights into the temporal dynamics of behavior change prior to clinical recognition and defining a potential window for early detection. In real-world settings, these findings suggest that longitudinal monitoring of PA using consumer wearable devices may offer scalable and actionable signals to support risk stratification, targeted prevention, and earlier intervention for MDD. However, careful consideration of confounding and subgroup heterogeneity remains essential.

Acknowledgments

The authors gratefully acknowledge the All of Us participants for their contributions, without whom this research would not have been possible. They also thank the National Institutes of Health's All of Us Research Program for making available the participant data examined in this study. The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the Generative Artificial Intelligence Delegation Taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing. The GenAI tool used was GPT-4.5 (OpenAI). Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes. GPT-4.5 was used only for language polishing (improving grammar, clarity, and conciseness) during manuscript preparation. The authors reviewed and edited all artificial intelligence–assisted text and take full responsibility for the final content. This use complies with the JMIR Publications editorial policy on the use of artificial intelligence tools.

Funding

RJBD is supported by the following institutions: (1) National Institute for Health and Care Research (NIHR) Biomedical Research Centre (BRC) at South London and Maudsley National Health Service (NHS) Foundation Trust and King's College London; (2) Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh government), Public Health Agency (Northern Ireland), British Heart Foundation, and Wellcome Trust; (3) the BigData@Heart consortium, funded by the Innovative Medicines Initiative 2 Joint Undertaking (which receives support from the European Union's Horizon 2020 research and innovation program and the European Federation of Pharmaceutical Industries and Associations, in collaboration with 20 academic and industry partners and the European Society of Cardiology); (4) the NIHR University College London Hospitals BRC; (5) the NIHR BRC at South London and Maudsley (related to attendance at the American Medical Informatics Association) NHS Foundation Trust and King's College London; (6) the UK Research and Innovation London Medical Imaging & Artificial Intelligence Centre for Value-Based Healthcare; (7) the NIHR Applied Research Collaboration South London at King's College Hospital NHS Foundation Trust; and (8) the Wellcome Trust. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, and approval of the manuscript; or decision to submit the manuscript for publication.

Data Availability

The datasets analyzed during this study are not publicly available because individual-level participant data from the All of Us Research Program cannot be publicly released under program privacy policies. The data are available to approved researchers through the All of Us Researcher Workbench [68] following registration, completion of required training, and attestation of the data use agreement.

XSL•FO
RenderX

Authors' Contributions

YZ had full access to all study data and takes responsibility for the integrity of the data and the accuracy of the data analysis. YZ, AAF, and RJBD contributed to the study concept and design. Data acquisition, analysis, and interpretation were performed by YZ, AAF, AP, HLE, RZ, CS, HS, HK, SS, and RJBD. YZ drafted the manuscript and conducted the statistical analysis. Funding was obtained by RJBD and AAF. All authors critically reviewed the manuscript for important intellectual content.

Conflicts of Interest

AP is an employee of Boehringer Ingelheim Pharmaceuticals, Inc. AAF and RJBD are cofounders of Onsentia. AAF holds shares in Google. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Supplementary figure and tables presenting Fitbit wear-time trajectories, diagnostic code definitions, additional model estimates, subgroup analyses, and sensitivity analyses.
[DOCX File, 301 KB - jmir_v28i1e93164_app1.docx]

References

1. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct 17;396(10258):1204-1222. [doi: 10.1016/S0140-6736(20)30925-9] [Medline: 33069326]
2. Walker ER, McGee RE, Druss BG. Mortality in mental disorders and global disease burden implications: a systematic review and meta-analysis. JAMA Psychiatry 2015 Apr;72(4):334-341. [doi: 10.1001/jamapsychiatry.2014.2502] [Medline: 25671328]
3. Cambridge OR, Knight MJ, Mills N, Baune BT. The clinical relationship between cognitive impairment and psychosocial functioning in major depressive disorder: a systematic review. Psychiatry Res 2018 Nov;269:157-171. [doi: 10.1016/j.psychres.2018.08.033] [Medline: 30149273]
4. Vancampfort D, Correll CU, Galling B, et al. Diabetes mellitus in people with schizophrenia, bipolar disorder and major depressive disorder: a systematic review and large scale meta-analysis. World Psychiatry 2016 Jun;15(2):166-174. [doi: 10.1002/wps.20309] [Medline: 27265707]
5. Patel V, Chisholm D, Parikh R, et al. Addressing the burden of mental, neurological, and substance use disorders: key messages from Disease Control Priorities, 3rd edition. Lancet 2016 Apr;387(10028):1672-1685. [doi: 10.1016/S0140-6736(15)00390-6]
6. Firth J, Solmi M, Wootton RE, et al. A meta-review of “lifestyle psychiatry”: the role of exercise, smoking, diet and sleep in the prevention and treatment of mental disorders. World Psychiatry 2020 Oct;19(3):360-380. [doi: 10.1002/wps.20773] [Medline: 32931092]
7. Kraus C, Kadriu B, Lanzenberger R, Zarate CAJ, Kasper S. Prognosis and improved outcomes in major depression: a review. Transl Psychiatry 2019 Apr 3;9(1):127. [doi: 10.1038/s41398-019-0460-3] [Medline: 30944309]
8. Gurrieri R, Russomanno G, Gambini M, Pescini E, Sità G, Marazziti D. The duration of untreated illness in major depressive disorder: clinical relevance, predictors, and strategies for early intervention. Psychiatry Res Commun 2025 Dec;5(4):100231. [doi: 10.1016/j.psycom.2025.100231]
9. Young AS, Klap R, Sherbourne CD, Wells KB. The quality of care for depressive and anxiety disorders in the United States. Arch Gen Psychiatry 2001 Jan;58(1):55-61. [doi: 10.1001/archpsyc.58.1.55] [Medline: 11146758]
10. Schuch FB, Vancampfort D, Firth J, et al. Physical activity and incident depression: a meta-analysis of prospective cohort studies. Am J Psychiatry 2018 Jul 1;175(7):631-648. [doi: 10.1176/appi.ajp.2018.17111194] [Medline: 29690792]
11. Pearce M, Garcia L, Abbas A, et al. Association between physical activity and risk of depression: a systematic review and meta-analysis. JAMA Psychiatry 2022 Jun 1;79(6):550-559. [doi: 10.1001/jamapsychiatry.2022.0609] [Medline: 35416941]
12. Bizzozero-Peroni B, Díaz-Goñi V, Jiménez-López E, et al. Daily step count and depression in adults: a systematic review and meta-analysis. JAMA Netw Open 2024 Dec 2;7(12):e2451208. [doi: 10.1001/jamanetworkopen.2024.51208] [Medline: 39680407]
13. Azevedo Da Silva M, Singh-Manoux A, Brunner EJ, et al. Bidirectional association between physical activity and symptoms of anxiety and depression: the Whitehall II study. Eur J Epidemiol 2012 Jul;27(7):537-546. [doi: 10.1007/s10654-012-9692-8] [Medline: 22623145]
14. Wüthrich F, Nabb CB, Mittal VA, Shankman SA, Walther S. Actigraphically measured psychomotor slowing in depression: systematic review and meta-analysis. Psychol Med 2022 May;52(7):1208-1221. [doi: 10.1017/S0033291722000903] [Medline: 35550677]
15. Levihn-Coon A, Smits JAJ, Beevers CG. The longitudinal relations between physical activity, inflammation, and depression in the Health and Retirement Study. PLOS Ment Health 2025;2(4):e0000211. [doi: 10.1371/journal.pmen.0000211] [Medline: 41661985]

XSL•FO
RenderX

16. Mahindru A, Patil P, Agrawal V. Role of physical activity on mental health and well-being: a review. Cureus 2023 Jan;15(1):e33475. [doi: 10.7759/cureus.33475] [Medline: 36756008]
17. Qiu S, Xing Z. Dose-response relationship between accelerometer-measured physical activity and depression: evidence from the UK Biobank. Transl Psychiatry 2025 Aug 19;15:297. [doi: 10.1038/s41398-025-03543-9]
18. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annu Rev Clin Psychol 2017 May 8;13:23-47. [doi: 10.1146/annurev-clinpsy-032816-044949] [Medline: 28375728]
19. Zapata-Lamana R, Lalanza JF, Losilla JM, Parrado E, Capdevila L. mHealth technology for ecological momentary assessment in physical activity research: a systematic review. PeerJ 2020;8:e8848. [doi: 10.7717/peerj.8848] [Medline: 32257648]
20. De Angel V, Lewis S, White K, et al. Digital health tools for the passive monitoring of depression: a systematic review of methods. NPJ Digit Med 2022 Jan 11;5(1):3. [doi: 10.1038/s41746-021-00548-8] [Medline: 35017634]
21. Vallance JK, Winkler EA, Gardiner PA, Healy GN, Lynch BM, Owen N. Associations of objectively-assessed physical activity and sedentary time with depression: NHANES (2005-2006). Prev Med 2011 Oct;53(4-5):284-288. [doi: 10.1016/j.ypmed.2011.07.013] [Medline: 21820466]
22. Zhang Y, Stewart C, Ranjan Y, et al. Large-scale digital phenotyping: identifying depression and anxiety indicators in a general UK population with over 10,000 participants. J Affect Disord 2025 Apr;375:412-422. [doi: 10.1016/j.jad.2025.01.124]
23. Huang J, Zhang Y, Zhou D, et al. Leveraging low-cost app-based step count data to assess depression and anxiety in university students: a cross-sectional mobile health study. J Affect Disord 2026 Mar 15;397:120880. [doi: 10.1016/j.jad.2025.120880] [Medline: 41407112]
24. All of Us Research Program Investigators, Denny JC, Rutter JL, et al. The “All of Us” research program. N Engl J Med 2019 Aug 15;381(7):668-676. [doi: 10.1056/NEJMsr1809937] [Medline: 31412182]
25. Mayo KR, Basford MA, Carroll RJ, et al. The All of Us data and research center: creating a secure, scalable, and sustainable ecosystem for biomedical research. Annu Rev Biomed Data Sci 2023 Aug 10;6:443-464. [doi: 10.1146/annurev-biodatasci-122120-104825] [Medline: 37561600]
26. Master H, Kouame A, Marginean K, Basford M, Harris P, Holko M. How Fitbit data are being made available to registered researchers in All of Us Research Program. Pac Symp Biocomput 2023;28:19-30. [Medline: 36540961]
27. User support hub. All of Us Research Program. URL: https://support.researchallofus.org/hc/en-us [accessed 2026-04-28]
28. Tudor-Locke C, Craig CL, Brown WJ, et al. How many steps/day are enough? For adults. Int J Behav Nutr Phys Act 2011 Jul 28;8:79. [doi: 10.1186/1479-5868-8-79] [Medline: 21798015]
29. Bull FC, Al-Ansari SS, Biddle S, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. Br J Sports Med 2020 Dec;54(24):1451-1462. [doi: 10.1136/bjsports-2020-102955] [Medline: 33239350]
30. Semanik P, Lee J, Pellegrini CA, Song J, Dunlop DD, Chang RW. Comparison of physical activity measures derived from the Fitbit Flex and the ActiGraph GT3X+ in an employee population with chronic knee symptoms. ACR Open Rheumatol 2020 Jan;2(1):48-52. [doi: 10.1002/acr2.11099] [Medline: 31943970]
31. Brewer W, Swanson BT, Ortiz A. Validity of Fitbit’s active minutes as compared with a research-grade accelerometer and self-reported measures. BMJ Open Sport Exerc Med 2017;3(1):e000254. [doi: 10.1136/bmjsem-2017-000254] [Medline: 29018543]
32. Kozey SL, Lyden K, Howe CA, Staudenmayer JW, Freedson PS. Accelerometer output and MET values of common physical activities. Med Sci Sports Exerc 2010 Sep;42(9):1776-1784. [doi: 10.1249/MSS.0b013e3181d479f2] [Medline: 20142781]
33. Master H, Annis J, Huang S, et al. Association of step counts over time with the risk of chronic disease in the All of Us research program. Nat Med 2022 Nov;28(11):2301-2308. [doi: 10.1038/s41591-022-02012-w] [Medline: 36216933]
34. Brittain EL, Han L, Annis J, et al. Physical activity and incident obesity across the spectrum of genetic risk for obesity. JAMA Netw Open 2024 Mar 4;7(3):e243821. [doi: 10.1001/jamanetworkopen.2024.3821] [Medline: 38536175]
35. Yakdan S, Benedict B, Singh P, et al. Association of activity with the risk of developing musculoskeletal pain in the All of Us research program. J Pain 2025 Oct;35:105516. [doi: 10.1016/j.jpain.2025.105516] [Medline: 40774444]
36. Zheng NS, Annis J, Master H, et al. Sleep patterns and risk of chronic disease as measured by long-term monitoring with commercial wearable devices in the All of Us research program. Nat Med 2024 Sep;30(9):2648-2656. [doi: 10.1038/s41591-024-03155-8] [Medline: 39030265]
37. Klann JG, Joss MAH, Embree K, Murphy SN. Data model harmonization for the All Of Us research program: transforming i2b2 data into the OMOP common data model. PLoS ONE 2019;14(2):e0212463. [doi: 10.1371/journal.pone.0212463] [Medline: 30779778]
38. Akhtar-Danesh N, Landeen J. Relation between depression and sociodemographic factors. Int J Ment Health Syst 2007 Sep 4;1(1):4. [doi: 10.1186/1752-4458-1-4] [Medline: 18271976]
39. Wagner M, Helmer C, Tzourio C, Berr C, Proust-Lima C, Samieri C. Evaluation of the concurrent trajectories of cardiometabolic risk factors in the 14 years before dementia. JAMA Psychiatry 2018 Oct 1;75(10):1033-1042. [doi: 10.1001/jamapsychiatry.2018.2004] [Medline: 30043038]
40. Wu Z, Cribb L, Wolfe R, et al. Cardiometabolic trajectories preceding dementia in community-dwelling older individuals. JAMA Netw Open 2025 Feb 3;8(2):e2458591. [doi: 10.1001/jamanetworkopen.2024.58591] [Medline: 39918818]

XSL•FO
RenderX

41. Jia J, Ning Y, Chen M, et al. Biomarker changes during 20 years preceding Alzheimer's disease. N Engl J Med 2024 Feb 22;390(8):712-722. [doi: 10.1056/NEJMoa2310168] [Medline: 38381674]
42. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 1995 Jan;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]
43. Pede VB, Jaiswal SV, Sawant VA. Study of prodromal and residual symptoms of depression. Ind Psychiatry J 2017;26(2):121-127. [doi: 10.4103/ipj.ipj_19_18] [Medline: 30089957]
44. Iacoviello BM, Alloy LB, Abramson LY, Choi JY. The early course of depression: a longitudinal investigation of prodromal symptoms and their relation to the symptomatic course of depressive episodes. J Abnorm Psychol 2010 Aug;119(3):459-467. [doi: 10.1037/a0020114] [Medline: 20677835]
45. Benasi G, Fava GA, Guidi J. Prodromal symptoms in depression: a systematic review. Psychother Psychosom 2021;90(6):365-372. [doi: 10.1159/000517953] [Medline: 34350890]
46. Kredlow MA, Capozzoli MC, Hearon BA, Calkins AW, Otto MW. The effects of physical activity on sleep: a meta-analytic review. J Behav Med 2015 Jun;38(3):427-449. [doi: 10.1007/s10865-015-9617-6] [Medline: 25596964]
47. Gardner AW, Montgomery PS, Wang M, Shen B. Minimal clinically important differences in daily physical activity outcomes following supervised and home-based exercise in peripheral artery disease. Vasc Med 2022 Apr;27(2):142-149. [doi: 10.1177/1358863X211072913] [Medline: 35164605]
48. Feig EH, Harnedy LE, Celano CM, Huffman JC. Increase in daily steps during the early phase of a physical activity intervention for type 2 diabetes as a predictor of intervention outcome. Int J Behav Med 2021 Dec;28(6):834-839. [doi: 10.1007/s12529-021-09966-0] [Medline: 33575971]
49. Solomon E, Rees T, Ukoumunne OC, Metcalf B, Hillsdon M. Personal, social, and environmental correlates of physical activity in adults living in rural south-west England: a cross-sectional analysis. Int J Behav Nutr Phys Act 2013 Nov 21;10:129. [doi: 10.1186/1479-5868-10-129] [Medline: 24261335]
50. Humpel N, Owen N, Leslie E. Environmental factors associated with adults' participation in physical activity: a review. Am J Prev Med 2002 Apr;22(3):188-199. [doi: 10.1016/s0749-3797(01)00426-3] [Medline: 11897464]
51. Salazar de Pablo G, Solmi M, Vaquerizo-Serrano J, et al. Primary prevention of depression: an umbrella review of controlled interventions. J Affect Disord 2021 Nov 1;294:957-970. [doi: 10.1016/j.jad.2021.07.101] [Medline: 34375224]
52. Hallgren M, Stubbs B, Vancampfort D, Lundin A, Jääkallio P, Forsell Y. Treatment guidelines for depression: greater emphasis on physical activity is needed. Eur Psychiatry 2017 Feb;40:1-3. [doi: 10.1016/j.eurpsy.2016.08.011] [Medline: 27837666]
53. Schuch FB, Vancampfort D, Richards J, Rosenbaum S, Ward PB, Stubbs B. Exercise as a treatment for depression: a meta-analysis adjusting for publication bias. J Psychiatr Res 2016 Jun;77:42-51. [doi: 10.1016/j.jpsychires.2016.02.023] [Medline: 26978184]
54. Kandola A, Ashdown-Franks G, Hendrikse J, Sabiston CM, Stubbs B. Physical activity and depression: towards understanding the antidepressant mechanisms of physical activity. Neurosci Biobehav Rev 2019 Dec;107:525-539. [doi: 10.1016/j.neubiorev.2019.09.040] [Medline: 31586447]
55. Mammen G, Faulkner G. Physical activity and the prevention of depression: a systematic review of prospective studies. Am J Prev Med 2013 Nov;45(5):649-657. [doi: 10.1016/j.amepre.2013.08.001] [Medline: 24139780]
56. Teychenne M, Ball K, Salmon J. Physical activity and likelihood of depression in adults: a review. Prev Med 2008 May;46(5):397-411. [doi: 10.1016/j.ypmed.2008.01.009] [Medline: 18289655]
57. Caspersen CJ, Pereira MA, Curran KM. Changes in physical activity patterns in the United States, by sex and cross-sectional age. Med Sci Sports Exerc 2000;32(9):1601-1609. [doi: 10.1097/00005768-200009000-00013]
58. Guthold R, Stevens GA, Riley LM, Bull FC. Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1·9 million participants. Lancet Glob Health 2018 Oct;6:e1077-e1086.
59. Seidler ZE, Dawes AJ, Rice SM, Oliffe JL, Dhillon HM. The role of masculinity in men's help-seeking for depression: a systematic review. Clin Psychol Rev 2016 Nov;49:106-118. [doi: 10.1016/j.cpr.2016.09.002] [Medline: 27664823]
60. Yousaf O, Grunfeld EA, Hunter MS. A systematic review of the factors associated with delays in medical and psychological help-seeking among men. Health Psychol Rev 2015;9(2):264-276. [doi: 10.1080/17437199.2013.840954] [Medline: 26209212]
61. Azevedo MR, Araújo CL, Reichert FF, Siqueira FV, da Silva MC, Hallal PC. Gender differences in leisure-time physical activity. Int J Public Health 2007;52(1):8-15. [doi: 10.1007/s00038-006-5062-1] [Medline: 17966815]
62. Dohrn IM, Gardiner PA, Winkler E, Welmer AK. Device-measured sedentary behavior and physical activity in older adults differ by demographic and health-related factors. Eur Rev Aging Phys Act 2020;17:8. [doi: 10.1186/s11556-020-00241-x] [Medline: 32537028]
63. Lohne-Seiler H, Hansen BH, Kolle E, Anderssen SA. Accelerometer-determined physical activity and self-reported health in a population of older adults (65-85 years): a cross-sectional study. BMC Public Health 2014 Mar 27;14:284. [doi: 10.1186/1471-2458-14-284] [Medline: 24673834]
64. Baillot A, Chenail S, Barros Polita N, et al. Physical activity motives, barriers, and preferences in people with obesity: a systematic review. PLoS One 2021;16(6):e0253114. [doi: 10.1371/journal.pone.0253114] [Medline: 34161372]

XSL•FO
RenderX

65. Wang PS, Simon G, Kessler RC. The economic burden of depression and the cost-effectiveness of treatment. Int J Methods Psychiatr Res 2003;12(1):22-33. [doi: 10.1002/mpr.139] [Medline: 12830307]
66. Le LKD, Esturas AC, Mihalopoulos C, et al. Cost-effectiveness evidence of mental health prevention and promotion interventions: a systematic review of economic evaluations. PLoS Med 2021 May;18(5):e1003606. [doi: 10.1371/journal.pmed.1003606] [Medline: 33974641]
67. Zhang Y, Folarin AA, Ranjan Y, et al. Assessing seasonal and weather effects on depression and physical activity using mobile health data. Npj Ment Health Res 2025 Apr 18;4(1):11. [doi: 10.1038/s44184-025-00125-x] [Medline: 40251379]
68. All of Us Researcher Workbench. URL: https://workbench.researchallofus.org/login [accessed 2026-04-30]

Abbreviations

AoURP: All of Us Research Program
EHR: electronic health record
MDD: major depressive disorder
MVPA: moderate to vigorous physical activity
PA: physical activity

Edited by S Brini; submitted 09.Feb.2026; peer-reviewed by D Shin, EW Lampe; revised version received 14.Apr.2026; accepted 15.Apr.2026; published 04.May.2026.

Please cite as:
Zhang Y, Folarin AA, Ellis HL, Zhong R, Stewart C, Sankesara H, Kim H, Sun S, Pratap A, Dobson RJB
Characterizing Physical Activity Trajectories Preceding Incident Major Depressive Disorder Diagnosis With Consumer Wearable Devices in the All of Us Research Program: Retrospective Nested Case-Control Study
J Med Internet Res 2026;28:e93164
URL: https://www.jmir.org/2026/1/e93164
doi:10.2196/93164

© Yuezhou Zhang, Amos A Folarin, Hugh Logan Ellis, Rongrong Zhong, Callum Stewart, Heet Sankesara, Hyunju Kim, Shaoxiong Sun, Abhishek Pratap, Richard JB Dobson. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 4.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Proposed Participatory Framework for Explainable AI in mHealth: Mixed Methods Study Integrating User and Stakeholder Requirements

Farzana Islam, BSc, MSc; Ashraful Islam, BSc, MS, PhD; M Ashraful Amin, BSc, MSc, PhD; Moinul Zaber, BSc, MS, PhD

Center for Computational & Data Sciences, Independent University, Bangladesh, Plot 16, Block B, Aftab Uddin Ahmed Road, Bashundhara R/A, Dhaka, Dhaka Division, Bangladesh

Corresponding Author:
Ashraful Islam, BSc, MS, PhD
Center for Computational & Data Sciences, Independent University, Bangladesh, Plot 16, Block B, Aftab Uddin Ahmed Road, Bashundhara R/A, Dhaka, Dhaka Division, Bangladesh

Abstract

Background: Artificial intelligence (AI) integration in mobile health (mHealth) apps offers health care access opportunities in low-resource settings, yet opaque AI recommendations undermine trust and adoption. Existing explainable AI (XAI) frameworks, designed in Western contexts, fail to address the linguistic, cultural, and infrastructural realities of South Asian populations, creating barriers where users cannot understand AI recommendations, clinicians cannot validate outputs, and developers lack implementation guidance. Thus, understanding explainability requirements among educated, digitally literate populations provides foundational insights for future development of inclusive mHealth technologies.

Objective: This study aims to (1) investigate stakeholder perceptions of trust and explainability in AI-driven mHealth in Bangladesh; (2) identify demographic predictors of trust; and (3) develop and propose a context-adapted framework benefiting developers, policymakers, clinicians, and end users in resource-constrained settings.

Methods: This study used a sequential mixed methods design that combined a quantitative survey (n=137) with a qualitative phase involving 20 stakeholders. This qualitative cohort consisted of developers (n=4), XAI experts (n=6), and clinicians (n=10) who participated through either focus groups or individual interviews. We used statistical analysis to examine demographic predictors and applied thematic analysis to identify explainability needs specific to each stakeholder group.

Results: Education level showed a significant effect on trust ($F_{3, 133}$=2.81, P=.042). Completed undergraduate students reported lower trust (mean 3.14, SD 1.10) compared with current undergraduates (mean 3.66, SD 0.93), suggesting that undergraduate completion develops critical evaluation skills that may decrease uncritical acceptance of AI systems. Despite recognizing AI's utility for preliminary guidance, users emphasized the necessity of human validation and expressed concerns about understanding AI's decision-making logic. Interviews with different stakeholder groups revealed critical gaps. Developers acknowledged minimal explainability implementation in current mHealth apps, while medical professionals unanimously prioritized clinical judgment over automated outputs and advocated for physician-mediated AI systems. Synthesizing findings across all stakeholder groups revealed five core requirements: (1) Human–AI collaboration and clinical validation, (2) Transparent logic paths, (3) Contextual personalization, (4) Cultural and linguistic relevance, and (5) Trust calibration and ethical safeguards.

Conclusions: The framework bridges stakeholder misalignments and offers actionable guidance for design, deployment, and policy alignment in resource-constrained environments. By situating explainability within the sociocultural realities of South Asia, this research advances XAI beyond algorithmic transparency toward equity, inclusion, and user empowerment in digital health.

(J Med Internet Res 2026;28:e87158) doi:10.2196/87158

KEYWORDS

artificial intelligence; explainable AI; mHealth; digital health; human-centered design; trust in AI; Global South; responsible AI; clinical decision support; AI transparency; participatory design

XSL•FO
RenderX

Introduction

Background

Artificial intelligence (AI) is the ability of a computer system to learn and make decisions based on what it has learned from a given set of data, and it is transforming almost every aspect of health care [1]. It is now routinely applied to disease prediction, clinical decision support, and population-scale analytics, often outperforming traditional statistical methods. In parallel, mobile health–related apps (mHealth apps) have risen dramatically, offering symptom checkers, medication reminders, and personalized lifestyle advice [2]. The World Health Organization (WHO) reports that digital health adoption accelerated by over 70% following the COVID-19 pandemic, with AI-enabled mobile services showing the fastest growth [3]. The global mHealth market now exceeds 387 million active users, with forecasts surpassing 450 million by 2026, and more than 36,000 health apps available in the Google Play Store [4]. However, as AI-driven systems proliferate, their trustworthiness, explainability, and equity are increasingly questioned [5,6].

AI systems are often described as "black boxes" that generate recommendations without revealing their underlying logic [7]. Explainable AI (XAI) seeks to mitigate this opacity by providing transparent reasoning, feature importance, and confidence metrics [8,9]. Research shows that well-designed explanations can enhance clinicians' cognitive and affective trust, whereas poor or overly complex explanations can undermine confidence [10]. In mHealth contexts, explainability challenges are amplified by continuous personal data collection, real-time decision-making, and the need to communicate complex health information through small screens to users with differing literacy levels. Unlike clinical AI systems built for professionals, mHealth apps must balance technical accuracy with lay comprehension while delivering health-critical recommendations. The lack of meaningful explanations in such apps has been linked to reduced engagement, poor adherence, and premature abandonment [11,12]. Consequently, explainability is not merely a desirable feature but a prerequisite for safe, effective, and trustworthy AI adoption in mHealth care [6,13].

The global XAI research landscape reveals a striking geographic imbalance. While the field has matured rapidly in high-income countries [14-16], only a few published studies have examined XAI in Global South contexts [17]. This bias has produced frameworks ill-suited for low-resource environments characterized by limited infrastructure, small datasets, and low AI literacy [18,19]. The resulting "AI divide" reflects disparities in investment, institutional readiness, and the cultural relevance of AI technologies [20-24]. Fear and uncertainty further impede AI literacy; global surveys reveal nearly equal proportions of adults expressing nervousness (52%) and excitement (54%) about AI [25]. Without deliberate inclusion, the AI revolution risks widening existing digital and social inequalities rather than bridging them.

Within this global context, Bangladesh exemplifies both progress and persistent challenges. With over 188 million mobile connections and a penetration rate of 108%, the country has achieved remarkable digital connectivity [26]. Yet digital literacy, affordability, and sociocultural barriers continue to hinder meaningful participation in AI-driven health technologies [27,28]. The health care system faces severe workforce shortages, with around 5 physicians and 2 nurses per 10,000 population, with 75% of clinicians serving urban areas where only 38% of people live [29-31]. As a result, mHealth apps often act as the first point of contact for health care advice, especially in rural regions. However, if users cannot understand or trust AI recommendations, the potential of such systems to expand health care access remains unrealized.

The explainability requirements differ between the stakeholder groups in AI mHealth. End users prioritize concise and actionable explanations that foster confidence without technical complexity [32,33]. Developers and data scientists seek diagnostic insights into model behavior and performance to ensure reliability [34,35]. Clinicians demand clinically valid evidence pathways aligned with medical standards [6,27], while policymakers require auditability and ethical accountability [36]. Addressing these divergent needs necessitates a participatory, multistakeholder approach that integrates technical and human perspectives. Human-centered XAI emphasizes designing explanations that align with users' mental models, cognitive capacities, and cultural contexts rather than exposing raw model mechanics [37]. Personalized, context-aware explanations have proven more effective, particularly among populations with limited digital literacy [38]. In low-resource settings, where mHealth may substitute for clinical care, poorly designed or culturally mismatched explanations can lead to dangerous misinterpretations. Effective explanations must therefore not only inform but also empower users to make safe, contextually appropriate health decisions while maintaining trust and acknowledging uncertainty.

Bangladesh provides a compelling research context for exploring these issues. It exhibits features representative of South Asia, as well as the Global South, an emerging digital health industry, socioeconomic diversity, and varied literacy levels [39,40]. The nation's demographic and infrastructural patterns mirror those of countries such as India, Nigeria, and Indonesia [41-43], making insights transferable to similar settings. These parallels position Bangladesh as a microcosm for studying how AI explainability can be designed to accommodate constrained resources, cultural diversity, and evolving digital ecosystems.

Although mHealth apps promise to democratize health care, opaque AI components risk undermining their effectiveness, especially where literacy, infrastructure, and trust barriers persist. Most XAI research has been conducted in Western contexts [14,15,44,45], leaving a critical gap in understanding explainability for low-resource users [19,38,46]. Addressing this gap requires empirical evidence on how different stakeholders perceive and require explainability in mHealth apps.

Accordingly, this study aims to investigate stakeholder perceptions of trust and explainability in AI-driven mHealth apps in Bangladesh and to synthesize findings into a human-centered design framework for the South Asian community. While our broader goal is to inform XAI design

XSL•FO
RenderX

for diverse populations, including underserved communities, the current absence of Bangla-language AI-based health apps with XAI features necessitated studying English-proficient, educated users as an essential first step. These findings represent the perspectives of digital natives and early adopters in Bangladesh, providing foundational insights that can inform the future development of vernacular mHealth apps for broader population segments, including low-literacy monolingual users.

Significance and Motivation

This study addresses a critical gap at the intersection of AI, global health equity, and human-centered design. While XAI research has matured in high-income countries, South Asian populations remain underrepresented in both the design and evaluation of XAI systems [17,47]. This geographic imbalance has profound implications for health equity: AI systems trained on Western data and designed for Western users systematically fail to meet the needs of populations who could benefit most from accessible digital health technologies [44,48].

Bangladesh exemplifies the challenges facing AI-driven health care in low-resource settings. The country has widespread mobile connectivity (108% penetration [26]) yet faces severe health care workforce shortages (5 physicians per 10,000 population compared with 36 in high-income countries [29,42]) alongside significant linguistic and cultural diversity.

Three interconnected concerns motivate this work. First, the transparency paradox: while technical XAI methods such as LIME (local interpretable model-agnostic explanations), SHAP (Shapley additive explanations), and attention mechanisms proliferate in research literature [49-51], these tools remain inaccessible to end users and fail to address fundamental questions of trust, cultural appropriateness, and actionability in health care contexts [17,32]. Second, the stakeholder gap: existing frameworks rarely incorporate perspectives from developers facing resource constraints, clinicians navigating local treatment protocols, or users with limited digital literacy [45,47,51]. This results in solutions that are technically sound but practically unusable. Third, the equity imperative: as AI systems increasingly mediate access to health information and services, the absence of explainability adapted to diverse populations threatens to widen rather than narrow global health disparities [18,44,50].

This research is significant for 4 interconnected reasons. Methodologically, it provides rare empirical evidence on XAI requirements from a South Asian context, addressing what scholars have termed “algorithmic colonialism,” or the imposition of Western AI paradigms on non-Western populations [44]. Theoretically, it demonstrates that explainability is not purely a technical or cognitive challenge but fundamentally a sociocultural one, shaped by language, institutional trust, health care norms, and power dynamics [8,50]. Practically, it offers actionable guidance for developers, policymakers, and health institutions navigating the complex terrain of responsible AI deployment in resource-constrained environments. Ethically, it foregrounds the voices of users, clinicians, and developers who are typically excluded from AI design conversations, embodying principles of participatory and inclusive technology development [47,51].

By investigating explainability requirements across multiple stakeholder groups, including users with diverse educational backgrounds, developers with limited resources, clinicians prioritizing patient safety, and XAI experts balancing technical feasibility with human needs, this study generates insights that bridge research, practice, and policy. The resulting framework operationalizes explainability not as an algorithmic property but as an ecosystem of transparency, trust, cultural resonance, and accountability.

Prior Work

Explainable Artificial Intelligence

AI is increasingly integrated into health care, enabling data-driven diagnosis, prediction, and personalized care. Yet, as systems grow more complex, the opacity of their decision-making limits user understanding and trust. XAI addresses this challenge by making algorithmic reasoning interpretable to humans. XAI refers to a collection of procedures and techniques that enable people to understand and have trust in the output and results produced by AI’s machine learning (ML) algorithms. Doshi-Velez and Kim [11] conceptualize XAI as aligning algorithmic logic with human cognition to foster trust and accountability. Similarly, Gunning et al [52] emphasize “glass box” systems that reveal objectives, justifications, and reasoning in ways lay users can comprehend.

Common XAI techniques include post hoc explanation models such as LIME and SHAP, which approximate model behavior for interpretation [49,50], saliency and attention mechanisms that visualize model focus [51], and rule-based surrogate models that simplify complex systems. These methods, however, are primarily designed for technical audiences such as developers and data scientists. Recent scholarship argues for human-centered XAI, which prioritizes usability, accessibility, and contextual appropriateness over technical transparency (Liao et al [53]). Despite this shift, research continues to lack a systematic evaluation of XAI effectiveness for nontechnical users or populations in non-Western, resource-constrained contexts.

AI and mHealth: Potential and Pitfalls

AI-powered mHealth apps use ML to support real-time health monitoring, self-diagnosis, and preventive care. These tools expand access to health care, especially in regions where formal medical services are limited. Studies by Deniz-Garcia et al [2] and Singh et al [54] highlight the growing reliance on AI-backed symptom checkers such as Ada, WebMD, and Symptomate, which use natural-language processing and probabilistic reasoning to suggest possible conditions.

Existing reviews have highlighted persistent challenges in ensuring the usability, transparency, and effectiveness of AI-enabled mHealth apps, and researchers warn that premature or blind trust in AI systems may pose ethical and safety risks. Gerdes [34] and Rudin [13] caution that opaque AI in high-stakes health care can yield inappropriate advice and harm patient safety. Beede et al [45] further demonstrate that technically sound AI systems may still fail in real-world deployment due to workflow misalignment, low explainability, and user distrust. Regulatory frameworks for AI in health care

XSL•FO
RenderX

remain nascent, leaving uncertainty around accountability and validation. Thus, mHealth apps must balance innovation with transparency and cultural relevance to ensure user safety and equitable access.

Trust, Transparency, and Human-Centered XAI

Trust is central to AI adoption in health care. Shin [9] finds that explainability enhances user trust by revealing causes and reasoning behind outputs. Trust depends not only on accuracy but also on how results are communicated. Kizilcec [55] shows that users engage more confidently with systems that disclose their reasoning, particularly when users lack domain expertise.

Ehsan et al [8] advocate "social transparency," emphasizing narrative and contextual explanations over technical details to humanize AI interactions. This is especially relevant for health care, where patients need empathetic and understandable rationales rather than statistical evidence. Visser et al [12] demonstrate that explanation preferences differ among stakeholders: clinicians seek evidence-based validation, while patients value intuitive, emotionally reassuring feedback. Consequently, one-size-fits-all explanations are ineffective; adaptive and customizable approaches are required.

Participatory design has emerged as a promising approach to bridge these gaps. Liao et al [53] and Larasati et al [32] argue that effective XAI must be co-designed with patients, clinicians, and developers. Engaging stakeholders in the design process ensures that AI explanations meet real-world cognitive, professional, and ethical expectations, thus strengthening user trust and accountability.

XAI in South Asia and Global South: Challenges and Contextual Needs

Despite progress in XAI research, most studies are grounded in Global North contexts. Scholars such as Okolo [47] and Sambasivan et al [48] describe this imbalance as a form of "algorithmic colonialism," where models and design assumptions from high-income countries are exported to environments with distinct cultural and infrastructural realities. AI trained on Western datasets may fail to generalize, such as dermatology models that misdiagnose darker skin tones or provide explanations that are linguistically or culturally inappropriate.

Islam et al [38] show that users in South Asia, including Bangladesh, often interact with AI technologies without fully understanding them due to limited digital literacy. This can lead to both over-reliance and distrust, depending on prior experience. Ismail and Kumar [56] likewise observed among Indian community health workers that sociocultural context, language accessibility, and human mediation are essential for AI acceptance.

Okolo et al [47] emphasize the absence of localized explainability strategies that account for language diversity, cultural norms, and expectations around medical authority. To improve credibility and adoption, XAI systems in South Asia should incorporate:

- Endorsement by local health institutions, linking AI recommendations to trusted authorities.
- Integration of national clinical guidelines, ensuring contextual medical validity.
- Local language and visual support, accommodating low literacy.
- Consideration of drug availability and affordability, enhancing actionability.

Without such adaptation, AI systems risk being viewed as foreign, untrustworthy, or irrelevant. The lack of culturally grounded XAI approaches threatens to widen rather than close global health disparities.

Research Gap

The literature demonstrates substantial technical progress in XAI but limited focus on contextual, user-centered design. In health care, where AI recommendations influence real decisions, explanations must be culturally sensitive, transparent, and trust-building. Most existing work, however, reflects Global North priorities and assumes high digital literacy, overlooking the heterogeneous realities of users in the Global South as well as South Asia. Stakeholder voices such as doctors, developers, and patients are seldom integrated into XAI design, creating a critical research gap. This study addresses these limitations by:

- Investigating how users, developers, and health care professionals in Bangladesh, a country in South Asia, perceive and require explainability in mHealth AI systems.
- Identifying design principles that balance technical feasibility with cultural and cognitive accessibility.
- Proposing a participatory, context-aware framework that operationalizes XAI for equitable health care delivery.

By embedding social, cultural, and institutional dimensions into XAI design, this work advances an inclusive, human-centered approach to AI in global health, ensuring that mHealth technologies empower rather than exclude the populations they aim to serve.

Research Questions and Objectives

Research objectives are as follows: (1) to characterize perceptions of trust and explainability across users, developers, and health care professionals; (2) to evaluate explanation formats suitable for varying literacy levels; and (3) to construct a participatory design framework for explainable mHealth AI.

Research questions are as follows: (1) How do different stakeholders perceive the trustworthiness and explainability of AI-generated health recommendations? (2) What explanation formats and approaches are most meaningful for users with differing literacy levels? (3) What expectations and concerns do developers, clinicians, and experts have regarding explainability in mHealth AI? (4) How can these needs be synthesized into a framework for explainable mHealth AI in low-resource settings?

By situating XAI within Bangladesh's mHealth ecosystem, this research contributes empirical evidence to an underrepresented region and offers practical guidance to develop transparent, equitable, and reliable AI systems.

XSL•FO
RenderX

Study Contributions

This research makes 3 primary contributions, as shown in the following sections.

Empirical Contributions

- First comprehensive multistakeholder evidence on XAI requirements from a South Asian context (n=157 across users, developers, clinicians, and experts).
- Quantitative evidence that education level (*P*=.042) predicts trust in AI explainability, challenging "digital native" assumptions.
- Systematic documentation of stakeholder misalignments revealing gaps between user needs, developer constraints, and clinical requirements.

Theoretical Contributions

The theoretical contributions are as follows:

- Reconceptualization of explainability as a sociocultural construct requiring cultural resonance, linguistic accessibility, and institutional validation, not merely algorithmic transparency.
- Evidence supporting the "social explainability" theory: users prioritize human validation and institutional endorsement over technical interpretability metrics.

Practical Contributions

The practical contributions are as follows:

- Five-pillar participatory framework operationalized for resource-constrained environments (human-AI collaboration, transparent logic, contextual personalization, cultural-linguistic relevance, and ethical safeguards).
- Actionable implementation guidance for developers addressing real constraints (limited resources and absence of localized NLP tools).
- Evidence-based for policymakers establishing AI governance standards and clinical validation protocols in health care contexts.

By centering voices from contexts where 3+ billion people share similar infrastructural realities, this work advances responsible AI toward equity, inclusion, and user empowerment in global digital health.

Methods

Overview

This study used a human-centered mixed methods approach to explore how explainability can be effectively designed in AI-driven mHealth apps. The aim was to identify what makes explanations clear, trustworthy, and contextually relevant in the sociocultural environment of Bangladesh. By integrating quantitative and qualitative data, the research captured perspectives from end users, health care professionals, app developers, and XAI researchers. A sequential design was used: survey findings informed subsequent interviews, ensuring both breadth and depth in understanding explainability needs.

Study Design

The study was conducted in 2 phases, each focusing on a distinct stakeholder group.

- User Survey: Explored perceptions of trust, transparency, and explanation preferences among mHealth users.
- Expert and Developer Interviews: Investigated technical and design considerations related to XAI implementation.
- Key Informant Interviews (KIIs): Engaged medical professionals to assess clinical validity, usability, and trust concerns.

In Figure 1, the detailed recruitment process is shown.

Multimedia Appendix 1 presents an interview guide for stakeholder interviews with developers, clinicians, and XAI experts. Multimedia Appendix 2 presents the survey questionnaires administered to end users.

Table 1 summarizes participants and methods, and Figure 2 shows the sequential mixed methods design used in this study. Quantitative survey findings informed the development of interview and focus group discussion guides. The qualitative responses were analyzed thematically and integrated with survey insights to develop the proposed explainable mHealth AI framework.

XSL•FO
RenderX

Figure 1. Study design and participant recruitment flow diagram. XAI: explainable AI.

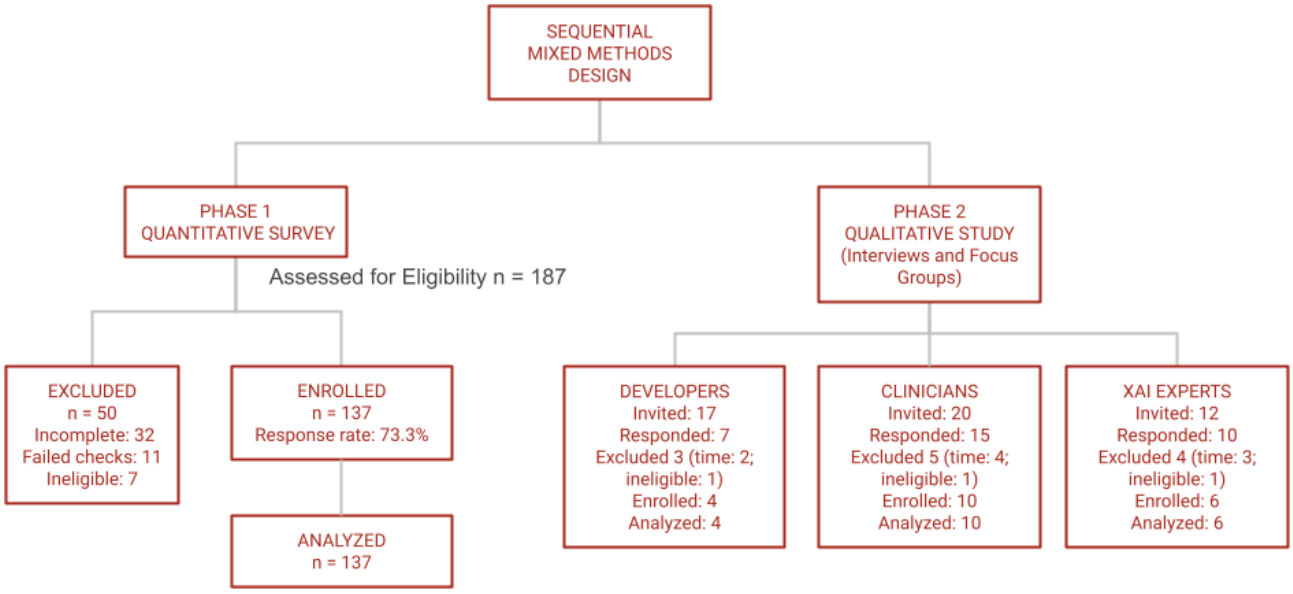

Table . Summary of study participants and methods.

Stakeholder group	Number of participants	Method	Mode	Duration (mins)	Notes
mHealth[a] users	137	Survey	Online	15 - 20	Used Ada, WebMD, and Symptomate apps
XAI[b] experts	6	Interview	Online + In-person	30 - 40	Global North and South representation
App developers	4	Interview	Online + In-person	30 - 40	Local and international mHealth experience
Medical professionals	10 (1+9)	Interview + Focus Group	Online + In-person	45 - 60	Practicing in Bangladeshi hospitals

[a]mHealth: mobile health.

[b]XAI: explainable artificial intelligence.

XSL•FO
RenderX

Figure 2. Sequential mixed methods research design used in this study. AI: artificial intelligence; FGD: focus group discussion; mHealth: mobile health; XAI: explainable AI.

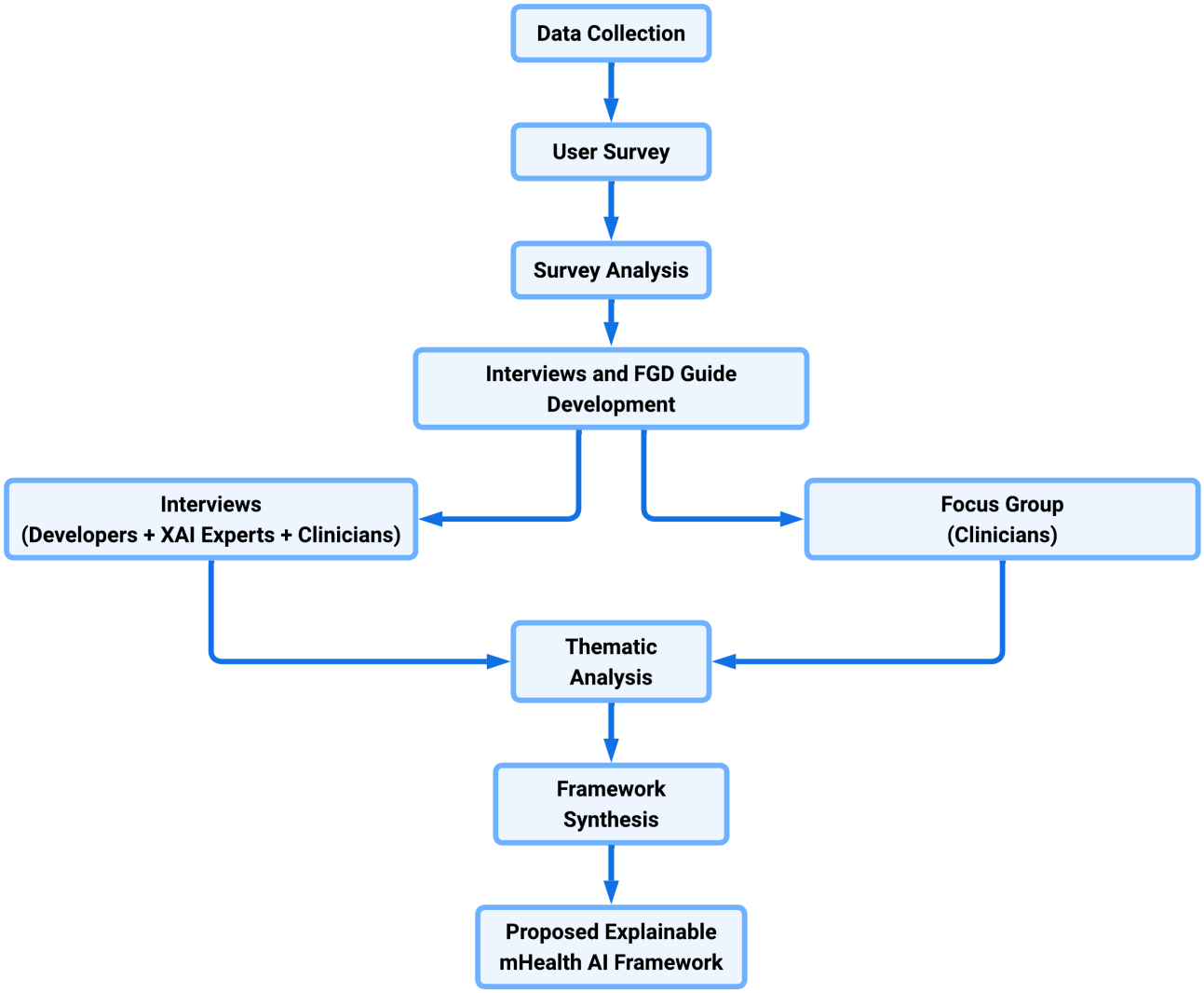

User Survey and Pilot Testing

Overview

The user survey examined app usage trends, demographic factors, and perceptions of explainability in AI-driven mHealth apps. We used Google Forms for anonymous online data collection. Participation was voluntary, with informed consent obtained prior to access. The survey instrument was pilot tested with 15 participants matching the target demographic before full deployment. Pilot participants completed the survey after using one of the mHealth apps and provided feedback on question clarity, length, and comprehensibility. Based on pilot results, the survey was revised to simplify technical terminology, reduce redundant items, and clarify instructions. This pilot testing ensured the instrument was appropriate for the target population and capable of capturing meaningful explainability and trust perceptions.

Survey Structure

The questionnaire included three parts:

1. Demographics and Background: Collected age, gender, region, education, and AI familiarity to examine how explainability perceptions varied across groups.
2. AI Health App Usage Context: Explored frequency, familiarity, and trust in AI-based symptom checkers.
3. Perceptions of Explainability: Combined 5-point Likert-scale items [57] with open-ended questions [58] such as “Why do you trust or distrust the app’s recommendation?” and “What would make the app’s suggestions more explainable or trustworthy?”

To ensure participants understood the concept of explainability regardless of their familiarity with technical terminology, we implemented a structured approach. After asking whether participants were familiar with “Explainable AI,” all participants received a clear, jargon-free explanation. The explanation described explainability as “the app’s ability to show you WHY it gave a particular recommendation and HOW it reached that conclusion,” using everyday language and culturally relevant health examples. For instance, we explained, “If an app suggests you might have the flu, explainability means it tells you which symptoms led to this suggestion, rather than just giving you the result.” This ensured participants rated actual explainability features (transparency of AI reasoning) rather than general UI satisfaction.

XSL•FO
RenderX

App Selection

Three widely used AI health apps, namely Symptomate, Ada Health, and WebMD Symptom Checker [59], were selected based on (1) free availability on both iOS and Android platforms, ensuring broad accessibility; (2) diversity in explanation design: Ada offers Bayesian "white-box" reasoning, Symptomate provides dynamic questioning, and WebMD uses traditional list-based diagnostics [54]; and (3) established user bases with more than 3 million users worldwide. Table 2 gives the comparative analysis of these 3 apps.

Table . Comparison of selected artificial intelligence (AI) health apps.

Characteristics	Symptomate	Ada Health	WebMD Symptom Checker
User interface	Interactive questionnaire with chatbot functionality	Conversational interface similar to ChatGPT	Traditional list-based approach with text input
Interaction model	Dynamic, personalized questionnaire	Conversational AI with follow-up questions	Symptom selection and body-based navigation
AI architecture	Interactive questionnaire with personalized algorithms	White box Bayesian probabilistic system	Conventional diagnostic algorithm
Explainability level	Moderate - implicit explainability via questioning process	High - white box Bayesian probabilistic reasoning	Low - minimal algorithmic transparency
User base	3+ million users globally	13+ million users worldwide	Extensive (part of WebMD ecosystem)
Geographic reach	Global accessibility	148 countries, 11 languages	Primarily English-speaking markets
Primary features	Symptom analysis, health insights, decision support	Health assessment, telemedicine integration	Symptom checking, condition lookup, doctor finder
Assessment approach	Unlimited symptom input with follow-up prompts explaining logic	Medical dictionary comparison using differential diagnosis	Symptom-to-condition matching with minimal reasoning
Output format	Guided health insights	Personalized assessment reports	Potential conditions list with treatment options
Explainability features	Question flow logic, symptom pathway transparency	Bayesian reasoning display, medical logic explanation	Limited - basic condition information
User manual transparency	Process explanation via user experience	White-box system documentation	Minimal AI process explanation
Regulatory compliance	Standard health app guidelines	FDA AI/ML[a] framework compliance	Medical information standards
Trust indicators	Process transparency through guided questioning	Clinical validation with explicit medical reasoning	Medical brand credibility over algorithmic transparency
User experience focus	Accessibility and ease of use	Comprehensive health companion	Professional medical information

[a]ML: machine learning.

Participant Recruitment and Sample Characteristics

Survey Participants

Survey participants (n=137) were recruited using convenience sampling [60]. All had internet access and smartphones. The majority (93/137, 67.9%) were aged 18 - 24 years, and 64.2% (n=88) were male. Educational levels were predominantly undergraduate (121/137, 88.3%). About 94% (n=129) were familiar with AI concepts, while 39.4% (n=54) recognized the term "explainable AI." Approximately half were first-time mHealth users. Tables 3 and 4 summarize key demographics.

Table . Age and gender distribution of participants.

Age group (years)	Male	Female	Total, n (%)
18 - 24	60	33	93 (67.9)
25 - 30	21	11	31 (22.6)
31 - 40	6	3	9 (6.6)
Older than 40	2	2	4 (2.9)
Total, n (%)	88 (64.2)	49 (35.8)	137 (100)

XSL•FO
RenderX

Table . Health app usage patterns among participants.

Usage pattern	Values, n (%)	Implication for study
Regular users	5 (3.6)	Experienced with health apps
Very rare users	66 (48.2)	Limited prior experience
First-time users	66 (48.2)	New to health apps
Total	137 (100.0)	Diverse experience levels

Interview and Focus Group Participants

Interview and focus group participants (20 total: 4 developers, 10 clinicians, and 6 XAI experts) were recruited through convenience and snowball sampling. Initial contacts were made through the lead researcher's professional and academic networks in the Bangladeshi health technology and medical sectors. Subsequent participants were recruited through referrals from initial participants. Participation was voluntary. While no financial compensation was provided, light refreshments (tea and snacks) were offered during interviews and focus groups as a gesture of appreciation.

Language and Sampling Considerations

Participants were selected based on the fact that all mHealth apps assessed in this study (Ada, WebMD, and Symptomate) are only available in English due to the lack of Bangla-language mHealth apps with AI and XAI features. Consequently, all participant interactions with these apps occurred in English. This linguistic limitation necessitated recruiting participants with English proficiency, which inherently biased the sample toward educated, digitally literate individuals. The decision to recruit English-proficient participants allowed us to isolate trust and explainability issues stemming from AI opacity and design inadequacies, rather than confounding these factors with basic language comprehension barriers. By ensuring participants could understand the English interface, we aimed to evaluate whether the apps' explanations were inherently unclear, insufficient, or culturally inappropriate. This approach represents a "best-case scenario" for assessing current international mHealth apps in the Bangladeshi context, as any trust or explainability deficits identified among English-proficient users would likely be substantially amplified among non–English-speaking populations.

Data Collection

Participants installed and used one app as per their health needs to perform structured symptom-checking tasks before completing the survey. Specifically, participants input their own real-time health data, current symptoms, health concerns, or wellness queries they were genuinely experiencing, rather than hypothetical scenarios. This approach ensured feedback reflected authentic personal experiences. The protocol involved (1) downloading one assigned app (Ada, WebMD, or Symptomate), (2) using the app to check at least one real health symptom or concern, (3) interacting with the AI-driven symptom checker by answering its questions, (4) reviewing the app's recommendations and explanations, and (5) immediately completing the survey about their experience. This real-time approach captured genuine user perceptions of trust, explainability, and usability that hypothetical scenarios would not provide. Data included quantitative Likert-scale ratings and open-ended qualitative reflections.

Expert and Developer Interviews

To complement user insights with technical perspectives, semistructured interviews were conducted with 6 XAI experts and 4 app developers. Experts were selected for their work in explainable ML, responsible AI, or human-computer interaction. Developers were practitioners from mHealth companies with experience in deploying AI-based tools in low-resource contexts.

Interviews (30 - 60 min) were conducted online or in person using a semistructured protocol. Participants provided informed consent and were anonymized. Topics included limitations of existing XAI tools (eg, LIME, SHAP), trade-offs between accuracy and interpretability, and adaptation of explanations for nontechnical audiences. Developer interviews focused on implementation challenges, interface design, and contextual constraints such as literacy and language barriers.

Interview data were manually coded using an inductive thematic approach. Themes such as technical feasibility, explanation usability, and trust calibration emerged, revealing gaps between research-driven XAI tools and real-world deployment realities.

The KIIs

KIIs [61] were used to capture clinical perspectives on AI explainability. Ten physicians participated: one in-depth individual interview and a focus group of 9 practitioners from hospitals in Bangladesh. KIIs explored perceptions of AI reliability, trust, patient misuse, and preferred explanation formats.

Sessions were conducted in Bangla (in person or online) and lasted 40 - 60 minutes. Doctors expressed cautious optimism toward AI tools but stressed that explanations should align with clinical logic and national treatment protocols. Themes such as clinical trust boundaries, alignment with diagnostic reasoning, and explanation expectations were identified through inductive coding.

Data Analysis

A mixed methods analytical approach was used, integrating quantitative statistical analysis of survey data with qualitative thematic analysis of interview and focus group transcripts.

Quantitative Data Analysis

Survey data (n=137) were analyzed using both descriptive and inferential statistics. Descriptive statistics, including means, SDs, medians, and frequencies, were calculated for all Likert-scale items [57] and demographic variables to provide

XSL•FO
RenderX

an overview of trust perceptions, explainability ratings, and belief patterns across the sample, summarized in Table 1.

Inferential Statistical Analysis

To examine differences in trust and explainability perceptions across demographic groups, one-way ANOVA tests were conducted to compare trust ratings (measured on a 5-point Likert scale) across age categories (18 - 24, 25 - 30, 31 - 40, and older than 40 years) and education levels (higher secondary, undergraduate running, undergraduate completed, postgraduate). Given the potential for violations of parametric test assumptions with the sample distribution, Kruskal-Wallis H tests were also performed as nonparametric alternatives to ANOVA for all group comparisons. Both parametric (ANOVA) and nonparametric (Kruskal-Wallis) results are reported to ensure robustness of findings. Effect sizes were calculated for significant ANOVA results using eta-squared (η^2), with interpretations following Cohen's conventions: small (0.01 - 0.06), medium (0.06 - 0.14), and large (≥0.14) effects. Chi-square tests of independence examined associations between categorical variables, specifically: (1) education level and belief in app decisions (Yes/No/Not sure), (2) XAI familiarity and belief in app decisions, (3) age group and belief in app decisions, and (4) age group and willingness to trust app results over physicians. Statistical significance was set at α of .05 for all tests. All quantitative analyses were conducted using Python (version 3.10) with the SciPy statistical library (version 1.9) and Pandas for data manipulation.

Qualitative Data Analysis

Analytical Framework

Open-ended survey responses and interview data were analyzed using Braun and Clarke's [62] reflexive thematic analysis, following their 6-phase process [63]. This method was selected for its flexibility in capturing diverse stakeholder perspectives while maintaining systematic rigor.

Six-Phase Analysis Process

Phase 1: Familiarization

All data were read repeatedly to develop deep familiarity. Survey open-ended responses (n=137, in English) and interview notes (in English) were reviewed multiple times. The Bangla focus group transcript was read repeatedly in the original language.

Phase 2: Generating Initial Codes

Coding was performed systematically by the lead researcher. English survey responses and interview notes were coded directly. The Bangla focus group transcript was coded in Bangla to preserve cultural-linguistic nuances. A codebook [64] was iteratively developed and refined through peer debriefing with supervisors. Codes captured both explicit meanings (semantic) and underlying assumptions (latent).

Phase 3: Generating Initial Themes

Codes were organized into candidate themes by identifying patterns and relationships across the dataset. Visual mapping (tables and clusters) explored connections between codes and potential overarching themes. Preliminary themes were discussed with supervisors who provided critical feedback on theme coherence and boundaries.

Phase 4: Reviewing Themes

Candidate themes were reviewed for internal coherence and distinctiveness through regular peer debriefing sessions with supervisors. Coded data extracts were reexamined to ensure adequate support for each theme. Supervisors challenged interpretations and suggested alternative organizations. Themes lacking sufficient data or coherence were revised, merged, or discarded based on this collaborative review.

Phase 5: Defining and Naming Themes

Each theme was clearly defined, specifying its scope, boundaries, and relationship to research questions. Theme names and definitions were refined through supervisor feedback. Theme names were crafted to be concise yet evocative.

Phase 6: Producing the Report

Final themes were presented with representative quotes. Bangla focus group themes and quotes were translated to English for manuscript presentation (see Translation Pipeline below).

Translation and Coding Pipeline

The survey was conducted entirely in English, as all evaluated mHealth apps (Ada, WebMD, Symptomate) operate in English. Open-ended survey responses were in English and required no translation.

Individual interviews (n=11: 4 developers, 1 clinician, 6 experts) were conducted in Bangla, with the researcher taking notes in English during interviews to facilitate analysis. One focus group with clinicians (n=9) was conducted in Bangla, audio-recorded, and transcribed verbatim in Bangla by a professional transcriber.

Coding was performed on the Bangla focus group transcript by the lead researcher (native Bangla speaker) to preserve cultural-linguistic authenticity. While initial codes were assigned to Bangla text segments, themes were developed directly in English to facilitate integration with other data sources (English survey responses and interview notes). Representative quotes from the focus group were translated from Bangla to English for manuscript presentation, with back-translation verification (Bangla→English→Bangla comparison) performed to ensure semantic equivalence. We acknowledge the potential loss of cultural-linguistic nuance in this translation process.

Data Adequacy and Saturation

Formal data saturation assessment (Hennink et al [65]) [66] was not conducted.

Sample adequacy is characterized as:

- User survey responses (n=137): Achieved informational redundancy for primary trust and explainability concerns.
- Developer interviews (n=4): Small sample yields preliminary insights rather than saturated themes.
- Clinician perspectives (n=10 total: 1 interview+9 focus groups): Achieved thematic adequacy for core professional concerns.
- Expert interviews (n=6): Achieved thematic adequacy for technical-ethical considerations.

XSL•FO
RenderX

Developer Sample Size and Thematic Characterization

While the developer sample was small (n=4), we characterize findings as "themes" rather than "observations" based on several methodological considerations. First, based on the research team's engagement with Bangladesh's mHealth ecosystem through professional networks and industry observations, we estimate there are approximately 10 to 15 active mHealth development teams in the country. Our sample of 4 developers, therefore, represents roughly 30% to 40% of this community. However, no comprehensive industry census or registry exists to definitively establish the total number of active mHealth developers in Bangladesh. This estimate should be interpreted as a qualitative assessment of sampling adequacy rather than a statistically validated market share.

Second, we observed strong thematic convergence despite organizational diversity: all 4 developers, working independently across different companies, applications, and technical infrastructures, spontaneously identified similar constraints (resource limitations, lack of Bangla-compatible NLP tools, and market pressures), implementation challenges (explainability deprioritization and technical debt), and design trade-offs.

Third, developer insights were triangulated with other stakeholder findings. User complaints about lack of explanations (n=137), expert observations about implementation gaps (n=6), and clinician concerns about system opacity (n=10) all corroborated developer-reported constraints, strengthening confidence that themes reflected actual industry practices. Fourth, we applied the concept of Malterud et al [67] "information power" rather than traditional saturation criteria, which emphasizes study aim specificity, sample specificity, theoretical application, dialogue quality, and analysis strategy over sample size alone.

Fifth, the 4 developers represented meaningful diversity: different organizational sizes, app types (symptom checkers vs wellness apps), technical approaches (rule-based vs ML), and target populations. However, we acknowledge this sample cannot capture all variations, particularly from larger organizations, international developers, or those with specialized AI/ML expertise. We use "dominant themes" to indicate patterns emerging consistently across our participants within this specific context, not to claim universal prevalence. These findings illuminate current industry constraints in the Bangladeshi mHealth context but require validation with larger, more diverse samples before generalizing beyond this setting.

Methodological Rigor and Trustworthiness

We used multiple strategies to enhance trustworthiness. Credibility was established through triangulation across surveys, interviews, and focus groups; peer debriefing with supervisors who critically examined interpretations; thick description providing contextual detail; and analysis grounded in extensive illustrative quotes. Dependability was ensured through an audit trail documenting analytical decisions via dated codebook versions, regular progress reports, and meeting notes capturing supervisor feedback. Confirmability was strengthened by grounding findings in participant quotations, actively seeking negative cases (eg, some users preferred simplicity over detailed explanations), and supervisor review of coding and themes. Transferability is supported through a detailed description of the Bangladeshi health care context, participant characteristics, app features, and sociocultural factors, enabling readers to assess applicability to similar settings. Throughout the analysis, regular peer debriefing meetings with supervisors provided critical feedback, questioned assumptions, and challenged analytical decisions, enhancing rigor and reducing individual researcher bias.

Reflexivity

The lead researcher is a Bangladeshi computer scientist familiar with local health care and technology contexts. This insider status facilitated cultural interpretation and participant rapport but may have led to overlooking aspects obvious to insiders yet notable to outsiders. Technical background aided understanding of developer constraints but required conscious effort to foreground user and clinician perspectives. As a researcher advocating for explainability, I held prior assumptions that transparency improves trust; through reflexive practice and supervisor feedback, I actively sought disconfirming evidence, documenting instances where participants valued simplicity over transparency or expressed concerns that excessive information could overwhelm users.

Limitation of Single-Coder Approach

While coding was performed by a single researcher, the peer debriefing process with supervisors provided external checks on interpretations. However, the lack of independent intercoder reliability testing remains a limitation. Future research would benefit from dual independent coding with systematic reliability assessment.

Integration and Triangulation

Triangulation [68] integrated findings across quantitative and qualitative datasets to strengthen validity and develop a comprehensive understanding of explainability needs. Quantitative analysis revealed that education level significantly predicted trust ratings (P=.042, Figure 3), while qualitative data provided contextual explanations for why completed undergraduate students expressed lower trust (eg, heightened critical evaluation skills, concerns about AI limitations). User confusion about unclear app logic in qualitative responses paralleled clinician concerns about patient overreliance, while developers highlighted implementation limits of multilingual explanations. Statistical patterns identified in survey data informed the interpretation of qualitative themes, while qualitative narratives provided contextual depth to statistical findings. This cross-stakeholder synthesis informed the development of the proposed XAI framework, ensuring it addressed both measurable trust deficits and nuanced stakeholder concerns.

XSL•FO
RenderX

Figure 3. Distribution of trust ratings across education levels.

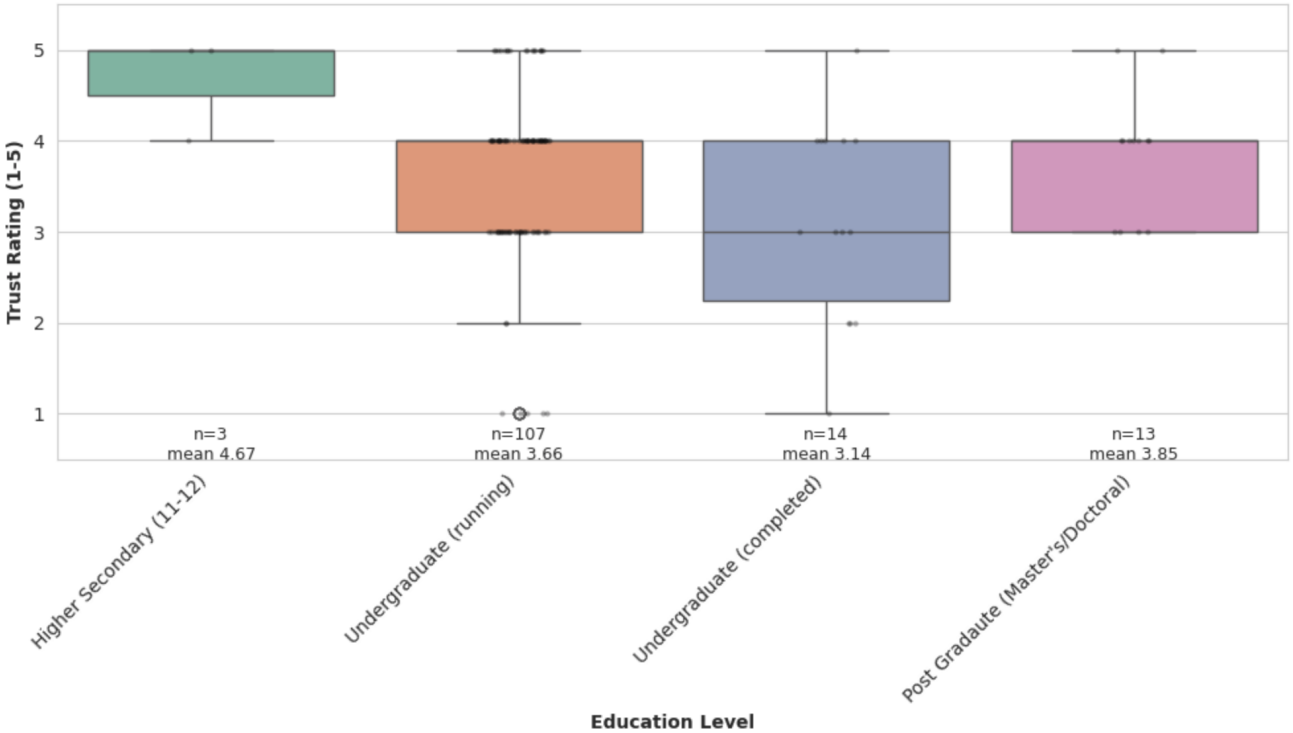

Ethical Considerations

This study received full ethical approval from the Institutional Review Board at Independent University, Bangladesh (approval number: 2025-SETS-NSR:03). All research procedures were conducted in accordance with the Declaration of Helsinki and institutional guidelines for human subjects' research. Care was taken to ensure psychological comfort; sensitive or technical questions were optional.

Beyond procedural ethics, broader concerns were addressed:

- Digital Divide and Equity: To avoid marginalization of low-literacy users, the study emphasized inclusive participation.
- Informed Consent and Trust: Clear language was used to explain study aims and AI-related concepts.
- Algorithmic Bias: Awareness was raised about potential misdiagnosis due to Western-trained datasets.
- Accountability: Discussions highlighted the need for governance mechanisms clarifying responsibility for AI-generated medical advice.

By integrating these principles, the research upheld ethical transparency and participant autonomy throughout all stages of data collection and analysis.

Prior to participation, all participants were provided with a detailed information sheet explaining the study's purpose, procedures, potential risks and benefits, the voluntary nature of participation, and their right to withdraw at any time without penalty.

For the survey component (Phase 1), participants provided electronic informed consent by checking a box at the beginning of the survey indicating they had read and understood the information sheet and agreed to participate voluntarily.

For interviews and focus groups (Phase 2), participants provided written informed consent prior to data collection. The consent form explicitly informed participants that sessions would be audio-recorded, and participants had the option to decline recording (though none did). Participants were reminded of their right to skip any questions or withdraw from the study at any time.

To protect participant confidentiality, all data were anonymized immediately upon collection. Survey responses were collected anonymously using Google Forms, with no personally identifiable information requested other than the ones mentioned, such as age, gender, etc. Interview and focus group participants were assigned unique pseudonymous identifiers (eg, AD [application developer] 1 for Developer 1, D1 for Clinician 1, P1 for Expert 1) used throughout data analysis and reporting. Audio recordings of interviews and focus groups were stored on password-protected, encrypted devices accessible only to the research team. Transcripts were de-identified before analysis. All data files are stored securely in accordance with Independent University, Bangladesh's data protection policies and will be retained for (X years, eg, 5 years) as per institutional and regulatory requirements, after which they will be securely destroyed. Screenshot uploads were requested to verify that participants had installed and used the mHealth app. While the survey questionnaire asked users to upload screenshots showing symptom check results, these images were used solely for verification purposes and were not analyzed, coded, or included in any research findings. All uploaded screenshots were securely stored in password-protected cloud storage accessible only to the research team, reviewed solely to confirm app usage, and permanently deleted after verification. No personal health information from screenshots was extracted, recorded, or retained for research purposes.

XSL•FO
RenderX

Participation was voluntary and noncompensated. However, all participants were offered access to study results upon completion. The study posed minimal/no risk to participants. Potential risks included minor discomfort from discussing professional challenges or concerns about AI technologies. Participants were informed they could skip any questions they felt uncomfortable answering and could withdraw at any time. No participant reported adverse events or exercised their right to withdraw during data collection. Contact information for the research team and the Institutional Review Board was provided to all participants for any questions, concerns, or complaints regarding the research.

Results

Stakeholder Perspectives on XAI in mHealth

User Perspectives on XAI in mHealth Apps

Overview

The user study examined how people in Bangladesh perceive and interact with AI-driven mHealth apps, focusing on explainability, trust, and expectations. A total of 137 participants used one of 3 AI-enabled health apps, Ada, WebMD, or Symptomate [59], before completing a survey that included open-ended and Likert-scale questions. The aim was to understand their experiences with app recommendations and their perceptions of the clarity, reliability, and usefulness of explanations.

Quantitative Findings: Trust Perceptions Across Demographics

Survey Response Overview

A total of 137 valid survey responses were analyzed. Participants were predominantly young adults aged 18 - 24 years (93/137, 67.9%), male (88/137, 64.2%), and currently enrolled in undergraduate programs (107/137, 78.1%). Approximately 94% (n=129) reported familiarity with AI concepts, while only 39.4% (n=54) recognized the term “explainable AI.” Overall, participants rated app explainability with a mean score of 3.65 (SD 0.94, median 4.0) on a 5-point Likert scale, indicating moderate trust in app explainability. Trust ratings ranged from 1 to 5, demonstrating substantial variability in user perceptions.

Education Level as a Primary Predictor of Trust

One-way ANOVA revealed a statistically significant difference in trust ratings across education levels ($F_{3, 133}$=2.81, P=.042, η^2=0.060), confirmed by Kruskal-Wallis test (H(3)=7.88, P=.049). Users with higher secondary education reported the highest trust (mean 4.67, SD 0.58, n=3), followed by postgraduate holders (mean 3.85, SD 0.69, n=13) and current undergraduates (mean 3.66, SD 0.93, n=107). Completed undergraduate students showed the lowest trust (mean 3.14, SD 1.10, n=14). The small effect size (η^2=0.060) indicates that education accounts for approximately 6% of the variance in trust ratings, suggesting that other factors also influence perceptions. Figure 3 illustrates these differences visually.

Box plots showing trust ratings (1 - 5 Likert scale) across 4 education categories. The box represents the interquartile range (IQR), with the horizontal line indicating the median. Whiskers extend to the minimum and maximum values within 1.5× IQR. Individual data points are overlaid as dots. Sample sizes are as follows: higher secondary (n=3), undergraduate (running; n=107), undergraduate (completed; n=14), and postgraduate (n=13; $F_{3, 133}$= 2.81, P=.042, η^2=0.060 obtained through one-way ANOVA).

Age: No Significant Differences

Contrary to initial observations, ANOVA revealed no statistically significant difference in trust ratings across age groups ($F_{3, 133}$=0.62, P=.60), confirmed by the Kruskal-Wallis test (H(3)=2.72, P=.44). Trust remained consistent across ages: 18 - 24 (mean 3.63, SD 0.96) years, 25 - 30 (mean 3.74, SD 0.89) years, 31 - 40 (mean 3.33, SD 0.71) years, and older than 40 (mean 4.00, SD 1.41) years (n=4). While descriptive means vary, differences were not statistically significant.

Other Demographic Factors

No significant differences emerged by gender, XAI familiarity, app type, or usage frequency. This pattern indicates that education level, rather than age or other demographics, was the primary predictor of trust differences.

Categorical Associations: Belief Patterns

Chi-square tests revealed 2 significant associations. Education level is significantly associated with belief in app decisions (χ^2_6=17.08, P=.009), with 69% of postgraduate holders believing recommendations compared with 43% of completed undergraduates. XAI familiarity is also significantly associated with belief (χ^2_4=11.50, P=.02), with 63% of XAI-familiar users expressing belief versus 33% of partially familiar users. No significant associations emerged between age and belief patterns.

Language and Cultural Context

It is important to note that these trust ratings reflect user perceptions of English-language apps. Although the participants knew English and could comprehend the app’s content, the language barrier, which is different from AI opacity, may have led to lower trust ratings in a number of ways: (1) reduced cultural resonance of English medical terminology and explanations, (2) increased cognitive load when processing health information in a nonnative language, (3) lack of emotional connection that native-language interactions provide in sensitive health contexts, and (4) contextual inappropriateness of recommendations generated from Western-centric medical knowledge bases. Therefore, observed trust deficits likely reflect a combination of AI explainability gaps and language-cultural barriers.

Qualitative Findings: User Perspectives

Trust and Clarity of Recommendations

Users valued the convenience of mHealth apps for providing preliminary diagnoses but frequently questioned the clarity and trustworthiness of the underlying logic. As one participant explained,

> *The app diagnosed me with a condition after I provided my symptoms, but it didn't explain how or why it arrived at this conclusion.* [P21]

XSL•FO
RenderX

Many participants expressed skepticism about accuracy and a desire for human validation.

> *I'm not sure because I'd want a doctor to confirm any treatment.* [P11]
>
> *I cannot trust what a robot tells me to do.* [P5]

Participants often viewed AI outputs as uncertain or incomplete, "Sometimes it may not be true" (P4), but believed transparency could mitigate distrust. Users requested explanations that disclosed how recommendations were generated, what data sources were used, and whether clinical guidelines or expert input were applied:

> *It would be helpful if the app could provide a brief explanation or reasoning behind its recommendations, including relevant studies or sources.* [P35]
>
> *The app should provide clear information about the sources of its data and how its AI models work.* [P53]

These findings reveal a need for explicit, contextually relevant explanations to strengthen user confidence in AI-based health assessments.

AI for Early Guidance, Human for Final Judgment

Despite skepticism, most users acknowledged the utility of AI tools as a first point of guidance rather than a replacement for medical expertise.

> *The app identified my problem correctly.* [P6]
>
> *At least it gives me an insight or an idea about what is going on with my health.* [P39]

Participants described AI as useful for early insight but emphasized the need for human oversight:

> *This app might be helpful for decision-making, but you have to go to a doctor finally.* [P100]
>
> *To make the apps more trustworthy, it could use clear explanations for its decisions and include user feedback to keep improving.* [P11]

These reflections underscore that explainability enhances perceived utility but does not fully substitute for professional judgment.

Preference for Human Judgment

A strong majority preferred human expertise for final decision-making, emphasizing empathy, contextual reasoning, and physical examination:

> *Doctors can consider a broader range of factors and provide more personalized insights than an app.* [P27]
>
> *A doctor can physically examine the situation and give proper treatment, unlike an app that just collects symptom data.* [P132]

Participants perceived AI as limited in handling rare or complex conditions, as they said, "AI cannot do what humans can do" (P48), reflecting a boundary between computational reasoning and human care.

Perceived Benefits and Boundaries of AI Guidance

Participants recognized AI's analytical strength but remained aware of its limits:

> *The app made decisions based on the data, which might be helpful for decision-making.* [P119]

However, they insisted that transparency and validation were necessary safeguards:

> *Transparent explanation of recommendations: The app could provide detailed explanations on how it arrives at its recommendations, including highlighting the key symptoms, medical history, or patterns it considered.* [P68]
>
> *Explainability features such as providing a rationale for each recommendation can make it easier to trust.* [P25]

These responses highlight that users accept AI as an assistant rather than an authority and that clear reasoning enhances, but does not guarantee, trust.

Differences Among Apps and Familiarity Effects

Although users were randomly assigned one app, differences among Ada, Symptomate, and WebMD shaped perceptions. Chatbot-based systems (Ada, Symptomate) were often viewed as more engaging and transparent than static list-based tools (WebMD). Prior exposure also influenced reactions. First-time users tended to be more forgiving and impressed by novelty, while experienced users were more critical, expecting deeper explanations and medical validation. These variations suggest that interface design and familiarity meaningfully shape how explainability is experienced.

Improving Explainability Through User-Centered Design

Participants offered constructive recommendations to enhance AI explainability, emphasizing that transparency must be clear, contextual, and human-connected rather than purely technical.

Clear and Contextualized Explanations

> *Including a clear, concise explanation of the potential risks and benefits associated with each recommendation could make the app's suggestions feel more reasoned and cautious.* [P77]

Users desired plain-language rationales that disclose limitations and uncertainties, enabling informed decision-making.

Human Oversight and Clinical Endorsement

> *A reliable approach would be to offer a paid chat feature that connects users with a live doctor, allowing them to verify whether the app's results align with a medical professional's diagnosis.* [P12]

Participants strongly favored hybrid models combining AI efficiency with clinical validation, aligning with the broader principle of human-in-the-loop explainability.

XSL•FO
RenderX

Personalized and Longitudinal Tracking

> *The app should include a section to log symptoms and track health records for more personalized guidance.* [P3]

Customization and data continuity were seen as critical for relevance and user confidence.

User Feedback Loops and Adaptive Systems

> *Allowing users to provide feedback on whether the recommendations were helpful or accurate could help the AI learn and improve over time.* [P80]

Participants envisioned bidirectional learning, where user input refines future recommendations and ensures the system stays aligned with local clinical standards.

Overall, users valued AI-driven mHealth apps for their accessibility and informational benefits but were cautious about relying on them without human confirmation. Trust hinged on transparent reasoning, contextual clarity, and human oversight. Explainability was not merely about exposing algorithms; it needed to reflect social credibility, cultural fit, and communicative empathy.

These insights underscore that in South Asia (as well as in the Global South), explainability must move beyond technical interpretability to socially meaningful transparency, where AI recommendations are understandable, verifiable, and situated within trusted human frameworks.

Expert Perspectives on Explainable AI in mHealth Apps

Overview

Interviews with 6 XAI experts (P1-P6) yielded rich insights into the role, limitations, and design priorities of explainability in AI-driven mHealth systems. Participants represented diverse disciplinary and geographic backgrounds spanning AI research, clinical integration, and product design for low-resource settings such as Bangladesh. Thematic analysis revealed several interconnected themes: trust, explanation modality, interpretability and transparency, personalization and contextualization, the role of medical professionals, technical limitations and ethics, and design priorities.

Trust as the Cornerstone of Explainability

All experts agreed that user trust fundamentally determines whether AI recommendations are accepted or dismissed. Trust, however, was described as socially situated rather than purely technical. P1 observed that “when people can relate their thinking and the app decision, the user’s trust increases.” P5 added that trust grows when recommendations are “short, reference-based, and clearly attributed to professionals, like-‘doctor said it, so people trust it.’”

Several experts linked trust to repeated accuracy and social validation. P4 noted that even clinicians “begin to trust tools like GPT after seeing multiple times correct results.” Yet, the contextual fragility of trust was evident in P3’s account: “In Bangladesh, especially low-literate people do not trust doctors… so how can this group trust AI apps?” Trust was therefore portrayed as a fragile construct influenced by professional reputation, algorithmic credibility, prior experience, and broader social attitudes toward health care.

Preferred Forms of Explanation

Experts emphasized that effective explainability depends on how information is communicated. Participants favored concise, reference-supported, and context-sensitive explanations. P1 suggested highlighting relevant data: “It should show the relevant data part, like for COVID.” P2 added that explanations should cite evidence: “Just showing the problem is not enough; references should also be given.”

Multimodal presentation was strongly advocated. Visual explanations, such as CAM-based overlays or color-coded graphs, were recommended by P1, P2, and P5 to simplify complex model outputs. P6 proposed “videos or images” for low-literacy users, while P4 recommended “case-study-based explanations” and conversational, narrative formats that mimic physician dialogue. Comparing GPT interactions with brief doctor consultations, P4 noted, “A doctor gives five minutes, but I can talk with GPT for an hour.”

These perspectives indicate a preference for dialogue-driven, story-like explanations that engage users emotionally and cognitively, rather than abstract technical justifications.

Interpretability and Transparency

Participants distinguished between interpretability as understanding how the model functions and explainability as understanding why it produced a given output. P1 stressed that most mHealth apps “fail because their internal decision-making is opaque.” They argued that interpretability requires insight into the model’s structure and data sources: “Explainability is based on which data.”

Experts also cautioned that existing tools, such as LIME and SHAP, remain developer-centric. P1 stated bluntly, “They are for developers, not for users.” P3 highlighted “feature-changing issues and bias” as obstacles to consistent interpretation.

Transparency was universally seen as beneficial yet context-dependent. P1 remarked, “People believe what they can understand by nature,” while P5 added, “If the app shows how, it reaches its conclusion, that increases trust.” However, P1 warned that excessive or poorly framed transparency could backfire in misinformation-prone environments: “This is directly related to taboo. For misinformation, social media plays a role.”

Overall, experts advocated layered transparency, offering high-level summaries for general users and deeper interpretive options for clinicians or developers.

Personalization and Contextualization

All experts emphasized the importance of tailoring explanations to user literacy, language, and context. Contextual data such as GPS, demographics, and prior history were viewed as crucial for both diagnostic accuracy and relevance. P4 illustrated this need: “In the query, I have to provide the context, like GPS, profile, all in one, to get the correct result.”

XSL•FO
RenderX

P5 proposed a 2-step communication model: first, deliver the analytic result, then supplement it with a context-aware explanation that accounts for individual circumstances. P1 similarly suggested offering "multiple options in a single app" to accommodate diverse user backgrounds.

Inclusivity emerged as a recurring concern. P1 noted a "major opportunity to work with illiterate people," while P3 cautioned that most current systems are "made for high-rich people," excluding low-income and low-literacy groups. Experts thus framed personalization not merely as a usability enhancement but as an ethical imperative for equitable AI.

Integrating Medical Professionals

Experts strongly endorsed involving clinicians as validators and communicators within AI systems. P1 proposed that "app developers can hire doctors to validate," and P4 envisioned "a live server connection with a doctor after several conversations."

Several participants (P2, P5, P6) advocated for embedded live-chat features with health care professionals to enhance trust. As P5 stated, "If a doctor says it, people trust it more; if a developer says it, no one will." P6 described this feature as promising, though logistically challenging.

Interestingly, P4 observed that XAI could also fill gaps in clinical communication: "In Bangladesh, doctors don't explain much. If the app can explain properly, it will close this gap and increase trust." These insights support the notion of human-AI collaboration, where explainable systems complement rather than replace professional judgment.

Technical Limitations and Bias

Participants openly discussed the technical fragility of current mHealth AI systems. Bias in training data, feature drift, and limited generalization were cited as major causes of diagnostic error. P3 referred to persistent "feature-changing issues" and "biased results that make explanations difficult."

Several experts described failure cases, such as misdiagnoses in fracture detection:

> *AI made the wrong detection in the fracture case. doctors use cumulative knowledge, but apps use only trained data.* [P1, P2, P3]

P6 criticized generic outputs like "These are all possible causes; seek medical help," arguing that such vague responses undermine informed decision-making.

Experts thus underscored the need for robust datasets, domain adaptation, and explainability techniques that acknowledge uncertainty rather than overstate confidence.

Ethics, Privacy, and Governance

Concerns over privacy and ethics were universal. Experts stressed that user satisfaction should never compromise confidentiality. As P4 asserted, "Confidential information can't be shared." P1 acknowledged regulatory progress, "rules are updating regularly," while P3 added, "If the company maintains privacy, then it's okay."

However, reliance on company-level policies was viewed as insufficient; participants emphasized the need for clear governance, data use, accountability, and regulatory enforcement. The discussions revealed that ethical explainability, making not only decisions but also data practices transparent, is crucial to sustaining public trust.

Design Priorities for Inclusive and Actionable Explainability

Experts called for participatory, feedback-driven design that accommodates varied literacy levels and promotes ongoing engagement. P5 urged developers to "deploy the system to minor groups and see their thinking," highlighting participatory evaluation. P6 pointed to missing feedback loops: "Every hour the app says drink water, but it never asks if I had water later."

Recommendations converged on several principles:

- Interactive feedback mechanisms so users can confirm, question, or rate recommendations.
- Multioption, inclusive interfaces accommodating text, audio, and visual modalities.
- Localization through language support, cultural cues, and regionally relevant medical references.
- Purpose-specific design, ensuring explainability features match the clinical or advisory scope of the app.

While some experts prioritized data-driven transparency (P1, P2), others favored narrative and conversational forms of explanation (P4, P6). This divergence underscores the need for adaptive explainability, offering multiple modes suited to both lay users and professionals.

Developer Perspectives on Explainable AI in mHealth Apps

Overview

Interviews with 4 AI application developers (AD1-AD4) provided practical insights into the feasibility, value, and challenges of implementing XAI in mHealth apps within Bangladesh. Developers reflected on technical limitations, contextual barriers, and design priorities shaped by local infrastructural and user constraints. Eight dominant themes emerged: developer skepticism and need for validation, gaps in explanation features, designable explanation mechanisms, traceability and transparency, user-centric adaptation, localization barriers, security and data control, and the feasibility of hybrid human-AI systems.

Developer Skepticism and Human Validation

Across all interviews, developers expressed conditional trust in AI-generated health recommendations. They considered AI a useful assistive tool but insisted that outputs must be validated by medical professionals. AD1 emphasized that guidance is "reliable and acceptable if doctors verified," while AD3 admitted, "As developers, we don't trust AI fully; we verify."

AD2 noted that "AI hasn't reached expert doctor level" and should therefore be restricted to "normal situations." Even AD4's neutral comment that AI is "good at performing work" lacked endorsement of its medical reliability. Collectively, developers framed trust as contingent upon expert validation, reinforcing the need for human oversight in AI-driven health systems.

XSL•FO
RenderX

Gaps in Explanation Features

Developers uniformly acknowledged that explainability remains largely absent from current mHealth apps. AD1 candidly stated, “Explainability-focused work was not done,” while AD2 added, “No explanation-based app I made.” AD3 observed that most apps provide only “minimal explanation,” and AD4 confirmed inconsistency: “Some apps give hints or descriptions; some don’t.”

These remarks reveal that explainability is often deprioritized during development, overshadowed by core functionality and time constraints. As a result, most deployed mHealth systems in Bangladesh lack consistent or meaningful explanation features, leaving users with a limited understanding of how AI-generated advice is produced.

Designable Explanation Mechanisms

Despite current gaps, developers proposed actionable strategies to improve XAI integration. Their suggestions emphasized multimodal, interactive, and context-sensitive explanation systems. AD1 recommended incorporating “LLM-based reasoning” to generate natural-language explanations that feel human and conversational. AD2 proposed a tiered design with “a button to expand features” and options for “short or long answers” supported by external resources.

AD3 described an ideal interface that “shows how the decision was made, visuals, text, video, or audio,” while AD4 advocated contextual modality: “During morning walk, I prefer audio; other times maybe visuals.” They also noted that “the image is good, but text should also be included.”

Together, these ideas point to interactive and multimodal explainability, where users can select how and when explanations are presented, balancing cognitive load with accessibility.

Traceability and Transparency

Transparency was recognized as essential for both user trust and developer confidence. AD1 simply asserted “transparency” as a trust factor, while AD4 confirmed, “Yes, I want to know,” when asked if visibility into model reasoning mattered. Developers viewed transparency as a shared requirement across user and technical domains; it enables debugging, verification, and accountability.

This emphasis on traceability suggests that explainability must function not only as a user interface feature but also as an engineering principle, helping developers monitor model behavior and assure quality in sensitive contexts such as health advice.

Customizable and User-Centric Design

Developers favored adaptable systems that personalize explanations based on user context and literacy. AD3 summarized this vision: “Users can choose what type of explainability they want.” Similarly, AD2 emphasized tailoring output “based on user level” and noted that explanation mechanisms should “improve over time” through user interaction.

AD4 added that explanations could be filtered by situational context to avoid overwhelming users. These perspectives converge on a model of adaptive explainability, where AI dynamically adjusts the complexity and modality of its explanations to fit the user’s background, language, and situational needs.

Localization and Usability Barriers

All developers highlighted the importance of localization for mHealth adoption. Language diversity, digital literacy, and socioeconomic factors were identified as key constraints. AD1 recommended “multiple languages and tutorials needed” for onboarding, and AD2 argued that systems must reflect “how people understand local healthcare.”

Financial accessibility also emerged as a concern. AD4 proposed a donation-based model: “Affordability should be addressed. Donations from those who can afford it will support those who can’t.”

These reflections underscore that effective explainability cannot be detached from cultural and economic realities; it must be embedded in an inclusive design that aligns with local capacities and norms.

Security, Device-Level Control, and Data Integrity

Developers expressed awareness of privacy and data security challenges, particularly regarding remote processing. AD1 advocated “on-device models” with “encrypted data,” while cautioning that “before explainability, we need to ensure awareness among users.”

AD3 warned that “API can be intercepted,” underscoring risks in cloud-dependent architectures. By contrast, AD4 claimed “no privacy concerns,” reflecting either overconfidence or limited exposure to data governance issues.

Collectively, these insights highlight the dual nature of security in mHealth XAI: it is both technical (encryption, on-device processing) and social, requiring user awareness and consent.

Feasibility of Human–AI Hybrid Systems

Developers diverged sharply on whether integrating medical professionals into AI workflows enhances or undermines system value. AD1 advocated hybrid models in which “doctors are in the loop, governed, and revenue given.” AD3 similarly envisioned “initial diagnosis from doctor, follow-up via app,” conditional on sustainable funding.

Others were skeptical. AD2 asked, “If I need to invite professionals, why use AI at all?” questioning the efficiency of hybrid systems. AD4 raised concerns about authenticity and liability: “We need to verify the actual doctor is behind the system. Developers won’t take responsibility if it’s fake.”

These conflicting views illustrate ongoing tension between automation and professional oversight, with feasibility hinging on governance, authentication, and clear role boundaries.

XSL•FO
RenderX

Key Informants' Perspectives on XAI in mHealth Apps

Overview

Insights from one in-depth interview and a focus group with 9 practicing physicians (D1-D10) in Bangladesh revealed cautious optimism toward the use of XAI in health care. While doctors recognized the potential of AI for clinical assistance, they emphasized that its use must remain under human supervision. Nine interrelated themes summarized their perspectives on trust, safety, localization, and explanation design.

Conditional Trust and Clinical Authority

Doctors expressed conditional trust in AI, consistently prioritizing clinical judgment over automated outputs. As one participant noted,

> *Even if AI gives a correct prediction, I will rely on my own judgment.* [D1]

Participants recounted errors such as misdiagnosed ECGs or inaccurate symptom assessments, which eroded confidence. Several warned that conflicting outputs between AI and physicians could confuse patients and undermine professional credibility:

> *If the app gives something different from what I say, the patient may think I'm wrong.* [D7]

Trust, therefore, hinged on AI complementing, not competing with, human expertise.

AI as Assistive, Not Autonomous

All participants viewed AI as a supportive tool rather than a replacement for clinical reasoning. It was seen as useful for tasks such as dosage verification and chronic disease monitoring, but unsafe for unsupervised use by patients. As D4 stated, "It can assist doctors, but for users, it is not feasible. They might get overwhelmed." This reflects a preference for physician-mediated AI, where apps serve as background aids for efficiency and reminders rather than independent diagnostic systems.

Deficits in Explanation and Reasoning

Doctors criticized current mHealth apps for lacking clear or actionable explanations. D3 remarked, "If I don't already know the topic, how will I judge if AI's answer is right?" While language models like ChatGPT could reference guidelines when prompted, most mobile apps offered opaque outputs with no interpretive reasoning. Explanation quality was thus linked to user expertise by limiting its accessibility to both nonspecialist clinicians and lay users.

Dependence on Input Quality and User Literacy

Participants underscored that AI reliability depends on the precision of user inputs, which most patients cannot provide. As D3 explained, "Chest pain could mean ten things; only a doctor can tell which by asking follow-up questions." Several doctors advocated for access restrictions or literacy-based user segmentation, arguing that open public use could heighten misinterpretation risks.

Cultural and Behavioral Barriers

Low public awareness of AI and entrenched cultural beliefs were cited as major barriers to adoption. "Most people don't know what AI is; that's why they don't use it," said D1. Others noted that people who distrust allopathic medicine were deemed unlikely to trust AI-driven advice to start medication in this direction, reflecting how social beliefs influence technology acceptance.

Localization and Contextual Relevance

Strong consensus emerged that Western-trained AI models fail to reflect Bangladesh's medical realities. D1 observed, "Lesions in BD patients are different; AI trained on Western data won't work." Participants stressed that systems must incorporate local disease profiles, drug availability, treatment costs, and national clinical guidelines to build trust and usability.

Governance and Ethical Oversight

Institutional endorsement was viewed as essential but insufficient on its own. "If the government released it, I'd compare its output with my own judgment," said D1, while D9 added, "Even if endorsed, the app must be for doctors, not for patients." Physicians called for strict regulation, data accountability, and clear boundaries defining who may use and interpret AI outputs.

Human-AI Collaboration and Live Consultation

Doctors generally supported live chat or hybrid systems combining AI with human input for mild or follow-up cases. However, they cautioned that remote advice cannot replace physical examination:

> *Only by talking, without examining, can we diagnose.* [D8]

Compensation and verification mechanisms for participating doctors were seen as essential to sustain such systems.

Preferred Explanation Design

Participants favored concise, prioritized outputs that present ranked differential diagnoses linked to guidelines and sources. As D3 urged, "Don't give 10 - 20 diagnoses; show what's most likely and explain why." Clinically structured, evidence-linked explanations were considered necessary to support both doctors' decision-making and patient understanding.

Overall, medical professionals characterized explainability as a clinical, not merely a technical, requirement. They envisioned AI as a supervised decision-support tool grounded in localized data, professional validation, and clear communication of uncertainty. Explainability, from their perspective, is essential for ensuring safety, contextual accuracy, and maintaining public trust in digital health systems.

Stakeholder Gap Analysis: Misalignments in XAI Expectations for mHealth Apps

Overview

This analysis integrates insights from users, developers, XAI experts, and medical professionals to identify where expectations diverge regarding XAI in mHealth apps. Although all

XSL•FO
RenderX

stakeholders recognize the importance of explainability for trust and adoption, their perspectives differ sharply in validation standards, explanation preferences, and perceived responsibility. Understanding these misalignments is essential for designing inclusive, context-appropriate XAI systems.

Trust and Validation

Trust remains the most consistent yet variably defined factor across stakeholders. Users often express willingness to accept AI recommendations when accompanied by simple reasoning or credible sources (eg, "helpful if the app could provide reasoning, including relevant studies" [P35]). In contrast, medical professionals adopt a far stricter stance: "Even if the AI gives a correct prediction, I will rely on my own judgment" (D1), underscoring the clinical accountability gap between end users and physicians. Developers and experts align more closely, viewing trust as contingent on transparency and iterative validation (AD3; P1). However, developers emphasize practical verification ("we don't trust AI fully, we verify") rather than the epistemic or cognitive alignment valued by experts.

Explanation Depth and User Segmentation

Stakeholders differ in how complex explanations should be. Users prefer brief, plain-language rationales describing potential risks and benefits (P77), while experts advocate multimodal or narrative explanations (eg, case-based or visual) (P2, P4). Developers propose user-selectable formats, "short or long explanations" (AD2) and adaptive delivery "based on user level." Medical professionals, however, argue that explainability should be contingent on user capability:

> *Apps should restrict who can use them based on education.* [D6]

These perspectives reveal that technical flexibility alone cannot solve the deeper issue of cognitive diversity across audiences.

Technical Implementation and Usability

Experts critique existing explainability tools such as LIME and SHAP as "for developers, not for users" (P1), whereas developers acknowledge that "explainability-focused work was not done" (AD1). This indicates a persistent translation gap between technical affordances and user-centered design. While users value understandability, developers confront resource constraints, making advanced interpretability difficult to operationalize. Hence, current mHealth AI systems often underdeliver on transparency despite available techniques.

Human–AI Collaboration and Accountability

All stakeholder groups endorse human involvement but differ in form and feasibility. Users favor AI-guided systems with "paid chat features that connect with live doctors" (P12). Experts also recommend doctor-in-the-loop approaches (P1), while physicians express caution, like "Only by talking, without examining, we can't diagnose" (D8), stressing physical evaluation as indispensable. Developers, however, question the practicality of hybrid models: "If professionals are required, why use AI at all?" (AD2). Responsibility and liability further complicate collaboration: doctors refuse accountability for unverified AI advice, and developers disclaim legal responsibility ("developers won't take responsibility if it's fake" [AD4]). This shared reluctance leaves accountability diffuse, a challenge for regulatory integration.

Localization and Cultural Context

Localization emerged as a shared priority, but with different emphases. Users implicitly expect culturally aligned recommendations, while experts advocate context-aware features integrating GPS and profile data (P4). Developers focus on multilingual interfaces and affordability (AD1 and AD2), and medical professionals emphasize clinical relevance:

> *Lesions in BD patients are different; AI trained on Western data won't work.* [D1]

Misalignment here reflects different interpretations of "localization": for experts and developers, it is technical adaptation; for clinicians, it is diagnostic and therapeutic validity.

Privacy, Governance, and Sustainability

Privacy received uneven attention: experts stressed ethical handling of "confidential information" (P4); developers were split between advocating encrypted, on-device models (AD1) and indifference (AD4); doctors emphasized institutional oversight rather than data protection. Economic sustainability also revealed friction: users expect free services, doctors require compensation for hybrid systems (D10), and developers proposed donation or subsidy models (AD4). This underscores that ethical and financial sustainability are as crucial as algorithmic transparency.

Table 5 provides a consolidated view of stakeholder perspectives across key explainability dimensions. It highlights both overlapping concerns, such as the demand for contextualized explanations, and critical misalignments, particularly around trust, responsibility, and economic feasibility.

XSL•FO
RenderX

Table . Stakeholder perspectives on key explainable AI (XAI) dimensions in mobile health (mHealth) apps.

Dimension	Key insights	Representative quotes / Examples	Stakeholders
Trust in AI[a]	Users accept simple explanations; doctors require clinical validation; developers verify; experts stress alignment with human reasoning.	"Helpful if the app could provide reasoning... including relevant studies." [P35] "Even if AI gives a correct prediction, I will rely on my own judgment." [D1] "As developers, we verify." [AD3]	Users, Doctors, Developers, Experts
Explainability format	Users prefer clear text; doctors need structured logic; developers rely on UI[b] features; experts suggest visual and narrative formats.	"Clear, concise explanation of potential risks and benefits." [P77] "Case study-based explanations..." [P2] "Provide prioritized differential diagnoses." [D3]	Users, Experts, Doctors, Developers
Personalization and context	Experts suggest context-aware systems using GPS and profile data; developers mention adaptive explanation levels; doctors highlight age/literacy needs.	"Apps should restrict use based on age or education." [D6] "Explanation delivery should improve over time." [AD2] "Include GPS, profile to get correct result." [P4]	Doctors, Developers, Experts
Explainability tools gap	Tools like LIME[c] and SHAP[d] are not accessible to end users; developers admit little use of explainability tools.	"They are for developers, not for users." [P1] "Explainability-focused work was not done." [AD1]	Experts, Developers
Human-AI collaboration	Users favor doctor-in-the-loop models; experts support it; doctors warn about lack of physical exam; developers see conflict with AI autonomy.	"Offer paid chat feature with live doctor." [P12] "Only by talking, without examining, we can't diagnose." [D8] "If I need professionals, why use AI?" [AD2]	Users, Doctors, Developers, Experts
Responsibility and liability	All stakeholders avoid full accountability for AI errors; doctors want clear validation; developers worry about legal risks.	"Developers won't take responsibility if it's fake." [AD4] "I won't be held liable without proper validation." [D3]	Doctors, Developers
Localization and cultural fit	Doctors cite mismatch between AI training and local clinical conditions; developers focus on language/UI; experts urge regional adaptation.	"Lesions in BD patients are different, AI trained on Western data won't work." [D1] "Understand how people interpret local healthcare." [AD1]	Doctors, Developers, Experts
Privacy concerns	Users rarely mention it; developers split on its importance; doctors prioritize clinical responsibility; experts warn about ethics.	"Confidential information can't be shared." [P4] "On-device model, encrypted data." [AD1]	Users, Experts, Developers, Doctors
Cost and sustainability	Users expect free access; developers suggest donation models; doctors require fair compensation; experts rarely discuss economics.	"Doctors need remuneration if we're asked to be involved." [D10] "Donation model for those who can't pay." [AD4]	Users, Developers, Doctor

[a]AI: artificial intelligence.

[b]UI: user interface.

[c]LIME: local interpretable model-agnostic explanations.

[d]SHAP: Shapley additive explanations.

Overall, while all stakeholders converge on the value of explainability, they diverge on what "good explanation" means in practice. These tensions highlight that effective XAI design in health care must be co-created with all actors, balancing technical transparency with human judgment, cultural nuance, and systemic accountability. This analysis directly informs the framework proposed in the following section, which aims to reconcile these gaps through participatory, context-sensitive design principles.

XSL•FO
RenderX

Proposed Framework for Explainable AI in mHealth Apps

Overview

Building upon insights from various stakeholders, we developed a conceptual 5-pillar framework for explainability in AI-driven mHealth apps (Figure 4). The framework positions explainability as a dynamic process emerging from interaction among human understanding, algorithmic transparency, and institutional accountability.

Figure 4. Hierarchical proposed framework for explainable mHealth apps. AI: artificial intelligence; mHealth: mobile health; XAI: explainable AI.

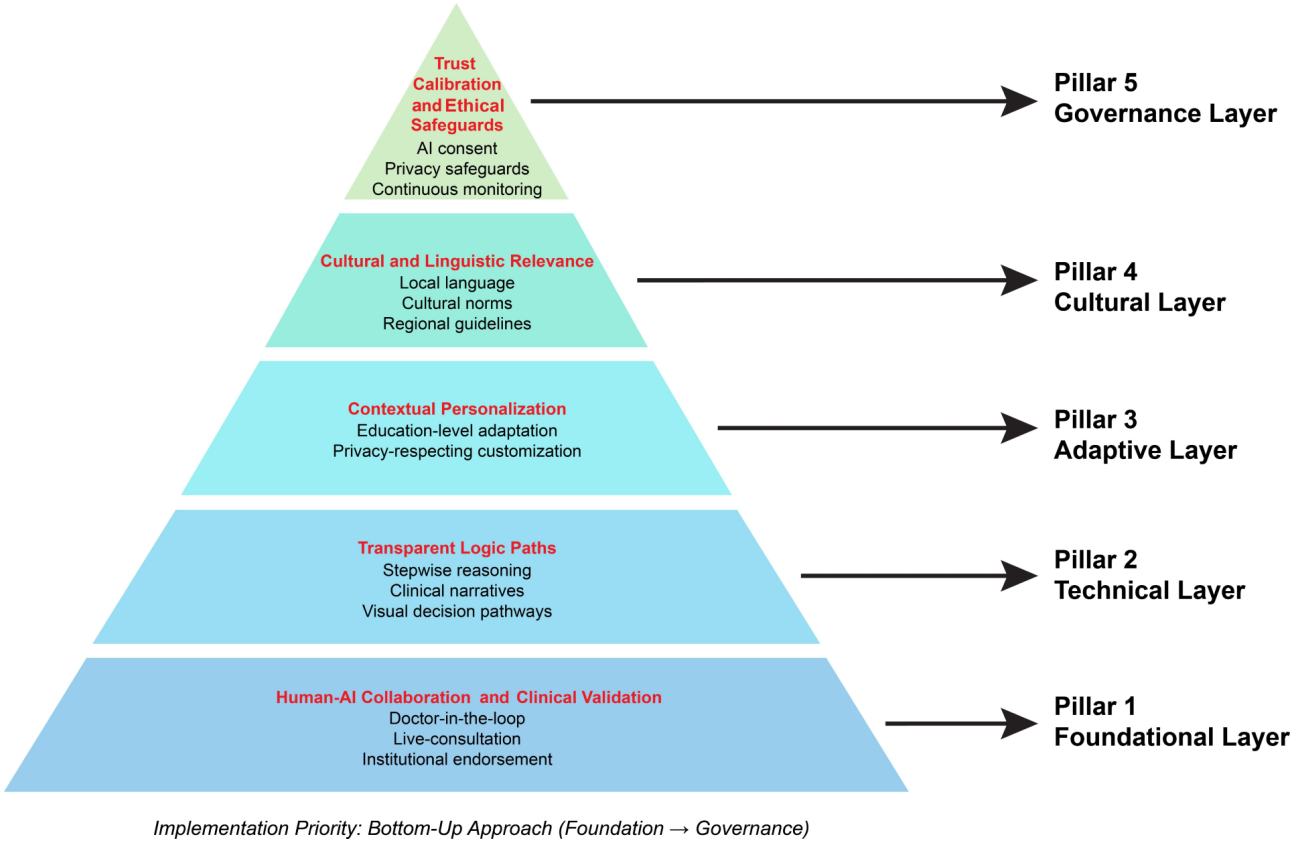

Pillar 1: Human-AI Collaboration and Clinical Validation

All stakeholder groups prioritized human oversight over autonomous AI decision-making. Medical professionals emphasized that clinical reasoning must remain the final arbiter of health recommendations. Users preferred apps offering access to professional consultation for validation. Developers and experts endorsed “doctor-in-the-loop” models integrating live chat, institutional endorsement, or verification mechanisms to anchor trust and mitigate overreliance risks.

Pillar 2: Transparent Logic Paths

Participants sought clarity about how AI systems reached conclusions, valuing stepwise cause-and-effect reasoning over abstract confidence scores or probability distributions. Users wanted narratives mirroring clinical thought processes (eg, “you reported these symptoms; therefore, these conditions are likely”) rather than technical explanations. This principle requires intuitive reasoning traces, visual summaries, or simplified decision pathways, making the model’s internal logic observable.

Pillar 3: Contextual Personalization

Users expected AI systems to reflect personal health profiles, local treatment access, and daily routines. Developers emphasized that personalization enhances perceived relevance and trust when privacy is safeguarded. The framework envisions adaptive explanation systems adjusting depth, tone, and modality (text, visual, and audio) based on user literacy, familiarity, and prior interaction as identified through stakeholder consultation. Stakeholders emphasized that explainability requirements vary substantially across users, with personalization extending to clinical adaptation considering region-specific disease prevalence, local drug availability, and affordability.

Pillar 4: Cultural and Linguistic Relevance

Users and clinicians highlighted how Western-centric terminology, English-only interfaces, and unfamiliar medical constructs can alienate users. Trust depends on linguistic and cultural resonance. This principle requires explanations in native languages supported by voice narration, culturally familiar metaphors, examples from local health practices, and integration of national clinical guidelines and approved medication lists.

Pillar 5: Trust Calibration and Ethical Safeguards

Users expressed privacy concerns, while experts emphasized communicating AI limitations clearly. This pillar advocates for explicit confidence indicators (“high certainty” and “low data confidence”), clear data use disclosure, visible ethical endorsements from professional or governmental bodies, feedback mechanisms for flagging incorrect outputs, regular

XSL•FO
RenderX

audits, and public reporting of accuracy metrics establishing accountability beyond interface design into governance.

Pillar Relationships and Implementation Dependencies

While Figure 4 presents the pillars in a circular arrangement emphasizing their interconnectedness during operation, successful implementation requires understanding their dependencies. Pillar 1 (Human-AI Collaboration) serves as the foundational layer with no prerequisites: stakeholders universally prioritized human oversight as nonnegotiable. Pillars 2 (Transparent Logic) and 3 (Contextual Personalization) build upon this foundation and can be developed in parallel, addressing complementary needs: how the AI reasons and how explanations adapt to individual users. Pillar 4 (Cultural-Linguistic Relevance) transforms how existing functionality is expressed within local contexts, requiring the lower pillars to function properly first. Pillar 5 (Trust Calibration and Ethical Safeguards) serves as the governance layer monitoring the complete system, validating that human-AI collaboration follows ethical guidelines, transparent logic reveals true reasoning, personalization respects privacy, and cultural adaptation is authentic. This hierarchical structure informs implementation priorities: establish human oversight first (Pillar 1), build technical and adaptive capabilities (Pillars 2 - 3), integrate cultural context (Pillar 4), and implement continuous governance (Pillar 5). We need to provide AI-specific informed consent explaining how algorithms use the GPS, health history, and demographics to generate explanations; data minimization with on-device processing where feasible; user control over personalization settings; and adherence to frameworks, including the European Union's AI Act's requirements for high-risk health AI systems [36] and WHO guidance on AI ethics [6]. Continuous monitoring must audit for potential harms, including privacy breaches, discriminatory personalization, and trust miscalibration through regular privacy impact assessments.

Framework Synthesis

The framework addresses identified stakeholder misalignments by reconciling users' need for accessible explanations with clinicians' demand for evidence-based validation, developers' feasibility constraints, and experts' interpretability requirements. These 5 principles form an interconnected system where user experiences refine explanations, clinical validation improves reliability, and contextual adaptation enhances engagement.

Discussion

Principal Findings

This study makes 5 principal contributions. First, education level significantly predicted trust in AI explainability ($F_{3,\ 133}$=2.81, P=.042), with critical thinking skills mattering more than demographic factors. Second, systematic stakeholder misalignments emerged: users prioritized human validation, developers faced resource constraints limiting XAI implementation, clinicians demanded clinical validation pathways, and experts confirmed technical methods (LIME, SHAP) inadequately serve lay users. Third, all stakeholders universally prioritized human oversight over autonomous AI, validating "social explainability" [8] in health care contexts. Fourth, current technical XAI methods fail to translate into comprehensible explanations for users with varying education levels. Fifth, we developed a five-pillar participatory framework addressing these gaps: (1) human-AI collaboration, (2) transparent logic paths, (3) contextual personalization, (4) cultural-linguistic relevance, and (5) trust calibration with ethical safeguards.

Interpretation of Statistical Findings

Education level significantly predicted trust in AI explainability ($F_{3,\ 133}$=2.81, P=.042, η^2=0.060 vs $F_{3,\ 133}$=0.62, P=.60), challenging "digital native" assumptions about generational technology adoption differences. The education-trust relationship revealed an unexpected nonlinear pattern. Completed undergraduate students reported lower trust (mean 3.14) than current undergraduates (mean 3.66) and higher secondary students (mean 4.67). This "critical evaluation dip" may reflect heightened critical thinking skills without yet gaining specialized knowledge that contextualizes AI's appropriate health care role. Postgraduate holders showed higher trust (mean 3.85), possibly reflecting greater AI exposure, and understanding of AI as an assistive rather than a replacement tool. The small effect size (η^2=0.060) indicates education accounts for only 6% of variance, suggesting multiple factors influence trust. The absence of age-related differences (P=.60) was contrary to expectations. Despite 67.9% of participants being aged 18 - 24 years, trust remained consistent across age groups, suggesting explainability requirements are universally important rather than generation-specific. Chi-square analyses reinforced these patterns: education level (χ^2_6=17.08, P=.009) and XAI familiarity (χ^2_4=11.50, P=.02) significantly predicted belief in app decisions. Postgraduate holders showed 69% belief rates versus 43% among completed undergraduates. Notably, users "somewhat familiar" with XAI showed the lowest belief (33%), suggesting superficial knowledge increases skepticism without providing sufficient context, indicating explainability education must be comprehensive rather than superficial.

These findings have important design implications. Rather than age-based personalization, mHealth systems should adapt explanations based on educational background and critical thinking orientation. Apps should include educational components explaining what explainability means and why it matters. These quantitative patterns align with qualitative findings, where recent graduates expressed heightened skepticism while postgraduate holders demonstrated more nuanced AI understanding, strengthening confidence in our conclusions.

Contextual and Relational Explainability

Users appreciated mHealth accessibility but doubted recommendations, citing vague, culturally disconnected outputs. Trust depended on understandable, localized explanations endorsed by professionals, extending prior work [47,51] while highlighting how explainability is reshaped in South Asia, where health literacy and infrastructure vary widely [49,55]. Experts and clinicians suggested that visual explanations and culturally grounded metaphors may be more effective than statistical confidence levels for users with varying literacy levels. These

XSL•FO
RenderX

findings parallel Ismail and Kumar's [56] work with Indian community health workers who require sociocultural context and human mediation, as well as Nigerian/Kenyan digital health, which faces similar infrastructure constraints and language gaps [18,49]. Our framework's human oversight and cultural adaptation principles appear transferable across South Asian contexts while requiring local implementation adaptation.

User Focus on Output Explainability Rather Than Technical Implementation

A critical finding emerged from comparing user responses across the 3 apps. Despite Ada using Bayesian reasoning, Symptomate using moderate algorithmic transparency, and WebMD relying on conventional rule-based logic, users showed no significant differences in trust ratings across these apps. This was not a limitation of our study design but rather an intentional research decision that revealed fundamental insights about what users care about in health technology.

We deliberately selected these 3 apps with varying technical architectures to test whether users differentiate based on underlying AI implementation. The results definitely show they do not. Users are not concerned with whether an app uses ML, Bayesian networks, or traditional algorithms. Instead, they evaluate systems based on whether the output is explained clearly and whether it can be validated by trusted human authorities.

This finding directly supports our framework's design priorities. Users consistently demanded the same things across all 3 apps, regardless of technical implementation: clear explanations of recommendations, understanding of why specific conclusions were reached, and access to doctor validation. Whether the system used sophisticated AI or simple rule-based logic made no difference to users. What mattered was whether they understood the output and could get human confirmation.

This validates our framework's structure. We positioned Human-AI Collaboration as the foundational first pillar precisely because users prioritize human validation over algorithmic sophistication. The technical architecture is invisible and irrelevant to users. They want to know what the system concluded, why it matters for their health, and whether a doctor agrees. Similarly, our fourth pillar emphasizes Cultural-Linguistic Relevance because users need explanations in familiar terms and local context, not technical details about model architecture.

The lack of differentiation across apps reveals that the AI explainability problem in health care is fundamentally about human communication, not technical transparency. Users do not want to see LIME visualizations or SHAP values. They want plain language explanations and professional endorsement. Our framework addresses this by prioritizing social explainability mechanisms (human validation, clear communication, and cultural adaptation) over purely technical interpretability approaches.

This finding challenges the dominant narrative in XAI research that focuses on making algorithms more transparent. Our results suggest that for health care users, algorithmic transparency is neither necessary nor sufficient for trust. What users need is output transparency combined with human oversight, which is exactly what our framework provides through its hierarchical pillar structure.

Economic Feasibility of Human-AI Collaboration

Developer stakeholders questioned the financial viability of human-in-the-loop systems, noting that paying clinicians contradicts mHealth's cost-saving purpose. On the other hand, the users were willing to pay for expert advice. However, multiple business models can be adapted to achieve economic viability.

A tiered service model could offer users choices between paid premium services (immediate doctor consultation) and free government or NGO (nongovernmental organization)-funded validation through layered review (community health workers escalating to physicians) or asynchronous consultation within 24 - 48 hours. This freemium approach, where paying users subsidize infrastructure while disadvantaged users access free services, is common in digital health platforms globally and could be adapted for AI-validated mHealth.

Bangladesh Government infrastructure already supports subsidized services. The Ministry of Health received 41,908 crore BDT (US $3.5 billion) in FY2025-26, with 4166 crore BDT (US $350 million) for free services [69]. The Digital Health Strategy operates 96 telemedicine centers [70], and Shasthyo Batayon handled 10 million calls at 3.6 BDT (US $0.0289) per consultation during COVID-19 [71]. WHO, United Nations Children's Fund, United Nations Population Fund [72], World Bank (US $200 million) [73], and NGOs like BRAC [74] provide additional funding streams that could support free validation.

Additional models include subscription-based access, insurance integration, and risk stratification that reserves physician validation for high-risk cases while AI handles low-risk assessments.

Sociocultural Dimensions and Stakeholder Gaps

Stakeholder expectations diverged: users preferred concise, reassuring messages; clinicians sought traceable logic aligned with guidelines; developers prioritized scalability; and experts stressed ethical clarity. These divergent priorities confirm that explainability must function simultaneously as a cognitive, ethical, and cultural interface.

Design Trade-Offs and Implications

Developing explainable mHealth AI involves unavoidable trade-offs: simplicity improves comprehension but risks omitting nuance; personalization increases relevance but raises resource demands; transparency can overwhelm low-literacy users; and cultural adaptation strengthens trust but challenges standardization. These tensions necessitate layered, adaptive explanations varying in depth according to user capability and task sensitivity, a design principle embedded in our 5-pillar framework integrating human-AI collaboration, transparent logic, contextual personalization, cultural alignment, and ethical safeguards.

XSL•FO
RenderX

Practical and Ethical Relevance

The framework's contribution lies in translating explainability into a responsible design principle. It operationalizes human oversight, transparency, and ethical accountability within mHealth ecosystems where literacy and infrastructure constraints persist. Developers in the study confirmed the feasibility for incremental deployment, starting with foundational safeguards, then progressively adding personalization and cultural adaptation. This approach promotes long-term sustainability while addressing the AI equity gap by enabling inclusive, interpretable, and locally meaningful digital health tools.

Limitations and Future Directions

Sample Representativeness

A critical limitation is the discrepancy between our motivating context (low-resource, underserved populations) and actual participants (predominantly young, highly educated, and English-proficient). This gap exists because evaluated apps (Ada, WebMD, Symptomate) are English-only; Bangla-language mHealth apps with XAI features do not currently exist. This technological constraint necessitated recruiting English-proficient participants representing educated early adopters rather than the general population, rural communities, or low-literacy groups who would benefit most from accessible mHealth solutions. Another limitation concerns sample size imbalance across education categories. The Higher Secondary group contained only 3 participants, while the Undergraduate (running) group had 107 students. Although the ANOVA showed statistical significance ($F_{3,\ 133}$=2.81, P=.042), the extremely small Higher Secondary sample means this finding should be considered anecdotal rather than definitive. The more reliable observation is that completed undergraduate students showed lower trust (mean 3.14, SD 1.10) compared with current undergraduates (mean 3.66, SD 0.93), as both groups had adequate sample sizes for comparison. Future research needs larger and more balanced samples across all education levels to confirm whether education truly predicts trust in AI explainability.

Selection Bias

Interview and focus group participants (n=20) were recruited through professional networks and snowball sampling, likely selecting technology-interested individuals, and over-representing optimistic perspectives. The "cautiously optimistic" characterization should be interpreted; accordingly, real-world skepticism may be greater than findings indicate. Based on expert and clinician insights, framework applicability to low-literacy populations may require adaptations: visual/audio explanations, oral health traditions, local language dependence, and community health worker intermediaries. Direct validation with low-literacy users is needed. This study was conducted in Bangladesh with educated, urban participants. Findings may not generalize across all South Asian contexts given regional variations in infrastructure, literacy, and cultural practices.

Language Barriers

Even among English-proficient participants, English-language interfaces may have acted as trust barriers distinct from AI opacity, reducing cultural resonance and increasing cognitive load. Trust deficits likely reflect both AI design inadequacies and linguistic-cultural barriers. Findings cannot be generalized to illiterate, semiliterate, or non–English-speaking populations whose perspectives remain critically underrepresented.

Methodological Limitations

Small subgroup samples (above 40 years [n=4]; higher secondary education [n=3]) limit statistical power. Education's small effect size (η^2=0.060) indicates it accounts for only 6% of trust variance, highlighting the influence of unmeasured factors. The developer sample (n=4) was small; findings represent preliminary insights requiring validation. Qualitative analysis used single-coder coding without independent inter-coder reliability testing, though peer debriefing was used.

Additionally, nearly half of the respondents (48.2%) were first-time users who installed the app immediately before survey completion. Their trust ratings likely reflect initial usability impressions rather than sustained trust developed through repeated use and verification of AI accuracy against actual health outcomes. Future research should examine how trust evolves across multiple usage sessions to distinguish between first-impression reactions and experience-based confidence.

Furthermore, WebMD uses rule-based logic rather than ML, meaning approximately one-third of participants (n≈46) evaluated a traditional algorithmic system rather than AI. Their trust and explainability ratings reflect mHealth interface usability rather than reactions to AI opacity. While we intentionally included WebMD to test whether users differentiate based on technical architecture, this means not all user responses address AI-specific explainability. The lack of differentiation between AI and non-AI systems is theoretically significant, but readers should note that the user sample includes evaluations of both system types.

Framework Validation

This framework represents a conceptual model synthesized from stakeholder requirements and existing XAI literature. While grounded in empirical multistakeholder data and theoretically informed by participatory design principles, the framework has not been empirically validated through expert evaluation (eg, cognitive walkthroughs, heuristic evaluation), prototype implementation, or usability testing with end users. Validation through expert review, pilot implementation, and user testing is required before operationalization in real-world mHealth apps to assess whether the framework successfully addresses identified gaps and whether users comprehend and trust explanations designed according to these principles.

Future Research Priorities

Urgent needs include (1) developing Bangla-language mHealth apps with culturally adapted XAI for low-literacy populations; (2) participatory design with illiterate, semiliterate, and non–English-speaking users to understand their unique explainability needs; (3) prototype implementation and expert evaluation of the framework; (4) longitudinal research evaluating how culturally tailored XAI influences health outcomes,

XSL•FO
RenderX

engagement, and trust over time; (5) cross-cultural validation across diverse Global South as well as South Asian contexts.

Comparison With Existing mHealth Apps

To assess current practices, widely used apps like Ada, WebMD, and Symptomate [59] were reviewed. These tools offer basic symptom guidance but lack structured explainability, localized communication, and clinical validation. None incorporates human oversight or clear reasoning paths, often providing opaque or English-only feedback. In contrast, the proposed framework directly addresses these deficiencies through clinician integration, traceable logic, contextual personalization, and ethical transparency. Table 6 summarizes these features.

Table . Comparison of existing mobile health (mHealth) apps with the proposed 5-pillar framework.

Pillars	Existing apps (Ada, WebMD, Symptomate)	Five-pillar framework
Human-AI[a] collaboration and clinical validation	No integration with human oversight. All 3 apps are fully automated with no clinician handoff or validation [75-77].	Encourages human-in-the-loop systems, including clinician referrals or expert review during critical decision points.
Transparent logic paths	Opaque decision logic. Ada uses Bayesian networks but does not show reasoning to users; WebMD and Symptomate provide no logic traceability [76-78].	Recommends user-facing, traceable decision paths through visual logic flows, confidence scores, or symptom highlighting.
Contextual personalization	Basic personalization (eg, age, gender). No adaptation to literacy, cultural norms, or prior user behavior [79,80].	Supports layered explanation complexity tailored to digital literacy, history of use, and cultural behaviors.
Cultural and linguistic relevance	Mostly English-based; limited localization. Explanations are generic, lacking cultural alignment [75-77].	Includes native language support, culturally familiar metaphors, and visual storytelling.
Trust calibration and ethical safeguards	Disclaimers are included, but lack detail on model accuracy, limitations, or fairness [81,82].	Emphasizes ethical design with clear disclaimers, data transparency, fairness disclosures, and user control.

[a]AI: artificial intelligence.

These contrasts underscore the practical value of the framework as a blueprint for improving explainability standards in mHealth. By situating AI explanations within social, cultural, and clinical realities, it offers a model that enhances trust, interpretability, and user safety.

Comparison With Existing mHealth Evaluation Frameworks

Table 7 situates our framework within the existing mHealth XAI literature, comparing this study's approach and contributions to 16 prior works across methodological, theoretical, and practical dimensions.

XSL•FO
RenderX

Table . Comparison of this study with prior explainable AI (XAI) and mobile health (mHealth) research.

Study	Context and sample	Methods	Key findings	Limitations	How this study advances globally
Kim et al (2024) [83]	South Korea; Clinicians & patients	Scenario-based needs assessment	Clinicians and patients have distinct explanation needs; designed stakeholder-tailored interfaces	Single disease (health management); 2 stakeholder groups only; no developers; controlled scenario	Multistakeholder including developers (who implement XAI); real-world mHealth apps; identifies implementation barriers (resource constraints, lack of tools); findings apply to any 2-sided health platform globally
Health care (MDPI) (2025) [84]	Western context: n=35 (5 experts, 30 users)	Usability evaluation of ADA, Mediktor, WebMD	All apps failed XAI heuristics; no confidence scores or rationales provided; critical transparency gaps	Identifies WHAT gaps exist but not WHY; no developer/clinician input; limited geographic diversity	Explains WHY XAI gaps exist through developer interviews (resource constraints, lack of localized tools); applicable to startups/SMEs[a] globally
Noor et al (2025) [85]	Global review; 1226 studies (2000 - 2023)	Systematic review	Categorized XAI methods (LIME[b], SHAP[c], GradCAM); technical taxonomy for health care AI[d]	Technical focus; no mHealth-specific analysis; no user comprehension studies	Validates user comprehension across education levels; demonstrates LIME/SHAP inadequate for lay users universally; education predicts trust (P=.042)
MDPI Study (2025) [86]	Global review; recent evidence (2024 - 2025)	Narrative review	Shift toward mental health AI; integration challenges identified across contexts	Review only; no actionable framework; no primary data	Provides actionable 5-pillar framework addressing integration challenges; stakeholder misalignments (universal across health care systems)
Doshi-Velez and Kim (2017) [11]	USA; conceptual	Theoretical framework	Defined interpretability evaluation criteria for ML[e]	No empirical validation; assumes universal technical audience	Empirical validation with diverse stakeholders; interpretability needs vary by education (universal), not geography
Rudin (2019) [13]	USA; conceptual	Position paper	Advocacy for inherently interpretable models versus post-hoc explanations	Theoretical; no user studies; model-choice focus	Users across education levels prefer interpretable outputs; validates Rudin's thesis with empirical data; applicable to high-stakes AI globally
Ribeiro et al (2016) [49] LIME	USA; technical validation	Post-hoc method	LIME algorithm for local feature importance	Developer-centric tool; not designed for end users	Expert/developer consensus: LIME inadequate for lay users universally; technical-lay user gap exists globally
Lundberg and Lee (2017) [50] SHAP	USA; technical validation	Unified interpretation method	SHAP values for game-theoretic feature importance	Technical complexity; assumes mathematical literacy	Developer testimony: "SHAP for developers, not users"; mathematical complexity barrier is universal, not cultural

XSL•FO
RenderX

Study	Context and sample	Methods	Key findings	Limitations	How this study advances globally
Liao et al (2020) [53]	USA; 20 practitioners	Qualitative interviews	Design patterns for XAI user experience	Practitioners only; no end users; high-resource assumptions	Extends with end user perspectives (n=137); universal needs (transparency, actionability) transcend geographic boundaries
Larasati et al (2023) [32]	Netherlands; 24 patients	Mixed methods	Trust influenced by explanation meaningfulness	High-resource, Western health care; patients only	Validates trust meaningfulness link across resource contexts; extends to multistakeholder (developers, clinicians, experts)
Visser et al (2023) [12]	Germany; review of 47 studies	Systematic review	Trust depends on explanation quality and user expertise	Geographic bias (45/47 Western); limited demographic analysis	Empirical demographic data; education (not age) predicts trust—likely universal pattern; tests Western assumptions in non-Western context
He et al (2022) [33]	China; user study	User-centered design	User needs vary by medical knowledge level	Single country; limited sample; no framework	Corroborates knowledge-level finding; education-based personalization universal; 5-pillar framework operationalizes insights globally
Rosenbacke et al (2024) [10]	Global review; 22 studies (20 Western)	Systematic review	XAI can increase OR decrease clinician trust depending on design	Severe Western bias; clinicians only; no user perspectives	Non-Western empirical data; validates universal finding: poor XAI decreases trust; clinician-user gap universal across contexts
Shin (2021) [9]	South Korea; 405 users	Quantitative experiment	Explainability enhances trust and perceived control	Not health care-specific; single country	Health care-specific validation; trust-explainability link holds across contexts (universal psychological mechanism)
Okolo et al (2022) [17]	Global South review; 63 papers	Systematic review	Identified severe XAI research gap in Global South	Conceptual; no primary empirical data	Fills identified gap; but findings (education predicts trust, stakeholder misalignments) are universal patterns applicable to underserved populations globally
Sambasivan et al (2021) [48]	India; 67 participants	Ethnographic study	Fairness assumptions from West fail in non-Western contexts	Data work focus; not health care XAI	Extends fairness critique to explainability; Western XAI assumptions (technical transparency sufficient, age predicts trust) fail universally

XSL•FO
RenderX

Study	Context and sample	Methods	Key findings	Limitations	How this study advances globally
This Study	Bangladesh; 157 total: Users (n=137), Developers (n=4), Clinicians (n=10), XAI Experts (n=6)	Sequential mixed methods: quantitative surveys, qualitative interviews/focus groups; statistical+ thematic analysis	Five-pillar universal framework: (1) Human-AI collaboration, (2) Transparent logic, (3) Contextual personalization, (4) Cultural-linguistic relevance, (5) Ethical safeguards. Education (not age) predicts trust (*P*=.042). Stakeholder misalignments documented. LIME/SHAP inadequate for lay users.	Single site (but generalizable patterns); English-proficient sample; 4 developers; selection bias	Universal Contributions: (1) Education predicts trust globally, applies to elderly care (Japan/Germany), health literacy programs (worldwide), digital inclusion (everywhere); (2) Stakeholder misalignments exist in ALL health care systems (US/European systems face same gaps); (3) LIME/SHAP inadequacy universal, not context-specific; (4) Framework applicable to ANY context: immigrant populations (US/Europe), Indigenous communities (Australia/Canada), rural populations (all continents), multilingual regions (Southeast Asia, Africa, South America), low-resource settings (rural US, Eastern Europe); (5) Developer constraints (limited resources, lack of tools) affect startups globally; (6) Replicable multistakeholder methodology for any context

[a]SME: small and medium-sized enterprise.

[b]LIME: local interpretable model-agnostic explanations.

[c]SHAP: Shapley additive explanations.

[d]AI: artificial intelligence.

[e]ML: machine learning.

Conclusions

This multistakeholder study identified five principal findings: (1) Education significantly predicts trust in AI explainability (*P*=.042); (2) Systematic misalignments exist between user needs, developer constraints, and clinician requirements; (3) Human oversight is universally prioritized over autonomous AI; (4) Technical XAI methods inadequately serve lay users; (5) A 5-pillar participatory framework addresses these gaps. These findings benefit mHealth developers (implementation guidance for resource-constrained contexts), health care providers (evidence for human-AI collaboration models), policymakers (principles for AI governance and digital literacy programs), XAI researchers (identification of technical-lay user gap), end users (advocacy for understandable health AI), and global health equity (framework preventing Western-centric paradigms and addressing challenges affecting billions worldwide).

Explainability in AI-driven mHealth apps is not merely a question of technical transparency but a social, cultural, and ethical necessity. This study, grounded in the South Asian context, reveals that users, developers, and medical professionals conceptualize explainability in distinct yet interconnected ways: users seek reassurance and clarity, developers emphasize feasibility and performance, and clinicians demand evidence-based accountability. The proposed 5-pillar framework responds to these diverse expectations by integrating human–AI collaboration, transparent logic, contextual personalization, cultural relevance, and ethical safeguards into a unified model for trustworthy mHealth design.

Beyond advancing technical discourse, this work highlights that explainability in health care AI is fundamentally about equity and inclusion. In resource-limited settings, the ability to understand and question AI-generated recommendations determines whether digital tools empower or marginalize users. By situating explainability within local linguistic and cultural realities, this research contributes a context-sensitive foundation

XSL•FO
RenderX

for designing AI systems that are not only intelligent but also understandable, credible, and humane.

Ultimately, the study calls for a paradigm shift from viewing explainability as a computational problem to treating it as a participatory design responsibility. Building trustworthy mHealth AI requires more than transparent models; it demands collaboration across disciplines and cultures. The findings and framework presented here aim to inform that transformation, guiding the development of AI systems that strengthen, not replace, the relationship between humans and technology in the pursuit of better health.

Acknowledgments

We thank the editor and reviewers for their constructive feedback, which significantly improved the quality of this manuscript. We also thank all participants of this study for their time and valuable insights, who supported data collection. We declare the use of generative artificial intelligence to assist with language editing and proofreading. Under full human supervision, Claude AI (Anthropic) was used for improving English language clarity and readability. All research design, data collection, statistical analysis, interpretation, and intellectual contributions are solely the work of the authors. Responsibility for the final manuscript lies entirely with the authors, and the AI tool is not listed as an author.

All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article.

Funding

The Independent University, Bangladesh (IUB) provided APF support for the publication of this work.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide for stakeholder interviews with developers, clinicians, and XAI experts.
[DOCX File, 1678 KB - jmir_v28i1e87158_app1.docx]

Multimedia Appendix 2
Survey questionnaires administered to end users.
[DOCX File, 2044 KB - jmir_v28i1e87158_app2.docx]

References

1. Koski E, Murphy J. AI in healthcare. In: Nurses and Midwives in the Digital Age: IOS Press; 2021:295-299. [doi: 10.3233/SHTI210726]
2. Deniz-Garcia A, Fabelo H, Rodriguez-Almeida AJ, et al. Quality, usability, and effectiveness of mHealth apps and the role of artificial intelligence: current scenario and challenges. J Med Internet Res 2023 May 4;25:e44030. [doi: 10.2196/44030] [Medline: 37140973]
3. Global Digital Health Monitor. The state of digital health report 2023. Global Digital Health Monitor. 2023. URL: https://digitalhealthmonitor.org/stateofdigitalhealth23 [accessed 2025-01-14]
4. Market.biz. MHealth apps statistics by usage, revenue and facts. Market.biz. 2025. URL: https://market.biz/mhealth-apps-statistics/#mhealth-app-statistics [accessed 2025-01-14]
5. Mehrotra S, Degachi C, Vereschak O, Jonker CM, Tielman ML. A systematic review on fostering appropriate trust in human-AI interaction: trends, opportunities and challenges. ACM J Responsib Comput 2024 Dec 31;1(4):1-45. [doi: 10.1145/3696449]
6. World Health Organization. World Health Organization. 2024. URL: https://www.who.int/news/item/18-01-2024-who-releases-ai-ethics-and-governance-guidance-for-large-multi-modal-models [accessed 2025-01-14]
7. Rai A. Explainable AI: from black box to glass box. J Acad Mark Sci 2020 Jan;48(1):137-141. [doi: 10.1007/s11747-019-00710-5]
8. Ehsan U, Liao QV, Muller M, Riedl MO, Weisz JD. Expanding explainability: towards social transparency in AI systems. Presented at: CHI Conference on Human Factors in Computing Systems; May 8-13, 2021; Yokohama Japan p. 82. [doi: 10.1145/3411764.3445188]

XSL•FO
RenderX

9. Shin D. The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI. Int J Hum Comput Stud 2021 Feb;146:102551. [doi: 10.1016/j.ijhcs.2020.102551]
10. Rosenbacke R, Melhus Å, McKee M, Stuckler D. How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: systematic review. JMIR AI 2024 Oct 30;3:e53207. [doi: 10.2196/53207] [Medline: 39476365]
11. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv. Preprint posted online on 2017. [doi: 10.48550/arXiv.1702.08608]
12. Visser R, Peters TM, Scharlau I, Hammer B. Trust, distrust, and appropriate reliance in (x) AI: a survey of empirical evaluation of user trust. arXiv. Preprint posted online on 2023. [doi: 10.48550/arXiv.2312.02034]
13. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019 May;1(5):206-215. [doi: 10.1038/s42256-019-0048-x] [Medline: 35603010]
14. Dhanorkar S, Wolf CT, Qian K, Xu A, Popa L, Li Y. Who needs to know what, when?: broadening the explainable AI (XAI) design space by looking at explanations across the AI lifecycle. Presented at: DIS '21: Proceedings of the 2021 ACM Designing Interactive Systems Conference; Jun 28 to Jul 2, 2021; Virtual Event USA p. 1591-1602. [doi: 10.1145/3461778.3462131]
15. Lakkaraju H, Slack D, Chen Y, Tan C, Singh S. Rethinking explainability as a dialogue: a practitioner's perspective. arXiv. Preprint posted online on Feb 3, 2022. [doi: 10.48550/arXiv.2202.01875]
16. Longo L, Brcic M, Cabitza F, et al. Explainable artificial intelligence (XAI) 2.0: a manifesto of open challenges and interdisciplinary research directions. Information Fusion 2024 Jun;106:102301. [doi: 10.1016/j.inffus.2024.102301]
17. Okolo CT, Dell N, Vashistha A. Making AI explainable in the global south: a systematic review. Presented at: COMPASS '22: Proceedings of the 5th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies; Jun 29 to Jul 1, 2022; Seattle WA USA p. 439-452. [doi: 10.1145/3530190.3534802]
18. Abebe R, Aruleba K, Birhane A, et al. Narratives and counternarratives on data sharing in africa. Presented at: FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; Mar 3-10, 2021; Virtual Event Canada p. 329-341. [doi: 10.1145/3442188.3445897]
19. Okolo CT. AI in the global south: opportunities and challenges towards more inclusive governance. : The Brookings Institution; 2023. [doi: 10.48028/iiprds/ijiraet.v2.i1.09]
20. Stepnov IM, Kasatkin PI, Kolesnikova TV. Digital divide: understanding the disparities between global north and global south. In: Polycrisis and Economic Development in the Global South: Routledge; 2024:171-183. [doi: 10.4324/9781032694337-17]
21. Kitsara I. Artificial intelligence and the digital divide: from an innovation perspective. In: Platforms and Artificial Intelligence: Springer; 2022:245-265. [doi: 10.1007/978-3-030-90192-9_12]
22. Krishna K. Artificial intelligence & the global south: bridging or exacerbating the digital divide? SSRN. Preprint posted online on Apr 20, 2024. [doi: 10.2139/ssrn.4866372]
23. Bentley SV, Naughtin CK, McGrath MJ, Irons JL, Cooper PS. The digital divide in action: how experiences of digital technology shape future relationships with artificial intelligence. AI Ethics 2024 Nov;4(4):901-915. [doi: 10.1007/s43681-024-00452-3]
24. Berg J, Snene M, Velasco L. Mind the AI Divide: Shaping a Global Perspective on the Future of Work: International Labour Organization; 2024.
25. Gonzales S. AI literacy and the new digital divide — a global call for action. UNESCO. 2024 Aug 6. URL: https://www.unesco.org/en/articles/ai-literacy-and-new-digital-divide-global-call-action [accessed 2025-01-14]
26. GSMA. Bangladesh: advancing digital inclusion. GSMA. 2021. URL: https://www.gsma.com/about-us/regions/asia-pacific/gsma_resources/bangladesh-digital-inclusion/ [accessed 2025-01-14]
27. Ahmed T, Rizvi SJR, Rasheed S, et al. Digital health and inequalities in access to health services in Bangladesh: mixed methods study. JMIR Mhealth Uhealth 2020 Jul 21;8(7):e16473. [doi: 10.2196/16473] [Medline: 32706736]
28. Sikder AS. Artificial intelligence-enabled transformation in Bangladesh: overcoming challenges for socio-economic empowerment. IJIST 2023;1(1):77-96. [doi: 10.70774/ijist.v1i1.7]
29. Ahmed SM, Hossain MA, Rajachowdhury AM, Bhuiya AU. The health workforce crisis in Bangladesh: shortage, inappropriate skill-mix and inequitable distribution. Hum Resour Health 2011 Jan 22;9(1):3. [doi: 10.1186/1478-4491-9-3] [Medline: 21255446]
30. Das Pooja S, Nandonik AJ, Ahmed T, Kabir ZN. "Working in the Dark": experiences of frontline health workers in Bangladesh during COVID-19 pandemic. J Multidiscip Healthc 2022;15:869-881. [doi: 10.2147/JMDH.S357815] [Medline: 35496717]
31. Nuruzzaman M, Zapata T, McIsaac M, et al. Informing investment in health workforce in Bangladesh: a health labour market analysis. Hum Resour Health 2022 Oct 12;20(1):73. [doi: 10.1186/s12960-022-00769-2] [Medline: 36224554]
32. Larasati R, De Liddo A, Motta E. Meaningful explanation effect on user's trust in an AI medical system: designing explanations for non-expert users. ACM Trans Interact Intell Syst 2023 Dec 31;13(4):1-39. [doi: 10.1145/3631614]
33. He X, Hong Y, Zheng X, Zhang Y. What are the users' needs? design of a user-centered explainable artificial intelligence diagnostic system. Int J Hum Comput Interact 2023 Apr 21;39(7):1519-1542. [doi: 10.1080/10447318.2022.2095093]

XSL•FO
RenderX

34. Gerdes A. The role of explainability in AI-supported medical decision-making. Discov Artif Intell 2024;4(1):29. [doi: 10.1007/s44163-024-00119-2]
35. Tomsett R, Braines D, Harborne D, Preece A, Chakraborty S. Interpretable to whom? a role-based model for analyzing interpretable machine learning systems. arXiv. Preprint posted online on Jun 20, 2018. [doi: 10.48550/arXiv.1806.07552]
36. van Kolfschooten H, van Oirschot J. The EU artificial intelligence act (2024): implications for healthcare. Health Policy 2024 Nov;149:105152. [doi: 10.1016/j.healthpol.2024.105152] [Medline: 39244818]
37. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 2020 Jun;58:82-115. [doi: 10.1016/j.inffus.2019.12.012]
38. Islam F, Mayeesha TT, Ahmed N. Know your users: towards explainable AI in Bangladesh. Presented at: UbiComp '24: Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing; Oct 5-9, 2024; Melbourne VIC Australia p. 890-893. [doi: 10.1145/3675094.3679002]
39. Ahmed T, Bloom G, Iqbal M, et al. E-health and m-health in bangladesh: opportunities and challenges. : Evidence Report; 2014 60.
40. Heeks R. Information and Communication Technology for Development (ICT4D): Routledge; 2017. [doi: 10.4324/9781315652603]
41. World bank open data. World Bank. 2023. URL: https://data.worldbank.org/indicator/SP.RUR.TOTL.ZS [accessed 2025-01-14]
42. Global health workforce statistics database. World Health Organization. 2023. URL: https://www.who.int/data/gho/data/themes/topics/health-workforce [accessed 2025-01-14]
43. The state of mobile internet connectivity report 2023. GSMA. 2023. URL: https://www.gsma.com/r/somic-2023/ [accessed 2025-01-14]
44. Algorithmic Colonization of Africa. SCRIPT-ed 2020;17(2):389. [doi: 10.2966/scrip.170220.389]
45. Beede E, Baylor E, Hersch F, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. Presented at: CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; Apr 25-30, 2020; Honolulu HI USA p. 1-12. [doi: 10.1145/3313831.3376718]
46. Ali A, Raza AA, Qazi IA. Validated digital literacy measures for populations with low levels of internet experiences. Dev Eng 2023 Nov;8:100107. [doi: 10.1016/j.deveng.2023.100107]
47. Okolo CT. AI explainability in the global south: towards an inclusive praxis for emerging technology users. Ithaca, NY, USA: Cornell University; 2023. [doi: 10.7298/c1dw-j313]
48. Sambasivan N, Arnesen E, Hutchinson B, Doshi T, Prabhakaran V. Re-imagining algorithmic fairness in india and beyond. Presented at: FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; Mar 1-10, 2021; Virtual Event Canada p. 315-328. [doi: 10.1145/3442188.3445896]
49. Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?" explaining the predictions of any classifier. Presented at: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 URL: https://dl.acm.org/doi/proceedings/10.1145/2939672 [accessed 2026-04-30] [doi: 10.1145/2939672.2939778]
50. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017; Long Beach, CA, USA p. 4768-4777.
51. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv. Preprint posted online on Dec 20, 2013. [doi: 10.48550/arXiv.1312.6034]
52. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI-explainable artificial intelligence. Sci Robot 2019 Dec 18;4(37):eaay7120. [doi: 10.1126/scirobotics.aay7120] [Medline: 33137719]
53. Liao QV, Gruen D, Miller S. Questioning the AI: informing design practices for explainable AI user experiences. Presented at: CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; Apr 25-30, 2020; Honolulu HI USA p. 1-15. [doi: 10.1145/3313831.3376590]
54. Singh TV, Dogra N, Saxena A, Katare DP, Mani RJ. Critical analysis of current healthcare applications for diagnosis of diseases: pitfalls and future. In: Handbook of AI-Based Models in Healthcare and Medicine: CRC Press; 2024:303-329. [doi: 10.1201/9781003363361-16]
55. Kizilcec RF. How much information?: effects of transparency on trust in an algorithmic interface. Presented at: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems; May 7-12, 2016; San Jose, CA, USA2390 p. 2390-2395. [doi: 10.1145/2858036.2858402]
56. Ismail A, Kumar N. Empowerment on the margins: the online experiences of community health workers. Presented at: CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; May 4-9, 2019; Glasgow, Scotland, UK p. 1-15. [doi: 10.1145/3290605.3300329]
57. Joshi A, Kale S, Chandel S, Pal DK. Likert scale: explored and explained. BJAST 2015;7(4):396-403. [doi: 10.9734/BJAST/2015/14975]
58. Geer JG. What do open-ended questions measure? Public Opin Q 1988 ;52(3):365. [doi: 10.1086/269113]
59. Feeling sick? there's an app for that! – the big symptom checker review. The Medical Futurist. 2024. URL: https://medicalfuturist.com/the-big-symptom-checker-review/ [accessed 2025-01-14]

XSL•FO
RenderX

60. Golzar J, Noor S, Tajik O. Convenience sampling. Int J Educ Lang Stud 2022;1(2):72-77. [doi: 10.22034/ijels.2022.162981]
61. Tremblay MA. The key informant technique: a nonethnographic application. Am Anthropol 1957 Aug;59(4):688-701. [doi: 10.1525/aa.1957.59.4.02a00100]
62. Clarke V, Braun V. Thematic analysis. J Posit Psychol 2017 May 4;12(3):297-298. [doi: 10.1080/17439760.2016.1262613]
63. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
64. MacQueen KM, McLellan E, Kay K, Milstein B. Codebook development for team-based qualitative analysis. CAM J 1998 May;10(2):31-36. [doi: 10.1177/1525822X980100020301]
65. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: a systematic review of empirical tests. Soc Sci Med 2022 Jan;292:114523. [doi: 10.1016/j.socscimed.2021.114523] [Medline: 34785096]
66. Bryant A, Charmaz K. Coding for grounded theory. In: The SAGE Handbook of Current Developments in Grounded Theory: SAGE; 2019:167-185. [doi: 10.4135/9781526485656.n10]
67. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
68. Hartley RI, Sturm P. Triangulation. Comput Vis Image Underst 1997 Nov;68(2):146-157. [doi: 10.1006/cviu.1997.0547]
69. Budget FY26: health sector gets Tk41,908cr. The Business Standard. 2025 Jun 2. URL: https://www.tbsnews.net/economy/budget/fy26-budget-health-sector-gets-tk41908cr-1158031 [accessed 2026-02-14]
70. Bangladesh digital health strategy 2023-2027. Dhaka: Directorate General of Health Services, Ministry of Health and Family Welfare.: Government of Bangladesh; 2023. URL: https://objectstorage.ap-dcc-gazipur-1.oraclecloud15.com/n/axvjbnqprylg/b/V2Ministry/o/office-dghs/2024/12/5c9bc203e34d4d0b9e577de0c8315dd6.pdf [accessed 2026-02-14]
71. Ahmed NU, Rahman MH, Roy SS, Chowdhury SA, Chowdhury R. Telemedicine services of 'Shasthyo Batayon 16263' during COVID-19 pandemic: opportunities and challenges. Bangladesh Med Res Counc Bull 2020;46(3):240-244. [doi: 10.3329/bmrcb.v46i3.52259]
72. Advancing digital health strategy in Bangladesh. WHO Bangladesh.: World Health Organization; 2021 Jan 3. URL: https://www.who.int/bangladesh/news/detail/03-01-2021-advancing-digital-health-strategy-in-bangladesh [accessed 2026-02-14]
73. Fink S, Haner M. The world needs to help Bangladesh keep up its global health progress. NPR Goats and Soda. 2024 Sep 27. URL: https://www.npr.org/2024/09/27/g-s1-24547/bangladesh-global-health [accessed 2025-01-14]
74. Uddin Quadery SE, Hasan M, Khan MM. Consumer side economic perception of telemedicine during COVID-19 era: a survey on Bangladesh's perspective. Inform Med Unlocked 2021;27:100797. [doi: 10.1016/j.imu.2021.100797] [Medline: 34869827]
75. Ada health overview. Wikipedia. URL: https://en.wikipedia.org/wiki/Ada_Health [accessed 2025-01-14]
76. WebMD symptom checker. WebMD. URL: https://symptoms.webmd.com [accessed 2025-01-14]
77. Symptomate: symptom checker. Symptomate. URL: https://symptomate.com [accessed 2025-01-14]
78. Better health starts with Ada. Ada Health. URL: https://about.ada.com/ [accessed 2025-01-14]
79. Symptomate on Apple app store. Symptomate. URL: https://apps.apple.com/us/app/symptomate-symptom-checker/id837725433 [accessed 2025-01-14]
80. WebMD symptom checker review. WebMD. URL: https://www.prnewswire.com/news-releases/webmd-launches-redesigned-state-of-the-art-symptom-checker-300624752.html [accessed 2025-01-14]
81. Hern A. Eye doctors find that WebMD symptom checker was wrong more than half the time. Ars Technica. 2018. URL: https://arstechnica.com/science/2018/10/eye-doctors-find-that-webmd-symptom-checker-was-wrong-more-than-half-the-time/ [accessed 2025-08-24]
82. Accuracy of the WebMD symptom checker. American Academy of Ophthalmology.: EyeNet Magazine URL: https://www.aao.org/eyenet/article/accuracy-of-the-webmd-symptom-checker [accessed 2025-01-14]
83. Kim M, Kim S, Kim J, Song TJ, Kim Y. Do stakeholder needs differ? - designing stakeholder-tailored explainable artificial intelligence (XAI) interfaces. Int J Hum Comput Stud 2024 Jan;181:103160. [doi: 10.1016/j.ijhcs.2023.103160]
84. Alduhailan HW, Alshamari MA, Wahsheh HAM. A comprehensive comparison and evaluation of AI-powered healthcare mobile applications' usability. Healthcare (Basel) 2025 Jul 26;13(15):1829. [doi: 10.3390/healthcare13151829] [Medline: 40805861]
85. Noor AA, Manzoor A, Mazhar Qureshi MD, Qureshi MA, Rashwan W. Unveiling explainable AI in healthcare: current trends, challenges, and future directions. WIREs Data Min Knowl 2025 Jun;15(2):e70018 [FREE Full text] [doi: 10.1002/widm.70018]
86. Morelli S, Giansanti D. Recent advances in AI-driven mobile health enhancing healthcare-narrative insights into latest progress. Bioengineering (Basel) 2025 Dec 31;13(1):54. [doi: 10.3390/bioengineering13010054] [Medline: 41595987]

Abbreviations

AD: application developer
AI: artificial intelligence
KII: Key Informant Interview

XSL•FO
RenderX

LIME: local interpretable model-agnostic explanations
mHealth: mobile health
ML: machine learning
NGO: nongovernmental organization
SHAP: Shapley additive explanations
WHO: World Health Organization
XAI: explainable AI

Edited by A Coristine; submitted 12.Nov.2025; peer-reviewed by M Rabbani, T Alam; revised version received 09.Mar.2026; accepted 18.Mar.2026; published 04.May.2026.

Please cite as:
Islam F, Islam A, Amin MA, Zaber M
A Proposed Participatory Framework for Explainable AI in mHealth: Mixed Methods Study Integrating User and Stakeholder Requirements
J Med Internet Res 2026;28:e87158
URL: https://www.jmir.org/2026/1/e87158
doi:10.2196/87158

© Farzana Islam, Ashraful Islam, M Ashraful Amin, Moinul Zaber. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 4.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Smartphone App–Based Music-Facilitated Pulmonary Rehabilitation Program Integrating Rhythm-Guided Walking and Singing for Patients With Chronic Obstructive Pulmonary Disease: Multicenter Randomized Controlled Trial

Minghui Shi[1,2,3,4*], MD; Shiwei Qumu[2,3,4*], MD; Xiaoning Bu[5], MD; Jie Zhang[6], MD; Kai Liu[7], MD; Jinxiang Wang[8], MD; Yinan Zhang[9], MD; Jieping Lei[10], MD; Siyuan Wang[11], MD; Wei Li[2,3,4], MD; Xingyao Tang[1,2,3,4], MD; Jisong Yan[2,3,4], MD; Yaodie Peng[2,3,4,12], MD; Chen Wang[2,3,4], MD; Ting Yang[1,2,3,4*], MD; Ke Huang[2,3,4*], MD

[1]China-Japan Friendship Clinical Medical College, Capital Medical University, Beijing, China

[10]Department of Clinical Research and Data Management, Center of Respiratory Medicine, China-Japan Friendship Hospital, Beijing, China

[11]Department of Rehabilitation Medicine, China-Japan Friendship Hospital, Beijing, China

[12]China-Japan Friendship Clinical Medical College, Peking University, Beijing, China

[2]National Center for Respiratory Medicine, State Key Laboratory of Respiratory Health and Multimorbidity, National Clinical Research Center for Respiratory Diseases, China-Japan Friendship Hospital, Beijing, China

[3]Institute of Respiratory Medicine, Chinese Academy of Medical Sciences and Peking Union Medical College, China-Japan Friendship Hospital, Beijing, China

[4]Department of Pulmonary and Critical Care Medicine, Center of Respiratory Medicine, China-Japan Friendship Hospital, 2 Yinghua East Street, Chaoyang District, Beijing, China

[5]Department of Pulmonary and Critical Care Medicine, Beijing Tiantan Hospital, Beijing, China

[6]Department of Pulmonary and Critical Care Medicine, The Second Affiliated Hospital of Xi’an Jiaotong University, Xi' an, China

[7]Department of Rehabilitation Medicine, Qingdao Municipal Hospital, Qing Dao, China

[8]Department of Pulmonary and Critical Care Medicine, Beijing Luhe Hospital, Beijing, China

[9]Psychiatric Rehabilitation Center, Wuxi Mental Health Center, The Affiliated Mental Health Center of Jiangnan University, Wuxi, China

[*]these authors contributed equally

Corresponding Author:
Ke Huang, MD
National Center for Respiratory Medicine, State Key Laboratory of Respiratory Health and Multimorbidity, National Clinical Research Center for Respiratory Diseases, China-Japan Friendship Hospital, Beijing, China

Abstract

Background: Pulmonary rehabilitation (PR) is a cornerstone for the management of chronic obstructive pulmonary disease (COPD), yet global uptake remains low due to geographic and resource barriers. Digital health technologies, specifically smartphone apps, offer a promising platform for delivering accessible home-based PR. In addition, music-assisted interventions not only offer unique physiological and psychological benefits but may also serve as an innovative approach to enhancing patient engagement and improving the effectiveness of rehabilitation in home settings.

Objective: This study aimed to evaluate the effectiveness of a smartphone app–based, music-facilitated multicomponent PR program (integrating rhythm-guided walking [RW] and singing) for improving exercise capacity and other clinical outcomes in patients with COPD compared with usual care (UC).

Methods: This 3-arm, parallel-group, multicenter randomized controlled trial included 70 participants in China. Participants were randomized into a multimodule training (MT) group, which included a multicomponent PR program integrating RW and singing training (n=25); a RW group, which included RW training (n=23); or a UC group (n=22). The MT and RW groups received 12-week asynchronous home-based training via a smartphone app, and all arms received structured patient education. The primary outcome was the distance achieved in the incremental shuttle walking test (ISWT) at 12 weeks. The secondary outcomes included dyspnea, quality of life, and pulmonary function.

Results: The modified intention-to-treat principle was used to analyze the 70 study patients. At 12 weeks, the ISWT distance was significantly greater in the MT group than in the UC group (mean difference [MD] 56.35 m, 95% CI 6.66-106.04 m; *P*=.03; Cohen *d*=0.30). Significant improvements were observed in the MT group compared with the UC group in the modified Medical Research Council dyspnea scale (mMRC) score (MD −0.44, 95% CI −0.80 to −0.08; *P*=.02), COPD Assessment Test score (MD −3.23, 95% CI −6.18 to −0.29; *P*=.03), Hospital Anxiety and Depression Scale-anxiety subscale score (MD −2.31, 95% CI −3.99

XSL•FO
RenderX

to −0.63; *P*=.008), and inspiratory capacity (MD 15.98% predicted, 95% CI 4.76 to 27.21; *P*=.01). However, no significant differences were found between the RW and UC groups in primary or secondary outcomes. Compared with RW, MT was significantly better at decreasing the mMRC score (*P*=.03).

Conclusions: The findings of this study demonstrate that our smartphone app–based music-facilitated multicomponent PR program (including tempo-guided walking and singing) caused clinically meaningful improvements in exercise capacity among patients with COPD compared to UC. Moreover, secondary outcomes, including dyspnea, quality of life, psychological status, and inspiratory capacity, showed better improvements with MT than with UC.

Trial Registration: ClinicalTrials.gov NCT05832814; https://clinicaltrials.gov/study/NCT05832814

(J Med Internet Res 2026;28:e81707) doi:10.2196/81707

KEYWORDS

chronic obstructive pulmonary disease; pulmonary rehabilitation; digital health; smartphone app; music

Introduction

Background

Chronic obstructive pulmonary disease (COPD) represents a major global public health challenge, which is characterized by persistent respiratory symptoms, progressive airflow limitation, and systemic manifestations, resulting in reduced physical activity, deteriorating quality of life, and psychological symptoms [1,2]. Despite advances in pharmacotherapy, many patients continue to experience substantial symptom burden and functional limitation [3]. This highlights the need for comprehensive nonpharmacological interventions, with pulmonary rehabilitation (PR) established as a cornerstone, supported by robust evidence of improved clinical outcomes [4,5]. However, traditional center-based PR remains critically underutilized globally, with only 2% - 4% of eligible patients with COPD accessing programs, and completion rates frequently below 50% [6-8]. Access is limited by multiple barriers, including a shortage of therapists, long travel distances, high treatment costs, exercise-induced dyspnea, and the monotony of conventional training [8,9].

To address these implementation gaps, digital health technologies—such as mobile health (mHealth) rehabilitation, video-based telerehabilitation, and web-based platform rehabilitation—have emerged as promising alternatives or supplements to traditional PR [10-13]. Among these, mHealth rehabilitation offers greater flexibility and reduces human resource demands compared with other modalities while simultaneously enhancing accessibility, patient engagement, and long-term adherence [14,15]. Recent reviews and randomized controlled trials (RCTs) have indicated that smartphone app–based PR programs can sustainably improve symptom scores, health-related quality of life (HRQoL), and daily physical activity levels [14-18]. However, evidence regarding objective exercise capacity remains inconsistent [14-19]. For instance, one study found that a 12-week app-based PR program significantly improved 1-minute sit-to-stand test results in patients with COPD [14], whereas another study reported no significant effect of a 12-week app-guided PR program on 6-minute walk distance in patients with chronic respiratory disease [19]. Additionally, simply digitizing the delivery of PR leaves certain patient-related barriers unaddressed. Conventional training, which primarily focuses on physical exercise, often induces fear and resistance to rehabilitation among patients with COPD owing to exertional dyspnea and fatigue. Moreover, patients often struggle to adhere because the training is perceived as monotonous and physically demanding [9].

Recent studies have identified the incorporation of music into rehabilitation training as a promising approach to address this challenge [20-25]. Receptive music can serve as a distractive auditory stimulus during aerobic exercise, facilitating auditory-motor coupling and neural entrainment to reduce perceived dyspnea and improve endurance [26,27]. Analogous to respiratory training in PR, singing or playing wind instruments promotes controlled expiratory flow, reduces dynamic hyperinflation, and strengthens respiratory muscles such as the diaphragm [28,29]. Furthermore, the inherent appeal of music itself can benefit adherence [24,25]. Nevertheless, previous research has primarily been conducted in outpatient or community care settings, with limited studies in a home-based setting [20,23,30,31]. Most studies have focused on the effects of individual music interventions, which may not adequately address the complex pathophysiological changes in COPD, including weakness of the skeletal muscles, pulmonary hyperinflation, anxiety, and depression [20,23,30-33]. The best practices for smartphone apps supporting a composite music-facilitated PR program are not yet well established.

We conducted a multicenter, 3-arm, parallel, pilot RCT to explore the preliminary efficacy and feasibility of a smartphone app–based music-facilitated multicomponent PR program with regard to clinical health outcomes in patients with COPD (COPDMELODY [Music Therapy for COPD Rehabilitation]). Prior to this study, a smartphone app (Qiyue) was developed by our research team to deliver the online music-facilitated PR program. The program integrates core PR components, including tempo-guided walking (aerobic exercise), structured singing training (respiratory training), and systematic education. Appropriate audio accompaniments and songs were carefully preselected by qualified music therapists and preinstalled in the corresponding training modules of the app. Additionally, to ensure usability and technical feasibility, the app underwent preliminary testing and feedback collection at the lead study center among a small group of patients with COPD who had prior experience with outpatient PR. The scientific rationale and detailed methodology of this music-facilitated smartphone app–based rehabilitation approach have been previously outlined in our study protocol [34].

XSL•FO
RenderX

Objective

The objective of our study was to investigate the effectiveness of a smartphone app–based music-facilitated multicomponent PR program (incorporating rhythm-guided walking [RW] and singing) with regard to exercise capacity (primary outcome), and dyspnea, HRQoL, psychological status, and pulmonary function (secondary outcomes) in patients with COPD, compared with usual care (UC). Additionally, as an exploratory aim, we assessed the potential added benefits of combining singing training with RW versus RW alone.

Methods

Study Design

COPDMELODY was a multicenter, pragmatic, 3-arm, parallel-group, single-blind pilot RCT. The study protocol was registered at ClinicalTrials.gov (NCT05832814). Participants with moderate to severe stable COPD were randomized into 1 of the following 3 groups: multimodule training (MT), RW, and UC. Online training sessions were delivered over 12 weeks after randomization. Outcomes were assessed at baseline, 4 weeks, 8 weeks, and 12 weeks at hospital sites. The clinical trial protocol is presented in Multimedia Appendix 1. The CONSORT (Consolidated Standards of Reporting Trials) checklist is provided in Checklist 1.

Participants

Participants were recruited from 5 hospitals across 3 administrative regions in China (Figure S3 and Table S11 in Multimedia Appendix 2). Participants were considered eligible if they met the following criteria: age 40-75 years, diagnosis of COPD, postbronchodilator percentage of predicted forced expiratory volume in 1 second (FEV_1) lower than 80%, and FEV_1 to forced vital capacity ratio lower than 0.7. Participants were required to have stable symptoms for at least 2 weeks and to be able to independently use a smartphone, as judged by their ability to perform mobile payments and use the WeChat app (Tencent). Participants were excluded if they had unstable angina or acute myocardial infarction within the past 4 weeks or had other comorbidities that would preclude participation in exercise training.

Recruitment advertisements were placed in the respiratory outpatient departments of each participating hospital, allowing interested patients to contact the research team directly. Additionally, study information leaflets were distributed to hospital medical staff, who could then refer potential candidates to the study team. Potential eligible patients who expressed interest in participation were invited to an introductory meeting—either face-to-face or via telephone—where the research team explained the study details and addressed any questions. Participation in this study was completely voluntary. Detailed descriptions of the recruitment strategies, screening process, and inclusion/exclusion criteria are provided in the study protocol [34].

Randomization, Allocation Concealment, and Masking

Once written consent was received, a trial register form was completed, and the participant was assigned an ID number. Participants were randomly assigned to 1 of 3 groups in a 1:1:1 ratio, using a random number table generated in the SPSS statistical package (IBM Corp) by an independent statistical analyst. Randomization was conducted with permuted blocks, stratified by site. Group allocation information was stored in opaque envelopes and remained concealed until randomization occurred following the baseline assessment. Assessors and statisticians were blinded to group allocation throughout the trial; however, participants could not be blinded owing to the nature of the interventions.

Interventions

Participants in the MT and RW groups underwent a 12-week multicomponent PR program on an individual basis at their homes through the Qiyue smartphone app. This app was specially developed for this study to deliver the PR program, including a RW module, a singing training module, and an educational module. Each module incorporated carefully preselected songs chosen by qualified music therapists to guide the corresponding training (Table S1 in Multimedia Appendix 2). The PR program for the MT group included three 30-minute RW sessions (aerobic exercise) and three 25-minute singing sessions (respiratory training) per week (total of 6 sessions per week), in addition to a standardized educational program. The PR program for the RW group only included three 30-minute RW sessions per week, in addition to the standardized educational program. The intervention was primarily asynchronous and delivered via prerecorded modules in the app, allowing patients to train at their convenience. However, the study team held a video conference with each patient during the first training session to ensure proper technique and protocol fidelity. Participants in the MT and RW groups were provided with a smartphone preinstalled with the Qiyue app, earphones, and a sports wristwatch. All devices were retrieved at the end of the 12-week intervention period.

Participants in the UC group were not provided with a smartphone (having no access to the app), but they received the same educational program as the intervention groups through WeChat messages from the research team during the study period. After the 12-week assessment, patients in the UC group were also offered an individualized PR program (a waiting list control group). Participants continued their prescribed medications (eg, inhaled drugs) throughout the study. The study team did not intervene in any physical activities the participants engaged in autonomously as part of their daily routines.

RW Training

During RW training, patients performed RW while synchronizing their steps to music tracks with fixed tempos delivered by the Qiyue app. The intensity of walking was individualized and recalculated at baseline, week 4, and week 8, using the incremental shuttle walking test (ISWT). The calculation of the target music tempo followed a four-step process (Table 1):

1. The target walking speed was set at 75% of the maximum speed achieved during the ISWT, as recommended by the American Thoracic Society/European Respiratory Society (ATS/ERS) guidelines [4,5,35-38].

XSL•FO
RenderX

2. Individual step length was determined from the ISWT data. During the test, patients wore a camcorder to record the number of steps per shuttle. Step length was calculated by dividing the distance per shuttle (ie, 10 m) by the steps per shuttle.
3. The target step frequency was then calculated based on the target walking speed and the individual step length.
4. The music tempo was set equal to the target step frequency to guide walking.

Table . Equations for calculating individualized rhythm-guided walking tempo.

Step	Equation
1	Target walking speed ($km \cdot h^{-1}$) = 0.75 × peak speed in the incremental shuttle walking test ($km \cdot h^{-1}$)
2	Step length ($m \cdot step^{-1}$) = The distance per shuttle (m) ÷ the steps per shuttle
3	Target step frequency ($steps \cdot min^{-1}$) = Target walking speed ($km \cdot h^{-1}$) × 1000 ÷ step length ($m \cdot step^{-1}$) ÷ 60
4	Music tempo ($beats \cdot min^{-1}$) = Target step frequency ($steps \cdot min^{-1}$)

After completing the ISWT assessment, the Qiyue app automatically calculated the target tempo for each patient using the ISWT parameters and generated a personalized walking prescription (eg, three 30-minute sessions per week at a target tempo of 105 beats/min for 4 weeks). Upon returning home, participants logged into their personal account in the app to access their individualized weekly training schedule. When they opened the walking module, a series of music tracks with fixed tempos was played, and participants were instructed to walk in real time while synchronizing each step with the musical beat. Participants also wore a sports wristwatch connected to the app, which provided real-time feedback on steps, distance covered, and heart rate (Figure S1 in Multimedia Appendix 2).

Singing Training

The singing training program consisted of 3 parts: a 5-minute video-guided breathing exercise, a 10-minute audio-guided vocal warm-up, and a 15-minute singing session. During the singing session, audio tracks of various songs were played, allowing patients to sing along in real time (Figure S2 in Multimedia Appendix 2). This structure was designed based on the latest consensus statement for Singing for Lung Health [22]. Patients' vocals were picked up by the earphones and reported to the app when they sang.

Completed sessions, including both walking training (steps, duration, distance covered, and frequency) and singing training (vocals, duration, and frequency), were automatically recorded by the app, and aggregated data were provided to the study team to monitor and assess compliance with the intervention.

Educational Program

The educational program covered key topics, including medication adherence, smoking cessation, maintaining physical activity, and vaccination. Participants in the MT and RW groups received 1 themed push notification per week through the Qiyue app, while those in the UC group received the same content via weekly WeChat messages sent by the study team.

Primary Outcome

The primary outcome was the distance achieved in the ISWT, a widely recognized and frequently used measure of exercise capacity in patients with COPD. The ISWT was performed at baseline, 4 weeks, 8 weeks, and 12 weeks, according to the ATS/ERS technical standard. At each assessment, 2 tests were conducted, and the best distance was recorded to minimize the learning effect [35,39]. The primary endpoint was ISWT distance at 12 weeks after randomization. The minimal clinically important difference (MCID) for the ISWT is 35 meters [40].

Secondary Outcomes

Several secondary outcomes were assessed. First, dyspnea symptoms were assessed using the modified Medical Research Council dyspnea scale (mMRC) at baseline and 4, 8, and 12 weeks (MCID=0.5) [41]. Second, the COPD Assessment Test (CAT) score, which indicates how symptoms impact daily life, was measured at baseline and 4, 8, and 12 weeks (MCID=1.6) [42]. Third, HRQoL was measured using the St. George's Respiratory Questionnaire (SGRQ) and EQ-5D-5L [43]. The MCID for the SGRQ is 4 points [44]. Fourth, anxiety and depression were evaluated with the Hospital Anxiety and Depression Scale (HADS), which comprises anxiety (HADS-A) and depression (HADS-D) subscales (MCID=1.3 for the HADS-A and MCID=1.5 for the HADS-D) [45]. Fifth, lung function measurements were obtained at baseline and 12 weeks [46]. Sixth, respiratory muscle function was assessed by measuring the maximal inspiratory pressure and maximal expiratory pressure using the Gio Digital Pressure Gauge (GaleMed) [47]. Seventh, participant adherence (MT and RW groups only) was measured as the percentage of sessions completed (derived using app data) versus sessions planned [10]. Patients who completed ≥75% of the planned sessions for ≥75% of the weeks were defined as adherent users. Eighth, patient satisfaction with the app was assessed at the 12-week follow-up using a single-item global satisfaction question ("Overall, how satisfied are you with the app?") rated on 5 levels ranging from "very dissatisfied" to "very satisfied." Finally, adverse events, including exacerbations, exercise-related injuries, and falls, were recorded during the study period. Reliability, validity, and responsiveness of the measures have been presented in the study protocol [34].

XSL•FO
RenderX

Sample Size

The details of the sample size calculation and statistical analysis are presented in Multimedia Appendix 2 [34]. Given the absence of previous studies exploring the effect of smartphone app–based music-facilitated multicomponent PR programs (integrating music-guided aerobic exercise and singing training), the sample size for this pilot study was calculated based on relevant studies applying similar rehabilitation interventions [48,49]. The sample size was based on the change in our primary outcome (ie, ISWT distance) from baseline to week 12. As suggested by similar studies [48,49], we tested for mean changes of 40 m, 27 m, and −15 m with combined SDs of 35 m, 38 m, and 40 m in the MT, RW, and UC groups, respectively. A sample size of 9 patients per group would yield a power of 90% at a 2-sided significance level of 5% with ANOVA. Previous RCTs evaluating outpatient training interventions reported relatively high dropout rates, ranging from 45% to 63% [12,50,51]. Therefore, assuming a conservative dropout rate of 50%, enrollment of up to 54 participants was expected (18 participants per study group). For practical purposes, considering multicenter recruitment feasibility and resource constraints, we initially planned to enroll up to 75 participants. However, due to substantial recruitment difficulties at several participating centers and limitations in time and funding (exacerbated by the COVID-19 pandemic), we were unable to enroll the originally planned number of patients in our protocol [34]. Enrollment was terminated after randomizing the 70th participant.

Statistical Analysis

The primary analysis followed the modified intention-to-treat (mITT) principle, such that comparisons were according to the randomized group, irrespective of compliance. Categorical variables have been presented as number (percentage), and continuous variables have been presented as mean (SD). Between-group comparisons of baseline characteristics were performed using 1-way ANOVA for continuous variables and the chi-square test or Fisher exact test for categorical variables, as appropriate.

The primary outcome was evaluated by a linear mixed model (LMM) incorporating all time points, and a compound-symmetry covariance structure for the residuals was used. A single model was used for the analysis, and the primary endpoint was reported by pairwise comparisons of estimated marginal means of the 3 groups at week 12. The model included treatment group, time point of measurement, treatment by time point interaction, age, gender, BMI, and ISWT distance at baseline as fixed effects. The model included the recruitment site as a random effect. To detect differences in the proportions of participants achieving the MCID of the ISWT in the 3 groups, we also used a generalized LMM that included the same factors as the LMM described above. Evaluation of the secondary outcomes was performed analogously to the evaluation described above for the ISWT. Two sensitivity analyses were performed to evaluate the robustness of the primary analysis. One analysis excluded participants who did not complete the follow-up. A second analysis used multiple imputation to evaluate sensitivity for missing data under the assumption that data were missing at random. An iterative Markov chain Monte Carlo method was used to simulate 5 imputed values from the posterior predictive distribution of a multivariate normal model. To explore the predictors of clinical response and attrition, we performed 2 separate multivariable logistic regression models: one examining the baseline characteristics associated with achieving the MCID for ISWT distance, and another examining the factors associated with dropping out during follow-up. All statistical tests were performed at a 2-sided 5% significance level. We did not correct for multiple testing in this exploratory study. All analyses were performed in R (version 4.4.2; R Core Team).

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki. Ethical approval was granted by the Ethics Review Committee of the China-Japan Friendship Hospital (approval number: 2022-KY-024) and by the ethics committees of all participating centers. Written informed consent was obtained from all participants prior to enrollment. To protect participant privacy and confidentiality, all data were deidentified before analysis and storage. No financial or other compensation was provided to participants.

Results

Recruitment and Baseline Characteristics

Recruitment began in January 2023, with the first randomization occurring on February 7, 2023, and the final follow-up completed on October 23, 2023. Of the 102 patients screened, 16 were not eligible (7 failed lung function requirements; 5 had their diagnosis revised to asthma; and 4 were excluded due to unstable angina, recent exacerbation, or low digital literacy). Moreover, 16 patients did not participate in the trial for various reasons, such as “not interested” and “had no time.” Finally, 70 patients met the inclusion and exclusion criteria and were randomized into the MT, RW, or UC groups. Data from all 70 patients were included in the mITT analysis (Figure 1; Table S11 in Multimedia Appendix 2). The randomized patients with COPD had a mean age of 64.3 (SD 6.7) years and had moderate airflow obstruction, with a mean percentage of predicted FEV_1 of 54.5% (SD 14.7%; Table 2). Of the 70 participants, 55 (79%) completed the 12-week follow-up. Attrition was higher in the UC group (8/22, 36%) than in the MT group (5/25, 20%) and RW group (2/23, 9%).

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram. mITT: modified intention-to-treat; MT: multimodule training; RW: rhythm-guided walking; UC: usual care.

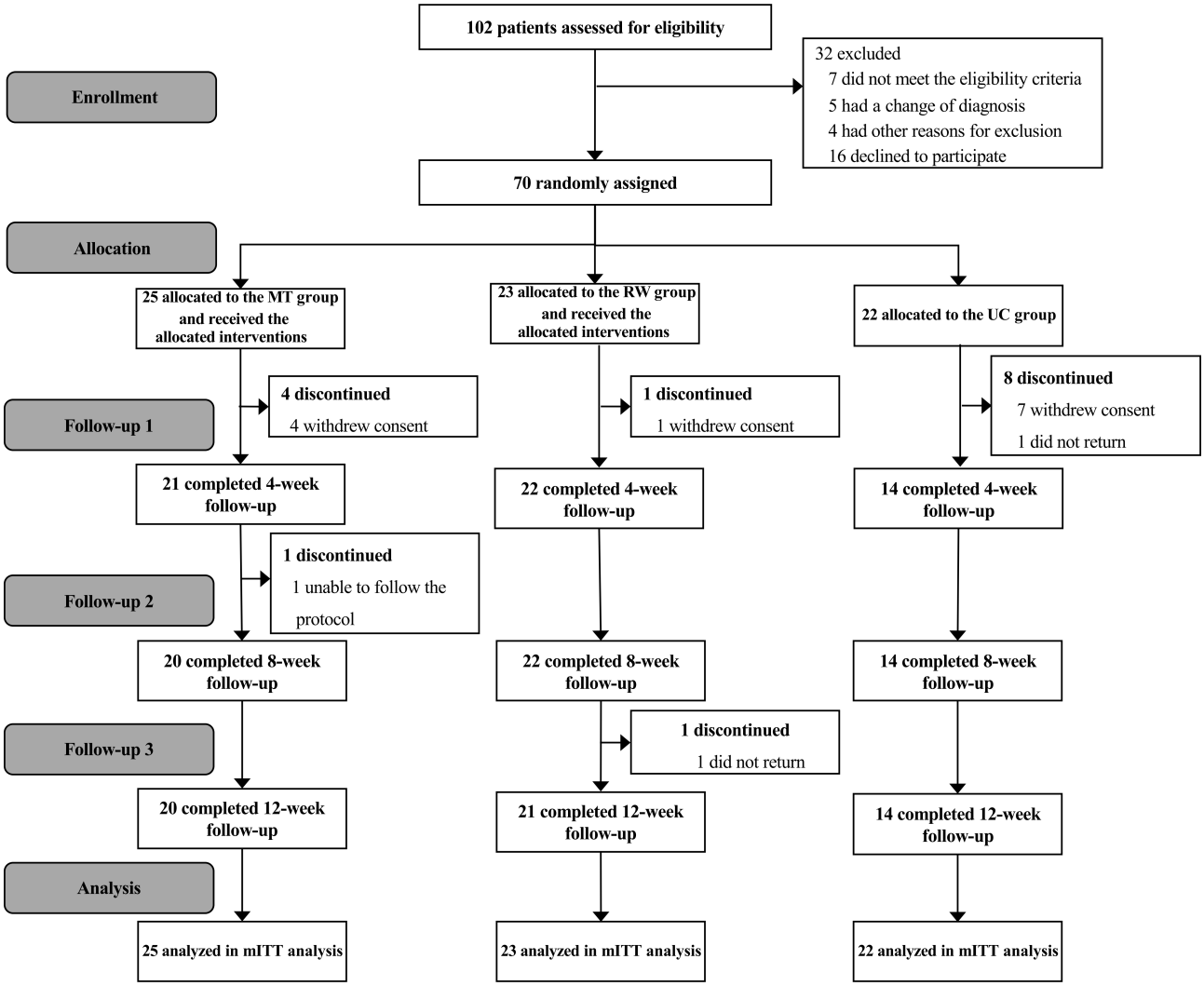

XSL•FO
RenderX

Table . Baseline characteristics of the study participants (N=70).

Variable	MT[a] group (n=25)	RW[b] group (n=23)	UC[c] group (n=22)	*P* value[d]
Male, n (%)	20 (80)	16 (70)	19 (86)	.43
Age (years), mean (SD)	65.6 (5.3)	64.6 (5.3)	62.7 (9.0)	.33
BMI, mean (SD)	24.6 (3.5)	25.2 (4.4)	25.6 (2.6)	.65
Smoking history, n (%)				.79
Currently smoking	5 (20)	6 (26)	5 (23)	
Previously smoked	12 (48)	10 (44)	13 (59)	
Never smoked	8 (32)	7 (30)	4 (18)	
FEV_1[e] (%pred), mean (SD)	52.5 (14.2)	57.2 (12.6)	53.9 (17.5)	.53
GOLD[f] classification, n (%)				.64
Class 2	16 (64)	16 (70)	12 (55)	
Class 3	7 (28)	7 (30)	8 (36)	
Class 4	2 (8)	0 (0)	2 (9)	
ISWT[g] distance (m), mean (SD)	384.7 (126.6)	365.0 (123.5)	367.7 (111.6)	.83
mMRC[h] score, mean (SD)	1.7 (0.9)	1.2 (0.4)	1.2 (0.7)	.02[i]
CAT[j] score, mean (SD)	14.0 (6.6)	12.4 (8.1)	12.5 (5.6)	.65
Medication, n (%)				.09
LAMA[k]	7 (28)	11 (48)	2 (9)	
LABA[l] + LAMA	5 (20)	3 (13)	4 (18)	
ICS[m] + LABA	6 (24)	7 (30)	10 (46)	
ICS + LABA + LAMA	7 (28)	2 (9)	6 (27)	

[a]MT: multimodule training.

[b]RW: rhythm-guided walking.

[c]UC: usual care.

[d]*P* values represent overall comparisons across the 3 groups (chi-square test or Fisher exact test for categorical variables; ANOVA for continuous variables).

[e]FEV_1: forced expiratory volume in 1 second.

[f]GOLD: Global Initiative for Chronic Obstructive Lung Disease.

[g]ISWT: incremental shuttle walking test.

[h]mMRC: modified Medical Research Council dyspnea scale.

[i]*P*<.05 (significant).

[j]CAT: COPD (chronic obstructive pulmonary disease) Assessment Test.

[k]LAMA: inhaled long-acting muscarinic antagonist.

[l]LABA: inhaled long-acting β2-agonist.

[m]ICS: inhaled corticosteroid.

Exercise Capacity

Changes in exercise capacity as measured by the ISWT over the 12-week period across the 3 groups are presented in Table 3 and Figure 2. At 12 weeks, the ISWT distance differed significantly among the 3 groups. Pairwise comparisons showed that the ISWT distance was significantly greater in the MT group than in the UC group (mean difference [MD] 56.35 m, 95% CI 6.66-106.04 m; *P*=.03; Cohen *d*=0.3), with the difference being clinically meaningful (Table 3; Table S2 in Multimedia Appendix 2). Although the RW group (the other intervention arm) showed a trend toward a greater ISWT distance compared with the UC group, the difference did not reach statistical significance (MD 38.91 m, 95% CI −10.91 to 88.74 m; *P*=.13). There was no significant difference in the ISWT distance between the MT and RW groups (*P*=.45; Table 3; Table S2 in Multimedia Appendix 2). The ISWT distance increased significantly from baseline to 12 weeks in both the MT group (*P*<.001) and RW group (*P*=.005) but not in the UC group (*P*=.39; Table 4). Additionally, the proportion of patients meeting the ISWT responder criteria was higher in the MT

XSL•FO
RenderX

group (13/20, 65%) and RW group (13/21, 62%) than in the UC group (4/14, 29%; Table 5).

Table . Comparison of the ISWT[a] distance among the MT[b], RW[c], and UC[d] groups at weeks 4, 8, and 12 (modified intention-to-treat population)[e].

Variable	Mean estimates (N=70)			MT group vs UC group		MT group vs RW group		RW group vs UC group	
	MT group (n=25)	RW group (n=23)	UC group (n=22)	Mean difference[f] (95% CI)	*P* value[g]	Mean difference[f] (95% CI)	*P* value[g]	Mean difference[f] (95% CI)	*P* value[g]
ISWT distance (m)									
4 weeks	377.4	371.9	374.8	2.55 (−46.78 to 51.88)	.92	5.46 (−38.11 to 49.04)	.81	−2.29 (−52.22 to 46.39)	.91
8 weeks	422.3	408.4	411.5	10.82 (−38.87 to 60.51)	.67	13.92 (−30.34 to 58.18)	.54	−3.10 (−52.40 to 46.20)	.90
12 weeks	445.2	427.7	388.8	56.35 (6.66 to 106.04)	.03[h]	17.44 (−27.35 to 62.23)	.45	38.91 (−10.91 to 88.74)	.13

[a]ISWT: incremental shuttle walking test.

[b]MT: multimodule training.

[c]RW: rhythm-guided walking.

[d]UC: usual care.

[e]The linear mixed model included treatment group, time point of measurement, treatment by time point interaction, age, gender, BMI, and ISWT at baseline (fixed effects), and site (random effect).

[f]Differences at weeks 4, 8, and 12 between the groups are based on the differences in the marginal means of the linear mixed model.

[g]*P* values represent pairwise comparisons from the linear mixed model.

[h]*P*<.05 (significant).

Figure 2. Changes in the primary outcome (incremental shuttle walking test [ISWT] distance) over 12 weeks in the multimodule training (MT), rhythm-guided walking (RW), and usual care (UC) groups. Data points show means, and error bars represent 95% CIs. **P*<.05 (MT group vs UC group).

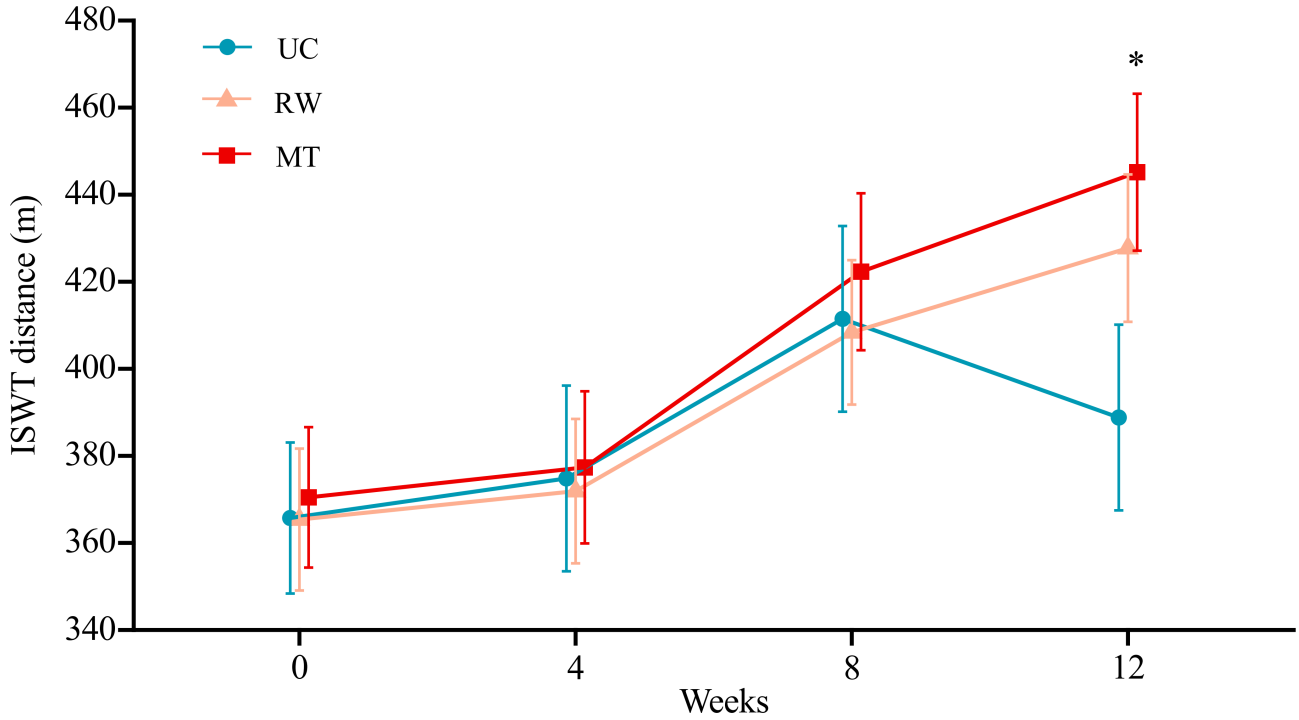

XSL•FO
RenderX

Table . Changes in primary and secondary outcomes from baseline to 12 weeks in the MT[a], RW[b], and UC[c] groups (N=70)[d].

Variable	MT group (n=25)	RW group (n=23)	UC group (n=22)
Exercise capacity			
ISWT[e] distance (m)			
Baseline, mean[f] (SD)	370.5 (16.1)	365.4 (16.3)	365.7 (17.4)
Follow-up at 12 weeks, mean[f] (SD)	445.2 (18.0)	427.7 (16.9)	388.8 (21.3)
Change from baseline[g], mean (95% CI)	74.69 (31.96 to 117.42)[h]	62.35 (19.51 to 105.18)[h]	23.08 (−25.54 to 71.70)
Dyspnea			
mMRC[i] score			
Baseline, mean[f] (SD)	1.5 (0.2)	1.3 (0.2)	1.3 (0.2)
Follow-up at 12 weeks, mean[f] (SD)	0.7 (0.2)	1.1 (0.2)	1.2 (0.2)
Change from baseline[g], mean (95% CI)	−0.73 (−1.04 to −0.42)[h]	−0.21 (−0.52 to 0.10)	−0.15 (−0.50 to 0.20)
Quality of life			
CAT[j] score			
Baseline, mean[f] (SD)	13.0 (1.2)	12.8 (1.2)	12.8 (1.3)
Follow-up at 12 weeks, mean[f] (SD)	7.4 (1.3)	9.9 (1.3)	10.6 (1.5)
Change from baseline[g], mean (95% CI)	−5.65 (−8.18 to −3.12)[h]	−2.87 (−5.41 to −0.33)[h]	−2.21 (−5.10 to 0.67)
SGRQ[k] score			
Baseline, mean[f] (SD)	37.8 (3.1)	37.0 (3.1)	36.2 (3.3)
Follow-up at 12 weeks, mean[f] (SD)	27.5 (3.3)	26.9 (3.2)	30.1 (3.7)
Change from baseline[g], mean (95% CI)	−10.30 (−16.23 to −4.37)[h]	−10.07 (−16.03 to −4.11)[h]	−6.02 (−12.78 to 0.74)
EQ-5D-5L			
Baseline, mean[f] (SD)	0.16 (0.02)	0.14 (0.02)	0.13 (0.03)
Follow-up at 12 weeks, mean[f] (SD)	0.04 (0.03)	0.05 (0.03)	0.07 (0.03)
Change from baseline[g], mean (95% CI)	−0.12 (−0.19 to −0.06)[h]	−0.09 (−0.15 to −0.03)[h]	−0.05 (−0.13 to 0.02)
Mental condition			
HADS-A[l] score			
Baseline, mean[f] (SD)	4.3 (0.7)	4.1 (0.7)	4.0 (0.7)
Follow-up at 12 weeks, mean[f] (SD)	2.7 (0.7)	3.9 (0.7)	5.0 (0.8)
Change from baseline[g], mean (95% CI)	−1.61 (−3.05 to −0.17)[h]	−0.21 (−1.66 to 1.24)	1.00 (−0.64 to 2.64)
HADS-D[m] score			
Baseline, mean[f] (SD)	4.7 (0.7)	4.6 (0.7)	4.1 (0.8)
Follow-up at 12 weeks, mean[f] (SD)	2.8 (0.8)	3.9 (0.7)	4.3 (0.9)
Change from baseline[g], mean (95% CI)	−1.91 (−3.50 to −0.31)[h]	−0.69 (−2.29 to 0.91)	0.18 (−1.63 to 2.00)
Respiratory muscle function			
MIP[n] (cmH_2O)			
Baseline, mean[f] (SD)	79.5 (4.2)	82.5 (4.3)	79.2 (4.5)
Follow-up at 12 weeks, mean[f] (SD)	92.2 (4.7)	82.3 (4.4)	85.5 (5.5)

XSL•FO
RenderX

Variable	MT group (n=25)	RW group (n=23)	UC group (n=22)
Change from baseline[g], mean (95% CI)	12.76 (2.14 to 23.38)[h]	−0.12 (−10.78 to 10.55)	6.37 (−5.73 to 18.48)
MEP[o] (cmH_2O)			
Baseline, mean[f] (SD)	85.7 (5.0)	89.1 (5.0)	82.3 (5.4)
Follow-up at 12 weeks, mean[f] (SD)	102.6 (5.5)	87.3 (5.2)	95.8 (6.5)
Change from baseline[g], mean (95% CI)	16.87 (4.33 to 29.40)[h]	−1.80 (−14.39 to 10.79)	13.53 (−0.76 to 27.82)

[a]MT: multimodule training.

[b]RW: rhythm-guided walking.

[c]UC: usual care.

[d]The linear mixed model included treatment group, time point of measurement, treatment by time point interaction, age, gender, BMI, and ISWT at baseline (fixed effects), and site (random effect).

[e]ISWT: incremental shuttle walking test.

[f]Means at baseline and week 12 are marginal means of the linear mixed model.

[g]Differences between week 12 and baseline within each group are based on the differences in the marginal means of the linear mixed model.

[h]$P<.05$ (significant); the P value comparison is for the outcome at baseline vs that at week 12 within each group.

[i]mMRC: modified Medical Research Council dyspnea scale.

[j]CAT: COPD (chronic obstructive pulmonary disease) Assessment Test.

[k]SGRQ: St. George's Respiratory Questionnaire.

[l]HADS-A: Hospital Anxiety and Depression Scale-anxiety subscale.

[m]HADS-D: Hospital Anxiety and Depression Scale-depression subscale.

[n]MIP: maximal inspiratory pressure.

[o]MEP: maximal expiratory pressure.

XSL•FO
RenderX

Table . Proportion of participants in the MT[a], RW[b], and UC[c] groups achieving the MCID[d] for the primary and secondary outcomes at week 12[e].

Variable	Participants (N=55), n (%)			MT group vs UC group		MT group vs RW group		RW group vs UC group	
	MT group (n=20)	RW group (n=21)	UC group (n=14)	Relative risk (95% CI)	*P* value	Relative risk (95% CI)	*P* value	Relative risk (95% CI)	*P* value
ISWT[f]	13 (65)	13 (62)	4 (29)	6.72 (1.19 - 38.16)	.03	1.14 (0.26 - 4.99)	.86	5.87 (1.00 - 34.60)	.05
mMRC[g]	14 (70)	6 (29)	4 (29)	5.63 (1.02 - 31.08)	.048	4.47 (0.95 - 21.01)	.06	1.26 (0.21 - 7.59)	.80
CAT[h]	14 (70)	12 (57)	10 (71)	1.99 (1.97 - 2.00)	<.001	2.15 (2.12 - 2.18)	<.001	0.92 (0.92 - 0.93)	<.001
SGRQ[i]	14 (70)	13 (62)	7 (50)	4.34 (0.64 - 29.33)	.13	1.17 (0.22 - 6.28)	.85	3.71 (0.52 - 26.47)	.19
HADS-A[j]	10 (50)	7 (33)	3 (21)	3.23 (0.43 - 24.11)	.25	2.78 (0.46 - 16.73)	.26	1.16 (0.12 - 10.97)	.90
HADS-D[k]	11 (55)	8 (38)	5 (36)	3.59 (0.45 - 28.90)	.23	4.72 (0.65 - 34.43)	.13	0.76 (0.09 - 6.10)	.80
MIP[l]	5 (25)	2 (10)	4 (29)	1.04 (0.15 - 7.25)	.97	2.23 (0.22 - 22.39)	.49	0.46 (0.05 - 4.52)	.51

[a]MT: multimodule training.

[b]RW: rhythm-guided walking.

[c]UC: usual care.

[d]MCID: minimal clinically important difference.

[e]The generalized linear mixed model included treatment group, time point of measurement, treatment by time point interaction, age, gender, BMI, and baseline values (fixed effects), and site (random effect).

[f]ISWT: incremental shuttle walking test.

[g]mMRC: modified Medical Research Council dyspnea scale.

[h]CAT: COPD (chronic obstructive pulmonary disease) Assessment Test.

[i]SGRQ: St. George's Respiratory Questionnaire.

[j]HADS-A: Hospital Anxiety and Depression Scale-anxiety subscale.

[k]HADS-D: Hospital Anxiety and Depression Scale-depression subscale.

[l]MIP: maximal inspiratory pressure.

Dyspnea

At 12 weeks, the mMRC score was significantly lower in the MT group than in the UC group (MD −0.44, 95% CI −0.80 to −0.08; *P*=.02) and RW group (MD −0.36, 95% CI −0.69 to −0.03; *P*=.03), indicating a reduced symptom burden in the MT group. However, there was no significant difference between the RW and UC groups (*P*=.64; Table 6; Figure 3). The mMRC score decreased significantly from baseline to 12 weeks in the MT group only (*P*<.001; Table 4). Additionally, the proportion of patients achieving the MCID was higher in the MT group (14/20, 70%) than in the RW group (6/21, 29%) and UC group (4/14, 29%; Table 5). However, there was no significant difference between the RW and UC groups (*P*=.80).

XSL•FO
RenderX

Table . Comparison of the secondary outcomes among the MT[a], RW[b], and UC[c] groups at week 12 (modified intention-to-treat population)[d].

Variable	Mean estimates (N=70)			MT group vs UC group		MT group vs RW group		RW group vs UC group	
	MT group (n=25)	RW group (n=23)	UC group (n=22)	Mean difference[e] (95% CI)	*P* value[f]	Mean difference[e] (95% CI)	*P* value[f]	Mean difference[e] (95% CI)	*P* value[f]
Dyspnea									
mMRC[g]	0.7	1.1	1.2	−0.44 (−0.80 to −0.08)	.02[h]	−0.36 (−0.69 to −0.03)	.03[h]	−0.09 (−0.44 to 0.27)	.64
Quality of life									
CAT[i]	7.4	9.9	10.6	−3.23 (−6.18 to −0.29)	.03[h]	−2.54 (−5.21 to 0.12)	.06	−0.69 (−3.66 to 2.28)	.65
SGRQ[j]	27.5	26.9	30.1	−2.67 (−9.59 to 4.24)	.45	0.54 (−5.70 to 6.77)	.87	−3.21 (−10.14 to 3.72)	.37
EQ-5D-5L	0.04	0.05	0.07	−0.04 (−0.11 to 0.04)	.33	−0.01 (−0.08 to 0.05)	.68	−0.02 (−0.09 to 0.05)	.55
Mental condition									
HADS-A[k]	2.7	3.9	5.0	−2.31 (−3.99 to −0.63)	.008[h]	−1.27 (−2.79 to 0.24)	.10	−1.04 (−2.72 to −0.65)	.23
HADS-D[l]	2.8	3.9	4.3	−1.44 (−3.29 to 0.42)	.13	−1.08 (−2.76 to 0.59)	.21	−0.35 (−2.21 to 1.51)	.71
Respiratory muscle function									
MIP[m] (cmH_2O)	92.2	82.3	85.5	6.70 (−5.68 to 19.07)	.29	9.90 (−1.28 to 21.07)	.08	−3.20 (−15.64 to 9.24)	.61
MEP[n] (cmH_2O)	102.6	87.3	95.8	6.73 (−7.91 to 21.38)	.37	15.24 (2.06 to 28.42)	.02[h]	−8.51 (−23.25 to 6.24)	.26
Prebronchodilator lung function (%pred)									
FEV_1[o]	62.4	59.8	61.0	1.41 (−3.74 to 6.55)	.59	2.62 (−1.98 to 7.23)	.27	−1.22 (−6.48 to 4.04)	.65
FVC[p]	92.3	91.4	86.9	5.35 (−0.39 to 11.10)	.07	0.82 (−4.24 to 5.87)	.75	4.54 (−1.25 to 10.33)	.13
FEV_1/FVC	51.9	50.8	48.4	3.42 (−0.81 to 7.65)	.12	1.06 (−2.74 to 4.86)	.59	2.36 (−1.94 to 6.66)	.29
PEF[q]	68.1	62.0	59.9	8.23 (−0.19 to 16.65)	.06	6.06 (−1.52 to 13.64)	.12	2.16 (−6.44 to 10.77)	.62
IC[r]	95.7	89.0	79.7	15.98 (4.76 to 27.21)	.01[h]	6.66 (−2.48 to 15.79)	.16	9.33 (−2.19 to 20.84)	.12
RV[s]	106.5	107.3	100.7	5.83 (−12.56 to 24.22)	.54	−0.80 (−17.14 to 15.53)	.92	6.63 (−12.11 to 25.38)	.49
FRC[t]	101.3	102.5	106.0	−4.71 (−15.90 to 6.47)	.41	−1.16 (−11.53 to 9.22)	.83	−3.56 (−14.84 to 7.72)	.54
TLC[u]	92.0	88.0	86.8	5.18 (−2.55 to 12.91)	.19	4.02 (−2.97 to 11.01)	.26	1.16 (−6.71 to 9.02)	.77

[a]MT: multimodule training.

[b]RW: rhythm-guided walking.

[c]UC: usual care.

[d]The linear mixed model included treatment group, time point of measurement, treatment by time point interaction, age, gender, BMI, and baseline values (fixed effects), and site (random effect).

[e]Differences at week 12 between the groups are based on the differences in the marginal means of the linear mixed model.

XSL•FO
RenderX

[f]*P* values represent pairwise comparisons from the linear mixed model.
[g]mMRC: modified Medical Research Council dyspnea scale.
[h]*P*<.05 (significant).
[i]CAT: COPD (chronic obstructive pulmonary disease) Assessment Test.
[j]SGRQ: St. George's Respiratory Questionnaire.
[k]HADS-A: Hospital Anxiety and Depression Scale-anxiety subscale.
[l]HADS-D: Hospital Anxiety and Depression Scale-depression subscale.
[m]MIP: maximal inspiratory pressure.
[n]MEP: maximal expiratory pressure.
[o]FEV_1: forced expiratory volume in 1 second.
[p]FVC: forced vital capacity.
[q]PEF: peak expiratory flow.
[r]IC: inspiratory capacity.
[s]RV: residual volume.
[t]FRC: functional residual capacity.
[u]TLC: total lung capacity.

Figure 3. Changes in the key secondary outcomes over 12 weeks in the multimodule training (MT), rhythm-guided walking (RW), and usual care (UC) groups: (A) dyspnea symptoms (modified Medical Research Council dyspnea scale [mMRC] score), (B) quality of life (COPD Assessment Test [CAT] score), (C) anxiety symptoms (Hospital Anxiety and Depression Scale [HADS] Anxiety score), and (D) inspiratory capacity. Data points show means, and error bars represent 95% CIs. **P*<.05 (MT group vs UC group), ***P*<.05 (MT group vs RW group).

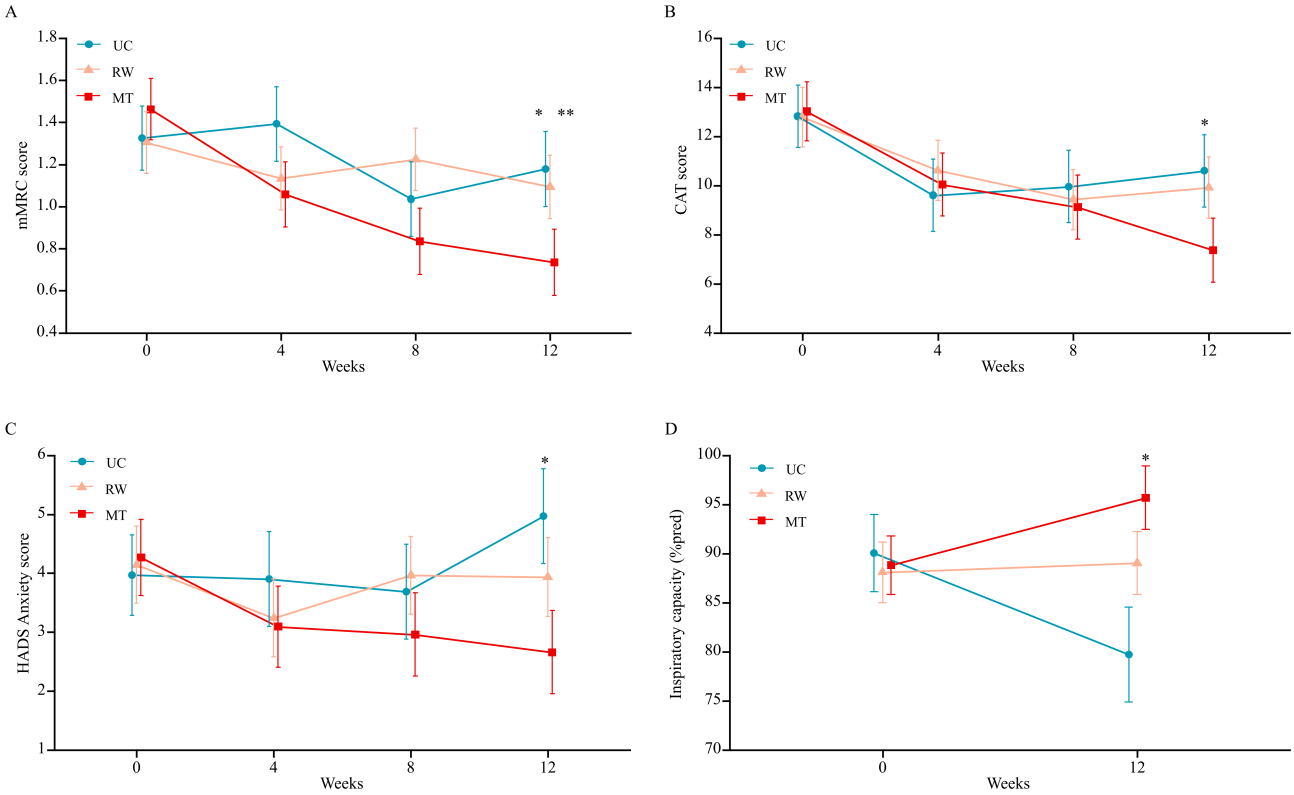

Assessment of HRQoL

For the SGRQ and EQ-5D-5L scores, no significant differences were observed among the 3 groups at 12 weeks (all *P*≥.05; Table 6). Nevertheless, clinically meaningful improvements from baseline to 12 weeks were observed in both the MT group (SGRQ: *P*<.001; EQ-5D-5L: *P*<.001) and RW group (SGRQ: *P*=.001; EQ-5D-5L: *P*=.004); however, changes in the UC group were not significant (SGRQ: *P*=.10; EQ-5D-5L: *P*=.14; Table 4). The proportion of participants achieving the MCID for the SGRQ did not differ significantly among the 3 groups (Table 5). For the CAT scores at 12 weeks, a significant improvement exceeding the MCID was observed in the MT group compared with the UC group (MD −3.23, 95% CI −6.18 to −0.29; *P*=.03; Table 6; Figure 3). However, there was no significant difference between the RW and UC groups (*P*=.65) or between the MT and RW groups (*P*=.06). In terms of the change in each group from baseline to 12 weeks, the CAT score improved significantly in both the MT group (*P*<.001) and RW group (*P*=.03) but not in the UC group (*P*=.19; Table 4).

Anxiety and Depression

Regarding anxiety symptoms, pairwise comparisons revealed statistically significant differences among the 3 groups at 12

XSL•FO
RenderX

months of follow-up (Table 6; Figure 3). Specifically, the MT group demonstrated significantly greater reductions in anxiety compared with the UC group (MD −2.31, 95% CI −3.99 to −0.63; *P*=.008). There were no significant differences between the RW and UC groups (*P*=.23) or between the MT and RW groups (*P*=.10). The HADS-A score improved significantly from baseline to 12 weeks in the MT group (*P*=.03) but not in the RW group (*P*=.80) or UC group (*P*=.24; Table 4). The proportion of patients achieving the MCID for the HADS-A score was higher in both the MT group (10/20, 50%) and RW group (7/21, 33%) than in the UC group (3/14, 21%). However, these differences did not reach statistical significance (MT vs UC: *P*=.23; RW vs UC: *P*=.80; Table 5). A similar trend was observed for depression scores, though the difference across the 3 groups did not reach statistical significance (Tables 5 and 6).

Lung Function Measurements

Among all pulmonary function parameters assessed at 12 weeks, only inspiratory capacity (IC) showed a statistically significant difference in pairwise comparisons among the 3 groups. Specifically, IC was significantly higher in the MT group than in the UC group (MD 15.98% predicted, 95% CI 4.76%-27.21%; *P*=.007). However, no significant differences were observed between the RW and UC groups (*P*=.12) or between the MT and RW groups (*P*=.16; Table 6; Figure 3). The MT group also tended to be better at increasing the peak expiratory flow compared with the UC group at 12 weeks, though the difference was not statistically significant (*P*=.06; Table 6).

App Adherence and Satisfaction

During the 12 weeks, except for patients who dropped out midway, all patients in the 2 intervention groups who completed the 12-week follow-up remained adherent users (MT: 20/25, 80%; RW: 21/23, 91%). The mean training session completion rate was 96.1% (SD 7.1%) in the MT group and 97.1% (SD 4.3%) in the RW group (Table S9 in Multimedia Appendix 2). Patient satisfaction with the app was very high. Except for patients who dropped out midway, all patients in the 2 intervention groups who completed the 12-week follow-up (MT: 20/25, 80%; RW: 21/23, 91%) were satisfied or very satisfied with the app (Table S10 in Multimedia Appendix 2). No participant withdrew from this study due to unresolved technical issues.

Sensitivity Analysis

Sensitivity analyses for missing data did not show marked differences from the primary analyses (Table S6 in Multimedia Appendix 2). From the per-protocol analysis, the ISWT distance was significantly higher among participants adhering to MT than among participants receiving UC at 12 weeks (*P*=.03). There was no significant difference between the RW and UC groups (*P*=.61) or between the MT and RW groups (*P*=.46; Table S4 in Multimedia Appendix 2). The results for secondary outcomes were also similar to the findings of the mITT analysis (Table S5 in Multimedia Appendix 2).

No significant correlations were identified between baseline characteristics or other factors and achieving the MCID for the ISWT (all *P*≥.05; Table S7 in Multimedia Appendix 2). Additionally, there were no clear predictors of noncompletion among the collected characteristics (Table S8 in Multimedia Appendix 2).

Discussion

Principal Findings

Our study demonstrated that the use of a 12-week music-facilitated PR program (including tempo-guided walking and singing) based on mHealth technology resulted in clinically meaningful improvements in exercise capacity among patients with COPD compared to UC, with a moderate effect size [52]. Secondary outcomes also showed benefits with MT compared to UC, including improved dyspnea, quality of life, psychological status, and IC.

Our findings align with recent research on music interventions for COPD [20,22,23,30-32,53]. However, the findings are not directly comparable as previous studies involved outpatient settings or community care, and the intervention was delivered to individuals or small groups by qualified music therapists and physiotherapists [24,25]. In COPDMELODY, music-facilitated PR was home-based and provided through predesigned online programs with a smartphone app. The consistent findings suggest that home-based music interventions may be as effective as hospital- or community-based interventions targeting COPD, aided by well-designed training sessions and remote technology. This might help with the great imbalance between the large COPD population and restricted rehabilitation resources.

Following the Medical Research Council framework, the MT program represents a multicomponent complex intervention where various elements interact to produce clinical benefits [54]. The foundation of this intervention is aerobic exercise through tempo-guided walking. Within this framework, music added to exercise serves as a critical distractive auditory stimulus, shifting the patient's focus away from an attention-demanding sensation (dyspnea) toward an external one (listening to music), which enables them to reach and maintain target exercise intensities more stably [20,21]. Singing acts as a key "active ingredient" specifically targeting respiratory mechanics. By controlling expiratory flow, lowering operating lung volumes, and doming the diaphragm to support exhalation, singing is more effective in reducing dyspnea symptoms [29]. This was supported by analyses of lung function indicating that MT was superior to UC at decreasing pulmonary hyperinflation, which is one of the main causes of dyspnea in COPD [55,56]. Furthermore, the intervention's context was optimized through mHealth technology. The integration of smartphone apps and wearable wristwatches has been shown to reduce the complexity of delivering MT in a home setting while ensuring high protocol fidelity [14,19]. From a behavioral perspective, the inherent enjoyment and pleasure derived from the music-based program enhanced motivation and adherence [57-59], which was supported by the training data objectively measured by the app.

In contrast, RW failed to significantly improve clinical outcomes compared to UC, despite causing significant within-group improvements from baseline to 12 weeks in exercise capacity (ISWT distance) and quality of life (CAT and SGRQ scores).

XSL•FO
RenderX

The findings do not align with previous studies suggesting that tempo-guided walking alone could improve exercise capacity [20,53], a discrepancy that may be attributed to the fact that the duration of walking training in those studies was substantially longer than that in our trial. Consequently, these findings suggest that the duration and frequency of walking sessions in our trial may need to be increased or supplemented with other active training modules.

Regarding the difference between MT and RW, MT demonstrated a potentially greater benefit in alleviating dyspnea compared to RW. This likely reflects the additive value of singing as a critical "active ingredient" within the complex intervention framework [22,29]. Such a multimodule approach appears to address the compound pathophysiological changes in COPD more comprehensively; however, these preliminary observations warrant further validation in future studies with adequate statistical power. We also explored the baseline predictors of achieving the MCID for the ISWT distance and those of study dropout. These exploratory analyses were conducted to identify which patients are most likely to benefit from smartphone app–based music-facilitated PR and which patients are at the greatest risk of attrition—key questions in the field. No significant predictors were identified in either analysis. The absence of clear predictors in our sample may reflect the relatively modest sample size, the homogeneity of our stable COPD cohort, and the lack of music-related personal characteristics collected. However, several previous PR studies have demonstrated that high adherence is the strongest driver of achieving the MCID for the 6-Minute Walk Test, whereas current smoking, poor lung function, and high symptom burden are associated with an increased risk of dropout [23,60]. These findings highlight the need for larger-scale and more refined future studies to better identify patient subgroups that are most likely to benefit from this kind of PR program.

Strengths and Limitations

Our research has several advantages. COPDMELODY is the translation of research evidence for music-facilitated exercise and singing training from a hospital or community context to a home context. Most studies focus on a single kind of intervention, while an optimal rehabilitation program relies on diverse components targeting compound pathophysiological changes in COPD [33]. Following the framework for complex interventions, we innovatively leveraged the synergistic effects of combining receptive and active music interventions, enabling patients to receive the maximum benefit. Predesigned music used to guide training was provided by qualified music therapists and experienced clinicians. The training time and intensity of each patient were captured by the app and wristwatch. Besides a face-to-face course on how to perform tempo-guided walking and singing, patients had a videoconference with us during their first training session. This approach helped our trial achieve high intervention protocol fidelity [61]. To ensure the personalization and incrementality of PR, we performed the ISWT for each patient every 4 weeks [35]. Finally, this study will inform a larger trial in terms of methodology, eligibility criteria, recruitment process, sample size, randomization procedure, data collection, and data analysis process. The trial will allow for a review of the case definitions of primary outcomes, an overview of participant retention rates, and the identification of unanticipated challenges that can be addressed when preparing for a larger trial.

Our study also has some limitations. Our study was a proof-of-concept study on the impact of home-based music-facilitated PR. This trial was limited by its design and small sample size, and it lacked adequate power to demonstrate significant effects for all measured outcomes. The study made multiple comparisons between arms, which can lead to a false inference. However, as this was a pilot study, the results will not be used to make treatment decisions, but the findings can be used to generate hypotheses in subsequent, adequately powered trials. Although the final sample size exceeded the minimum target calculated in the sample size estimation, enrollment was terminated earlier than originally planned in the protocol due to practical constraints. In addition, the outcomes were evaluated over a relatively short period; thus, the intervention effect and its sustainability need to be validated in future studies. During the follow-up, the ISWT distance increased steadily among participants in the MT group but showed a fluctuating trend among participants in the UC group, whose exercise capacity might be influenced by autonomous physical activities [62]. Other clinical outcomes, including subjective and objective outcomes, supported the benefits of our PR program. Furthermore, the study population was limited to smartphone users, potentially excluding those with lower digital literacy. However, with the continuously growing number of smartphone users globally, our program has the potential to provide a supplement to or substitute (according to individual needs) for traditional center-based PR [63]. The music tracks and songs were predefined rather than based on individual choice. Although we curated "golden oldies" (1950s to 1970s) and popular hits from Chinese streaming platforms, the lack of personalization may have influenced engagement. Future research could further enhance this model by allowing participants to select music to maximize the benefits of the program.

Conclusions

The findings of this study demonstrate that our smartphone app–based music-facilitated multicomponent PR program, involving RW and singing, has a positive effect on physiological and psychological outcomes among patients with COPD compared to UC. Specifically, exercise capacity, dyspnea, quality of life, anxiety, and IC showed significantly better improvements with MT than with UC. There is a need for further longitudinal, multicenter studies with fully powered samples to confirm sustained benefits and identify additional statistically significant events.

XSL•FO
RenderX

Acknowledgments

We acknowledge CHIC Health Technology (Shanghai, China) for their technical support in the design and development of the mini-program in the Qiyue app used in our study. We thank the research staff of the COPDMELODY (Music Therapy for COPD Rehabilitation) research group for their dedication.

Funding

This study was supported by the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2023ZD0506003), the CAMS Innovation Fund for Medical Sciences (CIFMS) (2021-I2M-1-049, 2022-I2M-C&T-B-107), and research projects of China-Japan Friendship Hospital (2022-HX-16, ZRJY2024-GG06).

Data Availability

Data may be made available according to the data sharing policy of local partners and ethics committees upon reasonable request to the corresponding author.

Authors' Contributions

M.S. and S.Q. conceived and designed the study. X.B., J.Z., K.L., J.W., Y.Z., J.L., S.W., W.L., X.T., J.Y. and Y.P. contributed significantly to data collection, analysis and interpretation. K.H., T.Y. and C.W. supervised the study and critically revised the manuscript for important intellectual content. K.H. and T.Y. contributed equally as co-corresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Clinical trial protocol.
[DOC File, 473 KB - jmir_v28i1e81707_app1.doc]

Multimedia Appendix 2
Additional material to support the study.
[DOCX File, 6963 KB - jmir_v28i1e81707_app2.docx]

Checklist 1
CONSORT checklist.
[DOC File, 223 KB - jmir_v28i1e81707_app3.doc]

References

1. Momtazmanesh S, Moghaddam SS, Ghamari SH. Global burden of chronic respiratory diseases and risk factors, 1990-2019: an update from the Global Burden of Disease Study 2019. EClinicalMedicine 2023 May;59:101936. [doi: 10.1016/j.eclinm.2023.101936] [Medline: 37229504]
2. Wang C, Xu J, Yang L, et al. Prevalence and risk factors of chronic obstructive pulmonary disease in China (the China Pulmonary Health [CPH] study): a national cross-sectional study. Lancet 2018 Apr 28;391(10131):1706-1717. [doi: 10.1016/S0140-6736(18)30841-9] [Medline: 29650248]
3. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. : Global Initiative for Chronic Obstructive Lung Disease, Inc; 2026 URL: https://goldcopd.org/wp-content/uploads/2026/01/GOLD-REPORT-2026-v1.3-8Dec2025_WMV2.pdf [accessed 2026-04-15]
4. Spruit MA, Singh SJ, Garvey C, et al. An official American Thoracic Society/European Respiratory Society statement: key concepts and advances in pulmonary rehabilitation. Am J Respir Crit Care Med 2013 Oct 15;188(8):e13-e64. [doi: 10.1164/rccm.201309-1634ST] [Medline: 24127811]
5. Rochester CL, Alison JA, Carlin B, et al. Pulmonary rehabilitation for adults with chronic respiratory disease: an official American Thoracic Society clinical practice guideline. Am J Respir Crit Care Med 2023 Aug 15;208(4):e7-e26. [doi: 10.1164/rccm.202306-1066ST] [Medline: 37581410]
6. Lindenauer PK, Stefan MS, Pekow PS, et al. Association between initiation of pulmonary rehabilitation after hospitalization for COPD and 1-year survival among Medicare beneficiaries. JAMA 2020 May 12;323(18):1813-1823. [doi: 10.1001/jama.2020.4437] [Medline: 32396181]
7. Press VG, Nici L. Improving pulmonary rehabilitation participation: it's time to examine the system. Annals ATS 2024 Dec;21(12):1655-1656. [doi: 10.1513/AnnalsATS.202410-1003ED]

XSL•FO
RenderX

8. Rochester CL, Vogiatzis I, Holland AE, et al. An official American Thoracic Society/European Respiratory Society policy statement: enhancing implementation, use, and delivery of pulmonary rehabilitation. Am J Respir Crit Care Med 2015 Dec 1;192(11):1373-1386. [doi: 10.1164/rccm.201510-1966ST] [Medline: 26623686]
9. Keating A, Lee A, Holland AE. What prevents people with chronic obstructive pulmonary disease from attending pulmonary rehabilitation? A systematic review. Chron Respir Dis 2011;8(2):89-99. [doi: 10.1177/1479972310393756] [Medline: 21596892]
10. Uzzaman MN, Agarwal D, Chan SC, et al. Effectiveness of home-based pulmonary rehabilitation: systematic review and meta-analysis. Eur Respir Rev 2022 Sep 30;31(165):220076. [doi: 10.1183/16000617.0076-2022] [Medline: 36130789]
11. Alwashmi M, Hawboldt J, Davis E, Marra C, Gamble JM, Abu Ashour W. The effect of smartphone interventions on patients with chronic obstructive pulmonary disease exacerbations: a systematic review and meta-analysis. JMIR Mhealth Uhealth 2016 Sep 1;4(3):e105. [doi: 10.2196/mhealth.5921] [Medline: 27589898]
12. Hansen H, Bieler T, Beyer N, et al. Supervised pulmonary tele-rehabilitation versus pulmonary rehabilitation in severe COPD: a randomised multicentre trial. Thorax 2020 May;75(5):413-421. [doi: 10.1136/thoraxjnl-2019-214246] [Medline: 32229541]
13. Cox NS, Dal Corso S, Hansen H, et al. Telerehabilitation for chronic respiratory disease. Cochrane Database Syst Rev 2021 Jan 29;1(1):CD013040. [doi: 10.1002/14651858.CD013040.pub2] [Medline: 33511633]
14. Gloeckl R, Spielmanns M, Stankeviciene A, et al. Smartphone application-based pulmonary rehabilitation in COPD: a multicentre randomised controlled trial. Thorax 2025 Mar 18;80(4):209-217. [doi: 10.1136/thorax-2024-221803] [Medline: 39706685]
15. Spielmanns M, Gloeckl R, Jarosch I, et al. Using a smartphone application maintains physical activity following pulmonary rehabilitation in patients with COPD: a randomised controlled trial. Thorax 2023 May;78(5):442-450. [doi: 10.1136/thoraxjnl-2021-218338] [Medline: 35450945]
16. Chung C, Lee JW, Lee SW, Jo MW. Clinical efficacy of mobile app-based, self-directed pulmonary rehabilitation for patients with chronic obstructive pulmonary disease: systematic review and meta-analysis. JMIR Mhealth Uhealth 2024 Jan 4;12:e41753. [doi: 10.2196/41753] [Medline: 38179689]
17. Zhuang M, Hassan II, W Ahmad WMA, et al. Effectiveness of digital health interventions for chronic obstructive pulmonary disease: systematic review and meta-analysis. J Med Internet Res 2025 May 26;27:e76323. [doi: 10.2196/76323] [Medline: 40418567]
18. Dai Y, Huang H, Zhang Y, He N, Shen M, Li H. The effects of telerehabilitation on physiological function and disease symptom for patients with chronic respiratory disease: a systematic review and meta-analysis. BMC Pulm Med 2024 Jun 28;24(1):305. [doi: 10.1186/s12890-024-03104-8] [Medline: 38943129]
19. Chung C, Kim DK, Lee JK, et al. Smartphone app-guided pulmonary rehabilitation in chronic respiratory diseases: randomized controlled trial. JMIR Mhealth Uhealth 2025 Aug 25;13:e64884. [doi: 10.2196/64884] [Medline: 40854157]
20. Lee AL, Desveaux L, Goldstein RS, Brooks D. Distractive auditory stimuli in the form of music in individuals with COPD: a systematic review. Chest 2015 Aug;148(2):417-429. [doi: 10.1378/chest.14-2168] [Medline: 25741661]
21. Lee AL, Dolmage TE, Rhim M, Goldstein RS, Brooks D. The impact of listening to music during a high-intensity exercise endurance test in people with COPD. Chest 2018 May;153(5):1134-1141. [doi: 10.1016/j.chest.2017.12.001] [Medline: 29253555]
22. Lewis A, Cave P, Stern M, et al. Singing for Lung Health-a systematic review of the literature and consensus statement. NPJ Prim Care Respir Med 2016 Dec 1;26:16080. [doi: 10.1038/npjpcrm.2016.80] [Medline: 27906158]
23. Kaasgaard M, Rasmussen DB, Andreasson KH, et al. Use of Singing for Lung Health as an alternative training modality within pulmonary rehabilitation for COPD: a randomised controlled trial. Eur Respir J 2022 May;59(5):2101142. [doi: 10.1183/13993003.01142-2021] [Medline: 34625480]
24. Feng X, Gao Y, Hu H, et al. The effects of music therapy on patients with chronic obstructive pulmonary disease: a systematic review and meta-analysis. Physiother Theory Pract 2025 Jul;41(7):1478-1494. [doi: 10.1080/09593985.2024.2420010] [Medline: 39487969]
25. Huang J, Yuan X, Zhang N, Qiu H, Chen X. Music therapy in adults with COPD. Respir Care 2021 Mar;66(3):501-509. [doi: 10.4187/respcare.07489] [Medline: 33144384]
26. Hernandez A, Bronas UG, Steffen AD, et al. Rhythmic auditory stimulation increases 6-minute walk distance in individuals with COPD: a repeated measures study. Heart Lung 2020;49(3):324-328. [doi: 10.1016/j.hrtlng.2020.01.006] [Medline: 32067724]
27. Liang Z, Yip H, Sena Moore K, et al. Self-managed music-guided exercise intervention improved upper and lower extremity muscle strength for ICU survivors-a pilot randomized controlled study. Biol Res Nurs 2022 Apr;24(2):145-151. [doi: 10.1177/10998004211050297] [Medline: 34738474]
28. Koelsch S. Brain correlates of music-evoked emotions. Nat Rev Neurosci 2014 Mar;15(3):170-180. [doi: 10.1038/nrn3666] [Medline: 24552785]
29. Lewis A, Philip KEJ, Lound A, Cave P, Russell J, Hopkinson NS. The physiology of singing and implications for “Singing for Lung Health” as a therapy for individuals with chronic obstructive pulmonary disease. BMJ Open Respir Res 2021 Nov;8(1):e000996. [doi: 10.1136/bmjresp-2021-000996] [Medline: 34764199]

XSL•FO
RenderX

30. Liu H, Song M, Zhai ZH, Shi RJ, Zhou XL. Group singing improves depression and life quality in patients with stable COPD: a randomized community-based trial in China. Qual Life Res 2019 Mar;28(3):725-735. [doi: 10.1007/s11136-018-2063-5] [Medline: 30612266]
31. Hart MK, Stewardson E, Jamil AK, Tecson KM, Millard MW. Usefulness of harmonica playing to improve outcomes in patients with chronic obstructive pulmonary disease. Proc (Bayl Univ Med Cent) 2020 Apr;33(2):178-182. [doi: 10.1080/08998280.2019.1704135] [Medline: 32313456]
32. Canga B, Azoulay R, Raskin J, Loewy J. AIR: Advances in Respiration - music therapy in the treatment of chronic pulmonary disease. Respir Med 2015 Dec;109(12):1532-1539. [doi: 10.1016/j.rmed.2015.10.001] [Medline: 26522499]
33. Holland AE, Cox NS, Houchen-Wolloff L, et al. Defining modern pulmonary rehabilitation. an official American Thoracic Society workshop report. Ann Am Thorac Soc 2021 May;18(5):e12-e29. [doi: 10.1513/AnnalsATS.202102-146ST] [Medline: 33929307]
34. Shi M, Yang L, Qumu S, et al. Efficacy and safety of a music-therapy facilitated pulmonary telerehabilitation program in COPD patients: the COPDMELODY study protocol. Front Med (Lausanne) 2024;11:1361053. [doi: 10.3389/fmed.2024.1361053] [Medline: 38523907]
35. Singh SJ, Morgan MD, Scott S, Walters D, Hardman AE. Development of a shuttle walking test of disability in patients with chronic airways obstruction. Thorax 1992 Dec;47(12):1019-1024. [doi: 10.1136/thx.47.12.1019] [Medline: 1494764]
36. Endurance - lower limb. Australian Pulmonary Rehabilitation - Lung Foundation Australia. URL: https://pulmonaryrehab.com.au/exercise-training/exercise-prescription/endurance-lower-limb [accessed 2026-04-15]
37. Palermo P, Corrà U. Exercise prescriptions for training and rehabilitation in patients with heart and lung disease. Ann Am Thorac Soc 2017 Jul;14(Supplement_1):S59-S66. [doi: 10.1513/AnnalsATS.201702-160FR] [Medline: 28719753]
38. Singh SJ, Morgan MD, Hardman AE, Rowe C, Bardsley PA. Comparison of oxygen uptake during a conventional treadmill test and the shuttle walking test in chronic airflow limitation. Eur Respir J 1994 Nov;7(11):2016-2020. [Medline: 7875275]
39. Holland AE, Spruit MA, Troosters T, et al. An official European Respiratory Society/American Thoracic Society technical standard: field walking tests in chronic respiratory disease. Eur Respir J 2014 Dec;44(6):1428-1446. [doi: 10.1183/09031936.00150314] [Medline: 25359355]
40. Evans RA, Singh SJ. Minimum important difference of the incremental shuttle walk test distance in patients with COPD. Thorax 2019 Oct;74(10):994-995. [doi: 10.1136/thoraxjnl-2018-212725] [Medline: 31147399]
41. Bestall JC, Paul EA, Garrod R, Garnham R, Jones PW, Wedzicha JA. Usefulness of the Medical Research Council (MRC) dyspnoea scale as a measure of disability in patients with chronic obstructive pulmonary disease. Thorax 1999 Jul;54(7):581-586. [doi: 10.1136/thx.54.7.581] [Medline: 10377201]
42. Dodd JW, Hogg L, Nolan J, et al. The COPD assessment test (CAT): response to pulmonary rehabilitation. A multicentre, prospective study. Thorax 2011 May;66(5):425-429. [doi: 10.1136/thx.2010.156372] [Medline: 21398686]
43. Rutten-van Mölken M, Oostenbrink JB, Tashkin DP, Burkhart D, Monz BU. Does quality of life of COPD patients as measured by the generic EuroQol five-dimension questionnaire differentiate between COPD severity stages? Chest 2006 Oct;130(4):1117-1128. [doi: 10.1378/chest.130.4.1117] [Medline: 17035446]
44. Xu W, Collet JP, Shapiro S, et al. Validation and clinical interpretation of the St George's Respiratory Questionnaire among COPD patients, China. Int J Tuberc Lung Dis 2009 Feb;13(2):181-189. [Medline: 19146745]
45. Bjelland I, Dahl AA, Haug TT, Neckelmann D. The validity of the Hospital Anxiety and Depression Scale. an updated literature review. J Psychosom Res 2002 Feb;52(2):69-77. [doi: 10.1016/s0022-3999(01)00296-3] [Medline: 11832252]
46. Miller MR, Crapo R, Hankinson J, et al. General considerations for lung function testing. Eur Respir J 2005 Jul;26(1):153-161. [doi: 10.1183/09031936.05.00034505] [Medline: 15994402]
47. Laveneziana P, Albuquerque A, Aliverti A, et al. ERS statement on respiratory muscle testing at rest and during exercise. Eur Respir J 2019 Jun;53(6):1801214. [doi: 10.1183/13993003.01214-2018] [Medline: 30956204]
48. Leung RWM, McKeough ZJ, Peters MJ, Alison JA. Short-form Sun-style t'ai chi as an exercise training modality in people with COPD. Eur Respir J 2013 May;41(5):1051-1057. [doi: 10.1183/09031936.00036912]
49. Wang CH, Chou PC, Joa WC, et al. Mobile-phone-based home exercise training program decreases systemic inflammation in COPD: a pilot study. BMC Pulm Med 2014 Aug 30;14(1):142. [doi: 10.1186/1471-2466-14-142] [Medline: 25175787]
50. Alghamdi SM, Rajah AMA, Aldabayan YS, Aldhahir AM, Alqahtani JS, Alzahrani AA. Chronic obstructive pulmonary disease patients' acceptance in e-health clinical trials. Int J Environ Res Public Health 2021 May 14;18(10):5230. [doi: 10.3390/ijerph18105230] [Medline: 34069028]
51. Spielmanns M, Fuchs-Bergsma C, Winkler A, Fox G, Krüger S, Baum K. Effects of oxygen supply during training on subjects with COPD who are normoxemic at rest and during exercise: a blinded randomized controlled trial. Respir Care 2015 Apr;60(4):540-548. [doi: 10.4187/respcare.03647] [Medline: 25516993]
52. Kinney AR, Eakman AM, Graham JE. Novel effect size interpretation guidelines and an evaluation of statistical power in rehabilitation research. Arch Phys Med Rehabil 2020 Dec;101(12):2219-2226. [doi: 10.1016/j.apmr.2020.02.017] [Medline: 32272106]
53. Ho CF, Maa SH, Shyu YIL, Lai YT, Hung TC, Chen HC. Effectiveness of paced walking to music at home for patients with COPD. COPD 2012 Aug;9(5):447-457. [doi: 10.3109/15412555.2012.685664] [Medline: 22643033]

XSL•FO
RenderX

54. Skivington K, Matthews L, Simpson SA, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061. [doi: 10.1136/bmj.n2061] [Medline: 34593508]
55. de Freitas A, Belo LF, Martinez L, Hernandes NA, Pitta F. Use of the Borg dyspnea scale to identify dynamic hyperinflation during the 6-minute walking test in individuals with moderate-severe COPD: a pilot study. Pulmonology 2023;29(4):335-337. [doi: 10.1016/j.pulmoe.2023.02.003] [Medline: 36907815]
56. Philip KEJ, Owles H, McVey S, et al. An online breathing and wellbeing programme (ENO Breathe) for people with persistent symptoms following COVID-19: a parallel-group, single-blind, randomised controlled trial. Lancet Respir Med 2022 Sep;10(9):851-862. [doi: 10.1016/S2213-2600(22)00125-4] [Medline: 35489367]
57. Baker FA, Pac Soo V, Bloska J, et al. Home-based family caregiver-delivered music and reading interventions for people living with dementia (HOMESIDE trial): an international randomised controlled trial. EClinicalMedicine 2023 Nov;65:102224. [doi: 10.1016/j.eclinm.2023.102224] [Medline: 38106552]
58. Gómez-Gallego M, Gómez-Gallego JC, Gallego-Mellado M, García-García J. Comparative efficacy of active group music intervention versus group music listening in Alzheimer's disease. Int J Environ Res Public Health 2021 Jul 30;18(15):8067. [doi: 10.3390/ijerph18158067] [Medline: 34360360]
59. Terry PC, Karageorghis CI, Curran ML, Martin OV, Parsons-Smith RL. Effects of music in exercise and sport: a meta-analytic review. Psychol Bull 2020 Feb;146(2):91-117. [doi: 10.1037/bul0000216] [Medline: 31804098]
60. Bamonti PM, Boyle JT, Goodwin CL, et al. Predictors of outpatient pulmonary rehabilitation uptake, adherence, completion, and treatment response among male U.S. veterans with chronic obstructive pulmonary disease. Arch Phys Med Rehabil 2022 Jun;103(6):1113-1121. [doi: 10.1016/j.apmr.2021.10.021] [Medline: 34856155]
61. Velez M, Lugo-Agudelo LH, Patiño Lugo DF, et al. Factors that influence the provision of home-based rehabilitation services for people needing rehabilitation: a qualitative evidence synthesis. Cochrane Database Syst Rev 2023 Feb 10;2(2):CD014823. [doi: 10.1002/14651858.CD014823] [Medline: 36780267]
62. Paixão C, Rocha V, Brooks D, Marques A. Unsupervised physical activity interventions for people with COPD: a systematic review and meta-analysis. Pulmonology 2024;30(1):53-67. [doi: 10.1016/j.pulmoe.2022.01.007] [Medline: 35151622]
63. 2026 tech trends and adults 50-plus. : AARP Research; 2025 URL: https://www.aarp.org/content/dam/aarp/research/topics/technology/internet-media-devices/2026-technology-trends-older-adults.doi.10.26419-2fres.01020.001.pdf [accessed 2026-04-15]

Abbreviations

ATS/ERS: American Thoracic Society/European Respiratory Society
CAT: COPD Assessment Test
CONSORT: Consolidated Standards of Reporting Trials
COPD: chronic obstructive pulmonary disease
COPDMELODY: Music Therapy for COPD Rehabilitation
FEV_1: forced expiratory volume in 1 second
HADS: Hospital Anxiety and Depression Scale
HADS-A: Hospital Anxiety and Depression Scale-anxiety subscale
HADS-D: Hospital Anxiety and Depression Scale-depression subscale
HRQoL: health-related quality of life
IC: inspiratory capacity
ISWT: incremental shuttle walking test
LMM: linear mixed model
MCID: minimal clinically important difference
MD: mean difference
mHealth: mobile health
mITT: modified intention-to-treat
mMRC: modified Medical Research Council dyspnea scale
MT: multimodule training
PR: pulmonary rehabilitation
RCT: randomized controlled trial
RW: rhythm-guided walking
SGRQ: St. George's Respiratory Questionnaire
UC: usual care

XSL•FO
RenderX

Edited by A Mavragani; submitted 01.Aug.2025; peer-reviewed by L Houchen-Wolloff; revised version received 17.Mar.2026; accepted 24.Mar.2026; published 27.Apr.2026.

Please cite as:
Shi M, Qumu S, Bu X, Zhang J, Liu K, Wang J, Zhang Y, Lei J, Wang S, Li W, Tang X, Yan J, Peng Y, Wang C, Yang T, Huang K Smartphone App–Based Music-Facilitated Pulmonary Rehabilitation Program Integrating Rhythm-Guided Walking and Singing for Patients With Chronic Obstructive Pulmonary Disease: Multicenter Randomized Controlled Trial
J Med Internet Res 2026;28:e81707
URL: https://www.jmir.org/2026/1/e81707
doi:10.2196/81707

© Minghui Shi, Shiwei Qumu, Xiaoning Bu, Jie Zhang, Kai Liu, Jinxiang Wang, Yinan Zhang, Jieping Lei, Siyuan Wang, Wei Li, Xingyao Tang, Jisong Yan, Yaodie Peng, Chen Wang, Ting Yang, Ke Huang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Perspectives Regarding the Privacy, Security, and Confidentiality of Data Collected via mHealth Apps in Saudi Arabia: Qualitative Analysis

Nasser Alhammad[1,2], PhD; Mohannad Alajlani[3], PhD; Alaa Abd-alrazaq[4], PhD; Theodoros Arvanitis[5], PhD; Gregory Epiphaniou[3], PhD

[1]Department of Health Informatics, Saudi Electronic University, Jeddah, Saudi Arabia
[2]Institute of Digital Healthcare, Warwick Manufacturing Group (WMG), University of Warwick, Millburn House, Coventry, United Kingdom
[3]Warwick Manufacturing Group (WMG), University of Warwick, Coventry, United Kingdom
[4]AI Center for Precision Health, Weill Cornell Medicine, Doha, Qatar
[5]School of Engineering, University of Birmingham, Birmingham, United Kingdom

Corresponding Author:
Nasser Alhammad, PhD
Department of Health Informatics, Saudi Electronic University, Jeddah, Saudi Arabia

Abstract

Background: As Saudi Arabia aims to enhance the adoption and utilization of mobile health (mHealth) apps in the health care sector, it is vital to identify the potential challenges that may be faced in the process and how to address them.

Objective: This study aimed to explore patients' and stakeholders' perspectives regarding the privacy, security, and confidentiality of data collected via mHealth apps in Saudi Arabia.

Methods: A qualitative research design was used by conducting an in-depth interview with 25 participants, comprising health care workers (HCWs), patients, and mHealth app developers. The interview questions focused on awareness of mHealth apps and issues relating to data security, privacy and confidentiality, benefits of mHealth apps, challenges faced, and features that may improve the utilization of such digital health technologies. Thematic analysis was performed using NVivo software (version 12; QSR International).

Results: Patients were relatively less informed regarding events associated with data privacy and security than HCWs. Specific factors were identified to influence patient awareness with HCWs, suggesting the need for patient education and collaboration with cybersecurity organizations. Participants posited that advanced security features, user-friendly features, online consultation for emergencies, remote monitoring features, and taking patient needs into account are important facilitators of mHealth apps. Shared experiences mainly revolved around the accessibility to information and reliability of mHealth apps, while the challenges were related to poor usability, technical difficulties, data security, and data breaches. Patient autonomy, remote monitoring, self-care management, medical adherence, and time and cost-saving summed up the perceived benefits of the mHealth app.

Conclusions: These findings may assist policymakers in developing strategies to improve Saudi users' and patients' adoption of mHealth apps and address the concerns raised to benefit significantly from these advanced health care modalities.

(J Med Internet Res 2026;28:e83381) doi:10.2196/83381

KEYWORDS

awareness; data privacy; confidentiality; security; health care; patients; mHealth; mobile apps; mobile phone

Introduction

With the introduction of smartphones and efforts to improve the delivery of health care services, mobile health (mHealth) apps have witnessed a wider acceptance by the public and health care workers (HCWs) [1,2]. The outbreak of COVID-19 also catalyzed the dependence on digital health as mHealth apps were used in diverse aspects of health care delivery, such as disease management, self-monitoring, health information gathering, supervision of behavioral changes, fitness management, and reminding patients of rehabilitation and medication schedules [3,4]. mHealth apps are also helpful in managing health data, providing easy access to health records, and offering opportunities to execute mobile consultation and remote monitoring either during treatment or follow-up assessment [5].

HCWs and patients constitute the predominant stakeholders within the health care sector. mHealth apps facilitate the sharing of clinical data and information between these 2 bodies [6]. Such clinical data are available in the clinical information

XSL•FO
RenderX

systems (CISs), which are specific domains in the hospital information systems, directly linked to inpatient and outpatient care. These systems are usually connected to other subsystems within the health care facility, such as department units and laboratories [7]. In situations where CISs are the main sources of decision-making and patient management for clinicians, both the accuracy of information provided by patients and how such data are stored and secured are pertinent. Despite the significant roles of integrating mHealth apps and CISs, particularly in terms of hospital workflows, issues relating to data security, confidentiality, and privacy remain a significant challenge. A few studies have demonstrated the adverse effects of data confidentiality, privacy, security, and regulatory supervision of mHealth apps on the adoption and implementation of these digital health technologies [8].

Patients' and HCWs' awareness about data confidentiality and privacy may influence the risk of data breaches. Moreover, these end users have an obligation for the privacy and security of their data to be guaranteed [4]. This data-related relationship between HCWs and patients is contractual and should comply with data privacy regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR). Developers of mHealth apps are also encouraged to implement secure authentication and data encryption and carry out timely risk assessments to ensure data security and privacy [9,10]. The HIPAA entails the administrative, physical, and technical aspects to facilitate the privacy of security of personal health data, whereas the GDPR emphasizes the documentation of users' consent before recording their personal data and preserving their right to access, amend, delete, or restrict processing their data [10,11]. While mHealth providers are more likely to be aware of the security measures, privacy, and confidentiality related to patients' health data, patients and end users appear to be lagging in these important aspects [4,10]. Users' knowledge and awareness of these aforementioned provisions may influence their perspectives and adoption of mHealth apps.

In Saudi Arabia's context, the increasing population of smartphone users has facilitated the implementation of several mHealth apps aligning with the country's Vision 2030 goals [11,12]. This transformation to digital health management commenced in 2017 to enhance the delivery of health care services in public and private health [13,14]. However, the adoption of mHealth apps in Saudi Arabia's health care sector is yet to meet the expected benchmark, with issues relating to data privacy and security as one of the mitigating factors [11,15]. While this problem is evident from recent studies using qualitative research methods [11,15,16], the underlying reasons are not fully understood, particularly in the context of Saudi Arabia [16].

Another issue regarding existing literature on mHealth apps in Saudi Arabia is that most previous studies have focused more on health care professionals, whereas patients' perspectives have received little attention [15]. Both developers' and users' perspectives need to be balanced to ensure that mHealth apps achieve their targeted goals in enhancing health care service delivery. Qualitative approaches, such as in-depth interviews, focus group discussions, and interpretative phenomenological techniques, offer diverse opportunities to elucidate the underlying events contributing to data privacy and security issues. Such information is pertinent in addressing users' needs and developing patient-friendly mHealth apps in Saudi Arabia. This study aims to explore patients' and stakeholders' perspectives regarding the privacy, confidentiality, and security of data collected via mHealth apps in Saudi Arabia.

Methods

Study Design and Location

To achieve the aim of exploring the use and perceptions of mHealth applications among health care workers and patients, a qualitative study was conducted. Specifically, interviews were carried out to gather in-depth information from HCWs and patients using mHealth apps. The research took place in regions where hospitals feature cutting-edge facilities and currently use mHealth apps for patient management (Riyadh, Madinah, Mecca, Tabuk, Najran, Hall, Northern, Eastern, Al Jouf, Asir, Qassim, Jazan, and Al Baha). These apps are mainly categorized into 4 types: government apps, hospital, insurance, and private clinic apps [14]. Examples of these medical applications include Sehhaty, Qariboon, Health Volunteering, SFHP Riyadh, Tawuniya, Labayh, Al Mostajeeb, and Meena Health, which were introduced in the last 6 years. The Sehhaty app, also known as "My health," is a robust digital health platform that offers personal health information and promotes knowledge about healthy lifestyles and public health. It was later used for booking COVID-19 vaccine appointments [17,18]. In terms of core functionalities, most of the mHealth apps are designed for general health, such as support or wellness, care delivery, and telemedicine services. Meanwhile, platforms such as Labayh emphasize health education and mental health consultations, addressing specific areas in the health care service delivery [14,17].

Study Population

The study population and unit of analysis in this study encompass all the HCWs, users and nonusers, and developers of mHealth apps who fulfill the inclusion criteria.

Participant Selection

The study used a nonprobability sampling method, specifically purposive sampling, to select participants. Eligible participants included HCWs from hospitals offering mHealth apps for patient management in the identified provinces, patients who are either currently using these apps or nonusers, and developers of these apps. We decided to recruit nonusers because they can equally provide important insights, particularly regarding the barriers to adoption of mHealth apps. mHealth app users are generally classified as either active or passive according to their present usage status. Active users are highly engaged and use the complete provisions in mHealth apps, whereas passive users use the apps occasionally and restrict themselves to a few features [19]. Additionally, participants needed to be aged 18 years or older and willing to give both oral and written consent to take part in the research. For qualitative research, sample sizes are not definite but are still required in selecting the study participants [20]. According to Leedy and Omrod [20], the

XSL•FO
RenderX

sample size in qualitative research should range from 5 to 30. Therefore, our target was to recruit about 20 - 30 respondents, depending on the attainment of data saturation. Multidimensional approaches, including emails and social media posts, were used to identify and invite participants in this study.

Data Collection Method

To ensure systematic and organized interviews, an interview guide was developed by the principal investigator (NA) and the supervisor team. Specifically, questions in the interview session were synthesized based on an in-depth review of previous literature and discussion between the researchers. Various studies using either interviews or focus group discussions were reviewed before selecting and modifying the topics, aligning with this study context [4,21-24].

Overall, 2 interview guide sections were developed: 1 for the users and nonusers of mHealth apps and 1 for the HCWs and app developers. For the latter, the questions were further divided to suit the HCWs and app developers. The first part contains their personal sociodemographic information. The second part contained specific open-ended questions on both HCWs' and app developers' perspectives regarding users' awareness and understanding of mHealth apps. Meanwhile, HCWs were asked specific questions on their experience with the use of mHealth apps installed in patients' personal devices, knowledge of patients' health data security and privacy, benefits of mHealth apps to their patients, and potential issues that may prompt the patient to stop using the apps. App developers were probed on the security features that may increase the risk of data breaches and those that could be used to mitigate such risks. For the participating patients, similar questions were asked, except that the items were directed to them. The interview questions were presented in English and no translation to Arabic was performed.

Interviews were scheduled with the participants upon communicating with them via email or contact directories. Participants were provided with detailed information about the study and their rights, including the right to withdraw at any time without penalty. The interviews were conducted via videoconferencing platforms such as Zoom (Zoom Communications, Inc) or Google Meet, as preferred by the participants. Once the interview session commenced, participants were allowed to share their views on data privacy, confidentiality, and security issues relating to mHealth app usage. Interviews were audio-recorded, professionally transcribed, and checked for accuracy by another reviewer (NA). Data saturation was attained when no additional information was provided by the participant. It took approximately 25 minutes for each participant to complete the interview.

Ethical Considerations

The ethical approval for this research was obtained from the Biomedical and Scientific Research Ethics Committee, University of Warwick (BSREC 03/22 - 23) and the Medical Research & Ethics Committee, Ministry of Health in Saudi Arabia. The interview was designed to ensure that participants' confidentiality was maintained throughout the research. Informed consent was obtained from all participants prior to the interviews. Videos were turned off during the interview, and only the audio was recorded. Thus, no identifying details were obtained from the participants during the research process. The researcher also stated the confidentiality of all information provided by the participants. Participants were informed that they could decide to withdraw from the study at any time without any penalty. No compensation or incentives were given to the participants. All collected data were stored on a personal computer with a secure password and were accessible only to authorized parties (the supervisory team and the researcher).

Data Analysis

Data obtained from the interviews were analyzed using the NVivo software (version 12; QSR International)—a qualitative data analysis software used to organize, code, and identify themes from transcribed data. A thematic analysis was performed in which participants' responses during the interview session were analyzed in-depth and checked for any interconnection. Some of the quotations were translated from Arabic to English because a few participants responded to some questions in Arabic despite being informed that the interview would be conducted in English. We also edited the quotations lightly for clarity while ensuring that the content is maintained. Transcription was performed by a research assistant who was not involved in the interview process.

An inductive approach was used to synthesize themes directly from the transcribed data by following the procedures described by Braun and Clarke [25], which is considered a robust systematic guide for thematic analysis. First, the data were scrutinized for familiarity, followed by coding systematically and synthesizing subthemes and themes. The generated themes were cross-checked with the participants' verbatim statements, leading to the development of an initial thematic map. Next, the themes were refined and regrouped to address issues with some unsuitable codes and yield metathemes for more detailed grouping. Finally, the themes were defined and given a name. Thematic analysis was conducted by the first author (NA), and validation of the codes and themes was checked by the second author (MA). All other authors discussed the codes and themes, while discrepancies were addressed through consensus. All the experts and members of the research team discussed and agreed on the emerging themes and interrelated responses.

Results

Respondents' Demographic Characteristics

This study involved 25 participants, comprising HCWs (n=11), mHealth app developers (n=2), and users (n=12). As shown in Table 1, male participants accounted for more than two-thirds of the samples, but they were evenly distributed in terms of age groups. All the participants were married, employed, and with at least a university or higher educational qualification. All the HCWs and app developers were active users of mHealth apps. Although 3 of the 12 patients or users were passive users, they were all recruited from hospitals where they are currently seeking medical assistance. However, despite considering nonusers of mHealth apps in this study, none of the recruited participants was identified as a nonuser. Apart from HCWs and app developers, the remaining participants were patients visiting

XSL•FO
RenderX

the selected hospitals in which mHealth apps are primarily used for clinician-patient communication.

Table . Demographic characteristics of the respondents[a].

Variables	HCWs[b] and app developers	Patients/users
Sex		
Male	10 (76.9)	8 (66.7)
Female	3 (23.1)	4 (33.3)
Age (years)		
18 - 30	5 (38.4)	4 (33.3)
31 - 40	4 (30.7)	5 (41.7)
≥41 years	4 (30.7)	3 (25.0)
Marital status		
Single	0 (0.0)	0 (0.0)
Married	13 (100.0)	12 (100.0)
Educational qualification		
Primary/secondary	0 (0.0)	0 (0.0)
University/higher education	8 (61.5)	10 (83.3)
Postgraduate	5 (38.5)	2 (16.7)
Employment status		
Employed	13 (100.0)	12 (100.0)
Unemployed	0 (0.0)	0 (0.0)
Type of user		
Active	13 (100.0)	9 (75.0)
Passive	0 (0.0)	3 (25.0)
Nonuser	0 (0.0)	0 (0.0)
Currently a patient/HCW attending a hospital in Saudi Arabia		
Yes	11 (100.0)	12 (100.0)
No	0 (0.0)	0 (0.0)

[a]Values are presented as n (%).

[b]HCWs: health care workers.

Thematic Findings

The main themes synthesized from the qualitative study are the following: (1) discordant perspectives on the understanding and awareness of mHealth apps, (2) perspectives and awareness of security and privacy of mHealth apps, (3) perceived benefits of mHealth apps, (4) challenges experienced when using mHealth apps (technical difficulties, data security, and data breach), and (5) suggestions to improve the security of mHealth apps. Textbox 1 shows the theme and respective subthemes.

XSL•FO
RenderX

Textbox 1. Themes and subthemes synthesized from the qualitative analysis.

Theme 1: Discordant perspectives on the understanding and awareness of mobile health (mHealth) apps

Theme 2: Perspectives and awareness of security and privacy of mHealth apps

Subthemes

- Positive perception
- Negative perception
- Health care providers' responsibilities
- Poor knowledge and the need for patient or user education
- Collaboration with cybersecurity

Theme 3: Perceived benefits of mHealth apps

Subthemes

- Patient autonomy
- Remote monitoring
- Self-care management and medication adherence
- Time and cost-saving

Theme 4: Challenges experienced when using mHealth apps

Subthemes

- Data security and privacy issues
- Data breaches
- Technical difficulties

Theme 5: Suggestions to improve the security of mHealth apps

- Improved security features
- User-friendly features and tailoring toward patient needs
- Online consultation for emergencies
- Remote monitoring features

Discordant Perspectives on the Understanding and Awareness of mHealth Apps

The first question assessed participants' understanding and awareness of mHealth apps. Discordant perspectives were shared by respondents. Specifically, all HCWs and developers of mHealth apps demonstrated a good understanding and awareness, while patients demonstrated otherwise. Overall, HCWs depicted good mHealth app awareness and understanding while emphasizing that such a level of exposure was lacking among their patients. The main subthemes include factors influencing the usage of mHealth apps, poor awareness among users or patients, and the need for patient education. Some of the comments highlighting this theme are illustrated in the following quotations:

> *I understand that there is low awareness of mHealth apps among patients. Patients have negative attitudes toward the use of mHealth applications because of various issues, including cost, security concerns, and uncertainty regarding which apps to choose. There is a need to promote mHealth app use among patients.* [Health care professional, No. 13; Health care worker 13]

> *My patient has been using a mobile health application for the last couple of years and is well acquainted with the usability of mhealth apps.* [Health care worker 11]

Most users who reflected a good understanding and awareness of mHealth apps emphasized the role of health care providers in data collection, using the garnered data for remote health monitoring, the positive impact of COVID-19 disease on increased usage of telehealth, and general information on the app operations and functionalities.

> *These apps can provide a range of services, including helping users track their physical activity and nutrition, providing educational resources on various health topics, and offering tools to manage conditions such as diabetes or mental health.* [mHealth app user, No. 10: US 10]

On the other hand, a few users opined that they were unsure of the reason and purpose for collecting their data. Examples of

XSL•FO
RenderX

the responses reflecting users' poor understanding and awareness of mHealth apps are presented as follows:

> *I do not have a deep understanding of mobile health applications, but little that I know is that Mobile health (mHealth) applications are software programs that run on smartphones, tablets, and other mobile devices and are designed to support health and healthcare-related functions.* [US 7]

Perspectives and Awareness of Security and Privacy of mHealth Apps

Overview

Five subthemes were synthesized from the users' responses regarding their perspectives and awareness of mHealth apps' data security and privacy. These subthemes included collaboration with cybersecurity organizations, positive perception, negative perception, health care providers' responsibilities, and poor knowledge of data security and privacy.

Positive Perception

Six of the 10 users perceived that the security and privacy of their data are strictly protected by health care providers and used for the purpose they were collected. These respondents also posited that the database is secure and prevents unauthorized or third-party access to their data. Furthermore, users cited examples of the security features that prompted their positive feedback on mHealth apps, as well as relaying the actions taken by health care providers to ensure the confidentiality and security of patient data.

> *I believe to a large extent my data are secured and well-protected because I take my time to navigate through the apps and see the security features available in the app before I start using it.* [US 9]

> *The security features strongly protect users' data because the app developers provide a range of data-protective mechanisms. Moreover, the data requested is basic regarding my health condition.* [US 2]

Negative Perception

On the other hand, 4 of the 10 users opined that the security and privacy of patient data were far from being protected, thus increasing the risk of data breaches and unauthorized access. Among the issues raised by respondents are potential usage of data for nonmedical purposes, ambiguous privacy policies that make it difficult for users to understand how data privacy and security are handled, and a lack of sensitization by health care providers. Therefore, this group of respondents recommended that developers of mHealth apps and health care providers should collaborate to ensure enhanced protection of devices.

> *I am afraid of the usage of my data for other things besides the medical and communication with healthcare professionals. This is a critical time whereby data can easily leak to hackers.* [US 6]

> *It's a two-way event because data can be used to improve the delivery of healthcare services, but at the same time, unauthorised access to such data may have serious implications, especially if manipulated and used for other purposes.* [US 10]

Health Care Providers' Responsibility

This theme was synthesized mainly from HCWs' responses to the second interview question. Five of the health care professionals highlighted that it is their responsibility to enlighten patients on the available security features on their mHealth apps. They also claimed to be responsible for any data breach since they are considered the custodians of patient data.

Health care professionals also raised concerns regarding data security and confidentiality, especially challenges in managing patient data. They highlighted the limitations of the current security features in mHealth apps, which may heighten the risk of data breaches and violation of patient privacy.

> *It is our duty to ensure that patients' data are protected and used solely for medical purposes. Most of these patients are naïve and not really exposed to mHealth apps because of their backgrounds. Hence, we have to educate them when we have the opportunity to do so.* [Health care worker 5]

> *Although the app developers are responsible for designing the security and data privacy features, I believe our role as clinicians is important to mitigate any practices that may lead to data confidentiality and privacy issues. The level of access to patient clinical data needs to be moderated across levels of healthcare services.* [Health care worker 2]

Poor Knowledge and the Need for Patient Education

While some users demonstrated either positive or negative attitudes toward data security and privacy, a few participants exhibited poor knowledge of these issues. This theme was reflected in users' lack of understanding of the security features available in their devices and their role in ensuring that their data are not shared with unauthorized parties. This finding is reflected in the following statements:

> *What is more important to me is to manage my health condition. I don't think I deeply understand the provisions in the apps to address the possibility of my personal data falling into the hands of authorised persons. Seriously, I want to learn more because it's the primary duty of the clinicians and app designers to protect my data, right?* [US 8]

> *I can't even recall any information relating to the privacy and security of these apps, despite being a user for almost 2 years. I just focus on using it to book appointments and update my treatment progress.* [US 4]

As a result, health care professionals and mHealth app developers emphasized the need for more patient education. Accordingly, such interventions should be tailored toward patient needs upon identifying their shortcomings in data security and privacy. This includes proper usage of security features, strong passwords, and who they can share their data with to prevent data breaches.

XSL•FO
RenderX

We have to educate patients on how to navigate through the security features and use strong passwords to protect against unauthorised data access. This is more important at the moment because there are several websites and apps other than mHealth apps requesting patient personal data. [Health care worker 5]

Collaboration With Cybersecurity Organizations

The fifth theme uncovers the need for health care providers to collaborate with cybersecurity institutions to improve the security, privacy, and confidentiality of data collected via mHealth apps. These respondents did not relay either a positive or negative perspective toward data security and privacy but mainly recommended the approach that can be adopted by health care providers to address the issue. A similar position was shared by the stakeholders who advocated for networking and collaboration between various health care bodies to enhance data security and privacy. The following statements reflect the emerging theme:

I strongly recommend working together with mHealth app developers, given our experience and understanding of the issues raised by patients and users of these apps. We are in the best position to share the information with the developers since we communicate frequently with patients. [Health care worker 8]

It will be challenging for mHealth app developers to enhance the security features of the digital technology without our input as clinicians from the point of patient-clinician relationships. Patients don't have access to the developers of these apps; hence, they often share their concerns with other users or healthcare workers during consultation. [Health care worker 1]

Challenges Experienced When Using mHealth Apps (Technical Difficulties, Data Security, and Data Breach)

Pertinent information was gleaned regarding the challenges faced by health care professionals and patients while using mHealth apps. These barriers ranged from data security and privacy issues to data breaches and technical difficulties. Poor usability encompasses problems with navigation, understanding the features, and slow operation. These problems made some patients prefer visiting their doctors since the app was not effective in delivering mHealth services. These findings are captured in the following statement:

I had some challenges navigating through some features. Sometimes I preferred visiting the doctor since I needed to utilise the application to suit my needs effectively. When using the application, I was unsure if my data records were safe since there have been increased hacking and theft cases that could easily compromise my medical records. On rare occasions, the application could freeze, and I needed more time to get what I needed. [US7]

A similar scenario was highlighted by health care professionals, who claimed that the app could be misleading with usability, especially when notifications are delayed and during data updates. These challenges are also linked to the issues of violation of data privacy, incidents of data breach, and technical difficulties.

The mHealth app can sometimes be misleading, especially when the notifications are delayed. I have also experienced misfunctions of the app when updating data and printing some information. I have heard instances where the app stopped operating, asking for updates, which could lead to catastrophe in the case of a terminally ill patient. [Health care worker 2]

I have been approached by many patients who wanted me to teach them how to use the application since they did not know how it functions. Even after teaching some of them how to use the application, I experience other issues, such as security and privacy and inaccurate results that patients experience. Some of the applications did not function well, and they frustrated the users. [Health care worker 12]

Perceived Benefits of mHealth Apps

Overview

It is well established that mHealth apps offer several advantages and benefits, especially during and in the post–COVID-19 era. Four subthemes were identified, namely, patient autonomy, remote monitoring, self-care management and medical adherence, and time and cost-saving. These themes encompass both HCWs' and patients' responses to the interview questions. The following subsection discusses each theme.

Patient Autonomy

The term "patient autonomy" was used by 3 HCWs in this study while describing the personalization of health care access and services for patients. This theme also elucidates patient control over their data. Likewise, 1 user agreed that mHealth apps have enabled them to make personal changes to their medical data, which can be detected only by the health care provider.

Patients have a higher degree of autonomy. The applications have also helped many patients to save money and time while quickly accessing medical help. [Health care worker 5]

Mobile health applications are an essential tool that has enabled me to personalise my healthcare access. It is a time-saving tool that helps me quickly get the necessary healthcare services. I can easily book appointments and manage my prescriptions and other healthcare services. [US 6]

There is a tendency for patients to perceive that their data are secure once they experience the benefits of mHealth apps in terms of treatment and management of their health conditions. This is what I have noticed over time since we introduced these apps to clients. [Health care worker 4]

XSL•FO
RenderX

Self-Care Management and Medication Adherence

This theme was among the most recurring information provided by health care professionals (n=6) and users (n=5) of mHealth apps in this study. Apart from improving patient awareness of the functions and importance of mHealth apps, several users relayed that the technology has assisted them in adhering to various clinicians' recommendations and disease management modalities. This includes adhering to recommended diet, medication, and lifestyle activities, therefore culminating in self-care management. Most importantly, the users with this view posited that their health has improved significantly due to the self-care management features provided on the app. The following statements reflect the present theme:

> *The mHealth app has proven to be an invaluable tool in my journey towards better health. It has allowed me to adhere to a strict diet, keep track of doctors' appointments and medication schedules, and monitor my metabolism and heart rate.* [US 7]

> *The Mhealth application provided me with timely medication, which led to the improvement of my health condition. I was able to attend to my health using the application easily.* [US 9]

Remote Monitoring

The fact that mHealth apps provide the opportunity for patients and health care professionals to communicate without physical visits was mentioned as a key benefit by both parties. Stakeholders reported that this benefit contributed to the increasing trend of mHealth adoption and usage following the COVID-19 pandemic. Thus, clinicians were able to monitor and track their patients irrespective of their locations. Patients also benefit from communicating directly with clinicians without hospital visits.

> *The app has helped me to stick to certain routines, allowing efficiency. The app has also helped me to monitor patients, such as those at home, without having to visit them. The app has also allowed me to keep my record by downloading history easily without having to ask the patient.* [Health care worker 3]

> *The mHealth app has helped me to monitor the progress of the patient without having to be there. The app allows sending suggestions and prescriptions, and the recording of data more dynamically and easily.* [Health care worker 2]

> *The application kept my health in check while reducing my re-visits. I also saved time costs related to physical visits to the hospital.* [US 4]

Time-Saving and Cost-Saving

In line with the provisions for remote monitoring and self-care management, mHealth apps offered a crucial benefit by facilitating real-time information access without any physical interaction between the health care professional and patients. These events save both parties from expending time and resources associated with hospital visits. According to health care professionals, this provision of mHealth apps is crucial in the management of patients with chronic diseases and those requiring home-based therapy.

> *The application kept my health in check while reducing my re-visits. I also saved time costs related to physical visits to the hospital.* [US 4]

> *The app has also provided a convenient way for me to track my progress, and my doctor has been able to monitor my health remotely, saving me both time and money.* [US 9]

Suggestions to Improve the Security of mHealth Apps

Overview

Based on the diverse issues raised by users and stakeholders regarding the challenges faced in using mHealth apps, the respondents were probed on the features that need to be integrated into these advanced health technologies to address the identified issues. Resultantly, 4 subthemes emerged from the thematic analysis as follows: improved security features; user-friendly features and tailoring toward patient needs; online consultation for emergencies, tailoring toward patient needs; and remote monitoring features. Each of these themes is presented in the following section.

Improved Security Features

Several participants (HCWs: n=4, users: n=5, and App developers: n=1) suggested the need to add more security features to enhance the privacy, security, and confidentiality of data collected via mHealth apps. Different security options and features were mentioned by respondents, including IP address, notification, and second-step verification such as OTP, thumbprint, or facial recognition. Some respondents stated that the app notification system should be enhanced such that the notification is received by the doctor and the patient. These positions were held by patients and health care professionals as presented in the following comments.

> *The greatest threat to these apps is security and privacy. To ensure that this is checked, the app should have a change of IP address notification whenever a new device is used, seeking verification. The app should also have a second-step verification whenever accessing the app, such as a thumbprint or OTP.* [US 6]

> *Let's improve the security of operating these apps by allowing thumbprint or facial recognition as a form of password to manage these apps. The apps should also come with a printing option so the patient can print, for example, a week's data on their progress.* [Health care worker 5]

> *We are currently working on how notifications can be sent to the healthcare provider and patient, as a way of enhancing the security system and data shared between both parties.* [App developer 2]

User-Friendly Features and Tailoring Toward Patient Needs

Another important theme emerging from the analysis is to make the mHealth app user-friendly. Participants posited that

XSL•FO
RenderX

user-friendly features may redirect patients' focus to the usability of the mHealth apps rather than focusing on security and privacy issues. This suggestion was presented by 3 users and 4 HCWs who emphasized the features to encourage the usability of the app and shape the users' attitudes and perceptions toward the security and privacy of such apps. Examples of the features include font size, navigation icons, language preference, and login procedures. A few verbatim statements reflecting this theme are shown in the following quotations:

> *The features I expect to be added include making the app user-friendly. This involves personalising the app depending on the patient's characteristics, such as age, language preference, and technical literacy. This can help advance the usability of the app.* [Health care worker 4]

> *A user-friendly interface should be incorporated into the mobile health application. This would help ensure that the system is easy to use.* [US 2]

This subtheme also reflects the need to consider patient needs when introducing any new feature to the current mHealth apps used in Saudi Arabia. This position was mostly held by HCWs as they suggested tailoring the app to suit users' age, language preference, and technical expertise. Home-based health care services were also suggested to be integrated into mHealth apps to accommodate patient needs. In addition, 1 app developer recommended designing specific features to encourage self-care management among patients with chronic diseases. Respondents claimed that these features will not only improve users' interactivity and usability but also provide a medium for social support and increase patient control of their data.

> *Patients with chronic illness can have their basic pulse and BP readings monitored after integration with their Fitbit app. This feature will allow them to quickly get the assistance they need at any place and at any time. The patients can read their pulse during their regular duties and check their blood pressure.* [App Developer 1]

Online Consultation for Emergencies

Three users and 2 HCWs advocated for a feature specifically designed to enable online consultation for emergency patients. This suggestion stemmed from the challenges associated with the usability of the apps, which may be time-consuming and non–user-friendly. The corresponding comments are presented in the following quotations:

> *An online consultation will be only for emergencies for our registered patients.* [US 1]

> *I believe a function in the app should be dedicated to emergency cases, to enable faster response.* [Health care worker 5]

Remote Monitoring Features

Respondents (n=4) opined that more features are required to ensure effective remote monitoring via mHealth apps. Some respondents felt that the features are limited probably to reduce the risk of data security and privacy issues. This theme also emerged from respondents who had negative perceptions of current mHealth apps used in the country. Apart from improving data security and preventing data breaches, this group of respondents suggested better features to monitor both admitted patients and outpatients, as well as motivating patients to engage in self-care management.

> *Maybe it is because of the concerns regarding the security of these apps, but I think it does not really help for online communication with healthcare providers. The apps need to be improved for real-time tracking of outpatients and follow-up without exposing our data to unauthorised channels.* [US 3]

> *I am using the NahdiCare apps, which are mainly for monitoring care. There is no provision for detailed interaction with the clinician. Most times, I have to visit the hospital for issues that should be resolved via the app.* [US 10]

Discussion

Principal Findings

This study constitutes the first attempt to explore users' perspectives on issues relating to the privacy and security of data collected via mHealth apps in Saudi Arabia. Besides, there is limited information on how these challenges may affect the adoption and utilization of digital health technologies. In this study, we focused on the users and nonusers of mHealth apps, as well as HCWs and app developers who constitute the predominant stakeholders of these advanced digital technologies in Saudi Arabia.

Awareness and Understanding of mHealth Apps

Regarding the awareness and understanding of mHealth apps, stakeholders in the health care sector, such as HCWs and app developers, demonstrated a good understanding of issues relating to data security and privacy, whereas users and patients were perceived to lack such knowledge. Moreover, patients also demonstrated a diverse understanding of mHealth apps, which aligns with several studies reporting users' concerns about the security and privacy of sensitive data, including images of body parts and those revealing their identity [22,24,26]. A systematic review by Aljohani and Chandran [13] reported that the ease of access and operation of mHealth apps were among the reasons for hospitalized patients' concerns about unauthorized access to their health data [13]. However, some patients were not worried about these issues either because of their confidence in data safety and features to prevent data breaches, or they were completely unaware of the potential risks involved. This constitutes an important finding, given that most users in this study were highly educated, indicating possible disturbing levels of ignorance and awareness among medium to noneducated users in the general public. Users who are unaware of data privacy and security features and provisions in mHealth apps may be unable to differentiate between risky behaviors, depending on the type of private data collected by various technologies [27].

On the contrary, the understanding and awareness shown by HCWs are expected since they are more conversant with

XSL•FO
RenderX

mHealth apps. They are also more exposed and educated, thereby increasing their perception of the risks and benefits associated with digital health technologies. In order to address the low awareness level among patients, most HCWs suggested the need for patient education and enlightening them on pertinent areas. Numerous studies have taken the same stand by emphasizing brief training or education on data security and assuring data confidentiality and privacy [27-29]. Both Nurgalieva et al [30] and Lewis and Wyatt [31] recommended that patients should be trained and educated to address the external risk factors associated with the violation of users' privacy and security in mHealth apps.

Recommendations to Improve mHealth Apps' Data Privacy and Security

All the groups of participants in this study provided insights into measures to improve the privacy and security of data used in mHealth apps. Examples include features to improve data security and remote monitoring, making the app more user-friendly and tailored toward the patients' needs, and providing online consultation for emergencies. Specific security features such as 2-step verification and authentication, as well as the use of thumbprints and strong password requirements, were mentioned by health care professionals and end users. In the same vein, Alwashmi et al [32] opined for the use of an alphanumeric passcode to ensure the protection of mHealth apps rather than using a 4-digit pin, as well as wiping data from the mobile device after a specific number of failed passcode attempts. Nurgalieva et al [30] also emphasized the frequency of notifications and alerts programmed into mHealth apps.

As observed in this study, participants recommended sending notifications to both the health care provider and the patient to ensure extensive review of login attempts or data sharing. A discreet or private notification was advised to prevent any distress to users, particularly in situations where someone mistakenly views the icon. Overall, these recommendations reecho the need for users to have complete control in using their smartphone for mHealth and avoid any intrusion into their daily lives. This event coincides with the personalization and tailoring of newly added features according to patient needs, as mentioned by some respondents. Certain adjustments to the security and privacy features of mHealth apps need to be incorporated by the app developers upon considering patients' and users' demographics, as evident in previous studies [4]. However, given the low number of studies reporting the underlying factors influencing patients' views on data privacy, confidentiality, and security issues in mHealth apps, more research is needed to elucidate the relationships.

Another key recommendation, especially by users, was the need for real-time online consultations for emergency cases. This suggestion stems from the challenges faced in navigating through mHealth apps, which may be time-consuming and ineffective for health care service delivery. Therefore, a specific platform on mHealth apps to facilitate easy and prompt communication with the health care provider's emergency unit was advocated. Such an intervention will come with several benefits and challenges, particularly in terms of educating users. Regarding the benefits and experience garnered from using mHealth apps, participants mentioned diverse ways in which they have benefited from mHealth apps. It is well established that mHealth apps offer several advantages and benefits, especially during and in the post–COVID-19 era. The subthemes included patient autonomy, remote monitoring, self-care management and medical adherence, and time and cost-saving.

Patient autonomy refers to equipping patients or users, as the case may be, with substantial control of their data. Several users and health care professionals highlighted this view either by using personalization of patient data or by giving them unlimited access. On the other hand, remote monitoring entails online interaction and examination between patients and health care professionals. Consequently, patients were motivated to engage in self-care management and medication adherence, ranging from assessing physical parameters to restricting diet. These events contributed significantly to efficient disease progress monitoring, good communication between patients and health care professionals, and better patient outcomes. Our results corroborate the reports from previous studies, whereby improved health status [33] and better patient-health care professional relationships [34,35] were stated as the main benefits of using mHealth apps.

In terms of participants' experiences with mHealth, although encouraging comments were highlighted by several users, important barriers were mentioned that require further discussion. These barriers ranged from poor usability to data security and privacy issues, data breaches, and technical difficulties. Concerns about data privacy and confidentiality may discourage patients and end users from using mHealth apps [35]. Likewise, the low levels of privacy and security of patient data contributed significantly to the low adoption rate among the intended users [13]. Past reviews also reported security incidents associated with mHealth apps, such as vulnerabilities to data breaches and malware attacks. The concerns raised by the participants are plausible based on the extensive usage of mobile devices with location sensors in Saudi Arabia, which has facilitated easy access to location-based services [33]. Given the extensive privacy and security issues raised by the stakeholders in this study, cyberspace security can be further explored to improve the security of mHealth environments [4].

Surprisingly, the challenges faced appeared not to deter participants from using the technology, as all the respondents who shared their concerns were active users of mHealth apps. A recent study conducted among mature adults also shared a similar view, as none of the participants stopped using the technology despite their data privacy concerns [27]. This aligns with the widely discussed privacy paradox [36], whereby people engage with technology even when they acknowledge its potential privacy loss issues. Research involving social media, e-commerce, and mobile apps has also demonstrated this phenomenon [37,38]. This paradoxical behavior is likely driven by users' perception that the immediate benefits of using mHealth apps outweigh their potential or hypothetical risks [27].

Strengths and Limitations

This study represents the first attempt to qualitatively explore patients' and stakeholders' perspectives regarding the privacy

XSL•FO
RenderX

and security of data collected via mHealth apps in Saudi Arabia. Our findings offer insights on possible strategies to improve end users' and patients' adoption of mHealth apps and addressing concerns relating to data privacy and security. In terms of sampling, purposive sampling was used to recruit participants with diverse backgrounds (age, gender, app experience, and multiple hospitals). This approach aligns with qualitative research goals of capturing variation and depth. Moreover, qualitative studies are not designed to achieve representativeness or collect data from a representative sample.

Nevertheless, this study has several limitations that need to be acknowledged and considered when extrapolating the findings. We developed the interview guide and conducted the interviews in English, and only a slight translation was performed during the data transcription. Based on this criterion, most of the respondents were highly educated. As for the HCWs and app developers, this procedure is unlikely to introduce a selection bias since these professionals are expected to hold a higher educational level. However, most users and patients also had at least a university degree, indicating that their views on mHealth and privacy may not be representative of the general Saudi population. Nevertheless, it is interesting to note that several users were still unaware and had a limited understanding of privacy and security of mHealth apps—possibly reflecting a more serious problem in the general population.

Despite our plan to include nonusers in the study, all the participants were either active or passive users of mHealth apps. The patients recruited were from hospitals in which mHealth apps are actively used for consultation and follow-up, making it difficult to identify those who are yet to adopt these digital health apps. Therefore, our findings present limited perspectives on barriers among nonusers. This limitation narrows the study scope and insights. Future studies may consider recruiting nonusers of mHealth apps from the Saudi community or hospitals where patients are not compelled to use such apps.

Our interview guide was limited in terms of exploring security issues, which led to the superficial technical aspects raised by the app developers and HCWs. While patients were less aware of these technical aspects, as evidenced from the interviews, app developers and HCWs occasionally mentioned the need for more comprehensive access controls, whereas other critical aspects, such as data storage, encryption, and server-side protection, did not emerge from their responses. Apart from the limited questions that may trigger deeper insights into security-related issues, only 2 app developers were recruited, and HCWs may have limited technical awareness while focusing on health-related data and care delivery when using mHealth apps. Future studies may bridge this research gap by recruiting a larger number of stakeholders comprising HCWs and app developers who are technologically savvy for a nuanced understanding of technical and security issues associated with these digital health technologies.

Research and Practical Implications

The results from this study have important implications for mHealth app stakeholders in Saudi Arabia. Both HCWs and developers of mHealth apps have vital roles in addressing the challenges identified in this study, particularly issues relating to data privacy, confidentiality, and security that may affect the adoption and usage of digital health technologies.

Since HCWs acknowledge patients' low awareness and understanding of mHealth apps and the data collected in the process, they can assist by educating patients to address such knowledge deficits. Patients could be briefed during routine consultations by highlighting the need for recording sensitive data and the purpose of collecting such information. Regarding unauthorized third-party access and potential data breaches, users may benefit from information on the privacy requirements and meeting the standards set by the GDPR and HIPAA, which are designed to ensure that such privacy issues and data breaches are prevented.

Another important aspect is training and educating end users on the existing features and measures to ensure that their data remain secure and confidential. Nevertheless, educational interventions need to be effectively tailored toward the user's level of awareness and understanding. As evident from this study, users' less concern about their data privacy and security may arise from being completely ignorant of these potential challenges, or they may perceive that the benefits of using mHealth outweigh the potential risks.

mHealth app developers can also benefit from the research findings, as the concerns raised by patients and end users may culminate in low adoption and usage of mHealth apps [18]. While users are expected to express concerns regarding data privacy and security, it remains unclear whether such concerns influence their behavior and actual use of such apps. This gray area needs more understanding and research. mHealth app developers, particularly in Saudi Arabia, have the opportunity to digest the issues raised by users and strategize on how to improve their current security and privacy measures. Such strategies should be designed and implemented to meet the requirements of HIPAA.

Finally, policymakers can use the research findings in identifying the areas to focus on when developing policies and strategies to facilitate the wide adoption of mHealth apps and achieve the desired outcomes. Accordingly, policymakers can emphasize features to enhance better patient–health care professional relationships, trust, and personal data protection and ensure that health status or medical conditions are not disclosed, as well as brief training or education on data security. Although these points can be relayed to users, trust is a feature that has to be earned, which has to be driven by HCWs and the government.

Acknowledgments

Generative artificial intelligence was not used to draft any portion of this manuscript.

XSL•FO
RenderX

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The full data extraction sheet is available from the corresponding author upon request.

Authors' Contributions

Conceptualization: NA, TA, GE (equal)
Formal analysis: NA (lead), MA, AA (supporting)
Investigation: NA (lead), MA, AA (supporting)
Methodology: NA, TA, GE (equal)
Data curation: NA (lead), MA, AA (supporting)
Project administration: NA (lead), TA, GE, MA (supporting)
Writing – original draft: NA (lead), MA, AA (supporting)
Writing – review & editing: NA, TA, GE, MA, AA
Supervision: TA, GE (equal)
Funding acquisition: NA, TA, GE (equal)

Conflicts of Interest

None declared.

References

1. Chen J, Lieffers J, Bauman A, Hanning R, Allman-Farinelli M. The use of smartphone health apps and other mobile health (mHealth) technologies in dietetic practice: a three country study. J Hum Nutr Diet 2017 Aug;30(4):439-452. [doi: 10.1111/jhn.12446] [Medline: 28116773]
2. Hussain M, Zaidan AA, Zidan BB, et al. Conceptual framework for the security of mobile health applications on Android platform. Telematics Inform 2018 Aug;35(5):1335-1354. [doi: 10.1016/j.tele.2018.03.005]
3. Fagherazzi G, Goetzinger C, Rashid MA, Aguayo GA, Huiart L. Digital health strategies to fight COVID-19 worldwide: challenges, recommendations, and a call for papers. J Med Internet Res 2020 Jun 16;22(6):e19284. [doi: 10.2196/19284] [Medline: 32501804]
4. Zhou L, Bao J, Watzlaf V, Parmanto B. Barriers to and facilitators of the use of mobile health apps from a security perspective: mixed-methods study. JMIR Mhealth Uhealth 2019 Apr 16;7(4):e11223. [doi: 10.2196/11223] [Medline: 30990458]
5. Kao CK, Liebovitz DM. Consumer mobile health apps: current state, barriers, and future directions. PM R 2017 May;9(5S):S106-S115. [doi: 10.1016/j.pmrj.2017.02.018] [Medline: 28527495]
6. Li J, Silvera-Tawil D, Varnfield M, Hussain MS, Math V. Users' perceptions toward mHealth technologies for health and well-being monitoring in pregnancy care: qualitative interview study. JMIR Form Res 2021 Dec 2;5(12):e28628. [doi: 10.2196/28628] [Medline: 34860665]
7. S. Bhuyan S, Kim H, Isehunwa OO, et al. Privacy and security issues in mobile health: current research and future directions. Health Policy Technol 2017 Jun;6(2):188-191. [doi: 10.1016/j.hlpt.2017.01.004]
8. Sampat BH, Prabhakar B. Privacy risks and security threats in mHealth apps. J Int Technol Inform Manage 2017;26(4):126-153. [doi: 10.58729/1941-6679.1353]
9. Asiri E, Asiri H, Househ M. Exploring the concepts of privacy and the sharing of sensitive health information. In: Integrating Information Technology and Management for Quality of Care: IOS Press; 2014:161-164.
10. Esposito M, Minutolo A, Megna R, Forastiere M, Magliulo M, De Pietro G. A smart mobile, self-configuring, context-aware architecture for personal health monitoring. Eng Appl Artif Intell 2018 Jan;67:136-156. [doi: 10.1016/j.engappai.2017.09.019]
11. Alanzi T. mHealth for diabetes self-management in the Kingdom of Saudi Arabia: barriers and solutions. J Multidiscip Healthc 2018;11:535-546. [doi: 10.2147/JMDH.S174198] [Medline: 30349285]
12. Young Y, Alharthy A, Hosler AS. Transformation of Saudi Arabia's health system and its impact on population health: what can the USA learn? Saudi J Health Syst Res 2021 Oct 5;1(3):93-102. [doi: 10.1159/000517488]
13. Aljohani N, Chandran D. The adoption of mobile health applications by patients in developing countries: a systematic review. Int J Adv Comput Sci Appl 2021;12(4):4. [doi: 10.14569/IJACSA.2021.0120403]
14. Alharthi SA. mHealth Applications in Saudi Arabia: current features and future opportunities. Healthcare (Basel) 2025 Jun 11;13(12):1392. [doi: 10.3390/healthcare13121392] [Medline: 40565422]
15. Aljedaani B, Ahmad A, Zahedi M, Babar MA. End-users' knowledge and perception about security of clinical mobile health apps: a case study with two Saudi Arabian mHealth providers. J Syst Softw 2023 Jan;195:111519. [doi: 10.1016/j.jss.2022.111519]

XSL•FO
RenderX

16. Kassem LM, Alhabib B, Alzunaydi K, Farooqui M. Understanding patient needs regarding adverse drug reaction reporting smartphone applications: a qualitative insight from Saudi Arabia. Int J Environ Res Public Health 2021 Apr 7;18(8):3862. [doi: 10.3390/ijerph18083862] [Medline: 33917014]
17. Thapa S, Nielsen JB, Aldahmash AM, Qadri FR, Leppin A. Willingness to use digital health tools in patient care among health care professionals and students at a university hospital in Saudi Arabia: quantitative cross-sectional survey. JMIR Med Educ 2021 Feb 19;7(1):e18590. [doi: 10.2196/18590] [Medline: 33605896]
18. Alzahrani A, Gay V, Alturki R. Exploring Saudi individuals' perspectives and needs to design a hypertension management mobile technology solution: qualitative study. Int J Environ Res Public Health 2022 Oct 10;19(19):12956. [doi: 10.3390/ijerph191912956] [Medline: 36232254]
19. Wang Y, Wu T, Chen Z. Active usage of mobile health applications: cross-sectional study. J Med Internet Res 2021 Dec 22;23(12):e25330. [doi: 10.2196/25330] [Medline: 34941545]
20. Leedy PD, Ormrod JE. Practical Research: Planning and Design, 9th edition: Pearson Education, Inc, Upper Saddle River; 2010:67.
21. Dhagarra D, Goswami M, Kumar G. Impact of trust and privacy concerns on technology acceptance in healthcare: an Indian perspective. Int J Med Inform 2020 Sep;141:104164. [doi: 10.1016/j.ijmedinf.2020.104164] [Medline: 32593847]
22. Liu Y, Lu X, Zhao G, Li C, Shi J. Adoption of mobile health services using the unified theory of acceptance and use of technology model: self-efficacy and privacy concerns. Front Psychol 2022;13:944976. [doi: 10.3389/fpsyg.2022.944976]
23. Natsiavas P, Kakalou C, Votis K, Tzovaras D, Koutkias V, et al. Citizen perspectives on cross-border eHealth data exchange: a European survey. In: Ohno-Machado L, Séroussi B, editors. MEDINFO 2019: Health and Wellbeing e-Networks for All: IOS Press; 2019:719-723. [doi: 10.3233/SHTI190317]
24. Dang Y, Guo S, Guo X, Wang M, Xie K. Privacy concerns about health information disclosure in mobile health: questionnaire study investigating the moderation effect of social support. JMIR Mhealth Uhealth 2021 Feb 8;9(2):e19594. [doi: 10.2196/19594] [Medline: 33555266]
25. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
26. Atienza AA, Zarcadoolas C, Vaughon W, et al. Consumer attitudes and perceptions on mHealth privacy and security: findings from a mixed-methods study. J Health Commun 2015;20(6):673-679. [doi: 10.1080/10810730.2015.1018560] [Medline: 25868685]
27. Schroeder T, Haug M, Gewald H. Data privacy concerns using mHealth apps and smart speakers: comparative interview study among mature adults. JMIR Form Res 2022 Jun 13;6(6):e28025. [doi: 10.2196/28025] [Medline: 35699993]
28. Fernández-Alemán JL, Señor IC, Lozoya P, Toval A. Security and privacy in electronic health records: a systematic literature review. J Biomed Inform 2013 Jun;46(3):541-562. [doi: 10.1016/j.jbi.2012.12.003] [Medline: 23305810]
29. Zhou L, Parmanto B, Alfikri Z, Bao J. A mobile app for assisting users to make informed selections in security settings for protecting personal health data: development and feasibility study. JMIR Mhealth Uhealth 2018 Dec 11;6(12):e11210. [doi: 10.2196/11210] [Medline: 30538088]
30. Nurgalieva L, O'callaghan D, Doherty G. Security and privacy of mHealth applications: a scoping review. IEEE Access 2016;4:10433-10446. [doi: 10.1109/ACCESS.2016.2637622]
31. Lewis TL, Wyatt JC. mHealth and mobile medical apps: a framework to assess risk and promote safer use. J Med Internet Res 2014 Sep 15;16(9):e210. [doi: 10.2196/jmir.3133] [Medline: 25223398]
32. Alwashmi MF, Fitzpatrick B, Farrell J, et al. Perceptions of patients regarding mobile health interventions for the management of chronic obstructive pulmonary disease: mixed methods study. JMIR Mhealth Uhealth 2020 Jul 23;8(7):e17409. [doi: 10.2196/17409] [Medline: 32706697]
33. Zhang D, Lim J, Zhou L, Dahl AA. Breaking the data value-privacy paradox in mobile mental health systems through user-centered privacy protection: a web-based survey study. JMIR Ment Health 2021 Dec 24;8(12):e31633. [doi: 10.2196/31633] [Medline: 34951604]
34. Kayyali R, Peletidi A, Ismail M, Hashim Z, Bandeira P, Bonnah J. Awareness and use of mHealth apps: a study from England. Pharmacy (Basel) 2017 Jun 14;5(2):33. [doi: 10.3390/pharmacy5020033] [Medline: 28970445]
35. Korpershoek YJG, Vervoort S, Trappenburg JCA, Schuurmans MJ. Perceptions of patients with chronic obstructive pulmonary disease and their health care providers towards using mHealth for self-management of exacerbations: a qualitative study. BMC Health Serv Res 2018 Oct 4;18(1):757. [doi: 10.1186/s12913-018-3545-4] [Medline: 30286761]
36. Norberg PA, Horne DR, Horne DA. The privacy paradox: personal information disclosure intentions versus behaviors. J Consum Aff 2007 Jun;41(1):100-126. [doi: 10.1111/j.1745-6606.2006.00070.x]
37. Sutanto J, Palme E, Tan CH, Phang CW. Addressing the personalization–privacy paradox: an empirical assessment from a field experiment on smartphone users1. MIS Q 2013 Dec 1;37(4):1141-1164. [doi: 10.25300/MISQ/2013/37.4.07]
38. Kokolakis S. Privacy attitudes and privacy behaviour: a review of current research on the privacy paradox phenomenon. Comput Secur 2017 Jan;64:122-134. [doi: 10.1016/j.cose.2015.07.002]

XSL•FO
RenderX

Abbreviations

CIS: clinical information system
GDPR: General Data Protection Regulation
HCW: health care worker
HIPAA: Health Insurance Portability and Accountability Act
mHealth: mobile health

Edited by A Stone; submitted 01.Sep.2025; peer-reviewed by A Amanna, A Bustam, M Chakit; revised version received 14.Jan.2026; accepted 15.Jan.2026; published 23.Apr.2026.

Please cite as:
Alhammad N, Alajlani M, Abd-alrazaq A, Arvanitis T, Epiphaniou G
Perspectives Regarding the Privacy, Security, and Confidentiality of Data Collected via mHealth Apps in Saudi Arabia: Qualitative Analysis
J Med Internet Res 2026;28:e83381
URL: https://www.jmir.org/2026/1/e83381
doi:10.2196/83381

© Nasser Alhammad, Mohannad Alajlani, Alaa Abd-alrazaq, Theodoros Arvanitis, Gregory Epiphaniou. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Gamified Mobile Health Intervention to Promote Physical Activity, Executive Function, and Mental Health in College Students: Randomized Controlled Trial

Ming Yang[1,2], PhD; Yingjun Guo[3], MEd; Zhen Li[4], PhD; Lei Jiang[1], PhD; Yanan Gao[4], PhD; Sen Yang[1], PhD; Zhixiong Zhou[1], PhD

[1]Institute of Artificial Intelligence in Sports, Capital University of Physical Education and Sports, No.11, North Third Ring West Road, Haidian District, Beijing, China

[2]School of Physical Education, Yantai University, Yantai, Shandong, China

[3]Beijing Sport University Press, Beijing Sport University, Beijing, Beijing, China

[4]School of Physical Education and Sport Science, Fujian Normal University, Fuzhou, Fujian, China

Corresponding Author:
Zhixiong Zhou, PhD
Institute of Artificial Intelligence in Sports, Capital University of Physical Education and Sports, No.11, North Third Ring West Road, Haidian District, Beijing, China

Abstract

Background: College students commonly experience suboptimal health conditions, including insufficient physical activity (PA), excessive body weight, and declining physical fitness. Traditional interventions face low adherence, while gamified mobile health (mHealth) programs may improve engagement and outcomes.

Objective: This study aimed to evaluate the feasibility and effectiveness of a novel gamified, incentive-based mHealth intervention on primary outcomes (PA and adherence) and secondary outcomes (physical fitness, body composition, executive function [EF], and mental health).

Methods: A 2-arm parallel-group randomized controlled trial (RCT) was conducted in 2025 at Yantai University with 160 college students (18 - 25 years; BMI 18.5 - 30.0) who were randomized 1:1 (computer-generated, sex-stratified blocks of 4; concealed allocation) to the intervention group (IG) or control group (CG; n=80 each); major exclusions were contraindications to exercise, severe physical/mental illness, recent PA interventions, or psychotropic medication use. Both used the same fitness watch–app system and identical PA targets (≥150 min moderate-to-vigorous physical activity [MVPA] per week or ≥900 metabolic equivalent-minutes [MET-min] per week); IG additionally received team-based gamification (competition, points/leaderboards, feedback, and rewards), while CG received monitoring only. PA and adherence were monitored throughout the 8-week intervention; other outcomes were assessed at baseline and 8 weeks (fitness, body composition, EF, and mental health). Open-label with blinded outcome assessors/analysts; intention-to-treat (ITT) with multiple imputation.

Results: At 8 weeks, data were available for 154 participants (IG 78; CG 76); all 160 were analyzed per ITT. Compared to the CG, the IG demonstrated significantly higher mean levels in all primary PA outcomes over 8 weeks (daily steps: mean 10,356, SD 1245 versus 8242, SD 1087; Δ=2114; d=1.81, 95% CI 1.44 - 2.18; P<.001; daily MVPA: mean 71, SD 15 versus 43, SD 12 min; Δ=28 min; d=2.06, 95% CI 1.68 - 2.45; P<.001; and weekly MET-min: mean 1650, SD 310 versus 1340, SD 285; Δ=310; d=1.04, 95% CI 0.71 - 1.37; P<.001). Adherence was also higher in the IG (both P<.001). Key secondary outcomes, including skeletal muscle mass (Δ +0.54 kg), body fat percentage (Δ −0.46 percentage points), 800/1000 m run time (Δ −6.5 s), and depressive symptoms (Δ −2.0 points), significantly favored the IG (all P<.001). No serious adverse events occurred.

Conclusions: This RCT demonstrated that integrating gamification into an mHealth platform significantly enhanced PA, intervention adherence, and selected health outcomes among college students. The key innovation lies in the trial design, which held the technology platform and activity goals constant across groups, thereby allowing a clearer estimate of the incremental contribution of gamification. Unlike many existing studies, this approach reduces confounding from co-occurring intervention components in between-group comparisons. It therefore provides more comparable and methodologically robust evidence on the specific efficacy of gamification in mHealth interventions. For real-world implementation, this model leverages the existing smartphone and wearable ecosystem, offering a low-cost, scalable strategy for university health-promotion programs.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2500108950; https://www.chictr.org.cn/showproj.html?proj=285460

(J Med Internet Res 2026;28:e82769) doi:10.2196/82769

XSL•FO
RenderX

KEYWORDS

mHealth; eHealth; gamification; physical fitness; psychological well-being; cognitive function; adherence

Introduction

Physical inactivity has become a prevalent issue among today's college student population, posing a significant threat to both their physical and mental well-being as well as their overall quality of life. Recent national estimates indicate that the combined prevalence of overweight and obesity among Chinese adults aged ≥18 years exceeds 50% [1], with college students being particularly affected. This cluster of risks, including insufficient physical activity (PA), imbalanced diet, excess weight, poor sleep, and psychological distress, not only undermines individual development but also poses challenges to national public health policies, such as the "Healthy China 2030" and "Weight Management Year" campaigns [2,3].

Traditional exercise programs often underperform because they are monotonous, provide delayed feedback, and lack sustained motivational scaffolding, which collectively erode adherence and limit scalability [4]. By contrast, mobile health (mHealth) leverages smartphones, wearables, and connected data to enable real-time, personalized monitoring and feedback on activity, heart rate, sleep, and related indicators [5]. Recent reviews and empirical studies have confirmed that mHealth interventions can produce small-to-moderate improvements in PA and related health behaviors [6]. In parallel, gamification is the purposeful use of points, badges, leaderboards, challenges, and achievement systems in nongame contexts and has gained traction in health promotion for strengthening motivation and social engagement [7]. However, evidence for the impact of gamification is mixed: while some studies report significant benefits, others show minimal effects, highlighting the need for further research on the mechanisms and optimal design of gamified interventions [8].

Self-determination theory (SDT) provides a framework for designing more effective interventions. SDT posits that satisfying the psychological needs for autonomy, competence, and relatedness supports motivation internalization and sustained behavior change. In our design, autonomy was supported via self-selection and personalized goals, competence via real-time feedback and progress dashboards, and relatedness via team-based collaboration and competition [9]. Similarly, the theory of planned behavior (TPB) contends that behavior is driven by intention, which is shaped by attitude, subjective norms, and perceived behavioral control [10].

In our gamification context, these determinants were strengthened as follows: positive outcome expectancies aimed to improve attitudes; peer norms embedded in teams and leaderboards helped establish subjective norms; and clear weekly targets were designed to enhance participants' perceived behavioral control [11].

Recent studies and the World Health Organization (WHO) Global Action Plan on Physical Activity 2018‑2030 both emphasize that such theory-driven, cross-sectoral approaches are essential for promoting sustainable PA [12]. However, there remains a scarcity of research that systematically integrates mHealth with gamification under a combined SDT-TPB framework and rigorously evaluates its efficacy among college students. Furthermore, many existing studies rely on self-reported outcomes or focus on a narrow set of physiological measures, lacking comprehensive evaluation across physical, cognitive, and mental health domains within a single trial [13].

Integrating gamification with mHealth, therefore, offers a promising approach that combines real-time, data-driven monitoring with motivational and socially interactive elements to transform health behaviors into a more visualized, task-oriented, and engaging process. Within this framework, TPB helps explain the formation of behavioral motives and intentions, whereas SDT elucidates how short-term motivation is converted into long-term engagement [14]. This synergy is expected not only to promote PA but also to support mental well-being and, via goal setting and self-monitoring demands, to engage core domains of executive function [15]. College students, who have high smartphone penetration and digital literacy, are therefore an ideal population for such an approach [16].

Therefore, the aim of this study was to evaluate the feasibility and effectiveness of a novel, theory-based gamified mHealth intervention incorporating an incentive mechanism. This intervention integrates real-time monitoring, team-based competition, and personalized feedback and was comprehensively evaluated using a randomized controlled trial (RCT) design [17]. We prespecified two primary outcomes, (1) PA levels and (2) intervention adherence, to capture the core efficacy of the model. Secondary outcomes included physical fitness, body composition, mental health, and executive function (EF) to further assess the potential broader benefits for physical and psychological well-being. We hypothesized that, relative to control, participants receiving the gamified intervention would demonstrate higher PA levels and better adherence after the intervention and would show more favorable changes across secondary outcomes [18]. By addressing these objectives, this study aims to provide a scientific basis and practical strategies for college health promotion in the context of the "Healthy China" initiative and contributes to the innovation of health education systems in higher education institutions.

Methods

Trial Design

This study used a 2-arm, parallel-group (RCT) design to evaluate the feasibility and effectiveness of a gamified, competition-based intervention using mHealth technology (intervention group [IG]) versus a conventional intervention approach (control group [CG]) in promoting physical and mental health among college students aged 18‑25 years. The formal intervention period lasted for 8 weeks. Baseline assessments (T0) were conducted one week prior to the intervention, and postintervention assessments (T1) were carried out one week after the intervention concluded.

XSL•FO
RenderX

Prior to the formal intervention, all participants underwent a 1-week preintervention phase, during which the accuracy of the wearable smartwatches was validated (tolerance ≤3%), and participants were familiarized with the "Shouti Fitness" app (Tianbao Zhu) testing procedures and measurement protocols.

The study was designed and reported in accordance with the CONSORT (Consolidated Standards Of Reporting Trials) 2025 statement for randomized trials, the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth; Checklist 1) extension for web- and mobile-based health interventions, and the CONSORT extension for abstracts [19-21].

Trial Setting and Eligibility Criteria

Participants were recruited between February 04 and 10, 2025, at Yantai University, Shandong Province, through campus announcements, social media platforms, and health education lectures. A total of 224 volunteers were initially screened. Eligibility criteria were the following: full-time college students aged 18 - 25 years owning a smartphone; BMI between 18.5 and 30.0; and willingness to provide written informed consent. Exclusion criteria included the presence of severe physical or mental illnesses or contraindications to exercise; participation in other PA interventions within the past 3 months; and the use of psychotropic medications at the time of this writing. After screening, 160 eligible participants were enrolled.

Eligibility Criteria for Sites and Individuals Delivering the Interventions

No eligibility criteria were prespecified for individuals delivering the interventions. The interventions were primarily technology-mediated and self-administered by participants. The role of the research team was limited to initial instruction, technical support, and data monitoring, rather than direct delivery of the exercise intervention itself.

Intervention and Comparator

The intervention system comprised the following hardware and software components.

- Hardware: participants wore fitness watches equipped with high-precision accelerometers to monitor key metrics such as step count (Figure 1A-1B), duration of moderate-to-vigorous physical activity (MVPA), and metabolic equivalent (MET).
- Software: a customized mobile app, the "Shouti Fitness" app (compatible with Android and iOS; Figure 1C), was employed. The app features AI-driven exercise guidance, personalized exercise prescription delivery, real-time feedback, and synchronization of PA data, supporting both intervention implementation and adherence monitoring (Figure 1D).

XSL•FO
RenderX

Figure 1. Technical platform used in this study. (**A**) High-precision fitness watch. (**B**) Data transmission station. (**C**) “Shouti Fitness” App. (**D**) Health behavior monitoring system. (**E**) InBody H30 body composition analyzer. (**F**) NIH Toolbox cognitive function assessment tool on the iPad platform.

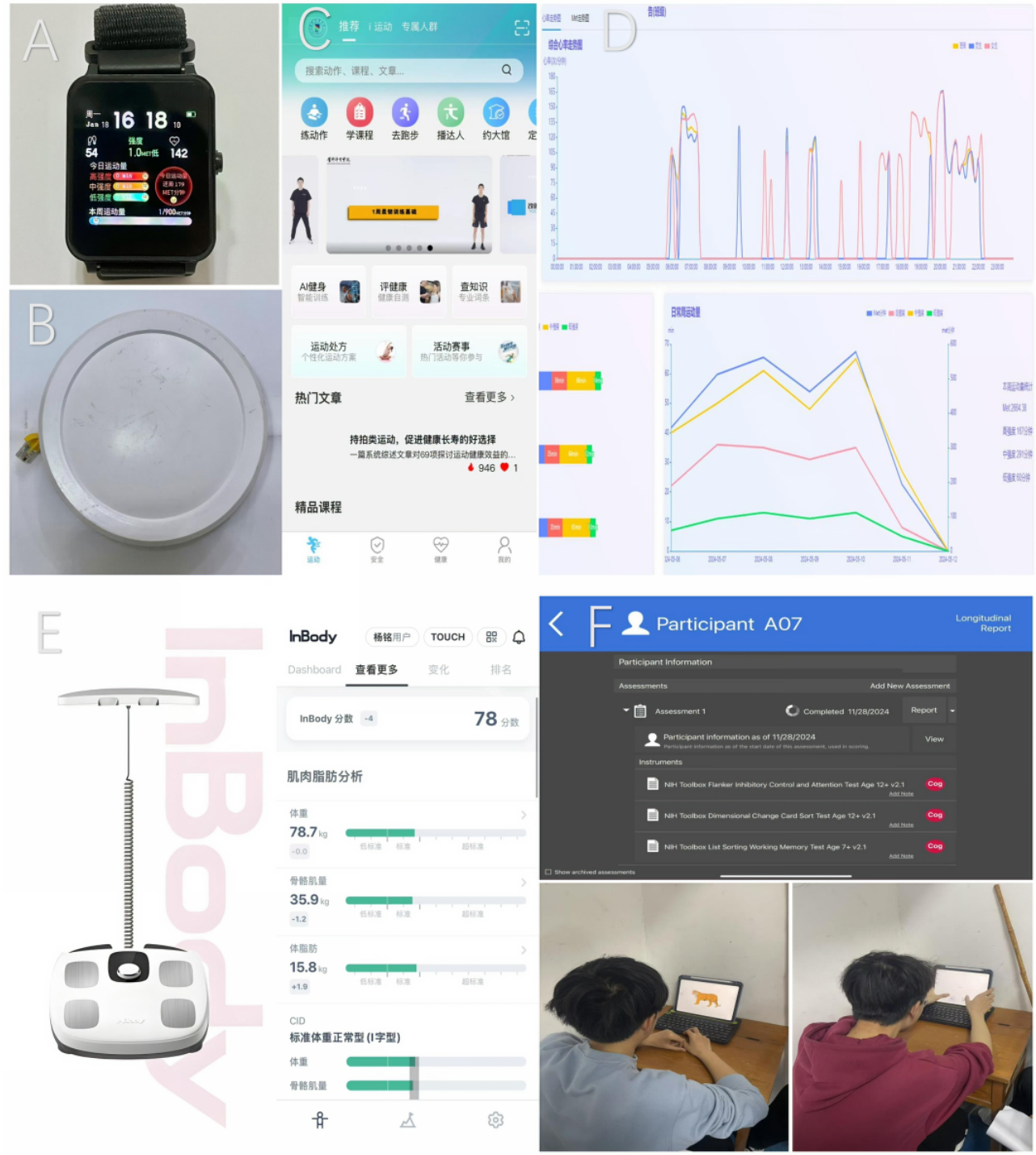

All participants were instructed to download and register on the “Shouti Fitness” app and were provided with a study-issued fitness watch. They then completed a 1-week familiarization phase (February 17 - 23, 2025) to learn the procedures and adapt to the device. The formal intervention ran for 8 weeks, from March 3 to April 27, 2025 (Figure 2).

XSL•FO
RenderX

Figure 2. Workflow diagram of the intervention mechanism used in this study. Overview of the 8-week randomized controlled trial (RCT) workflow in college students, including T0/T1 assessments, physical activity targets, and intervention group-only gamified components.

XSL•FO
RenderX

During the intervention, the fitness watches estimated exercise intensity and energy expenditure (metabolic equivalent-minutes [MET-min]) using a convolutional neural network-long short-term memory (CNN-LSTM) model and recorded durations of MVPA [22]. Participants were required to complete daily recommended PA targets (≥30 minutes of MVPA or ≥180 MET-min), with weekly goals set at ≥150 minutes of MVPA or ≥900 MET-min [23]. Data synchronization occurred twice daily at fixed data stations (07:30 AM and 11:30 PM), with activity data automatically uploaded to the exercise behavior monitoring system.

In IG (n=80) participants were assigned to 4 teams, 2 single-sex teams (1 all-male or 1 all-female), and 2 mixed-sex teams (10 males and 10 females each). Beyond meeting the baseline activity goals, a team-based competition was implemented; weekly rankings of teams based on average activity levels awarded points as follows. 1st place: 40 points, 2nd: 30 points, 3rd: 20 points, and 4th: 10 points. Additionally, the top 10% of individuals within each team (overall top 8 performers) contributed an extra 20 points to their team. Real-time activity progress bars (MVPA duration and MET-min) were displayed via the fitness watch and monitoring platform. Weekly midweek updates announced leading teams and individual champions every Wednesday, with complete team and individual leaderboards published each Sunday. Feedback and incentives included cumulative volunteer hours, honorary certificates, academic credit coordination, and material rewards of 400 Chinese yuan (Renminbi; approximately US$55.53 or €49.31, based on the exchange rate in April 2025) for individuals ranking in the top 10%.

CG (n=80) received an active, technology-based PA program without any gamification or social competition. Participants used the same technical platform, fitness watches, and app system as the IG and received identical daily and weekly personalized activity goal prompts. This design controlled for the nonspecific effects of platform use, goal setting, and researcher attention. The key distinction was the absence of all gamification and team-based competition elements, including team assignments; the point-based scoring/ranking system; real-time progress bars; midweek updates; and weekly leaderboards; as well as any performance-based feedback, incentives, or rewards.

Outcomes

The outcome measures were prespecified as primary and secondary. Primary outcomes directly assess the core efficacy of the mHealth-gamification intervention; secondary outcomes examine potential downstream health benefits.

Primary Outcomes

The primary outcomes were PA levels and intervention adherence. These indicators are selected to test the core hypothesis of this study, that “mHealth+gamification” intervention model would effectively increase participants’ PA and adherence.

- PA levels: daily steps, daily minutes of MVPA, and weekly MET-min were derived from the self-developed wearable device and a companion sports application. Intensity classification followed contemporary MET cut-points (moderate 3.0 - 5.9 METs; vigorous ≥6.0 METs), and weekly behavioral targets were benchmarked to the 2020 WHO adult guidelines (≥150 - 300 min per week of moderate activity, or ≥75 - 150 min per week of vigorous activity, or an equivalent combination) [23].
- Adherence: calculated from the health behavior monitoring system, including daily MVPA≥30 min or MET-min≥180; weekly MVPA≥150 min or MET-min≥900, aligning with WHO guidelines [23]. The data synchronization rate was the percentage of weeks with successful device-data synchronization on ≥5 days. Group-level results are presented as the mean (SD) of these individual percentages.

Secondary Outcomes

The secondary outcomes evaluated the potential far-end health benefits of increased PA, including physical fitness, mental health, EF, and body composition. We prespecified mental health and EF as secondary outcomes because (1) increasing PA is associated with lower depressive and anxiety symptoms in adult and college populations, and exercise shows small-to-moderate benefits in randomized trials [24] and (2) team-based, goal-driven gamification requires planning, monitoring, inhibition, and cognitive flexibility, core executive domains plausibly sensitive to change during an 8-week intervention [25].

- Physical fitness: physical fitness was field-tested on the Changxiang 100 platform in accordance with the National Student Physical Fitness Standards (GB/T 18041 - 2022). The assessment comprised 7 tests to evaluate multiple facets of physical fitness: 1000 m run (men) / 800 m run (women), vital capacity, sit-and-reach test, 1-min sit-ups (women) / pull-ups (men), 50 m sprint, and standing long jump. This standardized national protocol ensures the assessments’ widespread applicability and comparability.
- Mental health: depressive and anxiety symptoms were assessed online (Wenjuanxing) using the Zung Self-Rating Depression Scale (SDS) and the Zung Self-Rating Anxiety Scale (SAS). These scales are well-established tools with demonstrated good reliability and validity (eg, SAS Cronbach α=.84; SDS Cronbach α=.88). Recent studies in Chinese student samples and broader Chinese cohorts support acceptable internal consistency and the applicability of Chinese norms for SDS/SAS. The SDS comprises 20 items, each rated on a 4-point scale, with the total raw score converted to an index. Similarly, the SAS contains 20 items rated on a 4-point scale. For both scales, higher standardized scores indicate a greater severity of symptoms. We report raw and standardized scores; higher scores indicate greater symptom burden [26].
- EFs: EFs were measured on iPad using the National Institutes of Health (NIH) Toolbox Cognition Battery (Figure 1F). This computerized assessment battery was selected due to its strong psychometric properties, standardization, and ability to efficiently measure core components of EF. The battery specifically targeted 3 key domains: the dimensional change card sort test, which assesses cognitive flexibility by requiring individuals to shift between matching rules; the flanker inhibitory control

XSL•FO
RenderX

and attention test, which measures inhibitory control and attention by assessing the ability to suppress responses to distracting stimuli; and the List Sorting Working Memory Test, which evaluates working memory capacity through the temporary storage and manipulation of information. Scoring used age-adjusted T-scores per the latest NIH Toolbox V3 technical guidance (renormed to the 2020 US Census and validated against gold-standard measures). Recent meta-analyses and reviews indicate that persistent PA and movement-based programs can yield small-to-moderate improvements in EF in youth and young adults, especially in inhibitory control and cognitive flexibility; acute-exercise effects are mixed, underscoring the focus on multiweek training [27].

- Body composition: body fat percentage, skeletal muscle mass, BMI, and basal metabolic rate (BMR) were obtained with a multifrequency bioelectrical impedance analyzer (InBody H30 (InBody Co, Ltd; Figure 1E). We adhered to standard pretest controls (eg, no vigorous exercise or large meals prior to testing). This objective bioelectrical impedance analysis ensured accurate tracking of changes in body composition resulting from the PA intervention.

Sample Size

Based on findings from previous comparable studies [28-30], the sample size was determined using G*Power 3.1 (Heinrich Heine University Düsseldorf). A 2-tailed significance level of α=.05, statistical power (1-β)=.80, and an anticipated medium effect size (d=0.5) were specified. For an independent samples 2-tailed t test comparing 2 groups, using a 2-tailed test, the minimum required sample size was 64 participants per group (total N=128). Allowing for an estimated 20% attrition, the final target was set at 80 participants per group (total N=160).

Randomization

Random allocation sequences were generated using the PROC PLAN procedure in SAS 9.4 (SAS Institute Inc) by an independent researcher who was not involved in participant recruitment, outcome assessment, or data analysis. We used sex-stratified block randomization with a block size of 4 and a 1:1 allocation ratio to the IG (n=80) and CG (n=80) to ensure balance in sex distribution. Allocation concealment was maintained using sequentially numbered, opaque, sealed envelopes prepared off-site; after completion of baseline assessments, research staff opened the next envelope in sequence to assign participants to groups. Recruitment staff had no access to the allocation sequence; assignment was revealed when the next sealed envelope was opened after baseline assessments.

For the IG only, a postrandomization team structure was created to implement the gamified competition mechanics. The IG was subdivided into 4 equal-size teams (n=20 each). An independent coordinator used constrained random allocation based only on sex and headcount (no outcome data) to form 2 single-sex teams (Team A: 20 males; Team B: 20 females) and 2 mixed-sex teams (Teams C-D: 10 males and 10 females each). Each team functioned as an independent gamification unit for scoring and ranking. In contrast, CG participants were not assigned to any teams and did not participate in gamified competitions; they only completed the baseline PA tasks set on the same platform. This subdivision served two purposes: (1) to create the stable in-app subgroups required for points, leaderboards, and weekly challenges and (2) to promote competition fairness and mitigate structural influences arising from sex-related differences in PA and fitness [31,32].

Blinding

Given the visible nature of the intervention components, namely the gamified team competition and group structure, this study adopted an open-label design, and participant blinding was not feasible. Participants were informed that they would be randomly assigned either to the IG, which involved team-based gamified competition, or to the CG, which involved only meeting the basic PA targets. Outcome assessors were blinded. Physical-fitness testers, InBody H30 technicians, and NIH Toolbox/SDS/SAS administrators used deidentified testing schedules, were trained not to discuss allocation, and logged any accidental unmasking. Data analysts worked on a deidentified dataset with groups masked as A/B; the statistical analysis plan was finalized and code frozen before unmasking, which occurred only after database lock and completion of primary analyses.

Statistical Methods

Data analysis followed the intention-to-treat (ITT) principle. Continuous variables were expressed as mean (SD), and categorical variables as frequency and percentage (n%). Baseline differences were assessed using independent-samples t tests (or Mann-Whitney U tests, as appropriate) for continuous variables and χ^2 tests (or Fisher exact tests) for categorical variables.

The primary outcomes (PA and adherence) were summarized as mean values over the 8-week intervention period. Between-group differences in these postintervention means were then assessed using ANCOVA, with the corresponding baseline values included as covariates. For secondary outcomes (physical fitness, body composition, SDS/SAS, and EF), ANCOVA with baseline adjustment was likewise used, with mixed effects models as sensitivity analyses. Adherence-related indicators (eg, goal completion and data synchronization) were summarized descriptively and compared between groups; when applicable, ANCOVA with baseline adjustment was applied. Sex-based comparisons were not prespecified, and the trial was not powered for sex-related effect modification. To avoid multiplicity and over-interpretation, no inferential subgroup analyses (eg, group×sex interactions) were conducted; inference focuses on overall between-group effects.

To address missing data, Little test for missing completely at random (MCAR) was applied to all outcomes measured at T0 and T1. The results showed no significant departure from MCAR (χ^2_{15}=11.236; P=.73). Given the low proportion of missing values (<5%) and to preserve statistical power and reduce potential bias associated with listwise deletion, multiple imputation (m=20) using chained equations was performed under a missing-at-random assumption. The imputation model included group assignment (IG vs CG), sex, age, baseline values of each outcome, and adherence indicators (eg, percentage of days meeting the PA goal) as predictors. Robustness was confirmed by reporting both complete-case and multiple

XSL•FO
RenderX

imputation–pooled estimates for primary outcomes and through sensitivity analyses that excluded algorithm-flagged suspect data, which yielded consistent conclusions. All statistical tests were 2-sided, with $P<.05$ considered statistically significant.

Analyses were performed using SPSS (version 26.0; IBM Corp).

Ethical Considerations

The study received ethics approval (Approval No. 2025A125, approved on January 15, 2025) from the Ethics Committee of Capital University of Physical Education and Sports. The study protocol was finalized before the start of participant recruitment (February 4, 2025) and is identical to the protocol retrospectively registered in the Chinese Clinical Trial Registry (ChiCTR2500108950) with respect to trial design, eligibility criteria, prespecified primary and secondary outcomes, and planned statistical analyses. No changes were made to these elements after trial commencement.

Trial registration was completed retrospectively in the Chinese Clinical Trial Registry (ChiCTR2500108950) on September 9, 2025, after enrollment had begun, because prospective registration was inadvertently not completed during trial setup due to an administrative oversight; this was not an attempt to avoid registration. To mitigate concerns about selective reporting, the institutional review board (IRB)–approved protocol was finalized prior to recruitment and remained unchanged throughout the trial with respect to the trial design, eligibility criteria, prespecified outcomes, and planned statistical analyses. An official letter from the IRB confirming that the submitted study design is identical to the protocol assessed and approved before trial initiation has been provided to the editorial office for review.

All procedures were conducted in accordance with the principles of the Declaration of Helsinki. Before enrollment, all participants received written and verbal information about the study purpose, procedures, potential risks, and anticipated benefits and subsequently provided written informed consent. Participants were informed that participation was voluntary and that they could withdraw from the study at any time without providing a reason. This withdrawal would incur no penalty or detriment to their academic standing. To ensure privacy and confidentiality, all participant data were deidentified upon collection and assigned a unique study ID. The analytical dataset contained only these anonymized IDs. The master list linking IDs to personal information was stored separately on a secure, password-protected server with access restricted to the principal investigator and the data manager. Participants who completed the entire intervention cycle received nonmonetary compensation (accumulated volunteer hours, a certificate of honor, and coordination with academic credit where applicable). Additionally, the top 10% of performers received a material reward of 400 Chinese yuan (Renminbi; approximately US $55.53 or €49.31, based on the exchange rate in April 2025). No identifiable images of participants are presented in this manuscript or in the supplementary materials; the participants pictured in Figure 1F provided written consent for publication of the deidentified photographs for academic purposes.

Data Security and Risk Management

To prevent cheating and ensure data integrity, each smartwatch (serial number/MAC) was identity-verified and bound 1:1 to a unique participant account. Weekly, a random 10% sample of participants underwent raw-data audits in which 7-day device traces (steps, moderate-to-vigorous physical activity [MVPA], metabolic equivalent-minutes [MET-min], and wear time with timestamps) were cross-checked against server sync logs; discrepant days were documented and flagged for sensitivity analyses. For privacy and confidentiality, analytical datasets were deidentified, access followed role-based access control under a least-privilege policy, and data were used solely for academic research as approved by the ethics committee; all access and exports were audit-logged. To reduce attrition, participants received a standardized briefing on study requirements before randomization; withdrawals were allowed at any time, and reasons were recorded and summarized in the CONSORT flow.

Harms

Safety was monitored through weekly checks and audits using predefined triggers; events were logged by a blinded assessor. Serious adverse events required immediate IRB reporting and discontinuation as prespecified in the protocol. Patients or members of the public were not involved in this trial. No important changes to the trial protocol, outcomes, or analyses were made after trial initiation.

Results

Participant Flow

A total of 224 volunteers were screened for eligibility. After excluding 64 individuals who did not meet the inclusion criteria, 160 participants were enrolled and randomly assigned in a 1:1 ratio to the IG and CG, with 80 participants in each group (Figure 3).

XSL•FO
RenderX

Figure 3. CONSORT (Consolidated Standards of Reporting Trials) flow diagram, participant flow through the randomized trial. CG: control group; IG: intervention group; PA: physical activity;

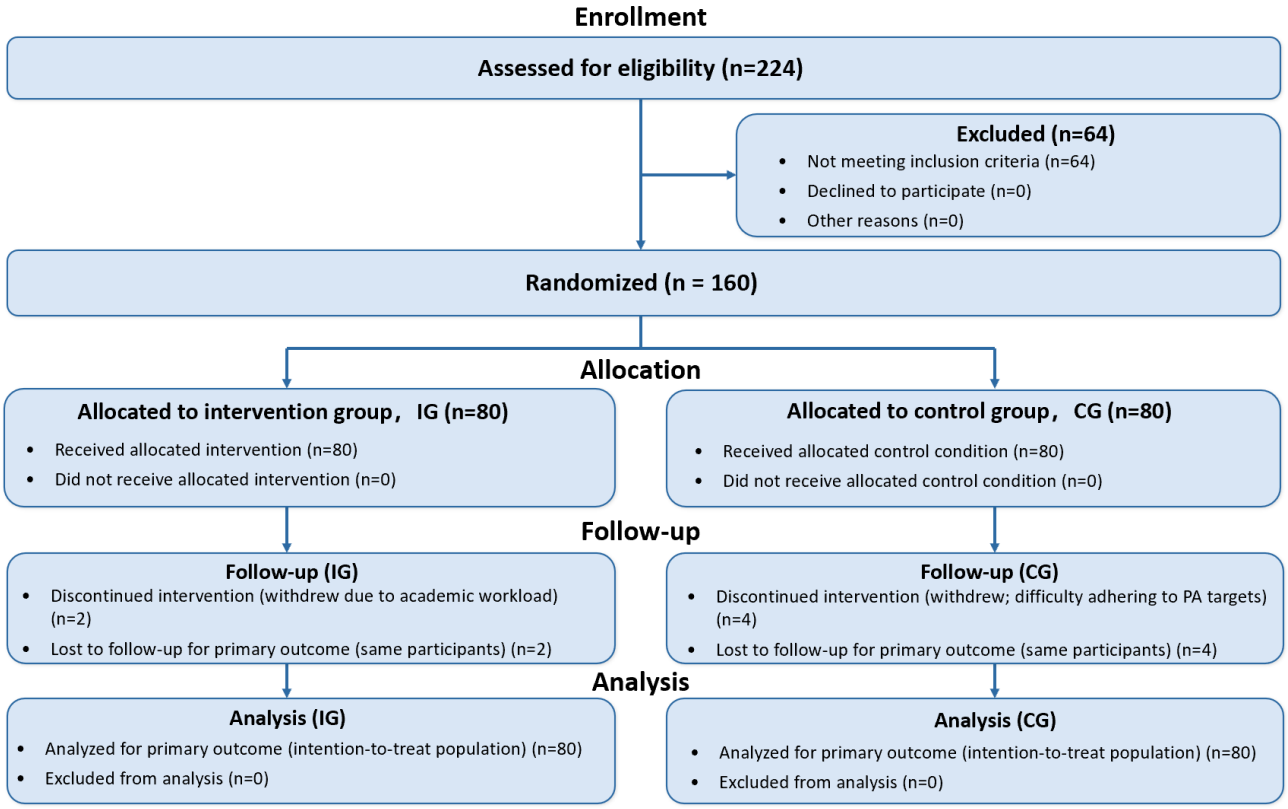

During the 8-week intervention period, 6 participants withdrew from the study (IG: 2 participants, 2.5%; CG: 4 participants, 5%). The primary reasons for dropout were academic workload (n=2) and failure to meet PA targets (n=4). Overall, this resulted in 3.8% missingness in post-intervention outcomes. Little MCAR test across all primary and secondary outcomes indicated that this missingness did not significantly depart from MCAR (χ^2_{15}=11.236; *P*=.73).

The trial ended as planned after completion of the prespecified 8-week intervention and postintervention assessments and was not stopped early.

Baseline Data

There were no statistically significant differences between the 2 groups in baseline demographic characteristics (age, sex, and BMI) or psychological scale scores (*P*>.05), indicating good comparability between group (Table 1).

XSL•FO
RenderX

Table . Baseline characteristics of participants.

Characteristic	IG[a] (n=80)	CG[b] (n=80)	Total (N=160)	*P* value
Age (years), mean (SD)	20.5 (1.2)	20.8 (1.3)	20.6 (1.3)	.15
Sex, n (%)				
Female	40 (50.0)	40 (50.0)	80 (50.0)	>.99
Male	40 (50.0)	40 (50.0)	80 (50.0)	>.99
Body composition, mean (SD)				
Body weight, (kg)	66.4 (12.5)	65.8 (11.8)	66.1 (12.1)	.76
BMI (kg/m²)	22.4 (2.5)	22.1 (2.3)	22.2 (2.4)	.42
Mental health, mean (SD)				
SAS[c] score	44.3 (7.2)	45.1 (5.9)	44.7 (6.6)	.43
SDS[d] score	46.4 (7.7)	45.9 (8.1)	46.2 (7.9)	.67
PA[e]				
Daily steps, mean (SD)	7842 (1,603)	7921 (1,425)	7882 (1,514)	.74
MVPA[f] (min/day), mean (SD)	37.2 (11.5)	37.4 (10.7)	37.3 (11.1)	.91
Weekly MET-min[g], mean (SD)	1020 (205)	1035 (220)	1028 (212)	.66

[a]IG: intervention group.
[b]CG: control group.
[c]SAS: Self-Rating Anxiety Scale.
[d]SDS: Self-Rating Depression Scale.
[e]PA: physical activity.
[f]MVPA: moderate-to-vigorous physical activity.
[g]MET-min: metabolic equivalent-minute.

Intervention and Comparator Delivery

Both arms were technology-mediated and primarily self-administered; the study team provided standardized onboarding, technical support, and data monitoring. No crossovers or major deviations from the intended delivery were identified, and adherence indicators (goal completion and data synchronization) are summarized in Table 2.

XSL•FO
RenderX

Table . Outcome during the 8-week intervention and change from T0 to T1.

Domain and outcome[a]	IG[b] (n=80), mean (SD)	CG[c] (n=80), mean (SD)	Δ (IG-CG)	Δ IG-CG (%)	*P* value	Effect size, *d*[d] (95% CI)
PA[e] levels[f]						
Daily steps	10,356 (1245)	8242 (1087)	+2114	25.6	<.001	1.81 (1.44 to 2.18)
Daily MVPA[g] (min)	71 (15)	43 (12)	+28	65.1	<.001	2.06 (1.68 to 2.45)
Weekly MET-min[h]	1650 (310)	1340 (285)	+310	23.1	<.001	1.04 (0.71 to 1.37)
Adherence (%)						
Goal completion rate	88.9 (8.5)	77.5 (10.2)	+11.4	14.7	<.001	1.21 (0.88 to 1.55)
Data sync rate	89.1 (7.7)	75.5 (11.3)	+13.6	18.0	<.001	1.41 (1.06 to 1.75)
Body composition						
Δ Skeletal muscle (kg)	+0.86 (0.23)	+0.32 (0.11)	+0.54	168.8	<.001	3.00 (2.54 to 3.45)
Δ Body fat (%)	−1.46 (0.51)	−1.00 (0.32)	−0.46	46.0	<.001	−1.08 (−1.41 to −0.75)
Δ BMR[i] (kcal/d)	+20.7 (4.9)	+12.9 (4.8)	+7.8	60.5	<.001	1.61 (1.25 to 1.97)
Δ Body weight (kg)	−1.17 (1.45)	−0.91 (1.45)	−0.26	28.6	.26	−0.18 (−0.49 to 0.13)
Δ BMI (kg/m²)	−0.45 (1.10)	−0.26 (1.04)	−0.19	73.1	.26	−0.18 (−0.49 to 0.13)
Physical fitness						
Δ 800/1000 m run (s)	−13.0 (6.0)	−6.5 (4.0)	−6.5	100.0	<.001	−1.27 (−1.62 to −0.93)
Δ Vital capacity (mL)	+470 (110)	+280 (80)	+190	67.9	<.001	1.98 (1.60 to 2.35)
Δ Sit-and-reach (cm)	+5.2 (1.1)	+3.4 (1.2)	+1.8	52.9	<.001	1.56 (1.21 to 1.92)
Δ Sit-ups (female), counts (n=40 per group)	+10.5 (2.0)	+6.0 (1.5)	+4.5	75.0	<.001	2.55 (1.95 to 3.14)
Δ Pull-ups (male, counts, n=40 per group)	+0.6 (0.5)	+0.4 (0.5)	+0.2	50.0	.08	0.40 (−0.04 to 0.84)
Δ 50 m sprint (s)	−0.30 (0.30)	−0.25 (0.30)	−0.05	20.0	.28	−0.17 (−0.48 to 0.14)
Δ Standing long jump (cm)	+8.5 (13.1)	+4.5 (14.5)	+4.0	88.9	.07	0.29 (−0.02 to 0.60)
EF[j]						
Δ Cognitive flexibility (ms)	−104 (26)	−75 (32)	−29	38.7	<.001	−0.99 (−1.32 to −0.67)
Δ Inhibitory control (errors)	−1.4 (2.5)	−1.0 (2.5)	−0.4	40.0	.31	−0.16 (−0.47 to 0.15)
Δ Working memory (correct)	+0.5 (1.5)	+0.3 (1.3)	+0.2	66.7	.38	0.14 (−0.17 to 0.45)
Mental health						
Δ SDS[k] (depression)	−4.5 (2.0)	−2.5 (1.5)	−2.0	80.0	<.001	−1.13 (−1.47 to −0.80)

XSL•FO
RenderX

Domain and outcome[a]	IG[b] (n=80), mean (SD)	CG[c] (n=80), mean (SD)	Δ (IG-CG)	Δ IG-CG (%)	*P* value	Effect size, *d*[d] (95% CI)
Δ SAS[l] (anxiety)	−2.2 (3.5)	−1.3 (3.1)	−0.9	69.2	.09	−0.27 (−0.58 to 0.04)

[a]Values are mean (SD). Analyses followed intention-to-treat with multiple imputation

[b]IG: intervention group.

[c]CG: control group.

[d]Effect sizes are Cohen *d* (pooled SD); the sign of *d* follows Δ (IG−CG).

[e]PA: physical activity.

[f]PA and adherence outcomes were summarized over the 8-week intervention period; outcomes prefixed with Δ are change scores (T1−T0).

[g]MVPA: moderate-to-vigorous physical activity.

[h]MET-min: metabolic equivalent-minutes.

[i]BMR: basal metabolic rate.

[j]EF: executive function.

[k]SDS: Self-Rating Depression Scale.

[l]SAS: Self-Rating Anxiety Scale.

No concomitant care was provided by the study team, and concomitant interventions outside the trial that could influence outcomes were not systematically monitored.

Numbers Analyzed

Primary and secondary outcomes were analyzed according to the ITT principle (IG n=80; CG n=80). Missing postintervention (T1) values were handled using multiple imputation as prespecified. Outcome data were available at T1 for 78/80 participants in the IG (97.5%) and 76/80 in the CG (95.0%). Table 2 presents MI-pooled descriptive statistics and between-group inferences based on the primary ITT analysis.

Outcomes and Estimation

Primary Outcomes

PA Levels

Across the 8-week period, both arms showed the typical early rise-mid plateau-late rebound pattern. Adjusted between-group differences consistently favored IG for steps, daily MVPA, and weekly MET-min throughout the intervention (all $P<.001$). These results indicate that the gamified mHealth arm elicited additional PA beyond baseline platform targets; see Table 2 for detailed outcomes and 95% CIs and Figure 4 for effect sizes.

XSL•FO
RenderX

Figure 4. Effect sizes (*d*) and 95% CIs for each outcome measure (IG vs CG): positive values favor IG (directions standardized). The dashed vertical line indicates no between-group difference (*d*=0); CIs crossing 0 indicate nonsignificance. MET-min: metabolic equivalent-minute; MVPA: moderate-to-vigorous physical activity; SAS: Self-Rating Anxiety Scale; SDS: Self-Rating Depression Scale.

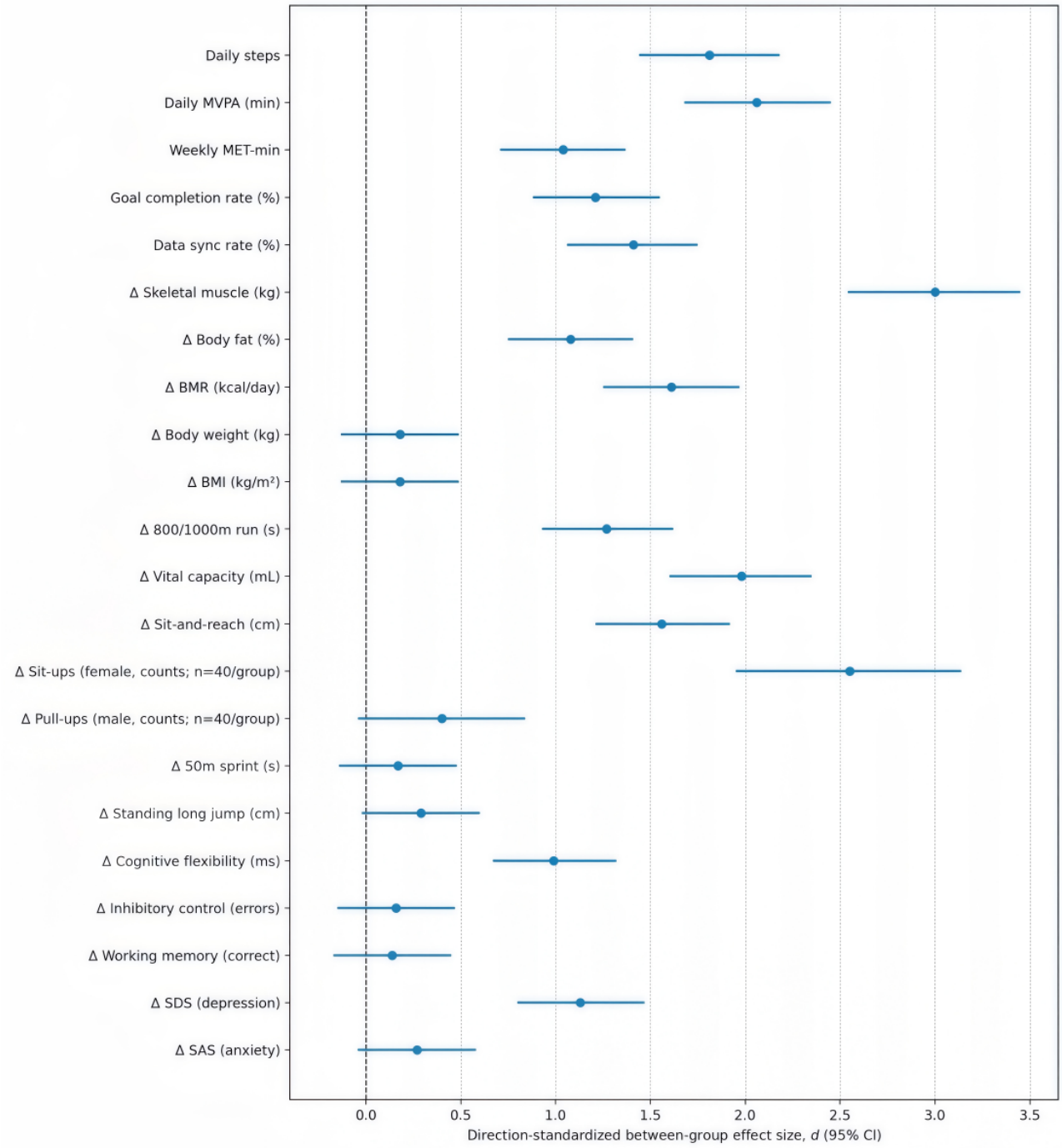

Intervention Adherence

IG demonstrated significantly higher adherence—for both goal completion and data synchronization—than CG (all $P<.001$). Exact means, between-group differences (Δ), and effect sizes with 95% CIs are reported in Table 2.

Secondary Outcomes

Physical Fitness

The intervention produced differential effects across fitness levels. For physical fitness, IG improved more than CG in aerobic endurance (800/1000 m run), vital capacity, muscular endurance (female sit-ups), and flexibility (sit-and-reach) (all $P<.001$), while no between-group differences emerged for 50 m sprint, pull-ups, or standing long jump ($P>.05$). See Table 2 and Figure 4 for detailed outcomes and 95% CIs.

Body Composition

Relative to CG, IG showed larger body-composition gains, higher skeletal muscle mass, lower body fat percentage, and higher BMR (all $P<.001$), while changes in body weight and BMI showed no significant between-group differences ($P>.05$). See Table 2 for detailed outcomes and 95% CIs and Figure 4 for effect sizes.

XSL•FO
RenderX

EF

EF showed a domain-specific pattern, cognitive flexibility improved in IG ($P<.001$), whereas working memory and inhibitory control showed no between-group differences ($P>.05$). See Table 2 and Figure 4 for detailed outcomes and 95% CIs.

Mental Health Status

Depressive symptoms (SDS) decreased more in IG than in CG ($P<.001$), whereas anxiety (SAS) changes were small and not different between group ($P>.05$). Individual trajectories varied; see Table 2 and Figure 4 for detailed outcomes and 95% CIs.

Ancillary Analyses

Given the lack of prespecification for sex-based subgroup inference, no inferential sex comparisons are reported; primary conclusions rely on overall effects.

Harms

No study-related serious adverse events; no device-related dermatologic events requiring care; and no major technical failures or data losses affecting outcome interpretation were recorded.

Discussion

Principal Results

This RCT evaluated the incremental effect of adding team-based gamification and incentives to a standard wearable- and app-based PA program for college students. Compared with an active control using the identical platform and activity goals, the IG showed greater improvements in the prespecified primary outcomes—objectively measured PA and intervention adherence—and in selected secondary outcomes related to physical fitness and body composition, whereas effects on EF and anxiety outcomes were limited. This overall pattern is consistent with meta-analytic evidence that gamified mHealth interventions tend to increase daily steps/MVPA and yield small-to-moderate improvements in adiposity-related indicators compared with nongamified comparators [33,34]. Notably, by holding the app-wearable platform and activity goals constant across groups, this study may help isolate the incremental contribution of gamification and reduce confounding that is common when platforms and components differ across conditions [35]. The findings indicate that the added value of gamification in this context may operate primarily through strengthening adherence and engagement, leading to downstream benefits for selected physical outcomes [36].

Mechanisms

Improvements in PA and adherence may be interpreted within SDT and TPB, in that need-supportive gamification could facilitate motivational internalization and help translate intention into action [37]. First, SDT posits that autonomy, competence, and relatedness are core components of intrinsic motivation [38,39]. In this study, team-based competition, real-time feedback, and visible progress cues may have supported competence and relatedness while maintaining task structure to promote more sustained engagement in PA.

In addition, the intervention combined support for basic psychological needs (autonomy, competence, and relatedness) with extrinsic incentives (points and leaderboards) [40]. From an SDT perspective, a key issue is how such external incentives are framed and delivered [41]. When points and leaderboards are presented as informational signals of personal progress rather than controlling demands, they may better support autonomy and align with the regulatory continuum of motivational internalization [42], which could help explain the observed improvements in adherence during the intervention period.

More specifically, short-term gains in PA and adherence may reflect a synergy between extrinsic prompts and need-supportive features: the former may catalyze initial participation and short-term persistence, whereas the latter may enhance self-efficacy and belonging by satisfying competence and relatedness needs [43]. Nevertheless, the salience and material nature of certain incentives may slow internalization toward more integrated regulation [44]. In this RCT, in addition to designing nonmaterial rewards such as cumulative volunteer hours, honorary certificates, and academic credit coordination, we also incorporated a conditional monetary reward for the IG, tied to leaderboard performance. This external economic incentive served as a salient extrinsic goal and could bolster adherence and help sustain PA, particularly during the mid-to-late stages of the intervention. However, its effect may attenuate after the incentive is withdrawn and could further limit motivational internalization [45]. Such incomplete internalization could partly limit the transfer of benefits to domains that typically require greater cognitive engagement, longer accumulation, or additional components beyond increasing activity volume (eg, strength-related fitness and selected EF domains) [46]. Evidence from multi-arm randomized trials directly comparing gamification, financial incentives, and their combination supports the plausibility that each component can increase PA and that combining them may yield larger improvements than either component alone, consistent with additive or synergistic effects [47]. Therefore, the intervention benefits observed in the present study are more likely the result of the combined action of the engaging environment created by gamification and the salient extrinsic goals set by monetary rewards.

Overall, the SDT-based account primarily addresses motivation formation and internalization. TPB provides a complementary perspective by clarifying how attitudes, subjective norms, and perceived behavioral control jointly shape the translation of behavioral intention into actual behavior [48]. In the present intervention, digital reminders, social-comparison signals, and personalized feedback may have strengthened perceived control and normative support, thereby facilitating daily goal execution and adherence [49].

Interpretation and Boundaries

Unlike prior physical fitness–promotion interventions that typically report small-to-moderate effect sizes [7], the present mHealth-based intervention with added gamification showed comparatively large standardized effect sizes for objectively measured physical activity and selected fitness-related outcomes. This pattern may reflect the intervention's efficacy under

XSL•FO
RenderX

relatively controlled trial conditions. First, the single-university student cohort was relatively homogeneous in age range and daily routines, which may reduce within-sample outcome variability and thereby inflate standardized effect size estimates for a given raw mean difference [50]. Second, delivery was supported by a technology platform with high implementation intensity (eg, frequent monitoring, scheduled synchronization, and staged feedback) and high adherence during the intervention period (Table 2), which likely improved intervention fidelity and strengthened the effective "dose contrast" between groups. Prior evidence suggests that higher engagement/adherence in digital interventions is associated with larger behavioral effects [51]. However, when such interventions are scaled from controlled trial settings to more heterogeneous real-world populations with lower monitoring/support intensity, observed effect sizes may attenuate [52]. Building on this overall interpretation, we now turn to the domain-specific findings and key boundaries.

The higher frequency of MVPA in IG may have contributed to improvements in aerobic capacity [53]. This is plausible because the intervention primarily promoted cardiovascular adaptation by increasing activity volume, and the achieved activity pattern was broadly aligned with WHO-recommended PA levels [23]. In contrast, between-group differences were not statistically significant for strength- or power-related outcomes (eg, 50-m sprint, standing long jump, pull-ups). A primary explanation may lie in the structural emphasis of the intervention: the program was designed to promote MVPA via mHealth, with common modalities such as jogging and cycling that primarily load cardiorespiratory and energy-metabolic systems, but it did not include a systematic resistance-training component targeting large muscle groups [54]. Improvements in strength/power generally require more targeted resistance training, and baseline capability (eg, the ability to perform pull-ups) may also influence the magnitude of improvement [55,56].

Although the intervention did not yield statistically significant changes in body weight or BMI, IG showed significant improvements in body-composition indicators (eg, skeletal muscle mass, body fat percentage, and BMR) relative to CG. This pattern may be explained by the fact that, in the absence of strict dietary control and a standardized resistance-training dose, increasing PA can preferentially improve body composition [57]. Moreover, because adipose tissue has a lower density than muscle tissue, concurrent fat loss and muscle gain may result in a "weight substitution effect," whereby meaningful changes in body shape occur with minimal change in total body weight [58]. In addition, reductions in fat mass together with increases in lean mass may elevate resting metabolic rate, consistent with the metabolic benefits associated with increases in fat-free mass [59].

For EF, IG showed a statistically significant improvement only in cognitive flexibility, whereas changes in working memory and inhibitory control did not reach statistical significance. This suggests that cognitive flexibility may be a domain more readily influenced by this PA-focused intervention. One possible mechanism is the acute and cumulative effects of moderate-intensity aerobic exercise on prefrontal hemodynamics and neural signaling [60]. By contrast, limited changes in working memory and inhibitory control are also consistent with the general "dose-response" and domain-specific nature of EF adaptation [61], whereby these domains may require higher intensity, longer duration, or more cognitively engaging activity to change meaningfully [62,63].

The intervention showed limited effects on anxiety symptoms, which may reflect that the protocol primarily targeted PA behavior and did not incorporate psychological components (eg, cognitive behavioral therapy, mindfulness, or stress-management modules) that more directly modulate emotional processes [64]. In addition, gamification mechanisms may be double-edged: competitive elements can enhance motivation, but they may also thwart autonomy and relatedness for some participants, potentially constraining psychological benefits [65]. Prior evidence also suggests substantial heterogeneity in mental health responses to gamified interventions, which may relate to individual sensitivity to social comparison and competitive pressure [66]. Moreover, PA tends to show stronger effects on depressive symptoms than on anxiety, and benefits may be more apparent among participants with milder baseline symptom levels, implying conditional efficacy [67].

Taken together, changes in EF and mental health may be interpreted as potential, conditional downstream benefits of PA-focused interventions rather than inevitable effects of gamification per se [62]. This further suggests that when interventions aim to target mental health or EF, it is important to align mechanisms, dose/intensity, and the target population to achieve more precise matching and maximize the likelihood of benefit [68].

Limitations

This study used an RCT design and leveraged mHealth technology to enable real-time tracking and feedback of intervention behaviors. Nevertheless, several limitations should be acknowledged.

First, the 8-week intervention duration, while sufficient for assessing short-term efficacy, may be insufficient to observe long-term behavioral maintenance and sustained neural or physiological adaptations. Future studies with longer follow-up periods are needed to evaluate the sustainability of the effects [69]. Second, the absence of systematic control over dietary intake and sleep patterns represents a potential confounding factor for the observed changes in body composition, as these lifestyle components are known to interact with PA [70]. Third, all participants were recruited from a single university, which may limit generalizability to other collegiate populations and demographic groups; the relatively homogeneous cohort may also have reduced outcome variability [71]. Fourth, the team sex composition was implemented for fairness and feasibility rather than for testing sex differences; the study was neither powered to detect effect modification by sex, nor were such comparisons prespecified. Additionally, because conditional financial incentives were bundled with gamification only in the IG, we could not disentangle the independent effects of incentives versus gamification [72]. Moreover, although the protocol had been finalized before recruitment and remained unchanged throughout the trial, the retrospective registration

XSL•FO
RenderX

of the trial due to an administrative oversight may still be considered a limitation in terms of methodological transparency and potential risk of bias. Finally, the evaluation of EF and mental health, although we cited literature [73], is limited by the lack of baseline impairment, which limits sensitivity to change and raises the possibility of ceiling effects.

Implications

Overall, this 8-week RCT showed that, compared with a nongamified (monitoring only) intervention with identical activity goals, a gamified mHealth exercise intervention delivered within the same app-wearable system yielded clear incremental benefits: it increased PA and intervention adherence and improved selected physical fitness and body-composition indicators. The key innovation lies in quantifying the "incremental effect" attributable to gamification on top of a unified activity goal and the same technical platform, thereby allowing a clearer estimate of the incremental contribution of gamification. This approach differed from much of the prior literature, where between-group comparisons were often confounded by co-occurring intervention components [35], and it reduced such confounding by holding the technology platform and activity goals constant across groups. In terms of evaluation, by integrating objective behavioral data with a multidimensional outcome set, the study offered more robust and comparable evidence and helped delineate where gamification is most likely to be effective and where its effects may be bounded. From a real-world perspective, the model could be scaled at low cost by leveraging existing smartphone apps and wearable devices, supporting feasible implementation in university health-promotion and campus health management programs and offering an accessible digital health strategy to address physical inactivity among college students and inform related public health practice and policy [74].

Future Directions

Building on our findings and limitations, we propose several avenues for future research: (1) investigate integration of multimodal components, such as the integration of structured resistance training components to more effectively target muscular strength and power, which showed limited improvement here [75]; (2) extend intervention exposure (eg, >12 - 16 wk) and prespecify dose-response targets for PA to test whether benefits for domains like working memory and inhibition emerges with a greater cumulative dose [76]; (3) explore synergies with evidence-based psychological approaches. This includes embedding digital cognitive behavioral therapy elements or mindfulness training within the mHealth platform to directly target mechanisms underlying anxiety, depression, and cognitive regulation [77]; (4) testing the adapted intervention in more diverse populations, including individuals with baseline subclinical symptoms or lower digital literacy, will be crucial for determining its broader public health applicability [78]; (5) add pragmatic diet/sleep capture (eg, 1 - 2 24-hour dietary recalls per week including one weekend day, photo-assisted intake logs, and 7-day sleep diary + actigraphy/wearable-derived sleep/heart rate) to strengthen causal attribution for body-composition outcomes [79]; (6) preregister sex-related hypotheses, power accordingly, and conduct planned interaction/subgroup analyses to evaluate potential heterogeneity of effects by sex and to align analytic decisions with study aims [80]; and (7) use multiarm or factorial designs (eg, gamification-only, incentive-only, both, and control), with postincentive follow-up, to isolate component-specific effects and assess maintenance [47].

Conclusions

This RCT demonstrates that an mHealth intervention integrating gamification effectively enhanced physical fitness, body composition, and adherence to PA among college students. While its impact on psychological health and EF was limited, the approach demonstrates strong potential for scalable, technology-driven health promotion in young adult populations.

However, because this study did not include post-intervention follow-up, the sustainability of the effects was not assessed. Future research should evaluate maintenance and scalability in more diverse real-world settings, ideally using multiarm/factorial designs with postincentive follow-up. Studies should also test longer intervention exposure, multimodal components, and dietary strategies.

Acknowledgments

This work was supported by the Emerging Interdisciplinary Platform for Medicine and Engineering in Sports (EIPMES).

Funding

This study was funded by the Emerging Interdisciplinary Platform for Medicine and Engineering in Sports (EIPMES) and the teaching reform program of Yantai University (Grant No. JYXM2024025Y).

Data Availability

The trial protocol, statistical analysis plan, de-identified individual participant data (including the data dictionary), statistical code, and other study materials are available from the corresponding author on reasonable request.

Authors' Contributions

MY and ZZ performed conceptualization. MY, Y Guo, ZZ, SY, ZL, Y Gao, and JL performed methodology. Y Gao conducted the literature search. MY and ZZ wrote the original draft. ZZ performed visualization. All authors contributed to writing – review & editing.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Editorial Notice

This randomized study was only retrospectively registered because the lack of prospective registration was due to an administrative oversight during trial setup, which resulted in the registration being omitted from the trial-start checklist. Once the omission was identified while checking the journal's submission and reporting requirements for randomized controlled trials, the authors retrospectively registered the trial in the Chinese Clinical Trial Registry. The editor granted an exception from ICMJE rules mandating prospective registration of randomized trials because the risk of bias appears low and because the authors provided an official letter from the institutional review board confirming that the submitted study design is identical to the protocol assessed and approved before trial initiation. However, readers are advised to carefully assess the validity of any potential explicit or implicit claims related to primary outcomes or effectiveness, as retrospective registration does not prevent authors from changing their outcome measures retrospectively.

Checklist 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File, 245 KB - jmir_v28i1e82769_app1.pdf]

References

1. Jia X, Hu C, Xu Y, et al. Revisiting obesity thresholds for cardiovascular disease and mortality risk in Chinese adults: age- and gender-specific insights. Cell Rep Med 2025 Sep 16;6(9):102309. [doi: 10.1016/j.xcrm.2025.102309] [Medline: 40858104]
2. Sun Q, Yu D, Fan J, et al. Healthy lifestyle and life expectancy at age 30 years in the Chinese population: an observational study. Lancet Public Health 2022 Dec;7(12):e994-e1004. [doi: 10.1016/S2468-2667(22)00110-4] [Medline: 35926549]
3. Sun Z, Sun M, Wei W, Peng W, Wang Y. China launches National Obesity Campaign. Lancet Diabetes Endocrinol 2025 Jun;13(6):465-466. [doi: 10.1016/S2213-8587(25)00131-7] [Medline: 40404276]
4. Lee SA, Park JH. Systematic review and meta analysis of standalone digital behavior change interventions on physical activity. NPJ Digit Med 2025 Jul 14;8(1):436. [doi: 10.1038/s41746-025-01827-4] [Medline: 40659761]
5. Beratarrechea A, Lee AG, Willner JM, Jahangir E, Ciapponi A, Rubinstein A. The impact of mobile health interventions on chronic disease outcomes in developing countries: a systematic review. Telemed J E Health 2014 Jan;20(1):75-82. [doi: 10.1089/tmj.2012.0328] [Medline: 24205809]
6. Mönninghoff A, Kramer JN, Hess AJ, et al. Long-term effectiveness of mHealth physical activity interventions: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2021 Apr 30;23(4):e26699. [doi: 10.2196/26699] [Medline: 33811021]
7. Mazeas A, Duclos M, Pereira B, Chalabaev A. Evaluating the effectiveness of gamification on physical activity: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2022 Jan 4;24(1):e26779. [doi: 10.2196/26779] [Medline: 34982715]
8. Looyestyn J, Kernot J, Boshoff K, Ryan J, Edney S, Maher C. Does gamification increase engagement with online programs? A systematic review. PLoS ONE 2017;12(3):e0173403. [doi: 10.1371/journal.pone.0173403] [Medline: 28362821]
9. Autin KL, Herdt ME, Garcia RG, Ezema GN. Basic psychological need satisfaction, autonomous motivation, and meaningful work: a self-determination theory perspective. J Career Assess 2022 Feb;30(1):78-93. [doi: 10.1177/10690727211018647]
10. Abou Kamar M, Maher A, Salem IE, Elbaz AM. Gamification impact on tourists' pro-sustainability intentions: integration of technology acceptance model (TAM) and the theory of planned behaviour (TPB). TR 2024 Feb 6;79(2):487-504. [doi: 10.1108/TR-04-2023-0234]
11. Girelli L, Hagger M, Mallia L, Lucidi F. From perceived autonomy support to intentional behaviour: testing an integrated model in three healthy-eating behaviours. Appetite 2016 Jan 1;96:280-292. [doi: 10.1016/j.appet.2015.09.027] [Medline: 26423363]
12. Jang D, Kim I, Kwon S. Motivation and intention toward physical activity during the COVID-19 pandemic: perspectives from integrated model of self-determination and planned behavior theories. Front Psychol 2021;12:714865. [doi: 10.3389/fpsyg.2021.714865] [Medline: 34393954]
13. Wang Q, Liu J, Zhou L, et al. Usability evaluation of mHealth apps for elderly individuals: a scoping review. BMC Med Inform Decis Mak 2022 Dec 2;22(1):317. [doi: 10.1186/s12911-022-02064-5] [Medline: 36461017]
14. Schmidt-Kraepelin M, Warsinsky S, Thiebes S, Sunyaev A. The role of gamification in health behavior change: a review of theory-driven studies. Presented at: Hawaii International Conference on System Sciences; Jan 7-10, 2020. [doi: 10.24251/HICSS.2020.155]
15. Damghanian M, Alijanzadeh M. Theory of planned behavior, self-stigma, and perceived barriers explains the behavior of seeking mental health services for people at risk of affective disorders. Asian J Soc Health Behav 2018;1(2):54-61. [doi: 10.4103/SHB.SHB_27_18]

XSL•FO
RenderX

16. Kim HM, Cho I, Kim M. Gamification aspects of fitness apps: implications of mHealth for physical activities. Int J Hum Comput Interact 2023 Jun 15;39(10):2076-2089. [doi: 10.1080/10447318.2022.2073322]
17. Yu T, Parry M, Yu T, et al. Effectiveness of mobile health-based gamification interventions for improving physical activity in individuals with cardiovascular diseases: systematic review and meta-analysis of randomized controlled trials. JMIR Serious Games 2025 Jan 24;13:e64410. [doi: 10.2196/64410] [Medline: 39854099]
18. Xu L, Li J, Zhang X, et al. Mobile health-based gamification intervention to increase physical activity participation among patients with coronary heart disease: study protocol of a randomised controlled trial. BMJ Open 2022 Jan 31;12(1):e054623. [doi: 10.1136/bmjopen-2021-054623] [Medline: 35105640]
19. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ 2025 Apr 14;389:e081123. [doi: 10.1136/bmj-2024-081123] [Medline: 40228833]
20. Eysenbach G, Group CE. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
21. Hopewell S, Clarke M, Moher D, et al. CONSORT for reporting randomized controlled trials in journal and conference abstracts: explanation and elaboration. PLoS Med 2008 Jan 22;5(1):e20. [doi: 10.1371/journal.pmed.0050020] [Medline: 18215107]
22. Zhang J, Li Z, Li M, et al. High-precision monitoring of fitness behavior and intelligent recommendation of personalized fitness programs. J Capital Univ Phys Educ Sports 2023;35(3):276-283. [doi: 10.14036/j.cnki.cn11-4513.2023.03.004]
23. WHO Guidelines on Physical Activity and Sedentary Behaviour: World Health Organization; 2020. URL: https://www.who.int/publications/i/item/9789240015128 [accessed 2025-03-20]
24. Contreras-Osorio F, Ramirez-Campillo R, Cerda-Vega E, et al. Effects of physical exercise on executive function in adults with depression: a systematic review and meta-analysis. Int J Environ Res Public Health 2022 Nov 18;19(22):15270. [doi: 10.3390/ijerph192215270] [Medline: 36429985]
25. Ángel Latorre-Román P, Berrios-Aguayo B, Aragón-Vela J, Pantoja-Vallejo A. Effects of a 10-week active recess program in school setting on physical fitness, school aptitudes, creativity and cognitive flexibility in elementary school children. A randomised-controlled trial. J Sports Sci 2021 Jun 3;39(11):1277-1286. [doi: 10.1080/02640414.2020.1864985]
26. Yue T, Li Q, Wang R, et al. Comparison of Hospital Anxiety and Depression Scale (HADS) and Zung Self-Rating Anxiety/Depression Scale (SAS/SDS) in evaluating anxiety and depression in patients with psoriatic arthritis. Dermatology (Basel) 2020;236(2):170-178. [doi: 10.1159/000498848] [Medline: 31434087]
27. Bollimbala A, James PS, Ganguli S. The impact of physical activity intervention on creativity: role of flexibility vs persistence pathways. Think Skills Creat 2023 Sep;49:101313. [doi: 10.1016/j.tsc.2023.101313]
28. Schmidt-Kraepelin M, Toussaint PA, Thiebes S, Hamari J, Sunyaev A. Archetypes of gamification: analysis of mHealth apps. JMIR Mhealth Uhealth 2020 Oct 19;8(10):e19280. [doi: 10.2196/19280] [Medline: 33074155]
29. Zolfaghari M, Shirmohammadi M, Shahhosseini H, Mokhtaran M, Mohebbi SZ. Development and evaluation of a gamified smart phone mobile health application for oral health promotion in early childhood: a randomized controlled trial. BMC Oral Health 2021 Jan 7;21(1):18. [doi: 10.1186/s12903-020-01374-2] [Medline: 33413304]
30. Cechetti NP, Bellei EA, Biduski D, Rodriguez JPM, Roman MK, De Marchi ACB. Developing and implementing a gamification method to improve user engagement: a case study with an m-Health application for hypertension monitoring. Telematics and Informatics 2019 Aug;41:126-138. [doi: 10.1016/j.tele.2019.04.007]
31. Wang M, Xu J, Zhou X, Li X, Zheng Y. Effectiveness of gamification interventions to improve physical activity and sedentary behavior in children and adolescents: systematic review and meta-analysis. JMIR Serious Games 2025 Sep 18;13:e68151. [doi: 10.2196/68151] [Medline: 40966596]
32. Fernandez-Rio J, de las Heras E, González T, Trillo V, Palomares J. Gamification and physical education. Viability and preliminary views from students and teachers. Phys Educ Sport Pedagogy 2020 Sep 2;25(5):509-524. [doi: 10.1080/17408989.2020.1743253]
33. Gao Y, Zhang J, He Z, Zhou Z. Feasibility and usability of an artificial intelligence-powered gamification intervention for enhancing physical activity among college students: quasi-experimental study. JMIR Serious Games 2025 Mar 24;13:e65498. [doi: 10.2196/65498] [Medline: 40127464]
34. Patel MS, Small DS, Harrison JD, et al. Effectiveness of behaviorally designed gamification interventions with social incentives for increasing physical activity among overweight and obese adults across the United States: the STEP UP randomized clinical trial. JAMA Intern Med 2019 Dec 1;179(12):1624-1632. [doi: 10.1001/jamainternmed.2019.3505] [Medline: 31498375]
35. Romeo A, Edney S, Plotnikoff R, et al. Can smartphone apps increase physical activity? Systematic review and meta-analysis. J Med Internet Res 2019 Mar 19;21(3):e12053. [doi: 10.2196/12053] [Medline: 30888321]
36. Nishi SK, Kavanagh ME, Ramboanga K, et al. Effect of digital health applications with or without gamification on physical activity and cardiometabolic risk factors: a systematic review and meta-analysis of randomized controlled trials. EClinicalMedicine 2024 Oct;76:102798. [doi: 10.1016/j.eclinm.2024.102798] [Medline: 39764571]
37. Li L, Hew KF, Du J. Gamification enhances student intrinsic motivation, perceptions of autonomy and relatedness, but minimal impact on competency: a meta-analysis and systematic review. Education Tech Research Dev 2024 Apr;72(2):765-796. [doi: 10.1007/s11423-023-10337-7]

XSL•FO
RenderX

38. Patricio R, Moreira AC, Zurlo F. Gamification in innovation teams. Int J Innov Stud 2022 Sep;6(3):156-168. [doi: 10.1016/j.ijis.2022.05.003]
39. Wang T, Fan L, Zheng X, et al. The impact of gamification-induced users' feelings on the continued use of mHealth apps: a structural equation model with the self-determination theory approach. J Med Internet Res 2021 Aug 12;23(8):e24546. [doi: 10.2196/24546] [Medline: 34387550]
40. Sotos-Martínez VJ, Ferriz-Valero A, García-Martínez S, Tortosa-Martínez J. The effects of gamification on the motivation and basic psychological needs of secondary school physical education students. Phys Educ Sport Pedagogy 2024 Mar 3;29(2):160-176. [doi: 10.1080/17408989.2022.2039611]
41. Rigby CS, Ryan RM. Self-determination theory in human resource development: new directions and practical considerations. Adv Dev Hum Resour 2018 May;20(2):133-147. [doi: 10.1177/1523422318756954]
42. Yang Y, Goh KY, Teo HH, et al. Compete with me? The impact of online gamified competition on exercise behavior. JAIS 2023;24(3):912-935. [doi: 10.17705/1jais.00806]
43. Zainuddin Z, Chu SKW, Perera CJ. Gamification in education. In: Gamification in a Flipped Classroom: Pedagogical Methods and Best Practices: Springer Nature; 2024:67-113. [doi: 10.1007/978-981-97-2219-8_3]
44. Zhang Q, Yu Z. Meta-analysis on investigating and comparing the effects on learning achievement and motivation for gamification and game-based learning. Education Research International 2022 Aug 1;2022:1-19. [doi: 10.1155/2022/1519880]
45. Luong MLN, Hall M, Bennell KL, Kasza J, Harris A, Hinman RS. The impact of financial incentives on physical activity: a systematic review and meta-analysis. Am J Health Promot 2021 Feb;35(2):236-249. [doi: 10.1177/0890117120940133] [Medline: 32672050]
46. Shi P, Feng X. Motor skills and cognitive benefits in children and adolescents: relationship, mechanism and perspectives. Front Psychol 2022;13:1017825. [doi: 10.3389/fpsyg.2022.1017825] [Medline: 36478944]
47. Fanaroff AC, Patel MS, Chokshi N, et al. Effect of gamification, financial incentives, or both to increase physical activity among patients at high risk of cardiovascular events: the BE ACTIVE randomized controlled trial. Circulation 2024 May 21;149(21):1639-1649. [doi: 10.1161/CIRCULATIONAHA.124.069531] [Medline: 38583084]
48. Egashira M, Son D, Ema A. Serious game for change in behavioral intention toward lifestyle-related diseases: experimental study with structural equation modeling using the theory of planned behavior. JMIR Serious Games 2022 Feb 21;10(1):e28982. [doi: 10.2196/28982] [Medline: 35188465]
49. Zhu Y, Long Y, Wang H, Lee KP, Zhang L, Wang SJ. Digital behavior change intervention designs for habit formation: systematic review. J Med Internet Res 2024 May 24;26:e54375. [doi: 10.2196/54375] [Medline: 38787601]
50. Lakens D. Sample size justification. Collabra 2022 Mar 22;8(1):33267. [doi: 10.1525/collabra.33267]
51. Akinosun AS, Polson R, Diaz-Skeete Y, et al. Digital technology interventions for risk factor modification in patients with cardiovascular disease: systematic review and meta-analysis. JMIR Mhealth Uhealth 2021 Mar 3;9(3):e21061. [doi: 10.2196/21061] [Medline: 33656444]
52. Moshe I, Terhorst Y, Philippi P, et al. Digital interventions for the treatment of depression: a meta-analytic review. Psychol Bull 2021 Aug;147(8):749-786. [doi: 10.1037/bul0000334] [Medline: 34898233]
53. Mora-Gonzalez J, Pérez-López IJ, Esteban-Cornejo I, Delgado-Fernández M. A gamification-based intervention program that encourages physical activity improves cardiorespiratory fitness of college students: "The Matrix rEFvolution Program". Int J Environ Res Public Health 2020 Jan 30;17(3):877. [doi: 10.3390/ijerph17030877] [Medline: 32019259]
54. Chu XL, Zhao XX, Liu SY, et al. Research progress in different physical therapies for treating peripheral nerve injuries. Front Neurol 2025;16:1508604. [doi: 10.3389/fneur.2025.1508604] [Medline: 40260135]
55. Santos PDG, Vaz JR, Correia J, Neto T, Pezarat-Correia P. Long-term neurophysiological adaptations to strength training: a systematic review with cross-sectional studies. J Strength Cond Res 2023 Oct 1;37(10):2091-2105. [doi: 10.1519/JSC.0000000000004543] [Medline: 37369087]
56. Lu Y, Wiltshire HD, Baker JS, Wang Q. The effects of running compared with functional high-intensity interval training on body composition and aerobic fitness in female university students. IJERPH 2021;18(21):11312. [doi: 10.3390/ijerph182111312]
57. Dakin C, Beaulieu K, Hopkins M, Gibbons C, Finlayson G, Stubbs RJ. Do eating behavior traits predict energy intake and body mass index? A systematic review and meta-analysis. Obes Rev 2023 Jan;24(1):e13515. [doi: 10.1111/obr.13515] [Medline: 36305739]
58. Fu J, Sun S, Zhu S, et al. Relationship between 24-h activity behavior and body fat percentage in preschool children: based on compositional data and isotemporal substitution analysis. BMC Public Health 2024 Apr 16;24(1):1063. [doi: 10.1186/s12889-024-18570-2] [Medline: 38627707]
59. Jayedi A, Soltani S, Emadi A, Zargar MS, Najafi A. Aerobic exercise and weight loss in adults: a systematic review and dose-response meta-analysis. JAMA Netw Open 2024 Dec 2;7(12):e2452185. [doi: 10.1001/jamanetworkopen.2024.52185] [Medline: 39724371]
60. Shen QQ, Hou JM, Xia T, et al. Exercise promotes brain health: a systematic review of fNIRS studies. Front Psychol 2024;15:1327822. [doi: 10.3389/fpsyg.2024.1327822] [Medline: 38659667]

XSL•FO
RenderX

61. Huang TY, Chen FT, Li RH, et al. Effects of acute resistance exercise on executive function: a systematic review of the moderating role of intensity and executive function domain. Sports Med Open 2022 Dec 8;8(1):141. [doi: 10.1186/s40798-022-00527-7] [Medline: 36480075]
62. Feng X, Zhang Z, Jin T, Shi P. Effects of open and closed skill exercise interventions on executive function in typical children: a meta-analysis. BMC Psychol 2023 Nov 30;11(1):420. [doi: 10.1186/s40359-023-01317-w] [Medline: 38037184]
63. Kolovelonis A, Goudas M. Acute enhancement of executive functions through cognitively challenging physical activity games in elementary physical education. Eur Phy Educ Rev 2023 May;29(2):268-285. [doi: 10.1177/1356336X221135139]
64. Gkintoni E, Vantaraki F, Skoulidi C, Anastassopoulos P, Vantarakis A. Gamified health promotion in schools: the integration of neuropsychological aspects and CBT—a systematic review. Med Bogota Colomb 2024;60(12):2085. [doi: 10.3390/medicina60122085]
65. Cao C, Cheng S. Does competition in online games foster green behaviour in life? The relationship between gamification affordances and engagement in green behaviours. Int J Hum Comput Interact 2025;41(5):3491-3508. [doi: 10.1080/10447318.2024.2338657]
66. Singh B, Olds T, Curtis R, et al. Effectiveness of physical activity interventions for improving depression, anxiety and distress: an overview of systematic reviews. Br J Sports Med 2023 Sep;57(18):1203-1209. [doi: 10.1136/bjsports-2022-106195] [Medline: 36796860]
67. Wanjau MN, Möller H, Haigh F, et al. Physical activity and depression and anxiety disorders: a systematic review of reviews and assessment of causality. AJPM Focus 2023 Jun;2(2):100074. [doi: 10.1016/j.focus.2023.100074] [Medline: 37790636]
68. Recchia F, Bernal JDK, Fong DY, et al. Physical activity interventions to alleviate depressive symptoms in children and adolescents: a systematic review and meta-analysis. JAMA Pediatr 2023 Feb 1;177(2):132-140. [doi: 10.1001/jamapediatrics.2022.5090] [Medline: 36595284]
69. Jeong J, Pitchik HO, Fink G. Short-term, medium-term and long-term effects of early parenting interventions in low- and middle-income countries: a systematic review. BMJ Glob Health 2021 Mar;6(3):e004067. [doi: 10.1136/bmjgh-2020-004067] [Medline: 33674266]
70. Chaput JP, McHill AW, Cox RC, et al. The role of insufficient sleep and circadian misalignment in obesity. Nat Rev Endocrinol 2023 Feb;19(2):82-97. [doi: 10.1038/s41574-022-00747-7] [Medline: 36280789]
71. Almeqdad QI, Alodat AM, Alquraan MF, Mohaidat MA, Al-Makhzoomy AK. The effectiveness of universal design for learning: A systematic review of the literature and meta-analysis. Cogent Education 2023 Dec 31;10(1):2218191. [doi: 10.1080/2331186X.2023.2218191]
72. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine--reporting of subgroup analyses in clinical trials. N Engl J Med 2007 Nov 22;357(21):2189-2194. [doi: 10.1056/NEJMsr077003] [Medline: 18032770]
73. Lund JI, Toombs E, Radford A, Boles K, Mushquash C. Adverse childhood experiences and executive function difficulties in children: a systematic review. Child Abuse Negl 2020 Aug;106:104485. [doi: 10.1016/j.chiabu.2020.104485] [Medline: 32388225]
74. Damaševičius R, Maskeliūnas R, Blažauskas T. Serious games and gamification in healthcare: a meta-review. Information 2023;14(2):105. [doi: 10.3390/info14020105]
75. Treacy C, Levenstein JM, Jefferies A, et al. The LEISURE study: a longitudinal randomized controlled trial protocol for a multi-modal lifestyle intervention study to reduce dementia risk in healthy older adults. J Alzheimers Dis 2023;94(2):841-856. [doi: 10.3233/JAD-230193] [Medline: 37334601]
76. Ludyga S, Held S, Rappelt L, Donath L, Klatt S. A network meta-analysis comparing the effects of exercise and cognitive training on executive function in young and middle-aged adults. Eur J Sport Sci 2023 Jul;23(7):1415-1425. [doi: 10.1080/17461391.2022.2099765] [Medline: 35848068]
77. Gkintoni E, Vassilopoulos SP, Nikolaou G. Next-generation cognitive-behavioral therapy for depression: integrating digital tools, teletherapy, and personalization for enhanced mental health outcomes. Medicina (Kaunas) 2025 Feb 28;61(3):431. [doi: 10.3390/medicina61030431] [Medline: 40142242]
78. Victoria LW, Oberlin LE, Ilieva IP, et al. A digital intervention for cognitive deficits following COVID-19: a randomized clinical trial. Neuropsychopharmacology 2024 Dec;50(2):472-479. [doi: 10.1038/s41386-024-01995-z] [Medline: 39358543]
79. Wilson K, St-Onge MP, Tasali E. Diet composition and objectively assessed sleep quality: a narrative review. J Acad Nutr Diet 2022 Jun;122(6):1182-1195. [doi: 10.1016/j.jand.2022.01.007] [Medline: 35063665]
80. Tschon M, Contartese D, Pagani S, Borsari V, Fini M. Gender and sex are key determinants in osteoarthritis not only confounding variables. A systematic review of clinical data. J Clin Med 2021 Jul 19;10(14):3178. [doi: 10.3390/jcm10143178] [Medline: 34300344]

Abbreviations

BMR: basal metabolic rate
CG: control croup
EF: executive function
IG: intervention group

XSL•FO
RenderX

IRB: institutional review board
ITT: intention-to-treat
MCAR: missing completely at random
MET-min: metabolic equivalent-minute
mHealth: mobile health
MVPA: moderate-to-vigorous physical activity
NIH: National Institutes of Health
PA: physical activity
RCT: randomized controlled trial
SAS: Self-Rating Anxiety Scale
SDS: Self-Rating Depression Scale
SDT: self-determination theory
TPB: theory of planned behavior
WHO: World Health Organization

Edited by S Brini; submitted 21.Aug.2025; peer-reviewed by F Coelho, K Peng, W Vieira; revised version received 04.Feb.2026; accepted 06.Mar.2026; published 07.Apr.2026.

Please cite as:
Yang M, Guo Y, Li Z, Jiang L, Gao Y, Yang S, Zhou Z
A Gamified Mobile Health Intervention to Promote Physical Activity, Executive Function, and Mental Health in College Students: Randomized Controlled Trial
J Med Internet Res 2026;28:e82769
URL: https://www.jmir.org/2026/1/e82769
doi:10.2196/82769

© Ming Yang, Yingjun Guo, Zhen Li, Lei Jiang, Yanan Gao, Sen Yang, Zhixiong Zhou. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Preferences for Personalized Text Message Appointment Reminders Among Outpatients in a Universal Health System: Cross-Sectional Study

Shih-Huai Hsiao[1,2,3], MPH; Yu-Chen Chuang[4*], MSc; Shu-Ling Ma[5], MSc; Hui-Tzu Lin[6], MPH; I-Chen Lee[7*], PhD; Pei-Shih Chen[2], PhD

[1]Tele-Healthcare Center, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan

[2]Department of Public Health, College of Health Science, Kaohsiung Medical University, No. 100 Shih-Chuan 1st Road, Sammin District, Kaohsiung, Taiwan

[3]Taiwan Health Insurance Association, Taipei, Taiwan

[4]Cardiovascular Division, Department of Internal Medicine, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan

[5]Department of Nursing, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan

[6]Clinical Skill Training Center, Department of Clinical Education and Skill Training, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan

[7]Department of Healthcare Administration and Medical Informatics, College of Health Sciences, Kaohsiung Medical University, Kaohsiung, Taiwan

[*]these authors contributed equally

Corresponding Author:
Pei-Shih Chen, PhD
Department of Public Health, College of Health Science, Kaohsiung Medical University, No. 100 Shih-Chuan 1st Road, Sammin District, Kaohsiung, Taiwan

Abstract

Background: SMS text messaging reminders are widely used to reduce missed outpatient appointments; however, evidence remains limited regarding which types of reminder content patients prefer, particularly within East Asian universal health systems. In Taiwan, minimal financial barriers to care and unrestricted access to secondary and tertiary hospitals contribute to high outpatient visit volumes and persistent no-show rates. These contextual features underscore the need for behaviorally informed and demographically tailored reminder strategies rather than uniform messaging approaches.

Objective: This study aimed to examine patient preferences for 6 theory-guided SMS appointment reminder types and to identify the predictors of reminder preference related to demographic characteristics and health care utilization, with the goal of informing personalized reminder design for a forthcoming randomized controlled trial.

Methods: We conducted a cross-sectional online survey among adults in Taiwan with prior outpatient experience. Six SMS reminder prototypes were developed based on behavioral communication principles and validated by a multidisciplinary expert panel using item-level content validity indices. Participants selected their preferred SMS reminder type and reported sociodemographic characteristics and recent health care utilization. Bivariate associations were examined using chi-square tests and one-way ANOVA, with Benjamini-Hochberg false discovery rate correction applied to control for multiple testing. To identify independent predictors of SMS reminder preference while adjusting for potential confounding, we fitted a multinomial logistic regression model with all covariates entered simultaneously.

Results: A total of 1095 respondents completed the survey. General reminders and messages referencing prior missed appointments were most frequently preferred, whereas empathy-based or relationally framed messages were selected less often. In false discovery rate–adjusted univariate analyses, both age and sex were associated with SMS reminder preference. However, in the fully adjusted multinomial logistic regression model, age emerged as the only statistically significant independent predictor. Participants younger than 50 years were significantly more likely to prefer alternative reminder message types compared with the general reminder (adjusted odds ratio 1.64, 95% CI 1.18 - 2.28; P=.003). Sex did not retain statistical significance after multivariable adjustment. Other sociodemographic characteristics and health care utilization variables, including education level, employment status, residential region, outpatient visit frequency, and recent missed appointments history, were not independently associated with reminder preference.

Conclusions: Preferences for outpatient SMS reminder content vary systematically, with age representing the most robust independent predictor. Across the sample, concise and behavior-focused reminders were preferred over empathy-oriented or relational formats. These findings support age-informed tailoring of SMS reminder content and provide content-validated SMS prototypes for use in subsequent interventional research. The results offer formative evidence to guide the design of randomized

XSL•FO
RenderX

trials aimed at reducing outpatient no-shows and improving the efficiency of ambulatory care delivery in Taiwan's universal health care system.

(J Med Internet Res 2026;28:e81324) doi:10.2196/81324

KEYWORDS

SMS reminders; outpatient no-shows; patient preferences; behavioral design; health communication; personalized messaging; universal health coverage

Introduction

Timely attendance at outpatient appointments is fundamental to maintaining care continuity, optimizing health care resource utilization, and achieving favorable clinical outcomes [1]. However, missed appointments (MAs), commonly referred to as "no-shows," remain a persistent global challenge, with no-show rates in outpatient settings averaging approximately 23% [2]. These disruptions not only compromise the efficiency of health care delivery but also contribute to inequities in access and outcomes, particularly among patients facing socioeconomic disadvantages, transportation barriers, or prolonged wait times in specialty care services, such as cardiology and dermatology [3-6].

To address this issue, SMS reminders have been widely implemented and demonstrated to be effective in reducing no-show rates across various clinical contexts, including gastrointestinal procedures [7-9] and surgeries requiring anesthesia [10]. Compared with traditional telephone calls, which are labor-intensive and less scalable, SMS reminders offer a cost-effective and scalable solution for promoting appointment adherence [11]. Moreover, behaviorally informed message designs, such as emotionally salient language, appeals to social responsibility, or moral framing, have been shown to enhance patient engagement and accountability [7,12-14]. The inclusion of personalized content, such as references to prior MAs, has also been associated with further reductions in no-show rates [10,15,16].

Despite the growing body of evidence supporting the efficacy of SMS reminders, limited research has examined the optimal design of message content. Specifically, there is a lack of systematic investigation into which core components should be included, how linguistic and emotional framing influences patient behavior, and to what extent content should be tailored to individual characteristics. These characteristics may include insurance type, age, occupation, socioeconomic status, prior attendance history, disease severity, or the need for chronic follow-up care. Personalization strategies of this nature are likely critical to maximizing the impact of reminder interventions but have been insufficiently explored in the literature.

This gap is particularly salient in East Asian health care systems, such as Taiwan's, where universal health coverage and unrestricted access to secondary and tertiary care may lead to distinct patterns of health care utilization. Under Taiwan's National Health Insurance program, patients can access higher-level hospitals directly without referral and with minimal out-of-pocket expense, often resulting in frequent outpatient visits for relatively minor conditions [17]. The ease of access and low perceived urgency of such visits may contribute to elevated no-show rates; one tertiary medical center in Taiwan reported a no-show rate as high as 16.8% [18]. These contextual factors highlight the need for more effective and targeted reminder strategies that are responsive to patient behavior and health care system dynamics.

Specifically, the study aimed to (1) assess whether there was significant heterogeneity in preferences for various types of SMS content and (2) examine the extent to which these preferences were associated with patients' sociodemographic characteristics, socioeconomic status, and recent health care utilization patterns.

Methods

Study Design and Participant Recruitment

This cross-sectional study used a nonprobability convenience sampling method. Participants were recruited through online platforms (eg, Facebook, LINE) and professional networks, including outreach by health care professionals.

Participants were eligible if they were adults aged 20 years or older, resided in Taiwan, had experience using SMS-capable mobile phones, and completed the online questionnaire with informed consent. Individuals were excluded if they lacked SMS access, had known cognitive impairments, or reported working in health care–related professions (eg, hospitals or clinics) to minimize potential response bias. The structured questionnaire was distributed electronically, and the data were collected anonymously over a 14-day period from February 25 to March 10, 2025. A total of 1152 responses were received, of which 1095 were valid, yielding a valid response rate of 95.05%.

Instrument Development and Content Validity

SMS Reminder Messages Development

Grounded in behavioral theory and prior empirical research, this study developed 6 SMS appointment reminder prototypes that incorporate cognitive, emotional, and social cues to enhance patient relevance and behavioral efficacy. Message design was guided by the SMS content framework proposed by Acharya et al [19], with 4 core principles applied: the use of plain language, a neutral tone, explicit appointment details, and clear sender identification. Three principles, namely the inclusion of external links, preparation instructions, and emphasis on the consequences of nonattendance, were considered less applicable to general outpatient contexts. While the prototypes largely conformed to established design standards, contextual adaptation remains essential when applying screening-focused frameworks

XSL•FO
RenderX

to broader clinical settings. The 6 SMS message types are summarized in Table 1 and are described as follows:

1. General reminder: a neutral notification containing appointment details [9-11,20].
2. Prior MA: referencing the patient's history of missed visits and institutional policies. This type can further reduce the no-show rate by approximately 1% compared to standard reminders [14] and promote behavioral change through increased awareness of institutional norms [10].
3. Provider-focused empathy: highlighting the emotional burden and workload of health care staff, aiming to evoke moral responsibility and cooperation, especially in high-burden clinical settings [12].
4. Patient-focused empathy: emphasizing how MA may limit others' access to care, particularly under Taiwan's hospital-specific global budgeting system, where physician time and service volume are constrained. Such awareness may foster social responsibility and encourage patients to proactively cancel or reschedule. Empathy-based reminders have also been shown to reduce provider-patient tension [21].
5. Physician follow-up: simulating concern from the attending physician by referencing the patient's upcoming visit and inviting them to return on the scheduled date. This tone, regardless of prior clinical contact, may enhance engagement and continuity of care, especially for patients with chronic conditions or long-term follow-up needs [22].
6. Supportive reminder: using friendly, conversational language to reduce psychological barriers and ease anxiety related to clinic visits. Warm, supportive messaging may improve follow-up motivation and engagement, particularly for patients who perceive medical visits as stressful [14].

The 6 SMS messages in their original Chinese form have been included in Multimedia Appendix 1.

Table . Theory-guided SMS reminder types with message content and technical specifications for outpatient appointment adherence.

Type of SMS reminder	Wording of SMS content in this study
General reminders	Hi! You have an appt. with [Dr Name] on the [Morning/Afternoon/Evening] of [Date]. To cancel or reschedule, please visit [clinic or hospital] booking system. We care about your health!
Prior missed appointment behavior	Hi! You have an appt. with [Dr Name] on the [Morning/Afternoon/Evening] of [Date]. Your Appt. No. is [Appt. No.]. If this MA results in a total of three no-shows without prior cancellation, you will no longer be eligible for remote booking. Future appt. must be made in person at the [clinic or hospital]. To cancel or change, please visit [clinic or hospital] booking system. We care about your health!
Joint empathy communication type 1: provider-focused	Hi! You have an appt. with [Dr Name] on the [Morning/Afternoon/Evening] of [Date]. Your Appt. No. is [Appt. No.]. If you don't show up and don't cancel, it may cause trouble for medical staff. To cancel or reschedule, please visit [clinic or hospital] booking system. We care about your health!
Joint empathy communication type 2: patient-focused	Hi! You have an appt. with [Dr Name] on the [Morning/Afternoon/Evening] of [Date]. Your Appt. No. is [Number]. If you miss your appt. without cancellation, it may affect other patients' access to care. To cancel or reschedule, please visit [clinic or hospital booking system]. We care about your health!
Gentle and friendly (physician follow-up)	Hi! [Dr Name] cares about you. You have a [specialty] appt. on the [Morning/Afternoon/Evening] of [Date]. To cancel or reschedule, please visit [clinic or hospital] booking system. We care about your health!
Supportive reminder	Hope you're doing well! [Dr Name] cares about your health. You have a [specialty] appt. on the [Morning/Afternoon/Evening] of [Date]. Your Appt. No. is [number]. To cancel or reschedule, please visit [clinic or hospital] booking system. We care about your health!

Content Validity Examination of SMS Message

To evaluate the content validity of the SMS messages, a multidisciplinary panel of 12 experts was convened. The panel included 2 clinicians, 3 managers from patient registration and cashier services, 2 language educators, 3 nursing supervisors responsible for outpatient care, and 2 professionals from insurance and customer service departments. All panel members had prior personal experience as patients or caregivers and demonstrated strong communication competencies, ensuring a broad range of perspectives from clinical, administrative, linguistic, and consumer viewpoints. No significant demographic differences were found among the expert subgroups (P>.05).

Each SMS prototype was assessed using a 5-point Likert scale across 3 domains: semantic clarity, content relevance, and ease of comprehension. The item-level content validity index for all 6 messages exceeded the standard threshold of 0.78 [17,23], confirming adequate content validity.

Statistical Analysis

Descriptive statistics summarized participant characteristics. Chi-square tests and one-way ANOVA were used to examine bivariate associations between participant characteristics and

XSL•FO
RenderX

SMS reminder preferences. To control for inflated type 1 error due to multiple comparisons, the Benjamini-Hochberg false discovery rate (FDR) procedure was applied to the bivariate tests [24].

To adjust for potential confounding and identify independent predictors of SMS reminder preference, we fitted a multinomial logistic regression model with SMS reminder preference as the outcome. The general reminder SMS was specified as the reference category. Demographic characteristics and health care utilization variables were entered simultaneously as covariates, and the results are reported as adjusted odds ratios with 95% CI. Statistical significance was set at $P<.05$ (2-tailed). All analyses were conducted using SPSS (version 29.0.1.0).

Ethical Considerations

This study was reviewed and approved by the Institutional Review Board of Kaohsiung Medical University Chung-Ho Memorial Hospital (IRB number: KMUHIRB-E(I)-20240047; approval date: February 3, 2025). Electronic informed consent was obtained from all participants prior to survey submission. Participation was voluntary, anonymity was ensured, and no identifiable personal information was collected. All study procedures adhered to institutional ethical guidelines and the principles of the Declaration of Helsinki.

Results

Participant Characteristics

Demographic and Socioeconomic Status

Among the 1095 participants, female participants predominated (n=851, 77.7%). The mean age was 43.2 years (SD 12.37), with 65.8% (n=720) aged below 50 years. Most (n=726, 66.3%) respondents held a university or college degree, followed by postgraduate qualifications (n=272, 24.8%), and secondary-level education or below (n=97, 8.9%).

The largest occupational groups were professionals (n=352, 32.1%) and clerical and administrative staff (n=254, 23.2%), with smaller proportions in sales or service roles, technical support, senior management, and unemployed categories. Nearly half (n=515, 47%) resided in the Kaohsiung-Pingtung region, and the majority (n=800, 73.1%) lived with others. A minority (n=216, 19.7%) were unmarried or living alone (n=79, 7.2%; Table 2).

XSL•FO
RenderX

Table . Demographic and socioeconomic distribution of the respondents.

Variable	Respondents
Sex, n (%)	
Male	244 (22.3)
Female	851 (77.7)
Age (y), mean (SD)	43.2 (12.38)
Age group (y), n (%)	1095 (100)
<50	720 (65.8)
≥50	375 (34.2)
Education level, n (%)	
Senior high school or lower education	97 (8.9)
College or university education	726 (66.3)
Postgraduate degree	272 (24.8)
Occupation category, n (%)	
Senior executives and decision-makers	114 (10.4)
Professionals	352 (32.1)
Technicians and assistants	96 (8.8)
Clerical and administrative staff	254 (23.2)
Sales and service workers	147 (13.4)
Other occupations	54 (4.9)
Unemployed	78 (7.1)
Employment status, n (%)	
Retired	109 (10)
Employed	897 (81.9)
Unemployed	89 (8.1)
Region of residence, n (%)	
Kaohsiung-Pingtung region	515 (47)
Outside Kaohsiung-Pingtung region	580 (53)
Living arrangement, n (%)	
Unmarried	216 (19.7)
Living alone	79 (7.2)
Living with others	800 (73.1)

XSL•FO
RenderX

Variable	Respondents
Outpatient visits in the past 3 months, n (%)	
0 visit	192 (17.5)
1 - 3 visits	658 (60.1)
4 or more visits	245 (22.4)
Missed outpatient appointment in the past 3 months, n (%)	
No missed appointment	923 (84.3)
Missed appointment at least once	172 (15.7)
Frequency of missed appointments in the past 3 months, n (%)	
Fewer than 3 times	1074 (98.1)
Three times or more	21 (1.9)

Health Care Utilization

In the preceding 3 months, 60.1% (n=658) of the participants had 1 - 3 outpatient visits, 22.4% (n=245) reported 4 or more visits, and 17.5% (n=192) had not accessed outpatient services. Additionally, 15.7% (n=172) reported at least 1 MA, while 84.3% (n=923) had no such record.

Of note, 17.5% (n=192) of the participants reported no outpatient visits in the past 3 months, yet 15.7% (n=172) indicated at least 1 MA. This discrepancy may reflect recall bias or differences in participants' interpretation of MAs, such as counting canceled or forgotten appointments that were never formally registered as visits.

Distribution of SMS Reminder Preferences

As shown in Table 3, nearly half (n=520, 47.5%) of the respondents preferred reminders referencing "Prior MAs" behavior, followed by general reminders (n=341, 31.1%). Fewer participants selected messages with an empathetic tone, including provider-focused empathy (n=48, 4.4%) and patient-focused empathy (n=60, 5.5%). Relational reminders, such as physician follow-up (n=53, 4.8%) and supportive reminders (n=73, 6.7%), accounted for a combined total of 11.5% of the preferences. These results indicate a predominant preference for practical and behavior-linked reminder formats over affective or relational messaging.

Because 4 SMS categories (provider-focused empathy, patient-focused empathy, physician follow-up, and supportive reminders) were selected by relatively few participants, the corresponding subgroup cell sizes in the chi-square analyses were small. Therefore, statistical comparisons involving these categories should be interpreted with caution.

Male respondents showed relatively higher proportions selecting supportive or relational message types, although these differences were not statistically significant after FDR correction.

Table . Distribution of preferred SMS reminder types (N=1095).

SMS reminder type	Values, n (%)
Prior missed appointment reminder	520 (47.5)
General reminder	341 (31.1)
Supportive reminder	73 (6.7)
Patient-focused empathy	60 (5.5)
Physician follow-up	53 (4.8)
Provider-focused empathy	48 (4.4)
Total	1095 (100)

XSL•FO
RenderX

Associations Between Demographic, Socioeconomic, and Health Care Utilization Characteristics and Preferred SMS Reminder Type

Analyses of demographic and health care utilization characteristics showed several initial associations with SMS reminder type preferences. As summarized in the cross-tabulations, gender, age group, education level, occupation, residential region, and living arrangement each demonstrated significant differences across message categories in unadjusted chi-square and ANOVA analyses (Table 4). For example, women and younger adults more frequently selected general or prior missed appointment reminders, whereas older adults showed comparatively higher proportions selecting empathetic or supportive formats. Occupational variation was also observed, with clerical and professional workers tending toward standard reminders, while senior executives reported a greater share of emotionally supportive preferences. These subgroup differences, however, did not persist following correction for multiple comparisons.

Health care utilization variables exhibited similar patterns. Participants with 1 to 3 outpatient visits in the past 3 months tended to prefer neutral, task-oriented reminders, while those with at least 1 recent MA were more likely to choose messages explicitly referencing prior absences (Table 5).

Table . Associations between demographic characteristics and preferred SMS reminder type.

Variable	General reminder, n (%)	Prior missed appointment behavior, n (%)	Provider-focused empathy, n (%)	Patient-focused empathy, n (%)	Physician follow-up, n (%)	Supportive reminder, n (%)	Chi-square or *F* test (*df*)	*P* value
Sex							26.554[a] (5)	.001
Male	67 (19.6)	102 (19.6)	12 (25.0)	21 (35.0)	11 (20.8)	31 (42.5)		
Female	274 (80.4)	418 (80.4)	36 (75.0)	39 (65.0)	42 (79.2)	42 (57.5)		
Age (y)								
Mean age (SD)	45.82 (12.43)	40.05 (12.33)	43.13 (12.41)	45.62 (11.93)	50.45 (11.19)	45.53 (13.39)	19.54[b]	.001
<50, n (%)	197 (57.8)	395 (76.0)	30 (62.5)	32 (53.3)	24 (45.3)	42 (57.5)		
≥50, n (%)	144 (42.2)	125 (24.0)	18 (37.5)	28 (46.7)	29 (54.7)	31 (42.5)		
Education								
Senior high or less	28 (8.2)	46 (8.8)	3 (6.3)	8 (13.3)	8 (15.1)	8 (11.0)	5.977[b] (10)	.82
College or university	225 (66.0)	356 (68.5)	31 (64.6)	35 (58.3)	35 (66.0)	44 (60.3)		
Postgraduate	88 (25.8)	118 (22.7)	14 (29.2)	17 (28.3)	14 (26.4)	21 (28.8)		

[a]Chi-square test.

[b]*F* test.

XSL•FO
RenderX

Table . Associations between health care utilization characteristics and preferred SMS reminder type.

Variable	General reminder, n (%)	Prior MA[a] behavior, n (%)	Provider-focused empathy, n (%)	Patient-focused empathy, n (%)	Physician follow-up, n (%)	Supportive reminder, n (%)	Chi-square (*df*)	*P* value
Outpatient visits (past 3 months)							5.112 (10)	.88
0 visits	56 (5.1)	89 (8.1)	10 (9.4)	12 (11.0)	11 (10.4)	14 (13.3)		
1 - 3 visits	201 (58.9)	315 (60.6)	31 (64.6)	37 (61.7)	33 (62.3)	41 (56.2)		
≥4 visits	84 (24.6)	116 (22.3)	7 (14.6)	11 (18.3)	9 (17.0)	18 (24.7)		
MAs (past 3 months)							3.267 (5)	.66
None	286 (83.9)	446 (85.8)	41 (85.4)	50 (83.3)	44 (83.0)	57 (78.1)		
≥1 MA	56 (16.4)	74 (14.2)	7 (14.6)	10 (16.7)	9 (17.0)	16 (21.9)		
Frequency of MAs							10.464 (5)	.06
<3 MAs	331 (97.1)	490 (94.2)	47 (97.9)	59 (98.3)	52 (98.1)	69 (94.5)		
≥3 MAs	10 (2.9)	4 (0.8)	1 (2.1)	1 (1.7)	1 (1.9)	4 (5.5)		

[a]MA: missed appointment.

To address the potential inflation of type 1 error due to multiple comparisons, the Benjamini-Hochberg FDR correction was applied across all demographic and utilization variables. After adjustment, only age and gender remained statistically significant predictors of SMS preference. All other variables, including education, occupation, region, living arrangement, outpatient visit frequency, and missed appointment history, no longer met the adjusted significance threshold. Multinomial logistic regression yielded consistent results, confirming age and sex as the only independent predictors after controlling for all covariates.

Although initial chi-square and ANOVA results suggested that several demographic and health care utilization variables were associated with SMS reminder preferences, these patterns did not remain statistically stable after controlling for multiple comparisons. After applying the Benjamini-Hochberg FDR correction, only age and sex remained significant predictors, whereas education level, occupation, residential region, living arrangement, outpatient visit frequency, and missed appointment history were no longer statistically significant (Table 6).

Table . Summary of false discovery rate (FDR)–corrected significance across all demographic and health care utilization variables.

Variable	Chi-square and *F* test (*df*)	Original *P* value	FDR-adjusted *P* value	Significant after FDR?
Sex	26.554 (df)[a]	.001	.006	Yes
Age (y; continuous)	19.54 (df)[b]	.001	.003	Yes
Education	5.977 (df) [a]	.82	.98	No
Outpatient visits (past 3 months)	5.112 (df)[a]	.88	.88	No
Missed appointments (past 3 months)	3.267 (df)[a]	.66	.99	No
Frequency of missed appointments	10.464 (df)[a]	.06	.13	No

[a]Chi-square.

[b]*F* test.

To identify independent predictors of SMS reminder preference while simultaneously adjusting for all covariates, a multinomial logistic regression model was constructed. The full results of this multivariable analysis are presented in Table 7, including adjusted odds ratios (aORs), 95% CI, and *P* values.

As shown in Table 7, the multinomial logistic regression analysis identified age as the only statistically significant independent predictor of SMS reminder preference after simultaneous adjustment for all covariates. Participants younger than 50 years were significantly more likely to prefer alternative reminder message types compared with the general reminder SMS (aOR 1.64, 95% CI 1.18 - 2.28; *P*=.003). In contrast, sex did not remain independently associated with reminder preference after multivariable adjustment (aOR 1.23, 95% CI 0.87 - 1.73; *P*=.24). Other sociodemographic characteristics and health care utilization variables, including education level, employment status, residential region, living arrangement, outpatient visit frequency, and recent MAs history, were not statistically significant in the fully adjusted model.

XSL•FO
RenderX

Table . Multinomial logistic regression analysis of factors associated with SMS reminder preference[a].

Predictor and category	aOR[b]	95% CI	*P* value
Sex			
Male (vs female)	1.23	0.87-1.73	.24
Age group			
<50 years (vs ≥50 years)	1.64	1.18-2.28	.003
Education level			
Less than or equal to high school (vs graduate+)	1.58	0.90-2.78	.11
College (vs graduate+)	1.12	0.81-1.55	.50
Employment status			
Retired (vs unemployed)	1.33	0.64-2.78	.44
Employed (vs unemployed)	1.65	0.90-3.01	.11
Residential region			
Kaoping region (vs non-Kaoping)	0.79	0.60-1.03	.08
Living arrangement			
Unmarried (vs with others)	1.31	0.91-1.89	.15
Living alone (vs with others)	1.21	0.72-2.06	.47
Outpatient visits (past 3 months)			
0 visits (vs ≥4)	1.19	0.78-1.81	.42
1-3 visits (vs ≥4)	1.15	0.84-1.58	.40
Missed appointments (past 3 months)			
≤3 times (vs ≥3)	2.08	0.86-5.04	.10

[a]Outcome variable: preference for SMS reminder type. Reference category: general reminder SMS. Model: multinomial logistic regression (all covariates entered simultaneously).

[b]aOR: adjusted odds ratio.

Discussion

Principal Findings

The primary objective of this study was to examine public preferences regarding SMS reminder content in preparation for a forthcoming large-scale randomized controlled trial. In the planned randomized controlled trial, participants will be randomly allocated to either an intervention group, which will receive a standardized SMS reminder seven days prior to their scheduled appointment containing details such as the appointment date, time, department, and queue number, or a control group that receives no reminder. Insights from this study will inform the development of effective SMS content strategies to enhance outpatient appointment adherence. The following sections address several key issues that merit further discussion.

Sample Representativeness

The observed preference for concise, behaviorally framed SMS reminders, particularly the general reminder and prior MA formats, was most prominent among women, younger adults, and highly educated respondents. These patterns align with the demographic profile of the study sample, in which the majority were female (n=851, 77.7%), under age 50 (n=720, 65.8%), and held a college degree or above (91.1%).

Prior literature indicates that women are more likely than men to assume caregiving responsibilities within households, maintain greater engagement with preventive health services, and demonstrate higher willingness to participate in self-administered online questionnaires related to health behaviors and health care use [25-27]. These factors may contribute to their increased representation in digital health surveys.

Furthermore, women consistently report higher health literacy and mobile engagement for health-related purposes, including appointment scheduling, medication reminders, and symptom monitoring, compared with men of the same age groups [28]. Studies have also shown that women exhibit stronger preferences for receiving health information via SMS or other asynchronous communication channels, which may further elevate their likelihood of responding to an online survey focused on SMS reminder content. In contrast, men tend to participate at lower rates in health communication surveys unless participation is tied to an active medical condition or clinical follow-up requirement [29].

Collectively, the predominance of women in this sample likely reflects well-documented patterns in digital health engagement rather than a study-specific recruitment artifact. Nevertheless,

XSL•FO
RenderX

the skewed sex distribution requires the cautious interpretation of aggregated preference patterns. To address this concern, sex was explicitly incorporated as a key covariate in the multinomial logistic regression analysis. After multivariable adjustment, sex did not remain independently associated with SMS reminder preference, indicating that the observed preference patterns were not solely driven by sample composition.

Potential Data Inconsistency in Appointment and Missed Visit Reporting

A minor discrepancy was observed between outpatient visit reports and MA data. Specifically, 17.5% (n=192) of the respondents reported no outpatient visits in the past 3 months, but 15.7% (n=172) reported at least 1 MA. This inconsistency may reflect recall bias or variable interpretations of MAs, such as canceled or forgotten visits not formally captured by hospital systems. Our findings echo those of de Reuver and Bouwman [30], who observed that type 1 and type 2 reporting errors are common in self-reported utilization data.

Preference Heterogeneity

The chi-square analyses indicated that participants most frequently selected general reminders and prior MA messages, while empathy-oriented or relationally framed messages were selected far less often. This suggests that patients prioritize clarity and actionable information over affective framing when receiving appointment-related communication. These findings are consistent with Hallsworth et al [16], who demonstrated that reminders containing cost-related information and clear behavioral cues were more effective than emotionally framed messages.

Preferences Differ by Sociodemographic and Health Care Utilization Characteristics

The results demonstrated meaningful variation in SMS reminder preferences by age and sex, whereas other demographic and behavioral characteristics did not show stable associations after correction for multiple testing.

After correction for multiple testing, age and sex remained statistically significant in FDR-adjusted univariate analyses. However, when all covariates were entered simultaneously into the multinomial logistic regression model, age emerged as the only independent predictor of SMS reminder preference. Sex did not retain statistical significance after multivariable adjustment.

After multivariable adjustment, age emerged as the only independent predictor of SMS reminder preference, whereas sex did not retain statistical significance. Other variables, including education, occupation, residential region, living arrangement, outpatient visit frequency, and recent MAs, lost statistical significance following the FDR adjustment.

Descriptively, men demonstrated a more even distribution of preferences across message types; however, these differences did not remain statistically significant after multivariable adjustment and should be interpreted cautiously.

These findings indicate that designing personalized SMS strategies based primarily on age and sex may be more appropriate than tailoring messages by education level or utilization history. This pattern is consistent with Crutchfield et al [31] and Mohammed Selim et al [32], who observed that sociodemographic characteristics influence responsiveness to reminder messages. An Israeli randomized controlled trial further found that older adults and those with prior MAs were more responsive to emotionally framed messages [33]. Junod Perron et al [34], however, emphasized that communication mode (SMS vs telephone) may influence engagement differently than message framing.

Because multiple chi-square tests were conducted, the Benjamini-Hochberg FDR procedure was applied to reduce the risk of type 1 error. After correction, only age and sex remained statistically significant predictors of SMS preference. Residential region, living arrangement, and all health care utilization variables, including outpatient visit frequency and recent MAs, were not significant after correction.

These results were corroborated by the multinomial logistic regression model, in which age and sex remained the only independent predictors after mutual adjustment. Overall, these findings indicate that although unadjusted analyses suggested broader subgroup differences, only age- and sex-based differences demonstrated sufficient stability to warrant consideration in tailored SMS design.

It should be noted that several message types, particularly empathy-based formats, had limited subgroup counts. The observed associations for these categories may therefore lack statistical robustness and should be interpreted conservatively. The revised text avoids overstating preference differences in low-frequency groups.

Content Validity Assessment

To ensure the appropriateness and clarity of SMS content prior to trial implementation, this study used item-level content validity indices based on 9 communication design principles derived from existing literature [17,19,23]. A multidisciplinary expert panel of 12 evaluators assessed the messages, exceeding recommended minimum panel sizes for content validation.

Although the panel did not include formal patient representatives, all members had extensive experience interacting with patients in outpatient or hospital settings. These perspectives provided practical insights grounded in real-world communication processes. This limitation is acknowledged, yet expert-based content validation remains standard practice in early formative research.

Rationale for Age Cutoff at 50 Years

The decision to dichotomize age at 50 years was justified by both statistical and contextual considerations. Statistically, this threshold yielded significant differences in SMS preferences. Contextually, adults around age 50 years in Taiwan frequently assume dual caregiving responsibilities for both aging parents and children, which may influence their engagement with health care communication [35].

XSL•FO
RenderX

Limitations

This study relied on convenience sampling through an online survey, which may limit representativeness. The predominance of female respondents reflects a well-documented pattern in online and digital health surveys, in which women are more likely than men to participate, particularly in studies related to health behaviors and health care communication [36,37]. To address this imbalance, sex was explicitly incorporated as a covariate in both univariate analyses and the multinomial logistic regression model. After multivariable adjustment, sex did not remain independently associated with SMS reminder preference, suggesting that the observed patterns were not solely driven by sample composition. Nevertheless, absolute preference proportions should be interpreted with caution, particularly for male patients. Future studies may enhance representativeness through targeted recruitment strategies and the inclusion of more inclusive gender response options, in line with emerging best practices in survey methodology [38].

Conclusion

This study demonstrates that preferences for SMS appointment reminder content vary systematically among outpatients, with age emerging as the most robust independent predictor. General reminders and prior missed appointment messages were most consistently preferred, whereas empathy-based or relational formats were selected far less frequently. After applying FDA correction and conducting multinomial logistic regression, age remained the only stable independent predictor of SMS reminder preference, whereas sex did not retain statistical significance after multivariable adjustment. Beyond predictor effects, the observed preference for concise, action-oriented messages underscores the potential importance of clarity and behavioral specificity when designing SMS reminders in Taiwan's universal health coverage context. Content validity assessment further confirmed that the 6 SMS prototypes possess adequate semantic clarity and cultural appropriateness for use in subsequent interventional research.

Although convenience sampling limits generalizability, the results offer valuable formative evidence to guide the development of SMS interventions aimed at improving appointment adherence. Future randomized controlled trials should evaluate the real-world effectiveness of these demographically tailored and behaviorally informed reminder formats in reducing outpatient no-show rates and supporting more efficient use of ambulatory care resources.

Acknowledgments

The authors would like to thank the Kaohsiung Medical University Hospital and Kaohsiung Medical University for their generous support.

Funding

This study was supported by a research grant from Kaohsiung Municipal Ta-Tung Hospital, Taiwan (Grant No. KMTTH-112-036).

Authors' Contributions

Conceptualization: P-SC, S-HH
Data curation: H-TL, S-HH, S-LM, Y-CC
Formal analysis: S-HH, Y-CC
Funding acquisition: P-SC
Investigation: H-TL, S-HH, S-LM, Y-CC
Methodology: P-SC, S-HH
Project administration: P-SC, S-HH
Resources: P-SC
Software: S-HH, Y-CC
Supervision: P-SC
Validation: H-TL, S-HH, S-LM, Y-CC
Visualization: P-SC, S-HH, Y-CC
Writing–original draft: S-HH, Y-CC
Writing–review and editing: I-CL, P-SC, S-HH

Conflicts of Interest

The authors declare no conflicts of interest.

Multimedia Appendix 1
The original 6 prototypes of SMS reminders (Chinese and English).
[DOCX File, 20 KB - jmir_v28i1e81324_app1.docx]

References

XSL•FO
RenderX

1. Li DG, Pournamdari AB, Liu KJ, Laskowski K, Joyce C, Mostaghimi A. Evaluation of point-of-care decision support for adult acne treatment by primary care clinicians. JAMA Dermatol 2020 May 1;156(5):538-544. [doi: 10.1001/jamadermatol.2020.0135] [Medline: 32129799]
2. Creadore A, Desai S, Li SJ, et al. Insurance acceptance, appointment wait time, and dermatologist access across practice types in the US. JAMA Dermatol 2021 Feb 1;157(2):181-188. [doi: 10.1001/jamadermatol.2020.5173] [Medline: 33439219]
3. Ayanian JZ, Sequist TD, Zaslavsky AM, Johannes RS. Physician reminders to promote surveillance colonoscopy for colorectal adenomas: a randomized controlled trial. J Gen Intern Med 2008 Jun;23(6):762-767. [doi: 10.1007/s11606-008-0576-2] [Medline: 18386103]
4. Lam TYT, Hui AJ, Sia F, et al. Short Message Service reminders reduce outpatient colonoscopy nonattendance rate: a randomized controlled study. J Gastroenterol Hepatol 2021 Apr;36(4):1044-1050. [doi: 10.1111/jgh.15218] [Medline: 32803820]
5. Deng X, Wang Y, Zhu T, Zhang W, Yin Y, Ye L. Short message service (SMS) can enhance compliance and reduce cancellations in a sedation gastrointestinal endoscopy center: a prospective randomized controlled trial. J Med Syst 2015 Jan;39(1):169. [doi: 10.1007/s10916-014-0169-z] [Medline: 25476268]
6. Kooij FO, Vos N, Siebenga P, Klok T, Hollmann MW, Kal JE. Automated reminders decrease postoperative nausea and vomiting incidence in a general surgical population. Br J Anaesth 2012 Jun;108(6):961-965. [doi: 10.1093/bja/aes024] [Medline: 22382396]
7. Chen LC, Schafheutle EI, Noyce PR. The impact of nonreferral outpatient co-payment on medical care utilization and expenditures in Taiwan. Res Social Adm Pharm 2009 Sep;5(3):211-224. [doi: 10.1016/j.sapharm.2008.08.003] [Medline: 19733822]
8. Sallis A, Sherlock J, Bonus A, et al. Pre-notification and reminder SMS text messages with behaviourally informed invitation letters to improve uptake of NHS Health Checks: a factorial randomised controlled trial. BMC Public Health 2019 Aug 22;19(1):1162. [doi: 10.1186/s12889-019-7476-8] [Medline: 31438908]
9. Lacy NL, Paulman A, Reuter MD, Lovejoy B. Why we don't come: patient perceptions on no-shows. Ann Fam Med 2004;2(6):541-545. [doi: 10.1370/afm.123] [Medline: 15576538]
10. Hwang AS, Atlas SJ, Cronin P, et al. Appointment "no-shows" are an independent predictor of subsequent quality of care and resource utilization outcomes. J Gen Intern Med 2015 Oct;30(10):1426-1433. [doi: 10.1007/s11606-015-3252-3] [Medline: 25776581]
11. Husain-Gambles M, Neal RD, Dempsey O, Lawlor DA, Hodgson J. Missed appointments in primary care: questionnaire and focus group study of health professionals. Br J Gen Pract 2004 Feb;54(499):108-113. [Medline: 14965389]
12. Smith A. In: Haakonssen K, editor. Adam Smith: The Theory of Moral Sentiments: Boston University; 2012. [doi: 10.1017/CBO9780511800153]
13. Berliner Senderey A, Kornitzer T, Lawrence G, et al. It's how you say it: systematic A/B testing of digital messaging cut hospital no-show rates. PLoS One 2020;15(6):e0234817. [doi: 10.1371/journal.pone.0234817] [Medline: 32574181]
14. Hu M, Xu X, Li X, Che T. Managing patients' no-show behaviour to improve the sustainability of hospital appointment systems: exploring the conscious and unconscious determinants of no-show behaviour. J Clean Prod 2020 Oct;269:122318. [doi: 10.1016/j.jclepro.2020.122318]
15. Smith A. The Theory of Moral Sentiments: Penguin; 2010.
16. Hallsworth M, Berry D, Sanders M, et al. Stating appointment costs in SMS reminders reduces missed hospital appointments: findings from two randomised controlled trials. PLoS One 2015;10(9):e0137306. [doi: 10.1371/journal.pone.0137306] [Medline: 26366885]
17. Haynes SN, Richard DCS, Kubany ES. Content validity in psychological assessment: a functional approach to concepts and methods. Psychol Assess 1995;7(3):238-247. [doi: 10.1037/1040-3590.7.3.238]
18. Tsai WC, Lee WC, Chiang SC, Chen YC, Chen TJ. Factors of missed appointments at an academic medical center in Taiwan. J Chin Med Assoc 2019 May;82(5):436-442. [doi: 10.1097/JCMA.0000000000000068] [Medline: 30907780]
19. Acharya A, Judah G, Ashrafian H, et al. Investigating the national implementation of SMS and mobile messaging in population screening (the SIPS study). EBioMedicine 2023 Jul;93:104685. [doi: 10.1016/j.ebiom.2023.104685] [Medline: 37384997]
20. Neal RD, Hussain-Gambles M, Allgar VL, Lawlor DA, Dempsey O. Reasons for and consequences of missed appointments in general practice in the UK: questionnaire survey and prospective review of medical records. BMC Fam Pract 2005 Nov 7;6:47. [doi: 10.1186/1471-2296-6-47] [Medline: 16274481]
21. Barrera Ferro D, Brailsford S, Bravo C, Smith H. Improving healthcare access management by predicting patient no-show behaviour. Decis Support Syst 2020 Nov;138:113398. [doi: 10.1016/j.dss.2020.113398]
22. Watson J, Woodward T. Jean Watson's theory of human caring. In: Nursing Theories and Nursing Practice, 3rd edition: Davis Company; 2010:351-369 URL: https://www.fadavis.com/product/nursing-research-theory-practice-parker-smith-3 [accessed 2026-02-03]
23. Lynn MR. Determination and quantification of content validity. Nurs Res 1986;35(6):382-385. [Medline: 3640358]
24. Sarkar SK, Tang CY. Adjusting the Benjamini–Hochberg method for controlling the false discovery rate in knockoff-assisted variable selection. Biometrika 2022 Nov 29;109(4):1149-1155. [doi: 10.1093/biomet/asab066]

XSL•FO
RenderX

25. DiMatteo MR. Social support and patient adherence to medical treatment: a meta-analysis. Health Psychol 2004 Mar;23(2):207-218. [doi: 10.1037/0278-6133.23.2.207] [Medline: 15008666]
26. Dobransky K, Hargittai E. The disability divide in internet access and use. Inf Commun Soc 2006 Jun;9(3):313-334. [doi: 10.1080/13691180600751298]
27. Xie B. Effects of an eHealth literacy intervention for older adults. J Med Internet Res 2011 Nov 3;13(4):e90. [doi: 10.2196/jmir.1880] [Medline: 22052161]
28. Bidmon S, Terlutter R. Gender differences in searching for health information on the internet and the virtual patient-physician relationship in Germany: exploratory results on how men and women differ and why. J Med Internet Res 2015 Jun 22;17(6):e156. [doi: 10.2196/jmir.4127] [Medline: 26099325]
29. Terrero Ledesma NE, Nájera A, Reolid-Martínez RE, Escobar-Rabadán F. Internet health information seeking by primary care patients. Rural Remote Health 2022 Oct;22(4):6585. [doi: 10.22605/RRH6585] [Medline: 36192837]
30. de Reuver M, Bouwman H. Dealing with self-report bias in mobile Internet acceptance and usage studies. Inf Manage 2015 Apr;52(3):287-294. [doi: 10.1016/j.im.2014.12.002]
31. Crutchfield TM, Kistler CE. Getting patients in the door: medical appointment reminder preferences. Patient Prefer Adherence 2017;11:141-150. [doi: 10.2147/PPA.S117396] [Medline: 28182131]
32. Mohammed Selim S, Senanayake S, McPhail SM, Carter HE, Naicker S, Kularatna S. Consumer preferences for a healthcare appointment reminder in Australia: a discrete choice experiment. Patient 2024 Sep;17(5):537-550. [doi: 10.1007/s40271-024-00692-9] [Medline: 38605246]
33. Teo AR, Niederhausen M, Handley R, et al. Using nudges to reduce missed appointments in primary care and mental health: a pragmatic trial. J Gen Intern Med 2023 Jul;38(Suppl 3):894-904. [doi: 10.1007/s11606-023-08131-5] [Medline: 37340264]
34. Junod Perron N, Dao MD, Righini NC, et al. Text-messaging versus telephone reminders to reduce missed appointments in an academic primary care clinic: a randomized controlled trial. BMC Health Serv Res 2013 Apr 4;13(1):125. [doi: 10.1186/1472-6963-13-125] [Medline: 23557331]
35. Do EK, Cohen SA, Brown MJ. Socioeconomic and demographic factors modify the association between informal caregiving and health in the Sandwich Generation. BMC Public Health 2014 Apr 15;14(1):362. [doi: 10.1186/1471-2458-14-362] [Medline: 24735495]
36. Becker R. Gender and survey participation: an event history analysis of the gender effects of survey participation in a probability-based multi-wave panel study with a sequential mixed-mode design. Methods Data Anal 2022;16(1):3-32. [doi: 10.12758/mda.2021.08]
37. Weber AM, Gupta R, Abdalla S, Cislaghi B, Meausoone V, Darmstadt GL. Gender-related data missingness, imbalance and bias in global health surveys. BMJ Glob Health 2021 Nov;6(11):e007405. [doi: 10.1136/bmjgh-2021-007405] [Medline: 34836912]
38. Albaugh QM, Harell A, Loewen PJ, Rubenson D, Stephenson LB. Measuring transgender and nonbinary identities in online surveys: evidence from two national election studies. PS: Polit Sci Polit 2025 Jul;58(3):406-412. [doi: 10.1017/S1049096525000137]

Abbreviations

aOR: adjusted odds ratio
FDR: false discovery rate
MA: missed appointment

Edited by A Schwartz, M Balcarras; submitted 26.Jul.2025; peer-reviewed by K Le, YW Chiu; accepted 18.Dec.2025; published 07.Apr.2026.

Please cite as:
Hsiao SH, Chuang YC, Ma SL, Lin HT, Lee IC, Chen PS
Preferences for Personalized Text Message Appointment Reminders Among Outpatients in a Universal Health System: Cross-Sectional Study
J Med Internet Res 2026;28:e81324
URL: https://www.jmir.org/2026/1/e81324
doi:10.2196/81324

© Shih-Huai Hsiao, Yu-Chen Chuang, Shu-Ling Ma, Hui-Tzu Lin, I-Chen Lee, Pei-Shih Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet

XSL•FO
RenderX

Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Personalized Smartphone Messaging for Secondary Prevention After Percutaneous Coronary Intervention: Randomized Controlled Trial

Sunki Lee[1*], MD; Ji Bak Kim[2*], MD, PhD; Seung Eon Shin[1], RN; Jeonghoon Ahn[3], PhD; Hyun-Ghang Jeong[4], MD, PhD; Kyung Ho Jung[5], BSc; Younghoon Song[5], MBA; Eung Ju Kim[1], MD, PhD

[1]Department of Cardiology, Korea University Guro Hospital, 148, Gurodong-Ro, Guro-gu, Seoul, Republic of Korea

[2]Division of Cardiology, Department of Internal Medicine, Hanil General Hospital, Seoul, Republic of Korea

[3]Department of Health Convergence, Ewha Womans University, Seoul, Republic of Korea

[4]Department of Psychiatry, Korea University Guro Hospital, Korea University College of Medicine, Seoul, Republic of Korea

[5]Department of IT Business, Hanmi Pharmaceutical (South Korea), Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Eung Ju Kim, MD, PhD
Department of Cardiology, Korea University Guro Hospital, 148, Gurodong-Ro, Guro-gu, Seoul, Republic of Korea

Abstract

Background: Cardiac rehabilitation improves outcomes after percutaneous coronary intervention (PCI), but participation remains suboptimal due to barriers such as distance, cost, and lack of referral. Mobile health interventions may enhance accessibility and engagement.

Objective: This study evaluated the effectiveness and user acceptability of a smartphone messaging app (AnSim) in improving cardiovascular risk factors among patients who recently underwent PCI.

Methods: A 2-arm randomized controlled trial was conducted at 2 Korean hospitals; 120 patients who had undergone PCI within the prior month were randomized (1:1) to receive AnSim plus usual care (intervention) or usual care alone (control). The intervention comprised personalized messages tailored by behavioral stage and risk profile, delivered 6 times/week for 6 months. Outcomes were assessed at baseline, 6 months, and 9 months. The primary outcome was change in blood pressure (BP) at 6 and 9 months. Secondary outcomes were lipid profiles, hemoglobin A_{1c}, BMI, smoking status, physical activity, and achievement of guideline-recommended risk-factor targets. User acceptability was assessed by questionnaire.

Results: Overall, 115 (95.8%; intervention: n=58, control: n=57) participants completed follow-up at 6 months and 110 (91.7%; intervention: n=54, control: n=56) at 9 months. At 6 months, systolic BP was 124.9 (SD 16.7) mmHg in the intervention group and 124.8 (SD 13.1) mmHg in the control group (between-group difference 0.1 mmHg, 95% CI −5.4 to 5.6), and diastolic BP was 80.2 (SD 11.4) mmHg vs 78.4 (SD 10.6) mmHg (difference 1.8 mmHg, 95% CI −2.3 to 5.9). At 9 months, systolic BP was 126 (SD 14.2) mmHg vs 128 (SD 15.4) mmHg (difference −2 mmHg, 95% CI −7.6 to 3.6), and diastolic BP was 80.8 (SD 13.2) mmHg vs 81.3 (SD 12.0) mmHg (difference −0.5 mmHg, 95% CI −5.3 to 4.3). No significant between-group differences were observed in secondary outcomes. In post hoc analyses, participants with above-median message reading were more likely to achieve ≥4 of 5 guideline-recommended risk-factor targets at 9 months (69% vs 19.2%; relative risk 3.59, 95% CI 1.57 - 8.18; $P<.001$) and logged more diary entries (mean difference 203.4 entries, 95% CI 79.3 - 327.5; $P=.001$). BP responders at 9 months read messages on more days than nonresponders (110.3 vs 83.3 d; mean difference 27 d, 95% CI 4.6 - 49.3; $P=.02$). Among survey respondents (54/60, 90%), 87% (47/54) found the messages helpful and 81.4% (44/54) wished to continue. Two events occurred in the control group ($P=.50$).

Conclusions: This trial found no between-group improvement in BP with low-intensity personalized messaging compared to an attention-matched diary app, but exploratory analyses suggest engagement may be a key determinant of potential benefit. Unlike comprehensive app-based cardiac rehabilitation programs, AnSim provides scalable behavioral reinforcement that can be integrated into post-PCI follow-up with minimal burden. These findings support future pragmatic trials optimizing engagement strategies and evaluating longer-term clinical and implementation outcomes.

Trial Registration: Clinical Research Information Service (CRIS) KCT0002361; https://cris.nih.go.kr/cris/search/detailSearch.do?seq=7230

(J Med Internet Res 2026;28:e81524) doi:10.2196/81524

XSL•FO
RenderX

KEYWORDS

cardiac rehabilitation; coronary artery disease; mobile health; mHealth; smartphone app; text messaging; risk factor modification; secondary prevention

Introduction

Cardiovascular (CV) disease, including coronary artery disease (CAD) and stroke, remains a leading cause of morbidity and mortality worldwide [1,2]. Even after successful percutaneous coronary intervention (PCI), patients with CAD remain at substantial risk of recurrent myocardial infarction, rehospitalization, and premature death, contributing to ongoing health care use and costs [3,4]. Primary prevention aims to prevent the first occurrence of CV events, whereas secondary prevention focuses on reducing recurrence and improving outcomes in patients with established CV disease. In the post-PCI setting, secondary prevention strategies, such as cardiac rehabilitation (CR), are important for reducing recurrence and improving patient outcomes.

CR is a comprehensive, guideline-based intervention designed to promote healthy lifestyles and address modifiable CV risk factors through education, exercise training, dietary counseling, and stress management [5]. Participation in CR has been associated with significantly reduced CV morbidity and mortality and improved quality of life, with evidence supporting its cost-effectiveness [6-8]. However, despite its clear benefits, global CR participation remains low, reported in 20% to 50% of patients [9]. This reflects both limited availability and access in some settings and low uptake or sustained adherence even when CR is offered, and these factors may vary by health system and geography [10,11]. Patients with multiple CV risk factors or low health literacy, a particularly at-risk population, are among the least likely to attend [12] and adhere to CR [13].

Barriers to CR participation and sustained secondary prevention are multifactorial. System-level barriers include limited referral and program capacity, geographic inaccessibility, costs, and insurance constraints [14,15], whereas patient-level barriers include limited knowledge of CR, low perceived benefit or motivation, competing demands, physical limitations, and challenges maintaining health behaviors over time [16,17]. Accordingly, scalable interventions that reinforce education, motivation, and self-management outside clinic visits may be valuable, particularly when structured CR participation is limited. Mobile communication technologies offer opportunities to deliver low-cost, individualized support at scale. SMS text messaging has been used for health behavior interventions and, in patients with CAD, has been associated with improvements in medication adherence and risk-factor management in several studies [18,19]. As smartphone capabilities have expanded, mobile apps can deliver tailored content and support self-monitoring, though evidence remains mixed regarding their clinical impact on sustained risk-factor control beyond usual care [20].

Although prior digital secondary prevention and CR interventions have shown promise, many rely on more resource-intensive components such as structured exercise modules, real-time monitoring, or intensive coaching, which may limit scalability in routine post-PCI care [21,22]. Evidence remains limited on whether a lightweight, theory-informed, personalized messaging intervention can improve risk-factor control while remaining feasible for integration into usual follow-up workflows.

In this context, we developed AnSim, a smartphone-based, patient-specific messaging program designed to support post-PCI secondary prevention through personalized educational and motivational messages, with optional self-monitoring features. AnSim was not intended to deliver comprehensive CR but rather to provide scalable reinforcement of guideline-recommended behaviors and risk-factor management following PCI. Our primary objective was to determine whether this personalized messaging intervention improves blood pressure control compared to usual care. We also explored whether higher levels of patient engagement with the app would be positively associated with the attainment of guideline-recommended CV risk-factor targets.

Methods

Trial Design

The clinical trial was designed as a single-blinded, 2-arm randomized controlled trial with a total follow-up period of 9 months, comprising 6 months of intervention and 3 months of additional observation. The trial was conducted at 2 sites in Korea: a secondary general hospital (Sejong General Hospital) and a tertiary academic hospital (Korea University Guro Hospital). The trial was registered at Clinical Research Information Service (CRIS) (KCT0002361) [23]. Registration occurred prospectively on August 10, 2016, and the first participant was enrolled on December 17, 2016. This study is reported in accordance with the CONSORT (Consolidated Standards of Reporting Trials) 2025 statement and the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist (Checklist 1) for electronic health and mobile health randomized trials [24,25], and the completed CONSORT 2025 checklist is provided in Checklist 2.

Participants and Setting

Eligible participants were adults (≥18 y) who had undergone PCI within the previous 1 month and were able to use a smartphone. Patients with cognitive dysfunction (eg, stroke or dementia) or significant psychiatric disorders (eg, anxiety, depression) were excluded. PCI was performed in patients with significant anatomic stenosis (>50% in the left main or >70% in non–left main coronary arteries) or physiologic ischemia (fractional flow reserve <0.80) and included cases of ST-segment elevation myocardial infarction, non–ST elevation acute coronary syndrome, unstable angina, and stable angina. All patients received guideline-directed medical therapy. Exclusion criteria included inability to use a smartphone, nonfluency in Korean, prior coronary artery bypass graft

XSL•FO
RenderX

surgery, uncontrolled arrhythmia or heart failure, severe chronic obstructive pulmonary disease or asthma, end-stage renal disease requiring dialysis, or terminal malignancy.

Randomization and Blinding

Participants were randomly assigned to either the intervention or control group using a centralized, computer-generated randomization table with a 1:1 allocation ratio and a block size of 4. To minimize performance and measurement bias, all participants were provided with a general health monitoring app ("Heart Keeper") developed by the Korean Society of Cardiology. To support participant blinding, the Heart Keeper app was installed for both groups; Heart Keeper functioned as a simple diary-style tool for recording clinical information and did not deliver structured behavior-change messaging or coaching. The intervention group additionally received AnSim. A comparison between the 2 apps is listed in Table S1 in Multimedia Appendix 1. Due to the nature of the intervention (personalized smartphone messages), participants could not be fully blinded after group allocation. Blinding was maintained by ensuring that outcome assessors and investigators were unaware of group allocation. Research nurses trained in allocation concealment and participant instruction managed all app installations. Participants were instructed not to disclose their app usage to study staff during follow-up visits.

Ethical Considerations

The study protocol was approved by the Institutional Review Board of Korea University Guro Hospital (approval number MD16037-002). All participants provided written informed consent prior to enrollment; the consent process included permission for use of collected data for the prespecified analyses reported in this manuscript. Study data were deidentified and stored on encrypted, access-restricted institutional servers, and only authorized study personnel had access. Participants were provided with contact details to report any privacy-related concerns. Participants did not receive financial compensation for their participation in this study. No identifiable images of individual participants or users are included in the manuscript or supplementary materials.

Interventions

A detailed description of the AnSim app development and usability testing has been published previously [26]. Briefly, AnSim was developed through multidisciplinary collaboration (cardiology, behavioral science, nursing, pharmacy, nutrition, rehabilitation, and app development), informed by focus group interviews, and implemented as a structured message bank spanning 5 domains (CV health/medications, nutrition, physical activity, stress management, and smoking cessation). Messages were tailored according to simplified transtheoretical model stages and mapped to 26 behavior change techniques [27]; iterative refinement incorporated expert review and feedback from potential users, followed by pilot testing of the delivery system.

Intervention Development: Focus Group Interviews

To capture patient needs and inform message development, a focus group interview was conducted with 8 patients who had previously undergone PCI and used a smartphone. Participants were purposefully sampled to reflect a diversity of age, sex, and education level. The subject of the interview consisted of five categories: (1) degree of use of the smartphone app, (2) exercise, (3) nutrition, (4) stress management, and (5) knowledge about CAD and prevention. Interviews were recorded with patient consent and textualized in verbatim form, and qualitative analysis was performed using NVivo (QSR International).

Intervention Development: Message Drafting and Refinement

Based on insights from the focus group interview, health experts drafted messages in five thematic categories: (1) general CV health and medications, (2) nutrition, (3) physical activity, (4) stress management, and (5) smoking cessation. Each message contained 40 to 140 Korean characters and was adapted from international guidelines and official educational resources from CV health–related academic societies.

A total of 450 messages (90 per category) was initially developed and reviewed for clinical accuracy, literacy level, and relevance. Subsequently, 200 potential users evaluated each message using a 5-point Likert scale for readability and usefulness. Most messages received favorable scores (mean 3.95, SD 0.39 for readability and mean 3.91, SD 0.39 for usefulness), and 98% (n=441) scored ≥3 on both metrics. Messages were refined accordingly, and examples are provided in Table S2 in Multimedia Appendix 1.

Intervention Development: Theoretical Framework

The messaging content was informed by 26 established behavior change techniques derived from models such as the information-motivation-behavioral skills model, theory of planned behavior, social cognitive theory, and operant conditioning [27]. Multimedia elements (eg, video clips, illustrations) were integrated when possible to support learning and reinforce motivation.

To match message content to individual readiness for change, we applied a simplified transtheoretical model with three behavioral stages: (1) precontemplation, (2) contemplation and preparation, and (3) action and maintenance [26]. According to these simplified steps, behavior change techniques were categorized, and 30 messages were developed for each. Each patient's behavioral stage for each message category was reassessed monthly through the app, and messaging was adapted accordingly.

Intervention Delivery

Participants in the intervention group received 6 messages per week over a 24-week period. Messages were sent randomly at chosen times (9:00 AM, 12:00 PM, or 3:00 PM) from Monday through Saturday. The content was personalized based on baseline characteristics (eg, presence of diabetes, smoking status) and behavioral stage. Messages were drawn from the 5 categories mentioned above, with 5 standard messages per week plus 1 additional message targeting an individual's weakest behavior domain (eg, nutrition or smoking cessation). Messages were never repeated for the same individual. Both the AnSim and Heart Keeper apps allowed optional tracking of health metrics (eg, blood pressure, glucose, exercise, diet). To support

XSL•FO
RenderX

participant blinding, Heart Keeper was installed in both groups and functioned as a diary-style app for recording clinical information; it did not deliver structured behavior-change messaging, coaching, or interactive rehabilitation modules. All participants received brief onboarding at enrollment. To enhance engagement, participants in the intervention group also received weekly feedback messages based on their logged health data, sent by a designated health care provider. No changes in app functionality, message content, or technical downtimes occurred during the intervention period.

Outcomes

Primary and Secondary Outcomes

Participants were assessed at baseline, 6 months, and 9 months. The primary outcome was the change in blood pressure, measured after 10 minutes of seated rest using a standardized automatic device (HEM-7080IT, OMRON Healthcare). Secondary outcomes were lipid profiles (total cholesterol, low-density lipoprotein cholesterol [LDL-C], high-density lipoprotein cholesterol, triglycerides), hemoglobin A_{1c} (HbA_{1c}), 6-minute walk distance, self-reported smoking status, and venous carboxyhemoglobin levels as a biochemical indicator of smoking [28]. Exploratory (post hoc) analyses were performed within the intervention group to examine associations between engagement metrics and risk-factor target achievement. Participants were classified as blood pressure responders if their blood pressure was less than or equal to 130/80 mmHg at 9 months. These exploratory analyses were considered hypothesis-generating.

Process Outcomes

The process evaluation comprised of (1) objective app usage metrics (eg, message delivery success, message reading frequency, and health data entry frequency) and (2) app acceptability and usability, assessed at 6 months using a postintervention self-report questionnaire evaluating satisfaction with and perceived utility of the messages on a 5-point Likert scale.

Sample Size

The sample size was calculated based on the TEXT-ME (Tobacco, Exercise and Diet Messages) study [18], which reported a 6% reduction in systolic blood pressure (SBP; approximately 8 mmHg difference) in the intervention group compared to the control group. Based on these parameters (α=.05, β=.20, power=80%), the minimum required sample size was calculated as 121 using MedCalc (version 16.2.1; MedCalc Software). Although the calculated target was 121, this study enrolled a total of 120 patients (60 per group), which approximates the required power. To evaluate the missingness mechanism across prespecified continuous outcomes (SBP/diastolic blood pressure [DBP], BMI, lipid profile, and HbA_{1c} measured at baseline, 6 mo, and 9 mo), we performed the Little test for missing completely at random (MCAR).

Statistical Analysis

All analyses were performed on an intention-to-treat basis using SPSS version 20.0 (IBM Corp). Continuous variables are presented as mean (SD); categorical variables are reported as n (%). Normality was assessed using the Shapiro-Wilk test and variance homogeneity by *F* test. Between-group comparisons for continuous variables were performed using independent-samples *t* tests (2-tailed). Categorical variables were compared using Pearson chi-square or Fisher exact tests, as appropriate. To examine group-by-time interactions for repeated measures, 2-way mixed ANOVA was used. Relative risks and 95% CIs for outcomes were estimated using Cox proportional hazards models. All statistical tests were 2-sided, and $P<.05$ was considered statistically significant.

Results

Participant Flow and Recruitment

A total of 120 patients were enrolled between December 2016 and April 2018 and randomized into the intervention group (n=60) and the control group (n=60). Participants were recruited consecutively from the participating PCI clinics during the enrollment period (ie, a convenience sample of eligible patients presenting to the sites), and randomization occurred after written informed consent and baseline assessment. Follow-up completion rates were high: 96.7% (58/60) in the intervention group and 95% (57/60) in the control group at 6 months, and 90% (54/60) and 93.3% (56/60), respectively, at 9 months. By 6 months, 2 participants in the intervention group and 3 in the control group did not complete follow-up. Between 6 and 9 months, additional attrition occurred in 4 and 1 participants, respectively. Overall, reasons for attrition included consent withdrawal (2 in each group) and loss to follow-up (4 in the intervention group, 2 in the control group) (Figure 1).

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram of participant enrollment, randomization, follow-up, and analysis. A total of 120 patients were randomized to the intervention group (n=60) or the control group (n=60), and all received the allocated condition. Follow-up completion was 96.7% (58/60) in the intervention group and 95% (57/60) in the control group at 6 months, and 90% (54/60) and 93.3% (56/60), respectively, at 9 months. Reasons for noncompletion at 6 months and additional attrition between 6 and 9 months are shown. The number of patients assessed for eligibility before consent was not prospectively recorded in the original trial dataset and, therefore, could not be reported.

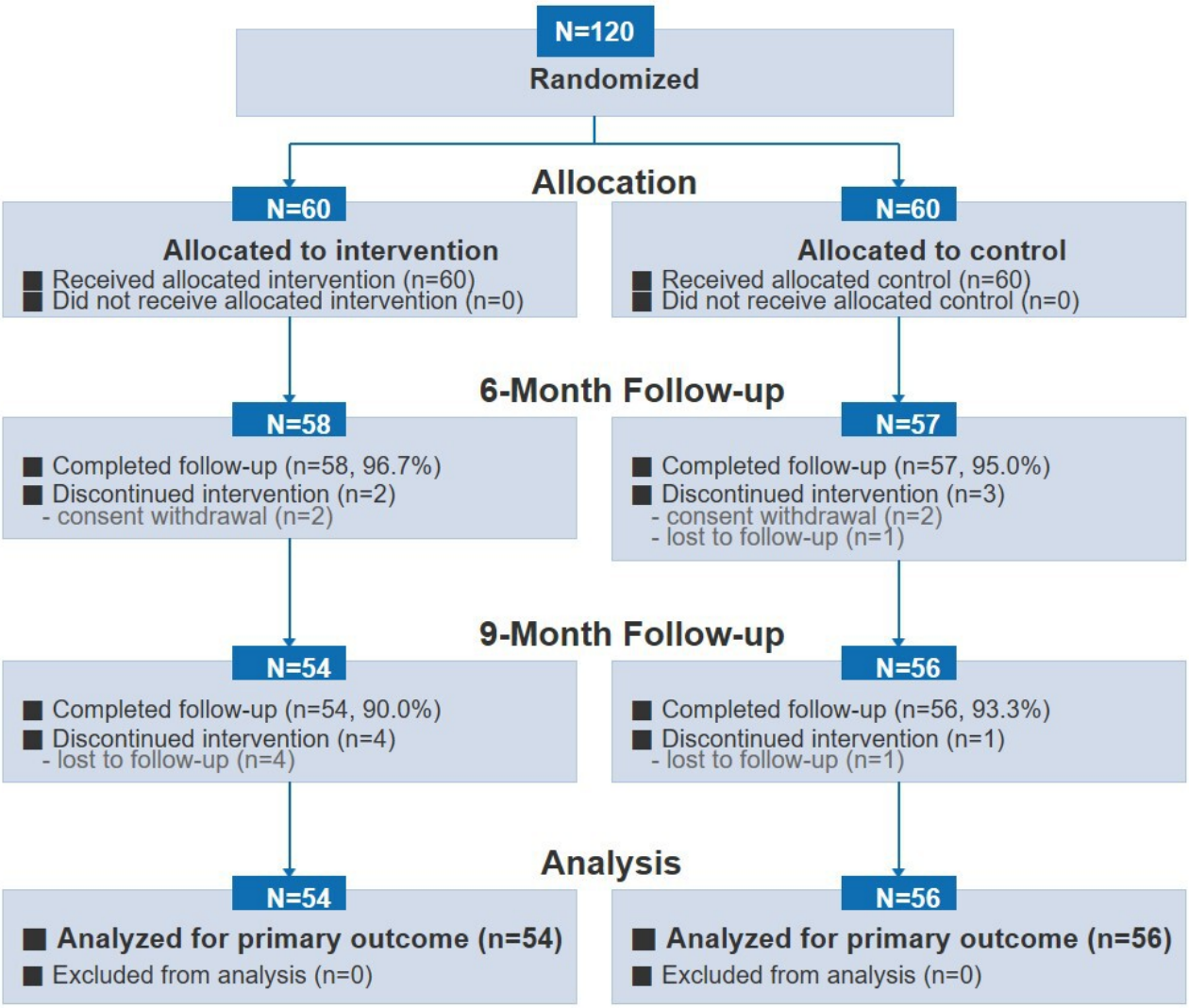

Baseline Data

Baseline characteristics are shown in Table 1. The mean age of participants was 58.5 (8.8) years, and 84.2% (101/120) were male. However, the prevalence of diabetes mellitus was significantly lower in the intervention group (21.7% vs 41.7%, *P*=.02), resulting in a lower baseline HbA_{1c} (6.13% vs 6.60%, *P*=.048). No significant differences were observed in baseline blood pressure, lipid profiles, BMI, medication usage, or proportion achieving guideline-recommended goals. The medications prescribed in each group at the time of discharge after index coronary intervention were similar. Baseline clinical variables were nearly complete (mean missingness 0.83%). Missingness increased at follow-up (mean missingness 12.40% at 6 mo and 23.14% at 9 mo across the prespecified continuous outcomes) and was broadly similar between groups (eg, SBP at 9 mo: 16.7% vs 20%; HbA_{1c} at 9 mo: 23.3% vs 30%). While the Little MCAR test across 24 continuous clinical variables did not reject the MCAR assumption (χ^2_{137}=152.83, *P*=.17), indicating insufficient evidence of a systematic missingness pattern, this result suggests rather than definitively proves that the data are missing completely at random.

XSL•FO
RenderX

Table . Baseline characteristics.

Characteristics	Intervention (n=60)	Control (n=60)	Total (N=120)	*P* value[a]
Demographics				
Age (y), mean (SD)	59.1 (8.2)	57.2 (7.9)	58.5 (8.8)	.22
Male sex, n (%)	51 (85)	50 (83.3)	101 (84.2)	.80
Clinical data, mean (SD)				
Height (cm)	166.4 (7.6)	166.1 (7.8)	166.3 (7.7)	.84
Weight (kg)	71.7 (11.4)	70.3 (10.8)	71 (11.1)	.50
BMI (kg/m²)	25.8 (3)	25.4 (3)	25.6 (3)	.49
Clinical risk factors				
Acute coronary syndrome, n (%)	35 (58.3)	40 (66.7)	75 (62.5)	.35
Hypertension, n (%)	31 (51.7)	36 (60)	67 (55.8)	.36
Diabetes mellitus, n (%)	13 (21.7)	25 (41.7)	38 (31.7)	.02
Dyslipidemia, n (%)	28 (46.7)	30 (50)	58 (48.3)	.72
Heart failure, n (%)	2 (3.3)	2 (3.3)	4 (3.3)	>.99
Stroke, n (%)	0 (0)	0 (0)	0 (0)	NA[b]
Cholesterol (mg/dL), mean (SD)				
Total cholesterol	170.4 (43.4)	175 (47.8)	172.7 (45.5)	.58
LDL-C[c]	97.7 (37)	102.6 (35.4)	100.1 (36.1)	.46
HDL-C[d]	44.3 (12.3)	45.1 (10.1)	44.7 (11.2)	.69
Triglycerides	160.73 (137.8)	162.7 (203.7)	161.7 (173.2)	.95
Systolic BP[e] (mmHg), mean (SD)	124.9 (16.1)	124.8 (15)	124.8 (15.5)	.97
Diastolic BP (mmHg), mean (SD)	73.2 (9.4)	73.1 (9.3)	73.1 (9.3)	.99
HbA_{1c}[f] (%), mean (SD)	6.13 (1)	6.6 (1.4)	6.4 (1.2)	.048
Carboxyhemoglobin (%), mean (SD)	0.6 (0.6)	0.9 (1.2)	0.7 (1)	.15
Ejection fraction (%), mean (SD)	59.1 (8.2)	57.2 (7.9)	58.1 (8.1)	.22
6-min walk distance (m), mean (SD)	421.7 (126.9)	404.3 (123)	413 (124.7)	.45
Current smoker, n (%)	17 (28.3)	17 (28.3)	34 (28.3)	>.99
Guideline, n (%)				
LDL-C <70 mg/dL	18 (30)	12 (20)	30 (25)	.21
BP <140/90 mmHg	51 (85)	47 (78.3)	98 (81.7)	.35
Regular exercise	26 (43.3)	23 (38.3)	49 (40.8)	.58
Nonsmoker	43 (71.7)	43 (71.7)	86 (71.7)	>.99
BMI <25 kg/m^2	22 (36.7)	26 (43.3)	48 (40)	.46
Achieving 5 of 5	0 (0)	1 (1.7)	1 (0.8)	.32
Achieving 4 of 5	17 (28.3)	10 (16.7)	27 (22.5)	.13
Medications, n (%)				
Dual antiplatelet therapy	59 (98.3)	60 (100)	119 (99.2)	>.99

XSL•FO
RenderX

Characteristics	Intervention (n=60)	Control (n=60)	Total (N=120)	*P* value[a]
Renin-angiotensin system blocker	34 (56.7)	39 (65)	73 (60.8)	.35
Beta-blocker	37 (61.7)	41 (68.3)	78 (65)	.44
Statin	56 (95.8)	59 (98.3)	115 (95.8)	.36
All 4 medication classes	21 (35)	32 (53.3)	53 (43.2)	.04

[a]Independent *t* test between groups.

[b]N/A: not available.

[c]LDL-C: low-density lipoprotein cholesterol.

[d]HDL-C: high-density lipoprotein cholesterol.

[e]BP: blood pressure.

[f]HbA_{1c}: hemoglobin A_{1c}.

Outcomes and Estimation

Changes in blood pressure over time were not significantly different between groups at either 6 or 9 months (Figure 2; Table 2). At 9 months, mean SBP was 126 (14.2) mmHg in the intervention group vs 128 (15.4) mmHg in the control group (*P*=.94); diastolic pressure was 80.8 (13.2) mmHg vs 81.3 (12) mmHg (*P*=.85). There was no significant interaction effect between time and group for systolic ($F_{2,95}$=0.029; *P*=.97) or diastolic ($F_{2,95}$=0.333; *P*=.72) blood pressure. Similarly, no significant differences were observed in secondary outcomes, including lipid levels, HbA_{1c}, BMI, walking distance, or smoking rates at either follow-up point. The proportion of patients achieving all 5 guideline-recommended targets (LDL-C <70 mg/dL, BP <140/90 mmHg, regular exercise, nonsmoking status, BMI <25 kg/m²) did not differ significantly between groups at 6 or 9 months (Table 3).

Figure 2. Changes in blood pressure over time between intervention and control groups. (A) Systolic blood pressure and (B) diastolic blood pressure were assessed at 6 and 9 months. Mean values and SDs are shown for both groups. Points denote group means, and error bars indicate SDs. BP: blood pressure.

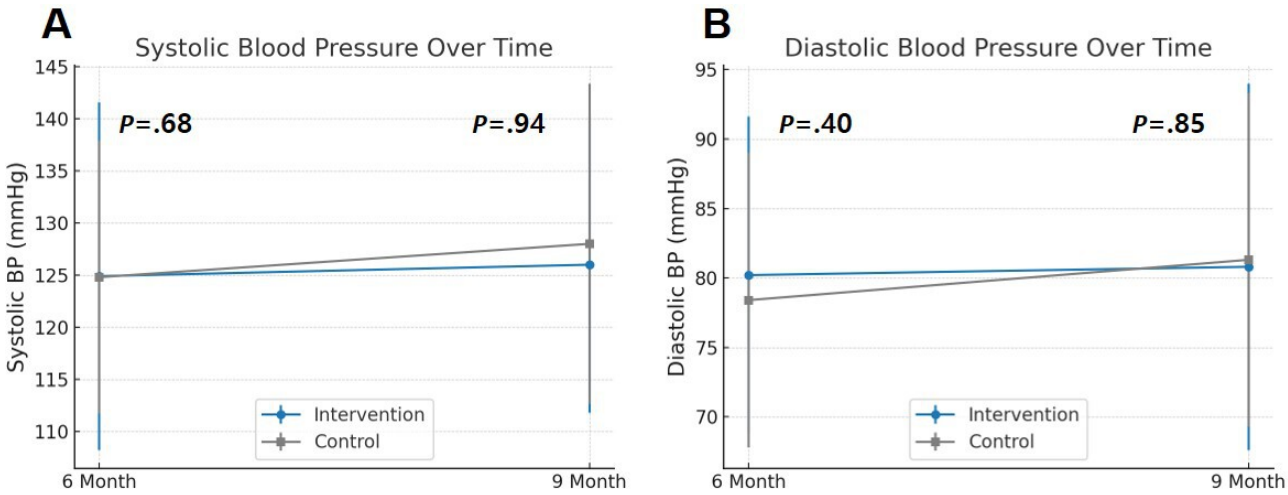

XSL•FO
RenderX

Table . Primary and secondary outcomes analyses at 6 and 9 months of follow-up.

Variable	Intervention (6 months: n=58; 9 months: n=54)	Control (6 months: n=57; 9 months: n=56)	*P* value[a]
Primary outcome			
Systolic BP[b] (mmHg), mean (SD)			
6 months	124.9 (16.7)	124.8 (13.1)	.68
9 months	126 (14.2)	128 (15.4)	.94
Diastolic BP (mmHg), mean (SD)			
6 months	80.2 (11.4)	78.4 (10.6)	.40
9 months	80.8 (13.2)	81.3 (12)	.85
Secondary outcome			
Total cholesterol (mg/dL), mean (SD)			
6 months	130.1 (26.2)	136.3 (27.6)	.24
9 months	134.4 (31.5)	132.5 (20)	.74
Triglycerides (mg/dL), mean (SD)			
6 months	146.1 (97.8)	144.3 (98.5)	.93
9 months	137.9 (71.3)	144.5 (85.7)	.69
HDL-C[c] (mg/dL), mean (SD)			
6 months	44.1 (9.7)	47.2 (11.8)	.14
9 months	46.8 (17.8)	45.3 (10.2)	.63
LDL-C[d] (mg/dL), mean (SD)			
6 months	67 (20.6)	70.1 (21)	.45
9 months	70.5 (29.4)	69.5 (17.1)	.86
HbA_{1c}[e] (%), mean (SD)			
6 months	6 (0.7)	6.3 (0.9)	.10
9 months	6.1 (0.7)	6.3 (0.9)	.18
Body weight (kg), mean (SD)			
6 months	71 (10.8)	68.7 (9.8)	.25
9 months	72.2 (11.4)	68.8 (9.9)	.19
BMI (kg/m^2), mean (SD)			
6 months	25.8 (2.9)	25 (3)	.29
9 months	25.8 (2.9)	25 (3)	.15
6-min walk distance (m), mean (SD)			
6 months	492.4 (87.8)	478.6 (83.9)	.43
9 months	509.6 (89.5)	498.6 (76.3)	.54
Current smoking, n (%)			
6 months	11 (20.4)	10 (19.2)	.88
9 months	9 (20.5)	10 (21.7)	.69
Carboxyhemoglobin (%), mean (SD)			
6 months	1.4 (1.6)	2.1 (2.7)	.10
9 months	1.4 (1.1)	1.4 (1.7)	.98

XSL•FO
RenderX

[a]Independent *t* test between groups.
[b]BP: blood pressure.
[c]HDL-C: high-density lipoprotein cholesterol.
[d]LDL-C: low-density lipoprotein cholesterol.
[e]HbA_{1c}: hemoglobin A_{1c}.

Table . Achievement of guideline-recommended goals at the 6 and 9 months of follow-up.

Variables	Intervention, n (%)	Control, n (%)	Relative risk (95% CI)	*P* value[a]
6 months (intervention: n=58; control: n=57)				
LDL-C[b] <70 mg/dL	31 (53.4)	32 (56.1)	0.95 (0.68 - 1.33)	.85
BP[c] <140/90 mmHg	39 (67.2)	38 (66.7)	1.01 (0.78 - 1.30)	>.99
Regular exercise	41 (70.7)	39 (68.4)	1.03 (0.81 - 1.32)	.84
Nonsmoker	43 (74.1)	42 (73.7)	1.01 (0.81 - 1.25)	>.99
BMI <25 kg/m^2	22 (37.9)	31 (54.4)	0.70 (0.46 - 1.05)	.093
Achieving 5 of 5	5 (8.6)	11 (19.3)	0.45 (0.17 - 1.20)	.11
Achieving 4 of 5	23 (39.7)	31 (54.4)	0.73 (0.49 - 1.08)	.14
9 months (intervention: n=54; control: n=56)				
LDL-C <70 mg/dL	24 (44.4)	28 (50)	0.89 (0.60 - 1.32)	.57
BP <140/90 mmHg	29 (53.7)	30 (53.6)	1 (0.71 - 1.42)	.99
Regular exercise	36 (66.7)	31 (55.4)	1.20 (0.89 - 1.63)	.25
Nonsmoker	42 (77.8)	38 (67.9)	1.15 (0.91 - 1.44)	.29
BMI <25 kg/m^2	16 (29.6)	24 (42.9)	0.69 (0.42 - 1.15)	.17
Achieving 5 of 5	6 (11.1)	9 (16.1)	0.69 (0.26 - 1.81)	.58
Achieving 4 of 5	19 (35.2)	22 (39.3)	0.90 (0.55 - 1.46)	.70

[a]Relative risk and 95% CIs were calculated with Cox proportional-hazard models.
[b]LDL-C: low-density lipoprotein cholesterol.
[c]BP: blood pressure.

Harms

During the 9-month follow-up, 2 clinical events occurred, both in the control group (1 coronary revascularization and 2 readmissions), but the difference between groups was not significant (*P*=.50; Table S3 in Multimedia Appendix 1). No deaths, myocardial infarctions, or heart failure hospitalizations were reported in either group.

Ancillary Analysis: Message Engagement

Since the intervention using the AnSim app was message-mediated, the efficacy of the intervention could be dependent on participant message use. Among intervention group participants, the frequency of message reading was analyzed as a marker of engagement. Patients were stratified into high (upper 50%) and low (lower 50%) message readers (Table 4). High readers accessed more than twice as many messages on average (130.5 vs 50.6 messages; *P*<.001) and had significantly greater engagement in health diary input (250.2 vs 46.8 entries; *P*=.001). Although no significant differences were observed in blood pressure or lipid levels between subgroups, a significantly larger proportion of high readers achieved at least 4 of 5 guideline-recommended goals at 9 months (69% vs 19.2%; *P*<.001).

XSL•FO
RenderX

Table . Comparison according to frequency of messages read.

Parameters	Upper 50% (6 months: n=31; 9 months: n=29)	Lower 50% (6 months: n=27; 9 months: n=26)	*P* value[a]
Baseline characteristics			
Age (y), mean (SD)	57.2 (7.6)	59.9 (10.7)	.25
Male sex, n (%)	25 (80.6)	26 (96.3)	.11
Acute coronary syndrome, n (%)	20 (64.5)	15 (55.6)	.59
Achieving ≥4 goals, n (%)	23 (74.2)	20 (74.1)	>.99
Days messages were read (d), mean (SD)	130.5 (11.4)	50.6 (35)	<.001
Health diary entries (n), mean (SD)	250.2 (325.6)	46.8 (93.5)	.001
Satisfaction with AnSim app, mean (SD)	3.7 (0.7)	3.6 (0.5)	.62
Primary outcome			
Systolic BP[b] (mmHg), mean (SD)			
6th month	125.5 (13)	122.5 (20)	.65
9th month	123.9 (13.1)	130 (15.4)	.18
Diastolic BP (mmHg), mean (SD)			
6th month	80.3 (10.2)	80.1 (12.8)	.97
9th month	80 (11.7)	81.6 (14.9)	.22
Secondary outcome			
Total cholesterol (mg/dL), mean (SD)			
6th month	132.3 (23.7)	127.8 (29)	.27
9th month	133.7 (32.4)	135.2 (31.3)	.86
Triglycerides (mg/dL), mean (SD)			
6th month	135.6 (60.1)	157.4 (126.9)	.74
9th month	141.2 (60.2)	134.1 (83.7)	.32
HDL-C[c] (mg/dL), mean (SD)			
6th month	45.3 (10.7)	42.8 (8.7)	.62
9th month	48.9 (22.7)	44.3 (8.7)	.85
LDL-C[d] (mg/dL), mean (SD)			
6th month	68.7 (21.9)	65.1 (19.3)	.36
9th month	67.8 (30.9)	73.4 (28.1)	.35
HbA_{1c}[e] (%), mean (SD)			
6th month	5.9 (0.8)	6.1 (0.7)	.10
9th month	6.1 (8.3)	6.1 (0.6)	.53
Body weight (kg), mean (SD)			
6th month	71 (11.3)	70.9 (10.6)	.83
9th month	72.6 (12.1)	71.7 (10.8)	.89
BMI (kg/m^2), mean (SD)			
6th month	25.6 (2.9)	25.4 (2.5)	.99
9th month	26 (3)	25.6 (2.8)	.70

XSL•FO
RenderX

Parameters	Upper 50% (6 months: n=31; 9 months: n=29)	Lower 50% (6 months: n=27; 9 months: n=26)	*P* value[a]
6-min walk distance (m), mean (SD)			
6th month	483.3 (87.8)	502.7 (88.6)	.41
9th month	511.3 (70.2)	507.8 (108)	.78
Current smoking, n (%)			
6th month	3 (9.7)	8 (29.6)	.09
9th month	3 (10.3)	6 (23.1)	.28
Carboxyhemoglobin (%), mean (SD)			
6th month	1.2 (1.2)	1.6 (1.9)	.73
9th month	1.2 (1.2)	1.7 (1.1)	.046
Achieving 4 of 5, n (%)			
6th month	13 (41.9)	10 (37)	.79
9th month	15 (51.7)	4 (15.4)	.01
Achieving 5 of 5, n (%)			
6th month	3 (9.7)	2 (7.4)	>.99
9th month	5 (17.2)	1 (3.8)	.20
Achieving ≥4 of 5, n (%)			
6th month	16 (51.6)	12 (44.4)	.61
9th month	20 (69)	5 (19.2)	<.001

[a]Independent *t* test between groups. Data are presented as mean (SD) for continuous variables and n (%) for categorical variables.

[b]BP: blood pressure.

[c]HDL-C: high-density lipoprotein cholesterol.

[d]LDL-C: low-density lipoprotein cholesterol.

[e]HbA_{1c}: hemoglobin A_{1c}.

Ancillary Analysis: Blood Pressure Responders

In a subgroup analysis, patients in the intervention group were categorized based on whether they demonstrated improvement in blood pressure by 9 months (n=23) or not (n=33). Patients were categorized as blood pressure responders if their follow-up blood pressure was less than or equal to 130/80 mmHg at 9 months. Responders had significantly higher baseline systolic and DBPs but showed greater reductions over time. At 9 months, mean SBP was lower in the responder group (121.4 vs 130.8 mmHg; *P*=.02), and DBP was similarly reduced (75.9 vs 84.3 mmHg; *P*=.03; Table 5). Responders also read significantly more messages (110.3 vs 83.3 messages, *P*=.02) and achieved better overall CV risk factor control. They were more likely to meet LDL-C targets at 6 and 9 months (*P*=.06 and .05) and maintain BP less than 140/90 mmHg at 9 months (*P*=.052). At 6 months, 73.9% of the BP-improved group met at least 4 guideline-recommended goals compared to 31.4% in the nonimproved group (*P*=.04, Table S4 in Multimedia Appendix 1). At 9 months, this pattern persisted, with 71.4% of the improved group versus 30.3% of the nonimproved group achieving at least 4 of 5 goals (*P*=.045, Table S4 in Multimedia Appendix 1).

XSL•FO
RenderX

Table . Blood pressure (BP) improvement at 9 months.

Variables	BP-improved group (baseline: n=23; 6 months: n=23; 9 months: n=21)	BP not improved group (baseline: n=37; 6 months: n=35; 9 months: n=33)	*P* value[a]
Days messages were read (d), mean (SD)	110.26 (34.4)	83.30 (49.1)	.02
Systolic BP (mmHg), mean (SD)			
Baseline	135.54 (14.23)	118.32 (13.57)	<.001
6th month	125 (16.40)	123.48 (16.81)	.74
9th month	121.35 (14.82)	130.75 (13)	.02
Delta	−14.98 (10.42)	13.93 (13.08)	<.001
Diastolic BP (mmHg), mean (SD)			
Baseline	79.23 (12.13)	69.43 (8.11)	<.001
6th month	79.36 (12.13)	80.79 (11.05)	.65
9th month	75.95 (12.30)	84.25 (12.92)	.03
Delta	−2.89 (9.43)	14.65 (13)	<.001
Age (y), mean (SD)	58.8 (7.4)	58.3 (9.8)	.82
Male sex, n (%)	21 (91.3)	30 (81.1)	.28
Acute coronary syndrome, n (%)	12 (52.2)	30 (81.1)	.35
Hypertension, n (%)	13 (56.5)	23 (62.2)	.36
Diabetes mellitus, n (%)	6 (26.1)	7 (18.9)	.02
Dyslipidemia, n (%)	10 (43.5)	18 (48.6)	.72
Heart failure, n (%)	1 (4.3)	1 (2.7)	>.99
Stroke, n (%)	0 (0)	0 (0)	N/A[b]
Total cholesterol (mg/dL)			
Baseline	172.22 (51.15)	169.19 (38.47)	.80
6th month	123.36 (22.32)	134.75 (27.78)	.12
9th month	128.70 (31.29)	138.96 (31.62)	.28
LDL-C[c] (mg/dL), mean (SD)			
Baseline	101.48 (39.57)	95.30 (35.62)	.54
6th month	62.77 (19.01)	69.84 (21.41)	.22
9th month	69.10 (26.70)	71.55 (32)	.79
HDL-C[d] (mg/dL), mean (SD)			
Baseline	39.57 (7.38)	47.19 (13.86)	.02
6th month	42.91 (9.46)	44.94 (10)	.46
9th month	41.10 (8.09)	51.28 (21.92)	.05
Triglycerides (mg/dL), mean (SD)			
Baseline	184.26 (207.04)	146.11 (65.90)	.30
6th month	133.91 (53.19)	154.44 (119.35)	.45
9th month	142.85 (68.24)	133.88 (74.88)	.07
HbA_{1c}[e] (%), mean (SD)			
Baseline	6.13 (1.09)	6.02 (0.99)	.07
6th month	6.05 (0.80)	6 (0.70)	.83
9th month	6.12 (0.90)	6.07 (0.61)	.85

XSL•FO
RenderX

Variables	BP-improved group (baseline: n=23; 6 months: n=23; 9 months: n=21)	BP not improved group (baseline: n=37; 6 months: n=35; 9 months: n=33)	*P* value[a]
6-min walk distance (m), mean (SD)			
Baseline	433.35 (92.92)	414.46 (144.77)	.54
6th month	481.95 (80.26)	500.18 (93.70)	.48
9th month	510.56 (66.86)	508.87 (105.41)	.95

[a]Independent *t* test between groups.

[b]N/A: not applicable.

[c]LDL-C: low-density lipoprotein cholesterol.

[d]HDL-C: high-density lipoprotein cholesterol.

[e]HbA_{1c}: hemoglobin A_{1c}.

Process Outcomes: App Acceptability and User Satisfaction

Among 53 intervention participants who completed the satisfaction survey at 6 months, overall acceptability of the messaging intervention was high. Most respondents reported that the messages were easy to understand and helpful (n=45, 84.9%), expressed willingness to continue receiving messages (n=42, 79.2%), and indicated that they would recommend the app to others (84.9%). Detailed item-level acceptability and usability ratings are provided in Table S5 in Multimedia Appendix 1. We further explored whether engagement varied by participant characteristics. Linear regression analyses showed weak, nonsignificant associations between age and engagement metrics: age was not significantly associated with message-reading behavior (R^2=0.04; *P*=.13; Figure 3A) or with diary engagement (R^2=0.02; *P*=.34; Figure 3B). However, message reading frequency differed by education level, with elementary school graduates reading significantly fewer messages (*P*=.045; Table S6 in Multimedia Appendix 1). Satisfaction and health diary input frequency did not significantly differ by education level.

Figure 3. Correlation between age and AnSim app usage. (A) A scatter plot depicting the relationship between participant age and the number of days messages were read. (B) A scatter plot showing the relationship between participant age and the number of health diary entries recorded. Each point represents one participant, and the solid line represents the fitted linear regression.

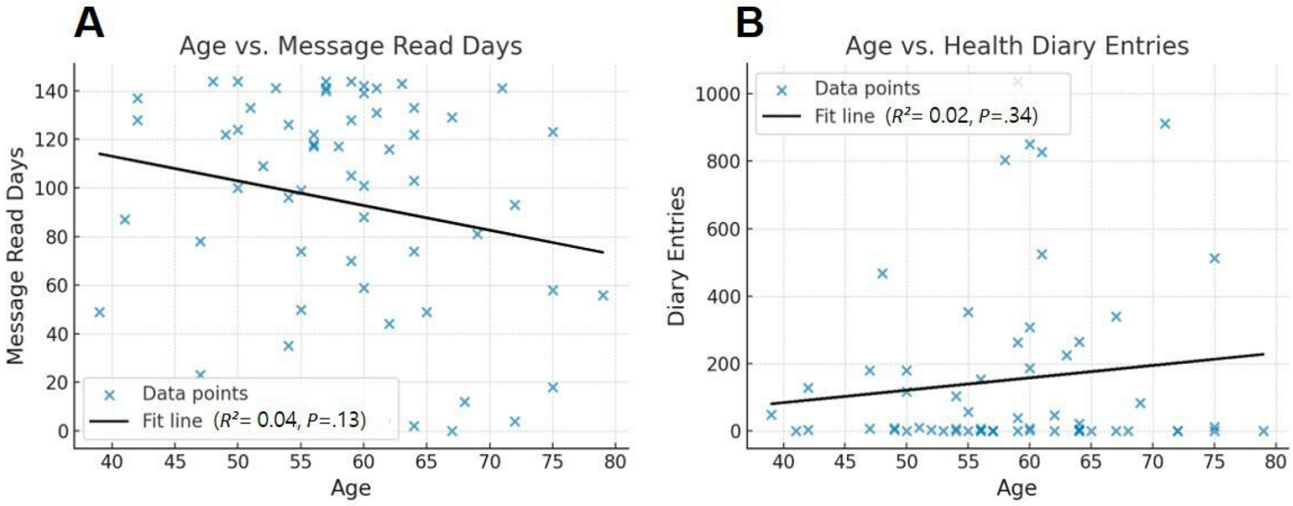

Discussion

Principal Findings and Interpretation

In this single-blinded randomized controlled trial, we evaluated the effectiveness of a patient-specific, smartphone-based messaging app (AnSim) on CV risk factor control in patients following PCI. Relative to our primary objectives, there were no significant between-group differences in changes in SBP or DBP. Within the intervention arm, greater engagement with messaging was associated with a higher likelihood of achieving guideline-recommended secondary prevention targets. Participants who experienced clinically relevant reductions in blood pressure at 9 months read more messages and demonstrated greater health behavior engagement, suggesting a potential dose–response relationship between digital engagement and risk factor control.

Several factors may have contributed to the lack of between-group differences in blood pressure. The participants had relatively well-controlled baseline blood pressure and lipid profiles, likely due to recent coronary intervention and optimal medical therapy. These ceiling effects may have limited the observable impact of the intervention. Moreover, adherence to medication and behavior change tend to be highest immediately following PCI, narrowing the margin for improvement and attenuating between-group differences [29]. A staged or adaptive

XSL•FO
RenderX

approach, with reinforcement delivered later in recovery or targeted to patients with suboptimal baseline control, may be more likely to demonstrate incremental effects.

Several digital health interventions have demonstrated beneficial effects on CV outcomes, medication adherence, and behavior change [30,31]. Delivery of a comprehensive, app-supported rehabilitation program in the SMART-REHAB (SMARTphone-based early cardiac REHABilitation) trial was associated with improvements in physical recovery outcomes and greater enrollment and adherence to CR [32]. Compared to SMART-REHAB, which used structured exercise and real-time tracking, our AnSim intervention relied on passive, message-based behavioral prompting. This design distinction may explain differences in observed outcomes. Additionally, the baseline blood pressure of our participants was relatively well controlled due to recent PCI and optimal medical therapy, similar to the findings in SMART-REHAB where prescription rates for secondary prevention pharmacotherapy were consistently high. Accordingly, any incremental benefit of a low-intensity, message-only intervention on blood pressure is likely to be modest and may require longer follow-up or a higher-risk subgroup to be detectable.

Our results are consistent with the TEXT-ME study [18], which found that messaging helped maintain favorable risk profiles over time rather than lowering already well-controlled parameters. Additionally, our findings highlight the heterogeneity in responsiveness to digital interventions. Consistent with prior trials [33,34], our results demonstrate that engagement, not mere exposure, is a primary driver of digital intervention efficacy. Participants in the top half of message readership demonstrated greater health diary input and a higher likelihood of meeting at least 4 CV risk factor targets. Similarly, those who experienced improvements in blood pressure had significantly higher message exposure. This engagement-dependent effect parallels findings from a secondary analysis of a large CV text messaging trial, in which participants who actively engaged with the messages demonstrated significantly higher 12-month medication adherence than nonresponders, despite overall neutral trial results [35]. Such results support the view that the “dose” of digital engagement matters, and simply offering mobile health solutions may be insufficient without ensuring user uptake and sustained participation. Practically, it is essential to sustain participation through tailored content and timing, and to add a feedback loop that converts passive messages into actionable tasks (self-monitoring, goal tracking, and reinforcement). For users showing early drop-off, a stepped-care approach, such as automated reminders followed by brief coaching, may help restore engagement [36].

Acceptability, Usability, and Durability of Effects

This trial represents an early real-world evaluation of AnSim, as acceptability and usability are central to interpreting scalability. Overall acceptability of the messaging content was high (eg, clarity/helpfulness, willingness to continue, and recommendation intent), supporting feasibility; however, overall usability ratings were comparatively lower, suggesting the presence of workflow friction that may limit sustained engagement [37]. To ensure digital equity, we implemented a simplified, user-friendly interface with large fonts and intuitive navigation, and provided face-to-face training at enrollment to overcome initial technical barriers. These measures likely contributed to our finding that age was not a significant determinant of message engagement. In contrast, education level remained an influencing factor; participants with lower education levels tended to read fewer messages, highlighting a persistent gap. This suggests that while simplified interfaces can bridge the age gap, addressing literacy barriers requires further tailoring [38]. Future interventions should focus on minimizing workflow friction and incorporating visual content with adaptive education levels to improve inclusivity and reduce disparities.

We observed a sustained and increased effect of the intervention over follow-up, which may reflect delayed behavior adoption or reinforcement over time. As shown in Table 4, the proportion of participants in the intervention group achieving multiple guideline-recommended risk factor targets was sustained or even slightly improved after the cessation of the 6-month messaging intervention. In contrast, the control group showed a marked decline in target achievement over the same period. These patterns are hypothesis-generating and suggest that the behavioral improvements facilitated by the AnSim messaging intervention may persist beyond the active intervention phase, supporting the hypothesis of a delayed or sustained effect. Given that secondary prevention requires lifelong adherence, interventions that promote sustained behavior change and long-term engagement may be particularly valuable, warranting longer-term evaluation.

Strengths and Limitations

We used a randomized design with blinded outcome assessment and standardized follow-up to evaluate the incremental effect of a messaging intervention on risk-factor control after PCI. Strengths include randomized design, theory-based messaging content, integration of behavior change staging, and concurrent evaluation of clinical outcomes, acceptability, and intervention engagement. To support participant blinding, the Heart Keeper app was installed in both groups; Heart Keeper functioned as a simple diary-style tool for recording clinical information and did not deliver structured behavior-change messaging or coaching.

However, there are several limitations. First, the modest sample size and short follow-up period limited the power to detect differences in long-term CV outcomes. Second, we did not systematically capture detailed reasons for nonenrollment (including inability to use a smartphone), which limits the assessment of representativeness and generalizability, particularly for populations with lower digital access or literacy. Third, providing a diary-style logging application to both groups may have encouraged self-monitoring and reduced between-group separation. Fourth, our intervention lacked multimedia interactivity, real-time feedback, or structured exercise modules, which may have reduced efficacy. Regulatory restrictions in Korea also limited bidirectional communication, a component emphasized in SMART-REHAB as essential for behavior reinforcement. Furthermore, because Heart Keeper

XSL•FO
RenderX

does not generate message-exposure metrics comparable to AnSim, objective engagement data were not available for the control group, precluding direct between-group comparisons of digital engagement intensity. Finally, we did not conduct qualitative interviews among participants with low engagement; future studies should incorporate user-centered qualitative methods to identify modifiable barriers and optimize equitable uptake.

Implications for Practice and Future Research

The current study underscores the importance of personalization and engagement in digital secondary prevention support. While app-based interventions can extend the reach of CR programs, their impact depends heavily on user activation. Future versions of AnSim should incorporate co-design with patients and clinicians, formative usability testing, and stakeholder engagement to optimize fit-for-purpose delivery. Adaptive intervention timing (reinforcement initiated later in recovery or triggered by loss of control/engagement) may better align with patients' evolving needs. Integrating app data with electronic medical records and care teams could facilitate more proactive medical adjustments, similar to nurse-led programs or digital case management models [39].

Larger trials with longer follow-up are needed to determine whether these interventions improve clinical outcomes such as hospital readmissions and mortality. Comparative effectiveness studies between full-scale digital rehabilitation programs and lighter interventions like AnSim will also help clarify the optimal design and intensity of mobile CR. Although messaging interventions may be low-cost per participant, implementation requires development, maintenance, and monitoring resources; future work should assess cost-effectiveness, including staff workload if escalation or bidirectional communication is introduced, and evaluate scalability in routine post-PCI care.

Conclusions

This randomized trial demonstrated that a personalized, smartphone-based messaging intervention did not significantly improve blood pressure or lipid levels after PCI compared with usual care. In exploratory post hoc analyses within the intervention group, greater engagement with the AnSim app was associated with better risk-factor target attainment, with signals suggesting a larger potential benefit among participants with higher baseline blood pressure. These findings should be interpreted as hypothesis-generating and underscore engagement as a key determinant of potential impact for low-intensity mobile health strategies. Unlike comprehensive digital CR programs that incorporate structured exercise modules, real-time monitoring, or bidirectional coaching, AnSim was intentionally designed as a lightweight, message-mediated intervention that can be deployed with minimal resource burden. As a practical implication, these results suggest that real-world impact will likely depend less on message availability and more on implementation strategies that improve uptake and sustained engagement (eg, adaptive timing, stepped support, and integration into routine follow-up workflows). Future studies should test interventions designed to improve uptake and sustained participation, evaluate targeted or adaptive delivery for patients at higher residual risk, ensure digital equity for individuals with lower digital literacy, and assess longer-term clinical outcomes.

Acknowledgments

The authors sincerely thank all participants who took part in the AnSim trial. They gratefully acknowledge the contributions of Yeonghun Song and Gyeongho Jung from the IT business division at Hanmi Healthcare for their work in app development and study operations. They also thank Seungeon Shin for her assistance with participant recruitment and study coordination.

Generative artificial intelligence (AI) tools were used under full human supervision only for language editing (clarity, grammar, readability). According to the GAIDeT taxonomy (2025), these tasks were supported using ChatGPT (OpenAI; GPT-5.2). The authors reviewed and edited all AI-assisted text and take full responsibility for the final manuscript; no generative AI tool is listed as an author or bears responsibility for the study design, analyses, results, or conclusions. This disclosure was submitted by SL.

Funding

This research was supported by a grant from the Korea Health Industry Development Institute, funded by the Ministry of Health and Welfare, Republic of Korea (grant HI16C0483). The funders had no role in the design of the study, data collection and analysis, interpretation of data, or writing of the manuscript.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to institutional review board restrictions and participant privacy considerations. Deidentified data may be available from the corresponding author upon reasonable request, contingent on institutional approval and execution of a data use agreement.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Comparison of the 2 apps, examples of developed messages, comparison of clinical events, guideline-recommended goals according to blood pressure improvement, acceptability and usability of the AnSim app at 6 months, and app utilization analysis according to education level.
[DOCX File, 38 KB - jmir_v28i1e81524_app1.docx]

Checklist 1
CONSORT-EHEALTH checklist (version 1.6.1).
[PDF File, 1438 KB - jmir_v28i1e81524_app2.pdf]

Checklist 2
CONSORT 2025 checklist.
[PDF File, 281 KB - jmir_v28i1e81524_app3.pdf]

References

1. Nichols M, Townsend N, Scarborough P, Rayner M. Cardiovascular disease in Europe 2014: epidemiological update. Eur Heart J 2014 Nov 7;35(42):2950-2959. [doi: 10.1093/eurheartj/ehu299] [Medline: 25139896]
2. Kang HJ, Wang J, Cho EJ, et al. Addressing cardiovascular diseases challenges in South Korea: strategies to improve outcomes. Korean Circ J 2025 Jul;55(7):557-583. [doi: 10.4070/kcj.2024.0385] [Medline: 40345831]
3. Zeitouni M, Clare RM, Chiswell K, et al. Risk factor burden and long-term prognosis of patients with premature coronary artery disease. J Am Heart Assoc 2020 Dec 15;9(24):e017712. [doi: 10.1161/JAHA.120.017712] [Medline: 33287625]
4. Kodeboina M, Piayda K, Jenniskens I, et al. Challenges and burdens in the coronary artery disease care pathway for patients undergoing percutaneous coronary intervention: a contemporary narrative review. Int J Environ Res Public Health 2023 Apr 25;20(9):5633. [doi: 10.3390/ijerph20095633] [Medline: 37174152]
5. Brown TM, Pack QR, Aberegg E, et al. Core components of cardiac rehabilitation programs: 2024 update: a scientific statement from the American Heart Association and the American Association of Cardiovascular and Pulmonary Rehabilitation. Circulation 2024 Oct 29;150(18):e328-e347. [doi: 10.1161/CIR.0000000000001289] [Medline: 39315436]
6. Anderson L, Oldridge N, Thompson DR, et al. Exercise-based cardiac rehabilitation for coronary heart disease: cochrane systematic review and meta-analysis. J Am Coll Cardiol 2016 Jan 5;67(1):1-12. [doi: 10.1016/j.jacc.2015.10.044] [Medline: 26764059]
7. Eijsvogels TMH, Maessen MFH, Bakker EA, et al. Association of cardiac rehabilitation with all-cause mortality among patients with cardiovascular disease in the Netherlands. JAMA Netw Open 2020 Jul 1;3(7):e2011686. [doi: 10.1001/jamanetworkopen.2020.11686] [Medline: 32716516]
8. Shields GE, Wells A, Doherty P, Heagerty A, Buck D, Davies LM. Cost-effectiveness of cardiac rehabilitation: a systematic review. Heart 2018 Sep;104(17):1403-1410. [doi: 10.1136/heartjnl-2017-312809] [Medline: 29654096]
9. Sugiharto F, Nuraeni A, Trisyani Y, Melati Putri A, Aghnia Armansyah N. Barriers to participation in cardiac rehabilitation among patients with coronary heart disease after reperfusion therapy: a scoping review. Vasc Health Risk Manag 2023;19:557-570. [doi: 10.2147/VHRM.S425505] [Medline: 37671387]
10. Turk-Adawi K, Supervia M, Lopez-Jimenez F, et al. Cardiac rehabilitation availability and density around the globe. EClinicalMedicine 2019;13:31-45. [doi: 10.1016/j.eclinm.2019.06.007] [Medline: 31517261]
11. Taylor RS, Fredericks S, Jones I, et al. Global perspectives on heart disease rehabilitation and secondary prevention: a scientific statement from the Association of Cardiovascular Nursing and Allied Professions, European Association of Preventive Cardiology, and International Council of Cardiovascular Prevention and Rehabilitation. Eur Heart J 2023 Jul 21;44(28):2515-2525. [doi: 10.1093/eurheartj/ehad225] [Medline: 37477626]
12. Wang L, Liu J, Fang H, Wang X. Factors associated with participation in cardiac rehabilitation in patients with acute myocardial infarction: a systematic review and meta-analysis. Clin Cardiol 2023 Nov;46(11):1450-1457. [doi: 10.1002/clc.24130] [Medline: 37594292]
13. Turk-Adawi KI, Oldridge NB, Tarima SS, Stason WB, Shepard DS. Cardiac rehabilitation patient and organizational factors: what keeps patients in programs? J Am Heart Assoc 2013 Oct 21;2(5):e000418. [doi: 10.1161/JAHA.113.000418] [Medline: 24145743]
14. Beatty AL, Beckie TM, Dodson J, et al. A new era in cardiac rehabilitation delivery: research gaps, questions, strategies, and priorities. Circulation 2023 Jan 17;147(3):254-266. [doi: 10.1161/CIRCULATIONAHA.122.061046] [Medline: 36649394]
15. Duncan MS, Robbins NN, Wernke SA, et al. Geographic variation in access to cardiac rehabilitation. J Am Coll Cardiol 2023 Mar 21;81(11):1049-1060. [doi: 10.1016/j.jacc.2023.01.016] [Medline: 36922091]
16. Bomtempo APD, Ghisi GLDM. Exploring cardiac rehabilitation barriers across health literacy levels. Health Lit Res Pract 2025 Apr;9(2):e72-e77. [doi: 10.3928/24748307-20241127-02] [Medline: 40489385]

XSL•FO
RenderX

17. Ramesh SH, Jull D, Fournier H, Rajabiyazdi F. Exploring barriers to patients' progression in the cardiac rehabilitation journey from health care providers' perspectives: qualitative study. Interact J Med Res 2025 Feb 21;14:e66164. [doi: 10.2196/66164] [Medline: 39983120]
18. Chow CK, Redfern J, Hillis GS, et al. Effect of lifestyle-focused text messaging on risk factor modification in patients with coronary heart disease: a randomized clinical trial. JAMA 2015;314(12):1255-1263. [doi: 10.1001/jama.2015.10945] [Medline: 26393848]
19. Li Z, Long Y, Yang Q, Liu J, Wang Y. Smartphone-based interventions in the secondary prevention of cardiovascular disease: a meta-analysis. Technol Health Care 2023;31(5):1935-1948. [doi: 10.3233/THC-230022] [Medline: 37125588]
20. Spaulding EM, Marvel FA, Piasecki RJ, Martin SS, Allen JK. User engagement with smartphone apps and cardiovascular disease risk factor outcomes: systematic review. JMIR Cardio 2021 Feb 3;5(1):e18834. [doi: 10.2196/18834] [Medline: 33533730]
21. Lunz L, Würth S, Kulnik ST. Health care professionals' use of digital technology in the secondary prevention of cardiovascular disease in Austria: online survey study. JMIR Cardio 2025 Jun 25;9:e71366. [doi: 10.2196/71366] [Medline: 40561497]
22. Golbus JR, Lopez-Jimenez F, Barac A, et al. Digital technologies in cardiac rehabilitation: a science advisory from the American Heart Association. Circulation 2023 Jul 4;148(1):95-107. [doi: 10.1161/CIR.0000000000001150] [Medline: 37272365]
23. The effects of patient-specific smartphone application for cardiac rehabilitation on improvement of risk factors of coronary artery disease and medical expenses reduction. Clinical Research Information Service (CRIS). URL: https://cris.nih.go.kr/cris/search/detailSearch.do?seq=7230&search_page=L [accessed 2026-04-17]
24. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ 2025 Apr 14;389:e081123. [doi: 10.1136/bmj-2024-081123] [Medline: 40228833]
25. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
26. Choi JY, Kim JB, Lee S, et al. A smartphone app (AnSim) with various types and forms of messages using the transtheoretical model for cardiac rehabilitation in patients with coronary artery disease: development and usability study. JMIR Med Inform 2021 Dec 7;9(12):e23285. [doi: 10.2196/23285] [Medline: 34878987]
27. Abraham C, Michie S. A taxonomy of behavior change techniques used in interventions. Health Psychol 2008 May;27(3):379-387. [doi: 10.1037/0278-6133.27.3.379] [Medline: 18624603]
28. Schimmel J, George N, Schwarz J, Yousif S, Suner S, Hack JB. Carboxyhemoglobin levels induced by cigarette smoking outdoors in smokers. J Med Toxicol 2018 Mar;14(1):68-73. [doi: 10.1007/s13181-017-0645-1] [Medline: 29285649]
29. Mathews R, Peterson ED, Honeycutt E, et al. Early medication nonadherence after acute myocardial infarction: insights into actionable opportunities from the Treatment With ADP Receptor Inhibitors: longitudinal Assessment of Treatment Patterns and Events After Acute Coronary Syndrome (TRANSLATE-ACS) study. Circ Cardiovasc Qual Outcomes 2015 Jul;8(4):347-356. [doi: 10.1161/CIRCOUTCOMES.114.001223] [Medline: 26038524]
30. Widmer RJ, Collins NM, Collins CS, West CP, Lerman LO, Lerman A. Digital health interventions for the prevention of cardiovascular disease: a systematic review and meta-analysis. Mayo Clin Proc 2015 Apr;90(4):469-480. [doi: 10.1016/j.mayocp.2014.12.026] [Medline: 25841251]
31. Azizi Z, Broadwin C, Islam S, et al. Digital health interventions for heart failure management in underserved rural areas of the United States: a systematic review of randomized trials. J Am Heart Assoc 2024 Jan 16;13(2):e030956. [doi: 10.1161/JAHA.123.030956] [Medline: 38226517]
32. Yudi MB, Clark DJ, Tsang D, et al. SMARTphone-based, early cardiac REHABilitation in patients with acute coronary syndromes: a randomized controlled trial. Coron Artery Dis 2021 Aug 1;32(5):432-440. [doi: 10.1097/MCA.0000000000000938] [Medline: 32868661]
33. Kennedy F, Smith S, Beeken RJ, et al. An app-based intervention with behavioral support to promote brisk walking in people diagnosed with breast, prostate, or colorectal cancer (APPROACH): process evaluation study. JMIR Cancer 2025 Feb 10;11(1):e64747. [doi: 10.2196/64747] [Medline: 39928926]
34. Bress JN, Falk A, Schier MM, et al. Efficacy of a mobile app-based intervention for young adults with anxiety disorders: a randomized clinical trial. JAMA Netw Open 2024 Aug 1;7(8):e2428372. [doi: 10.1001/jamanetworkopen.2024.28372] [Medline: 39163044]
35. Shore-Plavec R, Zucker R, Glorioso TJ, et al. Engagement with text messaging improves cardiovascular medication adherence: secondary analysis of a randomized controlled trial. J Med Internet Res 2025 Nov 24;27:e80794. [doi: 10.2196/80794] [Medline: 41284985]
36. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462. [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]

XSL•FO
RenderX

37. Jakob R, Harperink S, Rudolf AM, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371. [doi: 10.2196/35371] [Medline: 35612886]
38. Amouzadeh E, Dianat I, Faradmal J, Babamiri M. Optimizing mobile app design for older adults: systematic review of age-friendly design. Aging Clin Exp Res 2025 Aug 14;37(1):248. [doi: 10.1007/s40520-025-03157-7] [Medline: 40804492]
39. Lin J, Bates SM, Allen LN, Wright M, Mao L, Kidd M. Integrating mobile health app data into electronic medical or health record systems and its impact on health care delivery and patient health outcomes: scoping review. JMIR mHealth uHealth 2025 Jun 23;13:e66650. [doi: 10.2196/66650] [Medline: 40549960]

Abbreviations

BP: blood pressure
CAD: coronary artery disease
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
CR: cardiac rehabilitation
CRIS: Clinical Research Information Service
CV: cardiovascular
DBP: diastolic blood pressure
HbA_{1c}: hemoglobin A_{1c}
LDL-C: low-density lipoprotein cholesterol
MCAR: missing completely at random
PCI: percutaneous coronary intervention
SBP : systolic blood pressure
SMART-REHAB: SMARTphone-based early cardiac REHABilitation
TEXT-ME: Tobacco, Exercise and Diet Messages

Edited by S Brini; submitted 30.Jul.2025; peer-reviewed by L Allan, S Payne; revised version received 13.Mar.2026; accepted 17.Mar.2026; published 23.Apr.2026.

Please cite as:
Lee S, Kim JB, Shin SE, Ahn J, Jeong HG, Jung KH, Song Y, Kim EJ
Personalized Smartphone Messaging for Secondary Prevention After Percutaneous Coronary Intervention: Randomized Controlled Trial
J Med Internet Res 2026;28:e81524
URL: https://www.jmir.org/2026/1/e81524
doi:10.2196/81524

© Sunki Lee, Ji Bak Kim, Seung Eon Shin, Jeonghoon Ahn, Hyun-Ghang Jeong, Kyung Ho Jung, Younghoon Song, Eung Ju Kim. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Peer Support to Increase Walking Among Older Adults: Cluster Randomized Trial

Atsushi Nakagomi[1], MD, PhD; Noriyuki Abe[1], PhD; Takayuki Ueno[2], PhD; Gemmei Izuka[1], MD, PhD; Yohei Kawasaki[3], PhD; Katsunori Kondo[1,4], MD, PhD

[1]Department of Social Preventive Medical Sciences, Center for Preventive Medical Sciences, Chiba University, 1-33 Yayoicho, Inage-ku, Chiba, Japan

[2]Research Team for Social Participation and Healthy Aging, Tokyo Metropolitan Institute for Geriatrics and Gerontology, Tokyo, Japan

[3]Department of Biostatistics, Graduate School of Medicine, Saitama Medical University, Saitama, Japan

[4]Research Department, Institute for Health Economics and Policy, Tokyo, Japan

Corresponding Author:
Atsushi Nakagomi, MD, PhD
Department of Social Preventive Medical Sciences, Center for Preventive Medical Sciences, Chiba University, 1-33 Yayoicho, Inage-ku, Chiba, Japan

Abstract

Background: As the population ages, older adults face an increasing risk of physical inactivity and related health complications, highlighting the need for scalable interventions. Smartphone-based programs have emerged as a promising strategy to support sustained physical activity among older adults.

Objective: This study aimed to evaluate whether a smartphone lecture program incorporating a digital peer support app would increase physical activity among older adults, compared to a conventional smartphone lecture program.

Methods: This 2-arm, 1:1 parallel-arm, cluster-randomized trial was conducted in 2 urban regions of Japan (Sumida Ward, Tokyo, and Chiba City, Chiba). Eligible participants were community-dwelling adults aged ≥60 years, able to walk independently, and smartphone users; exclusion criteria included prior use of the peer support app or medical restrictions on walking. Participants were recruited offline during community smartphone lectures (closed-group recruitment). The intervention combined face-to-face lectures with app-based peer support, while outcomes were assessed both objectively (via smartphones) and through self-administered paper questionnaires. All participants received a baseline smartphone lecture. Intervention participants attended 2 additional sessions using a digital peer support app (Minchalle; A10 Lab Inc), which included features such as daily step goals, peer sharing, and group encouragement. Control participants attended 2 standard follow-up smartphone lectures. The primary outcome was the change in weekly average daily step count from baseline to Week 12. Secondary outcomes included total metabolic equivalent of task (MET)–minutes per week (assessed via the International Physical Activity Questionnaire), walking time (≥30 minutes per day), daily smartphone use, and number of smartphone use purposes.

Results: A total of 156 community-dwelling older adults were grouped into 40 clusters and randomized (20 intervention clusters, n=80 and 20 control clusters, n=76). In total, 124 participants (79.5%) completed the follow-up, and valid step data were available for 117 participants, with missing data ranging from 5.1% to 29.1%. Baseline daily steps averaged 3951 (SD 1686) in controls versus 4583 (SD 1973) in the intervention arm. An unadjusted mixed model for repeated measures showed significantly higher step changes for intervention participants at Week 12 (difference=579, 95% CI 36-1123; *P*=.04). No significant differences emerged for total METs (difference=646 MET-min per week, 95% CI –12 to 1303; *P*=.054) or walking ≥30 minutes per day (odds ratio [OR] 1.56, 95% CI 0.63-3.90; *P*=.33). However, the intervention arm demonstrated a significant increase in daily smartphone use (OR 4.10, 95% CI 1.15-14.6; *P*=.03) and in the number of smartphone use purposes (difference=0.58, 95% CI 0.12-1.05; *P*=.01).

Conclusions: A smartphone lecture program integrated with app-based peer support led to modest but meaningful improvements in step counts among older Japanese adults, at Week 12 of the 12-week intervention. Future research should investigate long-term maintenance, additional measures of physical activity, and subpopulation responses to optimize digital health programs for older adults.

Trial Registration: UMIN-CTR UMIN000051904; https://tinyurl.com/4m4zm99v

(J Med Internet Res 2026;28:e75708) doi:10.2196/75708

KEYWORDS

peer support; smartphone; digital health; physical activity; older adults; digital divide

XSL•FO
RenderX

Introduction

Physical activity is a key factor in healthy aging, helping prevent chronic diseases, maintain functional independence, and improve quality of life [1,2]. In Japan, where a rapidly aging population has raised concerns about inactivity and health care costs, the Ministry of Health, Labor, and Welfare (MHLW) has taken steps to address the issue. As part of the Healthy Japan 21 (3rd term) initiative, the MHLW set a target of 6000 steps per day for older adults by 2032, based on 2019 data showing averages of 5396 steps for men and 4656 for women [3]. The MHLW also released the Physical Activity and Exercise Guide for Health Promotion 2023 [4], emphasizing that even small increases in light-intensity activity yield significant health benefits, particularly for sedentary individuals [5]. As Japan's older population continues to grow, scalable strategies to promote physical activity are an urgent public health need.

Evidence suggests behavior change strategies can promote sustained physical activity among older adults [6,7]. Intrapersonal strategies—such as action planning, goal setting, commitment, and self-monitoring—are commonly used in physical activity interventions [8,9]. Interpersonal strategies, especially peer support, are also gaining traction [10-12]. Social cognitive theory (SCT) provides a strong theoretical basis for these strategies, as it emphasizes the dynamic interaction between individual, behavioral, and social-environmental factors [13]. Key constructs of SCT include self-efficacy (confidence in one's ability to maintain behavior), observational learning (acquiring new behaviors by observing others), and reinforcement (sustaining behaviors through social and environmental feedback). By drawing attention to how individual agency and social context interact, SCT offers a useful framework for understanding how both intrapersonal and interpersonal strategies can promote lasting engagement in physical activity among older adults.

With the growth of digital health technologies, many mobile apps now integrate behavior change techniques [14]. Alongside intrapersonal features, peer-support elements such as group challenges and shared goals are increasingly common [15-17]. For example, peer support can be structured in a smartphone app (Minchalle; A10 Lab Inc) through small groups of 3-5 participants who collaboratively select step goals, commit to posting updates at regular intervals, and encourage each other's progress [17,18]. Integrating such peer support with intrapersonal strategies within the app has been shown to enhance user engagement and adherence to health goals [17,18].

However, older adults often face barriers—known as the "digital divide"—that hinder their use of technology-based interventions [19]. In Japan, local governments and community groups have introduced smartphone lectures to bridge this gap, yet sustained smartphone use remains a challenge. Digital peer support may offer a more lasting solution by providing ongoing group-based reinforcement, promoting both technology use and healthier behaviors. Still, few studies have compared smartphone lectures with and without digital peer support, highlighting a gap in empirical evidence.

This study evaluates whether adding digital peer support to smartphone lectures increases physical activity among older Japanese adults. Using a cluster-randomized controlled trial, we compared a conventional smartphone lecture to one combined with a peer support app grounded in behavioral science. A cluster design was required because the intervention was implemented within peer groups of 3-5 participants; randomizing individuals within the same group would have compromised the integrity of the peer support mechanism and risked contamination between trial arms. The primary objective of the trial was to evaluate changes in daily step counts at the individual participant level, while randomization occurred at the cluster (group) level to preserve the integrity of the peer support intervention. Participants were recruited in community settings and randomly assigned to either group. This design allowed us to examine whether ongoing digital peer support enhances motivation, sustains smartphone engagement, and improves physical activity outcomes.

Methods

Overview

The study was conducted in 2 urban municipalities within the Tokyo metropolitan region, Sumida Ward (Tokyo) and Chiba City (Chiba Prefecture). The trial was first implemented in Sumida Ward from September to December 2023 in collaboration with the municipal government and subsequently replicated in Chiba City from June to September 2024 through collaborations with local senior clubs. Although implemented in different municipalities, both settings are urban communities located near central Tokyo, allowing the study procedures to be applied consistently. This manuscript follows the 2010 CONSORT (Consolidated Standards of Reporting Trials) guidelines for cluster randomized trials (see Table S1 in Multimedia Appendix 1) [20] and the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) guidelines (Checklist 1) [21].

Design

This was a 2-arm, 1:1 parallel-arm, cluster-randomized trial that included a 2-week baseline measurement period followed by a 12-week intervention period. The unit of randomization (the cluster) was a peer support group consisting of 3-5 participants. In the intervention arm, these clusters functioned as peer groups within the Minchalle app, while in the control arm, participants were grouped to maintain comparability but did not engage in peer support during the trial. Control participants were informed that they would be offered the opportunity to form peer support groups after the trial (waiting-list design). Randomization at the cluster level was required to preserve the integrity of arm allocation.

Participants and Recruitment

Recruitment took place from September 2023 to June 2024, in partnership with the Sumida Ward local government and senior clubs in Chiba City. Individuals were recruited in community venues during smartphone lectures hosted by the local municipal government or senior clubs, which aimed to bridge the digital

XSL•FO
RenderX

divide and promote physical activity to prevent frailty among older adults. The local government and senior clubs outsourced lecture services to A10 Lab Inc., developer of the “Minchalle” app (Challenge with All), and Active Senior Outreach Nonprofit Organization, a nonprofit organization specializing in smartphone lectures for older adults.

These smartphone lectures were implemented as part of the annual programs conducted by the municipalities and senior clubs and were held in public community venues, typically accommodating 15 - 30 older adults per session. Participants were approached in person at the beginning of the first session. Researchers informed all attendees about the study using slides and printed information sheets. Attendees were eligible to participate regardless of prior interest, and every individual attending the session was approached, ensuring a consistent and nonselective recruitment process.

Participants were eligible if they (1) were ≥60 years old, (2) could communicate in Japanese, (3) could read and provide informed consent, and (4) owned a smartphone at the time of consent. Participants were excluded if they (1) had any condition making participation infeasible (eg, inability to walk with a smartphone), (2) had a medical restriction on walking (eg, per doctor’s advice), or (3) had already used the Minchalle app.

Interested individuals were then screened by the research team based on the exclusion criteria. Those who met all criteria received a detailed explanation of the study procedures and provided written informed consent on-site. No participant was recruited through individual referrals, advertisements, or convenience sampling outside of the lecture setting.

Clusters were formed by grouping 3-5 participants who attended the same lecture session. No eligibility criteria were applied at the cluster level, and only individuals who provided informed consent were included. Individuals who did not meet the inclusion criteria or declined to participate were allowed to attend the lecture activities but were not included in the study cohort.

Initial Session and Assessment

Figure 1 outlines the study procedures. Consenting participants were organized into groups of 3 - 5 and attended a smartphone lecture, followed by completion of baseline questionnaires. The lecture included a 15-minute presentation by a medical doctor on the health benefits of walking and a one-hour training session on installing and using health-tracking apps—Google Fit (Google LLC; Android) and Apple Health app (Apple Inc; iOS). Participants were encouraged to use basic app features, such as step counting and daily goal setting, incorporating intrapersonal behavior change strategies (action planning, goal setting, commitment, and self-monitoring). Participants were instructed to carry their smartphones daily to ensure continuous step count recording.

Figure 1. Timeline of study procedures.

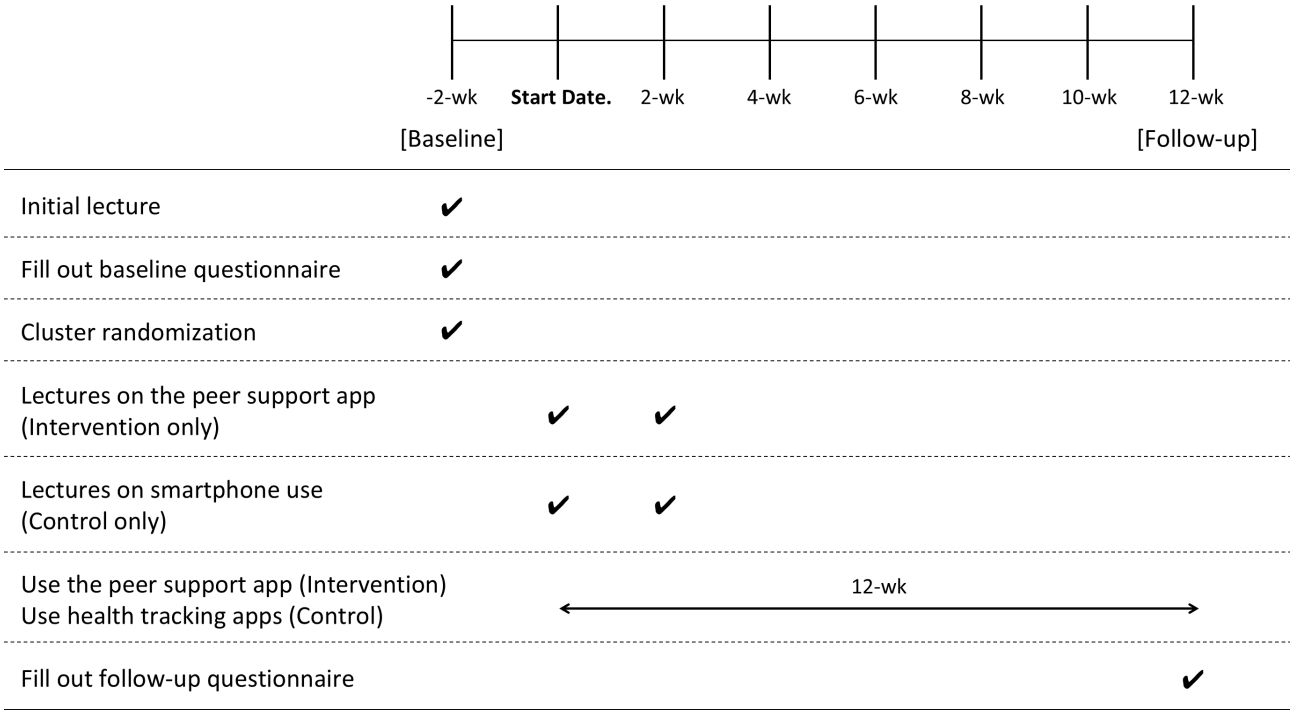

Before leaving the session, research staff verified successful installation of the app and confirmed that step-count tracking was active. Baseline paper questionnaires were completed on-site and collected immediately by trained staff. Baseline questionnaire data were linked to each participant using a deidentified study ID.

Follow-Up Data Collection

At the 12-week follow-up session, participants completed follow-up paper questionnaires. During this session, research staff guided participants through the in-app procedures required to export and send their accumulated step count data to the Minchalle server using each participant’s deidentified study ID. Only participants who attended the follow-up session and completed these in-app actions were able to transmit their step

XSL•FO
RenderX

data. The app provider then securely transferred the deidentified step-count dataset to the research team, where it was merged with questionnaire responses using the study ID. No personally identifiable information was shared, and all data were stored on secure, access-restricted servers.

Randomization

Following the initial session and baseline assessment, randomization was conducted at the cluster level, with each group of 3-5 participants serving as the unit of allocation. The random allocation sequence was generated electronically using a random number generator with block sizes of 2-4 clusters. An independent allocation officer from the Institute for Assistance of Academic and Education implemented the sequence. The allocation sequence and block size were concealed from study personnel during enrollment, and arm assignments were revealed only after enrollment was finalized.

Individual participants were enrolled during community smartphone lectures and grouped pragmatically into clusters of 3-5 based on attendance at the same session. Randomization was therefore performed at the cluster level, while allocation concealment and consent were handled at the individual participant level.

Control arm participants were informed they would receive the intervention content after the trial (ie, a waiting-list control design).

Blinding

Because of the nature of the intervention, participants and local staff could not be blinded to arm assignment. However, outcome assessment was based on step count data automatically recorded by participants' smartphones, and data analysts were blinded to allocation, reducing the risk of bias.

The 2 study arms differed in that only the intervention arm received access to the Minchalle app and related training. Therefore, the interventions could not be made similar in appearance or content.

Intervention

The intervention was a smartphone lecture program incorporating the digital peer support app Minchalle. All smartphone lecture programs were conducted in community venues in Sumida Ward (Tokyo) and Chiba City (Chiba Prefecture), which are typical venues for older adult smartphone lectures. The app, developed by A10 Lab Inc. and launched in 2015, was provided free of charge. Illustrations of the app are shown in Figure S1 in Multimedia Appendix 1. A structured description of the intervention is also provided according to the Template for Intervention Description and Replication (TIDieR) checklist (see Table S2 in Multimedia Appendix 1).

Minchalle integrates several behavior change techniques (BCTs) as defined by the BCT Taxonomy v1 [22]. The core element of Minchalle was the peer support component, in which groups of 3-5 members supported one another to achieve their goals by posting at least one photo and one message daily in the group chat (social support: general [BCT 3.1]). In addition, groups of 3-5 members selected daily step goals of 2500, 5000, or 10,000 steps (goal setting [BCT 1.1] and action planning [BCT 1.4]). Groups also selected a posting interval (every 4, 8, or 15 days). Members who failed to post within the agreed interval were removed (with the option to rejoin), reinforcing commitment (BCT 1.9: commitment). Step counts were automatically tracked and displayed in the chat, enabling self-monitoring of behavior (BCT 2.3) and promoting feedback and social comparison (BCT 6.2, 6.3) through the visibility of group achievements. This design was grounded in SCT, where peer encouragement, observational learning, and reinforcement are expected to enhance self-efficacy and sustain physical activity.

Two weeks after the initial baseline session, participants in the intervention arm attended 2 additional weekly lectures led by A10 Lab Inc, the developer of Minchalle. In these lectures, participants were instructed on how to install the app and configure initial settings, including selecting group step goals and defining posting intervals, after which members who failed to post within the interval would be removed. They also learned how to take photos, type messages, and post them in the group chat, supported by manuals provided during the sessions.

The intervention was delivered primarily at the cluster level (peer support groups of 3-5 participants), but individual participants used the app within their assigned peer support groups.

Control

Participants in the control arm also attended 2 smartphone lectures, akin to typical sessions for older adults in Japan. These lectures covered basic smartphone functions, including email, web browsing, use of the messenger app LINE, and official local government apps or accounts. As in the initial lecture, they were encouraged to continue using Google Fit or Apple Health for step tracking. The final assessment for the control arm was conducted 3 months later, simultaneously with the intervention arm but in different locations.

Outcomes

The primary outcome was the change in the weekly average of daily step counts from baseline to Week 12 (3 months after installation of the peer support app). After the baseline assessment, participants were asked to carry their smartphones at all times to capture daily steps. During the 2-week baseline period leading up to the second lecture, baseline step counts were measured. To minimize mismeasurement, any day with fewer than 1000 steps was treated as missing [23]. No changes to the prespecified primary outcome were made after trial commencement. All primary outcome data pertained to the individual participant level, although randomization was conducted at the cluster level.

Secondary outcomes included weekly step counts at other intervention weeks (Weeks 1 - 11) and monthly averages (Months 1 - 3), total physical activity (metabolic equivalent of task [MET] assessed by the International Physical Activity Questionnaire [IPAQ] short form) [24], self-reported walking time (≥30 minutes per day), daily smartphone use, and the number of smartphone use purposes at follow-up. The IPAQ short form was administered at baseline and follow-up to assess frequency and duration of vigorous, moderate, and walking

XSL•FO
RenderX

activities (≥10 min bouts). Walking time was dichotomized as <30 or ≥30 minutes per day. Smartphone use was dichotomized as daily (almost daily) versus not daily. Smartphone use purposes were assessed from a checklist and summed to yield a total score from information search and browsing, work-related tasks, communication (eg, email, LINE, or video calls), taking photos or videos, using maps or traffic information, online banking, shopping, social media (eg, Facebook [Meta], Twitter [subsequently rebranded X; X Corp]), health apps (eg, step counters), entertainment (eg, video streaming, e-books, or music), gaming, and other uses.

Statistical Analysis

The sample size was calculated assuming a 1-tailed test based on preliminary usage data with the same app; however, all subsequent analyses were conducted using 2-tailed tests at the 0.05 significance level. From pilot data provided by the app developer, we anticipated a medium effect size (Cohen d≈0.52) for changes in step counts. With α=.05, an intraclass correlation coefficient of 0.05 to account for clustering (mean cluster size of 4), and an expected attrition rate of 30%, we estimated that a total of 200 participants (100 per arm) would provide 80% power to detect a significant effect.

Baseline characteristics were summarized by treatment arm. Continuous variables were reported as mean (SD) and compared via 2-tailed 2‐sample t tests, while categorical variables were shown as frequencies (percentages) and analyzed via chi‐square or Fisher exact tests, using the *CreateTableOne* package in R (R Core Team).

Days with <1000 steps and other missing data were imputed; based on evidence, these often reflect incomplete capture [23,25]. Multiple imputation was performed using the package in R, incorporating baseline covariates and primary and secondary outcomes. Ten imputed datasets were generated, and estimates were combined using Rubin rules.

Separate imputation models were created for step count data (n=117) and questionnaire-based outcomes (n=124).

The primary analysis estimated the between-arm difference in the preregistered primary outcome—the change in weekly average daily step count from baseline to Week 12—using an unadjusted mixed model for repeated measures (MMRM). This model included treatment arm, week (categorical: baseline and Weeks 1‐12), and their interaction, with an unstructured covariance matrix at the participant level and cluster-robust standard errors to account for clustering of participants within groups.

Several sensitivity analyses were conducted to examine the robustness of the findings. These included (1) adjusted MMRM, (2) unadjusted, and (3) adjusted linear mixed-effects models with random intercepts at the participant and group levels, (4) unadjusted and (5) adjusted ANCOVA models with the change from baseline to Week 12 as the dependent variable, and (6) unadjusted MMRM using unimputed data, excluding days with missing or <1000 steps. For adjusted models, baseline covariates were selected a priori based on theoretical relevance and prior literature. Operating system (Android or iOS), study field (Sumida or Chiba), gender (men or women), education (high school or less vs college or more), marital status (married or unmarried), baseline smartphone use (daily vs not daily), and walking ≥30 minutes per day were treated as categorical variables. Age, self-rated health (0‐10 scale), and total MET-minutes per week were treated as continuous variables and modeled linearly.

Analyses of weekly step counts for Weeks 1‐11 and monthly averages (Months 1‐3) were conducted as secondary outcomes using the same unadjusted MMRM framework as the primary analysis. These analyses were not adjusted for multiple comparisons and were not part of the primary estimand, but are included to aid interpretation of temporal patterns.

Secondary outcomes obtained from the Week-12 follow-up questionnaire were analyzed using unadjusted models to maintain consistency with the primary analytic approach. Total physical activity (MET-minutes per week) and the number of smartphone use purposes were examined using ANCOVA, adjusting only for the baseline value of each outcome. Walking ≥30 minutes per day and daily smartphone use were analyzed using logistic regression with the same minimal adjustment.

Because a baseline age imbalance between arms was observed, we conducted an exploratory subgroup analysis stratified at the median age of 80 years. For each age stratum, the primary outcome was re-estimated using the unadjusted MMRM model described above. An interaction model including treatment × week × age (binary: ≥80 vs <80 years) was further examined to assess whether age modified the intervention effect.

Ethical Considerations

This study received ethical approval from the Chiba University Ethical Committee (protocol number: M10626) and was registered with the University Hospital Medical Information Network (UMIN000051904). Written informed consent was obtained from all participants before randomization. All procedures adhered to the principles of the Declaration of Helsinki. Participant privacy and confidentiality were protected by deidentifying all data prior to analysis. Data were stored securely on password-protected servers accessible only to authorized study personnel. Only aggregated results are reported in this manuscript. Participants did not receive monetary compensation; however, they were allowed to use the smartphone app provided as part of the intervention, and no additional incentives were offered.

Results

Overview

Recruitment ended when the smartphone lecture series provided by the municipality and senior clubs—our sole recruitment venue—was completed, making additional enrollment infeasible. At that point, 177 participants had been enrolled. The analysis was conducted at this stage because recruitment had necessarily concluded, although the final sample size did not reach the originally planned statistical power.

At the analysis, 40 groups comprising 156 participants were allocated either to an intervention arm (n=80, 20 groups) or a control arm (n=76, 20 groups; see Figure 2). Of the 156 enrolled

XSL•FO
RenderX

participants, 124 (79.5%) completed the 3-month follow-up evaluation (64 in the intervention arm and 60 in the control arm), with step-count data successfully linked for 117 participants.

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flow diagram for study clusters and participants.

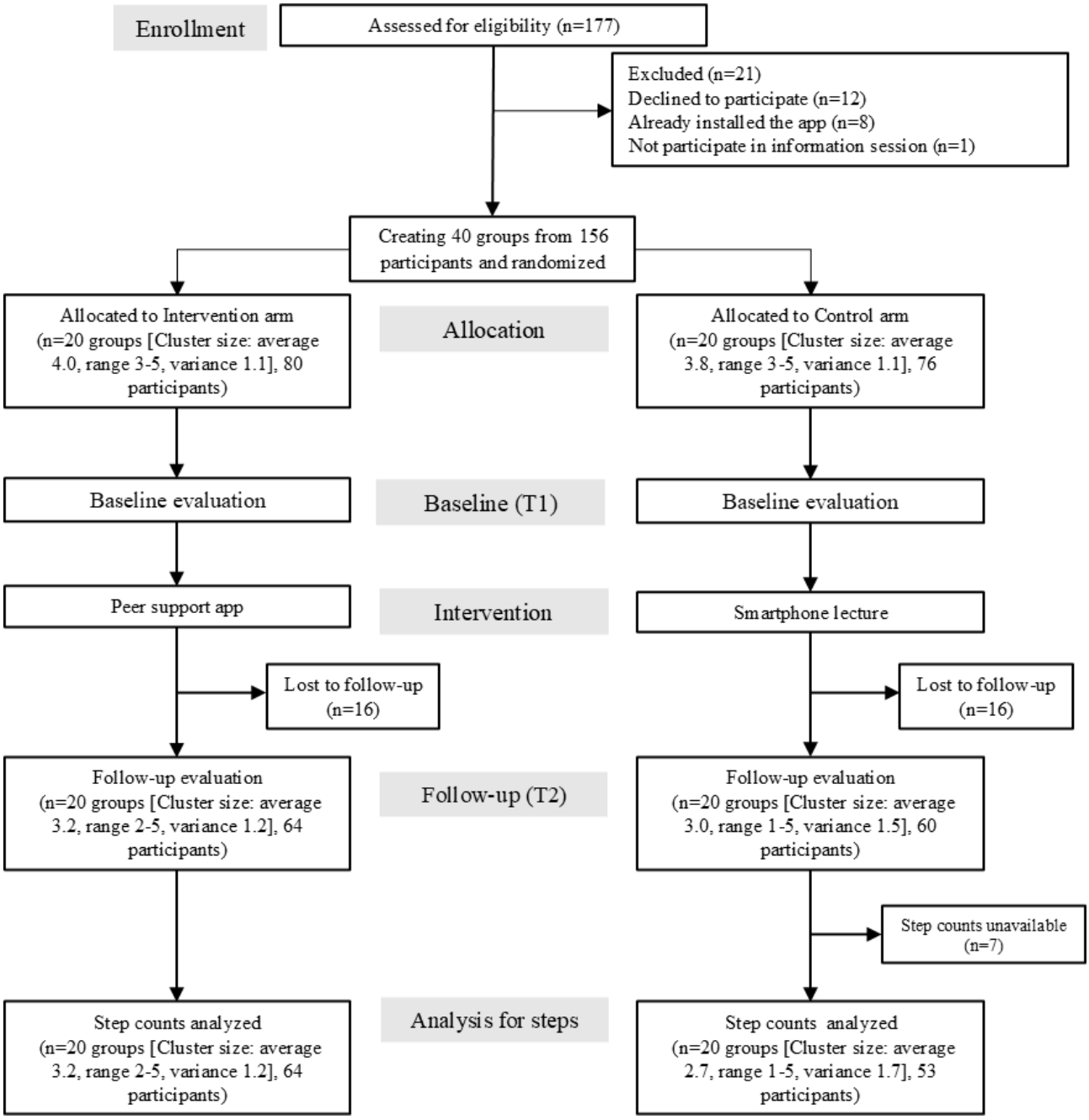

Missing step data ranged from 5.1% to 29.1%. As shown in Table 1, baseline characteristics were generally well balanced between arms, except for a significant difference in mean age (P=.009). Mean (SD) baseline daily steps were 3951 (1686) for the control arm and 4583 (1976) for the intervention arm.

XSL•FO
RenderX

Table . Characteristics of study participants.

Variables	Control (n=53)	Intervention (n=64)
Age (year), mean (SD)	81.2 (4.7)	79.0 (4.5)
Gender (Men), n (%)	17 (32.1)	18 (28.1)
OS (iOS), n (%)	9 (17.0)	18 (28.1)
Daily use of smartphone, n (%)	41 (77.4)	55 (85.9)
Education (college or more), n (%)	25 (47.2)	20 (31.7)
Married, n (%)	26 (49.1)	31 (48.4)
Self-rated health, mean (SD)	6.8 (2.0)	6.8 (1.9)
Total MET[a], mean (SD)	3274 (1850)	3372 (2373)
Walking more than 30 minutes, n (%)	37 (69.8)	43 (67.2)
Baseline weekly average steps, mean (SD)	3951 (1686)	4583 (1976)

[a]MET: metabolic equivalent of task.

Primary Outcome

The unadjusted differences in mean daily steps by week and study arm are illustrated in Figure 3. Across most time points, the intervention arm recorded higher step counts than the control.

Figure 3. Unadjusted differences in mean daily steps by week and study arm.

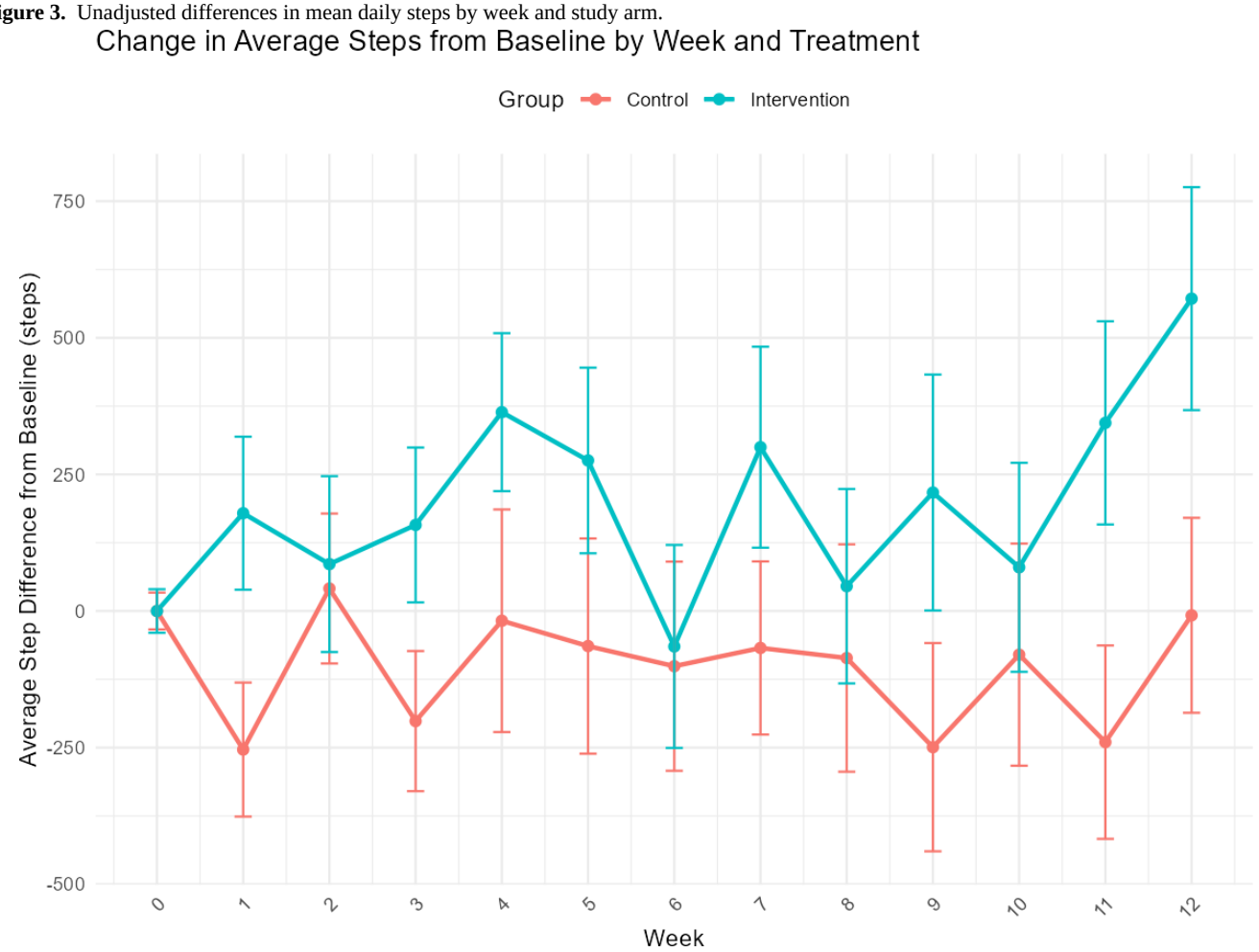

The primary outcome was the change in average daily steps at week 12 compared with baseline. In the unadjusted MMRM model, the intervention arm recorded significantly higher steps than the control arm at Week 12 (difference=579 steps, 95% CI 36-1123;*P*=.04) (Table 2). The estimated cluster-level variance for the primary outcome was essentially zero (intraclass correlation coefficient ≈ 0.00), indicating minimal between-cluster variability after adjustment.

XSL•FO
RenderX

Table . Estimated between-arm differences in step counts by week (intervention minus control), unadjusted mixed-effects model for repeated measures.

Variables	Estimate[a], β (95% CI)	*P* value
Intervention × Week 1[b]	433 (74 to 791)	.02
Intervention × Week 2	45 (−385 to 474)	.84
Intervention × Week 3	359 (−26 to 744)	.07
Intervention × Week 4	382 (−180 to 943)	.18
Intervention × Week 5	340 (−273 to 953)	.28
Intervention × Week 6	36 (−564 to 637)	.91
Intervention × Week 7	367 (−173 to 908)	.18
Intervention × Week 8	131 (−420 to 683)	.64
Intervention × Week 9	466 (−222 to 1154)	.18
Intervention × Week 10	160 (−434 to 754)	.60
Intervention × Week 11	584 (17 to 1151)	.04
Intervention × Week 12	579 (36 to 1123)	.04

[a]Estimates represent unadjusted between-arm differences at each week, derived from arm×week interaction terms.

[b]Intervention×week=product terms in the model.

The overall pattern remained consistent with the main analysis in sensitivity analysis: adjusted MMRM model (difference=539, 95% CI 92-985, *P*=.02); unadjusted mixed effects model (difference=589, 95% CI 110-1069; *P*=.02); adjusted mixed effects model (difference=526, 95% CI 65-987; *P*=.03); unadjusted ANCOVA model (difference=589, 95% CI 76-1101; *P*=.02); adjusted ANCOVA model (difference=489, 95% CI 33-946; *P*=.04); unadjusted MMRM model with complete case analysis (difference=1295, 95% CI 107-2483; *P*=.03).

Secondary Outcomes

Exploratory analyses of weekly step counts (Weeks 1 - 11) showed that the intervention arm had higher steps at Week 1 and Week 11 (Week 1: difference=433, 95% CI 74-791; *P*=.02; Week 11: difference=584, 95% CI 17-1151; *P*=.04, respectively) (Table 2). In a further exploratory analysis aggregating daily steps by month, the intervention arm showed greater, but not significant, differences at each month (Table S3 in Multimedia Appendix 1).

No significant arm differences were observed for physical activity measured by total METs or for self-reported walking time (≥30 minutes per day). The unadjusted difference in total METs was 646 MET-minutes per week (95% CI −12 to 1303; *P*=.054), and the odds ratio (OR) for walking ≥30 minutes per day was 1.56 (95% CI 0.63-3.90; *P*=.33; Table S4 in Multimedia Appendix 1). However, there was a suggestive trend toward an increase in daily smartphone use (OR 4.10, 95% CI 1.15-14.6; *P*=.03) and the number of smartphone use purposes (mean difference 0.58, 95% CI 0.12-1.05; *P*=.01) among participants in the intervention arm.

Exploratory Post-Hoc Analyses

In exploratory subgroup analyses stratified by the median age of 80 years (younger: n=57; older: n=58), baseline age was well balanced within subgroups (younger: mean 76.2 vs 76.3 years; older: mean 84.2 vs 83.1 years). Intervention effects varied by age. Among older participants (≥80 years; n=58), significant treatment effects were observed, with between-arm differences exceeding 1000 steps at Week 12. In contrast, younger participants (<80 years; n=57) showed inconsistent and occasionally negative effects, with no sustained significant benefits. Consistent with these findings, the treatment × week × age interaction model indicated that intervention effects became stronger among participants aged 80 years or older in multiple weeks. Detailed estimates from the subgroup and interaction analyses are provided in Table S5 in Multimedia Appendix 1.

Discussion

Short Summary

This cluster-randomized controlled trial demonstrates that a smartphone lecture program, combined with a mobile app–based peer support component, may produce modest but meaningful increases in step counts among older adults in Japan compared to a conventional smartphone lecture program. The intervention arm recorded a statistically significant increase in average daily steps at Week 12 (the preregistered primary end point), with supportive differences observed at Weeks 1 and 11. All other weekly comparisons and monthly averages were exploratory and should be interpreted with caution, as no adjustment for multiple testing was applied. In addition, smartphone use and purposes were greater in the intervention arm. These results suggest that integrating peer support features into smartphone-based health initiatives could be a promising strategy for sustaining engagement and encouraging physical activity and smartphone use over time.

Principal Findings

The peer support app appeared to stimulate progressive increases in step counts, particularly noticeable during the latter weeks. An important question is why intervention effects became most apparent in the latter part of the program (Weeks 11 - 12). Novelty effects typically lead to high engagement at the outset

XSL•FO
RenderX

of an intervention but tend to diminish over time, making it unlikely that novelty alone explains the late increase in steps. A more plausible explanation is that repeated goal setting, self-monitoring, and peer reinforcement gradually strengthened self-efficacy and supported the formation of walking habits, consistent with SCT. In this framework, the app's design may have enhanced self-efficacy (through repeated goal attainment and encouragement), promoted observational learning (by making peers' behaviors visible), and provided reinforcement (through accountability and supportive feedback).

In addition, seasonal and contextual factors may have contributed. The study was conducted in Sumida Ward, Tokyo, from September to December 2023. December includes cultural and social events that may have motivated participants to walk more and share photos with peers in the app. Also, in Chiba City, these were held from June to September 2024. Older adults might avoid walking outside in hot weather in July and August, while cooler weather in September may have encouraged outdoor activity. These contextual influences, alongside habit formation, may help explain why intervention effects were observed more strongly toward the end of the study period.

Comparison to Prior Work

Our findings align with and extend prior evidence that interpersonal and peer-support components are particularly important for promoting and sustaining physical activity among older adults. McMahon et al [12] conducted a factorial randomized trial comparing intrapersonal and interpersonal behavior change strategies among community-dwelling adults aged ≥70 years and demonstrated that interventions including interpersonal components produced greater improvements in objectively measured physical activity, whereas intrapersonal components alone were less effective. Our results are consistent with this framework, while extending it to a pragmatic community education setting by embedding peer support within routinely delivered smartphone lectures rather than a research-intensive intervention.

Importantly, Tabira et al [17] evaluated the same digital peer-support smartphone app (Minchalle) used in our study in a nonrandomized controlled design among adults aged ≥65 years. They reported high feasibility, retention, and increases in physical activity, supporting the app's acceptability and potential effectiveness in older populations. However, their study focused on feasibility and short-term behavioral change under a nonrandomized design, leaving uncertainty regarding causal effects. By contrast, our randomized controlled trial using the same app extends this prior work by strengthening causal inference and demonstrating effectiveness when Minchalle is implemented as part of existing smartphone lecture programs that are already widely disseminated across Japan.

In addition, Patel et al [18] showed in a large randomized trial that socially structured incentives embedded in behaviorally designed gamification can increase step counts, although the magnitude and persistence of effects varied by incentive type. While conducted primarily among working-age adults in a remote, workplace-adjacent context, their findings reinforce the broader principle that social and peer-based design features meaningfully influence physical activity behavior. Our study complements this literature by demonstrating that similar social mechanisms can be effectively operationalized in older adults using commercially available apps without requiring specialized devices or intensive program infrastructure.

Exploratory Post-Hoc Analyses

This study suggests that older participants (≥80 years) may derive greater benefit from app-based peer support, a group often most vulnerable to the digital divide [26]. By facilitating engagement with smartphones through peer support, our approach may help promote physical activity while also addressing inequities in digital participation. These findings are exploratory and not prespecified, and given the limited sample size, should be interpreted with caution.

Secondary Findings

Despite the significant gains in step counts, we observed no significant arm differences in total METs or self-reported walking time (≥30 minutes per day). These discrepancies may stem from the inherently coarse nature of subjective physical activity measures, which can miss nuances that more objective step-tracking captures [12,27]. Indeed, correlations between self-report and objectively measured physical activity in older adults are positive, yet often moderate at best [28]. Future studies integrating more comprehensive, sensor-based activity assessments will help clarify whether peer support can robustly influence other domains of physical activity beyond daily step counts.

There was also a suggestive trend toward increased daily smartphone use in the intervention arm, implying that digital peer support may facilitate broader technology adoption. This is consistent with the primary aim of smartphone lectures for older adults and suggests that digital peer support apps can effectively enhance participants' confidence and willingness to explore additional smartphone features.

Limitations

Several limitations merit consideration. First, the final sample size was smaller than planned, which reduced statistical power, particularly for secondary outcomes. Attrition (≈20%) further reduced the analytic sample, mainly due to nonattendance at the 3-month follow-up lecture, and step-count data from 7 participants were lost because of a technical linkage error. Second, missing data and imputation procedures, while consistent with prior interventions, could introduce bias if data were not missing at random. Third, digital peer support may not uniformly appeal to all participants, potentially limiting generalizability. Future studies should examine heterogeneous responses to peer support for physical activity and smartphone adoption. Fourth, although we reported step count differences at multiple weekly and monthly time points, only the between-arm difference at Week 12 was prespecified as the confirmatory primary end point. All other weekly and monthly comparisons were exploratory and not adjusted for multiplicity; therefore, these findings should be interpreted with caution. Fifth, no formal arm sequential testing procedure (eg, alpha-spending function) was applied to adjust for multiple looks at the data. Recruitment was concluded after 177 participants had been enrolled, based on the completion of the

XSL•FO
RenderX

smartphone lectures organized by the local municipal government and senior clubs. The absence of a formal stopping rule represents a limitation with respect to strict error rate control. Sixth, SCT constructs were not directly measured in this trial, and the hypothesized mechanisms (eg, self-efficacy, observational learning, and reinforcement) remain speculative. Seventh, no process measures such as outcome expectancies or social support perceptions were collected, and detailed app usage metrics (eg, posting volume, response rates, and network analyses) were not available, limiting the ability to fully evaluate the peer support component. Eighth, baseline age differed modestly between arms (79.0 years in intervention vs 81.2 years in control). However, exploratory analyses suggested that older participants benefited more from the intervention. Therefore, this imbalance would likely have biased estimates toward underestimation of the true effect rather than exaggeration. These age-stratified analyses were exploratory, not prespecified, and should be interpreted with caution given the limited sample size. Ninth, although the Minchalle app interface and features were identical on Android and iOS, potential platform-specific differences in step data capture (eg, Google Fit vs Apple Health) cannot be excluded. Due to the small number of iOS users (n=27), subgroup analyses by operating system were not feasible, and future studies should address this more explicitly. Tenth, individual-level attrition resulted in smaller cluster sizes at follow-up, particularly in the control arm, where some clusters ultimately had only one participant contributing data. Participants who did not attend the follow-up session were unable to provide questionnaire or step-count data, and in the intervention arm, we were also unable to monitor their continued use of the peer-support app. As a result, it remains uncertain whether any intervention clusters functionally shrank to very small active groups during the trial period. This uncertainty may influence the interpretation of peer-related effects, as reduced group size could weaken the social facilitation or mutual-accountability mechanisms central to the intervention. Finally, the external validity of our findings should be interpreted with caution. This trial was conducted in 2 urban areas of Japan, where walkable environments and access to digital infrastructure are relatively well developed. The results may not directly translate to rural areas with limited walkability or to contexts with lower smartphone and internet penetration. In addition, cultural differences in peer support and technology use may influence the effectiveness of similar interventions in international settings. These contextual factors should be carefully considered when applying our findings to broader populations.

Conclusions

In conclusion, our results suggest that a mobile peer support intervention combined with a smartphone lecture can yield significant increases in step counts among older Japanese adults. While secondary measures of total physical activity and self-reported walking time did not show a comparable effect, the observed gains in daily steps underscore the potential of app-based peer support for promoting sustained walking behavior. Moreover, a marginally significant increase in daily smartphone use and a significant increase in the number of smartphone use purposes highlight the possibility that a digital peer support program can enhance technology adoption among older adults. Future research should explore longer-term maintenance and identify which segments of the older population stand to benefit the most in order to optimize digital health interventions for this rapidly growing demographic.

Acknowledgments

We would like to thank Sumida-Ward and senior clubs in Chiba city for giving us opportunities to recruit participants. We would like to thank Active Senior Outreach Nonprofit Organization for holding smartphone lectures. We would like to thank the Institute for Assistance of Academic and Education for carrying out the randomization.

During the preparation of this work, the authors used ChatGPT in order to improve language and readability. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

This study was funded by A10 Lab Inc. AN and KK have received research funding from A10 Lab Inc. to evaluate the Minchalle app and provide feedback on service improvement based on study findings. The sponsor's role was limited to introducing municipalities for recruitment and providing raw step count data. The sponsor had no involvement in the study design, data analysis, data interpretation, manuscript writing, or the decision to submit the article for publication. To safeguard independence, all statistical analyses were conducted solely by academic investigators (NA and AN), with independent oversight by YK.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: AN, TU, NA
Methodology: AN, NA, TU
Data Curation: TU, NA, GI
Formal Analysis: NA, AN
Validation: YK

XSL•FO
RenderX

Supervision: AN, KK
Writing – Original Draft Preparation: AN
Writing – Review & Editing: AN, NA, TU, GI, YK, KK
Funding Acquisition: AN, KK
All authors had full access to the dataset, contributed to the interpretation of the data, approved the final version of the manuscript, and agree to be accountable for all aspects of the work.

Conflicts of Interest

AN and KK have received research funding from A10 Lab Inc. The sponsor had no involvement in the design of the study; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

Multimedia Appendix 1
Supplementary figure and additional statistical analysis tables.
[DOCX File, 4631 KB - jmir_v28i1e75708_app1.docx]

Checklist 1
CONSORT-eHEALTH checklist (V 1.6.1)
[PDF File, 57248 KB - jmir_v28i1e75708_app2.pdf]

References

1. Lee IM, Shiroma EJ, Lobelo F, et al. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. Lancet 2012 Jul 21;380(9838):219-229. [doi: 10.1016/S0140-6736(12)61031-9] [Medline: 22818936]
2. Bull FC, Al-Ansari SS, Biddle S, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. Br J Sports Med 2020 Dec;54(24):1451-1462. [doi: 10.1136/bjsports-2020-102955] [Medline: 33239350]
3. Healthy Japan 21. Ministry of Health Law. 2024 Mar 26. URL: https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/kenkou/kenkounippon21_00006.html [accessed 2025-03-26]
4. Physical activity and exercise guide for health promotion 2023. Ministry of Health Law. 2023. URL: https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/kenkou/undou/index.html [accessed 2025-03-26]
5. Murakami H, Tripette J, Kawakami R, Miyachi M. “Add 10 min for your health”: the new Japanese recommendation for physical activity based on dose-response analysis. J Am Coll Cardiol 2015 Mar 24;65(11):1153-1154. [doi: 10.1016/j.jacc.2014.10.080] [Medline: 25790890]
6. García-Hermoso A, Ramirez-Vélez R, Sáez de Asteasu ML, et al. Safety and effectiveness of long-term exercise interventions in older adults: a systematic review and meta-analysis of randomized controlled trials. Sports Med 2020 Jun;50(6):1095-1106. [doi: 10.1007/s40279-020-01259-y] [Medline: 32020543]
7. Piercy KL, Troiano RP, Ballard RM, et al. The physical activity guidelines for Americans. JAMA 2018 Nov 20;320(19):2020-2028. [doi: 10.1001/jama.2018.14854] [Medline: 30418471]
8. French DP, Olander EK, Chisholm A, Mc Sharry J. Which behaviour change techniques are most effective at increasing older adults’ self-efficacy and physical activity behaviour? A systematic review. Ann Behav Med 2014 Oct;48(2):225-234. [doi: 10.1007/s12160-014-9593-z] [Medline: 24648017]
9. Zubala A, MacGillivray S, Frost H, et al. Promotion of physical activity interventions for community dwelling older adults: a systematic review of reviews. PLoS ONE 2017;12(7):e0180902. [doi: 10.1371/journal.pone.0180902] [Medline: 28700754]
10. Dennis CL. Peer support within a health care context: a concept analysis. Int J Nurs Stud 2003 Mar;40(3):321-332. [doi: 10.1016/s0020-7489(02)00092-5] [Medline: 12605954]
11. Ginis KAM, Nigg CR, Smith AL. Peer-delivered physical activity interventions: an overlooked opportunity for physical activity promotion. Transl Behav Med 2013 Dec;3(4):434-443. [doi: 10.1007/s13142-013-0215-2] [Medline: 24294332]
12. McMahon SK, Lewis BA, Guan W, et al. Effect of intrapersonal and interpersonal behavior change strategies on physical activity among older adults: a randomized clinical trial. JAMA Netw Open 2024 Feb 5;7(2):e240298. [doi: 10.1001/jamanetworkopen.2024.0298] [Medline: 38421648]
13. Bandura A. Social cognitive theory: an agentic perspective. Annu Rev Psychol 2001;52:1-26. [doi: 10.1146/annurev.psych.52.1.1] [Medline: 11148297]
14. Laranjo L, Ding D, Heleno B, et al. Do smartphone applications and activity trackers increase physical activity in adults? Systematic review, meta-analysis and metaregression. Br J Sports Med 2021 Apr;55(8):422-432. [doi: 10.1136/bjsports-2020-102892] [Medline: 33355160]
15. Zhang J, Jemmott Iii JB. Mobile app-based small-group physical activity intervention for young African American women: a pilot randomized controlled trial. Prev Sci 2019 Aug;20(6):863-872. [doi: 10.1007/s11121-019-01006-4] [Medline: 30788692]

XSL•FO
RenderX

16. Xu L, Shi H, Shen M, et al. The effects of mHealth-based gamification interventions on participation in physical activity: systematic review. JMIR Mhealth Uhealth 2022 Feb 3;10(2):e27794. [doi: 10.2196/27794] [Medline: 35113034]
17. Tabira K, Oguma Y, Yoshihara S, et al. Digital peer-supported app intervention to promote physical activity among community-dwelling older adults: nonrandomized controlled trial. JMIR Aging 2024 May 30;7:e56184. [doi: 10.2196/56184] [Medline: 38814686]
18. Patel MS, Small DS, Harrison JD, et al. Effectiveness of behaviorally designed gamification interventions with social incentives for increasing physical activity among overweight and obese adults across the United States: the STEP UP randomized clinical trial. JAMA Intern Med 2019 Dec 1;179(12):1624-1632. [doi: 10.1001/jamainternmed.2019.3505] [Medline: 31498375]
19. Deursen AJ, Helsper EJ. The third-level digital divide: who benefits most from being online? Communication and information technologies annual: Emerald Group Publishing Limited; 2015:29-52. [doi: 10.1108/S2050-206020150000010002]
20. Campbell MK, Piaggio G, Elbourne DR, Altman DG, CONSORT Group. Consort 2010 statement: extension to cluster randomised trials. BMJ 2012 Sep 4;345:e5661. [doi: 10.1136/bmj.e5661] [Medline: 22951546]
21. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
22. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
23. Kang M, Rowe DA, Barreira TV, Robinson TS, Mahar MT. Individual information-centered approach for handling physical activity missing data. Res Q Exerc Sport 2009 Jun;80(2):131-137. [doi: 10.1080/02701367.2009.10599546] [Medline: 19650377]
24. Lee PH, Macfarlane DJ, Lam TH, Stewart SM. Validity of the International Physical Activity Questionnaire Short Form (IPAQ-SF): a systematic review. Int J Behav Nutr Phys Act 2011 Oct 21;8(115):115. [doi: 10.1186/1479-5868-8-115] [Medline: 22018588]
25. Bassett DR, Wyatt HR, Thompson H, Peters JC, Hill JO. Pedometer-measured physical activity and health behaviors in U.S. adults. Med Sci Sports Exerc 2010 Oct;42(10):1819-1825. [doi: 10.1249/MSS.0b013e3181dc2e54] [Medline: 20305579]
26. OECD. OECD digital economy outlook 2024 (volume 1): digital technology diffusion and data. : OECD Publishing; 2024 URL: https://www.oecd.org/en/publications/oecd-digital-economy-outlook-2024-volume-1_a1689dc5-en/ [accessed 2026-02-28]
27. Grande GD, Oliveira CB, Morelhão PK, et al. Interventions promoting physical activity among older adults: a systematic review and meta-analysis. Gerontologist 2020 Nov 23;60(8):583-599. [doi: 10.1093/geront/gnz167] [Medline: 31868213]
28. Skender S, Ose J, Chang-Claude J, et al. Accelerometry and physical activity questionnaires - a systematic review. BMC Public Health 2016 Jun 16;16:515. [doi: 10.1186/s12889-016-3172-0] [Medline: 27306667]

Abbreviations

BCT: behavior change technique
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
IPAQ: International Physical Activity Questionnaire
MET: metabolic equivalent of task
MHLW: Ministry of Health, Labor, and Welfare
MMRM: mixed model for repeated measures
OR: odds ratio
SCT: social cognitive theory
TIDieR: Template for Intervention Description and Replication

Edited by A Stone, N Cahill; submitted 09.Apr.2025; peer-reviewed by M Dolezel, P Liv; revised version received 05.Feb.2026; accepted 10.Feb.2026; published 10.Mar.2026.

Please cite as:
Nakagomi A, Abe N, Ueno T, Izuka G, Kawasaki Y, Kondo K
Digital Peer Support to Increase Walking Among Older Adults: Cluster Randomized Trial
J Med Internet Res 2026;28:e75708
URL: https://www.jmir.org/2026/1/e75708
doi:10.2196/75708

XSL•FO
RenderX

© Atsushi Nakagomi, Noriyuki Abe, Takayuki Ueno, Gemmei Izuka, Yohei Kawasaki, Katsunori Kondo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Enhancing Adherence to Home-Based Expiratory Muscle Strength Training in Parkinson Disease: Randomized Controlled Trial of an mHealth Intervention

Martin Srp[1*], PhD; Martina Hoskovcova[1*], PhD; Rebeka Lagnerova[1], BSc; Katerina Dvorakova[1], MSc; Radim Kliment[2], PhD; Jan Muzik[2], PhD; Radim Krupicka[3], PhD; Ota Gal[1], PhD; Evzen Ruzicka[1], MD, DrSc

[1]Department of Neurology and Center of Clinical Neuroscience, General University Hospital and First Faculty of Medicine, Charles University, Katerinska 30, Prague, Czech Republic

[2]Department of Information and Communication Technologies in Medicine, Faculty of Biomedical Engineering, Czech Technical University in Prague, Prague, Czech Republic

[3]Department of Biomedical Informatics, Faculty of Biomedical Engineering, Czech Technical University in Prague, Kladno, Czech Republic

[*]these authors contributed equally

Corresponding Author:
Evzen Ruzicka, MD, DrSc
Department of Neurology and Center of Clinical Neuroscience, General University Hospital and First Faculty of Medicine, Charles University, Katerinska 30, Prague, Czech Republic

Abstract

Background: Aspiration pneumonia is a leading cause of death in Parkinson disease (PD). Expiratory muscle strength training (EMST) is a promising intervention for respiratory and swallowing dysfunction. However, long-term EMST adherence is frequently poor in PD.

Objective: This study aims to determine whether mobile health (mHealth)–assisted EMST with the *SpiroGym* app (Czech Technical University) improves long-term adherence and physiological outcomes versus conventional EMST among participants at risk for nonadherence.

Methods: In this single-center, parallel, phase 2 randomized controlled trial, 75 individuals with PD were randomized 1:1 to conventional EMST (control; n=38) or the same protocol enhanced with the *SpiroGym* app (experimental; n=37), using a simple computer-generated randomization sequence. The *SpiroGym* is an mHealth app that provides real-time performance monitoring, direct visual feedback, and longitudinal progress tracking. All participants completed 8 weeks of semisupervised intensive EMST with biweekly in-person reassessments, followed by 16 weeks of unsupervised maintenance training. The primary outcome was adherence during weeks 8 to 24 among participants at risk for nonadherence, defined a priori at week 8 as Self-Efficacy for Home Exercise Program Scale (SEHEPS) less than 59. Because risk status was determined at week 8 and all participants subsequently entered the unsupervised phase, individuals not classified as at-risk were not excluded. Their data from week 8 onward were reported alongside the at-risk group. Secondary outcomes were changes in maximum expiratory pressure and SEHEPS.

Results: No study-related adverse events occurred. Groups were well matched at baseline (control vs experimental: mean disease duration 7.0 (SD 5.7) vs 7.3 SD 4.7) y; mean Hoehn-Yahr 1.97 (SD 0.6) vs 2.0 (SD 0.5)). The mixed-effects model showed no significant 3-way interaction (group×interval×SEHEPS risk; P=.14). At week 24, the at-risk category for the nonadherence cohort comprised 34 participants (control, n=17; experimental, n=17). In this at-risk cohort, the experimental group demonstrated a smaller decline in adherence during weeks 8 to 24 than controls (β=496.9, 95% CI 130.7 - 863.3; P=.008), completing 1073 (95% CI 643 - 1502) expiratory maneuvers versus 525 (95% CI 358 - 692). Maximum expiratory pressure increased in both groups from weeks 0 to 24, with larger gains in the experimental group (+43.1, 95% CI 32.4 - 53.8 cmH_2O) than in controls (+22.8, 95% CI 13.8 - 31.8 cmH_2O; P=.006; Cohen d=0.74). SEHEPS improved after intensive training in both groups, but only the experimental group exceeded the 12-point minimal detectable change at the 95% confidence limit.

Conclusions: This is the first randomized controlled trial to integrate mHealth with EMST. Unlike prior studies in the EMST field, we focused on sustaining long-term exercise adherence. *SpiroGym*-assisted EMST resulted in higher long-term adherence and greater gains in expiratory muscle strength than conventional EMST. In real-world PD care, assessing self-efficacy after the supervised EMST phase may help identify individuals who would benefit from digital support, making mHealth-assisted EMST a practical approach for maintaining exercise adherence.

Trial Registration: ClinicalTrials.gov NCT05728099; https://clinicaltrials.gov/study/NCT05728099

XSL•FO
RenderX

(J Med Internet Res 2026;28:e78022) doi:10.2196/78022

KEYWORDS

mobile health; randomized controlled trial; Parkinson disease; expiratory muscle strength training; exercise adherence

Introduction

Neurodegenerative disorders are now the leading cause of disability worldwide [1]. Among them, Parkinson disease (PD) is the fastest-growing condition [2]. PD is a complex, progressive neurodegenerative disorder, with symptoms primarily arising from the loss of dopaminergic neurons in the brain [3]. Aspiration pneumonia is the leading cause of death in PD [4-6]. Among nonpharmacological interventions, expiratory muscle strength training (EMST) has shown promise in improving airway protection [7,8]. However, long-term adherence to EMST, which is critical to sustaining therapeutic benefits [9], is often poor in people with chronic conditions such as PD [10]. This challenge is compounded by the home-based, unsupervised nature of EMST, limiting opportunities for professional feedback and monitoring [11]. Additionally, dopaminergic dysfunction may reduce motivation, interest in pleasurable activities, and confidence in one's ability to exercise, further potentiating nonadherence to EMST [12].

The widespread use of smartphones has enabled the development of health apps that track and manage symptoms, thereby strengthening self-care interventions for people with chronic illness [13-15]. Digital health technologies can expand access to home-based therapy and extend the benefits of in-person physiotherapy by providing real-time feedback, goal tracking, and performance summaries [14,16]. Feedback on past performance is particularly effective in strengthening exercise self-efficacy [17], which is defined as an individual's belief in their ability to successfully perform goal-directed behaviors and stands out as a consistent predictor of exercise adherence across diverse populations [18-20], including people with PD [21]. Digital health features may be especially relevant for EMST, where patients may have limited insight into whether maneuvers are performed with adequate effort and technique, and where reinforcement is minimal when training alone at home.

To address barriers to long-term EMST, we developed the *SpiroGym* app using mobile health (mHealth) technologies to provide real-time performance monitoring, direct visual feedback, and longitudinal progress tracking. A pilot study [22] demonstrated the feasibility of *SpiroGym*-assisted EMST in PD. Just 2 weeks of training significantly improved maximum expiratory pressure (MEP) [22], achieving results comparable to longer EMST protocols [23,24]. These rapid gains were likely facilitated by increased training effort driven by real-time feedback. Further validation came from a recent multicenter study showing high usability and acceptance of *SpiroGym* across diverse international PD cohorts [25]. A subsequent proof-of-concept study reported excellent adherence during a 16-week unsupervised home-based EMST program using *SpiroGym*, even among participants unaccustomed to strength training [26]. That study also reported significant improvements in the Self-Efficacy for Home Exercise Program Scale (SEHEPS) [26]. Greater confidence on this scale correlated with long-term EMST adherence, highlighting self-efficacy's critical role in sustained engagement [21,27,28]. While these findings underscore the potential of *SpiroGym*-assisted EMST, previous feasibility and proof-of-concept studies did not include an active-controlled randomized comparison against conventional EMST. Therefore, an active-controlled randomized controlled trial (RCT) is needed to determine whether mHealth-supported EMST improves adherence beyond conventional home-based EMST.

The primary objective of this RCT was to compare adherence during the unsupervised 16-week training period between conventional EMST and *SpiroGym*-assisted EMST in participants at risk for nonadherence. Secondary objectives were to evaluate changes in MEP and SEHEPS. Based on prior findings [22,26], we hypothesized that *SpiroGym*-assisted EMST would demonstrate superior results across these domains.

Methods

Study Design

We conducted a single-center, parallel, phase 2, randomized, active-controlled clinical trial that was retrospectively registered at ClinicalTrials.gov (NCT05728099) in 2023, after recruitment began in 2022 but before trial completion in 2025, and reported in accordance with the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) guidelines (Checklist 1) [29]. The institutional review board (IRB; the ethics committee of the General University Hospital in Prague) confirmed that the protocol and study design reported in this paper are consistent with the study protocol and design that were reviewed and approved by the ethics committee before the trial started. A side-by-side comparison of key protocol elements across the IRB submission, the ClinicalTrials.gov record, and this paper is provided in Multimedia Appendix 1. This trial was conducted at the Department of Neurology, First Faculty of Medicine, Charles University and General University Hospital in Prague, Czech Republic, between March 2022 and January 2025. Participants were randomly assigned, with allocation concealment, to either the control group following the conventional EMST protocol or to the experimental group, following the same protocol enhanced with the *SpiroGym* app.

Participants

Participants were recruited through convenience sampling from the Movement Disorders Center, Department of Neurology, First Faculty of Medicine, Charles University and General University Hospital in Prague, Czech Republic, between March 2022 and August 2024. The inclusion criteria were: (1) a diagnosis of PD confirmed by movement disorder specialist using the UK Brain Bank criteria [30], (2) a modified Hoehn and Yahr stage of I-IV in ON medication state, (3) an age of 40 to 80 years, and (4) stable dopaminergic medication (stable dose

XSL•FO
RenderX

for at least 1 mo). The exclusion criteria were: (1) suspected parkinsonism due to causes other than idiopathic PD, (2) significant cognitive impairment (Montreal Cognitive Assessment score <19), (3) respiratory disorders or diseases, (4) current or previous history of head and neck cancer, (5) smoking within the past 5 years, (6) uncontrolled hypertension, and (7) previous experience with EMST.

Randomization and Blinding

A simple computer-generated random allocation sequence (1:1 ratio) was generated by the trial statistician prior to study initiation using a web-based randomization tool. All outcome assessors and data analysts were blinded to group allocation. However, the physiotherapists overseeing the EMST training could not be blinded, as the use of the *SpiroGym* app in the experimental group was apparent. Participants were informed that the study compared 2 EMST programs. They were not given specific details regarding the nature of each program or the hypothesis that one approach might be superior, thereby maintaining participant blinding as effectively as possible under these circumstances.

Study Outcome Parameters

The primary aim of the study was to compare EMST adherence during the unsupervised home-training phase (weeks 8 - 24) between a control group and an experimental group among patients at risk for nonadherence, defined as a SEHEPS score below 59 points at the 8-week visit. The decision to identify at-risk participants at week 8 was based on previous research [26], demonstrating no correlation between baseline SEHEPS scores and later adherence, indicating that week 8 is a more reliable point to detect adherence risk. Because the SEHEPS classification was obtained at week 8, randomization at baseline could not be stratified by SEHEPS. Allocation was not altered thereafter, and participants remained in their originally randomized arms. Secondary outcomes included changes in MEP after the intensive EMST phase (weeks 0 - 8) and the maintenance phase (weeks 8 - 24), as well as changes in SEHEPS scores following the intensive phase (weeks 0 - 8). Since the risk of nonadherence was defined at week 8, and all participants proceeded from that visit into the unsupervised phase, we did not exclude those classified as not at risk. Their data from week 8 onward are reported alongside the at-risk group to provide a comprehensive view.

Assessment Visits

The same assessment protocol was completed at pre- and posttreatment visits. Outcome examinations were carried out at the same time of day in the ON medication state (1 h after regular dopaminergic medication) at the Department of Neurology, First Faculty of Medicine, Charles University and General University Hospital in Prague, Czech Republic.

Adherence (Primary Outcome)

Adherence during the unsupervised home-training phase (weeks 8 - 24) was defined as the cumulative number of completed expiratory maneuvers. The prescribed total was 800 maneuvers. In the experimental group, adherence was monitored using the *SpiroGym* app, which automatically recorded the number of completed EMST maneuvers. In the control group, adherence was tracked via self-reported training diaries (Multimedia Appendix 2), where participants recorded the number of EMST maneuvers completed.

MEP and SEHEPS (Secondary Outcomes)

MEP assessments were performed using a flanged rubber mouthpiece connected to a pressure manometer (Micro RPM, Vyaire Medical). The procedure followed standards for respiratory muscle testing [31]. Seated upright, participants inhaled to total lung capacity and then exhaled forcefully into the manometer. Participants received verbal encouragement throughout the MEP assessment. Nose clips were applied to prevent air leakage through the nose. Participants were explicitly instructed to support their cheeks and lips with their hands during the MEP maneuver, as any cheek bulging and lip leakage can dissipate pressure and result in artificially reduced values [32]. One week before baseline testing, a training MEP session was conducted to minimize learning effects [33]. Before each formal assessment, participants completed a warm-up of 6 MEP maneuvers at 50% maximal effort [34] to account for neural facilitation from repeated efforts [33]. At least 3 measurements with a less than 10% variation between them were taken, and the highest value (cmH_2O) was recorded as the MEP.

The exercise self-efficacy was assessed using the SEHEPS [35]. The test-retest reliability of SEHEPS has been demonstrated as excellent with an intraclass correlation coefficient (ICC) of 0.96 (95% CI 0.91 - 0.98) in patients with PD [26]. The minimal detectable change at the 95% confidence limit (MDC_{95}) for SEHEPS is 12 points [35]. A total self-efficacy score is derived by summing responses to 12 items, resulting in a range from 0 to 72, where higher scores indicate greater exercise self-efficacy. A SEHEPS score of less than 59 points identifies individuals at risk of nonadherence (defined as <70% of exercise adherence) to home exercise programs [35]. At baseline, patients completed the SEHEPS based on their confidence in undertaking any long-term home exercise program. At week 8, they completed the SEHEPS again, focusing specifically on their confidence in continuing long-term EMST.

Training Protocols

Conventional EMST

Patients performed EMST at home using an expiratory muscle trainer (the EMST150; Aspire Products, LLC). During the intensive phase, therapy sessions consisted of 5 sets of 5 forceful expirations, completed on 5 self-selected days per week. The EMST150 was set at 75% of each participant's MEP value. If 75% was not initially achievable, participants were instructed to temporarily lower the threshold to a tolerable level and progressively readjust it upward across subsequent sessions, with the aim of reaching 75% as soon as feasible. Participants were instructed to occlude their nose with nose clips, take a big breath in, and blow as forcefully as possible into the device to open the valve. The expiratory effort should last a couple of seconds for the air to move through the device. A typical training session lasted approximately 10 minutes. Every 2 weeks during the intensive phase, participants attended in-person visits. At each visit, a physiotherapist reassessed MEP, readjusted the device to 75% of the participant's current MEP, and briefly

XSL•FO
RenderX

verified technique (1 set of 5 expiratory maneuvers). A full 25-repetition training session was not conducted. The same procedure was applied in both groups (control and experimental). After the initial 8-week period, patients continued EMST for an additional 16 weeks (maintenance phase) without scheduled visits. In the maintenance phase, the prescription was 5 sets of 5 forceful expirations twice weekly, with optional extra sessions to mirror real-world practice. In our clinical experience, some patients prefer to exercise more often than twice a week. During the maintenance phase, participants were instructed to maintain the threshold set at the week-8 visit. Self-adjustment was not permitted. If they felt they could tolerate a higher resistance or if resistance needed to be decreased (eg, after an intercurrent illness), they were instructed to contact the study team for an in-person visit to reassess MEP and have the device reset by the physiotherapist. The same guidance applied to both groups. This protocol was adapted from our earlier proof-of-concept study [26].

SpiroGym-Assisted EMST

Participants in the experimental group followed the same EMST schedule and intensity as the control group, but their EMST150 was coupled with the *SpiroGym* app on a study smartphone (Samsung Galaxy A20e) preloaded with the app. The app monitors expiratory training in real time by analyzing acoustic signals captured via an external microphone mounted to the EMST device using a custom 3D-printed ring fabricated by our team. A download link to the ring's build instructions and a photo of the correct attachment on the EMST150 are available in [36]. The attachment does not alter resistance, airflow, or the valve and standardizes the microphone's position and angle. The microphone was wired to the phone via a 3.5-mm jack. When the training is performed correctly, the trainer's valve opens, generating increased airflow and sound. These signals are converted by the app into a visual feedback curve displayed on the participant's smartphone (Multimedia Appendix 3). Additionally, the app automatically records session data, creating a digital training diary that enables patients to review their performance over time. Although *SpiroGym* can send exercise reminders, this feature was not enabled in the present trial because participants were loaned a study device and were unlikely to carry it during the day, making push notifications unreliable. Technical details of the app have been described in detail elsewhere [22].

Sample Size

Sample size calculations were based on the primary outcome (adherence to unsupervised EMST). Because no prior controlled studies had compared EMST alone with *SpiroGym*-assisted EMST, we estimated the effect size from an internal pilot study. We analyzed week 8 to 24 adherence in the first 10 PD participants in each arm with SEHEPS less than 59 at week 8. Their adherence data yielded a Cohen *d* of 0.92 (unpublished internal data). A power analysis using G*Power (version 3.1.9.3; Heinrich Heine University Düsseldorf) [37] indicated that 16 participants at risk for nonadherence per group would achieve 80% power at a 1-sided 5% type 1 error rate to detect this effect size. Accounting for a 20% attrition rate, each group was thus targeted to include 19 participants.

Safety

Safety was assessed by monitoring adverse events throughout the study. During the intensive phase, safety was monitored via bi-weekly in-person visits. During the maintenance phase, participants were instructed to contact the study team by phone if any adverse effects occurred.

Statistical Analysis

Baseline variables were tested for normality (Kolmogorov-Smirnov test). Normally distributed data were compared with independent 2-tailed *t* tests; nonnormal data with the Mann-Whitney *U* test. Sex distribution was analyzed by the Fisher exact test. A linear mixed-effects model was used to analyze adherence across time intervals, groups, and SEHEPS subgroups, accounting for repeated measures within individuals. The model formula was:

$$\text{adherence} \sim \text{group} \times \text{interval} \times \text{SEHEPS group} + (1 \mid \text{ID})$$

where group (control/experimental), interval (wk 0-8/8-24), and SEHEPS group (<59/≥59) were included as fixed effects and their interactions, and participant ID was included as a random intercept to account for within-individual correlation. Model assumptions were checked using residual diagnostics (Shapiro-Wilk and Breusch-Pagan tests). Post hoc analyses were performed to assess group differences at each interval. Adherence was compared between groups for weeks 0 to 8 and 8 to 24, and within SEHEPS subgroups (<59 vs ≥59) using the Mann-Whitney *U* test. Four-week interval (8-12, 12-16, 16-20, and 20-24) adherences were also analyzed using the same approach. MEP changes (0 - 8, 8 - 24, and 0 - 24 wk) were assessed with paired *t* tests within groups and independent *t* tests between groups. Effect size was expressed as Cohen *d* (95 % CI), interpreted as moderate (≥0.5) and large (≥0.8). Associations between adherence and SEHEPS were examined with Spearman ρ. Our sample-size planning used a 1-sided hypothesis based on prior evidence [26] supporting a directional expectation (*SpiroGym* >control). However, for the inferential analyses, we adopted 2-sided tests to provide a more conservative interpretation. Missing data were examined for randomness using the Little Missing Completely at Random test to determine whether the missingness was random or systematic. The missing data originated from participants who did not complete the therapy. $P<.05$ was considered statistically significant. All analyses were performed in Python (version 3.12; Python Software Foundation) using the *scipy* and *statsmodels* packages.

Ethical Considerations

This study was conducted in accordance with the principles of the Declaration of Helsinki. The study was approved prospectively by the IRB of the General University Hospital in Prague (protocol 223/20 S-IV). All participants provided written informed consent prior to any study procedures, including consent for collection, analysis, and publication of deidentified data. Participation was entirely voluntary, and participants could withdraw at any time without consequence. Participants did not receive any financial compensation. All data were deidentified before analysis and stored in password-protected files accessible only to the research team in accordance with institutional data

XSL•FO
RenderX

protection policies. The paper and supplementary materials contained no identifiable images of participants. However, Multimedia Appendix 3 shows an identifiable study researcher (not a participant). Written consent for publication from this individual has been obtained.

Results

Participants

No study-related adverse events occurred during the study period. To achieve the target of 19 participants with PD in each arm with low self-efficacy, recruitment continued until this quota was achieved, which did not occur until enrolling 75 individuals. Patients in the control group (n=38) and *SpiroGym*-assisted EMST group (n=37) were well-matched at baseline. Baseline demographics also did not differ significantly between the at-risk (SEHEPS <59) and not-at-risk (SEHEPS ≥59) subgroups. A summary of demographic and clinical characteristics is provided in Table 1. A total of nine participants from the control group and seven participants from the experimental group did not complete the study (detailed reasons are provided in Figure 1), resulting in missing data. The Little Missing Completely at Random test was nonsignificant (P>.05), indicating that the data were missing completely at random. Over the 24-week period, the mean MDS-UPDRS-III motor score increased by 1.8 (SD 5.2) points in the experimental group and by 1.7 (SD 3.9) points in the control group, with no significant difference between the groups (P=.95).

XSL•FO
RenderX

Table . Baseline demographic characteristics for all randomized participants and subgroups defined at week 8 by risk for nonadherence using the Self-Efficacy for Home Exercise Program Scale (SEHEPS): at-risk (SEHEPS <59) and not at-risk (SEHEPS ≥59).

Measure	All patients			At-risk patients, SEHEPS <59			Not-at-risk patients, SEHEPS ≥59		
	Control group	Experimental group	*P* value	Control group	Experimental group	*P* value	Control group	Experimental group	*P* value
Sex (male/female), n	21/17	24/13	.48	10/8	10/9	>.99	7/6	13/3	.15
Age (y), mean (SD)	65.4 (9.5)	65.5 (8.9)	.99	66.2 (9.0)	64.8 (8.4)	.62	67.8 (9.8)	66.2 (9.6)	.68
Height (cm), mean (SD)	174 (8.9)	173 (9.3)	.68	174.8 (8.8)	171.0 (9.1)	.21	171.8 (8.4)	175.2 (9.0)	.31
Weight (kg), mean (SD)	78.3 (19.2)	79.8 (18.5)	.74	76.2 (17.0)	79.0 (17.7)	.63	78.9 (17.5)	82.6 (20.2)	.61
BMI, mean (SD)	25.8 (5.0)	26.5 (4.4)	.51	24.8 (4.4)	26.7 (4.0)	.17	26.7 (5.4)	26.7 (4.8)	.99
Education (y), mean (SD)	14.9 (2.9)	15.6 (2.9)	.25	15.6 (3.0)	15.1 (2.7)	.60	14.3 (2.6)	15.8 (2.7)	.10
Disease duration (y), mean (SD)	7 (5.7)	7.3 (4.7)	.58	6.2 (6.2)	7.5 (4.5)	.49	6.9 (4.3)	7.6 (5.1)	.72
Hoehn-Yahr, mean (SD)	1.97 (0.6)	2 (0.5)	.84	2.0 (0.5)	2.1 (0.3)	.46	2.2 (0.6)	1.9 (0.7)	.38
MDS-UPDRS III[a] (ON), mean (SD)	17.1 (8.8)	17.5 (9.5)	.86	17.3 (9.3)	20.2 (7.5)	.30	17.2 (9.4)	15.1 (11.2)	.61
MoCa[b], mean (SD)	26.3 (3.0)	25.9 (2.9)	.64	25.7 (3.2)	26.3 (2.4)	.56	27.4 (3.1)	25.5 (3.3)	.13
LEDD[c] (mg/d), mean (SD)	672 (300)	696 (295)	.75	690.0 (313.6)	673.2 (347.9)	.89	697 (280.0)	726.4 (253.0)	.78
SEHEPS, mean (SD)	40.7 (15.4)	39.8 (17.3)	.81	32.8 (12.3)	30.0 (13.8)	.52	48.3 (16.5)	49.6 (14.6)	.83
Maximum expiratory pressure (cmH_2O), mean (SD)	130 (50)	137 (42.9)	.56	133.5 (44.7)	130.9 (45.8)	.86	128.7 (54.5)	141.4 (41.1)	.48

[a]MDS-UPDRS III: Movement Disorder Society—Unified Parkinson's Disease Rating Scale, Part III: Motor Examination

[b]MoCA: Montreal Cognitive Assessment.

[c]LEDD: levodopa equivalent daily dose.

XSL•FO
RenderX

Figure 1. Flowchart of the patients with Parkinson disease in the study. SEHEPS: Self-Efficacy for Home Exercise Program Scale.

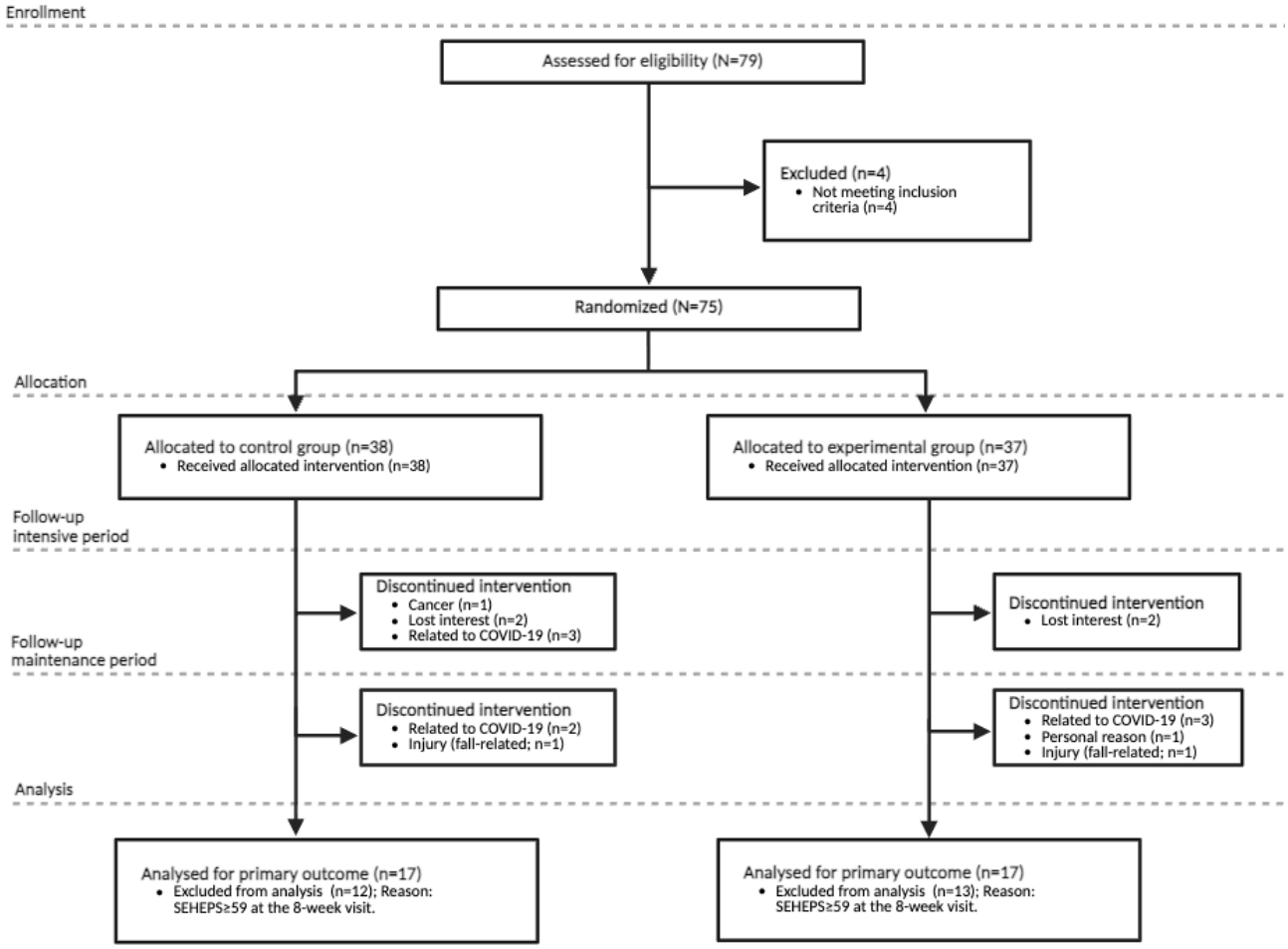

Primary Outcome: Adherence

The primary analysis using a linear mixed-effects model did not show a significant 3-way interaction between group, time interval, and baseline risk for nonadherence (group×interval×SEHEPS group; *P*=.14; Multimedia Appendix 4). However, a planned secondary analysis of participants at high risk for nonadherence (SEHEPS <59 at week 8) revealed a significant group×interval interaction. In this subgroup, the intervention group showed a significantly smaller decline in adherence during weeks 8 to 24 compared to the control group, as confirmed by a post hoc test (β=496.9, 95% CI 130.7-863.3; *P*=.008), indicating that the intervention was effective in sustaining adherence specifically in this at-risk population.

Post Hoc Analyses: Adherence

Among participants at risk for nonadherence (n=34), adherence during the maintenance phase (wk 8 - 24) was significantly higher in the experimental group than in the control group (*P*=.04; Cohen *d*=0.80; Figure 2 and Table 2). Within this group, 47% (8/17) of *SpiroGym* participants versus 18% (3/17) of controls achieved the minimum prescribed 800 or more repetitions during weeks 8 to 24. A 4-week interval analysis revealed that between-group differences in adherence reached significance during weeks 12 to 16 (*P*=.04; Cohen *d*=0.84) and weeks 20 to 24 (*P*=.01; Cohen *d*=1.03), with no significant differences during weeks 8 to 12 or 16 to 20 (Figure 3). Among patients (n=25) without risk for nonadherence, no difference in adherence was observed during the maintenance period between the control and experimental groups (*P*=.34; Cohen *d*=0.22; Table 2; Multimedia Appendix 5). For the overall sample (n=59), adherence during the intensive phase (wk 0 - 8) did not differ significantly between groups (*P*=.05; Cohen *d*=0.49), while during the maintenance phase (wk 8 - 24), the experimental group achieved significantly greater adherence (*P*=.02; Cohen *d*=0.58; Table 2).

XSL•FO
RenderX

Figure 2. Adherence to expiratory muscle strength training during weeks 8 to 24 in participants at risk for nonadherence based on the Self-Efficacy for Home Exercise Program Scale (SEHEPS <59). The values are presented as mean (SD).

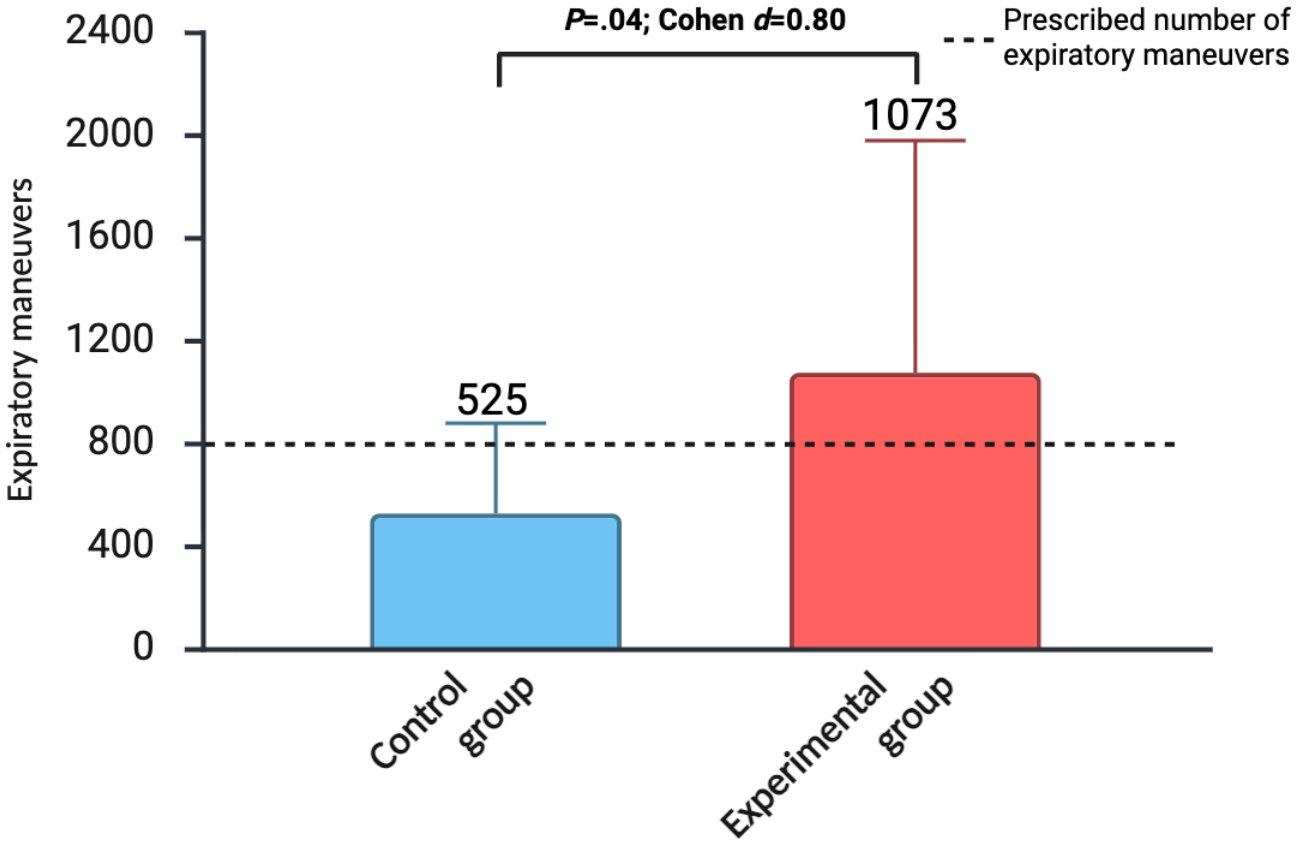

XSL•FO
RenderX

Table . Adherence to expiratory muscle strength training by study interval: weeks 0 to 8 (intensive phase) and weeks 8 to 24 (maintenance), for the full sample and for subgroups defined at week 8 by Self-Efficacy for Home Exercise Program Scale (SEHEPS) risk (at-risk: SEHEPS <59; not at-risk: SEHEPS ≥59).

Time interval and group	Participants, n	Repetitions, mean (SD)	95% CI	Between-group effect	
				P value	Cohen *d* (95% CI)
Overall adherence (all patients)					
Weeks 0-8				.05	0.49 (0.01 to 0.97)
CG[a]	32	953 (83.2)	924 to 982		
EG[b]	35	998 (95.7)	966 to 1030		
Weeks 8-24				.02	0.58 (0.06 to 1.10)
CG	29	771 (538)	575 to 967		
EG	30	1139 (712)	884 to 1393		
Adherence analysis based on SEHEPS at week 8					
Weeks 8-24				.04	0.80 (0.10 to 1.50)
CG SEHEPS <59	17	525 (351)	358 to 692		
EG SEHEPS <59	17	1073 (903)	643 to 1502		
Weeks 8-24				.34	0.22 (–0.57 to 1.0)
CG SEHEPS ≥59	12	1120 (577)	794 to 1446		
EG SEHEPS ≥59	13	1225 (352)	1034 to 1416		
Four-week interval adherence analysis in patients at risk of nonadherence					
Weeks 8-12				.15	0.74 (–0.01 to 1.4)
CG	17	165 (98.4)	118 to 212		
EG	17	277 (204)	180 to 374		
Weeks 12-16				.04	0.84 (0.14 to 1.54)
CG	17	135 (108)	84 to 186		
EG	17	301 (258)	178 to 423		
Weeks 16-20				.06	0.75 (0.05 to 1.45)
CG	17	107 (118)	51 to 163		
EG	17	255 (252)	135 to 375		
Weeks 20-24				.01	1.03 (0.32 to 1.75)
CG	17	74 (90.3)	31 to 116		
EG	17	240 (208)	141 to 339		

[a]CG: control group.

[b]EG: experimental group.

XSL•FO
RenderX

Figure 3. Four-week interval adherence to expiratory muscle strength training during weeks 8 to 24 in patients at risk for nonadherence (Self-Efficacy for Home Exercise Program Scale [SEHEPS] <59). The values are presented as mean (SD). CG: control group; EG: experimental group.

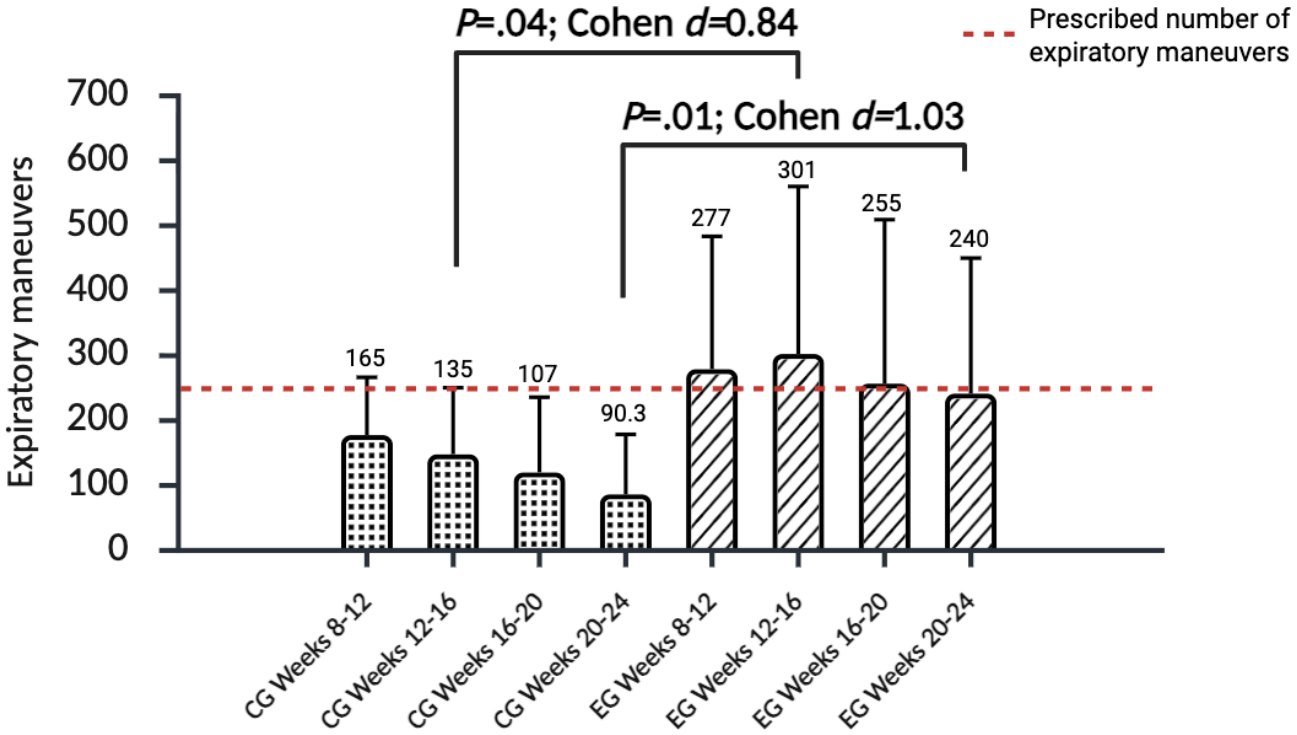

Secondary Outcomes

Maximum Expiratory Pressure

From baseline to week 8, MEP increased by 28.4 cmH_2O in the control group (*P*<.001; Cohen *d*=0.58) and by 42.3 cmH_2O in the experimental group (*P*<.001; Cohen *d*=0.89), with the experimental group showing significantly greater improvement (*P*=.03; Cohen *d*=0.54). During the unsupervised period (wk 8 - 24), MEP decreased by 6.8 cmH_2O in the control group (*P*=.11; Cohen *d*=–0.15) and increased by 0.1 cmH_2O in the experimental group (*P*=0.96, Cohen *d*=0.004), with no significant between-group difference (*P*=.16; Cohen *d*=0.37). Over the entire 24-week study period, MEP increased by 22.8 cmH_2O in the control group (*P*<.001; Cohen *d*=0.50) and by 43.1 cmH_2O in the experimental group (*P*<.001; Cohen *d*=0.91), with the experimental group demonstrating significantly greater improvements (*P*=.006; Cohen *d*=0.74). Table 3 summarizes MEP outcomes.

XSL•FO
RenderX

Table . Overall maximum expiratory pressure changes by study interval: weeks 0 to 8 (intensive phase) and weeks 8 to 24 (maintenance), for the full sample and for subgroups defined at week 8 by Self-Efficacy for Home Exercise Program Scale (SEHEPS) risk (at-risk: SEHEPS <59; not at-risk: SEHEPS ≥59).

Time interval and group	Participants (n)	Preinterval, mean (SD)	Postinterval, mean (SD)	Within-group change		Between-group effect	
				Mean (SD) Δ[a]	95% CI Δ	*P* value	Cohen *d* (95% CI)
All patients							
Weeks 0-8						.03	0.54 (0.05 to 1.03)
CG[b]	31	131.5 (48.2)	159.9 (50.3)	+28.4 (25.8)	19.3 to 37.5		
EG[c]	35	135.7 (43.4)	178.1 (51.7)	+42.3 (25.9)	33.8 to 50.9		
Weeks 8-24						.16	0.37 (–0.14 to 0.89)
CG	29	161.3 (50.4)	154.6 (42.6)	–6.8 (22.4)	–15 to 1.4		
EG	30	179.7 (51.1)	179.9 (52.8)	+0.1 (13.7)	–4.8 to 5.0		
Weeks 0-24						.006	0.74 (0.21 to 1.27)
CG	29	131.8 (48.1)	154.6 (42.6)	+22.8 (24.7)	13.8 to 31.8		
EG	30	136.8 (41.4)	179.9 (52.8)	+43.1 (29.9)	32.4 to 53.8		
Subgroup analysis based on exercise self-efficacy (SEHEPS at wk 8)							
Weeks 8-24 (SEHEPS <59)						.04	0.73 (0.04 to 1.43)
CG	17	168.1 (53.4)	153.5 (45.4)	–14.6 (21)	–24.6 to –4.6		
EG	17	180.1 (49.9)	177.9 (50.7)	–2.2 (11.3)	–7.6 to 3.1		
Weeks 8-24 (SEHEPS ≥59)						.90	–0.06 (–0.84 to 0.73)
CG	12	151.8 (46.4)	156.1 (40)	+4.2 (20.4)	–7.3 to 15.8		
EG	13	179.2 (54.5)	182.5 (57.5)	+3.2 (16.3)	–5.6 to 12.1		

[a]Δ: change.

[b]CG: control group.

[c]EG: experimental group.

Self-Efficacy for Home Exercise Program Scale

From baseline to week 8, the mean SEHEPS score significantly increased by 11.2 (SD 13.6, 95% CI 6.36-15.96; *P*<.001) points in the control group and by 14.5 (SD 14.0, 95% CI 9.91-19.2; *P*<.001) points in the experimental group. Furthermore, at week 8, SEHEPS scores correlated positively with adherence during weeks 8 to 24 in both groups (control: ρ=0.60, *P*=.001; experimental: ρ=0.44, *P*=.01).

Discussion

Principal Findings

This phase 2 RCT compared the efficacy of a conventional EMST protocol with a *SpiroGym*-assisted protocol. For the primary objective, participants at risk for EMST nonadherence who trained with *SpiroGym* sustained greater adherence during the unsupervised phase than those following conventional EMST. Regarding secondary objectives, the *SpiroGym* group achieved greater improvements in MEP and a clinically meaningful increase in self-efficacy, whereas the control group resulted in smaller but still significant gains.

Building on earlier findings [26] demonstrating sustained user engagement with the *SpiroGym* app during long-term EMST, our study adds further evidence that *SpiroGym* can effectively support adherence over time. Notably, participants with sufficient self-efficacy (SEHEPS ≥59) performed 30% more expiratory maneuvers than the minimum prescribed target during the maintenance period in both groups. In contrast, among those with low self-efficacy (SEHEPS <59), adherence in the control group dropped from 83% (wk 8 - 12) to 37% (wk 20 - 24), while the *SpiroGym* group consistently (every 4-wk interval) maintained adherence 20% above the prescribed maneuver target. These findings align with Chung et al [38], who reported higher adherence and enhanced self-efficacy in an mHealth-supported exercise program compared to a paper-based approach. Similarly, Sørensen et al [39] reported that feedback-based inspiratory muscle training yielded better

XSL•FO
RenderX

adherence than self-reported protocols over a 12-week unsupervised period.

In line with our prior study [26], we observed a positive correlation between SEHEPS scores at week 8 and adherence during weeks 8 to 24 in both groups, suggesting that self-efficacy assessed after the intensive phase may predict long-term EMST compliance. Although SEHEPS scores improved significantly in both protocols, only the *SpiroGym* group exceeded the 12-point MDC_{95} threshold [35], indicating a true increase in exercise self-efficacy beyond measurement error—a finding supported by previous research [40].

These findings have direct clinical implications for tailoring EMST. Patients should be assessed for self-efficacy following the intensive phase. Those with low self-efficacy may benefit from enhanced supervision or motivational support, such as the *SpiroGym* app, to sustain adherence during the maintenance period. Conversely, conventional EMST may be sufficient for patients with higher self-efficacy. Beyond self-efficacy, other factors such as apathy or caregiver support may plausibly influence adherence and treatment outcomes and may guide treatment selection in practice [41,42]. Screening for these factors alongside SEHEPS may help triage patients to conventional versus supported EMST.

Another key finding relates to real-time visual feedback and its impact on MEP. After the 8-week intensive phase, both groups achieved meaningful MEP gains, but the increase was significantly larger in the *SpiroGym* group (P=.03; Cohen d=0.54). On average, MEP increased by 22% in the control group and 31% in the experimental group, consistent with prior reports of approximately 24% MEP gains in PD [8]. Based on the observation that, during weeks 0 - 8, the experimental group showed greater MEP gains despite comparable adherence, we hypothesize that real-time visual feedback may enhance motivational aspects of learning and thereby facilitate better skill acquisition. This interpretation is consistent with prior studies showing amplified respiratory muscle activation under visual feedback conditions [43]. Given the role of EMST in airway protection [4,6], these superior MEP outcomes suggest that SpiroGym-assisted EMST may translate into improved clinical outcomes for individuals with dysphagia or impaired cough, although this remains to be confirmed in future research.

A further notable result pertains to long-term MEP maintenance in patients at risk for nonadherence (SEHEPS <59). While the control group showed an average 9% decline in MEP between weeks 8 and 24, the *SpiroGym* group maintained stable values. Previous studies report detraining-related MEP declines of 2% to 20% over 1 to 18 months in PD [9,44,45]. Long-term maintenance training at least twice per week is generally recommended [46]. In our study, control group participants missed an average of 11 out of 32 scheduled sessions, falling below the minimum recommended dose and likely contributing to the observed MEP reduction. In contrast, adherence in the experimental group was sufficient to prevent detraining effects. On the other hand, resistance-training adaptations typically extend beyond 8 weeks in both healthy adults and people with PD. One likely explanation for the plateau in MEP gains is that no participant requested recalibration during the maintenance period. As a result, training thresholds remained fixed, reducing progressive resistance. This contrasts with evidence from a 2-year randomized trial, in which Corcos et al [47] showed that people with PD who trained twice weekly with progressively increased resistance continued to improve upper-limb strength up to the 24-month visit, whereas those following a nonprogressive program did not. A second contributor may be the device load ceiling. In a subset of experimental participants (n=11), week-8 MEP exceeded 200 cmH_2O, placing the 75% training target above the EMST150's maximum load. To sustain MEP progression, future protocols should implement routine recalibration during the maintenance period and, for participants approaching the ceiling, employ devices with a higher resistance range.

Limitations

This study has several limitations. First, adherence in the control group was based on self-reported training diaries, which might be prone to inaccuracy [48]. Second, we did not include objective measures of swallowing function or cough efficacy, limiting conclusions about the broader clinical impact of the observed increase in the MEP. Third, the sample primarily consisted of individuals with mild-to-moderate PD, which limits the generalizability of findings to those with more advanced PD or cognitive impairment. Fourth, we did not collect person-centered outcomes (eg, treatment burden and quality of life). Although a separate study reports acceptability of *SpiroGym*, it does not provide a head-to-head comparison with conventional EMST [25,26]. Consequently, it remains unclear whether SpiroGym-supported EMST reduces perceived burden or improves quality of life relative to standard care. Fifth, during the maintenance phase, participants were instructed to perform five sets of five expirations twice weekly, with the option to train more frequently to reflect real-world practice. This flexibility allowed some individuals to accumulate higher totals of expiratory maneuvers, which may have influenced the effect size.

Future studies should involve larger cohorts, longer follow-up periods, and objective assessments of airway protection. In addition, qualitative and mixed methods studies comparing *SpiroGym* with traditional EMST are warranted to understand patient experience (eg, satisfaction, motivation, enjoyment, and usability).

To support further research and clinical adoption, the *SpiroGym* app is freely available at [36], along with detailed hardware specifications. A user manual is included in Multimedia Appendix 6.

Conclusions

This is the first RCT to integrate mHealth with EMST. Unlike prior studies in the EMST field, we focused on long-term exercise adherence, which is critical for sustaining therapeutic effects. Compared with conventional EMST, *SpiroGym*-assisted training led to higher adherence, most notably among patients prone to nonadherence, and produced greater expiratory muscle strength gains. In routine PD care, evaluating self-efficacy after the supervised EMST phase may help identify patients who are likely to benefit from digital support, positioning

XSL•FO
RenderX

mHealth-assisted EMST as a practical strategy to sustain exercise adherence.

Acknowledgments

The authors thank the individuals with Parkinson disease and their families who participated in this study for their immense dedication and contribution to this work. The authors also thank Pavel Dusek for his assistance with patient recruitment and his methodological guidance. The authors declare the use of generative artificial intelligence in the research and writing process. According to GAIDeT (Generative AI Delegation Taxonomy; 2025) [49], the following tasks were delegated to generative artificial intelligence tools under full human supervision: proofreading and editing. ChatGPT 5.2 (GPT; OpenAI's large-scale language model) was used for these purposes.

Funding

The study was supported by the Czech Health Research Council (project NW24-04-00259) and by General University Hospital in Prague (project MH CZ-DRO-VFN64165).

Data Availability

The authors have deposited an open-access deidentified participant-level dataset along with the source code for statistical analysis on GitHub [50].

Authors' Contributions

MS was involved in the conception, organization, and execution of the research project; design, execution, review, and critique of statistical analysis; and writing the original draft of the manuscript. MH was involved in the conception, organization, and execution of the research project; design, execution, review, and critique of statistical analysis; and writing the original draft of the manuscript. RL was involved in the conception and organization of the research project and reviewing and critiquing the manuscript. KD was involved in the execution of the research project, reviewing and critiquing the statistical analysis, and reviewing and critiquing the manuscript. R Kliment was involved in the conception of the research project and reviewing and critiquing the manuscript. JM was involved in the conception of the research project and reviewing and critiquing the manuscript. R Krupicka was involved in the conception and organization of the research project and reviewing and critiquing the manuscript. OG was involved in the conception of the research project, reviewing and critiquing statistical analysis, and reviewing and critiquing the manuscript. ER was involved in the conception of the research project, reviewing and critiquing statistical analysis, and reviewing and critiquing the manuscript.

Conflicts of Interest

MS, RL, R Kliment, and ER are creators of the *SpiroGym* app. All other authors declare no conflicts of interest.

Editorial Notice

This randomized study was only retrospectively registered because of an initial lack of awareness of the requirement for prospective registration prior to study commencement, with the need for registration recognized only during the course of the study. The editor granted an exception from the International Committee of Medical Journal Editors' rules mandating prospective registration of randomized trials because the risk of bias appears low and because the authors provided an official letter from the institutional review board confirming that the submitted study design is identical to the protocol assessed and approved before trial initiation. However, readers are advised to carefully assess the validity of any potential explicit or implicit claims related to primary outcomes or effectiveness, as retrospective registration does not prevent authors from changing their outcome measures retrospectively.

Multimedia Appendix 1
A side-by-side comparison of key protocol elements across the institutional review board submission, the ClinicalTrials.gov record (NCT05728099), and published manuscript.
[DOCX File, 18 KB - jmir_v28i1e78022_app1.docx]

Multimedia Appendix 2
Self-monitoring training diaries.
[DOCX File, 17 KB - jmir_v28i1e78022_app2.docx]

Multimedia Appendix 3
Visual feedback provided by the SpiroGym app during expiratory muscle strength training.
[MP4 File, 10843 KB - jmir_v28i1e78022_app3.mp4]

XSL•FO
RenderX

Multimedia Appendix 4
Linear mixed-effects model for adherence.
[DOCX File, 18 KB - jmir_v28i1e78022_app4.docx]

Multimedia Appendix 5
Four-week interval adherence analysis in patients without risk of nonadherence.
[DOCX File, 17 KB - jmir_v28i1e78022_app5.docx]

Multimedia Appendix 6
SpiroGym user manual.
[DOCX File, 438 KB - jmir_v28i1e78022_app6.docx]

Checklist 1
CONSORT-EHEALTH checklist.
[PDF File, 1247 KB - jmir_v28i1e78022_app7.pdf]

References

1. Dorsey ER, Bloem BR. The Parkinson pandemic—a call to action. JAMA Neurol 2018 Jan 1;75(1):9-10. [doi: 10.1001/jamaneurol.2017.3299] [Medline: 29131880]
2. Steinmetz JD, Seeher KM, Schiess N, et al. Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet Neurol 2024 Apr;23(4):344-381. [doi: 10.1016/S1474-4422(24)00038-3]
3. Magrinelli F, Picelli A, Tocco P, et al. Pathophysiology of motor dysfunction in Parkinson's disease as the rationale for drug treatment and rehabilitation. Parkinsons Dis 2016;2016(1):9832839. [doi: 10.1155/2016/9832839] [Medline: 27366343]
4. D'Amelio M, Ragonese P, Morgante L, et al. Long-term survival of Parkinson's disease: a population-based study. J Neurol 2006 Jan;253(1):33-37. [doi: 10.1007/s00415-005-0916-7] [Medline: 16021349]
5. Fall PA, Saleh A, Fredrickson M, Olsson JE, Granérus AK. Survival time, mortality, and cause of death in elderly patients with Parkinson's disease: a 9-year follow-up. Mov Disord 2003 Nov;18(11):1312-1316. [doi: 10.1002/mds.10537] [Medline: 14639673]
6. Pennington S, Snell K, Lee M, Walker R. The cause of death in idiopathic Parkinson's disease. Parkinsonism Relat Disord 2010 Aug;16(7):434-437. [doi: 10.1016/j.parkreldis.2010.04.010]
7. Claus I, Muhle P, Czechowski J, et al. Expiratory muscle strength training for therapy of pharyngeal dysphagia in Parkinson's disease. Mov Disord 2021 Aug;36(8):1815-1824. [doi: 10.1002/mds.28552] [Medline: 33650729]
8. Troche MS, Curtis JA, Sevitz JS, et al. Rehabilitating cough dysfunction in Parkinson's disease: a randomized controlled trial. Mov Disord 2023 Feb;38(2):201-211. [doi: 10.1002/mds.29268] [Medline: 36345090]
9. Troche MS, Rosenbek JC, Okun MS, Sapienza CM. Detraining outcomes with expiratory muscle strength training in Parkinson disease. J Rehabil Res Dev 2014;51(2):305-310. [doi: 10.1682/JRRD.2013.05.0101] [Medline: 24933728]
10. Ellis T, Cavanaugh JT, Earhart GM, et al. Factors associated with exercise behavior in people with Parkinson disease. Phys Ther 2011 Dec;91(12):1838-1848. [doi: 10.2522/ptj.20100390] [Medline: 22003171]
11. Isernia S, Pagliari C, Jonsdottir J, et al. Efficiency and patient-reported outcome measures from clinic to home: the human empowerment aging and disability program for digital-health rehabilitation. Front Neurol 2019;10:1206. [doi: 10.3389/fneur.2019.01206] [Medline: 31824398]
12. Stevens A, Stanton R, Rebar AL. Helping people with Parkinson disease build exercise self-efficacy. Phys Ther 2020 Feb 7;100(2):205-208. [doi: 10.1093/ptj/pzz160] [Medline: 31665447]
13. Lee J, Yeom I, Chung ML, Kim Y, Yoo S, Kim E. Use of mobile apps for self-care in people with Parkinson disease: systematic review. JMIR Mhealth Uhealth 2022 Jan 21;10(1):e33944. [doi: 10.2196/33944] [Medline: 35060910]
14. Templeton JM, Poellabauer C, Schneider S, et al. Modernizing the staging of Parkinson disease using digital health technology. J Med Internet Res 2025 Apr 4;27:e63105. [doi: 10.2196/63105] [Medline: 40184612]
15. Willemse IHJ, Schootemeijer S, van den Bergh R, Dawes H, Nonnekes JH, van de Warrenburg BPC. Smartphone applications for movement disorders: towards collaboration and re-use. Parkinsonism Relat Disord 2024 Mar;120:105988. [doi: 10.1016/j.parkreldis.2023.105988] [Medline: 38184466]
16. Litman L, Rosen Z, Spierer D, Weinberger-Litman S, Goldschein A, Robinson J. Mobile exercise apps and increased leisure time exercise activity: a moderated mediation analysis of the role of self-efficacy and barriers. J Med Internet Res 2015 Aug 14;17(8):e195. [doi: 10.2196/jmir.4142] [Medline: 26276227]
17. Ashford S, Edmunds J, French DP. What is the best way to change self-efficacy to promote lifestyle and recreational physical activity? A systematic review with meta-analysis. Br J Health Psychol 2010 May;15(Pt 2):265-288. [doi: 10.1348/135910709X461752] [Medline: 19586583]

XSL•FO
RenderX

18. Jancey J, Lee A, Howat P, Clarke A, Wang K, Shilton T. Reducing attrition in physical activity programs for older adults. J Aging Phys Act 2007 Apr;15(2):152-165. [doi: 10.1123/japa.15.2.152] [Medline: 17556782]
19. Oman RF, King AC. Predicting the adoption and maintenance of exercise participation using self-efficacy and previous exercise participation rates. Am J Health Promot 1998;12(3):154-161. [doi: 10.4278/0890-1171-12.3.154] [Medline: 10176088]
20. Rhodes RE, Martin AD, Taunton JE. Temporal relationships of self-efficacy and social support as predictors of adherence in a 6-month strength-training program for older women. Percept Mot Skills 2001 Dec;93(3):693-703. [doi: 10.2466/pms.2001.93.3.693] [Medline: 11806588]
21. Schootemeijer S, van der Kolk NM, Ellis T, et al. Barriers and motivators to engage in exercise for persons with Parkinson's disease. J Parkinsons Dis 2020;10(4):1293-1299. [doi: 10.3233/JPD-202247] [Medline: 32925106]
22. Srp M, Korteová R, Kliment R, Jech R, Růžička E, Hoskovcová M. Expiratory muscle strength training in patients with Parkinson's disease: a pilot study of mobile monitoring application. Mov Disord Clin Pract 2021 Oct;8(7):1148-1149. [doi: 10.1002/mdc3.13313] [Medline: 34631956]
23. Reyes A, Castillo A, Castillo J. Effects of expiratory muscle training and air stacking on peak cough flow in individuals with Parkinson's disease. Lung 2020 Feb;198(1):207-211. [doi: 10.1007/s00408-019-00291-8] [Medline: 31720808]
24. Reyes A, Castillo A, Castillo J, Cornejo I. The effects of respiratory muscle training on peak cough flow in patients with Parkinson’s disease: a randomized controlled study. Clin Rehabil 2018 Oct;32(10):1317-1327. [doi: 10.1177/0269215518774832] [Medline: 29756459]
25. Pérez-Rangel D, Monje MHG, Grosjean S, et al. Insights from a mixed methods analysis of 3 health technologies used in patients with Parkinson disease: mixed methods study. J Med Internet Res 2025 Aug 1;27:e67986. [doi: 10.2196/67986] [Medline: 40749223]
26. Srp M, Ferro Á, Ferreira J, et al. mHealth-assisted expiratory muscle strength training in Parkinson’s disease patients: a proof-of-concept study. J Parkinsons Dis 2024 Nov;14(8):1623-1630. [doi: 10.1177/1877718X241296013] [Medline: 39957190]
27. Brassington GS, Atienza AA, Perczek RE, DiLorenzo TM, King AC. Intervention-related cognitive versus social mediators of exercise adherence in the elderly. Am J Prev Med 2002 Aug;23(2 Suppl):80-86. [doi: 10.1016/s0749-3797(02)00477-4] [Medline: 12133741]
28. Neupert SD, Lachman ME, Whitbourne SB. Exercise self-efficacy and control beliefs: effects on exercise behavior after an exercise intervention for older adults. J Aging Phys Act 2009 Jan;17(1):1-16. [doi: 10.1123/japa.17.1.1] [Medline: 19299835]
29. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
30. Hughes AJ, Daniel SE, Kilford L, Lees AJ. Accuracy of clinical diagnosis of idiopathic Parkinson’s disease: a clinico-pathological study of 100 cases. J Neurol Neurosurg Psychiatry 1992 Mar 1;55(3):181-184. [doi: 10.1136/jnnp.55.3.181]
31. Laveneziana P, Albuquerque A, Aliverti A, et al. ERS statement on respiratory muscle testing at rest and during exercise. Eur Respir J 2019 Jun;53(6):1801214. [doi: 10.1183/13993003.01214-2018] [Medline: 30956204]
32. Rubinstein I, Slutsky AS, Rebuck AS, et al. Assessment of maximal expiratory pressure in healthy adults. J Appl Physiol 1988 May;64(5):2215-2219. [doi: 10.1152/jappl.1988.64.5.2215] [Medline: 3391920]
33. Windisch W, Criée CP. Maximum respiratory pressure assessment: standardization is required. Chest 2022 Oct;162(4):738-739. [doi: 10.1016/j.chest.2022.05.033] [Medline: 36210100]
34. Marion R, Niebuhr BR. Effect of warm-up prior to maximal grip contractions. J Hand Ther 1992 Jul;5(3):143-146. [doi: 10.1016/S0894-1130(12)80349-8]
35. Picha KJ, Lester M, Heebner NR, et al. The self-efficacy for home exercise programs scale: development and psychometric properties. J Orthop Sports Phys Ther 2019 Sep;49(9):647-655. [doi: 10.2519/jospt.2019.8779]
36. SpiroGym. GitHub. URL: https://fbme-ictm.github.io/spirogym [accessed 2026-02-13]
37. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007 May;39(2):175-191. [doi: 10.3758/bf03193146] [Medline: 17695343]
38. Chung BPH, Chiang WKH, Lau H, et al. Pilot study on comparisons between the effectiveness of mobile video-guided and paper-based home exercise programs on improving exercise adherence, self-efficacy for exercise and functional outcomes of patients with stroke with 3-month follow-up: a single-blind randomized controlled trial. Hong Kong Physiother J 2020 Jun;40(1):63-73. [doi: 10.1142/S1013702520500079] [Medline: 32489241]
39. Sørensen D, Svenningsen H. Adherence to home-based inspiratory muscle training in individuals with chronic obstructive pulmonary disease. Appl Nurs Res 2018 Oct;43:75-79. [doi: 10.1016/j.apnr.2018.07.005] [Medline: 30220368]
40. Theis S, Schäfer D, Haubrich C, et al. Perceived self-efficacy in Parkinson’s disease through mobile health monitoring. In: Stephanidis C, Antona M, Gao Q, Zhou J, editors. HCI International 2020 – Late Breaking Papers: Universal Access and Inclusive Design: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 2020:749-762. [doi: 10.1007/978-3-030-60149-2_57]

XSL•FO
RenderX

41. Cools CI, Kotz SA, Bloem BR, Duits AA, de Vries NM. Motivation matters: elucidating factors driving exercise in people with Parkinson disease. Phys Ther 2025 Jun 2;105(6):pzaf048. [doi: 10.1093/ptj/pzaf048] [Medline: 40207869]
42. Kim M, Kim E, Kim M, et al. Motivators and barriers affecting exercise in patients with Parkinson's disease. J Clin Neurol 2025 Jan;21(1):13-20. [doi: 10.3988/jcn.2024.0328] [Medline: 39778563]
43. Park HK, Kim YJ, Kim TH. The role of visual feedback in respiratory muscle activation and pulmonary function. J Phys Ther Sci 2015 Sep;27(9):2883-2886. [doi: 10.1589/jpts.27.2883] [Medline: 26504317]
44. Huang CC, Lai YR, Wu FA, et al. Detraining effect on pulmonary and cardiovascular autonomic function and functional outcomes in patients with Parkinson's disease after respiratory muscle training: an 18-month follow-up study. Front Neurol 2021;12:735847. [doi: 10.3389/fneur.2021.735847] [Medline: 34744975]
45. Saleem AF, Sapienza CM, Okun MS. Respiratory muscle strength training: treatment and response duration in a patient with early idiopathic Parkinson's disease. NeuroRehabilitation 2005;20(4):323-333. [Medline: 16403998]
46. Sapienza C, Hoffman B. Respiratory Muscle Strength Training: Plural Publishing; 2020.
47. Corcos DM, Robichaud JA, David FJ, et al. A two-year randomized controlled trial of progressive resistance exercise for Parkinson's disease. Mov Disord 2013 Aug;28(9):1230-1240. [doi: 10.1002/mds.25380] [Medline: 23536417]
48. Schmidt M, Paul SS, Canning CG, et al. The accuracy of self-report logbooks of adherence to prescribed home-based exercise in Parkinson's disease. Disabil Rehabil 2022 Apr;44(8):1260-1267. [doi: 10.1080/09638288.2020.1800106] [Medline: 32762573]
49. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (Generative AI Delegation Taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res 2025 Aug 8:1-27. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]
50. Expiratory muscle strength training in Parkinson's disease - analysis. GitHub. URL: https://github.com/kbi-fbmi/articles--2025expiratory [accessed 2026-02-26]

Abbreviations

CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
EMST: expiratory muscle strength training
ICC: intraclass correlation coefficient
IRB: institutional review board
MDC95: minimal detectable change at the 95% confidence level
MEP: maximum expiratory pressure
mHealth: mobile health
PD: Parkinson disease
RCT: randomized controlled trial
SEHEPS: Self-Efficacy for Home Exercise Program Scale

Edited by S Brini; submitted 24.May.2025; peer-reviewed by A Dakin, J Sevitz; revised version received 17.Jan.2026; accepted 18.Jan.2026; published 11.Mar.2026.

Please cite as:
Srp M, Hoskovcova M, Lagnerova R, Dvorakova K, Kliment R, Muzik J, Krupicka R, Gal O, Ruzicka E
Enhancing Adherence to Home-Based Expiratory Muscle Strength Training in Parkinson Disease: Randomized Controlled Trial of an mHealth Intervention
J Med Internet Res 2026;28:e78022
URL: https://www.jmir.org/2026/1/e78022
doi:10.2196/78022

© Martin Srp, Martina Hoskovcova, Rebeka Lagnerova, Katerina Dvorakova, Radim Kliment, Jan Muzik, Radim Krupicka, Ota Gal, Evzen Ruzicka. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Comparing the Associations of Internet Addiction and Internet Gaming Disorder With Psychopathological Symptoms: Cross-Sectional Study of Three Independent Adolescent Samples

Ying-ying Li[1*], BMBS; A-qian Hu[1*], MM; Ling-li Yi[2,3*], MM; Zi-xin Mao[3,4], MM; Qiu-yue Lü[3], MD, PhD; Juan Wang[1], PhD; Wei Wei[1], MD, PhD; Yue-qi Huang[1], MM; Shu Huang[1], BMBS; Wen-jing Dai[1,5], BMBS; Meng-xuan Qiao[1], MM; Jia-jun Xu[3], MD, PhD; Qiang Wang[3], MD, PhD; Xiao-jing Li[1,3,6,7], MD, PhD; Fu-gang Luo[1], BMBS; Wei Deng[1,7], MD, PhD; Yu-zheng Hu[6,7], PhD; Tao Li[1,3,5,6,7], MD, PhD; Wan-jun Guo[1,3,5,6,7], MD, PhD

[1]Affiliated Mental Health Center & Hangzhou Seventh People's Hospital, School of Brain Science and Brain Medicine, Zhejiang University School of Medicine, 305 Tiamnyshan Rd, Hangzhou, Zhejiang, China

[2]MOE Key Lab for Neuroinformation, The Clinical Hospital of Chengdu Brain Science Institute, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

[3]Mental Health Center and Psychiatric Laboratory, State Key Laboratory of Biotherapy, West China Hospital of Sichuan University, Chengdu, Sichuan, China

[4]Department of Psychiatry, Yichang Mental Health Center, Hubei, China

[5]School of Mental Health, Wenzhou Medical University, 325035, Wenzhou, China

[6]Department of Psychology and Behavioral Science, Zhejiang University, Hangzhou, 148 Tianmushan Road, Xihu District, Hangzhou, China

[7]Liangzhu Laboratory, MOE Frontier Science Center for Brain Science and Brain-Machine Integration, State Key Laboratory of Brain-Machine Intelligence, Zhejiang University, Hangzhou, China

[*]these authors contributed equally

Corresponding Author:
Wan-jun Guo, MD, PhD
Affiliated Mental Health Center & Hangzhou Seventh People's Hospital, School of Brain Science and Brain Medicine, Zhejiang University School of Medicine, 305 Tiamnyshan Rd, Hangzhou, Zhejiang, China

Abstract

Background: Both internet gaming disorder (IGD) and internet addiction (IA) have been associated with diverse psychopathological symptoms. However, how the 2 conditions relate to each other and which is more strongly associated with psychopathology remain unclear.

Objective: This study aimed to examine the association between IGD and IA and compare the strength of their associations with various types of psychopathological symptoms.

Methods: This cross-sectional study surveyed 3 independent samples of Chinese adolescents: the first sample (S1) comprised 8194 first-year undergraduates at a comprehensive university in Chengdu, the second sample (S2) comprised 1720 students from a high school in Hangzhou, and the third sample (S3) comprised 551 inpatients aged 13 to 19 years recruited from 2 tertiary psychiatric hospitals in Hangzhou and Chengdu. IGD was defined as a score of 22 or more on the Internet Gaming Disorder Scale–Short Form (IGDS9-SF), whereas IA was defined as a score of 50 or more on Young's 20-item Internet Addiction Test (IAT-20). Symptoms of depression, anxiety, psychoticism, paranoid ideation, and attention-deficit or hyperactivity were assessed using internationally validated scales including 9-item the Patient Health Questionnaire, 7-item Generalized Anxiety Disorder, psychoticism and paranoid ideation subscales of the Symptom Checklist 90 (absent for S2), and Adult ADHD Self-Report Scale (absent for S1), through online surveys in S1 (October 2020) and S3 (January 2022 to February 2025) and via an offline survey in S2 (March 2024).

Results: The prevalence estimates (95% CI) of IGD were 4.8% (4.3% - 5.2%) in S1, 15.8% (14.0% - 17.5%) in S2, and 32.3% (28.4% - 36.2%) in S3, whereas prevalence estimates (95% CI) of IA were consistently higher across samples, ranging from 7.3% (6.8% - 7.9%) in S1 and 18.8% (17.0% - 20.6%) in S2 to 45.9% (41.8% - 50.1%) in S3. The IGDS9-SF and the IAT-20 were moderately correlated (Pearson r=0.51 - 0.57; all P<.001) and were associated with the severity of most psychopathological symptom domains, with consistently stronger associations observed for IAT-20 scores. In multivariate models including all psychopathological symptoms as independent variables, the coefficients of determination (R^2, 95% CIs) were consistently higher for the IAT-20 than for the IGDS9-SF in S1 (0.33, 0.30 - 0.35 vs 0.13, 0.11 - 0.16) and S2 (0.44, 0.39 - 0.49 vs 0.23, 0.18 - 0.27), with a similar but nonsignificant pattern observed in S3 (0.13, 0.06 - 0.26 vs 0.06, 0.03 - 0.16). Post hoc analyses indicated that

XSL•FO
RenderX

psychopathological symptoms were generally more severe in individuals with IA, either alone or comorbid with IGD, than in those with IGD only.

Conclusions: This study provides additional evidence that IGD and IA are distinct yet interrelated constructs, and further demonstrates that IA consistently exhibits stronger associations with the severity of psychopathological symptoms than IGD. These findings underscore the importance of recognizing and addressing compulsive and problematic online behaviors that extend beyond gaming, highlighting the need to refine diagnostic frameworks and prioritize targeted clinical interventions.

(J Med Internet Res 2026;28:e82414) doi:10.2196/82414

KEYWORDS

adolescent; attention deficit; hyperactivity; internet addiction; internet gaming disorder; psychopathology

Introduction

In 2023, approximately 5.4 billion people were using the internet [1], many of whom were teenagers and young adults. These groups are at a critical stage of growth and development, which can be endangered by excessive or inappropriate use of the internet, potentially leading to internet gaming disorder (IGD) and internet addiction (IA) [2,3]. Recent systematic reviews and meta-analyses indicate that IA and IGD are increasingly prevalent worldwide, particularly among adolescents and young adults, and are associated with a growing burden of mental health and functional impairments [4,5].

IGD has been proposed in the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* as a condition warranting further study [6], which specifies 9 dimensions of preoccupation, withdrawal, tolerance, unsuccessful attempts to stop or limit gaming, loss of interest in previous activities, continued use despite harm, avoidance, deception, and harm. The 11th edition of the *International Classification of Diseases, Eleventh Revision* (*ICD-11*) has defined gaming disorder, including predominantly online gaming, emphasizing impaired control, increasing priority given to gaming over other activities, and continued gaming despite negative consequences [7]. Both definitions imply significant social dysfunction arising from excessive online gaming. Studies based on one of the most widely used validated instruments for assessing internet gaming disorder that was proposed by a meta-analysis for screening internet gaming disorder, the Internet Gaming Disorder Scale–Short Form (IGDS9-SF) [8], suggest that the condition affects 9.3%-18.2% of adolescents in various countries [9,10]. Recent epidemiological and longitudinal studies have consistently shown that IGD in adolescents is associated with a wide range of adverse outcomes, including depressive and anxiety symptoms, attention-deficit or hyperactivity problems, aggression, sleep disturbances, poorer quality of life, and suicidality [11-13].

In contrast to IGD, IA has yet to be recognized as a bona fide disorder in the *Diagnostic and Statistical Manual* or other consensus guidelines [14]. It is typically defined as a behavioral addiction characterized by impaired control over internet use or online behaviors that results in clinically significant distress and functional impairment [15]. Studies using the most widely used instrument for assessing IA, the 20-item Young's Internet Addiction Test (IAT-20) [16], suggest that up to 18% of adolescents may suffer from moderate IA and up to 1.5% may suffer from severe IA [17]. Studies based on the IAT-20 have linked IA to diverse psychopathological symptoms including depression, anxiety, obsessive-compulsive disorder, and suicidal ideation [18,19].

How IA and IGD relate to each other is unclear, with studies proposing that they are one and the same, that they are entirely separate from each other, or that one is a subtype of the other [20,21]. Although recent work has begun to assess generalized problematic internet use, gaming-specific problems, and other online behaviors within the same samples, highlighting both shared risk factors (eg, negative affect, maladaptive coping) and partially distinct symptom profiles across these conditions [22,23], their associations with psychopathological symptoms have not been systematically compared between IGD and IA within the same samples. Therefore, this study employed internationally validated instruments to assess IA and IGD in 3 adolescent samples from distinct settings in China, with the goal of examining the association between the 2 conditions and determining which is more strongly related to psychopathological symptom severity.

Methods

Participants

This study involved participants from 3 independent samples. Sample 1 (S1) comprised first-year undergraduates enrolled in 2020 at Sichuan University, a large comprehensive university in southwestern China that recruits students from all provincial-level administrative regions nationwide. As part of the Development of Psychological Health Assessment and Crisis Alarm and Intervention System project, all freshmen (n=9409) were invited to complete an online, self-administered psychosomatic health questionnaire via the university's Online Psychosomatic Health Survey in October 2020. This study used the data from 8125 freshmen who provided valid responses to the Online Psychosomatic Health Survey, representing a response rate of 86.4%. Sample 2 (S2) comprised 1720 respondents (response rate: 78.2%) out of 2200 students invited to participate in a school-wide mental health survey at a high school in Hangzhou, China, conducted in March 2024 using paper-based questionnaires. Sample 3 (S3) comprised 551 respondents (response rate: 86.8%) out of 635 psychiatric inpatients aged 13‑19 years recruited from 2 psychiatric hospitals in Hangzhou and Chengdu, China, to complete online self-report scales of psychiatric symptoms between January 2022 and February 2025.

XSL•FO
RenderX

Individuals were excluded from the study if they were younger than 13 or older than 19 years, if they did not complete all questionnaire items, if their responses seemed unreliable (eg, endorsing the same symptom severity across all items), or if they used the same personal identification number as another participant. Undergraduates were excluded if they failed to submit their questionnaires properly through the online system or submitted them after the designated deadline.

Measurements

The questionnaire packets contained queries on sociodemographic background along with various internationally validated assessment scales, including the IGDS9-SF, the IAT-20, the Patient Health Questionnaire-9 (PHQ-9), Generalized Anxiety Disorder-7 (GAD-7), psychoticism and paranoid Ideation Subscales of the Symptom Checklist 90 (administered only for S1 and S3), and Adult ADHD Self-Report Scale (ASRS; administered only for S2 and S3). These instruments were used to assess IGD, IA, and 5 domains of psychopathology (depression, anxiety, psychoticism, paranoid ideation, and attention-deficit or hyperactivity symptoms), respectively.

Internet Gaming Disorder Scale–Short Form

The IGDS9-SF [8] assesses the frequency of symptoms of IGD during the previous 12 months. The symptoms correspond to the diagnostic criteria of the 5th edition of the *Diagnostic and Statistical Manual* and the features defined by the 11th edition of the *International Classification of Diseases*. The IGDS9-SF is one of the few questionnaires assessing IGD that includes the criterion of "continued internet use despite harm." Participants respond to each of the 9 items on a 5-point Likert scale from 1 ("rarely") to 5 ("always"), yielding a total score ranging from 0 to 45. Scores ≥22 were defined as indicative of IGD [24]. Cronbach α in all 3 samples of this study ranged from 0.89 to 0.93.

Twenty-Item Young's Internet Addiction Test

The IAT-20 [25] is the most frequently used self-report measure of problematic internet use [26]. Participants respond to each of the 20 items on a 5-point Likert scale from 1 ("rarely") to 5 ("always") or by entering the value 0 ("not applicable"), yielding a total score from 0 to 100. Scores ≥50 were considered indicative of IA [27]. In this study, Cronbach α across the 3 samples ranged from 0.93 to 0.94.

Nine-Item Patient Health Questionnaire

The PHQ-9 assesses the severity of depressive symptoms over the past 2 weeks. Participants respond to each of the 9 items using a 4-point scale ranging from 0 ("not at all") to 3 ("nearly every day"), yielding a total score range of 0 - 27. Higher scores indicate greater depressive symptom severity. The scale has demonstrated satisfactory psychometric properties among Chinese populations [28]. In this study, Cronbach α across the 3 samples ranged from 0.86 to 0.91.

Seven-Item Generalized Anxiety Disorder

The GAD-7 [29] assesses the severity of anxiety symptoms during the previous 2 weeks. Participants respond to each of 7 items on a 4-point scale from 0 ("not at all") to 3 ("nearly every day"), yielding a total score range of 0-21. Higher scores indicate greater anxiety symptom severity. The Chinese version has demonstrated satisfactory validity and reliability. Cronbach α in all 3 samples of this study ranged from 0.90 to 0.93.

Psychoticism and Paranoid Ideation Subscales of the Symptom Checklist-90

The severity of psychoticism and paranoid ideation was assessed using the corresponding subscales of the Symptom Checklist-90. Participants respond to each of the 10 items on the psychoticism subscale or 6 items on the paranoid ideation subscale using a 5-point Likert scale ranging from 0 ("not at all") to 4 ("extremely"). The SCL-90 has demonstrated robust validity and reliability in Chinese populations [30]. These subscales were administered only in S1 (undergraduates) and S3 (inpatients). Cronbach α was 0.83 (S1) and 0.79 (S3) for psychoticism and 0.87 (S1) and 0.84 (S3) for paranoid ideation.

Adult ADHD Self-Report Scale

The ASRS, widely used clinically to screen adults for attention-deficit or hyperactivity disorder, is based on the 18 criteria in the "TR" revision of the 4th edition of the *Diagnostic and Statistical Manual of Mental Disorders* [31]. It examines symptoms over the past 6 months. Participants respond to 18 items on a Likert scale from 0 ("never") to 4 ("very often"), giving a total score from 0 to 72. Higher scores indicate greater ADHD symptom severity. The Chinese version has demonstrated both reliability and validity in young adults in Taiwan [32]. This survey was administered only in S2 (high school students) and S3 (inpatients). Cronbach α was 0.93 in S2 and 0.91 in S3.

Analysis

The data were analyzed using SPSS 27.0 (IBM) and R (version 4.4.3; R Core Team). Descriptive statistics, correlation analyses, independent-samples *t* tests, ANOVA, and post hoc comparisons were conducted in SPSS, whereas regression-related analyses, including the computation of confidence intervals for R^2 (not available in SPSS), were performed in R. The significance of univariate associations was assessed using, as appropriate, Pearson correlation coefficients (r), 2-tailed t tests, chi-square tests, or ANOVA. ANOVA results were adjusted using Tukey honest significant difference procedure to control for multiple comparisons. Potential relationships among variables were explored using multiple linear and logistic regression. To control for the confounding effects of demographics, we used standardized residuals (ZREs) of psychopathology scores adjusted for age and sex in regression and post hoc analyses; descriptive statistics were based on raw scores to facilitate comparability with prior studies. To evaluate whether multicollinearity among independent variables (eg, $r>0.7$ between the PHQ-9 and the GAD-7) affected the reliability of our regression models, we obtained tolerance and variance inflation factors before multivariate regression analyses. Where appropriate, associations were expressed as odds ratios (ORs) with corresponding 95% CIs. The results were considered statistically significant at a 2-tailed $P<.05$. The missing data in S2 were handled using complete-case analysis without multiple imputation after testing for missing completely at random, as

XSL•FO
RenderX

item-level missingness was minimal (0.2% - 3.0% per item). S1 and S3 used forced-response online questionnaires, resulting in no item-level missingness. The ASRS data in S3 were available only for 241 participants because the scale was introduced midway through the study. Accordingly, analyses involving the ASRS in S3 included only these 241 participants.

Ethical Considerations

The study procedures were carried out in accordance with the Declaration of Helsinki, and the study was conducted and reported in accordance with the Journal Article Reporting Standards [33]. The study was approved by the Ethics Committee of West China Hospital, Sichuan University (approval No. 2016 - 171), and the Ethics Committee of Hangzhou Seventh People's Hospital (approval No. 2023 - 064). All S1 participants provided electronic informed consent; for S2 and S3, both participants and their parents received detailed study information and provided written informed consent. All data were anonymized and stored securely on password-protected servers, and only aggregated results were reported. Participants did not receive any financial or material compensation for their participation in this study.

Results

Demographics

Male participants constituted a slight majority in S1 (n=8194, 54.5%), were slightly underrepresented in S2 (n=1720, 43.8%), and were substantially underrepresented in S3 (n=551, 26.0%). Mean ages (95% CI) in years were 18.1 (18.1-18.1) among undergraduates, 17.3 (17.0-17.6) among high school students, and 15.6 (15.4-15.7) among inpatients.

Measurement Scores of IGD and IA and Their Correlation With Demographics

The mean IGDS9-SF score (95% CI) in S1 was 11.8 (11.7-12.0), which was significantly lower than in S2 (15.1, 14.8-15.4), and S2 was significantly lower than S3 (18.8, 17.8-19.7). Similar trends were observed across samples for IA. The mean IAT-20 score (95% CI) in S1 was 31.7 (31.5-32.0), which was significantly lower than in S2 (32.4, 31.5-33.3), and S2 was significantly lower than S3 (47.7, 45.5-49.9; Table 1).

XSL•FO
RenderX

Table . Mean scores of measurements of internet gaming disorder (IGD), internet addiction (IA), and psychopathology, and the Pearson correlation matrix among demographic variables and these measures, across the 3 samples[a].

Measure	Scores, mean (95% CI)	Pearson correlation coefficients (*r*) among demographic variables and measures of IGD, IA, and psychopathology								
		Age	Gender	IAT-20[b]	IGDS9-SF[c]	PHQ-9[d]	GAD-7[e]	SCL90-psychoticism[f]	SCL90-paranoid ideation[g]	ASRS[h]
Sample 1 (n=8194)										
IGDS9-SF	11.8 (11.7-12.0)	0.02	–0.28[i]	0.51[i]	1.00	0.27[i]	0.24[i]	0.32[i]	0.27[i]	—[j]
IAT-20	31.7 (31.5-32.0)	–0.02[k]	0.08[i]	1.00	0.51[i]	0.50[i]	0.46[i]	0.52[i]	0.48[i]	—
PHQ-9	3.0 (2.9-3.1)	–0.00	0.12[i]	0.50[i]	0.27[i]	1.00	0.77[i]	0.63[i]	0.56[i]	—
GAD-7	2.0 (1.9-2.1)	–0.00	0.10[i]	0.46[i]	0.24[i]	0.77[i]	1.00	0.61[i]	0.56[i]	—
SCL90-psychoticism	1.2 (1.2-1.3)	–0.01	0.03[l]	0.52[i]	0.32[i]	0.63[i]	0.61[i]	1.00	0.80[i]	—
SCL90-paranoid ideation	1.3 (1.2-1.3)	–0.02	0.06[i]	0.48[i]	0.27[i]	0.56[i]	0.56[i]	0.80[i]	1.00	—
Sample 2 (n=1720)										
IGDS9-SF	15.1 (14.8-15.4)	–0.02	–0.24[i]	0.55[i]	1.00	0.35[i]	0.30[i]	—	—	0.41[i]
IAT-20	32.4 (31.5-33.3)	0.05[k]	0.09[i]	1.00	0.55[i]	0.54[i]	0.48[i]	—	—	0.62[i]
PHQ-9	6.3 (6.0-6.5)	0.03	0.11[i]	0.54[i]	0.35[i]	1.00	0.69[i]	—	—	0.55[i]
GAD-7	4.1 (3.9-4.2)	0.02	0.07[l]	0.48[i]	0.30[i]	0.69[i]	1.00	—	—	0.55[i]
ASRS	22.8 (22.3-23.3)	–0.01	0.05[k]	0.62[i]	0.41[i]	0.55[i]	0.55[i]	—	—	1.00
Sample 3										
IGDS9-SF (n=551)	18.8 (17.8-19.7)	–0.10[k]	–0.14[l]	0.57[i]	1.00	0.13	0.08	0.14[k]	0.16[k]	0.24[i]
IAT-20 (n=551)	47.7 (45.5-49.9)	–0.09[k]	0.03	1.00	0.57[i]	0.24[i]	0.16[k]	0.26[i]	0.18[l]	0.35[i]
PHQ-9 (n=551)	18.2 (17.4-19.1)	–0.16[i]	0.13[l]	0.24[i]	0.13	1.00	0.77[i]	0.67[i]	0.57[i]	0.57[i]
GAD-7 (n=551)	13.5 (12.9-14.1)	–0.14[l]	0.11[l]	0.16[k]	0.08	0.77[i]	1.00	0.69[i]	0.58[i]	0.53[i]
SCL90-psychoticism (n=551)	2.5 (2.3-2.6)	–0.18[i]	0.06	0.26[i]	0.14[k]	0.67[i]	0.69[i]	1.00	0.77[i]	0.67[i]
SCL90-paranoid ideation (n=551)	2.5 (2.3-2.7)	–0.17[i]	0.07	0.18[l]	0.16[k]	0.57[i]	0.58[i]	0.77[i]	1.00	0.59[i]
ASRS (n=241)[m]	54.0 (51.8-56.2)	–0.17[l]	0.12	0.35[i]	0.24[i]	0.57[i]	0.53[i]	0.67[i]	0.59[i]	1.00

[a]Sample 1 comprises undergraduate freshmen enrolled at Sichuan University. Sample 2 comprises students recruited from a high school in Hangzhou.

XSL•FO
RenderX

Sample 3 comprises inpatients aged 13 - 19 years recruited from 2 tertiary mental health centers in Hangzhou and Chengdu. The Internet Gaming Disorder Scale–Short Form, 20-item Young's Internet Addiction Test, Patient Health Questionnaire-9, and Generalized Anxiety Disorder-7 were administered in all 3 samples; the SCL-90 Psychoticism and Paranoid Ideation subscales were administered only in S1 and S3, and the Adult ADHD Self-Report Scale was administered only in S2 and S3.

[b]IAT-20: 20-item Young's Internet Addiction Test.

[c]IGDS9-SF: Internet Gaming Disorder Scale–Short Form.

[d]PHQ-9: 9-item Patient Health Questionnaire.

[e]GAD-7: 7-item Generalized Anxiety Disorder.

[f]SCL90-psychoticism: subscale of the Symptom Checklist-90 to measure psychoticism.

[g]SCL90–paranoid ideation: subscale of the Symptom Checklist-90 to measure paranoid ideation.

[h]ASRS: Adult ADHD Self-Report Scale.

[i]$P<.001$.

[j]Not available.

[k]$P<.05$.

[l]$P<.01$.

[m]The Adult ADHD Self-Report Scale data in sample 3 were available only for 241 participants because the scale was introduced midway through the study.

IGDS9-SF and IAT-20 scores showed either statistically significant but modest correlations with age or nonsignificant associations across all 3 samples (r=–0.10 to 0.10; Table 1). The scores of the IGDS9-SF were significantly higher in male than in female participants in all 3 samples (t_{8192}=26.2, $P<.001$ in S1; t_{1718}=9.92, $P<.001$ in S2; and t_{549}=3.18, P=.002 in S3). Conversely, the scores of the IAT-20 were significantly higher in female than in male participants in S1 (t_{8192}=–7.00; $P<.001$) and S2 (t_{1718}=–3.67; $P<.001$), whereas no significant sex differences were observed in S3 (t_{549}=0.61; P=.55).

Prevalence Rates of IGD and IA

The prevalence (95% CI) of IGD in S1, based on the predefined cutoff (IGDS9-SF ≥22), was 4.8% (4.3%-5.2%), which was significantly lower than in S2 (15.8%, 14.0%-17.5%), and in turn significantly lower than in S3 (32.3%, 28.4%-36.2%). The prevalence (95% CI) of IA in S1, based on the predefined cutoff (IAT-20 ≥50), was 7.3% (6.8%-7.9%), which was significantly lower than in S2 (18.8%, 17.0%-20.6%), and in turn was significantly lower than that in S3 (45.9%, 41.8%-50.1%).

In other words, in S1 (n=8194), most undergraduates (n=7408, 90.4%, 95% CI 89.8%-91.1%) had neither IGD nor IA, whereas small proportions (n=184, 2.24%, 95% CI 1.94%-2.59%) had IGD only, IA only (n=396, 4.83%, 95% CI 4.38%-5.32%), or both (n=206, 2.51%, 95% CI 2.19%-2.88%). In S2 (n=1720), a smaller majority (n=1287, 74.83%, 95% CI 72.75%-76.83%) of the high school students had neither disorder, with correspondingly higher prevalence of only IGD (n=109, 6.34%, 95% CI 5.27%-7.59%), only IA (n=163, 9.48%, 95% CI 8.17%-10.94%), and both (n=161, 9.36%, 95% CI 8.06%-10.83%). In contrast, in S3 (n=551), a slight majority of the inpatients had one or both disorders: only 253 (45.92%, 95% CI 41.71%-50.18%) had neither disorder, whereas 45 (8.17%, 95% CI 6.08%-10.86%) had IGD only, 120 (21.78%, 95% CI 18.45%-25.51%) had IA only, and 133 (24.14%, 95% CI 20.67%-27.98%) had both (Table 1).

Associations of Measurements and Prevalence Between IGD and IA

Univariate analyses indicated that IGDS9-SF and IAT-20 scores were moderately correlated in all 3 samples (r=0.51, $P<.001$ in S1; r=0.55, $P<.001$ in S2; and r=0.51, $P<.001$ in S3). These correlations remained significant after controlling for sex and age (Table 1). Similarly, multivariable analyses controlling for sex and age indicated that the odds of IA were significantly higher among participants with IGD in all 3 samples, with adjusted ORs (95% CI) of 29.6 (23.2-37.9) in S1, 14.1 (10.3-19.6) in S2, and 6.5 (4.3-9.9) in S3.

Scores of Psychopathological Symptoms

The mean PHQ-9 score (95% CI) in S1 was 3.0 (2.9-3.1), which was significantly lower than in S2 (6.3, CI 6.0-6.5) and in turn significantly lower than in S3 (18.2, 17.4-19.1). Similarly, the mean GAD-7 score (95% CI) in S1 was 2.0 (1.9-2.1), which was significantly lower than in S2 (4.1, 3.9-4.2), and in turn was significantly lower than in S3 (13.5, 12.9-14.1). The mean SCL-90 psychoticism subscale score (95% CI) in S1 was 1.2 (1.2-1.3), which was significantly lower than in S3 (2.5, 2.3-2.6). Similarly, the mean SCL-90 paranoid ideation subscale score (95% CI) in S1 was 1.3 (1.2-1.3), which was significantly lower than that in S3 (2.5, 2.3-2.7). The mean ASRS score (95% CI) in S2 was 22.8 (22.3-23.3), which was significantly lower than that in S3 (54.0, 51.8-56.2).

Psychopathology severity was not associated with age in S1 or S2 but showed small negative correlations with age in S3 (r=–0.09 to –0.18). All measured types of psychopathological symptoms in S1 and S2 were significantly more severe in female than in male participants, whereas only depression (PHQ-9) and anxiety (GAD-7) exhibited this sex difference in S3 (Table 1).

Associations of IGD or IA With Severity of Psychopathological Symptoms

To examine whether IGD or IA was associated with psychopathology, we performed pairwise comparisons between participants with and without IGD and between participants with and without IA; these analyses revealed statistically significant differences across all psychopathological variables

XSL•FO
RenderX

in each sample. Notably, participants with IA exhibited significantly higher dimensional psychopathology scores than those with IGD on the PHQ-9, GAD-7, and the SCL-90 paranoid ideation subscale in S1. A similar trend of differences was observed in S2 and S3, although the differences did not reach statistical significance (Table 2).

Table . Mean scores (95% CI) of measurements for internet gaming disorder (IGD), internet addiction (IA), and other psychopathologies among all participants, those with IGD and IA, and those without IGD or IA, across the 3 samples[a].

Sample and psychopathologies	Scores among participants with IGD[b], mean (95% CI)	Scores among participants with IA[c], mean (95% CI)	Scores among participants without IGD[d], mean (95% CI)	Scores among participants without IA[e], mean (95% CI)
Sample 1				
IGDS9-SF[f]	25.2 (24.7-25.7)	17.6 (17.0-18.2)	11.2 (11.1-11.2)	11.4 (11.3-11.5)
IAT-20[g]	50.8 (49.3-52.3)	58.2 (57.5-58.8)	30.8 (30.5-31.0)	29.6 (29.4-29.8)
PHQ-9[h]	6.0 (5.5-6.5)	7.2 (6.9-7.6)	2.8 (2.8-2.9)	2.7 (2.6-2.7)
GAD-7[i]	4.6 (4.1-5.1)	5.6 (5.3-6.0)	1.9 (1.9-2.0)	1.8 (1.7-1.8)
SCL90-psychoticism[j]	1.6 (1.5-1.7)	1.7 (1.6-1.7)	1.2 (1.2-1.2)	1.2 (1.2-1.2)
SCL90–paranoid ideation[k]	1.7 (1.6-1.7)	1.8 (1.7-1.8)	1.3 (1.3-1.3)	1.2 (1.2-1.3)
Sample 2				
IGDS9-SF	26.5 (25.9-27.0)	21.3 (20.4-22.2)	13.0 (12.8-13.2)	13.6 (13.4-13.8)
IAT-20	53.4 (51.3-55.4)	62.8 (61.8-63.9)	28.4 (27.5-29.4)	32.4 (31.9-32.8)
PHQ-9	9.8 (9.1-10.5)	10.6 (10.0-11.2)	5.7 (5.4-5.9)	5.2 (5.0-5.4)
GAD-7	6.8 (6.1-7.5)	7.4 (6.9-7.9)	3.5 (3.3-3.8)	3.2 (3.1-3.3)
ASRS[l]	31.9 (30.4-33.5)	33.9 (32.7-35.0)	21.0 (20.3-21.6)	20.0 (19.4-20.6)
Sample 3				
IGDS9-SF	30.0 (29.0-30.9)	23.1 (21.9-24.4)	13.5 (13.0-13.9)	15.1 (14.3-15.8)
IAT-20	60.9 (58.1-63.8)	67.2 (65.6-68.8)	41.4 (39.4-43.5)	31.2 (29.7-32.6)
PHQ-9	17.9 (17.0-18.9)	18.9 (17.8-19.9)	18.4 (17.7-19.1)	17.7 (16.9-18.5)
GAD-7	13.2 (12.4-14.1)	14.0 (13.3-14.7)	13.6 (13.0-14.3)	13.1 (12.4-13.8)
SCL90-psychoticism	2.6 (2.4-2.7)	2.7 (2.6-2.9)	2.4 (2.3-2.5)	2.2 (2.1-2.3)
SCL90–paranoid ideation	2.7 (2.5-2.9)	2.8 (2.7-2.9)	2.4 (2.3-2.5)	2.3 (2.1-2.4)
ASRS[m]	56.5 (54.0-59.0)	56.7 (54.9-58.5)	52.4 (50.3-54.4)	50.0 (46.3-53.7)

[a]Internet gaming disorder was defined as a total score ≥22 on the IDSG9-SF, and internet addiction was defined as a total score ≥50 on Young's 20-item Internet Addiction Test (IAT-20).

[b]Sample sizes for participants with internet gaming disorder: sample 1: n=393; sample 2: n=271; sample 3: n=178.

[c]Sample sizes for participants with internet addiction: sample 1: n=598; sample 2: n=323; sample 3: n=253.

[d]Sample sizes for participants without internet gaming disorder: sample 1: n=7801; sample 2: n=1,449; sample 3: n=373.

[e]Sample sizes for participants without internet addiction: sample 1: n=7596; sample 2: n=1397; sample 3: n=298.

[f]IDSG9-SF: Internet Gaming Disorder Scale–Short Form.

[g]IAT-20: 20-item Young's Internet Addiction Test.

[h]PHQ-9: 9-item Patient Health Questionnaire.

[i]GAD-7: 7-item Generalized Anxiety Disorder.

[j]SCL90-psychoticism: subscale of the Symptom Checklist-90 to measure psychoticism.

[k]SCL90–paranoid ideation: subscale of the Symptom Checklist-90 to measure paranoid ideation.

[l]ASRS: Adult ADHD Self-Report Scale.

[m]For the ASRS in sample 3, there were 88 participants with IGD, 10 with IA, 59 without IGD, and 84 without IA.

Next, we examined whether IGDS9-SF and IAT-20 scores were associated with psychopathological symptom scores. In these analyses, the independent variables were the ZREs of psychopathological symptom scores, adjusted for age and sex; the dependent variable was the ZRE of the IGDS9-SF or IAT-20 score. Univariate models indicated that both IGDS9-SF and

XSL•FO
RenderX

IAT-20 scores were significantly associated with all psychopathological symptom scores. These associations were significantly stronger for the IAT-20 than for the IGDS9-SF, as indicated by higher R^2 values, in S1 and S2. A similar trend was observed in S3, though most differences in associations did not reach statistical significance in this sample. Consistently, multivariate models, in which multicollinearity was acceptable (all tolerances >0.20, 95% CI 0.28‑0.53, and all variance inflation factors <5, 95% CI 1.55‑3.55), showed that psychopathological symptom scores were related to a larger proportion of changes in IAT-20 scores than IGDS9-SF scores in S1 (0.33, 95% CI 0.30-0.35 vs 0.13, 95% CI 0.11-0.16) and S2 (0.44, 95% CI 0.39-0.49 vs 0.23, 95% CI 0.18-0.27). A similar trend was observed in S3, although it did not reach statistical significance.

We also performed regression analyses in which IGD or IA status (present vs absent), defined using the prespecified IGDS9-SF and IAT-20 cutoff scores, served as the dependent variables, and the ZREs of psychopathological symptom scores served as the independent variables. In univariate models, both IGD and IA were significantly associated with all psychopathological symptom scores in S1 and S2, and with some of these symptom scores in S3. These associations were significantly stronger for IA than for IGD, as indicated by larger ORs, in S1 and S2. A similar trend was observed in S3, although most differences in ORs did not reach statistical significance in this sample. Multivariate analyses showed a similar pattern in which ORs were generally larger for IA than for IGD, particularly in S1 and S2, although most differences did not reach statistical significance.

Notably, in S2, IAT-20 scores showed a stronger univariate association with attention-deficit or hyperactivity symptom severity (ASRS scores) (R^2, 95% CI; 0.39, 0.33-0.43) than with depression (0.28, 0.23-0.32) or anxiety (0.22, 0.18-0.26). IGDS9-SF scores (R^2, 95% CI) also showed a stronger univariate association with attention deficit or hyperactivity symptom severity (0.19, 0.14-0.23) than with anxiety (0.10, 0.07-0.14). Similar trends were observed in S3, although they did not reach statistical significance. In the logistic regression analyses, univariate models showed that ASRS scores had the strongest associations with IGD and IA among all psychopathological measures in both S2 and S3, although differences in association strength, in terms of ORs, did not reach statistical significance. In multivariate models adjusting for the intercorrelation among other psychopathological measures, higher ASRS scores were independently associated with increased odds of IGD (OR, 95% CI; S2: 2.17, 1.80‑2.65; S3: 1.43, 0.99-2.09) and IA (S2: 3.18, 2.57‑3.96; S3: 1.75, 1.19-2.64). By contrast, most associations between other psychopathological symptoms and IGD or IA that were significant in univariate models were no longer statistically significant after multivariate adjustment (Table 3).

XSL•FO
RenderX

Table . Associations of Internet Gaming Disorder Scale–Short Form (IGDS9-SF), 20-item Young's Internet Addiction Test (IAT-20), and identified internet gaming disorder (IGD) and internet addiction (IA) with the severities of other psychopathologies across three samples[a].

Sample and psychopathology	IGDS9-SF[b]		IAT-20[b]		IGD		IA	
	R^2 (95% CI) in univariate model	R^2 (95% CI) in multivariate model[c]	R^2 (95% CI) in univariate model	R^2 (95% CI) in multivariate model[c]	OR[d] (95% CI)	Adjusted OR[c] (95% CI)	OR (95% CI)	Adjusted OR[c] (95% CI)
Sample 1		0.13 (0.11-0.16)		0.33 (0.30-0.35)				
PHQ-9[b,e]	0.10 (0.08-0.12)		0.25 (0.23-0.27)		1.89 (1.76-2.04)	1.33 (1.16-1.53)	2.50 (2.33-2.69)	1.50 (1.34-1.69)
GAD-7[b,f]	0.08 (0.06-0.10)		0.21 (0.19-0.23)		1.76 (1.64-1.89)	1.09 (0.95-1.23)	2.23 (2.09-2.39)	1.22 (1.09-*1.35)*
SCL90-psychoticism[b,g]	0.12 (0.10-0.14)		0.27 (0.25-0.30)		1.87 (1.75-2.00)	1.24 (1.08-1.42)	2.40 (2.25-2.58)	1.54 (1.36-1.74)
SCL90–paranoid ideation[b,h]	0.09 (0.07-0.11)		0.23 (0.21-0.25)		1.88 (1.75-2.01)	1.32 (1.16-1.51)	2.21 (2.07-2.36)	1.16 (1.03-1.30)
Sample 2		0.23 (0.18-0.27)		0.44 (0.39-0.49)				
PHQ-9[b]	0.15 (0.11-0.19)		0.28 (0.23-0.32)		2.20 (1.93-2.52)	1.56 (1.28-1.91)	3.08 (2.66-3.58)	1.89 (1.54-2.31)
GAD-7[b]	*0.10 (0.07-0.14)*		0.22 (0.18-0.26)		1.90 (1.68-2.15)	1.02 (0.84-1.23)	2.47 (2.17-2.83)	1.09 (0.90-1.31)
ASRS[b,i]	0.19 (0.14-0.23)		0.39 (0.33-0.43)		2.85 (2.43-3.38)	2.17 (1.80-2.65)	4.60 (3.80-5.64)	3.18 (2.57-3.96)
Sample 3		0.06 (0.03-0.16)		0.13 (0.06-0.26)				
PHQ-9[b]	0.00 (0.00-0.01)		0.01 (0.00-0.03)		0.93 (0.78-1.12)	0.92 (0.60-1.41)	1.16 (0.98-1.37)	1.21 (0.79-1.85)
GAD-7[b]	0.00 (0.00-0.00)		0.00 (0.00-0.02)		0.93 (0.78-1.11)	0.89 (0.57-1.40)	1.13 (0.95-1.34)	0.85 (0.54-1.33)
SCL90-psychoticism[b]	0.02 (0.00-0.05)		0.10 (0.06-0.16)		1.16 (0.97-1.40)	0.98 (0.59-1.63)	1.71 (1.43-2.06)	1.17 (0.69-1.99)
SCL90–paranoid ideation[b]	0.03 (0.01-0.07)		0.11 (0.06-0.16)		1.32 (1.09-1.59)	1.21 (0.78-1.87)	1.65 (1.37-1.99)	0.80 (0.50-1.24)
ASRS[b]	0.06 (0.01-0.14)		0.11 (0.02-0.22)		1.40 (1.07-1.85)	1.43 (0.99-2.09)	1.73 (1.30-2.36)	1.75 (1.19-2.64)

[a]IGD was defined based on the total score of the IGDS9-SF, with a cutoff value of ≥21. IA was defined based on the total score of IAT-20, with a cutoff value of ≥50. R^2 represents the square of the correlation coefficient (ie, *r*).

[b]Standardized residuals (ZREs) adjusted for the confounding effects of age and sex of measurement scores were used as dependent or independent variables in the regression models.

[c]These multivariate models included ZREs of other psychopathological symptom measures as independent variables.

[d]ORs: odds ratios.

[e]PHQ-9: 9-item Patient Health Questionnaire.

[f]GAD-7: 7-item Generalized Anxiety Disorder.

[g]SCL90-psychoticism: subscale of the Symptom Checklist-90 to measure psychoticism.

[h]SCL90–paranoid ideation: subscale of the Symptom Checklist-90 to measure paranoid ideation.

[i]ASRS: Adult ADHD Self-Report Scale. The ASRS data in S3 were available for only 241 participants because the scale was added midway through the study.

Finally, we classified participants into 4 groups according to IGD and IA status: both IGD and IA, neither condition, IGD only, or IA only. ANOVA tests indicated that ZREs of all measured psychopathological symptoms differed significantly

XSL•FO
RenderX

across the 4 groups in all 3 samples. Post hoc analyses in S1 revealed a significant stepwise increase in depressive, anxiety, psychoticism, and paranoid ideation symptom severity from participants with neither disorder to those with IGD only, then to those with IA only, and finally to those with both disorders. Similar trends were observed in S2 for the severity of depression, anxiety, and attention-deficit or hyperactivity, although the differences in depressive and anxiety symptom severity between the IA-only group and the comorbid group were not statistically significant. In S3, participants with IA (with or without comorbid IGD) tended to have higher psychopathological symptom scores than those with neither disorder nor IGD only, although some post hoc comparisons did not reach statistical significance (Figure 1).

Figure 1. Post hoc comparisons of psychopathological symptom severity (*z*-standardized residuals adjusted for age and sex) among participants with neither internet gaming disorder (IGD) nor internet addiction (IA; group A), IGD only (group B), IA only (group C), or comorbid IGD and IA (group D) across 3 independent samples: S1 (first-year undergraduates at Sichuan University), S2 (high school students recruited in Hangzhou), and S3 (inpatients aged 13 - 19 years recruited from 2 tertiary mental health centers in Hangzhou and Chengdu). ASRS: Adult ADHD Self-Report Scale; GAD-7: 7-item Generalized Anxiety Disorder; PHQ-9: 9-item Patient Health Questionnaire; SCL-90-paranoid: subscale of the Symptom Checklist-90 to measure paranoid ideation; SCL-90-psychotism: subscale of the Symptom Checklist-90 to measure psychoticism. ***P*<.01; ****P*<.001.

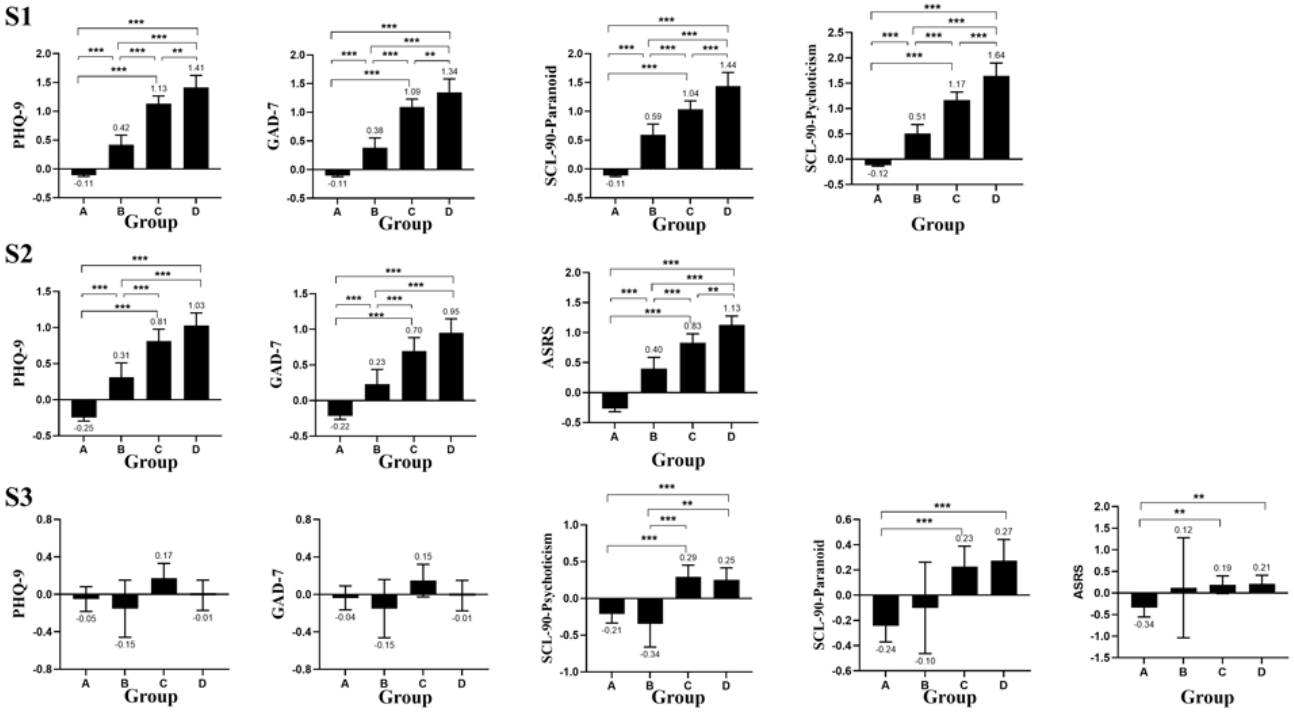

Discussion

Findings

To our knowledge, this is the first study to examine a potential relationship between IGD and IA, as well as to compare their associations with a broad range of psychopathological symptoms within the same samples. In addition, we cross-validated the main findings across 3 independent samples. Across samples, we found that IGD and IA were distinct yet moderately correlated and that IA was more prevalent and more strongly associated with psychopathological symptom severity.

These key findings have important implications for clarifying the nosological relationship between IGD and IA. The moderate association we detected between the 2 disorders was similar to that reported between so-called “problematic internet use” and “problematic online gaming” [20], but it was weaker than the associations between depressive and anxiety symptoms observed in this study and in previous work in other Chinese samples [34,35]. These observations support considering IGD and IA as distinct yet correlated entities, rather than as a single entity or as a subtype of a broader construct [21]. Consistent with this interpretation, we identified participants who met the most widely accepted definition of IGD but not IA, and vice versa. In our sample, IA, whether occurring alone or comorbid with IGD, was associated with more severe psychopathological symptoms than IGD alone. Furthermore, our univariate correlation and multivariate regression analyses indicated that IA severity was consistently more strongly associated with psychopathological symptom severity than IGD severity. These findings suggest that compulsive, problematic online behaviors extending beyond gaming warrant greater attention. Although “gaming” is currently the only type of online activity codified in the *International Classification of Diseases* and listed as a condition for further study in the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (*DSM-5*), other forms of online entertainment, such as social media and algorithm-driven short-video platforms, may pose comparable or even greater risks to mental health and may undermine psychological well-being, academic achievement, and family dynamics [36,37]. It may be more appropriate to refer to broader constructs such as “internet entertainment disorder” or “internet entertainment addiction” to better capture the spectrum of addictive behaviors related to internet use. These constructs may require further expansion or adaptation to reflect the growing use of immersive technologies, such as virtual or

XSL•FO
RenderX

augmented reality and artificial intelligence, which may further facilitate addictive behaviors.

In our study, both IGD and IA were more strongly associated with symptoms of attention-deficit or hyperactivity symptoms than with depressive or anxiety symptoms, consistent with findings from a study of adults aged 20 - 40 years in Taiwan [38]. These findings suggest that ADHD-related inattention and impulsivity, as indexed by ASRS scores, may partially account for the observed associations between IGD or IA and other psychopathological symptoms. Evidence from adolescents in several countries further indicates that attention-deficit or hyperactivity disorder frequently co-occurs with IGD and IA and may even predict their onset [39]. Core features of attention-deficit or hyperactivity, including impulsivity, inattention, heightened sensation seeking, and poor regulatory control, may increase the risk of problematic internet use [40-42]. More broadly, both disorders may be more strongly associated with externalizing symptoms characteristic of attention-deficit or hyperactivity than with internalizing symptoms, such as depression or anxiety, a pattern supported by recent meta-analytic evidence, although the pooled effect sizes were modest [43]. This closer association with externalizing symptoms may be related to the immediate gratification and heightened sensory stimulation inherent in gaming and other forms of online entertainment [44,45]. In contrast, internalizing symptoms may instead reflect maladaptive coping mechanisms for psychological distress [46,47].

We consider our data to be reliable because we assessed IGD and IA using widely validated, commonly used instruments and because we were able to replicate key associations across the 3 independent samples of adolescents. In addition, the differences in psychopathological symptoms observed across samples were consistent with expectations: hospitalized adolescents exhibited high levels of psychopathological symptoms similar to those reported in a study of Caucasian adolescents [48], and these levels were higher than those in high school students, which in turn were higher than those observed in undergraduates. Our high school students may have experienced heightened stress due to the upcoming national university entrance examination (gaokao), which may help explain their higher levels of depression and anxiety [49]. By contrast, our first-year undergraduates had already performed sufficiently well on the entrance examination to secure admission to a top-tier university. This may have contributed to their lower levels of depressive and anxiety symptoms. More generally, students admitted to top-tier universities in China may possess more effective coping strategies and problem-solving skills, which could buffer against psychopathological symptoms [50].

In our samples, male sex appeared to be associated with a higher risk of internet gaming disorder but a lower risk of IA, consistent with previous work in Chinese and US populations in which the 2 disorders were conceptualized as separate constructs [51,52]. Future research should examine whether and through what mechanisms sex influences the risk of either disorder, particularly given well-documented sex differences in internet use: male participants tend to engage more in computer gaming, whereas female participants tend to engage more in social networking and social media [53]. Previous studies have reported inconsistent findings regarding whether sex influences the risk of IGD or IA [54,55]. These inconsistencies may reflect differences in the relative proportions of individuals with IGD only, IA only, or comorbid IGD and IA.

Limitations

Our findings should be interpreted with caution in light of several limitations. First, the cross-sectional design of this study precludes causal inference. Second, all data were obtained from self-report questionnaires, which may increase the risk of social desirability and recall biases. Third, the sample comprised exclusively Chinese participants residing in China, which may limit the generalizability of our findings to other cultural and geographic contexts. In addition, the inpatient sample size was relatively small, and many participants scored near the maximum possible values on several symptom scales, a phenomenon known as the ceiling effect. These factors may have reduced statistical power and, together with other sample-specific characteristics (eg, differences in sex ratios across samples), may help explain why several associations observed in the university and high school samples did not reach statistical significance in the inpatient sample. Nevertheless, most analyses conducted among inpatients showed patterns similar to those observed in the other 2 samples.

It should also be noted that our findings reflect IGD as a global construct without differentiating between specific game genres (eg, real-time strategy, massively multiplayer online role-playing games, sports games, or first-person shooters). Given evidence that the prevalence and psychological correlates of IGD may vary by game genre [56,57], future studies should systematically characterize predominant game types and examine genre-specific associations with psychopathology. In addition, the measures of psychopathological symptoms included in the 3 samples were not entirely consistent across the 3 samples (eg, the absence of SCL-90 in S2 and ASRS in S1), due to considerations such as survey timing constraints and primary study objectives. This inconsistency may have reduced the comparability of certain results (eg, associations involving ASRS) across samples. However, the key findings—namely, that IGD and IA are distinct yet moderately correlated constructs, with IA more strongly associated with the severity of psychopathological symptoms—were robust and consistent across all 3 samples.

Conclusions

This study provides additional evidence that IGD and IA represent distinct yet interrelated constructs and further demonstrates that IA consistently exhibits a stronger association with the severity of psychopathological symptoms than IGD. These findings underscore the importance of recognizing and addressing compulsive and problematic online behaviors that extend beyond gaming, contributing to ongoing debates regarding the classification and clinical significance of behavioral addictions related to internet use, and highlighting the need for further refinement of diagnostic frameworks and the prioritization of targeted, evidence-based clinical interventions.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank all the patients who participated in this study. The authors declare the use of generative artificial intelligence (GAI) in the writing process. According to GAIDeT (Generative AI Delegation Taxonomy; 2025), translation was delegated to GAI tools under full human supervision. The GAI tool used was ChatGPT (OpenAI, GPT-4.5). Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was supported by the National Natural Science Foundation of China (grant No. 82171487), “Pioneer” and “Leading Goose” R&D Program of Zhejiang (2024C03006), and the Leading Innovation and Entrepreneurship Team of Hangzhou (TD2024003). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

YYL and WJG conceived and designed the study and drafted the manuscript. YYL, AQH, and LLY curated and analyzed the data. AQH and WW performed data visualization, YYL, LLY, AQH, JW, and ZXM collected the data. QYL, WJD, MXQ, and JJX validated the results. YQH, SH, FGL, and XJL provided resources. JJX, QW, XJL, TL, and WJG supervised the project. TL and WJG acquired funding and administered the project. QW, WD, YZH, TL, and WJG reviewed and edited the manuscript. All authors have full access to the study dataset and are responsible for ensuring its integrity and the accuracy of the resulting analyses.

Conflicts of Interest

None declared.

References

1. Statistics. International Telecommunication Union. 2023. URL: https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx [accessed 2026-01-20]
2. León Méndez M, Padrón I, Fumero A, Marrero RJ. Effects of internet and smartphone addiction on cognitive control in adolescents and young adults: a systematic review of fMRI studies. Neurosci Biobehav Rev 2024 Apr;159:105572. [doi: 10.1016/j.neubiorev.2024.105572] [Medline: 38320657]
3. Bersani FS, Barchielli B, Ferracuti S, et al. The association of problematic use of social media and online videogames with aggression is mediated by insomnia severity: a cross-sectional study in a sample of 18- to 24-year-old individuals. Aggress Behav 2022 May;48(3):348-355. [doi: 10.1002/ab.22008] [Medline: 34870339]
4. Kim HS, Son G, Roh EB, et al. Prevalence of gaming disorder: a meta-analysis. Addict Behav 2022 Mar;126:107183. [doi: 10.1016/j.addbeh.2021.107183] [Medline: 34864436]
5. Soriano-Molina E, Limiñana-Gras RM, Patró-Hernández RM, Rubio-Aparicio M. The association between internet addiction and adolescents’ mental health: a meta-analytic review. Behav Sci (Basel) 2025 Jan 23;15(2):116. [doi: 10.3390/bs15020116] [Medline: 40001747]
6. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5): American Psychiatric Association; 2013. [doi: 10.1176/appi.books.9780890425596]
7. ICD-11: International Classification of Diseases 11th Revision. 2022. URL: https://icd.who.int/en/ [accessed 2026-01-20]
8. King DL, Chamberlain SR, Carragher N, et al. Screening and assessment tools for gaming disorder: a comprehensive systematic review. Clin Psychol Rev 2020 Apr;77:101831. [doi: 10.1016/j.cpr.2020.101831] [Medline: 32143109]
9. Rahman M, Sarkar P, Haque SE, et al. Technology addiction: effects of electronic games and social media use on academic performance and symptoms of psychiatric disorders among school-age adolescents. Health Sci Rep 2025 Jul;8(7):e71045. [doi: 10.1002/hsr2.71045] [Medline: 40687541]
10. Severo RB, Soares JM, Affonso JP, et al. Prevalence and risk factors for internet gaming disorder. Braz J Psychiatry 2020;42(5):532-535. [doi: 10.1590/1516-4446-2019-0760] [Medline: 32785455]
11. Wartberg L, Kriston L, Kramer M, Schwedler A, Lincoln TM, Kammerl R. Internet gaming disorder in early adolescence: associations with parental and adolescent mental health. Eur Psychiatry 2017 Jun;43:14-18. [doi: 10.1016/j.eurpsy.2016.12.013] [Medline: 28365463]
12. Wang CY, Wu YC, Su CH, Lin PC, Ko CH, Yen JY. Association between Internet gaming disorder and generalized anxiety disorder. J Behav Addict 2017 Dec 1;6(4):564-571. [doi: 10.1556/2006.6.2017.088] [Medline: 29280398]

XSL•FO
RenderX

13. Falcione K, Weber R. Psychopathology and gaming disorder in adolescents. JAMA Netw Open 2025 Jul 1;8(7):e2528532. [doi: 10.1001/jamanetworkopen.2025.28532] [Medline: 40728787]
14. Petry NM, Zajac K, Ginley MK. Behavioral addictions as mental disorders: to be or not to be? Annu Rev Clin Psychol 2018 May 7;14:399-423. [doi: 10.1146/annurev-clinpsy-032816-045120] [Medline: 29734827]
15. Young KS. Caught in the Net: How to Recognize the Signs of Internet Addiction—and a Winning Strategy for Recovery: John Wiley & Sons; 1998. URL: https://www.wiley.com/en-us/Caught+in+the+Net%3A+How+to+Recognize+the+Signs+of+Internet+Addiction--and+a+Winning+Strategy+for+Recovery-p-9780471191599 [accessed 2026-01-20]
16. Meng SQ, Cheng JL, Li YY, et al. Global prevalence of digital addiction in general population: a systematic review and meta-analysis. Clin Psychol Rev 2022 Mar;92:102128. [doi: 10.1016/j.cpr.2022.102128] [Medline: 35150965]
17. Guo W, Tao Y, Li X, et al. Associations of internet addiction severity with psychopathology, serious mental illness, and suicidality: large-sample cross-sectional study. J Med Internet Res 2020 Aug 11;22(8):e17560. [doi: 10.2196/17560] [Medline: 32780029]
18. Kitazawa M, Yoshimura M, Murata M, et al. Associations between problematic internet use and psychiatric symptoms among university students in Japan. Psychiatry Clin Neurosci 2018 Jul;72(7):531-539. [doi: 10.1111/pcn.12662] [Medline: 29652105]
19. Thorens G, Achab S, Billieux J, et al. Characteristics and treatment response of self-identified problematic internet users in a behavioral addiction outpatient clinic. J Behav Addict 2014 Mar;3(1):78-81. [doi: 10.1556/JBA.3.2014.008] [Medline: 25215217]
20. Király O, Griffiths MD, Urbán R, et al. Problematic internet use and problematic online gaming are not the same: findings from a large nationally representative adolescent sample. Cyberpsychol Behav Soc Netw 2014 Dec;17(12):749-754. [doi: 10.1089/cyber.2014.0475] [Medline: 25415659]
21. Young KS. Psychology of computer use: XL. Addictive use of the Internet: a case that breaks the stereotype. Psychol Rep 1996 Dec;79(3 Pt 1):899-902. [doi: 10.2466/pr0.1996.79.3.899] [Medline: 8969098]
22. Moreno M, Riddle K, Jenkins MC, Singh AP, Zhao Q, Eickhoff J. Measuring problematic internet use, internet gaming disorder, and social media addiction in young adults: cross-sectional survey study. JMIR Public Health Surveill 2022 Jan 27;8(1):e27719. [doi: 10.2196/27719] [Medline: 34081596]
23. Sánchez-Fernández M, Borda-Mas M, Horvath Z, Demetrovics Z. Similarities and differences in the psychological factors associated with generalised problematic internet use, problematic social media use, and problematic online gaming. Compr Psychiatry 2024 Oct;134:152512. [doi: 10.1016/j.comppsych.2024.152512] [Medline: 38955108]
24. Severo RB, Barbosa APPN, Fouchy DRC, et al. Development and psychometric validation of Internet Gaming Disorder Scale-Short-Form (IGDS9-SF) in a Brazilian sample. Addict Behav 2020 Apr;103:106191. [doi: 10.1016/j.addbeh.2019.106191] [Medline: 31887719]
25. Young KS, de Abreu CN, editors. Internet Addiction: A Handbook and Guide to Evaluation and Treatment: John Wiley & Sons P&T; 2011. URL: https://www.vitalsource.com/products/internet-addiction-v9780470892244?srsltid=AfmBOor1sLx0xQsKzvwxfzRwAyYkPEBTF3lqPhLP6JTMTZvhFugoBrS_ [accessed 2026-01-20]
26. Durkee T, Kaess M, Carli V, et al. Prevalence of pathological internet use among adolescents in Europe: demographic and social factors. Addiction 2012 Dec;107(12):2210-2222. [doi: 10.1111/j.1360-0443.2012.03946.x] [Medline: 22621402]
27. Endomba FT, Demina A, Meille V, et al. Prevalence of internet addiction in Africa: a systematic review and meta-analysis. J Behav Addict 2022 Sep 26;11(3):739-753. [doi: 10.1556/2006.2022.00052] [Medline: 35984734]
28. Yu X, Tam WWS, Wong PTK, Lam TH, Stewart SM. The Patient Health Questionnaire-9 for measuring depressive symptoms among the general population in Hong Kong. Compr Psychiatry 2012 Jan;53(1):95-102. [doi: 10.1016/j.comppsych.2010.11.002] [Medline: 21193179]
29. Löwe B, Decker O, Müller S, et al. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. Med Care 2008 Mar;46(3):266-274. [doi: 10.1097/MLR.0b013e318160d093] [Medline: 18388841]
30. Preti A, Carta MG, Petretto DR. Factor structure models of the SCL-90-R: replicability across community samples of adolescents. Psychiatry Res 2019 Feb;272:491-498. [doi: 10.1016/j.psychres.2018.12.146] [Medline: 30611969]
31. Diagnostic and Statistical Manual of Mental Disorders, 4th Edition: American Psychiatric Association; 2022. [doi: 10.1176/appi.books.9780890425787]
32. Yeh CB, Gau SSF, Kessler RC, Wu YY. Psychometric properties of the Chinese version of the adult ADHD Self-report Scale. Int J Methods Psychiatr Res 2008;17(1):45-54. [doi: 10.1002/mpr.241] [Medline: 18286465]
33. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
34. Xie Y, Tang L. The symptom network of internet gaming addiction, depression, and anxiety among children and adolescents. Sci Rep 2024 Nov 29;14(1):29732. [doi: 10.1038/s41598-024-81094-7] [Medline: 39614079]

XSL•FO
RenderX

35. Teng Z, Pontes HM, Nie Q, Griffiths MD, Guo C. Depression and anxiety symptoms associated with internet gaming disorder before and during the COVID-19 pandemic: a longitudinal study. J Behav Addict 2021 Mar 10;10(1):169-180. [doi: 10.1556/2006.2021.00016] [Medline: 33704085]
36. Chao M, Lei J, He R, Jiang Y, Yang H. TikTok use and psychosocial factors among adolescents: comparisons of non-users, moderate users, and addictive users. Psychiatry Res 2023 Jul;325:115247. [doi: 10.1016/j.psychres.2023.115247] [Medline: 37167877]
37. Feng T, Wang B, Mi M, et al. The relationships between mental health and social media addiction, and between academic burnout and social media addiction among Chinese college students: a network analysis. Heliyon 2025 Feb 15;11(3):e41869. [doi: 10.1016/j.heliyon.2025.e41869] [Medline: 39959490]
38. Yen JY, Király O, Griffiths MD, Demetrovics Z, Ko CH. A case-control study for psychiatric comorbidity and associative factors of gaming disorder and hazardous gaming based on ICD-11 criteria: cognitive control, emotion regulation, and reinforcement sensitivity. J Behav Addict 2024 Dec 30;13(4):1014-1027. [doi: 10.1556/2006.2024.00066] [Medline: 39636323]
39. Peeters M, Koning I, van den Eijnden R. Predicting internet gaming disorder symptoms in young adolescents: a one-year follow-up study. Comput Human Behav 2018 Mar;80:255-261. [doi: 10.1016/j.chb.2017.11.008]
40. Dullur P, Krishnan V, Diaz AM. A systematic review on the intersection of attention-deficit hyperactivity disorder and gaming disorder. J Psychiatr Res 2021 Jan;133:212-222. [doi: 10.1016/j.jpsychires.2020.12.026] [Medline: 33360866]
41. Wang BQ, Yao NQ, Zhou X, Liu J, Lv ZT. The association between attention deficit/hyperactivity disorder and internet addiction: a systematic review and meta-analysis. BMC Psychiatry 2017 Jul 19;17(1):260. [doi: 10.1186/s12888-017-1408-x] [Medline: 28724403]
42. Koronczai B, Kökönyei G, Griffiths MD, Demetrovics Z. The relationship between personality traits, psychopathological symptoms, and problematic internet use: a complex mediation model. J Med Internet Res 2019 Apr 26;21(4):e11837. [doi: 10.2196/11837] [Medline: 31025955]
43. Eirich R, McArthur BA, Anhorn C, McGuinness C, Christakis DA, Madigan S. Association of screen time with internalizing and externalizing behavior problems in children 12 years or younger: a systematic review and meta-analysis. JAMA Psychiatry 2022 May 1;79(5):393-405. [doi: 10.1001/jamapsychiatry.2022.0155] [Medline: 35293954]
44. Wang L, Wu L, Lin X, et al. Dysfunctional default mode network and executive control network in people with Internet gaming disorder: independent component analysis under a probability discounting task. Eur Psychiatry 2016 Apr;34:36-42. [doi: 10.1016/j.eurpsy.2016.01.2424] [Medline: 26928344]
45. Ko CH, Wang PW, Liu TL, Chen CS, Yen CF, Yen JY. The adaptive decision-making, risky decision, and decision-making style of Internet gaming disorder. Eur Psychiatry 2017 Jul;44:189-197. [doi: 10.1016/j.eurpsy.2017.05.020] [Medline: 28646731]
46. Kardefelt-Winther D. A conceptual and methodological critique of internet addiction research: towards a model of compensatory internet use. Comput Human Behav 2014 Feb;31:351-354. [doi: 10.1016/j.chb.2013.10.059]
47. Miao S, Xu L, Gao S, Bai C, Huang Y, Peng B. The association between anxiety and internet addiction among left-behind secondary school students: the moderating effect of social support and family types. BMC Psychiatry 2024 May 30;24(1):406. [doi: 10.1186/s12888-024-05855-4] [Medline: 38811914]
48. Torres-Rodríguez A, Griffiths MD, Carbonell X, Oberst U. Internet gaming disorder in adolescence: psychological characteristics of a clinical sample. J Behav Addict 2018 Sep 1;7(3):707-718. [doi: 10.1556/2006.7.2018.75] [Medline: 30264606]
49. Zhou J, Liu Y, Ma J, et al. Prevalence of depressive symptoms among children and adolescents in China: a systematic review and meta-analysis. Child Adolesc Psychiatry Ment Health 2024 Nov 19;18(1):150. [doi: 10.1186/s13034-024-00841-w] [Medline: 39563377]
50. Zhang W, Gao W, Liu X. Does attending elite colleges matter in the relationship between self-esteem and general self-efficacy of students in China? Heliyon 2022 Jun;8(6):e09723. [doi: 10.1016/j.heliyon.2022.e09723] [Medline: 35756109]
51. Peng C, Guo T, Cheng J, et al. Sex differences in association between Internet addiction and aggression among adolescents aged 12 to 18 in mainland of China. J Affect Disord 2022 Sep 1;312:198-207. [doi: 10.1016/j.jad.2022.06.026] [Medline: 35728679]
52. Ohayon MM, Roberts L. Internet gaming disorder and comorbidities among campus-dwelling U.S. university students. Psychiatry Res 2021 Aug;302:114043. [doi: 10.1016/j.psychres.2021.114043] [Medline: 34129998]
53. Chen B, Liu F, Ding S, Ying X, Wang L, Wen Y. Gender differences in factors associated with smartphone addiction: a cross-sectional study among medical college students. BMC Psychiatry 2017 Oct 10;17(1):341. [doi: 10.1186/s12888-017-1503-z] [Medline: 29017482]
54. Zhuang X, Zhang Y, Tang X, Ng TK, Lin J, Yang X. Longitudinal modifiable risk and protective factors of internet gaming disorder: a systematic review and meta-analysis. J Behav Addict 2023 Jun 29;12(2):375-392. [doi: 10.1556/2006.2023.00017] [Medline: 37224007]
55. Kiviruusu O. Excessive internet use among Finnish young people between 2017 and 2021 and the effect of COVID-19. Soc Psychiatry Psychiatr Epidemiol 2024 Dec;59(12):2291-2301. [doi: 10.1007/s00127-024-02723-0] [Medline: 38985326]

XSL•FO
RenderX

56. Na E, Choi I, Lee TH, et al. The influence of game genre on Internet gaming disorder. J Behav Addict 2017 Jun 29;6(2):1-8. [doi: 10.1556/2006.6.2017.033] [Medline: 28658960]
57. Kim D, Nam JK, Keum C. Adolescent Internet gaming addiction and personality characteristics by game genre. PLoS ONE 2022;17(2):e0263645. [doi: 10.1371/journal.pone.0263645]

Abbreviations

ASRS: Adult ADHD Self-Report Scale
***DSM-5*:** *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*
GAD-7: 7-item Generalized Anxiety Disorder
IA: internet addiction
IAT-20: 20-item Young's Internet Addiction Test
***ICD-11*:** *International Classification of Diseases, 11th Revision*
IGD: internet gaming disorder
IGDS9-SF: Internet Gaming Disorder Scale–Short Form
OR: odds ratio
PHQ-9: 9-item Patient Health Questionnaire
S1: sample 1
S2: sample 2
S3: sample 3
ZRE: standardized residual

Edited by S Brini; submitted 14.Aug.2025; peer-reviewed by KW Tay, X Xiang, Y Feng; accepted 07.Jan.2026; published 03.Feb.2026.

Please cite as:
Li YY, Hu AQ, Yi LL, Mao ZX, Lü QY, Wang J, Wei W, Huang YQ, Huang S, Dai WJ, Qiao MX, Xu JJ, Wang Q, Li XJ, Luo FG, Deng W, Hu YZ, Li T, Guo WJ
Comparing the Associations of Internet Addiction and Internet Gaming Disorder With Psychopathological Symptoms: Cross-Sectional Study of Three Independent Adolescent Samples
J Med Internet Res 2026;28:e82414
URL: https://www.jmir.org/2026/1/e82414
doi:10.2196/82414

© Ying-ying Li, A-qian Hu, Ling-li Yi, Zi-xin Mao, Qiu-yue Lü, Juan Wang, Wei Wei, Yue-qi Huang, Shu Huang, Wen-jing Dai, Meng-xuan Qiao, Jia-jun Xu, Qiang Wang, Xiao-jing Li, Fu-gang Luo, Wei Deng, Yu-zheng Hu, Tao Li, Wan-jun Guo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 3.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Phenotyping for Adolescent Mental Health: Feasibility Study Using Machine Learning to Predict Mental Health Risk From Active and Passive Smartphone Data

Balasundaram Kadirvelu[1*], PhD; Teresa Bellido Bel[2*], MD; Aglaia Freccero[2], MSc; Martina Di Simplico[2], PhD, MD; Dasha Nicholls[2], MD; A Aldo Faisal[1,3], PhD

[1]Brain & Behaviour Lab, Department of Computing and Department of Bioengineering, Imperial College London, Royal School of Mines, London, United Kingdom

[2]Division of Psychiatry, Department of Brain Sciences, Imperial College London, London, England, United Kingdom

[3]Chair in Digital Health & Data Science, Faculty of Life Sciences, University of Bayreuth, Bayreuth, Bavaria, Germany

[*]these authors contributed equally

Corresponding Author:
A Aldo Faisal, PhD
Brain & Behaviour Lab, Department of Computing and Department of Bioengineering, Imperial College London, Royal School of Mines, London, United Kingdom

Abstract

Background: Adolescents are particularly vulnerable to mental disorders, with over 75% of lifetime cases emerging before the age of 25 years. Yet most young people with significant symptoms do not seek support. Digital phenotyping, leveraging active (self-reported) and passive (sensor-based) data from smartphones, offers a scalable, low-burden approach for early risk detection. Despite this potential, its application in school-going adolescents from general (nonclinical) populations remains limited, leaving a critical gap in community-based prevention efforts.

Objective: This study evaluated the feasibility of using a smartphone app to predict mental health risks in nonclinical adolescents by integrating active and passive data streams within a machine learning (ML) framework. We examined the utility of this approach for identifying risks related to internalizing and externalizing difficulties, eating disorders, insomnia, and suicidal ideation.

Methods: Participants (n=103; mean age 16.1 years, SD 1.0) from 3 UK secondary schools used the Mindcraft app (Brain and Behaviour Lab) for 14 days, providing daily self-reports (eg, mood, sleep, and loneliness) and continuous passive sensor data (eg, location, step count, and app usage). We developed a deep learning model incorporating contrastive pretraining with triplet margin loss to stabilize user-specific behavioral patterns, followed by supervised fine-tuning for binary classification of 4 mental health outcomes, namely, the Strengths and Difficulties Questionnaire (SDQ)-high risk, insomnia, suicidal ideation, and eating disorder. Performance was assessed using leave-one-subject-out cross-validation (LOSO-CV), with balanced accuracy as the primary metric. Comparative analyses were conducted using CatBoost (Yandex) and multilayer perceptron (MLP) models without pretraining. Feature importance was assessed using Shapley Additive Explanations (SHAP) values, and associations between key digital features and clinical scales were analyzed.

Results: Integration of active and passive data outperformed single-modality models, achieving mean balanced accuracies of 0.71 (0.03) for SDQ-high risk, 0.67 (0.04) for insomnia, 0.77 (0.03) for suicidal ideation, and 0.70 (0.03) for eating disorder. The contrastive learning approach improved representation stability and predictive robustness. SHAP analysis highlighted clinically relevant features, such as negative thinking and location entropy, underscoring the complementary value of combining subjective and objective data. Correlation analyses confirmed meaningful associations between key digital features and mental health outcomes. Performance in an independent external validation cohort (n=45) achieved balanced accuracies of 0.63 - 0.72 across outcomes, suggesting generalizability to new settings.

Conclusions: This study demonstrates the feasibility and utility of smartphone-based digital phenotyping for predicting mental health risks in nonclinical, school-going adolescents. By integrating active and passive data with advanced machine modeling techniques, this approach shows promise for early detection and scalable intervention strategies in community settings.

(J Med Internet Res 2026;28:e72501) doi:10.2196/72501

XSL•FO
RenderX

KEYWORDS

youth mental health; digital health; mobile applications; mobile health; mHealth; smartphone sensing; ecological momentary assessment; EMA; artificial intelligence; early intervention

Introduction

Children and young people are particularly vulnerable to mental health problems due to critical developmental changes in emotion, behavior, and cognition [1], with over 75% of mental disorders emerging before the age of 25 years [2]. The World Health Organization estimates that one in 7 adolescents aged 10 - 19 years at the time of this writing lives with a diagnosable mental disorder [3].

Adolescent mental health symptoms are typically classified as internalizing (eg, anxiety, depression, and suicidal thoughts) or externalizing (eg, aggression and impulsivity) [4,5]. These symptoms can impair social, academic, and interpersonal functioning, and if left unaddressed, may lead to long-term psychiatric disorders [6]. Anxiety, depression, and eating disorders are among the most prevalent conditions, with sleep problems often acting as both a symptom and a risk factor [7,8]. Among the most severe manifestations of internalizing psychopathology is suicidal ideation, a particularly serious concern during adolescence; when unaddressed, it may progress to suicide, which ranks as the third leading cause of death globally among individuals aged 15-29 years [3]. Yet only 18% - 34% of young people with significant symptoms seek professional help [9]. This critical gap in receiving care underscores an urgent need for scalable, accessible, and youth-friendly solutions in mental health care.

Given the central role of schools in adolescents' daily lives, digital health interventions in schools offer a promising, cost-effective, and scalable way to support mental well-being in children and young people [10,11]. The proliferation of smartphones has enabled a new class of digital mental health interventions that leverage their unique data collection capabilities [12-14]. Smartphones can gather active data (subjective self-reports of behaviors and experiences) and passive data (sensor-based measures such as GPS, step count, and ambient light). These behavioral markers, collectively termed "digital phenotyping," reveal dynamic interactions between individuals and their environments [15-17].

Active data capture internal states such as mood, sleep quality, or loneliness, but it depends on user engagement and is vulnerable to recall bias. In contrast, passive data offer continuous, objective insights into daily routines, capturing behavioral patterns that may reflect underlying mental health states. For instance, lower location entropy and reduced movement are associated with depression [18], fewer steps with lower mood [19,20], and variations in ambient light with circadian disruption [21]. However, passive data alone may miss crucial psychological context and can be challenging to interpret in isolation [22].

Although several studies [23-41] have explored either active or passive data for mental health monitoring, few have examined their integration, particularly in community-based adolescent populations. Combining these modalities provides a more comprehensive view of mental health, connecting how young people feel with how they behave in real-world settings [42]. This fusion can enhance model robustness, uncover hidden patterns, and address mismatches between self-report and behavior. For example, a young person may report feeling fine (active data) while showing signs of social withdrawal or sleep disruption (passive data), a critical challenge for unimodal approaches. Multimodal approaches are better suited to detecting such discrepancies, enabling earlier and more nuanced detection in nonclinical settings [42,43].

The complexity of digital phenotyping data, such as its high dimensionality, multimodal nature, and nonlinear interaction patterns, poses challenges for traditional statistical approaches, which rely on strong parametric assumptions and may require extensive a priori feature engineering. In contrast, machine learning (ML) can model complex, heterogeneous data and automatically learn nonlinear latent patterns without predefined hypotheses [44,45]. This makes ML particularly well-suited to adolescent mental health prediction using digital data, where subtle behavioral signals may be distributed across multiple features and modalities [17].

Despite growing interest in applying ML to digital mental health [23-41], existing studies often present one or more limitations in the context of adolescent-focused research: (1) they are primarily conducted in adult populations, limiting relevance to younger age groups such as adolescents; (2) they typically involve clinical samples, reducing generalizability to community-based or non–help-seeking populations; (3) they focus on a single condition (eg, depression or anxiety); and (4) they rely on either active or passive data, but not both. These limitations constrain the generalizability, robustness, and real-world utility of current approaches for children and young people.

This study advances the field in several ways. First, it uses a nonclinical adolescent sample, enabling evaluation in a real-world prevention context, which is critical for early detection and scalable screening. Second, it investigates 4 distinct mental health outcomes (internalizing and externalizing difficulties, insomnia, eating disorder risk, and suicidal ideation) rather than focusing on a single condition. Third, it integrates both subjective (active) and objective (passive) smartphone data to develop multimodal predictive models that capture patterns that may be missed when using either modality alone. Finally, it uses a contrastive learning framework, a novel ML technique for learning stable behavioral representations, enhancing generalizability.

Specifically, this study addresses the following research questions: (1) Can integrating active and passive smartphone data improve the prediction of mental health risks in a nonclinical adolescent population using ML? (2) How well does this multimodal approach generalize across different mental health outcomes? (3) Which digital features are most relevant to predicting specific mental health outcomes?

XSL•FO
RenderX

Methods

Recruitment and Data Collection

Participants were recruited from secondary schools in northwest London between November 2022 and July 2023. We contacted schools via email and followed up via telephone. Three schools that expressed interest in taking part in our study were recruited. The inclusion criteria were young people aged 14 - 18 years attending 10-13 years with a sufficient level of English to respond to the study instrument and use the app and who had access to an iOS- or Android-compatible smartphone.

Students initially completed an online survey accessed via a Qualtrics link included in the promotional materials. This survey began with the Strengths and Difficulties Questionnaire (SDQ) [46], a screening tool whose predictions have been largely consistent with clinical diagnoses with good levels of internal consistency and test-retest stability. To detect eating disorders, we included the Eating Disorder-15 Questionnaire (ED - 15), which has been described as a valuable tool to assess eating disorder psychopathology in young individuals quickly [47]. We excluded the compensatory behaviors section to simplify the data collection process. Its ability to detect changes early in treatment means that it could be used as a routine outcome measure within therapeutic contexts. We also incorporated a question from the Patient Health Questionnaire version 9 (PHQ-9) [48], which is validated for young people, to identify suicidal ideation [49]. Finally, we used the Sleep Condition Indicator (SCI), a brief scale to evaluate insomnia disorder in everyday clinical practice [50,51].

Upon completing the online survey, participants received a link to download the Mindcraft app (Brain and Behaviour Lab) [52] from the App Store or Play Store, along with a unique login. Participants were asked to use the app for at least 2 weeks. The Mindcraft app is a user-friendly mobile app designed to collect self-reported well-being updates (active data) and phone sensor data (passive data). Participants set their data-sharing preferences during onboarding and can adjust them at any time through the app's settings. Detailed technical specifications of the Mindcraft app are available in the reference [52].

Active and Passive Data Features

Once participants began using the app, we gathered active data and 8 categories of raw passive data sourced from phone sensors and usage metrics. Active data responses (eg, mood, sleep quality, and loneliness), scored on a scale of 1-7, were directly incorporated as features for the ML model.

Passive sensing data were collected via the Mindcraft mobile app, which continuously monitored phone-based sensors and system events in the background. Sensor sampling frequencies were set to balance data resolution with device energy efficiency. All sensors except the location sensor were collected at 15-minute intervals. GPS location data were triggered by significant location changes (as determined by the operating system). From the passive data, we engineered 92 distinct features (Table 1). Location-derived features such as total distance, radius of gyration, and maximum distance from the day's center of mass were derived from GPS logs using Haversine distance metrics and filtered for spurious location jumps. Features were aggregated over 2 time windows, the full 24-hour day and the nighttime period, defined as 10 PM to 6 AM.

A complete list of all active and passive features, along with descriptions and correlations with mental health scales, is provided in Multimedia Appendices 1 and 2, respectively. We highlight the key active and passive data features most strongly associated with each mental health outcome (the union of the top 3 per modality per outcome, ranked by absolute Spearman correlation) in Table 2.

Passive features that were unavailable due to user permission settings or OS-level constraints were set to a sentinel value of –1, indicating "sensor unavailable." This allowed the model to learn platform- or user-specific absence patterns without biasing distributions. All continuous features were *z* score normalized using training set means and SDs. Binary indicator features were kept in their original form.

XSL•FO
RenderX

Table . Summary of features engineered from passive data sensors.

Passive sensor	Number of features	Feature list
Ambient light (Android only)	8	Total, mean, median, and SD of ambient light reading in the day; total, mean, median, and SD of ambient light reading during night hours
App usage (Android only)	36	Total app usage count; unique apps; total, mean, and median time usage in the day; total app usage count; unique apps; total, mean, and median time usage during the night hours; total time in app categories of camera, communication, entertainment, gaming, physical health, mental health, Mindcraft, news, productivity, and social media; percentage time in app categories of camera, communication, entertainment, gaming, physical health, mental health, Mindcraft, news, productivity, and social media
Background noise level (Android only)	10	Total, median, mean, maximum, and SD of background noise levels in the day; total, median, mean, maximum, and SD of background noise levels during night hours
Battery	8	Min, maximum, mean, and median of battery level; number of charges per day; mean battery use per hour; time below 20 percent; nighttime usage count
Location	15	Mean latitude; mean longitude; total distance traveled in a day; location count; maximum distance from home; mean distance from home; median distance from home; nighttime movement; radius of gyration; SD of latitude; SD of longitude; location entropy; time spent at home
Mindcraft usage	3	First hour of use; last hour of use; nighttime usage;
Screen brightness (iOS only)	8	Total, mean, median, SD of screen brightness sensor reading in the day; total, mean, median, SD of screen brightness sensor reading during night hours
Step count	4	Daily step count; is daily step count greater than 5000 steps or 7500 steps or 10,000 steps?

XSL•FO
RenderX

Table . Key active data and passive data features and their Spearman correlations with mental health outcomes (Strengths and Difficulties Questionnaire [SDQ], Sleep Condition Indicator [SCI], Eating Disorder-15 Questionnaire [ED-15], and suicidal ideation).

Feature type	Feature	SDQ[a]	SCI[b]	Suicidal ideation	ED-15[c]
Active	Loneliness - "How lonely are you feeling today?"	0.48[d]	–0.42[d]	0.46[d]	0.31[d]
Active	Negative thinking - "How negative do you think today?"	0.48[d]	–0.47[d]	0.57[d]	0.46[d]
Active	Racing thoughts - "Are you experiencing racing thoughts today?"	0.45[d]	–0.44[d]	0.52[d]	0.48[d]
Active	Sleep quality - "How did you sleep last night?"	–0.37[d]	0.44[d]	–0.34[d]	–0.20[d]
Active	Self-care - "How is your self-care today?"	–0.37[d]	0.33[d]	–0.24[d]	–0.42[d]
Passive	Max background noise levels over the full day	–0.49[e]	0.36	–0.39	–0.1
Passive	Mean latitude of GPS samples over the full day	0.38[f]	–0.23	0.44[e]	–0.03
Passive	Number of entertainment app usage events over the full day	0.37[e]	–0.37[f]	0.51[d]	0.41[e]
Passive	Median ambient light at night	0.06	–0.39[e]	0.14	0.27
Passive	Mean longitude of GPS samples over the full day	0.19	–0.38[f]	0.25	–0.01
Passive	SD of background noise levels over the full day	–0.43	0.36	–0.63[d]	–0.31
Passive	Median app session duration over the full day	0.37[f]	–0.19	0.44[d]	0.19
Passive	Mean background noise level at night	0.2	–0.23	0.42	0.46[f]
Passive	Total number of app usage events at night	0.09	–0.06	0.27	0.39[e]

[a]SDQ: Strengths and Difficulties Questionnaire.

[b]SCI: Sleep Condition Indicator.

[c]ED-15: Eating Disorder-15 Questionnaire.

[d]$P<.001$.

[e]$P<.01$.

[f]$P<.05$.

To reduce day-to-day variability and enhance the stability of daily feature measurements, we computed the cumulative median of each feature for every participant, that is, the median of all values up to each day. This approach progressively aggregates behavioral signals over time, dampening the influence of short-lived anomalies (eg, a one-day spike due to sensor glitches or atypical behavior, which are common in smartphone data) while preserving sustained trends. Multimedia Appendix 3 illustrates how the cumulative median stabilizes noisy input without suppressing genuine behavioral shifts, such as a consistent drop in activity due to worsening mood. In preliminary analyzes, using the same model architecture, hyperparameters, and evaluation protocol, we compared models trained on raw daily features versus those trained on cumulative median-aggregated features. We found that the latter consistently improved balanced accuracy across all outcomes (Multimedia Appendix 4). We further compared the cumulative median and the cumulative mean aggregation. While both methods yielded comparable predictive accuracy, we selected the cumulative median as the primary aggregation function due to its superior

XSL•FO
RenderX

robustness to outliers, a common artifact in passive mobile sensing, thereby ensuring greater stability in user representation.

Using the engineered features, we developed an ML model for each of the 4 mental health outcomes, namely SDQ risk, insomnia, suicidal ideation, and eating disorders. We used 3 distinct feature sets, including active data, passive data, and a combination of both. This design enabled us to assess the predictive strength of each feature set individually and in combination, allowing us to systematically quantify each modality's contribution to prediction performance across mental health outcomes.

Participants were classified as high-risk or low-risk for each outcome using validated thresholds specific to each mental health measure, framing the prediction task as a binary classification problem. Each scale's total score range is as follows: SDQ score 0‑40, SCI 0‑32, ED-15 0‑6, and suicidal ideation 0‑4 (based on response frequency to the question, "Over the last two weeks, how often have you been bothered by thoughts that you would be better off dead or of hurting yourself in some way?"). High-risk classifications were defined as follows: for SDQ, a self-reported score of ≥16 [53]; for insomnia, if their SCI score was ≤16 [51]; for suicidal ideation, if they responded at least once to the question regarding frequency of suicidal thoughts over the last 2 weeks, and for eating disorders, an ED-15 total score exceeded 2.69, which corresponds to the mean plus one SD in a nonclinical population [54]. The proportions of participants classified as high-risk for each mental health outcome are summarized in Table 3.

Table . Demographics and mental health measures of the study population (N=103).

Variables	Values
Sex (Female), n (%)	73 (70.9)
Age (years), mean (SD)	16.1 (1)
Strengths and Difficulties Questionnaire (SDQ) score, mean (SD)	12.8 (6.2)
High-risk SDQ category (SDQ score≥16), n (%)	31 (30.1)
Eating Disorder (ED-15) scale, mean (SD)	2.2 (1.8)
High-risk eating disorder category (ED-15 score≥2.7), n (%)	38 (36.9)
Sleep Condition Indicator (SCI) score, mean (SD)	19.9 (7.8)
High-risk insomnia category (SCI score<17), n (%)	34 (33)
"Over the last two weeks, how often have you been bothered by thoughts that you would be better off dead or of hurting yourself in some way?" mean (SD)	0.6 (0.9)
High-risk suicidal ideation category (≥1), n (%)	38 (36.9)

ML Workflow and Model Development

Figure 1A outlines our ML pipeline, starting with active and passive data collection via the Mindcraft app. The data were preprocessed and engineered to create a comprehensive feature set, which was subjected to a pretraining phase with contrastive learning using triplet margin loss. This pretraining step clustered user-specific features from different days, minimizing day-to-day variability and preserving individual behavioral patterns. The resulting stable embeddings were fine-tuned on labeled data in a supervised setting to predict mental health outcomes. This end-to-end pipeline, combining pretraining and fine-tuning, enabled the development of a predictive model evaluated using balanced accuracy and additional metrics.

XSL•FO
RenderX

Figure 1. Overview of the machine learning (ML) framework for predicting adolescent mental health outcomes using smartphone-based digital phenotyping. (A) Workflow of the ML pipeline, from data acquisition to mental health outcome prediction, incorporating contrastive pretraining with triplet loss and fine-tuning. (B) t-SNE visualization of feature embeddings for a sample of 10 test users before (left) and after (right) contrastive pretraining, showing enhanced user-specific clustering following pretraining. t-SNE: t-distributed Stochastic Neighbor Embedding.

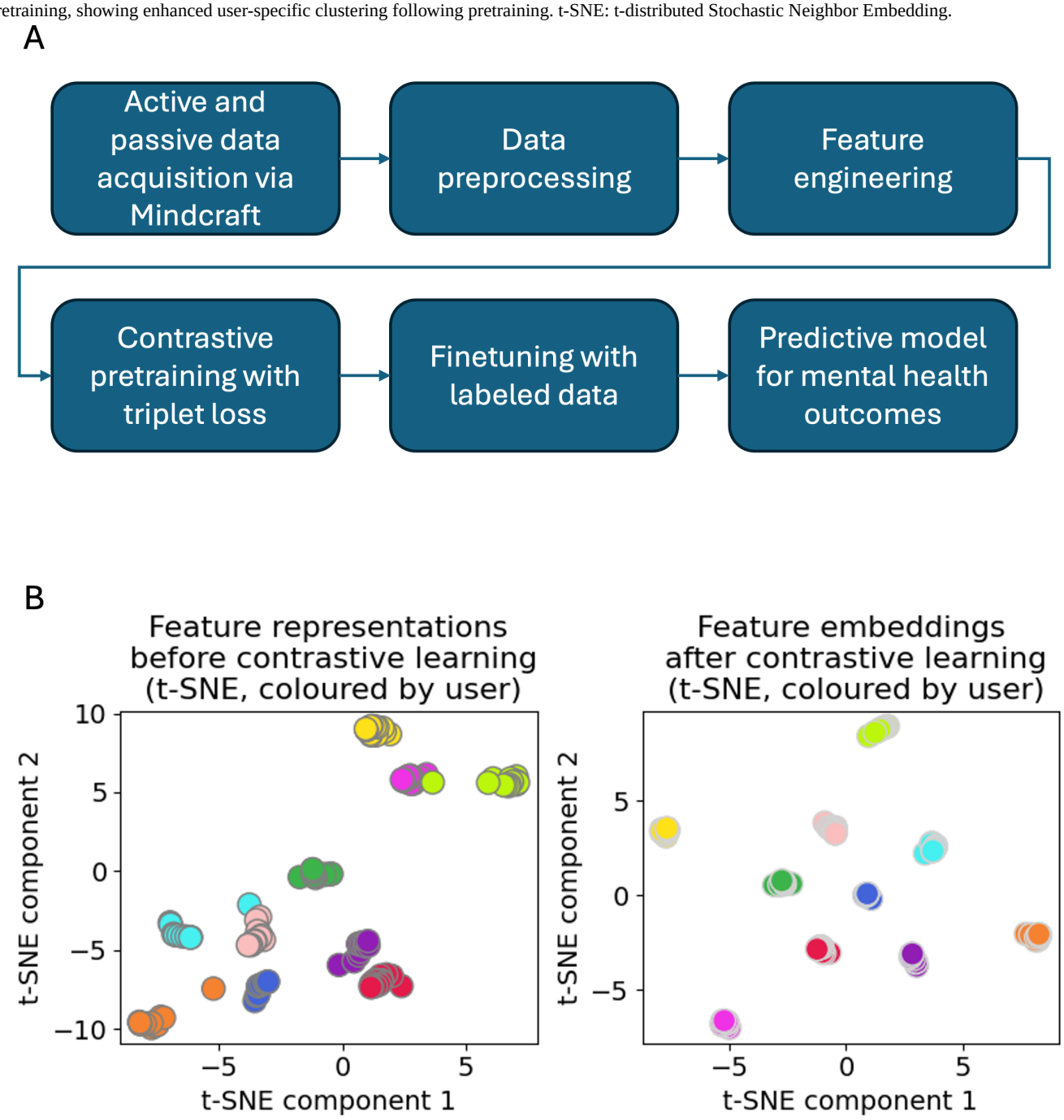

Contrastive Pretraining Model Architecture and Training

To accurately predict mental health risks, it is essential to distinguish between a user's stable behavioral patterns and random daily fluctuations. Raw smartphone data can vary significantly day to day due to external factors (such as school holidays) unrelated to mental health. To address this, we used an ML technique known as contrastive learning with triplet margin loss. In simple terms, this technique teaches the model to recognize a user's unique digital fingerprint. It does so by examining 3 data points (a "triplet") simultaneously:

- Anchor: data from a specific user on a particular day (eg, User A, Monday).
- Positive: data from the same user on a different day (eg, User A, Thursday).
- Negative: data from a different user entirely (eg, User B, Monday).

The model is trained to minimize the distance between the Anchor and the Positive (pulling them together in mathematical space) while maximizing the distance between the Anchor and the Negative (pushing them apart). Repeating this process across thousands of triplets many times helps the model learn to ignore irrelevant daily noise and form a stable, underlying behavioral

XSL•FO
RenderX

fingerprint unique to that user. These stable representations are then used for the subsequent supervised prediction of mental health risk. A simplified visual explanation of this idea is provided in Figure 1B, which helps understand why contrastive pretraining is valuable in datasets with naturally high daily variability, such as smartphone-based behavioral data.

We implemented a custom contrastive learning pipeline in PyTorch (Meta AI), based on established principles of triplet-based metric learning [55,56], with the objective of learning robust, user-specific feature embeddings that remain consistent across days, thereby reducing intrasubject variability while maximizing intersubject separability. For this pretraining phase, we selected triplet margin loss over other contrastive objectives (eg, InfoNCE) due to its suitability for instance-level metric learning, where the aim is to ensure that embeddings from the same user are closer in the latent space than those from different users. Compared to InfoNCE, triplet loss is more stable with moderate mini-batch sizes and is better suited to high-dimensional tabular data [55,57].

Triplets were constructed without using mental health outcome labels, as the pretraining phase was fully self-supervised. A user was first randomly selected from the training set, and a single day from their data was sampled as the Anchor. The Positive sample was randomly chosen from a different day for the same user, while the Negative was drawn from a random day of a different user. The triplet loss requires that an Anchor data point is closer to a Positive data point than it is to a Negative data point, by at least a specified margin m.

The contrastive learning model consisted of a 2-layer multilayer perceptron (MLP) encoder, which mapped input features to a latent embedding space, followed by a 2-layer MLP projection head. The projection head was trained using the triplet loss, allowing the encoder to retain generalizable behavioral representations while the projection space focused on the contrastive objective [58]. The Adam optimizer was used for pretraining with a fixed learning rate of 1×10^{-3}. The model was pretrained for 3 epochs with a batch size of 256 triplets. The margin m was set to 1.0 in our experiments.

Supervised Fine-Tuning Model Architecture and Training

Following contrastive pretraining, the learned encoder was used as a fixed base for the downstream classification tasks. Its weights were frozen to preserve the stable user-specific embeddings. A new 2-layer MLP classification head was added on top of the frozen encoder to perform binary classification for each mental health outcome. The fine-tuning model was trained using the Adam optimizer with a learning rate of 1×10^{-3} for 20 epochs and a batch size of 1024. Binary cross-entropy with logits loss was used as the training objective, with class weights applied to account for imbalance. These weights were calculated based on the inverse frequency of each class within the training set of each cross-validation fold.

Key architectural hyperparameters (eg, number of layers, dimensions of embeddings, and projections) and training hyperparameters (eg, learning rate, batch size, triplet margin, and number of epochs) were selected based on common practice in contrastive learning and refined through exploratory tuning on a validation split of the training data within the first cross-validation fold. No test user data from any fold was used during hyperparameter selection. All selected hyperparameters were then fixed across folds to ensure methodological consistency and reproducibility and to prevent data leakage.

To demonstrate the effect of contrastive pretraining, Figure 1B shows t-distributed Stochastic Neighbor Embedding visualizations of feature embeddings for a sample of 10 test users. Before pretraining (left plot), user-specific data points were scattered with minimal clustering, reflecting high day-to-day variability. After applying contrastive learning with triplet loss (right plot), data points from the same user formed tighter clusters, indicating enhanced user-specific feature stability.

Evaluation and Benchmarking

Day-wise prediction probabilities were averaged for each test user to obtain a single, user-level prediction. Model performance was evaluated using balanced accuracy as the primary metric, along with other relevant metrics such as area under the receiver operating characteristic curve (AUC) and F_1-score, to assess classification outcomes comprehensively. For interpretability, Shapley additive explanations (SHAP) values [59] were computed for test folds using DeepExplainer [60], offering insight into the contributions of specific features to classification outcomes.

We validated the model's performance using leave-one-subject-out cross-validation (LOSO-CV), where data from all but one user were used for training and validation, with the excluded user's data serving as the test set. This approach ensures the model's generalizability to new individuals, closely simulating real-world applications where accurate predictions for unseen users are critical.

To benchmark our model, we compared it with a CatBoost (Yandex) classifier [61] and an MLP network that had a similar number of parameters but was trained without pretraining. Both benchmarks used class-weight balancing to address the class imbalance in the dataset. CatBoost was selected for its strong performance on tabular data and its built-in capability to handle class imbalance [62], making it well-suited for datasets like ours. These comparisons isolated the effect of contrastive pretraining on model performance.

External Validation on an Independent Cohort

We collected data from an additional cohort of 90 adolescents across 5 new secondary schools in London. To create an independent external validation set, we randomly split this new cohort into 2 halves. One half (n=45) was added to the original dataset to form an expanded training set, while the remaining half (n=45) served as a fully held-out external validation cohort. Importantly, no model architecture, hyperparameters, preprocessing, or feature engineering steps were modified after adding the new data; the complete pipeline, including contrastive pretraining and supervised fine-tuning, remained fixed. This design allowed us to assess generalizability to users from different schools, collected at a later time point, with different demographic characteristics and smartphone-sensor enabling

XSL•FO
RenderX

patterns. Model performance on this external cohort was evaluated using balanced accuracy following the same user-level prediction protocol used in the original LOSO-CV analysis.

Ethical Considerations

Ethics approval for the study was granted by the Imperial College London Research Ethics Committee (reference number: ICREC 20IC6132). All procedures complied with relevant national and institutional ethical standards and with the Declaration of Helsinki. Informed consent was obtained digitally from all participants at the start of the survey; for participants younger than 16 years, parental consent was obtained before participation. Participants could withdraw from the study at any time before the start of data analysis. Study-specific identification numbers were used to maintain participant anonymity, and data handling complied with the General Data Protection Regulation (GDPR) for health and care research. Participation was voluntary and uncompensated. Multimedia Appendix 5 provides a CONSORT (Consolidated Standards of Reporting Trials)-style flow diagram outlining the subject inclusion process from the initial 205 survey respondents.

Results

Recruitment and App Usage

Figure 2A provides a conceptual overview of the study, demonstrating how active data (eg, sleep quality, mood, and loneliness) and passive data (eg, location, app usage, and noise levels) collected via the Mindcraft app are integrated into a contrastive learning-based deep neural network to predict mental health outcomes, including SDQ risk, insomnia, suicidal ideation, and eating disorders.

A total of 103 students from 3 London schools downloaded and installed the Mindcraft app. The average age was 16.1 years (SD 1), with 71% identifying as female, 25% as male, and 4% as other or nonbinary. The skew in gender distribution is partially due to one of the participating schools being girls-only. Of the participants, 78 used the app on iPhones, and 25 used Android phones. Table 3 provides demographic information and mental health outcome scores, and Multimedia Appendix 6 illustrates the distribution of the different mental health measures across our study population.

Participants contributed active data via self-reported measures and passive data through smartphone sensors. Active data included daily ratings of mental well-being measures such as sleep quality, mood, confidence, and loneliness on a 1 - 7 scale. Passive data comprised data from phone sensors like location, app usage, ambient noise, and step count. Figure 2B shows user engagement patterns over the 14-day study period. Initial engagement was high, with all participants contributing at baseline. However, active data engagement declined more rapidly than passive data, with 14 users contributing active data and 36 users contributing passive data on day 14.

Engagement with active data measures (Figure 2C) remained consistent across users, with slight variations reflecting individual preferences. In contrast, passive data collection exhibited substantial variability (Figure 2D). While 36 users opted not to enable any sensors, others enabled multiple categories. The most frequently enabled sensors were step count and battery usage, followed by Mindcraft usage and screen brightness (Figure 2E). The heatmap visualization of passive data coverage by users and sensor type (Figure 2F) underscores substantial interuser variability, with some users providing comprehensive coverage across multiple sensors and others contributing sporadically.

XSL•FO
RenderX

Figure 2. Study overview and participant engagement with active and passive data collection using the Mindcraft app. (A) Conceptual overview of the study. (B) User engagement trends for active and passive data over the 14-day period. (C) User participation across active data questions. (D) Distribution of enabled passive sensors among users. (E) User participation across different passive sensor types. (F) Heatmap of passive data completeness by user and sensor type. SDQ: Strengths and Difficulties Questionnaire.

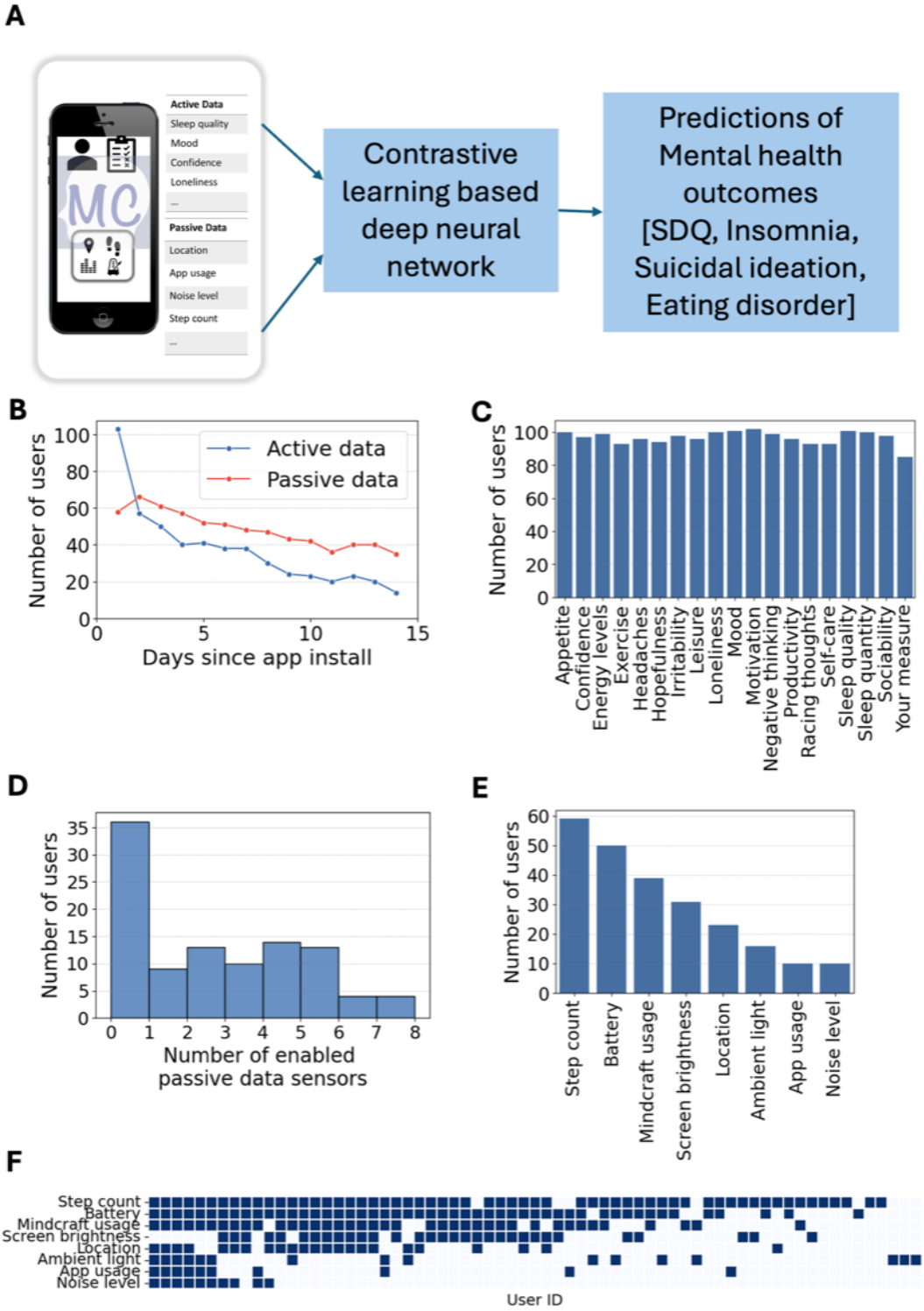

Exploratory Analysis of Active and Passive Data Features

Figure 3 provides an overview of descriptive statistics and correlations among active and passive data features collected from users. Figure 3A illustrates the distribution of self-reported active data features on a scale of 1-7. Positive indicators, such as mood, motivation, and confidence, had higher mean values than negative indicators, such as negative thinking, racing thoughts, and irritability.

XSL•FO
RenderX

Figure 3. Descriptive statistics of active and passive digital phenotyping features. (A) Distribution of responses across active data features. (B) Correlation heatmap of active data features. (C) Frequency distributions of passive data features: daily step count. (D) Number of unique apps opened per day. (E) Location entropy reflecting movement variability. (F) Mean background noise levels at night.

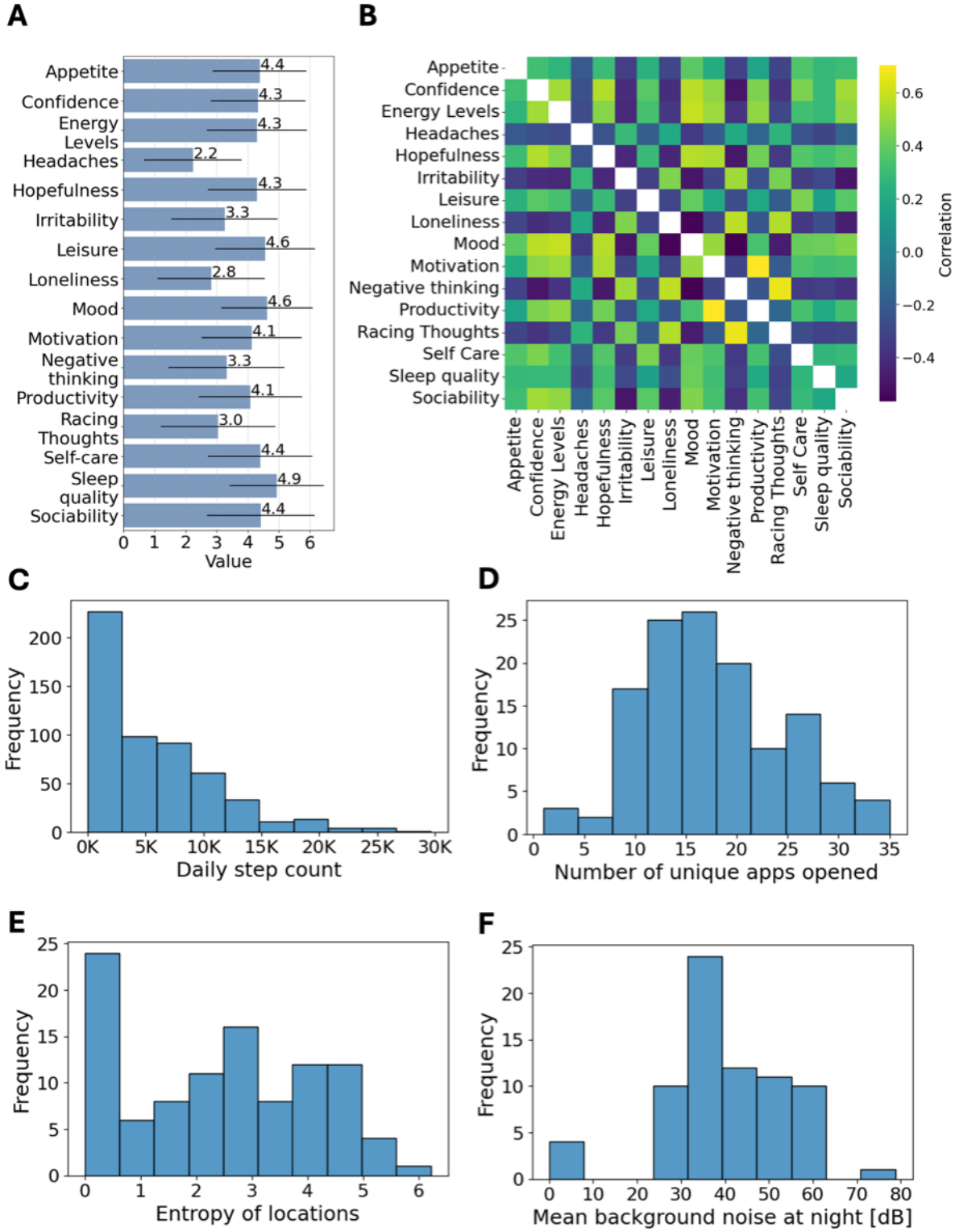

The correlation heatmap (Figure 3B) highlights relationships among active data features. A fully annotated, high-resolution version of this heatmap is provided in Multimedia Appendix 7. The strongest correlation (r=0.7) was observed between motivation and productivity, followed by a strong association between negative thinking and racing thoughts (r=0.66). Positive correlations were also observed between 2 well-being indicators, energy levels and mood (r=0.58), and between 2 distress indicators, loneliness and negative thinking (r=0.57). Conversely, negative correlations were seen, such as between mood and negative thinking (r=–0.56) and irritability and sociability (r=–0.49).

Figures 3C-3F illustrate 4 of the 92 engineered passive data features. The distribution of daily step counts (Figure 3C) is

XSL•FO
RenderX

right-skewed, with most users taking fewer than 10,000 steps per day. Figure 3D shows the frequency of unique apps opened daily, peaking at 15 - 20 apps, indicating varying levels of mobile engagement. The entropy of locations visited (Figure 3E) reflects movement variability, with higher values suggesting diverse activity patterns. Finally, Figure 3F highlights the distribution of mean background noise levels at night, clustering between 30 and 50 decibels.

Associations Between Active and Passive Features and Clinical Outcomes

To assess the alignment between digital phenotyping features and clinical mental health symptoms, we examined Spearman correlations between active and passive features and continuous scores on the SDQ, SCI, ED-15, and suicidal ideation frequency. Among active features, negative thinking consistently showed the strongest associations across all outcomes (ρ=0.48 with SDQ, ρ=–0.47 with SCI, ρ=0.57 with suicidal ideation, ρ=0.48 with ED-15, all $P<.001$). Higher levels of racing thoughts and loneliness were associated with more severe mental health symptoms, whereas greater confidence and hopefulness were linked to reduced risk. Key passive features also demonstrated moderate associations with outcomes. For instance, greater entertainment app usage consistently showed associations across all outcomes (ρ=0.37 with SDQ, ρ=–0.37 with SCI, ρ=0.51 with suicidal ideation, ρ=0.41 with ED-15, all $P<.05$). Nighttime ambient light exposure and location variability (latitude and longitude) were also relevant across multiple outcomes, particularly for insomnia and eating disorder symptoms. Key active and passive features, along with their correlations with each outcome, are summarized in Table 2. Multimedia Appendices 1 and 2 list detailed descriptions and Spearman correlations for all active and passive features used in the predictive modeling pipeline, offering an overview of their associations with SDQ, SCI, ED-15, and suicidal ideation.

To further illustrate the discriminative capacity of key digital phenotyping features, we compared the top 5 active and passive features between high- and low-risk groups for each mental health outcome. Multimedia Appendix 8 shows the variability in these features across the 4 outcomes (SDQ, insomnia, suicidal ideation, and eating disorder). The observed group differences are consistent with the correlation-based findings and reinforce the predictive relevance of both active and passive data streams in distinguishing individuals at elevated mental health risk.

Performance of Models Predicting Mental Health Outcomes

Building on these associations, we evaluated how well ML models could predict mental health outcomes using active, passive, and combined data. Figure 4A illustrates the balanced accuracy of predictive models for the 4 mental health outcomes (SDQ-high risk, insomnia, suicidal ideation, and eating disorder) evaluated across 10 repetitions of LOSO-CV. The analysis involved 3 feature sets, including passive data, active data, and a combination of both. This evaluation was restricted to the 67 participants who provided both active and passive data to ensure a fair comparison. The red dashed line indicates chance-level performance, and statistically significant differences are denoted by asterisks (*$P<.05$, **$P<.01$, ***$P<.001$; Wilcoxon signed-rank test).

XSL•FO
RenderX

Figure 4. Model performance for predicting mental health outcomes using active and passive digital phenotyping data. (A) Balanced accuracy of mental health outcome predictions (SDQ-high risk, insomnia, suicidal ideation, and eating disorder) using passive, active, and combined data. (B) Comparison of balanced accuracy for models with contrastive pretraining, without pretraining, and a CatBoost model, showing the performance benefit of pretraining (*P*<.001, paired *t* test). (C) Confusion matrices for the combined data model's predictions, showing true-positive and true-negative classifications across mental health outcomes. *: *P*<.05; **: *P*<.01; ***: *P*<.001; SDQ: Strengths and Difficulties Questionnaire.

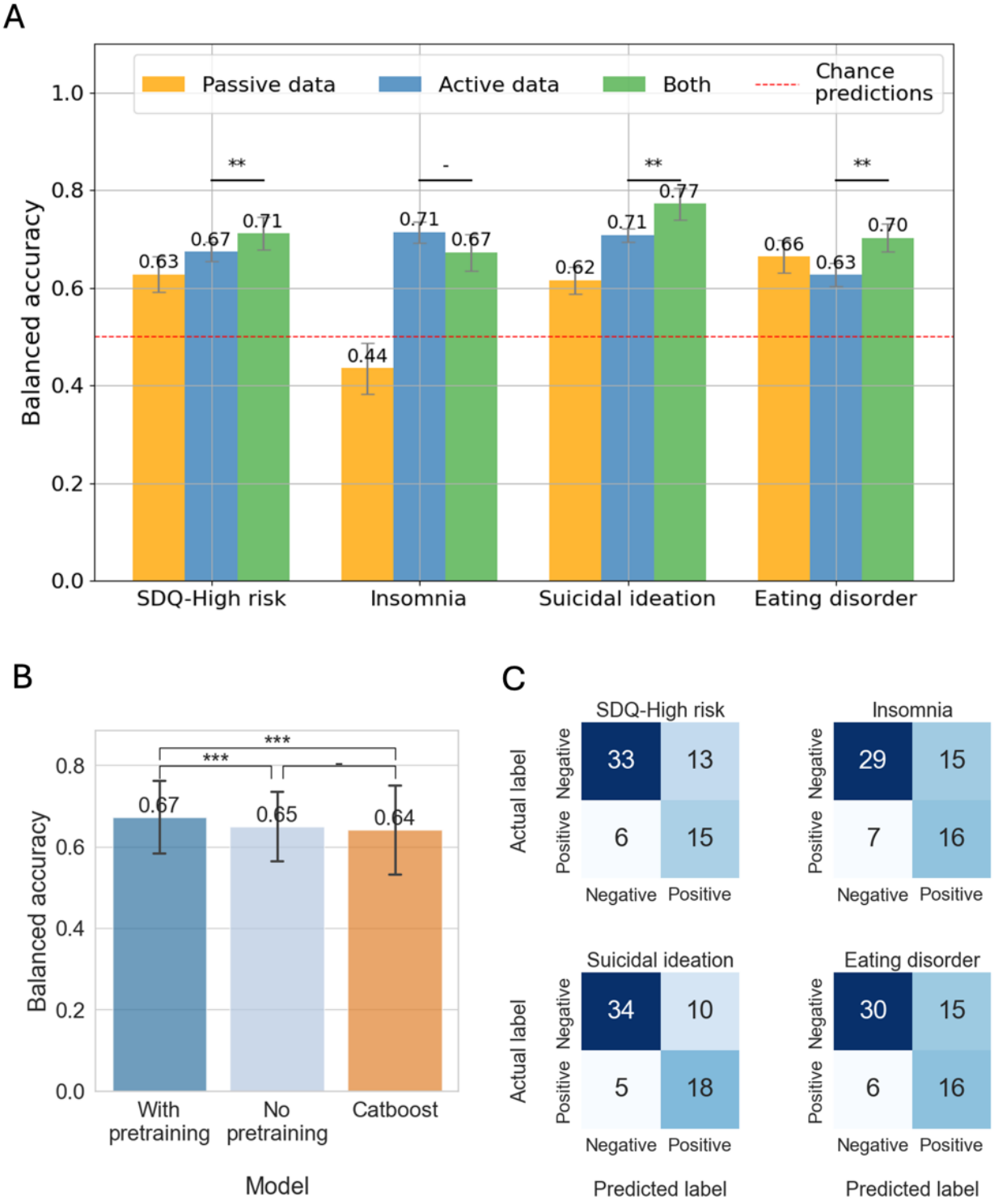

For SDQ-high risk, the model using passive data achieved a balanced accuracy of 0.63, while active data alone reached 0.67 (Multimedia Appendices 1 and 2). The combined model, leveraging both data types, achieved a significantly higher balanced accuracy of 0.71 compared to active data alone (*P*=.003, Wilcoxon signed-rank test). Similarly, the combined data model outperformed the active data alone for eating disorder predictions, with balanced accuracies of 0.70 and 0.63, respectively (*P*=.003; Wilcoxon signed-rank test). In predicting insomnia, the combined model achieved a balanced accuracy of 0.67, while passive data alone performed below the chance level (0.44). For suicidal ideation, the combined model achieved

XSL•FO
RenderX

the highest balanced accuracy of 0.77, significantly outperforming both active data (0.71) and passive data (0.62; *P*=.003; Wilcoxon signed-rank test). Table 4 summarizes additional performance metrics, including the AUC, the area under the precision-recall curve, F_1-scores, sensitivity, specificity, precision, and recall for each mental health outcome.

Table . Detailed performance metrics for mental health outcome predictions.

Metric	SDQ[a]-high risk, mean (SD)	Insomnia, mean (SD)	Suicidal ideation, mean (SD)	Eating disorder, mean (SD)
Balanced accuracy	0.71 (0.03)	0.67 (0.04)	0.77 (0.03)	0.70 (0.03)
AUC[b]	0.77 (0.03)	0.74 (0.02)	0.82 (0.03)	0.73 (0.02)
AUC-PR[c]	0.53 (0.04)	0.52 (0.05)	0.64 (0.05)	0.52 (0.03)
*F*1-score	0.61 (0.04)	0.59 (0.04)	0.70 (0.04)	0.61 (0.03)
F1 macro	0.69 (0.03)	0.66 (0.04)	0.76 (0.03)	0.68 (0.03)
Sensitivity	0.71 (0.06)	0.68 (0.03)	0.78 (0.05)	0.74 (0.03)
Specificity	0.71 (0.05)	0.66 (0.07)	0.77 (0.04)	0.67 (0.04)
Precision	0.53 (0.04)	0.52 (0.05)	0.64 (0.05)	0.52 (0.03)
Recall	0.71 (0.06)	0.68 (0.03)	0.78 (0.05)	0.74 (0.03)

[a]SDQ: Strengths and Difficulties Questionnaire.

[b]AUC: area under the receiver operating characteristic curve.

[c]AUC-PR: area under the precision-recall curve.

Figure 4B demonstrates the effectiveness of contrastive pretraining. Models with pretraining achieved the highest balanced accuracy (0.67), significantly outperforming both the model without pretraining (0.65; *P*<.001; paired 2-tailed *t* test) and the CatBoost model (0.64; *P*<.001; paired 2-tailed *t* test).

Figure 4C presents the confusion matrices for the combined data models. For SDQ-high risk, the model correctly identified 33 negatives and 15 positives, with 6 false negatives and 13 false positives. The model had higher misclassification rates for insomnia, with 7 false negatives and 15 false positives. In predicting suicidal ideation, the model demonstrated strong performance, correctly classifying 34 negatives and 18 positives, with only 5 false negatives and 10 false positives. Similarly, for eating disorders, the model accurately identified 30 negatives and 16 positives, with 6 false negatives and 15 false positives.

To assess whether the active-only model's predictive performance generalizes across user subgroups, we compared balanced accuracy for the subset (n=67) who provided both active and passive data with that of the full cohort (n=103), which includes participants who shared only active data (Multimedia Appendix 9). The consistent results across these 2 groups indicate that the active-only model's performance is stable and not restricted to a specific subgroup, thereby confirming that restricting the analysis to the subset of users with both active and passive data (as done in Figure 4A) has not introduced bias in model performance.

External Validation Performance on an Independent Cohort

Model performance on the held-out external validation cohort (n=45 adolescents from 5 additional schools; none of these adolescents' data were used for model development) showed a similar pattern to the LOSO-CV results, albeit with reduced accuracy for some outcomes (Figure 5). Figure 5 shows balanced accuracy for models trained using passive data only, active data only, and combined active and passive data; error bars indicate standard deviation across 10 repeated runs. The red dashed line indicates chance-level performance, and statistically significant differences are denoted by asterisks (**P*<.05, ***P*<.01, ****P*<.001; Wilcoxon signed-rank test). Using combined active and passive data, balanced accuracy in the external cohort was 0.72 for SDQ-high risk, 0.63 for insomnia, 0.63 for suicidal ideation, and 0.63 for eating disorder risk (vs 0.71, 0.67, 0.77, and 0.70, respectively, in the LOSO-CV analysis). Models trained using passive data only and active data only showed the same qualitative pattern, with active and combined data generally outperforming passive data alone, but with lower accuracies than in LOSO-CV, particularly for suicidal ideation and eating disorders. A full set of evaluation metrics for the combined model in the external validation sample is provided in Multimedia Appendix 10. Overall, the external validation results showed a similar performance pattern to the LOSO-CV analysis, with reduced accuracy for suicidal ideation and eating disorder.

XSL•FO
RenderX

Figure 5. External validation performance on an independent cohort. Balanced accuracy for predicting SDQ high risk, insomnia, suicidal ideation, and eating disorder risk in an external validation sample (n=45 adolescents not used for model development). Error bars indicate SD across 10 repeated runs. *: *P*<.05; **: *P*<.01; ***: *P*<.001; SDQ: Strengths and Difficulties Questionnaire.

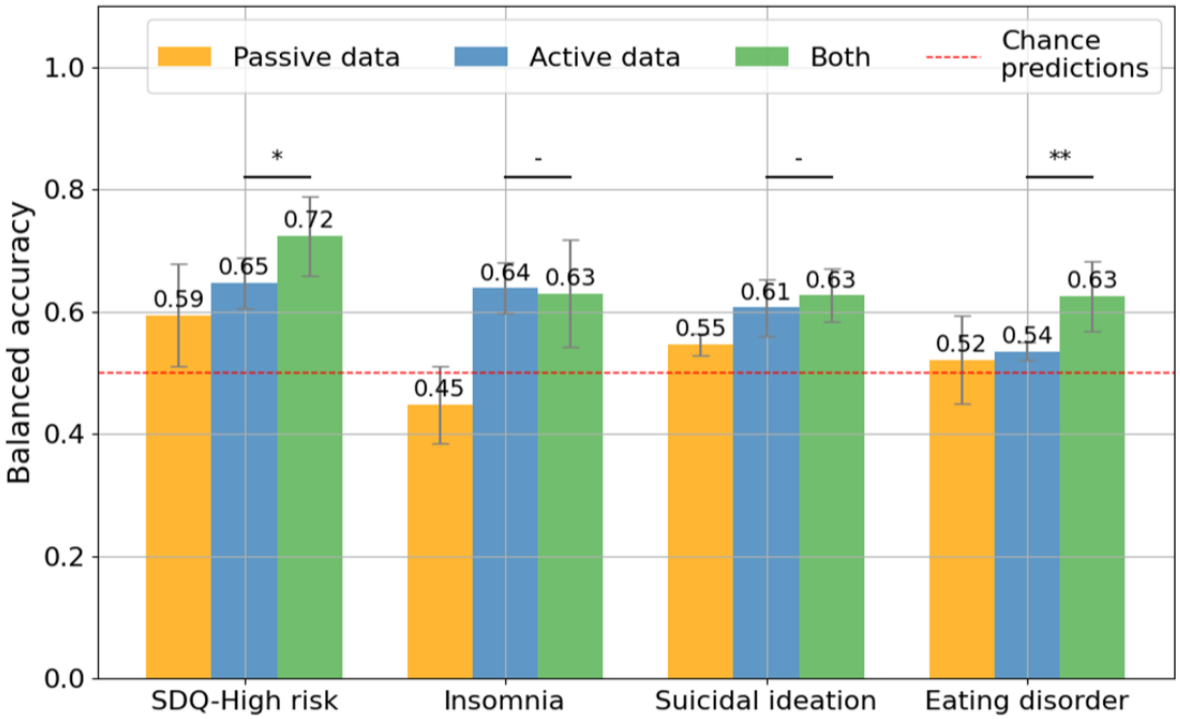

Model Fairness Across Gender and School Contexts

Given known differences in smartphone use and mental health between demographic groups, we examined whether model performance varied systematically by gender (Multimedia Appendix 11). Participants were grouped into female (n=46) and male and other (n=21). For the combined active and passive model, balanced accuracy was broadly comparable across gender for all 4 outcomes, with differences modest in magnitude and inconsistent in direction (eg, suicidal ideation: 0.73 vs 0.74). Error bars overlapped for all outcomes, and no subgroup was consistently disadvantaged, suggesting no evidence of systematic gender-related performance degradation in this sample.

Because individual-level socioeconomic status (SES) data were unavailable, we used school as a contextual proxy for both SES and the institutional environment. The 3 participating schools differed in selectivity, gender composition, and catchment-area deprivation: School 1 (girls-only, partially selective grammar in a mid-SES area), School 2 (mixed-gender community school in a deprived urban area), and School 3 (mixed-gender selective sixth-form college in a mid-to-higher SES area). For each school, we compared balanced accuracy for students from that school with that for students from the other 2 schools combined (Multimedia Appendix 11). Across all 4 outcomes, differences were varied in direction rather than systematically favoring or disadvantaging any single school (eg, School 1 vs others: SDQ 0.71 vs 0.73; insomnia 0.76 vs 0.59; suicidal ideation 0.68 vs 0.76; eating disorder 0.65 vs 0.71), with overlapping CIs in all cases. Overall, these analyses provide no evidence of systematic performance degradation associated with school context or school-level SES, although the study was not powered to detect more subtle or intersectional fairness effects.

Robustness to Heterogeneous Sensor Activation

To evaluate whether missing passive-sensor streams introduced systematic bias, we stratified participants based on the number of passive sensors enabled. The median number of sensors in the dataset was 3; therefore, we compared model performance between users with ≤3 sensors enabled and those with >3 sensors enabled (Multimedia Appendix 12). Model performance was evaluated separately for each subgroup for all 4 outcomes using the combined active-and-passive model. Differences in balanced accuracy were inconsistent in direction across outcomes, indicating no evidence of systematic performance degradation for users with more missing sensor data. Notably, for suicidal ideation, the 2 subgroups performed nearly identically (0.74 vs 0.72). For eating disorder risk, the ≤3-sensor group performed better (0.80 vs 0.56), suggesting that the core predictive signals for this outcome are captured within the most commonly enabled sensors and that additional sensors may introduce noise for this specific target. These results provide no evidence that heterogeneous sensor activation introduced systematic group-level bias against users with fewer sensors. They also suggest that the combined model is reasonably robust to the observed patterns of missing passive data. However, our sample size limits the detection of more subtle fairness effects related to specific sensor combinations or intersectional subgroups.

XSL•FO
RenderX

Predictive Accuracy Across Mental Health Risk Groups

Figure 6 illustrates model accuracy in predicting mental health risks using combined active and passive data, segmented by risk levels for various mental health measures. The model performed exceptionally well at extreme risk levels, achieving near-perfect accuracy for high-risk groups (eg, SCI scores 0 - 8 and EDQ scores 4 - 6) and low-risk groups (eg, SDQ scores 1 - 8). However, accuracy decreased significantly in ranges near thresholds (eg, SDQ scores 9 - 16 and SCI scores 9 - 16).

Figure 6. Accuracy of mental health risk prediction across different levels of (A) Strengths and Difficulties Questionnaire (SDQ) total score. (B) Sleep Condition Indicator (SCI) score. (C) Frequency of suicidal ideation thoughts. (D) Eating Disorder (ED-15) total score.

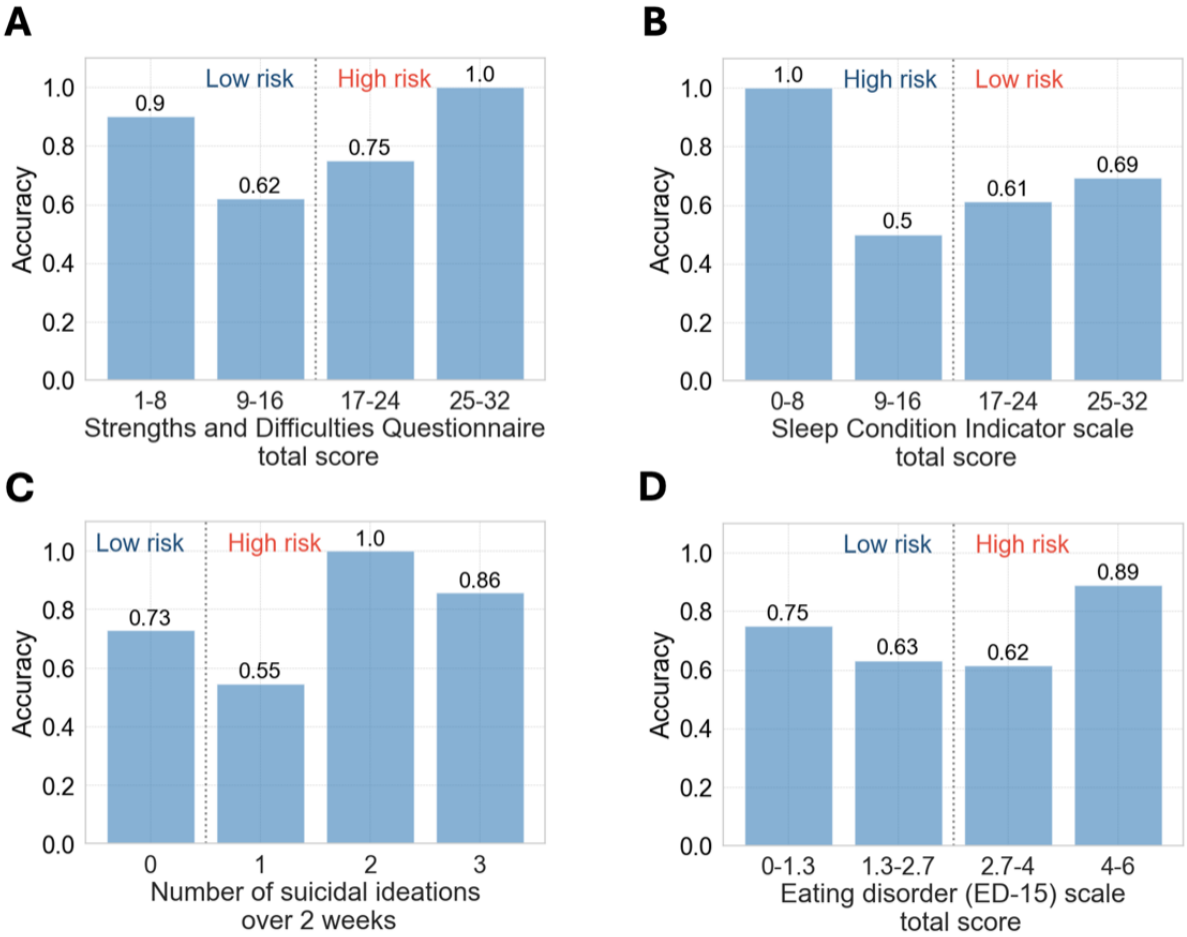

Model Interpretability: Active and Passive Data Contributions

Figure 7A illustrates the feature importance calculated using SHAP values for predicting the SDQ high-risk category using a combination of both active and passive data, with passive data aggregated by sensor type and active data shown individually. The top predictors included negative thinking, location features, app usage, racing thoughts, and self-care. Cognitive and emotional indicators (eg, negative thinking and racing thoughts) ranked highest among active data features, while movement and environmental stability (eg, location entropy and step count) dominated passive data contributions. Corresponding SHAP-based feature importance plots for predicting insomnia, suicidal ideation, and eating disorders are provided in Multimedia Appendices 13-15, respectively.

XSL•FO
RenderX

Figure 7. Feature importance analysis for predicting the SDQ high-risk category using both active and passive data. (A) SHAP-based feature importances, with passive data aggregated by sensor type and active data shown individually. (B) Distribution of the top five active data features across SDQ risk categories. (C) Distribution of the top five passive data features across SDQ risk categories. Statistically significant differences between low-risk and high-risk groups are indicated (*: *P*<.05, **: *P*<.01, ***: *P*<.001, *t* test). SDQ: Strengths and Difficulties Questionnaire; SHAP: Shapley Additive Explanations.

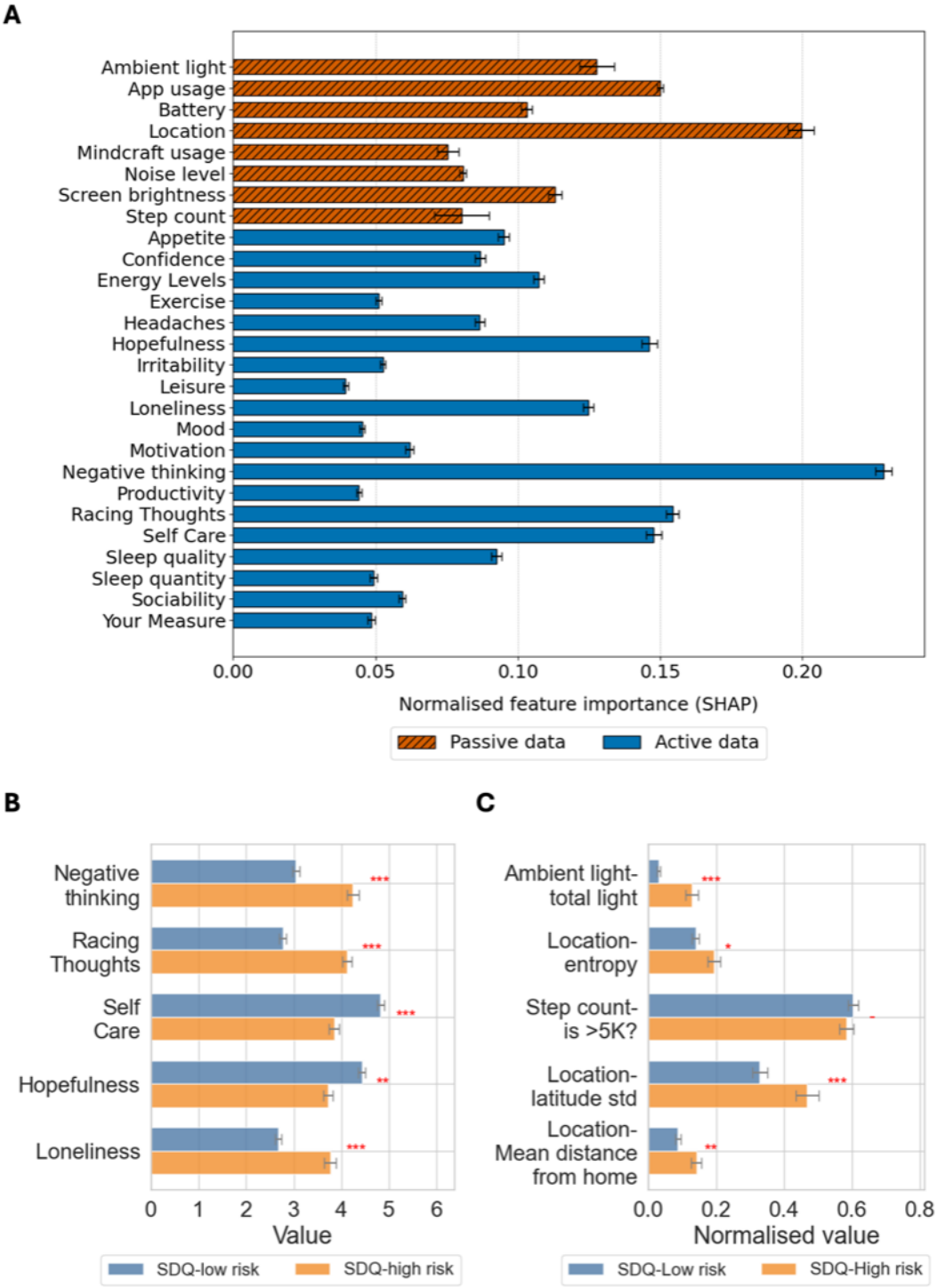

The distribution of the top five active data features (negative thinking, racing thoughts, self-care, hopefulness, and loneliness) showed clear distinctions between low- and high-risk SDQ groups (Figure 7B). Negative thinking and racing thoughts were significantly higher in the high-risk group (*P*<.001; *t* test). Conversely, self-care (*P*<.001; *t* test) and hopefulness (*P*<.001; *t* test) were significantly lower. Loneliness was also notably higher in the high-risk group (*P*<.001; *t* test).

The distribution of the top 5 passive data features (ambient light, location entropy, step count, latitude SD, and mean distance from home) highlighted significant differences between risk

XSL•FO
RenderX

groups (Figure 7C). High-risk individuals showed greater ambient light exposure (P<.001; t test), potentially reflecting greater exposure to light at night and sleep disruptions. They also exhibited higher location entropy (P=.02; t test) and latitude variability (P<.001; t test). Additionally, fewer high-risk individuals exceeded 5000 daily steps (P=.002; t test).

Discussion

Principal Findings

Our study demonstrated the effectiveness of integrating active self-reported and passive smartphone sensor data to predict adolescent mental health risks using a novel ML framework. By leveraging data collected via the Mindcraft app, we evaluated predictions across 4 critical mental health outcomes, namely SDQ-high risk, insomnia, suicidal ideation, and eating disorders. Combined models consistently outperformed unimodal approaches, achieving competitive balanced accuracies across all outcomes (eg, 0.77 for suicidal ideation and 0.71 for SDQ-high risk), even in a broad, nonclinical adolescent sample (Figure 4A). These results highlight the complementary value of passive data, which unobtrusively captures continuous behavioral patterns that enrich the subjective insights provided by active data.

Importantly, these results were obtained in a nonclinical, school-based adolescent cohort that was deliberately broad, behaviorally diverse, and not selected for help-seeking status. Such populations introduce substantial intragroup variability, natural fluctuations in engagement, and noise in both active and passive measures, making predictive modeling especially challenging. Previous digital phenotyping studies in non-clinical adolescent or student samples have reported the AUC or F_1-scores typically ranging from 0.60 to 0.80 when predicting stress, depression, or anxiety using mobile sensing [35-41]. Many of these studies also involved smaller or more homogeneous samples and often focused on a single outcome. Our balanced accuracies of 0.70‑0.77 are therefore competitive, especially given that our study simultaneously addressed 4 distinct mental health outcomes in a naturalistic, non–help-seeking adolescent sample. The external validation results suggest that the learned behavioral representations generalize reasonably well to new schools and time periods, while also highlighting some loss of accuracy for the more challenging outcomes.

User engagement patterns indicated sustained utility of passive data collection, underscoring its lower participant burden and feasibility in scalable longitudinal mental health monitoring. Participants preferred less intrusive metrics, including step count, battery usage, and screen brightness, emphasizing the importance of prioritizing user-friendly data collection methods. Our innovative contrastive learning approach effectively addressed variability inherent in daily behavioral data, stabilizing user-specific feature representations. This methodological advancement yielded improved performance and increased confidence in the model's applicability to the real world.

Comparison With Prior Work

Previous work has largely focused on adults or clinical populations [23-40], limiting its applicability to adolescents in community settings. Even among adolescent-focused studies [25,33,34,63,64], most relied solely on passive sensing and targeted depression or anxiety, limiting the generalizability of their findings to broader, nonclinical groups.

Digital self-monitoring has a potential role in multiple stages of the clinical pathway, from prevention to clinical intervention. Our work addresses a broader range of mental health outcomes of internalizing and externalizing disorders, eating disorders, insomnia, and the presence of suicidal ideation in a nonclinical, non–help-seeking school-going adolescent population. MacLeod et al [63], the closest study to ours, included younger adolescents from clinical and nonclinical settings but relied solely on passive sensing. To our knowledge, this study is the first to use ML to accurately predict mental health risk across a broad range of outcomes in low- and higher-risk school-going adolescents, using a combination of active and passive data in a general, nonclinical population.

To interpret our results in light of previous literature, we discuss the findings for each mental health outcome in our study individually below.

SDQ

SHAP analysis highlighted key passive features (Figure 7C), including lower step count, increased location entropy, and elevated ambient light exposure, which were linked to behavioral and environmental patterns associated with mental health risk, consistent with prior studies on physical inactivity, disrupted routines, and light exposure [18-21]. Fewer high-risk individuals exceeded 5000 daily steps, reinforcing associations with sedentary behavior [19,20]. Location entropy, capturing variability in movement, may signal a lack of daily structure [18], while ambient light patterns could reflect disturbed circadian rhythms [21]. In parallel, active features such as negative thinking, racing thoughts, and poor self-care also ranked highly (Figure 7B) and aligned with literature on internalizing and externalizing symptomatology [65-67].

Insomnia

Predictions were driven by active features such as self-care, negative thinking, and appetite, and by passive measures such as app usage hours, nighttime movement, and ambient light (Multimedia Appendix 13). This aligns with earlier research highlighting subjective perceptions, such as increased negative thinking, loneliness, and decreased self-care and hopefulness, as critical factors in sleep disturbances [68,69]. The significance of ambient light exposure further supports evidence linking circadian disruptions to poor sleep quality in adolescents [70,71].

Suicidal Ideation

Key active features (Multimedia Appendix 10), including negative thinking, loneliness, and reduced hopefulness, reflect core cognitive-affective vulnerabilities and are consistent with prior evidence linking these traits to elevated suicide risk in adolescents [72,73]. Passive features (Multimedia Appendix 13) such as screen brightness variability and late-night Mindcraft

XSL•FO
RenderX

app usage may indicate disrupted circadian rhythms, which have been associated with poorer mental health outcomes and suicidal ideation [70,74-76].

Eating Disorder

High-risk individuals reported lower appetite and energy, poorer self-care, and more negative thinking (Multimedia Appendix 11), consistent with links between cognitive-emotional dysregulation, somatic symptoms, and disordered eating in adolescents [77,78]. Passive data (Multimedia Appendix 11) showed consistently elevated screen brightness and earlier app use in the high-risk group, possibly indicating compulsive nighttime device use and disrupted sleep-wake cycles, both associated with emotional dysregulation and body image concerns [79,80].

Strengths and Limitations

This study offers several strengths that advance the field of adolescent digital mental health. By integrating active self-reports and passive smartphone sensor data via the Mindcraft app, we provide a scalable, unobtrusive, and practical framework for early risk detection. Notably, our models maintained strong performance despite high attrition in active data, underscoring the robustness and low participant burden of passive sensing. Additionally, our use of contrastive learning to stabilize day-to-day behavioral features enhanced model robustness. SHAP-based interpretability increased transparency and clinical relevance, both of which are key attributes for adoption in real-world settings.

Several limitations warrant consideration. First, the sample was relatively small and drawn from 3 London-based schools, which may limit the generalizability of our findings, particularly regarding socioeconomic and regional representativeness. Prior research has shown that digital phenotyping features, such as smartphone usage patterns and affective expression, as well as the manifestation and reporting of mental health symptoms, can differ by gender, geography, and cultural context [81-83]. Our geographic concentration in an urban, high-resource setting may therefore not capture behavioral or contextual variability observed in rural or culturally distinct environments, where access to technology, school structures, and sociocultural norms may differ substantially. Furthermore, the gender imbalance driven by the inclusion of a girls-only school may have biased the learned representations toward behavioral and emotional patterns more characteristic of female adolescents, potentially reducing performance for male or nonbinary young people. Consequently, while our findings provide novel insights into adolescents' behavioral monitoring, caution is warranted when extrapolating these results to nonurban, gender-balanced, and resource-limited populations. Future work should therefore include socioeconomically and demographically diverse samples to assess the generalizability of this approach more robustly.

Second, passive sensor data quality varied across device types, operating systems, permission settings, and user engagement, with some participants not enabling key sensors, such as location or app usage, resulting in heterogeneous data completeness. We applied sentinel values to flag missing sensor inputs, enabling the model to learn from patterns of missingness. However, this may not fully eliminate systematic differences, as participants with more complete data may differ meaningfully from those with limited data, potentially skewing model learning. Future work should incorporate fairness-aware modeling and stratified evaluation to ensure equitable performance across subgroups [84]. While cumulative median aggregation improves robustness to short-term noise, it may dampen sensitivity to clinically meaningful abrupt behavioral changes, such as those observed during acute mood episodes or crisis events. Hybrid strategies such as combining cumulative medians with volatility-sensitive features may better capture both stable patterns and sudden shifts.

Finally, this study did not include broader biological and environmental factors, such as genetic risk, socioeconomic status, family history, adverse childhood experiences, or neurodevelopmental profiles, that are known to influence adolescent mental health [81,85]. The absence of such contextual and historical information may limit the model's ability to capture all relevant sources of variance, potentially constraining predictive accuracy. Future studies should consider incorporating these factors to enhance explanatory power and clinical utility.

Implications and Recommendations

Active data engagement declined markedly by day 14 (Figure 2B), underscoring a key longitudinal feasibility challenge. Future iterations of Mindcraft will therefore incorporate specific design adaptations to reduce burden and sustain engagement. These include shifting from fixed-time prompts to context-aware adaptive sampling triggered by behavioral anomalies detected through passive data; incorporating light gamification (such as streaks) to maintain motivation; and closing the feedback loop by providing personalized behavioral insights and just-in-time recommendations informed by each user's active and passive data. Together, these adaptations aim to shift Mindcraft from a one-way data collection tool to a personalized support platform, thereby improving long-term engagement and feasibility.

Real-time digital phenotyping at a population level can complement traditional screening methods by identifying and prioritizing high-risk individuals and enabling tailored prevention and early intervention strategies [76,86]. The use of a mobile app for digital phenotyping is particularly valuable for children and young people, for whom early identification and intervention are essential to prevent the onset of more severe mental health issues in adulthood. When implemented in schools, it addresses barriers such as stigma and accessibility, offering adolescents a preventive tool that empowers them to manage their mental health. Digital phenotyping provides the opportunity to inform school-based digital interventions that might be central to early intervention and prevention of mental health problems in the community [87]. While adherence to GDPR and secure data protocols provides a legal baseline, the ethical landscape of adolescent digital phenotyping extends far beyond regulatory compliance [88,89]. Collecting continuous passive data from minors introduces distinct tensions regarding autonomy and informed consent; specifically, the invisible nature of passive sensing means adolescents may habituate to the monitoring, potentially eroding their ongoing awareness of data sharing [89]. Furthermore, the deployment of predictive

XSL•FO
RenderX

risk models carries the risk of digital labeling, where algorithmic outputs, if misinterpreted or generating false positives, could lead to stigma, unnecessary anxiety, or oversurveillance [88]. Addressing these complexities demands more than static consent forms; it requires ongoing participatory approaches involving adolescents, parents, and clinicians, alongside governance mechanisms that ensure transparency, prioritize interpretability over black-box predictions, and maintain human clinical oversight of algorithmic outputs [88]. To this end, future work must empirically evaluate how adolescents understand, experience, and respond to continuous passive sensing, and determine how ethical frameworks can best support safe, acceptable integration in school settings.

Achieving real-world feasibility requires distinguishing between scientific interpretability and user-centric explainability [90,91]. While the SHAP values presented in this study provide necessary transparency for model validation, raw feature importance scores are unlikely to be meaningful to non-expert stakeholders such as adolescents, parents, or educators. For broad deployment, these technical outputs must be bridged by a translation layer that converts granular risk estimates into accessible, actionable narratives. For instance, rather than displaying a high SHAP value for "location entropy," a user-facing interface should translate this into an intuitive insight, such as detecting "significant changes in your daily routine." Future implementation work must prioritize the co-design of these explanatory interfaces to ensure that algorithmic transparency supports understanding rather than overwhelming users.

Digital biomarkers from sensors and ML have shown accuracy in predicting disease progression [92,93], underscoring the broader potential of sensor-based technologies for personalized healthcare. Smartphones, being both ubiquitous and affordable, enable continuous, real-time data collection even in low-resource settings, reducing reliance on clinical supervision [34,94]. Digital phenotyping offers a scalable mechanism for continuous, context-aware monitoring [94], which could feed into well-being dashboards accessible to pastoral staff, school counselors, or clinical teams. In school settings, such dashboards could support early identification of distress and enable stepped-care approaches that match support intensity to need. In clinical pathways, risk predictions could support triage decisions and monitoring between appointments, complementing existing services rather than replacing them [86]. Achieving this will require implementation research on workflow integration, governance, and alignment with school well-being policies and child mental health services.

Traditional platforms such as Childline rely on proactive engagement from children and young people, creating barriers for disengaged users. In contrast, Mindcraft's passive tracking capabilities offer a proactive approach by identifying early signs of poor mental health and prompting timely professional interventions. Building on this pilot work, we have developed a school-based intervention study that evaluates the effectiveness of personalized artificial intelligence recommendations delivered through the Mindcraft app across schools in the United Kingdom (ISRCTN11686798). With this further development, Mindcraft is evolving into a comprehensive platform that delivers in-app recommendations informed by active and passive data, leveraging user profiles [95,96] to tailor suggestions to individual needs, enhance engagement, and improve intervention effectiveness. Subsequent phases will focus on scaling and validation across diverse school settings and on integrating Mindcraft with existing education and mental health pathways to support sustainable, real-world implementation. This integration of proactive detection and tailored intervention could help address significant gaps in traditional mental health support systems.

Conclusion

In conclusion, this study underscores the transformative potential of integrating active and passive smartphone data to predict adolescent mental health. By leveraging innovative ML techniques, such as contrastive learning, and the scalability of tools like the Mindcraft app, we present a robust framework for early risk detection across diverse mental health outcomes. These findings lay the groundwork for more inclusive, accessible, and personalized early detection and intervention strategies in adolescent mental health.

Funding

TBB is supported by a Fellowship funded by the Fundación Alicia Koplowitz. MDS is supported by the NIHR Imperial Biomedical Research Collaboration and NIHR Mental Health Translational Research Collaboration. DN is supported by the National Institute for Health Research (NIHR) Applied Research Collaboration Northwest London and NIHR Imperial Biomedical Research Collaboration. AAF acknowledges support from the United Kingdom Research and Innovation Turing AI Fellowship (EP/V025449/1).

Data Availability

The study data are not publicly available because of the General Data Protection Regulation restrictions and privacy policies outlined in the study information provided to participants. The data sets generated during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

DN, MDS, and AAF conceptualized the study. TBB, DN, and MDS gained ethical approval and conducted recruitment and data collection. BK curated the data and performed data analysis and machine learning under the supervision of AAF. BK, TBB, and

XSL•FO
RenderX

AF drafted the original manuscript. MDS, DN, and AAF critically reviewed and edited the manuscript. All authors reviewed and approved the final version of the manuscript for publication.

BK and TBB contributed equally as joint first authors; additionally, DN and AAF contributed equally as joint last authors. DN, the clinical corresponding author, can be reached at d.nicholls@imperial.ac.uk.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Active data features, including feature descriptions and Spearman correlations with mental health outcomes (SDQ, SCI, ED-15, and suicidal ideation).

[DOCX File, 47 KB - jmir_v28i1e72501_app1.docx]

Multimedia Appendix 2

Passive data features engineered from smartphone sensors, including feature descriptions and Spearman correlations with mental health outcomes (SDQ, SCI, ED-15, and suicidal ideation).

[DOCX File, 55 KB - jmir_v28i1e72501_app2.docx]

Multimedia Appendix 3

Example illustrating the behavior of the cumulative median compared to raw daily step count values for a single participant. The cumulative median smooths isolated anomalies (eg, a one-day spike on Day 3) while remaining sensitive to sustained behavioral shifts (eg, a persistent drop beginning on Day 7).

[PNG File, 186 KB - jmir_v28i1e72501_app3.png]

Multimedia Appendix 4

Balanced accuracy (mean [SD] across 10 different runs of the experiment) for models trained on raw daily features, cumulative mean-aggregated and cumulative median–aggregated features of the combined active and passive data for the mental health outcomes.

[DOCX File, 45 KB - jmir_v28i1e72501_app4.docx]

Multimedia Appendix 5

CONSORT (Consolidated Standards of Reporting Trials)-style flow diagram showing participant inclusion for the main analysis. The main analysis included 67 users who provided at least one active self-report and enabled at least one passive data stream.

[PNG File, 174 KB - jmir_v28i1e72501_app5.png]

Multimedia Appendix 6

Distribution of mental health assessment scores across participants. (A) Strengths and Difficulties Questionnaire (SDQ) scores, representing overall mental health and behavioral difficulties. (B) Sleep Condition Indicator (SCI) scores, assessing sleep quality and potential insomnia. (C) Frequency of self-reported suicidal ideations over a 2-week period. (D) Eating Disorder Examination Questionnaire (ED-15) scores, evaluating symptoms associated with eating disorders.

[PNG File, 151 KB - jmir_v28i1e72501_app6.png]

Multimedia Appendix 7

Fully annotated correlation heatmap of active data features, with Spearman correlation coefficients displayed in each cell. This supplementary version is provided to support the complete numerical transparency of the simplified heatmap shown in Figure 3B of the main manuscript.

[PNG File, 539 KB - jmir_v28i1e72501_app7.png]

Multimedia Appendix 8

Mean (SE) *z* score–normalized values of the top five active and passive features that significantly differed between low- and high-risk groups across 4 mental health outcomes: SDQ, insomnia, suicidal ideation, and eating disorder. Group differences were assessed using one-sided t-tests with Benjamini–Hochberg correction for multiple comparisons. Features were selected based on adjusted p-values. Red asterisks indicate statistical significance (*$P \leq .05$, **$P \leq .01$, ***$P \leq .001$). Normalization was applied to enable comparison across features measured on different scales.

[PNG File, 241 KB - jmir_v28i1e72501_app8.png]

XSL•FO
RenderX

Multimedia Appendix 9
Comparison of balanced accuracy of mental health outcome predictions using active data only for all users in the study (N=103) vs users who also enabled passive data collection (N=67).
[PNG File, 156 KB - jmir_v28i1e72501_app9.png]

Multimedia Appendix 10
Detailed performance metrics for predicting SDQ high risk, insomnia, suicidal ideation, and eating disorder risk using combined active and passive data in an external validation sample (N=45 adolescents not used for model development).
[DOCX File, 46 KB - jmir_v28i1e72501_app10.docx]

Multimedia Appendix 11
Fairness analysis across gender and school context for the combined active and passive data model. Balanced accuracy for predicting SDQ high risk, insomnia, suicidal ideation, and eating disorder risk stratified by (A) gender (female vs male/other) and (B–D) school context. Error bars represent SD across 10 different runs of the experiment.
[PNG File, 225 KB - jmir_v28i1e72501_app11.png]

Multimedia Appendix 12
Fairness analysis of the combined active and passive data model across passive-sensor availability. Balanced accuracy for users with less than or equal to 3 enabled sensors and greater than 3 enabled sensors for each mental health outcome. Error bars represent the SD across 10 different runs of the experiment.
[PNG File, 98 KB - jmir_v28i1e72501_app12.png]

Multimedia Appendix 13
Feature importance analysis for predicting the insomnia high-risk category using both active and passive data. (A) SHAP-based feature importances, with passive data aggregated by sensor type and active data shown individually. (B) Distribution of the top five active data features across insomnia risk categories. (C) Distribution of the top five passive data features across insomnia risk categories. Statistically significant differences between low-risk and high-risk groups are indicated (*P<.05, **P<.01, ***P<.001, t test).
[PNG File, 261 KB - jmir_v28i1e72501_app13.png]

Multimedia Appendix 14
Feature importance analysis for predicting the suicidal ideation high-risk category using both active and passive data. (A) SHAP-based feature importances, with passive data aggregated by sensor type and active data shown individually. (B) Distribution of the top five active data features across suicidal ideation risk categories. (C) Distribution of the top five passive data features across suicidal ideation risk categories. Statistically significant differences between low-risk and high-risk groups are indicated (*P<.05, **P<.01, ***P<.001, t test).
[PNG File, 276 KB - jmir_v28i1e72501_app14.png]

Multimedia Appendix 15
Feature importance analysis for predicting the eating disorder high-risk category using both active and passive data. (A) SHAP-based feature importances, with passive data aggregated by sensor type and active data shown individually. (B) Distribution of the top five active data features across eating disorder risk categories. (C) Distribution of the top five passive data features across eating disorder risk categories. Statistically significant differences between low-risk and high-risk groups are indicated (*P<.05, **P<.01, ***P<.001, t test).
[PNG File, 271 KB - jmir_v28i1e72501_app15.png]

References

1. Mughal F, England E. The mental health of young people: the view from primary care. Br J Gen Pract 2016 Oct;66(651):502-503. [doi: 10.3399/bjgp16X687133] [Medline: 27688487]
2. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005 Jun;62(6):593-602. [doi: 10.1001/archpsyc.62.6.593] [Medline: 15939837]
3. Mental health of adolescents. World Health Organization. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health [accessed 2025-05-22]
4. Mathiesen KS, Sanson A, Stoolmiller M, Karevold E. The nature and predictors of undercontrolled and internalizing problem trajectories across early childhood. J Abnorm Child Psychol 2009 Feb;37(2):209-222. [doi: 10.1007/s10802-008-9268-y] [Medline: 18766436]

XSL•FO
RenderX

5. Samek DR, Hicks BM. Externalizing disorders and environmental risk: mechanisms of gene-environment interplay and strategies for intervention. Clin Pract (Lond) 2014;11(5):537-547. [doi: 10.2217/CPR.14.47] [Medline: 25485087]
6. Kim-Cohen J, Caspi A, Moffitt TE, Harrington H, Milne BJ, Poulton R. Prior juvenile diagnoses in adults with mental disorder: developmental follow-back of a prospective-longitudinal cohort. Arch Gen Psychiatry 2003 Jul;60(7):709-717. [doi: 10.1001/archpsyc.60.7.709] [Medline: 12860775]
7. Polanczyk GV, Salum GA, Sugaya LS, Caye A, Rohde LA. Annual research review: a meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. J Child Psychol Psychiatry 2015 Mar;56(3):345-365. [doi: 10.1111/jcpp.12381] [Medline: 25649325]
8. Murray CJL. The global burden of disease study at 30 years. Nat Med 2022 Oct;28(10):2019-2026. [doi: 10.1038/s41591-022-01990-1] [Medline: 36216939]
9. Gulliver A, Griffiths KM, Christensen H. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. BMC Psychiatry 2010 Dec 30;10:1-9. [doi: 10.1186/1471-244X-10-113] [Medline: 21192795]
10. Bonell C, Jamal F, Harden A, et al. Systematic review of the effects of schools and school environment interventions on health: evidence mapping and synthesis. Public Health Research 2013;1(1):1-320. [doi: 10.3310/phr01010]
11. Anderson JK, Ford T, Soneson E, et al. A systematic review of effectiveness and cost-effectiveness of school-based identification of children and young people at risk of, or currently experiencing mental health difficulties. Psychol Med 2019 Jan;49(1):9-19. [doi: 10.1017/S0033291718002490] [Medline: 30208985]
12. Larsen ME, Huckvale K, Nicholas J, et al. Using science to sell apps: evaluation of mental health app store quality claims. NPJ Digit Med 2019;2(1):18. [doi: 10.1038/s41746-019-0093-1] [Medline: 31304366]
13. Donker T, Petrie K, Proudfoot J, Clarke J, Birch MR, Christensen H. Smartphones for smarter delivery of mental health programs: a systematic review. J Med Internet Res 2013 Nov 15;15(11):e247. [doi: 10.2196/jmir.2791] [Medline: 24240579]
14. Eisenstadt M, Liverpool S, Infanti E, Ciuvat RM, Carlsson C. Mobile apps that promote emotion regulation, positive mental health, and well-being in the general population: systematic review and meta-analysis. JMIR Ment Health 2021 Nov 8;8(11):e31170. [doi: 10.2196/31170] [Medline: 34747713]
15. Oudin A, Maatoug R, Bourla A, et al. Digital phenotyping: data-driven psychiatry to redefine mental health. J Med Internet Res 2023 Oct 4;25:e44502. [doi: 10.2196/44502] [Medline: 37792430]
16. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA 2017 Oct 3;318(13):1215-1216. [doi: 10.1001/jama.2017.11295] [Medline: 28973224]
17. Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacol 2016 Jun;41(7):1691-1696. [doi: 10.1038/npp.2016.7]
18. Shin J, Bae SM. A systematic review of location data for depression prediction. Int J Environ Res Public Health 2023 May 29;20(11):5984. [doi: 10.3390/ijerph20115984] [Medline: 37297588]
19. Kandola A, Lewis G, Osborn DPJ, Stubbs B, Hayes JF. Depressive symptoms and objectively measured physical activity and sedentary behaviour throughout adolescence: a prospective cohort study. Lancet Psychiatry 2020 Mar;7(3):262-271. [doi: 10.1016/S2215-0366(20)30034-1] [Medline: 32059797]
20. Brown HE, Pearson N, Braithwaite RE, Brown WJ, Biddle SJH. Physical activity interventions and depression in children and adolescents: a systematic review and meta-analysis. Sports Med 2013 Mar;43(3):195-206. [doi: 10.1007/s40279-012-0015-8] [Medline: 23329611]
21. Obayashi K, Saeki K, Iwamoto J, Ikada Y, Kurumatani N. Exposure to light at night and risk of depression in the elderly. J Affect Disord 2013 Oct;151(1):331-336. [doi: 10.1016/j.jad.2013.06.018] [Medline: 23856285]
22. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annu Rev Clin Psychol 2017 May 8;13(1):23-47. [doi: 10.1146/annurev-clinpsy-032816-044949] [Medline: 28375728]
23. Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health. Psychiatr Rehabil J 2015 Sep;38(3):218-226. [doi: 10.1037/prj0000130] [Medline: 25844912]
24. Rickard N, Arjmand HA, Bakker D, Seabrook E. Development of a mobile phone app to support self-monitoring of emotional well-being: a mental health digital innovation. JMIR Ment Health 2016 Nov 23;3(4):e49. [doi: 10.2196/mental.6202] [Medline: 27881358]
25. Cao J, Truong AL, Banu S, Shah AA, Sabharwal A, Moukaddam N. Tracking and predicting depressive symptoms of adolescents using smartphone-based self-reports, parental evaluations, and passive phone sensor data: development and usability study. JMIR Ment Health 2020 Jan 24;7(1):e14045. [doi: 10.2196/14045] [Medline: 32012072]
26. Faurholt-Jepsen M, Vinberg M, Frost M, Christensen EM, Bardram J, Kessing LV. Daily electronic monitoring of subjective and objective measures of illness activity in bipolar disorder using smartphones--the MONARCA II trial protocol: a randomized controlled single-blind parallel-group trial. BMC Psychiatry 2014 Nov 25;14(1):309. [doi: 10.1186/s12888-014-0309-5] [Medline: 25420431]
27. Berrouiguet S, Ramírez D, Barrigón ML, et al. Combining continuous smartphone native sensors data capture and unsupervised data mining techniques for behavioral changes detection: a case series of the evidence-based behavior (eB2) study. JMIR Mhealth Uhealth 2018 Dec 10;6(12):e197. [doi: 10.2196/mhealth.9472] [Medline: 30530465]

XSL•FO
RenderX

28. Wang R, Aung MS, Abdullah S, et al. CrossCheck: toward passive sensing and detection of mental health changes in people with schizophrenia. 2016 Presented at: Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing; Sep 12-16, 2016. [doi: 10.1145/2971648.2971740]
29. Beiwinkel T, Kindermann S, Maier A, et al. Using smartphones to monitor bipolar disorder symptoms: a pilot study. JMIR Ment Health 2016 Jan 6;3(1):e2. [doi: 10.2196/mental.4560] [Medline: 26740354]
30. Jacobson NC, Chung YJ. Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones. Sensors (Basel) 2020 Jun 24;20(12):3572. [doi: 10.3390/s20123572] [Medline: 32599801]
31. Opoku Asare K, Visuri A, Ferreira DST. Towards early detection of depression through smartphone sensing. 2019 Sep 9 Presented at: UbiComp '19 URL: https://dl.acm.org/doi/proceedings/10.1145/3341162 [doi: 10.1145/3341162.3347075]
32. Dargél AA, Mosconi E, Masson M, et al. Toi Même, a mobile health platform for measuring bipolar illness activity: protocol for a feasibility study. JMIR Res Protoc 2020 Aug 18;9(8):e18818. [doi: 10.2196/18818] [Medline: 32638703]
33. Mullick T, Radovic A, Shaaban S, Doryab A. Predicting depression in adolescents using mobile and wearable sensors: multimodal machine learning-based exploratory study. JMIR Form Res 2022 Jun 24;6(6):e35807. [doi: 10.2196/35807] [Medline: 35749157]
34. Maharjan SM, Poudyal A, van Heerden A, et al. Passive sensing on mobile devices to improve mental health services with adolescent and young mothers in low-resource settings: the role of families in feasibility and acceptability. BMC Med Inform Decis Mak 2021 Apr 7;21(1):117. [doi: 10.1186/s12911-021-01473-2] [Medline: 33827552]
35. Thakur SS, Roy RB. Predicting mental health using smart-phone usage and sensor data. J Ambient Intell Human Comput 2021 Oct;12(10):9145-9161. [doi: 10.1007/s12652-020-02616-5]
36. Wang W, Nepal S, Huckins JF, et al. First-gen lens: assessing mental health of first-generation students across their first year at college using mobile sensing. Proc ACM Interact Mob Wearable Ubiquitous Technol 2022 Jul;6(2):1-32. [doi: 10.1145/3543194] [Medline: 36561350]
37. Acikmese Y, Alptekin SE. Prediction of stress levels with LSTM and passive mobile sensors. Procedia Comput Sci 2019;159:658-667. [doi: 10.1016/j.procs.2019.09.221]
38. Rhim S, Lee U, Han K. Tracking and modeling subjective well-being using smartphone-based digital phenotype. In: Rhim S, Lee U, Han K, editors. 2020 Jul 7 Presented at: UMAP '20 URL: https://dl.acm.org/doi/proceedings/10.1145/3340631 [accessed 2025-12-23] [doi: 10.1145/3340631.3394855]
39. Xu X, Liu X, Zhang H, Wang W, Nepal S, Sefidgar Y, et al. GLOBEM: cross-dataset generalization of longitudinal human behavior modeling. Proc ACM Interact Mob Wearable Ubiquitous Technol 2022;6(4):1-34. [doi: 10.1145/3569485]
40. Currey D, Torous J. Digital phenotyping correlations in larger mental health samples: analysis and replication. BJPsych Open 2022 Jun 3;8(4):e106. [doi: 10.1192/bjo.2022.507] [Medline: 35657687]
41. Xu X, Chikersal P, Doryab A, et al. Leveraging routine behavior and contextually-filtered features for depression detection among college students. Proc ACM Interact Mob Wearable Ubiquitous Technol 2019 Sep 9;3(3):1-33. [doi: 10.1145/3351274]
42. Cornet VP, Holden RJ. Systematic review of smartphone-based passive sensing for health and wellbeing. J Biomed Inform 2018 Jan;77:120-132. [doi: 10.1016/j.jbi.2017.12.008] [Medline: 29248628]
43. Choo M, Park D, Cho M, Bae S, Kim J, Han DH. Exploring a multimodal approach for utilizing digital biomarkers for childhood mental health screening. Front Psychiatry 2024;15:1348319. [doi: 10.3389/fpsyt.2024.1348319]
44. Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. Biol Psychiatry Cogn Neurosci Neuroimaging 2018 Mar;3(3):223-230. [doi: 10.1016/j.bpsc.2017.11.007] [Medline: 29486863]
45. Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. Psychol Med 2019 Jul;49(9):1426-1448. [doi: 10.1017/S0033291719000151] [Medline: 30744717]
46. Goodman R. The strengths and difficulties questionnaire: a research note. J Child Psychol Psychiatry 1997 Jul;38(5):581-586. [doi: 10.1111/j.1469-7610.1997.tb01545.x] [Medline: 9255702]
47. Tatham M, Turner H, Mountford VA, Tritt A, Dyas R, Waller G. Development, psychometric properties and preliminary clinical validation of a brief, session-by-session measure of eating disorder cognitions and behaviors: the ED-15. Int J Eat Disord 2015 Nov;48(7):1005-1015. [doi: 10.1002/eat.22430] [Medline: 26011054]
48. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
49. Fonseca-Pedrero E, Díez-Gómez A, Pérez-Albéniz A, Al-Halabí S, Lucas-Molina B, Debbané M. Youth screening depression: validation of the Patient Health Questionnaire-9 (PHQ-9) in a representative sample of adolescents. Psychiatry Res 2023 Oct;328:115486. [doi: 10.1016/j.psychres.2023.115486] [Medline: 37738682]
50. Espie CA, Kyle SD, Hames P, Gardani M, Fleming L, Cape J. The sleep condition indicator: a clinical screening tool to evaluate insomnia disorder. BMJ Open 2014 Mar 18;4(3):e004183. [doi: 10.1136/bmjopen-2013-004183] [Medline: 24643168]
51. Espie CA, Farias Machado P, Carl JR, et al. The sleep condition indicator: reference values derived from a sample of 200 000 adults. J Sleep Res 2018 Jun;27(3):e12643 [FREE Full text] [doi: 10.1111/jsr.12643]

XSL•FO
RenderX

52. Kadirvelu B, Bellido Bel T, Wu X, et al. Mindcraft, a mobile mental health monitoring platform for children and young people: development and acceptability pilot study. JMIR Form Res 2023 Jun 26;7:e44877. [doi: 10.2196/44877] [Medline: 37358901]
53. Goodman R, Ford T, Simmons H, Gatward R, Meltzer H. Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. Br J Psychiatry 2000 Dec;177(6):534-539. [doi: 10.1192/bjp.177.6.534] [Medline: 11102329]
54. Rodrigues T, Vaz AR, Silva C, Conceição E, Machado PPP. Eating Disorder-15 (ED-15): factor structure, psychometric properties, and clinical validation. Eur Eat Disord Rev 2019 Nov;27(6):682-691. [doi: 10.1002/erv.2694] [Medline: 31257707]
55. Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering. In: Schroff F, Kalenichenko D, Philbin J, editors. 2015 Presented at: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [doi: 10.1109/CVPR.2015.7298682]
56. Hoffer E, Ailon N, editors. Deep metric learning using triplet network. 2015 Presented at: Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015; Oct 12-14, 2015. [doi: 10.1007/978-3-319-24261-3_7]
57. Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification. arXiv. Preprint posted online on 2017. [doi: 10.48550/arXiv.1703.07737]
58. Chen T, Kornblith S, Norouzi M. A simple framework for contrastive learning of visual representations. In: Hinton G, editor. 2020 Presented at: Proceedings of the 37th International Conference on Machine Learning (ICML 2020); Jul 13-18, 2020.
59. Lundberg S. A unified approach to interpreting model predictions. . Preprint posted online on May 22, 2017. [doi: 10.48550/arXiv.1705.07874]
60. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2(1):56-67. [doi: 10.1038/s42256-019-0138-9]
61. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst 2018;31:6638-6648. [doi: 10.5555/3327757.3327770]
62. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. J Big Data 2020;7(1):94. [doi: 10.1186/s40537-020-00369-8] [Medline: 33169094]
63. MacLeod L, Suruliraj B, Gall D, et al. A mobile sensing app to monitor youth mental health: observational pilot study. JMIR Mhealth Uhealth 2021 Oct 26;9(10):e20638. [doi: 10.2196/20638] [Medline: 34698650]
64. Ware S, Yue C, Morillo R, et al. Predicting depressive symptoms using smartphone data. Smart Health (2014) 2020 Mar;15:100093. [doi: 10.1016/j.smhl.2019.100093]
65. LeMoult J, Gotlib IH. Depression: a cognitive perspective. Clin Psychol Rev 2019 Apr;69:51-66. [doi: 10.1016/j.cpr.2018.06.008] [Medline: 29961601]
66. Piguet C, Dayer A, Kosel M, Desseilles M, Vuilleumier P, Bertschy G. Phenomenology of racing and crowded thoughts in mood disorders: a theoretical reappraisal. J Affect Disord 2010 Mar;121(3):189-198. [doi: 10.1016/j.jad.2009.05.006] [Medline: 19515428]
67. Town R, Hayes D, March A, Fonagy P, Stapley E. Self-management, self-care, and self-help in adolescents with emotional problems: a scoping review. Eur Child Adolesc Psychiatry 2024 Sep;33(9):2929-2956. [doi: 10.1007/s00787-022-02134-z] [Medline: 36641785]
68. Di Benedetto M, Towt CJ, Jackson ML. A cluster analysis of sleep quality, self-care behaviors, and mental health risk in Australian University students. Behav Sleep Med 2020;18(3):309-320. [doi: 10.1080/15402002.2019.1580194] [Medline: 30821507]
69. Kurina LM, Knutson KL, Hawkley LC, Cacioppo JT, Lauderdale DS, Ober C. Loneliness is associated with sleep fragmentation in a communal society. Sleep 2011 Nov 1;34(11):1519-1526. [doi: 10.5665/sleep.1390] [Medline: 22043123]
70. Chang AM, Aeschbach D, Duffy JF, Czeisler CA. Evening use of light-emitting eReaders negatively affects sleep, circadian timing, and next-morning alertness. Proc Natl Acad Sci USA 2015 Jan 27;112(4):1232-1237. [doi: 10.1073/pnas.1418490112]
71. Auger RR, Burgess HJ, Dierkhising RA, Sharma RG, Slocumb NL. Light exposure among adolescents with delayed sleep phase disorder: a prospective cohort study. Chronobiol Int 2011 Dec;28(10):911-920. [doi: 10.3109/07420528.2011.619906] [Medline: 22080736]
72. Franklin JC, Ribeiro JD, Fox KR, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol Bull 2017 Feb;143(2):187-232. [doi: 10.1037/bul0000084] [Medline: 27841450]
73. Gallagher M, Prinstein MJ, Simon V, Spirito A. Social anxiety symptoms and suicidal ideation in a clinical sample of early adolescents: examining loneliness and social support as longitudinal mediators. J Abnorm Child Psychol 2014 Aug;42(6):871-883. [doi: 10.1007/s10802-013-9844-7] [Medline: 24390470]
74. Littlewood DL, Kyle SD, Carter LA, Peters S, Pratt D, Gooding P. Short sleep duration and poor sleep quality predict next-day suicidal ideation: an ecological momentary assessment study. Psychol Med 2019 Feb;49(3):403-411. [doi: 10.1017/S0033291718001009] [Medline: 29697037]

XSL•FO
RenderX

75. Pizzoli SFM, Monzani D, Conti L, Ferraris G, Grasso R, Pravettoni G. Issues and opportunities of digital phenotyping: ecological momentary assessment and behavioral sensing in protecting the young from suicide. Front Psychol 2023;14:1103703. [doi: 10.3389/fpsyg.2023.1103703]
76. Melcher J, Hays R, Torous J. Digital phenotyping for mental health of college students: a clinical review. Evid Based Mental Health 2020 Nov;23(4):161-166. [doi: 10.1136/ebmental-2020-300180]
77. Paulus FW, Ohmann S, Möhler E, Plener P, Popow C. Emotional dysregulation in children and adolescents with psychiatric disorders. A narrative review. Front Psychiatry 2021;12:628252. [doi: 10.3389/fpsyt.2021.628252] [Medline: 34759846]
78. Lavender JM, Wonderlich SA, Engel SG, Gordon KH, Kaye WH, Mitchell JE. Dimensions of emotion dysregulation in anorexia nervosa and bulimia nervosa: a conceptual review of the empirical literature. Clin Psychol Rev 2015 Aug;40:111-122. [doi: 10.1016/j.cpr.2015.05.010] [Medline: 26112760]
79. Allison KC, Spaeth A, Hopkins CM. Sleep and eating disorders. Curr Psychiatry Rep 2016 Oct;18(10):1-8. [doi: 10.1007/s11920-016-0728-8] [Medline: 27553980]
80. Yang H, Wang JJ, Tng GYQ, Yang S. Effects of social media and smartphone use on body esteem in female adolescents: testing a cognitive and affective model. Children (Basel) 2020 Sep 21;7(9):148. [doi: 10.3390/children7090148] [Medline: 32967376]
81. Reiss F. Socioeconomic inequalities and mental health problems in children and adolescents: a systematic review. Soc Sci Med 2013 Aug;90:24-31. [doi: 10.1016/j.socscimed.2013.04.026]
82. Salk RH, Hyde JS, Abramson LY. Gender differences in depression in representative national samples: meta-analyses of diagnoses and symptoms. Psychol Bull 2017 Aug;143(8):783-822. [doi: 10.1037/bul0000102] [Medline: 28447828]
83. Zainal NH, Wang V, Garthwaite B, Curtiss JE. What factors are related to engagement with digital mental health interventions (DMHIs)? A meta-analysis of 117 trials. Health Psychol Rev 2025 Aug 26:1-21. [doi: 10.1080/17437199.2025.2547610] [Medline: 40857362]
84. Buolamwini J, Gebru T, editors. Gender shades: intersectional accuracy disparities in commercial gender classification. Presented at: Conference on fairness, accountability and transparency; Feb 23-24, 2018 URL: https://proceedings.mlr.press/v81/buolamwini18a.html [accessed 2026-01-22]
85. Patel V, Flisher AJ, Hetrick S, McGorry P. Mental health of young people: a global public-health challenge. The Lancet 2007 Apr;369(9569):1302-1313. [doi: 10.1016/S0140-6736(07)60368-7]
86. Bufano P, Laurino M, Said S, Tognetti A, Menicucci D. Digital phenotyping for monitoring mental disorders: systematic review. J Med Internet Res 2023 Dec 13;25:e46778. [doi: 10.2196/46778] [Medline: 38090800]
87. Cohen KA, Ito S, Ahuvia IL, et al. Brief school-based interventions targeting student mental health or well-being: a systematic review and meta-analysis. Clin Child Fam Psychol Rev 2024 Sep;27(3):732-806. [doi: 10.1007/s10567-024-00487-2] [Medline: 38884838]
88. Martinez-Martin N, Greely HT, Cho MK. Ethical development of digital phenotyping tools for mental health applications: delphi study. JMIR Mhealth Uhealth 2021 Jul 28;9(7):e27343. [doi: 10.2196/27343] [Medline: 34319252]
89. Wies B, Landers C, Ienca M. Digital mental health for young people: a scoping review of ethical promises and challenges. Front Digit Health 2021;3:697072. [doi: 10.3389/fdgth.2021.697072] [Medline: 34713173]
90. Fleming W, Coutts A, Pochard D, Trivedi D, Sanderson K. Human-centered design and digital transformation of mental health services. JMIR Hum Factors 2025 Aug 11;12:e66040. [doi: 10.2196/66040] [Medline: 40789172]
91. Miller T. Explanation in artificial intelligence: insights from the social sciences. Artif Intell 2019 Feb;267:1-38. [doi: 10.1016/j.artint.2018.07.007]
92. Kadirvelu B, Gavriel C, Nageshwaran S, et al. A wearable motion capture suit and machine learning predict disease progression in Friedreich's ataxia. Nat Med 2023 Jan;29(1):86-94. [doi: 10.1038/s41591-022-02159-6] [Medline: 36658420]
93. Ricotti V, Kadirvelu B, Selby V, et al. Wearable full-body motion tracking of activities of daily living predicts disease trajectory in Duchenne muscular dystrophy. Nat Med 2023 Jan;29(1):95-103. [doi: 10.1038/s41591-022-02045-1] [Medline: 36658421]
94. Khosravi M, Azar G. A systematic review of reviews on the advantages of mHealth utilization in mental health services: a viable option for large populations in low-resource settings. Glob Ment Health (Camb) 2024;11:e43. [doi: 10.1017/gmh.2024.39] [Medline: 38690573]
95. Pieritz S, Khwaja M, Faisal AA. In: Matic A, editor. Personalised Recommendations in Mental Health Apps: The Impact of Autonomy and Data Sharing: Association for Computing Machinery (ACM); 2021. [doi: 10.1145/3411764.3445523]
96. Khwaja M, Pieritz S. In: Matic A, editor. Personality and Engagement with Digital Mental Health Interventions: Association for Computing Machinery (ACM); 2021. [doi: 10.1145/3450613.3456823]

Abbreviations

AUC: area under the receiver operating characteristic curve
CONSORT: Consolidated Standards of Reporting Trials
ED-15: Eating Disorder-15 questionnaire
GDPR: General Data Protection Regulation

XSL•FO
RenderX

LOSO CV: leave-one-subject-out cross-validation
ML: machine learning
MLP: multilayer perceptron
PHQ-9: Patient Health Questionnaire version 9
SCI: Sleep Condition Indicator
SDQ: Strengths and Difficulties Questionnaire
SES: socioeconomic status
SHAP: Shapley Additive Explanations

Edited by J Sarvestan; submitted 11.Feb.2025; peer-reviewed by CY Chen, JH Song; revised version received 04.Dec.2025; accepted 05.Dec.2025; published 04.Feb.2026.

Please cite as:
Kadirvelu B, Bellido Bel T, Freccero A, Di Simplico M, Nicholls D, Faisal AA
Digital Phenotyping for Adolescent Mental Health: Feasibility Study Using Machine Learning to Predict Mental Health Risk From Active and Passive Smartphone Data
J Med Internet Res 2026;28:e72501
URL: https://www.jmir.org/2026/1/e72501
doi:10.2196/72501

© Balasundaram Kadirvelu, Teresa Bellido Bel, Aglaia Freccero, Martina Di Simplico, Dasha Nicholls, A Aldo Faisal. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 4.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Effectiveness of the Headspace App for Improving Sleep: Randomized Controlled Trial

Zoltan Andre Torok, PhD; Larisa Gavrilova, PhD; Amish Patel, BA; Matthew Jason Zawadzki, PhD

Department of Psychological Sciences, University of California, 5200 North Lake Road, Merced, CA, United States

Corresponding Author:
Zoltan Andre Torok, PhD
Department of Psychological Sciences, University of California, 5200 North Lake Road, Merced, CA, United States

Abstract

Background: Improving sleep is critical for optimizing short-term and long-term health. Although in-person meditation training has been shown to impact sleep positively, there is a gap in our understanding of whether apps that teach self-guided meditation are also effective.

Objective: This study aims to test whether Headspace (Headspace, Inc) improves sleep quality, tiredness, sleep duration, and sleep efficiency.

Methods: Staff employees (N=135; mean age 38.1, SD 10.9; 75.0% female; 59.3% non-Hispanic White; 27.1% Hispanic) from a university in California's San Joaquin Valley participated in the study. Participants were randomized to complete 10 minutes of daily meditation via the Headspace app for 8 weeks or waitlist control. Sleep assessments were taken for 4 consecutive days at baseline, and then for 4-day bursts at 2, 5, and 8 weeks after randomization. Sleep quality and subjective sleep duration were assessed each morning with a sleep diary, tiredness was assessed throughout the day using ecological momentary assessment, and objective sleep duration and efficiency were measured using a Fitbit Charge 2.

Results: Both subjective and objective sleep outcomes improved. For subjective sleep outcomes, multilevel modeling revealed that those in the Headspace condition, compared to the control group, reported better sleep quality at sessions 2 (β=0.48, SE=0.12; *P*<.001), 5 (β=0.91, SE=0.13; *P*<.001), and 8 (β=0.69, SE=0.15; *P*<.001) compared to baseline, and a decrease in tiredness at session 5 (β=−0.58, SE=0.19; *P*=.001) compared to baseline, but not at sessions 2 or 8. For objective sleep outcomes, those in the Headspace condition compared to the control group had longer sleep durations at session 5 (β=23.96, SE=12.19; *P*=.04) compared to baseline, but not at sessions 2 or 8. There were no significant effects for sleep efficiency.

Conclusions: This study continues adding to the ever-developing field of mobile health apps by demonstrating that Headspace can positively impact sleep quality, tiredness, and duration.

Trial Registration: ClinicalTrials.gov NCT03652168; https://clinicaltrials.gov/study/NCT03652168

(J Med Internet Res 2026;28:e56287) doi:10.2196/56287

KEYWORDS

Headspace; sleep quality; tiredness; sleep duration; sleep efficiency; ecological momentary assessment; app-based; sleep; subjective; objective; outcome; meditation; randomized controlled trial; RCT; application; self-guided; female; women; United States; mHealth; mobile health; Fitbit; wearable

Introduction

Overview

Training in mindfulness meditation has shown the ability to improve sleep outcomes, such as sleep quality and efficiency [1,2]. Most of this research has tested meditation as traditional in-person mindfulness meditation practices, including mindfulness-based stress reduction and mindfulness-based therapy for insomnia [3,4]. There has been a shift in recent years toward using mobile health (mHealth) apps, such as Headspace (Headspace, Inc) and Calm (Calm.com, Inc), to learn meditation and mindfulness. Therefore, because of their increased popularity, these apps are critical to study as users are seeking these apps to improve their health, including sleep [5]. In addition, they have the potential to be more accessible and available than in-person practices. However, these benefits may come with a trade-off, as these apps are generally self-guided and may not be as effective in teaching meditation or mindfulness practices. App designers often focus on usability and acceptability rather than effectiveness in improving health outcomes [6]. Thus, it is necessary to rigorously examine whether mHealth apps for meditation can effectively improve sleep outcomes. Furthermore, there is a need to research when these positive sleep effects emerge, enabling the ability to track the efficacy of interventions during engagement rather than waiting until a postintervention evaluation period.

XSL•FO
RenderX

Sleep Is a Complex Construct

Sleep is a complex, multidimensional construct that can be measured subjectively and objectively. A well-supported approach to studying sleep identifies 5 dimensions: sleep quality, sleep duration, sleep continuity, sleep timing, and sleepiness [7]. Given that each dimension of sleep can change independently of the others depending on the intervention, it is advisable to measure multiple dimensions. Subjective sleep is self-reported and evaluated using retrospective surveys such as the Pittsburgh Sleep Quality Index and sleep diaries [7]. Objective sleep measurements are assessed using behavioral and physiological technology, and thus are generally outside the participant's control regarding the data collected [7]. Both objective and subjective measures tap into different parts of the sleep experience. For example, several studies have shown modest to null correlations between objective and subjective assessments of the same dimension [8-10]. Therefore, collecting both can provide a more complete picture of how sleep is affected by engaging in mindfulness meditation.

mHealth Apps and Sleep

Research examining the impact of mHealth apps on meditation is in its early stages, with only a few randomized controlled trials testing the efficacy of these apps. Most of this work has been tested using the Calm app. For example, in a sample of adults with sleep disturbance, researchers found that Calm users had reduced self-reported daytime fatigue, sleepiness, and presleep arousal compared to a control group [11]. In a study among employees of a large retail company, researchers found that participants using the Calm app reported decreased daytime sleepiness compared to a control group [12]. Cross-sectional studies indicate that individuals who use Calm report longer and better sleep than those who do not use the app [13].

Yet, there is a dearth of knowledge about whether the Headspace app would also effectively improve sleep outcomes. Headspace is vital to study on its own, given its widespread use. For example, in September 2023 alone, Headspace was downloaded 400,000 times and generated $4 million in revenue [14]. Yet, while Calm is pitched as a multimodal app aimed at improving sleep using techniques that include meditation, Headspace operates from a different perspective. Namely, it was developed to provide step-by-step guidance on the basic principles of mindfulness meditation practice. In this way, Headspace serves as a mHealth corollary to the in-person Mindfulness-Based Stress Reduction training program. Indeed, research has demonstrated that the beneficial changes that occur directly from learning mindfulness meditation, such as decentering [15], acceptance [16], and nonreactivity to inner experience [17,18], are present after using Headspace [19]. As such, it is plausible that improvements in sleep may also ensue after using Headspace. Some initial evidence from quasi-experimental research has shown that children with attention-deficit hyperactivity disorder who used Headspace had reduced self-reported sleep disturbance after a 4-week intervention compared to baseline [20]. However, there is a need to rigorously investigate the impact of Headspace on various measures of sleep to further inform prospective users of its effectiveness.

Additional gaps exist in the current literature on mHealth meditation apps and sleep. Most research has focused on how mHealth meditation apps affect individuals with diagnosable sleep disorders. Yet, the usage rates of mHealth apps indicate that they are being used by the general population, who are likely not meeting clinical levels of sleep disturbance. Yet, poor sleep outcomes are still regularly reported among this population daily. For example, according to the 2020 Behavioral Risk Factor Surveillance System, 33.2% of adults report getting less than 7 hours of sleep per 24 hours [21]. Second, there is a need to test the effectiveness of Headspace using subjective and objective sleep measures to reduce the potential for expectancy effects to influence self-reports, which could account for significant effects. To this end, testing the effects of Headspace in daily life at several time points during the intervention can provide a richer picture of when mHealth apps affect sleep outcomes.

The Present Study

This study tested whether the mHealth app Headspace affects several dimensions of sleep, including self-reported quality via a morning diary and daily tiredness via ecological momentary assessment (EMA), and objectively assessed duration and efficiency via a Fitbit Charge 2 wearable. We hypothesize that the use of a mHealth app will have a positive impact on both sleep quality and sleep duration. We also hypothesize that sleep quality and duration improvements will be recognized early on, possibly by the week 2 assessment. Finally, we do not have a clear hypothesis on how the improvements will last, as there is very little research on how sleep improvements change over time.

Methods

Participants

In this randomized controlled trial, participants were staff employees at a university in the San Joaquin Valley of California. Participants were excluded if they were not university employees or if they were university faculty, younger than 18 years old, not fluent in English, did not have access to a smartphone, or had prior experience in meditation, defined as having participated in mindfulness meditation 2 times per week for 10 minutes over the previous 3 months. To determine the sample size, we conducted a power analysis to compare 2 groups, assuming a moderate effect size, an alpha of .05, and a power of 0.80. The moderate effect size was used, given prior research suggesting such effects could be observed with this kind of mHealth app [22]. The power analysis results indicated that the required sample size to detect the effect was 128 participants. We aimed to enroll 140 participants to account for potential dropouts.

Procedure

Participants were recruited through posted flyers on campus, presentations at departmental and university staff assembly meetings, and word-of-mouth referrals. Interested participants were directed to a secure website, where they read more about the study and were directed to a screener delivered through Qualtrics.

XSL•FO
RenderX

Participants attended an in-lab orientation lasting about 60 minutes. This orientation introduced participants to the assessment procedure. First, they downloaded the RealLife Exp app (Life Data Corporation) on their smartphone and loaded the survey. Participants were guided through a start-up session that demonstrated how to navigate the app and view examples of each question they would encounter during the study. Participants practiced how to answer and were able to ask questions about the process. After practicing, participants were instructed that when a survey was ready, it would appear as a push notification.

To assess subjective sleep quality, participants completed a daily diary upon waking. The survey was available for completion by 6:00 AM and remained open for 4 hours. Participants completed an EMA protocol throughout the day to assess tiredness. EMA surveys were completed randomly within each block: 8:00-10:00 AM, 10:30 AM-12:30 PM, 1:00-3:00 PM, 3:30-5:30 PM, and 6:00-8:00 PM. If participants did not immediately respond to a survey, they had up to 1 hour to complete it, with a reminder notification 20 minutes after the initial prompt. Surveys took between 3 and 4 minutes to complete on average.

Four consecutive days of data were collected via daily diary (quality) and EMA (tiredness). Each orientation took place on a Monday, Tuesday, or Wednesday morning. All data were then scheduled to begin data collection on Wednesday to ensure relative commonality across participants and to ensure both working (Wednesday through Friday) and nonworking (Saturday) days. Participants completed the 4 days of data collection (daily diary and EMA) as part of the baseline session and then at 2, 5, and 8 weeks after the first day of data collection at baseline.

Until this point, the researcher and participant were masked as to their condition. After randomization, both the researcher and the participant were unmasked for the duration of the study. Participants needed to be unmasked as they were instructed to engage with the Headspace app or not. Although researchers were also unmasked, they did not interact with participants outside of standardized messages sent during the study. Per instructions described during the in-lab orientation, participants were only randomized into the Headspace or control condition after the initial 4 days of data collection. Thus, weeks 2, 5, and 8 of data are postrandomization. Allocation to condition occurred at a 2:1 ratio, with more people in the Headspace condition, to account for the potential higher dropout rate in the intervention group as observed in mHealth studies [23]. A random number generator in Microsoft Excel was used to develop the order in which participants would be randomly allocated to a condition. To mimic the experience participants would normally encounter when downloading Headspace, minimal training was provided to participants by research staff. Participants in the Headspace condition were sent an email with download instructions for the app and a code to enter that would grant 12 months of access to Headspace. Headspace provided all the codes but had no say over data collection procedures, analyses, or dissemination. Participants were instructed to use Headspace for 10 minutes daily for the 8-week intervention period. The instructions specified that participants should complete each of the 3 Basic packs (with 10 units each) and then complete the Stress pack (with 30 units). To ensure compliance, we tracked downloads and initial use of the app, ensuring that all participants completed their first session within the first week postrandomization.

For sleep quality assessed via diary, at week 0, there were 462 assessments (mean 4.90, SD 0.56), 418 assessments at week 2 (mean 4.93, SD 0.44), 330 assessments at week 5 (mean 4.85, SD 0.76), and 237 assessments at week 8 (mean 4.85, SD 0.74). For tiredness assessed via EMA, at week 0, there were 2556 assessments (mean 18.66, SD 2.82), 2256 assessments at week 2 (mean 18.05, SD 3.83), 1832 assessments at week 5 (mean 18.87, SD 3.40), and 1326 assessments at week 8 (mean 18.46, SD 4.45). For sleep duration and efficiency assessed via Fitbit at week 0, there were 681 assessments (mean 4.80, SD 0.83), 487 assessments at week 2 (mean 4.87, SD 0.63), 315 assessments at week 5 (mean 4.50, SD 1.32), and 159 assessments at week 8 (mean 4.68, SD 1.17).

Participants in the control condition were given a 1-year subscription to Headspace after completing the 4-month waitlist period. During these 4 months on the waitlist, they were asked not to participate in any mindfulness activities such as yoga or meditation during this time. They completed the same daily diary and EMA protocol as participants in the Headspace condition.

Measures

Baseline Measures

During the baseline assessment, participants first completed demographic information, including their gender (coded for analysis as 0=male, 1=female), age (in years), and ethnicity (coded for analysis as 0=non-Hispanic and Latino, 1=Hispanic and Latino), as well as other measures not relevant to this study.

Daily Diary and EMA Measures

Subjective sleep quality and tiredness were measured using EMA, based on methods from previous research that used EMA to assess physical well-being [24] and developed a standard daily sleep diary [25]. To assess subjective sleep quality, each morning, participants were asked, “How well did you sleep?,” on a scale from 0 (not at all well) to 10 (extremely well). To assess subjective tiredness, participants were asked, “Right now, how tired do you feel?”, rated from 0 (not at all) to 10 (extremely).

Fitbit

A Fitbit Charge 2 wearable was used to estimate sleep duration and efficiency. Fitbit is a popular fitness tracker with a microelectronic triaxial accelerometer to capture body motion in 3-dimensional space. These motion data are then analyzed using proprietary algorithms to identify motion patterns. The Fitbit Charge 2 also monitors heart rate activity through a patented photoplethysmography technology called PurePulse. PurePulse uses light-emitting diodes on the skin-facing surface to continuously estimate heart rate by monitoring blood volume changes [26]. Fitbit estimates steps, calories burned, and sleep through the estimated heart rate. Participants received training on the proper use of the Fitbit and the Fitbit app (Google LLC).

XSL•FO
RenderX

To standardize this, participants were asked to wear the device continuously (including sleep) throughout the study. In addition, they were instructed to sync and charge it on the same days, Tuesday and Sunday, and to ensure the device was placed back on their wrist by the evening of each of these days. Fitabase (Small Steps Labs LLC) was used to extract and process the Fitbit data for the study. Fitabase measures duration as the total number of minutes asleep. Fitabase estimates efficiency by dividing the total time in bed by the total number of minutes asleep. To ensure data were valid, any sleep duration that was less than 2 hours or more than 12 hours was deleted from analyses (although patterns are similar when either or both criteria are not applied), as well as any efficiency scores greater than 100.

Headspace

The intervention was delivered through the commercially available mindfulness meditation app Headspace, widely used in previous intervention studies [27,28]. Headspace provides a variety of formal guided and unguided meditation practices, with instructions delivered through short, animated training videos. The intervention group was instructed to start meditating using the Basic pack. This pack is designed as an introduction to mindfulness meditation and can be used as an opportunity for participants to get familiar with the Headspace teaching style. Each lesson in the Basic pack is approximately 11 minutes and is led by a teacher of the user's choosing. In the first part of the session, the teacher gives the user a short tutorial on the concept of mindfulness. Then, the teacher proceeds with a guided meditation. There are 3 total Basic packs, and each one has 10 sessions, with each session lasting approximately 11 minutes. If participants follow the directions of the study, they will finish all the Basic packs within 30 days. Once participants completed the Basic, they were instructed to move on to the Stress pack, which lasted for 30 sessions. Like the Basic pack, the Stress pack used a teacher and combined visualization and body scanning to help users learn to accept their emotions and pay close attention to the present moment.

Participants were encouraged to complete the meditation quietly for each session without distraction. They were instructed to use the app to meditate for 10 minutes a day for 8 weeks. This duration was chosen based on prior work, which showed that just 10 days of practicing mindfulness for 10 minutes a day successfully reduced stress, negative affect, and improved well-being among a range of sample types [29-32].

Analytic Plan

We used multilevel modeling to account for the 3-level structure of our data, with multiple daily diary and EMA observations nested within assessment sessions (0, 2, 5, or 8) nested within participants. Models were tested using the PROC MIXED command in SAS v.9.4 (SAS Institute Inc), which tests a linear mixed model analysis. We used restricted maximum likelihood estimates to handle missing data, which is the recommended approach given its robustness in addressing missing data that is often nested (eg, a participant who dropped out in session 8 will have missing data for all diary and EMA observations) [33]. This procedure does not impute missing data but uses available data to calculate maximum likelihood estimates.

Analyses followed an intention-to-treat approach, assuming those randomized to the condition completed the intervention material as instructed [34]. As such, we did not factor in specific measures of usage of Headspace in analyses (an issue we return to in the "Discussion"). To test our main research question, we included the following variables as predictors: session, condition (0=control group, 1=Headspace group), and the interaction of session by condition. In the presented results, the terms labeled session refer to the effects of the control condition at that session relative to baseline. The interaction term of session by condition refers to the effects of comparing the Headspace to the control group at that session relative to baseline. Sleep was the outcome variable; each dimension of sleep was tested in a separate model. The session was modeled as a categorical variable to interpret whether the sleep data from sessions 2, 5, and 8 were statistically different than the session 0 data.

Finally, at the between-person level, models controlled for gender and age, and at the within-person level, models controlled for the day of the week as either a nonweekend (0) or weekend (1) day.

Ethical Considerations

The local Institutional Review Board approved all the study procedures and measures (IRB #UCM2018). This study was also registered on clinicaltrials.gov (NCT03652168). There were no deviations from the preregistered protocol, and the analyses presented are those for the secondary set of outcomes proposed. Data from the primary set of outcomes testing the effect of Headspace on mechanisms of mindfulness have previously been reported [19]. The participant completed informed consent at the initial intake session. Participants were reminded at all sessions that they could refrain from answering any questions they wished or could opt out of study procedures without penalty. All participants were given a study code to allow deidentification in the datasets. As compensation, participants received a 1-year subscription to Headspace. For each weekly survey (weeks 0, 2, 5, and 8), participants received US $15. In addition, participants could receive up to a US $20 bonus for a high completion rate (ie, over 80% of surveys completed) across the study.

Results

Table 1 provides descriptive statistics for demographics and the sleep variables. The final sample (N=135) was between the ages of 21 and 65 years old and self-identified as primarily White or Hispanic, and female. Participants in the control condition (mean 40.88, SD 10.84, t_{137}= 2.02; P=.05) were slightly older than participants in the Headspace condition (mean 36.91, SD 10.84, t_{137}=2.02; P=.05). There were similar proportions of participants identified as Hispanic and female across the 2 conditions, χ^2=2.29; P=.13, and χ^2=0.28; P=.60, respectively. On average, participants reported sleeping moderately well each morning and feeling somewhat tired. They were recorded as sleeping about 5 and a half hours a night, and their sleep efficiency was below normal.

XSL•FO
RenderX

Table . Descriptive statistics for sleep outcomes.

Variable	Headspace group	Control group	Overall
Age (years), mean (SD)	36.91 (10.84)	41.12 (10.71)	38.19 (10.94)
Ethnicity, n (%)			
American Indian or Alaska Native	1 (1.04)	2 (4.76)	3 (2.17)
Asian	7 (7.29)	4 (9.52)	11 (7.97)
Black	3 (3.13)	1 (2.38)	4 (2.90)
Hispanic	30 (31.25)	8 (19.05)	38 (27.54)
Pacific Islander	2 (2.08)	1 (2.28)	3 (2.17)
White	57 (59.38)	26 (61.90)	83 (60.14)
Gender, n (%)			
Women	74 (77.08)	30 (71.43)	104 (75.36)
Men	22 (22.92)	12 (28.57)	34 (24.64)
Sleep quality, mean (SD)	5.87 (1.40)	5.31 (1.29)	5.68 (1.40)
Tiredness, mean (SD)	4.08 (1.65)	4.51 (1.44)	4.19 (1.56)
Total minutes asleep, mean (SD)	329.54 (63.70)	312.28 (44.34)	322.78 (71.52)
Efficiency, mean (SD)	74.74 (9.33)	75.03 (8.30)	73.93 (10.32)
Total time in bed, mean (SD)	449.36 (50.51)	417.13 (57.22)	447.70 (70.22)

Figure 1 provides a diagram of the process used to determine the final participant number for the study. An initial 291 employees were screened, with 271 eligible. Of the 271 eligible participants, 186 people provided informed consent. Before the study began, 43 participants dropped out, mainly citing time demands. One hundred forty-three participants were randomized, but 8 additional participants were dropped for not having any data. Randomization was completed using a random number generator via Excel to create the order in which participants would be enrolled and allocated to a condition.

At week 0, our sample consisted of 135 participants, with 93 randomized into the Headspace group and 42 into the waitlist control group using a planned 2:1 allocation strategy. At week 2, 87 (93.5% of the original Headspace group) and 40 (95.2% of the original control group) participants completed at least one daily diary and EMA survey. At week 5, 68 (71.5% of the original Headspace group) and 35 (83.3% of the original control group) participants completed at least one daily diary and EMA survey. At week 8, 49 (54.3% of the original Headspace group) and 26 (61.9% of the original control group) participants completed at least one daily diary and EMA survey.

We tested whether participants in the Headspace condition experienced improved subjective and objective sleep outcomes compared to those in the control condition, examining when these effects emerged over the 8-week intervention period.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram of participant flow.

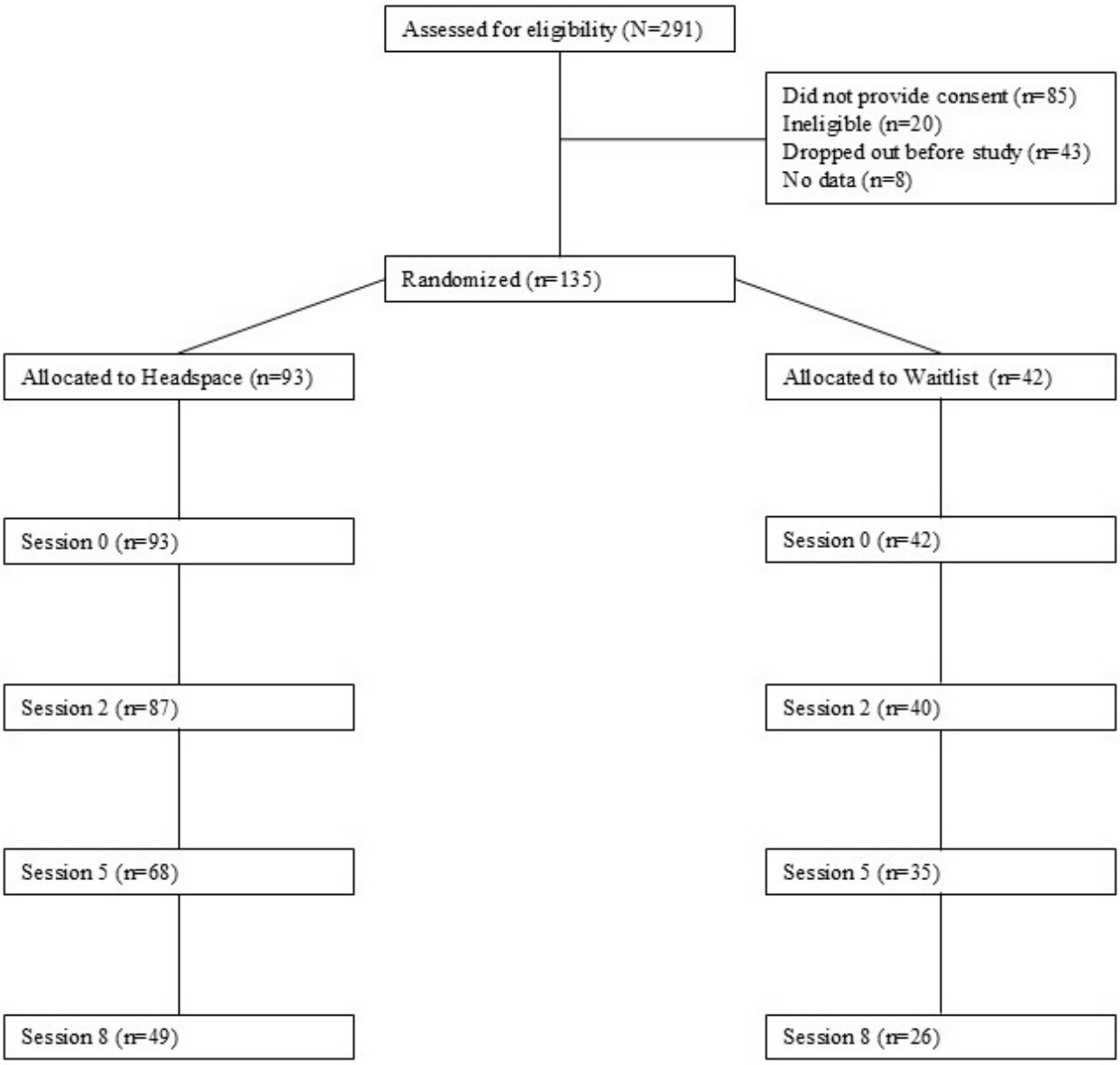

We tested quality and tiredness for the subjective sleep outcomes with results reported in Table 2. The effects for the control group over time are represented as the session terms, whereas the effects for the Headspace condition compared to the control group over time are represented as session x Headspace terms. For quality, compared to the baseline session, those in the control condition reported worse sleep at session 5 (P<.001), with no differences at sessions 2 (P=.41) or 8 (P=.99). There were significant interaction effects of condition by session. Those in the Headspace condition, compared to the control group, reported better sleep at sessions 2 (P<.001), 5 (P<.001), and 8 (P<.001), with each session compared to baseline. For tiredness, those in the control condition reported, compared to the baseline session, an increase in tiredness at session 5 (P=.04), but not at sessions 2 (P=.85) or 8 (P=.79). There was a significant interaction effect of condition by session for session 5. Compared to the control group, those in the Headspace condition reported a decrease in tiredness at session 5 (P=.001), but no difference at sessions 2 (P=.97) or 8 (P=.97), with each session compared to baseline.

XSL•FO
RenderX

Table . Unstandardized beta estimates (SEs) for subjective sleep outcomes with Headspace, session, and interaction as predictors.

Subjective sleep outcomes	Subjective quality		Subjective tiredness	
	β (SE)	*P* value	β (SE)	*P* value
Random effects				
Intercept	1.63[a] (0.21)[a]	<.001	2.29[a] (0.30[a])	<.001
Residual	3.33[a] (0.06)[a]	<.001	4.41[a] (0.09[a])	<.001
Fixed effects				
Intercept	5.08[a] (0.51[a])	<.001	4.76[a] (0.61[a])	<.001
Female	−0.40 (0.26)	.12	−0.16 (0.31)	.61
Age	0.01 (0.01)	.18	−0.003 (0.01)	.79
Weekend	0.35[a] (0.05[a])	<.001	−0.32[a] (0.07[a])	<.001
Session 2	−0.08 (0.10)	.41	−0.03 (0.13)	.85
Session 5	−0.41[a] (0.11[a])	<.001	0.30[a] (0.15[a])	.04
Session 8	−0.002 (0.12)	.99	−0.04 (0.16)	.79
Headspace	0.24 (0.25)	.34	−0.34 (0.31)	.27
Session 2x Headspace	0.48[a] (0.12[a])	<.001	0.01 (0.16)	.97
Session 5x Headspace	0.91[a] (0.13[a])	<.001	−0.58[a] (0.18[a])	.001
Session 8x Headspace	0.69[a] (0.15[a])	<.001	0.01 (0.19)	.97
Model effects				
Pseudo r^2	0.034	—[b]	0.017	—

[a]Coefficients significant at $P<.05$.

[b]Not applicable.

For the objective sleep outcomes, we tested duration and efficiency with results reported in Table 3. For duration, compared to the baseline session, those in the control condition had shorter sleep durations at session 5 (*P*=.03) compared to baseline, but not at sessions 2 (*P*=.94) or 8 (*P*=.20). There was a significant interaction effect of condition by session for session 5. Compared to the control group, those in the Headspace condition had longer sleep durations at session 5 (*P*=.04), but no difference at sessions 2 (*P*=.59) or 8 (*P*=.88), with each session compared to baseline. For sleep efficiency, compared to the baseline session, those in the control condition had a higher efficiency at sessions 2 (*P*=.04) and 5 (*P*=.002) compared to baseline, but not at session 8 (*P*=.33). There were no significant interaction effects of condition by session for sessions 2 (*P*=.37), 5 (*P*=.36), and 8 (*P*=.11).

XSL•FO
RenderX

Table . Unstandardized beta estimates (SEs) for objective sleep outcomes with Headspace, session, and interaction as predictors.

Objective sleep outcomes	Objective duration		Objective efficiency	
	β (SE)	*P* value	β (SE)	*P* value
Random effects				
Intercept	2404.70[a] (571.52[a])	<.001	4.19[a] (0.99[a])	<.001
Residual	4011.98[a] (151.83[a])	<.001	9.95[a] (0.41[a])	<.001
Fixed effects				
Intercept	353.93[a] (32.26[a])	<.001	89.70[a] (1.39[a])	<.001
Female	29.27 (15.04)	.05	0.80 (0.65)	.21
Age	0.57 (0.67)	.40	−0.04 (0.03)	.12
Weekend	11.20[a] (4.12[a])	.007	−0.37 (0.23)	.11
Session 2	−0.59 (8.16)	.94	0.98[a] (0.46[a])	.04
Session 5	−23.27[a] (10.55[a])	.03	1.77[a] (0.58[a])	.002
Session 8	−14.37 (11.21)	.20	0.59 (0.60)	.33
Headspace	23.32 (16.81)	.17	−0.69 (0.74)	.35
Session 2x Headspace	−5.26 (9.68)	.59	−0.49 (0.54)	.37
Session 5x Headspace	23.96[a] (12.19[a])	.04	−0.61 (0.66)	.36
Session 8x Headspace	2.12 (14.58)	.88	1.26 (0.79)	.11
Model effects				
Pseudo r^2	0.058	—[b]	0.050	—

[a]Coefficients significant at $P<.05$.

[b]Not applicable.

Discussion

Principal Findings

This study used daily diaries, EMA, and Fitbit wearables to investigate the impact of Headspace on several dimensions of sleep. Our findings indicate that Headspace had the most significant effect on sleep quality, with some impact on tiredness and sleep duration, and no significant effects on sleep efficiency. Notably, improvements in subjective sleep (quality and tiredness) and objective sleep outcomes (duration) did not occur simultaneously but rather at different times during the study. These findings suggest the continued viability of using mHealth meditation apps to improve sleep and the likelihood that subjective and objective benefits may not happen in tandem. Furthermore, they are consistent with previous research on the Calm app [11,13] that illustrates sleep is multidimensional, with some dimensions improving while others fail to reach significance.

Sleep Quality

The most consistent finding of this study was the rapid improvement in sleep quality. Participants reported better sleep quality within 2 weeks of using Headspace. This finding is impactful because there is emerging evidence that sleep quality is considered a superior sleep index to sleep quantity for assessing sleep [35]. Getting enough sleep is crucial for psychological and physical health, as there are negative consequences when sleep duration decreases past a certain amount of time [36]. What is interesting about sleep quality is that several studies have demonstrated that even if a person gets a whole night's sleep, they can still report their sleep quality as poor and say they were unsatisfied with the night's sleep [37-39]. These findings suggest that objective sleep measures do not fully account for subjective sleep quality. Studies have shown that sleep quality is a good indicator of psychological and overall health [40-42]. This also demonstrates that a person's perceived efficacy of an intervention (which we suggest is as essential a dimension as objective measures) is a significant component that impacts adherence to the intervention [43]. As a future research direction, testing how the perceived efficacy of Headspace impacts adherence is of specific interest.

The improvements in sleep quality continued throughout the study. This promising finding shows that Headspace can alter a person's thinking outside of novel effects. This may not seem profound, but if one considers that it is not uncommon for a person to claim an intervention changed them, then to only renounce that claim quickly thereafter. This is an example of the novelty effect, a well-known phenomenon further discussed below [44-46]. The fact that Headspace users reported continued improvements in sleep quality for 8 weeks indicates that this finding is more than likely not because of novelty and that Headspace does alter a person's quality of sleep positively. This study did not investigate possible mechanisms; however, it has been suggested that meditation may improve sleep quality by

XSL•FO
RenderX

reducing sleep-interfering cognitive processes [47], altering sleep architecture [48], changing perceptions of stress [49], and morphometric and connectivity alterations in sleep-related brain regions [50,51]. Taken together, these findings raise many unanswered questions, fueling future investigation.

Intervention Length

A surprising finding is the convergence of improvements that peaked at week 5 in both subjective measurements, sleep quality and tiredness, and in one of the objective measurements, duration. This finding is novel in suggesting that there may be a previously unidentified point in time at which effectiveness peaks. It has been shown that 8 weeks, but not 4 weeks, of meditation is needed to see benefits in sleep quality and mood [52]. Equally, in non–sleep-related studies, 8 weeks of meditation increased brain structure and function [53] and mental health in university students [54], while 12 weeks of meditation improved depression, anxiety, and stress [55]. However, these studies measure 1 time point or compare 2 time points and cannot accurately pinpoint when effects emerge.

Although the 5-week time point may seem brief, research has shown the potential of mindfulness to have an acute positive effect even after a single use [56]. By allowing users to engage in meditation precisely when needed, such as before bed on a stressful day, and to use it frequently throughout the day in small but frequent doses (thus allowing microdose training), mHealth apps may accelerate the accumulation process. Indeed, other work in this sample found changes to the mindfulness mechanisms of acceptance, attention, and nonreactivity in as little as 2 weeks [19]. More broadly, these findings show the benefit of this study design to test for effectiveness during the engagement with Headspace and not simply wait for a postintervention evaluation period. This approach could be expanded in future studies to determine, with greater detail, the acute and longer-term effects of Headspace in each dimension of sleep. In the long term, this additional information will hopefully replace the “do ten minutes a day of Headspace” with a more nuanced approach.

It was unexpected that Headspace’s effects appeared to wane after 5 weeks. As suggested by other research, it was assumed that sleep benefits would continue throughout the 8-week intervention [51]. A possible explanation for this waning could be that the improvement was a placebo effect. A recent study showed that a sham meditation could increase state mindfulness, mindful observing, state decentering, and mindful nonreacting, similar to an actual meditation condition [57]. Yet, simply chalking up findings to a placebo effect seems unsatisfying. The continuous measurement of sleep daily, within and across days, would suggest an incredibly robust and omnipresent effect. Moreover, it is unclear why the placebo would wear off after 5 weeks. Nevertheless, future research should continue to probe how expected benefits from meditation may prime certain short-term perceived benefits.

There is another, broader hypothesis that may better explain the post-5-week waning: the well-known phenomenon that user engagement with mHealth apps decreases over time. For example, among users measured, 53% uninstall the app within 30 days [58]. Although determining the exact reasons for waning or discontinuing is challenging, research has shown that the main reasons are a loss of interest or declining motivation, a lack of desired features, and the app not being fun [59,60]. Ultimately, the novelty effect significantly influences user abandonment; users may be drawn to the new app out of curiosity but lose interest once it wears off [44-46]. Finally, attrition is a common phenomenon in longitudinal studies, with the literature suggesting that rates from 30% to 70% are not uncommon [61,62].

Of concern is that this pattern of waning, discontinued use, or attrition was observed despite the initial robust effects. The ability of these apps to run autonomously allows for their incredible reach [63]. However, without continued user involvement and self-management, behavioral changes and health improvements tend to dissipate [64]. Therefore, there is a need to continue developing features and behavior change techniques that sustain usage levels for long-term maintenance. In a systematic review of sleep apps, feedback and monitoring were used most often as a behavior change technique [65]. Yet, a recent meta-analysis identified 7 main areas that impact user engagement, with personalization—the ability to make technology act in a particular way depending on user preferences—being the element that users cited as most influencing their engagement [66]. Thus, there may be a disconnect between what app developers pursue as features to implement and what techniques may benefit long-term outcomes. App designers often focus on usability and acceptability rather than effectiveness in improving health outcomes [6]. Due to the increasing shift toward the use of mHealth apps and the vast number of apps available, user engagement is likely to become more critical as app developers strive to attract users. This will further the field of behavior change techniques and is a possible direction for future research.

Limitations and Future Directions

A limitation of this study is that we were unable to track Headspace usage accurately and reliably. Therefore, we have yet to determine whether people in the Headspace group were meditating and are only relying on random assignments to the experimental condition to drive expected changes (ie, intention-to-treat analyses). In this way, we may be underestimating the effect of Headspace, as there may be nonusers in the Headspace group. In addition, tracking engagement would have allowed us to identify a type of person who was more likely to continue completing the intervention and to tailor the intervention content better to these individuals. Research from clinical psychology has shown that younger age, lower education, poorer social problem-solving, lower levels of persistence, and greater avoidance coping are factors of noncompletion [67]. Although not all these factors apply to the current intervention, it demonstrates distinct factors when examining noncompletion and why it is essential to track engagement.

As with most mHealth experimental research, attrition was a limitation, which may have led to the selection of certain traits related to study completion and sleep behaviors. For example, a recent meta-analysis found that the pooled attrition rate in young adults across 15 mHealth trials was 26% [67], and another

XSL•FO
RenderX

found a pooled attrition rate of 43% [22]. This study's attrition rate aligns with the attrition rates identified in these meta-analyses. Therefore, resolving or limiting attrition remains an unresolved issue, and addressing this would greatly benefit research aimed at determining interventions to reduce attrition.

Conclusion

This study demonstrated that Headspace can improve sleep, showing that sleep dimensions do not all change with the same amount of meditation. Some sleep dimensions improved quickly, while others took longer to change. There was a significant convergence of benefits at the 5-week time point and increased attrition. These convergences uncovered 2 critical pieces of information: that there might be a lower-than-expected threshold of meditation needed to experience sleep benefits and that future research using mHealth apps must include factors that improve adherence and engagement. The current study showed that mHealth apps, specifically Headspace, are viable for those seeking to enhance sleep with minimal life disruption. The results from this study can be used by future scientists wanting to learn how to dose mHealth meditation interventions better and practitioners looking to use mHealth meditation to improve patient sleep outcomes.

Acknowledgments

Portions of this work were presented at the American Psychosomatic Society conference in March 2023 [68].

Funding

Funding for data collection was partially supported by a seed grant from the Healthy Campus Network at the University of California, San Francisco (PI: JZ). ZAT's time was supported by the National Heart, Lung, and Blood Institute under the award number 3R15HL161681-01S1 (PI: JZ).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: ZAT, LG, MJZ
Data curation: LG, MJZ
Formal analysis: ZAT, MJZ
Funding acquisition: ZAT, MJZ
Investigation: ZAT, LG, MJZ
Methodology: LG, MJZ
Project administration: LG, MJZ
Resources: MJZ
Supervision: MJZ
Validation: MJZ
Visualization: MJZ
Writing – original draft: ZAT, LG, AP, MJZ
Writing – review & editing: ZAT, AP, MJZ

Conflicts of Interest

LG is now employed at Service Titan. No conflicts of interest were declared for the remaining authors.

Checklist 1
CONSORT checklist.
[DOCX File, 33 KB - jmir_v28i1e56287_app1.docx]

References

1. Kanen J, Nazir R, Sedky K, Pradhan B. The effects of mindfulness-based interventions on sleep disturbance: a meta-analysis. . 2015(2) p. 105-115. [doi: 10.2174/2210676605666150311222928]
2. Chen TL, Chang SC, Hsieh HF, Huang CY, Chuang JH, Wang HH. Effects of mindfulness-based stress reduction on sleep quality and mental health for insomnia patients: a meta-analysis. J Psychosom Res 2020 Aug;135:110144. [doi: 10.1016/j.jpsychores.2020.110144] [Medline: 32590218]
3. Ong JC, Manber R, Segal Z, Xia Y, Shapiro S, Wyatt JK. A randomized controlled trial of mindfulness meditation for chronic insomnia. Sleep 2014 Sep 1;37(9):1553-1563. [doi: 10.5665/sleep.4010] [Medline: 25142566]

XSL•FO
RenderX

4. Rusch HL, Rosario M, Levison LM, et al. The effect of mindfulness meditation on sleep quality: a systematic review and meta-analysis of randomized controlled trials. Ann N Y Acad Sci 2019 Jun;1445(1):5-16. [doi: 10.1111/nyas.13996] [Medline: 30575050]
5. Lee E, Han S, Jo SH. Consumer choice of on-demand mHealth app services: context and contents values using structural equation modeling. Int J Med Inform 2017 Jan;97:229-238. [doi: 10.1016/j.ijmedinf.2016.10.016] [Medline: 27919381]
6. Chib A, Lin SH. Theoretical advancements in mHealth: a systematic review of mobile apps. J Health Commun 2018;23(10-11):909-955. [doi: 10.1080/10810730.2018.1544676] [Medline: 30449261]
7. Buysse DJ. Sleep health: can we define it? Does it matter? Sleep 2014 Jan 1;37(1):9-17. [doi: 10.5665/sleep.3298] [Medline: 24470692]
8. Argyropoulos SV, Hicks JA, Nash JR, et al. Correlation of subjective and objective sleep measurements at different stages of the treatment of depression. Psychiatry Res 2003 Sep 30;120(2):179-190. [doi: 10.1016/s0165-1781(03)00187-2] [Medline: 14527649]
9. Armitage R, Trivedi M, Hoffmann R, Rush AJ. Relationship between objective and subjective sleep measures in depressed patients and healthy controls. Depress Anxiety 1997;5(2):97-102. [doi: 10.1002/(sici)1520-6394(1997)5:2<97::aid-da6>3.0.co;2-2] [Medline: 9262940]
10. Rotenberg VS, Indursky P, Kayumov L, Sirota P, Melamed Y. The relationship between subjective sleep estimation and objective sleep variables in depressed patients. Int J Psychophysiol 2000 Sep;37(3):291-297. [doi: 10.1016/s0167-8760(00)00110-0] [Medline: 10858574]
11. Huberty JL, Green J, Puzia ME, et al. Testing a mindfulness meditation mobile app for the treatment of sleep-related symptoms in adults with sleep disturbance: a randomized controlled trial. PLoS One 2021;16(1):e0244717. [doi: 10.1371/journal.pone.0244717] [Medline: 33411779]
12. Huberty JL, Espel-Huynh HM, Neher TL, Puzia ME. Testing the pragmatic effectiveness of a consumer-based mindfulness mobile app in the workplace: randomized controlled trial. JMIR Mhealth Uhealth 2022 Sep 28;10(9):e38903. [doi: 10.2196/38903] [Medline: 36169991]
13. Huberty J, Puzia ME, Larkey L, Vranceanu AM, Irwin MR. Can a meditation app help my sleep? A cross-sectional survey of Calm users. PLoS One 2021;16(10):e0257518. [doi: 10.1371/journal.pone.0257518] [Medline: 34679078]
14. Headspace revenue & app download estimates from sensor tower - Apple app store. Headspace Inc. URL: https://app.sensortower.com/ios/publisher/publisher/384434796 [accessed 2024-10-03]
15. Shapiro SL, Carlson LE, Astin JA, Freedman B. Mechanisms of mindfulness. J Clin Psychol 2006 Mar;62(3):373-386. [doi: 10.1002/jclp.20237] [Medline: 16385481]
16. Brown KW, Ryan RM. The benefits of being present: mindfulness and its role in psychological well-being. J Pers Soc Psychol 2003 Apr;84(4):822-848. [doi: 10.1037/0022-3514.84.4.822] [Medline: 12703651]
17. Baer RA, Carmody J, Hunsinger M. Weekly change in mindfulness and perceived stress in a mindfulness-based stress reduction program. J Clin Psychol 2012 Jul;68(7):755-765. [doi: 10.1002/jclp.21865] [Medline: 22623334]
18. Bishop SR. What do we really know about mindfulness-based stress reduction? Psychosom Med 2002;64(1):71-83. [doi: 10.1097/00006842-200201000-00010] [Medline: 11818588]
19. Gavrilova L, Zawadzki MJ. Examining how headspace impacts mindfulness mechanisms over an 8-week app-based mindfulness intervention. Mindfulness (N Y) 2023 Sep;14(9):2236-2249. [doi: 10.1007/s12671-023-02214-4]
20. Fried R, DiSalvo M, Farrell A, Biederman J. Using a digital meditation application to mitigate anxiety and sleep problems in children with ADHD. J Atten Disord 2022 May;26(7):1033-1039. [doi: 10.1177/10870547211025616] [Medline: 34865550]
21. Pankowska MM, Lu H, Wheaton AG, et al. Prevalence and geographic patterns of self-reported short sleep duration among US adults, 2020. Prev Chronic Dis 2023 Jun 29;20:E53. [doi: 10.5888/pcd20.220400] [Medline: 37384831]
22. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567. [doi: 10.2196/14567] [Medline: 31573916]
23. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020 Sep 29;22(9):e20283. [doi: 10.2196/20283] [Medline: 32990635]
24. Zawadzki MJ, Hussain M, Kho C. Comparing multidimensional facets of stress with social, emotional, and physical well-being using ecological momentary assessment among a Hispanic sample. Stress Health 2022 Apr;38(2):375-387. [doi: 10.1002/smi.3098] [Medline: 34494721]
25. Carney CE, Buysse DJ, Ancoli-Israel S, et al. The consensus sleep diary: standardizing prospective sleep self-monitoring. Sleep 2012 Feb 1;35(2):287-302. [doi: 10.5665/sleep.1642] [Medline: 22294820]
26. Fitbit charge: wireless activity wristband product manual. Fitbit Inc.: N.p. Fitbit, Inc. n.d.
27. Bostock S, Crosswell AD, Prather AA, Steptoe A. Mindfulness on-the-go: effects of a mindfulness meditation app on work stress and well-being. J Occup Health Psychol 2019 Feb;24(1):127-138. [doi: 10.1037/ocp0000118] [Medline: 29723001]
28. Champion L, Economides M, Chandler C, ed. The efficacy of a brief app-based mindfulness intervention on psychosocial outcomes in healthy adults: a pilot randomised controlled trial. PLoS One 2018;13(12):e0209482. [doi: 10.1371/journal.pone.0209482] [Medline: 30596696]

XSL•FO
RenderX

29. Economides M, Martman J, Bell MJ, Sanderson B. Improvements in stress, affect, and irritability following brief use of a mindfulness-based smartphone app: a randomized controlled trial. Mindfulness (N Y) 2018;9(5):1584-1593. [doi: 10.1007/s12671-018-0905-4] [Medline: 30294390]
30. Foley T, Lanzillotta-Rangeley J. Stress reduction through mindfulness meditation in student registered nurse anesthetists. AANA J 2021 Aug;89(4):284-289. [Medline: 34342565]
31. Zollars I, Poirier TI, Pailden J. Effects of mindfulness meditation on mindfulness, mental well-being, and perceived stress. Curr Pharm Teach Learn 2019 Oct;11(10):1022-1028. [doi: 10.1016/j.cptl.2019.06.005] [Medline: 31685171]
32. Schwartz JE, Stone AA. Strategies for analyzing ecological momentary assessment data. Health Psychol 1998;17(1):6-16. [doi: 10.1037/0278-6133.17.1.6]
33. Altman DG. Missing outcomes in randomized trials: addressing the dilemma. Open Med 2009 May 12;3(2):e51-e53. [Medline: 19946393]
34. Kohyama J. Factors affecting the quality of sleep in children. Children (Basel) 2021 Jun 12;8(6):499. [doi: 10.3390/children8060499] [Medline: 34204755]
35. Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem: National Academies Press; 2006:11617. [doi: 10.17226/11617]
36. Faerman A, Kaplan KA, Zeitzer JM. Subjective sleep quality is poorly associated with actigraphy and heart rate measures in community-dwelling older men. Sleep Med 2020 Sep;73:154-161. [doi: 10.1016/j.sleep.2020.04.012] [Medline: 32836083]
37. Kaplan KA, Hardas PP, Redline S, Zeitzer JM, Sleep Heart Health Study Research Group. Correlates of sleep quality in midlife and beyond: a machine learning analysis. Sleep Med 2017 Jun;34:162-167. [doi: 10.1016/j.sleep.2017.03.004] [Medline: 28522086]
38. Kaplan KA, Hirshman J, Hernandez B, et al. When a gold standard isn't so golden: lack of prediction of subjective sleep quality from sleep polysomnography. Biol Psychol 2017 Feb;123:37-46. [doi: 10.1016/j.biopsycho.2016.11.010] [Medline: 27889439]
39. Kohyama J. Which is more important for health: sleep quantity or sleep quality? Children (Basel) 2021 Jun 24;8(7):542. [doi: 10.3390/children8070542] [Medline: 34202755]
40. Lao XQ, Liu X, Deng HB, et al. Sleep quality, sleep duration, and the risk of coronary heart disease: a prospective cohort study with 60,586 adults. J Clin Sleep Med 2018 Jan 15;14(1):109-117. [doi: 10.5664/jcsm.6894] [Medline: 29198294]
41. Moore PJ, Adler NE, Williams DR, Jackson JS. Socioeconomic status and health: the role of sleep. Psychosom Med 2002;64(2):337-344. [doi: 10.1097/00006842-200203000-00018] [Medline: 11914451]
42. Al-Hashmi I, Hodge F, Nandy K, Thomas E, Brecht ML. The effect of a self-efficacy-enhancing intervention on perceived self-efficacy and actual adherence to healthy behaviours among women with gestational diabetes mellitus. Sultan Qaboos Univ Med J 2019;18(4):e513-e319. [doi: 10.18295/squmj.2018.18.04.014]
43. Bai S, Hew KF, Huang B. Does gamification improve student learning outcome? Evidence from a meta-analysis and synthesis of qualitative data in educational contexts. Educational Research Review 2020 Jun;30:100322. [doi: 10.1016/j.edurev.2020.100322]
44. Sailer M, Sailer M. Gamification of in‐class activities in flipped classroom lectures. Brit J Educational Tech 2021 Jan;52(1):75-90. [doi: 10.1111/bjet.12948]
45. Schmidt-Kraepelin M, Thiebes S, Stepanovic S, Mettler T, Sunyaev A. Gamification in health behavior change support systems - a synthesis of unintended side effects. Presented at: 52nd Hawaii International Conference on System Sciences (HICSS 2019); Jan 8-11, 2019.
46. Winbush NY, Gross CR, Kreitzer MJ. The effects of mindfulness-based stress reduction on sleep disturbance: a systematic review. Explore (NY) 2007;3(6):585-591. [doi: 10.1016/j.explore.2007.08.003] [Medline: 18005910]
47. Nagendra RP, Maruthai N, Kutty BM. Meditation and its regulatory role on sleep. Front Neurol 2012;3:54. [doi: 10.3389/fneur.2012.00054] [Medline: 22529834]
48. Omidi A, Zargar F. Effects of mindfulness-based stress reduction on perceived stress and psychological health in patients with tension headache. J Res Med Sci 2015 Nov;20(11):1058-1063. [doi: 10.4103/1735-1995.172816] [Medline: 26941809]
49. Fox KCR, Nijeboer S, Dixon ML, et al. Is meditation associated with altered brain structure? A systematic review and meta-analysis of morphometric neuroimaging in meditation practitioners. Neurosci Biobehav Rev 2014 Jun;43:48-73. [doi: 10.1016/j.neubiorev.2014.03.016] [Medline: 24705269]
50. Hasenkamp W, Barsalou LW. Effects of meditation experience on functional connectivity of distributed brain networks. Front Hum Neurosci 2012;6:38. [doi: 10.3389/fnhum.2012.00038] [Medline: 22403536]
51. Basso JC, McHale A, Ende V, Oberlin DJ, Suzuki WA. Brief, daily meditation enhances attention, memory, mood, and emotional regulation in non-experienced meditators. Behav Brain Res 2019 Jan 1;356:208-220. [doi: 10.1016/j.bbr.2018.08.023] [Medline: 30153464]
52. Kral TRA, Schuyler BS, Mumford JA, Rosenkranz MA, Lutz A, Davidson RJ. Impact of short- and long-term mindfulness meditation training on amygdala reactivity to emotional stimuli. Neuroimage 2018 Nov 1;181:301-313. [doi: 10.1016/j.neuroimage.2018.07.013] [Medline: 29990584]
53. Cassar L, Fischer M, Valero V. Keep calm and carry on: the short- vs. long-run effects of mindfulness meditation on (academic) performance. 2022.

XSL•FO
RenderX

54. Nikkhah Ravari O, Mousavi SZ, Babak A. Evaluation of the effects of 12 weeks mindfulness-based stress reduction on glycemic control and mental health indices in women with diabetes mellitus type 2. Adv Biomed Res 2020;9(1):61. [doi: 10.4103/abr.abr_133_20] [Medline: 33457344]
55. Johnson KT, Merritt MM, Zawadzki MJ, Di Paolo MR, Ayazi M. Cardiovascular and affective responses to speech and anger: proactive benefits of a single brief session of mindfulness meditation. J Appl Biobehavioral Res 2019 Sep;24(3) [FREE Full text] [doi: 10.1111/jabr.12167]
56. Davies JN, Colagiuri B, Sharpe L, Day MA. Placebo effects contribute to brief online mindfulness interventions for chronic pain: results from an online randomized sham-controlled trial. Pain 2023 Oct 1;164(10):2273-2284. [doi: 10.1097/j.pain.0000000000002928] [Medline: 37310492]
57. The uninstall threat: 2020 app uninstall benchmarks. AppsFlyer. URL: https://www.appsflyer.com/infograms/2019-app-uninstall-benchmarks/ [accessed 2023-12-13]
58. Attig C, Franke T. I track, therefore I walk – exploring the motivational costs of wearing activity trackers in actual users. Int J Hum Comput Stud 2019 Jul;127:211-224. [doi: 10.1016/j.ijhcs.2018.04.007]
59. Krebs P, Duncan DT. Health app use among US mobile phone owners: a national survey. JMIR Mhealth Uhealth 2015 Nov 4;3(4):e101. [doi: 10.2196/mhealth.4924] [Medline: 26537656]
60. Miller RB, Wright DW. Detecting and correcting attrition bias in longitudinal family research. J Marriage Fam 1995 Nov;57(4):921. [doi: 10.2307/353412]
61. Tambs K, Rønning T, Prescott CA, et al. The Norwegian Institute of Public Health twin study of mental health: examining recruitment and attrition bias. Twin Res Hum Genet 2009 Apr;12(2):158-168. [doi: 10.1375/twin.12.2.158] [Medline: 19335186]
62. Ben-Zeev D, Scherer EA, Gottlieb JD, et al. mHealth for schizophrenia: patient engagement with a mobile phone intervention following hospital discharge. JMIR Ment Health 2016 Jul 27;3(3):e34. [doi: 10.2196/mental.6348] [Medline: 27465803]
63. Bakker D, Rickard N. Engagement in mobile phone app for self-monitoring of emotional wellbeing predicts changes in mental health: MoodPrism. J Affect Disord 2018 Feb;227:432-442. [doi: 10.1016/j.jad.2017.11.016] [Medline: 29154165]
64. Arroyo AC, Zawadzki MJ. The implementation of behavior change techniques in mHealth apps for sleep: systematic review. JMIR Mhealth Uhealth 2022 Apr 4;10(4):e33527. [doi: 10.2196/33527] [Medline: 35377327]
65. Wei Y, Zheng P, Deng H, Wang X, Li X, Fu H. Design features for improving mobile health intervention user engagement: systematic review and thematic analysis. J Med Internet Res 2020;22(12):e21687. [doi: 10.2196/21687] [Medline: 33295292]
66. McMurran M, Huband N, Overton E. Non-completion of personality disorder treatments: a systematic review of correlates, consequences, and interventions. Clin Psychol Rev 2010 Apr;30(3):277-287. [doi: 10.1016/j.cpr.2009.12.002] [Medline: 20047783]
67. Prior E, Dorstyn D, Taylor A, Rose A. Attrition in psychological mHealth interventions for young people: a meta-analysis. J technol behav sci 2023;9(4):639-651. [doi: 10.1007/s41347-023-00362-x]
68. Torok Z, Gavrilova L, Patel A, Zawadzki M. Does headspace improve sleep quality in daily life: a daily diary study of app-based mindfulness meditation and sleep. Presented at: American Psychosomatic Society 80th Annual Scientific Meeting Challenging the Future: Towards a Better Biopsychosocial Health; Mar 8-11, 2023 URL: https://thesbsm.org/wp-content/uploads/2023/03/APS-2023-Abstract-Book.pdf [accessed 2026-02-02]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
EMA: ecological momentary assessment
mHealth: mobile health

Edited by TDA Cardoso; submitted 11.Jan.2024; peer-reviewed by D Wilson, SAA Massar, T Nakamura; revised version received 02.Jul.2025; accepted 07.Jul.2025; published 04.Feb.2026.

Please cite as:
Torok ZA, Gavrilova L, Patel A, Zawadzki MJ
The Effectiveness of the Headspace App for Improving Sleep: Randomized Controlled Trial
J Med Internet Res 2026;28:e56287
URL: https://www.jmir.org/2026/1/e56287
doi:10.2196/56287

© Zoltan Andre Torok, Larisa Gavrilova, Amish Patel, Matthew Jason Zawadzki. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 4.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and

XSL•FO
RenderX

reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Associations Between Smartphone-Based Finger Tapping and Cognitive Performance in Older Adults: Observational Study

Huitong Ding[1,2], PhD; Taylor A Orwig[3], MS; Jian Rong[4], PhD; Heaven Y Tatere[3], MD; Chunyu Liu[5], PhD; Eric Schramm[6], PharmD; Chathurangi H Pathiravasan[7], PhD; Joanne M Murabito[2,8*], MD, ScM; Honghuang Lin[3*], PhD

[1]Department of Anatomy and Neurobiology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, United States

[2]The Framingham Heart Study, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, United States

[3]Department of Medicine, University of Massachusetts Chan Medical School, 55 Lake Avenue North, Worcester, MA, United States

[4]Department of Neurology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, United States

[5]Department of Biostatistics, Boston University School of Public Health, Boston, MA, United States

[6]Care Evolution, Ann Arbor, MI, United States

[7]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

[8]Department of Medicine, Section of General Internal Medicine, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, United States

[*]these authors contributed equally

Corresponding Author:
Honghuang Lin, PhD
Department of Medicine, University of Massachusetts Chan Medical School, 55 Lake Avenue North, Worcester, MA, United States

Abstract

Background: Finger tapping tasks assess fine motor control and have been proposed as potential markers of cognitive function. With smartphones widely available, these tasks can be easily administered at home or in other nonclinical settings. However, the relationship between smartphone-based finger tapping measures and cognitive performance is still not well understood.

Objective: This study aimed to examine the association between smartphone digital finger tapping features and cognitive performance in an aging population.

Methods: Participants were enrolled in the study as part of the electronic Framingham Heart Study. They were instructed to perform a 2-finger tapping task every 2 months over a 1-year period. In each session, they tapped with 2 fingers of the left hand for 10 seconds and then with 2 fingers of the right hand for 10 seconds, for a total of 20 seconds. Global cognition performance and 4 cognitive domains, including memory, executive function, language, and visuospatial function, were assessed using a standardized neuropsychological battery. Fifteen tapping features were extracted to capture aspects of motor performance, including the mean, SD, skewness, and slope of the intertap interval (ITI). Associations between tapping-derived features and cognitive performance were assessed using linear regression models, adjusting for age, sex, handedness, education, cohort, and the time interval between cognitive and tapping assessments.

Results: A total of 302 participants (mean age 74.7, SD 6.3 years; n=169, 56% female participants, and n=40, 13.2% non-White participants) completed the digital finger tapping tasks. Eleven tapping features such as basic temporal properties (eg, number of taps and mean ITI), temporal variability (eg, SD and coefficient of variation of ITI, ITI range, and Microfluctuation Index), and fatigue/temporal drift (ITI slope) were significantly associated with global cognitive function (all $P<.001$). Each 1 SD increase in the number of taps was associated with a 0.14-unit higher global cognitive function score ($\beta=.14$, 95% CI 0.07-0.21). Furthermore, finger tapping features were significantly associated with multiple cognitive domains, with the greatest number of associations observed for executive function (11 significant features), including a strong association with mean ITI ($\beta=-0.27$, 95% CI −0.33 to −0.20). Stratified analyses by hand showed consistent effect directions across both hands. The aggregated composite scores derived from finger tapping features demonstrated significant associations with their respective cognitive test scores, with partial correlation coefficients ranging from 0.21 for memory function to 0.41 for global cognitive function.

Conclusions: Digital finger tapping features, particularly those reflecting ITI variability, are significantly associated with cognitive performance. These findings suggest that finger tapping tasks may serve as noninvasive, scalable tools for assessing cognitive performance. Further research is warranted to validate their use for monitoring cognitive health in aging populations.

(J Med Internet Res 2026;28:e82463) doi:10.2196/82463

XSL•FO
RenderX

KEYWORDS

finger tapping test; smartphone; digital features; cognitive performance; cohort

Introduction

Motor performance has emerged as a promising indicator of cognitive aging. Growing evidence suggests its potential utility as an early marker of cognitive decline [1]. While gait-related metrics have been extensively explored and consistently associated with impairments in memory, executive function, and visuospatial function [2,3], there is increasing recognition that motor biomarkers should also include upper-limb function. Recent studies have suggested that abnormalities in upper-limb motor function, such as slower movement speed and irregular rhythm, may reflect early neurodegenerative changes and have been associated with mild cognitive impairment (MCI) and dementia [4-6].

Among upper-limb assessments, the finger tapping test (FTT) is a simple and widely used measure of psychomotor speed, motor coordination, and central nervous system integrity [4,7]. Early versions of the FTT used a telegraph key device [8]. Later adaptations moved to devices such as the computer mouse [9] and keyboard [10]. With the widespread availability of smartphones, the FTT can now be administered on touchscreen interfaces [11,12]. This approach is low-cost and scalable. And it allows fine motor function to be assessed outside traditional clinical settings. Smartphones are also well-suited for population-level digital phenotyping because they can support at-home assessments that are more convenient and accessible for diverse and underserved groups. In addition, digital platforms capture high-resolution temporal data. This makes it possible to derive a wide range of tapping features that may help identify subtle motor changes related to early neurodegenerative processes.

Neuropsychological testing remains the gold standard for assessing domain-specific cognitive performance including memory and executive function [13]. However, high-resolution digital motor measures may provide complementary information about brain-behavior relationships. Examining how fine motor measures derived from widely accessible technologies relate to cognitive performance could improve the way we characterize cognitive aging and support the development of scalable digital biomarkers. Linking accessible, real-world digital tools with rigorous neuropsychological assessment is an important step toward more precise evaluation of cognitive function in older adults and may ultimately contribute to earlier identification of subtle changes.

In this study, we examined the relationship between smartphone-based finger tapping performance and multiple cognitive domains in participants from the Framingham Heart Study (FHS), a well-characterized multigenerational cohort. We hypothesized that poorer tapping performance, as captured by digital tapping-derived metrics, would be associated with lower cognitive performance. By linking these digital motor measures with standard neuropsychological tests, our findings may help support the use of smartphone-based tapping tasks as scalable tools for characterizing cognitive function in aging.

Methods

Study Population

This study included Offspring and Omni 1 cohort participants from the FHS, a longitudinal and multigenerational cohort study initiated in 1948 [14-16]. Offspring participants were enrolled from 1971 to 1975 (n=5124), and the multiethnic Omni 1 participants were enrolled from 1994 to 1998 (n=506) to reflect the increasing diversity of the city of Framingham, MA. Both cohorts were examined in the FHS research center every 4 to 8 years using the same protocols. Eligible participants were invited to enroll in the electronic FHS (eFHS) at the time of the last research exam (Offspring examination 10 and Omni 1 examination 5, 2019 - 2022). English-speaking participants who owned a smartphone (iPhone with iOS 10 or higher or Android version 7 or higher) were invited to download the MyDataHelps app, which included a number of surveys and tasks. A research assistant assisted participants with app download in the research center and provided written instructions if participants preferred to download the app at home. Further details of the eFHS protocol have been previously described [17]. A total of 620 participants enrolled in the eFHS study. Of these, 9 participants who enrolled using a computer and 133 participants who enrolled using an Android device were excluded, as the 2-finger tapping test was only available on the iPhone platform. Among the remaining participants, 445 completed at least one 2-finger tapping test. For the present analysis, we further restricted the sample to individuals who had completed at least one 2-finger tapping assessment and at least one neuropsychological assessment within a 5-year window. This additional criterion yielded a final analytic sample of 302 participants. Participants provided electronic consent within the eFHS mobile app and gave written consent for the parent FHS at each in-person exam visit and neuropsychological testing session. Figure 1 presents the graphical abstract of this study.

Cognitive and clinical characteristics were assessed using standardized procedures. Cognitive status was adjudicated by a review committee that included at least 1 neurologist and 1 neuropsychologist [18]. Dementia was diagnosed according to the criteria outlined in the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition [19]. MCI was defined as cognitive decline in one or more domains without functional impairment and not meeting criteria for dementia [20]. Details of the FHS diagnostic procedures have been described in a prior study [21]. Parkinson disease (PD) status was determined using 3 sources: (1) self-report to the question “have you been told by a doctor you have PD?” (2) examiner-administered health history interviews; and (3) responses to the question “for what medical conditions do you use cannabis?” where PD was listed as an option. Self-rated health was measured using a question from the self-administered Short Form-12 Health Survey. In this survey, participants rated their overall health on a 4-point scale ranging from “poor” to “excellent” [22].

XSL•FO
RenderX

Figure 1. Graphical abstract.

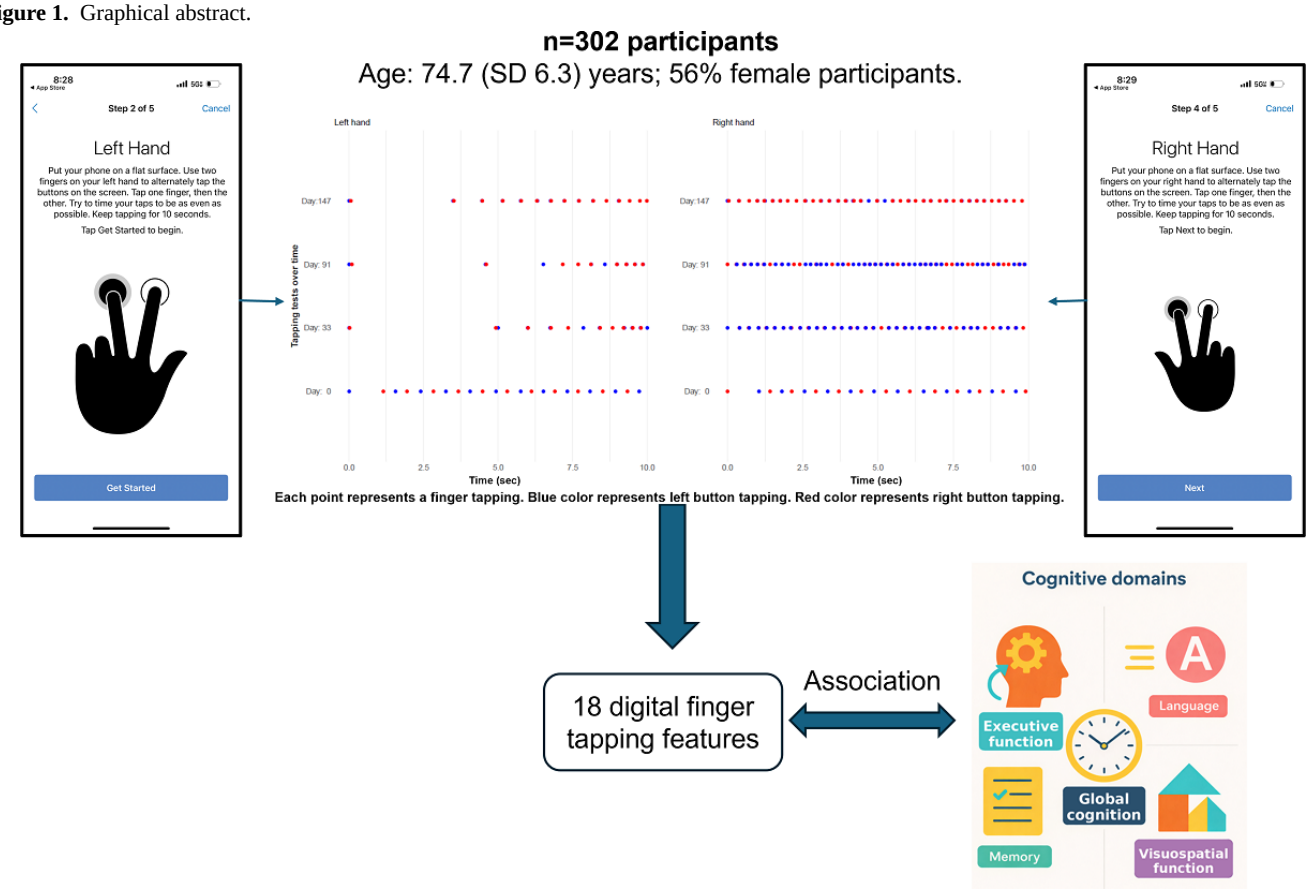

Ethical Considerations

This research received ethical approval from the institutional review board of Boston University Medical Center (H-32132). Participants provided electronic consent within the eFHS mobile app and gave written consent for the parent FHS at each in-person exam visit and neuropsychological testing session. The original informed consent and institutional review board approval allowed secondary analysis of the data without requiring additional participant consent. All data analyzed in this study were deidentified. No financial compensation was provided to participants.

Two-Finger Tapping Test and Feature Extraction

As part of the eFHS Offspring and Omni 1 protocol, participants were sent the 2-finger tapping test at enrollment and then every 2 months over the course of 1 year. The 2-finger tap task informed participants that the task measured tapping speed and instructed participants to put their phone on a flat surface to use 2 fingers to alternately tap the buttons on the screen while trying to time the taps to be as even as possible (Figure S1 in Multimedia Appendix 1, screenshot of task). Each test session lasted 20 seconds, during which participants alternated tapping on designated areas of the smartphone screen with 2 fingers of the left hand for 10 seconds, followed by 2 fingers of the right hand for 10 seconds, as quickly and consistently as possible. Data were included if the app recorded a valid 10-second trial with at least 1 tap event. We did not set additional performance-based exclusion thresholds, such as a minimum number of taps. Our goal was to capture the full range of tapping behavior, including slow, irregular, or instruction-nonadherent tapping, because these patterns may reflect cognitive function. A few examples of finger tapping results are shown in Figure S2 in Multimedia Appendix 1. A comprehensive set of features was extracted from the intertap interval (ITI) data to capture various aspects of motor performance. (1) Basic temporal features included the total number of taps per session as an indicator of effort, endurance, and overall speed, the mean ITI as measures of central tendency, and the SD of ITI as a measure of temporal variability. We also calculated the coefficient of variation of ITI to account for differences in overall tapping speed. (2) Asymmetry features included the alternating tap ratio (the proportion of left-right alternating taps), and the mean and SD of ITI for each finger. Skewness and kurtosis of the ITI distribution were also calculated to capture irregular tapping patterns. (3) Consistency metrics included ITI range. This feature reflected the stability of performance. (4) Fatigue and temporal drift were characterized using the ITI slope (the change in ITI over time during the trial) and the last-to-first ITI ratio. These features indicated whether tapping slowed or changed over the course of the session. (5) Microfluctuation feature included the Microfluctuation Index. This index was defined as the SD of successive ITI differences to capture subtle motor instability.

Neuropsychological Assessment

The methodology for administering neuropsychological tests in FHS has been previously described [23]. Cognitive testing was first introduced in 1976, through brief cognitive screenings and expanded to a more comprehensive neuropsychological battery beginning in 1999 [13,24]. In the present study, we used 21 commonly administered neuropsychological tests to evaluate

XSL•FO
RenderX

cognitive function across 4 domains (Table S1 in Multimedia Appendix 1) [25]. To address missing values of neuropsychological tests, we used multiple imputation by chained equations [26]. Distributions of the Hooper Visual Organization Test, Trail Making Test part A, and Trail Making Test part B were normalized using natural logarithmic transformation. To aid interpretation, the transformed scores were directionally adjusted so that higher values reflected better cognitive performance. All neuropsychological test scores were subsequently standardized to *z*-scores (mean 0, SD 1). These neuropsychological test scores were then summarized into 4 cognitive domains by averaging *z*-scores of neuropsychological tests within each domain: memory, executive function, language, and visuospatial function. A global cognitive score was further derived based on the mean values of all 4 cognitive domain scores.

Statistical Analyses

Linear regression models were used to assess the associations between finger tapping performance and cognitive domain scores. For each participant, the mean value of each finger tapping feature across all available trials was calculated and used as the independent variable. Global cognitive function served as the primary dependent variable. All finger tapping features were standardized to *z*-scores to enable comparability across metrics. Models were adjusted for age at the time of the tapping assessment, sex, education, cohort (Offspring or Omni 1), handedness, and the time interval (in years) between the neuropsychological assessment and the baseline tapping test. As secondary analyses, we further explored associations with 4 specific cognitive domains: memory, executive function, language, and visuospatial function.

To further characterize hand-specific relationships, we performed stratified analyses for right- and left-hand tapping data. Sensitivity analyses were also conducted within the subset of participants identified as right-handed to reduce heterogeneity due to hand preference. Within this group, we reassessed the associations between each finger tapping metric and cognitive domain scores, additionally stratifying by the hand used during the tapping assessment (right vs left). In addition, we grouped the trials according to whether they were performed with the dominant or nondominant hand and repeated the association analyses within these groups. To explore the impact of noncompliant trials, we conducted a sensitivity analysis excluding trials (3729/4277, 12.8%) with an alternating tap ratio less than 0.80. To evaluate the potential impact of practice effects arising from different numbers of completed sessions, we performed a further sensitivity analysis to examine the associations between finger tapping measures and cognitive domain scores using each participant's first 3 sessions. Bonferroni adjustment was used to correct multiple testing, and a significance threshold was determined with $P<.05/15=.003$, with the total number of finger tapping features. As an exploratory analysis, we examined whether cohort modified the association between finger tapping measures and cognition. In each linear regression model, we included an interaction term between the tapping feature and cohort.

We further used the least absolute shrinkage and selection operator (LASSO) regression to derive a composite score for global cognitive function based on finger tapping features. Specifically, we implemented a 5-fold cross-validation loop. For each training set, an inner 5-fold cross-validation was used to select the optimal regularization parameter (α). The composite score was defined as a weighted linear combination of the tapping features. The weights were the LASSO-derived coefficients within each training fold. The fitted model was then applied to the corresponding held-out test fold to obtain out-of-sample predicted global cognitive scores. Then, we assessed the association between the composite score and global cognitive function using partial correlation analysis adjusting for age, sex, and education. In addition, domain-specific composite tapping scores were constructed for each cognitive domain following the same procedure. All analyses were performed using *R* (version 4.0.3; R Core Team) or Python (version 3.6.8; Python Software Foundation).

Results

Cohort Descriptive

This study included 302 eFHS participants (mean age 74.7, SD 6.3 years; n=169, 56%, were female participants, n=192, 63.6%, had college degree or higher, and n=40, 13.2%, were non-White participants; Table 1). As expected, the majority (247/302, 81.8%) were right-handed, while 11.9% (36/302) were left-handed and 5% (15/302) were ambidextrous. On average, participants had completed 7 (SD 4) FTT sessions with both hands administered. Neurologically, the sample was generally healthy—the prevalence of PD was 0.3% (1/302). The average Mini-Mental State Exam (MMSE) score was 28.8 (SD 1.3), with 2.3% (7/302) of participants meeting criteria for MCI and 0.3% (1/302) for dementia. Self-rated health status was favorable—74.1% (224/302) of participants rated their general health as "very good" or "excellent."

XSL•FO
RenderX

Table . Demographics of the study participants[a].

Variable	Total (N=302)	Global cognitive performance		
		First tertile (n=101)	Second tertile (n=100)	Third tertile (n=101)
Age (y), mean (SD)	74.7 (6.3)	76.0 (7.0)	75.0 (6.0)	73.3 (5.5)
Sex, n (%)				
Females	169 (56)	52 (51.5)	52 (52)	65 (64.4)
Males	133 (44)	49 (48.5)	48 (48)	36 (35.6)
Omni cohort, n (%)	40 (13.2)	27 (26.7)	7 (7)	6 (5.9)
Education, n (%)				
High school	30 (9.9)	15 (14.9)	10 (10)	5 (5)
Some college	80 (26.5)	39 (38.6)	23 (23)	18 (17.8)
College and higher	192 (63.6)	47 (46.5)	67 (67)	78 (77.2)
Handedness, n (%)				
Right	247 (81.8)	81 (80.2)	84 (84)	82 (81.2)
Left	36 (11.9)	14 (13.9)	11 (11)	11 (10.9)
Ambidextrous	15 (5)	5 (5)	4 (4)	6 (5.9)
Unknown/missing	4 (1.3)	1 (1)	1 (1)	2 (2)
Time interval between the finger tapping task and the neuropsychological tests, mean (SD)	1.7 (1.2)	1.7 (1.2)	1.6 (1.1)	1.8 (1.3)
Number of finger tapping test session administered, mean (SD)	7 (4)	6 (4)	7 (4)	8 (3)
MMSE[b], mean (SD)	28.8 (1.3)	28.1 (1.6)	29.0 (1.1)	29.4 (0.8)
Prevalent MCI[c], n (%)	7 (2.3)	7 (6.9)	0	0
Prevalent dementia, n (%)	1 (0.3)	1 (1)	0	0
Prevalent Parkinson disease, n (%)	1 (0.3)	0	1 (1)	0
General health SF[d]-12, n (%)				
Poor	1 (0.3)	1 (1)	0	0
Fair	10 (3.3)	6 (5.9)	2 (2)	2 (2)
Good	67 (22.2)	26 (25.7)	25 (25)	16 (15.8)
Very good	146 (48.3)	45 (44.6)	47 (47)	54 (53.5)
Excellent	78 (25.8)	23 (22.8)	26 (26)	29 (28.7)

[a]Categorical variables are presented as counts (proportions). Continuous variables are presented as mean (SD).

[b]MMSE: Mini-Mental State Exam.

[c]MCI: mild cognitive impairment.

[d]SF: Short Form.

When stratified by tertiles of global cognitive function, participants in the third tertile (n=101) were younger (mean age 73.3 vs 76 years in the first tertile), more likely to be female participants (n=65, 64.4% vs n=52, 51.5%), and more highly educated (n=78, 77.2% with college or higher vs n=47, 46.5% in the first tertile). The proportion of participants from the Omni cohort (representing greater racial/ethnic diversity) was higher in the first tertile (n=27, 26.7%) compared to the highest (n=6, 5.9%). MMSE scores followed the expected gradient, increasing from 28.1 (SD 1.6) in the first tertile to 29.4 (SD 0.8) in the third tertile. Similarly, MCI and dementia cases were observed only in the first tertile group. Participants in the third tertile also completed more FTTs on average (17 vs 13). This suggested possible differences in adherence or engagement.

XSL•FO
RenderX

Participants who completed the digital finger tapping assessment were younger and more highly educated compared to those in the broader Offspring/Omni 1 cohort at Exam 10/5. Additional demographic comparisons with the full eFHS and Offspring/Omni 1 samples are presented in Table S2 in Multimedia Appendix 1. The distributions of finger tapping features and cognitive domain scores, including 25th percentile, median, and 75th percentile values, are shown in Tables S3 and S4, respectively, in Multimedia Appendix 1.

Association Between Digital Finger Tapping Features and Cognitive Performance

Among the 15 finger tapping features evaluated, 11 were significantly associated with global cognitive function (*P* values<.001; Table 2). The strongest associations were observed for tapping variability metrics derived from ITI, such as SD, range, and Microfluctuation Index. For example, a 1 SD increase in mean ITI (average time differences between successive taps) was associated with a 0.23 decrease in global cognitive function score (β=−.23; 95% CI −0.30 to −0.17; *P*<.001). In addition to variability-based features, performance-related features such as number of taps (β=.14; 95% CI 0.07-0.21) were positively associated with global cognitive function.

Table . The association between digital finger tapping features and global cognitive function[a].

Group and finger tapping features	Description	β (95% CI)	*P* value
Basic temporal features			
Number of taps	Indicator of effort, endurance, and speed	0.14 (0.07 to 0.21)	<.001
Mean ITI[b]	Average time differences between successive taps	–0.23 (–0.30 to –0.17)	<.001
SD of ITI	Measures tapping variability	–0.22 (–0.28 to –0.15)	<.001
CV[c] of ITI	CV of ITI	–0.14 (–0.20 to –0.07)	<.001
Asymmetry features			
Skewness of ITI	Skewness of ITI	0.00 (–0.07 to 0.06)	.95
Kurtosis of ITI	Kurtosis of ITI	0.03 (–0.04 to 0.09)	.38
Alternating tap ratio	Fraction of alternating left-right taps	0.05 (–0.01 to 0.12)	.11
Mean of left ITI	Mean of tap times for the left finger	–0.22 (–0.29 to –0.16)	<.001
SD of left ITI	SD of tap times for the left finger	–0.18 (–0.25 to –0.11)	<.001
Mean of right ITI	Mean of tap times for the right finger	–0.20 (–0.27 to –0.14)	<.001
SD of right ITI	SD of tap times for the right finger	–0.12 (–0.19 to –0.05)	<.001
Consistency metrics			
ITI range	Difference between the longest and shortest ITI	–0.23 (–0.30 to –0.17)	<.001
Fatigue and temporal drift			
ITI slope	Change in ITI overtime	–0.11 (–0.17 to –0.05)	<.001
Last to first ITI	Ratio of last ITI to first ITI	–0.06 (–0.13 to 0.00)	.05
Microfluctuation feature			
Microfluctuation Index	SD of successive ITI differences	–0.16 (–0.22 to –0.09)	<.001

[a]Significant associations were claimed if *P*<.05/15=.0033. Features without laterality labels were computed using all taps across both buttons within each trial. Features labeled “left” or “right” were calculated separately for taps on the left and right buttons, respectively.

[b]ITI: intertap interval.

[c]CV: coefficient of variation.

Beyond global cognitive function, we further examined the associations between finger tapping features and 4 domain-specific cognitive outcomes: memory, executive function, language, and visuospatial function (Table 3, Table S5 in Multimedia Appendix 1). Similar to findings for global cognition, variability in ITI emerged as a consistent predictor across domains. Executive function demonstrated the strongest associations, with 11 out of 15 finger tapping features reaching significance. The most significant association was observed for mean ITI (β=−0.27; 95% CI −0.33 to −0.20; *P*<.001). Visuospatial function showed fewer associations (6 significant features), yet ITI range again emerged as the strongest predictor (β=−0.26; 95% CI −0.38 to −0.15; *P*<.001). When we applied a false discovery rate correction across all 75 finger

XSL•FO
RenderX

tapping-cognition tests, the overall pattern remained similar. Ten features remained significant for memory, 11 for executive function, 10 for language, 10 for visuospatial function, and 11 for global cognitive function (all with false discovery rate <.05). After excluding trials with an alternating tap ratio less than 0.8, the pattern of the associations between tapping metrics and cognitive domains remained similar (Table S6 in Multimedia Appendix 1). In additional sensitivity analysis, we used the first 3 sessions of each participant. The directions of the associations between finger tapping measures and cognitive domains were largely consistent, but fewer associations were statistically significant (Table S7 in Multimedia Appendix 1). Across all domains, 10 (13.3%) interactions reached nominal significance (Table S8 in Multimedia Appendix 1). For global cognitive function, the most significant interaction was for ITI slope (*P*=.03). We conducted a sensitivity analysis using complete-case data by excluding participants with missing neuropsychological test scores (Table S9 in Multimedia Appendix 1). Most associations remained significant with the same direction of effects.

Table . The most significant digital finger tapping feature for each cognitive domain[a].

Cognitive domain	Number of significant finger tapping features, n	Most significant association		
		Finger tapping features	β (95% CI)	*P* value
Memory	9	ITI[b] range	–0.19 (–0.26 to –0.12)	<.001
Executive function	11	Mean ITI	–0.27 (–0.33 to –0.20)	<.001
Language	9	Mean of right ITI	–0.24 (–0.32 to –0.16)	<.001
Visuospatial function	6	ITI Range	–0.26 (–0.38 to –0.15)	<.001

[a]Significant associations were claimed if *P*<.05/15=.003.

[b]ITI: intertap interval.

Given the presence of multiple finger tapping features associated with cognitive performance, we constructed a composite tapping score for each cognitive test domain using LASSO regression. The finger tapping composite scores were significantly correlated with cognitive domain scores (all with *P*<.001). The strength of these correlations differed by domain. Partial correlation coefficients were 0.21 for memory, 0.22 for executive function, 0.38 for language, 0.23 for visuospatial function, and 0.41 for global cognition.

We further examined the associations between finger tapping features and cognitive function by analyzing FTTs administered separately to the left and right hands (Tables S5, S10, and S11 in Multimedia Appendix 1). Overall, the direction and significance of associations observed in the combined-hand models were largely consistent with those identified in hand-specific analyses, supporting the robustness of the findings across tapping conditions. For global cognitive function, 9 features remained significantly associated in both left-hand and right-hand models. For example, 2 key features of temporal control, including the mean ITI and ITI range, were consistently associated with lower global cognition across all 3 models. Across domain-specific outcomes, effect directions were again broadly consistent. For executive function, the number of significant tapping features was similar across hands (10 in left-hand and 8 in right-hand models), with mean ITI or mean right-hand ITI consistently emerging as the most predictive features. Language and memory domains also demonstrated robust associations, particularly with right-hand tapping metrics such as ITI range and mean ITI. In contrast, associations for visuospatial function were more pronounced in the right-hand model (6 significant features) than the left-hand model (1 significant feature). This asymmetry may reflect functional lateralization of spatial processing or task engagement differences between hands. Additionally, the predominance of right-handed individuals in the sample (247/302, 81.8%) may have contributed to the stronger associations observed in right-hand tapping performance.

To further examine the robustness of associations and minimize potential confounding from handedness, we conducted subgroup analyses restricted to right-handed participants (n=247; Tables S12-S14 in Multimedia Appendix 1). While differences in the number of significant tapping features emerged across cognitive domains, the direction of effects, particularly for temporal variability features, remained consistent. When we reanalyzed the data by dominant versus nondominant hand, the overall pattern of associations was similar (Tables S15 and S16 in Multimedia Appendix 1). For executive function, however, more finger tapping measures were significantly associated with performance when tests were administered with the nondominant hand (n=10) than with the dominant hand (n=8). In dominant-hand trials, mean ITI showed the most significant associations across all cognitive domains. In nondominant hand trials, mean ITI again showed the strongest associations with memory, executive function, and language. However, the ITI range, reflecting temporal variability, was most significantly associated with visuospatial and global cognitive function.

Discussion

Principal Results

In this study of predominantly cognitively intact, community-based older adults, several digital finger tapping features were significantly associated with cognitive performance. These results support the use of motor-based digital measures as a tool for evaluating cognitive health. Notably, variability features such as the SD, coefficient of variation, ITI range, and the Microfluctuation Index were consistently associated with cognitive domains. These results

XSL•FO
RenderX

suggest that temporal features derived from simple tapping tasks may serve as sensitive and scalable indicators of general cognitive functioning.

Executive function emerged as the most consistently associated domain, showing strong relationships with 13 finger tapping features. Given that executive function is a multidimensional construct encompassing key cognitive processes such as processing speed, planning, working memory, and inhibitory control [27], this finding provides important insight into the potential cognitive mechanisms underlying fine motor performance. Although the finger tapping task appears simple, its successful execution in a timed context requires the ability to sustain and update task goals (working memory) [28], initiate and maintain a consistent motor rhythm (planning and motor initiation) [29], and suppress irrelevant motor impulses (inhibitory control) [30]. The strong associations between temporal tapping features, particularly variability-based features, and executive function suggest that digital motor timing tasks may serve as sensitive proxies for detecting early executive dysfunction. Interestingly, language performance was most strongly associated with mean tap times for the right finger of both right and left hands. This suggested that beyond just timing or variability, the dynamics of motor execution may have a specific connection with language processing abilities. This finding is consistent with prior research indicating shared neural substrates for motor and language functions [31]. And it provides behavioral evidence for the functional interdependence between these domains.

Stratified analyses by hand and by handedness revealed generally consistent associations between finger tapping features and cognitive performance. Across both left- and right-hand administered tests, key temporal metrics such as SD of ITI, ITI range, and Microfluctuation Index remained associated with global cognitive function. However, subtle differences emerged in the number and strength of significant associations across hands and cognitive domains. For instance, visuospatial associations were more prominent in right-hand tests, while left-hand results showed fewer significant features overall. Analyses restricted to right-handed participants largely confirmed the main findings, with core tapping features such as mean ITI and ITI range consistently emerging as the informative predictors across domains. These results reinforce the generalizability of tapping-derived digital biomarkers, while also suggesting that motor laterality and hand-specific administration may modestly influence the association. While many tapping features showed associations with cognition, 4 features, including skewness and kurtosis of ITI, alternating tap ratio, and last-to-first ITI ratio, were not significantly related to any cognitive domain. The alternating tap ratio is primarily an index of adherence to the alternating-tap instruction. However, maintaining an alternating pattern relies on attention and executive function. Therefore, variation in this measure may carry cognitively relevant information rather than simply indicating noncompliance to following instructions. For this reason, we included it in our analyses, although we did not observe significant associations with any cognitive domain. These findings highlight the need to prioritize the most informative digital biomarkers when evaluating cognitive health.

Although these associations reached statistical significance, the effect sizes were small. This suggests that finger tapping measures explain only a modest proportion of variation in cognitive performance. From a clinical standpoint, these findings also underscore the growing relevance of digital phenotyping as a complementary approach to traditional cognitive screening [32]. Early signs of executive dysfunction and cognitive decline are often subtle and can be overlooked, especially in high-functioning or highly educated individuals. The associations identified between fine motor irregularities and cognitive performance suggest that brief, app-based tapping tasks may serve as passive indicators of early neurocognitive vulnerability.

This study offers several notable strengths. First, it leverages a comprehensive battery of neuropsychological assessments covering four major cognitive domains, enabling robust evaluation of cognitive function. Second, the repeated administration of digital FTTs at regular intervals provides a more detailed characterization of motor performance over time and its association with cognitive outcomes. While digital finger tapping tasks have been explored in prior studies, they have predominantly focused on targeted clinical populations, such as individuals with PD or other neurodegenerative disorders [10,33,34]. In contrast, our study is among the first to investigate smartphone-based finger tapping performance in a large, community-based sample of older adults. The majority of participants were subjectively healthy and free of diagnosed cognitive impairment. This distinction enhances the generalizability of our findings and supports the broader application of digital motor assessments for population-level cognitive monitoring. The use of a smartphone-based tapping task, coupled with the derivation of multiple temporal and variability-based digital metrics, represents a novel, noninvasive, and scalable approach for assessing fine motor control in real-world settings. Notably, a high proportion of older iPhone users completed and returned the tapping assessments. Many participants completed multiple administrations. Given the well-documented challenges of adherence in digital health studies, this level of engagement suggests that the task was both accessible and acceptable to participants. Additionally, stratification of tapping performance by left and right hand allows for the exploration of potential lateralization effects. This offers further insight into the relationship between motor function and cognitive domains.

Comparison With Prior Work

Our findings align with prior research using the traditional method of assessing finger tapping speed [35]. In the earlier study, finger tapping was featured by instructing participants to repeatedly press a mechanical lever with an attached counter using the index finger of each hand over a fixed 10-second interval. It demonstrated that individuals who later converted to MCI exhibited a faster decline in tapping speed compared to nonconverters. Similarly, in our study, a lower number of taps (slower tapping speed) was significantly associated with lower global and domain-specific cognitive performance, supporting motor slowing as a behavioral marker of cognition.

XSL•FO
RenderX

Limitations

First, this study was cross-sectional and conducted in a sample consisting predominantly of cognitively intact older adults. Although finger tapping features were collected longitudinally, the cognitive assessments were not temporally aligned with each tapping session. Therefore, we summarized finger tapping performance by averaging repeated trials into a single mean value per participant. This approach allowed us to focus on between-person differences in mean tapping performance but did not capture within-person variability or change over time. In addition, the time interval between neuropsychological test and finger tapping (mean 1.7, SD 1.2 y) limited inferences about how well tapping performance reflects real-time cognitive status. As longitudinal follow-up data become available, future work will evaluate whether tapping metrics can predict subsequent cognitive decline or discriminate between cognitively impaired and unimpaired individuals.

Second, the first step of the 2-finger tap app entitled tapping speed informed the participant that "this activity measures your tapping speed." However, the instructions for the left hand and right hand in steps 2 and 4 include the statement "try to make your taps as even as possible" (Figure S1 in Multimedia Appendix 1). These instructions were designed to encourage fast tapping with a stable rhythm. In practice, this may have led some participants to focus more on speed and others more on regularity. In this context, ITI variability may reflect both intrinsic motor control and individual task strategies. Our findings should be interpreted accordingly. Future studies could use separate speed-focused and regularity-focused conditions to better disentangle these components as both may reflect early neurodegenerative processes. Besides, we examined only the alternating 2-finger version of the FTT with a fixed 10-second trial duration. Because of this short duration, the ITI slope may capture rhythm establishment rather than physiological fatigue. It should be interpreted with caution. Future work should compare alternating and single-finger tapping protocols and systematically vary test duration to better separate rhythm establishment from physiological fatigue. This also can help determine the optimal design for digital tapping-based cognitive assessment.

Third, while the smartphone-based tapping task provides a standardized and practical assessment platform, performance may still be influenced by external factors such as participants' familiarity with digital devices or technical variability in smartphone hardware and software. Although this study adjusted for age, sex, education, and handedness, conditions such as arthritis, motor fatigue, medication use, or depression may independently affect tapping speed, rhythm, or consistency. In some cases, motor slowing may reflect physical rather than cognitive decline. Therefore, these factors should be considered when interpreting tapping metrics in clinical practice, especially in people with multiple comorbidities.

Finally, the tapping task was only available on iOS devices. Therefore, 133 Android users were excluded. Because iOS and Android ownership differs by socioeconomic and demographic factors, this restriction may introduce socioeconomic selection bias. In addition, participants in our sample were generally well-educated and reported good to excellent subjective health. These characteristics may limit the generalizability of our findings to more diverse populations with different demographic or clinical profiles. Moreover, cognitive performance was high in this cohort (mean MMSE 28.8, SD 1.3; dementia prevalence 0.3%), suggesting possible ceiling effects in the cognitive measures. Such ceiling effects may attenuate observed associations and underestimate the true relationships that may be more apparent in clinically diverse populations. In addition, the iPhone models and variation in touchscreen sampling rates or input latency across devices may introduce measurement noise into fine-grained variability metrics (eg, SD of ITI and Microfluctuation Index). Further research in larger and more diverse populations with standardized hardware is needed to confirm these findings.

Conclusions

In summary, this study identified several finger tapping features that were significantly associated with performance across multiple cognitive domains and highlighted the potential of digital finger tapping tasks as informative biomarkers for assessing cognitive function. These findings suggest that finger tapping tasks may be used as noninvasive, scalable tools to assess cognitive performance.

Acknowledgments

The authors sincerely thank the Framingham Heart Study participants for their dedication and invaluable contribution, without which this study would not have been possible. They also appreciate the Framingham Heart Study researchers and staff for their continued efforts in conducting participant examinations.

Funding

This work was supported by grants from the National Institute on Aging (R01AG076757 and R03AG095992) and National Heart, Lung, and Blood Institute (R01HL141434). The Framingham Heart Study is funded through the National Heart, Lung, and Blood Institute Contract (N01-HC-25195, HHSN268201500001I, 75N92019D00031, and 75N92025D00012), and the neuropsychological testing data and brain health data is supported by grants R21HL175584 and by grants from the National Institute on Aging (R01AG083735, R01AG080670, and U01AG068221). The funding agencies had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

XSL•FO
RenderX

Data Availability

The data used in this study can be accessed through the Framingham Heart Study upon reasonable request and in compliance with their data sharing guidelines [36].

Authors' Contributions

HD, JMM, and HL conceptualized and designed the study. HD and HL performed the analysis. HD and HL drafted the initial manuscript. TAO, JR, HYT, CL, ES, and CHP critically reviewed the final draft. All authors reviewed and edited the manuscript and have read and agreed to the version for publication.

Conflicts of Interest

ES is an employee of CareEvolution. The other authors have no conflicts of interest to report.

Multimedia Appendix 1
Supplemental tables and figures.
[DOCX File, 650 KB - jmir_v28i1e82463_app1.docx]

References

1. Yaqub A, Darweesh SKL, Dommershuijsen LJ, et al. Risk factors, neuroimaging correlates and prognosis of the motoric cognitive risk syndrome: a population-based comparison with mild cognitive impairment. Eur J Neurol 2022 Jun;29(6):1587-1599. [doi: 10.1111/ene.15281] [Medline: 35147272]
2. Collyer TA, Murray AM, Woods RL, et al. Association of dual decline in cognition and gait speed with risk of dementia in older adults. JAMA Netw Open 2022 May 2;5(5):e2214647. [doi: 10.1001/jamanetworkopen.2022.14647] [Medline: 35639376]
3. Grande G, Rizzuto D, Vetrano DL, et al. Cognitive and physical markers of prodromal dementia: a 12-year-long population study. Alzheimers Dement 2020 Jan;16(1):153-161. [doi: 10.1002/alz.12002] [Medline: 31914224]
4. Rudd KD, Lawler K, Callisaya ML, Alty J. Investigating the associations between upper limb motor function and cognitive impairment: a scoping review. Geroscience 2023 Dec;45(6):3449-3473. [doi: 10.1007/s11357-023-00844-z] [Medline: 37337026]
5. Kueper JK, Speechley M, Lingum NR, Montero-Odasso M. Motor function and incident dementia: a systematic review and meta-analysis. Age Ageing 2017 Sep 1;46(5):729-738. [doi: 10.1093/ageing/afx084] [Medline: 28541374]
6. Miyazaki A, Ito Y, Okuyama T, et al. Association between upper limb movements during drumming and cognition in older adults with cognitive impairment and dementia at a nursing home: a pilot study. Front Rehabil Sci 2023;4:1079781. [doi: 10.3389/fresc.2023.1079781] [Medline: 37347105]
7. Gopal A, Hsu WY, Allen DD, Bove R. Remote assessments of hand function in neurological disorders: systematic review. JMIR Rehabil Assist Technol 2022 Mar 9;9(1):e33157. [doi: 10.2196/33157] [Medline: 35262502]
8. Cousins MS, Corrow C, Finn M, Salamone JD. Temporal measures of human finger tapping: effects of age. Pharmacol Biochem Behav 1998 Feb;59(2):445-449. [doi: 10.1016/s0091-3057(97)00443-7] [Medline: 9476994]
9. Hubel KA, Reed B, Yund EW, Herron TJ, Woods DL. Computerized measures of finger tapping: effects of hand dominance, age, and sex. Percept Mot Skills 2013 Jun;116(3):929-952. [doi: 10.2466/25.29.PMS.116.3.929-952] [Medline: 24175464]
10. Akram N, Li H, Ben-Joseph A, et al. Developing and assessing a new web-based tapping test for measuring distal movement in Parkinson's disease: a Distal Finger Tapping test. Sci Rep 2022 Jan 10;12(1):386. [doi: 10.1038/s41598-021-03563-7] [Medline: 35013372]
11. Bot BM, Suver C, Neto EC, et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. Sci Data 2016 Mar 3;3:160011. [doi: 10.1038/sdata.2016.11] [Medline: 26938265]
12. Omberg L, Chaibub Neto E, Perumal TM, et al. Remote smartphone monitoring of Parkinson's disease and individual response to therapy. Nat Biotechnol 2022 Apr;40(4):480-487. [doi: 10.1038/s41587-021-00974-9] [Medline: 34373643]
13. Ding H, Wang B, Hamel AP, et al. Exploring cognitive progression subtypes in the Framingham Heart Study. Alzheimers Dement (Amst) 2024;16(1):e12574. [doi: 10.1002/dad2.12574] [Medline: 38515438]
14. Feinleib M, Kannel WB, Garrison RJ, McNamara PM, Castelli WP. The Framingham offspring study. Design and preliminary data. Prev Med 1975 Dec;4(4):518-525. [doi: 10.1016/0091-7435(75)90037-7] [Medline: 1208363]
15. Andersson C, Johnson AD, Benjamin EJ, Levy D, Vasan RS. 70-year legacy of the Framingham Heart Study. Nat Rev Cardiol 2019 Nov;16(11):687-698. [doi: 10.1038/s41569-019-0202-5] [Medline: 31065045]
16. Splansky GL, Corey D, Yang Q, et al. The third generation cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. Am J Epidemiol 2007 Jun 1;165(11):1328-1335. [doi: 10.1093/aje/kwm021] [Medline: 17372189]

XSL•FO
RenderX

17. Rong J, Pathiravasan CH, Zhang Y, et al. Baseline smartphone app survey return in the electronic Framingham Heart Study Offspring and Omni 1 Study: eCohort study. JMIR Aging 2024 Dec 31;7(1):e64636. [doi: 10.2196/64636] [Medline: 39740111]
18. Seshadri S, Wolf PA. Lifetime risk of stroke and dementia: current concepts, and estimates from the Framingham Study. Lancet Neurol 2007 Dec;6(12):1106-1114. [doi: 10.1016/S1474-4422(07)70291-0] [Medline: 18031707]
19. Guze SB. Diagnostic and Statistical Manual of Mental Disorders, (DSM-IV). Am J Psychiatry 1995 Aug;152(8):1228-1228. [doi: 10.1176/ajp.152.8.1228]
20. Albert MS, DeKosky ST, Dickson D, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011 May;7(3):270-279. [doi: 10.1016/j.jalz.2011.03.008] [Medline: 21514249]
21. Yuan J, Maserejian N, Liu Y, et al. Severity distribution of Alzheimer's disease dementia and mild cognitive impairment in the Framingham Heart Study. J Alzheimers Dis 2021;79(2):807-817. [doi: 10.3233/JAD-200786] [Medline: 33361590]
22. Lee J, Walker ME, Gabriel KP, Vasan RS, Xanthakis V. Associations of accelerometer-measured physical activity and sedentary time with chronic kidney disease: the Framingham Heart Study. PLoS One 2020;15(6):e0234825. [doi: 10.1371/journal.pone.0234825] [Medline: 32542048]
23. Au R, Seshadri S, Wolf PA, et al. New norms for a new generation: cognitive performance in the Framingham Offspring cohort. Exp Aging Res 2004;30(4):333-358. [doi: 10.1080/03610730490484380] [Medline: 15371099]
24. De Anda-Duran I, Sunderaraman P, Searls E, et al. Comparing cognitive tests and smartphone-based assessment in 2 US community-based cohorts. J Am Heart Assoc 2024 Jan 16;13(2):e032733. [doi: 10.1161/JAHA.123.032733] [Medline: 38226519]
25. Chen J, Doyle MF, Fang Y, et al. Eripheral inflammatory biomarkers are associated with cognitive function and dementia: Framingham Heart Study Offspring cohort. Aging Cell 2023 Oct;22(10):e13955. [doi: 10.1111/acel.13955] [Medline: 37584418]
26. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011 Mar;20(1):40-49. [doi: 10.1002/mpr.329] [Medline: 21499542]
27. Karr JE, Areshenkoff CN, Rast P, Hofer SM, Iverson GL, Garcia-Barrera MA. The unity and diversity of executive functions: a systematic review and re-analysis of latent variable studies. Psychol Bull 2018 Nov;144(11):1147-1185. [doi: 10.1037/bul0000160] [Medline: 30080055]
28. de Paula JJ, Malloy-Diniz LF, Romano-Silva MA. Reliability of working memory assessment in neurocognitive disorders: a study of the Digit Span and Corsi Block-Tapping tasks. Braz J Psychiatry 2016;38(3):262-263. [doi: 10.1590/1516-4446-2015-1879] [Medline: 27579598]
29. Omata K, Ito S, Takata Y, Ouchi Y. Similar neural correlates of planning and execution to inhibit continuing actions. Front Neurosci 2018;12:951. [doi: 10.3389/fnins.2018.00951] [Medline: 30631263]
30. Prigatano GP, de Oliveira SB, Goncalves CWP, Denucci SM, Pereira RM, Braga LW. Inhibitory control of adjacent finger movements while performing a modified version of the Halstead Finger Tapping Test: effects of age, education and sex. J Int Neuropsychol Soc 2021 Sep;27(8):813-824. [doi: 10.1017/S1355617720001101] [Medline: 33190661]
31. Brozzoli C, Roy AC, Lidborg LH, Lövdén M. Language as a tool: motor proficiency using a tool predicts individual linguistic abilities. Front Psychol 2019 Jul 16;10:1639. [doi: 10.3389/fpsyg.2019.01639] [Medline: 31379674]
32. Insel TR. Digital phenotyping: a global tool for psychiatry. World Psychiatry 2018 Oct;17(3):276-277. [doi: 10.1002/wps.20550] [Medline: 30192103]
33. Lee CY, Kang SJ, Hong SK, Ma HI, Lee U, Kim YJ. A validation study of a smartphone-based finger tapping application for quantitative assessment of bradykinesia in Parkinson's disease. PLoS ONE 2016 Jul 28;11(7):e0158852. [doi: 10.1371/journal.pone.0158852] [Medline: 27467066]
34. Roalf DR, Rupert P, Mechanic-Hamilton D, et al. Quantitative assessment of finger tapping characteristics in mild cognitive impairment, Alzheimer's disease, and Parkinson's disease. J Neurol 2018 Jun;265(6):1365-1375. [doi: 10.1007/s00415-018-8841-8] [Medline: 29619565]
35. Buracchio T, Dodge HH, Howieson D, Wasserman D, Kaye J. The trajectory of gait speed preceding mild cognitive impairment. Arch Neurol 2010 Aug;67(8):980-986. [doi: 10.1001/archneurol.2010.159] [Medline: 20697049]
36. For researchers. Framingham Heart Study. URL: https://www.framinghamheartstudy.org/fhs-for-researchers [accessed 2026-01-24]

Abbreviations

eFHS: electronic Framingham Heart Study
FHS: Framingham Heart Study
FTT: finger tapping test
ITI: intertap interval
LASSO: least absolute shrinkage and selection operator
MCI: mild cognitive impairment

XSL•FO
RenderX

MMSE: Mini-Mental State Exam
PD: Parkinson disease

Edited by A Coristine; submitted 15.Aug.2025; peer-reviewed by G Souza, W Myung; revised version received 20.Dec.2025; accepted 20.Dec.2025; published 18.Feb.2026.

Please cite as:
Ding H, Orwig TA, Rong J, Tatere HY, Liu C, Schramm E, Pathiravasan CH, Murabito JM, Lin H
Associations Between Smartphone-Based Finger Tapping and Cognitive Performance in Older Adults: Observational Study
J Med Internet Res 2026;28:e82463
URL: https://www.jmir.org/2026/1/e82463
doi:10.2196/82463

© Huitong Ding, Taylor A Orwig, Jian Rong, Heaven Y Tatere, Chunyu Liu, Eric Schramm, Chathurangi H Pathiravasan, Joanne M Murabito, Honghuang Lin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Impact of Push Notifications on Physical Activity and Sodium Intake Among Patients with Hypertension: Microrandomized Trial of a Just-in-Time Adaptive Intervention

Jessica R Golbus[1,2*], MD, MS; Michael P Dorsch[3*], MS, PharmD; Yuxuan Chen[4], MS; Tanima Basu[1], MA, MS; Evan Luff[1], MS; Predrag Klasnja[5], PhD; Mark W Newman[5], PhD; Lesli E Skolarus[6], MD, MS; Walter Dempsey[4], PhD; Brahmajee K Nallamothu[1,2], MPH, MD

[1]Division of Cardiovascular Medicine, Department of Internal Medicine, University of Michigan, 2723 Cardiovascular Center, 1500 E Medical Center Dr. SPC 5853, Ann Arbor, MI, United States

[2]Michigan Integrated Center for Health Analytics and Medical Prediction (MiCHAMP), University of Michigan, Ann Arbor, MI, United States

[3]Department of Clinical Pharmacy, College of Pharmacy, University of Michigan, Ann Arbor, MI, United States

[4]Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, United States

[5]School of Information and Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, United States

[6]Division of Stroke Vascular Neurology, Davee Department of Neurology, Northwestern University Feinberg School of Medicine, Chicago, IL, United States

[*]these authors contributed equally

Corresponding Author:
Jessica R Golbus, MD, MS
Division of Cardiovascular Medicine, Department of Internal Medicine, University of Michigan, 2723 Cardiovascular Center, 1500 E Medical Center Dr. SPC 5853, Ann Arbor, MI, United States

Abstract

Background: Achieving adequate blood pressure control is challenging for patients and clinicians. Digital hypertension management solutions that use push notifications to promote lifestyle management have been proposed as an approach, but their effectiveness remains unknown.

Objective: This analysis was designed to interrogate the independent and short-term effects of push notifications, tailored to participant and environmental factors, and on physical activity levels and sodium intake among individuals with hypertension.

Methods: The myBPmyLife study was a 6-month randomized controlled trial of participants with self-reported hypertension recruited from an academic medical center and federally qualified health centers. A core component of the intervention consisted of microrandomized push notifications promoting lifestyle modifications that were randomly delivered at 4 daily time points and focused on physical activity and dietary sodium intake. Our primary outcome for this secondary analysis was the step count 60 minutes after a physical activity notification and lower-sodium food choices 24 hours after a dietary notification. This analysis focuses on the results of the microrandomized trial and used a centered and weighted least squares method adapted for 2 or more treatments.

Results: A total of 298 participants were randomized to the intervention arm, of whom 287 had data available for analysis. Participants' mean age was 59.5 (SD 13.1) years, 138 (48.1%) were women, and 206 (71.8%) were White. Participants were randomized at 187,517 time points over 6 months, which led to 0.96 (SD 0.86) push notifications per day divided between activity (50.4%; SD 0.4) and dietary (49.8%; SD 0.4) notifications. Activity notifications did not increase step count in the 60 minutes after a notification (estimate 1.01, 95% CI 0.98 - 1.04; P=.40). Similarly, dietary notifications did not impact the number of lower-sodium food choices in the subsequent 24 hours (estimate 0.93, 95% CI 0.83 - 1.04; P=.23), but in exploratory post hoc analyses, did increase mobile app use by 95.5% (95% CI 1.81 - 2.10; P<.001), mobile app clicks or searches by 93.7% (95% CI 1.72% - 2.16%; P<.001), and low sodium searches by 113.0% (95% CI 1.73 - 2.53; P<.001), all within 60 minutes.

Conclusions: In patients with hypertension, push notifications did not impact short-term physical activity levels or dietary sodium intake but did improve intervention engagement.

Trial Registration: ClinicalTrials.gov NCT05154929; https://clinicaltrials.gov/study/NCT05154929

International Registered Report Identifier (IRRID): RR2-10.1161/JAHA.123.031234

(J Med Internet Res 2026;28:e78218) doi:10.2196/78218

XSL•FO
RenderX

KEYWORDS

digital health; hypertension; mHealth; mobile health; physical activity; just-in-time adaptive intervention

Introduction

Nearly half of all US adults have hypertension, though only 1 in 4 have their blood pressure (BP) adequately treated [1]. Achieving adequate BP control is challenging for patients and clinicians given the episodic nature of clinical encounters, high patient volumes, and the silent nature of the disease process, promoting clinical inertia [2]. Digital hypertension management solutions aim to improve BP control by promoting patient self-management and expanded access to care. These solutions often center around connected BP cuffs for self-monitoring, with or without medication management, and may be paired with behavior and lifestyle change interventions in the form of mobile apps, text messages or push notifications, and phone calls [2]. Although these methods could enhance BP control, the effectiveness of each digital intervention component—especially within a multicomponent framework—continues to be a significant and unresolved issue.

We recently completed the myBPmyLife study, which evaluated a mobile health (mHealth) intervention designed to improve BP control by promoting increased physical activity and the selection of lower-sodium food choices. The study recruited participants from an academic medical center and federally qualified health centers [3]. The intervention consisted of a mobile app designed to facilitate goal setting and feedback and contextually tailored push notifications delivered as a part of a microrandomized trial, a novel experimental design in which participants are serially randomized to different types or levels of an intervention (eg, push notifications) [4]. While the intervention reduced sodium intake and improved physical activity levels, it failed to lower systolic BP (SBP) compared to a control group that engaged in BP self-monitoring alone [5].

One important question that remained was whether push notifications led to immediate short-term effects or mediated behavior change that occurred on a longer timescale. To address this question, we present the myBPmyLife study's 6-month microrandomized trial results, which focuses on important secondary, mechanistic outcomes of this trial. This analysis was designed to interrogate the independent and short-term effects of push notifications, tailored to participant and environmental factors, on physical activity levels and sodium intake. Such an approach allows us to isolate the causal effects of push notifications relative to the larger intervention package. We hypothesized that tailored push notifications would increase participants' step counts in the 60 minutes following delivery and decrease sodium intake in the subsequent 24 hours.

Methods

Study Design and Participants

The myBPmyLife study was a prospective, remotely administered, randomized-controlled trial (ClinicalTrials.gov NCT05154929) of 602 participants with self-reported hypertension (CONSORT [Consolidated Standards of Reporting Trials] guidelines; Checklist 1). Participants were recruited from the University of Michigan Health and the Hamilton Community Health Network, a series of federally qualified health centers in Flint, Michigan. The analysis described here focuses on the results of the microrandomized trial, which was embedded within the intervention arm of the randomized controlled trial, and which was composed of push notifications promoting increased physical activity and the selection of lower-sodium food choices. The data and code that support this study are openly available through GitHub.

Participants were eligible for the study if they were 18 years or older with self-reported hypertension, had no changes in their antihypertensive therapies in the preceding 4 weeks (if on medical therapies), consumed >1500 mg of daily sodium, and owned a compatible smartphone. Sodium intake was assessed by the Block Sodium Screener [6] after informed consent was obtained, and those who consumed <1500 mg/d of sodium were ineligible. Participants with contraindications to physical activity or to following a low-sodium diet were excluded from the study. The full inclusion and exclusion criteria have been previously published [3].

Ethical Considerations

The study was approved by the University of Michigan Health IRB (HUM00205845). All participants participated in an informed consent process and signed consent forms. All study data are deidentified and remain anonymous. Participants received up to US $100 over 6 months for their time completing study tasks. All participants were provided with a Bluetooth-connected smartwatch (Fitbit Versa 2) and a Bluetooth BP monitor (Omron Evolv BP7000) for the purposes of study participation.

Study Procedures

The myBPmyLife study was conducted between December 2021 and July 2023. Study procedures have been previously published. Briefly, participants underwent remote screening, recruitment, and consent processes. Consent forms were signed using the mobile study app myDataHelps (CareEvolution, LLC). Following the completion of the Block Sodium Screener, eligible participants were randomized 1:1 to the intervention arm, which received a BP monitor and an mHealth intervention delivered through a mobile app and a smartwatch, or to the enhanced usual arm, which received a BP monitor and smartwatch alone. Randomization schemes were created by study statisticians in Randomize.net and performed by study staff using a permuted block design with variable block sizes of 2-6, stratified by study site. Researchers and participants were unblinded to allocation assignment given the nature of the intervention.

Following informed consent, participants in both study arms were mailed a Fitbit Versa 2 and a Bluetooth BP monitor (Omron Evolv BP7000) and provided instructions on downloading the associated mobile apps to enable data sharing. All participants subsequently underwent remote enrollment appointments, at which time they were assisted with pairing

XSL•FO
RenderX

their smartwatches and smartphones as needed. Participants randomized to the intervention arm additionally answered a series of questions to facilitate intervention tailoring, including those on mobility, confidence in selecting lower-sodium food choices, preferred name, times of day for push notification (ie, morning, lunch, afternoon, and evening times), and preferred day for grocery shopping. These preferences could be changed during the study upon participant request. Following enrollment but before intervention receipt, participants experienced a baseline period in which no interventions were delivered to establish baseline step counts.

Description of the Intervention

The myBPmyLife study was designed to increase physical activity levels and promote the selection of lower-sodium food choices through an mHealth intervention grounded in the behavior change strategies of goal setting, prompts, visualizations, and feedback. The intervention consisted of both static (ie, mobile app with a central visualization promoting strategic feedback and goal setting) and dynamic components, with the latter referring to push notifications delivered as part of the microrandomized trial [5]. Microrandomized trials are an experimental approach to guide the design of just-in-time adaptive interventions. Just-in-time interventions, within the context of health behavior change, aim to provide support at times of opportunity or risk when individuals are receptive to change [7]. As such, just-in-time interventions aim to shape behavior change both in the moment and over time.

Microrandomized trials are an experimental design in which participants are serially randomized to receive (or not receive) different types or levels of an intervention (eg, push notifications) at different time points (ie, decision points) over the length of a study [4]. Intervention efficacy is ascertained by 1 or more proximal outcomes, which refer to short-term outcomes hypothesized to mediate a desired long-term effect. By leveraging intraindividual contrasts, microrandomization is a powerful experimental design for determining causal effects. In the myBPmyLife trial, microrandomized push notifications comprised a core component of the intervention (Figure S1 in Multimedia Appendix 1). They were designed to promote low-level physical activity and the selection of lower-sodium food choices. Activity notifications were tailored based on the time of day (ie, morning, lunch, afternoon, and evening), day of week (ie, weekend vs weekday), weather, and mobility. Similarly, dietary notifications were tailored according to the time of day, day of the week (ie, weekend, weekday, or grocery day), and participants' confidence in selecting lower-sodium food choices. Both notification types consisted of expert-generated and community-generated notifications, with the latter tailored based on participants' site of enrollment (ie, University of Michigan Health or Hamilton Community Health Network). A subset of notifications was personalized using participants' preferred names [8].

The study was designed so that participants would receive 1 activity or dietary notification per day, on average, throughout the study. Thus, participants had a 25% probability of receiving a notification at each decision point, divided equally between activity and dietary notifications. The proximal outcomes for the microrandomized trial serve as key secondary outcomes from the overarching clinical trial. The proximal outcome for activity notifications was step count 60 minutes after a decision point as determined by the smartwatch or mobile phone, as these messages were intended to be actionable in real time. For dietary notifications, the proximal outcome was self-reported lower-sodium food choices in the mobile app within 24 hours of a decision point, as these messages were intended to be applicable at a future time around meals or snacks and when at restaurants or grocery stores. For the dietary analysis, the decision was made post hoc to explore a series of more proximal engagement measures, as dietary notifications encouraged participants to use the mobile app to identify lower-sodium alternatives to commonly consumed or purchased foods. These exploratory outcomes included mobile app use (yes or no), the number of mobile app searches or clicks, and the number of low salt searches within the mobile app, all within 60 minutes of a decision point.

Statistical Analysis

Baseline clinical characteristics are described as means and SD for continuous symmetric variables and median with IQR for skewed continuous variables. Categorical variables are presented as counts and percentages. We performed Student 2-tailed *t* tests for bivariate comparisons between continuous variables and chi-square tests for comparisons across categorical variables.

For both the physical activity and dietary analyses, the intervention period was defined as the time of participants' first decision point at or beyond day 8 and lasting until day 180 (ie, maximum study period of 173 d). This was defined to account for variability in the duration of the baseline period (as determined by smartwatch wear time) and in the completion of the final surveys.

For the physical activity analysis, we evaluated the effect of delivering an activity notification compared to no notification on step count 60 minutes after a decision point. As step counts were positively skewed, 0.5 was added to all counts before natural log transformation. Analyses used a version of the weighted and centered least squares (WCLS) method. The WCLS approach is designed for microrandomized trials, where treatments occur frequently and moderators may be affected by prior interventions. This method involves centering time-varying treatments using their conditional means based on past history and applying weights similar to inverse probability of treatment weighting in causal inference. By combining centering and weighting, WCLS offers robustness against the misspecification of the working model of weighted outcome condition on history [9]. Our analyses used a version of the WCLS method adapted to account for 2 or more treatments [9]. Such an approach increases statistical power by avoiding treatment-specific models and guarantees the robust, unbiased inclusion of covariates to reduce noise. Models were adjusted for age (<65 vs >65 y), gender, race (White vs non-White), baseline mean daily step count (dichotomized on mean value), step count 30 minutes before a decision point (standardized as mean or SD), and time (d). At each decision point, participants were considered available to receive the intervention if they were wearing their smartwatches consistent with prior studies [10]. This was

XSL•FO
RenderX

operationalized by requiring that participants have at least 1 heart rate measurement recorded on their smartwatches in the 30 minutes before a decision point. Models included decision points at which participants were available to receive notifications, and step count data were available in the 30 minutes before and 60 minutes after a decision point (ie, excluded 5807, 3.1% of decision points). Subsequently, we conducted a series of exploratory analyses to understand the impact of key demographic (ie, age, gender, and community) and clinical (ie, baseline step count) characteristics on our primary outcome by evaluating the interaction between these covariates and treatment. An additional exploratory model evaluated the impact between time as a quadratic term and treatment. Finally, as outcome data can be present in instances of missing wear time as data can come from both the mobile phone and smartwatch, we conducted a sensitivity analysis in which we excluded the data from participants with zero or missing wear time 3 hours after a decision point. This allowed us to disentangle instances of smartwatch nonwear from zero step counts.

For the dietary analysis, we evaluated the effect of delivering a dietary notification compared to the composite of an activity notification or no notification on lower-sodium choices (as denoted in the mobile app) within 24 hours of a decision point. Given the potential overlap in push notifications over 24 hours, the assessment of treatment effect included only decision points at which participants received a dietary notification, and no additional dietary notifications were sent in the subsequent 24 hours (excluded 42,313, 22.6% decision points). These were contrasted to decision points at which no dietary notification was sent at that time or in the subsequent 24 hours. As lower-sodium choices were zero-inflated, we used the method proposed by Liu et al [11] and used the R (R Foundation for Statistical Computing) function *emee* in package *MRTAnalysis*. Models were adjusted for age (<65 vs >65 y), gender, race (White vs non-White), baseline sodium intake as measured by the Block Sodium Screener (dichotomized on mean value), low salt choices 30 minutes before a decision point, and time (d). Participants were considered available at all decision points given the extended time window for outcome assessment. As with the physical activity analysis, we also conducted a series of exploratory analyses to understand the potential impact of key demographics (ie, age, gender, and community) and baseline characteristics (eg, baseline sodium intake) on sodium intake by evaluating the interaction between these covariates and treatment. An additional exploratory model evaluated the impact between time as a quadratic term and treatment. Similar models were created for our exploratory outcomes of mobile app use (yes or no), number of mobile app clicks or searches, and number of low sodium searches within the mobile app, all within 60 minutes of a decision point. These similarly used the method proposed by Liu et al [11], though it included all decision points (N=187,517), unlike the prior analysis. Covariates in these models included age (<65 vs >65 y), gender, race (White vs Non-white), and either mobile app use, number of search or clicks, or number of low sodium choices 30 minutes before a decision point, respectively. In all cases, $P<.05$ was considered statistically significant. All analyses were conducted in R (version 4.3.1; R Foundation for Statistical Computing).

Results

Study Population

Between December 2021 and July 2023, 912 participants were screened for eligibility, of whom 602 were ultimately enrolled. A total of 298 participants were randomized to the intervention arm of the study, 4 of whom withdrew before intervention receipt during the initial 5 days of the study (Figure 1). An additional 15 participants withdrew from the study, though they allowed their data to be used up until their withdrawal date (Figure 1). Post hoc, an additional 7 participants were excluded from the analysis (6 due to missing baseline step count and 1 missing step count 30 min prior to the decision point), leaving 287 participants available for the primary analysis.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram. Between December 2021 and July 2023, 912 participants were screened for eligibility. Following informed consent, participants were randomized 1:1 to the intervention and enhanced usual care arms with 298 participants randomized to the intervention arm of the trial. Of these, 4 participants withdrew before intervention receipt (postrandomization exclusions), and an additional 15 participants withdrew from the study but allowed their data to be used up until their withdrawal date. A total of 287 participants were included in the analysis of the microrandomized trial.

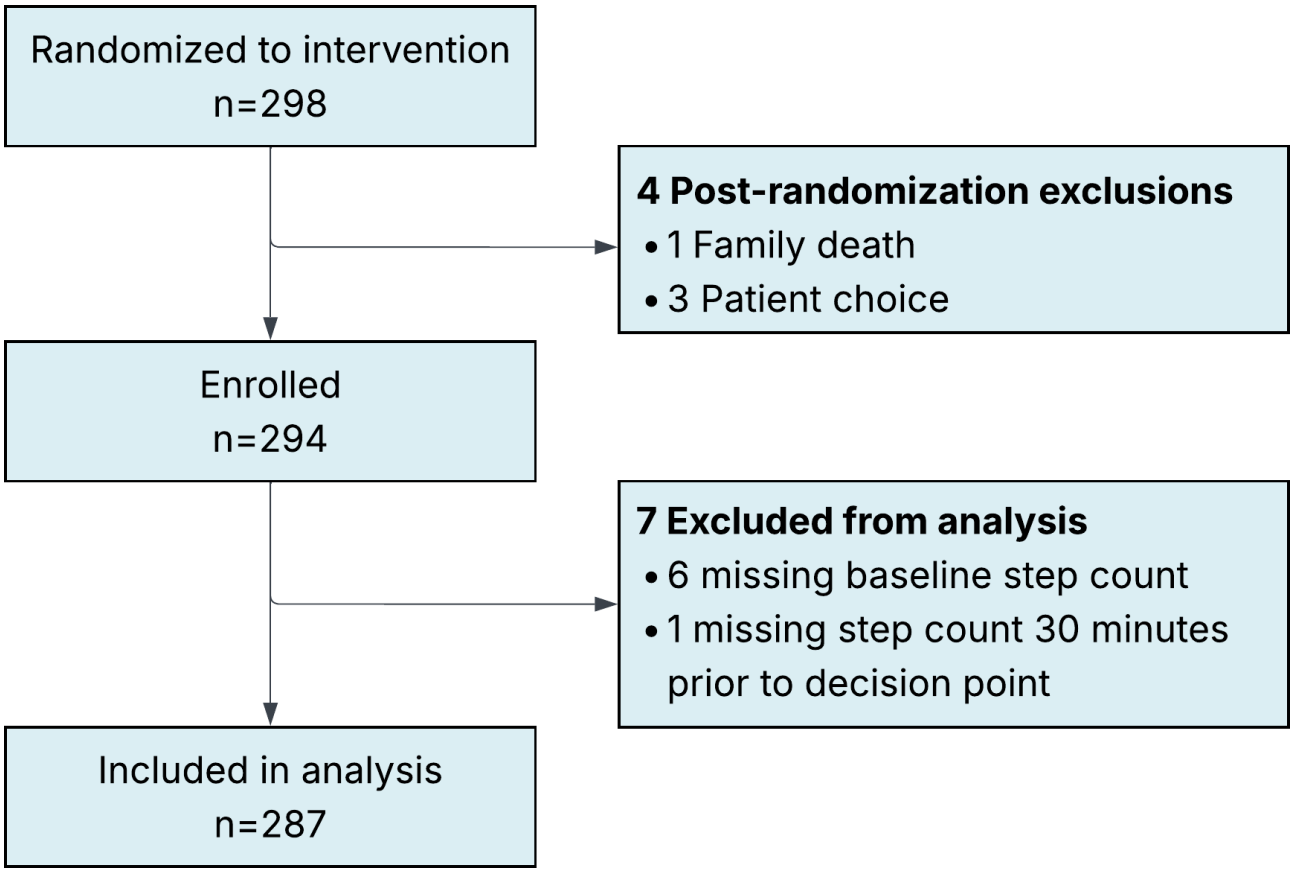

Participants were enrolled in the study for a median of 180 (IQR 37 - 180) days with a median intervention duration of 171 (IQR 25 - 173) days. The baseline characteristics of the population are displayed in Table 1. Participants had a mean age of 59.5 (SD 13.1) years, 138 (48.1%) were women, and 206 (71.8%) were White. Most were from the University of Michigan Health (242/287, 84.3%). Participants' baseline step count was 7408.1 (SD 3611.9) steps per day, and baseline sodium intake was 3072.7 (SD 1049.9) mg/d.

XSL•FO
RenderX

Table . Demographic and clinical characteristics of study population (N=287).

Characteristics	Ann Arbor (n=242)	Flint (n=45)	Overall
Age (y), mean (SD)	61.35 (12.5)	49.24 (11.6)	59.5 (13.1)
Gender, n (%)			
Woman	106 (43.8)	32 (71.1)	138 (48.1)
Man	136 (56.2)	13 (28.9)	149 (51.9)
Race, n (%)			
Asian	28 (11.6)	0 (0)	28 (9.8)
Black	20 (8.3)	20 (44.4)	40 (13.9)
Multiple	2 (0.8)	4 (8.9)	6 (2.1)
Other[a]	5 (2.1)	2 (4.4)	7 (2.4)
White	187 (77.3)	19 (42.2)	206 (71.8)
Ethnicity, n (%)			
Hispanic	11 (4.5)	1 (2.2)	12 (4.2)
Non-Hispanic	231 (95.5)	44 (97.8)	275 (95.8)
Baseline step count (steps/d) mean (SD)	7633.09 (3564.1)	6197.87 (3666.5)	7408.06 (3611.9)
Baseline estimated dietary sodium intake (mg/d), mean (SD)	3054.95 (1016.0)	3167.96 (1223.4)	3072.67 (1049.7)
Self-reported comorbid conditions, n (%)			
Chronic kidney disease	12 (5.0)	4 (8.9)	16 (5.6)
High cholesterol	109 (45.0)	16 (35.6)	125 (43.6)
Depression	39 (16.1)	13 (28.9)	52 (18.1)
Coronary artery disease	9 (3.7)	1 (2.2)	10 (3.5)
Stroke	7 (2.9)	2 (4.4)	9 (3.1)
Diabetes mellitus	57 (23.6)	9 (20.0)	66 (23.0)

[a]Other race: American Indian, Native Hawaiian or Other Pacific Islander, or Other or refused to answer race question

Study Execution

Over the duration of the intervention, participants were randomized to receive or not receive a push notification at 187,517 decision points (ie, time points; median 679, IQR 89‐694 decision points per participant). On average, participants had 4.0 (SD 0.2) decision points per day and were randomized to receive a push notification at 24.9% (SD 0.2) of decision points. These were nearly equally split between activity notifications (50.4%; SD 0.4) and dietary notifications (49.8%; SD 0.4), consistent with the study design. On average, participants were randomized to receive 0.96 (SD 0.86) push notifications per day, distributed nearly equally over the 4 time points (Tables S1-S3 in Multimedia Appendix 1). After accounting for participants' availability (considered available at mean 81.3%, SD 0.3 of decision points), participants were presumed to have received 0.79 (SD 0.83) push notifications each day.

Activity Intervention

The mean step count in the 60 minutes following a decision point was 456.1 (SD 707.0) steps. In a multivariable model accounting for the demographic and baseline characteristics of participants, activity notifications did not significantly impact 60-minute step count (estimate 1.01, 95% CI 0.98‐1.04; *P*=.40; Tables S4 and S5 in Multimedia Appendix 1; Figure 2). In a subsequent sensitivity analysis excluding decision points at which participants had zero hour or missing wear time in the 3 hours after a decision point (17,452, 9.6% of decision points), the results were similar with no significant change in 60-minute step count (Table S6 in Multimedia Appendix 1). To understand potential time-varying effects, we performed an exploratory analysis in which we modeled the intervention effect as a quadratic term over time, which did not significantly impact intervention efficacy (Figure 2). Similarly, in a sequence of univariable moderator analyses, age, community, and gender did not significantly impact intervention efficacy. In an exploratory analysis, however, push notifications were more effective in the subgroup of participants that were less active at baseline (dichotomized based on mean daily step count), increasing 60-minute step count by 4% (estimate 1.04, 95% CI 1.00‐1.08; *P*=.03).

XSL•FO
RenderX

Figure 2. Impact of receiving (A) an activity notification on step count 60 minutes after a decision point or (B) a dietary notification on lower-sodium food choices 24 hours after a decision point. The models show treatment effect over time, with time modeled as a quadratic term. Activity notifications and dietary notifications did not significantly impact step count or the number of lower-sodium food choices, respectively. The results were exponentiated for interpretability.

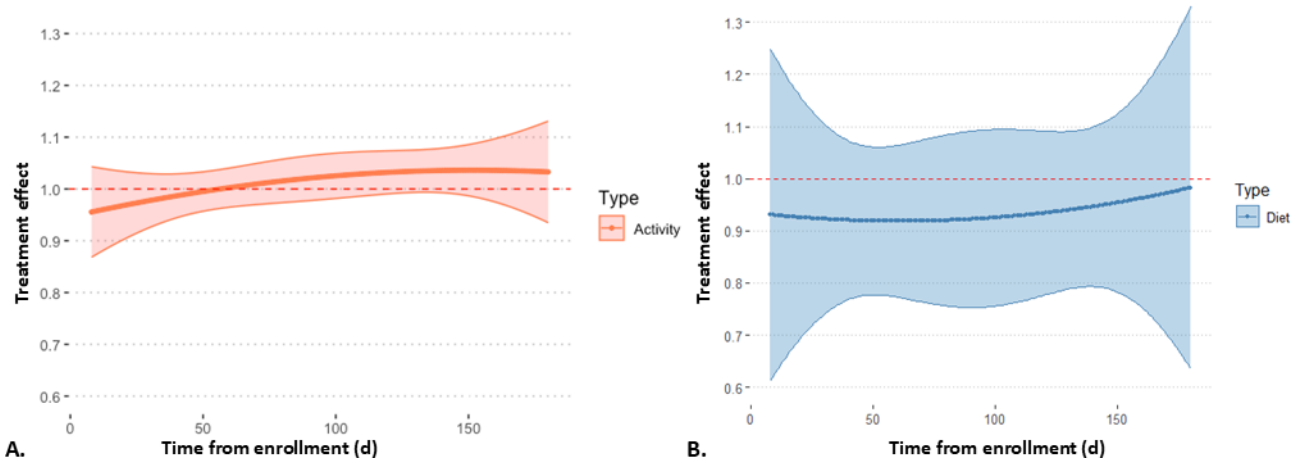

Dietary Intervention

In a multivariable model accounting for demographic and baseline characteristics, dietary notifications did not change the number of lower-sodium food choices in the 24 hours after a decision point (estimate 0.93, 95% CI 0.83 - 1.04; *P*=.23; Tables S4 and S7 in Multimedia Appendix 1; Figure 2). The results were similar in all subgroups of the population and did not change over time.

Subsequently, we conducted an exploratory analysis to evaluate the impact of push notifications on immediate measures of intervention engagement. Dietary notifications increased mobile app use in the subsequent 60 minutes by 95.5% (estimate 1.96, 95% CI 1.81 - 2.10; *P*< .001; Figure 3). Similarly, dietary push notifications increased the number of mobile app clicks or searches by 93.7% (estimate 1.94, 95% CI 1.72% - 2.16%; *P*<.001) and increased the number of low sodium food searches by 113% (estimate 2.13, 95% CI 1.73 - 2.53; *P*<.001), 60 minutes after a decision point (Figure 3). In all cases, the effect of dietary push notifications was the highest at the beginning of the study and decreased over time, though it remained significant throughout the study period.

Figure 3. Intervention engagement after dietary notifications. Dietary notifications significantly increased (**A**) mobile app use, (**B**) number of mobile app searches/clicks, and (**C**) number of low sodium searches, all within 60 minutes of a decision point. The models show treatment effect over time, with time modeled as a quadratic term. The results were exponentiated for interpretability.

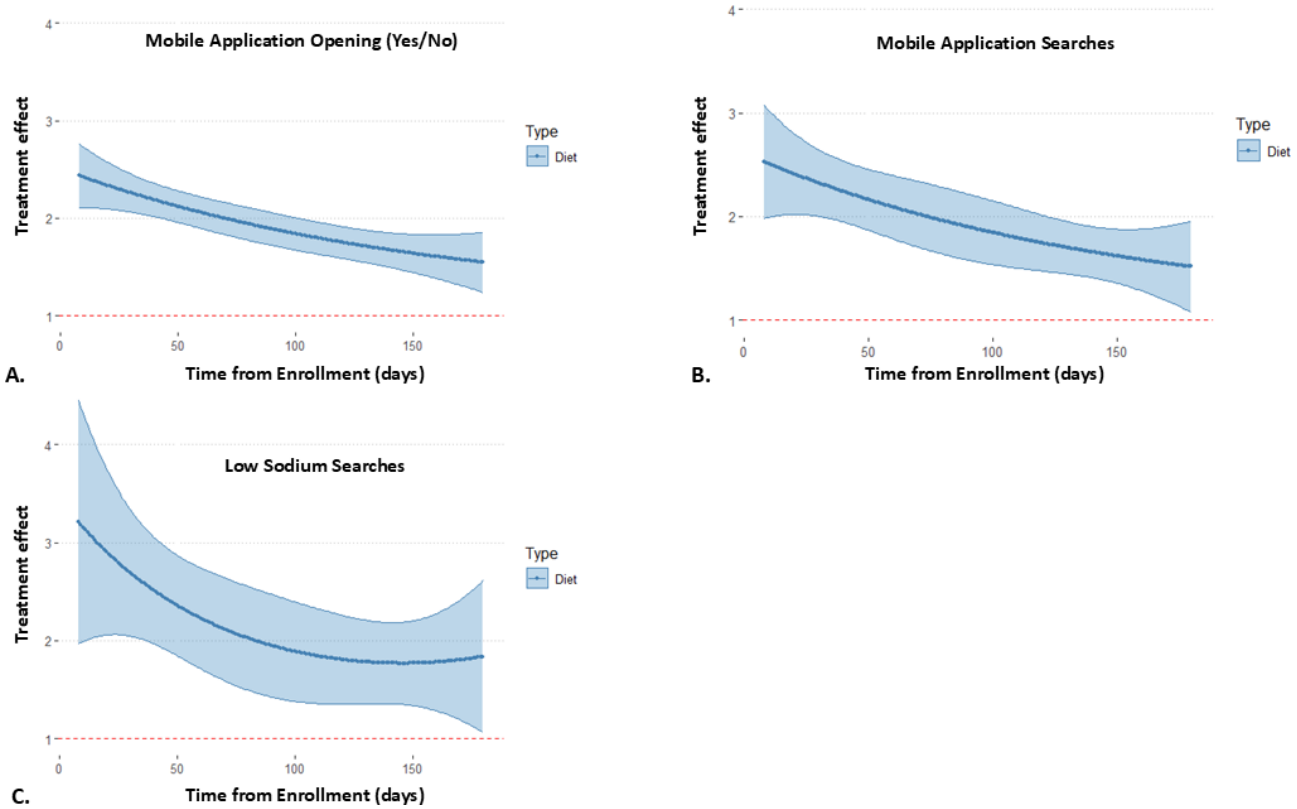

XSL•FO
RenderX

Discussion

Principal Findings

Digital hypertension solutions, typically delivered as multicomponent interventions, have shown promise for enhancing BP control [2,12,13]. However, the independent effects of push notifications when offered as part of a comprehensive digital hypertension intervention remain unknown. This is especially true with regard to the impact of push notifications on behavior change. We found through this trial that push notifications, tailored to participant (eg, community, mobility, and time) and environmental (eg, weather) factors, did not increase short-term physical activity levels or reduce sodium intake despite improving overall measures of these outcomes at 6 months. In an exploratory analysis, however, push notifications promoting physical activity were more effective for less active individuals, though this effect was overall very small, and for the overall cohort, the intervention was insufficient to overcome a participant's existing behavioral inertia. We also identified through an exploratory analysis that notifications that encouraged participants to interact with the mobile app and select lower-sodium food choices led to a nearly 2-fold increase in mobile app engagement, a measure of mechanical engagement though not necessarily effective behavioral engagement. While this effect persisted over the 6-month study period, the magnitude of effect decremented over time, suggesting either internalization of the behavior of interest (ie, engagement with the mobile app was no longer necessary for promoting adherence to a lower-sodium diet) or habituation.

Interpreting these results is best done within the context of those from the larger randomized controlled trial in which we found that the digital intervention package, consisting of both a mobile app and push notifications, increased physical activity levels (mean difference 489 steps, 95% CI 22-956) and reduced sodium intake (mean difference –285 mg, 95% CI −462 to –108) at 6 months [5]. One possible explanation for these results is that the interaction with the mobile app, which included multiple behavior change techniques (ie, goal setting, self-monitoring, and feedback) but not necessarily the content of the push notifications themselves, mediated the observed effects. This is supported by prior literature demonstrating positive associations between engagement with digital health interventions and health outcomes [14-16]. An alternative explanation, however, would be that push notifications did mediate the observed results, but that the proximal outcomes did not adequately capture their long-term impact. This could be due to lagged effects (ie, increase in physical activity beyond 60 min), incompletely captured effects (ie, failure to record lower-sodium food choices within the mobile app), or slower-developing mediators of the distal outcome (ie, knowledge about lower-sodium food choices, salience of activity goals). Future interventions should consider incorporating both proximal and intermediate outcomes within their study design, with the latter potentially capturing lagged effects (eg, 24 h step count and lower-sodium food choices over 48 h) or more slowly developing mediators of the target behavior.

Comparison With Prior Work

In general, digital hypertension solutions have focused on promoting BP self-monitoring, medication management, and lifestyle modification [2]. However, a challenge within the field has been in understanding the relative contributions of different digital intervention components, particularly when delivered as 1 part of a comprehensive intervention. In the pragmatic, randomized controlled trial BP Home, for example, over 2000 patients were randomized to BP self-monitoring using a standard device or to enhanced self-monitoring using a connected smartphone app [17]. Both groups experienced similarly large reductions in SBP (>10 mm Hg), without significant differences between the 2 groups. A second study found no significant differences in SBP reduction at 6 months using an artificial intelligence–enhanced conversational smartphone app promoting self-management, BP measurement, and lifestyle management, compared to a regular smartphone app paired with a home BP monitor [18]. There was, however, greater self-confidence in controlling BP in the intervention group and a trend toward greater self-reported physical activity.

To disentangle the effects of varying intervention components, Tucker et al [12] conducted an individual participant meta-analysis with over 7000 patients from 15 studies. Self-monitoring was associated with reduced SBP at 12 months, though this effect was strongly influenced by the intensity of co-interventions, with no effect with self-monitoring alone and a greater effect when combined with more intensive interventions, such as lifestyle counseling or systematic medication titration. These results have been supported by those from other studies, which, while mixed, suggest in aggregate that successful interventions are those with more comprehensive functionality and that incorporate medication management in conjunction with a care team [2,13]. A limitation of these and other studies, however, has been their focus on evaluating the distal effects of an intervention package, often delivered through a mobile app with multiple behavioral components. Our microrandomized trial analysis overcomes these limitations by isolating the impact of push notifications on shorter-term mediators of the desired long-term effect (eg, lifestyle management) and suggests how the mobile intervention package may impact long-term behavior change.

Study Strengths

Our study has several strengths. First, this is one of the first microrandomized trials that we are aware of among individuals with hypertension. Microrandomized trials are a novel experimental design, and as such, few studies have been performed among individuals with cardiovascular disease or cardiovascular risk factors. By incorporating serial randomization, microrandomized trials can provide insight into the causal effects of an intervention component over time in such patients. This serves to advance the science of behavioral interventions, which have traditionally been optimized using a series of randomized controlled trials. Randomized controlled trials, however, are designed to assess the average effect of an intervention package on a behavior of interest and not to investigate which components of an intervention are most efficacious, their time-varying effects, or what psychosocial or

XSL•FO
RenderX

contextual factors impact their efficacy [4,19]. Second, we delivered the trial remotely and enrolled participants from 2 sites, including a series of federally qualified health centers, enhancing the diversity of the study population and the generalizability of our findings. In general, the data around mHealth interventions for individuals with hypertension with digital barriers or from underrepresented groups have been limited [20]. Finally, we followed participants for 6 months, addressing a critical limitation of many mHealth interventions, namely the short time horizon.

Limitations

This study should be interpreted within the context of its limitations, which relate both to the study population and the intervention design and analysis. First, we only enrolled participants with a compatible smartphone. Smartphone ownership, however, is common across age, race, and socioeconomic groups, and we ensured the use of a broad range of possible devices [21]. Second, participants were active at baseline (mean 7408 [SD 3611.9] steps per day), had only mildly elevated BP, and were connected to their health care system. Given that this population was already active at baseline, this may have led to a ceiling effect. Similarly, the intervention may have been perceived as less salient to participants with relatively well-controlled BPs. Whether the results would be similar with less active individuals or to those with poorly controlled BP is unknown. Third, push notifications were tailored on a limited number of environmental and participant-level factors. It is unknown whether notifications would be more effective in promoting lifestyle modification if tailored to psychosocial constructs or an extended set of participant factors (eg, recent physical activity). We also did not evaluate notification efficacy based on message framing, nor did we evaluate treatment moderation based on attributes of the notifications (eg, time of day, weather). For example, in a 2×2 randomized experiment of over 500 participants designed to evaluate message framings, participants preferred the ability to choose message framing (autonomy-supportive vs controlling language) over the presumed preference for autonomy-supportive language [22]. Future studies should consider an expanded set of tailoring variables and evaluate for treatment moderation by the characteristics of the notifications. Fourth, we required that participants self-report low sodium choices within the mobile app. It is thus possible that dietary notifications increased the number of low sodium food choices, though these were not reported by participants in the mobile app. Fifth, it is possible that the effect size of notifications may have been less than anticipated, leading us to underpower the study for our proximal outcomes. Finally, we required that participants be wearing their smartwatches to be considered available for the physical activity intervention. We assumed that participants were wearing their smartwatches if they had 1 or more heart rate measurements in the 30 minutes before a message was sent. Although this increased the rigor of our analyses, we cannot confirm that participants were wearing their smartwatches when push notifications were sent. It is also possible that push notifications had a delayed or unmeasured effect for participants not wearing their smartwatches or who received the push notifications at a later time.

Conclusions and Future Directions

In conclusion, in this microrandomized trial, we demonstrated that tailored push notifications did not increase short-term physical activity levels or reduce sodium intake among individuals with hypertension. We did observe greater mobile app engagement following push notifications, which may have mediated the observed longer-term effects on sodium intake and physical activity levels. These results suggest that push notifications may be effective in promoting intervention engagement. Additional studies, however, are needed to identify which individuals benefit most from push notifications and how to optimally tailor push notifications based on environmental and psychosocial factors, particularly when delivered as part of multicomponent interventions. Furthermore, future studies may consider mediation analyses to enhance our understanding of the impact of intervention engagement on longer-term measures of behavior change and which evaluate the impact of novel methodological approaches such as reinforcement learning algorithms on notification efficacy.

Funding

This study received funding from American Heart Association, Inc (AWD014891).

Data Availability

The data and code that support this study are openly available through GitHub [23].

Authors' Contributions

JRG, MPD, PK, MWN, LES, WD, and BKN contributed to the conceptualization of the project, investigation, and methodology. MPD, MWN, LES, and BKN were responsible for funding acquisition. EL was responsible for data curation. YC and TB were responsible for the formal analysis under the oversight of JRG, WD, and BKN. JRG and MPD drafted the original manuscript (writing – original draft) which was reviewed and edited by all co-authors (writing – review & editing).

Conflicts of Interest

MPD receives funding from the Agency for Health Research and Quality, the National Heart, Lung, and Blood Institute, and the American Heart Association. JRG receives funding from the National Institutes of Health (NIH; 1K23HL168220) and the Patient-Centered Outcomes Research Institute and from Blue Cross and Blue Shield of Michigan for her role in the Blue Cross

XSL•FO
RenderX

Blue Shield of Michigan Cardiovascular Consortium. LES receives funding from the National Institute of Neurologic Diseases and Stroke, the National Institute of Minority Health and Health Disparities, and the American Heart Association. BKN receives funding from the NIH, VA HSR&D, American Heart Association, and Janssen. He also receives compensation as Editor-in-Chief of Circulation: Cardiovascular Quality & Outcomes, a journal of the American Heart Association. Finally, he is a co-inventor on US Utility Patent US15/356,012 (US20170148158A1) entitled "Automated Analysis of Vasculature in Coronary Angiograms," which uses software technology with signal processing and machine learning to automate the reading of coronary angiograms, held by the University of Michigan. The patent is licensed to AngioInsight, Inc., where BKN holds ownership shares and receives consultancy fees. MWN receives funding from the National Heart, Lung, and Blood Institute and the National Institute on Drug Abuse. PK is a principal investigator or a co-investigator on research grants from the NIH.

Multimedia Appendix 1
Supplemental tables and figures.
[DOCX File, 238 KB - jmir_v28i1e78218_app1.docx]

Checklist 1
CONSORT 2025 checklist.
[PDF File, 216 KB - jmir_v28i1e78218_app2.pdf]

References

1. Martin SS, Aday AW, Almarzooq ZI, et al. 2024 heart disease and stroke statistics: a report of US and global data from the American Heart Association. Circulation 2024 Feb 20;149(8):e347-e913. [doi: 10.1161/CIR.0000000000001209] [Medline: 38264914]
2. Digital hypertension management solutions. Peterson Health Technology Institute. 2024. URL: https://phti.org/assessment/digital-hypertension-management-solutions/ [accessed 2025-12-26]
3. Golbus JR, Jeganathan VSE, Stevens R, et al. A physical activity and diet just-in-time adaptive intervention to reduce blood pressure: the myBPmyLife study rationale and design. J Am Heart Assoc 2024 Jan 16;13(2):e031234. [doi: 10.1161/JAHA.123.031234] [Medline: 38226507]
4. Golbus JR, Dempsey W, Jackson EA, Nallamothu BK, Klasnja P. Microrandomized trial design for evaluating just-in-time adaptive interventions through mobile health technologies for cardiovascular disease. Circ Cardiovasc Qual Outcomes 2021 Feb;14(2):e006760. [doi: 10.1161/CIRCOUTCOMES.120.006760] [Medline: 33430608]
5. Dorsch MP, Golbus JR, Stevens R, et al. Physical activity and diet just-in-time adaptive intervention to reduce blood pressure: a randomized controlled trial. NPJ Digit Med 2025 Jul 14;8(1):438. [doi: 10.1038/s41746-025-01844-3] [Medline: 40659778]
6. Tangney CC, Rasmussen HE, Richards C, Li M, Appelhans BM. Evaluation of a brief sodium screener in two samples. Nutrients 2019 Jan 14;11(1):166. [doi: 10.3390/nu11010166] [Medline: 30646541]
7. Nahum-Shani I, Hekler EB, Spruijt-Metz D. Building health behavior models to guide the development of just-in-time adaptive interventions: a pragmatic framework. Health Psychol 2015 Dec;34S:1209-1219. [doi: 10.1037/hea0000306] [Medline: 26651462]
8. Hellem AK, Casetti A, Bowie K, et al. A community participatory approach to creating contextually tailored mHealth notifications: myBPmyLife project. Health Promot Pract 2024 May;25(3):417-427. [doi: 10.1177/15248399221141687] [Medline: 36704967]
9. Boruvka A, Almirall D, Witkiewitz K, Murphy SA. Assessing time-varying causal effect moderation in mobile health. J Am Stat Assoc 2018;113(523):1112-1121. [doi: 10.1080/01621459.2017.1305274] [Medline: 30467446]
10. Golbus JR, Shi J, Gupta K, et al. Text messages to promote physical activity in patients with cardiovascular disease: a micro-randomized trial of a just-in-time adaptive intervention. Circ Cardiovasc Qual Outcomes 2024 Jul;17(7):e010731. [doi: 10.1161/CIRCOUTCOMES.123.010731] [Medline: 38887953]
11. Liu X, Qian T, Bell L, Chakraborty B. Incorporating nonparametric methods for estimating causal excursion effects in mobile health with zero-inflated count outcomes. Biometrics 2024 Mar 27;80(2):ujae054. [doi: 10.1093/biomtc/ujae054] [Medline: 38837902]
12. Tucker KL, Sheppard JP, Stevens R, et al. Self-monitoring of blood pressure in hypertension: a systematic review and individual patient data meta-analysis. PLoS Med 2017 Sep;14(9):e1002389. [doi: 10.1371/journal.pmed.1002389] [Medline: 28926573]
13. Alessa T, Abdi S, Hawley MS, de Witte L. Mobile apps to support the self-management of hypertension: systematic review of effectiveness, usability, and user satisfaction. JMIR mHealth uHealth 2018 Jul 23;6(7):e10723. [doi: 10.2196/10723] [Medline: 30037787]
14. Mclaughlin M, Delaney T, Hall A, et al. Associations between digital health intervention engagement, physical activity, and sedentary behavior: systematic review and meta-analysis. J Med Internet Res 2021 Feb 19;23(2):e23180. [doi: 10.2196/23180] [Medline: 33605897]

XSL•FO
RenderX

15. Alshurafa N, Jain J, Alharbi R, Iakovlev G, Spring B, Pfammatter A. Is more always better?: Discovering incentivized mHealth intervention engagement related to health behavior trends. Proc ACM Interact Mob Wearable Ubiquitous Technol 2018 Dec;2(4):153. [doi: 10.1145/3287031] [Medline: 32318650]
16. Schoeppe S, Alley S, Van Lippevelde W, et al. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review. Int J Behav Nutr Phys Act 2016 Dec 7;13(1):127. [doi: 10.1186/s12966-016-0454-y] [Medline: 27927218]
17. Pletcher MJ, Fontil V, Modrow MF, et al. Effectiveness of standard vs enhanced self-measurement of blood pressure paired with a connected smartphone application: a randomized clinical trial. JAMA Intern Med 2022 Oct 1;182(10):1025-1034. [doi: 10.1001/jamainternmed.2022.3355] [Medline: 35969408]
18. Persell SD, Peprah YA, Lipiszko D, et al. Effect of home blood pressure monitoring via a smartphone hypertension coaching application or tracking application on adults with uncontrolled hypertension: a randomized clinical trial. JAMA Netw Open 2020 Mar 2;3(3):e200255. [doi: 10.1001/jamanetworkopen.2020.0255] [Medline: 32119093]
19. Klasnja P, Hekler EB, Shiffman S, et al. Microrandomized trials: an experimental design for developing just-in-time adaptive interventions. Health Psychol 2015 Dec;34S:1220-1228. [doi: 10.1037/hea0000305] [Medline: 26651463]
20. Khoong EC, Olazo K, Rivadeneira NA, et al. Mobile health strategies for blood pressure self-management in urban populations with digital barriers: systematic review and meta-analyses. NPJ Digit Med 2021 Jul 22;4(1):114. [doi: 10.1038/s41746-021-00486-5] [Medline: 34294852]
21. Mobile fact sheet. Pew Research Center. 2025. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2025-12-30]
22. Smit ES, Zeidler C, Resnicow K, de Vries H. Identifying the most autonomy-supportive message frame in digital health communication: a 2x2 between-subjects experiment. J Med Internet Res 2019 Oct 30;21(10):e14074. [doi: 10.2196/14074] [Medline: 31670693]
23. MRT_myBPmyLife. GitHub. URL: https://github.com/WIRED-L-um/MRT_myBPmyLife/blob/main/README.md [accessed 2025-12-29]

Abbreviations

BP: blood pressure
CONSORT : Consolidated Standards of Reporting Trials
mHealth: mobile health
SBP: systolic blood pressure
WCLS: weighted and centered least square

Edited by A Coristine; submitted 29.May.2025; peer-reviewed by AA Lopez-Gonzalez, RJ Katz; revised version received 10.Dec.2025; accepted 10.Dec.2025; published 07.Jan.2026.

Please cite as:
Golbus JR, Dorsch MP, Chen Y, Basu T, Luff E, Klasnja P, Newman MW, Skolarus LE, Dempsey W, Nallamothu BK
Impact of Push Notifications on Physical Activity and Sodium Intake Among Patients with Hypertension: Microrandomized Trial of a Just-in-Time Adaptive Intervention
J Med Internet Res 2026;28:e78218
URL: https://www.jmir.org/2026/1/e78218
doi:10.2196/78218

© Jessica Rachel Golbus, Michael P Dorsch, Yuxuan Chen, Tanima Basu, Evan Luff, Predrag Klasnja, Mark W Newman, Lesli E Skolarus, Walter Dempsey, Brahmajee K Nallamothu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effects of an eHealth Cardiac Exercise Rehabilitation Platform for Patients After Percutaneous Coronary Intervention Based on the Persuasive Systems Design Model: Randomized Controlled Trial

Yang Liu[1], MSN; Xiting Huang[1,2], MM; Ziying Dai[1], MM; Zhili Jiang[3], MSN; Wenxiao Wu[1], MSN; Jing Wang[1], MSN; Zhiqian Wang[1], MSN; Luyao Yu[1], MSN; Hanyu Li[1], MSN; Lihua Huang[1], MBBS

[1]Department of Nursing, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China

[2]Department of Nursing, Sun Yat-sen University Cancer Center, Guangzhou, China

[3]School of Humanities and Management, Zhejiang Chinese Medical University, Hangzhou, China

Corresponding Author:
Lihua Huang, MBBS
Department of Nursing
The First Affiliated Hospital
Zhejiang University School of Medicine
79 Qingchun Road
Shangcheng District
Hangzhou, 310003
China
Phone: 86 13867129329
Email: lihuahuang818@zju.edu.cn

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2026/1/e99326/

Abstract

Background: Cardiac exercise rehabilitation is an important intervention for disease management of patients with coronary heart disease (CHD) after percutaneous coronary intervention (PCI). Still, the participation and compliance with exercise rehabilitation remain suboptimal. Mobile health technology is a promising approach to promoting involvement in cardiac exercise rehabilitation. Remote rehabilitation can overcome the problems existing in traditional rehabilitation.

Objective: This study aimed to evaluate the effects of an eHealth cardiac rehabilitation (CR) platform based on the persuasive systems design model in addition to standard CR after PCI on physical activity (PA), exercise endurance, self-perceived fatigue, exercise self-efficacy (ESE), and quality of life for patients after PCI.

Methods: A single-blinded, parallel, randomized controlled trial design was used. The study was conducted in the Department of Cardiology of a tertiary hospital in Hangzhou, China. A total of 180 eligible patients with CHD were enrolled from June to December 2023. Participants were randomly assigned (1:1) to the intervention group or the control group, with 90 patients in each group. The study is a 24-week eHealth CR program. The primary outcome was PA level; the secondary outcomes included exercise endurance, self-perceived fatigue, ESE, and quality of life. Data on the primary and secondary outcome measures were collected at baseline (T0), at 12 weeks of intervention (T1), and at 4 (T2), 8 (T3), and 12 (T4) weeks of follow-up. The generalized estimating equation model was used to examine changes in the outcome variables between the 2 groups across the study end points.

Results: Generalized estimating equation analyses revealed significant group-by-time interactions for all outcome measures (all $P<.001$). At T4, compared with the control group, the intervention group demonstrated statistically significant improvements in the following outcomes: PA: median 1723.00 versus 805.50 Metabolic Equivalent Task minutes per week (β coefficient=937.29, 95% CI 867.61-1006.97); 6-minute walk distance: median 436.00 versus 405.00 m (β coefficient=31.00); self-perceived fatigue: median 9.00 versus 10.00 (β coefficient=–1.00, indicating reduced fatigue); ESE: 61.11 versus 27.78 (β coefficient=33.33); Short Form of 36 Health Survey Questionnaire score: 91.19 versus 84.13 (β coefficient=7.06; all $P<.001$). Notably, there was no significant difference in self-perceived fatigue between the 2 groups at T1 ($P=.50$).

XSL•FO
RenderX

Conclusions: The findings of this study demonstrate the effectiveness of the eHealth CR based on the persuasive systems design model in addition to standard CR after PCI in improving the PA level, exercise endurance, ESE, quality of life, and self-perceived fatigue of patients. These findings also provide insights into the application of an eHealth cardiac exercise rehabilitation interventions to enhance the rehabilitation of patients with CHD.

Trial Registration: China Clinical Trials Registry (ChiCTR) ChiCTR2300071666; https://www.chictr.org.cn/showprojEN.html?proj=197908

(J Med Internet Res 2026;28:e71450) doi:10.2196/71450

KEYWORDS

coronary heart disease; percutaneous coronary intervention; PCI; exercise rehabilitation; persuasive systems design; eHealth

Introduction

Background

Cardiovascular diseases (CVDs) have long been a global health scourge, exacting a heavy toll on both individual well-being and health care systems worldwide. The World Health Organization (WHO) highlights that CVDs claim approximately 17.9 million lives annually, constituting a staggering 32% of all global deaths [1]. In China, the situation is equally disconcerting. With the rapid pace of urbanization, changes in lifestyle, and an aging population, the prevalence of CVDs, especially coronary heart disease (CHD), has been on an upward trajectory [2,3]. This has led to a substantial increase in the number of patients undergoing percutaneous coronary intervention (PCI), a common and effective treatment for CHD.

Despite the remarkable advancements in PCI technology, which significantly improve the acute condition of patients by restoring blood flow to the heart, it is not a cure-all solution. Patients post PCI often face a plethora of challenges. Exercise intolerance is a prevalent issue, as a large proportion of these patients experience a decline in their physical capacity, which restricts their daily activities and quality of life [4]. Self-perceived fatigue is another common complaint, which can be attributed to the physiological stress of the intervention, underlying cardiac damage, and the body's recovery process [5]. Moreover, many patients struggle with suboptimal exercise self-efficacy (ESE), lacking the confidence to engage in regular physical activity (PA), which is crucial for their long-term cardiac health [6].

Cardiac rehabilitation (CR) has emerged as an essential component of post-PCI care. It is a comprehensive, multidisciplinary program that encompasses exercise training, risk factor modification, psychological counseling, and patient education [7]. Extensive research has demonstrated its effectiveness in improving exercise capacity, reducing cardiovascular risk factors, enhancing psychological well-being, and ultimately decreasing mortality and morbidity rates among patients post PCI [8]. For instance, a network meta-analysis found that multiple exercise interventions enhanced peak oxygen consumption, with high-intensity interval training showing the greatest effect, followed by combined water-based and moderate-intensity continuous training, and other interventions like combined aerobic and resistance exercise also demonstrated benefits [9].

However, traditional CR programs, which are typically centered around in-hospital or clinic-based services, are plagued with limitations. Geographical barriers pose a significant challenge, especially for patients living in rural or remote areas, who may have to travel long distances to access these services [10]. Time constraints are another hurdle, as many patients, particularly those who are still working or have family responsibilities, find it difficult to fit regular rehabilitation sessions into their busy schedules. Additionally, the cost associated with in-person rehabilitation, including transportation and potential loss of income during treatment, can be a deterrent for some patients [11]. The proportion of patients participating in the CR program is only 25% to 35% [12].

In recent years, remote cardiac rehabilitation (RCR) has been advancing rapidly, and the potential exists for the challenges of traditional facility-based CR programs to be addressed by delivering care to patients in the convenience of their own homes with real-time, personalized support [11]. The RCR uses different methods such as the internet, wearable devices, and mobile apps [13]. Despite the positive outcomes observed in the application of digital health interventions for CR, such as SMS, remote electrocardiographic monitoring, and mobile or web portal tools, these advancements have predominantly remained in research settings and have not yet been adopted widely in clinical practice [14]. A systematic review by Duff et al [15] assessed the application of behavior change techniques (BCTs) in eHealth interventions designed to increase PA in CVD, and identified feedback and monitoring as the most common BCT category implemented in these interventions. Behavior change theory proponents note that these theories explain how behavior change happens [16]. However, information system developers often prioritize using BCTs over understanding the underlying theories [17]. Systems lacking a strong theoretical basis may have conflicting mechanisms, harming long-term effectiveness.

Such systems are designed to form, alter, or reinforce the attitudes, behaviors, or compliance of their users voluntarily. A key element in behavior and attitude change is persuasion. The persuasive systems design (PSD) model addresses this gap by guiding the analysis of the persuasion context, including recognizing the intent of the persuasion, understanding the persuasion event, and defining the strategies in use [18]. It describes how to inject persuasive characteristics into the system to stimulate behavior change in the design process. In the functional design of the persuasive system, the characteristics of the system are divided into main task support, dialogue support, system credibility support, and social support [19], and a total of 28 persuasive principles are included. It fully explains

XSL•FO
RenderX

how the design principles are translated into software requirements and further manifests the system characteristics. The information system developed and designed under the guidance of the framework can motivate users to implement self-management and trigger health behavior change [20].

Aims of This Research

In the context of post-PCI CR, there is a paucity of well-designed mobile health (mHealth) interventions that fully leverage the PSD model to comprehensively address patients' multifaceted needs. Existing studies either focus on single-aspect interventions or fail to fully capitalize on the potential of the PSD model. Therefore, this study aimed to develop and evaluate an eHealth CR platform grounded in the PSD model in addition to standard CR for patients post PCI. We hypothesized that such a platform could effectively improve patients' PA levels, exercise endurance, ESE, and quality of life, while reducing self-perceived fatigue.

Methods

Study Design

The study was a single-blinded, parallel, randomized controlled clinical trial. The study protocol complied with the Declaration of Helsinki. The report was in accordance with the CONSORT (Consolidated Standards of Reporting Trials) guidelines.

Participants

Participants were recruited in the Department of Cardiology of a tertiary hospital in Hangzhou, a provincial capital city in eastern China, from June to December 2023.

The inclusion criteria were as follows: (1) patients were older than 18 years old; (2) met the WHO diagnostic criteria of CHD, underwent PCI with 1 or 2 stents at first time, and were diagnosed as low-risk factors according to the risk classification of exercise rehabilitation of CHD; (3) patients with a left ventricular ejection fraction of 50% to 70%; (4) patients with cardiac function grade Ⅰ or Ⅱ; (5) transradial or ulnar artery puncture; (6) patients and their families agreed and actively cooperated with the exercise rehabilitation treatment; and (7) no language disorder, could read and speak Chinese, had no prescribed PA restrictions. Participants were excluded from the study if they (1) had presenting with cardiogenic shock, severe arrhythmia and cardiac function grade Ⅲ or above; (2) had severe liver and kidney dysfunction (liver dysfunction of Child-Pugh class B or C; estimated glomerular filtration rate <30 ml/[min·1.73m^2]), severe anemia (hemoglobin<60 g/L); (3) had severe pulmonary disease, such as chronic obstructive pulmonary disease, emphysema, bronchiectasis, and pulmonary heart disease; (4) had a history of falls within half a month, or were unable to exercise independently; (5) took antianxiety or depression drugs; (6) had previous history of cerebral hemorrhage; and (7) could not use WeChat (Tencent Holdings Limited) skillfully.

Sample Size

The sample size was determined by PA level based on a pilot trial, in which the SD was 133.81, and the mean difference between the 2 groups was 131. A total of 150 participants were required to detect a difference between the 2 groups at a 5% (2-sided) significant level with a power of 80%. Considering the loss to follow-up, the sample size of each group was at least 90, and the total sample size was at least 180, allowing a 20% dropout.

Randomization and Blinding

A total of 180 eligible participants were randomly assigned to the intervention or control group at a 1:1 ratio via block randomization (block size=4) to ensure balanced group sizes. The randomization was implemented by an independent statistician (not part of the core study team, nor involved in recruitment, assessment, or intervention) to avoid selection bias.

The statistician used SPSS (version 26.0; IBM Corp) to generate a random number sequence, pairing each number with a unique participant ID. Group assignment was predefined: even numbers for the intervention group and odd numbers for the control group. For allocation concealment, each ID, random number, and group assignment was sealed in a sequentially numbered opaque envelope, stored in a locked cabinet. Upon recruitment, the research coordinator, unaware of the sequence, retrieved the next envelope to inform the participant of their group.

Finally, 90 participants were allocated to each group. The intervention lasted for 12 weeks, with an additional 12-week follow-up period. Notably, an eHealth cardiac exercise rehabilitation based on the PSD model was provided in addition to standard CR; the volume and content of standard rehabilitation were consistent between the 2 groups to ensure that group differences in outcomes could be attributed to the eHealth platform intervention. Due to the distinguishable eHealth platform intervention, blinding of participants and intervention deliverers was unfeasible, but outcome assessors remained blinded. Thus, this was a single-blinded, randomized controlled trial.

Interventions

Both groups received standard CR to isolate the effect of the additional mHealth intervention in the intervention group. Standard care included in-hospital education, exercise guidance, and follow-up support.

In in-hospital education, a 60-minute group session was conducted by a cardiac nurse, covering PCI postcare knowledge included medication adherence, dietary management, warning signs of adverse events (AEs), and basic exercise principles.

In exercise guidance, a 30-minute one-on-one session was conducted with a physiotherapist to teach low-to-moderate intensity aerobic exercises, including brisk walking, stationary cycling, and resistance training, such as hand grip exercises suitable for home practice, with a recommended weekly volume of 150 minutes of moderate-intensity activity.

In the follow-up support, monthly 15-20 minutes each telephone check-ins were conducted to address patient questions, remind them of exercise and medication adherence, and collect brief feedback on physical status.

XSL•FO
RenderX

Intervention Group: Standard Care+mHealth Platform

Overview

The intervention group received a PSD model–based eHealth exercise rehabilitation in addition to standard CR. We established a development team composed of cardiology experts, physiotherapists, clinical nursing experts, evidence-based nursing experts, software engineers, and artists. According to the design content, we constructed an exercise rehabilitation WeChat applet for patients with CHD after PCI. The specific construction process of the WeChat mini program is detailed in Figure 1.

Figure 1. Schematic diagram of the persuasive systems design–based eHealth application development process for patients with coronary heart disease after percutaneous coronary intervention.

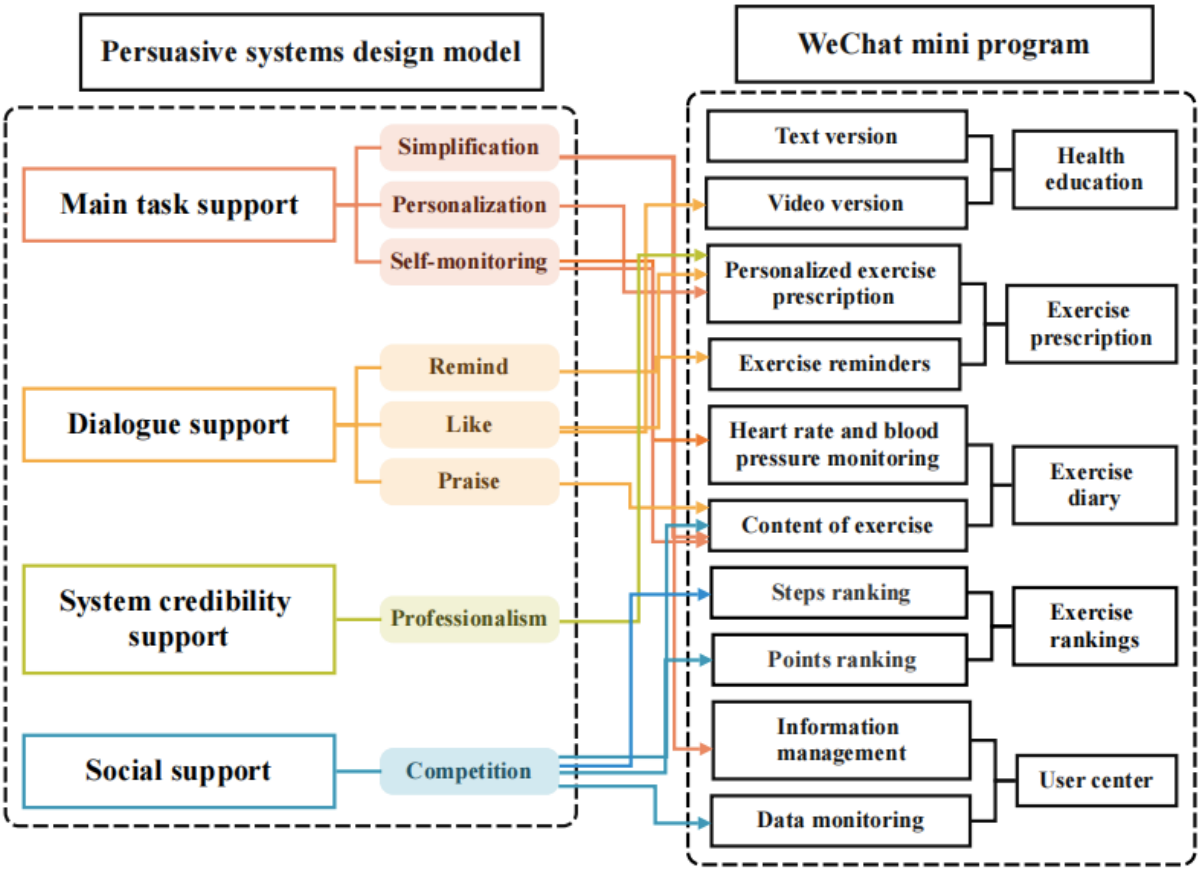

The eHealth platform was developed exclusively under the guidance of the PSD model—a theoretical framework specifically tailored for designing digital tools to stimulate behavior change. The platform's functional modules and design logic (schematic diagram illustrated in Figure 1) directly align with the 4 core components of the PSD model (main task support, dialogue support, system credibility support, and social support). The detailed integration of the PSD model into the intervention, its connection to global eHealth behavior change frameworks, and how it guided the platform's design are described further in this study.

Main Task Support: Simplifying Exercise Rehabilitation Adherence

PSD Theoretical Basis

The PSD model defines "main task support" as designing features to reduce user effort in completing core goals via principles like simplification and personalization. This component addresses a key barrier to digital health adherence—complex or nontailored tasks that increase user burden.

Integration Into the Platform

Simplification Principle

The home page was designed with a single "Exercise Prescription" button as the core entry point, reducing navigation steps from an average of 4 clicks to 1. Postexercise logging only requires 3 inputs (duration, heart rate, and discomfort), avoiding redundant data entry.

Personalization Principle

Based on baseline 6-minute walk distance (6MWD), left ventricular ejection fraction, and exercise preferences, the platform automatically generates individualized exercise plans (eg, 30-minute brisk walking for patients with 6MWD > 350 m vs 20-minute slow walking for 6MWD < 300 m)—aligned with PSD's focus on "matching tasks to user capabilities." This personalization is distinct from generic plans in non-PSD eHealth tools, which often use fixed templates.

XSL•FO
RenderX

Self-Monitoring Principle

Patients could record the data before and after each exercise in a diary, and track the progress of setting goals over time, so that patients could clearly understand their performance or status in the exercise rehabilitation process, and the rehabilitation platform can support patients to achieve self-reporting and motivate patients to achieve exercise goals.

Dialogue Support: Enhancing Real-Time Interaction and Feedback

PSD Theoretical Basis

"Dialogue support" in the PSD model refers to interactive features that guide behavior via reminders, praise, and preference matching—critical for maintaining user engagement in long-term rehabilitation.

Integration Into the Platform

Reminder Principle

Patients set customizable exercise reminders (eg, "7 PM daily walk") with pop-up and vibration alerts. The platform also sends adaptive reminders (eg, "You haven't exercised in 3 days—start with a 15-minute walk today") if adherence drops, addressing the "forgetfulness" barrier common in patients post PCI.

Praise Principle

After each exercise session, the platform generates immediate positive feedback (eg, "Great job! You completed 100% of today's plan") and weekly "adherence badges" (eg, "3-week Consistent Exercise Award"). This aligns with PSD's goal of reinforcing desired behaviors through positive reinforcement.

Like Principle

Patients can also receive "likes" from the research team and anonymous peers (via the "Exercise Ranking" module) for completing exercise plans or achieving personal goals. The platform sends real-time notifications of likes (eg, "3 peers praised your consistent exercise!"), leveraging social approval to enhance motivation—consistent with PSD's dialogue support, focus on interactive feedback.

System Credibility Support: Building Trust in Digital Guidance

PSD Theoretical Basis

"System credibility support" ensures users perceive the platform as reliable via professionalism and transparency—a prerequisite for accepting digital health advice, especially in cardiac care.

Integration Into the Platform: Professionalism Principle

All exercise prescriptions and educational content were co-developed by a team of 2 cardiologists and 1 CR nurse, with references to the guidelines of the European Society of Cardiology [21] displayed in the "Health Education" section.

Social Support: Leveraging Social Influence for Adherence

PSD Theoretical Basis

"Social support" in the PSD model uses competitive or cooperative features to motivate behavior via social comparison—proven effective in digital health interventions for exercise adherence.

Integration Into the Platform: Competitive Principle

The "Exercise Ranking" module displays anonymized weekly exercise completion rates of other participants (eg, "You are in the top 25% of users this week")—without personal identifiers to protect privacy. This feature leverages mild social comparison to boost motivation, consistent with PSD's focus on positive social influence [22] and avoiding the pressure of direct competition.

Iterative Phases

The PSD model–based eHealth platform was developed in 4 iterative phases, with pilot testing integrated to refine functionality:

Phase 1: Framework Design (Month 1-2)

Guided by the PSD model, we identified four core persuasive features for CR:

1. Primary task support: personalized exercise prescriptions aligned with frequency, intensity, time, and type principles to simplify task completion.

2. Dialogue support: real-time feedback and reminders to enhance user engagement, including praise, likes, and adaptive reminders.

3. Credibility support: evidence-based educational content cited from the guidelines of the European Society of Cardiology [21] to build trust.

4. Social support: using competitive features to motivate behavior, a multidisciplinary team collaborated to draft the platform's functional modules (health education, exercise prescription, exercise diary, and exercise rankings) and user interface wireframes.

Phase 2: Prototype Development (Month 3-4)

Based on the framework, a prototype version (V1.0) was developed with basic functionalities—exercise plan generation based on baseline data input, manual heart rate logging (preliminary version before integrating wearable device connectivity), and static educational articles (n=10) on post-PCI care.

Phase 3: Pilot Testing and Revision (Month 5)

To validate usability and feasibility, we conducted a pilot test with 20 patients post PCI (mean age 62.3, SD 7.5 years; 12/20, 60% male) who met the study's inclusion criteria.

Pilot Testing Procedures

Usability assessment: Patients used V1.0 for 4 weeks, then completed the System Usability Scale (SUS) [23].

XSL•FO
RenderX

Functional evaluation: The research team recorded technical issues, such as loading delays and calculation errors relating to exercise volume, as well as adherence metrics, such as the daily login and module completion rates.

Key Revisions After Pilot Testing

Technical fixes: In total, 2 critical bugs were resolved (eg, an incorrect MET [Metabolic Equivalent Task]-min/week calculation), and the loading speed was optimized by 40%.

Usability improvements: Simplified the exercise logging interface (reduced input fields from 5 to 3) and added video tutorials for first-time users.

Content expansion: Increased educational materials to 20 (added 5 videos on exercise form correction based on patient feedback).

The revised version (V2.0) achieved a posttest SUS score of 78.5 (SD 6.2) versus 62.3 (SD 8.1) for V1.0 ($P<.01$), indicating acceptable usability.

Phase 4: Finalization (Month 6)

Incorporated feedback from the pilot test to finalize the platform (V3.0), which included all features personalized, simplification, self-monitoring, and so on. A 1-day training session was conducted for researchers to ensure consistent data collection during the formal study.

Establishment of the Exercise Prescription Library

To ensure the safety, feasibility, and effectiveness of exercise interventions for patients post PCI, we developed a tailored exercise prescription library based on evidence summary and multidimensional considerations, as illustrated in Figure 2.

Figure 2. Exercise prescription library for cardiac rehabilitation in patients with coronary heart disease after percutaneous coronary intervention in a tertiary hospital in Hangzhou, China.

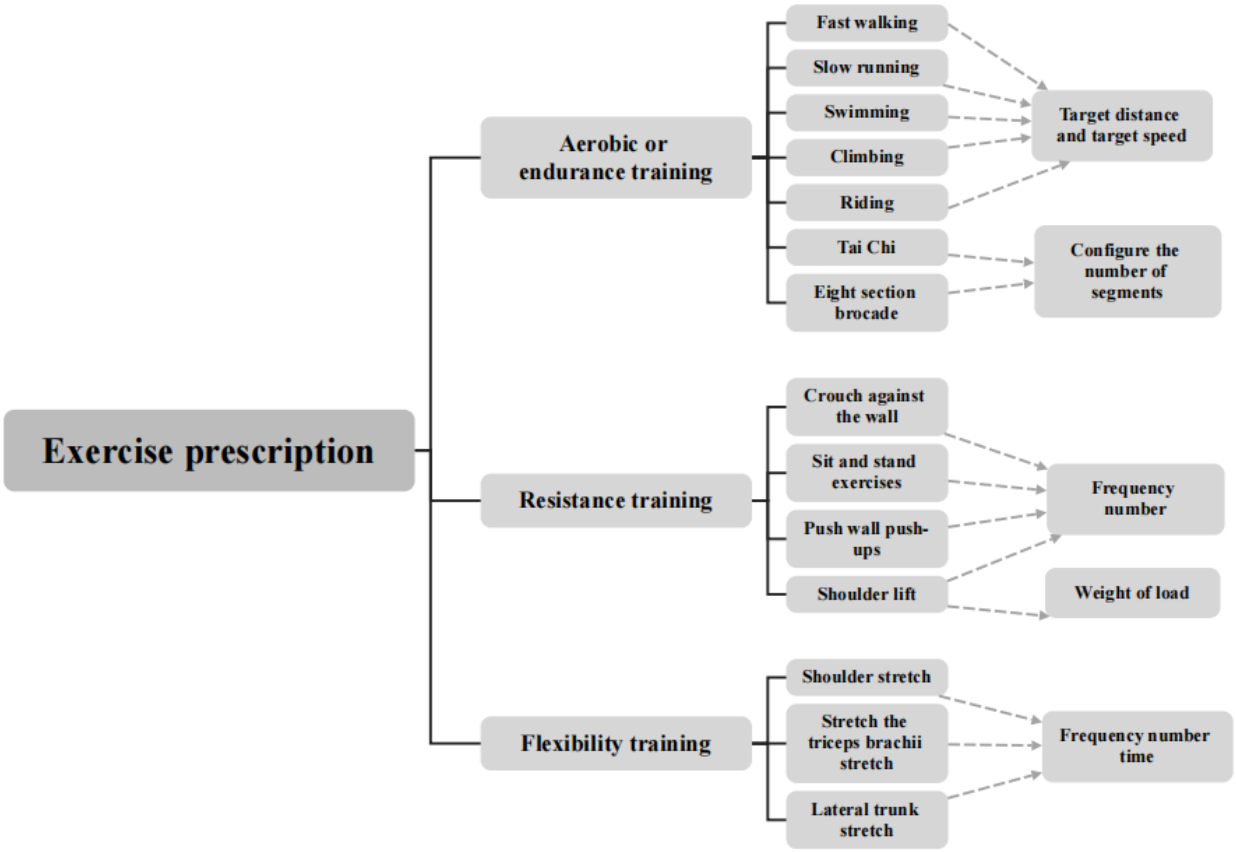

Evidence Summary for Prescription Development

First, we conducted a systematic review of guidelines and studies on post-PCI exercise rehabilitation, including (1) 2021 ESC Guidelines for CR [21], which recommend 150 minutes/week of moderate-intensity aerobic exercise for patients with CHD; (2) a network meta-analysis by Gomes-Neto et al [9], confirming that structured exercise improves exercise capacity in patients post PCI; and (3) Chinese expert consensus on CR [24], emphasizing gradual intensity progression and individualization for Chinese populations.

From these, we extracted core parameters, that is, suitable intensity (40-60% of maximal oxygen uptake), optimal frequency (3-5 sessions/week), and safe duration (20-40 minutes/session for moderate-intensity activity).

The details of the multidimensional considerations in library construction are presented in Multimedia Appendix 1.

The library was reviewed and approved by a panel of 3 CR specialists (with more than 10 years of experience) to ensure clinical appropriateness.

XSL•FO
RenderX

There were 5 basic modules in the program, namely, health education, exercise prescription, exercise diary, exercise ranking, and personal center. The user interface design diagram is illustrated in Figure 3.

Figure 3. User interface diagram of the persuasive systems design–based WeChat mini program for exercise rehabilitation in patients with coronary heart disease after percutaneous coronary intervention.

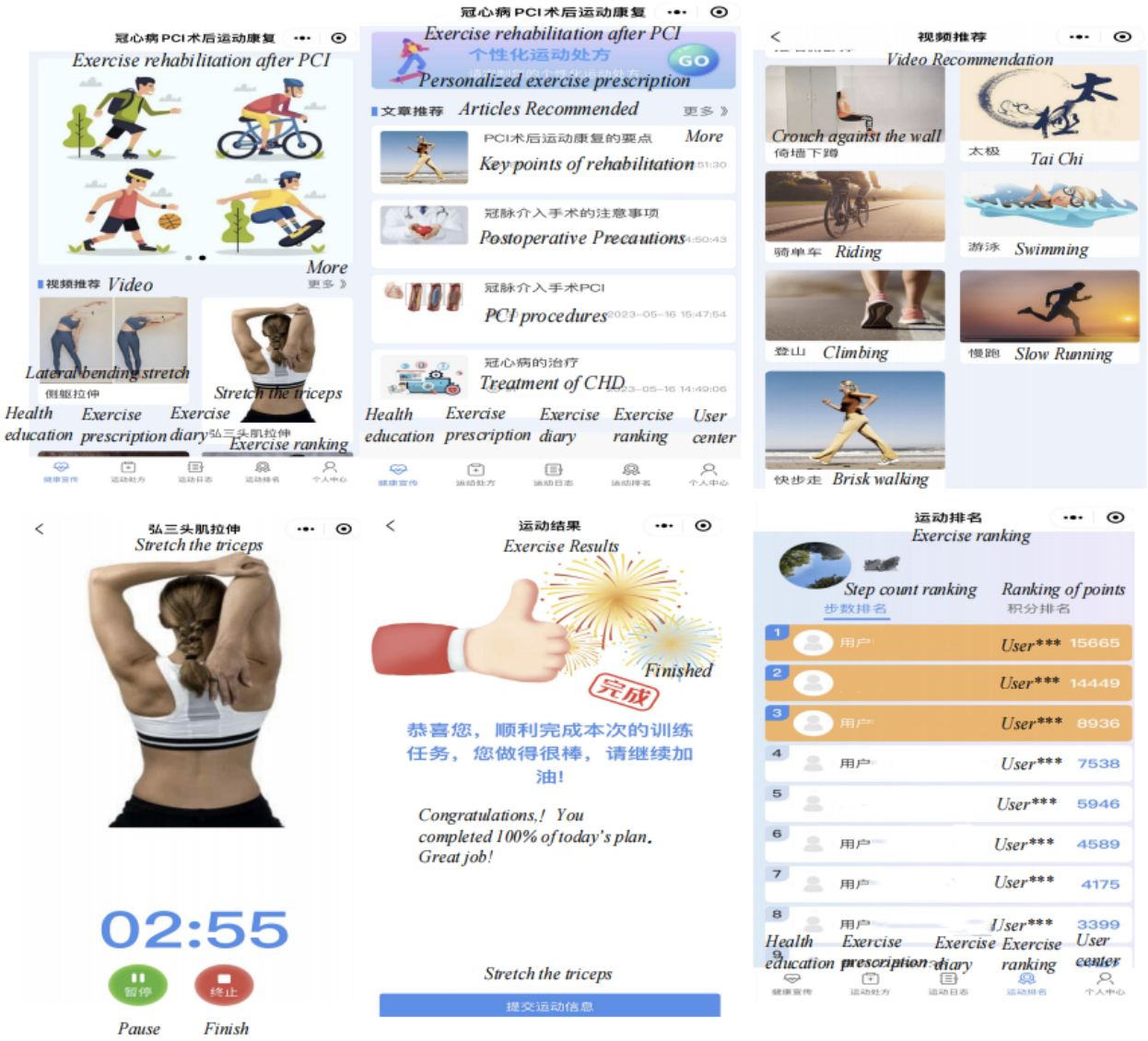

We introduced the use of a WeChat mini program to patients before discharge, ensuring that patients could successfully operate the mini program. The mini program's health education module allows patients to access disease knowledge through text and video formats, complete cardiac function grading assessments, 6-minute walking tests, Borg fatigue ratings, and monitor relevant vital signs within 1 week after discharge. Personalized exercise prescriptions are then tailored for patients based on comprehensive data evaluation. Patients can set weekly exercise reminders and follow the guidance process of the mini program to start their exercise rehabilitation. After each session, vital signs and discomfort symptoms during exercise are recorded in the exercise log module. Medical staff can timely adjust patients' exercise prescriptions according to their weekly situation and provide feedback on problems. Participants were required to continue to use the WeChat mini program for 12 weeks, and the step tracking function of the WeChat mini program was active during the 12-week intervention period. The duration of using the WeChat mini program, the frequency of weekly login to the platform, and the specific exercise data recorded in the exercise log could also be monitored through the data background of the medical and nursing side.

According to the frequency, intensity, time, and type principle, the specific structure of the selected exercise items, namely the frequency, intensity, and time of exercise, can be adjusted, which can realize the personalized adjustment of exercise tasks and task details, and provide a basis for the customization of personalized exercise prescription. Low-intensity exercise was defined as 30%-49% of maximal heart rate (or 3-5 METs), requiring aerobic or endurance training twice a week; moderate-intensity exercise as 50%-69% of maximal heart rate (or 6-8 METs), with aerobic or endurance training 3 times a week, resistance training once a week, and flexibility training once a week. High-intensity exercise, as ≥70% of maximal heart rate (or ≥9 METs), consisted of aerobic or endurance training

XSL•FO
RenderX

5 times a week, resistance training twice a week, and flexibility training twice a week.

Measurements

Overview

A self-designed demographic questionnaire was used to collect baseline sociodemographic data, including age, sex, marital status, education, employment status, living conditions, course of disease, and so on. Clinical data, including comorbidities, number of diseased vessels, cardiac function classification, PA level, scores of exercise endurance, self-perceived fatigue, ESE, and quality of life, were retrieved from the medical records. Primary and secondary outcomes were assessed with validated questionnaires. Questionnaires were collected by trained researchers through a pencil-and-paper survey. A previous intervention study [25] suggested that the exercise rehabilitation effect of patients with CHD after PCI should be intervened for at least 12 weeks, and the difference was significant. Therefore, the intervention period of this study was 12 weeks, and follow-up was conducted at 4 weeks, 8 weeks, and 12 weeks after the end of the intervention. Data collection time points were: within 3 days of admission (T0), at 12 weeks of intervention (T1), 4 weeks of follow-up (T2), 8 weeks of follow-up (T3), and 12 weeks of follow-up (T4).

Primary Outcome

The primary outcome focused on increased PA, which was assessed at baseline (T0), at 12 weeks of intervention (T1), 4 weeks of follow-up (T2), 8 weeks of follow-up (T3), and 12 weeks of follow-up (T4). PA comprises body movements that use energy. PA level was defined as weekly exercise volume, quantified by MET-min/week via the International Physical Activity Questionnaire (IPAQ; Long Form) [26]. The 27-item IPAQ-Long Form evaluates 4 domains (walking, moderate- to vigorous-intensity activity, and sitting time). Participants reported activity frequency (days/week) and duration (minutes/day), which was converted to MET-min/week using standard values (3.5 METs for walking, 4.0 for moderate, and 8.0 for vigorous activity). It has cross-cultural reliability (test-retest r=0.75) and validity across 12 countries. The Chinese version showed internal consistency (Cronbach α=0.78) and criterion validity (r=0.42 with accelerometer data, P<.01), suitable for Chinese patients post PCI [27].

Secondary Outcomes

The secondary outcomes focused on exercise endurance, self-perceived fatigue, ESE, and quality of life, which were also measured at baseline (T0), at 12 weeks of intervention (T1), and at 4 weeks of follow-up (T2), 8 weeks of follow-up (T3), and 12 weeks of follow-up (T4) weeks of follow-up.

Exercise Endurance

Exercise endurance was measured by the 6-minute walk test (6MWT), a standardized field test originally developed by Guyatt et al [28] to assess functional exercise capacity in patients with chronic respiratory or cardiac problems. Following American Thoracic Society guidelines [29], participants walked back-and-forth along a 30 m flat corridor (red turn markers); researchers instructed “fastest comfortable pace” and gave standardized encouragement every 1 minute. The total 6MWD was recorded. A portable pulse oximeter monitored peripheral oxygen saturation (SPO_2)—test paused or terminated if SPO_2<88% or severe discomfort occurred. The 6MWT correlates with maximal oxygen uptake (r=0.6-0.8) in patients with cardiac problems [29].

Self-Perceived Fatigue

Self-perceived fatigue was assessed using the Borg fatigue rating scale. According to the Borg fatigue rating scale, 6-8 points indicate very very easy, 9-10 points indicate very easy, 11-12 points indicate easy, 13-14 points indicate slight exertion, 15-16 points indicate exertion, 17-18 points indicate very exerted, and 19-20 points indicate very very exerted [30].

ESE

Measured with the Exercise Self-Efficacy Scale (ESES), developed by Resnick and Jenkins [31] for older adults and validated in CR. The 10-item scale assesses confidence in exercise tasks, such as “walking 1 km” (1=“not at all confident” to 10=“completely confident”). Total score=sum of items; higher scores=greater self-efficacy. The ESES had passed the reliability and validity test, with a Cronbach α of 0.92. The reliability coefficients of each dimension ranged from 0.72 to 0.91, indicating good reliability. The structural validity of factor analysis showed that the cumulative variance of the 4 factors was 70.19%, and the factor load of each item was 0.606-0.909, indicating good validity [31].

Quality of Life

Quality of life was assessed using the Short Form of 36 Health Survey Questionnaire (SF-36), which evaluates the influence of CHD on individuals’ physical, emotional, and social well-being. The SF-36 questionnaire has a total of 8 dimensions and 36 items. Except for physiological function and emotional function, which are answered with “yes” and “no,” the other items are scored according to 3-6 levels, and the scores of each dimension are finally converted to 0-100 points. Cronbach α coefficient of the total volume table is 0.91. The scale has been validated in Chinese populations, demonstrating good reliability and validity [32].

Usability Indicator

Usability evaluation forms an integral part of the electronic product development process. It acts as a crucial factor in ensuring the successful implementation of telemedicine programs and provides a vital means to promote user acceptance and improve compliance among users [33]. System usability was measured at 12 weeks of intervention (T1) using the SUS. The 10-item scale includes 5 positive (eg, “I would like to use this system frequently”) and 5 negative (eg, “I found the system unnecessarily complex”) statements; 5-point Likert scale (1=“strongly disagree” to 5=“strongly agree”) [34]. Scoring criteria: (1) for odd-numbered items (1, 3, 5, 7, and 9): assign a score equal to the raw item score (ie, score= raw score); (2) for even-numbered items (2, 4, 6, 8, and 10): reverse score by calculating 5 minus the raw item score (ie, score=5–raw score). Compute the total sum of all item scores, then multiply by 2.5 to obtain the final score (sum×2.5=final score; range 0-100).

XSL•FO
RenderX

Higher final scores indicate better usability of the platform. A score above 68 points is considered to be above average, a score below 68 is considered below average [23].

Safety Indicator

The incidence of AEs was used to evaluate the exercise safety of patients during the whole study period. If the patient has chest tightness, chest pain, arrhythmia, pale face, dizziness, gait instability, soft tissue injury, dislocation of bone and joint, fracture, and other conditions during exercise, the occurrence of one or more symptoms is considered as an AE, and the incidence of AEs is as follows: (Number of AEs÷total number of exercises)×100%. If there are adverse reactions, stop exercising immediately and seek help from medical personnel in time. The cause of the event will be analyzed by the research team, and all AEs will be reported to the Ethics Committee as required.

Statistical Analysis

Data analysis was performed using SPSS (version. 26.0), with 2-sided $P<.05$ considered statistically significant. An intention-to-treat approach was used. All randomly assigned participants (n=180, 90 per group) were included in the analysis, regardless of whether they completed the intervention or withdrew from follow-up. Missing data were handled via the generalized estimating equation (GEE) model, which inherently accounts for longitudinal data dependency and preserves the intention-to-treat principle. Descriptive statistics were used to summarize the participants' characteristics—continuous variables as mean (SD) or median (IQR; Shapiro-Wilk test for normality) and categorical variables as n (%). Baseline between-group comparisons used independent samples *t* test (normal continuous data), chi-square test (categorical data), or Mann-Whitney *U* test (nonnormal continuous data); intragroup nonnormal data comparisons used paired Wilcoxon rank sum test. The GEE model analyzed all primary and secondary outcomes across T0-T4, adjusting for baseline outcome values, age, sex, and education level with an exchangeable correlation structure. Given the nonnormal distribution of outcome data (confirmed via the Shapiro-Wilk test), we initially considered distributional assumptions for skewed continuous data. However, after sensitivity analysis, a Gaussian distribution with an identity link function was ultimately selected for all outcomes (PA, 6MWD, ESES, SF-36, and Borg fatigue scale). This choice was justified by two key considerations. First, the large sample size (n=180) provides robustness to violations of normality via the central limit theorem; and second, the identity link yields interpretable additive β coefficients (directly reflecting between-group mean differences), which align with clinical relevance for rehabilitation outcomes (eg, absolute changes in MET-min/week or meters). The reference category was defined as "baseline (T0)+control group" for clear effect interpretation [35]. Bonferroni correction adjusted for multiple comparisons to ensure rigor.

Ethical Considerations

This study was conducted in compliance with the Declaration of Helsinki and approved by the Hospital Institutional Review Board (approval IIT20230069B-R2). Written, fully informed consent was obtained from all participants before their enrollment, detailing the study purpose, procedures, potential risks, and benefits. Participants retained the right to withdraw from the study at any time without penalty or impact on their subsequent medical care.

All participant data were managed in strict adherence to privacy and confidentiality protocols. Identifiable personal information was deidentified during data processing and storage, and access to the dataset was restricted to authorized research personnel only.

Participants were provided with a digital health management tool as compensation for their time and participation. The compensation arrangement was fully disclosed in the informed consent document, and participants voluntarily agreed to the terms before study participation.

Results

Patient Flow and Baseline Characteristics

A total of 180 participants were randomly allocated to the intervention group (n=90) or control group (n=90). Overall, 159 (88.3%) participants completed the follow-up. The CONSORT flow diagram is presented in Figure 4. As illustrated in Table 1, no statistically significant differences were observed between the intervention group and control group in all baseline characteristics ($P>.05$), including sociodemographic variables (age: mean 63.86, SD 9.17 y vs mean 62.68, SD 8.58 y; $P=.21$; and sex (male): 73/90, 81.1% vs 67/90, 74.4%; $P=.28$), clinical indicators (cardiac function grade I: 46/90, 51.1% vs 38/90, 42.2%; $P=.23$), and baseline values of outcome measures (6MWD: median 383.50, IQR 370.00-405.00 m vs median 380.00, IQR 373.00-386.00 m; $P=.16$; and PA: median 693.00, IQR 396.00-1031.25 MET-min/week vs median 648.00, IQR 495.00-1041.75 MET-min/week; $P=.60$). Consistent with the CONSORT guidelines, this baseline balance indicates that the randomization process was effective, and any subsequent differences in outcome measures between the 2 groups can be attributed to the intervention rather than preexisting group disparities.

XSL•FO
RenderX

Figure 4. CONSORT (Consolidated Standards of Reporting Trials) flow diagram of the randomized controlled trial involving patients with coronary heart disease after percutaneous coronary intervention over the 24-week study period (12-week intervention+12-week follow-up) in Hangzhou, China. Analysis was conducted using an intention-to-treat approach, including all 180 randomly assigned participants regardless of intervention completion or follow-up status. Missing data were handled via the generalized estimating equation model.

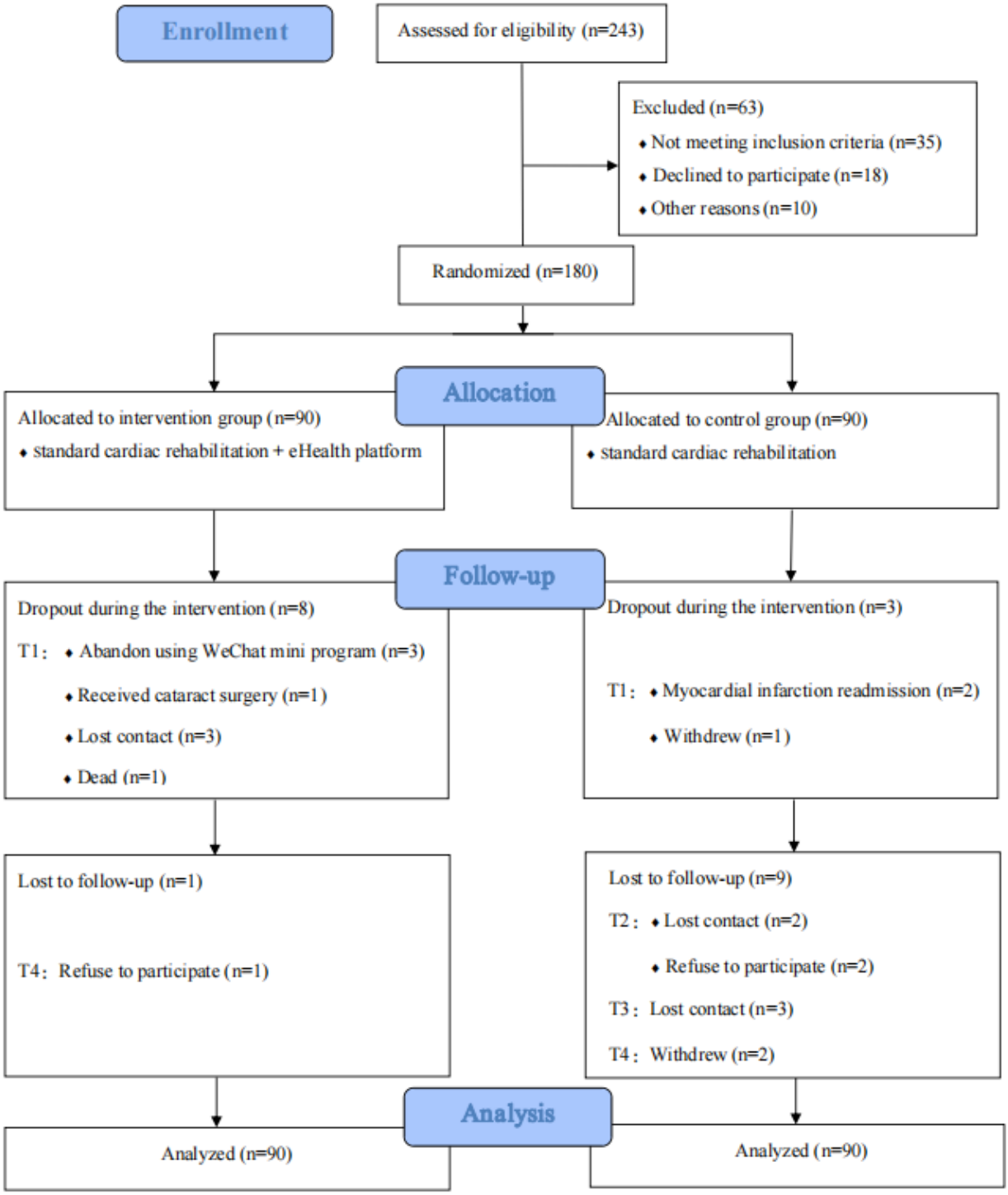

XSL•FO
RenderX

Table 1. Baseline demographic, clinical characteristics, and outcome variables of patients with coronary heart disease after percutaneous coronary intervention in the intervention and control groups (Hangzhou, China).

Variable	Control group (n=90)	Intervention group (n=90)	*t* test (*df*), Chi-square (*df*), or *z* score	*P* value
Age (y), mean (SD)	63.86 (9.17)	62.68 (8.58)	–1.266 (157)[a]	.21[b]
BMI (kg/m^2), mean (SD)	23.85 (2.37)	24.35 (2.07)	–1.425 (157)[a]	.16[b]
Sex, n (%)			1.2 (1)[c]	.28[d]
Male	73 (81.1)	67 (74.4)		
Female	17 (18.9)	23 (25.6)		
Marital status, n (%)			—[e]	.12[f]
Married	88 (97.8)	87 (96.7)		
Unmarried	2 (2.2)	0 (0)		
Widowed	0 (0)	3 (3.3)		
Residence, n (%)			2.3 (1)[c]	.13[d]
Town	31 (34.4)	41 (45.6)		
Countryside	59 (65.6)	49 (54.4)		
Education level, n (%)			0.5 (1)[c]	.98[d]
Primary school or below	40 (44.4)	38 (42.2)		
Junior high school	28 (31.1)	30 (33.3)		
High school or technical secondary school	9 (10)	11 (12.2)		
Junior college	7 (7.8)	6 (6.7)		
Undergraduate or above	6 (6.7)	5 (5.6)		
Professional types, n (%)			3.7 (4)[c]	.46[d]
Enterprises or institutions	7 (7.8)	2 (2.2)		
Farmer or worker	36 (40)	38 (42.2)		
Freelance	9 (10)	13 (14.4)		
Retired	32 (35.5)	32 (35.6)		
Unemployed	6 (6.7)	5 (5.6)		
Health insurance, n (%)			—	.44[f]
Medical insurance for employees	23 (25.6)	25 (27.8)		
Medical insurance for urban and rural residents	66 (73.3)	61 (67.8)		
Other payment methods	1 (1.1)	4 (4.4)		
Duration of illness (y), n (%)			0.8 (1)[c]	.93[d]
≤1	54 (60)	55 (61.1)		
1-3	19 (21.1)	18 (20)		
3-5	4 (4.4)	6 (6.7)		
5-10	7 (7.8)	7 (7.8)		
>10	6 (6.7)	4 (4.4)		
Monthly income (¥; ¥1=US $0.14), n (%)			1.6 (2)[c]	.46[d]
≤3000	22 (24.4)	20 (22.2)		

XSL•FO
RenderX

Variable	Control group (n=90)	Intervention group (n=90)	*t* test (*df*), Chi-square (*df*), or *z* score	*P* value
3001-4999	38 (42.2)	32 (35.6)		
≥5000	30 (33.3)	38 (42.2)		
Smoking , n (%)			2.0 (2)[c]	.37[d]
Yes	25 (27.8)	22 (24.4)		
Ever	48 (53.3)	43 (47.8)		
No	17 (18.9)	25 (27.8)		
Drinking , n (%)			2.4 (2)[c]	.30[d]
Yes	8 (8.9)	15 (16.7)		
Ever	59 (65.6)	54 (60)		
No	23 (25.6)	21 (23.3)		
Comorbidities, n (%)			6.2 (4)[c]	.18[d]
Hypertension	64 (71.1)	62 (68.9)		
Diabetes	31 (34.4)	33 (36.7)		
Dyslipidemia	5 (5.6)	17 (18.9)		
Others	67 (74.4)	71 (78.9)		
No	5 (5.6)	7 (7.8)		
Blocked vessels, n (%)			0.4 (1)[c]	.54[d]
Single	13 (14.4)	16 (17.8)		
Multiple	77 (85.6)	74 (82.2)		
Cardiac function grade, n (%)			1.4 (1)[c]	.23[d]
I	46 (51.1)	38 (42.2)		
II	44 (48.9)	52 (57.8)		
6MWD[g], median (IQR)	383.50 (370.00-405.00)	380.00 (373.00-386.00)	–1.412[h]	.16[i]
Borg fatigue rating, median (IQR)	12.00 (10.00-12.00)	12.00 (10.75-13.00)	–1.376[h]	.17[i]
ESES[j], median (IQR)	25.00 (22.22-27.78)	25.00 (22.22-31.25)	–0.173[h]	.86[i]
PA[k], median (IQR)	693.00 (396.00-1031.25)	648.00 (495.00-1041.75)	–0.522[h]	.60[i]
SF-36[l], median (IQR)	57.13 (53.44-67.09)	60.09 (50.94-66.72)	–0.963[h]	.34[i]

[a]*t* test.
[b]Independent samples *t* test.
[c]Chi-square.
[d]Pearson chi-square.
[e]Not applicable.
[f]Fisher exact test.
[g]6MWD: 6-minute walk distance.
[h]*z* score.
[i]Mann-Whitney *U* test.
[j]ESES: Exercise Self-Efficacy Scale.
[k]PA: physical activity.
[l]SF-36: Short Form of 36 Health Survey Questionnaire.

XSL•FO
RenderX

Primary Outcome

The primary outcome, PA level, was demonstrated in Figure 5 and Table 2. For the primary outcome (PA level), GEE analysis was performed with baseline (T0)+control group as the reference category to clarify group differences and time-dependent intervention effects. The model adjusted for baseline PA level and accounted for the correlation of repeated measurements within participants using an exchangeable correlation structure. Baseline PA levels were balanced between groups in unadjusted analysis (Table 1; *P*=.60), confirming effective randomization. The GEE model's "group (baseline, adjusted)" effect (β coefficient=48.254; *P*=.002) reflects the residual group difference after adjusting for age, sex, and education level—not an inherent baseline imbalance. This adjusted effect does not undermine randomization validity but enhances the model's precision by accounting for potential confounding. The control group showed a modest natural increase in PA over time (eg, T4 vs T0: β coefficient=125.358, 95% CI 101.798-148.918; *P*<.001). In contrast, the intervention group exhibited significant additional improvements at all postbaseline time points, with the strongest effect at T4 (group×time interaction: β coefficient=937.288, 95% CI 867.609-1006.967; *P*<.001). The consistent significance of group×time interactions (all *P*<.001) confirmed that the intervention's PA-enhancing effect was maintained throughout the study period.

Figure 5. Changes in physical activity levels (MET-min/week) over the 24-week study period in patients with coronary heart disease after percutaneous coronary intervention.

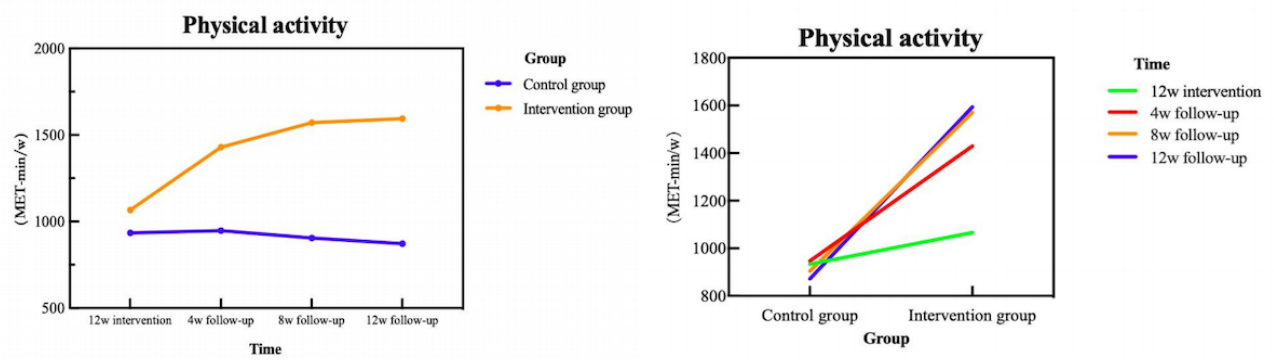

Table 2. Generalized estimating equation analysis for physical activity levels (MET-min/week) over the 24-week study period in patients with coronary heart disease after percutaneous coronary intervention.

Parameter category[a] (main effects) and parameter description	β coefficient (95% CI)	SE	*P* value
Time (control group)			
T1 (12-week intervention) vs T0	23.521 (–0.149 to 47.191)	11.683	.05
T2 (4-week follow-up) vs T0	46.857 (23.819 to 69.895)	11.748	<.001
T3 (8-week follow-up) vs T0	31.523 (8.412 to 54.634)	11.809	.008
T4 (12-week follow-up) vs T0	125.358 (101.798 to 148.918)	12.032	<.001
Group (baseline, adjusted)			
Intervention Group vs Control Group (T0, adjusted for covariates)	48.254 (18.021 to 78.487)	15.389	.002
Group×time interaction[b] (intervention additional effect)			
T1 (12-week intervention)	861.109 (783.934 to 938.284)	39.376	<.001
T2 (4-week follow-up)	899.386 (826.259 to 972.514)	37.311	<.001
T3 (8-week follow-up)	879.404 (811.213 to 947.596)	34.792	<.001
T4 (12-week follow-up)	937.288 (867.609 to 1006.967)	35.551	<.001

[a]Reference category: baseline (T0)+control group.

[b]β coefficients for group×time interactions represent the additional intervention effect (beyond the control group's natural change) in the intervention group at each time point relative to baseline.

The GEE model used a Gaussian distribution with an identity link function, yielding additive β coefficients (absolute between-group differences). "Group (baseline, adjusted)" represents the adjusted group difference in PA at T0 after accounting for covariates (age, sex, and education level). This adjusted effect does not contradict the unadjusted baseline balance (Table 1, *P*=.60), as it reflects the residual group difference after covariate adjustment (a common observation in longitudinal models) and confirms no failure of randomization.

Secondary Outcomes

Secondary outcomes are illustrated in Table 3. GEE analysis confirmed significant group-time interaction effects (all *P*<.001).

XSL•FO
RenderX

Compared with the control group, the intervention group had significantly longer 6MWD at all time points (eg, T4: β coefficient=45.278, 95% CI 41.084-49.472; *P*<.001), lower self-perceived fatigue from T2 onwards (T4: β coefficient=–1.058, 95% CI –1.243 to –0.873; *P*<.001), and higher ESE (T4: β coefficient=31.015, 95% CI 29.019-33.010; *P*<.001) and quality of life (T4: β coefficient=6.857, 95% CI 6.171-7.543; *P*<.001) across all postbaseline assessments. No significant difference in self-perceived fatigue was observed at T1 (β coefficient=0.062, 95% CI –0.116 to 0.240; *P*=.50).

Table 3. Effects of the persuasive systems design–based exercise rehabilitation platform on secondary outcomes in patients with coronary heart disease after percutaneous coronary intervention over the 24-week study period.

Outcome measures	Control group, median (IQR)	Intervention group, median (IQR)	GEE[a] β coefficient[b] (95% CI)	*P* value
6MWD[c]				
12 weeks of intervention	410.00 (389.00-420.00)	435.00 (423.00-448.00)	33.760 (29.311 to 38.208)	<.001
4 weeks of follow-up	413.00 (395.00-421.00)	440.00 (430.00-448.00)	37.577 (33.346 to 41.808	<.001
8 weeks of follow-up	406.00 (385.25-416.00)	439.50 (431.50-448.25)	44.495 (40.334 to 48.656)	<.001
12 weeks of follow-up	405.00 (385.00-413.50)	436.00 (429.00-449.00)	45.278 (41.084 to 49.472)	<.001
Borg fatigue rating				
12 weeks of intervention	10.00 (9.00-11.00)	10.00 (10.00-11.00)	0.062 (–0.116 to 0.240)	0.50
4 weeks of follow-up	10.00 (9.00-10.00)	9.50 (9.00-10.00)	–0.331 (–0.512 to –0.150)	<.001
8 weeks of follow-up	10.00 (9.00-10.00)	9.00 (9.00-9.00)	–0.707 (–0.881 to –0.533)	<.001
12 weeks of follow-up	10.00 (9.00-10.00)	9.00 (8.00-9.00)	–1.058 (–1.243 to –0.873)	<.001
ESES[d]				
12 weeks of intervention	33.33 (27.78-33.33)	58.33 (50.00-66.67)	23.355 (20.865 to 25.845)	<.001
4 weeks of follow-up	30.56 (25.00-30.56)	61.11 (52.78-66.67)	28.865 (26.534 to 31.195)	<.001
8 weeks of follow-up	30.56 (25.00-33.33)	61.11 (54.87-66.67)	28.012 (25.965 to 30.060)	<.001
12 weeks of follow-up	27.78 (25.00-30.56)	61.11 (54.17-66.67)	31.015 (29.019 to 33.010)	<.001
SF-36[e]				
12 weeks of intervention	80.19 (77.58-81.88)	89.44 (87.84-90.69)	9.520 (8.732 to 10.308)	<.001
4 weeks of follow-up	81.13 (78.71-83.00)	90.56 (89.31-91.31)	9.443 (8.733 to 10.153)	<.001
8 weeks of follow-up	82.56 (80.75-85.05)	90.88 (89.94-91.55)	8.138 (7.390 to 8.885)	<.001
12 weeks of follow-up	84.13 (82.38-86.63)	91.19 (90.41-91.81)	6.857 (6.171 to 7.543)	<.001

[a]GEE: generalized estimating equation.

[b]β coefficients represent the estimated difference in outcome values between the intervention and control groups (positive β=higher outcome in the intervention group and negative β=lower outcome in the intervention group). All generalized estimating equation models were adjusted for baseline values of the respective outcome and sociodemographic covariates (age, sex, and education level).

[c]6MWD: 6-minute walk distance.

[d]ESES: Exercise Self-efficacy Scale.

[e]SF-36: Short Form of 36 Health Survey Questionnaire.

Usability Indicator

The average SUS score of the WeChat mini program was 85.72 (SD 4.26) points (the highest score was 92 and the lowest score was 74.6, n=82), and its usability score was higher than the average level (≥68), indicating that the usability, ease of use, and learnability of the system were acceptable. The usability scores for the 10 SUS items and the average ratings for all items are summarized in Figure 6. Lighter shades of green indicate negative responses to the items (eg, “strongly disagree” for negative items and “strongly agree” for positive items), while darker shades of green indicate positive responses to the items (eg, “strongly agree” for negative items and “strongly disagree” for positive items).

XSL•FO
RenderX

Figure 6. Summary plot of the System Usability Scale scores for the persuasive systems design–based WeChat mini program in patients with coronary heart disease after percutaneous coronary intervention (12-week intervention period). SUS: System Usability Scale.

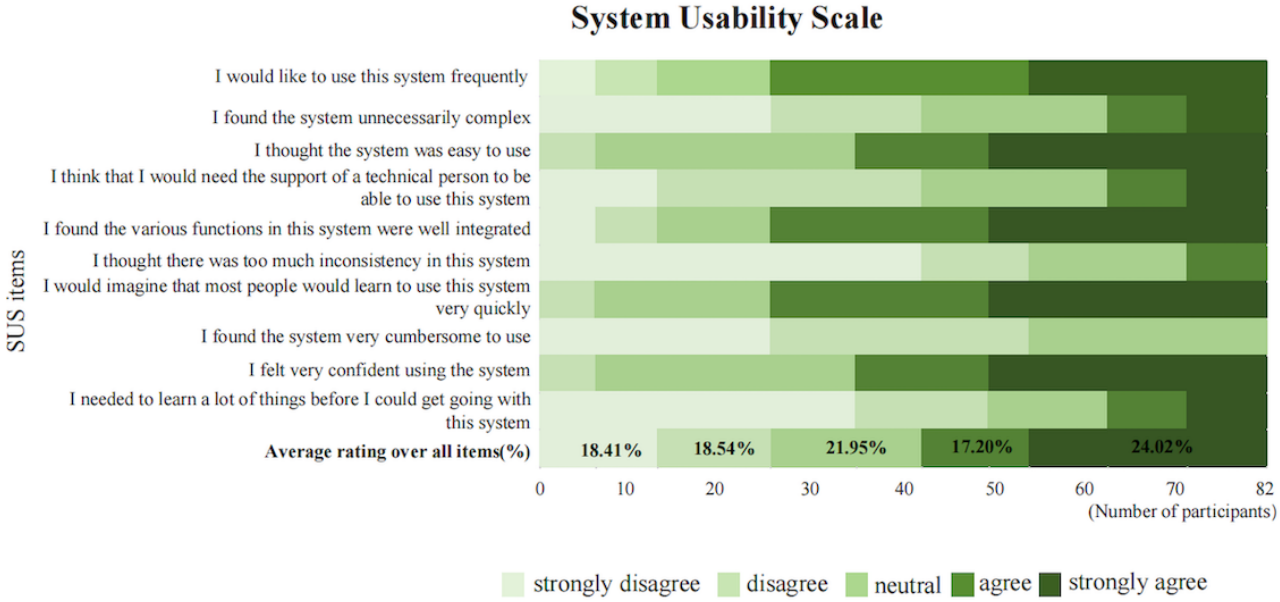

Safety Indicator

AEs during the study period were recorded comprehensively. As illustrated in Figure 4, three serious AEs were reported—2 cases of myocardial infarction readmission in the control group and 1 death in the intervention group. The 2 readmissions were unrelated to standard CR, and the death (acute cerebral hemorrhage in an 82-year-old male with hypertension and diabetes) was attributed to underlying comorbidities, not the exercise intervention. No other AEs (eg, chest tightness and musculoskeletal injury) occurred in either group. Mild muscle soreness in the early intervention stage resolved spontaneously. All serious AEs were documented, reported to the Ethics Committee, and managed per protocol.

Discussion

Principal Findings

This study confirmed that the PSD model–based mHealth platform provides added value beyond standard CR: both groups received consistent usual care, but the intervention group showed greater improvements in PA, exercise endurance, and ESE.

A key strength is the consistent delivery of standard care across both groups, which eliminated confounding from unequal usual care and clearly isolated the effect of the mHealth intervention. Our exclusive use of the PSD model addresses a key gap in CR eHealth studies. The platform's strength lies in strict alignment between PSD principles and every functional module, rather than combining multiple theories such as social cognitive theory, self-efficacy theory, theory of planned behavior, and control theory [17]. For example, our platform embeds persuasive principles within intervention designs, ensuring theoretical consistency. This focused approach avoids the "theoretical fragmentation" common in multitheory interventions. Globally, RCR interventions primarily fall into 3 categories. Basic telemonitoring systems rely on wearable devices (eg, Google Fitbit or Apple Watch) to track steps and heart rate. These systems provide limited active guidance, focusing more on data collection than behavior change [36]. App-based educational tools provide static exercise videos and reminders, but they lack personalization and include fixed exercise plans regardless of the patient's functional capacity [37]. Telehealth with human coaching combines weekly video consultations with physiotherapists and basic app tracking, but it is resource-intensive, limiting scalability [38]. In contrast, our intervention integrates PSD model–driven persuasive design with dynamic personalization—a gap in most RCR tools. Unlike basic telemonitoring or static apps, our WeChat mini program uses PSD-derived BCTs (self-monitoring via exercise diaries, social support via exercise ranking, reminders, and praise feedback) to actively motivate adherence.

The findings of PA are consistent with a previous study by Sankaran et al [39], which selected personalized and persuasive design principles to develop the HeartHab, a smartphone-based mobile app (developed by Hasselt University, Diepenbeek, Belgium) for telerehabilitation, by incorporating novel persuasive techniques to motivate patients with CHD to reach personalized PA targets. Compared with conventional rehabilitation, the effect of telerehabilitation treatment decreased at a lower rate after completion. A systematic review by Aldenaini et al [40] evaluated the effectiveness of persuasive techniques used to promote PA and reduce sedentary behavior and found that mobile and handheld devices were the most commonly used platforms, while the persuasive strategies most commonly used to achieve desired behavioral outcomes in the PA domain were self-monitoring and reminder. Salvi et al [41] found that by using the HeartCycle mobile medical system (a remote cardiovascular monitoring system developed by the research team of the University of Coimbra, Coimbra, Portugal; under a European Union–funded project coordinated by Philips, with participation of 17 partners including communication enterprises, hospitals, and universities, eg, Jessa Hospital,

XSL•FO
RenderX

Hasselt, Belgium) to guide exercise, taking into account the design principles of the persuasive system and incorporating the principles of reward, reminder, suggestion, and rehearsal into the implementation process of the system, the exercise monitoring, guidance, and feedback are provided to patients to motivate them to persist in completing the rehabilitation plan, and the exercise time of patients is significantly increased. Exercise habits also improved after 6 months of follow-up, with high user acceptance and perceived usefulness. Psychological research shows that effective implementation of the behavior change intervention helps habit formation and helps for a long time to maintain motivation [42-44]. Moreover, when it comes to technology-supported interventions, it is important to design user-friendly and accessible systems. By considering the needs and perspectives of the user to enhance the user experience, human-computer interaction research has mainly focused on usability and interaction techniques when designing technology-supported rehabilitation systems. Consider a patient's specific needs and ideas to help adjust these interventions based on technology, to accurately meet the target user, and to minimize the loss [42,45].

The 6MWT is a safe and effective tool to test exercise endurance, which has been widely used in clinical trials and the outcome measurement of cardiopulmonary rehabilitation [46]. The 6MWD was the main outcome index of the 6MWT. The study by Taylor et al [47] proposed that a within-group 6MWD improvement of at least 54 m constitutes a minimum clinically important difference (MCID) for patients with cardiac disease, reflecting meaningful functional recovery. It is important to distinguish this within-group change from the net intervention effect (between-group difference) reported in our study. For the intervention group, the within-group improvement in 6MWD from baseline (T0: 380.00 m) to T4 (436.00 m) was 56.00 m, exceeding the 54 m MCID threshold and indicating clinically meaningful functional gain for individual patients; the GEE β coefficients in Table 3 represent the net intervention effect (additional benefit of the PSD-based platform beyond the control group's natural recovery), which ranged from 33.76 m (T1) to 45.28 m (T4). While this net effect did not reach the 54 m MCID for between-group comparisons, it remains clinically relevant for several reasons. First, the control group also exhibited natural recovery (within-group improvement of 21.50 m from T0 to T4), and the intervention group's net gain still contributed to meaningful functional improvement at the individual level. Second, previous studies in CR have noted that even modest 6MWD improvements (30-45 m) are associated with better long-term prognosis [48,49]. Third, the intervention's net effect was maintained across follow-up, indicating durable functional benefits. A systematic review found that the 6MWT could be used to determine clinical response to cardiac interventions and is likely a good marker of prognosis, and demonstrated an inverse relationship between 6MWD and the likelihood of a major adverse cardiac event (MACE) [48]. Moreover, a study demonstrated that each 50 m increase in 6MWT distance was associated with a lower odds ratio and hazard ratio for the MACE. The 6MWT has a prognostic value for predicting MACE in patients with prevalent CVDs [49]. Conversely, another researcher [50] reported that 6MWT may not be sensitive enough, unable to monitor early intervention for patients with normal functions. They indicated that familiarity with walking routes or a good walking pace could lead to an artificially increased walking distance. They suggested that a familiar experiment should be carried out, the second test to establish a baseline value. In future studies, we will consider multiple repeated measurements in more populations and incorporate familiarization tests into the 6MWT to further improve the accuracy of assessment and testing and to present more precise clinical value and significance.

A study by Cavalheri et al [51] has pointed out that fatigue is one of the most frequently cited barriers for patients to adhere to exercise programs, and that exercises or tools to alleviate fatigue symptoms are very important for patients and health care workers and should be the target of exercise interventions. Exercise interventions in populations, such as lung transplantation [52] and rheumatoid arthritis [53], have also been found to have a good effect on improving the degree of fatigue in patients. However, in a systematic review of the effects of exercise rehabilitation or exercise rehabilitation on health-related quality of life and fatigue in patients with lung cancer [54], none of the studies reported statistically significant reductions in fatigue. This may be due to the large heterogeneity of the physical exercise program, the short duration of the intervention in some studies, and the poor treatment effect in most studies. This may explain why self-perceived fatigue in this study did not improve significantly during the intervention period but decreased significantly during the follow-up period.

The exercise rehabilitation platform based on the PSD model significantly improves the ESE of patients after PCI during the intervention period and the follow-up period, and has a good maintenance effect. This is consistent with the intervention effect of ESE of patients in the sedentary behavior intervention program based on the behavior change wheel theory [55]. In the intervention process, the research participants are guided to actively participate in the change of sedentary behavior, and self-monitoring is carried out, so that their real sense is affected by the positive change after the interruption of sedentary behavior, and the self-efficacy, participation intention, and compliance of sedentary patients are improved. Alhasani et al [22] found that self-monitoring to support self-management of health problems related to the application is crucial; it enhances the user's insight into their health, inspires them to seek medical help when necessary, and enhances patient confidence in the disease intervention process. In addition, the persuasive principle of praise is also selected, using incentives as a way of feedback to users. By evaluating the completion of exercise tasks, patients are sent praise messages, and certain points are rewarded, which will also strengthen the action of patients to take health behavior changes to a certain extent. Success depends on a digital self-management system related to health-promoting factors, including self-adjusting strategies, such as cognitive and emotional adjustment, goal setting, effective response, and problem-solving skills, and enhanced confidence in self-adjusting [56]. With time, changing the established behaviors and habits may be challenging, but in the future, eHealth interventions, combining health behaviors with persuasive strategies, may contribute to a sustainable behavior change in health management.

XSL•FO
RenderX

In Europe, especially Nordic countries, physiotherapists have long been core leaders in CR—driving clinical practice, program innovation, and research as key members of multidisciplinary teams, a role rooted in regional health care traditions [57]. In China, physiotherapists' involvement in CR is evolving, but is more focused on clinical execution rather than leadership or research. This is partly due to a shortage of specialized rehabilitation professionals and underdeveloped training systems, with CR programs still predominantly being physician-led [58]. This difference reflects varied health care models. Nordic systems prioritize physiotherapists as CR innovators, while China's developing CR infrastructure currently positions them as implementers—highlighting a need for workforce development to fully leverage their expertise.

Limitations

This study has limitations that should be considered. First, participants were recruited from a single tertiary hospital in Hangzhou, which may limit the generalizability of findings to patients from primary or secondary hospitals or rural areas. Second, PA was assessed by the self-reported IPAQ, and while outcome assessors were blinded to group assignments, participants knew their group allocation due to the unblinded exercise intervention. This awareness may have influenced self-reporting (eg, overestimation or underestimation of activity), a bias that cannot be fully eliminated by only blinding data collectors [59-61]. Third, the study design compared "standard CR + eHealth platform" versus "standard CR alone" and thus cannot evaluate the standalone effectiveness of the eHealth platform. Fourth, the GEE model assumes that missing data are missing at random. However, some dropouts in the control group were due to myocardial infarction (a serious AE related to the underlying disease), which constitutes informative censoring (missing not at random). This assumption may introduce bias, as the reasons for dropout are associated with the outcome of interest (cardiac function and PA). Future studies could use sensitivity analyses (eg, inverse probability weighting) or alternative statistical models (eg, joint models for longitudinal data and survival) to account for informative censoring and enhance the robustness of the results. Other limitations include the application environment, cost budget, time, and the ability of the technical team [62], which also become the obstacle factors for the application of the WeChat mini program of exercise rehabilitation to a certain extent.

Future Work

Future research directions should address these limitations. Priority could be given to direct comparative studies between center-based CR and the PSD model–based eHealth platform—a design that would clarify whether the eHealth tool can serve as a viable alternative for patients unable to attend in-person rehabilitation. Future studies could use objective tools (eg, accelerometers) to improve outcome reliability. Additionally, multicenter trials with larger sample sizes are needed to validate the effectiveness of the platform across diverse patient populations. Long-term follow-up should also be incorporated to assess the sustainability of outcomes, such as exercise adherence and quality of life.

Conclusions

The findings of this study demonstrate the effectiveness of the eHealth CR platform based on the PSD model in improving key rehabilitation outcomes for patients after PCI. Specifically, when provided in addition to standard CR, the PSD model–based eHealth platform further enhanced patients' PA level, exercise endurance, ESE, and quality of life, while also reducing self-perceived fatigue during the follow-up period. These findings provide practical insights for optimizing CR delivery in clinical settings. For health care providers, integrating PSD model–based eHealth tools into existing CR programs may serve as a feasible strategy to address suboptimal adherence and enhance long-term rehabilitation outcomes. For patients, the platform offers a flexible, accessible means to reinforce in-person rehabilitation guidance, particularly during postintervention follow-up.

Funding

This work was supported by a grant from the General Public Welfare Projects of the Science and Technology Department of Zhejiang Province, China (LGF20G030009).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author (LH) on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Multi-dimensional Considerations in Exercise Prescription Library Construction.
[DOCX File , 13 KB - jmir_v28i1e71450_app1.docx]

Multimedia Appendix 2
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1253 KB - jmir_v28i1e71450_app2.pdf]

XSL•FO
RenderX

References

1. Cardiovascular diseases (CVDs). World Health Organization. 2025. URL: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) [accessed 2025-10-08]
2. Liu MB, He XY, Yang XH, Wang ZW. Interpretation of Annual Report on Cardiovascular Health and Diseases in China 2024. Biomed Environ Sci 2025 Aug 20;38(8):893-917 [FREE Full text] [doi: 10.3967/bes2025.099] [Medline: 40928269]
3. China Health Statistics Yearbook 2021. 2021. URL: http://www.tjnjw.com/hangyefb/w/zhongguo-weishengjiankang-tongjinianjian-2022.html [accessed 2025-12-09]
4. Myers J, Prakash M, Froelicher V, Do D, Partington S, Atwood JE. Exercise capacity and mortality among men referred for exercise testing. N Engl J Med 2002;346(11):793-801. [doi: 10.1056/NEJMoa011858] [Medline: 11893790]
5. Piepoli MF, Hoes AW, Agewall S, Albus C, Brotons C, Catapano AL, ESC Scientific Document Group. 2016 European Guidelines on cardiovascular disease prevention in clinical practice: The Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts) Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). Eur Heart J 2016;37(29):2315-2381 [FREE Full text] [doi: 10.1093/eurheartj/ehw106] [Medline: 27222591]
6. Howarter AD, Bennett KK, Barber CE, Gessner SN, Clark JMR. Exercise self-efficacy and symptoms of depression after cardiac rehabilitation: predicting changes over time using a piecewise growth curve analysis. J Cardiovasc Nurs 2014;29(2):168-177. [doi: 10.1097/JCN.0b013e318282c8d6] [Medline: 23416936]
7. Fletcher GF, Ades PA, Kligfield P, Arena R, Balady GJ, Bittner VA, American Heart Association Exercise, Cardiac Rehabilitation,Prevention Committee of the Council on Clinical Cardiology, Council on Nutrition, Physical ActivityMetabolism, Council on CardiovascularStroke Nursing,Council on EpidemiologyPrevention. Exercise standards for testing and training: a scientific statement from the American Heart Association. Circulation 2013;128(8):873-934. [doi: 10.1161/CIR.0b013e31829b5b44] [Medline: 23877260]
8. Dibben GO, Faulkner J, Oldridge N, Rees K, Thompson DR, Zwisler A, et al. Exercise-based cardiac rehabilitation for coronary heart disease: a meta-analysis. Eur Heart J 2023;44(6):452-469 [FREE Full text] [doi: 10.1093/eurheartj/ehac747] [Medline: 36746187]
9. Gomes-Neto M, Durães AR, Conceição LSR, Saquetto MB, Alves IG, Smart NA, et al. Some types of exercise interventions are more effective than others in people with coronary heart disease: systematic review and network meta-analysis. J Physiother 2024;70(2):106-114 [FREE Full text] [doi: 10.1016/j.jphys.2024.02.018] [Medline: 38503676]
10. Kotseva K, Rydén L, De Backer G, De Bacquer D, Wood D. EURObservational research programme: EUROASPIRE. Eur Heart J 2015;36(16):950-951 [FREE Full text] [doi: 10.1093/eurheartj/ehv047] [Medline: 25899410]
11. Wongvibulsin S, Habeos EE, Huynh PP, Xun H, Shan R, Porosnicu Rodriguez KA, et al. Digital health interventions for cardiac rehabilitation: systematic literature review. J Med Internet Res 2021;23(2):e18773 [FREE Full text] [doi: 10.2196/18773] [Medline: 33555259]
12. Thakker R, Khan M, Al-Hemyari B. Cardiac rehabilitation after hospitalization for acute coronary syndrome. Curr Cardiol Rep 2023;25(12):1699-1703. [doi: 10.1007/s11886-023-02010-5] [Medline: 38063996]
13. Masterson Creber R, Dodson JA, Bidwell J, Breathett K, Lyles C, Harmon Still C, American Heart Association Cardiovascular Disease in Older Populations Committee of the Council on Clinical Cardiologythe Council on CardiovascularStroke Nursing; Council on Quality of CareOutcomes Research;Council on Peripheral Vascular Disease. Telehealth and health equity in older adults with heart failure: a scientific statement from the American Heart Association. Circ Cardiovasc Qual Outcomes 2023;16(11):e000123 [FREE Full text] [doi: 10.1161/HCQ.0000000000000123] [Medline: 37909212]
14. Bairey Merz CN, Alberts MJ, Balady GJ, Ballantyne CM, Berra K, Black HR, American Academy of Neurology, American Association of CardiovascularPulmonary Rehabilitation, American College of Preventive Medicine, American College of Sports Medicine, American Diabetes Association, American Society of Hypertension, Association of Black Cardiologists, Centers for Disease ControlPrevention, National Heart, Lung,Blood Institute, National Lipid Association, Preventive Cardiovascular Nurses Association. ACCF/AHA/ACP 2009 competence and training statement: a curriculum on prevention of cardiovascular disease: a report of the American College of Cardiology Foundation/American Heart Association/American College of Physicians Task Force on Competence and Training (Writing Committee to Develop a Competence and Training Statement on Prevention of Cardiovascular Disease): developed in collaboration with the American Academy of Neurology; American Association of Cardiovascular and Pulmonary Rehabilitation; American College of Preventive Medicine; American College of Sports Medicine; American Diabetes Association; American Society of Hypertension; Association of Black Cardiologists; Centers for Disease Control and Prevention; National Heart, Lung, And Blood Institute; National Lipid Association; and Preventive Cardiovascular Nurses Association. Circulation 2009;120(13):e100-e126. [doi: 10.1161/CIRCULATIONAHA.109.192640] [Medline: 19770387]
15. Duff OM, Walsh DM, Furlong BA, O'Connor NE, Moran KA, Woods CB. Behavior change techniques in physical activity eHealth interventions for people with cardiovascular disease: systematic review. J Med Internet Res 2017;19(8):e281 [FREE Full text] [doi: 10.2196/jmir.7782] [Medline: 28768610]
16. Johnston M, Carey RN, Connell Bohlen LE, Johnston DW, Rothman AJ, de Bruin M, et al. Development of an online tool for linking behavior change techniques and mechanisms of action based on triangulation of findings from literature synthesis

XSL•FO
RenderX

and expert consensus. Transl Behav Med 2021;11(5):1049-1065 [FREE Full text] [doi: 10.1093/tbm/ibaa050] [Medline: 32749460]
17. Agyei EEYF, Ekpezu A, Oinas-Kukkonen H. Persuasive systems design trends in coronary heart disease management: scoping review of randomized controlled trials. JMIR Cardio 2024;8:e49515 [FREE Full text] [doi: 10.2196/49515] [Medline: 38896840]
18. Chatterjee S, Price A. Healthy living with persuasive technologies: framework, issues, and challenges. J Am Med Inform Assoc 2009;16(2):171-178 [FREE Full text] [doi: 10.1197/jamia.M2859] [Medline: 19074300]
19. Oyibo K, Morita PP. Designing better exposure notification apps: the role of persuasive design. JMIR Public Health Surveill 2021;7(11):e28956 [FREE Full text] [doi: 10.2196/28956] [Medline: 34783673]
20. Hayes Watson C, Nuss H, Celestin M, Tseng TS, Parada N, Yu Q, et al. Health beliefs associated with poor disease self-management in smokers with asthma and/or COPD: a pilot study. J Asthma 2019;56(9):1008-1015. [doi: 10.1080/02770903.2018.1509990] [Medline: 30285498]
21. Authors/Task Force Members:, McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, ESC Scientific Document Group. 2021 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure: developed by the task force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). With the special contribution of the heart failure association (HFA) of the ESC. Eur J Heart Fail 2022;24(1):4-131 [FREE Full text] [doi: 10.1002/ejhf.2333] [Medline: 35083827]
22. Alhasani M, Mulchandani D, Oyebode O, Baghaei N, Orji R. A systematic and comparative review of behavior change strategies in stress management apps: opportunities for improvement. Front Public Health 2022;10:777567 [FREE Full text] [doi: 10.3389/fpubh.2022.777567] [Medline: 35284368]
23. Kortum P, Acemyan CZ, Oswald FL. Is it time to go positive? Assessing the positively worded System Usability Scale (SUS). Hum Factors 2021;63(6):987-998. [doi: 10.1177/0018720819881556] [Medline: 31913715]
24. He J, Jia J, Chen X, Dong A, Ding Y, Liao W. Expert consensus on full-cycle rehabilitation of elderly patients with chronic heart failure. Journal of Rehabilitation 2025;35(02):112-123 [FREE Full text] [doi: 10.3724/sp.j.1329.2025.02002]
25. Widmer RJ, Senecal C, Allison TG, Lopez-Jimenez F, Lerman LO, Lerman A. Dose-response effect of a digital health intervention during cardiac rehabilitation: subanalysis of randomized controlled trial. J Med Internet Res 2020;22(2):e13055 [FREE Full text] [doi: 10.2196/13055] [Medline: 32130116]
26. Craig CL, Marshall AL, Sjöström M, Bauman AE, Booth ML, Ainsworth BE, et al. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 2003;35(8):1381-1395. [doi: 10.1249/01.MSS.0000078924.61453.FB] [Medline: 12900694]
27. Qu N, Li J. Study on the reliability and validity of international physical activity questionnaire (Chinese Vision, IPAQ). Zhonghua Liu Xing Bing Xue Za Zhi 2004 Mar;25(3):265-268. [Medline: 15200945]
28. Guyatt GH, Sullivan MJ, Thompson PJ, Fallen EL, Pugsley SO, Taylor DW, et al. The 6-minute walk: a new measure of exercise capacity in patients with chronic heart failure. Can Med Assoc J 1985;132(8):919-923 [FREE Full text] [Medline: 3978515]
29. Enright PL, Sherrill DL. Reference equations for the six-minute walk in healthy adults. Am J Respir Crit Care Med 1998;158(5 Pt 1):1384-1387. [doi: 10.1164/ajrccm.158.5.9710086] [Medline: 9817683]
30. Morishita S, Tsubaki A, Nashimoto S, Fu JB, Onishi H. Face scale rating of perceived exertion during cardiopulmonary exercise test. BMJ Open Sport Exerc Med 2018;4(1):e000474 [FREE Full text] [doi: 10.1136/bmjsem-2018-000474] [Medline: 30622732]
31. Resnick B, Jenkins LS. Testing the reliability and validity of the self-efficacy for exercise scale. Nurs Res 2000;49(3):154-159. [doi: 10.1097/00006199-200005000-00007] [Medline: 10882320]
32. Li J, Wang Y, Shen L. Development and psychometric tests of a Chinese version of the SF-36 Health Survey Scales. Zhonghua Yu Fang Yi Xue Za Zhi 2002 Mar;36(2):109-113. [Medline: 12410965]
33. Leng M, Sun Y, Li C, Han S, Wang Z. Usability evaluation of a knowledge graph-based dementia care intelligent recommender system: mixed methods study. J Med Internet Res 2023;25:e45788 [FREE Full text] [doi: 10.2196/45788] [Medline: 37751241]
34. Hyzy M, Bond R, Mulvenna M, Bai L, Dix A, Leigh S, et al. System Usability Scale benchmarking for digital health apps: meta-analysis. JMIR Mhealth Uhealth 2022;10(8):e37290 [FREE Full text] [doi: 10.2196/37290] [Medline: 35980732]
35. de Melo MB, Daldegan-Bueno D, Menezes Oliveira MG, de Souza AL. Beyond ANOVA and MANOVA for repeated measures: advantages of generalized estimated equations and generalized linear mixed models and its use in neuroscience research. Eur J Neurosci 2022;56(12):6089-6098. [doi: 10.1111/ejn.15858] [Medline: 36342498]
36. Thorup C, Hansen J, Grønkjær M, Andreasen JJ, Nielsen G, Sørensen EE, et al. Cardiac patients' walking activity determined by a step counter in cardiac telerehabilitation: data from the intervention arm of a randomized controlled trial. J Med Internet Res 2016;18(4):e69 [FREE Full text] [doi: 10.2196/jmir.5191] [Medline: 27044310]
37. Xu L, Li F, Zhou C, Li J, Hong C, Tong Q. The effect of mobile applications for improving adherence in cardiac rehabilitation: a systematic review and meta-analysis. BMC Cardiovasc Disord 2019;19(1):166 [FREE Full text] [doi: 10.1186/s12872-019-1149-5] [Medline: 31299903]

XSL•FO
RenderX

38. Rosenstrøm S, Cecilie Tjustrup N, Kallemose T, Risom SS, Hove JD, Brødsgaard A. Evaluating a co-created model for video consultations in cardiac rehabilitation: impact on health literacy, quality of life and family support-a study protocol. BMJ Open 2025;15(10):e101099 [FREE Full text] [doi: 10.1136/bmjopen-2025-101099] [Medline: 41043840]
39. Sankaran S, Dendale P, Coninx K. Evaluating the impact of the HeartHab app on motivation, physical activity, quality of life, and risk factors of coronary artery disease patients: multidisciplinary crossover study. JMIR Mhealth Uhealth 2019;7(4):e10874 [FREE Full text] [doi: 10.2196/10874] [Medline: 30946021]
40. Aldenaini N, Alqahtani F, Orji R, Sampalli S. Trends in persuasive technologies for physical activity and sedentary behavior: a systematic review. Front Artif Intell 2020;3:7 [FREE Full text] [doi: 10.3389/frai.2020.00007] [Medline: 33733127]
41. Salvi D, Ottaviano M, Muuraiskangas S, Martínez-Romero A, Vera-Muñoz C, Triantafyllidis A, et al. An m-Health system for education and motivation in cardiac rehabilitation: the experience of HeartCycle guided exercise. J Telemed Telecare 2018;24(4):303-316 [FREE Full text] [doi: 10.1177/1357633X17697501] [Medline: 28350282]
42. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. J Med Internet Res 2017;19(6):e232 [FREE Full text] [doi: 10.2196/jmir.7126] [Medline: 28663162]
43. Adkisson RV. Nudge: improving decisions about health, wealth and happiness. The Social Science Journal 2019;45(4):700-701. [doi: 10.1016/j.soscij.2008.09.003]
44. Stawarz K, Cox A, Blandford A, Assoc CM. Beyond self-tracking and reminders: designing smartphone apps that support habit formation. In: editors. 2015 Presented at: CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems; 2015 April 18 - 23; Seoul Republic of Korea p. 2653-2662. [doi: 10.1145/2702123.2702230]
45. Schweitzer R, Schlögl S, Schweitzer M. Technology-supported behavior change-applying design thinking to mHealth application development. Eur J Investig Health Psychol Educ 2024;14(3):584-608 [FREE Full text] [doi: 10.3390/ejihpe14030039] [Medline: 38534900]
46. Du H, Wonggom P, Tongpeth J, Clark RA. Six-minute walk test for assessing physical functional capacity in chronic heart failure. Curr Heart Fail Rep 2017;14(3):158-166. [doi: 10.1007/s11897-017-0330-3] [Medline: 28421409]
47. Taylor RS, Long L, Mordi IR, Madsen MT, Davies EJ, Dalal H, et al. Exercise-based rehabilitation for heart failure: cochrane systematic review, meta-analysis, and trial sequential analysis. JACC Heart Fail 2019;7(8):691-705 [FREE Full text] [doi: 10.1016/j.jchf.2019.04.023] [Medline: 31302050]
48. Coulshed A, Coulshed D, Pathan F. Systematic review of the use of the 6-minute walk test in measuring and improving prognosis in patients with ischemic heart disease. CJC Open 2023;5(11):816-825 [FREE Full text] [doi: 10.1016/j.cjco.2023.08.003] [Medline: 38020329]
49. Sohn S, Jeon J, Lee JE, Park SH, Kang DO, Park EJ, et al. Prognostic value of the six-minute walk test in patients with cardiovascular disease. Sci Rep 2025;15(1):20817 [FREE Full text] [doi: 10.1038/s41598-025-04480-9] [Medline: 40593905]
50. Schmidt K, Vogt L, Thiel C, Jäger E, Banzer W. Validity of the six-minute walk test in cancer patients. Int J Sports Med 2013;34(7):631-636. [doi: 10.1055/s-0032-1323746] [Medline: 23444095]
51. Cavalheri V, Burtin C, Formico VR, Nonoyama ML, Jenkins S, Spruit MA, et al. Exercise training undertaken by people within 12 months of lung resection for non-small cell lung cancer. Cochrane Database Syst Rev 2019;6(6):CD009955 [FREE Full text] [doi: 10.1002/14651858.CD009955.pub3] [Medline: 31204439]
52. Diamond JM, Courtwright AM, Balar P, Oyster M, Zaleski D, Adler J, et al. Mobile health technology to improve emergent frailty after lung transplantation. Clin Transplant 2021;35(4):e14236. [doi: 10.1111/ctr.14236] [Medline: 33527520]
53. Azeez M, Clancy C, O'Dwyer T, Lahiff C, Wilson F, Cunnane G. Benefits of exercise in patients with rheumatoid arthritis: a randomized controlled trial of a patient-specific exercise programme. Clin Rheumatol 2020;39(6):1783-1792. [doi: 10.1007/s10067-020-04937-4] [Medline: 32036584]
54. Voorn MJJ, Driessen EJM, Reinders RJEF, van Kampen-van den Boogaart VEM, Bongers BC, Janssen-Heijnen MLG. Effects of exercise prehabilitation and/or rehabilitation on health-related quality of life and fatigue in patients with non-small cell lung cancer undergoing surgery: a systematic review. Eur J Surg Oncol 2023;49(10):106909 [FREE Full text] [doi: 10.1016/j.ejso.2023.04.008] [Medline: 37301638]
55. Chen D, Zhang H, Wu J, Xue E, Guo P, Tang L, et al. Effects of an individualized mhealth-based intervention on health behavior change and cardiovascular risk among people with metabolic syndrome based on the behavior change wheel: quasi-experimental study. J Med Internet Res 2023;25:e49257 [FREE Full text] [doi: 10.2196/49257] [Medline: 38019579]
56. Phelan S, Halfman T, Pinto AM, Foster GD. Behavioral and psychological strategies of long-term weight loss maintainers in a widely available weight management program. Obesity (Silver Spring) 2020;28(2):421-428 [FREE Full text] [doi: 10.1002/oby.22685] [Medline: 31970912]
57. Terbraak M, Major M, Jørstad H, Scholte Op Reimer W, van der Schaaf M. Home-based cardiac rehabilitation in older adults: expert-recommendations for physiotherapist-led care to improve daily physical functioning and reduce comorbidity-related barriers. Eur J Physiother 2024;26(5):288-298. [doi: 10.1080/21679169.2023.2276712] [Medline: 39380594]
58. Terbraak M, Verweij L, Jepma P, Buurman B, Jørstad H, Scholte Op Reimer W, et al. Feasibility of home-based cardiac rehabilitation in frail older patients: a clinical perspective. Physiother Theory Pract 2023;39(3):560-575 [FREE Full text] [doi: 10.1080/09593985.2022.2025549] [Medline: 35068322]

XSL•FO
RenderX

59. Healey EL, Allen KD, Bennell K, Bowden JL, Quicke JG, Smith R. Self-report measures of physical activity. Arthritis Care Res (Hoboken) 2020;72 Suppl 10(Suppl 10):717-730 [FREE Full text] [doi: 10.1002/acr.24211] [Medline: 33091242]
60. Adams SA, Matthews CE, Ebbeling CB, Moore CG, Cunningham JE, Fulton J, et al. The effect of social desirability and social approval on self-reports of physical activity. Am J Epidemiol 2005;161(4):389-398 [FREE Full text] [doi: 10.1093/aje/kwi054] [Medline: 15692083]
61. Prince SA, Adamo KB, Hamel ME, Hardt J, Connor Gorber S, Tremblay M. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. Int J Behav Nutr Phys Act 2008;5:56 [FREE Full text] [doi: 10.1186/1479-5868-5-56] [Medline: 18990237]
62. Sporrel K, De Boer RDD, Wang S, Nibbeling N, Simons M, Deutekom M, et al. The design and development of a personalized leisure time physical activity application based on behavior change theories, end-user perceptions, and principles from empirical data mining. Front Public Health 2020;8:528472 [FREE Full text] [doi: 10.3389/fpubh.2020.528472] [Medline: 33604321]

Abbreviations

6MWD: 6-minute walk distance
6MWT: 6-minute walk test
AE: adverse event
BCT: behavior change technique
CHD: coronary heart disease
CONSORT: Consolidated Standards of Reporting Trials
CR: cardiac rehabilitation
CVD: cardiovascular disease
ESE: exercise self-efficacy
ESES: Exercise Self-Efficacy Scale
GEE: generalized estimating equation
IPAQ: International Physical Activity Questionnaire
MACE: major adverse cardiac event
MCID: minimum clinically important difference
MET: Metabolic Equivalent Task
mHealth: mobile health
PCI: percutaneous coronary intervention
PSD: persuasive systems design
PA: physical activity
RCR: remote cardiac rehabilitation
SUS: System Usability Scale
SF-36: Short Form of 36 Health Survey Questionnaire
SPO2: peripheral oxygen saturation
WHO: World Health Organization

Edited by A Coristine; submitted 20.Jan.2025; peer-reviewed by M Hollings, T Sjögren; comments to author 11.Jul.2025; revised version received 02.Dec.2025; accepted 03.Dec.2025; published 14.Jan.2026.

Please cite as:
Liu Y, Huang X, Dai Z, Jiang Z, Wu W, Wang J, Wang Z, Yu L, Li H, Huang L
Effects of an eHealth Cardiac Exercise Rehabilitation Platform for Patients After Percutaneous Coronary Intervention Based on the Persuasive Systems Design Model: Randomized Controlled Trial
J Med Internet Res 2026;28:e71450
URL: https://www.jmir.org/2026/1/e71450
doi:10.2196/71450
PMID:

©Yang Liu, Xiting Huang, Ziying Dai, Zhili Jiang, Wenxiao Wu, Zhiqian Wang, Jing Wang, Luyao Yu, Hanyu Li, Lihua Huang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the

XSL•FO
RenderX

Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Feasibility, Acceptability, and Perspectives Regarding the Use of Activity Tracking Wearable Devices Among Home Health Aides: Mixed Methods Study

Ian René Solano-Kamaiko[1], MS; Michael Dicpinigaitis[2], BS; Melissa Tan[1], MS; Irene Yang[1], MS; Kexin Cheng[1], MS; Ronica Peramsetty[2], BS; Michelle Shum[2], BS; Yanira Escamilla[3], LCSW; Jennifer Bayly[2], MPH, MD; Meghan Reading Turchioe[4], RN, MPH, PhD; Ariel Avgar[5], PhD; Aditya Vashistha[5], PhD; Nicola Dell[1,5], PhD; Madeline R Sterling[2], MPH, MD, MS, FAHA

[1]Cornell Tech, New York, NY, United States
[2]Weill Cornell Medicine, New York, NY, United States
[3]1199SEIU Funds, New York, NY, United States
[4]Columbia University, New York, NY, United States
[5]Cornell University, Ithaca, NY, United States

Corresponding Author:
Madeline R Sterling, MPH, MD, MS, FAHA
Weill Cornell Medicine
420 East 70th Street, LH-357
New York, NY
United States
Phone: 1 646 962 5029
Email: mrs9012@med.cornell.edu

Abstract

Background: Home health aides and attendants (HHAs) provide in-home care to the growing population of older adults who want to age in place. Despite their vital role in patient care, HHAs are an underserved and vulnerable population of health care professionals who often experience poor health themselves. Activity tracking devices offer a promising way to improve HHAs' health-related awareness and promote health behavior change, particularly regarding physical activity and sleep quality, 2 areas in which the workforce struggles.

Objective: This study aimed to understand how feasible it is for HHAs to use activity tracking devices and assess their perceptions of such devices for improving their health. Specifically, we conducted (1) a field study to assess the use, feasibility, and acceptability of these devices among HHAs and (2) a qualitative study to understand HHAs' perspectives on and reactions to activity trackers on and off the job.

Methods: We partnered with the 1199 Service Employees International Union Training and Employment Fund to conduct a field study with home care agency–employed HHAs working in New York City, New York. Participants wore activity tracking devices for 4 weeks that collected data on physical activity and sleep. The HHAs were subsequently interviewed on their experiences with and attitudes toward the devices and asked to reflect on personalized visualizations of their data to prompt them to think aloud. Quantitative data were analyzed using descriptive statistics. Qualitative data were analyzed using grounded theory.

Results: A total of 17 HHAs participated; their mean age was 48.7 (SD 12.2) years, 15 (88%) were women, 11 (65%) identified as Black, 5 (29%) identified as Hispanic or Latinx, and they had worked as HHAs for a mean of 11.7 (SD 7.5) years. In total, 94% (n=16) of the HHAs wore their activity trackers for the full 28-day study period. Participants took a mean of 10,230 (SD 3586) daily steps during the study period and slept for a mean of 6.27 (SD 0.58) hours per night. Overall, 4 key themes emerged: (1) activity tracking devices enhanced participants' health awareness by providing empirical data for self-reflection; (2) this increased awareness led to positive behavior changes, including setting and achieving health-related goals; (3) HHAs believed that these devices could improve not only their own health but also that of their patients through positive behavior changes; and (4) despite this optimism, participants emphasized that their ability to modify sleep and activity patterns was constrained by social and occupational determinants, with sleep improvements being particularly challenging.

XSL•FO
RenderX

Conclusions: Our findings suggest that appropriately designed personal tracking interventions could offer a promising approach to supporting positive health-related changes in this historically overlooked workforce, potentially improving their well-being and the quality of care they provide to their patients.

(J Med Internet Res 2026;28:e77510) doi:10.2196/77510

KEYWORDS

home health aides and attendants; home health care; home care; worker well-being; frontline work; low-wage work; data governance; data advocacy; wearable activity trackers; passive sensing

Introduction

Home health aides and attendants (HHAs) are an essential frontline workforce providing in-home care to a growing number of older adults who wish to age in place [1-3]. HHAs provide personal care to older adults, including assistance with activities of daily living and instrumental activities of daily living as well as medically oriented care (monitoring symptoms and vital signs and counseling patients on physical activity [PA], diet, and medication adherence), all while providing emotional support [4].

Despite their critical role in patient care, HHAs are an underserved and vulnerable group of health care professionals who often have poor health themselves [5-7]. Predominantly comprising middle-aged women from racial minority groups earning substandard wages (national median of US $12 per hour), this workforce numbers over 3.5 million in the United States and has a high burden of cardiovascular disease risk factors, including obesity, hypertension, hyperlipidemia, physical inactivity, smoking, and poor sleep [4,8,9]. Our recent national study revealed that HHAs' health and health behaviors are notably worse than those of other frontline health care workers not employed in the home (ie, nursing homes and hospitals), with over one-third reporting fair or poor general health, nearly 40% being obese, and more than half being physically inactive [10]. Contributing factors to this include limited health insurance and opportunities to seek medical care, as well as little time to engage in positive health-promoting behaviors themselves. Left unaddressed, these occupational health disparities among HHAs could potentially harm them and compromise the quality of care they provide.

Activity tracking devices, also known as wearables, may be uniquely positioned to promote positive health behavior change in a workforce that has rarely been the focus of interventions and technology innovation. These devices, which include Fitbit trackers, Oura Rings, and others, show promise in potentially improving PA and sleep, 2 behaviors HHAs struggle to achieve optimal levels of, based on our prior studies [4,10]. Consistent with social cognitive theory and the theory of planned behavior, wearable devices allow users to visualize and reflect on their daily activity and sleep data, enhancing self-efficacy and promoting positive behavior change [11,12]. By improving their own health behaviors, HHAs may also model and encourage similar changes among their patients. To date, research on activity trackers in this context has primarily focused on their use among older adults [13,14], family caregivers [15,16], and facility-based care workers in international settings [17-19] and has never focused on HHAs, a growing workforce in constant contact with patients who also struggle to reach sufficient levels of PA and sleep. Furthermore, while these adjacent groups share caregiving roles, HHAs differ in that they are paid, agency-employed workers who face unique structural and occupational barriers to prioritizing their own health [20,21]. Therefore, studying HHAs extends existing work on caregiver- and worker-focused interventions by addressing a critical and understudied population within the US health care system.

To understand the potential for activity trackers to promote positive health behavior change, we conducted (1) a field study to assess the use, feasibility, and acceptability of these devices for daily use and (2) a qualitative study to understand HHAs' perspectives on and reactions to activity trackers on and off the job.

Methods

Overview and Study Setting

We conducted this field and qualitative study from July 19, 2024, to October 11, 2024, in collaboration with 1199 Service Employees International Union Training and Employment Fund (1199SEIU-TEF), a labor management fund of the 1199 Service Employees International Union (SEIU) United Healthcare Workers East, the largest health care union in the United States [22]. The 1199SEIU-TEF provides education and training to more than 55,000 HHAs who are employed by over 50 licensed and certified home care agencies in New York City, New York [23].

Study Participants and Recruitment

To be eligible, HHAs had to (1) be aged ≥18 years, (2) be English speaking, (3) own a smartphone (iOS or Android), and (4) be employed as an HHA in New York City. To recruit our intended sample of 25 to 30 study participants, 1199SEIU-TEF staff reached out to HHAs who had previously expressed interest in participating in research initiatives and invited them to take part in our study using standardized recruitment materials (email script and flyer). If potential participants were interested, staff referred them to the study team to be screened for eligibility. HHAs who were eligible for the study provided written consent. All aspects of this study took place in person at 1199SEIU-TEF headquarters in New York City.

Data Collection and Study Procedures

Before conducting the field study, the research team met with participants to elicit their initial perspectives on activity tracking devices and the device form factor (wristband vs ring). From these conversations, we found that HHAs were enthusiastic about the potential benefits of activity trackers for increasing

XSL•FO
RenderX

their health awareness. HHAs preferred wristband devices over rings as they felt that the wristband's small screen would be easier to check discreetly during the workday than their phone and its form factor would not interfere with their daily tasks.

The field study involved giving HHAs Fitbit Charge 6 devices, which they were asked to wear for 28 days. Before providing the devices, we conducted a 1-hour onboarding session where participants were introduced to the research team; asked to respond to a few self-reported demographic questions (age, gender, race, Hispanic or Latinx ethnicity, educational level, and years of experience as an HHA); and guided through setting up the Fitbit device and mobile app, highlighting key features including activity and sleep tracking. After the participants created their Fitbit accounts, we walked them through the authorization process to allow the research team remote access to their data through the Fitbit application programming interface.

Participants were then asked to wear the Fitbit daily for 4 weeks, during which the research team contacted them weekly via their preferred method (SMS text message). Before the field study, we confirmed participants' contact preferences (eg, SMS text message, email, or phone), and all preferred SMS text messages. Prior work has also supported this modality for communicating with HHAs [24].

SMS text message reminders prompted participants to wear the wristband, review their data, and contact the team with questions or issues. However, these reminders did not substantially affect engagement or acceptance. Most HHAs either ignored the messages or briefly acknowledged their receipt. In interviews, participants reported wearing the devices consistently because they valued the insights into their health and well-being, with SMS text messages serving primarily as a channel for troubleshooting or procedural clarifications (eg, how to access participant gift cards). Following the 28-day field study, we used the Fitbit application programming interface to collect participants' data to create personalized visualizations showing each participant's activities in relation to (1) the National Institutes of Health's (NIH) recommended PA and sleep guidelines for adults and (2) their broader group of peers in the study.

Four investigators from the research team (IRS-K, MT, IY, and KC) then conducted 1-hour semistructured interviews with each participant to understand their experiences, data preferences, and insights into the wearable devices (Multimedia Appendix 1). Along with participants' general perspectives, the investigators focused on device acceptability—as illustrated by participants' attitudes toward the devices' usability and perceived value. Following these discussions, we introduced the personalized visualizations of their data and prompted participants to engage in a think-aloud process [25]. Participants reflected from multiple perspectives: as individuals balancing complex personal and work lives, as members of a distributed peer group, and as part of a broader community advocating for improved working conditions. Interviews ended when thematic saturation was reached.

Quantitative Data Analysis

Participant characteristics were analyzed using descriptive statistics. We collected participants' Fitbit data, comprising 187 JSON files containing metrics on activity levels (sedentary, light, fairly active, and very active), step count, distance, heart rate, and sleep duration and stages (wake, light, rapid eye movement, and deep), each tagged with precise date and time stamps. We focused on sleep duration and step count and assessed use feasibility, defined as participants' ability to consistently engage with the devices, by examining the date and time stamps associated with these metrics to determine whether the devices captured participants' data on a daily basis.

We selected sleep duration and step count because these metrics were intuitive for participants (ie, required little explanation, were immediately interpretable, and were clearly actionable) and have clear policy relevance, such as the No More 24 campaign [26] advocating to end 24-hour HHA shifts. Furthermore, prior research indicates that Fitbit-derived sleep duration and step counts are reasonably reliable and accurate; systematic reviews and validation studies have found that Fitbit devices provide sleep duration estimates comparable to those of research-grade accelerometers [27,28], perform well in distinguishing wake from sleep [28,29], and meet acceptable validity and reliability standards for tracking PA in free-living conditions [30].

Participants' Fitbit data (steps and sleep duration) were processed using exploratory data analysis methods in Jupyter Notebook [31]. As mentioned above, during the exploratory data analysis process, we generated data visualizations for the reflection interviews, which also incorporated auxiliary data such as participants' sleep and work schedules along with the NIH's PA and sleep recommendations to create benchmarks for optimal levels.

Qualitative Data Analysis

Each participant interview lasted approximately 60 minutes, resulting in 17 hours of audio-recorded interview data. The interviews were professionally transcribed using NoScribe (noScribe.ai) [32], an open-source artificial intelligence tool run on the research team's computers. We verified each transcript against the original recordings, correcting errors and redacting identifying information. Our analysis used grounded theory and, as such, used an inductive open coding approach [25] with 4 investigators (IRS-K, MT, IY, and KC). We established a baseline code set by jointly coding one interview and then validated our approach by coding a second interview in 2 separate pairs (IRS-K and KC as well as MT and IY), meeting to reconcile differences. The remaining 15 interviews were coded independently, with regular meetings to resolve disagreements and refine the coding scheme. Finally, we conducted affinity diagramming to synthesize these codes into high-level themes with representative quotations.

Ethical Considerations

All participants gave informed consent to take part in the study, including consent for note taking, photo taking, audio recording of interviews, and collection of participants' Fitbit data. Identifiable details in the data and features in participant images

XSL•FO
RenderX

were removed to ensure participant privacy and confidentiality. Participants were compensated for their time with US $25 gift cards for each of the 3 components of data collection (preliminary discussions, field study, and final interview), for a total possible compensation of US $75 in addition to the Fitbit device, which they were allowed to keep. Cornell University's institutional review board classified the study as expedited and reviewed and approved the project under protocol IRB0148598.

Results

Participant Characteristics

A total of 17 HHAs participated in this study. They had a mean age of 48.7 (SD 12.2) years, 88% (n=15) were women, 65% (n=11) identified as Black, 29% (n=5) identified as Hispanic or Latinx, and they had worked as HHAs for a mean of 11.6 (SD 7.5) years (Table 1).

Table 1. Sample characteristics (N=17).

Characteristic	Values
Age (years), mean (SD)	48.7 (12.2)
Gender, n (%)	
Women	15 (88)
Men	2 (12)
Race, n (%)	
African American or Black	11 (65)
White	1 (6)
Ethnicity, n (%)	
Hispanic or Latinx	5 (29)
Educational level, n (%)	
High school diploma	3 (18)
Some college	3 (18)
Associate's degree	4 (24)
Bachelor's degree	5 (29)
Graduate degree	2 (12)
Experience (years), mean (SD)	11.6 (7.5)

Field Study Findings

Overall, 94% (16/17) of the HHAs wore their activity trackers for the full 28-day study period, with only 6% (1/17) wearing the device for 26 days, and 82% (14/17) indicated continued use after the study's conclusion. As shown in Table 2, participants took a mean of 10,230 (SD 3586) daily steps during the study period, ranging from 3855 to 20,528 steps. In total, 53% (9/17) of the participants surpassed the NIH recommendations for 10,000 daily steps for healthy adults [33]. Table 2 shows that participants slept for a mean of 6.27 (SD 0.58) hours per night, ranging from 5.40 to 7.38 hours, with 18% (3/17) exceeding the NIH recommendations for 7 hours or more of daily sleep duration for healthy adults [34].

XSL•FO
RenderX

Table 2. Wearable device-measured mean steps and sleep duration among study participants (N=17). Comparison of mean daily step count and mean daily sleep (in hours) by participant. The National Institutes of Health recommends that healthy adults should walk 10,000 steps [33] and sleep 7 hours or more per day [34].

Participants	Daily step count, mean (SD)	Daily sleep (hours), mean (SD)
P1	3855 (2345)	6.59 (2.36)
P2	12,683 (4577)	6.58 (2.05)
P3	11,680 (2552)	5.82 (1.87)
P4	6867 (3705)	7.38 (1.58)
P5	7935 (5063)	5.75 (3.04)
P6	7133 (2782)	5.87 (1.95)
P7	20,528 (6384)	7.30 (0.99)
P8	10,262 (3442)	5.79 (1.56)
P9	10,739 (3386)	5.74 (1.66)
P10	11,928 (3351)	6.46 (1.30)
P11	11,133 (3217)	6.13 (1.95)
P12	9212 (3556)	6.36 (0.80)
P13	12,980 (7196)	5.40 (1.84)
P14	7026 (2396)	7.15 (1.85)
P15	7863 (1894)	5.64 (1.23)
P16	9063 (3976)	6.55 (1.17)
P17	13,024 (3123)	6.09 (1.76)

Major Themes From Qualitative Interviews

A total of 4 major themes arose. Each theme is presented below with accompanying quotations that represent key concepts.

Theme 1: Activity Tracking Devices Were Feasible to Wear and Improved HHAs' Health Awareness

Most HHAs found the use of the devices to be seamless and feasible to incorporate into their day-to-day activities. Many cited the personal value they derived from using the devices. Participant P9 shared their experience:

> *It's becoming part of me, since one month I've been on it, so it's like I'm getting accustomed to it.... I love wearing it because the information it provided...was very helpful to my life.*

While many participants reported having a general awareness of their own health and well-being, a number stated how the activity tracker helped enhance their understanding of their own health. For instance, P4 said:

> *The difference is that with the Fitbit, I can see exactly how long I sleep or walk.... I know that sometimes I wasn't sleeping, but I can [only] say maybe I slept five or six hours when I count. But with the Fitbit, it's going to let you know exactly how long was the deeper sleep, the lighter sleep, and that stuff.*

Participants also reflected on the fact that, after being shown their data, this informed their perceptions of their own health behaviors. For example, P5 reflected:

> *A lot of us are going to be surprised because in our mind, we're sleeping a lot…. If this didn't come about, I might have said, "oh, yes, I get enough sleep." But this now is in fact showing me, "oh, girl, you're not sleeping."*

This revelation was common among participants, who found that the tracking data offered a concrete way to reflect on their habits and gain a deeper understanding of how their work and home life impacted aspects of their well-being, such as sleep. P16 reflected:

> *When I look, sometimes I see I have four hours, 40 minutes of sleep.... I'm tired and everything is bothering me. But prior to the Fitbit, I didn't realize what was bothering me.*

Theme 2: Activity Tracking Devices Helped HHAs Set and Achieve Health Goals

Many participants shared how increased awareness of their health led to positive changes in behavior. Several reported adopting healthier habits because of heightened awareness of their sleep patterns. Examples they shared included limiting television and phone use (ie, screen time), adjusting their bedtime and wake-up times, and incorporating relaxation techniques such as warm showers and listening to calming music. For example, P6 explained:

> *When I watched my sleep time (on my device), I was sleeping between three and five [hours].... [Now I] try and see if I could fall to sleep earlier.... I started getting eight hours, seven hours. I started trying to focus on going to sleep now, [no] TV, [no] phone.*

XSL•FO
RenderX

Reflecting on their activity data also had a motivating effect for many participants, helping them re-evaluate their existing behaviors and adopt new habits. Several mentioned that these changes had a compounding effect, where improvements in one area, such as increased walking, made them feel better and motivated them to walk even more. P5 shared their experience:

> *If I'm out there...for five, six minutes and I don't see the bus, I'm going down the road. [Before] I wouldn't do that. I'd stay there and wait for the bus. But it [the Fitbit] really encourages me now to walk. And as I said, the more I walk, the lighter I feel. The more encouraged I feel.*

Other participants highlighted how greater health awareness helped them set and achieve specific health goals. For example, P3 discussed how she aimed to improve her PA at work to meet the step count goal (which was preprogrammed into the device). She said:

> *Sometimes I go to work, when I see I didn't have the number of steps required, when I put my patient to bed, I went out to finish the steps. To have the score. So that was very interesting because I never did it before.*

Theme 3: Activity Tracking Devices May Have Potential to Improve the Health of HHAs and Their Patients

HHAs were acutely aware of the negative impacts of poor sleep on their health and well-being. Beyond personal health improvements, they expressed concerns about their ability to perform their high-stakes job, which demands constant vigilance, when sleep deprived. For instance, P5 highlighted this challenge, stating:

> *If you don't get enough rest, you cannot function effectively.... So, you as a [home care agency] admin, you want your workers to perform effectively. What will you do now for us so we can get adequate sleep?*

At the same time, several HHAs were optimistic about the potential of activity tracking devices to enhance both their personal health and the health of their patients while on the job. For example, P8 said:

> *If you get a case that will let you [and] the patient walk. So the two of you can walk around and do more exercise.*

Theme 4: HHAs' Abilities to Make Sleep and Activity Changes Were Impacted by Social and Occupational Determinants

HHAs reported that certain factors related to their social environments or working conditions impacted their health and made it challenging to make desired health changes. For example, participants working live-in cases or overnight shifts mentioned that their sleep schedules were dictated by patients' demands. P17 explained:

> *In live-in case we, by the law, [we] are supposed to sleep five hours not interrupted, but it's not possible.... You can't sleep this amount of hours.... They [patients] put TV on high volume and you can't sleep.*

This demonstrates the limited control that HHAs have over their sleep schedules. Additionally, HHAs expressed the need to remain vigilant, attending to patients' needs and staying alert for emergencies. This constant hypervigilance negatively affected their ability to relax and sleep during overnight shifts.

HHAs also noted the challenges of nontraditional work schedules. Shift work, subject to frequent changes based on patient needs, made it difficult for many to regulate their sleep, even on days off. For instance, P4 described the stability of working with the same patient vs their current unpredictable schedule:

> *Like last year or years before, I have a stable patient. I know that I was working four days a week with the same patient.... I know exactly what was my schedule at that time. But right now, I have no fixed schedule.*

Others, such as P15, faced difficulties regardless of their work schedule (daytime or nighttime), expressing difficulties balancing responsibilities in addition to work, such as school, family, and more, which impacted their ability to stay healthy:

> *Once I'm at home, I want to like catch up [on] my [school] assignments or cause currently I'm doing my internship as well. I still have to keep up with my personal life, my kids, my husband.... I have three kids and they are just like babies.... So when I'm at home, I think I tend to do more cause, if I can sleep, they have to like sleep before I can get to do anything.*

Beyond sleep, HHAs also felt that their PA was often dictated by patients' needs. P14 explained:

> *My patients, some of them, they want company.... They say, "let's watch a movie, or let's go to the theater," and they call an Uber. So you don't walk that much. You walk during the time that you're going to work, and then you go home.*

While some HHAs found that their PA was limited by their cases, others felt that their activity levels were more malleable. Many developed strategies to increase exercise, such as walking instead of taking the bus, neighborhood walks on days off, and encouraging patients to go outside for exercise with them during downtime.

Discussion

Principal Findings

To our knowledge, this is the first study to examine how activity tracking devices influence HHAs' awareness of their health and well-being using empirical data from the devices and qualitative reflection interviews with participants. Our findings demonstrate that the use of activity trackers was both feasible and acceptable among HHAs. Feasibility was reflected quantitatively by participants' ability to consistently engage with the devices, with 94% (16/17) wearing them for the full 4-week field study period. Acceptability was shown qualitatively through participants' reflections during follow-up interviews, where they expressed positive attitudes toward the devices' usability and perceived value for increasing health awareness. This level of engagement is noteworthy as previous research has shown

XSL•FO
RenderX

lower participant adherence to consumer wearables [35]. The high engagement observed among HHAs in this study suggests that activity tracking interventions may be particularly effective for supporting the health and well-being of HHAs more broadly, although further research is necessary to validate this assumption.

Additionally, the Fitbit Charge 6's wristband form factor and small screen provided participants with a discreet, functional user experience. Beyond use and feasibility, HHAs not only found that the devices offered them more awareness of key aspects of health—PA and sleep—but even in a short time, they reflected that they were empowered to make positive behavior changes because of wearing the devices. Finally, we found that the devices may have health benefits not only for HHAs but also for the patients for whom they care.

Our study builds on scientific literature in a few key ways. First, while many studies have focused on the acceptance and impact of wearables among older adults [13,14], few have addressed their use among family caregivers and facility-based health care workers [15-19], and even fewer have examined these applications with HHAs [36].

For example, several studies conducted in Finland and Norway have examined the use of wearables among family caregivers and facility-based home care nurses, nursing assistants, and occupational therapists [17-19]. However, these studies often relied on technologies such as accelerometers [17,18] and electrode-based monitors [19], which can be difficult to set up and may not offer real-time or easily interpretable health feedback for lay users, unlike consumer-friendly devices such as Fitbit trackers.

To date, there has been one prior case study in which Fitbits were deployed among caregivers at the SEIU 775 Benefits Group in Washington state [36]. The study found that using Fitbit devices, combined with health coaching, resulted in improvements in health behaviors over a 4-month period. These improvements included increased PA (measured via step counts and active minutes), weight loss, and reduced blood pressure. The findings were based on multiple blood pressure readings, self-reported weight loss, and recorded activity data from the devices. However, this was a pilot program conducted by SEIU 775 Benefits Group in collaboration with Fitbit, Inc (a Google subsidiary), resulting in a self-authored "case study" rather than a peer-reviewed journal article. While the case study presents promising findings, its details and depth are limited due to its intended audience. Our study builds on this preliminary work by outlining a detailed and robust study methodology focused on Fitbit devices for improving HHAs' awareness of their health and well-being. Additionally, our study incorporates both wearable data and qualitative reflections, aimed at providing a deeper understanding of how HHAs navigate barriers to improving their health and how tracking devices may help support these efforts.

Prior research has shown that wearable devices support health monitoring and facilitate behavior change, particularly in PA and sleep [13,16], and our study extends the literature to HHAs. In doing so, it highlights how these devices have the potential to improve the health of a vulnerable population and shape workplace dynamics and patient care within home care settings. Importantly, we found that wearables can not only track behavior but also offer a simple and feasible way to readily influence workplace routines, patient interactions, and broader occupational health strategies in home care settings. This finding is notable as previous research has shown that influencing behavior change is often challenging and may reflect HHAs' perception of being in control of their own health [37].

In recent years, there has been a growing call for better understanding and improving the health of HHAs [4,7,9,10]. However, empirical studies that aim to improve their health behaviors and health are scarce. Our study is a first step toward this through harnessing technology and offering HHAs real-time access to their own health data. While we found striking benefits, it is also important to note the challenges. For example, while these devices can offer HHAs personalized feedback and enhance user motivation and self-efficacy [13,16], we also found that external and structural barriers in the environments in which the HHAs live and work can limit the extent to which HHAs can act on their health feedback, making it difficult to prioritize meaningful and sustainable behavioral health changes. For example, unpredictable work schedules, multiple jobs, and familial obligations all left them with limited time and resources to obtain sufficient sleep.

Nevertheless, HHAs found the intervention valuable, using wearable devices to self-monitor, reflect on current health habits, and make changes to their behaviors. Some participants adjusted their sleep routines, such as modifying bedtimes and reducing screen time before sleep, whereas others increased their PA in response to real-time device feedback. Consistent with social cognitive theory and the theory of planned behavior, these self-directed changes suggest that wearables can increase health awareness and motivation to change, representing a promising approach to promote health behavior change among HHAs [11,12].

Although structural barriers such as power asymmetries among workers, patients, and agencies; low wages; and isolated work environments limit HHAs' ability to improve their health and well-being, wearable data could be integrated into existing peer support programs that aim to reduce worker isolation and strengthen community ties [38,39]. For example, in our prior studies, we have found that HHAs are more likely to engage in PA if they know peers or their clients are engaged as well. Our findings in this study may help bolster and extend these efforts, which have largely relied on HHAs' self-reported experiences. Aggregated wearable data could more concretely illustrate the challenges that HHAs face, enrich qualitative accounts, and guide the design of peer support programs by grounding discussions and activities in workers' collective experiences to identify strategies for navigating or mitigating these structural constraints.

Notably, the findings also suggest that PA may be more modifiable than sleep for the HHAs and have clearer positive implications for their patients. That is, several participants found that wearing their activity trackers reminded them of the importance of PA on the job and motivated them to spend more time walking with their patients to increase both their step

XSL•FO
RenderX

counts. This is an important observation as previous research has identified challenges in improving PA among older adults receiving home care [40] and researchers have suggested that activities such as walking are valuable for promoting PA outside of structured exercise programs [41].

In future studies, this relationship warrants investigation. Key questions include the following: does increasing HHAs' step count in turn increase their patients' PA and mobility and reduce the risk of falls? Does increasing time spent together moving foster closer relationships between HHAs and their patients through shared activity? Our work has previously found that higher levels of mutuality between HHAs and patients result in higher job satisfaction among HHAs [42], and this too could be formally examined after introducing technology. Another important area for exploration is the affordability of wearable devices for a workforce with limited disposable income (eg, the Fitbit Charge 6 costs approximately US $150 as of this writing). If these devices are shown to significantly impact worker and patient outcomes, they may be justified for inclusion in future workforce wellness programs. Alternatively, existing devices such as smartphones could offer a more cost-effective means of collecting data on PA and sleep. However, unlike wearable devices, smartphones are not continuously worn on the body, which may limit the accuracy of the data they capture. Regardless, the feasibility of using smartphones should be further examined.

Beyond improving personal health, these devices may help further center workers' experiences by introducing an additional layer of granular quantitative data to complement workers' qualitative narratives. This collection of data sources may help bolster ongoing advocacy efforts, such as supporting improving wages or better working conditions. For example, collective data can strengthen labor union and worker advocacy campaigns such as the push for Fair Pay for Home Care [43], which advocates for livable wages and job stability for HHAs. Similarly, the No More 24 campaign [26], which seeks to end the harmful practice of requiring HHAs to work 24-hour shifts, where HHAs work for 24 hours but are only paid for 13 hours under the premise that they receive at least 5 hours of uninterrupted sleep and other breaks, is directly relevant to HHAs' poor sleep levels. Our research suggests that tracking HHAs' activities could generate data-driven evidence on sleep patterns and quality, which, when combined with worker-centered reflections that contextualize these data, could provide compelling arguments for policymakers and bolster organizational advocacy efforts. Future research should explore how to responsibly manage and share HHAs' collective data in ways that center their needs and experiences to improve their health and well-being.

Limitations

We acknowledge that our study has several limitations. We conducted a small-scale, short-term empirical mixed methods study in a single geographic location, New York City. Further research is needed to understand how our findings might generalize to larger samples and other locales, including rural and non-US settings, particularly those where driving is a more frequent mode of transportation than walking or public transit. Participants were recruited through their affiliation with a labor management sector of a health care union and may have been more motivated to engage with the wearable devices than nonunionized HHAs, which could confound our findings and limit generalizability to the broader HHA workforce.

Our study did not ask separate questions regarding sex and gender or race and ethnicity, which may obscure participants' intersectional experiences. Future work should aim to include these perspectives. Additionally, researchers should investigate the results of a longitudinal study to understand how participants' experiences might change over time and how they objectively impact the patients they serve. Our study also has limitations inherent to qualitative interviews, including potential participant response bias [44]. Further research might explore techniques such as journaling and remotely distributed surveys to decrease the impact of researchers' positionality.

Conclusions

Our findings suggest that personal tracking devices offer a feasible and acceptable approach to supporting positive health-related changes in this historically overlooked workforce. These devices have the potential to improve both the health behaviors and well-being of HHAs and, potentially, of the patients for whom they care. These concepts should be formally tested in future clinical trials.

Acknowledgments

The authors would like to thank the 1199 Service Employees International Union Training and Employment Fund team for their assistance with recruitment. They would also like to thank the study participants for sharing their time and experience to make this study possible.

Funding

This work was supported in part by a Google Cyber NYC Research Award, an American Heart Association grant (23SCISA1142170), and the Initiative on Home Care Work at the Center for Applied Research at Cornell.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: IRS-K, AV, ND, MRS
Data curation: IRS-K, IY
Formal analysis: IRS-K, MD, MT, IY, KC, ND, MRS
Funding acquisition: AA, AV, ND, MRS
Investigation: IRS-K, MD, MT, IY, KC, RP, ND, MRS
Methodology: IRS-K, MD, MT, IY, KC, YE, AA, AV, ND, MRS
Project administration: IRS-K, MD, RP, MS, YE, ND, MRS
Resources: IRS-K, IY, YE
Software: IRS-K, IY
Supervision: IRS-K, JB, MRT, AA, AV, ND, MRS
Validation: IRS-K, MT, IY, RP, MS, ND, MRS
Visualization: IRS-K, MD, MT, IY, AV, ND, MRS
Writing—original draft: IRS-K, MD, RP, JB, MRT, ND, MRS
Writing—review and editing: IRS-K, MD, RP, ND, MRS

Conflicts of Interest

MRT would like to disclose her start-up, Iris OB Health Inc. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Qualitative interview guide.
[DOCX File , 1902 KB - jmir_v28i1e77510_app1.docx]

References

1. Markkanen P, Quinn M, Galligan C, Chalupka S, Davis L, Laramie A. There's no place like home: a qualitative study of the working conditions of home health care providers. J Occup Environ Med 2007 Mar;49(3):327-337. [doi: 10.1097/JOM.0b013e3180326552] [Medline: 17351519]
2. Occupational outlook handbook: home health aides and personal care aides. U.S. Bureau of Labor Statistics. URL: https://www.bls.gov/ooh/healthcare/home-health-aides-and-personal-care-aides.htm [accessed 2025-05-29]
3. Sterling MR, Shaw AL. Sharing the care-a patient and her caregivers. JAMA Intern Med 2019 Dec 01;179(12):1617-1618. [doi: 10.1001/jamainternmed.2019.4231] [Medline: 31566655]
4. Cho J, Toffey B, Silva AF, Shalev A, Safford MM, Phillips E, et al. To care for them, we need to take care of ourselves: a qualitative study on the health of home health aides. Health Serv Res 2023 Jun;58(3):697-704 [FREE Full text] [doi: 10.1111/1475-6773.14147] [Medline: 36815290]
5. Sterling MR, Kern LM, Safford MM, Jones CD, Feldman PH, Fonarow GC, et al. Home health care use and post-discharge outcomes after heart failure hospitalizations. JACC Heart Fail 2020 Dec;8(12):1038-1049 [FREE Full text] [doi: 10.1016/j.jchf.2020.06.009] [Medline: 32800510]
6. Thompson A, Fleischmann KE, Smilowitz NR, de Las Fuentes L, Mukherjee D, Aggarwal NR, Peer Review Committee Members. 2024 AHA/ACC/ACS/ASNC/HRS/SCA/SCCT/SCMR/SVM guideline for perioperative cardiovascular management for noncardiac surgery: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. Circulation 2024 Nov 05;150(19):e351-e442 [FREE Full text] [doi: 10.1161/CIR.0000000000001285] [Medline: 39316661]
7. Sterling MR, Silva AF, Leung PB, Shaw AL, Tsui EK, Jones CD, et al. "It's like they forget that the word 'health' is in 'home health aide'": understanding the perspectives of home care workers who care for adults with heart failure. J Am Heart Assoc 2018 Dec 04;7(23):e010134 [FREE Full text] [doi: 10.1161/JAHA.118.010134] [Medline: 30571599]
8. Bercovitz A, Moss A, Sengupta M, Park-Lee EY, Jones A, Harris-Kojetin LD. An overview of home health aides: United States, 2007. Natl Health Stat Report 2011 May 19(34):1-31 [FREE Full text] [Medline: 21688727]
9. Silver S, Boiano J, Li J. Patient care aides: differences in healthcare coverage, health-related behaviors, and health outcomes in a low-wage workforce by healthcare setting. Am J Ind Med 2020 Jan;63(1):60-73 [FREE Full text] [doi: 10.1002/ajim.23053] [Medline: 31631375]
10. Sterling MR, Li J, Cho J, Ringel JB, Silver SR. Prevalence and predictors of home health care workers' general, physical, and mental health: findings from the 2014 2018 behavioral risk factor surveillance system. Am J Public Health 2021 Dec;111(12):2239-2250. [doi: 10.2105/AJPH.2021.306512] [Medline: 34878879]
11. Bandura A. Health promotion from the perspective of social cognitive theory. Psychol Health 1998 Jul;13(4):623-649. [doi: 10.1080/08870449808407422]
12. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]

XSL•FO
RenderX

13. Moore K, O'Shea E, Kenny L, Barton J, Tedesco S, Sica M, et al. Older adults' experiences with using wearable devices: qualitative systematic review and meta-synthesis. JMIR Mhealth Uhealth 2021 Jun 03;9(6):e23832 [FREE Full text] [doi: 10.2196/23832] [Medline: 34081020]
14. Kononova A, Li L, Kamp K, Bowen M, Rikard RV, Cotten S, et al. The use of wearable activity trackers among older adults: focus group study of tracker perceptions, motivators, and barriers in the maintenance stage of behavior change. JMIR Mhealth Uhealth 2019 Apr 05;7(4):e9832 [FREE Full text] [doi: 10.2196/mhealth.9832] [Medline: 30950807]
15. Aparicio A, Inostroza-Correa MA, Miranda P, Salinas C, Herskovic V, Chackiel C, et al. The impact of technology use for care by informal female caregivers on their well-being: a scoping review. Syst Rev 2025 Apr 17;14(1):89 [FREE Full text] [doi: 10.1186/s13643-025-02817-z] [Medline: 40241162]
16. Mahmood A, Kim H, Kedia S, Dillon P. Wearable activity tracker use and physical activity among informal caregivers in the United States: quantitative study. JMIR Mhealth Uhealth 2022 Nov 24;10(11):e40391 [FREE Full text] [doi: 10.2196/40391] [Medline: 36422886]
17. Mänttäri S, Säynäjäkangas P, Selander K, Laitinen J. Increased physical workload in home care service is associated with reduced recovery from work. Int Arch Occup Environ Health 2023 Jul 18;96(5):651-660 [FREE Full text] [doi: 10.1007/s00420-023-01960-1] [Medline: 36808567]
18. Lohne FK, Fimland MS, Rasmussen CL, Liaset IF, Fischer H, Redzovic S. Is patients' activities of daily living self-care score in Norwegian home care a proxy for workers standing at work? BMC Health Serv Res 2024 May 09;24(1):565 [FREE Full text] [doi: 10.1186/s12913-024-10897-1] [Medline: 38724977]
19. Tjøsvoll SO, Wiggen Ø, Gonzalez V, Seeberg TM, Elez Redzovic S, Frostad Liaset I, et al. Assessment of physical work demands of home care workers in Norway: an observational study using wearable sensor technology. Ann Work Expo Health 2022 Nov 15;66(9):1187-1198 [FREE Full text] [doi: 10.1093/annweh/wxac052] [Medline: 35959647]
20. Grønoset Grasmo S, Frostad Liaset I, Redzovic SE. Home health aides' experiences of their occupational health: a qualitative meta-synthesis. Home Health Care Serv Q 2021 May 05;40(2):148-176 [FREE Full text] [doi: 10.1080/01621424.2021.1921650] [Medline: 33949920]
21. Yanez Hernandez M, Kuo EF, Henriquez Taveras Y, Lee A, Ramos A, Ringel J, et al. Mental health and well-being among home health aides. JAMA Netw Open 2024 Jun 03;7(6):e2415234 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.15234] [Medline: 38842806]
22. 1199SEIU benefit and pension funds. 1199SEIU Funds. URL: https://www.1199seiubenefits.org [accessed 2025-04-28]
23. MyTEF sign-up: 1199SEIU benefit and pension funds. 1199SEIU Funds. URL: https://www.1199seiubenefits.org/mytef-sign-up/ [accessed 2025-04-28]
24. Sterling M, Espinosa C, Vergez S, McDonald MV, Ringel J, Tobin J, et al. Home health aides caring for adults with heart failure: a pilot randomized clinical trial. JAMA Netw Open 2025 Nov 03;8(11):e2548121 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.48121] [Medline: 41213039]
25. Charters E. The use of think-aloud methods in qualitative research an introduction to think-aloud methods. Brock Educ J 2003 Jul 01;12(2):33. [doi: 10.26522/brocked.v12i2.38]
26. Home page. No More 24!. URL: https://nomore24.org/ [accessed 2025-05-29]
27. Diaz KM, Krupka DJ, Chang MJ, Peacock J, Ma Y, Goldsmith J, et al. Fitbit®: an accurate and reliable device for wireless physical activity tracking. Int J Cardiol 2015 Apr 15;185:138-140 [FREE Full text] [doi: 10.1016/j.ijcard.2015.03.038] [Medline: 25795203]
28. Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. Accuracy of wristband Fitbit models in assessing sleep: systematic review and meta-analysis. J Med Internet Res 2019 Nov 28;21(11):e16273 [FREE Full text] [doi: 10.2196/16273] [Medline: 31778122]
29. Feehan LM, Geldman J, Sayre EC, Park C, Ezzat AM, Yoo JY, et al. Accuracy of Fitbit devices: systematic review and narrative syntheses of quantitative data. JMIR Mhealth Uhealth 2018 Aug 09;6(8):e10527 [FREE Full text] [doi: 10.2196/10527] [Medline: 30093371]
30. Lee XK, Chee NI, Ong JL, Teo TB, van Rijn E, Lo JC, et al. Validation of a consumer sleep wearable device with actigraphy and polysomnography in adolescents across sleep opportunity manipulations. J Clin Sleep Med 2019 Sep 15;15(9):1337-1346 [FREE Full text] [doi: 10.5664/jcsm.7932] [Medline: 31538605]
31. Project Jupyter. Jupyter. URL: https://jupyter.org [accessed 2025-04-28]
32. Dröge K. kaixxx / noScribe. GitHub. URL: https://github.com/kaixxx/noScribe [accessed 2025-04-28]
33. How many steps for better health? National Institutes of Health. 2019. URL: https://www.nih.gov/news-events/nih-research-matters/how-many-steps-better-health [accessed 2025-04-28]
34. How much sleep do you need? National Center for Chronic Disease Prevention and Health Promotion. 2019. URL: https://magazine.medlineplus.gov/article/how-much-sleep-do-you-need [accessed 2025-04-28]
35. Beukenhorst AL, Howells K, Cook L, McBeth J, O'Neill TW, Parkes MJ, et al. Engagement and participant experiences with consumer smartwatches for health research: longitudinal, observational feasibility study. JMIR Mhealth Uhealth 2020 Jan 29;8(1):e14368 [FREE Full text] [doi: 10.2196/14368] [Medline: 32012078]
36. Caring for caregivers: a case study. Fitbit. URL: https://assets.ctfassets.net/0ltkef2fmze1/6Ja9Y4zhx0ubn0SFuNHkfQ/aac0b22552e9803009596c8e1e408bd0/FE_Customer-Case-Study_SEIU.pdf [accessed 2025-04-28]

XSL•FO
RenderX

37. Reading Turchioe M, Burgermaster M, Mitchell EG, Desai PM, Mamykina L. Adapting the stage-based model of personal informatics for low-resource communities in the context of type 2 diabetes. J Biomed Inform 2020 Oct;110:103572 [FREE Full text] [doi: 10.1016/j.jbi.2020.103572] [Medline: 32961309]
38. Olson R, Elliot D, Hess J, Thompson S, Luther K, Wipfli B, et al. The COMmunity of Practice And Safety Support (COMPASS) Total Worker Health™ study among home care workers: study protocol for a randomized controlled trial. Trials 2014 Oct 27;15:411 [FREE Full text] [doi: 10.1186/1745-6215-15-411] [Medline: 25348013]
39. Poon A, Guerrero L, Loughman J, Luebke M, Lee A, Sterling M, et al. Designing for peer-led critical pedagogies in computer-mediated support groups for home care workers. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023 Presented at: CHI '23; April 23-28, 2023; Hamburg, Germany p. 1-18 URL: https://dl.acm.org/doi/10.1145/3544548.3580881 [doi: 10.1145/3544548.3580881]
40. Burton E, Lewin G, Boldy D. Physical activity preferences of older home care clients. Int J Older People Nurs 2015 Sep;10(3):170-178. [doi: 10.1111/opn.12065] [Medline: 25400233]
41. Burton E, Farrier K, Galvin R, Johnson S, Horgan NF, Warters A, et al. Physical activity programs for older people in the community receiving home care services: systematic review and meta-analysis. Clin Interv Aging 2019;14:1045-1064 [FREE Full text] [doi: 10.2147/CIA.S205019] [Medline: 31239654]
42. Shalev A, Ringel JB, Riegel B, Vellone E, Stawnychy MA, Safford M, et al. Does connectedness matter? The association between mutuality and job satisfaction among home health aides caring for adults with heart failure. J Appl Gerontol 2023 Apr;42(4):747-757 [FREE Full text] [doi: 10.1177/07334648221146772] [Medline: 36541188]
43. Fair pay for home care. New York Caring Majority. URL: https://www.nycaringmajority.org [accessed 2025-04-28]
44. Dell N, Vaidyanathan V, Medhi I, Cutrell E, Thies W. "Yours is better!": participant response bias in HCI. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2012 Presented at: CHI '12; May 5-10, 2012; Austin, TX p. 1321-1330 URL: https://dl.acm.org/doi/10.1145/2207676.2208589 [doi: 10.1145/2207676.2208589]

Abbreviations

1199SEIU-TEF: 1199 Service Employees International Union Training and Employment Fund
HHA: home health aide or attendant
NIH: National Institutes of Health
PA: physical activity
SEIU: Service Employees International Union

Edited by A Stone; submitted 14.May.2025; peer-reviewed by F Materia, LF Belo; comments to author 14.Oct.2025; accepted 23.Dec.2025; published 26.Jan.2026.

Please cite as:
Solano-Kamaiko IR, Dicpinigaitis M, Tan M, Yang I, Cheng K, Peramsetty R, Shum M, Escamilla Y, Bayly J, Turchioe MR, Avgar A, Vashistha A, Dell N, Sterling MR
Feasibility, Acceptability, and Perspectives Regarding the Use of Activity Tracking Wearable Devices Among Home Health Aides: Mixed Methods Study
J Med Internet Res 2026;28:e77510
URL: https://www.jmir.org/2026/1/e77510
doi:10.2196/77510
PMID:

©Ian René Solano-Kamaiko, Michael Dicpinigaitis, Melissa Tan, Irene Yang, Kexin Cheng, Ronica Peramsetty, Michelle Shum, Yanira Escamilla, Jennifer Bayly, Meghan Reading Turchioe, Ariel Avgar, Aditya Vashistha, Nicola Dell, Madeline R Sterling. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Characteristics Influencing Support for the National Health Service COVID-19 App in England and Wales: Findings From a Longitudinal Survey

Josephine Exley[1], PhD; Paul Boadu[1], PhD; Kasim Allel[2], PhD; Bob Erens[1], MA; Nicholas Mays[1], MA; Mustafa Al-Haboubi[1], PhD

[1]London School of Hygiene & Tropical Medicine, London, United Kingdom

[2]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom

Corresponding Author:
Mustafa Al-Haboubi, PhD
London School of Hygiene & Tropical Medicine
15-17 Tavistock Place
London, WC1H 9SH
United Kingdom
Phone: 44 (0)20 7636 ext 8636
Email: Mustafa.Al-Haboubi@lshtm.ac.uk

Abstract

Background: The use of proximity (contact) tracing mobile phone apps during the COVID-19 pandemic to support manual contact tracing was novel. Uptake of the app was lower than expected.

Objective: We sought to identify distinct subgroups of individuals based on their level of support for the National Health Service (NHS) COVID-19 app in the first 15 months of the app's implementation, and to identify the attitudes and characteristics associated with membership of more and less supportive groups.

Methods: We conducted 8 waves of a longitudinal survey data of smartphone users, recruited from an online panel (n=2023 at baseline and n=1198 at survey wave 6) between October 14, 2020, and December 13, 2021. We used latent class analysis to identify subgroups of individuals with different inclinations of support for the NHS COVID-19 app. Sankey diagram analysis was used to assess individuals whose subgroup changed over the study period. We estimated population-weighted multinomial logistic regression models using sociodemographic characteristics as independent variables.

Results: We identified 4 subgroups in survey waves 1 to 4—"not supportive" (1765/7210, 25%), "ambivalent" (2124/7210, 30%), "somewhat supportive" (1421/7219, 20%), and "completely supportive" (1900/7210, 26%). At wave 5, a total of 3 subgroups of support for the app emerged—"not supportive" (549/1613, 34%), "ambivalent" (497/1613, 31%), and "supportive" (567/1613, 35%). From wave 6 onward, the results showed 4 subgroups emerging—"least supportive" (1568/6952, 23%), "less supportive" (1179/6952, 17%), "ambivalent" (2105/6952, 30%), and "supportive" (2100/6952, 29%). The majority of respondents remained within their identified subgroups between survey waves. Among those who moved into different subgroups, most moved into a less supportive subgroup. Exceptions to this were from waves 2 to 3 and from waves 3 to 4, when higher percentages of respondents moved into more supportive subgroups. The biggest movement to less supportive subgroups occurred after wave 1 (October 2020), when 38% (2740/7210) of respondents moved into a less supportive subgroup. The biggest movement to more supportive subgroups, on the other hand, occurred after wave 2, when 22% (1586/7210) of respondents moved into more supportive subgroups. Over the course of the 8 waves, the percentage of respondents in supportive subgroups declined from 56% (3353/5988) to 29% (1737/5988). Key characteristics of more supportive individuals included having higher levels of trust in the government to control the spread of COVID-19 and having the app installed, while those less concerned about the risk COVID-19 posed to the country were more likely to be unsupportive ($P<.05$).

Conclusions: When the app was launched, just over half of respondents were supportive, but this declined over the following 15 months. The attrition in support poses important challenges for governments to the use of apps in future pandemics. A potential reason was mistrust in the government's handling of the pandemic.

(J Med Internet Res 2026;28:e76863) doi:10.2196/76863

XSL•FO
RenderX

KEYWORDS

contact tracing apps; COVID-19; infection prevention and control; public health; test and trace

Introduction

As part of the response to the COVID-19 pandemic, several countries deployed novel proximity (contact) tracing mobile phone apps to supplement manual contact tracing efforts. They keep a temporary record of app users who have been in close contact with each other to alert them in case one user subsequently reports through the app contracting the virus. Evidence highlights the potential for apps to reduce the spread of COVID-19 [1-7]. For example, the National Health Service (NHS) COVID-19 app was estimated to have averted approximately 1 million cases and 10,000 deaths in England and Wales in its first year [8], and for every percentage increase in uptake the number of cases was estimated to have been reduced by up to 2% [9].

The effectiveness of such technology depends on the level of uptake among the population [1,10-12]. In settings where their use was not mandated, uptake was reliant on individuals' willingness to install and use the app. However, uptake of contact tracing apps in many countries was found to be lower than anticipated [13-17], although surveys conducted before the introduction of apps suggested high levels of public support for the use of these apps as part of the pandemic response [18,19]. Evidence from our own work following a cohort of smartphone users in England and Wales during the first year of implementation of the NHS COVID-19 app found that despite changes in policy and case numbers, installation of the app remained relatively stable at around 50% [20]. The majority of those who had ever installed the app did so soon after the app's launch.

With scientists predicting a growing probability of future pandemics [21], contact tracing apps are likely to become an increasingly ubiquitous public health tool. As such, there is a need to learn how to improve the implementation and uptake of such apps in the face of likely future pandemics. In this analysis of a longitudinal 8-wave survey of a representative sample of adult smartphone users in England and Wales, we aimed to identify distinct subgroups of individuals based on their level of support for the NHS COVID-19 app in the first 15 months of the app's implementation and to identify the attitudes and characteristics associated with membership of more and less supportive groups.

Methods

This study draws on longitudinal (prospective cohort) survey data of smartphone users aged 18 to 79 years in England and Wales. Details of the study have been described previously [20].

Setting

The NHS COVID-19 app became part of the NHS Test and Trace Programme in July 2020 and was launched in England and Wales on September 24, 2020 (Textbox 1). It alerted close contacts of individuals who later tested positive for coronavirus to self-isolate. Additional features allowed users to check their symptoms, book a COVID-19 test, and check in at a venue. The survey started 2 weeks after the app was launched and coincided with rising COVID-19 cases and the tightening of COVID-19 restrictions. Multimedia Appendix 1 maps the 8 rounds of the online survey against the number of new COVID-19 cases reported and the key policy changes introduced by the governments in England and Wales during the study period. COVID-19 cases were highest during survey wave 3 (conducted between December 28, 2020, and January 6, 2021) and survey wave 8 (conducted between November 25, 2021, and December 13, 2021). Cases were lowest during survey wave 5 (conducted between March 15 and 31, 2021).

Textbox 1. Overview of the National Health Service (NHS) COVID-19 app and the key COVID-19 restrictions in place in England and Wales during the first year of the app's launch.

The NHS COVID-19 contact tracing app, using a decentralized model, was launched in England and Wales on September 24, 2020, for users aged 16 years and older and was made available in 12 languages.

It required a recent smartphone operating system and only collected users' postcode areas to provide local risk information. By July 2022, it had over 31 million downloads [22]. The app sent anonymous alerts to close contacts (determined by proximity and duration algorithms) of users who reported positive COVID-19 tests in the app [23]. Unlike public health authorities' self-isolation instructions (which ended in February 2022), app guidance was not legally binding. Users needed Bluetooth enabled for contact tracing, with an option within the app to temporarily disable it. Other features included symptom checking, test ordering, venue check-ins by QR codes (mandatory in England from September 2020 to February 2022 and a recommendation in Wales), and a self-isolation timer.

The app's launch coincided with rising COVID-19 cases and the tightening of restrictions [24]. Wales introduced a 2-week firebreak on October 23, while England entered a second national lockdown on November 5. Both countries briefly returned to tiered restrictions in December before reimposing lockdowns by year-end. In early 2021, England entered a third lockdown, and Wales maintained the highest restriction level ("stay at home"). Restrictions were slowly eased throughout spring and summer 2021, with most lifted by July 19, including the need to self-isolate if double vaccinated. The emergence of the Omicron variant in November 2021 once again marked the tightening of restrictions.

As shown in Multimedia Appendix 1, survey wave 1 (October 14-22, 2020) took place after the introduction of the local 3-tier system of COVID-19 restrictions; at that time, it was a legal requirement for venues in England to take the contact details of visitors either manually or by scanning a QR code, while in Wales this was a recommendation only. Survey wave 2 (November 12-23, 2020) was conducted during the second national lockdown, introduced on November 5, and followed a 2-week firebreak in Wales from October 23 to November 8 in which the public were instructed to stay at home and the

XSL•FO
RenderX

hospitality industry and nonessential restaurants had to close. Survey wave 3 (December 28, 2020, to January 6, 2021) was conducted over the Christmas period when the country had returned to the tier system; the southeast of England and the whole of Wales moved to a newly created tier 4 ("stay at home"), the first vaccine was administered in the United Kingdom on December 8, 2020, and England entered the third national lockdown on January 6, 2021. Survey wave 4 (February 1-15, 2021) occurred while England remained in the third national lockdown and Wales remained at tier 4. Survey wave 5 (March 15-31, 2021) took place at the start of the easing of restrictions, after the government in England outlined a 4-step plan; by the start of the survey, schools had reopened and outdoor mixing in groups of up to 6 was allowed from March 29. Survey wave 6 (July 1-18, 2021) occurred ahead of the final step of easing restrictions in both England and Wales, when indoor venues had reopened and the public were allowed to meet in groups of up to 30 outdoors and 6 indoors. Survey wave 7 (August 31-September 13, 2021) took place after all legal limits on social contact had been lifted, including the requirement to check in to venues in England, and those who had been double vaccinated were no longer required to self-isolate if they had come into contact with someone who tested positive for COVID-19, provided they did not have any symptoms. Survey wave 8 (November 25-December 13, 2021) occurred at the same time as the Omicron variant was first being reported; a total of 6 southern African countries were placed on the travel red list and the first cases were detected in the United Kingdom, and on December 8 the government announced that Plan B measures were to be reintroduced, including mandatory face coverings in most indoor settings, working from home, and use of the NHS COVID Pass for entry into nightclubs and settings where large crowds gather. Data on daily new cases were obtained from the UK Coronavirus Dashboard, and information on lockdown and restriction timelines from the Institute for Government's coronavirus timeline.

Participants

We were interested in looking at differences in the usage of the app between different sociodemographic groups, such as age, level of education, social grade, health status, and minoritized ethnic groups. We expected to find differences in attitudes between members of our sample in different social grades (nonmanual vs manual). In order to detect a 6% difference (at 80% power and a 95% significance level) in attitudes between these 2 subgroups, a sample size of 1657 is required. More importantly, we wanted to detect differences in app use over time. Assuming about two-thirds of the sample at baseline were using the app, we would be able to detect about a 3% change in the overall sample between baseline/wave 1 and survey wave 2 with 95% confidence.

A representative sample of smartphone users aged 18 to 79 years was recruited through YouGov's volunteer online panel, with quotas set on age, gender, and social grade. Panel members are recruited from different sources, including through advertising and partnerships with a range of websites. Sociodemographic data are collected when they join the panel. Participants are invited from the panel in a way that seeks to generate a nationally-representative sample [25].

In total, 2023 participants were recruited at the start of the study (referred to as the "baseline sample"). Panel members who completed the baseline survey were invited to all follow-up waves. After the fifth round of data collection, we recruited an additional 1198 participants to increase the sample size (referred to as the "additional sample"). At survey wave 8, 61% (1233/2023) of the main sample and 71% (848/1198) of the additional sample responded. The response rate for each sample at each survey wave is provided in Table S1 in Multimedia Appendix 2.

Data Collection and Variables of Interest

Online surveys were undertaken roughly every 6 weeks between October 14, 2020, and December 13, 2021 (Multimedia Appendix 1). The survey was developed by the study team and adapted from previous surveys. It included questions on attitudes toward the app, including the level of support for it (measured on a 5-point Likert scale from 1 "not at all supportive" to 5 "completely supportive"), as well use of the app, perceptions of COVID-19, governments' response to the pandemic, and demographic information (ie, age, sex, ethnicity, income, employment, etc). The core content of the questionnaire remained the same across all waves, but some questions were added or removed in response to the evolving pandemic.

The survey was administered by YouGov and sent to respondents by email. Participants who completed the first survey (survey wave 1 for the baseline sample or wave 6 for the additional sample) were invited to participate in all subsequent waves of the survey.

Analysis

Determining Latent Classes of Support for NHS COVID-19 App

We used latent class analysis (LCA) to identify distinct subgroups of individuals with similar levels of support for the app at each survey wave [26] using the *gsem* function in Stata [27]. We considered LCA an optimal strategy due to its robustness against potential measurement errors, such as potential biases introduced by the timing of the questionnaire, and its ability to parsimoniously model restricted latent groups across multiple levels of support. This approach also helps prevent measurement errors that may arise from assigning respondents to groups based on their survey responses without recourse to statistical justification. To determine the optimal number of classes to include in the LCA model at each survey wave, we estimated the Bayesian information criteria (BIC) and Akaike information criteria (AIC) maximum likelihood values, with lower values indicating a better fit for the number of subgroups [28,29]. The results were similar when the test was conducted with and without covariates, including general health, disability, vulnerability to COVID-19, region, household income, tenure, app installation, trust in government, age, ethnicity, and sex. The results of the maximum likelihood test are presented in Table S2 in Multimedia Appendix 2.

Sankey Diagram Analysis

To assess the movement of individuals between subgroups of support for the app over survey waves, we constructed a Sankey

XSL•FO
RenderX

diagram of the wave-specific estimated classes and examined individual transitions between identified support subgroups over the survey period [30,31]. The analysis was restricted to individuals who responded to 2 consecutive waves.

Regression Analysis

To examine the factors associated with belonging to a given subgroup of support for the app, we estimated population-weighted multinomial logistic regression models at three timepoints: survey wave 1, survey wave 5, and survey wave 8. We opted to estimate the regression model at the study start and endpoints, but additionally included wave 5 as it represents a notable change in the distribution of LCA results, in that the number of subgroups reduced from 4 to 3. The definition of dependent and independent variables included in the multinomial logistic regression model are presented in Table S3 in Multimedia Appendix 2.

To select the independent variables for the model and mitigate multicollinearity, a correlation matrix (Tables S4-S6 in Multimedia Appendix 2) and variance inflation factors (VIF) (Tables S7-S9 in Multimedia Appendix 2) were generated. Based on the results, the following explanatory variables were included in the model: age, gender, ethnicity, self-reported health status, whether day-to-day activities were limited because of a health problem or disability, perceived vulnerability to COVID-19, region of residence, household income, household ownership, whether the app was installed at the current survey wave, whether the respondent had a COVID-19 infection since the previous survey, extent of trust in the government to control the spread of COVID-19, extent of concern about the risk COVID-19 poses to oneself, and extent of concern about the risk COVID-19 poses to the country. Fisher statistics are reported for measuring the model's goodness-of-fit. All statistical analyses were performed using Stata Standard Edition (version 18; StataCorp LLC) software. The research is reported in line with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) Statement for cohort studies [32] (Multimedia Appendix 3).

Ethical Considerations

Consent was sought before starting the first survey but not at subsequent waves. Responses were anonymized by YouGov before being passed to the researchers. Participants were free to withdraw at any time without needing to provide a reason. Ethics approval was obtained from the London School of Hygiene & Tropical Medicine Research Ethics Committee (reference number 22483). Participants receive "points" for completing surveys, which can subsequently be converted into cash payments.

Results

Overview

Overall, a total of 3221 participants were recruited to the study (2023 at survey wave 1 and 1198 at survey wave 6). The characteristics of participants are presented in Table 1. About 53% (1118/2023) of participants at wave 1 and 53% (619/1198) of participants at wave 6 were female, 42% (810/2023) and 39% (467/1198) were aged 40 years or younger, 86% (2775/2023) and 87% (1045/1198) were White, and 61% (1235/2023) and 57% (679/1198) had received higher education in the baseline and additional samples, respectively.

XSL•FO
RenderX

Table 1. Characteristics of respondents to an 8-wave longitudinal survey on use and views on the National Health Service (NHS) COVID-19 app (October 2020 to December 2021). A currency exchange rate of GBP £1=US $1.32 is applicable.

Characteristics	Baseline sample (n=2023), n (%)	Additional sample (n=1198), n (%)
Sex		
Male	905 (47.2)	579 (48.3)
Female	1118 (52.8)	619 (51.7)
Age group (years)		
18-29	439 (23.1)	218 (18.2)
30-39	371 (18.4)	249 (20.8)
40-49	381 (20.8)	220 (18.4)
50-59	381 (17.5)	210 (17.5)
60-69	298 (12.1)	192 (16.0)
70-79	153 (8.1)	109 (9.1)
Ethnicity		
White	1775 (86.1)	1045 (87.2)
All other ethnic groups	248 (13.9)	153 (12.8)
Region		
North East	103 (4.4)	30 (2.5)
North West	252 (12.4)	102 (8.5)
Yorkshire and the Humber	185 (9.1)	94 (7.8)
East Midlands	174 (8.0)	74 (6.2)
West Midlands	176 (9.9)	63 (5.3)
East of England	209 (10.4)	84 (7.0)
London	283 (15.2)	163 (13.6)
South East	320 (15.8)	101 (8.4)
South West	197 (9.5)	69 (5.8)
Wales	124 (5.4)	418 (34.9)
Highest level of education attainment		
No formal qualifications	78 (3.5)	42 (3.5)
GCSE[a] or equivalent	232 (11.3)	143 (11.9)
A-level or equivalent	293 (15.6)	200 (16.7)
Higher education	1235 (60.8)	679 (56.7)
Other	137 (6.4)	92 (7.7)
Prefer not to say or did not answer	48 (2.5)	42 (3.5)
Employment status		
Currently working	1151 (56.8)	693 (57.9)
Not currently working	92 (4.7)	33 (2.8)
Voluntary work or carer	96 (4.6)	58 (4.8)
Unemployed or permanent sick leave	222 (11.0)	102 (8.5)
Education	81 (5.7)	80 (6.7)
Retired	344 (15.6)	215 (18.0)
Other	37 (1.7)	17 (1.4)
Household income (£)		
<14,999	224 (10.8)	130 (10.9)

XSL•FO
RenderX

Characteristics	Baseline sample (n=2023), n (%)	Additional sample (n=1198), n (%)
15,000-24,999	285 (14.0)	197 (16.4)
25,000-34,999	251 (12.7)	154 (12.9)
35,000-60,000	431 (20.9)	227 (19.0)
>60,000	346 (17.5)	188 (15.7)
Prefer not to say or did not answer	486 (24.1)	302 (25.2)
Personal income (£)		
<14,999	510 (25.7)	300 (25.0)
15,000-24,999	382 (18.4)	237 (19.8)
25,000-34,999	319 (15.6)	157 (13.1)
35,000-60,000	267 (12.7)	154 (12.9)
>60,000	107 (5.5)	43 (3.6)
Prefer not to say or did not answer	438 (22.0)	307 (25.6)
IMD[b]		
1–most deprived	351 (17.9)	218 (18.2)
2	373 (18.0)	246 (20.6)
3	408 (20.3)	233 (19.5)
4	442 (21.3)	231 (19.3)
5–least deprived	448 (22.5)	269 (22.5)
Living arrangements		
lives alone	372 (18.3)	211 (17.6)
other adult(s), no children	1151 (55.4)	633 (52.8)
children, no other adults	49 (2.4)	32 (2.7)
other adult(s) and children	451 (23.9)	322 (26.9)
Household ownership		
Own	1169 (55.9)	683 (60.1)
Rent	589 (29.5)	292 (25.7)
Live with friends or family	233 (13.0)	133 (11.7)
Other	32 (1.6)	28 (2.5)
Self-reported health status		
Very good	535 (27.1)	297 (24.8)
Good	988 (49.2)	572 (47.8)
Fair	352 (16.6)	234 (19.5)
Bad or very bad	117 (5.4)	72 (6.0)
Not answered	31 (1.7)	23 (1.9)
Day-to-day activities limited because of a health problem or disability lasting (or expected to last) at least 12 months		
Limited a lot	159 (7.3)	108 (9.0)
Limited a little	325 (15.5)	195 (16.3)
No	1498 (74.9)	863 (72.0)
Not answered	41 (2.2)	32 (2.7)

[a]GCSE: General Certificate of Secondary Education.

[b]IMD: index of multiple deprivation based on respondent's usual place of residence [33].

XSL•FO
RenderX

Latent Classes of Support for the App

The test results (BIC/AIC) of the number of latent classes show 4 subgroups of population with differences in support for the NHS COVID-19 app for all survey waves, except for wave 5, where 3 subgroups were identified (Table S2 in Multimedia Appendix 2). The distribution of the identified subgroups in each survey wave is presented in Table 2.

Table 2. Distribution of subgroups of support for the National Health Service (NHS) COVID-19 app across 8 survey waves (W1–W8) from an 8-wave longitudinal survey conducted between October 2020 and December 2021 (n=number of observations per wave; W=survey wave).

Level of support	W1 (Oct 14-22, 2020; n=2023), %	W2 (Nov 12-23, 2020; n=1781), %	W3 (Dec 28, 2020-Jan 6, 2021; n=1732), %	W4 (Feb 1-15, 2021; n=1674), %	W5 (Mar 15-31, 2021; n=1613), %	W6 (Jul 1-18, 2021; n=2667), %	W7 (Aug 31-Sep 13, 2021; n=2204), %	W8 (Nov 25-Dec 13, 2021; n=2081), %
Least supportive	16[a]	28[a]	28[a]	27[a]	25[a]	20	23	25
Less supportive	—	—	—	—	—	17	18	16
Ambivalent	28	32	29	29	36	32	29	30
Somewhat supportive	20	19	20	19	39[b]	32[b]	29[b]	29[b]
Completely supportive	36	21	22	25	—	—	—	—

[b]This percentage value applies to both the “least supportive” and “less supportive” subgroups; in Waves 1-5 these categories were statistically measured as a single latent class.

[b]This percentage value applies to both the “somewhat supportive” and “completely supportive” subgroups; in Waves 5-8 these categories were statistically measured as a single latent class.

This percentage value applies to both the “least supportive” and “less supportive” subgroups; in Waves 1-5 these categories were measured as a single latent class.

The characteristic of individuals belonging to the identified subgroups were similar for survey waves 1-4 and survey waves 6-8, so they were presented and analyzed together.

From survey waves 1 to 4 (October 14, 2020, to February 15, 2021), we identified 4 subgroups of support for the app. Table 3 shows the characteristics of the 4 population subgroups with different levels of support for the NHS COVID-19 app. Individuals in subgroup 1 were either not supportive at all or least supportive of the app, with good health and no disability, not vulnerable to COVID-19, generally equally distributed across all income groups and regions, and the majority were homeowners. Most had never installed the app and had no trust in the government to control the spread of COVID-19. The average age of individuals in subgroup 1 was 46 years, and they comprised similar proportions of individuals from all ethnic groups, with higher proportions of males. Based on these characteristics, we labeled this population subgroup as “not supportive.”

XSL•FO
RenderX

Table 3. Summary of characteristics of individuals belonging to the 4 subgroups of support for the National Health Service (NHS) COVID-19 app (survey wave 1 to wave 4 [October 14, 2020, to February 15, 2021]). Table S10 in Multimedia Appendix 2 provides underlying statistics.

Characteristic	Subgroup 1: Not Supportive	Subgroup 2: Ambivalent	Subgroup 3: Somewhat Supportive	Subgroup 4: Completely Supportive
Support	Not at all supportive, 51% (901/1765) or least supportive, 49% (863/1765)	Indifferent about support for the app, 100% (2124/2124)	Somewhat supportive, 100% (1421/1421)	Completely supportive of the app, 100% (1900/1900)
General health	Majority in good health, 52% (813/1765)	Majority in good health, 53% (1134/2124)	Majority in good health, 53% (754/1421)	Higher proportion with good health, 46% (871/1900)
Disability	Majority without disability, 75% (1317/1765)	Majority without disability, 75% (1598/2124)	Majority without disability, 75% (1060/1421)	Majority without disability, 72% (1359/1900)
Vulnerability to COVID-19	Majority not vulnerable, 59% (1039/1765)	Majority not vulnerable, 59% (1258/2124)	Majority not vulnerable, 57% (815/1421)	Majority vulnerable, 51% (960/1900)
Region	Generally evenly spread across all regions	Generally evenly spread across all regions	Generally evenly spread across all regions	Generally evenly spread across all regions
Household income	Generally evenly spread across income groups	Generally evenly spread across income groups	Generally evenly spread across income groups	Generally evenly spread across income groups
Tenure	Majority homeowners, 58% (1015/1765)	Majority homeowners, 56% (1181/2124)	Majority homeowners, 58% (831/1421)	Majority homeowners, 64% (1213/1900)
App installation	Majority never installed, 62% (1093/1765)	Higher proportion never installed, 44% (935/2124)	Majority installed, 59% (845/1421)	Majority installed, 74% (1409/1900)
Trust	Higher proportion with not at all trust, 84% (854/1765)	Higher proportion with little trust, 38% (803/2124)	Higher proportion with little trust, 37% (521/1421)	Higher proportion with fair amount of trust, 34% (646/1900)
Age (average)	46 years	45 years	45 years	49 years
Ethnicity	Similar proportion across all ethnic groups	Higher proportions among other ethnic groups, 35% (298/855)	Similar proportion across all ethnic groups	Higher proportions among White individuals, 27% (1714/6355)
Sex	Higher proportion of males, 27% (859/3232)	Higher proportion of females, 33% (1281/3978)	Similar proportions of males and females	Higher proportion of males, 28% (890/3232)

Most of the individuals in subgroup 2 neither supported nor opposed the NHS COVID-19 app and indicated that they had good health and were not vulnerable to COVID-19. Most were homeowners, had never installed the app, and did not have very much trust in the government to control the spread of COVID-19. Most were from ethnic groups other than White, with a higher proportion being female, and their average age was 45 years.

They were equally distributed across regions and household incomes, with a higher proportion being female. We labeled this subgroup “Ambivalent” based on their characteristics.

The third subgroup was labeled “Somewhat supportive” because individuals in this group had some level of support for the app. Most were in good health, had no disability, and did not consider themselves vulnerable to COVID-19. A higher proportion had never installed the app and did not have much trust in the government to control the spread of COVID-19. Individuals in this subgroup were generally equally distributed across regions, income groups, sex, and ethnic groups, and the majority were homeowners, with an average age of 45 years.

Individuals in the fourth subgroup were completely supportive of the NHS COVID-19 app, a higher proportion were in good health, and the majority had no disability. However, the majority also indicated they were vulnerable to COVID-19, were homeowners, had installed the app, and had a fair amount of trust in the government to control the spread of COVID-19. Similar to the other subgroups, individuals in this subgroup were generally equally distributed across all income groups and regions. A higher proportion of individuals in this group were of White ethnic group and were male. The average age of this subgroup was 49 years.

The number of subgroups identified in wave 5 (March 15-31, 2021) was reduced by one to 3. Table 4 shows the characteristics of individuals in each of the identified subgroups.

XSL•FO
RenderX

Table 4. Summary of characteristics of individuals belonging to the 3 subgroups of support for the National Health Service (NHS) COVID-19 app (survey wave 5 [March 15-31, 2021]). Table S11 in Multimedia Appendix 2 provides underlying statistics.

Characteristic	Subgroup 1: Not Supportive	Subgroup 2: Ambivalent	Subgroup 3: Supportive
Support	Not at all supportive, 52% (288/549) or least supportive, 48% (261/549)	Indifferent about support for app 100% (497/497)	Somewhat supportive, 46% (258/567) or completely supportive of the app, 54% (309/567)
App installation	Majority never installed, 56% (308/549)	Higher proportion either installed, 41% (205/497) or never uninstalled, 37% (186/497)	Majority installed, 73% (413/567)
Trust	Higher proportion with no trust at all, 35% (193/549)	Higher proportion with a fair amount of trust, 41% (202/497)	Higher proportion with a fair amount of trust, 42% (236/567)
Age (average)	47 years	45 years	48 years
Sex	Higher proportion of males, 37% (268/721)	Higher proportion of females, 34% (306/892)	Higher proportion of males, 36% (262/721)

Unlike the subgroups identified from study waves 1-4 (October 14, 2020, to February 15, 2021), characteristics such as general health, disability status, vulnerability to COVID-19, region, household income, and tenure were not statistically different among the 3 identified subgroups in wave 5. Based on the characteristics of individuals in subgroup 1, we labeled it "Not supportive." Those in this subgroup were not at all or least supportive of the NHS COVID-19 app, the majority had never installed the app, and a higher proportion had no trust at all in the government to control the spread of COVID-19. The average age of this subgroup was 47 years, and a higher proportion were male. Following a similar approach, subgroups 2 and 3 were labeled "Ambivalent" and "Supportive," respectively.

Furthermore, the results from survey waves 6-8 (July 1, 2021, to December 13, 2021) show the reemergence of 4 subgroups of support for the NHS COVID-19 app, with some subgroups showing different characteristics than those identified in survey waves 1-4 (Table 5). Based on the characteristics of the individuals in the identified subgroups, we labeled subgroups 1, 2, 3, and 4 as "Least supportive," "Less supportive," "Ambivalent," and "Supportive," respectively (Table 5).

Table 5. Summary of characteristics of individuals belonging to the 4 subgroups of support for the National Health Service (NHS) COVID-19 app (survey waves 6-8 [July 1, 2021, to December 13, 2021]). Table S12 in Multimedia Appendix 2 provides underlying statistics.

Characteristic	Least Supportive	Less Supportive	Ambivalent	Supportive
Support	Not at all supportive, 100% (1568/1568)	Least supportive, 100% (1179/1179)	Indifferent about support for app, 100% (2105/2105)	Somewhat supportive, 48% (999/2100) or completely supportive of app, 52% (1101/2100)
General health	Higher proportion in good health, 48% (745/1568)	Higher proportion in good health, 51% (607/1179)	Higher proportion in good health, 50% (1043/2105)	Higher proportion in good health, 48% (1016/2100)
Disability	Majority without disability, 71% (1109/1568)	Majority without disability, 77% (903/1179)	Majority without disability, 70% (1484/2105)	Majority without disability, 69% (1444/2100)
Region	Higher proportion in Wales, 24% (318/1301)	Higher proportions in London and South and the North	Higher proportion in Midlands and East of England	Higher proportion in London and South and Wales
Household income	Generally evenly spread across income groups	Generally evenly spread across income groups	Generally evenly spread across income groups	Generally evenly spread across income groups
Tenure	Majority homeowners, 64% (1008/1568)	Majority homeowners, 61% (721/1179)	Majority homeowners, 59% (1246/2105)	Majority homeowners, 64% (1335/2100)
App installation	Majority never installed, 68% (1074/1568)	Higher proportion never installed, 47% (549/1179)	Higher proportion installed, 43% (901/2105)	Majority installed, 76% (1604/2100)
Trust	Higher proportion with no trust at all, 40% (630/1568)	Higher proportion with little trust, 37% (436/1179)	Higher proportion with a fair amount of trust, 37% (781/2105)	Higher proportion with a fair amount of trust, 40% (834/2100)
Age (average)	49 years	46 years	48 years	49 years
Sex	Higher proportion of males, 26% (856/3289)	Higher proportion of females, 18% (675/3663)	Higher proportion of females, 33% (1216/2105)	Higher proportion of males, 32% (1040/3289)

The Sankey diagram presented in Multimedia Appendix 4 shows the movement of respondents between classes over the course of the surveys. Between each survey wave, more than half of respondents did not change class. Among those who did, most moved into a less supportive class (measured by the thickness of the palette), except between survey waves 2 to 3 and 3 to 4,

XSL•FO
RenderX

where more individuals moved upwards into more supportive classes than downwards. The biggest movement downwards occurred after survey wave 1 (October 2020) when 38% (275/723) of individuals moved into a less supportive class, half of whom (130/723, 18%) moved out of the completely supportive class to the somewhat supportive class, and over a third (108/723, 15%) moved into the Ambivalent subgroup. The biggest movement upwards occurred after survey wave 2 (November 2020) with 22% (233/1060) of individuals moving to more supportive classes. As a result of the greater number moving into less supportive classes, the percentage of individuals in "supportive" subgroups decreased over time, from 56% (1131/2023; "completely" or "somewhat supportive") at survey wave 1 to 29% (611/2081; "supportive") by survey wave 8.

Factors Associated With Class Membership

To quantify the factors associated with the likelihood of an individual's membership in a particular subgroup at survey waves 1 (October 2020), 5 (March 2021), and 8 (December 2021), we calculated the relative risk ratio (RRR) of variables being present in a particular subgroup relative to the reference group (completely supportive). An RRR<1 suggests a lower likelihood relative to the reference group, while an RRR>1 indicates a higher likelihood of the variable's presence. The full results of the multinomial logistic regression are presented in Tables S13-S15 in Multimedia Appendix 2 and summarized in Multimedia Appendix 5.

At all 3 waves, individuals who never had the app installed and had less trust in government to control the spread of COVID-19 were associated with a relative reduction in the likelihood of being in the most supportive subgroups compared with all other subgroups. For example, never having the app installed reduced the likelihood of being in the most supportive group compared with the least supportive group (RRR 0.03, 95% CI 0.02-0.05 at survey wave 1; RRR 0.07, 95% CI 0.06-0.10 at survey wave 5; and RRR 0.03, 95% CI 0.02-0.05 at survey wave 8). Similarly, having little or no trust in the government to control the spread of COVID-19 reduced the likelihood of being in the most supportive group compared with the least supportive group (RRR 0.19, 95% CI 0.12-0.29 at survey wave 1; RRR 0.27, 95% CI 0.20-0.38 at survey wave 5; and RRR 0.30, 95% CI 0.19-0.47 at survey wave 8).

At all 3 waves, not being at all concerned about the risks COVID-19 posed to the country was associated with a reduced likelihood of being in the most supportive subgroups compared with the least supportive subgroup. Compared with individuals who were very concerned about the risks COVID-19 posed to the country, being not at all concerned increased the likelihood of being in the "not supportive" group at survey wave 1 (RRR 10.87, 95% CI 2.63-44.85) and survey wave 5 (RRR 9.92, 95% CI 3.03-32.48) and increased the likelihood of being in the "least supportive" subgroup at survey wave 8 (RRR 13.72, 95% CI 2.56-73.62).

Discussion

Overview

We examined how levels of support for the COVID-19 app changed in the 15 months following its launch in England and Wales. We found just over half of respondents were supportive of the app around the time of its launch and that support declined over time, with a notable drop-off in those who were completely supportive occurring between the first (October 14-22, 2020) and second (November 12-23, 2020) survey wave. From survey wave 6 (July 1-18, 2021), more respondents were unsupportive than supportive of the app. Key characteristics of individuals who were more supportive included having higher levels of trust in the government to control the spread of COVID-19, having the app installed, and being more concerned about the risk COVID-19 posed to the country.

The level of support observed in this study was lower than might have been expected based on studies conducted before the app was launched [18,19]. A contact tracing app was initially framed as a central component of the government's strategy to control the spread of COVID-19, and media coverage in early 2020 was positive [34,35]. Yet following technical issues during piloting, including abandoning the first design [36], the government's framing of the app changed from a technological solution to a more experimental technology [35]. The development process received prominent critical media coverage, which often presented contact tracing apps as controversial [37] and narrowed the debate to privacy concerns around what access authorities would have to personal data on users' phones [38,39]. Among the cohort included in this study, being worried about privacy was the key reason for not downloading the app [20].

Lower levels of support by the time of the national launch also likely reflect wider public perceptions of the government's handling of the pandemic. By autumn 2020, trust in the government to handle the pandemic was low [40,41], and its launch coincided with rising COVID-19 cases and the tightening of restrictions (Multimedia Appendix 1). Soon after the app's launch, a range of new restrictions were introduced, including a second national lockdown and a modified tiered system [42]. The government's approach was widely characterized as confusing and chaotic [43,44], exemplified by the last-minute scrapping of plans in England to relax restrictions over the 2020 Christmas period, and likely contributed to negative attitudes toward the government's handling of the pandemic. In turn, this may have contributed to the perception that the app was ineffective at controlling the spread of COVID-19 and did not live up to earlier positive expectations [45,46].

Unlike levels of trust, which were found to fluctuate over the study period [41], levels of support consistently declined. Following the app's launch, individuals' mobility and social contact were highly restricted, with nonessential indoor leisure facilities and outdoor hospitality only reopening in April 2021. This might have fed perceptions that the app had limited utility at an individual level. Evidence on uptake indicates that the venue check-in feature was a driver of uptake [20]. Yet by July 2021, most social distancing requirements (including the

XSL•FO
RenderX

requirement to check in to venues) were dropped. The shift observed at survey wave 6 (July 1-18, 2021), with more people reporting they were unsupportive than supportive for the first time, potentially indicates that individuals were no longer placing much value on the app. Among those who uninstalled the app, not finding it useful was a key reason for doing so [20]. When restrictions were introduced again in autumn 2021 in response to the Omicron variant, the requirement to check in to venues was not reintroduced, which potentially reinforced perceptions that the app did not have a major role to play in controlling the spread of COVID-19.

Declining support might also reflect perceptions around the reliability of the app. Issues experienced by users in the weeks following the app's launch, including "ghost notifications," were widely reported in the mainstream media [47], and analysis of reviews posted in app stores showed that problems associated with the app's functionality were a key driver of negative comments [48,49]. As England and Wales approached so-called "freedom day" in July 2021, there was an increase in app users being told to self-isolate, coined the "pingdemic" in the media, and often reported as a shortcoming of the app rather than the direct result of a spike in infections due to the newly circulating Delta variant [50]. In response, the government decreased the sensitivity of the app [51]. This decision was criticized by the Opposition [52], and it likely reinforced the (incorrectly held) perception that the app was not working effectively.

Individuals who used the app were found to be more supportive throughout the study period. Based on our findings, it is not possible to determine whether being more supportive was a reason for installing the app or whether using the app increased support, though the fact that most people installed the app for the first time around the app's launch in October 2020 suggests that the main direction of effect is likely from support in principle to take up the app. Since individuals who were more concerned about the risk COVID-19 posed to the country were more supportive of the app [53], these individuals potentially prioritized steps to control the spread of the virus over other concerns [54]. Evidence from elsewhere indicates that individuals who were more concerned about the risk of COVID-19 were more likely to engage in protective behaviors, including app use [20,55,56].

Strengths and Limitations

The strength of this study is that it includes 8 waves of data collected over a 15-month period after the launch of the NHS COVID-19 app. At each timepoint, participants were asked about their level of support for the app, allowing us to examine changing trends in support.

The sample was representative of the general population of smartphone users in terms of age (up to 79 years), gender, social grade, and region. The sample were younger and more highly educated than the general population. This partly reflects smartphone ownership, which tends to be higher among younger individuals and those from higher-income households [57], but is also a reflection of the population that participates in nonprobability online panels [58]. One particular weakness of the sample is the underrepresentation of some minoritized groups. This is particularly important in light of evidence that some groups at higher risk of contracting COVID-19 were less likely to use the app [20].

The attrition rate was relatively high, with 61% of the original sample responding at survey wave 8. Nevertheless, we recruited new panel members ahead of survey wave 6 to top up the sample.

The question on support for the app could have been interpreted in different ways, and the results likely capture respondents' support in principle for apps to control the spread of COVID-19 as well as specific support for the app as deployed. Using mixed methods that incorporate qualitative components would enhance surveillance systems by providing deeper insights into the public's perceptions of contact tracing apps and which aspects of the app people are more or less supportive of.

Implications

Evidence demonstrates that apps had a positive impact on controlling the spread of COVID-19 [8,9]. However, uptake of the app was disappointing compared to expectations based on reports from the start of the pandemic, and by the time the app launched, only 36% of our sample were completely supportive. As with many other countries, the deployment of a contact tracing app was novel in England and Wales. As such, some of the teething problems experienced during the development of the app are unlikely to be repeated in the future. However, while most countries, including England and Wales, have ended their COVID-19 contact tracing programs, it is likely that apps will continue to be included in future pandemic preparedness plans.

The results presented in this study provide important lessons and potential strategies as to how the governments in England and Wales might increase the impact of apps in the future. First, there is a need to develop a communication strategy that considers the different phases from app development through deployment and maintenance. In the future, when relying on untested technology, there is a need to be open and transparent during the development phase about the potential risks and benefits [45], avoiding overly optimistic messaging before the technology is proven. To improve public perceptions of effectiveness, an initial step would be to publicize the growing body of evidence that apps made a positive contribution to reducing transmission in the COVID-19 pandemic and be very clear from the outset about the functionality of the app, including what data the app collects and who these are shared with [59]. Given support was strongly influenced by trust in government, messaging might be better delivered by individuals and organizations in whom the public places greater trust [45,60]. More widely, support for an app is likely to depend on the degree to which the government of the day is trusted to be doing a reasonable job in responding to any pandemic or public health emergency.

There is a need to understand the optimal timing for launching an app to increase public perceptions that apps are an effective strategy to control the spread of COVID-19 at the population level, while also benefiting individuals in safely managing their daily lives. Given the wider contextual factors in autumn 2020, the launch of the app may have been mistimed. It might have been easier for the public to understand its benefit and utility if

XSL•FO
RenderX

it had it been introduced as part of the government's strategy to exit the lockdown between March and July 2021. Future research should examine how similar apps could be adapted to improve user experience and satisfaction [48].

Conclusions

In this study among smartphone users in England and Wales, we found that the level of support for the NHS COVID-19 app was highest at launch and declined over the 15 months after rollout. A potential reason for the low and declining level of support was mistrust in both the Welsh and Westminster government's handling of the pandemic, which likely contributed to a lack of support for the technology. The attrition in levels of support observed poses important challenges for the use of apps in future pandemics. However, our findings also show that individuals who installed the app and were more concerned about the risk of COVID-19 to the country were more supportive, suggesting that there is room to build support for apps especially among those concerned about the potential harm of a pandemic to others.

Acknowledgments

The authors would like to thank the survey respondents for their time and effort. No generative AI was used in any portion of the manuscript writing.

Data Availability

The datasets generated or analyzed during this study are available on reasonable request. Requests should be directed to PIRU-pm@lshtm.ac.uk.

Funding

This research was funded by the National Institute for Health and Care Research (NIHR) Policy Research Programme through the Policy Innovation and Evaluation Research Unit (PR-PRU-1217-20602). The views expressed are those of the authors and are not necessarily those of the NIHR or the Department of Health and Social Care.

Authors' Contributions

JE contributed to data curation, formal analysis, investigation, methodology, project administration, visualization, writing—original draft, and writing—review and editing. PB contributed to data curation, formal analysis, methodology, visualization, writing—original draft, and writing—review and editing. KA contributed to data curation, formal analysis, methodology, project administration, visualization, writing—original draft, and writing—review and editing. BE contributed to conceptualization, data curation, formal analysis, methodology, writing—original draft, and writing—review and editing. NM contributed to conceptualization, data curation, formal analysis, funding acquisition, methodology, writing—original draft, and writing—review and editing. MAH contributed to conceptualization, data curation, funding acquisition, methodology, project administration, writing—original draft, and writing—review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of number of daily new cases of COVID-19 and key changes in policy measures in England and Wales (September 2020 to December 2021) [61,62].
[PNG File , 73 KB - jmir_v28i1e76863_app1.png]

Multimedia Appendix 2
Additional tables.
[DOCX File , 155 KB - jmir_v28i1e76863_app2.docx]

Multimedia Appendix 3
STROBE checklist.
[PDF File (Adobe PDF File), 199 KB - jmir_v28i1e76863_app3.pdf]

Multimedia Appendix 4
Sankey diagram of respondents' movement between subgroups over the course of 8 waves of a longitudinal survey on use and views of the National Health Service (NHS) COVID-19 app (October 2020 to December 2021), with n=1781 from wave 1 to wave 2, n=1612 from wave 2 to wave 3, n=1584 from wave 3 to wave 4, n=1509 from wave 4 to wave 5, n=1353 from wave 5 to wave 6, n=1249 from wave 6 to wave 7, and n=1124 from wave 7 to wave 8.

XSL•FO
RenderX

[PNG File , 343 KB - jmir_v28i1e76863_app4.png]

Multimedia Appendix 5
Summary of multinomial logistic regression model.
[DOCX File , 22 KB - jmir_v28i1e76863_app5.docx]

References

1. Ferretti L, Wymant C, Kendall M, Zhao L, Nurtay A, Abeler-Dörner L. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. Science 2020;368(6491):eabb6936 [FREE Full text] [doi: 10.1126/science.abb6936] [Medline: 32234805]
2. Leung KY, Metting E, Ebbers W, Veldhuijzen I, Andeweg SP, Luijben G, et al. Effectiveness of a COVID-19 contact tracing app in a simulation model with indirect and informal contact tracing. Epidemics 2024;46:100735 [FREE Full text] [doi: 10.1016/j.epidem.2023.100735] [Medline: 38128242]
3. Kretzschmar ME, Rozhnova G, Bootsma MCJ, van Boven M, van de Wijgert JHHM, Bonten MJM. Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. Lancet Public Health 2020;5(8):e452-e459 [FREE Full text] [doi: 10.1016/S2468-2667(20)30157-2] [Medline: 32682487]
4. Burdinski A, Brockmann D, Maier BF. Understanding the impact of digital contact tracing during the COVID-19 pandemic. PLOS Digit Health 2022;1(12):e0000149 [FREE Full text] [doi: 10.1371/journal.pdig.0000149] [Medline: 36812611]
5. Unim B, Zile-Velika I, Pavlovska Z, Lapao L, Peyroteo M, Misins J. The role of digital tools and emerging devices in COVID-19 contact tracing during the first 18 months of the pandemic: a systematic review. Eur J Public Health 2024;34(Supplement_1):i11-i28. [doi: 10.1093/eurpub/ckae039] [Medline: 38946444]
6. Elder AS, Arrouzet CJ, Miljacic L, Karras BT, Higgins A, West LM. Evaluation of the effectiveness of Washington State's digital COVID-19 exposure notification system over one pandemic year. Front Public Health 2024;12:1408178 [FREE Full text] [doi: 10.3389/fpubh.2024.1408178] [Medline: 39206001]
7. Daniore P, Nittas V, Ballouz T, Menges D, Moser A, Höglinger M. Performance of the Swiss digital contact-tracing app over various SARS-CoV-2 pandemic waves: repeated cross-sectional analyses. JMIR Public Health Surveill 2022;8(11):e41004 [FREE Full text] [doi: 10.2196/41004] [Medline: 36219833]
8. Kendall M, Tsallis D, Wymant C, Di Francia A, Balogun Y, Didelot X. Epidemiological impacts of the NHS COVID-19 app in England and Wales throughout its first year. Nat Commun 2023;14(1):858 [FREE Full text] [doi: 10.1038/s41467-023-36495-z] [Medline: 36813770]
9. Wymant C, Ferretti L, Tsallis D, Charalambides M, Abeler-Dörner L, Bonsall D. The epidemiological impact of the NHS COVID-19 app. Nature 2021;594(7863):408-412. [doi: 10.1038/s41586-021-03606-z] [Medline: 33979832]
10. Hinch R, Probert W, Nurtay A, Kendall M. Effective Configurations of a Digital Contact Tracing App: A report to NHSX. 2020. URL: https://cdn.theconversation.com/static_files/files/1009/Report_-_Effective_App_Configurations.pdf [accessed 2025-11-15]
11. Yasaka TM, Lehrich BM, Sahyouni R. Peer-to-peer contact tracing: development of a privacy-preserving smartphone app. JMIR Mhealth Uhealth 2020;8(4):e18936 [FREE Full text] [doi: 10.2196/18936] [Medline: 32240973]
12. Elmokashfi A, Sundnes J, Kvalbein A, Naumova V, Reinemo S, Florvaag PM. Nationwide rollout reveals efficacy of epidemic control through digital contact tracing. Nat Commun 2021;12(1):5918 [FREE Full text] [doi: 10.1038/s41467-021-26144-8] [Medline: 34635661]
13. von Wyl V, Höglinger M, Sieber C, Kaufmann M, Moser A, Serra-Burriel M. Drivers of acceptance of COVID-19 proximity tracing apps in Switzerland: panel survey analysis. JMIR Public Health Surveill 2021;7(1):e25701 [FREE Full text] [doi: 10.2196/25701] [Medline: 33326411]
14. Horvath L, Banducci S, Blamire J, Degnen C, James O, Jones A. Adoption and continued use of mobile contact tracing technology: multilevel explanations from a three-wave panel survey and linked data. BMJ Open 2022;12(1):e053327 [FREE Full text] [doi: 10.1136/bmjopen-2021-053327] [Medline: 35039293]
15. Grill E, Eitze S, De Bock F, Dragano N, Huebl L, Schmich P. Sociodemographic characteristics determine download and use of a corona contact tracing app in Germany-results of the COSMO surveys. PLoS One 2021;16(9):e0256660 [FREE Full text] [doi: 10.1371/journal.pone.0256660] [Medline: 34473733]
16. Johnson B. Nearly 40% of Icelanders are using a covid app—and it hasn't helped much. MIT Technology Review. 2020. URL: https://www.technologyreview.com/2020/05/11/1001541/iceland-rakning-c19-covid-contact-tracing/ [accessed 2025-11-15]
17. Chen AT, Thio KW. Exploring the drivers and barriers to uptake for digital contact tracing. Soc Sci Humanit Open 2021;4(1):100212 [FREE Full text] [doi: 10.1016/j.ssaho.2021.100212] [Medline: 34642660]
18. Guillon M, Kergall P. Attitudes and opinions on quarantine and support for a contact-tracing application in France during the COVID-19 outbreak. Public Health 2020;188:21-31 [FREE Full text] [doi: 10.1016/j.puhe.2020.08.026] [Medline: 33059232]

XSL•FO
RenderX

19. Altmann S, Milsom L, Zillessen H, Blasone R, Gerdon F, Bach R. Acceptability of app-based contact tracing for COVID-19: cross-country survey study. JMIR Mhealth Uhealth 2020;8(8):e19857 [FREE Full text] [doi: 10.2196/19857] [Medline: 32759102]
20. Al-Haboubi M, Exley J, Allel K, Erens B, Mays N. One year of digital contact tracing: Who was more likely to install the NHS COVID-19 app? Results from a tracker survey in England and Wales. Digit Health 2023;9:20552076231159449. [doi: 10.1177/20552076231159449]
21. Marani M, Katul GG, Pan WK, Parolari AJ. Intensity and frequency of extreme novel epidemics. Proc Natl Acad Sci U S A 2021;118(35):e2105482118 [FREE Full text] [doi: 10.1073/pnas.2105482118] [Medline: 34426498]
22. NHS COVID-19 app statistics. NHS Test and Trace. 2022. URL: https://stats.app.covid19.nhs.uk/ [accessed 2022-08-09]
23. NHS COVID-19 app support: what the app does. NHS Test and Trace. 2022. URL: https://www.covid19.nhs.uk/what-the-app-does.html [accessed 2022-03-07]
24. Coronavirus timeline: Welsh and UK governments' response. Welsh Parliament. 2021. URL: https://research.senedd.wales/research-articles/coronavirus-timeline-welsh-and-uk-governments-response/ [accessed 2022-03-20]
25. Methodology. YouGov. 2025. URL: https://yougov.co.uk/about/panel-methodology [accessed 2025-10-10]
26. Goodman LA. Latent class analysis: the empirical study of latent types, latent variables, and latent structures. In: Hagenaars JA, McCutcheon AL, editors. Applied Latent Class Analysis. Cambridge: Cambridge University Press; 2002:3-55.
27. Bartus T. Multilevel multiprocess modeling with Gsem. Stata J 2017;17(2):442-461. [doi: 10.1177/1536867x1701700211]
28. Boxall PC, Adamowicz WL. Understanding heterogeneous preferences in random utility models: a latent class approach. Environ Resour Econ 2002;23(4):421-446. [doi: 10.1023/a:1021351721619]
29. Swait J. A structural equation model of latent segmentation and product choice for cross-sectional revealed preference choice data. J Retail Consum Serv 1994;1(2):77-89. [doi: 10.1016/0969-6989(94)90002-7]
30. Jann B. Color palettes for Stata graphics: an update. Stata J 2023;23(2):336-385. [doi: 10.1177/1536867x231175264]
31. Schonlau M. The clustergram: A graph for visualizing hierarchical and nonhierarchical cluster analyses. Stata J 2002;2(4):391-402. [doi: 10.1177/1536867x0200200405]
32. Cuschieri S. The STROBE guidelines. Saudi J Anaesth 2019;13(5):31. [doi: 10.4103/sja.sja_543_18]
33. English indices of deprivation 2019. Ministry of Housing. 2019. URL: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019 [accessed 2025-11-15]
34. NHS COVID-19 app launches across England and Wales. Department of Health and Social Care. 2020. URL: https://www.gov.uk/government/news/nhs-covid-19-app-launches-across-england-and-wales [accessed 2025-11-15]
35. Samuel G, Sims R. The UK COVID-19 contact tracing app as both an emerging technology and public health intervention: the need to consider promissory discourses. Health (London) 2023;27(4):625-644 [FREE Full text] [doi: 10.1177/13634593211060768] [Medline: 34812092]
36. Kelion L. UK Virus-Tracing App Switches to Apple-Google Model. 2020. URL: https://www.bbc.co.uk/news/technology-53095336 [accessed 2025-11-15]
37. Amann J, Sleigh J, Vayena E. Digital contact-tracing during the Covid-19 pandemic: an analysis of newspaper coverage in Germany, Austria, and Switzerland. PLoS One 2021;16(2):e0246524 [FREE Full text] [doi: 10.1371/journal.pone.0246524] [Medline: 33534839]
38. Lueks W, Benzler J, Bogdanov D, Kirchner G, Lucas R, Oliveira R. Toward a common performance and effectiveness terminology for digital proximity tracing applications. Front Digit Health 2021;3:677929 [FREE Full text] [doi: 10.3389/fdgth.2021.677929] [Medline: 34713149]
39. Samuel G, Lucivero F. Framing ethical issues associated with the UK COVID-19 contact tracing app: exceptionalising and narrowing the public ethics debate. Ethics Inf Technol 2022;24(1):5 [FREE Full text] [doi: 10.1007/s10676-022-09628-z] [Medline: 35110970]
40. Davies B, Lalot F, Peitz L, Heering MS, Ozkececi H, Babaian J. Changes in political trust in Britain during the COVID-19 pandemic in 2020: integrated public opinion evidence and implications. Humanit Soc Sci Commun 2021;8(1):166. [doi: 10.1057/s41599-021-00850-6]
41. Boehm C, Boadu P, Exley J, Al-Haboubi M, Mays N. Public trust in the government to control the spread of COVID-19 in England after the first wave-a longitudinal analysis. Eur J Public Health 2023;33(6):1155-1162 [FREE Full text] [doi: 10.1093/eurpub/ckad148] [Medline: 37579239]
42. Timeline of UK government coronavirus lockdowns and restrictions. Institute for Government. 2022. URL: https://www.instituteforgovernment.org.uk/charts/uk-government-coronavirus-lockdowns [accessed 2025-11-15]
43. Haddon C, Sasse T, Nice A. Science advice in a crisis. Institute of Government. 2020. URL: https://www.instituteforgovernment.org.uk/publication/report/science-advice-crisis [accessed 2025-11-15]
44. Coronavirus: lessons learned to date report published. House of Commons Health and Social Care. 2021. URL: https://committees.parliament.uk/committee/81/health-and-social-care-committee/news/157991/coronavirus-lessons-learned-to-date-report-published [accessed 2025-11-15]
45. Samuel G, Lucivero F, Johnson S, Diedericks H. Ecologies of public trust: the NHS COVID-19 contact tracing app. J Bioeth Inq 2021;18(4):595-608 [FREE Full text] [doi: 10.1007/s11673-021-10127-x] [Medline: 34609676]

XSL•FO
RenderX

46. Pepper C, Reyes-Cruz G, Pena AR, Dowthwaite L, Babbage CM, Wagner H. Understanding trust and changes in use after a year with the NHS covid-19 contact tracing app in the united kingdom: longitudinal mixed methods study. J Med Internet Res 2022;24(10):e40558 [FREE Full text] [doi: 10.2196/40558] [Medline: 36112732]
47. Cellan-Jones R. Disappearing Covid-19 App Alerts Cause Alarm. 2020. URL: https://www.bbc.co.uk/news/technology-54389083 [accessed 2022-07-08]
48. Garousi V, Cutting D. What do users think of the UK's three COVID-19 contact-tracing apps? A comparative analysis. BMJ Health Care Inform 2021;28(1):E100320 [FREE Full text] [doi: 10.1136/bmjhci-2021-100320] [Medline: 34281994]
49. Dixon EL, Joshi SM, Ferrell W, Volpp KG, Merchant RM, Guntuku SC. COVID-19 contact tracing app reviews reveal concerns and motivations around adoption. PLoS One 2022;17(9):e0273222 [FREE Full text] [doi: 10.1371/journal.pone.0273222] [Medline: 36084078]
50. Callaway E. Delta coronavirus variant: scientists brace for impact. Nature 2021;595(7865):17-18. [doi: 10.1038/d41586-021-01696-3] [Medline: 34158664]
51. Davies C, Stewart H. NHS Covid App to Be Tweaked to Cut Need for Self-Isolation. 2021. URL: https://www.theguardian.com/world/2021/jul/08/sunak-hints-at-changes-to-test-and-trace-app-to-cut-numbers-told-to-isolate [accessed 2025-11-15]
52. Stewart H. Keir Starmer: tweaking NHS Covid app ‘like taking batteries out of smoke alarm’. 2021. URL: https://www.theguardian.com/world/2021/jul/09/keir-starmer-tweaking-nhs-covid-app-taking-batteries-smoke-alarm [accessed 2025-11-15]
53. Dowthwaite L, Fischer J, Perez Vallejos E, Portillo V, Nichele E, Goulden M. Public adoption of and trust in the NHS COVID-19 contact tracing app in the United Kingdom: quantitative online survey study. J Med Internet Res 2021;23(9):e29085 [FREE Full text] [doi: 10.2196/29085] [Medline: 34406960]
54. Williams SN, Armitage CJ, Tampe T, Dienes K. Public attitudes towards COVID-19 contact tracing apps: a UK-based focus group study. Health Expect 2021;24(2):377-385 [FREE Full text] [doi: 10.1111/hex.13179] [Medline: 33434404]
55. Smith LE, Potts HWW, Aml t R, Fear NT, Michie S, Rubin GJ. Patterns of social mixing in England changed in line with restrictions during the COVID-19 pandemic (September 2020 to April 2022). Sci Rep 2022;12(1):10436 [FREE Full text] [doi: 10.1038/s41598-022-14431-3] [Medline: 35729196]
56. Zabel S, Schlaile MP, Otto S. Breaking the chain with individual gain? Investigating the moral intensity of COVID-19 digital contact tracing. Comput Human Behav 2023;143:107699 [FREE Full text] [doi: 10.1016/j.chb.2023.107699] [Medline: 36818428]
57. OFCOM Nations & Regions Technology Tracker - 2020, 9th January to 7th March 2020. 2020. URL: https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/data/statistics/2020/technology-tracker-2020/technology-tracker-2020-uk-data-tables?v=324794 [accessed 2025-11-15]
58. Erens B, Burkill S, Couper MP, Conrad F, Clifton S, Tanton C. Nonprobability web surveys to measure sexual behaviors and attitudes in the general population: a comparison with a probability sample interview survey. J Med Internet Res 2014;16(12):e276 [FREE Full text] [doi: 10.2196/jmir.3382] [Medline: 25488851]
59. Sujarwoto S, Maharani A. Facilitators and barriers to the adoption of mHealth apps for COVID-19 contact tracing: a systematic review of the literature. Front Public Health 2023;11:1222600 [FREE Full text] [doi: 10.3389/fpubh.2023.1222600] [Medline: 38145061]
60. Liccardi I, Alekseyev J, Woltz VLA, McLean JE, Zurko ME. Public willingness to engage with COVID-19 contact tracing, quarantine, and exposure notification. Public Health Rep 2022;137(2_suppl):90S-95S [FREE Full text] [doi: 10.1177/00333549221125891] [Medline: 36255241]
61. UKHSA data dashboard. URL: https://coronavirus.data.gov.uk/details/cases [accessed 2025-11-20]
62. Timeline of UK government coronavirus lockdowns and restrictions. URL: https://www.instituteforgovernment.org.uk/data-visualisation/timeline-coronavirus-lockdowns [accessed 2025-11-20]

Abbreviations

AIC: Akaike information criteria
BIC: Bayesian information criteria
LCA: latent class analysis
NHS: National Health Service
RRR: relative risk ratio
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
VIF: variance inflation factors

XSL•FO
RenderX

Edited by A Mavragani, T de Azevedo Cardoso; submitted 02.May.2025; peer-reviewed by J Igboanugo, V Surasani; comments to author 10.Jun.2025; revised version received 16.Oct.2025; accepted 16.Oct.2025; published 28.Jan.2026.

Please cite as:
Exley J, Boadu P, Allel K, Erens B, Mays N, Al-Haboubi M
Characteristics Influencing Support for the National Health Service COVID-19 App in England and Wales: Findings From a Longitudinal Survey
J Med Internet Res 2026;28:e76863
URL: https://www.jmir.org/2026/1/e76863
doi:10.2196/76863
PMID:

©Josephine Exley, Paul Boadu, Kasim Allel, Bob Erens, Nicholas Mays, Mustafa Al-Haboubi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patterns and Characteristics of Mobile App Use to Promote Wellness and Manage Illness: Cross-Sectional Study

Hayriye Gulec[1,2], PhD; David Smahel[1], PhD; Yi Huang[1,3], PhD

[1]Interdisciplinary Research Team on Internet and Society, Faculty of Social Studies, Masaryk University, Brno, Czech Republic

[2]Department of Psychology, Faculty of Arts and Sciences, Bursa Uludag University, Bursa, Turkey

[3]Centre for Research and Development, Pan-European University, Bratislava, Slovakia

Corresponding Author:
Hayriye Gulec, PhD
Interdisciplinary Research Team on Internet and Society
Faculty of Social Studies
Masaryk University
Joštova 10
Brno, 620 00
Czech Republic
Phone: 420 549 49 4180
Email: hayriyegulec@uludag.edu.tr

Abstract

Background: Mobile health (mHealth) apps target diverse health behaviors, but engagement may vary by purpose.

Objective: This study examined the prevalence, usage patterns, and user characteristics of mHealth apps among Czech adults with internet access, focusing on sociodemographics, digital knowledge and use, and health indicators predicting wellness- and illness-related app use.

Methods: Overall, 4775 Czech adults (2365/4775, 49.53% women) aged 18-95 (mean 45.37, SD 16.40) years completed an online survey. Sociodemographic factors included age, gender, education, and income. Digital knowledge and use were measured using the eHealth Literacy Scale and the passive/active use of social networking sites (SNS) for health information. Health indicators covered symptom severity, physical activity, BMI, and eating disorder–related risk propensity (body dissatisfaction, dietary restraint, and weight/shape overvaluation). Participants reported app use for sports, number of steps, nutrition, vitals, sleep, diagnosed conditions, reproductive health, diagnosis assistance, mood and mental well-being, and emergency care guidance. Multivariate hierarchical binary logistic regression analysis identified characteristics of app users. Exploratory structural equation modeling (ESEM) clustered apps into “promoting wellness” and “managing illness” and examined the predictors of frequency of use.

Results: Of 4440 respondents, 2172 (48.92%) used mHealth apps. Users were younger (odds ratio [OR] 0.98, 95% CI 0.98-0.99, $P<.001$), had a monthly income more than 50,000 CZK (1 CZK=US $0.048; vs ≤20,000 CZK: OR 0.54, 95% CI 0.41-0.7, $P<.001$; 20,001-35,000 CZK: OR 0.78, 95% CI 0.65-0.93, $P=.006$; 35,001-50,000 CZK: OR 0.83, 95% CI 0.7-0.99, $P=.03$), were more eHealth literate (OR 1.17, 95% CI 1.06-1.3, $P=.003$), used SNS passively for health information (OR 1.35, 95% CI 1.21-1.51, $P<.001$), and had higher eating disorder risk (OR 1.18, 95% CI 1.12-1.25, $P<.001$) and physical activity (OR 1.18, 95% CI 1.13-1.23, $P<.001$) than nonusers. Step-counting apps were most common; 65.99% (1430/2167) used them daily or several times a day, followed by apps for sleep (691/2163, 31.95%), vitals (611/2165, 28.22%), and sports (407/2158, 18.86%). ESEM confirmed a 2-factor structure (“promoting wellness” and “managing illness”; $\chi^2_{26}=71.9$, comparative fit index=0.99, Tucker-Lewis index=0.99, root-mean-square error of approximation=0.03, and standardized root-mean-square residual=0.03). Frequent use of wellness apps was associated with younger age (standardized $\beta=-0.22$, $P<.001$), higher eHealth literacy (standardized $\beta=0.10$, $P<.001$), and physical activity (standardized $\beta=0.15$, $P<.001$). Illness-management app use was associated with active use of SNS for health information (standardized $\beta=0.62$, $P<.001$) and eating disorder risk (standardized $\beta=0.11$, $P<.001$). Digital knowledge, digital use, and health indicators mediated the association between age and mHealth app use.

Conclusions: mHealth app engagement reflects broader social, digital, and psychological inequalities rather than individual preferences alone. Encouraging digital inclusion and addressing body image- and diet-related use may help ensure that mHealth technologies do not exacerbate existing health inequalities across age and user groups.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e71363) doi:10.2196/71363

KEYWORDS

mobile health; mHealth; eHealth literacy; eating disorder; ESEM; exploratory structural equation modeling; SNS; social networking site

Introduction

Background

The rapid advancements in mobile health (mHealth) app technologies and the continued improvement to content have resulted in the proliferation of health-related areas addressed via mHealth apps and wearables in the last decade [1-3]. Now, smartphone users can monitor and regulate daily routines (eg, number of steps and circadian rhythms) and autonomic indicators (eg, heart rate, blood pressure, body temperature, and respiratory rate). They can use apps to modify lifestyle behaviors (eg, physical activity, nutritional intake, and weight management), improve mental health and well-being (eg, meditation and mindfulness), get health assistance (eg, diagnosis assistance, reproductive health, and emergency care guidance), and manage diagnosed conditions (eg, diabetes and depression).

Two users who download the same app to track and regulate a particular health behavior (eg, nutritional intake) may use it for separate purposes: one to maintain a healthy lifestyle and the other to manage a chronic health condition (eg, diabetes and obesity). The characteristics of users might differ depending on their purpose of use. Although previous literature has linked mHealth app usage to user intentions related to self-management of wellness and illness [4-7], empirical studies have primarily focused on app adoption in general, overlooking user characteristics associated with the purpose of use [8]. Additionally, only a few studies have examined the prevalence and characteristics of mHealth app usage in large-scale, representative samples [9-12].

Therefore, the aim of this study is 2-fold. First, we examine the prevalence and characteristics of app users within a representative sample of Czech adults with internet access. We focus on sociodemographics, digital knowledge and use, and health indicators as user characteristics. Second, we cluster mHealth apps using a data-driven approach under the “promoting wellness” and “managing illness” dimensions. Finally, we determine the user characteristics that predict the app usage frequency for each dimension. This study reports on a large representative adult sample concerning the prevalence, usage patterns, and user characteristics of various types of apps. It also provides initial evidence for user characteristics as predictors of wellness- and illness-related app usage.

Prior large-scale studies examining adult mHealth app users were primarily conducted in the United States using different waves of data from the National Cancer Institute’s Health Information National Trends Survey. The prevalence of mHealth app use among smartphone owners ranged between 34% and 56%, and they consistently indicated younger age and higher educational status as significant predictors [9-12]. Females were more likely to own and use mHealth apps in general [10-12], although no gender difference was reported in one study [9]. A recent scoping review identified age, gender, educational level, and income status as significant factors [8]. This study examines the prevalence and sociodemographic characteristics of mHealth app use among a representative sample of Czech adults with internet access. It also reveals whether these characteristics differ when using apps to promote wellness and manage illness.

eHealth literacy is the perceived knowledge and skills to obtain, understand, and evaluate digital health information [13]. Previous studies reported eHealth literacy as a significant predictor of mHealth app use in general adult samples [14,15] and individuals with chronic health conditions [16,17]. In addition, higher eHealth literacy was associated with the perception of mHealth apps as effective and their frequent usage for health-related behavioral change [18]. This study provides initial insights into whether app usage to promote wellness and manage illness differs as a function of eHealth literacy.

Digital platforms, including social networking sites (SNS; eg, Facebook, Instagram, and X, formerly known as Twitter), are frequently used for health-related purposes [19,20]. Individuals who use social media for health information are more likely to seek health information through mHealth apps, indicating that both media might be used complementarily [21]. Moreover, prior research has shown that health messages delivered via SNS can shape users’ health-related decisions and behaviors [21,22]. Therefore, a positive relationship between SNS activity and the use of mobile apps for health purposes is plausible. This study aims to provide initial evidence on whether health-related SNS use is linked to the adoption of health apps and the frequency of their use for promoting wellness and managing illness.

Prior studies have examined general health status as a predictor of mHealth app use; however, these studies have assessed health status using a single item, yielding mixed findings [9-12]. This study adopts a more nuanced approach to address this limitation, assessing perceived somatic-symptom severity across multiple domains, including gastrointestinal, cardiopulmonary, musculoskeletal, and sleep/energy areas. In parallel, an active lifestyle—as reflected in physical activity levels—has received growing attention in mHealth research, with evidence suggesting that app use may support health promotion among physically active individuals [10,18,23]. By jointly examining both physical activity and detailed health status indicators, this study contributes to understanding how these factors shape the adoption and use of apps for promoting wellness and managing illness.

Lastly, only a few studies examined BMI and eating disorder–related risk factors as correlates of mHealth app use. This study focused on BMI because it is a significant predictor of health outcomes, including diabetes and obesity [24]. Evidence suggests that individuals with a higher BMI are more likely to use mHealth apps [10,18,25] and to use them for a

XSL•FO
RenderX

health behavior goal [9,26]. A higher BMI is also a significant predictor of eating disorder symptoms, including body dissatisfaction, dietary restraint, and weight and shape overvaluation [27]. Previous research has reported a connection between using mHealth apps to track diet and physical activity and their associations with eating disorder–related risks and symptoms, including BMI [28-30]. These findings underscore the importance of evaluating core risk factors for eating disorders together with BMI to understand how they might play roles in using apps that address behaviors related to diet and physical exercise (ie, health promotion) but also those related to monitoring health (eg, vitals) and managing disease (eg, obesity and diabetes). In this study, we evaluate body dissatisfaction, dietary restraint, and weight and shape overvaluation to determine participants' risk propensity for an eating disorder.

This Study

This study investigated the prevalence and frequency of mHealth app use in a representative sample of Czech adults with internet access. Our first objective was to identify user characteristics that predict app adoption and usage. We examined a comprehensive set of predictors, including sociodemographic variables (age, gender, education, and income), digital knowledge and use (eHealth literacy and SNS use for health information), and health indicators (somatic symptom severity, physical activity, BMI, and eating disorder risk propensity). This multidimensional approach enabled us to move beyond basic demographic correlates and gain a deeper understanding of the factors influencing engagement with mHealth apps. Our second objective was to cluster mHealth apps into two key dimensions—promoting wellness and managing illness—as suggested by the prior literature [4-7]. Rather than relying solely on predefined app classifications, we used a data-driven approach to cluster mHealth app usage into categories related to wellness and illness. Furthermore, we determined user characteristics associated with the frequency of use within each category. This approach provided a more nuanced understanding of user characteristics connected with using wellness and illness app types.

During the analyses, we observed that the effect size of age on app use was consistently reduced after entering the variables related to digital knowledge, digital use, and health indicators. This finding prompted us to conduct additional exploratory mediation analyses to test whether these factors mediated the association between age and app use. These exploratory analyses provide insight into potential mechanisms underlying age-related differences in mHealth engagement, informing future strategies to reduce disparities in app adoption and use for wellness promotion and illness management.

Methods

Recruitment

This study used cross-sectional data collected through an online survey targeting adult Czech internet users. The study design adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for observational cross-sectional research. Eligible participants were adults aged 18 years or older who reported using the internet. Data collection was conducted by STEM/MARK, a Czech market research agency that manages the Czech National Panel (part of National Sample s.r.o.). STEM/MARK is a member of the European Society for Opinion and Market Research and follows its data protection and panel management standards. Data collection occurred between October 2 and 16, 2023, as part of the NPO "Systemic Risk Institute" project.

The Czech National Panel includes 64,000 individuals and is designed to capture a broad spectrum of the population, including harder-to-reach subgroups. Eligible participants were selected from an initial pool of 58,000 participants. A quota-sampling strategy was applied with respect to age, educational attainment, household income, municipality size, and administrative region, following the Nomenclature of Territorial Units for Statistics (NUTS3) classification, as per Eurostat data. The gender distribution was balanced, with a tolerance of ±7% based on the registered gender within the panel. Our recruitment goal was to maximize sample size within the project's funding constraints, with a minimum target of 3500 respondents to ensure meaningful representation even from the smallest regions.

Participants were recruited through online and face-to-face invitations, and the questionnaire was delivered via computer-assisted web interviewing. The survey was opened by 5480 panelists. Of these, 5135 filled in the items to verify the respondent's suitability for data collection (eg, quotas). Due to noncompliance with research requirements or failure to fulfill quotas, 214 respondents were rejected. The number of panelists who completed the questionnaire was 4921 individuals. An additional 146 questionnaires were discarded due to excessive missing data (>10%), unrealistically fast completion times, or logically inconsistent answers. The final sample consisted of 4775 adult Czech internet users (2365/4775, 49.53% women) aged 18-95 years (mean 45.37, SD 16.40 years).

All survey instruments underwent cognitive pretesting before data collection. Feedback was obtained from 5 individuals representing diverse genders, ages, and educational backgrounds (aged 26-73 years). Their input focused on evaluating the clarity, comprehensibility, and overall content of the questionnaire.

Measures

Sociodemographic Characteristics

Participants reported their gender and responded to an open-ended question about their age. They indicated their highest education with the following response options: (1) no education or primary school, (2) secondary vocational school, (3) secondary school, and (4) university (or higher vocational school). The income status was assessed by asking, "In which category would you put the average monthly income of your household for all its members? Please consider the income from employment, part-time jobs, business, and social security." The response options were (1) up to 20,000 CZK, (2) 20,001-35,000 CZK, (3) 35,001-50,000 CZK, and (4) 50,001 CZK and more. [A currency exchange rate of CZK 1=US $0.048 is applicable.] Higher scores were indicative of better financial status.

XSL•FO
RenderX

Digital Knowledge and Use Factors

eHealth Literacy

eHealth literacy was assessed by the eHealth Literacy Scale [31]. The items inquired about knowledge of online health information sources (eg, I know what health resources are available on the internet), how to navigate the internet to obtain answers to health-related questions (eg, I know where to find helpful health resources on the internet), the perceived skills to evaluate the quality of online health information (eg, I can tell high-quality health resources from low-quality health resources on the internet), and ability to apply health information for health purposes (eg, I know how to use the health information I find on the internet to help me). The response options ranged from 1 (strongly disagree) to 5 (strongly agree). Higher scores indicated better eHealth literacy skills. The internal consistency was adequate (Cronbach α=0.89).

Health-Related SNS Use

The health-related use of SNS was assessed using items developed in a previous study to measure both active and passive social media use [32]. We adapted the items to tap into passive and active use of SNS for health information. Passive health-related SNS use refers to the passive engagement with health information. Participants responded to how often they "watched videos or pictures posted about health or illness," "followed health- or illness-related pages," and "read health- or illness-related discussions, user comments, ratings, or reviews" on SNS. Active health-related SNS use items evaluated active engagement with the health information and asked how often participants "posted health- or illness-related content," "chatted or interacted with others about health or illness," and "commented on, liked, or shared others' health- or illness-related posts" on SNS. The response options were (1) never, (2) a few times at most, (3) several times a month, (4) several times a week, (5) every day, and (6) several times a day. The scores were calculated separately for passive and active use; the Cronbach alphas were 0.9 for both scales.

Health Indicators

Symptom Severity

The perceived severity of symptoms was assessed using items from the Patient Health Questionnaire-15, which evaluates 15 somatic symptoms encountered in outpatient settings [33]. It is a widely used brief screening instrument with good psychometric properties for assessing somatic symptoms and screening for somatization [34]. The original scale asks participants to rate how much they were bothered by each symptom in the preceding month. For brevity, we grouped the symptoms under gastrointestinal (ie, stomach pain, constipation, loose bowels, diarrhea, nausea, gas, and indigestion), cardiopulmonary (ie, chest pain, feeling your heart pound or race, and shortness of breath), musculoskeletal (ie, back pain, pain in your arms or legs, and pain in joints like knees and hips), and sleep/energy (ie, headaches, dizziness, fainting spells, feeling tired, having low energy, and trouble sleeping) areas. For each symptom area, the participants evaluated the extent to which they were bothered by any of the symptoms in the preceding 4 weeks. The items were rated on a 5-point Likert scale that ranged from 1 (never) to 5 (always). Higher scores were indicative of more severe symptoms. The internal consistency of the scale was satisfactory (Cronbach α=0.72).

Physical Activity

Participants responded to the question—"How many hours a week do you usually exercise to the extent that you sweat and feel shortness of breath?"—to evaluate their physical activity. The response options included (1) less than half an hour per week, (2) about half an hour per week, (3) about 1 hour per week, (4) about 2-3 hours per week, (5) about 4-6 hours per week, and (6) about 7 hours or more per week.

BMI

Participants responded to open-ended questions about their height (in centimeters) and weight (in kilograms). The BMI was calculated by weight (kg)/height (m^2)×10,000 formula.

Eating Disorder–Related Risk Propensity

Eating disorder–related risk propensity was assessed with the modified 7-item version [35] of the Eating Disorder Examination Questionnaire [36], a widely used measure of eating pathology. It consists of 3 factors that evaluate dietary restraint, shape/weight overvaluation, and body dissatisfaction. Body dissatisfaction (ie, dissatisfaction with weight and shape) and weight/shape overvaluation (ie, the influence of weight and shape on self-worth) are concerned with the cognitive-evaluative aspects of body image. Dietary restraint assesses the frequency of rigid behaviors to influence weight and shape. The scale is rated with a 7-point forced-choice format that considers the preceding 28 days. Higher scores indicate a greater frequency of dietary restraint (ie, from zero to every day) and a greater severity of body dissatisfaction and weight/shape overvaluation (ie, from "not at all" to "extremely"). The internal consistencies of the factors ranged from 0.89 to 0.91, and the factor loadings and item intercepts were invariant for sex and overweight status in a previous study with adults [35]. The scale's internal consistency was satisfactory in this study (Cronbach α=0.87).

mHealth App Use

App use was determined by asking, "You can use various applications on your phone, tablet, or other mobile devices, like smartwatches. Have you used applications to monitor health and exercise (eg, counting steps, tracking calories, weight, sports activities, eating/drinking, stress, or sleep) in the last year?" The response options were (1) no and (2) yes. mHealth app users responded to an additional question about the frequency of using different types of apps, which included sports (eg, workouts, exercise, running, and strengthening), number of steps, nutrition (eg, calorie intake or expenditure, weight, and nutritional facts), vitals (eg, blood pressure, heart rate, body temperature, and respiratory rate), sleep (eg, sleep patterns), mood and mental well-being (eg, mindfulness, meditation, mental health, and self-esteem), support or control of a diagnosed health condition (eg, diabetes management and depression management), reproductive health (eg, pregnancy, ovulation, menstruation, and sexual health), diagnosis assistance (eg, symptom checking), and emergency care guidance. The response options included (1) never, (2) a few times at most, (3) several times a month, (4) several times a week, (5) every

XSL•FO
RenderX

day, and (6) several times a day. The scale's internal consistency was satisfactory (Cronbach α=0.82).

Statistical Analysis

Descriptive statistics were run for the sociodemographic characteristics, digital knowledge and use factors, and health indicators. They included means, SDs, and frequencies for the entire sample, app users, and app nonusers. Participants who reported using an mHealth app in the previous year were identified as app users. A multivariate hierarchical binary logistic regression analysis was conducted to examine the characteristics of app users. Sociodemographic factors (age, gender, education, and income) were included in the first step. The second step added the digital knowledge and use factors (eHealth literacy, active and passive use of SNS) and health indicators (perceived symptom severity, physical activity, BMI, and eating disorder–related risk propensity).

The remaining analyses were based on app users' responses regarding their app use patterns. First, we presented usage frequencies for each app type. Then, we examined the factor structure of app types based on usage frequencies. Guided by the mHealth app categories suggested in the previous literature [4-7], we assumed that mHealth app usage could be divided into two dimensions: mHealth app usage for "promoting wellness" and "managing illness." We initially ran the exploratory factor analysis (EFA), and based on the factor loading patterns (see the "mHealth_EFA" worksheet in Multimedia Appendix 1), we found that specific mHealth apps (related to nutrition and mental health) appeared to load on two dimensions. Furthermore, the strictly constrained confirmatory factor analysis (CFA) model (see the "mHealth_CFA" worksheet in Multimedia Appendix 1), where, psychometrically speaking, each item is specified to load exclusively on a single factor, yielded an unsatisfactory model fit (comparative fit index=0.86, Tucker-Lewis index=0.81, and root-mean-square error of approximation=0.13). Thus, we subsequently decided to apply the exploratory structural equation modeling (ESEM) method to test this hypothesis. The ESEM used usage frequencies of each app type to determine the underlying factors of use. Although ESEM operates within the traditional CFA framework, it notably allows for the free estimation of cross-loadings while still displaying targeted and constrained factors [37]. This less restrictive approach provides an alternative when CFA does not achieve an adequate model fit. The flexibility of ESEM is advantageous because it more effectively captures the multidimensional aspects of a measure [37,38]. In this study, the ESEM enabled us to assess the model's fit for our targeted factors (ie, promoting wellness and managing illness). In addition, it could capture the multidimensional aspects of app use by facilitating the cross-loading of app types for both dimensions.

The ESEM approach was also used to predict user characteristics associated with mHealth app usage for our targeted factors (ie, wellness and illness). The dependent variables in the first ESEM model were the frequency of mHealth app usage for promoting wellness and the frequency of mHealth app usage for managing illness. The independent variables were sociodemographics, including age, gender, educational level, and income status. In the second model, we added the predictors of digital knowledge and use factors (eHealth literacy, active and passive use of SNS) and health indicators (symptom severity, physical activity, BMI, eating disorder–related risk propensity).

Lastly, we conducted two separate mediation analyses to test whether digital knowledge, digital use, and health indicators mediated the association between age and app usage. The first mediation analysis examined the relationship between age and wellness app use frequency, and the second examined the relationship between age and illness app use frequency. The analyses were conducted using SPSS software (version 28.0, IBM Corp), the R package "lavaan" (The R Foundation), and the open-source software Jamovi (version 2.6), which provides an R-based graphical user interface. Within Jamovi, the analyses used the R packages "psych," "lavaan," and "lm." Mahalanobis distances were computed to identify multiple outliers, and data from 1.11% (53/4775) of the participants with significant Mahalanobis distance values at P<.001 were deleted.

Ethical Considerations

This study was reviewed and approved by the Research Ethics Committee of Masaryk University (EKV-2023-102). The study complied with ethical standards for human subjects research. Before completing the questionnaires, participants were informed about the purpose of the survey, its anonymity, and their right to refuse participation or skip any question by selecting "I don't know" or "I prefer not to say." Written informed consent was obtained before participation. All data were collected and analyzed anonymously, with no personally identifiable information retained. Respondents received monetary compensation for participation, determined by the panel's standard reward system based on the questionnaire length; the exact amount was not disclosed to the researchers. No images or other materials that could reveal participant identities are included in this manuscript.

Results

Data Availability

The sample size consisted of 4775 adults. Of these, data on mHealth app usage were available for 4440 participants and were included in the analyses. Those who did not provide data on mHealth app use (n=282) were more likely to report an older age (t_{4720}=17.18, P<.001; Cohen d=1.055). All t values are 2-tailed. There were also significant differences regarding educational level ($\chi^2_{3,4722}$=19.5, P<.001; ϕ_c=0.064) and income status ($\chi^2_{3,4722}$=117.2, P<.001; ϕ_c=0.158). Data providers were more likely to have a university or higher vocational school education and a monthly income of more than 50,000 CZK than data nonproviders. On the other hand, data nonproviders reported secondary vocational school education more frequently, and they were more likely to have an annual income of less than 20,000 CZK than data providers. No significant gender differences were observed ($\chi^2_{1,4718}$=3.04, P=.08; φ=0.025).

XSL•FO
RenderX

Prevalence of mHealth App Use and Sample Characteristics

Around half of the participants reported using mHealth apps (2172/4440, 48.92%). The characteristics of the total sample, mHealth app users, and nonusers on sociodemographics, digital knowledge and use factors, and health indicators are shown in Table 1.

Table 1. Sample characteristics of Czech internet users participating in this cross-sectional study[a].

Variables	Total sample (N=4440)	App users (n=2172)	App nonusers (n=2268)
Sociodemographics			
Age (years), mean (SD)	44.43 (15.97)	41.36 (15.01)	47.38 (16.31)
Gender, woman, n (%)	2185 (49.26)	1108 (51.08)	1077 (47.51)
Educational level, n (%)			
None/primary education	286 (6.44)	141 (6.49)	145 (6.39)
Secondary vocational	1324 (29.82)	568 (26.15)	756 (33.33)
Secondary school	1733 (39.03)	857 (39.46)	876 (38.62)
University (or higher vocational)	1097 (24.71)	606 (27.90)	491 (21.65)
Income status, n (%)			
Up to 20,000 CZK[b]	454 (10.23)	175 (8.06)	279 (12.30)
20,001-35,000 CZK	1364 (30.72)	623 (28.68)	741 (32.67)
35,001-50,000 CZK	1455 (32.77)	715 (32.92)	740 (32.63)
50,001 CZK and more	1167 (26.28)	659 (30.34)	508 (22.40)
Digital knowledge and use, mean (SD)			
eHealth literacy	3.73 (0.67)	3.81 (0.65)	3.65 (0.68)
Passive SNS use[c]	1.9 (0.94)	2.07 (1)	1.74 (0.83)
Active SNS use[d]	1.49 (0.82)	1.59 (0.9)	1.4 (0.73)
Health indicators, mean (SD)			
Symptom severity	2.55 (0.83)	2.6 (0.81)	2.5 (0.84)
Physical activity	2.82 (1.63)	3.10 (1.59)	2.52 (1.62)
BMI	27.66 (5.7)	27.49 (5.49)	27.82 (5.9)
ED-related risk[e]	2.73 (1.4)	2.97 (1.41)	2.5 (1.34)

[a]All values are calculated based on available (nonmissing) data; denominators may therefore vary across variables.

[b]A currency exchange rate of 1 CZK=US $0.048 is applicable.

[c]Passive SNS use: passive use of social networking sites for health information.

[d]Active SNS use: active use of social networking sites for health information.

[e]ED-related risk: eating disorder–related risk propensity (measured by body dissatisfaction, dietary restraint, and weight/shape overvaluation).

Characteristics of mHealth App Users

The multivariate hierarchical binary logistic regression analysis to evaluate the factors associated with mHealth app use is shown in Table 2. Model 1 reports the sociodemographic variables—age, gender, educational level, and income status. Model 2 tests the direct effects of digital knowledge and use factors (eHealth literacy and active and passive use of SNS for health information) and health indicators (symptom severity, physical activity, BMI, and eating disorder–related risk propensity). All models report odds ratios with 95% CIs.

XSL•FO
RenderX

Table 2. Sociodemographic, digital, and health-related predictors of mHealth app use based on multivariate hierarchical binary logistic regression analysis in a cross-sectional sample of Czech internet users (n=3923).

Predictors	Model 1[a]		Model 2[b]	
	OR[c] (95% CI)	*P* value	OR (95% CI)	*P* value
Sociodemographics				
Age (years)	0.98 (0.97-0.98)	<.001	0.98 (0.98-0.99)	<.001
Gender (reference: male)	1.24 (1.09-1.42)	.001	1.13 (0.98-1.3)	.09
Educational level				
None/primary education	0.75 (0.56-1.01)	.06	0.8 (0.59-1.09)	.15
Secondary vocational	0.87 (0.73-1.04)	.14	0.93 (0.77-1.12)	.44
Secondary school	0.92 (0.78-1.08)	.29	0.96 (0.81-1.13)	.61
University (or higher vocational)	Reference	Reference	Reference	Reference
Income status				
Up to 20,000 CZK[d]	0.56 (0.43-0.71)	<.001	0.54 (041-0.7)	<.001
20,001-35,000 CZK	0.78 (0.66-0.93)	<.001	0.78 (0.65-0.93)	.006
35,001-50,000 CZK	0.83 (0.7-0.98)	.006	0.83 (0.7-0.99)	.03
50,001 CZK and more	Reference	Reference	Reference	Reference
Digital knowledge and use				
eHealth literacy	__[e]	—	1.17 (1.06-1.3)	.003
Passive SNS use[f]	—	—	1.35 (1.21-1.51)	<.001
Active SNS use[g]	—	—	0.91 (0.81-1.02)	.10
Health indicators				
Symptom severity	—	—	1.09 (0.99-1.19)	.07
Physical activity	—	—	1.18 (1.13-1.23)	<.001
BMI	—	—	1 (0.99-1.02)	.57
ED-related risk[h]	—	—	1.18 (1.12-1.25)	<.001

[a]Nagelkerke R^2=0.056.

[b]Nagelkerke R^2=0.128.

[c]OD: odds ratio.

[d]A currency exchange rate of CZK 1=US $0.048 is applicable.

[e]Not applicable.

[f]Passive SNS use: passive use of social networking sites for health information.

[g]Active SNS use: active use of social networking sites for health information.

[h]ED-related risk: eating disorder–related risk propensity (measured by body dissatisfaction, dietary restraint, and weight/shape overvaluation).

The significant predictors of having an app were age, gender, and income status in model 1. Participants who were younger, female, and had a monthly income of more than 50,000 CZK (vs up to 20,000 CZK, between 20,001 and 35,000 CZK, or between 35,001 and 50,000 CZK) were more likely to use mHealth apps. In model 2, digital knowledge, use factors, and health indicators were entered. The role of gender became insignificant when additional variables were included in model 2. After adjusting for the roles of sociodemographic characteristics, the results demonstrated that app users had higher eHealth literacy and were more likely to engage in passive use of SNS for health information. Among the health indicators, higher physical activity and a propensity for risk related to eating disorders were predictive of mHealth app use.

mHealth App Use by Type of App

The frequency of use for each mHealth app type among app users is shown in Table 3. The most frequently used mHealth app type was the number of steps. The percentage of participants who count steps daily or several times a day was 65.99% (1430/2167). It was followed by mHealth apps to track sleep, vitals, and sports. The percentage of participants who used apps daily or several times a day for monitoring sleep was 31.95% (691/2163), 28.22% (611/2165) for monitoring vitals, and 18.86% (407/2158) for sports. App users were least likely to have ever used apps for emergency care guidance, diagnosis

XSL•FO
RenderX

assistance, or to manage diagnosed conditions. The percentages of participants who reported nonuse of apps for emergency care guidance, diagnosis assistance, and diagnosed conditions were 76.72% (1654/2156), 71.62% (1542/2153), and 70.49% (1517/2155), respectively.

Table 3. Frequency of using different types of mHealth apps in a cross-sectional sample of Czech internet users (n=2175).

Type of app	Never, n (%)	A few times at most, n (%)	Several times a month, n (%)	Several times a week, n (%)	Daily, n (%)	Several times a day, n (%)
Sports (n=2158)	449 (20.83)	489 (22.66)	394 (18.26)	419 (19.42)	328 (15.2)	79 (3.66)
Number of steps (n=2167)	60 (2.77)	189 (8.72)	224 (10.34)	264 (12.18)	1174 (54.18)	256 (11.81)
Nutrition (n=2161)	803 (37.16)	565 (26.15)	299 (13.84)	219 (10.13)	222 (10.27)	53 (2.45)
Vitals (n=2165)	382 (17.64)	418 (19.31)	392 (18.11)	362 (16.72)	503 (23.23)	108 (4.99)
Sleep (n=2163)	530 (24.5)	408 (18.86)	277 (12.81)	257 (11.88)	644 (29.77)	47 (2.17)
Mood and well-being (n=2156)	1103 (51.16)	447 (20.73)	222 (10.3)	177 (8.21)	180 (8.35)	27 (1.25)
Diagnosed conditions (n=2155)	1517 (70.49)	266 (12.36)	136 (6.32)	104 (4.83)	105 (4.88)	24 (1.12)
Reproductive health (n=2152)	1327 (61.52)	233 (10.8)	314 (14.56)	125 (5.8)	131 (6.07)	27 (1.25)
Diagnosis assistance (n=2153)	1542 (71.62)	323 (15)	134 (6.22)	81 (3.76)	58 (2.69)	15 (0.7)
Emergency care guidance (n=2156)	1654 (76.72)	267 (12.38)	90 (4.17)	78 (3.62)	56 (2.6)	11 (0.51)

Factor Analysis for mHealth Apps

The results confirmed the 2-factor structure of the mHealth app measure with good construct validity (χ^2_{26}=71.9, comparative fit index=0.99, Tucker-Lewis index=0.99, root-mean-square error of approximation=0.03, and standardized root-mean-square residual=0.03). The factor loadings are shown in Figure 1. The results revealed that mHealth usage could be categorized into two dimensions: mHealth app usage for promoting wellness and mHealth app usage for managing illness.

XSL•FO
RenderX

Figure 1. Two-factor structure of mobile health app usage identified using exploratory structural equation modeling in a cross-sectional sample of Czech internet users (n=2152). The numerical values represent factor loadings. **P<.001.

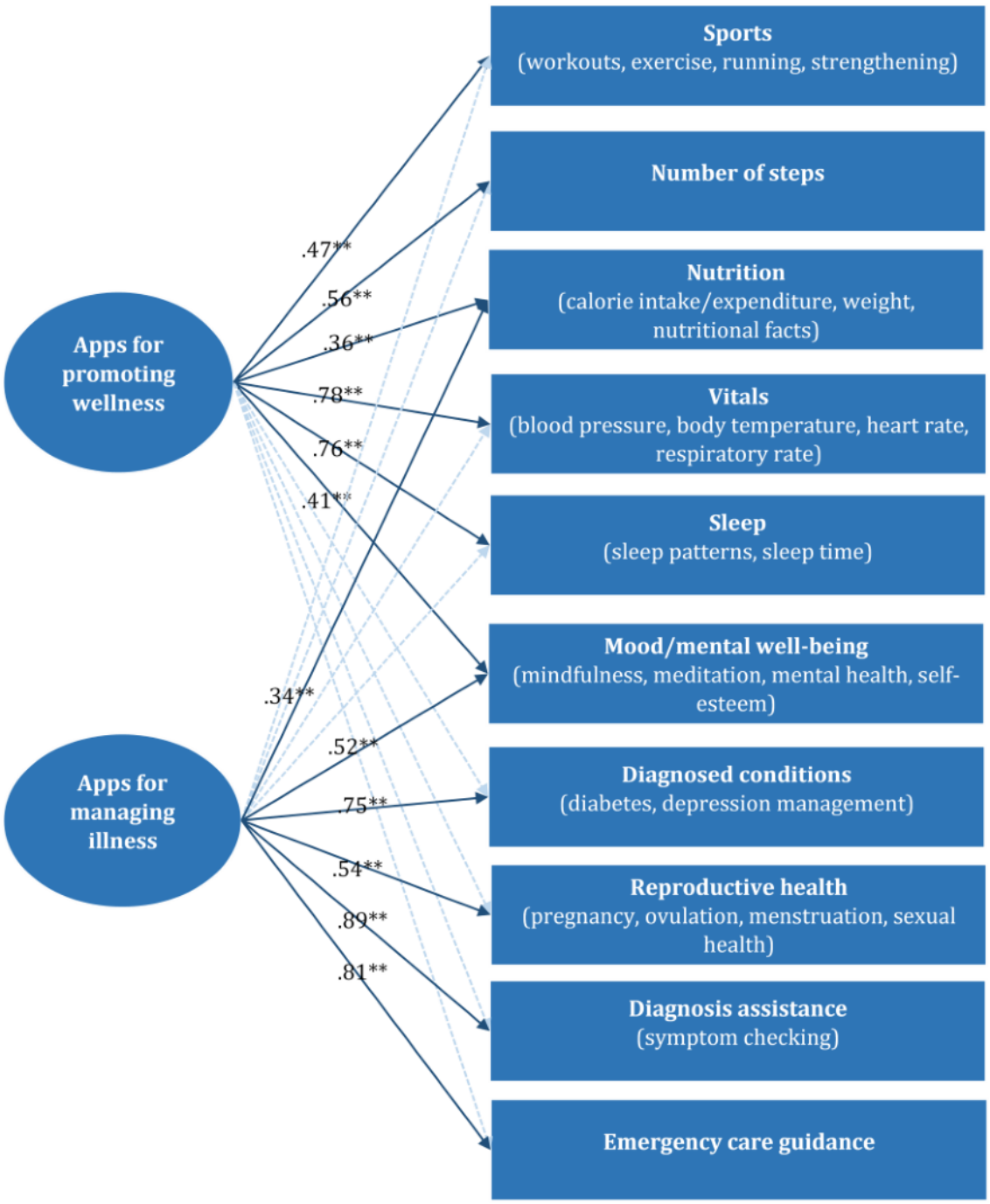

Characteristics of Wellness and Illness App Users

Table 4 presents the correlations between mHealth app usage and the predictors.

XSL•FO
RenderX

Table 4. Spearman correlations between mHealth app use and sociodemographic, digital, and health-related variables in a cross-sectional sample of Czech internet users[a].

Variables	Wellness apps[b]	Illness apps[c]	Age	Gender	Education	Income	eHealth literacy	SNS active[d]	SNS passive[e]	Sx[f]	BMI	PA[g]	ED risk[h]
Wellness apps													
ρ [i]	1												
P value	—[j]												
n	—												
Illness apps													
ρ	0.645	1											
P value	<.001	—											
n	2154	—											
Age													
ρ	–0.207	–0.27	1										
P value	<.001	<.001	—										
n	2178	2161	—										
Gender													
ρ	0.031	–0.08	0.009	1									
P value	0.15	<.001	0.55	—									
n	2175	2159	4771	—									
Education													
ρ	0.007	–0.078	–0.077	–0.053	1								
P value	0.75	<.001	<.001	<.001	—								
n	2178	2161	4775	4771	—								
Income													
ρ	0.043	–0.042	–0.169	0.148	0.273	1							
P value	0.05	0.05	<.001	<.001	<.001	—							
n	2178	2161	4775	4771	4775	—							
eHealth literacy													
ρ	0.123	0.086	–0.193	0.006	0.183	0.155	1						
P value	<.001	<.001	<.001	0.69	<.001	<.001	—						
n	2174	2157	4756	4752	4756	4756	—						
SNS active													
ρ	0.202	0.47	–0.298	–0.07	–0.057	–0.051	0.1	1					
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—					
n	2164	2147	4732	4728	4732	4732	4715	—					
SNS passive													
ρ	0.223	0.462	–0.243	–0.155	–0.012	–0.04	0.165	0.668	1				
P value	<.001	<.001	<.001	<.001	0.4	0.005	<.001	<.001	—				
n	2165	2148	4740	4736	4740	4740	4723	4731	—				
Sx													
ρ	0.099	0.242	0.009	–0.146	–0.106	–0.147	–0.032	0.245	0.304	1			
P value	<.001	<.001	0.52	<.001	<.001	<.001	0.03	<.001	<.001	—			
n	2178	2161	4766	4762	4766	4766	4747	4724	4732	—			

XSL•FO
RenderX

Variables	Wellness apps[b]	Illness apps[c]	Age	Gender	Education	Income	eHealth literacy	SNS active[d]	SNS passive[e]	Sx[f]	BMI	PA[g]	ED risk[h]
BMI													
ρ	0.005	–0.047	0.279	0.105	–0.112	–0.019	–0.085	–0.06	–0.057	0.086	1		
P value	0.83	0.03	<.001	<.001	<.001	0.2	<.001	<.001	<.001	<.001	—		
n	2178	2161	4775	4771	4775	4775	4756	4732	4740	4764	—		
PA													
ρ	0.179	0.093	–0.19	0.072	0.13	0.128	0.16	0.088	0.108	–0.055	–0.137	1	
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	
n	2034	2019	4299	4295	4299	4299	4283	4259	4265	4295	4299	—	
ED risk													
ρ	0.199	0.307	–0.116	–0.14	0.02	–0.009	0.053	0.227	0.296	0.307	0.276	0.154	1
P value	<.001	<.001	<.001	<.001	0.18	0.52	<.001	<.001	<.001	<.001	<.001	<.001	—
n	2164	2147	4731	4727	4731	4731	4712	4689	4697	4726	4731	4273	—

[a]Sample sizes vary by correlation pair; Ns are shown within the table.

[b]Wellness apps: frequency of using mHealth apps for promoting wellness.

[c]Illness apps: frequency of using mHealth apps for managing illness.

[d]SNS active: active use of social networking sites (SNS) for health information.

[e]SNS passive: passive use of social networking sites (SNS) for health information.

[f]Sx: perceived severity of symptoms.

[g]PA: physical activity.

[h]ED risk: eating disorder–related risk propensity (measured by body dissatisfaction, dietary restraint, and weight/shape overvaluation).

[i]ρ: Spearman rank order correlation.

[j]Not applicable.

Aligned with the ESEM method for factor analysis, the current model did not restrict the factor loadings of mHealth apps to a particular factor. Table 5 shows the results. When interpreting the results, we focused not only on statistical significance but also on the effect sizes. Following widely acknowledged guidelines, standardized path coefficients (β) between 0.1 and 0.3 were interpreted as indicating a small effect, those between 0.3 and 0.5 as medium, and those above 0.5 as large effects [39]. In the first model, which included only sociodemographic independent variables, app use to promote wellness was predicted by age (standardized β=–0.22, *P*<.001), indicating that older age was correlated with less mHealth app use to promote wellness. Except for gender, sociodemographic characteristics, including age, educational level, and income status, were significantly linked to app use in managing illness. However, the effect size of income was negligible (retaining three decimal places, standardized β=0.097). Likewise, the effect of education on illness-related app use was also at the boundary of a negligible effect size (retaining three decimal places, standardized β=–0.106). Furthermore, after including health-, digital knowledge-, and use-related variables, none of the sociodemographic variables remained significant. We did not interpret them further because both education level and income were already at the boundary of a negligible effect size when only sociodemographic variables were considered, and adding additional variables merely pushed them fully into the negligible range. Notably, however, the reduction in effect size of age was considerable for both app types, which prompted us to conduct further analyses to examine the mediating roles of digital knowledge, digital use, and health indicators in the association between age and wellness- and illness-related app use.

XSL•FO
RenderX

Table 5. Predictors of the frequency of using mHealth apps to promote wellness and manage illness based on exploratory structural equation modeling in a cross-sectional sample of Czech internet users[a].

	Model 1 (n=2152)			Model 2 (n=1985)		
Predictors	Standardized β	SE	*P* value	Standardized β	SE	*P* value
Apps to promote wellness						
Age	–0.22	0.00	<.001	–0.18	0.00	<.001
Gender	0.03	0.03	.07	0.04	0.05	.04
Education	0.03	0.02	.15	0.02	0.03	.24
Income	0.02	0.02	.27	0.02	0.02	.42
eHealth literacy	__[b]	—	—	0.10	0.02	<.001
Active SNS use[c]	—	—	—	0.07	0.04	.12
Passive SNS use[d]	—	—	—	0.08	0.05	.07
Symptom severity	—	—	—	0.04	0.03	.08
BMI	—	—	—	0.08	0.03	<.001
Physical activity	—	—	—	0.15	0.02	<.001
ED-related risk[e]	—	—	—	0.09	0.03	<.001
Apps to manage illness						
Age	–0.23	0.00	<.001	–0.04	0.00	.07
Gender	0.02	0.03	.31	0.01	0.08	.75
Education	–0.11	0.02	<.001	–0.05	0.04	.05
Income	–0.10	0.02	<.001	–0.02	0.04	.40
eHealth literacy	—	—	—	0.00	0.04	.93
Active SNS use[c]	—	—	—	0.62	0.15	<.001
Passive SNS use[d]	—	—	—	0.06	0.11	.49
Symptom severity	—	—	—	0.04	0.04	.20
BMI	—	—	—	–0.07	0.04	.01
Physical activity	—	—	—	0.02	0.03	.46
ED-related risk[e]	—	—	—	0.11	0.04	<.001

[a]Sample sizes vary across models; Ns are reported in the table.

[b]Not applicable.

[c]SNS active: active use of social networking sites for health information.

[d]SNS passive: passive use of social networking sites for health information.

[e]ED-related risk: eating disorder–related risk propensity (measured by body dissatisfaction, dietary restraint, and weight/shape overvaluation).

The results also suggested that eHealth literacy and physical activity were significantly correlated with the frequency of mHealth app usage to promote wellness. Regarding the use of mHealth apps to manage illness, the active use of SNS for health-related information and eating disorder–related risk propensity were significant predictors of the frequency of use.

For a more robust interpretation, we additionally performed basic linear regression analyses using the ordinary least squares method (see Table 6). The standardized estimates did not differ substantially from the main analyses, except that active use of SNS for health-related information (SNS active) was significant in the simple linear regression model but was not significantly associated with mHealth app usage for promoting wellness in the full model. It is worth noting, however, that the effect size was small (standardized β=0.13, *P*<.001). Similarly, in the ESEM model, the effect size of eating disorder risk on wellness-related mHealth app usage was negligible (standardized β=0.09), and in the ordinary least squares regression model, it was 0.12—still very small, bordering on a negligible effect size. Such a minor difference cannot be considered meaningful. Considering that the ESEM model accounted for factor loadings and that the effect size in the simple regression remained small, we regarded this discrepancy as acceptable. Therefore, the results from the ESEM model (Table 5) should be prioritized for interpretation.

XSL•FO
RenderX

Table 6. Predictors of the frequency of using mHealth apps to promote wellness and manage illness based on simple regression models in a cross-sectional sample of Czech internet users[a].

	Model 1			Model 2		
Predictors	Standardized β	SE	*P* value	Standardized β	SE	*P* value
Apps to promote wellness (n=2005)						
Age	–0.22	0.00	<.001	–0.15	0.00	<.001
Gender (2-1)[b]	0.07	0.04	.08	0.10	0.04	.02
Education						
3-4[c]	–0.01	0.05	.91	0.01	0.05	.81
2-4[d]	0.05	0.06	.41	0.05	0.06	.37
1-4[e]	–0.12	0.10	.23	–0.13	0.09	.18
Income	0.00	0.02	.88	0.02	0.02	.48
eHealth literacy	__[f]	—	—	0.08	0.03	<.001
Active SNS use[g]	—	—	—	0.13	0.03	<.001
Passive SNS use[h]	—	—	—	0.09	0.03	.004
Symptom severity	—	—	—	0.04	0.03	.12
BMI	—	—	—	0.04	0.00	.12
Physical activity	—	—	—	0.15	0.01	<.001
ED-related risk[i]	—	—	—	0.12	0.02	<.001
Apps to manage illness (n=1991)						
Age	–0.26	0.00	<.001	–0.09	0.00	<.001
Gender (2-1)[b]	–0.02	0.04	.64	0.00	0.03	.85
Education						
3-4[c]	0.03	0.05	.62	0.02	0.04	.62
2-4[d]	0.26	0.06	<.001	0.14	0.05	.004
1-4[e]	0.13	0.09	.17	0.02	0.07	.84
Income	–0.08	0.02	<.001	–0.01	0.02	.50
eHealth literacy	—	—	—	0.02	0.02	.19
Active SNS use[g]	—	—	—	0.51	0.02	<.001
Passive SNS use[h]	—	—	—	0.07	0.02	.01
Symptom severity	—	—	—	0.04	0.02	.05
BMI	—	—	—	–0.04	0.00	.03
Physical activity	—	—	—	0.05	0.01	.003
ED-related risk[i]	—	—	—	0.14	0.01	<.001

[a]Sample sizes vary across models; Ns are reported in the table.
[b]2-1: male-female.
[c]3-4: secondary school – university (or higher vocational school).
[d]2-4: secondary vocational school – university (or higher vocational school).
[e]1-4: none/primary education – university (or higher vocational school).
[f]Not applicable.
[g]SNS active: active use of social networking sites for health information.
[h]SNS passive: passive use of social networking sites for health information.
[i]ED-related risk: eating disorder–related risk propensity (measured by body dissatisfaction, dietary restraint, and weight/shape overvaluation).

XSL•FO
RenderX

Mediation Analysis

Due to the decrease in the effect size of age on two dependent variables after the addition of digital knowledge, digital use, and health indicators, it was worthwhile to explore further whether these variables mediated the relationship between age and mHealth usage. The mediation analysis suggested that older age correlated with more frequent use of wellness apps through "eHealth literacy," "active SNS use," "passive SNS use," "physical activity," and "eating disorder–related risk propensity" (see Figure 2). On the other hand, "active SNS use," "passive SNS use," and "eating disorder–related risk propensity" mediated the relationship between age and frequency of mHealth app use to manage illness (see Figure 3). It is worth noting that the correlations between active and passive SNS use and illness-related mHealth app usage were more robust than those of wellness-related mHealth app usage.

Figure 2. Mediation model depicting associations between age and wellness-related mobile health app use based on exploratory structural equation modeling in a cross-sectional sample of Czech internet users (n=1985). Only statistically significant paths (*P*<.05) are shown. Standardized coefficients are displayed. *SNS active: active use of social networking sites for health information. **SNS passive: passive use of social networking sites for health information. ***ED-related risk: eating disorder–related risk propensity (measured by body dissatisfaction, dietary restraint, and weight/shape overvaluation).

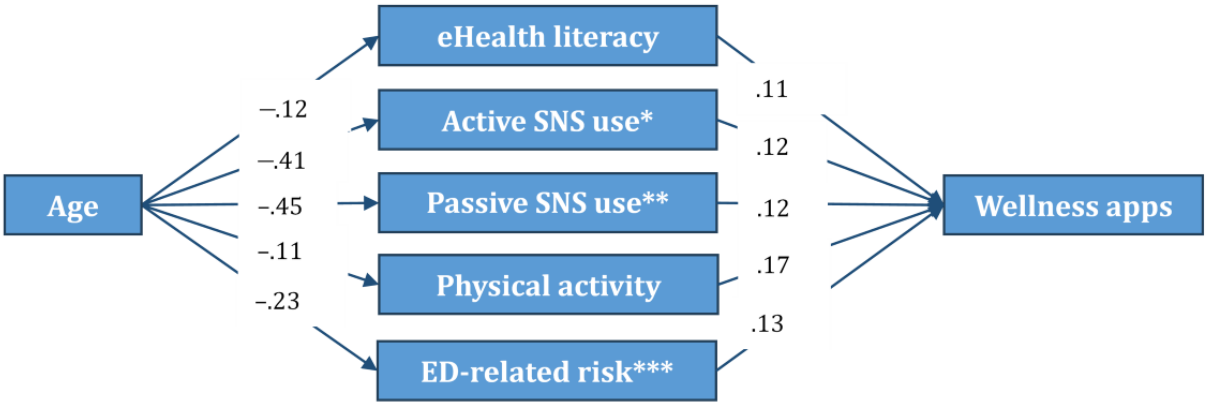

Figure 3. Mediation model depicting associations between age and illness-related mobile health app use based on exploratory structural equation modeling in a cross-sectional sample of Czech internet users (n=1985). Only statistically significant paths (*P*<.05) are shown. Standardized coefficients are displayed. *SNS active: active use of social networking sites for health information. **SNS passive: passive use of social networking sites for health information. ***ED-related risk: eating disorder–related risk propensity (measured by body dissatisfaction, dietary restraint, and weight/shape overvaluation).

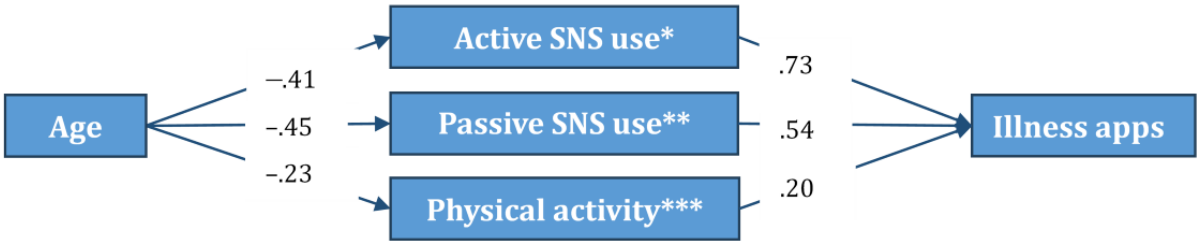

Discussion

Overview

This study evaluated the prevalence, user characteristics, and usage patterns of various mHealth apps in a representative sample of Czech adults with internet access. Furthermore, it distinguished between wellness and illness apps using a data-driven approach and identified user characteristics associated with the frequency of use of each dimension. This study provided significant insights into the relationship between user characteristics, intentions, and engagement with mHealth apps.

Principal Results

Prevalence of mHealth App Use

Around half of the participants with internet access (48.92%) had been using mHealth apps. Previous studies with adults have reported prevalence rates ranging from 34.1% to 56.6% in the United States [9-12]. Lower rates were reported in European and Far Eastern countries. For instance, 20.5% of German

XSL•FO
RenderX

smartphone owners in 2015 [18] and 24.1% of Hong Kongian smartphone or tablet owners in 2016 [23] used mHealth apps in population-based samples. More up-to-date findings are needed for a more accurate comparison of the prevalence rates across countries.

Characteristics of mHealth App Users

Sociodemographic Characteristics

Age was significantly associated with using mHealth apps. In line with earlier studies [8], younger adults were more likely to use them. This finding highlights that older adults are underserved in terms of mHealth app use. Lack of awareness, lack of motivation, low self-efficacy, mistrust of technology, and the absence of required technical skills are frequently reported barriers to using mHealth apps among older adults [40,41]. Additionally, cognitive impairment, reduced physical ability, and visual changes associated with aging can compromise the usability of app interfaces in this population [42-45]. Dispositional, attitudinal, and aging-related factors must be considered when addressing this population via apps.

Participants with monthly incomes exceeding 50,000 CZK were more likely to have mHealth apps than those in other income categories (ie, up to 20,000 CZK, between 20,001 and 35,000 CZK, and between 35,001 and 50,000 CZK). This finding is consistent with previous studies that have reported higher annual income as a significant predictor of mHealth app adoption and usage [9,11,12,23]. Higher-income status may facilitate access to apps that require payment to download and use, whereas such apps may be less affordable for low-income individuals. The association between educational level and mHealth app usage was not statistically significant. This finding contrasts with previous studies conducted in the United States [9-12] and Hong Kong [23], but it is similar to an earlier study conducted in Germany [18]. Future studies will clarify whether educational background will influence the adoption and use of apps as they become more affordable, user-friendly, and accessible.

Digital Knowledge and Use Factors

Adults with higher eHealth literacy were more likely to have mHealth apps. This finding aligns with previous studies [14,46] and confirms that eHealth literacy is a crucial digital skill for understanding mHealth app use behaviors. Individuals with higher eHealth literacy tend to seek online health information more frequently [47], are more health information-oriented [48], and place greater importance on health-related information [49]. They are also more likely to report competence in self-care [50] and their intention to promote health [51]. These findings underscore that perceived digital skills in finding, understanding, evaluating, and applying online health information in a health context might be precursors to having and using apps.

This study provided initial evidence in support of the association between SNS activity for health information and having mHealth apps. We found that passive use (ie, reading, watching, or following health content) was significantly associated with the use of apps. However, the association between active use (ie, creating, sharing, and commenting on health content) and app use was insignificant. Our results reveal that adopting mHealth apps might be more closely related to consuming health-related information on SNS (ie, passive use) than engaging with online networks for health-related information (ie, active use). This finding aligns with a previous study that reported a significant association between health information-seeking behaviors on social media and the use of mHealth apps for health information [52]. Health information–oriented individuals may use SNS and mHealth apps in a complementary manner to meet their health information needs. Additionally, SNS provides a platform for promoting various functions of mHealth apps [53], which may help meet the personal health care needs of consumers and contribute to the adoption of mHealth apps. These findings suggest that SNS could be a significant promotional channel.

Health Indicators

In line with previous research [10,18,23], participants with higher physical activity were more likely to have mHealth apps, which provides further evidence for the association between a physically active lifestyle and mHealth app adoption. We found a significant association between eating disorder–related risks and mHealth app use. Participants who reported higher dietary restraint, weight and shape overvaluation, and body dissatisfaction were more likely to have and use mHealth apps. Prior studies reporting an association between eating disorder–related risks and app usage have primarily focused on apps that promote weight loss and physical activity [28-30]. This study shows that body dissatisfaction, weight and shape overvaluation, and restrictive eating behaviors correlate with overall app adoption.

BMI was not associated with using apps, contrasting with earlier studies that reported a significant association between app usage and a higher BMI [10,18,25]. It is possible that eating disorder–related risks (ie, body dissatisfaction, weight/shape overvaluation, and dieting behaviors) capture app usage behaviors better than body weight. We did not observe a significant association between the use of apps and symptom severity. This finding aligns with studies that found no significant association between health status and mHealth app usage [10-12]. This study focused on a detailed measurement of perceived health status by assessing somatic symptom severity in multiple domains.

mHealth App Use by Type of App

Apps that count the number of steps were most frequently used. This finding aligns with global statistics, which indicate that the most downloaded app in 2022 was a step counter [54]. One reason for the frequent use of step counter apps could be related to their simplicity. Previous research showed this to be a significant factor in determining the frequency of using apps for physical activity [25]. Apps that count the number of steps may take less effort from consumers to integrate them into everyday life. These apps are often built-in apps delivered in the latest releases of major smartphone companies, and their availability could enhance increased adoption and use. Apps that track sleep, vitals, and sports activities were also popular. In a recent qualitative study, app users reported that their continued use of apps with tracking functions (eg, tracking steps, exercise, heart rate, and sleep) was associated with the perceived usefulness of these apps for maintaining and improving their health and well-being [55].

XSL•FO
RenderX

Factor Analysis for mHealth Apps

Using a data-driven approach with ESEM, we confirmed the 2-factor structure of mHealth app usage in the analyses. The apps under the "promoting wellness" factor were related to lifestyle behaviors (eg, sports apps, nutrition apps, and number of steps apps), monitoring of health indicators (eg, vitals apps and sleep apps), and the enhancement of mental vitality (eg, mood and mental well-being apps). This factor consisted of apps focusing on preventive health behaviors to promote wellness and reduce the risk of illness. The apps under the "managing illness" factor focused on maintaining health (eg, apps for reproductive health and diagnosis assistance) and providing support for illnesses (eg, apps for diagnosed conditions and emergency care guidance). We found that nutrition apps and mood and mental well-being apps cross-loaded on both factors. Several health conditions (eg, diabetes) require close monitoring of nutritional intake [56]. Therefore, it is reasonable that app users might monitor nutrition not only for promoting wellness but also for managing diagnosed health conditions. Similarly, mood and mental well-being are broad domains, and app users' intentions may have dual relevance, encompassing both the promotion of mood and mental well-being for wellness purposes and the management of mental health problems as part of a treatment regimen [57]. In conclusion, the factor structure of our mHealth measure was feasible. It might guide future research studies on app usage for health promotion and illness management.

Characteristics of Wellness and Illness App Users

Age was significantly associated with the frequency of using wellness apps. Older adults used them less frequently to promote wellness. This finding is disappointing because older adults are more susceptible to chronic conditions like diabetes, hypertension, and cardiovascular disease [58]. Effective characteristics of wellness apps that enhance the adherence of older adults should be identified to combat their digital exclusion.

eHealth literacy was significantly associated with more frequent use of apps to promote wellness, but no such association was observed for apps that manage illness. Previous research has confirmed that higher eHealth literacy is associated with perceiving mHealth apps as effective and using them more frequently for health-related behavioral change [14,15,45]. Our findings indicate that this association might be particularly relevant for behaviors that promote health. Individuals with higher eHealth literacy are more motivated to engage in behaviors to regulate their diet, sleep, and physical activity [59]. They are more likely to adhere to recommendations to reduce the risk of illness [60] and to manage stress [51]. Having the necessary digital skills to navigate within apps might reinforce the regular use of wellness apps in this population, which is inherently more motivated to engage in health-promoting behaviors.

We found a significant association between active SNS use for health information and the more frequent usage of apps to manage illness. This finding suggests that proactive engagement with online social networks for health information plays a crucial role in the frequent use of illness apps. A recent study reported a significant association between having a friend to discuss health-related issues and the use of mHealth apps [12]. We consider that our finding aligns with this previous study because the active use of SNS promises a platform for individuals to interact with their close networks about health and connect with others who might have similar health concerns. Users of illness apps might use SNS complementarily to exchange health information, share experiences, discuss health-related decisions, and provide and receive support from their online social networks [61]. The active use of these networks can synergistically encourage compliance with apps to manage illness, as they offer real-time feedback and a direct line of communication to those who may face similar health concerns.

Higher physical activity was associated with a more frequent use of wellness apps. Wellness apps (eg, Fitbit) offer a wide range of functions that physically active individuals can use to monitor and regulate their physical activity, conditioning, and fitness [62]. They also offer a platform to manage their diet, vitals, energy levels, and mental state [62], which addresses well-being holistically. These apps may be particularly appealing to physically active adults who may be inherently more motivated to pursue a healthy lifestyle [63,64] and thus use them more regularly to enhance their well-being.

This study provided initial evidence to indicate that individuals with elevated risks of eating disorders use mHealth apps more frequently to manage illness. Body image concerns and dieting behaviors may be influenced by various health conditions beyond eating disorders (eg, obesity, diabetes, and depression) [65,66]. The patterns and consequences of using apps can differ substantially depending on the health domain for which they are used. Previous studies focused on weight loss and physical activity apps to examine the association between unintended uses and the relative consequences of using these apps in exacerbating or triggering adverse outcomes among individuals with body image concerns and dieting behaviors [67-69]. These previous findings underscored that underweight BMI goals had a detrimental impact on eating disorder–related symptoms and behaviors, such as more restrictive eating behaviors, control, and obsession with eating and exercise. In contrast, more realistic goals regarding weight and shape, along with motivation toward healthy lifestyle choices, were associated with improved food choices, exercise behaviors, and health outcomes. Therefore, future research should specify health behavior goals and the consequences of using apps to manage illness in individuals with elevated body image and eating concerns.

Mediation Analysis

As mentioned earlier, while conducting the analyses, we observed that the effect size of age was considerably reduced for both app types after including digital knowledge, digital use, and health-related variables in the equation. Therefore, although it was not among our initial research aims, we decided to explore whether any of these variables mediated the association between age and app usage. The results revealed important insights into the association between older age and the less frequent use of apps for wellness and illness purposes.

The association between age and mHealth app use was negatively mediated by the active and passive use of SNS for

XSL•FO
RenderX

health information. In other words, older adults were less likely to use SNS to seek health information (ie, passive use) and to interact with online networks about health (ie, active use), which was significantly associated with less frequent app use. This association was significant for both app types and notably stronger for apps to manage illness. Compared with younger adults, older adults use social media and search engines less frequently for health information [52]. They are also more likely to prefer traditional media (eg, television and newspaper) [70]. Our findings indicate that a lack of orientation toward seeking and exchanging health information on SNS could hinder older adults' use of mHealth apps. SNS may provide know-how about app functions and facilitate social support for sustained use of mHealth apps.

The association between age and wellness app use was negatively mediated by eHealth literacy. Older adults had lower eHealth literacy, which was significantly associated with less frequent use of wellness apps. Previous studies have confirmed the associations between older age, lower eHealth literacy, and the use of mHealth apps [8,71]. Our finding shows that enhancing eHealth literacy skills could significantly increase app compliance for health behaviors that promote well-being in this population.

The association between age and wellness app use was negatively mediated by physical activity. Older adults reported lower physical activity levels, which were significantly associated with less frequent use of wellness apps. Older age is characterized by reduced physical activity [72], and older adults are less likely to adhere to recommendations for physical exercise [73,74]. However, a physically active lifestyle is associated with a reduced risk of all-cause mortality, a better quality of life, and improved cognitive functioning in older age [75]. Wellness apps offer novel opportunities to promote physical activity, reduce sedentary behaviors, and enhance physical and mental well-being in this population [76]. Thus, our finding highlights a significant gap in their dissemination to older adults, who might need such strategies most. The obstacles associated with lower acceptance and the usage of wellness apps and strategies to empower older adults' self-efficacy regarding healthy aging should be identified.

Finally, the association between age and mHealth app use was negatively mediated by eating disorder–related risk propensity. This association was significant for both app types. In other words, older adults were less likely to report body image concerns and dieting behaviors, which, in turn, were significantly associated with less frequent use of apps for wellness and illness purposes. The importance placed on physical appearance is reduced by increasing age. Body dissatisfaction, weight and shape overvaluation, and dieting behaviors are reported less frequently after the age of 40 [77]. In general, lower body image preoccupation is associated with healthier lifestyle behaviors, psychological adjustment, and better management of health conditions that require close monitoring of health indicators and dietary behaviors [65,78,79]. Therefore, it is interesting to find that a lower eating disorder–related risk is associated with less frequent use of apps in older adults. Further research is needed to understand the body image perceptions, dieting behaviors, and weight- and shape-related concerns of older adults, as well as their associations with the use of apps.

Limitations and Future Research

The findings of this study should be interpreted in light of its limitations. Due to the cross-sectional study design, temporal and causal associations between the variables could not be assessed. For instance, we cannot ascertain whether eating disorder–related risks determine mHealth app use or whether they are exacerbated due to the use of mHealth apps. Our mHealth app measure consisted of apps frequently downloaded in app stores for health purposes. However, it is not exhaustive and can be improved in later studies. We used a data-driven approach (ESEM) to categorize app usage into wellness and illness dimensions. This approach enabled us to demonstrate the multidimensional structure of mHealth app usage. Wellness and illness dimensions also drew parallels between preventive health behaviors and illness behaviors in the health context [80]. Although ESEM allows for more flexible estimation of latent constructs compared with traditional CFA, its results can still be sensitive to model specification and sample characteristics. Thus, despite ESEM being a better option in this study, the findings should be interpreted with caution, and future studies are encouraged to replicate the model using different samples or model structures.

We determined the usage patterns of apps based on their frequency of use. This approach limited our ability to identify the healthy and unhealthy uses of apps. For instance, excessive use of sports apps can be dysfunctional when users engage with them to the point of overexercising. Weight loss apps might have adverse consequences for at-risk individuals if they are used frequently due to an obsession with eating and body image concerns.

The assessments were based on self-reports. Thus, we could not control the response characteristics of the participants. We used adapted versions of frequently used questionnaires to assess symptom severity and SNS use for health information. Additionally, eating disorder–related risk propensity and eHealth literacy measures have not been previously adequately validated in the Czech context. Nevertheless, all the measures had adequate internal consistency in this study. Besides, we conducted cognitive testing with a subsample of participants to ensure the comprehensibility of items and the quality of the assessment procedure.

Future research is needed to investigate the health behaviors and outcomes associated with the use of apps. Usage intentions and motivations are essential to understanding the healthy and unhealthy uses of mHealth apps. Therefore, future studies should consider the health behavior goals associated with using apps and identify the factors that contribute to unintended uses and consequences.

Conclusions

This study demonstrates that sociodemographic, digital, and health-related factors might jointly shape engagement with mHealth apps. eHealth literacy and health-related SNS activity emerged as key digital correlates of use, highlighting the potential of promoting digital inclusion to support equitable

XSL•FO
RenderX

participation in digital health. Eating disorder–related risk propensity—reflecting body image concerns and dietary restraint—was associated with the overall app adoption and frequent use of apps to manage illness, underscoring the importance of examining app use purposes and health outcomes among individuals with elevated body image and dietary concerns, and the need for future research and app development to promote health-supportive use while minimizing appearance-driven behaviors. Indirect associations of age through eHealth literacy, health-related SNS use, physical activity, and eating disorder–related risk further highlight multidimensional barriers older adults face in adopting and maintaining mHealth app use. Collectively, these findings indicate that mHealth engagement reflects broader social, digital, and psychological inequalities rather than individual preferences alone. Enhancing digital skills, promoting inclusive app design, and addressing body image- and diet-related use may foster equitable mHealth participation and help prevent widening health disparities across age and user groups.

Data Availability

The data and the analysis scripts are shared via the Open Science Framework [81].

Funding

The research reported in this publication was supported by the NPO “Systemic Risk Institute” grant LX22NPO5101, funded by the European Union—Next Generation EU (Ministry of Education, Youth and Sports, NPO: EXCELES). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional analysis of mobile health app usage.
[XLSX File (Microsoft Excel File), 20 KB - jmir_v28i1e71363_app1.xlsx]

References

1. Wells C, Spry C. CADTH Horizon Scan: an overview of smartphone apps. Can J Health Technol 2022;2(2):1-26. [doi: 10.51731/cjht.2022.269]
2. Peng C, He M, Cutrona SL, Kiefe CI, Liu F, Wang Z. Theme trends and knowledge structure on mobile health apps: bibliometric analysis. JMIR Mhealth Uhealth 2020;8(7):e18212 [FREE Full text] [doi: 10.2196/18212] [Medline: 32716312]
3. Wu CC, Huang CW, Wang YC, Islam M, Kung W, Weng Y, et al. mHealth research for weight loss, physical activity, and sedentary behavior: bibliometric analysis. J Med Internet Res 2022;24(6):e35747 [FREE Full text] [doi: 10.2196/35747] [Medline: 35675126]
4. Sama PR, Eapen ZJ, Weinfurt KP, Shah BR, Schulman KA. An evaluation of mobile health application tools. JMIR Mhealth Uhealth 2014;2(2):e19 [FREE Full text] [doi: 10.2196/mhealth.3088] [Medline: 25099179]
5. Stec M, Arbour MW. Wellness and disease self-management mobile health apps evaluated by the Mobile Application Rating Scale. Adv Fam Pract Nurs 2020;2:87-102. [doi: 10.1016/j.yfpn.2020.01.003]
6. Xu W, Liu Y. mHealth apps: a repository and database of mobile health apps. JMIR Mhealth Uhealth 2015;3(1):e28 [FREE Full text] [doi: 10.2196/mhealth.4026] [Medline: 25786060]
7. Iribarren SJ, Akande TO, Kamp KJ, Barry D, Kader YG, Suelzer E. Effectiveness of mobile apps to promote health and manage disease: systematic review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2021;9(1):e21563 [FREE Full text] [doi: 10.2196/21563] [Medline: 33427672]
8. Bao H, Lee EWJ. Examining theoretical frameworks and antecedents of health apps and wearables use: a scoping review. Health Commun 2024;39(12):2671-2681. [doi: 10.1080/10410236.2023.2283655] [Medline: 37968803]
9. Bhuyan SS, Lu N, Chandak A, Kim H, Wyant D, Bhatt J, et al. Use of mobile health applications for health-seeking behavior among US adults. J Med Syst 2016;40(6):153. [doi: 10.1007/s10916-016-0492-7] [Medline: 27147516]
10. Carroll JK, Moorhead A, Bond R, LeBlanc WG, Petrella RJ, Fiscella K. Who uses mobile phone health apps and does use matter? A secondary data analytics approach. J Med Internet Res 2017;19(4):e125 [FREE Full text] [doi: 10.2196/jmir.5604] [Medline: 28428170]
11. Rising CJ, Jensen RE, Moser RP, Oh A. Characterizing the US population by patterns of mobile health use for health and behavioral tracking: analysis of the National Cancer Institute's Health Information National Trends Survey data. J Med Internet Res 2020;22(5):e16299 [FREE Full text] [doi: 10.2196/16299] [Medline: 32406865]
12. Tundealao S, Titiloye T, Sajja A, Egab I, Odole I, Alufa O, et al. Factors associated with the non-use of mobile health applications among adults in the United States. J Public Health (Berl.) 2023;33(7):1575-1581. [doi: 10.1007/s10389-023-02132-8]

XSL•FO
RenderX

13. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
14. Vervier LS, Valdez AC, Ziefle M. "Attitude"- mHealth apps and users' insights: an empirical approach to understand the antecedents of attitudes towards mHealth applications. In: Proceedings of the 5th International Conference on Information and Communication Technologies for Ageing Well and e-Health (ICT4AWE 2019). Setúbal, Portugal: SciTePress; 2019:213-221.
15. Durmuş A. The influence of digital literacy on mHealth app usability: the mediating role of patient expertise. Digit Health 2024;10:20552076241299061 [FREE Full text] [doi: 10.1177/20552076241299061] [Medline: 39600388]
16. Shaw G, Castro BA, Gunn LH, Norris K, Thorpe RJ. The association of eHealth literacy skills and mHealth application use among US adults with obesity: analysis of health information national trends survey data. JMIR Mhealth Uhealth 2024;12:e46656 [FREE Full text] [doi: 10.2196/46656] [Medline: 38198196]
17. Guo SH, Hsing HC, Lin JL, Lee CC. Relationships between mobile eHealth literacy, diabetes self-care, and glycemic outcomes in Taiwanese patients with type 2 diabetes: cross-sectional study. JMIR Mhealth Uhealth 2021;9(2):e18404 [FREE Full text] [doi: 10.2196/18404] [Medline: 33544088]
18. Ernsting C, Dombrowski SU, Oedekoven M, O Sullivan JL, Kanzler M, Kuhlmey A, et al. Using smartphones and health apps to change and manage health behaviors: a population-based survey. J Med Internet Res 2017;19(4):e101 [FREE Full text] [doi: 10.2196/jmir.6838] [Medline: 28381394]
19. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013;15(4):e85 [FREE Full text] [doi: 10.2196/jmir.1933] [Medline: 23615206]
20. Zhao Y, Zhang J. Consumer health information seeking in social media: a literature review. Health Info Libr J 2017;34(4):268-283 [FREE Full text] [doi: 10.1111/hir.12192] [Medline: 29045011]
21. Liu PL. COVID-19 information seeking on digital media and preventive behaviors: the mediation role of worry. Cyberpsychol Behav Soc Netw 2020;23(10):677-682. [doi: 10.1089/cyber.2020.0250] [Medline: 32498549]
22. Iftikhar R, Abaalkhail B. Health-seeking influence reflected by online health-related messages received on social media: cross-sectional survey. J Med Internet Res 2017;19(11):e382 [FREE Full text] [doi: 10.2196/jmir.5989] [Medline: 29146568]
23. Shen C, Wang MP, Chu JT, Wan A, Viswanath K, Chan SSC, et al. Health app possession among smartphone or tablet owners in Hong Kong: population-based survey. JMIR Mhealth Uhealth 2017;5(6):e77 [FREE Full text] [doi: 10.2196/mhealth.7628] [Medline: 28583905]
24. Kramer H, Cao G, Dugas L, Luke A, Cooper R, Durazo-Arvizu R. Increasing BMI and waist circumference and prevalence of obesity among adults with type 2 diabetes: the National Health and Nutrition Examination Surveys. J Diabetes Complications 2010;24(6):368-374. [doi: 10.1016/j.jdiacomp.2009.10.001] [Medline: 19914095]
25. Oba T, Takano K, Katahira K, Kimura K. Use patterns of smartphone apps and wearable devices supporting physical activity and exercise: large-scale cross-sectional survey. JMIR Mhealth Uhealth 2023;11:e49148 [FREE Full text] [doi: 10.2196/49148] [Medline: 37997790]
26. Liu D, Maimaitijiang R, Gu J, Zhong S, Zhou M, Wu Z, et al. Using the unified theory of acceptance and use of technology (UTAUT) to investigate the intention to use physical activity apps: cross-sectional survey. JMIR Mhealth Uhealth 2019;7(9):e13127 [FREE Full text] [doi: 10.2196/13127] [Medline: 31507269]
27. Rø Ø, Reas DL, Rosenvinge J. The impact of age and BMI on Eating Disorder Examination Questionnaire (EDE-Q) scores in a community sample. Eat Behav 2012;13(2):158-161. [doi: 10.1016/j.eatbeh.2011.12.001] [Medline: 22365803]
28. Elavsky S, Smahel D, Machackova H. Who are mobile app users from healthy lifestyle websites? Analysis of patterns of app use and user characteristics. Transl Behav Med 2017;7(4):891-901 [FREE Full text] [doi: 10.1007/s13142-017-0525-x] [Medline: 28929368]
29. Hahn SL, Hazzard VM, Larson N, Klein L, Loth KA, Neumark-Sztainer D. Correlates of weight-related self-monitoring application use during emerging adulthood in a population-based sample. Eat Weight Disord 2022;27(6):2107-2119 [FREE Full text] [doi: 10.1007/s40519-021-01349-4] [Medline: 35040079]
30. Linardon J, Messer M. My fitness pal usage in men: associations with eating disorder symptoms and psychosocial impairment. Eat Behav 2019;33:13-17. [doi: 10.1016/j.eatbeh.2019.02.003] [Medline: 30772765]
31. Norman CD, Skinner HA. eHEALS: the eHealth literacy scale. J Med Internet Res 2006;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
32. Escobar-Viera CG, Shensa A, Bowman ND, Sidani JE, Knight J, James AE, et al. Passive and active social media use and depressive symptoms among United States adults. Cyberpsychol Behav Soc Netw 2018;21(7):437-443. [doi: 10.1089/cyber.2017.0668] [Medline: 29995530]
33. Kroenke K, Spitzer RL, Williams JBW. The PHQ-15: validity of a new measure for evaluating the severity of somatic symptoms. Psychosom Med 2002;64(2):258-266. [doi: 10.1097/00006842-200203000-00008] [Medline: 11914441]
34. Kroenke K, Spitzer RL, Williams JBW, Löwe B. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. Gen Hosp Psychiatry 2010;32(4):345-359. [doi: 10.1016/j.genhosppsych.2010.03.006] [Medline: 20633738]

XSL•FO
RenderX

35. Grilo CM, Reas DL, Hopwood CJ, Crosby RD. Factor structure and construct validity of the eating disorder examination-questionnaire in college students: further support for a modified brief version. Int J Eat Disord 2015;48(3):284-289 [FREE Full text] [doi: 10.1002/eat.22358] [Medline: 25346071]
36. Fairburn CG, Beglin SJ. Assessment of eating disorders: interview or self-report questionnaire? Int J Eat Disord 1994;16(4):363-370. [Medline: 7866415]
37. van Zyl LE, Olckers C, Roll LC. The psychometric properties of the Grit-O scale within the Twente region in Netherlands: an ICM-CFA vs. ESEM approach. Front Psychol 2020;11:796 [FREE Full text] [doi: 10.3389/fpsyg.2020.00796] [Medline: 32457679]
38. Jordan SL, Ferris GR, Hochwarter WA, Wright TA. Toward a work motivation conceptualization of Grit in organizations. Group Organ Manag 2019;44(2):320-360. [doi: 10.1177/1059601119834093]
39. Nieminen P. Application of standardized regression coefficient in meta-analysis. BioMedInformatics 2022;2(3):434-458. [doi: 10.3390/biomedinformatics2030028]
40. Aranha M, James K, Deasy C, Heavin C. Exploring the barriers and facilitators which influence mHealth adoption among older adults: a literature review. Gerontechnology 2021;20(2):1-16. [doi: 10.4017/gt.2021.20.2.424.06]
41. Ahmad NA, Mat Ludin AF, Shahar S, Mohd Noah SA, Mohd Tohit N. Willingness, perceived barriers and motivators in adopting mobile applications for health-related interventions among older adults: a scoping review. BMJ Open 2022;12(3):e054561 [FREE Full text] [doi: 10.1136/bmjopen-2021-054561] [Medline: 35264349]
42. Wildenbos GA, Peute LW, Jaspers MWM. A framework for evaluating mHealth tools for older patients on usability. Stud Health Technol Inform 2015;210:783-787. [Medline: 25991261]
43. Wildenbos GA, Peute L, Jaspers M. Aging barriers influencing mobile health usability for older adults: a literature based framework (MOLD-US). Int J Med Inform 2018;114:66-75. [doi: 10.1016/j.ijmedinf.2018.03.012] [Medline: 29673606]
44. Wildenbos GA, Jaspers MWM, Schijven MP, Dusseljee-Peute LW. Mobile health for older adult patients: using an aging barriers framework to classify usability problems. Int J Med Inform 2019;124:68-77. [doi: 10.1016/j.ijmedinf.2019.01.006] [Medline: 30784429]
45. Ramdowar H, Khedo KK, Chooramun N. A comprehensive review of mobile user interfaces in mHealth applications for elderly and the related ageing barriers. Univ Access Inf Soc 2023;23(4):1613-1629. [doi: 10.1007/s10209-023-01011-z]
46. Cho J, Park D, Lee HE. Cognitive factors of using health apps: systematic analysis of relationships among health consciousness, health information orientation, eHealth literacy, and health app use efficacy. J Med Internet Res 2014;16(5):e125 [FREE Full text] [doi: 10.2196/jmir.3283] [Medline: 24824062]
47. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. J Med Internet Res 2012;14(1):e19 [FREE Full text] [doi: 10.2196/jmir.1619] [Medline: 22357448]
48. Lwin MO, Panchapakesan C, Sheldenkar A, Calvert GA, Lim LK, Lu J. Determinants of eHealth literacy among adults in China. J Health Commun 2020;25(5):385-393. [doi: 10.1080/10810730.2020.1776422] [Medline: 32552607]
49. Zakar R, Iqbal S, Zakar MZ, Fischer F. COVID-19 and health information seeking behavior: digital health literacy survey amongst university students in Pakistan. Int J Environ Res Public Health 2021;18(8):4009 [FREE Full text] [doi: 10.3390/ijerph18084009] [Medline: 33920404]
50. Kim S, Oh J. The relationship between e-Health literacy and health-promoting behaviors in nursing students: a multiple mediation model. Int J Environ Res Public Health 2021;18(11):5804 [FREE Full text] [doi: 10.3390/ijerph18115804] [Medline: 34071469]
51. Cho H, Han K, Park BK. Associations of eHealth literacy with health-promoting behaviours among hospital nurses: a descriptive cross-sectional study. J Adv Nurs 2018;74(7):1618-1627. [doi: 10.1111/jan.13575] [Medline: 29575085]
52. Zhang L, Qin Y, Li P. Media complementarity and health information acquisition: a cross-sectional analysis of the 2017 HINTS-China survey. J Health Commun 2020;25(4):291-300. [doi: 10.1080/10810730.2020.1746868] [Medline: 32255740]
53. Moungui HC, Nana-Djeunga HC, Anyiang CF, Cano M, Ruiz Postigo JA, Carrion C. Dissemination strategies for mHealth apps: systematic review. JMIR Mhealth Uhealth 2024;12:e50293 [FREE Full text] [doi: 10.2196/50293] [Medline: 38180796]
54. Ceci L. Leading health and fitness apps worldwide in 2022, by downloads. Statista. 2024 Feb 5. URL: https://www.statista.com/statistics/1284875/global-top-health-and-fitness-apps/ [accessed 2025-01-07]
55. El-Gayar O, Elnoshokaty A. Factors and design features influencing the continued use of wearable devices. J Healthc Inform Res 2023;7(3):359-385. [doi: 10.1007/s41666-023-00135-4] [Medline: 37637719]
56. Malik VS, Hu FB. The role of sugar-sweetened beverages in the global epidemics of obesity and chronic diseases. Nat Rev Endocrinol 2022;18(4):205-218 [FREE Full text] [doi: 10.1038/s41574-021-00627-6] [Medline: 35064240]
57. Anthes E. Mental health: there's an app for that. Nature 2016;532(7597):20-23. [doi: 10.1038/532020a] [Medline: 27078548]
58. Maresova P, Javanmardi E, Barakovic S, Barakovic Husic J, Tomsone S, Krejcar O, et al. Consequences of chronic diseases and other limitations associated with old age – a scoping review. BMC Public Health 2019;19(1):1431 [FREE Full text] [doi: 10.1186/s12889-019-7762-5] [Medline: 31675997]
59. Britt RK, Collins WB, Wilson K, Linnemeier G, Englebert AM. eHealth literacy and health behaviors affecting modern college students: a pilot study of issues identified by the American College Health Association. J Med Internet Res 2017;19(12):e392 [FREE Full text] [doi: 10.2196/jmir.3100] [Medline: 29258979]

XSL•FO
RenderX

60. Guo Z, Zhao SZ, Guo N, Wu Y, Weng X, Wong JY, et al. Socioeconomic disparities in eHealth literacy and preventive behaviors during the COVID-19 pandemic in Hong Kong: cross-sectional study. J Med Internet Res 2021;23(4):e24577 [FREE Full text] [doi: 10.2196/24577] [Medline: 33784240]
61. Zhang L, Jung EH. WeChatting for health: an examination of the relationship between motivations and active engagement. Health Commun 2019;34(14):1764-1774. [doi: 10.1080/10410236.2018.1536942] [Medline: 30358416]
62. Duncan M, Murawski B, Short CE, Rebar AL, Schoeppe S, Alley S, et al. Activity trackers implement different behavior change techniques for activity, sleep, and sedentary behaviors. Interact J Med Res 2017;6(2):e13 [FREE Full text] [doi: 10.2196/ijmr.6685] [Medline: 28807889]
63. Hong H. An extension of the extended parallel process model (EPPM) in television health news: the influence of health consciousness on individual message processing and acceptance. Health Commun 2011;26(4):343-353. [doi: 10.1080/10410236.2010.551580] [Medline: 21416420]
64. Pu B, Zhang L, Tang Z, Qiu Y. The relationship between health consciousness and home-based exercise in China during the COVID-19 pandemic. Int J Environ Res Public Health 2020;17(16):5693 [FREE Full text] [doi: 10.3390/ijerph17165693] [Medline: 32781751]
65. Himmerich H, Mirzaei K. Body image, nutrition, and mental health. Nutrients 2024;16(8):1106 [FREE Full text] [doi: 10.3390/nu16081106] [Medline: 38674797]
66. Chao H, Lao I, Hao L, Lin C. Association of body image and health beliefs with health behaviors in patients with diabetes: a cross-sectional study. Diabetes Educ 2012;38(5):705-714. [doi: 10.1177/0145721712452796] [Medline: 22814357]
67. Eikey EV. Unintended users, uses, and consequences of mobile weight loss apps: using eating disorders as a case study. In: Sezgin E, Yildirim S, Özkan-Yildirim S, Sumuer E, editors. Current and Emerging mHealth Technologies: Adoption, Implementation, and Use. Cham, Switzerland: Springer International Publishing; 2018:119-133.
68. Eikey EV. Effects of diet and fitness apps on eating disorder behaviours: qualitative study. BJPsych open 2021;7(5):e176. [doi: 10.1192/bjo.2021.1011]
69. Eikey EV, Reddy MC, Booth KM, Kvasny L, Blair JL, Li V, et al. Desire to be underweight: exploratory study on a weight loss app community and user perceptions of the impact on disordered eating behaviors. JMIR Mhealth Uhealth 2017;5(10):e150 [FREE Full text] [doi: 10.2196/mhealth.6683] [Medline: 29025694]
70. Lin J, Dutta MJ. A replication of channel complementarity theory among internet users in India. Health Commun 2017;32(4):483-492. [doi: 10.1080/10410236.2016.1140268] [Medline: 27301884]
71. Jiang X, Wang L, Leng Y, Xie R, Li C, Nie Z, et al. The level of electronic health literacy among older adults: a systematic review and meta-analysis. Arch Public Health 2024;82(1):204 [FREE Full text] [doi: 10.1186/s13690-024-01428-9] [Medline: 39511667]
72. Bauman AE, Reis RS, Sallis JF, Wells JC, Loos RJF, Martin BW, Lancet Physical Activity Series Working Group. Correlates of physical activity: why are some people physically active and others not? Lancet 2012;380(9838):258-271. [doi: 10.1016/S0140-6736(12)60735-1] [Medline: 22818938]
73. Nelson ME, Rejeski WJ, Blair SN, Duncan PW, Judge JO, King AC, et al. Physical activity and public health in older adults: recommendation from the American College of Sports Medicine and the American Heart Association. Med Sci Sports Exerc 2007;39(8):1435-1445. [doi: 10.1249/mss.0b013e3180616aa2] [Medline: 17762378]
74. Jefferis BJ, Sartini C, Lee I, Choi M, Amuzu A, Gutierrez C, et al. Adherence to physical activity guidelines in older adults, using objectively measured physical activity in a population-based study. BMC Public Health 2014;14:382 [FREE Full text] [doi: 10.1186/1471-2458-14-382] [Medline: 24745369]
75. Cunningham C, O'Sullivan R, Caserotti P, Tully MA. Consequences of physical inactivity in older adults: a systematic review of reviews and meta-analyses. Scand J Med Sci Sports 2020;30(5):816-827. [doi: 10.1111/sms.13616] [Medline: 32020713]
76. Sohaib Aslam A, van Luenen S, Aslam S, van Bodegom D, Chavannes NH. A systematic review on the use of mHealth to increase physical activity in older people. Clin eHealth 2020;3:31-39. [doi: 10.1016/j.ceh.2020.04.002]
77. Peat CM, Peyerl NL, Muehlenkamp JJ. Body image and eating disorders in older adults: a review. J Gen Psychol 2008;135(4):343-358. [doi: 10.3200/GENP.135.4.343-358] [Medline: 18959226]
78. Paxton SJ, Phythian K. Body image, self-esteem, and health status in middle and later adulthood. Aust Psycho 2007;34(2):116-121. [doi: 10.1080/00050069908257439]
79. Chatelan A, Carrard I. Diet quality in middle-aged and older women with and without body weight dissatisfaction: results from a population-based national nutrition survey in Switzerland. J Nutr Sci 2021;10:e38 [FREE Full text] [doi: 10.1017/jns.2021.32] [Medline: 34367623]
80. Glanz K. Health behavior and risk factors. In: Quah SR, editor. International Encyclopedia of Public Health. 3rd ed. New York: Academic Press; 2025:175-181.
81. Gulec H, Smahel D, Huang Y. Open Science Framework. 2024 Dec 23. URL: https://osf.io/ubazy/?view_only=ab5bbaad18824c248d6227fb02b1653340 [accessed 2026-01-08]

XSL•FO
RenderX

Abbreviations

CFA: confirmatory factor analysis
EFA: exploratory factor analysis
ESEM: exploratory structural equation modeling
mHealth: mobile health
NUTS3: Nomenclature of Territorial Units for Statistics
OR: odds ratio
SNS: social networking site
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani; submitted 22.Jan.2025; peer-reviewed by A Eisingerich; comments to author 03.Sep.2025; revised version received 17.Nov.2025; accepted 30.Dec.2025; published 30.Jan.2026.

Please cite as:
Gulec H, Smahel D, Huang Y
Patterns and Characteristics of Mobile App Use to Promote Wellness and Manage Illness: Cross-Sectional Study
J Med Internet Res 2026;28:e71363
URL: https://www.jmir.org/2026/1/e71363
doi:10.2196/71363
PMID:

©Hayriye Gulec, David Smahel, Yi Huang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effects of an 8-Week App-Based Mindfulness Intervention on Mental Health in Working Women: Randomized Controlled Trial

Riko Uwagawa[1*], MA; Koichiro Adachi[1*], MA; Mariko Shimoda[1*], MA; Ryu Takizawa[1,2*], MD, PhD

[1]Department of Clinical Psychology, Graduate School of Education, The University of Tokyo, Tokyo, Japan

[2]MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[*]all authors contributed equally

Corresponding Author:
Ryu Takizawa, MD, PhD
Department of Clinical Psychology
Graduate School of Education
The University of Tokyo
7-3-1, Hongo, Bunkyo-ku
Tokyo, 1130033
Japan
Phone: 81 3 5841 1397
Email: takizawar-tky@umin.ac.jp

Abstract

Background: Although working women experience increased work-related stress, preventive interventions to reduce its negative effects on their mental health are insufficient.

Objective: This study evaluated the effectiveness of an 8-week mindfulness-based self-help intervention via a smartphone app across 4 domains (general psychological, work-related, family-related, and work-to-conflict) among working women.

Methods: This study recruited women workers via various media sources, such as crowdsourcing sites and social networking services. Participants were randomly assigned to the intervention (n=106) or waitlist control groups (n=107). Participants in the intervention group practiced guided mindfulness meditation every day at their convenience via an app on their cell phones for 8 weeks. The app provides an 8-week program with 4 meditation contents per 2 weeks. Participants in the waitlist control group lived as usual for 8 weeks. We conducted web-based questionnaires to assess participants' general psychological (life satisfaction, perceived stress, depressive and anxiety symptoms, trait anger, and mindfulness), work-related (work performance, job satisfaction, quantitative job overload, and job control), family-related (family satisfaction and partner satisfaction), and work-to-family conflict indicators.

Results: An analysis of covariance, controlled for preintervention scores, revealed that the intervention significantly increased life satisfaction (b=1.47, β=0.11; P=.005) and decreased perceived stress (b=–2.00, β=–0.17; P=.01), depressive and anxiety symptoms (b=–1.24, β=–0.15; P=.02), and trait anger (reaction; b=–0.59, β=–0.11; P=.04). The intervention group demonstrated significantly increased life satisfaction (t_{93}=–3.36; P=.001) and decreased depressive and anxiety symptoms (t_{93}=2.35; P=.02).

Conclusions: The app was effective in reducing perceived stress, depressive and anxiety symptoms, and trait anger (reaction), and in improving life satisfaction among working women. However, to improve work- and family-related indicators, higher-intensity interventions may be required, such as modifying the intervention content or extending its duration.

Trial Registration: University Hospital Medical Information Network Clinical Trials Registry (UMIN-CTR) UMIN000051796; https://center6.umin.ac.jp/cgi-open-bin/ctr_e/ctr_view.cgi?recptno=R000059110

(J Med Internet Res 2026;28:e62814) doi:10.2196/62814

KEYWORDS

mindfulness; mobile apps; randomized controlled trial; women's health; mental health; subjective well-being; health promotion; mHealth; application; applications; work-related stress; stress; intervention; interventions; women; mobile phone

XSL•FO
RenderX

Introduction

The impact of work-related stress on workers' mental health has been recently investigated, and its significant social impact has become an issue [1]. According to the World Health Organization, work-related stress refers to "the response people may have when presented with work demands and pressures that are not matched to their knowledge and abilities and which challenge their ability to cope" [2].

Working women experience increased work-related stress compared with working men. The American Psychological Association found that women consistently exhibited higher levels of stress than men and had additional difficulty in coping [3]. Furthermore, women are more likely to develop stress-related symptoms owing to neurobiological differences, a sense of burden from the dual roles of balancing work and family, and exposure to job insecurity [4-6]. Work-family conflict of working women has a negative impact on their stress and on their physical and mental health, and a framework regarding the relationship between these is presented [7]. Work-related factors may affect women and men differently, with women possibly being further affected owing to their work and family roles. With the global aim of gender parity in the labor market [8], the number of women in the working population is expected to increase. Therefore, preventive interventions to reduce the negative effects of work-related stress on women's mental health are required. However, such support is insufficient [9,10].

Traditionally, psychiatry has focused on the treatment of mental disorders rather than prevention. However, mental health is more than the absence of mental illness [10,11]. Therefore, interventions that focus on preventing mental health problems among women workers before they worsen and improving positive aspects, such as life satisfaction, could have positive effects on women's well-being and their work, family, and society as a whole.

Mindfulness meditation is an effective intervention strategy for improving mental health and well-being. Mindfulness is the awareness that emerges from deliberate, nonjudgmental attention to experiences as they unfold moment-by-moment [12]. As mindfulness-based interventions reduce symptoms of depression, anxiety, and perceived stress and improve sleep quality and well-being [13-15], they are attracting attention as a preventive intervention strategy.

Additionally, the effectiveness of mindfulness meditation provided by smartphone apps has been recently highlighted. According to the International Telecommunication Union, there are over 8.89 billion mobile subscriptions worldwide [16]. Therefore, mobile technology can be used to provide preventive health care interventions to numerous people.

A traditional mindfulness-based program is high-intensity (8 weekly sessions of 2.5 hours per session and 30-40 minutes of practice per day) and time-constrained, which creates a participation barrier for nonclinical working women. The mobile-based mindfulness intervention is an app-based, voice-guided meditation practice that allows users to practice at their own convenience, which offers the advantages of high convenience and low cost [17-19]. Furthermore, online mindfulness interventions are effective in improving depression, anxiety, stress, rumination, and well-being [20,21].

However, no studies have examined the effects of mobile-based mindfulness meditation on working women from work, family, and work–family conflict aspects, as well as general measures. To our knowledge, only two studies have examined the effectiveness of mobile-based mindfulness meditation among working women. Santos et al [10] found that an app-based mindfulness and positive psychology intervention effectively reduced perceived stress and anxiety symptoms in working women. Coelhoso et al [11] revealed that a well-being mobile app designed to handle psychological stress based on relaxation training, breathing techniques, meditation (mindfulness, loving meditation, such as mindfulness, loving, kindness, and empathetic joy), and positive psychology principles improved working women's work-related well-being and reduced their work-related and overall stress.

Conversely, no study has examined working women's well-being from the 4 aspects of general psychological, work-related, family-related, and work-family conflict indicators. Examining their effects is essential for the future applications of mindfulness meditation, as it will help us comprehensively understand how mindfulness meditation works for women workers.

Therefore, this study aimed to evaluate the effectiveness of an 8-week mindfulness meditation intervention via a smartphone app among women workers through a randomized controlled trial (RCT). Effectiveness was examined via 4 indicators: general psychological, work-related, family-related, and work-to-conflict measures. Furthermore, we examined the measures that would effectively influence. We hypothesized that participants in the intervention group (self-care mindfulness meditation via the smartphone app) would have a higher level of general psychological (life satisfaction, perceived stress, depressive and anxiety symptoms, trait anger, and mindfulness), work-related (work performance, job satisfaction, quantitative job overload, and job control), family-related (family satisfaction and partner satisfaction), and work-to-family conflict indicators compared with those in the waitlist control group.

Methods

Participants

A power analysis was conducted to determine the sample size needed for this study (significance=.05; statistical power=.8; effect size=0.4), and a sample size of 100 participants per group, for a total of 200 participants, was needed. The effect size demonstrated in the meta-analysis of the effects of online mindfulness-based interventions on mental health was used as reference (depression Hedges g=0.34; stress Hedges g=0.44) [21].

This study recruited 397 women workers via various media sources, such as crowdsourcing sites and social networking services. Inclusion criteria included those who were (1) biologically female, (2) employed for at least 20 hours per week,

XSL•FO
RenderX

(3) owned an iPhone (for convenience of the app used), and (4) aged 18-64 years. Exclusion criteria included those who (1) received treatment for a mental disorder, (2) scored ≥13 on the 6-item Kessler Psychological Distress Scale (K6) Japanese version, (3) were on leave, and (4) were currently pregnant or likely to become pregnant within six months. Among the participants, 95 did not meet the inclusion and exclusion criteria. Hence, 302 women workers who met the criteria were asked to respond to the preintervention assessment, and 215 who completed the assessment were randomly assigned to the intervention (n=107) or waitlist control group (n=108). Randomization was computerized using a blocked randomization scheme (block size 10). A total of 8 working women dropped out. Of the 8 participants, 2 participants (intervention group, n=1; wait-list control group, n=1) declined to participate in this study, 2 participants in the intervention group opted out of the intervention, and 4 participants (intervention group, n=2; wait-list control group, n=2) could not be contacted. After 8 weeks, the participants were asked to respond to the postintervention assessment, and 196 women workers completed the assessment (intervention group, n=95; waitlist control group, n=101). Of 215 participants who completed the preintervention assessment, 4 who worked <19 hours per week on average in the preintervention assessment were excluded from analysis (intervention group, n=1; waitlist control group, n=3). Therefore, of the 215 participants who were randomized, data from 209 participants (intervention group, n=105; waitlist control group, n=104) were finally analyzed, excluding 2 participants who declined to participate in this study and 4 participants who worked <19 hours per week on average in the preintervention assessment.

Procedure

Overview

This study was designed as a parallel-design RCT. Randomization was computerized independently by research staff using a blocked randomization scheme (block size 10). Participants were expected to be randomized in a ratio of 1:1 to the intervention or waitlist control group. This study was an open-label RCT as it was not possible to blind the allocation.

This study was conducted from July 2023 to January 2024 via web forms. Participants in the intervention group installed the app for meditation after the preintervention assessment. Participants practiced guided mindfulness meditation via the app on their cell phones every day at their convenience for 8 weeks. After 8 weeks, the participants received the postintervention questionnaire via the app and email. The participants in the waitlist control group lived as usual for 8 weeks after the preintervention assessment. After 8 weeks, they also responded to the postintervention questionnaire via email.

This study was registered in the University Hospital Medical Information Network (UMIN) Clinical Trials Registry (UMIN000051796).

8-Week Mindfulness-Based Self-Help Intervention via the Smartphone App

Mindfulness meditation was conducted via the iOS app, with the content changed every 2 weeks (Table 1). The app displayed the day's meditation content and explanation on the home screen. After viewing this screen, the participants pressed the play button to hear the guided audio and practiced meditation. Figure 1 illustrates the display of the app. In addition, the psychoeducation pages on mindfulness and self-compassion were created and inserted on the app (Figure 2).

Table 1. Content of the 8-week self-help mindfulness-based meditation.

Week	Types of meditation	Duration (minutes)
1 and 2	Meditation of breath	7
3 and 4	Body scan	7
5 and 6	Meditation of breath, sound, and body	12
7 and 8	Loving-kindness meditation	12

XSL•FO
RenderX

Figure 1. Display of the smartphone app.

Figure 2. Display of the psychoeducation.

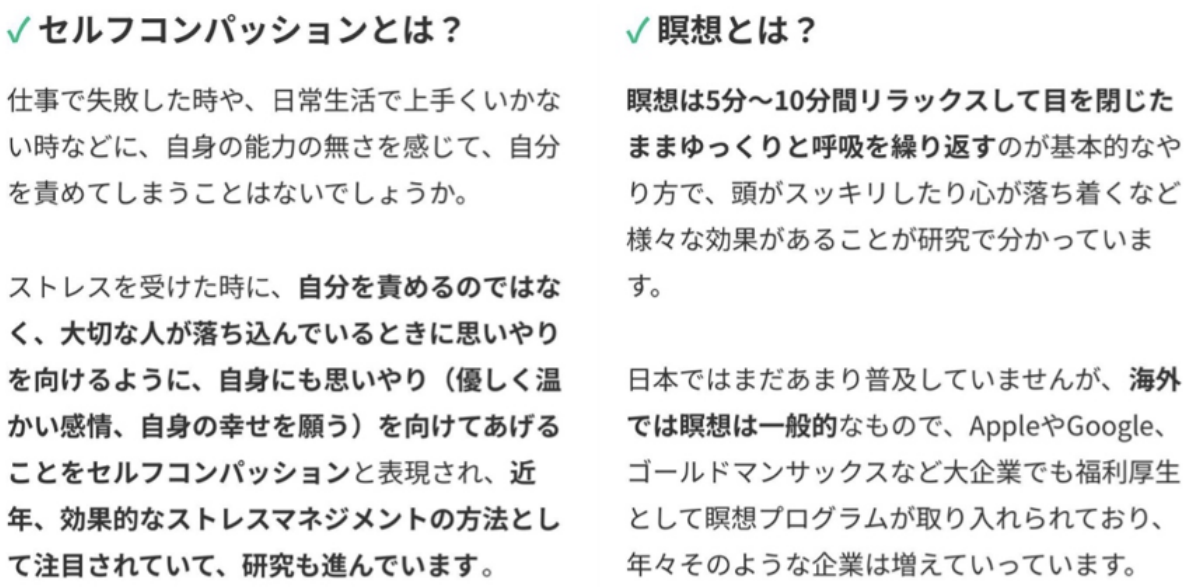

The content included “meditation of breath,” “meditation of breath, sound, and body,” and “body scan meditation,” based on previous studies [22]. As the “body scan” was partially included in “meditation of breath, sound, and body,” in this study, the latter was conducted after the former. Furthermore, as the effectiveness of interventions that incorporated elements of self-compassion was recently highlighted, “loving-kindness meditation” was ultimately added. As a daily 13-minute meditation was effective after 8 weeks [23], the intervention period was designed to be 8 weeks.

Measurements

General Psychological Domain: Well-Being

Well-being was assessed as life satisfaction using the 5-item Satisfaction with Life Scale. Participants evaluated their subjective life satisfaction on a 7-point Likert scale that ranged from 1 (strongly disagree) to 7 (strongly agree) [24,25]. This measurement was developed by Diener et al [24]. The development of the Japanese version used in this study and its validity and reliability were studied by Sumino [25]. Sample items included “In most ways, my life is close to my ideal.” The total score was a sum of all the individual item scores, and

XSL•FO
RenderX

higher scores indicated greater life satisfaction. In this study, Cronbach α was 0.85 and 0.81 for the pre- and postintervention assessments, respectively.

Mental Health Outcomes

Perceived Stress

The 10-item Perceived Stress Scale was used to assess perceived stress. Participants rated how unpredictable, uncontrollable, and overloaded they found their lives on a 5-point Likert scale that ranged from 0 (never) to 4 (very often) [26,27]. This measurement was developed by Cohen et al [26]. The development of the Japanese version used in this study and its validity and reliability were studied by Sumi [27]. Sample items included "How often have you been upset because of something that happened unexpectedly?" The total score was a sum of the individual item scores, and higher scores indicated greater perceived stress. Cronbach α was 0.69 and 0.79 for pre- and postintervention, respectively.

Depressive and Anxiety Symptoms

K6 was used to assess depression and anxiety symptoms. Participants described how often they experienced depressive symptoms in the past 30 days on a 5-point Likert scale that ranged from 0 (none of the time) to 4 (all of the time) [28-30]. This measurement was developed by Kessler et al [28], and the Japanese version of it used in this study was developed by Furukawa et al [29]. The validity and reliability were studied by Furukawa et al [29] and by Sakurai et al [30]. Sample items included "How often did you feel nervous?" and "How often did you feel restless or fidgety?" The total score was a sum of all the individual item scores, and higher scores indicated a greater severity of depression and anxiety. Cronbach α was 0.83 and 0.80 for pre- and postintervention, respectively.

Trait Anger

"Trait anger (T-Ang; 10-item)," a subscale of the 57-item State-Trait Anger Expression Inventory 2 (STAXI-2), was used to assess the traits of anger reaction [31-33]. This measurement was developed by Spielberger [31]. The development of the Japanese version used in this study and its reliability were studied by Mine and Ohki [32] and by Mine and Sato [33]. Participants evaluated their perceptions of anger proneness on a 4-point Likert scale that ranged from 1 (strongly disagree) to 4 (strongly agree). Sample items included "I am quick-tempered." T-Ang included two subfactors: T-Ang/Temperament (T-Ang/T; trait of feeling anger with or without stimulus) and T-Ang/Reaction (T-Ang/R; frequency of experiencing feelings of anger in situations involving irritation or negative evaluation). The total score within each subfactor and all items was calculated by summing the item scores. Higher scores indicated greater trait anger. Cronbach α for preintervention was T-Ang Cronbach α=0.84, T-Ang/T Cronbach α=0.79, and T-Ang/R Cronbach α=0.77, and for postintervention was T-Ang Cronbach α=0.83, T-Ang/T Cronbach α=0.84, and T-Ang/R Cronbach α=0.76.

Mindfulness

The 15-item Mindful Attention Awareness Scale was used to assess dispositional mindfulness [34,35]. This measurement was developed by Brown and Ryan [34]. The development of the Japanese version used in this study and its validity and reliability were studied by Fujino et al [35]. Participants rated the degree to which they functioned without awareness of the present experience in daily life on a 6-point scale that ranged from 1 (almost never) to 6 (almost always). Sample items included "I could be experiencing some emotion and not be conscious of it until sometime later." All items were reversed as they assessed the lack of mindful attention and awareness. The total score was a sum of all the reversed-item scores, and higher scores indicated greater mindful attention and awareness. Cronbach α was 0.81 and 0.87 for pre- and postintervention, respectively.

Work-Related Domain

Work Performance

The World Health Organization Health and Work Performance Questionnaire Short Form was used to assess work performance. The questions included: "On a scale of 0-10, where 0 is the worst job performance anyone could have at your job, and 10 is the performance of a top worker, how would you rate the usual performance of most workers in a job similar to yours?" (possible performance) and "Using the same 0-10 scale, how would you rate your overall job performance on the days you worked during the past four weeks?"(actual performance) [36-38]. This measurement was developed by Kessler et al [36]. The development of the Japanese version used in this study and its validity and reliability were studied by Kawakami et al [38]. Participants evaluated the workplace costs of health problems regarding self-reported sickness leaves and reduced job performance (presenteeism). Presenteeism was assessed by "absolute" and "relative presenteeism." "Absolute presenteeism" was calculated by multiplying the score of actual performance by 10. Higher scores indicated greater performance. "Relative presenteeism" was calculated by the ratio of actual performance to possible performance (restricted to the range of 0.25–2.0, where values <0.25 and >2.0 were converted to 0.25 and 2.0, respectively). Higher scores indicated greater performance.

Job Satisfaction

Job satisfaction was assessed via a single item from the Brief Job Stress Questionnaire (BJSQ) [39]. The development of this measurement used in this study and its validity and reliability were studied by Inoue et al [39]. Participants rated the degree to which they agreed with the item, "I am satisfied with my job," on a 4-point Likert scale that ranged from 1 (satisfied) to 4 (dissatisfied). The item was reversed as it assessed the high level of job satisfaction. Higher scores indicated greater job satisfaction.

Quantitative Job Overload

"Quantitative job overload (3-item)," a subscale of the BJSQ, was used to assess job overload [39]. Participants rated the degree of their job overload on a 4-point Likert scale that ranged from 1 (agree) to 4 (disagree). Sample items included "I have a lot of work to do." All items were reversed as they assessed the high level of job overload. The total score was a sum of all the reversed-item scores, and higher scores indicated a greater

XSL•FO
RenderX

job overload. Cronbach α was 0.64 and 0.56 for pre- and postintervention, respectively.

Job Control

"Job control (3-item)," a subscale of the BJSQ, was used to assess job control [39]. Participants rated the degree of their job control on a 4-point Likert scale that ranged from 1 (agree) to 4 (disagree). Sample items included "I can work at my own pace." All items were reversed as they assessed the high level of job control. The total score was a sum of all the reversed-item scores, and higher scores indicated a greater sense of job control. Cronbach α was 0.60 and 0.64 for pre- and postintervention, respectively.

Family-Related Domain

Family Satisfaction

Family satisfaction was assessed via a single item from the BJSQ [39]. Participants rated the degree to which they agreed with the item, "I am satisfied with my family life," on a 4-point Likert scale that ranged from 1 (satisfied) to 4 (dissatisfied). The item was reversed as it assessed the high level of family satisfaction. Higher scores indicated greater family satisfaction.

Partner Satisfaction

Partner satisfaction was assessed via a single item: "Using the 10-point scale, how would you rate your current level of satisfaction with your relationship with your partner?" Only participants who lived with their partners were asked to respond. Participants rated the degree of their satisfaction with their partner on a 10-point Likert scale that ranged from 1 (dissatisfied) to 10 (satisfied). Higher scores indicated greater satisfaction.

Work-to-Family Conflict Domain

The 22-item Survey Work-Home Interaction-Nijmegen was used to assess the 4 subscales that reflected the underlying dimensions of work–family spillover: (1) work-family negative spillover (WFNS, 8 items; eg, "You do not have the energy to engage in leisure activities with your spouse/family/friends because of your job."), (2) family-work negative spillover (FWNS, 4 items; eg, "You do not feel like working because of problems with your spouse/family/friends."), (3) work-family positive spillover (WFPS, 5 items; eg, "You fulfill your domestic obligations better because of the things you have learned on your job."), (4) family-work positive spillover (FWPS, 5 items; eg, "You have greater self-confidence at work because you have your home life well organized") [40,41]. This measurement was developed by Geurts et al [40] in 2005. The development of the Japanese version used in this study and its validity and reliability were studied by Shimada et al [41]. Responses were rated on a 4-point Likert scale that ranged from 0 (never) to 3 (always). The total score of each subscale was calculated as a sum of all the individual item scores. Higher scores on the positive (WFPS and FWPS) and negative spillover subscales (WFNS and FWNS) indicated greater positive and negative impacts, respectively. For preintervention, the Cronbach α were WFNS Cronbach α=0.88, FWNS Cronbach α=0.79, WFPS Cronbach α=0.73, and FWPS Cronbach α=0.78, and for postintervention, it was WFNS Cronbach α=0.89, FWNS Cronbach α=0.76, WFPS Cronbach α=0.79, and FWPS Cronbach α=0.83.

Statistical Analysis

We conducted Chi-squared, *t* tests, and the Fisher exact test in order to examine whether there are differences in demographic variables and psychological indices between the intervention and control groups. Subsequently, we conducted 2-tailed *t* tests to examine whether there were differences in demographic variables and psychological indices of participants in the intervention and waitlist control groups, respectively.

For the intervention effects, we conducted an analysis of covariance (ANCOVA; independent variables: intervention group=1 and waitlist control group=0) that used the least squares method as an estimation method, controlled for preintervention scores. We conducted an ANCOVA that used the least squares estimation method, controlled for preintervention scores, age, employment status (regular employment: employed full time with no fixed term of employment; nonregular employment: not regular employment), psychiatric history, education, and marital status. In this study, the participants were randomly assigned to the intervention and control groups. However, because of the possibility that the intervention effect might not be properly detected due to group differences in preintervention scores and demographic data, we controlled for them. Additionally, paired *t* tests were conducted to determine any differences in the pre- and postintervention assessments within each group. An intention-to-treatment analysis was used. R (version 4.3.2; R Foundation for Statistical Computing) was used for statistical analysis.

Ethical Considerations

This study was approved by the Life Science Research Ethics and Safety Committee, the University of Tokyo (23-144, 23-227, and 24-020).

Results

Baseline

Figure 3 illustrates the CONSORT (Consolidated Standards for Reporting Trials) flow diagram (the CONSORT checklist is provided in Multimedia Appendix 1).

Table 2 shows the participants' demographic characteristics. Chi-squared and *t* tests revealed no differences in demographic variables and psychological indices between the intervention and waitlist control groups ($P>.05$).

XSL•FO
RenderX

Figure 3. CONSORT (Consolidated Standards for Reporting Trials) flowchart for participants.

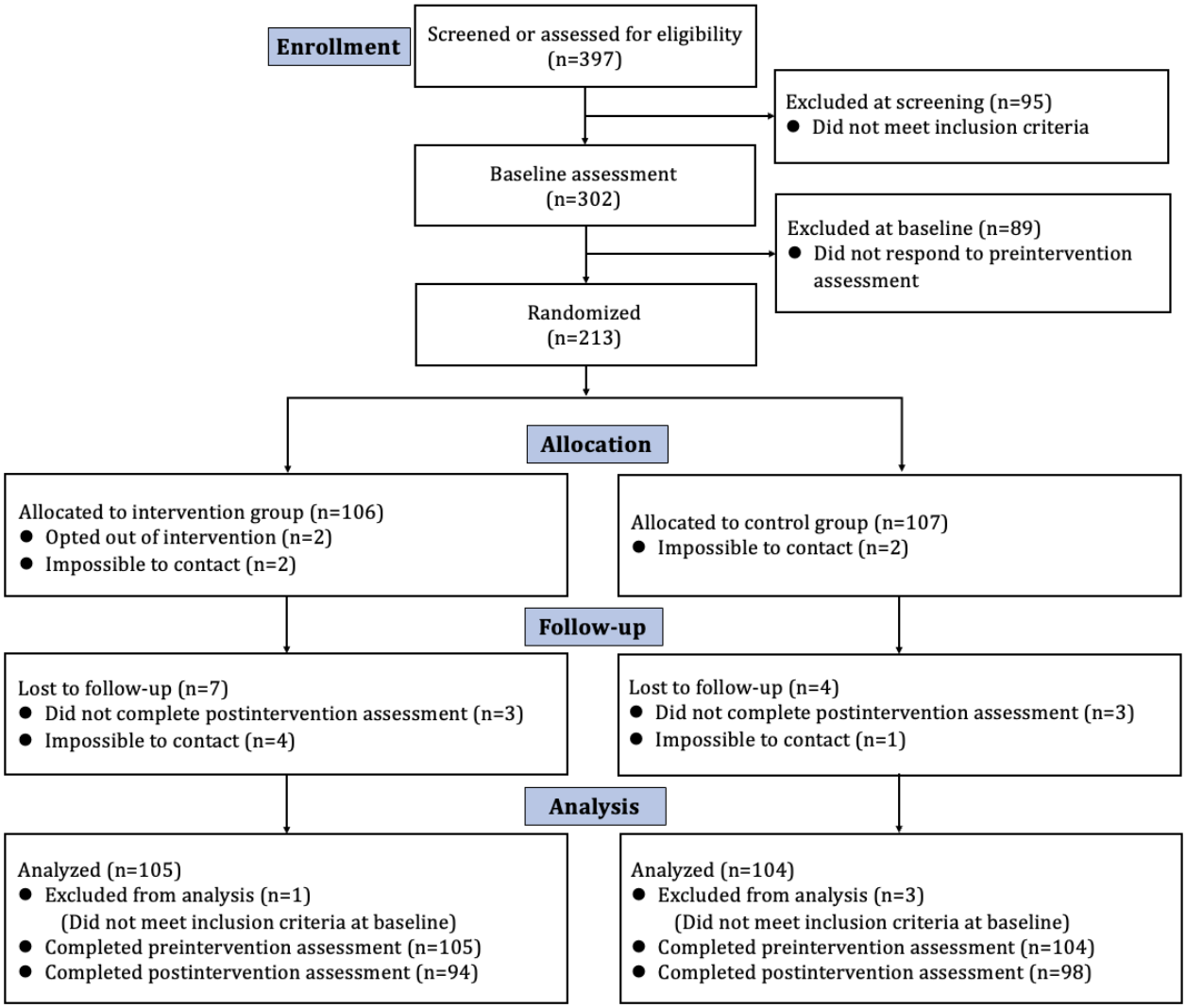

XSL•FO
RenderX

Table 2. Participants' demographic information.

Participant characteristics	Intervention group (n=105)	Waitlist control group (n=104)	Difference statistic: t test (*df*) or chi-square (*df*)	*P* value
Age (years), mean (SD)	36.81 (10.82)	36.81 (10.70)	0.0 (207)[a]	.99
Education level, n (%)			Fisher exact test	.09
Less than a bachelor's degree	46 (43.6)	31 (29.8)		
Bachelor's degree	48 (45.7)	60 (57.7)		
Master's degree	11 (10.6)	11 (10.6)		
Doctoral degree	0 (0)	2 (1.9)		
Marital status, n (%)			Fisher exact test	.44
Married	46 (46.8)	50 (48.1)		
Single	51 (44.7)	42 (40.4)		
Divorced	6 (6.4)	11 (10.6)		
Widowed	2 (2.1)	1 (1)		
Employment status[b], n (%)			2.1 (1)[c]	.14
Regular employment	66 (62.9)	54 (51.9)		
Nonregular employment	39 (37.1)	50 (48.1)		
Psychiatric history[d], n (%)			2.6 (1)[c]	.11
Yes	7 (6.7)	15 (14.4)		
No	98 (93.3)	89 (85.6)		
Living with a partner, n (%)			0.0 (1)[c]	.83
Yes	52 (49.5)	54 (51.9)		
No	53 (50.5)	50 (48.1)		
Youngest child age in years, n (%)			2.1 (3)[c]	.56
0-2	8 (7.6)	5 (4.8)		
3-18	30 (28.6)	30 (28.8)		
19+	4 (3.8)	8 (7.7)		
None	63 (6)	61 (58.7)		

[a]*t* test (*df*).

[b]Employment status indicates whether the individual is a regular employee.

[c]Chi-square (*df*).

[d]Psychiatric history indicates whether the individual has a history of visiting a psychosomatic medicine or psychiatric clinic.

Comparing Completers and Dropouts Within Each Group

In the intervention group, no statistically significant differences were observed in demographic information and psychological measurement scores between the dropouts and participants who completed the postintervention assessment ($P>.05$). In the waitlist control group, there were differences in age ($t_{102}=2.66$; $P=.009$), T-Ang ($t_{102}=-2.22$; Cohen $d=0.93$; $P=.03$), T-Ang/R ($t_{102}=-2.26$; Cohen $d=0.95$; $P=.03$), and job control ($t_{102}=-2.57$; Cohen $d=1.08$; $P=.01$). Dropouts were significantly younger (mean 25.83, SD 4.62), more angry (T-Ang: mean 24.00, SD 9.06; T-Ang/R: mean 11.00, SD 4.65), and perceived an additional sense of job control (mean 9.17, SD 1.60) compared with the retained participants (age mean 37.48, SD 10.61; T-Ang: mean 18.76, SD 5.39; T-Ang/R: mean 8.34, SD 2.68; job control: mean 6.56, SD 2.45).

Practice Frequency

Participants in the intervention group used the app for a mean of 42.32 days (75.57%, SD 15.63) in 8 weeks.

Outcomes

Group Effects

Table 3 presents the scores of the pre- and postintervention assessments. Table 4 presents the results of ANCOVA. The ANCOVA, controlled for preintervention scores, revealed significant group effects on life satisfaction ($b=1.47$, $\beta=0.11$; $P=.005$), perceived stress ($b=-2.00$, $\beta=-0.17$; $P=.01$), depressive and anxiety symptoms ($b=-1.24$, $\beta=-0.15$; $P=.02$), and T-Ang/R

XSL•FO
RenderX

(b=–0.59, β=–0.11; P=.04). The ANCOVA, controlled for pre-intervention scores and demographic data (age, employment status, psychiatric history, education, marital status), revealed significant group effects on life satisfaction (b=1.35, β=0.10; P=.02), perceived stress (b=–1.91, β=–0.16; P=.02), depressive and anxiety symptoms (b=–1.13, β=–0.13; P=.03), and T-Ang/R (b=–0.71, β=–0.13; P=.02).

Table 3. Scores of the pre- and postintervention assessments.

	Intervention group				Waitlist control group				*t* test	
	Mean (SD)		*P* value	Cohen *d* (95% Cl)	Mean (SD)		*P* value	Cohen *d* (95% Cl)	*P* value	Cohen *d* (95% Cl)
	Pre	Post			Pre	Post				
General psychological domain										
Life satisfaction	17.70 (6.68)	18.80 (6.76)	.001	0.22 (0.09 to 0.35)	18.62 (6.89)	18.58 (6.32)	.38	0.05 (–0.06 to 0.15)	.82	0.03 (–0.25 to 0.31)
Perceived stress	18.71 (5.14)	18.65 (5.65)	.67	0.05 (–0.19 to 0.29)	17.62 (5.95)	19.94 (6.38)	<.001	0.40 (0.21 to 0.60)	.14	0.21 (–0.07 to 0.50)
Depressive and anxiety symptoms	6.16 (4.06)	5.37 (3.61)	.02	0.27 (0.04 to 0.50)	6.20 (4.89)	6.41 (4.71)	.33	0.08 (–0.08 to 0.25)	.09	0.25 (–0.04 to 0.53)
Trait anger	18.50 (5.16)	18.27 (5.31)	.17	0.10 (–0.04 to 0.23)	19.06 (5.73)	18.92 (5.58)	.70	0.03 (–0.12 to 0.18)	.41	0.12 (–0.17 to 0.40)
Trait anger (temperament)	7.09 (2.62)	6.94 (2.54)	.06	0.12 (–0.01 to 0.24)	7.49 (2.78)	7.30 (2.95)	.56	0.04 (–0.10 to 0.19)	.37	0.13 (–0.15 to 0.41)
Trait anger (reaction)	8.50 (2.65)	8.28 (2.71)	.24	0.10 (–0.07 to 0.27)	8.49 (2.86)	8.71 (2.74)	.07	0.14 (–0.01 to 0.29)	.27	0.16 (–0.12 to 0.44)
Mindfulness	43.36 (10.65)	44.05 (11.31)	.70	0.03 (–0.10 to 0.07)	43.74 (9.85)	44.33 (11.75)	.72	0.02 (–0.11 to 0.16)	.87	0.02 (–0.26 to 0.31)
Work										
Work performance										
Absolute presenteeism	61.43 (19.24)	62.02 (18.64)	.73	0.04 (–0.19 to 0.27)	61.06 (19.5)	61.84 (18.69)	.67	0.05 (–0.17 to 0.27)	.95	0.01 (–0.27 to 0.29)
Relative presenteeism	1.02 (0.32)	1.05 (0.27)	.50	0.09 (–0.17 to 0.34)	1.00 (0.33)	1.01 (0.32)	.67	0.05 (–0.19 to 0.30)	.35	0.14 (–0.15 to 0.42)
Job satisfaction	2.70 (0.72)	2.67 (0.79)	.90	0.01 (–0.21 to 0.24)	2.81 (0.87)	2.81 (0.83)	.53	0.05 (–0.10 to 0.20)	.25	0.17 (–0.12 to 0.45)
Quantitative job overload	8.10 (2.36)	8.11 (2.43)	.82	0.02 (–0.17 to 0.22)	7.68 (2.59)	7.85 (2.30)	.43	0.07 (–0.10 to 0.23)	.45	0.11 (–0.17 to 0.39)
Job control	8.36 (2.33)	8.54 (2.23)	.17	0.11 (–0.05 to 0.26)	8.29 (2.48)	8.40 (2.29)	.84	0.02 (–0.15 to 0.18)	.66	0.06 (–0.22 to 0.35)
Family										
Family satisfaction	3.00 (0.77)	3.06 (0.81)	.18	0.13 (–0.06 to 0.33)	2.99 (0.82)	2.96 (0.88)	.45	0.07 (–0.12 to 0.26)	.40	0.12 (–0.16 to 0.41)
Partner satisfaction	7.52 (2.14)	7.39 (2.25)	.82	0.02 (–0.15 to 0.19)	7.69 (2.05)	7.15 (2.43)	.02	0.25 (0.04 to 0.45)	.62	0.10 (–0.29 to 0.49)
Work-to-family conflict										
Work-family negative spillover	5.62 (4.92)	5.63 (4.48)	.84	0.02 (–0.14 to 0.17)	5.51 (5.51)	5.18 (5.23)	.20	0.08 (–0.04 to 0.20)	.53	0.09 (–0.19 to 0.37)
Family-work negative spillover	1.25 (1.71)	1.20 (1.72)	.64	0.05 (–0.16 to 0.26)	1.19 (1.66)	1.36 (1.85)	.33	0.09 (–0.09 to 0.28)	.55	0.09 (–0.20 to 0.37)
Work-family positive spillover	7.08 (3.04)	7.02 (3.05)	.92	0.01 (–0.18 to 0.20)	6.77 (3.30)	7.13 (3.72)	.34	0.08 (–0.09 to 0.25)	.82	0.03 (–0.25 to 0.32)
Family-work positive spillover	7.09 (3.51)	7.22 (3.55)	.66	0.04 (–0.13 to 0.21)	7.19 (3.94)	7.28 (4.06)	.98	0.00 (–0.17 to 0.16)	.92	0.01 (–0.27 to 0.30)

XSL•FO
RenderX

Table 4. Comparison between the control and the intervention groups.

	Controlling for prescores					Controlling for prescores and demographic data				
	b	β	SE	*t* test (*df*)	*P* value	*b*	β	SE	*t* test (*df*)	*P* value
General psychological domain										
Life satisfaction	1.47	0.11	0.52	2.82 (188)	.005	1.35	0.10	0.55	2.47 (181)	.02
Perceived stress	–2.00	–0.17	0.79	–2.55 (188)	.01	–1.91	–0.16	0.82	–2.34 (181)	.02
Depressive and anxiety symptoms	–1.24	–0.15	0.51	–2.43 (188)	.02	–1.13	–0.13	0.53	–2.14 (181)	.03
Trait anger	–0.66	–0.06	0.53	–1.26 (188)	.21	–0.88	–0.08	0.54	–1.64 (181)	.10
Trait anger (temperament)	–0.22	–0.04	0.25	–0.85 (188)	.40	–0.23	–0.04	0.26	–0.89 (181)	.37
Trait anger (reaction)	–0.59	–0.11	0.29	–2.03 (188)	.04	–0.71	–0.13	0.29	–2.41 (181)	.02
Mindfulness	–0.02	0.00	1.03	–0.01 (188)	.99	0.11	0.00	1.08	0.10 (181)	.92
Work										
Work performance										
Absolute presenteeism	0.05	0.00	2.49	0.02 (188)	.98	–0.33	–0.01	2.60	–0.13 (181)	.90
Relative presenteeism	0.03	0.06	0.04	0.80 (188)	.42	0.02	0.04	0.04	0.53 (181)	.60
Job satisfaction	–0.03	–0.02	0.10	–0.26 (188)	.79	–0.01	0.00	0.10	–0.07 (181)	.94
Quantitative job overload	–0.01	0.00	0.28	–0.05 (188)	.96	0.00	0.00	0.29	0.01 (181)	.99
Job control	0.24	0.05	0.24	0.99 (188)	.32	0.25	0.06	0.24	1.03 (181)	.30
Family										
Family satisfaction	0.14	0.08	0.10	1.37 (188)	.17	0.14	0.08	0.10	1.32 (181)	.19
Partner satisfaction	0.50	0.11	0.30	1.64 (188)	.10	0.38	0.08	0.32	1.20 (181)	.23
Work-to-family conflict										
Work-family negative spillover	0.37	0.04	0.44	0.84 (188)	.40	0.28	0.03	0.46	0.60 (181)	.55
Family-work negative spillover	–0.21	–0.06	0.22	–0.93 (188)	.35	–0.23	–0.06	0.23	–1.00 (181)	.32
Work-family positive spillover	–0.21	–0.03	0.40	–0.53 (188)	.60	–0.06	–0.01	0.41	–0.15 (181)	.88
Family-work positive spillover	0.07	0.01	0.42	0.16 (188)	.87	–0.04	0.00	0.43	–0.09 (181)	.93

Differences Between Pre- and Postintervention Assessment Within Each Group

Regarding the intervention group, the postintervention scores of life satisfaction were significantly higher ($mean_{pre}$ 17.70, SD_{pre} 6.68; $mean_{post}$ 18.80, SD_{post} 6.76; t_{93}=–3.36; Cohen *d*=0.22, 95% CI 0.09-0.35; *P*=.001) and those of depressive and anxiety symptoms were significantly lower ($mean_{pre}$ 6.16, SD_{pre} 4.06; $mean_{post}$ 5.37, SD_{post} 3.61; t_{93}=2.35; Cohen *d*=0.27, 95% CI 0.04-0.50; *P*=.02) than the preintervention scores.

Regarding the waitlist control group, the postintervention scores of perceived stress were significantly higher ($mean_{pre}$ 17.62, SD_{pre} 5.95; $mean_{post}$ 19.94, SD_{post} 6.38; t_{97}=–4.20; Cohen *d*=0.40, 95% CI 0.21-0.60; *P*<.001) and those of partner satisfaction were significantly lower than the preintervention scores ($mean_{pre}$ 7.69, SD_{pre} 2.05; $mean_{post}$ 7.15, SD_{post} 2.43; t_{50}=2.41; Cohen *d*=0.25, 95% CI 0.04-0.45; *P*=.02).

Discussion

Principal Findings

This study examined the effectiveness of a mindfulness meditation intervention via a smartphone app among healthy women workers. To our knowledge, this was the first study that examined the effects of the mindfulness meditation intervention via a smartphone app on 4 domains (psychological, work, family, and work-to-family conflict) among women workers. Women workers who received the intervention demonstrated higher postintervention scores on the general psychological indicators (life satisfaction, perceived stress, depressive and anxiety symptoms, and trait anger (reaction) than those in the waitlist control group, controlled for preintervention scores as well as age, employment status, psychiatric history, education, and marital status. However, the intervention was not effective in the other 3 domains (work, family, and work-to-family

XSL•FO
RenderX

conflict). In particular, life satisfaction and depression, and anxiety symptoms significantly improved in the intervention group.

Our results corroborated the findings of Santos et al [10] and Coelhoso et al [11] that app-based mindfulness interventions reduced perceived stress and anxiety symptoms and improved subjective well-being among working women. Additionally, we found that app-based mindfulness interventions were useful for reducing reactive anger in working women. Working women are more likely to experience stress owing to neurobiological differences and balancing work and family than men, which may impair their well-being [3-7]. Mindfulness interventions enhance acceptance and observation skills by halting in daily life, paying attention to what is happening "here and now," and observing and accepting things as they are [12,42,43]. Therefore, acceptance and observation skills enable working women to pause and look at things as they are without being overwhelmed by negative thoughts and feelings when they are burdened by work and family in their daily lives. This is likely to calm their anger, lower their subjective stress, and increase their sense of well-being.

Conversely, this study observed no improvements in work-related, family-related, and work-to-family conflict indicators after the intervention. Previous studies have reported that mindfulness interventions increase family satisfaction among elementary and secondary school teachers, partner satisfaction among participants in a romantic relationship, and work satisfaction and performance among workers, and decrease the work-to-family conflict among workers [44-50]. The inconsistency of our results with those of previous studies could be owing to differences in sex, intervention duration, and meditation time per session. Previous studies examining the effects of preventive online mindfulness interventions for nonclinical populations on perceived stress and mindfulness have shown substantial differences across studies regarding design, setting, participants' age, gender ratio, intervention characteristics, and outcome measures [51]. These factors have been suggested to potentially influence the magnitude of the observed effects. Moreover, this preventive intervention may have been too short to reduce burden in the 3 work- or family-related domains. Furthermore, the intervention content was aimed at general meditation ("mindfulness of breath," "body scan," "mindfulness of breath, sound, and body," and "loving-kindness meditation"), rather than work- or family-specific content, and was implemented in a specific order. Therefore, the 8-week low-intensity meditation intervention could have led to an improvement in the individual's general well-being; however, the effect on work- or family-related indicators may have occurred after a few months. Alternatively, higher-intensity interventions may be required, such as modifying the intervention's contents or extending its duration. Additionally, only 1 item was used to measure work performance and job, family, and partner satisfaction, whereas 3 items were used to measure job overload and control. Therefore, the number of questions may have been too small to detect significant differences.

Limitations and Future Directions

This study has some limitations: a lack of subgroup analysis, an intervention not designed specifically for the target or context, and problems with generalizability and variability in the intensity of the intervention due to the application and problems with the scales used. First, in this study, subgroup analyses were not conducted to examine the impact of the subjects' traits on intervention effects. The effects of our mindfulness intervention on general psychological, work-related, family-related, and work-to-family conflict indicators may differ based on other factors. Some participants may have benefited from work-related, family-related, or work-to-conflict indicators. Therefore, it is necessary to examine the factors that moderate the effect of mindfulness interventions.

Second, the mindfulness intervention used in this study was not designed as target- and context-specific. Previous studies have developed target- and context-specific mindfulness interventions, such as for the workplace and parenting. Therefore, future studies should be designed specifically for working women, with an aim to increase the effects on work- and family-related indicators.

Third, there are two limitations of using an app for the intervention: the quality of the intervention cannot be assessed, and generalizability is limited due to restrictions on participant conditions. Since a self-help app was used as the intervention in this study, it was not possible to assess how well participants were focused on meditation, which may have resulted in variability in the effectiveness of the intervention. In addition, the limitations of the app used for the intervention limited the participants in this study to iPhone users, which may have biased the sample and limited generalizability. Therefore, future studies should address compliance issues to address these limitations caused by the app without limiting participants to iPhone users.

Fourth, some of the scales used had few items. The small number of items might have prevented the detection of significant differences. Therefore, future research should increase the number of items used in the survey.

Conclusion

This study examined the effects of mindfulness interventions via a smartphone app on women workers' general psychological, work-related, family-related, and work-to-family conflict indicators through an RCT. Our results revealed that the intervention increased life satisfaction and reduced perceived stress, depressive and anxiety symptoms, and anger reactions.

Acknowledgments

We would like to thank the participants for their time and effort, and Mai Sugie, Kohki Kaji, Yukari Kimura, Takumu Kurosawa, and Kei Minoura for their contributions and technical assistance in data collection. RU, KA, and MS were supported in part by Japan Science and Technology Agency SPRING (JPMJSP2108).

XSL•FO
RenderX

Funding

This work was supported by the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (JP16H05653, JP19K03278, 22H01091, and 22K18582 to RT), the Royal Society and the British Academy (AL150003 to RT), and the University of Tokyo Social Cooperation Program "Fulfillment through Work" (to RT). The funders had no role in data collection or analyses, the decision to publish, or preparation of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-EHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 480 KB - jmir_v28i1e62814_app1.pdf]

References

1. Hassard J, Teoh KRH, Visockaite G, Dewe P, Cox T. The cost of work-related stress to society: a systematic review. J Occup Health Psychol 2018;23(1):1-17. [doi: 10.1037/ocp0000069] [Medline: 28358567]
2. Occupational health: stress at the workplace. World Health Organization. 2020. URL: https://tinyurl.com/dkrj3f2c [accessed 2024-04-24]
3. Stress in America 2023. A nation recovering from collective trauma. American Psychological Association. 2023. URL: https://www.apa.org/news/press/releases/stress/2023/collective-trauma-recovery [accessed 2024-04-24]
4. Kawakami N, Haratani T, Kobayashi F, Ishizaki M, Hayashi T, Fujita O, et al. Occupational class and exposure to job stressors among employed men and women in Japan. J Epidemiol 2004;14(6):204-211 [FREE Full text] [doi: 10.2188/jea.14.204] [Medline: 15617394]
5. Savic I, Perski A, Osika W. MRI shows that exhaustion syndrome due to chronic occupational stress is associated with partially reversible cerebral changes. Cereb Cortex 2018;28(3):894-906. [doi: 10.1093/cercor/bhw413] [Medline: 28108490]
6. Peristera P, Westerlund H, Magnusson Hanson LL. Paid and unpaid working hours among Swedish men and women in relation to depressive symptom trajectories: results from four waves of the Swedish longitudinal occupational survey of health. BMJ Open 2018;8(6):e017525 [FREE Full text] [doi: 10.1136/bmjopen-2017-017525] [Medline: 29880559]
7. Poms L, Fleming L, Jacobsen K. Work-family conflict, stress, and physical and mental health: a model for understanding barriers to and opportunities for women's well-being at home and in the workplace. World Med Health Policy 2016;8(4):444-457. [doi: 10.1002/wmh3.211]
8. Global gender gap report 2023. World Economic Forum. 2023. URL: https://www3.weforum.org/docs/WEF_GGGR_2023.pdf [accessed 2024-04-24]
9. Ornek OK, Esin MN. Effects of a work-related stress model based mental health promotion program on job stress, stress reactions and coping profiles of women workers: a control groups study. BMC Public Health 2020;20(1):1658 [FREE Full text] [doi: 10.1186/s12889-020-09769-0] [Medline: 33148247]
10. Santos FRMD, Lacerda SS, Coelhoso CC, Barrichello CR, Tobo PR, Kozasa EH. The integration of meditation and positive psychology practices to relieve stress in women workers (Flourish): effects in two pilot studies. Behav Sci (Basel) 2021;11(4):43 [FREE Full text] [doi: 10.3390/bs11040043] [Medline: 33810304]
11. Coelhoso CC, Tobo PR, Lacerda SS, Lima AH, Barrichello CRC, Amaro E, et al. A new mental health mobile app for well-being and stress reduction in working women: randomized controlled trial. J Med Internet Res 2019;21(11):e14269 [FREE Full text] [doi: 10.2196/14269] [Medline: 31697244]
12. Kabat-Zinn J. Mindfulness-based interventions in context: past, present, and future. Clinical Psychology: Science and Practice 2003;10(2):144-156. [doi: 10.1093/clipsy.bpg016]
13. Khoury B, Sharma M, Rush SE, Fournier C. Mindfulness-based stress reduction for healthy individuals: a meta-analysis. J Psychosom Res 2015;78(6):519-528. [doi: 10.1016/j.jpsychores.2015.03.009] [Medline: 25818837]
14. Galante J, Friedrich C, Dawson AF, Modrego-Alarcón M, Gebbing P, Delgado-Suárez I, et al. Mindfulness-based programmes for mental health promotion in adults in nonclinical settings: a systematic review and meta-analysis of randomised controlled trials. PLoS Med 2021;18(1):e1003481 [FREE Full text] [doi: 10.1371/journal.pmed.1003481] [Medline: 33428616]
15. Rusch H, Rosario M, Levison L, Olivera A, Livingston WS, Wu T, et al. The effect of mindfulness meditation on sleep quality: a systematic review and meta-analysis of randomized controlled trials. Ann N Y Acad Sci 2019;1445(1):5-16 [FREE Full text] [doi: 10.1111/nyas.13996] [Medline: 30575050]
16. Key ICT indicators for the world and special regions (totals and penetration rates). The International Telecommunication Union. 2024. URL: https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ITU_regional_global_Key_ICT_indicator_aggregates_Nov_2023.xlsx [accessed 2024-04-24]

XSL•FO
RenderX

17. Plaza I, Demarzo MMP, Herrera-Mercadal P, García-Campayo J. Mindfulness-based mobile applications: literature review and analysis of current features. JMIR Mhealth Uhealth 2013;1(2):e24 [FREE Full text] [doi: 10.2196/mhealth.2733] [Medline: 25099314]
18. Flynn S, Hastings RP, Burke C, Howes S, Lunsky Y, Weiss JA, et al. Online mindfulness stress intervention for family carers of children and adults with intellectual disabilities: feasibility randomized controlled trial. Mindfulness 2020;11(9):2161-2175. [doi: 10.1007/s12671-020-01436-0]
19. Liu C, Chen H, Zhou F, Long Q, Wu K, Lo L, et al. Positive intervention effect of mobile health application based on mindfulness and social support theory on postpartum depression symptoms of puerperae. BMC Womens Health 2022;22(1):413 [FREE Full text] [doi: 10.1186/s12905-022-01996-4] [Medline: 36217135]
20. Rigabert A, Motrico E, Moreno-Peral P, Resurrección DM, Conejo-Cerón S, Cuijpers P, et al. Effectiveness of online psychological and psychoeducational interventions to prevent depression: Systematic review and meta-analysis of randomized controlled trials. Clin Psychol Rev 2020;82:101931 [FREE Full text] [doi: 10.1016/j.cpr.2020.101931] [Medline: 33137611]
21. Sommers-Spijkerman M, Austin J, Bohlmeijer E, Pots W. New Evidence in the booming field of online mindfulness: an updated meta-analysis of randomized controlled trials. JMIR Ment Health 2021;8(7):e28168 [FREE Full text] [doi: 10.2196/28168] [Medline: 34279240]
22. Armstrong L. An investigation into the effects of a short-term mindfulness intervention on stress and emotion regulation in undergraduate students: understanding mechanisms of action. Manchester Metropolitan University. 2012. URL: https://www.semanticscholar.org/paper/An-investigation-into-the-effects-of-a-short-term-Armstrong/96cbd0ad6b2a2c1f19fb88d76fa247626d1bc860 [accessed 2026-01-08]
23. Basso JC, McHale A, Ende V, Oberlin DJ, Suzuki WA. Brief, daily meditation enhances attention, memory, mood, and emotional regulation in non-experienced meditators. Behav Brain Res 2019;356:208-220. [doi: 10.1016/j.bbr.2018.08.023] [Medline: 30153464]
24. Diener E, Emmons RA, Larsen RJ, Griffin S. The satisfaction with life scale. J Pers Assess 1985;49(1):71-75. [doi: 10.1207/s15327752jpa4901_13] [Medline: 16367493]
25. Sumino Z. Development of the Japanese version of the satisfaction with life scale. 1994 Presented at: Proceedings of the 36th Annual Meeting of the Japanese Association of Educational Psychology; September 28-30, 1994; Kyoto, Japan p. 28-30.
26. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983;24(4):385. [doi: 10.2307/2136404]
27. Sumi K. Reliability and validity of the Japanese version of the perceived stress scale [Article in Japanese]. Jpn J Health Psychol 2006;19(2):44-53. [doi: 10.11560/jahp.19.2_44]
28. Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SLT, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002;32(6):959-976. [doi: 10.1017/s0033291702006074] [Medline: 12214795]
29. Furukawa TA, Kawakami N, Saitoh M, Ono Y, Nakane Y, Nakamura Y, et al. The performance of the Japanese version of the K6 and K10 in the world mental health survey Japan. Int J Methods Psychiatr Res 2008;17(3):152-158 [FREE Full text] [doi: 10.1002/mpr.257] [Medline: 18763695]
30. Sakurai K, Nishi A, Kondo K, Yanagida K, Kawakami N. Screening performance of K6/K10 and other screening instruments for mood and anxiety disorders in Japan. Psychiatry Clin Neurosci 2011;65(5):434-441 [FREE Full text] [doi: 10.1111/j.1440-1819.2011.02236.x] [Medline: 21851452]
31. Spielberger CD. State-Trait Anger Expression Inventory (Staxi): Professional Manual. Tampa, FL: Psychological Assessment Resources, Inc; 1999.
32. Mine H, Ohki M. Applicability of anger expression inventory (STAXI-2) to Japanese. 2001 Presented at: Proceedings of the 14th Annual Convention of the Japanese Association of Health Psychology; November 3-4, 2001; Miyagi, Japan.
33. Mine H, Sato S. An attempt for the assessment of anger expression behavior (STAXI-2)--- comparison of anger expression behavior with U.S. and Japan ---. 2005 Presented at: The 18th Annual Convention of the Japanese Association of Health Psychology; September 1-3, 2005; Hyogo, Japan.
34. Brown KW, Ryan RM. The benefits of being present: mindfulness and its role in psychological well-being. J Pers Soc Psychol 2003;84(4):822-848. [doi: 10.1037/0022-3514.84.4.822] [Medline: 12703651]
35. Fujino M, Kajimura S, Nomura M. Development and validation of the Japanese version of the mindful attention awareness scale using item response theory analysis [Article in Japanese]. Jpn J Personal 2015;24(1):61-76. [doi: 10.2132/personality.24.61]
36. Kessler RC, Barber C, Beck A, Berglund P, Cleary PD, McKenas D, et al. The world health organization health and work performance questionnaire (HPQ). J Occup Environ Med 2003;45(2):156-174. [doi: 10.1097/01.jom.0000052967.43131.51] [Medline: 12625231]
37. Kessler R, Ames M, Hymel P, Loeppke R, McKenas DK, Richling DE, et al. Using the world health organization health and work performance questionnaire (HPQ) to evaluate the indirect workplace costs of illness. J Occup Environ Med 2004;46(6 Suppl):S23-S37. [doi: 10.1097/01.jom.0000126683.75201.c5] [Medline: 15194893]

XSL•FO
RenderX

38. Kawakami N, Inoue A, Tsuchiya M, Watanabe K, Imamura K, Iida M, et al. Construct validity and test-retest reliability of the world mental health japan version of the world health organization health and work performance questionnaire short version: a preliminary study. Ind Health 2020;58(4):375-387 [FREE Full text] [doi: 10.2486/indhealth.2019-0090] [Medline: 32173661]
39. Inoue A, Kawakami N, Shimomitsu T, Tsutsumi A, Haratani T, Yoshikawa T, et al. Development of a short version of the new brief job stress questionnaire. Ind Health 2014;52(6):535-540 [FREE Full text] [doi: 10.2486/indhealth.2014-0114] [Medline: 24975108]
40. Geurts SAE, Taris TW, Kompier MAJ, Dikkers JSE, Van Hooff MLM, Kinnunen UM. Work-home interaction from a work psychological perspective: development and validation of a new questionnaire, the SWING. Work & Stress 2005;19(4):319-339. [doi: 10.1080/02678370500410208]
41. Shimada K, Shimazu A, Geurts SAE, Kawakami N. Reliability and validity of the Japanese version of the survey work–home interaction – NijmeGen, the SWING (SWING-J). Community, Work & Family 2018;22(3):267-283. [doi: 10.1080/13668803.2018.1471588]
42. Siegel RD, Germer CK, Olendzki A. Mindfulness: What is it? Where did it come from? In: Clinical Handbook of Mindfulness. New York, NY: Springer New York; 2009.
43. Hölzel BK, Lazar SW, Gard T, Schuman-Olivier Z, Vago DR, Ott U. How does mindfulness meditation work? Proposing mechanisms of action from a conceptual and neural perspective. Perspect Psychol Sci 2011;6(6):537-559. [doi: 10.1177/1745691611419671] [Medline: 26168376]
44. Crain TL, Schonert-Reichl KA, Roeser RW. Cultivating teacher mindfulness: effects of a randomized controlled trial on work, home, and sleep outcomes. J Occup Health Psychol 2017;22(2):138-152. [doi: 10.1037/ocp0000043] [Medline: 27182765]
45. Espel-Huynh H, Baldwin M, Puzia M, Huberty J. The indirect effects of a mindfulness mobile app on productivity through changes in sleep among retail employees: secondary analysis. JMIR Mhealth Uhealth 2022;10(9):e40500 [FREE Full text] [doi: 10.2196/40500] [Medline: 36169994]
46. Hülsheger UR, Alberts HJEM, Feinholdt A, Lang JWB. Benefits of mindfulness at work: the role of mindfulness in emotion regulation, emotional exhaustion, and job satisfaction. J Appl Psychol 2013;98(2):310-325. [doi: 10.1037/a0031313] [Medline: 23276118]
47. Kappen G, Karremans JC, Burk WJ. Effects of a short online mindfulness intervention on relationship satisfaction and partner acceptance: the moderating role of trait mindfulness. Mindfulness 2019;10(10):2186-2199. [doi: 10.1007/s12671-019-01174-y]
48. Kiburz KM, Allen TD, French KA. Work–family conflict and mindfulness: investigating the effectiveness of a brief training intervention. J Organ Behavior 2017;38(7):1016-1037. [doi: 10.1002/job.2181]
49. Nicklin JM, Shockley KM, Dodd H. Self-compassion: implications for work-family conflict and balance. J Vocat Behav 2022;138:103785. [doi: 10.1016/j.jvb.2022.103785]
50. Slutsky J, Chin B, Raye J, Creswell JD. Mindfulness training improves employee well-being: a randomized controlled trial. J Occup Health Psychol 2019;24(1):139-149. [doi: 10.1037/ocp0000132] [Medline: 30335419]
51. Jayewardene WP, Lohrmann DK, Erbe RG, Torabi MR. Effects of preventive online mindfulness interventions on stress and mindfulness: a meta-analysis of randomized controlled trials. Prev Med Rep 2017;5:150-159 [FREE Full text] [doi: 10.1016/j.pmedr.2016.11.013] [Medline: 28050336]

Abbreviations

ANCOVA: analysis of covariance
BJSQ: Brief Job Stress Questionnaire
CONSORT: Consolidated Standards for Reporting Trials
FWNS: family-work negative spillover
FWPS: family-work positive spillover
K6: 6-item Kessler Psychological Distress Scale
RCT: randomized controlled trial
STAXI-2: State-Trait Anger Expression Inventory 2
T-Ang/R: Trait-anger/reaction
T-Ang/T: Trait-anger/temperament
UMIN: University Hospital Medical Information Network
WFNS: work-family negative spillover
WFPS: work-family positive spillover

XSL•FO
RenderX

Edited by T McCall; submitted 01.Jun.2024; peer-reviewed by U Sinha, D Abdel-Hady, ID Yucel; comments to author 13.Jan.2025; revised version received 07.Nov.2025; accepted 24.Nov.2025; published 02.Feb.2026.

Please cite as:
Uwagawa R, Adachi K, Shimoda M, Takizawa R
Effects of an 8-Week App-Based Mindfulness Intervention on Mental Health in Working Women: Randomized Controlled Trial
J Med Internet Res 2026;28:e62814
URL: https://www.jmir.org/2026/1/e62814
doi:10.2196/62814
PMID:41627885

©Riko Uwagawa, Koichiro Adachi, Mariko Shimoda, Ryu Takizawa. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Changing Habits With the Happy Hands App: Qualitative Focus Group Study of a Hand Osteoarthritis Self-Management Intervention

Kristine Aasness Fjeldstad[1,2], MSc; Anne Therese Tveter[1], PhD; Eivor Rasmussen[3], BSc; Lena Olden[4], BSc; Sissel Nyheim[5], MSc; Thalita Blanck[1], MSc; Rikke Munk Killingmo[2], PhD; Ingvild Kjeken[1], PhD

[1]Health Service Research and Innovation Unit, Center for Treatment of Rheumatic and Musculoskeletal Diseases (REMEDY), Diakonhjemmet Hospital, Oslo, Norway

[2]Faculty of Health Sciences, Department of Rehabilitaion Science and Health Technology, OsloMet – Oslo Metropolitan University, Oslo, Norway

[3]Hand Therapy Clinic, Kirkenes Hospital, Kirkenes, Norway

[4]Department of Physiotherapy and Occupational Therapy, Levanger Hospital, Levanger, Norway

[5]Norwegian Rheumatism Association, Oslo, Norway

Corresponding Author:
Kristine Aasness Fjeldstad, MSc
Health Service Research and Innovation Unit
Center for Treatment of Rheumatic and Musculoskeletal Diseases (REMEDY)
Diakonhjemmet Hospital
Postboks 23 Vinderen
Oslo, 0319
Norway
Phone: 47 22451500
Email: KristineAasness.Fjeldstad@diakonsyk.no

Abstract

Background: People with hand osteoarthritis represent a large patient group with limited access to recommended treatment. In recent years, there has been a notable shift in health care delivery, with increased use of digital technologies. The Happy Hands app (The University Information Technology Center [USIT]) is a digital self-management intervention developed to provide evidence-based treatment for people with hand osteoarthritis, with the goal of empowering them to self-manage their disease. Participants' experiences and perceptions of using this digital intervention are crucial for the adoption and continued use of the Happy Hands app.

Objective: The objective of this qualitative study was to explore participants' experience with using the Happy Hands app, focusing on whether and how it empowered them to self-manage their hand osteoarthritis.

Methods: The study is embedded within a randomized controlled trial (RCT). The participants were recruited from the intervention group in the RCT, who got access to the Happy Hands app. The 12-week self-management intervention included a hand exercise program and informational videos about hand osteoarthritis. Focus groups were conducted in various geographical areas in Norway. The focus groups were transcribed verbatim, coded, and analyzed inductively using reflexive thematic analysis.

Results: Seven focus groups, with a total of 26 participants, were recruited from both specialist and primary health care. The mean age was 67 years. Three themes were developed from the analysis. The first theme, "Being acknowledged," highlights the essential role of recognition for people with hand osteoarthritis. It suggests that the Happy Hands app provided participants with a sense of validation and support. The second theme, "Changed perception of hand osteoarthritis," indicates that participants gained insights and knowledge about their condition. This new understanding empowered them to make more informed decisions about their care, fostering a sense of hope and motivation by demonstrating that effective measures are available to manage the disease. The third theme, "Changing habits with the Happy Hands app," describes how participants developed new habits after using the self-management intervention delivered through the app. The exercise program was experienced as motivating, flexible, well-structured, and committing. Some challenges were reported, including experiencing pain during or after exercising. The new habits included performing hand exercises and implementing ergonomic working methods, which were tailored to meet the individual needs and integrated into the participants' daily lives and routines.

XSL•FO
RenderX

Conclusions: The findings suggest that the Happy Hands app is a valuable tool for supporting people with hand osteoarthritis in managing their disease by helping them integrate hand osteoarthritis management into their daily lives.

Trial Registration: ClinicalTrials.gov NCT05568875; https://clinicaltrials.gov/study/NCT05568875

(J Med Internet Res 2026;28:e82773) doi:10.2196/82773

KEYWORDS

mHealth; eHealth; osteoarthritis; hand exercises; self-management; social cognitive theory; qualitative research

Introduction

Hand osteoarthritis is a prevalent and debilitating condition. Nearly half of all women and a quarter of all men will develop hand osteoarthritis during their lifetime [1], with symptoms often emerging around middle adulthood. People with hand osteoarthritis typically experience pain, stiffness, and reduced grip strength [2]. These symptoms challenge individuals' ability to perform everyday activities, such as household tasks, hobbies, and work, and may reduce health-related quality of life. Although there is no cure for hand osteoarthritis, various effective measures can help alleviate symptoms [3]. Access to treatment is, however, often limited, and the quality of care is suboptimal [4], with patients often being referred to surgery without having received recommended treatment [5].

First-line treatments for people with hand osteoarthritis include patient education, hand exercises, and the use of assistive devices [3,6]. Patient education is central, as much of the management of hand osteoarthritis relies on actions individuals have to perform themselves. However, supporting individuals in self-managing a chronic condition like hand osteoarthritis can be challenging [4], particularly in maintaining adherence [7].

There has been a significant shift in the delivery of health care in recent years, with an increased emphasis on the use of digital devices in the provision of care. Government documents have highlighted eHealth as a key strategy to optimize resource use, streamline patient pathways, and ensure that individuals get access to health information [8].

Previous studies have demonstrated that digital delivery for hand osteoarthritis is a good alternative to traditional in-person consultations [9] and effective in improving hand function and reducing pain compared to usual care [10]. Qualitative studies have described participants being mostly positive to digital delivery of osteoarthritis management [11-13].

The use of mobile apps offers a valuable tool for ensuring access to treatment and supporting individuals in self-managing their hand osteoarthritis [6]. Mobile apps can offer tailored information and guided exercise routines, which may also improve exercise adherence [14,15].

The Happy Hands app (The University Information Technology Center [USIT]) contains a 12-week self-management intervention for people with hand osteoarthritis, developed to ensure access to evidence-based treatment. The goal of the app is to empower people with hand osteoarthritis to self-manage their disease [15]. Before this study, the app underwent feasibility testing aimed at refining and improving its design. The results indicated that the Happy Hands app contributed to reductions in pain and stiffness, as well as improvements in activity performance and grip strength. In focus group interviews, participants reported that the app was useful and highlighted several areas for further enhancement [15]. In this study, the aim was to explore participants' experience with using the Happy Hands app, focusing on whether and how it empowered them to self-manage their hand osteoarthritis.

Methods

Study Design

This qualitative study is conducted as part of a randomized controlled trial (RCT) investigating the effect and cost-effectiveness of the Happy Hands app. Focus groups were conducted with participants from the intervention group in the RCT to explore their experiences of using a digital intervention for hand osteoarthritis.

Happy Hands App

The Happy Hands app was developed by a research group at Diakonhjemmet hospital [15]. The development of the app was guided by social cognitive theory [16]. The behavior change taxonomy, a classification system categorizing different techniques used to change behavior [17], was used to classify the elements in the app.

The app contains a self-management intervention with informational videos and a hand exercise program. The informational videos include different themes, such as information about hand osteoarthritis, hand exercises, use of assistive devices and orthoses, medication and surgical options, and how to cope with everyday life. Furthermore, the self-management program includes a hand exercise program with videos showing how the patients should warm up and perform exercises to improve mobility, strength, and coordination, and a stretching exercise. The informational videos and exercise program are delivered in a progressive order across 12 weeks. The participants could choose 3 days a week to use the app. On these designated days they received notifications on their smartphones with that week's informational videos and hand exercises. The participants had the opportunity to tick off each completed exercise. Encouragement, motivational messages, and quizzes were provided each week to enhance continued adherence to the intervention [18].

Setting

This qualitative study was nested within the Happy Hands study, a multicenter RCT that evaluated whether a self-management intervention delivered through the Happy Hands app, in addition to usual care, was more effective than usual care alone for

XSL•FO
RenderX

people with hand osteoarthritis [18]. The RCT was conducted between November 2022 and February 2024. Participants in the RCT were recruited from 14 hospitals, 2 rehabilitation centers, 3 physiotherapy clinics, and 1 private rheumatology center, following their regular consultations at these sites. In total, 376 participants were enrolled and randomly allocated to either the control group (n=185) or the intervention group (n=191).

The control group received a one-page information sheet and usual care. Usual care for people with hand osteoarthritis varies across settings and can include no treatment, a one-day educational course, or consultations with an occupational therapist and/or a rheumatologist. The intervention group received the same one-page information sheet and usual care; in addition, they were given access to the Happy Hands app, which provided them with a 12-week self-management program.

Participants and Recruitment

In the RCT, eligible participants had symptomatic hand osteoarthritis diagnosed by health care personnel, owned a smartphone, and could read and understand Norwegian. Participants were excluded from the RCT if they had cognitive deficits, were scheduled for hand surgery within 3 months after inclusion, or had serious comorbidities or inflammatory rheumatic disease. Participants were recruited through clinical practice. Clinicians applied the inclusion and exclusion criteria to determine eligibility. Information on serious comorbidities was obtained through participant self-report, and cognitive function was assessed through clinical evaluation.

Participants in the focus groups were recruited from the intervention group in the RCT. When they were enrolled in the RCT and signed the consent form, participants could indicate their willingness to be contacted for an interview about their experiences with the app. A purposive sampling strategy was used to ensure diversity in geographic location and health care setting. Participants were recruited from various regions across Norway to ensure geographic diversity and from both primary and specialist care. Individuals in the intervention group who consented to an interview were contacted by phone. In total, 28 participants agreed to participate. However, 2 of the participants declined before the focus groups were conducted due to unknown reasons. Thus, a total of 26 participants participated in 7 focus groups.

The focus groups took place between 7 and 11 months after recruitment for the RCT began. All focus groups were conducted in 2023, the first 6 in June and the last one in October. We chose this timing to ensure that participants had completed the intervention, enabling us to explore their reflections on the full intervention period. One participant, however, joined a focus group after one month because of late enrollment, whereas the rest were interviewed after 3 or more months. Adherence to app use throughout the 12-week program period was not an inclusion criterion.

Data Collection

Focus groups were chosen as the data collection method in this study because they are useful to obtain in-depth understanding of a topic by taking advantage of the group dynamics and discussions that can occur in a group setting [19].

All 7 focus groups were conducted by the first author (KAAF). Either the last author (IK), a master student associated with the project, or one of the clinicians working at the recruitment sites assisted as moderators in the focus groups by taking notes and asking follow-up questions. KAAF is a female PhD candidate and nurse with some experience in qualitative research, while IK is an occupational therapist with extensive experience in both clinical practice and research. The master student is also an occupational therapist and was writing a thesis concerning the Happy Hands app. The clinicians were working as either a physiotherapist or occupational therapist at the recruitment sites and assisted in recruiting patients in their clinical practice to the RCT and have experience working with individuals with hand osteoarthritis. KAAF had previously met a few of the participants in connection with their inclusion in the RCT, while IK and the master student had no prior relationship with the participants before the study. The clinicians had previously met some of the participants during consultations.

The focus groups were held at the site where the participants were recruited to the RCT, at a hospital or clinic, except for one focus group that was held in a conference room at a hotel. Before the focus groups began, participants were informed about the aim of the study. They were told that any experience, both positive and negative, they had had with the Happy Hands app was welcomed and were encouraged to discuss and ask each other questions. This aligns with Liamputtong [20], who describes that focus groups allow for more topics to be initiated by participants, as they steer the conversation more than the researcher does in individual interviews.

A first version of the interview guide was developed by KAAF, ATT, and IK and focused on participants' experiences, problems with using the app, perceived benefits, and future use (Multimedia Appendix 1). It was reviewed by patient research partners (SN and TB), who suggested adding a question about initial experiences, including whether users received advice and support. They also recommended using the term "challenging" instead of "difficult" to frame questions more positively. The interview guide was thereafter tested in a pilot focus group. As no major revisions were necessary, the data from this session was included in the final dataset.

At the end of the focus group, the participants were asked if they had anything to add. Each focus group consisted of 3-5 participants and lasted between 50 minutes and one and a half hours. The audio recordings were transcribed verbatim by KAAF and a research assistant. Notes were taken by the moderator during each focus group. Following each session, the moderator and interviewer (KAAF) discussed their impressions of the content and clarified any points that required further understanding.

We applied the concept of information power to determine whether we had a sufficient sample size. This approach suggests that factors such as the study aim, participant specificity, theoretical framework, interview quality, and analysis strategy should guide decisions about when sufficient information has been obtained [21]. Our study sought to explore user experience

XSL•FO
RenderX

in a broad context, which typically necessitates a diverse sample. We therefore decided to include focus groups from different regions in Norway, covering various levels of care. Accordingly, 6 focus groups were conducted in different geographical areas, encompassing both primary and specialist health care settings. We anticipated that this strategy would yield sufficient data to address our research questions. Following the initial round of analysis, we noted a lack of information regarding negative experiences with the app and how it would be used in the future. To strengthen our understanding of this aspect, we conducted one additional focus group. We then considered that the focus groups provided a comprehensive understanding of the participants' experiences and perspectives. The data had sufficient depth and richness to address the study aim, and thus, we concluded that information power was achieved.

Theoretical Framework

Following an initial phase of inductive analysis and coding, social cognitive theory (SCT) [16] was applied to explore whether participants had benefited from the intervention and achieved behavior change. SCT outlines key determinants for behavior change: knowledge about health risks, outcome expectations, goal-setting and strategies to achieve them, and perceived facilitators and barriers. A crucial aspect is self-efficacy, which is the belief that individuals have in their ability to achieve change through their actions [16]. Additionally, observational learning, learning a new behavior by watching others perform it, is a central concept [22].

To complement this behavioral perspective, the common-sense model of illness representations developed by Leventhal et al [23], was also applied to the analysis in this study. While SCT focuses on the mechanisms underlying behavior change, the Common-Sense Model helps to explain how individuals understand and make sense of their illness. It posits that illness representations are shaped by various sources of information, such as societal lay knowledge, advice from health care professionals, and personal experiences with the illness. Illness representation can be divided into 5 different dimensions. "Cause" considers beliefs about what is causing the disease. "Consequences" addresses beliefs regarding its impact, "identity" is beliefs regarding the symptoms, and "timeline" refers to beliefs about the duration and progress. "Cure or control" is the belief about the potential for recovery or management and the individual's ability to influence their condition. These dimensions collectively shape how patients perceive their illness and have significance for the way individuals will seek help and for adopting a coping strategy for the disease [24].

Involvement of Patient Research Partners

Two patient research partners (SN and TB), both of whom completed the 12-week app program, were involved from the outset of the study and contributed to the review of the interview guide. Following a presentation of the results, a discussion was held between them and IK and KAAF. The patient research partners indicated that they identified with the results and the developed themes. They are both coauthors of this article and have reviewed and provided feedback on the manuscript.

Data Analysis

Reflexive thematic analysis was applied to analyze the interviews, aiming to explore, interpret, and develop patterns of meaning across the dataset. This method involves 6 key phases, including familiarizing oneself with the data, generating initial codes, constructing themes, reviewing themes, defining and naming themes, and producing the final report. A key concept in this method is reflexivity, which involves continuously reflecting on our assumptions and practices and how these affect the research [25]. Analysis was performed by the first author (KAAF), the second author (ATT), and the last author (IK). Throughout the research process, we actively engaged in reflexivity through several discussions. These discussions enabled us to examine our perceptions and how they might have influenced the research.

After conducting the first 6 focus groups, we performed an initial analysis by reading through the transcripts multiple times. The aim of this familiarization phase was to gain an overall understanding of the data. Notes were written throughout this process to identify interesting elements and possible patterns across the dataset and as a tool to be able to reflect on interpretations [25]. Following this familiarization phase, we also decided to conduct one additional focus group.

Each transcript was coded by KAAF using NVivo software (QSR International Pty Ltd). Text segments relevant to the aim of the study were tagged with distinct codes for different meanings. The initial analysis was conducted inductively. Coding was performed at a semantic level, staying close to the participants' language**.** In a second round of coding, a deductive approach was applied, looking for segments of text relevant to the theories we planned to include. The second and last authors (ATT and IK) each read half of the transcripts, and the team met to discuss reflections and interpretations on content in data and potential themes.

The codes were gathered, and 3 initial themes were developed from them. The tentatively developed themes were reviewed against all the codes clustered around each theme. A discussion was held with the patient research partners, IK and KAAF, following a presentation about the results. They expressed recognition of the developed themes and highlighted the importance of feeling seen and heard, access to information, and how exercises had become a natural part of their daily routines, facilitated by the app. Quotes were selected from the transcripts to illustrate the findings. Finally, the themes were defined and given names. All coauthors reviewed the drafts and provided feedback.

Reflexive notes were written throughout the research process to increase awareness of preconceptions and to enhance transparency. Initial preconceptions were documented at the beginning of the study and later compared with the final results. The considerable differences between them can be understood as an indication that the first author's preconceptions did not prevent new insights from emerging.

Ethical Considerations

This study was approved by the Regional Committees for Health Research Ethics (477746), the Data Protection Officer (00660),

XSL•FO
RenderX

and the local research committee at Diakonhjemmet hospital. Informed consent was obtained from all participants. The interviews were audio recorded by the Nettskjema-Diktafon app (University of Oslo), which safely transfers encrypted audio files to Services for Sensitive data (TSD) at the University of Oslo. TSD is a platform for collecting and storing sensitive data in compliance with the Norwegian privacy regulation. The qualitative study is briefly described in the trial description for the RCT registered on ClinicalTrials.gov (NCT05568875). We followed the Consolidated Criteria for Reporting Qualitative Research (COREQ; Multimedia Appendix 2) [26]. Participants did not receive compensation for taking part in the focus groups but were offered reimbursement for travel expenses.

Results

Participants' Characteristics

A total of 26 participants (18 women and 8 men) participated in 7 focus groups. The mean age was 67 years. Participants had hand osteoarthritis diagnosis for between one and 25 years (Table 1).

Table 1. Participant characteristics (N=26).

Characteristic	Participants
Age (years), mean (SD)	67 (7)
Gender, n (%)	
Women	18 (69)
Men	8 (31)
Years living with hand osteoarthritis, n (%)	
0-9	13 (50)
10-19	11 (42)
More than 19	2 (8)
Geographical area, n (%)	
Southeast	15 (58)
West	4 (15)
Central	3 (12)
North	4 (15)

Through the analysis we developed 3 themes. The theme "Being acknowledged" highlights the importance informants placed on having their hand osteoarthritis recognized and validated. Meanwhile, the theme "Changed perception of hand osteoarthritis" reflects how their perceptions of hand osteoarthritis changed after gaining insights from the self-management intervention in the app. Finally, "Changing habits with the Happy Hands" app reports how people with hand osteoarthritis integrated hand exercising and ergonomic working methods into their daily life and routines.

Being Acknowledged

A discussion about a feeling of not getting support and recognition for their hand osteoarthritis emerged in one of the focus groups. A few participants described a feeling of not receiving adequate health care. For example, one participant described having had pain for several years and experiencing a prolonged wait before finally receiving a diagnosis:

> *I've actually had pain for many years, but no one understood what it was. [...] I go to the doctor, no, it's not that, I take tests, nothing happens, but this year I got the diagnosis.* [Woman, 57 years]

In one focus group it was discussed that osteoarthritis is a disease that is given little attention in society, even though it affects many people. The participants described lacking support from the workplace and that they did not talk much about their disease with family or friends. Moreover, participants described a sense of dealing with the burden of hand osteoarthritis alone:

> *It is not talked about a lot, you go around with pains and you don't talk about it. We're supposed to manage it, it's not a problem, we women can fix it.* [Woman, 72 years]

Being part of the project and getting access to the app was therefore perceived as meaningful, as these experiences provided a sense of validation and acknowledgment. This was described in one of the focus groups.

I want to take a slightly different approach, which has actually been my problem. I've had pain and limitations in activity for a very long time, and that leads to, now I'm almost about to cry, poor sleep, and a low mood for a long time […] you feel like it's just something we have to live with […]So when [the doctor] said that I could be part of this, I was really happy […] That someone cares and someone sees you. Yes. [Woman, 75 years]

Furthermore, the focus group revealed that after getting the app and becoming part of the Happy Hands study, participants were increasingly willing to discuss hand osteoarthritis with their friends and family, having previously avoided doing so.

> *But what is currently the most important thing is all my female friends […] it has started to become a topic*

XSL•FO
RenderX

> *of conversation. […] There have been times when we've sat together and done some exercises at gatherings, saying, 'Now we're stretching our fingers,' and turning it into a little show.* [Woman, 72 years]

This willingness to share experiences with others reflects an increased openness about the condition, which may contribute to a sense of acknowledgment and support.

The theme highlights how some participants perceive a lack of support for their hand osteoarthritis, making their involvement in the study and access to the app feel like a genuine acknowledgment of their condition. This experience encouraged them to open up about their struggles with friends and family, indicating that it developed a sense of community and support.

Changed Perception of Hand Osteoarthritis

Through the self-management intervention delivered through the app, participants gained knowledge and insights about hand osteoarthritis, which seemed to have changed their perception of the disease in different ways. Participants described that the app gave them hope and motivation by demonstrating that different measures could be taken to manage the disease effectively.

> *It's important that one can do something about it, because I think it is what has been most discouraging in a way, that there isn't anything to do about it. It just has to run its course. And then, just the fact that you see that there is something you can do.* [Woman, 65 years]

Moreover, knowledge about the disease and its mechanisms reduced the feeling of uncertainty, for example, by getting an understanding of why they had pain and why some activities suddenly were difficult to perform.

> *You are now more aware of why it is there, right [...] what is it now, what have I done, so the uncertainty is gone, now you sort of know what you have to deal with.* [Man, 66 years]

The participants emphasized that it was important to receive knowledge about recommended treatment options for patients with hand osteoarthritis. This could be information about surgery, medication, use of orthoses and how to modify activity performance. This information could empower them to make informed decisions about their care based on updated knowledge. For example, in one of the focus groups, it was a consensus that they appreciated knowing that surgery is recommended as the last treatment option, since they did not want to have surgery. One participant was referred to surgery but decided to wait and see if he would benefit from doing hand exercises.

> *So, there were a number of things that became clear now […] it was like, you were given the steps and that surgery was presented as the last option, right. […] And... I liked that, because I don't want to be cut open.* [Man, 66 years]

A few participants described gaining knowledge that using their hands can have a preventive effect and that it is not harmful to use their hands even when experiencing pain:

> *I was a bit skeptical at first because I was afraid that doing something painful might be harmful. I thought I was causing damage, so I was very cautious. But now, I am willing to tolerate some pain as long as it doesn't get worse. In a way, it's reassuring to know this, and it has helped me normalize the experience.* [Woman, 64 years]

This theme describes how participants gained valuable insights about hand osteoarthritis, which shifted their perceptions of the disease. They expressed that increased knowledge about treatment options and their understanding of their condition alleviated uncertainty, supported them to make more informed decisions about their care, and provided hope and motivation by demonstrating that there exist measures that can help manage the disease.

Changing Habits With the Happy Hands App

This theme explores how participants changed their habits after completing the self-management intervention in the Happy Hands app, particularly in exercising and activity performance.

Developing New Habits: Exercising

The app's exercise component helped participants learn hand exercises and integrate them into their daily lives. Participants unanimously agreed that it was easy to use the app, with no one reporting any difficulties in understanding its technical aspects. Many participants emphasized the flexibility of the app as a key factor in its success, appreciating the ability to perform the exercises anywhere and at a time that was convenient for them.

> *You can exercise almost anywhere. I have exercised outdoors, I have exercised indoors in different places, I've kind of brought along a small box with equipment. It is very practical not to be tied to a specific table or chair or anything like that.* [Woman, 74 years]

The participants described that the app instilled a sense of commitment by requiring them to complete specific exercises within a designated timeframe. This sense of obligation was further reinforced by their participation in a research project and the requirement to tick off each completed exercise within the app. Several participants expressed that receiving notifications on their smartphones served as helpful reminders to complete their exercises. Moreover, scheduling 3 fixed days per week provided structure and consistency.

The app includes videos that demonstrate how to perform the exercises, which participants noted improved their ability to execute them correctly. A few even described feeling a sense of connection with the person in the videos, as if they were exercising alongside someone, which they found motivating.

> *I think it's very important that you see a person demonstrating the exercises, rather than just sitting and reading something. For me, there's a significant difference between the two. […] The physical or psychological—I'm not sure exactly what to call*

XSL•FO
RenderX

> *it—connection makes a big difference when it comes to actually doing those exercises.* [Woman, 72 years]

The possibility to track their progress in the app, complete quizzes, and receive feedback was also described as a motivation by some of the participants. Moreover, these features encouraged them to put in extra effort to both understand and retain the app' content.

Participants in the focus groups reported experiencing various improvements. When they experienced improvement, for instance, in improved strength, mobility, or ability to perform activities, they were motivated to carry out the training program.

> *I found it very motivating when I noticed that, wow, I have much more mobility, it didn't come the first week, not the second week, and not the third week, but maybe in the fourth and fifth week I started to get better. I can bend all my fingers now, which I couldn't do before, and then you understand, this is working, and it is also motivating to continue.* [Woman, 57 years]

There were, however, also some challenges described when it came to using the Happy Hands app. These included experiencing pain during or after exercising. Some participants said that their hand osteoarthritis was too advanced for effective exercising. Not everyone reported improvement. It was suggested that the app should include guidance on how to manage pain during exercise.

> *But my experience has been, in short, that I have a few fingers that have very mild osteoarthritis. And I can feel how good this app is for those fingers. At the same time, I notice that it's not very good for my bad fingers, as they actually get provoked and irritated by some of them [the exercises].* [Woman, 65 years]

There were also other reasons for not continuing to perform hand exercises. For example, one participant said she had stopped doing exercises simply because she had more going on in her life and she had forgotten about it. Other diseases or injuries could also take away the focus of conducting hand exercises. Additionally, some participants expressed that they wished that new exercises had appeared in the app, as this would have been motivating.

The results suggest that the Happy Hands app enabled participants to learn an exercise program for hand osteoarthritis, to integrate it into their daily routines, and ultimately to establish it as a habit. Once the participants had memorized the exercises, it was no longer necessary to watch the app while performing them. This made it possible to integrate exercising into their daily life by performing them while engaged in other activities.

> *Now I notice that I sit and watch the news while doing it with a tennis ball and perform many of the exercises a bit automatically in my daily life. When you go for a walk, […] many of them are easy to continue with.* [Woman, 67 years]

They could also adjust the exercising to when and how it suited the individual. For example, some preferred following the exercise program less structured. Other participants preferred to continue to exercise in a structured way, 3 times a week with the app in front of them.

Moreover, the participants adjusted the exercises because of other factors. For example, conducting exercises with less intensity when experiencing pain or conducting exercises differently when finding the exercise difficult to carry out.

> *… squeezing that ball has also started to hurt more, as I said you should squeeze as hard as you can for five seconds, five times, per hand. So, I'm trying to be a bit more careful; I'm still trying to do it and hold it, but maybe I'm not squeezing as hard simply because it hurts more than it did before.* [Man, 59 years]

The majority of the participants said they planned to continue to exercise. These results suggest that the app facilitated making exercising a habit.

This theme discusses how the app helped participants integrate hand exercises into their daily routines. Participants found the app easy to use and appreciated its flexibility. A sense of commitment was supported by the need to complete the exercise on time, reminders from the app, and scheduled sessions. Seeing a person demonstrate the exercises gave participants motivation, as well as making them easier to understand. Many experienced improvements that enhanced motivation to continue. However, some participants faced challenges, such as pain during or after exercising or osteoarthritis too advanced to benefit from exercise. The app facilitated the development of new habits for many participants, enabling them to integrate exercises into their daily life and adjust them to their individual needs and circumstances.

Developing New Habits: Adapting Daily Activities

The findings also reveal that participants made changes to how they performed everyday activities, including ergonomic working methods and use of assistive devices as suggested in the app. The adaptations they learned through the app made some activities easier to perform. For example, participants reported learning new techniques for holding and lifting objects, such as avoiding carrying items solely with their fingers or warming up their hands before use. These changes seemed to have had a significant impact on their daily life, as one participant described,

> *I do notice a difference, yes, [...] and then I think about warming up when I'm going to write, for example. Once I've warmed up, it's not so embarrassing to write my name anymore. It was embarrassing [...] it was almost like being naked, I would say, sixty-seven, and I could hardly write my name.* [Man, 67 years]

Additionally, participants found the information about different assistive devices to be helpful. For example, they mentioned tools like a nutcracker for opening soda bottles, as well as other devices such as a cheese cutter and a bread knife with ergonomic handles. However, a few participants expressed some reluctance to using assistive devices, as they associated them with "being old."

XSL•FO
RenderX

This theme described how participants also changed their performance of everyday activities after insights they had gained from the app.

Discussion

Principal Findings

The aim of this study was to explore participants' experience with using the Happy Hands app, focusing on whether and how it empowered them to self-manage their hand osteoarthritis. Our findings indicate that participants previously had experienced a lack of support and recognition for their hand osteoarthritis. Hence, to become part of the study and get access to the app provided them with a sense of validation and acknowledgment. Furthermore, the results suggest that participants' illness perception changed after gaining insights from the informational videos in the app, from viewing their condition as something insignificant and beyond their control to recognizing its importance, taking their own experiences seriously, and gaining insights into strategies they could use to alleviate pain and improve function.

Participants perceived the app as flexible, motivating, and well-structured, which facilitated both their learning and adherence to the exercise program as well as alternative working methods. As a result, they developed new habits, such as integrating hand exercises into their daily life and changing the way they perform everyday activities. However, for some participants, their condition hindered their ability to complete the exercises, often due to pain.

The findings in our study reveal that some participants felt unsupported and lacked recognition in their struggles with hand osteoarthritis by family members, health care providers, and society at large. This lack of understanding and acknowledgment is consistent with findings from other qualitative studies. Hill et al [27] noted that individuals with hand osteoarthritis perceive a lack of empathy from health professionals regarding the impact of the disease, whereas Gignac et al [28] found that symptoms were often dismissed as a normal part of aging. In our study, participants reported minimizing the severity of their condition, rarely discussing their hand osteoarthritis despite its significant impact on their daily life. This tendency to downplay symptoms aligns with the findings of Magnussen et al [29], who reported that people with hand osteoarthritis often feel undeserving of health care. Bukhave and Huniche [30] also found that participants did not seek medical care, despite experiencing a wide range of activity limitations. According to the common-sense model of illness representation [23], people's perceptions and understanding of their disease influence their behaviors and coping strategies. It is therefore essential for health care professionals to explore patients' previous experiences with the health care system and to actively acknowledge their perceived symptoms and functional challenges. Feeling that their condition is validated may have shifted participants' views from considering their hand osteoarthritis as insignificant and something they simply had to endure to recognizing its importance and the impact it has on their daily lives.

As outlined in the common-sense model of illness representations, various kinds of knowledge contribute to a person's perception of their illness [24]. The findings in our study suggest that participants changed their illness perception by acquiring new knowledge. They gained expert knowledge through the app, where health professionals provided valuable information about hand osteoarthritis. Additionally, they developed new personal knowledge through their own experiences, such as noticing improvement and regaining the ability to perform activities. This suggests that the newly acquired knowledge enabled participants to develop a new understanding of their illness, resulting in a shift in their illness perception.

According to the common-sense model of illness representations, illness perception can be categorized into different dimensions, including the "cause" dimension, which refers to what people believe to be the underlying reason for their illness [24]. Our findings suggest that participants altered their perception of this dimension by using the app, transitioning from uncertainty and fear about the origins of their pain, worried that activity might worsen their condition, to a more informed understanding. They came to recognize the actual cause of their illness and understood that using their hands would not exacerbate their pain.

Our results indicate that participants' perceptions shifted from believing nothing could be done about the disease to recognizing that effective measures are available. This aligns with the "cure or control" dimension of the common-sense model of illness representations, which describes a person's belief about how a disease can be managed [24]. The initial perception that nothing can be done aligns with findings from other studies, which report that individuals often view hand osteoarthritis as a natural result of wear and tear or aging [27,31], and believe that the condition is untreatable [29]. Participants in our study also changed their perception of their own ability to take action. For instance, information about treatment options helped them understand that delaying surgery is often a recommended approach. This shift in perception, from viewing hand osteoarthritis as untreatable to acknowledging the availability of effective management strategies, empowered participants to begin adopting these strategies.

An important finding in our study is that participants developed new habits by integrating hand exercises into their daily routines and adjusting them to their individual needs. The results further suggest that the Happy Hands app played a key role in facilitating this behavioral change. SCT, which is widely used in the development of health interventions [22], may help explain how the Happy Hands app contributed to this change. By supporting participants to gradually learn and implement the exercise program, the app enabled them to master activities that were previously difficult.

The findings further reveal that due to new behaviors, such as hand exercising, participants experienced positive outcomes, including reduced pain and an improved ability to perform previously challenging activities. In SCT, self-efficacy, a person's belief in their ability to make necessary changes to achieve a desired outcome, is a central concept. Self-efficacy

XSL•FO
RenderX

can be enhanced by empowering individuals to succeed with achievable actions that progressively become challenging [22]. Thus, the Happy Hands app may have played a significant role in enhancing participants' self-efficacy, which is critical for driving sustained behavior change [16].

According to SCT, self-regulation is a key strategy for facilitating successful behavior change. It comprises several elements, including self-monitoring, goal setting, feedback, and self-reward, and often includes the possibility to observe and record one's behavior [22]. These elements were integrated into the app and were highlighted by participants as important in helping them adhere to the exercise program. Participants emphasized the value of tracking their progress by ticking off completed exercises, which fostered a sense of commitment. Moreover, having 3 designated exercise days per week provided structure, while app notifications served as helpful reminders to complete the exercises. This aligns with the goal-setting element of SCT, which suggests that setting short-term, achievable goals motivates effort and guides actions [16]. Furthermore, participants found motivation in the app's feedback features, such as being able to track their progress and answer quizzes. Feedback is recognized in SCT as a crucial component in facilitating behavior change, reinforcing effort, and sustaining motivation.

The app includes videos that demonstrate how to perform the exercises, which was appreciated by the participants. First, observing someone performing the exercises helped them understand how they should be performed correctly. Second, participants expressed that doing the exercises while watching the videos felt as though they were training alongside someone, which increased their motivation. This finding can be further explained through SCT, where observational learning is a key concept, suggesting that seeing others perform a specific behavior encouraged app users to replicate that behavior [22].

Most participants expressed a positive attitude toward the intervention delivered through the Happy Hands app. However, some challenges were reported. These included experiencing pain during or after exercising, and for some participants, their hand osteoarthritis seemed too severe for them to benefit from the exercises. Such individuals may require individual face-to-face guidance, either as a substitute for or in combination with digitally delivered interventions. While this raises concerns about the usability of digital interventions for all patients, it is important to recognize that widespread use of digital solutions could free up valuable health care resources for those who require more intensive follow-up care from health professionals.

People with hand osteoarthritis are a large patient group with limited treatment options. To ensure a sustainable health care service, the integration of digital solutions and technologies is essential. The Happy Hands app has the potential to provide people with hand osteoarthritis access to reliable information, guidance, and recommended treatment.

Strengths and Limitations

A key strength of this study is that participants were recruited from both primary and specialist care and from various regions across Norway, ensuring representation across a broad geographical area and levels of care. Furthermore, we collaborated with patient research partners with hand osteoarthritis who had used the Happy Hands app for 12 weeks. These partners provided feedback on the interview guide and participated in a meeting to discuss the study results. They expressed that they could personally relate to the findings. Respondent validation, that is, sharing results with people who have experienced the phenomena being studied, serves as an important strategy to help prevent misinterpretations [32].

The focus groups consisted of participants who shared their perceptions and experiences, with the interviews being transcribed verbatim. This approach facilitated the collection of rich data, characterized by depth and diversity, thus providing a comprehensive understanding of the issues being studied [32]. Another strength was the collaborative analysis process, where both the PhD candidate and supervisors read through transcripts, participated in coding, and engaged in discussions regarding the content in the interviews. Such discussions and feedback are beneficial for identifying potential flaws in logic or methodology, thereby enhancing the rigor of the research [32].

Our study also has some limitations. One is that participants voluntarily chose to participate in the Happy Hands study and agreed to be interviewed. This may have introduced a selection bias, as they might have had a more positive attitude toward using an app and were more receptive of digitally delivered interventions. This could also indicate that the participants in our study had a higher level of eHealth literacy compared to the general population. Individuals who did not use the app may have been more likely to decline participation in the focus group interviews. Future studies should aim to include their perspectives and experiences, as this could yield valuable insights into whether and how the app could be improved to meet their needs, as well as inform alternative approaches to care delivery.

While focus groups are useful for exploring collective experiences and generating rich, interactive data, they also have limitations. Group dynamics can influence what is shared, with dominant voices potentially influencing the discussion. Some individuals may hesitate to express dissenting or personal views, particularly when discussing sensitive topics. Additionally, the time constraints of group discussions may limit the depth of individual narratives [20].

Conclusions

The findings indicate that the Happy Hands app initiated and facilitated a process that resulted in behavior change among the participants. The app provided validation and acknowledgment of the disease. Furthermore, insights from the app contributed to a more informed understanding of hand osteoarthritis, its causes, and its consequences. Features of the app supported participants in learning hand exercises and new working methods. Together with altered illness perception, this enabled participants to develop new behaviors. For a few participants, challenges such as pain during or after exercise were noted, indicating that some individuals may require more individualized face-to-face follow-up.

XSL•FO
RenderX

Patients with hand osteoarthritis represent a large patient group with limited access to recommended treatment. There is, therefore, a need for a new model for hand osteoarthritis care. The findings of this study demonstrate that participants benefited from the Happy Hands app. The app proved to be a valuable tool in empowering participants to better self-manage their condition. The result of this study suggests that the app can serve as a component in a treatment pathway for people with hand osteoarthritis.

Acknowledgments

We would like to thank all the participants who took the time to participate in the focus groups. We also thank Christine Hillestad Hestevik for reviewing the manuscript and for providing valuable feedback. GPT UiO (OpenAI's GPT models within University of Oslo's privacy requirements) [33] was used to translate quotes from Norwegian to English. The authors carefully reviewed the quotes afterwards to ensure they were translated correctly. GPT UiO was also used to improve English language. All suggested changes were evaluated by authors.

Data Availability

The dataset generated in this study is not available due to ensure participants anonymity and to not reveal sensitive information.

Funding

The project is funded by Foundation Dam (2022/FO387170). The funder was not involved in the study.

Authors' Contributions

ATT and IK planned the study. KAAF, ATT, SN, TB and IK developed the interview guide. KAAF recruited participants and conducted the focus groups. ER and LO assisted during focus groups. KAAF, ATT and IK performed the main analysis. SN and TB provided feedback on the results. KAAF, ATT, ER, LO, SN, TB, RMK and IK provided feedback on drafts of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[DOCX File , 18 KB - jmir_v28i1e82773_app1.docx]

Multimedia Appendix 2
COREQ checklist.
[DOC File , 72 KB - jmir_v28i1e82773_app2.doc]

References

1. Qin J, Barbour KE, Murphy LB, Nelson AE, Schwartz TA, Helmick CG, et al. Lifetime Risk of Symptomatic Hand Osteoarthritis: The Johnston County Osteoarthritis Project. Arthritis Rheumatol 2017 Jun;69(6):1204-1212 [FREE Full text] [doi: 10.1002/art.40097] [Medline: 28470947]
2. Marshall M, Watt FE, Vincent TL, Dziedzic K. Hand osteoarthritis: clinical phenotypes, molecular mechanisms and disease management. Nat Rev Rheumatol 2018;14(11):641-656. [doi: 10.1038/s41584-018-0095-4] [Medline: 30305701]
3. Kloppenburg M, Kroon FP, Blanco FJ, Doherty M, Dziedzic KS, Greibrokk E, et al. 2018 update of the EULAR recommendations for the management of hand osteoarthritis. Ann Rheum Dis 2019;78(1):16-24. [doi: 10.1136/annrheumdis-2018-213826] [Medline: 30154087]
4. Dziedzic KS, Allen KD. Challenges and controversies of complex interventions in osteoarthritis management: recognizing inappropriate and discordant care. Rheumatology (Oxford) 2018;57(suppl_4):iv88-iv98 [FREE Full text] [doi: 10.1093/rheumatology/key062] [Medline: 29684219]
5. Gravås EMH, Tveter AT, Nossum R, Eide REM, Klokkeide Å, Matre KH, et al. Non-pharmacological treatment gap preceding surgical consultation in thumb carpometacarpal osteoarthritis - a cross-sectional study. BMC Musculoskelet Disord 2019;20(1):180 [FREE Full text] [doi: 10.1186/s12891-019-2567-3] [Medline: 31039774]
6. Kjeken I, Bordvik DH, Osteras N, Haugen IK, Aasness Fjeldstad K, Skaalvik I, et al. Efficacy and safety of non-pharmacological, pharmacological and surgical treatments for hand osteoarthritis in 2024: a systematic review. RMD Open 2025;11(1):e004963 [FREE Full text] [doi: 10.1136/rmdopen-2024-004963] [Medline: 39793978]
7. Pisters MF, Veenhof C, Schellevis FG, Twisk JWR, Dekker J, De Bakker DH. Exercise adherence improving long-term patient outcome in patients with osteoarthritis of the hip and/or knee. Arthritis Care Res (Hoboken) 2010;62(8):1087-1094 [FREE Full text] [doi: 10.1002/acr.20182] [Medline: 20235201]

XSL•FO
RenderX

8. Meld. St. 9 Nasjonal helse- og samhandlingsplan 2024–2027 - Vår felles helsetjeneste. Norwegian Ministry of Health and Care Services. 2024. URL: https://www.regjeringen.no/no/dokumenter/meld.-st.-9-20232024/id3027594/ [accessed 2025-04-01]
9. Walter MM, Sirard P, Nero H, Hörder H, Dahlberg LE, Tveter AT, et al. Digitally delivered education and exercises for patients with hand osteoarthritis-an observational study. Musculoskeletal Care 2023;21(4):1154-1160. [doi: 10.1002/msc.1796] [Medline: 37421256]
10. Rodríguez Sánchez-Laulhé P, Biscarri-Carbonero Á, Suero-Pineda A, Luque-Romero LG, Barrero García FJ, Blanquero J, et al. The effects of a mobile app-delivered intervention in people with symptomatic hand osteoarthritis: a pragmatic randomized controlled trial. Eur J Phys Rehabil Med 2023;59(1):54-64 [FREE Full text] [doi: 10.23736/S1973-9087.22.07744-9] [Medline: 36633498]
11. Cronström A, Dahlberg LE, Nero H, Ericson J, Hammarlund CS. 'I would never have done it if it hadn't been digital': a qualitative study on patients' experiences of a digital management programme for hip and knee osteoarthritis in Sweden. BMJ Open 2019;9(5):e028388 [FREE Full text] [doi: 10.1136/bmjopen-2018-028388] [Medline: 31129601]
12. Nelligan RK, Hinman RS, Teo PL, Bennell KL. Exploring attitudes and experiences of people with knee osteoarthritis toward a self-directed ehealth intervention to support exercise: qualitative study. JMIR Rehabil Assist Technol 2020;7(2):e18860 [FREE Full text] [doi: 10.2196/18860] [Medline: 33242021]
13. Ezzat AM, Bell E, Kemp JL, O'Halloran P, Russell T, Wallis J, et al. "Much better than I thought it was going to be": telehealth delivered group-based education and exercise was perceived as acceptable among people with knee osteoarthritis. Osteoarthr Cartil Open 2022;4(3):100271 [FREE Full text] [doi: 10.1016/j.ocarto.2022.100271] [Medline: 36474949]
14. Gell NM, Smith PA, Wingood M. Physical therapist and patient perspectives on mobile technology to support home exercise prescription for people with arthritis: a qualitative study. Cureus 2024;16(3):e55899 [FREE Full text] [doi: 10.7759/cureus.55899] [Medline: 38601402]
15. Tveter AT, Varsi C, Maarnes MK, Pedersen SJ, Christensen BS, Blanck TB, et al. Development of the happy hands self-management app for people with hand osteoarthritis: feasibility study. JMIR Form Res 2024;8:e59016 [FREE Full text] [doi: 10.2196/59016] [Medline: 39470716]
16. Bandura A. Health promotion by social cognitive means. Health Educ Behav 2004;31(2):143-164. [doi: 10.1177/1090198104263660] [Medline: 15090118]
17. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
18. Tveter AT, Fjeldstad KA, Varsi C, Maarnes MK, Pedersen SJ, Christensen BS, et al. Evaluation of an e-self-management intervention (Happy Hands app) for hand osteoarthritis: study protocol for a multicentre randomised controlled trial. Rheumatol Int 2025;45(1):30 [FREE Full text] [doi: 10.1007/s00296-025-05787-6] [Medline: 39820549]
19. Bowling A. Research Methods in Health: Investigating Health and Health Services. Maidenhead: McGraw-Hill Education; 2014.
20. Liamputtong P. Focus Group Methodology : Principles and Practice. London: SAGE; 2011.
21. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
22. McAlister AL, Perry CL, Parcel GS. How individuals, environments, and health behaviors interact: social cognitive theory. In: Glanz K, Rimer BK, Viswanath K, editors. Health Behavior and Health Education : Theory, Research, and Practice. San Francisco, CA: Jossey-Bass; 2008:169-188.
23. Leventhal H, Meyer D, Nerenz D. The common sense representation of illness danger. In: Rachman S, editor. Medical Psychology. Elmsford, NY: Pergamon Press; 1980:7-30.
24. Hagger MS, Orbell S. A meta-analytic review of the common-sense model of illness representations. Psychology Health 2003;18(2):141-184. [doi: 10.1080/0887044031000081321]
25. Braun V, Clarke V. Thematic Analysis: A Practical Guide. Los Angeles, CA: SAGE; 2022.
26. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
27. Hill S, Dziedzic KS, Nio Ong B. Patients' perceptions of the treatment and management of hand osteoarthritis: a focus group enquiry. Disabil Rehabil 2011;33(19-20):1866-1872. [doi: 10.3109/09638288.2010.550381] [Medline: 21859421]
28. Gignac MAM, Davis AM, Hawker G, Wright JG, Mahomed N, Fortin PR, et al. "What do you expect? You're just getting older": a comparison of perceived osteoarthritis-related and aging-related health experiences in middle- and older-age adults. Arthritis Rheum 2006;55(6):905-912. [doi: 10.1002/art.22338] [Medline: 17139636]
29. Magnussen HJ, Kjeken I, Pinxsterhuis I, Sjøvold TA, Hennig T, Thorsen E, et al. Participation in healthcare consultations: a qualitative study from the perspectives of persons diagnosed with hand osteoarthritis. Health Expect 2023;26(3):1276-1286 [FREE Full text] [doi: 10.1111/hex.13744] [Medline: 36916677]
30. Bukhave EB, Huniche L. Activity problems in everyday life--patients' perspectives of hand osteoarthritis: "try imagining what it would be like having no hands". Disabil Rehabil 2014;36(19):1636-1643. [doi: 10.3109/09638288.2013.863390] [Medline: 24308906]

XSL•FO
RenderX

31. Grime J, Richardson JC, Ong BN. Perceptions of joint pain and feeling well in older people who reported being healthy: a qualitative study. Br J Gen Pract 2010;60(577):597-603 [FREE Full text] [doi: 10.3399/bjgp10X515106] [Medline: 20822692]
32. Maxwell JA. Qualitative Research Design: An Interactive Approach. Los Angeles, CA: Sage; 2013.
33. GPT UiO-UiOs privacy friendly GPT chat. University of Oslo. 2025. URL: https://www.uio.no/english/services/it/ai/gpt-uio/ [accessed 2025-08-01]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
RCT: randomized controlled trial
SCT: social cognitive theory
USIT: The University Information Technology Center

Edited by A Stone; submitted 25.Aug.2025; peer-reviewed by W Ogundare, GA McHugh; comments to author 17.Oct.2025; accepted 24.Dec.2025; published 02.Feb.2026.

Please cite as:
Fjeldstad KA, Tveter AT, Rasmussen E, Olden L, Nyheim S, Blanck T, Killingmo RM, Kjeken I
Changing Habits With the Happy Hands App: Qualitative Focus Group Study of a Hand Osteoarthritis Self-Management Intervention
J Med Internet Res 2026;28:e82773
URL: https://www.jmir.org/2026/1/e82773
doi:10.2196/82773
PMID:

©Kristine Aasness Fjeldstad, Anne Therese Tveter, Eivor Rasmussen, Lena Olden, Sissel Nyheim, Thalita Blanck, Rikke Munk Killingmo, Ingvild Kjeken. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Facilitators and Barriers to Implementing Mobile Mental Health Interventions: Qualitative Study of the Consolidated Framework for Implementation Research in Pediatric Oncology Providers

Shannon J H Hong[1], BS; Michaela Patton[2], PhD; Krysta S Barton[3], PhD; Tonya M Palermo[2,4], PhD; Karen Mulholland[2]; Eric J Chow[5,6], MPH, MD; Nancy Lau[2,7], PhD

[1]Department of Psychology, University of California, Berkeley, Berkeley, CA, United States
[2]Center for Child Health, Behavior and Development, Seattle Children's Research Institute, Seattle, WA, United States
[3]Stanford Center for Biomedical Ethics, Stanford University School of Medicine, Stanford, CA, United States
[4]Department of Anesthesiology and Pain Medicine, University of Washington School of Medicine, Seattle, WA, United States
[5]Department of Pediatrics, University of Washington School of Medicine, Seattle, WA, United States
[6]Clinical Research and Public Health Sciences Divisions, Fred Hutchinson Cancer Center, Seattle, WA, United States
[7]Department of Psychiatry and Behavioral Sciences, University of Washington School of Medicine, Seattle, WA, United States

Corresponding Author:
Shannon J H Hong, BS
Department of Psychology
University of California, Berkeley
2121 Berkeley Way
Berkeley, CA, 94704
United States
Phone: 1 510 642 5292
Email: s.hong@berkeley.edu

Abstract

Background: Adolescent and young adult (AYA) cancer survivors experience unique psychosocial needs during and after treatment. Mobile health (mHealth) interventions are an emerging area of research to help address unmet psychosocial needs. However, few studies have examined provider perspectives on the design-to-implementation pipeline.

Objective: Guided by the Consolidated Framework for Implementation Research (CFIR), our study aimed to examine provider perspectives on facilitators and barriers to implementing mHealth apps in routine clinical care.

Methods: AYA oncology providers participated in a semistructured 1:1 interview on facilitators and barriers to incorporating mHealth apps as psychosocial standard of care. We conducted a directed content analysis of the interviews using a standardized CFIR codebook and construct definitions, with codebook adaptations for mHealth innovations and the population of AYAs with cancer.

Results: A total of 20 providers (mean 39, SD 7.0 years; 80% female and 70% non-Hispanic White) representing various medical and psychosocial roles participated in the interviews. The data were analyzed with 16 CFIR constructs. We identified the following facilitators to mHealth implementation across four CFIR domains: (1) Innovation: alignment with patient needs, patient-centered co-design, strong research evidence, and user-friendly design; (2) Outer Setting: shared commitment to addressing mental health needs and openness to mHealth use; (3) Inner Setting: openness to training on mHealth use; and (4) Individuals: engaging key implementation partners such as bedside nurses and social workers and strong clinical team buy-in. We identified the following barriers to mHealth implementation across three CFIR domains: (1) Innovation: associated costs, (2) Outer Setting: heavy clinical workloads, and (3) Inner Setting: lack of cross-team collaboration and communication and clinical workflow integration.

Conclusions: Our findings highlight key considerations for mHealth co-design, the adoption of mHealth apps into routine care, implementation strategies, and provider training opportunities in the context of AYA cancer care. Partnering with AYA cancer survivors, families, and providers will be crucial for developing and implementing mHealth apps with the ultimate goal of advancing universally accessible evidence-based digital health care.

(J Med Internet Res 2026;28:e87533) doi:10.2196/87533

XSL•FO
RenderX

KEYWORDS

adolescents; young adults; cancer; digital health; mHealth; mental health; mobile phone, psychosocial intervention; implementation science

Introduction

Approximately 80,000 adolescents and young adults (AYAs) are diagnosed with cancer annually in the United States [1]. Cancer disrupts key developmental milestones for AYAs, such as identity formation, transitioning to independence, relationship development, and achieving educational and vocational goals [2,3]. AYAs living with cancer and AYA cancer survivors (hereafter referred to as AYA cancer survivors) also have unique psychosocial needs that may be unmet throughout their cancer experience [4,5]. Consequently, AYA cancer survivors experience a myriad of mental health issues, including anxiety, depression, and distress, which often persist into long-term survivorship [6,7].

Evidence-based interventions for AYA cancer survivors are a new and emerging area of research [8-11]. The number of AYA-specific psychosocial intervention trials conducted annually has increased significantly from 2 during the years 2007-2014 to 11 during the years 2015-2021 [9]. This growing research activity highlights a commitment to developing evidence-based, developmentally tailored interventions that address the unique needs of this population [12]. As new interventions continue to emerge, leveraging technology has the potential to address barriers that AYAs face in accessing (eg, lack of service availability and financial challenges) and using (eg, time limitations and discomfort in sharing personal problems) face-to-face services [13,14]. As digital natives who have been immersed in technology from a young age, AYAs frequently use digital tools in their daily lives for communication, entertainment, and to access health information and support [15,16]. In recent years, there has been growing research on digital health interventions designed for AYA cancer survivors, with most studies focusing on symptom management, behavior change, self-care, and emotional health [17]. Digital health interventions have demonstrated feasibility and acceptability for AYA cancer survivors [18]. Compared with standard care, digital psychosocial interventions have been found to significantly reduce anxiety and depression and increase resilience in children and AYAs undergoing cancer treatment [10,19]. With 95% of teenagers having access to smartphones [20], mobile health (mHealth) interventions have the potential to expand evidence-based psychosocial care.

On average, it takes 17 years for research evidence to be translated into clinical practice, and only 1 in 5 evidence-based interventions are eventually implemented into routine clinical care [21-23]. While mHealth apps align with AYAs' digital preferences and hold promise in expanding access to psychosocial care, their integration into clinical care remains a significant research-to-practice gap. Although it is well recognized that successful mHealth implementation depends on the commitment and support of both patients and providers, little prior research has examined provider perspectives [24]. Recent implementation research studies have examined provider perspectives on mHealth tools such as symptom-tracking apps for patients, decision support apps for providers, and cancer triage systems for the general public [25-27]. However, studies have not yet examined provider perspectives on the implementation of psychosocial mHealth interventions. Providers are key stakeholders and partners whose insights are essential for understanding what influences the uptake and implementation of mHealth apps as standard of care within clinical settings.

To address these gaps, we conducted a directed content analysis of qualitative interview data from AYA oncology providers. The primary objectives of our study were to identify facilitators and barriers to implementing new and emerging mHealth apps into routine clinical care from the perspective of AYA oncology providers. Our analysis was guided by the Consolidated Framework for Implementation Research (CFIR), a widely used determinant framework to examine contextual factors influencing implementation effectiveness [28].

Methods

Ethical Considerations

This study was approved by the Seattle Children's Hospital (SCH) Institutional Review Board (number STUDY00004006). All participants provided written informed e-consent via REDCap (Research Electronic Data Capture; Vanderbilt University). Any and all quotes have been deidentified. Providers received a small token of appreciation (ie, water tumbler) for participation in the study. No other monetary incentives were provided.

Design, Setting, and Participants

Eligible participants were health care providers in the SCH Cancer and Blood Disorders Clinic. All patient-facing providers in the clinic, regardless of role, were eligible to participate in the study. This included psychosocial clinicians (social workers, therapists, and psychologists), oncologists (physicians and fellows), and nurses (nurse practitioners, inpatient, and outpatient) who would potentially integrate mHealth apps into their clinical workflows. Past studies have focused primarily on the perspectives of oncologists. In the interest of inclusivity, we used provider networks across medical and psychosocial staff to try to get equal representation across 6 different patient-facing roles (occupational therapist, pediatric oncologist, psychologist, nurse, social worker, or mental health therapist). We advertised the study on SCH oncology clinic listservs and newsletters and used snowball sampling to identify additional potential participants for recruitment. Participants were queried for convenient appointment times and scheduled for participation in a 1:1 interview. Before the interview, participants completed a brief demographics survey via REDCap.

Research Team

Authors' backgrounds included health services research (TMP and NL), implementation science (NL), digital health (SJHH, MP, TMP, and EJC), clinical and pediatric psychology (SJHH,

XSL•FO
RenderX

MP, TMP, and NL), pediatric oncology (EJC), psycho-oncology research (MP and NL), and qualitative research (KSB and NL). Authors have extensive background and experience in use of the CFIR framework and qualitative methods (SJHH, MP, TMP, KSB, and NL), which informed study design and research objectives. Authors also have extensive clinical experience with AYA cancer survivors (MP, EJC, and NL), which informed the interview guide, codebook development, and clinical implications drawn from the study.

Interviews

Semistructured 1:1 interviews were conducted with 20 health care providers via videoconference and were audio-recorded. This sample size was determined based on qualitative research guidelines recommending group sizes of 15-20 to achieve thematic saturation (no new themes or qualifying patterns) [29,30]. Interviews were deidentified and transcribed verbatim. A clinical psychologist trained in qualitative research methods (NL) conducted interviews from July 2024 to September 2024. Individual interviews were 30-50 minutes in length (average 40 minutes) and focused on facilitators and barriers to incorporating emerging mHealth apps as psychosocial standard of care. The interview script is shown in Multimedia Appendix 1. The interviewer memoed during the data collection process. Interviews were conducted at a single cross-sectional timepoint. We adapted a standardized CFIR interview guide for mHealth and the population of AYAs with cancer focused on the CFIR Innovation, Outer Setting, and Inner Setting domains [28].

Qualitative Analysis

We followed the SRQR (Standards for Reporting Qualitative Research) guidelines for complete reporting of qualitative research data collection and analysis procedures [31]. The completed SRQR checklist is provided in Multimedia Appendix 2. The overall analysis approach was a directed content analysis aimed at identifying facilitators and barriers to implementation of mHealth apps in routine care. We used a standardized CFIR codebook with CFIR domains and constructs, and CFIR definitions. We then developed coding definitions that were specific to mHealth innovations and the population of AYAs with cancer. Using this initial codebook, SJHH, MP, and NL independently coded the first 3 interview transcripts to determine relevance, validity, and fit of each relevant CFIR domain and construct. Over consecutive coding meetings, the analysis team (SJHH, MP, and NL) independently coded transcripts in sets of 3 and iteratively discussed and addressed codebook discrepancies and refined definitions until the codebook was finalized.

The final version of the codebook consisted of 16 of the 67 constructs from the CFIR that were most appropriate to digital mental health interventions. Standardized protocol for applying the CFIR framework is to select a small subset of relevant constructs and domains, specifically operationalized for each research project in accordance with the CFIR qualitative study design guidelines [32]. The most relevant 16 constructs operationalized for our study spanned across all 4 CFIR domains: Innovation (Source, Evidence-Base, Relative Advantage, Adaptability, Design, and Cost), Outer Setting (Local Attitudes and Local Conditions), Inner Setting (Information Technology Infrastructure, Work Infrastructure, Relational Connections, Communications, Recipient Centeredness, Compatibility, and Access to Knowledge and Information), and Individuals (Implementation Team Members). A primary coder (SJHH) and secondary coders (MP and NL) independently coded all transcripts, and all coders met twice monthly to discuss discrepancies and reach consensus. Interview transcripts were managed and analyzed using DeDoose software (SocioCultural Research Consultants).

The interviewer and coder (NL) had prior research relationships with some providers who participated in the study. The interview guide was designed so that questions and probes placed equal emphasis on facilitators and barriers, ensuring that participant feedback covered both domains. Research relationships and solicitations for feedback have historically fostered constructive dialogue, including weighing the pros and cons of different approaches. Dissenting opinions were shared openly and were explicitly encouraged. Reflexive memos and extensive audit trails documenting coding decisions were further used to enhance analytic rigor.

Results

Demographics

Participants were 20 oncology providers (mean 39, SD 7.0, range 29-56 years; Table 1). Most of the sample was female (80%) and self-identified as non-Hispanic White (85%). In terms of provider roles, 60% were medical providers and 40% were psychosocial providers. Provider demographics were representative of the overall staff within the clinic. The mean number of years participants had provided clinical care for oncology patients was 10 (SD 9.2, range 1-24) years. On average, participants reported providing care for 20 AYA patients per month (SD 21.0; range 4-100).

XSL•FO
RenderX

Table 1. Participant demographic characteristics (N=20).

Participant characteristics	Values, n (%)
Sex	
Female	16 (80)
Male	4 (20)
Race	
Asian	3 (15)
White	14 (70)
Multiracial	3 (15)
Ethnicity	
Hispanic or Latino	1 (5)
Not Hispanic or Latino	19 (95)
Role	
Occupational therapist	1 (5)
Pediatric oncologist	8 (40)
Psychologist	2 (10)
Nurse (registered nurse or advanced practice provider)	4 (20)
Mental health therapist	2 (10)
Social worker	3 (15)

CFIR Framework

We identified facilitators and barriers to mHealth implementation by using constructs across CFIR domains (Innovation, Outer Setting, Inner Setting, and Individuals) and indicated whether each CFIR construct was a facilitator or barrier using (+) or (–), respectively. Thematic saturation was reached with our study sample. A visual representation of the interrelationships among the Innovation, Outer Setting, Inner Setting, and Individuals domains is shown in Figure 1. We demarcated each CFIR construct as a facilitator or barrier to provide a holistic synthesis of our findings. All findings were relevant to AYA cancer survivors across the cancer care continuum from diagnosis to long-term survivorship. Exemplar quotes for each CFIR construct are shown in Table 2.

XSL•FO
RenderX

Figure 1. Identified Consolidated Framework for Implementation Research facilitators (+) and barriers (–) to implementing mHealth apps in pediatric oncology settings across the cancer care continuum.

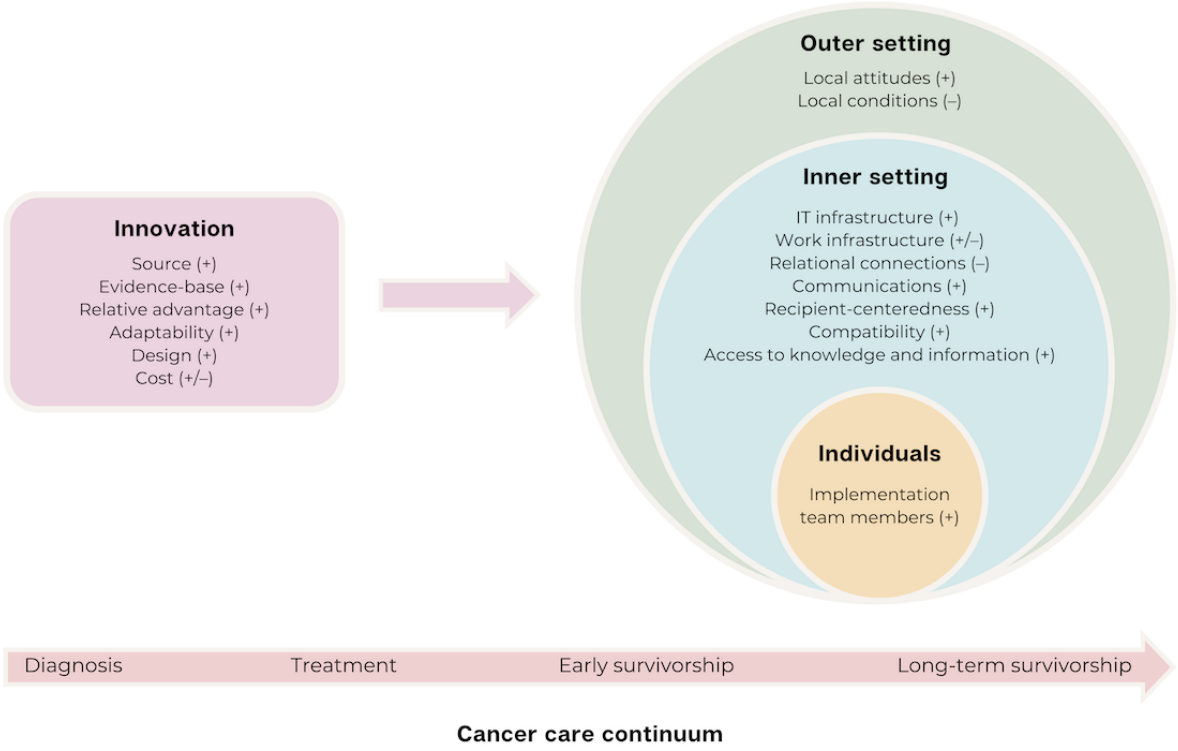

XSL•FO
RenderX

Table 2. Exemplar quotes of Consolidated Framework for Implementation Research constructs by domain.

Domain and construct	Representative Quotes
Innovation	
Source	• "As a professional recommending something to patients, I would want to do better than just generating the type of recommendation patients could find...just Googling. So I would want for someone with some expertise in the area to have evaluated the apps...Good resources would be things [that are] evidence based, sound, scientific foundations, have been used with [an] AYA[a] cancer population." [P9][b] • "I think about programs...that were born out of things that patients were involved in. I think about our AYA, advisory board. People get excited when our own patients are involved in the creation and implementation of things...if patients could take some ownership of this as well and be involved in the beginning stages." [P14]
Evidence-base	• "I would want a product that was vetted by psychologists and other clinicians and had some sort of proven efficacy...I just want to make sure that anything that I am recommending...is something that is going to provide some sort of medical benefit." [P1] • "A lot of it, of course, is dependent on how engaged the patient is...There are a lot of apps out there that are not evidence based, and so whether, you know, patients are being exposed to concepts or interventions that are anecdotally useful for people but not really evidence based." [P3]
Relative advantage	• "I think a digital app can be presented and interacted with much more frequently than a face-to-face interaction with a clinician. And you know, because of that, you can get maybe a bit more longitudinal interaction and tracking." [P1] • "We provide really good psychosocial care to our AYA population, which something like this would just enhance. But of the 80,000-ish AYAs that are diagnosed in the US annually...I think something like this app is particularly valuable for reaching those patients who are a sizable proportion of the AYA patient population nationally." [P9]
Adaptability	• "One of the things that I think is always beneficial are those med reminders [and] symptom management tools where they can document things. Just different tools to help...address that...developmental stage that they're in and...seeking out more of their independence, autonomy, all those different types of things...plays really deeply into their mental health." [P7] • "It's very scalable. It would allow AYAs to move through the material at their own pace, according to their own interests on their own time...It potentially could be delivered in a variety of languages with the content potentially tailored to a variety of cultures. So I think it's very scalable, which is important for potential use outside of just our patient population in our institution." [P9]
Design	• "It's my clinical experience that [AYAs] like to receive information by video rather than by spoken word or written word. They like content that is short snippets, rapidly paced, attention getting, highly personalized to them in terms of their interests, their issues, their questions...Something that feels more like that is probably going to be more acceptable to them, and so they'd probably be more likely to use it." [P9] • "I feel receptive to an app, especially if it's well designed and relatively intuitive." [P12]
Cost	• "I would consider cost and what kind of burdens [exist] for signing up. You know, everybody always has to create a username and all this stuff...but I would want a product that was really easy to sign up for and register and wasn't collecting tons of overly invasive stuff or posing extra burdens for an AYA to sign up for." [P1] • "I think the cost and availability of the app would definitely be...in the forefront. If a patient has to commit or get a subscription, I may be more hesitant to recommend that." [P4]
Outer setting	
Local attitudes	• "As I've discussed with other nurses and we do see a huge need for mental health support for our patient population and I think that the hospital and the [clinic] program have done a good job of trying to improve upon that and we've had more resources and more people join that team." [P17] • "A gap in our psychosocial services that maybe a tool could help us with is transition planning for survivors [with] assessments and then also like resources and interventions for survivors, I don't think we do a great job." [P12]
Local conditions	• "Nurses do feel like they're overburdened with tasks to complete throughout the day and so I do worry that some would and kind of cast it off to the side...It's just a reminder fatigue, I guess in that sense." [P17] • "One [barrier] is definitely the time and the fact that there's many different providers involved and things can fall through the cracks when they don't feel like the top priority to everyone involved, as there's a lot of things going on." [P4]
Inner setting	

XSL•FO
RenderX

Domain and construct	Representative Quotes
IT infrastructure	• "We have the Get Well Network which is part of our TV system...that included videos...about...what it means to have your counts dropped, what it means to get a central line, all of those things. And alerts pop up on the TV for families to watch those videos...So I'm not sure how you could embed it into that, but I think that could be really helpful to say "reminder, don't forget to do your [app]." [P17] • "Making it a standardized thing in our clinic, whether it's on the iPads that are in the rooms or...just having the visibility of...the [app] in our clinic, I think would be helpful, especially if it's proven to be effective. I think having easier access to it, within our clinical space and showing that the clinic backs it up, maybe that would help more patients download it and utilize it at home and take out that a little bit of that stigma piece to it." [P7] • "When you talk to someone about an app, it's harder to really understand what it looks like or what it entails...[Using TVs] could be a nice opportunity to reach families...while they're inpatient." [P18]
Work infrastructure	• "How can there be something that feels...automated and part of the system that then can be sort of supported for patients who have questions about it or who bring it up rather than relying on one person remembering to mention this at a visit or at a different time point." [P4] • "I'm learning a lot more in terms of being a supervisor and how to support patients and family or support my team...when you find the people who have the passion behind the work, things tend to happen more rather than just kind of sitting in this creative process." [P7]
Relational connections	• "There is a lack of collaboration between our social work and our embedded behavioral and mental health teams. And so, because of that, it's made a little bit challenging to understand who's doing what and how and where there's reason for overlap...and best practices of how things are delivered and who is delivering it or who's responsible for it." [P2] • "Factors include, I guess a lack of clarity with differences in clarity with roles and scope of practice. I really think it's like we're in such extreme silos, if there is just kind of this often this lack of understanding of what other services do in comparison to your own." [P14]
Communications	• "Half the battle is gonna be making people aware of [the apps] and buying into their benefit. I think any kind of 1 pager or a few minutes at a meeting, but that both talks about the benefits and the research behind why this is a good approach." [P5] • "The psychosocial group has tried to incorporate...a monthly email that comes out, that's just a reminder...And then I can refer back to the email and it has bullet points of how to refer and what the criteria are and the dates of the groups. And so something like that could also be helpful." [P4]
Recipient centeredness	• "I always think about independence and helping [AYAs] have some autonomy in their decision making aspects and or learning the skills so that when they reach the age 18 that it's not such a shock and preparing them for that transition. So I think helping them create a sense of identity or independence helps their mental health, and especially in this kind of this disease based populations, specifically when they rely so much on their caregivers and their parents or any other adults in their lives." [P7] • "We transition to patients about like a year and a half to two years after therapy to the survivorship for long term follow up team and that might be another point where [the resource] could be helpful...some patients find challenging reintegrating into school or thinking about the fact that scans are upcoming, and each time scans come up that can induce some anxiety. So maybe highlighting some things that other patients or families have cited as stressful and thinking about how these apps could support patients and sort of coping or having increased resilience through that." [P4] • "There's so much isolation...lack of autonomy, and a lot of issues with self-image that I just would want there to be some sort of other avenue for them to be able to engage in [apps] and be like 'Hey, you're not the only person who's feeling this' and 'Here's some other people that you can talk to about this.'" [P11]
Compatibility	• "For a lot of the patients that I meet with, they're sort of at the top of the pyramid in terms of their clinical need. Those types of tools could be especially helpful sort of at that universal tier where they're not necessarily connected to an individual mental health provider, but could get some support that way or do some more kind of self directed skills based learning." [P8] • "In a survivorship clinic visit...it's been my experience [that] almost universally the first six months off therapy are actually harder than on therapy cancer experience itself. So my thoughts are that almost immediately after that end of therapy visit and scans, that is where...these tools from digital health are potentially most critical and allow for kind of continued way to kind of check in and communicate back. And for the team to get an insight on what's going on such that when patients come back, they...haven't been suffering from mental health issues and stress and decreased resilience for three to six months by the time they come back to see us for scans." [P2]
Access to knowledge and information	

XSL•FO
RenderX

Domain and construct	Representative Quotes
	• "[What] would be the most helpful is if [innovation developers] came probably to one of our meetings and just presented the information to our team. And just kind of let us know what you need from us, and how we can best promote this to the patients. I think that would probably be the first step. I could bring it back to our team and then see if there's interest, which I'm assuming there would be. And then [innovation developers] could come to our team meeting, and we could talk about it as a as a group about how to go about it." [P15] • "I think better training on [app] functionality and kind of the interventions that they utilize. As well as I think more, kind of research data around their efficacy would be really helpful." [P7]
Individuals	
Implementation team members	• "Social workers for sure often have a really good finger on the pulse and you know, I also think potentially some bedside nursing members or nursing champions on the inpatient unit around that new diagnosis time period could be people who could potentially help in a way." [P1] • "I think people would be incredibly receptive. I know that my nursing colleagues at the bedside would be really excited by something like this. I think that my medical colleagues would also be, you know, social work and psychology, obviously. I think child life is another group who would be very excited. I think a group that we may not sometimes think of as being invested in the mental health of AYAs, but that truly are is PT and OT...we think of them obviously as being involved in promoting physical health, but working with them closely on the inpatient unit, so much of their work focuses on mental health...so much of what PT OT does is really promoting mental health that I think very group who would be very excited by this as well." [P9]

[a]AYA: adolescent and young adult.

[b]P: provider.

Innovation

Source (+)

Most providers emphasized the importance of mHealth apps being recommended by trusted sources such as colleagues on their clinical team or national medical specialty associations. Providers also highlighted the value of AYA involvement in developing these apps and incorporating their perspectives and feedback, as "people get excited when our own patients are involved in the creation and implementation of [programs and initiatives]" [P14].

Evidence-Base (+)

Many providers valued robust scientific evidence demonstrating the effectiveness of mHealth apps and being able to see whether apps would be most beneficial for AYA cancer survivors. One provider stated, "[It would] feel like a wasted opportunity to be able to introduce a useful tool if it's not grounded in any good evidence and theory" [P3].

Relative Advantage (+)

Providers reported several advantages to using mHealth apps compared with traditional clinical practice. Providers emphasized the developmental appropriateness of apps and stated that AYA cancer survivors can "engage within the app in a way that may actually speak their language a bit better than an oncologist sitting in front of them" and that can "facilitate them being able to express themselves in a different kind of way that is probably a little bit more fitting to them" [P3]. Providers also described how mHealth apps could expand access to psychosocial care for AYA cancer survivors following cancer treatment as they transitioned back to a new normal.

Adaptability (+)

All providers discussed the adaptability of mHealth apps. Given that AYAs have unique needs and interests, "an app that has really good uptake is [going to be] one that is able to address [what] an AYA is most interested in" [P1]. A common suggestion from providers for tailoring mHealth app content to AYAs' unique psychosocial needs was incorporating social connection features (eg, connecting with AYA peers with cancer).

Design (+)

Providers stated that mHealth apps should be well designed and intuitive to use. One provider stated, "I think it's a great way to get health content...to this population as long as A: the content is sound and B: the delivery is engaging to them. I think those are the two key ingredients" [P9]. Based on their experiences with AYA cancer survivors, providers suggested that content should be presented in an interactive and time-efficient format and be highly personalized to their interests. Some app features of interest include short informational videos, gamification elements (eg, leaderboards and streaks), and mood and symptom monitoring.

Cost (+/–)

Nearly half of the sample stated that they were concerned about cost and exclusively interested in recommending mHealth apps that were available for free. Providers emphasized the importance of affordability, with one provider stating, "I wouldn't recommend ones that are not free, especially if there were a large charge" [P20].

Outer Setting

Local Attitudes (+)

Providers shared a commitment to meeting AYA cancer survivors' mental health needs and were receptive to using mHealth apps to support clinical needs. One provider stated, "A lot of providers probably...sense the impact of their patients' mental health on care and may not have the tools...that may be able to help their patients in this way. So [apps] may actually help them feel sort of empowered to at least provide something

XSL•FO
RenderX

that the patient could use in the interim before meeting with a mental health provider" [P4].

Local Conditions (–)

Some providers identified challenges within the current conditions of the clinic that could hinder the implementation of mHealth apps. A common challenge, particularly among nurses and social workers, was the time constraints in their clinical workload. To address this, providers emphasized the need for systemic support of clinic-wide mHealth implementation efforts. One provider stated, "Some of the buy-in would also have to be systemic support...It would have to come through management to say we want to provide this opportunity for anyone that's interested to have this time carved out of their clinical day to put this work into it" [P19].

Inner Setting

IT Infrastructure (+)

Many providers recommended integrating information from mHealth apps (eg, data that show that the patient is improving or struggling) into electronic medical record systems. One provider stated, "Incorporating the results or critical information from that tool back into the environment and EMR...is a critical element for the clear communication and for longevity of the app" [P2].

Providers also suggested leveraging existing technological tools in patient rooms, such as iPads and TVs, to support adoption. For example, they described how educational videos and alerts are preloaded onto TVs for patients and their families. Similarly, TVs could be used to display tutorials of the functions and features of mHealth apps and reminders to download and use them.

Work Infrastructure (+/–)

Providers highlighted how mHealth apps could help with their limited bandwidth, as "having another tool that doesn't actually require really additional provider time that we may not have would be super useful" [P3]. However, they emphasized that tasks and responsibilities would need to be integrated or automated into their existing workflow to streamline the adoption of mHealth apps.

Relational Connections (–)

Providers noted operating in silos and a lack of awareness about what other clinical teams do and how they collaborate. One provider described having "a lack of clarity...with roles and scope of practice" and "how "we're in such extreme silos...[that] there is just kind of this lack of understanding of what other services do in comparison to your own" [P14]. As a result, it has been "challenging to understand who's doing what and how, where there's reason for overlap...and best practices of how things are delivered and who is delivering it or who's responsible for it" [P2].

Communications (+)

Although there was no consensus about the frequency of sharing information about mHealth apps among providers, they mentioned that a combination of face-to-face and written communication methods, such as email, one-pagers, and presence at team meetings, would be helpful. One provider stated, "Those Friday AM talks...have always been really well-received...I just feel more visibility about these things and the work that we're doing or that is being done to support our patients from the psychosocial mental health perspective...I feel like the providers are craving it and needing it...And I think it would actually create a lot of buy-in" [P7].

Recipient Centeredness (+)

Providers demonstrated shared values around supporting and tailoring mHealth apps to the needs and interests of AYA cancer survivors. A commonly perceived need was social connection. Providers also described how AYAs have previously engaged in gamified activities. For example, one provider shared that AYAs "would get so competitive and see that another teen had gotten a certain number of points and they wanted to try to beat them" [P20]. They suggested that a similar leaderboard feature could be incorporated into mHealth apps to enhance engagement.

Compatibility (+)

Providers stated that mHealth apps could "play a role across [the] treatment continuum," particularly for AYAs getting off treatment and into survivorship as "mental health concerns persist or can absolutely worsen during that time" [P8]. Providers stated that mHealth apps could be "a tool to use especially in moments of distress, which are common in survivorship" or when patients are "coming for scans or labs and have a really high increase of anxiety or stress" [P19]. AYA cancer survivors receiving active treatment also "have a lot of downtime" and mHealth apps would be "a really great way to...harness some of that when they're between appointments or waiting after labs and before the provider comes in the clinic" [P5]. Providers also described how mHealth apps could be universally used with all AYA cancer survivors, which would "be less of a strain on...the more limited resources that we have for individual patients" [P8].

Access to Knowledge and Information (+)

Providers expressed a need for guidance on the benefits of mHealth apps and how to promote them to AYA cancer survivors. One provider stated that "very basic education about...what it's for and why it's important" would be needed so that providers "feel engaged and invested" [P4].

Individuals: Implementation Team Members (+)

Providers recommended bedside nurses and social workers as the clinical team members who would need to be involved as the most ideal potential clinic partners for the successful implementation of mHealth apps. Providers also stated that buy-in and strong engagement from clinical team members would be needed. One provider stated, "You could have created the best thing ever and it won't make a difference. What you need is a boots-on-the-ground clinic champion to make an impact at the bedside" [P9].

XSL•FO
RenderX

Discussion

Principal Findings

To our knowledge, this is the first study to examine facilitators and barriers to implementing mHealth apps in routine clinical care from the perspective of AYA oncology providers. Constructs across the CFIR Innovation, Outer Setting, Inner Setting, and Individuals domains were identified as facilitators and barriers to mHealth implementation. Our study addresses a key gap in the design-to-implementation pipeline of emerging mHealth interventions.

In terms of facilitators to mHealth implementation, a central theme was strong alignment with patient needs. Providers were interested in collaborating on the development of mHealth apps and emphasized that they should be co-designed with AYAs. Strong research evidence and user-friendly design were other key facilitators, with providers stressing that mHealth apps needed to be both effective and engaging for AYA cancer survivors. Providers viewed mental health care as a valuable component of care in oncology and felt optimistic about the potential of mHealth apps to improve patient outcomes and advance mental health equity. They also expressed openness to receiving guidance and training on using and recommending mHealth apps to AYA cancer survivors. Providers identified bedside nurses and social workers as key implementation partners and highlighted the importance of strong clinical team engagement and buy-in.

Providers also identified several implementation barriers. Cost and accessibility were common concerns, with providers expressing hesitancy in recommending subscription-based or high-fee apps. Usability was another common barrier—providers emphasized that AYA cancer survivors would not want to use mHealth apps with a slow user interface. In terms of organizational barriers, providers reported that heavy clinical workloads and limited cross-team collaboration would make it challenging to implement mHealth apps. Providers further acknowledged the potential challenges of integrating these practices into existing clinic workflows, often referencing the high task demands and the time burden associated with supporting mHealth app implementation.

Comparison With Prior Work

Our findings align with prior research with health care providers who have found positive attitudes and receptivity toward mHealth apps across various health care conditions and disease management functions. A growing body of research also supports the use of mHealth apps in oncology care, which provide opportunities to address patient needs and help them feel engaged throughout treatment [24]. However, to our knowledge, this is the first qualitative study to examine oncology provider perspectives on mHealth implementation. It is also the first to focus specifically on the AYA age group and addressing psychosocial needs in this population. Previous studies that examined provider perspectives have focused on co-designing mHealth interventions with patients and families. Features such as social connection, symptom tracking, and appointment management have emerged as important considerations for these interventions [33,34]. With regard to facilitators of mHealth adoption, our findings align with prior research that has found that providers emphasize the need for evidence-based content and user-friendly design [35]. Our findings support previous recommendations to involve both patients and providers throughout the app development process, ensuring a balance of user and clinical expertise [36,37]. Beyond intervention design, providers offered insights on the integration of mHealth tools and interventions into clinical workflows through electronic health records and existing electronic systems [38]. Providers also shared suggestions for establishing partnerships with boots-on-the-ground clinical champions for large-scale adoptions.

With regard to barriers, our findings align with existing literature on concerns about data privacy, provider liability, and time burden associated with adopting new digital tools [24,34,39]. Our study extends this work by identifying previously unreported barriers, namely, fragmented communication between clinical teams and unclear ownership of implementation tasks. Our findings also highlight opportunities to integrate both medical and psychosocial provider roles into the implementation process and the use of mHealth tools across the cancer care continuum. This includes using mHealth tools during inpatient or outpatient care and throughout the active treatment and posttreatment phases. Additionally, a qualitative review conducted by Jacob et al [40] found that only 3 out of 50 studies examining clinician adoption of mHealth tools used the CFIR framework. This highlights the added value of our study in applying a rigorous implementation science approach to understanding factors that enable and hinder mHealth implementation. While the studies that used the CFIR framework primarily identified structural and organizational factors influencing implementation [41-43], our findings uniquely underscore the value of examining team-level processes when implementing mHealth tools in complex, multidisciplinary care environments.

Limitations

The findings of our study should be interpreted in the context of several limitations. First, our study was conducted at a single academic hospital, which may limit the generalizability of our results. Providers at this institution may have more experience with or interest in digital health than providers in other clinical settings. Second, our sample lacked demographic diversity, with most participants identifying as female and non-Hispanic White. Although this reflects the provider population in the geographic location in which our study was conducted, it may limit the applicability of our findings to more diverse clinical settings. Third, we intentionally predefined 16 CFIR constructs that were applicable to facilitators and barriers within our settings. With the CFIR having 67 constructs total, future research may explore additional factors that could impact implementation, such as policies, partnerships, or external pressures. Fourth, implementation contexts are dynamic. Our interviews captured the perspectives of providers on mHealth implementation as of 2024.

Conclusions

Our findings synthesized key implementation considerations for incorporating mHealth apps as a standard of care in

XSL•FO
RenderX

psychosocial support for AYA cancer survivors. Based on these findings, recommendations for practice include co-designing mHealth apps with both patients and providers to ensure alignment with user preferences and clinical needs. These tools should then be embedded within existing clinical workflows such as electronic health records to reduce provider burden and ensure seamless integration. Improvements in cross-team collaboration and communication can also be achieved by clearly delineating roles and responsibilities across clinical teams. Additionally, concerns around data privacy, provider liability, usability, and cost should be addressed to support the sustained use and implementation of mHealth apps. Implementation strategies should also include provider training on the benefits of mHealth apps and best practices for recommending these tools to AYA cancer survivors.

Our findings have important clinical implications. Strong partnerships among AYA cancer survivors, families, and providers will be essential for developing mHealth tools that are patient-centered and evidence-based [44]. Importantly, providers of all represented clinical roles believed that mHealth tools had clinical utility and wanted to be involved in the innovation development through implementation pipeline. They were also ready and willing to partner with researchers in early adoption efforts and had suggestions on how to do so. Future research should focus not only on evaluating the effectiveness of mHealth apps but also on implementing strategies that integrate into clinical workflows and minimize provider burden [45,46]. Understanding how these strategies impact provider uptake and patient outcomes will be crucial to moving toward universal evidence-based digital health care using innovative digital health technologies.

Acknowledgments

The authors thank the providers who participated in the study and generously shared their insights.

Funding

This work is supported by the National Cancer Institute at the National Institutes of Health (grant 1K08CA263474). The opinions herein represent those of the authors and not necessarily the funders.

Data Availability

The data used in the research cannot be publicly posted and cannot be shared privately with any person because participants did not consent to data sharing. Any shared data, even if deidentified, may result in breach of confidentiality.

Authors' Contributions

Conceptualization was done by SJHH, KSB, and NL. Data were curated by SJHH, KM, and NL. Formal analysis was handled by SJHH, MP, and NL. Investigation was done by SJHH, KSB, TMP, EJC, and NL. Methodology was handled by SJHH, KSB, and NL. Project administration and supervision were done by NL. Visualization was done by SJHH and KM. Writing of the original draft was done by SJHH. Review and editing of the writing were done by SJHH, MP, KSB, TMP, KM, EJC, and NL.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Provider interview guide.
[DOCX File , 18 KB - jmir_v28i1e87533_app1.docx]

Multimedia Appendix 2
SRQR checklist.
[DOCX File , 17 KB - jmir_v28i1e87533_app2.docx]

References

1. US cancer statistics data visualizations tool. US Cancer Statistics Working Group. 2024. URL: https://www.cdc.gov/united-states-cancer-statistics/dataviz/index.html [accessed 2025-05-30]
2. Bibby H, White V, Thompson K, Anazodo A. What are the unmet needs and care experiences of adolescents and young adults with cancer? A systematic review. J Adolesc Young Adult Oncol 2017;6(1):6-30. [doi: 10.1089/jayao.2016.0012] [Medline: 27454408]
3. Itzep N, Roth M. Psychosocial distress due to interference of normal developmental milestones in AYAs with cancer. Children (Basel) 2022;9(3):309 [FREE Full text] [doi: 10.3390/children9030309] [Medline: 35327680]

XSL•FO
RenderX

4. Zebrack BJ, Block R, Hayes-Lattin B, Embry L, Aguilar C, Meeske KA, et al. Psychosocial service use and unmet need among recently diagnosed adolescent and young adult cancer patients. Cancer 2013;119(1):201-214 [FREE Full text] [doi: 10.1002/cncr.27713] [Medline: 22744865]
5. Avutu V, Lynch KA, Barnett ME, Vera JA, Glade Bender JL, Tap WD, et al. Psychosocial needs and preferences for care among adolescent and young adult cancer patients (ages 15-39): a qualitative study. Cancers (Basel) 2022;14(3):710 [FREE Full text] [doi: 10.3390/cancers14030710] [Medline: 35158975]
6. Hall AE, Boyes AW, Bowman J, Walsh RA, James EL, Girgis A. Young adult cancer survivors' psychosocial well-being: a cross-sectional study assessing quality of life, unmet needs, and health behaviors. Support Care Cancer 2012;20(6):1333-1341. [doi: 10.1007/s00520-011-1221-x] [Medline: 21720746]
7. Barnett M, McDonnell G, DeRosa A, Schuler T, Philip E, Peterson L, et al. Psychosocial outcomes and interventions among cancer survivors diagnosed during adolescence and young adulthood (AYA): a systematic review. J Cancer Surviv 2016;10(5):814-831 [FREE Full text] [doi: 10.1007/s11764-016-0527-6] [Medline: 26920873]
8. Rosenberg AR, Bradford MC, McCauley E, Curtis JR, Wolfe J, Baker KS, et al. Promoting resilience in adolescents and young adults with cancer: results from the PRISM randomized controlled trial. Cancer 2018;124(19):3909-3917 [FREE Full text] [doi: 10.1002/cncr.31666] [Medline: 30230531]
9. Murphy KM, Siembida E, Lau N, Berkman A, Roth M, Salsman JM. A systematic review of health-related quality of life outcomes in psychosocial intervention trials for adolescent and young adult cancer survivors. Crit Rev Oncol Hematol 2023;188:104045. [doi: 10.1016/j.critrevonc.2023.104045] [Medline: 37269881]
10. Salsman JM, McLouth LE, Tooze JA, Little-Greene D, Cohn M, Kehoe MS, et al. An eHealth, positive emotion skills intervention for enhancing psychological well-being in young adult cancer survivors: results from a multi-site, pilot feasibility trial. Int J Behav Med 2023;30(5):639-650. [doi: 10.1007/s12529-023-10162-5] [Medline: 36890329]
11. Salsman JM, Rosenberg AR. Fostering resilience in adolescence and young adulthood: considerations for evidence-based, patient-centered oncology care. Cancer 2024;130(7):1031-1040 [FREE Full text] [doi: 10.1002/cncr.35182] [Medline: 38163249]
12. Thornton CP, Ruble K, Kozachik S. Psychosocial interventions for adolescents and young adults with cancer: an integrative review. J Pediatr Oncol Nurs 2020;37(6):408-422. [doi: 10.1177/1043454220919713] [Medline: 32452711]
13. Holland LR, Walker R, Henney R, Cashion CE, Bradford NK. Adolescents and young adults with cancer: barriers in access to psychosocial support. J Adolesc Young Adult Oncol 2021;10(1):46-55. [doi: 10.1089/jayao.2020.0027] [Medline: 32486879]
14. Lau N, Parsa AG, Walsh C, Yi-Frazier JP, Weiner BJ, Curtis JR, et al. Facilitators and barriers to utilization of psychosocial care in adolescents and young adults with advanced cancer: integrating mobile health perspectives. J Adolesc Young Adult Oncol 2021;10(4):476-482 [FREE Full text] [doi: 10.1089/jayao.2020.0129] [Medline: 32985937]
15. van de Graaf DL, Vlooswijk C, Bol N, Krahmer EJ, Bijlsma R, Kaal S, et al. AYAs' online information and eHealth needs: a comparison with healthcare professionals' perceptions. Cancer Med 2023;12(2):2016-2026 [FREE Full text] [doi: 10.1002/cam4.5048] [Medline: 35879825]
16. Abrol E, Groszmann M, Pitman A, Hough R, Taylor RM, Aref-Adib G. Exploring the digital technology preferences of teenagers and young adults (TYA) with cancer and survivors: a cross-sectional service evaluation questionnaire. J Cancer Surviv 2017;11(6):670-682 [FREE Full text] [doi: 10.1007/s11764-017-0618-z] [Medline: 28634734]
17. McCann L, McMillan KA, Pugh G. Digital interventions to support adolescents and young adults with cancer: systematic review. JMIR Cancer 2019;5(2):e12071 [FREE Full text] [doi: 10.2196/12071] [Medline: 31368438]
18. Devine KA, Viola AS, Coups EJ, Wu YP. Digital health interventions for adolescent and young adult cancer survivors. JCO Clin Cancer Inform 2018;2:1-15 [FREE Full text] [doi: 10.1200/CCI.17.00138] [Medline: 30652583]
19. Ozturk CS, Toruner EK. Effectiveness of technology-based psychosocial interventions for child, adolescents and young adults undergoing cancer treatment: a meta-analysis of randomised controlled trials. Eur J Cancer Care (Engl) 2022;31(1):e13515. [doi: 10.1111/ecc.13515] [Medline: 34632662]
20. Vogels EA, Gelles-Watnick R, Massarat N. Teens, social media and technology 2022. Pew Research Center. 2022. URL: https://www.pewresearch.org/internet/2022/08/10/teens-social-media-and-technology-2022/ [accessed 2025-06-13]
21. Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. J R Soc Med 2011;104(12):510-520 [FREE Full text] [doi: 10.1258/jrsm.2011.110180] [Medline: 22179294]
22. Kilbourne AM, Glasgow RE, Chambers DA. What can implementation science do for you? Key success stories from the field. J Gen Intern Med 2020;35(Suppl 2):783-787 [FREE Full text] [doi: 10.1007/s11606-020-06174-6] [Medline: 33107001]
23. Rubin R. It takes an average of 17 years for evidence to change practice-the burgeoning field of implementation science seeks to speed things up. JAMA 2023;329(16):1333-1336. [doi: 10.1001/jama.2023.4387] [Medline: 37018006]
24. Berkowitz CM, Zullig LL, Koontz BF, Smith SK. Prescribing an App? Oncology providers' views on mobile health apps for cancer care. JCO Clin Cancer Inform 2017;1:1-7. [doi: 10.1200/CCI.17.00107] [Medline: 30657404]
25. Gregory ME, Cao W, Rahurkar S, Jonnalagadda P, Stock JC, Ghazi SM, et al. Exploring the incorporation of a novel cardiotoxicity mobile health app into care of patients with cancer: qualitative study of patient and provider perspectives. JMIR Cancer 2023;9:e46481 [FREE Full text] [doi: 10.2196/46481] [Medline: 38085565]

XSL•FO
RenderX

26. Jacob C, Sanchez-Vazquez A, Ivory C. Clinicians' role in the adoption of an oncology decision support app in europe and its implications for organizational practices: qualitative case study. JMIR Mhealth Uhealth 2019;7(5):e13555 [FREE Full text] [doi: 10.2196/13555] [Medline: 31066710]
27. Sangers TE, Wakkee M, Moolenburgh F, Nijsten T, Lugtenberg M. Mobile health apps for skin cancer triage in the general population: a qualitative study on healthcare providers' perspectives. BMC Cancer 2025;25(1):851 [FREE Full text] [doi: 10.1186/s12885-025-14244-3] [Medline: 40346534]
28. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated consolidated framework for implementation research based on user feedback. Implement Sci 2022;17(1):75 [FREE Full text] [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]
29. Creswell JW. Qualitative Inquiry and Research Design: Choosing Among Five Approaches. Cham: SAGE; 2013.
30. Moser A, Korstjens I. Series: practical guidance to qualitative research. Part 3: sampling, data collection and analysis. Eur J Gen Pract 2018;24(1):9-18 [FREE Full text] [doi: 10.1080/13814788.2017.1375091] [Medline: 29199486]
31. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
32. Reardon CM, Damschroder LJ, Ashcraft LE, Kerins C, Bachrach RL, Nevedal AL, et al. The Consolidated Framework for Implementation Research (CFIR) user guide: a five-step guide for conducting implementation research using the framework. Implement Sci 2025;20(1):39 [FREE Full text] [doi: 10.1186/s13012-025-01450-7] [Medline: 40819067]
33. Hou SHJ, Henry B, Drummond R, Forbes C, Mendonça K, Wright H, et al. Co-designing priority components of an mhealth intervention to enhance follow-up care in young adult survivors of childhood cancer and health care providers: qualitative descriptive study. JMIR Cancer 2025;11:e57834 [FREE Full text] [doi: 10.2196/57834] [Medline: 40279633]
34. Jibb LA, Sivaratnam S, Hashemi E, Chu CH, Nathan PC, Chartrand J, et al. Parent and clinician perceptions and recommendations on a pediatric cancer pain management app: a qualitative co-design study. PLOS Digit Health 2023;2(11):e0000169 [FREE Full text] [doi: 10.1371/journal.pdig.0000169] [Medline: 38019890]
35. Ardito V, Golubev G, Ciani O, Tarricone R. Evaluating barriers and facilitators to the uptake of mHealth apps in cancer care using the consolidated framework for implementation research: scoping literature review. JMIR Cancer 2023;9:e42092 [FREE Full text] [doi: 10.2196/42092] [Medline: 36995750]
36. Hartzler A, Pratt W. Managing the personal side of health: how patient expertise differs from the expertise of clinicians. J Med Internet Res 2011;13(3):e62 [FREE Full text] [doi: 10.2196/jmir.1728] [Medline: 21846635]
37. Tarricone R, Cucciniello M, Armeni P, Petracca F, Desouza KC, Hall LK, et al. Mobile health divide between clinicians and patients in cancer care: results from a cross-sectional international survey. JMIR Mhealth Uhealth 2019;7(9):e13584 [FREE Full text] [doi: 10.2196/13584] [Medline: 31493318]
38. Upadhyay VA, Landman AB, Hassett MJ. Landscape analysis of oncology mobile health applications. JCO Clin Cancer Inform 2021;5:579-587. [doi: 10.1200/CCI.20.00156] [Medline: 34033510]
39. Kessel KA, Vogel MME, Schmidt-Graf F, Combs SE. Mobile apps in oncology: a survey on health care professionals' attitude toward telemedicine, mHealth, and oncological apps. J Med Internet Res 2016;18(11):e312 [FREE Full text] [doi: 10.2196/jmir.6399] [Medline: 27884810]
40. Jacob C, Sanchez-Vazquez A, Ivory C. Understanding clinicians' adoption of mobile health tools: a qualitative review of the most used frameworks. JMIR Mhealth Uhealth 2020;8(7):e18072 [FREE Full text] [doi: 10.2196/18072] [Medline: 32442132]
41. Lord S, Moore SK, Ramsey A, Dinauer S, Johnson K. Implementation of a substance use recovery support mobile phone app in community settings: qualitative study of clinician and staff perspectives of facilitators and barriers. JMIR Ment Health 2016;3(2):e24 [FREE Full text] [doi: 10.2196/mental.4927] [Medline: 27352884]
42. Possemato K, Kuhn E, Johnson EM, Hoffman JE, Brooks E. Development and refinement of a clinician intervention to facilitate primary care patient use of the PTSD Coach app. Transl Behav Med 2017;7(1):116-126 [FREE Full text] [doi: 10.1007/s13142-016-0393-9] [Medline: 27234150]
43. Varsi C, Ekstedt M, Gammon D, Ruland CM. Using the consolidated framework for implementation research to identify barriers and facilitators for the implementation of an internet-based patient-provider communication service in five settings: a qualitative study. J Med Internet Res 2015;17(11):e262 [FREE Full text] [doi: 10.2196/jmir.5091] [Medline: 26582138]
44. Abernethy A, Adams L, Barrett M, Bechtel C, Brennan P, Butte A, et al. The promise of digital health: then, now, and the future. NAM Perspect 2022;2022 [FREE Full text] [doi: 10.31478/202206e] [Medline: 36177208]
45. Ramsey WA, Heidelberg RE, Gilbert AM, Heneghan MB, Badawy SM, Alberts NM. eHealth and mHealth interventions in pediatric cancer: a systematic review of interventions across the cancer continuum. Psychooncology 2020;29(1):17-37. [doi: 10.1002/pon.5280] [Medline: 31692183]
46. Parikh RB, Basen-Enquist KM, Bradley C, Estrin D, Levy M, Lichtenfeld JL, et al. Digital health applications in oncology: an opportunity to seize. J Natl Cancer Inst 2022;114(10):1338-1339 [FREE Full text] [doi: 10.1093/jnci/djac108] [Medline: 35640986]

XSL•FO
RenderX

Abbreviations

AYA: adolescent and young adult
CFIR: Consolidated Framework for Implementation Research
mHealth: mobile health
REDCap: Research Electronic Data Capture
SCH: Seattle Children's Hospital
SRQR: Standards for Reporting Qualitative Research

Edited by A Stone; submitted 10.Nov.2025; peer-reviewed by T Jimoh, WNA Wan Adnan; comments to author 04.Dec.2025; revised version received 07.Jan.2026; accepted 14.Jan.2026; published 23.Feb.2026.

Please cite as:
Hong SJH, Patton M, Barton KS, Palermo TM, Mulholland K, Chow EJ, Lau N
Facilitators and Barriers to Implementing Mobile Mental Health Interventions: Qualitative Study of the Consolidated Framework for Implementation Research in Pediatric Oncology Providers
J Med Internet Res 2026;28:e87533
URL: https://www.jmir.org/2026/1/e87533
doi:10.2196/87533
PMID:

©Shannon J H Hong, Michaela Patton, Krysta S Barton, Tonya M Palermo, Karen Mulholland, Eric J Chow, Nancy Lau. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Mobile Health App Attitudes and Adoption Among Oncology Providers: Cross-Sectional National Survey

Nancy Lau[1,2], PhD; Kavin Srinakarin[1], BS; Shannon JH Hong[3], BS; Homer Aalfs[1], BS; Michael E Roth[4], MD; Karly M Ingram[5], PhD; Amy Berkman[6], MD; Sherilynn Chan[1,2,8], PhD; Joanna Patten[7], PsyD; Wade Iwata[8], LICSW; Mallory R Taylor[9,10], MD, MS; Jesse R Fann[2], MD, MPH; Eric J Chow[09,10,11], MD, MPH; Tonya M Palermo[1,12], PhD

[1]Center for Child Health, Behavior and Development, Seattle Children's Research Institute, Seattle, WA, United States
[2]Department of Psychiatry and Behavioral Sciences, University of Washington School of Medicine, Seattle, WA, United States
[3]Department of Psychology, University of California, Berkeley, Berkeley, CA, United States
[4]Department of Pediatrics, MD Anderson Cancer Center, Houston, TX, United States
[5]Department of Psychology, East Carolina University, Greenville, NC, United States
[6]Department of Epidemiology and Cancer Control, St Jude Children's Research Hospital, Memphis, TN, United States
[7]Division of Psychiatry and Behavioral Medicine, Seattle Children's Hospital, Seattle, WA, United States
[8]Cancer and Blood Disorders Center, Seattle Children's Hospital, Seattle, WA, United States
[9]Ben Towne Center for Childhood Cancer and Blood Disorders Research, Seattle Children's Research Institute, Seattle, WA, United States
[10]Department of Pediatrics, University of Washington School of Medicine, Seattle, WA, United States
[11]Clinical Research and Public Health Sciences Divisions, Fred Hutch Cancer Center, Seattle, WA, United States
[12]Department of Anesthesiology & Pain Medicine, University of Washington School of Medicine, Seattle, WA, United States

Corresponding Author:
Nancy Lau, PhD
Center for Child Health, Behavior and Development
Seattle Children's Research Institute
1920 Terry Ave
Seattle, WA, 98101
United States
Phone: 1 206 884 8238
Email: nancy.lau@seattlechildrens.org

Abstract

Background: Mobile health (mHealth) apps can address health inequities and enhance access to care for individuals with immunocompromising conditions. Although hundreds of oncology apps exist, research on provider perspectives regarding their use in clinical care remains limited.

Objective: This study aimed to describe oncology providers' recommended apps, mHealth attitudes and beliefs, and perceived barriers to and facilitators of mHealth adoption. Exploratory aims examined differences based on provider type (medical vs psychosocial), provider age (<45 vs ≥45 years old), and patient population (pediatric vs adult).

Methods: We conducted a cross-sectional survey administered via REDCap (Research Electronic Data Capture) to oncology providers across the United States between June and November 2024. Data were summarized using descriptive statistics. Pearson's chi-square analyses examined exploratory group differences based on provider type, provider age, and patient age.

Results: Of 188 respondents, the majority self-identified as female (150/188, 79.8%), White (161/188, 85.6%), and non-Hispanic/Latino (174/188, 92.6%). Nearly all providers (178/188, 94.7%) reported either recommending or using mHealth apps with their patients, with primary use for patient-provider communication (139/188, 73.9%). Providers perceived potential benefit across a broad spectrum of holistic care functions. Providers, on average, reported a growth mindset and confidence in their ability to learn mHealth tools and in its potential to improve care access. Key facilitators included alignment with patient needs, increased accessibility, and cost-effectiveness, while barriers included disparities in technology access, digital health literacy, and data security and privacy. Exploratory analyses showed some significant group differences by provider role, provider age, and patient age. Psychosocial providers were significantly more likely to recommend or use apps for pain management (χ^2_1=14.34, P<.001, ϕ=0.28), mental health (χ^2_1=50.54, P<.001, ϕ=0.53), and sleep health (χ^2_1=25.47, P<.001, ϕ= 0.38).

XSL•FO
RenderX

Psychosocial providers also perceived higher benefit for sleep health apps (χ^2_1=6.40, *P*=.01, ϕ=0.19). Medical providers were significantly more likely to perceive medication management apps as potentially beneficial (χ^2_1=10.93, *P*<.001, ϕ=0.25). Older providers (16/88, 18.2%) and adult care providers (8/32, 25%) were significantly more likely to recommend or use disease management apps compared to younger providers (5/100, 5%; χ^2_1=8.20, *P*=.004, ϕ=0.21) and pediatric care providers (6/101, 5.9%; χ^2_2=9.22, *P*=.01, Cramer V=0.22), respectively. Pediatric care providers (83/101, 82.2%) were more likely to recommend or use medical team communication apps compared to adult care providers (15/32, 46.9%; χ^2_2=15.66, *P*<.001, Cramer V=0.29).

Conclusions: Our study underscores the opportunity to develop inclusive mHealth solutions tailored to the diverse needs of individuals across the cancer care continuum, including those in active treatment and survivorship care. Engaging diverse medical and psychosocial providers is essential to inform clinical integration of mHealth technologies in oncology care.

(J Med Internet Res 2026;28:e85583) doi:10.2196/85583

KEYWORDS

digital health technology; digital health technologies; mobile application; mobile app; mobile health; mHealth; health personnel; oncology, psycho-oncology; delivery of health care; health care delivery

Introduction

Over the past decade, the worldwide adoption and use of smartphone technologies has grown exponentially. The total number of smartphone users exceeds 6.7 billion worldwide [1]. Smartphones have become ubiquitous, regardless of gender, race, ethnicity, level of education, and socioeconomic status [2-5]. Over 350,000 mobile health (mHealth) apps (software apps downloadable on mobile devices) are available for public download on Google Play and Apple App Store [6]. mHealth apps have the potential to reduce health disparities for underresourced populations and to improve accessibility for those with immunocompromised or debilitating health conditions [7-15] .mHealth apps may perform a wide range of health care delivery and information dissemination functions.

A 2016 systematic review identified 539 mHealth oncology apps available for public download [16]. Content categories across the oncology apps included teaching and information, medication help, monitoring, prevention, screening, smoking cessation, and patient support. A more recent 2021 systematic review identified 794 mHealth oncology apps available for public download [17], a 47% increase in new apps created over a 5-year time frame. Content categories across oncology apps included education, clinical decision support, and patient support. However, most available oncology apps have not been developed or tested in research settings, and few have Food and Drug Administration (FDA) approval as “digital therapeutics” [16-23]. A 2020 systematic review identified only nine publicly available oncology apps with scientific publications supporting their feasibility or efficacy, suggesting an industry-research gap [18].

Recent research suggests that patients with cancer are interested in using mHealth apps. Studies have found positive patient attitudes toward mHealth apps for a range of functions over the cancer care continuum [24-27], such as medication adherence, early detection and diagnostics, patient-provider communication, lifestyle interventions, weight and diet management, exercise facilitation, symptom monitoring, medical health literacy, disease self-management, and medical decision-making [28-38]. Results of a 2018 oncology patient survey showed that over 60% of survey respondents regularly used the internet to research information about disease prevention, healthy lifestyles, health care, and their cancer treatment [39]. Over 80% of respondents stated that they would download and use an app if their health care provider recommended it to them, highlighting the importance of health care provider involvement in patient adoption of mHealth tools [39].

To date, relatively few studies have examined oncology health care provider attitudes toward mHealth apps [21,40,41]. An international survey of oncologists showed that 80% of respondents used mHealth, primarily for the following functions: to conduct literature reviews, to interact with colleagues, for patient communication, to access patient electronic health records, to collect test results, and for follow-up care [40]. A single-site provider survey of oncologists and nurses in Germany found that the mHealth functions rated as most useful were appointment and medication reminders, assessment of side effects, and collecting patient-reported outcomes [41]. Most survey respondents considered mHealth apps to be a welcome complement to traditional health care and perceived benefits to implementing mobile apps as part of the standard of care. Provider concerns endorsed by a minority of respondents included legal/medical responsibility, data privacy, and data security. A qualitative interview study [21] found that oncology providers expressed openness to mHealth apps and acknowledged their potential to fill care gaps. Perceived implementation barriers included limited access to technologies, time burden, and concerns about the trustworthiness and credibility of app developers [21]. Providers prioritized app usability and patient preferences over app efficacy or effectiveness [21].

Early research on provider attitudes toward mHealth apps focused primarily on oncologist perspectives and disease management functions. Here, we filled a research gap by incorporating perspectives from providers across diverse health care roles, including medical and psychosocial providers. We included providers who work with patients across the age continuum who may have different attitudes toward and familiarity with mHealth. We explored feedback on mHealth content that extends beyond disease management to include

XSL•FO
RenderX

behavioral health and holistic patient care. Holistic patient care is a "whole person" approach to medicine that encompasses physical, mental, emotional, lifestyle, and social needs—not just treating symptoms of disease—in order to promote overall well-being and improve long-term health outcomes [42-44]. This broader focus is critical for understanding how technology can support a more comprehensive approach to patient outcomes. Consequently, this study sought to report an update and extension of prior research by answering the following research questions (RQs): (1) What types of apps have oncology providers used with their patients or would consider potentially beneficial? (2) What are the attitudes, beliefs, facilitators, and barriers related to the adoption of mHealth apps in oncology settings? (3) (Exploratory) How do provider preferences differ based on provider role (medical vs psychosocial), provider age (<45 vs ≥45 years), and patient populations served (pediatric vs adult)? The selection of exploratory variables was informed by existing mHealth and technology acceptance literature [45-50]. Findings will inform recommendations for apps that providers perceive to have clinical utility, future directions for app development, and methods to promote implementation and uptake of mHealth apps.

Methods

Ethical Considerations

All study procedures were approved by the Seattle Children's Hospital Institutional Review Board (IRB approval number STUDY00004962; approval date June 11, 2024). The study was conducted under IRB-exempt status in which survey information obtained is recorded in a manner that the identity of the survey respondents cannot be ascertained. Survey participation was voluntary and anonymous, and no compensation was offered for completing the survey. A one-page REDCap (Research Electronic Data Capture) Study Information Sheet containing purpose of the study, investigator information, study description, confidentiality statement, data storage details, and contact information was provided on the main screen of the online questionnaire. Survey respondents were informed via the Study Information Sheet that their responses would remain anonymous and research records would be stored exclusively within secure institutional intranet databases. The survey was administered through Seattle Children's Hospital REDCap, a secure web-based survey platform that uses password-protected access and role-based permissions to prevent unauthorized data access.

Study Design

We disseminated an online survey called the "Oncology Provider Experiences National Survey (OPENS): Mobile Health" (hereafter referred to as "OPENS") requesting responses on provider attitudes and experiences toward the use of mHealth apps in oncology. The target population for this study consisted of health care professionals currently practicing in oncology settings within the United States. Participation was open to all who self-identified as an oncology provider. Although we did not include an eligibility screener, eligibility was controlled with targeted recruitment through oncology provider list servers (listservs). Our survey was distributed nationally across the United States to members of the American Psychosocial Oncology Society (APOS) and provider listservs at the home institutions of several of the coauthors (eg, Seattle Children's Hospital Cancer and Blood Disorders Center, Fred Hutchinson Cancer Center, MD Anderson Cancer Center, St Jude Children's Research Hospital). We distributed the survey to several organizations that are pediatric-focused: Children's Oncology Group (COG) Behavioral Science Committee, Society of Pediatric Psychology Hematology/Oncology/Bone Marrow Transplant Special Interest Group, and Association of Pediatric Oncology Social Workers (APOSW). Permission to distribute our online survey was obtained by research committees or administrators, as appropriate for each respective listserv. Each listserv was sent a single request with a designated survey end date and with no follow-up reminders. We staggered survey distribution pending submission and review of materials by each listserv's respective research committee and administrators. Prospective participants were informed via the survey dissemination email and the REDCap Study Information Sheet that the survey was intended for oncology providers only. The survey took approximately 10 minutes to complete. The survey remained open from June 2024 to December 2024. This study was reported in accordance with CHERRIES (Checklist for Reporting Results of Internet E-Surveys). See Multimedia Appendix 1 for the complete checklist.

Oncology Provider Experiences National Survey: Mobile Health

To the best of our knowledge, no validated measure currently exists to assess oncology provider attitudes toward mHealth apps. We developed the 30-item OPENS to describe mHealth-related usage, attitudes, perceived benefit, facilitators, and barriers. OPENS was designed as a descriptive survey to assess provider perspectives and usage patterns rather than as a psychometric scale intended to measure a single latent construct. These survey items directly address our study aims to inform the clinical integration of mHealth technologies in oncology care. Internally developed surveys are a standard approach in digital health research when established measures are unavailable [51-54]. Survey items were collaboratively developed by a multidisciplinary research team (including experts in psychology, psychiatry, social work, oncology, and digital health) following a targeted literature review. We established content validity using a modified Delphi approach [55-58]. This process involved asynchronous feedback in which collaborators independently reviewed survey items for clarity and relevance to both the study aims and the oncology clinical context. Facilitated group consensus meetings were held with external oncology experts to refine survey items. Eight items assessing mHealth attitudes and beliefs were adapted from the "Questionnaire to Assess Public Health Staff Viewpoints about Mobile Phone Use in the Delivery of Their Services" [59]. These items were updated for oncology relevance. Survey items were finalized asynchronously among the research team members.

Before national dissemination, the survey underwent pilot-testing with a small group of oncology providers to assess technical functionality within REDCap, item comprehension, and user burden. All 33 items were distributed across 4 pages and administered in a forced-response format to ensure complete

XSL•FO
RenderX

data capture. Questions included single-choice items, multiple-choice items, “5-point Likert scale” items (eg, strongly disagree to strongly agree), and “select all that apply” items (including “none of the above” and “other” options). Respondents were able to review and modify their responses prior to survey submission, with no time limits or automatic expiration imposed for completion. Consistent with the scoring method of the validated instrument from which eight survey items were adapted [48], Likert-scale responses were summarized using means (SDs) to characterize the strength and variability of providers’ beliefs and attitudes on mHealth apps. The brief close-ended survey was designed to reduce participant burden and maximize response rates for uncompensated national listserv recruitment methods, targeting a completion time of approximately 10 minutes. To ensure that our study met IRB-exempt criteria, we did not include open-ended survey questions in which human subjects’ identifying information may have inadvertently been obtained. See Multimedia Appendix 2 for all survey items.

Statistical Analysis

We summarized provider demographics and other survey items using descriptive statistics and data visualizations created in Canva. Exploratory statistical analyses were performed using IBM SPSS Statistics (version 29). Study results are presented as frequency counts, percentages, and means (SDs). We also conducted exploratory analyses on survey responses broadly categorized by provider type (medical providers vs psychosocial providers) who may differ in their attitudes toward and the use of mHealth tools for disease management and behavioral/emotional support based on their training backgrounds and areas of expertise. For the purposes of this study, medical providers were defined as those whose primary training and clinical focus address disease treatment and the management of physiological symptoms. This group included those who identified their role as oncologist, nurse or advanced practice provider, physician, or health care administrator. Psychosocial providers were defined as clinicians whose primary training and clinical focus address the psychological, social, and emotional aspects of the cancer experience. This group included those who identified their role as psychologist, social worker, mental health therapist, or occupational therapist. This categorization also allowed for a balanced distribution of providers for comparison (90/188, 47.9%, medical vs 89/188, 47.3%, psychosocial). We performed chi-square analyses to examine differences between medical providers and psychosocial providers on the use of mHealth apps in clinical practice, mHealth apps endorsed as potentially helpful or beneficial, and the perceived facilitators of and barriers to mHealth app adoption. We also performed chi-square analyses to examine differences between younger providers (age≤45 years) vs older providers (age≥46 years) and pediatric care providers (children, teens, or young adults) vs adult care providers (adults, older adults) on the use of mHealth apps in clinical practice. Given that the study used categorical data, traditional assumptions of normality were not applicable. As all chi-square analyses were exploratory and intended for hypothesis generation rather than formal hypothesis testing, we did not control for multiple comparisons.

Results

Demographic and Provider Role Characteristics

A total of 188 providers responded to the survey (Table 1). Participants ranged in age from under 24 to over 65 years. Over one-third (70/188, 37.2%) were between the ages of 36 and 45 years. Most providers identified as female (150/188, 79.8%), White (161/188, 85.6%), and not Hispanic or Latino (174/188, 92.6%).

XSL•FO
RenderX

Table 1. Demographic characteristics of oncology providers (N=188).

Characteristics	Respondents, n (%)
Age group (years)	
<24	2 (1.1)
24-35	28 (14.9)
36-45	70 (37.2)
46-55	42 (22.3)
56-65	29 (15.4)
>65	17 (9.0)
Gender identity	
Female	150 (79.8)
Male	37 (19.7)
Prefer not to respond	1 (0.5)
Race	
American Indian or Alaska Native	2 (1.1)
Asian	13 (6.9)
Black or African American	0 (0.0)
Hispanic or Latino	7 (3.7)
Middle Eastern or North African	1 (0.5)
Native Hawaiian or Pacific Islander	0 (0.0)
White	161 (85.6)
Not listed	2 (1.1)
Prefer not to respond	4 (2.1)
Ethnicity	
Hispanic or Latino	12 (6.4)
Not Hispanic or Latino	174 (92.6
Prefer not to respond	2 (1.1)

Of the 188 respondents, 47.9% (90/188) were medical providers, 47.3% (89/188) were psychosocial providers, and 4.8% (9/188) had administrative staff or research roles. Psychologists represented the largest group of survey respondents (66/188, 35.1%), followed by oncologists (54/188, 28.7%) and nurses or advanced practice providers (32/188, 17%), as shown in Table 2. Most providers practiced in academic or teaching hospitals (155/188, 82.4%) and urban settings (130/188, 69.1%). Over half of the sample provided care for children (120/188, 63.8%), teens (124/188, 66%), and young adults (143/188, 76.1%).

XSL•FO
RenderX

Table 2. Role characteristics of oncology providers (N=188).

Characteristics	Respondents, n (%)
Current role[a]	
Oncologist	54 (28.7)
Nurse or advanced practice provider	32 (17.0)
Psychologist	66 (35.1)
Social worker	13 (6.9)
Health care administrator	2 (1.1)
Researcher	33 (17.6)
Physician (other specialty)	4 (2.1)
Other	14 (7.4)
Practice setting[a]	
Private practice	11 (5.9)
Academic or teaching hospital	155 (82.4)
Community hospital	16 (8.5)
Hospice care	0 (0.0)
Other	15 (8.0)
Practice location	
Rural	15 (8.0)
Suburban	35 (18.6)
Urban	130 (69.1)
Other	8 (4.3)
Duration of experience providing care (years)	
<1	1 (0.5)
1-5	39 (20.7)
6-10	41 (21.8)
11-15	38 (20.2)
>16	68 (36.2)
Prefer not to respond	1 (0.5)
Patient age demographic[a]	
Children	120 (63.8)
Teens	124 (66.0)
Young adults	143 (76.1)
Adults	82 (43.6)
Older adults	50 (26.6)

[a]Questions on the respondents' current role, practice setting, and patient age demographic were multiple-response items; providers could select all applicable options.

mHealth Apps Used in Clinical Practice

Nearly all providers (178/188, 94.7%) reported recommending or using at least one mHealth app with their patients (Figure 1). Most providers recommended or used telehealth and virtual consultation apps (139/188, 73.9%). Other types of mHealth apps that were commonly recommended to or used with patients included communication with the medical team (136/188, 72.3%) and mental health and wellness (98/188, 52.1%) apps. Only a small proportion of providers reported recommending or using cancer and disease management (21/188, 11.2%) or health literacy (15/188, 8%) mHealth apps with their patients. Both medical and psychosocial providers predominantly used mHealth apps for medical team communication or telehealth (Figure 2 and Figure 3, respectively). Psychosocial providers (25/89, 28.1%) were significantly more likely than medical providers (6/90, 6.7%) to recommend or use pain management

XSL•FO
RenderX

apps (χ^2_1=14.34, P<.001, ϕ=0.28). Psychosocial providers (70/89, 78.7%) were significantly more likely than medical providers (23/90, 25.6%) to recommend or use mental health apps (χ^2_1=50.54, P<.001, ϕ=0.53). Psychosocial providers (48/89, 53.9%) were significantly more likely than medical providers (16/90, 17.8%) to recommend or use sleep health apps with their patients (χ^2_1=25.47, P<.001, ϕ=0.38). All other chi-square analyses comparing medical and psychosocial provider groups were nonsignificant.

Figure 1. Comparison of mHealth apps used in clinical practice vs potentially useful or beneficial mHealth apps for patients. mHealth: mobile health.

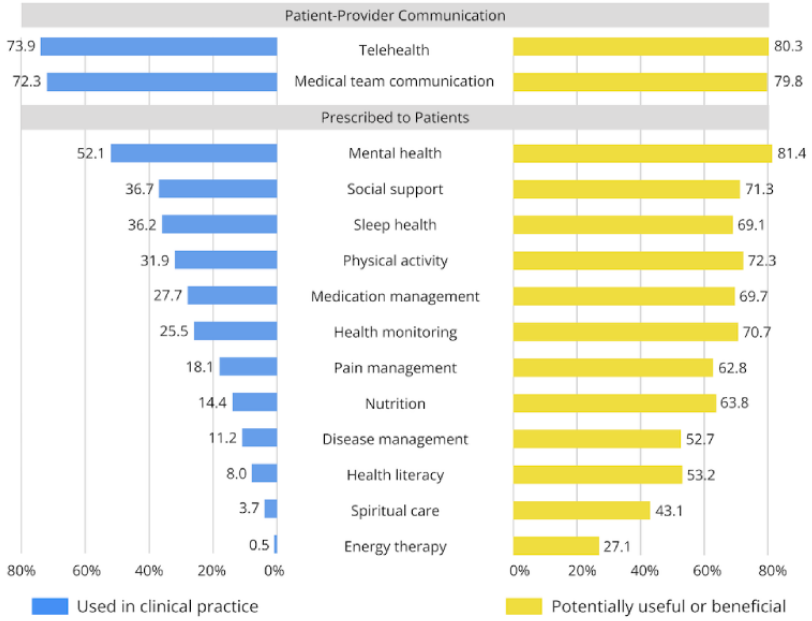

Figure 2. Comparison of mHealth apps used in clinical practice vs potentially useful or beneficial mHealth apps for medical providers. mHealth: mobile health.

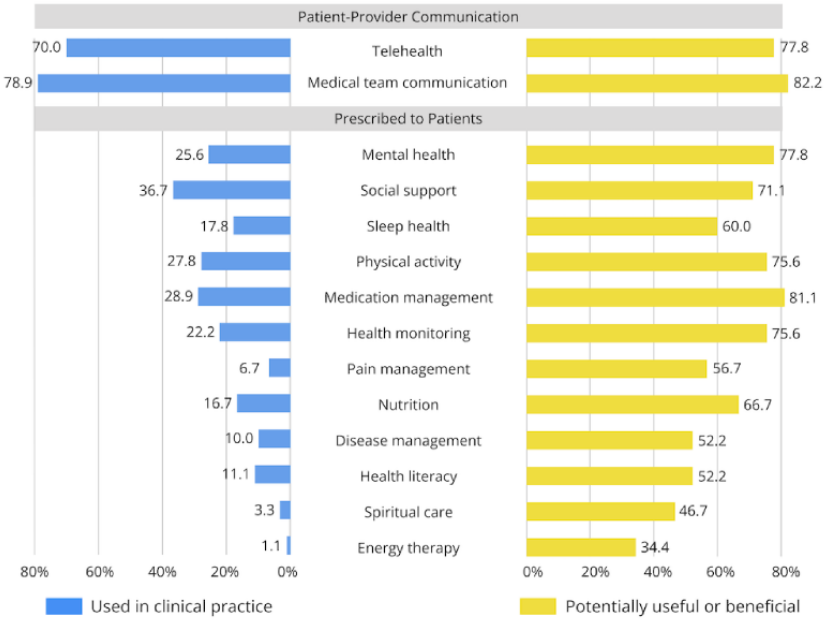

XSL•FO
RenderX

Figure 3. Comparison of mHealth apps used in clinical practice vs potentially useful or beneficial mHealth apps for psychosocial providers. mHealth: mobile health.

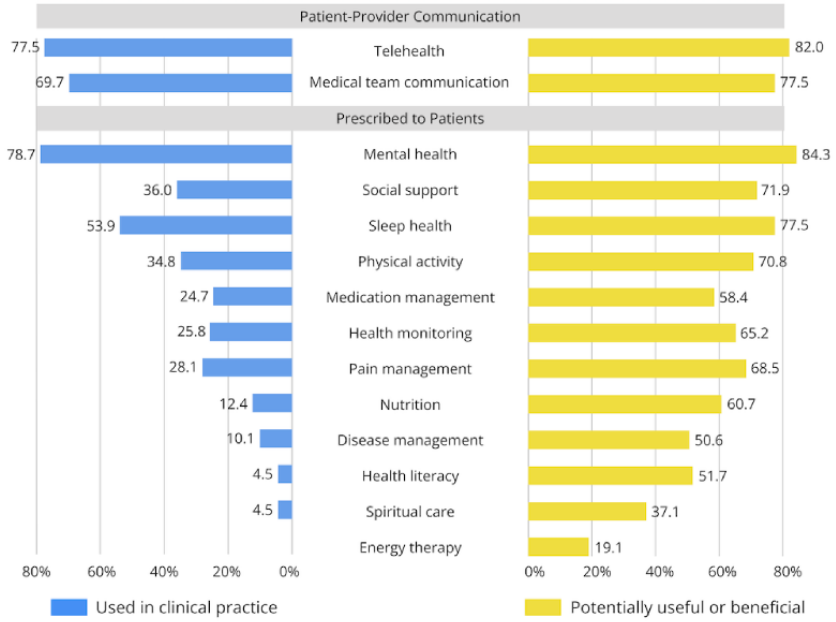

Providers stratified by age (≤45 and ≥46 years) recommended or used similar types of mHealth apps with patients (Multimedia Appendix 3). Older providers (16/88, 18.2%) were significantly more likely than younger providers (5/100, 5%) to recommend or use disease management apps with their patients (χ^2_1=8.20, *P*=.005, ϕ=0.21). All other chi-square analyses comparing older and younger providers were nonsignificant. Providers working with pediatric and adult patients also recommended or used similar types of mHealth apps (Multimedia Appendix 4). Adult care providers (8/32, 25%) were significantly more likely than pediatric care providers (6/101, 5.9%) to recommend or use disease management apps (χ^2_2=9.22, *P*=.01, Cramer V=0.22). Pediatric care providers (83/101, 82.2%) were more likely than adult care providers (15/32, 46.9%) to recommend or use medical team communication apps (χ^2_2=15.66, *P*<.001, Cramer V=0.29). All other chi-square analyses comparing adult and pediatric care providers were nonsignificant.

Potentially Useful or Beneficial mHealth Apps

All providers endorsed mHealth apps to be potentially useful or beneficial for patients (Figure 1). Most providers perceived mental health and wellness apps to be potentially useful (153/188, 81.4%). Other types of mHealth apps that were perceived as potentially useful were telehealth and virtual consultation (151/188, 80.3%), communication with the medical team (150/188, 79.8%), exercise and physical activity (136/188, 72.3%), social support (134/188, 71.3%), health monitoring and medical symptom tracking (133/188, 70.7%), medication management (131/188, 69.7%), and sleep health (130/188, 69.1%). Although providers predominantly used mHealth apps for patient communication (eg, medical team communication, telehealth), they endorsed interest in prescribing mHealth apps for a broad range of holistic care functions.

Interest in a wide array of psychosocial support functions was fairly consistent across medical and psychosocial providers (Figure 2 and Figure 3, respectively). Medical providers (73/90, 81.1%) were significantly more likely than psychosocial providers (52/89, 58.4%) to perceive medication management apps as potentially beneficial for their patients (χ^2_1=10.93, *P*<.001, ϕ=0.25). Medical providers (31/90, 34.4%) were significantly more likely than psychosocial providers (17/89, 19.1%) to perceive energy therapy apps as potentially beneficial (χ^2_1=5.37, *P*=.02, ϕ=0.17). In contrast, psychosocial providers (69/89, 77.5%) were significantly more likely than medical providers (54/90, 60%) to endorse sleep health apps as potentially helpful for their patients (χ^2_1=6.40, *P*=.01, ϕ=0.19). All other chi-square analyses comparing medical and psychosocial provider groups were nonsignificant.

mHealth Attitudes and Beliefs

Providers had positive attitudes and beliefs toward specific facets of mHealth apps (Table 3). Providers had the highest mean scores for specific items pertaining to health equity, indicating that they "agree" or "strongly agree" that mHealth apps reduce travel costs for services (mean 4.18/5, SD 0.68) and expands the range of health services (mean 4.18/5, SD 0.69). Providers had the highest mean scores for specific items pertaining to confidence and a growth mindset, indicating that they "agree" or "strongly agree" that they are confident in their ability to navigate mobile apps (mean 4.14/5, SD 0.83) and can adapt to health technology changes over time (mean 4.18/5, SD 0.70). Providers had the lowest mean scores for specific items pertaining to the patient-provider relationship and training, indicating that they "neither agree nor disagree" that mHealth improves provider-patient interactions (mean 3.21/5, SD 0.95) or are trained and informed about mHealth regulatory and ethical issues in health care (mean 3.24/5, SD 1.07).

XSL•FO
RenderX

Table 3. Attitudes and beliefs toward mHealth[a] apps among oncology providers.

Item	Mean (SD)
Training	
I feel that I am trained and informed about regulatory and ethical issues related to technology use in health care.	3.24 (1.07)
I can effectively teach patients how to use mHealth technologies for their care.	3.44 (1.03)
I am interested in receiving more training or specialized training in the use of digital tools in my health care practice.	3.88 (0.96)
Confidence and growth mindset	
I can adapt to health technology changes over time.	4.18 (0.70)
I feel confident in my ability to navigate mobile apps.	4.14 (0.83)
I am comfortable incorporating technology into patient care practices.	3.74 (0.90)
Quality of care	
The use of mobile technology improves the quality of public health services.	3.92 (0.68)
Mobile technology can help improve patients' quality of life.	3.94 (0.66)
Mobile technology improves self-management in public health care professionals.	3.65 (0.76)
Mobile technology can improve patients' health literacy and encourage them to actively engage in their care.	3.86 (0.73)
Health equity	
Using mobile technology reduces the cost of providing public health services.	3.64 (0.83)
The use of mobile technology expands the range of health services.	4.18 (0.69)
The use of mobile technology creates equal access to facilities and services for the general public.	3.31 (0.98)
Using mobile technology reduces the number of visits to health centers.	3.77 (0.83)
The use of mobile technology reduces travel costs for services.	4.18 (0.68)
Patient-provider relationship	
Mobile technology improves interpersonal interactions between the provider and the recipient.	3.21 (0.95)
Integrating mobile technology can improve patient-provider interactions.	3.81 (0.70)

[a]mHealth: mobile health.

Facilitators of mHealth Adoption

The top five perceived facilitators of mHealth adoption in clinical settings were "alignment with patients' expressed interest and needs" (147/188, 78.2%), "improved health care accessibility" (135/188, 71.8%), "cost-effectiveness for patients and health care systems" (127/188, 67.6%), "seamless integration with existing health care systems and/or electronic medical records" (113/188, 60.1%), and "improved job performance and efficiency" (109/188, 58%); see Multimedia Appendix 5. Medical and psychosocial providers reported similar top facilitators overall (Figure 4). However, medical providers (47/90, 52.2%) were significantly more likely than psychosocial providers (23/89, 25.8%) to endorse patient health literacy as a perceived facilitator (χ^2_1=13.08, P<.001, ϕ=0.27).

Figure 4. Common facilitators of adopting mHealth apps in health care settings among medical and psychosocial providers. mHealth: mobile health.

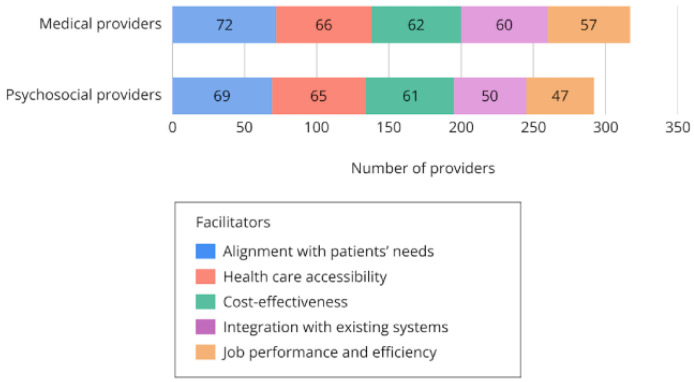

XSL•FO
RenderX

Barriers to mHealth Adoption

The top five perceived barriers to mHealth adoption in clinical settings were "disparities in access to technology" (132/188, 70.2%), "digital literacy" (ie, technical knowledge to access, understand, and use digital resources and tools; 117/188, 62.2%), "data privacy and safety concerns" (114/188, 60.6%), "lack of knowledge regarding which digital tools to recommend or where to find relevant information" (110/188, 58.5%), and "being overwhelmed by too many digital health technology options" (105/188, 55.9%); see Multimedia Appendix 6. Medical and psychosocial providers reported similar top barriers overall (Figure 5). However, medical providers (52/90, 57.8%) were more likely than psychosocial providers (38/89, 42.7%) to endorse cost as a perceived barrier (χ^2_1=4.07, P=.04, ϕ=0.15).

Figure 5. Common barriers to adopting mHealth apps in health care settings among medical and psychosocial providers. mHealth: mobile health.

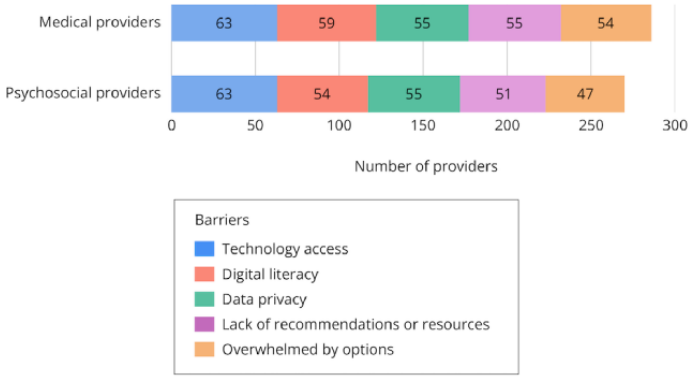

Discussion

Principal Findings

This study provided an update to and extension of prior research regarding the integration of mHealth apps in oncology care. Regarding mHealth app usage and its perceived benefit (RQ1), oncology providers commonly use mHealth apps, with an emphasis on patient-provider communication tools. Although current usage is primarily logistical, there is clear interest in whole person health–supporting apps that providers consider potentially beneficial but that currently see lower rates of adoption. With regard to provider attitudes, facilitators, and barriers related to mHealth adoption (RQ2), providers consistently endorsed a growth mindset, reporting high confidence in their ability to learn and adopt digital tools. Providers recognized the potential for mHealth to improve health equity by expanding access to care. The top perceived facilitators of mHealth adoption were alignment with patient needs, improved health care accessibility, and cost-effectiveness. The top perceived barriers to mHealth adoption were disparities in access to technology, digital health literacy, and data privacy.

With regard to exploratory differences by clinical role, provider age, and patient age (RQ3), we found several group differences. Reflecting their respective training backgrounds, medical providers were more likely to endorse the potential benefit of medication management apps, whereas psychosocial providers were more likely to use and/or perceive potential benefit in pain management, mental health, and sleep health apps. The use of mHealth apps was generally consistent across oncology provider age groups (≤45 and ≥46 years) and patient population (pediatric care providers and adult care providers). However, older providers (age≥46 years) and those providing care to adult patients used more disease management tools, whereas pediatric providers used more medical team communication apps.

Comparison With Prior Work

Our study expands on the background literature on oncology providers' attitudes toward mHealth apps by soliciting perspectives of those in various clinical roles, not exclusively oncologists; expanding mHealth app functions beyond disease management; and identifying perceived barriers to and facilitators of implementation in oncology care settings. Our findings are consistent with previous studies that have reported that oncologists primarily use mHealth tools for telehealth and medical team communication and rarely use mHealth tools for disease management and follow-up care [40]. Of note, we also found that medical and psychosocial providers alike are interested in using mHealth apps focusing on a wide range of psychosocial care functions to their patients and perceived them to have clinical utility. This was consistent with the literature on patients with cancer and those who have survived cancer, showing high rates of acceptance of mHealth apps but low rates of adoption [60-64]. Specific patient facilitators and barriers beyond demographic and disease characteristics have not yet been examined. Studies examining patient preferences have predominantly focused on disease management apps [60,62-66]. Consistent with the background literature, we found that a primary barrier identified by providers is disparities in access to technology [21,40,67]. However, research has shown that mHealth technologies are universally accessible to individuals regardless of income level, rurality, or marginalized status [3,4,68]. In the United States, 79%-94% of individuals from low-income households (<US $30,000) have access to smartphones, computers, and tablets [3,4,69]. Digitally deployed health interventions have also been proposed for populations/persons experiencing homelessness as viable alternatives to in-person services to overcome traditional access care barriers [68,70,71]. In the United States, 91% of individuals own a smartphone, 95% use the internet, and 90% have a high-speed internet subscription at home [1,3,4,69]. Universal use of smartphones holds across the age continuum, with 95%

XSL•FO
RenderX

of teens/young adults, 95%-97% of adults aged>45 years [2-4,72], and 76%-79% of older adults aged>65 years owning a smartphone [4,69,72]. Digital technologies (smartphones, computers, tablets) have become significantly more affordable in recent years, and their rapid adoption has been further accelerated by the COVID-19 pandemic.

Consistent with the background literature, a primary barrier to mHealth adoption, as identified by providers, was limited digital health literacy [61]. Previous studies have shown that health literacy and digital health literacy are significant mediators of the efficacy of digital health interventions [73] and are not fixed but malleable patient characteristics [74]. Thus, providers may be more likely to recommend and use mHealth apps with intuitive user interfaces, accessibility for those with impairments, provision of health information at lower reading levels, and multimedia options (eg, text-to-speech and animations) [61,75,76]. Recommendations for mHealth may also be paired with apps that identify a user's health literacy level and provide tailored definitions of medical terminologies to their reading level (eg, MediReader [77]).

Previous research has found that younger oncologists (<45 years old) and younger cancer patients and survivors (<45 years old) are more likely to use internet-enabled mHealth tools for disease management and cancer care [40]. Our results showed consistent use of mHealth apps in younger as well as in older providers (>45 years old) and among providers of patients across the age continuum. For younger pediatric patients, caregivers may primarily interact with mHealth apps, while teens and young adults may be more independent app users [78]. Digital inclusion for older adults is recognized as an important social determinant of health and a crucial component of a healthy aging society [79-88]. These findings highlight the importance of developing, testing, and implementing mHealth apps that are tailored to the diverse needs of patients with cancer and those who have survived cancer, of all ages.

The Technology Acceptance Model (TAM) proposes that perceived usefulness (ie, the belief that using a particular technology enhances job performance) and perceived ease of use (ie, the belief that using a particular technology is effort free) are the primary drivers of user adoption of new technologies [89,90]. Interpreted through the TAM, our findings suggest high readiness for mHealth adoption across the oncology workforce. Specifically, providers' endorsement of mHealth apps for reducing costs and expanding care access aligns with high perceived usefulness. Providers' reported growth mindset and confidence in their ability to learn and adapt to mHealth technologies reflect high self-efficacy, which aligns with high perceived ease of use. Although TAM pathways were not formally tested in our study, the alignment between our findings and these TAM constructs suggest that oncology providers have the positive attitudes and behavioral intentions necessary to facilitate mHealth adoption.

Future Directions

mHealth apps that address holistic patient care have the potential to improve patient outcomes across the cancer care continuum and patient age continuum. Despite the narrow use of mHealth apps in current practice, the perceived utility of mHealth across medical and psychosocial domains suggests important opportunities for broader integration. Guided by the RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) framework [91-93], future research should focus on expanding *reach* by partnering with medical and psychosocial providers on codesign strategies of mHealth tools in the innovation design-to-implementation pipeline. Future research should also enhance *effectiveness* via rigorous clinical testing and addressing the complex ethical and regulatory landscape of mHealth apps. Although oncology providers are interested in using mHealth tools, the lack of standardized clinical guidelines for recommending non-FDA-approved apps remains an important barrier. Endorsing or recommending mHealth clinical tools that lack rigorous scientific testing poses risks to patient safety and data confidentiality. Future research and policy work should focus on developing clinical guidelines to identify evidence-based mHealth apps and pathways for determining their clinical efficacy/effectiveness. In addition, future research should support *adoption* and *implementation* by examining providers' nuanced perspectives on the perceived barrier of technology access and investigate the specific sources of these concerns (eg, financial burden of cancer, lack of reliable cellular or internet service in rural areas). Efforts to mitigate barriers, such as concerns about data privacy and security, will also be critical to ensuring adoption of mHealth solutions. Successful integration of mHealth apps into oncology care will require accessible, user-centered design (eg, intuitive interfaces, appropriate reading levels, multimedia supports) and shared decision-making informed by ongoing examination of concordant and discordant beliefs and attitudes toward mHealth apps among providers and patients. Long-term *maintenance* will depend on sustained institutional and policy support (reimbursement pathways, standardized evaluation criteria, regulatory guidance on data privacy and security), engagement of local providers or stakeholder champions to promote uptake through education and shared learning in navigating digital health systems, and preservation of alternative care options (eg, in-person services) as technologies continues to evolve. Future research using a mixed methods approach is also warranted to provide richer context to these exploratory findings.

Limitations

This study has several limitations. First, this is a convenience sample of oncology providers, which introduces the potential for selection and response bias due to voluntary participation. It is possible that providers with a pre-existing interest in mHealth were more motivated to participate in the study, potentially leading to nonresponse bias where favorable attitudes are overrepresented. In addition, social desirability bias may have influenced participants to overreport favorable attitudes to align with current digital health trends. To mitigate potential bias, we distributed the survey through multiple national professional listservs representing diverse professional roles. In addition, survey questions assessed diverse app functions, barriers, and facilitators, which resulted in variability in study findings. This suggests that the survey captured an array of provider perspectives. Second, most respondents identified as White and female, practiced in urban settings, and did not see older adults in their clinical practice. This limits the

XSL•FO
RenderX

generalizability and transferability of our findings to underrepresented providers because of a lack of diversity in our sample. Future studies should implement recruitment strategies to target provider demographic diversity, rural health settings, and those in older adult and geriatric care to ensure findings are representative of all clinical contexts. Third, this is a small sample relative to the number of oncology providers who practice nationally, and we relied on our institutional and national professional organization listservs to distribute the survey. Sample size was determined by convenience sampling within the study recruitment time frame. As the primary aim of this study was descriptive, we did not establish a predetermined target sample size. Fourth, as there are no well-validated measures of provider attitudes toward mHealth interventions with established cutoff scores, we used an internally developed survey. This influenced the study by limiting our ability to categorize and compare scores with existing studies. To mitigate the lack of a pre-existing validated scale, we used a modified Delphi process involving a multidisciplinary team and external oncology experts to ensure content validity. Fifth, to comply with IRB-approved protocols and IRB-exempt status, we did not collect identifiers from survey respondents, so it is not possible to determine response rates from specific organizations or listservs or analyze group differences based on affiliations. Future research could include such identifiers in order to assess representativeness and potential cohort effects. Sixth, due to the exploratory nature of our analyses, we did not control for multiple comparisons or include demographic covariates. Although this approach increases the risk of type 1 errors, we mitigated the risk of overinterpretation by explicitly framing these analyses as hypothesis generating, and our findings should be viewed as preliminary. Finally, this was a cross-sectional study who findings provide a snapshot in time of provider attitudes in a post-COVID-19 pandemic landscape that has normalized digital health care. Because attitudes and perspectives may continue to shift over time and technology evolves rapidly, future longitudinal research is warranted.

Conclusion

The findings of this study reveal a significant “readiness gap” in oncology care. Although current use remains largely focused on telehealth services and patient-provider communication, providers expressed interest in integrating a wider array of mHealth apps encompassing health information, disease and symptom monitoring, mental health resources, and psychosocial support. Our findings suggest that the oncology workforce is prepared to move beyond logistical communication tools toward an integrated holistic model of oncology care that encompasses broad disease management and psychosocial support functions. Because these positive attitudes are consistent across clinical roles and patient populations, mHealth interventions have broad appeal and should be implemented as interdisciplinary standards of care. Health care systems can leverage existing provider buy-in to create a more equitable and accessible oncology care model.

Acknowledgments

We thank the providers who participated in the study and shared their insights.

Data Availability

The study was conducted under Institutional Review Board (IRB)-exempt status, and data generated are not publicly available. Data may be made available from the corresponding author (NL) upon reasonable request and subject to a formal data use agreement.

Funding

NL is supported by the National Cancer Institute at the National Institutes of Health (grant number 1K08CA263474). The opinions herein represent those of the authors and not necessarily the funders.

Authors' Contributions

Conceptualization was handled by NL, KS, HA, MER, KMI, AB, EJC, and TMP; data curation by KS, SJHH, and HA; formal analysis and writing—original draft by NL, KS, and SJHH; funding acquisition by NL; methodology by NL, KS, and TMP; visualization by SJHH; and investigation and writing—review and editing by all authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Checklists for Reporting Results of Internet E-Surveys (CHERRIES).
[DOCX File , 26 KB - jmir_v28i1e85583_app1.docx]

Multimedia Appendix 2
Oncology Provider Experiences National Survey (OPENS): Mobile Health.
[DOCX File , 27 KB - jmir_v28i1e85583_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Types of mobile health (mHealth) apps providers prescribed to or used with patients stratified by provider age (≤45 vs ≥46 years).
[PNG File , 135 KB - jmir_v28i1e85583_app3.png]

Multimedia Appendix 4
Types of mobile health (mHealth) apps providers prescribed to or used with patients stratified by patient age (pediatric vs adult).
[PNG File , 136 KB - jmir_v28i1e85583_app4.png]

Multimedia Appendix 5
Facilitators of adopting mobile health (mHealth) apps in health care settings.
[PNG File , 165 KB - jmir_v28i1e85583_app5.png]

Multimedia Appendix 6
Barriers to adopting mobile health (mHealth) apps in health care settings.
[PNG File , 120 KB - jmir_v28i1e85583_app6.png]

References

1. Laricchia F. Smartphone penetration worldwide as share of global population 2016-2023. Statista. 2024. URL: https://www.statista.com/statistics/203734/global-smartphone-penetration-per-capita-since-2005/ [accessed 2024-11-24]
2. Bashir U. Americans owning a smartphone in the U.S., by age 2024. Statista. 2024. URL: https://www.statista.com/statistics/231612/number-of-cell-phone-users-usa/ [accessed 2025-01-08]
3. Bazen A, Schmitt K. Cell phone statistics 2025. ConsumerAffairs. 2024. URL: https://www.consumeraffairs.com/cell_phones/cell-phone-statistics.html [accessed 2025-01-08]
4. Gelles-Watnick R. Americans' use of mobile technology and home broadband. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/2024/01/31/americans-use-of-mobile-technology-and-home-broadband/ [accessed 2025-01-08]
5. Laricchia F. Smartphone ownership in the U.S. 2024, by gender. Statista. 2024. URL: https://www.statista.com/forecasts/1363291/smartphone-ownership-by-gender-us [accessed 2025-01-08]
6. Digital health trends 2021. IQVIA. 2021. URL: https://www.iqvia.com/insights/the-iqvia-institute/reports-and-publications/reports/digital-health-trends-2021 [accessed 2024-11-24]
7. Buneviciene I, Mekary RA, Smith TR, Onnela JP, Bunevicius A. Can mHealth interventions improve quality of life of cancer patients? A systematic review and meta-analysis. Crit Rev Oncol Hematol 2021 Jan;157:103123 [FREE Full text] [doi: 10.1016/j.critrevonc.2020.103123] [Medline: 33190065]
8. Qudah B, Luetsch K. The influence of mobile health applications on patient - healthcare provider relationships: a systematic, narrative review. Patient Educ Couns 2019 Jun;102(6):1080-1089 [FREE Full text] [doi: 10.1016/j.pec.2019.01.021]
9. Rivers BM, Bernhardt JM, Fleisher L, Green BL. Opportunities and challenges of using technology to address health disparities. Future Oncol 2014 Mar;10(4):519-522 [FREE Full text] [doi: 10.2217/fon.14.17] [Medline: 24754582]
10. Saevarsdottir SR, Gudmundsdottir SL. Mobile apps and quality of life in patients with breast cancer and survivors: systematic literature review. J Med Internet Res 2023 Jul 26;25:e42852 [FREE Full text] [doi: 10.2196/42852] [Medline: 37494111]
11. Schliemann D, Tan MM, Hoe WMK, Mohan D, Taib NA, Donnelly M, et al. mHealth interventions to improve cancer screening and early detection: scoping review of reviews. J Med Internet Res 2022 Aug 15;24(8):e36316 [FREE Full text] [doi: 10.2196/36316] [Medline: 35969450]
12. Sibelius K. 2011 mHealth summit. 2011 Presented at: 3rd Annual mHealth Summit; December 6-8, 2011; Washington, DC URL: https://conference.researchbib.com/view/event/10146 [doi: 10.5040/9798765195659]
13. Steinhubl SR, Muse ED, Topol EJ. Can mobile health technologies transform health care? JAMA 2013 Dec 11;310(22):2395-2396 [FREE Full text] [doi: 10.1001/jama.2013.281078] [Medline: 24158428]
14. Stiles-Shields C, Reyes KM, Archer J, Lennan N, Zhang J, Julion WA, et al. mHealth uses and opportunities for teens from communities with high health disparities: a mixed-methods study. J Technol Behav Sci 2022 Sep 13:1-13 [FREE Full text] [doi: 10.1007/s41347-022-00278-y] [Medline: 36117748]
15. van Veen T, Binz S, Muminovic M, Chaudhry K, Rose K, Calo S, et al. Potential of mobile health technology to reduce health disparities in underserved communities. West J Emerg Med 2019 Aug 06;20(5):799-802 [FREE Full text] [doi: 10.5811/westjem.2019.6.41911] [Medline: 31539337]
16. Brouard B, Bardo P, Bonnet C, Mounier N, Vignot M, Vignot S. Mobile applications in oncology: is it possible for patients and healthcare professionals to easily identify relevant tools? Ann Med 2016 Nov 27;48(7):509-515 [FREE Full text] [doi: 10.1080/07853890.2016.1195010] [Medline: 27348761]
17. Upadhyay VA, Landman AB, Hassett MJ. Landscape analysis of oncology mobile health applications. JCO Clin Cancer Inform 2021 May;5:579-587 [FREE Full text] [doi: 10.1200/CCI.20.00156] [Medline: 34033510]

XSL•FO
RenderX

18. Ana F, Loreto M, José LMM, Pablo SM, María Pilar MJ, Myriam SA. Mobile applications in oncology: a systematic review of health science databases. Int J Med Inform 2020 Jan;133:104001 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.104001] [Medline: 31706229]
19. Pandey A, Hasan S, Dubey D, Sarangi S. Smartphone apps as a source of cancer information: changing trends in health information-seeking behavior. J Cancer Educ 2013 Mar;28(1):138-142 [FREE Full text] [doi: 10.1007/s13187-012-0446-9] [Medline: 23275239]
20. Bender JL, Yue RYK, To MJ, Deacken L, Jadad AR. A lot of action, but not in the right direction: systematic review and content analysis of smartphone applications for the prevention, detection, and management of cancer. J Med Internet Res 2013 Dec 23;15(12):e287 [FREE Full text] [doi: 10.2196/jmir.2661] [Medline: 24366061]
21. Berkowitz CM, Zullig LL, Koontz BF, Smith SK. Prescribing an app? Oncology providers' views on mobile health apps for cancer care. JCO Clin Cancer Inform 2017 Nov;1(1):1-7 [FREE Full text] [doi: 10.1200/CCI.17.00107] [Medline: 30657404]
22. Nasi G, Cucciniello M, Guerrazzi C. The role of mobile technologies in health care processes: the case of cancer supportive care. J Med Internet Res 2015 Feb 12;17(2):e26 [FREE Full text] [doi: 10.2196/jmir.3757] [Medline: 25679446]
23. Wolf JA, Moreau JF, Akilov O, Patton T, English JC, Ho J, et al. Diagnostic inaccuracy of smartphone applications for melanoma detection. JAMA Dermatol 2013 Apr;149(4):422-426 [FREE Full text] [doi: 10.1001/jamadermatol.2013.2382] [Medline: 23325302]
24. Blair CK, Harding E, Wiggins C, Kang H, Schwartz M, Tarnower A, et al. A home-based mobile health intervention to replace sedentary time with light physical activity in older cancer survivors: randomized controlled pilot trial. JMIR Cancer 2021 Apr 13;7(2):e18819 [FREE Full text] [doi: 10.2196/18819] [Medline: 33847588]
25. Chow EJ, Doody DR, Di C, Armenian SH, Baker KS, Bricker JB, et al. Feasibility of a behavioral intervention using mobile health applications to reduce cardiovascular risk factors in cancer survivors: a pilot randomized controlled trial. J Cancer Surviv 2021 Aug 10;15(4):554-563 [FREE Full text] [doi: 10.1007/s11764-020-00949-w] [Medline: 33037989]
26. King-Dowling S, Psihogios AM, Hill-Kayser C, Szalda D, O'Hagan B, Darabos K, et al. Acceptability and feasibility of survivorship care plans and an accompanying mobile health intervention for adolescent and young adult survivors of childhood cancer. Pediatr Blood Cancer 2021 Mar;68(3):e28884 [FREE Full text] [doi: 10.1002/pbc.28884] [Medline: 33416214]
27. Quintiliani LM, Mann DM, Puputti M, Quinn E, Bowen DJ. Pilot and feasibility test of a mobile health-supported behavioral counseling intervention for weight management among breast cancer survivors. JMIR Cancer 2016;2(1):e4 [FREE Full text] [doi: 10.2196/cancer.5305] [Medline: 27761518]
28. Allicock M, Kendzor D, Sedory A, Gabriel KP, Swartz MD, Thomas P, et al. A pilot and feasibility mobile health intervention to support healthy behaviors in African American breast cancer survivors. J Racial Ethn Health Disparities 2021 Feb 8;8(1):157-165 [FREE Full text] [doi: 10.1007/s40615-020-00767-x] [Medline: 32385847]
29. Beck SL, Smith R, Mindes J, Beck K, Leah Kim J, Weitzman M, et al. Feasibility and usability of EnergyPoints: a mobile health app to guide acupressure use for cancer symptom management. Integr Cancer Ther 2024;23:15347354231223965 [FREE Full text] [doi: 10.1177/15347354231223965] [Medline: 38284345]
30. Birkhoff S, Cantrell M, Moriarty H, Lustig R. The usability and acceptability of a patient-centered mobile health tracking app among a sample of adult radiation oncology patients. ANS Adv Nurs Sci 2018;41(3):243-259 [FREE Full text] [doi: 10.1097/ANS.0000000000000202] [Medline: 29474226]
31. Gelgoot EN, Kruglova K, Chan P, Lo K, Rosberger Z, Chown P, et al. Evaluation of a mobile health app offering fertility information to male patients with cancer: usability study. JMIR Cancer 2022 May 04;8(2):e33594 [FREE Full text] [doi: 10.2196/33594] [Medline: 35507408]
32. Lee H, Uhm KE, Cheong IY, Yoo JS, Chung SH, Park YH, et al. Patient satisfaction with mobile health (mHealth) application for exercise intervention in breast cancer survivors. J Med Syst 2018 Nov 06;42(12):254 [FREE Full text] [doi: 10.1007/s10916-018-1096-1] [Medline: 30402781]
33. Lee H, Mtengezo JT, Makin MS, Shi L, Malata A, Fitzpatrick J, et al. Mobile health-delivered narrative intervention to increase cervical cancer screening among Malawian women living with HIV: a pilot randomized controlled trial. Asia Pac J Oncol Nurs 2024 May;11(5):100448 [FREE Full text] [doi: 10.1016/j.apjon.2024.100448] [Medline: 38784066]
34. Pereira-Salgado A, Westwood JA, Russell L, Ugalde A, Ortlepp B, Seymour JF, et al. Mobile health intervention to increase oral cancer therapy adherence in patients with chronic myeloid leukemia (the REMIND system): clinical feasibility and acceptability assessment. JMIR Mhealth Uhealth 2017 Dec 06;5(12):e184 [FREE Full text] [doi: 10.2196/mhealth.8349] [Medline: 29212628]
35. Schrier E, Xiong N, Thompson E, Poort H, Schumer S, Liu JF, et al. Stepping into survivorship pilot study: harnessing mobile health and principles of behavioral economics to increase physical activity in ovarian cancer survivors. Gynecol Oncol 2021 May;161(2):581-586 [FREE Full text] [doi: 10.1016/j.ygyno.2021.02.023] [Medline: 33637350]
36. Somers TJ, Abernethy AP, Edmond SN, Kelleher SA, Wren AA, Samsa GP, et al. A pilot study of a mobile health pain coping skills training protocol for patients with persistent cancer pain. J Pain Symptom Manage 2015 Oct;50(4):553-558 [FREE Full text] [doi: 10.1016/j.jpainsymman.2015.04.013] [Medline: 26025279]

XSL•FO
RenderX

37. Li Y, Chen W, Liang Y, Yang L, Hou L. Evaluation of mobile health technology interventions for the postdischarge management of patients with head and neck cancer: scoping review. JMIR Mhealth Uhealth 2023 Oct 23;11:e49051 [FREE Full text] [doi: 10.2196/49051] [Medline: 37870887]
38. Wang T, Huang R, Yang S, Chou C, Chen L. Evaluating the effects of a mobile health app on reducing patient care needs and improving quality of life after oral cancer surgery: quasiexperimental study. JMIR Mhealth Uhealth 2020 Jul 27;8(7):e18132 [FREE Full text] [doi: 10.2196/18132] [Medline: 32716303]
39. Collado-Borrell R, Escudero-Vilaplana V, Calles A, Garcia-Martin E, Marzal-Alfaro B, Gonzalez-Haba E, et al. Oncology patient interest in the use of new technologies to manage their disease: cross-sectional survey. J Med Internet Res 2018 Oct 23;20(10):e11006 [FREE Full text] [doi: 10.2196/11006] [Medline: 30355554]
40. Tarricone R, Cucciniello M, Armeni P, Petracca F, Desouza KC, Hall LK, et al. Mobile health divide between clinicians and patients in cancer care: results from a cross-sectional international survey. JMIR Mhealth Uhealth 2019 Sep 06;7(9):e13584 [FREE Full text] [doi: 10.2196/13584] [Medline: 31493318]
41. Kessel KA, Vogel MME, Schmidt-Graf F, Combs SE. Mobile apps in oncology: a survey on health care professionals' attitude toward telemedicine, mHealth, and oncological apps. J Med Internet Res 2016 Nov 24;18(11):e312 [FREE Full text] [doi: 10.2196/jmir.6399] [Medline: 27884810]
42. Ventegodt S, Kandel I, Ervin D, Merrick J. Concepts of holistic care. In: Rubin IL, Merrick J, Greydanus DE, Patel DR, editors. Health Care for People with Intellectual and Developmental Disabilities across the Lifespan. Cham: Springer; 2016:1935-1941.
43. Ayash EJ, Alasiri MA, Alfarhan FAH, Alshammmari HHR, Alharfi HHK, Alazmi SI, et al. Improving patient outcomes through holistic approaches and collaboration between different specialties in hospital settings. J Popul Ther Clin Pharmacol 2022 Oct 10;29(4):3209-3218. [doi: 10.53555//jptcp.v29i04.5828]
44. Strandberg EL, Ovhed I, Borgquist L, Wilhelmsson S. The perceived meaning of a (w)holistic view among general practitioners and district nurses in Swedish primary care: a qualitative study. BMC Fam Pract 2007 Mar 08;8(1):8 [FREE Full text] [doi: 10.1186/1471-2296-8-8] [Medline: 17346340]
45. Kuhn E, Eftekhari A, Hoffman JE, Crowley JJ, Ramsey KM, Reger GM, et al. Clinician perceptions of using a smartphone app with prolonged exposure therapy. Adm Policy Ment Health 2014 Nov;41(6):800-807 [FREE Full text] [doi: 10.1007/s10488-013-0532-2] [Medline: 24398700]
46. Castor C, Lindkvist R, Hallström IK, Holmberg R. Health care professionals' experiences and views of eHealth in pediatric care: qualitative interview study applying a theoretical framework for implementation. JMIR Pediatr Parent 2023 Oct 18;6(1):e47663 [FREE Full text] [doi: 10.2196/47663] [Medline: 37851500]
47. Lieser T, Huang Y, Sezgin E. The current state of mobile apps owned by large pediatric hospitals in the United States: systematic search and analysis on Google Play and Apple App Stores. JMIR Pediatr Parent 2022 Oct 06;5(4):e38940 [FREE Full text] [doi: 10.2196/38940] [Medline: 36201385]
48. Nievas Soriano BJ, Uribe-Toril J, Ruiz-Real JL, Parrón-Carreño T. Pediatric apps: what are they for? A scoping review. Eur J Pediatr 2022 Apr 20;181(4):1321-1327 [FREE Full text] [doi: 10.1007/s00431-021-04351-1] [Medline: 35048178]
49. Ramsey WA, Heidelberg RE, Gilbert AM, Heneghan MB, Badawy SM, Alberts NM. eHealth and mHealth interventions in pediatric cancer: a systematic review of interventions across the cancer continuum. Psychooncology 2020 Jan;29(1):17-37 [FREE Full text] [doi: 10.1002/pon.5280] [Medline: 31692183]
50. Wen X, Nguyen L, Kowalski-Dobson T, Mohammed L, Woolford S. Employing smartphone health apps in weight management for adolescents: analysis on experiences and perspectives from pediatric providers. Clin Pediatr (Phila) 2022 Oct;61(10):692-698 [FREE Full text] [doi: 10.1177/00099228221097277] [Medline: 35581724]
51. Burmann A, Tischler M, Faßbach M, Schneitler S, Meister S. The role of physicians in digitalizing health care provision: web-based survey study. JMIR Med Inform 2021 Nov 11;9(11):e31527 [FREE Full text] [doi: 10.2196/31527] [Medline: 34545813]
52. Kong T, Scott MM, Li Y, Wichelman C. Physician attitudes towards-and adoption of-mobile health. Digit Health 2020;6:2055207620907187 [FREE Full text] [doi: 10.1177/2055207620907187] [Medline: 32128235]
53. Machleid F, Kaczmarczyk R, Johann D, Balčiūnas J, Atienza-Carbonell B, von Maltzahn F, et al. Perceptions of digital health education among European medical students: mixed methods survey. J Med Internet Res 2020 Aug 14;22(8):e19827 [FREE Full text] [doi: 10.2196/19827] [Medline: 32667899]
54. Tajirian T, Stergiopoulos V, Strudwick G, Sequeira L, Sanches M, Kemp J, et al. The influence of electronic health record use on physician burnout: cross-sectional survey. J Med Internet Res 2020 Jul 15;22(7):e19274 [FREE Full text] [doi: 10.2196/19274] [Medline: 32673234]
55. Robles N, Puig EP, Gómez-Calderón C, Saigí-Rubió F, Cambra GC, Zamora A, et al. Evaluation criteria for weight management apps: validation using a modified Delphi process. JMIR Mhealth Uhealth 2020 Jul 22;8(7):e16899 [FREE Full text] [doi: 10.2196/16899] [Medline: 32706689]
56. Potter A, Munsch C, Watson E, Hopkins E, Kitromili S, O'Neill IC, et al. Identifying research priorities in digital education for health care: umbrella review and modified Delphi method study. J Med Internet Res 2025 Feb 19;27:e66157 [FREE Full text] [doi: 10.2196/66157] [Medline: 39969988]

XSL•FO
RenderX

57. de la Torre E, Montane C, Rubio O, Sampietro-Colom L, Camacho-Mahamud A, Grau-Corral I. Integrating patient perspectives into the digital health technology readiness framework: Delphi study. J Med Internet Res 2025 Aug 20;27(1):e71600 [FREE Full text] [doi: 10.2196/71600] [Medline: 40835414]
58. Nwosu AC, McGlinchey T, Sanders J, Stanley S, Palfrey J, Lubbers P, et al. Identification of digital health priorities for palliative care research: modified Delphi study. JMIR Aging 2022 Mar 21;5(1):e32075 [FREE Full text] [doi: 10.2196/32075] [Medline: 35311674]
59. Erfannia L, Barman MP, Hussain S, Barati R, Arji G. How mobile health affects primary healthcare? Questionnaire design and attitude assessment. Digit Health 2020;6:2055207620942357 [FREE Full text] [doi: 10.1177/2055207620942357] [Medline: 32742715]
60. Jiang Y, West BT, Barton DL, Harris MR. Acceptance and use of eHealth/mHealth applications for self-management among cancer survivors. Stud Health Technol Inform 2017;245:131-135 [FREE Full text] [Medline: 29295067]
61. Kemp E, Trigg J, Beatty L, Christensen C, Dhillon HM, Maeder A, et al. Health literacy, digital health literacy and the implementation of digital health technologies in cancer care: the need for a strategic approach. Health Promot J Austr 2021 Feb;32 Suppl 1:104-114 [FREE Full text] [doi: 10.1002/hpja.387] [Medline: 32681656]
62. Abdelwhab Z, Elshwey R, Adel N, Eid R, Shash E. Evaluation of acceptance and perceived barriers to mobile health applications among breast cancer patients and caregivers in a low- to middle-income country. BMC Digit Health 2025 Sep 05;3(1):58 [FREE Full text] [doi: 10.1186/s44247-025-00203-2]
63. Richards R, Kinnersley P, Brain K, Staffurth J, Wood F. The preferences of patients with cancer regarding apps to help meet their illness-related information needs: qualitative interview study. JMIR Mhealth Uhealth 2019 Jul 31;7(7):e14187 [FREE Full text] [doi: 10.2196/14187] [Medline: 31368446]
64. Steeb T, Wessely A, Mastnik S, Brinker TJ, French LE, Niesert A, et al. Patient attitudes and their awareness towards skin cancer-related apps: cross-sectional survey. JMIR Mhealth Uhealth 2019 Jul 02;7(7):e13844 [FREE Full text] [doi: 10.2196/13844] [Medline: 31267978]
65. Vaffis S, Whaley S, Axon DR, Hall-Lipsy E, Hincapie A, Slack M, et al. Features of cancer mHealth apps and evidence for patient preferences: scoping literature review. JMIR Cancer 2023 Apr 28;9(1):e37330 [FREE Full text] [doi: 10.2196/37330] [Medline: 37115587]
66. Armbruster C, Knaub M, Farin-Glattacker E, von der Warth R. Predictors of adherence to cancer-related mHealth apps in cancer patients undergoing oncological or follow-up treatment—a scoping review. Int J Environ Res Public Health 2022 Oct 21;19(20):13689 [FREE Full text] [doi: 10.3390/ijerph192013689] [Medline: 36294265]
67. LeBaron V, Adhikari A, Bennett R, Chapagain Acharya S, Dhakal M, Elmore CE, et al. A survey of cancer care institutions in Nepal to inform design of a pain management mobile application. BMC Palliat Care 2021 Nov 05;20(1):171 [FREE Full text] [doi: 10.1186/s12904-021-00824-0] [Medline: 34740339]
68. Heaslip V, Richer S, Simkhada B, Dogan H, Green S. Use of technology to promote health and wellbeing of people who are homeless: a systematic review. Int J Environ Res Public Health 2021 Jun 25;18(13):6845 [FREE Full text] [doi: 10.3390/ijerph18136845] [Medline: 34202242]
69. Sidoti O, Dawson W, Gelles-Watnick R, Faverio M, Atske S, Radde K, et al. Mobile fact sheet. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2025-02-25]
70. Kershaw K, Martelly L, Stevens C, McInnes DK, Silverman A, Byrne T, et al. Text messaging to increase patient engagement in a large health care for the homeless clinic: results of a randomized pilot study. Digit Health 2022;8:20552076221129729 [FREE Full text] [doi: 10.1177/20552076221129729] [Medline: 36238754]
71. Polillo A, Gran-Ruaz S, Sylvestre J, Kerman N. The use of eHealth interventions among persons experiencing homelessness: a systematic review. Digit Health 2021 Feb 02;7:2055207620987066 [FREE Full text] [doi: 10.1177/2055207620987066] [Medline: 33598308]
72. Laricchia F. Smartphone ownership in the US by age 2023. Statista. 2024. URL: https://www.statista.com/statistics/489255/percentage-of-us-smartphone-owners-by-age-group/ [accessed 2025-02-25]
73. Jiang S, Hong YA. Mobile-based patient-provider communication in cancer survivors: the roles of health literacy and patient activation. Psychooncology 2018 Mar;27(3):886-891 [FREE Full text] [doi: 10.1002/pon.4598] [Medline: 29193503]
74. Gilstad H, Sand K, Solbjør M, Melby L. Deconstructing (e)health literacy: aspects that promote and inhibit understanding of health information in breast cancer patient pathways. Int J Qual Stud Health Well-being 2022 Dec;17(1):2137961 [FREE Full text] [doi: 10.1080/17482631.2022.2137961] [Medline: 36268568]
75. Kim H, Xie B. Health literacy in the eHealth era: a systematic review of the literature. Patient Educ Couns 2017 Jun;100(6):1073-1082 [FREE Full text] [doi: 10.1016/j.pec.2017.01.015] [Medline: 28174067]
76. Smith B, Magnani JW. New technologies, new disparities: the intersection of electronic health and digital health literacy. Int J Cardiol 2019 Oct 01;292:280-282 [FREE Full text] [doi: 10.1016/j.ijcard.2019.05.066] [Medline: 31171391]
77. Hendawi R, Alian S, Li J. A smart mobile app to simplify medical documents and improve health literacy: system design and feasibility validation. JMIR Form Res 2022 Apr 01;6(4):e35069 [FREE Full text] [doi: 10.2196/35069] [Medline: 35363142]

XSL•FO
RenderX

78. Mueller EL, Cochrane AR, Bennett WE, Carroll AE. A survey of mobile technology usage and desires by caregivers of children with cancer. Pediatr Blood Cancer 2018 Nov 17;65(11):e27359 [FREE Full text] [doi: 10.1002/pbc.27359] [Medline: 30015371]
79. Hasnan S, Aggarwal S, Mohammadi L, Koczwara B. Barriers and enablers of uptake and adherence to digital health interventions in older patients with cancer: a systematic review. J Geriatr Oncol 2022 Nov;13(8):1084-1091 [FREE Full text] [doi: 10.1016/j.jgo.2022.06.004] [Medline: 35752605]
80. Hwang M, Jiang Y. Personalization in digital health interventions for older adults with cancer: a scoping review. J Geriatr Oncol 2023 Nov;14(8):101652 [FREE Full text] [doi: 10.1016/j.jgo.2023.101652] [Medline: 37866009]
81. Nahm E, McQuaige M, Steacy K, Zhu S, Seong H. The impact of a digital cancer survivorship patient engagement toolkit on older cancer survivors' health outcomes. Comput Inform Nurs 2025 Jan 01;43(1):e01199 [FREE Full text] [doi: 10.1097/CIN.0000000000001199] [Medline: 39365650]
82. Newton L, Monkman H, Fullerton C. Exploring older adult cancer survivors' digital information needs: qualitative pilot study. JMIR Cancer 2025 Feb 27;11:e59391 [FREE Full text] [doi: 10.2196/59391] [Medline: 40014827]
83. Paillaud E, Galvin A, Doublet S, Poisson J, Gay P, Perrin C, et al. Health literacy and the use of digital tools in older patients with cancer and their younger counterparts: a multicenter, nationwide study. Patient Educ Couns 2025 Jan;130:108420 [FREE Full text] [doi: 10.1016/j.pec.2024.108420] [Medline: 39303502]
84. Rivière C, Brureau L, Parnot C, Becherirat S, Duverger C, Picchi H, et al. Effectiveness of a digital telemonitoring platform for cancer care of older patients: the ConnectElderlyPatientToDoctor study. Int J Cancer 2023 Feb 01;152(3):504-510 [FREE Full text] [doi: 10.1002/ijc.34196] [Medline: 35770377]
85. Sanders AB, Conroy DE, Schmitz KH, Gusani N. Physical activity and sedentary behavior in older gastrointestinal cancer survivors: need and acceptability of digital health interventions. J Gastrointest Cancer 2018 Jun 18;50:703-708 [FREE Full text] [doi: 10.1007/s12029-018-0128-x] [Medline: 29911290]
86. Verma R, Saldanha C, Ellis U, Sattar S, Haase KR. eHealth literacy among older adults living with cancer and their caregivers: a scoping review. J Geriatr Oncol 2022 Jun;13(5):555-562 [FREE Full text] [doi: 10.1016/j.jgo.2021.11.008] [Medline: 34810146]
87. Wankah P, Chandra S, Lofters A, Mohamednur N, Osei B, Makuwaza T, et al. Improving digital cancer care for older black adults: qualitative study. J Med Internet Res 2025 Feb 19;27:e63324 [FREE Full text] [doi: 10.2196/63324] [Medline: 39970422]
88. Zhou W, Cho Y, Shang S, Jiang Y. Use of digital health technology among older adults with cancer in the United States: findings from a national longitudinal cohort study (2015-2021). J Med Internet Res 2023 May 31;25:e46721 [FREE Full text] [doi: 10.2196/46721] [Medline: 37256672]
89. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989 Sep;13(3):319-340 [FREE Full text] [doi: 10.2307/249008]
90. Nadal C, Sas C, Doherty G. Technology acceptance in mobile health: scoping review of definitions, models, and measurement. J Med Internet Res 2020 Jul 06;22(7):e17256 [FREE Full text] [doi: 10.2196/17256] [Medline: 32628122]
91. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327 [FREE Full text] [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
92. Glasgow RE, Harden SM, Gaglio B, Rabin B, Smith ML, Porter GC, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. Front Public Health 2019 Mar;7:64 [FREE Full text] [doi: 10.3389/fpubh.2019.00064] [Medline: 30984733]
93. Kwan BM, McGinnes HL, Ory MG, Estabrooks PA, Waxmonsky JA, Glasgow RE. RE-AIM in the real world: use of the RE-AIM framework for program planning and evaluation in clinical and community settings. Front Public Health 2019 Nov 22;7:345 [FREE Full text] [doi: 10.3389/fpubh.2019.00345] [Medline: 31824911]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
FDA: Food and Drug Administration
IRB: Institutional Review Board
listserv: list server
mHealth: mobile health
OPENS: Oncology Provider Experiences National Survey
REDCap: Research Electronic Data Capture
RQ: research question
TAM: Technology Acceptance Model

XSL•FO
RenderX

Edited by A Stone, N Cahill; submitted 09.Oct.2025; peer-reviewed by S Ajayi, M Lemonde; comments to author 18.Nov.2025; revised version received 13.Feb.2026; accepted 18.Feb.2026; published 23.Mar.2026.

Please cite as:
Lau N, Srinakarin K, Hong SJH, Aalfs H, Roth ME, Ingram KM, Berkman A, Chan S, Patten J, Iwata W, Taylor MR, Fann JR, Chow EJ, Palermo TM
Mobile Health App Attitudes and Adoption Among Oncology Providers: Cross-Sectional National Survey
J Med Internet Res 2026;28:e85583
URL: https://www.jmir.org/2026/1/e85583
doi:10.2196/85583
PMID:

©Nancy Lau, Kavin Srinakarin, Shannon JH Hong, Homer Aalfs, Michael E Roth, Karly M Ingram, Amy Berkman, Sherilynn Chan, Joanna Patten, Wade Iwata, Mallory R Taylor, Jesse R Fann, Eric J Chow, Tonya M Palermo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patient Sharing of Digital Health Data in the Veterans Health Administration: Cross-Sectional Analysis

Mark S Zocchi[1,2,3], PhD; Jessica M Lipschitz[1,4,5], PhD; Felicia R Bixler[1,6], MS; Stephanie A Robinson[1,2,7], PhD; Bella Etingen[1,8,9], PhD; Timothy P Hogan[1,2,9], PhD; Ndindam Ndiwane[1,2], MPH; Saige Calkins[1,2], MA; Terry Newton[10], MD; Nilesh Shah[10,11], MD; Ben Kragen[1,2], MBA, PhD; Alexander S Young[1,12,13], MD; Stephanie L Shimada[1,2,3,14], PhD

[1]eHealth Partnered Evaluation Initiative, VA Bedford Healthcare System, Bedford, MA, United States

[2]Center for Health Optimization and Implementation Research, VA Bedford Healthcare System, Bedford, MA, United States

[3]Division of Health Informatics and Implementation Science, Department of Population and Quantitative Health Sciences, UMass Chan Medical School, Worcester, MA, United States

[4]Department of Psychiatry, Brigham and Women's Hospital, Boston, MA, United States

[5]Department of Psychiatry, Harvard Medical School, Boston, MA, United States

[6]Center of Innovation for Complex Chronic Healthcare, Edward Hines Jr. VA Hospital, Hines, IL, United States

[7]The Pulmonary Center, Boston University School of Medicine, Boston, MA, United States

[8]Research and Development Service, Dallas VA Medical Center, Dallas, TX, United States

[9]Department of Health Economics, Systems, and Policy, Peter O'Donnell Jr School of Public Health, The University of Texas Southwestern Medical Center, Dallas, TX, United States

[10]Office of Connected Care, Veterans Health Administration, Washington, DC, United States

[11]School of Medicine, University of California San Diego, La Jolla, CA, United States

[12]Mental Illness Research, Education, and Clinical Center, Veterans Affairs Greater Los Angeles Healthcare System, Los Angeles, CA, United States

[13]Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, United States

[14]Department of Health Law, Policy, and Management, Boston University School of Public Health, Boston, MA, United States

Corresponding Author:
Mark S Zocchi, PhD
Center for Health Optimization and Implementation Research
VA Bedford Healthcare System
200 Springs Road, Bldg. 70
Bedford, MA, 01730
United States
Phone: 1 781 687 4737
Email: Mark.Zocchi@va.gov

Abstract

Background: The integration of patient-generated health data (PGHD) into health care has the potential to significantly transform patient care and clinical practice. PGHD includes health-related data created by patients, enabling the collection of health data beyond traditional health care settings. The Veterans Health Administration (VA) has taken proactive steps to incorporate PGHD into health care through its Share My Health Data (SMHD) mobile app. Launched in 2023, the SMHD app allows veterans to securely share data from their personal digital health devices with the VA for clinical and research use. However, data characterizing patients who use such tools in real-world health care systems are lacking, creating an evidence gap for implementing PGHD-informed care equitably.

Objective: This study aimed to identify the characteristics of patients using the VA SMHD mobile app, which allows veterans to share PGHD with the VA.

Methods: We conducted a cross-sectional analysis of veterans who began using SMHD between October 2023 and September 2024 (n=3157, "SMHD users"). We collected demographic information, including age, sex, race/ethnicity, and rurality, and clinical information, including physiological and mental health conditions, from VA administrative data. We compared characteristics of SMHD users to a 10% random sample of veterans from the same underlying administrative data cohort that

XSL•FO
RenderX

had never used the app (n=632,187, “nonusers”). Statistical analyses were performed using chi-square tests, independent t tests, and multivariable regression to assess the relationship between use and key characteristics.

Results: Middle-aged veterans were more likely to be SMHD users (40-49 years: odds ratio [OR] 1.55, *P*<.001; 50-59 years: OR 1.37, *P*<.001), while veterans aged 60 years and over were less likely (60-69 years: OR 0.72, *P*<.001; ≥70 years: OR 0.24, *P*<.001). Female (OR 1.23, *P*<.001) and married (OR 1.31, *P*<.001) veterans were more likely to be SMHD users than male and unmarried veterans. In contrast, Black or African American (OR 0.62, *P*<.001) and rural (OR 0.82, *P*<.001) veterans were less likely to be SMHD users than White and urban veterans. Veterans in higher-income zip codes (OR 1.36, *P*<.001) were more likely to have used the app than those in lower-income zip codes. Clinically, SMHD users were more likely to have a service-connected disability (OR 1.81, *P*<.001), multiple physiological conditions (OR 1.86, *P*<.001), and multiple mental health diagnoses (OR 1.35, *P*<.001) versus none.

Conclusions: Veterans who used the SMHD app differed significantly from nonusers across several demographic and clinical characteristics. These insights identify specific demographic and clinical subgroups with higher and lower app adoption, providing an evidence base to inform targeted implementation and outreach and support strategies to promote enhanced engagement in PGHD-informed care.

(J Med Internet Res 2026;28:e80517) doi:10.2196/80517

KEYWORDS

patient-generated health data; veterans; mobile health; mHealth; digital phenotyping; wearable electronic devices

Introduction

The health care landscape is rapidly evolving with the integration of digital patient-generated health data (PGHD) into medical practice [1]. The Office of the National Coordinator for Health Information Technology defines PGHD as “health-related data created, recorded, or gathered by or from patients (or family members or other caregivers) to help address a health concern” [2]. PGHD can include health-related data, such as physical activity, vital signs, sleep, and other physiological metrics collected through wearable devices (eg, smartwatches), mobile health apps (eg, Apple Health), and other digital devices (eg, Bluetooth-enabled blood pressure cuffs).

The use of digital PGHD is part of a broader movement, sometimes called digital phenotyping [3] or personal sensing [4], to understand how to use data from personal digital devices to improve health care outcomes and personalization of treatment. Capturing health data in patients’ everyday environments has the potential to positively impact care delivery and health outcomes. These data provide frequent, objective insights into patients’ health status, offering a more complete view of health and well-being than periodic clinical encounters. This may allow for more personalized, proactive, and responsive care, especially for management of chronic conditions and preventive care [5]. In addition, PGHD has the potential to enhance patient engagement in self-management. Researchers have found many use cases for PGHD, such as predicting onset of new symptoms [6], monitoring chronic disease [7], detection of physiological and psychological conditions [8], and predicting response to treatment [9].

Despite the many use cases and opportunities offered by PGHD, health care system implementation of PGHD-informed care has been limited. Survey research suggests that many patient populations (eg, psychiatric, arthritis) have interest in digital symptom monitoring [10-12]. However, we do not yet know which patients may be most likely to actually use digital tools that support the collection of PGHD. Understanding the characteristics of these users could help app developers and clinical leaders identify what clinical populations to target for initial PGHD-informed care initiatives and also offer insight into ways to increase uptake in patient populations that show lower propensity to use these kinds of tools. Therefore, characterizing patients who adopt PGHD-sharing tools within health care systems is a critical piece of implementation science necessary for guiding rollout and tailoring support strategies.

To move toward implementing PGHD-informed care, the Veterans Health Administration (VA) launched the Share My Health Data (SMHD) mobile app in 2023 [13]. The SMHD app allows veterans to voluntarily share data from personal digital health devices, such as Bluetooth-enabled glucometers and third-party health apps, such as Apple Health, Fitbit, and Garmin [14]. Once paired with the SMHD app, veterans’ PGHD are securely stored in a dedicated VA PGHD database [15]. The contents of this database can, in turn, be applied in support of clinical care and related research. For example, the VA is currently pilot-testing the integration of these data into clinical dashboards, allowing providers to view PGHD directly from their patients during or between appointments.

VA’s SMHD app offers a unique opportunity to examine the passive digital health–monitoring initiative within a large, integrated health care system. This study compared the demographic characteristics, clinical conditions, and health care costs of SMHD users to a nationally representative random sample of veterans who have not used the app. By describing the demographic and clinical profile of SMHD users and comparing them to the broader VA population, we aimed to generate foundational evidence to guide targeted engagement strategies, identify clinical populations for pilot programs, and highlight potential disparities in digital health tool adoption. These insights can inform efforts to broaden patient participation in PGHD-informed care and support decisions about which clinical populations may be most appropriate for early implementation within the VA and in other health care systems.

XSL•FO
RenderX

Methods

Design

This was a retrospective, observational, cross-sectional analysis using routinely collected data from national VA data sources. This evaluation focused on SMHD users and compared characteristics of those who used the app to those who did not. The main objective was descriptive, with no intervention assignment or randomization.

Setting

VA's Office of Connected Care launched the SMHD app on May 1, 2023. The app was added to the VA App Store, featured on its home page, and made available to download on iOS and Android devices [16]. After downloading and accepting the end-user agreement, users can pair their digital health devices through VA's secure authentication process. Other than instructional materials provided on the VA Mobile App Store, no additional training, outreach, or incentives were provided.

Cohort Selection

We examined veterans who began using the app shortly after it was launched, between October 1, 2023, and September 30, 2024 ("SMHD users") and compared them to a random sample of veterans who did not use the app prior to or during this same time period ("nonusers").

Our analytic cohort included SMHD users, who we identified as having at least one measurement in the PGHD database submitted through the SMHD app (N=4202). This definition of "use" indicates SMHD app download, linkage to a digital health device, and initiation of data sharing; it does not measure intensity or continuity of use over time. Exclusions were nonveterans (n=62, 1.5%), those not included in the fiscal year 2023 VA Nosos risk score dataset (n=318, 7.6%) [17], or participants in early field tests of the SMHD app (n=665, 15.8%). Early field test participants were excluded to focus this evaluation on patients who discovered, downloaded, and began using the app on their own, independent of recruitment into a controlled field test.

The comparison group of nonusers was selected using a 10% random sample of veterans who were included in the fiscal year 2023 VA Nosos risk score dataset and who had never used the SMHD app prior to September 30, 2024, the end of the evaluation period. Of note, all veterans who received care in VA during fiscal year 2023 (October 1, 2022-September 30, 2023) are included in the fiscal year 2023 VA Nosos risk score dataset. This sampling approach ensured all veterans in the sample received VA or VA-paid care and allowed for sufficient statistical power to detect differences between SMHD users and nonusers, while maintaining feasibility for data processing and analysis. The final cohort consisted of 3157 SMHD users and 632,187 nonusers.

Data Sources

We obtained SMHD app usage data from VA's PGHD database. PGHD were extracted via secure SQL queries and validated for appropriate SMHD source linkage and timestamp accuracy. PGHD with missing timestamps or insufficient source metadata were removed prior to cohort selection. Patient demographics and clinical encounter data were obtained from the VA's Corporate Data Warehouse (CDW). Zip code–level household income data came from the US Census Bureau American Community Survey (2023 5-year estimates) [18].

Preuse "baseline" measures of comorbidity and cost came from the fiscal year 2023 VA Nosos risk score dataset, which includes indicators for chronic physiological and mental health conditions based on the Centers for Medicare and Medicaid Services' Hierarchical Condition Categories (HCC) [19] and a psychiatric case-mix system (PsyCMS) developed by Sloan et al [20] and later updated for *International Classification of Diseases* (ICD)-10 diagnosis codes [21]. The Nosos dataset uses diagnosis data (ICD-9 or ICD-10) from VA outpatient encounters and inpatient stays during the entire fiscal year, including VA-purchased community care data, to calculate comorbidities. Comorbidity data are combined with use data to calculate an overall "score," which is standardized to center around 1, so a value of 1 means that the veteran is expected to have costs that are equivalent to the national average for VA patients.

Variables

Variables were selected based on availability in the electronic medical record and prior work examining adoption and use of other digital health technologies in VA [22-24]. Demographic variables included age, race/ethnicity, sex, marital status, rurality, zip code–level household income, and service-connected disability. A service-connected disability refers to a physical or mental health condition that was incurred or aggravated during active military service and is formally recognized by VA. Clinical variables included counts of physiological and mental health conditions (eg, 0, 1, ≥2). To avoid overlap with the mental health counts, we excluded HCC codes from the overall physiological condition counts that were for mental health conditions (eg, alcohol use disorder, schizophrenia, major depressive disorder). For the mental health counts, we excluded PsyCMS codes indicating conditions in remission (eg, alcohol use remission). Relative cost outcomes were estimated using the fiscal year 2023 Nosos concurrent risk score [19].

Data Analysis

We compared SMHD users with the random sample of nonusers across all variables of interest using means (SDs), frequencies, and percentages. Frequencies of categorical variables were compared using chi-square tests. Means of continuous variables were compared using independent *t* tests. Standardized mean differences (SMDs) were calculated for all variables to aid interpretation of the magnitude of differences between groups, with an absolute value of >0.1 conventionally considered a small but meaningful difference.

A multivariable logistic regression was used to estimate odds ratios (ORs) associated with being an SMHD user, with SEs adjusted for potential clustering of veterans within VA regions. Because our objective was descriptive and did not have an a priori hypothesis about effect moderation, we did not include any interaction terms. To avoid multicollinearity, we checked covariates' variance inflation factor and included a term for

XSL•FO
RenderX

total chronic condition counts (eg, 0, 1, ≥2) but excluded individual comorbidities in the model. As a conventional check for model fit, we reported the Hosmer-Lemeshow goodness-of-fit test (10 groups). A complete case analysis was used for the multivariable model. The proportion of missing data was small (5%), and sensitivity analyses using alternative missing data methods were not conducted.

All analyses were conducted using R version 4.4.1 (R Foundation for Statistical Computing). Key packages included *stats* for regression modeling and *dplyr* for data manipulation.

Ethical Considerations

This work was reviewed by the Institutional Review Boards of the VA Bedford Healthcare System in Bedford, Massachusetts, and the Edward Hines Jr VA Hospital in Hines, Illinois, and designated as program evaluation for quality improvement purposes, exempting it from further oversight (VA Handbook 1058.05).

Results

Characteristics of SMHD Users

SMHD users differed from nonusers in several ways (Table 1). SMHD users were younger than nonusers by 9 years (users: mean 53.2, SD 13.4; nonusers: mean 62.3, SD 17.2, years), and a greater proportion were female (users: n=539, 17.1%; nonusers: n=67,401, 10.7%). In addition, race/ethnicity differed between SMHD users and nonusers. A larger proportion of SMHD users were Hispanic or Latino (users: n=322, 10.2%; nonusers: n=47,265, 7.5%), while a smaller proportion were non-Hispanic Black or African American (users: n=477, 15.1%; nonusers: n=109,459, 17.3%). Furthermore, a smaller proportion of SMHD users lived in rural areas (users: n=914, 29%; nonusers: n=214,608, 33.9%), while a larger proportion were married (users: n=1878, 59.5%; nonusers: n=349,908, 55.3%), had a service-connected disability (users: n=2791, 88.4%; nonusers: n=460,129, 72.8%), and, on average, lived in higher-income zip codes (mean household income for users: US $79,731; nonusers: US $75,071).

Table 2 presents a comparison of health conditions between SMHD users and nonusers. Clinically, SMHD users had a significantly higher burden of mental health conditions, such as posttraumatic stress disorder, depression, anxiety, adjustment disorder, sleep disorder, attention deficit disorder, and mood disorder. In contrast, they had a lower incidence of several major physiological conditions, such as chronic obstructive pulmonary disorder, heart arrhythmias, and congestive heart failure (all $P<.001$), though they showed higher rates of obesity and arthritis. Rates of alcohol dependence or abuse did not differ significantly between SMHD users and nonusers; however, rates of nicotine dependence were slightly lower among SMHD users (users: n=358, 11.3%; nonusers: n=86,180, 13.6%; $P<.001$). In terms of total VA health care costs, SMHD users had higher mean Nosos risk scores than nonusers (users: 1.13, SD 1.34; nonusers: 1.05, SD 1.64; $P=.005$), indicating VA health care costs were about 13% higher than the VA population average for SMHD users.

XSL•FO
RenderX

Table 1. Unadjusted differences in demographic characteristics among SMHD[a] users and nonusers.

Characteristics	SMHD users (n=3157)	Nonusers (n=632,187)	SMD[b]	*P* value
Age (years), mean (SD)	53.2 (13.4)	62.3 (17.2)	0.589	<.001
Age group (years), n (%)			0.716	<.001
<40	538 (17.0)	86,458 (13.7)	__[c]	—
40-49	765 (24.2)	72,881 (11.5)	—	—
50-59	852 (27.0)	90,594 (14.3)	—	—
60-69	561 (17.8)	117,407 (18.6)	—	—
>69	441 (14.0)	264,842 (41.9)	—	—
Female sex, n (%)	539 (17.1)	67,401 (10.7)	0.186	<.001
Race/ethnicity, n (%)			0.116	<.001
White, not Hispanic	2016 (63.9)	402,055 (63.6)	—	—
Black or African American, not Hispanic	477 (15.1)	109,459 (17.3)	—	—
Hispanic	322 (10.2)	47,265 (7.5)	—	—
Other race, not Hispanic	107 (3.4)	19,222 (3.0)	—	—
Unknown	235 (7.4)	54,186 (8.6)	—	—
Rural residence n (%)	914 (29.0)	214,608 (33.9)	0.122	<.001
Zip code median income (US $), mean (SD)	79,731 (27,953)	75,071 (26,646)	0.171	<.001
Zip code median income (US $) category, n (%)			0.169	<.001
<50,000	328 (10.4)	83,355 (13.2)	—	—
50,000-69,999	1011 (32.0)	221,976 (35.1)	—	—
70,000-89,999	828 (26.6)	157,757 (25.0)	—	—
≥90,000	948 (30.0)	146,026 (23.1)	—	—
Married, n (%)	1878 (59.5)	349,908 (55.3)	0.083	<.001
Service-connected disability, n (%)	2791 (88.4)	460,129 (72.8)	0.403	<.001

[a]SMHD: Share My Health Data.

[b]SMD: standardized mean difference.

[c]Not applicable.

XSL•FO
RenderX

Table 2. Unadjusted differences in health conditions among SMHD[a] users and nonusers.

Health conditions	SMHD users (n=3157), n (%)	Nonusers (n=632,187), n (%)	SMD[b]	*P* value
Nosos Risk Score, Mean (SD)	1.13 (1.34)	1.05 (1.64)	0.055	.005
Physiological conditions (15 most common)				
Diabetes with chronic complications	450 (14.3)	88,718 (14.0)	0.006	.74
Diabetes without complication	360 (11.4)	65,279 (10.3)	0.035	.05
Chronic obstructive pulmonary disease	185 (5.9)	63,720 (10.1)	0.156	<.001
Specified heart arrhythmias	235 (7.4)	62,671 (9.9)	0.088	<.001
Congestive heart failure	154 (4.9)	45,954 (7.3)	0.100	<.001
Vascular disease	158 (5.0)	44,802 (7.1)	0.087	<.001
Breast, prostate, and other cancers and tumors	96 (3.0)	25,993 (4.1)	0.058	.003
Chronic kidney disease, moderate (stage 3)	94 (3.0)	24,418 (3.9)	0.049	.01
Morbid obesity	308 (9.8)	24,068 (3.8)	0.238	<.001
Acute renal failure	68 (2.2)	20,267 (3.2)	0.065	.001
Rheumatoid arthritis and inflammatory connective tissue disease	149 (4.7)	20,011 (3.2)	0.080	<.001
Cardiorespiratory failure and shock	43 (1.4)	17,325 (2.7)	0.097	<.001
Coagulation defects and other specified hematological disorders	80 (2.5)	17,161 (2.7)	0.011	.57
Other significant endocrine and metabolic disorders	80 (2.5)	15,280 (2.4)	0.008	.71
Ischemic or unspecified stroke	56 (1.8)	14,715 (2.3)	0.039	.05
Mental health conditions (15 most common)				
Posttraumatic stress disorder	775 (24.5)	103,665 (16.4)	0.203	<.001
Nicotine dependence	358 (11.3)	86,180 (13.6)	0.069	<.001
Major depressive disorder, recurrent	474 (15.0)	56,132 (8.9)	0.19	<.001
Anxiety, not elsewhere classified	316 (10.0)	41,227 (6.5)	0.127	<.001
Adjustment disorder	339 (10.7)	39,001 (6.2)	0.165	<.001
Generalized anxiety	306 (9.7)	32,517 (5.1)	0.174	<.001
Other mental health code	124 (3.9)	22,660 (3.6)	0.018	.32
Major depressive disorder, single episode	160 (5.1)	20,080 (3.2)	0.095	<.001
Alcohol dependence	90 (2.9)	19,126 (3.0)	0.01	.60
Alcohol abuse	69 (2.2)	16,206 (2.6)	0.025	.20
Dementia	8 (0.3)	14,143 (2.2)	0.18	<.001
Sleep disorders	138 (4.4)	13,647 (2.2)	0.125	<.001
Attention deficit disorder	216 (6.8)	13,488 (2.1)	0.229	<.001
Mood disorder, not elsewhere classified	103 (3.3)	12,369 (2.0)	0.082	<.001
Bipolar disorders	104 (3.3)	10,730 (1.7)	0.103	<.001
Number of physiological conditions			0.065	.001
0	1500 (47.5)	306,388 (48.5)	__[c]	—
1	831 (26.3)	149,436 (23.6)	—	—
≥2	826 (26.2)	176,363 (27.9)	—	—
Number of mental health conditions			0.324	<.001
0	1171 (37.1)	323,307 (51.1)	—	—
1	695 (22.0)	138,703 (21.9)	—	—

XSL•FO
RenderX

Health conditions	SMHD users (n=3157), n (%)	Nonusers (n=632,187), n (%)	SMD[b]	*P* value
≥2	1291 (40.9)	170,177 (26.9)	—	—

[a]SMHD: Share My Health Data.

[b]SMD: standardized mean difference.

[c]Not applicable.

Predictors of SMHD App Use

The results from the multivariable logistic regression model are presented in Table 3. Of the 635,344 patients in the cohort, 95% (n=600,590) had complete data for the multivariable model. The Hosmer-Lemeshow test indicated no evidence of lack of fit (*P*=.19), and the variance inflation factor values ranged from 1.1 to 1.5, which is significantly below the commonly used threshold of 5 to identify multicollinearity [25]. Several characteristics were significantly associated with reduced odds of being an SMHD user. Compared to younger veterans (<40 years), older veterans (>60 years) were less likely to be SMHD users (60-69 years: OR 0.72, 95% CI 0.63-0.83; ≥70 years: OR 0.24, 95% CI 0.21-0.28). Compared to White non-Hispanic veterans, Black non-Hispanic (OR 0.62, 95% CI 0.56-0.69) and other race/ethnicity (OR 0.79, 95% CI 0.65-0.97) veterans were less likely to be SMHD users. Veterans living in rural areas (OR 0.82, 95% CI 0.75-0.89) were less likely to be SMHD users compared to veterans living in more urban areas.

Conversely, several characteristics were found to be associated with increased odds of being an SMHD user. Compared to younger veterans (<40 years), middle-aged veterans were more likely to be SMHD users (40-49 years: OR 1.55, 95% CI 1.38-1.73; 50-59 years: OR 1.37, 95% CI 1.22-1.54). Female veterans (OR 1.23, 95% CI 1.11-1.35) were more likely to be SMHD users compared to male veterans, and married veterans (OR 1.31, 95% CI 1.22-1.41) were more likely to use SMHD than unmarried veterans (ie, divorced, separated, widowed, or single). In addition, compared to veterans living in zip codes with mean incomes less than US $50,000, veterans living in higher-income zip codes were more likely to be SMHD users (US $70,000-$89,999: OR 1.15, 95% CI 1.01-1.31; ≥US $90,000: OR 1.36, 95% CI 1.19-1.55). Finally, having a service-connected disability (OR 1.81, 95% CI 1.61-2.04), having one (OR 1.54, 95% CI 1.41-1.68) or two or more (OR 1.86, 95% CI 1.68-2.05) physiological conditions (vs none), and having two or more mental health conditions (OR 1.35, 95% CI 1.24-1.47) versus none were all associated with increased odds of being an SMHD user.

XSL•FO
RenderX

Table 3. Multivariable logistic regression model predicting odds of being an SMHD[a] user (N=600,590).[b]

Characteristics	OR[c] (95% CI)	*P* value
Age group (years; reference: <40 years)		
40-49	1.55 (1.38-1.73)	<.001
50-59	1.37 (1.22-1.54)	<.001
60-69	0.72 (0.63-0.83)	<.001
>69	0.24 (0.21-0.28)	<.001
Female (reference: male)	1.23 (1.11-1.35)	<.001
Race/ethnicity (reference: White)		
Black or African American	0.62 (0.56-0.69)	<.001
Hispanic	0.94 (0.83-1.07)	.35
Other	0.79 (0.65-0.97)	.02
Unknown	0.82 (0.71-0.95)	.01
Rural (reference: urban)	0.82 (0.75-0.89)	<.001
Zip code median income category (reference: <US $50,000)		
50,000-69,999	1.06 (0.93-1.20)	.36
70,000-89,999	1.15 (1.01-1.31)	.04
≥90,000	1.36 (1.19-1.55)	<.001
Married (reference: not married)	1.31 (1.22-1.41)	<.001
Service-connected disability	1.81 (1.61-2.04)	<.001
Number of physiological conditions (reference: 0)		
1	1.54 (1.41-1.68)	<.001
≥2	1.86 (1.68-2.05)	<.001
Number of mental health conditions (reference: 0)		
1	1.09 (0.99-1.20)	.09
≥2	1.35 (1.24-1.47)	<.001

[a]SMHD: Share My Health Data.

[b]Note that 34,754 (5%) patients were excluded from the model due to missing data. SEs were adjusted for clustering at the Veterans Health Administration (VA) regional level. Hosmer-Lemeshow goodness-of-fit test: χ^2_8=11.2, *P*=.19.

[c]OR: odds ratio.

Discussion

Principal Findings

PGHD-guided care has the potential to be the future of medicine. Effective future implementation requires understanding which patients are open to using digital devices and other tools to collect and share PGHD. To the best of our knowledge, this is the first evaluation to look at users of a mobile app intended to support the collection and sharing of PGHD across a health care system. We found that SMHD users are more likely to be female, be married, and live in a higher-income zip code and are less likely to be aged 60 years and over, be Black or from another race, or live in a rural area. We also found that SMHD users have higher health care costs and more medical and mental health comorbidities.

Our findings about the demographic characteristics of SMHD users are similar to the literature on users of other digital health technologies. Studies on the use of other digital health technologies have also found that women and higher-income and urban-dwelling patients are more likely to use digital health technologies [26-30]. Studies have also found that patients who are from minority races (in this case non-Hispanic Black and other races), older, and divorced or separated are less likely to use digital health technologies [31-33]. These patterns may reflect broader determinants, such as differential access to technology and broadband [29,30], variations in digital health literacy, and potential questions of trust in data-sharing initiatives [31]. As PGHD tools and PGHD-informed care initiatives are designed and implemented over the coming years, it will be important to consider how to make these tools accessible to all patients to promote health and reduce health care disparities.

In addition, these findings could inform future decisions regarding what clinical populations could be appropriate targets for initial pilots of PGHD-guided care models. In this study,

XSL•FO
RenderX

SMHD users were more likely to have, for example, arthritis, morbid obesity, multiple comorbidities, and mental health diagnoses. These findings corroborate those from survey studies indicating high interest in monitoring apps and sensing technologies among these populations [10-12], but until now, data on actual uptake have been limited. Thus, our findings can help inform health care systems about what patient populations to consider first, as they start to develop and test PGHD-informed care models.

Our findings indicate that SMHD users have higher health care costs and greater rates of both mental and physiological conditions compared to the general VA population. The association between app use and higher clinical complexity suggests that SMHD adopters represent a subgroup that could stand to benefit most from more precise, personalized, and proactive PGHD-informed care models. This presents a compelling opportunity to evaluate whether integrating PGHD into routine care can improve outcomes and optimize resource use within this more complex patient population.

As health care systems increasingly adopt PGHD tools, it will be important to ensure that implementation strategies account for differences in patient characteristics and technology use. Our findings highlight demographic and clinical differences between SMHD users and nonusers, which can inform targeted outreach and support efforts. Future work should explore barriers that may limit use among certain patient groups, and identify approaches to encourage broader engagement with PGHD tools across the VA population. For example, future implementation strategies for SMHD and similar tools could integrate targeted support, such as peer-led digital coaching for older adults, partnerships with rural community centers to facilitate access, and explicit messaging addressing data privacy concerns.

Limitations

This analysis was limited to characteristics found in the VA electronic health record. For the multivariable models, we only used complete case data, which may introduce bias if data are not completely missing at random; however, the amount of missing data was relatively small (5%). Our focus was on SMHD users for sharing PGHD, and we did not evaluate the frequency, quality, or validity of the shared data.

In addition, this paper describes the use of the SMHD app before clinical displays or broader PGHD-informed care models have been systematically implemented in VA clinical settings. The connection to care received, and, consequently, the value of sharing PGHD, is likely different at this stage of implementation than it would be in the context of provider recommendation and systematic care team engagement with the PGHD shared by patients. In addition, as implementation efforts move forward, organizational factors may play a role in adoption and use and will be important to consider in future work.

It is also important to keep in mind that this was an exploratory cross-sectional analysis, and as such, it does not make any causal claims about SMHD use and subsequent health outcomes. By design, the comorbidity and cost measures represent a preuse baseline status and therefore do not capture any changes that may have occurred after the study period began. Although we used a large nationwide sample and included several relevant covariates in our model, we cannot eliminate the possibility of unobserved confounders. For example, smartphone ownership, internet access, and perceptions of data privacy and security are not available in VA administrative data. Feasibility and acceptability were inferred from use of the SMHD app; however, no direct user feedback was collected. Finally, although this evaluation draws on a large, nationwide sample of veterans, the findings may have limited generalizability to non-VA health care settings or populations.

Conclusion

This national evaluation of the VA's SMHD app reveals that app adopters are more likely to be middle-aged, female, married, urban-dwelling veterans with service-connected disabilities and multiple health conditions, while being less likely to be veterans who are older, Black, or from rural or lower-income areas. These findings provide the first large-scale, real-world evidence of who actively shares PGHD within an integrated health care system, moving beyond prior studies that relied primarily on patient self-report.

Understanding these adoption patterns is critical for implementation initiatives. They identify natural adopter populations for pilot programs and highlight demographic groups at risk for digital exclusion, informing the design of equitable engagement strategies. As PGHD integration evolves, ongoing monitoring of adoption trends and focused research on barriers will be essential to ensure these transformative tools benefit all patients.

Finally, organizational factors, such as local clinic promotion strategies, provider endorsement, and the degree of PGHD integration into clinical workflows, may significantly influence future adoption. Addressing these factors alongside patient-level disparities will be key to achieving scalable PGHD-informed care across VA and other health care systems.

Acknowledgments

The views expressed in this paper are those of the authors and do not necessarily reflect the position or policy of the US Department of Veterans Affairs or of the US government. This work was supported by the US Department of Veterans Affairs, Office of Connected Care and Office of Research and Development, Health Services Research and Development Service, Quality Enhancement Research Initiative Program (PEC 15–470; principal investigator: TPH). The authors of this paper are affiliated with both the Office of Connected Care and the Quality Enhancement Research Initiative Program.

XSL•FO
RenderX

Data Availability

The US Department of Veterans Affairs prohibits unauthorized sharing of data. The data used for this evaluation are not permitted to be shared outside the Veterans Health Administration (VA) firewall without a Data Use Agreement. This limitation is consistent with other studies based on VA data; however, VA data are available to researchers behind the VA firewall with an approved VA study protocol. For more information, please visit VA Information Resource Center (VIReC) [34].

Authors' Contributions

All authors have made substantial contributions to the conception or design of the work. MSZ and JML wrote the initial draft, and FRB, SAR, BE, TPH, SC, TN, NS, and SLS contributed to substantive revisions. MSZ and NN conducted the statistical analysis and managed the data. TPH, BE, TN, NS, BK, ASY, and SLS critically reviewed and provided important intellectual content to the final draft. All authors were involved in data interpretation and approved the final version submitted for publication. MSZ had access to the data in the evaluation and takes responsibility for data integrity and accuracy.

Conflicts of Interest

JML has a consulting relationship with Sway Behavioral Health. Additionally, she receives grant funding from Alkermes, Inc. and BD2 (Breakthrough Discoveries for thriving with Bipolar Disorder).

References

1. Smuck M, Odonkor CA, Wilt JK, Schmidt N, Swiernik MA. The emerging clinical role of wearables: factors for successful implementation in healthcare. NPJ Digit Med 2021 Mar 10;4(1):45 [FREE Full text] [doi: 10.1038/s41746-021-00418-3] [Medline: 33692479]
2. Patient-generated health data. Office of the National Coordinator for Health Information Technology. URL: https://www.healthit.gov/topic/scientific-initiatives/pcor/patient-generated-health-data-pghd [accessed 2025-05-14]
3. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. NPJ Digit Med 2019 Sep 6;2(1):88 [FREE Full text] [doi: 10.1038/s41746-019-0166-1] [Medline: 31508498]
4. Mohr DC, Shilton K, Hotopf M. Digital phenotyping, behavioral sensing, or personal sensing: names and transparency in the digital age. NPJ Digit Med 2020 Mar 25;3(1):45 [FREE Full text] [doi: 10.1038/s41746-020-0251-5] [Medline: 32219186]
5. Nazi KM, Newton T, Armstrong CM. Unleashing the potential for patient-generated health data (PGHD). J Gen Intern Med 2024 Feb 22;39(Suppl 1):9-13 [FREE Full text] [doi: 10.1007/s11606-023-08461-4] [Medline: 38252246]
6. Quer G, Radin JM, Gadaleta M, Baca-Motes K, Ariniello L, Ramos E, et al. Wearable sensor data and self-reported symptoms for COVID-19 detection. Nat Med 2021 Jan 29;27(1):73-77 [FREE Full text] [doi: 10.1038/s41591-020-1123-x] [Medline: 33122860]
7. Gashi S, Oldrati P, Moebus M, Hilty M, Barrios L, Ozdemir F, PHRT Consortium, et al. Modeling multiple sclerosis using mobile and wearable sensor data. NPJ Digit Med 2024 Mar 11;7(1):64 [FREE Full text] [doi: 10.1038/s41746-024-01025-8] [Medline: 38467710]
8. Chakrabarti S, Biswas N, Jones LD, Kesari S, Ashili S. Smart consumer wearables as digital diagnostic tools: a review. Diagnostics (Basel) 2022 Aug 31;12(9):2110 [FREE Full text] [doi: 10.3390/diagnostics12092110] [Medline: 36140511]
9. Masoumian Hosseini M, Masoumian Hosseini ST, Qayumi K, Hosseinzadeh S, Sajadi Tabar SS. Smartwatches in healthcare medicine: assistance and monitoring; a scoping review. BMC Med Inform Decis Mak 2023 Nov 03;23(1):248 [FREE Full text] [doi: 10.1186/s12911-023-02350-w] [Medline: 37924029]
10. Torous J, Chan SR, Yee-Marie Tan S, Behrens J, Mathew I, Conrad EJ, et al. Patient smartphone ownership and interest in mobile apps to monitor symptoms of mental health conditions: a survey in four geographically distinct psychiatric clinics. JMIR Ment Health 2014 Dec;1(1):e5 [FREE Full text] [doi: 10.2196/mental.4004] [Medline: 26543905]
11. Barrera A, Ghenciulescu A, Owens N, Cortez A, De Giorgi R, Cowen P, et al. A survey of preferences for sensing technologies in people with severe mental illness admitted to an acute psychiatric unit. Ann Biomed Eng 2025 Sep 25;53(9):2372-2383 [FREE Full text] [doi: 10.1007/s10439-025-03769-0] [Medline: 40560456]
12. Bergmann J, Chandaria V, McGregor A. Wearable and implantable sensors: the patient's perspective. Sensors 2012 Dec 05;12(12):16695-16709 [FREE Full text] [doi: 10.3390/s121216695] [Medline: 23443394]
13. Hogan TP, Sherman SE, Dardashti N, McMahon N, Slightam C, Zulman DM. Realizing virtual care in VA: supporting the healthcare system's journey towards enhanced access, engagement, and outcomes. J Gen Intern Med 2024 Feb 23;39(Suppl 1):1-4 [FREE Full text] [doi: 10.1007/s11606-024-08618-9] [Medline: 38393612]
14. Share My Health Data. U.S. Department of Veterans Affairs. URL: https://mobile.va.gov/app/share-my-health-data [accessed 2025-05-14]
15. Newton TJ, Shah N, Lewis K, Zocchi MS, Bixler FR, Etingen B, et al. Developing infrastructure to realize the value of patient-generated health data in a large integrated health care system: the Veterans Health Administration experience. J Med Internet Res 2025 Jun 06;27:e70755-e70755 [FREE Full text] [doi: 10.2196/70755] [Medline: 40479588]
16. VA App Store. U.S. Department of Veterans Affairs. URL: https://mobile.va.gov/appstore [accessed 2025-11-20]

XSL•FO
RenderX

17. Wagner T, Moran E, Shen M, Gehlert E. HERC's guide to the Nosos risk adjustment score. Health Economics Resource Center, U.S. Department of Veterans Affairs. 2024 May. URL: https://www.herc.research.va.gov/include/page.asp?id=guidebook-nosos [accessed 2025-05-14]
18. American community survey 5-year data (2009-2023). United States Census Bureau. URL: https://www.census.gov/data/developers/data-sets/acs-5year.html [accessed 2025-05-19]
19. Nosos risk adjustment. U.S. Department of Veterans Affairs. URL: https://www.herc.research.va.gov/include/page.asp?id=risk-adjustment [accessed 2025-05-19]
20. Sloan K, Montez-Rath M, Spiro A, Christiansen CL, Loveland S, Shokeen P, et al. Development and validation of a psychiatric case-mix system. Med Care 2006 Jun;44(6):568-580 [FREE Full text] [doi: 10.1097/01.mlr.0000215819.76050.a1] [Medline: 16708006]
21. Wagner T, Gehlert E, Rosen A. Updating the psychiatric case mix system (PsyCMS) mental health and substance use grouper for ICD-10-CM. Technical Report 31. Health Economics Resource Center, U.S. Department of Veterans Affairs. 2016 Nov. URL: https://www.herc.research.va.gov/include/page.asp?id=technical-report-psycms-icd10 [accessed 2025-05-14]
22. Hogan TP, Etingen B, McMahon N, Bixler FR, Am L, Wacks RE, et al. Understanding adoption and preliminary effectiveness of a mobile app for chronic pain management among US military veterans: pre-post mixed methods evaluation. JMIR Form Res 2022 Jan 20;6(1):e33716 [FREE Full text] [doi: 10.2196/33716] [Medline: 35049515]
23. Lipschitz J, Miller CJ, Hogan TP, Burdick KE, Lippin-Foster R, Simon SR, et al. Adoption of mobile apps for depression and anxiety: cross-sectional survey study on patient interest and barriers to engagement. JMIR Ment Health 2019 Jan 25;6(1):e11334 [FREE Full text] [doi: 10.2196/11334] [Medline: 30681968]
24. Griffin A, Troszak L, Van Campen J, Midboe A, Zulman D. Tablet distribution to veterans: an opportunity to increase patient portal adoption and use. J Am Med Inform Assoc 2022 Dec 13;30(1):73-82 [FREE Full text] [doi: 10.1093/jamia/ocac195] [Medline: 36269168]
25. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: With Applications in R. 1st ed. Corr. 7th printing 2017 edition. Cham: Springer; 2013.
26. Lindsay JA, Caloudas A, Hogan J, Ecker AH, Day S, Day G, et al. Getting connected: a retrospective cohort investigation of video-to-home telehealth for mental health care utilization among women veterans. J Gen Intern Med 2022 Sep 30;37(Suppl 3):778-785 [FREE Full text] [doi: 10.1007/s11606-022-07594-2] [Medline: 36042096]
27. Ferguson JM, Goldstein KM, Zullig LL, Zulman DM. Gender differences in adoption and frequency of virtual primary care among men and women veterans. J Womens Health (Larchmt) 2024 Jun 01;33(6):749-757. [doi: 10.1089/jwh.2023.0781] [Medline: 38629443]
28. Liu SK, Osborn AE, Bell S, Mecchella JN, Hort S, Batsis JA. Patient characteristics and utilization of an online patient portal in a rural academic general internal medicine practice. BMC Med Inform Decis Mak 2022 Feb 16;22(1):42 [FREE Full text] [doi: 10.1186/s12911-022-01778-w] [Medline: 35172805]
29. Greenberg AJ, Haney D, Blake KD, Moser RP, Hesse BW. Differences in access to and use of electronic personal health information between rural and urban residents in the United States. J Rural Health 2018 Feb 11;34(Suppl 1):s30-s38 [FREE Full text] [doi: 10.1111/jrh.12228] [Medline: 28075508]
30. Vogels E. Digital divide persists even as Americans with lower incomes make gains in tech adoption. Pew Research Center. 2021 Jun. URL: https://pewrsr.ch/2TRM7cP [accessed 2025-05-19]
31. Guajardo E, Amspoker AB, Stanley MA, Hogan J, Ecker AH, Lindsay JA. Patterns of telehealth use for mental health treatment among Hispanic veterans. Telemed J E Health 2023 May 01;29(5):788-792. [doi: 10.1089/tmj.2022.0300] [Medline: 36282802]
32. Javier SJ, Carey EP, Gunzburger E, Chen HP, Zeliadt SB, Williams EC, et al. Pain care disparities and the use of virtual care among racial-ethnic minority groups during COVID-19. J Gen Intern Med 2024 Feb 22;39(Suppl 1):68-78 [FREE Full text] [doi: 10.1007/s11606-023-08473-0] [Medline: 38252245]
33. Day G, Ecker AH, Amspoker AB, Dawson DB, Walder A, Hogan JB, et al. Black veteran use of video telehealth for mental health care. Psychol Serv 2024 Aug;21(3):478-488. [doi: 10.1037/ser0000827] [Medline: 38059987]
34. The researcher's guide to VA data. VA Information Resource Center. URL: https://www.virec.research.va.gov [accessed 2026-01-25]

Abbreviations

HCC: Hierarchical Condition Categories
ICD: International Classification of Diseases
OR: odds ratio
PGHD: patient-generated health data
PsyCMS: psychiatric case-mix system
SMD: standardized mean difference
SMHD: Share My Health Data
VA: Veterans Health Administration

XSL•FO
RenderX

Edited by J Sarvestan; submitted 14.Jul.2025; peer-reviewed by C Turvey, CK Yoo, S Hosseini; comments to author 01.Aug.2025; revised version received 30.Jan.2026; accepted 31.Jan.2026; published 26.Mar.2026.

Please cite as:
Zocchi MS, Lipschitz JM, Bixler FR, Robinson SA, Etingen B, Hogan TP, Ndiwane N, Calkins S, Newton T, Shah N, Kragen B, Young AS, Shimada SL
Patient Sharing of Digital Health Data in the Veterans Health Administration: Cross-Sectional Analysis
J Med Internet Res 2026;28:e80517
URL: https://www.jmir.org/2026/1/e80517
doi:10.2196/80517
PMID:

©Mark S Zocchi, Jessica M Lipschitz, Felicia R Bixler, Stephanie A Robinson, Bella Etingen, Timothy P Hogan, Ndindam Ndiwane, Saige Calkins, Terry Newton, Nilesh Shah, Ben Kragen, Alexander S Young, Stephanie L Shimada. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Use and Influencing Factors of mHealth Services Among Adult Survivors of Cancer: Cross-Sectional Survey Study

Taiguo Liu[1*], MD; Lina Duo[2*], MD; Musi Zhang[3*], MNS; Jianjun Zhang[4*], MD; Jingna Zhang[5*], MD; Xiao Sun[2], MM; Haoyue Deng[2], MM; Ji Cheng[2]; Ya Sun[2], MM

[1]Department of Oncology, Chengdu Seventh People's Hospital (Affiliated Cancer Hospital of Chengdu Medical College), Chengdu, China

[2]Department of Dermatology & Medical Aesthetics, Chengdu Seventh People's Hospital (Affiliated Cancer Hospital of Chengdu Medical College), Chengdu, China

[3]Department of Radiation Oncology, Sichuan Cancer Hospital&Institute, Sichuan Cancer Center, University of Electronic Science and Technology of China, Chengdu, China

[4]Department of Gynecology and Obstetrics, West China Second Hospital, Sichuan University/Key Laboratory of Birth Defects and Related Diseases of Women and Children (Sichuan University), Ministry of Education, Chengdu, China

[5]Department of Palliative Medicine, West China Fourth Hospital, Sichuan University, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Taiguo Liu, MD
Department of Oncology
Chengdu Seventh People's Hospital (Affiliated Cancer Hospital of Chengdu Medical College)
51 Zhimin Road, Wuhou District
Chengdu, 610000
China
Phone: 86 18780041235
Email: liutaiguo1125@163.com

Abstract

Background: The growing number of survivors of cancer in China has created an increasing need for survivorship care as many survivors face ongoing physical, psychological, and social challenges after treatment. Mobile health (mHealth) services, which are delivered through mobile devices and apps, have emerged as potential tools to support self-management, facilitate access to care, and improve quality of life. However, evidence on the prevalence, use patterns, and determinants of mHealth adoption among Chinese survivors of cancer remains limited.

Objective: This study aimed to examine the prevalence and patterns of mHealth use among adult survivors of cancer in China and identify sociodemographic and clinical factors associated with adoption.

Methods: We conducted a multicenter cross-sectional survey study between February 13, 2024, and September 21, 2024, at 4 tertiary cancer care centers in China. Adult survivors of cancer were recruited using convenience sampling. Data were collected through structured face-to-face questionnaires assessing sociodemographic and clinical characteristics, mHealth use, perceived needs, concerns, and user experience. Missing data were limited and handled using complete-case analysis after the Little test supported consistency with missing completely at random. Descriptive analyses summarized participant characteristics and mHealth-related variables. Group differences were examined using chi-square tests and 2-tailed independent-sample *t* tests. Significant variables in univariate analyses were entered into a multivariable logistic regression model.

Results: Of 1152 participants, 364 (31.6%) reported prior mHealth use. Use was concentrated in practical functions, particularly appointment booking (301/364, 82.7%), online consultation (244/364, 67%), and viewing examination or laboratory reports (215/364, 59.1%), with WeChat-based platforms being the most commonly used access channel (244/364, 67%). Participants reported high demand for clinical guidance (917/1152, 79.6%) and direct communication with health care professionals (901/1152, 78.2%), whereas common concerns included leakage of private information (694/1152, 60.2%) and inaccurate illness judgment (633/1152, 54.9%). In multivariable analysis, mHealth use was significantly associated with younger age, higher educational level, annual household income of at least ¥100,000 (US $14,527.90), widowed or divorced marital status, living alone, and treatment dissatisfaction; cancer type and time since diagnosis were not significant predictors.

Conclusions: mHealth use among adult survivors of cancer in China is established but uneven, with use concentrated in practical service functions rather than comprehensive survivorship support. Uptake was significantly associated with age, socioeconomic

XSL•FO
RenderX

position, social circumstances, and treatment experience but not with clinical characteristics. Survivors reported a strong demand for clinically relevant and communication-oriented functions but also expressed substantial concerns about privacy, accuracy, reimbursement, and physician authenticity. Future survivorship mHealth services should prioritize clinical relevance, trust, integration with formal care, and equitable implementation to achieve broader and more meaningful use.

(J Med Internet Res 2026;28:e82902) doi:10.2196/82902

KEYWORDS

survivor of cancer; mobile health; mHealth; adoption; influencing factor; China; cross-sectional study

Introduction

The growing population of survivors of cancer has made survivorship care an increasingly important public health priority worldwide, with the number of adult survivors of cancer rising rapidly due to advances in early detection, treatment, and supportive care. In China, survivorship care is shaped by a combination of rapid growth in the survivor population, persistent urban-rural disparities in health care access and use, variable continuity of follow-up care, and uneven distribution of digital and supportive care resources, all of which may influence how survivors manage long-term physical, psychological, and social consequences after treatment [1,2]. Mobile health (mHealth) services, defined as the use of mobile devices and apps to deliver health information and support, have emerged as promising tools to address the complex needs of survivors of cancer by facilitating self-management, improving access to care, and enhancing quality of life (QOL) [3-5].

The literature on mHealth interventions for survivors of cancer is extensive and rapidly evolving. Systematic reviews and meta-analyses have consistently demonstrated that mHealth interventions can improve health-related QOL, self-efficacy, and psychological outcomes such as anxiety and depression among adult survivors of cancer [3,6-8]. For example, a recent meta-analysis of randomized controlled trials found that mobile app–based interventions significantly improved QOL and self-efficacy and reduced anxiety, depression, and distress, with short-term and interactive interventions showing the greatest benefits [3]. Similarly, mHealth interventions targeting physical activity and dietary behaviors have shown promise in promoting healthy lifestyles and improving behavioral outcomes among survivors of cancer, although most studies have focused on short-term effects and predominantly included college-educated participants [9-12]. The use of mHealth for symptom monitoring and management has also been shown to be feasible and acceptable, with high adherence rates and positive associations between reduced symptom burden and improved health-related QOL [13-15].

Despite encouraging findings, important gaps remain. Engagement and sustained use of mHealth tools are often suboptimal. Uptake and adherence may be influenced by technical barriers, competing priorities, age, digital literacy, and psychological status [7,16,17]. Rural-urban disparities have also been documented, with rural survivors of cancer being less likely than urban survivors to own mobile devices and use mHealth apps [18-20]. In China, studies have highlighted suboptimal health care use among survivors of cancer, particularly in rural areas, and identified socioeconomic status, health insurance coverage, and clinical factors as key determinants of service use [1,21]. While mHealth has the potential to bridge some of the gaps, research on its adoption and influencing factors among Chinese survivors of cancer remains limited.

Furthermore, the quality and effectiveness of mHealth interventions are highly variable, with significant heterogeneity in intervention design, delivery methods, theoretical underpinnings, and outcome measures [17,22]. Reviews have called for more rigorous, theory-driven research to establish best practices for mHealth intervention development, implementation, and evaluation, as well as greater attention to issues of equity, usability, and patient-centeredness [17,22,23]. The digital divide persists, with older adults, ethnic minority groups, and those with lower socioeconomic status less likely to benefit from mHealth innovations [9,19,20]. In the Chinese context, the adoption and impact of mHealth services are further shaped by urban-rural differences in service availability, differences in insurance and reimbursement arrangements, the central role of widely used super app platforms such as WeChat, and variations in health and digital literacy across survivor groups [1,2,21].

Despite the proliferation of mHealth interventions worldwide, there is a notable research gap regarding the use patterns and determinants of mHealth service adoption among adult survivors of cancer in China. Most existing studies have been conducted in Western populations or have focused on specific cancer types, short-term outcomes, or feasibility rather than comprehensive assessments of use and influencing factors in diverse Chinese settings [9,21,23]. There is a critical need for empirical data on how adult survivors of cancer in China engage with mHealth services, what factors facilitate or hinder their use, and how the services can be optimized to address the unique needs of this population.

The aim of this study was to examine the prevalence, use patterns, and correlates of mHealth service use among adult survivors of cancer in China. Using a multicenter cross-sectional design, we sought to identify sociodemographic, clinical, and care-related factors associated with mHealth adoption, as well as survivors' needs, concerns, and perceived barriers.

Methods

Study Design and Participants

An institution-based cross-sectional survey study was conducted at 4 tertiary cancer care centers in China between February 13, 2024, and September 21, 2024. Convenience sampling was used to recruit adult survivors of cancer attending outpatient services

XSL•FO
RenderX

at the participating centers. Eligible individuals were aged 18 years or older, had a confirmed cancer diagnosis, demonstrated adequate communication ability, were able to accurately report their personal health information, and provided written informed consent. Individuals with cognitive impairment who were unable to understand the questionnaire after explanation by the investigators were excluded.

Sample Size Estimation

The sample size was calculated using the following formula:

where n is the required sample size, Z is the standard normal deviate at a 95% confidence level (Z=1.96), p is the expected prevalence of mHealth use, and δ is the allowable error. On the basis of prior literature, the prevalence of mHealth use among survivors of cancer was assumed to be 30% (P=.30). The allowable error was set as a δ value of 0.1 and p value of 0.03 to ensure moderate precision. Thus, a minimum of 896 participants was required. To account for potential invalid responses, the sample size was inflated by 20%, resulting in a final target of approximately 1075 participants.

Questionnaire Development and Validation

The questionnaire was developed based on prior literature on mHealth use among survivors of cancer and chronic disease populations [24-27] and was refined through consultation with 3 oncology specialists and 2 digital health researchers to ensure content validity. A pilot test was conducted with 30 adult survivors of cancer recruited from the Department of Oncology, Chengdu Seventh People's Hospital, who were not included in the final analytic sample. The pilot confirmed face validity and comprehensibility and led to minor wording revisions. In the final sample, internal consistency was acceptable for the multi-item domains, with a Cronbach α of 0.842 for needs, 0.813 for concerns, and 0.764 for satisfaction. The final questionnaire contained 47 items. See Multimedia Appendix 1 for the questionnaire.

Data Collection

Data were collected face-to-face using a structured questionnaire administered by trained health care professionals. All investigators underwent standardized training to ensure consistent questionnaire administration. Questionnaires did not include direct personal identifiers. Data were entered into a study dataset accessible only to the research team. To minimize duplicate participation, each participant completed the questionnaire once during on-site recruitment under investigator supervision.

Questionnaire Measures

All study variables were collected through self-report using the structured questionnaire.

Sociodemographic and Clinical Characteristics

Information was collected on sex, age, educational level, annual household income, marital status, residence (urban or rural), occupation, cohabitation status, smoking and drinking habits, cancer type (solid vs nonsolid), time since diagnosis (<12 or >12 months), treatment modalities (surgery, chemotherapy, radiotherapy, targeted therapy, immunotherapy, or others), comorbid chronic conditions, and satisfaction with treatment.

mHealth Use

Participants were asked whether they had ever used mHealth services. For users, additional information was collected, including frequency of use, average duration per session, main platforms or apps used, functions accessed (such as online consultation, appointment booking, health information, or medication reminders), and satisfaction with the services.

Needs and Expectations for mHealth

Participants were asked about their perceived needs for various mHealth functions, including clinical guidance (such as medication, rehabilitation, and psychological support), communication with physicians, lifestyle guidance, home-based disease management, access to health information, regular follow-up, family assistance, peer support, electronic medical record access, and online purchasing of medicines or rehabilitation equipment.

Perceptions of Potential Problems

Concerns regarding mHealth use were assessed, including data inaccuracy, privacy risks, misjudgment of illness, reimbursement restrictions, authenticity of physician information (identity, credentials, and work experience), limitations in purchasing medicines, and service costs.

User Experience and System Performance Measures

Participants evaluated ease of use, system stability, prevalence of advertisements, concerns about fraud, operability of the system, ability to deliver disease-specific treatments, and integration with hospital information systems.

Statistical Analysis

Overview

Data were entered into a database using Microsoft Excel and analyzed using SPSS software (version 26.0; IBM Corp). Descriptive statistics were used to summarize participant characteristics and mHealth-related variables, with continuous variables presented as means and SDs and categorical variables presented as frequencies and percentages. Group differences between mHealth users and nonusers were examined using chi-square tests for categorical variables and 2-tailed independent-sample t tests for continuous variables. Variables associated with mHealth use in univariate analyses at a P value of less than .05 were entered into the multivariable logistic regression model to identify factors independently associated with mHealth use. Adjusted odds ratios, 95% CIs, and P values are reported. A 2-sided P value of less than .05 was considered statistically significant.

Missing Data Handling

Missing data were assessed for all variables included in descriptive and regression analyses. The proportion of missingness for individual variables ranged from 0.3% to 4.6%, and the overall proportion of missing values across the analyzed variables was 2.1%. As the extent of missingness was low, we

XSL•FO
RenderX

formally evaluated the missing data mechanism before selecting the analytic approach. The Little test suggested that the missing data pattern was consistent with missing completely at random (MCAR; χ^2_{47}=52.8; *P*=.26). On that basis, and because the proportion of missing data was small, we used complete-case analysis. No score prorating was performed.

Ethical Considerations

The research protocol was reviewed and approved by the ethics committee of Chengdu Seventh People's Hospital (approval KY0182). All participants provided written informed consent after being informed of the purpose, procedures, risks, and benefits of the study at the beginning of the survey. Participants' privacy and confidentiality were protected by assigning unique study codes, storing data on encrypted servers with restricted access, and removing all personal identifiers prior to analysis. No identifiable participant information appears in the manuscript or supplementary files. Participants received no compensation. Data handling complied with local and international regulations on personal data protection, including the Helsinki Declaration.

Results

Participant Recruitment and Sample Characteristics

Participants completed the questionnaire in an average of 614 (range 472-922) seconds. Of 1273 questionnaires distributed, 1152 (90.5%) valid questionnaires were included in the final analysis. A total of 121 questionnaires were excluded because of substantial item nonresponse (n=54, 44.6%), duplicate completion (n=11, 9.1%), ineligibility identified after screening (n=9, 7.4%), internally inconsistent responses (n=27, 22.3%), or patterned responding suggestive of inattentive completion (n=20, 16.5%; Figure 1).

Figure 1. Participant recruitment and inclusion flowchart.

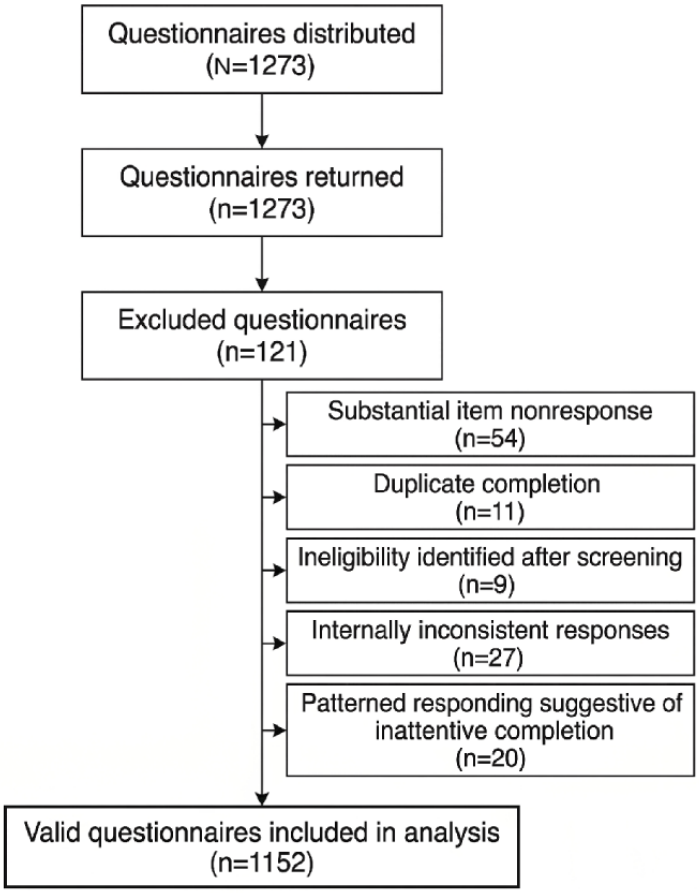

Among the 1152 participants, the mean age was 55.7 (SD 11.8) years. Overall, 52% (n=599) of the participants were female, 81.8% (n=942) resided in urban areas, 79.4% (n=915) were married, and 82.7% (n=953) had solid tumors. mHealth use was reported by 31.6% (n=364) of the participants, whereas 68.4% (n=788) reported no prior use. Variable-level missingness across the analyzed variables was low, and complete data on the outcome and all covariates required for the multivariable model were available for 97% (n=1118) of the participants.

As shown in Table 1, mHealth users were younger on average than nonusers (mean 52.4, SD 11.1 years vs mean 57.2, SD 11.9 years; *P*<.001) and were more likely to have undergraduate or postgraduate education (174/364, 47.8% vs 249/788, 31.6%; *P*<.001), an annual household income of at least ¥100,000 (US $14,527.90; 71/364, 19.5% vs 77/788, 9.8%; *P*<.001), and widowed or divorced marital status (72/364, 19.8% vs 100/788, 12.7%; *P*=.002) and to live alone (59/364, 16.2% vs 88/788, 11.2%; *P*=.02). Treatment dissatisfaction was also more common among users than nonusers (79/364, 21.7% vs 67/788, 8.5%; *P*<.001). No statistically significant group differences were observed for sex, residence, employment status, smoking, alcohol consumption, cancer type, or time since diagnosis.

XSL•FO
RenderX

Table 1. Participant characteristics overall and by mobile health (mHealth) use status (N=1152)[a].

Characteristic	Total sample	mHealth users (n=364)	Nonusers (n=788)	*P* value
Age (y), mean (SD)	55.7 (11.8)	52.4 (11.1)	57.2 (11.9)	<.001
Female sex, n (%)	599 (52)	197 (54.1)	402 (51)	.33
Urban residence, n (%)	942 (81.8)	309 (84.9)	633 (80.3)	.07
Educational level, n (%)				<.001
High school or lower	729 (63.3)	190 (52.2)	539 (68.4)	
Undergraduate	357 (31)	138 (37.9)	219 (27.8)	
Postgraduate or higher	66 (5.7)	36 (9.9)	30 (3.8)	
Annual household income, n (%)				<.001
<¥50,000 (US $7263.95)	621 (53.9)	164 (45.1)	457 (58)	
¥50,000 to ¥99,999 (US $7263.95 to US $14,527.80)	383 (33.2)	129 (35.4)	254 (32.2)	
≥¥100,000 (US $14,527.90)	148 (12.8)	71 (19.5)	77 (9.8)	
Marital status, n (%)				.002
Married	915 (79.4)	272 (74.7)	643 (81.6)	
Unmarried	65 (5.6)	20 (5.5)	45 (5.7)	
Widowed or divorced	172 (14.9)	72 (19.8)	100 (12.7)	
Living with others, n (%)	1005 (87.2)	305 (83.8)	700 (88.8)	.02
Currently employed, n (%)	243 (21.1)	85 (23.4)	158 (20.1)	.18
Current or former smoker, n (%)	182 (15.8)	63 (17.3)	119 (15.1)	.31
Current alcohol use, n (%)	201 (17.4)	73 (20.1)	128 (16.2)	.10
Solid tumor, n (%)	953 (82.7)	300 (82.4)	653 (82.9)	.84
Time since diagnosis of ≥12 mo, n (%)	676 (58.7)	222 (61)	454 (57.6)	.29
Treatment satisfaction, n (%)				<.001
Satisfied	472 (41)	120 (33)	352 (44.7)	
Neutral	534 (46.4)	165 (45.3)	369 (46.8)	
Dissatisfied	146 (12.7)	79 (21.7)	67 (8.5)	

[a]*P* values were derived from independent-sample *t* tests for continuous variables and chi-square tests for categorical variables.

Perceived Needs and Concerns Regarding mHealth Services Among All Participants

Across the full sample, perceived need for mHealth support was high. The most frequently endorsed needs were clinical guidance (917/1152, 79.6%), direct communication with health care professionals (901/1152, 78.2%), lifestyle guidance (846/1152, 73.4%), regular follow-up and evaluation (801/1152, 69.5%), and home-based disease management (780/1152, 67.7%). More than two-thirds of the participants (774/1152, 67.2%) also reported a need for access to reliable health information, whereas approximately half endorsed needs related to family assistance in disease management (628/1152, 54.5%), electronic medical record access (594/1152, 51.6%), peer support (562/1152, 48.8%), and online medication or equipment ordering (514/1152, 44.6%).

Concerns about mHealth services were also common. The most frequently reported concerns were leakage of private information (694/1152, 60.2%), inaccurate illness judgment (633/1152, 54.9%), and inaccurate information collection (596/1152, 51.7%). Reimbursement restrictions were endorsed by 42.9% (494/1152) of the participants, and 37.5% (432/1152) expressed concern about fraudulent or inauthentic physician information. Difficulty purchasing medicines through the platform was reported by 36.3% (418/1152) of the participants, whereas high service costs were reported by 27% (311/1152). Detailed frequencies are presented in Table 2.

XSL•FO
RenderX

Table 2. Perceived needs and concerns regarding mobile health services among all participants (N=1152).

Variable	Participants, n (%)
Perceived needs[a]	
Clinical guidance	917 (79.6)
Direct communication with health care professionals	901 (78.2)
Lifestyle guidance	846 (73.4)
Regular follow-up and evaluation	801 (69.5)
Home-based disease management	780 (67.7)
Access to reliable health information	774 (67.2)
Family assistance in disease management	628 (54.5)
Electronic medical record access	594 (51.6)
Peer support	562 (48.8)
Online medication or equipment ordering	514 (44.6)
Concerns and barriers[a]	
Leakage of private information	694 (60.2)
Inaccurate illness judgment	633 (54.9)
Inaccurate information collection	596 (51.7)
Reimbursement restrictions	494 (42.9)
Fraudulent or inauthentic physician information	432 (37.5)
Difficulty purchasing medicines through the platform	418 (36.3)
High service costs	311 (27)

[a]Multiple responses were permitted; percentages do not add up to 100%.

Factors Associated With mHealth Use

Variables significantly associated with mHealth use in univariate analyses were entered into the multivariable logistic regression model. The adjusted model was based on 1118 participants with complete data on the outcome and all included covariates. In the final model, younger age, higher educational level, higher annual household income, widowed or divorced marital status, living alone, and treatment dissatisfaction were independently associated with higher odds of mHealth use. The model showed acceptable fit (Hosmer-Lemeshow P=.60) and moderate discrimination (area under the receiver operating characteristic curve=0.74).

In the adjusted model, the strongest positive associations with mHealth use were observed for postgraduate educational level, treatment dissatisfaction, and annual household income of at least ¥100,000 (US $14,527.90). Undergraduate educational level, widowed or divorced marital status, and living alone were also significantly associated with higher odds of use, whereas older age was significantly associated with lower odds of use (P≤.046 in all cases). In contrast, annual household income of ¥50,000 (US $7263.95) to ¥99,999 (US $14,527.80), unmarried status, and neutral treatment satisfaction were not statistically significant (P≥.10 in all cases; Table 3; Figure 2).

XSL•FO
RenderX

Table 3. Multivariable logistic regression analysis of factors associated with mobile health use (n=1118).

Variable	Adjusted OR[a] (95% CI)	*P* value
Age per 10-y increase	0.82 (0.73-0.92)	<.001
Educational level		
High school or lower	Reference	__[b]
Undergraduate	1.73 (1.29-2.33)	<.001
Postgraduate or higher	3.12 (1.82-5.34)	<.001
Annual household income		
<¥50,000 (US $7263.95)	Reference	—
¥50,000 to ¥99,999 (US $7263.95 to US $14,527.80)	1.24 (0.92-1.67)	.16
≥¥100,000 (US $14,527.90)	2.21 (1.45-3.37)	<.001
Marital status		
Married	Reference	—
Unmarried	1.02 (0.56-1.84)	.95
Widowed or divorced	1.68 (1.17-2.42)	.005
Living alone	1.49 (1.01-2.21)	.046
Treatment satisfaction		
Satisfied	Reference	—
Neutral	1.28 (0.95-1.72)	.10
Dissatisfied	3.08 (2.02-4.68)	<.001

[a]OR: odds ratio.
[b]Not applicable.

XSL•FO
RenderX

Figure 2. Adjusted odds ratios for factors associated with mobile health (mHealth) use. Forest plot of adjusted odds ratios from the multivariable logistic regression model computed from the dataset. Younger age, higher educational level, higher annual household income, widowed or divorced marital status, and treatment dissatisfaction were associated with greater odds of mHealth use. Error bars represent 95% CIs.

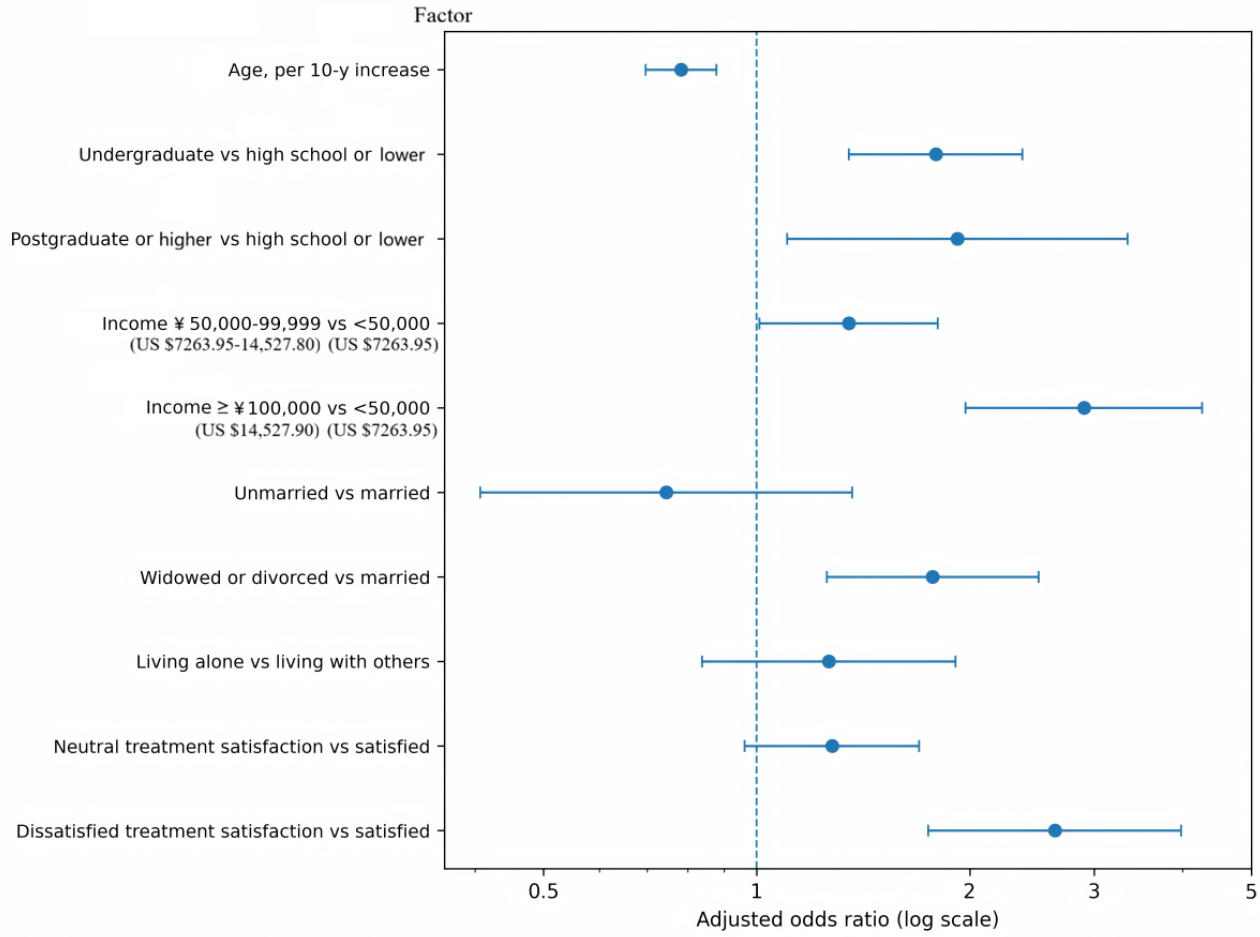

Patterns of mHealth Use Among Users

Among the 364 mHealth users, the most commonly reported frequency of use was monthly (n=126, 34.6%), followed by weekly (n=118, 32.4%), less than monthly (n=62, 17%), and daily (n=58, 15.9%). The mean duration of a typical use session was 22.4 (SD 10.2) minutes. Most users (n=244, 67%) primarily accessed services through WeChat-based platforms, whereas 25% (n=91) primarily used dedicated mHealth apps and 8% (n=29) primarily used hospital web-based portals or hospital apps.

Appointment booking (301/364, 82.7%) and online consultation (244/364, 67%) were the most commonly used functions, followed by viewing examination or laboratory reports (215/364, 59.1%), accessing health information (185/364, 50.8%), medication reminders (128/364, 35.2%), and symptom self-monitoring (117/364, 32.1%). Peer support or patient community functions were less commonly used (69/364, 19%). Overall satisfaction with mHealth services was favorable, with 67.3% (245/364) of the users reporting that they were satisfied or very satisfied with their experience (Table 4).

XSL•FO
RenderX

Table 4. Patterns of mobile health (mHealth) use among users (n=364).

Variable	Participants, n (%)
Frequency of use	
Daily	58 (15.9)
Weekly	118 (32.4)
Monthly	126 (34.6)
Less than monthly	62 (17)
Typical session duration (min), mean (SD)	22.4 (10.2)
Primary platform used	
WeChat-based platform	244 (67)
Dedicated mHealth app	91 (25)
Hospital website or hospital app	29 (8)
Commonly used functions[a]	
Appointment booking	301 (82.7)
Online consultation	244 (67)
Viewing examination or laboratory reports	215 (59.1)
Accessing health information	185 (50.8)
Medication reminders	128 (35.2)
Symptom self-monitoring	117 (32.1)
Peer support or community interaction	69 (19)
Overall satisfaction with mHealth services	
Very satisfied	61 (16.8)
Satisfied	184 (50.5)
Neutral	93 (25.5)
Dissatisfied	26 (7.1)

[a]Multiple responses were permitted for commonly used functions; percentages for that section do not add up to 100%.

User Experience and System Performance Among mHealth Users

User-reported experience with mHealth services was generally positive but not uniformly so. Among users, 66.2% (241/364) described the services as easy or very easy to use, whereas 8.5% (31/364) described them as difficult. In terms of convenience, 68.4% (249/364) of the users rated the services as somewhat or very convenient, whereas 7.1% (26/364) rated them as inconvenient. System stability was moderate: 38.7% (141/364) of the users reported never experiencing crashes, 37.6% (137/364) reported crashes rarely, 19.2% (70/364) reported them sometimes, and 4.4% (16/364) reported them often.

Perceptions of clinical utility were more mixed. Only 40.9% (149/364) of users believed that the mHealth services they used could provide disease-specific treatment support, whereas 40.7% (148/364) believed that they could not and 18.4% (67/364) were uncertain. Perceived integration with hospital services was more favorable, with 61.3% (223/364) of users reporting that the service they used was integrated with hospital-based care, 17.6% (64/364) being uncertain, and 21.2% (77/364) reporting no integration. Additional user experience indicators are presented in Table 5.

XSL•FO
RenderX

Table 5. User experience and system performance among mobile health users (n=364).

Variable	Participants, n (%)
Ease of use	
Very easy	76 (20.9)
Easy	165 (45.3)
Neutral	92 (25.3)
Difficult	31 (8.5)
Overall convenience	
Very convenient	68 (18.7)
Somewhat convenient	181 (49.7)
Neutral	89 (24.5)
Inconvenient	26 (7.1)
Frequency of system crashes	
Never	141 (38.7)
Rarely	137 (37.6)
Sometimes	70 (19.2)
Often	16 (4.4)
Advertising burden	
Few advertisements	158 (43.4)
Moderate number of advertisements	112 (30.8)
Many advertisements	94 (25.8)
Can provide disease-specific treatment support	
Yes	149 (40.9)
Uncertain	67 (18.4)
No	148 (40.7)
Integrated with hospital services	
Yes	223 (61.3)
Uncertain	64 (17.6)
No	77 (21.2)

Discussion

Principal Findings

In this multicenter cross-sectional survey study of adult survivors of cancer in China, approximately one-third of participants (364/1152, 31.6%) reported prior mHealth use. Use was concentrated on practical care navigation functions, especially appointment booking, online consultation, and report viewing, and most users (244/364, 67%) accessed services through WeChat-based platforms rather than stand-alone apps. In adjusted analyses, mHealth use was associated primarily with age, educational level, income, social circumstances, and treatment dissatisfaction rather than with clinical characteristics. At the same time, survivors expressed a strong demand for clinically relevant functions, especially guidance and direct communication with health care professionals, while also reporting substantial concerns about privacy, accuracy, reimbursement, and physician authenticity. Taken together, the findings suggest that mHealth is already a meaningful component of survivorship care for some patients, but its use remains selective rather than broad based. Current engagement appears to be driven more by immediate practical needs and perceived gaps in care than by mature adoption of comprehensive digital survivorship support.

Comparison With Prior Work and Interpretation of Correlates

The adjusted associations observed in this study suggest that mHealth use among adult survivors of cancer in China is shaped less by clinical characteristics than by digital readiness and perceived unmet needs. Younger age, higher educational level, and higher income were associated with greater mHealth use, a pattern that is consistent with that observed in prior literature showing that digital health uptake is socially patterned and facilitated by greater digital access, confidence, and ability to navigate app-based care [28,29]. In contrast, cancer type and time since diagnosis were not significant in the adjusted model,

XSL•FO
RenderX

suggesting that engagement with mHealth in this sample was not primarily determined by the disease profile itself [30].

The social and care experience correlates point in a complementary direction. Widowed or divorced marital status, living alone, and treatment dissatisfaction were associated with higher mHealth use, which may reflect greater motivation to seek supplementary support when day-to-day support is more limited or when conventional care is perceived as insufficient [24,31]. Viewed together, these findings suggest that the significant correlates are connected rather than coincidental: some factors may increase capacity to engage with mHealth, whereas others may increase motivation to use it [28,31]. This interpretation helps explain why uptake appears to be socially and experientially patterned rather than primarily disease driven, and it suggests that implementation strategies should not rely on clinical stratification alone because real-world use may depend more on who is able and motivated to engage with mHealth [24,32].

Perceived Needs for mHealth Services

The needs profile observed in this study suggests that survivors are seeking clinically meaningful support rather than purely informational content [33,34]. The highest-demand functions centered on guidance, communication, follow-up, and home-based disease management, indicating that survivors view mHealth as a potential extension of survivorship care rather than merely a source of general health information [33,34].

The prominence of communication-related needs is especially important. The finding that 78.2% (901/1152) of the participants desired direct communication with health care professionals suggests that digital health tools may be most acceptable when they enhance continuity and responsiveness rather than when they merely provide static educational content [33,34]. Similarly, the high demand for clinical guidance and follow-up indicates that survivors may value digital services most when they support ongoing management rather than one-time information retrieval [33]. This interpretation is reinforced by the use pattern findings, where appointment booking, online consultation, and result viewing were much more common than peer support functions. Together, these results suggest that survivors prioritize utility, accessibility, and clinical relevance over more discretionary or socially oriented features [33].

The demand for electronic medical record access, reliable health information, and family assistance in disease management also points to the importance of integrating mHealth into broader care systems [14,35,36]. Survivors are not simply asking for more digital tools; they appear to be asking for tools that help them coordinate care, interpret medical information, and manage illness in the home environment [24,37]. This has implications for future design. mHealth interventions for cancer survivorship are likely to be more valuable when they are linked to clinical workflows, adapted to family involvement, and structured around specific self-management tasks rather than around generic health promotion messages [17,36,38].

Concerns, Barriers, and User Experience

Although the perceived need for mHealth services was high, concerns and barriers were also prominent. The most common concerns centered on privacy, accuracy, and reliability, indicating that survivors are attentive not only to what digital services can offer but also to the risks associated with using them [39,40]. Privacy concerns are particularly important in oncology settings, where personal health information may be perceived as sensitive and stigma may remain relevant in some contexts [41]. Concerns about inaccurate judgment and inaccurate information collection point to a deeper issue of clinical trust [42,43]. Survivors may be willing to use digital services for administrative functions while remaining uncertain about their reliability for individualized assessment or treatment-related decision support [39,44,45].

The concern about reimbursement restrictions, reported by 42.9% (494/1152) of the participants, should also be taken seriously. This is not a marginal finding. Rather, it suggests that payment arrangements and insurance integration remain practical barriers to wider adoption [24,40,44]. Even well-designed digital services may remain underused if they are not financially accessible or if their status within formal reimbursement structures is unclear [39,40]. Similarly, more than one-third of participants (432/1152, 37.5%) expressed concern about fraudulent or inauthentic physician information, which highlights the role of physician verification and institutional credibility. In this context, trust is not only a technical issue but also a governance issue [35,39].

The user experience findings suggest that current mHealth services are generally usable and convenient but are not yet consistently trusted as tools for disease-specific clinical support. Therefore, improving usability alone is unlikely to be sufficient; stronger clinical linkage, transparent quality assurance, and clearer integration with formal oncology care may be necessary to improve trust and sustain engagement [39,42,45].

Implications for Practice and Research

These findings have several practical implications. Survivorship-oriented mHealth services should prioritize functions that survivors clearly value, particularly clinical guidance, communication, follow-up, and home-based disease management. Because uptake appears to be facilitated by familiarity and convenience, integration into commonly used platforms such as WeChat may be more effective than reliance on stand-alone apps alone. At the same time, the social patterning of use suggests that implementation should include explicit support for older adults and survivors with fewer educational or financial resources.

Trust building should also be central to implementation. Privacy protection, physician authentication, content quality control, and clearer reimbursement pathways may influence uptake as much as interface design. Future research should examine patterns of sustained use over time, discontinuation, and whether mHealth engagement improves survivorship outcomes in specific patient subgroups.

Strengths and Limitations

This study has several strengths. It used a multicenter design, included a relatively large sample of adult survivors of cancer, and achieved a high effective response rate. It also examined not only whether survivors used mHealth services but also how

XSL•FO
RenderX

they used them, what they wanted from them, and what concerns they had. This broader perspective is valuable because uptake alone provides an incomplete picture of digital health implementation in survivorship care.

Several limitations should also be considered. First, the cross-sectional design precludes causal inference, and the observed associations should not be interpreted as directional. For example, treatment dissatisfaction may increase mHealth use, but mHealth use may also shape how survivors evaluate their care experiences. Second, participants were recruited using convenience sampling from tertiary cancer centers, which may limit generalizability to survivors receiving care in other settings, particularly rural or lower-resource environments. Although the effective response rate was high, some degree of self-selection bias likely remains because participation depended on survivors being present, eligible, and willing to complete the survey. Third, all variables were self-reported and, therefore, are subject to recall bias and response bias. This is especially relevant for prior mHealth use, perceived needs, treatment satisfaction, and user experience items. Fourth, missing data were limited, and the Little MCAR test supported the use of complete-case analysis, but this approach may still have reduced precision and could introduce bias if the true missing data mechanism deviated from MCAR. Finally, although the questionnaire was developed from prior literature, reviewed by experts, piloted, and showed acceptable internal consistency for several multi-item domains, it was not a fully validated instrument. Accordingly, some measurement error or limitations in construct validity may remain.

Conclusions

mHealth use among adult survivors of cancer in China is established but uneven, with use concentrated on practical functions such as appointment booking, online consultation, and result viewing rather than comprehensive survivorship support. Uptake was significantly associated with younger age, higher educational level, higher income, widowed or divorced marital status, and treatment dissatisfaction but not with clinical characteristics, suggesting that digital engagement reflects both resource-related differences and perceived unmet needs. Survivors reported a strong demand for clinically relevant functions, especially guidance, communication, follow-up, and home-based management, but also expressed substantial concerns about privacy, accuracy, reimbursement, and physician authenticity. These findings indicate that future survivorship mHealth services should prioritize clinical relevance, trust, integration with formal care, and equitable implementation if they are to achieve broader and more meaningful use.

Acknowledgments

We used Kimi chat (Moonshot AI) and ChatGPT (OpenAI) to translate the initial draft and polish the English language in the final draft.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Funding

No external funding was received.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire (English translation).
[PDF File (Adobe PDF File), 97 KB - jmir_v28i1e82902_app1.pdf]

References

1. Wang H, Hua X, Yao N, Zhang N, Wang J, Anderson R, et al. The urban-rural disparities and associated factors of health care utilization among cancer patients in China. Front Public Health 2022 Mar 4;10:842837 [FREE Full text] [doi: 10.3389/fpubh.2022.842837] [Medline: 35309211]
2. Francis Ye J, Zou W, Huang L, Lai YK, Zhang L, Zhao X. Health information acquisition and lifestyle choices among cancer survivors in China: examining the roles of patient-centered communication, patient trust, and system trust. J Health Psychol 2025 Nov;30(13):3865-3879. [doi: 10.1177/13591053251321782] [Medline: 40079257]
3. Qin M, Chen B, Sun S, Liu X. Effect of mobile phone app-based interventions on quality of life and psychological symptoms among adult cancer survivors: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2022 Dec 19;24(12):e39799 [FREE Full text] [doi: 10.2196/39799] [Medline: 36534460]
4. Buneviciene I, Mekary RA, Smith TR, Onnela JP, Bunevicius A. Can mHealth interventions improve quality of life of cancer patients? A systematic review and meta-analysis. Crit Rev Oncol Hematol 2021 Jan;157:103123 [FREE Full text] [doi: 10.1016/j.critrevonc.2020.103123] [Medline: 33190065]

XSL•FO
RenderX

5. Zhang X, Sun S, Jiangenuer L, Zhao P, Lei H, Xu Z, et al. Effect of mobile health (mHealth) on improving anxiety, depression and quality of life in cancer patients: a systematic review and meta-analysis. J Affect Disord 2025 Apr 01;374:11-25. [doi: 10.1016/j.jad.2025.01.016] [Medline: 39793615]
6. Luo X, Chen Y, Chen J, Zhang Y, Li M, Xiong C, et al. Effectiveness of mobile health-based self-management interventions in breast cancer patients: a meta-analysis. Support Care Cancer 2022 Mar;30(3):2853-2876. [doi: 10.1007/s00520-021-06568-0] [Medline: 34561732]
7. Furness K, Sarkies MN, Huggins CE, Croagh D, Haines TP. Impact of the method of delivering electronic health behavior change interventions in survivors of cancer on engagement, health behaviors, and health outcomes: systematic review and meta-analysis. J Med Internet Res 2020 Jun 23;22(6):e16112 [FREE Full text] [doi: 10.2196/16112] [Medline: 32574147]
8. Sarbaz M, Manouchehri Monazah F, Eslami S, Kimiafar K, Mousavi Baigi SF. Effect of mobile health interventions for side effects management in patients undergoing chemotherapy: a systematic review. Health Policy Technol 2022 Dec;11(4):100680. [doi: 10.1016/j.hlpt.2022.100680]
9. Wang L, Langlais CS, Kenfield SA, Chan JM, Graff RE, Allen IE, et al. mHealth interventions to promote a healthy diet and physical activity among cancer survivors: a systematic review of randomized controlled trials. Cancers (Basel) 2022 Aug 06;14(15):3816 [FREE Full text] [doi: 10.3390/cancers14153816] [Medline: 35954479]
10. Chow EJ, Doody DR, Di C, Armenian SH, Baker KS, Bricker JB, et al. Feasibility of a behavioral intervention using mobile health applications to reduce cardiovascular risk factors in cancer survivors: a pilot randomized controlled trial. J Cancer Surviv 2021 Aug;15(4):554-563 [FREE Full text] [doi: 10.1007/s11764-020-00949-w] [Medline: 33037989]
11. Valle CG, Heiling HM, Deal AM, Diamond MA, Hales DP, Nezami BT, et al. Examining sociodemographic and health-related characteristics as moderators of an mHealth intervention on physical activity outcomes in young adult cancer survivors. J Cancer Surviv 2025 Oct;19(5):1611-1622. [doi: 10.1007/s11764-024-01577-4] [Medline: 38607515]
12. Gregory ME, Cao W, Rahurkar S, Haroun F, Stock JC, Ghazi SM, et al. Effectiveness of mobile health for exercise promotion on cardiorespiratory fitness after a cancer diagnosis: a systematic review and meta-analysis. Cancer Med 2024 Sep;13(17):e7079 [FREE Full text] [doi: 10.1002/cam4.7079] [Medline: 39248010]
13. Howell KE, Baedke JL, Bagherzadeh F, McDonald A, Nathan PC, Ness KK, et al. Using mHealth technology to evaluate daily symptom burden among adult survivors of childhood cancer: a feasibility study. Cancers (Basel) 2024 Aug 27;16(17):2984 [FREE Full text] [doi: 10.3390/cancers16172984] [Medline: 39272842]
14. Howell KE, Shaw M, Santucci AK, Rodgers K, Ortiz Rodriguez I, Taha D, et al. Using an mHealth approach to collect patient-generated health data for predicting adverse health outcomes among adult survivors of childhood cancer. Front Oncol 2024 May 10;14:1374403 [FREE Full text] [doi: 10.3389/fonc.2024.1374403] [Medline: 38800387]
15. Liao T, Qiu L, Zhu J, Li J, Zhang Y, Yang L. A mHealth-based nursing model for assessing the health outcomes of the discharged patients with nasopharyngeal carcinoma: a pilot RCT. BMC Nurs 2022 Aug 01;21(1):210 [FREE Full text] [doi: 10.1186/s12912-022-00993-0] [Medline: 35915490]
16. Psihogios A, King-Dowling S, O'Hagan B, Darabos K, Maurer L, Young J, et al. Contextual predictors of engagement in a tailored mHealth intervention for adolescent and young adult cancer survivors. Ann Behav Med 2021 Nov 18;55(12):1220-1230 [FREE Full text] [doi: 10.1093/abm/kaab008] [Medline: 33674863]
17. Dhar E, Bah AN, Chicchi Giglioli IA, Quer S, Fernandez-Luque L, Núñez-Benjumea FJ, et al. A scoping review and a taxonomy to assess the impact of mobile apps on cancer care management. Cancers (Basel) 2023 Mar 15;15(6):1775 [FREE Full text] [doi: 10.3390/cancers15061775] [Medline: 36980661]
18. Shi J, Robinson TA, Loomba P, Murley B, Strong LL, Basen-Engquist K, et al. "Everything cannot be handled virtually": understanding intention and use of digital health technologies among rural adults and rural cancer survivors. J Rural Health 2025 Jan;41(1):e12926. [doi: 10.1111/jrh.12926] [Medline: 39878383]
19. Mahmood A, Mahmood A, Kedia S, Chang CF. Rural-urban disparities in mobile health application ownership and utilization among cancer survivors. Med Care 2025 Feb 01;63(2):111-116. [doi: 10.1097/MLR.0000000000002092] [Medline: 39791845]
20. Fareed N, Swoboda CM, Jonnalagadda P, Huerta TR. Persistent digital divide in health-related internet use among cancer survivors: findings from the Health Information National Trends Survey, 2003-2018. J Cancer Surviv 2021 Feb;15(1):87-98 [FREE Full text] [doi: 10.1007/s11764-020-00913-8] [Medline: 32671557]
21. Deng P, Fu Y, Chen M, Si L. Factors associated with health care utilization and catastrophic health expenditure among cancer patients in China: evidence from the China Health and Retirement Longitudinal Study. Front Public Health 2022 Nov 10;10:943271 [FREE Full text] [doi: 10.3389/fpubh.2022.943271] [Medline: 36438282]
22. Tune T, Goh S, Williams PA, Koczwara B. How is quality of mHealth interventions for cancer survivors defined and described? An umbrella review. JCO Clin Cancer Inform 2022 May;6:e2100203 [FREE Full text] [doi: 10.1200/CCI.21.00203] [Medline: 35623020]
23. Lee K, Kim S, Kim SH, Yoo SH, Sung JH, Oh EG, et al. Digital health interventions for adult patients with cancer evaluated in randomized controlled trials: scoping review. J Med Internet Res 2023 Jan 06;25:e38333 [FREE Full text] [doi: 10.2196/38333] [Medline: 36607712]

XSL•FO
RenderX

24. Ardito V, Golubev G, Ciani O, Tarricone R. Evaluating barriers and facilitators to the uptake of mHealth apps in cancer care using the Consolidated Framework for Implementation Research: scoping literature review. JMIR Cancer 2023 Mar 30;9:e42092 [FREE Full text] [doi: 10.2196/42092] [Medline: 36995750]
25. Monteiro-Guerra F, Signorelli GR, Rivera-Romero O, Dorronzoro-Zubiete E, Caulfield B. Breast cancer survivors' perspectives on motivational and personalization strategies in mobile app-based physical activity coaching interventions: qualitative study. JMIR Mhealth Uhealth 2020 Sep 21;8(9):e18867 [FREE Full text] [doi: 10.2196/18867] [Medline: 32955446]
26. Richards R, Kinnersley P, Brain K, Staffurth J, Wood F. The preferences of patients with cancer regarding apps to help meet their illness-related information needs: qualitative interview study. JMIR Mhealth Uhealth 2019 Jul 31;7(7):e14187 [FREE Full text] [doi: 10.2196/14187] [Medline: 31368446]
27. Budenz A, Sleight AG, Klein WM. A qualitative study of online information-seeking preferences among cancer survivors. J Cancer Surviv 2022 Aug;16(4):892-903 [FREE Full text] [doi: 10.1007/s11764-021-01082-y] [Medline: 34505207]
28. Melhem SJ, Nabhani-Gebara S, Kayyali R. Digital trends, digital literacy, and e-health engagement predictors of breast and colorectal cancer survivors: a population-based cross-sectional survey. Int J Environ Res Public Health 2023 Jan 13;20(2):1472 [FREE Full text] [doi: 10.3390/ijerph20021472] [Medline: 36674237]
29. Onyeaka HK, Zambrano J, Longley RM, Celano CM, Naslund JA, Amonoo HL. Use of digital health tools for health promotion in cancer survivors. Psychooncology 2021 Aug;30(8):1302-1310 [FREE Full text] [doi: 10.1002/pon.5677] [Medline: 33742737]
30. Armbruster C, Knaub M, Farin-Glattacker E, von der Warth R. Predictors of adherence to cancer-related mHealth apps in cancer patients undergoing oncological or follow-up treatment-a scoping review. Int J Environ Res Public Health 2022 Oct 21;19(20):13689 [FREE Full text] [doi: 10.3390/ijerph192013689] [Medline: 36294265]
31. Zhang M, Chu J, Sheng W, Bai J, Song Y, Fan L. Cancer survivors' experiences with mHealth interventions for PA: a meta-synthesis of qualitative studies. Front Public Health 2025 Dec 3;13:1715103 [FREE Full text] [doi: 10.3389/fpubh.2025.1715103] [Medline: 41415235]
32. Tarricone R, Cucciniello M, Armeni P, Petracca F, Desouza KC, Hall LK, et al. Mobile health divide between clinicians and patients in cancer care: results from a cross-sectional international survey. JMIR Mhealth Uhealth 2019 Sep 06;7(9):e13584 [FREE Full text] [doi: 10.2196/13584] [Medline: 31493318]
33. Shi N, Ching Wong AK, Yuet Wong FK, Zhang N, Zhu W, Shen K, et al. Feasibility of a mobile health app-based self-management program for Chinese patients with breast cancer receiving chemotherapy: a randomized controlled pilot study. Digit Health 2024 Feb 25;10:20552076241231560 [FREE Full text] [doi: 10.1177/20552076241231560] [Medline: 38410790]
34. Lin H, Ye M, Lin Y, Chen F, Chan S, Cai H, et al. Mobile app for gynecologic cancer support for patients with gynecologic cancer receiving chemotherapy in China: multicenter randomized controlled trial. J Med Internet Res 2023 Nov 13;25:e49939 [FREE Full text] [doi: 10.2196/49939] [Medline: 37955943]
35. Sun Q, Li D, Cai T, Luo J, Jia Q, Yu X, et al. Design and implementation of mHealth applications for older people with cancer: a scoping review. Asia Pac J Oncol Nurs 2025 May 17;12:100726 [FREE Full text] [doi: 10.1016/j.apjon.2025.100726] [Medline: 40611823]
36. Lin J, Bates SM, Allen LN, Wright M, Mao L, Kidd M. Integrating mobile health app data into electronic medical or health record systems and its impact on health care delivery and patient health outcomes: scoping review. JMIR Mhealth Uhealth 2025 Jun 23;13:e66650 [FREE Full text] [doi: 10.2196/66650] [Medline: 40549960]
37. Arioz U, Smrke U, Šafran V, Lin S, Nateqi J, Bema D, et al. Evaluating the benefits and implementation challenges of digital health interventions for improving self-efficacy and patient activation in cancer survivors: single-case experimental prospective study. Appl Sci 2025 Apr 24;15(9):4713. [doi: 10.3390/app15094713]
38. Hou SH, Henry B, Drummond R, Forbes C, Mendonça K, Wright H, et al. Co-designing priority components of an mHealth intervention to enhance follow-up care in young adult survivors of childhood cancer and health care providers: qualitative descriptive study. JMIR Cancer 2025 Apr 25;11:e57834 [FREE Full text] [doi: 10.2196/57834] [Medline: 40279633]
39. Mahalingam L, Tan BK, Bee PC, Teoh CH, Sellappans R, Ng DL, et al. Mobile health applications for supporting self-management of oral systemic anticancer therapy in haematological malignancies: a systematic review of acceptability and efficacy. Digit Health 2025 Jul 21;11:20552076251361221 [FREE Full text] [doi: 10.1177/20552076251361221] [Medline: 40718402]
40. Leidong W, Monachino M, Lloyd-Williams D, Nguyen TL, Dsouza B, Moreira JP. Mobile apps for cancer patients: identifying positive impacts and concerns. Digit Health 2025 Mar 03;11:20552076241305707 [FREE Full text] [doi: 10.1177/20552076241305707] [Medline: 40041393]
41. Wu W, Graziano T, Salner A, Chen MH, Judge MP, Cong X, et al. Acceptability, effectiveness, and roles of mHealth applications in supporting cancer pain self-management: integrative review. JMIR Mhealth Uhealth 2024 Jul 18;12:e53652 [FREE Full text] [doi: 10.2196/53652] [Medline: 39024567]
42. Alzghaibi H. Examining healthcare practitioners' perceptions of virtual physicians, mHealth applications, and barriers to adoption: insights for improving patient care and digital health integration. Int J Gen Med 2025 Apr 01;18:1865-1885 [FREE Full text] [doi: 10.2147/IJGM.S515448] [Medline: 40191233]

XSL•FO
RenderX

43. Alzghaibi H. Healthcare practitioners' perceptions of mHealth application barriers: challenges to adoption and strategies for enhancing digital health integration. Healthcare (Basel) 2025 Feb 25;13(5):494 [FREE Full text] [doi: 10.3390/healthcare13050494] [Medline: 40077056]
44. Tandon A, Dhir A, Islam N. Mobile health interventions for cancer care and support: the next level of digitalization in healthcare? IEEE Trans Eng Manag 2023 Mar 17;71:6173-6189. [doi: 10.1109/tem.2023.3243724]
45. Wang X, Wang B, Tew WY, Yang X, Xu X, Gao Y, et al. Exploring mHealth interventions for medication management: a scoping review of digital tools, implementation barriers, and patient outcomes. PeerJ Comput Sci 2025 Sep 17;11:e3190. [doi: 10.7717/peerj-cs.3190] [Medline: 40989380]

Abbreviations

MCAR: missing completely at random
mHealth: mobile health
QOL: quality of life

Edited by A Stone, N Cahill; submitted 24.Aug.2025; peer-reviewed by I Spildo; comments to author 23.Sep.2025; revised version received 03.Apr.2026; accepted 03.Apr.2026; published 23.Apr.2026.

Please cite as:
Liu T, Duo L, Zhang M, Zhang J, Zhang J, Sun X, Deng H, Cheng J, Sun Y
Use and Influencing Factors of mHealth Services Among Adult Survivors of Cancer: Cross-Sectional Survey Study
J Med Internet Res 2026;28:e82902
URL: https://www.jmir.org/2026/1/e82902
doi:10.2196/82902
PMID:

©Taiguo Liu, Lina Duo, Musi Zhang, Jianjun Zhang, Jingna Zhang, Xiao Sun, Haoyue Deng, Ji Cheng, Ya Sun. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness and Cost-Effectiveness of an App-Based Mindfulness Breast Care Program for Reducing Body Image Distress and Stigma Among Survivors of Breast Cancer: Randomized Controlled Trial

Huicong Lin[1*], MSc; Yujing Zhong[2*], MSc; Shuang Zheng[2,3*], MSc; Xiaoying Jiang[4*], BSc; Yun Xu[5,6*], MSc; Sally Chan[7], PhD; Shengjie Liu[2,8], MSc; Ka Ming Chow[9], PhD; Qingmo Yang[10], PhD; Wenhe Huang[11,12], BSc; Jiemin Zhu[13], PhD

[1]Department of Obstetrics and Gynecology, Women and Children's Hospital, School of Medicine, Xiamen University, Xiamen, China
[2]School of Medicine, Xiamen University, Xiamen, China
[3]Department of Nuclear Medicine, General Hospital of Central Theater Command, Wuhan, China
[4]School of Nursing, Fujian Medical University, Fuzhou, China
[5]School of Social and Behavioral Sciences, Nanjing University, Nanjing, China
[6]Center for Mental Health Education and Research, Nanjing University, Nanjing, China
[7]President Office, Tung Wah College, Hong Kong, China
[8]Department of Thyroid and Breast Surgery, Weifang People's Hospital, Shandong Second Medical University, Weifang, China
[9]The Nethersole School of Nursing, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China
[10]Department of Breast Surgery, The First Affiliated Hospital, Xiamen University, Xiamen, China
[11]Cancer Centre and Department of Breast and Thyroid Surgery, Xiang’an Hospital, Xiamen University, Xiamen, China
[12]Xiamen Research Centre of Clinical Medicine in Breast & Thyroid Cancers, Xiamen University, Xiamen, China
[13]Department of Nursing, Women and Children's Hospital, School of Medicine, Xiamen University, Xiamen, China
[*]these authors contributed equally

Corresponding Author:
Jiemin Zhu, PhD
Department of Nursing, Women and Children's Hospital
School of Medicine
Xiamen University
Alice North building 111, No.4221, Xiang'an South Road
Xiang'an District
Xiamen, 361102
China
Phone: 86 15960212649
Email: jieminzhu@xmu.edu.cn

Abstract

Background: Breast cancer surgery and corresponding treatments have significant residual effects on survivors of breast cancer in China. Body image distress and stigma are persistent challenges that negatively affect their quality of life. Accessible, sustainable, and cost-effective support remains scarce.

Objective: This study aimed to evaluate the effectiveness and cost-effectiveness of an app-based mindfulness breast care (MBC) program in addressing body image distress and stigma for survivors of breast cancer.

Methods: We carried out a randomized controlled trial in 2 university-affiliated hospitals in China. Survivors of breast cancer who had completed primary treatments and had mobile phone internet access were recruited and randomly assigned at a 1:1 ratio to the intervention (3-month MBC program plus routine care) or the control group (routine care alone). Under the conceptualization of mindfulness-based cognitive therapy, the MBC program was developed, including three modules: (1) Library, (2) Mindfulness Yoga, and (3) Mindfulness Practices. The primary outcomes measured were body image distress and stigma, and secondary outcomes included sleep quality, social support, and quality of life (physical and mental well-being). Assessments were conducted

XSL•FO
RenderX

at baseline, 3 months (T1), and 6 months (T2). Multiple imputation was used to handle missing data and generalized estimating equations were fitted to evaluate the effectiveness. The incremental cost per quality-adjusted life year (QALY) gained was used to measure cost-effectiveness.

Results: A total of 192 survivors of breast cancer participated in the baseline assessment, with 155 completing the 2 follow-up surveys. The median total usage duration was 199.60 (IQR 70.90-451.31; mean 360.59, SD 511.72) minutes, and total login frequency was 39.50 (IQR 19.00-86.50; mean 57.02, SD 50.26) times. The reduction in body image distress at T2 (adjusted mean difference −1.91; 95% CI −3.40 to −0.42; *P*=.01; *d*=−0.31), the reduction in stigma at T1 (adjusted mean difference −5.83; 95% CI −8.46 to −3.20; *P*<.001; *d*=−0.61) and T2 (adjusted mean difference −7.79; 95% CI −10.62 to −4.97; *P*<.001; *d*=−0.82), and the improvement in mental well-being at T1 (adjusted mean difference 4.44; 95% CI 1.70 to 7.18; *P*=.002; *d*=0.43) were statistically significantly greater in the intervention group compared with the control group. No statistically significant group differences were observed regarding sleep quality, social support, and physical well-being. The cost-effectiveness analysis showed that the intervention group gained more QALYs than the control group at T2 (adjusted mean difference 0.008; 95% CI 0.004 to 0.016; *P*=.01). The incremental cost per QALY gained at T2 was US $19,431.25, indicating a 57% probability that the MBC program is a cost-effective intervention at a threshold of US $37,530, three times the 2023 gross domestic product per capita of China.

Conclusions: An app-based MBC program was effective and potentially cost-effective and had the promise to be scalable for clinical practice.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2200059952; https://www.chictr.org.cn/showproj.html?proj=167247

(J Med Internet Res 2026;28:e85913) doi:10.2196/85913

KEYWORDS

app-based mindfulness intervention; body image distress; breast cancer; cost-effectiveness; quality of life; randomized controlled trial; stigma

Introduction

With early detections and advancements in medical technology, the number of survivors of breast cancer has been continuously increasing [1]. However, breast cancer surgery and corresponding treatments have significant residual effects on survivors of breast cancer. The entire or partial loss of breasts after surgery and adjuvant therapy–induced menopause or vaginal dryness often significantly negatively impacts women's feminine identity and body image, causing body image distress [2-4]. Body image distress is typically conceptualized as the negative mental representation of an individual's body, as well as their perceived dissatisfied state of being themselves [5]. Previous studies show that many survivors of breast cancer experience body image distress [6], even many years after recovery [7]. Sustained body image distress often leads to low self-esteem, resulting in stigma [8].

Stigma is proposed as a deeply discreditable attribute that associates an individual with an undesirable stereotype [9]. More than 80% of survivors of breast cancer in China perceive stigma [10]. Cancer-related stigma may be culturally based, especially in traditional Chinese medicine, where breast cancer is connected with some unfavorable personality traits (eg, narrow-mindedness) [10]. Some survivors of breast cancer even believe that they are to blame for their disease because of their immoral behavior or misfortune [11]. Stigma often prevents survivors of breast cancer from accepting the kindness of others and participating in social activities [12], results in poor sleep quality [13], and ultimately adversely affects their quality of life (QoL) [14].

Various approaches have been explored to alleviate body image distress and stigma among survivors of breast cancer. Psychoeducational therapies and physical activities have been proven to be effective interventions for alleviating body image distress and stigma among survivors of breast cancer [15,16]. Recently, mindfulness-based cognitive therapy (MBCT) has also emerged as an increasingly popular approach for supporting survivors of breast cancer [17,18]. MBCT applied cognitive behavioral therapy skills to practice mindfulness-based stress reduction [19]. Cognitive behavioral therapy provides psychoeducation about the interactions between thoughts, emotions, and behaviors [20], while mindfulness-based stress reduction emphasizes enhancing mindfulness skills through regular meditation practices [21]. The combination of the 2 can contribute to better control of feelings, thoughts, and behavior and ultimately to achieve mind stability [19]. MBCT guided some group-based skills training programs, involving psychoeducation, yoga, and regular mindfulness practices, to help patients deal with chronic conditions [22,23]. These programs have been proven to alleviate patients' body image distress for young patients with cancer [24], stigma for female patients with schizophrenia [25], and psychological distress for patients with cancer [26]. However, there remains a paucity of such programs specially targeting survivors of breast cancer to address their unique psychological challenges in relation to body image distress and stigma.

Further, traditional face-to-face programs delivering MBCT are time-intensive for patients and resource-intensive for health care providers [27]. With the advantage of a broad reach, the internet may provide a convenient and cost-effective platform. Several internet-based programs have been developed to offer MBCT [27,28], but few such programs include health economic evaluation [29]. Compen et al [30] demonstrated the potential cost-effectiveness of internet-based individual MBCT in alleviating psychological distress in patients with cancer. However, this internet-based program was delivered individually and recruited patients with various types of cancers [30]. The

XSL•FO
RenderX

cost-effectiveness of internet-based group MBCT remains unknown for survivors of breast cancer.

In 2023, nearly 1.079 billion individuals in China could surf the internet through their mobile phones [31]. WeChat, the most popular free mobile social platform in China, offers an easily accessible channel to deliver support and service to survivors of breast cancer [32]. Under the conceptualization of MBCT, we developed an app-based mindfulness breast care (MBC) program on the WeChat platform. We hypothesized that the MBC program (lasting 3 months) plus routine care (the intervention group) would outperform routine care alone (the control group) in reducing body image distress and stigma, while enhancing sleep quality, social support, and QoL. Additionally, the MBC program would be cost-effective.

Methods

Study Design and Participants

We conducted a randomized controlled trial (RCT; Multimedia Appendices 1 and 2). The trial was registered with the Chinese Clinical Trial Registry (trial number ChiCTR2200059952; date of registration: May 14, 2022).

Eligible participants were women (1) with breast cancer who had completed primary treatments (surgical operation, chemotherapy, and/or radiotherapy) and were ready to be discharged, (2) who had mobile phone internet access, and (3) who were literate in Mandarin. Participants diagnosed with other concurrent cancers (such as ovarian cancer) or psychiatric disorders were excluded.

We recruited participants from 2 university-affiliated hospitals. One hospital is located in a central city and performs 400-600 breast surgeries each year. The other hospital is located in a rural area and has about 100 breast surgeries annually. Of these, approximately 50% of survivors of breast cancer met the aforementioned inclusion criteria. After primary treatments, survivors of breast cancer were often followed up every 3 months after discharge. There were no other app-based programs to support survivors of breast cancer at the 2 participating hospitals during the study period.

Procedures

The breast surgeons introduced the MBC program to eligible women in the Breast Cancer Unit. If women showed interest, the researchers approached them, confirmed their eligibility, and explained the program in detail. Eligible participants signed the consent form, completed a paper-based baseline assessment (T0), and were randomly assigned to different groups.

Participants in the intervention group scanned the Quick Scan QR code, downloaded the app of the MBC program from the WeChat platform, and used their accounts (phone numbers) and passports (changeable later) to log in to the program. Afterwards, trained researchers provided approximately 20-minute onsite program training, including reading contents in the Library module, watching videos in the Mindfulness Yoga module, and participating in videoconferences in the Mindfulness Practices module. The usernames expired 3 months after activation.

All participants were invited to complete follow-up paper questionnaires at 3 months (T1) and 6 months (T2). These evaluation time points were chosen because most internet-based interventions documented the greatest benefits within 3 months [33], and some benefits might be endured at 6 months [34]. If some participants did not come back to their hospitals at either follow-up time point to complete the paper questionnaires, the researchers sent the online survey with the same questionnaires via the “Questionnaire Star” platform, which had a feature that only when all questions were answered, the online questionnaire was able to be submitted. If several scales were ticked with the same answers consecutively, the questionnaires were returned to participants to double-check or refill. Participants received an appreciation gift of approximately US $3 after each evaluation (direct delivery of a gift after completing the paper questionnaires and WeChat red packet after completing the online questionnaires).

Randomization and Masking

Internet Research Randomizer [35], with selected block sizes of 4, 6, and 8, was used to generate the group allocation. Participants were randomly assigned at a 1:1 ratio to the intervention (MBC program plus routine care) or the control group (routine care alone) according to sequential participants’ recruitment. Neither the data collectors nor the participants were blinded to group allocation.

MBC Program

A multidisciplinary health care team and a Chinese IT company developed the MBC program. The development of the MBC program has been published [36]. A previous similar study on an internet-based MBCT program indicated that a 3-month intervention period was sufficient to improve psychological outcomes [37]. Thus, a 3-month MBC program was designed.

The MBC program was based on MBCT. MBCT highlights the awareness of current feelings and thoughts, observes these with a nonjudgmental attitude, and ultimately develops individuals’ ability to achieve mind stability and live in the present (Figure 1) [19].

XSL•FO
RenderX

Figure 1. Framework of an app-based mindfulness breast care program for addressing body image distress and stigma.

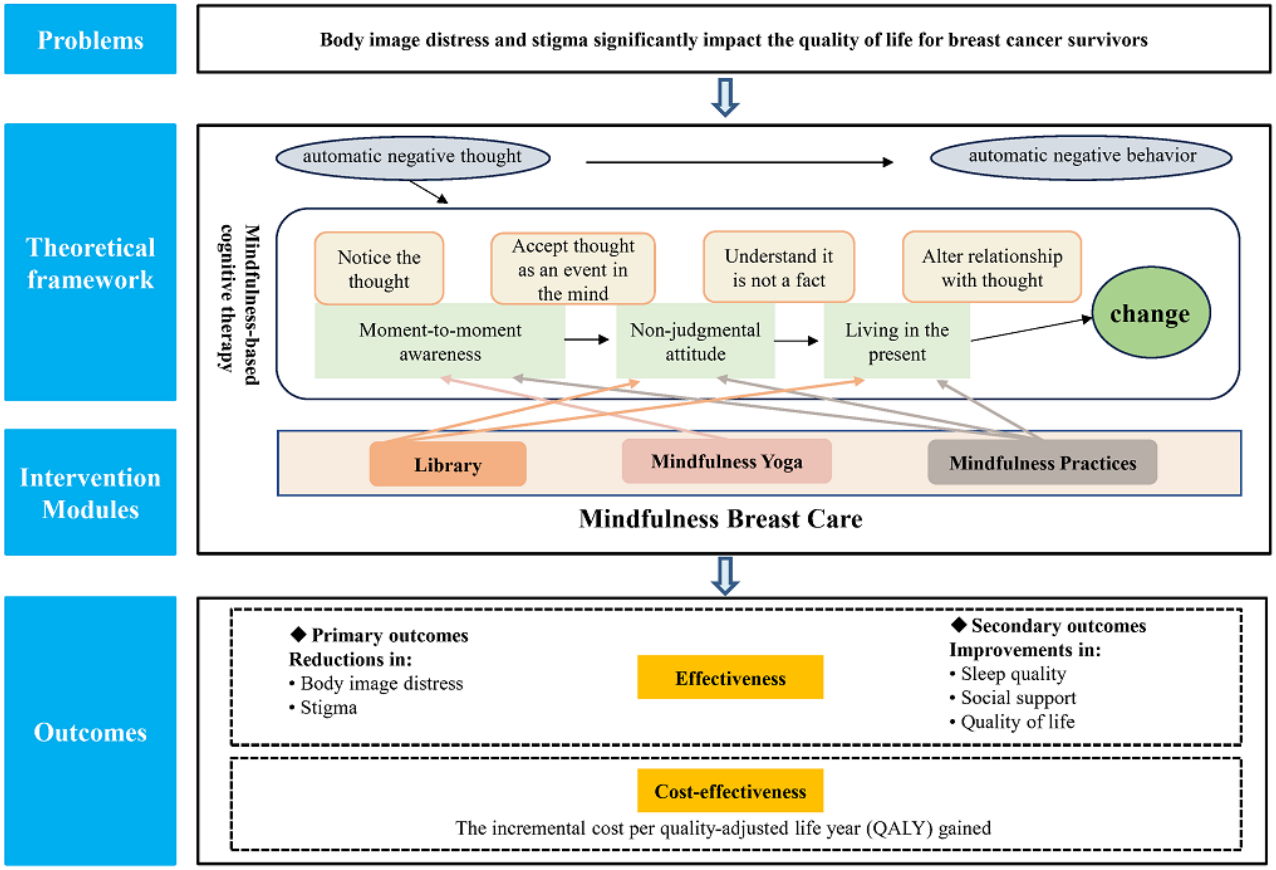

This program included three modules: (1) Library, (2) Mindfulness Yoga, and (3) Mindfulness Practices (Figure 2). The Library module provided evidence-based educational information on breast cancer, the source of stigma, altered body image, strategies to manage stigma and body image distress, mindfulness, healthy lifestyle, and recovery [38-40].

XSL•FO
RenderX

Figure 2. Screenshots of an app-based mindfulness breast care program.

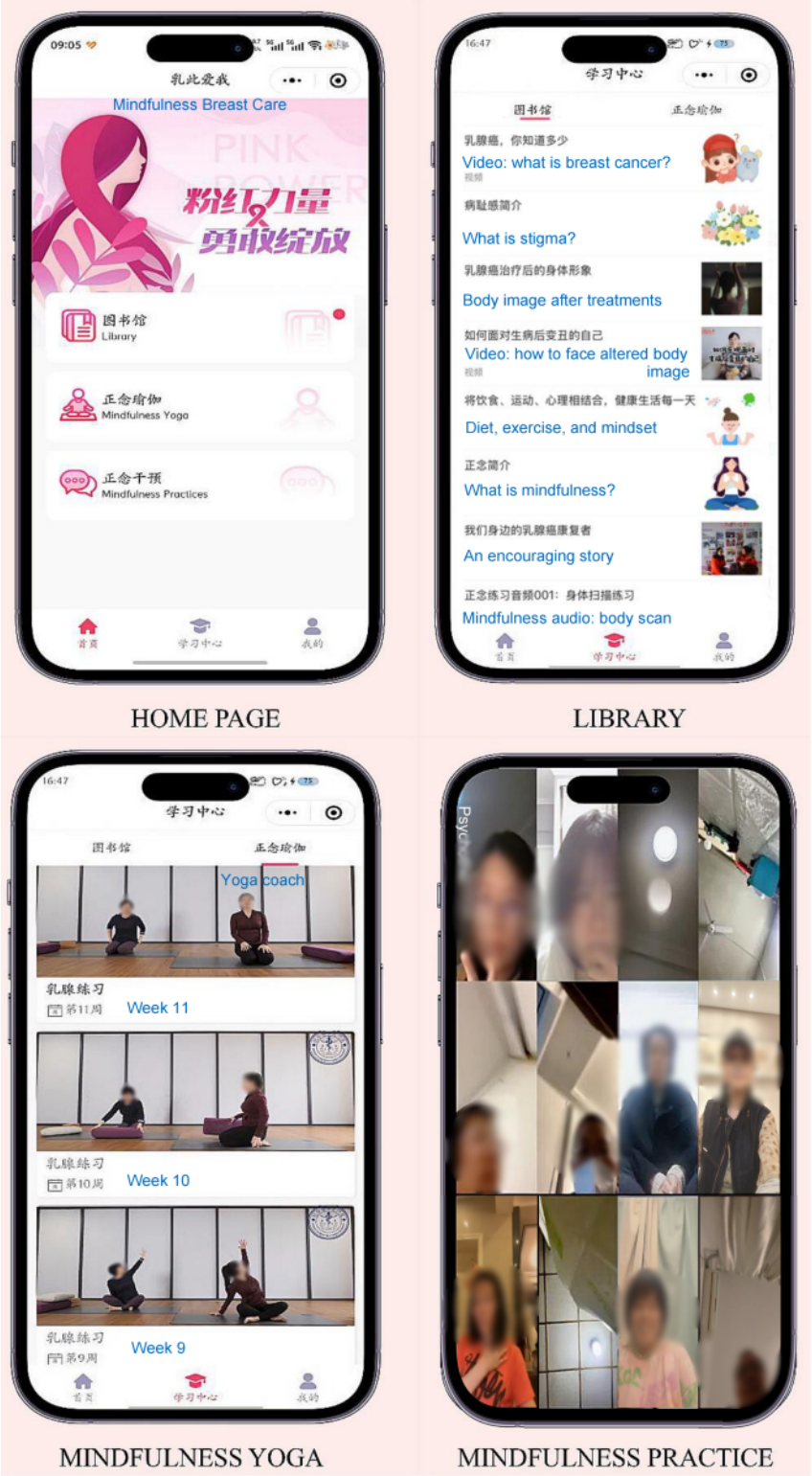

The Mindfulness Yoga module consisted of 12 short videos of Hatha yoga, following the principles of MBCT, which incorporated mindful awareness into physical postures and breathing exercises [39]. During mindfulness yoga practice, women were guided to observe their bodies and be aware of their thoughts instead of judging their bodily sensations.

The Mindfulness Practices module was offered in 12 weekly 45-minute videoconferences. In a prior face-to-face MBCT intervention, participants were placed in groups of 8-12 participants and 1 certified mindfulness psychotherapist per subgroup [41]. Thus, according to the recruitment sequence, approximately 10 participants were allocated to a corresponding subgroup with 1 psychotherapist for better communication. Based on the Mindful Way Workbook [42], full communication between the 2 certified psychotherapists was conducted to guarantee standardized mindfulness practices for each group. Mindfulness practices involved three main components (1) awareness-guided meditation to focus on bodily and breathing

XSL•FO
RenderX

sensations, such as sitting meditation; (2) connection to the present moment by observing somatic sensations and movements through practices like a guided body scan; and (3) development of a nonjudgmental attitude by observing mental activity, including emotions, impulses, and thought patterns, through practices such as 2 ways of knowing.

Various strategies were adopted to enhance the MBC program fidelity, including keeping the intervention consistent and enhancing participant engagement. To keep intervention consistency, the researchers ensured the standard delivery of mindfulness practice through full communication between 2 psychotherapists. To increase participants' engagement, the content in the Library and Mindfulness Yoga modules was updated each week, with an instant WeChat messaging reminder of the incoming new information and a new Yoga video sent to the participants.

The quality of the MBC program was assessed using the Mobile App Rating Scale (user version) among 5 survivors of breast cancer for pilot feedback [43], with an overall mean score of 3.96 out of 5, indicating that the program is generally credible, enjoyable, usable, and highly relevant to survivors of breast cancer's medical conditions.

Routine Care

Participants in 2 groups received routine care. Before discharge, nurses offer oral and written instructions on routine information after cancer treatment and follow-up education, including exercises to prevent upper limb lymphedema, infection control, nutrition, healthy lifestyle, and information on medical follow-up. At every follow-up visit, oncologists conduct medical assessments and adjust medications as needed. Nurses give general advice on monitoring and managing symptoms post treatment.

Measures

Sample Characteristic Measures

A self-designed questionnaire was applied to collect participants' baseline sociodemographic and clinical characteristics.

Primary Outcome Measures

The primary outcomes included body image distress and stigma. Body image distress was evaluated using the 10-item Body Image Scale (BIS) [44]. A higher total score on the BIS (range 0-30) indicates a greater level of body image distress. The Chinese version of the BIS has good reliability in individuals with breast cancer [45]. The baseline Cronbach α was 0.94 in this study.

Stigma was evaluated using the 24-item Social Impact Scale (SIS) [9]. A higher total score on the SIS (range 24-96) indicates a greater perceived stigma. The Chinese version of the SIS has good reliability in individuals with breast cancer [10]. The baseline Cronbach α was 0.85 in this study.

Secondary Outcome Measures

The secondary outcomes included sleep quality, social support, and QoL, which were evaluated using the Pittsburgh Sleep Quality Index (PSQI) [46], Multidimensional Scale of Perceived Social Support (MSPSS) [47], and the 36-item Short-Form Health Survey (SF-36) [48]. The SF-36 is summarized into 2 component scores: the Physical Component Summary (PCS) and the Mental Component Summary (MCS). Higher total scores on the aforementioned questionnaires (PSQI range 0-21; MSPSS range 12-84; and SF-36 PCS/MCS range 0-100) indicate worse sleep quality, better social support, and better physical and mental well-being, respectively. These questionnaires have been validated in Chinese individuals with breast cancer [49-51]. In this study, the baseline Cronbach α values of PSQI, MSPSS, and SF-36 were 0.74, 0.94, and 0.91, respectively.

Economic Evaluation Measures

The costs were considered from a societal perspective [52], including (1) fixed costs of the development of the app-based MBC program; (2) direct medical costs; and (3) indirect costs, namely income and time lost due to medical treatment [53]. Furthermore, the scores of the SF-36 were converted into a health utility value based on the Short-Form Six-Dimension (SF-6D) to calculate the quality-adjusted life years (QALYs) [54].

Usage Data

The background thread of the MBC program tracked and recorded 3-month usage data, which were measured as login frequency (number of times a participant logged in the program) and usage duration (total time from login to logout) of the entire program and each module. If one participant's mobile phone was in standby mode or one participant was browsing other websites together, the MBC program ran in the background operational mode and it stopped calculating the participant's usage data. Based on a prior web-based intervention for survivors of breast cancer [55], the minimal login frequency and minimal usage duration of expected use were defined as 10 times and 50 minutes during the whole course of the MBC program.

Sample Size

The primary outcomes of body image distress and stigma were used to estimate effect size. Previous literature reported an effect size of 0.28 for body image distress [56] and 0.36 for stigma among patients with cancer [57]. Thus, the effect size of 0.28 was chosen to calculate the sample size. Using G*Power (version 3.1.9.7), a total of 154 participants (77 each group) were needed for 90% power at a significance level of .05 (2-sided). Accounting for a 20% dropout rate, a target sample size of 192 (96 per group) was determined.

Statistical Analyses

Overview

Statistical analysis was performed using SPSS (version 26; IBM Corp) and R (version 4.3.3; R Foundation for Statistical Computing). Exploration of the missing-data structure indicated that the missingness was consistent with a missing-at-random mechanism. Therefore, to ensure robustness of parameter estimation, multiple imputation was conducted to handle missing data. Baseline comparisons were performed using the chi-square test or Fisher exact test for categorical variables, as well as independent-samples 2-sided *t* tests for continuous variables.

XSL•FO
RenderX

Effectiveness

Generalized estimating equations were applied to analyze repeated measurements at baseline, 3 months, and 6 months with group, time, and the group × time interaction as fixed effects. Demographic and clinical characteristics that significantly differed between groups (endocrine therapy) and baseline outcomes were adjusted during data analysis. An exchangeable working correlation structure was specified to account for within-subject correlation scores across repeated assessments, and robust standard errors were used to obtain valid inference even if the correlation structure was not specified. *P* values were adjusted with the Bonferroni correction to reduce the risk of inflated type I error. The adjusted estimated mean differences, 95% CIs, significance levels, and adjusted effect sizes were reported. The adjusted effect size (Cohen *d*) was calculated as the adjusted mean difference divided by the pooled unadjusted SD at baseline [58]. The minimally important clinical differences of the primary outcomes were evaluated as one-half the SD of the baseline score [59] using the distribution-based method [60]. To examine the robustness of the results regarding missing values for the primary outcomes, a sensitivity analysis was conducted with Markov chain Monte Carlo simulation imputation.

Cost-Effectiveness

An economic evaluation was performed to explore the cost-effectiveness of the MBC program in terms of incremental cost per QALY gained [61]. QALYs gained were calculated from health utility scores [54] using the area-under-the-curve method [62]. Total costs were calculated as the sum of the following three types of costs: (1) fixed costs associated with the app development, (2) direct medical costs, and (3) indirect costs.

Fixed costs included (1) costs for app development and maintenance, (2) costs of video production, and (3) consultation fees charged by mindfulness therapists (approximately US $44.78 per hour). According to the CHEERS (Consolidated Health Economic Evaluation Reporting Standards) statement (Multimedia Appendix 2) [52], the app was assumed to remain operational for 5 years. Fixed costs were amortized using a 3% discount rate with the following calculation formula [63] and then converted into 6-month average costs.

Consistent with standard trial-based economic evaluation methods [64], direct costs were extracted from medical records strictly within the predefined 6-month study for each participant. Indirect costs were calculated as the average labor income multiplied by the time spent in the hospital from T0 to T2, including inpatient and outpatient days (times of clinic visits × 0.5 day per visit) [65] following the human capital approach [66]. According to 2023 statistics, the per capita annual income in Xiamen city was US $10,606.27, yielding a daily average income of US $40.80 [67]. All costs were converted to 2023 US dollars using the mean exchange rate averaged across 2022 and 2023 provided by the Bank of China foreign exchange rate.

Generalized estimating equations were fitted to evaluate the direct and indirect costs (adjusted for endocrine therapy) and the QALYs gained (adjusted for endocrine therapy and baseline health utility scores). Total costs and QALYs gained were used to estimate the incremental cost-effectiveness ratio (ICER) and net monetary benefit. According to the recommendation of the World Health Organization [68], 3 times the 2023 gross domestic product (GDP) per capita (US $12,510) was applied as the threshold [69]. The interpretation was as follows: if the ICER is less than 3 times the GDP (US $37,530), it is deemed cost-effective [68].

Sensitivity analysis was used to determine how the ICER was influenced by altering specific assumptions or parameters. The 2-tailed *t* tests used a bias-corrected and accelerated nonparametric bootstrapping method with 10,000 replications to estimate the 95% CI of incremental costs and QALYs at T2 [70]. The cost-effectiveness plane was graphically represented to illustrate the uncertainty surrounding the ICER at the threshold. Cost-effectiveness acceptability curves were used to display the net monetary benefits of the MBC program at various willingness-to-pay thresholds, ranging from US $0 to US $200,000 per QALY gained [71].

Usage Data and Its Relationships With Measurement Outcomes

Due to the highly skewed nature, the usage duration (minutes) and login frequency of the program were presented as median, mean, and maximum values. Spearman rank correlation was used to calculate the associations between MBC usage data and measurement outcomes at baseline, T1, and T2.

Ethical Considerations

The trial was conducted in accordance with the Declaration of Helsinki. This study received approval from the ethics committees of First Affiliated Hospital, Xiamen University (approval number FA-2022-30) and Xiang'an Hospital, Xiamen University (approval number XAHLL2022027). All participants were assured of voluntary participation and provided written informed consent. All collected data regarding participants were confidential and anonymous. Participants received compensation (approximately US $3) each time they returned their questionnaires. The patients provided written informed consent to allow their image to be published.

Results

Sample Characteristics

We approached a total of 202 participants for eligibility assessment, and 192 (95.2%) completed baseline assessments. A total of 155 participants completed all follow-up assessments at T1 and T2, yielding a response rate of 80.7% (155/192; Figure 3). The participants (mean age 45.59, SD 8.74 years) were primarily diagnosed with stage II breast cancer (104/192, 54.2%). Most participants had an education at the senior high school level or above (114/192, 59.4%). Almost all participants (187/192, 97.4%) were diagnosed with breast cancer within 2 years. All participants (192/192, 100%) had surgical treatment and chemotherapy. Most participants underwent axillary lymph node dissection (153/192, 79.7%), endocrine therapy (114/192,

XSL•FO
RenderX

59.4%), targeted therapy (104/192, 54.2%), and radiotherapy (103/192, 53.6%). No group differences were found in baseline characteristics except for endocrine therapy (*P*=.04; Table 1).

Figure 3. Trial profile (CONSORT flow diagram). CONSORT: Consolidated Standards of Reporting Trials; MBC: mindfulness breast care; RC: routine care.

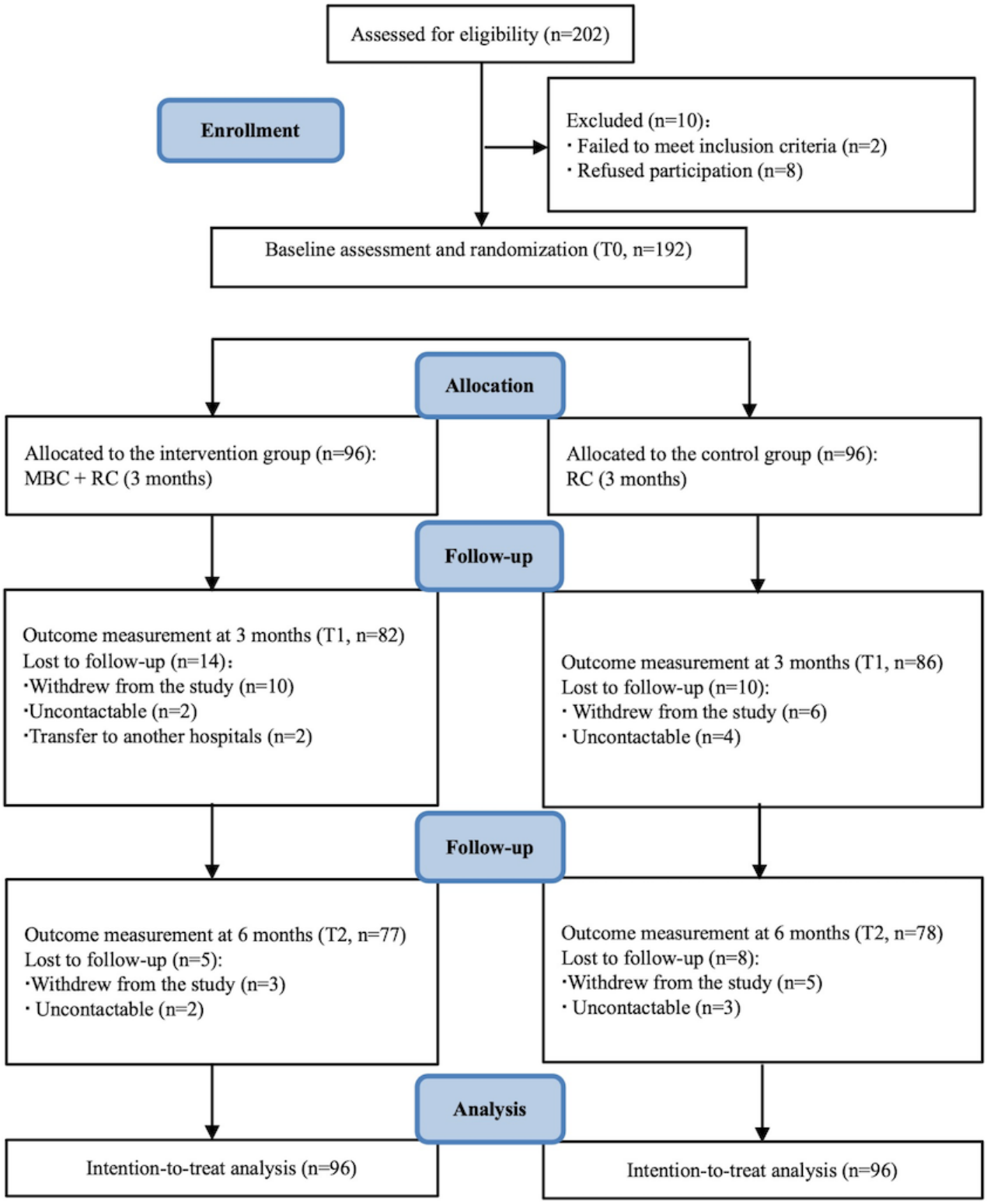

XSL•FO
RenderX

Table 1. Participants' characteristics and baseline health outcomes.

Variables	Overall (n=192)	Intervention (n=96)	Control (n=96)
Age (years), mean (SD)	45.59 (8.74)	44.64 (8.92)	46.55 (8.49)
Marital status, n (%)			
Married or cohabitation	175 (91.1)	88 (91.7)	87 (90.6)
Separated or divorced or single	17 (8.9)	8 (8.3)	9 (9.4)
Education, n (%)			
Junior high school or lower	78 (40.6)	37 (38.5)	41 (42.7)
Senior high school or higher	114 (59.4)	59 (61.5)	55 (57.3)
Current employment, n (%)			
Employed	101 (52.6)	54 (56.3)	47 (49)
Unemployed or retired	91 (47.4)	42 (43.8)	49 (51)
Monthly family income (US $), n (%)			
<1492.54	144 (75)	73 (76)	71 (74)
≥1492.54	48 (25)	23 (24)	25 (26)
Payment methods, n (%)			
Medical insurance	179 (93.2)	87 (90.6)	179 (93.2)
Out of pocket money	13 (6.8)	9 (9.4)	13 (6.8)
Stage of breast cancer, n (%)			
I	46 (24)	24 (25)	22 (22.9)
II	104 (54.2)	55 (57.3)	49 (51)
III	42 (21.9)	17 (17.7)	25 (26)
Time form diagnosis, n (%)			
T ≤ 2 years	186 (97.4)	94 (97.9)	93 (96.9)
2< T ≤ 5 years	5 (2.6)	2 (2.1)	3 (3.1)
Surgery, n (%)			
Yes	192 (100)	96 (100)	96 (100)
No	0 (0)	0 (0)	0 (0)
Surgical methods, n (%)			
Breast-conserving surgery	40 (20.8)	19 (19.8)	21 (21.9)
Mastectomy	83 (43.2)	41 (42.7)	42 (43.8)
Mastectomy + breast reconstruction	69 (35.9)	36 (37.5)	33 (34.4)
Axillary lymph node dissection, n (%)			
Yes	153 (79.7)	76 (79.2)	77 (80.2)
No	39 (20.3)	20 (20.8)	19 (19.8)
Chemotherapy, n (%)			
Yes	192 (100)	96 (100)	96 (100)
No	0 (0)	0 (0)	0 (0)
Radiotherapy, n (%)			
Yes	103 (53.6)	49 (51.0)	54 (56.3)
No	89 (46.4)	47 (49.0)	42 (43.8)
Endocrine therapy, n (%)			
Yes	114 (59.4)	50 (52.1)	64 (66.7)
No	78 (40.6)	46 (47.9)	32 (33.3)

XSL•FO
RenderX

Variables	Overall (n=192)	Intervention (n=96)	Control (n=96)
Targeted therapy, n (%)			
Yes	104 (54.2)	54 (56.3)	50 (52.1)
No	88 (45.8)	42 (43.8)	46 (47.9)
Traditional Chinese medicine treatment, n (%)			
Yes	35 (18.2)	21 (21.9)	14 (14.6)
No	157 (81.8)	75 (78.1)	82 (85.4)
Health outcomes, mean (SD)			
Body image distress (BIS[a] scores)	6.71 (6.09)	6.76 (5.83)	6.67 (6.37)
Stigma (SIS[b] scores)	52.58 (9.50)	52.88 (9.09)	52.29 (9.93)
Sleep quality (PSQI[c] scores)	7.54 (3.88)	7.65 (3.69)	7.43 (4.08)
Social support (MSPSS[d] scores)	66.04 (9.34)	65.88 (8.54)	66.21 (10.12)
Physical well-being (SF-36[e] PCS[f] scores)	36.01 (10.38)	36.54 (9.12)	35.47 (11.53)
Mental well-being (SF-36 MCS[g] scores)	40.76 (10.37)	40.94 (10.27)	40.57 (10.51)

[a]BIS: Body Image Scale.
[b]SIS: Social Impact Scale.
[c]PSQI: the Pittsburgh Sleep Quality Index.
[d]MSPSS: Multidimensional Scale of Perceived Social Support.
[e]SF-36: 36-item Short-Form Health Survey.
[f]PCS: the Physical Component Summary.
[g]MCS: the Mental Component Summary.

Effectiveness

The intervention group showed a statistically significantly decrease in body image distress at T2 (adjusted mean difference, −1.91; 95% CI, −3.40 to −0.42; P=.01, d=−0.31), decrease in stigma at T1 (adjusted mean difference, −5.83; 95% CI, −8.46 to −3.20; P<.001, d=−0.61) and T2 (adjusted mean difference, −7.79; 95% CI, −10.62 to −4.97; P<.001, d=−0.82), and improvement in mental well-being at T1 (adjusted mean difference, 4.44; 95% CI, 1.70 to 7.18; P=.002, d=0.43) when compared to the control group (Table 2). No statistically significant group differences were observed regarding sleep quality, social support, and physical well-being. In our trial, only the between-group adjusted mean difference in stigma at T1 and T2 exceeded the minimally important clinical differences of stigma (the one-half SD of the baseline score). Sensitivity analysis yielded similar findings regarding the primary outcomes, which confirmed the robustness of the results (Table 3).

XSL•FO
RenderX

Table 2. Effects of mindfulness breast care program on health outcomes (N=192).

Outcomes[a]	Intervention group mean (SE)	Control group mean (SE)	Adjusted mean difference (95% CI)	*P* value[b]	Cohen *d*[c]
Primary outcomes					
Body image distress (BIS scores)[d,e]					
T1	6.33 (0.52)	7.32 (0.52)	−0.99 (−2.43 to 0.46)	.18	−0.16
T2	6.32 (0.45)	8.23 (0.60)	−1.91 (−3.40 to −0.42)	.01	−0.31
Stigma (SIS scores)[e,f]					
T1	50.50 (0.99)	56.33 (1.01)	−5.83 (−8.46 to −3.20)	<.001	−0.61
T2	48.13 (0.99)	55.92 (1.01)	−7.79 (−10.62 to −4.97)	<.001	−0.82
Secondary outcomes					
Sleep Quality (PSQI scores)[e,g]					
T1	7.09 (0.33)	7.29 (0.35)	−0.20 (−1.14 to 0.73)	.67	−0.05
T2	7.11 (0.34)	7.03 (0.34)	0.07 (−0.86 to 1.02)	.87	0.02
Social support (MSPSS scores)[h,i]					
T1	64.32 (1.14)	65.76 (1.12)	−1.44 (−4.63 to 1.74)	.37	−0.15
T2	64.72 (1.03)	61.91 (1.04)	2.71 (−0.15 to 5.58)	.06	0.29
Physical well-being (SF-36 PCS scores)[i,j,k]					
T1	39.69 (0.90)	38.11 (0.98)	1.58 (−1.03 to 4.19)	.24	0.15
T2	41.74 (0.95)	39.72 (0.92)	2.02 (−0.60 to 4.64)	.13	0.19
Mental well-being (SF-36 MCS scores)[i,l]					
T1	46.23 (1.02)	41.79 (0.95)	4.44 (1.70 to 7.18)	.002	0.43
T2	44.36 (0.96)	42.22 (0.94)	2.14 (−0.49 to 4.77)	.11	0.21

[a]The intention-to-treat analysis was conducted using the multiple imputation method. Adjusted for endocrine therapy and baseline outcomes, generalized estimating equations were used to examine the effectiveness of the MBC program. Pairwise comparisons were conducted with Bonferroni adjustment.

[b]All *P* values for pairwise comparisons were Bonferroni-adjusted.

[c]Cohen *d* was calculated by dividing the adjusted mean difference by the pooled unadjusted standard deviation at baseline [58].

[d]BIS: Body Image Scale.

[e]An increase implies worsening and a decrease implies improvement.

[f]SIS: Social Impact Scale.

[g]PSQI: the Pittsburgh Sleep Quality Index.

[h]MSPSS: Multidimensional Scale of Perceived Social Support.

[i]An increase implies improvement, and a decrease implies worsening.

[j]SF-36: 36-item Short-Form Health Survey.

[k]PCS: the Physical Component Summary.

[l]MCS: the Mental Component Summary.

XSL•FO
RenderX

Table 3. Sensitivity analysis of primary outcomes.

Primary outcomes[a]	Intervention group mean (SE)	Control group mean (SE)	Adjusted mean difference (95% CI)	*P* value[b]	Cohen *d*[c]
Body image distress (BIS scores)[d,e]					
T1	4.41 (1.10)	5.54 (1.10)	−1.13 (−2.74 to 0.46)	.10	−0.19
T2	4.62 (1.14)	6.27 (1.16)	−1.65 (−2.97 to −0.18)	.02	−0.27
Stigma (SIS scores)[d,f]					
T1	47.02 (2.10)	52.80 (2.14)	−5.78 (−8.79 to −3.26)	<.001	−0.61
T2	44.99 (2.14)	50.61 (2.18)	−5.62 (−8.18 to −3.20)	<.001	−0.59

[a]The intention-to-treat analysis was conducted using the Markov chain Monte Carlo simulation imputation. Generalized estimating equations were used to examine the effectiveness of the MBC program on primary outcomes. Pairwise comparisons were conducted with Bonferroni adjustment.

[b]All *P* values for pairwise comparisons were Bonferroni-adjusted.

[c]Cohen *d* was calculated by dividing the adjusted mean difference by the pooled unadjusted standard deviation at baseline [58].

[d]An increase implies worsening and a decrease implies improvement.

[e]BIS: Body Image Scale.

[f]SIS: Social Impact Scale.

At T1, a statistically significant time effect was found in stigma (T1:=−2.39; *P*=.001). At T2, statistically significant time effects were detected in social support (=−5.83; *P*=.004). Statistically significant time group interaction effects were identified for stigma at T1 and T2 (Group T1: =−5.89; *P*=.009; Group T2: =−6.64; *P*=.005) (Figure 4).

XSL•FO
RenderX

Figure 4. Adjusted mean values of the primary and secondary outcomes between two groups from T0 to T2. SIS: Social Impact Scale; BIS: Body Image Scale; PSQI: Pittsburgh Sleep Quality Index; MSPSS: Multidimensional Scale of Perceived Social Support; SF-36: Short Form-36 items; PCS: Physical Component Summary; MCS: Mental Component Summary. a An increase implies worsening and a decrease implies improvement. b An increase implies improvement and a decrease implies worsening.

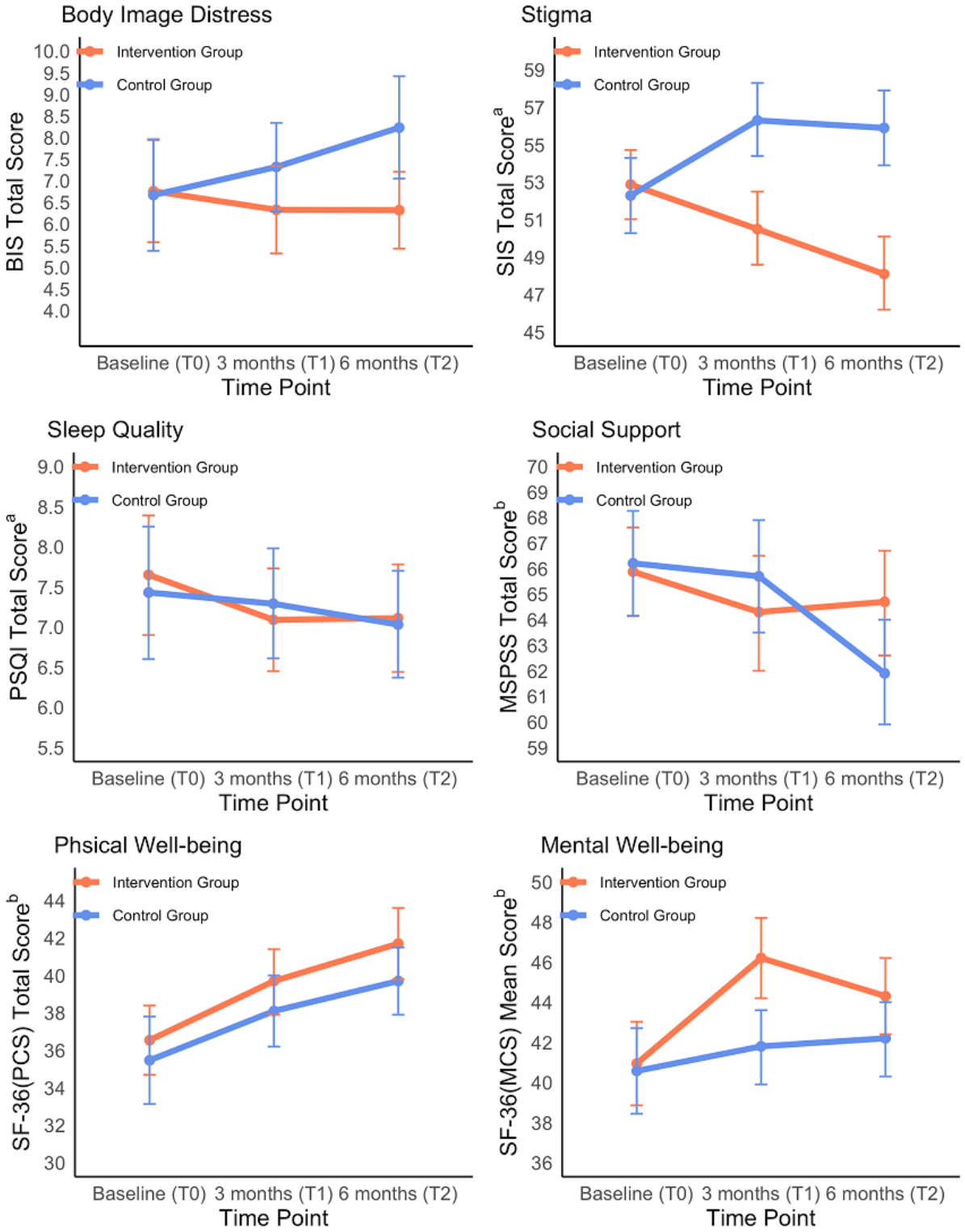

Cost-Effectiveness

Fixed costs primarily comprised the following: (1) app costs totaling US $15,014.92, (2) video production amounting to US $746.27, and (3) first-year consultation fees charged by mindfulness therapists equaling US $4298.51. The total fixed cost over 5 years was US $36,037.68, with US $37.54 per participant for six months. There is no difference in adjusted direct medical costs and indirect costs between two groups at T2. The total costs were estimated at US $9295.45 for the intervention group and US $9140.00 for the control group with no statistically significant group difference (*P*=.92). The intervention group gained more QALYs than the control group at T2 (adjusted mean difference 0.008, 95% CI 0.004 to 0.016;

XSL•FO
RenderX

P=.01). The incremental cost per QALY gained at T2 was US $19,431.25, making our program cost-effective at a threshold of three times of China's 2023 GDP per capita (US $37,530; Table 4).

Table 4. Cost-effectiveness analysis for cost, health utility score, and quality adjusted life years gained by study time-point.

Variables[a]	Intervention group mean (SE)	Control group mean (SE)	Adjusted mean difference (95% CI)	*P* value
Cost (US $), baseline to 6 months				
Fixed costs	37.54	0	__[b]	—
Direct medical costs	9113.58 (1034.33)	9002.69 (1022.39)	111.04 (−2803.43 to 3025.37)	.94
Indirect costs	144.33 (8.43)	137.31 (8.30)	7.03 (−16.10 to 30.15)	.55
Total costs for 6 months	9295.45	9140.00	155.45 (−2773.28 to 3084.48)	.92
SF-6D[b] health utility score				
SF-6D value at T0	0.663 (0.014)	0.663 (0.014)	0.0001 (−0.019 to 0.019)	.997
SF-6D value at T1	0.685 (0.014)	0.667 (0.014)	0.018 (−0.004 to 0.039)	.10
SF-6D value at T2	0.681 (0.014)	0.674 (0.014)	0.007 (−0.013 to 0.021)	.06
QALYs[c] gained, baseline to 6 months				
QALYs (T0 to T1)	0.189 (0.002)	0.187 (0.002)	0.002 (–0.003 to 0.006)	.47
QALYs (T0 to T2)	0.377 (0.003)	0.369 (0.003)	0.008 (0.004 to 0.016)	.01

[a]The data were analyzed using generalized estimating equations. The scores of the SF-36 were converted into a health utility value based on the SF-6D to calculate QALYs [54]. Costs were adjusted for endocrine therapy. The QALYs gained were adjusted for endocrine therapy and the SF-6D utility score at T0.

[b]Not applicable.

[c]SF-6D: Short Form Six-Dimension.

[d]QALYs: quality adjusted life years.

The robustness of this finding was confirmed by the sensitivity analyses. The cost-effectiveness plane showed the ICERs of QALYs gained were predominantly located in the southeast and northeast quadrants, indicating that the cost-effectiveness of these outcomes may be dominant or can be evaluated by comparison with the threshold (US $37,530; Figure 5). The cost-effectiveness acceptability curve predicted that our program had a 57% probability of being cost-effective in QALYs gained (Figure 6).

XSL•FO
RenderX

Figure 5. Cost-effectiveness plane of QALYs gained comparing intervention group to control group at T2. QALYs: quality-adjusted life years; MBC: mindfulness breast care.

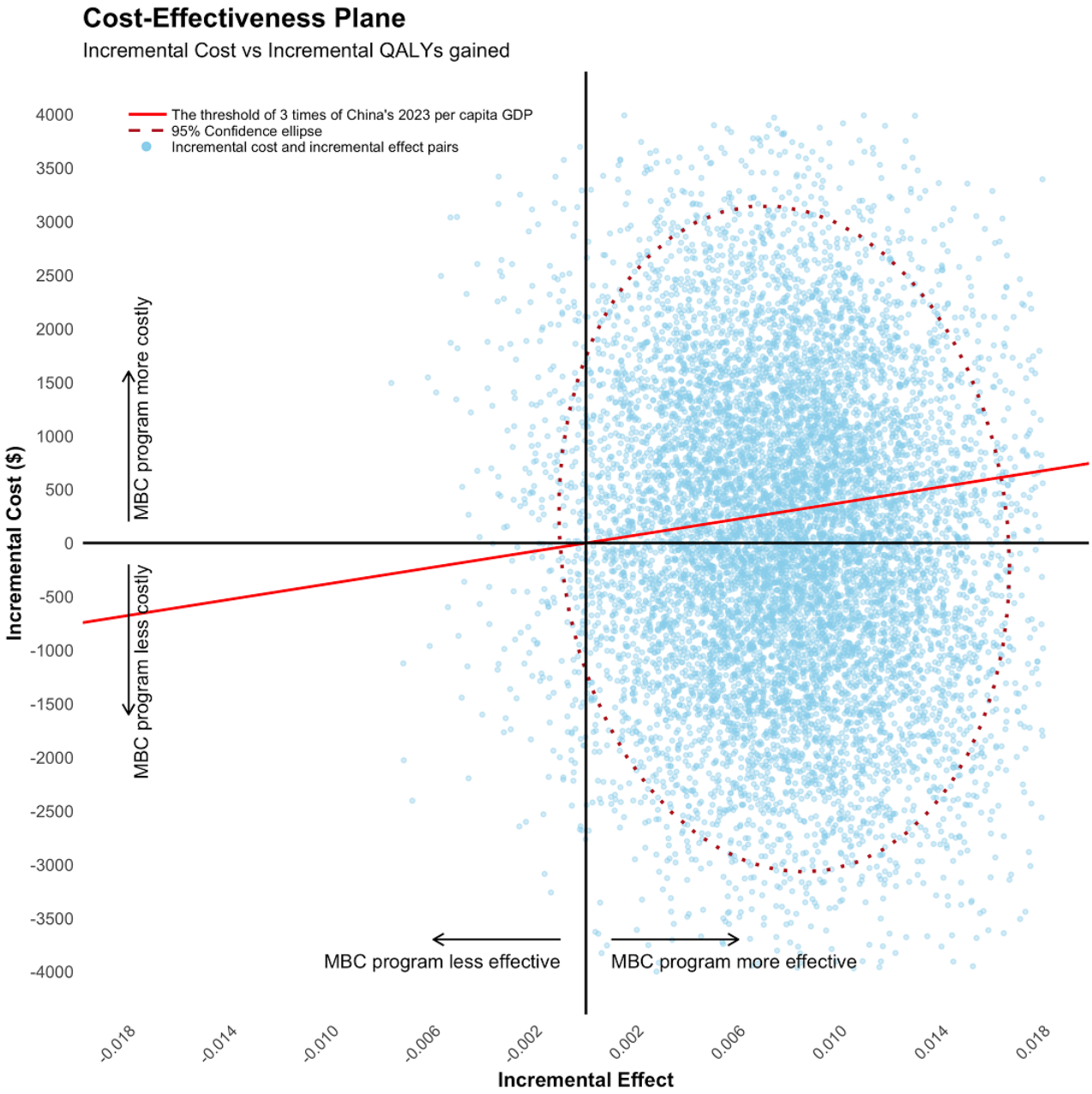

XSL•FO
RenderX

Figure 6. Cost-effectiveness acceptability curve of QALYs gained comparing intervention group to control group at T2. QALYs: quality-adjusted life years.

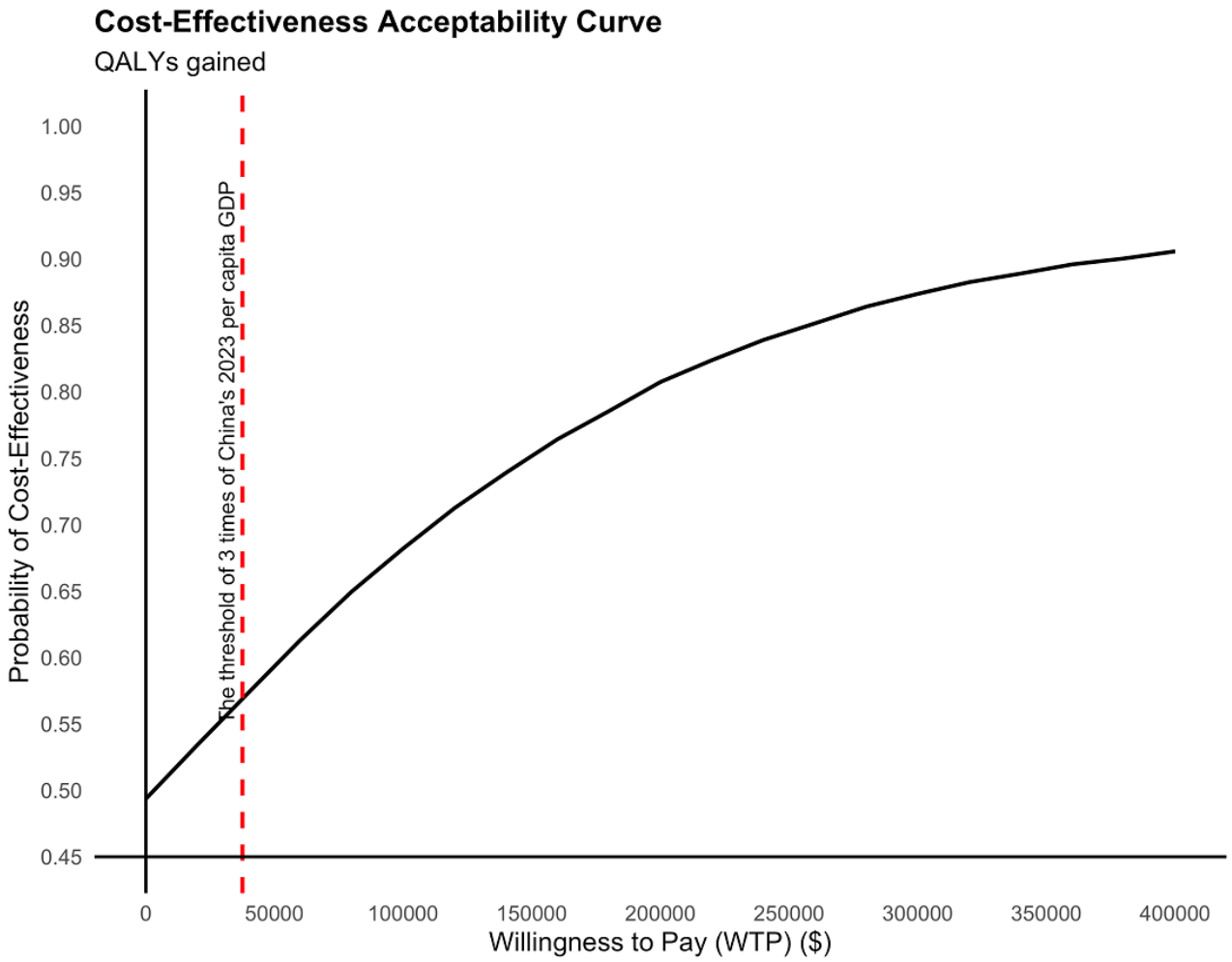

Usage Data and Its Relationships With Measurement Outcomes

The 3-month usage data of MBC program varied considerably (Table 5). The median of the total usage duration (minutes) was 199.60 (IQR 70.90-451.31, mean 360.59, SD 511.72), and the total login frequency (time) was 39.50 (IQR 19.00-86.50, mean 57.02, SD 50.26). In our intervention, a total of 76 participants (76/96, 79.2%) met the expected use as specified beforehand. The Mindfulness Practices module was most popular among participants (median usage duration 124.06, IQR 8.45-330.33; login frequency 22, IQR 8.25-43.50), followed by the Library (median usage duration 25.54, IQR 5.82-111.75; login frequency 11, IQR 4-26) and Mindfulness Yoga modules (median usage duration 3.89, IQR 0.12-16.91; login frequency 2, IQR 1-6) (Table 5). No moderate or strong correlation between MBC usage data and measurement outcomes was found.

Table 5. Usage data of the Mindfulness Breast Care program.

Program and modules	Usage duration (minutes)			Login frequency (times)		
	Median (IQR)	Mean (SD)	Maximum	Median (IQR)	Mean (SD)	Maximum
Entire MBC[a] program	199.60 (70.90-451.31)	360.59 (511.72)	3382.82	39.50 (19-86.50)	57.02 (50.26)	239
Library	25.54 (5.82-111.75)	144.32 (427.31)	3177.62	11 (4-26)	22.05 (28.34)	142
Mindfulness yoga	3.89 (0.12-16.91)	17.97 (32.37)	160.58	2 (1-6)	4.81 (7.58)	57
Mindfulness practices	124.06 (8.45-330.33)	198.31 (218.73)	1088.80	22 (8.25-43.50)	30.16 (28.27)	128

[a]MBC: Mindfulness Breast Care.

XSL•FO
RenderX

Discussion

Principal Results

This is the first RCT to investigate the effectiveness and cost-effectiveness of an app-based program for survivors of breast cancer in China. Our study illustrates that MBC program can decrease body image distress and stigma, and improve mental well-being for survivors of breast cancer. Our study also demonstrates the potential cost-effectiveness of our program.

Our program statistically significantly reduced body image distress at six months, not at three months (post intervention). Prior study shows that a weekly 90-minute mindfulness-based meditation and yoga intervention significantly reduced body image distress for patients with breast cancer immediately post intervention [72]. Our study only offered weekly 45-minute videoconference on mindfulness practice and may not be sufficient to achieve optimal effects immediately. However, survivors of breast cancer can gradually enhance their mindfulness and yoga skills through cumulative practice, thereby better observing and accepting body changes at follow-up [73]. Meanwhile, the trial did not show a clinically meaningful difference of body image distress. A systematic review recommends combining psychological interventions with physical activities and cosmetic procedures for better results [74].

Our program statistically significantly decreased stigma at three and six months, which was consistent with the results from two prior studies demonstrating that face-to-face MBCT theory–driven intervention significantly decreased stigma of patients with schizophrenia [25,75]. Additionally, an important clinical difference at three and six months was discovered, indicating that the impact of MBC program on stigma was clinically meaningful. Mindfulness practices and yoga could promote awareness of negative thoughts among survivors of breast cancer, such as stigma, and treat them in a non-judgmental manner, ultimately leading to stigma reduction [12]. Notably, statistically significant time effects, as well as time and group interaction effects regarding stigma were also found, which indicated that the impact of our program was sustained and might be endured over time.

In contrast to the prior 2 studies [76,77], our program did not statistically significantly improve sleep quality. The difference between our study and the prior two studies may be explained by the density and format of interventions, with weekly 45-minute video conferencing sessions in our intervention and weekly two-hour face-to-face group sessions in the prior two interventions [76,77]. Longer immersion in mindfulness practices and face-to-face communication may enhance the impact of mindfulness on sleep quality.

Our program improved mental well-being for participating survivors of breast cancer at three months but not sustained at six months. Mindfulness practices may be helpful for participants to regulate feelings consciously, deal with emotions peacefully, and improve their mental stability [78,79]. However, such effects would not be sustained when participants felt unsupported after the intervention was completed [80]. Hence, more scientific researches are encouraged to extend the duration of app-based MBCT-guided programs for reaching longer-term mental well-being.

Time effect on social support was found at the 6-month follow-up. Although the group difference in social support at T2 did not reach statistical significance (P=.06), this study suggested a trend toward improvement. Mindfulness and stigma reduction may facilitate BCSs to build stronger relationships with others, thus promoting their social support [81].

Participating survivors of breast cancer reported better QALYs in our program. A systematic review found that all app-based programs were cost-effective and reported greater QALYs gained compared to routine care [82]. Given the stretched resources in health care provision [83] and the continuous rise in the number of cancer survivors [1], an easily accessible, sustainable, clinically effective, and relatively cost-effective strategy for posttreatment care is much needed [84,85]. Our study provides further evidence for health care budget holders when considering continuity of care after discharge.

In our study, the probability that the MBC program is cost-effective was 57% at a willingness-to-pay threshold of US $37,530 per QALY, which is in line with the concept of "more value for money" [86]. However, the cost-effectiveness probability of 57% reflects the inherent uncertainty of trial-based economic evaluations, particularly given the modest sample size and short analytic horizon [52]. Prior methodological literature indicates that probabilities of cost-effectiveness around 50–60% are common in exploratory or early-phase studies, and should be interpreted as decision uncertainty rather than evidence against cost-effectiveness [87,88]. Furthermore, owing to the lack of a guideline on the perceived acceptable probability of cost-effectiveness of internet programs by health care policymakers [89], future economic evaluations of app-based programs are needed to help draw a firm conclusion.

In our 12-week program, the average login frequency and usage duration of the entire program were 57.02 times and 360.59 minutes, respectively. This means that within the 12-week intervention period, women logged into MBC program approximately 4-5 times per week and used the program 30 minutes per week. This intervention showed better usage data than previous studies [55,90]. A prior 16-week web-based program for BCSs indicated the average of total login frequency and usage duration as 11 times and 337.2 minutes [55], and another 24-week app-based program for patients with gynecologic cancer reported the average of total login frequency to be 67.90 times and usage duration to be 85.23 minutes [90]. The better usage data in our program could be partially explained by the weekly 45-minute videoconference, where all participants were invited to practice mindfulness with the therapist during the videoconference. Participants' interactivity positively influenced their engagement in an app-based intervention [91].

In our program, the relatively large difference between mean and median revealed usage polarization among survivors of breast cancer. Usage polarization of the MBC program could be due to the variation in participants' education level, with 38.5% of participants' educational level at junior high school

XSL•FO
RenderX

or lower. Participants with less educated might not be familiar with novel technology, resulting in less use [92]. Future researchers should consider offering more instruction on app usage with plain language and applying various strategies to make app-based programs more understandable and acceptable.

Limitations

Certain limitations warrant consideration. First, the sample size was calculated exclusively on the effectiveness parameter while neglecting the economic parameter (ie, QALYs), rendering our study unlikely to be adequately robust for an economic analysis [93]. Second, we did not incorporate a face-to-face mindfulness intervention into our design, making it impossible to clarify the differences in costs and effects between different formats (app-based or face-to-face) of MBCT-guided intervention. However, internet-based interventions offer a potential platform for one mindfulness therapist to serve more cancer survivors, as well as save participants' transportation costs and time to achieve cost savings [27]. Third, owing to limited resources, only the cost-benefit within 6 months was evaluated, and the long-term effects were not considered. A short period may not fully demonstrate their cost-effectiveness, especially for health care interventions with high upfront costs and potential long-term effects [94]. Fourth, data collectors and participants were not blinded to group allocation, which may increase the risk of unconscious and conscious bias during data collection in an RCT. Further, self-report outcomes may be subject to reporting bias [95].

Conclusions

Our trial reveals that an app-based MBC program is effective and potentially cost-effective. Our study indicates that offering this program in clinical practice reduces body image distress and stigma, improves mental well-being, and saves societal costs. It is recommended that this program be incorporated into routine care to support cancer patients after primary treatment. Health care policymakers should consider allocating more resources to eHealth.

Acknowledgments

The authors thank all participants in two hospitals: Xiang'an Hospital, Xiamen University and First Affiliated Hospital, Xiamen University.

Data Availability

All data collected have been shared in Mendeley Data (DOI: 10.17632/7hsp3jnwh8.1). All data analysis results of this study are included in this article.

Funding

This study was funded by the National Natural Science Foundation of China (71974162 and 7231101009).

Authors' Contributions

HL, YZ, and SZ contributed to conceptualization, data curation, formal analysis, investigation, methodology, project administration, validation, visualization, and writing–original draft. XJ contributed to conceptualization, project administration, supervision, validation, and writing – review and editing. YX contributed to conceptualization, project administration, validation, and writing – original draft. SC and KMC contributed to supervision, validation, and writing – review and editing. SL contributed to the investigation, software, and writing – original draft. QY and WH contributed to resources and writing – review and editing. JZ contributed to conceptualization, data curation, formal analysis, funding acquisition, methodology, project administration, resources, software, supervision, validation, visualization, and writing – review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 906 KB - jmir_v28i1e85913_app1.pdf]

Multimedia Appendix 2
CHEERS 2022 checklist.
[PDF File (Adobe PDF File), 220 KB - jmir_v28i1e85913_app2.pdf]

References

1. Allemani C, Matsuda T, Di Carlo V, Harewood R, Matz M, Nikšić M, CONCORD Working Group. Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. Lancet 2018;391(10125):1023-1075 [FREE Full text] [doi: 10.1016/S0140-6736(17)33326-3] [Medline: 29395269]

XSL•FO
RenderX

2. Fang S, Chang H, Shu B. Objectified body consciousness, body image discomfort, and depressive symptoms among breast cancer survivors in taiwan. Psychology of Women Quarterly 2014;38(4):563-574. [doi: 10.1177/0361684314552652]
3. Kołodziejczyk A, Pawłowski T. Negative body image in breast cancer patients. Adv Clin Exp Med 2019;28(8):1137-1142 [FREE Full text] [doi: 10.17219/acem/103626] [Medline: 30817097]
4. Proctor M, Cassisi JE, Dvorak RD, Decker V. Medical tattooing as a complementary cosmetic intervention to reduce body-image distress and mental health symptoms in U.S. breast cancer survivors. Support Care Cancer 2024;32(9):600. [doi: 10.1007/s00520-024-08809-4] [Medline: 39167227]
5. Przezdziecki A, Sherman KA, Baillie A, Taylor A, Foley E, Stalgis-Bilinski K. My changed body: Breast cancer, body image, distress and self-compassion. Psychooncology 2013;22(8):1872-1879. [doi: 10.1002/pon.3230] [Medline: 23203842]
6. Hsu H, Juan C, Chen J, Hsieh H. Mediator roles of social support and hope in the relationship between body image distress and resilience in breast cancer patients undergoing treatment: A modeling analysis. Front Psychol 2021;12:695682 [FREE Full text] [doi: 10.3389/fpsyg.2021.695682] [Medline: 34630208]
7. Kang D, Choi EK, Kim IR, Nam SJ, Lee JE, Im YH, et al. Distress and body image due to altered appearance in posttreatment and active treatment of breast cancer patients and in general population controls. Palliat Support Care 2018;16(2):137-145. [doi: 10.1017/S1478951517000062] [Medline: 28262081]
8. Amini-Tehrani M, Zamanian H, Daryaafzoon M, Andikolaei S, Mohebbi M, Imani A, et al. Body image, internalized stigma and enacted stigma predict psychological distress in women with breast cancer: A serial mediation model. J Adv Nurs 2021;77(8):3412-3423. [doi: 10.1111/jan.14881] [Medline: 33969915]
9. Fife BL, Wright ER. The dimensionality of stigma: A comparison of its impact on the self of persons with HIV/AIDS and cancer. J Health Soc Behav 2000;41(1):50-67. [Medline: 10750322]
10. Jin R, Xie T, Zhang L, Gong N, Zhang J. Stigma and its influencing factors among breast cancer survivors in China: A cross-sectional study. Eur J Oncol Nurs 2021;52:101972. [doi: 10.1016/j.ejon.2021.101972] [Medline: 33991869]
11. Bedi M, Devins GM. Cultural considerations for South Asian women with breast cancer. J Cancer Surviv 2016;10(1):31-50. [doi: 10.1007/s11764-015-0449-8] [Medline: 25859626]
12. Jiang N, Zhang YX, Zhao J, Shi HY, Wang T, Jin W, et al. The mediator role of stigma in the association of mindfulness and social engagement among breast cancer survivors in China. Support Care Cancer 2022;30(6):5007-5015 [FREE Full text] [doi: 10.1007/s00520-022-06882-1] [Medline: 35192056]
13. Wu IHC, Tsai W, McNeill LH, Lu Q. The associations of self-stigma, social constraints, and sleep among Chinese American breast cancer survivors. Support Care Cancer 2020;28(8):3935-3944. [doi: 10.1007/s00520-019-05233-x] [Medline: 31858250]
14. Akin-Odanye EO, Husman AJ. Impact of stigma and stigma-focused interventions on screening and treatment outcomes in cancer patients. Ecancermedicalscience 2021;15:1308 [FREE Full text] [doi: 10.3332/ecancer.2021.1308] [Medline: 34824631]
15. Lewis-Smith H, Diedrichs PC, Rumsey N, Harcourt D. Efficacy of psychosocial and physical activity-based interventions to improve body image among women treated for breast cancer: A systematic review. Psychooncology 2018;27(12):2687-2699. [doi: 10.1002/pon.4870] [Medline: 30161285]
16. Zheng S, Liu S, Yang Q, Chan S, Huang W, Jiang X, et al. The effectiveness of interventions to reduce cancer-related stigma: An integrative review. J Clin Nurs 2024;33(7):2438-2455. [doi: 10.1111/jocn.17014] [Medline: 38345136]
17. Zhang J, Xu R, Wang B, Wang J. Effects of mindfulness-based therapy for patients with breast cancer: A systematic review and meta-analysis. Complement Ther Med 2016;26:1-10. [doi: 10.1016/j.ctim.2016.02.012] [Medline: 27261975]
18. Chayadi E, Baes N, Kiropoulos L. The effects of mindfulness-based interventions on symptoms of depression, anxiety, and cancer-related fatigue in oncology patients: A systematic review and meta-analysis. PLoS One 2022;17(7):e0269519 [FREE Full text] [doi: 10.1371/journal.pone.0269519] [Medline: 35834503]
19. Sipe WEB, Eisendrath SJ. Mindfulness-based cognitive therapy: Theory and practice. Can J Psychiatry 2012;57(2):63-69. [doi: 10.1177/070674371205700202] [Medline: 22340145]
20. Hofmann SG, Gómez AF. Mindfulness-based interventions for anxiety and depression. Psychiatr Clin North Am 2017;40(4):739-749 [FREE Full text] [doi: 10.1016/j.psc.2017.08.008] [Medline: 29080597]
21. Chang YC, Lin GM, Tseng TA, Vitale E, Yang CH, Yang YL. The experience of mindfulness-based stress reduction on menopausal symptoms, sleep disturbance, and body image among patients with breast cancer-a qualitative study. Curr Oncol 2023;30(1):1255-1266 [FREE Full text] [doi: 10.3390/curroncol30010097] [Medline: 36661746]
22. Musial F, Büssing A, Heusser P, Choi KE, Ostermann T. Mindfulness-based stress reduction for integrative cancer care: A summary of evidence. Forsch Komplementmed 2011;18(4):192-202. [doi: 10.1159/000330714] [Medline: 21934319]
23. Rizwana A. Mindfulness-based cognitive therapy for depression: A review. Int J Sci Healthc Res 2020;5(4):77-84. [doi: 10.1037/v00751-001]
24. Nissim RS, Roth A, Gupta AA, Elliott M. Mindfulness-based cognitive therapy intervention for young adults with cancer: A pilot mixed-method study. J Adolesc Young Adult Oncol 2020;9(2):256-261. [doi: 10.1089/jayao.2019.0086] [Medline: 31621473]

XSL•FO
RenderX

25. Tang Q, Yang S, Liu C, Li L, Chen X, Wu F, et al. Effects of mindfulness-based cognitive therapy on stigma in female patients with schizophrenia. Front Psychiatry 2021;12:694575 [FREE Full text] [doi: 10.3389/fpsyt.2021.694575] [Medline: 34366925]
26. Compen F, Bisseling E, Schellekens M, Donders R, Carlson L, van der Lee M, et al. Face-to-face and internet-based mindfulness-based cognitive therapy compared with treatment as usual in reducing psychological distress in patients with cancer: A multicenter randomized controlled trial. J Clin Oncol 2018;36(23):2413-2421. [doi: 10.1200/JCO.2017.76.5669] [Medline: 29953304]
27. Rung AL, Oral E, Berghammer L, Peters ES. Feasibility and acceptability of a mobile mindfulness meditation intervention among women: Intervention study. JMIR Mhealth Uhealth 2020;8(6):e15943 [FREE Full text] [doi: 10.2196/15943] [Medline: 32442147]
28. Nissen ER, O'Connor M, Kaldo V, Højris I, Borre M, Zachariae R, et al. Internet-delivered mindfulness-based cognitive therapy for anxiety and depression in cancer survivors: A randomized controlled trial. Psychooncology 2020;29(1):68-75 [FREE Full text] [doi: 10.1002/pon.5237] [Medline: 31600414]
29. Saevarsdottir SR, Gudmundsdottir SL. Mobile apps and quality of life in patients with breast cancer and survivors: Systematic literature review. J Med Internet Res 2023;25:e42852 [FREE Full text] [doi: 10.2196/42852] [Medline: 37494111]
30. Compen F, Adang E, Bisseling E, van der Lee M, Speckens A. Cost-utility of individual internet-based and face-to-face mindfulness-based cognitive therapy compared with treatment as usual in reducing psychological distress in cancer patients. Psychooncology 2020;29(2):294-303 [FREE Full text] [doi: 10.1002/pon.5246] [Medline: 31650662]
31. 2023 China internet users data. China Marketing Corporation. URL: https://chinamarketingcorp.com/blog/2023-china-internet-user-data/ [accessed 2025-02-18]
32. Zhou K, Wang W, Zhao W, Li L, Zhang M, Guo P, et al. Benefits of a weChat-based multimodal nursing program on early rehabilitation in postoperative women with breast cancer: A clinical randomized controlled trial. Int J Nurs Stud 2020;106:103565. [doi: 10.1016/j.ijnurstu.2020.103565] [Medline: 32272281]
33. Leslie M, Beatty L, Hulbert-Williams L, Pendrous R, Cartwright T, Jackson R, Finding My Way UK Trial Steering Group, et al. Web-based psychological interventions for people living with and beyond cancer: Meta-review of what works and what does not for maximizing recruitment, engagement, and efficacy. JMIR Cancer 2022;8(3):e36255 [FREE Full text] [doi: 10.2196/36255] [Medline: 35802418]
34. Smith IS, Bind MA, Weihs KL, Bei B, Wiley JF. Targeting emotional regulation using an Internet-delivered psychological intervention for cancer survivors: A randomized controlled trial. Br J Health Psychol 2023;28(4):1185-1205. [doi: 10.1111/bjhp.12679] [Medline: 37437963]
35. Urbaniak GC, Plous S. Research Randomizer (Version 4.0)[Computer program]. URL: http://www.randomizer.org/ [accessed 2022-04-23]
36. Zheng S, Huang W, Zhang X, Hua Y, Chan S, Liu S, et al. A mindfulness breast care app to reduce body image distress and stigma for breast cancer survivors: Development and preliminary quality assessment. Int J Nurs Sci 2025;12(3):203-209 [FREE Full text] [doi: 10.1016/j.ijnss.2025.04.007] [Medline: 40529460]
37. Kahlmann V, Moor CC, van Helmondt SJ, Mostard RLM, van der Lee ML, Grutters JC, et al. Online mindfulness-based cognitive therapy for fatigue in patients with sarcoidosis (TIRED): A randomised controlled trial. Lancet Respir Med 2023;11(3):265-272. [doi: 10.1016/S2213-2600(22)00387-3] [Medline: 36427515]
38. Chinese Anti-Cancer Association-Committee of Breast Cancer Society. Breast cancer diagnostic and treatment (2021 Edition) (article in Chinese). China Oncology 2021;31(10):954-1040 [FREE Full text]
39. Crane B. Mindfulness-Based Cognitive Therapy: Distinctive Features. London and New York: Routledge; 2017.
40. Chinese Nutrition Society. Dietary Guidelines for Chinese Residents (2022) (article in Chinese). Beijing: The People's Health Press; 2022.
41. Ter Avest MM, Huijbers MJ, Horjus CS, Römkens TEH, Witteman EM, van Dop WA, et al. Group-delivered mindfulness-based cognitive therapy to reduce psychological distress and improve sleep in patients with inflammatory bowel diseases: A multicenter randomized controlled trial (MindIBD). Inflamm Bowel Dis 2025;31(11):3021-3032. [doi: 10.1093/ibd/izaf116] [Medline: 40658641]
42. Teasdale J, Williams M, Segal Z. The Mindful Way Workbook: An 8-Week Program to Free Yourself from Depression and Emotional Distress. New York: The Guilford Press; 2014.
43. Stoyanov SR, Hides L, Kavanagh DJ, Wilson H. Development and validation of the user version of the mobile application rating scale (uMARS). JMIR Mhealth Uhealth 2016;4(2):e72 [FREE Full text] [doi: 10.2196/mhealth.5849] [Medline: 27287964]
44. Hopwood P, Fletcher I, Lee A, Al Ghazal S. A body image scale for use with cancer patients. Eur J Cancer 2001;37(2):189-197. [doi: 10.1016/s0959-8049(00)00353-1] [Medline: 11166145]
45. Chiu HC, Lin CY, Kuo YL, Hou WL, Shu BC. Resilience among women with breast cancer surviving longer than five years: The relationship with illness perception and body image. Eur J Oncol Nurs 2023;62:102254. [doi: 10.1016/j.ejon.2022.102254] [Medline: 36621263]

XSL•FO
RenderX

46. Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh sleep quality index: A new instrument for psychiatric practice and research. Psychiatry Res 1989;28(2):193-213. [doi: 10.1016/0165-1781(89)90047-4] [Medline: 2748771]
47. Zimet GD, Dahlem NW, Zimet SG, Farley GK. The multidimensional scale of perceived social support. Journal of Personality Assessment 1988;52(1):30-41. [doi: 10.1207/s15327752jpa5201_2]
48. Ware JE, Gandek B. Overview of the SF-36 health survey and the international quality of life assessment (IQOLA) project. J Clin Epidemiol 1998;51(11):903-912. [doi: 10.1016/s0895-4356(98)00081-x] [Medline: 9817107]
49. Yao ZW, Zhao BC, Yang X, Lei SH, Jiang YM, Liu KX. Relationships of sleep disturbance, intestinal microbiota, and postoperative pain in breast cancer patients: A prospective observational study. Sleep Breath 2021;25(3):1655-1664 [FREE Full text] [doi: 10.1007/s11325-020-02246-3] [Medline: 33211236]
50. Zhu J, Ebert L, Liu X, Wei D, Chan SWC. Mobile breast cancer e-support program for chinese women with breast cancer undergoing chemotherapy (part 2): Multicenter randomized controlled trial. JMIR Mhealth Uhealth 2018;6(4):e104 [FREE Full text] [doi: 10.2196/mhealth.9438] [Medline: 29712622]
51. Wang T, Huang YY, Liu XL, Molassiotis A, Yao LQ, Zheng SL, et al. Prevalence and correlates of joint pain among Chinese breast cancer survivors receiving aromatase inhibitor treatment. Support Care Cancer 2022;30(11):9279-9288 [FREE Full text] [doi: 10.1007/s00520-022-07345-3] [Medline: 36065027]
52. Husereau D, Drummond M, Augustovski F, de Bekker-Grob E, Briggs AH, Carswell C, CHEERS 2022 ISPOR Good Research Practices Task Force. Consolidated health economic evaluation reporting standards 2022 (CHEERS 2022) statement: Updated reporting guidance for health economic evaluations. BMJ 2022;376:e067975 [FREE Full text] [doi: 10.1136/bmj-2021-067975] [Medline: 35017145]
53. Ding D, Lawson KD, Kolbe-Alexander TL, Finkelstein EA, Katzmarzyk PT, van Mechelen W, Lancet Physical Activity Series 2 Executive Committee. The economic burden of physical inactivity: A global analysis of major non-communicable diseases. Lancet 2016;388(10051):1311-1324. [doi: 10.1016/S0140-6736(16)30383-X] [Medline: 27475266]
54. McGhee SM, Brazier J, Lam CLK, Wong LC, Chau J, Cheung A, et al. Quality-adjusted life years: Population-specific measurement of the quality component. Hong Kong Med J 2011;17 Suppl 6:17-21 [FREE Full text] [Medline: 22147354]
55. van den Berg SW, Peters EJ, Kraaijeveld JF, Gielissen MFM, Prins JB. Usage of a generic web-based self-management intervention for breast cancer survivors: Substudy analysis of the BREATH trial. J Med Internet Res 2013;15(8):e170 [FREE Full text] [doi: 10.2196/jmir.2566] [Medline: 23958584]
56. Esplen MJ, Wong J, Warner E, Toner B. Restoring body image after cancer (ReBIC): Results of a randomized controlled trial. J Clin Oncol 2018;36(8):749-756. [doi: 10.1200/JCO.2017.74.8244] [Medline: 29356610]
57. Sun L, Liu X, Weng X, Deng H, Li Q, Liu J, et al. Narrative therapy to relieve stigma in oral cancer patients: A randomized controlled trial. Int J Nurs Pract 2022;28(4):e12926. [doi: 10.1111/ijn.12926] [Medline: 33817894]
58. Kyle SD, Siriwardena AN, Espie CA, Yang Y, Petrou S, Ogburn E, et al. Clinical and cost-effectiveness of nurse-delivered sleep restriction therapy for insomnia in primary care (HABIT): A pragmatic, superiority, open-label, randomised controlled trial. Lancet 2023;402(10406):975-987 [FREE Full text] [doi: 10.1016/S0140-6736(23)00683-9] [Medline: 37573859]
59. Rogers RG, Meyer I, Smith AL, Ackenbom M, Barden L, Korbly N, et al. Improved body image after uterovaginal prolapse surgery with or without hysterectomy. Int Urogynecol J 2022;33(1):115-122 [FREE Full text] [doi: 10.1007/s00192-021-04954-0] [Medline: 34432089]
60. Rejas J, Pardo A, Ruiz MA. Standard error of measurement as a valid alternative to minimally important difference for evaluating the magnitude of changes in patient-reported outcomes measures. J Clin Epidemiol 2008;61(4):350-356. [doi: 10.1016/j.jclinepi.2007.05.011] [Medline: 18313559]
61. Jones K, Weatherly H, Birch S, Castelli A, Chalkley M, Dargan A. Unit Costs of Health and Social Care 2023 Manual. 2023. URL: https://kar.kent.ac.uk/105685/ [accessed 2024-04-06]
62. McClure NS, Paulden M, Ohinmaa A, Johnson JA. Modifying the quality-adjusted life year calculation to account for meaningful change in health-related quality of life: Insights from a pragmatic clinical trial. Eur J Health Econ 2021;22(9):1441-1451. [doi: 10.1007/s10198-021-01324-x] [Medline: 34089409]
63. China guidelines for pharmacoeconomic evaluations (2020). Chinese Pharmaceutical Association. 2020. URL: https://www.cpa.org.cn/cpadmn/attached/file/20211207/1638846707310195.pdf [accessed 2024-04-01]
64. Zhao M, Shao T, Yin Y, Fang H, Shao H, Tang W. Adverse event costs and cost-effectiveness analyses of anticancer drugs: A systematic review. JAMA Netw Open 2025;8(5):e2512455 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.12455] [Medline: 40423968]
65. Drummond ME, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. Methods for the Economic Evaluation of Health Care Programmes. Oxford, UK: Oxford University Press; 2005.
66. Ramsey SD, Willke RJ, Glick H, Reed SD, Augustovski F, Jonsson B, et al. Cost-effectiveness analysis alongside clinical trials II-An ISPOR good research practices task force report. Value Health 2015;18(2):161-172 [FREE Full text] [doi: 10.1016/j.jval.2015.02.001] [Medline: 25773551]
67. Statistical bulletin on national economic and social development of Xiamen city in 2023. Xiamen Municipal Bureau of Statistics. URL: https://tjj.xm.gov.cn/zfxxgk/zfxxgkml/tjsjzl/ndgb/202403/t20240320_2829912.htm [accessed 2024-07-11]

XSL•FO
RenderX

68. Marseille E, Larson B, Kazi DS, Kahn JG, Rosen S. Thresholds for the cost-effectiveness of interventions: Alternative approaches. Bull World Health Organ 2015;93(2):118-124 [FREE Full text] [doi: 10.2471/BLT.14.138206] [Medline: 25883405]
69. GDP per capita, current prices (U.S. dollars per capita). International Monetary Fund. URL: https://www.imf.org/external/datamapper/NGDPDPC@WEO/OEMDC/ADVEC/WEOWORLD/CHN?year=2023 [accessed 2024-09-08]
70. Thompson SG, Barber JA. How should cost data in pragmatic randomised trials be analysed? BMJ 2000;320(7243):1197-1200 [FREE Full text] [doi: 10.1136/bmj.320.7243.1197] [Medline: 10784550]
71. Fenwick E, Claxton K, Sculpher M. Representing uncertainty: The role of cost-effectiveness acceptability curves. Health Econ 2001;10(8):779-787. [doi: 10.1002/hec.635] [Medline: 11747057]
72. Pehlivan M, Eyi S. The impact of mindfulness-based meditation and yoga on stress, body image, self-esteem, and sexual adjustment in breast cancer patients undergoing modified radical mastectomy: A randomized controlled trial. Cancer Nurs 2025;48(3):190-199. [doi: 10.1097/NCC.0000000000001427] [Medline: 39773822]
73. Cillessen L, Schellekens MPJ, Van de Ven MOM, Donders ART, Compen FR, Bisseling EM, et al. Consolidation and prediction of long-term treatment effect of group and online mindfulness-based cognitive therapy for distressed cancer patients. Acta Oncol 2018;57(10):1293-1302 [FREE Full text] [doi: 10.1080/0284186X.2018.1479071] [Medline: 29932784]
74. Morales-Sánchez L, Luque-Ribelles V, Gil-Olarte P, Ruiz-González P, Guil R. Enhancing self-esteem and body image of breast cancer women through interventions: A systematic review. Int J Environ Res Public Health 2021;18(4):1640 [FREE Full text] [doi: 10.3390/ijerph18041640] [Medline: 33572137]
75. Dai J, Sun D, Li B, Zhang Y, Wen M, Wang H, et al. Mixed-mode mindfulness-based cognitive therapy for psychological resilience, self esteem and stigma of patients with schizophrenia: A randomized controlled trial. BMC Psychiatry 2024;24(1):179 [FREE Full text] [doi: 10.1186/s12888-024-05636-z] [Medline: 38439012]
76. Andersen SR, Würtzen H, Steding-Jessen M, Christensen J, Andersen KK, Flyger H, et al. Effect of mindfulness-based stress reduction on sleep quality: Results of a randomized trial among Danish breast cancer patients. Acta Oncol 2013;52(2):336-344. [doi: 10.3109/0284186X.2012.745948] [Medline: 23282113]
77. Bower JE, Crosswell AD, Stanton AL, Crespi CM, Winston D, Arevalo J, et al. Mindfulness meditation for younger breast cancer survivors: A randomized controlled trial. Cancer 2015;121(8):1231-1240 [FREE Full text] [doi: 10.1002/cncr.29194] [Medline: 25537522]
78. Zhang JY, Li SS, Meng LN, Zhou YQ. Effectiveness of a nurse-led mindfulness-based tai chi chuan (MTCC) program on posttraumatic growth and perceived stress and anxiety of breast cancer survivors. Eur J Psychotraumatol 2022;13(1):2023314 [FREE Full text] [doi: 10.1080/20008198.2021.2023314] [Medline: 35140880]
79. YuYu L, Shan Z, JingJun P. The mediating effect of mindfulness on demoralization syndrome and quality of life of thyroid cancer patients: A correlational study. Medicine (Baltimore) 2023;102(7):e32719 [FREE Full text] [doi: 10.1097/MD.0000000000032719] [Medline: 36800585]
80. Bogosian A, Hurt CS, Hindle JV, McCracken LM, Vasconcelos E Sa DA, Axell S, et al. Acceptability and feasibility of a mindfulness intervention delivered via videoconferencing for people with parkinson's. J Geriatr Psychiatry Neurol 2022;35(1):155-167 [FREE Full text] [doi: 10.1177/0891988720988901] [Medline: 33504245]
81. Schellekens MPJ, Jansen ETM, Willemse HHMA, van Laarhoven HWM, Prins JB, Speckens AEM. A qualitative study on mindfulness-based stress reduction for breast cancer patients: How women experience participating with fellow patients. Support Care Cancer 2016;24(4):1813-1820 [FREE Full text] [doi: 10.1007/s00520-015-2954-8] [Medline: 26446701]
82. Gentili A, Failla G, Melnyk A, Puleo V, Tanna GLD, Ricciardi W, et al. The cost-effectiveness of digital health interventions: A systematic review of the literature. Front Public Health 2022;10:787135 [FREE Full text] [doi: 10.3389/fpubh.2022.787135] [Medline: 36033812]
83. Bergerot C, Jacobsen PB, Rosa WE, Lam WWT, Dunn J, Fernández-González L, et al. Global unmet psychosocial needs in cancer care: Health policy. EClinicalMedicine 2024;78:102942 [FREE Full text] [doi: 10.1016/j.eclinm.2024.102942] [Medline: 39634034]
84. Massoudi B, Holvast F, Bockting CLH, Burger H, Blanker MH. The effectiveness and cost-effectiveness of e-health interventions for depression and anxiety in primary care: A systematic review and meta-analysis. J Affect Disord 2019;245:728-743. [doi: 10.1016/j.jad.2018.11.050] [Medline: 30447572]
85. Zhang N, Li Q, Chen S, Wu Y, Xin B, Wan Q, et al. Effectiveness of nurse-led electronic health interventions on illness management in patients with chronic heart failure: A systematic review and meta-analysis. Int J Nurs Stud 2024;150:104630. [doi: 10.1016/j.ijnurstu.2023.104630] [Medline: 38029453]
86. Banke-Thomas AO, Madaj B, Charles A, van den Broek N. Social return on investment (SROI) methodology to account for value for money of public health interventions: A systematic review. BMC Public Health 2015;15:582 [FREE Full text] [doi: 10.1186/s12889-015-1935-7] [Medline: 26099274]
87. Hollinghurst S, Peters TJ, Kaur S, Wiles N, Lewis G, Kessler D. Cost-effectiveness of therapist-delivered online cognitive-behavioural therapy for depression: Randomised controlled trial. Br J Psychiatry 2010;197(4):297-304. [doi: 10.1192/bjp.bp.109.073080] [Medline: 20884953]

XSL•FO
RenderX

88. Donker T, Petrie K, Proudfoot J, Clarke J, Birch MR, Christensen H. Smartphones for smarter delivery of mental health programs: A systematic review. J Med Internet Res 2013;15(11):e247 [FREE Full text] [doi: 10.2196/jmir.2791] [Medline: 24240579]
89. Kolovos S, van Dongen JM, Riper H, Buntrock C, Cuijpers P, Ebert DD, et al. Cost effectiveness of guided Internet-based interventions for depression in comparison with control conditions: An individual-participant data meta-analysis. Depress Anxiety 2018;35(3):209-219 [FREE Full text] [doi: 10.1002/da.22714] [Medline: 29329486]
90. Lin H, Ye M, Lin Y, Chen F, Chan S, Cai H, et al. Mobile app for gynecologic cancer support for patients with gynecologic cancer receiving chemotherapy in china: Multicenter randomized controlled trial. J Med Internet Res 2023;25:e49939 [FREE Full text] [doi: 10.2196/49939] [Medline: 37955943]
91. Crafoord M, Fjell M, Sundberg K, Nilsson M, Langius-Eklöf A. Engagement in an interactive app for symptom self-management during treatment in patients with breast or prostate cancer: Mixed methods study. J Med Internet Res 2020;22(8):e17058 [FREE Full text] [doi: 10.2196/17058] [Medline: 32663140]
92. Crouch E, Gordon NP. Prevalence and factors influencing use of internet and electronic health resources by middle-aged and older adults in a us health plan population: Cross-sectional survey study. JMIR Aging 2019;2(1):e11451 [FREE Full text] [doi: 10.2196/11451] [Medline: 31518256]
93. Hollingworth W, McKell-Redwood D, Hampson L, Metcalfe C. Cost-utility analysis conducted alongside randomized controlled trials: Are economic end points considered in sample size calculations and does it matter? Clin Trials 2013;10(1):43-53. [doi: 10.1177/1740774512465358] [Medline: 23231969]
94. Abbott JH, Wilson R, Pryymachenko Y, Sharma S, Pathak A, Chua JYY. Economic evaluation: A reader's guide to studies of cost-effectiveness. Arch Physiother 2022;12(1):28 [FREE Full text] [doi: 10.1186/s40945-022-00154-1] [Medline: 36517825]
95. Bridgman S, Engebretsen L, Dainty K, Kirkley A, Maffulli N, ISAKOS Scientific Committee. Practical aspects of randomization and blinding in randomized clinical trials. Arthroscopy 2003;19(9):1000-1006. [doi: 10.1016/j.arthro.2003.09.023] [Medline: 14608321]

Abbreviations

BIS: Body Image Scale
CHEERS: Consolidated Health Economic Evaluation Reporting Standards
GDP: gross domestic product
ICER: incremental cost-effectiveness ratio
MBC: mindfulness breast care
MBCT: mindfulness-based cognitive therapy
MCS: Mental Component Summary
MSPSS: Multidimensional Scale of Perceived Social Support
PCS: Physical Component Summary
PSQI: Pittsburgh Sleep Quality Index
QALY: quality-adjusted life year
QoL: quality of life
RCT: randomized controlled trial
SF-6D: Short-Form Six-Dimension
SF-36: 36-item Short-Form Health Survey
SIS: Social Impact Scale

Edited by A Stone; submitted 15.Oct.2025; peer-reviewed by M Knobf, H He, Y Hu; comments to author 13.Feb.2026; revised version received 13.Mar.2026; accepted 13.Mar.2026; published 24.Apr.2026.

Please cite as:
Lin H, Zhong Y, Zheng S, Jiang X, Xu Y, Chan S, Liu S, Chow KM, Yang Q, Huang W, Zhu J
Effectiveness and Cost-Effectiveness of an App-Based Mindfulness Breast Care Program for Reducing Body Image Distress and Stigma Among Survivors of Breast Cancer: Randomized Controlled Trial
J Med Internet Res 2026;28:e85913
URL: https://www.jmir.org/2026/1/e85913
doi:10.2196/85913

©Huicong Lin, Yujing Zhong, Shuang Zheng, Xiaoying Jiang, Yun Xu, Sally Chan, Shengjie Liu, Ka Ming Chow, Qingmo Yang, Wenhe Huang, Jiemin Zhu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org),

XSL•FO
RenderX

24.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Agile Development and Testing of a Gamified Human Milk Feeding Education Mobile App for Participants of the Special Supplemental Nutrition Program for Women, Infants, and Children: Co-Design Approach

Yarisbel Melo Herrera[1], RDN, PhD; Shannon Mitchell[2], BS; Amy Malinowski[3], RD; Catherine Wright[1], MS; Marguerite Dibble[2], BS; Patricia Cassi[3], BSS, IBCLC; Naina Qayyum[1], MS; Erin Hennessy[1], MPH, PhD

[1]Gerald J. and Dorothy R. Friedman School of Nutrition Science and Policy, Tufts University, Boston, MA, United States

[2]GameTheory Co, Burlington, VT, United States

[3]Special Supplemental Nutrition Program for Women, Infants, and Children (WIC), Vermont Department of Health, Waterbury, VT, United States

Corresponding Author:
Erin Hennessy, MPH, PhD
Gerald J. and Dorothy R. Friedman School of Nutrition Science and Policy
Tufts University
150 Harrison Ave
Boston, MA, 02111
United States
Phone: 1 617 636 3636
Email: erin.hennessy@tufts.edu

Abstract

Background: Human milk feeding and breastfeeding are the recommended standards for infant feeding. Nevertheless, breastfeeding rates in the United States remain below target levels, with disparities across racial, ethnic, and income groups. The Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) plays a substantial role in reducing these disparities by providing lactation support to individuals with low income. With ongoing WIC modernization efforts, there is an opportunity to create and optimize technology solutions responsive to WIC participants' and staff's needs to increase access to the program and its services.

Objective: This study aimed to describe the development and pilot testing of Daily Drop, a gamified, low-bandwidth mobile app to provide human milk feeding education and support for WIC participants.

Methods: Guided by the 5-stage model for comprehensive research on telehealth, Daily Drop underwent 3 stages: concept development, service design, and preimplementation. Using a mixed methods approach, the project team sought feedback from WIC leadership and staff at the state and local levels, state IT staff, and WIC participants at each stage. Suggestions from stages 1 and 2 were incorporated into the testable app before field testing (stage 3). During field testing, participants and staff completed surveys across multiple time points and qualitative interviews to evaluate the app's feasibility, usability, and acceptability. Quantitative data were analyzed using descriptive statistics, and qualitative data were thematically analyzed.

Results: Key feedback from WIC participants and staff included providing flexibility for a variety of human milk feeding approaches (stage 1); and providing easily accessible educational information throughout gameplay, diversifying progress tracking to emphasize knowledge growth and expertise development, and including supportive or affirming messages for users (stage 2). During field testing (stage 3), >67% of WIC participants agreed with 7 out of 12 acceptability, satisfaction, and usability questions about the app, reiterated in interviews where they highlighted the simplicity of the app and how it increased their confidence to feed human milk. However, barriers to app use included a lack of reminders and repetitive information for parents with previous human milk feeding experience. Similarly, for WIC staff, mean scores for acceptability and feasibility were 3.8 (SD 1.0) and 4.4 (SD 0.6), respectively (max 5) at the early phase, but these scores declined over time. Staff recommendations included providing further, in-depth training to increase their familiarity with the app and reporting, and integrating the reports into WIC's management information system.

Conclusions: The development of Daily Drop followed an agile development, co-design approach with the involvement of relevant key partners at all stages of development. Overall, Daily Drop was deemed acceptable, usable, and engaging by WIC

XSL•FO
RenderX

participants and staff. Future research could focus on testing its effectiveness in improving human milk feeding behaviors and cost-effectiveness.

(J Med Internet Res 2026;28:e80330) doi:10.2196/80330

KEYWORDS

agile development; app; breastfeeding; gamification; human milk feeding; mHealth; software testing; WIC

Introduction

Overview

Breastfeeding and human milk feeding [1] are the recommended standards for infant feeding, given their short- and long-term health benefits for both the infant and birthing parent [2]. More specifically, the World Health Organization and the American Academy of Pediatrics recommend exclusive human milk feeding for the first 6 months of life, with continued human milk feeding alongside complementary solid foods until 2 years or beyond [2-4]. Nevertheless, rates of exclusive human milk feeding in the United States remain below Healthy People 2030 targets—by 15.2% at 6 months and 14.6% at 12 months [5-7]—and significant disparities exist across racial, ethnic, and income groups [5]. There is strong and consistent evidence of barriers impeding continued human milk feeding, including a lack of knowledge about human milk feeding, a lack of parental leave and unsupportive work environments, and a lack of access to health care or unsupportive health care practices, among others [8-10].

Therefore, one of the main goals of the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) is to provide lactation support, through access to lactation health professionals and peer counselors [11], to encourage human milk feeding initiation and extended duration among families with low income, and by extension, help reduce racial and ethnic disparities in these outcomes [12,13].

Nevertheless, as access to and use of technology and web-based and mobile apps has become more widespread in the United States [14], many parents have turned to these technologies to access information on human milk feeding [15,16]. Furthermore, the COVID-19 pandemic accelerated the shift toward online platforms and solutions within the WIC program, including remote nutrition education and lactation support appointments (eg, via phone or video call), remote benefit issuance, online nutrition education platforms, and online grocery ordering [17-19]. Since then, several reports have found that WIC participants accept and support routine technology use to better access WIC services [20,21]. Therefore, the US Department of Agriculture (USDA) Food and Nutrition Service launched WIC modernization efforts, which include initiatives to improve information management systems, create and optimize data-sharing solutions, and enhance websites and apps [22].

Recent systematic investigations of mobile health (mHealth) apps and interventions targeting human milk feeding behaviors among the general and low-income populations show that while apps are highly used and acceptable to users, their effectiveness on improving breastfeeding outcomes is highly variable [23-26]. This variability is likely due to significant heterogeneity in the app's purpose (eg, provision of information vs tracking), active components (eg, videos, push notifications, and forum discussions), and target users (eg, birthing parents vs fathers), among other factors [23,26]. Furthermore, less is known about the extent to which end users were involved in the development process of these apps and the impact of this involvement on breastfeeding outcomes. Among studies evaluating user feedback, key components of apps include personalized education, peer support [25,26], encouragement, and audiovisual depictions of proper breastfeeding techniques [16]. Thus, incorporating user feedback during the mHealth app development is imperative to ensure maximum acceptability and uptake among intended users.

Agile development is a creative methodology that centers the perspective of a product's audience through iterative testing and flexible development that adapts to new opportunities and challenges [27]. Key principles of agile development include (1) cross-disciplinary collaboration between audience members, subject matter experts, and creative professionals; (2) rapid, iterative development for early and frequent product testing; and (3) intentional reflection and decision-making through transparent testing, communication, and collaboration across key partner groups. Although agile development has widespread adoption across technology fields and has been shown to promote clarity, stability, and relevancy of products [28], it is still an underused process in many fields and programs, including WIC [29].

To our knowledge, no mHealth app focusing on human milk feeding behaviors has been developed specifically for WIC participants. Therefore, as part of the USDA/Tufts Telehealth Intervention Strategies for WIC (THIS-WIC) grant, the Vermont State WIC agency (VT WIC) collaborated with GameTheory Co to develop Daily Drop, a low-bandwidth mHealth app for WIC participants to use during pregnancy and through the early weeks postpartum for human milk feeding support. Daily Drop used a "serious game" design approach, which capitalizes on "traditional" game mechanics and technologies to promote learning and practicing skills, and by extension, evoke behavior change [30]. Consistent with an agile development approach, app content and functionality were informed by evidence-based WIC guidelines and refined iteratively through collaboration with subject matter experts, WIC leadership, and WIC participants. This manuscript reports on the agile development and pilot testing of Daily Drop.

Conceptual Framework: 5-Stage Model for Comprehensive Research on Telehealth

The 5-stage model for comprehensive research on telehealth [31] informed the development and pilot testing of Daily Drop. The stages include (1) concept development, (2) service design,

XSL•FO
RenderX

(3) preimplementation, (4) implementation, and (5) postimplementation. This manuscript reports the methodology and results of the first 3 stages (Figure 1 [31]), by stage. Throughout all stages, the core project team used respect-based design principles, which bridge elements of agile, user-centered, and empathetic co-design practices into a centralized theory for creating an mHealth solution that feels approachable, relevant, and most importantly, respectful of users' identities and experiences.

Figure 1. Timeline of the agile codevelopment process and testing of Daily Drop, a gamified human milk feeding education app for Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) participants, in alignment with the 5-stage model for comprehensive research on telehealth.

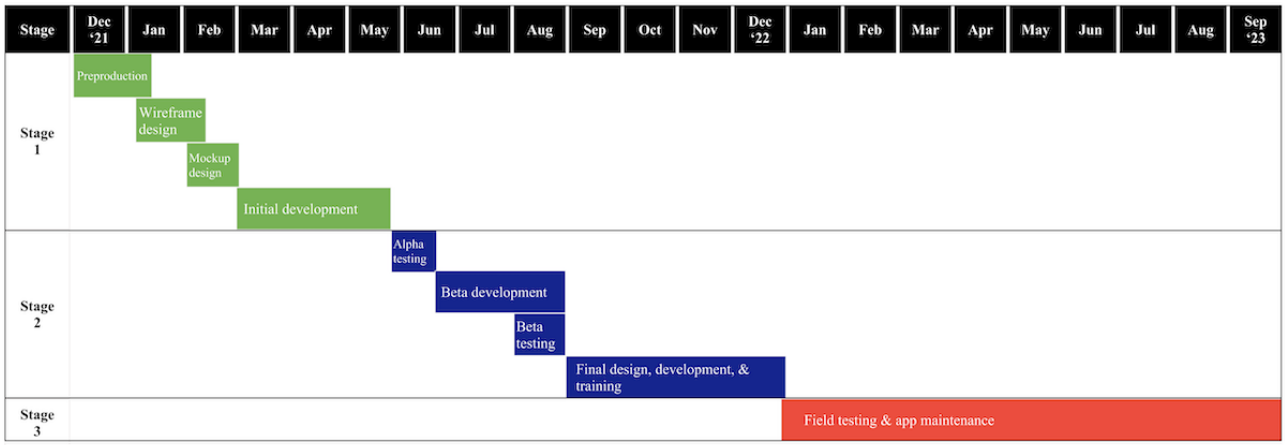

Stage	Milestone	Summary	Timeline
Stage 1	Preproduction	Discovery sessions with key stakeholders to set goals, personas, and constraints	December 6, 2021 → January 14, 2022
	Wireframe design	Wireframe review and iteration with the core project team	January 14, 2022 → February 8, 2022
	Mockup design	Clickable mockup review and iteration with the Advisory Council	February 8, 2022 → March 1, 2022
	Initial development	Core functionality is built into the app	March 1, 2022 → May 24, 2022
Stage 2	Alpha testing	Conduct testing with WIC staff and participants	May 24, 2022 → June 7, 2022
	Beta development	Iterative buildout using alpha findings & feedback from the core project team	June 7, 2022 → August 30, 2022
	Beta testing	Conduct testing with WIC staff, professionals, and participants	August 2, 2022 → August 30, 2022
	Final design, development, & training	Iterative redesign using beta testing feedback; complete quality assurance; train staff	August 30, 2022 → January 1, 2023
Stage 3	Field testing & app maintenance	Users have access to the app; the developers provide quarterly maintenance	January 1, 2023 → September 30, 2023

Stage 1: Preproduction, Design, and Initial Development

Methods

Preproduction

During this phase (December 2021 to January 2022), core project team members (software development team and VT WIC leadership) envisioned possibilities for the project, set goals and key performance indicators, and identified constraints, assumptions, and essential requirements for project success. Furthermore, to inform all aspects of the project, VT WIC convened an advisory council that included individuals representing the State of Vermont IT staff, VT WIC IT department, local agency WIC staff, and WIC participants. The core project team met with the advisory council to gather their insights on potential users, app benefits and risks, and to receive feedback on the alpha-, beta-, and field-testing plans, as well as evaluation-related questions.

Using this input, the software development team developed "User Personas," fictional profiles leveraged in user-centered design to describe potential users' motivations, concerns, needs, and values [32]. User personas are widely used in health care and in the co-design of mHealth and eHealth technologies [32-34] to ensure that the proposed concept and planned functionality align with the needs of diverse potential users, rather than relying solely on the perspectives of those directly involved in the design process. For this project, User Personas were collaboratively designed to represent a wide range of perspectives, including technology access limitations, minimal baseline knowledge of the learning material, and socioeconomic characteristics and experiences not well represented within the core project team or the advisory council. Then, the software development team created accompanying "User Stories," short feature descriptions written from a user perspective with information on the value each feature may provide to a specific User Persona (Figure 2) [35]. Together, the User Personas and User Stories provided a shared reference point for key partners in this stage and, in later stages, served as an evaluative framework to assess how the product design mapped against the range of qualities, experiences, sentiments, and perceptions of both potential end users and key partner groups in later stages.

Lastly, VT WIC leadership served as subject matter experts to identify key sources of evidence-based content for the app,

XSL•FO
RenderX

which included USDA's WIC Breastfeeding Curriculum and relevant resources from other state WIC agencies, USDA, Centers for Disease Control and Prevention, National Institutes of Health, La Leche League, and the Institute for the Advancement of Breastfeeding and Lactation Education.

Figure 2. Example of a User Persona and Story used in Daily Drop's preproduction (stage 1, December 2021 to January 2022). Image source: Flickr, published under Creative Commons Attribution 4.0 International License [36].

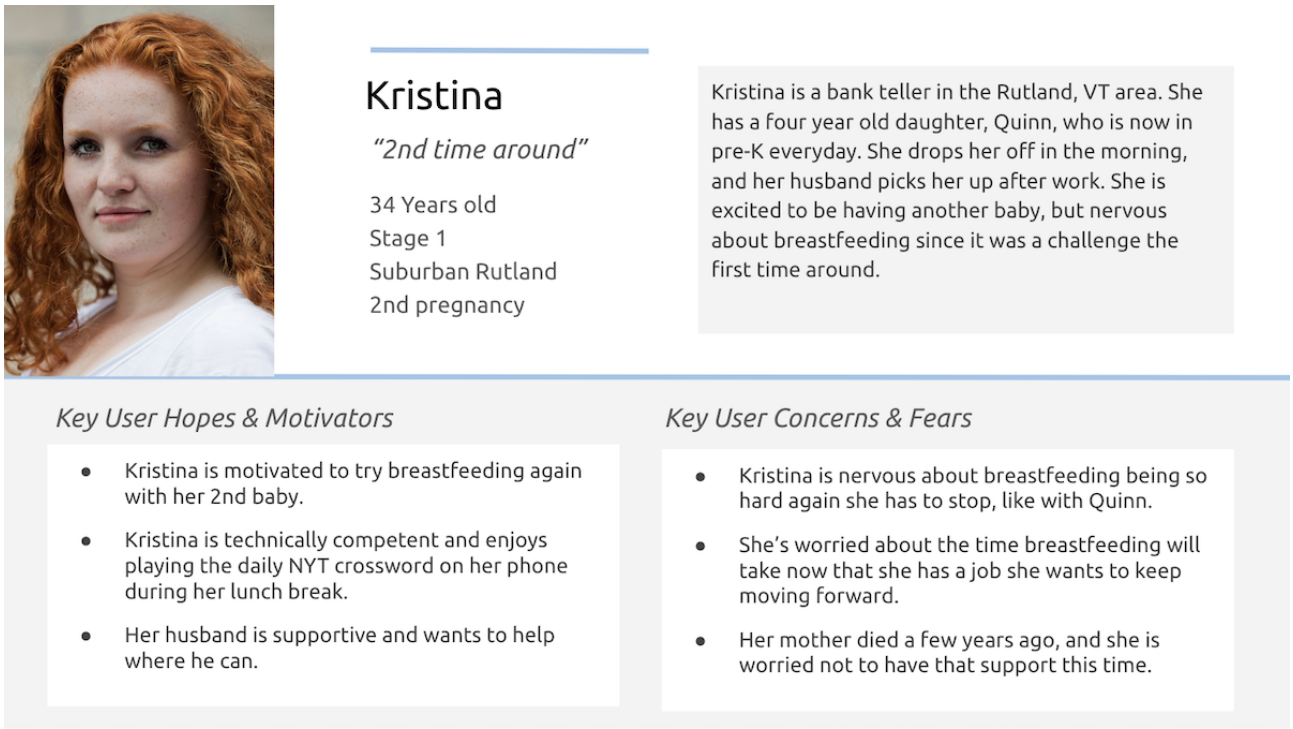

Wireframe and Mockup Design

This phase was conducted between January and March 2022. The software development team leveraged the findings from preproduction into a wireframing product screen by screen, without aesthetic or brand styling applied (Figure 3). The purpose of a wireframe is to define simple screens that can be quickly and easily changed based on review and feedback [37]. The software development team developed simple screen concepts based on the user stories and presented them to the core project team for review and feedback. The software development team also worked on stylistic development to establish a visual style and aesthetic approach for the product. Various style concepts were created and narrowed down based on the core project team's feedback.

Together with the findings from the wireframe and style review sessions, the software development team refined and redeveloped the screens to create a clickable mockup to serve as a prototype (Figure 3). The clickable mockups were designed to showcase the desired product experience, intentionally excluding custom user input, final content, and any user data, thereby encouraging additional review. The software development team presented the clickable mockups to the advisory council for feedback early in the development process in order to identify areas for improvement.

XSL•FO
RenderX

Figure 3. Wireframe (A) and clickable mockup (B) used in Daily Drop's design (stage 1, January-February 2022).

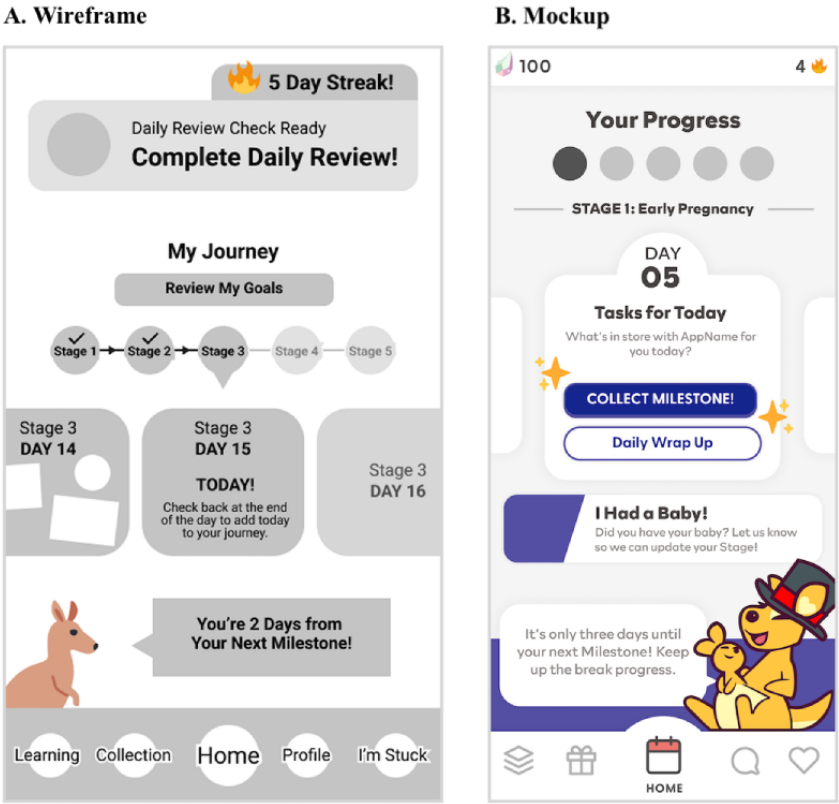

Initial Development and Game Design

Following the wireframe and mockup design phase, the software development team worked to establish an early technical prototype between March and May 2022. In the manner of prototype, test, and evolve, the software development team created a horizontal slice, which focuses on building simple versions of the app features first. It excludes learning material and visuals to focus on the "must-have" features and systems. Once completed, the software development team included the key content selected by the subject matter experts to achieve an alpha level of completion. Software is considered alpha complete at the earliest point at which all core functionality and systems are ready for review by key partners.

Ethical Considerations

The Vermont Agency of Human Services Institutional Review Board (IRB) reviewed stage 1 activities and deemed them non–human subjects' research. Still, the core project team reiterated to advisory council members that their contributions were voluntary and that they could stop their participation at any time. Advisory council members who were WIC participants received US $25 gift cards per meeting attended, whereas employees of the State of Vermont did not receive financial incentives in compliance with federal and/or state policies.

Results

Wireframe and Mockup Design

Key feedback at this phase included ensuring that games provided flexibility for a variety of human milk feeding approaches, allowing users to complete the introductory, setup questions for the app in an intuitive and easy way, increasing customization options, and leaning into the microlearning approach as an alternative to other digital information approaches. This feedback largely validated the initial wireframe design, while resulting in the design of a mascot with further customization options and game results that were neutral or open-ended rather than having strict right-or-wrong answers in all cases. The software development team applied the feedback from the advisory council to the clickable mockups and used the design as a blueprint as the project moved into the development phase. This early review served to refine and mitigate critical product risks at an early development stage, thus reducing overall project expenses.

Initial Development and Game Design

The User Personas established a preference for fast, pick-up-and-play experiences that could fit within busy schedules. Thus, the app was designed to operate through short microinteractions within a playlist, with rapid rewards for each interaction and the ability to advance game progress over relatively short durations. Minigames were selected for their recognizability and simplicity, relying on common interaction patterns such as drag-and-drop, matching, image comparison, and multiple-choice. These interaction styles reflect widely used user experience heuristics and were intended to reduce cognitive load while supporting a wide range of human milk feeding learning content.

Based on these design principles, 6 types of minigames were developed: matching games (eg, matching images and text), identifying differences in images, card sorting, and games to identify myths or facts about human milk feeding. When users opened the app, they accessed a "Daily Playlist" consisting of 5 minigames, enabling progress through short, self-contained

XSL•FO
RenderX

sessions. User performance was tracked through points awarded for correct responses and deductions for incorrect answers.

At the start of the app experience, users completed the Breastfeeding Attrition Prediction Tool (BAPT) [38], a validated survey to assess readiness and supports around human milk feeding among pregnant individuals routinely used during lactation services by VT WIC [1,39]. The gamification pathway was personalized based on the user's BAPT score domains, which relate to an individual's knowledge of human milk feeding, support, and confidence. Each minigame was tagged to 1 of the 3 BAPT domains, and domain scores could range between –50 and 50 points depending on user responses. For the first 10 active days of gameplay, the "Daily Playlist" presented mostly minigames tagged to the user's lowest domain score. After the first 10 days, minigames were randomly selected from a list of the 10 least recently shown minigames across all domains, and the 10 least recently shown minigames within the user's lowest domain, supporting both reinforcement and content variety.

To reinforce app engagement, the app's reward framework was informed by the Quantic Foundries Gaming Motivation Framework [40], leveraging "Completion" and "Achievement" motivation clusters. Completionist users are motivated to finish everything the game has to offer, for instance, complete every mission, find every collectible, and unlock all content. This motivation cluster has the broadest appeal across age and gender and aligns well with pick-up-and-play designs that unfold over multiple sessions. In Daily Drop, completionist engagement was supported through in-game rewards, including "Gems" earned for correct responses and redeemable for mascot costumes and accessories, randomized item unlocks, item rarity, playlist performance progress tracking and scoring, a daily emotion log, a knowledge library, and a minigame library. Additional commercially successful engagement mechanics were incorporated into the app, including streak tracking and a "Challenge Playlist" that allowed users to retry higher-difficulty content.

Stage 2: Alpha and Beta Testing

Methods

Alpha Testing

Alpha testing is an assessment typically done in-house as a quality check to ensure accessibility, usability, accuracy of information, and alignment with technical, legal, and ethical standards [41]. However, to remain consistent with the agile development approach, the software development team invited potential users and key partners between May and June 2022 to respond to an early version of the project to identify potential access and usage barriers that could limit product impact [42]. Alpha testing procedures included conducting 4 focus groups via video call with 2 groups: WIC participants (n=6) and WIC staff, including full-time nutritionists (n=2) and breastfeeding peer counselors (n=1). WIC staff in 5 local offices recruited WIC participants who met the inclusion criteria (aged 18 years or older and pregnant) during routine appointments. If WIC participants were interested in taking part in alpha testing, they provided their contact information, which was shared with the project team. A project team member then reached out to formally invite them to participate. The project team also emailed all WIC staff and breastfeeding peer counselors to invite them to participate in the focus groups.

Each focus group had 1-2 moderators and an observer. Focus groups began with an overview of the project, a demo of the early prototype, and examples of the planned look and feel. Then, focus group participants were asked questions related to (1) experiences with human milk feeding, (2) current technology use and barriers to technology use, (3) preferences on the art style, (4) the alignment of the prototype with project goals and staff and user needs, and (5) overall suggestions for improvements. Finally, focus group participants completed a survey regarding app usability and motivations for future use.

Beta Development

Following alpha testing, the software development team incorporated feedback through weekly development, testing, and review with the core project team between June and August 2022. Changes and updates were checked against the available project scope before being applied. If any changes were deemed out of scope, design work was completed to identify alternative mechanisms to address the underlying feedback. The product gradually achieved greater stability through this process, and its functionality was expanded and refined until it reached a nearly final state. Quality assurance testing was also conducted toward the end of the beta development phase to identify any high-priority issues that need resolution before beta testing [43-45]. Examples of procedures included testing the app with older phones or unstable internet connection and ensuring appropriate functionality for all app features.

Beta Testing

Beta testing aims to validate the product in real-world conditions before release, evaluating user experiential learning and interactivity to ensure alignment with the project's goals [41]. In August 2022, WIC participants (n=12), staff (n=5), and professionals (ie, WIC experts outside of VT WIC, n=5) served as beta testers. Some WIC participants who were alpha testers were invited to participate as beta testers, in addition to new testers. WIC staff recruited new WIC participant testers during routine appointments at 5 local offices, provided they met the inclusion criteria (aged 18 years or older and currently or recently pregnant). If WIC participants were interested in taking part in beta testing, they provided their contact information, which was shared with the project team. A project team member then reached out to formally invite them to participate. WIC participants were involved in beta testing over a 4-week period, which included 2 meetings with asynchronous testing sessions in between.

The first meeting of beta testing was an onboarding session for groups of no more than 4-6 WIC participants. During this meeting, the software development team walked WIC participants through downloading the game onto their smartphones, demonstrated how to use the app, and discussed the testing purpose in this phase. WIC participants were given access to the game and asked to use it for 4 weeks. While testing, data on their app usage was collected and linked to their WIC ID. Collected metadata included time spent in the app, game

XSL•FO
RenderX

engagement streaks (ie, consecutive days using the app), and scores in each BAPT domain (knowledge, support, and confidence). At the close of the testing window, the VT WIC team conducted qualitative interviews with each WIC participant via video call. Interview questions asked about testers' barriers and facilitators to using the product, motivators for future use, and overall experience. After the interview, WIC participants also completed a survey (developed by the core project team) regarding the acceptability and ease of use of the app and changes in perceptions regarding their human milk feeding experience, informed by key feedback from alpha testing.

WIC staff and professionals were invited to participate in beta testing via email. However, the instructions for testing were more flexible for WIC staff and professionals than for participants, encouraging them to explore the app to become familiar with its features and educational framework. Then, WIC staff and professionals who tested the app were invited to complete the same survey as WIC participants. WIC staff did not receive financial incentives in compliance with federal and/or state policies.

Final Design, Development, and Staff Training

The core project team collaborated in a final development phase for approximately 4 months (August 2022 to January 2023) to design upgrades and changes to the app based on the findings of beta testing. These designs were built into the app, and the software development team completed additional iterative rounds to ensure the user-identified issues were addressed. Lastly, the team completed a rigorous quality assurance phase during which the tool was tested on various devices and conditions to ensure the app was stable and ready for the pilot trial (Figure 4).

During this phase, the core project team solidified the data transfer and storage processes, which included establishing the file structure, creating the data dictionary, and developing a data visualization dashboard. The dashboard was vetted with the advisory council and revised based on their feedback. This dashboard was used to track key components of users' app experience, such as time in app, time per session, and changes in learning scores during the field-testing stage (stage 3).

Lastly, the core project team hosted a 2-part training for WIC staff in preparation for the field-testing stage. Part 1 included introducing the app, its purpose and features, a preview of the clinical report tool, and instructions on how to download and use the test app. Part 2 included providing details on the pilot trial and their role in it, how to promote the app and document its use in WIC's management information system, and practice using the clinical report tool in breakout rooms.

Figure 4. Screenshots of Daily Drop's final design (stage 2, September-December 2022). From left to right: (A) initial screen with motivational message and the app's mascot, (B) overview of the "Learn" tab, allowing access to Daily and Challenge Playlists and skills library, (C) overview of the "Home" tab, displaying daily tasks and progress, with a message on how to connect with the local WIC office.

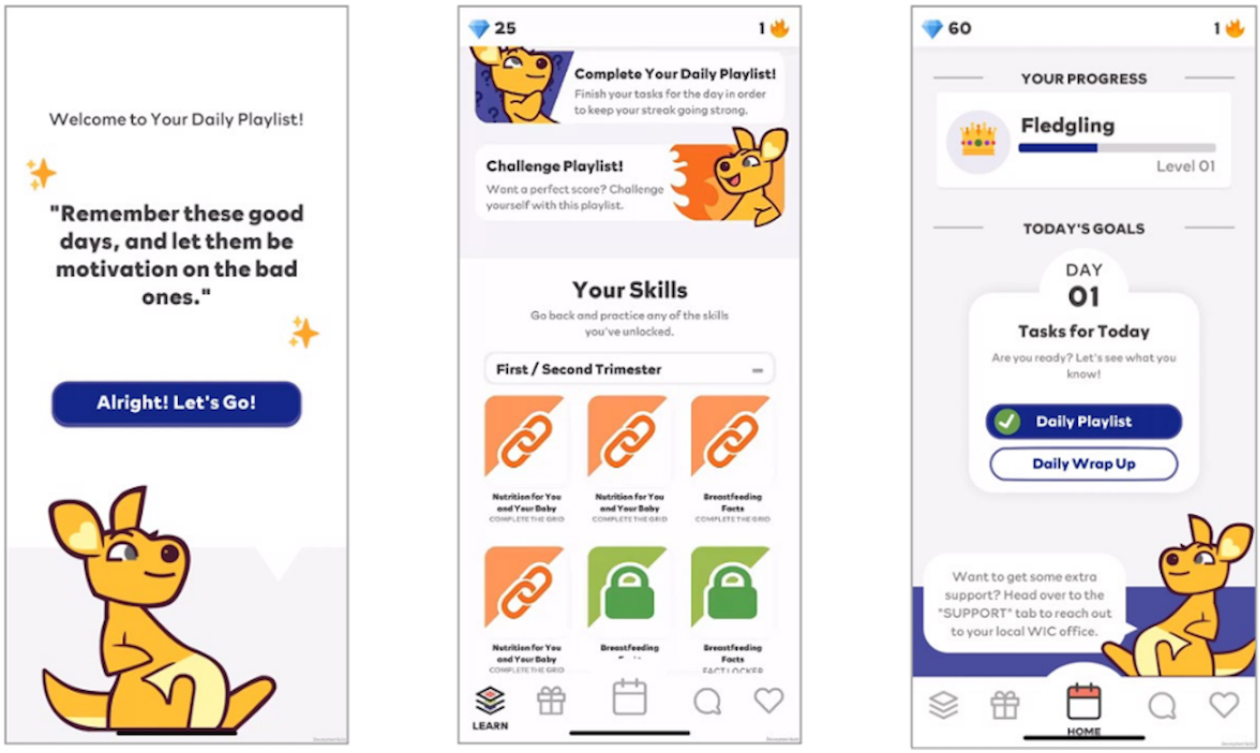

Quantitative Data Analysis

Analyses were conducted in SAS (version 9.8; SAS Institute). Descriptive statistics were performed to summarize responses to alpha- and beta-testing survey items (via means and SDs) among WIC participants, staff, and professionals.

Qualitative Data Analysis

A thematic analysis [46] of alpha testing focus groups was conducted using a deductive approach. Two members of the software development team developed a codebook and summary template in alignment with focus group questions (behavioral motivations for app engagement, validation of design and

XSL•FO
RenderX

mechanical choices, content suggestions, and functionality requests), and summarized the content of focus groups transcripts into the template. To ensure trustworthiness and credibility of findings [47], and alignment with the project goals, themes, subthemes, and exemplar quotes were member-checked with the advisory council and members of the core project team and discussed until consensus was reached.

Ethical Considerations

The Vermont Agency of Human Services IRB reviewed stage 2 activities and deemed them non–human subjects' research. Still, the core project team reiterated to testers that participation was voluntary, and that they could stop their participation at any point. WIC participants who attended alpha testing focus groups received a US $25 gift card, and those who completed all portions of beta testing received US $100 gift cards in compensation for their time. WIC staff did not receive financial incentives for their participation in stage 2 activities in compliance with federal and/or state policies.

Results

Alpha Testing

Survey results indicated that alpha testers perceived the app as very useful for pregnant parents both in rural Vermont and overall, and indicated that "building skills and knowledge" and "having a quick way to get answers to questions or help with challenges" were the most prominent factors to make them interested in using the app (data not shown). Key suggestions from focus groups included (1) allowing app users to complete the BAPT questionnaire within the first couple of days using the app vs completing it during sign-up due to its length, (2) providing educational information after incorrectly answering questions, and (3) adding affirmations throughout gameplay to encourage users and keep the app feeling supportive.

Beta Testing

Survey results indicated that all WIC participants and staff considered their app experience as positive (mean scores 5.0-6.0 out of 7), highlighting the app's potential to improve the breastfeeding experience (mean scores 5.4-6.0 out of 7; Table 1). However, the mean score of 4.0 for the statement "Was the application motivating for you to keep playing over time?" was indicative of neutral motivation for WIC participants. This finding was supported during qualitative interviews, during which WIC participants indicated wanting the app to provide more answer clarity and content diversity. This feedback led to the following improvements: (1) diversification of progress tracking or reward mechanics, particularly emphasizing knowledge growth and expertise development, (2) adjustments to the variety and length of the daily content, (3) improvements to navigation and user experience, and (4) updates to improve user experience on a broader range of mobile devices.

Table 1. Post–beta testing of Daily Drop survey results among WIC[a] participants, staff, and professionals (stage 2, August 2022).

Statement	WIC[a] participants (n=12), mean (SD)	WIC staff (n=5), mean (SD)	WIC professionals (n=5), mean (SD)
Overall, how would you rate your experience with the application?[b]	5.3 (0.9)	6.0 (0.6)	5.0 (0.7)
How easy was it to find the information you wanted in the application?[c]	4.8 (1.9)	5.4 (1.0)	5.4 (1.3)
Was the application motivating for you to keep playing over time? (WIC participants)[d]	4.0 (1.3)	5.2 (0.4)	4.8 (1.1)
How motivating did you perceive the application would be for WIC participants? (WIC staff and professionals)[d]	4.0 (1.3)	5.2 (0.4)	4.8 (1.1)
How do you feel playing the application when you were pregnant/breastfeeding would have influenced your breastfeeding experience? (WIC participants)[e]	5.7 (1.2)	6.0 (0.9)	5.4 (0.55)
How do you feel a WIC user's breastfeeding experience would have improved or worsened because of the application? (WIC staff and professionals)[e]	5.7 (1.2)	6.0 (0.9)	5.4 (0.55)
Was the amount of breastfeeding information presented in the daily playlist an appropriate amount for your learning?[f]	5.2 (0.9)	N/A[g]	N/A
How do you feel your ability to provide breastfeeding education/counseling would be improved or worsened because of the application?[e]	N/A	5.6 (0.8)	5.4 (0.5)

[a]WIC: Special Supplemental Nutrition Program for Women, Infants, and Children.

[b]7-point Likert scale where 1=very negative and 7=very positive.

[c]7-point Likert scale where 1=very hard and 7=very easy.

[d]7-point Likert scale where 1=very demotivating and 7=very motivating.

[e]7-point Likert scale where 1=greatly worsened and 7=greatly improved.

[f]7-point Likert scale where 1=not enough information and 7=too much information.

[g]N/A: not applicable.

XSL•FO
RenderX

Stage 3: Field Testing (Pilot Trial)

Methods

Overview

The project team pilot tested the (1) acceptability, feasibility, and usability of Daily Drop among WIC participants and (2) the acceptability and usability of Daily Drop to support lactation services among WIC staff. Between January 2023 and September 2023, 5 WIC local agencies participated in the pilot. WIC participants were recruited using multiple methods. First, WIC staff recruited WIC participants during their routine appointments, provided they met the inclusion criteria (aged 18 years or older, currently or recently pregnant, and had access to a smartphone or tablet, n=339). If WIC participants were interested in taking part in field testing, they provided their contact information, which was shared with the project team. A project team member then reached out to formally invite them to participate, provided them with an electronic study information sheet, and obtained their consent. Second, WIC staff also posted physical flyers at the 5 local WIC offices and sent emails and SMS text message blasts inviting potentially eligible WIC participants. Interested WIC participants could contact the project team to receive an electronic study information sheet and consent to participate.

Upon consenting, WIC participants were invited to download the app on their mobile device and asked to provide information to guide them through gameplay. This information included WIC ID, local agency, preferred term (breastfeeding or chestfeeding), pregnancy status, anticipated due date, and baby's birth date (if they indicated they recently gave birth). After the initial setup, users were asked to complete the BAPT questionnaire. Additionally, WIC staff responsible for delivering lactation support at participating agencies were invited to participate in the evaluation.

WIC Participants

WIC participants were asked to complete up to 3 surveys during the evaluation, including a baseline survey (n=74), a 1-month user survey (n=49; those who played Daily Drop at least once) or a 1-month nonuser survey (n=25), and a 3-month postpartum user survey (n=42). Electronic surveys with embedded consent language were programmed using Alchemer (Alchemer), and links were sent by VT WIC staff via email or SMS text message. Up to 2 reminders were sent for the baseline and 1-month surveys. One reminder was sent for the 3-month postpartum survey following the initial text/email. This manuscript presents results for questions related to acceptability, satisfaction, and usability among app users in the 1-month survey, which were adapted from valid and reliable tools [48-51]. Adaptations were necessary in this context, given that no existing tools captured all the aspects of WIC human milk feeding education, support, and lactation staff-participant interactions.

Moreover, a subset (n=26) of WIC participants who used Daily Drop were invited to an interview at the end of the evaluation period to provide additional insights about their app use. This subset was selected using random stratified sampling based on their app use (n=13 per stratum; stratum 1: low use [fewer than 10 sessions]; stratum 2: moderate to high use [10 or more sessions]). The semistructured interview guide was tested with members of the advisory council and refined based on their feedback. Interviews were scheduled for 30 minutes and digitally recorded and transcribed. Interviews were conducted until saturation was reached (ie, no new emergent themes) [52].

WIC Staff

Eligible WIC staff (n=24) were invited to participate in the evaluation via email with an embedded survey link. Staff surveys with embedded consent information were distributed electronically via Qualtrics (Qualtrics) 3 times during the evaluation: once in the early phase after project implementation (January-February 2023; n=16), once at the midphase (July 2023; n=13), and once in the late phase after the pilot concluded (October 2023; n=8). Up to 2 reminders were sent via email to eligible staff who did not complete a survey, 1 week and 2 weeks after the initial email invitation. Similar to WIC participant surveys, staff survey items were adapted from valid and reliable tools [48,53-55] to ensure that all the aspects of WIC human milk feeding education, support, and lactation staff-participant interactions were captured in responses.

Additionally, WIC staff and local agency directors at participating agencies were invited to participate in interviews. A semistructured interview guide, informed by the Reach, Effectiveness, Adoption, Implementation, and Maintenance [56] and the Consolidated Framework for Implementation Research (CFIR) [57] implementation frameworks to assess key implementation aspects (eg, relative advantage, compatibility, complexity, and trialability), was used to guide interviews. Interviews were conducted via video call in the early (March 2023, n=6), mid (August-September 2023, n=6), and late phases (October 2023, n=2) of the implementation period. The interviews were scheduled for 1 hour and were digitally recorded and transcribed. Given that there was a small number of eligible staff, the number of interviews at each phase was dictated by the number of staff willing to participate rather than saturation.

Quantitative Data Analysis

Analyses were conducted in Stata (version 18; StataCorp LLC). Descriptive statistics were performed to summarize acceptability, satisfaction, and feasibility (via frequencies and percentages, and means and SDs) among WIC participants and staff.

Qualitative Data Analysis

A thematic analysis [46] of WIC participants and staff interviews was performed using a deductive approach, with a preliminary codebook developed using CFIR [57] and the Evaluation Framework for Telemedicine [58] prior to analysis. A total of 10 trained qualitative researchers coded interview transcripts, with each transcript being independently coded by 2 researchers using NVivo (version 13; Lumivero). Coders met regularly to discuss coding disagreements, reach consensus, and update the codebook, if needed.

Ethical Considerations

For research activities in stage 3, The Vermont Agency of Human Services IRB reviewed and approved the protocol related

XSL•FO
RenderX

to the collection of WIC participant data, including surveys, metadata from Daily Drop, and WIC's management information system data (#308). The Tufts University IRB reviewed and approved the protocols related to WIC staff data, including surveys and qualitative interviews (#00003228). All individuals who took part in these research activities provided written informed consent, and all procedures were performed in accordance with the ethical standards of the approving IRBs and the 2013 Declaration of Helsinki.

WIC participants who used the app received a US $50 gift card for completing the baseline survey, a US $50 gift card for completing the 1-month survey, and a US $50 gift card for completing the qualitative interviews. WIC participants who did not use the app and completed the nonuser survey received a US $10 gift card. Incentives were not offered for the 3-month postpartum survey. WIC staff did not receive incentives for completing surveys or interviews, in compliance with federal and/or state policies.

Results

WIC Participants

The results of the 1-month user survey indicated that app users enjoyed using it (35/49, 71.4%), appreciated its simplicity (46/49, 93.9%), and improved their breastfeeding learning experience (39/49, 79.6%; Table 2). Users also expressed satisfaction with the app, with 71.4% (35/49) indicating playing Daily Drop was a good use of their time. About 89% (44/49) suggested WIC offices should use Daily Drop. About two-thirds (33/49, 67.4%) of respondents stated that Daily Drop allowed them to learn about breastfeeding more quickly and easily. Qualitative interviews largely supported these findings, with users discussing that it was "fun," "helpful," and "trustworthy." Users mentioned that the app increased their self-efficacy: "It gave me more confidence in my ability because I had no idea how to breastfeed or if I was doing it right or even to know like if my baby was full. It gave a lot of knowledge that I didn't have before."

However, some users noted that the app might be less beneficial for individuals with previous human milk feeding experience, as the information could feel repetitive and may not provide as much value compared to first time parents: "I don't think I would share it with somebody that's had seven kids and breastfed all of them, but somebody that's new to breastfeeding or new to just motherhood. I would share it because, like I said, there's a lot of information on there that you can learn from it." Lastly, the absence of push notification features and the inability to play more than one playlist per day were highlighted as barriers to using the app: "My only issue was remembering to do it because it didn't have any reminders, or if it did, I didn't have them on."

Table 2. Acceptability and satisfaction with Daily Drop among WIC[a] participants who completed the 1-month user survey (n=49; stage 3, January-September 2023).

Construct and statement	Strongly agree or agree, n (%)	Neither agree or disagree, n (%)	Strongly disagree or disagree, n (%)
Acceptability			
I had more fun learning about breastfeeding because of using Daily Drop.	35 (71.4)	11 (22.4)	3 (6.1)
Using Daily Drop made learning about breastfeeding a better experience than it would have been otherwise.	39 (79.6)	8 (16.3)	2 (4)
I would like to play the Daily Drop more.	29 (59.2)	13 (26.5)	7 (14.3)
I learned about breastfeeding more quickly and easily because of using Daily Drop.	33 (67.4)	14 (28.6)	2 (4)
Playing Daily Drop made me want to use more WIC breastfeeding support resources.	23 (46.9)	18 (36.7)	8 (16.3)
Using Daily Drop made what I was learning from my IBCLC[b] or peer counselor feel more relevant to me.	16 (32.7)	30 (61.2)	3 (6.1)
My WIC staff person provided helpful feedback based on my Daily Drop scores.	12 (24.5)	20 (40.8)	17 (34.7)
Satisfaction			
I am glad I played Daily Drop	38 (77.5)	9 (18.4)	2 (4)
Using the Daily Drop was a good use of my time.	35 (71.4)	11 (22.4)	3 (6.1)
I would recommend Daily Drop to other WIC participants.	41 (83.7)	6 (12.2)	2 (4)
I think other WIC offices should offer Daily Drop.	44 (89.8)	4 (8.2)	1 (2)
Ease of Use			
Daily Drop was simple for me to use.	46 (93.9)	2 (4.1)	1 (2)

[a]WIC: Special Supplemental Nutrition Program for Women, Infants, and Children.

[b]IBCLC: International Board–Certified Lactation Consultant.

XSL•FO
RenderX

WIC Staff

WIC staff agreed that Daily Drop was an acceptable way to provide WIC services (Table 3). The average agreement score for the usefulness of the app declined over time, but the difference was not significant. Mean interest scores in using Daily Drop feedback for breastfeeding support were neutral in the midphase and low in the late phase. However, feasibility measures declined steadily over time. Early phase scores for learning gameplay feedback on breastfeeding education approached agreement on ease of use, while late-phase scores indicated disagreement. Average scores for understanding reports and experience with Daily Drop reports showed low feasibility for providing client feedback in the clinic.

These findings were further supported during qualitative interviews, where WIC staff mentioned that the app would benefit WIC participants due to its interactive features, which help deliver lactation information in an engaging manner. Staff also appreciated the self-paced nature of the app, which eased the burden of information overload often experienced during lactation support appointments:

> *We're really excited about having another way to present breastfeeding information to parents...it's user-friendly, and it's tailoring to what families are, what information they're looking for.* [Early phase]

Furthermore, staff indicated that BAPT scores and gameplay reports allowed them to provide even more tailored discussions and conversations based on the participants' needs:

> *We use the (app) data mostly at the mid-certification visit for pregnant participants... We also use it to check our BAPT survey, which is done on the Daily Drop app also separately. And mostly it's used as a tool to then follow up about topics that are either of peak interest or topics that don't seem to be as much on a family's radar.* [Midphase]

Nevertheless, during mid- and late-phase interviews, staff noted that more comprehensive training on using the app and integrating gameplay in the lactation appointments would have made it more likely for them to use it. Staff's concerns were less about the app's design and more about its implementation and their readiness to incorporate it into their routine WIC appointments:

> *And I do think that if we continue to use it, that at least I would get the hang of both reviewing it, reviewing those reports ahead of the appointments and incorporating it into the actual appointment itself pretty effectively.* [Late phase]

Table 3. Feasibility and acceptability of Daily Drop among WIC[a] staff who completed the early- (March 2023), mid- (August-September 2023), and late-phase (October 2023) surveys (stage 3).

Construct and statement	Early phase (n=14), mean (SD)[b]	Midphase (n=6), mean (SD)[b]	Late phase (n=3), mean (SD)[b]
Acceptability			
Daily Drop is an acceptable way to provide WIC breastfeeding education.	4.4 (0.6)	3.7 (0.5)	3.3 (0.6)
Daily Drop reports are useful for me as WIC staff.	4.3 (0.7)	3.0 (1.3)	2.7 (1.5)
I would like to continue using Daily Drop gameplay feedback in my breastfeeding support sessions.	N/A[b]	3.0 (1.1)	2.0 (1.0)
Feasibility			
Learning to use gameplay feedback for breastfeeding education and support from Daily Drop was easy for me.	3.8 (1.0)	3.0 (1.1)	2.0 (1.7)
I find the Daily Drop reports easy to understand.	3.9 (0.9)	2.8 (1.3)	3.0 (1.0)
I have a lot of experience giving Daily Drop gameplay feedback in my clinic.	N/A	1.5 (0.5)	1.3 (0.6)
I am good at giving Daily Drop gameplay feedback in my clinic appointments.	N/A	2.5 (1.4)	1.7 (1.2)

[a]WIC: Special Supplemental Nutrition Program for Women, Infants, and Children.

[b]Mean and SD are from 5-point Likert scales, where 1=strongly disagree and 5=strongly agree.

[b]N/A: not applicable.

Discussion

Overview

This manuscript describes the development and testing of a gamified, human milk feeding education app designed for pregnant and postpartum WIC participants. Consistent with the agile development approach [27], the core project team integrated mixed methods feedback from key partner groups, including WIC participants and staff, at all stages of app development (concept development, service design, and preimplementation [31]). Active user involvement is critical at these stages, as it can increase the perceived value of the mHealth product and future successful uptake [59]. This iterative, cyclical process can help ensure high-quality mHealth software is usable and acceptable for end users, increasing its

XSL•FO
RenderX

potential to improve health-related behaviors and outcomes [60,61]. Furthermore, using a "serious game" design [30] enabled the transformation of extensive content into easily digestible, minigame formats—an innovative approach within the WIC context, compared to traditional methods of delivering lactation education and support.

Key feedback from WIC participants and staff included providing easily accessible educational information throughout gameplay, diversifying progress tracking to emphasize knowledge growth and expertise development, and including supportive or affirming messages to users, among others. This feedback was incorporated before the field-testing stage of development, and preliminary results from mixed methods evaluations with WIC participants and staff suggested that the app was acceptable, user-friendly, and engaging, and that the lactation information provided was easier to understand compared to more traditional education materials.

These findings align with previous research evaluating implementation outcomes (eg, acceptability, feasibility, and engagement) of mHealth apps targeting human milk feeding behaviors among the general and low-income populations. Several systematic reviews have shown that such apps are highly acceptable and have moderate to high levels of engagement [23-26]. Nevertheless, apps had varying levels of effectiveness in improving human milk feeding behaviors, often attributed to the heterogeneity of the apps' engagement, whether the app focused on the provision of information vs interactive support (eg, from lactation consultants or peers) vs monitoring or tracking behaviors, and participant characteristics (eg, otherwise healthy pregnant individuals vs those with pregnancy complications). While this manuscript does not describe behavioral outcomes, it is worth emphasizing that apps may not be as effective in evoking behavior change in the absence of multipronged strategies to reduce the structural barriers to human milk feeding [10,62], particularly those faced by parents with low income. Given WIC's eligibility requirement of medical or nutrition-related risks, participants may face additional or more severe human milk feeding challenges compared to income-eligible populations. Therefore, mHealth strategies seamlessly integrated into WIC services, paired with already established initiatives (lactation professionals, classes and support groups, peer counselors, among others [63]), have the potential to achieve higher reach and scalability among those who may need it most, and increase the program's impact on reducing income, racial, and ethnic disparities in human milk feeding outcomes.

A recent systematic review showed that WIC participants consider mHealth diet interventions (eg, apps) highly acceptable and feasible, although no app focused solely on human milk feeding behaviors [64]. Furthermore, in various states, apps focusing on nutrition or lactation education can also be used to fulfill nutrition education requirements instead of attending in-person classes, further reducing transportation and scheduling-related barriers to WIC participation. In sum, our findings contribute to the growing evidence supporting the increased use of technologies within the WIC context to improve access to the program and its services and potentially improve health-related outcomes, ensuring WIC's continued success in promoting the health and well-being of its participants.

Limitations

The findings of this research must be interpreted in the context of the following limitations. First, the core project team used a convenience sampling approach, and sample sizes across all development and testing stages were fairly small (n<75), especially for WIC staff (n<14); thus, results may not be as generalizable to the larger VT WIC participant population or lactation staff. Nevertheless, this was proof-of-concept research, and preliminary results demonstrated the acceptability and usability of the app among WIC participants and staff. Therefore, a necessary next step for our research is to better understand how effective the app is at improving human milk feeding behaviors compared to usual care (eg, electronic or printed handouts) with a larger, more diverse sample of WIC participants.

Second, while survey items were adapted to the WIC context from valid and reliable tools and reviewed by subject matter experts, they were not validated with the target population, further limiting their generalizability. Additionally, the use of multiple assessment tools with varying Likert scale formats may introduce complexity in cross-measure comparisons. Thus, as mobile apps become more widely integrated into WIC routine care as a result of ongoing modernization efforts [22], future research should focus on developing and validating tools to standardize implementation outcomes in this context.

Lastly, the app was developed only in English. While 94% of VT WIC households list English as their preferred spoken language [50], future work is needed to ensure appropriate cultural and linguistic adaptations in the app to ensure fair access for all WIC participants.

Lessons Learned

Several areas for improvement and lessons learned emerged from our process. From a logistics perspective, for software-related projects specifically for federal nutrition assistance programs, it is critical to ensure sufficient time is planned to procure and establish a contract with a vendor with expertise in leading a user-centered design project, as technical procurements within federally funded agencies involve additional levels of internal review to ensure security standards are met. Furthermore, developing and maintaining strong relationships with internal IT staff is crucial for implementing the necessary components within the government environment. These include facilitating the technical procurement review process, meeting government security standards, and establishing data flow and storage of game use data while preserving user privacy. Lastly, having a dedicated project management staff to coordinate with the vendor and internal government partners, assist with recruitment, and coordinate testers in each stage was key to our success.

From the WIC staff's perspective, our results suggest that comprehensive training is crucial for equipping staff to demo the app to participants, access and interpret game reports, and use them during lactation support appointments, all while ensuring that these tasks do not substantially increase staff

XSL•FO
RenderX

burden, which is critical for staff buy-in. Moreover, from the WIC participants' perspective, engagement data suggested that users interacted with the app for a particular duration and reported experiencing benefits. Thus, explicitly communicating a specific time frame for app use (eg, how much, how often, and for how long) can help users feel motivated and perceive more benefits, potentially increasing the app's impact on human milk feeding behaviors.

Implications for Future Research and Practice

With the ongoing WIC modernization efforts [22], this research can serve as a guide for other state and local WIC agencies wanting to incorporate technology solutions into their routine practices. The co-design, user-centered approach involving WIC leadership, staff, participants, and software developers resulted in an acceptable and user-friendly app. It is worth noting that another key advantage of agile development and end-user feedback during the early stages relates to cost savings, particularly for government agencies, like WIC. "Traditional" software development usually involves a "waterfall" or "top-down" approach, during which end-user feedback is not incorporated until the last phase—when the software is operating in a live environment [28]. Contrarily, by gathering end-user feedback early and throughout all stages of development, and providing interactive materials (eg, early prototypes of the app) instead of "hypothetical" questions (eg, "what would you like to see in the app"?), software developers can ensure that the final product is closely aligned with end-user requirements, thus increasing the likelihood of greater acceptability and engagement and decreasing the need for costly updates once the product is "live." Nevertheless, further research evaluating the cost-effectiveness of these technologies is warranted.

Conclusions

Daily Drop is a gamified mobile app that provides human milk feeding education and support specifically for WIC participants. Using an agile development and co-design approach involving key partner groups at all stages of development, Daily Drop was deemed acceptable, usable, and engaging by end users. Future research could focus on statewide implementation, testing in a randomized controlled trial, and evaluating the cost-effectiveness of the app. The widespread implementation of mHealth solutions in the WIC program is a vital step forward into the increasingly digital health landscape, to support WIC's mission of improving the health and well-being of families with low income.

Acknowledgments

The authors would like to thank Noah Praamsma for his support in program management throughout the US Department of Agriculture/Tufts Telehealth Intervention Strategies for Special Supplemental Nutrition Program for Women, Infants, and Children (THIS-WIC) project period; the THIS-WIC team for their guidance and support in all stages of the project; Weathered Sweater for their contributions during software development; the advisory council for their enthusiasm for this project and valuable feedback; and testers and study participants for taking part in this study.

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing. The GenAI tool used was Grammarly AI. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was supported by a subaward from the US Department of Agriculture (USDA)/Tufts Telehealth Intervention Strategies for Special Supplemental Nutrition Program for Women, Infants, and Children (THIS-WIC), a cooperative agreement between the USDA Food and Nutrition Service and Tufts University (cooperative agreement number #OPS-WIC Telehealth-FY19-Tufts-MA). The findings and conclusions in this manuscript are those of the authors and should not be construed to represent any official USDA or US Government determination or policy.

Data Availability

The datasets generated and analyzed during this study are not publicly available. Data were collected and shared between the Trustees of Tufts College and the State of Vermont WIC under a Collaborative Data Transfer and Use Agreement that classifies all shared data as deidentified data about human subjects, with no personally identifiable information per the National Institute of Standards and Technology Special Publication 800-122 and HIPAA (Health Insurance Portability and Accountability Act) deidentification standards. The agreement does not include public access requirements for the data. Deidentified quantitative data (survey responses and aggregated app usage metadata) are stored on HIPAA-compliant servers and in password-protected institutional repositories, and may be made available upon reasonable request to the corresponding author, subject to a data use agreement and approval from the relevant institutional review boards (Vermont Agency of Human Services; Tufts University). Pursuant to the US Department of Agriculture (USDA) cooperative agreement governing this project, the USDA Food and Nutrition Service retains the right to obtain, reproduce, and publish generated data, including in response to Freedom of Information Act requests. Qualitative interview transcripts are not available due to the risk of participant reidentification given the small sample sizes and rural study setting. The Daily Drop app was developed by GameTheory Co as proprietary software; source code is not publicly available.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: AM, CW, PC, EH
Formal analysis: YMH, NQ
Funding acquisition: EH
Investigation: SM, AM, MD, PC
Software: SM, MD
Supervision: AM, EH
Visualization: YMH, SM, NQ
Writing — original draft: YMH
Writing — review & editing: SM, AM, CW, MD, PC, NQ, EH

Conflicts of Interest

YMH, AM, CW, PC, NQ, and EH have no conflicts of interest to declare. SM and MD are the Chief Operating Officer and Chief Executive Officer, respectively, of GameTheory Co and were involved in the development of the mobile app described in this manuscript. However, they were not involved in the field testing of the app to ensure that the data collection and analysis remained unbiased and solely focused on scientific integrity and Special Supplemental Nutrition Program for Women, Infants, and Children–related outcomes associated with US Department of Agriculture's funding.

References

1. Kinney R, Praamsma N, Malinowski A, Cassi P, Hennessy E. Testing inclusive language revisions of the breastfeeding attrition prediction tool using cognitive interviewing: a pilot study. J Hum Lact 2023;39(3):529-539. [doi: 10.1177/08903344231174221] [Medline: 37272601]
2. Meek JY, Noble L, Section on Breastfeeding. Policy statement: breastfeeding and the use of human milk. Pediatrics 2022;150(1):e2022057988. [doi: 10.1542/peds.2022-057988] [Medline: 35921640]
3. U.S. Department of Agriculture and U.S. Department of Health and Human Services. Dietary Guidelines for Americans, 2020-2025. 2024. URL: https://www.dietaryguidelines.gov [accessed 2024-09-24]
4. Infant and young child feeding. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/infant-and-young-child-feeding [accessed 2026-01-20]
5. NIS-child data results. Breastfeed data. CDC. 2024. URL: https://www.cdc.gov/breastfeeding-data/survey/results.html [accessed 2025-03-12]
6. Increase the proportion of infants who are breastfed exclusively through age 6 months-Data. Healthy People 2030. URL: https://odphp.health.gov/healthypeople/objectives-and-data/browse-objectives/infants/increase-proportion-infants-who-are-breastfed-exclusively-through-age-6-months-mich-15/data [accessed 2025-03-27]
7. Increase the proportion of infants who are breastfed at 1 year- MICH-16. Healthy People 2030. URL: https://odphp.health.gov/healthypeople/objectives-and-data/browse-objectives/infants/increase-proportion-infants-who-are-breastfed-1-year-mich-16 [accessed 2025-03-27]
8. Sriraman NK, Kellams A. Breastfeeding: what are the barriers? Why women struggle to achieve their goals. J Womens Health (Larchmt) 2016;25(7):714-722. [doi: 10.1089/jwh.2014.5059] [Medline: 27111125]
9. Odom EC, Li R, Scanlon KS, Perrine CG, Grummer-Strawn L. Reasons for earlier than desired cessation of breastfeeding. Pediatrics 2013;131(3):e726-e732 [FREE Full text] [doi: 10.1542/peds.2012-1295] [Medline: 23420922]
10. Barriers to breastfeeding in the United States. The surgeon general's call to action to support breastfeeding. In: Office of the Surgeon General (US). Rockville, MD: Centers for Disease Control and Prevention (US); 2011.
11. Special Supplemental Nutrition Program for Women, Infants, and Children (WIC). USDA-FNS. URL: https://www.fns.usda.gov/wic [accessed 2021-07-19]
12. Thoma ME, De Silva DA, Kim J, Hodges L, Guthrie J. Breastfeeding initiation trends by Special Supplemental Nutrition Program for Women, Infants, and Children participation and race/ethnicity among medicaid births. J Nutr Educ Behav 2023;55(3):170-181 [FREE Full text] [doi: 10.1016/j.jneb.2022.09.006] [Medline: 36642586]
13. The state of WIC: celebrating 50 years of impact. National WIC Association. URL: https://media.nwica.org/state%20of%20wic%20report%202025.pdf [accessed 2025-03-07]
14. Internet, broadband fact sheet. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/fact-sheet/internet-broadband/ [accessed 2025-03-27]
15. Lau Y, Htun TP, Tam WSW, Klainin-Yobas P. Efficacy of e-technologies in improving breastfeeding outcomes among perinatal women: a meta-analysis. Matern Child Nutr 2016;12(3):381-401 [FREE Full text] [doi: 10.1111/mcn.12202] [Medline: 26194599]
16. Demirci JR, Cohen SM, Parker M, Holmes A, Bogen DL. Access, use, and preferences for technology-based perinatal and breastfeeding support among childbearing women. J Perinat Educ 2016;25(1):29-36 [FREE Full text] [doi: 10.1891/1058-1243.25.1.29] [Medline: 26848248]

XSL•FO
RenderX

17. Dunn C, Kenney E, Bleich S, Fleischhacker S. Strengthening WIC's impact during and after the COVID-19 pandemic. Healthy Eating Research. 2020. URL: https://healthyeatingresearch.org/research/strengthening-wics-impact-during-and-after-the-covid-19-pandemic/ [accessed 2026-03-21]
18. Zimmer M, McElrone M, Anderson Steeves ET. J Acad Nutr Diet 2021;121(12):2464-2474.e1. [doi: 10.1016/j.jand.2021.05.015] [Medline: 34219049]
19. Morris ME, Aguilera A. Mobile, social, and wearable computing and the evolution of psychological practice. Prof Psychol Res Pr 2012;43(6):622-626 [FREE Full text] [doi: 10.1037/a0029041] [Medline: 25587207]
20. Federal nutrition programs during the COVID-19 pandemic: Special Supplemental Nutrition Program for Women, Infants, and Children (WIC). Healthy Eating Research. 2021. URL: https://healthyeatingresearch.org/wp-content/uploads/2021/12/HER_WICSummaryBrief_121521.pdf [accessed 2025-03-07]
21. The state of WIC: investing in the next generation. National WIC Association. 2022. URL: https://thewichub.org/the-state-of-wic-investing-in-the-next-generation/ [accessed 2026-03-21]
22. WIC modernization. USDA-FNS. 2024. URL: https://www.fns.usda.gov/wic/modernization [accessed 2024-07-17]
23. Ziebart M, Kammermeier M, Koletzko B, Patro-Golab B. Mobile applications for promoting and supporting breastfeeding: systematic review and meta-analysis. Matern Child Nutr 2025;21(1):e13733. [doi: 10.1111/mcn.13733] [Medline: 39390945]
24. Griffin LB, López JD, Ranney ML, Macones GA, Cahill AG, Lewkowitz AK. Effect of novel breastfeeding smartphone applications on breastfeeding rates. Breastfeed Med 2021;16(8):614-623 [FREE Full text] [doi: 10.1089/bfm.2021.0012] [Medline: 33826418]
25. Eppes EV, Augustyn M, Gross SM, Vernon P, Caulfield LE, Paige DM. Engagement with and acceptability of digital media platforms for use in improving health behaviors among vulnerable families: systematic review. J Med Internet Res 2023;25:e40934 [FREE Full text] [doi: 10.2196/40934] [Medline: 36735286]
26. Almohanna AA, Win KT, Meedya S. Effectiveness of internet-based electronic technology interventions on breastfeeding outcomes: systematic review. J Med Internet Res 2020;22(5):e17361 [FREE Full text] [doi: 10.2196/17361] [Medline: 32469315]
27. USDS. What is Agile software development. TechFAR Hub Handbook. URL: https://techfarhub.usds.gov/pre-solicitation/agile-overview/ [accessed 2025-03-28]
28. Islam AKMZ, Ferworn DA. A comparison between agile and traditional software development methodologies. Glob J Comput Sci Technol 2020;20(2):7-42. [doi: 10.34257/gjcstcvol20is2pg7]
29. 2023 WIC technology landscape report: using technology to improve the enrollment and certification experience for participants and WIC agencies. Nava Public Benefit Corporation, National WIC Association. 2023. URL: https://media.nwica.org/2023%20wic%20tech%20report%204.5.23.pdf [accessed 2025-04-10]
30. Maxim RI, Arnedo-Moreno J. Identifying key principles and commonalities in digital serious game design frameworks: scoping review. JMIR Serious Games 2025;13:e54075 [FREE Full text] [doi: 10.2196/54075] [Medline: 40053743]
31. Fatehi F, Smith AC, Maeder A, Wade V, Gray LC. How to formulate research questions and design studies for telehealth assessment and evaluation. J Telemed Telecare 2017;23(9):759-763. [doi: 10.1177/1357633X16673274] [Medline: 29070001]
32. Ten Klooster I, Wentzel J, Sieverink F, Linssen G, Wesselink R, van Gemert-Pijnen L. Personas for better targeted eHealth technologies: user-centered design approach. JMIR Hum Factors 2022;9(1):e24172 [FREE Full text] [doi: 10.2196/24172] [Medline: 35289759]
33. Hurde J, McCollough A, Chalil Madathil S, Khasawneh M. Enhancing healthcare with patient personas: a systematic review. J Healthc Risk Manag 2025;45(2):33-44. [doi: 10.1002/jhrm.70014] [Medline: 41063419]
34. Dominello A, Cheek C, Clay-Williams R. The use of personas to involve consumers in healthcare co-design: a scoping review. BMC Health Serv Res 2025;25(1):649 [FREE Full text] [doi: 10.1186/s12913-025-12786-7] [Medline: 40329348]
35. Amna AR, Poels G. Systematic literature mapping of user story research. IEEE Access 2022;10:51723-51746. [doi: 10.1109/access.2022.3173745]
36. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2026-04-29]
37. Hamm MJ. Wireframing Essentials. Birmingham, UK: Packt Publishing Ltd; 2014.
38. Janke JR. Prediction of breast-feeding attrition: instrument development. Appl Nurs Res 1992;5(1):48-53. [doi: 10.1016/s0897-1897(05)80086-2] [Medline: 1570961]
39. Bortree L, Decher L, Flynn K. Supporting long-term breastfeeding with the new WIC food packages: you can do it/WIC can help. USDA. Burlington, VT: Vermont Agency of Human Services; 2013. URL: https://wicworks.fns.usda.gov/sites/default/files/media/document/VT_WIC-SPG_Final-Report.pdf [accessed 2025-04-18]
40. Yee N. 7 things we learned about primary gaming motivations from over 250,000 gamers. Quantic Foundry. 2016. URL: https://quanticfoundry.com/2016/12/15/primary-motivations/ [accessed 2026-02-05]
41. Hai-Jew S. Alpha testing, beta testing, and customized testing. In: Designing Instruction for Open Sharing. Cham: Springer International Publishing; 2019:381-428.
42. Ahonen EQ, Watson DP, Adams EL, McGuire A. Alpha test results for a housing first eLearning strategy: the value of multiple qualitative methods for intervention design. Pilot Feasibility Stud 2017;3:46 [FREE Full text] [doi: 10.1186/s40814-017-0187-y] [Medline: 29093826]

XSL•FO
RenderX

43. Silva Farias R, Ahmed I, de Almeida ES. What makes a great software quality assurance engineer? IIEEE Trans Softw Eng 2025;51(4):1153-1172. [doi: 10.1109/tse.2025.3542763]
44. Chemuturi M. Mastering Software Quality Assurance: Best Practices, Tools and Techniques for Software Developers. Plantation, Florida, United States: J Ross Publishing; 2010.
45. Hossain M. Challenges of software quality assurance and testing. IJSECS 2018;4(1):133-144. [doi: 10.15282/ijsecs.4.1.2018.11.0044]
46. Clarke V, Braun V. Thematic analysis. J Posit Psychol 2016;12(3):297-298. [doi: 10.1080/17439760.2016.1262613]
47. Jones SR, Torres V, Arminio J. Negotiating the Complexities of Qualitative Research in Higher Education: Essential Elements and Issues. New York: Routledge; 2021.
48. Parmanto B, Lewis AN, Graham KM, Bertolet MH. Development of the Telehealth Usability Questionnaire (TUQ). Int J Telerehabil 2016;8(1):3-10 [FREE Full text] [doi: 10.5195/ijt.2016.6196] [Medline: 27563386]
49. Agha Z, Schapira RM, Laud PW, McNutt G, Roter DL. Patient satisfaction with physician-patient communication during telemedicine. Telemed J E Health 2009;15(9):830-839. [doi: 10.1089/tmj.2009.0030] [Medline: 19919189]
50. Au LE, Whaley S, Rosen NJ, Meza M, Ritchie LD. Online and in-person nutrition education improves breakfast knowledge, attitudes, and behaviors: a randomized trial of participants in the Special Supplemental Nutrition Program for Women, Infants, and Children. J Acad Nutr Diet 2016;116(3):490-500. [doi: 10.1016/j.jand.2015.10.012] [Medline: 26669795]
51. Nestor B, McKenzie J, Hasan N, AbuSabha R, Achterberg C. Client satisfaction with the nutrition education component of the California WIC program. J Nutr Educ 2001;33(2):83-94. [doi: 10.1016/s1499-4046(06)60171-1] [Medline: 12031188]
52. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: a systematic review of empirical tests. Soc Sci Med 2022;292:114523 [FREE Full text] [doi: 10.1016/j.socscimed.2021.114523] [Medline: 34785096]
53. Rho MJ, Choi IY, Lee J. Predictive factors of telemedicine service acceptance and behavioral intention of physicians. Int J Med Inform 2014;83(8):559-571. [doi: 10.1016/j.ijmedinf.2014.05.005] [Medline: 24961820]
54. Gao S, Krogstie J, Siau K. Developing an instrument to measure the adoption of mobile services. Mob Inf Syst 2011;7(1):45-67. [doi: 10.1155/2011/831018]
55. Shiferaw KB, Mengiste SA, Gullslett MK, Zeleke AA, Tilahun B, Tebeje T, et al. Healthcare providers' acceptance of telemedicine and preference of modalities during COVID-19 pandemics in a low-resource setting: an extended UTAUT model. PLoS One 2021;16(4):e0250220 [FREE Full text] [doi: 10.1371/journal.pone.0250220] [Medline: 33886625]
56. Glasgow RE, Harden SM, Gaglio B, Rabin B, Smith ML, Porter GC, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. Front Public Health 2019;7:64 [FREE Full text] [doi: 10.3389/fpubh.2019.00064] [Medline: 30984733]
57. Keith RE, Crosson JC, O'Malley AS, Cromp D, Taylor EF. Using the Consolidated Framework for Implementation Research (CFIR) to produce actionable findings: a rapid-cycle evaluation approach to improving implementation. Implement Sci 2017;12(1):15 [FREE Full text] [doi: 10.1186/s13012-017-0550-7] [Medline: 28187747]
58. Chang H. Evaluation framework for telemedicine using the logical framework approach and a fishbone diagram. Healthc Inform Res 2015;21(4):230-238 [FREE Full text] [doi: 10.4258/hir.2015.21.4.230] [Medline: 26618028]
59. Støme LN, Wilhelmsen CR, Kværner KJ. Enabling guidelines for the adoption of eHealth solutions: scoping review. JMIR Form Res 2021;5(4):e21357 [FREE Full text] [doi: 10.2196/21357] [Medline: 33929330]
60. Obigbesan O, Graham K, Benzies KM. Software testing of eHealth interventions: existing practices and the future of an iterative strategy. JMIR Nurs 2024;7:e56585 [FREE Full text] [doi: 10.2196/56585] [Medline: 39028552]
61. Wilson K, Bell C, Wilson L, Witteman H. Agile research to complement agile development: a proposal for an mHealth research lifecycle. NPJ Digit Med 2018;1:46 [FREE Full text] [doi: 10.1038/s41746-018-0053-1] [Medline: 31304326]
62. Pérez-Escamilla R, Tomori C, Hernández-Cordero S, Baker P, Barros AJD, Bégin F, 2023 Lancet Breastfeeding Series Group. Breastfeeding: crucially important, but increasingly challenged in a market-driven world. Lancet 2023;401(10375):472-485 [FREE Full text] [doi: 10.1016/S0140-6736(22)01932-8] [Medline: 36764313]
63. How the WIC program supports breastfeeding. National WIC Association. URL: https://media.nwica.org/breastfeeding_infographic2018.pdf [accessed 2023-10-25]
64. Crespo-Bellido M, Fernandez Ong J, Yaroch A, Byker Shanks C. E-health dietary interventions for participants of SNAP and WIC: a systematic review. Curr Dev Nutr 2024;8(4):102099 [FREE Full text] [doi: 10.1016/j.cdnut.2024.102099] [Medline: 38706511]

Abbreviations

BAPT: Breastfeeding Attrition Prediction Tool
CFIR: Consolidated Framework for Implementation Research
HIPAA: Health Insurance Portability and Accountability Act
IRB: institutional review board
mHealth: mobile health
THIS-WIC: USDA/Tufts Telehealth Intervention Strategies for WIC
USDA: US Department of Agriculture

XSL•FO
RenderX

VT WIC: Vermont State WIC agency
WIC: Special Supplemental Nutrition Program for Women, Infants, and Children

Edited by A Stone; submitted 08.Jul.2025; peer-reviewed by L Furman, P Reimers; comments to author 06.Jan.2026; accepted 09.Mar.2026; published 29.Apr.2026.

Please cite as:
Melo Herrera Y, Mitchell S, Malinowski A, Wright C, Dibble M, Cassi P, Qayyum N, Hennessy E
Agile Development and Testing of a Gamified Human Milk Feeding Education Mobile App for Participants of the Special Supplemental Nutrition Program for Women, Infants, and Children: Co-Design Approach
J Med Internet Res 2026;28:e80330
URL: https://www.jmir.org/2026/1/e80330
doi:10.2196/80330
PMID:

©Yarisbel Melo Herrera, Shannon Mitchell, Amy Malinowski, Catherine Wright, Marguerite Dibble, Patricia Cassi, Naina Qayyum, Erin Hennessy. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Posttraumatic Symptoms as Predictors of Engagement With a Mobile App for Coping After Military Sexual Trauma: Public Usage Data Analysis Study

Shilpa Hampole[1*], MA; Shannon McCaslin-Rodrigo[1,2*], PhD; Valerie Carr[3*], PhD; David Schuster[3*], PhD; Sofia Reyes[1*], BA; Sarah Senti[1*], PhD; Kelly M Ramsey[1*]; Jason E Owen[1*], MPH, PhD; W Christopher Skidmore[4,5*], PhD; Amy E Street[4,5,6*], PhD; Katherine Taylor[1*], MPH, PsyD; Margaret-Anne Mackintosh[1*], PhD

[1]Dissemination and Training Division, National Center for PTSD, VA Palo Alto Health Care System, Menlo Park, CA, United States

[2]Department of Psychiatry and Behavioral Sciences, School of Medicine, Stanford University, Stanford, CA, United States

[3]Department of Psychology, San José State University, San Jose, CA, United States

[4]Office of Mental Health, Veterans Health Administration, Washington, DC, United States

[5]VA Boston Healthcare System, Boston, MA, United States

[6]Department of Psychiatry, Chobanian & Avedisian School of Medicine, Boston University, Boston, MA, United States

[*]all authors contributed equally

Corresponding Author:
Shilpa Hampole, MA
Dissemination and Training Division
National Center for PTSD
VA Palo Alto Health Care System
MPD-334B201
795 Willow Rd
Menlo Park, CA, 94025
United States
Phone: 1 650 614 9997 ext 24408
Email: shilpa.hampole@va.gov

Abstract

Background: Military sexual trauma (MST) can have significant adverse effects on mental health and well-being, often leading to posttraumatic stress disorder (PTSD) symptoms and maladaptive beliefs. Although effective psychotherapies exist, stigma, confidentiality concerns, and systemic barriers often hinder help-seeking among service members and veterans. Mobile mental health apps offer an accessible and anonymous support alternative, potentially addressing such barriers. However, app effectiveness depends on user engagement and emerging evidence suggests that engagement may be shaped by symptom severity.

Objective: This retrospective observational study aimed to explore the relationship between posttraumatic symptom severity and user engagement with Beyond MST (US Department of Veterans Affairs [VA] National Center for PTSD), an app for individuals who experienced MST. Specific aims included (1) characterizing trauma-related symptom levels and app engagement among users who completed in-app assessments, and (2) evaluating how PTSD symptom severity, negative posttraumatic cognitions, and mental well-being relate to objective measures of engagement.

Methods: Anonymous usage data from 27,517 users collected between March 11, 2021 and July 29, 2024, were analyzed. Three subsamples were identified: those who completed the in-app PTSD checklist for *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [*Fifth Edition*]; PCL-5, n=3689), the Posttraumatic Maladaptive Beliefs Scale (PMBS; n=2197), and the Warwick-Edinburgh Mental Well-Being Scale (WEMWBS; n=2160). Engagement metrics included duration of use (ie, days of use and minutes of use), frequency of feature access (ie, coping tool and psychoeducation access), and frequency of feature use (ie, total assessment completions). Regression analyses, including quadratic terms, were conducted to evaluate how symptom severity and well-being levels influenced engagement and identify possible curvilinear trends.

Results: Median engagement levels ranged across subsamples as follows: 3-4 days of use (IQR 5-6), 22-30 minutes of use (IQR 33.7-42.9), 1-5 feature accesses (IQR 6-9), and 2-3 assessment completions (IQR 2). Subsamples were highly symptomatic. Analyses revealed that moderate PTSD symptom and negative posttraumatic cognition severity were associated with higher

XSL•FO
RenderX

engagement relative to users with very low and very high symptom levels, particularly for days of use and frequency of coping tool access. Conversely, higher mental well-being scores were generally linked to increased app engagement with linear effects. Effect sizes were small, suggesting limited clinical impact.

Conclusions: This study highlights the possible challenges in engaging highly symptomatic individuals with digital mental health interventions. Although Beyond MST successfully reaches its targeted population, very low or high symptom levels and lower well-being may hinder sustained engagement. These findings suggest that symptom levels should be considered in app development (ie, personalization) and when integrating apps into professional care. Interpretation is limited by the anonymous nature of the data, which prevented characterization of users and their trauma histories. Further research is needed to clarify how symptom patterns influence engagement, especially in trauma contexts.

International Registered Report Identifier (IRRID): RR2-10.31979/etd.882a-5fcx

(J Med Internet Res 2026;28:e85098) doi:10.2196/85098

KEYWORDS

military sexual trauma; posttraumatic stress disorder; trauma; veterans; mental health; mobile app; mHealth; digital health; public health

Introduction

Prevalence of Military Sexual Trauma

Experiences of sexual assault or harassment during military service, known as military sexual trauma (MST) [1], are common and can have lasting negative effects on military service members and veterans, including impacts on military preparedness, postservice readjustment, and long-term mental and physical health. In a national survey of 20,563 recent veterans conducted from 2009 to 2011 by the US Department of Veterans Affairs (VA), approximately 41% of women and 4% of men reported experiences of MST [2]. However, the actual prevalence of MST may be even higher (eg, as high as 74% in women veterans [3]) due to underreporting and the substantial heterogeneity in definitions and methods for assessing MST across studies [4,5].

Implications of MST on Mental Health

MST can have profound mental health implications, often leading to severe psychological distress, poor psychosocial functioning, and reduced quality of life. The most common mental health condition associated with MST is posttraumatic stress disorder (PTSD) [5]. Military personnel are at high risk for developing PTSD due to combat and other traumatic exposures experienced during military service, and experiencing MST further exacerbates this risk [6]. In addition, experiences of MST are associated with other mental health conditions such as substance use [7], depression, and suicidality [8], all of which compound the psychological burden on affected individuals. These findings underscore the critical need for interventions that support individuals coping with the aftereffects of MST.

MST, like other forms of trauma, can also lead to increased dysfunctional beliefs or maladaptive thinking patterns. Research suggests that negative posttraumatic cognitions (NPCs, ie, distorted or maladaptive thoughts and beliefs that develop following trauma) play a crucial role in the mechanisms of action of PTSD [9,10]. Specifically, trauma can disrupt an individual's core beliefs about themselves, others, or the world, which can lead to maladaptive beliefs (eg, "I don't feel safe anywhere anymore" [11]) and loss of meaning in life [9,12]. Even when controlling for other posttraumatic risk factors, NPCs can intensify other PTSD symptoms such as avoidance and hypervigilance [10]. Overall, NPCs can hinder recovery after traumatic events and may be especially problematic following experiences of MST. Notably, MST is frequently associated with feelings of self-stigma, or the internalization of negative stereotypes that can lead to guilt, shame, and self-blame. This can manifest in NPCs such as, "The event happened to me because of the sort of person I am," or "Somebody else would not have gotten into this situation" [13]. The distinctive nature of the experience of MST highlights the importance of tailored, evidence-based approaches that address NPCs and other symptoms of sexual trauma.

Implications of MST on Help-Seeking

Although numerous psychotherapy interventions exist to address symptoms of PTSD following traumatic experiences such as MST (eg, Cognitive Processing Therapy, Prolonged Exposure, and Acceptance and Commitment Therapy) [14], individuals who have experienced sexual trauma face significant barriers to help-seeking. Societal stigma around sexual trauma can translate into fear of judgment in survivors, and victim-blaming and disbelief from others. For example, one study found that higher self-stigma is associated with a lower likelihood of disclosing MST in health care screenings [15]. In addition, the setting in which MST occurs (ie, service members living in close quarters on a long-term basis; hierarchical rank structures) can produce fears around subsequent violence, work retaliation, and ostracism by colleagues [16]. Consequently, negative institutional experiences may prompt confidentiality concerns, feelings of distrust toward providers, and discomfort disclosing MST experiences in health care settings [17-19]. These kinds of barriers to help-seeking, nondisclosure with providers, and underuse of health care services can reduce an individual's opportunities for recovery.

Apps for Self-Managing Mental Health

Fortunately, widely accessible technology-based solutions, such as mental health apps, have the potential to expand access to psychoeducation, coping tools, and resources for self-management purposes [20]. They can also bolster patient autonomy, offer 24/7 support when traditional care is unavailable, and preserve anonymity to mitigate

XSL•FO
RenderX

stigma—benefits that prove especially valuable for individuals coping with sensitive traumatic events.

Increasingly, researchers have demonstrated the effectiveness of mental health apps in improving coping skills, reducing mental health symptoms, and enhancing users' sense of well-being and quality of life across populations and mental health challenges [21]. The PTSD Coach app, developed by the US Department of Veterans Affairs (VA) National Center for PTSD, is one well-researched example that has demonstrated feasibility, acceptability, and effectiveness across 16 studies [22]. Digital mental health technologies for addressing specific trauma types, such as sexual trauma and MST [23,24], have also emerged. For example, apps for sexual violence (especially intimate partner violence) are widely available, but typically serve informational, preventative, or peer support purposes [25-27].

The Beyond MST app (US Department of Veterans Affairs [VA] National Center for PTSD) is uniquely designed to help users self-manage MST-related psychosocial challenges (eg, mental health symptoms, behavioral issues, and relationship challenges), enhance their general well-being, and find hope via evidence-based psychoeducation and coping tools [28,29]. To date, no prior research on digital mental health technologies for individuals who experienced MST, including Beyond MST, has been conducted. While usability and effectiveness studies are important, research is also needed to identify factors that may limit users' engagement with app content and consequently impact their effectiveness [30].

App Engagement

Low levels of user engagement with mental health technologies in real-world, nonresearch settings are a significant problem, as many users discontinue use after only a few interactions. For instance, a review of 93 popular mental health apps revealed a median 15-day user retention rate of just 3.9% (IQR 10.3%) [31]. This pattern mirrors the issue of psychotherapy dropout, which can impede patient recovery; similarly, insufficient engagement may hinder users from achieving their full therapeutic benefit.

While early research on digital mental health technology engagement primarily examined its link to health outcomes, recent efforts have shifted toward uncovering the underlying factors that impede or promote user engagement. These include setting-, intervention-, and user-related predictors. For instance, a large body of research has examined intervention-level factors such as the addition of human coaching, gamification, and persuasive technology [30]. However, user-level predictors—particularly symptom severity—remain underexplored.

A systematic review of 208 studies on digital mental health interventions found that only 19 studies examined the relationship between symptom severity and objective measures of user engagement. Most indicated that higher symptom severity, especially related to depression and fatigue, was linked to lower engagement [32]. Yet, findings are mixed: another review reported that while 2 studies associated greater symptom severity with increased usage, 3 found the opposite [30]. Conflicting results may be due to the range of mental health conditions (eg, depression, anxiety, bipolar disorder, and schizophrenia) and digital interventions (eg, web-based services, telehealth programs, and mobile apps) assessed, as well as diversity in the engagement measures used.

The relationship between symptom severity and mental health app engagement may be further complicated by potential curvilinear patterns [30], yet very few studies have modeled symptom severity as a continuous predictor that includes quadratic (or higher-order) terms, leaving such nonlinear relationships largely untested. There is some evidence of possible nonlinear relationships between depression symptoms and mental health app engagement, with moderate symptom levels associated with relatively higher engagement, and mild and severe symptom levels associated with relatively lower engagement (eg, [33]). Possible inverted-U patterns may reflect varying barriers across symptom levels—for instance, low perceived need among users with mild symptoms and increased cognitive burden among users with severe symptoms.

Furthermore, research exploring whether posttraumatic symptom severity predicts engagement with digital mental health technologies remains limited. While some evidence suggests that higher PTSD symptoms are associated with greater intentions to use such technologies, self-reported intentions may not translate into actual behavior (eg, [32,34]). This underscores the importance of analyzing objective engagement metrics. For instance, usage data from COVID Coach (US Department of Veterans Affairs [VA] National Center for PTSD)—an app offering coping tools and resources during the COVID-19 pandemic—revealed that users with significant PTSD symptoms were more likely to use the app for more days on average than those with subthreshold symptoms [33]. Taken together, the relationship between symptom severity and engagement with digital health technologies is unclear and may vary based on symptom type and engagement measures used. Further research is needed to delineate the relationships between trauma-related symptom measures and engagement with trauma-focused apps, especially following experiences of MST where no research has been conducted.

Study Objectives

This retrospective observational study is the first to evaluate a mental health app specifically designed for veterans and service members who have experienced MST. Given the limited research in this area, an exploratory approach was taken to (1) characterize trauma-related symptom levels and app engagement among subsamples of users who completed in-app assessments, and (2) for those subsamples, examine the relationship between trauma-related measures and objective app engagement measures. Usage data from a large sample of Beyond MST app users were processed and analyzed. Two primary predictors were examined: PTSD symptom severity and the severity of NPCs. Mental well-being was included as a secondary predictor to serve as a comparator. App engagement was measured via 3 types of variables: measures of duration of app use, frequencies of feature access, and frequency of feature use. Overall, findings may enhance our understanding of trauma-related predictors of

XSL•FO
RenderX

digital mental health engagement among real-world users coping with MST.

Methods

Description of the Beyond MST Mobile App

The Beyond MST app was developed by the VA National Center for PTSD in collaboration with the national VA MST Support Team [28,29]. It is also part of a larger portfolio of mental health apps developed and maintained by the National Center for PTSD [35]. Beyond MST can be used for self-care purposes (eg, to self-manage MST-related symptoms) or in conjunction with professional health care. However, the app is not intended to replace professional therapy or treatment. While app content was curated for individuals who experienced MST, it may also be helpful to those who have experienced other types of sexual trauma outside of military settings (eg, intimate partner violence).

Beyond MST is publicly available and can be downloaded for free from the Google Play and Apple App stores on smartphones or tablets. It was first released in March 2021 and has since undergone ongoing maintenance and app updates. Once downloaded, the app can be used without an internet connection. Additionally, no user account is required. The app is also private and secure, meaning all usage data collected are anonymous, and users can opt out of providing usage data.

Users can access the following four content areas: (1) the "Home" screen with inspirational quotes, user-selected bookmarks, and a programmed conversational feature for app guidance, (2) "Tools" for developing and practicing active coping skills, (3) "Learn" for psychoeducation topics and resources (eg, Veterans Crisis Line), and (4) "Progress" for tracking progress over time through in-app assessments and a goal-setting feature (Figure 1). The Tools section includes 54 coping tools (eg, "Managing Reactions" and "Finding My Values"). The Learn section includes 49 primary psychoeducation topics (eg, "Self-Blame and Anger" and "Relationships and Health"), and 29 subtopics embedded within the primary topics (eg, Types of Support). Features were designed based on Cognitive Behavioral Therapy (eg, Cognitive Processing Therapy) and Mindfulness-Based Therapy (eg, Acceptance and Commitment Therapy) principles and practices. Content is also informed by findings from MST research studies as cited within the "About" section of the app.

Figure 1. Screenshots from the Beyond MST (military sexual trauma) app, version 1.1.1(125), showing the 4 main content areas (Home, Tools, Learn, and Progress) and an example question item screen from an in-app assessment.

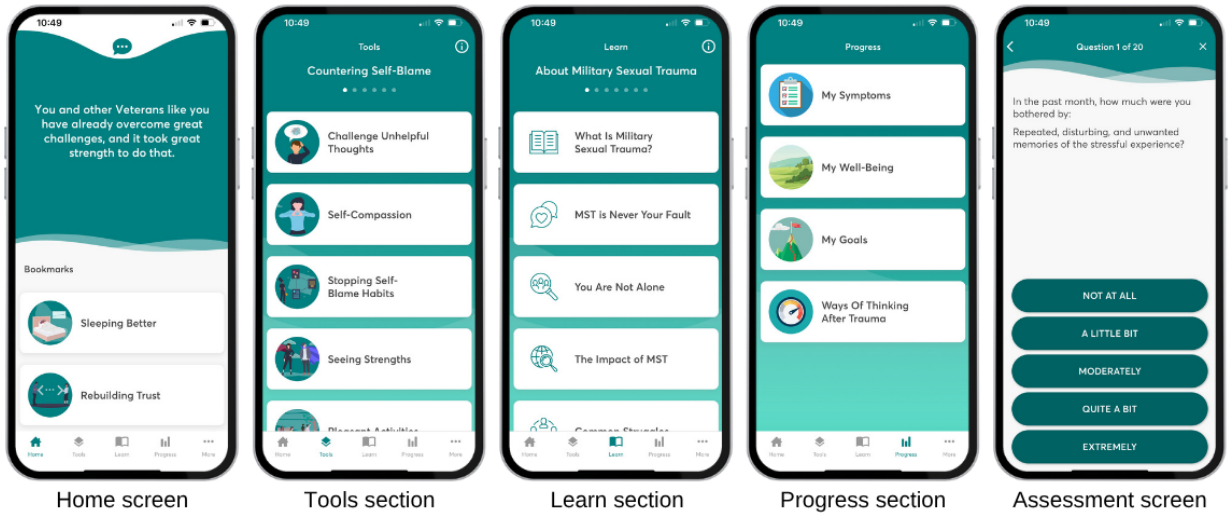

Data Source

JavaScript Object Notation (JSON) formatted usage data were downloaded through a secure data management platform. This platform was developed by the VA National Center for PTSD for quality improvement purposes in accordance with VA's Technical Reference Model [36]. All data collected were anonymous and, thus, did not include any identifying information about users, such as demographic information. The dataset comprised "event-level" data in which each row (or set of rows) of data represented one in-app event (ie, one app interaction, such as a button press or a swipe). The dataset included the following variables: (1) an anonymous computer-generated alphanumeric identification (ID) string assigned when the app was first downloaded, (2) the operating system (ie, Android or iOS), (3) coded event and metadata information describing the interaction (eg, which button was pressed on which screen), and (4) a timestamp indicating the date and time the event occurred.

ID numbers served as proxies for the identification of individual users. It is possible that multiple ID numbers were associated with a single user if they deleted and redownloaded the app, reset the app, or used the app on multiple devices. Given the inability to differentiate between actual users due to anonymity, and the likely rarity of such occurrences, unique ID numbers represented individual users for the purposes of this study.

Procedure

Beyond MST usage data from March 11, 2021 through July 29, 2024, were downloaded on July 30, 2024. Data were cleaned, processed, and analyzed using R statistical software (v4.4.1; R Core Team). As part of data cleaning, errors in the raw

XSL•FO
RenderX

event-level dataset were identified via descriptive statistics, graphical visualization, and simple univariate and bivariate analyses. Research and testing data were removed. Problematic data (eg, due to technical issues) were removed or corrected. Next, a data dictionary mapping user-app interaction data was constructed to guide summarization of the event-level data into user-level variables with one row per user. User-level variables captured total scores on in-app assessments, duration of app use, frequency of feature access, and frequency of feature use.

Measures

In-App Assessments

Three measures were examined based on available usage data from the three assessments found in the Progress section of the app: (1) severity of PTSD symptoms, (2) severity of negative posttraumatic cognitions that are associated with PTSD symptoms, and (3) level of mental well-being.

Primary Measure: Severity of PTSD Symptoms

Severity of PTSD symptoms was measured using the in-app PTSD Checklist for *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders [Fifth Edition]*; PCL-5) [37]. This 20-item measure assesses how much an individual has been bothered by PTSD-related symptoms in the last 1 month. Items are rated on a 5-point Likert scale ranging from 0 (not at all) to 4 (extremely). Total scores range from 0 to 80 with higher scores indicating higher symptom severity. To help characterize the subsample of users who completed the in-app PCL-5, a cutoff score of ≥31 indicated a probable PTSD diagnosis [38]. However, the total PCL-5 score was included as a continuous variable in subsequent analyses. Internal consistency for this study was very good (Cronbach α=0.94).

A technical error found in the Android version of the app resulted in item 3 (“Suddenly feeling or acting as if the stressful experience were actually happening again (as if you were actually back there reliving it”) appearing twice, and item 4 (“Feeling very upset when something reminded you of the stressful experience”) missing from the in-app PCL-5. For Android users, responses to the first iteration of item 3 were retained. Due to item 4 being missing, total scores were prorated (for each user, their 19-item responses were averaged, then multiplied by 20). Finally, only scores from users’ first completion of the PCL-5 were included in analyses.

Primary Measure: Severity of Negative Posttraumatic Cognitions

Severity of NPCs was measured via the 15-item in-app Posttraumatic Maladaptive Beliefs Scale (PMBS) [11]. The PMBS is used to assess unhelpful and potentially harmful beliefs about current life circumstances that may develop following trauma exposure. Seven items represent potential maladaptive beliefs (eg, “I don’t feel safe anywhere anymore”) and 8 reverse-coded items represent potential healthy beliefs (eg, “I feel as though I can depend on other people”). Items are rated on a 7-point scale ranging from 1 (not true) to 7 (completely true). Total scores range from 15 to 105, with higher scores indicating more severe NPCs. Internal consistency in this study was good (Cronbach α=0.88). Scores from users’ first completion of the PMBS were included in analyses.

Secondary Measure: Mental Well-Being

A secondary measure of mental well-being was included as a comparator and to, more broadly, assess positive aspects of mental health in contrast to symptoms of mental illness. This was measured through the in-app 14-item Warwick-Edinburgh Mental Well-Being Scale (WEMWBS) [39]. Each item represents a positive feeling or behavior associated with general well-being (eg, “I’ve been feeling optimistic about the future,” or “I’ve been dealing with problems well”). Users reported the frequency at which each statement was experienced in the past 2 weeks on a scale from 1 (none of the time) to 5 (all of the time). Scores range from 14 to 70, with higher scores indicating higher levels of well-being. Internal consistency in this study was very good (Cronbach α=0.94). Scores from users’ first completion of the WEMWBS were included in analyses.

App Engagement

Measures of user engagement with the Beyond MST app were grouped into three categories: (1) duration of app use, (2) frequency of feature access, and (3) frequency of feature use. Duration of app use was assessed via 2 metrics for the number of distinct days and the number of minutes of app use within the observation period. For minutes of use, serial events greater than 10 minutes apart were recoded to indicate the end of a session and the start of the next session. This prevented overinflation when users navigated away from the app without shutting it down or ending the session. Frequency of feature access was assessed using 2 metrics for the frequency of tool access and the frequency of Learn (psychoeducation) topic access within the observation period. Finally, feature use was assessed using 1 metric for frequency of assessment completion (ie, the number of times the in-app PCL-5, PMBS, and WEMWBS were completed in total).

Data Analytic Plan

Analytic Samples

The following three subsamples of users who completed the in-app symptom assessments were identified: (1) users who completed at least one PCL-5, (2) users who completed at least one PMBS, and (3) users who completed at least one WEMWBS. These subsamples were analyzed separately since users varied in terms of which assessments they completed. For each, we calculated the number of iOS and Android users, the number of users meeting criteria for subthreshold PTSD versus probable PTSD, and summary statistics related to assessment scores and app engagement. Due to the sampling method and properties of the outcome variables, the datasets did not have missing data—except for the previously noted missing value in PCL-5. Since PCL-5 scores for Android users were prorated, a Wilcoxon rank-sum test was also conducted to check for large differences in scores between Android and iOS users.

Variable Diagnostics

Predictor and outcome variable distributions were examined descriptively, statistically, and visually to inform model selection for predicting app engagement. Findings indicated significant

XSL•FO
RenderX

deviations from normality, including positive skewness and overdispersion (ie, observed variance exceeding the expected variance under a Poisson model), violations of residual normality assumptions for linear regression models, as well as the presence of extreme outliers across all outcomes. For 4 outcomes (ie, distinct days of use, frequency of tool access, frequency of Learn topic access, and frequency of assessment completion), this was expected, as these are count variables. Most count variables have the majority of observations occurring at the lower end of the range, with fewer instances of higher counts, leading to a long right tail in the distribution.

Further examination of extreme outliers revealed that high counts reflected acceptable variations in user behavior (eg, "superusers," repeated use over time) and were thus retained in the dataset. Winsorization, a statistical technique that reduces the potential effect of outliers by replacing values beyond a specified percentile with the nearest value within that percentile range, was applied to all outcome variables. A conservative approach was taken to address extreme outliers above the 99.7th to 99.9th percentile.

Predicting App Engagement

Several types of models were constructed to correctly evaluate the impacts of symptom levels on app engagement metrics. Careful attention was given to the data type (ie, count and continuous) and distribution patterns of the outcome variables to inform the optimal model. Rather than include all 3 predictors in a single model, separate models were constructed for each of the 3 predictors on each of the 5 outcomes to maximize sample size and maintain power across the 3 subsamples. Additionally, a quadratic term was included in all models to assess for possible curvilinear relationships.

For 4 count variables (distinct days of use, frequency of tool access, frequency of Learn topic access, and frequency of assessment completion), Poisson and negative binomial models with log-links were evaluated. For distinct days of use and frequency of assessment completions, zero-truncated Poisson and negative binomial models were evaluated as a zero value was not possible. Model fit was compared based on the Akaike Information Criteria statistic, where lower index values indicate better fit [40]. Results from the zero-truncated and standard negative binomial regression models provided the best fit overall and for addressing overdispersion [40]. Thus, those modeling results were used to assess the impacts of predictors on outcome variables. Results are reported as unstandardized coefficients (B). To facilitate interpretation, coefficients were mathematically adjusted as relative percentage changes in expected app engagement using the transformation $(e^{B} - 1) \times 100$.

For the continuous outcome, minutes of use, robust linear regression was used, which lessens the impact of outliers and nonnormality. Unstandardized regression coefficients (B) were interpreted as the expected change in minutes of use for each 1-point change in the predictor. All 3 predictors were mean-centered to reduce multicollinearity and simplify the interpretation of the regression coefficients.

Figure 2 illustrates the types of individual regression models selected. All analyses were performed using R statistical software (v4.4.1; R Core Team), with a significance level set at $P<.05$.

XSL•FO
RenderX

Figure 2. Conceptual models illustrating the predictors and outcome variables examined in a retrospective observational study of Beyond MST (military sexual trauma) app usage data. PCL-5: PTSD (posttraumatic stress disorder) Checklist for DSM-5 (Diagnostic and Statistical Manual of Mental Disorders [Fifth Edition]); PMBS: Posttraumatic Maladaptive Beliefs Scale; WEMWBS: Warwick-Edinburgh Mental Well-Being Scale.

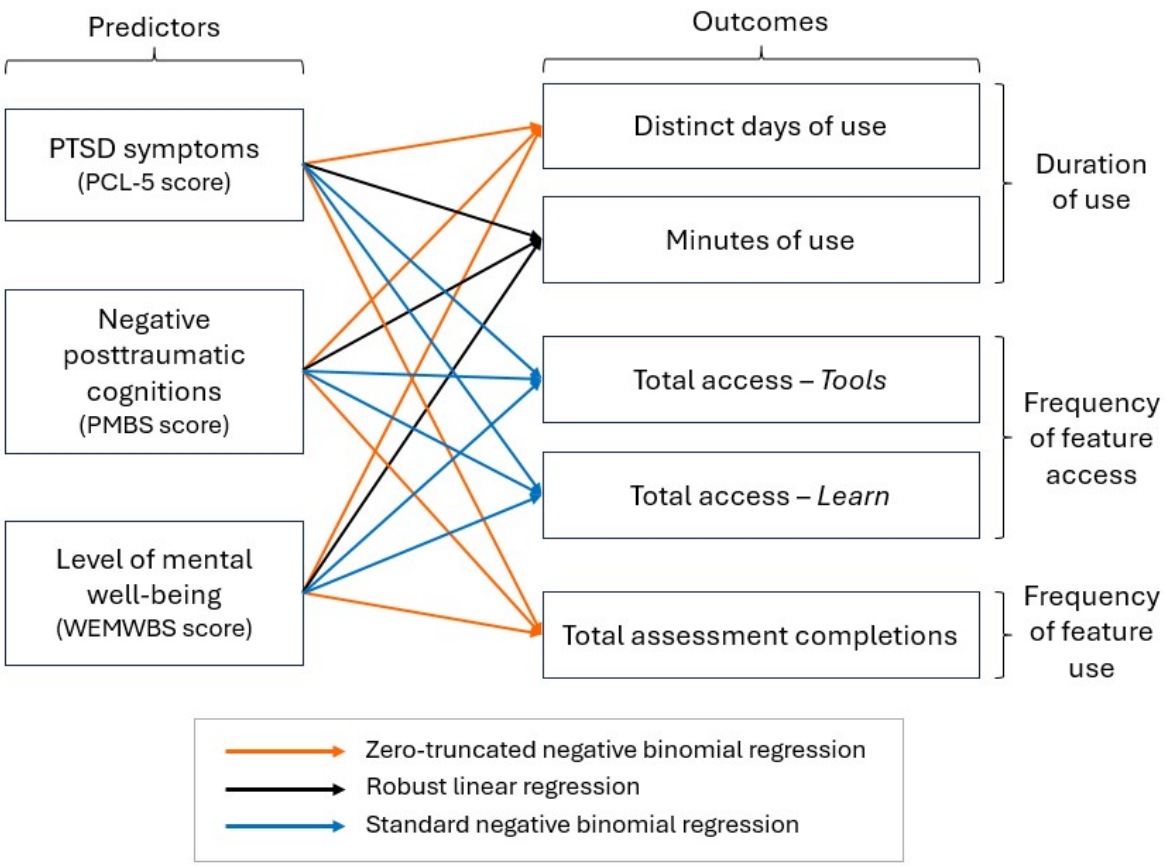

Ethical Considerations

The Stanford University Institutional Review Board determined that this project is quality improvement and is not classified as human subjects research and was, therefore, excluded from oversight (protocol 75160).

Results

Sample Description

Subsamples were obtained from a larger sample of users (N=27,517). Since the data were anonymous, sociodemographic information about users was not available. Among a subsample of users who completed at least one PCL-5 (3689/27,517, 13.4%), 1778 (48.2%) were Android users and 1911 (51.8%) were iOS users. Based on the PCL-5, a large majority of users (3293/3689, 89.3%) met criteria for probable PTSD (score of ≥31). In addition, the distribution of PCL-5 scores across users in this subsample was negatively skewed (Figure 3). To account for prorated responses in the PCL-5, a Wilcoxon rank-sum test was performed to assess potential differences in median PCL-5 scores between Android and iOS users. Results indicated a statistically significant difference in scores, W=1,622,530; P=.02, with median scores of 53.7 (IQR 22.1) for Android users and 55.0 (IQR 23.0) for iOS users. However, the effect size was small (r=0.04, indicating that 0.16% of variance in PCL-5 scores can be attributed to differences in Android vs iOS users), suggesting the observed difference (1.3 points) was statistically and clinically negligible. Consequently, data from different operating systems were combined for the main analyses. Descriptive statistics for all subsamples and engagement measures are shown in Table 1.

XSL•FO
RenderX

Figure 3. Distributions of PTSD (posttraumatic stress disorder) Checklist for DSM-5 (Diagnostic and Statistical Manual of Mental Disorders [Fifth Edition]) (PCL-5; n=3689), Posttraumatic Maladaptive Beliefs Scale (PMBS; n=2197), and Warwick-Edinburgh Mental Well-Being Scale (WEMWBS; n=2160) scores for subsamples of users who completed the Beyond MST (military sexual trauma) in-app assessments during the observation period.

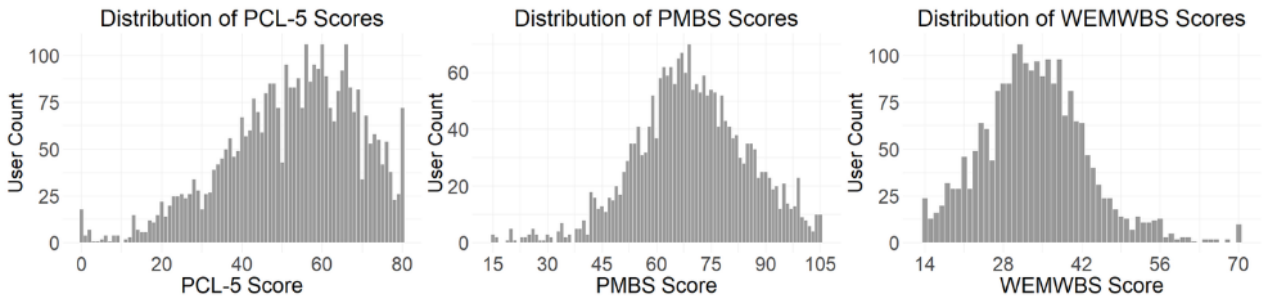

Table 1. Descriptive statistics for subsamples of Beyond MST (military sexual trauma) users based on completed in-app assessment.

Variable	Mean (SD)	95% CI	Median (IQR)	Range
PCL-5[a] subsample (n=3689)				
PCL-5 score	52.38 (16.25)	51.85-52.90	54 (22.89)	0-80
Distinct days of use	6.43 (14.24)	5.97-6.89	3 (5)	1-416
Minutes of use	46.58 (147.62)	41.81-51.34	22.15 (33.72)	0.45-5112.62
Frequency of tool access	7.99 (20.77)	7.32-8.67	3 (7)	0-709
Frequency of Learn topic access	5.14 (16.77)	4.60-5.68	1 (6)	0-637
Frequency of assessment completion	2.80 (4.40)	2.66-2.94	2 (2)	1-151
PMBS[b] subsample (n=2197)				
PMBS score	69.49 (15.43)	68.84-70.13	69 (20)	15-105
Distinct days of use	7.74 (17.61)	7.00-8.48	4 (6)	1-416
Minutes of use	62.36 (188.87)	54.46-70.26	29.78 (42.85)	0.52-5112.62
Frequency of tool access	11.27 (27.74)	10.11-12.43	5 (9)	0-709
Frequency of Learn topic access	6.92 (23.34)	5.95-7.90	2 (7)	0-637
Frequency of assessment completion	3.62 (4.47)	3.43-3.81	3 (2)	1-131
WEMWBS[c] subsample (n=2160)				
WEMWBS score	33.66 (9.39)	33.26-34.05	33 (11)	14-70
Distinct days of use	7.96 (18.80)	7.17-8.76	4 (6)	1-416
Minutes of use	61.74 (188.99)	53.76-69.71	29.21 (41.99)	0.75-5112.62
Frequency of tool access	10.90 (27.02)	9.76-12.04	5 (9)	0-709
Frequency of Learn topic access	6.55 (21.13)	5.66-7.45	1 (7)	0-637
Frequency of assessment completion	3.84 (5.65)	3.60-4.07	3 (2)	1-151

[a]PCL-5: PTSD (posttraumatic stress disorder) Checklist for DSM-5 (Diagnostic and Statistical Manual of Mental Disorders [Fifth Edition]).

[b]PMBS: Posttraumatic Maladaptive Beliefs Scale.

[c]WEMWBS: Warwick-Edinburgh Mental Well-Being Scale.

Users who completed at least one PMBS (2197/27,517, 8%) included 1030 (46.9%) Android and 1167 (53.1%) iOS users. As with PCL-5 scores, the distribution of PMBS scores across users in this subsample was also negatively skewed (Figure 3).

Finally, the subsample of users who competed at least one WEMWBS (2160/27,517, 7.8%) included 1043 (48.3%) Android and 1117 (51.7%) iOS users. In contrast to the distributions for PCL-5 and PMBS scores, the distribution for WEMWBS scores was positively skewed (Figure 3).

XSL•FO
RenderX

Relationship Between Trauma-Related Measures and App Engagement

Severity of PTSD Symptoms

Model results for PCL-5 scores are presented in Table 2 and graphs depicting the impacts of PCL-5 scores on outcomes are provided in Figure 4A-E. For distinct days of use, PCL-5 scores showed significant linear (B=−0.01, *z*=−2.7; *P*=.008), and quadratic (B=−0.0003, *z*=−3.2; *P*=.001) effects, such that days of use were lowest at very low symptom levels, highest at moderate PCL-5 levels, and declined again at high PCL-5 symptom levels. For 3 additional predictors of PCL-5 scores (ie, minutes of use, frequency of tool access, and frequency of assessment completion), similar inverted-U patterns of statistically significant negative linear (B<−3.1; *P*<.001) and quadratic (B<−0.004; *P*<.001) effects were found with the lowest and highest levels of PCL-5 scores predicting less engagement, access, and use, while moderate PCL-5 scores were related to increased levels of engagement, access, and use of the app. Finally, for frequency of Learn topic access, a 1-point increase in PCL-5 was linked to a 0.6% reduction in expected counts (B=−0.006, *z*=−2.8; *P*=.005), though the quadratic term was not significant (B=−0.0001, *z*=−1.3; *P*=.20), indicating a primarily linear relationship, where Learn topic access declined as PCL-5 scores increased. In summary, across 4 of the 5 metrics, moderate levels of PCL-5 scores were associated with the highest levels of engagement, access, and use, compared with lower and higher levels of symptoms.

Table 2. Results from models assessing relationships between PCL-5[a] (n=3690), PMBS[b] (n=2197), and WEMWBS[c] (n=2160) scores and measures of app engagement for Beyond MST[d] users who completed in-app assessments.

Predictor and outcomes	Linear term			Quadratic term		
	B[e]	Test statistic value[f]	*P* value	B	Test statistic value[f]	*P* value
PCL-5						
Distinct days of use	–0.01	–2.7	.008	–0.0003	–3.2	.001
Minutes of use	–0.09	–4.5	<.001	–0.004	–4.0	<.001
Frequency of tool access	–0.01	–9.6	<.001	–0.0003	–4.1	<.001
Frequency of Learn topic access	–0.006	–2.8	.005	–0.0001	–1.3	.20
Frequency of assessment completion	–0.008	–3.1	.002	–0.0003	–4.9	<.001
PMBS						
Distinct days of use	–0.01	–2.2	.03	–0.0004	–2.7	.01
Minutes of use	–0.02	–0.7	.49	–0.002	–1.2	.22
Frequency of tool access	–0.008	–4.8	<.001	–0.0002	–2.3	.02
Frequency of Learn topic access	–0.003	–1.2	.23	–0.00009	–0.8	.40
Frequency of assessment completion	0.0005	0.31	.75	–0.00008	–1.9	.06
WEMWBS						
Distinct days of use	0.03	3.4	<.001	–0.0005	–1.7	.09
Minutes of use	0.11	1.9	.06	–0.004	–1.0	.30
Frequency of tool access	0.02	6.2	<.001	–0.0003	–1.5	.13
Frequency of Learn topic access	0.009	1.9	.06	0.0001	0.5	.62
Frequency of assessment completion	–0.003	–1.5	.13	–0.00009	–0.7	.49

[a]PCL-5: PTSD (posttraumatic stress disorder) Checklist for DSM-5 (Diagnostic and Statistical Manual of Mental Disorders [Fifth Edition]).

[b]PMBS: Posttraumatic Maladaptive Beliefs Scale.

[c]WEMWBS: Warwick-Edinburgh Mental Well-Being Scale.

[d]MST: military sexual trauma.

[e]B: unstandardized coefficient.

[f]z-statistics are presented for zero-truncated negative binomial and standard negative binomial regression models. T-statistics are presented for robust linear regression models.

XSL•FO
RenderX

Figure 4. Predicted engagement outcomes by PTSD (posttraumatic stress disorder) Checklist for DSM-5 (Diagnostic and Statistical Manual of Mental Disorders [Fifth Edition]) (PCL-5) score for the subsample of Beyond MST (military sexual trauma) users who completed the in-app PCL-5 (n=3689).

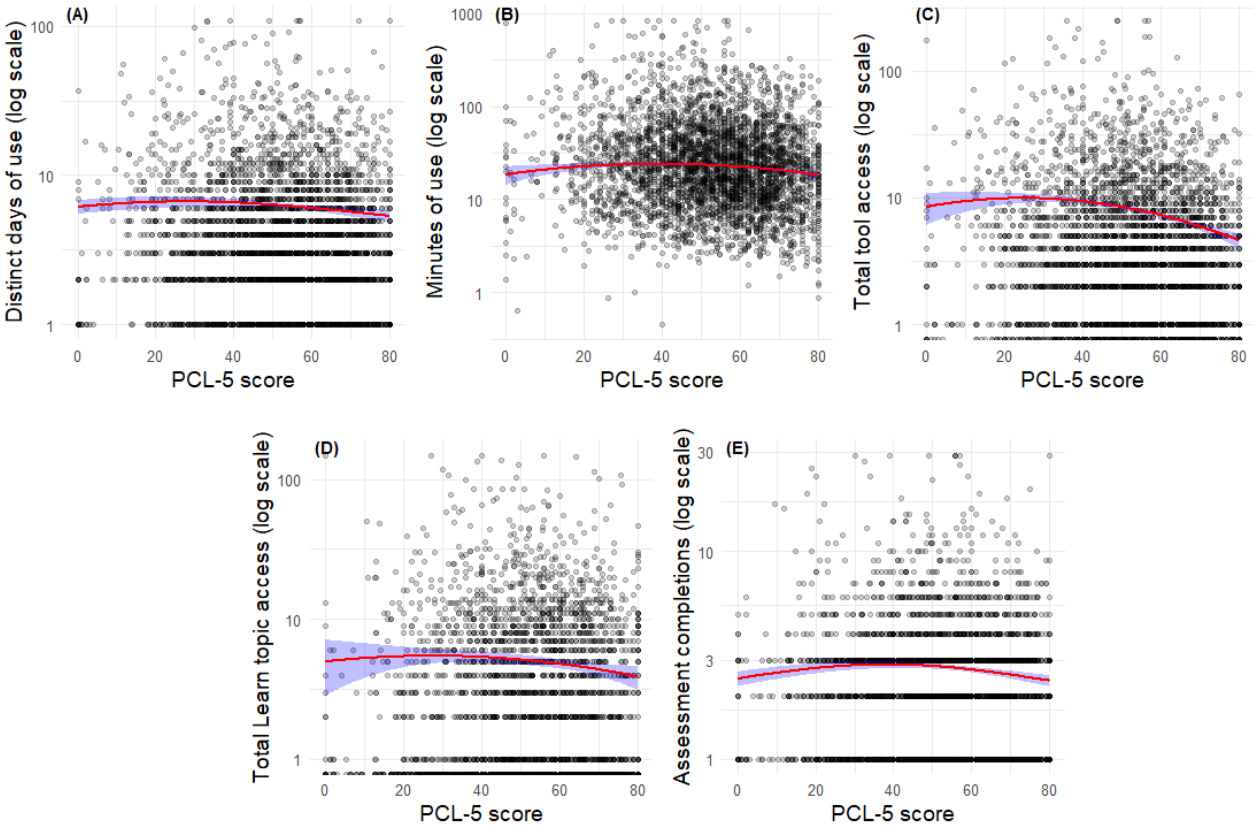

In Figure 4, red lines represent model-estimated values based on quadratic negative binomial, zero-truncated negative binomial, and robust linear regression models. Shaded blue regions indicate 95% CIs for predicted values. All plots display outcome variables on a log-transformed scale for interpretability. Panels A-E depict the results for PCL-5 score on each outcome: (A) distinct days of use, (B) minutes of use, (C) frequency of tool access, (D) frequency of Learn topic access, and (E) frequency of assessment completion.

Severity of Negative Posttraumatic Cognitions

Table 2 and Figure 5A-E depict the results for the PMBS models. For distinct days of use, PMBS scores significant linear (B=−0.01, z=−2.2; P=.03) and quadratic (B=−0.0004, z=−2.7; P=.006) effects were identified. As with the PCL-5, this suggests that days of use were lowest at very low PMBS levels, highest at moderate PMBS levels, and declined again at high PMBS levels. A similar inverted-U pattern was observed for frequency of tool access, with statistically significant linear (B=−0.008, z=−4.8; P<.001) and quadratic effects (B=−0.0002, z=−2.3; P=.02). In contrast, statistically significant relationships were not observed between PMBS scores and the other 3 outcomes (ie, minutes of use, frequency of Learn topic access, and frequency of assessment completion; all P>.05). In summary, for days of use and tool access, moderate levels of PMBS scores were associated with the highest levels of engagement and access, compared with lower and higher levels of symptoms, while no relationships were found between PMBS scores and the other metrics.

XSL•FO
RenderX

Figure 5. Predicted engagement outcomes by Posttraumatic Maladaptive Beliefs Scale (PMBS) score for the subsample of Beyond MST (military sexual trauma) users who completed the in-app PMBS (n=2197).

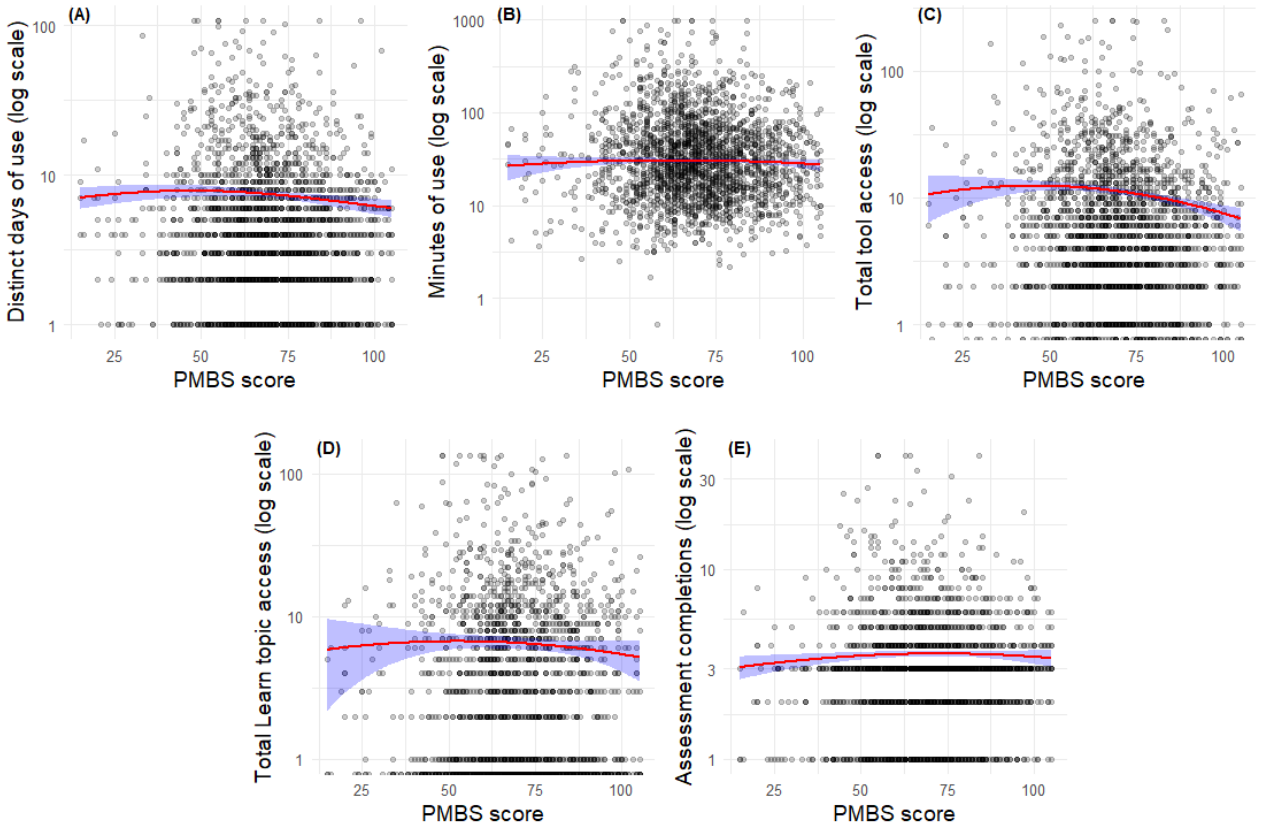

In Figure 5, red lines represent model-estimated values based on quadratic negative binomial, zero-truncated negative binomial, and robust linear regression models. Shaded blue regions indicate 95% CIs for predicted values. All plots display outcome variables on a log-transformed scale for interpretability. Panels A-E depict the results for PMBS score on each outcome: (A) distinct days of use, (B) minutes of use, (C) frequency of tool access, and (D) frequency of Learn topic access, and (E) frequency of assessment completion.

Level of Mental Well-Being

Table 2 and Figure 6A-E depict the results for the WEMWBS models. In contrast to the PCL-5 and PMBS findings, higher WEMWBS scores, reflecting greater mental well-being, were associated with increased app engagement for 2 of the 5 outcomes. For distinct days of use, each 1-point increase in WEMWBS scores predicted a 3% increase in expected usage days (B=0.03, z=3.4; P<.001). A similar positive linear relationship was found for frequency of tool access, with a 2% increase in expected counts per 1-point increase in WEMWBS scores (B=0.02, z=6.2; P<.001). No significant effects were observed for minutes of use, frequency of Learn topic access, and frequency of assessment completion (P>.05). In summary, predicted usage days and frequency of tool access increased as WEMWBS scores increased, while no relationships were found for the other metrics.

XSL•FO
RenderX

Figure 6. Predicted engagement outcomes by Warwick-Edinburgh Mental Well-Being Scale (WEMWBS) score for the subsample of Beyond MST (military sexual trauma) users who completed the in-app WEMWBS (n=1160).

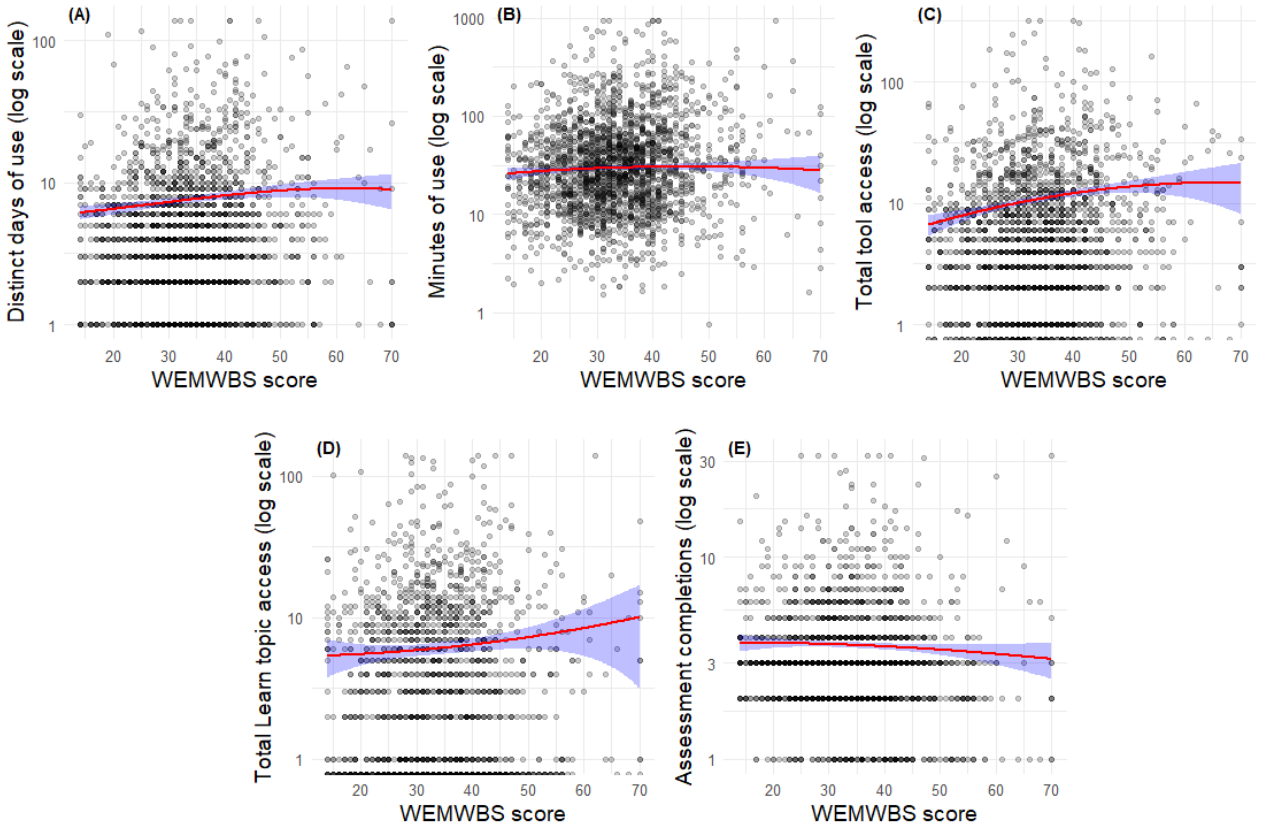

In Figure 6, red lines represent model-estimated values based on quadratic negative binomial, zero-truncated negative binomial, and robust linear regression models. Shaded blue regions indicate 95% CIs for predicted values. All plots display outcome variables on a log-transformed scale for interpretability. Panels A-E depict the results for WEMWBS score on each outcome: (A) distinct days of use, (B) minutes of use, (C) frequency of tool access, (D) frequency of Learn topic access, and (E) frequency of assessment completion.

Discussion

Principal Findings and Comparisons With Prior Work

Mobile app engagement is a challenging construct to analyze, particularly when examining factors that influence human behavior in naturalistic settings. Researchers have made significant strides in understanding how aspects related to the apps themselves (eg, usability and acceptability) impact mental health app engagement. However, less is known about which user-level factors impact engagement. Emerging research highlights the potential influence of factors such as age, gender, and digital literacy [32]. However, few studies have evaluated how symptom severity might impact engagement with digital health technologies, especially mental health apps for trauma and PTSD. Studies examining other mental health conditions (eg, depression) suggest that symptom severity may play a crucial role [30-33]. This exploratory study furthers our understanding of how posttraumatic symptoms, negative cognitions, and levels of mental well-being may be associated with engagement with an app designed for those who experienced MST.

Based on usage data, more than 27,000 users engaged with Beyond MST from its launch on March 11, 2021, through July 29, 2024. A small subsample of users completed the in-app PCL-5 (3689/27,517, 13.4%), PMBS (2197/27,517, 7.8%), or WEMWBS (2160/27,517, 7.9%) at least once during the observation period. These assessment completion rates are consistent with data from studies on other publicly available mental health apps [33,41-43]. In addition, the subsamples were nearly split in terms of Android (46.9%-48.3%) versus iOS (51.5%-53.1%) users.

The results suggest that Beyond MST is reaching the targeted populations, including those with high levels of PTSD-related concerns, including negative posttraumatic cognitions and poorer mental well-being. Specifically, summary statistics for the subsamples revealed that scores for the PCL-5 and PMBS were positively skewed, indicating relatively severe PTSD symptoms and NPCs, whereas scores for the WEMWBS were negatively skewed, indicating relatively poor mental well-being overall. In addition, almost 90% (3293/3689) of users who completed the PCL-5 met criteria for probable PTSD. Taken together, it is likely that these subsamples of users are highly symptomatic, with substantial posttraumatic symptoms that may negatively impact their well-being and quality of life, similar to findings from studies of the PTSD Coach app [41,42].

XSL•FO
RenderX

Consistent with research on engagement with mental health apps, generally there was limited engagement with the Beyond MST app in terms of duration of app use as well as frequency of feature access and use. The median distinct days of use was just 3-4 days (IQR 5-6), and the median minutes of use was about 22-30 minutes (IQR 33.7-42.9) across samples. The median frequency of coping tool access and Learn (ie, psychoeducation) topic access was 3-5 (IQR 7-9) and 1-2 (IQR 6-7) times, respectively. Finally, the median frequency of assessment completion was 2-3 (IQR 2). Considering the tools are designed to be used multiple times to support coping skills practice and the assessments are meant to be completed repeatedly for progress tracking, it follows that median rates of tool access and assessment completion would be higher compared to psychoeducation access. Both the relatively low rates of engagement and the differences in tool versus psychoeducation access are consistent with findings from evaluations of other mental health apps with similar designs and content areas [33,41-43].

Results also uncovered curvilinear patterns related to trauma-related measures and app engagement. The nonlinear relationships were most clearly seen in relation to PTSD symptom severity and NPCs. For users with the most and least severe symptom levels (eg, PCL-5 score ≥65 or <25), engagement with Beyond MST was lower compared to users reporting moderate symptom levels (eg, scores of 30-40). These effects were consistent for 2 engagement outcomes, distinct days of use and frequency of tool access. This curvilinear pattern is consistent with the Yerkes-Dodson principle [44,45], describing the inverted-U shaped relationship between arousal and performance, with optimal levels of performance, or engagement in this case, at moderate levels of arousal and with decrements when arousal is too low or too high.

For minutes of app use and frequency of assessment completion, the relationship between symptom severity and app engagement was less consistent and clear. For minutes of app use, the results were mixed. Like with PTSD symptom severity and distinct days of use, there was a similar curvilinear relationship between PTSD symptom severity and minutes of use. However, no statistically significant relationship was found for NPC severity and minutes of use. It is unclear why consistent effects were not observed across similar duration of use metrics. One potential factor is the difference in the average minutes of app use between the subsamples (PCL-5 sample: mean 46.6, SD 147.6 minutes; PMBS sample: mean 62.4, SD 188.9 minutes) and increased variability. In contrast, consistent effects were found for distinct days of use where the subsamples had more similar means (PCL-5 sample: mean 6.4, SD 14.2 days; PMBS sample: mean 7.7, SD 17.6 days) with less variability.

These same patterns were observed for frequency of assessment completion where the highest and lowest levels of PTSD symptom severity predicted lower completion rates for all scales compared to those with moderate PTSD symptom levels, whereas NPC severity showed no significant linear or quadratic associations. Assessment completion reflects a complex, multistep behavioral sequence, requiring initiation, sustained attention, and follow-through. Therefore, it is possible that PTSD symptom severity may more strongly disrupt the cognitive and motivational processes needed to complete structured tasks, while NPCs alone may not exert the same level of interference. Other factors, such as the length of the scale (15 vs 20 items) and the trauma-related symptom content of the PCL-5 items may influence engagement.

The pattern of results for psychoeducation access was also inconsistent. There were no curvilinear relationships for either symptom severity measure. In addition, the results regarding linear effects were mixed. Increasing PTSD symptom severity predicted less psychoeducation access, but there was no relationship between NPC severity and psychoeducation access. Although the reasons for this discrepancy are unclear, it is possible that those experiencing particular PTSD symptoms at higher levels are seeking in-the-moment symptom relief versus wanting to learn about MST.

In contrast, greater mental well-being was associated with increased app engagement, generally following linear trends. These patterns were strongest for distinct days of use and tool access, while effects for psychoeducation access were more modest or inconsistent. Notably, those with the lowest levels of well-being accessed the tools less frequently compared to users with higher well-being. Taken together, this suggests that in this sample of primarily highly symptomatic users, individuals displaying better well-being–related factors (eg, positive emotions, life satisfaction, and functioning) may continue to find value in the app over time, using it for ongoing support; however, users with fewer internal resources may find slightly more difficulty sustaining engagement.

Overall, the novel findings of possible inverted-U patterns prompt a need for further assessment of curvilinear relationships in digital mental health engagement research. Potentially, users with lower posttraumatic symptom severity may not perceive a pressing need for support, while those with more severe symptoms may experience avoidance, emotional numbing, or cognitive overload that impairs their ability to engage meaningfully with digital tools. It also follows that emerging evidence of such relationships has been found for depression symptoms (eg, [33]) since depression often co-occurs with MST [5]. This also parallels psychotherapy research showing that both low perceived need and high avoidance can undermine treatment engagement and retention [46-48]. Therefore, it is possible that other symptoms not assessed are driving these relationships. Taken together, engagement is not simply a function of need, but also of readiness and psychological bandwidth.

However, it is important to interpret these findings in the context of clinical significance. According to the clinical guidelines for the PCL-5 [49], a 10-point change in the PCL-5 suggests a clinically significant change in PTSD symptoms. This can be used as a benchmark for understanding clinically meaningful differences in app engagement. For example, based on the model in this study, a 10-point increase on PCL-5 predicts approximately a 10% reduction in the expected number of distinct days of use, which translates to roughly half a day less engagement from a baseline median of 5 days. Similarly, for tools, psychoeducation, and assessments, a 10-point increase corresponds to modest decreases in the expected frequency of

XSL•FO
RenderX

access or use. While these effects are statistically significant, the absolute changes in engagement are small, suggesting that even clinically meaningful differences in PTSD symptoms may only slightly reduce platform use. These small effect sizes highlight the importance of interpreting behavioral outcomes in real-world contexts where even brief or infrequent engagement may still yield therapeutic benefit. It is also possible that symptom levels indirectly affect engagement through more direct predictors such as motivation level [50]. This is especially pertinent for types of engagement that require sustained attention and repeated behavior, such as in-app symptom tracking over time or tool use for ongoing skills practice. Finally, since the subsamples displayed high levels of symptoms overall, it is important to interpret these findings in the context of highly symptomatic individuals. For example, compared to users with high levels of symptoms who may still benefit from self-guided use, perhaps those with the highest levels of symptoms might display improved engagement when provided additional support (eg, human coaching) when using apps such as Beyond MST.

Given the curvilinear relationships found, future work may also benefit from using analytic methods capable of detecting more complex or heterogeneous patterns of engagement. The current models tested only linear and quadratic associations assuming a homogeneous population. However, the overdispersion and presence of outliers suggest that the sample may include multiple populations, such as “typical users” and “super users,” who engage with the app differently. Future studies could use alternate methods, such as mixture modeling approaches to determine whether subgroups of users follow distinct engagement patterns or whether more nuanced relationships between symptom levels and app engagement exist.

In sum, PTSD symptom severity and, to a somewhat lesser degree, negative posttraumatic cognitions, may be important factors to consider when addressing user engagement with an app for MST. This has important implications for app developers and mental health professionals. Specifically, personalized in-app treatment plans or stepped care pathways based on baseline symptom level, such as increased human support for those with more severe symptoms, may improve app engagement. Providers using apps in care should also consider symptom severity when making recommendations (eg, psychoeducation vs tool use). Additionally, the insights generated from these analyses suggest how symptom severity can be used to understand digital interventions along the continuum of care. This is especially crucial for apps such as Beyond MST that can be systematically integrated with care in large health care systems serving diverse populations with a range of mental health symptoms and trauma types [51]. Although these findings may suggest potential clinical implications for the subsamples examined, limitations of the current study should be considered when interpreting these results more broadly.

Limitations

Several study limitations are inherent to the nature of the data collected as well as the subsamples and methods used. First, due to the data being anonymous, user characteristics were not accessible, making it impossible to control for potential confounding factors such as sociodemographics. This may have been problematic because some demographic factors (eg, female gender [4] and young age [52]) are disproportionately associated with risk of MST victimization. More generally, studies have found that women are also more likely to adopt and engage with digital mental health interventions compared to men [32]. Therefore, it is possible that the subsamples examined in this study were not demographically diverse, limiting the generalizability of these results to the broader population. Clinical trials should be conducted to explore the relationship between posttraumatic symptom severity and app engagement while controlling for potential confounding factors.

Second, this study focused on a subset of users who engaged with the in-app assessments, which may limit generalizability to the complete user sample. This potentially biases the results, as these subsamples of users may be more engaged when compared to the overall user base. Nevertheless, there remains value in the evaluation. First, the data provide insights into the real-world use patterns and relationships among large user samples that engaged with a free and publicly available app. For-profit companies often do not release such detailed information around mobile app engagement. These results can inform hypothesis generation for future studies and potential methods for increasing engagement. Regardless, further user evaluation and controlled studies are needed to identify possible differences in engagement between the broader sample of users and the subsets examined in this study.

Third, the in-app assessments and measures of app engagement used in this study may not fully capture the scope of posttraumatic symptom severity and its effects on app engagement. Several different symptoms are associated with trauma and PTSD, so the use of PCL-5 scores does not inform how individual posttraumatic symptoms might predict app engagement. For example, avoidant behavior, a hallmark symptom of PTSD, might negatively impact app engagement when compared to other symptoms [5]. Also, this study focused on objective metrics of app engagement. Future user experience studies with Beyond MST should explore subjective measures of app engagement such as self-reported intentions to engage. This would provide a more comprehensive picture of how users intended to engage, how they actually engaged, and possible barriers to engagement.

Relatedly, there are numerous app engagement measures that were not explored due to the constraints of the app’s design and availability of data. For example, apart from feature access and assessment feature use, other types of feature use could not be assessed due to limitations in the backend of Beyond MST. Importantly, researchers have urged the field to move beyond just measuring “clicks and downloads” when examining app engagement [53]. Engagement is a multidimensional construct that not only includes behavioral aspects (eg, duration of use and frequency of use), but also subjective experiential aspects (eg, focused attention, emotional involvement, and usability) [53] which are often best understood through qualitative and mixed methods approaches. The extent to which users meaningfully apply skills learned or practiced via the app in day-to-day life should also be explored in future work. For mental health apps, this means defining what constitutes

XSL•FO
RenderX

clinically meaningful app engagement. For example, the frequency of accessing the Deep Breathing tool in Beyond MST does not fully capture the experiences of users who adopted the skill with minimal app use but continued to apply that skill in daily life. In other words, clinically meaningful app engagement may vary from person to person. Longitudinal studies are also needed to understand how symptom severity may impact engagement over time. Overall, further research is needed that captures the breadth of these complex constructs, providing a more nuanced understanding of how posttraumatic symptoms might impact mental health app engagement.

Conclusions

The Beyond MST app has the potential to support a diverse range of service members and veterans who experienced military sexual trauma, whether they are using the app for self-care purposes, or are using it in conjunction with professional care. Even highly symptomatic users were found to engage with Beyond MST. However, some small effects on user engagement based on posttraumatic symptom severity levels were found which may limit the potential effectiveness of Beyond MST in helping some users cope with trauma, find hope, and enhance their well-being, particularly users with the highest levels of symptoms. Additional research on possible user-level predictors of app engagement, and how specific factors might promote or impede engagement, is needed. Findings from this study suggest statistically significant curvilinear and linear relationships between posttraumatic symptom severity and app engagement, but the clinical implications of these results remain unclear without additional study. Careful examination of the potential predictive power of individual posttraumatic symptoms on a wider range of app engagement measures is needed. Nevertheless, this study may serve as a model for future research in this area and highlights the importance of understanding symptom severity as it relates to user engagement to optimize digital solutions for mental health conditions. Continued exploration in this area will be crucial for developing more effective mental health apps for people who have experienced MST and other types of traumatic stress.

Acknowledgments

Preliminary results from this study were published online as part of SRH's master's thesis through San Jose State University. The authors would like to acknowledge Virginia McCaughey, Emma Archibald, and the VA's MST Support Team for supporting the development of Beyond MST. The authors would also like to acknowledge Jessica Mow, Kathryn Macia, and Craig Rosen for their support with this project. The views expressed in this paper are those of the authors and do not reflect the official policy or position of the institutions with which the authors are affiliated, the US Department of Veterans Affairs, or the US Government.

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: literature search, code generation for data processing and analysis; code optimization for data processing and analysis; proofreading and editing. The GAI tools used were VA's ChatGPT-4.5 and Microsoft Copilot. Responsibility for the final manuscript lies entirely with the authors. This declaration is submitted by SH.

Data Availability

The participants in this quality improvement project did not provide written consent for their data to be shared publicly. Therefore, due to the sensitive nature of the research, supporting data are not available.

Funding

No grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article. This research was partially supported with funding and resources from the National Center for PTSD, Dissemination and Training Division at the US Department of Veterans Affairs.

Authors' Contributions

SH and MAM contributed to writing the original draft and visualization. SH, SMR, VC, DS, and MAM contributed to conceptualization, formal analysis, methodology, review and editing of the manuscript. SMR, VC, DS, and MAM contributed to supervision. SR contributed to data curation. SS, KMR, JEO, WCS, AES, and KT contributed to data curation, resources, and software related to the backend data and Beyond MST app. All authors reviewed, edited, and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Street A, Skidmore C, Gyuro L, Bell M. Military sexual trauma. U.S. Department of Veterans Affairs. 2024. URL: https://www.ptsd.va.gov/professional/treat/type/sexual_trauma_military.asp [accessed 2024-06-03]

XSL•FO
RenderX

2. Barth SK, Kimerling RE, Pavao J, McCutcheon SJ, Batten SV, Dursa E, et al. Military sexual trauma among recent veterans: correlates of sexual assault and sexual harassment. Am J Prev Med 2016;50(1):77-86. [doi: 10.1016/j.amepre.2015.06.012] [Medline: 26232906]
3. Gibson CJ, Gray KE, Katon JG, Simpson TL, Lehavot K. Sexual assault, sexual harassment, and physical victimization during military service across age cohorts of women veterans. Womens Health Issues 2016;26(2):225-231. [doi: 10.1016/j.whi.2015.09.013] [Medline: 26632005]
4. Wilson LC. The prevalence of military sexual trauma: a meta-analysis. Trauma Violence Abuse 2018;19(5):584-597. [doi: 10.1177/1524838016683459] [Medline: 30415636]
5. Galovski TE, Street AE, Creech S, Lehavot K, Kelly UA, Yano EM. State of the knowledge of VA military sexual trauma research. J Gen Intern Med 2022;37(Suppl 3):825-832 [FREE Full text] [doi: 10.1007/s11606-022-07580-8] [Medline: 36042078]
6. Kerr K, Varker T, Meurk C, Williams G. Synergistic effects of military sexual trauma and military combat exposure on posttraumatic stress disorder in veterans. a systematic review. Traumatology 2024;30(4):509-516. [doi: 10.1037/trm0000464]
7. Forkus SR, Weiss NH, Goncharenko S, Mammay J, Church M, Contractor AA. Military sexual trauma and risky behaviors: a systematic review. Trauma Violence Abuse 2021;22(4):976-993 [FREE Full text] [doi: 10.1177/1524838019897338] [Medline: 31920168]
8. Galovski TE, McSweeney LB, Woolley MG, Alpert E, Nillni YI. The relative impact of different types of military sexual trauma on long-term PTSD, depression, and suicidality. J Interpers Violence 2023;38(15-16):9465-9491. [doi: 10.1177/08862605231168823] [Medline: 37102588]
9. Ehlers A, Clark DM. A cognitive model of posttraumatic stress disorder. Behav Res Ther 2000;38(4):319-345. [doi: 10.1016/s0005-7967(99)00123-0] [Medline: 10761279]
10. Brown LA, Belli GM, Asnaani A, Foa EB. A review of the role of negative cognitions about oneself, others, and the world in the treatment of PTSD. Cogn Ther Res 2018;43(1):143-173. [doi: 10.1007/s10608-018-9938-1]
11. Vogt DS, Shipherd JC, Resick PA. Posttraumatic maladaptive beliefs scale: evolution of the personal beliefs and reactions scale. Assessment 2012;19(3):308-317. [doi: 10.1177/1073191110376161] [Medline: 20660470]
12. Janoff-Bulman R. Shattered Assumptions: Towards a New Psychology of Trauma. New York, NY: Simon and Schuster; 1992.
13. Carroll KK, Lofgreen AM, Weaver DC, Held P, Klassen BJ, Smith DL, et al. Negative posttraumatic cognitions among military sexual trauma survivors. J Affect Disord 2018;238:88-93 [FREE Full text] [doi: 10.1016/j.jad.2018.05.024] [Medline: 29864715]
14. Matt L, Thompson K, Lofgreen AM, Van Horn R. Treatment of posttraumatic stress disorder related to military sexual trauma. Psychiatr Ann 2020;50(10):432-436. [doi: 10.3928/00485713-20200916-01]
15. Andresen FJ, Blais RK. Higher self-stigma is related to lower likelihood of disclosing military sexual trauma during screening in female veterans. Psychol Trauma 2019;11(4):372-378 [FREE Full text] [doi: 10.1037/tra0000406] [Medline: 30284856]
16. Dardis CM, Reinhardt KM, Foynes MM, Medoff NE, Street AE. “Who are you going to tell? who’s going to believe you?”: women’s experiences disclosing military sexual trauma. Psychol Women Q 2018;42(4):414-429. [doi: 10.1177/0361684318796783]
17. Holland KJ, Rabelo VC, Cortina LM. Collateral damage: military sexual trauma and help-seeking barriers. Psychol Violence 2016;6(2):253-261. [doi: 10.1037/a0039467]
18. Holliday R, Monteith LL. Seeking help for the health sequelae of military sexual trauma: a theory-driven model of the role of institutional betrayal. J Trauma Dissociation 2019;20(3):340-356. [doi: 10.1080/15299732.2019.1571888] [Medline: 30714879]
19. Street AE, Shin MH, Marchany KE, McCaughey VK, Bell ME, Hamilton AB. Veterans' perspectives on military sexual trauma-related communication with VHA providers. Psychol Serv 2021;18(2):249-259 [FREE Full text] [doi: 10.1037/ser0000395] [Medline: 31621356]
20. Neary M, Schueller SM. State of the field of mental health apps. Cogn Behav Pract 2018;25(4):531-537 [FREE Full text] [doi: 10.1016/j.cbpra.2018.01.002] [Medline: 33100810]
21. Wang K, Varma DS, Prosperi M. A systematic review of the effectiveness of mobile apps for monitoring and management of mental health symptoms or disorders. J Psychiatr Res 2018;107:73-78. [doi: 10.1016/j.jpsychires.2018.10.006] [Medline: 30347316]
22. Bröcker E, Suliman S, Olff M, Seedat S. Feasibility, acceptability, and effectiveness of web-based and mobile PTSD coach: a systematic review and meta-analysis. Eur J Psychotraumatol 2023;14(2):2209469 [FREE Full text] [doi: 10.1080/20008066.2023.2209469] [Medline: 37226885]
23. Hardeberg Bach M, Ahrens C, Olff M, Armour C, Krogh SS, Hansen M. EHealth for sexual assault: a systematic scoping review. Trauma Violence Abuse 2024;25(1):102-116. [doi: 10.1177/15248380221143355] [Medline: 36632639]
24. Nichols JL. Reaching those who are difficult to reach: exploring online interventions for survivors of military sexual trauma. J Mil Gov Couns 2015;3(4):2-4 [FREE Full text]

XSL•FO
RenderX

25. Moret JD, Todd A, Rose L, Pollitt E, Anderson J. Mobile phone apps for intimate partner and sexual violence prevention and response: systematic search on app stores. JMIR Form Res 2022;6(2):e28959 [FREE Full text] [doi: 10.2196/28959] [Medline: 35133285]
26. Collaton J, Barata P, Lewis SP. Understanding discussions of sexual assault in young women on a peer support mental health app: a content analysis. J Interpers Violence 2022;37(23-24):NP22811-NP22833 [FREE Full text] [doi: 10.1177/08862605211073112] [Medline: 35195468]
27. Alhusen J, Bloom T, Clough A, Glass N. Development of the myPlan safety decision app with friends of college women in abusive dating relationships. J Technol Hum Serv 2015;33(3):263-282. [doi: 10.1080/15228835.2015.1037414]
28. Street AE, Senti S, Skidmore WC, Taylor K, Archibald EA, Brown T, et al. Beyond MST [Mobile application software]. Version 1.1.1(125). Google Play Store. 2024. URL: https://play.google.com [accessed 2026-03-26]
29. Street AE, Senti S, Skidmore WC, Taylor K, Archibald EA, Brown T, et al. Beyond MST [Mobile application software]. Version 1.1.1(125). Apple App Store. 2024. URL: https://itunes.apple.com [accessed 2026-03-27]
30. Lipschitz JM, Pike CK, Hogan TP, Murphy SA, Burdick KE. The engagement problem: a review of engagement with digital mental health interventions and recommendations for a path forward. Curr Treat Options Psychiatry 2023;10(3):119-135 [FREE Full text] [doi: 10.1007/s40501-023-00297-3] [Medline: 38390026]
31. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
32. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
33. Jaworski BK, Taylor K, Ramsey KM, Heinz A, Steinmetz S, Pagano I, et al. Exploring usage of COVID coach, a public mental health app designed for the COVID-19 pandemic: evaluation of analytics data. J Med Internet Res 2021;23(3):e26559 [FREE Full text] [doi: 10.2196/26559] [Medline: 33606656]
34. Yeager CM, Shoji K, Luszczynska A, Benight CC. Engagement with a trauma recovery internet intervention explained with the health action process approach (HAPA): longitudinal study. JMIR Ment Health 2018;5(2):e29 [FREE Full text] [doi: 10.2196/mental.9449] [Medline: 29636323]
35. Jaworski BK, Ramsey KM, Taylor K, Heinz AJ, Senti S, Mackintosh M, et al. Mental health apps and U.S. military veterans: perceived importance and utilization of the National Center for Posttraumatic Stress Disorder app portfolio. Psychol Serv 2024;21(3):538-551. [doi: 10.1037/ser0000806] [Medline: 37917475]
36. VA Technical Reference Model v24.7. U.S. Department of Veterans Affairs. 2024. URL: https://www.oit.va.gov/Services/TRM/TRMHomePage.aspx [accessed 2024-09-08]
37. Blevins CA, Weathers FW, Davis MT, Witte TK, Domino JL. The posttraumatic stress disorder checklist for DSM-5 (PCL-5): development and initial psychometric evaluation. J Trauma Stress 2015;28(6):489-498. [doi: 10.1002/jts.22059] [Medline: 26606250]
38. Bovin MJ, Marx BP, Weathers FW, Gallagher MW, Rodriguez P, Schnurr PP, et al. Psychometric properties of the PTSD Checklist for Diagnostic and Statistical Manual of Mental Disorders-Fifth Edition (PCL-5) in veterans. Psychol Assess 2016;28(11):1379-1391. [doi: 10.1037/pas0000254] [Medline: 26653052]
39. Tennant R, Hiller L, Fishwick R, Platt S, Joseph S, Weich S, et al. The Warwick-Edinburgh mental well-being scale (WEMWBS): development and UK validation. Health Qual Life Outcomes 2007;5:63 [FREE Full text] [doi: 10.1186/1477-7525-5-63] [Medline: 18042300]
40. Green JA. Too many zeros and/or highly skewed? A tutorial on modelling health behaviour as count data with Poisson and negative binomial regression. Health Psychol Behav Med 2021;9(1):436-455 [FREE Full text] [doi: 10.1080/21642850.2021.1920416] [Medline: 34104569]
41. Hallenbeck HW, Jaworski BK, Wielgosz J, Kuhn E, Ramsey KM, Taylor K, et al. PTSD Coach Version 3.1: a closer look at the reach, use, and potential impact of this updated mobile health app in the general public. JMIR Ment Health 2022;9(3):e34744 [FREE Full text] [doi: 10.2196/34744] [Medline: 35348458]
42. Owen JE, Jaworski BK, Kuhn E, Makin-Byrd KN, Ramsey KM, Hoffman JE. mHealth in the wild: using novel data to examine the reach, use, and impact of PTSD coach. JMIR Ment Health 2015;2(1):e3935 [FREE Full text] [doi: 10.2196/mental.3935] [Medline: 26543913]
43. Kozlov E, Bantum E, Pagano I, Walser R, Ramsey K, Taylor K, et al. The reach, use, and impact of a free mHealth mindfulness app in the general population: mobile data analysis. JMIR Ment Health 2020;7(11):e23377 [FREE Full text] [doi: 10.2196/23377] [Medline: 33245289]
44. Yerkes RM, Dodson JD. The relation of strength of stimulus to rapidity of habit‐formation. J Comp Neurol Psychol 2004;18(5):459-482. [doi: 10.1002/cne.920180503]
45. Teigen KH. Yerkes-Dodson: a law for all seasons. Theory Psychol 1994;4(4):525-547. [doi: 10.1177/0959354394044004]
46. Shayani DR, Canale CA, Sloan DM, Hayes AM. Predictors of dropout in cognitive processing therapy for PTSD: an examination of in-session treatment processes. Behav Res Ther 2023;171:104428. [doi: 10.1016/j.brat.2023.104428] [Medline: 37952284]

XSL•FO
RenderX

47. Smith NB, Sippel LM, Rozek DC, Hoff RA, Harpaz-Rotem I. Predictors of dropout from residential treatment for posttraumatic stress disorder among military veterans. Front Psychol 2019;10:362 [FREE Full text] [doi: 10.3389/fpsyg.2019.00362] [Medline: 30873081]
48. Mackenzie CS, Pagura J, Sareen J. Correlates of perceived need for and use of mental health services by older adults in the collaborative psychiatric epidemiology surveys. Am J Geriatr Psychiatry 2010;18(12):1103-1115. [doi: 10.1097/jgp.0b013e3181dd1c06]
49. Weathers FW, Litz BT, Keane TM, Palmieri PA, Marx BP, Schnurr PP. The PTSD checklist for DSM-5 (PCL-5). U.S. Department of Veterans Affairs. 2024. URL: https://www.ptsd.va.gov/professional/assessment/adult-sr/ptsd-checklist.asp [accessed 2024-06-03]
50. Boucher EM, Raiker JS. Engagement and retention in digital mental health interventions: a narrative review. BMC Digit Health 2024;2(1):52. [doi: 10.1186/s44247-024-00105-9]
51. Hampole SR, Becket-Davenport CM, McCaslin-Rodrigo SE, Bosch JO, Jamison AL, Mackintosh M, et al. Facilitators and barriers to implementing mental health apps among interdisciplinary staff in the Veterans Health Administration. Psychol Serv 2025;22(4):664-687. [doi: 10.1037/ser0000934] [Medline: 39836148]
52. Hendrikx LJ, Williamson V, Murphy D. Adversity during military service: The impact of military sexual trauma, emotional bullying and physical assault on the mental health and well-being of women veterans. BMJ Mil Health 2023;169(5):419-424. [doi: 10.1136/bmjmilitary-2021-001948] [Medline: 34697241]
53. O'Brien HL, Morton E, Kampen A, Barnes SJ, Michalak EE. Beyond clicks and downloads: a call for a more comprehensive approach to measuring mobile-health app engagement. BJPsych Open 2020;6(5):e86 [FREE Full text] [doi: 10.1192/bjo.2020.72] [Medline: 32778200]

Abbreviations

DSM-5: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
ID: identification
JSON: JavaScript Object Notation
MST: military sexual trauma
NPC: negative posttraumatic cognition
PCL-5: PTSD (posttraumatic stress disorder) Checklist for DSM-5 (Diagnostic and Statistical Manual of Mental Disorders [Fifth Edition])
PMBS: Posttraumatic Maladaptive Beliefs Scale
PTSD: posttraumatic stress disorder
VA: US Department of Veterans Affairs
WEMWBS: Warwick-Edinburgh Mental Well-Being Scale

Edited by A Mavragani; submitted 30.Sep.2025; peer-reviewed by C Baxter, R Omachonu; comments to author 20.Jan.2026; revised version received 03.Mar.2026; accepted 25.Mar.2026; published 07.May.2026.

Please cite as:
Hampole S, McCaslin-Rodrigo S, Carr V, Schuster D, Reyes S, Senti S, Ramsey KM, Owen JE, Skidmore WC, Street AE, Taylor K, Mackintosh MA
Posttraumatic Symptoms as Predictors of Engagement With a Mobile App for Coping After Military Sexual Trauma: Public Usage Data Analysis Study
J Med Internet Res 2026;28:e85098
URL: https://www.jmir.org/2026/1/e85098
doi:10.2196/85098
PMID:

©Shilpa Hampole, Shannon McCaslin-Rodrigo, Valerie Carr, David Schuster, Sofia Reyes, Sarah Senti, Kelly M Ramsey, Jason E Owen, W Christopher Skidmore, Amy E Street, Katherine Taylor, Margaret-Anne Mackintosh. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Use of Health and Well-Being Technology, Basic Psychological Needs, and the Mediating Role of Technological Identity in 6 European Countries: Prospective Longitudinal Survey Study

Moona Heiskari[1], MSocSci; Aki Koivula[2], DrSocSci; Magdalena Celuch[1], PhD; Iina Savolainen[1], DrSocSci; Teijo Osma[1], MSocSci; Atte Oksanen[1], DrSocSci

[1]Faculty of Social Sciences, Tampere University, Tampere, Finland
[2]Faculty of Social Sciences, University of Turku, Turku, Finland

Corresponding Author:
Moona Heiskari, MSocSci
Faculty of Social Sciences
Tampere University
Kalevantie 5
Tampere, 33014
Finland
Phone: 358 504717890
Email: moona.heiskari@tuni.fi

Abstract

Background: Digital health technologies are increasingly used to monitor and improve personal health and well-being. Simultaneously, they can influence user behavior and self-understanding. As health technologies advance and are embedded in everyday life, understanding their broader psychological impacts, as well as the role of identity in shaping these outcomes, is crucial.

Objective: This prospective longitudinal survey study examined how usage of health and well-being technologies predicts experiences of new technology-related basic psychological needs over time, reflecting the psychological outcomes of digital health technologies. Furthermore, we investigated whether technological social identity serves as a pathway through which health technology use is associated with basic needs.

Methods: We used 3-wave survey data (2022-2024) collected from participants aged 18-75 years in Finland (n=1541), France (n=1561), Germany (n=1529), Ireland (n=1112), Italy (n=1530), and Poland (n=1533). Participants were recruited through a survey research panel, and follow-up data were collected from the same cohort. The sample was representative of the target populations by age (average age 46.79, SD 15.50 years) and gender (4304/8806, 48.88% male). The measure of digital health technology usage included smartphone health and well-being apps, well-being coaching apps, fitness trackers or watches, and smart rings. We applied a dynamic panel model within a structural equation modeling framework to examine both contemporaneous and lagged effects of technology use on the outcome variables—autonomy frustration, competence frustration, and relatedness satisfaction, measured with the Technology Effects on Need Satisfaction in Life scale. Mediation analysis was performed to assess whether social identification as a new technology user explained the relationship between technology use and need experiences.

Results: Usage of health and well-being technologies was associated with higher technology-mediated relatedness (β=.14, 95% CI 0.11-0.18; *P*<.001). Autonomy frustration (β=.06, 95% CI 0.02-0.10; *P*=.003) and competence frustration (β=.06, 95% CI 0.02-0.10; *P*=.008) also demonstrated small but positive connections with technology use. We found no statistically significant differences across countries. Mediation analysis revealed that the relationships between technology use and psychological needs were largely explained by social identification as a technology user.

Conclusions: Unlike most existing studies, this research focuses on the technology-related psychological effects of everyday health and well-being technologies and provides longitudinal and cross-national evidence of how they shape broader outcomes beyond health content and goals. The findings demonstrate that health technologies can both support users' social needs and undermine their personal sense of agency. They also provide insights into the significance and dual-edged role of technology users' social identity in transmitting these effects. The study highlights the central role of technology-related basic needs and

XSL•FO
RenderX

identity processes in understanding the larger outcomes of rapidly advancing digital health technologies. It offers valuable insights for health technology designs and implementations that aim to enable need-supportive technology engagement.

(J Med Internet Res 2026;28:e83054) doi:10.2196/83054

KEYWORDS

digital health; health information technology; mobile apps; psychological needs; psychological theory; technological identity

Introduction

Overview

Personal health and wellness management has become increasingly digitized over the past 20 years [1-3], providing personalized health-related content, predictive analytics, and recommendations, which are increasingly artificial intelligence– and algorithm-based [4-7]. Understanding the psychological aspects of technology-mediated well-being management is critical in an era of individuals relying on advanced technologies in monitoring and improving their personal health, both in personal and clinical health care settings [8,9]. Particularly, there is limited knowledge about the psychological outcomes of human-technology interaction in the context of digital health technologies and the mechanisms underlying this relationship [2,10,11].

The aim of this study is to investigate how using health and well-being technologies affects the new technology-related basic psychological needs—autonomy frustration, competence frustration, and relatedness satisfaction—and whether in-group identification as a new technology user mediates these relationships. Technology-related autonomy reflects engaging with new technologies in ways that support one's values and goals, whereas frustration arises when there is a sense of pressure to use technologies [12,13]. Technology-related competence refers to the extent to which new technologies help users feel capable and effective, whereas frustration reflects feelings of incapability and powerlessness [12,13]. Technology-enabled relatedness captures feeling socially connected to other people and communities through the use of new technologies [12]. Social identity processes are likely to shape these need-related experiences as technologies are increasingly embedded in users' self-concepts and social group memberships. Building on this insight, we propose that in-group identification may help explain why health technology use supports or frustrates psychological needs.

Well-Being Outcomes of Digital Health Technologies

Previous investigations on wearable and self-tracking technologies and mobile health apps have demonstrated positive associations between health technologies and well-being [14]. There is a significant amount of evidence that health application interventions are efficient in enhancing users' quality of life and have positive effects in terms of reducing stress, anxiety, and depression [15-17]. Simultaneously, a growing body of research is studying the negative effects of self-tracking and their causes. For instance, self-quantification can decrease self-esteem, which may then reduce well-being [14]. Failing to achieve goals or engaging in social comparison when using health technology can also elicit negative emotions, such as guilt, anxiety, and stress [18]. The features of technologies, such as app notifications and human-like characteristics, might also cause stress, annoyance, and negative behavioral outcomes [19,20].

Health and well-being technology usage is not solely an individual practice; it also includes a clear social component in terms of offline and online interaction [21-23]. Social networking and sharing data with other users are part of health technologies, their tactics of behavior change, and how they motivate users [2,24,25]. Social networking can, however, yield varied outcomes. For example, reciprocal support in fitness apps may enhance autonomous engagement with health activities, while social recognition may foster compulsive behaviors [26].

Basic Psychological Needs in the Use of Health Technologies

Self-determination theory (SDT) is a well-established theory that defines basic psychological needs of autonomy, competence, and relatedness as vital nutriments for well-being [27,28]. Extensive evidence has shown that the fulfillment of these needs is linked to well-being and happiness [29], while need thwarting predicts ill-being and negative outcomes [30]. Need satisfaction and need frustration are empirically distinct experiences, meaning that lack of need satisfaction does not necessarily equate to experiencing frustration and vice versa [13,27].

The motivation, engagement, and thriving in user experience model describes basic psychological needs as key mediators of the relationship between technology and well-being [12,31]. Existing research indicates that digital health technology can predominantly support users' competence and autonomy needs, especially when it increases their knowledge, awareness, and control over healthy behaviors and lifestyles [21,32-34]. However, the available evidence suggests that the significance of feedback provided by wearable activity trackers diminishes over time, and that the initial feelings of inspiration may fade, leading to a lack of self-responsibility [21,34]. In terms of social connectedness, trackers can foster a sense of belonging and social support from family and friends who use the trackers together as well as from other users in online communities and platforms [21]. Prior research has shown that among users of online fitness communities, self-regulatory motivation—stemming from experiences of autonomy and competence—is particularly important for new users, whereas social motivation and enjoyment become more prominent for long-term users [35].

XSL•FO
RenderX

Technological Social Identity as a Mediator of the Relationship Between Health Technology Use and Basic Needs

Personal IT identity has been shown to influence behavioral outcomes in technology use, and higher identification with technology has been found to predict broader and more explorative use [36,37]. Although the role of identity is increasingly recognized in research as human-technology interaction intensifies [38], evidence of the role of social identification in health technology use is lacking. The social identity approach emphasizes the human tendency to identify with social groups to which individuals perceive themselves as belonging, even when minimal cues of in- and out-group categorizations are present [39,40].

A large body of research evidence supports the fact that social identity predicts well-being outcomes, both positively and negatively, depending on the group's qualities and status [41-43]. Studies on basic psychological needs have demonstrated similar associations [44,45]. One recent cross-national study demonstrated that in-group identification with technology users is systematically associated with basic psychological need outcomes, highlighting identification as a meaningful social context of technology use [46]. In terms of digital health technologies, early technology adopters can form a significant reference group with other early adopters and create a community—concrete or imaginary—among similar-minded people. When internalized as part of oneself, this group membership is prone to shape one's interpretation and behavior when promoting uniformity in perception among group members, sharing values and norms, seeing oneself as similar to a group prototype, and comparing oneself to the group norm [47-49]. Thus, identifying as a technology user, which includes the norms and expectations one associates with the group, might influence the relationship and behaviors one has with health technology use, subsequently influencing the psychological outcomes.

As a mechanism linking health technology use and basic psychological needs, in-group identification as a technology user may support the user's autonomy and mitigate their frustration as technology use is a part of one's self-concept. Thus, aligning with group values and norms can promote internalized motivation to use technology [44]. In-group identification may enhance health technology users' feelings of efficiency and mastery and decrease the competence frustration, as groups can help their members to feel more capable and personally in control [50]. Also, the group's prototypical qualities can increase the individual's sense of competence [49]. The relatedness of health technology users may be supported by identification with other technology users, particularly because it emphasizes belonging to a group of people with similar values, behaviors, and goals in terms of technology use [47,48]. As health technologies involve social components, identifying with other technology users can make an individual feel more connected and allow them to share experiences and obtain social validation from similar-minded individuals.

This Study

The objective of this prospective longitudinal survey study was to investigate how health and well-being technology use is associated with technology-related basic psychological needs, and whether in-group identification as a technology user mediates this relationship. Our study design and hypotheses were preregistered in the Open Science Framework before the data analysis [51]. In the preregistration, we described SDT variables as need satisfaction. However, after revisiting the operationalization of our measures, we recognized that the autonomy and competence measures are more accurately conceptualized as need frustration. The hypotheses were therefore refined to reflect this operationalization. The fourth hypothesis was included after preregistration. The following hypotheses were proposed:

- H1: Health and well-being technology usage predicts lower new technology-related autonomy frustration.
- H2: Health and well-being technology usage predicts lower new technology-related competence frustration.
- H3: Health and well-being technology usage predicts higher new technology-related relatedness satisfaction.
- H4: The associations between the use of health and well-being technology and technology-related needs are mediated by in-group identification as a new technology user.

Methods

Participant Characteristics and Study Design

This study was reported following the American Psychological Association's Journal Article Reporting Standards for Quantitative Research [52]. Our longitudinal survey data were collected as part of the Self & Technology research project, targeting participants aged 18-75 years from 6 European countries: Finland, France, Germany, Ireland, Italy, and Poland. The included European Union (EU) countries share major cultural and societal similarities but differ significantly when it comes to digitalization and the use of new technologies, making them suitable for investigating the role of health and well-being technologies in psychological functioning. The structured online survey was first conducted in October-November 2022, with a total sample size of 8806: Finland (n=1541), France (n=1561), Germany (n=1529), Ireland (n=1112), Italy (n=1530), and Poland (n=1533). Follow-up surveys were conducted in October-November 2023 and 2024 and targeted the same respondents who took part at time point 1. At T2, the response rate was 71% (1095/1541) in Finland, 65% (1014/1561) in France, 59% (900/1529) in Germany, 53% (588/1112) in Ireland, 72% (1099/1530) in Italy, and 63% (967/1533) in Poland. At T3, the response rate reached 84% (924/1095) in Finland, 86% (872/1014) in France, 82% (736/900) in Germany, 80% (468/588) in Ireland, 86% (946/1099) in Italy, and 76% (731/967) in Poland. The longitudinal sample flow, detailing participation and attrition across T1-T3 is presented in Figure 1.

XSL•FO
RenderX

Figure 1. Study participant flowchart.

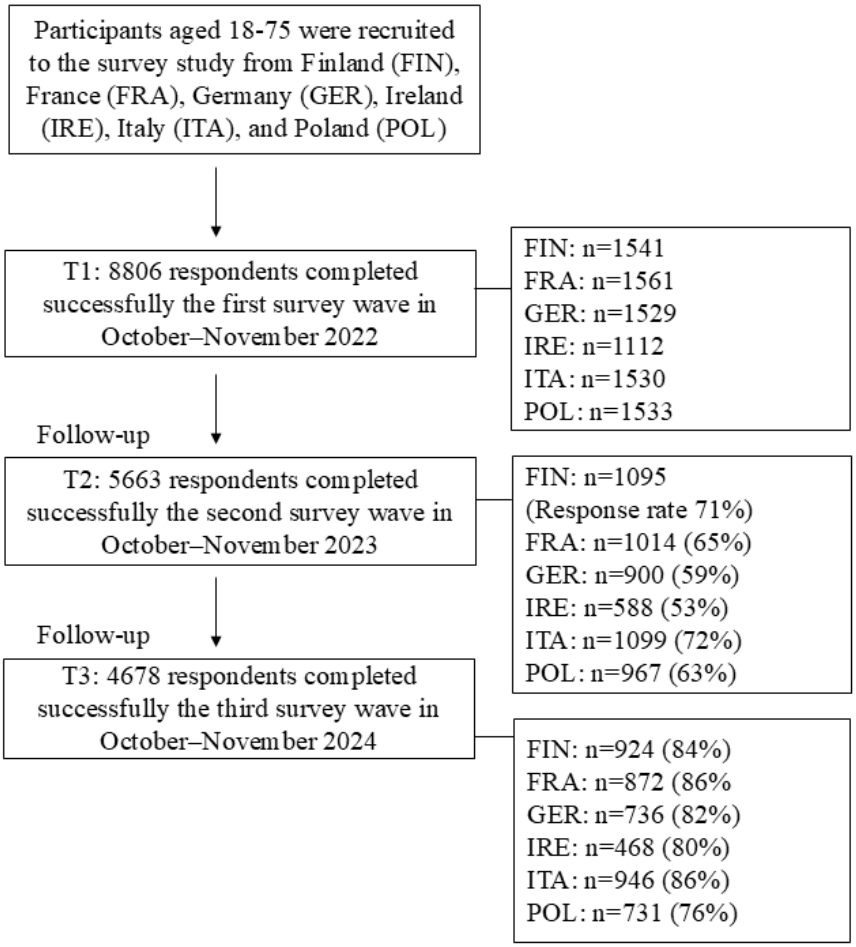

The samples represent the age and gender demographics of the target populations in their respective countries [53]. Response rates were high at both follow-up times, 64% (5663/8806) of all the participants returned to the study at T2. In longitudinal studies, higher attrition typically occurs during the first follow-up, which was also visible in our data overall. At T3, a total of 53% (4678/8806) of T1 participants responded to the survey. Upon inspection, younger participants were more likely to drop out. This pattern of attrition is common in longitudinal studies and is often explained by changes in life circumstances over time. Missing data were examined with Little's missing at completely random test and addressed in the analyses, as described in the Statistical Analyses section.

Data Collection

Data collection was carried out by Norstat, a data solution provider. In the first study round, respondents were recruited from Norstat's panel of voluntary research participants. Participants from the panel were invited via email to fill out the questionnaire about Europeans' use of technology. Respondents to the first survey were reinvited via email to respond to the second and third study waves. The survey was originally designed in English and then expertly translated into the necessary languages by native speakers using the back-translation technique. Validated scales were used in each language when available. Data quality checks, such as patterned responses and attention checks, were conducted after each round of data collection to ensure data quality. Based on this, any questionable respondents were dropped from the final sample.

Ethical Considerations

The study protocol was approved in 2022 by the Ethics Committee of the Tampere Region, which stated that the study did not involve ethical concerns (statement 115/2022). The survey respondents were provided with information about the study and the use of data for research purposes, highlighting that by completing the survey, respondents provided their informed consent to take part in the study. Respondents received compensation for their participation from Norstat. The company awards reward coins for each completed survey; the monetary value was approximately €2.60 (US $3.04) per survey in Finland, Poland, and Ireland, and €2.50 (US $2.93) in Germany, France, and Italy. These coins can be exchanged for gift cards, donated to charity, or used to enter prize draws. Participation in the study was voluntary, and the participants were informed about their right to withdraw their participation at any time without consequences. Participants were anonymous, and no information in the dataset allows for their identification. No individual participant can be identified from any part of the study materials. In addition, the study adheres to the EU's General Data Protection Regulation.

Measures and Covariates

Our measures of new technology-related autonomy, competence, and relatedness were adapted from the Technology Effects on

XSL•FO
RenderX

Need Satisfaction in Life scale [12], which is based on the well-validated Basic Psychological Need Satisfaction and Frustration Scale [54]. We examined the degree to which respondents perceived new technologies to have influenced their sense of autonomy, competence, and relatedness in life overall. Ten statements were rated on a scale from 1 (does not describe me at all) to 7 (describes me completely). Four items measured sense of autonomy frustration (eg, "The new technologies intrude in my life"). One item was dropped to gain a better fit value in the confirmatory factor analysis conducted with structural equation modeling (SEM), resulting in a 3-item autonomy variable. Three items measured sense of competence frustration (eg, "Using the new technologies has made me feel insecure about my abilities"), and 3 items measured sense of relatedness satisfaction (eg, "Because of these new technologies, I feel closer to some others"). The full list of items is included in Multimedia Appendix 1. Higher scores for autonomy and competence express higher need frustration, while higher scores for relatedness indicate higher need satisfaction. The final scales ranged from 3 to 21. The reliability of the autonomy, competence, and relatedness measures varied from good to excellent at all 3 time points and across all 6 countries. See Multimedia Appendix 2 for the McDonald's reliability figures of our measures.

As the independent variable, participants' use of digital health and well-being technologies (wtech) was measured, including the use of smartphones and well-being apps, well-being coaching apps, fitness trackers and watches, and smart rings. The frequency of using these technologies was measured on a scale of 0-5 (0=do not use, 5=uses multiple times a day). We developed a sum variable to investigate the overall intensity of respondents' usage of health and well-being technologies.

In-group identification with new technology users (IGI) was measured using a scale modified from the hierarchical model of in-group identification [55]. Out of the 5 components of in-group identification, we used the satisfaction dimension to measure the extent to which participants expressed positive feelings toward a group of new technology users and their membership in it. Positive feelings, conceptualized as satisfaction, are a direct and clear indicator of in-group identification that also indicate affective ties to the group and group-level self-investment [55,56]. Participants evaluated 4 statements (eg, "It is pleasant to be a user of new technologies") on a scale of 1 (strongly disagree) to 7 (strongly agree). The reliability of the measure was excellent at all time points and across all countries.

For the control variables, we included the participants' overall health status. We asked, "How would you evaluate your present health?" The participants were asked to provide an answer on a scale of 1 (very poor) to 7 (very good). As a straightforward and subjective assessment of one's health, the self-rated health item effectively acts as a summary measure for different health dimensions, allowing respondents to select their own emphasis [57]. To assess self-esteem, we asked participants to rate the statement "I have high self-esteem" on a scale of 1 (not very true of me) to 7 (very true of me). The Single-Item Self-Esteem Scale used here has been found to be valid and has high convergent validity with longer global self-esteem scales in adult populations [58]. We also measured participants' orientation for social comparison with a 6-item version of the Iowa-Netherlands Comparison Orientation Measure [59]. This scale is a shortened version of the 11-item version created by Gibbons and Buunk [60]. Finally, we asked study participants to evaluate the question "All things considered, how happy would you say you are?" on a scale of 1 (extremely unhappy) to 7 (extremely happy). Measuring the general sense of happiness with this single item has been found to be reliable and valid, as well as widely applicable, across different countries and cultures [61].

Sociodemographic variables included age, gender, and education. Age was measured in years. For gender, participants chose from the following options: "male," "female," or "other." For the analysis, we created a dummy variable to reflect male vs other genders (0=not male gender, 1=male gender). For education level, participants were asked whether they had a university degree (at least a bachelor's degree) and coded as 0 (no higher degree) and 1 (higher degree).

Statistical Analyses

We performed dynamic panel models using the xtdpdml command in Stata (version 17; Stata Corp LLC) and conducted full information maximum likelihood estimation with standardized coefficients [62,63]. Within the SEM framework, this multilevel SEM method allowed us to study how the independent variable—use of health and well-being technology—influenced the dependent variables (ie, each technology-related need), while accounting for the previous level of need frustration or satisfaction. It further enabled us to estimate the lagged effects of health and well-being technology use. Simultaneously, with the within-level fixed effects, the dynamic panel model method allowed us to control for unobservable confounders [62,64]. We also included country-level dummy variables in our models to account for potential country differences. As the reference category, Finland was omitted, and the model included dummy variables for all other countries. The dummies were included as time-invariant covariates using the inv() command with xtdpdml. All models had a good fit based on traditional SEM thresholds (Multimedia Appendix 3), but these have limited relevance in dynamic panel models. For a more detailed analysis of country differences, we used random-effects within-between models to examine how within-person variation in the use of health technology is associated with basic needs across different countries.

Finally, we used a Monte Carlo simulation approach to examine whether IGI mediated the contemporaneous effect of health and well-being technology use (wtech) on the outcomes of technology-related autonomy, competence, and relatedness. For each outcome, we estimated the effect of wtech on IGI (path a) and the effect of IGI on the outcome variable (path b) using xtdpdml, which accounts for the lagged effects of the outcomes, individual-level fixed effects, and country differences. We then simulated 20,000 draws from normal distributions defined by the estimated coefficients and their SEs to generate a distribution of the indirect effect ($a \times b$). From this distribution, we derived the mean indirect effect, 95% CIs, and 2-tailed P values.

XSL•FO
RenderX

Measurement invariance for the SDT and IGI factors was tested across countries using multigroup confirmatory factor analysis with maximum likelihood estimation using robust SEs in R software (version 4.5.2; R Foundation for Statistical Computing) with the *lavaan* package (0.6-20). We evaluated the overall fit of the initial model using widely applied fit indices and criteria, including the chi-square test of model fit (χ^2), root-mean-square error of approximation (RMSEA)<.06, comparative fit index (CFI)<.95, and standardized root-mean-square residual (SRMR)<.08 [65]. Because the chi-square test is highly sensitive to large sample sizes and may inflate RMSEA, we relied on alternative fit indices and interpreted RMSEA with caution. When adding constraints and comparing nested models, we used the criteria of ΔCFI<0.01, ΔRMSEA<0.015, and ΔSRMR<0.03 to indicate support for each level of invariance [66].

Missing Data

Missing data resulted from survey dropout across waves. We performed Little's missing completely at random test on the continuous variables in the analytic models. The result indicated that the data were not missing completely at random (χ^2_{71}=879.42; *P*<.001). Therefore, we handled missing data with full information maximum likelihood, which uses all available data and provides valid estimates under the missing at random assumption. No auxiliary variables were included.

Results

Table 1 presents the results of measurement invariance analyses. The initial configural model demonstrated good fit to the data (χ^2_{354}=2660.75; CFI=0.96; RMSEA 0.081, 90% CI 0.078-0.083; SRMR=0.062). RMSEA indicated acceptable fit given the large sample size, and the other fit indices suggested good fit, thus supporting further testing. Constraining the loadings across groups and comparing the fit indices supported metric invariance (ΔCFI=-0.001; ΔRMSEA=–0.004; ΔSRMR=0.001). Further constraining the intercepts (ΔCFI=–0.008; ΔRMSEA=0.004; and ΔSRMR=0.002) and residuals (ΔCFI=0.010; ΔRMSEA=0.003; and ΔSRMR=0.001) supported a strict level of invariance across countries.

Table 1. Measurement invariance results for basic psychological needs and in-group identification with new technology users across the countries.

	Chi-square (*df*)	Δ chi-square	*P* value (Δ chi-square)	CFI[a]	ΔCFI	RMSEA[b] (90% CI)	ΔRMSEA	SRMR[c]	ΔSRMR
Configural	2660.75 (354)	N/A[d]	N/A	0.96	N/A	0.08 (0.08-0.08)	N/A	0.06	N/A
Metric	2831.07 (399)	117.85	<.001	0.96	–0.001	0.08 (0.07-0.08)	0.00	0.06	0.00
Scalar	3534.48 (444)	849.86	<.001	0.96	–0.008	0.08 (0.08-0.08)	0.00	0.07	0.00
Strict	3899.14 (509)	429.07	<.001	0.95	–0.010	0.08 (0.08-0.09)	0.00	0.07	0.00

[a]CFI: comparative fit index.

[b]RMSEA: root-mean-square error of approximation.

[c]SRMR: root-mean-square residual.

[d]N/A: not applicable.

Tables 2 and 3 present a descriptive summary of the variables included in this study, covering data from Finland, France, Germany, Ireland, Italy, and Poland. Across countries, the average score for technology-related autonomy frustration reached 8.50 at T1 and 7.61 at T3. A fixed-effects analysis showed a statistically significant decrease in within-person effects across all countries (*P*<.001). Competence frustration scores were 7.53 at T1 and 7.47 at T3, but did not show a statistically significant change. Relatedness satisfaction scored 9.45 at T1 and decreased to 8.11 at T3 (*P*<.001). The use of health and well-being technologies (wtech) showed a low level of 3.28 at T1 but moderately increased to 3.41 at T3 (*P*=.01). Finally, IGI scored, on average, 13.91 at T1 and decreased to 13.34 at T3 (*P*<.001). The fixed effects of the main variables are reported in detail in Multimedia Appendix 4.

XSL•FO
RenderX

Table 2. Descriptive statistics by country (Finland: n=1541; France: n=1561; and Germany: n=1529). Values for continuous variables are reported as mean, between-person SD (SDB), and within-person SD (SDW) across all time points. Age and categorical variables are reported from T1. Sample characteristics: Finland—male, 766 (47.71%); higher education, 413 (26.80%). France—male, 744 (47.66%); higher education, 572 (36.64%). Germany—male, 761 (49.77%); higher education, 450 (28.43%).

Variable		Finland			France			Germany		
	Range	Mean	SD_B	SD_W	Mean	SD_B	SD_W	Mean	SD_B	SD_W
Autonomy frustration	3-21	7.30	3.43	1.98	8.64	4.06	2.37	7.52	4.09	2.28
Competence frustration	3-21	6.62	3.59	2.02	8.35	4.20	2.55	6.87	4.07	2.25
Relatedness satisfaction	3-21	8.04	3.95	2.27	8.74	4.39	2.31	7.76	4.39	2.30
Wtech	0-16	3.21	3.10	1.23	2.27	3.00	1.22	2.48	3.11	1.17
IGI[a]	4-28	12.26	5.75	2.17	14.23	6.08	2.14	13.27	6.47	2.23
Self-rated health	1-7	4.68	1.24	0.56	4.80	1.29	0.53	4.57	1.41	0.56
Self-esteem	1-7	4.82	1.42	0.56	4.09	1.51	0.64	4.70	1.57	0.58
Social comparison	5-35	18.04	5.74	2.57	17.30	5.61	2.74	17.18	6.02	2.60
Happiness	1-7	4.89	1.11	0.52	4.67	1.18	0.57	4.66	1.27	0.71
Age (years)	18-75	46.35	16.34	N/A[b]	46.88	15.84	N/A	47.36	15.16	N/A

[a]IGI: in-group identification with new technology users.

[b]N/A: not applicable.

Table 3. Descriptive statistics by country (Ireland: n=1112; Italy: n=1530; and Poland: n=1533). Values for continuous variables are reported as mean, between-person SD (SDB), and within-person SD (SDW) across all time points. Age and categorical variables are reported from T1. Sample characteristics: Ireland—male, 540 (48.56%); higher education, 564 (50.72%). Italy—male, 749 (48.95%); higher education, 592 (38.69%). Poland—male, 744 (48.53%); higher education, 835 (54.47%).

Variable		Ireland			Italy			Poland		
	Range	Mean	SD_B	SD_W	Mean	SD_B	SD_W	Mean	SD_B	SD_W
Autonomy frustration	3-21	8.07	4.09	2.17	8.75	3.85	2.29	7.95	3.86	2.27
Competence frustration	3-21	6.85	3.99	2.26	8.25	4.19	2.54	7.77	4.17	2.38
Relatedness satisfaction	3-21	8.70	4.46	2.36	10.86	4.36	2.39	8.79	4.46	2.48
Wtech	0-16	3.74	3.34	1.23	3.39	3.48	1.56	3.64	3.78	1.63
IGI[a]	4-28	14.63	5.86	2.07	14.90	6.05	2.21	14.42	6.16	2.32
Self-rated health	1-7	4.91	1.31	0.53	4.97	1.21	0.55	4.58	1.25	0.55
Self-esteem	1-7	4.22	1.58	0.64	4.39	1.57	0.70	4.08	1.56	0.73
Social comparison	5-35	19.29	5.29	2.32	21.22	4.54	2.59	18.99	5.44	2.64
Happiness	1-7	4.87	1.25	0.69	4.65	1.23	0.59	4.74	1.25	0.64
Age (years)	18-75	46.77	14.50	N/A[b]	47.67	15.34	N/A	45.69	15.41	N/A

[a]IGI: in-group identification with new technology users.

[b]N/A: not applicable.

Table 4 reports the findings from Model 1, with separate analyses conducted for new technology-related autonomy frustration, competence frustration, and relatedness satisfaction. The wtech had a small but positive association with autonomy frustration (β=.06, 95% CI 0.02-0.10; *P*=.003), the CI not including zero, indicating a small but consistently positive population effect. Competence frustration similarly indicated a statistically significant positive effect (β=.06, 95% CI 0.02-0.10; *P*=.008). Conversely, wtech showed a positive association with relatedness satisfaction (β=.14, 95% CI 0.11-0.18; *P*<.001), the estimate remaining positive across the confidence bound. The effect of wtech on autonomy frustration, competence frustration, and relatedness satisfaction remained similar even when other factors were controlled (Model 4 in the Multimedia Appendix 5). With the lagged predictor, we found no evidence of a lagged effect of wtech use on autonomy, competence, or relatedness (Model 3 in the Multimedia Appendix 6).

XSL•FO
RenderX

Table 4. Dynamic panel models predicting the effect of health and well-being technology usage on technology-related basic psychological needs, presented without (Model 1) and with the mediator, in-group identification with new technology users (IGI; Model 2). Longitudinal time frame including T1-T3 (N=8806).

	Model 1			Model 2		
	β (SE)	*P* value	95% CI	β (SE)	*P* value	95% CI
Autonomy frustration						
Autonomy lag	0.06 (0.03)	.02	0.01 to 0.11	0.05 (0.02)	.03	0.01 to 0.10
Wtech	0.06 (0.02)	.003	0.02 to 0.10	0.01 (0.02)	.58	–0.03 to 0.05
IGI	N/A[a]	N/A	N/A	0.38 (0.02)	<.001	0.34 to 0.43
Country (reference=Finland)						
France	0.12 (0.01)	<.001	0.09 to 0.14	0.06 (0.01)	<.001	0.04 to 0.09
Germany	–0.01 (0.01)	.36	–0.04 to 0.01	–0.04 (0.01)	.005	–0.06 to –0.01
Ireland	0.04 (0.01)	.004	0.01 to 0.06	–0.01 (0.01)	.46	–0.04 to 0.02
Italy	0.12 (0.01)	<.001	0.10 to 0.15	0.06 (0.01)	<.001	0.03 to 0.08
Poland	0.04 (0.01)	.002	0.01 to 0.07	–0.01 (0.01)	.48	–0.04 to 0.02
Competence frustration						
Competence lag	0.09 (0.03)	.003	0.03 to 0.15	0.09 (0.03)	.004	0.03 to 0.14
Wtech	0.06 (0.02)	.008	0.02 to 0.10	0.03 (0.02)	.16	–0.01 to 0.07
IGI	N/A	N/A	N/A	0.21 (0.03)	<.001	0.17 to 0.26
Country (reference=Finland)						
France	0.13 (0.01)	<.001	0.11 to 0.16	0.11 (0.01)	<.001	0.08 to 0.13
Germany	0.01 (0.01)	.40	–0.02 to 0.04	0.00 (0.01)	.86	–0.03 to 0.02
Ireland	0.01 (0.01)	.50	–0.02 to 0.03	–0.02 (0.01)	.22	–0.04 to 0.01
Italy	0.13 (0.01)	<.001	0.10 to 0.15	0.09 (0.01)	<.001	0.06 to 0.12
Poland	0.09 (0.01)	<.001	0.06 to 0.11	0.06 (0.01)	<.001	0.03 to 0.09
Relatedness satisfaction						
Relatedness lag	0.04 (0.03)	.09	–0.01 to 0.09	0.00 (0.02)	.96	–0.04 to 0.03
Wtech	0.14 (0.02)	<.001	0.11 to 0.18	0.06 (0.02)	.001	0.02 to 0.09
IGI	N/A	N/A	N/A	0.65 (0.02)	<.001	0.61 to 0.69
Country (reference=Finland)						
France	0.09 (0.01)	<.001	0.06 to 0.11	0.00 (0.01)	.75	–0.02 to 0.02
Germany	–0.01 (0.01)	.63	–0.03 to 0.02	–0.05 (0.01)	<.001	–0.07 to –0.03
Ireland	0.04 (0.01)	.004	0.01 to 0.06	–0.04 (0.01)	<.001	–0.07 to –0.02
Italy	0.23 (0.01)	<.001	0.20 to 0.26	0.13 (0.01)	<.001	0.11 to 0.15
Poland	0.06 (0.01)	<.001	0.04 to 0.09	–0.02 (0.01)	.07	–0.04 to 0.00

[a]N/A: not applicable.

Model 1 illustrates the country differences of need frustration and satisfaction, with Finland acting as a reference point. Compared to Finland, autonomy was lower in all other countries ($P<.01$) except Germany. Competence was lower in France, Italy, and Poland ($P<.001$). The other countries, except Germany, showed higher relatedness than Finland ($P<.01$). Figure 2 depicts the country-level patterns in the within-individual (person-mean centered) association between the use of digital health and well-being technologies (wtech) and autonomy frustration (left panel), competence frustration (middle panel), and relatedness satisfaction (right panel). The lines represent predicted values derived from the random-effect within-between models, with other variables held constant. The x-axis reflects within-person deviations in technology use, and the y-axis shows standardized predicted levels of the respective need outcomes.

XSL•FO
RenderX

Figure 2. Within-person associations between health and well-being technology use and technology-related basic psychological needs across 6 countries (N=8806).

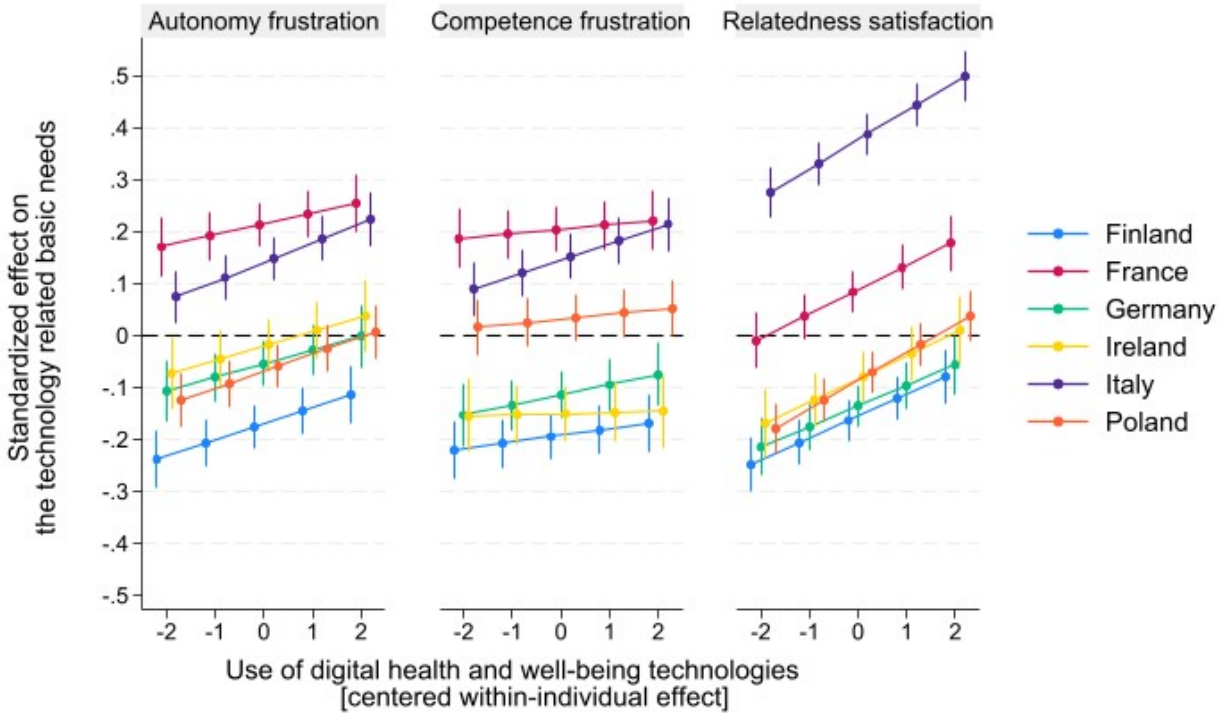

The results indicate that countries have broadly similar trends in their within-level associations between wtech and autonomy frustration and wtech and relatedness satisfaction. In most countries, predicted autonomy frustration increases as within-individual technology use increases. At the same time, predicted relatedness satisfaction increases consistently with higher levels of technology use across all 6 countries. In contrast, the pattern for competence frustration is more heterogeneous. A statistically significant positive association was observed in Italy (B=0.03, 95% CI 0.02-0.05; *P*<.001) and a marginally significant association in Germany (B=0.02, 95% CI 0.00-0.04; *P*=.07), whereas the association was negligible in the remaining countries.

Overall, in the random-effects within-between model (Multimedia Appendix 5), the main within-person effect of wtech on competence frustration became nonsignificant after including the country interaction terms. However, the reduction in the coefficient was approximately 15%, indicating that the association remained largely similar in magnitude and that cross-national variation in the within-person effect was modest. Moreover, formal tests of the cross-level interaction terms showed that none of the countries differed significantly from Finland in the strength of the within-individual association between technology use and autonomy frustration, competence frustration, or relatedness satisfaction (Multimedia Appendix 5). Thus, despite minor differences in predicted trends, the within-person associations did not significantly vary across countries relative to the reference category.

Finally, Table 4 presents Model 2, which includes the IGI variable. Fully illustrated in Figure 3, mediation analyses showed that IGI is a significant factor explaining the association between wtech and technology-related need frustration for autonomy and competence and satisfaction for relatedness. Using a Monte Carlo simulation approach, we found that the indirect effect of IGI on autonomy frustration was positive, with the lower CI boundary above zero (β=.05, 95% CI 0.04-0.06; *P*<.001), accounting for 85.7% of the total effect. For competence frustration, the indirect effect was also positive, the CI excluding zero (β=.03, 95% CI 0.02-0.04; *P*<.001), accounting for 48.7% of the total effect. The indirect effect of IGI on relatedness was positive (β=.09, 95% CI 0.07-0.10; *P*<.001), accounting for 59.6% of the total effect.

XSL•FO
RenderX

Figure 3. Mediation models showing the indirect association between health and well-being technology use and psychological needs through in-group identification with new technology users (N=8806).

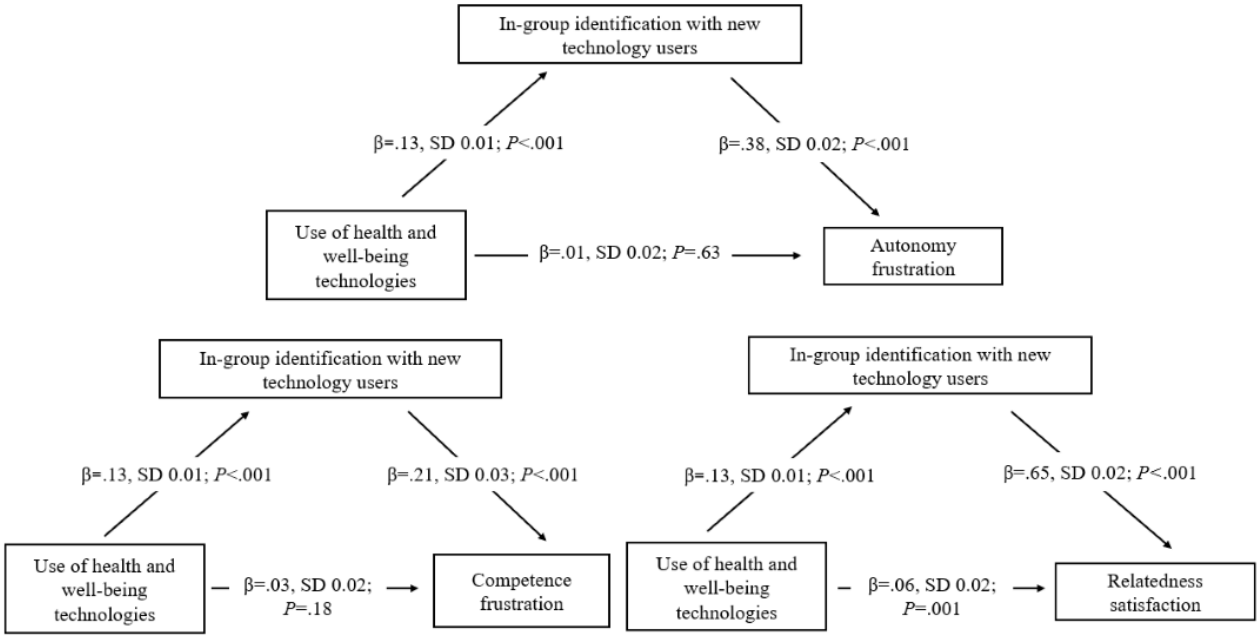

Discussion

Main Findings

This prospective longitudinal survey study investigated the influence of health and well-being technology use, such as the use of smart rings, well-being apps, and fitness trackers, on basic psychological needs related to technology. Furthermore, it examined whether IGI mediated this association. Across 6 countries—Finland, France, Germany, Ireland, Italy, and Poland—we hypothesized that more intensive health and well-being technology usage would predict lower levels of new technology-related autonomy frustration (H1) and competence frustration (H2), and higher relatedness satisfaction (H3) over time. As we expected, higher health technology use was associated with higher relatedness. However, more intensive technology usage was associated with higher frustration with autonomy and competence. The mediation analysis revealed that IGI largely explained the outcomes of health and well-being technology use in terms of technological need frustration and satisfaction.

The 6 countries showed similar within-level associations between health technology use and autonomy and relatedness, with increased use generally linked to higher autonomy frustration and higher relatedness satisfaction. In terms of competence frustration, we found notable within-level associations only in Germany and Italy. However, with Finland as a reference country, no significant variation between countries was found in terms of the impact of health technology use on need frustration and satisfaction in general. Additionally, our investigation of the lagged effect of health technology use revealed that the contemporaneous effect of technology usage is more significant than the delayed impact on needs frustration and satisfaction.

Theoretical and Practical Implications

Previous research has found positive impacts of health technologies on users' health and well-being [14-16] but has also increasingly identified their adverse effects [18]. Our results indicate that health technologies' overall effects on well-being are not solely determined by measurable outcomes. Rather, they are also closely tied to psychological processes that address individual needs. Based on SDT, our study offers a social psychological perspective on the role of health and well-being technology in responding to users' basic psychological needs, showing that users may experience new technology-related autonomy and competence frustration despite simultaneously feeling that their relatedness needs are supported. Previous studies have demonstrated that digital health technologies can enhance users' autonomy and competence when fostering a sense of control and providing information and resources [21,32-34]. Our study indicates that digital health and well-being technology users may experience their engagement with technologies as lacking self-endorsement and undermining their perceived sense of capability. This frustration of human agency is linked to the adverse effects associated with health technologies [18,31,67].

Focusing particularly on new technology-related needs, our study provides knowledge on technology's role in psychological outcomes of health and well-being technologies that are increasingly sophisticated and incorporating artificial intelligence–driven features and advanced analytics [4-6]. This approach provides a novel understanding of technology-mediated processes of digital health management and their role on users' basic psychological needs in life overall [12,31]. Understanding life-level impacts of new technologies is crucial, especially because health and well-being technologies are closely embedded into everyday routines, targeting users' behavior and self-understanding. Overall, a more nuanced understanding of the psychological effects of human-technology interaction in digital health is needed, since most of the studies

XSL•FO
RenderX

in the field have examined health technology acceptance and adoption [68].

Our study also investigated everyday health and well-being technologies more broadly, allowing us to identify general patterns, acknowledging that users may simultaneously use multiple health-related devices, applications, and platforms. This differs from the majority of health technology research that has examined the impacts related to tasks and behaviors with a focus on specific applications, interventions, and health conditions [14,69]. The longitudinal design also allowed for observing long-term trends, which is crucial in light of previous scholarship outlining the positive influence of digital health technologies may diminish over time [21,34].

This study also aimed to examine social identity–related processes underlying health technology use, and particularly how possessing a technology user's identity contributes to well-being outcomes. IGI reflects a group-based identity that stems from the social categorizations and positioning of an individual among similarly minded technology users [39]. Our study tested whether this technological social identity mediates the association between well-being technology use and psychological needs. The social identity approach supports the notion that group membership affects the perception and behaviors of individuals [70]. In the context of technologies, norms and expectations of the technology users' group are likely to shape users' engagement with technologies, motivations, and future outcomes of health and well-being technology use. Our results indicated that the technology-enabled relatedness of health technology users is likely to be supported by IGI. Belonging to a group of people who share similar values, behaviors, and goals in terms of new technology seems to help satisfy relatedness in digital settings [43,44]. As health technologies involve social components, identifying with technology users might help an individual feel connected with other users and allow them to share experiences and receive social validation from like-minded individuals. This aligns with previous findings that users with a stronger personal IT identity seek more social engagement, as they are more willing to share their health information with others in health apps [71].

IGI did not support autonomy in health technology use; instead, it mediated higher autonomy frustration. Rather than promoting autonomous and self-directed technology use, technological social identity was linked to higher autonomy frustration, indicating more external pressure to use new technologies and a lack of personal control [12]. Similarly, technological identity mediated higher competence frustration, which suggests that instead of supporting mastery and efficiency in technology use, social technology user identity seemed to diminish technology-enabled competence feelings, connected to experiences of incapability and inadequacy [13]. Together, these results suggest that when technology is an integrated part of the social self, it can promote group norms and pressure an individual to use technology in a certain manner [37,42]. The results also suggest that when facing failure or negative comparisons to others—particularly prototypical or ideal technology users—satisfaction of autonomy and competence is diminished. This may be a potential downside for people with a stronger personal IT identity who use technology in a more intensive and explorative way [37,72]. IT identity has been similarly linked to embeddedness and enjoyment of technology use, but also to maladaptive emotions and problematic use [73]. Additionally, among new technology users, there may be more pressure to stay on track, master new technologies, and keep updated with the constant developments.

Our findings highlight the need to acknowledge the social identity perspective in health technology use, which typically involves social components and social gamification features [35,74,75]. Well-being supportive technology design should consider how to better enhance user agency in terms of competence and autonomy in technology use, but also take into account the group processes formed around technology use, including attitudes, norms, and expectations that shape how users experience and engage with technologies. Leveraging the social aspects of health and well-being technologies is a significant strength, as it can provide users with meaningful group memberships and a sense of belonging, thus supporting well-being beyond numerical tracking or goal achievement as well as long-term usage [22]. Future research should also investigate the factors shaping the overall psychological outcomes of health and well-being technology use in more detail.

Strengths and Limitations

This study is based on longitudinal and cross-country evidence, both of which are needed in the field of digital health and well-being technology research. Our longitudinal and cross-national method enabled a nuanced investigation of the relationships between health and well-being technology use and the technology-related basic psychological needs, while also accounting for the role of IGI. However, this study is limited to self-reported data and to a European context; thus, any generalizations should be made with caution. Health and well-being technology use was operationalized through the use of smartphones, well-being apps, coaching apps, fitness trackers and watches, and smart rings. Hence, we did not inquire about the specific purposes for which these devices and apps were used. Future research should explore more detailed usage patterns and specific types of health technologies in relation to the basic psychological needs and technology user identity.

Conclusion

This study offers novel insights into the psychological outcomes of everyday health and well-being technologies, focusing on new technology-related basic psychological needs. Our approach highlights that health technologies do not only passively transfer health content or collect data, but they also have a significant role in shaping how users experience the overall effects. The results showed that health technology use was associated with higher frustration of autonomy and competence, but also with higher satisfaction of relatedness. Our investigation also revealed that IGI largely transmitted these relationships. This finding advances the understanding of underlying factors that drive the psychological outcomes of health technology use and highlights the significance of technology users' social identification. It particularly shows that this identification can support health technology users' social needs but impair their sense of agency. The study provides important practical implications, highlighting

XSL•FO
RenderX

how future health technology design needs to support users' sense of personal control when adopting digital health technologies and fostering their sense of competence.

Funding

The study was funded by the Finnish Cultural Foundation (Self and Technology 2024, principal investigator [PI]: AO; decision number 00240914); the DigiSus research platform at Tampere University (Digital Wellbeing—Sustainability Amid Disruption project; PIs: AO and IS); Tampere University (internal project grant; PI: AO), the Finnish Foundation For Economic Education (personal working grant, MH); the Jenny and Antti Wihuri Foundation (grant for postdoctoral research, MC); and the Research Council of Finland (INVEST Research Flagship; decision number 345546). Funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

Data will be made available to the Finnish Social Science Data Archive after the end of the research projects. The data are available from the authors upon reasonable request.

Authors' Contributions

Conceptualization: MH, AK, IS, AO
Methodology: MH, AK, AO
Formal analysis: MH, AK, IS, TO, AO
Data curation: MC, IS, AO
Writing—original draft: MH, AK, MC, IS, TO, AO
Writing—review and editing: MH, AK, MC, IS, TO, AO
Supervision: AO
Project administration: AO
Funding acquisition: AO

Conflicts of Interest

None declared.

Multimedia Appendix 1
Measurement items.
[DOCX File , 20 KB - jmir_v28i1e83054_app1.docx]

Multimedia Appendix 2
Reliability figures, McDonald's omega (ω).
[PDF File (Adobe PDF File), 75 KB - jmir_v28i1e83054_app2.pdf]

Multimedia Appendix 3
Fit statistics for dynamic panel models.
[DOCX File , 23 KB - jmir_v28i1e83054_app3.docx]

Multimedia Appendix 4
Fixed effects showing temporal variation of the main variables between T1 and T3 (N=8806).
[PDF File (Adobe PDF File), 92 KB - jmir_v28i1e83054_app4.pdf]

Multimedia Appendix 5
Country interaction effects between health technology use and autonomy and competence frustration, and relatedness satisfaction in the random-effects within–between model (N=8806).
[DOCX File , 28 KB - jmir_v28i1e83054_app5.docx]

Multimedia Appendix 6
Dynamic panel models predicting lagged effect of health and well-being technology usage on technology-related basic psychological needs, estimated with and without control variables (N=8806). Note: Models 1 and 2 are included in the main text.
[DOCX File , 27 KB - jmir_v28i1e83054_app6.docx]

XSL•FO
RenderX

References

1. Crawford K, Lingel J, Karppi T. Our metrics, ourselves: a hundred years of self-tracking from the weight scale to the wrist wearable device. Eur J Cult Stud 2015;18(4-5):479-496. [doi: 10.1177/1367549415584857]
2. Lupton D. Quantifying the body: monitoring and measuring health in the age of mHealth technologies. Crit Public Health 2013;23(4):393-403. [doi: 10.1080/09581596.2013.794931]
3. Wieczorek M, Rossmaier LWS. Healthiness as a virtue: the healthism of mHealth and the challenges to public health. Public Health Ethics 2023;16(3):219-231 [FREE Full text] [doi: 10.1093/phe/phad019] [Medline: 38333766]
4. Bhatt P, Liu J, Gong Y, Wang J, Guo Y. Emerging artificial intelligence-empowered mHealth: scoping review. JMIR Mhealth Uhealth 2022;10(6):e35053 [FREE Full text] [doi: 10.2196/35053] [Medline: 35679107]
5. Khan ZF, Alotaibi SR. Applications of artificial intelligence and big data analytics in m-health: a healthcare system perspective. J Healthc Eng 2020;2020:8894694 [FREE Full text] [doi: 10.1155/2020/8894694] [Medline: 32952992]
6. Li A, Ding Z, Sun C, Liu Y. Recommending AI based on quantified self: investigating the mechanism of consumer acceptance of AI recommendations. Electron Mark 2024;34(1):57. [doi: 10.1007/s12525-024-00739-7]
7. Ma Z, Mei Y, Su Z. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. AMIA Annu Symp Proc 2023;2023:1105-1114 [FREE Full text] [Medline: 38222348]
8. Hameed K, Naha R, Hameed F. Digital transformation for sustainable health and well-being: a review and future research directions. Discov Sustain 2024;5(1):104. [doi: 10.1007/s43621-024-00273-8]
9. Lupton D. Digital Health: Critical and Cross-Disciplinary Perspectives. London: Routledge; 2017.
10. Constantiou I, Mukkamala A, Sjöklint M, Trier M. Engaging with self-tracking applications: how do users respond to their performance data? Eur J Inf Syst 2022;32(6):941-961. [doi: 10.1080/0960085x.2022.2081096]
11. Lupton D. The Quantified Self: A Sociology of Self-Tracking. Cambridge: Polity Press; 2016.
12. Peters D, Calvo RA, Ryan RM. Designing for motivation, engagement and wellbeing in digital experience. Front Psychol 2018;9:797 [FREE Full text] [doi: 10.3389/fpsyg.2018.00797] [Medline: 29892246]
13. Vansteenkiste M, Soenens B, Ryan RM. Basic psychological needs theory: a conceptual and empirical review of key criteria. In: The Oxford Handbook of Self-Determination Theory. Oxford, UK: Oxford University Press; 2023:84-123.
14. Jain A, Pradhan D, Kuanr A, Singh S. Self‐quantification and consumer well‐being: a meta‐analytic review. Psychol Market 2024;42(2):514-538. [doi: 10.1002/mar.22141]
15. Zhang X, Sun S, Jiangenuer L, Zhao P, Lei H, Xu Z, et al. Effect of mobile health (mHealth) on improving anxiety, depression and quality of life in cancer patients: a systematic review and meta-analysis. J Affect Disord 2025;374:11-25. [doi: 10.1016/j.jad.2025.01.016] [Medline: 39793615]
16. Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry 2019;18(3):325-336 [FREE Full text] [doi: 10.1002/wps.20673] [Medline: 31496095]
17. Rathbone AL, Prescott J. The use of mobile apps and SMS messaging as physical and mental health interventions: systematic review. J Med Internet Res 2017;19(8):e295 [FREE Full text] [doi: 10.2196/jmir.7740] [Medline: 28838887]
18. Chen H, Schoefer K, Manika D, Tzemou E. The “dark side” of general health and fitness-related self-service technologies: a systematic review of the literature and directions for future research. J Public Policy Mark 2024;43(2):151-170. [doi: 10.1177/07439156231224731]
19. Fronczek LP, Mende M, Scott ML, Nenkov GY, Gustafsson A. Friend or foe? Can anthropomorphizing self-tracking devices backfire on marketers and consumers? J Acad Mark Sci 2022;51(5):1075-1097. [doi: 10.1007/s11747-022-00915-1]
20. Honary M, Bell BT, Clinch S, Wild SE, McNaney R. Understanding the role of healthy eating and fitness mobile apps in the formation of maladaptive eating and exercise behaviors in young people. JMIR Mhealth Uhealth 2019;7(6):e14239 [FREE Full text] [doi: 10.2196/14239] [Medline: 31215514]
21. Karapanos E, Gouveia R, Hassenzahl M, Forlizzi J. Wellbeing in the making: peoples' experiences with wearable activity trackers. Psychol Well Being 2016;6:4 [FREE Full text] [doi: 10.1186/s13612-016-0042-6] [Medline: 27376017]
22. Russell HC, Potts C, Nelson E. “If it's not on Strava it didn’t happen”: perceived psychosocial implications of Strava use in collegiate club runners. Recreat Sports J 2022;47(1):15-25. [doi: 10.1177/15588661221148170]
23. Stollfuß S. Communitainment on Instagram: fitness content and community-driven communication as social media entertainment. Sage Open 2020;10(2):215824402091953. [doi: 10.1177/2158244020919535]
24. Arigo D, Brown MM, Pasko K, Suls J. Social comparison features in physical activity promotion apps: scoping meta-review. J Med Internet Res 2020;22(3):e15642 [FREE Full text] [doi: 10.2196/15642] [Medline: 32217499]
25. Hamari J, Hassan L, Dias A. Gamification, quantified-self or social networking? Matching users’ goals with motivational technology. User Model User-Adap Inter 2018;28(1):35-74. [doi: 10.1007/s11257-018-9200-2]
26. Whelan E, Clohessy T. How the social dimension of fitness apps can enhance and undermine wellbeing. Inf Technol People 2020;34(1):68-92. [doi: 10.1108/itp-04-2019-0156]
27. Deci EL, Ryan RM. The "What" and "Why" of goal pursuits: human needs and the self-determination of behavior. Psychol Inq 2000;11(4):227-268. [doi: 10.1207/s15327965pli1104_01]
28. Ryan RM, Deci EL. Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness. New York: Guilford Press; 2017:1462538967.

XSL•FO
RenderX

29. Ryan RM, Huta V, Deci EL. Living well: a self-determination theory perspective on eudaimonia. J Happiness Stud 2006;9(1):139-170. [doi: 10.1007/s10902-006-9023-4]
30. Vansteenkiste M, Ryan RM. On psychological growth and vulnerability: basic psychological need satisfaction and need frustration as a unifying principle. J Psychother Integr 2013;23(3):263-280. [doi: 10.1037/a0032359]
31. Burnell R, Peters D, Ryan RM, Calvo RA. Technology evaluations are associated with psychological need satisfaction across different spheres of experience: an application of the METUX scales. Front Psychol 2023;14:1092288 [FREE Full text] [doi: 10.3389/fpsyg.2023.1092288] [Medline: 37275684]
32. Wannheden C, Stenfors T, Stenling A, von Thiele Schwarz U. Satisfied or frustrated? A qualitative analysis of need satisfying and need frustrating experiences of engaging with digital health technology in chronic care. Front Public Health 2020;8:623773 [FREE Full text] [doi: 10.3389/fpubh.2020.623773] [Medline: 33585391]
33. Tikkanen H, Heinonen K, Ravald A. Smart wearable technologies as resources for consumer agency in well-being. J Interact Mark 2023;58(2-3):136-150. [doi: 10.1177/10949968221143351]
34. Lin Y, Tudor-Sfetea C, Siddiqui S, Sherwani Y, Ahmed M, Eisingerich AB. Effective behavioral changes through a digital mHealth app: exploring the impact of hedonic well-being, psychological empowerment and inspiration. JMIR Mhealth Uhealth 2018;6(6):e10024 [FREE Full text] [doi: 10.2196/10024] [Medline: 29907557]
35. Stragier J, Vanden Abeele M, Mechant P, De Marez L. Understanding persistence in the use of online fitness communities: comparing novice and experienced users. Comput Hum Behav 2016;64:34-42. [doi: 10.1016/j.chb.2016.06.013]
36. Carter M, Grover V. Me, my self, and I(T): conceptualizing information technology identity and its implications. MIS Q 2015;39(4):931-957. [doi: 10.25300/misq/2015/39.4.9]
37. Carter M, Petter S, Grover V, Thatcher JB. Information technology identity: a key determinant of IT feature and exploratory usage. MIS Q 2020;44(3):983-1021. [doi: 10.25300/misq/2020/14607]
38. Mosafer H, Sarabadani J. Identity in the digital age: a review of information technology identity (ITID) research in information systems. Inf Syst J 2021. [doi: 10.24251/hicss.2021.321]
39. Tajfel H. Experiments in intergroup discrimination. Sci Am 1970;223(5):96-102. [doi: 10.1038/scientificamerican1170-96]
40. Tajfel H, Turner JC. An integrative theory of intergroup conflict. In: Austin WG, Worchel S, editors. The Social Psychology of Intergroup Relations. Monterey, CA: Brooks/Cole; 1979:33-37.
41. Haslam C, Jetten J, Cruwys T, Dingle GA, Haslam SA. The New Psychology of Health: Unlocking the Social Cure. London: Routledge; 2018.
42. Jetten J, Haslam SA, Cruwys T, Greenaway KH, Haslam C, Steffens NK. Advancing the social identity approach to health and well‐being: progressing the social cure research agenda. Eur J Social Psych 2017;47(7):789-802. [doi: 10.1002/ejsp.2333]
43. Steffens NK, LaRue CJ, Haslam C, Walter ZC, Cruwys T, Munt KA, et al. Social identification-building interventions to improve health: a systematic review and meta-analysis. Health Psychol Rev 2021;15(1):85-112. [doi: 10.1080/17437199.2019.1669481] [Medline: 31530154]
44. Kachanoff FJ, Wohl MJA, Koestner R, Taylor DM. Them, Us, and I: how group contexts influence basic psychological needs. Curr Dir Psychol Sci 2019;29(1):47-54. [doi: 10.1177/0963721419884318]
45. Greenaway KH, Cruwys T, Haslam SA, Jetten J. Social identities promote well‐being because they satisfy global psychological needs. Eur J Social Psychol 2015;46(3):294-307. [doi: 10.1002/ejsp.2169]
46. Heiskari M, Celuch M, Koivula A, Savolainen I, Oksanen A. Technological identity and basic psychological needs in the use of new technologies: a two-wave cross-national survey study. Int J Inf Manag 2025;84:102926. [doi: 10.1016/j.ijinfomgt.2025.102926]
47. Haslam SA, Jetten J, Postmes T, Haslam C. Social identity, health and well‐being: an emerging agenda for applied psychology. Appl Psychol 2008;58(1):1-23. [doi: 10.1111/j.1464-0597.2008.00379.x]
48. Stets JE, Burke PJ. Identity theory and social identity theory. Soc Psychol Q 2000;63(3):224 [FREE Full text] [doi: 10.2307/2695870]
49. Turner JC, Reynolds KJ. The story of social identity. In: Rediscovering Social Identity. New York, NY: Psychology Press; 2010:13-32.
50. Greenaway KH, Haslam SA, Cruwys T, Branscombe NR, Ysseldyk R, Heldreth C. From “we” to “me”: group identification enhances perceived personal control with consequences for health and well-being. J Pers Soc Psychol 2015;109(1):53-74. [doi: 10.31234/osf.io/agtqm]
51. A self-determination perspective on digital health and well-being technologies. OSF. URL: https://osf.io/pz53b/
52. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: the APA publications and communications board task force report. Am Psychol 2018;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
53. Eurostat. Digitalisation in Europe - 2024 edition. European Commission. 2024. URL: https://ec.europa.eu/eurostat/web/interactive-publications/digitalisation-2024 [accessed 2025-06-15]
54. Chen B, Vansteenkiste M, Beyers W, Boone L, Deci EL, Van der Kaap-Deeder J, et al. Basic psychological need satisfaction, need frustration, and need strength across four cultures. Motiv Emot 2014;39(2):216-236. [doi: 10.1007/s11031-014-9450-1]

XSL•FO
RenderX

55. Leach CW, van Zomeren M, Zebel S, Vliek MLW, Pennekamp SF, Doosje B, et al. Group-level self-definition and self-investment: a hierarchical (multicomponent) model of in-group identification. J Pers Soc Psychol 2008;95(1):144-165 [FREE Full text] [doi: 10.1037/0022-3514.95.1.144] [Medline: 18605857]
56. Ellemers N, Kortekaas P, Ouwerkerk JW. Self-categorisation, commitment to the group and group self-esteem as related but distinct aspects of social identity. Eur J Soc Psychol 1999;29(2-3):371-389. [doi: 10.1002/(sici)1099-0992(199903/05)29:2/3<371::aid-ejsp932>3.0.co;2-u]
57. Jylhä M, Guralnik JM, Ferrucci L, Jokela J, Heikkinen E. Is self-rated health comparable across cultures and genders? J Gerontol B Psychol Sci Soc Sci 1998;53(3):S144-S152. [doi: 10.1093/geronb/53b.3.s144] [Medline: 9602839]
58. Robins RW, Hendin HM, Trzesniewski KH. Measuring global self-esteem: construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. Pers Soc Psychol Bull 2001;27(2):151-161. [doi: 10.1177/0146167201272002]
59. Schneider SM, Schupp J. Individual differences in social comparison and its consequences for life satisfaction: introducing a short scale of the Iowa–Netherlands Comparison Orientation Measure. Soc Indic Res 2013;115(2):767-789. [doi: 10.1007/s11205-012-0227-1]
60. Gibbons FX, Buunk BP. Individual differences in social comparison: development of a scale of social comparison orientation. J Pers Soc Psychol 1999;76(1):129-142. [doi: 10.1037//0022-3514.76.1.129] [Medline: 9972558]
61. Abdel-Khalek AM. Measuring happiness with a single-item scale. Soc Behav Pers 2006;34(2):139-150. [doi: 10.2224/sbp.2006.34.2.139]
62. Allison PD, Williams R, Moral-Benito E. Maximum likelihood for cross-lagged panel models with fixed effects. Socius 2017;3:2378023117710578. [doi: 10.1177/2378023117710578]
63. Moral-Benito E, Allison P, Williams R. Dynamic panel data modelling using maximum likelihood: an alternative to Arellano-Bond. Appl Econ 2018;51(20):2221-2232. [doi: 10.1080/00036846.2018.1540854]
64. Williams R, Allison PD, Moral-Benito E. Linear dynamic panel-data estimation using maximum likelihood and structural equation modeling. Stata J 2018;18(2):293-326. [doi: 10.1177/1536867x1801800201]
65. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Modeling 1999;6(1):1-55. [doi: 10.1080/10705519909540118]
66. Chen FF. Sensitivity of goodness of fit indexes to lack of measurement invariance. Struct Equ Modeling 2007;14(3):464-504. [doi: 10.1080/10705510701301834]
67. Martela F. Being as having, loving, and doing: a theory of human well-being. Pers Soc Psychol Rev 2024;28(4):372-397 [FREE Full text] [doi: 10.1177/10888683241263634] [Medline: 39056545]
68. Bao H, Lee EWJ. Examining theoretical frameworks and antecedents of health apps and wearables use: a scoping review. Health Commun 2024;39(12):2671-2681. [doi: 10.1080/10410236.2023.2283655] [Medline: 37968803]
69. Feng S, Mäntymäki M, Dhir A, Salmela H. How self-tracking and the quantified self promote health and well-being: systematic review. J Med Internet Res 2021;23(9):e25171 [FREE Full text] [doi: 10.2196/25171] [Medline: 34546176]
70. Turner JC, Oakes PJ. The socially structured mind. In: McGarty C, Haslam SA, editors. The Message of Social Psychology: Perspectives on Mind in Society. Oxford, Uk: Blackwell; 1997:355-373.
71. Esmaeilzadeh P. The influence of gamification and information technology identity on postadoption behaviors of health and fitness app users: empirical study in the United States. JMIR Serious Games 2021;9(3):e28282 [FREE Full text] [doi: 10.2196/28282] [Medline: 34812736]
72. Ogbanufe O, Gerhart N. The mediating influence of smartwatch identity on deep use and innovative individual performance. Inf Syst J 2020;30(6):977-1009. [doi: 10.1111/isj.12288]
73. Gong X, Cheung CMK, Zhang KZK, Chen C, Lee MKO. A dual-identity perspective of obsessive online social gaming. J Assoc Inf Syst 2021;22(5):1245-1284. [doi: 10.17705/1jais.00693]
74. Tong HL, Laranjo L. The use of social features in mobile health interventions to promote physical activity: a systematic review. NPJ Digit Med 2018;1:43 [FREE Full text] [doi: 10.1038/s41746-018-0051-3] [Medline: 31304323]
75. Cheng VWS, Davenport T, Johnson D, Vella K, Hickie IB. Gamification in apps and technologies for improving mental health and well-being: systematic review. JMIR Ment Health 2019;6(6):e13717 [FREE Full text] [doi: 10.2196/13717] [Medline: 31244479]

Abbreviations

CFI: comparative fit index
EU: European Union
IGI: in-group identification with new technology users
RMSEA: root-mean-square error of approximation
SDT: self-determination theory
SEM: structural equation modeling
SRMR: root-mean-square residual

XSL•FO
RenderX

Edited by S Brini; submitted 27.Aug.2025; peer-reviewed by A Eisingerich, Y Kurebayashi, LT De Beer; comments to author 09.Feb.2026; accepted 10.Apr.2026; published 19.May.2026.

Please cite as:
Heiskari M, Koivula A, Celuch M, Savolainen I, Osma T, Oksanen A
Use of Health and Well-Being Technology, Basic Psychological Needs, and the Mediating Role of Technological Identity in 6 European Countries: Prospective Longitudinal Survey Study
J Med Internet Res 2026;28:e83054
URL: https://www.jmir.org/2026/1/e83054
doi:10.2196/83054
PMID:

©Moona Heiskari, Aki Koivula, Magdalena Celuch, Iina Savolainen, Teijo Osma, Atte Oksanen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Daily Stress and Heart Rate Variability Among Mindfulness Meditation Practitioners: mHealth Observational Study

Jo Takezawa[1,2], MOT, MS(AI), MBA, JD; Shixian Geng[1], PhD; Masahiro Fujino[3], PhD; Mika Miyake[4], MS; Kazutoshi Sasahara[2], PhD; Koji Yatani[1], PhD; Atsushi Niida[5], PhD

[1]Interactive Intelligent Systems Laboratory, The University of Tokyo, Tokyo, Japan

[2]Department of Innovation Science, School of Environment and Society, Institute of Science Tokyo, Tokyo, Japan

[3]Human Information Science Laboratory, NTT (Nippon Telegraph and Telephone Corporation) Communication Science Laboratories, Atsugi, Kanagawa, Japan

[4]Garmin Health, Garmin Japan Ltd., Yokohama, Kanagawa, Japan

[5]Laboratory of Molecular Medicine, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan

Corresponding Author:
Atsushi Niida, PhD
Laboratory of Molecular Medicine, Human Genome Center
The Institute of Medical Science
The University of Tokyo
General Research Building, 7th Floor
4-6-1 Shirokanedai, Minato-ku
Tokyo, 108-8639
Japan
Phone: 81 3 5449 5237
Email: aniida@ims.u-tokyo.ac.jp

Abstract

Background: Mindfulness meditation has been reported to reduce stress and enhance well-being. However, its effects on heart rate variability (HRV)—a physiological marker of stress—remain underexplored.

Objective: This study aimed to examine how meditation practice is associated with subjective stress, HRV, and their interaction, using mobile health technologies.

Methods: This 3-week observational study included 90 participants—19 meditation practitioners (meditation group), 32 recreational runners as an active control group characterized by lower stress and higher HRV (running group), and 39 individuals without regular meditation or exercise habits (control group). HRV was continuously recorded using Garmin smartwatches. Subjective stress levels and activity states were assessed 3 times daily through a smartphone-based experience sampling method, yielding a total of 4557 responses (mean 50.6, SD 22.8 per participant). From the meditation group, start and end times of 632 daily meditation sessions (mean 33.3, SD 18.3 per participant) were also collected via the app. Standardized questionnaires on stress and related measures were administered at the end of the study period.

Results: The questionnaire survey confirmed that stress levels were significantly lower in both the meditation and running groups compared with controls (median Perceived Stress Scale scores: meditation 21, IQR 17-24; running 22, IQR 19-25; control 25, IQR 21-30; Kruskal-Wallis P=.02; adjusted Wilcoxon P=.05 and .04, respectively). Smartwatch-derived HRV (root-mean-square of successive differences [RMSSD]) was elevated in the running group relative to controls (median 47.0, IQR 44.0-54.2 and median 42.0, IQR 34.2-47.8, respectively; P<.001), whereas no significant difference was observed between the meditation and control groups (median 40.8, IQR 35.5-44.4 and median 42.0, IQR 34.2-47.8, respectively). Bayesian analysis of the experience sampling method data indicated that higher subjective stress levels were associated with a concurrent RMSSD reduction of –2.24 (95% credible interval –3.97 to –0.26) milliseconds. Although this pattern was consistent across groups, the steeper decline of –3.94 (95% CI –7.04 to –0.74) milliseconds was observed only in the running group, likely reflecting their elevated baseline HRV. Additionally, Bayesian modeling of 632 logged meditation sessions revealed an acute RMSSD increase of +4.68 (95% CI 2.96 to 6.38) milliseconds during meditation, with effects maintained for at least 30 minutes post practice.

Conclusions: Although HRV among meditation practitioners did not appear elevated in overall daily life, the ability to increase HRV at arbitrary timings, along with the prolonged residual effect, may correlate with stress reduction. This hypothesis requires

XSL•FO
RenderX

further exploration with adequate controls. Despite the preliminary nature of this study due to its limited sample size, our findings highlight the potential of mobile health–based methodologies to capture stress and HRV dynamics in real-world settings.

(J Med Internet Res 2026;28:e78244) doi:10.2196/78244

KEYWORDS

mindfulness meditation; stress reduction; heart rate variability; experience sampling method; smartwatch

Introduction

Mindfulness refers to a psychological state characterized by present-moment awareness and nonjudgmental acceptance of one's experiences [1]. Although originally rooted in Buddhist traditions, mindfulness meditation (hereafter referred to simply as meditation) has been redefined as a secular mental practice aimed at cultivating mindfulness, and is now widely adopted in nonreligious contexts for its potential to reduce stress and enhance well-being [2,3]. Numerous studies have demonstrated the effectiveness of meditation in reducing perceived stress through self-report measures [4-6]. However, these studies have primarily assessed stress levels in constrained laboratory settings or before and after fixed intervention periods, leaving the impact of daily meditation practice on real-world stress largely unexplored.

Psychological stress is commonly conceptualized as a transactional process in which environmental demands are appraised as taxing or exceeding one's coping resources, thereby eliciting emotional and physiological responses [7]. In this study, we focus on perceived stress (ie, subjective appraisal) as captured by self-report measures [8]. In addition to retrospective questionnaires, subjective stress levels can be assessed using the experience sampling method (ESM), which is also referred to as ecological momentary assessment [9,10]. While questionnaires are suitable for capturing general trends in perceived stress, ESM enables participants to report their thoughts, emotions, and behaviors multiple times per day in real time, within their natural environments. This approach enables the simultaneous collection of contextual information, facilitating the examination of temporal relationships between stress and other variables in everyday life. Recently, ESM has been used in meditation research to evaluate the stress reduction effects in real-world contexts [11,12]. For example, a recent study used ESM to examine the effects of a brief mindfulness-based intervention and demonstrated that, although no acute changes were observed during the intervention phase, stress decreased at postintervention and follow-up [13].

Heart rate variability (HRV) is widely used as a physiological marker of stress, reflecting autonomic nervous system balance through fluctuations in beat-to-beat intervals (BBIs). Higher stress levels are generally associated with reduced HRV [14,15]. Although HRV has traditionally been measured using electrocardiography (ECG), recent advances in mobile health (mHealth) technologies have enabled widespread use of consumer-grade wearable devices that rely on photoplethysmography (PPG) [16,17]. PPG-based measurements are more susceptible to noise than ECG, particularly under conditions involving motion. Studies comparing wearable-derived PPG measures with ECG in daily-life settings show that agreement is condition-dependent—often higher during low-motion periods, such as sleep—and that PPG-derived indices are not always interchangeable with ECG-derived HRV, underscoring the need for appropriate quality control and cautious interpretation [18,19]. Nevertheless, PPG-based wearables offer practical advantages, including portability, long battery life, and feasibility for long-term monitoring in real-world environments. Their low cost and widespread adoption further enable large-scale population studies, and recent evidence indicates that PPG-derived HRV metrics are sufficiently reliable to capture population-level patterns of autonomic function across age, sex, and circadian rhythms, supporting their use in stress-related research [20].

Laboratory-based studies have consistently demonstrated an inverse association between HRV and perceived stress, particularly under experimentally induced stress conditions. A recent laboratory investigation using PPG-derived HRV from wearable devices has replicated these findings, demonstrating reliable reductions in HRV during controlled cognitive stress tasks and strong concordance with ECG-based measures as well as validated psychological stress scales [21]. However, findings in everyday life settings have been inconsistent [22-25]. More recently, a study integrating ESM with PPG-based wearable devices for continuous HRV monitoring has begun to clarify this relationship in real-world contexts; although effect sizes are generally modest, elevated perceived stress has been shown to be significantly associated with concurrent reductions in HRV [26]. A recent ambulatory study in a clinical population also demonstrated that HRV-based stress detection can meaningfully capture real-world stress states, particularly when individual characteristics, such as age and BMI, are taken into account, supporting the practical utility of wearable HRV for stress assessment beyond laboratory settings [27].

Although transient increases in HRV during meditation have been consistently reported in laboratory settings [28,29], whether long-term meditation practice yields maintained enhancements in HRV remains uncertain, as findings from intervention studies have been mixed. Indeed, an ambulatory randomized controlled trial found no significant effects of mindfulness-based intervention on daily-life HRV indices [30], and a recent meta-analysis reported only small-to-moderate and highly heterogeneous effects on vagally mediated HRV [31]. Consistent with this uncertainty, a recent systematic review of brief mindfulness interventions concluded that evidence for long-term HRV enhancement remains limited and of low quality [32]. Moreover, recent repeated-measures data show that meditation-related HRV modulation emerges transiently during practice, is stronger among experienced practitioners, and does not manifest as elevated baseline or daily-life HRV, indicating predominantly state-dependent and experience-dependent rather

XSL•FO
RenderX

than trait-level effects [33]. Notably, most of these insights are derived from laboratory-based measurements or structured intervention designs, and there has been little to no observational research systematically examining HRV patterns in daily-life settings or during routine meditation practice among experienced meditation practitioners. In contrast, there is clear evidence from another behavioral domain—regular physical activity—that long-term, habitual engagement can be reflected in elevated everyday HRV. Physical activity is well established to confer stress-reducing benefits [34,35], and athletes consistently exhibit higher resting and daily-life HRV compared with nonathletic individuals [36,37], providing a useful reference for understanding how continuous practices may shape autonomic regulation in naturalistic settings.

In summary, despite growing evidence linking meditation to stress reduction and HRV elevation, the lack of real-world monitoring has limited our understanding of how these relationships unfold in daily life. We therefore conducted an observational study to investigate the relationship between stress and HRV in experienced meditation practitioners, integrating smartwatch-based HRV monitoring and ESM. The study was conducted on 90 participants consisting of 3 groups—19 meditation practitioners (meditation group), 32 recreational runners as an active control group characterized by lower stress and higher HRV (running group), and 39 individuals with neither a regular meditation nor exercise routine as a passive control group (control group). Over a period of 3 weeks, participants continuously wore Garmin smartwatches to monitor HRV. In parallel, smartphone-based ESM was used to assess momentary perceived stress and activity states throughout daily life. At the end of the monitoring period, participants completed a web-based questionnaire evaluating subjective stress and related factors. Furthermore, for the meditation group, the timing of each daily meditation session was logged via the smartphone app, allowing the capture of HRV dynamics during and after practice. These multimodal datasets were integrated to address the following key questions:

1. Do meditation practitioners exhibit lower levels of perceived stress in daily life, consistent with previous findings?

2. Are HRV levels in daily life elevated among meditation practitioners compared with control participants?

3. Do real-time associations of stress with activity states and HRV differ between meditation practitioners and control participants?

4. During regular meditation sessions, do meditation practitioners show increases in HRV, as reported by laboratory-based studies, and if so, how long are such increases maintained after meditation sessions?

Taken together, this study seeks to quantitatively assess how meditation practice is associated with subjective stress, HRV, and their interaction in real-world settings, leveraging mHealth technologies to extend beyond the constraints of controlled laboratory environments (Figure 1).

Figure 1. Graphical Summary of our study design. Data were obtained from 90 participants divided into 3 groups—19 meditation practitioners (meditation group), 32 recreational runners (running group), and 39 individuals with neither regular meditation nor exercise routines (control group). Over a 3-week period, participants continuously wore Garmin smartwatches to monitor heart rate variability. In parallel, a smartphone-based ESM was used to assess momentary stress and activity in daily life. At the end of the study, participants completed a web-based questionnaire on subjective stress and related psychological variables. For the meditation group, daily practice times were logged via a smartphone app to enable fine-grained analysis of heart rate variability dynamics during and after meditation. ESM: experience sampling method; HRV: heart rate variability.

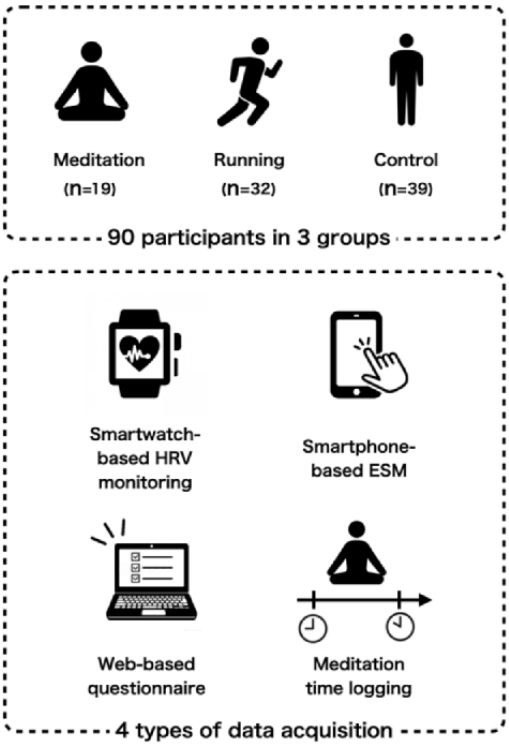

XSL•FO
RenderX

Methods

Participants and Recruitment

We recruited healthy adults aged 18 years and older who did not receive medical treatment or take regular medication. We grouped participants based on the following criteria:

1. Meditation group: Participants must have at least 1 year of experience in secular mindfulness meditation (eg, mindfulness-based stress reduction and mindfulness-based cognitive therapy) or meditation based on traditional schools of Buddhism (eg, Theravada and Zen). They must practice meditation for at least 20 minutes a day, 5 days a week, and meet at least one of the following criteria:

a. Participation in an intensive meditation retreat of at least 1 week.

b. Completion of a teacher training course in the Method of Clinical Meditation [38].

c. Completion of an 8-week mindfulness program (eg, mindfulness-based stress reduction and mindfulness-based cognitive therapy).

2. Running group: Participants must have no regular practice of meditation or yoga, no participation in meditation courses (retreats or multiweek programs), and must run for at least 30 minutes a day, 4 days a week, for more than a year.

3. Control group: Participants must have no regular practice of meditation or yoga, no participation in meditation courses, and exercise no more than 2 days a week for a maximum of 30 minutes per day.

All participants were required to own an iPhone (Apple Inc), as data collection and experience sampling were conducted using UTracker, an in-house iOS app. The app is compatible with Garmin watch models Vivosmart 4, ForeAthlete 245/245M, and ForeAthlete 745. We provided Vivosmart 4 to participants who did not own compatible Garmin watches at the time of the data collection, while those who already owned one of the supported models used their own devices.

We prepared a project website describing the recruitment criteria and study conditions, along with a link to the application page. On the application page, participants provided information regarding their health profiles and gave informed consent electronically before enrollment. Participant recruitment was conducted through multiple channels. Press releases announcing the launch of the project were issued, which also included a call for research participants together with a link to the project description page. The same recruitment information was also disseminated via social media platforms, including Facebook. In addition, recruitment announcements were distributed through a Garmin-operated mailing list targeting Garmin users, which proved particularly effective for enrolling participants in the running group. In addition to these standard recruitment approaches, participants in the meditation group were also recruited through direct contact via the authors' personal networks.

During a 2-month recruitment period from August 2 to September 30, 2022, a total of 301 individuals applied to participate in the study. To ensure a comparable health profile across the 3 groups, additional selection criteria were applied to the applicants:

- No ongoing medical treatment or regular medication use.
- Age between 25 and 65 years.
- Smoking fewer than 20 cigarettes per day.
- Consuming alcohol fewer than 4 times per week and less than 100 grams per drink.

This selection process resulted in the enrollment of 140 participants (meditation: n=23, running: n=57, and control: n=60) at the start of data collection.

Data Collection

We collected 3 types of data from 3 participant groups: physiological data from smartwatches, self-reported stress data assessed via ESM, and questionnaire data. For participants in the meditation group, we additionally recorded the start and end times of each meditation session.

We collected physiological and ESM data using our UTracker app, which leverages the Garmin Health SDK to capture BBIs and step counts from Garmin smartwatches, transmitting data to a server. Participants used this app to log data for any 3 weeks between October 12, 2022, and December 21, 2022. The app invoked ESM notifications randomly 3 times within their awake time, prompting participants to report their daily activity states (eating, work, housework, leisure, travel, rest, or other) and subjective stress on 4 levels (no, low, middle, or high stress). Responses submitted within 15 minutes were considered valid. Meditation group participants also recorded meditation session timings via an in-app feature. After the 3-week period, participants completed web-based questionnaires, which were administered via LimeSurvey (LimeSurvey GmbH).

Of the recruited participants, 90 individuals (meditation: 19, running: 32, and control: 39) provided valid smartwatch and ESM data and completed the questionnaires. We therefore used the data from these 90 participants in the subsequent data analysis. The recruitment process, device usage, and demographic characteristics, including group-wise distributions of sex, age, and BMI, are summarized in the supplementary visualization (Multimedia Appendix 1).

Questionnaire Measures

Physical activity was assessed using the International Physical Activity Questionnaire (IPAQ) short form (7 items) [39]. In accordance with the IPAQ scoring protocol, responses were converted to total metabolic equivalent task minutes per week, with higher values indicating greater physical activity. Sleep quality was assessed using the Pittsburgh Sleep Quality Index (PSQI; 19 items) [40], yielding a global score ranging from 0 to 21, where higher scores indicate poorer sleep quality. Dispositional mindfulness was assessed using the Mindful Attention Awareness Scale (MAAS) [41]. The MAAS consists of 15 items rated on a 6-point Likert scale ranging from 1 (almost always) to 6 (almost never). To align with the study's analysis strategy, scores were calculated as the total sum of the

XSL•FO
RenderX

15 items (possible range: 15-90), with higher scores indicating higher levels of dispositional mindfulness. Perceived stress was assessed using the 10-item Perceived Stress Scale (PSS) [42], with total scores ranging from 0 to 40, where higher scores indicate greater perceived stress. Well-being was assessed using the World Health Organization Well-Being Index (WHO-5; 5 items) [43], with total scores ranging from 0 to 25, where higher scores indicate greater well-being. All questionnaires were administered in Japanese using validated Japanese versions where available [39-43].

Calculation of HRV Indexes

From smartwatch data, we derived time-domain indices of HRV as objective markers of stress levels. These HRV indices were used to compare HRV among the 3 groups and to evaluate the association of HRV with ESM-derived stress levels and meditation (hereafter, referred to as the comparative HRV analysis, the ESM-HRV analysis, and the meditation-HRV analysis, respectively).

To calculate the HRV indices, we used the BBI time series provided via the Garmin Health SDK, which is derived from wrist PPG. The "Enhanced BBI" includes a per-beat binary confidence indicator; beats can be flagged as low confidence mainly due to motion artifacts or poor signal quality, and no beat is provided during periods of very high motion [44]. To ensure data reliability, we applied additional rule-based quality filters. We segmented heart rate (HR) interval data into nonoverlapping 5-minute windows starting at 00:00, and retained only those with a beat count between 150 and 1000 and a standard deviation of normal-to-normal interval (SDNN) between 10 milliseconds and 200 milliseconds. We also excluded windows for which the UTracker app returned no step-count values (missing activity flag), as these could not be classified into physical activity states.

Because continuous monitoring occasionally exceeded the prespecified observation period, data length was harmonized by selecting, for each participant, a contiguous 21-day window (3 weeks) that maximized the number of available 5-minute windows. In the comparative HRV analysis, the nonoverlapping 5-minute windows starting at 00:00 were used. For ESM-HRV and meditation-HRV analyses, we prepared different 5-minute windows contained by the 21-day measurement period: the ESM-HRV analysis used 5-minute windows centered on ESM reporting times, while the meditation-HRV analysis used 5-minute windows aligned with the reported start or end times of meditation sessions.

Across all participants, a total of 409,608 5-minute windows containing "Enhanced BBI" data were obtained via the Garmin Health SDK (mean 4551, SD 1927 per participant). Among these, 5242 windows (mean 58, SD 47 per participant) were excluded according to the 3 predefined exclusion criteria: beat count, SDNN, and assignment of a physical activity state. The remaining 322,220 windows (mean 3580, SD 1148 per participant) that fell within the selected 21-day measurement periods were used for the subsequent analyses (Multimedia Appendix 2).

Throughout all the 3 analyses, we primarily focused on the root-mean-square of successive differences (RMSSD), which is known to strongly correlate with high-frequency (HF) HRV and to be sensitive to changes in parasympathetic activity [14]. RMSSD was computed within each 5-minute window by calculating successive differences between adjacent intervals, squaring these differences, averaging the square values, and taking the square root of the mean.

For the comparative HRV analysis, we evaluated the distribution of mean RMSSD across the 3 groups. Mean RMSSD was calculated by averaging RMSSD values from all valid windows during the measurement period for each participant. To ensure the robustness of our findings, additional short-term HRV metrics, including SDNN, percentage of adjacent NN intervals greater than 50 milliseconds, and mean of normal-to-normal intervals (meanNN), were similarly assessed. Long-term HRV was also evaluated using the SDNN, calculated by averaging BBIs within 5-minute windows over a 24-hour period and determining the SD of these means. SDNN provided a stable measure of long-term autonomic activity, complementing short-term metrics by mitigating the influence of transient fluctuations and artifacts [45].

To further investigate HRV differences while accounting for variations in physical activity levels across the groups, we stratified RMSSD by physical activity states. Using smartwatch-derived step counts and self-reported bedtime and wake-up time collected via the UTracker app, we categorized 5-minute windows into 4 states: run (≥400 steps), walk (50-399 steps), rest (10-49 steps or <10 steps outside sleep), and sleep (<10 steps during sleep). Windows lacking step information could not be assigned a physical activity state and were therefore excluded during the data-cleaning process. From windows assigned to each state, we calculated the mean RMSSD for each participant and compared the distributions among the 3 groups.

In addition to the HRV indices, we calculated HR for each physical activity state (sleep, rest, walk, and run) from the meanNN (in milliseconds) by converting it to beats per minute (HR=60,000/meanNN). For each participant, state-specific HR values were summarized as the median across all eligible 5-minute windows within each state over the 21-day measurement period. Resting HR was operationalized as the median HR during the rest state.

In the meditation-HRV analysis, we examined alterations in autonomic nervous system activity using frequency-domain metrics. Low-frequency (LF; 0.04-0.15 Hz) and HF (0.15-0.40 Hz) power were calculated using Welch periodogram after cubic spline resampling at 4 Hz for each 5-minute window. However, frequency-domain metrics derived from wrist PPG are subject to greater measurement uncertainty than those obtained from ECG [46,47]. Therefore, these metrics are reported as supplementary descriptive indices rather than primary outcomes.

Statistical Analysis

Normality was first assessed using the Shapiro-Wilk test. Because several variables deviated from normality and included outliers despite preprocessing, nonparametric tests were used to compare questionnaire scores and HRV indices among the

XSL•FO
RenderX

3 groups. To ensure comparability across outcomes, the same nonparametric approach was applied to all outcomes. Shapiro-Wilk *P* values, as well as the numbers of missing observations and IQR-based outliers, are provided in Multimedia Appendix 3. For group comparisons, Kruskal-Wallis tests were used for overall differences, followed by pairwise Wilcoxon rank-sum tests with multiple-comparison adjustment using the Shaffer procedure [48]. Effect sizes were calculated as epsilon-square (ε^2) for Kruskal-Wallis tests, and rank-biserial correlation (*r*) together with the common-language effect size for pairwise Wilcoxon tests. As a sensitivity analysis, we additionally conducted parametric tests (Welch ANOVA followed by Games-Howell post hoc tests), which yielded results consistent with those from the nonparametric analyses.

To account for the effect of age on HRV, we also conducted an analysis of covariance (ANCOVA) with age included as a covariate (age entered as a continuous variable; group as the main factor). Effect sizes were quantified as partial eta-square (η^2), and a group-by-age interaction term was examined to assess the assumption of homogeneous regression slopes. In addition, given the unequal group sizes and the potential presence of heteroskedasticity, we performed a sensitivity analysis using ANCOVA with HC3 heteroskedasticity-consistent SEs.

For analyses of ESM-reported stress levels, stress ratings were dichotomized (no-low vs middle-high). Furthermore, 2-sided binomial tests were conducted at the pooled and group-specific levels for each activity state to examine whether the observed proportion of middle-high stress differed from the expected proportion, using the corresponding pooled or group-specific proportion as the binomial parameter. Separately, group-level comparisons of the meditation and running groups against the control group were performed using binomial tests with the control group proportion as the reference parameter. *P* values from all binomial tests were adjusted for multiple comparisons using the Holm procedure. Effect sizes were quantified using Cohen *h*.

The *P* values and effect sizes for all statistical tests are provided in Multimedia Appendices 3-11. All statistical analyses were performed using Python (version 3.12) with the *SciPy* and *statsmodels* libraries.

Hierarchical Bayesian Analysis

In the ESM-HRV analysis, we examined differences in HRV across the 2 stress states, no-low and middle-high stress, by using a hierarchical Bayesian model. The model was specifically designed to isolate HRV variations associated with stress states while accounting for interindividual differences in baseline HRV (Multimedia Appendix 12). In the model, *m*, *i*, and *j* index stress states, participants, and observations, respectively. As the observed variable, the model has m, i, and e, which represent the stress state, participant, and RMSSD value in the *j*-th observation. The model assumes that e is sampled from a normal distribution with a mean of β defined as the sum of the mean RMSSD for the observed state (α) and the residual associated with the observed participant (r). By focusing on α as a parameter corresponding to the mean RMSSD for state *m*, we estimated the posterior distributions of α and the differences in α by running Markov Chain Monte Carlo (MCMC).

Additionally, we extended the analysis to examine HRV differences between the 2 stress states within each of the 3 participant groups. For this purpose, the model was adapted such that m indexes 6 states, defined by the combination of the 2 stress levels and the 3 groups. In the meditation-HRV analysis, we examined HRV changes across the following meditation phases: the 10-minute before meditation phase, the during meditation phase, and the after meditation phases with 10-minute intervals up to 60 minutes post meditation. Similarly, the model was adjusted to allow *m* to represent the respective meditation phases.

In the meditation-HRV analysis, we further investigated time-series HRV profiles across meditation sessions of varying durations. For this purpose, we extended the model to assume temporal dependencies among parameters corresponding to mean RMSSD (Multimedia Appendix 13). In the model, *t*, *k*, *i*, and *j* index time points, meditation duration, participants, and observations, respectively. As the observed variable, the model has t, k, i, and e, which represent the time points, meditation duration, participant, and RMSSD value in the *j*-th observation. The parameter corresponding to the mean RMSSD for the time point *t* and the meditation length *k*, α, is sampled from a normal distribution with a mean of the parameter of the previous time point, α_1, though the parameter of the initial time point, α_1, depends on a hyper-parameter α. Taking the start or end of meditation as time 0, RMSSD was computed using start or end time-aligned windows spanning from –20 to +55 minutes. After grouping meditation sessions by duration in 10-minute intervals, MCMC sampling was conducted to estimate posterior medians and 95% credible intervals of α, generating time-series HRV profiles for each duration group.

In all the analyses, MCMC sampling was performed on RStan (version 2.32.6; Stan Development Team), with parameter settings including 4 chains, 1000 burn-in iterations, and 2000 total iterations. MCMC convergence was confirmed using the Gelman-Rubin diagnostic.

Ethical Considerations

This study was approved by the Ethics Review Board of the Institute of Medical Science, The University of Tokyo (approval 2022-11-0629) and the Human Subject Research Ethics Review Committee of Tokyo Institute of Technology (approval 2022154). Participants provided informed consent by submitting their application after reading the study explanation on the recruitment website. Participation was voluntary, with withdrawal permitted at any time without justification.

As compensation, participants who were provided with Garmin smartwatches retained the devices after the data collection period. Participants who used their own compatible devices received a Garmin electronic coupon.

XSL•FO
RenderX

Results

Questionnaire Data Analysis

We first compare subjective stress levels and related factors across the 3 groups, based on scores from the 5 questionnaires (Figure 2; refer to Multimedia Appendices 4 and 5 for nonparametric analyses). The IPAQ results confirmed that the running group demonstrated significantly higher physical activity than the control group (median 2625, IQR 1885-3871 and median 990, IQR 424-2075, respectively; Kruskal-Wallis test $P<.001$; Wilcoxon test with Shaffer correction $P<.001$), with the meditation group scoring between the two (median 1548, IQR 1089-3150). In the MAAS questionnaire, the running group had significantly higher mindfulness scores than the control group (median 70.5, IQR 64-81 and median 60, IQR 55-68, respectively; Kruskal-Wallis $P=.006$; Wilcoxon $P=.007$). The meditation group also showed higher scores, although the difference did not reach the significance cutoff (median 68, IQR 61-76; Wilcoxon $P=.10$).

The PSS results indicate that both the meditation and running groups experienced significantly lower subjective stress compared to the control group (meditation: median 21, IQR 17-24; running: median 22, IQR 19-25; and control: median 25, IQR 21-30; Kruskal-Wallis $P=.02$; Wilcoxon $P=.05$ and .04, respectively). In the WHO-5 questionnaire, a significant difference existed among the 3 groups (mediation: median 18, running: median 16, and control: median 14; Kruskal-Wallis $P=.04$), and both the meditation and running groups showed higher well-being scores than the control group with marginal significance (Wilcoxon $P=.07$ and .08, respectively). We observed no significant patterns in the PSQI questionnaire (mediation: median 3, IQR 3-5; running: median 4, IQR 3-6; control: median 5, IQR 3-6; Kruskal-Wallis $P=.18$).

These findings suggest that both meditation and physical exercise contributed to the reduction of stress, together with enhanced mindfulness and well-being.

Figure 2. Comparative analysis of the 5 questionnaire scores. The horizontal line within each violin plot denotes the median, while the vertical line or shaded region represents the IQR, encompassing values from the 25th to 75th percentiles. Points represent scores from each participant in the control, meditation, and runner groups. (A) High International Physical Activity Questionnaire scores indicate high physical activity levels, (B) high Mindful Attention Awareness Scale scores indicate high mindfulness levels, (C) low Perceived Stress Scale scores indicate low perceived stress levels, (D) high World Health Organization Well-Being Index scores indicate high psychological well-being levels, and (E) low PSQI scores indicate better sleep quality. Significant differences were assessed using Kruskal-Wallis tests followed by Wilcoxon test. IPAQ: International Physical Activity Questionnaire; K-W: Kruskal-Wallis; MAAS: Mindful Attention Awareness Scale; n.s.: not significant ($P>.05$); PSQI: Pittsburgh Sleep Quality Index; PSS: Perceived Stress Scale; WHO-5: World Health Organization Well-being Index. *$P<.05$, **$P<.01$, ***$P<.001$ in the Wilcoxon test.

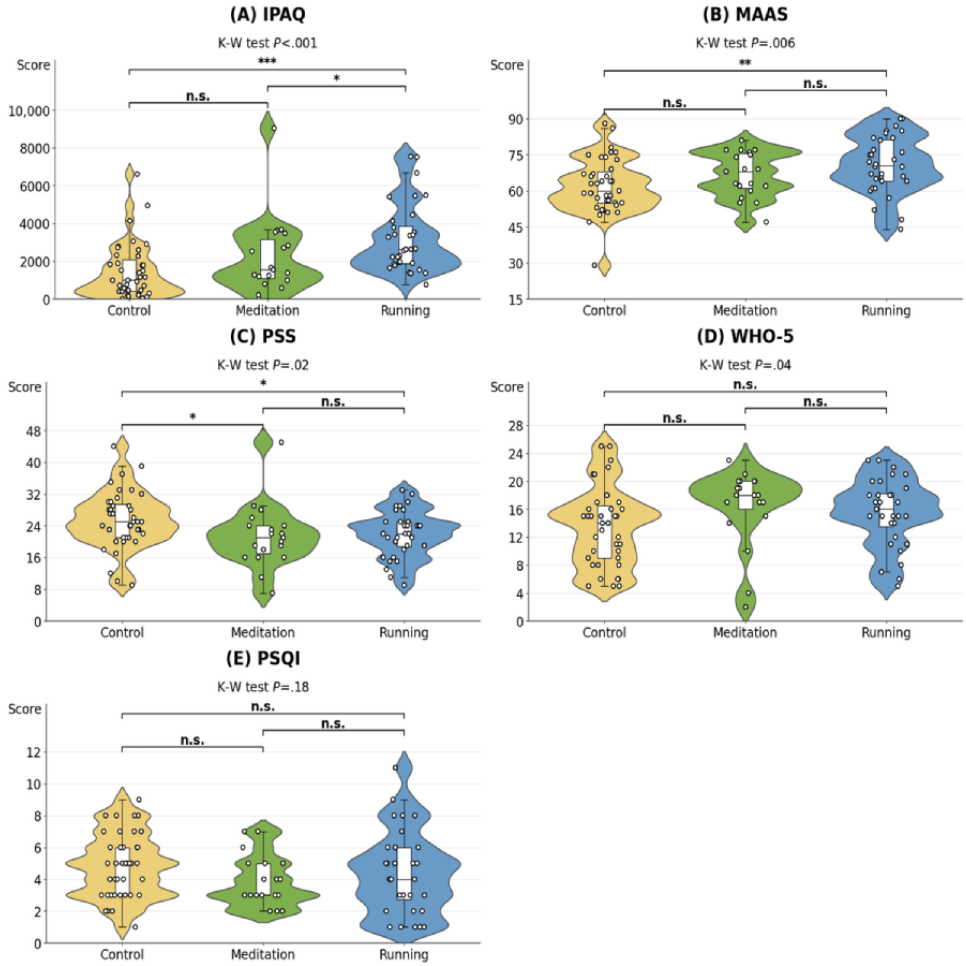

XSL•FO
RenderX

HRV Data Analysis

We next compared smartwatch-derived HRV among the 3 groups (Figure 3; refer to Multimedia Appendices 4 and 6 for nonparametric analyses, and Multimedia Appendices 9-11 for parametric analyses). Comparison of median RMSSD showed that the running group exhibited significantly higher RMSSD than the control group (Figure 3A; running: median 47.0, IQR 44.0-54.2, control: median 42.0, IQR 34.2-47.8, respectively; Kruskal-Wallis $P<.001$; Wilcoxon $P<.001$), In contrast, the meditation group exhibited RMSSD levels comparable with those of the control group (mediation: median 40.8, IQR 35.5-44.4, and control: median 42.0, IQR 34.2-47.8). We also found a consistent pattern for other short-term HRV metrics, and SDNN, an HRV metric calculated over 24-hour periods (Figure 3B).

Since HRV is affected by the intensity of physical activity, it is possible that these results merely reflect the fact that the running group engaged in different physical activity intensities. To correct for such effects, we categorized 5-minute windows into 4 activity states—sleep, rest, walk, and run—using smartwatch step counts and self-reported sleep data. We calculated mean RMSSD for each activity state and compared group-level distributions (Figure 3C). The running group consistently displayed higher RMSSD during the rest state (running: median 49.1, IQR 43.7-54.5, control: median 42.0, IQR 34.1-47.2, meditation: median 41.2, IQR 36.8-43.9; Kruskal-Wallis $P<.001$; Wilcoxon $P<.001$) and the walk state (running: median 43.5, IQR 39.7-47.4, control: median 36.5, IQR 31.1-40.6, meditation: median 34.4, IQR 32.4-38.5; Kruskal-Wallis $P<.001$; Wilcoxon $P<.001$), compared with the other groups.

It is well-established that HRV is associated with age [49]. To account for potential age-related effects when examining group differences in HRV, we conducted an ANCOVA with age included as a covariate (Multimedia Appendix 7). The results of the ANCOVA indicated that the observed group differences in HRV remained statistically significant after adjusting for age, confirming the robustness of our findings to age correction. Because RMSSD is known to be associated with HR [50], we also additionally compared HR across the 3 groups within each activity state (Multimedia Appendix 14). HR showed group differences that were consistent with those observed for RMSSD; notably, a significantly lower resting HR was observed exclusively in the running group. These findings indicate that group differences in RMSSD are accompanied by corresponding differences in HR, consistent with known HR-HRV coupling.

Taken together, our results suggest that physical exercise is associated with higher HRV, while meditation practice has no apparent effects on overall HRV patterns in daily life.

Figure 3. Comparative analysis of the heart rate variability indexes. For each participant, the median values of various heart rate variability indices were calculated from smartwatch-derived beat-to-beat intervals, and their distributions were compared across 3 groups—control, meditation, and runner. The indices include (A) 5-min root-mean-square of successive differences, (B) 24-hour standard deviation of normal-to-normal interval, and (C) root-mean-square of successive differences stratified by 4 physical activity states (sleep, rest, walk, and run), categorized using smartwatch-derived step counts and self-reported sleep times. Figure presentation and statistical tests were conducted in the same manner as in Figure 2. Significant differences were assessed using Kruskal-Wallis tests followed by Wilcoxon test. K-W: Kruskal-Wallis; n.s.: not significant ($P>.05$). *$P<.05$, **$P<.01$, ***$P<.001$ in the Wilcoxon test.

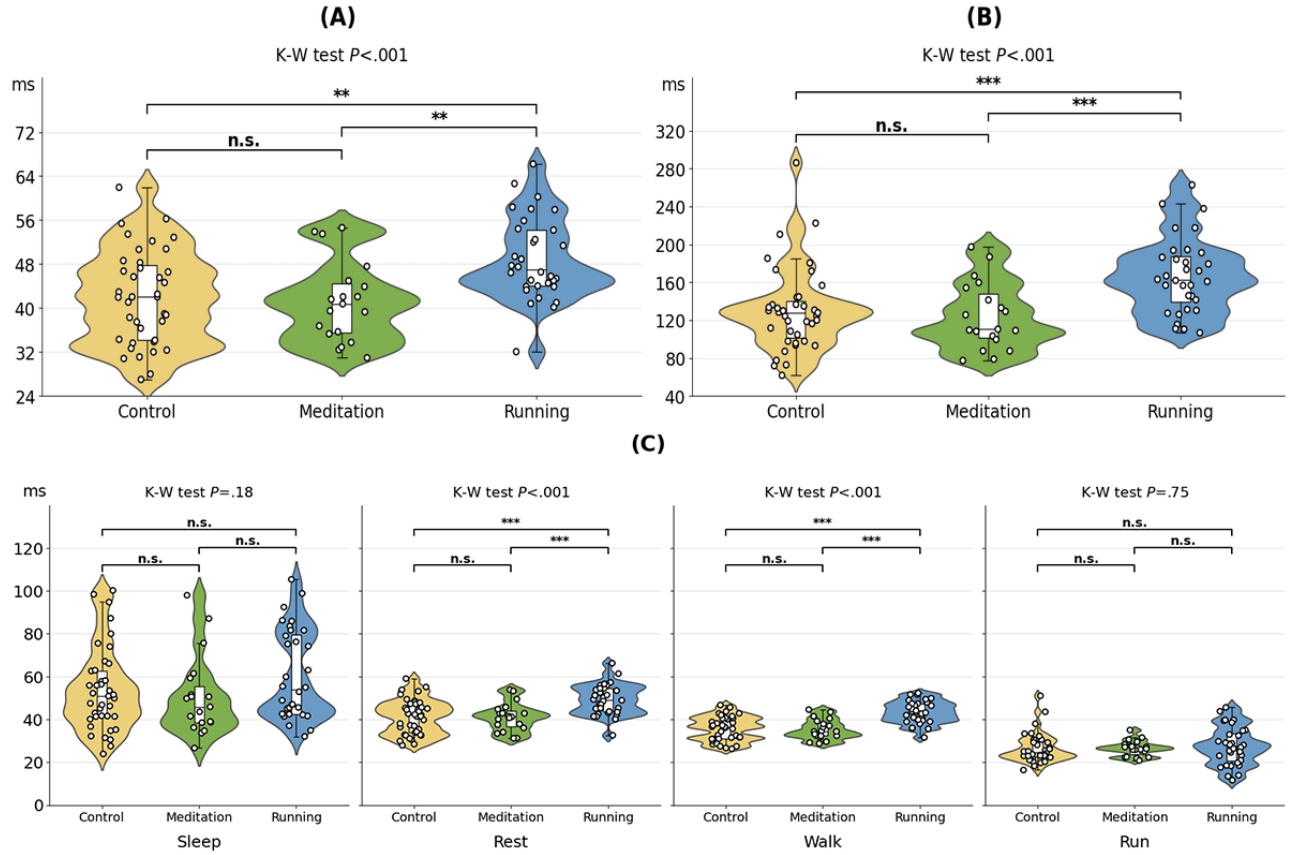

XSL•FO
RenderX

ESM Data Analysis

In this study, we also collected daily activity states and stress levels by ESM to analyze their associations (Figure 4A; Multimedia Appendices 8 and 15). ESM samplings were randomly administered 3 times per day, generating 4557 responses (51.78 per participant on average). Among all the responses, 10.3% (471/4557) reported middle-high stress. Using the overall proportion of middle-high stress responses as a baseline, stress levels were higher during housework (54/386, 14%; adjusted P=.05, based on a binomial test) and work (280/1352N, 20.7%; P<.001), and lower during leisure (20/864, 2.3%; P<.001), rest (21/734, 2.9%; P<.001), and eating (16/472, 3.4%; P<.001). This pattern was consistent across groups, although the proportion of middle-high stress responses was lower in the running group (121/1680, 7.2%; P<.001) and meditation group (76/882, 8.6%; P<.001) compared with the control group (274/1995, 13.7%), in line with the questionnaire findings. We also examined the relationship between self-reported stress levels and RMSSD recorded at reporting times (Figure 4B).

To account for interparticipant variations in baseline HRV, we used a hierarchical Bayesian approach to estimate posterior distributions of the average RMSSD for the 2 different stress states—no-low stress and middle-high stress (Multimedia Appendix 12). The analysis revealed that RMSSD was lower during middle-high stress states compared with no-low stress states, confirming the ecological validity of HRV as a stress indicator. We additionally sampled the posterior distribution of the difference between the 2 parameters, whose posterior median estimated the RMSSD difference as −2.242 (95% credible interval −3.967 to −0.260) milliseconds (Multimedia Appendix 16). This trend was observed across all groups, with the running group showing a more pronounced decrease of −3.939 (95% credible interval −7.041 to −0.737) milliseconds, likely due to higher baseline RMSSD values.

Collectively, our ESM analysis demonstrated that subjective stress levels are associated with both daily activity states and HRV.

XSL•FO
RenderX

Figure 4. Associations between stress levels, daily activities, and heart rate variability. (A) Stress level distribution by activity type. The proportion of self-reported stress levels across 7 daily activity states, derived from 4557 experience sampling method responses. The vertical axis denotes the proportions of 4-staged stress levels, while the horizontal axis represents activity types. Horizontal black lines across the bars indicate the overall middle-high stress proportion. (B) root-mean-square of successive differences (RMSSD) density plots by stress levels. Posterior density plots of averaged RMSSD under no-low stress and middle-high stress conditions. RMSSD was calculated using 5-minute windows centered on reporting times. A hierarchical Bayesian model (Multimedia Appendix 12) was used to account for variability in response frequency and baseline heart rate variability across participants. Posterior densities of α were estimated using Markov Chain Monte Carlo, with shaded regions denoting 95% credible intervals and vertical dotted lines indicating posterior medians. C: control; M: meditation; R: running.

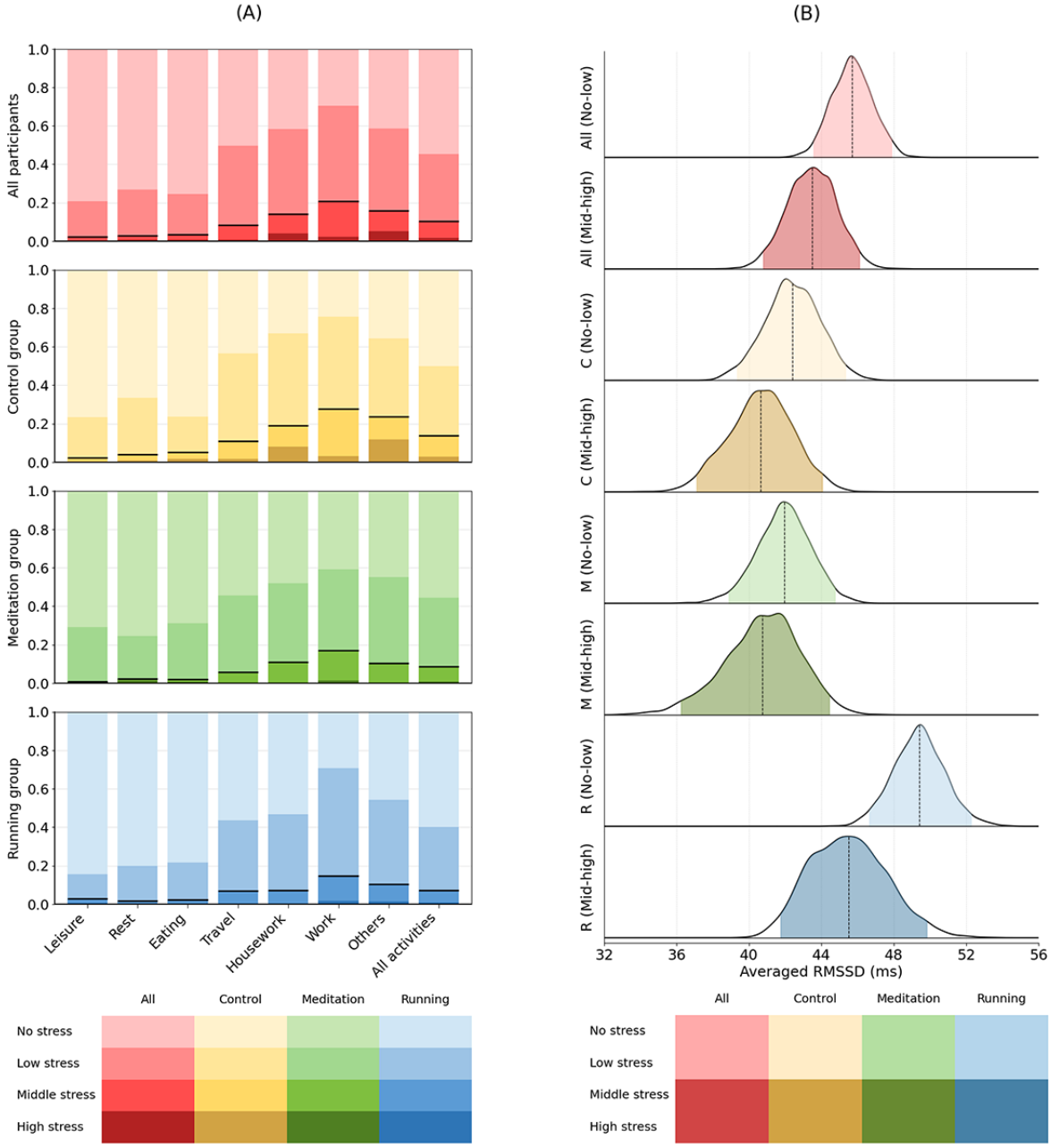

Analysis of HRV and Meditation

From the meditation group, we collected information about start and end times of 632 daily meditation sessions (33 per participant on average) using the smartphone app. We investigated HRV changes during meditation by integrating the meditation timing information with the heartbeat data obtained from smartwatches (Multimedia Appendix 17).

From each participant in the meditation group, RMSSD was computed for the 10-minute period before meditation, the period throughout meditation, and the periods in 10-minute intervals up to 60 minutes after meditation. Similarly to the ESM data analysis, we used hierarchical Bayesian models (Multimedia Appendix 12; also refer to Multimedia Appendix 13 for the time-series extension) to account for interparticipant variations in baseline HRV and calculate posterior distributions of parameters corresponding to averaged RMSSD values for each period (Figure 5A). As a result, RMSSD showed an increase during meditation, with the posterior median estimated at 4.675 (95% credible interval 2.962-6.380; Multimedia Appendix 18). Remarkably, RMSSD required more than 30 minutes to return

XSL•FO
RenderX

to baseline levels after meditation, suggesting a prolonged residual effect.

We also examined RMSSD profiles across meditation sessions of varying durations. Meditation sessions were stratified into 10-minute intervals based on their duration, and time-series RMSSD trajectories were computed using a hierarchical Bayesian approach (Figure 5B). The analysis revealed a progressive increase in RMSSD that persisted throughout the meditation period. Upon stopping meditation, RMSSD gradually decreased toward baseline levels, underscoring the persistence of residual effects.

To further characterize the physiological changes during meditation, we analyzed frequency-domain indices, evaluating both their absolute distributions and their baseline-adjusted shifts (Multimedia Appendix 19). During meditation, both HF and LF power showed slight increases, with the elevations most pronounced in the early postmeditation period (after 0-10 min). Thereafter, HF gradually approached baseline levels, whereas LF returned to baseline after 10-20 min and subsequently decreased below baseline. These patterns also suggest transient alterations in autonomic nervous system activity following meditation.

In summary, our analysis revealed that during daily meditation practice, practitioners experience an increase in HRV, which may persist for a few tens of minutes after meditation.

Figure 5. Variation of heart rate variability levels during daily meditation practices. (A) Posterior density plots illustrating the average root-mean-square of successive differences (RMSSD) across different phases of meditation practice. RMSSD data were categorized as follows: the 10-minute before meditation phase, the during meditation phase, and the after meditation phases with 10-minute intervals up to 60 minutes post meditation. RMSSD was computed using start time-aligned 5-minute windows for the before and during phases, while end time-aligned windows were used for the after phases. The data were analyzed and visualized as in Figure 4B. (B) Start-aligned time-series RMSSD profiles across meditation practices of varying durations. Assuming the start of meditation as time 0, RMSSD was calculated using start time-aligned windows spanning from −20 to +55 minutes. Profiles were grouped by meditation duration in 10-minute intervals, and Markov Chain Monte Carlo–based hierarchical Bayesian analysis (Multimedia Appendix 13) was used to estimate posterior medians and 95% credible intervals of α , represented by solid and dotted lines, respectively. Horizontal time scales were aligned with the centers of the RMSSD windows. The shaded regions between the dotted lines denote the during meditation phase, while the lightly shaded areas indicate that the results are based on data partially originating from outside the during meditation phase. (C) End-aligned time-series RMSSD profiles assuming the end of meditation as time 0 and based on end time-aligned windows, visualized in the same manner as in panel (B). RMSSD: root-mean-square of successive differences.

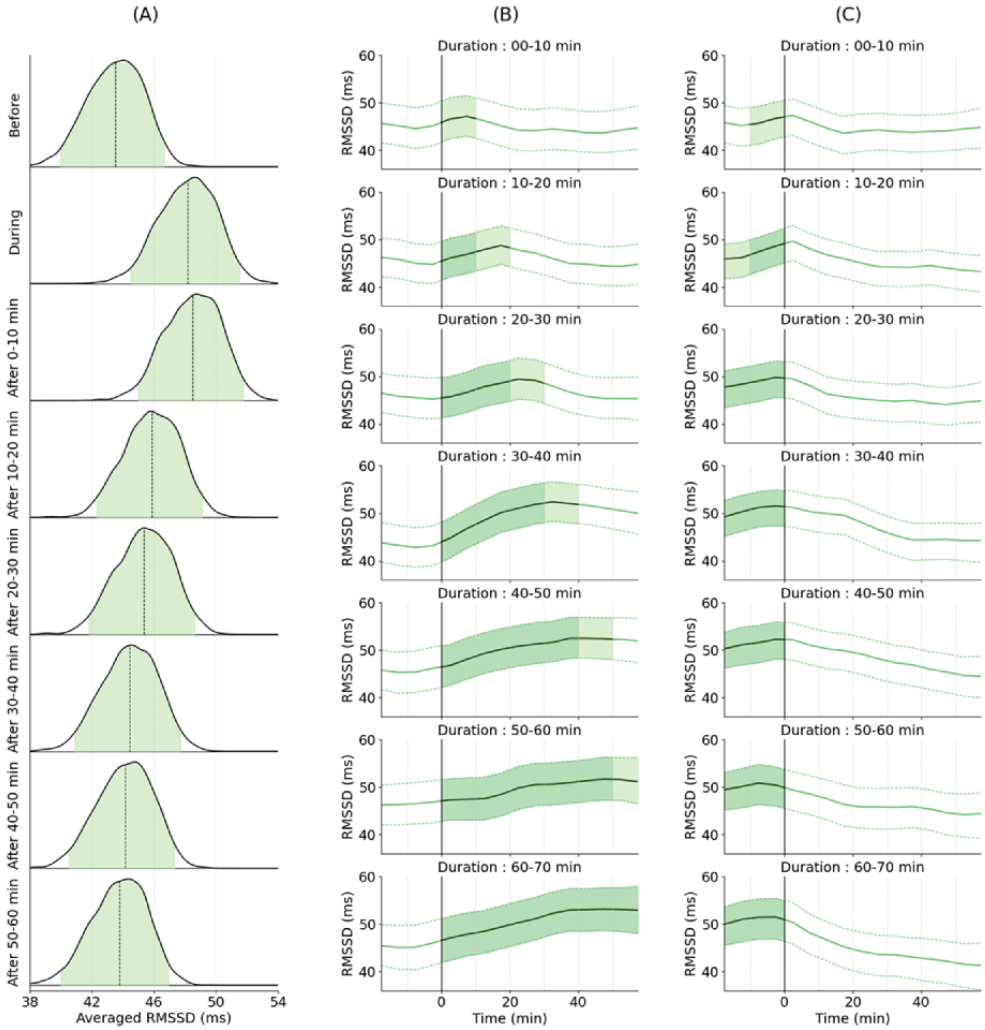

XSL•FO
RenderX

Discussion

Principal Findings

In this mHealth-based observational study, we integrated 3 weeks of continuous smartwatch-derived HRV monitoring with smartphone-based experience sampling to examine stress-HRV dynamics in meditation practitioners, recreational runners, and controls. Questionnaire and ESM data consistently indicated lower perceived stress in both the meditation and running groups relative to controls. However, daily-life HRV patterns differed; recreational runners showed higher HRV than controls, whereas meditation practitioners exhibited daily-life HRV levels comparable with controls. Across all groups, higher momentary stress was associated with concurrent reductions in HRV. Finally, within the meditation group, HRV increased during meditation sessions and remained elevated for tens of minutes after practice, suggesting a residual physiological effect detectable in naturalistic settings.

Interpretation in the Context of Previous Studies

Our questionnaire-based assessments indicated that perceived stress levels were lower in both the meditation and running groups compared with the control group, consistent with previous studies reporting stress-reducing effects of both meditation practice [2,3] and regular physical activity [34,35]. However, reduced subjective stress did not necessarily correspond to higher smartwatch-derived HRV in daily life. In fact, daily-life HRV levels were comparable between meditation practitioners and controls, suggesting that perceived stress and average daily HRV are not simply or linearly related.

Although multiple intervention studies have examined the impact of meditation practice on HRV, their findings have been mixed; some report increases in daily-life HRV among meditation groups relative to controls, while others observe no significant differences [28,31,32]. In our observational study involving experienced meditation practitioners, we found no evidence of a general increase in HRV throughout daily life. This interpretation is consistent with recent repeated-measures evidence showing that meditation primarily modulates HRV at the state level during meditation, particularly among experienced practitioners, even in the absence of elevated baseline HRV [33].

In contrast, and consistent with previous reports [36,37], the running group exhibited higher HRV as well as lower resting HR. This HRV elevation may reflect long-term cardiovascular and autonomic adaptations associated with endurance training, often referred to as an athlete's heart [51], including increased stroke volume, resting bradycardia, and enhanced vagal tone. Such physiological adaptations are likely to exert a stronger and more persistent influence on daily-life autonomic regulation than psychological training alone, potentially explaining the long-term HRV elevation observed in the running group.

To further explore the relationship between stress and HRV in daily life, we used a smartphone-based ESM integrated with smartwatch-derived HR monitoring. ESM also captured daily activities, allowing us to examine the links between stress and activity states. Across all groups, elevated stress levels were associated with reduced real-time HRV, in line with a previous study [26], as well as with specific stress-related activities such as work.

Consistent with the questionnaires, ESM confirmed that both meditation and running groups experienced lower stress levels in daily life compared with controls. While recreational runners showed more pronounced stress-related reductions in HRV, likely due to elevated baseline levels, meditation practitioners and controls exhibited comparatively smaller shifts.

Previous work has demonstrated that meditation training can attenuate HRV decreases during standardized cognitive stress tasks [52], suggesting that meditation may modulate stress appraisal rather than baseline autonomic function per se. Together, these findings raise the possibility that meditation practitioners experience identical daily-life stressors as less physiologically taxing, resulting in attenuated HRV responses. The absence of clear group differences in overall daily-life HRV does not necessarily contradict this interpretation. In naturalistic settings, stressors of sufficient intensity to elicit marked autonomic responses may occur relatively infrequently, and averaging across heterogeneous daily contexts may obscure group-level differences that emerge primarily under acute stress conditions.

For the meditation group, we also examined changes in HRV during daily meditation practice by integrating HRV data with meditation timing information collected via a smartphone app. We observed a significant increase in HRV during meditation. This finding is broadly consistent with laboratory-based studies reporting acute HRV increases during meditation [28,29,32,33], and extends the previous work by demonstrating such changes in naturalistic, daily-life settings.

Notably, our analysis also demonstrated that the HRV increase persisted for a few tens of minutes afterward, suggesting a substantial residual effect. To our knowledge, such a residual HRV elevation after meditation has not been previously documented. By capturing meditation practices in a naturalistic, daily-life context, this study provides unique insights into the physiological dynamics of meditation. Meditation practitioners' ability to increase HRV at an arbitrary timing, along with the prolonged residual effect, may correlate with stress reduction; however, this hypothesis warrants further investigation with adequate controls.

In the meditation-HRV analysis, changes were observed not only in the time-domain but also in the frequency-domain HRV indices. The concurrent increases in HF and LF observed during meditation and immediately after meditation are consistent with laboratory studies reporting enhanced parasympathetic activity and elevated overall HRV during meditative states [53]. HF gradually returned toward baseline after meditation, suggesting recovery from this parasympathetic-dominant state, while the delayed decrease in LF below baseline may reflect a transient reorganization of autonomic regulation after meditation ended. However, these interpretations should be made with caution. Frequency-domain indices derived from wrist-worn PPG may be less precise than those obtained from ECG [54,55], and the physiological interpretation of frequency-domain indices as markers of autonomic activity remains debated [14,56].

XSL•FO
RenderX

Therefore, although these observations may also support the residual effect of meditation, we refrain from drawing strong conclusions regarding the underlying mechanisms.

In addition, the observed HRV increase may be partly attributable to respiration-related heart rate variability (RespHRV) induced by changes in breathing pace [57,58]. Under typical conditions, RespHRV is primarily reflected in the HF band; however, slow and regular breathing, which commonly occurs during meditation, can shift respiratory oscillations toward lower frequencies and thereby contribute to increases in LF power. Because respiratory data were not collected in this study, we were unable to directly assess the contribution of RespHRV to the observed HRV changes. Recent studies have demonstrated that emerging wearable technologies, such as chest-mounted or textile-based respiratory sensors, enable direct and continuous measurement of breathing patterns in daily-life settings [59]. Incorporating such respiratory sensing technologies in future studies will help disentangle respiration-driven HRV modulation from other autonomic mechanisms and further clarify the physiological basis of meditation-related HRV changes.

From a methodological perspective, this study highlights the utility of Bayesian analysis for wearable PPG-derived HRV data. While most HRV studies have relied on ECG-based devices, this study used a Garmin smartwatch with PPG sensors, enabling the collection of 3 weeks of continuous data. Although the reliability of wearable PPG-derived HRV metrics has been debated due to susceptibility to noise [16-19], the combination of extensive data collection and robust statistical approaches allowed us to extract meaningful physiological insights. In particular, the hierarchical Bayesian approach used in our ESM-HRV and meditation-HRV analyses was well-suited to this setting, as it explicitly accounts for the nested and repeated-measures structure of the data. By partially pooling information across individuals, the model reduces the influence of noisy or uncertain observations at the individual level while preserving between-participant heterogeneity in baseline HRV. This framework enables more stable estimation of covariate-associated effects under realistic measurement uncertainty, which is especially relevant for wearable PPG-derived signals collected in daily life. Aligned with a recently published work [60], our achievement highlights the potential of Bayesian approaches to enhance the utility of wearable PPG-based devices in HRV research.

Limitations

Despite these promising results, our study remains preliminary, with several limitations requiring attention. First, unequal sample sizes across the 3 groups may have introduced imbalances in statistical power and estimation precision, which should be considered when interpreting between-group comparisons. Moreover, our recruitment procedures may have introduced potential biases in the composition of participant groups. A substantial proportion of the running group used their own Garmin smartwatches for data collection, while the control and meditation groups primarily used study-provided devices (Multimedia Appendix 1). This discrepancy arose from recruiting runners via Garmin users' mailing list. This could influence the results due to differences in device models and the users' familiarity with the devices, although we confirmed that recreational runners wearing different device models showed no significant difference in HRV levels (data not shown).

The meditation group also harbored potential recruitment biases. Participants in the meditation group were recruited through calls on social media and direct outreach to potential candidates. In the recruitment process, we noticed that more experienced meditation practitioners were often reluctant to wear a smartwatch on a daily basis, which may have biased our sample toward individuals with less meditation experience. Additionally, the meditation group included participants practicing different meditation styles, which may have introduced heterogeneity. In the questionnaire analysis, the mindfulness scores showed no significant difference, which may reflect not only the small sample size but also these concerns related to participant recruitment.

It is also important to consider that the 3 groups may have differed in covariates that could have influenced our analyses (Multimedia Appendix 1). Although age-adjusted sensitivity analyses using ANCOVA yielded results consistent with the primary nonparametric analyses, future studies should recruit larger and more balanced samples and incorporate more comprehensive covariate adjustment to strengthen the robustness and interpretability of the findings.

Finally, because respiration was not directly measured, we cannot fully disentangle stress-related autonomic changes from respiration-driven HRV modulation, particularly during meditation sessions. Moreover, the frequency-domain metrics reported in this study (HF and LF) are derived from PPG and are subject to greater estimation error than ECG-based measures, particularly under free-living conditions. Therefore, the interpretation of these indices should ideally be supported by validation against ECG-derived HRV measurements together with concurrent respiratory monitoring.

Conclusions

In this study, we used an mHealth framework integrating smartwatch-based physiological monitoring with smartphone-based experience sampling to examine how meditation practice is associated with subjective stress, HRV, and their interaction in real-world settings. We showed that meditation practitioners reported lower perceived stress but did not exhibit higher daily-life HRV than controls. Real-time associations between stress, activity states, and HRV were also comparable between the meditation and control groups. HRV analysis integrated with meditation timing logs showed that daily meditation practice was accompanied by transient increases in HRV, consistent with laboratory-based studies. Notably, we extend previous work by demonstrating that the HRV elevation persists for several tens of minutes after meditation practice.

Building on these findings, this study illustrates that mHealth-based approaches can capture dynamic relationships among subjective stress, daily activities, and autonomic physiology outside laboratory environments. Continuous wearable sensing yielded dense physiological time series, and

XSL•FO
RenderX

the integration of these data with in situ self-reports enabled the examination of stress-HRV associations at fine temporal resolution. Such multimodal data allowed us to characterize both baseline daily-life patterns and short-term physiological changes associated with meditation practice under naturalistic conditions.

From a digital health perspective, quantifying acute and short-lived physiological changes surrounding meditation sessions provides a basis for evaluating the timing and immediate after-effects of daily practice. The findings of this study suggest that mobile systems combining wearable sensors and self-reports may enable individual-level assessment of stress-HRV coupling and short-term recovery signals following meditation. Future studies incorporating additional physiological signals, such as respiration, and intervention-based designs will be necessary to further elucidate underlying mechanisms and to assess the practical utility of these integrated mHealth approaches for stress management.

Acknowledgments

We thank Yuki Ooishi (NTT Communication Science Laboratories) for insightful discussion about heart rate variability analysis, and Shunsuke Norihama (The University of Tokyo) for his valuable advice during the revision process.

Generative artificial intelligence tools were used to assist with English language editing and with coding for statistical analyses and figure generation in Python. All artificial intelligence–generated text and code were reviewed by the authors.

Funding

This study was supported by research funding from The University of Tokyo Excellent Young Researcher. No external research funding was provided.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JT: Data processing and analysis, and manuscript drafting
SG: App development, and data collection
MF: Study design, participant recruitment, and data analysis supervision
MM: Participant recruitment
KS: Data analysis supervision
KY: App development supervision
AN: Project management, study design, participant recruitment, data collection, data analysis, and manuscript refinement

Conflicts of Interest

MM is an employee of Garmin Japan Corporation and contributed solely to participant recruitment. She was not involved in the study design, data collection, data analysis, data interpretation, or manuscript preparation. Garmin Japan Corporation provided access to the Garmin Health SDK and electronic coupons for participants but did not provide any direct financial support for this research.

Multimedia Appendix 1
Characteristics of the recruited participants. (A) Recruitment flow chart of participants assigned to three groups: meditation (M), running (R), and control (C). (B) Sankey diagram illustrating the relationships between Garmin usage, group classification, and sex distribution. GU represents Garmin users, who participated in the study with their own smartwatches, while NGU represents non-Garmin users, who participated using provided smartwatches. The width of the flows corresponds to the number of participants transitioning between categories, with numbers in parentheses indicating the total participants in each category. (C) Age and (D) Body Mass Index (BMI) distribution across the three groups and differences in Garmin usage. In the violin plots (as presented in Figure 2), white dots represent female participants, while black dots represent male participants. BMI was calculated as weight (kg) divided by the square of height (m^2).
[PNG File , 243 KB - jmir_v28i1e78244_app1.png]

Multimedia Appendix 2
(A) Stacked bar chart showing the distribution of all 5-minute windows for each of the 90 participants. Colors indicate windows included in the selected 21-day analysis period (dark gray; 322,220 windows in total), valid windows outside this period (light gray; 82,146), and windows excluded by quality filters (red; 5,242). (B) Breakdown of excluded windows: invalid beat count (<150 or >1,000; 3,344), invalid SDNN (<10 ms or >200 ms; 1,857), and missing activity flag (41). Participants are grouped as

XSL•FO
RenderX

control (C), meditation (M), and running (R) and ordered by descending total window count. Dashed vertical lines indicate group boundaries.
[PNG File , 115 KB - jmir_v28i1e78244_app2.png]

Multimedia Appendix 3
Normality assessment and data quality summary for questionnaire scores and HRV/HR variables. Shapiro–Wilk test *P* values are reported for the full sample and for each group (C, M, and R). The number of missing observations and the number of IQR-based outliers are also shown.
[PNG File , 150 KB - jmir_v28i1e78244_app3.png]

Multimedia Appendix 4
Kruskal–Wallis tests comparing questionnaire scores and HRV/HR variables among the three groups (C: n=39; M: n=19; R: n=32). The table reports the Kruskal–Wallis test statistic (H), two-sided *P* values, and effect sizes quantified as epsilon squared (ε^2).
[PNG File , 120 KB - jmir_v28i1e78244_app4.png]

Multimedia Appendix 5
Pairwise group comparisons for questionnaire scores and overall HRV indices using Wilcoxon rank-sum tests. The table reports the U statistic, two-sided *P* values, Shaffer-corrected *P* values for multiple comparisons, and effect sizes (rank-biserial correlation, r; and common-language effect size). Group sizes were C=39, M=19, and R=32.
[PNG File , 266 KB - jmir_v28i1e78244_app5.png]

Multimedia Appendix 6
Pairwise group comparisons for activity-state RMSSD and heart rate (sleep, rest, walk, and run) using Wilcoxon rank-sum tests. The table reports the U statistic, two-sided *P* values, Shaffer-corrected *P* values for multiple comparisons, and effect sizes (rank-biserial correlation, r; and common-language effect size).
[PNG File , 231 KB - jmir_v28i1e78244_app6.png]

Multimedia Appendix 7
Age-adjusted group comparisons using analysis of covariance (ANCOVA). For each outcome, the table reports the F statistic for the group effect, two-sided *P* values, and effect sizes quantified as partial eta squared (η^2). As a sensitivity analysis, *P* values based on HC3 heteroskedasticity-consistent standard errors are also provided. The *P* value for the Group × Age interaction term is reported to assess the homogeneity of regression slopes assumption.
[PNG File , 133 KB - jmir_v28i1e78244_app7.png]

Multimedia Appendix 8
Binomial tests on the proportions of mid-high stress. ESM stress ratings were dichotomized (no-low vs mid-high). For each activity, the proportion of mid-high stress is shown for all participants (All) and for the control (C), meditation (M), and running (R) groups, together with the number of events (k), total observations (n), and estimated proportions ($\hat{p}$). Two-sided binomial tests assessed differences from the corresponding overall group proportion (Total), with *P* values adjusted using the Holm procedure and effect sizes reported as Cohen's h. Rows labeled All_ref, C_ref, M_ref, or R_ref indicate reference proportions and therefore have no *P* values or effect sizes. The bottom two rows (M vs C and R vs C) compare the meditation and running groups with the control group across all activities, using C_ref as the reference.
[PNG File , 249 KB - jmir_v28i1e78244_app8.png]

Multimedia Appendix 9
Welch's ANOVA comparing questionnaire scores and HRV/HR variables among the three groups (C: n=39; M: n=19; R: n=32) as a parametric sensitivity analysis. The table reports the Welch's F statistic with between-groups degrees of freedom fixed at 2, adjusted within-groups degrees of freedom (Adjusted df), two-sided *P* values, and effect sizes quantified as partial eta squared (ηp^2).
[PNG File , 14 KB - jmir_v28i1e78244_app9.png]

Multimedia Appendix 10
Games-Howell post hoc tests comparing questionnaire scores and overall heart rate variability (HRV) indices among the three groups (C: control; M: meditation; R: running). The table reports the t statistic, adjusted degrees of freedom (Adjusted df), two-sided *P* values, and effect sizes quantified as Hedges' g.
[PNG File , 22 KB - jmir_v28i1e78244_app10.png]

XSL•FO
RenderX

Multimedia Appendix 11
Games-Howell post hoc tests comparing activity-specific heart rate variability (HRV) and heart rate (HR) indices among the three groups (C: control; M: meditation; R: running). The table reports the t statistic, adjusted degrees of freedom (Adjusted df), two-sided *P* values, and effect sizes quantified as Hedges' g.
[PNG File , 19 KB - jmir_v28i1e78244_app11.png]

Multimedia Appendix 12
Probabilistic graphical representation of the model used for the comparison of RMSSD between stress levels or meditation phases. Circles and rectangles represent random variables and fixed constants, respectively. Shaded squares indicate observed data while a shaded circle indicates a variable of interest. Solid and dashed arrows represent probabilistic and deterministic dependencies, respectively. For definition of variables, see Methods.
[PNG File , 96 KB - jmir_v28i1e78244_app12.png]

Multimedia Appendix 13
Probabilistic graphical representation of the model used for the time-series analysis of RMSSD along meditation practices. Presented as in Multimedia Appendix 12.
[PNG File , 120 KB - jmir_v28i1e78244_app13.png]

Multimedia Appendix 14
Comparative analysis of additional HRV indices and activity-state heart rate. Violin plots illustrate group comparisons of (A) SDNN, (B) meanNN, (C) pNN50, and (D) HR stratified by four physical activity states (sleep, rest, walk, and run; derived from meanNN). Figure presentation and statistical tests were conducted in the same manner as in Figure 2.
[PNG File , 352 KB - jmir_v28i1e78244_app14.png]

Multimedia Appendix 15
Overviews of the ESM data. (A) A bar chart showing the counts of ESM responses across participants. Bar height reflects the total number of responses for each participant. Groups are presented from left to right as control (C), meditation (M), and running (R), with participants within each group ordered by the number of responses in descending order. (B) A stacked bar chart showing the proportions of seven daily activity states for each of the three groups. Bar widths are scaled according to the total number of ESM responses from participants in each group.
[PNG File , 95 KB - jmir_v28i1e78244_app15.png]

Multimedia Appendix 16
Posterior density plots of differences in averaged RMSSD at mid-high stress conditions with reference to no-low stress conditions. In the MCMC sampling process used to generate Figure 4B, posterior distributions were simultaneously estimated for differences in parameters corresponding to averaged RMSSD (α) between the different stress levels. The posterior densities are presented as in Figure 4B. The vertical highlighted dotted line indicates no difference, and negative values on the horizontal axis indicate a reduction in RMSSD under mid-high stress compared to no-low stress.
[PNG File , 107 KB - jmir_v28i1e78244_app16.png]

Multimedia Appendix 17
Overviews of the meditation records. (A) A stacked bar chart showing the number of meditation practices recorded by participants in the meditation groups. Each bar is composed of stacked segments, with each segment representing a single meditation practice from an individual participant. The height of each bar reflects the total count of meditation practices, with participants arranged in descending order along the horizontal axis based on their total count. Within each bar, stacked segments are also ordered in descending duration, with darker shades representing longer meditation practices. (B) A bar chart showing the distribution of meditation practice durations. Practices recorded by all the meditation group members are categorized into 10-minute intervals based on their duration, with the height of each bar representing the total number of practices whose durations fall within each interval. Bar colors correspond to the coding scheme shown in (A).
[PNG File , 98 KB - jmir_v28i1e78244_app17.png]

Multimedia Appendix 18
Posterior density plots of differences in averaged RMSSD at each meditation phase with reference to the "before" meditation phase. In the same manner as in Multimedia Appendix 16, posterior density plots of parameters α were generated from the MCMC sampling process used to generate Figure 5A. A positive value on the horizontal axis indicates an increase in RMSSD under each meditation phase compared to the before meditation phase.
[PNG File , 119 KB - jmir_v28i1e78244_app18.png]

XSL•FO
RenderX

Multimedia Appendix 19
Frequency-domain HRV indices during and after meditation. (A) and (B) Posterior distributions of log-scaled HF and LF are shown in the same manner as Figure 5A. (C) and (D) Baseline-adjusted shifts (differences from the pre-meditation baseline) of log-scaled HF and LF are shown in the same manner as Multimedia Appendix 18.
[PNG File , 339 KB - jmir_v28i1e78244_app19.png]

References

1. Kabat-Zinn J. Mindfulness-based interventions in context: past, present, and future. Clin Psychol Sci Pract 2003;10(2):144-156. [doi: 10.1093/clipsy.bpg016]
2. Brown KW, Ryan RM. The benefits of being present: mindfulness and its role in psychological well-being. J Pers Soc Psychol 2003;84(4):822-848. [doi: 10.1037/0022-3514.84.4.822] [Medline: 12703651]
3. Grossman P, Niemann L, Schmidt S, Walach H. Mindfulness-based stress reduction and health benefits. a meta-analysis. J Psychosom Res 2004;57(1):35-43. [doi: 10.1016/S0022-3999(03)00573-7] [Medline: 15256293]
4. Chiesa A, Serretti A. Mindfulness-based stress reduction for stress management in healthy people: a review and meta-analysis. J Altern Complement Med 2009;15(5):593-600. [doi: 10.1089/acm.2008.0495] [Medline: 19432513]
5. Khoury B, Sharma M, Rush SE, Fournier C. Mindfulness-based stress reduction for healthy individuals: a meta-analysis. J Psychosom Res 2015;78(6):519-528. [doi: 10.1016/j.jpsychores.2015.03.009] [Medline: 25818837]
6. Vibe M, Bjørndal A, Fattah S, Dyrdal GM, Halland E, Tanner‐Smith EE. Mindfulness‐based stress reduction (MBSR) for improving health, quality of life and social functioning in adults: a systematic review and meta‐analysis. Campbell Systematic Reviews 2017;13(1):1-264. [doi: 10.4073/csr.2017.11]
7. Lazarus RS, Folkman S. Stress, Appraisal, and Coping. New York, NY: Springer Publishing Company; 1984.
8. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983;24(4):385-396. [doi: 10.2307/2136404]
9. Csikszentmihalyi M, Larson R. Validity and reliability of the experience-sampling method. J Nerv Ment Dis 1987;175(9):526-536. [Medline: 3655778]
10. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
11. Bai S, Elavsky S, Kishida M, Dvořáková K, Greenberg MT. Effects of mindfulness training on daily stress response in college students: ecological momentary assessment of a randomized controlled trial. Mindfulness 2020;11(6):1433-1445 [FREE Full text] [doi: 10.1007/s12671-020-01358-x] [Medline: 33343764]
12. Matko K, Sedlmeier P, Bringmann HC. Differential effects of ethical education, physical hatha yoga, and mantra meditation on well-being and stress in healthy participants-an experimental single-case study. Front Psychol 2021;12:672301 [FREE Full text] [doi: 10.3389/fpsyg.2021.672301] [Medline: 34421729]
13. Sommerhoff A, Ehring T, Takano K. Effects of induced mindfulness at night on repetitive negative thinking: ecological momentary assessment study. JMIR Ment Health 2023;10:e44365 [FREE Full text] [doi: 10.2196/44365] [Medline: 37467038]
14. Shaffer F, Ginsberg JP. An overview of heart rate variability metrics and norms. Front Public Health 2017;5:258 [FREE Full text] [doi: 10.3389/fpubh.2017.00258] [Medline: 29034226]
15. Kim HG, Cheon EJ, Bai DS, Lee YH, Koo BH. Stress and heart rate variability: a meta-analysis and review of the literature. Psychiatry Investig 2018;15(3):235-245 [FREE Full text] [doi: 10.30773/pi.2017.08.17] [Medline: 29486547]
16. Pinheiro N, Couceiro R, Henriques J, Muehlsteff J, Quintal I, Goncalves L, et al. Can PPG be used for HRV analysis? 38th Annu Int Conf IEEE Eng Med Biol Soc 2016;2016:2945-2949. [doi: 10.1109/EMBC.2016.7591347] [Medline: 28268930]
17. Li K, Cardoso C, Moctezuma-Ramirez A, Elgalad A, Perin E. Heart rate variability measurement through a smart wearable device: another breakthrough for personal health monitoring? Int J Environ Res Public Health 2023;20(24):7146 [FREE Full text] [doi: 10.3390/ijerph20247146] [Medline: 38131698]
18. Dewig HG, Cohen JN, Renaghan EJ, Leary ME, Leary BK, Au JS, et al. Are wearable photoplethysmogram-based heart rate variability measures equivalent to electrocardiogram? a simulation study. Sports Med 2024;54(11):2927-2934. [doi: 10.1007/s40279-024-02066-5] [Medline: 38935328]
19. Rehman RZU, Chatterjee M, Manyakov NV, Daans M, Jackson A, O'Brisky A, et al. Assessment of physiological signals from photoplethysmography sensors compared to an electrocardiogram sensor: a validation study in daily life. Sensors (Basel) 2024;24(21):6826 [FREE Full text] [doi: 10.3390/s24216826] [Medline: 39517723]
20. Natarajan A, Pantelopoulos A, Emir-Farinas H, Natarajan P. Heart rate variability with photoplethysmography in 8 million individuals: a cross-sectional study. Lancet Digit Health 2020;2(12):e650-e657 [FREE Full text] [doi: 10.1016/S2589-7500(20)30246-6] [Medline: 33328029]
21. Tsai YY, Chen YJ, Lin YF, Hsiao FC, Hsu CH, Liao LD. Photoplethysmography-based HRV analysis and machine learning for real-time stress quantification in mental health applications. APL Bioeng 2025;9(2):026103 [FREE Full text] [doi: 10.1063/5.0256590] [Medline: 40191605]

XSL•FO
RenderX

22. Kageyama T, Nishikido N, Kobayashi T, Kurokawa Y, Kaneko T, Kabuto M. Self-reported sleep quality, job stress, and daytime autonomic activities assessed in terms of short-term heart rate variability among male white-collar workers. Ind Health 1998;36(3):263-272 [FREE Full text] [doi: 10.2486/indhealth.36.263] [Medline: 9701906]
23. Sin NL, Sloan RP, McKinley PS, Almeida DM. Linking daily stress processes and laboratory-based heart rate variability in a national sample of midlife and older adults. Psychosom Med 2016;78(5):573-582 [FREE Full text] [doi: 10.1097/PSY.0000000000000306] [Medline: 26867082]
24. Hynynen E, Konttinen N, Kinnunen U, Kyröläinen H, Rusko H. The incidence of stress symptoms and heart rate variability during sleep and orthostatic test. Eur J Appl Physiol 2011;111(5):733-741. [doi: 10.1007/s00421-010-1698-x] [Medline: 20972879]
25. Rieger A, Stoll R, Kreuzfeld S, Behrens K, Weippert M. Heart rate and heart rate variability as indirect markers of surgeons intraoperative stress. Int Arch Occup Environ Health 2014;87(2):165-174. [doi: 10.1007/s00420-013-0847-z] [Medline: 23370764]
26. Martinez GJ, Grover T, Mattingly SM, Mark G, D'Mello S, Aledavood T, et al. Alignment between heart rate variability from fitness trackers and perceived stress: perspectives from a large-scale in situ longitudinal study of information workers. JMIR Hum Factors 2022;9(3):e33754 [FREE Full text] [doi: 10.2196/33754] [Medline: 35925662]
27. Fletcher R, Zeng K, Yang MY, Pietrzak A, Eddie D. Development of a heart rate variability based ambulatory stress detection model for clinical populations. Appl Psychophysiol Biofeedback 2026;51(1):19-28. [doi: 10.1007/s10484-025-09714-0] [Medline: 40411737]
28. Krygier JR, Heathers JAJ, Shahrestani S, Abbott M, Gross JJ, Kemp AH. Mindfulness meditation, well-being, and heart rate variability: a preliminary investigation into the impact of intensive Vipassana meditation. Int J Psychophysiol 2013;89(3):305-313. [doi: 10.1016/j.ijpsycho.2013.06.017] [Medline: 23797150]
29. Tang Y, Ma Y, Fan Y, Feng H, Wang J, Feng S, et al. Central and autonomic nervous system interaction is altered by short-term meditation. Proc Natl Acad Sci USA 2009;106(22):8865-8870 [FREE Full text] [doi: 10.1073/pnas.0904031106] [Medline: 19451642]
30. Grossman P, Deuring G, Walach H, Schwarzer B, Schmidt S. Mindfulness-Based intervention does not influence cardiac autonomic control or the pattern of physical activity in fibromyalgia during daily life: an ambulatory, multimeasure randomized controlled trial. Clin J Pain 2017;33(5):385-394. [doi: 10.1097/AJP.0000000000000420] [Medline: 27518489]
31. Brown L, Rando AA, Eichel K, Van Dam NT, Celano CM, Huffman JC, et al. The effects of mindfulness and meditation on vagally mediated heart rate variability: a meta-analysis. Psychosom Med 2021;83(6):631-640 [FREE Full text] [doi: 10.1097/PSY.0000000000000900] [Medline: 33395216]
32. Barbry A, Gál É, Carton A, Orianne P, De Raedt R. Effect of brief mindfulness meditation interventions on heart rate variability in adults: a systematic review. Appl Psychophysiol Biofeedback 2026;51(1):3-18. [doi: 10.1007/s10484-025-09724-y] [Medline: 40715924]
33. Wei Y, Xu Y, Chen W, Zheng J, Chen H, Chen S. Can heart rate variability demonstrate the effects and the levels of mindfulness? a repeated-measures study on experienced and novice mindfulness practitioners. BMC Complement Med Ther 2025;25(1):231 [FREE Full text] [doi: 10.1186/s12906-025-04972-1] [Medline: 40611081]
34. Mikkelsen K, Stojanovska L, Polenakovic M, Bosevski M, Apostolopoulos V. Exercise and mental health. Maturitas 2017;106:48-56. [doi: 10.1016/j.maturitas.2017.09.003] [Medline: 29150166]
35. Herbert C, Meixner F, Wiebking C, Gilg V. Regular physical activity, short-term exercise, mental health, and well-being among university students: the Results of an online and a laboratory study. Front Psychol 2020;11:509 [FREE Full text] [doi: 10.3389/fpsyg.2020.00509] [Medline: 32528333]
36. Kiss O, Sydó N, Vargha P, Vágó H, Czimbalmos C, Édes E, et al. Detailed heart rate variability analysis in athletes. Clin Auton Res 2016;26(4):245-252. [doi: 10.1007/s10286-016-0360-z] [Medline: 27271053]
37. Plews DJ, Laursen PB, Stanley J, Kilding AE, Buchheit M. Training adaptation and heart rate variability in elite endurance athletes: opening the door to effective monitoring. Sports Med 2013;43(9):773-781. [doi: 10.1007/s40279-013-0071-8] [Medline: 23852425]
38. Oshita D. Introduction to the Clinical Meditation Method: Its Merits And a Practical Guide for Therapists and Care Providers. Sagamihara: Being-Net Press; 2025.
39. Guidelines for data processing and analysis of the International Physical Activity Questionnaire - short and long forms. International Physical Activity Questionnaire (IPAQ) Research Committee. 2005. URL: https://drive.google.com/file/d/1gehdq1-04eSWfbxscwtzXa1MUlD8Mffa/view [accessed 2026-04-16]
40. Doi Y, Minowa M, Okawa M, Uchiyama M. Development of the japanese version of the pittsburgh sleep quality index. Jpn J Psychiat Treat 2005;35(7):530-534. [doi: 10.1007/s00595-004-2977-0] [Medline: 15976948]
41. Fujino M, Kajimura S, Nomura M. Development and validation of the japanese version of the mindful attention awareness scale using item response theory analysis. The Japanese Journal of Personality 2015;24(1):61-76. [doi: 10.2132/personality.24.61]
42. Sumi K. Reliability and validity of the Japanese version of the perceived stress scale. Jpn J Health Psychol 2006;19(2):44-53. [doi: 10.11560/jahp.19.2_44]

XSL•FO
RenderX

43. World Health Organization. The World Health Organization-five well-being index (WHO-5). 2024. URL: https://www.who.int/publications/m/item/WHO-UCN-MSD-MHE-2024.01 [accessed 2025-05-17]
44. Garmin Health. Garmin enhanced beat-to-beat interval (BBI): technical white paper. Garmin ltd. 2020. URL: https://www8.garmin.com/garminhealth/news/Garmin-Enhanced-BBI_Final.pdf [accessed 2025-05-17]
45. Kleiger RE, Stein PK, Bigger JT. Heart rate variability: measurement and clinical utility. Ann Noninvasive Electrocardiol 2005;10(1):88-101 [FREE Full text] [doi: 10.1111/j.1542-474X.2005.10101.x] [Medline: 15649244]
46. Sarhaddi F, Kazemi K, Azimi I, Cao R, Niela-Vilén H, Axelin A, et al. A comprehensive accuracy assessment of Samsung smartwatch heart rate and heart rate variability. PLoS One 2022;17(12):e0268361 [FREE Full text] [doi: 10.1371/journal.pone.0268361] [Medline: 36480505]
47. Lam E, Aratia S, Wang J, Tung J. Measuring heart rate variability in free-living conditions using consumer-grade photoplethysmography: validation study. JMIR Biomed Eng 2020;5(1):e17355. [doi: 10.2196/17355]
48. Shaffer JP. Modified sequentially rejective multiple test procedures. J Am Stat Assoc 2012;81(395):826-831. [doi: 10.1080/01621459.1986.10478341]
49. Umetani K, Singer DH, McCraty R, Atkinson M. Twenty-four hour time domain heart rate variability and heart rate: relations to age and gender over nine decades. J Am Coll Cardiol 1998;31(3):593-601 [FREE Full text] [doi: 10.1016/s0735-1097(97)00554-8] [Medline: 9502641]
50. de Geus EJC, Gianaros PJ, Brindle RC, Jennings JR, Berntson GG. Should heart rate variability be "corrected" for heart rate? biological, quantitative, and interpretive considerations. Psychophysiology 2019;56(2):e13287 [FREE Full text] [doi: 10.1111/psyp.13287] [Medline: 30357862]
51. Maron BJ, Pelliccia A. The heart of trained athletes: cardiac remodeling and the risks of sports, including sudden death. Circulation 2006;114(15):1633-1644. [doi: 10.1161/circulationaha.106.613562]
52. Shearer A, Hunt M, Chowdhury M, Nicol L. Effects of a brief mindfulness meditation intervention on student stress and heart rate variability. Int J Stress Manag 2016;23(2):232-254. [doi: 10.1037/a0039814]
53. Nesvold A, Fagerland MW, Davanger S, Ellingsen Ø, Solberg EE, Holen A, et al. Increased heart rate variability during nondirective meditation. Eur J Prev Cardiol 2012;19(4):773-780. [doi: 10.1177/1741826711414625] [Medline: 21693507]
54. Theurl F, Schreinlechner M, Sappler N, Toifl M, Dolejsi T, Hofer F, et al. Smartwatch-derived heart rate variability: a head-to-head comparison with the gold standard in cardiovascular disease. Eur Heart J Digit Health 2023;4(3):155-164 [FREE Full text] [doi: 10.1093/ehjdh/ztad022] [Medline: 37265873]
55. Mejía-Mejía E, Budidha K, Abay TY, May JM, Kyriacou PA. Heart rate variability (HRV) and pulse rate variability (PRV) for the assessment of autonomic responses. Front Physiol 2020;11:779 [FREE Full text] [doi: 10.3389/fphys.2020.00779] [Medline: 32792970]
56. Billman GE. The LF/HF ratio does not accurately measure cardiac sympatho-vagal balance. Front Physiol 2013;4:26 [FREE Full text] [doi: 10.3389/fphys.2013.00026] [Medline: 23431279]
57. Menuet C, Ben-Tal A, Linossier A, Allen AM, Machado BH, Moraes DJA, et al. Redefining respiratory sinus arrhythmia as respiratory heart rate variability: an international expert recommendation for terminological clarity. Nat Rev Cardiol 2025;22(12):978-984. [doi: 10.1038/s41569-025-01160-z] [Medline: 40328963]
58. Lehrer PM, Gevirtz R. Heart rate variability biofeedback: how and why does it work? Front Psychol 2014;5:756 [FREE Full text] [doi: 10.3389/fpsyg.2014.00756] [Medline: 25101026]
59. Hussain T, Ullah S, Fernández-García R, Gil I. Wearable sensors for respiration monitoring: a review. Sensors (Basel) 2023;23(17):7518 [FREE Full text] [doi: 10.3390/s23177518] [Medline: 37687977]
60. Corponi F, Li BM, Anmella G, Valenzuela-Pascual C, Pacchiarotti I, Valentí M, et al. A Bayesian analysis of heart rate variability changes over acute episodes of bipolar disorder. Npj Ment Health Res 2024;3(1):44 [FREE Full text] [doi: 10.1038/s44184-024-00090-x] [Medline: 39363115]

Abbreviations

ANCOVA: analysis of covariance
BBI: beat-to-beat interval
ECG: electrocardiography
ESM: experience sampling method
HF: high-frequency
HR: heart rate
HRV: heart rate variability
IPAQ: International Physical Activity Questionnaire
LF: low-frequency
MAAS: Mindful Attention Awareness Scale
MCMC: Markov Chain Monte Carlo
meanNN: mean of normal-to-normal interval
mHealth: mobile health

XSL•FO
RenderX

PPG: photoplethysmography
PSQI: Pittsburgh Sleep Quality Index
PSS: Perceived Stress Scale
RespHRV: respiration-related heart rate variability
RMSSD: root-mean-square of successive differences
SDNN: standard deviation of normal-to-normal interval
WHO-5: World Health Organization Well-being Index

Edited by A Stone; submitted 29.May.2025; peer-reviewed by R Gevirtz, K Pham; comments to author 09.Dec.2025; accepted 08.Apr.2026; published 29.May.2026.

Please cite as:
Takezawa J, Geng S, Fujino M, Miyake M, Sasahara K, Yatani K, Niida A
Daily Stress and Heart Rate Variability Among Mindfulness Meditation Practitioners: mHealth Observational Study
J Med Internet Res 2026;28:e78244
URL: https://www.jmir.org/2026/1/e78244
doi:10.2196/78244
PMID:

©Jo Takezawa, Shixian Geng, Masahiro Fujino, Mika Miyake, Kazutoshi Sasahara, Koji Yatani, Atsushi Niida. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Understanding mHealth Engagement Among Patients With 30-Day Hospital Revisits: Secondary Analysis of a Randomized Clinical Trial

Susan Landon[1], MSc, MD; Angira Mondal[2], MS; Aiden Ahn[1], MD; Klea Profka[3,4], BA; Anna U Morgan[1,2], MSc, MD; Eric Bressman[1,2,5], MSHP, MD

[1]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[2]Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, PA, United States
[3]Center for Health Incentives and Behavioral Economics, University of Pennsylvania, Philadelphia, PA, United States
[4]Department of Medical Ethics and Health Policy, University of Pennsylvania, Philadelphia, PA, United States
[5]Philadelphia VA Medical Center, Philadelphia, PA, United States

Corresponding Author:
Susan Landon, MSc, MD
Department of Medicine
Perelman School of Medicine
University of Pennsylvania
3400 Spruce St
Philadelphia, PA, 19104
United States
Phone: 1 215 349 5200
Email: Susan.Landon@pennmedicine.upenn.edu

Abstract

Background: Reducing 30-day hospital readmissions has been a long-standing goal across health systems in the United States. While nurse-led phone outreach has been widely adopted to support transitional care, its reach is constrained by staffing and time limitations. Mobile health (mHealth) interventions, such as automated SMS text messaging and patient portals, offer scalable alternatives but have shown mixed effectiveness in reducing readmissions. Understanding how patients engage with mHealth after discharge may help optimize these tools for postdischarge care.

Objective: This study aimed to characterize patients in an mHealth transitional care program who experienced hospital revisits within 30 days of discharge, comparing demographic and clinical characteristics, intervisit interactions, and revisit features between those who engaged with mHealth and those who did not.

Methods: We conducted a secondary analysis of patients in the intervention arm of the Mobile Outreach to Reduce Emergencies-Primary Care randomized clinical trial. Participants received automated SMS text messages for 30 days after discharge alongside usual transitional care. We identified patients with a 30-day hospital revisit and conducted manual chart reviews to assess mHealth engagement and other forms of health care contact. We compared patient characteristics, intervisit interactions, and revisit features (time to revisit, relatedness to index hospitalization, and predictability of revisit) between mHealth users and nonusers.

Results: Among 496 patients with a 30-day revisit, 185 (37%) engaged with mHealth before their return. mHealth users were younger (n=47, 26% aged <50 years vs n=41, 14% among nonusers; *P*=.004) and more likely to have commercial insurance (n=43, 23% mHealth users vs n=35, 11% mHealth nonusers; *P*=.005). Revisits were more likely to be rated highly or somewhat predictable among mHealth users compared to nonusers (n=105, 56% vs n=152, 49%; *P*=.04), while relatedness to the index hospitalization was similar (n=98, 53% vs n=173, 55%; *P*=.09). mHealth users had a longer mean time to revisit than nonusers (15.2, SD 8.1 vs 11.3, SD 8.4 days; *P*<.001) and were more likely to contact their practices via telephone (n=100, 54% vs n=136, 44%; *P*=.03) or attend a clinic visit (n=112, 61% vs n=131, 42%; *P*<.001).

Conclusions: Among patients enrolled in an mHealth postdischarge program who experienced hospital revisits, fewer than half engaged with mHealth prior to their return. Revisits among mHealth users occurred later and were more predictable, suggesting

XSL•FO
RenderX

that engagement may enhance situational awareness but not necessarily prevent revisits. Future work should focus on strategies to increase engagement across groups and integrate mHealth with existing transitional care infrastructure.

(J Med Internet Res 2026;28:e89067) doi:10.2196/89067

KEYWORDS

mobile health; mHealth; readmissions; transitional care management; quality improvement

Introduction

In 2012, the Centers for Medicare and Medicaid Services introduced the Hospital Readmissions Reduction Program, creating financial penalties for hospitals with elevated 30-day readmission rates and formalizing readmissions as a core quality metric [1,2]. Despite this policy, between 2016 and 2020, hospital readmission rates remained relatively steady around 14% [3]. Nonetheless, in an effort to reduce readmissions, health systems have focused on improving transitions of care, recognizing that these points in time are particularly vulnerable to error and miscommunication [4]. To date, health systems have largely focused on nurse-led phone outreach to strengthen transitional care management, but the reach of such programs is limited by staffing and time constraints [1,5,6].

Opportunities for remote interactions, however, extend beyond telephone calls. Mobile health (mHealth), defined by the World Health Organization as "medical and public health practice supported by mobile devices, such as mobile phones, patient monitoring devices, personal digital assistants, and other wireless devices," offers a scalable alternative [7]. With expanding technology capabilities, mHealth merits particular attention for improvement initiatives. During transitions of care, mHealth interventions may take the form of, for example, a mobile app, web-based application, SMS text messaging, or remote monitoring [8,9]. Early studies of postdischarge mHealth interventions suggest possible modest readmission reductions [10-12]. They have also shown potential for improving adherence to postdischarge medication regimens and follow-up instructions, as well as increased patient-reported outcomes during the transitional care period [13,14]. In addition to aiding in adherence, it is possible that mHealth interventions may aid in early identification of complications or clinical deteriorations in transitional care periods. Particularly in cases where an index hospitalization and readmission are directly related, enhanced communication between patients and providers may help to increase predictability of a possible readmission with the intention of identifying potential avenues for intervention to avoid hospital revisits where feasible.

To overcome barriers of call-based strategies, the Mobile Outreach to Reduce Emergencies-Primary Care (MORE-PC) trial, one of the largest postdischarge mHealth trials to date, tested implementation of an SMS text messaging–based platform to reduce 30-day readmissions, showing no significant reduction in readmissions [15]. In this secondary analysis, we study patients in the intervention arm with 30-day hospital revisits. Comparing patients who used mHealth and patients who did not, our objectives were to analyze patient characteristics, interaction patterns leading up to revisits, and revisit characteristics to inform future development of mHealth transitional care interventions.

Methods

Study Design

We conducted a subgroup analysis using data from the MORE-PC randomized clinical trial and manual chart review. Study reporting followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines (Multimedia Appendix 1) [16].

Setting

The MORE-PC trial took place between March 2022 and January 2023 within an academic health system in Philadelphia, Pennsylvania. Participants in the intervention arm received automated SMS text messages for 30 days alongside standard transitional care management. Participants in the control arm received standard transitional care management. Full details on the study design can be found in the original manuscript [15].

Participants

Patients included in the MORE-PC trial were adults who received care within a University of Pennsylvania Health System's primary care practice and had been discharged home from any regional acute care facility. Patients in the intervention arm who had a hospital revisit (return emergency department [ED] visit or readmission) within 30 days of index discharge were included. We included all patients who met these criteria in our analyses.

Ethical Considerations

This study was considered a part of the larger MORE-PC study, which was reviewed and approved by the University of Pennsylvania Institutional Review Board (protocol number 24-0431). The trial was conducted under a waiver of informed consent based on meeting criteria that both arms receive usual care, SMS text messages involved no more than minimal risk to patients, patients were free to opt out at any time after enrollment, and requiring informed consent would limit external validity of the study [15]. All data were stored on the Penn network to protect privacy. In addition, following completion of chart reviews, patient identifiers were removed for subsequent analyses. Trial participants received no compensation.

Variables, Data Sources, and Measurement

The primary independent variable of interest was mHealth engagement, defined as using the MORE-PC SMS text messaging platform or patient portal, identified through MORE-PC data and patient charts. We used data collected on patients' age, sex, race, ethnicity, insurer, and comorbidities from MORE-PC data files. Data on intervisit health care

XSL•FO
RenderX

interactions—mHealth use, phone calls, and clinic visits—were collected from MORE-PC data files and manual chart review. We also assessed 3 primary outcome variables: time to revisit, relatedness of index hospitalization, and revisit and predictability of revisit. Time to revisit was calculated as days from discharge to revisit and was a classified as a continuous variable. Relatedness and predictability were classified as ordinal variables. Relatedness was defined as any connection between index and revisit diagnoses, as determined through chart review. For example, 2 visits for decompensated heart failure would be considered directly related. Similarly, an index visit for deep vein thrombosis and revisit for gastrointestinal bleed after starting anticoagulation would be considered directly related. An admission for pancreatitis and an admission for alcohol withdrawal would be considered indirectly related as alcohol use disorder underlies both presentations. Predictability describes the extent to which a revisit could have been anticipated based on information available to the outpatient team through review of medical records considering issues remaining at discharge and concerns brought up between visits. For example, if a patient was discharged after experiencing a bleeding gastrointestinal ulcer, reported recurrence of blood in the stool at a follow-up primary care visit, then represented to the ED, this would be considered predictable. By contrast, if a patient with an index hospitalization for pyelonephritis who had no noted complications posthospitalization, then represented to the ED for an acute stroke, this would be considered unpredictable.

Chart Review

We developed a template for systematically reviewing and extracting data from electronic health records of the study population (Multimedia Appendix 2) in an effort to minimize bias in this process. We tested the template on 15 records to identify areas of ambiguity. Subsequently, 3 researchers (SL, AA, and KP) reviewed the same 5 records to ensure agreement across data elements. Remaining records were divided among these researchers for data extraction.

Classification of relatedness and predictability were determined through review of index visit, intervisit, and revisit documentation.

Statistical Methods

We examined patient-level and encounter-level characteristics, comparing those engaged in mHealth and those not engaged. We assessed differences in baseline characteristics and outcome variables across groups using Wilcoxon tests for continuous variables and χ^2or Fisher exact tests for categorical variables. For comparing intervisit interactions across groups, we performed 2-sample proportion testing. Patients with missing data for a given variable were excluded for that analysis. Analyses were conducted using SAS Studio (version 3.81; SAS Institute) and R software (version 4.3.2; R Foundation for Statistical Computing).

Results

Overview

A total of 496 patients in the intervention arm of the MORE-PC trial had a hospital revisit within 30 days of discharge. Of these, 296 (60%) were female, 217 (44%) were aged ≥71 years, 283 (57%) were Black, and 315 (64%) were insured by Medicare (Table 1). A total of 185 (37%) patients engaged with mHealth prior to their revisit.

XSL•FO
RenderX

Table 1. Characteristics of patients with 30-day hospital revisits and revisit-related outcomes, overall and stratified by mobile health (mHealth) engagement (N=496).

	Overall	Engaged in mHealth (n=185)	Not engaged in mHealth (n=311)	*P* value
Demographic characteristics				
Female, n (%)	296 (60)	106 (57)	190 (61)	.40
Age group (years), n (%)				.004
<30	15 (3)	7 (4)	8 (3)	
31-50	73 (15)	40 (22)	33 (11)	
51-70	192 (39)	70 (38)	122 (39)	
≥71	217 (44)	69 (37)	148 (48)	
Race, n (%)				.54
Black	283 (57)	98 (53)	185 (60)	
White	187 (38)	78 (42)	109 (35)	
Asian or Pacific Islander	3 (1)	1 (1)	2 (1)	
Other	23 (5)	8 (4)	15 (5)	
Ethnicity, n (%)				.59
Hispanic	16 (3)	7 (4)	9 (3)	
Non-Hispanic	480 (97)	178 (96)	302 (97)	
Insurance, n (%)				.005
Medicare	315 (64)	109 (59)	206 (66)	
Medicaid	95 (19)	31 (17)	64 (21)	
Private	78 (16)	43 (23)	35 (11)	
Other	8 (2)	2 (1)	6 (2)	
Clinical characteristics				
Comorbidity score[a], μ (SE)	5.22 (3.0)	5.13 (3.1)	5.27 (3.0)	.50
Index length of stay, μ (SE)	5.51 (12.6)	5.04 (5.2)	5.79 (15.4)	.75
Acute care visits within 12 months before index, μ (SE)				
Emergency department visits	3.58 (5.8)	3.08 (3.7)	3.89 (6.8)	.45
Admissions	1.65 (2.2)	1.66 (2.2)	1.63 (2.2)	.88
Outcomes				
Revisit type[b], n (%)				.37
Emergency department visit	191 (39)	68 (37)	123 (40)	
Readmission	302 (61)	117 (63)	185 (59)	
Predictability of revisit, n (%)				.04
Highly	115 (23)	47 (25)	68 (22)	
Somewhat	142 (29)	58 (31)	84 (27)	
Unknown	11 (2)	0 (0)	11 (4)	
Not at all	228 (46)	80 (43)	148 (48)	
Revisit related to index admission, n (%)				.09
Directly	213 (43)	69 (37)	144 (46)	
Indirectly	58 (12)	29 (16)	29 (9)	
Unrelated	165 (33)	64 (35)	101 (32)	
Unsure	60 (12)	23 (12)	37 (12)	

XSL•FO
RenderX

	Overall	Engaged in mHealth (n=185)	Not engaged in mHealth (n=311)	*P* value
Patient referred back to emergency department, n (%)	160 (32)	66 (36)	94 (30)	.21
Time from discharge to revisit, μ (SE)	12.77 (8.53)	15.23 (8.13)	11.31 (8.44)	<.001

[a]n=462 (171 engaged in mHealth and 298 not engaged in mHealth).

[b]n=493 (185 engaged in mHealth and 308 not engaged in mHealth).

Patients who engaged with mHealth were younger (n=47, 26% aged <50 years vs n=41, 14% of those not engaged; *P*=.004 across age groups) and more likely to have commercial insurance (n=43, 23% vs n=35, 11%; *P*=.01). Patients in both groups had similar rates of ED visits and hospitalizations in the year prior to index admission (ED visits, mHealth users vs nonusers: 3.08 vs 3.89, *P*=.45; hospital admissions: 1.66 vs 1.63, *P*=.88).

Outcomes

Across both groups, hospital readmissions were nearly twice as common as standalone ED returns (117/185, 63% of revisits among mHealth users vs 185/311, 59% among nonusers). Revisits were more predictable among patients who engaged with mHealth (105/185, 56% highly or somewhat predictable vs 152/311, 49% among non-mHealth users; *P*=.04). A similar proportion of revisits were directly or indirectly related to the index admission among mHealth users and nonusers (98/185, 53% vs 173/311, 55%; *P*=.09). mHealth users and nonusers had similar rates of ED referrals from their outpatient teams (66/185, 36% vs 94/311, 30%; *P*=.21). On average, those engaged with mHealth had more time between their index discharge and revisit (15.2 vs 11.3 days; *P*<.001).

Intervisit Interactions

Among mHealth users, similar numbers of patients used SMS text messaging versus the patient portal (111/185, 60% vs 107/185, 58%; Figure 1, Multimedia Appendix 3). mHealth users were more likely than nonusers to engage with practices via non-mHealth modalities (telephone: 54% vs 44%, *P*=.03; clinic visits: 112/185, 61% vs 131/311, 42%; *P*<.001). Both groups had similar rates of home care visits (72/185, 39% vs 114/311, 37%; *P*=.60).

Figure 1. Summary of intervisit health care interactions. Categories are not mutually exclusive.**P* value significant at *P*<.05 (only applied for interaction categories relevant to both study arms). mHealth: mobile health.

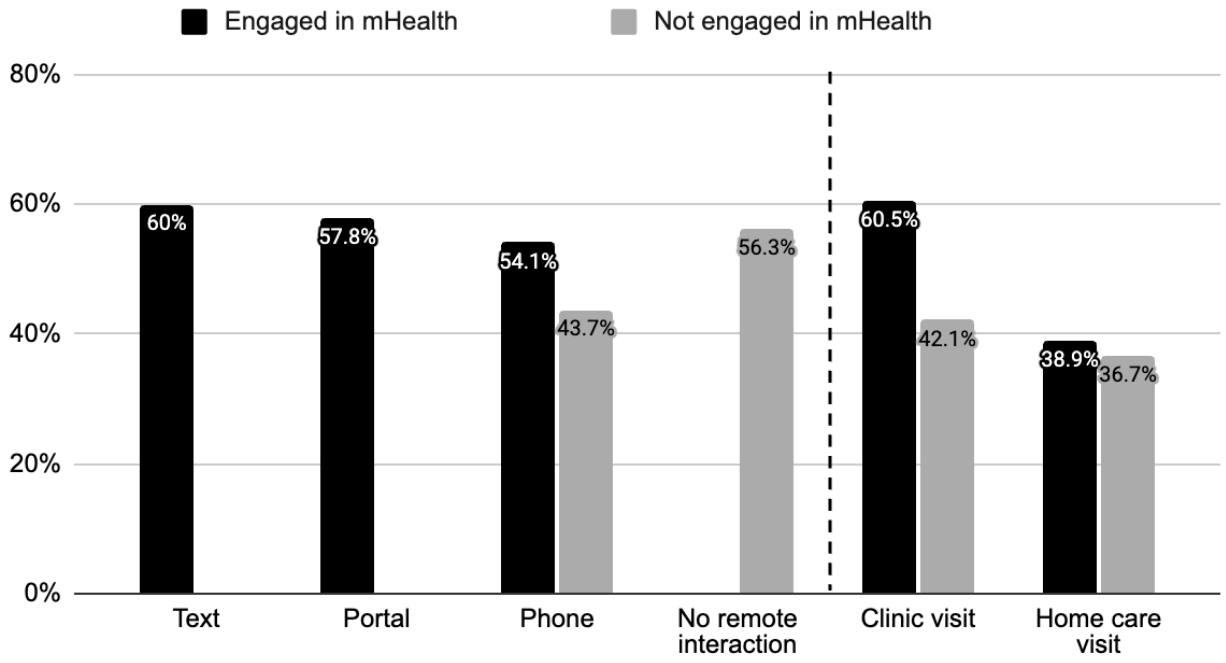

Discussion

Principal Findings

We examined patient, hospitalization, and intervisit health care interaction features among patients enrolled in an mHealth program, MORE-PC, with 30-day hospital revisits. We found that despite active outreach offered by the MORE-PC program, a minority of patients in the intervention arm used mHealth to communicate with outpatient providers before returning to the hospital. Notably, the parent MORE-PC trial found no significant reduction in 30-day readmissions, suggesting that the presence of an mHealth outreach program alone may be insufficient to reduce revisits without sufficient patient engagement. mHealth users tended to be younger and more likely to have commercial insurance. Predictability of revisits was high among those who used mHealth, and time from index discharge to revisit was longer as compared to those who did

XSL•FO
RenderX

not use mHealth. In addition, those who used mHealth were also more likely to use other means of communication with outpatient providers.

Similar to our study, previous research has shown disparities in mHealth use by age, socioeconomic status, and insurer [17,18]. One potential explanation for the differences in usage by age may be that older patients have more difficulty using technology or be less inclined to adopt new technology [19,20]. Insurance and socioeconomic differences likely go hand in hand, and may be associated with mHealth use through mechanisms such as education or affordability of technology [18]. Interestingly, we did not find significant differences in mHealth use by race or ethnicity, though prior work has shown disparities by race and ethnicity [17,18]. With future development and expansion of mHealth interventions, it is crucial to consider ways of ensuring equitable implementation of mHealth interventions given early evidence of disparities across age, racial or ethnic, and socioeconomic groups.

Unsurprisingly, we found higher levels of revisit predictability among those who used mHealth. Though expected with greater communication with practices, this is encouraging as it suggests increased awareness of issues prompting a revisit. Though predictability may not have any bearing on preventability of revisits, timely identification of areas for potential intervention may help to avert a subset of revisits. In line with this, we found time from discharge to revisit was, on average, longer for mHealth users, suggesting mHealth use may delay but not prevent revisits. In line with previous work, we found no major difference in relatedness of index hospitalizations and revisits [21].

mHealth users interacted with outpatient teams by telephone and clinic visits at higher rates than nonusers. This raises concern that mHealth offerings may be additive, rather than substitutive, to traditional communication modalities, benefitting engaged groups more than reaching disengaged groups. For example, although it appears promising that mHealth use is associated with increased predictability and prolonged time to revisit, it is possible that this is more reflective of patient engagement than true advantages of mHealth. Previous research has explored such challenges, highlighting engagement of diverse patients for development of health applications and usage of community health workers to encourage uptake, though more work is needed in this area [22,23].

Limitations

First, this study included patients within a single health system, and thus results may not be generalizable. Second, chart review methods may be subject to bias. We created data extraction templates and reviewed an initial sample as a group to try to minimize discrepancies across reviewers but recognize this bias cannot be entirely eliminated. Third, we categorized mHealth engagement as responding to SMS text messages or sending portal messages, but definitions of mHealth vary, and it is possible that those who engage via SMS text message may be different from those who engage via portal. Moreover, defining engagement through a binary variable may not capture meaningful differences across a spectrum of mHealth engagement. Finally, this study focuses on a single sample of mHealth users and nonusers with 30-day revisits but does not capture items such as patient satisfaction or strength of provider handoffs that may also be used to define success mHealth in transitions of care.

Conclusions

In this secondary analysis of patients enrolled in an mHealth intervention who experienced 30-day revisits, mHealth engagement was limited. mHealth users were more likely to be younger, commercially insured, and to interact with their practices by traditional modalities (telephone and clinic visits), raising concerns about equity of mHealth interventions and their potential to increase disparities. We also found revisits among mHealth users occurred farther from their index discharge and were considered more predictable, suggesting potential benefits of mHealth engagement. Further research is needed to better define meaningful engagement with mHealth, identify outcomes most important and relevant to mHealth use, and understand how to leverage mHealth to equitably engage patients during transitions of care.

Funding

The authors report research funding for this work from Optum, the research and development arm of UnitedHealth Group.

Data Availability

Data analyzed in this study are not publicly accessible due to inclusion of patient-level data from the electronic medical record.

Authors' Contributions

Conceptualization: SL, AM, AA, KP, AUM, EB
Formal analysis: AM
Investigation: SL, AA, KP
Methodology: SL, AM, AA, KP, EB
Writing – original draft: SL
Writing – review & editing: SL, AM, AA, KP, AUM, EB

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
STROBE checklist.
[DOCX File , 21 KB - jmir_v28i1e89067_app1.docx]

Multimedia Appendix 2
Chart review extraction template.
[DOCX File , 16 KB - jmir_v28i1e89067_app2.docx]

Multimedia Appendix 3
Table of intervisit interactions.
[DOCX File , 14 KB - jmir_v28i1e89067_app3.docx]

References

1. Plan all-cause readmissions. National Committee for Quality Assurance. URL: https://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality-report/plan-all-cause-readmissions-pcr/ [accessed 2024-08-26]
2. Hospital Readmissions Reduction Program. CMS. URL: https://www.cms.gov/medicare/payment/prospective-payment-systems/acute-inpatient-pps/hospital-readmissions-reduction-program-hrrp [accessed 2024-08-26]
3. Kumar V, Dolan RD, Yang AL, Jin DX, Banks PA, McNabb-Baltar J. Characteristics of 30-day all-cause hospital readmissions among patients with acute pancreatitis and substance use. Dig Dis Sci 2022 Dec;67(12):5500-5510 [FREE Full text] [doi: 10.1007/s10620-022-07463-2] [Medline: 35348968]
4. Horwitz LI, Moriarty JP, Chen C, Fogerty RL, Brewster UC, Kanade S, et al. Quality of discharge practices and patient understanding at an academic medical center. JAMA Intern Med 2013 Oct 14;173(18):1715-1722 [FREE Full text] [doi: 10.1001/jamainternmed.2013.9318] [Medline: 23958851]
5. Tyler N, Hodkinson A, Planner C, Angelakis I, Keyworth C, Hall A, et al. Transitional care interventions from hospital to community to reduce health care use and improve patient outcomes: a systematic review and network meta-analysis. JAMA Netw Open 2023 Nov 01;6(11):e2344825 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.44825] [Medline: 38032642]
6. Jayakody A, Bryant J, Carey M, Hobden B, Dodd N, Sanson-Fisher R. Effectiveness of interventions utilising telephone follow up in reducing hospital readmission within 30 days for individuals with chronic disease: a systematic review. BMC Health Serv Res 2016 Aug 18;16(1):403 [FREE Full text] [doi: 10.1186/s12913-016-1650-9] [Medline: 27538884]
7. mHealth: new horizons for health through mobile technologies: based on the findings of the second global survey on eHealth. World Health Organization. 2011. URL: https://iris.who.int/server/api/core/bitstreams/ad1b13c0-7c82-47b4-8dd5-f0a26c3a3cc3/content [accessed 2025-03-25]
8. Backman C, Papp R, Tonjock Kolle A, Papp S, Visintini S, Schaefer Ferreira de Mello AL, et al. Platform-based patient-clinician digital health interventions for care transitions: scoping review. J Med Internet Res 2024 Dec 30;26:e55753 [FREE Full text] [doi: 10.2196/55753] [Medline: 39753212]
9. Tan SY, Sumner J, Wang Y, Wenjun Yip A. A systematic review of the impacts of remote patient monitoring (RPM) interventions on safety, adherence, quality-of-life and cost-related outcomes. NPJ Digit Med 2024 Jul 18;7(1):192 [FREE Full text] [doi: 10.1038/s41746-024-01182-w] [Medline: 39025937]
10. Mashhadi SF, Hisam A, Sikander S, Rathore MA, Rifaq F, Khan SA, et al. Post discharge mHealth and teach-back communication effectiveness on hospital readmissions: a systematic review. Int J Environ Res Public Health 2021 Oct 04;18(19):10442 [FREE Full text] [doi: 10.3390/ijerph181910442] [Medline: 34639741]
11. Kitsiou S, Vatani H, Paré G, Gerber BS, Buchholz SW, Kansal MM, et al. Effectiveness of mobile health technology interventions for patients with heart failure: systematic review and meta-analysis. Can J Cardiol 2021 Aug;37(8):1248-1259. [doi: 10.1016/j.cjca.2021.02.015] [Medline: 33667616]
12. Johnson AE, Routh S, Taylor CN, Leopold M, Beatty K, McNamara DM, et al. Developing and implementing an mHealth heart failure self-care program to reduce readmissions: randomized controlled trial. JMIR Cardio 2022 Mar 21;6(1):e33286 [FREE Full text] [doi: 10.2196/33286] [Medline: 35311679]
13. Indraratna P, Biswas U, McVeigh J, Mamo A, Magdy J, Vickers D, et al. A smartphone-based model of care to support patients with cardiac disease transitioning from hospital to the community (TeleClinical Care): pilot randomized controlled trial. JMIR Mhealth Uhealth 2022 Feb 28;10(2):e32554 [FREE Full text] [doi: 10.2196/32554] [Medline: 35225819]
14. Pooni A, Brar MS, Anpalagan T, Schmocker S, Rashid S, Goldstein R, et al. Home to stay: a randomized controlled trial evaluating the effect of a postdischarge mobile app to reduce 30-day readmission following elective colorectal surgery. Ann Surg 2023 May 01;277(5):e1056-e1062. [doi: 10.1097/SLA.0000000000005527] [Medline: 35815882]

XSL•FO
RenderX

15. Bressman E, Long JA, Burke RE, Ahn A, Honig K, Zee J, et al. Automated text message-based program and use of acute health care resources after hospital discharge: a randomized clinical trial. JAMA Netw Open 2024 Apr 01;7(4):e243701 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.3701] [Medline: 38564221]
16. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007 Oct 20;335(7624):806-808 [FREE Full text] [doi: 10.1136/bmj.39335.541782.AD] [Medline: 17947786]
17. Chen Y, Kruahong S, Elias S, Turkson-Ocran RA, Commodore-Mensah Y, Koirala B, et al. Racial disparities in shared decision-making and the use of mHealth technology among adults with hypertension in the 2017-2020 health information National Trends Survey: cross-sectional study in the United States. J Med Internet Res 2023 Sep 13;25:e47566 [FREE Full text] [doi: 10.2196/47566] [Medline: 37703088]
18. Yao R, Zhang W, Evans R, Cao G, Rui T, Shen L. Inequities in health care services caused by the adoption of digital health technologies: scoping review. J Med Internet Res 2022 Mar 21;24(3):e34144 [FREE Full text] [doi: 10.2196/34144] [Medline: 35311682]
19. van Acker J, Maenhout L, Compernolle S. Older adults' user engagement with mobile health: a systematic review of qualitative and mixed-methods studies. Innov Aging 2023 Jan 30;7(2):igad007 [FREE Full text] [doi: 10.1093/geroni/igad007] [Medline: 37007638]
20. Gordon NP, Crouch E. Digital information technology use and patient preferences for internet-based health education modalities: cross-sectional survey study of middle-aged and older adults with chronic health conditions. JMIR Aging 2019 Apr 04;2(1):e12243 [FREE Full text] [doi: 10.2196/12243] [Medline: 31518291]
21. Rosen AK, Chen Q, Shin MH, O'Brien W, Shwartz M, Mull HJ, et al. Medical and surgical readmissions in the Veterans Health Administration: what proportion are related to the index hospitalization? Med Care 2014 Mar;52(3):243-249. [doi: 10.1097/MLR.0000000000000081] [Medline: 24374424]
22. Liu P, Astudillo K, Velez D, Kelley L, Cobbs-Lomax D, Spatz ES. Use of mobile health applications in low-income populations: a prospective study of facilitators and barriers. Circ Cardiovasc Qual Outcomes 2020 Sep;13(9):e007031 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.120.007031] [Medline: 32885681]
23. Brewer LC, Fortuna KL, Jones C, Walker R, Hayes SN, Patten CA, et al. Back to the future: achieving health equity through health informatics and digital health. JMIR Mhealth Uhealth 2020 Jan 14;8(1):e14512 [FREE Full text] [doi: 10.2196/14512] [Medline: 31934874]

Abbreviations

ED: emergency department
mHealth: mobile health
MORE-PC: Mobile Outreach to Reduce Emergencies-Primary Care
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Stone; submitted 05.Dec.2025; peer-reviewed by S Horman, K Khashu; comments to author 09.Mar.2026; accepted 04.May.2026; published 29.May.2026.

Please cite as:
Landon S, Mondal A, Ahn A, Profka K, Morgan AU, Bressman E
Understanding mHealth Engagement Among Patients With 30-Day Hospital Revisits: Secondary Analysis of a Randomized Clinical Trial
J Med Internet Res 2026;28:e89067
URL: https://www.jmir.org/2026/1/e89067
doi:10.2196/89067
PMID:

©Susan Landon, Angira Mondal, Aiden Ahn, Klea Profka, Anna U Morgan, Eric Bressman. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Smartphone Keystroke Biomarkers as Predictors of Adverse Neuropsychiatric Sequelae After Trauma in Trauma Survivors: Prospective Observational Cohort Study

Nicole A Short[1], PhD; Xinming An[2], PhD; Yinyao Ji[2], PhD; Qinghua Li[2], PhD; Thomas C Neylan[3,4], MD; Gari D Clifford[5], PhD; Stacey L House[6], MD, PhD; Francesca L Beaudoin[7], MD, PhD; Jennifer S Stevens[5], PhD; Sarah D Linnstaedt[2], PhD; Laura T Germine[8], PhD; John P Haran[9], MD, PhD; Alan B Storrow[10], MD; Christopher Lewandowski[11], MD; Paul I Musey Jr[12], MD; Phyllis L Hendry[13], MD; Sophia Sheikh[13], MD; Christopher W Jones[14], MD; Brittany E Punches[15], PhD; Jose L Pascual[16], MD, PhD; Mark J Seamon[16], MD; Erica Harris[17], MD; Claire Pearson[18], MD; Roland C Merchant[19], MD, MPH, SCD; Robert M Domeier[20], MD; Niels K Rathlev[21], MD; Brian J O'Neil[18], MD; Paulina Sergot[22], MD; Leon D Sanchez[19], MD, MPH; Steven E Bruce[23], PhD; Ronald C Kessler[24], PhD; Karestan C Koenen[24], PhD; Kerry J Ressler[8], MD, PhD; Samuel A McLean[2], MD, MPH

[1]Department of Psychology, University of Nevada, Las Vegas, Las Vegas, NV, United States
[2]Department of Anesthesiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[3]Department of Psychiatry, University of California, San Francisco, San Francisco, CA, United States
[4]Department of Neurology, University of California, San Francisco, San Francisco, CA, United States
[5]Department of Psychiatry and Behavioral Sciences, Emory University, Atlanta, GA, United States
[6]Department of Emergency Medicine, Washington University in St. Louis, St. Louis, MO, United States
[7]Department of Epidemiology, Brown University, Providence, RI, United States
[8]Department of Psychiatry, McLean Hospital, Belmont, MA, United States
[9]Department of Emergency Medicine, University of Massachusetts Chan Medical School, Worcester, MA, United States
[10]Department of Emergency Medicine, Vanderbilt University Medical Center, Nashville, TN, United States
[11]Department of Emergency Medicine, Henry Ford Health System, Detroit, MI, United States
[12]Department of Emergency Medicine, Indiana University School of Medicine, Indianapolis, IN, United States
[13]Department of Emergency Medicine, University of Florida, Jacksonville, FL, United States
[14]Department of Emergency Medicine, Cooper Medical School of Rowan University, Camden, SC, United States
[15]Department of Emergency Medicine, The Ohio State University, Columbus, OH, United States
[16]Department of Surgery, University of Pennsylvania, Philadelphia, PA, United States
[17]Department of Emergency Medicine, Einstein Medical Center Philadelphia, Philadelphia, PA, United States
[18]Department of Emergency Medicine, Wayne State University, Detroit, MI, United States
[19]Department of Emergency Medicine, Brigham and Women's Hospital, Boston, MA, United States
[20]Department of Emergency Medicine, Trinity Health Ann Arbor Hospital, Ypsilanti, MI, United States
[21]Department of Emergency Medicine, University of Massachusetts Chan Medical School, Springfield, MA, United States
[22]Department of Emergency Medicine, The University of Texas, Houston, TX, United States
[23]Department of Psychological Sciences, University of Missouri–St. Louis, St. Louis, MO, United States
[24]Department of Epidemiology, Harvard University, Boston, MA, United States

Corresponding Author:
Nicole A Short, PhD
Department of Psychology
University of Nevada, Las Vegas
Mail Stop 5030, 4505 South Maryland Parkway
Las Vegas, NV, 89154
United States
Phone: 1 702 895 0606
Email: nicole.short@unlv.edu

XSL•FO
RenderX

Abstract

Background: Adverse posttraumatic neuropsychiatric sequelae are common after trauma. Early identification of individuals at risk for these outcomes could enable the deployment of preventive interventions to survivors at greatest risk. Smartphone keystroke biomarkers show promise in identifying individuals with neuropsychiatric symptoms; however, to our knowledge, no research has examined whether they can be used to identify symptoms in the aftermath of trauma.

Objective: This study evaluates whether passively collected keystroke data from smartphone use in daily life could identify individuals with high symptom levels, as well as worsening or recovery of symptoms, after trauma exposure.

Methods: Data from a diverse cohort of individuals presenting to 27 emergency departments after trauma were analyzed. Inclusion criteria were presenting to the emergency department within 72 hours of trauma, age 18-75, and the ability to speak and read English. Exclusion criteria were solid organ injury, significant hemorrhage, operative intervention, or likely admission for over 72 hours. Participants installed an app that passively collected keystroke data during use of any app on their smartphone, beginning in the emergency department. Participants also completed serial symptom assessments over 8 weeks after trauma exposure.

Results: A total of 3445 patients met study criteria, provided informed consent, and completed assessments in the emergency department. Of these, 1072 (mean age 40, SD 13; 616/1072, 57.46%, women; 565/1072, 52.71% non–Hispanic Black) installed the app on their Android smartphone and completed the 8-week assessment and were therefore included in analyses. Keystroke biomarkers related to typing speed, identified using bivariate linear mixed models controlling for false discovery rates, were associated with elevated pain, reexperiencing, and mental fatigue (absolute values of rs=0.22-0.25, Ps=.02). Separate change-of-operation and scrolling keystroke biomarkers were associated with increased reexperiencing symptoms (r=0.18, *P*=.047) and mental fatigue (rs=0.18-0.19, Ps=.031-.047). Further, changes in specific keystroke biomarkers were associated with worsening or recovery of pain (rs=0.07-0.10, Ps=.02), somatic symptoms (rs=0.02, Ps=.02), mental fatigue (rs=0.02-0.04, Ps=.02), sleep disturbance (absolute rs=0.07-0.09, Ps=.02), reexperiencing (rs=0.02-0.04, Ps=.02), and hyperarousal (rs=0.02-0.04, Ps=.02).

Conclusions: In general, slower typing and scrolling speeds were associated with higher symptom levels, with small to medium effect sizes. Keystroke data passively collected via smartphone use may help identify individuals with significant or changing posttraumatic symptoms. Future research should continue to explore these keystroke biomarkers and whether they can be leveraged to connect vulnerable trauma survivors to appropriate services. Overall, these results add to the literature, indicating that passively collected keystroke data may help identify individuals with neuropsychiatric symptoms or changes and are, to our knowledge, the first to test whether keystroke biomarkers are useful in the aftermath of trauma. This represents a critical period during which preventive interventions could be deployed to reduce the long-term burden of trauma-related sequelae.

(J Med Internet Res 2026;28:e73771) doi:10.2196/73771

KEYWORDS

trauma; keystroke; posttraumatic stress; pain; somatic symptoms; sleep; mHealth

Introduction

Problem

Approximately 9 in 10 Americans experience at least one traumatic stressor during their lifetimes [1]. Most individuals recover naturally; however, adverse posttraumatic neuropsychiatric symptoms (APNSs) are common and morbid [2]. Common APNS domains include pain and other somatic symptoms; depression and anxiety; posttraumatic stress (PTS) symptoms such as reexperiencing, avoidance, and hyperarousal; as well as sleep disruption and nightmares. APNSs are associated with negative consequences, including emotional distress, functional impairments [3,4], and reduced quality of life [5,6].

The vast majority of individuals experience some level of pain or mental health symptoms, particularly PTS, in the immediate aftermath of trauma exposure [7-9]. For most, these symptoms resolve naturally on their own [7,9], whereas for others they become chronic and debilitating conditions [7]. Certain traumas, such as sexual assault, result in a higher risk for specific APNS [10,11]. Despite this, even after “minor” traumas, such as motor vehicle collision (MVC) without serious injury, a large proportion of survivors develop persistent APNSs. Up to half develop persistent pain [8,12-21], and approximately a quarter develop chronic PTS symptoms [22]. Unfortunately, many trauma survivors do not receive trauma-related services outside of emergency care [23]. This is particularly true for individuals who identify as racially minoritized or have lower levels of education and income [24], who are also at higher risk for trauma exposure [25-27]. However, early treatments for such disorders [28,29] could help avoid decades of suffering for trauma survivors, underscoring the importance of identifying at-risk survivors.

Review of Relevant Scholarship

Nearly 9 in 10 Americans own a smartphone, up from 4 in 10 just 10 years ago [30]. Apps can passively collect data in daily life, including user keystroke behavior. Keystroke patterns may help identify individuals’ risk for mental and physical health problems, including APNSs. Smartphone keystroke behavior data are particularly appealing because they do not require any additional device (eg, wearable), making them feasible to collect given the ubiquity of smartphone use. Research in this area is

XSL•FO
RenderX

in nascent stages but suggests that keystroke behavior data, alone [31-33] or in combination with other smartphone data [34], are useful for identifying increased depressive symptoms and cognitive functioning/impairments [35,36] among individuals with depression and bipolar disorder, as well as among healthy controls and military veterans [37]. Furthermore, keystroke dynamics can identify individuals with fine motor impairments [38]. This area of research has tremendous promise, as smartphone use is ubiquitous and could be leveraged to screen for and monitor APNSs in the aftermath of trauma, and to connect those in need to further assessment or care. Although studies have used other types of passively sensed data to detect PTS-related symptoms [39], no large-scale study, to our knowledge, has evaluated the predictive utility of smartphone keystroke behavior for APNSs after trauma.

Hypotheses, Aims, and Objectives

Given the potential of keystroke behavior data to identify and predict APNSs, the goal of this study was to use longitudinal keystroke data collected from an app installed on participants' smartphones after trauma exposure to predict APNS outcomes. Specifically, the aims were to test whether keystroke data collected from smartphone use in daily life in a socioeconomically and racially diverse population of trauma survivors presenting to emergency departments (EDs) after traumatic stress exposure can predict risk for, and recovery from or worsening of, APNSs. Individuals with lower socioeconomic status and those identifying with racially/ethnically minoritized groups are at relatively high risk for trauma exposure [25,27] and may experience difficulties accessing appropriate care after the ED visit [23], making them an important sample for targeting preventive interventions. Research Domain Criteria [40]–defined APNSs (ie, pain, reexperiencing, avoidance, hyperarousal, mental fatigue, sleep disturbance, somatic symptoms, nightmares, depression, anxiety) were assessed in the 8 weeks after trauma exposure (a critical period for the development of chronic symptoms) [7]. We hypothesized that it would be possible to derive and validate keystroke behaviors associated with each APNS outcome (ie, trait biomarkers stable over time), and that changes in such behaviors could accurately predict worsening of or recovery from APNSs over time (ie, state symptoms measured at the current assessment and subject to fluctuation) as well as trait symptoms (stable over time). To enhance rigor and replicability, journal article reporting standards were used in this study [41].

Methods

Study Design and Setting

Data for these analyses were obtained from the AURORA (Advancing Understanding of Recovery After Trauma) study, a prospective observational cohort study of a diverse sample of trauma survivors recruited from EDs across the United States in the early aftermath of trauma. Most EDs were part of academic medical centers and concentrated in the Northeast and Midwest of the United States. The full methodology of the AURORA study has been published elsewhere [2]. AURORA enrollment began in September 2017 and continued through June 30, 2020. Individuals were eligible to participate if they presented to 1 of 27 EDs within the national AURORA network. Participants were recruited from EDs and followed using a variety of remote follow-up procedures, including smartphone-based surveys that collected self-reported APNS data over 8 weeks after trauma exposure, and an app that collected keystroke data (SonderMind; formerly Mindstrong).

Participants and Sample Size

Participants were recruited if they presented to an ED within the AURORA network after experiencing qualifying traumatic events (ie, MVC, physical assault, sexual assault, fall >10 ft, or mass casualty incidents). Participants were included if they presented within 72 hours of the trauma, were aged 18-75, and were able to speak and read English. Individuals were excluded if they had a solid organ injury grade >1 per the American Association for the Surgery of Trauma, significant hemorrhage, required operative intervention, or were likely to be admitted for >72 hours. A total of 3445 patients met these criteria, provided informed consent, and completed ED assessments. Of the 3445 patients enrolled at baseline, 2626 remained enrolled for at least 67 days or completed 90% or more of the week 8 survey and did not become pregnant or incarcerated. Of the 2626 participants, 1072 were included in this analysis because they had an Android operating system on their device, had keystroke data, and completed APNS measures. Only Android users were included because iOS users could opt out of the passive collection of keyboard behaviors, resulting in a lack of keystroke data (of note, Android users could also opt out of the app portion of the study while still participating in the overall study, if desired). See Tables S1 and S2 in Multimedia Appendix 1 for a comparison of Android versus iOS users. In brief, there were no clinically significant differences between Android and iOS users; however, Android users were, on average, 7 years older, less likely to be women, more racially and ethnically diverse, less educated, more likely to have been married, and more likely to have experienced traumas other than an MVC. Purposive sampling was used to recruit trauma survivors presenting at EDs, and the sample size for the parent study was determined based on power analysis for its aims [2].

Participants were, on average, 40 (SD 13) years of age (Table 1), with a slight majority being women (616/1072, 57.46%). Most participants self-identified as non–Hispanic Black (565/1072, 52.71%), followed by non–Hispanic White (353/1072, 32.93%), Hispanic (114/1072, 10.63%), and other (34/1072, 3.17%). Most reported having some college education (457/1072, 42.63%), a total annual family income of less than US $35,000 (653/1072, 60.91%), and being single/never married (578/1072, 53.92%).

XSL•FO
RenderX

Table 1. Sample demographic and trauma characteristics of trauma survivors participating in an observational cohort study (N=1072).

Characteristics	Values
Age (years), mean (SD)	39.4 (12.6)
Female, n (%)	616 (57.46)
Total family income (US $), n (%)	
≤19K	348 (32.46)
19,000-35,000	305 (28.45)
35,001-50,000	129 (12.03)
>50,000	174 (16.23)
Race, n (%)	
Hispanic	114 (10.63)
Non–Hispanic White	353 (32.93)
Non–Hispanic Black	565 (53.71)
Non–Hispanic other	34 (3.17)
Education status, n (%)	
High school or less	422 (39.37)
Some college	457 (42.63)
College or more	189 (17.63)
Marital status, n (%)	
Married	245 (22.85)
Separated, divorced, widowed, or annulled	238 (22.20)
Never been married	578 (53.92)
Trauma type, n (%)	
Motor vehicle collision	772 (72.01)
Physical assault	118 (11.01)
Sexual assault	4 (0.37)
Fall	83 (7.74)
Nonmotorized collision	21 (1.96)
Animal-related	20 (1.87)
Other (including poisoning, burns, mass trauma exposure)	54 (5.04)

Assessments and Data Sources

APNS Data Collection and Preparation

Sociodemographic characteristics measured in the ED were assessed via survey items [2]. Following the ED visit, participants completed a rotating battery of smartphone-based questionnaires consisting of brief assessments of 10 common APNS domains: pain [42], depression and anxiety symptoms [43], sleep [44], nightmares [45], somatic symptoms [46], difficulty with concentration/thinking/fatigue (mental fatigue) [46], and reexperiencing, avoidance, and hyperarousal [47]. Each survey item was administered at 6 time points within the first 8 weeks after trauma exposure using the app. Survey items, and the study day on which each item was administered, are presented in Table S3 in Multimedia Appendix 1.

These survey items were used as indicator variables to develop measurement models for each APNS domain, and factor scores for each symptom were computed for each participant at every time point. Joint measurement models that included all 6 time points within the first 8 weeks after trauma exposure were developed to define each symptom domain. Temporal correlations among these indicator variables were introduced to improve model fit when temporal autocorrelations were not fully explained by the joint measurement model. Model fit indices (eg, Comparative Fit Index, Tucker-Lewis Index, standardized root mean square residual) were used to evaluate the fit of each measurement model [48].

Keystroke Data Collection

Longitudinal keystroke behavior data were collected from an app installed on participants' phones after trauma exposure (SonderMind; formerly Mindstrong). Keystroke data were collected from any smartphone use across apps and were not limited to study-specific interactions or the use of any specific app. Naturally occurring typing was captured, and users did not

XSL•FO
RenderX

receive prompts to type. Typing word content was also collected and is reported in a separate manuscript [24]. Several aspects of keystroke behavior were collected: typing speed (the time elapsed between typing one character [or making one action] and the next); change of operations (switching from one process [eg, typing characters] to another [eg, deleting]); and scrolling (in any context, the distance scrolled in a list before clicking on an item). The data were processed within the app; thus, only the features reported in this manuscript were available for analysis.

Ethical Considerations

The AURORA study protocol was approved by the Institutional Review Board (approval number 17-0703) at the University of North Carolina at Chapel Hill. All participants provided informed consent at the time of enrollment during their ED visit. All data reported are fully deidentified and anonymous. Participants were compensated for their involvement as follows: US $60 for the ED assessment, US $30 for app installation, and US $5 per serial assessment. No identification of individual participants or users in any images in the manuscript or multimedia appendices is possible. Shared data from the AURORA study are deidentified, with no patient identifiers included.

Data Analysis: Keystroke Feature Extraction

The passive data collection app collected 41 features derived from participants' interactions with their smartphones. Each feature had 23 signals, resulting in a total of 943 possible feature variables. Analyses began by pairing keystroke features with APNS data. Daily keystroke feature variables were merged with each of the 10 APNS constructs, whereby keystroke data collected on the same day or the day before the flash survey day were retained. The means of the keystroke feature variables across the 2 days were used to summarize the data in relation to the relevant APNS construct.

Second, within- and between-person correlations of keystroke features with each of the 10 APNS constructs were calculated. The top 50 variables with the highest absolute correlation values (either within- or between-persons) that were statistically significant (adjusted P value <.05, controlling for false discovery rate) were selected.

Third, the aggregated data were randomly divided into 2 equal parts: 1 for biomarker identification and 1 for validation. A bivariate linear mixed model approach was used to model the cross-sectional and longitudinal associations with each of the 10 APNS constructs. Feature variables that had significant associations (either cross-sectional or longitudinal) with any of the APNS constructs were then validated using the same bivariate linear mixed model in the remaining 50% of the data. Multiple testing was controlled by adjusting P values with a false discovery rate for the identification step and a Bonferroni correction for the validation step.

Fourth, we conducted an additional analysis for features that passed the identification and validation steps. For state biomarkers, we evaluated how accurately they predicted the direction of change in the corresponding APNS constructs (eg, worsening vs recovery).

Missing values for these 10 APNS constructs ranged from 250 out of 2438 (10.25%) to 285 out of 2554 (11.16%) during the first week after enrollment and from 1251 out of 2547 (49.12%) to 1479 out of 2438 (60.66%) around 6 months after enrollment. To explore the potential impact of missing data, we evaluated correlations between main outcomes (eg, posttraumatic stress disorder and pain) at different time points (eg, before trauma, 2 weeks, and 8 weeks after trauma) and completion rates of 4 main tasks (eg, watch wearing, flash survey, neurocognitive tests, and full surveys). All correlations were weak (<0.1; see Table S8 in Multimedia Appendix 1), suggesting that missing data were not driven by these outcomes. Thus, missing data for these APNS construct scores were considered missing at random and were imputed using the joint measurement model.

Of note, due to the large number of highly intracorrelated features associated with various APNSs, results were simplified as follows. For similar signal processing techniques (eg, measures of central tendency such as mean and median; first-, second-, and third-order spectral moments) that were highly intracorrelated with one another (rs>0.80), only the signal processing technique variable with the highest correlation with the relevant APNS was retained in the results and tables (eg, if both mean and median frequency were associated with pain, only mean frequency was retained as a predictor when it had the higher correlation with pain). Full results, including all variables, are available in Tables S4-S7 in Multimedia Appendix 1.

Results

Participant Flow and Sociodemographic, Trauma Exposure, and Clinical Characteristics

The final sample consisted of 1072 trauma survivors who used the passive data collection app on their Android smartphones (see Figure S1 in Multimedia Appendix 1). The majority of participants experienced an MVC as their trauma type (772/1072, 72.01%), followed by physical assault (118/1072, 11.01%), fall >10 ft (83/1072, 7.74%), other (54/1072, 5.04%; eg, poisoning, burn, mass trauma), nonmotorized collision (21/1072, 1.96%), animal-related (20/1072, 1.87%), and sexual assault (4/1072, 0.37%).

Derivation of Trait Biomarkers

First, we evaluated whether trait-level keystroke biomarkers could identify individuals experiencing high levels of specific APNS (Table 2) at the trait level (ie, symptoms stable over time). Within- and between-person correlations indicated that a total of 24 unique keystroke biomarkers related to typing speed were associated with pain (Table 2; absolute values of rs=0.22-0.25, Ps=.02; α level set at P<.05, applied to Bonferroni-adjusted P values for these and all analyses), indicating that, overall, slower character typing speed was associated with higher levels of pain. Additionally, 1 keystroke biomarker was associated with increased reexperiencing symptoms (ie, slower speed when typing the space bar; r=0.18, P=.047). Finally, 5 scroll-related biomarkers were associated with mental fatigue (rs=0.18-0.19, Ps=.031-.047), indicating that slower scrolling was associated with higher levels of mental

XSL•FO
RenderX

fatigue. Correlations between biomarkers and specific APNS were in the small- to medium-effect size range and were statistically significant at $P<.05$ after Bonferroni adjustment.

XSL•FO
RenderX

Table 2. Trait keystroke biomarkers captured via the app and associations with adverse posttraumatic neuropsychiatric sequelae among trauma survivors participating in an observational cohort study.

Construct and theme	Type	Signal processing	Correlation (*r*)	Bonferroni-adjusted *P* value
Pain				
Typing speed[a]	Two characters in a row	SD	0.22	.02
Typing speed	Two characters in a row	Mean log	0.24	.02
Typing speed	Two characters in a row	80th percentile	0.24	.02
Typing speed	Two characters in a row	Total (sum) power	–0.25	.02
Typing speed	Two characters in a row	First spectral moment	–0.25	.02
Typing speed	Two characters	Mean log	0.25	.02
Typing speed	Two characters	10th percentile	0.24	.02
Typing speed	Two characters	Total (sum) power	–0.25	.02
Typing speed	Two characters	Mean frequency	–0.25	.02
Typing speed	Two characters	First spectral moment	–0.25	.02
Typing speed	Two characters	Second spectral moment	–0.25	.02
Typing speed	Two characters	Mean log	0.24	.02
Typing speed	Two characters	50th percentile	0.25	.02
Typing speed	Two characters	Total (sum) power	–0.25	.02
Typing speed	Two characters	First spectral moment	–0.25	.02
Typing speed	Three characters or fewer in a row	Mean log	0.24	.02
Typing speed	Three characters or fewer in a row	50th percentile	0.24	.02
Typing speed	Three characters or fewer in a row	Total (sum) power	–0.25	.02
Typing speed	Three characters or fewer in a row	Second spectral moment	–0.25	.02
Typing speed	More than 3 characters in a row	Mean	0.24	.02
Typing speed	More than 3 characters in a row	Mean log	0.25	.02
Typing speed	More than 3 characters in a row	80th percentile	0.25	.02
Typing speed	More than 3 characters in a row	Total (sum) power	–0.26	.02
Typing speed	More than 3 characters in a row	First spectral moment	–0.26	.02
Reexperiencing				
Change of operations[b]	Character to space	Mean power	–0.18	.047
Mental fatigue				
Scroll[c]	Scroll then click	Mean	0.18	.047
Scroll	Scroll then click	Mean log	0.19	.03
Scroll	Scroll then click	*a* (intercept of linear regression of sorted time differences)	0.18	.047
Scroll	Scroll then click	Total (sum) power	–0.19	.03
Scroll	Scroll then click	Third spectral moment	–0.19	.03

[a]Typing speed is the length of time elapsed from typing 1 character (or making 1 action) to another (higher values=slower typing speed).

[b]Change of operations refers to switching from 1 process (eg, typing characters) to another (eg, deleting; thought to measure task or set shifting).

[c]Scroll refers to the distance scrolled in a list before clicking on the item they are looking for, or the speed in scrolling through a list (thought to measure processing speed). When the keystroke type description includes “characters in a row,” this refers to typing several characters in sequence that are not interrupted by a space key press, deletion, or other action. Correlation can be used as a measure of effect size for linear relationships, where a

XSL•FO
RenderX

positive/negative correlation of *r* between variables *X* and *Y* indicates that a 1 SD change in *X* is associated with *r* SD change in *Y* in the same/opposite direction.

Derivation of State Biomarkers

Second, we evaluated whether changes in keystroke behaviors were associated with worsening or recovery of state-level (ie, measured at the current assessment and subject to fluctuations) specific APNS during the initial 8 weeks after trauma using bivariate linear mixed models (Table 3). State-level biomarkers were identified for changes in pain, sleep disturbance, hyperarousal, reexperiencing, somatic symptoms, and mental fatigue. For pain, 18 separate keystroke-related biomarkers were identified (*r*s=0.07-0.10, *P*s=.02; 17 related to typing speed and 1 related to scrolling speed). Overall, slower typing and scrolling speeds were associated with higher levels of pain. For sleep disturbance, 11 state biomarkers related to typing speed were identified, with absolute *r* values ranging from 0.07 to 0.09 (*P*s=.02). Again, slower typing speed was associated with increased sleep disturbance. For somatic symptoms, 19 typing speed–related biomarkers were identified (*r*s=0.02), along with 4 biomarkers associated with mental fatigue (2 related to scrolling speed and 2 related to typing speed; *r*s=0.02-0.04). Overall, slower scrolling and typing speeds were associated with increased somatic symptoms and mental fatigue.

XSL•FO
RenderX

Table 3. State keystroke biomarkers captured via the app and associations with adverse posttraumatic neuropsychiatric sequelae among trauma survivors participating in an observational cohort study.

Construct and theme	Type	Signal processing	Correlation (*r*)	Bonferroni-adjusted *P* value
Pain				
Typing speed[a]	Two characters in a row	Total (sum) power	–0.07	.02
Typing speed	Two characters in a row	Third spectral moment	–0.09	.02
Typing speed	Two characters	Total (sum) power	–0.07	.02
Typing speed	Two characters	Mean frequency	–0.09	.02
Typing speed	Two characters	Third spectral moment	–0.08	.02
Typing speed	Two characters	Total (sum) frequency	–0.07	.02
Typing speed	Two characters	Mean frequency	–0.09	.02
Typing speed	Two characters	Third spectral moment	–0.09	.02
Typing speed	Consecutive typing events	Mean log	0.08	.02
Typing speed	Consecutive typing events	50th percentile	0.07	.02
Typing speed	Consecutive typing events	Total (sum) power	–0.09	.02
Typing speed	Consecutive typing events	Median frequency	–0.08	.02
Typing speed	Consecutive typing events	Mean frequency	–0.10	.02
Typing speed	Consecutive typing events	Third spectral moment	–0.10	.02
Scroll[b]	Scroll twice to item	Maximum fractal length	–0.09	.02
Typing speed	Three characters or fewer in a row	Mean frequency	–0.07	.02
Typing speed	More than 3 characters in a row	Mean frequency	–0.07	.02
Typing speed	More than 3 characters in a row	Third spectral moment	–0.07	.02
Sleep disturbance				
Typing speed	Two characters in a row	First spectral moment	–0.07	.02
Typing speed	Two characters	Mean log	0.07	.02
Typing speed	Two characters	Second spectral moment	–0.07	.02
Typing speed	Two characters	First spectral moment	–0.07	.02
Typing speed	Consecutive typing events	Mean log	0.08	.02
Typing speed	Consecutive typing events	Total (sum) power	–0.08	.02
Typing speed	Consecutive typing events	Mean frequency	–0.08	.02
Typing speed	Consecutive typing events	Second spectral moment	–0.08	.02
Typing speed	More than 3 characters in a row	Mean	0.08	.02
Typing speed	More than 3 characters in a row	80th percentile	0.08	.02
Typing speed	More than 3 characters in a row	Mean power	–0.09	.02
Hyperarousal				

XSL•FO
RenderX

Construct and theme	Type	Signal processing	Correlation (*r*)	Bonferroni-adjusted *P* value
Typing speed	Character to space	Second spectral moment	–0.07	.02
Typing speed	Two characters in a row	Mean frequency	–0.09	.02
Typing speed	Two characters in a row	First spectral moment	–0.10	.02
Typing speed	Two characters	First spectral moment	–0.09	.02
Typing speed	Consecutive typing events	Second spectral moment	–0.09	.02
Typing speed	More than 3 characters in a row	20th percentile	0.07	.02
Typing speed	More than 3 characters in a row	Total (sum) power	–0.08	.02
Typing speed	More than 3 characters in a row	First spectral moment	–0.08	.02
Typing speed	The second set of at least 5 characters in a row	10th percentile	0.07	.02
Typing speed	The second set of at least 5 characters in a row	80th percentile	0.08	.02
Typing speed	The second set of at least 5 characters in a row	Total (sum) power	–0.08	.02
Typing speed	The second set of at least 5 characters in a row	Mean frequency	–0.07	.02
Typing speed	The second set of at least 5 characters in a row	First spectral moment	–0.08	.02
Reexperiencing				
Typing speed	Character to space	90th percentile	0.07	.04
Typing speed	Character to space	Second spectral moment	–0.08	.02
Typing speed	Character to space	Median frequency	–0.07	.02
Typing speed	Consecutive typing events	Mean log	0.07	.02
Typing speed	Consecutive typing events	Third spectral moment	–0.09	.02
Typing speed	Three characters or fewer in a row	Total (sum) power	–0.09	.02
Typing speed	Four characters or fewer, then space	Mean frequency	–0.07	.02
Somatic symptoms				

XSL•FO
RenderX

Construct and theme	Type	Signal processing	Correlation (*r*)	Bonferroni-adjusted *P* value
Typing speed	Two characters in a row	Total (sum) power	–0.08	.02
Typing speed	Two characters in a row	Mean frequency	–0.08	.02
Typing speed	Two characters in a row	Second spectral moment	–0.08	.02
Typing speed	Two characters	Total (sum) power	–0.07	.02
Typing speed	Two characters	Mean frequency	–0.08	.02
Typing speed	Two characters	Third spectral moment	–0.08	.02
Typing speed	Two characters	Median frequency	–0.08	.02
Typing speed	Two characters	Third spectral moment	–0.09	.02
Typing speed	Consecutive typing events	80th percentile	0.07	.02
Typing speed	Consecutive typing events	Total (sum) power	–0.08	.02
Typing speed	Consecutive typing events	Mean frequency	–0.09	.02
Typing speed	Consecutive typing events	Second spectral moment	–0.08	.02
Typing speed	Three characters or fewer in a row	Total (sum) power	–0.06	.04
Typing speed	Three characters or fewer in a row	Second spectral moment	–0.07	.02
Typing speed	More than 3 characters in a row	Mean log	0.07	.02
Typing speed	More than 3 characters in a row	50th percentile	0.07	.02
Typing speed	More than 3 characters in a row	Total (sum) power	–0.09	.02
Typing speed	More than 3 characters in a row	Mean frequency	–0.10	.02
Typing speed	More than 3 characters in a row	Second spectral moment	–0.10	.02
Mental fatigue				
Scroll	Speed of scroll to item	50th percentile	0.06	.04
Scroll	Speed of scroll to item	Median frequency	–0.07	.02
Typing speed	More than 3 characters in a row	Mean frequency	–0.07	.02
Typing speed	More than 3 characters in a row	Third spectral moment	–0.07	.02

[a]Typing speed is the length of time elapsed from typing 1 character (or making 1 action) to another (higher values=slower typing speed).

[b]Scroll refers to the distance scrolled in a list before clicking on the item they are looking for, or the speed in scrolling through a list (thought to measure processing speed). When the keystroke type description includes "characters in a row," this refers to typing several characters in sequence that are not interrupted by a space key press, deletion, or other action. Correlation can be used as a measure of effect size for linear relationships, where a positive/negative correlation of *r* between variables *X* and *Y* indicates that a 1 SD change in *X* is associated with *r* SD change in *Y* in the same/opposite direction. Regarding PTS, 13 typing speed–related biomarkers were associated with state changes in hyperarousal symptoms (*r*s=.02-.04), and 7 typing speed–related biomarkers were associated with state changes in reexperiencing symptoms (*r*s=.02-.04). In general, slower typing speed was associated with elevated PTS. All correlations for state biomarkers were in the small effect size range and were statistically significant at $P<.05$ after Bonferroni adjustment.

Utility of State Biomarkers

Finally, we examined the potential utility of keystroke biomarkers as screening tools for worsening or improving APNSs after trauma using bivariate linear mixed models (Tables 4 and 5; only statistically significant biomarkers are presented). Worsening and improvement in self-report symptoms were defined as symptom severity at 6 months minus symptom severity in the first week >0 and <0, respectively. High positive predictive values for symptom recovery and negative predictive values for worsening suggest that simple keystroke measures, passively collected via smartphone, may have utility as initial screening tools for mental fatigue, pain, and somatic symptom outcomes among diverse trauma survivors.

Specifically, for identifying worsening PTS symptoms, we found that 11 typing speed–related biomarkers were associated with hyperarousal symptoms, and 7 typing speed–related biomarkers were associated with reexperiencing symptoms. A total of 2 scroll-related and 2 typing speed–related biomarkers were associated with worsening mental fatigue. For pain, 16 typing speed–related and 1 scroll-related biomarker were associated with worsening symptoms. Eleven typing speed–related biomarkers were predictive of worsening sleep. Finally, 19 typing speed–related biomarkers were predictive of worsening somatic symptoms.

XSL•FO
RenderX

Table 4. Prediction of worsening adverse posttraumatic neuropsychiatric symptoms using state keystroke biomarkers captured via the app among trauma survivors participating in an observational cohort study.

Construct and worsening, *n* (%)	Theme	Type	Signal processing	Sensitivity	Specificity	Positive predictive value	Negative predictive value
Hyperarousal (N=509)							
263 (51.7)	Typing speed[a]	Character to space	Second spectral moment	0.52	0.48	0.53	0.47
263 (51.7)	Typing speed	Two characters in a row	First spectral moment	0.51	0.49	0.52	0.49
263 (51.7)	Typing speed	Two characters	First spectral moment	0.5	0.5	0.52	0.48
263 (51.7)	Typing speed	Consecutive typing events	Second spectral moment	0.5	0.5	0.52	0.48
263 (51.7)	Typing speed	More than 3 characters in a row	20th percentile	0.51	0.47	0.51	0.47
263 (51.7)	Typing speed	More than 3 characters in a row	Total (sum) power	0.54	0.47	0.53	0.48
263 (51.7)	Typing speed	More than 3 characters in a row	First spectral moment	0.54	0.47	0.53	0.48
263 (51.7)	Typing speed	The second set of at least 5 characters in a row	80th percentile	0.53	0.52	0.59	0.46
263 (51.7)	Typing speed	The second set of at least 5 characters in a row	Total (sum) power	0.51	0.48	0.57	0.43
263 (51.7)	Typing speed	The second set of at least 5 characters in a row	Mean frequency	0.53	0.49	0.58	0.44
263 (51.7)	Typing speed	The second set of at least 5 characters in a row	First spectral moment	0.51	0.48	0.56	0.43
Mental fatigue (N=574)							
160 (27.9)	Scroll[b]	Speed of scroll to item	50th percentile	0.54	0.5	0.3	0.74
160 (27.9)	Scroll	Speed of scroll to item	Median frequency	0.56	0.5	0.3	0.75
160 (27.9)	Typing speed	More than 3 characters in a row	Mean frequency	0.49	0.56	0.3	0.74
160 (27.9)	Typing speed	More than 3 characters in a row	Third spectral moment	0.54	0.55	0.32	0.76
Pain (N=692)							
159 (23.0)	Typing speed	Two characters in a row	Total (sum) power	0.41	0.61	0.24	0.78
159 (23.0)	Typing speed	Two characters in a row	Third spectral moment	0.4	0.6	0.23	0.77
159 (23.0)	Typing speed	Two characters	Total (sum) power	0.39	0.59	0.22	0.77
159 (23.0)	Typing speed	Two characters	Mean frequency	0.36	0.65	0.23	0.77
159 (23.0)	Typing speed	Two characters	Third spectral moment	0.38	0.62	0.23	0.77
159 (23.0)	Typing speed	Two characters	Total (sum) power	0.43	0.62	0.25	0.79
159 (23.0)	Typing speed	Two characters	Mean frequency	0.36	0.62	0.22	0.76
159 (23.0)	Typing speed	Two characters	Third spectral moment	0.41	0.63	0.25	0.78
159 (23.0)	Typing speed	Consecutive typing events	Mean log	0.37	0.65	0.24	0.78
159 (23.0)	Typing speed	Consecutive typing events	50th percentile	0.36	0.64	0.23	0.77
159 (23.0)	Typing speed	Consecutive typing events	Total (sum) power	0.39	0.61	0.23	0.77
159 (23.0)	Typing speed	Consecutive typing events	Mean frequency	0.39	0.62	0.23	0.77
159 (23.0)	Typing speed	Consecutive typing events	Third spectral moment	0.44	0.61	0.25	0.79

XSL•FO
RenderX

Construct and worsening, *n* (%)	Theme	Type	Signal processing	Sensitivity	Specificity	Positive predictive value	Negative predictive value
159 (23.0)	Scroll	Scroll twice to item	Maximum fractal length	0.36	0.59	0.21	0.75
159 (23.0)	Typing speed	Three characters or fewer in a row	Mean frequency	0.44	0.58	0.24	0.77
159 (23.0)	Typing speed	More than 3 characters in a row	Mean frequency	0.42	0.55	0.22	0.77
159 (23.0)	Typing speed	More than 3 characters in a row	Third spectral moment	0.46	0.54	0.23	0.77
Reexperiencing (N=479)							
226 (47.2)	Typing speed	Character to space	90th percentile	0.53	0.48	0.46	0.55
226 (47.2)	Typing speed	Character to space	Second spectral moment	0.55	0.49	0.48	0.56
226 (47.2)	Typing speed	Character to space	Median frequency	0.55	0.54	0.5	0.59
226 (47.2)	Typing speed	Consecutive typing events	Mean log	0.51	0.46	0.46	0.52
226 (47.2)	Typing speed	Consecutive typing events	Third spectral moment	0.5	0.5	0.47	0.53
226 (47.2)	Typing speed	Three characters or fewer in a row	Total (sum) power	0.54	0.51	0.49	0.55
226 (47.2)	Typing speed	Four characters or fewer, then space	Mean frequency	0.56	0.47	0.48	0.56
Sleep (N=487)							
269 (55.2)	Typing speed	Two characters in a row	First spectral moment	0.51	0.49	0.55	0.45
269 (55.2)	Typing speed	Two characters	Mean log	0.51	0.5	0.55	0.45
269 (55.2)	Typing speed	Two characters	Second spectral frequency	0.5	0.48	0.54	0.44
269 (55.2)	Typing speed	Two characters	First spectral frequency	0.53	0.5	0.57	0.47
269 (55.2)	Typing speed	Consecutive typing events	Mean log	0.52	0.52	0.57	0.47
269 (55.2)	Typing speed	Consecutive typing events	Total (sum) power	0.51	0.51	0.56	0.46
269 (55.2)	Typing speed	Consecutive typing events	Mean frequency	0.52	0.51	0.57	0.46
269 (55.2)	Typing speed	Consecutive typing events	Second spectral moment	0.51	0.51	0.56	0.46
269 (55.2)	Typing speed	More than 3 characters in a row	Mean	0.52	0.48	0.56	0.44
269 (55.2)	Typing speed	More than 3 characters in a row	80th percentile	0.5	0.5	0.56	0.44
269 (55.2)	Typing speed	More than 3 characters in a row	Mean power	0.5	0.46	0.54	0.42
Somatic symptoms (N=572)							
130 (22.7)	Typing speed	Two characters in a row	Total (sum) power	0.46	0.54	0.23	0.77
130 (22.7)	Typing speed	Two characters in a row	Mean frequency	0.48	0.57	0.25	0.79
130 (22.7)	Typing speed	Two characters in a row	Second spectral moment	0.45	0.53	0.22	0.77
130 (22.7)	Typing speed	Two characters	Total (sum) power	0.48	0.52	0.23	0.78
130 (22.7)	Typing speed	Two characters	Mean frequency	0.42	0.54	0.21	0.76
130 (22.7)	Typing speed	Two characters	Third spectral moment	0.48	0.53	0.23	0.78
130 (22.7)	Typing speed	Two characters	Median frequency	0.45	0.54	0.22	0.77

XSL•FO
RenderX

Construct and worsening, *n* (%)	Theme	Type	Signal processing	Sensitivity	Specificity	Positive predictive value	Negative predictive value
130 (22.7)	Typing speed	Two characters	Third spectral moment	0.49	0.54	0.24	0.78
130 (22.7)	Typing speed	Consecutive typing events	80th percentile	0.47	0.54	0.23	0.78
130 (22.7)	Typing speed	Consecutive typing events	Total (sum) power	0.48	0.56	0.25	0.79
130 (22.7)	Typing speed	Consecutive typing events	Mean frequency	0.43	0.54	0.22	0.76
130 (22.7)	Typing speed	Consecutive typing events	Second spectral moment	0.47	0.57	0.24	0.79
130 (22.7)	Typing speed	Three characters or fewer in a row	Total (sum) power	0.49	0.56	0.25	0.79
130 (22.7)	Typing speed	Three characters or fewer in a row	Second spectral moment	0.5	0.55	0.25	0.79
130 (22.7)	Typing speed	More than 3 characters in a row	Mean log	0.54	0.55	0.26	0.8
130 (22.7)	Typing speed	More than 3 characters in a row	50th percentile	0.54	0.53	0.25	0.8
130 (22.7)	Typing speed	More than 3 characters in a row	Total (sum) power	0.53	0.54	0.26	0.8
130 (22.7)	Typing speed	More than 3 characters in a row	Mean frequency	0.52	0.57	0.26	0.8
130 (22.7)	Typing speed	More than 3 characters in a row	Second spectral moment	0.54	0.55	0.26	0.8

[a]Typing speed is the length of time elapsed from typing 1 character (or making 1 action) to another (higher values=slower typing speed).

[b]Scroll refers to the distance scrolled in a list before clicking on the item they are looking for, or the speed in scrolling through a list (thought to measure processing speed). When the keystroke type description includes “characters in a row,” this refers to typing several characters in sequence that are not interrupted by a space key press, deletion, or other action.

XSL•FO
RenderX

Table 5. Prediction of recovering adverse posttraumatic neuropsychiatric symptoms using state keystroke biomarkers captured via the app among trauma survivors participating in an observational cohort study.

Construct and recovering, n (%)	Theme	Type	Signal processing	Sensitivity	Specificity	Positive predictive value	Negative predictive value
Hyperarousal (N=509)							
246 (48.3)	Typing speed[a]	Character to space	Second spectral moment	0.52	0.48	0.47	0.53
246 (48.3)	Typing speed	Two characters in a row	First spectral moment	0.51	0.49	0.48	0.51
246 (48.3)	Typing speed	Two characters	First spectral moment	0.5	0.5	0.48	0.52
246 (48.3)	Typing speed	Consecutive typing events	Second spectral moment	0.5	0.5	0.48	0.52
246 (48.3)	Typing speed	More than 3 characters in a row	20th percentile	0.53	0.49	0.49	0.53
246 (48.3)	Typing speed	More than 3 characters in a row	Total (sum) power	0.53	0.46	0.47	0.52
246 (48.3)	Typing speed	More than 3 characters in a row	First spectral moment	0.53	0.46	0.47	0.52
246 (48.3)	Typing speed	The second set of at least 5 characters in a row	80th percentile	0.48	0.47	0.41	0.54
246 (48.3)	Typing speed	The second set of at least 5 characters in a row	Total (sum) power	0.52	0.49	0.43	0.57
246 (48.3)	Typing speed	The second set of at least 5 characters in a row	Mean frequency	0.51	0.47	0.42	0.56
246 (48.3)	Typing speed	The second set of at least 5 characters in a row	First spectral moment	0.52	0.49	0.44	0.57
Mental fatigue (N=573)							
413 (72.1)	Scroll[b]	Speed of scroll to item	50th percentile	0.5	0.46	0.7	0.26
413 (72.1)	Scroll	Speed of scroll to item	Median frequency	0.5	0.44	0.7	0.25
413 (72.1)	Typing speed	More than 3 characters in a row	Mean frequency	0.44	0.51	0.7	0.26
413 (72.1)	Typing speed	More than 3 characters in a row	Third spectral moment	0.45	0.46	0.68	0.24
Pain (N=692)							
533 (77.0)	Typing speed	Two characters in a row	Total (sum) power	0.39	0.59	0.76	0.22
533 (77.0)	Typing speed	Two characters in a row	Third spectral moment	0.4	0.6	0.77	0.23
533 (77.0)	Typing speed	Two characters	Total (sum) power	0.41	0.61	0.78	0.23
533 (77.0)	Typing speed	Two characters	Mean frequency	0.35	0.64	0.77	0.23
533 (77.0)	Typing speed	Two characters	Third spectral moment	0.38	0.62	0.77	0.23
533 (77.0)	Typing speed	Two characters	Total (sum) power	0.38	0.57	0.75	0.21
533 (77.0)	Typing speed	Two characters	Mean frequency	0.38	0.64	0.78	0.24
533 (77.0)	Typing speed	Two characters	Third spectral moment	0.37	0.59	0.75	0.22
533 (77.0)	Typing speed	Consecutive typing events	Mean log	0.35	0.63	0.76	0.23
533 (77.0)	Typing speed	Consecutive typing events	50th percentile	0.36	0.64	0.77	0.23
533 (77.0)	Typing speed	Consecutive typing events	Total (sum) power	0.39	0.61	0.77	0.23
533 (77.0)	Typing speed	Consecutive typing events	Mean frequency	0.38	0.61	0.77	0.23
533 (77.0)	Typing speed	Consecutive typing events	Third spectral moment	0.39	0.56	0.75	0.21
533 (77.0)	Scroll	Scroll twice to item	Maximum fractal length	0.41	0.64	0.79	0.25
533 (77.0)	Typing speed	Three characters or fewer in a row	Mean frequency	0.42	0.56	0.76	0.23
533 (77.0)	Typing speed	More than 3 characters in a row	Mean frequency	0.45	0.58	0.78	0.23
533 (77.0)	Typing speed	More than 3 characters in a row	Third spectral moment	0.46	0.54	0.77	0.23

XSL•FO
RenderX

Construct and recovering, n (%)	Theme	Type	Signal processing	Sensitivity	Specificity	Positive predictive value	Negative predictive value
Reexperiencing (N=479)							
253 (52.8)	Typing speed	Character to space	90th percentile	0.52	0.47	0.54	0.45
253 (52.8)	Typing speed	Character to space	Second spectral moment	0.51	0.45	0.52	0.44
253 (52.8)	Typing speed	Character to space	Median frequency	0.46	0.45	0.5	0.41
253 (52.8)	Typing speed	Consecutive typing events	Mean log	0.54	0.49	0.54	0.48
253 (52.8)	Typing speed	Consecutive typing events	Third spectral moment	0.5	0.5	0.53	0.47
253 (52.8)	Typing speed	Three characters or fewer in a row	Total (sum) power	0.49	0.46	0.51	0.45
253 (52.8)	Typing speed	Four characters or fewer, then space	Mean frequency	0.53	0.44	0.52	0.44
Sleep (N=487)							
218 (44.8)	Typing speed	Two characters in a row	First spectral moment	0.51	0.49	0.45	0.55
218 (44.8)	Typing speed	Two characters	Mean log	0.5	0.49	0.45	0.55
218 (44.8)	Typing speed	Two characters	Second spectral moment	0.52	0.5	0.46	0.56
218 (44.8)	Typing speed	Two characters	First spectral moment	0.5	0.47	0.43	0.53
218 (44.8)	Typing speed	Consecutive typing events	Mean log	0.48	0.48	0.43	0.53
218 (44.8)	Typing speed	Consecutive typing events	Total (sum) power	0.49	0.49	0.44	0.54
218 (44.8)	Typing speed	Consecutive typing events	Mean frequency	0.49	0.48	0.43	0.54
218 (44.8)	Typing speed	Consecutive typing events	Second spectral moment	0.49	0.49	0.44	0.54
218 (44.8)	Typing speed	More than 3 characters in a row	Mean	0.52	0.48	0.44	0.56
218 (44.8)	Typing speed	More than 3 characters in a row	80th percentile	0.5	0.5	0.44	0.56
218 (44.8)	Typing speed	More than 3 characters in a row	Mean power	0.54	0.5	0.46	0.58
Somatic symptoms (N=572)							

XSL•FO
RenderX

Construct and recovering, n (%)	Theme	Type	Signal processing	Sensitivity	Specificity	Positive predictive value	Negative predictive value
442 (77.3)	Typing speed	Two characters in a row	Total (sum) power	0.46	0.54	0.77	0.23
442 (77.3)	Typing speed	Two characters in a row	Mean frequency	0.43	0.52	0.75	0.21
442 (77.3)	Typing speed	Two characters in a row	Second spectral moment	0.47	0.55	0.78	0.23
442 (77.3)	Typing speed	Two characters	Total (sum) power	0.48	0.52	0.77	0.22
442 (77.3)	Typing speed	Two characters	Mean frequency	0.46	0.58	0.79	0.24
442 (77.3)	Typing speed	Two characters	Third spectral moment	0.47	0.52	0.77	0.22
442 (77.3)	Typing speed	Two characters	Median frequency	0.46	0.55	0.78	0.23
442 (77.3)	Typing speed	Two characters	Third spectral moment	0.46	0.51	0.76	0.22
442 (77.3)	Typing speed	Consecutive typing events	80th percentile	0.46	0.53	0.77	0.22
442 (77.3)	Typing speed	Consecutive typing events	Total (sum) power	0.44	0.52	0.75	0.21
442 (77.3)	Typing speed	Consecutive typing events	Mean frequency	0.46	0.57	0.78	0.24
442 (77.3)	Typing speed	Consecutive typing events	Second spectral moment	0.43	0.53	0.76	0.21
442 (77.3)	Typing speed	Three characters or fewer in a row	Total (sum) power	0.44	0.51	0.75	0.21
442 (77.3)	Typing speed	Three characters or fewer in a row	Second spectral moment	0.45	0.5	0.75	0.21
442 (77.3)	Typing speed	Fewer than 3 characters in a row	Mean log	0.45	0.46	0.74	0.2
442 (77.3)	Typing speed	Fewer than 3 characters in a row	50th percentile	0.47	0.46	0.75	0.2
442 (77.3)	Typing speed	Fewer than 3 characters in a row	Total (sum) power	0.46	0.47	0.74	0.2
442 (77.3)	Typing speed	Fewer than 3 characters in a row	Mean frequency	0.43	0.48	0.74	0.2
442 (77.3)	Typing speed	Fewer than 3 characters in a row	Second spectral moment	0.45	0.46	0.74	0.2

[a]Typing speed is the length of time elapsed from typing 1 character (or making 1 action) to another (higher values=slower typing speed).

[b]Scroll refers to the distance scrolled in a list before clicking on the item they are looking for, or the speed in scrolling through a list (thought to measure processing speed). When the keystroke type description includes “characters in a row,” this refers to typing several characters in sequence that are not interrupted by a space key press, deletion, or other action.

For identifying recovering PTS symptoms, 11 typing speed–related biomarkers were predictive of hyperarousal, and 7 typing speed–related biomarkers were associated with reexperiencing symptoms. Similarly, 2 scroll-related and 2 typing speed–related biomarkers were associated with recovery from mental fatigue. A total of 17 biomarkers (16 typing speed–related and 1 scroll-related) were predictive of recovery from pain. For recovery from sleep disturbance, there were 11 typing speed–related biomarkers. Finally, there were 19 typing speed–related biomarkers associated with recovery from somatic symptoms.

Discussion

Principal Findings

Millions of Americans present to emergency care each year following exposure to a traumatic stressor [49]. Although the majority recover from these events, a significant minority go on to develop chronic APNSs, including PTS, depression, anxiety, sleep disturbance, nightmares, pain, somatic symptoms, and mental fatigue [2]. The ability to identify these individuals and, ultimately, intervene to prevent the development of chronic APNSs would advance public health and reduce substantial personal suffering. Approximately 90% of Americans own a smartphone [30], and data from smartphone use can be collected passively in daily life with minimal effort from the user. This study is, to our knowledge, the first to suggest that such smartphone data may serve as biomarkers for the development of chronic APNSs in the 8 weeks after traumatic event exposure and presentation to emergency care. Further, keystroke data, which index cognitive and motor behavior, may be useful for tracking worsening and recovery of APNSs. Specifically, smartphone data measuring typing and scrolling behaviors during daily life among 1072 trauma survivors who presented for emergency care helped identify those with high levels of pain, reexperiencing, and mental fatigue in the 8 weeks after trauma exposure. In addition, smartphone indicators predicted state changes in pain, somatic symptoms, mental fatigue, sleep disturbance, and PTS (ie, reexperiencing and hyperarousal), including identifying individuals with worsening versus recovering symptoms. Effects were typically in the small- to medium-effect size range. Importantly, the sample was diverse, with more than half identifying as Black and socioeconomically disadvantaged, with a third reporting a total annual family income of less than US $19,000. This contrasts with other studies in the literature using smartphone biomarkers, which tend to include predominantly White, highly educated samples with higher socioeconomic status [33,50].

XSL•FO
RenderX

Comparison With Prior Work

Regarding trait biomarkers, our results suggest that 18 keystroke biomarkers could identify individuals with elevated pain in the 8 weeks after trauma exposure. Specifically, individuals with slower typing speeds demonstrated higher levels of pain. Frequency-domain features (eg, total [sum] power, mean frequency, and spectral moments) were the most common signal processing techniques observed in typing speed–related biomarkers, suggesting that these techniques may be particularly useful. Change-of-operation behaviors (eg, time spent switching from typing to pressing the space bar) were associated with reexperiencing symptoms. This is consistent with prior research indicating that PTS symptoms may be associated with difficulties in set shifting [51,52]; however, this is the first study, to our knowledge, to extend such findings to passively collected smartphone data. Finally, scrolling behaviors were associated with mental fatigue, such that longer time spent scrolling to identify desired items was associated with higher levels of mental fatigue. Effects were all in the small- to medium-effect size range. These results are the first to suggest that keystroke biomarkers can identify trauma survivors with pain, PTS symptoms, and mental fatigue. However, the findings are consistent with prior studies in other populations [33], suggesting that keystroke behaviors, including slower typing speeds, can identify individuals with depressive symptoms [37], cognitive impairment [35,36], and pain and clinical disability [50]. These results also align with research indicating that other passively sensed data (eg, sleep and physiological metrics) may serve as potential biomarkers of posttraumatic stress disorder symptoms [39].

State keystroke biomarkers were also identified that could detect within-person changes in APNSs, suggesting improvement or worsening of symptoms. Changes in physical APNSs, such as pain, somatic symptoms, mental fatigue, and sleep disturbance, were predicted by typing and scrolling speed. Changes in mental health–related APNSs (ie, reexperiencing, hyperarousal) were predicted by typing speed. These findings are consistent with prior research indicating that changes in keystroke behaviors can predict within-person changes in mood [31-34]. However, the study findings are, to our knowledge, the first to suggest that keystroke biomarkers can predict worsening or recovery of APNSs among a large sample of trauma survivors followed over the acute posttrauma period.

This study has important clinical implications. Although the sensitivity and specificity of the results indicate that keystroke biomarkers may not be sufficient on their own, they could be useful as part of a suite of heterogeneous biomarkers. The ability to identify individuals at risk for poor outcomes after trauma exposure by leveraging passively collected smartphone data in daily life could facilitate early identification of those at risk for developing APNSs. In turn, such individuals could be connected with appropriate health care services before their symptoms become chronic. More research is needed to determine the most effective preventive interventions after trauma exposure, particularly given the diverse range of potential negative outcomes. However, research has indicated that cognitive behavioral preventive interventions delivered in the early aftermath of trauma may be promising for reducing PTS and related symptoms [28]. Utilizing keystroke biomarkers could also assist in clinical research by identifying and recruiting individuals at risk for APNSs after trauma exposure. Furthermore, the ability to detect state changes in APNSs is particularly valuable. Pending further research, Figure 1 depicts an example of a clinical tool that could be developed to leverage state data from passively collected biomarkers to complement assessments conducted by providers during clinical visits. Ongoing passive data collection indicating state changes in APNSs may ultimately be informative for health care providers and researchers for several reasons. First, providers are often only able to assess changes between appointments, which may occur at long intervals. Second, APNSs are often difficult or impossible to assess “objectively.” Third, APNSs are challenging to report retrospectively with accuracy [53]. Data derived from keystroke biomarkers that capture trajectories of APNSs over time could enhance assessment and improve the ability to evaluate treatment efficacy.

XSL•FO
RenderX

Figure 1. An example of a clinical tool that could be developed to leverage state data from passively collected biomarkers to complement assessments conducted by providers during clinical visits. mo: month; w: week.

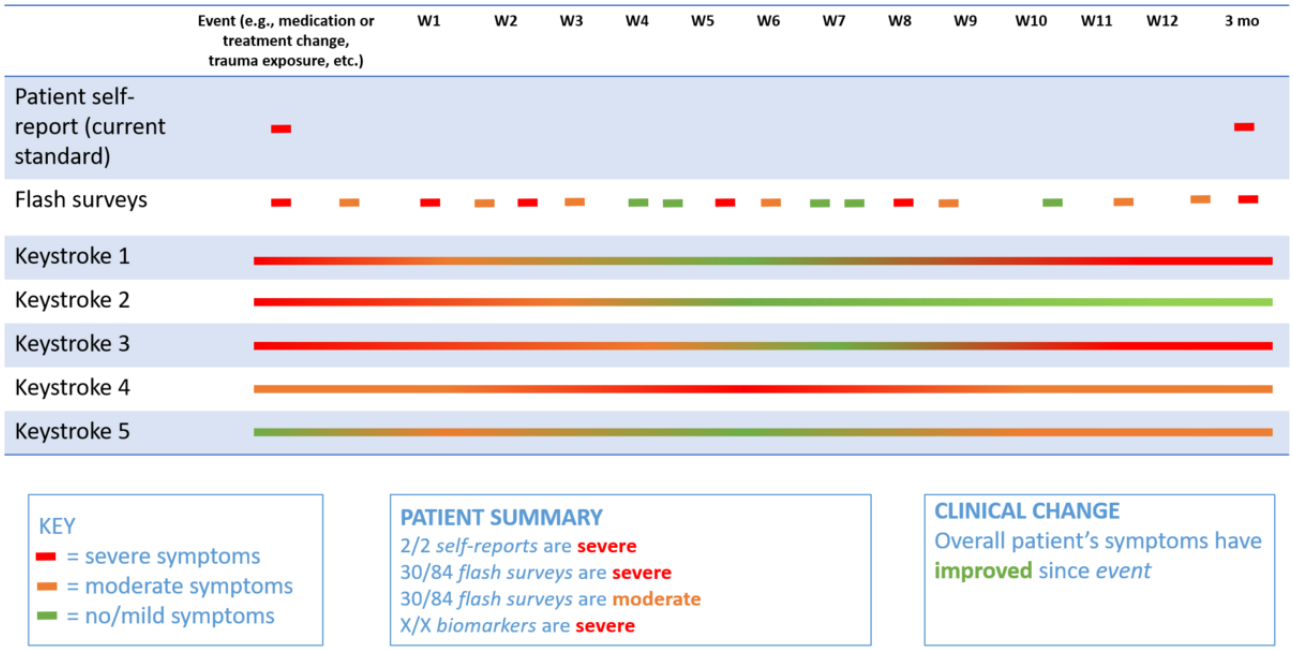

Strengths and Limitations

Results of this study should be considered in the context of its strengths and limitations. Strengths include the large sample size of diverse trauma survivors presenting to EDs who were followed prospectively after trauma exposure. Regarding limitations, first, only individuals who presented for emergency care after trauma exposure were included. Many trauma survivors do not seek emergency care, and it is unclear whether these findings generalize to those populations. Second, the majority of the sample consisted of MVC survivors. This is consistent with rates of trauma exposure in the United States, where MVC is one of the most common traumas experienced [1]. However, compared with other trauma types, MVCs may confer a relatively lower risk of developing APNSs [10]. Therefore, results may not generalize to survivors of all trauma types. Third, only Android users were included because iOS users could opt out of keystroke data collection. Although we did not detect differences in clinical symptoms between Android and iOS users, restricting the sample to Android users may affect generalizability. Notably, demographic differences were observed between iOS and Android users; however, these differences suggested that Android users represented a more racially and ethnically diverse sample, which may enhance generalizability. Relatedly, an important consideration in digital phenotyping research is the balance between patient privacy and the potential clinical utility of passively sensed smartphone data. Although substantially more research is needed before keystroke biomarkers can be implemented in clinical settings, further work is also required to ensure the development of secure and trustworthy apps that adequately safeguard private information. Even with such safeguards, some individuals may choose not to use these apps due to privacy concerns. Ethical considerations surrounding these data sources remain ongoing, and the balance of risks and benefits should be discussed with all patients and participants who may use such technologies. Fourth, participants were followed for 8 weeks after trauma exposure to assess APNS outcomes. Research indicates that by this point in the posttrauma period, most individuals with ongoing APNSs will continue to experience chronic symptoms unless they receive treatment [7]. However, future research may benefit from longer follow-up periods after a traumatic event.

Conclusions

Up to 90% of Americans experience a traumatic event at some point in their lives [1], and many go on to develop chronic APNSs [2]. To improve recovery after trauma, it is critical to identify individuals at risk of developing these symptoms. Our findings among 1072 diverse trauma survivors presenting to EDs nationwide indicate the potential utility of passively collected smartphone keystroke data to identify both individuals at risk for developing chronic pain, reexperiencing, and mental fatigue symptoms, as well as state-level worsening or recovery of pain, sleep disturbance, hyperarousal, reexperiencing, somatic symptoms, and mental fatigue in the 8 weeks after trauma exposure. In general, slower typing and scrolling speeds were associated with higher symptom levels, with small- to medium-effect sizes. Keystroke data passively collected via smartphone use in daily life show promise and warrant further research to determine whether they may serve as useful biomarkers for identifying trauma survivors at risk for APNSs. Future research should continue to explore associations between these keystroke biomarkers and APNSs, and assess whether they can be leveraged to connect vulnerable trauma survivors to appropriate health care services, preventive interventions, or both. Overall, these results add to the literature across other conditions indicating that passively collected keystroke data may help identify individuals with neuropsychiatric symptoms or changes. However, to our knowledge, this is the first study to test whether keystroke biomarkers are useful in identifying individuals at risk for neuropsychiatric symptoms in the aftermath of trauma—a critical period during which preventive

XSL•FO
RenderX

interventions could be deployed to reduce the long-term burden of trauma-related sequelae.

Acknowledgments

The investigators thank the trauma survivors participating in the AURORA Study. Their time and effort during a challenging period of their lives make efforts to improve recovery for future trauma survivors possible.

Generative artificial intelligence was not used in any portion of the manuscript writing.

Funding

This project was supported by the National Institute of Mental Health (NIMH) under U01MH110925, the US Army MRMC, One Mind, and The Mayday Fund. The content is solely the responsibility of the authors and does not necessarily represent the official views of any of the funders. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or of the submitters who contributed original data to NDA.

Data Availability

Data used in this manuscript are available through the National Institute of Mental Health (NIMH) Data Archive (NDA). The NDA collection for the AURORA (Advancing Understanding of Recovery After Trauma) Project can be found online [54].

Conflicts of Interest

TCN has received research support from the National Institutes of Health (NIH), United States Department of Veterans Affairs (VA), and the Rainwater Charitable Foundation, and consulting income from Otsuka Pharmaceuticals. In the past three years, GDC has received research funding from the National Science Foundation (NSF), the National Institutes of Health, and LifeBell AI, and unrestricted donations from AliveCor Inc, Amazon Research, the Center for Discovery, the Gates Foundation, Google, the Gordon and Betty Moore Foundation, MathWorks, Microsoft Research, NextSense Inc, One Mind Foundation, and the Rett Research Foundation. GDC has a financial interest in AliveCor Inc and NextSense Inc; is also the Chief Technology Officer (CTO) of MindChild Medical; and holds significant stock (all these relationships are unrelated to the current work). LTG receives funding from the National Institute of Mental Health (grant R01 MH121617) and serves on the board of the Many Brains Project; her family also has equity in Intelerad Medical Systems, Inc. CWJ has no competing interests related to this work, although he has been an investigator on studies funded by AstraZeneca, Vapotherm, Abbott, and Ophirex. JLP is president-elect of the Society of Critical Care Medicine. In the past three years, RCK has served as a consultant for Cambridge Health Alliance, Canandaigua VA Medical Center, Holmusk, Partners Healthcare, Inc., RallyPoint Networks, Inc., and Sage Therapeutics; and also holds stock options in Cerebral Inc., Mirah, PYM, and Roga Sciences. KJR has provided scientific consultation for Bioxcel, Bionomics, Acer, and Jazz Pharmaceuticals; serves on Scientific Advisory Boards for Sage, Boehringer Ingelheim, Senseye, and the Brain Research Foundation; and has received sponsored research support from Alto Neuroscience. KCK has served as a paid scientific consultant for the US Department of Justice and Covington & Burling, LLP over the past three years; and receives royalties from Guilford Press and Oxford University Press. SAM has served as a consultant for Walter Reed Army Institute for Research, Arbor Medical Innovations, and BioXcel Therapeutics, Inc.

Multimedia Appendix 1
Additional analysis.
[DOCX File , 513 KB - jmir_v28i1e73771_app1.docx]

References

1. Kilpatrick DG, Resnick HS, Milanak ME, Miller MW, Keyes KM, Friedman MJ. National estimates of exposure to traumatic events and PTSD prevalence using DSM-IV and DSM-5 criteria. J Trauma Stress 2013 Oct 22;26(5):537-547 [FREE Full text] [doi: 10.1002/jts.21848] [Medline: 24151000]
2. McLean S, Ressler K, Koenen K, Neylan T, Germine L, Jovanovic T, et al. The AURORA Study: a longitudinal, multimodal library of brain biology and function after traumatic stress exposure. Mol Psychiatry 2020 Feb;25(2):283-296 [FREE Full text] [doi: 10.1038/s41380-019-0581-3] [Medline: 31745239]
3. Momartin S, Silove D, Manicavasagar V, Steel Z. Comorbidity of PTSD and depression: associations with trauma exposure, symptom severity and functional impairment in Bosnian refugees resettled in Australia. Journal of Affective Disorders 2004 Jun;80(2-3):231-238. [doi: 10.1016/s0165-0327(03)00131-9]
4. Dobie DJ, Kivlahan DR, Maynard C, Bush KR, Davis TM, Bradley KA. Posttraumatic stress disorder in female veterans: association with self-reported health problems and functional impairment. Arch Intern Med 2004 Feb 23;164(4):394-400. [doi: 10.1001/archinte.164.4.394] [Medline: 14980990]

XSL•FO
RenderX

5. Clapp J, Beck J, Palyo S, Grant D. An examination of the synergy of pain and PTSD on quality of life: additive or multiplicative effects? Pain 2008 Aug 31;138(2):301-309 [FREE Full text] [doi: 10.1016/j.pain.2008.01.001] [Medline: 18280044]
6. Gormsen L, Rosenberg R, Bach FW, Jensen TS. Depression, anxiety, health-related quality of life and pain in patients with chronic fibromyalgia and neuropathic pain. Eur J Pain 2010 Feb 16;14(2):127.e1-127.e8. [doi: 10.1016/j.ejpain.2009.03.010] [Medline: 19473857]
7. Rothbaum BO, Foa EB, Riggs DS, Murdock T, Walsh W. A prospective examination of post-traumatic stress disorder in rape victims. J Trauma Stress 1992 Jul;5(3):455-475. [doi: 10.1007/bf00977239]
8. Bortsov A, Platts-Mills TF, Peak D, Jones J, Swor R, Domeier R, et al. Pain distribution and predictors of widespread pain in the immediate aftermath of motor vehicle collision. Eur J Pain 2013 Sep 20;17(8):1243-1251 [FREE Full text] [doi: 10.1002/j.1532-2149.2013.00285.x] [Medline: 23335385]
9. Vollmer Dahlke D, Fair K, Hong YA, Beaudoin CE, Pulczinski J, Ory MG. Apps seeking theories: results of a study on the use of health behavior change theories in cancer survivorship mobile apps. JMIR Mhealth Uhealth 2015 Mar 27;3(1):e31 [FREE Full text] [doi: 10.2196/mhealth.3861] [Medline: 25830810]
10. Smith HL, Summers BJ, Dillon KH, Cougle JR. Is worst-event trauma type related to PTSD symptom presentation and associated features? J Anxiety Disord 2016 Mar;38:55-61. [doi: 10.1016/j.janxdis.2016.01.007] [Medline: 26826984]
11. Kessler RC, Aguilar-Gaxiola S, Alonso J, Benjet C, Bromet EJ, Cardoso G, et al. Trauma and PTSD in the WHO World Mental Health Surveys. Eur J Psychotraumatol 2017 Jul 31;8(sup5):1353383 [FREE Full text] [doi: 10.1080/20008198.2017.1353383] [Medline: 29075426]
12. Holm LW, Carroll LJ, Cassidy JD, Skillgate E, Ahlbom A. Widespread pain following whiplash-associated disorders: incidence, course, and risk factors. J Rheumatol 2007 Jan;34(1):193-200. [Medline: 17143964]
13. Berglund A, Alfredsson L, Jensen I, Cassidy J, Nygren. The association between exposure to a rear-end collision and future health complaints. J Clin Epidemiol 2001 Aug;54(8):851-856. [doi: 10.1016/s0895-4356(00)00369-3] [Medline: 11470396]
14. Carroll LJ, Holm LW, Hogg-Johnson S, Côté P, Cassidy JD, Haldeman S, et al. Course and prognostic factors for neck pain in whiplash-associated disorders (WAD). Spine 2008;33(Supplement):S83-S92. [doi: 10.1097/brs.0b013e3181643eb8]
15. Cassidy JD, Carroll L, Côté P, Berglund A, Nygren. Low back pain after traffic collisions. Spine 2003;28(10):1002-1009. [doi: 10.1097/01.brs.0000061983.36544.0d]
16. Côté P, Cassidy JD, Carroll L. The factors associated with neck pain and its related disability in the Saskatchewan population. Spine (Phila Pa 1976) 2000 May 01;25(9):1109-1117. [doi: 10.1097/00007632-200005010-00012] [Medline: 10788856]
17. Hincapié CA, Cassidy J, Côté P, Carroll L, Guzmán J. Whiplash injury is more than neck pain: a population-based study of pain localization after traffic injury. J Occup Environ Med 2010 Apr;52(4):434-440. [doi: 10.1097/JOM.0b013e3181bb806d] [Medline: 20357684]
18. Kasch H, Bach FW, Jensen TS. Handicap after acute whiplash injury: a 1-year prospective study of risk factors. Neurology 2001 Jun 26;56(12):1637-1643. [doi: 10.1212/wnl.56.12.1637] [Medline: 11425927]
19. Sterling M, Jull G, Vicenzino B, Kenardy J, Darnell R. Physical and psychological factors predict outcome following whiplash injury. Pain 2005 Mar;114(1-2):141-148. [doi: 10.1016/j.pain.2004.12.005] [Medline: 15733639]
20. Pilet C, Galinski M, Lafont S. Chronic pain prevalence two years after a road crash and its biopsychosocial risk factors - results from the ESPARR cohort. Journal of Transport & Health 2023 May;30:101613. [doi: 10.1016/j.jth.2023.101613]
21. Jones GT, Nicholl BI, McBeth J, Davies KA, Morriss RK, Dickens C, et al. Role of road traffic accidents and other traumatic events in the onset of chronic widespread pain: results from a population-based prospective study. Arthritis Care Res (Hoboken) 2011 May 06;63(5):696-701 [FREE Full text] [doi: 10.1002/acr.20417] [Medline: 21425478]
22. Vollmer Dahlke D, Fair K, Hong YA, Beaudoin CE, Pulczinski J, Ory MG. Apps seeking theories: results of a study on the use of health behavior change theories in cancer survivorship mobile apps. JMIR Mhealth Uhealth 2015 Mar 27;3(1):e31 [FREE Full text] [doi: 10.2196/mhealth.3861] [Medline: 25830810]
23. Short NA, Lechner M, McLean BS, Tungate AS, Black J, Buchanan JA, et al. Health care utilization by women sexual assault survivors after emergency care: results of a multisite prospective study. Depress Anxiety 2021 Jan 08;38(1):67-78 [FREE Full text] [doi: 10.1002/da.23102] [Medline: 33032388]
24. Vizer L, Pierce J, Ji Y, Bucher MA, Liu M, Ungar L, et al. Smartphone language features may help identify adverse post-traumatic neuropsychiatric sequelae and their trajectories. NPP Digit Psychiatry Neurosci 2025 May 20;3(1):8. [doi: 10.1038/s44277-025-00028-x] [Medline: 40406207]
25. López CM, Andrews AR, Chisolm AM, de Arellano MA, Saunders B, Kilpatrick DG. Racial/ethnic differences in trauma exposure and mental health disorders in adolescents. Cultural Diversity & Ethnic Minority Psychology 2017 Jul;23(3):382-387 [FREE Full text] [doi: 10.1037/cdp0000126] [Medline: 27786496]
26. Carbone JT, Holzer KJ, Vaughn MG. Posttraumatic stress disorder among low-income adolescents experiencing family-neighborhood income disparities. J Trauma Stress 2019 Dec 17;32(6):899-907. [doi: 10.1002/jts.22452] [Medline: 31623017]
27. McLaughlin KA, Alvarez K, Fillbrunn M, Green JG, Jackson JS, Kessler RC, et al. Racial/ethnic variation in trauma-related psychopathology in the United States: a population-based study. Psychol Med 2018 Oct 31;49(13):2215-2226. [doi: 10.1017/s0033291718003082]

XSL•FO
RenderX

28. Short NA, Morabito DM, Gilmore AK. Secondary prevention for posttraumatic stress and related symptoms among women whohave experienced a recent sexual assault: a systematic review and meta-analysis. Depress Anxiety 2020 Oct 22;37(10):1047-1059 [FREE Full text] [doi: 10.1002/da.23030] [Medline: 32442345]
29. Bryant RA, Moulds ML, Nixon RV. Cognitive behaviour therapy of acute stress disorder: a four-year follow-up. Behav Res Ther 2003 Apr;41(4):489-494. [doi: 10.1016/s0005-7967(02)00179-1] [Medline: 12643970]
30. Mobile fact sheet. Pew Research Center. 2021. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2025-07-01]
31. Vesel C, Rashidisabet H, Zulueta J, Stange J, Duffecy J, Hussain F, et al. Effects of mood and aging on keystroke dynamics metadata and their diurnal patterns in a large open-science sample: a BiAffect iOS study. J Am Med Inform Assoc 2020 Jul 01;27(7):1007-1018 [FREE Full text] [doi: 10.1093/jamia/ocaa057] [Medline: 32467973]
32. Zulueta J, Piscitello A, Rasic M, Easter R, Babu P, Langenecker SA, et al. Predicting mood disturbance severity with mobile phone keystroke metadata: a biaffect digital phenotyping study. J Med Internet Res 2018 Jul 20;20(7):e241 [FREE Full text] [doi: 10.2196/jmir.9775] [Medline: 30030209]
33. Mastoras R, Iakovakis D, Hadjidimitriou S, Charisis V, Kassie S, Alsaadi T, et al. Touchscreen typing pattern analysis for remote detection of the depressive tendency. Sci Rep 2019 Sep 16;9(1):13414 [FREE Full text] [doi: 10.1038/s41598-019-50002-9] [Medline: 31527640]
34. Cao B, Zheng L, Zhang C, Yu P, Piscitello A, Zulueta J. Deepmood: modeling mobile phone typing dynamics for mood detection. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017 Presented at: 23rd ACM SIGKDD International Conference; August 13–17, 2017; Halifax, NS, Canada. [doi: 10.48550/arXiv.1803.08986]
35. Ajilore O, Bark JS, Demos AP, Zulueta J, Stange J, Duffecy J, et al. Assessment of cognitive function in bipolar disorder with passive smartphone keystroke metadata: a BiAffect digital phenotyping study. Front Psychiatry 2025 Mar 6;16:1430303 [FREE Full text] [doi: 10.3389/fpsyt.2025.1430303] [Medline: 40115651]
36. Ning E, Estabrook R, Tulabandhula T, Zulueta J, Ross MK, Kabir S, et al. Predicting cognitive functioning in mood disorders through smartphone typing dynamics. Journal of Psychopathology and Clinical Science 2025 Nov;134(8):998-1019. [doi: 10.1037/abn0001052]
37. Schultz L, Murphy M, Donegan M, Knights J, Baker J, Thompson M, et al. Evaluating the acceptability and feasibility of collecting passive smartphone data to estimate psychological functioning in U.S. Service Members and Veterans: a pilot study. Mil Med 2025 Jan 16;190(1-2):285-292. [doi: 10.1093/milmed/usae144] [Medline: 38619334]
38. Matias P, Afonseca P, Nunes F, Henriques A, Gonçalves C, Rodrigues A. Smartphone keyboard typing for rheumatic disease identification: a machine learning approach. 2025 Presented at: EPIA 2025: Progress in Artificial Intelligence - 24th EPIA Conference on Artificial Intelligence; October 1–3, 2025; Faro, Portugal. [doi: 10.1007/978-3-032-05176-9_7]
39. Zhu N, Sarawgi A, Bühner M, Baumeister H, Garatva P, Ehring T, et al. The relation between passively collected data and PTSD: a systematic review and meta-analysis. NPJ Digit Med 2025 Jul 05;8(1):413 [FREE Full text] [doi: 10.1038/s41746-025-01825-6] [Medline: 40617877]
40. Insel T, Cuthbert B, Garvey M, Heinssen R, Pine D, Quinn K, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am J Psychiatry 2010 Jul;167(7):748-751. [doi: 10.1176/appi.ajp.2010.09091379] [Medline: 20595427]
41. APA Style Journal Article Reporting Standards (APA Style JARS). American Psychological Association (APA). 2025. URL: https://apastyle.apa.org/jars [accessed 2025-07-01]
42. Bijur P, Latimer C, Gallagher E. Validation of a verbally administered numerical rating scale of acute pain for use in the emergency department. Acad Emergency Med 2003 Apr;10(4):390-392. [doi: 10.1111/j.1553-2712.2003.tb01355.x]
43. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, PROMIS Cooperative Group. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol 2010 Nov;63(11):1179-1194 [FREE Full text] [doi: 10.1016/j.jclinepi.2010.04.011] [Medline: 20685078]
44. Morin C, Belleville G, Bélanger L, Ivers H. The Insomnia Severity Index: psychometric indicators to detect insomnia cases and evaluate treatment response. Sleep 2011 May 01;34(5):601-608 [FREE Full text] [doi: 10.1093/sleep/34.5.601] [Medline: 21532953]
45. Weathers FW, Bovin MJ, Lee DJ, Sloan DM, Schnurr PP, Kaloupek DG, et al. The Clinician-Administered PTSD Scale for DSM–5 (CAPS-5): development and initial psychometric evaluation in military veterans. Psychological Assessment 2018 Mar;30(3):383-395. [doi: 10.1037/pas0000486]
46. Auvergne L, Bortsov AV, Ulirsch JC, Peak DA, Jones JS, Swor RA, et al. Association of epidemiologic factors and genetic variants influencing hypothalamic-pituitary-adrenocortical axis function with postconcussive symptoms after minor motor vehicle collision. Psychosom Med 2016 Jan;78(1):68-78 [FREE Full text] [doi: 10.1097/PSY.0000000000000253] [Medline: 26588823]
47. Blevins CA, Weathers FW, Davis MT, Witte TK, Domino JL. The Posttraumatic Stress Disorder Checklist for DSM-5 (PCL-5): development and initial psychometric evaluation. J Trauma Stress 2015 Dec;28(6):489-498. [doi: 10.1002/jts.22059] [Medline: 26606250]

XSL•FO
RenderX

48. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal 1999 Jan;6(1):1-55. [doi: 10.1080/10705519909540118]
49. Niska R, Bhuiya F, Xu J. National Hospital Ambulatory Medical Care Survey: 2007 emergency department summary. Natl Health Stat Report 2010 Aug 06(26):1-31 [FREE Full text] [Medline: 20726217]
50. Lam K, Meijer K, Loonstra F, Coerver E, Twose J, Redeman E, et al. Real-world keystroke dynamics are a potentially valid biomarker for clinical disability in multiple sclerosis. Mult Scler 2020 Nov 05;27(9):1421-1431. [doi: 10.1177/1352458520968797]
51. Walter KH, Palmieri PA, Gunstad J. More than symptom reduction: changes in executive function over the course of PTSD treatment. Journal of Traumatic Stress 2010 Apr 06;23(2):292-295. [doi: 10.1002/jts.20506]
52. Simmons A, Strigo I, Matthews S, Paulus M, Stein M. Initial evidence of a failure to activate right anterior insula during affective set shifting in posttraumatic stress disorder. Psychosom Med 2009 May;71(4):373-377 [FREE Full text] [doi: 10.1097/PSY.0b013e3181a56ed8] [Medline: 19398499]
53. Nahleen S, Nixon RD, Takarangi MK. Current PTSD symptomatology distorts memory for past symptoms. Psychiatry Res 2019 Apr;274:330-334. [doi: 10.1016/j.psychres.2019.02.052] [Medline: 30849715]
54. National Institutes of Health. URL: https://nda.nih.gov/edit_collection.html?id=2526 [accessed 2026-04-06]

Abbreviations

APNS: adverse posttraumatic neuropsychiatric symptom
AURORA: Advancing Understanding of Recovery After Trauma
ED: emergency department
MVC: motor vehicle collision
PTS: posttraumatic stress

Edited by S Brini; submitted 13.Mar.2025; peer-reviewed by E Harris, O Tutsoy, M Ross; comments to author 26.Nov.2025; revised version received 25.Mar.2026; accepted 31.Mar.2026; published 01.Jun.2026.

Please cite as:
Short NA, An X, Ji Y, Li Q, Neylan TC, Clifford GD, House SL, Beaudoin FL, Stevens JS, Linnstaedt SD, Germine LT, Haran JP, Storrow AB, Lewandowski C, Musey Jr PI, Hendry PL, Sheikh S, Jones CW, Punches BE, Pascual JL, Seamon MJ, Harris E, Pearson C, Merchant RC, Domeier RM, Rathlev NK, O'Neil BJ, Sergot P, Sanchez LD, Bruce SE, Kessler RC, Koenen KC, Ressler KJ, McLean SA
Smartphone Keystroke Biomarkers as Predictors of Adverse Neuropsychiatric Sequelae After Trauma in Trauma Survivors: Prospective Observational Cohort Study
J Med Internet Res 2026;28:e73771
URL: https://www.jmir.org/2026/1/e73771
doi:10.2196/73771
PMID:

©Nicole A Short, Xinming An, Yinyao Ji, Qinghua Li, Thomas C Neylan, Gari D Clifford, Stacey L House, Francesca L Beaudoin, Jennifer S Stevens, Sarah D Linnstaedt, Laura T Germine, John P Haran, Alan B Storrow, Christopher Lewandowski, Paul I Musey Jr, Phyllis L Hendry, Sophia Sheikh, Christopher W Jones, Brittany E Punches, Jose L Pascual, Mark J Seamon, Erica Harris, Claire Pearson, Roland C Merchant, Robert M Domeier, Niels K Rathlev, Brian J O'Neil, Paulina Sergot, Leon D Sanchez, Steven E Bruce, Ronald C Kessler, Karestan C Koenen, Kerry J Ressler, Samuel A McLean. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Associations Between Mukbang Watching and Appetite, Nutrition, and Quality of Life in Pediatric Patients with Cancer: Intensive Longitudinal Study

Yuxue Xiao[1*], MSN; Jiani Wang[1,2*], MSN; Ting Zhong[3*], BSN; Xiao Wang[1,4], MSN; Dongyan Tang[3], BSN; Wei Xia[1], MPhil, PhD

[1]School of Nursing, Sun Yat-sen University, RM 613, Block of School of Nursing, Sun Yat-Sen University North Campus, No. 74, 2nd Yat-Sen Road,Yuexiu District, Guangzhou, China

[2]School of Nursing, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, Southern, China (Hong Kong)

[3]Department of Pediatric Hematologic Oncology, Sun Yat-sen Memorial Hospital, Guangzhou, China

[4]Department of Gerontology, Peking University Sixth Hospital, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Wei Xia, MPhil, PhD
School of Nursing, Sun Yat-sen University, RM 613, Block of School of Nursing, Sun Yat-Sen University North Campus, No. 74, 2nd Yat-Sen Road,Yuexiu District, Guangzhou, China

Abstract

Background: Malnutrition is prevalent among children with cancer and is often exacerbated by diminished appetite. To combat this, hospitalized children are increasingly adopting mukbang, a popular online eating show genre.

Objective: This study aimed to assess the associations between mukbang watching and appetite, nutrition, and quality of life in children with cancer.

Methods: From September 2022 to June 2023, an intensive longitudinal study involved 179 children undergoing radiotherapy or chemotherapy at 2 tertiary hospitals in Guangzhou, China. Daily mukbang watching and appetite data were collected for 5 consecutive days. Nutrition and quality of life were assessed pretreatment and posttreatment. Latent class analysis and generalized estimating equation (GEE) models were used. Sensitivity analyses were conducted by stratifying participants based on baseline appetite levels using adjusted GEE models.

Results: Among participants, 63.7% (114/179) watched mukbang during hospitalization, forming low (C2), medium (C3), and high (C4) watching classes. The adjusted appetite GEE model indicated a significant interaction effect of time × class in both the medium (β=0.08, 95% CI 0.02 - 0.13; *P*=.006) and high classes (β=0.10, 95% CI 0.04 - 0.17; *P*=.003). No significant association was observed between mukbang watching and nutritional status over time. A significantly higher quality of life was observed in the medium (β=19.18, 95% CI 7.84 - 30.53; *P*=.001) and high (β=13.63, 95% CI 1.15 - 26.12; *P*=.03) classes compared to the never class, with significant negative interactions of time × class observed in the low (β=−7.58, 95% CI −14.78 to −0.37; *P*=.04) and medium classes (β=−14.00, 95% CI −22.06 to −5.94; *P*=.001). Sensitivity analyses confirmed that appetite maintenance was robust across baseline appetite subgroups.

Conclusions: Medium- and high-frequency mukbang watching were associated with better appetite maintenance over time among children undergoing treatment. Medium and high classes showed higher overall quality of life compared to the never class; however, the low and medium classes exhibited greater declines over time. These findings suggest the potential for mukbang in appetite-related symptom management while highlighting the need to address the psychological gap between virtual stimulation and physical reality in future research.

Trial Registration: ClinicalTrials.gov NCT05493020; https://clinicaltrials.gov/ct2/show/NCT05493020

(J Med Internet Res 2026;28:e80932) doi:10.2196/80932

KEYWORDS

appetite; mukbang; nutrition; pediatric oncology; quality of life

XSL•FO
RenderX

Introduction

Globally, the incidence of pediatric cancer is increasing, with approximately 400,000 children and adolescents aged 0 to 19 years diagnosed with cancer worldwide [1]. Significant advancements in diagnosis, treatment, and nursing care have raised the average cure rate to 52% and the 5-year survival rate to 80% for children with cancer [2]. Standard cancer treatments, such as surgery and radiation, cause side effects that impact the physical, psychological, and social functioning of pediatric patients with cancer [3]. Consequently, these patients often experience compromised well-being and a reduced quality of life [3].

Evidence shows that malnutrition, particularly inadequate nutrition, can lead to decreased treatment tolerance, impaired functional capacity, prolonged hospital stays, and substantially reduced quality of life [4]. Malnutrition affects up to 75% of children with cancer, primarily due to low dietary intake and excessive disease-related energy expenditure [5]. Some studies have explored nutritional interventions for pediatric patients with cancer, focusing on hospital-based nutritional support, targeted dietary modifications, and structured exercise regimens [6]. However, these interventions have often yielded unsatisfactory outcomes. The ineffectiveness of many nutritional interventions can be attributed to diminished appetite and inadequate food intake in children [7]. Appetite, as the primary regulator of food intake, is crucial for improving nutrition [8]. Therefore, developing strategies to stimulate appetite and promote food intake is essential to prevent malnutrition and maintain quality of life in pediatric patients with cancer.

Pediatric patients with cancer often experience appetite loss [7], and social media content such as mukbang is increasingly being explored by children to boost their food intake. Mukbang, a globally recognized phenomenon, is an online culinary spectacle that captivates viewers with a fusion of auditory and visual experiences and fosters dynamic, interactive dialogue between the host and the audience [9]. Food-related visual and auditory stimuli have been found to enhance the exploration of various sensory dimensions, including those linked to physical tasting encounters [10,11]. Evidence shows that individuals experience positive emotional changes and report diverse taste sensations when exposed to various mukbang videos [12]. A recent scoping review also emphasized that mukbang may have positive effects on viewers' appetites, food choices, and psychological well-being [11]. Mukbang watching is a deliberate, self-motivated choice among children to invigorate their appetites, reflecting a growing acceptance of interactive culinary experiences.

However, evidence has also linked problematic mukbang watching to disordered eating behaviors, showing a significant positive correlation [13]. A previous qualitative study further explored the psychological and physiological experiences of youth watching mukbang, finding that while it may alleviate mental health concerns and stimulate appetite, long-term watching might have adverse health effects [14]. Given the uncertainty surrounding the impact of mukbang-watching behavior among pediatric patients with cancer, further research is needed.

Appetite's transient and subjective nature suggests that cross-sectional methods may be subject to recall biases. An intensive longitudinal approach, tracking appetite and watching habits among pediatric patients with cancer in natural settings, provides more insights than retrospective studies. Therefore, this study aimed to establish evidence for the potential of mukbang as an intervention to enhance nutrition and appetite in children with cancer by exploring its impact on the appetite, nutrition, and quality of life of pediatric patients through an intensive longitudinal study.

Methods

Study Design

We conducted an intensive longitudinal study using ecological momentary assessment [15] to assess the association between mukbang watching and appetite, nutritional status, and quality of life in children with cancer undergoing radiotherapy or chemotherapy. Intensive longitudinal designs rely on high-frequency repeated assessments over short periods to capture within-person fluctuations. Nutritional status and quality of life were measured at baseline and at the end of treatment. Mukbang watching and appetite were assessed 15 times over 5 consecutive days, with assessments conducted 3 times daily at mealtimes to capture real-time changes. This design minimized recall bias [16] and accurately reflected the transient nature of appetite. The protocol has been prospectively registered on ClinicalTrials.gov (NCT05493020).

Participants

Recruitment was conducted between September 2022 and June 2023 using convenience sampling. Children hospitalized in the pediatric oncology wards of 2 tertiary hospitals in China were invited to participate. Eligible participants met the following criteria: (1) aged 3 to 18 years; (2) diagnosed with cancer according to the *Notice on Diagnosis and Treatment Specifications for 12 Pediatric Hematological and Malignancy Diseases (2021 Edition)*, issued by the General Office of the China National Health and Health Commission [17]; and (3) had received at least 1 session of radiotherapy or chemotherapy.

Exclusion criteria included (1) severe cardiopulmonary disorders, serious infections, significant organ damage, genetic metabolic disorders, immune deficiencies, or mental illnesses; (2) inability to watch or listen to videos due to physical dysfunction; and (3) restricted diet due to illness or treatment factors. Dropout criteria included transfer or death, physical weakness preventing participation, and discontinuation of radiation or chemotherapy for any reason during the study.

Sample Size

The sample size was calculated using an online calculator developed for intensive longitudinal studies based on the framework described by Pirla et al [18]. Using the affect dynamics measure with an effect size of 0.33, a power of 0.9, and an α level of .05 [19], a minimum of 110 participants was required for 15 repeated measures. Accounting for a 10%

XSL•FO
RenderX

dropout rate at the end of follow-up, this study required at least 122 participants.

Measures

Overview

A structured questionnaire was used to collect data. Demographic details (age and sex), disease characteristics (diagnosis, treatment type, and phase), and nutritional indicators (serum albumin, hemoglobin, and arm circumference) were obtained from the medical records of the patients. Mukbang watching behaviors, therapy side effects, appetite, nutritional status, and quality of life were evaluated using various questionnaires and scales. The selected variables assess the pathway from behavioral triggers (mukbang watching) to health outcomes (appetite, nutrition, and quality of life) according to a previous qualitative study [14]. All these instruments were utilized with formal permission from their respective developers or copyright holders.

Characteristics of Mukbang-Watching Behavior

Mukbang-watching behaviors were recorded using a self-designed questionnaire adapted from our prior qualitative research on Chinese youth [14]. It included details on watching duration, frequency, preferences, and experiences during mukbang sessions. To ensure the clarity and feasibility of the items, the tool was pilot-tested with 5 pediatric patients, and the questionnaire was further refined based on their feedback before formal administration.

Therapy-Related Symptom Checklist for Children

Therapy-related symptoms were assessed using the 26-item TRSC-C (Therapy-Related Symptom Checklist for Children) (Checklist 1) [20]. Each item is rated on a 5-point scale, with a higher score indicating more severe symptoms. The scale has shown good internal consistency (Cronbach α=0.87) among Chinese children [21].

Cancer Appetite and Symptom Questionnaire

Appetite was measured using the 12-item Cancer Appetite and Symptom Questionnaire (CASQ). The total score of the questionnaire is 48 points [22], with lower total scores indicating poor appetite. The Chinese version of the CASQ demonstrated good split-half reliability, with a score of 0.85 [23].

Subjective Global Nutritional Assessment

Subjective global nutritional assessment was adopted to measure the nutritional status of participants [24]. The standard includes (1) well-nourished, when the child is growing normally, with adequate food consumption and without gastrointestinal symptoms; (2) moderately malnourished, when the patient shows signs of weight loss or loss of food consumption, functional capacity, and reduced muscle mass, demonstrating nutritional status impairment when it was previously normal; and (3) severely malnourished, when the child has progressive malnutrition, with weight loss, reduction of muscle and fat mass, and loss of food consumption. Participants were classified as well-nourished, moderately malnourished, or severely malnourished. The Chinese version of the subjective global nutritional assessment demonstrated a substantial Cronbach α coefficient of 0.736 [25].

Pediatric Quality of Life Inventory Measurement Models 3.0

Quality of life was measured using the 27-item Pediatric Quality of Life Inventory (PedsQL) 3.0 Cancer Module [26]. Each item is rated on a 5-point Likert scale. Items were reverse-scored and linearly transformed to a 0 - 100 scale (0=100, 1=75, 2=50, 3=25, and 4=0) following the standard PedsQL scoring protocol. The total score was calculated as the mean of all items, with higher scores representing a better quality of life. The PedsQL has demonstrated reliability in evaluating the quality of life in childhood cancer survivors, exhibiting Cronbach α coefficients between 0.87 and 0.95 across its subscales [27].

Data Collection

Baseline data were collected on demographics, disease characteristics, therapy-related symptoms, mukbang-watching behavior, appetite, nutritional status, and quality of life. Participants and their guardians were given a self-recorded brochure (Multimedia Appendix 1) to track their appetite and mukbang-watching behavior during meals for 5 consecutive days starting from the first day of radiotherapy or chemotherapy. To maximize data completeness, research staff performed a brief quality check upon collection to ensure no mealtimes were inadvertently skipped. If any missing entries were identified, parents were invited to provide the information while their memory was still fresh. Nutritional status and quality of life were assessed at the end of the current treatment phase. The schedule for data collection and the flowchart of sample selection are presented in Multimedia Appendix 2.

Statistical Analysis

Categorical variables are shown as frequencies and percentages, while continuous variables and normally distributed data are presented as mean (SD). The Wilcoxon and chi-square tests were used to identify factors differentiating the mukbang watching and nonwatching classes.

The frequency of mukbang-watching behavior was classified using latent class analysis (LCA) with R software (version 4.2.0; R Foundation for Statistical Computing). Models with 1 to 5 latent classes were repeatedly fitted, increasing the number of latent classes in a stepwise manner. The models were evaluated using log likelihood, Akaike information criterion, adjusted Bayesian information criterion, entropy, and the Vuong-Lo-Mendell-Rubin ratio test. Lower log likelihood, Bayesian information criterion, and Akaike information criterion values indicate a better model fit [28], and relative entropy values greater than 0.80 indicate sufficient certainty in classification [28]. The features of the latent classes were compared using a chi-square test. Subsequently, unadjusted and adjusted generalized estimating equation (GEE) models were used to explore the association between mukbang-watching behavior and changes in appetite, nutritional status, and quality of life over time. This approach was chosen to effectively account for the correlation between repeated measures and to provide robust population-averaged estimates for the longitudinal data.

XSL•FO
RenderX

Additionally, to assess the robustness of our primary findings against baseline appetite imbalances, we conducted a sensitivity analysis by stratifying the participants based on CASQ scores into 3 groups: low (0 - 16), medium (17-32), and high (33-48) baseline appetite. Given the restricted distribution of actual scores, in our sample, no participants fell into the low appetite category at baseline. We conducted the sensitivity analysis within the medium and high baseline appetite subgroups and re-evaluated the associations between mukbang watching and appetite, nutrition, and quality of life using adjusted GEE models.

Except for LCA, all statistical analyses were performed using SPSS (version 26.0; IBM Corp). Statistical significance was set at 2-sided $P<.05$. No missing values were found in this study due to careful checking of the data for each collection.

Ethical Considerations

Ethical approval for the study was obtained from the institutional review board of the School of Nursing, Sun Yat-sen University, following the Declaration of Helsinki (reference: L2022SYSU-HL-089). Written informed consent was obtained from all participants and their guardians prior to their participation, after they were fully informed of the study's purpose and procedures. To ensure privacy and confidentiality, all personal identifiers were removed, and a unique study identification number was assigned to each participant for data entry, management, and analysis.

Results

The study enrolled 179 eligible participants, with no dropouts during the follow-up period. As shown in Table 1, participants had a mean age of 7.7 (SD 3.7) years, with 60.3% (n=108) being male and 65.4% (n=117) diagnosed with leukemia. Most (n=67, 37.4%) were in their first phase of treatment. At baseline, participants reported an average of 4.1 (SD 4.0) days of appetite loss during their last treatment, and 83.8% (n=150) demonstrated adequate nutritional status. The mean quality of life score was 72.8 (SD 12.5).

XSL•FO
RenderX

Table . Demographic and disease-related data of participants at baseline (N=179).

Item	Values			*P* value
	Total (N=179)	Watch mukbang (n=114)	Do not watch mukbang (n=65)	
Age (y)[a], mean (SD)	7.7 (3.7)	7.5 (3.7)	8.0 (3.7)	.42
3, n (%)	82 (45.8)	55 (48.2)	27 (41.5)	
7, n (%)	32 (17.9)	17 (14.9)	15 (23.1)	
10 - 18, n (%)	65 (36.3)	42 (36.8)	23 (35.4)	
Sex, n (%)				.31
Male	108 (60.3)	72 (63.2)	36 (55.4)	
Female	71 (39.7)	42 (36.8)	29 (44.6)	
Height (cm), mean (SD)	125.6 (25.1)	125.4 (26.2)	125.8 (23.5)	.91
Arm circumference (cm), mean (SD)	17.5 (3.8)	17.5 (3.9)	17.5 (3.7)	.93
Weight (kg), mean (SD)	27.3 (14.7)	27.7 (15.9)	26.5 (12.3)	.58
Disease diagnosis, n (%)				.009
Leukemia	117 (65.4)	83 (72.8)	34 (52.3)	
Lymphoma	16 (8.9)	8 (7)	8 (12.3)	
Sarcoma	16 (8.9)	8 (7)	8 (12.3)	
Neuroblastoma	10 (5.6)	8 (7)	2 (3.1)	
Others[b]	20 (11.2)	7 (6.1)	13 (20)	
Treatment phase, n (%)				>.99
I	67 (37.4)	43 (37.7)	24 (36.9)	
II	57 (31.8)	36 (31.6)	21 (32.3)	
III	31 (17.3)	20 (17.5)	11 (16.9)	
IV	24 (13.4)	15 (13.2)	9 (13.8)	
Albumin (g/L)[c], n (%)				.04
<40	82 (45.8)	59 (51.8)	23 (35.4)	
40 - 55	97 (54.2)	55 (48.2)	42 (64.6)	
Hemoglobin (g/L)[d], n (%)				.02
<110	152 (84.9)	102 (89.5)	50 (76.9)	
110 - 160	27 (15.1)	12 (10.5)	15 (23.1)	
Duration of the loss of appetite during the last therapy day, mean (SD)	4.1 (4.0)	3.9 (4.0)	4.5 (4.1)	.35
Watch other kinds of videos before or during meals (yes), n (%)	127 (70.9)	90 (70.9)	37 (29.1)	.002
Playing video games before or during meals (yes), n (%)	73 (40.8)	47 (64.4)	26 (35.6)	.87
TRSC-C[e], mean (SD)	9.4 (6.7)	8.9 (5.4)	10.2 (8.5)	.22
Appetite[f], mean (SD)	30.4 (3.8)	31.4 (3.6)	28.7 (3.6)	<.001
Nutritional status (adequate nutrition)[g], n (%)	150 (83.8)	99 (86.8)	51 (78.5)	.14

XSL•FO
RenderX

Item	Values			*P* value
	Total (N=179)	Watch mukbang (n=114)	Do not watch mukbang (n=65)	
Quality of life[h], mean (SD)	72.8 (12.5)	74.4 (11.3)	70.1 (14.0)	.08
The frequency of watching mukbang[i], n (%)				
Less than once a week	—[j]	12 (10.5)	—	—
Once a week-once a day	—	39 (34.2)	—	—
Once a day-once a meal	—	48 (42.1)	—	—
More than one meal	—	15 (13.2)	—	—
The duration of watching mukbang[i], n (%)				
Less than half an hour	—	58 (50.9)	—	—
Half an hour-an hour	—	38 (33.3)	—	—
More than an hour	—	18 (15.8)	—	—
The time point for watching mukbang[k], n (%)				
Eating	—	51 (44.7)	—	—
Nauseated	—	5 (4.4)	—	—
Bored	—	90 (78.9)	—	—
Insomnia	—	2 (1.8)	—	—
The experience of watching mukbang[i,k], n (%)				
Increased appetite	—	51 (44.7)	—	—
Decreased nausea	—	4 (3.5)	—	—
Increased pleasure	—	85 (74.6)	—	—
Increased contentment	—	30 (26.3)	—	—

[a]Categorized according to the Erikson stages of development: 3-6 years (preschool), 7-9 years (middle school), and 10-18 years (adolescence).

[b]Medulloblastoma, hepatoblastoma, nephroblastoma, teratoma, intracranial nongerm cell tumors, left ovarian yolk sac tumor, Langerhans cell histiocytosis, myelodysplastic syndrome, malignant mesenchymal tumors, hemophagocytic syndrome, and sellar region low-grade glioma.

[c]Pediatric reference interval for albumin: <40 g/L (low), 40 - 55 g/L (normal), and >55 g/L (high).

[d]Pediatric reference interval for hemoglobin: <110 g/L (low), 110 - 160 g/L (normal), and >160 g/L (high).

[e]TRSC-C: Therapy-Related Symptom Checklist for Children. This checklist aimed to enhance comprehension of the specific needs of children and comprises 26 items to cover therapy-related symptoms. Each item is assessed using a numerical scale ranging from 0 to 4, where higher scores correspond to a more pronounced manifestation of symptoms.

[f]Appetite of children with cancer was measured by the Cancer Appetite and Symptom Questionnaire. A lower total score indicates a worse appetite in a patient.

[g]Nutritional status was measured by using the subjective global nutritional assessment.

[h]Quality of life was measured by using the Pediatric Quality of Life Inventory (PedsQL) 3.0 Cancer Module. Higher scores indicate a better quality of life.

[i]Only children who watched mukbang were included (n=114).

[j]Not applicable.

[k]More than 1 answer is possible.

Regarding mukbang watching habits, 63.7% (114/179) of participants reported watching mukbang. Among these viewers, 42.1% (48/114) watched once daily per meal, and 50.9% (58/114) watched for less than half an hour. Notably, a high proportion engaged in mukbang watching due to boredom (90/114, 78.9%) and reported an enhanced sense of pleasure while watching (74.6%, 85/114). Further details regarding the demographic and disease-related characteristics of the participants are presented in Table 1.

The results of the LCA demonstrated that a 4-class model best fits the data, as indicated by the fit statistics shown in Table 2. This model provided a better fit than the 5-class model, with an entropy of 0.935, indicating good separation between the classes. The average posterior probabilities for the classes were 0.84 (class 1), 0.88 (class 2), 0.86 (class 3), and 0.87 (class 4), suggesting minimal classification uncertainty. Classes 1 (C1) to 4 (C4) comprised 66 (36.9%), 26 (14.5%), 51 (28.5%), and 36 (20.1%) of the 179 participants, respectively.

XSL•FO
RenderX

Table . Model fit statistics for latent class analysis regarding watching mukbang behavior (N=179).

Classes	LL[a]	AIC[b]	BIC[c]	Entropy	VLMR[d]
1	−1934.317	3928.635	4024.256	__[e]	—
2	−1327.031	2776.061	2970.492	0.903	<.001
3	−1256.492	2696.985	2990.224	0.931	<.001
4	−1214.411	2674.822	3066.871	0.935	<.001
5	−1188.508	2685.016	3175.873	0.924	<.001

[a]LL: log likelihood.

[b]AIC: Akaike information criterion.

[c]BIC: Bayesian information criterion.

[d]VLMR: Vuong-Lo-Mendell-Rubin likelihood ratio test.

[e]Not applicable.

Figure 1 illustrates mukbang watching frequency, sorted by the degree of separation between classes. C1 was defined by the absence of mukbang watching among participants at baseline and throughout the follow-up period. For simplicity, the 4 classes were referred to as never (C1), low (C2), medium (C3), and high (C4) frequencies of mukbang watching.

Baseline demographic and disease characteristics of participants across the 4 classes are detailed in Table 3.

Figure 1. Latent class analysis for the frequency of watching mukbang (N=179). The y-axis displays the daily average of participants' mukbang watching duration per meal, derived from the original categorical codes (0 for no watching, 1 for less than 0.5 h, 2 for 0.5 to 1 h, and 3 for more than 1 h).

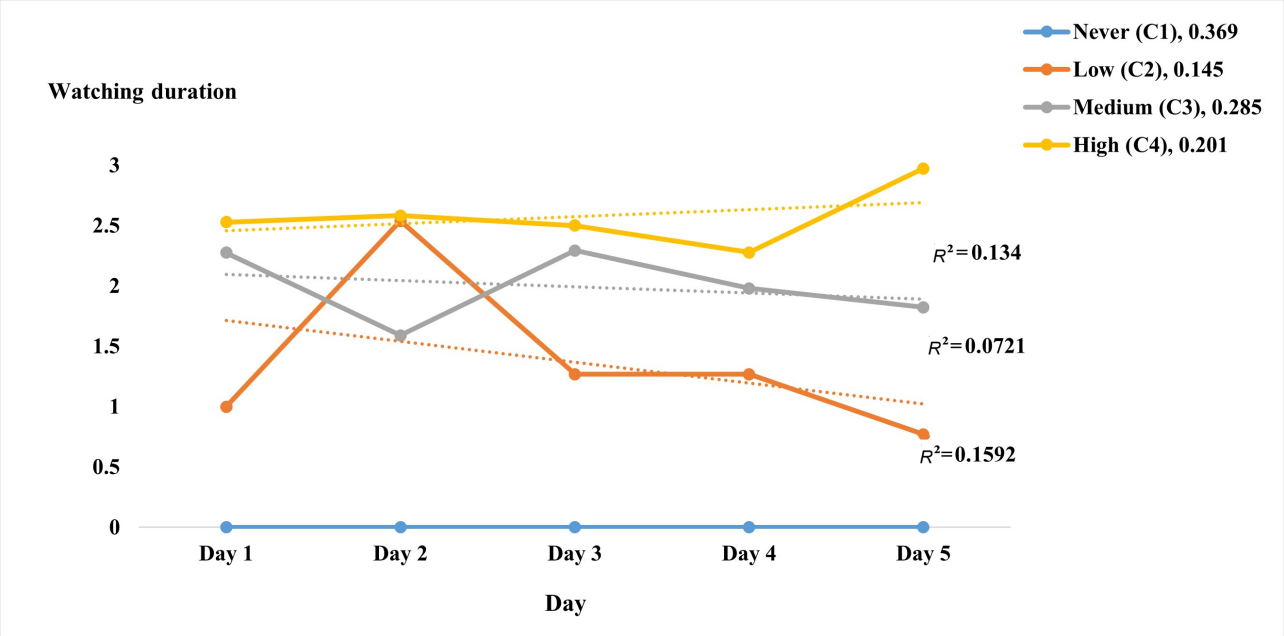

XSL•FO
RenderX

Table . Demographic and disease-related data of 4-class participants at baseline (N=179).

Item	Values				
	Never (n=66)	Low (n=26)	Medium (n=51)	High (n=36)	*P* value
					.36
Age (y)[a], mean (SD)	7.8 (3.7)	8.1 (3.9)	8.1 (3.7)	6.8 (3.5)	
3, n (%)	30 (45.5)	10 (38.5)	22 (43.1)	20 (55.6)	
7, n (%)	14 (21.2)	5 (19.2)	7 (13.7)	6 (16.7)	
10 - 18, n (%)	22(33.3)	11(42.3)	22(43.1)	10(27.8)	
Sex, n (%)					.61
Male	37 (56.1)	16 (61.5)	30 (58.8)	25 (69.4)	
Female	29 (43.9)	10 (38.5)	21 (41.2)	11 (30.6)	
Height, cm, mean (SD)	124.9 (23.5)	129.4 (29.6)	129.3 (26.8)	118.7 (21.5)	.22
Arm circumference (cm), mean (SD)	17.5 (3.6)	18.3 (4.6)	18.0 (4.1)	16.2 (2.9)	.10
Weight (kg), mean (SD)	26.1 (12.4)	31.1 (17.7)	29.9 (17.9)	23.1 (9.7)	.08
Disease diagnosis, n (%)					.02
Leukemia	37 (56.1)	20 (76.9)	36 (70.6)	24 (66.7)	
Lymphoma	8 (12.1)	2 (7.7)	4 (7.8)	2 (5.6)	
Sarcoma	7 (10.6)	0 (0)	6 (11.8)	3 (8.3)	
Neuroblastoma	1 (1.5)	1 (3.8)	2 (3.9)	6 (16.7)	
Others[b]	13 (19.7)	3 (11.5)	3 (5.9)	1 (2.8)	
Treatment phase, n (%)					.08
I	23 (34.8)	14 (53.8)	14 (27.5)	16 (44.4)	
II	22 (33.3)	5 (19.2)	16 (28.1)	14 (38.9)	
III	13 (19.7)	5 (19.2)	13 (25.5)	0 (0)	
IV	8 (12.1)	2 (7.7)	8 (15.7)	6 (16.7)	
Albumin (g/L)[c], n (%)					.43
<40	25 (37.9)	13 (50)	25 (49)	19 (23.2)	
40 - 55	41 (62.1)	13 (50)	26 (51)	17 (47.2)	
Hemoglobin (g/L)[d], n (%)					.17
<110	51 (77.3)	22 (84.6)	46 (90.2)	33 (91.7)	
110 - 160	15 (22.7)	4 (15.4)	5 (9.8)	3 (8.3)	
Duration of the loss of appetite during the last therapy (day), mean (SD)	4.7 (4.7)	3.6 (4.6)	3.3 (3.1)	4.3 (3.2)	.24
Watching other kinds of videos before or during meals (yes), n (%)	36 (54.5)	18 (69.2)	45 (88.2)	28 (77.8)	.001
Playing video games before or during meals (yes), n (%)	26 (39.4)	14 (53.8)	22 (43.1)	11 (30.6)	.31
TRSC-C[e], mean (SD)	10.8 (8.7)	7.9 (5.4)	8.8 (5.8)	8.8 (3.8)	.18
Appetite[f], mean (SD)	28.6 (3.6)	31.4 (3.2)	31.9 (3.7)	30.9 (3.5)	<.001

XSL•FO
RenderX

Item	Values				
	Never (n=66)	Low (n=26)	Medium (n=51)	High (n=36)	*P* value
Nutritional status (adequate nutrition)[g], n (%)	51 (77.3)	22 (84.6)	45 (88.2)	32 (88.9)	.22
Quality of life[h], mean (SD)	69.7 (14.1)	74.2 (10.0)	77.1 (11.6)	71.5 (10.8)	.02

[a]Categorized according to the Erikson stages of development: 3 - 6 years (preschool), 7 - 9 years (middle school), and 10 - 18 years (adolescence).

[b]Medulloblastoma, hepatoblastoma, nephroblastoma, teratoma, intracranial nongerm cell tumors, left ovarian yolk sac tumor, Langerhans cell histiocytosis, myelodysplastic syndrome, malignant mesenchymal tumors, hemophagocytic syndrome, and sellar region low-grade glioma.

[c]Pediatric reference interval for albumin: <40 g/L (low), 40 - 55 g/L (normal), and >55 g/L (high).

[d]Pediatric reference interval for hemoglobin: <110 g/L (low), 110 - 160 g/L (normal), and >160 g/L (high).

[e] TRSC-C: Therapy-Related Symptom Checklist for Children. Higher scores correspond to a more pronounced manifestation of symptoms.

[f]Appetite of children with cancer was measured by the Cancer Appetite and Symptom Questionnaire. A lower total score indicates a worse patient's appetite.

[g]Nutritional status was measured by the subjective global nutritional assessment.

[h]Quality of life was measured by the Pediatric Quality of Life Inventory (PedsQL) 3.0 Cancer Module. Higher scores indicate better quality of life.

At baseline, appetite and quality of life significantly differed among the 4 classes, while no significant baseline differences were observed in nutritional status. Longitudinal analysis (Table S1 in Multimedia Appendix 3) indicated that while participants across all 4 classes numerically improved in quality of life compared to baseline, the extent of this improvement varied across classes, highlighting heterogeneity in the observed changes. Meanwhile, overall nutritional status among the 4 classes exhibited a decreasing trend.

As shown in Table 4, compared to the baseline, there were significant differences among the 4 classes in appetite and nutritional status at the end of treatment (all *P*s<.001). The adjusted appetite GEE model indicated significant effects of time × class in both the medium class (β=0.08, 95% CI 0.02 - 0.13; *P*=.006) and the high class (β=0.10, 95% CI 0.04 - 0.17; *P*=.003). Both the medium class (β=19.18, 95% CI 7.84 - 30.53; *P*=.001) and the high class (β=13.63, 95% CI 1.15 - 26.12; *P*=.03) showed a significantly higher quality of life compared to the never classes. However, a significant decline in quality of life was also observed in the low class (β=−7.58, 95% CI −14.78 to −0.37; *P*=.04) and the medium class (β=−14.00, 95% CI −22.06 to −5.94; *P*=.001). Changes in appetite, quality of life, and adequate nutrition at baseline (T_0) and during follow-up (T_1-T_{end}) in the 4 classes, based on unadjusted GEE models, are shown in Table S2 in Multimedia Appendix 3.

XSL•FO
RenderX

Table . Changes in appetite, quality of life, and adequate nutrition at baseline (T_0) and during follow-up (T_1-T_{end}) in 4 classes based on adjusted generalized estimating equation models.

Item	Values[a]	*P* value
Appetite[b], adjusted β (95% CI)		
Time	−0.28 (−0.30 to −0.25)	<.001
Class (reference: never)		
Low	0.07 (−0.50 to 0.65)	.81
Medium	0.33 (−0.13 to 0.80)	.16
High	0.01 (−0.54 to 0.55)	.98
Time × class (reference: time × never)		
Time × low	0.03 (−0.04 to 0.10)	.45
Time × medium	0.08 (0.02 to 0.13)	.006
Time × high	0.10 (0.04 to 0.17)	.003
Nutrition status[b], adjusted OR (95% CI)		
Time	0.20 (0.11 to 0.37)	<.001
Class (reference: never)		
Low	0.45 (0.04 to 4.78)	.51
Medium	0.98 (0.13 to 7.67)	.99
High	0.94 (0.11 to 7.93)	.95
Time × class (reference: time × never)		
Time × low	1.71 (0.54 to 5.46)	.36
Time × medium	1.48 (0.50 to 4.38)	.48
Time × high	1.91 (0.63 to 5.79)	.25
Quality of life[c], adjusted β (95% CI)		
Time	4.70 (−0.34 to 9.74)	.07
Class (reference: never)		
Low	9.81 (1.12 to 20.75)	.08
Medium	19.18 (7.84 to 30.53)	.001
High	13.63 (1.15 to 26.12)	.03
Time × class (reference: time × never)		
Time × low	−7.58 (−14.78 to −0.37)	.04
Time × medium	−14.00 (−22.06 to −5.94)	.001
Time × high	−8.01 (−16.34 to 0.32)	.06

[a]β coefficients were reported for appetite and quality of life; odds ratios were reported for nutritional status.

[b]Adjusted for age, gender, disease diagnosis, treatment phase, playing video games before or during meals, watching other kinds of videos before or during meals, baseline appetite, and TRSC-C (Therapy-Related Symptom Checklist for Children) score.

[c]Adjusted for age, gender, disease diagnosis, treatment phase, playing video games before or during meals, watching other kinds of videos before or during meals, baseline appetite, TRSC-C score, and quality of life.

As shown in Table S3 in Multimedia Appendix 3, the sensitivity analysis stratified by baseline appetite yielded results consistent with those of the main analysis. Regarding appetite, the adjusted GEE model indicated significant interaction effects in both the medium (β=.07, 95% CI 0.01 - 0.13; *P*=.03) and high baseline appetite subgroups (β=.16, 95% CI 0.05 - 0.28; *P*=.005). For quality of life, the medium baseline appetite subgroup showed significant class differences (medium: β=19.18, 95% CI 7.84 - 30.53, *P*=.001; high: β=13.63, 95% CI 1.15 - 26.12, *P*=.03) and significant interaction effects (low: β=−7.58, 95% CI −14.78 to −0.37, *P*=.04; medium: β=−14.00, 95% CI −22.06 to −5.94, *P*=.001). In contrast, the high baseline appetite subgroup only exhibited a significant decline over time (β=−8.72, 95% CI −16.73 to −0.71; *P*=.03).

XSL•FO
RenderX

Discussion

This study explored the associations between mukbang watching and appetite, nutrition, and quality of life in pediatric patients with cancer using an intensive longitudinal design. These results may provide preliminary evidence for appetite-related symptom management in pediatric patients with cancer.

Reflecting on a broader trend, our study found that most pediatric patients (63.7%) engaged in mukbang, aligning with recent national data showing that 61.8% and 65.1% of junior and senior high school students in China actively use short video platforms, and 72.1% and 77% engage with social media, respectively [29]. In our research, a prevalent pattern of mukbang engagement was observed among pediatric patients, with 89.5% tuning in weekly and 49.1% spending over 30 minutes per session. This trend closely echoes the data of Tayla et al [19], which revealed a committed viewer base, with 34.0% engaging daily for a comparable duration. This trend likely arises from the limited entertainment options and prolonged treatment times experienced by hospitalized children, leading them to seek media that meets their social and companionship needs.

Furthermore, our findings highlighted a notable preference for watching mukbang content while eating, with a considerable proportion of pediatric patients (44.7%) engaging in this behavior—a rate significantly higher than the specific content-based engagement reported in a previous general pediatric study [30]. The combination of video watching with mealtime was notably more common among children with cancer compared to their healthy peers [30]. Guardians often permit pediatric patients with cancer to eat while watching videos as a practical strategy to help mitigate treatment-related issues such as nausea and appetite loss. This approach aligns with broader evidence suggesting that mukbang can enhance mood and provide sensory stimulation [11], potentially serving as a digital tool to overcome clinical barriers to eating.

This study found that appetite in pediatric patients with cancer declined throughout treatment, with significant variations at different times, matching the results of an earlier study [31]. In contrast, the findings also showed that frequent mukbang watching helped mitigate this decline over time. Notably, a graded association was observed, in which higher frequencies of mukbang watching were associated with greater appetite maintenance. This indicates that consistent mukbang watching may be a contributing factor in the stabilization of appetite trajectories. Interestingly, despite the video-watching habits of some participants in the Never class, their appetite levels did not show a positive trend compared to those in other classes actively watching mukbang. The auditory and visual stimuli in mukbang, including the chewing and swallowing sounds and the appealing display of delicious dishes, stimulate the brain to evoke a desire to eat [11]. Frequent exposure to food content on social media can alter eating habits driven by visual attraction [32]. Hunger may activate brain areas, such as the prefrontal and orbitofrontal cortices and the amygdala, when watching food images, sparking intense cravings [33]. Evidence indicates that auditory cues in videos can enhance the perceived flavor of food, which in turn may stimulate the viewer's appetite [34]. Mukbang creates an enticing ambiance that stimulates appetite.

Beyond sensory stimulation, research has linked loss of appetite in children to various lifestyle factors [35], and mukbang could help address this issue in pediatric patients with cancer. Previous evidence shows that emotional factors often drive food cravings, with eating choices and pace sometimes influenced more by emotions than by physical hunger [36]. The engaging and calming content of mukbang eases the strain of treatment for pediatric patients with cancer, improving their mood and stimulating their appetite. This encourages them to return to these videos during periods of emotional distress or reduced appetite. The graded association suggests that frequent engagement with mukbang potentially amplifies the emotional regulation benefits that alleviate treatment-related appetite loss. Evidence indicates that mukbang, as a type of food-focused media, can address food cravings and anxieties in those on restrictive diets, effectively satisfying appetite-related desires [37]. Mukbang fosters a sense of community for pediatric patients, reducing loneliness and enhancing social support. Its interactive nature strengthens viewer connections by simulating shared dining experiences, which can boost appetite and food consumption.

Despite the observed significant interaction effect on appetite trajectories, no significant changes were found in nutritional status over time. A previous study revealed that mukbang watching is associated with increased food intake [12], though this discrepancy may be explained by the clinical context of cancer treatment. During chemotherapy or other intensive therapies, children often experience treatment-related dietary restrictions, gastrointestinal side effects, mucositis, or infection-control dietary limitations [38], which may prevent increased appetite from translating into greater caloric intake. Moreover, appetite represents a subjective motivational state, whereas nutritional status reflects objective physiological outcomes that may require sustained dietary changes over a longer period to demonstrate measurable improvement [39]. In addition, cancer-related metabolic alterations and systemic inflammation may counteract the potential benefits of increased food desire [40]. Future research could integrate structured nutritional management strategies alongside mukbang watching to examine whether enhanced appetite, when supported by appropriate dietary interventions, contributes to measurable improvements in nutritional outcomes.

A significant baseline difference in quality of life was observed across classes, suggesting that initial well-being and mukbang engagement may be associated. After adjusting for baseline quality of life, our study showed that the observed negative interactions in the low and medium classes indicate a greater decline in quality of life compared to the never class. Sensitivity analyses confirmed the longitudinal trends observed in our study. While the medium baseline appetite subgroup followed a pattern similar to the main analysis, the high baseline appetite subgroup exhibited a distinct decline over time. These results indicate that the association between mukbang watching frequency and quality of life varies according to the patient's baseline appetite. This discrepancy suggests that the physiological and psychological impact of mukbang watching

XSL•FO
RenderX

is sensitive to the patient's baseline appetite, which may be explained by the psychological frustration arising when enhanced appetite meets the physical constraints of treatment [41]. Specifically, the strong desire for food stimulated by mukbang may conflict with patients' dietary restrictions or gastrointestinal discomfort. This contrast between virtual enjoyment and physical reality may create a psychological gap and a sense of deprivation, thereby hindering rapid gains in overall well-being [42]. Beyond that, repeated exposure to highly rewarding food cues without actual consumption may increase reward prediction error and emotional dysregulation [43]. Mukbang may also function as a passive coping strategy that temporarily alleviates distress but limits engagement in more adaptive coping behaviors. Future research should further explore the long-term impact of mukbang watching on both physiological and psychological outcomes.

This study had certain limitations, including a small sample size and a short tracking period focusing on a single pediatric oncology treatment cycle. The relatively smaller sample size of the high baseline appetite subgroup (n=59) may limit the robustness of our sensitivity analyses. While the single-center design may limit the generalizability of the findings, it allowed for consistent data collection protocols. Furthermore, although we implemented real-time data verification to minimize measurement bias, the subjective nature of the CASQ scale and the involvement of parents may cause reporting bias, as caregivers' perceptions may not fully reflect the child's subjective experience. This study also did not explore the relationship between mukbang content and appetite due to the absence of a standard definition for mukbang. Future qualitative follow-up research is recommended to better understand mukbang watching among pediatric patients with cancer.

Despite its limitations, the findings provide preliminary insights for clinical practice and research implications. Our study represents a novel exploration of the associations between mukbang watching and the appetite, nutritional status, and quality of life of pediatric patients with cancer. By using an intensive longitudinal study, we documented the real-time behaviors and appetite of pediatric patients with cancer watching mukbang, offering valuable insights into their evolving needs throughout treatment. These results provide preliminary empirical evidence suggesting that mukbang could be a potential supportive strategy to manage appetite in this vulnerable population.

Our findings show that frequent mukbang watching is associated with better appetite maintenance in pediatric patients with cancer, particularly in the medium and high classes. However, the longitudinal patterns of quality of life were less favorable, with the low and medium classes exhibiting a slower recovery process compared to the never class. Sensitivity analyses revealed that the association between mukbang watching and quality of life varies by baseline appetite. While the medium baseline appetite subgroup showed significant longitudinal differences, the high baseline appetite subgroup was characterized primarily by a significant decline over time. These results indicate that the association between mukbang watching and quality of life is not uniform but is sensitive to the patient's initial appetite status. Further research is needed to identify the factors that may contribute to these varying associations and to explore whether mukbang can be optimized as a supportive strategy for pediatric patients with cancer.

Acknowledgments

The authors give special recognition to Xingcai Zhang for his contributions to the latent class analysis, which greatly enriched their findings. The authors would also like to thank Editage for editing the English language.

Funding

This work was supported by the Natural Science Foundation of Guangdong Province (2024A1515012147; WX), Guangzhou Basic Research Program—Basic and Applied Basic Research Projects, General Projects (202201011311; WX), and the Nursing Innovation Development Research Project of the Guangdong Nursing Association and the Lingnan Nightingale Nursing Research Institute (GDSHLXHYJYY202601; XW). The funder had no involvement in the study design, datacollection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The data will be shared upon reasonable request to the corresponding author.

Authors' Contributions

Conceptualization: YX, TZ, WX
Data curation: YX, XW, DT
Formal analysis: YX, JW, WX
Funding acquisition: WX
Investigation: XW, DT, WX
Methodology: JW, WX
Project administration: YX, TZ, XW, WX
Resources: TZ, DT, WX

XSL•FO
RenderX

Software: XW
Supervision: WX
Validation: JW, DT, WX
Visualization: JW, TZ, XW, WX
Writing – original draft: YX, JW, TZ
Writing – review & editing: WX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Self-reported brochure.
[DOCX File, 486 KB - jmir_v28i1e80932_app1.docx]

Multimedia Appendix 2
Data collection schedule.
[DOCX File, 224 KB - jmir_v28i1e80932_app2.docx]

Multimedia Appendix 3
Participants' profiles of appetite, quality of life, and nutritional status at baseline and during follow-up (Table S1), the longitudinal results based on unadjusted generalized estimating equations (Table S2), and adjusted generalized estimating equations stratified by baseline appetite subgroups (Table S3).
[DOCX File, 36 KB - jmir_v28i1e80932_app3.docx]

Checklist 1
TRSC-C checklist.
[PDF File, 255 KB - jmir_v28i1e80932_app4.pdf]

References

1. Childhood cancer. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/cancer-in-children [accessed 2025-07-01]
2. Lam CG, Howard SC, Bouffet E, Pritchard-Jones K. Science and health for all children with cancer. Science 2019 Mar 15;363(6432):1182-1186. [doi: 10.1126/science.aaw4892] [Medline: 30872518]
3. van Gorp M, van Erp LME, Maas A, et al. Increased health-related quality of life impairments of male and female survivors of childhood cancer: DCCSS LATER 2 psycho-oncology study. Cancer 2022 Mar 1;128(5):1074-1084. [doi: 10.1002/cncr.34003] [Medline: 34726782]
4. Triarico S, Rinninella E, Cintoni M, et al. Impact of malnutrition on survival and infections among pediatric patients with cancer: a retrospective study. Eur Rev Med Pharmacol Sci 2019 Jan;23(3):1165-1175. [doi: 10.26355/eurrev_201901_17009] [Medline: 30779086]
5. Joffe L, Ladas EJ. Nutrition during childhood cancer treatment: current understanding and a path for future research. Lancet Child Adolesc Health 2020 Jun;4(6):465-475. [doi: 10.1016/S2352-4642(19)30407-9] [Medline: 32061318]
6. Harrison ME, Norris ML, Robinson A, Spettigue W, Morrissey M, Isserlin L. Use of cyproheptadine to stimulate appetite and body weight gain: a systematic review. Appetite 2019 Jun 1;137:62-72. [doi: 10.1016/j.appet.2019.02.012] [Medline: 30825493]
7. Childs DS, Jatoi A. A hunger for hunger: a review of palliative therapies for cancer-associated anorexia. Ann Palliat Med 2019 Jan;8(1):50-58. [doi: 10.21037/apm.2018.05.08] [Medline: 29860861]
8. Hariyanto TI, Kurniawan A. Appetite problem in cancer patients: pathophysiology, diagnosis, and treatment. Cancer Treat Res Commun 2021;27:100336. [doi: 10.1016/j.ctarc.2021.100336] [Medline: 33607591]
9. Strand M, Gustafsson SA. Mukbang and disordered eating: a netnographic analysis of online eating broadcasts. Cult Med Psychiatry 2020 Dec;44(4):586-609. [doi: 10.1007/s11013-020-09674-6] [Medline: 32277331]
10. Weidner F, Maier JE, Broll W. Eating, smelling, and seeing: investigating multisensory integration and (in)congruent stimuli while eating in VR. IEEE Trans Vis Comput Graph 2023 May;29(5):2423-2433. [doi: 10.1109/TVCG.2023.3247099] [Medline: 37027726]
11. Wang X, Xiao Y, Nam S, et al. Use of mukbang in health promotion: scoping review. J Med Internet Res 2025 Mar 27;27:e56147. [doi: 10.2196/56147] [Medline: 40146986]
12. James MN, Ranasinghe N, Tang A, Oehlberg L. Flavor-videos: enhancing the flavor perception of food while eating with videos. Presented at: IMX '22: Proceedings of the 2022 ACM International Conference on Interactive Media Experiences; Jul 22-24, 2022. [doi: 10.1145/3505284.3529967]

XSL•FO
RenderX

13. Kircaburun K, Yurdagül C, Kuss D, Emirtekin E, Griffiths MD. Problematic mukbang watching and its relationship to disordered eating and internet addiction: a pilot study among emerging adult mukbang watchers. Int J Ment Health Addiction 2021 Dec;19(6):2160-2169. [doi: 10.1007/s11469-020-00309-w]
14. Xiao Y, Zhong T, Li L, et al. Psychological and physiological experiences of youth watching Mukbang in China: a qualitative study. Child Youth Serv Rev 2024 Aug;163:107729. [doi: 10.1016/j.childyouth.2024.107729]
15. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
16. Stone AA, Shiffman S. Capturing momentary, self-report data: a proposal for reporting guidelines. Ann Behav Med 2002;24(3):236-243. [doi: 10.1207/S15324796ABM2403_09] [Medline: 12173681]
17. Notice of the general office of the national health commission on the issuance of diagnosis and treatment norms for 12 diseases related to children's blood diseases and malignant tumors (2021 version) [Report in Chinese]. : National Health Commission of the People's Republic of China; 2021 URL: https://wjw.xinjiang.gov.cn/hfpc/ylzl/202105/78e6de081cd14b289c4e624225c16355.shtml [accessed 2025-07-01]
18. Pirla S, Taquet M, Quoidbach J. Measuring affect dynamics: an empirical framework. Behav Res Methods 2023 Jan;55(1):285-300. [doi: 10.3758/s13428-022-01829-0] [Medline: 35381955]
19. von Ash T, Huynh R, Deng C, White MA. Associations between mukbang viewing and disordered eating behaviors. Int J Eat Disord 2023 Jun;56(6):1188-1198. [doi: 10.1002/eat.23915] [Medline: 36807921]
20. Williams PD, Schmideskamp J, Ridder EL, Williams AR. Symptom monitoring and dependent care during cancer treatment in children: pilot study. Cancer Nurs 2006;29(3):188-197. [doi: 10.1097/00002820-200605000-00004] [Medline: 16783117]
21. Xia W, Li HCW, Lam KWK, et al. The impact of hematologic cancer and its treatment on physical activity level and quality of life among children in mainland China: a descriptive study. Cancer Nurs 2019;42(6):492-500. [doi: 10.1097/NCC.0000000000000661] [Medline: 30433896]
22. Halliday V, Porock D, Arthur A, Manderson C, Wilcock A. Development and testing of a cancer appetite and symptom questionnaire. J Hum Nutr Diet 2012 Jun;25(3):217-224. [doi: 10.1111/j.1365-277X.2012.01233.x] [Medline: 22348402]
23. Zhao X, Yan C, Du H, Xie Y, Luo B, Li Z. Reliability and validity research of Chinese version of Cancer Appetite and Symptom Questionnaire [Article in Chinese]. Cancer Res Prev Treat 2018;45(11):874-878. [doi: 10.3971/j.issn.1000-8578.2018.18.0194]
24. Secker DJ, Jeejeebhoy KN. Subjective global nutritional assessment for children. Am J Clin Nutr 2007 Apr;85(4):1083-1089. [doi: 10.1093/ajcn/85.4.1083] [Medline: 17413109]
25. Peng T, Zhao Y, Fu C, et al. A study of validity and reliability for subjective global nutritional assessment in outpatient children with cerebral palsy. Nutr Neurosci 2022 Dec;25(12):2570-2576. [doi: 10.1080/1028415X.2021.1990463] [Medline: 34663203]
26. Vetsch J, Wakefield CE, Robertson EG, et al. Health-related quality of life of survivors of childhood acute lymphoblastic leukemia: a systematic review. Qual Life Res 2018 Jun;27(6):1431-1443. [doi: 10.1007/s11136-018-1788-5] [Medline: 29372438]
27. Eiser C, Vance YH, Horne B, Glaser A, Galvin H. The value of the PedsQLTM in assessing quality of life in survivors of childhood cancer. Child Care Health Dev 2003 Mar;29(2):95-102. [doi: 10.1046/j.1365-2214.2003.00318.x] [Medline: 12603354]
28. Aflaki K, Vigod S, Ray JG. Part II: A step-by-step guide to latent class analysis. J Clin Epidemiol 2023 Jul;159:348-351. [doi: 10.1016/j.jclinepi.2023.05.025] [Medline: 37286148]
29. Liu M, Zhuang A, Norvilitis JM, Xiao T. Usage patterns of short videos and social media among adolescents and psychological health: a latent profile analysis. Comput Human Behav 2024 Feb;151:108007. [doi: 10.1016/j.chb.2023.108007]
30. Sacramento JT, Menezes CD, Brandão MD, Broilo MC, Vinholes DB, Raimundo FV. Association between time of exposure to screens and food consumption of children aged 2 to 9 years during the COVID-19 pandemic. Rev Paul Pediatr 2022;41:e2021284. [doi: 10.1590/1984-0462/2023/41/2021284] [Medline: 36102401]
31. Meryk A, Kropshofer G, Hetzer B, et al. Use of daily patient-reported outcome measurements in pediatric cancer care. JAMA Netw Open 2022 Jul 1;5(7):e2223701. [doi: 10.1001/jamanetworkopen.2022.23701] [Medline: 35881395]
32. Wu Y, Kemps E, Prichard I. Digging into digital buffets: a systematic review of eating-related social media content and its relationship with body image and eating behaviours. Body Image 2024 Mar;48:101650. [doi: 10.1016/j.bodyim.2023.101650] [Medline: 38039952]
33. Killgore WDS, Young AD, Femia LA, Bogorodzki P, Rogowska J, Yurgelun-Todd DA. Cortical and limbic activation during viewing of high- versus low-calorie foods. Neuroimage 2003 Aug;19(4):1381-1394. [doi: 10.1016/s1053-8119(03)00191-5] [Medline: 12948696]
34. Kawai N, Guo Z, Nakata R. A human voice, but not human visual image makes people perceive food to taste better and to eat more: "social" facilitation of eating in a digital media. Appetite 2021 Dec 1;167:105644. [doi: 10.1016/j.appet.2021.105644] [Medline: 34416287]
35. Freitas A, Albuquerque G, Silva C, Oliveira A. Appetite-related eating behaviours: an overview of assessment methods, determinants and effects on children's weight. Ann Nutr Metab 2018;73(1):19-29. [doi: 10.1159/000489824] [Medline: 29843129]

XSL•FO
RenderX

36. Reichenberger J, Kuppens P, Liedlgruber M, et al. No haste, more taste: an EMA study of the effects of stress, negative and positive emotions on eating behavior. Biol Psychol 2018 Jan;131:54-62. [doi: 10.1016/j.biopsycho.2016.09.002] [Medline: 27654506]
37. Lin J, Portingale J, Krug I. Virtual feasts and their impact: a pilot study on the influence of mukbang videos on eating behaviours, body image, and emotional wellbeing. Body Image 2026 Mar;56:102004. [doi: 10.1016/j.bodyim.2025.102004] [Medline: 41308585]
38. Bahari H, Ghanam A, Hajaj H, Elouali A, Babakhouya A, Rkain M. Chemotherapy-induced mucositis in pediatric oncology: experience from 145 cases at the Mohammed VI University Hospital, Oujda. Cureus 2025 Apr;17(4):e82885. [doi: 10.7759/cureus.82885] [Medline: 40416099]
39. Gomes CC, Silva CCGD, Nascimento PRPD, et al. Nutritional status and appetite-regulating hormones in early treatment of acute lymphoblastic leukemia among children and adolescents: a cohort study. Sao Paulo Med J 2020 Mar;138(2):118-125. [doi: 10.1590/1516-3180.2019.0307.r1.19112019] [Medline: 32491080]
40. Li L, Ling ZQ. Mechanisms of cancer cachexia and targeted therapeutic strategies. Biochim Biophys Acta Rev Cancer 2024 Nov;1879(6):189208. [doi: 10.1016/j.bbcan.2024.189208] [Medline: 39542382]
41. Ferrari M, Ripamonti CI, Hulbert-Williams NJ, Miccinesi G. Relationships among unmet needs, depression, and anxiety in non-advanced cancer patients. Tumori 2019 Apr;105(2):144-150. [doi: 10.1177/0300891618765546] [Medline: 29714666]
42. Clarke E, Pugh G, van den Heuvel E, et al. Understanding the patient and family experience of nutrition and dietetic support during childhood cancer treatment. Support Care Cancer 2023 May 8;31(6):326. [doi: 10.1007/s00520-023-07787-3] [Medline: 37154869]
43. Houben K, Dibbets P. Taming temptations: comparing the effectiveness of counterconditioning and extinction in reducing food cue reactivity. Appetite 2025 Apr 1;208:107932. [doi: 10.1016/j.appet.2025.107932] [Medline: 40020972]

Abbreviations

CASQ: Cancer Appetite and Symptom Questionnaire
GEE: generalized estimating equation
LCA: latent class analysis
PedsQL: Pediatric Quality of Life Inventory
SGNA: Subjective Global Nutritional Assessment
TRSC-C: Therapy-Related Symptom Checklist for Children

Edited by M Balcarras; submitted 19.Jul.2025; peer-reviewed by B Satravada, SL Coven; revised version received 29.Apr.2026; accepted 29.Apr.2026; published 22.May.2026.

Please cite as:
Xiao Y, Wang J, Zhong T, Wang X, Tang D, Xia W
Associations Between Mukbang Watching and Appetite, Nutrition, and Quality of Life in Pediatric Patients with Cancer: Intensive Longitudinal Study
J Med Internet Res 2026;28:e80932
URL: https://www.jmir.org/2026/1/e80932
doi:10.2196/80932

© Yuxue Xiao, Jiani Wang, Ting Zhong, Xiao Wang, Dongyan Tang, Wei Xia. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Thematic Mapping and Evolution of Social Media Mining in Health Research: Hybrid Bibliometric Synthesis

Mia Jiming Yang, MSc; Sabine Bohnet-Joschko, Prof Dr

Chair of Management and Innovation in Healthcare, Faculty of Management, Economics and Society, Witten/Herdecke University, Alfred-Herrhausen-Str. 50, Witten, North Rhine-Westphalia, Germany

Corresponding Author:
Mia Jiming Yang, MSc
Chair of Management and Innovation in Healthcare, Faculty of Management, Economics and Society, Witten/Herdecke University, Alfred-Herrhausen-Str. 50, Witten, North Rhine-Westphalia, Germany

Related Article:

Comment in: https://www.jmir.org/2026/1/e102159

Abstract

Background: Social media platforms offer extensive data, as they are widely used globally. Social media mining (SMM) enables real-time monitoring of user-reported health information and serves as a supplement to traditional health data analytics. However, the rapid proliferation of literature has produced fragmentation, and a comprehensive knowledge map regarding SMM is lacking. Further, existing bibliometric reviews in health fields are easily undermined by synonym fragmentation and parameter settings, reducing their robustness. Thus, a more robust, reproducible, and decision-oriented bibliometric framework is required.

Objective: This study aimed to (1) outline key thematic clusters in health-related SMM and map their dynamic evolution, and (2) methodologically demonstrate how machine learning–based bibliometric analysis can strengthen the robustness, transparency, and foresight capacity of evidence synthesis.

Methods: This study designed a fully automated and reproducible bibliometric analysis of PubMed journal articles published from 2015 to 2025 (n=250) and analyzed records with both abstracts and keywords (n=189). We performed cleaning and standardization for titles, abstracts, author keywords, and MeSH terms, and carried out an exploratory descriptive analysis to obtain preliminary insights into publication patterns. Subsequently, we used SPECTER2 and PubMedBERT embeddings with keywords and abstracts to construct a hybrid similarity matrix. Then, we applied Uniform Manifold Approximation and Projection for dimensionality reduction, followed by Hierarchical Density-Based Spatial Clustering of Applications with Noise for thematic clustering, and visualized the results in a 3D strategic coordinate system (maturity, influence, and recency). We performed intercluster relationship analysis and time-slice analysis to examine thematic intersections and evolution. To ensure robustness and enhance interpretability, we implemented dual-level validation.

Results: We identified 6 thematic clusters: cluster 1 (candidate incubator pool of peripheral cross-cutting topics in health-related SMM), cluster 2 (computational methods in health informatics), cluster 3 (public attitudes and sociopsychological determinants), cluster 4 (infodemiology and the COVID-19 information ecosystem), cluster 5 (health communication and public health engagement), and cluster 6 (social media analysis and network methods). Strategic 3D mapping revealed that methodological clusters (clusters 2 and 6) occupied high-maturity and high-influence positions, while application-driven clusters (clusters 3 and 4) occupied high-influence and high-recency positions, representing rapidly expanding frontiers. Clusters 1 and 5 demonstrated strong potential for further growth. Temporal slicing confirmed a trajectory moving from methodological consolidation and thematic diversification to a renewed focus on convergence and problem-solving. Validation showed strong semantic coherence and robustness of the methods and findings.

Conclusions: We developed a semantic-structural hybrid bibliometric framework with dual-level validation, reducing synonym fragmentation and parameter sensitivity inherent in traditional approaches. The resulting decision-oriented knowledge map offers strategic guidance for infodemiology-informed and audience-segmented public health communication, research priority settings, and the deployment and evaluation of real-world surveillance and pharmacovigilance workflows while supporting evidence-driven and patient-centered decision-making in public health and health care.

(J Med Internet Res 2026;28:e86200) doi:10.2196/86200

KEYWORDS

social media mining; bibliometrics; machine learning; strategic mapping; health research

XSL•FO
RenderX

Introduction

Social media has become a ubiquitous component of modern life, with over 5 billion people worldwide using social platforms as of February 2025 [1]. The value of social media data in health research lies in its volume, velocity, and variety, and the insights derived from user-reported content. Unlike traditional historical data, social media offers a large-scale, real-time representation of the natural state of public discourse and behavior. Content from platforms, such as Twitter and Reddit, constitutes a rich data source for gaining insights into population health trends, behaviors, and opinions [2-4]. Researchers have recognized that the computational analysis of social media data—often referred to as social media mining (SMM)—enables real-time health monitoring and health-related interactions that are unattainable through traditional data sources [5]. In fact, SMM is flourishing as an emerging science aimed at providing health care stakeholders with additional evidence for decision-making.

Over the past 5 years, there has been explosive growth in digital health research utilizing social media data, spanning areas such as epidemiological studies and pharmacovigilance [5-8]. Moreover, SMM techniques have been applied in health communication and promotion (eg, evaluating public responses to health campaigns), mental health monitoring (eg, identifying depression-related discourse online), and behavioral research (eg, analyzing lifestyle-related topics) [2,9,10]. Mapping such a fragmented and rapidly evolving field has become especially important. Therefore, a literature review in this area can provide an overview of existing research.

Four reviews have previously addressed the use of social media and the analysis of social media data for health purposes. A review of the application of SMM in health outcomes included 19 papers [11]. However, it was published in 2019, prior to the major expansion of social media data mining, and only provided an overview of literature related to medication and treatment side effects. The authors suggested at the time that adverse drug events and health-related quality of life would be future research directions. Another review on social text mining conducted in the early stages of the pandemic focused on the lack of ethical consensus and guidelines in this emerging field, thus emphasizing ethical standards in the use of social media text data analysis [12]. Subsequently, a 2021 review of the literature on social media for health purposes identified at least 10 applications in the health domain—spanning health interventions, public communication, surveillance, education, and research—highlighting the breadth of the field [13]. While this review did not provide an overview of SMM, it illustrated the diverse potential application scenarios of SMM. Another scoping review on the analysis of social media data in health care searched 5 academic databases, including PubMed, Web of Science, and Embase, and ultimately included 134 papers on social media data analysis [14]. However, only 55 of these papers fell under the category of SMM, with the rest being traditional qualitative studies. Although this review only briefly described the methods used in SMM in the section on computational tools, it pointed out that future analysis of social media data will inevitably move in a technology-driven direction.

In terms of research methodology, existing bibliometric reviews on health research and medical innovation predominantly follow 2 typical approaches. The first approach uses co-word networks and thematic mapping based on keyword co-occurrence, combined with descriptive metrics, such as publication volume, country- and institution-level collaboration, and core journals, to summarize the research landscape [15-22]. The second approach uses topic modeling techniques, such as latent Dirichlet allocation (LDA) and structured topic modeling, to extract latent themes from textual data [15,21,23-28]. While valuable, these methods often exhibit 2 limitations when applied to rapidly evolving research fields. First, pure co-occurrence statistics rely excessively on string sequences and frequency counts, frequently overlooking synonymy, near-synonymy, and conceptual fragmentation caused by variations in expression, and the results are highly sensitive to threshold selection [24,29-31]. Furthermore, traditional topic models typically rely on the bag-of-words assumption and require predefined topic counts, potentially compromising comparability and robustness across different time windows [23,24,29]. Addressing these methodological gaps, this study proposes a hybrid bibliometric framework integrating semantic and structural analysis [30]. We introduce semantic similarity while preserving the co-occurrence network structure to mitigate synonym fragmentation [30]. Aligned with this approach, a recent study has integrated structural and semantic information within citation networks to more precisely characterize the influence of publications and their capacity for interdisciplinary knowledge diffusion [32]. Density-based clustering enables topic cluster identification without prespecifying the cluster numbers [33]. By integrating time slicing, burst detection, and multidimensional strategic coordinates (maturity, influence, and recency), we systematically characterize a topic's knowledge accumulation, real-world impact, and recent growth potential. This provides a more decision-oriented, reproducible evidentiary landscape for the field. To ensure the robustness and reproducibility of this hybrid framework, we further detail the rationale for key parameter settings in the Methods section and demonstrate the validity and stability of the methodology and results through multiple validation tests.

In summary, prior studies have involved highly diverse application scenarios of social media data and have made fragmented statements. While previous reviews have cataloged applications and thematic summarization, they have rarely integrated advanced computational approaches for a structured synthesis. In particular, bibliometric reviews often rely on co-occurrence statistics or topic models with limited robustness. This leads to a limitation in sustaining a prospective strategic vision. To address this, our study developed an automated, reproducible pipeline for bibliometric analysis of SMM in health care, introducing several methodological innovations.

Furthermore, this study aimed to not only identify hot topics but also trace their temporal evolution. Additionally, this study attempted to explore the future hotspots of research innovation and reveal the composition of the innovation incubation pool, thereby helping researchers, funding agencies, and policymakers to anticipate future directions and foster interdisciplinary integration. Thus, this study had the following questions:

XSL•FO
RenderX

1. What are the main topics of research on SMM in health care and their latent patterns?
2. What are their developmental stages, their strategic positioning, and the future potential research themes?

To answer these questions, we used the following guiding goals: (1) substantively outline key thematic clusters in health-related SMM, map their pre-, during-, and postexplosive growth period evolution, and evaluate the implications for fostering social equity, effective health management, and informed policy, and (2) methodologically demonstrate how advanced bibliometric methods can strengthen the robustness, transparency, and foresight capacity of evidence synthesis.

Methods

Study Design

This study aimed to systematically review and analyze the research progress and thematic evolution of SMM in the context of health by constructing a fully automated and reproducible bibliometric analysis framework. To achieve this objective, we designed a multistage methodological pathway in accordance with BIBLIO (Preliminary Guideline for Reporting Bibliometric Reviews of the Biomedical Literature) [34]. Figure S1 in Multimedia Appendix 1 illustrates the research process. The methods are elaborated in more detail in Figure S2 in Multimedia Appendix 1. Additionally, the BIBLIO checklist has been provided in Checklist 1.

Data Retrieval and Processing

The first phase of this study involved constructing a comprehensive literature dataset. We selected the PubMed database as the primary data source, as it is one of the most extensive and authoritative bibliographic databases in the field of health research. We summarized the search strategy, record identification, and inclusion process using an adapted PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram. Following the PRISMA-S extension for reporting literature searches, we have presented the search process in Checklist 2 to strengthen the transparency and reproducibility of the search [35]. To ensure the comprehensiveness of the search, we used a search strategy utilizing Boolean logic operators to combine free-text terms and Medical Subject Headings (MeSH) terms. The query included terms for social media platforms (eg, “social media,” “Twitter,” and “Reddit”), combined with terms related to data mining and analysis (eg, “mining” and “data mining”). We also applied filters to include only journal articles from January 01, 2015, to July 31, 2025 (excluding news and retracted articles). The literature download process was automated using the Entrez module. Necessary parameters, including *Entrez.email* and *Entrez.api_key*, were configured to comply with National Institutes of Health application programming interface (API) access guidelines and prevent connection errors during high-frequency requests. All retrieved results were stored to ensure transparency and traceability throughout the research process. Furthermore, to mitigate data loss due to API updates or network instability, we implemented comprehensive logging after each batch download and recorded completion timestamps, the search strategy used, the number of records returned, and other relevant metadata, thereby guaranteeing data integrity and reproducibility.

In addition to the metadata obtained from PubMed, this study integrated supplementary metrics from external databases and official journal websites. Specifically, we collected the H-index of relevant journals from the SCImago database to reflect their academic prestige and long-term impact. Concurrently, the most recently released impact factor (IF) and 5-year IF were acquired from official journal websites and matched with the publication records. Then, we retrieved the relative citation ratio (RCR) of the included articles, which is a normalized measure of citation influence across fields and time. The RCR was used only for post-hoc external comparison and was not included in embedding, clustering, parameter selection, or labeling. The integration of these supplementary data laid a solid foundation for subsequent external validity checks.

During data processing, we extracted the core metadata, including title, keywords, abstract, publication year, journal, etc. Then, we performed keyword standardization by unifying plurals and synonyms, expanding abbreviations, and mapping entry terms to MeSH descriptors with the NLM MeSH descriptor dataset (accessed on July 31, 2025; latest record revision date: July 15, 2025).

Exploratory Descriptive Analysis

After processing the bibliographic dataset, we conducted an exploratory descriptive analysis to characterize SMM in health research. This stage provided a phenomenon-level overview to guide subsequent structural modeling. We explored annual publication trends with autoregressive integrated moving average (ARIMA)-based exploratory projection, collaboration networks between institutions and researchers, countries, citation impact, and keyword dynamics (keyword statistics and Kleinberg burst detection). Overall, this multifaceted descriptive analysis established an overview of the research. The methods of keyword burst detection and parameter robustness assessment are described in Multimedia Appendix 2.

Thematic Clustering and Strategic Mapping

This study proposes an automated bibliometric pipeline specifically designed for literature on SMM in the health domain. Although we performed normalization, residual terminological variations may still exist. Therefore, we developed an automated process integrating structural and semantic similarity, where semantic proximity is derived from contextual embeddings (PubMedBERT and SPECTER2) computed from titles and abstracts to help reduce fragmentation caused by near-synonymous or variably expressed terms and to capture contextual semantics. A hybrid similarity matrix was reduced with Uniform Manifold Approximation and Projection (UMAP) and clustered via Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) under multiple parameter profiles. Outputs included interpretable topic clusters with representative keywords and their positions in a 3D strategic diagram (maturity, influence, and recency). In this study, the number of papers published for a keyword represents maturity, the number of connections a keyword has in co-occurrence networks represents influence, and the proportion

XSL•FO
RenderX

of papers published within the last 5 years represents recency. Details of parameter tuning, thresholds, and ablation tests are provided in Multimedia Appendix 3. After clustering, we performed an intercluster relationship analysis on the weighted keyword co-occurrence network to quantify cross-theme connectivity. We computed a cluster-to-cluster coupling metric based on the total cross-cluster co-occurrence strength, ranked cluster pairs to identify the most strongly connected themes, and interpreted these links (information is provided in the Results section). To explain what drives these cross-cluster connections, we further identified pair-specific bridging keywords contributing the most to the coupling between each highly coupled cluster pair.

Dynamic Temporal Slicing

For the dynamic analysis, the study period was divided into 3 slices with the cutoff of explosive growth: 2015 - 2019, 2020 - 2023, and 2024 - 2025. Within each slice, we performed spectral clustering on the keyword co-occurrence network. This allowed us to observe the emergence, growth, and decline of research themes over time. Details of the process are provided in Multimedia Appendix 3.

Validation

This study conducted either macro-level or micro-level validation. To assess the robustness of our methods, we conducted internal validation through a series of sensitivity analyses. Moreover, we validated the findings with external checks using RCR alignment.

In addition, we conducted micro-level interpretive triangulation through evidence mapping and expert content analysis. The 3D strategic diagram was divided into 8 quadrants based on a reference point (x=0, y=3.51, z=0.5). From each cluster in each quadrant, we sampled 1 representative keyword that was the farthest from the reference point. For each sampled keyword, we retrieved 2 full-text articles: 1 with the highest RCR and 1 published in a journal with the highest IF. Two reviewers examined the full texts to extract evidence on research introduction, methods, social media sites, contributions, challenges, and future outlook. The protocols for the analysis of the selected articles are described in Multimedia Appendix 4. This interpretive triangulation was performed after the mathematical analyses were finalized and was not used to fine-tune parameters, select models, modify cluster assignments, or remove records and keywords.

Tools and Software Utilized

All analyses were conducted primarily in Python (via PyCharm IDE), with the exception of burst detection, which was implemented in RStudio (Posit). Microsoft Excel was used for supplementary manual inspection and verification of selected outputs, and Microsoft Office Visio and Draw.io were used to draw flow charts.

Ethical Considerations

This study was a bibliometric and secondary analysis of bibliographic data about academic publications from an open-access database. Neither humans nor animals were involved in this work. Therefore, ethical approval was not required. Moreover, the original data for this study consisted entirely of publicly available bibliographic information. During the research process, we did not collect any individual-level participant data or recruit human participants. Therefore, informed consent was not required, and no compensation was provided. Furthermore, the study exclusively processed publicly available article-level metadata and reported the results in summary form. None of the data collected, processed, or included in the manuscript text and supplementary materials contain any personally identifiable information. Consequently, there is no risk of individual privacy disclosure, and no additional authorization or consent documents are required.

Results

Publications Identified

In the initial stage, we retrieved 250 publications from PubMed covering the period between January 1, 2015, and July 31, 2025. After the exclusion of articles without an abstract or author-supplied keywords, a total of 189 publications were retained for systematic analysis (Multimedia Appendix 5). The PRISMA flow diagram is presented in Figure 1. Moreover, details of the search strategy and variations of the key search terms are presented in Figure 1.

XSL•FO
RenderX

Figure 1. Adapted PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for bibliometric dataset construction using PubMed. Exclusions were based on objective metadata completeness (missing author keywords/abstract), without manual title/abstract screening or full-text eligibility assessment. The image has been reproduced from Page et al [36], which is published under Creative Commons Attribution 4.0 International License [37].

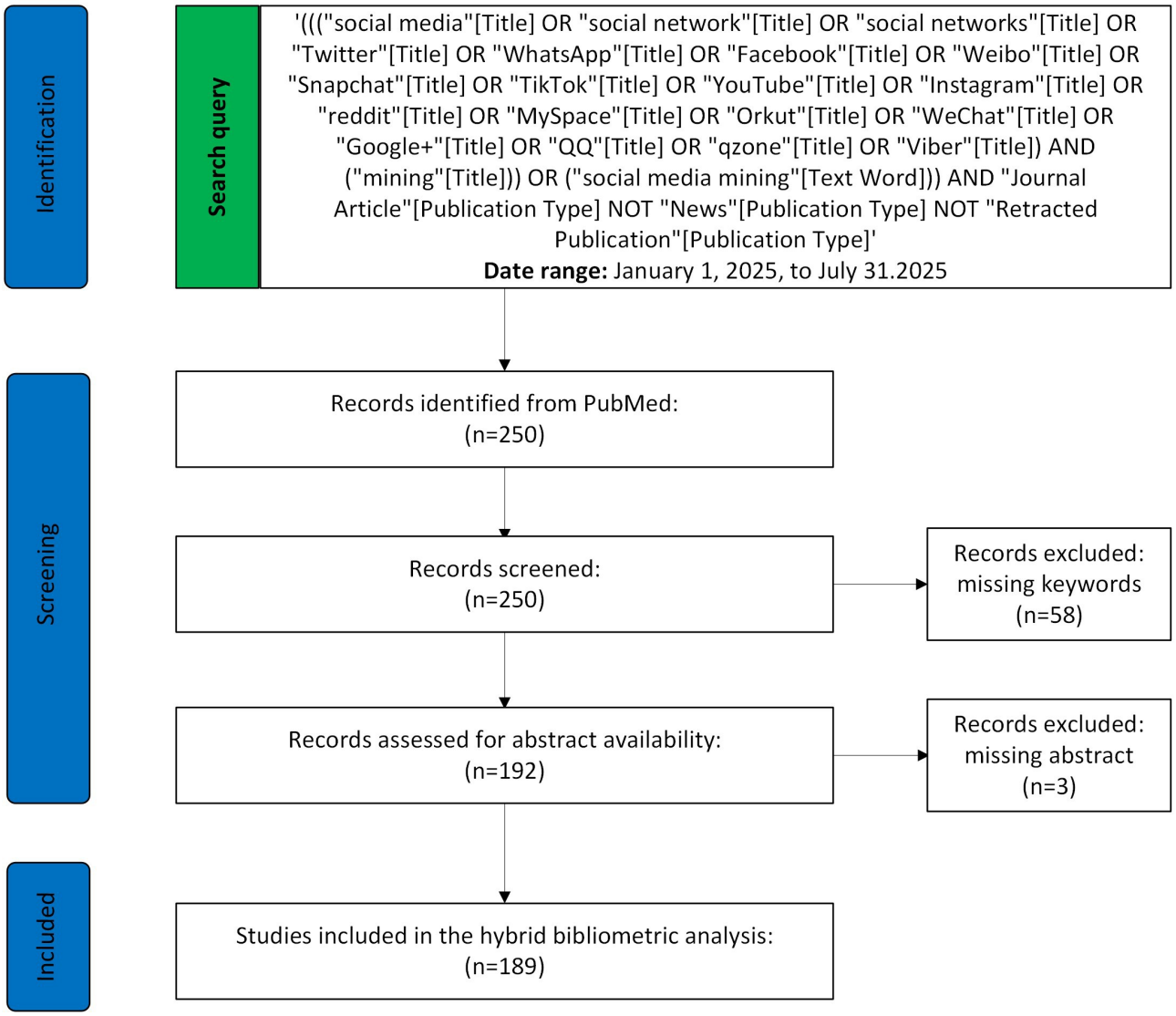

Exploratory Descriptive Analysis

As shown in Figure 2 and Table S1 in Multimedia Appendix 6, annual publication trends in SMM exhibited a consistent upward trajectory over the past decade. Between 2015 and 2024, the annual number of PubMed-indexed articles averaged 17.4. By July 31, 2025, 15 publications had already been recorded. Using an ARIMA-based exploratory projection fitted to the annual series, we estimated approximately 32 publications for 2025, with a wide 95% prediction interval (18-45) reflecting uncertainty due to the short annual time series. Overall, these findings indicate sustained growth and underscore the growing scholarly interest in the field.

XSL•FO
RenderX

Figure 2. Annual publication trends and autoregressive integrated moving average (ARIMA)-based exploratory projection (January 1, 2015, to July 31, 2025, including the predicted number for 2025). In this study, each year is treated as a single observation, resulting in a short time series based on annual aggregation. The ARIMA model serves only as an exploratory projection, yielding a broad prediction interval. The predicted counts for 2025 should be interpreted with caution.

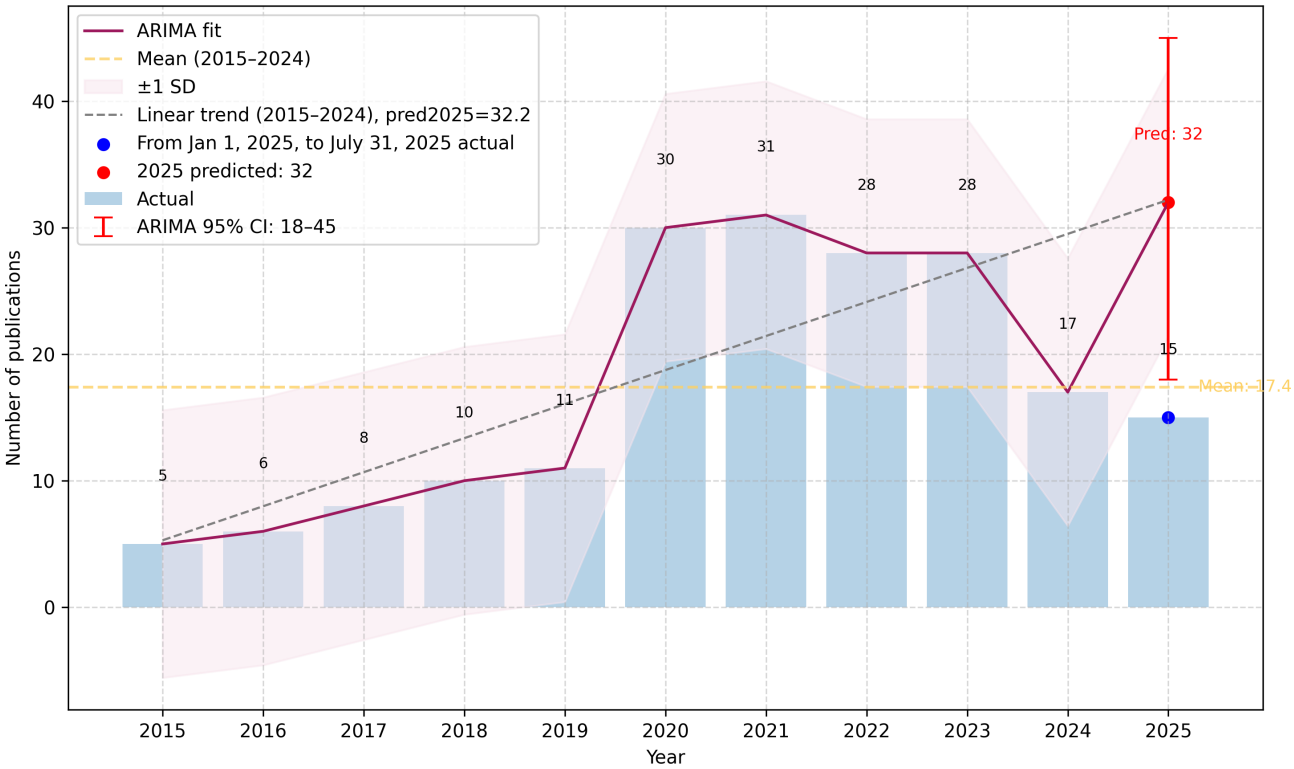

The distributions of countries, institutions, and journals are summarized in Table 1. The United States was the leading contributor (54/189, 28.6%), followed by China (27/189, 14.3%), Italy (11/189, 5.8%), and France (10/189, 5.3%). Among institutions, the University of Texas at Austin (5/189, 2.7%) and Kazan Federal University (3/189, 1.6%) appeared most frequently. The Journal of Medical Internet Research was the dominant journal (21/189, 11.1%), followed by the International Journal of Environmental Research and Public Health (7/189, 3.7%) and JMIR Public Health and Surveillance (7/189, 3.7%). Beyond these core contributors, a long tail of less-represented countries, institutions, and journals reflects the field's broadening boundaries and thematic heterogeneity.

XSL•FO
RenderX

Table . Top 10 entities (countries, affiliations, and journals) by frequency in social media mining publications.

Variable	Value (N=189), n (%)
Country	
United States	54 (28.6)
China	27 (14.3)
Italy	11 (5.8)
France	10 (5.3)
United Kingdom	7 (3.7)
Canada	6 (3.2)
Korea	6 (3.2)
Republic of China (Taiwan)	5 (2.7)
Spain	5 (2.7)
Japan	4 (2.1)
Affiliation	
Center for Health Communication, The University of Texas at Austin, Austin, Texas, United States	5 (2.7)
Kazan Federal University, Kazan, Russian Federation	3 (1.6)
Center Régional de Pharmacovigilance, Hôpital Européen Georges-Pompidou, Paris, France	2 (1.1)
College of Nursing and Public Health, Adelphi University, Garden City, New York, United States	2 (1.1)
Department of Psychiatry, Singapore General Hospital, Singapore	2 (1.1)
Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Boston, Massachusetts, United States	2 (1.1)
Division of Rheumatology, Allergy and Immunology, Massachusetts General Hospital, Boston, Massachusetts, United States	2 (1.1)
Institute of Informatics and Telematics (IIT), National Research Council (CNR), Pisa, Italy	2 (1.1)
Kap Code, Paris, France	2 (1.1)
National Research University Higher School of Economics, Moscow, Russian Federation	2 (1.1)
Journal	
Journal of Medical Internet Research	21 (11.1)
International Journal of Environmental Research and Public Health	7 (3.7)
JMIR Public Health and Surveillance	7 (3.7)
Journal of Biomedical Informatics	7 (3.7)
Frontiers in Psychology	6 (3.2)
Journal of the American Medical Informatics Association: JAMIA	6 (3.2)
PeerJ Computer Science	6 (3.2)
Studies in Health Technology and Informatics	6 (3.2)
JMIR Infodemiology	4 (2.1)
Social Network Analysis and Mining	4 (2.1)

The author collaboration network in Figure S3 in Multimedia Appendix 1 displays a pattern in which a few highly connected researchers, such as Graciela Gonzalez-Hernandez, Davy Weissenbacher, and Abeed Sarker, occupy central hub positions, maintaining multiple cross-institutional collaborations. Their cooperative studies are summarized in Tables S2 and S3 in Multimedia Appendix 6, highlighting broader collaboration portfolios, notably those of Nathalie Texier and colleagues. Overall, the field is shaped by a small number of core investigators complemented by a long tail of occasional

XSL•FO
RenderX

contributors. Importantly, this authorship structure is mirrored in citation patterns, which further reflect the influence and reach of these core groups.

Citation analyses highlighted both globally influential work and highly interconnected contributions within the field. As shown in Figure S4 in Multimedia Appendix 1, the ranking of publications by RCR identified a set of cornerstone articles, with articles by Ayyoubzadeh et al [38] on COVID-19 trend prediction and Nikfarjam et al [39] on adverse drug reactions consistently leading the field. Other high-RCR papers addressed vaccine hesitancy, mental health, and methodological advances, such as machine learning (ML) and natural language processing (NLP), reflecting both topical diversity and methodological innovation among the most impactful studies. Within the internal citation network (Figure S5 in Multimedia Appendix 1), the article by Nikfarjam et al [39] again occupied the central position with the most citations, followed by articles by Tapi Nzali et al [40] on breast cancer and Lazard et al [41] on public communication. Several COVID-19–related studies (eg, Li et al [42] and Zhang et al [43]) also ranked prominently, underscoring the pandemic's role in driving recent scholarly influence.

The keyword statistics in Table 2 and Figure S6 in Multimedia Appendix 1 revealed both stable core terms and dynamically emerging concepts. As shown in Figure S6 in Multimedia Appendix 1, "social media," "data mining," and "natural language processing" consistently ranked among the most frequent keywords forming a foundation in SMM. In contrast, the occurrence of terms, such as "COVID-19," "vaccine hesitancy," and "mental health," increased sharply in specific years, highlighting the responsiveness of the literature to external events and public health concerns. Certain topics (eg, "sentiment analysis") extended beyond 2025, suggesting potential growth.

Table . Top 20 keywords in social media mining publications (2015 - 2025).

Rank	Keyword	Overall frequency, n
1	Social media	84
2	Data mining	58
3	COVID-19	43
4	Natural language processing	36
5	Twitter	34
6	Sentiment analysis	25
7	Social media mining	24
8	Machine learning	22
9	Infodemiology	14
10	Topic modeling	10
11	Deep learning	9
12	Pandemics	9
13	Pharmacovigilance	9
14	Coronavirus	8
15	Mental health	7
16	Big data	6
17	Latent Dirichlet allocation	6
18	Public health	6
19	Reddit	6
20	Social networking	6

We applied the Kleinberg burst detection algorithm to capture topic dynamics. The resulting heatmap (Figure 3) and timeline (Figure 4) identified multiple high-intensity bursts, with peak activity observed in 2020 - 2021. Burst events were not limited to pandemic-related terms, and method-focused keywords, such as "topic modeling" and "ML," indicated that methodological innovation engages in dynamic interaction with real-world demands.

XSL•FO
RenderX

Figure 3. Heatmap of keyword burst events in PubMed literature. Each row represents a keyword identified from PubMed articles, and columns denote 2015 to 2025. The central heatmap displays burst periods detected by the Kleinberg burst detection algorithm under parameter configuration G0.5_W3_L1. Burst levels are shown in 10 discrete tiers with specific hex colors, where darker shades indicate stronger bursts (eg, level ≥18 in navy). Blue stars mark individual burst periods. The top bar plot shows the total number of keyword bursts per year, with counts labeled above each bar. The right bar plot indicates the total number of burst occurrences per keyword across all years. These bursts often correspond to emerging or intensifying topics in biomedical and social media research.

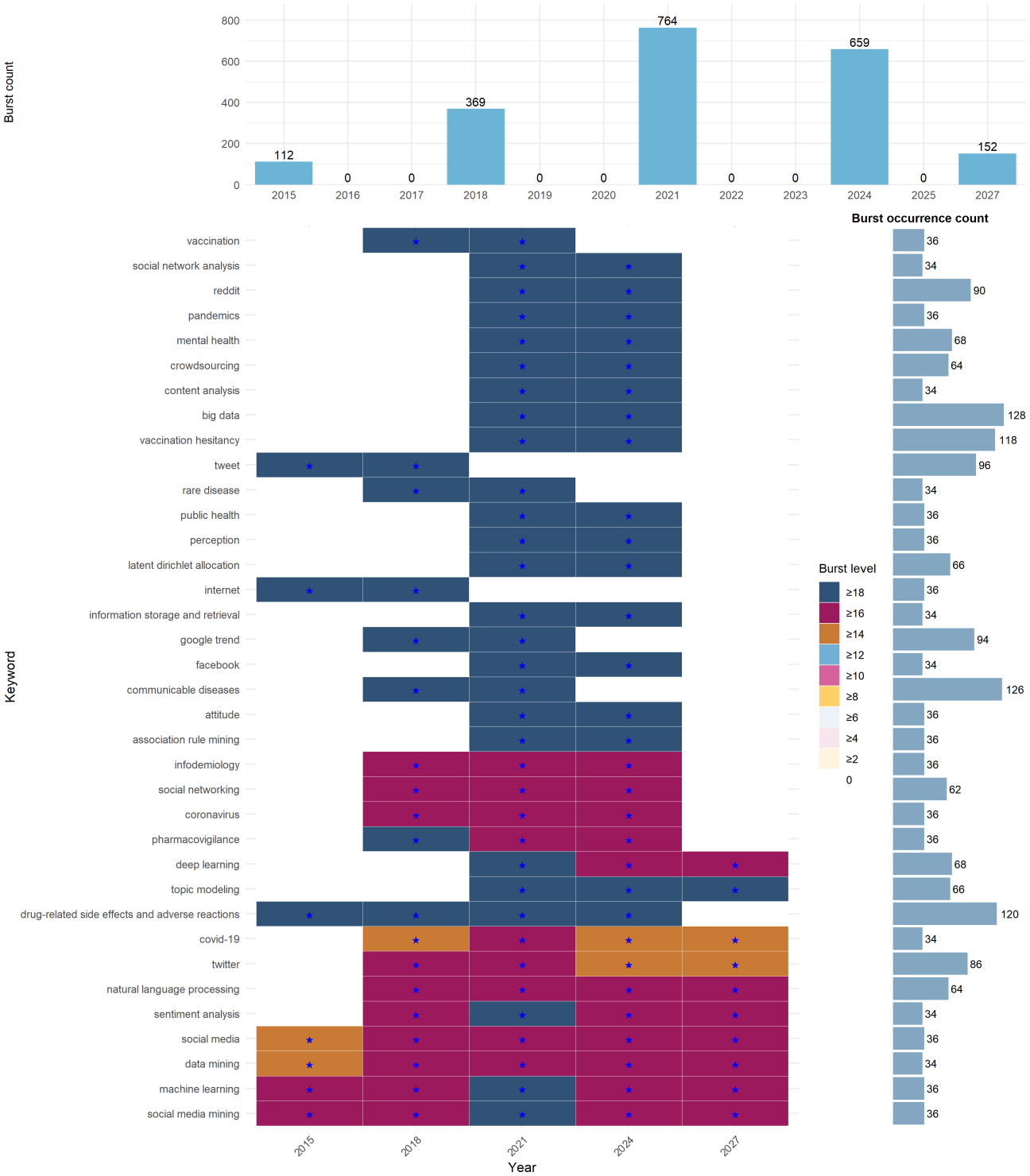

XSL•FO
RenderX

Figure 4. Burst timelines of keywords with burst levels ≥2 in social media mining from 2015 to 2027. The image shows the temporal burst patterns of 37 keywords identified as most significant in terms of burst strength. Each keyword is represented by a horizontal timeline from the start of 2015 to the end of 2027 (blue lines), with red solid lines indicating the duration of burst periods and red dots marking the onset of each burst. If a burst extends beyond the year 2025, the line becomes dashed to signify projected or ongoing trends. The visualization highlights the emergence and persistence of research interests within the field of digital health and social media analytics.

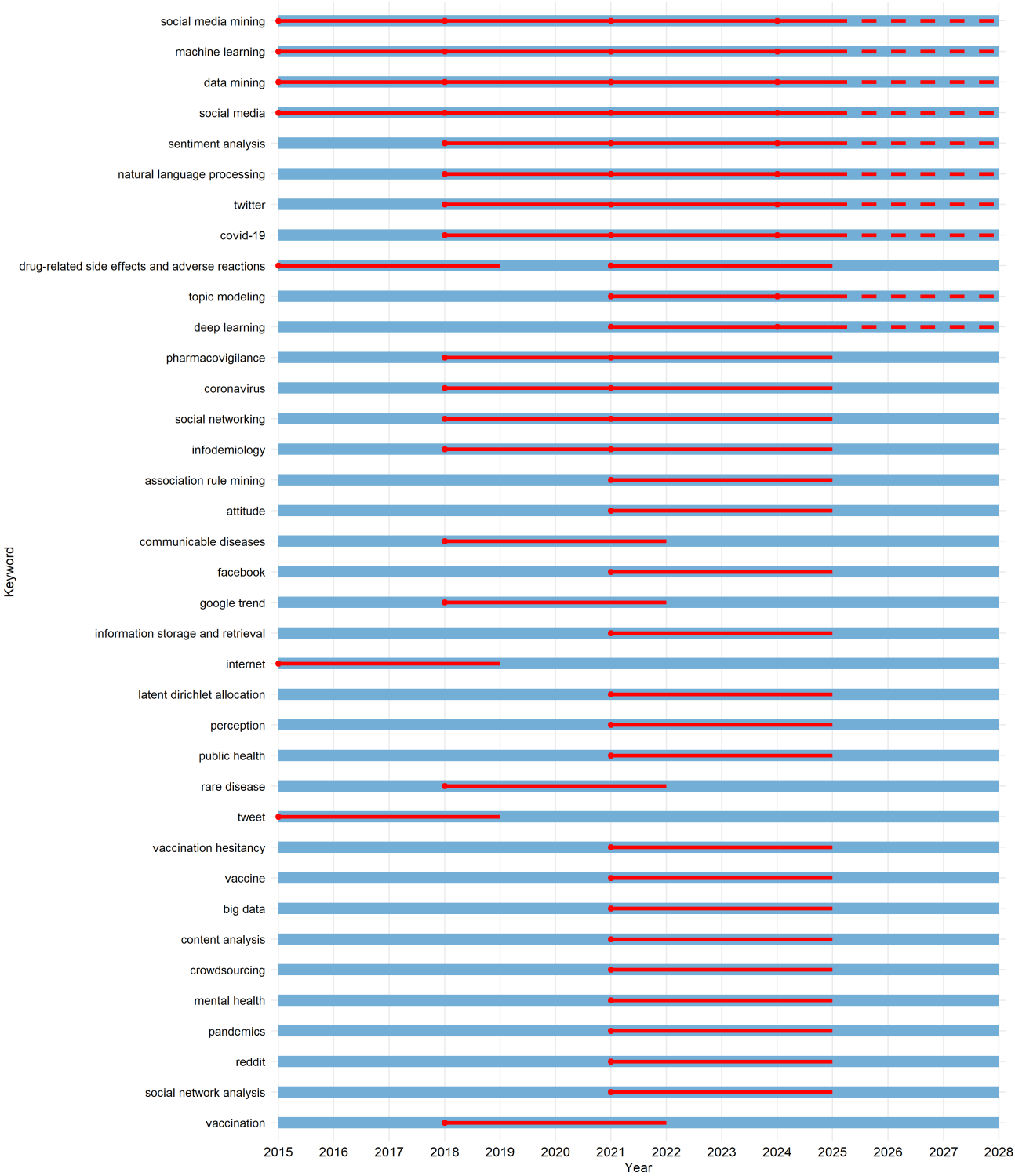

Static Thematic Clustering, Strategic Positioning, and Intercluster Relationships

Overview

A total of 1157 author keywords were present in the included literature. Initially, low-frequency words with an occurrence frequency of ≤2 were removed, along with the 2 most frequently occurring words. This approach prevents noise from affecting the co-occurrence network and clustering results while also avoiding overdominating the network outcomes. Then, to prevent isolated or incidentally collinear edges from affecting structural similarity, a minimum co-occurrence threshold of 1 was set. Only keywords with edges present in both the structural

XSL•FO
RenderX

and semantic spaces were retained. Finally, 78 high-quality keywords (Table 3) were included for subsequent analysis.

XSL•FO
RenderX

Table . Detailed metrics of keywords and clusters.

Cluster and keywords	Freq[a], n	Overall degree	Network strength	Centroid distance[b]	Quadrant
Cluster 1[c] (noise; n=19)					
Mental health	7	11	4.195	2.019	Core emerging hotspot (x≥0, y≥3.51, z≥0.5)
Big data	6	10	3.829	1.903	Potential breakthrough theme (x<0, y≥3.51, z≥0.5)
Public health	6	24	9.121	0.872	Core emerging hotspot (x≥0, y≥3.51, z≥0.5)
Depression	3	5	1.722	2.099	Peripheral mature theme (x≥0, y<3.51, z<0.5)
Google trend	3	7	2.460	1.330	Peripheral mature theme (x≥0, y<3.51, z<0.5)
Perception	3	15	5.166	1.918	Potential breakthrough theme (x<0, y≥3.51, z≥0.5)
Cyberincivility	2	3	1.705	1.758	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Digital health	2	3	1.199	1.822	Peripheral emerging theme (x≥0, y<3.51, z≥0.5)
Drug repositioning	2	6	0.983	2.100	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
LSTM	2	7	2.198	0.980	Peripheral emerging theme (x≥0, y<3.51, z≥0.5)
Medium	2	1	0.492	0.989	Peripheral obsolete theme (x<0, y<3.51, z<0.5)
Neoplasms	2	5	1.979	1.688	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Nurse	2	6	2.240	1.669	Peripheral emerging theme (x≥0, y<3.51, z≥0.5)
Republic of Korea	2	3	0.944	1.542	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Social networking site	2	3	1.705	1.661	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Suicide	2	4	1.426	2.450	Peripheral emerging theme (x≥0, y<3.51, z≥0.5)
Thematic analysis	2	6	2.165	1.401	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Web scraping	2	3	0.757	1.685	Exploratory emerging theme (x<0, y<3.51, z≥0.5)

XSL•FO
RenderX

Cluster and keywords	Freq[a], n	Overall degree	Network strength	Centroid distance[b]	Quadrant
YouTube	2	3	1.127	1.599	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Cluster 2[d] (n=15)					
Natural language processing	36	41	13.465	0.752	Core emerging hotspot (x≥0, y≥3.51, z≥0.5)
Social media mining	24	27	8.489	0.067	Core emerging hotspot (x≥0, y≥3.51, z≥0.5)
Machine learning	22	28	8.820	0.473	Core emerging hotspot (x≥0, y≥3.51, z≥0.5)
Pharmacovigilance	9	8	2.776	0.511	Peripheral mature theme (x≥0, y<3.51, z<0.5)
Information storage and retrieval	5	6	2.311	0.511	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Drug-related side effects and adverse reactions	4	6	1.841	0.404	Peripheral emerging theme (x≥0, y<3.51, z≥0.5)
Rare disease	3	4	1.080	0.452	Peripheral obsolete theme (x<0, y<3.51, z<0.5)
Computational social science	2	2	0.360	0.551	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Epidemiology	2	4	1.285	0.788	Peripheral obsolete theme (x<0, y<3.51, z<0.5)
Patient forum	2	5	1.318	0.469	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Patient-reported outcome measures	2	6	1.959	0.840	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Post-acute COVID-19 syndrome	2	5	1.755	0.925	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Psychological	2	3	0.786	1.153	Peripheral emerging theme (x≥0, y<3.51, z≥0.5)
Quality of life	2	1	0.181	0.507	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Symptom	2	5	2.282	0.939	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Cluster 3[e] (n=8)					
Pandemics	9	23	8.704	0.567	Core emerging hotspot (x≥0, y≥3.51, z≥0.5)
Vaccination	6	18	7.074	0.347	Core emerging hotspot (x≥0, y≥3.51, z≥0.5)
Vaccination hesitancy	4	7	2.706	0.163	Peripheral emerging theme (x≥0, y<3.51, z≥0.5)

XSL•FO
RenderX

Cluster and keywords	Freq[a], n	Overall degree	Network strength	Centroid distance[b]	Quadrant
Attitude	3	14	5.578	0.340	Potential breakthrough theme (x<0, y≥3.51, z≥0.5)
SARS-CoV-2	3	13	5.132	0.150	Potential breakthrough theme (x<0, y≥3.51, z≥0.5)
Human	2	4	1.927	0.480	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Influenza	2	4	1.927	0.480	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Twitter mining	2	4	1.447	0.411	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Cluster 4[f] (n=13)					
COVID-19	43	49	16.294	0.326	Core emerging hotspot (x≥0, y≥3.51, z≥0.5)
Twitter	34	37	13.168	0.829	Core emerging hotspot (x≥0, y≥3.51, z≥0.5)
Infodemiology	14	25	8.831	0.220	Core emerging hotspot (x≥0, y≥3.51, z≥0.5)
Topic modeling	10	18	6.685	0.678	Core emerging hotspot (x≥0, y≥3.51, z≥0.5)
Coronavirus	8	13	5.151	0.173	Core emerging hotspot (x≥0, y≥3.51, z≥0.5)
Tweet	6	15	5.508	0.603	Immature but declining theme (x<0, y≥3.51, z<0.5)
Public opinion	4	5	1.906	0.800	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Communicable diseases	3	4	1.704	0.291	Peripheral mature theme (x≥0, y<3.51, z<0.5)
Internet	3	5	2.005	0.673	Peripheral mature theme (x≥0, y<3.51, z<0.5)
Communication	2	7	2.648	0.245	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Electronic nicotine delivery systems	2	4	1.250	0.668	Peripheral obsolete theme (x<0, y<3.51, z<0.5)
Geolocation	2	6	1.965	0.770	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Neural network	2	6	2.124	0.273	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Cluster 5[g] (n=10)					
Content analysis	5	28	10.239	0.354	Potential breakthrough theme (x<0, y≥3.51, z≥0.5)

XSL•FO
RenderX

Cluster and keywords	Freq[a], n	Overall degree	Network strength	Centroid distance[b]	Quadrant
Facebook	5	10	2.980	0.519	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Adolescent	3	9	3.117	0.618	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Association rule mining	3	7	2.490	0.518	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Health communication	2	3	1.296	0.175	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Health promotion	2	5	1.930	0.876	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Smoking	2	9	2.878	0.387	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Social media analysis	2	6	1.984	0.450	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Tobacco	2	8	2.620	0.498	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
User engagement	2	3	1.255	0.311	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Cluster 6[h] (n=13)					
Sentiment analysis	25	21	7.476	0.462	Core emerging hotspot (x≥0, y≥3.51, z≥0.5)
Deep learning	9	15	5.526	0.351	Potential breakthrough theme (x<0, y≥3.51, z≥0.5)
Latent Dirichlet allocation	6	19	7.123	0.805	Core emerging hotspot (x≥0, y≥3.51, z≥0.5)
Reddit	6	14	4.554	0.699	Potential breakthrough theme (x<0, y≥3.51, z≥0.5)
Social networking	6	9	3.350	0.627	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Social network analysis	5	3	1.074	0.382	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Vaccine	5	19	7.156	1.018	Potential breakthrough theme (x<0, y≥3.51, z≥0.5)
Crowdsourcing	4	9	2.570	0.603	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Network analysis	3	8	2.729	0.637	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
HPV	2	4	1.603	0.453	Peripheral obsolete theme (x<0, y<3.51, z<0.5)

XSL•FO
RenderX

Cluster and keywords	Freq[a], n	Overall degree	Network strength	Centroid distance[b]	Quadrant
Recurrent neural networks	2	3	0.789	0.382	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Social media data	2	3	0.746	0.202	Exploratory emerging theme (x<0, y<3.51, z≥0.5)
Word embedding	2	2	0.671	0.639	Exploratory emerging theme (x<0, y<3.51, z≥0.5)

[a]Freq: frequency.

[b]Representative terms can be identified from centroid distance values, providing both structural and semantic insights into each cluster. A smaller centroid distance indicates that a keyword is closer to the cluster center and thus more representative of the cluster's overall theme.

[c]Cluster 1: Candidate incubator pool of peripheral cross-cutting topics in health-related social media mining.

[d]Cluster 2: Computational methods in health informatics.

[e]Cluster 3: Public attitudes and sociopsychological determinants.

[f]Cluster 4: Infodemiology and the COVID-19 information ecosystem.

[g]Cluster 5: Health communication and public health engagement.

[h]Cluster 6: Social media analysis and network methods.

The UMAP-HDBSCAN clustering of the author keywords yielded 6 thematic clusters (Figure 5 and Table 3) with detailed statistics (Table S4 in Multimedia Appendix 6). Their 3D strategic positioning has been summarized in Figure 6 (2D projections are shown in Figures S7-S9 in Multimedia Appendix 1) and Table 3.

XSL•FO
RenderX

Figure 5. Keyword clusters identified by Uniform Manifold Approximation and Projection (UMAP)–Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) with cluster-level metrics. This image illustrates the thematic structure obtained by performing semantic-structural integration of keywords, followed by UMAP to reduce high-dimensional similarity relationships in 2D space and HDBSCAN for density-based clustering. The horizontal and vertical axes represent the 2 UMAP embedding coordinates (UMAP-1 and UMAP-2), which are precomputed as 2D representations of distances. The relative positions of points on the plane indicate the comprehensive similarity between keywords (closer distances indicate greater semantic and co-occurrence similarity). Different colors denote distinct cluster labels (clusters 2 - 6 represent assigned thematic clusters). The blue cross labeled cluster 1 (noise; HDBSCAN label=−1) indicates a set of keywords not stably assigned to any high-density cluster by HDBSCAN (unassigned/peripheral set), which does not equate to meaninglessness. Point size is proportional to keyword frequency (Freq) in the sample literature, reflecting each keyword's relative importance and coverage within the dataset. Six labels were identified under this parameter configuration: 5 nonnoise thematic clusters and 1 unassigned set.

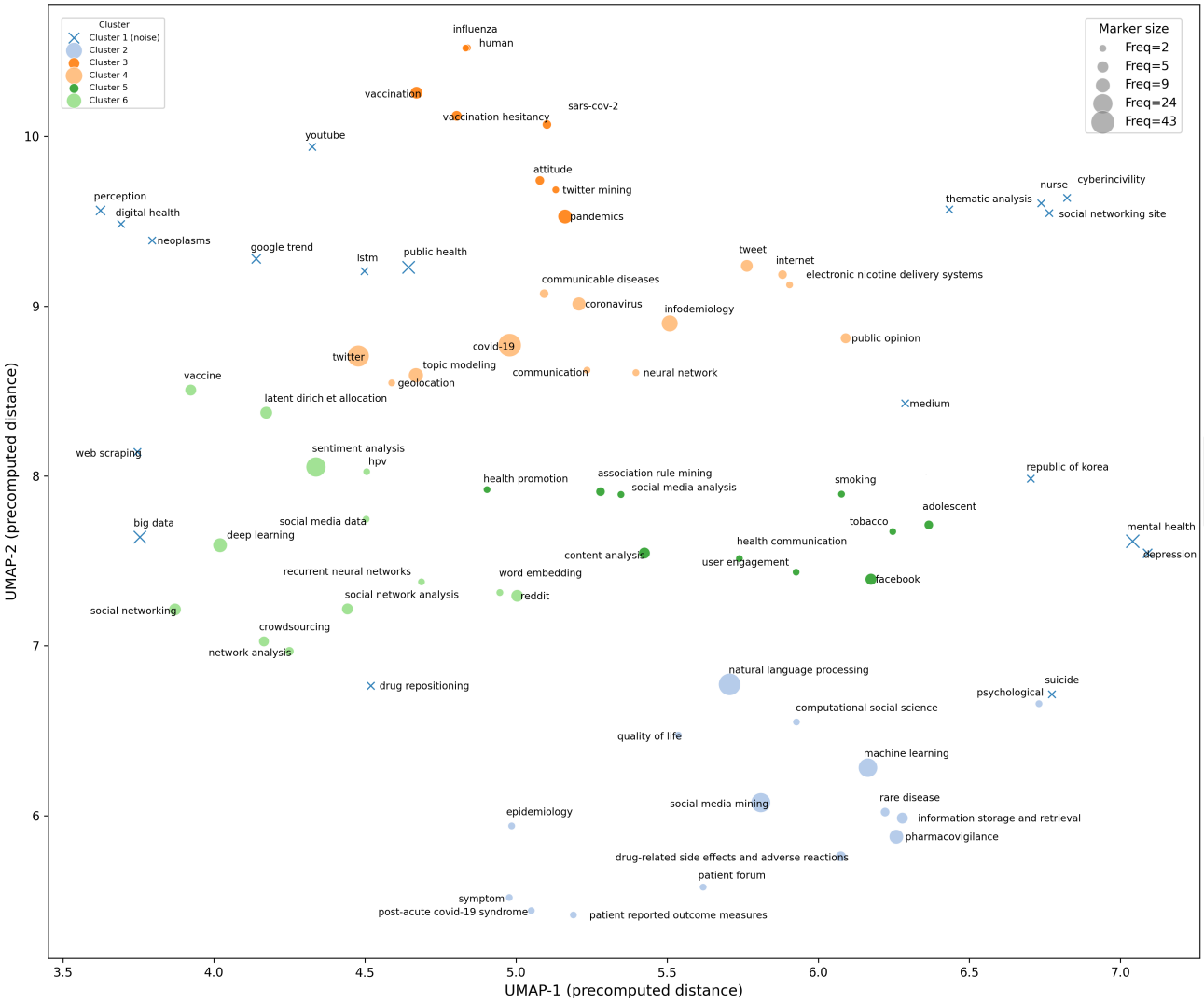

XSL•FO
RenderX

Figure 6. Three-dimensional thematic map of research on social media mining from January 1, 2015, to July 31, 2025. The image maps the keyword thematic structure identified by Uniform Manifold Approximation and Projection (UMAP)–Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) onto a 3D strategic coordinate system to assess the historical accumulation, structural influence, and recent growth potential of different research directions. The x-axis (maturity) represents the degree of research accumulation associated with each keyword, and it is characterized by the number of papers associated with the keyword, namely its occurrence frequency, and is normalized. The y-axis (influence) represents the structural importance of the keywords within the co-occurrence networks, and it is represented by network connection strength, namely centrality metrics, and is normalized. The z-axis (recency) represents the recent activity level of the keywords, and it is represented by recency metrics, such as the proportion of publications in the last 5 years, and is normalized. Each point in the image represents a keyword, with color denoting its assigned cluster. Clusters 2 - 6 represent assigned thematic clusters. Cluster 1 (noise) includes peripheral keywords that are not stably assigned, and it is used to present cross-thematic or sparse yet potentially valuable research signals. The red coordinate axis and central red dot denote the reference origin in 3D space. The black dashed planes mark the overall mean thresholds for the 3 axes (eg, mean of x=0.00, mean of y=3.51, and mean of z=0.66), which are used to divide the space into multiple strategic quadrants. This identifies distinct research directions such as core hotspots, potential breakthroughs, emerging frontiers, and declining themes.

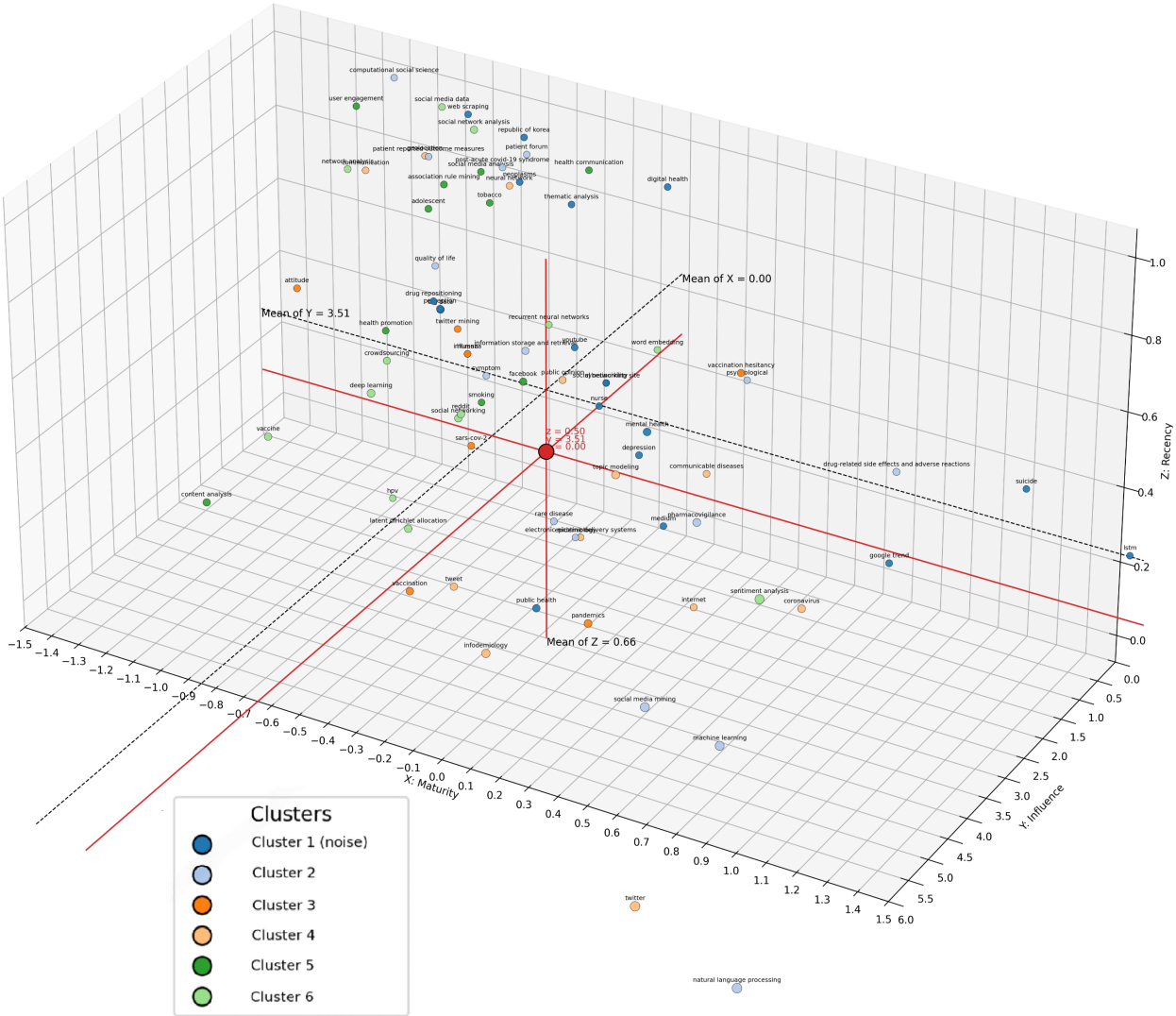

Cluster 1 (Noise): Candidate Incubator Pool of Peripheral Cross-Cutting Topics in Health-Related SMM

The UMAP-HDBSCAN analysis identified 6 thematic clusters, with 1 cluster identified as an unassigned peripheral noise set (cluster 1). We retained this unassigned set as a cluster for structural interpretation, as it contained meaningful cross-thematic elements, keywords disrupted by major event windows, and early-stage keywords. Its retention helped prevent information loss and aided in the capture of potentially meaningful cross-cutting and early-stage signals that did not form stable dense regions in the embedding space. The supporting rationale is presented in Multimedia Appendix 7. This cluster is a collection of heterogeneous, peripheral, and thematically varied research that cannot be categorized into any other single specialized cluster. The themes in this group are highly fragmented, with representative keywords ranging from "public health," "mental health," and "perception" to specific technologies like "big data" and "LSTM." However, this "noise" cluster is not without value. It represents the emerging and cross-disciplinary signals of the field, serving as an "incubator" for potential research directions and an important reference for defining the boundaries of current core research areas. The majority of keywords in this group fell within the *exploratory emerging theme* quadrant (x<0, y<3.51, z≥0.5), representing recently active, immature research topics such as "neoplasms," "YouTube," etc. Therefore, we treated cluster 1 as a candidate incubator pool of peripheral cross-cutting topics that may

XSL•FO
RenderX

contain topics that could evolve into coherent thematic clusters as the field matures, and its detailed interpretation is provided in Multimedia Appendix 7.

Cluster 2: Computational Methods in Health Informatics

Cluster 2 represents the technical and methodological foundations of the entire research field. It focuses on the application of core computational technologies, particularly "SMM," "NLP," and "ML," to diverse health scenarios. Its research involves the development and application of these technologies to address specific issues such as "drug-related side effects and adverse reactions," "pharmacovigilance," identification of "rare diseases," and analysis of "patient-reported outcomes" and "quality of life." It aims to efficiently extract valuable health insights from vast amounts of unstructured text data, serving as the technical foundation for all other applied clusters. Most topics in this cluster were located in the *exploratory emerging theme* quadrant ($x<0$, $y<3.51$, $z\geq0.5$), particularly the keywords representing patient-reported outcomes and the topic about symptoms.

Cluster 3: Public Attitudes and Sociopsychological Determinants

Cluster 3 represents a mature field dedicated to analyzing public psychology and reactions during health crises, particularly "pandemics" and "vaccination hesitancy." Using methods like "Twitter mining," it delves into public "attitudes," sentiments, and concerns. This cluster aims to uncover the deep-seated sociopsychological drivers behind public health attitudes, providing scientific evidence for developing targeted communication strategies and public policies.

Cluster 4: Infodemiology and the COVID-19 Information Ecosystem

Cluster 4 represents the most critical and dynamic research frontier, driven by the major global health crisis of "COVID-19." This cluster focuses on "infodemiology," examining how various types of information, including misinformation, spread, evolved, and influenced public opinion on social media platforms like Twitter during the pandemic. This is a quintessential crisis-driven research domain characterized by high timeliness and strategic importance. It aims to provide real-time monitoring tools and response strategies for addressing information challenges during public health emergencies.

Cluster 5: Health Communication and Public Health Engagement

Cluster 5 represents a highly application-oriented proactive intervention field, centered on designing and evaluating effective digital health communication campaigns. Using platforms like Facebook as its primary arena and focusing on specific issues, such as tobacco control, it uses "user engagement" as a key metric. Through methods like "content analysis," it explores which information frameworks most effectively reach and influence target audiences (eg, adolescents). Its findings offer direct practical guidance for public health institutions to optimize routine online health promotion efforts. All the themes in this group represent highly promising research trends for the coming decades.

Cluster 6: Social Media Analysis and Network Methods

Cluster 6 represents a group of advanced analytical techniques, deepening and advancing the methodologies explored in cluster 2, which involves general computational methods. It incorporates established technologies like "sentiment analysis" and "topic modeling" while introducing more sophisticated approaches such as "deep learning" and "social network analysis." Its core distinction lies in focusing not merely on the content of the text, but on revealing the mechanism of information dissemination. By analyzing user connections and the network structures of information flow, it uncovers the transmission dynamics of complex digital discourse.

Global Thematic Landscape and Cross-Cluster Bridging

Overall, the representative keywords in each quadrant clearly outline the stratified research landscape: flagship core themes, such as NLP, ML, and COVID-19, drive the development of the entire field; traditional pillars like content analysis and health communication maintain high influence but have waned in popularity; emerging potential topics, such as vaccination hesitancy and deep learning, are on the rise; and marginal incubation and dormant areas provide a window for observing future dynamic evolution.

To further explore the relationships between different thematic clusters, we measured the cross-cluster coupling strength between thematic clusters in the keyword co-occurrence network, defined as the sum of weights on cross-cluster edges connecting 2 distinct clusters, and identified bridging keywords contributing the most to coupling (Figure S10 in Multimedia Appendix 1 and Tables S5-S8 in Multimedia Appendix 6). The results indicated that cluster 4 occupied a central hub position in cross-cluster interactions, exhibiting strong coupling with multiple thematic clusters.

Among all cluster pairs, the highest coupling strength was observed between cluster 1 and cluster 4 (cross-cluster strength=10.649), followed by cluster 4 and cluster 6 (10.404), cluster 3 and cluster 4 (7.989), cluster 2 and cluster 4 (7.216), and cluster 2 and cluster 6 (6.967; Table S6 in Multimedia Appendix 6). Overall, these highly coupled clusters primarily revolve around cross-cluster co-occurrence links between cluster 4 and methodological clusters (cluster 2/6). It should be noted that cluster 1 consists of a heterogeneous set of peripheral keywords annotated by HDBSCAN. Therefore, its high coupling with core clusters more likely reflects conceptual convergence and enhanced cross-domain connections within specific event windows, such as the peak of pandemic-related research, rather than a single cohesive theme.

Furthermore, we analyzed the top 5 cross-cluster strength pairs by identifying their pair-specific bridging keywords (Table 4 and Table S7 in Multimedia Appendix 6). For the cluster 1-cluster 4 pair, the primary bridging keywords were COVID-19 (contribution=3.052), public health (2.104), Twitter (2.018), topic modeling (1.549), and coronavirus (1.429). For cluster 4 - cluster 6, the bridging keywords were Twitter (2.724), COVID-19 (2.540), deep learning (2.271), sentiment analysis (1.650), and topic modeling (1.646). For cluster 3-cluster 4, the bridging keywords were COVID-19 (3.013), pandemics (2.182),

XSL•FO
RenderX

Twitter (1.973), attitude (1.532), and tweet (1.440). For cluster 2-cluster 4, the bridging keywords primarily included NLP (3.031), COVID-19 (2.081), information epidemiology (1.541), Twitter (1.208), and ML (1.135). For cluster 2-cluster 6, the bridging keywords included NLP (2.476), SMM (2.436), deep learning (1.360), sentiment analysis (1.359), and LDA (1.288). These findings reflect significant tool and algorithm sharing across methodological clusters.

Table . Top 5 intercluster pairs (ranked by cross-cluster strength) with their top 5 pair-specific bridging keywords.

Cluster pair and keyword	Source cluster/keyword cluster	Keyword pair contribution[a]	Cross-cluster strength[b]
Cluster 1[c] (noise)-cluster 4[d]			10.649
COVID-19	Cluster 4	3.052	
Public health	Cluster 1 (noise)	2.104	
Twitter	Cluster 4	2.018	
Topic modeling	Cluster 4	1.549	
Coronavirus	Cluster 4	1.429	
Cluster 4-cluster 6[e]			10.404
Twitter	Cluster 4	2.724	
COVID-19	Cluster 4	2.540	
Deep learning	Cluster 6	2.271	
Sentiment analysis	Cluster 6	1.650	
Topic modeling	Cluster 4	1.646	
Cluster 3[f]-cluster 4			7.989
COVID-19	Cluster 4	3.013	
Pandemics	Cluster 3	2.182	
Twitter	Cluster 4	1.973	
Attitude	Cluster 3	1.532	
Tweet	Cluster 4	1.440	
Cluster 2[g]-cluster 6			7.216
Natural language processing	Cluster 2	2.476	
Social media mining	Cluster 2	2.436	
Deep learning	Cluster 6	1.360	
Sentiment analysis	Cluster 6	1.359	
Latent Dirichlet allocation	Cluster 6	1.288	
Cluster 2-cluster 4			6.967
Natural language processing	Cluster 2	3.031	
COVID-19	Cluster 4	2.081	
Infodemiology	Cluster 4	1.541	
Twitter	Cluster 4	1.208	
Machine learning	Cluster 2	1.135	

[a]Keyword pair contribution represents the contribution of the keyword to the cross-cluster coupling strength of the cluster, which is part of the cumulative contribution to cross-cluster edge weights.

[b]Cross-cluster strength is the sum of weights of cross-cluster edges connecting keywords from 2 distinct thematic clusters, which is used to quantify the degree of overlap between thematic clusters.

[c]Cluster 1: Candidate incubator pool of peripheral cross-cutting topics in health-related social media mining.

[d]Cluster 4: Infodemiology and the COVID-19 information ecosystem.

[e]Cluster 6: Social media analysis and network methods.

[f]Cluster 3: Public attitudes and sociopsychological determinants.

[g]Cluster 2: Computational methods in health informatics.

XSL•FO
RenderX

At the global level, keywords with the highest bridging strength clustered around event-driven themes and generic methodological/platform elements (Table S8 in Multimedia Appendix 6): COVID-19 (cross-cluster strength=12.520), NLP (10.642), Twitter (9.408), public health (7.669), content analysis (7.472), and pandemics (7.035). These bridging keywords jointly constructed a cross-cluster connection framework centered on event topics, platform data, and computational methods.

Dynamic Temporal Analysis

We divided the study period into 3 slices using the explosive growth time stamp as the temporal boundary: 2015 - 2019, 2020 - 2023, and 2024-2025 (July 31, 2025). For each time slice, we scanned K∈[2.20]. The Silhouette and Calinski-Harabasz (CH) indices were noninformative in the spectral space (yielding values near zero or remaining constant), and thus, we prioritized modularity and used the Davies-Bouldin Index elbow as a safeguard against overfragmentation. The decision to fix the number of spectral clusters at K=3 (2015 - 2019), K=8 (2020 - 2023), and K=5 (2024 - 2025) was guided by systematic evaluations across multiple criteria. While the Silhouette and CH indices were largely uninformative in the spectral embedding space, modularity consistently exhibited clear local maxima at these values, and Davies-Bouldin curves indicated elbows in the same vicinity. Selecting K at these points avoided both undersegmentation and overfragmentation while maximizing structural separation and interpretability. This choice ensured that each slice captured the dominant thematic structure of the period—technology-driven methodological consolidation before 2020, problem-driven thematic diversification from 2020 to 2023, and thematic convergence and strategic priorities after 2023. Figures S11-S14 in Multimedia Appendix 1 reveal the thematic changes across the 3 time slices.

Technology-Driven Methodological Consolidation (2015-2019)

During this phase, research themes emerged from the convergence of ML and NLP technological advancements with specific needs in pharmacovigilance and rare diseases. Though studies were relatively fragmented, they exhibited strong purposefulness, with technology driving applications. While this marked only the foundational stage of methodology and emerging applications, core methodological frameworks began to come into focus.

Problem-Driven Thematic Diversification (2020-2023)

The COVID-19 pandemic, as a black swan event, provided unprecedented data, attention, and application scenarios for the field. Consequently, COVID-19 became the absolute domain focus, leading to explosive growth in research scope and rapid diversification of research themes. This phase was problem-driven—specifically pandemic-driven—mobilizing all methodological approaches (ML, NLP, and sentiment analysis) and application directions (vaccines, information monitoring, and public attitudes) to meet its requirements. This process fostered a mature, complex, and highly differentiated research ecosystem.

Thematic Convergence and Strategic Priorities (2024-2025)

As the pandemic's impact began to recede, research themes started to reconsolidate. Methods like ML, NLP, and content analysis remained central. This indicates that technologies validated and refined during the pandemic, such as sentiment analysis and deep learning, are maturing and being systematically applied to health research. Furthermore, the latest research no longer pursues isolated hot topics, and instead, it applies validated methodologies to prior, long-term public health challenges like "vaccines" and "mental health."

Temporal Dynamics and Their Alignment With the Global Structure

Temporal spectral clusters aligned strongly with the global UMAP-HDBSCAN structure, with most spectral clusters mapping to a single global cluster at Jaccard scores above 0.3. The methodological cluster consistently matched the global "ML/NLP-driven social media analysis" domain. On the other hand, COVID-19 and mental health clusters overlapped with the global "Public health and pandemic response" and "Psychosocial health" domains, respectively.

Compared with the 3D strategic coordinate map, temporal slices provided a dynamic counterpart. The 2015 - 2019 clusters coincided with the map's high-maturity but low-recency quadrant, representing methodological foundations. The 2020 - 2023 diversification overlapped with clusters in the high-influence and high-recency space, reflecting crisis-driven impact. The 2024 - 2025 convergence mapped onto the high-maturity and moderate-recency zone, indicating thematic stabilization.

Validation

We conducted sensitivity analyses on burst detection and clustering hyperparameters to validate the methods and findings at the macro-level.

In the parameter sensitivity analysis for burst detection, we selected the configuration with γ=0.5, a slice width of 3 years, and a minimum burst length of 1, which demonstrated strong overall performance. This setting achieved optimal values in coverage (1.0) and realism (1.8), ensuring that detected bursts were both comprehensive and plausible. Although its standalone stability scores (Jaccard=0.108; Spearman=0.259) appeared modest, cross-parameter comparisons (Table S9 in Multimedia Appendix 6) revealed that alternative configurations yielded Jaccard similarities ≥0.97 and Spearman correlations of approximately 0.70 with the baseline, confirming that the results are consistently reproducible across the parameter space. Thus, G0.5_W3_L1 was chosen as the primary setting.

The internal metrics for clustering evaluation are presented in Table S10 in Multimedia Appendix 6. As can be seen from the table, the balanced version with K=6 showed slightly better internal separation perception. In contrast, the fine version maintained "good" performance across major metrics while improving interpretability by an order of magnitude. Additionally, this study conducted 3 independent runs of fine_K6, all yielding nearly identical scores and key metrics.

XSL•FO
RenderX

This indicates that the performance of fine_K6 is robust and insensitive to minor feature enhancements.

To assess the external validity of our clustering, we examined the relationship involving keyword network centrality and compared RCR distributions across clusters. As shown in Figure S15 in Multimedia Appendix 1, keyword strength was significantly positively correlated with mean RCR (Spearman ρ=0.49; $P<.05$), indicating that keywords occupying more central positions in the co-occurrence network tended to be associated with a higher citation impact (Figure S15 in Multimedia Appendix 1). Moreover, RCR distributions varied substantially across clusters (Figure S15 in Multimedia Appendix 1). Cluster 2 exhibited the highest citation impact (median RCR of approximately 1.5, with maximum values approaching 3), clearly exceeding other clusters. In contrast, clusters 1 and 4 showed lower mean RCR values (approximately 0.5 - 0.7), while noise (cluster 1) had the lowest citation impact (mean RCR of approximately 0.45). Together, these findings demonstrate that the identified clusters not only reflect structural differences within the keyword network but also align with external citation impact, thereby validating the robustness and scientific relevance of our approach.

To enhance the interpretability of the clusters beyond quantitative indicators, we performed micro-level interpretive triangulation by systematically analyzing the introduction, contribution, limitations, and future directions reported in 28 representative studies (Multimedia Appendix 4). The content analysis suggested that the thematic boundaries identified by clustering correspond closely to coherent conceptual domains. First, studies assigned to cluster 2 (NLP/ML-driven social media analysis) consistently emphasized methodological innovation. This validates the cluster's positioning as a methodological core in the strategic coordinate map where ML and NLP occupy high-influence and high-novelty regions [39,44]. Second, studies linked to clusters 1 and 2 highlighted empirical applications such as psychological health and suicide-related discussions [9,45]. Their reported contributions emphasized the utility of social media for monitoring population health behaviors, aligning with our dynamic slicing results where such content clusters emerged strongly during the COVID-19 period (2020 - 2023) and then diversified into related subthemes. Third, the limitations articulated in these studies mirror the structural signals captured by clustering. Frequent reliance on Twitter and US-based data explains why Twitter appeared as a central hub in cluster 2 [7,46], while non-Twitter platforms (eg, Weibo, Facebook, and Google Maps reviews) were less frequently studied, corresponding to their marginal or emerging placement in the coordinate map. Similarly, methodological constraints in sentiment analysis, cross-lingual generalizability, and demographic representativeness corroborated the lower maturity scores of several peripheral clusters. Finally, the future research directions proposed by the authors include cross-platform data integration [7,46]. Deep learning methods, population stratification, and longitudinal prediction directly overlap with the potential breakthroughs located in the high-influence but low-maturity quadrants of our strategic map (eg, vaccine attitudes, digital health, and Weibo-based research). This convergence suggests that our framework not only captures the current structure of the field but also anticipates trajectories already envisioned by domain experts.

Collectively, this micro-level evidence mapping supports the interpretability of the thematic clusters and provides qualitative context that complements the quantitative indicators and external validation results, without being used as confirmatory evidence of model correctness. It also demonstrates that the clusters are semantically meaningful and consistent with the intellectual agenda of the field, thereby strengthening confidence in both the static and dynamic analyses.

Discussion

Principal Findings

This study used a hybrid methodology that combined ML-based clustering with qualitative evidence mapping to systematically draw a strategic research landscape of SMM. The findings clearly demonstrated that this field has evolved into a complex and coherent ecosystem. This ecosystem revolves around "infodemiology" as its absolute core, powered by a dual engine of "computational methods" and "analytical techniques," extending its influence toward multiple critical application frontiers, including pharmacovigilance, public listening, and health management. Our findings not only depict the current state of the field but also reveal its underlying driving mechanisms and future developmental trajectory.

This study systematically analyzed health-related SMM literature collected in PubMed from 2015 to mid-2025, revealing significant evolutionary trends over the past decade. Overall, publication output showed steady growth, with explosive increases from 2019 to 2020, indicating SMM's rising strategic importance as a tool for digital health and public health research. By integrating natural language model embeddings with network algorithm clustering, we identified 6 relatively independent yet interconnected thematic clusters. These spanned methodological anchors (NLP, ML, and social network analysis), socially driven topics (vaccine hesitancy and public listening), and priority themes reconverging in the postexplosive growth phase (health communication, mental health, and oncology). The distribution of these themes across a strategic 3D coordinate system (maturity, influence, and recency) further revealed the historical accumulation, current impact, and future potential of different research directions.

Time-slice analysis indicated that the knowledge structure of this field underwent three phased transitions: (1) the methodological consolidation from 2015 to 2019 laid the technical foundation; (2) the pandemic-driven surge from 2020 to 2023 accelerated thematic diversification, concentrating on various public health issues; and (3) by 2024 - 2025, research refocused and converged on long-term strategic issues like cancer care and mental health. Through multiple validations, this study not only confirmed the robustness of clustering and strategic positioning but also demonstrated that these patterns align closely with the field's actual development trajectory. Collectively, the study's key findings provide a dynamic knowledge map for the digital health and health management field, spanning macro trends to micro evidence and offering a

XSL•FO
RenderX

robust strategic reference for academic research, policy formulation, and practical applications.

Comparison With Prior Work

Our finding that NLP and ML constitute the backbone of SMM aligns with recent reviews cataloging the rapid growth of computational methods in digital health [47-49]. However, some overviews have framed NLP as experimental in health care [50,51]. Our findings empirically demonstrate that, from a structural perspective, these technological pipelines have become firmly established as core architectures for social media–based health analytics (Table 3 and Figure 6). Additional empirical studies reinforce this interpretation. For instance, Ren et al [52] demonstrated the effectiveness of NLP and emotion-based deep learning (ML) for depression detection on Reddit, while Low et al [53] used NLP to identify vulnerable mental health support groups in online communities. Based on the above examples, these applications illustrate that methods are not only conceptually important but also practically operationalized in diverse health contexts. However, such research is often constrained by single-platform dependency and insufficient sample representativeness [52,54,55]. This is precisely why this strategic map positions this cluster as mature but still requiring methodological improvement and extended applications. Its strategic implications, from a sociological perspective, suggest that the dominance of a single-language Twitter, Facebook, Reddit, or Weibo corpus may structurally exclude speakers of other languages and marginalized voices, thereby exacerbating health inequalities [56-58]. From a health management perspective, mature NLP and ML pipelines are now deployable for real-time pharmacovigilance and attitude monitoring (eg, vaccine hesitancy dashboards), shifting management challenges from feasibility to governance [59-62]. Public health practice can benefit from earlier detection of emerging risks, enabling targeted, cost-effective preventive measures that reduce systemic burdens like disease incidence costs [63,64]. Collectively, the fundamental methods serve as strategic infrastructure in SMM.

Moreover, our analysis captured the explosion of crisis-driven research themes since 2020. Research during this period exhibited pronounced diversification and fragmentation, characterized by the rapid emergence of numerous exploratory subfields. Previous research accurately commented on this explosive growth but fragmented landscape in SMM [65]. However, to our knowledge, we provide data-driven evidence of an evolutionary trend in this field, shifting from fragmentation toward a focus on solving priority problems, based on our burst detection and time-slice analysis. Starting in 2024, as shown in our time-slice analysis, SMM research entered an era of integration characterized by a focus on addressing critical health issues. This signifies the field's transition from an exploratory phase of data-driven research to a mission-driven phase oriented toward identifying problems to solve. While acknowledging the fragmented explosion as a necessary phase, we critically highlight the ongoing structural shift toward a more mature research paradigm.

Methodological Highlights

Another major contribution of this study lies in methodological innovation. Unlike previous bibliometric research that relied primarily on co-occurrence–based keyword analysis, we proposed a fully automated, reproducible multilevel hybrid bibliometric methodology [17,18,24,29]. This approach of embedding-based mapping significantly enhances the interpretability and strategic value of the results while ensuring transparency and robustness [66].

First, at the semantic level, we used 2 natural language models—SPECTER2 and PubMedBERT—to vectorize titles and abstracts, aggregating them at the keyword level [66-68]. This approach overcomes the limitations of traditional methods that rely solely on manual synonym merging, enabling topic identification to capture latent semantic patterns and contextual associations [30]. In addition, at the structural level, we combined keyword co-occurrence networks with embedded similarity. Building upon the foundation, we constructed a hybrid similarity matrix that better aligns with actual disciplinary structures, achieving dual integration of semantics and structure.

Second, we used the nonparametric UMAP-HDBSCAN method to classify research themes. Systematic parameter grid searches and multiple random seed runs ensured the stability of clustering outcomes [69,70]. Unlike most studies relying solely on single clustering outputs, we further conducted internal (Jaccard, Silhouette, and Adjusted Rand Index) and external (RCR) validation. This was supplemented by micro-level evidence mapping and content analysis of representative papers. Such double validation could enhance clustering reliability.

Finally, this study proposes a 3D strategic coordinate map. By mapping clustering results onto the axes of maturity, influence, and recency, this framework transforms clustering outcomes into a knowledge map with strategic implications. This approach provides an intuitive tool for identifying hotspots, emerging frontiers, and declining fields.

Future Research Trends

Based on the above analysis, we identified several key research gaps and propose the following future research directions:

- Research direction 1: While some studies have explored the use of SMM to understand cancer patients' use of and perceptions toward traditional, complementary, and integrative medicine through micro-level content analysis, systematic evaluations of SMM research in chronic disease fields—represented by cancer—remain lacking [20,71,72]. The presence of "neoplasms" in cluster 1 of the static cluster analysis further corroborates this conclusion.
- Research direction 2: Most studies focus on correlational descriptions, with some literature exploring responses to public health initiatives among different populations. However, causal inference remains lacking, leaving unanswered questions such as "Can social media content truly demonstrate the impact of health promotion initiatives on changing health behaviors?" and "How do these behaviors change?" [73]. Future research could use mixed methods, integrating qualitative approaches to explore the underlying logic of health behavior change [61].
- Research direction 3: Future potential hotspots lie within cluster 1, such as SMM research on mental health and psychological disorders. This has been confirmed by the

XSL•FO
RenderX

3D coordinate map. Proactively leveraging new technologies like large language models to explore this domain using social media data will drive a shift from retrospective summarization to prospective prevention [74,75].

- Research direction 4: The 3D strategic coordinate map showed topics like adolescent and tobacco positioned in niche yet high-potential zones. This indicates that targeted health communication and health behavior–related issues, particularly those addressing specific populations, may evolve into core hotspots in future research [76].
- Research direction 5: Keywords like vaccination and attitude occupied either the high-impact or high-novelty quadrant in the 3D strategic coordinate map, marking a strategic frontier hotspot. Topics, such as vaccination hesitancy and public listening, experienced short-term bursts during the pandemic but landed closer to the peripheral emerging zone, indicating limited long-term stability [60,77]. Nevertheless, vaccine-related issues and public listening remain core academic focuses and will continue to evolve.

The outlined future directions can be consolidated into a transformable research framework. At the data level, comparative analyses involving cross-databases, multilingual datasets, and cross-platform sources should be encouraged to mitigate representational biases arising from single-platform or single-language sources while testing the transferability of thematic structures [78]. At the methodological level, semantic-structural integration can expand from static co-occurrence networks to dynamic networks to better capture thematic evolution and cross-cluster migration [79]. At the validation level, time slices and burst detection results can be compared against external reference data, such as policy timelines and significant events, while exploring event study methods to enhance interpretability and real-world applicability [73,75,80].

Limitations

This study has several limitations. First, our analysis examined only English literature from PubMed, excluding other databases and other languages, which may not fully capture SMM applications across the entire field. In particular, this study relied on PubMed as its sole primary data source, resulting in a bias toward biomedical and clinical research. Consequently, representation from other fields was insufficient, potentially underestimating the scale of methodological innovations or interdisciplinary research. Second, due to the time lag in literature collection, recency may be underestimated. Third, the clustering analysis itself relied on keywords. While it considered the meaning of keywords within abstracts, it could not capture all the fine distinctions across studies. Moreover, we used keywords as analytical units in this study. Keywords themselves exhibit variations in naming and granularity, and have nonuniform usage. Hence, synonymous expressions may not be fully normalized. Although this study used PubMedBERT and SPECTER2 to learn structural and semantic similarity from title and abstract contexts, this approach only partially mitigates the fragmentation caused by relying on string matching. It cannot guarantee complete unification of all synonyms and polysemous terms, particularly when interdisciplinary terminology and task definitions are inconsistent. Beyond this, we did not conduct stratified external validation for each of the 3 time periods, such as comparing RCRs stage by stage. Such analyses may be statistically unstable with smaller samples and when reference metrics require time accumulation, potentially introducing systematic bias for the most recent stage, particularly the short and incomplete window of 2024 - 2025 (July 2025). Future studies may explore stage-by-stage external validation with longer follow-up intervals. Additionally, SMM research can be influenced by structural biases associated with social media platforms and their user demographics [81]. For example, different platforms have different user demographics, and regional disparities and the digital divide may also affect the visibility of research topics. Our included SMM literature often relied on data from platforms like Twitter, Reddit, and Weibo, which differ in user profiles, languages, and geographic distributions. Consequently, which topics receive more research attention largely depends on the ease of data accessibility and user activity on these platforms. Therefore, the research findings presented here should primarily be understood as reflecting the knowledge structure at the level of academic publications, rather than being directly representative of the complete picture in the real world. Finally, as a bibliometric analysis, this study aimed to reveal correlations and trends rather than validate strict causal relationships.

Conclusion

This study systematically revealed the knowledge structure, strategic positioning, and evolutionary trajectory of health-related SMM from 2015 to 2025 by developing a fully automated, reproducible hybrid bibliometric methodology. The innovation of this study lies in constructing a semantic-structural hybrid similarity matrix while enhancing method robustness through dual-level validation to reduce synonym fragmentation and parameter sensitivity. Unlike traditional bibliometric reviews that primarily rely on co-occurrence statistics or single-pass clustering, this study offers a decision-oriented, comprehensive analysis integrating interpretable thematic clustering, 3D strategic positioning, intercluster relationships, and time-slice analysis. Substantively, we found that the research evolved through 3 distinct periods: an initial phase of methodological consolidation, a middle phase where themes were rapidly diversified and fragmented, and a later phase of long-term priority problem-solving, such as mental health and cancer care. The 3D strategic coordinate map further indicates that differences in maturity, influence, and recency among various themes determine their potential and challenges for future development.

Our findings hold practical implications for public health, health research, and health care. For public health, the findings confirm the necessity of infodemiology and provide methodological support for developing precise, audience-segmented public health communication strategies. For health research, the knowledge map offered by this study can be used to strategically set research priorities and optimize the allocation of research resources. For health care, technologies in this field can be applied to support real-world health surveillance systems and data-driven decision-making. In particular, mature analytic

XSL•FO
RenderX

pipelines in SMM can support not only the development but also the validation of real-world surveillance and pharmacovigilance workflows and a better understanding of the patient journey through analysis of social media datasets, thereby supporting evidence-informed decision-making. Although this study provides an organized framework, further efforts are needed to support real-world decision-making. Regarding future work, the hybrid analytical framework proposed in this study will undergo continuous iteration across 3 dimensions. At the data level, we will conduct cross-database, multilingual, and cross-platform comparative analyses. Methodologically, we will extend semantic-structural integration from static co-occurrence networks to dynamic networks. For validation, we will combine temporal slicing analysis and burst detection with event analysis to enhance explainability and practical utility.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing, adapting and adjusting emotional tone, translation, and reformatting. The GAI tools used were ChatGPT-4o, ChatGPT-5.2, and DeepSeek-V3. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. For manuscript preparation, given that the authors are nonnative English speakers, writing assistance tools were used: DeepL for translation, Grammarly and Qullibot for grammar refinement, and large language models (GPT-4o, GPT-5, GPT-5.2, and DeepSeek-V3) for language polishing. These tools were used only for language editing.

Funding

This study was conducted as part of the ATLAS project "Innovation and Digital Transformation in Healthcare," funded by the State of North Rhine-Westphalia, Germany (grant number: ITG-1-1).

Data Availability

The original code is not publicly available because the repository contains personal API keys/credentials, local computer paths, and project-specific configuration information. The code can be obtained from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: MJY, SB-J
Data analysis: MJY
Supervision: SB-J
Writing – original draft: MJY
Writing – review & editing: MJY, SB-J

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary figures to support the study.
[PDF File, 4182 KB - jmir_v28i1e86200_app1.pdf]

Multimedia Appendix 2
Keyword burst detection (Kleinberg bursts) and parameter robustness assessment.
[PDF File, 236 KB - jmir_v28i1e86200_app2.pdf]

Multimedia Appendix 3
Mathematical derivation and computational process documentation for the hybrid semantic-structural bibliometric analysis pipeline.
[PDF File, 353 KB - jmir_v28i1e86200_app3.pdf]

Multimedia Appendix 4
Micro-level interpretive triangulation with selected articles (evidence mapping).
[PDF File, 632 KB - jmir_v28i1e86200_app4.pdf]

Multimedia Appendix 5
List of included publications.

XSL•FO
RenderX

[XLSX File, 157 KB - jmir_v28i1e86200_app5.xlsx]

Multimedia Appendix 6
Supplementary tables to support the study.
[PDF File, 394 KB - jmir_v28i1e86200_app6.pdf]

Multimedia Appendix 7
Interpretation and rationale for retaining the Hierarchical Density-Based Spatial Clustering of Applications with Noise unassigned noise set (cluster 1: candidate incubator pool of peripheral cross-cutting topics in health-related social media mining).
[PDF File, 188 KB - jmir_v28i1e86200_app7.pdf]

Checklist 1
BIBLIO checklist.
[PDF File, 162 KB - jmir_v28i1e86200_app8.pdf]

Checklist 2
PRISMA-S checklist.
[PDF File, 277 KB - jmir_v28i1e86200_app9.pdf]

References

1. Number of internet and social media users worldwide. Statista. URL: https://www.statista.com/statistics/617136/digital-population-worldwide [accessed 2025-05-10]
2. Shah HA, Househ M. Understanding loneliness through analysis of Twitter and Reddit data: comparative study. Interact J Med Res 2025 Mar 14;14:e49464. [doi: 10.2196/49464] [Medline: 40085832]
3. Valdez D, Mena-Meléndez L, Crawford BL, Jozkowski KN. Analyzing Reddit forums specific to abortion that yield diverse dialogues pertaining to medical information seeking and personal worldviews: data mining and natural language processing comparative study. J Med Internet Res 2024 Feb 14;26:e47408. [doi: 10.2196/47408] [Medline: 38354044]
4. Sinnenberg L, Buttenheim AM, Padrez K, Mancheno C, Ungar L, Merchant RM. Twitter as a tool for health research: a systematic review. Am J Public Health 2017 Jan;107(1):143-143. [doi: 10.2105/AJPH.2016.303512a]
5. Koss J, Rheinlaender A, Truebel H, Bohnet-Joschko S. Social media mining in drug development-fundamentals and use cases. Drug Discov Today 2021 Dec;26(12):2871-2880. [doi: 10.1016/j.drudis.2021.08.012] [Medline: 34481080]
6. Li J, Xu Q, Cuomo R, Purushothaman V, Mackey T. Data mining and content analysis of the Chinese social media platform Weibo during the early COVID-19 outbreak: retrospective observational infoveillance study. JMIR Public Health Surveill 2020 Apr 21;6(2):e18700. [doi: 10.2196/18700] [Medline: 32293582]
7. Lazard AJ, Saffer AJ, Wilcox GB, Chung AD, Mackert MS, Bernhardt JM. E-cigarette social media messages: a text mining analysis of marketing and consumer conversations on Twitter. JMIR Public Health Surveill 2016 Dec 12;2(2):e171. [doi: 10.2196/publichealth.6551] [Medline: 27956376]
8. Koss J, Bohnet-Joschko S. Social media mining of long-COVID self-medication reported by Reddit users: feasibility study to support drug repurposing. JMIR Form Res 2022 Oct 3;6(10):e39582. [doi: 10.2196/39582] [Medline: 36007131]
9. Park A, Conway M, Chen AT. Examining thematic similarity, difference, and membership in three online mental health communities from Reddit: a text mining and visualization approach. Comput Human Behav 2018 Jan;78(98-112):98-112. [doi: 10.1016/j.chb.2017.09.001] [Medline: 29456286]
10. Chi Y, Chen HY. Investigating substance use via Reddit: systematic scoping review. J Med Internet Res 2023 Oct 25;25:e48905. [doi: 10.2196/48905] [Medline: 37878361]
11. Ru B, Yao L. A literature review of social media-based data mining for health outcomes research. In: Bian J, Guo Y, He Z, Hu X, editors. Social Web and Health Research: Springer; 2019:1-14. [doi: 10.1007/978-3-030-14714-3_1]
12. Ford E, Shepherd S, Jones K, Hassan L. Toward an ethical framework for the text mining of social media for health research: a systematic review. Front Digit Health 2020;2:592237. [doi: 10.3389/fdgth.2020.592237] [Medline: 34713062]
13. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res 2021 May 12;23(5):e17917. [doi: 10.2196/17917] [Medline: 33978589]
14. Fu J, Li C, Zhou C, et al. Methods for analyzing the contents of social media for health care: scoping review. J Med Internet Res 2023 Jun 26;25:e43349. [doi: 10.2196/43349] [Medline: 37358900]
15. Cobelli N, Blasi S. Combining topic modeling and bibliometric analysis to understand the evolution of technological innovation adoption in the healthcare industry. Eur J Innov Manag 2024 Dec 16;27(9):127-149. [doi: 10.1108/EJIM-06-2023-0497]
16. Chien SC, Yen CM, Chang YH, et al. Use of artificial intelligence, internet of things, and edge intelligence in long-term care for older people: comprehensive analysis through bibliometric, Google trends, and content analysis. J Med Internet Res 2025 Mar 4;27:e56692. [doi: 10.2196/56692] [Medline: 40053718]

XSL•FO
RenderX

17. Hassan W, Duarte AE. Bibliometric analysis: a few suggestions. Curr Probl Cardiol 2024 Aug;49(8):102640. [doi: 10.1016/j.cpcardiol.2024.102640] [Medline: 38740289]
18. Marzi G, Balzano M, Caputo A, Pellegrini MM. Guidelines for bibliometric - systematic literature reviews: 10 steps to combine analysis, synthesis and theory development. Int J Management Reviews 2025 Jan;27(1):81-103. [doi: 10.1111/ijmr.12381]
19. Petit P, Vuillerme N. Leveraging administrative health databases to address health challenges in farming populations: scoping review and bibliometric analysis (1975-2024). JMIR Public Health Surveill 2025 Jan 9;11:e62939. [doi: 10.2196/62939] [Medline: 39787587]
20. Xie Y, Guo Z, Zeng X, et al. Social media-based cancer education: bibliometric and thematic analysis. JMIR Cancer 2025 Oct 6;11:e77214. [doi: 10.2196/77214] [Medline: 41052420]
21. Alhashmi SM, Hashem IAT, Al-Qudah I. Artificial Intelligence applications in healthcare: a bibliometric and topic model-based analysis. Intelligent Systems with Applications 2024 Mar;21:200299. [doi: 10.1016/j.iswa.2023.200299]
22. Hu J, Li C, Ge Y, Yang J, Zhu S, He C. Mapping the evolution of digital health research: bibliometric overview of research hotspots, trends, and collaboration of publications in JMIR (1999-2024). J Med Internet Res 2024 Oct 17;26:e58987. [doi: 10.2196/58987] [Medline: 39419496]
23. Hankar M, Kasri M, Beni-Hssane A. A comprehensive overview of topic modeling: techniques, applications and challenges. Neurocomputing 2025 May;628:129638. [doi: 10.1016/j.neucom.2025.129638]
24. Ogunleye B, Lancho Barrantes BS, Zakariyyah KI. Topic modelling through the bibliometrics lens and its technique. Artif Intell Rev 2025;58(3). [doi: 10.1007/s10462-024-11011-x]
25. Kukreja V. Comic exploration and Insights: recent trends in LDA-based recognition studies. Expert Syst Appl 2024 Dec;255:124732. [doi: 10.1016/j.eswa.2024.124732]
26. Sheng B, Wang Z, Qiao Y, Xie SQ, Tao J, Duan C. Detecting latent topics and trends of digital twins in healthcare: a structural topic model-based systematic review. Digit Health 2023;9:20552076231203672. [doi: 10.1177/20552076231203672] [Medline: 37846404]
27. Yu JH, Chauhan D. Trends in NLP for personalized learning: LDA and sentiment analysis insights. Educ Inf Technol 2025 Mar;30(4):4307-4348. [doi: 10.1007/s10639-024-12988-2]
28. Lee S, Kim L, Shim MS, Kim GS. Identifying health care services offered in the HIV care continuum via a machine learning-based topic modeling approach: exploratory literature review. JMIR Public Health Surveill 2025 Jul 9;11:e65081. [doi: 10.2196/65081] [Medline: 40632764]
29. Xie Q, Waltman L. A comparison of citation-based clustering and topic modeling for science mapping. Scientometrics 2025 May;130(5):2497-2522. [doi: 10.1007/s11192-025-05324-z]
30. Chen G, Hong S, Du C, Wang P, Yang Z, Xiao L. Comparing semantic representation methods for keyword analysis in bibliometric research. J Informetr 2024 Aug;18(3):101529. [doi: 10.1016/j.joi.2024.101529]
31. Deka P, Kalita H, Sarmah M. Author keyword occurrence and trends in digital media research: a bibliometric study. Alexandria: The Journal of National and International Library and Information Issues 2025. [doi: 10.1177/09557490251406852]
32. Kong M, Zhang Y, Sheng L, Hong K. Citation structural diversity: a novel metric combining structure and semantics for literature evaluation. Scientometrics 2025 Jul;130(7):4027-4060. [doi: 10.1007/s11192-025-05356-5]
33. Feng Y, Chen Z, Zhang Y, Huang W, Zhang X, He S. BERTopic_Teen: a multi-module optimization approach for short text topic modeling in adolescent health. Front Public Health 2025;13. [doi: 10.3389/fpubh.2025.1608241]
34. Montazeri A, Mohammadi S, M Hesari P, Ghaemi M, Riazi H, Sheikhi-Mobarakeh Z. Preliminary guideline for reporting bibliometric reviews of the biomedical literature (BIBLIO): a minimum requirements. Syst Rev 2023 Dec 15;12(1):239. [doi: 10.1186/s13643-023-02410-2] [Medline: 38102710]
35. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
36. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
37. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2026-04-30]
38. Ayyoubzadeh SM, Ayyoubzadeh SM, Zahedi H, Ahmadi M, R Niakan Kalhori S. Predicting COVID-19 incidence through analysis of Google trends data in Iran: data mining and deep learning pilot study. JMIR Public Health Surveill 2020 Apr 14;6(2):e18828. [doi: 10.2196/18828] [Medline: 32234709]
39. Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inform Assoc 2015 May;22(3):671-681. [doi: 10.1093/jamia/ocu041] [Medline: 25755127]
40. Tapi Nzali MD, Bringay S, Lavergne C, Mollevi C, Opitz T. What patients can tell us: topic analysis for social media on breast cancer. JMIR Med Inform 2017 Jul 31;5(3):e23. [doi: 10.2196/medinform.7779] [Medline: 28760725]
41. Lazard AJ, Wilcox GB, Tuttle HM, Glowacki EM, Pikowski J. Public reactions to e-cigarette regulations on Twitter: a text mining analysis. Tob Control 2017 Dec;26(e2):e112-e116. [doi: 10.1136/tobaccocontrol-2016-053295] [Medline: 28341768]

XSL•FO
RenderX

42. Li D, Chaudhary H, Zhang Z. Modeling spatiotemporal pattern of depressive symptoms caused by COVID-19 using social media data mining. Int J Environ Res Public Health 2020 Jul 10;17(14):4988. [doi: 10.3390/ijerph17144988] [Medline: 32664388]
43. Zhang J, Wang Y, Shi M, Wang X. Factors driving the popularity and virality of COVID-19 vaccine discourse on Twitter: text mining and data visualization study. JMIR Public Health Surveill 2021 Dec 3;7(12):e32814. [doi: 10.2196/32814] [Medline: 34665761]
44. Blumenthal KG, Topaz M, Zhou L, et al. Mining social media data to assess the risk of skin and soft tissue infections from allergen immunotherapy. J Allergy Clin Immunol 2019 Jul;144(1):129-134. [doi: 10.1016/j.jaci.2019.01.029] [Medline: 30721764]
45. Cheng Q, Li TM, Kwok CL, Zhu T, Yip PS. Assessing suicide risk and emotional distress in Chinese social media: a text mining and machine learning study. J Med Internet Res 2017 Jul 10;19(7):e243. [doi: 10.2196/jmir.7276] [Medline: 28694239]
46. Daughton AR, Shelley CD, Barnard M, et al. Mining and validating social media data for COVID-19-related human behaviors between January and July 2020: infodemiology study. J Med Internet Res 2021 May 25;23(5):e27059. [doi: 10.2196/27059] [Medline: 33882015]
47. Chen G, Tan B, Laham N, Tracey TJG, Lapinski S, Liu Y. A bibliometric review of natural language processing applications in psychology from 1991 to 2023. Basic Appl Soc Psych 2025 Mar 4;47(2):105-119. [doi: 10.1080/01973533.2024.2433720]
48. Habehh H, Gohel S. Machine learning in healthcare. Curr Genomics 2021 Dec 16;22(4):291-300. [doi: 10.2174/1389202922666210705124359] [Medline: 35273459]
49. Kim J, Lee D, Park E. Machine learning for mental health in social media: bibliometric study. J Med Internet Res 2021 Mar 8;23(3):e24870. [doi: 10.2196/24870] [Medline: 33683209]
50. Cunha Reis T. Artificial intelligence and natural language processing for improved telemedicine: before, during and after remote consultation. Aten Primaria 2025 Aug;57(8):103228. [doi: 10.1016/j.aprim.2025.103228] [Medline: 39955812]
51. Casey A, Davidson E, Poon M, et al. A systematic review of natural language processing applied to radiology reports. BMC Med Inform Decis Mak 2021 Jun 3;21(1):179. [doi: 10.1186/s12911-021-01533-7] [Medline: 34082729]
52. Ren L, Lin H, Xu B, Zhang S, Yang L, Sun S. Depression detection on Reddit with an emotion-based attention network: algorithm development and validation. JMIR Med Inform 2021 Jul 16;9(7):e28754. [doi: 10.2196/28754] [Medline: 34269683]
53. Low DM, Rumker L, Talkar T, Torous J, Cecchi G, Ghosh SS. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on Reddit during COVID-19: observational study. J Med Internet Res 2020 Oct 12;22(10):e22635. [doi: 10.2196/22635] [Medline: 32936777]
54. Shankar K, Chandrasekaran R, Jeripity Venkata P, Miketinas D. Investigating the role of nutrition in enhancing immunity during the COVID-19 pandemic: Twitter text-mining analysis. J Med Internet Res 2023 Jul 10;25:e47328. [doi: 10.2196/47328] [Medline: 37428522]
55. Zhang C, Xu S, Li Z, Hu S. Understanding concerns, sentiments, and disparities among population groups during the COVID-19 pandemic via Twitter data mining: large-scale cross-sectional study. J Med Internet Res 2021 Mar 5;23(3):e26482. [doi: 10.2196/26482] [Medline: 33617460]
56. Sametoğlu S, Pelt DHM, Eichstaedt JC, Ungar LH, Bartels M. The value of social media language for the assessment of wellbeing: a systematic review and meta-analysis. J Posit Psychol 2024 May 3;19(3):471-489. [doi: 10.1080/17439760.2023.2218341]
57. Western MJ, Smit ES, Gültzow T, et al. Bridging the digital health divide: a narrative review of the causes, implications, and solutions for digital health inequalities. Health Psychol Behav Med 2025;13(1):2493139. [doi: 10.1080/21642850.2025.2493139] [Medline: 40276490]
58. Wu M, Xue Y, Ma C. The association between the digital divide and health inequalities among older adults in China: nationally representative cross-sectional survey. J Med Internet Res 2025 Jan 15;27:e62645. [doi: 10.2196/62645] [Medline: 39813666]
59. Karapetiantz P, Audeh B, Redjdal A, Tiffet T, Bousquet C, Jaulent MC. Monitoring adverse drug events in web forums: evaluation of a pipeline and use case study. J Med Internet Res 2024 Jun 18;26:e46176. [doi: 10.2196/46176] [Medline: 38888956]
60. Boatman D, Starkey A, Acciavatti L, Jarrett Z, Allen A, Kennedy-Rea S. Using social listening for digital public health surveillance of human papillomavirus vaccine misinformation online: exploratory study. JMIR Infodemiology 2024 Mar 8;4:e54000. [doi: 10.2196/54000] [Medline: 38457224]
61. Timotin A, Paladi A, Mita V, Chihai V, Lozan O. Social listening applied to tailor communication on immunization in the Republic of Moldova. Eur J Public Health 2025 Apr 1;35(2):270-275. [doi: 10.1093/eurpub/ckae161] [Medline: 39470446]
62. Wehrli S, Irrgang C, Scott M, Arnrich B, Boender TS. The role of the (in)accessibility of social media data for infodemic management: a public health perspective on the situation in the European Union in March 2024. Front Public Health 2024;12:1378412. [doi: 10.3389/fpubh.2024.1378412] [Medline: 38651120]

XSL•FO
RenderX

63. Liscano Y, Anillo Arrieta LA, Montenegro JF, Prieto-Alvarado D, Ordoñez J. Early warning of infectious disease outbreaks using social media and digital data: a scoping review. Int J Environ Res Public Health 2025 Jul 13;22(7):1104. [doi: 10.3390/ijerph22071104] [Medline: 40724171]
64. Amin S, Zeb MA, Alshahrani H, Hamdi M, Alsulami M, Shaikh A. Social media-based surveillance systems for health informatics using machine and deep learning techniques: a comprehensive review and open challenges. CMES 2024;139(2):1167-1202. [doi: 10.32604/cmes.2023.043921]
65. Resnik P, De Choudhury M, Musacchio Schafer K, Coppersmith G. Bibliometric studies and the discipline of social media mental health research. Comment on “Machine learning for mental health in social media: bibliometric study”. J Med Internet Res 2021 Jun 17;23(6):e28990. [doi: 10.2196/28990] [Medline: 34137722]
66. González-Márquez R, Schmidt L, Schmidt BM, Berens P, Kobak D. The landscape of biomedical research. Patterns (N Y) 2024 Jun 14;5(6):100968. [doi: 10.1016/j.patter.2024.100968] [Medline: 39005482]
67. Menke JD, Kilicoglu H, Smalheiser NR. Publication type tagging using transformer models and multi-label classification. AMIA Annu Symp Proc 2024;2024:818-827. [Medline: 40417522]
68. Singh A, D’Arcy M, Cohan A, Downey D, Feldman S. SciRepEval: a multi-format benchmark for scientific document representations. Presented at: 2023 Conference on Empirical Methods in Natural Language Processing; Dec 6-10, 2023. [doi: 10.18653/v1/2023.emnlp-main.338]
69. Lezhnina O. Depression, anxiety, and burnout in academia: topic modeling of PubMed abstracts. Front Res Metr Anal 2023;8:1271385. [doi: 10.3389/frma.2023.1271385] [Medline: 38090103]
70. Bonakala S, Aupetit M, Bensmail H, El-Mellouhi F. A human-in-the-loop approach for visual clustering of overlapping materials science data. Digital Discovery 2024 Mar 13;3(3):502-513. [doi: 10.1039/D3DD00179B]
71. Lam CS, Zhou K, Loong HHF, Chung VCH, Ngan CK, Cheung YT. The use of traditional, complementary, and integrative medicine in cancer: data-mining study of 1 million web-based posts from health forums and social media platforms. J Med Internet Res 2023 Apr 21;25:e45408. [doi: 10.2196/45408] [Medline: 37083752]
72. Wu X, Lam CS, Hui KH, et al. Perceptions in 3.6 million web-based posts of online communities on the use of cancer immunotherapy: data mining using BERTopic. J Med Internet Res 2025 Feb 10;27:e60948. [doi: 10.2196/60948] [Medline: 39928933]
73. Dahabreh IJ, Bibbins-Domingo K. Causal inference about the effects of interventions from observational studies in medical journals. JAMA 2024 Jun 4;331(21):1845-1853. [doi: 10.1001/jama.2024.7741] [Medline: 38722735]
74. Hua Y, Na H, Li Z, et al. A scoping review of large language models for generative tasks in mental health care. NPJ Digit Med 2025 Apr 30;8(1):230. [doi: 10.1038/s41746-025-01611-4] [Medline: 40307331]
75. Deiner MS, Deiner NA, Hristidis V, et al. Use of large language models to assess the likelihood of epidemics from the content of tweets: infodemiology study. J Med Internet Res 2024 Mar 1;26:e49139. [doi: 10.2196/49139] [Medline: 38427404]
76. Vogel EA, Ranker LR, Harrell PT, et al. Characteristics of adolescents’ and young adults’ exposure to and engagement with nicotine and tobacco product content on social media. J Health Commun 2024 Jun 2;29(6):383-393. [doi: 10.1080/10810730.2024.2355291] [Medline: 38775659]
77. Xu J, Wu Z, Wass L, Larson HJ, Lin L. Mapping global public perspectives on mRNA vaccines and therapeutics. NPJ Vaccines 2024 Nov 14;9(1):218. [doi: 10.1038/s41541-024-01019-3] [Medline: 39543153]
78. Shen Y, Luo Z, Song X, Liu C. Research on the evolution of cross-platform online public opinion for public health emergencies considering stakeholders. PLoS ONE 2024;19(6):e0304877. [doi: 10.1371/journal.pone.0304877]
79. Ma X, Strube M, Zhao W. Graph-based clustering for detecting semantic change across time and languages. Presented at: 18th Conference of the European Chapter of the Association for Computational Linguistics; Mar 17-22, 2024. [doi: 10.18653/v1/2024.eacl-long.93]
80. Yue Z, Yu G. Effects of policy communication changes on social media: before and after policy adjustment. Systems 2025;13(4):248. [doi: 10.3390/systems13040248]
81. Mourtzouchou A, Marin AL, Laveneziana L, et al. Comparative analysis of public and expert perceptions of electrified vehicles in the European Union. Sci Rep 2025 Jul 1;15(1):21695. [doi: 10.1038/s41598-025-06071-0] [Medline: 40594306]

Abbreviations

API: application programming interface
ARIMA: autoregressive integrated moving average
BIBLIO: Preliminary Guideline for Reporting Bibliometric Reviews of the Biomedical Literature
CH: Calinski-Harabasz
HDBSCAN: Hierarchical Density-Based Spatial Clustering of Applications with Noise
IF: impact factor
LDA: latent Dirichlet allocation
MeSH: Medical Subject Headings
ML: machine learning

XSL•FO
RenderX

NLP: natural language processing
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCR: relative citation ratio
SMM: social media mining
UMAP: Uniform Manifold Approximation and Projection

Edited by S Brini; submitted 20.Oct.2025; peer-reviewed by B Xiang, D Zikos; revised version received 27.Mar.2026; accepted 02.Apr.2026; published 08.May.2026.

Please cite as:
Yang MJ, Bohnet-Joschko S
Thematic Mapping and Evolution of Social Media Mining in Health Research: Hybrid Bibliometric Synthesis
J Med Internet Res 2026;28:e86200
URL: https://www.jmir.org/2026/1/e86200
doi:10.2196/86200

© Mia Jiming Yang, Sabine Bohnet-Joschko. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Large Language Model–Based Analysis of Statin Therapy Discussions and Sentiment on Social Media: Cross-Sectional Observational Study

Siru Liu[1,2], PhD; Jialin Liu[3,4], MD

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Ave #1475, Nashville, TN, United States

[2]Department of Computer Science, Vanderbilt University, Nashville, TN, United States

[3]Department of Medical Informatics, West China Medical School, Chengdu, China

[4]Department of Otolaryngology-Head and Neck Surgery, West China Hospital, Sichuan University, Chengdu, China

Corresponding Author:
Siru Liu, PhD
Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Ave #1475, Nashville, TN, United States

Abstract

Background: Statin therapy, despite proven cardiovascular benefits, remains underused. Social media platforms may capture patient perspectives that are less visible in clinical encounters.

Objective: This study aimed to characterize themes, sentiment, and decision-making factors related to statin therapy through large language model (LLM)–based analysis of Reddit discussions.

Methods: This cross-sectional observational study analyzed English-language Reddit posts and comments mentioning statins from January 2022 to May 2025, identified via keyword-based Reddit application programming interface searches (≤1000 posts per keyword). A total of 5328 retrieved discussions (n=1661, 31.2% posts and n=3667, 68.8% keyword-containing comments) from public subreddits were included. Themes, sentiments (positive, neutral, or negative), guideline-informed clinical relevance, information-seeking behavior, adverse effect mentions, decision factors, and adherence-related content were extracted using an LLM-based pipeline.

Results: Among 5328 discussions, prominent topics included adverse effects (n=1697, 31.9%), decision-making references related to laboratory results and physician advice (n=2767, 51.9% and n=2034, 38.2%, respectively), and alternative approaches (n=2485, 46.6%). Overall sentiment was neutral in 34% (n=1812) of discussions, negative in 30.9% (n=1646), and positive in 16.9% (n=900); the remainder were mixed or unclear. Statin-directed sentiment was neutral in 44.1% (n=2350) of discussions, negative in 25.2% (n=1343), and positive in 12.5% (n=666); the remainder did not express statin-directed sentiment. High clinical relevance was identified in 12.6% (n=672) of discussions. Adherence-related issues were mentioned in 29.8% (n=1587) of discussions. Among adverse effect mentions, muscle pain (n=129, 7.6%) and fatigue (n=110, 6.5%) were common.

Conclusions: LLM-enabled analysis of Reddit discourse highlights substantial negative sentiment, adherence-related concerns, and adverse effect narratives surrounding statin therapy. These findings suggest opportunities for patient-centered communication and shared decision-making strategies that address symptom attribution, uncertainty, and information needs in digital information environments.

(J Med Internet Res 2026;28:e85057) doi:10.2196/85057

KEYWORDS

statin therapy; social media; large language model; sentiment analysis; Reddit; cardiovascular disease; patient perspectives; adherence; adverse effects; cohort study

Introduction

Statins are among the most widely prescribed medications globally, with more than 200 million individuals using these agents to reduce cardiovascular morbidity and mortality [1]. These 3-hydroxy-3-methylglutaryl coenzyme A reductase inhibitors remain the cornerstone of atherosclerotic cardiovascular disease (ASCVD) prevention, demonstrating well-established efficacy in both primary and secondary prevention settings [2,3]. Despite their proven benefits, real-world statin therapy is frequently limited by concerns regarding safety and tolerability. In routine care, statin-related events are commonly documented, with myalgia and myopathy representing the most frequent category, and these events often precipitate temporary discontinuation and subsequent

XSL•FO
RenderX

rechallenge [4]. Consequently, long-term use remains suboptimal; a recent systematic review and meta-analysis found that only approximately 62% of patients achieve good adherence (≥80% use) over a median follow-up of 24 months [5], while population-based primary care data show frequent discontinuation with substantial restarting, consistent with an intermittent use pattern in practice [6].

Traditional clinical research methods, including randomized controlled trials and observational studies, may not fully capture the breadth of patient experiences and concerns in routine care [7]. Effective communication between patients and health care providers is essential for shared decision-making and treatment adherence; however, barriers such as limited consultation time, health literacy challenges, and patient reluctance to disclose adverse experiences may constrain insight into the real-world challenges of medication use [8,9]. Patients may hesitate to report symptoms or concerns to clinicians due to fear of judgment, desire to avoid conflict with prescribers, or uncertainty about whether their experiences warrant clinical attention [9]. Social media platforms, especially Reddit (Reddit, Inc), provide access to unfiltered patient narratives that may not be expressed during clinical encounters [10-13]. Reddit hosts a wide range of health-related communities and has more than 50 million daily active users [14]. Within these forums, individuals share personal medical experiences, seek peer advice, and discuss treatment decisions. These discussions may provide valuable insights into patient beliefs, motivations, and barriers that influence medication adherence and cardiovascular risk management [15-17].

Recent advances in large language models (LLMs) offer scalable methods to analyze such unstructured, high-volume textual data, enabling systematic characterization of patient perspectives at a granularity that would be infeasible using manual qualitative approaches alone [18]. In the context of cardiovascular prevention, applying LLMs to social media discourse may provide novel insights into how patients interpret statin-related information, evaluate perceived risks and benefits, and make adherence decisions outside the clinical setting. Accordingly, this study aimed to systematically analyze Reddit discussions related to statin use using an LLM to characterize patient-reported experiences, identify recurrent concerns and misconceptions, and explore factors influencing decision-making and adherence. By integrating patient-generated narratives with computational text analysis, this work seeks to complement traditional evidence sources and inform more patient-centered approaches to cardiovascular risk management.

Methods

Study Design and Data Source

This cross-sectional study analyzed publicly available Reddit discussions about statin therapy posted between January 1, 2022, and May 1, 2025. Reddit was selected because it hosts large, topic-specific health communities in which users openly share medication experiences, treatment decisions, and interactions with health care providers in naturalistic settings. Data were accessed via the official Reddit application programming interface (API) using Python Reddit API Wrapper, limited to public subreddits, and no user contact or reidentification attempts were made.

Data Collection

Search Strategy

We retrieved content using “statin,” generic statin names (eg, atorvastatin, rosuvastatin, simvastatin, pravastatin, lovastatin, fluvastatin, and pitavastatin), and US brand names (eg, Lipitor, Crestor, Zocor, Pravachol, Livalo, Lescol, and Mevacor); full queries are provided in Multimedia Appendix 1.

API Procedure

Using Python Reddit API Wrapper (version 7.7.1) in Python (version 3.9; Python Software Foundation; May 15-May 20, 2025), we queried Reddit’s search end point (sorted by “new” where available). Due to indexing and API constraints, each keyword returned up to 1000 submissions. For each submission, we extracted metadata (submission ID, created_utc, subreddit, title, selftext, author, score, comment count, and permalink) and downloaded its comment tree (comment ID, parent ID, created_utc, author, body, and score). We retained submissions within the study window; pagination, time stamps, and rate-limit handling are described in Multimedia Appendix 1.

Eligibility

We included English-language posts and comments from public subreddits that matched 1 or more keywords (case-insensitive) within the study window and excluded removed and deleted placeholders, duplicates, non-English text (langdetect v1.0.9), and promotional and spam content (Multimedia Appendix 1).

Unit of Analysis

The unit of analysis was a Reddit discussion thread (eligible submission plus associated comments), enabling contextual interpretation of patient narratives and peer responses and supporting downstream LLM-based thematic extraction [19].

Data Cleaning and Preprocessing

We removed URLs and normalized whitespace while preserving the original wording (no stemming or lemmatization). Privacy protection included removing @mentions and replacing detected email addresses and phone numbers with placeholders (eg, “[EMAIL],” “[PHONE]”); subreddit names were retained as public metadata, with reidentification risk minimized via aggregate reporting and avoidance of traceable quotations [20]. English language posts were identified using fastText (lid.176.bin) [21] (retain ≥0.80, adjudicate 0.60-0.80 by 2 investigators, and exclude <0.60). For analysis, comments were retained if they contained 1 or more keywords; misspellings and variants were captured via regex-based partial matching with guardrails. Exact duplicates were removed via SHA-256 hashing, and near duplicates were excluded using MinHash with Jaccard similarity ≥0.85 [22]. All steps were deterministic and version controlled; a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)–style flow diagram is provided in Multimedia Appendix 1.

XSL•FO
RenderX

LLM-Based Content Extraction

Structured extraction used GPT-4.1 [23] via the OpenAI API (version 1.52.0) between May and June 2025 with standardized parameters (temperature=0.1, top_p=1.0, and max_tokens=4096). A single prompt template and predefined JSON schema were applied to all discussions. Two investigators iteratively developed the schema and decision rules through pilot testing on 50 purposively sampled discussions and consensus refinement of edge cases, consistent with clinical reasoning standards [18,24]. For 11 analytic domains (Table 1), outputs included structured variables, brief summaries, and verbatim evidence quotes supporting key classifications. Dietary changes and exercise were coded as alternative approaches only when explicitly discussed as substitutes for statin therapy (eg, attempting to avoid, delay, or discontinue statins); when described as adjuncts alongside statins, they were not classified as alternatives. Outputs failing JSON validation or missing mandatory evidence were automatically requeried up to 3 times; only schema-valid outputs were retained (Multimedia Appendix 2).

Table . GPT-4.1 content analysis framework with examples.

Domain	Description	Example extractions
Primary themes	1 to 3 main discussion topics	• Concerns about statin adverse effects, cholesterol management through lifestyle changes, and physician-patient communication
Statin medications	Specific statin names mentioned	• Generic: atorvastatin and rosuvastatin • Brand: Lipitor and Crestor
Experience type	Postclassification	• Personal experience, information request, and medical advice
Sentiment analysis	Multitarget emotional assessment	• Neutral, negative, or positive • Overall, toward statins and toward physicians
Clinical relevance	Actionable insight assessment based on AHA[a] and ACC[b] guidelines	• High: ASCVD[c] history (MI[d], stroke, and PAD[e]), LDL-C[f] ≥190 mg/dL, diabetes (in individuals aged 40 - 75 years), and 10-year ASCVD risk ≥7.5% • Medium: family history of premature ASCVD, chronic kidney disease, inflammatory conditions, CAC[g] score ≥100, and smoking • Low: general cardiovascular health discussions
Adverse effects	Adverse reactions mentioned	• Muscle pain, fatigue, brain fog, dizziness, nausea, and memory issues
Decision factors	Treatment choice influences	• Laboratory results, physician recommendation, and adverse effects
Information seeking	Behavioral patterns	• Asking for advice, sharing experience, and seeking alternatives
Adherence issues	Medication compliance	• Mentions discontinuation, compliance issues, and dose changes
Alternative treatments	Nonstatin options	• Diet changes, exercise, ezetimibe, and coenzyme Q10
Emotional indicators	Expressed emotions	• Frustration, confusion, hope, and anxiety

[a]AHA: American Heart Association.
[b]ACC: American College of Cardiology.
[c]ASCVD: atherosclerotic cardiovascular disease.
[d]MI: myocardial infarction.
[e]PAD: peripheral artery disease.
[f]LDL-C: low-density lipoprotein cholesterol.
[g]CAC: coronary artery calcium.

XSL•FO
RenderX

All content was analyzed as user-generated narratives. References to clinician recommendations within discussions (eg, "my doctor recommended starting a statin") were coded as patient-reported physician communication rather than independently verified clinical input. Posts containing general medical guidance without a described patient-clinician interaction were not coded under the "decision factors" physician recommendation category. Clinician perspectives were not independently ascertained; physician-related content was captured only through users' own accounts of clinical interactions.

Clinical Relevance Tiering

A deterministic rule-based algorithm aligned with American Heart Association, American College of Cardiology, and US Preventive Services Task Force indicators [25,26] assigned high, medium, or low clinical relevance (Multimedia Appendix 2). High relevance required explicit mention of ASCVD, low-density lipoprotein cholesterol ≥190 mg/dL, diabetes (in individuals aged 40 - 75 years), or elevated 10-year ASCVD risk, and medium relevance reflected risk-enhancing factors (eg, family history of premature ASCVD, chronic kidney disease, inflammatory conditions, and coronary artery calcium ≥100 Agatston units) [25,26]; otherwise, discussions were low relevance. Given self-reported data, absence of criteria was coded as "unreported," and tier assignment defaulted downward unless explicit evidence supported a higher tier.

Detection of LLM-Related Content

We searched for mentions of LLM tools (eg, ChatGPT [OpenAI], Claude [Anthropic], and Google Gemini) and categorized them as (1) seeking medical advice, (2) seeking general health information, (3) sharing LLM-generated content, or (4) incidental mentions. Categories 1 to 3 were flagged for sensitivity analyses (Multimedia Appendix 2).

Validation

LLM outputs were validated via expert review (Multimedia Appendix 3). A stratified random sample of 50 discussions (0.9% of 5328) spanning key domains was independently assessed by 2 reviewers (a clinical informaticist and a physician) using a 5-point accuracy scale (1=very poor to 5=excellent). Interrater agreement used Cohen κ. Reviewers also documented recurrent errors. Verbatim evidence requirements enabled direct verification of extracted labels.

Statistical Analysis

Multivariable logistic regression identified factors associated with negative sentiment toward statins. Cluster-robust SEs were estimated at the subreddit level. Covariates were prespecified (clinical relevance tier, adverse effect mentions, adherence-related content, information-seeking behavior, specific statins, and subreddit category). Absence of a mention was coded as "not reported." Diagnostics included variance inflation factors and influential observation checks. Results are reported as odds ratios with 95% CIs; 2-sided $P<.05$ indicated significance.

Ethical Considerations

This study analyzed publicly accessible Reddit posts and comments as secondary data and involved no direct contact, interaction, intervention, or attempt to reidentify users. Informed consent was not obtained because the study used publicly available online content, the investigators did not interact with users, and obtaining consent from all posters was not feasible. No compensation was provided because no participants were recruited or contacted. The study was exempt from institutional review board approval and aligned with the Association of Internet Researchers' Internet Research: Ethical Guidelines 3.0 [27].

Results

Dataset Characteristics and Distribution

After systematic filtering, the final dataset comprised 5328 discussions contributed by 4832 unique users across multiple Reddit communities. The most frequently represented forum was r/Cholesterol (n=2722, 51.1%). Nearly half of the discussions (n=2552, 47.9%) referenced a specific statin. The most commonly mentioned agents were rosuvastatin (n=1276, 23.9%) and atorvastatin (n=1013, 19%). Dataset characteristics and statin mentions are summarized in Table 2.

XSL•FO
RenderX

Table . Characteristics of Reddit discussions and statin mentions (N=5328).

Characteristic and category	Mentions, n (%)
Document type	
Submissions (posts)	1661 (31.2)
Comments	3667 (68.8)
Source community	
r/Cholesterol	2722 (51.1)
r/PeterAttia	576 (10.8)
r/AskDocs	300 (5.6)
r/keto	92 (1.7)
r/HeartAttack	87 (1.6)
r/diabetes_t2	82 (1.5)
r/stroke	46 (0.9)
Other subreddits	1423 (26.7)
Primary discussion type	
Personal experience sharing	3210 (60.2)
Information request	882 (16.6)
Medical advice provision	734 (13.8)
General discussion	502 (9.4)
Mention of a specific statin	
Yes	2552 (47.9)
No	2776 (52.1)
Most frequent statins[a]	
Rosuvastatin (Crestor)	1276 (23.9)
Atorvastatin (Lipitor)	1013 (19)
Pravastatin (Pravachol)	279 (5.2)
Simvastatin (Zocor)	211 (4)

[a]Generic and brand name mentions were combined.

Thematic Analysis and Primary Discussion Topics

Thematic analysis showed that discussions most commonly focused on treatment effectiveness (3311/14,007, 23.6% of thematic mentions), followed by safety and tolerability concerns (n=2366, 16.9%) and alternative lifestyle interventions (n=2081, 14.9%; Figure 1). Privacy-preserving, paraphrased exemplar posts (1 - 2 per major theme) and the corresponding LLM theme classifications are provided in Multimedia Appendix 4.

XSL•FO
RenderX

Figure 1. Quantitative distribution of themes in Reddit discussions related to statins (January 2022 to May 2025). The hierarchical tree map illustrates the primary discussion topics identified by the large language model pipeline across 5328 discussions, organized into 5 overarching thematic groups and 10 topic categories. Each discussion could receive up to 3 thematic codes; percentages represent the proportion of total thematic mentions (N=14,007). The 10 most frequent topic categories are shown; remaining thematic mentions (n=3828, 27.3%) comprised less frequent topics and are not displayed.

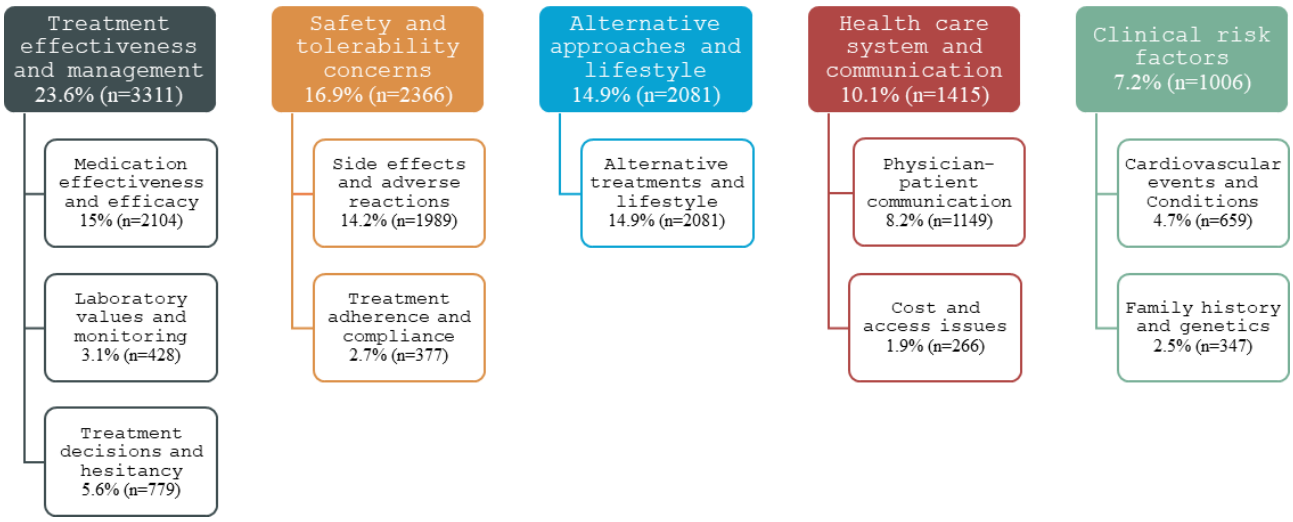

Sentiment Analysis

Overall sentiment was neutral in 34% of discussions, negative in 30.9%, and positive in 16.9%, with the remainder classified as mixed or unclear. Sentiment distributions varied significantly by community type (χ_8^2=335.7; P<.001). The proportion of discussions with negative sentiment was substantially higher in medical advice–seeking forums (eg, r/AskDocs: 77.3%) than in lifestyle-oriented forums (eg, r/Biohackers: 9.1%).

Clinical Relevance Assessment

Among 5328 discussions, using guideline-informed criteria, 12.6% (n=672) of discussions were classified as high clinical relevance, 22.4% (n=1193) as medium relevance, and 65% (n=3463) as low relevance. High-relevance content was most frequent in r/stroke (40/46, 87.0%), r/HeartAttack (41/87, 47.1%), r/diabetes_t2 (29/82, 35.4%), and r/AskDocs (82/300, 27.3%). Clinical information elements were reported as follows: laboratory values (n=2048, 38.4%), cardiovascular events (n=362, 6.8%), family history (n=908, 17%), and lifestyle factors (n=2616, 49.1%).

Adverse Effect Reports and Safety Concerns

Among 5328 discussions, adverse effects were reported in 31.9% (n=1697; 95% CI 30.7% - 33.2%) of discussions. Among discussions that referenced adverse effects, the most frequent included muscle pain (n=129, 7.6%) and fatigue (n=110, 6.5%), followed by cognitive symptoms such as “brain fog” (n=61, 3.6%). A detailed breakdown of adverse effects, including neuropsychiatric symptoms, is provided in Multimedia Appendix 5. In 0.3% (n=17) of discussions, users explicitly stated that no adverse effects occurred.

Decision-Making Factors and Treatment Influences

Of 5328 discussions, the most frequently coded factors associated with statin-related decisions were laboratory results (n=2767, 51.9%), physician recommendations (n=2034, 38.2%), adverse effects (n=1593, 29.9%), and family history of cardiovascular disease (n=822, 15.4%). Additional factors included lifestyle modifications (n=1092, 20.5%), online research (n=573, 10.8%), cost (n=316, 5.9%), and insurance coverage (n=72, 1.4%). Genetic predisposition and clinical guideline recommendations were each coded in 0.5% (n=26 for both) of discussions.

Alternative Treatment Discussions

Among 5328 discussions, alternative treatments were coded in 46.6% (n=2485) of discussions. The most frequent nonpharmacologic alternatives were dietary changes (n=449, 8.4%) and exercise (n=406, 7.6%). Pharmacologic alternatives included ezetimibe (n=168, 3.2%), Repatha (n=84, 1.6%), and Zetia (n=70, 1.3%). Supplements included fish oil (n=60, 1.1%), coenzyme Q10 (n=57, 1.1%), and red yeast rice (n=53, 1%). Weight loss was coded in 1.4% (n=73) of discussions.

Information-Seeking Behaviors and Community Engagement

Among 5328 discussions, experience sharing was coded in 78.2% (n=4167) of discussions and advice seeking was coded in 46.1% (n=2456). Discussions focused on alternative treatments occurred in 7.2% (n=386) of cases, while 5.3% (n=284) included questioning medication necessity. Advice seeking was most frequent in r/AskDocs (n=292, 97.3%).

Emotional Indicators and Adherence Issues

Of the 5328 discussions, emotional expressions were identified in 85.2% (n=4537) of discussions. The most frequently coded emotions were frustration (n=1705, 32%) and confusion (n=911, 17.1%). Medication adherence issues were present in 29.8% (n=1587) of the discussions, including discontinuation (n=522, 9.8%) and dose modification (n=437, 8.2%). Multiple codes per discussion were permitted; therefore, percentages may sum to more than 100%.

Mentions of LLM Tools in Reddit Discussions

Explicit mentions of LLM tools occurred in 23/5,328 (0.4%) discussions, totaling 38 mentions. Of the 38 mentions, ChatGPT accounted for 32 (84.2%), followed by other GPT variants (n=3,

XSL•FO
RenderX

7.9%) and Claude, Bard/Gemini, or DeepSeek (n=3, 7.9% combined). Mentions were coded as statin-related medical decision-making (n=2, 5.3%), information seeking (n=6, 15.8%), and general references (n=30, 78.9%). Discussions containing LLM mentions were most commonly observed in r/PeterAttia (6/23, 26.1%) and r/Cholesterol (5/23, 21.7%). Mentions increased from 6 in 2023 to 32 in January 2025 to May 2025.

Expert Validation Results

Expert validation of a stratified sample of 50 discussions showed a mean validation score of 4.67 (SD 0.74). Interrater agreement was substantial (Cohen κ=0.85; 95% CI 0.78-0.92). Expert feedback noted recurring coding challenges in distinguishing genetic risk factors (eg, lipoprotein[a]) from family history, as well as in identifying cardiovascular event–related content (eg, coronary artery calcium scores).

Statistical Associations

Sentiment distributions differed across communities (χ^2=335.7; *P*<.001). Negative sentiment was more frequent in medical advice–seeking communities than in lifestyle-focused communities (eg, r/AskDocs: 77.3% vs r/Biohackers: 9.1%; *P*<.001). Adverse effect reporting also differed by community type (χ_2^2=124.6; *P*<.001). Adverse effect mentions and adherence issues were associated with negative sentiment (*P*<.001 for both). In adjusted multivariable analysis, factors associated with negative sentiment included adverse effect mentions (adjusted odds ratio [aOR] 3.42, 95% CI 2.89 - 4.05; *P*<.001), posts in medical advice communities (aOR 2.76, 95% CI 2.31 - 3.30; *P*<.001), adherence issues (aOR 2.18, 95% CI 1.84 - 2.58; *P*<.001), and high clinical relevance content (aOR 1.47, 95% CI 1.18 - 1.84; *P*=.001; Table 3).

Table . Multivariable logistic regression of factors associated with negative sentiment.

Variable	Adjusted odds ratio (95% CI)	*P* value
Adverse effect mentions	3.42 (2.89 - 4.05)	<.001
Posts in medical advice communities	2.76 (2.31 - 3.30)	<.001
Adherence issues	2.18 (1.84 - 2.58)	<.001
High clinical relevance content	1.47 (1.18 - 1.84)	.001

Discussion

Main Findings

In this large-scale analysis of statin-related discussions on Reddit, patient discourse was shaped primarily by perceived adverse effects, uncertainty about benefits, and peer validation, with relatively limited reference to formal cardiovascular risk stratification. Although statins are strongly endorsed by clinical guidelines for ASCVD prevention [2,26], real-world patient narratives were frequently framed around experiential and emotional factors, including fear of long-term harm, symptom attribution, and ambivalence toward medical authority. This divergence between guideline-based evidence and patient-centered concerns helps explain persistently suboptimal statin adherence despite decades of robust trial data [28].

Notably, fewer than 1 in 8 discussions met criteria for high clinical relevance based on American Heart Association, American College of Cardiology, and US Preventive Services Task Force risk thresholds [2,26]. Instead, most conversations focused on nonspecific symptoms, laboratory fluctuations, or lifestyle considerations, suggesting that engagement with statin therapy often occurs outside a formal risk-benefit calculus. Therefore, effective statin counseling may require attention to patients' beliefs and concerns, beyond clinical risk factors alone.

The Adverse Effect Paradox: Perception vs Trial Evidence

Adverse effect concerns were prominent in our dataset (n=1697, 31.9% of discussions), which is not directly comparable to trial-based incidence estimates but may reflect a perception-evidence gap that shapes beliefs and adherence decisions. Evidence from blinded randomized trials and individual-participant meta-analyses suggests that the excess risk of muscle symptoms attributable to statins is small and that most muscle symptom reports under blinded conditions are not attributable to statin therapy [29]. This divergence is consistent with nocebo-related expectation and attribution mechanisms within the broader information environment. In the Anglo-Scandinavian Cardiac Outcomes Trial–Lipid-Lowering Arm, muscle-related adverse events were reported more frequently during the unblinded extension than during blinded treatment despite comparable exposure, supporting expectation-driven symptom attribution [30].

Cognitive concerns (eg, "brain fog") appeared in 3.6% (n=61) of discussions, although randomized trials have not demonstrated statin-attributable cognitive impairment [31]. Regulatory agencies have noted rare, generally reversible postmarketing reports, and the prominence of such concerns online may reflect expectation-driven symptom attribution [32]. Mechanistically, true statin myopathy occurs in a subset of patients, including those with *SLCO1B1* variants or drug-drug interactions, but the high salience of adverse effects in online communities likely reflects both heterogeneity in susceptibility and selection effects (symptomatic users preferentially posting) [33,34]. Clinicians should anticipate that patients initiate statins within an information ecosystem that can magnify harm narratives. Proactive counseling, clear differentiation of evidence-based risks, and structured symptom assessment (eg, Statin-Associated Muscle Symptom Clinical Index) may reduce nocebo-driven discontinuation [35,36].

XSL•FO
RenderX

Emotional Burden as a Hidden Driver of Nonadherence

Beyond physical symptoms, statin discussions carried substantial emotional content (n=4537, 85.2%), with frustration (n=1705, 32%) and confusion (n=911, 17.1%) predominating. These patterns suggest that statin decisions may be shaped by emotional responses (eg, frustration and anxiety) in addition to objective risk appraisal. Negative emotion is consistently associated with poorer adherence across chronic disease contexts. A meta-analysis by DiMatteo et al [37] found markedly higher nonadherence among patients with depression. Similarly, medication-specific emotional distress has been associated with treatment discontinuation, even after accounting for clinical depression [38,39]. In our dataset, adherence problems (n=1587, 29.8%) frequently co-occurred with adverse effect narratives and negative sentiment, suggesting a pathway in which perceived harms generate distress that undermines persistence. These emotional dimensions may be underelicited in routine care, where time-constrained visits often prioritize biomarker review and dose adjustment. Incorporating a brief assessment of medication-related distress into cardiovascular prevention workflows, paired with motivational interviewing, could identify patients at risk for emotionally driven discontinuation earlier in the treatment course.

Community as an Information Filter: The Ecology of Online Health Discourse

Sentiment and content varied substantially across Reddit communities. Negative sentiment was more prevalent in medical advice–seeking subreddits (eg, r/AskDocs: 77.3%) than in lifestyle-focused subreddits (eg, r/Biohackers: 9.1%). Condition-specific forums (eg, r/stroke and r/HeartAttack) contained the highest proportion of clinically relevant content, while dietary communities (eg, r/keto) more often reflected cholesterol skepticism that diverged from guideline framing [2].

These patterns are consistent with community selection as an “information filter,” in which users preferentially encounter narratives aligned with prevailing community norms and their own concerns; similar dynamics have been described in other online health communities [40,41]. Such community-specific patterns may reinforce users’ existing perspectives. For example, dietary forums showed more skepticism toward statins, while condition-specific communities (eg, r/HeartAttack) contained more secondary prevention content. Therefore, community-specific outreach may be more effective than generic education. Strategies could include engaging trusted voices within skeptical forums, tailoring evidence presentation to community values, and directly addressing prevalent misconceptions (eg, “cholesterol myth” narratives), thereby complementing traditional patient counseling.

The Emerging Role of Artificial Intelligence in Patient Decision-Making

Explicit mentions of LLMs were uncommon (n=23, 0.4%) but increased from 6 in 2023 to 32 in early 2025, suggesting growing uptake. Users described using tools such as ChatGPT to interpret laboratory results, weigh treatment options, and prepare questions for clinical encounters. To the best of our knowledge, few studies have quantified artificial intelligence (AI) tool use within real-world cardiovascular treatment discussions. Prior research has evaluated LLM performance on medical queries [42,43], but implications for patient expectations and decision-making remain incompletely characterized. These findings raise questions about how AI-mediated information may shape patient perspectives before clinical visits, warranting further investigation and practical guidance for clinicians.

Limitations

This study has several limitations. First, Reddit API retrieval imposes sampling constraints. Search results are not uniformly sampled. Each keyword query is capped (≤1000 submissions), and ranking algorithms can overrepresent newer posts. This may yield a recency- and visibility-biased corpus. Accordingly, theme and sentiment frequencies should be interpreted as visibility-weighted patterns within the retrieved dataset rather than population-level prevalence for all statin-related content on Reddit; this may inflate high-salience topics (eg, adverse effects, discontinuation, and emotionally charged narratives) and complicate temporal comparisons. Second, Reddit’s pseudonymous structure prevents verification of demographics, diagnoses, lipid values, comorbidities, and outcomes. Users are unlikely to be representative of the broader statin-using population, limiting generalizability—particularly to older adults and those with lower digital access or health literacy. Third, self-reported social media content is subject to recall bias and selective posting, potentially overrepresenting unusual or negative experiences. Fourth, although we used expert-guided prompt development and validation, LLM-based extraction may misclassify nuanced language (eg, sarcasm or irony) or clinical attribution, which could affect domain-specific estimates. Evidence-quote requirements and human validation mitigate but do not eliminate this risk. Fifth, findings are limited to 1 platform and English-language content; discourse may differ across platforms and languages. Finally, this cross-sectional design cannot establish causality, assess within-person changes, or link discourse to clinical outcomes. Because nondisclosure in social media reflects “unreported” rather than missing at random, our conservative handling of unreported variables may underestimate clinical relevance and attenuate associations. We mitigate these limitations through transparent reporting, stratified expert validation, and emphasis on associations rather than causal claims. Future work should consider time-stratified sampling, multiplatform triangulation, and linkage to external clinical data (eg, registries, claims, or electronic health record–linked cohorts) where feasible.

Conclusions

This study highlights a substantial misalignment between guideline-based statin risk stratification and the lived experiences expressed by patients online. Across 5328 Reddit discussions, discourse centered on perceived adverse effects (n=1697, 31.9%), emotional distress (n=4537, 85.2%), and mentions of “natural” alternatives (n=2485, 46.6%), patterns that may help explain persistent challenges in long-term adherence. These findings suggest that effective cardiovascular prevention must go beyond information provision to address

XSL•FO
RenderX

nocebo effects and treatment-related frustration that shape patients' responses to therapy. Methodologically, our validated LLM-enabled pipeline captures nuanced dimensions of treatment experience (eg, ambivalence and peer influence) that are poorly measured by conventional surveillance approaches. As patients increasingly navigate AI-mediated information environments, evidence-informed engagement with digital spaces may support shared decision-making and improve long-term outcomes.

Acknowledgments

During manuscript preparation, the authors used ChatGPT (OpenAI) for language editing to improve clarity and readability. The tool was used only to assist with wording and expression. All artificial intelligence–generated suggestions were critically reviewed and revised by the authors, who take full responsibility for the accuracy, integrity, and originality of the final manuscript.

Funding

This work was supported by the National Institutes of Health (grant R00LM014097).

Data Availability

The datasets generated or analyzed during this study are not publicly available because sharing Reddit-derived data could increase the risk of user traceability and would be inconsistent with the study's privacy protections.

Authors' Contributions

Conceptualization: SL, JL
Data curation: SL
Formal analysis: SL
Investigation: SL
Methodology: SL, JL
Software: SL
Writing—original draft: SL
Writing—review and editing: SL, JL

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data cleaning and filtering.
[DOCX File, 18 KB - jmir_v28i1e85057_app1.docx]

Multimedia Appendix 2
Large language model–based extraction protocol and prompt.
[DOCX File, 26 KB - jmir_v28i1e85057_app2.docx]

Multimedia Appendix 3
Expert validation protocol.
[DOCX File, 32 KB - jmir_v28i1e85057_app3.docx]

Multimedia Appendix 4
Example posts with large language model–derived theme and sentiment classifications.
[DOCX File, 17 KB - jmir_v28i1e85057_app4.docx]

Multimedia Appendix 5
Frequency of specific adverse effects mentioned in Reddit discussions on statin therapy.
[DOCX File, 20 KB - jmir_v28i1e85057_app5.docx]

References

1. Collins R, Reith C, Emberson J, et al. Interpretation of the evidence for the efficacy and safety of statin therapy. Lancet 2016 Nov 19;388(10059):2532-2561. [doi: 10.1016/S0140-6736(16)31357-5] [Medline: 27616593]
2. Grundy SM, Stone NJ, Bailey AL, et al. 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA guideline on the management of blood

XSL•FO
RenderX

cholesterol: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol 2019 Jun 25;73(24):3168-3209. [doi: 10.1016/j.jacc.2018.11.002] [Medline: 30423391]
3. Visseren FLJ, Mach F, Smulders YM, et al. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. Eur J Prev Cardiol 2022 Feb 19;29(1):5-115. [doi: 10.1093/eurjpc/zwab154] [Medline: 34558602]
4. Zhang H, Plutzky J, Skentzos S, et al. Discontinuation of statins in routine care settings: a cohort study. Ann Intern Med 2013 Apr 2;158(7):526-534. [doi: 10.7326/0003-4819-158-7-201304020-00004] [Medline: 23546564]
5. Basios A, Markozannes G, Ntzani EE, et al. Prevalence and determinants of adherence to statin therapy: a systematic review and meta-analysis. Eur J Prev Cardiol 2025 Dec 16:zwaf769. [doi: 10.1093/eurjpc/zwaf769] [Medline: 41401202]
6. Vinogradova Y, Coupland C, Brindle P, Hippisley-Cox J. Discontinuation and restarting in patients on statin treatment: prospective open cohort study using a primary care database. BMJ 2016 Jun 28;353:i3305. [doi: 10.1136/bmj.i3305] [Medline: 27353261]
7. Anker SD, Agewall S, Borggrefe M, et al. The importance of patient-reported outcomes: a call for their comprehensive integration in cardiovascular clinical trials. Eur Heart J 2014 Aug 7;35(30):2001-2009. [doi: 10.1093/eurheartj/ehu205] [Medline: 24904027]
8. Magnani JW, Mujahid MS, Aronow HD, et al. Health literacy and cardiovascular disease: fundamental relevance to primary and secondary prevention: a scientific statement from the American Heart Association. Circulation 2018 Jul 10;138(2):e48-e74. [doi: 10.1161/CIR.0000000000000579] [Medline: 29866648]
9. Brinton EA. Understanding patient adherence and concerns with STatins and MedicatION Discussions With Physicians (ACTION): a survey on the patient perspective of dialogue with healthcare providers regarding statin therapy. Clin Cardiol 2018 Jun;41(6):710-720. [doi: 10.1002/clc.22975] [Medline: 29749101]
10. Lardon J, Abdellaoui R, Bellet F, et al. Adverse drug reaction identification and extraction in social media: a scoping review. J Med Internet Res 2015 Jul 10;17(7):e171. [doi: 10.2196/jmir.4304] [Medline: 26163365]
11. Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inform Assoc 2015 May;22(3):671-681. [doi: 10.1093/jamia/ocu041] [Medline: 25755127]
12. Golder S, O'Connor K, Wang Y, Klein A, Gonzalez Hernandez G. The value of social media analysis for adverse events detection and pharmacovigilance: scoping review. JMIR Public Health Surveill 2024 Sep 6;10:e59167. [doi: 10.2196/59167] [Medline: 39240684]
13. Somani S, van Buchem MM, Sarraju A, Hernandez-Boussard T, Rodriguez F. Artificial intelligence-enabled analysis of statin-related topics and sentiments on social media. JAMA Netw Open 2023 Apr 3;6(4):e239747. [doi: 10.1001/jamanetworkopen.2023.9747] [Medline: 37093597]
14. Press - Reddit. Reddit. URL: https://www.redditinc.com/press [accessed 2025-05-20]
15. Rupert DJ, Gard Read J, Amoozegar JB, et al. Peer-generated health information: the role of online communities in patient and caregiver health decisions. J Health Commun 2016 Nov;21(11):1187-1197. [doi: 10.1080/10810730.2016.1237592] [Medline: 27805496]
16. Golder S, O'Connor K, Hennessy S, Gross R, Gonzalez-Hernandez G. Assessment of beliefs and attitudes about statins posted on Twitter: a qualitative study. JAMA Netw Open 2020 Jun 1;3(6):e208953. [doi: 10.1001/jamanetworkopen.2020.8953] [Medline: 32584408]
17. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85. [doi: 10.2196/jmir.1933] [Medline: 23615206]
18. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 2025 Jan 28;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
19. Goglia D, Vega D. Structure and dynamics of growing networks of Reddit threads. Appl Netw Sci 2024;9:48. [doi: 10.1007/s41109-024-00654-y]
20. Golder S, Ahmed S, Norman G, Booth A. Attitudes toward the ethics of research using social media: a systematic review. J Med Internet Res 2017 Jun 6;19(6):e195. [doi: 10.2196/jmir.7082] [Medline: 28588006]
21. Language identification. FastText. URL: https://fasttext.cc/docs/en/language-identification.html [accessed 2025-05-20]
22. Broder AZ. On the resemblance and containment of documents. Presented at: Compression and Complexity of SEQUENCES 1997; Jun 13, 1997. [doi: 10.1109/SEQUEN.1997.666900]
23. Introducing GPT-4.1 in the API. OpenAI. 2025. URL: https://openai.com/index/gpt-4-1/ [accessed 2025-04-14]
24. Croskerry P. A universal model of diagnostic reasoning. Acad Med 2009 Aug;84(8):1022-1028. [doi: 10.1097/ACM.0b013e3181ace703] [Medline: 19638766]
25. Grundy SM, Stone NJ, Bailey AL, et al. 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA guideline on the management of blood cholesterol: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Circulation 2019 Jun 18;139(25):e1082-e1143. [doi: 10.1161/CIR.0000000000000625] [Medline: 30586774]

XSL•FO
RenderX

26. US Preventive Services Task Force, Mangione CM, Barry MJ, et al. Statin use for the primary prevention of cardiovascular disease in adults: US Preventive Services Task Force recommendation statement. JAMA 2022 Aug 23;328(8):746-753. [doi: 10.1001/jama.2022.13044] [Medline: 35997723]
27. Franzke AS, Bechmann A, Zimmer M, Ess C. Internet research: ethical guidelines 3.0. Association of Internet Researchers. 2020 URL: https://aoir.org/reports/ethics3.pdf [accessed 2026-04-06]
28. Ingersgaard MV, Helms Andersen T, Norgaard O, Grabowski D, Olesen K. Reasons for nonadherence to statins - a systematic review of reviews. Patient Prefer Adherence 2020;14:675-691. [doi: 10.2147/PPA.S245365] [Medline: 32308373]
29. Cholesterol Treatment Trialists' Collaboration. Effect of statin therapy on muscle symptoms: an individual participant data meta-analysis of large-scale, randomised, double-blind trials. Lancet 2022 Sep 10;400(10355):832-845. [doi: 10.1016/S0140-6736(22)01545-8] [Medline: 36049498]
30. Gupta A, Thompson D, Whitehouse A, et al. Adverse events associated with unblinded, but not with blinded, statin therapy in the Anglo-Scandinavian Cardiac Outcomes Trial-Lipid-Lowering Arm (ASCOT-LLA): a randomised double-blind placebo-controlled trial and its non-randomised non-blind extension phase. Lancet 2017 Jun 24;389(10088):2473-2481. [doi: 10.1016/S0140-6736(17)31075-9] [Medline: 28476288]
31. Zhou Z, Ryan J, Ernst ME, et al. Effect of statin therapy on cognitive decline and incident dementia in older adults. J Am Coll Cardiol 2021 Jun 29;77(25):3145-3156. [doi: 10.1016/j.jacc.2021.04.075] [Medline: 34167639]
32. Cholesterol-lowering drugs get labeling changes. U.S. Food and Drug Administration. 2015. URL: https://www.fda.gov/drugs/special-features/cholesterol-lowering-drugs-get-labeling-changes [accessed 2025-11-10]
33. Cooper-DeHoff RM, Niemi M, Ramsey LB, et al. The Clinical Pharmacogenetics Implementation Consortium Guideline for SLCO1B1, ABCG2, and CYP2C9 genotypes and statin-associated musculoskeletal symptoms. Clin Pharmacol Ther 2022 May;111(5):1007-1021. [doi: 10.1002/cpt.2557] [Medline: 35152405]
34. SEARCH Collaborative Group, Link E, Parish S, et al. SLCO1B1 variants and statin-induced myopathy--a genomewide study. N Engl J Med 2008 Aug 21;359(8):789-799. [doi: 10.1056/NEJMoa0801936] [Medline: 18650507]
35. Howard JP, Wood FA, Finegold JA, et al. Side effect patterns in a crossover trial of statin, placebo, and no treatment. J Am Coll Cardiol 2021 Sep 21;78(12):1210-1222. [doi: 10.1016/j.jacc.2021.07.022] [Medline: 34531021]
36. Rosenson RS, Miller K, Bayliss M, et al. The Statin-Associated Muscle Symptom Clinical Index (SAMS-CI): revision for clinical use, content validation, and inter-rater reliability. Cardiovasc Drugs Ther 2017 Apr;31(2):179-186. [doi: 10.1007/s10557-017-6723-4] [Medline: 28421332]
37. DiMatteo MR, Lepper HS, Croghan TW. Depression is a risk factor for noncompliance with medical treatment: meta-analysis of the effects of anxiety and depression on patient adherence. Arch Intern Med 2000 Jul 24;160(14):2101-2107. [doi: 10.1001/archinte.160.14.2101] [Medline: 10904452]
38. Hoogendoorn CJ, Krause-Steinrauf H, Uschner D, et al. Emotional distress predicts reduced type 2 diabetes treatment adherence in the Glycemia Reduction Approaches in Diabetes: a comparative effectiveness study (GRADE). Diabetes Care 2024 Apr 1;47(4):629-637. [doi: 10.2337/dc23-1401] [Medline: 38227900]
39. Gonzalez JS, Shreck E, Psaros C, Safren SA. Distress and type 2 diabetes-treatment adherence: a mediating role for perceived control. Health Psychol 2015 May;34(5):505-513. [doi: 10.1037/hea0000131] [Medline: 25110840]
40. Chou WYS, Oh A, Klein WMP. Addressing health-related misinformation on social media. JAMA 2018 Dec 18;320(23):2417-2418. [doi: 10.1001/jama.2018.16865] [Medline: 30428002]
41. Bode L, Vraga EK. See something, say something: correction of global health misinformation on social media. Health Commun 2018 Sep;33(9):1131-1140. [doi: 10.1080/10410236.2017.1331312] [Medline: 28622038]
42. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. J Biomed Inform 2024 Mar;151:104620. [doi: 10.1016/j.jbi.2024.104620] [Medline: 38462064]
43. Liu M, Okuhara T, Chang X, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res 2024 Jul 25;26:e60807. [doi: 10.2196/60807] [Medline: 39052324]

Abbreviations

AI : artificial intelligence
aOR: adjusted odds ratio
API: application programming interface
ASCVD: atherosclerotic cardiovascular disease
LLM: large language model
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

XSL•FO
RenderX

Edited by A Mavragani; submitted 30.Sep.2025; peer-reviewed by L Wreyford, Y Zhu; revised version received 02.Feb.2026; accepted 24.Mar.2026; published 10.Apr.2026.

Please cite as:
Liu S, Liu J
Large Language Model–Based Analysis of Statin Therapy Discussions and Sentiment on Social Media: Cross-Sectional Observational Study
J Med Internet Res 2026;28:e85057
URL: https://www.jmir.org/2026/1/e85057
doi:10.2196/85057

© Siru Liu, Jialin Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

How Online Scheduling Platforms Affect Insurance-Based Disparities in Access to Specialist Outpatient Care in Berlin, Germany: Cross-Sectional Audit Study

Cordula Charlotte Josefine Kreuzenbeck[1,2], PhD; Silvia Angerer[2], PhD, Priv-Doz

[1]IU International University of Applied Science, Juri-Gagarin-Ring 152, Erfurt, Germany

[2]Department for Public Health, Health Services Research and Health Technology Assessment, UMIT TIROL—Private University for Health Sciences and Health Technology, Hall in Tirol, Austria

Corresponding Author:
Cordula Charlotte Josefine Kreuzenbeck, PhD
IU International University of Applied Science, Juri-Gagarin-Ring 152, Erfurt, Germany

Abstract

Background: Online scheduling platforms are increasingly chosen by patients for scheduling outpatient appointments. Due to payment for listing or platform decisions on listing visibility, they can amplify access inequalities. Especially in Germany's dual insurance system, the beneficiary difference in waiting times for private health insurance (PHI) patients compared to statutory health insurance (SHI) patients for specialist appointments might increase.

Objective: The objective of this study was to quantify differences in waiting times for the first available specialist outpatient appointments for SHI versus PHI profiles on a large commercial online scheduling platform (Doctolib) in Berlin and to assess whether practices offering equal appointment options for both insurance types are deprioritized by the platform.

Methods: We performed a cross-sectional, internet-based audit between January 6, 2025, and February 26, 2025. Two standardized simulated patient profiles (SHI and PHI) were used to query 1867 platform listings; practices with at least 1 bookable slot for both profiles were documented (n=492). The primary outcome was the waiting time in days for the earliest available appointment. Paired 1-tailed *t* tests and Wilcoxon signed-rank tests were used to compare within-provider waiting times, while exploratory subgroup analyses examined specialty-specific differences. Sensitivity analyses excluded the top 5% of the longest waits. Effect sizes (Hedges *g*) were computed.

Results: Across 492 practices, mean waiting times were 50.4 (SD 52.2) days for SHI and 22.6 (SD 36.8) days for PHI, yielding a mean difference of 27.8 days (95% CI 23.6 - 31.9 d; t_{491}=13.05; *P*<.001; Hedges *g*=0.59). After excluding the top 5% of waiting times (n=44), the difference remained significant (Δ=23.1 d; 95% CI 20.0 - 26.2 d; t_{447}=14.77; *P*<.001). Providers that offered earlier PHI appointments showed substantially longer SHI waiting times (61.8 vs 37.5 d; *P*<.001) and substantially shorter PHI waits (12.0 vs 38.2 d; *P*<.001). Platform ranking was correlated with earlier PHI availability (Spearman rank correlation *r*=0.574; *P*<.001).

Conclusions: On this commercial online scheduling platform in Berlin, SHI patients experienced substantially longer waits for the first available specialist appointments than PHI patients across multiple specialties. Potential mechanisms that could amplify pre-existing insurance-based access inequities are discussed. Transparency about the platform's ranking criteria should be considered to promote equitable access.

(J Med Internet Res 2026;28:e82452) doi:10.2196/82452

KEYWORDS

health services accessibility; health care disparities; insurance; algorithms; medical informatics applications; internet; health policy; health

Introduction

Disparities in access to medical care—particularly in scheduling outpatient appointments—are well-documented in health services research [1-3]. In Germany, this issue is especially pronounced due to its dual insurance system, which separates statutory health insurance (SHI) and private health insurance (PHI). Approximately 11% of the population is insured under PHI [4].

Field experiments have shown that SHI patients face significantly longer waiting times than PHI patients when attempting to schedule specialist appointments by phone, with reported differences ranging from 7 to 15 days [5-9]. Earlier data found that, in 2006, average waiting times for specialist

XSL•FO
RenderX

appointments ranged from 9.5 to 24.8 working days [10]. In contrast, waiting time disparities have not been observed to the same extent in the primary care sector in Germany [11].

Recently, online scheduling platforms have become increasingly common and may soon replace traditional phone-based scheduling in many settings [12,13]. Atherton et al [12] report a mixed methods investigation of patient use and experience with online scheduling platforms in primary care, focusing on adoption, usability, and user perceptions rather than audit outcomes [12]. Studies that examine online scheduling platforms find lower no-show rates and overall shorter waiting times but do not assess inequalities [13,14].

While online scheduling platforms offer convenience and the appearance of neutrality, emerging evidence suggests they may also perpetuate or even reinforce access inequities. A US-based study of an online scheduling platform in primary care found systematic discrimination against Medicaid patients. Kurtzman et al conducted an audit of online primary care appointment availability in the United States and identified differences in availability by insurance status (Medicaid vs private insurance). Although over 4000 listings were analyzed, this study did not simulate real bookings and did not focus on specialist care [15]. Our study adds a cross-specialty, Germany-based audit on a commercial online scheduling platform, extending the online-audit literature from primary care settings in the United States to specialist bookings in a European dual insurance system.

Online scheduling platforms can implement algorithms and rules that may be discriminatory by design, which can be enforced through the application of machine learning [16]. Particularly with the power to decide how the algorithm determines visibility on the listing page and the position in the list, online scheduling platforms can influence the metrics regarding how practices offer appointment slots to patients.

This study adds field evidence that disparities based on insurance status are present on a commercial online platform in practice. The analyzed platform, Doctolib, is a self-reported leading online scheduling platform in Europe. It introduces an active process of applying rules to the platform, through which patients can access appointments, and it comes with a monthly fee for physicians.

This study aims to investigate whether SHI patients experience longer waiting times than PHI patients when booking specialist appointments through an online scheduling platform. Specifically, we examine (1) the magnitude of waiting time differences between SHI and PHI patients across multiple medical specialties on the platform and (2) whether practices offering equal appointment options for both insurance types are deprioritized by platform algorithms. In doing so, our study contributes to the growing literature on digital health access by extending online audit methodologies beyond primary care and the US context. It provides the first cross-specialty, Germany-based audit of a commercial online scheduling platform, offering empirical evidence that platform design may potentially reinforce existing disparities in access to specialist care in a dual insurance system.

Methods

Study Design

We conducted a cross-sectional, internet-based audit study to evaluate differences in appointment availability for SHI versus PHI patients on Doctolib in Berlin, Germany. The study follows the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines (Checklist 1) for cross-sectional research and adopts a “simulated patient” methodology to minimize measurement bias [17].

Setting and Data Source

Data were collected between January 6, 2025, and February 26, 2025. The platform aggregates publicly accessible listings for outpatient medical specialists in Berlin and allows users to filter by insurance status (SHI vs PHI). To mitigate temporal variability, each specialty was assessed within 1 day for both the PHI and the SHI profiles. Due to frequent platform updates, the order of appearance changed throughout the assessment window. To ensure complete uptake of all available specialists, all search result pages were saved as static HTML snapshots at the moment of extraction. These snapshots were then double-checked to confirm that all possible appointments were incorporated. For validation, the assessment day was included as a possible confounder.

Participants and Sample Selection

The specialties were derived from the European Union Professional Qualifications Directive (2005/36/EC) of the European Parliament and of the Council of September 7, 2005, on the recognition of professional qualifications [18]. The study excluded the specialties of dentistry and orthodontics, which showed no observable insurance-based differences in a prior separate prestudy conducted in Frankfurt. Additionally, for these specialties, there was no option to select between SHI and PHI provided on Doctolib. We screened 1867 provider listings on the platform; inclusion criteria were as follows: (1) active online booking availability during the data collection window and (2) availability of appointments for both SHI and PHI. Providers offering only 1 insurance type or no online booking were excluded. A total of 492 practices met the inclusion criteria and were analyzed.

Audit Simulation and Data Collection Procedure

The following provides a step-by-step overview of the data collection procedure. First, 2 standardized profiles were created: profile 1 representing an SHI patient and profile 2 representing a PHI patient (profile 1: male, aged 42 y, no prior appointments or medical history on the platform; profile 2: female, aged 36 y, no prior appointments or medical history on the platform).

Next, the online scheduling platform Doctolib page was opened, and the specialty was chosen. According to the profile, the filter for PHI or not was chosen. For each specialty, 2 parallel searches—1 under each profile—were conducted on the same day and, wherever possible, within the same browser session to control for diurnal fluctuations in availability. For multiphysician practices or multilocation listings, each distinguishable physician or location combination was treated

XSL•FO
RenderX

as a unit. For each unit (listing of a physician), the earliest available appointment date without completing a booking by following the booking process was recorded. For simplicity, we refer to these units as "practices" throughout the paper, even though a single practice may have multiple listings.

To avoid systematic variation in the type of appointment requested, 1 of 2 standard booking options, identical for both patient profiles, was consistently used: (1) at priority 1, "first consultation new patient," and if this was not available, (2) "unspecific appointment within consultation hours" was chosen for all specialties, reflecting a typical new patient inquiry. The only exception was radiology, where we consistently selected magnetic resonance tomography or computed tomography (hand; magnetic resonance tomography priority 1 and computed tomography priority 2) because no general appointment categories were offered.

Before the final booking, the day available for booking was noted, and the process was stopped. No real appointments were made, ensuring that study activities did not impact providers' schedules. Each page was saved as a screenshot before moving on to the next page of listings. Doctolib provides 20 listings per page. All listings were documented (name of listing or practice and, in 3 specialties, also the address), including those that did not offer any appointments for any profile, which led to the final dataset with practices providing appointments for both PHI and SHI.

The process was repeated for each practice and profile until the last provider within Berlin. The first practice outside of Berlin was identified by the platform, which provided a note on the kilometer distance to Berlin and the name of the other city. When 1 specialty was fully assessed by both profiles and validation was done, the process was repeated for a different specialty on the same or another day.

After data collection for both profiles in each specialty, the insurance status was switched between the 2 profiles (female to SHI and male to PHI), and the first 20 listings were reassessed to investigate potential confounding with gender- or age-based discrimination due to profile construction.

The full list of variables extracted from Doctolib can be accessed in Multimedia Appendix 1. For the first research question on waiting time disparities, we used waiting times in days for the SHI and PHI profiles on an appointment for each practice as the primary outcome. Waiting times were defined as the difference in days between the day of the "First available appointment" and the "Day of data assessment." For subgroup analysis, we differentiated between the specialties of practice listings. The study also included the "Rank on Doctolib" for some specialties to address the second research question.

Bias Control

To reduce measurement bias and maintain consistency, the data collection was conducted in the order presented by the platform (first to last page and top to bottom), and static snapshots were archived.

Statistical Analysis

Descriptive statistics (mean and median) were computed for waiting times by insurance type. Paired 1-tailed *t* tests and nonparametric Wilcoxon signed-rank tests were used to assess within-provider differences in waiting times for SHI versus PHI. To explore heterogeneity, subgroup analyses stratified by specialty were conducted; however, given the small sample sizes, the results are descriptive only and should be interpreted cautiously. All analyses were 1-tailed with α=.05. To adjust for multiple comparisons when stratifying by specialty, a conservative Bonferroni correction was applied, which involves multiplying the unadjusted *P* values by the number of hypotheses (in our case, subgroups) tested [19]. Analyses were performed in SPSS, version 28 (IBM Corporation). Effect sizes (Hedges *g*) were computed.

Sensitivity Analysis

To evaluate the robustness of the findings, we excluded the top 5% of practices with the longest waiting times (outliers) and repeated the paired comparisons. This outlier exclusion did not materially alter the direction or significance of the overall results; however, some specialty subgroups lost sufficient power for significance.

Ethical Considerations

This study involved only publicly available booking information and did not involve any real patient data, physician data, or actual appointment bookings. Neither were provider schedules impacted, nor were provider names reported. Due to the study design being observational and not involving any human participants, no formal ethics approval was applied for. This follows the guidelines from the German Research Society, where none of the cases that require an ethical vote apply to this study design [20].

Results

Listings and Bookable Practices

A total of 1867 provider listings were screened on the platform between January 6, 2025, and February 26, 2025. After applying our exclusion criteria (no online booking, availability for only 1 insurance type, or specialty outside the predefined list), 492 practices remained with at least 1 bookable appointment slot for both SHI and PHI patients (bookable practices with appointments for both groups). Figure 1 illustrates the detailed inclusion and exclusion processes.

Multimedia Appendix 2 compares bookable practices with all available practices in Berlin using data from the Association of Statutory Health Insurance Physicians (Kassenärztliche Vereinigung [KV]) to estimate the platform's approximate local market share. Note that the Berlin KV registry lists all outpatient specialists authorized to treat statutory-insured patients; however, platform listings include multiple entries per physician and might omit some practices, so the KV comparison is approximate.

Overall, 492 of 4123 physicians listed on the platform offer services to SHI patients, accounting for approximately 12.6% of the available practices in the relevant specialties in Berlin,

XSL•FO
RenderX

according to the KV. This estimate should be considered a lower bound, as the lists do not correspond perfectly. In particular, the KV categories comprise broader specialty groupings (eg, neurology or psychiatry) than those represented on Doctolib, which leads to discrepancies in the comparison.

Figure 1. Extraction of listings, exclusions with exclusion criteria, and final dataset. Units: listings of practices. PHI: private health insurance; SHI: statutory health insurance.

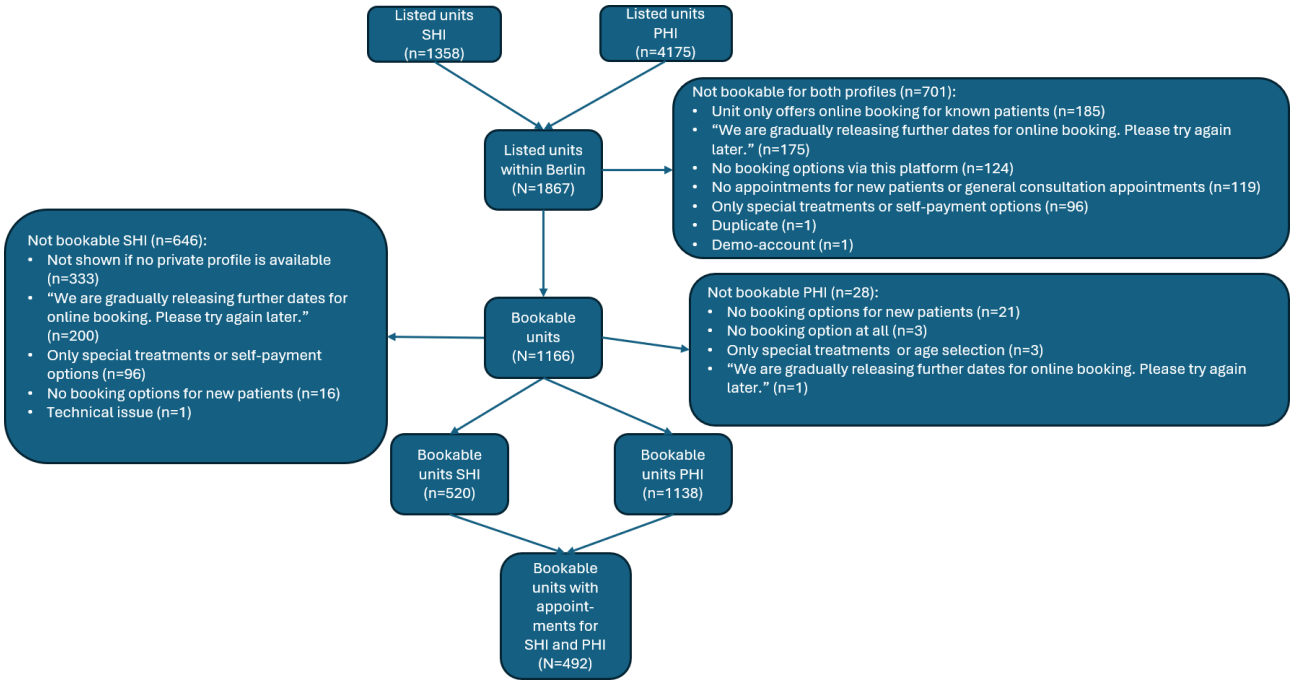

Provider Bookability by Specialty

Table 1 provides an overview of the distribution of cases among the specialties, the day of observation, and the descriptive results on waiting times.

The specialties with the highest SHI bookability rates were ENT (ear, nose, and throat), orthopedics, obstetrics and gynecology, and urology. Neurology had the lowest SHI bookability rate (0.5%), with only 2 out of 405 registered neurologists offering SHI appointments on this platform online.

XSL•FO
RenderX

Table . Descriptive statistics of waiting times for statutory health insurance (SHI) and private health insurance (PHI) patients by practice specialization.

Specialty	n	Day of observation	PHI<SHI	No difference	SHI<PHI	PHI<SHI (%)	SHI, mean (SD); median (IQR)	PHI, mean (SD); median (IQR)
General surgery	13	Tuesday	7	5	1	53.8	51.9 (40.0); 57 (21-69)	17.0 (21.9); 8 (2-21)
Ophthalmology	20	Sunday	8	11	1	40	48.6 (58.3); 19 (2-86)	17.2 (35.0); 3 (2-9)
Obstetrics and gynecology	103	Sunday	46	41	16	44.7	57.9 (54.7); 44 (15-86)	37.9 (48.6); 19 (4-58)
Vascular surgery	3	Sunday	0	2	1	0	39.7 (39.5); 38 (1-80)	51.7 (24.5); 38 (37-80)
Dermatology	18	Monday	6	11	1	33.3	63.0 (43.7); 55 (28-99)	47.4 (44.1); 36 (9-78)
ENT[a]	75	Wednesday	45	27	3	60	50.6 (35.9); 49 (21-78)	19.9 (29.5); 6 (1-23)
Cardiology	14	Monday	11	2	1	78.6	174.8 (123.2); 206 (52-283)	44.1 (95.8); 2 (0-23)
Pediatrics	12	Sunday	0	12	0	0.0	13.5 (18.6); 2 (1-26)	13.5 (18.6); 2 (1-26)
Neurology	2	Monday	1	1	0	50.0	82.0 (5.7); 82 (78-86)	81.5 (6.4); 82 (77-86)
Orthopedics	159	Tuesday	121	29	9	76.1	36.2 (31.4); 29 (13-55)	11.2 (16.8); 6 (1-13)
Radiology	38	Sunday	15	21	2	39.5	19.6 (25.2); 7 (2-25)	8.9 (16.8); 2 (1-8)
Urology	35	Saturday[b]	23	10	2	65.7	81.7 (52.5); 89 (47-107)	31.6 (32.2); 20 (4-58)
All specialties	492	—[c]	283	172	37	57.5	50.4 (52.2); 37 (10-74)	22.6 (36.8); 7 (1-29)

[a]ENT: ear, nose, and throat.

[b]Because the day following the observation was a Sunday—on which no bookings could be made—waiting times for both SHI and PHI patients were systematically reduced by 1 day.

[c]Not applicable.

Paired *t* tests and nonparametric Wilcoxon signed-rank tests were conducted across specialties to compare the first available appointment waiting times for SHI versus PHI profiles. Specialties with fewer than 3 evaluable cases (vascular surgery and neurology) were excluded from hypothesis testing. Pediatrics did not show any differences, so hypothesis testing could not be performed. Table 2 summarizes the results.

Across all specialties (including those with small sample sizes), SHI patients waited, on average, 27.8 (47.2) days longer than PHI patients (95% CI 23.6 - 31.9; t_{491}=13.05; P<.001; Hedges g=0.59), corresponding to approximately 4 weeks.

At the specialty level, statistically significant differences in waiting times between SHI and PHI patients were observed in 8 out of 9 specialties (P<.05). After applying a Bonferroni correction to account for multiple comparisons, 6 of these differences remained statistically significant (denoted as footnote “a” in Table 2), indicating that the overall pattern is not solely driven by chance findings. The largest disparities occurred in cardiology and urology. Results at the specialty level should, however, be interpreted cautiously, given the small sample sizes.

Given the skewed distribution of waiting times in some specialties, nonparametric Wilcoxon signed-rank tests are reported alongside *t* tests. The results remain qualitatively unchanged.

XSL•FO
RenderX

Table . Paired *t* test and Wilcoxon signed-rank test results for the differences between statutory health insurance (SHI) and private health insurance (PHI) waiting times for the first available appointment by practice specialization.

Specialty[a]	n[b]	SHI–PHI, mean difference[c] (SD)	95% CI	*t* test[d]		Wilcoxon signed-rank test[e], *P* value	Hedges *g*[f]
				t test (*df*)	*P* value		
General surgery	13	34.9 (42.7)	9.12 to 60.72	2.95 (12)	.01	.02	0.79
Ophthalmology	20	31.4 (46.5)	9.65 to 53.16	3.02 (19)	.01	.01	0.66
Obstetrics & gynecology	103	20.1 (43.7)	11.51 to 28.61	4.65 (102)	<.001	<.001	0.46
Dermatology	18	15.6 (31.6)	–0.10 to 31.33	2.10 (17)	.05	.05	—[g]
ENT[h]	75	30.7 (40.4)	21.43 to 40.02	6.59 (74)	<.001	<.001	0.76
Cardiology	14	130.7 (125.8)	58.11 to 203.32	3.89 (13)	.002	.003	1.01
Pediatrics	12	0	—	—	—	—	—
Orthopedics	159	25.0 (31.7)	20.00 to 29.93	9.93 (158)	<.001	<.001	0.79
Radiology	38	10.7 (22.5)	3.28 to 18.09	2.92 (37)	.01	.001	0.47
Urology	35	50.0 (56.6)	30.57 to 69.48	5.23 (34)	<.001	<.001	0.87
All specialties	492	27.8 (47.2)	23.57 to 31.93	13.05 (491)	<.001	<.001	0.59

[a]Specialties in obstetrics and gynecology, ENT, cardiology, orthopedics, radiology, and urology indicate statistically significant differences ($P<.05$) after Bonferroni correction.

[b]Reports the number of observations.

[c]Presents the mean difference in waiting times (in days) between the first available appointment for SHI and PHI patients.

[d]Report the results of the 1-tailed *t* test.

[e]Report the results of the Wilcoxon signed-rank test.

[f]Provides the effect size, calculated as Hedges *g*.

[g]Not available.

[h]ENT: ear, nose, and throat.

Figure 2 provides a boxplot showing the statistically significant difference in waiting times across all specialties between the SHI and PHI profiles.

XSL•FO
RenderX

Figure 2. Boxplot showing the distribution of waiting times for the first available appointment, comparing statutory health insurance (SHI) and private health insurance (PHI) insurance profiles.

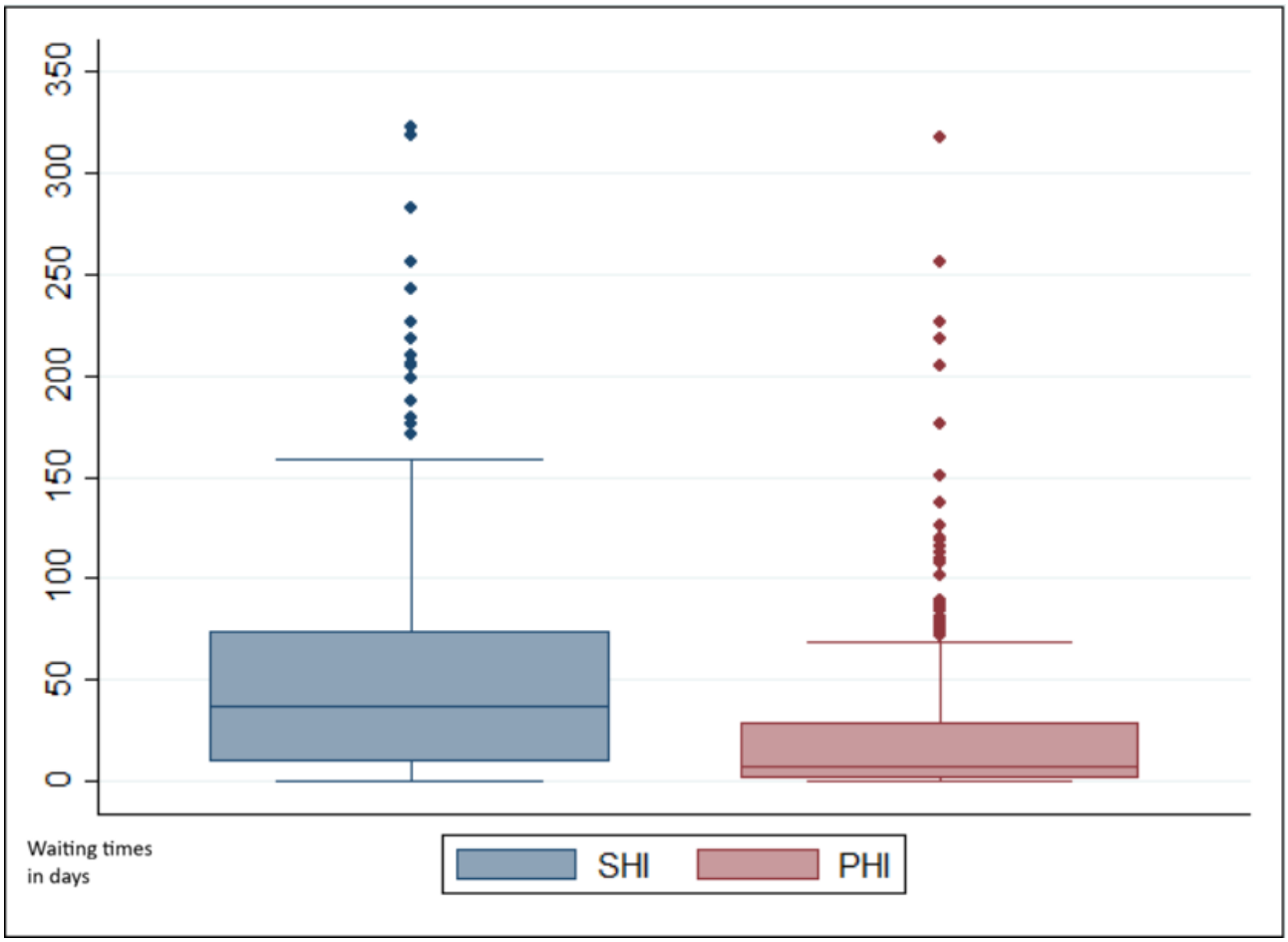

Comparison of Providers With and Without Faster PHI Appointments

Given that physicians differentiate appointment allocations between SHI and PHI patients, we investigated whether longer waiting times for SHI patients correspond to shorter waiting times for PHI patients. Therefore, we examined whether offering faster appointment slots to PHI patients was associated with waiting times for both SHI and PHI profiles. Providers were classified into 2 groups as follows:

- Group 0 (practices with equal SHI and PHI appointment timing): providers whose earliest SHI and PHI appointments did not differ (n=172).
- Group 1 (practices with earlier PHI than SHI availability): providers whose PHI appointment was earlier than their SHI appointment (n=283).

A total of 37 practices that offered an earlier appointment to SHI patients were excluded from this analysis.

Table 3 shows that SHI waiting times were significantly longer for group 1 (61.9 d) compared with group 0 (37.6 d), with a mean difference of −24.3 days (95% CI −33.6 to −14.9 d; t_{454}=−5.10, P<.001; Hedges g=−0.469). SHI patients wait 24.3 days less when they visit practices from group 0.

PHI waiting times were significantly shorter in group 1 (12.1 d) compared with group 0 (37.6 d), with a mean difference of 25.5 days (95% CI 18.2 - 32.9 d; t_{454}=6.81, P<.001; Hedges g=0.75). This shows that the group of practices with earlier PHI availability than SHI availability can offer appointments to PHI patients 25.5 days earlier than those that do not differentiate between SHI and PHI patients.

These results indicate that practices offering faster slots to PHI patients not only reduced PHI waiting times substantially but also correspondingly increased SHI waiting times, with large effect sizes in both comparisons.

For the specialties of obstetrics and gynecology, dermatology, urology, and ENT (n=801), we were able to reproduce the ranking on Doctolib exactly due to the exact name and address notation in our dataset and the corresponding screenshots from the search with the PHI profile. With 493 practices having available rankings and PHI appointments, we performed the following correlation analysis: rank on Doctolib (place of listing where 1 is the first listing on the page and counting) and waiting time for PHI have a highly significant (P<.001) Spearman rank correlation r of 0.702. This indicates a strong positive correlation. These findings demonstrate that the ranking order of entries on Doctolib is associated with appointment availability. Specifically, listings offering the earliest available appointments are positioned highest in the ranking.

Of the specialties with information on the rank, 175 also had appointments available for both PHI and SHI patients. An unpaired t test showed that practices in group 0 are, on average,

XSL•FO
RenderX

58 ranks behind compared to group 1 (95% CI 37.9-78.0; t_{174}=5.706, P<.001; Hedges g=0.86). As the listings on Doctolib have 20 practices per page, practices with equal SHI and PHI appointment timing will, on average, be placed almost 3 pages behind practices with earlier PHI than SHI availability on this platform in the view of PHI profiles.

Table . Descriptive statistics, tests, and effect sizes for waiting times by provider group[a].

Variable and group	n	Mean (SD)	Group 0–group 1, mean difference (95% CI)	1-tailed *t* test		Levene *F* test		Effect sizes: Hedges *g*[b]
				t test (*df*)	*P* value	*F* test (df)	*P* value	
Wait time (SHI)[c]			−24.3 (−33.60 to −14.91)	−5.104 (454)	<.001	4.208 (454)	.04	−0.469
Group 0	172	37.6 (45.1)						
Group 1	283	61.9 (55.2)						
Wait time (PHI)[d]			25.5 (18.15 to 32.92)	6.810 (454)	<.001	61.094 (454)	<.001	0.747
Group 0	172	37.6 (45.1)						
Group 1	283	12.1 (25.3)						

[a]Levene test assessed equality of variances. Mean difference (0–1): group 0 (practices with equal SHI and PHI appointment timing) mean minus group 1 (practices with earlier PHI than SHI availability) mean (so negative for SHI indicates longer waits in group 1).

[b] Effect sizes: Hedges *g* used pooled SD.

[c]SHI: statutory health insurance.

[d]PHI: private health insurance.

Validation Analysis: Exclusion of the Top 5% of Outliers

To confirm the robustness of our primary findings, we repeated the paired *t* tests and Wilcoxon signed-rank tests after excluding the top 5% (n=44) of the longest waiting times, yielding a sample of 448 practices. Table 4 presents the group statistics and inferential results for SHI versus PHI waiting times in this reduced sample by specialists and overall.

After removing extreme waiting time outliers in both SHI and PHI patients, SHI patients continued to experience significantly longer waits than PHI patients, with an average difference of 23.1 days (95% CI 20.0 - 26.2; t_{447}=14.77; P<.001). This validation analysis confirms the primary finding of a substantial disparity in appointment waiting times based on insurance status.

At the specialty level, statistically significant differences between SHI and PHI waiting times were observed in 7 of 12 specialties; 5 of these remained significant after Bonferroni correction (denoted as footnote “a” in Table 4). The largest disparities were found in the specialties of urology, general surgery, and ENT.

XSL•FO
RenderX

Table . Paired *t* test and Wilcoxon signed-rank test results for the difference between statutory health insurance (SHI) and private health insurance (PHI) waiting times for the first available appointment by practice specialization after excluding the top 5% of outliers.

Specialty[a]	n[b]	SHI–PHI, mean difference[c] (SD)	95% CI	*t* test[d] (df)		Wilcoxon signed-rank test[e], *P* value	Hedges *g*[f]
				t	*P* value		
General surgery	13	34.9 (42.7)	9.1 to 60.7	2.949 (12)	.01	.02	0.766
Ophthalmology	16	22.4 (33.0)	4.8 to 39.9	2.714 (15)	.02	.04	0.644
Obstetrics and gynecology	89	17.4 (32.3)	10.6 to 24.2	5.075 (88)	<.001	<.001	0.533
Vascular surgery	3	−12.0 (20.8)	—[g]	—	—	—	—
Dermatology	13	15.2 (25.9)	−0.4 to 30.9	2.119 (12)	.06	.02	0.550
ENT[h]	71	31.6 (36.8)	22.9 to 40.3	7.235 (70)	<.001	<.001	0.849
Cardiology	5	24.8 (34.0)	−17.4 to 67.0	1.630 (4)	.18	.17	0.582
Pediatrics	12	—	—	—	—	—	—
Neurology	1	—	—	—	—	—	—
Orthopedics	156	23.3 (27.7)	19.0 to 27.7	10.530 (155)	<.001	<.001	0.839
Radiology	38	10.7 (22.5)	3.3 to 18.1	2.924 (37)	.01	<.001	0.465
Urology	31	45.8 (48.7)	27.9 to 63.6	5.234 (30)	<.001	<.001	0.916
All specialties	448	23.1 (33.1)	20.0 to 26.2	14.77 (447)	<.001	<.001	0.59

[a]Specialties in obstetrics and gynecology, ENT, orthopedics, radiology, and urology indicate statistically significant differences ($P<.05$) after Bonferroni correction.

[b]Reports the number of observations.

[c]Presents the mean difference in waiting times (in days) between the first available appointment for SHI and PHI patients.

[d]Report the results of 1-tailed *t* test.

[e]Report the results of the Wilcoxon signed-rank test.

[f]Provides the effect size, calculated as Hedges *g*.

[g]Not available.

[h]ENT: ear, nose, and throat.

As with the main specialty-level analyses, these findings should be interpreted cautiously, given the small sample sizes within individual specialties, which may affect the precision and stability of the estimates.

Validation Analysis: Profile Construction and Potential Confounding

To assess potential confounding due to gender- and age-based discrimination in profile construction, we reassessed the first 20 listings within each specialty after switching the insurance status assigned to the 2 profiles. Across all specialties, the earliest available appointment date remained consistently linked to the insurance status indicated in the profile rather than to the gender or age of the profile. Specifically, when the male profile, aged 42 years, was assigned PHI status, it received the same earliest appointment option previously observed for the female PHI profile, aged 36 years, and vice versa for SHI. No deviations from this pattern were observed. These findings suggest that the differences in waiting times are most likely attributable to insurance status rather than gender- or age-related factors embedded in the profile design.

Discussion

Main Study Findings

In this cross-sectional, internet-based audit of 492 outpatient specialists listed on a major European booking platform, we found that SHI patients in Germany faced substantially longer waiting times for the first available appointments compared with PHI patients. To our knowledge, no previous study of comparable scale has examined appointment allocation via online scheduling platforms in a dual insurance system of PHI and SHI, specifically for specialist care.

Across multiple medical specialties, SHI patients waited, on average, 23 to 32 days longer than PHI patients. Additionally, for PHI patients, there were more than twice as many practices with offerings listed on the platform compared to SHI patients in the same region.

Although more than one-third of practices offered identical appointment dates to both profiles, these physicians appeared at a significantly later rank in search results and had significantly longer waiting times for PHI patients. The practices with earlier PHI than SHI availability, on the other hand, showed

XSL•FO
RenderX

significantly longer waiting times for SHI patients and had a higher rank in the listing. Thus, the preferential ranking of earlier appointment slots advantages practices whose PHI slots precede their SHI slots.

Analysis of the Findings

Our findings indicate that SHI patients face significant disparities in outcomes on the platform, both in terms of reduced appointment availability and longer waiting times. The observed difference in waiting times, averaging 23 to 32 days, is larger than what was demonstrated in earlier telephone-based audit studies in Germany [5-10]. For example, Breitenbach and Heinrich [5] documented waiting times for SHI patients that were, on average, 15 days longer than for PHI patients in major German cities. Similarly, Heinrich et al [6] reported differences of 14 to 23 days for elective procedures in a high-density urban region (excluding Berlin). Other studies found smaller gaps: in Bavaria, Muschol and Gissel [8] observed differences of 9 days in 2019 and 6 days in 2020, while Werbeck et al [9] reported an average difference of 13 weekdays across 36 counties in all 16 German federal states. Lungen et al [10] found differences ranging from 5 to 25 working days depending on the specialty and the requested intervention. This variation across specialties mirrors the pattern observed in our study, where disparities between SHI and PHI patients also differ substantially by medical specialty. By contrast, disparities in primary care appear much smaller; Luque Ramos et al reported differences of less than 1 day [11].

Taken together, these comparisons suggest that disparities observed through online scheduling platforms are at the upper end—or exceed—the range documented in prior German studies using telephone-based appointment scheduling. Evidence on insurance-based differences in appointment allocation on online scheduling platforms is scarce. The only study investigating insurance-based differences on online scheduling platforms was conducted in the United States and examined appointment scheduling with primary care physicians [12]. It found no differences in the timing of available appointments by insurance status; however, patients insured through Medicaid were directed to physicians located farther away, indicating disparities in geographic access. Comparability with this study is, however, limited, as the analysis was conducted within the US health care system and focused on primary care, whereas this study examines access to outpatient specialist appointments in the German context of statutory and private health insurance.

Turning to the mechanism that may help explain the observed disparities in appointment availability and waiting times: one explanation relates to economic incentives within the dual insurance system. Reimbursement levels for PHI patients are generally higher and less tightly regulated than those for SHI patients, potentially incentivizing practices to allocate scarce appointment slots preferentially to PHI patients. Consistent with this interpretation, our audit revealed that a notable subset of practices (n=96) presented additional self-pay or privately billed service options exclusively to SHI profiles. These offers often took the form of enhanced screening packages, extended consultation times, self-pay appointments, or cosmetic procedures, and were either absent or framed differently for the PHI profile. This pattern suggests that some practices actively promote out-of-pocket services to SHI patients as an alternative revenue stream. Such differential presentation of self-pay options not only adds another layer of complexity to access—where SHI patients face not only longer waiting times but also additional financial considerations—but may also reflect incentives to prioritize higher-margin services over standard insured care. From a health-economics perspective, these dynamics could exacerbate inequities by shifting SHI patients toward more costly, noncovered services.

Building on this, and in connection with our result that practices with earlier PHI than SHI availability are ranked higher in the platform listing, the platform's ranking logic—while not explicitly designed to favor PHI patients—could indirectly amplify differences in access by directing patients toward practices where PHI patients receive faster appointments. In this sense, ranking based on waiting times may introduce a second layer of prioritization: practices that already allocate appointments preferentially to PHI patients become more visible and attractive within the platform environment. This dynamic may create reinforcing incentives for practices to prioritize PHI patients, since shorter waiting times improve their relative ranking and potentially attract additional demand. Over time, such feedback mechanisms could widen inequalities in outpatient care access between SHI and PHI patients, even if the platform's design does not explicitly intend to differentiate by insurance status.

While differential reimbursement and platform ranking are potential mechanisms, the observed pattern may also reflect alternative explanations, such as differences in expected case complexity, expected consultation length, or differential demand patterns between SHI and PHI patients. Because the audit design cannot distinguish empirically between these alternative explanations, further research should investigate the underlying mechanisms and whether, and to what extent, ranking amplifies insurance-based disparities in access to specialist care.

Although the share of practices using Doctolib is still limited in Berlin, the use of online scheduling platforms is on the rise [12,13], and other platforms work in a similar manner (eg, Jameda). Thus, although this study was conducted in Berlin using Doctolib, the results are likely informative for a broader population of patients and providers actively using online scheduling platforms in other regions and on other platforms. Further research is needed to assess the generalizability of these findings across different regions and online scheduling platforms.

These findings also raise questions about the role of platform design in shaping access to care. From a policy perspective, this suggests that safeguards—such as algorithmic audits or insurance-neutral ranking criteria—could be considered to mitigate the risk of unequal access. Greater transparency regarding the factors that influence platform rankings may also be important to better understand how digital scheduling systems affect the distribution of appointments across patient groups.

XSL•FO
RenderX

Limitations

This study offers a valuable contribution by systematically quantifying insurance-based disparities in access to specialist care through online scheduling platforms, a growing channel in health care delivery. Nevertheless, several limitations warrant consideration.

First, our sampling frame comprised only practices listed on a single online platform; approximately 12.6% of KV - registered specialists offered SHI appointments on this platform. Our results, therefore, do not capture the full range of outpatient care pathways, including appointment scheduling via phone or in - person. Additionally, the sample is not representative as the opt-in to the paid use of this platform is likely to induce a selection bias by itself. Providers participating in online scheduling platforms are likely a nonrandom subset of all outpatient specialists. Adoption may be higher among digitally oriented practices, larger group practices, or those serving a higher share of PHI or self-paying patients. In addition, practices facing higher demand or capacity constraints may have stronger incentives to actively manage appointment allocation via digital tools. As a result, our findings are not generalizable to the appointment scheduling behavior of all outpatient physicians. This selection into the platform could affect the magnitude of the observed disparities. On the one hand, practices adopting digital scheduling tools may have more standardized appointment management and greater transparency in slot allocation, which could attenuate differences between SHI and PHI patients compared with traditional scheduling channels. On the other hand, digital platforms may facilitate more explicit segmentation of appointment slots or patient groups, potentially amplifying differences in waiting times if practices strategically allocate capacity. Compared with evidence from telephone-based appointment scheduling studies, the latter explanation seems more plausible [5,6,8-11]. Future research should, therefore, explicitly examine the determinants of providers' participation in online scheduling platforms and how such selection may influence observed patterns of access.

Nevertheless, even if our findings are not generalizable to the appointment scheduling behavior of all outpatient physicians, they do reflect the actual experiences of patients who rely on online booking platforms to access care. Given the growing use of digital scheduling systems, these platforms increasingly shape how patients search for and obtain appointments. Consequently, the observed patterns may become more relevant for patient access to care over time.

Second, the platform's proprietary ranking algorithms are opaque; while we see a correlation suggesting an "earliest - appointment" prioritization, other factors may influence the ranking, given that this association does not imply causation.

Third, our audit design captures only the first available slots and cannot account for cancellations, patient no - shows, or subsequent appointment opportunities. The "first available appointment" is a useful and reproducible audit metric for comparing what the platform exposes to different insurance profiles at a given moment. However, it is only an approximation of actual access. Real patient pathways may differ because patients can accept cancellations, request urgent slots, choose a preferred physician, switch providers, or obtain appointments through nondigital channels. As a result, the observed waiting-time gap may not translate one-to-one into realized delays in care, especially if SHI and PHI patients differ in their ability or willingness to use such alternative pathways. In particular, because the study measures visible first available slots rather than completed appointments, the estimated differences should be interpreted as differences in booking opportunity, not necessarily as direct differences in treatment receipt.

Fourth, 2 profiles were not perfectly identical because they were based on real first-time user accounts that differed in gender and age; however, the age difference of 6 years can be considered negligible. In addition, we conducted a robustness check by switching SHI and PHI status across the profiles and reassessed the first 20 listings. This analysis confirmed the observed differences by insurance status, suggesting that the gap between SHI and PHI patients is unlikely to be driven by gender or age. The validation, therefore, supports the robustness of the insurance gradient, but it does not perfectly rule out gender- or age-based confounding.

Fifth, distributions of waiting times were skewed, so while the mean reflects the average experience of the patient profile in this study, the findings are likely to differ for all patients. This is reflected in the outlier exclusion, where cardiology was no longer significant and was mainly driven by the outliers.

Finally, although specialty-level estimates were directionally consistent with the main finding, these subgroup analyses were exploratory and underpowered in several strata. Accordingly, apparent between-specialty differences should not be interpreted as definitive evidence of specialty-specific effects. The Bonferroni-adjusted results reduce concerns about multiple testing but do not eliminate the power issues.

Conclusions

Online scheduling platforms, while improving convenience, appear to reaffirm offline inequities in specialist access: statutory - insured patients wait substantially longer than their PHI counterparts. These disparities are consistent across multiple specialties and robust to validation analyses. Importantly, the observed inequality reflects the structural features of Germany's dual insurance system rather than being created by the platform itself. However, if platform rankings are influenced by waiting times, differences in waiting times between SHI and PHI patients could further amplify existing inequalities by making practices with earlier PHI than SHI availability more prominent on the platform. As online booking becomes an increasingly dominant mode of access to outpatient care, systematic monitoring of its distributional effects is essential. Policymakers and platform operators should consider regulatory and design interventions that promote equitable access and greater transparency in the factors influencing platform rankings, helping to ensure that digitalization contributes to fairness rather than reinforcing existing inequalities in health care access.

XSL•FO
RenderX

Acknowledgments

The authors used ChatGPT (free version; OpenAI; ChatGPT-5 Mini) to support the manuscript writing process. It was used to summarize and draft for formal elements of the submission process. All content, analyses, and interpretations were developed, verified, and approved by the authors. ChatGPT was not used for data generation, analysis, or the creation of original scientific content. The artificial intelligence (AI) use for this manuscript falls under Domain 1 of the "Recommendations for a Classification of AI Use in Academic Manuscript Preparation" document. Responsibility for the final manuscript lies entirely with the authors. Generative AI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit organizations for the research or authorship. Open Access publication was enabled by IU International University of Applied Sciences.

Data Availability

The entire original dataset is available from the corresponding author upon reasonable request.

Authors' Contributions

CCJK conceived the study, designed the audit protocol, collected and validated the data, performed the analyses, and drafted and revised the manuscript. SA revised the manuscript, validated the statistical analyses, and contributed to it.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Variable list.
[DOCX File, 21 KB - jmir_v28i1e82452_app1.docx]

Multimedia Appendix 2
Estimated Kassenärztliche Vereinigung market share of Doctolib in Berlin based on practice specialization and registry data.
[DOCX File, 16 KB - jmir_v28i1e82452_app2.docx]

Checklist 1
STROBE checklist.
[PDF File, 55 KB - jmir_v28i1e82452_app3.pdf]

References

1. Angerer S, Waibel C, Stummer H. Discrimination in health care: a field experiment on the impact of patients' socioeconomic status on access to care. Am J Health Econ 2019 Oct;5(4):407-427. [doi: 10.1162/ajhe_a_00124]
2. Kim CY, Wiznia DH, Roth AS, Walls RJ, Pelker RR. Survey of patient insurance status on access to specialty foot and ankle care under the Affordable Care Act. Foot Ankle Int 2016 Jul;37(7):776-781. [doi: 10.1177/1071100716642015] [Medline: 27026727]
3. Rhodes KV, Kenney GM, Friedman AB, et al. Primary care access for new patients on the eve of health care reform. JAMA Intern Med 2014 Jun;174(6):861-869. [doi: 10.1001/jamainternmed.2014.20] [Medline: 24710808]
4. von dem Knesebeck O, Klein J. Attitudes towards the dual health insurance system and inequalities in health care in Germany—results of a population survey. BMC Health Serv Res 2025 May 14;25(1):696. [doi: 10.1186/s12913-025-12847-x] [Medline: 40369528]
5. Breitenbach A, Heinrich M. Diskriminierung im deutschen Krankenversicherungssystem: Werden gesetzlich Versicherte bei der Terminvergabe von Fachärzten benachteiligt? [Report in German]. : Social Science Open Access Repository (SSOAR); 2023 URL: https://www.ssoar.info/ssoar/bitstream/handle/document/85085/ssoar-2023-breitenbach_et_al-Diskriminierung_im_deutschen_Krankenversicherungssystem_Werden.pdf?sequence=1&isAllowed=y&lnkname=ssoar-2023-breitenbach_et_al-Diskriminierung_im_deutschen_Krankenversicherungssystem_Werden.pdf [accessed 2026-05-16]
6. Heinrich N, Wübker A, Wuckel C. Waiting times for outpatient treatment in Germany: new experimental evidence from primary data. Jahrb Natl Ökon Stat 2018 Sep 25;238(5):375-394. [doi: 10.1515/jbnst-2018-0025]
7. Roll K, Stargardt T, Schreyögg J. Effect of type of insurance and income on waiting time for outpatient care. Geneva Pap Risk Insur Issues Pract 2012 Oct;37(4):609-632. [doi: 10.1057/gpp.2012.6]

XSL•FO
RenderX

8. Muschol J, Gissel C. COVID-19 pandemic and waiting times in outpatient specialist care in Germany: an empirical analysis. BMC Health Serv Res 2021 Oct 11;21(1):1076. [doi: 10.1186/s12913-021-07094-9] [Medline: 34635091]
9. Werbeck A, Wübker A, Ziebarth NR. Cream skimming by health care providers and inequality in health care access: evidence from a randomized field experiment. J Econ Behav Organ 2021 Aug;188:1325-1350. [doi: 10.1016/j.jebo.2021.05.028]
10. Lungen M, Stollenwerk B, Messner P, Lauterbach KW, Gerber A. Waiting times for elective treatments according to insurance status: a randomized empirical study in Germany. Int J Equity Health 2008 Jan 9;7(1):1. [doi: 10.1186/1475-9276-7-1] [Medline: 18184426]
11. Luque Ramos A, Hoffmann F, Spreckelsen O. Waiting times in primary care depending on insurance scheme in Germany. BMC Health Serv Res 2018 Mar 20;18(1):191. [doi: 10.1186/s12913-018-3000-6] [Medline: 29558925]
12. Atherton H, Eccles A, Poltawski L, Dale J, Campbell J, Abel G. Investigating patient use and experience of online appointment booking in primary care: mixed methods study. J Med Internet Res 2024 Jul 8;26:e51931. [doi: 10.2196/51931] [Medline: 38976870]
13. Zhao P, Yoo I, Lavoie J, Lavoie BJ, Simoes E. Web-based medical appointment systems: a systematic review. J Med Internet Res 2017 Apr 26;19(4):e134. [doi: 10.2196/jmir.6747] [Medline: 28446422]
14. Nabila N, Ayuningtyas D. The effectivity of outpatient waiting time in hospital through online or web-based reservation (literature review). Asian J Eng Soc Health 2024 Aug 10;3(8):1725-1739. [doi: 10.46799/ajesh.v3i8.370]
15. Kurtzman GW, Keshav MA, Satish NP, Patel MS. Scheduling primary care appointments online: differences in availability based on health insurance. Healthc (Amst) 2018 Sep;6(3):186-190. [doi: 10.1016/j.hjdsi.2017.07.002] [Medline: 28757308]
16. Samorani M, Harris S, Blount LG, Lu H, Santoro MA. Overbooked and overlooked: machine learning and racial bias in medical appointment scheduling. Manuf Serv Oper Manag 2021;24:2825-2842. [doi: 10.1287/msom.2021.0999]
17. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
18. Directive 2005/36/EC of the European Parliament and of the Council of 7 September 2005 on the Recognition of Professional Qualifications (text with EEA relevance). European Union. 2005 URL: https://eur-lex.europa.eu/eli/dir/2005/36/oj/eng [accessed 2026-05-28]
19. Bonferroni CE. Il calcolo delle assicurazioni su gruppi di teste [Article in Italian]. Studi in onore del professore salvatore ortu carboni 1935:13-60 [FREE Full text]
20. FAQ: humanities and social sciences. Deutsche Forschungsgemeinschaft. 2023. URL: https://www.dfg.de/en/research-funding/proposal-funding-process/faq/humanities-social-sciences [accessed 2026-04-07]

Abbreviations

ENT: ear, nose, and throat
KV: Kassenärztliche Vereinigung (Association of Statutory Health Insurance Physicians)
PHI: private health insurance
SHI: statutory health insurance
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Stone; submitted 19.Aug.2025; peer-reviewed by A Kodan, E Birisci; revised version received 22.Apr.2026; accepted 22.Apr.2026; published 15.Jun.2026.

Please cite as:
Kreuzenbeck CCJ, Angerer S
How Online Scheduling Platforms Affect Insurance-Based Disparities in Access to Specialist Outpatient Care in Berlin, Germany: Cross-Sectional Audit Study
J Med Internet Res 2026;28:e82452
URL: https://www.jmir.org/2026/1/e82452
doi:10.2196/82452

© Cordula Charlotte Josefine Kreuzenbeck, Silvia Angerer. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Social Support Mechanisms in an Online Type 1 Diabetes Community: Social Network Analysis of Stakeholder Diversity and Disease Duration

Yujia Zhu[1,2*], MJC; Yanxiang Zhang[2*], PhD; Chenxiao Zhao[2], MJC; Xifeng Zeng[3], BE; Xueying Zheng[1], MD

[1]Department of Endocrinology and Metabolism, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, No. 5 Building, South Campus (Anhui Provincial Hospital), Hefei, Anhui, China

[2]Department of Communication of Science and Technology, University of Science and Technology of China, Hefei, Anhui, China

[3]Shenzhen Aibowei Biotechnology Co., Ltd. (Tangtangquan), Shenzhen, Guangdong, China

[*]these authors contributed equally

Corresponding Author:
Xueying Zheng, MD
Department of Endocrinology and Metabolism, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, No. 5 Building, South Campus (Anhui Provincial Hospital), Hefei, Anhui, China

Abstract

Background: Online health communities (OHCs) have emerged as critical platforms for patients with type 1 diabetes (T1D) to exchange informational and emotional support. However, how stakeholder roles and disease duration jointly shape support dynamics and influence formation remains underexplored.

Objective: This study aimed to examine network-based social support mechanisms in a large T1D OHC, focusing on how stakeholder diversity and disease duration are associated with social support behaviors, subnetwork structures, and user influence.

Methods: This retrospective observational study analyzed digital trace data from China's largest T1D online community (January 1-May 20, 2024), comprising 43,788 posts and 145,423 comments contributed by 1393 users. We manually annotated 2000 randomly sampled posts and fine-tuned a GPT-4o-mini (OpenAI) to classify support type (informational or emotional, and seeking or providing), yielding 20,384 support-related posts and 56,953 comments from 1224 users. We constructed weighted directed informational and emotional interaction networks and modeled predictors of a composite influence metric (Relative Centrality) using a gamma log-link generalized linear model (including demographics, identity, sentiment, disease duration, posting orientation, and cyclical activity time). Analyses were conducted in Python (version 3.11; Python Software Foundation). Statistical significance was set at $P<.05$.

Results: Support predominantly flowed from longer-duration members (≥ y) to those at earlier stages (≤5 y). Both subnetworks exhibited multicentered, star-like structures; the informational subnetwork had broader participation (density 0.031, diameter 7), while the emotional network was denser (density 0.039, diameter 6). In the influence model, peer supporters had substantially higher influence than patients (exp(β)=34.79, 95% CI 18.94 - 64.08; $P<.001$), professionals lower (exp(β)=0.41, 95% CI 0.17 - 0.99; $P=.055$), and women higher than men (exp(β)=1.65, 95% CI 1.23 - 2.23; $P=.001$). Positive sentiment was associated with higher influence (exp(β)=1.91, 95% CI 1.22 - 2.97; $P=.005$), and negative lower (exp(β)=0.54, 95% CI 0.37 - 0.79; $P=.001$). Influence followed an inverted U-shaped trajectory over disease duration, peaking at approximately the 116th month (95% CI 43.25 - 188.91).

Conclusions: This study suggests that social support patterns and user influence in a T1D OHC vary by stakeholder role and disease duration. Users with shorter disease duration more often sought support, whereas longer-duration users more often provided support, and informational and emotional exchanges formed distinct interaction subnetworks. Peer supporters were the most influential users; influence was also associated with gender, sentiment, activity timing, and a nonlinear (inverted U-shaped) relationship with disease duration. These findings may inform peer-facilitated, stage-tailored community strategies, with professionals engaged in targeted, complementary roles. A patient-centered collaborative care approach integrating peer experience with multidisciplinary clinical input could be explored in future work.

(J Med Internet Res 2026;28:e82996) doi:10.2196/82996

KEYWORDS

diabetes mellitus type 1; social support; online health communities; social network analysis; large language models; sentiment analysis; peer group

XSL•FO
RenderX

Introduction

Background

Type 1 diabetes (T1D) commonly develops during adolescence but can occur at any age [1]. As there is currently no cure, individuals diagnosed with T1D require lifelong management. Because clinical characteristics, complication risks, and mortality are closely associated with disease duration [2,3], the long-term demands of intensive self-management and cumulative complication risk may increase patients' psychosocial burden, potentially contributing to emotional distress, behavioral difficulties, and poor treatment adherence [4]. As of 2024, an estimated 599,000 individuals are living with T1D in China, accounting for approximately 6.5% of the global total [5]. Although the incidence of T1D in China remains relatively low, the large population size results in one of the highest national case counts worldwide [1]. The primary disease burden among Chinese individuals with T1D includes poor glycemic control, high rates of complications, and considerable financial strain [6]. Understanding long-term management and support mechanisms for T1D in China is particularly important, given the uneven access to offline social support and the growing role of digital health platforms in providing continuous care.

T1D management primarily occurs outside clinical settings. Due to geographic dispersion, many individuals with T1D struggle to connect with peers who share similar medical experiences [7], making family members and friends their primary sources of support. In contrast, online health communities (OHCs) provide more accessible social environments that facilitate peer interaction and experience sharing, and thus serve as critical platforms for social support exchange [8]. Studies have shown that when individuals share their disease trajectories, treatment outcomes, and coping strategies in OHCs, it facilitates the exchange of diabetes-related knowledge and health information, which can substantially enhance disease management. Beyond peer interaction, OHCs also serve as venues where health care professionals and other stakeholders can engage individuals with T1D. For example, physicians and nurses who actively participate in online discussions can bridge medical expertise with lived experiences, fostering trust, promoting adherence, and ultimately improving health outcomes [9]. Despite growing interest in online social support, existing OHC-based research has paid limited attention to the co-evolution of social support behaviors and social network structures in long-term T1D management. As a result, identifying and empowering key actors in support networks, along with developing sustainable T1D-specific online social support systems, remains a critical challenge for advancing targeted and durable health communication and intervention.

Social Support From a Diverse Stakeholder Perspective

Social support refers to information that leads individuals to perceive themselves as cared for, loved, respected, and integrated into a mutual aid network [10]. It is typically categorized into informational, emotional, instrumental, and appraisal support [11,12]. Among these, informational and emotional support are the most prevalent forms on social media, playing a crucial role in facilitating health decision-making and improving health outcomes [13,14]. In online communities, informational support is typically sought through question asking and advice seeking, whereas emotional support is expressed through sharing illness experiences, emotional reactions, and coping narratives [15,16]. The type and emotional tone of posts influence the likelihood of receiving support, with emotionally expressive negative posts more likely to elicit supportive responses [17,18]. Moreover, even users who do not actively participate in discussions can benefit indirectly from observing others' emotional exchanges, thereby gaining valuable informational resources in the process [19]. These patterns highlight the heterogeneity of support-seeking behaviors in T1D online communities and suggest that variation in support sources is a key factor in understanding how social support functions [14,20].

Contemporary research on online social support in T1D follows 2 primary trajectories. The first centers on professional support offered by health care providers, emphasizing medically guided knowledge and structured intervention. Health care professionals are among the most reliable sources of informational support [21], with numerous studies showing that such information is typically more accurate, evidence-based, and trusted by individuals with T1D [22]. As a sustainable resource, professional informational support complements the emotional support more commonly exchanged among peers, jointly promoting improved self-care and behavioral outcomes [22]. The second line of research focuses on experiential support, exchanged among patients and caregivers through shared experiences and mutual understanding. According to the principle of network homophily, individuals are more likely to trust and adopt advice from those with whom they share similar health trajectories [23,24]. Unlike professionally sourced support, peer-generated content tends to be more personalized, emotionally resonant, and reflective of day-to-day management challenges [25]. This type of experiential knowledge, accumulated through lived experience and long-term observation of disease progression, positions patients and caregivers as potential lay experts in the T1D community [26]. Taken together, these 2 trajectories underscore the need to understand how professional and experiential support interact within broader community networks, rather than in isolation.

Gender is a salient factor influencing patterns of seeking and providing social support in online communities, with implications for health outcomes [27]. In terms of posting motivation, women are generally more active in maintaining relationships through written interactions and post more frequently than men [28]. At the content level, men's posts are more technical and information-focused, often using medical terminology, whereas women are more likely to seek emotional support and express subjective concerns or negative emotions [8,29]. Women also report receiving more emotional support from social networks, while men more often rely on their partners as their primary source [30]. These gendered tendencies shape both the type of support exchanged and the emergent network configuration, providing context for subsequent analyses of stakeholder roles.

XSL•FO
RenderX

Beyond individual attributes, stakeholder interactions in T1D communities involve diverse and often asymmetrical support dynamics. The psychosocial burden of T1D, particularly its impact on quality of life, extends beyond individuals with T1D to caregivers and family members [31], shaping how different groups engage and what types of support they prioritize. Yet, most studies remain focused on dyadic support exchanges, neglecting the broader collaborative mechanisms across heterogeneous actors. Understanding these complex, multistakeholder patterns is critical to designing more effective and sustainable support systems in OHCs.

Social Support From the Perspective of Disease Duration

Following diagnosis, individuals with T1D undergo a prolonged adaptation process during which disease characteristics, complication risks, and mortality vary by age and disease duration [2,3]. Many also experience diabetes-related stigma, contributing to emotional and psychological distress [32]. Previous research highlights the temporal dynamics of social support in OHCs. Informational support often fulfills immediate needs, while emotional support fosters long-term engagement. Previous work by Wang et al [33] suggests that informational support is effective for addressing short-term challenges, whereas emotional support strengthens interpersonal ties, promoting long-term engagement in support groups [34]. These temporal patterns indicate that support needs evolve over the disease duration, underscoring the importance of timing in community-based interventions.

Support behaviors in type 2 diabetes communities shift with disease duration; prediabetic users typically seek information, while insulin-treated users more frequently provide both informational and emotional support [35]. Similar patterns are observed in T1D communities. Wang et al [36] found that newcomers primarily sought emotional support, but gradually shifted toward providing informational support as their engagement deepened. Patients and caregivers tend to seek more external support during early diagnosis to cope with uncertainty and emotional distress [13]. As experience accumulates, reliance on external information often decreases [37]. Accordingly, it has been recommended that self-management of education and support for T1D be tailored to key time points, such as diagnosis, unmet treatment goals, psychosocial or physiological transitions, and major changes in care [38]. However, little is known about how support behaviors evolve across stakeholder groups across the disease course, leaving a critical gap in understanding the sustainability of online social support systems. This gap motivates our first research question (RQ) 1.

Social Support Networks Within Diabetes OHCs

In public health, social connection is a key determinant of effective health management and long-term behavioral change [39], with perceived social support widely used as a proxy for tie strength. Social network analysis (SNA) offers a systematic approach to mapping interaction structures, identifying influential actors, and understanding network formation and evolution. Previous studies show that individuals occupying central positions in diabetes-related networks play a greater role in information diffusion [40]. Gender also shapes network structure. Although men typically maintain smaller social networks [41], they often occupy more central positions [8]. Disease duration is another factor. Dias et al [42] found that users diagnosed for less than 2 years tended to be more active, whereas those with longer disease duration (≥2 y) were more likely to occupy central positions, suggesting a potential association between disease duration and network centrality. This evidence motivates the second RQ, which examines how stakeholder interactions shape informational and emotional subnetworks, and informs hypotheses about influence formation.

In recent years, researchers have increasingly applied SNA to examine support networks among T1D stakeholders. Earlier work primarily focused on professional medical networks, emphasizing the roles of health care providers in information exchange and care coordination [43]. In contrast, Wu et al [19] adopted an experimental design to construct a Facebook (Meta)-based virtual peer network consisting of 212 individuals with T1D, in which 10 highly active participants were trained as peer leaders. These leaders exhibited centrality scores nearly 10 times higher than regular members, positioning them as influential brokers within the network. Their high betweenness centrality allowed them to bridge subgroups and facilitate diffusion of health behaviors and information [44-46]. Identifying such high-influence nodes is therefore critical to enhancing the effectiveness and sustainability of online T1D self-management systems [47], which motivates the third RQ3.

Limited attention has been paid to the mechanisms by which informational and emotional support circulate across diverse stakeholder groups over the course of T1D. Using data from China's largest T1D OHC, we analyzed user-generated posts and comments. A manually annotated subset of posts was used to fine-tune a large language model (LLM) for automated classification of social support types, followed by SNA to map interaction networks and generalized linear model (GLM) regression to examine factors associated with user influence. Drawing on social support theory, this research integrates stakeholder diversity and disease duration to examine who seeks and provides support across stakeholder groups, how these behaviors evolve with disease duration, and through what mechanisms influence forms among support participants. The findings aim to provide empirical evidence to inform personalized health interventions and the development of sustainable online social support systems, ultimately supporting improved health management for patients with T1D. Accordingly, we pose the following RQs:

- RQ1: What are the characteristics of social support interactions among diverse stakeholders in T1D online communities, and how do these interactions evolve across different stages of disease duration?
- RQ2: How do stakeholder interactions collectively shape the structural characteristics of informational and emotional support subnetworks?
- RQ3: Which stakeholder groups hold the highest influence in T1D online communities, and what mechanisms drive the formation of their influence?

XSL•FO
RenderX

Methods

Data Collection and Cleaning

We analyzed user-generated digital traces from the largest online T1D community in China, covering the period from January 1 to May 20, 2024. The dataset comprised 43,788 posts, 145,423 top-level comments, and profiles for 1393 users. Metadata included post and comment time stamps, user identifiers (ID and nickname), registration dates, gender, community identity, and time since diagnosis.

Before analysis, we preprocessed the post texts following methods outlined by Kumar et al [48]. First, irrelevant content such as "@username," "#Topic," and URLs were removed, and duplicate entries were eliminated, resulting in a final dataset of 42,321 posts. From the cleaned dataset, a simple random sample of 2000 posts was drawn for manual annotation. The coding scheme was developed by the conceptualization of social support in networks by House et al [11], support dimensions by Cutrona and Russell [49], and typology of support types by Myrick et al [50]. A subset of 500 posts was first selected as a coding sample and independently annotated by 2 trained researchers (Y Zhu and CZ). Posts were initially classified as either related or unrelated to social support. Those deemed relevant were then further categorized as informational or emotional in content, and finally distinguished as either seeking or providing support (Cohen κ=0.713). An additional 500 posts were subsequently coded to refine and validate the coding scheme, ensuring clarity, reproducibility, and mutually exclusive categories (Cohen κ=0.776). Detailed operationalizations of the key terms used in this study are provided in Multimedia Appendix 1. The remaining 1000 posts were then annotated by the coders, yielding a manually labeled dataset of 2000 posts in total (Cohen κ=0.802).

We fine-tuned the GPT-4o-mini (OpenAI) model on the manually labeled dataset, using 75% of the data for training and 25% for validation. The model was trained over 3 epochs with a batch size of 16. The full validation loss across the entire validation set was 0.073. In addition, the model achieved a validation accuracy of 0.977. For a rigorous and unbiased performance evaluation, we conducted an independent inference test on fresh model weights, achieving an accuracy of 0.878 with a macro F_1-score of 0.869, indicating acceptable generalization performance and reliable classification of relevant posts. After removing unrelated posts, the dataset retained 20,384 posts containing either informational or emotional support, along with 89,352 corresponding comments. After further cleaning, which involved removing irrelevant content, emojis, and self-replies from the comments, a final set of 56,953 valid comments and data from 1224 users were obtained.

Ethical Considerations

All personally identifiable information was anonymized during preprocessing. Ethical approval for this study was granted by the Clinical Trial Ethics Committee of Anhui Provincial Hospital (review 2019KY-Lunshen-27). No compensation was provided because no participants were recruited or contacted.

Sentiment Analysis

Sentiment analysis (SA), a text mining technique, is designed to extract individual opinions and detect the polarity and intensity of emotions expressed in textual data [51]. In this study, the sentiment knowledge enhanced pretraining model was used to perform sentence-level sentiment classification. This approach enables the extraction of fine-grained semantic features and improves classification accuracy [52,53]. For each post and comment, we implemented a binary classification task in which the model outputs a sentiment label (positive or negative) along with a confidence score ranging from 0 to 1. Higher scores indicate greater confidence in the predicted sentiment class. To ensure comparability across all samples, we defined a normalized sentiment score (SentiScore; Multimedia Appendix 2).

To ensure the reliability and accuracy of sentiment classification, 1721 posts related to informational and emotional support were selected from a manually labeled dataset of 2000 posts. These posts were independently annotated using a binary sentiment classification scheme by 2 trained researchers (Y Zhu and CZ), achieving a high interrater agreement (Cohen κ=0.814), indicating the dataset's suitability for the study context. To evaluate model performance, 75% of the dataset was used for training and 25% for testing. On the test set, the model achieved an accuracy of 94.7%, comparable with its 95.4% accuracy on the Chinese SA dataset ChnSentiCorp [53]. The model also achieved a macro F_1-score of 0.943, mitigating concerns related to class imbalance [54]. Additionally, precision and recall were 96.6% and 94.8%, respectively, with an area under the receiver operating characteristic curve of 94.6%. These results demonstrate that the sentiment knowledge enhanced pretraining model performs robustly in SA within the T1D context and effectively identifies users' emotional expressions in posts.

SNA

In social networks, user interactions are typically conceptualized as the formation of links. Users who participate in discussion threads are treated as nodes, and interactions between them form the edges of the network. Subnetworks represent specific subsets within the overall structure [55]. Drawing on users' posting and commenting behaviors in the T1D online community, we constructed a weighted directed network $G=(V, E)$. A directed edge (u, v) indicates that u commented on a post by v, with edge weights determined by the number of comments exchanged [56]. Network density captures overall connectivity, defined as the ratio of observed directed ties to all possible ties. The diameter of the network refers to the longest shortest path between any 2 nodes, reflecting the network's structural span and reach.

Network centrality metrics are widely used to identify influential actors in information diffusion processes [57-61]. Eigenvector centrality captures a node's influence based not only on its direct connections but also on the importance of the nodes to which it is connected [62]. As such, ties to highly influential nodes substantially elevate a node's status in the overall network structure. Betweenness centrality, on the other hand, captures the extent to which a node lies on the shortest paths between other pairs of nodes [58]. Nodes with high betweenness

XSL•FO
RenderX

centrality often occupy key positions in information flow. Accordingly, we used the NetworkX library to compute centrality measures, weighting eigenvector centrality by comment count and betweenness centrality by its inverse. Network visualization was conducted using Gephi (Gephi Consortium) [63].

Statistical Analysis

In the first 2 years following diagnosis, patients with T1D typically experience a marked decline in C-peptide levels, leading to severe insulin deficiency and pronounced glycemic variability [31]. The recommended time frame for screening and intervention for complications, such as retinopathy and nephropathy, is between 3 and 5 years post diagnosis [3,64]. With longer disease duration, the risk of complications associated with poor metabolic control increases. Notably, the risk of long-term complications, such as diabetic peripheral neuropathy, rises significantly after 10 years post diagnosis [65], substantially elevating patients' mortality risk and adversely impacting both their physical and mental health. Dividing patients based on key milestones in disease duration helps better understand their social support needs, enabling targeted and individualized health interventions. In this study, users from the T1D online community were grouped by disease duration—short duration (0 - 2 y and 3-5 y) and long duration (6 - 10 y and 11+ y). Given the skewed distribution of posting frequency, we first used the Kruskal-Wallis test to robustly assess the main effects of disease duration and gender. Interaction effects were then examined using a 2-way ANOVA. When Levene test indicated homogeneity of variance (P>.05), standard F tests were applied; otherwise, Welch ANOVA was used to mitigate the risk of Type I error [66]. For post hoc comparisons, Tukey's Honestly Significant Difference test was used when variance was equal, and the Games-Howell test was applied when it was not. The threshold for statistical significance was set at P<.05.

Statistical Variables

To assess the impact of gender on supportive behavior, gender was coded as a binary variable (man=1, woman=0). Community identity was modeled as a 4-category variable, including patient, guardian, peer supporter, and professional, with patient serving as the reference group. The professional category includes health care providers, dietitians, mental health counselors, and diabetes education specialists; for these users, disease duration reflects years of participation in the OHC.

Demographic attributes can shape a node's influence within a network, and an individual's marginal influence often declines with increasing social exposure [67]. Accordingly, disease duration was modeled in 2 forms, that is, a continuous variable (months) representing time since diagnosis for patients or years of service for professionals, and a quadratic term (months^2) to capture potential diminishing returns to network embeddedness. To control for heterogeneity in disease duration across identity groups, the variable was standardized (z score) within each group.

Mallipeddi et al [68] noted a positive correlation between users' emotional tone and their engagement on social media; emotionally charged content tends to be shared more rapidly and widely than neutral content. Such interactions are often used to assess user influence [69]. Building on this, this study calculated the mean and SE of users' SentiScore derived from their posts and comments. A z test (α=.001, z=3.291) was used to construct a 99.9% CI to determine overall sentiment orientation. Users were classified as positive if the lower bound of the interval exceeded the neutral threshold of 0.5, negative if the upper bound was below 0.5, and neutral otherwise. For users contributing no more than 1 post or comment, the raw SentiScore was used for classification in order to mitigate the impact of outliers in small samples and improve the robustness of sentiment categorization.

Previous studies have highlighted the importance of content value in shaping user influence, with popular posts significantly enhancing a user's centrality within the network [70]. Given individual variation, users' posting preferences may influence the mechanisms through which they gain influence. To capture this, we constructed a 2×2 normalized matrix based on post types, categorized along 2 dimensions: "seeking versus providing support" and "informational versus emotional support." From this matrix, row and column vectors were extracted to represent 2 continuous variables, postcontent orientation (PCO) and postinteraction orientation (PIO). The principal argument (θ) between these vectors was calculated and linearly scaled to the (−1, 1) interval. A higher PCO indicates a stronger preference for informational support, while a higher PIO reflects a greater tendency to provide support.

Research has shown that users tend to receive more comments and shares when posting content during specific time periods (eg, morning or noon) on social media [71,72]. In this study, user activity across different time periods may vary by identity, and if key groups post during high-interaction periods, their network influence could increase. To quantify users' activity patterns for posts and comments throughout the day, we first converted the average active time for each user into a circular mean, with values ranging from 0 to 24, and mapped it to radians. This resulted in the creation of a combined cyclical variable, cyclical activity time (Sine/Cosine), to capture their activity characteristics within the daily cycle.

To assess users' structural influence within the T1D social support network, this study constructed a composite metric of network centrality. Eigenvector centrality is commonly used to measure global influence, as interacting with highly central nodes can enhance a user's own network position [73]. Betweenness centrality reflects efficiency-driven interaction paths [59], aligning with the need-based behaviors observed in T1D OHCs, where users tend to connect with others who can meet their informational or emotional needs [74]. We then developed a composite metric "Relative Centrality," which combines eigenvector centrality and betweenness centrality to identify key influencers in the T1D support network. Principal component (PC) analysis was applied to reduce the dimensionality of the 2 centrality measures [75]. The first PC had an eigenvalue of 1.68 and accounted for 83.97% of the variance, indicating that PC1 effectively captures the shared variance between the 2 metrics. Relative centrality, derived

XSL•FO
RenderX

from the loading coefficients of PC1, was thus used as the core indicator of a user's influence within the network.

A GLM was fitted with relative centrality as the dependent variable, using a gamma-distribution with a log-link function to accommodate the outcome variable's positive skewness [76], with statistical significance set at $P<.05$. The ratio of residual deviance to degrees of freedom was 10.87 (>1), indicating potential overdispersion. All predictor variables had variance inflation factors between 1.021 and 1.663, all below 2, indicating no multicollinearity [77]. All analyses were conducted in Python (version 3.11; Python Software Foundation).

Results

Descriptive Statistics

In the T1D online community, patients accounted for 62.5% (765/1224) of users, followed by guardians (428/1224, 35%), professionals (24/1224, 2%), and peer supporters (7/1224, 0.6%). Women made up 61.1% (748/1224) of users, and men 38.9% (476/1224). The mean disease duration among patients was 7.8 (SD 7.0; median 6.7, IQR 2.7-11.3) years. In terms of social support interactions, the distribution of the 4 support types varied significantly across disease duration groups ($\chi^2(9)=1459.8$; $P<.001$). Posts providing emotional support were most common (n=8603, 42.2%), followed by informational support (n=5807, 28.5%), seeking emotional support (n=3311, 16.2%), and seeking informational support (n=2663, 13.1%).

Figure 1 illustrates the distribution of support types across stakeholder identities. Patients and guardians are the primary seekers of both informational and emotional support. Peer supporters exhibited the highest frequency of informational support provision, whereas patients contributed most frequently to emotional support. Professionals participated in both informational and emotional support interactions. Across all support types, women demonstrated higher levels of participation than men.

Kruskal-Wallis tests revealed significant differences in posting frequency by both disease duration (H=554.59, $P<.001$) and gender (H=946.51, $P<.001$), suggesting that T1D stakeholders' engagement in social support interactions is influenced by these factors. To delve deeper into the nature of these interactions, we next examined the specific dimensions of seeking and providing support. Given significant results from Levene tests ($P<.05$), indicating heterogeneity of variance, the more robust Welch-type ANOVA was conducted to examine differences in seeking and providing support behaviors.

Figure 2 indicates significant main effects and the interaction effects of both disease duration and gender on types of social support. Figure 3 illustrates the interaction effect between disease duration and gender. Overall, in the early stage (0 - 2 y), users predominantly sought support, with women showing particularly high demand for emotional support. Users gradually transitioned from seeking to providing support, with provision peaking at 3 - 5 years. For informational support provision, women consistently outperformed men across all disease durations. In contrast, the participation of men increased steadily with disease duration, reaching the highest frequency of informational support-seeking during the 6 - 10 year stage, while the engagement of women declined after 3 - 5 years. Gender differences in emotional support provision were modest, with similar activity levels between men and women in the 6 - 10 year and longer duration groups.

Games-Howell post hoc tests showed that support-seeking behaviors peaked during the 3 - 5 year postdiagnosis period, with higher frequencies than the 0 - 2 year group for both emotional support seeking (mean 112.74, SD 94.62 vs mean 38.36, SD 40.70; $P<.001$) and informational support seeking (mean 44.00, SD 36.45 vs mean 25.72, SD 36.34; Hedges g=−0.5; $P<.001$). In contrast, support-providing behaviors increased with disease duration, particularly for emotional support provision (≥11 y: mean 218.57, SD 151.52 vs 0 - 2 y: mean 82.04, SD 59.47; Hedges g=1.07; $P<.001$), while informational support provision reached its highest level at ≥11 years (mean 346.12, SD 196.20 vs mean 239.82, SD 203.12 at 0 - 2 y; $P<.001$). Regarding gender differences, women provided substantially more informational support than men (women: mean 298.15, SD 200.84 vs men: mean 65.37, SD 64.40; Hedges g=1.25; $P<.001$).

XSL•FO
RenderX

Figure 1. Distribution of informational and emotional support posts across stakeholder identities and genders in a retrospective observational study of China's largest type 1 diabetes online health community (January 1-May 20, 2024).

Figure 2. Main and interaction effects of disease duration stage and gender on the posting frequency of informational and emotional support. Results are based on a Welch-type ANOVA analyzing digital traces from a retrospective observational study of China's largest type 1 diabetes online health community (January 1-May 20, 2024).

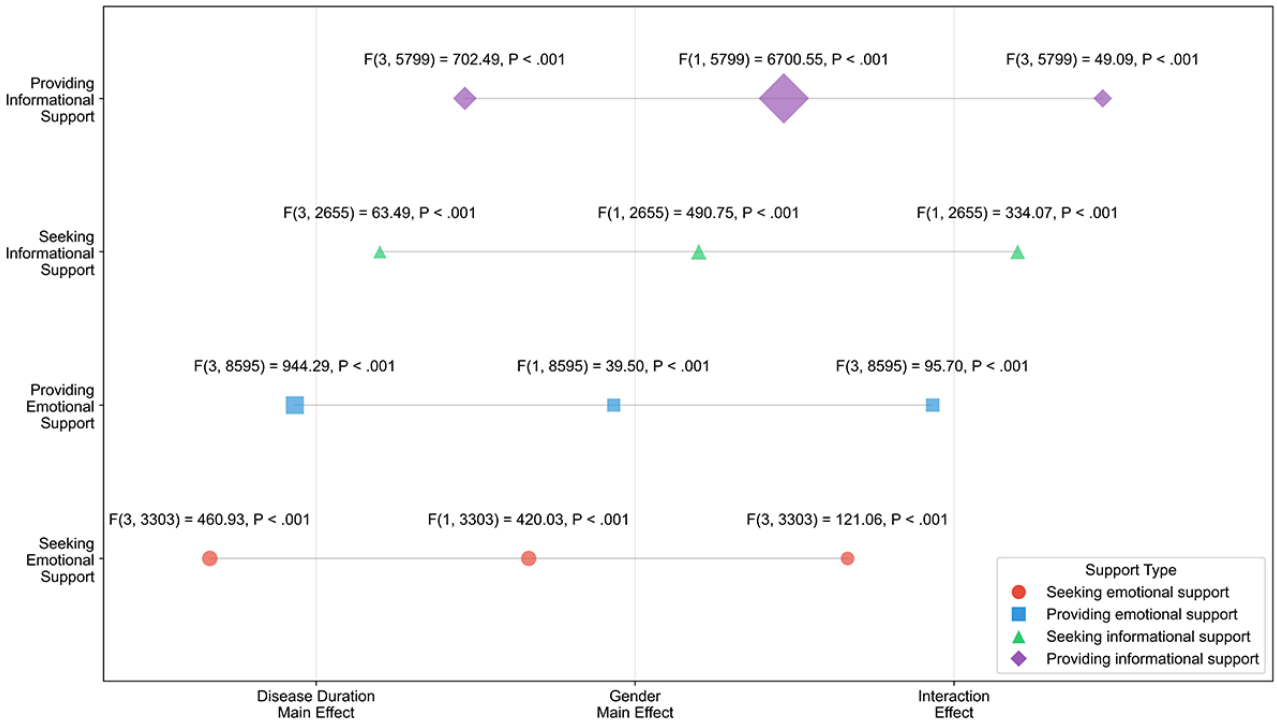

XSL•FO
RenderX

Figure 3. Interaction effects between disease duration stage and gender on social support behaviors (seeking vs providing, informational vs emotional) among users in China's largest type 1 diabetes online health community (January 1-May 20, 2024).

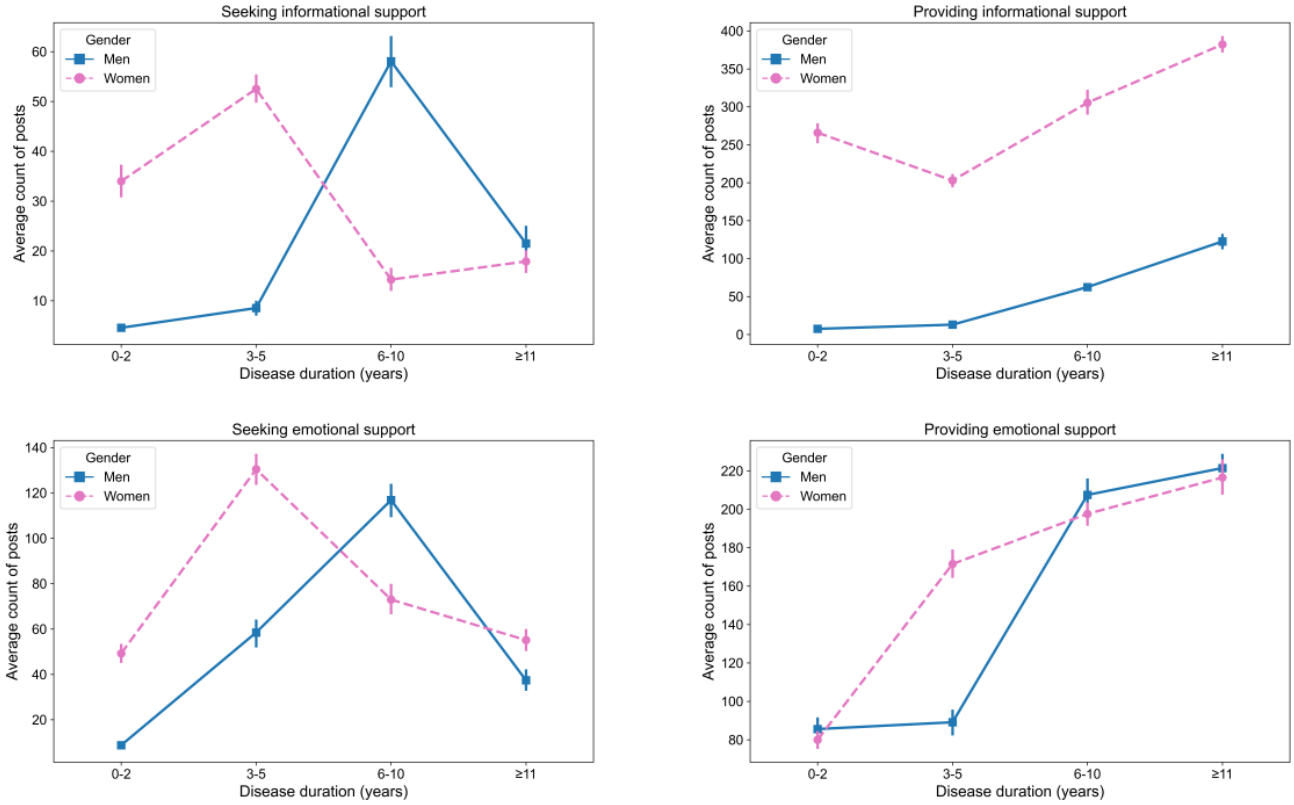

SNA

This study constructed a weighted directed network based on users' commenting behavior in the OHC, identifying key individuals through relative centrality measures. Table 1 illustrates the distinct structures of the informational and emotional support subnetworks. The informational support subnetwork involved a broader range of users; however, its lower density indicates that these interactions were relatively diffuse. In contrast, the emotional support subnetwork was smaller in scale and had a shorter network diameter. Yet, it featured more frequent interactions and higher network density.

Table . Structural characteristics of the informational and emotional social support subnetworks.

Subnetwork	Nodes	Edges	Density	Network diameter	Average degree
Informational Support	985	29,921	0.031	7	30.377
Emotional Support	829	27,032	0.039	6	32.608

Mapping Social Support Interactions Among Diverse Stakeholders

Figure 4 illustrates the social network structures of informational and emotional support interactions among diverse stakeholders. Both subnetworks exhibit a star-like topology, with multiple highly connected hubs linking to numerous peripheral nodes. Within both subnetworks, some peer supporters across different disease durations occupied central positions, forming dense local clusters and acting as key social influencers. Professionals, especially those with more than 5 years of service, were strategically placed and demonstrated high relative centrality in both networks. Some patients and guardians also formed local hubs. For example, guardians in the 0‑ to 2-year disease-duration group held influential positions in the emotional support network, while patients with ≥11 years of disease duration were central in the informational support subnetwork. In contrast, most patient and guardian nodes were sparsely connected, forming a typical core and periphery structure.

XSL•FO
RenderX

Figure 4. (A) Informational and (B) emotional social support interaction subnetworks from a retrospective observational study of China's largest type 1 diabetes online health community (January 1-May 20, 2024). The networks are weighted and directed, based on comment-to-post interactions. Each node represents a user, with size indicating Relative Centrality, color representing disease duration stage, and shape denoting stakeholder identity. Arrow colors indicate the type of support (an arrow from user X to user Y indicates user X provided or sought support from user Y).

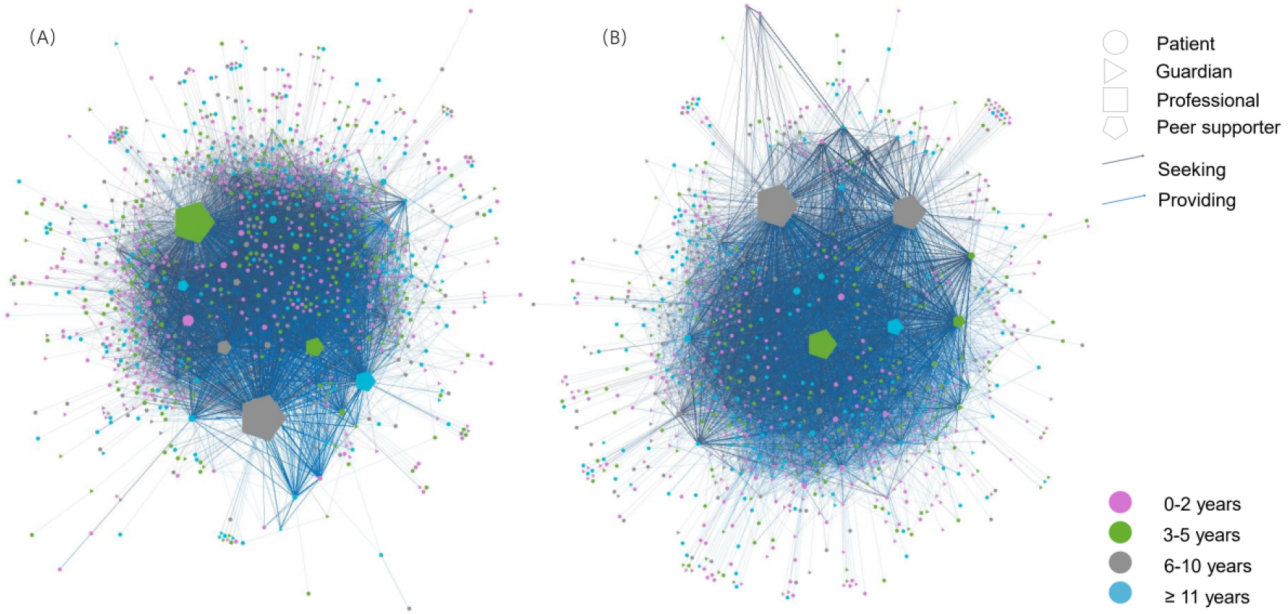

To further assess the flow of informational and emotional support among stakeholders with different disease durations, we applied Bayes theorem to infer conditional probabilities [78] and to estimate the conditional probability distributions of support-seeking and support-providing behaviors. Probabilities were normalized based on the total support interactions initiated by each disease-duration group toward other groups.

In social networks, posters typically initiate threads, while commenters engage in follow-up interactions that are often asynchronous and temporally lagged. To avoid interpretive ambiguities in statistical analysis, this study adopts 2 conditional perspectives—one based on the disease duration of posters and the other on that of commenters (refer to Multimedia Appendix 2 for details.) Figure 5 presents the conditional probability distributions of social support provision and seeking across different disease duration groups. In Figure 4, node colors correspond to different stages of disease duration, while node shapes represent different stakeholder identities within the community. Duration groups are denoted as G_i (i=1,2,3,4), with posting users as receive group R_{Gi}, and commenting users as send group S_{Gi}.

In the T1D OHC, social support behaviors vary significantly by disease duration. When a stakeholder posts a request for support, individuals in the 0 - 2 year group are the least likely to provide either informational or emotional support. For example, in seeking informational and emotional support, the conditional probabilities $P(S_{G2}|R_{G1})$ are 0.158 and 0.177, respectively. As disease duration increases, the probability of stakeholders providing support shows a stepwise upward trend. Those in the 6 - 10 year group exhibit the highest willingness to offer support, with probabilities of 0.353 for informational support and 0.368 for emotional support. Among stakeholders with ≥11 years of disease, the probability of providing informational support in response to requests from newly diagnosed users (0 - 2 y) $P(R_{G1}|S_{G4})$ was 0.409, significantly higher than in other groups (P<.001).

Social support interactions exhibited both reciprocity and homophilic mixing based on disease duration. In the informational support dimension, the conditional probability of stakeholders with 6 - 10 years of disease duration seeking support when commenting on posts by long-duration users (≥11 y) $P(R_{G4}|S_{G3})$ was 0.287. In contrast, emotional support provision and seeking were largely concentrated within the same disease duration groups. For instance, individuals with 3 - 5 years of disease were more likely to interact with peers at the same stage, reflecting a homophilic mixing effect driven by shared illness experiences.

XSL•FO
RenderX

Figure 5. Conditional probability distributions of seeking and providing informational and emotional support across varying disease duration stages.

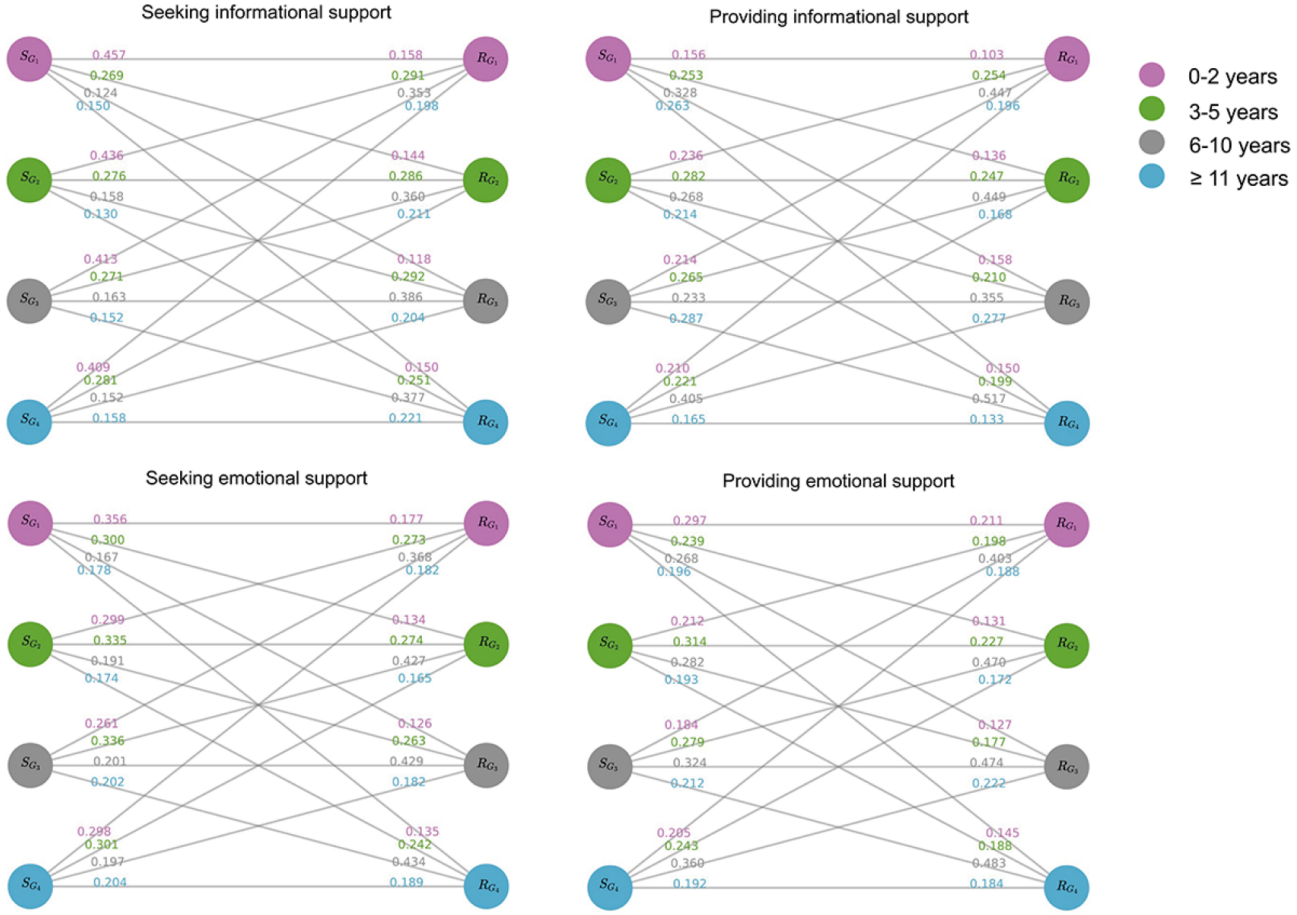

Regression Analysis

To further identify which T1D stakeholders exert greater influence within the network and to explore the mechanisms behind it, we developed a behavior-influence analytical framework. This framework links node attributes to network centrality, to examine how high-influence nodes emerge. We estimated a GLM with a gamma distribution and log link, appropriate for the positively skewed, nonnegative distribution of composite relative centrality, which served as the dependent variable. Predictor variables included user demographics, emotional tendencies, posting orientation, cyclical activity patterns, and disease duration. Figure 6 presents exponentiated regression coefficients (exp(β)), interpreted as multiplicative effects on relative centrality. Bars represent estimated coefficients, boxes show SEs, and whiskers denote 95% CIs. Dark gray indicates statistically significant effects.

XSL•FO
RenderX

Figure 6. Factors associated with user influence, measured by Relative Centrality, in a social support network. Exponentiated coefficients (exp[β]) from a Gamma-generalized linear model (log link) are presented, with whiskers denoting 95% CIs. **P*<.05; ***P*<.01; ****P*<.001.

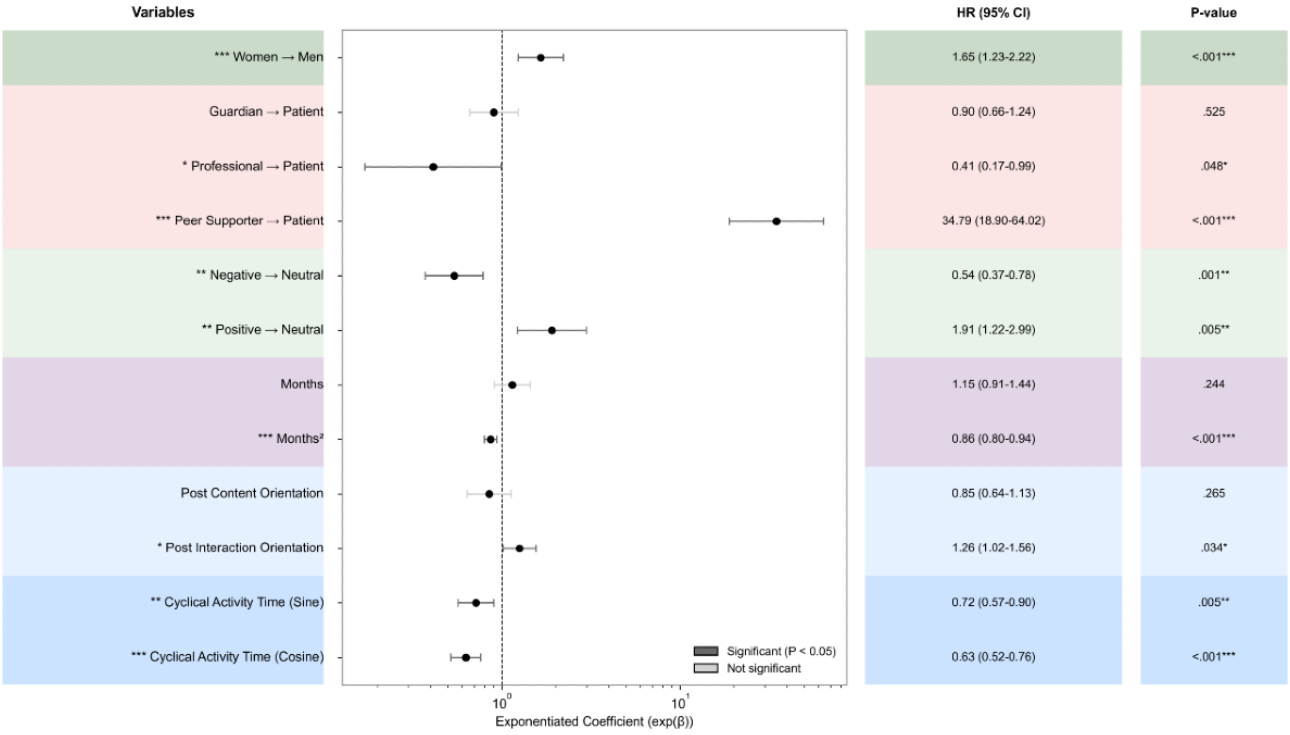

The results show that both gender and community identity significantly affect user influence within the network. Compared with men, women users demonstrated significantly higher influence (exp(β)=1.65; *P*=.001), indicating approximately 65% greater influence. Among identities, peer supporters exhibited the highest influence, with a relative centrality approximately 34.79 times that of the reference group (patients; exp(β)=34.79; *P*<.001). In contrast, professionals exhibited significantly lower influence, approximately 41% that of patients (exp(β)=0.41; *P*=.055), whereas the influence of guardians did not differ significantly from that of patients (*P*=.53).

Regarding posting orientation, PCO had no significant effect on relative centrality (*P*=.27), indicating that whether a post conveyed informational or emotional content did not affect a user's network influence. In contrast, PIO was positively associated with centrality (exp(β)=1.26; *P*=.03), suggesting that users who posted with a stronger intent to provide support had, on average, 26% higher influence. This implies that consistently offering support contributes to increased social influence within the community.

In terms of emotional orientation, users expressing more positive emotions exhibited significantly higher centrality, with an increase of 91% compared with those with neutral sentiment (exp(β)=1.91; *P*=.005). Conversely, negative emotional orientation significantly reduced centrality, lowering it to 54% of the neutral group (exp(β)=0.54, *P*=.001). This indicates that positive affect enhances user influence within the T1D community, suggesting that expressions of positivity help elevate one's position in the support network.

Additionally, the time of user activity showed a significant association with network influence. Using cyclical activity time (sine/cosine) variables to capture diurnal patterns, we found that both the cosine (exp(β)=0.63; *P*<.001) and sine (exp(β)=0.72; *P*=.005) terms had significant negative effects on relative centrality. This indicates that users active during off-peak hours, such as late at night or early in the morning, tend to have lower influence within the network. To further capture the cyclical effects of posting time, we modeled users' average active hours based on the estimated coefficients of the sine and cosine terms. The resulting curve (Figure 7) reveals a peak in relative centrality between 2 PM and 3 PM, suggesting that users active during this period tend to hold greater influence in the social support network.

XSL•FO
RenderX

Figure 7. Predicted user influence (Relative Centrality) based on average daily activity time (0 - 24 h) among users in China's largest type 1 diabetes online health community. The curve displays the marginal effect derived from cyclical time-of-day terms (sine and cosine) in a Gamma-generalized linear model (log link).

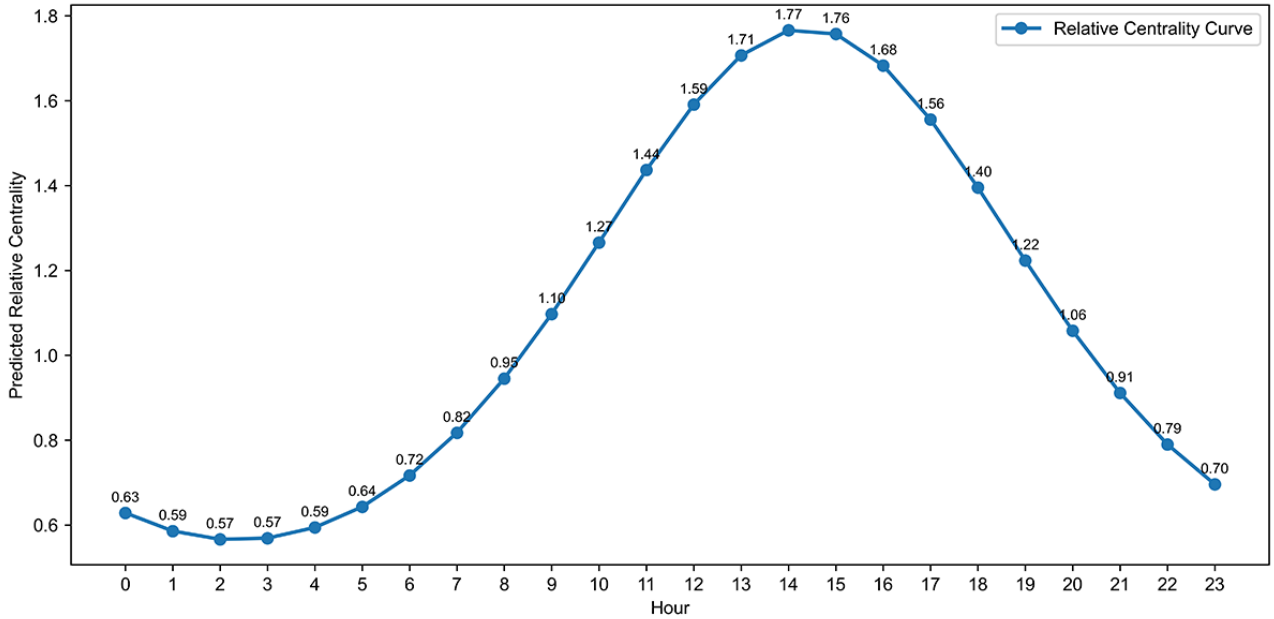

In the analysis of disease duration, the quadratic term of months since diagnosis (months^2) exhibited a significant negative effect on relative centrality (exp(β)=0.87; *P*=.001), while the linear term was not significant (*P*=.24). This suggests an inverted U-shaped relationship between disease duration and user influence, indicating that relative centrality increases in the early stages, peaks, and then declines as disease duration accumulates (Figure 8). Based on the regression coefficients, the peak of relative centrality in the social support network occurs at the 116th month (approximately 9 - 10 y) after T1D diagnosis, with a 95% CI of 43.25-188.91, supporting the robustness of the inverted U-shaped pattern [79]. Additionally, the peak influence of patients and peer supporters was observed at 127 months (10 - 11 y) and 173 months (14 - 15 y), respectively.

Figure 8. Nonlinear association between disease duration (months since diagnosis or years of service for professionals) and user influence (Relative Centrality), derived from a retrospective observational study of China's largest type 1 diabetes online health community. The inverted U-shaped marginal effect is calculated from linear and quadratic disease-duration terms in a Gamma-generalized linear model (log link).

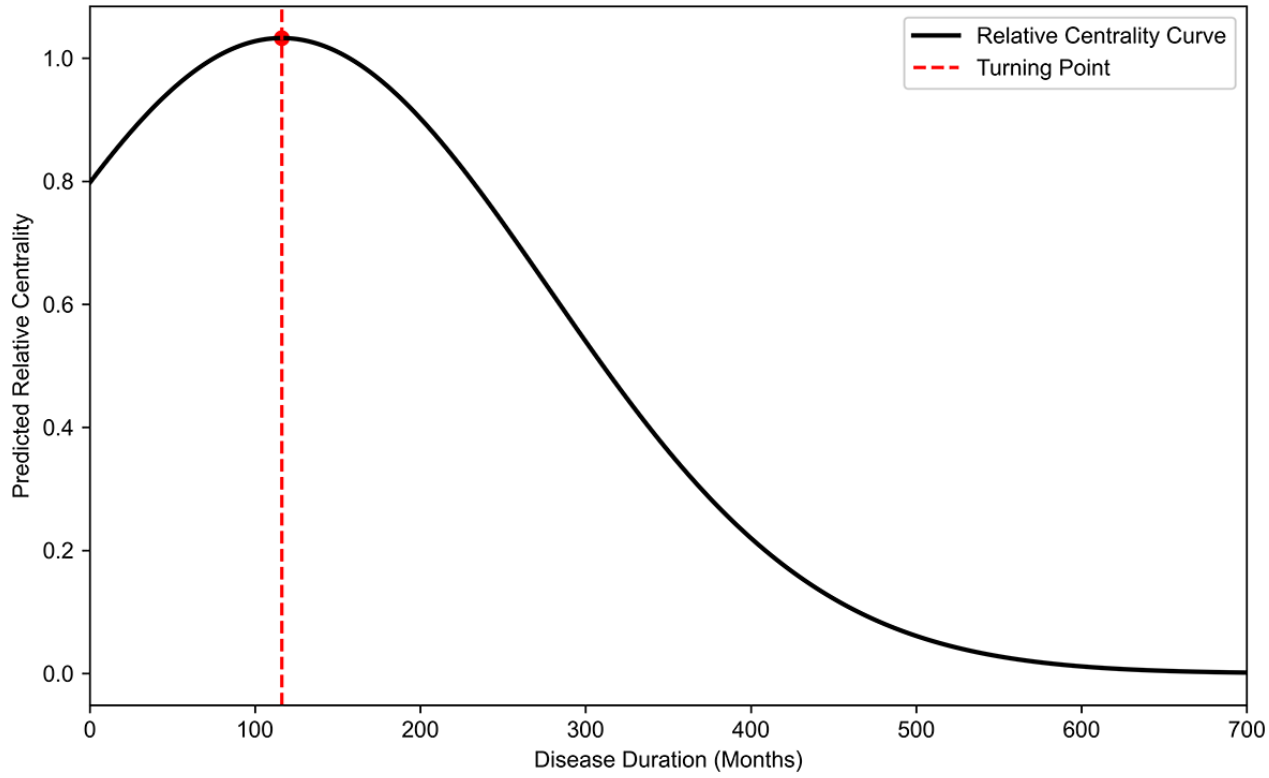

XSL•FO
RenderX

Model Validation and Robustness Check

The model achieved a Cox-Snell pseudo-R^2 of 0.214, indicating a moderate level of model fit. Model diagnostics based on deviance residuals suggested no pronounced systematic deviations. To test the robustness of the observed nonlinear effect, a bootstrap resampling procedure (1200 iterations) was conducted using bias-corrected confidence intervals. The quadratic term for disease duration (months^2) remained significant (β=–.286; *P*=.04), with a 95% CI of –0.545 to −0.013, which did not cross zero. This supports the statistical robustness of the inverted U-shaped relationship between disease duration and relative centrality, independent of normality and homoscedasticity assumptions. In addition, gender, stakeholder identity, and emotional tone also showed significant effects (*P*<.05).

Discussion

Principal Findings

This study analyzed the digital footprints from the largest online T1D community in China and applied SNA to construct social support networks involving diverse stakeholders, identifying key influential users and the mechanisms underlying their influence. The findings show that, as disease duration increases, users tend to shift from seeking support to providing support, peer supporters are the most active providers, primarily offering informational support, whereas patients and guardians more frequently seek support and engage in emotional support interactions. Both informational and emotional support subnetworks exhibit multicentric, star-like structures, with the informational subnetwork displaying broader connectivity and greater cross–disease-stage reciprocity, while the emotional subnetwork is smaller, denser, and characterized by homophilous mixing within similar disease stages. Peer supporters occupy the most central network positions, followed by highly active women patients. User influence is significantly associated with positive emotional expression, posting during afternoon time windows, support-oriented content, and disease duration, following a pronounced inverted U-shaped lifecycle trajectory that peaks at approximately 116 months after diagnosis.

Among diverse stakeholders, a clear division of roles emerges within the social support network. Peer supporters function as the most active providers, predominantly offering informational support, whereas patients and guardians primarily act as support seekers and tend to contribute emotional support. Professionals play a complementary role by participating in both informational and emotional support, providing expert guidance alongside peer-based exchanges. Together, these patterns suggest a collaborative and role-differentiated support ecology within the T1D OHC.

From a disease duration perspective, users in the early stages of diagnosis predominantly seek support and gradually transition toward providing support. This pattern aligns with social epidemiological perspectives that emphasize resilience in long-term diabetes management, defined as the capacity to adapt and function positively under chronic stress or prolonged trauma [80]. Existing research indicates that among chronic disease populations such as patients with cardiovascular problems, emotional or self-esteem support is the most significant dimension of social support in predicting quality of life. Its core function lies in enhancing patients' sense of being valued, cared for, and their psychological coping abilities [81]. In OHCs, informational and emotional support are vital resources for coping with uncertainty and are considered key factors in promoting resilience [82]. As disease experience accumulates, users gradually transition from "seekers" to "providers," reflecting increasing confidence and competence in disease self-management. This shift not only improves their quality of life [83], but also significantly enhances health literacy and attitudes [84], thereby reinforcing both individual resilience and the collective capacity for collaborative coping within the community.

At the network structure level, the T1D social support network exhibits a complex, multicentered collaborative pattern with cross-disease duration interactions, reflecting a low-centralization topology characteristic of complex networks. Both the subnetworks display star-like topologies with high overlap in node composition, yet significant structural differences exist between them. The informational support subnetwork has a wider coverage and lower density, which facilitates broader user engagement and multisource information diffusion. In contrast, the emotional support subnetwork is more tightly connected, fostering more frequent connections with stable node relationships. These 2 subnetworks jointly reinforce each other, contributing to the connectivity and stability of the online social support network system.

In terms of interaction patterns, informational support follows a reciprocal mechanism across disease durations, whereas emotional support tends to exhibit a homophilic mixing pattern across similar disease durations. Consistent with previous research, the informational support subnetwork is more "instrumental," with connections spanning multiple duration groups and facilitating knowledge transfer, thereby helping to improve health outcomes and reduce mortality risk [85,86]. The broader, cross-stage pattern of informational exchange reflects a reciprocity mechanism [87], in which users engage in mutual support exchanges across disease stages to seek and transmit practical self-management knowledge. By contrast, emotional support is more "expressive," relying on strong emotional bonds that enhance users' sense of community and long-term engagement [86]. The stronger within-stage clustering observed in emotional support exchanges aligns with network homophily, defined as the tendency for individuals to connect with similar others [23]. Moreover, some high-influence stakeholders may gradually evolve into resilience mediators, facilitating emotional adaptation and peer support [82]. Through these complementary informational and emotional exchanges across different disease durations, diverse stakeholders develop a functional division of labor that supports both information flow and emotional connectivity within the system, thereby laying the foundation for the structural resilience and sustainable development of the T1D online social support system.

This study develops a social behavior—node influence framework by systematically linking social support behaviors to network structures and examining how individual attributes

XSL•FO
RenderX

and behavioral traits shape relative centrality. In the T1D online social support system, peer supporters serve as the most influential actors while active patients tend to occupy core positions in the network, with influence significantly higher than that of professionals. This highlights the value of the peer "demonstration effect" in chronic disease management, as their continuous participation broadens the scope and depth of information sharing [88]. Compared with professional knowledge, experiential knowledge shared by patients rooted in lived experience and long-term observation of disease progression [26], tends to be more practical and adaptable in managing daily health behaviors and real-life coping. Although professionals, as authoritative information sources guiding health-related discussions and decision-making [89], their structural influence remains relatively limited, likely due to lower participation frequency and narrower interaction scope. This constraint reflects the platform's operational logic and community culture, which emphasize horizontal, peer-to-peer interactions over hierarchical communication, positioning professionals primarily as supplementary knowledge providers. Guardians occupy a distinct role in family-based care and risk prevention. While their influence does not significantly differ from that of patients, the elevated diabetes risk among patients' relatives [90,91], highlights the importance of further identifying guardians' needs and intervention points. This aligns with the current trend of integrating family caregivers into collaborative diabetes management and peer mentoring programs [92]. Furthermore, women demonstrate significantly higher influence than men, possibly due to advantages in metabolic control, emotional expressiveness, and disease self-management [93], thereby facilitating their dominant role in the provision of supportive information.

Beyond individual attributes, influence formation is jointly shaped by microlevel behavioral variables. First, T1D stakeholders who express positive emotions typically exhibit higher relative centrality consistent with evidence that emotional resilience facilitates flexible regulation and engagement in social support interactions [94]. Encouraging positive emotional expression can be an effective strategy to strengthen stakeholder engagement and self-management capabilities. Furthermore, user activity timing also influences their network influence; those active between 2 PM and 3 PM demonstrate markedly higher centrality, reflecting a time-window effect on information visibility and attention clustering [71,72]. Finally, the PCO of posts by T1D stakeholders shows no significant association with their relative centrality, while PIO of posts significantly predicts network position. This aligns with existing research indicating that providing social support enhances health literacy and attitudes more effectively than receiving it [84]. Identifying directional interaction patterns of support providers can help assess their leadership roles in the network, thereby pinpointing credible social influencers [95].

This study further reveals the dynamic trajectory of influence within the T1D online social support system as it evolves over disease duration. The findings indicate that users' relative centrality follows an inverted U-shaped trajectory correlated with disease duration, peaking at approximately 116 months post diagnosis (10 y), with distinct peaks for patients (127 months) and peer supporters (173 months). In the early stages after diagnosis, as T1D stakeholders gain experience in managing the disease, their influence gradually increases to a peak. However, as the disease progresses, the role of nodes in information brokerage may gradually diminish [57,96]; hyperconnected nodes often become less efficient information conduits due to brokerage overload effects [97]. At the behavioral level, users with longer disease durations often exhibit more established self-management routines and a reduced reliance on community-based informational and emotional support. This may lead to decreased participation frequency and, in some cases, signs of social withdrawal. Moreover, the individualized and long-term experiences they share may not align with the immediate needs of newly diagnosed patients, reducing the contextual relevance and dissemination value of their content. This inverted U-shaped trajectory highlights the dynamic nature of user influence over time and underscores the need for digital health platforms to tailor engagement strategies and knowledge structures according to users' lifecycle stages.

The findings highlight that identifying and empowering potential social influencers not only enhances community engagement and improves the disease management abilities of other users [98], but also fosters health literacy and behavior change through guided interactions and shared experiences [99]. Based on these findings, it is recommended to further leverage the "demonstration effect" of long-term users within the T1D social network. This would enhance patients' recognition of the long-term benefits of disease control and strengthen their self-efficacy and autonomy in glycemic management. To operationalize this, the study proposes an innovative patient-centered collaborative care model, led by peer supporters and supported by a multidisciplinary team. The aim is to activate experiential knowledge within the informal caregiving system, optimize information flow, and strengthen emotional connections. In addition, based on the inverted U-shaped lifecycle characteristics of influencers in OHC, it is recommended to establish a mechanism for updating opinion leaders within online communities. This should account for the influence of support behaviors, emotional tendencies, and active time periods on the distribution of influence, thereby continuously refining content design and algorithm-driven content delivery strategies on the platform.

Limitations

This study has several limitations. First, although the age of onset for T1D is an important factor influencing adult excess mortality [100], the inconsistency and unreliability of age information in user profiles, particularly when entered by caregivers, hinder the accurate reflection of the patient's age at diagnosis. Additionally, Wu et al [19] found no significant correlation between age at diagnosis and relative centrality, so this variable was excluded from the analysis due to data availability and theoretical relevance. Second, as T1D is an autoimmune disease with heterogeneous characteristics and onset mechanisms [101], this study analyzes disease duration from a general perspective and does not fully capture the individualized disease profiles or specific social support actions of different patients. The impact of social support on health

XSL•FO
RenderX

outcomes was also not thoroughly assessed. Future research could integrate longitudinal data and clinical health indicators to explore the role of social support in T1D management more comprehensively. Finally, this study relied solely on text-based data and, due to privacy concerns, did not include multimodal data such as blood glucose monitoring images or videos [48]. Future studies could enhance the robustness and adaptability of classification systems by incorporating diverse data sources and, in conjunction with clinical management data, further investigate the coupling mechanisms between social support network structures and health outcomes.

Conclusion

In summary, this study elucidates influence formation and social support mechanisms in a large Chinese T1D OHC by integrating stakeholder diversity, disease duration, and network structure. Influence is jointly associated with stakeholder roles, sentiment orientation, activity timing, and supportive interaction patterns. Informational support interactions demonstrate broader participation and stronger cross–disease-stage reciprocity, suggesting a bridging function for knowledge diffusion, whereas emotional support networks are denser and more homophilous, concentrating exchanges among users with similar disease durations. Peer supporters occupy the most central positions, highlighting the critical value of experiential knowledge for maintaining community-based self-management support. Moreover, user influence follows an inverted U-shaped trajectory over the disease duration, implying a lifecycle prominence that platforms may leverage for timely identification and renewal of community facilitators.

This work uses large-scale, real-world digital trace data and integrates disease duration with a multistakeholder perspective, using a reproducible analytic framework that links social support behaviors, network structure, and user influence. These findings may inform the identification and engagement of high-influence peer supporters and the design of disease-stage–tailored community strategies, with professionals involved in targeted complementary roles. Future studies should extend this work with longitudinal and cross-platform data, incorporate comment-level support and content-quality assessment, and link network-based measures to clinical and psychosocial outcomes to clarify how online support relates to health-relevant end points.

Acknowledgments

We gratefully acknowledge all the volunteers for their contributions to data annotation, and we thank the external experts for their valuable feedback and technical support.

Funding

This study is part of the China Type 1 Diabetes Registration and Management Project under the Science and Technology Innovation 2030 initiative. It was supported by the “Noncommunicable Chronic Diseases-National Science and Technology Major Project” (2023ZD0509100) and National Natural Science Foundation of China (No.82470870).

Data Availability

The datasets generated and/or analyzed during the current study are not publicly available due to privacy protection requirements for individuals with type 1 diabetes but are available from the corresponding author on reasonable request.

Authors' Contributions

Y Zhu was responsible for developing the study design, conducting data analysis, and interpretation. Y Zhang contributed to the conceptualization and design of the study, and provided overall academic guidance. CZ contributed to the early-stage conceptualization and development of the coding scheme, assisted in manuscript review, and was primarily responsible for English language editing and translation. X Zeng provided the original dataset used in this study. X Zheng, the corresponding author and a clinical expert in type 1 diabetes, offered critical clinical perspectives that informed the study design and contributed to the manuscript’s revision from initial drafting through to final approval. All authors have read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Coding scheme.
[DOCX File, 16 KB - jmir_v28i1e82996_app1.docx]

Multimedia Appendix 2
SentiScore formula and Bayesian conditional probability calculation method.
[PDF File, 128 KB - jmir_v28i1e82996_app2.pdf]

XSL•FO
RenderX

References

1. Li Y, Qian K, Wu D, et al. Incidence of childhood type 1 diabetes in Beijing during 2011-2020 and predicted incidence for 2025-2035: a multicenter, hospitalization-based study. Diabetes Ther 2023 Mar;14(3):519-529. [doi: 10.1007/s13300-023-01367-8] [Medline: 36701106]
2. Constantino MI, Molyneaux L, Limacher-Gisler F, et al. Long-term complications and mortality in young-onset diabetes: type 2 diabetes is more hazardous and lethal than type 1 diabetes. Diabetes Care 2013 Dec;36(12):3863-3869. [doi: 10.2337/dc12-2455] [Medline: 23846814]
3. Bjerg L, Hulman A, Carstensen B, Charles M, Witte DR, Jørgensen ME. Effect of duration and burden of microvascular complications on mortality rate in type 1 diabetes: an observational clinical cohort study. Diabetologia 2019 Apr;62(4):633-643. [doi: 10.1007/s00125-019-4812-6] [Medline: 30649599]
4. Zhou Z. Guidelines for the diagnosis and treatment of type 1 diabetes mellitus in China. Chin J Diabetes Mellit 2022;14(11):1143-1250. [doi: 10.3760/cma.j.cn115791-20220916-00474]
5. The diabetes atlas. International Diabetes Federation. URL: https://diabetesatlas.org [accessed 2025-09-14]
6. Weng J. The epidemic study and burden of type 1 diabetes in China. Sci Sin-Vitae 2018 Aug 1;48(8):834-839. [doi: 10.1360/N052018-00016]
7. van Belle TL, Coppieters KT, von Herrath MG. Type 1 diabetes: etiology, immunology, and therapeutic strategies. Physiol Rev 2011 Jan;91(1):79-118. [doi: 10.1152/physrev.00003.2010] [Medline: 21248163]
8. Liu X, Sun M, Li J. Research on gender differences in online health communities. Int J Med Inform 2018 Mar;111:172-181. [doi: 10.1016/j.ijmedinf.2017.12.019] [Medline: 29425630]
9. Min J, Chen Y, Wang L, He T, Tang S. Diabetes self-management in online health communities: an information exchange perspective. BMC Med Inform Decis Mak 2021 Jun 28;21(1):201. [doi: 10.1186/s12911-021-01561-3] [Medline: 34182977]
10. Cobb S. Social support as a moderator of life stress. Psychosom Med 1976 Sep;38(5):300-314. [doi: 10.1097/00006842-197609000-00003]
11. House JS, Umberson D, Landis KR. Structures and processes of social support. Annu Rev Sociol 1988 Aug;14(1):293-318. [doi: 10.1146/annurev.so.14.080188.001453]
12. Berkman LF, Glass T, Brissette I, Seeman TE. From social integration to health: Durkheim in the new millennium. Soc Sci Med 2000 Sep;51(6):843-857. [doi: 10.1016/s0277-9536(00)00065-4] [Medline: 10972429]
13. Yan L, Tan Y. Feeling blue? Go online: an empirical study of social support among patients. Information Systems Research 2014 Dec;25(4):690-709. [doi: 10.1287/isre.2014.0538]
14. Lin X, Zhang D, Li Y. Delineating the dimensions of social support on social networking sites and their effects: a comparative model. Comput Human Behav 2016 May;58:421-430. [doi: 10.1016/j.chb.2016.01.017]
15. Rodgers S, Chen Q. Internet community group participation: psychosocial benefits for women with breast cancer. J Comput-Mediat Commun 2005 Jul;10(4):00-00. [doi: 10.1111/j.1083-6101.2005.tb00268.x]
16. Chung JE. Social networking in online support groups for health: how online social networking benefits patients. J Health Commun 2014;19(6):639-659. [doi: 10.1080/10810730.2012.757396] [Medline: 23557148]
17. Liu X, Jiang S, Sun M, Chi X. Examining patterns of information exchange and social support in a web-based health community: exponential random graph models. J Med Internet Res 2020 Sep 29;22(9):e18062. [doi: 10.2196/18062] [Medline: 32990628]
18. Pham S, Churruca K, Ellis LA, Braithwaite J. Help-seeking, support, and engagement in gestational diabetes mellitus online communities on Facebook: content analysis. JMIR Form Res 2024 Feb 26;8(1):e49494. [doi: 10.2196/49494] [Medline: 38407949]
19. Wu N, Wang SJ, Brazeau AS, et al. Supporting and incentivizing peer leaders for an internet-based private peer community for youths with type 1 diabetes: social network and directed content analysis. J Med Internet Res 2023 Dec 12;25(1):e48267. [doi: 10.2196/48267] [Medline: 38085568]
20. Thoits PA. Conceptual, methodological, and theoretical problems in studying social support as a buffer against life stress. J Health Soc Behav 1982 Jun;23(2):145-159. [doi: 10.2307/2136511] [Medline: 7108180]
21. Bouchard EG, Prince MA, McCarty C, et al. Understanding social network support, composition, and structure among cancer caregivers. Psychooncology 2023 Mar;32(3):408-417. [doi: 10.1002/pon.6087] [Medline: 36588195]
22. Sun W, Guo X, Kwok RCW. Emotion or information? Differences and relationships between two types of professional support in online health communities. Information & Management 2024 Apr;61(3):103944. [doi: 10.1016/j.im.2024.103944]
23. Rueger J, Dolfsma W, Aalbers R. Perception of peer advice in online health communities: access to lay expertise. Soc Sci Med 2021 May;277:113117. [doi: 10.1016/j.socscimed.2020.113117] [Medline: 33865092]
24. Zhang J, Centola D. Social networks and health: new developments in diffusion, online and offline. Annu Rev Sociol 2019 Jul 30;45(1):91-109. [doi: 10.1146/annurev-soc-073117-041421]
25. Gage-Bouchard EA, LaValley S, Mollica M, Beaupin LK. Communication and exchange of specialized health-related support among people with experiential similarity on Facebook. Health Commun 2017 Oct;32(10):1233-1240. [doi: 10.1080/10410236.2016.1196518] [Medline: 27485860]

XSL•FO
RenderX

26. Akrich M, Rabeharisoa V. On the multiplicity of lay expertise: an empirical and analytical overview of patient associations' achievements and challenges. In: Eyal G, Medvetz T, editors. The Oxford Handbook of Expertise and Democratic Politics: Oxford University Press; 2023. [doi: 10.1093/oxfordhb/9780190848927.013.5]
27. Ginossar T. Online participation: a content analysis of differences in utilization of two online cancer communities by men and women, patients and family members. Health Commun 2008;23(1):1-12. [doi: 10.1080/10410230701697100] [Medline: 18443988]
28. Kimbrough AM, Guadagno RE, Muscanell NL, Dill J. Gender differences in mediated communication: women connect more than do men. Comput Human Behav 2013 May;29(3):896-900. [doi: 10.1016/j.chb.2012.12.005]
29. Zhang Y, Dang Y, Chen H. Research note: examining gender emotional differences in web forum communication. Decis Support Syst 2013 Jun;55(3):851-860. [doi: 10.1016/j.dss.2013.04.003]
30. Fuhrer R, Stansfeld SA. How gender affects patterns of social relations and their impact on health: a comparison of one or multiple sources of support from "close persons". Soc Sci Med 2002 Mar;54(5):811-825. [doi: 10.1016/S0277-9536(01)00111-3]
31. Greenbaum CJ. A key to T1D prevention: screening and monitoring relatives as part of clinical care. Diabetes 2021 May;70(5):1029-1037. [doi: 10.2337/db20-1112] [Medline: 33931405]
32. Housni A, Katz A, Kichler JC, Nakhla M, Brazeau AS. 624-P: portrayal of perceived stigma across ages in type 1 diabetes—A BETTER registry analysis. Diabetes 2023 Jun 20;72(Supplement_1). [doi: 10.2337/db23-624-P]
33. Wang YC, Kraut R, Levine JM. To stay or leave? the relationship of emotional and informational support to commitment in online health support groups. 2012 Presented at: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work; Feb 11-15, 2012; Seattle, Washington, USA p. 833-842. [doi: 10.1145/2145204.2145329]
34. Wang YC, Kraut RE, Levine JM. Eliciting and receiving online support: using computer-aided content analysis to examine the dynamics of online social support. J Med Internet Res 2015 Apr 20;17(4):e99. [doi: 10.2196/jmir.3558] [Medline: 25896033]
35. Da Moura Semedo C, Bath PA, Zhang Z. Social support in a diabetes online community: mixed methods content analysis. JMIR Diabetes 2023 Jan 6;8(1):e41320. [doi: 10.2196/41320] [Medline: 36607714]
36. Wang X, Zhao K, Street N. Analyzing and predicting user participations in online health communities: a social support perspective. J Med Internet Res 2017 Apr 24;19(4):e130. [doi: 10.2196/jmir.6834] [Medline: 28438725]
37. Hansen AH, Bradway M, Broz J, et al. Inequalities in the use of eHealth between socioeconomic groups among patients with type 1 and type 2 diabetes: cross-sectional study. J Med Internet Res 2019 May 29;21(5):e13615. [doi: 10.2196/13615] [Medline: 31144669]
38. ElSayed NA, Aleppo G, Aroda VR, et al. 5. Facilitating positive health behaviors and well-being to improve health outcomes: standards of care in diabetes—2023. Diabetes Care 2023 Jan 1;46(Supple 1):S68-S96. [Medline: 36507648]
39. Holt-Lunstad J. Social connection as a public health issue: the evidence and a systemic framework for prioritizing the "Social" in social determinants of health. Annu Rev Public Health 2022 Apr 5;43:193-213. [doi: 10.1146/annurev-publhealth-052020-110732] [Medline: 35021021]
40. Pérez-Aldana CA, Lewinski AA, Johnson CM, Vorderstrasse AA, Myneni S. Exchanges in a virtual environment for diabetes self-management education and support: social network analysis. JMIR Diabetes 2021 Jan 25;6(1):e21611. [doi: 10.2196/21611] [Medline: 33492236]
41. Brinkhues S, Dukers-Muijrers N, Hoebe C, et al. Socially isolated individuals are more prone to have newly diagnosed and prevalent type 2 diabetes mellitus - the Maastricht study. BMC Public Health 2017 Dec 19;17(1):955. [doi: 10.1186/s12889-017-4948-6] [Medline: 29254485]
42. Dias A, Chomutare T, Botsis T. Exploring the community structure of a diabetes forum. In: Quality of Life through Quality of Information: IOS Press; 2012:833-837. [doi: 10.3233/978-1-61499-101-4-833]
43. Ostovari M, Steele-Morris CJ, Griffin PM, Yu D. Data-driven modeling of diabetes care teams using social network analysis. J Am Med Inform Assoc 2019 Oct 1;26(10):911-919. [doi: 10.1093/jamia/ocz022] [Medline: 31045227]
44. Katz E. The two-step flow of communication: an up-to-date report on an hypothesis. Public Opin Q 1957;21(1):61. [doi: 10.1086/266687]
45. Aral S, Walker D. Identifying influential and susceptible members of social networks. Science 2012 Jul 20;337(6092):337-341. [doi: 10.1126/science.1215842] [Medline: 22722253]
46. Himelboim I, Golan GJ. A social network approach to social media influencers on Instagram: the strength of being a nano-influencer in cause communities. Journal of Interactive Advertising 2023 Jan 2;23(1):1-13. [doi: 10.1080/15252019.2022.2139653]
47. Tang X, Yang CC. Ranking user influence in healthcare social media. ACM Trans Intell Syst Technol 2012 Sep;3(4):1-21. [doi: 10.1145/2337542.2337558]
48. Kumar A, Srinivasan K, Cheng WH, Zomaya AY. Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. Inf Process Manag 2020 Jan;57(1):102141. [doi: 10.1016/j.ipm.2019.102141]
49. Cutrona CE, Russell DW. Type of social support and specific stress: toward a theory of optimal matching. In: Social Support: An Interactional View (Wiley Series on Personality Processes): John Wiley & Sons; 1990:319-366.

XSL•FO
RenderX

50. Myrick JG, Holton AE, Himelboim I, Love B. #Stupidcancer: exploring a typology of social support and the role of emotional expression in a social media community. Health Commun 2016;31(5):596-605. [doi: 10.1080/10410236.2014.981664] [Medline: 26453480]
51. Abbasi A, Chen H, Salem A. Sentiment analysis in multiple languages: feature selection for opinion classification in web forums. ACM Trans Inf Syst 2008;26(3):1-34. [doi: 10.1145/1361684.1361685]
52. Tian H, Gao C, Xiao X, et al. SKEP: sentiment knowledge enhanced pre-training for sentiment analysis. arXiv. Preprint posted online on May, 2020. [doi: 10.48550/arXiv.2005.05635]
53. Sun Y, Wang S, Li Y, et al. ERNIE: enhanced representation through knowledge integration. arXiv. Preprint posted online on Apr, 2019. [doi: 10.48550/arXiv.1904.09223]
54. Haibo H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng 2009;21(9):1263-1284. [doi: 10.1109/TKDE.2008.239]
55. Hawe P, Webster C, Shiell A. A glossary of terms for navigating the field of social network analysis. J Epidemiol Community Health 2004 Dec;58(12):971-975. [doi: 10.1136/jech.2003.014530] [Medline: 15547054]
56. Opsahl T, Agneessens F, Skvoretz J. Node centrality in weighted networks: generalizing degree and shortest paths. Soc Networks 2010 Jul;32(3):245-251. [doi: 10.1016/j.socnet.2010.03.006]
57. Burt RS. Structural holes and good ideas. American Journal of Sociology 2004 Sep;110(2):349-399. [doi: 10.1086/421787]
58. Freeman LC. A set of measures of centrality based on betweenness. Sociometry 1977 Mar;40(1):35. [doi: 10.2307/3033543]
59. Borgatti SP. Centrality and network flow. Soc Networks 2005 Jan;27(1):55-71. [doi: 10.1016/j.socnet.2004.11.008]
60. Dubois E, Gaffney D. The multiple facets of influence: identifying political influentials and opinion leaders on Twitter. Am Behav Sci 2014;58(10):1260-1277. [doi: 10.1177/0002764214527088]
61. Riquelme F, Vera JA. A parameterizable influence spread-based centrality measure for influential users detection in social networks. Knowl Based Syst 2022 Dec;257:109922. [doi: 10.1016/j.knosys.2022.109922]
62. Bonacich P. Factoring and weighting approaches to status scores and clique identification. J Math Sociol 1972 Jan;2(1):113-120. [doi: 10.1080/0022250X.1972.9989806]
63. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. ICWSM 2009;3(1):361-362. [doi: 10.1609/icwsm.v3i1.13937]
64. Atkinson MA, Eisenbarth GS, Michels AW. Type 1 diabetes. Lancet 2014 Jan 4;383(9911):69-82. [doi: 10.1016/S0140-6736(13)60591-7] [Medline: 23890997]
65. Mizokami-Stout KR, Li Z, Foster NC, et al. The contemporary prevalence of diabetic neuropathy in type 1 diabetes: findings from the T1D exchange. Diabetes Care 2020 Apr;43(4):806-812. [doi: 10.2337/dc19-1583] [Medline: 32029635]
66. Tomarken AJ, Serlin RC. Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. Psychol Bull 1986;99(1):90-99. [doi: 10.1037//0033-2909.99.1.90]
67. Katona Z, Zubcsek PP, Sarvary M. Network effects and personal influences: the diffusion of an online social network. Journal of Marketing Research 2011 Jun;48(3):425-443. [doi: 10.1509/jmkr.48.3.425]
68. Mallipeddi R, Kumar S, Sriskandarajah C, Zhu Y. A framework for analyzing influencer marketing in social networks: selection and scheduling of influencers. SSRN Journal. 2020 p. 75-104. [doi: 10.2139/ssrn.3255198]
69. Li YM, Shiu YL. A diffusion mechanism for social advertising over microblogs. Decis Support Syst 2012 Dec;54(1):9-22. [doi: 10.1016/j.dss.2012.02.012]
70. Cha M, Haddadi H, Benevenuto F, Gummadi K. Measuring user influence in Twitter: the million follower fallacy. ICWSM 2010;4(1):10-17. [doi: 10.1609/icwsm.v4i1.14033]
71. Villamediana J, Küster I, Vila N. Destination engagement on Facebook: time and seasonality. Annals of Tourism Research 2019 Nov;79:102747. [doi: 10.1016/j.annals.2019.102747]
72. Zhang M, Zhang D, Zhang Y, Yeager K, Fields TN. An exploratory study of Twitter metrics for measuring user influence. J Informetr 2023 Nov;17(4):101454. [doi: 10.1016/j.joi.2023.101454]
73. Bonacich P. Some unique properties of eigenvector centrality. Soc Networks 2007 Oct;29(4):555-564. [doi: 10.1016/j.socnet.2007.04.002]
74. Katz E, Blumler JG, Gurevitch M. Uses and gratifications research. Public Opin Q 1973;37(4):509. [doi: 10.1086/268109]
75. Abdi H, Williams LJ. Principal component analysis. WIREs Computational Stats 2010 Jul;2(4):433-459. [doi: 10.1002/wics.101]
76. Ng VKY, Cribbie RA. Using the gamma generalized linear model for modeling continuous, skewed and heteroscedastic outcomes in psychology. Curr Psychol 2017 Jun;36(2):225-235. [doi: 10.1007/s12144-015-9404-0]
77. Kutner MH, Nachtsheim CJ, Neter J, Li W. Applied Linear Statistical Models: McGraw-Hill; 2005.
78. van de Schoot R, Depaoli S, King R, et al. Bayesian statistics and modelling. Nat Rev Methods Primers 2021;1(1):1-26. [doi: 10.1038/s43586-020-00001-2]
79. Haans RFJ, Pieters C, He ZL. Thinking about U: theorizing and testing U- and inverted U-shaped relationships in strategy research. Strateg Manag J 2016;37(7):1177-1195. [doi: 10.1002/smj.2399]
80. Hilliard ME, Harris MA, Weissberg-Benchell J. Diabetes resilience: a model of risk and protection in type 1 diabetes. Curr Diab Rep 2012 Dec;12(6):739-748. [doi: 10.1007/s11892-012-0314-3] [Medline: 22956459]

XSL•FO
RenderX

81. Chollou KM, Shirzadi S, Pourrazavi S, Babazadeh T, Ranjbaran S. The role of perceived social support on quality of life in people with cardiovascular diseases. Ethiop J Health Sci 2022 Sep;32(5):1019-1026. [doi: 10.4314/ejhs.v32i5.17] [Medline: 36262697]
82. Park I, Sarnikar S, Cho J. Disentangling the effects of efficacy-facilitating informational support on health resilience in online health communities based on phrase-level text analysis. Information & Management 2020 Dec;57(8):103372. [doi: 10.1016/j.im.2020.103372]
83. Rains SA, Wright KB. Support and computer-mediated communication: a state-of-the-art review and agenda for future research. In: Communication Yearbook 40: Routledge; 2016. [doi: 10.4324/9781315627083]
84. Chen L, Baird A, Straub D. Fostering participant health knowledge and attitudes: an econometric study of a chronic disease-focused online health community. Journal of Management Information Systems 2019 Jan 2;36(1):194-229. [doi: 10.1080/07421222.2018.1550547]
85. Thoits PA. Mechanisms linking social ties and support to physical and mental health. J Health Soc Behav 2011 Jun;52(2):145-161. [doi: 10.1177/0022146510395592] [Medline: 21673143]
86. Turner JB, Turner RJ. Social relations, social integration, and social support. In: Handbook of the Sociology of Mental Health Handbooks of Sociology and Social Research, 2nd edition: Springer; 2013:341-356. [doi: 10.1007/978-94-007-4276-5_17]
87. Zhang X, Liu S, Deng Z, Chen X. Knowledge sharing motivations in online health communities: a comparative study of health professionals and normal users. Comput Human Behav 2017 Oct;75:797-810. [doi: 10.1016/j.chb.2017.06.028]
88. Tseng HT, Ibrahim F, Hajli N, Nisar TM, Shabbir H. Effect of privacy concerns and engagement on social support behaviour in online health community platforms. Technol Forecast Soc Change 2022 May;178:121592. [doi: 10.1016/j.techfore.2022.121592]
89. Hunt D, Koteyko N, Gunter B. UK policy on social networking sites and online health: from informed patient to informed consumer? Digit Health 2015;1:2055207615592513. [doi: 10.1177/2055207615592513] [Medline: 29942541]
90. Battaglia M, Anderson MS, Buckner JH, et al. Understanding and preventing type 1 diabetes through the unique working model of TrialNet. Diabetologia 2017 Nov;60(11):2139-2147. [doi: 10.1007/s00125-017-4384-2] [Medline: 28770323]
91. Mathieu C, Lahesmaa R, Bonifacio E, Achenbach P, Tree T. Immunological biomarkers for the development and progression of type 1 diabetes. Diabetologia 2018 Nov;61(11):2252-2258. [doi: 10.1007/s00125-018-4726-8] [Medline: 30209538]
92. Enlow P, Thomas C, Collins DA, et al. 694-P: Engaging family and health care provider stakeholders to refine a Type 1 diabetes peer mentoring program for equitable CGM use. Diabetes 2024 Jun 14;73(Supplement_1). [doi: 10.2337/db24-694-P]
93. Boettcher C, Tittel SR, Meissner T, et al. Sex differences over time for glycemic control, pump use and insulin dose in patients aged 10-40 years with type 1 diabetes: a diabetes registry study. BMJ Open Diabetes Res Care 2021 Dec;9(2):e002494. [doi: 10.1136/bmjdrc-2021-002494] [Medline: 34969693]
94. Bonanno GA. Resilience in the face of potential trauma. Curr Dir Psychol Sci 2005 Jun;14(3):135-138. [doi: 10.1111/j.0963-7214.2005.00347.x]
95. Valente TW, Pumpuang P. Identifying opinion leaders to promote behavior change. Health Educ Behav 2007 Dec;34(6):881-896. [doi: 10.1177/1090198106297855] [Medline: 17602096]
96. Reus B, Moser C, Groenewegen P. Expanding organisational knowledge online: the role of bridging members in knowledge expansion in online groups. Innovation 2022 Jan 2;24(1):144-167. [doi: 10.1080/14479338.2020.1840381]
97. Stewart SA, Abidi SSR. Applying social network analysis to understand the knowledge sharing behaviour of practitioners in a clinical online discussion forum. J Med Internet Res 2012 Dec 4;14(6):e170. [doi: 10.2196/jmir.1982] [Medline: 23211783]
98. Fergie G, Hunt K, Hilton S. Social media as a space for support: young adults' perspectives on producing and consuming user-generated content about diabetes and mental health. Soc Sci Med 2016 Dec;170:46-54. [doi: 10.1016/j.socscimed.2016.10.006] [Medline: 27750067]
99. Tang TS, Yip KW, Moore L, Klein G. 575-P: screening, selecting, and training peers to deliver mental health support via a virtual care platform to adults with type 1 diabetes living in rural and remote communities. Diabetes 2022 Jun 1;71(Supplement_1). [doi: 10.2337/db22-575-P]
100. Rawshani A, Sattar N, Franzén S, et al. Excess mortality and cardiovascular disease in young adults with type 1 diabetes in relation to age at onset: a nationwide, register-based cohort study. Lancet 2018 Aug 11;392(10146):477-486. [doi: 10.1016/S0140-6736(18)31506-X] [Medline: 30129464]
101. Ilonen J, Lempainen J, Veijola R. The heterogeneous pathogenesis of type 1 diabetes mellitus. Nat Rev Endocrinol 2019 Nov;15(11):635-650. [doi: 10.1038/s41574-019-0254-y] [Medline: 31534209]

Abbreviations

GLM: generalized linear model
LLM: large language model
OHC: online health community
PC: principal component

XSL•FO
RenderX

PCO: postcontent orientation
PIO: postinteraction orientation
RQ: research question
SA: Sentiment analysis
SentiScore: sentiment score
SNA: social network analysis
T1D: type 1 diabetes

Edited by A Stone; submitted 27.Aug.2025; peer-reviewed by C Dms, J Wei, S Ranjbaran; revised version received 09.Apr.2026; accepted 10.Apr.2026; published 15.Jun.2026.

Please cite as:
Zhu Y, Zhang Y, Zhao C, Zeng X, Zheng X
Social Support Mechanisms in an Online Type 1 Diabetes Community: Social Network Analysis of Stakeholder Diversity and Disease Duration
J Med Internet Res 2026;28:e82996
URL: https://www.jmir.org/2026/1/e82996
doi:10.2196/82996

© Yujia Zhu, Yanxiang Zhang, Chenxiao Zhao, Xifeng Zeng, Xueying Zheng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Reddit Discussions During the 2022 Mpox Outbreak: Observational Analysis of Sentiment, Topics, and Audience Engagement

Xi Ning Luo[1,2], BSc; Zahra Movahedi Nia[1,3,4,5], PhD; Jude Dzevela Kong[1,3,5,6,7,8], PhD

[1]Artificial Intelligence and Mathematical Modeling Lab, Dalla Lana School of Public Health, University of Toronto, 155 College St, Room 500, Toronto, ON, Canada

[2]Department of Computer Science, Faculty of Arts and Science, University of Toronto, Toronto, ON, Canada

[3]Global South Artificial Intelligence for Pandemic and Epidemic Preparedness and Response Network, Toronto, ON, Canada

[4]Department of Mathematics and Statistics, Faculty of Science, York University, North York, ON, Canada

[5]Africa-Canada Artificial Intelligence and Data Innovation Consortium, Toronto, ON, Canada

[6]Munk School of Global Affairs and Public Policy, University of Toronto, Toronto, ON, Canada

[7]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada

[8]Department of Mathematics, Bahen Centre for Information Technology, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Jude Dzevela Kong, PhD
Artificial Intelligence and Mathematical Modeling Lab, Dalla Lana School of Public Health, University of Toronto, 155 College St, Room 500, Toronto, ON, Canada

Abstract

Background: Public health crises often reshape online discourse by amplifying uncertainty, frustration, stigma, and misinformation, with important implications for risk communication.

Objective: This study examines these dynamics on Reddit (Reddit Inc) during a recent outbreak, using Mpox as a case study.

Methods: We analyzed sentiment, topical themes, and audience engagement in posts and comments drawn from 4 Mpox-related subreddits. Using natural language processing methods, we applied sentiment analysis and latent Dirichlet allocation to classify 1169 posts and 6571 comments (from July 21, 2021, to July 16, 2025) into sentiment categories and 9 distinct topics. Of the 1169 posts, 611 (52.3%) were neutral, 370 (31.6%) were negative, and 188 (16.1%) were positive. Among comments, 2825 of 6571 (43%) were neutral, 1962 (29.9%) were negative, and 1784 (27.1%) were positive. We then used Kruskal-Wallis tests, Dunn post hoc comparisons, and Vargha-Delaney A to assess relationships among sentiment, topic, and engagement metrics.

Results: Engagement differed significantly by sentiment ($P<.001$) and topic ($P<.001$). Negative posts had higher median scores (median 7, IQR 2-27) than positive ones (median 5, IQR 2-16; z score=6.02; adjusted $P<.001$; Vargha-Delaney A=0.55). Posts about systemic public health failures (Topic 4) received lower median scores (median 4, IQR 1.75-14.25) than other topics. Topic 9 accounted for 980 of 6571 (14.9%) comments, dominating discussions regardless of original post topic. Positive posts generated 284 of 922 (30.8%) positive comments, whereas negative posts received 526 of 1615 (32.6%) negative comments. Comments on positive posts had higher sentiment scores (Vargha-Delaney A=0.550), whereas comments on negative posts had lower sentiment scores (Vargha-Delaney A=0.463). Topic-level differences in comment sentiment were also observed: comments responding to posts on scientific- and policy-related debates (Topic 8) were more positive (Vargha-Delaney A=0.531), whereas those on systemic failures (Topic 4) were more negative (Vargha-Delaney A=0.478).

Conclusions: Overall, the findings highlight how audience reactions can amplify emotionally charged narratives and reframe technical information into socially and politically charged debates. These insights can inform public health communication strategies by anticipating likely audience responses, mitigating stigma and misinformation, and fostering constructive dialogue during health crises.

(J Med Internet Res 2026;28:e90152) doi:10.2196/90152

KEYWORDS

audience engagement; monkeypox; virus; infectious outbreak; Reddit; social media; sentiment analysis; topic modelling; natural language processing; public discourse

XSL•FO
RenderX

Introduction

Public health crises often reshape online discourse by amplifying uncertainty, frustration, stigma, and misinformation, with important implications for risk communication. Social media platforms—and Reddit (Reddit Inc) in particular—have become crucial spaces where the public seeks information, shares experiences, and reacts to evolving guidance during outbreaks. In this context, Mpox serves as a timely case study for understanding how online discussions unfold during a public health crisis.

Mpox, formerly known as Monkeypox, is a double-stranded DNA virus in the Orthopoxvirus genus, which contains smallpox, cowpox, and vaccinia. The disease first became endemic in the Democratic Republic of Congo in 1970 and for decades remained a rare zoonotic disease confined to Central and West Africa, mainly affecting children in rural areas. Over time, cases steadily increased, with a notable outbreak in the United States in 2003 linked to imported rodents. By the 2010s, outbreaks were reported in multiple African countries, with a major resurgence in Nigeria in 2017-2018 that also led to exported cases abroad. Since May 2022, however, Mpox has evolved into a global outbreak affecting over 70 countries, driven primarily by human-to-human transmission and presenting with atypical clinical features compared to earlier outbreaks [1].

The rapid global spread of Mpox has not only posed clinical and epidemiological challenges but has also fueled widespread uncertainty, frustration, stigma, and misinformation online. As seen in other public health crises, public discourse around outbreaks is often shaped by polarized sentiments, conspiracy theories, and shifting attitudes toward prevention measures. Social media platforms, in particular, have become crucial spaces for both the dissemination of health information and the circulation of misleading narratives. Past studies have shown how online discussions can capture nuanced public responses during pandemics. For instance, an analysis of Canadian city-based subreddits revealed that local sentiment toward COVID-19 vaccines varied across regions, with joy being the most associated emotion in vaccine-related conversations [2]. Within the context of Mpox, large-scale analyses of X (formerly Twitter; X Corp) and Facebook (Meta Platforms, Inc) posts found that discussions frequently linked the outbreak to lesbian, gay, bisexual, transgender, queer, and more (LGBTQ+) communities, with 8 of 10 identified topics actively stigmatizing sexual minorities [3]. Similarly, thematic analysis of Mpox-related Reddit posts in LGBTQ+ subreddits highlighted changes in sexual behavior, concerns about vaccine safety, and the negative mental health effects of stigma, while also showing how these communities relied on online platforms for information and mutual support [4]. Expanding on these findings, another study analyzed Reddit comments between June 1, 2022, and August 5, 2022, and identified dominant themes such as symptoms, transmission, vaccination, government interventions, and homophobia [5].

Beyond Reddit, deep learning and network-based analyses have shown that misinformation and online toxicity were defining features of Mpox-related discourse. One study mapped Twitter narratives and found that misleading or irrelevant information spread faster and reached wider audiences than official communications [6]. Likewise, another study identified widespread toxic discourse on Twitter centered on homophobia, racism, and politicization, illustrating how stigma and polarization were amplified during the outbreak [7]. Complementary qualitative research also emphasized the dual role of social media as both an effective communication tool and a challenge for public health response. One paper reported that social media platforms in the United Kingdom facilitated rapid information sharing and advocacy during the Mpox outbreak but simultaneously reproduced exclusionary dynamics and misinformation [8]. Similarly, another found that public health departments' Twitter communications promoted vaccination and community engagement but struggled to maintain visibility amid competing misinformation streams [9]. Collectively, these studies underscore both the potential and pitfalls of social media in shaping public understanding and communication during health crises.

While prior studies on Mpox have illuminated the thematic structure and emotional tone of Mpox discussions online, they rarely consider how different forms of audience engagement, such as upvotes, comments, shares, or views, interact with the sentiment and topics of the original content. This is a critical omission, as engagement metrics are not monolithic but reflect distinct audience behaviors. A study on Facebook news posts [10] demonstrated that commenting, sharing, and liking are separate modes of engagement, each driven by different content characteristics and emotional triggers. For example, politics-related posts were most likely to generate comments, while health-related posts were strongly associated with sharing, and positive-tone posts received more likes. Emotional triggers also mattered, with anger increasing commenting and sharing, while sadness encouraged sharing and reacting. However, less is known about how these dynamics manifest on Reddit during public health crises—particularly how engagement relates to the sentiment and topical framing of outbreak-related posts and how audiences respond through comments. Using Mpox as a case study, this work examines how the relationship between posts and audience reactions may shape discourse during a health crisis.

To address this gap, this study investigates two research questions (RQs):

1. How do audience engagement metrics relate to the sentiment and topic categories of social media posts about health emergency crises (Mpox as a case study)?
2. How do audience responses (comment, sentiment, and topical focus) differ for posts of varying sentiments and topic categories (Mpox as a case study)?

By bridging this gap, this study provides a more nuanced understanding of how online discussions evolve from the initial post to audience interaction, offering insights that can help public health communicators design more effective messaging, guide platform moderators in mitigating stigma and misinformation, and support vulnerable communities in accessing accurate information during public health crises.

XSL•FO
RenderX

Methods

Data Collection and Cleaning

The data used for this study were gathered from Reddit. Four subreddits with sufficient size and general discussions related to Mpox were selected: r/Monkeypox, r/monkeypoxpositive, r/worldnews, and r/news [11-14]. The 4 subreddits were selected to capture both specialized and general-audience discourse. r/Monkeypox and r/monkeypoxpositive focus specifically on outbreak-related discussions, whereas r/news and r/worldnews provide broader news-driven conversations and public reactions. This combination allowed examination of both community-specific and general public engagement. Other subreddits (eg, r/PublicHealth and r/Coronavirus) were not included because Mpox-related content appeared sporadically and was less consistently represented during preliminary screening.

This dataset was collected from July 2021 to July 2025 to capture discourse preceding the global outbreak, the peak transmission period in 2022 - 2023, and subsequent post-peak discussions, allowing examination of how sentiment and thematic emphasis evolved across emergency and later phases.

Since r/worldnews and r/news also contain discussions about other topics, relevant posts were filtered using the keywords "mpox" and "monkeypox." These keywords reflect the most commonly used labels during the study period. While related terms (eg, orthopoxvirus or strain-specific names) may have captured additional posts, preliminary searches indicated that such terminology appeared infrequently in general news subreddits.

By using ReddAPI (SeasonedCode; available through RapidAPI, Nokia) [15], posts and their respective comments were retrieved from these 4 subreddits. Key information retrieved for posts included subreddit, permalink, title, description, score (calculated by the number of upvotes minus downvotes), and created time. Key information retrieved for comments included subreddit, permalink, permalink of the parent post, text message, score, and created time.

Data processing and analysis were done using R (version 4.5.1; R Foundation for Statistical Computing) within RStudio Desktop (version 2025.05.1+513; Posit Software) [16,17]. Posts with empty entries in any column were removed, with the exception of description, which is optional on Reddit. Comments containing any empty entries were also excluded. After removing missing values, 1491 posts and 7706 comments remained. Language detection was then performed using the detect_language() function from the Compact Language Detector 3 library, which predicts the most likely language of each text string and returns "Not Applicable" if the language cannot be reliably determined. Language detection was applied to post titles and full comment bodies. English was identified as the dominant language, accounting for 1186 posts and 6802 comments. Only English-language content (language code "en") was retained for analysis.

Sentiment Analysis Methods

A sentiment score was assigned to each post and comment using the sentimentr library [18], which performs sentence-level scoring and accounts for contextual valence shifters such as negators, amplifiers, deamplifiers, and adversative conjunctions (eg, "but"). This approach was selected over simpler bag-of-words lexicons because Reddit posts and comments often contain mixed sentiment within a single message, informal language, and variable length. A general-purpose lexicon-based method was therefore considered appropriate for capturing contextual sentiment shifts in heterogeneous social media text. The sentimentr package computes sentence-level polarity scores using a lexicon-based approach that incorporates valence shifters (eg, negators such as "not," amplifiers such as "very," and adversative conjunctions) to adjust contextual sentiment intensity. Sentence-level scores are aggregated to produce a document-level sentiment score for each post or comment. Higher values indicate more positive sentiment, values near zero reflect neutrality, and lower values indicate more negative sentiment. Based on these scores, each observation was also assigned a categorical sentiment label: scores below −0.1 were classified as "negative," scores above 0.1 as "positive," and scores in between as "neutral." The thresholds of 0.1 and −0.1 were selected to distinguish substantively polarized text from near-neutral language, as very small absolute scores typically represent minimal lexical polarity.

Topic Modeling Methods

Topic modeling was conducted using the topicmodels library [19,20]. The topicmodels package implements probabilistic topic modeling algorithms, including latent Dirichlet allocation (LDA), which models each document as a mixture of topics and each topic as a probability distribution over words. Before modeling, posts (title and description) and comments were combined into a single corpus and preprocessed using standard text normalization procedures, including lowercasing, punctuation removal, number removal, stop-word removal, and whitespace trimming. A document-term matrix was then constructed, and sparse terms were removed to reduce noise. Documents with no remaining terms after preprocessing were excluded. LDA was applied to identify clusters of topics from the text data. For each document, LDA estimates a topic probability distribution (gamma matrix), and each document was assigned to the topic with the highest posterior probability. To determine the optimal number of topics, the k-value (the number of topics) was fine-tuned by testing values between 2 and 15. For each k-value, topic-word distributions (beta matrix) were extracted, and pairwise Jensen-Shannon divergence (JSD) was computed between topics. The mean JSD among the top 10 terms was then calculated, with higher values indicating more distinct topics, and the k-value with the highest mean JSD (k=9) was selected, as shown in Figure 1.

XSL•FO
RenderX

Figure 1. Mean Jaccard similarity among the top 10 terms of each topic as a function of the number of topics (k). For each candidate model (k=2-15), latent Dirichlet allocation (LDA) was fitted, and the average pairwise Jaccard similarity between the top terms of topics was computed. Lower values indicate greater separation between topics.

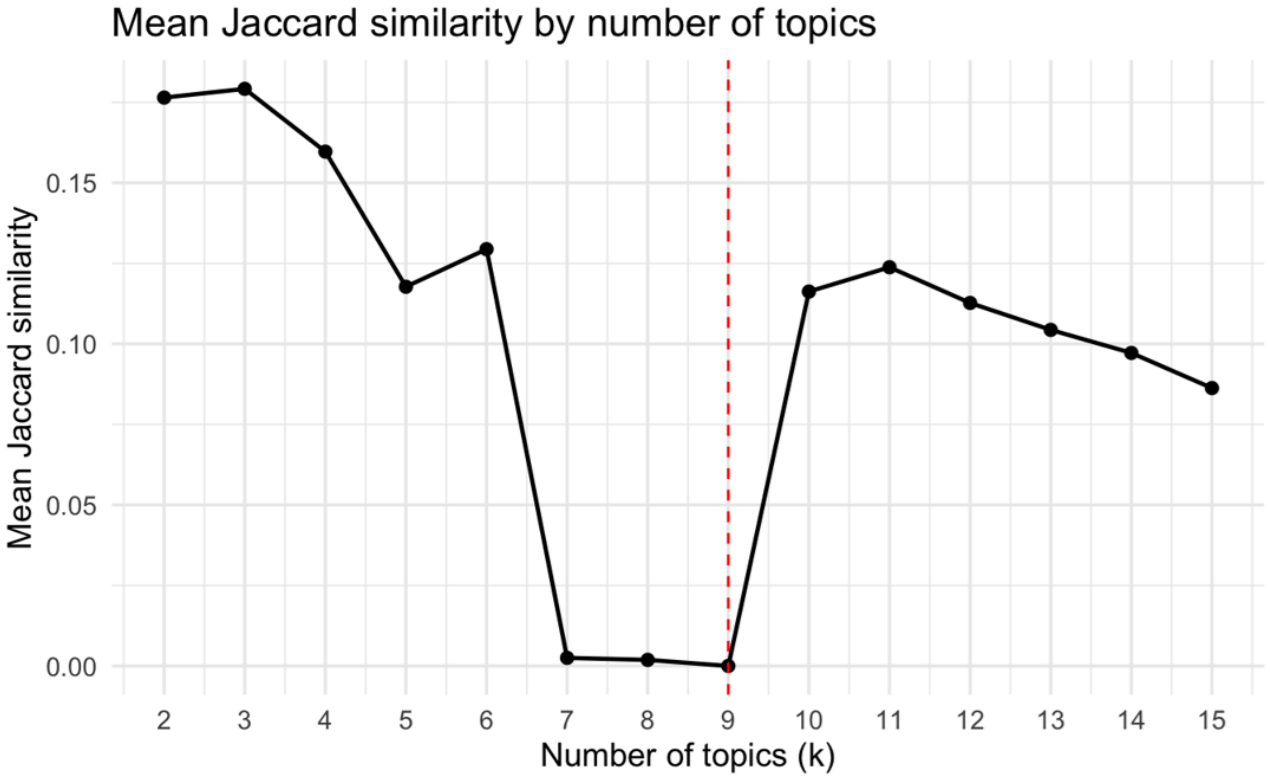

After statistical topic assignments were finalized, a large language model (LLM; Gemini 1.5 Flash; Google) was used solely to generate human-readable topic summaries. The LLM did not influence clustering or document assignment. For each topic, randomly sampled subsets of posts and comments were provided to the model across multiple runs. Across iterations, approximately 575‑669 unique messages per topic were examined, representing the majority of documents assigned to each topic. Generated summaries were reviewed and aggregated to produce final topic descriptions. This process focused on interpretive labeling rather than altering statistical topic assignments. The exact prompt structure used for topic labeling is described below.

The first step was a system-level instruction. The following system message was provided before any topic text was sent:

> *You are helping analyze Reddit discussions of Mpox. I have grouped Reddit posts/comments into statistically distinct topics. I will now send you several chunks of text per topic. Each chunk will be clearly labeled. Do not respond yet — just read and remember the content. At the end, I will ask for a summary of each topic's main theme.*

The second step was topic-specific text input. For each topic, 50 randomly sampled documents (posts and comments combined) were selected. The full text of each sampled document was provided. Each topic was labeled and structured as follows:

> *=== Topic X Chunk 1 ===*
>
> *[Full text of sampled posts and comments belonging to Topic X]*

No summarization or preprocessing beyond earlier text cleaning was performed before submission to the model.

The third step was a first-stage summarization instruction. After all topic-specific text had been provided, the following instruction was issued:

> *I finished sending all the Reddit text. Now summarize the main theme of each topic (1 to 9) in one clear, concise sentence.*
>
> *Format exactly as: Topic 1: [theme]; Topic 2: [theme]; … Topic 9: [theme]*
>
> *Focus on the most distinctive ideas across the chunks for each topic.*

The fourth step was a multi-run stability procedure. To reduce variability in single-model outputs, the above procedure was repeated 20 times using newly sampled documents for each topic. This resulted in 20 candidate summaries per topic.

The final step was the second-stage consolidation prompt. All 20 auto-generated summaries for each topic were then provided back to the model using the following instruction:

> *I am studying discussions related to Mpox. Here are 20 auto-generated theme summaries for Topic X:*
>
> *[Theme 1]*
>
> *[Theme 2]*
>
> *…*
>
> *[Theme 20]*

XSL•FO
RenderX

Now summarize the main theme of each topic (1-9) based on the summaries I sent.

Format your response as:

Topic 1 theme: [theme]; Topic 1 concise summary: [2 - 3 words];

Topic 2 theme: [theme]; Topic 2 concise summary: [2 - 3 words];

…

Each 'theme' should be a complete sentence capturing the topic's overall meaning. Each 'concise summary' should be 2 - 3 keywords.

The final theme labels were generated from this consolidation stage and subsequently reviewed by the research team to ensure conceptual clarity and distinguishability across topics. No manual reassignment of documents was performed. A small number of observations could not be assigned to a topic due to formatting issues, and these were removed from the analysis.

Statistical Analysis Methods

After processing, a total of 1169 posts and 6571 comments were used for the analysis. Both the earliest and latest creation dates for posts and comments are July 21, 2021, and July 16, 2025, respectively. Exploratory data analysis was then carried out using a combination of visualizations and summary tables to identify initial patterns and trends in the data. Subsequent analyses focused on derived attributes of each post or comment, such as sentiment scores, sentiment labels, assigned topic categories representing the main subject of the text, and engagement metrics, rather than on the full textual content itself. To assess differences in score distributions across sentiment labels and topics, the Kruskal-Wallis test was applied. The Dunn test and Kolmogorov-Smirnov statistics were used for pairwise comparisons to more comprehensively evaluate the extent of differences between groups. Nonparametric statistical tests were selected due to the nonnormal distribution and skewness observed in engagement metrics and sentiment scores. The Kruskal-Wallis test was used to compare differences across multiple independent topic groups, as it does not assume normality. When overall group differences were significant, the Dunn test with appropriate adjustment was applied for pairwise comparisons. The Kolmogorov-Smirnov test was used to assess distributional differences between groups, as it is sensitive to differences in both central tendency and distribution shape. All statistical tests were considered significant at $P<.05$, roughly equivalent to 2 standard deviations from the mean [21].

Post Content Divergence Analysis Methods

Pairwise comparisons were further evaluated using the Vargha-Delaney A (stochastic superiority, "A12") statistic, which estimates the probability that a randomly selected score from one group will exceed that from another. This measure was applied to compare scores across sentiment and topic groups, as well as to examine sentiment differences in comments associated with posts of varying sentiment and topic. This approach was chosen due to the skewed, nonnormal distribution of engagement and sentiment scores. The Vargha-Delaney A statistic was calculated as a nonparametric effect size measure, quantifying the probability that values from one group exceed those from another. The threshold for statistical significance was set at 0.05 for all tests [21].

Ethical Considerations

This study adhered to widely accepted ethical standards governing online research and used only publicly accessible data obtained from Reddit. Safeguarding user privacy was a central consideration throughout the data collection process. Information was sourced exclusively from public Reddit communities (subreddits), where posts and discussions are openly visible and are not shared under any expectation of confidentiality. To minimize intrusion, the collection procedures were automated and structured to remove identifying details from the dataset. No personal identifiers, including usernames, profile details, or IP addresses, were retained at any stage. Consequently, the final dataset consisted solely of anonymized comment text, posting dates and times, and the associated subreddit, ensuring that no individual users could be identified.

Results

Sentiment Analysis Results

The distribution of sentiment scores for Reddit posts is shown in Table 1 and Figures 2A and 2B. Across all posts, the mean sentiment score (−0.0326, SD 0.1450) was nearly neutral and slightly on the negative side, with a median close to zero (−0.0120, IQR −0.120 to 0.0404). The overall distribution was approximately symmetric and centered around zero, with a slight right skew as shown in Figure 2. When disaggregated by subreddit, r/Monkeypox [11] and r/monkeypoxpositive [12] showed similar distributions, while r/worldnews [13] posts tended to have more negative sentiment on average (mean −0.051, SD 0.128).

Table . Summary statistics of sentiment scores for posts in the entire dataset and for each of the 4 subreddits.

Subreddit	Count	Mean (SD) sentiment	Median (IQR) sentiment	Range (min-max) sentiment
ALL	1169	−0.0326 (0.1452)	−0.0120 (−0.1260 to 0.0431)	−0.5924 to 0.6114
r/Monkeypox [11]	703	−0.0309 (0.1458)	0.0000 (−0.1294 to 0.0483)	−0.5349 to 0.6114
r/monkeypoxpositive [12]	263	−0.0290 (0.1518)	−0.0199 (−0.1205 to 0.0404)	−0.5924 to 0.5093
r/news [14]	47	−0.0189 (0.1517)	0.0000 (−0.1106 to 0.0631)	−0.3408 to 0.3455
r/worldnews [13]	156	−0.0505 (0.1277)	−0.0461 (−0.1321 to 0)	−0.4184 to 0.3719

XSL•FO
RenderX

Figure 2. Distribution of sentiment scores (A) for posts in the entire dataset and (B) for each of the 4 subreddits.

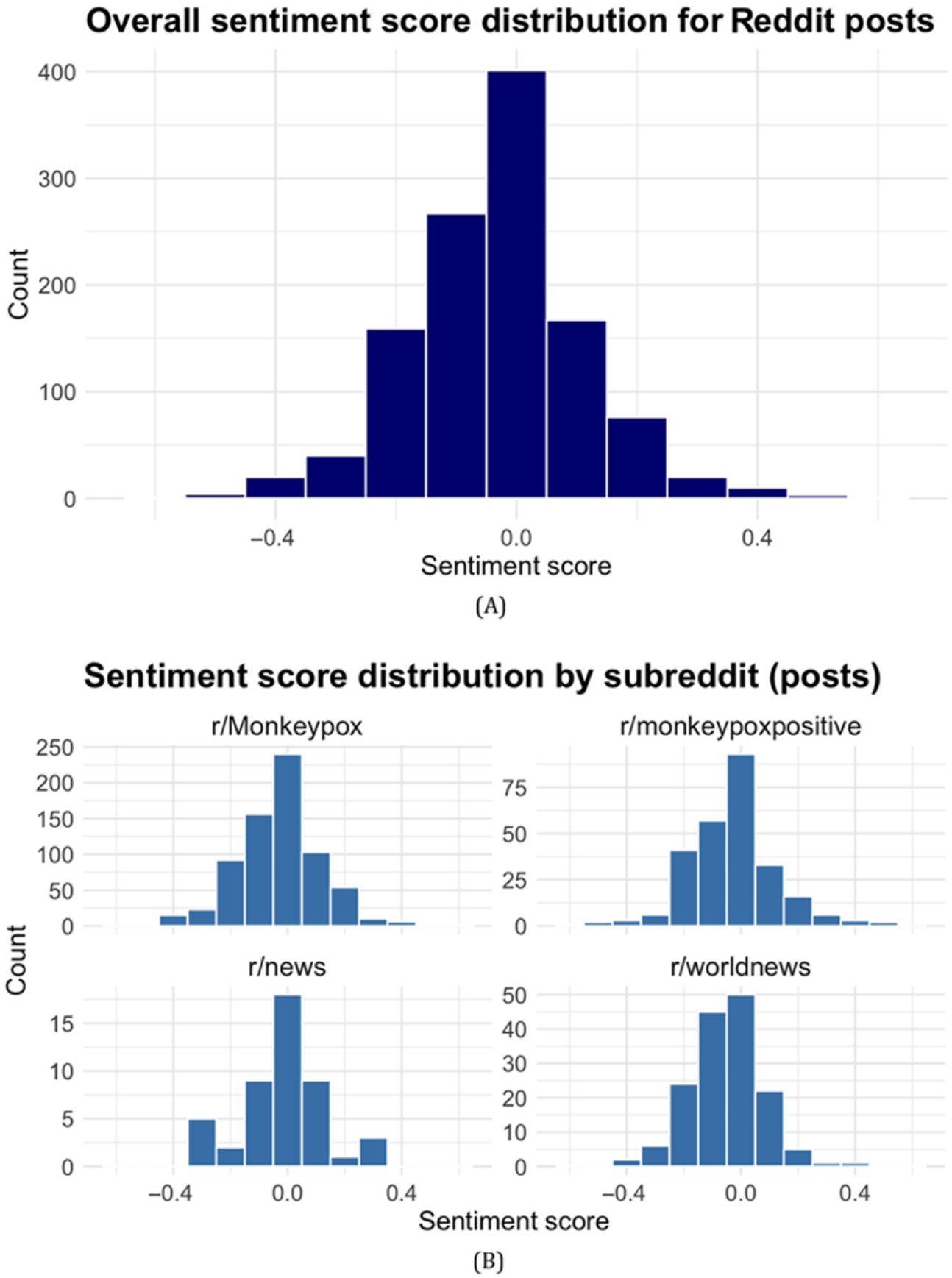

For comments, summary statistics are provided in Table 2, with distributions shown in Figures 3A and 3B. Overall, comments also centered around zero (mean −0.006, SD 0.2263; median 0, IQR −0.1335 to 0.1158), but their distribution appeared more symmetric than that of posts as shown in Figure 3A for the whole dataset and Figure 3B for each of the 4 subreddits. The

XSL•FO
RenderX

SDs of scores were higher, reflecting greater variation in comment sentiment. Subreddit-level distributions followed a similar pattern, with comments from r/Monkeypox [11] and r/monkeypoxpositive [12] exhibiting slightly more positive tendencies compared to r/news [14] and r/worldnews [13].

Table . Summary statistics of sentiment scores for comments in the entire dataset and for each of the 4 subreddits.

Subreddit	Count	Mean (SD) sentiment	Median (IQR) sentiment	Range (min-max) sentiment
ALL	6571	−0.0060 (0.2263)	0.0000 (−0.1335 to 0.1158)	−1.2478 to 1.3112
r/Monkeypox [11]	2572	0.0048 (0.2196)	0.0000 (−0.1114 to 0.1182)	−1.2478 to 1.3112
r/monkeypoxpositive [12]	1270	0.0238 (0.2175)	0.0000 (−1.090 to 0.1364)	−0.9550 to 1.0606
r/news [14]	964	−0.0287 (0.2240)	0.0000 (−0.1520 to 0.0970)	−1.2300 to 0.8573
r/worldnews [13]	1765	−0.0307 (0.2393)	0.0000 (−0.1746 to 0.1021)	−1.1976 to 1.0062

For Reddit posts, 611 of 1169 (52.3%) posts were neutral, followed by 370 (31.6%) negative and 188 (16.1%) positive posts. This suggests that although most posts adopted a neutral tone, negative sentiment was expressed about twice as often as positive sentiment. For Reddit comments, the distribution followed a similar pattern but with a greater proportion of positive content compared to posts. Neutral comments were most common, with a count of 2825 out of 6571 (43%), followed by 1962 (29.9%) negative and 1784 (27.1%) positive comments. Compared to posts, comments contained relatively less neutral and more positive sentiment.

XSL•FO
RenderX

Figure 3. Distribution of sentiment scores (A) for comments in the entire dataset and (B) for each of the 4 subreddits.

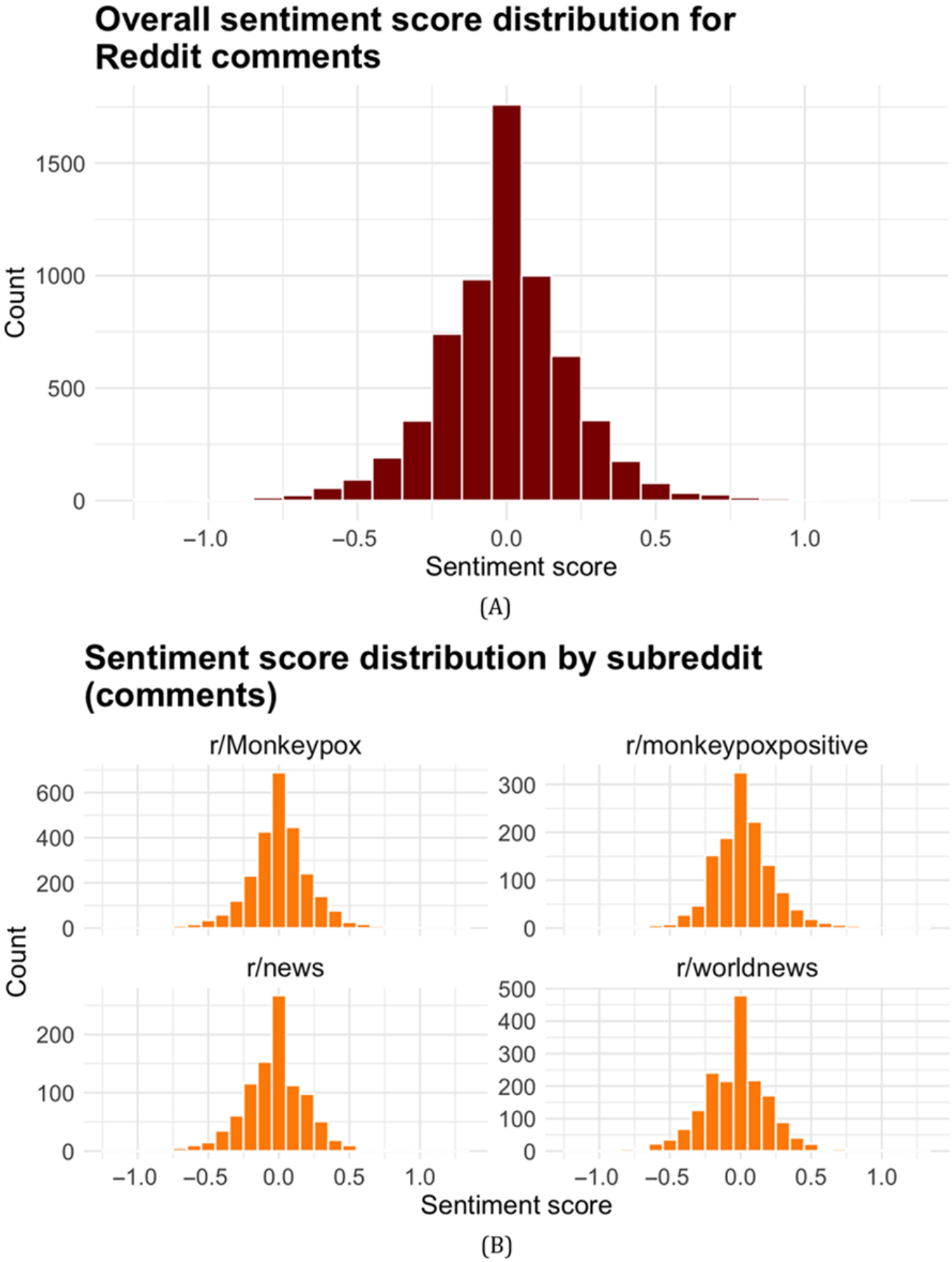

Topic Modeling Results

A total of 9 distinct topics were identified through topic modeling:

1. Discussions focus on global Mpox transmission concerns, public health responses (especially vaccine distribution and messaging), and stigma, with emphasis on LGBTQ+

XSL•FO
RenderX

communities and personal experiences of symptoms or issues. Sample text:

> *Post: Monkeypox, COVID, and Racism... Journalists and editors of global North media outlets badly need training on how to not be racist and stigmatizing in their reporting [of] Ebola, COVID-19, monkeypox, tweeted Dr. Madhu Pai…*
>
> *Comment: I'm a week past my last TPOXX dose and my anorectal lesions are still healing. They are getting better every day but it's just taking a while. Externally it looks pretty much back to normal but I can still feel the internal ones. Definitely will not be bottoming for at least another week or two. No pain when having a bowel movement, at least.*

2. Debates center on Mpox spread among MSM, vaccine access challenges, and comparisons to COVID-19, alongside personal symptom accounts and frustration with public health failures. Sample text:

> *Post: The CDC Scientist Who Couldn't Get Monkeypox Treatment*
>
> *Comment: Wasn't the vaccine two dose vaccine? How effective will it be now that you're only getting 20% of the original vaccine? How long before it becomes effective against the virus? The original dosage you get protection after 14 days of the second dose.*

3. Postpeak concerns include declining cases, lingering public health challenges (eg, misinformation and isolation fatigue), and scientific analyses of mutations and long-term impacts. Sample text:

> *Post: WHO recommends gay and bisexual men limit sexual partners to reduce the spread of monkeypox*
>
> *Comment: I'm so sorry you're feeling this and hopefully it ends soon (and doctors do something.) I'm following your story and thank you so much for sharing this, even though it's awful. Sadly people need the awareness that we can ALL get this.*

4. Criticism of systemic public health failures (eg, vaccine disparities and bureaucratic delays) and anxiety about broadening transmission (eg, pediatric cases and workplaces). Sample text:

> *Post: Thailand says mpox case recorded in foreigner traveling from Africa*
>
> *Comment: Now I'm expecting a rightwing conspiracy claim that humans cannot catch monkey pox, it is a hoax to trick people into believing in evolution.*

5. Fears of Mpox resurgence or endemicity, debates over vaccine strategies, and critiques of global disparities in outbreak response. Sample text:

> *Post: Two people diagnosed with monkeypox in London, health officials say*
>
> *Comment: Sure. Remind me how many vaxs are the US giving? 50,000?*

6. Emergence of high-risk Mpox variants (eg, Clade I), underreported global spread, and debates over containment strategies beyond MSM communities. Sample text:

> *Post: Monkeypox Can Spread In Three Ways. Here's What To Know And How To Avoid Infection.*
>
> *Comment: But we are not shutting down the sex clubs for 2 weeks to flatten the curve?*

7. Cross-border spread of strains, personal accounts of severe symptoms, and critiques of institutional response delays and inequities. Sample text:

> *Post: Is this mpox? Day of 3 since I noticed this pimple-like bump on my thigh. Not feeling any of the known mpox symptoms but the bump looks suspicious :(*
>
> *Comment: What good is 40,000 doses in a country with 80+ million?*

8. Scientific- and policy-related debates on mutations and reinfection, stigma's role in public perception, and advocacy for marginalized groups amid systemic gaps. Sample text:

> *Post: Sex workers are hit hard by Congo's mpox outbreak but say their only option is to keep working*
>
> *Comment: somehow smallpox, a more contagious and deadlier cousin, was contained. sounds like motivated reasoning from so-called experts*

9. Sociopolitical reactions dominate, including stigma, naming controversies, and comparisons to past pandemics, alongside frustration with institutional mismanagement. Sample text:

> *Post: As Chicago's Market Days approaches, doctors and organizers share monkeypox concerns, advice*
>
> *Comment: This little gem caught me by surprise with the level of accuracy and simplicity that their guide manages to achieve in describing mpox to clients who may be affected by it or interested in it as a result of living in group homes or congregate settings, and working or living independently with close caretakers. Compassionate, clear and concise. Kudos to ASAN. I'm impressed that someone put this much effort into developing materials like this specifically about mpox for a vulnerable, population.*

The distribution of posts and comments across these topics is provided in Table 3. Across all subreddits, posts were most frequently associated with Topic 8, which accounted for 371 posts. This was followed by Topic 7 with 203 posts. By contrast, Topic 4 was the least represented among posts, with only 39 entries. For comments, the distribution was somewhat different. Topic 9 dominated with 980 comments, followed by Topic 3 with 865 comments and Topic 4 with 809 comments. In comparison, Topic 8, which was the most common for posts, contained only 465 comments, the lowest among all topics.

Subreddit-level distributions revealed further variation. On r/Monkeypox [11] and r/monkeypoxpositive [12], posts again concentrated in Topic 8, while comments were more evenly distributed, with Topics 3, 4, and 9 consistently drawing higher engagement. In contrast, r/news [14] and r/worldnews [13],

XSL•FO
RenderX

which contained fewer total posts, showed a relatively weaker emphasis on Topic 8.

Table . Number of posts and comments per topic for each subreddit and the whole dataset.

Topic no	r/Monkeypox		r/monkeypoxpositive		r/news		r/worldnews		Whole dataset	
	Posts	Comments	Posts	Comments	Posts	Comments	Posts	Comments	Posts	Comments
1	75	306	21	119	7	112	24	190	127	727
2	45	277	22	125	2	122	9	192	78	716
3	40	325	21	233	3	118	16	189	80	865
4	23	308	4	259	5	54	7	188	39	809
5	35	296	16	127	8	102	16	214	75	739
6	90	235	12	117	4	109	13	198	119	659
7	134	263	39	76	8	109	22	163	203	611
8	225	241	110	63	5	61	31	100	371	465
9	36	321	18	151	5	177	18	331	77	980

Relationships Between Engagement, Sentiment, and Topic

Engagement scores were highly right-skewed, with a small number of posts receiving disproportionately high levels of engagement. To improve interpretability, Figures 4A and 4B display engagement on a log10 scale. Despite this skewed distribution, clear differences in median engagement were observed across both sentiment labels and topics.

XSL•FO
RenderX

Figure 4. Distribution of engagement scores by (A) sentiment label and (B) topic. Engagement scores are displayed on a log10 scale to account for the highly right-skewed distribution. Boxplots show median and IQRs, with points representing individual posts and comments. A total of 1052 outliers were excluded based on the 1.5×IQR criterion before plotting.

XSL•FO
RenderX

The Kruskal-Wallis test confirmed that engagement scores differed significantly across sentiment and topic categories for the dataset overall, as shown in Table 4. At the subreddit level, patterns varied: on r/Monkeypox [11], both sentiment and topic distributions differed significantly, while on r/monkeypoxpositive [12], no significant differences were observed. On r/news [14], engagement varied only by sentiment label, whereas on r/worldnews [13] differences were significant only by topic.

Table . Kruskal-Wallis test results of score distributions for sentiment label and topic.

Subreddit	Variable	*P* value	Count
ALL	Sentiment Label	<.001[a]	7740
ALL	Topic	<.001[a]	7740
r/Monkeypox [11]	Sentiment Label	<.001[a]	3275
r/Monkeypox [11]	Topic	<.001[a]	3275
r/monkeypoxpositive [12]	Sentiment Label	.52	1533
r/monkeypoxpositive [12]	Topic	.54	1533
r/news [14]	Sentiment Label	.02[b]	1011
r/news [14]	Topic	.36	1011
r/worldnews [13]	Sentiment Label	.61	1921
r/worldnews [13]	Topic	<.001[a]	1921

[a]*P*<.001.
[b]*P*<.05.

Pairwise comparisons provided additional insights. Tables 5 and 6 show Dunn test results for pairwise score distribution differences among sentiment labels and topics, respectively. These results include *z* scores and Benjamini-Hochberg adjusted *P* values. Only statistically significant (*P*<.05) comparisons are reported in Table 6. Dunn test revealed that posts with positive and negative sentiment differed the most in score distributions, while neutral content showed moderate differences compared to both groups, according to Table 5. Topic 8 differed in distribution the most compared to other topics, according to Table 6. Kolmogorov-Smirnov statistics indicated that Topics 7 and 8 stood out, exhibiting the largest differences in score distributions from other topics, although they were relatively similar to each other, as provided in Figures 5A and 5B for the sentiments and the topics, respectively.

Table . Pairwise comparison of sentiment scores (Dunn test).

Group 1	Count 1, n	Median (IQR) 1	Group 2	Count 2, n	Median (IQR) 2	*z* score	*P* value adjusted
Negative	2332	7.0 (2-27)	Positive	1972	5.0 (2-16)	6.0217	<.001
Neutral	3436	5.5 (2-22)	Positive	1972	5.0 (2-16)	3.7881	<.001
Negative	2332	7.0 (2-27)	Neutral	3436	5.5 (2-22)	2.8774	.004

XSL•FO
RenderX

Table . Pairwise comparison of topic scores (Dunn test).

Group 1	N 1	Median (IQR) 1	Group 2	N 2	Median (IQR) 2	*z* score	*P* value adjusted
3	945	4 (2-16)	8	836	10 (2-37.25)	−6.7162	<.001
4	848	4 (1.75–14.25)	8	836	10 (2-37.25)	−7.9924	<.001
1	854	5 (2-21)	8	836	10 (2-37.25)	−5.6745	<.001
2	794	5 (2-19)	8	836	10 (2-37.25)	−5.8514	<.001
5	814	5 (2-20)	8	836	10 (2-37.25)	−5.0353	<.001
8	836	10 (2-35.25)	9	1057	5 (2-20)	5.7662	<.001
3	945	4 (2-16)	7	814	8 (2-31)	−4.9035	<.001
4	949	4 (1.75–14.25)	7	814	8 (2-31)	−6.2184	<.001
6	778	6 (2-19)	8	836	10 (2-37.25)	−4.7381	<.001
2	794	5 (2-19)	7	814	8 (2-31)	−4.1211	<.001
1	854	5 (2-21)	7	814	8 (2-31)	−3.9130	<.001
7	814	8 (2-31)	9	1057	5 (2-20)	3.9131	<.001
5	814	5 (2-20)	7	814	8 (2-31)	−3.2993	.002
4	848	4 (1.75-14.25)	6	778	6 (2-19)	−3.0922	.005
6	778	6 (2-19)	7	814	8 (2-31)	−3.0240	.006
4	848	4 (1.75–14.25)	5	814	5 (2-20)	−2.8855	.008
4	848	4 (1.75–14.25)	9	1057	5 (2-20)	−2.6605	.02
1	854	5 (2-21)	4	848	4 (1.75-14.25)	2.3403	.04

XSL•FO
RenderX

Figure 5. Pairwise Kolmogorov-Smirnov statistics comparing score distributions across (A) sentiment labels, and (B) topics. KS D: Kolmogorov-Smirnov D.

XSL•FO
RenderX

Vargha-Delaney A statistics, which estimate the probability that a random score from one group exceeds that of another, highlighted systematic trends, as shown in Figures 6A and 6B for the posts and the comments, respectively. Across all topics, negative posts tended to attract higher engagement than positive ones, with this trend particularly prominent for posts. Among posts, Topic 4 achieved the highest scores on average, while Topic 6 had the lowest. In contrast, for comments, Topics 7, 5, and 8 attracted the highest scores, whereas Topic 1 and Topic 4 had the lowest.

Figure 6. Vargha-Delaney A, comparing scores across different sentiment and topic combinations for (A) posts and (B) comments.

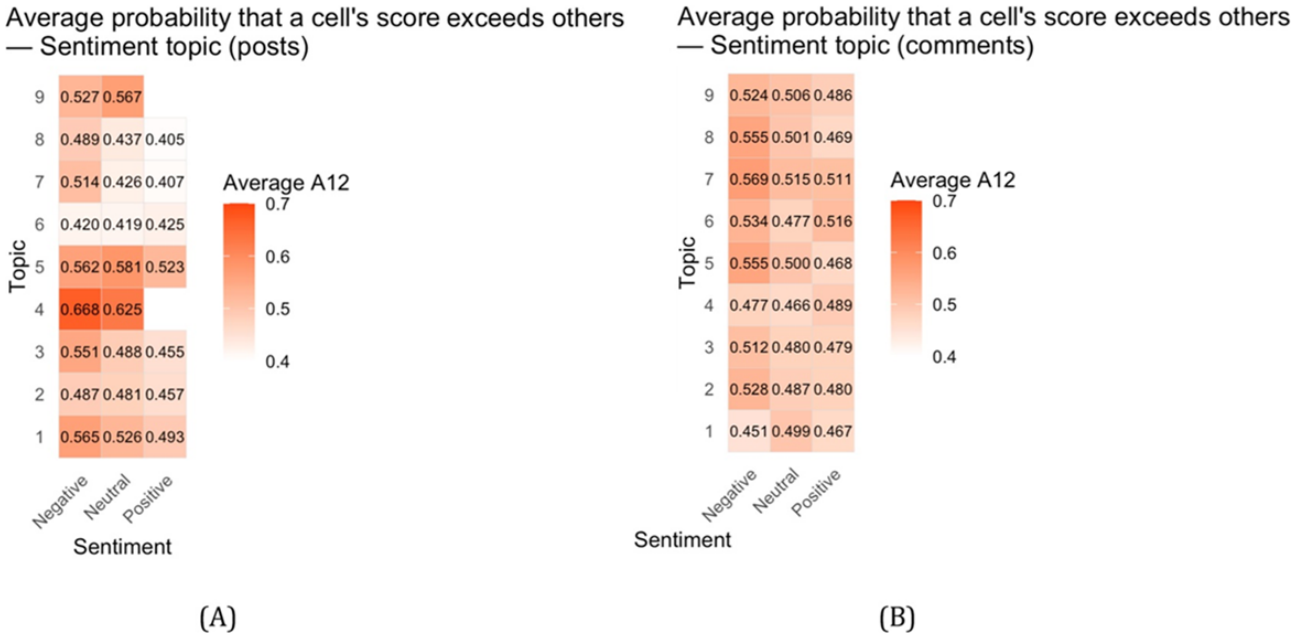

Divergence Between Post Content and Audience Responses

The relationship between post sentiment and the sentiment of associated comments is shown in Figure 7. Across all post types, neutral comments made up the majority of responses, though patterns varied with post sentiment. Positive posts tended to elicit a greater proportion of positive comments (284/922 comments, 30.8%), while negative posts attracted relatively more negative comments (526/1615 comments, 32.6%). Neutral posts drew a more balanced distribution of comment sentiments, with neutral comments remaining dominant (1194/2756 comments, 43.3%). Vargha-Delaney A statistics confirmed these patterns, showing that comments on positive posts had significantly higher sentiment scores, while comments on negative posts tended to be lower, as shown in Table 7.

XSL•FO
RenderX

Figure 7. Heatmap showing the proportion of comments in each sentiment label category for posts of different sentiment categories.

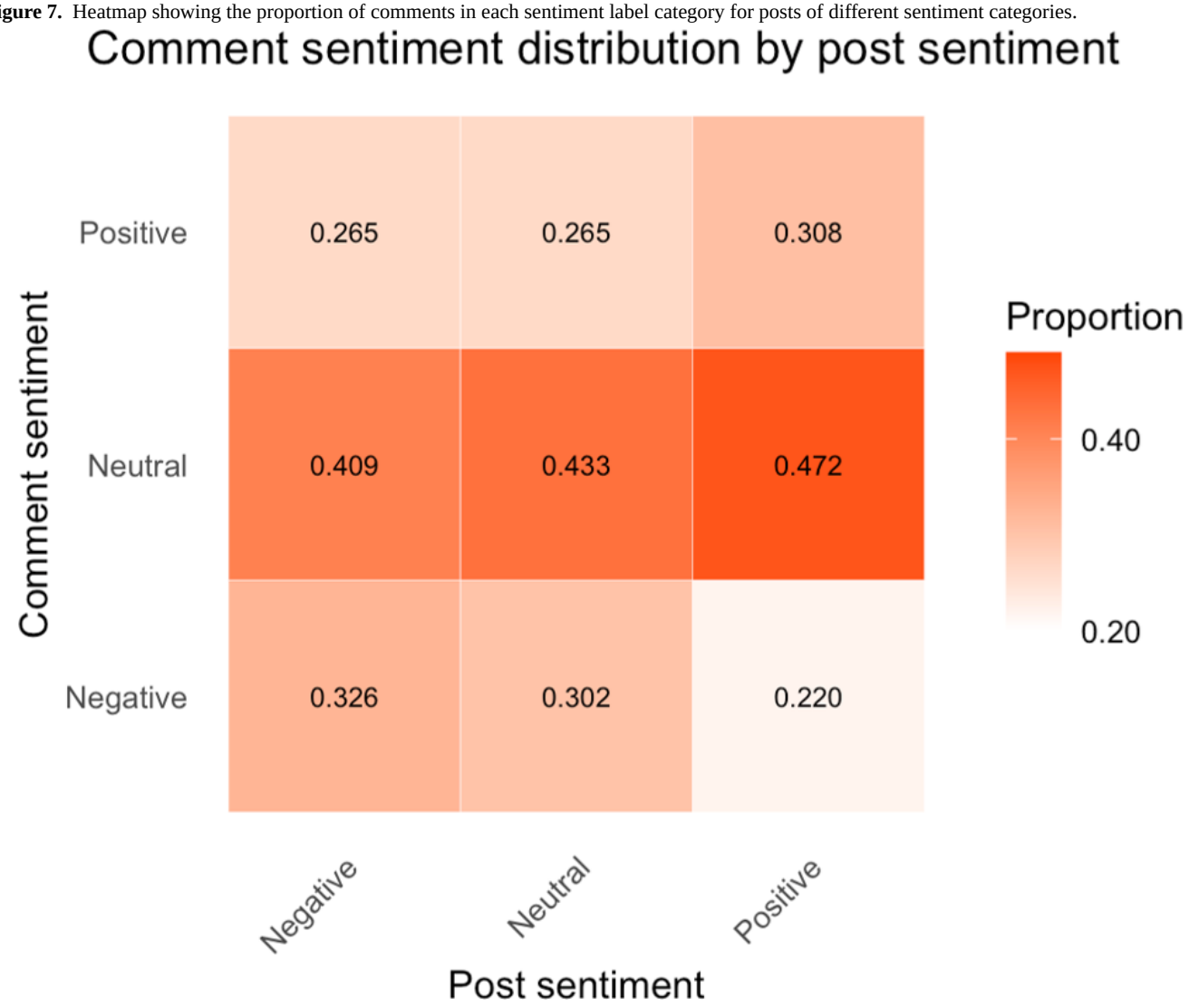

Table . Vargha-Delaney A comparing comment sentiment scores for posts of different sentiment labels.

Post sentiment category	A vs others, mean (SD)	A vs others, median (IQR)	Cliffs delta vs others, mean (SD)	Dominance share
Positive	0.5502 (0.0111)	0.5502 (0.5463-0.5542)	0.1005 (0.0222)	1.0
Neutral	0.4870 (0.0416)	0.4870 (0.4723-0.5017)	−0.0260 (0.0222)	0.5
Negative	0.4627 (0.0295)	0.4627 (0.4523-0.4732)	−0.0745 (0.0589)	0.0

Comment sentiment also varied by the topic of the original post, as shown in Table 8. Posts classified under Topic 8 elicited the highest overall comment sentiment scores, followed by Topics 2 and 1. In contrast, comments from posts on Topics 4 and 6 had the lowest sentiment scores, suggesting that discussions centered on systemic failures or new risks provoked more negative responses.

XSL•FO
RenderX

Table . Vargha-Delaney A comparing comment sentiment scores for posts of different topics.

Post sentiment label	A vs others, mean (SD)	A vs others, median (IQR)	Cliffs delta vs others, mean (SD)	Dominance share
8	0.5305 (0.0113)	0.5312 (0.5226-0.5380)	0.0609 (0.0225)	1.000
2	0.5140 (0.0152)	0.5167 (0.5085-0.5241)	0.0281 (0.0304)	0.875
1	0.5112 (0.0154)	0.5140 (0.5056-0.5208)	0.0223 (0.0308)	0.750
5	0.5019 (0.0151)	0.5059 (0.4911-0.5123)	0.0038 (0.0303)	0.625
9	0.4999 (0.0153)	0.5032 (0.4895-0.5107)	−0.0002 (0.0307)	0.500
7	0.4906 (0.0155)	0.4909 (0.4812-0.5028)	−0.0187 (0.0310)	0.375
3	0.4883 (0.0152)	0.4894 (0.4788-0.4988)	−0.0234 (0.0303)	0.250
6	0.4857 (0.0147)	0.4860 (0.4771-0.4965)	−0.0285 (0.0293)	0.125
4	0.4778 (0.0138)	0.4804 (0.4689-0.4887)	−0.0444 (0.0276)	0.000

Thematic alignment between posts and comments was further examined in Figure 8. Comment topics did not always mirror post topics, indicating that discussion often shifted focus. Regardless of the post topic, Topic 9 was the most common theme among comments. This reflects a broader trend of audiences reframing technical or health-focused posts into socially and politically charged conversations.

Figure 8. Heatmap showing the proportion of comments of each topic for posts of different topics.

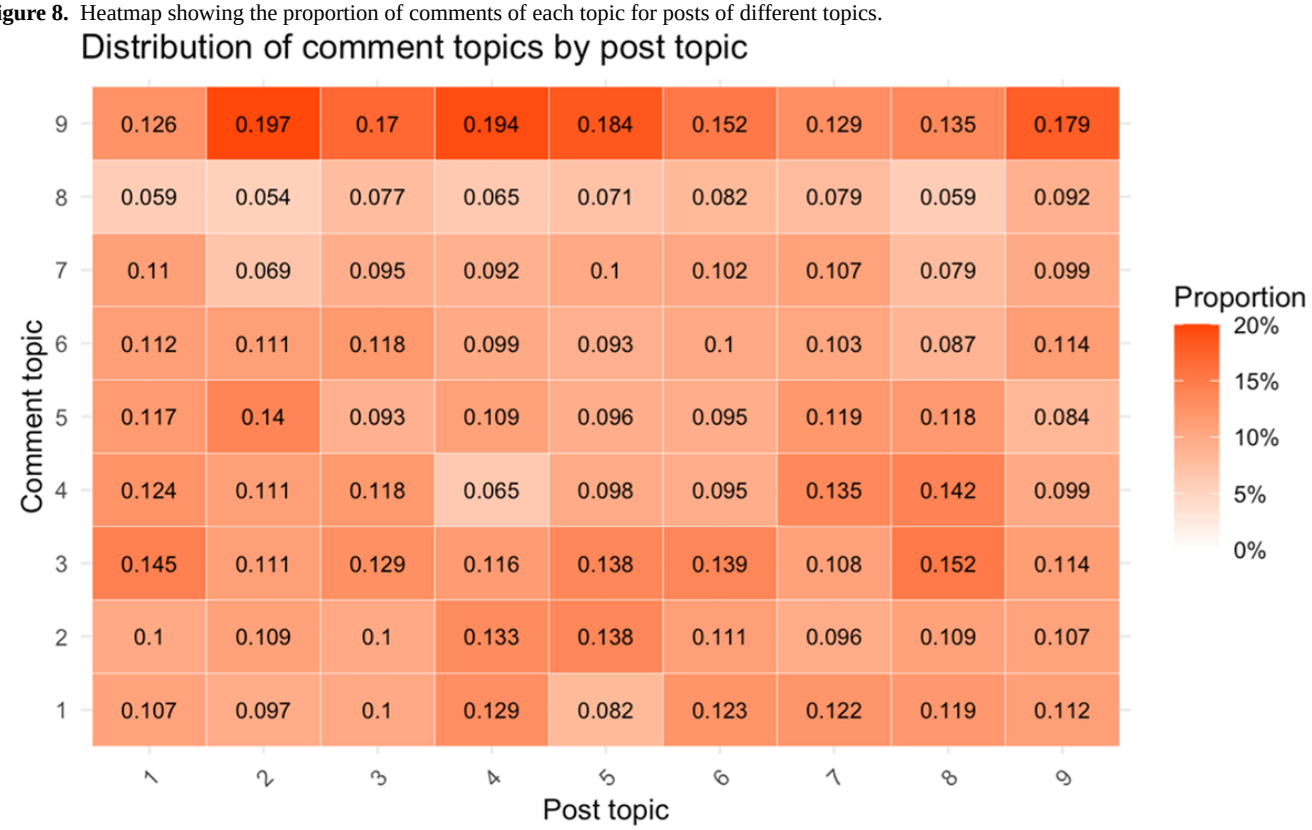

Discussion

Principal Findings

This study analyzed the topics and sentiments of Reddit posts regarding Mpox. Topic 9, which centers on sociopolitical reactions including stigma, naming controversies, and comparisons to past pandemics, dominated comments across the dataset. This heightened engagement may reflect multiple factors. First, the platform's culture likely encourages discussion and debate, particularly around topics with social or ethical implications. Second, sociopolitical and stigma-related content in health crises often elicits emotional responses, personal anecdotes, and opinion sharing, making these posts more likely to attract comments. Finally, attributes specific to the Mpox outbreak, such as attention to vulnerable populations, institutional mismanagement, and ongoing public discourse about naming and stigma, may have contributed to the higher comment volume. These factors together help explain why Topic 9 generated more discussion than other topics. This research examined how audience engagement metrics relate to the sentiment and topic types of Reddit posts about Mpox (RQ 1) and how the focus of the audience, reflected in comment sentiment and topic, varies in response to posts of different

XSL•FO
RenderX

sentiment and topic types (RQ 2). By linking original posts with the comments they generated, this study advances understanding of how public discourse around health crises evolves on social media platforms.

Relationships Between Engagement, Sentiment, and Topic

The analysis revealed that engagement scores were highly skewed, with a small number of posts receiving disproportionately high upvotes. Despite this skew, clear differences emerged across sentiment and topic categories. Negative posts tended to attract higher engagement than positive ones, and posts about systemic public health failures (Topic 4) and sociopolitical reactions (Topic 9) garnered especially high scores. By contrast, posts emphasizing technical or epidemiological details (Topic 6) received comparatively lower engagement.

These findings echo the past observation that engagement is not a uniform behavior but is shaped by specific content and emotional triggers [10]. Just as a previous study [10] found that emotionally charged political posts tended to elicit more active engagement than neutral ones, the present results suggest that negative sentiment and controversial or blame-oriented topics captured more attention on Reddit. This underscores the role of emotional salience in driving user reactions.

Divergence Between Post Content and Audience Responses

Although neutral comments were most common across all types of posts, positive posts elicited relatively more positive comments, and similarly, negative posts drew more negative comments. Furthermore, comments on posts about systemic failures (Topic 4) and high-risk variants (Topic 6) were more negative, whereas comments on posts about scientific and policy debates (Topic 8) tended to be more positive.

Thematic alignment between posts and comments was loose: comment topics often shifted away from the original post's topic, and sociopolitical reactions (Topic 9) dominated the comment space regardless of post topic. This pattern resonates with previous research on Mpox discourse, which has documented the tendency for online discussions to become infused with stigma and broader sociopolitical framings, especially around LGBTQ+ communities [3,4]. While prior studies primarily focused on the content of posts, these results extend that understanding by showing how audience responses can reframe even technically focused or neutral posts into socially charged debates.

These findings also suggest that emotionally charged and blame-oriented narratives may play a role in shaping audience engagement and subsequent discussion trajectories. Posts emphasizing institutional failures, perceived inequities, or stigmatized groups often attracted more polarized responses, which in turn redirected conversations toward broader sociopolitical concerns. Even when original posts focused on epidemiological updates or technical information, comment threads frequently introduced themes related to accountability, stigma, or policy critique. This dynamic indicates that audience engagement may not only amplify emotionally salient content but also actively reshape the framing of discussions beyond the initial informational focus.

Comparison With Previous Works

Our findings align with recent studies examining Mpox-related discourse on social media, which have reported predominantly neutral-to-negative sentiment and the presence of stigma-related narratives [3,22-26]. Prior analyses of Twitter and Reddit data during the 2022 outbreak observed that discussions frequently focused on transmission risk [22,27], vaccine access [5], and the association of Mpox with LGBTQ+ communities [3,28,29], often accompanied by misinformation and stigmatizing language [3,30-32]. These studies similarly noted that emotionally charged content tended to receive greater engagement, reinforcing the visibility of polarizing narratives [4].

The topic modeling results in this study extend previous work by identifying distinct thematic clusters spanning epidemiological updates, policy debates, and sociopolitical reactions. Earlier research has emphasized that public health discussions during outbreaks often evolve from technical information sharing to broader debates about governance, equity, and institutional trust [4,26]. Our observation that sociopolitical reactions dominated comment threads regardless of original post topic is consistent with this pattern and highlights how audience engagement can reshape discourse beyond initial informational intent.

The relationship between sentiment and engagement observed here also parallels findings from studies of COVID-19 and other health crises, where negative or controversial content was more likely to attract interaction [33]. Such dynamics may amplify emotionally salient narratives and contribute to the persistence of stigma or misinformation in online environments [34]. However, our analysis further demonstrates that technically oriented posts, particularly those focused on scientific or policy debates, were associated with relatively more positive audience responses, suggesting that constructive discourse can emerge when information is framed in evidence-based contexts.

Overall, these results complement recent research by demonstrating not only the thematic structure of Mpox discussions but also the divergence between original post content and audience responses. This interactional perspective provides additional insight into how public discourse evolves during health emergencies and underscores the importance of anticipating audience-driven reframing in public health communication strategies [35].

Limitations

However, several limitations should be noted when interpreting these findings. First, the data were drawn exclusively from Reddit, which is mostly popular in North America and other English-speaking countries. Therefore, the results may not reflect the global landscape of Mpox discourse or the communication norms of other platforms. Second, only 4 subreddits were included, and r/Monkeypox [11] contributed a disproportionately large number of posts and comments. This imbalance could bias the observed patterns toward the norms of that particular community, potentially underrepresenting the perspectives seen in more generalist spaces. Third, engagement

XSL•FO
RenderX

was measured using Reddit scores (upvotes minus downvotes), which capture one form of audience response, but do not fully represent other interaction behaviors such as commenting, sharing, or passive viewing. Fourth, engagement differences by sentiment may be influenced by unmeasured factors such as author visibility, posting frequency, or source type (eg, personal narratives vs news links), which were not explicitly controlled for in the analysis. Although no individual author contributed more than 25 posts out of the 1169 total (2.1% of the dataset), suggesting that no single author dominated the dataset, residual confounding by author-level or source-level characteristics cannot be ruled out. Fifth, the long temporal range of the dataset (2021 - 2025) spans both the early emergency phase of the outbreak and later periods of reduced transmission. As a result, the analysis may conflate discourse dynamics that differ between acute crisis communication and more endemic-phase discussion. Sixth, keyword-based filtering was used to identify Mpox-related posts, which may have omitted relevant discussions occurring in broader news or general-interest subreddits that did not explicitly use the selected terms.

In addition, concise topic theme labels were generated with assistance from an LLM after statistical topic assignments were finalized. While the clustering itself was produced using LDA, the summarization step introduces some degree of stochasticity and subjectivity inherent to LLM-generated outputs. To improve stability, the summarization process was repeated multiple times using different random samples of documents per topic, and the resulting summaries were aggregated, though the themes were not independently validated by multiple human coders or alternative models.

Furthermore, due to computational and token limitations, the LLM-based interpretation step relied on randomly sampled subsets of documents from each topic rather than the full set of posts and comments assigned to that topic. Although these samples were drawn repeatedly across multiple runs to capture representative content, a small proportion of topic-related messages may not have been included in the summaries used to derive the final theme labels.

Finally, sentiment analysis relied on a lexicon-based approach, which may not fully capture sarcasm, irony, or context-specific language commonly found in Reddit discussions. Such nuances may affect the precision of sentiment classification.

While these limitations restrict the generalizability of the findings, they also highlight the value of Reddit as a lens into globally circulating narratives, concerns, and framings of Mpox that may cross national boundaries.

Conclusion

This study investigated how audience engagement relates to the sentiment and topic types of Reddit posts about Mpox (RQ 1) and how the sentiment and topical focus of comments vary in response to posts of different sentiments and topics (RQ 2). The analysis revealed that engagement was unevenly distributed, with negative posts and those focused on systemic public health failures or sociopolitical reactions receiving higher scores, while posts emphasizing technical details or emerging variants drew lower engagement. Although neutral comments were most common, positive posts tended to elicit more positive comments and negative posts more negative comments. In addition, the focus of comments often differed from that of the original post, with sociopolitical discussions dominating regardless of the post topic.

These findings highlight the complex ways in which public discourse on health crises evolves from initial posts to audience responses. Emotionally charged and blame-oriented content appears more likely to attract engagement, while subsequent discussions frequently reframe posts around sociopolitical concerns. Understanding these dynamics can inform public health communication strategies by helping practitioners design messages that anticipate likely audience reactions, mitigate panic, stigma, and misinformation, as well as foster constructive dialogue during outbreaks. Future research could build on these findings by examining other platforms and expanding the scope beyond Reddit to capture more diverse public perspectives.

Acknowledgments

The generative artificial intelligence tool Gemini 1.5 Flash by Google was used during the topic modeling phase of the study to assist in generating human-readable summaries and labels for the identified topic categories. These summaries were reviewed, refined, and validated by the authors before inclusion in the analysis and paper.

Funding

This research was funded by the NSERC Discovery Grant (RGPIN-2022-04559). JDK acknowledges support from the NSERC Discovery Launch Supplement (DGECR-2022-00454), the SSHRC New Frontiers in Research Fund–Exploratory (NFRFE-2021-00879), and Canada's International Development Research Centre (IDRC; 109981), as well as IDRC and the Foreign, Commonwealth and Development Office (FCDO; 110554-001).

Data Availability

The datasets generated or analyzed during this study are available from reference [36].

XSL•FO
RenderX

Authors' Contributions

XNL contributed to conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing of the original draft, and review and editing of the paper. ZMN contributed to formal analysis, investigation, methodology, supervision, and review and editing of the paper. JDK contributed to conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, writing of the original draft, and review and editing of the paper.

Conflicts of Interest

None declared.

References

1. El Eid R, Allaw F, Haddad SF, Kanj SS. Human monkeypox: a review of the literature. PLOS Pathog 2022 Sep;18(9):e1010768. [doi: 10.1371/journal.ppat.1010768] [Medline: 36136979]
2. Yan C, Law M, Nguyen S, Cheung J, Kong J. Comparing public sentiment toward COVID-19 vaccines across Canadian cities: analysis of comments on Reddit. J Med Internet Res 2021 Sep 24;23(9):e32685. [doi: 10.2196/32685] [Medline: 34519654]
3. Movahedi Nia Z, Bragazzi N, Asgary A, Orbinski J, Wu J, Kong JP. Mpox panic, infodemic, and stigmatization of the two-spirit, lesbian, gay, bisexual, transgender, queer or questioning, intersex, asexual community: geospatial analysis, topic modeling, and sentiment analysis of a large, multilingual social media database. J Med Internet Res 2023 May 1;25:e45108. [doi: 10.2196/45108] [Medline: 37126377]
4. Hong C. Mpox on Reddit: a thematic analysis of online posts on mpox on a social media platform among key populations. J Urban Health 2023 Dec;100(6):1264-1273. [doi: 10.1007/s11524-023-00773-4] [Medline: 37580545]
5. Anoop VS, Sreelakshmi S. Public discourse and sentiment during MPOX outbreak: an analysis using natural language processing. Public Health (Fairfax) 2023 May;218:114-120. [doi: 10.1016/j.puhe.2023.02.018] [Medline: 37019026]
6. Edinger A, Valdez D, Walsh-Buhi E, et al. Misinformation and public health messaging in the early stages of the mpox outbreak: mapping the Twitter narrative with deep learning. J Med Internet Res 2023 Jun 6;25:e43841. [doi: 10.2196/43841] [Medline: 37163694]
7. Fan L, Li L, Hemphill L. Toxicity on social media during the 2022 mpox public health emergency: quantitative study of topical and network dynamics. J Med Internet Res 2024 Dec 12;26:e52997. [doi: 10.2196/52997] [Medline: 39666969]
8. Garcia-Iglesias J, May T, Pickersgill M, et al. Social media as a public health tool during the UK mpox outbreak: a qualitative study of stakeholders' experiences. BMJ Public Health 2023 Nov;1(1):e000407. [doi: 10.1136/bmjph-2023-000407] [Medline: 40017890]
9. Boyce MR, Gordon M, Piltch-Loeb R, Katz R. Chicago public health department social media communications on Twitter during the COVID-19 pandemic and the mpox epidemic: cross-sectional content analysis. J Med Internet Res 2025 Jul 18;27:e68200. [doi: 10.2196/68200] [Medline: 40680161]
10. Tenenboim O. Comments, shares, or likes: what makes news posts engaging in different ways. Social Media + Society 2022 Oct;8(4). [doi: 10.1177/20563051221130282]
11. R/monkeypox. Reddit. URL: https://www.reddit.com/r/Monkeypox [accessed 2026-02-24]
12. R/monkeypoxpositive. Reddit. URL: https://www.reddit.com/r/monkeypoxpositive [accessed 2026-02-24]
13. R/worldnews. Reddit. URL: https://www.reddit.com/r/worldnews [accessed 2026-02-24]
14. Reddit, r/news. URL: https://www.reddit.com/r/news [accessed 2026-02-24]
15. ReddAPI. Rapid. 2025. URL: https://rapidapi.com/SeasonedCode/api/reddapi [accessed 2026-06-04]
16. The R Project for Statistical Computing. The R Foundation. 2026. URL: https://www.R-project.org [accessed 2026-03-25]
17. Homepage. Posit. 2025. URL: http://www.posit.co [accessed 2026-03-25]
18. sentimentr. GitHub. 2021. URL: https://github.com/trinker/sentimentr [accessed 2026-06-04]
19. Grün B, Hornik K. topicmodels: an R package for fitting topic models. J Stat Softw 2011. [doi: 10.18637/jss.v040.i13]
20. Grün B, Hornik K. Topicmodels: topic models. CRAN (Comprehensive R Archive Network). 2024. URL: https://CRAN.R-project.org/package=topicmodels [accessed 2026-06-04]
21. Kennedy-Shaffer L. Before $p < 0.05$ to beyond $p < 0.05$: using history to contextualize p-values and significance testing. Am Stat 2019;73(sup1):82-90. [doi: 10.1080/00031305.2018.1537891] [Medline: 31413381]
22. Cooper LN, Radunsky AP, Hanna JJ, et al. Analyzing an emerging pandemic on Twitter: monkeypox. Open Forum Infect Dis 2023 Apr;10(4):ofad142. [doi: 10.1093/ofid/ofad142] [Medline: 37035497]
23. Thakur N. MonkeyPox2022Tweets: a large-scale Twitter dataset on the 2022 monkeypox outbreak, findings from analysis of tweets, and open research questions. Infect Dis Rep 2022 Nov 14;14(6):855-883. [doi: 10.3390/idr14060087] [Medline: 36412745]
24. Farahat RA, Head MG, Tharwat S, et al. Infodemic and the fear of monkeypox: call for action. Trop Med Health 2022 Sep 5;50(1):63. [doi: 10.1186/s41182-022-00459-8] [Medline: 36058955]

XSL•FO
RenderX

25. Ji-Xu A, Htet KZ, Leslie KS. Monkeypox content on TikTok: cross-sectional analysis. J Med Internet Res 2023 Jan 17;25:e44697. [doi: 10.2196/44697] [Medline: 36649057]
26. Dsouza VS, Rajkhowa P, Mallya BR, et al. A sentiment and content analysis of tweets on monkeypox stigma among the LGBTQ+ community: a cue to risk communication plan. Dialogues Health 2023 Dec;2:100095. [doi: 10.1016/j.dialog.2022.100095] [Medline: 36573228]
27. Isidro J, Borges V, Pinto M, et al. Phylogenomic characterization and signs of microevolution in the 2022 multi-country outbreak of monkeypox virus. Nat Med 2022 Aug;28(8):1569-1572. [doi: 10.1038/s41591-022-01907-y] [Medline: 35750157]
28. Coker DC, Al-Ahdal T. Social media and the response to mpox. In: The Scientific Basis of Mpox (Monkeypox): Features, Prevention, and Treatment 2025:101-113. [doi: 10.1016/B978-0-443-22123-1.00028-4]
29. Lai S, Dong J, Lin Z, et al. Sentiment analysis based on machine learning for infodemic during the mpox epidemic. Humanit Soc Sci Commun 2026;13(1):27. [doi: 10.1057/s41599-025-06324-3]
30. Calabrese C, Kumble S, Yu X. Stigmatization of mpox and the MSM community on Twitter: a computational approach to exploring the negative emotions expressed through stigmatizing language. Telemat Inform 2025 Oct;102:102326. [doi: 10.1016/j.tele.2025.102326]
31. Scalvini M, Fletcher S. The economy of suffering during the mpox crisis. Inf Commun Soc 2025 Oct 3;28(13):2413-2436. [doi: 10.1080/1369118X.2025.2526712]
32. Ng QX, Yau CE, Lim YL, Wong LKT, Liew TM. Public sentiment on the global outbreak of monkeypox: an unsupervised machine learning analysis of 352,182 Twitter posts. Public Health 2022 Dec;213:1-4. [doi: 10.1016/j.puhe.2022.09.008] [Medline: 36308872]
33. Tangcharoensathien V, Calleja N, Nguyen T, et al. Framework for managing the COVID-19 infodemic: methods and results of an online, crowdsourced WHO technical consultation. J Med Internet Res 2020 Jun 26;22(6):e19659. [doi: 10.2196/19659] [Medline: 32558655]
34. Spruijt I, Cronin A, Udeorji F, et al. Respected but stigmatized: healthcare workers caring for COVID-19 patients. PLoS One 2023;18(7):e0288609. [doi: 10.1371/journal.pone.0288609] [Medline: 37478112]
35. Anawade PA, Sharma D, Gahane S. Connecting health and technology: a comprehensive review of social media and online communities in healthcare. Cureus 2024 Mar;16(3):e55361. [doi: 10.7759/cureus.55361] [Medline: 38562335]
36. Mpox-reddit-sentiment-analysis. GitHub. URL: https://github.com/QQ1723/Mpox-Reddit-Sentiment-Analysis [accessed 2026-06-05]

Abbreviations

JSD: Jensen-Shannon divergence
LDA: latent Dirichlet allocation
LGBTQ+: lesbian, gay, bisexual, transgender, queer, and more
LLM: large language model
RQ: research question

Edited by C Argyropoulos; submitted 22.Dec.2025; peer-reviewed by A Girma, J Zhang, L Cooper; revised version received 31.Mar.2026; accepted 20.Apr.2026; published 23.Jun.2026.

Please cite as:
Luo XN, Movahedi Nia Z, Kong JD
Reddit Discussions During the 2022 Mpox Outbreak: Observational Analysis of Sentiment, Topics, and Audience Engagement
J Med Internet Res 2026;28:e90152
URL: https://www.jmir.org/2026/1/e90152
doi:10.2196/90152

© Xi Ning Luo, Zahra Movahedi Nia, Jude Dzevela Kong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Characteristics of Human-Like Virtual Profiles in Relation to Audience Reach and Engagement on Instagram: Secondary Data Analysis

Alexandra Maria Lebrecht, MSci; Winze Tam, MSci; Omar Merlo, PhD; Andreas Benedikt Eisingerich, PhD

Imperial Business School, Imperial College London, Exhibition Road, South Kensington Campus, London, United Kingdom

Corresponding Author:
Omar Merlo, PhD
Imperial Business School, Imperial College London, Exhibition Road, South Kensington Campus, London, United Kingdom

Abstract

Background: Prior research on human-like virtual profiles (VPs)—computer-generated imagery (CGI; displaying legible artificiality) or artificial intelligence (AI)–generated personas on social media—is often based on small samples and limited data. Furthermore, prior studies group highly photorealistic virtual influencers with abstract virtual characters, complicating comparisons. Recent generative AI has enabled large-scale production of synthetic content. These advances have drastically increased the quality and volume of synthetic media. Much of the prior literature predates these developments, leaving open questions about how posting behavior and VP design now relate to reach and engagement outcomes.

Objective: This study aimed to examine how posting behavior and key VP attributes, namely, photorealism; physical, behavioral, and narrative consistency; and human copresence related to reach and engagement across content formats (static images vs videos) on Instagram.

Methods: A total of 157 human-like female VPs were included in the final Instagram dataset. During the initial screening stage, 5 human-like male macro and mega VPs were identified but excluded because their number was too small to support meaningful subgroup comparison. Engagement was operationalized as like rate (likes or followers) for images and videos; reach was measured via examined absolute impressions. Performance was summarized using the single best-performing post and the arithmetic mean of the top 3 posts per format. Profiles and posts were further coded using predefined content variables covering visual realism, identity consistency, copresence structure, appearance patterns, and body-type representation.

Results: Higher-performing content was more commonly observed among larger profiles, particularly for video metrics. Among videos exceeding 20 million impressions, out of 12 videos, 7 (58%) were produced by CGI-like VPs, whereas high video like rates were more often observed among photorealistic or face-swapped profiles. In the image engagement analysis, out of 8 highest-like-rate posts, 6 (75%) featured VPs with dark hair and dark eyes, and out of the 8 posts, 4 (50%) included copresence of multiple subjects. Furthermore, of the 157 top-performing profiles reviewed, 122 (77.71%) demonstrated stable visual identity and consistent behavioral and narrative presentation, whereas noticeable inconsistencies were rarely observed in this subset. These findings represent descriptive patterns within established macro- and mega-level profiles rather than as causal predictors of growth.

Conclusions: Among established human-like female VPs on Instagram, higher engagement and reach frequently co-occurred with larger profile scale, identity coherence, and transparent virtual presentation than with photorealism alone. CGI-like aesthetics aligned more with algorithmic distribution, whereas photorealistic motion content featured more prominently among posts with high active engagement. These results suggest that audiences may respond more favorably to coherent and legible virtual identities than to ambiguous realism. Because the study was restricted to macro- and mega-level profiles, the observed traits should be interpreted as characteristics of successful incumbents rather than as general predictors of VP growth.

(J Med Internet Res 2026;28:e86233) doi:10.2196/86233

KEYWORDS

virtual influencers; artificial intelligence; generative AI; Instagram; audience engagement; audience reach; photorealism; content analysis; ethical design; computer-generated image

XSL•FO
RenderX

Introduction

Overview

A virtual human or virtual profile (VP) is a fictional character on social media, operated by a creative professional or agency, that defines the full spectrum of its online presence, including physical appearance, personality, and narrative identity [1]. When VP 2, one of the earliest and most prominent digital-born human-like virtual celebrities, debuted on Instagram in 2016, the idea of a convincingly human-like synthetic persona was still largely unfamiliar to mainstream audiences [2]. Since then, human-like VPs have moved from novelty to a visible creator category, supported by rapid improvements in generative artificial intelligence (AI) that have lowered production barriers and increased the visual quality and volume of synthetic content distributed on major platforms [3,4].

This technological shift matters for research because much of the previous virtual influencer (VI) literature was produced under materially different conditions. Earlier work often evaluated relatively small samples, relied on manual selection of posts, or aggregated fundamentally diverse character styles, including abstract non–human-like VPs into a single "VI" category [5,6]. In practice, however, even the human-like VP category occupies a spectrum ranging from explicitly computer-generated imagery (CGI; displaying legible artificiality)–like personas to highly photorealistic VPs. These distinctions are not merely aesthetic. They may influence how audiences interpret the account's intent (entertainment vs deception), how authenticity is inferred, and how platform governance (eg, disclosure expectations or AI-content labels) interacts with user response [7-9].

Against this backdrop, this study examines reach and engagement performance among human-like female VPs on Instagram using a contemporary secondary dataset. The study is motivated by 2 guiding principles of today's VP environment. First, video is increasingly central to platform distribution and creator strategy, and it offers impressions and reach information that is not systematically visible for static images on Instagram; consequently, studies that treat image-based engagement as representative of overall performance may understate the role of algorithmic amplification [10]. Second, audience literacy has changed: as users have become more familiar with synthetic media, the success conditions for VPs may have shifted from maximizing photorealism toward maximizing identity coherence, perceived transparency, and authenticity [7,8].

In line with these developments, the study evaluates 3 core design characteristics, namely, photorealism, physical and behavioral consistency, and copresence structure, and how they descriptively relate to reach and engagement across content formats. Accordingly, the following exploratory propositions were specified: (P1) Photorealism and explicit CGI-likeness co-occur with distinct reach and engagement patterns, rather than one style uniformly outperforming the other. (P2) Stable physical, ethnic, and behavioral presentation is frequently observed among high-reach and high-engagement profiles. (P3) Posts featuring copresence, whether human-VP or VP-VP, are frequently observed among high-performing posts, although solo VP posts also appear in this subset. (P4) Specific appearance patterns, such as dark hair, dark eyes, and exaggerated body proportions, recur among top-performing VPs.

By examining these exploratory propositions, this study aims to contribute to the growing body of research on human-AI interaction, shedding light on how audiences engage emotionally and behaviorally with AI-generated personas in digital environments. The study further advances knowledge by clarifying which VP content deployment strategies and design signals are most prominent among high-performing accounts on Instagram. It also strives to provide a more valid basis for comparison than approaches that rely primarily on high-performing outliers or noncomparable metrics [5,6,11].

Background

Prior research examining the degree of human-likeness in VIs and its association with user engagement has yielded mixed results. While some studies suggest that VIs can elicit lower perceived authenticity and trustworthiness than physical human influencers in certain contexts [7,12], other work indicates that engagement outcomes can be comparable in some settings [10] and may be shaped by novelty, curiosity, and creator control [6]. This divergence is consistent with findings in related fields such as embodied and conversational agents and digital health, where human-like cues can strengthen social presence and relational engagement but, consistent with the "uncanny valley" hypothesis, may also reduce affinity when high realism creates expectancy violations or perceptual discomfort [13-15]. Achieving the right balance of human-likeness is therefore crucial for maintaining trust, perceived authenticity, and engagement in human-virtual interactions [13].

Several studies have compared VIs with traditional human influencers to better understand their relative effectiveness, advantages, and limitations. These analyses have produced mixed results. Some research suggests that VIs may be perceived as less trustworthy or relatable than human influencers but benefit from novelty, originality, and design control [6,16]. Conversely, other studies emphasize that VIs often struggle to replicate the emotional connection and sense of authenticity that human influencers achieve with their audiences [12,17]. This phenomenon can be explained through social identity theory, which posits that individuals are more likely to identify with others who display similar human or social characteristics [18].

Authenticity, a central construct in influencer marketing and digital communication, has also been explored extensively in relation to VPs [7,8]. Prior work highlights that a VP's success depends not only on visual or behavioral realism but also on the consistency of these authentic cues across contexts [19]. More recent studies identify clusters of VI attributes that foster user compliance, identification, and internalization, and are linked to specific behavioral responses [20]. Together, these findings suggest that perceived authenticity and relational consistency are fundamental to sustaining engagement in human-AI interactions. Furthermore, navigating authenticity remains a primary challenge in influencer marketing, as creators and brands must continuously negotiate trust with their

XSL•FO
RenderX

audiences [21]; a dynamic that becomes even more complex when the influencer is entirely synthetic.

Insights regarding digital health show that relational and safety factors also shape human-virtual interactions, offering perspectives that complement research on VPs. In digital health, for instance, scholars have examined embodied conversational agents in clinical psychology and related areas. Embodied conversational agents are used to deliver automated human-support factors and are increasingly explored for psychotherapeutic interventions; however, controlled evidence of clinical effectiveness remains scarce [14]. This idea aligns with the notion of relational agents, computational systems intentionally designed to build and sustain long-term social-emotional relationships with users to facilitate ongoing engagement [15].

However, research in health communication also reveals the risks of deploying conversational AI without sufficient safeguards. An observational study of Siri, Alexa, and Google Assistant identified potential patient safety risks, as these systems occasionally provided incomplete or unsafe responses to medical queries [22]. Similarly, a systematic review of mental health chatbots found only 2 studies assessing safety, both reporting no adverse events yet highlighting the low overall quality of evidence [23]. More recent work on AI-based conversational agents for chronic conditions has reported generally positive results for usability, satisfaction, and engagement, although variability remains high and standardized evaluation is lacking [24]. A 1-year observational study of a chatbot for patients with breast cancer likewise found high engagement and satisfaction but emphasized the need for controlled trials to assess health outcomes [25]. Across these studies, design quality, authenticity, and relational engagement consistently emerge as key factors associated with user trust and sustained use.

Together, these streams of research demonstrate that the perceived authenticity and relational design of digital agents—whether health chatbots or VIs—are closely related to human engagement and trust. Yet, despite these parallels, the factors relating to engagement among highly photorealistic, AI-generated VPs remain underexplored. Many existing studies were conducted before the emergence of sophisticated generative AI tools capable of creating lifelike images and videos. Consequently, prior findings may not reflect current user perceptions and behaviors toward hyperrealistic virtual personas. This study addresses this gap by analyzing contemporary VPs created using advanced generative AI technologies to better understand how photorealism, behavioral consistency, and contextual alignment relate to user engagement and trust in digital environments.

Methods

Study Design

This research investigates how VP design and posting outcomes relate to 2 distinct performance concepts: interaction-based engagement and distribution-based reach. More specifically, engagement is operationalized as like rate (likes divided by follower count) for both images and videos, reflecting an observable, comparable interaction signal. Furthermore, reach is examined only for video content via absolute impressions, with exceptional virality for the category of human-like VPs defined as videos exceeding 20 million impressions, a threshold that captures rare, high-magnitude distribution events without normalizing by follower count. This distinction is conceptually important: impressions reflect algorithmic exposure enabled by positive feedback loops, whereas *likes* reflect active user response, and the 2 may not move in parallel across VP types or content strategies [6,10]. Because inferential statistical comparison was not feasible to be performed on the underlying dataset, the study's conclusions are limited to descriptive patterns and exploratory interpretation.

Virtual Content Design Coding

Overview

A structured content-coding procedure was applied across the full sample of 157 human-like female VPs included in the final dataset. Coding focused on stable profile-level attributes and recurring content design features that could plausibly relate to audience response. All coding variables were defined in advance and applied using fixed operational decision rules to improve transparency and reduce subjectivity. The complete coding framework, including operational definitions and decision rules, is provided in Table S1 in Multimedia Appendix 1. This full-sample coding provides the baseline descriptive distribution used to contextualize whether specific traits observed among high-performing outliers represent a descriptive concentration relative to the broader VP ecosystem. Following this full-sample profile-level coding, a more detailed postlevel structured review was conducted only for posts meeting the predefined performance thresholds specified in the research design.

Core Design Variables

Consistent with the study's exploratory propositions, the primary appearance-related differentiation concerned the visual realism of the VP. Each profile was coded as either CGI-like or photorealistic. Profiles were classified as CGI-like when their artificial nature was visually explicit through stylized textures, nonphotographic rendering, exaggerated proportions, or clearly synthetic facial features. Profiles were classified as photorealistic when visual cues closely resembled human photographic realism and could plausibly be mistaken for real human imagery, particularly in static image content.

In addition, profiles were coded for identity and content consistency using predefined observational criteria. Consistency was coded dichotomously (consistent vs inconsistent) based on the stability of facial identity, body representation, overall visual appearance, and behavioral or narrative presentation across posts. Profiles showing repeated and noticeable shifts in these features were coded as inconsistent. That is, minor variation attributable to pose, clothing, lighting, or camera angle was not sufficient for an inconsistency code unless it materially altered the apparent identity of the VP.

XSL•FO
RenderX

Copresence Coding

To assess the social structure of depicted content, each analyzed post was coded for copresence structure using mutually exclusive visual categories: (1) solo VP content, in which 1 focal VP appeared without any additional clearly identifiable actor; (2) VP-VP copresence, in which 2 or more VPs appeared together within the same frame; (3) VP-human copresence, in which a VP appeared alongside a clearly identifiable human individual; and (4) no codable copresence, in which visual material was insufficient to classify additional actors with confidence.

Because hybrid face-swapped content may otherwise conflate narrative structure with production technique, face-swapping was treated separately as a production-related characteristic rather than as a copresence category in itself. Accordingly, a face-swapped video showing only 1 apparent actor was coded as solo VP content, whereas a face-swapped video showing a VP alongside another clearly identifiable human was coded as VP-human copresence.

Exploratory Appearance Patterns

An additional exploratory coding layer was introduced following the observation of recurring visual patterns among high-performing profiles. Specifically, hair and eye color combinations were coded to examine whether certain appearance traits appeared repeatedly among the top-performing posts. Four categories were used: (1) dark hair and dark eyes, (2) dark hair and bright eyes, (3) bright hair and dark eyes, and (4) bright hair and bright eyes.

These appearance categories were treated as descriptive variables only and were not intended to imply fixed biological or cultural preference structures.

Body-Type Coding

Finally, profiles were coded for body-type representation using a binary distinction between exaggerated sexual dimorphism and nonexaggerated body representation. This variable was operationalized using observable visual cues, including markedly amplified bust-to-waist and/or hip-to-waist contrast relative to the surrounding body proportions depicted in the account's content. Because the final analytic sample included only human-like female VPs, this coding variable was developed for female-presenting bodies and should not be interpreted as a gender-neutral anthropometric classification. The variable was included as a descriptive coding dimension to capture recurring stylization patterns in the sample VP ecosystem rather than to assign normative judgments about body shape.

Because generative AI and CGI frequently produce hyperstylized anatomies that need not conform to human biological norms, applying standard clinical anthropometric thresholds (eg, World Health Organization waist-to-hip ratios) is not methodologically appropriate for synthetic media. Instead, our operational definition of exaggerated sexual dimorphism is grounded in media studies literature on the sexualization and unrealistic body representation of female characters in virtual environments and video games [26]. This approach ensures that the coding captures virtual hyperidealization rather than treating synthetic bodies as human medical cases.

Coding Procedure

All coding was conducted using predefined operational definitions and decision rules to minimize subjectivity. Ambiguous cases—such as unclear background figures, uncertain content provenance, or insufficient visual information—were excluded from variable-specific coding rather than resolved inferentially. The coded variables were subsequently used in descriptive pattern analysis.

VP design content was coded by the first author using a predefined coding framework. To assess reliability, a second trained coder independently coded a random subsample comprising 20% of the dataset. Intercoder reliability was substantial (Cohen κ=0.76), indicating good reproducibility of the coding approach.

Research Design

This study used a multiphase research design to systematically examine the relationships between virtual content formats (images and videos) and audience engagement patterns on Instagram. The design integrates quantitative engagement metrics with qualitative profile and content evaluations to better understand how design characteristics relate to user interaction with human-like VPs. Figure 1 provides an overview of the research process, which includes the sequential stages of data collection, engagement and reach analysis, performance visualization, content design analysis, and results synthesis.

XSL•FO
RenderX

Figure 1. Research design flowchart.

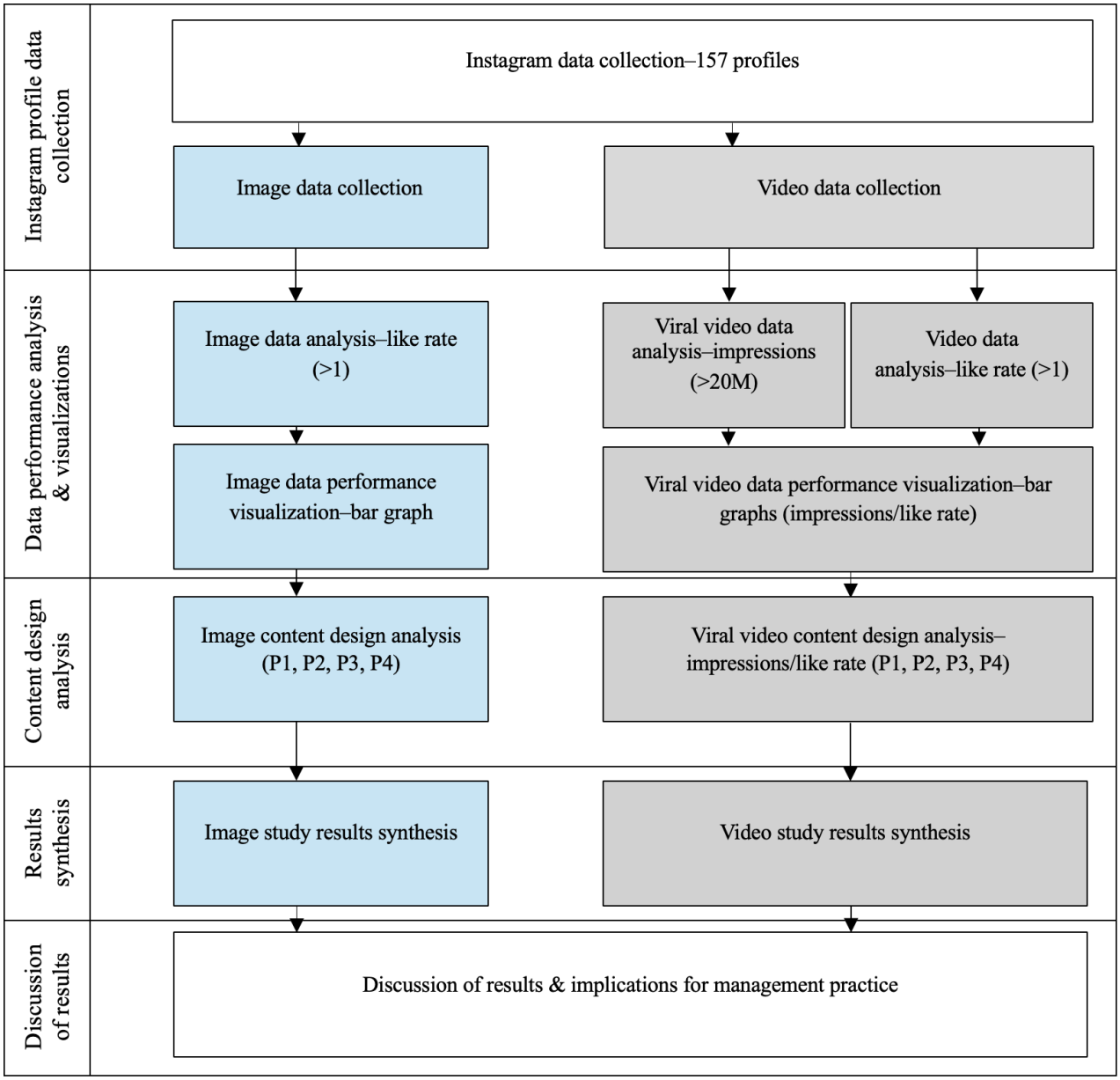

Following the data collection of 157 human-like female VPs, the study proceeded to the Engagement and Reach Analysis phase, which was divided into Image Analysis and Video Analysis components.

In the Image Analysis, engagement performance was assessed using like rate only, because Instagram does not provide observable impressions counts for image posts in the dataset. Like rate was treated as a limited proxy for passive interaction intensity rather than as a direct measure of relational bonding. Profiles were prioritized for closer review when at least 1 image exceeded a like rate of 1.0, indicating that the post generated more likes than the account's total follower base and therefore represented an extreme interaction outcome. For these cases, the analysis compared recurring profile characteristics, most notably CGI-identifiable versus photorealistic styling, the degree of character and behavioral consistency across visual content and captions, and the frequency of copresence in the highest performing image content.

The Video Analysis retained like rate as a relative interaction metric and additionally examined high-impression distribution events using Instagram's video impression data. Rather than relying on follower-normalized reach, the analysis concentrated on videos surpassing 20 million impressions and then profiled the accounts and posts responsible for these events. Because Instagram strategically prioritized short-form video (ie, Reels) during the study period, these reach outcomes should not be interpreted as reflecting only intrinsic content superiority; they likely also reflect platform-level distribution bias favoring Reels. Accordingly, the video analyses distinguish between interaction-based performance (like rate) and distribution-based performance (impressions).

The 20 million impression threshold was selected as a dataset-specific marker of extreme reach within the sampled population of female human-like VPs. The purpose of this threshold was to identify a subset of genuinely overperforming posts that was still sufficiently small to permit meaningful

XSL•FO
RenderX

pattern-based comparison. Lower cutoffs, such as 10 million impressions, captured a substantially larger share of the sample and therefore reduced the discriminative value of the category. Accordingly, the 20 million threshold was used as a pragmatic analytic criterion to isolate rare, high-salience distribution events within this specific VP niche, rather than as a universal benchmark of Instagram virality.

To limit sensitivity to one-off anomalies (eg, public relations–driven boosts or amplification through exceptionally large nonvirtual cocreators), each account's performance was analyzed using (1) its single best-performing image and video post, and (2) the mean of the 3 best posts per format for like rate and impressions.

Following performance identification and visualization, the content design analysis qualitatively assessed the design features of the most extreme-performing posts, with emphasis on identity coherence, CGI-likeness versus photorealistic ambiguity, human-VP copresence, and recurring patterns involving dark hair, dark eyes, and exaggerated female body types.

Finally, in the results synthesis and discussion, evidence from image engagement patterns, video engagement, and high-reach events, and the content design review was integrated to address the study's exploratory propositions and to derive empirically grounded implications for how human-like female VPs achieve engagement and reach on Instagram.

Ethical Considerations

Data collection was limited to publicly available profile- and postlevel information relevant to the research questions and was conducted in accordance with platform accessibility at the time of collection. One major ethical consideration was the responsible use of public information and the avoidance of unnecessary republication of identifiable account content beyond what was required for scientific analysis. Data collection was limited to information pertinent to the research objectives and was performed in strict accordance with Instagram's terms of service and public data policies. Furthermore, all information used for this study is freely available in the public domain; did not involve human participants, their tissues, and/or data; does not involve methods or subject matter that is sensitive; and is deemed not human subjects research [27]. Moreover, visual evidence is included strictly as analytical data. Because this study evaluates complex, nonstandard visual traits, such as explicit CGI-likeness, photorealistic ambiguity, and exaggerated sexual dimorphism, including representative visual examples is methodologically necessary to operationalize these variables for the reader and ensure the reproducibility of our content design analysis.

Participants and Procedure

The data collection process for this study was designed to identify public Instagram accounts representing human-like VPs for subsequent quantitative and qualitative analysis. For the purpose of this study, "human-like" was defined as exhibiting coherent human morphology (head, torso, and limbs) and facial features capable of recognizable human expression and gaze. This definition includes VPs with stylized aesthetics, such as anime-inspired avatars or exaggerated body proportions, provided that they retained a fundamentally human anatomical structure and human-like behavioral presentation. By contrast, abstract virtual characters—such as nonhumanoid objects, animals, or geometric entities lacking standard human biological features—were excluded.

During the initial screening stage, both human-like female VPs and human-like male VPs were identified. However, only 5 human-like male VPs met the eligibility criteria, which was too few to support meaningful subgroup comparison. To preserve analytical consistency and avoid unstable comparisons across highly imbalanced gender groups, only human-like female VPs were retained in the final dataset. Thus, the final analytic sample comprised 157 human-like female VPs, whereas the 5 human-like male VPs identified during screening were excluded from further analysis.

To reduce the likelihood of including heavily manipulated or commercially repurposed accounts with inflated audience metrics, only profiles with more than 100,000 followers were included in the final sample [28,29]. This threshold was selected because follower-purchase services have made it relatively easy to inflate lower-tier accounts, whereas manipulation of macro- and mega-level profiles is generally more detectable through content chronology gaps and third-party analytic indicators. Nevertheless, the possibility of residual manipulation cannot be excluded entirely [9,30,31].

The resulting dataset comprised 157 human-like female VPs. The restriction to macro- and mega-level accounts improved comparability within the final sample, but it also means that the study describes performance patterns among established incumbents rather than growth trajectories among emerging profiles.

Posts selected for detailed qualitative content review were identified using predefined quantitative thresholds applied to the full dataset. Specifically, posts were flagged for structured qualitative review when they met at least one of the following criteria: (1) image or video like rate greater than 1.0, indicating that the post received more likes than the account's total follower count; or (2) video impressions greater than 20 million, indicating an extreme distribution event within the sampled VP category. In this manuscript, "structured qualitative review" refers to manual assessment of coded visual and narrative variables using the predefined coding framework, including visual realism, identity consistency, copresence structure, appearance traits, body-type representation, and production technique where relevant. The process did not involve open-ended interpretive coding beyond these predefined variables.

Results

Overview

To contextualize the characteristics observed among the highest performing subsets, baseline descriptive statistics were compiled for the full coded sample of 157 human-like female VPs. The distribution of all coded profile-level variables, including visual realism, identity consistency, appearance patterns, and body-type representation, is provided in Table S2 in Multimedia Appendix

XSL•FO
RenderX

1. These full-sample descriptive statistics serve as the reference context for interpreting the traits observed among the highest performing profiles and posts.

Image Engagement Analysis (Like Rate)

The image engagement analysis identified 8 top-performing human-like female VPs that produced posts with like rates greater than 1.0, indicating unusually strong interaction intensity relative to audience size (Figure 2). By contrast, several legacy CGI-like profiles with very large follower bases performed strongly in terms of absolute like volume but less strongly relative to like rate terms, because interaction levels scale less than proportionally with follower count. This distinction separates relative engagement from scale-dependent absolute interaction and provides a descriptive basis for the subsequent content review of visual and narrative features commonly observed alongside unusually high like rate outcomes.

Figure 2. Image likes and engagement metrics. VP: virtual profile; IG: Instagram.

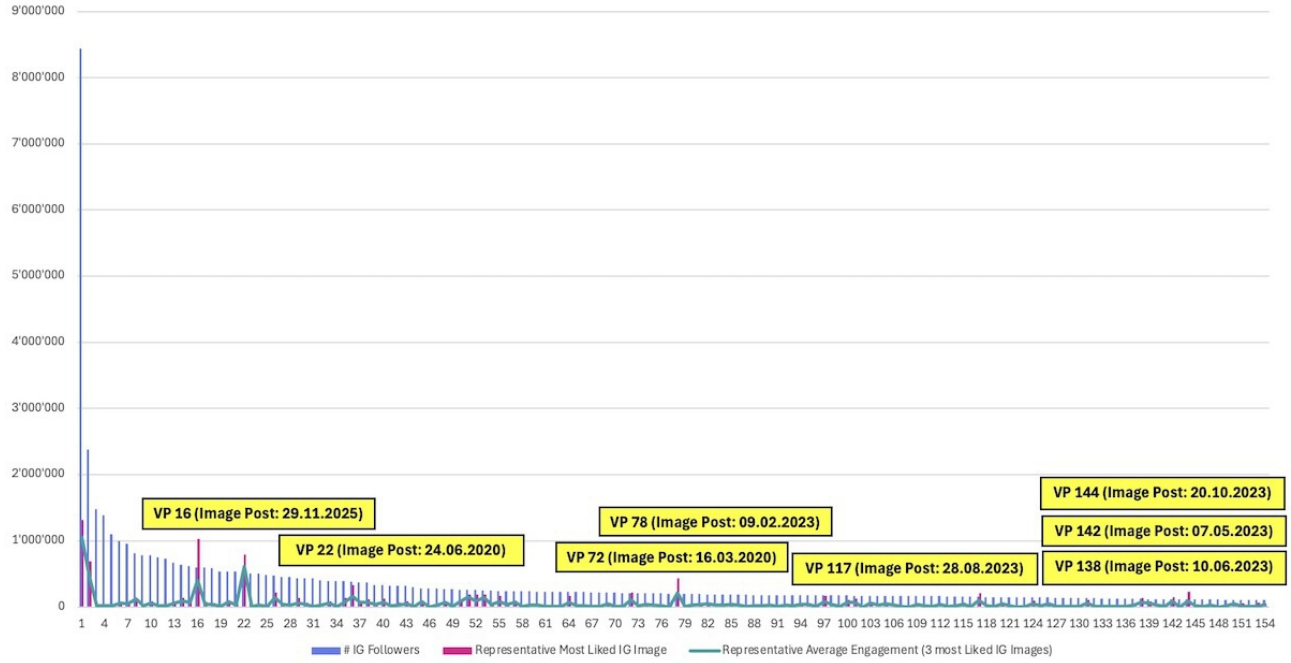

Image Content Design Analysis

The high like rate image posts examined in this analysis reveal several recurring visual and narrative patterns observed in posts with exceptional relative engagement (see Figure S1 in Multimedia Appendix 2 for visual evidence of these design variables). A prominent feature across the sample is copresence within the image frame: in half of the top-performing images, more than 1 subject is depicted, which represents a recurring pattern observed in posts with high relative engagement. Importantly, this pattern is not limited to the inclusion of real humans. While the informal car scene featuring the VP 16 alongside a prominent human celebrity athlete represents a particularly salient example of celebrity copresence appearing alongside higher engagement, it is not the only pattern associated with high performance.

Rather, the broader pattern suggests that multiple subjects—whether human-VP or VP-VP copresence—were frequently observed in posts with higher interaction intensity, possibly through increased perceived social density and enhancing narrative richness. In the case of CGI-like VP 72, relative engagement may also have been related to platform-level social signaling of mega VP 2 who is both visibly featured in the image post and liked the post too, which in combination was observed alongside higher relative engagement. This distinction underscores that social validation mechanisms operate both within and outside the visual frame.

Across the image set, stylistic and character consistency emerges as a common attribute, particularly among CGI-like profiles. Because such characters are typically maintained as stable 3D models or tightly controlled generative assets on creators' servers, they allow for high visual continuity across VP content. This technical consistency may contribute to recognizability and parasocial familiarity. Topical consistency in captions was also prevalent in these posts, as illustrated by VP 22, where narrative alignment between visual content and text co-occurred with stronger audience resonance and engagement. The CGI-identifiable VP 22 also achieved high like rate without relying on any character copresence, indicating that explicit artificiality does not preclude strong performance.

From an aesthetic standpoint, most high-performing images depict VPs with dark hair and dark eyes, with 6 of the 8 highlighted profiles fitting this pattern and the remaining 2 featuring blonde characters. Compared with the baseline frequency of 62.42% in the full sample (Table S2 in Multimedia Appendix 1), this represents a descriptive concentration of dark-featured VPs among the highest performing image posts. While causal inference is not possible from this descriptive analysis alone, this concentration may reflect prevailing audience taste structures within the dominant follower demographics. In addition, many of the images emphasize exaggerated female body features, particularly pronounced bust and waist-hip contrast, as exemplified by VP 142 (Figure S1 in Multimedia Appendix 2). Such visual exaggeration, bordering

XSL•FO
RenderX

on pornographic stylization, may reflect content designed for a predominantly male audience, a dynamic that is further illustrated by the VP 138 post, which situates 2 stylized human-like female VPs in a construction-site setting wearing safety vests and helmets. The juxtaposition of sexualized female forms with traditionally male-coded occupational contexts may attract attention through contrast and stereotype activation.

The image-level findings indicate that higher relative engagement was more commonly observed alongside a combination of physical and behavioral consistency, thematic coherence, and audience-targeted visual framing. Copresence, whether enacted visually or reinforced through platform-level interactions, emerges as a particularly robust pattern, suggesting that perceived social density and narrative context were frequently observed in posts with high relative engagement on Instagram.

Viral Video Reach Analysis (Impressions)

The analysis of viral video reach indicated a strong concentration of extreme virality within a limited subset of human-like VPs (Figure 3). Videos exceeding 20 million impressions were produced exclusively by a relatively small number of established accounts, underscoring a pronounced scale effect in which many of the most viral videos originated from profiles with the largest follower bases. At the same time, a very small number of lower-follower outliers also exceeded the 20 million impression threshold. These exceptions suggest that extreme video reach was not exclusively limited to the largest accounts, although some cases appeared likely to involve external amplification mechanisms such as a brand collaboration.

Figure 3. Video impressions and reach metrics. VP: virtual profile; IG: Instagram.

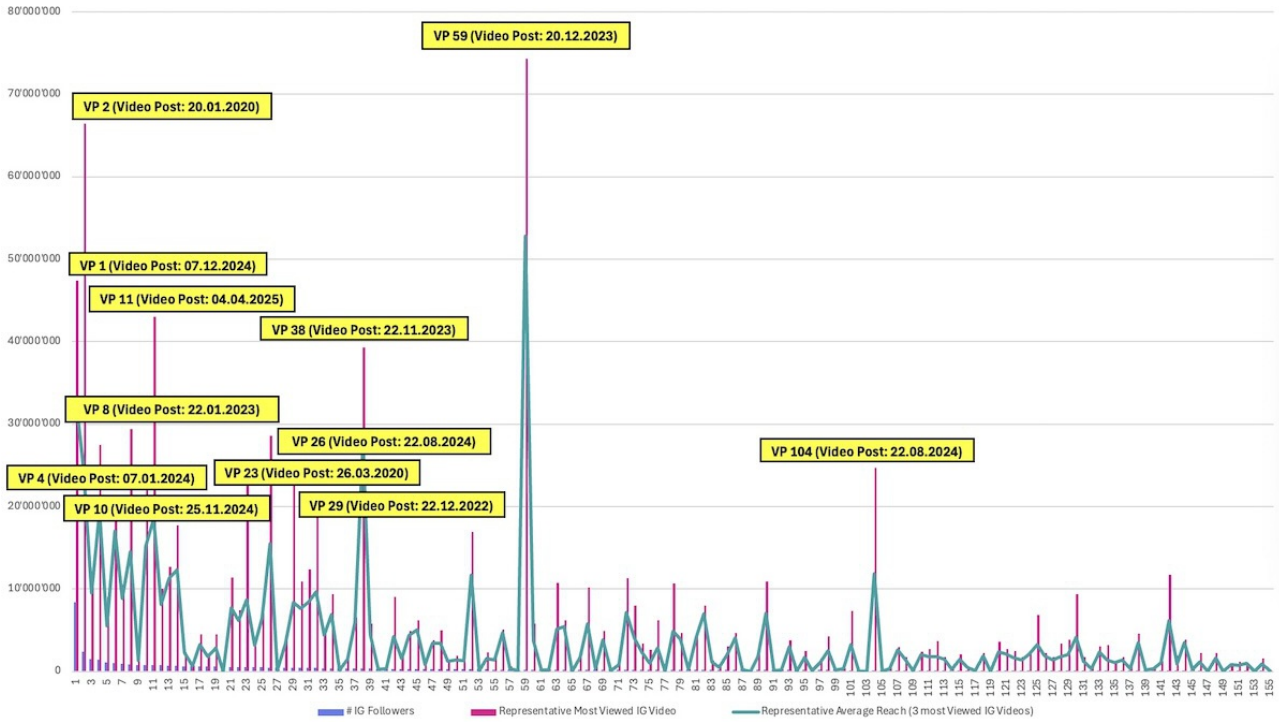

Viral Video Content Design Analysis (Impressions)

Analysis of viral video content exceeding 20 million impressions reveals a clear descriptive pattern observed among posts with extreme reach performance (see Figure S2 in Multimedia Appendix 2 for visual evidence of these design variables). Most notably, CGI-like VPs dominate the highest performing videos, with 7 of the 12 most viral videos produced by CGI-like VPs. This qualified majority indicates that explicitly legible CGI aesthetics were commonly observed among exceptionally viral videos. This prevalence of CGI aesthetics among viral videos (58%) is notably higher than the baseline of 8.92% observed in the full sample, suggesting that explicitly virtual identities were more common among extreme reach outcomes. In contrast to photorealistic VPs, whose motion-based content may expose perceptual inconsistencies, CGI-like profiles exhibit structurally stable identity representation, which was frequently observed in video contexts where frame-to-frame coherence is critical.

As discussed in the image engagement analysis, CGI-like VPs inherently exhibit a high degree of visual and behavioral consistency, as they are typically maintained as persistent 3D assets or tightly controlled generative models. This consistency may help describe their prominence among high-performing videos by reinforcing recognizability and thus perceived authenticity and may coincide with algorithmic amplification and sustained viewer attention. The dominance of CGI-like profiles among the most viral videos suggests that, for reach-driven outcomes, clarity and stability of artificial identity were more commonly observed than incremental photorealistic gains among these high-reach videos.

Within the viral video set, there is also evidence consistent with a positive descriptive role for VP-human copresence, consistent with prior findings [32]. One particularly salient example is the video by VP 38, which achieved nearly 40 million impressions and features the VP alongside a real human. Importantly, VP 38's appearance remains unmistakably CGI-like; rather than

XSL•FO
RenderX

diminishing performance, this explicit artificiality appears alongside high audience reach. The juxtaposition of a CGI-like character with a real person may be perceived as reinforcing authenticity, transparency, and originality, aligning with prior observation that human copresence has been linked to higher engagement when the virtual nature of the VP is not ambiguous [32].

Beyond copresence, the broader visual patterns observed in viral videos largely mirror those identified in the image content design analysis. Characters with dark hair and dark eyes appeared repeatedly among the top-performing posts examined in this subset. In addition, physically feminine body types, often emphasizing large busts and wide hip-waist ratios, as well as female stereotype activating contexts, were also commonly observed in this subset. These descriptive patterns may reflect dominant audience preferences.

Ultimately, the case of VP 104 represents a notable and theoretically informative outlier. Despite a comparatively small follower base, this profile achieved exceptional video reach, positioning it alongside much larger accounts. Like the CGI-like profiles, VP 104 is unambiguously virtual, but her morbidly obese body type departs sharply from conventional influencer aesthetics. This deviation may have co-occurred with high impression counts through a combination of radical transparency, visual novelty, and perceived authenticity. It may also reflect spectacle- or ridicule-based entertainment dynamics associated with exaggerated motion and bodily contrast. Regardless of the underlying mechanism, this case is consistent with a central observation: in contemporary Instagram video environments, explicit virtuality, rather than photorealistic ambiguity, was frequently observed as a characteristic of extreme video reach.

Video Engagement Analysis (Like Rate)

The video engagement analysis, which was based on relative like rate rather than absolute reach, identified a partially overlapping but clearly distinct group of high-performing VPs (Figure 4). Compared with the viral video reach analysis, only a subset of accounts appeared in both categories, indicating that distribution-based success and interaction-based success followed different descriptive patterns within the sample. This divergence is consistent with the distinction between reach and engagement: absolute impressions heavily represent profiles with large existing audiences and algorithmic momentum, whereas like rate normalizes interaction by follower count and therefore reduces the scale-dependent advantage.

Figure 4. Video likes and engagement metrics. VP: virtual profile; IG: Instagram.

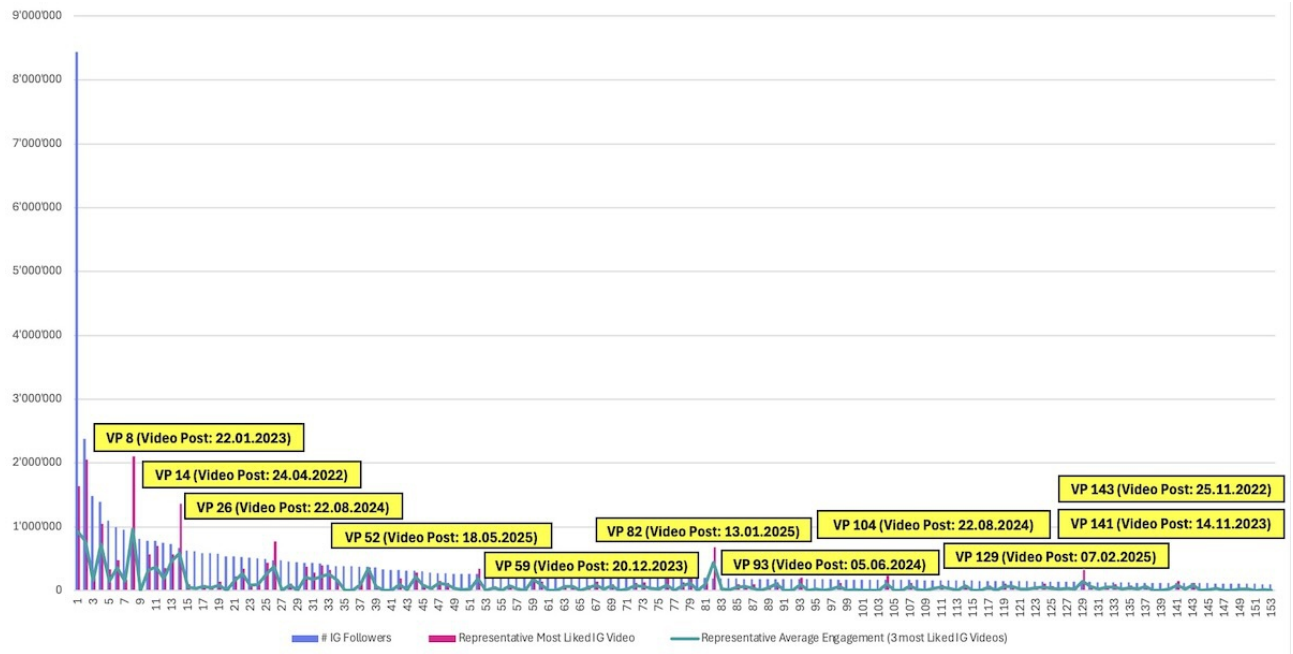

Video Content Design Analysis (Like Rate)

In contrast to the viral video reach analysis, the video engagement analysis reveals a markedly different pattern with regard to top-performing content design (see Figure S3 in Multimedia Appendix 2 for visual evidence of these production techniques). Only a small subset of CGI-like VPs, VP 59 and VP 82, appeared among the posts with the highest interaction intensity, indicating that CGI-like VPs were more commonly observed in high-reach distribution events than in high-engagement interactions. This divergence indicates that CGI-like profiles were commonly observed in large-scale exposure events, whereas their ability to convert visibility into active engagement is more limited.

Instead, the video engagement results are dominated by photorealistic VPs, including VP 8, VP 14, VP 26, VP 52, VP 93, VP 104, VP 129, VP 141, and VP 143. While photorealistic profiles constituted 91.08% (143/157) of the total sample, they represented only a minority (5 out of 12) of top-performing videos by reach rate but a majority (9 out of 11) of top-performing videos by like rate, indicating a descriptive shift toward realism for interaction-based success. These profiles are more commonly observed among the top-performing videos in terms of like rate and predominantly rely on generative AI video

XSL•FO
RenderX

techniques or face-swapping software, such as the HelloFace app, to map highly realistic virtual faces onto human bodies. This approach may be perceived as increasing realism, emotional expressiveness, and micro-level social cues, and these features were commonly observed in videos with high active user engagement. Notably, this finding represents the inverse of the reach-based results, where CGI-like profiles were dominant, underscoring that reach and engagement correspond to distinct design mechanisms.

An additional and noteworthy pattern concerns the strong engagement performance of VP 52, VP 104, and VP 141, all of which embody morbidly obese virtual personas. These profiles were observed among posts with exceptional like rates despite diverging sharply from conventional virtual beauty standards. The reasons underlying this outcome remain ambiguous and may involve perceived authenticity, parody, or even elements of ridicule-driven entertainment; however, the data do not allow for a definitive attribution.

Across these high-performing videos, several shared visual and behavioral characteristics emerge. The content typically zooms in on the VP's face and upper body, allowing for clear display of facial expressions, eye contact, and subtle emotional cues. These nonverbal behaviors are known to facilitate perceived social proximity, parasocial bonding, and active response [33], which aligns with our observation of these features in high-engagement videos. Consistent with both image engagement and viral video reach analyses, the high-performing video engagement content largely features VPs with dark hair and dark eyes. The primary exceptions are the blonde profile of VP 26, the stylized Asian avatar with pink-dyed hair and bright contact lenses of VP 82, and the light-eyed yet dark-haired profile of VP 14. These results reinforce the notion that certain visual traits recur across performance dimensions.

Discussion

Principal Findings

Overview

This study evaluated exploratory propositions concerning reach and engagement among human-like VPs on Instagram and identified several recurring descriptive patterns across image and video content. Overall, the findings indicate that identity and behavioral consistency were commonly observed among high-performing content across formats. Photorealism alone did not emerge as a uniform marker of stronger performance; rather, the observed patterns aligned closely with transparent virtual presentation, identity coherence, and format-specific content strategies. Human-VP and VP-VP copresence both appeared among high-performing posts, although their role varied by format and should be interpreted descriptively. Moreover, the study findings should be interpreted as descriptive associations within the sampled dataset rather than as statistically confirmed differences or causal effects.

Image Content Engagement Evaluation

The image engagement analysis indicates that high like rate was commonly observed alongside copresence within the image frame, strong identity and narrative consistency, and audience-targeted visual framing. CGI-like profiles also appeared among high-performing images, including in cases without copresence. Across profiles, dark hair and eyes and exaggerated feminine stylization recurred repeatedly, indicating a descriptive concentration of these traits among high-performing image posts.

From a theoretical perspective, the findings align with social identity theory, whereby contextual, physical, and behavioral congruence between VPs and their audiences fosters identification and self-affirmation. Prior research on parasocial relationships and digital intimacy suggests that sustained, 1-sided engagement with idealized figures may contribute to distorted relationship expectations, emotional overidentification, and social comparison effects [33,34]. The frequent emphasis on exaggerated body features may further intensify appearance pressures.

Identity coherence emerges as a foundational characteristic, particularly for CGI-like VPs. The use of stable 3D models or tightly controlled generative assets supports high visual and behavioral consistency, which may contribute to recognizability and parasocial familiarity over time. When captions aligned topically with imagery, stronger audience resonance was descriptively observed, indicating that narrative continuity across modalities was observed alongside stronger audience resonance. Importantly, CGI-like profiles that retain sufficient human resemblance to avoid "uncanny valley" discomfort were observed among high-performing posts [35], suggesting that stronger engagement was more commonly observed with a balance of authenticity and transparency rather than maximal photorealism alone.

Copresence operates both directly (image post human-VP or VP-VP cofeaturing) and indirectly through platform-level social signaling activity. These patterns show that higher relative image engagement was commonly observed alongside social-contextual cues that reflect high perceived social density and narrative consistency. Copresence increases interpretive depth and invites social inference, which may coincide with interaction. Platform-level validation (eg, likes from prominent accounts) further reinforces this pattern by signaling relevance and credibility beyond the visual content itself.

Overall, image engagement was more commonly observed alongside the interplay of consistent character and contextual consistency, copresence, and audience-aligned aesthetics. These insights clarify how relative engagement manifests on Instagram while underscoring the ethical importance of transparency and responsible VP design to mitigate potential risks to digital well-being.

Viral Video Content Reach Evaluation

The reach-based analysis of viral video content indicates a pronounced concentration of extreme absolute reach outcomes among CGI-like VPs. Seven of the 12 most viral videos stem from CGI-like VPs, suggesting that clear CGI aesthetics and the associated stability of artificial identity were commonly observed in this subset among the high-performing viral videos. Furthermore, across the viral videos, recurring appearance cues,

XSL•FO
RenderX

including dark hair and dark eyes, and highly feminized body presentations are also frequently observed, broadly aligning with the visual regularities identified in the high-engagement image content analysis.

The dominance of CGI-like VPs in viral video reach outcomes is theoretically consistent with the idea that clarity of artificial identity was frequently observed in video contexts. Motion content increases perceptual demand: viewers continuously evaluate facial dynamics, eye behavior, and body physics, meaning that near-photorealistic VPs may be penalized if small inconsistencies accumulate across frames. CGI-like VPs, by contrast, implicitly set a different expectation baseline and often exhibit higher frame-to-frame coherence because they are maintained as persistent 3D assets or tightly controlled digital low-rank adaptation models. This structural consistency may contribute to recognizability and lower perceptual friction, which in turn may coincide with greater watch time and reexposure, signals that frequently co-occur with algorithmic amplification and were frequently observed among viral video content on Instagram.

Finally, within this set, very high reach was also observed in selective cases featuring VP-human copresence, with VP 38 reaching nearly 40 million impressions in a hybrid scene, consistent with the observation by Cascio et al [32] that human copresence can strengthen VP content performance when virtuality remains unambiguous. However, the present findings suggest an important boundary condition: copresence appears most effective when it reinforces rather than obscures the VP's synthetic ontology. Meaning, hybrid compositions may be especially potent when the contrast between human and virtual increases perceived transparency, authenticity, and narrative tension, and may co-occur with stronger audience attention and algorithmic distribution. In summary, the viral video content reach analysis suggests that extreme video virality among human-like VPs was commonly observed alongside explicit virtual legibility, identity stability, and high-salience design cues, and that the "uncanny valley" concept may contribute to reach both by being avoided (through clear CGI stylization) and, in some cases, by being exploited as a mechanism of discomfort-driven reach.

Video Content Engagement Evaluation

Video content overall showed higher interaction intensity than static images, with 11 video posts exceeding a like rate of 1.0 compared with 8 image posts, underscoring the heightened persuasive power of motion-based content [33,34]. Contrary to the viral video reach analysis, the video engagement analysis reveals a distinct performance structure. Unlike viral video reach, high video engagement is not primarily driven by CGI-like profiles. Only a small subset of CGI-like VPs, VP 59 and VP 82, appears among the top performers in terms of like rate. Instead, the majority of high-engagement videos are produced by photorealistic VPs using advanced generative AI or face-swapping technologies, indicating that engagement and reach were characterized by different design patterns.

The divergence between reach and engagement outcomes suggests that algorithmic exposure and active user endorsement may reflect different perceptual pathways and platform dynamics. CGI-like profiles, which dominate extreme reach events, appear less commonly observed in converting visibility into proportional engagement. This pattern suggests that their virality may coincide with attentional capture, potentially through novelty, salience, or discomfort, rather than affective affinity. In line with the viral video reach evaluation, this is compatible with the post hoc interpretation that CGI-like virality may, in some cases, reflect attention capture associated with perceptual incongruity, whereby perceptual incongruity sustains attention and impressions without consistently translating into positive endorsement [35].

By contrast, video content engagement is clearly dominated by photorealistic VPs. Seven of the 11 top-performing videos with regard to like rate exhibit full photorealism consistent with face-swapping and advanced generative AI technologies, suggesting that such techniques were more commonly observed than more complex CGI tools among posts with high interaction. The relative visual homogeneity of these videos suggests that audiences may respond particularly strongly to realistic motion, facial expressiveness, and microemotional cues, which are known to facilitate parasocial bonding and active response [33].

A theoretically informative exception is VP 104, which performs exceptionally well in terms of both reach and engagement, unlike most CGI-like VPs that display a clear asymmetry between these metrics. Although photorealistic, the VP is unmistakably artificial and departs sharply from conventional influencer aesthetics through a morbidly obese body type. This combination of clear virtuality and extreme body type appears to mitigate the negative effects typically associated with the "uncanny valley," suggesting that perceptual discomfort does not necessarily suppress either attention or interaction in this case. Instead, VP 104—together with the thematically aligned profiles of VP 52 and VP 141—may benefit from radical transparency, novelty, and emotional arousal, co-occurring with both high visibility and active user response.

Across high-engagement videos, recurring design features further clarify these dynamics. Content typically centers on the VP's face and upper body, enabling direct gaze, visible facial expressions, and subtle emotional signaling, nonverbal cues commonly discussed in relation to perceived social proximity and parasocial interaction [36]. Notably, only one of the high-engagement videos features VP-VP copresence; instead, interaction intensity was often observed in solo-performer formats, where audience attention may be concentrated on a single virtual persona. This contrasts with the static image-based findings, where copresence played a more prominent role, and further underscores that image and video are characterized by different content design patterns. Consistent with image engagement and video reach analyses, VPs with dark hair and dark eyes appeared frequently in the high-performing subset, with limited stylistic exceptions (eg, the blonde VP 26, pink-dyed VP 82, and the bright-eyed VP 14), indicating that certain visual traits retain cross-format relevance.

Taken together, these findings indicate that video is the most effective medium for generating engagement, but that success depends critically on how virtuality is expressed. CGI-like content appears optimized for reach—potentially through reverse

XSL•FO
RenderX

"uncanny valley" dynamics—whereas photorealistic, face-swapped videos are more commonly observed in content that paired attention with interaction. The case of VP 104 demonstrates that when artificiality is both explicit and narratively coherent, the "uncanny valley" need not constrain performance, highlighting important boundary conditions for theories of virtual embodiment and audience response.

Additional Findings

Beyond the formal research propositions, the study yielded several additional findings that further illuminate how human-like VPs generate engagement and reach across image and video content formats. A central observation is that image and video performance feature fundamentally different VP types and design cues, indicating that success was not observed uniformly across media. The characteristics commonly observed in high-performing video content differed from those commonly observed in high-performing image content, reinforcing the importance of distinguishing algorithmic exposure from active user response.

Although no single configuration of physical or ethnic attributes emerged as a universal descriptive marker of success, a clear concentration of VPs with dark hair and dark eyes was observed among top-performing profiles across both image and video formats. We hypothesize that this pattern may correspond to the physical traits of major segments of the global population, particularly within Asian, African, Middle Eastern, and Latin American regions, which together constitute the numerical majority of social media users [34]. From a social identity perspective, audiences might therefore be more likely to engage with VPs that visually resemble their own or reflect familiar, normative appearance cues [18]. However, as this study did not use specific audience demographic data, we cannot confirm that this design trend results from audience self-identification. Consequently, this interpretation remains a theoretical possibility for future research rather than a conclusion supported by the current data.

Importantly, the findings indicate that there is no homogeneous blueprint for a universally high-performing human-like VP. While certain trends recur, the remaining design attributes among top-performing profiles are highly heterogeneous. This diversity suggests that engagement is contingent on contextual coherence and audience fit, rather than strict adherence to a fixed aesthetic formula. Consistent with this, no single VP appeared among both the top-performing image and video posts, pointing toward content format specialization rather than cross-format dominance.

Another noteworthy outcome is that some highly viral video posts were produced by VPs with average or below-average overall following, including VP 59 and VP 104. This challenges the misconception that VPs with only above-average-follower counts can generate viral content. Instead, it suggests that content-specific relevance, emotional salience, and algorithmic amplification may coincide with episodic virality beyond baseline profile performance.

From a digital health and behavioral science perspective, these dynamics highlight the emergence of nonlinear influence patterns in AI-mediated environments. The fact that below-average-follower VPs can intermittently generate extreme reach or engagement underscores how affective design cues and platform algorithms jointly shape exposure and psychological outcomes. Such mechanisms raise concerns related to the attention economy, emotional contagion, and the amplification of content that may provoke strong reactions without delivering informational or social value—dynamics previously linked to parasocial overidentification and mental well-being risks [36].

The linguistic analysis further revealed a strong dominance of English across all content formats, which may reflect algorithmic prioritization rather than pure audience preference. Whether this linguistic dominance aligns with user needs, particularly in health communication contexts, remains an open question.

Finally, the observation that top-performing videos were generally created more recently than top-performing images highlights the rapid evolution of generative video technologies and the historical lag that previously constrained AI video production. As these tools continue to mature, their persuasive capacity, and associated risks, will expand accordingly. Prior research has shown that realistic motion and expressive facial behavior can significantly enhance engagement [33], but such psychological potency also introduces ethical challenges related to emotional manipulation, misinformation susceptibility, and addictive engagement loops [37,38].

Taken together, these additional findings reinforce the need for cross-sector collaboration between technologists, health communication researchers, and policymakers. As human-like VPs become increasingly lifelike and socially influential, future research must move beyond surface-level engagement metrics to assess psychological, social, and ethical implications. Improving transparency in content provenance, strengthening verification protocols, and establishing responsible design standards will be critical to ensuring that AI-generated personas support digital engagement without undermining user well-being, trust, or public health outcomes.

Implications of Findings

This study offers substantial insights not only to traditional modeling agencies and human influencers but also to broader fields concerned with digital communication, health promotion, and the ethical design of AI-generated personas. The data underscore the importance of transparent, authentic, and narratively consistent content in relation to engagement outcomes, suggesting that creating coherent and identifiably artificial human-like VPs may help inform how commercial and public organizations navigate the ongoing digital transformation. In health-related contexts, similar approaches could be used to design perceived trustworthy and authentic virtual health educators or support agents that may capture attention and sustain engagement while maintaining transparency about their artificial nature.

Using VPs as a testing ground for audience response also offers opportunities beyond marketing. Agencies, and by extension, health communication practitioners, can identify which visual, behavioral, and linguistic attributes elicit trust, empathy, or

XSL•FO
RenderX

attention, providing valuable insights for developing more effective virtual counselors, digital therapy companions, and patient education interfaces. This mirrors evidence from human-computer interaction studies showing that perceived authenticity and emotional resonance strongly influence both adherence and behavior change [39].

For digital creators and VI agencies, the findings highlight that VPs combining consistent physical, behavioral, and narrative coherence are commonly observed among profiles showing sustained engagement. This can guide creators in refining design parameters that may be related to relatability, cultural alignment, and emotional comfort, all of which are equally relevant to the creation of AI health agents intended for diverse populations. The results also emphasize the potential of locally oriented VPs, those embedded within a specific cultural or linguistic context, to generate stronger audience trust through cultural congruity [40,41]. Virtual modeling and influencing agencies can leverage these principles to improve both marketability and ethical integrity, ensuring that virtual entities reflect diverse, inclusive, and psychologically safe representations.

Finally, the superior engagement performance of video content, especially when generated through face-swapping technology, suggests that motion and emotional expressiveness were frequently observed alongside high user attention. While this presents clear opportunities for content creators and educators, it also introduces psychological and ethical considerations. The same features that enhance engagement may heighten emotional persuasion or dependency. Consequently, developers of AI-driven media, including virtual health communicators, should integrate ethical safeguards, such as clear disclosure of synthetic identity, limits on emotional manipulation, and design standards that promote digital well-being.

Overall, these findings underscore the growing overlap between commercial and health communication ecosystems in an era of synthetic media. As virtual personas become more pervasive, interdisciplinary collaboration among designers, health professionals, and policymakers will be critical to ensure that the technologies driving engagement are harnessed responsibly to educate, empower, and protect users, rather than exploit their psychological vulnerabilities.

Comparison With Prior Work

The findings of this study broadly align with prior work on VIs and human-AI interactions while refining several points of interpretation. Rather than identifying visual realism itself as the primary feature of interest, the present results suggest that transparent virtual presentation, consistency of identity, and format-specific expressive cues were closely aligned with the observed higher performance. These conclusions remain descriptive and should be interpreted in relation to the present sample of established human-like female VPs.

Extant studies have reported that increasing human-likeness in VIs can, under certain conditions, reduce positive reactions and trigger uncanny valley effects. In contrast to interpretations that emphasize realism itself as problematic, our results show that ambiguity in virtual identity was the primary factor linked to apparent perceptual discomfort. Photorealistic VPs in this study did not exhibit reduced engagement when their artificial nature was clearly communicated, suggesting that uncanny valley effects are context-dependent and contingent on transparency rather than visual fidelity alone [5,11,35].

This finding suggests that concerns related to the uncanny valley may be highly context-sensitive: in social and influencer environments—where users anticipate artificial mediation—explicit virtual legibility can mitigate perceptual discomfort and frequently co-occurs with higher engagement [6,20]. From a health communication perspective, this insight has implications for designing trustworthy virtual health agents or AI health educators, where credibility is supported less by realism itself than by ethical framing, transparency, and behavioral consistency [14,15,27].

Authenticity and consistency also emerged as central factors prevalent among high-engagement content. Prior studies have identified authenticity as a core factor associated with VI success [7,8,19]. Our data extend this understanding by demonstrating that coherence across physical, ethnic, and behavioral cues was frequently observed among posts with stronger engagement outcomes [10,12]. This pattern aligns with social identity theory [18], which posits that individuals are more likely to connect with figures who reflect aspects of their own identity. The resemblance between high-performing VPs and the ethnic characteristics of their primary audiences exemplifies this process and translates it into measurable behavioral responses [16,20]. These results parallel findings in digital health research showing that cultural and linguistic alignment strengthens trust and adherence in AI-mediated patient interactions, underscoring the cross-domain relevance of identity congruence [24,25,39].

Our findings partially diverge, rather than fully contradict, prior expectations regarding human-virtual copresence. While earlier work suggested that human inclusion may enhance credibility and social realism [12,16], this study shows that human-VP copresence is not inherently superior to VP-VP copresence, and that both frequently co-occurred with strong engagement and reach when the VP's artificial nature remains clearly identifiable [20,32]. Notably, the most engaging image posts did not rely on human collaborators, suggesting that audiences may value virtual autonomy and narrative coherence over hybrid realism. In digital health contexts, this insight can inform the design of autonomous virtual counselors or therapeutic avatars that cultivate trust through identity stability and emotional presence rather than simulated human companionship [15,23,27].

Finally, the results reinforce earlier evidence of the growing dominance of video content [33]. Engagement with video posts far exceeded that of image posts, particularly when videos were generated using AI-driven face-swapping tools that produce emotionally legible motion rather than maximal photorealism [4,6,32]. This distinction aligns with the study's broader finding that expressive motion, microemotional cues, and behavioral realism were more consistently observed among top-performing content than visual realism alone. While this trend holds promise for enhancing health communication reach and interactivity, it also underscores the need for ethical guidelines governing the persuasive potential of emotionally powerful AI media—especially in domains where user vulnerability,

XSL•FO
RenderX

misinformation, or emotional influence may carry public health implications [9,27,37].

Limitations

A major limitation of this research lies in the potential algorithmic bias introduced by the usage of the authors' personal Instagram account during the data generation phase. Although significant efforts were made to ensure the consideration of a diverse range of profiles during data collection, it must be acknowledged that Instagram's algorithm may have skewed the selection process by potentially prioritizing profiles that relate to the authors' previous browsing activity. To mitigate this limitation in future research, it is advised to create novel unbiased Instagram accounts used for the purpose of research only. This would enable the collection of a more potent and representative dataset.

A further limitation concerns the operationalization of engagement solely through like rate. Likes reflect low-effort, passive approval and do not capture more active forms of relational response such as commenting, sharing, or sustained conversational interaction. In addition, like rate normalization may disadvantage legacy accounts with large numbers and inactive followers relative to newer accounts whose follower base is more active. Accordingly, the engagement findings should be interpreted as patterns in passive interaction intensity rather than as comprehensive measures of audience bonding or trust.

Furthermore, the exclusion of profiles with fewer than 100,000 followers introduces a survivor bias. By limiting the analysis to macro- and megainfluencers to avoid manipulated accounts, the study effectively examines the characteristics of established success rather than the factors contributing to growth. Traits observed among high-performing macro- and mega-level accounts may not generalize to early-stage profile development.

Additionally, the comparative analysis of content formats must account for platform-level distribution biases. During the data collection period, Instagram's algorithm heavily prioritized short-form video (ie, Reels) to compete with emerging market competitors, resulting in systemic amplification of video content over static images. Consequently, the superior reach and engagement observed for video posts may partially reflect this strategic platform prioritization rather than purely intrinsic superiority in content design or user preference. In addition, the comparison between image and video performance must be interpreted in light of platform-level distribution asymmetry. During the data collection period, Instagram's prioritization of Reels likely increased the exposure potential of video content independently of content design. The observed superiority of video on reach metrics therefore reflects both platform incentives and content characteristics.

An inherent challenge in studying VPs is their vulnerability to hacking, manipulation, or platform removal due to alleged content policy violations. This issue was exemplified by the cases of VP 5 and VP 6, whose accounts were deactivated and/or deleted during the research analysis phase. This uncertainty complicates the reliable scientific study of VPs, as there is a permanent risk of midstudy profile deletion. Fortunately, the study had included multiple alternative VPs, thus mitigating potential cluster risk, and so the analysis could be continued with immediate effect.

Moreover, the study does not directly measure audience perception of artificiality or realism in AI-generated content. As a result, it cannot empirically determine the extent to which users correctly identify photorealistic AI-generated images versus videos as synthetic, nor how such perceptions influence reach and engagement outcomes. This limitation is particularly relevant for interpreting the divergence observed between viral video reach and engagement, as well as the strong reach performance of CGI-like profiles. Without direct insight into audience-level perception, it remains unclear to what extent "uncanny valley" effects, perceptual discomfort, or ambiguity-driven attention contribute to the observed performance discrepancies.

Consequently, we offer a post hoc interpretation regarding the application of the "uncanny valley" theory [35]. While this construct typically predicts negative affect and user withdrawal, our data suggest a potential divergence between attention and affinity: the visual dissonance associated with perceived artificiality may concurrently trigger an orienting response driven by curiosity. This mechanism could elucidate why CGI-like profiles achieved high algorithmic visibility (reach) despite lower relative emotional resonance (engagement). We acknowledge this as a retrospective observation; the "uncanny valley" effect may be bidirectional in this context, where perceptual discomfort is associated with attentional capture rather than strict avoidance.

This interpretation also provides context for outliers such as VP 104. Although not CGI-like in the classical sense, the profile represents a significant deviation from conventional influencer aesthetics. Such content may achieve high reach through radical novelty or spectacle-based consumption, where viewership is motivated by expectancy violations. This suggests that for extreme reach outcomes, platform algorithms may preferentially amplify content that elicits high-arousal reactions, including potential aversion, due to their relation to retention and repeated exposure. Future research is required to experimentally isolate "attention-driven discomfort" as a distinct variable from positive narrative interest.

Finally, the speed of technological progress, combined with the fast-paced nature of the digital advertising and marketing landscape, presents a challenge for the longevity and relevance of this study's findings. As the industry and technology advance, there is a risk that these findings may become outdated more quickly than those in other areas of management research. However, it is worth emphasizing that many of the insights presented in this study are grounded in well-established social and organizational theories, which may help sustain their relevance even as the surrounding industry continues to evolve.

Future Research

Future research should explicitly examine user recognition, affective responses, and interpretive judgments of AI-generated content across static and motion formats. Experimental or survey-based methods are particularly well suited to clarify

XSL•FO
RenderX

whether attention toward virtual content is driven primarily by attraction, repulsion, or misattribution, especially in cases where reach and engagement diverge. Such work would substantially strengthen causal inference and improve understanding of how virtual content design shapes both algorithmic amplification and human response. This line of inquiry is especially relevant given this study's findings on CGI-like virality, where extreme reach may be associated with discomfort or novelty-based attention, rather than positive affect.

Conducting in-depth qualitative research, including audience surveys and interviews, could validate and enrich the findings of this study. Qualitative data would provide a more nuanced understanding of why certain physical attributes (such as dark hair and brown eyes within VPs) closely resonate with majority audiences, and how these aesthetic and cultural preferences might evolve over time. Rather than implying a deterministic advantage, such work could help disentangle whether these traits reflect audience self-identification, algorithmic reinforcement, or generative AI training biases. In addition to advancing marketing knowledge, such research could shed light on how exposure to idealized virtual appearances shapes user attitudes, well-being, and perceptions of authenticity.

Another area for improvement involves expanding the analysis to a broader spectrum of VP sizes, particularly micro and nano profiles. Examining smaller-scale VPs could yield valuable insights into how engagement rates vary across audience scales and the extent to which intimacy, familiarity, or social presence differs among less-followed profiles. Additionally, the creation and experimental testing of self-generated VPs on Instagram would allow researchers to systematically manipulate profile variables (such as explicit AI disclosure, identity consistency, degree of CGI stylization, or communication tone) and observe their direct effects on engagement, affect, and audience behavior.

The application of "uncanny valley" theory [35] in the context of AI-generated content remains a particularly promising line of inquiry. Building on this study's findings, future research should not only examine avoidance-based "uncanny valley" effects but also investigate conditions under which perceptual discomfort or "wrongness" produces increased attention and reach. Experimental designs exposing participants to AI-generated images and videos with varying degrees of realism and explicit artificiality could clarify whether discomfort suppresses engagement, amplifies reach, or operates differently across content formats. Such insights would be especially valuable for the design of transparent, authentic digital health agents that balance ethical compliance with psychological comfort.

An emerging area of relevance concerns the potential for VPs to establish and promote their own brands, mirroring prominent human influencers. Future studies could examine whether VP-led brands benefit from the same trust and identity-consistency mechanisms identified in this research, or whether commercialized virtual identities introduce additional ethical and authenticity challenges. Understanding this dynamic could offer both commercial and health-focused organizations opportunities for innovation in digital identity, persuasion, and trust formation [42].

As the VI space continues to evolve, further attention should be paid to emerging human-like subcategories, particularly human-like male VPs [43], who remain highly underrepresented in the current landscape. Moreover, a particularly underexplored but empirically relevant direction concerns alternative and nonnormative virtual body types. This study identifies VP 104 and VP 141 as rare cases of VPs embodying morbidly obese physiques that nevertheless achieve exceptional video engagement and reach.

These VPs demonstrate that deviation from dominant beauty norms does not inherently suppress performance and may instead activate engagement through novelty, transparency, or emotional arousal. Future research should systematically investigate such alternative body representations, extending analysis to additional minority VPs, such as those representing individuals with Down syndrome and thereby challenging conventional beauty standards in different ways. These investigations would advance understanding of authenticity, diversity, stigma, body image, user engagement with virtual content, user anxiety, and inclusivity within AI-generated identity spaces and their potential implications for psychological well-being [26,44,45].

A significant observation from this study is the underperformance of VPs representing ethnic minorities. Currently, only a small number of macro-level Black VPs, including VP 58, VP 94, VP 100, and VP 151, achieve substantial visibility. This raisesimportant questions about the sources of this underrepresentation—whether it reflects algorithmic bias in generative AI systems or deeper societal inequalities that are being reproduced in virtual spaces. Future research should examine these factors systematically to determine whether AI adjustments and data diversification can mitigate bias, or whether more structural, societal interventions are required to promote equity and representation in the evolving ecosystem of virtual personas.

Summary

Findings from the empirical analysis indicate that VP engagement varied notably across content formats and also co-occurred with physical and behavioral consistency, transparency, authenticity, and disclosure of artificiality. Among these factors, AI-generated video posts showed higher audience interaction and reach than static images in this dataset. These results highlight the importance of both technological sophistication and ethical compliance in shaping user response to AI-mediated social content.

Conclusions

This study contributes to the growing literature on VIs and AI-mediated communication by examining reach and engagement patterns among established human-like female VPs on Instagram. Across sampled accounts, high performance aligned closely with identity coherence, transparent virtual presentation, and format-specific content strategies, rather than with photorealism alone.

The results further suggest that image engagement, video engagement, and video reach should not be treated as interchangeable outcomes. CGI-like presentation was frequently

XSL•FO
RenderX

observed among the most widely distributed videos, whereas photorealistic and face-swapped motion content was more common among highly engaging videos by like rate. Copresence and recurring appearance stylizations also appeared repeatedly among high-performing posts, although these patterns should be interpreted as descriptive features of established accounts rather than as universal success factors.

Because the sample was restricted to macro- and mega-level human-like female VPs, the findings characterize successful incumbents and do not establish causal predictors of audience growth. Further work should extend analysis to smaller profiles, incorporate richer engagement measures, directly assess user perception of artificiality, and further examine the ethical implications of synthetic identity design in digitally mediated environments.

Finally, this study highlights the need for responsible innovation in the rapidly expanding domain of VPs. As AI-generated identities grow more pervasive and persuasive, interdisciplinary collaboration between technologists, behavioral scientists, and health communication experts will be essential to ensure that these systems are deployed not only to maximize engagement but also to safeguard user well-being, promote digital literacy, and foster trust in AI-mediated human interaction.

Acknowledgments

The authors used a generative artificial intelligence tool for language editing and structural suggestions during revision of this manuscript. All study design decisions, data collection, and data analysis were carried out by the authors.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Data Availability

The data for the study are available from the authors on reasonable request.

Authors' Contributions

The authors contributed equally.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Content coding framework and full-sample descriptive statistics.
[DOCX File, 22 KB - jmir_v28i1e86233_app1.docx]

Multimedia Appendix 2
Visual evidence of content design variables.
[DOCX File, 3134 KB - jmir_v28i1e86233_app2.docx]

References

1. Audrezet A, Koles B. Virtual influencer as a brand avatar in interactive marketing. In: Wang C, editor. The Palgrave Handbook of Interactive Marketing: Springer International Publishing; 2023:353-376. [doi: 10.1007/978-3-031-14961-0_16]
2. Drenten J, Brooks G. Celebrity 2.0: Lil Miquela and the rise of a virtual star system. Feminist Media Stud 2020 Nov 16;20(8):1319-1323. [doi: 10.1080/14680777.2020.1830927]
3. Brüns JD, Meißner M. Do you create your content yourself? Using generative artificial intelligence for social media content creation diminishes perceived brand authenticity. J Retailing Consum Serv 2024 Jul;79:103790. [doi: 10.1016/j.jretconser.2024.103790]
4. Solis B. Introducing the GenAI prism infographic: a framework for collaborating with generative AI.: Brian Solis; 2023 Dec. URL: https://briansolis.com/2023/12/introducing-the-genai-prism-infographic-a-framework-for-colalborating-with-generative-ai/ [accessed 2026-05-04]
5. Arsenyan J, Mirowska A. Almost human? A comparative case study on the social media presence of virtual influencers. Int J Hum Comput Stud 2021 Nov;155:102694. [doi: 10.1016/j.ijhcs.2021.102694]
6. Sands S, Campbell CL, Plangger K, Ferraro C. Unreal influence: leveraging AI in influencer marketing. Eur J Mark 2022 Jun 7;56(6):1721-1747. [doi: 10.1108/EJM-12-2019-0949]
7. Block E, Lovegrove R. Discordant storytelling, 'honest fakery', identity peddling: How uncanny CGI characters are jamming public relations and influencer practices. Public Relat Inq 2021 Sep;10(3):265-293. [doi: 10.1177/2046147X211026936]

XSL•FO
RenderX

8. Lou C, Kiew STJ, Chen T, Lee TYM, Ong JEC, Phua Z. Authentically fake? How consumers respond to the influence of virtual influencers. J Advert 2023 Aug 8;52(4):540-557. [doi: 10.1080/00913367.2022.2149641]
9. Buy Instagram accounts with real followers. Social Tradia. 2024. URL: https://socialtradia.com/ [accessed 2026-05-04]
10. Belanche D, Casaló LV, Flavián M. Human versus virtual influences, a comparative study. J Bus Res 2024 Feb;173:114493. [doi: 10.1016/j.jbusres.2023.114493]
11. Xie-Carson L, Magor T, Benckendorff P, Hughes K. All hype or the real deal? Investigating user engagement with virtual influencers in tourism. Tour Manag 2023 Dec;99:104779. [doi: 10.1016/j.tourman.2023.104779]
12. Franke C, Groeppel-Klein A, Müller K. Consumers' responses to virtual influencers as advertising endorsers: novel and effective or uncanny and deceiving? J Advert 2023 Aug 8;52(4):523-539. [doi: 10.1080/00913367.2022.2154721]
13. Mori M, MacDorman KF, Kageki N. The uncanny valley. IEEE Robot Autom Mag 2012;19(2):98-100. [doi: 10.1109/MRA.2012.2192811]
14. Provoost S, Lau HM, Ruwaard J, Riper H. Embodied conversational agents in clinical psychology: a scoping review. J Med Internet Res 2017 May 9;19(5):e151. [doi: 10.2196/jmir.6553] [Medline: 28487267]
15. Bickmore TW, Picard RW. Establishing and maintaining long-term human-computer relationships. ACM Trans Comput-Hum Interact 2005 Jun;12(2):293-327. [doi: 10.1145/1067860.1067867]
16. Stein JP, Linda Breves P, Anders N. Parasocial interactions with real and virtual influencers: The role of perceived similarity and human-likeness. New Media Soc 2024 Jun;26(6):3433-3453. [doi: 10.1177/14614448221102900]
17. Liu S, Aw ECX, Tan GWH, Ooi KB. Virtual influencers as the next generation of influencer marketing: identifying antecedents and consequences. In: Al-Sharafi M, Al-Emran M, Tan GWH, Ooi KB, editors. Current and Future Trends on Intelligent Technology Adoption: Springer; 2023:23-39. [doi: 10.1007/978-3-031-48397-4_2]
18. Turner JC. Social comparison and social identity: some prospects for intergroup behaviour. Eur J Soc Psychol 1975 Jan;5(1):1-34. [doi: 10.1002/ejsp.2420050102]
19. Um N. Predictors affecting effects of virtual influencer advertising among college students. Sustainability 2023;15(8):6388. [doi: 10.3390/su15086388]
20. Davlembayeva D, Chari S, Papagiannidis S. Virtual influencers in consumer behaviour: a social influence theory perspective. Br J Manage 2025 Jan;36(1):202-222. [doi: 10.1111/1467-8551.12839]
21. Duffek B, Eisingerich AB, Merlo O, Lee G. Authenticity in influencer marketing: how can influencers and brands work together to build and maintain influencer authenticity? J Mark 2025 Sep;89(5):21-46. [doi: 10.1177/00222429251319786]
22. Bickmore TW, Trinh H, Olafsson S, et al. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google Assistant. J Med Internet Res 2018 Sep 4;20(9):e11510. [doi: 10.2196/11510] [Medline: 30181110]
23. Abd-Alrazaq AA, Rababeh A, Alajlani M, Bewick BM, Househ M. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. J Med Internet Res 2020 Jul 13;22(7):e16021. [doi: 10.2196/16021] [Medline: 32673216]
24. Schachner T, Keller R, V Wangenheim F. Artificial intelligence-based conversational agents for chronic conditions: systematic literature review. J Med Internet Res 2020 Sep 14;22(9):e20701. [doi: 10.2196/20701] [Medline: 32924957]
25. Chaix B, Bibault JE, Pienkowski A, et al. When chatbots meet patients: one-year prospective study of conversations between patients with breast cancer and a chatbot. JMIR Cancer 2019 May 2;5(1):e12856. [doi: 10.2196/12856] [Medline: 31045505]
26. Anto A, Asif RO, Basu A, et al. Exploring the impact of social media on anxiety among university students in the United Kingdom: qualitative study. JMIR Form Res 2023 Jun 16;7:e43037. [doi: 10.2196/43037] [Medline: 37327030]
27. Ethics approval overview. Imperial, Research Ethics. URL: https://www.imperial.ac.uk/research-ethics-committee/ethics-approval-overview [accessed 2026-03-18]
28. Milne-Ives M, de Cock C, Lim E, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res 2020 Oct 22;22(10):e20346. [doi: 10.2196/20346] [Medline: 33090118]
29. Nizri E. Macro-influencer vs micro-influencer: the great social media debate. Forbes Technology Council. 2022 Mar 1. URL: https://www.forbes.com/councils/forbestechcouncil/2022/03/01/macro-influencers-vs-micro-influencers-the-great-social-media-debate/ [accessed 2026-05-04]
30. Our approach to labelling AI-generated content and manipulated media. Meta. 2024 Apr 5. URL: https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/ [accessed 2026-05-04]
31. WAICA–jury members & finalist. World AI Creator Awards. 2024. URL: https://www.waicas.com/ [accessed 2026-05-04]
32. Cascio Rizzo GL, Berger JA, Villarroel Ordenes F. What drives virtual influencer's impact? SSRN J. 2023. [doi: 10.2139/ssrn.4329150]
33. Sheikh M. 50+ social media video marketing statistics for 2024. Sprout Social. 2024. URL: https://sproutsocial.com/insights/social-media-video-statistics/ [accessed 2026-05-04]
34. Digital 2024–global overview report. : We Are Social Ltd; 2024 URL: https://wearesocial.com/uk/blog/2024/01/digital-2024/ [accessed 2026-05-04]
35. Mori M. Bukimi no tani. Energy (Oxf) 1970;7(4):33-35 [FREE Full text]
36. Jarzyna CL. Parasocial interaction, the COVID-19 quarantine, and digital age media. Hum Arenas 2021 Sep;4(3):413-429. [doi: 10.1007/s42087-020-00156-0]

XSL•FO
RenderX

37. Gaysynsky A, Senft Everson N, Heley K, Chou WYS. Perceptions of health misinformation on social media: cross-sectional survey study. JMIR Infodemiology 2024 Apr 30;4:e51127. [doi: 10.2196/51127] [Medline: 38687591]
38. Kaplan A, Haenlein M. Rulers of the world, unite! the challenges and opportunities of artificial intelligence. Bus Horiz 2020 Jan;63(1):37-50. [doi: 10.1016/j.bushor.2019.09.003]
39. Bickmore TW, Pfeifer LM, Jack BW. Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. Presented at: Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI 2009); Apr 4-9, 2009; Boston, MA, USA p. 1265-1274. [doi: 10.1145/1518701.1518891]
40. Riddall J. Utilizing local influencers for digital marketing success.: Search Engine J; 2024 Aug 20. URL: https://www.searchenginejournal.com/utilizing-local-influencers-for-digital-marketing-success/522895/ [accessed 2026-05-04]
41. Koles B, Audrezet A, Moulard JG, Ameen N, McKenna B. The authentic virtual influencer: authenticity manifestations in the metaverse. J Bus Res 2024 Jan;170:114325. [doi: 10.1016/j.jbusres.2023.114325]
42. Igarashi R, Bhoumik K, Thompson J. Investigating the effectiveness of virtual influencers in prosocial marketing. Psychol Mark 2024 Sep;41(9):2121-2135. [doi: 10.1002/mar.22031]
43. Platter D. Why even the world's richest men have fake followers (and what marketers can do about it). Forbes. 2021. URL: https://www.forbes.com/councils/forbescommunicationscouncil/2021/08/27/why-even-the-worlds-richest-men-have-fake-followers-and-what-marketers-can-do-about-it/ [accessed 2026-05-04]
44. Martins N, Williams DC, Harrison K, Ratan RA. A content analysis of female body imagery in video games. Sex Roles 2009 Dec;61(11-12):824-836. [doi: 10.1007/s11199-009-9682-9]
45. Lynch T, Tompkins JE, van Driel II, Fritz N. Sexy, strong, and secondary: a content analysis of female characters in video games across 31 years. J Commun 2016 Aug;66(4):564-584. [doi: 10.1111/jcom.12237]

Abbreviations

AI: artificial intelligence
CGI: computer-generated imagery (displaying legible artificiality)
VI: virtual influencer
VP: virtual profile

Edited by A Coristine, T Leung; submitted 21.Oct.2025; peer-reviewed by L Zhang, S Fasasi, Y Wang; revised version received 07.Apr.2026; accepted 07.Apr.2026; published 05.Jun.2026.

Please cite as:
Lebrecht AM, Tam W, Merlo O, Eisingerich AB
Characteristics of Human-Like Virtual Profiles in Relation to Audience Reach and Engagement on Instagram: Secondary Data Analysis
J Med Internet Res 2026;28:e86233
URL: https://www.jmir.org/2026/1/e86233
doi:10.2196/86233

© Alexandra Maria Lebrecht, Winze Tam, Omar Merlo, Andreas Benedikt Eisingerich. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Human Papillomavirus Vaccine Discourse and Sentiment on Reddit Before and After COVID-19: Mixed Methods Retrospective Cross-Sectional Study

Marcin Marciniak[1,2], BS; Sean A Setzen[1,3], MD; Deeya Bhattacharya[4], BS; Jamie Masliah[1], MD; Robin Powszok[1], MD, MS; Erich M Sturgis[5], MD, MPH; Vanessa C Stubbs[1], MD; Mihir K Bhayani[1], MD

[1]Department of Otorhinolaryngology-Head and Neck Surgery, Rush University Medical Center, 1611 W Harrison St Suite 550, Chicago, IL, United States

[2]Rosalind Franklin University of Science and Medicine, North Chicago, IL, United States

[3]Department of Otorhinolaryngology-Head and Neck Surgery, St. Luke's University Health Network, Bethlehem, PA, United States

[4]Rush University Medical College, Chicago, IL, United States

[5]Department of Otolaryngology-Head and Neck Surgery, Baylor College of Medicine, Houston, TX, United States

Corresponding Author:
Mihir K Bhayani, MD
Department of Otorhinolaryngology-Head and Neck Surgery, Rush University Medical Center, 1611 W Harrison St Suite 550, Chicago, IL, United States

Abstract

Background: Human papillomavirus (HPV) is a sexually transmitted virus that causes various oropharyngeal and anogenital cancers. The HPV vaccine provides protection against several strains of HPV and is a key preventative tool against HPV-related cancers; however, vaccination rates remain suboptimal in the United States due to variable state mandates and misperceptions of vaccine efficacy and risks. As social media becomes an increasingly popular avenue for health discussions, platforms such as Reddit offer opportunities to understand public vaccine discourse, particularly among underrepresented groups. Furthermore, vaccine hesitancy and mistrust generally increased during the COVID-19 pandemic, potentially impacting HPV vaccination rates.

Objective: This study aimed to characterize HPV vaccine–related discussions on Reddit by (1) identifying dominant themes, (2) assessing sentiment, and (3) examining shifts in themes and sentiment before and after the start of the COVID-19 pandemic.

Methods: This convergent mixed methods analysis used a Python-based script to extract Reddit posts and comments referencing the HPV vaccine from multiple subreddits from September 13, 2016, to February 13, 2025. Entries were manually labeled as intentional or incidental, with intentional entries further categorized into 10 thematic categories. Sentiment analysis was conducted using the VADER (Valence Aware Dictionary for Sentiment Reasoning) algorithm. Pre– vs post–COVID-19 comparisons used March 11, 2020, as the cutoff date. Temporal trends were assessed using pre– and post–COVID-19 stratification. Chi-square tests, Mann-Whitney *U* tests, and linear regression were used for statistical analysis.

Results: Of 4235 collected entries, 2801 (66.1%) intentional posts and comments were analyzed. The most common themes were factual claims (703/2801, 25.1%), personal experiences (571/2801, 20.4%), and offering advice (465/2801, 16.6%). Overall sentiment was 51.7% (1448/2801) positive (95% CI 49.8%-53.5%), 37.6% (1053/2801) negative, and 10.8% (300/2801) neutral, with a median sentiment score of 0.13 (IQR –0.49 to 0.71). Pre–COVID-19 entries (n=278) were 124 (44.6%) positive (95% CI 38.7%-50.7%), with a median sentiment of 0.00 (–0.66 to 0.58). Post–COVID-19 entries (n=2523) were 1324 (52.4%) positive (95% CI 50.5%-54.4%), with a median sentiment of 0.17 (–0.46 to 0.72). Sentiment increased significantly post–COVID-19 (P=.01). Theme distribution differed before vs after COVID-19 (χ^2_9=114.47, P<.001). Pre–COVID-19 discussions overrepresented barriers and social or cultural influences, whereas post–COVID-19 discussions more frequently reflected personal experiences, advice seeking, and advice offering. Posting volume increased by 50% per year throughout the study period (incidence rate ratio=1.50, 95% CI 1.47-1.53; P<.001), with a steeper increase post–COVID-19 (incidence rate ratio=1.60, 95% CI 1.56-1.65).

Conclusions: This study provides a post–COVID-19 perspective on patient-generated HPV vaccine discourse by integrating thematic content with sentiment and temporal trends. Prior studies have been limited by a narrower thematic scope and have focused on prepandemic periods. By examining discussions on an anonymous platform such as Reddit, this analysis captures personal conversations and assesses pandemic-related shifts in discourse. These findings add to the literature on digital vaccine communication and highlight opportunities for targeted public health engagement and strategies to leverage online communities.

(J Med Internet Res 2026;28:e83558) doi:10.2196/83558

XSL•FO
RenderX

KEYWORDS

papillomavirus vaccines; papillomavirus infections; social media; sentiment analysis; COVID-19

Introduction

Human papillomavirus (HPV) is a sexually transmitted virus and a well-established risk factor for various cancers in both men and women [1]. Most commonly associated with cervical cancer, HPV has also been increasingly linked to anogenital and oropharyngeal cancers [1,2]. High-risk HPV strains, particularly types 16 and 18, are responsible for over 70% of these cases, while low-risk strains cause benign conditions such as papillomas and warts [3,4].

Vaccines have been developed to protect against high- and low-risk HPV strains, reducing the likelihood of persistent infection. The first HPV vaccine, Gardasil, was approved by the US Food and Drug Administration in 2006 and provided protection against HPV 6, 11, 16, and 18 [5]. Subsequent vaccines, such as Gardasil 9, expanded coverage to include 5 additional high-risk strains (HPV 31, 33, 45, 52, and 58), offering broader protection against HPV-associated cancers [6]. Research has shown that HPV vaccination is highly effective in preventing certain cancers [7]. Despite this, vaccination coverage remains suboptimal, with only about 66% of eligible individuals completing the full vaccine series by age 17 and approximately 78% receiving at least 1 dose [8,9].

Unlike vaccines routinely required for school attendance, the HPV vaccine encounters distinct challenges. Mandates for HPV vaccination exist in only a few states, and even these often include opt-out provisions that hinder widespread uptake [10]. Globally, there is a varying level of support for the HPV vaccine among parents, which is particularly low in the United States. Parents of sons are less likely to promote HPV vaccination for their children or may be unaware of its existence or efficacy [11]. Studies have shown that education on HPV in schools led to a positive impact on parents supporting HPV vaccination for their children [12,13]. Since its introduction, the vaccine has also been embroiled in controversy, with public hesitancy fueled by concerns regarding its safety, efficacy, and perceived association with sexual activity [10,14]. A lack of awareness of HPV infection in men and its effects on them and their sexual partners may contribute to this skepticism, although research has shown that men benefit from HPV vaccination in preventing anogenital disease [15]. Moreover, the COVID-19 pandemic limited the availability of vaccines and coverage [13] and may have exacerbated this skepticism, as heightened public discourse around vaccines amplified misinformation, emotional responses, and confusion [16,17]. Data show that HPV vaccination rates were increasing until 2022 [18], which may have been affected by increased vaccine hesitancy during the COVID-19 pandemic.

While health care providers typically rely on evidence-based guidelines for vaccine advocacy, this approach can sometimes overlook the personal experiences, specific concerns, and misconceptions held by patients. Addressing these individual perspectives is crucial for improving HPV vaccination rates and for bridging the divide between scientific recommendations and public apprehension. Although research exists examining providers regarding HPV vaccine hesitancy based on their experiences with patients, patient perspectives are not well represented in the literature [17]. There are also many groups that may be underrepresented in this understanding of HPV vaccine hesitancy; for instance, there is a paucity of published literature on HPV vaccination among men who have sex with men [2]. In this context, social media platforms provide rich, readily available datasets for exploring patient viewpoints, offering advantages over traditional survey methods, which can be resource-intensive and unable to capture spontaneous, real-time discussions [19].

However, systematically reviewing individual social media posts and comments may hinder the efficacy of research efforts. This is where sentiment analysis is useful, as it is a methodology that uses natural language processing (NLP) to understand the opinions expressed within text and to systemically measure emotional tone [20]. One tool used for sentiment analysis is VADER (Valence Aware Dictionary for Sentiment Reasoning), which has been shown to be useful in deciphering sentiments within social media, such as comments on YouTube [21,22]. VADER operates by using a lexicon or dictionary with predetermined positive, negative, or neutral sentiment values, allowing code to efficiently parse through large amounts of text posted on social media [20].

Although previous research has extensively analyzed HPV vaccine discussions on Twitter, a platform popular among adolescents and young adults who are key demographics for the vaccine, less attention has been paid to corresponding discussions on Reddit [14,19,23-29]. Earlier work by Lama et al [30] examined Reddit discourse from 2007 to 2015 using topic modeling and identified themes such as political debate, sexual behavior, and vaccination schedules. More recent work by Du et al [31] and Meci et al [32] explored misinformation prevalence and conducted sentiment analysis of Reddit posts, respectively, but both studies were limited by narrow thematic scopes and did not assess how discourse evolved in response to the COVID-19 pandemic [31,32]. Recent research by Singh et al [33] evaluated perceptions of HPV vaccination among Indian users on Reddit but was limited to this specific demographic, suggesting that a broader study is necessary.

Consequently, there is a gap in understanding the longitudinal evolution of HPV vaccine discourse on Reddit, particularly regarding shifts in themes and sentiment surrounding major public health events such as the COVID-19 pandemic, analyzed through detailed manual coding. Recent work has shown that VADER is useful in parsing Reddit posts when examining views related to the COVID-19 pandemic [21,34]. This study addresses this gap by analyzing both Reddit posts and comments from 2016 to 2025 using a granular, 10-category manual coding scheme. It uniquely examines how discussion themes and overall sentiment changed before and after the onset of the COVID-19 pandemic.

Reddit offers a distinct environment for such analysis. Unlike platforms often linked to identifiable user accounts, Reddit

XSL•FO
RenderX

facilitates anonymous, long-form, asynchronous conversations within topic-specific communities called subreddits. These features foster more candid discussions and deeper exchanges between users. Prior research indicates that Reddit users are predominantly younger than 35 years, with 1 study reporting a mean user age of 25.7 years, which aligns closely with the target demographic for HPV vaccination and makes Reddit particularly relevant for examining vaccine attitudes [31,35]. The large amount of subreddit communities also allows researchers to capture discussions across diverse demographic and behavioral contexts. Because Reddit preserves entire discussion threads and allows posts to resurface through ongoing commenting, it provides a unique opportunity to observe how vaccine-related conversations evolve longitudinally. Analyzing these interactions can illuminate prevalent misconceptions, identify barriers to vaccination, and reveal opportunities for more effective public health communication.

This study aimed to address the following research questions: (1) What key thematic categories characterize Reddit discussions about the HPV vaccine? (2) What is the overall sentiment expressed toward the HPV vaccine on Reddit? and (3) How did themes, sentiment, and discourse change before versus after the start of the COVID-19 pandemic? To answer these questions, we conducted a retrospective cross-sectional analysis of Reddit posts, applying manual thematic categorization and VADER sentiment analysis and comparing temporal patterns prepandemic and post pandemic. Findings from this work may help inform more effective and empathetic communication strategies for health care professionals and public health officials.

Methods

Ethical Considerations

This study was reviewed by the Rush University Medical Center Institutional Review Board and was determined to be exempt from human subjects research (5511) due to the secondary use of publicly available data. No informed consent was required because all data were publicly accessible, contained no identifiable private information, and were analyzed in aggregate. No images included in this manuscript or supplementary materials contain identifiable information about individual users. All posts and comments were collected in accordance with Reddit's application programming interface Terms of Service. All study materials were stored in a secure institutional OneDrive folder. No compensation was provided, as no participants were recruited or directly involved in the study.

Research Design Overview

This study used a convergent mixed methods descriptive design and followed the APA Journal Article Reporting Standards for mixed methods research [36]. Mixed methods were used to allow the identification of dominant themes in HPV vaccine–related discourse, followed by quantitative analysis of sentiment, thematic frequencies, and temporal trends.

Qualitative and quantitative analyses were conducted on the same set of Reddit posts and comments from September 13, 2016, to February 13, 2025. Qualitative thematic analysis was performed first to classify posts as incidental or intentional HPV vaccine discussions and to assign thematic categories. Quantitative analyses were subsequently conducted to examine thematic distribution, sentiment scores, and posting volume. Integration occurred during analysis and interpretation, with qualitative themes serving as the framework for quantitative comparisons.

Qualitative Methods

Overview

A descriptive qualitative thematic analysis was conducted to identify and characterize patterns of HPV vaccine–related discourse on Reddit.

Data Sources and Collection

Posts and comments referencing the HPV vaccine within the study period were collected from multiple subreddits, along with their respective metadata. The subreddits analyzed are outlined in Table 1. Using Visual Studio Code (version 1.90.0 [Universal]), an integrated development environment developed by Microsoft (Multimedia Appendix 1), a data scraping script was written in Python (version 3.9.16) using the PRAW library to interact with the Reddit application programming interface and collect entries containing keywords from selected subreddits to identify those mentioning the HPV vaccine (Table 1). For each post and comment, metadata—including the subreddit name, text body, flair, date of submission, number of upvotes, post or comment ID, parent comment ID, entry type (post or comment), and anonymized author ID—were collected (Multimedia Appendices 2 and 3).

XSL•FO
RenderX

Table . Subreddits included in the study and keyword mapping strategy for identifying human papillomavirus vaccine–related posts.[a]

Subreddit category	Subreddits	Keywords applied
HPV[b]-focused	r/HPV, r/PreCervicalCancer, r/Warts, r/Cervical-Cancer	Vaccine-related keywords: • vaccine, vaccines, vaccinated, vaccination, vaccinations, vax, vaxxed, unvaxxed • Gardasil, Cervarix • anti-vax, antivax, antivaccine, anti-vaccine, pro-vax • jab, shot, immunization
Vaccine-focused	r/VACCINES, r/vaxxhappened, r/Vaccine,r/anti-vax, r/DebateVaccines,r/unvaccinated, r/skeptic, r/StopMandatoryVaccine	HPV-related keywords: • HPV, HPV vaccine, HPV shot • Gardasil, Cervarix • human papillomavirus
Medical and science	r/AskDocs, r/medicine, r/PeterAttia, r/Headand-NeckCancer,r/Health, r/ScienceBasedParenting, r/HNSCC	HPV+vaccine-related keywords: • HPV, HPV vaccine, HPV shot, Gardasil, Cervarix, human papillomavirus
Women's health	r/TwoXChromosomes, r/TheGirlSurvivalGuide, r/WomenOver40, r/AskWomenOver30	HPV+vaccine-related keywords: • vaccine, vaccines, vaccinated, vaccination, jab, shot, immunization
General and social	r/MensRights, r/AskGaybrosOver30, r/AskCon-servatives, r/askgaybros	HPV+vaccine-related keywords: • anti-vax, antivax, pro-vax

[a]This table lists all subreddits analyzed in this retrospective cross-sectional social media content analysis of human papillomavirus vaccine discussions on Reddit. It details each subreddit's overall category and the keyword strategy applied to identify human papillomavirus vaccine–related posts and comments collected from September 13, 2016, to February 13, 2025.

[b]HPV: human papillomavirus.

Data Selection and Preprocessing

Data were collected and preprocessed prior to analysis using the following steps:

1. Keyword mapping: Keywords were divided into 2 primary categories to identify relevant discussion—general vaccine–related terms and HPV-specific terms (Table 1). Keyword groups were mapped to subreddits based on thematic relevance. HPV-specific subreddits (eg, r/HPV) were filtered using the general vaccine–related keyword group, vaccine-focused subreddits (eg, r/VACCINES, r/vaxxhappened) were filtered using the HPV-specific keyword group, and broad medical subreddits (eg, r/AskDocs) used both keyword groups. Keywords were processed to prioritize longer, more specific terms (eg, "HPV vaccine" before "HPV") to prevent mismatches in which shorter keywords might appear within unrelated longer words or phrases. All retrieved entries and metadata were exported to Microsoft Excel for further processing.
2. Text cleaning and tokenization: Basic NLP preprocessing was performed to prepare entries for analysis. This process consisted of decoding HTML entities, converting all text to lowercase, and tokenization to ensure compatibility with the VADER sentiment tool.
3. Filtering and inclusion: Entries were screened and categorized as either "incidental" or "intentional" discussion. Intentional entries were defined as posts or comments in which the HPV vaccine was the primary focus of discussion. In contrast, incidental entries were defined as content in which a keyword was present but the HPV vaccine was mentioned only tangentially. Examples of incidental entries included listing "HPV vaccine" as part of a general medical history without further discussion or brief references within unrelated discussions in which the vaccine was not the central topic (Figure 1). Inclusion criteria were posts and comments labeled as "intentional" and those consisting of 10 words or more.

XSL•FO
RenderX

Figure 1. Flowchart of data collection, filtering, and manual labeling of Reddit posts related to the human papillomavirus (HPV) vaccine (2016 - 2025). This flowchart summarizes the workflow for identifying and processing Reddit posts and comments referencing the HPV vaccine in this retrospective cross-sectional social media content analysis. After keyword-based extraction (2016 - 2025), entries were screened for relevance and classified as intentional (directly discussing HPV vaccination) or incidental (mentioning keywords without substantive relevance and excluded). Intentional entries were then manually coded into 10 mutually exclusive thematic categories for analysis.

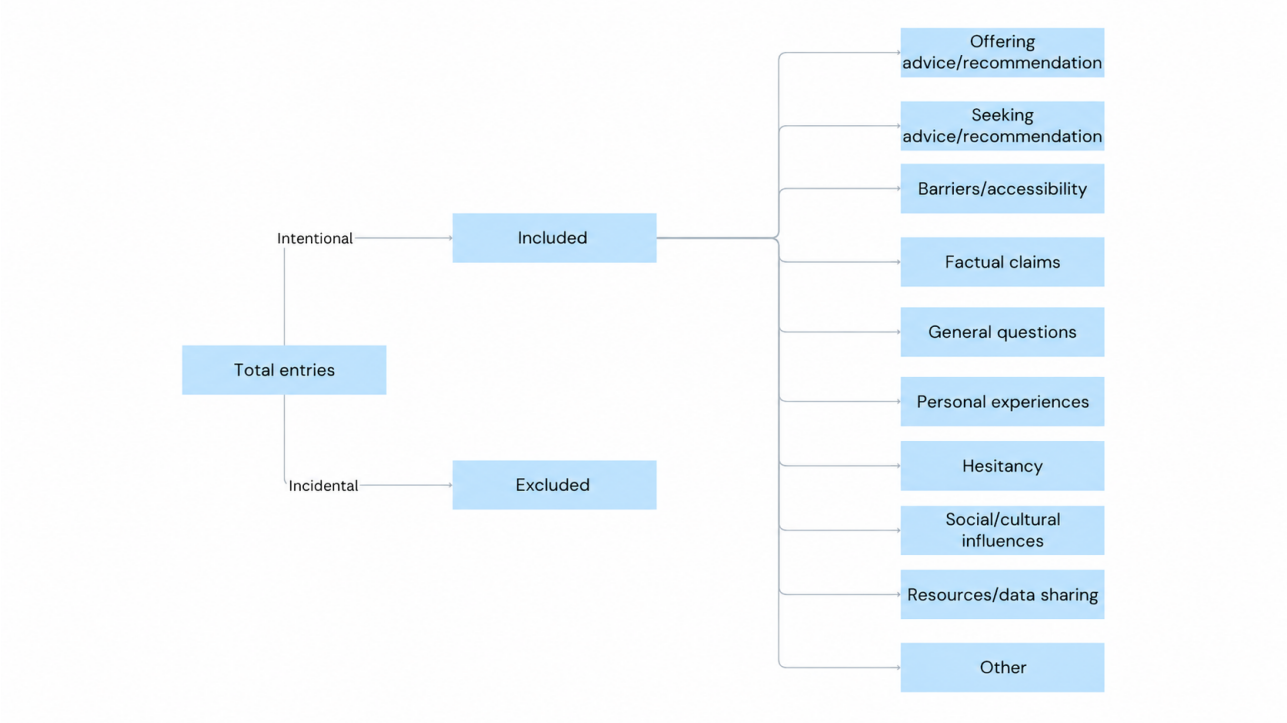

Thematic Analysis

Intentional entries underwent further classification into 10 thematic subcategories, which are listed and described in Table 2. These 10 categories were created after an initial screening of 200 entries. To ensure reliability, 2 authors (MM, DB) independently performed manual data labeling, and any discrepancies were resolved through discussion and consensus. Because the dataset was predefined and finite, categorization continued until all eligible content was labeled rather than terminating at the point of saturation. Thematic saturation was assessed retrospectively and was considered achieved when no new thematic categories emerged during the later stages of analysis.

XSL•FO
RenderX

Table . Definitions and examples of thematic categories used in manual coding of Reddit posts.[a]

Theme	Definition	Examples
Advice or recommendations (offering)	Posts that explicitly offer advice or recommendations regarding HPV[b] vaccination to others.	"You should get vaccinated," "My doctor recommended getting the vaccine"
Advice or recommendations (seeking)	Posts that request advice or opinions from others about receiving the HPV vaccine.	"Should I get the vaccine?," "Should I encourage my partner to get the vaccine?"
Barriers or accessibility	Posts that describe challenges in accessing the HPV vaccine, including financial, insurance, geographic, or age-related limitations.	"I wanted the vaccine 3 months before I was 26, insurance said no, my age disqualified coverage, but I hadn't turned 26 yet," "My daughter got the shot after she moved out and the first shot was over 400 dollars. I told her about in network and the 2nd was free."
Factual claims	Statements presented as facts, regardless of accuracy, typically lacking overt opinion or emotion. Intended to convey information or assert knowledge.	"Gardasil 9 protects against nine types of human papillomavirus (HPV): 6, 11, 16, 18, 31, 33, 45, 52, and 58," "There's zero data on actual cervical cancer rates declining due to HPV vaccines."
General questions	Posts that inquire about specific aspects of the HPV vaccine, including dosing, efficacy, safety, or eligibility.	"Is two shots enough to protect you?," "Why the difference between US CDC 3-dose and UK NHS 2-dose recommendation for HPV vaccination for older adults up to age 45?"
Hesitancy	Posts expressing concerns, skepticism, or fear about HPV vaccine safety, efficacy, or perceived risks. May include conspiracy or mistrust themes.	"I know someone who periodically needs to have her skull drained of excess fluid because of the HPV vaccine. They almost died after taking it due to fluid build up around her brains. I stupidly took it as an adult at 36 and have had autoimmune issues ever since.," "And my generation, 90% of us have HPV and don't know it and we're fine... U think doctors don't lie? Doctors aren't corrupt or capable of being corrupted? Just use common sense. Less vacinnes the better IMHO!"
Personal experiences	Narratives sharing personal or family experiences with HPV vaccination, including positive, negative, or neutral outcomes.	"I got my 1st dose of Gardasil vaccine on Thursday. I have a fever of 103.3 and I feel really sick. How are you doing?," "I was diagnosed with HPV after almost 20 years of marriage... I'm now getting vaccinated at 43, even though I had a hysterectomy because I want to make sure I'm protected from any possible further in the future.," "I had the vaccine aged 12, and got diagnosed with HPV at my first smear, ended up having a colposcopy to remove dodgy cells. It is not a guaranteed prevention."
Resources or data sharing	Posts that share articles, external links, or statistical information to support a claim, argument, or question.	"The CDC recommends that women get the HPV vaccine through age 26, if they haven't already been vaccinated. https://www.cancer.gov/types/cervical/causes-risk-prevention," "The risk of cancer for low risk strains is quite low. The high risk strains themselves comprise the majority of HPV-derived cancer. See [this table] (https://www.cdc.gov/cancer/hpv/statistics/cases.htm) for a breakdown of risk of cancer type by HPV strain."
Social or cultural influences	Posts discussing how culture, religion, family, or societal norms influence attitudes or decisions regarding HPV vaccination.	"I thought, because of what the Ministry of Health informed back then in my country, the HPV vaccine only worked in young girls." "I never got the vaccine when I was younger (due to being too young to make my own decision and my mom being against it). I'm obviously very upset but just happy to see this post and hopefully my HPV clears up and goes dormant once this is all over with"

XSL•FO
RenderX

Theme	Definition	Examples
Other	Posts that do not fit into other categories due to ambiguity, sarcasm, or lack of context.	"At what age did you get the vaccine?," "I'm not her, so no - her witness testimony is evidence of her claims & she's more credible than Gardasil."

[a]This table summarizes the 10 manually coded thematic categories applied to Reddit posts and comments intentionally discussing the human papillomavirus vaccine. Definitions and representative examples illustrate how each entry was categorized in this cross-sectional analysis of human papillomavirus vaccine discourse on Reddit (2016 - 2025).

[b]HPV: human papillomavirus.

Quantitative Methods

Inclusion and Exclusion Criteria

Analyses of sentiment scores, thematic frequencies, thematic comparisons, and temporal trends were restricted to posts classified as intentional HPV vaccine discussions with 10 words or more. For the analysis of time from a post to the first response, all posts and comments, including both intentional and incidental mentions, were included.

Data Sources and Characteristics

Quantitative data consisted of metadata and derived variables from Reddit posts and comments, including posting date, subreddit, upvote counts, response time to first comment, sentiment scores, and assigned thematic categories. Demographic characteristics were unavailable due to the anonymous nature of the platform.

Sample Size, Power, and Precision

The final analytic dataset included all posts and comments meeting the inclusion criteria. Given the descriptive and exploratory nature of the study, formal power calculations were not conducted. Precision of estimates was assessed using confidence intervals where applicable.

Measures and Data Collection

Sentiment analysis was conducted using the VADER (Valence Aware Dictionary and Sentiment Reasoning) algorithm, which was selected because it is a user-friendly, reliable, and human-validated lexicon-based tool optimized for interpreting sentiment in social media text [22]. VADER assigns each entry a compound score from −1 (most negative) to +1 (most positive). Based on established thresholds, entries were classified as positive (score >0.05), neutral (score between –0.05 and 0.05), or negative (score <–0.05).

Additional quantitative variables were derived from metadata collected at the time of data extraction and from theme frequency counts generated through the qualitative thematic categorization described above.

Analytic Strategy

Sentiment scores were analyzed across the entire dataset and within each thematic category. Due to the nonnormal distribution of these scores, the overall sentiment trend was summarized using the median and IQR. However, mean sentiment scores were reported for individual themes to facilitate interpretation and cross-theme comparisons.

To assess the potential impact of the COVID-19 pandemic, entries were temporally stratified based on the WHO pandemic declaration date (March 11, 2020). Entries posted before this date were designated pre–COVID-19, whereas those posted on or after this date were designated post COVID-19. Posting volume per year was also determined. User engagement was assessed by calculating the median response time (time elapsed from submission of an original post to the first collected comment) across all posts within the included subreddits, including both intentional and incidental entries.

Chi-square tests of independence were used to compare theme frequencies between the pre– and post–COVID-19 periods. Standardized residuals (SRs >|2|) were used to identify themes that were significantly overrepresented or underrepresented in each period. Temporal trends in annual posting volume were modeled using Poisson regression, with additional models stratified by pre– vs post–COVID-19 periods. Given the nonparametric distribution of sentiment scores and variability in group sizes, Mann-Whitney *U* tests were used to compare median sentiment overall and within each theme before vs after the pandemic. A linear regression model assessed whether thematic category predicted sentiment score, using "Factual Claims" as the reference due to its prevalence and informational neutrality.

There were no missing data in the final analytic dataset; therefore, no missing completely at random testing or imputation was required. All analyses were conducted in RStudio (version 2024.12.1+563), and statistical significance was defined as $P<.05$.

Results

Overview

A total of 4235 Reddit entries referencing the HPV vaccine were collected between September 13, 2016, and February 13, 2025. Of those, 2803 were classified as intentional posts or comments. After excluding 2 entries from 2011 that fell outside the defined study period, 2801 intentional entries were included in the final analysis, including 614 posts and 2187 comments.

Factual claims were most common (703/2801, 25.1%), followed by personal experiences (573/2801, 20.4%) and advice offering (466/2801, 16.6%) (Table 3). Overall, 1447 of 2801 (51.7%) entries exhibited positive sentiment (95% CI 49.8% - 53.5%), 1052 of 2801 (37.6%) negative, and 302 of 2801 (10.8%) neutral. The median sentiment score across all entries was 0.13 (IQR –0.49 to 0.71). At the theme level, advice seeking had the highest mean sentiment (0.17, SD 0.62), whereas "Other" entries were slightly negative (–0.04, SD 0.65).

XSL•FO
RenderX

Table . Frequencies of thematic categories and mean sentiment scores in Reddit discussions about the human papillomavirus vaccine before and after the World Health Organization COVID-19 pandemic declaration (2016 - 2025).[a]

Theme	Pre–COVID-19, n (%)[b]	Pre–COVID-19 sentiment, mean (SD)	Post–COVID-19, n (%)	Post–COVID-19 sentiment, mean (SD)	Total, n (%)	Total sentiment, mean (SD)
Advice or recommendations (offering)	34 (12.2)	0.05 (0.69)	432 (17.1)	0.12 (0.63)	466 (16.6)	0.12 (0.64)
Advice or recommendations (seeking)	5 (1.8)	−0.09 (0.30)	148 (5.9)	0.18 (0.63)	153 (5.5)	0.17 (0.62)
Barriers or accessibility	52 (18.7)	0.03 (0.71)	168 (6.7)	0.12 (0.63)	220 (7.8)	0.10 (0.65)
Factual claims	71 (25.5)	0.00 (0.62)	632 (25)	0.09 (0.65)	703 (25.1)	0.08 (0.65)
General questions	16 (5.8)	−0.01 (0.56)	199 (7.9)	0.08 (0.69)	215 (7.7)	0.07 (0.68)
Hesitancy	9 (3.2)	−0.27 (0.52)	132 (5.2)	0.09 (0.65)	141 (5)	0.07 (0.64)
Personal experiences	37 (13.3)	0.08 (0.62)	536 (21.2)	0.14 (0.64)	573 (20.4)	0.13 (0.64)
Resources or data sharing	21 (7.6)	0.05 (0.70)	161 (6.4)	0.10 (0.66)	182 (6.5)	0.09 (0.67)
Social or cultural influences	28 (10.1)	−0.17 (0.73)	62 (2.5)	0.14 (0.68)	90 (3.2)	0.04 (0.71)
Other	5 (1.8)	0.11 (0.67)	53 (2.1)	−0.05 (0.65)	58 (2.1)	−0.04 (0.65)
Total	278	0.00 (0.65)	2523	0.11 (0.65)	2801	0.10 (0.65)

[a]This table reports the total number of entries within each of the ten manually coded themes, along with theme-specific frequencies and mean (SD) sentiment scores stratified by time period, in a retrospective cross-sectional social media content analysis of Reddit posts and comments. Data include intentional human papillomavirus vaccine–related entries collected from publicly available Reddit discussions between September 2016 and February 2025 (N=2801). Pre–COVID-19 entries were defined as those posted before March 11, 2020, and post–COVID-19 entries as those posted on or after this date.

[b]Counts and percentages represent the total number of entries per theme in each period. Mean sentiment scores (SD) are shown alongside their respective theme frequencies.

Pre– vs Post–COVID-19 Volume, Themes, and Sentiment

Temporal analysis revealed increasing engagement with HPV vaccine topics over time (Figure 2). Annual post counts rose from 12 in 2016 to 132 in 2018 and then accelerated after 2021, peaking at 1337 in 2024 (Figure 2). Early 2025 data included 563 entries through February 13, suggesting that posting volume in 2025 is on pace to surpass 2024. A Poisson regression model assessing annual post and comment volume from 2016 to 2025 confirmed a significant upward trend. Each additional year was associated with a log-linear increase in count (β=.41, SE=0.01; IRR=1.50, 95% CI 1.47 - 1.53; *P*<.001), corresponding to a 50% increase in activity per year. When stratified, the pre–COVID-19 period showed a 13% annual increase (β=.12, SE=0.04; IRR=1.13, 95% CI 1.04 - 1.23; *P*=.01), while the post–COVID-19 period demonstrated a steeper 60% annual increase (β=.47, SE=0.01; IRR=1.60, 95% CI 1.56 - 1.65; *P*<.001).

Of the 2801 intentional entries, 278 occurred pre–COVID-19 (September 2016 to March 2020) and 2523 post–COVID-19 (April 2020 to February 2025). Theme distributions differed significantly pre– vs post–COVID-19 periods (χ^2_9=114.47, *P*<.001). SRs indicated that prepandemic entries were overrepresented in barriers or accessibility (SR=7.09) and social or cultural influences (SR=6.83), whereas postpandemic entries were overrepresented in personal experiences (SR=3.11), advice seeking (SR=2.83), and advice offering (SR=2.08).

XSL•FO
RenderX

Figure 2. Annual volume of Reddit posts and comments intentionally referencing the human papillomavirus (HPV) vaccine, September 2016-February 2025. This figure depicts the yearly count of publicly available Reddit posts and comments containing HPV vaccine–related keywords collected for this retrospective cross-sectional social media content analysis. Data include intentional entries collected from multiple subreddits between September 13, 2016, and February 13, 2025. Bars represent the total annual volume of intentional entries. A sharp increase in activity is observed beginning in 2021 and peaking in 2024, reflecting growing online engagement with HPV vaccine discourse over time. Importantly, counts for 2025 include only data through February 13 and already approach 600 entries, indicating that activity is on pace to surpass 2024 by the end of the year.

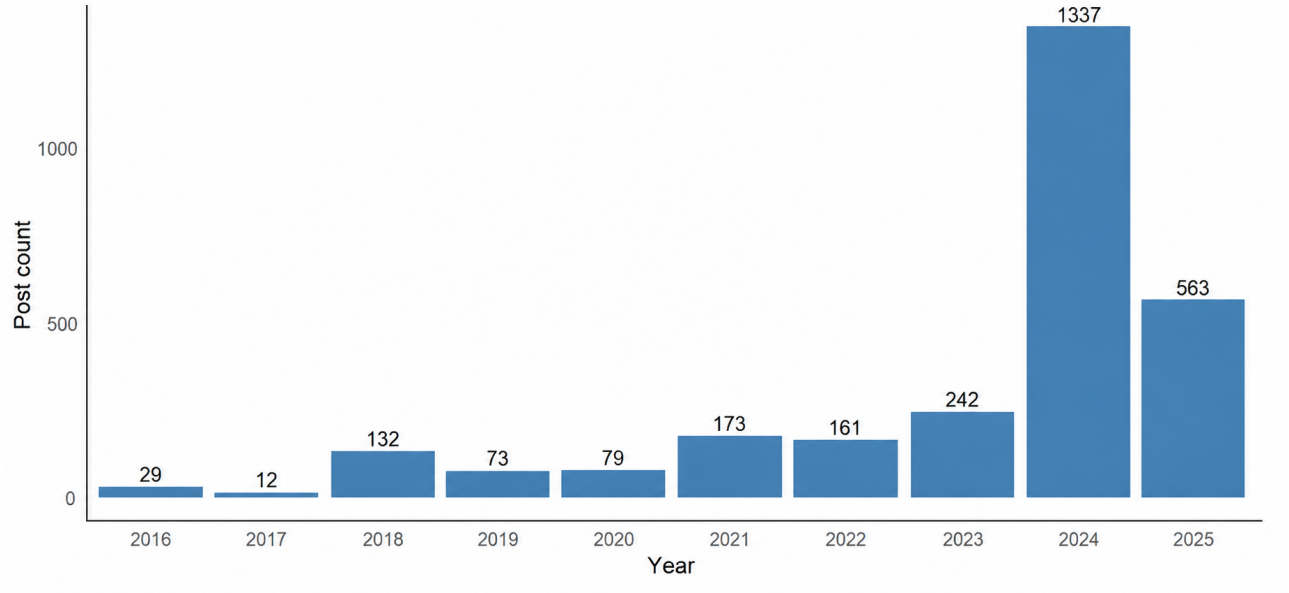

Sentiment also differed across periods. Pre–COVID-19 entries were 44.6% (124/278) positive (95% CI 38.7% - 50.7%), 43.9% (122/278) negative, and 11.5% (32/278) neutral sentiment, with a median score of 0.00 (–0.66 to 0.58). Post–COVID-19 sentiment saw an increase in positive sentiment, with entries being 52.4% (1324/2523) positive (95% CI 50.5% - 54.4%), 36.9% (931/2523) negative, and 10.7% (268/2523) neutral, with a median score of 0.17 (IQR –0.46 to 0.72). A Mann-Whitney *U* test confirmed a statistically significant increase in sentiment scores post–COVID-19 (W=384615, *P*=.01), suggesting a shift toward a more positive tone in HPV vaccine discussions since the onset of the pandemic.

When sentiment changes were assessed within individual themes, no category-specific comparisons reached statistical significance (all *P*>.05). However, some themes demonstrated notable shifts. For example, vaccine hesitancy changed from a mean of –0.27 to +0.09 (*P*=.09), and social or cultural influences shifted from –0.17 to 0.14 (*P*=.08). Although these changes were not statistically significant, they may indicate evolving tone and engagement, possibly reflecting more constructive or reflective dialog in the post–COVID-19 period.

Linear Regression of Themes Predicting Sentiment Scores

The linear regression model examining theme as a predictor of sentiment score was not statistically significant ($F_{9,2791}$=0.98, *P*=.45), indicating that thematic category did not significantly predict sentiment score (R^2=0.003) when controlling for other themes using “Factual claims” as the reference category. No individual theme coefficient was statistically significant.

Median Response Time

To assess user interaction and determine whether Reddit serves as a reliable platform for receiving timely responses to vaccine-related inquiries, we analyzed the median time from post to first comment. The overall median response time was 136.1 (IQR 36.0-613.1) minutes, suggesting that users typically received replies within a few hours. However, engagement varied by subreddit. Among the most active communities, r/HPV (850 unique posts) had a median response time of 148.2 (IQR 36.0-613.6) minutes, r/PreCervicalCancer (110 posts) had a response time of 215.6 (IQR 44.4-1184.1) minutes, and r/VACCINES (102 posts) showed a faster response time of 109.8 (IQR 36.1-501.9) minutes. Additional subreddits included r/Vaccine (29 posts), with a median time of 310.1 (IQR 62.1-1053.4) minutes; r/CervicalCancer (18 posts), with a median response time of 94.1 (IQR 37.3-180.6) minutes; r/DebateVaccines (17 posts), with a median response time of 115.7 (IQR 21.3-472.0) minutes; r/vaxxhappened (13 posts), with a median response time of 417.9 (IQR 74.3-748.3) minutes; and r/AskDocs (10 posts), with a median response time of 77.6 (IQR 54.6-80.0) minutes. Subreddits with fewer than 10 unique posts are summarized in Table 1 and were excluded from direct comparisons due to small sample size. This response time analysis included all collected entries, including both intentional and incidental entries, for posts initiating threads.

Discussion

Limitations

This study analyzed 2801 Reddit posts and comments to characterize public discourse surrounding the HPV vaccine, identify dominant themes and sentiment, and evaluate how these patterns changed following the start of the COVID-19 pandemic. Factual claims, personal experiences, and advice offering were the most common themes, and overall sentiment was more positive than negative, with a median sentiment score of 0.13. Posting volume increased substantially over time, rising by 50%

XSL•FO
RenderX

per year overall and more rapidly after 2020. Post–COVID-19 entries demonstrated significantly higher positive sentiment and a shift toward more advice seeking and personal experience–based discussions. Confidence interval patterns aligned with sample size distribution, with estimates for the pre–COVID-19 period being less precise due to smaller sample counts, whereas overall and post–COVID-19 estimates had narrow confidence intervals, reflecting high precision and more stable inference. These findings indicate that HPV vaccine conversations on Reddit have become more frequent, more positive in tone, and increasingly peer support driven.

Additionally, with an overall median post-to-first-comment time of just over 2 hours and subreddit communities with member counts in the tens or hundreds of thousands, users can quickly ask questions and receive feedback. This responsiveness is particularly relevant in the post–COVID-19 era, during which patient portal messages increased by 157%, potentially causing delays in direct communication with health care providers [37]. The marked rise in HPV vaccine discussions after 2020 (Figure 2) may therefore reflect a growing demand for timely, community-driven health information amid pandemic-related uncertainty. Because Reddit engagement is high and sentiment is increasingly supportive, it represents a critical and underutilized method for targeted interventions designed to address vaccine hesitancy, spread accurate information, and improve HPV vaccination uptake.

The COVID-19 pandemic created heightened vaccine awareness and public debate [38], which likely shaped the tone and content of HPV vaccine discussions observed in this study. While some studies report a decline in vaccine confidence post–COVID-19 [39], a recent literature review found increased uptake of routine adult vaccinations, such as influenza and pneumococcal vaccines, alongside decreased childhood vaccination rates [40]. Similarly, a cross-sectional survey showed that most participants had no change in opinion regarding the HPV vaccine following the pandemic [41]. Together, these contrasting findings underscore the complex and evolving nature of public attitudes toward vaccination across populations and vaccine types. The broad discussions surrounding vaccines likely influenced the volume and sentiment of discussions in this study. Heightened public uncertainty surrounding vaccines may have prompted more users to seek answers and advice on Reddit, along with other social media [42]. The increase in sentiment observed in the study was somewhat unexpected, as heightened skepticism and polarization during the pandemic might have suggested a decline in favorable attitudes [42]. One explanation is that oppositional views may have been redirected toward COVID-specific subreddits during this time, leaving HPV-focused discussions to adopt a more supportive and informational tone. Additionally, users in these subreddits may have felt more supportive and engaged in more constructive discussions surrounding the vaccine during a period marked by broader health uncertainty.

The shift toward more experience-driven content and increasingly positive sentiment may point to Reddit's growing role as an informal support network and a disseminator of health information [43]. The results suggest that individuals use the platform not only to access information but also to validate their concerns, share lived experiences, and support others navigating vaccine decisions. This trend is further reflected in the growing volume of HPV vaccine–related posts over time. Poisson regression modeling showed a 50% annual increase in post volume from 2016 to 2025, with a sharper rise post–COVID-19 (60% per year) compared to pre–COVID-19 (13% per year). However, these findings should be interpreted alongside broader changes in the platform itself. Reddit's user demographics, community structures, and overall engagement have evolved considerably over this period [35]. External sociopolitical events, such as the 2020 US presidential election, which coincided near our pandemic cutoff, may also have influenced online vaccine discourse [42]. Considering these factors, public health professionals may benefit from monitoring Reddit and similar platforms to identify emerging concerns, counter misinformation, and tailor outreach strategies to align with community-driven dialog.

However, the unmoderated nature of many social media discussions allows misinformation to spread rapidly. For instance, a prior 2021 study identified misinformation in 25.6% of HPV vaccine–related Reddit posts [31]. These falsehoods can complicate public health messaging and create confusion among users. More studies similar to the work of Du et al [31] should explore subreddit-specific misinformation trends to better understand the nature of these inaccuracies and to support the development of targeted interventions. For example, Reddit users often frame their concerns in personal or culturally specific terms [44]. Public health campaigns may benefit from tailored messaging that acknowledges these concerns and shares relatable narratives or culturally competent information [44]. Interventions might focus especially on younger adults, who make up a large proportion of Reddit users and who may still be within vaccine-eligible age groups [35].

Another notable finding was the shift in sentiment within the vaccine hesitancy and social or cultural influence themes. Although not statistically significant, the directional trend may still hold public health relevance. Both categories shifted from negative sentiment pre–COVID-19 to positive sentiment post–COVID-19, resulting in an overall neutral average. Although individual posts were not examined in detail, this shift may suggest a transition from overt skepticism to more reflective or constructive dialog. For example, users may have become more likely to express uncertainty rather than opposition, share stories about overcoming hesitancy, or describe how they navigate familial or cultural pressures. These evolving dynamics within Reddit HPV discussions suggest a possible shift toward more open and solution-oriented conversations. While this trend warrants further investigation across other platforms, it represents a meaningful opportunity for targeted public health messaging.

Interestingly, the proportion of positive sentiment (n=1447, 51.7%) approximates recent US HPV vaccination rates, with 66% of adolescents completing the full vaccine series by age 17 [8]. Although demographic data were not collected in this study, prior research shows that Reddit users are predominantly younger than 35 years, with one study citing a mean user age of 25.7 years, and most reside in English-speaking countries [31]. This demographic overlap strengthens the case for using

XSL•FO
RenderX

Reddit as a window into the attitudes of vaccine-eligible populations and underscores its value in identifying motivators or barriers to vaccine uptake.

Our findings expand on prior research examining HPV vaccine discourse online. Lama et al [30] analyzed Reddit posts from 2007 to 2015 and identified dominant topics such as political debate, sexual behavior, and vaccination schedules using automated topic modeling. While their study provided early insights into Reddit-based vaccine discussions, it did not assess sentiment or examine temporal shifts in discourse. In addition, Du et al [31] focused specifically on identifying misinformation related to the HPV vaccine on Reddit, offering insight into the prevalence of inaccurate claims within public discourse but without thematic categorization or sentiment analysis.

More recently, Meci et al [32] examined sentiment in Reddit posts but limited their analysis to only 3 broad themes—factual, personal, and conspiracy—and excluded user comments. Their study focused on discourse surrounding US Food and Drug Administration guidance encouraging HPV vaccination for head and neck cancer prevention. In contrast, the present study analyzed both posts and comments, employed manual coding across 10 distinct themes for greater granularity, and specifically investigated temporal shifts in theme distribution and sentiment surrounding the COVID-19 pandemic. This multifaceted approach offers a more nuanced and contemporary understanding of online vaccine discourse.

Overall, this study provides an innovative post–COVID-19 examination of public HPV vaccine discourse by integrating qualitative thematic analysis with quantitative sentiment and temporal trends on Reddit. Unlike prior studies that relied on prepandemic data or narrower thematic scope, this analysis offers a more detailed and contemporary view by examining both posts and comments across 10 distinct themes using manual labeling to enhance contextual accuracy. The findings demonstrate a shift toward more personal, advice-oriented, and supportive engagement following the pandemic, contributing to the growing literature on digital vaccine communication. Practically, these findings highlight the value of social media platforms such as Reddit for targeted public health outreach, identification of emerging concerns, and strategies to counter misinformation while supporting efforts to promote HPV vaccine uptake.

Limitations

This study has several limitations. First, selection and reporting bias are inherent to social media research, which limits generalizability. Although prior research suggests that Reddit users are predominantly younger than 35 years and primarily located in English-speaking countries, the platform's anonymity makes it difficult to draw definitive conclusions about user demographics [31,40]. Additionally, while this age group overlaps with key HPV vaccine–eligible populations, the findings may not fully capture the sentiments of all decision-makers, such as parents of adolescents or older adults up to age 45, potentially limiting the transferability of conclusions to these specific subpopulations. Second, Reddit is a worldwide platform, and some discussions may reference non-US health care systems. Countries may vary in the amount of discussion on the HPV vaccine, and differences in background and access barriers may shape discussions in ways that do not reflect the US context. These differences reduce external validity and limit generalizability to US populations. Third, the study's cross-sectional and retrospective design limits causal inference. User perspectives cannot be followed longitudinally, and shifts in discussion cannot be attributed to specific events with certainty. Additionally, some subreddits archive older posts, which may have contributed to the lower volume of retrieved entries during the earlier years of the study period. Fourth, thematic categorization required assigning each entry to a single dominant theme, which may oversimplify entries containing multiple topics. Although 2 independent coders resolved discrepancies by consensus to maximize consistency and reliability, some nuances may inevitably be lost. Finally, the reliance on an automated tool (VADER) for sentiment analysis presents limitations. While efficient for large datasets and social media, automated scoring may fail to capture sarcasm, medical jargon, context-dependent nuances, or complex emotional expressions accurately.

Future Directions

Future research could employ more advanced NLP techniques (eg, topic modeling within themes and stance detection) to further dissect nuanced discussion threads. This approach could enable more fine-grained identification of specific concerns, misinformation patterns, or information gaps within each theme. In turn, such insights could better inform public health campaigns and digital outreach efforts aimed at improving vaccine uptake, trust, and health literacy in online spaces.

Additionally, future studies could explore whether awareness of HPV-related oropharyngeal cancers, particularly in men, is increasing over time through social media discourse. Prior survey research has documented low public awareness of the link between HPV and oropharyngeal cancer in men [45]. Because HPV is often associated with cervical cancer, many men may not perceive the vaccine as relevant to them, contributing to lower uptake and greater vaccine hesitancy in this group. A broader assessment of these knowledge gaps across social media platforms could support the development of targeted, gender-specific educational campaigns. Tailored messaging that addresses misconceptions and highlights the relevance of HPV vaccination for men may be critical to improving coverage and informed decision-making in vaccine-hesitant populations.

Conclusion

This study provides a contemporary analysis of HPV vaccine discourse and sentiment on Reddit, highlighting significant shifts in both thematic focus and sentiment before and after the COVID-19 pandemic. Findings suggest that conversations have become increasingly personal and advice-driven over time, with a notable rise in positive sentiment and public engagement following 2020. These trends reflect a growing reliance on online communities for health-related dialogue and peer support, particularly within a digital landscape transformed by the pandemic. Reddit appears to serve not only as a platform for information exchange but also as a space for collective decision-making, emotional validation, and rapid feedback,

XSL•FO
RenderX

especially in the context of health care system delays and uncertainty. These insights underscore the importance of public health professionals monitoring and potentially engaging with evolving online conversations to promote accurate vaccine information and respond to emerging concerns in real time. This work builds on prior research while offering new perspectives on the dynamic, community-driven nature of digital health communication in the social media era.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: code generation, code optimization, creation of algorithms for data analysis. The GAI tool used was ChatGPT (GPT-5.2). Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted under collective responsibility. ChatGPT (GPT-5.2) was used for the development of the Python script used for data extraction and the R code used for statistical analysis.

Funding

No external funding was received for the conduct of this study. The authors acknowledge that the applicable article processing fee was covered by the Department of Otolaryngology-Head and Neck Surgery at Rush University Medical Center.

Data Availability

All data generated or analyzed during this study are included in this published article and the multimedia appendices.

Authors' Contributions

Conceptualization: MM, SAS, MKB
Data curation: MM, SAS
Formal analysis: MM, SAS
Investigation: MM, SAS, DB
Methodology: MM, SAS
Project administration: MM, SAS, MKB
Software: SAS, MM
Supervision: MKB
Validation: MM, SAS, MKB
Visualization: MM
Writing – original draft: MM, SAS, DB, JM, RP
Writing – review & editing: EMS, VCS, MKB, JM, RP

Conflicts of Interest

None declared.

Multimedia Appendix 1
Microsoft Excel spreadsheet containing all posts and comments scraped, along with their labels and sentiment scores.
[XLSX File, 3896 KB - jmir_v28i1e83558_app1.xlsx]

Multimedia Appendix 2
Natural language processing code used for sentiment analysis.
[PDF File, 168 KB - jmir_v28i1e83558_app2.pdf]

Multimedia Appendix 3
Python script used to scrape data from Reddit.
[PDF File, 100 KB - jmir_v28i1e83558_app3.pdf]

References

1. Szymonowicz KA, Chen J. Biological and clinical aspects of HPV-related cancers. Cancer Biol Med 2020 Nov 15;17(4):864-878. [doi: 10.20892/j.issn.2095-3941.2020.0370] [Medline: 33299640]
2. Guo Y, Liu X, Nicholas S, Maitland E, Liu R. Revisiting global HPV vaccination behavior and its determinants: a comprehensive review. Risk Manag Healthc Policy 2025;18:2675-2689. [doi: 10.2147/RMHP.S524223] [Medline: 40831587]

XSL•FO
RenderX

3. Chabeda A, Yanez RJR, Lamprecht R, Meyers AE, Rybicki EP, Hitzeroth II. Therapeutic vaccines for high-risk HPV-associated diseases. Papillomavirus Res 2018 Jun;5:46-58. [doi: 10.1016/j.pvr.2017.12.006] [Medline: 29277575]
4. Egawa N, Doorbar J. The low-risk papillomaviruses. Virus Res 2017 Mar 2;231:119-127. [doi: 10.1016/j.virusres.2016.12.017] [Medline: 28040475]
5. Ljubojević S. The human papillomavirus vaccines. Acta Dermatovenerol Croat 2006;14(3):208. [Medline: 17010274]
6. Harper DM, DeMars LR. HPV vaccines - a review of the first decade. Gynecol Oncol 2017 Jul;146(1):196-204. [doi: 10.1016/j.ygyno.2017.04.004] [Medline: 28442134]
7. Lehtinen M, Lagheden C, Luostarinen T, et al. Human papillomavirus vaccine efficacy against invasive, HPV-positive cancers: population-based follow-up of a cluster-randomised trial. BMJ Open 2021 Dec 30;11(12):e050669. [doi: 10.1136/bmjopen-2021-050669] [Medline: 35149535]
8. Pingali C, Yankey D, Chen M, et al. National vaccination coverage among adolescents aged 13-17 years - national immunization survey-teen, United States, 2023. MMWR Morb Mortal Wkly Rep 2024 Aug 22;73(33):708-714. [doi: 10.15585/mmwr.mm7333a1] [Medline: 39173168]
9. Cervical cancer elimination initiative. World Health Organization. 2025 Jan 29. URL: https://www.who.int/initiatives/cervical-cancer-elimination-initiative [accessed 2026-04-17]
10. Saulsberry L, Fowler EF, Nagler RH, Gollust SE. Perceptions of politicization and HPV vaccine policy support. Vaccine (Auckl) 2019 Aug 14;37(35):5121-5128. [doi: 10.1016/j.vaccine.2019.05.062] [Medline: 31296376]
11. Heyde S, Osmani V, Schauberger G, Cooney C, Klug SJ. Global parental acceptance, attitudes, and knowledge regarding human papillomavirus vaccinations for their children: a systematic literature review and meta-analysis. BMC Womens Health 2024 Sep 27;24(1):537. [doi: 10.1186/s12905-024-03377-5] [Medline: 39334328]
12. Davies C, Stoney T, Hutton H, et al. School-based HPV vaccination positively impacts parents' attitudes toward adolescent vaccination. Vaccine (Auckl) 2021 Jul 5;39(30):4190-4198. [doi: 10.1016/j.vaccine.2021.05.051] [Medline: 34127299]
13. Tobaiqy M, MacLure K. A systematic review of human papillomavirus vaccination challenges and strategies to enhance uptake. Vaccines (Basel) 2024 Jul 6;12(7):746. [doi: 10.3390/vaccines12070746] [Medline: 39066384]
14. Kornides ML, Badlis S, Head KJ, Putt M, Cappella J, Gonzalez-Hernadez G. Exploring content of misinformation about HPV vaccine on Twitter. J Behav Med 2023 Apr;46(1-2):239-252. [doi: 10.1007/s10865-022-00342-1] [Medline: 35896853]
15. Goldstone SE, Giuliano AR, Palefsky JM, et al. Efficacy, immunogenicity, and safety of a quadrivalent HPV vaccine in men: results of an open-label, long-term extension of a randomised, placebo-controlled, phase 3 trial. Lancet Infect Dis 2022 Mar;22(3):413-425. [doi: 10.1016/S1473-3099(21)00327-3] [Medline: 34780705]
16. Fowler EF, Nagler RH, Banka D, Gollust SE. Effects of politicized media coverage: experimental evidence from the HPV vaccine and COVID-19. Prog Mol Biol Transl Sci 2022;188(1):101-134. [doi: 10.1016/bs.pmbts.2021.11.008] [Medline: 35168740]
17. Hoang TN, Agana-Norman DFG, Rizvi AS, et al. Provider perspectives on HPV vaccine hesitancy in the post-COVID-19 era. Hum Vaccin Immunother 2025 Dec;21(1):2550054. [doi: 10.1080/21645515.2025.2550054] [Medline: 41503766]
18. Joshi R, Kilinsky A. HPV vaccine hesitancy in the United States. Curr Opin Pediatr 2025 Apr 1;37(2):198-204. [doi: 10.1097/MOP.0000000000001441] [Medline: 39882684]
19. Du J, Luo C, Shegog R, et al. Use of deep learning to analyze social media discussions about the human papillomavirus vaccine. JAMA Netw Open 2020 Nov 2;3(11):e2022025. [doi: 10.1001/jamanetworkopen.2020.22025] [Medline: 33185676]
20. Villanueva-Miranda I, Xie Y, Xiao G. Sentiment analysis in public health: a systematic review of the current state, challenges, and future directions. Front Public Health 2025;13:1609749. [doi: 10.3389/fpubh.2025.1609749] [Medline: 40620557]
21. Gandy LM, Ivanitskaya LV, Bacon LL, Bizri-Baryak R. Public health discussions on social media: evaluating automated sentiment analysis methods. JMIR Form Res 2025 Jan 8;9:e57395. [doi: 10.2196/57395] [Medline: 39773420]
22. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proc Int AAAI Conf Web Soc Media 2014 May;8(1):216-225. [doi: 10.1609/icwsm.v8i1.14550]
23. Shapiro GK, Surian D, Dunn AG, Perry R, Kelaher M. Comparing human papillomavirus vaccine concerns on Twitter: a cross-sectional study of users in Australia, Canada and the UK. BMJ Open 2017 Oct 5;7(10):e016869. [doi: 10.1136/bmjopen-2017-016869] [Medline: 28982821]
24. Du J, Xu J, Song HY, Tao C. Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data. BMC Med Inform Decis Mak 2017 Jul 5;17(Suppl 2):69. [doi: 10.1186/s12911-017-0469-6] [Medline: 28699569]
25. Dunn AG, Leask J, Zhou X, Mandl KD, Coiera E. Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: an observational study. J Med Internet Res 2015 Jun 10;17(6):e144. [doi: 10.2196/jmir.4343] [Medline: 26063290]
26. Surian D, Nguyen DQ, Kennedy G, Johnson M, Coiera E, Dunn AG. Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection. J Med Internet Res 2016 Aug 29;18(8):e232. [doi: 10.2196/jmir.6045] [Medline: 27573910]
27. Dunn AG, Surian D, Leask J, Dey A, Mandl KD, Coiera E. Mapping information exposure on social media to explain differences in HPV vaccine coverage in the United States. Vaccine (Auckl) 2017 May 25;35(23):3033-3040. [doi: 10.1016/j.vaccine.2017.04.060] [Medline: 28461067]

XSL•FO
RenderX

28. Zhou X, Coiera E, Tsafnat G, Arachi D, Ong MS, Dunn AG. Using social connection information to improve opinion mining: identifying negative sentiment about HPV vaccines on Twitter. Stud Health Technol Inform 2015;216:761-765. [Medline: 26262154]
29. Luo X, Zimet G, Shah S. A natural language processing framework to analyse the opinions on HPV vaccination reflected in Twitter over 10 years (2008-2017). Hum Vaccin Immunother 2019;15(7-8):1496-1504. [doi: 10.1080/21645515.2019.1627821] [Medline: 31194609]
30. Lama Y, Hu D, Jamison A, Quinn SC, Broniatowski DA. Characterizing trends in human papillomavirus vaccine discourse on Reddit (2007-2015): an observational study. JMIR Public Health Surveill 2019 Mar 27;5(1):e12480. [doi: 10.2196/12480] [Medline: 30916662]
31. Du J, Preston S, Sun H, et al. Using machine learning-based approaches for the detection and classification of human papillomavirus vaccine misinformation: infodemiology study of Reddit discussions. J Med Internet Res 2021 Aug 5;23(8):e26478. [doi: 10.2196/26478] [Medline: 34383667]
32. Meci A, Tseng CC, Jensen Z, Goyal N. Online discourse and trends surrounding HPV vaccination for head and neck cancer prevention. Otolaryngol Head Neck Surg 2025 Jun;172(6):1962-1970. [doi: 10.1002/ohn.1222] [Medline: 40105426]
33. Singh G, Dash NR, Shaju A, Chakkalakkunnan SS. Perceptions and sentiments associated with HPV vaccine uptake among Indian Reddit users: a qualitative social media analysis. BMC Public Health 2025 Nov 19;25(1):4037. [doi: 10.1186/s12889-025-25418-w] [Medline: 41257679]
34. Nagpal M, Jalali N, Sherifali D, Morita P, Cafazzo JA. Investigating Reddit data on type 2 diabetes management during the COVID-19 pandemic using latent Dirichlet allocation topic modeling and Valence Aware Dictionary for Sentiment Reasoning analysis: content analysis. JMIR Form Res 2025 Feb 21;9:e51154. [doi: 10.2196/51154] [Medline: 39983050]
35. Gjurković M, Karan M, Vukojević I, Bošnjak M, Snajder J. PANDORA talks: personality and demographics on reddit. Presented at: Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media; Jun 10, 2021. [doi: 10.18653/v1/2021.socialnlp-1.12]
36. Mixed methods article reporting standards (MMARS) information recommended for inclusion in manuscripts that report the collection and integration of qualitative and quantitative data. : American Psychological Association; 2020 URL: https://apastyle.apa.org/jars/mixed-table-1.pdf [accessed 2026-04-17]
37. Holmgren AJ, Downing NL, Tang M, Sharp C, Longhurst C, Huckman RS. Assessing the impact of the COVID-19 pandemic on clinician ambulatory electronic health record use. J Am Med Inform Assoc 2022 Jan 29;29(3):453-460. [doi: 10.1093/jamia/ocab268] [Medline: 34888680]
38. Malagoli LG, Stancioli J, Ferreira CHG, Vasconcelos M, Couto da Silva AP, Almeida JM. A look into COVID-19 vaccination debate on twitter. : ACM Presented at: WebSci '21: Proceedings of the 13th ACM Web Science Conference 2021; Jun 21-25, 2021; Virtual Event United Kingdom p. 225-233. [doi: 10.1145/3447535.3462498]
39. Siani A, Tranter A. Is vaccine confidence an unexpected victim of the COVID-19 pandemic? Vaccine (Auckl) 2022 Nov;40(50):7262-7269. [doi: 10.1016/j.vaccine.2022.10.061]
40. Vojtek I, van Wouw M, Thomson A. Impact of COVID-19 on vaccine confidence and uptake: a systematic literature review. Hum Vaccin Immunother 2024 Dec 31;20(1):2384180. [doi: 10.1080/21645515.2024.2384180] [Medline: 39106971]
41. Schelbar N, Ward CN, Phillips E, et al. Impact of COVID-19 pandemic and vaccine perceptions on HPV vaccine hesitancy. Am J Otolaryngol 2024;45(2):104172. [doi: 10.1016/j.amjoto.2023.104172] [Medline: 38103489]
42. Park B, Jang IS, Kwak D. Sentiment analysis of the COVID-19 vaccine perception. Health Informatics J 2024;30(1):14604582241236131. [doi: 10.1177/14604582241236131] [Medline: 38403926]
43. Record RA, Silberman WR, Santiago JE, Ham T. I sought it, I Reddit: examining health information engagement behaviors among Reddit users. J Health Commun 2018;23(5):470-476. [doi: 10.1080/10810730.2018.1465493] [Medline: 29718799]
44. Hartzler A, Pratt W. Managing the personal side of health: how patient expertise differs from the expertise of clinicians. J Med Internet Res 2011 Aug 16;13(3):e62. [doi: 10.2196/jmir.1728] [Medline: 21846635]
45. Davis EN, Doyle PC. An assessment of young adults' awareness and knowledge related to the human papillomavirus (HPV), oropharyngeal cancer, and the HPV vaccine. Cancers (Basel) 2025 Jan 21;17(3):344. [doi: 10.3390/cancers17030344] [Medline: 39941716]

Abbreviations

HPV: human papillomavirus
NLP: natural language processing
SR: standardized residual
VADER: Valence Aware Dictionary and Sentiment Reasoner

XSL•FO
RenderX

Edited by S Brini; submitted 05.Sep.2025; peer-reviewed by C Ahmadu, J Guo; revised version received 02.Mar.2026; accepted 03.Mar.2026; published 19.May.2026.

Please cite as:
Marciniak M, Setzen SA, Bhattacharya D, Masliah J, Powszok R, Sturgis EM, Stubbs VC, Bhayani MK
Human Papillomavirus Vaccine Discourse and Sentiment on Reddit Before and After COVID-19: Mixed Methods Retrospective Cross-Sectional Study
J Med Internet Res 2026;28:e83558
URL: https://www.jmir.org/2026/1/e83558
doi:10.2196/83558

© Marcin Marciniak, Sean A Setzen, Deeya Bhattacharya, Jamie Masliah, Robin Powszok, Erich M Sturgis, Vanessa C Stubbs, Mihir K Bhayani. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Interpretations of Menstrual Blood Appearance and Diagnostic Potential Among Social Media Users: Cross-Sectional Mixed Methods Social Media Listening Study

Ruth Niedermeier[1], Dr Med; Noemi Castelletti[1,2,3], PD, Dr rer nat; Andreas Bender[4,5], PhD; Michael Hoelscher[1,2,6,7], Prof Dr Med; Olena Ivanova[1,8], Dr med, PhD

[1]Institute of Infectious Diseases and Tropical Medicine, LMU University Hospital, LMU Munich, Leopoldstrasse 5, Munich, Germany

[2]Fraunhofer Institute for Translational Medicine and Pharmacology, ITMP, Immunology, Infection and Pandemic Research, Munich, Germany

[3]Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center, Johannes Gutenberg University, Mainz, Germany

[4]Department of Statistics, LMU Munich, Munich, Germany

[5]Munich Center for Machine Learning (MCML), LMU Munich, Munich, Germany

[6]German Centre for Infection Research (DZIF), partner site Munich, Munich, Germany

[7]Unit Global Health, Helmholtz Center Munich, German Research Center for Environmental Health (HMGU), Neuherberg, Germany

[8]SPHERE, NHMRC Centre of Research Excellence, Department of General Practice, School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia

Corresponding Author:
Ruth Niedermeier, Dr Med
Institute of Infectious Diseases and Tropical Medicine, LMU University Hospital, LMU Munich, Leopoldstrasse 5, Munich, Germany

Abstract

Background: Menstruation has long been framed primarily as a hygiene issue, with mainstream products and public messaging emphasizing concealment and disposal of menstrual blood (MB). This has contributed to a culture of silence in which conversations about menstrual health have been marginalized in public and clinical settings. Recent international guidance, including the World Health Organization's call to reframe menstruation as a health issue, underscores the need for more open discourse. Simultaneously, social media has become a prominent space where menstruating individuals share experiences, seek advice, and challenge stigma. The resurgence of reusable menstrual products has increased users' direct observation of MB, prompting questions about variations in color, texture, and smell. These developments highlight growing curiosity about MB yet reveal persistent information gaps regarding how MB is understood outside the clinical setting.

Objective: This study aimed to examine how MB is represented in social media discourse and to explore individuals' perceptions of MB's potential use as a diagnostic tool.

Methods: We conducted a cross-sectional, convergent mixed methods social listening study combining qualitative content analysis, social network analysis, sentiment analysis, and descriptive statistical analysis. Data were collected from TikTok (ByteDance), Facebook (Meta), Instagram (Meta), and Reddit using Mention and Apify. Between February 1 and 28, 2025, 6263 posts and videos were extracted using 3 strategies—group searches, hashtag searches, and social listening alerts. All data were anonymized, and demographic information was unavailable. After removing duplicates, non-English content, images, and posts without reference to blood, 349 posts were included. Coding followed a multistep deductive process in Atlas.ti. All posts were assigned with quotations, which were designated with one or more codes. Network analysis examined associations between appearance descriptors and reported health conditions. Sentiment analysis assessed perceptions of MB-based diagnostics.

Results: Among the included posts (n=349), most originated from Reddit and Facebook. *Seeking help* (154/349, 44.1%) was the most common type of post. Appearance descriptions (n=243 posts) focused on color, particularly brown, bright red, pink, and black; consistency, particularly coagulation; and smell, mainly unpleasant. Network analysis linked specific colors and textures to perceived conditions, including miscarriage, endometriosis, hormonal changes, polycystic ovary syndrome (PCOS), and infections. Discussion of MB as a diagnostic tool (n=80 posts) was less frequent but included predominantly positive quotations (110/115, 95.7%), emphasizing accessibility, noninvasiveness, and home-based sampling. Concerns (19/115, 16.5%) focused on inclusivity, stigma, and bodily autonomy.

Conclusions: This study demonstrates that social media serves as an important source for discussion on MB-related topics and highlights a gap between public information needs and the available scientific evidence. The findings also indicate a strong interest in MB characteristics and support further research into its diagnostic potential. To our knowledge, this is the first study to analyze social media discussions on MB characteristics and its diagnostic potential.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e85550) doi:10.2196/85550

KEYWORDS

menstrual blood; social media; menstrual health; diagnostic; color

Introduction

Description of Research Problem

Menstruation has long been treated as a private hygiene issue, often framed in advertising and public discourse as something to be concealed and swiftly managed. Sanitary pads, tampons, and similar products have historically emphasized the removal and disposal of menstrual blood (MB), reinforcing a culture of silence and invisibility around menstruation. As a result, open conversations about menstrual health have long been marginalized, both in public discourse and within medical research [1].

This narrative is beginning to shift. In 2022, the World Health Organization identified 3 critical actions to advance menstrual health globally; first, to reframe menstruation as a health issue rather than a hygiene matter; second, to guarantee access to menstrual information and education; and third, to institutionalize menstrual health within governmental policies and national budgets [2]. These recommendations reflect an evolving understanding of menstruation as an important public health concern with implications for well-being, equity, and access to care.

Simultaneously, social media has become an increasingly visible space where menstruating individuals share and discuss their menstrual health in public forums. These platforms have become spaces where users share personal experiences, ask questions in moments of uncertainty, provide informal education, promote menstrual products, and challenge persistent taboos [3-7]. One significant shift contributing to this change is the resurgence of the menstrual cup, reintroduced by Mooncup in 2002 [8]. Unlike disposable products, menstrual cups allow users to observe their MB more closely, leading to greater awareness of its color, texture, and volume. This visibility often prompts questions such as “What is considered normal?” and “When should one be concerned?” [5].

Recent research on young people’s media encounters with menstruation shows that while advertising often minimizes menstrual pain and promotes a narrow, idealized portrayal of menstruation, the increasing visibility of menstruation across social and news media is simultaneously contributing to its normalization as a topic of public discussion. This ambivalence underscores social media’s dual role as both a site of stigma reproduction and a space for contesting dominant menstrual narratives [9].

Parallel to evolving societal perceptions of menstruation, medical and scientific interest in MB has increased substantially over the past decade. Historically regarded as biological waste and culturally stigmatized, menstruation is now increasingly recognized as a valuable source of biological information, reflecting both local endometrial processes and systemic physiological states [1]. Advances in analytical techniques have further enabled detailed characterization of MB, revealing its complex composition, including endometrial and immune cells, proteins, nucleic acids, lipids, extracellular vesicles, and stem or progenitor cells [10,11].

Against this background, MB is increasingly explored as a noninvasive diagnostic matrix for reproductive and systemic diseases. Studies have demonstrated the feasibility of using MB to monitor glycemic control in patients with diabetes, highlighting its potential to capture systemic metabolic biomarkers outside conventional blood sampling frameworks [12]. Similarly, passive self-collected MB has been successfully used for screening high-risk human papillomavirus, underscoring its applicability in population-level screening and self-sampling approaches [13]. More recently, technological innovation has accelerated this field, exemplified by the development of wearable in-pad diagnostic devices capable of real-time detection of disease-associated biomarkers directly from menstrual flow [14].

Beyond diagnostics, MB has also attracted attention for its regenerative and therapeutic potential. MB-derived stem cells have been extensively reviewed as a readily accessible, ethically uncontroversial cell source with promising applications in regenerative medicine, tissue engineering, and immunomodulation [15,16]. As research continues to expand, MB is increasingly recognized not only as a window into endometrial biology but also as a scalable, patient-centered biospecimen with the potential to transform approaches to disease detection, monitoring, and treatment.

Despite this growing scientific interest and increasing visibility of MB in biomedical research, little is currently known about how MB is discussed, interpreted, and made sense of by menstruating individuals outside of clinical and research settings. In this context, social media listening (SML) has emerged as a promising research method to explore public discourse around health-related topics [17]. Traditionally developed and applied by companies to monitor consumer sentiment, emerging trends, and competitor positioning across platforms, such as Facebook (Meta), Instagram (Meta), TikTok (ByteDance), Reddit, and X (formerly Twitter), SML analyses what users say and how they engage with specific topics online [18]. While rooted in commercial market research, SML has recently gained traction in the academic and medical fields as a way to access and analyze sentiments, opinions, and discourse among patients and specific social groups. For example, Carneiro et al [3] examined the discourse surrounding infertility on Instagram in Brazil; Le Busque and Mellish [19] analyzed the impact of Endometriosis Awareness Month on Instagram content; Pepin et al [20] explored the use of machine learning to analyze Reddit discussions about endometriosis; Metwally et al [21] investigated patient sentiment toward procedures like colposcopy, mammography, and Pap smears; and Sloeasen et al [22] studied how individuals with chronic ocular pain shared experiences across X, blogs, and online news platforms. These studies demonstrate the potential of SML to capture real-time,

XSL•FO
RenderX

patient-centered insights that may otherwise remain undocumented in clinical settings. While scientific interest in the composition and clinical value of MB is expanding, public discourse on this topic remains largely unexplored.

Study Aim and Objectives

In this context, this mixed methods study aims to explore how MB is discussed on social media, with two specific research questions (RQs): (1) How is the appearance of MB—such as color, texture, or smell—described and interpreted? (2) How is the potential use of MB as a diagnostic tool perceived and discussed?

Methods

Research Design Overview

The study used a cross-sectional, convergent mixed methods design, integrating qualitative and quantitative approaches to examine how MB is discussed on social media. This mixed methods research was chosen for studying complex health-related discourse, as it allows for both in-depth interpretation of meanings and systematic analysis of patterns and associations [23].

Qualitative and quantitative data were collected from the same social media platforms during a single extraction period and analyzed in parallel. Findings from qualitative content analysis were integrated with results from descriptive statistical analysis, social network analysis, and sentiment analysis during interpretation. The qualitative objective was to explore how menstruating individuals describe and interpret the appearance of MB, including color, texture, and smell, in social media discourse and how individuals link MB characteristics to specific menstrual health conditions. The quantitative objective was to examine patterns in these descriptions and to assess sentiment and relational associations between MB characteristics and perceived health conditions using descriptive statistics and sentiment analysis. Ultimately, we integrated the qualitative interpretations with quantitative patterns to better understand how experiential knowledge, perceived diagnostic meaning, and attitudes toward MB-based diagnostics emerge and interact in public online discourse.

SML was selected because it enables the examination of naturally occurring discussions in public online environments. Hence, potential researcher bias could be avoided. Compared with surveys or interviews, SML captures spontaneous, user-generated discourse and reveals terminology, interpretations, and concerns that may not emerge in traditional qualitative research. However, SML also has methodological constraints, including platform-specific biases, variable data accessibility, and the absence of demographic information [17].

Data Sources and Sampling

The unit of analysis was individual social media posts rather than identifiable users. Data were collected from TikTok, Instagram, Facebook, and Reddit using Mention and Apify. These platforms were selected because they represent some of the most widely used social media globally, with Facebook reporting 3.07 billion monthly active users, Instagram 3 billion, TikTok 1.99 billion, and Reddit 765 million users [24]. Other major platforms (eg, YouTube [Google], WhatsApp [Meta], and WeChat [Tencent Holdings Limited]) were excluded due to limited data accessibility, privacy restrictions, and technical constraints, such as restricted downloading or transcription options. Posts represented the spontaneous discourse of platform users; individual-level demographic information was not available because all data were anonymized. No direct interaction with users occurred. Posts were eligible if they were publicly accessible, in English to ensure analytical consistency, contained explicit reference to MB, and were retrievable through one of the 3 extraction methods. Posts were excluded if they consisted of images or video without analyzable text, were in non-English languages, were duplicates, or did not mention the word “blood.” Visual data were excluded because consistent, valid visual analysis was not feasible in this study, and only textual descriptions of MB appearance were analyzed.

Data Collection

Data extraction was conducted following principles of SML analysis in health research as previously described [4,19,25]. New user profiles were created on Instagram, TikTok, Reddit, and Facebook to minimize algorithmic bias. The profiles were gender-neutral and unlabeled, with the exception of the Instagram account, which required a sex designation and was therefore registered as “female.” Data were collected between February 1 and February 28, 2025. However, this time frame refers to the time of extraction, rather than to the time of publication of the posts. To improve data quality and extend the time frame in which posts were published, we applied triangulation and used 3 methods to extract content from these social media platforms [26].

Group searches: Facebook and Reddit groups were preselected by searching 4 keywords, which were selected by the researchers, as they closely align with the study focus (“Menstrual blood,” “periods,” “period blood,” and “endometriosis” [this term chosen due to the main focus of MAI study (Menstrual Blood Analysis with AI]) and screening the top 10 groups or communities per keyword. After removing duplicates and assessing relevance, 3 Facebook groups (“Girl Talk With Dr. Amber: Menstrual Problems + Solution.Period.,” “PERIOD PROBLEM,” and “Menstruation (Period) share problem”) and 1 Reddit community (“r/Periods”) were included. Using Apify’s Facebook Group Scraper and Reddit Scraper Lite, the 4 preselected groups or communities were scraped.

Hashtag searches: In total, 2 hashtags (#periodblood and #menstrualblood), which are widely used and directly refer to MB-related keywords as selected in the group search, were chosen. Apify’s Hashtag Scraper tools retrieved all posts on TikTok and Instagram associated with these hashtags at the time of extraction. Because Apify’s hashtag scrapers limit retrieval to a maximum of 100 posts per hashtag, researchers could not influence or verify the completeness of the extracted content beyond initiating the search and, therefore, could not ensure that all relevant posts associated with each hashtag were captured.

Social listening alerts: In total, 6 alerts were created to capture MB-related discussions across platforms over 2 time frames. A

XSL•FO
RenderX

prospective search collected and downloaded posts from February 13 to 19, 2025; a retrospective search collected posts from February 2023 to January 2025. Alerts included combinations of menstrual-related terms (eg, “menstrual blood,” “period blood,” “clots,” and “color”). Platforms included in alerts were TikTok, Instagram, Facebook, and X (Twitter).

In total, 6213 posts and videos were retrieved before filtering. Figure 1 illustrates the data collection and preparation workflow. A detailed description of each search method can be found in Tables S1-S5 in Multimedia Appendix 1.

Figure 1. Identification and downloading process: Flowchart illustrating the data identification, extraction, and filtering process used in this mixed methods social media listening study. Data were retrieved from TikTok, Instagram, Facebook, and Reddit in a prospective search from February 13 to February 19, 2025, and a retrospective search from February 2023 to January 2025 using group searches, hashtag searches, and social listening alerts. The figure shows the steps from initial retrieval (6213 posts) to the final analytic dataset (349 posts).

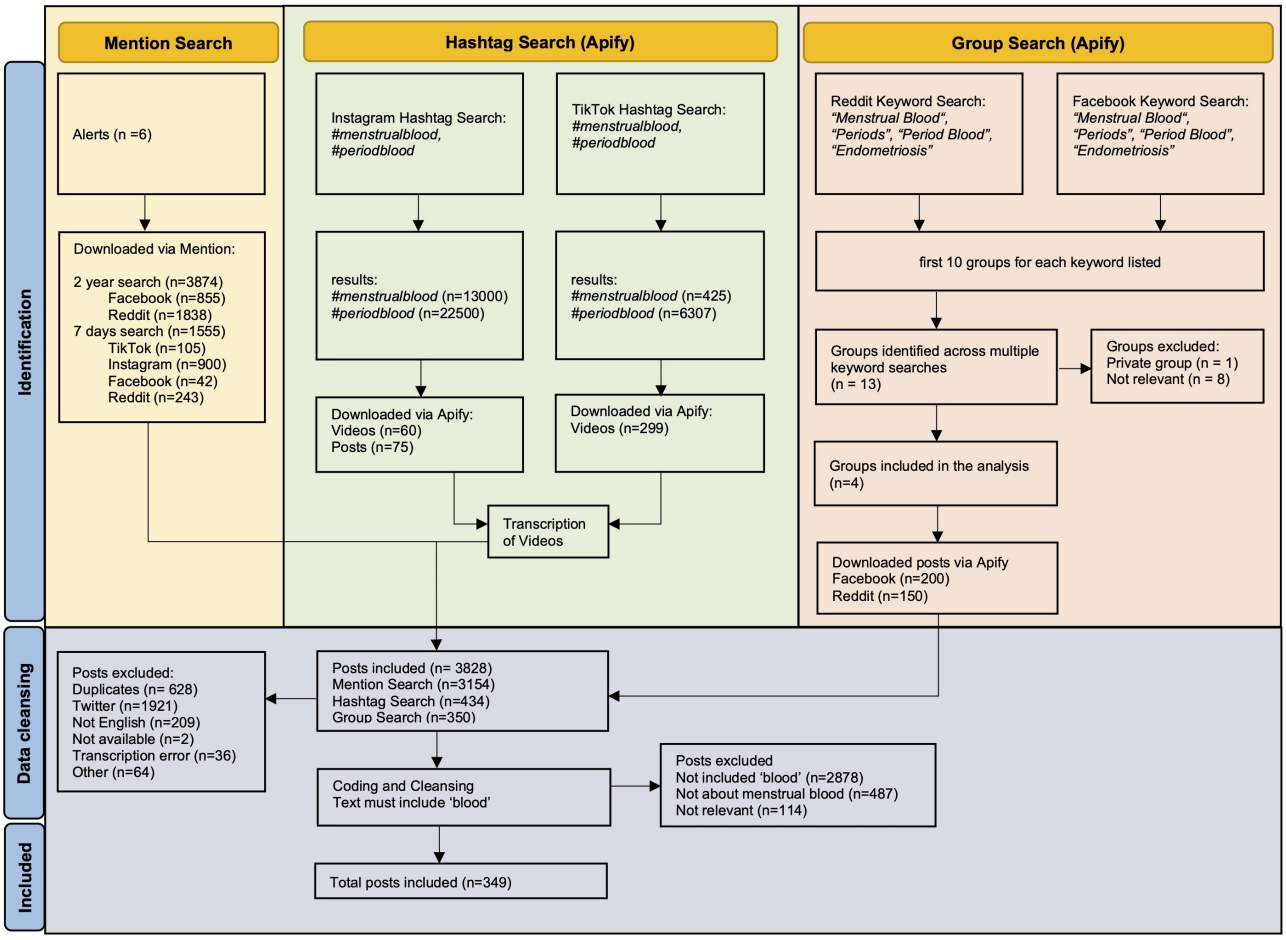

Transforming the Data

All data were anonymized before analysis. Usernames, profile information, timestamps, and identifiable details were removed. Content from X was excluded because it could be extracted exclusively through the alert search and was, therefore, lacking the quality principle of triangulation. After removing duplicates and X content, 3664 posts remained. Data were categorized by type (images, videos, and text). Images were excluded as they could not be objectively analyzed. A total of 392 videos from Instagram and TikTok were transcribed using OpenAI Whisper in Python. Posts in other languages rather than English were excluded. Using a keyword search for the term “blood,” 2890 nonrelevant posts were removed. A final relevance screening yielded 349 posts for analysis.

Data Analysis

Overview

Similar to recent SML studies [4,6,27,28], we used a mixed methods analysis to examine how MB was described (RQ1), which companies working with MB were mentioned, and what benefits and concerns were discussed regarding its diagnostic use (RQ2), with categories treated as nonmutually exclusive.

Qualitative Analysis

Content Analysis

Analysis followed a multistep deductive approach. The dataset was coded using an iterative, cycling approach. First, Atlas.ti’s AI Coding tool generated an initial set of codes. The first author (RN) manually read and reviewed the entire dataset, revising, refining, and verifying all artificial intelligence (AI)–generated codes before applying them to the dataset. Additionally, new codes were applied. Moreover, 2 authors (RN and OI) discussed

XSL•FO
RenderX

and refined the codes. All duplicates were removed, similar codes were merged, and a structured codebook was created.

For analytical clarity, 3 levels of analysis were distinguished: “posts” as the contextual unit of analysis, “quotations” as the coding unit, and “codes” as analytical categories. Posts represented complete social media entries and were each assigned to a single post type (eg, *explaining* and *seeking help*). Quotations referred to discrete text segments within posts that contained relevant content and could be assigned to one or more codes. Codes were applied at the quotation level to capture specific thematic elements, allowing multiple codes from different code groups to be assigned to the same quotation where applicable. As defined by Atlas.ti, groundedness describes the number of quotations coded by a specific code [29].

To address RQ1 on the appearance of MB, 4 code groups were developed: *consistency description*, *color description*, *smell description*, and *other appearance description*. For RQ2 on the diagnostic use of MB, 7 code groups were used: *benefits*, *companies*, *hopes for the future*, *concerns*, *diagnostic methods*, *research*, and *other*. Additional cross-cutting category codes included *type of post*, *sentiment*, and *reasons for MB changes*, supporting contextual analysis of content and tone.

In total, 209 individual codes were generated and are detailed in Tables S6 and S7 in Multimedia Appendix 1.

Given that all eligible posts were coded and the final stage of coding did not yield new categories beyond those already established, data saturation was considered achieved.

Network Analysis

To address RQ1, we performed a network analysis, a methodological approach that examined the relation (edges) between individual actors (nodes), in order to explore how specific characteristics of MB were associated with various health conditions [30,31]. Using Atlas.ti’s network analysis, we linked codes that described appearance descriptions to specific health conditions, as perceived by social media users.

Quantitative Analysis

Statistical analysis: To examine the distribution of posts by platform, RQ, search type, and country, we performed a descriptive statistical analysis.

Sentiment Analysis: Next, we also carried out a sentiment analysis to investigate public perceptions of using MB as a diagnostic tool (RQ2) [21,22]. Sentiment analysis was conducted manually. In addition to content-related codes, each post discussing MB-based diagnostics was categorized as positive, negative, or neutral based on explicit evaluative language by the first author (RN). The last author revised all sentiment-related codes. Positive posts included statements expressing trust, enthusiasm, or implied willingness to use MB as a diagnostic tool. Negative posts expressed concerns, distrust, skepticism, or implied aversion to the use of MB for diagnostics. Neutral posts discussed the potential use of MB without evaluating the wording.

Validity, Reliability, and Methodological Integrity

Methodological rigor was ensured through strategies appropriate to both qualitative and quantitative components of this mixed methods study, in line with Journal Article Reporting Standards for Qualitative and Quantitative Research [32].

For the qualitative analysis, an initial set of deductive codes was generated and refined through repeated engagement with the data. To further enhance reliability, all codes were discussed by 2 authors (RN and OI). Data saturation was considered achieved when no new codes or themes emerged during the final stages of analysis. The dataset was adequate to address the study aims, as it captured diverse, naturally occurring discussions of MB across multiple platforms (Reddit, Facebook, Instagram, and TikTok) and time frames. Although individual demographic information was unavailable, platform diversity and triangulated extraction strategies enhanced the breadth of perspectives represented.

For the quantitative components, reliability was supported through systematic and replicable procedures for data extraction, filtering, and categorization. Descriptive statistics, network analysis, and sentiment categorization were applied consistently across the full dataset. Network analysis followed established methodological principles to examine associations between MB characteristics and perceived health conditions, while sentiment analysis applied predefined evaluative criteria (positive, negative, and neutral) based on explicit user statements.

Ethical Considerations

This study involved the analysis of publicly available social media content and did not include interaction with users or the collection of identifiable personal information. The study protocol was reviewed and approved by the Ethical Commission of Ludwig-Maximilians-University Munich (Project 24 - 1112). Because we relied exclusively on publicly accessible posts and did not recruit participants, informed consent was not required. To protect privacy and confidentiality, all data were anonymized before analysis. Usernames, profile information, timestamps, images, and any potentially identifying material were removed during data preparation, and no attempts were made to reidentify individuals. Only deidentified textual content was retained for coding and analysis—consistent with ethical standards for social media research. As no human participants were recruited for the study, no compensation was provided. No identifiable images, profile screenshots, or user photographs were included in the manuscript or supplementary material. All figures and visualizations reflect aggregated or abstracted data. Therefore, no individual user can be identified, and no additional consent is required for publication.

Results

Quantitative Post Analysis

In total, 349 social media posts discussing MB were analyzed (n=349). Table 1 illustrates the distribution of search terms, sources, and geography. *Seeking help* was the most common post type (154/349, 44.1%), followed by *comment* (73/349, 20.9%), mainly on Reddit and Facebook. *Explaining* posts (47/349, 13.5%) were found across all platforms, especially

XSL•FO
RenderX

Facebook and Instagram, while *advertising* (35/349, 10%) occurred mostly on Facebook. Less common were *patient reports*, *spiritual*, and *menstrual awareness*.

XSL•FO
RenderX

Table . Search tools, sources, and geography of retrieved social media posts: summary of all extraction tools (Apify scrapers and Mention alerts), platforms included (TikTok, Instagram, Facebook, and Reddit), the type of search performed, and the geographic metadata available for retrieved posts. This quantitative data reflects filtered and preprocessed content collected during the extraction period (N describes the entire set of included posts and n/N describes the proportion within a specific subgroup).

	Values, n (%)	Values, n/N (%)
Total included posts	349	__[a]
Search tool and Search terms		
Apify	27 (7.7)	27/27 (100)
GS_Facebook	7 (2)	7/27 (25.9)
GS_Reddit	7 (2)	7/27 (25.9)
HS_TikTok	6 (1.7)	6/27 (22.2)
HS_Insta_Post	4 (1.1)	4/27 (14.8)
HS_Insta_Video	3 (0.9)	3/27 (11.1)
Mention	322 (92.3)	322/322 (100)
HS_MB + Endo	192 (55)	192/322 (59.6)
HS_PB and period facts	13 (3.7)	13/322 (4)
MB and Endo	102 (29.2)	102/322 (31.7)
MB+ #	9 (2.6)	9/322 (2.8)
MB+color+diagnostics	6 (1.7)	6/322 (1.9)
Source		
Facebook	92 (26.4)	—
Instagram	24 (6.9)	—
Reddit	226 (64.8)	—
TikTok	7 (2)	—
Geographies		
Unknown	278 (79.7)	—
Known	71 (20.3)	71/278 (100)
Austria	7 (2)	7/278 (9.9)
Botswana	1 (0.3)	1/278 (1.4)
Canada	2 (0.6)	2/278 (2.8)
Cameroon	2 (0.6)	2/278 (2.8)
Egypt	1 (0.3)	1/278 (1.4)
Spain	1 (0.3)	1/278 (1.4)
Great Britain	8 (2.3)	8/278 (11.3)
India	8 (2.3)	8/278 (11.3)
Luxembourg	3 (0.9)	3/278 (4.2)
Malaysia	1 (0.3)	1/278 (1.4)
Nigeria	5 (1.4)	5/278 (7)
Netherlands	1 (0.3)	1/278 (1.4)
New Zealand	2 (0.6)	2/278 (2.8)
Philippines	1 (0.3)	1/278 (1.4)
Portugal	1 (0.3)	1/278 (1.4)
Taiwan	1 (0.3)	1/278 (1.4)
United States	25 (7.2)	25/278 (35.2)
Zimbabwe	1 (0.3)	1/278 (1.4)

XSL•FO
RenderX

[a]Not applicable.

Platform-specific trends showed Reddit posts (n=226) were largely focused on *seeking help* (144/226, 63.7%), while Facebook (n=92) featured more *advertising* (28/92, 30.4%) and *explaining* (21/92, 22.8%). Instagram (n=24) and TikTok (n=7) posts were primarily *explaining* (13/24, 54.2% and 6/7, 85.7%, respectively; Table S8 Multimedia Appendix 1).

Regarding the research focus, most posts addressed appearance-related aspects of MB and were thus grouped into RQ1 (243/349, 69.6%), and 80 discussed its diagnostic potential (80/349, 22.9%) and were grouped into RQ2, among those only a few were overlapping (5/349, 1.4%), whereas some could not be assigned to a specific RQ (31/349, 8.9%). RQ1 content was mainly sourced from Reddit (183/243, 75.3%), whereas RQ2 was mostly discussed on Facebook (49/80, 61.3%) and Reddit (27/80, 33.8%).

RQ1: Qualitative and Quantitative Description of the Appearance of MB

On posts describing MB appearance (n=243), 80 codes from 4 code groups were distributed among 295 quotations: *color*, *consistency*, *smell*, and *other characteristics*.

Color, Consistency, Smell, and Other Descriptions

Color was the most frequently described aspect (189/295, 64.1%), with *brown* (71/189, 37.6%), *bright red* (37/189, 19.6%), *pink* (25/189, 13.2%), and *black* (22/189, 11.2%) among the most common. Other colors included *dark red*, *gray*, *purple*, *orange*, and unique terms like *cranberry red* and *terracotta*, as illustrated in Figure 2.

Figure 2. Tree map visualization of color descriptions mentioned in posts: Tree map visualization of color descriptors used by social media users to describe menstrual blood in the final dataset. The figure displays the relative frequency of specific color categories and illustrates how users characterize menstrual blood appearances. Each rectangle represents a different color, and the size is proportional to the number of posts mentioning that color. The number after the color indicates the number of mentions.

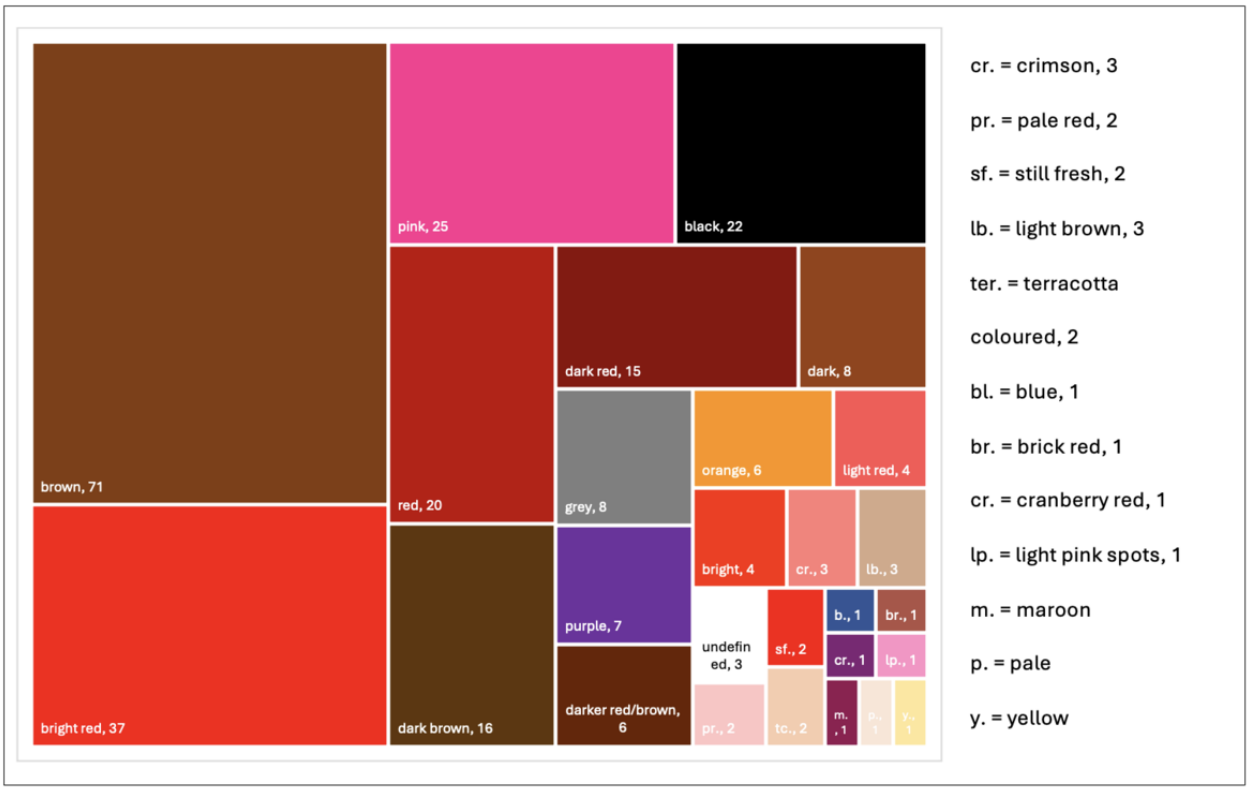

Consistency (123/295, 41.7%) was dominated by the term *coagulation* (87/123, 70.7%), followed by *thick (13/123, 10.6%)*, *light (11/123, 8.9%)*, and *watery (8/123, 6.5%)*. Less common were grainy clumps (4/123, 3.3%), thin (4/123, 3.3%), gloopy (2/123, 1.6%), coffee ground inclusions (2/123, 1.6%), slimy (2/123, 1.6%), sticky (2/123, 1.6%), tissuey flakes (2/123, 1.6%), *chunky (1/123, 0.8%)*, *pasty (1/123, 0.8%)*, and others.

Smell (23/295, 7.8%) was less frequently addressed but included terms such as *unpleasant (9/23, 39.1%)*, *irregular (4/23, 17.4%)*, *metallic (1/23, 4.3%)*, *sour (1/23, 4.3%)*, *sweet (2/23, 8.7%)*, and *yeast-like (1/23, 4.3%)*. Other appearance-related descriptors included *spotting*, *irregular periods*, and *missed periods*.

Appearance Descriptions by Post Type

Of the quotations coded with *colour*-related descriptions, nearly half occurred in *seeking help* posts (92/189, 48.7%). The second largest share of color descriptions appeared in *explaining* posts (52/189, 27.5%), followed by *comment* posts (14/189, 7.4%). A similar pattern was observed for quotations describing MB *consistency: seeking help* posts accounted for the majority of instances (67/123, 54.5%), followed by *explaining* (23/123,

XSL•FO
RenderX

18.7%) and *comment* posts (19/123, 15.4%). Descriptions of *smell* were predominantly found in *seeking help* posts (12/23, 52.2%). Figure 3 illustrates the distribution of posts describing MB appearance characteristics across post types.

Figure 3. Distribution of qualitative posts describing the appearance of menstrual blood across different post types: visualization of how descriptors of MB appearance (color, consistency, smell, and other) were distributed across post types. Gr: groundedness=number of all quotations coded by a specific code.

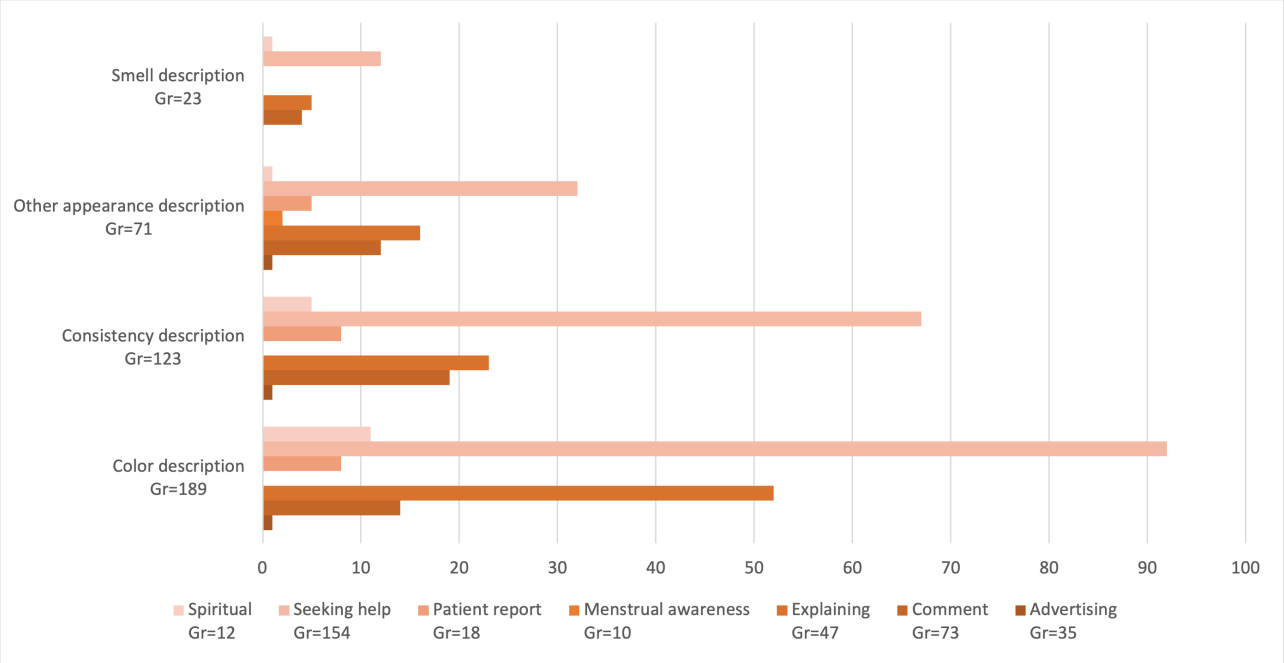

Association With Specific Health Conditions

A network analysis was conducted to explore the relationship between MB appearance and health conditions perceived by social media users. The analysis focused on *coagulation* and the 4 most frequently mentioned colors—*brown*, *bright red*, *pink*, and *black*. These were analyzed in relation to the code group *reasons for changes in MB*, excluding codes that were rare or unrelated. A visualization of both network analyses is presented in Figures 4 and 5.

Figure 4. Co-occurrence network linking menstrual blood color descriptors with perceived health conditions: conceptual map illustrating the association between the most frequent color descriptions (brown, bright red, pink, and black) and perceived health conditions mentioned by users. This figure represents perceived, not clinically validated, associations based on user discourse. The red, centered nodes represent color descriptions, the green nodes represent identified medical conditions that influence the appearance, and the turquoise nodes represent other major influencing factors. Arrows indicate the nature of the relationships, with labels specifying associations (eg, “is associated with,” “is cause of,” “contradicts,” and “even though”).

XSL•FO
RenderX

Figure 5. Co-occurrence network linking menstrual blood consistency descriptors with perceived health conditions: conceptual map illustrating relationships between coagulation and associated health conditions mentioned in user posts. The red, centered node represents consistency description, the green nodes represent identified medical conditions that influence the appearance, and the turquoise nodes represent other major influencing factors. Arrows indicate the nature of the relationships, with labels specifying associations (eg, "is associated with," "is cause of," "contradicts," and "even though").

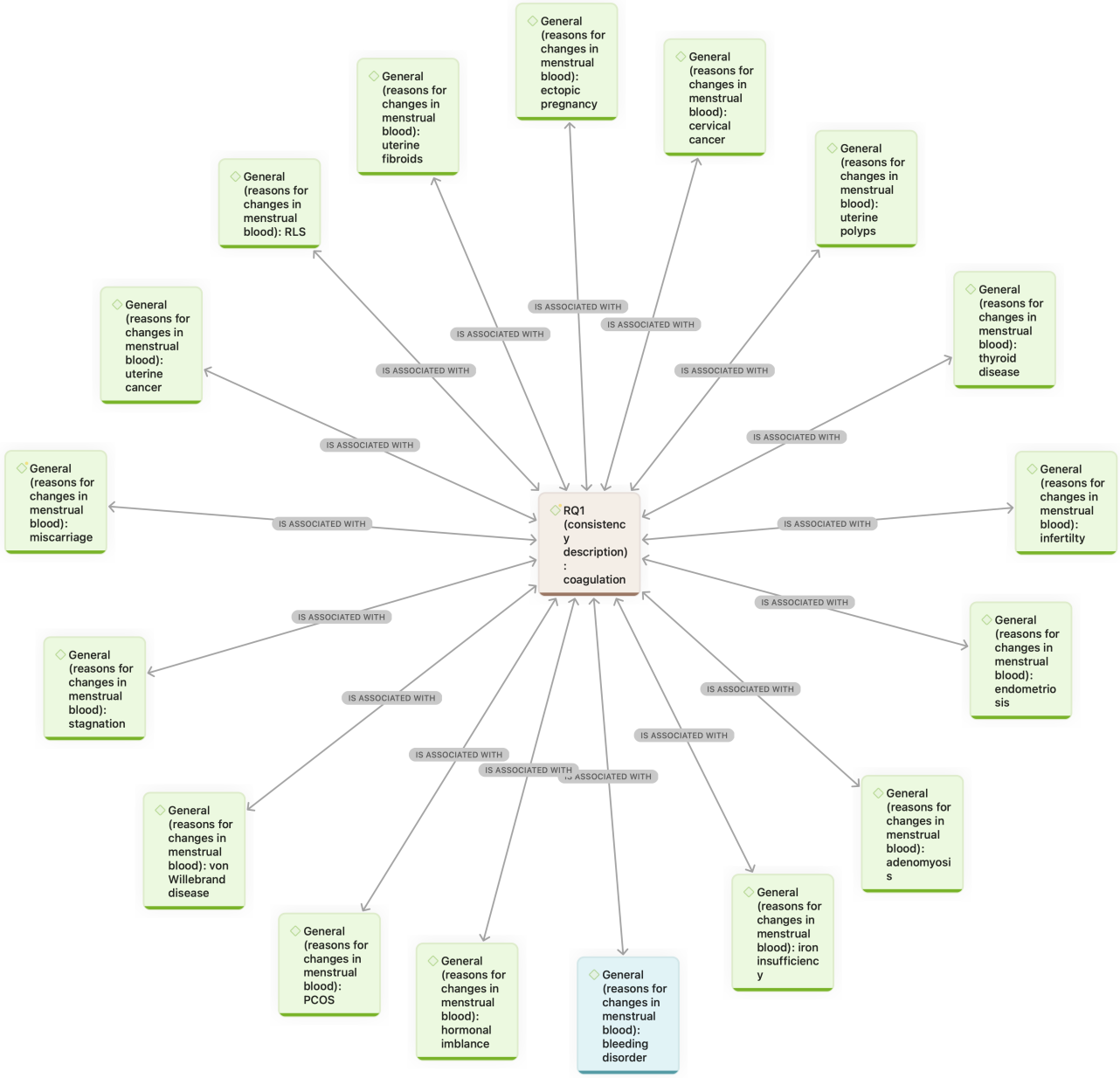

Brown MB was linked to conditions such as *low progesterone*, *ectopic pregnancy*, *PCOS* (polycystic ovary syndrome), *lochia*, *endometriosis*, *miscarriage*, and the *use of medication*. It was also associated with normal phases of the menstrual cycle (eg, *oxidized blood* or the *start and end of menstruation*). Notably, some individuals reported experiencing *brown* MB despite not expecting to menstruate due to contraceptive use. *Pink* blood was connected to *hormonal contraceptives*, *low estrogen*, *iron deficiency*, *ovarian cysts*, *implantation bleeding*, and *mixed with cervical fluid*. *Bright red* blood was often mentioned in relation to *miscarriage*, *cervical cancer*, *sexually transmitted infections* (eg, gonorrhea and chlamydia), and *surgery* (eg, ovary removal). *Black* MB appeared in relation to *endometriosis*, *miscarriage*, *medication use*, and *vaginal obstruction*, as well as *typical menstrual transitions*.

Coagulation was linked to both clinical and cyclical causes, including *uterine polyps*, *adenomyosis*, *endometriosis*, *thyroid issues*, *PCOS*, *hormonal imbalance*, and *uterine cancer*, along with factors like *medication use* and *menstrual phase*.

RQ2: Qualitative and Quantitative Perception of MB as a Diagnostic Tool

In total, 80 posts addressing the diagnostic potential of MB (RQ2) were coded with a set of 42 codes in 7 code groups (*benefits*, *companies*, *hopes for the future*, *concerns*, *diagnostic methods*, *Research*, and *Other*) in 115 quotations.

Companies and Research Initiatives

To assess how MB is discussed as a diagnostic tool, we identified 11 different companies and research groups, which

XSL•FO
RenderX

were named 34 times in social media posts (n=80). Frequently mentioned organizations were *Qvin* (15/34, 44.1%) and *Fertilysis* (6/34, 17.6%). Other commercial entities mentioned less frequently included *Hello Period* (1/34, 2.9%), *Theblood* (1/34, 2.9%), *Diamens* (3/23, 8.8%), and *Papcup* (1/34, 2.9%). In addition, users referred to several research-oriented initiatives, including *ROSE* (3/23, 8.8%), *NextGen Jane* (1/34, 2.9%), *MensEndoDiag* (1/34, 2.9%), and *Flowintell* (1/34, 2.9%). The distribution is illustrated in Figure 6.

Figure 6. Mentions of diagnostic companies and technologies across platforms: bar chart showing the number of posts mentioning diagnostic companies or research groups related to the use of menstrual blood for diagnostic purposes across 3 social media platforms (Facebook, Instagram, and Reddit).

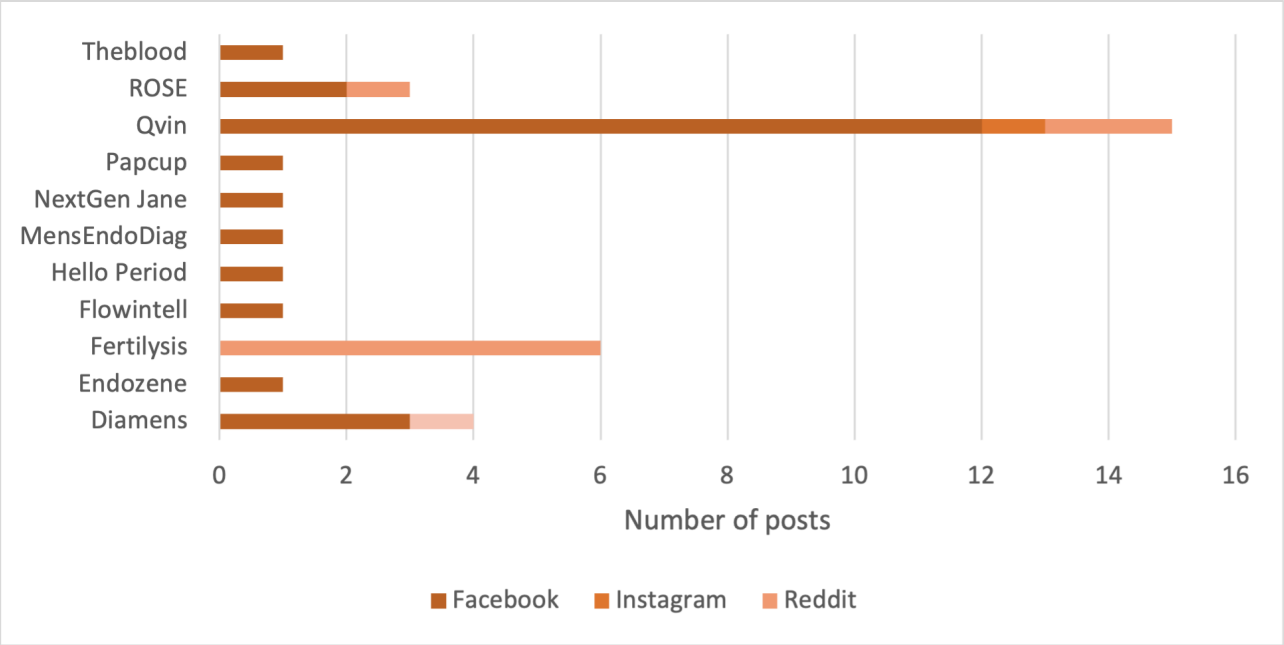

Sentiment Analysis

All quotations discussing diagnostic use of MB (n=115) were designated as positive (n=110), negative (n=19), or neutral (n=16), with the possibility of overlapping (eg, when expressing both positive feelings and concerns). Most posts expressed positive sentiment (110/115, 95.7%), indicating strong support and optimism for innovation in menstrual health. Neutral quotations (16/115, 13.9%) were mainly descriptive, while only a few users expressed negative feelings (19/115, 16.5%), with skepticism or concerns.

Perceived Benefits and Concerns

Table 2 summarizes 60 mentions of potential benefits, contrasted by 20 concerns. Additionally, 10 posts expressed surprise at the delayed exploration of the potential of MB.

XSL•FO
RenderX

Table . Perceived benefits and concerns regarding the use of menstrual blood for diagnostics: summary of qualitative themes describing users' perceived advantages (60 mentions) and concerns (19 mentions) about potential menstrual blood–based diagnostic tools. Themes were derived from 80 relevant posts.

Perception of theme	Count, n
Benefits	
Destigmatization	2
Early detection	2
Easily accessible	15
Fast	4
Home-based	12
Less intimidating	1
No ethical concerns	1
Noninvasive	10
Pain-free	3
Plentiful	1
Priceworthy	6
Therapeutic use	3
Total	60
Concerns	
Being sold off	1
Collection of MB	1
Expensive	2
Fewer stem cells than in PB	1
Losing body autonomy	2
Menopause or no MB	3
Not getting paid	2
Political climate	1
Rather be tested by a doctor	1
Stigma	3
Toxic if consumed	1
Long time until accessible for all women	1
Yuck factor	1
Total	20

Discussion

Principal Findings

This study examined how individuals discuss MB, its appearance, and diagnostic potential on social media. Discussions of MB appearance were primarily framed within help-seeking contexts, particularly on Facebook and Reddit, where users asked whether variations in color (eg, brown, red, or pink) and texture (eg, coagulation) should be considered normal. Alongside these questions, explanatory posts interpreted such variations as physiological and within the range of typical menstrual experiences. Discussions addressing the diagnostic potential of MB were comparatively limited; however, while concerns regarding the costs or not experiencing menstrual bleeding were expressed, the overall sentiment toward the prospective diagnostic use of MB was largely positive.

The predominance of help-seeking posts, followed by explanatory contributions, suggests that social media platforms function as alternative spaces for menstrual health inquiry and social support [33]. Previous research has shown that social media plays a significant role in individual well-being and can both benefit and harm users, depending on how these platforms are used [34]. While social comparison and isolation may negatively affect well-being, meaningful interactions, such as the exchange of personal menstrual health experiences, have the potential to foster social support and improve both general well-being and menstrual health in particular [33].

XSL•FO
RenderX

Regarding the accuracy of menstrual health information shared, verification remains challenging. Although licensed health care professionals may disseminate medical advice on social media platforms, it is often difficult for users to distinguish between self-identified "experts" and credentialed professionals. This ambiguity raises concerns about the reliability of menstrual health advice encountered online and highlights a potential gap in accessible, authoritative menstrual health communication. As a result, individuals may increasingly rely on peer-generated or unverifiable information to address menstrual health questions. Thus, health professionals need to be aware of the influence of social media on health decision-making [4,17].

Even though the medical accuracy of menstrual health advice shared on social media cannot be verified, users actively draw connections between specific MB appearances and perceived pathologies or physiological menstrual cycle changes. To date, scientific research and clinical guidelines have largely focused on menstrual cycle characteristics, such as cycle length, bleeding duration, and volume, while neglecting the appearance of MB itself [35-38]. Against this background, the associations observed in social media discourse show notable thematic overlap with the limited scientific literature. Beyond this, most explanations available to the public stem from a wide range of clinical websites (eg, Cleveland Clinic) but do not offer systematic empirical analysis [39].

The findings of our study indicate that descriptions of MB appearance on social media predominantly focus on colors. While brown, bright red, pink, and black were the most frequently described colors, a wide spectrum of additional color variations was also reported. Analysis by Musarofah and Mahmudah [40] of the 2 historical manuscripts "*Serat Piwulang Estri*" and "*Risalah Al-Mahîd*" described specific MB colors as indicators of weak (eg, yellow and gray) or strong (eg, blackish red and bright red) health, suggesting an early conceptual link between MB color and well-being. Similar patterns emerged in this study. Social media users in this study often associated bright red MB with adverse outcomes of physiological changes, but also with miscarriage, cervical cancer, or sexually transmitted diseases. Darker shades, such as brown, were described both in relation to potential health concerns and as normal variations of the menstrual cycle, explained by prolonged retention of blood in the uterus. An article of Cleveland Clinic shows similar interpretations of bright red (fresh and healthy) and dark red (old and oxidized) blood [39]. Comparable associations between MB color and reproductive health have also been reported by Xiping et al [41], who found that bright red MB was associated with significantly lower pregnancy rates and light red MB with an increased risk of miscarriage. Both Musarofah and Mahmudah [40] and social media users associated pink MB with low estrogen levels. Cleveland Clinic, however, associated pink MB with normal color appearance during the first day of a period [39].

Notably, although 53.8% (1135/2109) of participants in the study by Xiping et al [41] reported blood coagulation, no increased risk of infertility or miscarriage was observed. However, they reported various conditions, such as uterine polyps, uterine fibroids, endometriosis, adenomyosis, and hormonal imbalances in connection to coagulation, all of which were also linked by social media users to MB clots. According to the Cleveland Clinic, the size of a blood clot is the distinguishing factor between physiological coagulation and pathological changes [39]. These findings highlight the potential of drawing menstrual health information from the appearance of MB. However, due to limited literature, there is a need for systematic research into MB color and consistency characteristics.

Smell was rarely discussed, but when mentioned, it was framed as unusual or concerning. Research on MB odor is scarce; however, parallels can be drawn from clinical descriptions of vaginal discharge, where specific odors may accompany infections, as outlined in standard gynecology resources [42-44].

Discussion about the diagnostic use of MB was relatively limited in social media discourse (80 posts compared to 243 posts discussing MB's appearance characteristics). Zaheer et al [11] similarly reported limited public awareness of MB's biomedical potential. However, when present, it was predominantly framed in a positive and anticipatory manner. A limited, yet positive awareness of the potential of MB was also shown by Kaur et al [45], who reported that only 28.3% (141/499) of participants knew about the possibility of donating MB, but 64.7% (323/499) of participants saw benefits for society in doing so. Manley et al [46] found that 78% (80/102) of menstruating individuals were open to donating MB for stem cell research, with many reporting a more positive perception of their menstrual cycle following participation. Notably, women affected by menstrual disorders, especially endometriosis, are particularly willing to donate MB for research purposes [47], suggesting that populations most affected by menstrual morbidity may also be most receptive to MB-based diagnostics for personal health monitoring.

Recent technological developments illustrate this growing interest; companies such as *Qvin* [48] developed home-based MB collection devices for biomarker analysis, including hemoglobin A1c, while *Fertilysis* [49] is exploring MB-based diagnostics relevant to fertility and hormonal assessment. AI-supported tools, such as *MenstruAI* [14], further show how computational approaches may complement biological assays in future menstrual health diagnostics. Advantages, such as noninvasive sampling, easy access to the sample, and having the possibility of doing home-based sampling, expressed by social media users closely mirrored those described in scientific literature. Frequently cited benefits included the noninvasiveness of MB collection, its ease of access, home-based self-collection, and the potential elimination of discomfort associated with conventional diagnostic procedures [11]. At the same time, users raised concerns related to body autonomy, lack of financial compensation for MB donation, logistical challenges, such as transportation and processing, and the exclusion of individuals who do not menstruate, due to menopause or hormonal disorders. Similar limitations have been reported by Zaheer et al [11], including restricted sample volume, hormonal fluctuations, contamination risks, interference from menstrual products, stigma, and the missing menstrual bleeding of postmenopausal individuals, children, or individuals experiencing menstrual disorders.

XSL•FO
RenderX

Practical feasibility emerged as a central issue. Menstrual cups are widely reported as an acceptable and effective method for noninvasive MB collection [47,50,51], yet MB can only be collected once per menstrual cycle and within a limited time window, posing challenges particularly for individuals with irregular cycles, infertility, or amenorrhea [10]. Despite growing interest, no standardized clinical protocols currently exist for the collection, transport, processing, and storage of MB samples [10,11]. Nevertheless, emerging research highlights substantial diagnostic and therapeutic potential. MB offers alternative approaches for cervical cancer screening, the diagnosis of endometriosis, genital tuberculosis, diabetes mellitus, altered lipid profiles, and sexually transmitted diseases. Furthermore, due to their pluripotency and low immunogenicity, menstrual blood–derived stem cells have shown promising results in regenerative and therapeutic applications, including the treatment of infertility, premature ovarian failure, intrauterine adhesions, chronic liver disease, and musculoskeletal injuries [15].

These results suggest a marked gap between public interest and clinical implementation. Nonetheless, SML offers valuable insights into lived experiences, especially when clinical engagement is limited or absent. To our knowledge, this is the first study to look at MB discourse, revealing that many individuals already use the appearance of MB as an informal health indicator, particularly in contexts of diagnostic uncertainty or lack of medical recognition. Other social listening studies have already explored the potential of analyzing online discussions for gaining insights into societal or patients' behaviors. Based on this potential, we extended the insights into how individuals experience menstrual bleeding and are open to further establishment of using MB for diagnostics.

Future research should investigate the diagnostic potential of MB, focusing on visual features, and explore how menstrual knowledge is shaped across time, cultures, and biomedical contexts. Moreover, awareness campaigns are needed to bridge the gap between public interest and clinical application [45]. Engaging both patients and health care providers is essential to overcoming stigma and improving diagnostic pathways.

Limitations

This study is subject to several limitations that should be considered when interpreting the findings. First, although the analysis focused on the "appearance" of MB, only textual descriptions could be included because images and videos could not be reliably analyzed. Excluding these visual materials means that the findings for RQ1 are based entirely on users' self-reported descriptions rather than the appearance itself. This introduces an interpretive gap between users' observations and their articulation, reducing the precision with which appearance-related interpretations can be understood. A second limitation arises from the short extraction period and technical constraints due to downloading restrictions from Mention and Apify. Apify hashtag scrapers retrieved only the most recent 100 posts per hashtag, and Mention alerts captured content within predefined retrospective and prospective windows. As a result, the dataset should be interpreted as a snapshot of the public discourse of all MB-related discussions online. Posting activity patterns, platform algorithms, and seasonal or event-driven fluctuations may therefore influence the types of posts retrieved.

Third, only English-language posts were included to ensure analytic consistency and feasibility. This choice inevitably skews the dataset toward English-speaking users and may underrepresent cultural or regional differences in how MB is discussed, interpreted, and linked to health conditions. Fourth, social media primarily reflects the perspectives of younger, digitally active users [52]; the findings may therefore not fully represent the broader population, particularly individuals without consistent internet access or those less likely to engage in public online discourse. Additional limitations inherent to SML also apply. Due to restricted data access, limited sample size, and the lack of standardized protocols, bias control and representativeness remain challenging.

Conclusion

This study demonstrates that individuals increasingly turn to social media platforms to seek information and reassurance about MB appearance, underscoring a gap between the demand for evidence-based menstrual health information and the limited scientific literature addressing MB characteristics. The prominence of peer-guided explanations highlights both unmet informational needs and users' openness to engaging with menstrual health topics outside clinical settings. Together with the largely positive attitudes toward the potential diagnostic use of MB, these findings emphasize the need to expand research on MB characteristics and to systematically evaluate its diagnostic potential for menstrual health–related conditions.

Acknowledgments

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing, summarizing text, adapting and adjusting emotional tone, and reformatting

The GenAI tool used was ChatGPT-5.2 (OpenAI). Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

This study is part of the MAI study, a citizen science project analyzing menstrual blood images from women with and without endometriosis using AI. A project website shares background, study details, and social media listening results [53], and is regularly updated to ensure transparency and communication. This social media listening does not involve active patient or public participation in study design, analysis, or interpretation.

XSL•FO
RenderX

Funding

This research was funded by a scholarship from the "Deutsche Gesellschaft für Digitale Medizin" and "Deutsche Ärztefinanz." The funder was not involved in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

Data relevant to the study are included in the article or Multimedia Appendix 1. The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

RN and OI conceptualized the study. RN performed data extraction and analysis. NC, AB, and MH supervised the study. All coauthors read and approved the final manuscript. All authors were affiliated with the institutions listed at the time the study was conducted.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies.
[DOCX File, 80 KB - jmir_v28i1e85550_app1.docx]

References

1. Critchley HOD, Babayev E, Bulun SE, et al. Menstruation: science and society. Am J Obstet Gynecol 2020 Nov;223(5):624-664. [doi: 10.1016/j.ajog.2020.06.004] [Medline: 32707266]
2. WHO statement on menstrual health and rights. World Health Organization. URL: https://www.who.int/news/item/22-06-2022-who-statement-on-menstrual-health-and-rights [accessed 2025-04-12]
3. Carneiro MM, Farace BL, Ribeiro LDC, et al. Using social media to educate women and healthcare providers on endometriosis: preliminary results. JBRA Assist Reprod 2020 Jan 30;24(1):9-12. [doi: 10.5935/1518-0557.20190048] [Medline: 31361436]
4. Adler H, Lewis M, Ng CHM, et al. Social media, endometriosis, and evidence-based information: an analysis of Instagram content. Healthcare (Basel) 2024 Jan 4;12(1):121. [doi: 10.3390/healthcare12010121] [Medline: 38201027]
5. Kubovski A, Cohen Shabot S. Experience of menstruation with the use of reusable menstrual products: a qualitative study among menstruators in Israel. Cult Health Sex 2024 Jul;26(7):904-919. [doi: 10.1080/13691058.2023.2260465] [Medline: 37782548]
6. Metzler JM, Kalaitzopoulos DR, Burla L, Schaer G, Imesch P. Examining the influence on perceptions of endometriosis via analysis of social media posts: cross-sectional study. JMIR Form Res 2022 Mar 18;6(3):e31135. [doi: 10.2196/31135] [Medline: 35302501]
7. Carneiro MM, Koga CN, Mussi MCL, Fradico PF, Carneiro MM. #infertility in Brazil: an analysis of authorship and post content on Instagram. JBRA 2024;28(4). [doi: 10.5935/1518-0557.20240081]
8. Sanitary suspenders to mooncups: a brief history of menstrual products. V&A Vic Albert Mus. URL: https://www.vam.ac.uk/articles/a-brief-history-of-menstrual-products [accessed 2025-04-12]
9. Tomlinson MK. Seeing Menstruation in the Media: The News, Advertisements, and Memes That Teenagers Encounter via Social Media and the Mainstream Media: Springer Nature Switzerland; 2025:167-210. [doi: 10.1007/978-3-031-72195-3_4]
10. Tindal K, Filby CE, Cousins FL, et al. The composition of menstrual fluid, its applications, and recent advances to understand the endometrial environment: a narrative review. F&S Reviews 2024 Jul;5(3):100075. [doi: 10.1016/j.xfnr.2024.100075]
11. Zaheer A, Komel A, Abu Bakr MB, et al. Potential for and challenges of menstrual blood as a non-invasive diagnostic specimen: current status and future directions. Annals of Medicine & Surgery 2024 Jun 10;86(8):4591-4600. [doi: 10.1097/MS9.0000000000002261]
12. Naseri S, Brewster RCL, Blumenthal PD. Novel use of menstrual blood for monitoring glycaemic control in patients with diabetes: a proof-of-concept study. BMJ Sex Reprod Health 2022 Apr;48(2):123-127. [doi: 10.1136/bmjsrh-2021-201211] [Medline: 34759003]
13. Naseri S, Young S, Cruz G, Blumenthal PD. Screening for high-risk human papillomavirus using passive, self-collected menstrual blood. Obstet Gynecol 2022 Sep 1;140(3):470-476. [doi: 10.1097/AOG.0000000000004904] [Medline: 35926207]
14. Dosnon L, Rduch T, Meyer C, Herrmann IK. A wearable in-pad diagnostic for the detection of disease biomarkers in menstruation blood. Primary Care Research. Preprint posted online on Mar 24, 2024. [doi: 10.1101/2024.03.22.24304704]
15. Feng Y, He Y. The secrets of menstrual blood: emerging frontiers from diagnostic tools to stem cell therapies. Front Cell Dev Biol 2025;13:1623959. [doi: 10.3389/fcell.2025.1623959] [Medline: 40909174]

XSL•FO
RenderX

16. Rosy MM. Menstrual blood banking-a review. Interna Jour of Nur Edu and Rese 2017;5(2):213. [doi: 10.5958/2454-2660.2017.00046.1]
17. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85. [doi: 10.2196/jmir.1933] [Medline: 23615206]
18. Hultman M, Iveson A, Oghazi P. Talk less and listen more? The effectiveness of social media talking and listening tactics on export performance. J Bus Res 2023 Apr;159:113751. [doi: 10.1016/j.jbusres.2023.113751]
19. Le Busque B, Mellish S. Endometriosis awareness month on social media: a content analysis of images and captions on Instagram. Women 2023 Jan 30;3(1):82-94. [doi: 10.3390/women3010007]
20. Pepin K, Bologna F, Thalken R, Wilkens M. Endometriosis online communities: how machine learning can help physicians understand what patients are discussing online. J Minim Invasive Gynecol 2024 Dec;31(12):1011-1018. [doi: 10.1016/j.jmig.2024.08.001] [Medline: 39134239]
21. Metwally O, Blumberg S, Ladabaum U, Sinha SR. Using social media to characterize public sentiment toward medical interventions commonly used for cancer screening: an observational study. J Med Internet Res 2017 Jun 7;19(6):e200. [doi: 10.2196/jmir.7485] [Medline: 28592395]
22. Sloesen B, O'Brien P, Verma H, et al. Patient experiences and insights on chronic ocular pain: social media listening study. JMIR Form Res 2024 Feb 15;8:e47245. [doi: 10.2196/47245] [Medline: 38358786]
23. Creswell JW, Clark VLP. Designing and Conducting Mixed Methods Research, 3rd edition: SAGE Publications; 2018.
24. Global social media statistics. DataReportal – Global Digital Insights. URL: https://datareportal.com/social-media-users [accessed 2025-11-18]
25. Park JH, Christman MP, Linos E, Rieder EA. Dermatology on Instagram: an analysis of hashtags. J Drugs Dermatol 2018 Apr 1;17(4):482-484. [Medline: 29601627]
26. Carter N, Bryant-Lukosius D, DiCenso A, Blythe J, Neville AJ. The use of triangulation in qualitative research. Oncol Nurs Forum 2014 Sep;41(5):545-547. [doi: 10.1188/14.ONF.545-547] [Medline: 25158659]
27. Cole-Lewis H, Pugatch J, Sanders A, et al. Social listening: a content analysis of e-cigarette discussions on Twitter. J Med Internet Res 2015 Oct 27;17(10):e243. [doi: 10.2196/jmir.4969] [Medline: 26508089]
28. Trevisan F, Clifford S, Cooper V. P34 use of social media listening (SML) methods for understanding the patient experience of chronic disease: a scoping review. Value Health 2022 Jan;25(1):S8. [doi: 10.1016/j.jval.2021.11.033]
29. ATLAS.ti 22 windows - quick tour. ATLAS.ti. URL: https://doc.atlasti.com/QuicktourWin.v22/Codes/CodingDataBasicConcepts.html [accessed 2026-01-27]
30. Friese S. Qualitative Data Analysis with ATLASTi (Third Edition): SAGE Publications Ltd; 2019.
31. Introduction to social network methods. Analytic Technologies. URL: http://www.analytictech.com/networks.pdf [accessed 2025-11-20]
32. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
33. Lin XY, Lachman ME. Social media use and daily well-being: the role of quantity and quality of social support. Res Aging 2024;46(5-6):287-301. [doi: 10.1177/01640275241227575] [Medline: 38217507]
34. Clark JL, Algoe SB, Green MC. Social network sites and well-being: the role of social connection. Curr Dir Psychol Sci 2018 Feb;27(1):32-37. [doi: 10.1177/0963721417730833]
35. Wesselink AK, Wise LA, Hatch EE, et al. Menstrual cycle characteristics and fecundability in a North American preconception cohort. Ann Epidemiol 2016 Jul;26(7):482-487. [doi: 10.1016/j.annepidem.2016.05.006] [Medline: 27449569]
36. Dasharathy SS, Mumford SL, Pollack AZ, et al. Menstrual bleeding patterns among regularly menstruating women. Am J Epidemiol 2012 Mar 15;175(6):536-545. [doi: 10.1093/aje/kwr356] [Medline: 22350580]
37. Genazzani AR, Goldstein SR, Simoncini T, editors. Menstrual Bleeding and Pain Disorders from Adolescence to Menopause: Volume 11: Frontiers in Gynecological Endocrinology: Springer International Publishing; 2024. [doi: 10.1007/978-3-031-55300-4]
38. Munro MG, Critchley HOD, Broder MS, Fraser IS, FIGO Working Group on Menstrual Disorders. FIGO classification system (PALM‐COEIN) for causes of abnormal uterine bleeding in nongravid women of reproductive age. Intl J Gynecology & Obste 2011 Apr;113(1):3-13. [doi: 10.1016/j.ijgo.2010.11.011] [Medline: 21345435]
39. Kollikonda S. Period blood comes in different colors: here's what they mean. Cleveland Clinic | Health Essentials. 2025. URL: https://health.clevelandclinic.org/what-does-the-color-of-your-period-mean [accessed 2025-11-22]
40. Musarofah S, Mahmudah U. Reproduction health based on various kinds of women's blood in Serat Piwulang Estri and Risalah Al-Mahîd: philological and comparative studies. JSTARD 2020 Dec 31;2(3). [doi: 10.30880/jstard.2020.02.03.007]
41. Xiping L, Xiaqiu WU, Lirong B, Jin P, Hui KK. Menstrual cycle characteristics as an indicator of fertility outcomes: evidence from prospective birth cohort study in China. J Tradit Chin Med 2022 Apr;42(2):272-278. [doi: 10.19852/j.cnki.jtcm.2022.02.010] [Medline: 35473349]
42. Rao VL, Mahmood T. Vaginal discharge. Obstetrics, Gynaecology & Reproductive Medicine 2020 Jan;30(1):11-18. [doi: 10.1016/j.ogrm.2019.10.004]

XSL•FO
RenderX

43. Stauber M, Weyerstahl T, Beham A, Beham A. Gynäkologie und geburtshilfe: 213 tabellen. In: [mit Video-CD-ROM: Filme zu Spontangeburt und Sectio]. 2., aktualisierte Aufl. mit Video-CD-ROM: Thieme; 2005.
44. Trivedi AN, Pandey A. A Practical Guide to Gynaecology: Cambridge Scholars Publishing; 2024. URL: http://ebookcentral.proquest.com/lib/unigiessen/detail.action?docID=31212728 [accessed 2025-11-22]
45. Kaur KN, Nandi D, Ramachandran K, et al. Knowledge, attitude, and practice of menstrual blood-derived mesenchymal stem cells among female healthcare workers in India. Front Public Health 2023;11:1102016. [doi: 10.3389/fpubh.2023.1102016] [Medline: 37213620]
46. Manley H, Sprinks J, Breedon P. Menstrual blood-derived mesenchymal stem cells: women's attitudes, willingness, and barriers to donation of menstrual blood. J Womens Health (Larchmt) 2019 Dec;28(12):1688-1697. [doi: 10.1089/jwh.2019.7745] [Medline: 31397634]
47. Bouzid K, Bourdon M, Bartkowski R, et al. Menstrual blood donation for endometriosis research: a cross-sectional survey on women's willingness and potential barriers. Reprod Sci 2024 Jun;31(6):1617-1625. [doi: 10.1007/s43032-024-01481-3] [Medline: 38418666]
48. QvinTM introduces q-padtm: transforming women's health with FDA-cleared lab testing using menstrual blood. BioSpace. 2024. URL: https://www.biospace.com/qvin-introduces-q-pad-transforming-women-s-health-with-fda-cleared-lab-testing-using-menstrual-blood [accessed 2024-09-27]
49. Home. FERTILYSIS. 2022. URL: https://fertilysis.com/ [accessed 2025-11-22]
50. Koks CAM, Dunselman GAJ, de Goeij A, Arends JW, Evers JLH. Evaluation of a menstrual cup to collect shed endometrium for in vitro studies. Fertil Steril 1997 Sep;68(3):560-564. [doi: 10.1016/S0015-0282(97)00250-1]
51. van Eijk AM, Zulaika G, Lenchner M, et al. Menstrual cup use, leakage, acceptability, safety, and availability: a systematic review and meta-analysis. Lancet Public Health 2019 Aug;4(8):e376-e393. [doi: 10.1016/S2468-2667(19)30111-2] [Medline: 31324419]
52. Mazza M, Piperis M, Aasaithambi S, Chauhan J, Sagkriotis A, Vieira C. Social media listening to understand the lived experience of individuals in Europe with metastatic breast cancer: a systematic search and content analysis study. Front Oncol 2022;12:863641. [doi: 10.3389/fonc.2022.863641] [Medline: 35719996]
53. Niedermeier R. MAI-menstrual blood analysis with artificial intelligence. Canva. URL: https://menstrualbloodai.my.canva.site/mai [accessed 2026-04-04]

Abbreviations

AI: artificial intelligence
MB: menstrual blood
PCOS: polycystic ovary syndrome
RQ: research question
SML: social media listening

Edited by S Brini; submitted 14.Oct.2025; peer-reviewed by A Amanna, B Otesanya, O Enaworu; revised version received 02.Mar.2026; accepted 03.Mar.2026; published 04.May.2026.

Please cite as:
Niedermeier R, Castelletti N, Bender A, Hoelscher M, Ivanova O
Interpretations of Menstrual Blood Appearance and Diagnostic Potential Among Social Media Users: Cross-Sectional Mixed Methods Social Media Listening Study
J Med Internet Res 2026;28:e85550
URL: https://www.jmir.org/2026/1/e85550
doi:10.2196/85550

© Ruth Niedermeier, Noemi Castelletti, Andreas Bender, Michael Hoelscher, Olena Ivanova. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 4.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Leveraging Social Media to Achieve Population-Level Reach of Lung Cancer Screening-Eligible Individuals: A RE-AIM Framework Perspective

Lisa Carter-Bawa[1,2], MSN, MPH, PhD; Jamie S Ostroff[3], PhD; Susan M Rawl[4], PhD; Erin A Hirsch[1], PhD; Smita C Banerjee[3], PhD; Andrew Ciupek[5], PhD; Robert Skipworth Comer[6], MS; Minal Kale[7], MPH, MD; Katherine T Leopold[8], MD; Patrick O Monahan[9], PhD; James E Slaven Jr[9], MS; Francis Valenzona[1], BS; Renda Soylemez Wiener[10,11,12], MPH, MD; Ana Guadalupe Vielma[1], PhD

[1]Cancer Prevention Precision Control Institute, Center for Discovery & Innovation, Hackensack Meridian Health, 123 Metro Blvd, 6th Floor, 6400 Pod, Nutley, NJ, United States

[10]Center for Health Optimization & Implementation Research, VA Boston Healthcare System, Boston, MA, United States

[11]National Center for Lung Cancer Screening, Veterans Health Administration, Washington, DC, United States

[12]The Pulmonary Center, Boston University Chobanian & Avedesian School of Medicine, Boston, MA, United States

[2]Cancer Prevention & Control Program, Georgetown Lombardi Comprehensive Cancer Center, Washington, DC, United States

[3]Department of Psychiatry & Behavioral Science, Memorial Sloan Kettering Cancer Center, New York, NY, United States

[4]School of Nursing, Indiana University, Indianapolis, United States

[5]Go2 for Lung Cancer, Washington, United States

[6]School of Informatics, Indiana University, Indianapolis, United States

[7]Department of Medicine, Internal Medicine, Icahn School of Medicine at Mount Sinai, New York, United States

[8]Hackensack Meridian School of Medicine, Hackensack Meridian Health, Nutley, United States

[9]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, United States

Corresponding Author:
Lisa Carter-Bawa, MSN, MPH, PhD
Cancer Prevention Precision Control Institute, Center for Discovery & Innovation, Hackensack Meridian Health, 123 Metro Blvd, 6th Floor, 6400 Pod, Nutley, NJ, United States

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2026/1/e94664

Abstract

Background: Annual lung cancer screening (LCS) can decrease lung cancer–related mortality by finding cancer at earlier, more treatable stages, yet uptake remains abysmally low in the United States, especially among adults who seldom interact with the health system. Many eligible individuals are unaware that LCS exists, underscoring the critical need for scalable, population-level communication strategies that increase awareness and engagement.

Objective: The aim of this study was to evaluate *reach*, as defined by the Reach, Effectiveness, Adoption, Implementation, and Maintenance framework, as the extent to which the target population comes in contact with a social media–based strategy, Facebook-targeted advertisement (FBTA), designed to connect LCS-eligible individuals in the United States with a digital health communication message. The advertisement served as a digital outreach strategy for promoting engagement with *LungTalk*, an evidence-based intervention aimed at increasing awareness and informed decision-making about LCS.

Methods: As part of the INSPIRE-Lung Study (INnovating Social Media for Prevention: LUNG Cancer Screening Awareness, Knowledge, and Uptake), 5 FBTA campaigns were launched over a 79-day period throughout the United States. Advertisements targeted adults aged 50 - 80 years with interests related to smoking or smoking cessation and linked to a study website where participants could complete an eligibility screener and learn more about the trial. Facebook analytics were used to assess *reach*, defined by the number, proportion, and demographic characteristics of individuals exposed to and interacting with FBTA content. Key metrics included total reach, impressions, link clicks, and cost-efficiency.

Results: The FBTA campaigns reached 1,048,191 unique users and generated 3,109,482 impressions (total advertisement displays, including repeat exposures to the same user). A total of 24,816 individuals clicked on the advertisements (2.37% click-through rate), and 7117 completed the eligibility screener. Of those eligible, 1272 (17.9%) met lung screening criteria, and

XSL•FO
RenderX

of these, 483 (38% participation rate) enrolled in the trial. The cost per click was US $0.40, and the cost per enrolled participant was US $19.46. Individuals reached via FBTA were demographically diverse and included many who may be disconnected from traditional health care systems.

Conclusions: FBTA is a scalable, cost-effective strategy to achieve population-level *reach* of LCS-eligible adults. By conceptualizing *reach* as exposure to an upstream digital message rather than enrollment alone, this study illustrates how social media can broaden population access to evidence-based cancer prevention tools such as *LungTalk*. Future research should explore embedding intervention content directly into social media platforms and tracking downstream clinical outcomes.

Trial Registration: ClinicalTrials.gov NCT05824273; https://clinicaltrials.gov/study/NCT05824273

International Registered Report Identifier (IRRID): RR2-10.1186/s12889-023-15857-8

(J Med Internet Res 2026;28:e80281) doi:10.2196/80281

KEYWORDS

lung cancer screening; RE-AIM; Reach, Effectiveness, Adoption, Implementation, and Maintenance; digital health communication; Facebook-targeted advertisement; implementation science; public health outreach

Introduction

Lung cancer remains the leading cause of cancer-related death in the United States, claiming more lives each year than breast, cervical, colorectal, and prostate cancers combined [1]. Although low-dose computed tomography of the chest is an effective screening tool that can detect lung cancer at early, more treatable stages [2], uptake of lung cancer screening (LCS) remains low, particularly among populations at highest risk [3,4]. Many eligible individuals are unaware that LCS exists or do not know that they meet the criteria, underscoring the urgent need for scalable, population-level communication strategies that increase awareness and engagement [5-7].

The RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) framework provides a widely accepted lens for evaluating the population impact of health interventions. Within this framework, *reach* is defined as the absolute number, proportion, and representativeness of individuals who are willing to participate in a given initiative, intervention, or program [8]. However, in digital communication contexts, *reach* should also include those exposed to messaging that serves as a gateway to an intervention, particularly when the goal is to raise awareness and to generate activation toward health care engagement upstream from the health system. Upstream exposure is critical, as prior research demonstrates that initial exposure to preventive health information can subsequently enhance readiness for clinical engagement [9,10].

Social media platforms such as Facebook offer a powerful and cost-effective channel for health communication outreach. Facebook-targeted advertisement (FBTA) enables precise delivery of content to users based on demographic characteristics and behavioral interests, allowing public health practitioners to connect with specific high-risk audiences in real time. Although FBTA has often been used for participant recruitment [11,12], its potential to achieve population-level exposure to health communication interventions has not been fully explored within an implementation science framework.

The objective of this study was to evaluate the *reach* of an FBTA campaign designed to connect LCS-eligible individuals with a digital health communication intervention. Specifically, we examined (1) the number and demographic characteristics of individuals exposed to the FBTA; (2) engagement metrics including click-through rates (CTRs) and progression through the study enrollment pathway; and (3) cost efficiency of FBTA as a population-level outreach strategy. To our knowledge, this is the first study to examine FBTA as a scalable public health outreach strategy for connecting high-risk individuals with a health communication intervention designed to promote awareness of, and informed decision-making about, LCS within an implementation science framework.

Methods

Conceptual Framework

This study was guided by the RE-AIM framework, a widely used implementation science model for evaluating the public health impact of interventions delivered in real-world settings [8]. In digital outreach efforts, conceptually, *reach* can be broadened to account for upstream, preintervention exposure that can set the stage for subsequent engagement and trial enrollment [13,14]. Previous studies using digital platforms indicate that broadening the conceptualization of *reach* to include initial digital exposure is important to accurately capture the public health potential of such interventions [15]. In this study, we operationalized *reach* as (1) the number of unique individuals exposed to the FBTA campaign (total reach); (2) the proportion who engaged with the content (CTR); and (3) the demographic representativeness of those exposed compared to the target population for LCS. This operationalization extends the traditional clinic-based conceptualization to include upstream digital exposure as a meaningful public health end point.

This campaign was designed to reach adults likely to meet LCS eligibility criteria by leveraging Facebook's advertisement targeting features (eg, age, interests, and keywords related to smoking or tobacco use). The ads delivered brief, visually engaging content, with images that served as proxy exposures to the visuals used in the health communication intervention to which participants would later be randomized. For example, some ads featured images of clinical encounters, such as a patient speaking with a clinician or undergoing a lung exam with a stethoscope—images that mirrored scenes from the health communication intervention called *LungTalk* (Table 1). Briefly, *LungTalk* is a computer-tailored health communication and

XSL•FO
RenderX

decision support tool designed to educate users about lung health, lung cancer risks, and LCS. The accompanying advertisement text was designed to recruit eligible participants into the study and included statements such as “seeking 50 - 80-year-old men and women for online study opportunity about lung health,” and “Are you between 50 and 80 years old and smoke cigarettes or used to smoke? Click to learn more and help improve patient education around lung screening!” Second, we assess user engagement with the FBTA content using Facebook’s built-in analytics to measure visibility (impressions), interaction (click-throughs), and efficiency (cost per click and cost per consent).

XSL•FO
RenderX

Table . Demographics and Facebook analytics across 5 Facebook-targeted advertisement campaigns for the INSPIRE-Lung Study (INnovating Social Media for Prevention: LUNG Cancer Screening Awareness, Knowledge, and Uptake), a randomized controlled trial to increase lung cancer screening awareness among adults aged 50 - 80 years with smoking-related interests in the United States from August 2023 to February 2024 (n=1,048,191 unique users reached). Images in this table were sourced from the Microsoft 365 stock image library and are reproduced under the Microsoft 365 license for stock content for documentation of study materials.

Facebook metrics	Campaign #1	Campaign #2	Campaign #3	Campaign #4	Campaign #5	Overall
Campaign dates	Aug 1 - 19, 2023	Sep 4 - 15, 2023	Oct 16-Nov 27, 2023	Nov 28, 2023-Jan 28, 2024	Dec 9 - 16, 2023	__[a]
Duration, d	19	12	43	62	8	144[b]
Total reach, n (%[c])	402,817 (38.4)	343,043 (32.7)	75,742 (7.2)	186,544 (17.8)	40,045 (3.8)	1,048,191 (100)
Female	216,513 (53.7)	193,090 (56.3)	55,502 (73.3)	108,804 (58.3)	8175 (20.4)	582,084 (55.5)
Male	180,992 (44.9)	145,921 (42.5)	19,679 (26)	77,162 (41.4)	31,806 (79.4)	455,560 (43.5)
Unknown	5312 (1.3)	4032 (1.2)	560 (0.7)	578 (0.3)	64 (0.2)	10,546 (1.0)
Link clicks, n (%[c])	7871 (31.7)	5889 (23.7)	7906 (31.9)	2659 (10.7)	491 (2)	24,816 (100)
Female: 45 - 54 y	276 (3.5)	267 (4.5)	353 (4.5)	141 (5.3)	22 (4.5)	1059
Female: 55 - 64 y	1112 (14.1)	852 (14.5)	1773 (22.4)	471 (17.7)	47 (9.6)	4255
Female: 65+ y	3785 (48.1)	2623 (44.5)	3761 (47.6)	1014 (38.1)	73 (14.9)	11,256
Male: 45 - 54 y	132 (1.7)	147 (2.5)	102 (1.3)	118 (4.4)	44 (9)	543
Male: 55 - 64 y	505 (6.4)	466 (7.9)	511 (6.5)	333 (12.5)	109 (22.2)	1924
Male: 65+ y	1809 (23)	1476 (25.1)	1319 (16.7)	566 (21.3)	107 (21.8)	5277
Unknown	252 (3.2)	58 (0.98)	87 (1.1)	16 (0.6)	89 (18.1)	502
Impressions, n (%[c])	751,251 (36.07)	692,496 (33.25)	183,078 (8.79)	361,983 (17.38)	93,682 (4.5)	2,082,490 (100)
Female: 45 - 54 y	27,127 (3.6)	27,734 (4)	7216 (3.9)	18,505 (5.1)	4473 (4.8)	85,055
Female: 55 - 64 y	98,284 (13.1)	96,094 (13.9)	44,591 (24.4)	73,422 (20.3)	8455 (9)	320,846
Female: 65+ y	285,392 (38)	270,208 (39)	78,966 (43.1)	123,969 (34.2)	12,632 (13.5)	771,166
Male: 45 - 54 y	35,629 (4.7)	28,269 (4.1)	2795 (1.5)	20,038 (5.5)	20,339 (21.7)	107,070
Male: 55 - 64 y	99,562 (13.3)	81,489 (11.8)	15,279 (8.3)	54,944 (15.2)	27,618 (29.5)	278,892
Male: 65+ y	205,257 (27.3)	188,702 (27.2)	34,231 (18.7)	71,105 (19.6)	20,165 (21.5)	519,460
CTR[d], %	1.96	1.72	10.44	1.43	1.23	2.37
Cost per click, US $	0.32	0.35	0.38	0.63	1.43	0.40
Total spent, US $	2499.55	2084.82	3000.00	1676.76	701.85	9962.98
Advertisement image						—
Advertisement text	Seeking 50-80-year-old men and women for a study opportunity about lung health.	Are you between 50-80 years old and smoke cigarettes or used to smoke?	Seeking 50-80-year-old men and women to learn more about lung health and screening.	Are you between 50-80 years old and interested in learning more about lung health and screening?	Are you between 50-80 years old and smoke cigarettes or used to smoke?	—

[a]Not applicable.

[b]Total days reflect 79 unique recruitment days; some campaigns overlapped (campaign 4 and 5 ran concurrently).

XSL•FO
RenderX

[c]The percentages in this main row were calculated by using the "Overall" value as the denominator. The percentages for the subcategories were calculated by using the corresponding n values from this row as the denominator.

[d]CTR: click-through rate (CTR = link clicks/reach ×100).

Most important, FBTA was not the intervention in itself. Rather, it was a communication medium for a study advertisement that enabled scalable outreach to high-risk individuals—adults aged 50‑80 years with a history of smoking—prior to their engagement with the health care system. These individuals represent a population that may have limited engagement with traditional clinical touchpoints (eg, annual physical exam and wellness visits) [16]. Given that LCS awareness remains below 10% nationally [17] and that eligible individuals who smoke are less likely to engage with preventive care [18], FBTA provides an opportunity to deliver targeted health messaging outside traditional clinical encounters—meeting individuals where they already spend time online and potentially catalyzing initial awareness that may precede health care engagement. As such, exposure to the FBTA served as a critical first step in a targeted health communication intervention engagement pathway.

Anchoring this analysis in the RE-AIM framework allows us to expand beyond basic metrics of trial enrollment. Instead, we offer a nuanced understanding of population-level *reach* by quantifying how digital advertising can expand the frontiers of public health communication, particularly for LCS, where knowledge gaps and access barriers persist. While the RE-AIM framework encompasses 5 dimensions, researchers have noted that these components can be examined individually when the research question warrants focused analysis of a specific dimension [14,19]. The present analysis focuses exclusively on *reach* because (1) this paper evaluates the FBTA outreach strategy as a discrete implementation component; and (2) establishing population-level *reach* is a necessary precursor to evaluating downstream dimensions. The remaining RE-AIM dimensions—effectiveness, adoption, implementation, and maintenance—will be evaluated in subsequent publications from the INSPIRE-Lung Study (INnovating Social Media for Prevention: LUNG Cancer Screening Awareness, Knowledge, and Uptake).

Study Overview

The INSPIRE-Lung Study (ClinicalTrials.gov NCT05824273) is a national, community-based randomized controlled trial designed to increase awareness, knowledge, and uptake of LCS among high-risk individuals [20]. To evaluate reach, we measured exposure to an FBTA campaign that served as a proxy for the health communication intervention.

Ethical Considerations

This study was approved by the Hackensack Meridian Health Institutional Review Board (Protocol: Pro2022-0860). Informed consent was obtained electronically from all the participants who enrolled in the trial. For aggregated Facebook analytics data (reach, impressions, and clicks), a waiver of informed consent was granted, as these data are anonymized and do not contain personally identifiable information. No personal identifying information was retained from users who only viewed advertisements. All data were deidentified prior to analysis to protect participant privacy and confidentiality. Participants received US $50 for completing baseline and 1-week follow-up surveys and US $25 for completing the 6-month follow-up survey.

FBTA Campaigns

Between August 2023 and February 2024, 5 discrete FBTA campaigns were used to reach adults aged 50‑80 years in the United States. Campaign durations ranged from 8 to 62 days (Table 1). Campaigns varied in messaging and imagery to test different engagement approaches; however, the primary focus of this analysis is cumulative *reach* rather than comparative campaign effectiveness. Advertisements were developed using Facebook's advertisement manager platform and tailored to target users based on age, geographic location, and interests related to smoking (eg, cigarettes, vaping, smoking cessation, and nicotine replacement therapy).

Each advertisement linked to the study landing page that briefly described the purpose of the trial and invited users to complete a screening questionnaire to determine LCS eligibility. Although users were not exposed to the full health communication intervention unless they enrolled in the trial, the FBTA messaging was designed to reflect the tone and content focus of health messaging around LCS, emphasizing that screening can save lives and is available to eligible adults with a history of smoking. Thus, FBTA served as a population-level outreach tool and an upstream proxy exposure to the intervention.

Campaigns varied in wording, imagery, and call-to-action phrasing to optimize engagement. Prior to deployment, all ads were reviewed for cultural appropriateness, literacy level, and ethical compliance with best practices for digital recruitment. Real-time monitoring allowed for iterative campaign refinements based on audience response.

Reach: Definition and Metrics

To evaluate *reach* in alignment with the RE-AIM framework, we used both Facebook analytics [21] and internal study data. Facebook analytics [21] provided the real-time performance metrics as follows:

- *Total reach:* The number of unique users who viewed the FBTA content at least once
- *Impressions:* The total number of times the FBTA appeared on users' screens (which may include multiple exposures per user)
- *Link clicks:* The number of users who clicked the FBTA and were directed to the study website
- *CTR:* The percentage of viewers who clicked on the advertisement
- *Cost metrics:* Cost per click and cost per consented participant

An important consideration is that Facebook's targeting capabilities allow selection based on age and behavioral interests (eg, smoking-related content) but cannot verify clinical eligibility criteria such as the minimum 20 pack-year smoking

XSL•FO
RenderX

history. Therefore, the total reach metric represents exposure among individuals likely but not confirmed to meet screening eligibility. Most importantly, for the purpose of this aim, we conceptualized *reach* as population-level exposure to a public-facing health communication message (delivered via FBTA), independent of whether individuals proceeded to trial enrollment. This approach aligns with RE-AIM's emphasis on assessing exposure and potential engagement among the broader target population.

Analytic Strategy

Descriptive statistics were used to summarize campaign-level and cumulative metrics across the 5 FBTA campaigns. We calculated total reach, impressions, CTRs, and cost-efficiency measures. We also examined FBTA impressions by age and sex distribution (Table 1). To assess representativeness of digital engagement, we compared the demographic distribution of link clicks to impressions using chi-square tests. Because Facebook analytics provide aggregate demographic data on impressions rather than unique viewers, we used impression distribution as a proxy for viewer demographics in these comparisons. Additionally, we examined user progression through the digital pathway—from advertisement exposure to click-through to screener completion and trial enrollment—to contextualize how FBTA contributed to downstream engagement with the full health communication intervention.

Results

Over a 79-day recruitment period between August 2023 and February 2024, 5 US-based FBTA campaigns were launched nationally to reach high-risk individuals eligible for LCS. Across all campaigns, a total of 1,048,191 unique Facebook users were exposed to at least 1 advertisement, representing the total reach of the FBTA strategy. These users generated 3,109,482 total impressions, indicating repeated exposure to the content over time. Among those reached, 24,816 (2.37%) users clicked on the advertisement, which directed them to the study landing page. Of these, 7117 (28.6%) individuals completed the brief eligibility screening questionnaire. A total of 1272 (17.9% of screeners) individuals met LCS eligibility criteria, and of these, 483 (38%) enrolled in the randomized trial. Table 1 details demographic characteristics across the 5 campaigns. Table 2 details the demographic characteristics of participants enrolled in the study.

XSL•FO
RenderX

Table . Demographic characteristics of participants enrolled in the INSPIRE-Lung Study (INnovating Social Media for Prevention: LUNG Cancer Screening Awareness, Knowledge, and Uptake), a community-based randomized controlled trial to increase lung cancer screening awareness, recruited via Facebook-targeted advertisement in the United States from August 2023 to February 2024 (n=483).

Characteristic	Participants
Age (y), mean (SD)	63.3 (6.7)
Sex, n (%)	
Male	236 (48.9)
Female	246 (50.9)
Nonbinary	1 (0.2)
Race, n (%)	
White	376 (79.3)
Black	70 (14.8)
Asian	7 (1.5)
American Indian	12 (2.5)
Native Hawaiian/Alaskan	2 (0.4)
Multiracial	7 (1.5)
Hispanic, n (%)	
Yes	23 (4.8)
No	454 (95.2)
Education, n (%)	
Less than high school	10 (2.1)
High school graduate	56 (11.6)
Some college	144 (29.9)
College graduate or higher	272 (56.4)
Income, n (%)	
<US $25,000/y	102 (21.2)
$US 25,000-US $50,000/y	141 (29.3)
>US $50,000/y	238 (49.5)
Smoking history (y), mean (SD)	35.5 (10.3)
Currently smokes, n (%)	
Yes	313 (64.8)
No	170 (35.2)

These data demonstrated the substantial *reach* of the FBTA campaign as an upstream population-level exposure strategy. Table 3 presents key campaign metrics across the 5 FBTA iterations. The average cost per link click was US $0.40, and the total campaign cost was US $9962.98, yielding a cost per enrolled participant of US $19.46. Performance varied across campaigns, with Campaign #1 generating the highest total reach (38.4%) and the lowest cost per click (US $0.32). Campaign #3 generated the most link clicks (7906; 31.9%) despite lower total reach. These variations may reflect differences in messaging, imagery, timing, or audience saturation and will be explored in future work.

XSL•FO
RenderX

Table . Summary of Facebook-targeted advertisement campaign performance metrics for the INSPIRE-Lung Study (INnovating Social Media for Prevention: LUNG Cancer Screening Awareness, Knowledge, and Uptake), a randomized controlled trial targeting lung cancer screening–eligible adults in the United States from August 2023 to February 2024.

Metric	Value[a]
Total reach, number of unique Facebook users	1,048,191
Total impressions, n	3,109,482
Total link clicks, n	24,816
Click-through rate, %	2.36
Screener completers, n	7117
Screen-eligible respondents, n	1272
Enrolled participants, n	483
Total campaign cost, US $	9962.98
Cost per click, US $	0.40
Cost per consent, US $	19.46

[a]These values are total social media and study metrics.

Participants reached through the FBTA campaign were geographically and demographically diverse, with representation across all 50 US states (Figure 1). Among those who enrolled, the average age was 63.3 (SD 6.7) years, and 50.9% identified as female participants.

Figure 1. Geographic distribution of INSPIRE-Lung Study participants enrolled via Facebook-targeted advertisement across all 50 US states from August 2023 to February 2024 (n=483). INSPIRE-Lung Study: INnovating Social Media for Prevention: LUNG Cancer Screening Awareness, Knowledge, and Uptake.

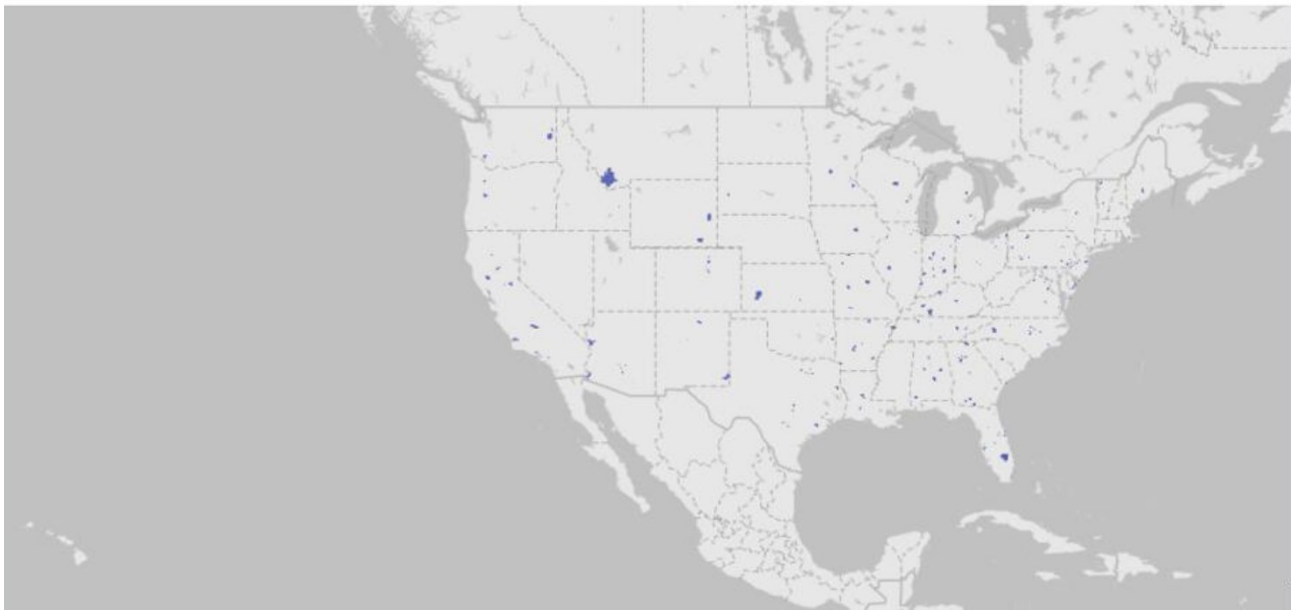

To examine representativeness, we compared the demographic distribution of individuals who clicked on the FBTA to those who viewed the advertisements (using impressions as a proxy for exposure). Chi-square tests revealed significant differences in engagement patterns by age, sex, and geography. Regarding age, individuals aged 65 years and older comprised 61.9% of total impressions but 68% of link clicks, indicating higher engagement among older viewers (χ^2_2=436.86; *P*<.001). CTRs increased with age: 0.83% for ages 45 - 54 years, 1.03% for ages 55 - 64 years, and 1.28% for ages 65+ years. Regarding sex, female participants comprised 56.5% of impressions but 68.2% of link clicks, demonstrating significantly higher engagement compared to male participants (χ^2_1=1337.74; *P*<.001). The CTR for female participants (1.41%) was notably higher than for male participants (0.86%). Geographic distribution also differed significantly across US Census regions (χ^2_3=26.89; *P*<.001), with the South accounting for the highest proportion of both impressions (36.2%) and clicks (37.1%).

However, even among individuals who did not proceed to enrollment, exposure to the FBTA message about LCS represents a meaningful first step toward increasing awareness and consideration of screening. This broader lens on *reach* captures the use of digital messaging strategies upstream from health care system engagement.

XSL•FO
RenderX

Discussion

Principal Findings

This study demonstrates the feasibility and use of FBTA as a digital strategy to achieve broad, population-level *reach* to LCS-eligible individuals using tailored health messaging. Guided by the RE-AIM framework [8], we operationalized *reach* not as trial enrollment but as exposure to a communication message that served as an upstream proxy for a health communication intervention. Across 5 campaigns in the United States, we reached over 1 million unique individuals within the target age range and smoking history risk profile for LCS. These findings highlight the potential of FBTA to serve as a scalable public health outreach tool that promotes engagement with evidence-based communication interventions.

Most importantly, this analysis distinguishes between *exposure* and *enrollment*. While a subset of users exposed to the FBTA ultimately enrolled in the trial and received the *LungTalk* intervention or control condition depending on randomization, the majority did not progress beyond the initial digital touchpoint. Nonetheless, this first contact with a tailored screening message represents a meaningful public health outcome, particularly in a context where lack of awareness is a persistent barrier to LCS uptake. Although FBTA effectively reached individuals who may not have engaged with traditional health care outreach, the extent to which initial digital exposure translates into sustained clinical behavior change remains uncertain and warrants further longitudinal investigation. In addition, given the fact that LCS awareness is so low in general (<10%) [17], increasing population level awareness regardless of screening eligibility is of great value. In line with RE-AIM, conceptualizing *reach* in this broader way captures the real-world potential of digital tools to connect individuals with preventive health information upstream of clinical interaction. The remaining RE-AIM dimensions—effectiveness, adoption, implementation, and maintenance—will be evaluated in subsequent publications from the INSPIRE-Lung Study. Specifically, effectiveness will be assessed through the primary RCT outcomes examining *LungTalk's* impact on LCS awareness, knowledge, and screening uptake. Adoption and implementation metrics will examine health care system and provider engagement, while maintenance will assess sustained behavior change at 6- and 12-month follow-ups.

To better understand the representativeness of the FBTA campaigns, we compared the demographics of enrolled participants with the screening population in the American College of Radiology's Lung Cancer Screening Registry (ACR LCSR), as reported by Silvestri et al [22]. In that study of over 1 million individuals screened for lung cancer with low-dose computed tomography between 2015 and 2019, the population was predominantly White (91.6%), with only 7.4% Black and 0.9% Asian individuals among those with known race/ethnicity data. In contrast, our FBTA-recruited sample was notably more diverse: 79.3% White, 14.8% Black (double the ACR LCSR proportion), and 4.8% Hispanic. This suggests that FBTA may be particularly effective at reaching racial and ethnic minority populations who have historically been underrepresented in clinic-based LCS.

Additional demographic patterns also differed from the ACR LCSR. While the registry population skewed older adults (42.6% aged 65+ y) and female participants (approximately 54%), our enrolled participants showed a more balanced sex distribution (50.9% female participants, 48.9% male participants) and a mean age of 63.3 (SD 6.7) years. The ACR LCSR also reported that 61.4% of screened individuals currently smoked; similarly, 64.8% of our enrolled participants currently smoked, indicating that FBTA successfully reached individuals with active smoking behaviors who represent a priority population for LCS outreach. Notably, Black and Hispanic individuals demonstrated lower adherence to annual screening in the ACR LCSR, suggesting that alternative outreach strategies like FBTA may help address these disparities at the initial awareness and engagement stage.

However, certain groups, such as those with lower educational attainment or without internet access, may still be underrepresented in digital outreach, underscoring the importance of multichannel approaches. In our sample, 56.4% had a college degree or higher, which exceeds national averages and suggests potential selection bias toward more educated individuals.

The FBTA strategy also proved highly cost-efficient. While the average cost per link click was US $0.40 and the cost per enrolled participant was US $19.46, the cost per view (or impression) was substantially lower, at just over US $0.003 per exposure. This underscores the potential of FBTA to deliver high-volume, targeted health communication at minimal per-person cost. Our cost per enrolled participant (US $19.46) compares favorably to published estimates for traditional recruitment methods in cancer prevention research. Community-based recruitment strategies have reported costs ranging from US $50 to over US $500 per participant, depending on intensity and population [23,24]. Direct mail campaigns for health research typically cost US $100 - 200 per enrolled participant [25]. Our prior work using FBTA for LCS research demonstrated similar cost efficiencies [11]. Compared to traditional outreach strategies such as print mailings, community events, telephone outreach, or provider referrals, which often require substantial personnel time and logistical coordination, FBTA offers real-time precision targeting and scalability. This is especially advantageous for reaching populations who may have limited engagement with the health care system. According to the Pew Research Center (2024), approximately 68% of the US adults use Facebook, with relatively consistent usage across income levels and geographic regions [26]. Most importantly, Facebook penetration among adults aged 50 - 64 (70%) years and 65+ (50%) years approaches or mirrors general population estimates, suggesting potential to reach individuals across the socioeconomic spectrum. While Facebook users as a whole may not be systematically disconnected from health care, our targeting strategy—focusing on individuals with smoking-related interests—may identify a subpopulation with elevated barriers to care, as prior research indicates that individuals who currently smoke are less likely to have a usual source of care and less likely to receive preventive services [18]. Moreover, the low rate of LCS uptake nationally (<6%) suggests substantial

XSL•FO
RenderX

numbers of eligible individuals are not engaging with screening through traditional clinical pathways. FBTA may be particularly valuable for reaching individuals living in rural areas, those without consistent primary care access, and those less likely to encounter LCS information through conventional clinical channels.

Limitations

Nevertheless, several limitations should be acknowledged. First, while FBTA allowed for targeted delivery of content based on self-reported profile characteristics and behavioral interests, we could not verify screening eligibility (eg, 20+ packs per year smoking history) until individuals clicked through to the screening questionnaire. As such, Facebook's *reach* metric represents exposure among individuals likely but not confirmed to meet screening eligibility criteria and may include a significant proportion who are not personally eligible for LCS. The fact that only 17.9% of those who completed the screener met eligibility criteria illustrates this gap between interest-based targeting and clinical eligibility. However, we conceptualize this exposure as meaningful from a public health perspective: raising awareness about LCS, even among some individuals who may not personally meet eligibility criteria, has value, as these individuals may share information with eligible family members or friends. Given the inherently social nature of Facebook, this kind of peer-sharing and diffusion of information could amplify the campaign's impact within networks of individuals with shared interests in smoking or cessation.

Second, although the study provides valuable insights into digital *reach*, it does not assess the quality or depth of engagement with the FBTA message among those who viewed but did not interact with the advertisement. Future studies could incorporate passive analytics (eg, time spent viewing or scrolling behavior) or embed brief *LungTalk*-derived content within the advertisement itself to assess this dimension.

Third, there are broader ethical and equity considerations when leveraging commercial social media platforms for health communication. While Facebook offers powerful precision-targeting capabilities, Facebook algorithms may inherently exclude or underrepresent certain demographic groups, potentially biasing exposure toward users already digitally active or engaged, which could inadvertently widen health disparities [27]. Ensuring equitable digital engagement will be essential as these strategies are scaled [27]. Future research should systematically monitor and adjust targeting criteria to mitigate this risk.

Fourth, we cannot definitively rule out exposure to fake or bot accounts, though Facebook's platform integrity measures are designed to minimize such accounts. Additionally, our downstream metrics (screener completion, eligibility verification, and trial enrollment) require human engagement that automated accounts cannot replicate, providing confidence that our enrollment figures represent genuine participants.

Despite these limitations, our findings have important implications for cancer prevention and public health communication. Reaching high-risk individuals who are not yet engaged with the health care system remains a major barrier to improving LCS rates. FBTA offers a powerful, scalable method to extend the reach of evidence-based interventions like *LungTalk* and to promote awareness among populations traditionally missed by clinic-based strategies.

Future efforts should explore how social media–based outreach can be integrated with other implementation strategies, such as patient navigation or community partnerships, to support screening uptake. Additionally, embedding the intervention content itself directly into the FBTA (rather than linking out) will allow for even greater scalability and immediacy of impact.

Next Steps and Future Directions

Building on the demonstrated feasibility and population-level *reach* of FBTA to connect with LCS-eligible individuals, several strategic opportunities emerge for advancing digital public health communication and implementation efforts.

First, future research should explore embedding core components of the intervention, such as tailored *LungTalk* messaging or *LungTalk* itself, directly within the FBTA interface. This would allow for immediate engagement with decision support content at the moment of exposure, reducing drop-off and increasing the likelihood that viewers absorb key information, even if they do not proceed to enrollment. Embedding brief video clips, testimonials, or tailored visual messages adapted from *LungTalk* could help bridge the gap between exposure and education. For example, a randomized comparative effectiveness study could assess whether embedding *LungTalk* intervention components directly within FBTA enhances immediate knowledge gains compared to ads linking externally to a study website.

Second, optimizing message design and delivery remains critical. Future campaigns may consider using systematic A or B testing of message framing, imagery, tone, and calls to action, with attention to behavioral science principles and cultural relevance. Tailoring content for priority populations, such as Black, Hispanic or Latino, or rural adults with elevated lung cancer risk, can further improve engagement and equity.

Third, future studies should examine how digital *reach* translates into downstream behavioral and clinical outcomes. For example, FBTA-driven exposure could be linked to subsequent engagement with patient navigation, telehealth consults, or direct referral systems to LCS programs. Implementation strategies that combine digital outreach with systems-level linkage to care hold promise for closing the gap between awareness and screening uptake.

Fourth, collaboration across outreach and communication sectors will be essential for scaling impact. Partnering with health systems, public health departments, advocacy organizations, and community-based coalitions can support broader dissemination, localized adaptation, and sustainable infrastructure for ongoing FBTA campaigns.

Finally, ensuring equity in digital health communication is imperative. As social media becomes an increasingly common tool for public health messaging, attention must be paid to disparities in internet access, digital literacy, and algorithmic exposure. Mixed methods evaluations, community-engaged

XSL•FO
RenderX

research, and monitoring of campaign delivery across demographic groups can help identify and address gaps. Most importantly, even users who are not personally eligible for LCS may serve as information conduits within their social networks, amplifying campaign effects through peer sharing and expanding awareness among others who may have missed the original exposure.

In addition, FBTA offers a scalable, adaptable, effective, and cost-efficient platform for delivering tailored health communication content upstream of clinical care. By refining and integrating this strategy with broader implementation efforts, future research and practice can more effectively promote equitable access to LCS information and ultimately reduce disparities in early detection and outcomes.

Conclusions

This study demonstrates the potential of FBTA to achieve meaningful population-level *reach* to individuals eligible for LCS, as defined by the RE-AIM framework. By delivering a tailored communication message that served as a proxy exposure to a health communication intervention that promotes awareness and informed decision-making about LCS, FBTA enabled contact with over 1 million adults within the eligible age range, many of whom are unlikely to encounter such information through traditional health care pathways.

By broadening *reach* to encompass initial exposure to health-promoting digital messages, this analysis highlights the importance of upstream engagement in implementation science and public health practice. FBTA proved to be a cost-efficient and scalable outreach strategy capable of penetrating social networks, initiating awareness, and opening pathways for subsequent engagement with LCS services.

As public health increasingly embraces digital platforms to extend its reach, integrating social media strategies with evidence-based interventions offers a promising direction for reducing disparities in screening and early detection. Future efforts should build on this foundation by embedding intervention content directly into advertisement formats, testing links to downstream care, and ensuring equitable access across diverse communities.

Funding

This work was supported by grant no. R01CA263662 (principal invesigator [PI]: LC-B) received from the National Cancer Institute. LC-B is supported in part by Georgetown University Lombardi CCSG P30CA051008 (PI: Louis Weiner). RSW is supported in part by resources from the VA Boston Healthcare system. JSO and SCB are supported in part by MSK CCSG P30CA008748 (PI: Selwyn Vickers).

Disclaimer

The views expressed herein are those of the authors and do not necessarily reflect the views of the Department of Veterans Affairs or the US government.

Data Availability

The Facebook analytics data presented in this study are available from the corresponding author upon reasonable request. Individual-level participant data from the INSPIRE-Lung Study (INnovating Social Media for Prevention: LUNG Cancer Screening Awareness, Knowledge, and Uptake) are not publicly available due to privacy and ethical restrictions but may be available upon reasonable request and with appropriate data use agreements.

Authors' Contributions

Conceptualization: LC-B, JSO, SMR
Data curation: LC-B, RSC, FV
Formal analysis: LC-B, POM, JES Jr
Funding acquisition: LC-B
Investigation: LC-B, EAH, AGV
Methodology: LC-B, JSO, SMR, SCB
Project administration: LC-B, FV
Resources: LC-B, AC
Software: RSC
Supervision: LC-B
Validation: LC-B, EAH
Visualization: LC-B, RSC
Writing – original draft: LC-B
Writing – review & editing: LC-B, JSO, SMR, EAH, SCB, AC, RSC, MK, KTL, POM, JES Jr, FV, RSW, AGV

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Cancer facts & figures 2023. : American Cancer Society, Inc; 2023 URL: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2023/2023-cancer-facts-and-figures.pdf [accessed 2026-02-03]
2. National Lung Screening Trial Research Team, Aberle DR, Adams AM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011 Aug 4;365(5):395-409. [doi: 10.1056/NEJMoa1102873] [Medline: 21714641]
3. Rivera MP, Katki HA, Tanner NT, et al. Addressing disparities in lung cancer screening eligibility and healthcare access. An official American Thoracic Society statement. Am J Respir Crit Care Med 2020 Oct 1;202(7):e95-e112. [doi: 10.1164/rccm.202008-3053ST] [Medline: 33000953]
4. Racial and ethnic disparities. American Lung Association, State of Lung Cancer. 2023. URL: https://www.lung.org/research/state-of-lung-cancer/racial-and-ethnic-disparities [accessed 2026-02-03]
5. Carter-Harris L, Ceppa DP, Hanna N, Rawl SM. Lung cancer screening: what do long-term smokers know and believe? Health Expect 2017 Feb;20(1):59-68. [doi: 10.1111/hex.12433] [Medline: 26701339]
6. Carter-Harris L, Tan ASL, Salloum RG, Young-Wolff KC. Patient-provider discussions about lung cancer screening pre- and post-guidelines: Health Information National Trends Survey (HINTS). Patient Educ Couns 2016 Nov;99(11):1772-1777. [doi: 10.1016/j.pec.2016.05.014] [Medline: 27241830]
7. Williamson TJ, Walsh LE, Rawl SM, Carter-Bawa L. Slipping through the cracks: who is eligible but does not receive a healthcare provider recommendation for lung cancer screening? Lung Cancer (Auckl) 2023 May;179:107185. [doi: 10.1016/j.lungcan.2023.107185] [Medline: 37023535]
8. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
9. Hibbard JH, Greene J. What the evidence shows about patient activation: better health outcomes and care experiences; fewer data on costs. Health Aff (Millwood) 2013 Feb;32(2):207-214. [doi: 10.1377/hlthaff.2012.1061] [Medline: 23381511]
10. Prochaska JO, DiClemente CC. Stages and processes of self-change of smoking: toward an integrative model of change. J Consult Clin Psychol 1983 Jun;51(3):390-395. [doi: 10.1037//0022-006x.51.3.390] [Medline: 6863699]
11. Carter-Harris L, Bartlett Ellis R, Warrick A, Rawl S. Beyond traditional newspaper advertisement: leveraging Facebook-targeted advertisement to recruit long-term smokers for research. J Med Internet Res 2016 Jun 15;18(6):e117. [doi: 10.2196/jmir.5502] [Medline: 27306780]
12. Walsh LE, Carter-Bawa L. Using social media to recruit individuals for health-related research: feasibility and lessons learned. J Health Psychol 2023 Jun;28(7):599-606. [doi: 10.1177/13591053221136928] [Medline: 36416197]
13. Kessler RS, Purcell EP, Glasgow RE, Klesges LM, Benkeser RM, Peek CJ. What does it mean to “employ” the RE-AIM model? Eval Health Prof 2013 Mar;36(1):44-66. [doi: 10.1177/0163278712446066] [Medline: 22615498]
14. Glasgow RE, Estabrooks PE. Pragmatic applications of RE-AIM for health care initiatives in community and clinical settings. Prev Chronic Dis 2018 Jan 4;15:E02. [doi: 10.5888/pcd15.170271] [Medline: 29300695]
15. Eppes EV, Augustyn M, Gross SM, Vernon P, Caulfield LE, Paige DM. Engagement with and acceptability of digital media platforms for use in improving health behaviors among vulnerable families: systematic review. J Med Internet Res 2023 Feb 3;25:e40934. [doi: 10.2196/40934] [Medline: 36735286]
16. Haddad DN, Sandler KL, Henderson LM, Rivera MP, Aldrich MC. Disparities in lung cancer screening: a review. Ann Am Thorac Soc 2020 Apr;17(4):399-405. [doi: 10.1513/AnnalsATS.201907-556CME] [Medline: 32017612]
17. Williamson TJ, Brymwitt WM, Hirsch EA, Reese MT, Carter-Bawa L. Patient-clinician discussions on lung cancer screening in the United States before and after 2021 guidelines. Transl Behav Med 2025 Jan 16;15(1):ibaf039. [doi: 10.1093/tbm/ibaf039] [Medline: 40803694]
18. Cornelius ME, Loretan CG, Jamal A, et al. Tobacco product use among adults - United States, 2021. MMWR Morb Mortal Wkly Rep 2023 May 5;72(18):475-483. [doi: 10.15585/mmwr.mm7218a1] [Medline: 37141154]
19. Holtrop JS, Estabrooks PA, Gaglio B, et al. Understanding and applying the RE-AIM framework: clarifications and resources. J Clin Transl Sci 2021;5(1):e126. [doi: 10.1017/cts.2021.789] [Medline: 34367671]
20. Carter-Bawa L, Banerjee SC, Comer RS, et al. Leveraging social media to increase lung cancer screening awareness, knowledge and uptake among high-risk populations (the INSPIRE-Lung Study): study protocol of design and methods of a community-based randomized controlled trial. BMC Public Health 2023 May 26;23(1):975. [doi: 10.1186/s12889-023-15857-8] [Medline: 37237339]
21. Meta. About reach. Facebook Business Help Center. URL: https://www.facebook.com/business/help/710746785663278 [accessed 2026-02-03]
22. Silvestri GA, Goldman L, Tanner NT, et al. Outcomes from more than 1 million people screened for lung cancer with low-dose CT imaging. Chest 2023 Jul;164(1):241-251. [doi: 10.1016/j.chest.2023.02.003] [Medline: 36773935]
23. Treweek S, Pitkethly M, Cook J, et al. Strategies to improve recruitment to randomised trials. Cochrane Database Syst Rev 2018 Feb 22;2(2):MR000013. [doi: 10.1002/14651858.MR000013.pub6] [Medline: 29468635]
24. UyBico SJ, Pavel S, Gross CP. Recruiting vulnerable populations into research: a systematic review of recruitment interventions. J Gen Intern Med 2007 Jun;22(6):852-863. [doi: 10.1007/s11606-007-0126-3] [Medline: 17375358]

XSL•FO
RenderX

25. Coley HL, Sadasivam RS, Williams JH, et al. Crowdsourced peer- versus expert-written smoking-cessation messages. Am J Prev Med 2013 Nov;45(5):543-550. [doi: 10.1016/j.amepre.2013.07.004] [Medline: 24139766]
26. Social media fact sheet. Pew Research Center. 2025 Nov 20. URL: https://www.pewresearch.org/internet/fact-sheet/social-media/ [accessed 2026-02-03]
27. Arigo D, Pagoto S, Carter-Harris L, Lillie SE, Nebeker C. Using social media for health research: methodological and ethical considerations for recruitment and intervention delivery. Digit Health 2018;4:2055207618771757. [doi: 10.1177/2055207618771757] [Medline: 29942634]

Abbreviations

ACR LCSR: American College of Radiology's Lung Cancer Screening Registry
CTR: click-through rate
FBTA: Facebook targeted advertisement
INSPIRE-Lung Study: INnovating Social Media for Prevention: LUNG Cancer Screening Awareness, Knowledge, and Uptake
LCS: lung cancer screening
RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance

Edited by A Mavragani; submitted 07.Jul.2025; peer-reviewed by I -Wen Pan, R Williams; revised version received 02.Jan.2026; accepted 20.Jan.2026; published 03.Mar.2026.

Please cite as:
Carter-Bawa L, Ostroff JS, Rawl SM, Hirsch EA, Banerjee SC, Ciupek A, Comer RS, Kale M, Leopold KT, Monahan PO, Slaven Jr JE, Valenzona F, Wiener RS, Vielma AG
Leveraging Social Media to Achieve Population-Level Reach of Lung Cancer Screening-Eligible Individuals: A RE-AIM Framework Perspective
J Med Internet Res 2026;28:e80281
URL: https://www.jmir.org/2026/1/e80281
doi:10.2196/80281

© Lisa Carter-Bawa, Jamie S Ostroff, Susan M Rawl, Erin A Hirsch, Smita C Banerjee, Andrew Ciupek, Robert Skipworth Comer, Minal Kale, Katherine T Leopold, Patrick O Monahan, James E Slaven Jr, Francis Valenzona, Renda Soylemez Wiener, Ana Guadalupe Vielma. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 3.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Diabetes Technologies in Ultra-Endurance Type 1 Diabetes: Qualitative Study

Jean-Charles Vauthier[1,2*], Dr med; Lucie Choley[1*]; Delphine Arduini[3]; Patrick Mas[4]; Bernard Kabuth[1,2,5], Prof Dr Med

[1]Faculty of Medicine, Maieutics and Health Professions, Université de Lorraine, 9 avenue de la Forêt de Haye, BP 20199, Vandoeuvre Lès Nancy, France

[2]Laboratoire INTERPSY EA 4432, Université de Lorraine, Nancy, France

[3]Independent Patient Partner, Saint Restitut, France

[4]Independent Patient Partner, Aigues Mortes, France

[5]Centre Psychothérapique de Nancy (CPN), Université de Lorraine, LAXOU, France

[*]these authors contributed equally

Corresponding Author:
Jean-Charles Vauthier, Dr med
Faculty of Medicine, Maieutics and Health Professions, Université de Lorraine, 9 avenue de la Forêt de Haye, BP 20199, Vandoeuvre Lès Nancy, France

Abstract

Background: Diabetes technologies—including continuous glucose monitoring (CGM), insulin pumps, and hybrid closed-loop systems—have profoundly transformed self-management in type 1 diabetes (T1D). While these technologies offer improved glycemic control and safety, their use in ultraendurance sports introduces specific cognitive, material, and organizational challenges that remain underexplored in digital health research.

Objective: This study aimed to explore how adults living with T1D experience and use diabetes technologies in ultraendurance sports, with particular attention to tensions between autonomy, mental load, and vulnerability.

Methods: We conducted semistructured interviews with 13 French-speaking adults with T1D who had completed at least one marathon or ultra-endurance event within the last 5 years and used ≥1 diabetes technology (CGM, pump, or hybrid closed loop). We adopted constructivist grounded theory (Charmaz), using iterative cycles of line-by-line and focused coding, constant comparison, and memo-writing to build and refine analytic categories. Sampling combined purposive strategies through associations and online communities with theoretical orientation (additional participants sought to elaborate emergent categories). Data collection ceased upon theoretical sufficiency, when further interviews no longer yielded substantively new insights for core categories. Two patient partners contributed to question framing, interim sense-checking, and manuscript review. Reporting followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.

Results: Five interrelated categories described how athletes negotiated technology in practice: (1) From episodic control to continuous anticipation (reframing glucose management through real-time visibility); (2) Gains in safety and performance (perceived benefits and expanded possibilities); (3) Redistributed mental work (hyper-vigilance, logistics, device management); (4) Keeping things working when they break (fragility in extreme conditions, redundancy, improvisation, and experiential expertise); and (5) Making diabetes visible (technologies mediating identity, solidarity, and stigma). Across categories, participants articulated a tension between optimization-oriented performance and a user-constructed robustness—the capacity to maintain function under uncertainty through redundancy and adaptive know-how.

Conclusions: In ultraendurance contexts, diabetes technologies act as both enablers and obligations: they open participation while shifting and sometimes intensifying cognitive and organizational work. A grounded account centered on robustness-in-use highlights practical implications for clinicians (pre-event routines, redundancy planning), designers (context-aware algorithms; improved physical durability), and policy makers (equitable access and exercise-specific education). These findings underscore the value of constructivist, practice-oriented inquiry to inform digital health tool design and support for people living with chronic illness.

(J Med Internet Res 2026;28:e86815) doi:10.2196/86815

KEYWORDS

diabetes mellitus, type 1; type 1 diabetes; continuous glucose monitoring; exercise; self-management; patient participation; qualitative research

XSL•FO
RenderX

Introduction

Diabetes technologies—including continuous glucose monitoring (CGM), insulin pumps, and increasingly automated insulin delivery or hybrid closed-loop (AID/HCL) systems—have reshaped self-management in type 1 diabetes (T1D) over the last decade [1]. Beyond well-documented glycemic benefits, these systems have been associated with improvements in quality of life for many users, although the extent and nature of these benefits vary across contexts and populations [1,2]. In ultraendurance sports, where physiological demands and environmental stressors are amplified, such technologies may enable participation but can also introduce distinct challenges in day-to-day decision-making [1,3].

Recent work in exercise physiology and diabetes technology has highlighted how prolonged aerobic activity, heat, altitude, or mechanical stress can alter insulin kinetics, sensor accuracy, and carbohydrate absorption, thereby challenging the promises of algorithmic safety and stability [4,5]. Similar concerns have been raised in endurance athletes without diabetes, where CGM accuracy decreases under movement, heat, and prolonged exertion [6]. Although AID/HCL systems offer more automation than previous generations, they still require substantial user input before, during, and after exercise, including anticipatory basal adjustments, activation of exercise modes, and flexible carbohydrate strategies [4]. These adjustments, critical for preventing hypoglycemia, impose organizational and cognitive work that may be heightened in long-duration events.

Qualitative and patient-centered studies show that diabetes technologies can produce both empowerment and burden: continuous data streams require interpretation, alarms may interrupt activity, and device maintenance generates ongoing logistical work [2,7,8]. Reports of data overload, trust calibration, adhesive failures, and the emotional labor of hyper-vigilance are increasingly documented in real-world settings, suggesting that technological autonomy is contingent and situated rather than inherent to device design [7,8]. These issues may be particularly salient in ultraendurance contexts, where environmental variability and long exposure magnify device fragility.

Recent syntheses emphasize that AID/HCL systems deliver robust gains in time in range and reduce hypoglycemia across age groups and life stages [1], yet they still require substantial user input (eg, carbohydrate announcements) and do not eliminate the need for anticipatory adjustments when exercising. Specifically for exercise, experts recommend pre-emptive basal reductions or activation of exercise modes 60 - 120 minutes before activity and additional carbohydrate intake adapted to intensity and duration—practices that impose planning work and may be difficult to implement during long events or unanticipated efforts [4]. These observations point to a persistent gap between algorithmic promises and lived exercise realities [1]. Similar limitations have also been observed in endurance athletes without diabetes, where CGM-informed carbohydrate strategies produced steadier glucose trends but did not improve performance during prolonged exercise [9], suggesting that even CGM-informed strategies may have context-dependent value when demands and exposure accumulate.

From a lived-experience perspective, qualitative and mixed-methods studies highlight both empowerment and cognitive burden associated with CGM and pump-AID/HCL use: constant data streams demand interpretation, alarms can be reassuring yet intrusive, and device maintenance generates logistical work [2,7]. Users report benefits in confidence and reduced diabetes distress [2,5], but also describe anxiety related to device reliability, data overload, or social visibility of devices [7]. The overall picture suggests that technologies may re-distribute rather than erase the mental work of self-management—an effect that prolonged effort, exposure to heat, moisture, or mechanical shocks could intensify in ultra-endurance conditions [1,4,8].

Despite advances in diabetes technologies and growing interest in exercise applications, little is known about how adults with T1D practically adapt these systems in ultra-endurance sports. Existing research has focused primarily on metabolic outcomes, device performance, or short-duration exercise protocols, leaving a gap regarding the situated practices, improvisations, cognitive load, and social negotiations that occur during long-duration efforts. No qualitative study has explored how athletes construct autonomy, robustness, and safety when technologies meet environmental uncertainty. This study addresses this gap by examining how adults with T1D experience and use diabetes technologies in ultra-endurance contexts.

Methods

Study Design and Theoretical Orientation

We adopted a constructivist grounded theory (CGT) approach [10], focusing on how meanings and practices are co-constructed between participants and researchers. CGT privileges iterative cycles of data collection and analysis, constant comparison, memo-writing, and category development toward conceptual density rather than hypothesis testing.

Setting

Participants were adults living with type 1 diabetes (T1D) who train and compete in ultra-endurance disciplines (eg, trail/ultramarathon, triathlon, long-distance cycling). Interviews were conducted between July 4, 2024, and December 19, 2024, via secure videoconference.

Participants and Recruitment

Eligibility criteria were: age ≥18 years; T1D duration ≥1 year; and completion of ≥1 marathon or ultraendurance event within the past 5 years. We used purposive sampling (maximum variation on age, T1D duration, sport, and technology configuration), complemented by a theoretical orientation of additional contacts to elaborate emerging categories when feasible. Recruitment proceeded through national patient associations, endurance communities, and online groups dedicated to T1D and ultraendurance, as well as limited personal networks. Administrators were contacted when required by platform policies; no group names are reported to preserve community privacy.

XSL•FO
RenderX

In France, the relevant online communities comprise at most ~200 individuals. Of these, ~30 eligible candidates received the study information, ~15 indicated willingness to participate, and 13 were ultimately interviewed. No second recruitment wave was required because interviews rapidly yielded theoretical sufficiency for the developing categories. We note that, according to experienced patient partners and community estimates, the pool of eligible T1D ultraendurance athletes in France is likely <100, which positions our sample as substantively large relative to this niche population. Sex distribution in our interviews mirrored what is commonly observed in endurance racing fields; we address implications of this imbalance in the Limitations section.

To ensure full transparency, we note that exact counts of initial outreach messages were not systematically logged because contact often occurred through community posts and administrator relays. The figures reported above therefore reflect the best available estimates based on administrator analytics and message tracking. This limitation is also acknowledged in the Limitations section.

Data Collection

Semistructured interviews (47‑68 min) were conducted via secure videoconference between July 4, 2024, and December 19, 2024, audio-recorded with consent and transcribed verbatim. The guide covered athletic and diabetes trajectories; technology configurations and daily/race-day practices; preparation and contingency routines; perceived benefits and constraints; cognitive load; device failures and improvisation; and interactions with clinicians/peers. Field notes and analytic memos were maintained throughout.

Researcher Positionality and Reflexivity

The lead researcher is a clinician-athlete living with T1D (insider position). This facilitated access and rapport but required sustained reflexivity [11]. We used memo-writing, peer debriefing, and collaborative coding discussions to surface assumptions and interrogate interpretive moves. In line with our constructivist grounded theory stance, we did not compute inter-coder reliability coefficients; meaning was negotiated through iterative discussions and documented decision trails.

Data Analysis and Coding Information

We followed constructivist grounded theory (Charmaz) with iterative line-by-line initial coding, focused coding, and constant comparison within and across interviews, supported by NVivo (version 1.7.2, Lumivero, 2023). Two researchers (BK and LC) independently coded during the initial round, then met to compare and reconcile codes, refine focused codes, and develop the category structure via memo-supported discussions. Disagreements were resolved through discussion and return-to-data; no third-party adjudication was required. We judged theoretical sufficiency when additional interviewing no longer contributed substantively new insights to our developing core categories (variation, conditions, consequences) [12]. An outline/codebook (labels, brief definitions, representative extracts) and an illustrative coding tree are provided in Multimedia Appendix 1.

Patient and Public Involvement (PPI)

Two patient partners were engaged before data collection (refining the research question), provided interim feedback on preliminary interpretations, and reviewed the final manuscript for clarity and relevance. Both patient partners also participated as interviewees (role overlap), a decision made to ensure that advisory input remained grounded in first-person experience; they did not participate in coding or editorial decisions. We reflect on potential influences of this overlap in the reflexivity subsection and in the Limitations.

Ethical Considerations

The study received ethics approval from the South-East VI Research Ethics Committee (Ref: 24.00907.000253). All participants provided written informed consent. Participants did not receive any financial or material compensation for their participation in the study. Data were anonymized and identifying details were removed or altered in excerpts. Reporting follows COREQ (Consolidated Criteria for Reporting Qualitative Research) [13]; the completed checklist is uploaded as a Multimedia Appendix 1.

Results

Participant Characteristics and Data Collection

Thirteen adults living with type 1 diabetes (T1D) participated in the study. Participants had diverse ultra-endurance profiles (eg, trail and ultratrail running, marathon running, triathlon, long-distance cycling) and varying durations of T1D and technology configurations. An overview of participant characteristics is provided in Table 1.

XSL•FO
RenderX

Table . Characteristics of study participants.

ID	Sex	Age	Age at diagnosis	Duration of T1D[a]	Treatment	CGM[b]	HbA1c	Sports practiced	Max distance
S1	M	36	21	15	Pump (Omnipod)	Dexcom	7.8	Ultra-trail, ski, tennis-marathon	200 km
S2	M	26	15	11	MDI[c] (pens)	Libre 2	N/A[d]	Trail, triathlon, rugby-marathon	80 km
S3	M	53	38	15	Pump (Omnipod Dash)	Dexcom	N/A	Ultratrail-marathon	108 km
S4	M	48	25	23	Pump (Omnipod, Diabeloop planned)	Dexcom G6	7.4	Trail, marathon	59 km
S5	F	49	23	27	Pump	Dexcom	5.8	Trail, ultra cycling, marathon, triathlon	42 km
S6	M	41	19	22	Pump	Dexcom	N/A	Trail-marathon	50 km
S7	M	66	40	26	MDI (pens)	Libre	6.7	Trail-marathon	50 km
S8	M	37	13	24	Pump	Dexcom	N/A	Trail-marathon	60 km
S9	M	46	42	4	MDI (pens)	Libre	N/A	Trail-marathon	42 km
S10	M	38	10	28	Pump	Dexcom	6.4 - 6.5	Trail-marathon	100 km
S11	M	44	19	25	Pump	Dexcom	6 - 6.2	Trail-marathon	80 km
S12	M	26	23	3	MDI (pens)	Libre	N/A	Trail-marathon	42 km
S13	F	46	16	30	Pump	Dexcom	N/A	Trail-marathon	50 km

[a]T1D: type 1 diabetes.

[b]CGM: continous glucose monitor

[c]MDI: multiple daily injection.

[d]N/A: not available.

This table summarizes demographic and clinical characteristics of participants, including age, diabetes duration, treatment type, technology use, and athletic background. HbA_{1c} values were not systematically collected; they are reported only when spontaneously mentioned by participants during interviews, in accordance with the protocol approved by the ethics committee.

Overview of Analytic Categories

Analysis yielded five interrelated analytic categories describing how participants experienced and enacted diabetes technologies in ultra-endurance contexts. These categories capture shifts in glucose management practices, perceived benefits and constraints of technology use, the redistribution of cognitive and logistical work, strategies to cope with technical fragility, and the social and identity implications of visible technologies. Across categories, participants described a persistent tension between performance optimization and robustness-in-use under conditions of uncertainty.

Figure 1 provides a conceptual overview of how technologies, practice-based strategies, and social dynamics interlock to shape robustness-in-use. This visual summary complements the five analytic categories detailed below.

This sociotechnical figure illustrates how technologies (CGM, insulin pump, AID/HCL), practice-based strategies (continuous anticipation; safety and performance; mental work; redundancy and improvisation), and social dynamics (visibility, peer support, stigma negotiation) interlock to sustain robustness-in-use. Together, these processes extend beyond performance-focused optimization to support “good-enough” control under uncertainty during ultraendurance efforts.

XSL•FO
RenderX

Figure 1. Conceptual map of robustness-in-use in ultra-endurance with T1D. AID/HCL: automated insulin delivery/hybrid closed-loop; CGM: continuous glucose monitoring; T1D: type 1 diabetes.

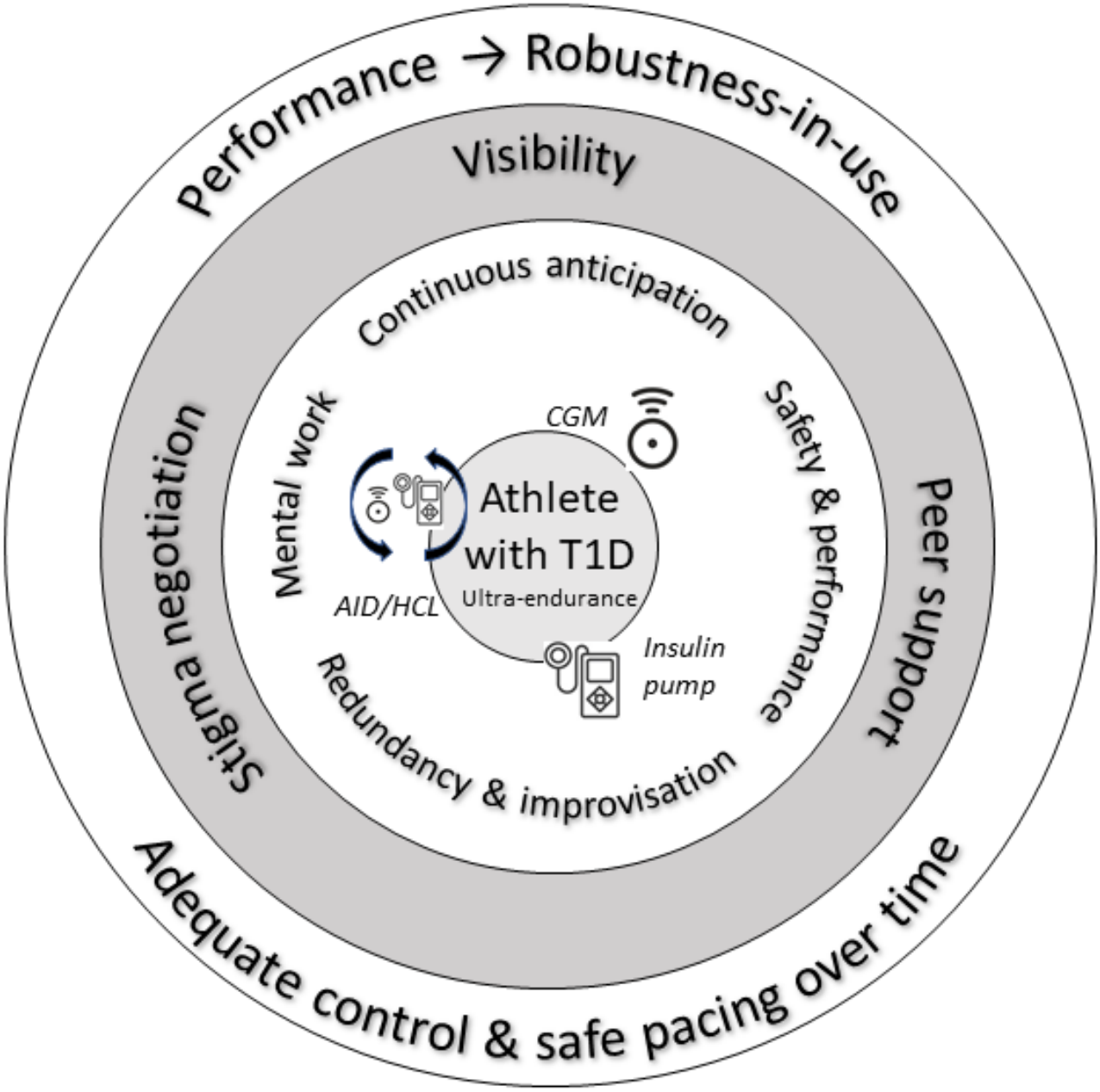

Analytic Category 1 — From Episodic Control to Continuous Anticipation

Participants described a shift from episodic control—sporadic glucose checks and retrospective corrections—to a mode of continuous anticipation made possible by real-time glucose technologies (CGM, pumps, sometimes AID/HCL). Glucose management became a prospective and dynamic process, in which athletes monitored trends, projected likely trajectories, and adjusted insulin and carbohydrate strategies in advance of anticipated physiological demands.

> *I see my glucose in real time and I can anticipate things, with the alerts… it's really convenient.* [S1]
>
> *Before, we were practically blind all the time.* [S5]

This anticipatory mode transformed glycemic regulation during ultraendurance efforts. Athletes described constantly looking ahead to the next climb, descent, or aid station, interpreting the CGM curve less as a momentary reading and more as a signal of what would happen minutes or hours later. The shift toward real-time visibility encouraged proactive adjustments—such as pre-emptive basal reductions, timed carbohydrate intake, or temporary targets—often activated before the body's physiological response became noticeable. For some, this capacity to plan ahead brought a sense of physiological stability and smoother pacing:

> *The pump was a small revelation for sports… before, I had lots of hypos as soon as I ran 12 or 15 km.* [S8]

However, the degree of anticipation varied with device configuration, terrain, and experience. While CGM alone enabled early warning and forward projection, athletes using

XSL•FO
RenderX

pumps—especially hybrid closed-loop systems—reported additional flexibility for fine-tuning insulin delivery. Yet several noted that algorithmic adjustments were "not designed" for prolonged or high-variability exercise, requiring manual overrides or personalized pre-settings. One participant noted:

> *Control-IQ added mental load… it didn't know how to manage sport at all* [S13.]

Anticipation was especially salient during long climbs, heat exposure, night running, or unpredictable sections of the course, where environmental stressors amplified the need to think ahead and compensate for potential CGM lag, sensor compression, or delayed carbohydrate absorption.

Experience shaped how athletes used data. Some alternated between data-driven projection and embodied sensations, relying on feelings of fatigue, heat, or digestive comfort to interpret CGM signals. Others prioritized bodily cues when technology felt out of sync. Across accounts, continuous anticipation reduced glycemic fluctuations and increased the sense of security but also redistributed cognitive work, requiring ongoing projection, parameter adjustment, and nutritional strategizing in parallel with the physical and navigational demands of ultraendurance effort.

Analytic Category 2 — Gains in Safety and Performance

Participants described diabetes technologies as enablers of performance, expanding what felt physiologically, medically, and symbolically possible. Devices provided a sense of security, reducing the fear of hypoglycemia, stabilizing glucose trends, and creating conditions that made prolonged or intense effort feel achievable. Several participants associated technology use with a renewed capacity to regulate their bodies during long events, thanks to tools that supported anticipation, stability, and fine-tuned adjustments in real time.

> *I can reduce my basal rate by 90% one hour before the race… it makes the experience much more comfortable, especially for digestion.* [S3]

For many, performance gains were entangled with a strengthened sense of mastery. Technologies enabled athletes to understand their glycemic patterns more deeply and to act with more precision. The access to continuous data, trend arrows, predictive insights, and adjustable insulin delivery contributed to what several described as a new "physiological literacy"—a clearer, more actionable grasp of how their bodies responded to exertion, terrain, food, and stress.

> *I understand the kinetics of my glucose levels now.* [S5]

Perceived clinical improvements reinforced this sense of control. Some participants noted reductions in HbA1c, fewer hypoglycemic episodes, or greater time in range, which they interpreted as signs that their training and technological adaptations were working together. These clinical gains were experienced not only as medical improvements but as psychological stabilizers, supporting confidence, self-efficacy, and long-term commitment to ultra-endurance sports.

> *With the closed-loop system, my HbA1c dropped to 6 or 6.2… I had never gone below 8% before.* [S11]

Technology also enabled participants to challenge external expectations, including medical advice and cultural assumptions about the capabilities of people with T1D. Several described their engagement in ultra-endurance sports as a redefinition of limits, where devices acted as both material and symbolic supports for pushing beyond narratives of fragility or constraint. The ability to complete marathons or ultra-trails was often framed as a form of self-affirmation, emancipation, or identity reconstruction.

> *They told me I wouldn't be able to run more than a half marathon… now there are no limits.* [S8]

For some, this expansion of possibility extended beyond sport and influenced their broader life trajectories and relationships. Technologies were woven into stories of belonging, growth, and unexpected positives related to diabetes itself.

> *Diabetes has brought me so many things… positive things, and lots of meaningful encounters.* [S1]

Across accounts, the perceived benefits of technology were profound but not unconditional. Performance, confidence, and empowerment depended on technical literacy, practical skills, and the capacity to adapt devices to real-world demands. These contingencies foreshadow the subsequent categories, which explore how gains in control coexist with cognitive, material, and emotional burdens, as well as the fragility of technologies under extreme conditions.

Analytic Category 3 — Redistributed Mental Work

Although diabetes technologies were largely described as empowering, participants also evoked a persistent—and sometimes intensified—mental load associated with their use. The constant availability of glucose data, while reassuring, created a form of hyper-vigilance that extended not only to the body but also to the devices themselves. Athletes emphasized that ultra-endurance sport already requires sustained attention to pace, terrain, nutrition, temperature, and safety; the addition of continuous glucose monitoring layered a parallel stream of cognitive work on top of these existing demands.

> *Diabetes is present all the time, all the time, all the time…* [S9]

> *Checking my glucose is the first thing I do when I wake up and the last thing before I sleep.* [S9]

This vigilance did not begin with the race itself but well before it. Preparing for endurance events required significant logistical work, including ensuring that pumps, sensors, adhesive patches, and supplies would last through long distances, unpredictable weather, sweat, or friction. Participants described lengthy and meticulous pre-race routines to secure, test, protect, and calibrate devices. Race-day preparation could be as cognitively taxing as the effort that followed:

> *You have to think about everything… it takes at least an hour and a half before the race.* [S9]

> *It's annoying to change the pump every three days, and to remember everything.* [S7]

XSL•FO
RenderX

Importantly, increased automation did not eliminate this burden. Hybrid closed-loop systems sometimes added new layers of complexity, particularly when algorithmic behavior clashed with the physiological realities of prolonged exertion. Athletes described having to monitor the device's decisions, override automated actions, or anticipate situations in which the algorithm would not adapt appropriately. Automation, in this sense, did not replace vigilance but redirected it toward the device itself.

> *Control IQ added mental load… it didn't know how to manage sport at all.* [S13]

Across accounts, mental load appeared less as an obstacle to performance than as a redistribution of cognitive labor. The work of glycemic management shifted from manual calculations to continuous monitoring, interpretation, and oversight of technological systems. This redistribution could feel worthwhile—because it enabled participation and enhanced safety—but it also demanded sustained attention, layered decision-making, and an acceptance that technology brought its own vulnerabilities. These dynamics connect directly to the next analytic category, which explores how athletes cope when devices fail or underperform during ultra-endurance efforts.

Analytic Category 4 — Keeping Things Working When They Break

Even as diabetes technologies were described as enabling participation and enhancing control, participants emphasized their material fragility during ultra-endurance events. Long distances, repeated mechanical shocks, perspiration, humidity, transitions in temperature, and the rhythm of movement all contributed to the vulnerability of devices. Pumps, pods, catheter sites, and CGM sensors were depicted as elements that could fail unexpectedly, disrupting both physical effort and the sense of safety cultivated during training.

> *My catheter got ripped off at the start.* [S7]

> *Sometimes the pod gets torn off… or the sensor readings don't match how I feel.* [S9]

These vulnerabilities necessitated a repertoire of redundancy strategies—practices that athletes developed over time to secure devices, anticipate weak points, and prepare for worst-case scenarios. Carrying multiple backup items (pods, catheters, adhesives, batteries, sensors), protecting sites with extra tape or foam, and choosing placement sites less exposed to friction were described as routine elements of pre-race logistics. Redundancy was not optional but part of the craft of running long distances with T1D.

> *I always take two spare pods.* [S4]

> *For big races, I bring two extra catheters just in case.* [S11]

When devices failed—whether through accidental detachment, adhesive degradation, CGM drift, or sensor-algorithm mismatch—participants relied on embodied knowledge accumulated over years of living and training with T1D. In the absence of reliable data, they turned to sensations of effort, bodily cues, fatigue patterns, digestive signals, and their understanding of how their glucose typically behaved under similar conditions. This adaptive turn away from technology did not represent a rejection of devices but an assertion of experiential expertise, a reservoir of tacit knowledge that filled technological gaps.

> *You have to go by feel… it takes years of experience and a certain confidence in yourself.* [S12]

Across accounts, these improvisational skills formed a kind of user-generated robustness. Robustness here did not reside in the devices themselves but in the capacity of athletes to adapt, improvise, and recover control when the technological layer faltered. This required not only planning and technical competence but also emotional steadiness in the face of uncertainty. Device failure could provoke frustration, anxiety, or fear—especially during long races where support is limited—but participants framed resilience as an ongoing process: a combination of preparation, calm evaluation, and embodied know-how.

This category highlights a central paradox of technology in ultra-endurance contexts: the more athletes rely on devices for stability, the more they must prepare for their instability. The work of "keeping things working" is therefore both material and cognitive, intertwining with the mental load described previously and foreshadowing the social and identity dimensions of technology use explored in the next category.

Analytic Category 5 — Making Diabetes Visible

Participants described the visibility of diabetes technologies—sensors, pumps, adhesive patches, and occasionally association-branded gear—as more than a technical feature: it was a social and symbolic act. Visibility served as a way to assert identity, challenge misconceptions, and embody a stance of openness rather than concealment. Several participants explained that displaying their devices during races or training felt like a form of ownership over the disease, a refusal of shame or silence.

> *I always said: I'm diabetic and I own it.* [S7]

Another described running as an opportunity to engage others:

> *It's about educating people while running.* [S1]

This dimension of visibility was often linked to a sense of responsibility and inspiration. Many participants saw themselves, willingly or not, as role models for younger people with T1D, for families navigating a new diagnosis, or for peers unsure of what was possible. Their presence in ultraendurance events, technologies in plain sight, was meant to broaden assumptions about what living with diabetes allows. For some, the act of being seen was explicitly oriented toward expanding collective horizons:

> *It's important to show that the disease doesn't stop us.* [S5]

> *We open up possibilities… it reassures parents.* [S8]

Visibility also generated connection. Encounters at races or during long training sessions often led to spontaneous conversations with other runners, spectators, or volunteers who recognized the devices. What might otherwise remain an isolating chronic condition became a channel for recognition, emotional support, and shared experience.

XSL•FO
RenderX

> *The gear gets noticed… it starts conversations.* [S1]
>
> *The community feels like family… it breaks the solitude.* [S2]

Although visibility offered connection and inspiration, participants also described moments of stigma negotiation. Showing one's devices could draw intrusive questions, misunderstandings, or paternalistic assumptions about fragility. Athletes navigated these encounters by reframing visibility as an act of resistance, turning potentially stigmatizing situations into opportunities for explanation, humor, or quiet defiance. At the same time, visibility created pathways for recognition and belonging: devices sparked spontaneous conversations, generated peer support, and signaled membership in a wider community of endurance athletes living with T1D. Through these interactions, technology became a social interface, contributing to the construction of a shared identity grounded in resilience, capability, and mutual encouragement. In this sense, visibility acted not only as a personal stance but also as a form of collective empowerment, expanding what is seen as possible for people with T1D and reshaping the social meanings of the condition in public athletic spaces.

Discussion

Principal Findings

Our findings show how diabetes technologies (CGM, pumps, AID/HCL) reconfigure self-management for adults with T1D engaged in ultra-endurance sports. Beyond glycemic gains, devices reshape the very conditions of participation: they enable anticipatory control (Category 1), unlock perceived performance and confidence (Category 2), redistribute mental work (Category 3), demand user-generated robustness in the face of breakdowns (Category 4), and mediate identity and solidarity (Category 5). Taken together, these categories illuminate the conditional nature of technological autonomy and the situated practices through which athletes keep care working under uncertainty.

Technological Autonomy Under Conditions

Consistent with recent syntheses on AID/HCL and real-world use, participants described devices as enablers that support tighter regulation and anticipatory adjustments, yet autonomy remained conditional on skills, planning, and context [1,4]. Athletes leveraged real-time data to project forward, pre-tune insulin/carbohydrate strategies, and secure pacing—when devices and environments cooperated. Autonomy, here, was not a default property of technology but the outcome of practice, contingent on technical literacy, race logistics, and environmental stressors. This conditionality reflects broader evidence that AID/HCL systems often struggle in prolonged aerobic exercise, requiring manual overrides, temporary targets, or carbohydrate adjustments to compensate for algorithmic mismatch [4,7].

Autonomy with a Cost: Redistributed Mental Work

Rather than eliminating workload, technology shifts cognitive demands. Continuous data streams required interpretation and oversight; athletes monitored battery life, adhesion, sensor accuracy, and supply chains, often before the race even started. These observations echo broader evidence that technologies can improve glycemic outcomes and contribute to diabetes distress when use becomes opaque or effortful [2,7]. Automation, especially in AID/HCL, did not erase vigilance; it redirected attention toward the device's decisions—aligning with the idea that digital systems change the kind of work patients do, not whether work is needed [1]. Similar patterns of redistributed cognitive work have been documented in qualitative studies of CGM and pump users, who describe increased vigilance around alarms, sensor accuracy, device maintenance, and algorithmic decisions, even when automation improves glycemic outcomes [2,5].

Device Fragility and User-Generated Robustness

Participants framed devices as vulnerable in ultra-settings—detachment, adhesive failure, CGM drift/lag, algorithm–physiology mismatch—calling for redundancy and improvisation. Real-world data and qualitative syntheses likewise note interruptions, inaccuracies, and trust challenges with CGM and pump systems, with glycemic consequences and anxiety when failures occur [7,8]. In our study, robustness did not primarily reside in devices but in users: carrying backups, securing sites, anticipating weak points, and reverting to embodied cues when data became unreliable. These observations are consistent with reports of reduced CGM accuracy under heat and movement and with qualitative accounts in ultra-endurance T1D showing how athletes build robustness through redundancy and on-the-spot improvisation [6,8,14].

Performance vs Robustness: A Pragmatic Lens, Not a Doctrine

The distinction between performance and robustness helps clarify how athletes actually keep self-management workable under ultra-endurance conditions. From a systems perspective, performance refers to optimization under expected conditions, whereas robustness denotes the ability to maintain function despite variability and perturbation—absorbing errors, adapting, and continuing to operate even when the environment deviates from design assumptions [15].

Our findings suggest that athletes did not pursue perfect optimization during ultra-endurance events. Instead, they enacted a form of robustness-in-use, aiming for "good-enough" stability across long durations, fluctuating terrain, and unpredictable bodily responses. This orientation was visible in anticipatory regulation (Category 1), where real-time visibility and projection enabled athletes to dampen destabilizing swings rather than chase precise targets. It also shaped the performance gains described in Category 2, where confidence and safety were valued because they could hold under shifting conditions, not because they achieved laboratory-level precision.

At the same time, Category 3 (Redistributed Mental Work) showed that robustness required continuous oversight of devices, interpretation of data, and readiness to intervene. Autonomous features of AID/HCL systems supported performance when conditions aligned with algorithmic expectations—but when they did not, robustness depended on the user's interpretive and corrective capacities. Category 4 (Keeping Things Working When They Break) further illustrated how robustness emerged

XSL•FO
RenderX

from redundancy, improvisation, and embodied knowledge: carrying spare equipment, securing adhesives, reverting to capillary checks or bodily cues when CGM data drifted or failed.

Finally, Category 5 (Making Diabetes Visible) highlighted a social dimension of robustness. Encounters with peers, recognition from other athletes, and the ability to reframe stigma supported the emotional steadiness required to troubleshoot under uncertainty. Robustness-in-use was therefore not only technical or cognitive but socially scaffolded, distributed across equipment, skills, and community interactions.

These patterns align with prior work on how digital self-tracking reshapes identity and social positioning and with qualitative accounts in ultra-endurance T1D emphasizing visibility as a form of connection and advocacy [7,14].Taken together, these findings suggest that evaluating diabetes technologies solely on performance metrics (HbA1c, TIR) overlooks a central dimension of real-world functioning: how well systems and users cope with perturbations. Robustness-in-use—anticipation, redundancy, adaptive plasticity, and the relational supports that sustain these practices—was central to autonomy during ultra-endurance, and may be equally relevant in everyday life.

Lay Expertise in Motion: Situated, Adaptive, Collective

Faced with algorithmic boundaries and environmental unpredictability, participants mobilized lay expertise—trial-and-error learning, embodied sensing, peer exchange—to fine-tune decisions moment by moment. This "practice-based knowledge" aligns with qualitative syntheses showing that users negotiate autonomy through on-the-ground adjustments rather than through stable, standardized routines [2,7]. It also echoes broader digital-health scholarship: self-tracking can foster autonomy through situated negotiation with uncertainty and personal values, not only through control [16]. Consistent with prior qualitative work, including studies in ultra-endurance T1D, such adaptive forms of expertise emerge when digital tools fall short and users must rely on embodied sensing and improvisation [2,14,16]

Practical Implications

- For clinicians: Translate technology features into exercise-specific routines: pre-emptive basal reductions or exercise modes 60 - 120 min before start; timed carbohydrates suited to intensity/duration; explicit redundancy checklists (spare pods/sets, tapes, batteries, backup CGM or capillary tests); scenario planning for adhesion/lag in heat, humidity, or night segments [1,4]. Acknowledge and validate the cognitive work patients do, offering strategies to simplify (eg, standard operating micro-routines, packing templates).
- For designers: Prioritize context-aware algorithms (detecting/proactively handling long aerobic efforts), user-controlled exercise targets that are simple to set and revert, and material durability (adhesives, housings, connectors) for sweat, friction, and temperature swings. Surface transparent device state (eg, adhesion/quality heuristics; lag alerts) to support trust and reduce oversight burden [1].
- For policy and services: Ensure equitable access to advanced systems and to education that explicitly includes endurance-exercise use cases. Support specialized clinics or peer-led workshops where experiential knowledge is shared safely and translated into practical guidance.

Transferability, Equity, and Diversity of Trajectories

Our sample reflects a male-skewed endurance field and a subset with access to advanced tech, which likely shapes emphasis on vigilance and robustness. Literature in sport/exercise consistently reports under-representation of women, with implications for how needs and design priorities are articulated; services and research agendas should proactively include women and other under-represented groups. Moreover, the "tech-privilege" in our cohort underscores the need to design for graduated complexity and to offer pathways that work for diverse resources, preferences, and bodies [2,7].

Rethinking Evaluation

These findings argue for evaluation frameworks that complement performance metrics (HbA_{1c}, TIR) with robustness-of-use indicators: stability of adhesion, transparency of device state, ease of switching to manual modes, workload to maintain function, preparedness for failure, and user confidence in troubleshooting. In ultra-endurance—an extreme but instructive case—robustness becomes visible; yet the same logic applies to daily life, where uncertainty, variability, and material constraints are the rule rather than the exception.

Strengths and Limitations

This study presents several strengths. It offers a deep, practice-based understanding of how adults with T1D use diabetes technologies in ultra-endurance contexts—a setting where physiological, environmental, and technological demands converge in distinctive ways. The use of constructivist grounded theory enabled a rich analysis of the processes through which participants enacted anticipation, robustness-in-use, and adaptive expertise. The study also benefited from the involvement of two patient partners, who contributed to the formulation of the research question, provided feedback on emerging interpretations, and reviewed the final manuscript, thereby strengthening the relevance and resonance of the findings.

However, the study also has limitations. The sample reflects the male-skewed composition of ultra-endurance sports, which may influence how certain forms of vigilance, risk management, or identity work are expressed. Access to advanced technologies was common in this cohort, which may limit transferability to athletes with fewer resources, different insurance coverage, or alternative device ecosystems. Recruitment occurred through associations and online communities, and although this strategy enabled access to a niche population, exact counts of initial reach could not be systematically recorded; the reported numbers reflect the best available estimates based on administrator analytics.

Two patient partners also participated as interviewees, creating a role overlap that may have influenced how certain experiences were emphasized. We mitigated this through reflexive memo-writing, peer debriefing, and collaborative coding

XSL•FO
RenderX

discussions, emphasizing interpretive transparency rather than inter-coder reliability metrics. Finally, interviews were conducted within a single national context, and ultra-endurance remains a highly selective activity; transferability to everyday physical activity or to broader T1D populations should therefore be considered with caution.

Despite these limitations, the study provides valuable insights into how technological autonomy is enacted under real-world constraints, and highlights dimensions of robustness, adaptation, and experiential knowledge that are relevant beyond ultra-endurance sport.

Conclusions

Diabetes technologies (CGM, pumps, AID/HCL) are enabling adults with T1D to engage in ultraendurance sports by shifting self-management from episodic control to continuous anticipation and by supporting perceived gains in safety and performance. In real-world practice, however, autonomy remains conditional: technologies redistribute mental work, exhibit material fragilities, and require user-generated robustness through anticipation, redundancy, improvisation, and socially scaffolded support. Rather than a quest for perfect optimization, athletes enact robustness-in-use—keeping care workable across long durations and fluctuating environments. These findings argue for clinical guidance that translates device features into exercise-specific routines, for design priorities that enhance context awareness and material durability, and for evaluation frameworks that complement HbA_{1c}/TIR with robustness-of-use indicators (eg, adhesion stability, transparency of device state, ease of switching modes, and troubleshooting confidence). What proves effective in ultraendurance may also inform everyday life with T1D, where uncertainty and variability are the norm.

Acknowledgments

The authors thank the participants for sharing their experiences and insights. We also acknowledge the support of the patient associations and running communities that facilitated recruitment. The authors used AI-assisted tools (Microsoft Copilot - GPT-5-chat) for minor language editing and grammar checks. All analytic decisions and interpretation of data were carried out manually by the research team.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: JCV, Patient Partners
Methodology: JCV, BK
Investigation: JCV
Formal analysis: JCV, LC
Data curation: JCV
Writing – original draft: JCV
Writing – review & editing: LC, BK, Patient Partners
Supervision: BK

Conflicts of Interest

None declared.

Multimedia Appendix 1
A radial coding tree. A central box labeled “Coding tree (T1D ultraendurance)” connects to five analytic categories: Continuous anticipation; Gains in safety & performance; Redistributed mental work; Keeping things working when they break; Making diabetes visible. Each category branches to focused codes, including Projection & anticipation, Stabilization & safety, Logistical & cognitive work, Redundancy & improvisation, Visibility & communication, Stigma & norm negotiation, and Peer & community support.
[PNG File, 96 KB - jmir_v28i1e86815_app1.png]

References

1. Kadiyala N, Hovorka R, Boughton CK. Closed-loop systems: recent advancements and lived experiences. Expert Rev Med Devices 2024 Oct;21(10):927-941. [doi: 10.1080/17434440.2024.2406901] [Medline: 39390689]

XSL•FO
RenderX

2. Speight J, Choudhary P, Wilmot EG, et al. Impact of glycaemic technologies on quality of life and related outcomes in adults with type 1 diabetes: A narrative review. Diabet Med 2023 Jan;40(1):e14944. [doi: 10.1111/dme.14944] [Medline: 36004676]
3. Da Silva J, Bosi E, Jendle J, et al. Real-world performance of the MiniMed. Diabetes Obes Metab 2021 Aug;23(8):1942-1949. [doi: 10.1111/dom.14424] [Medline: 33961340]
4. Perkins BA, Turner LV, Riddell MC. Applying technologies to simplify strategies for exercise in type 1 diabetes. Diabetologia 2024 Oct;67(10):2045-2058. [doi: 10.1007/s00125-024-06229-x] [Medline: 39145882]
5. Polonsky WH, Hessler D, Ruedy KJ, Beck RW, DIAMOND Study Group. The impact of continuous glucose monitoring on markers of quality of life in adults with type 1 diabetes: further findings from the DIAMOND randomized clinical trial. Diabetes Care 2017 Jun;40(6):736-741. [doi: 10.2337/dc17-0133] [Medline: 28389582]
6. Helleputte S, Podlogar T, Gonzalez J. Application potential of continuous glucose monitoring (CGM) in elite endurance athletes without diabetes: what do physiology and current evidence tell us? Perform Nutr 2025;1(1):13. [doi: 10.1186/s44410-025-00013-7]
7. Natale P, Chen S, Chow CK, et al. Patient experiences of continuous glucose monitoring and sensor-augmented insulin pump therapy for diabetes: a systematic review of qualitative studies. J Diabetes 2023 Dec;15(12):1048-1069. [doi: 10.1111/1753-0407.13454] [Medline: 37551735]
8. Markov AM, Krutilova P, Cedeno AE, McGill JB, McKee AM. Interruption of continuous glucose monitoring: frequency and adverse consequences. J Diabetes Sci Technol 2024 Sep;18(5):1096-1101. [doi: 10.1177/19322968231156572] [Medline: 36824046]
9. Poon ETC, Lau KW, Tsang JH, Zhang X, Wongpipit W. Comparative effects of continuous glucose monitoring-informed and traditional interval-based carbohydrate refueling protocols on endurance exercise responses: an exploratory study. J Int Soc Sports Nutr 2025 Dec;22(1):2561670. [doi: 10.1080/15502783.2025.2561670] [Medline: 40957601]
10. Constructing Grounded Theory.: SAGE Publications Ltd; 2026. URL: https://uk.sagepub.com/en-gb/eur/constructing-grounded-theory/book255601 [accessed 2026-01-31]
11. Dwyer SC, Buckle JL. The space between: on being an insider-outsider in qualitative research. Int J Qual Methods 2009 Mar;8(1):54-63. [doi: 10.1177/160940690900800105]
12. Saunders B, Sim J, Kingstone T, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907. [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
13. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
14. Vauthier JC, Choley L, Kabuth B. Chronic illness and extreme performance: type 1 diabetes in ultra-endurance. PLoS ONE 2025;20(12):e0338637. [doi: 10.1371/journal.pone.0338637] [Medline: 41417781]
15. Kitano H. Towards a theory of biological robustness. Mol Syst Biol 2007;3:137. [doi: 10.1038/msb4100179] [Medline: 17882156]
16. Sharon T. Self-tracking for health and the quantified self: re-articulating autonomy, solidarity, and authenticity in an age of personalized healthcare. Philos Technol 2017 Mar;30(1):93-121. [doi: 10.1007/s13347-016-0215-5]

Abbreviations

AID/HCL: automated insulin delivery / hybrid closed-loop
CGM: continuous glucose monitoring
CGT: constructivist grounded theory
COREQ: Consolidated Criteria for Reporting Qualitative Research
HbA1c: glycated hemoglobin
ID: participant identifier
MDI: multiple daily injections
NVivo: qualitative data analysis software
PPI: patient and public involvement
T1D: type 1 diabetes
TIR: Time in Range

XSL•FO
RenderX

Edited by A Stone; submitted 11.Nov.2025; peer-reviewed by AJ Umeano, A Famotire, E Omoregie, G Surianarayanan, T Jimoh; revised version received 19.Feb.2026; accepted 17.Mar.2026; published 08.May.2026.

Please cite as:
Vauthier JC, Choley L, Arduini D, Mas P, Kabuth B
Diabetes Technologies in Ultra-Endurance Type 1 Diabetes: Qualitative Study
J Med Internet Res 2026;28:e86815
URL: https://www.jmir.org/2026/1/e86815
doi:10.2196/86815

© Jean-Charles Vauthier, Lucie Choley, Delphine Arduini, Patrick Mas, Bernard Kabuth. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Cocreating Principles for Digital Health Equity: Cross-Sectional, Qualitative Study for Participatory Human-Centered Design in Catalonia

Jordi Piera-Jiménez[1,2], MBA, PhD; Núria Vilarasau Creus[3], MSc; Ada Maymó Costa[3], MA; Xabier Michelena[1,2,4], MD, PhD; Andrea Climent Fageda[2], MSc; Alèxia Farré[3], BSc; László Herczeg[3], MA; Lekshmy Parameswaran[3], MEng, MA; Gerard Carot-Sans[1,2], PhD; Luis Valle[2], MBA, PhD

[1]Digitalization for the Sustainability of the Healthcare System (DS3) research group, Barcelona, Spain

[2]Information Sysems Directorate, Catalan Health Service, Gran Via de les Corts Catalanes 587, Barcelona, Spain

[3]The Care Lab, Barcelona, Spain

[4]Rheumatology Research Group, Vall d'Hebron Research Institute, Barcelona, Spain

Corresponding Author:
Jordi Piera-Jiménez, MBA, PhD
Digitalization for the Sustainability of the Healthcare System (DS3) research group, Barcelona, Spain

Abstract

Background: Digital health technologies promise to democratize health care access yet often exacerbate existing inequalities when developed through traditional top-down approaches that prioritize technology implementation and exclude end users from design processes. The COVID-19 pandemic accelerated digital transformation while simultaneously exposing how technology can both bridge and widen gaps in health care access. Understanding how to systematically integrate equity considerations into digital health transformation across entire health systems has become increasingly urgent.

Objective: This study aims to cocreate actionable design principles for equitable digital health transformation through a large-scale participatory human-centered design (PHCD) process involving diverse stakeholders across Catalonia's health care ecosystem (northeast Spain), with the aim of establishing guidelines for information systems that support a person-centered, integrated, and longitudinal care delivery model.

Methods: We conducted a qualitative PHCD research study involving 265 participants representing diverse stakeholder groups: citizens and informal caregivers (n=106), health care professionals (n=83), health care managers and leaders (n=50), and experts representing various domains of digital health innovation (n=26). Through two sequential rounds of participation between June 2024 and April 2025, we used design thinking methodologies and cocreation tools in 24 sessions across Catalan geography and 7 topic-specific expert sessions. Data collection used innovative visual tools, including journey mapping, care model animations, future scenario storyboarding, and facilitated ideation techniques. Analysis followed an inductive-deductive approach combining affinity mapping, thematic synthesis, and participant validation to transform stakeholder proposals into actionable design principles.

Results: Participants identified critical barriers to digital health equity, including digital literacy gaps, fragmented information systems, a lack of user involvement in design, and insufficient consideration of vulnerable populations' needs. The cocreation process yielded 10 fundamental principles: (1) the person and their care circle at the center, (2) health for everyone, everywhere, (3) tools for more compassionate care, (4) a better professional experience, (5) an active role of the population, (6) personalized and proactive care, (7) systematic use of data for decision-making, (8) integrated quality data working for health, (9) an information system that builds trust, and (10) collaboration as a driver of innovation.

Conclusions: This study shows how large-scale, rigorously conducted PHCD can uncover and address equity barriers in health information systems. Beyond producing 10 actionable design principles, it highlights how engaging diverse stakeholders can turn complex inequities into practical guidance for equitable digital transformation. The resulting principles provide a framework for creating person-centered systems that are robust, inclusive, and accessible to all, while underscoring the need for enduring partnerships among public institutions, researchers, design experts, and communities as a foundation for sustainable and equitable digital health innovation.

(J Med Internet Res 2026;28:e84129) doi:10.2196/84129

KEYWORDS

digital health; equity; human-centred design; participatory design; digital health strategy; health information systems; co-creation; user experience

XSL•FO
RenderX

Introduction

The promise of digital health technologies to transform health care delivery has never been more apparent than in the wake of the COVID-19 pandemic. While the rapid adoption of telemedicine and mobile health apps demonstrated the potential of technology to maintain health care access during unprecedented disruption, it simultaneously revealed a troubling paradox: the very innovations designed to improve health care access often exacerbated existing inequities [1,2]. Vulnerable populations, including older adults, those with limited digital literacy, people with disabilities, and communities facing language barriers, frequently found themselves excluded from these digital advances, transforming potential bridges to care into barriers [3,4].

This paradox stems from a fundamental flaw in how health technologies are traditionally developed. Despite decades of evidence supporting user-centered approaches, health information systems continue to be designed primarily through top-down processes that prioritize technical requirements and organizational workflows over the needs, contexts, and capabilities of end users [5-7]. Health care professionals report spending up to 50% of their time on administrative tasks related to electronic health records (EHRs), reducing time available for direct patient care [8], while at the same time, patients struggle to navigate increasingly complex digital interfaces that assume levels of digital literacy many do not possess [9]. These challenges are particularly pronounced for vulnerable populations whose voices are rarely heard in technology design processes [10].

Participatory human-centered design (PHCD) offers a radically different approach to health technology development. Rooted in the principle that those who will use a system should be central to its design, PHCD engages users as partners throughout the development process, from initial problem definition through implementation and evaluation [11-14]. In health care contexts, this means involving patients, caregivers, health care professionals, and community members not merely as sources of requirements but as cocreators of solutions [15,16]. This orientation resonates with the broader global health literature describing human-centered design as a disciplined yet flexible practice that bridges design and implementation by iteratively engaging stakeholders and aligning innovation with human values and equity goals [17,18]. This approach is particularly crucial for addressing health equity, as it ensures that the voices of those most at risk of digital exclusion, often the same populations experiencing the greatest health disparities, are not just heard but actively shape the solutions developed [19,20]. Growing evidence demonstrates that health information systems developed through genuine participatory processes achieve higher adoption rates, better user satisfaction, and more equitable outcomes across diverse populations [21]. Yet, recent reviews highlight that while human-centered approaches can enhance equity for participants directly involved in cocreation, their broader, system-level impact often remains untested due to limited scaling and evaluation [22].

The concept of digital health equity extends far beyond mere access to technology. As defined by the World Health Organization (WHO), it encompasses the ability of all individuals to benefit from digital health solutions regardless of their socioeconomic status, geographic location, age, disability, or cultural background [23]. Operationalizing this vision requires addressing what recent collaborative frameworks identify as digital determinants of health (i.e., the complex, interconnected factors that shape whether digital health interventions produce equitable outcomes across populations) [24-27]. The adaptation of Richardson et al [26] of the National Institute of Minority Health and Health Disparities Research Framework [28] suggests that digital health equity operates across 4 different levels of influence (individual, interpersonal, community, and societal), each containing specific determinants that must be simultaneously addressed [26]. Similarly, recent work on equitable digital health design further proposes operational frameworks (eg, the Double Diamond and IDEAS models) that guide practitioners in deliberately embedding equity considerations throughout all design phases and in fostering structured collaboration with underserved groups [29].

Like many other countries worldwide, the COVID-19 crisis accelerated the digital transformation of health care in Catalonia (northeast Spain), with digital consultations increasing from 2% to more than 25% of all primary care visits between 2019 and 2021 [30,31]. However, this rapid digitalization revealed significant disparities across the region's 8 million inhabitants of the region. Rural areas struggled with inadequate internet infrastructure, older adult populations faced overwhelming digital interfaces, immigrant communities encountered language barriers in digital platforms, and health care professionals grappled with fragmented systems that hindered rather than helped care coordination [32]. These challenges acquired particular significance as Catalonia prepared to implement its ambitious Digital Health Strategy (2026 - 2031), a transformative initiative designed to support the future needs of health care delivery. Before embarking on this large-scale transformation, policymakers determined that equity principles must be embedded from the outset to avoid amplifying existing disparities.

Recognizing both the urgent need to address these digital equity challenges and the imperative to build equity into the upcoming Digital Health Strategy from its foundation, the Catalan Department of Health launched on a major participatory initiative to reimagine health information systems for the future. Rather than starting with predetermined technological solutions, the Department adopted a PHCD approach that would first identify gaps in the current health care delivery model and envision how future care should be organized, then determine how health information technologies could support this transformation. Following this approach, the idea was to reflect the growing recognition that creating truly equitable health information systems requires fundamentally rethinking not just what we build, but how we build it and who is involved in the building process [33].

This paper presents the results of the aforementioned large-scale PHCD process conducted across Catalonia, aimed at providing guidance to the design of a new and equitable health information

XSL•FO
RenderX

infrastructure that places the citizen at the center. The study addressed 2 main questions: (1) What are the lived experiences and challenges of diverse stakeholders navigating the Catalan health care system, particularly regarding barriers to equitable access and digital health engagement? (2) What fundamental principles and guidelines should inform the design of health information systems to ensure they are equitable, accessible, and responsive to the needs of all populations, especially those traditionally excluded from design processes?

Beyond these specific research questions, the study constitutes the first large-scale use case of user involvement in human-centered design for health care system transformation, providing valuable insights for replicating and scaling participatory design approaches across complex health systems.

Methods

Study Design and Conceptual Framework

We conducted a qualitative participatory design research study to develop a collective vision for Catalonia's future health care model and actionable design principles for digital health information systems. Our framework combined 2 complementary approaches. Human-centered design placed people at the core of the process, grounding design in their needs and contexts [34,35]. Participatory design ensured that diverse stakeholders were active cocreators rather than passive consultees [36]. This combined approach (ie, PHCD) was chosen because it directly addresses a fundamental cause of digital health inequity: traditional top-down development processes that systematically exclude the voices of those most affected by health disparities, particularly vulnerable populations who experience the greatest barriers to digital health access. PHCD engages stakeholders at all levels, from citizens and caregivers to frontline professionals, managers, and technical experts, capturing how barriers manifest and interact across individual, interpersonal, organizational, and systemic dimensions. This approach ensures a deep understanding of user needs before developing solutions, avoiding the premature solutioning that often leads to health technology failure [11,37].

The study design was structured around 2 sequential phases of participation, each addressing distinct but interconnected research questions. Both rounds used consistent participatory design methods, including journey mapping, system visualization, facilitated ideation, and collaborative prioritization, with Round 1 tools focusing on current care experiences and future care model visioning, and Round 2 tools adapted to translate identified needs into specific design principles for health information systems. Round 1 was deliberately designed to focus on health care delivery experiences and needs without introducing technology considerations, thereby avoiding the risk of biasing discussions toward predetermined technical solutions or constraining participant imagination to existing digital tools. This approach ensured that identified barriers and desired futures emerged authentically from lived experiences rather than being shaped by assumptions about technological feasibility or current system constraints. Only after establishing this grounded understanding of real citizen and professional needs did Round 2 shift focus to cocreating specific design principles for health information systems, ensuring that technology would serve human needs rather than drive the transformation agenda. The 2-phase approach enabled progressive deepening of understanding, moving from broad exploration of health care system experiences to focused cocreation of specific design principles for future health information systems. Finally, we deliberately adopted a multilevel scope encompassing technological design, organizational implementation, and broader system considerations. This approach was grounded on the fact that digital health equity failures occur across all these dimensions [26,38-40], and our PHCD methodology required understanding the full health care delivery context before narrowing to specific IT requirements.

The manuscript has been prepared according to the Standards for Reporting Qualitative Research (SRQR) guidelines [41].

Project Team Organization

The study was managed through a collaborative governance structure designed to ensure scientific rigor, meaningful stakeholder engagement, and operational excellence. A project steering board comprised representatives from the Catalan Health Service (JP-J, XM, LLV, and AC). The steering board was responsible for defining project objectives, monitoring progress against milestones, ensuring alignment with regional health transformation goals, making decisions about methodological adaptations based on emerging insights, and engaging with territorial health coordinators to recruit participants.

The operational implementation was carried out by a specialized design practice in multistakeholder engagement, human-centered design, and participatory design methodologies for enabling health and care system transformation. This team included supervisors (LP and LH) who provided methodological guidance and oversaw the overall design process to ensure consistency and quality. A dedicated tandem project management support structure coordinated logistics, including scheduling meetings across multiple territories, managing participant communications, providing technical support, and coordinating the overall design methodology and facilitation team (NVC). A team of 4 trained service-system designers who designed, facilitated, and synthesized the sessions, 2 lead facilitators who guided overall discussion flow and ensured all voices were heard (NVC and AMC), and 2 support facilitators who helped to document information during sessions, assisted participants with cocreation activities, and provided real-time analysis support (AF and AC). This team structure ensured comprehensive documentation while enabling smooth session flow and participant engagement.

The steering board and operational team established a coordinated governance structure with weekly alignment sessions to ensure methodological consistency, address implementation challenges, and maintain project objectives throughout the process.

Participants and Recruitment

Participants' recruitment and involvement adhered to the 4 core criteria outlined in the Catalan Department of Health's

XSL•FO
RenderX

Framework for Citizen Participation in Health, ensuring adequate methodological definition, maintaining scientific evidence and ethical standards, protecting vulnerable and minority populations, and preserving public health system sustainability [42].

Recruitment followed a purposive sampling strategy designed to maximize diversity and ensure representation of voices often excluded from technology design processes [43]. Sampling was operationalized through the health coordinators of each Catalan health region. These coordinators, who are managers within the Catalan Health Service, are responsible for planning and translating system-level health care policies into practice across the 10 health care regions of Catalonia. Once the study protocol was approved, and before fieldwork started, all 10 health coordinators attended a kick-off meeting in which the research team (ie, the authors) presented the study aims and procedures and provided detailed instructions for participant identification and recruitment. Comprehensive descriptions of the selection criteria and recruitment process are provided in the Multimedia Appendix 1. The health coordinators identified potential participants based on explicit diversity criteria, including gender and age distribution, organization types (public, private, and nonprofit sectors), health and social care service levels, and geographic representation across urban, semiurban, and rural territories. Health coordinators were explicitly asked to include, whenever possible and while meeting eligibility requirements, participants from potentially vulnerable groups such as older adults, immigrants, individuals with disabilities, and those with lower socioeconomic status. The project manager conducted weekly meetings with health coordinators for monitoring the recruitment process.

The study targeted four key stakeholder groups, each bringing essential perspectives to the cocreation process: (1) citizens and informal caregivers with diverse health care experiences (including expert patients and representatives of patient associations), who provide lived experience of navigating the health care system and represent the end user perspective; (2) health and social care professionals across different disciplines and care settings, who offer frontline insights into system functionality and daily operational challenges; (3) health and social care managers and leaders from various organizational levels, who contribute strategic and organizational perspectives on system implementation and sustainability; and (4) experts representing various domains of strategy, innovation, and technology, who provide technical expertise and broader sectoral knowledge to inform feasible and evidence-based solutions.

Territorial and Contextual Considerations

We recognized that territorial context significantly influences health care experiences and digital health needs [44,45]. Our sampling strategy deliberately included representation across urban, semiurban, and rural territories to ensure diverse contextual factors would be captured [3]. We also considered intersectional factors in our analysis framework, recognizing that age, gender, socioeconomic status, immigration status, and disability interact with geographic location to create unique barriers and opportunities [10]. We conducted an informed selection of the locations in consultation with the respective health regions to ensure comprehensive coverage of key regions and strategic areas across the Catalan territory.

Timeline and Methods

The study spanned from June 2024 to April 2025, beginning with preparation activities conducted from June to September 2024 (Figure 1). Data collection occurred through 2 sequential rounds of participation and analysis from October 2024 to March 2025, followed by consolidation of results and synthesis conducted from March to April 2025. Each round addressed distinct but interconnected research objectives, building progressively from understanding current health care experiences to cocreating specific design principles for future digital health information systems.

Figure 1. Project timeline, with activities (in blue) and outputs (in gray). R1: round 1; R2: round 2.

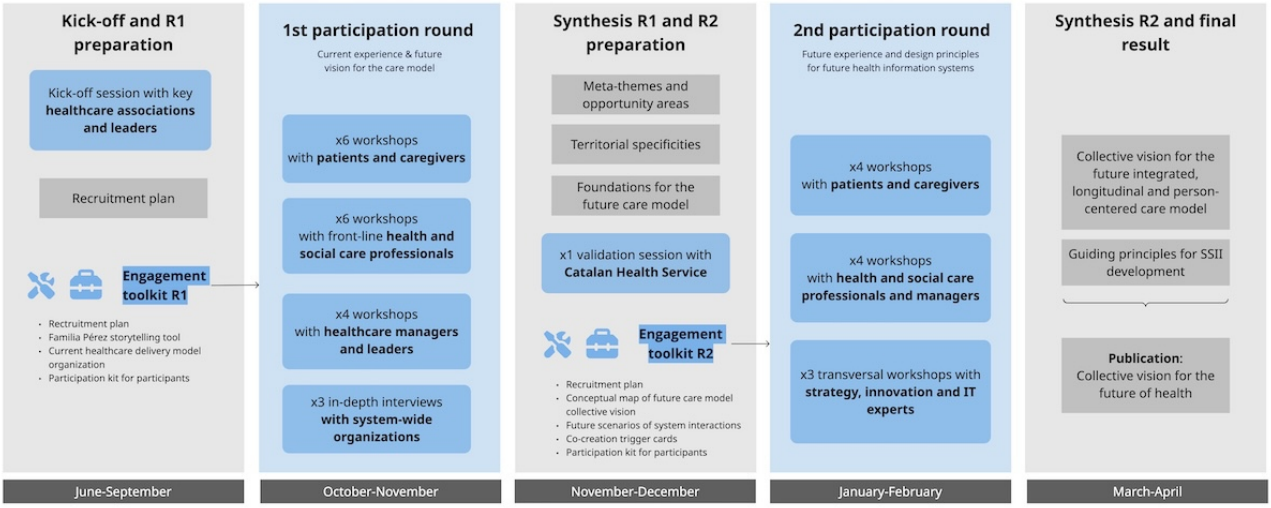

Round 1, conducted from October to November 2024, focused on collective reflection about the care model. This round aimed to understand current lived experiences with the health care system, identify what works well and what needs improvement, and envision desired futures. The central guiding questions included, “What do we understand by a person-centered, longitudinal, and integrated care model?” “What needs does this model address?” and “What must change in the current

XSL•FO
RenderX

system to achieve this vision?" Data were collected throughout 16 cocreation sessions conducted face-to-face and strategically selected to represent Catalonia's geographic and demographic diversity. Sessions were distributed across urban territories (Badalona, Barcelona, Girona, Lleida, and Tarragona), semiurban territories (Igualada, Manresa, Sant Cugat, Sant Feliu de Llobregat, Tortosa, and Vic), and rural territories (Amposta, Navàs, Palamós, Tremp, and Sort). Each 2-hour session included between 8 and 12 participants and followed a carefully structured format designed to maximize participation and insight generation. Additionally, 3 online one-hour group interviews with 2 participants each complemented the sessions across the territory, focusing on macrolevel perspectives and strategic insights into system-level challenges and opportunities. The organizations participating in these interviews were: the Agency for Quality and Assessment in Healthcare, the TICSalut and Social Foundation, and the Catalan Integrated Care Agency.

Round 2, conducted between January and February 2025, was devoted to cocreating specific design principles for digital health information systems. Building on insights from Round 1, this round transformed needs and barriers into actionable principles. The central question guiding this round was, "What design principles should guide the future development of health information systems to ensure they serve all citizens equitably?" This round started with a validation session with 7 experts from the Catalan Health Service, who reviewed Round 1 learnings and bridged into defining the focus for Round 2. It used 8 additional face-to-face sessions, maintaining geographic representation while focusing on the development of design principles. These sessions were distributed across urban territories (Barcelona), semiurban territories (Olot, Puigcerdà, Reus, Tàrrega, and Vilanova i la Geltrú), and rural territories (Mora d'Ebre, Solsona). Each session lasted 2 hours and included between 8 and 12 participants. Three topic-specific expert sessions brought together experts in specific domains identified from Round 1 learnings. Each session took 2 and a half hours and included 6 expert participants focusing on particular aspects of digital health transformation, such as strategy, innovation, and technology, providing a more macro perspective on the central question.

The locations of the 2 rounds across the Catalan geography are depicted in Figure 2.

XSL•FO
RenderX

Figure 2. Territorial distribution of the face-to-face participatory sessions across the 2 rounds.

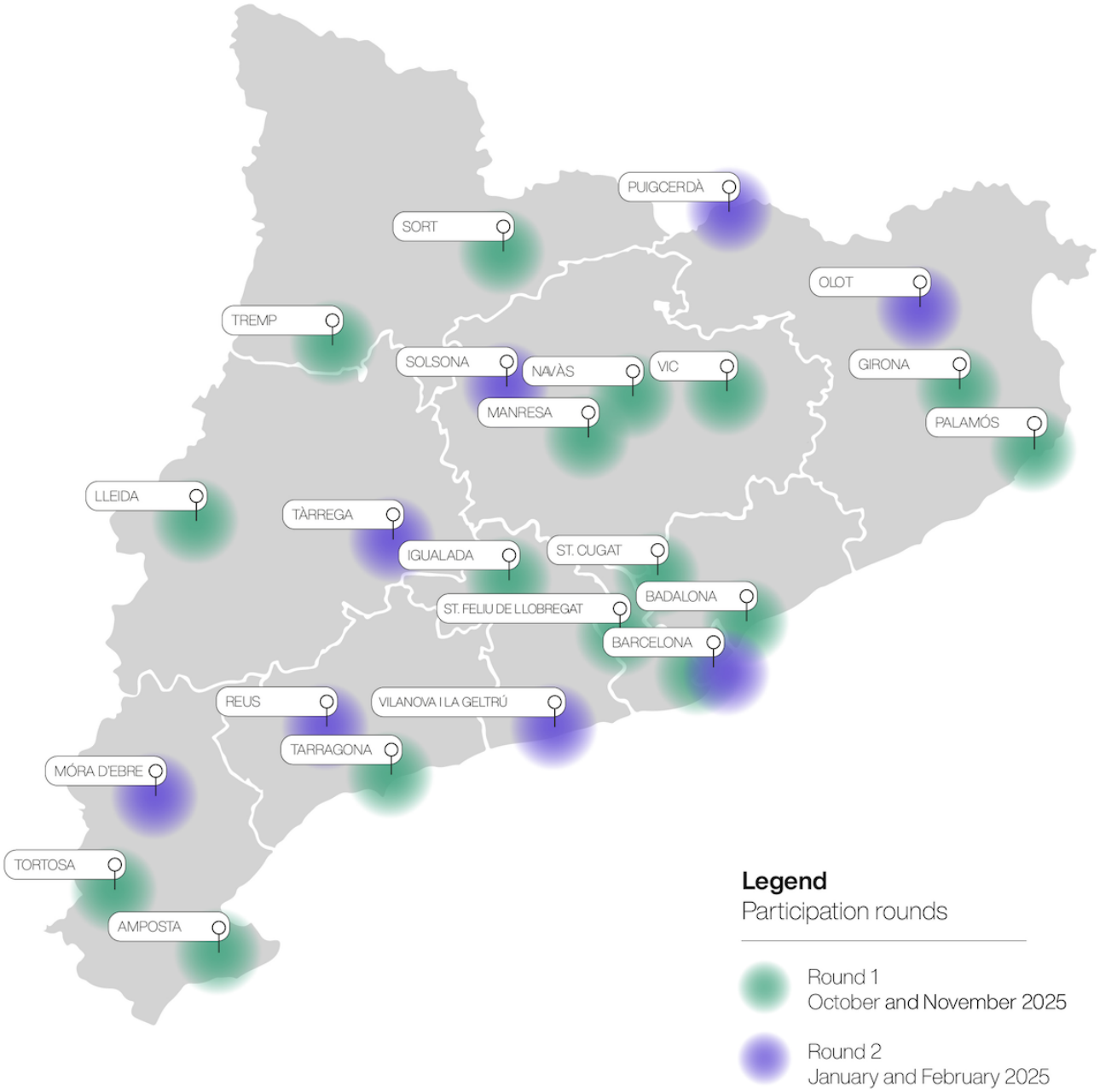

All sessions followed a consistent structure designed to create safe, inclusive spaces for participation. Sessions began with a 15-minute opening that established context, introduced facilitators, and established participation ground rules to ensure safe sharing of diverse perspectives. Core activities lasting 90‑105 minutes varied by round and session type but consistently used visual facilitation tools, including journey mapping, experience sharing, ideation techniques, and collaborative prioritization exercises. Sessions concluded with 10‑15 minutes for summarizing key insights, outlining next steps, and acknowledging participant contributions.

Sessions used innovative visual tools including journey mapping and storytelling to understand current health care experiences, scenario building for future visioning, and facilitated ideation techniques for developing design principles. Large-format posters, sticky notes, and collaborative boards enabled participants to contribute ideas both verbally and visually (reflecting individually and sharing collectively). Session logistics were carefully adapted to ensure inclusive participation across diverse contexts. Rural sessions addressed infrastructure challenges through flexible scheduling and accessible locations, while urban sessions accommodated participants' complex schedules while managing larger group dynamics. All venues provided step-free access, appropriate facilities for participants with disabilities, and professional translation services when needed. Transport support was provided where needed, particularly for participants from vulnerable populations or remote areas.

At the end of each session, a satisfaction questionnaire was provided to each participant to assess their participation experience and the perceived value provided by the workshop. The underpinning idea behind this feedback collection was to improve the forthcoming workshop dynamics in case any issue was identified (Multimedia Appendix 1).

XSL•FO
RenderX

Cocreation Tools

Central to the methodology was the development and use of innovative cocreation tools designed to make participation accessible and engaging for all stakeholders. A total of 4 tools were developed, 2 storytelling visual narratives were designed for the sessions with citizens, caregivers, and professionals. And 2 system maps for the sessions with health care managers and experts. The cocreation materials design drew from established journey mapping [46-48] and system visualization methodologies [49,50] in health care service design. The research team's approach was informed by their prior development of the WhocCares publication and the WeCare Toolkit (National Council of Social Service, Singapore and fuelfor) for caregiver support in Singapore [51], which demonstrated effective use of visual tools for engaging diverse stakeholders in the cocreation of health and social care service innovations. The full set of tools can be found in the Multimedia Appendix 1.

The "Pérez Family" journey mapping and storytelling tool consisted of a visual narrative depicting 8 health care interaction scenarios across the life course, from prenatal care through end-of-life support (Figure 3). This tool was designed with health care experts from the Catalan Health Service to ensure scenarios were realistic, relatable, and comprehensive, translating complex processes and multiple care pathways into accessible storylines to trigger exploration and discussion. The full poster of the "Pérez Family" tool is provided as supplementary material in Multimedia Appendix 1.

Figure 3. Translated sample of the "Pérez Family" journey mapping and storytelling tool used in face-to-face sessions with patients, caregivers, and professionals.

The system maps for professional and manager sessions included behind-the-scenes system processes and pain points (Figure 4).

XSL•FO
RenderX

Figure 4. Current health care delivery model organization used in face-to-face sessions with health care leaders and managers. IS3 (*Integrador de serveis de salut*): care process management; HC3 (*Història clínica compartida de catalunya*): shared care record*; eCAP:* primary care electronic health record; HIS: hospital information system; HS (*Historial social*): social care record; SITREM (*Sistema de tractament d'Emergències*): emergency treatment;

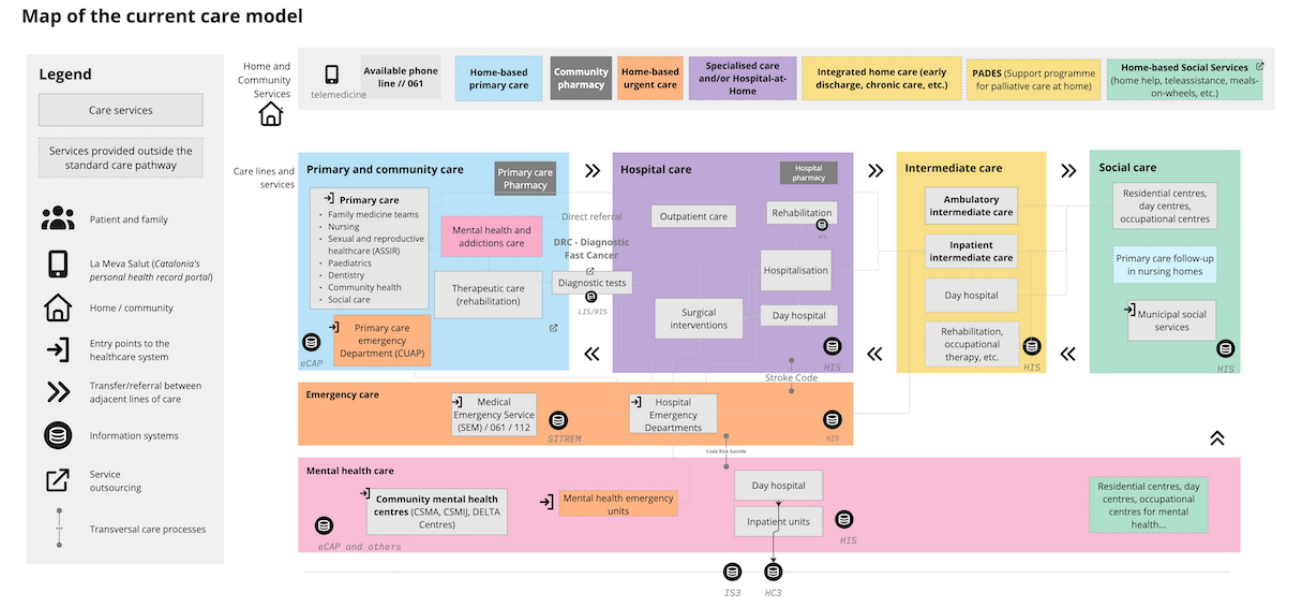

For Round 2, future scenarios were designed to portray a visual narrative of future interactions within 6 key themes emergent from Round 1 analysis: proximity care, empathic and quality interactions, collaborative and continuity care, personalized and proactive health prevention and promotion, active citizen participation in health, and holistic view of health. A conceptual map of the future care model vision resulting from Round 1 was another key visual used in the expert sessions.

Visual facilitation materials were carefully designed to support different communication styles and literacy levels. A facilitation toolkit was designed to enable issues to be seen at different levels of zoom or resolution, from lived experience to system overview, through technical backend, and overall delivery model. Large-format posters enabled collective visioning, while color-coded notes provided a simple yet effective method for categorizing inputs (Figure 5). Participants were given a participation kit with materials printed individually for better readability (Multimedia Appendix 1). All materials were tested for accessibility and refined based on feedback.

XSL•FO
RenderX

Figure 5. Example of a large-format poster for collective visioning with participants adding color-coded notes for item categorization.

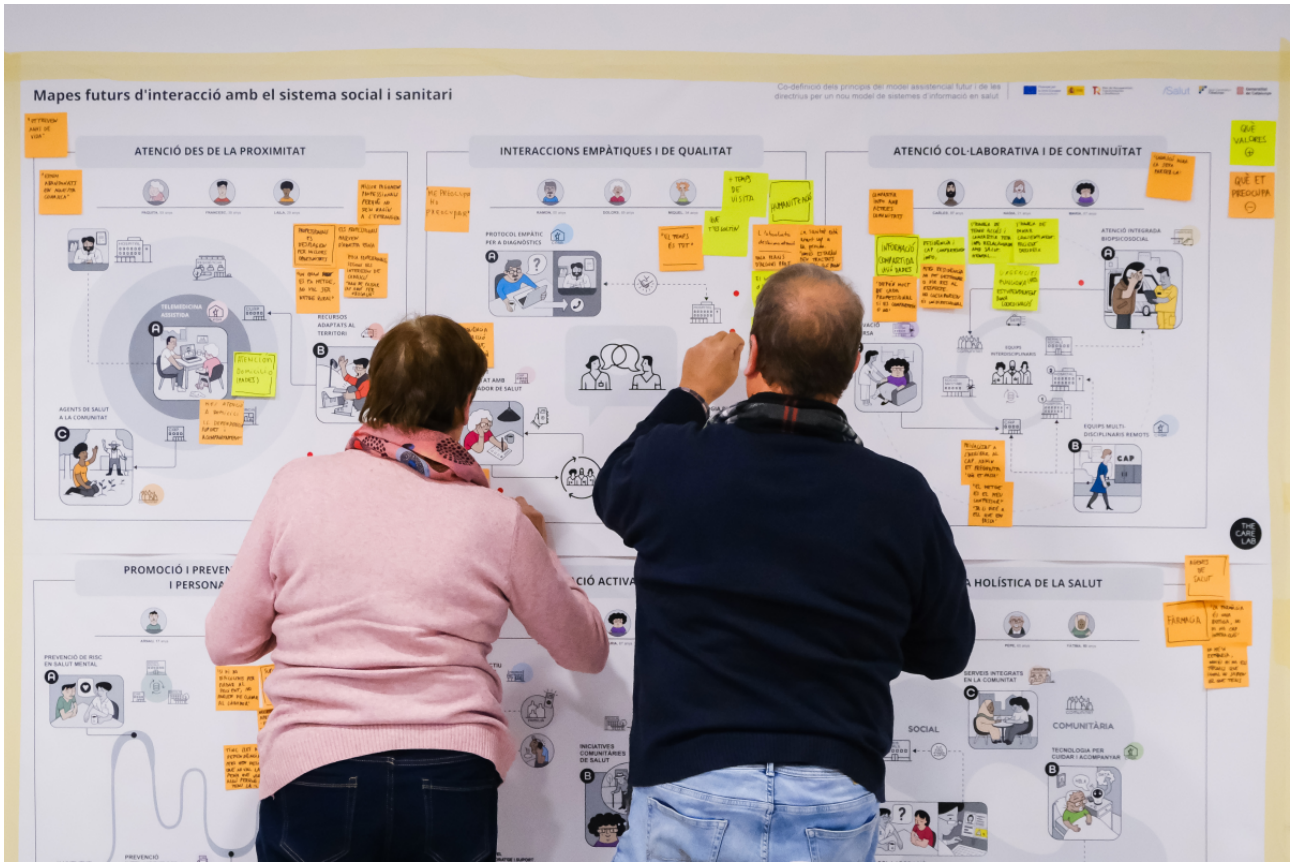

Data Collection

Data collection was designed to capture both the content and context of cocreation activities through multiple complementary approaches. The facilitation team used a structured approach to data capture, with each member having specific responsibilities.

The lead facilitator maintained a neutral stance while using follow-up questions to uncover the deeper reasoning and meaning behind participants' statements. The support facilitator observed and documented group dynamics, body language, nonverbal cues, power relations, and patterns of interaction. Dedicated notetakers captured verbatim statements, reasoning provided by participants, and contextual explanations. One notetaker focused exclusively on capturing verbatim quotes to preserve original language. Observer roles photographed and video-recorded key moments, interactions (such as voting, debating, and consensus-building), and visual outputs to document the atmosphere and participation patterns.

Multiple types of data were systematically collected from each session (1) physical artifacts including color-coded sticky notes, large-format posters, and flipcharts with participant handwriting and drawings; (2) documented narratives of participants' lived experiences and system interactions; (3) structured observational notes on group dynamics, including dominant voices, facilitation influence, and the sequence in which ideas emerged and built upon each other; (4) photographic and video documentation; and (5) verbatim capture of participants' oral explanations and reasoning. Table 1 summarizes the data collection methods, mapped to objectives and methods, according to project phases.

XSL•FO
RenderX

Table . Overview of methodological approaches and data collection tools by study phase.

Study phase	Timeline	Research objectives	Methods	Data collection tooling
Kick-off and Round 1 (R1) preparation	June-September 2024	• Develop cocreation tools. • Establish governance structure. • Recruit participants. • Design session protocols.	• Co-design of visual tools with health experts. • Tool testing and refinement.	• Recruitment tracking forms. • Tool design documentation. • Pilot session feedback.
Round 1: current experience and future vision for the care model	October-November 2024	• Understand current health care experiences. • Identify barriers and facilitators. • Envision desired future care model.	• Journey mapping. • System visualization. • Experience sharing. • Future visioning. • Collaborative discussion.	• "Pérez Family" journey tool. • Current care model system maps. • Color-coded sticky notes. • Large-format posters. • Verbatim capture. • Observational notes. • Photography and video. • Postsession debriefs. • Satisfaction questionnaires.
Synthesis R1 and Round 2 (R2) preparation	November-December 2024	• Validate Round 1 findings. • Define focus areas for Round 2.	• Expert validation session. • Collaborative synthesis.	• Summary presentations. • Validation discussion notes.
Round 2: future experience and design principles for future health information systems	January-February 2025	• Transform needs into actionable principles. • Define guidelines for health information systems. • Validate and refine principles.	• Scenario building. • Facilitated ideation. • Collaborative prioritization. • Principle refinement.	• Future scenario storyboards. • Future care model conceptual maps. • Color-coded sticky notes. • Large-format posters. • Verbatim capture. • Observational notes. • Photography and video. • Postsession debriefs. • Satisfaction questionnaires.
Synthesis R2 and final result	March-April 2025	• Synthesize findings across rounds. • Finalize design principles. • Prepare dissemination materials.	• Design synthesis. • Thematic analysis. • Principle validation.	• Digital collaboration boards (Miro). • Synthesis workshop outputs. • Final principle documentation. • Internal synthesis sessions.

Immediately following each session, the facilitation team conducted structured debriefs that captured observations on group dynamics (including identification of dominant participants, power relations, facilitation effects, and the order in which contributions emerged), surprising insights that emerged during sessions, patterns of agreement or tension, and participants' clarifications of their contributions. These immediate debriefs served to contextualize the visual outputs while observations were still fresh, ensuring that team observations about the process of cocreation informed the interpretation of the products (visual results). The debriefs used a standardized template covering key themes that emerged, surprising insights or unexpected perspectives, group dynamics observations (including any imbalances in participation or influence), emerging patterns across different voices, and new questions arising for subsequent sessions.

Data collection was extended at least until saturation, assessed through continuous monitoring of topics. As sessions progressed, we maintained a systematic list of topics and themes being discussed across workshops. The iterative download and synthesis process allowed clustering ideas and comments around

XSL•FO
RenderX

different themes for subsequent identification of both recurrent and unexplored topics. Saturation was reached when no new ideas or experiences were being captured.

Data Analysis

We used a systematic multistage analysis process combining inductive and deductive approaches, consistent with best practices in qualitative research [52]. Analysis began immediately after each session with the structured team debriefs (steering board and specialized design practice). These direct downloads enabled key takeaways to be captured whilst fresh and vivid, allowing the methodology to be tuned between sessions to optimize the approach and outcomes and boost learnings and insights. Our front-stage and backstage teams processed detailed data in agile loops as fieldwork progressed, which avoided a build-up of data and ensured learnings were taken quickly to deepen insights.

All physical materials, including colored notes, posters, and flipchart outputs, were photographed and systematically transcribed into digital collaboration boards using Miro software (Figure 6). This digitization process maintained original participant language and verbatim, preserved visual relationships between insights and ideas, and included contextual notes about discussion dynamics and nonverbal communications.

Figure 6. Example of digital boards (Miro software) used to analyze workshop outputs through axial coding. Contributions are mentioned in the Catalan language.

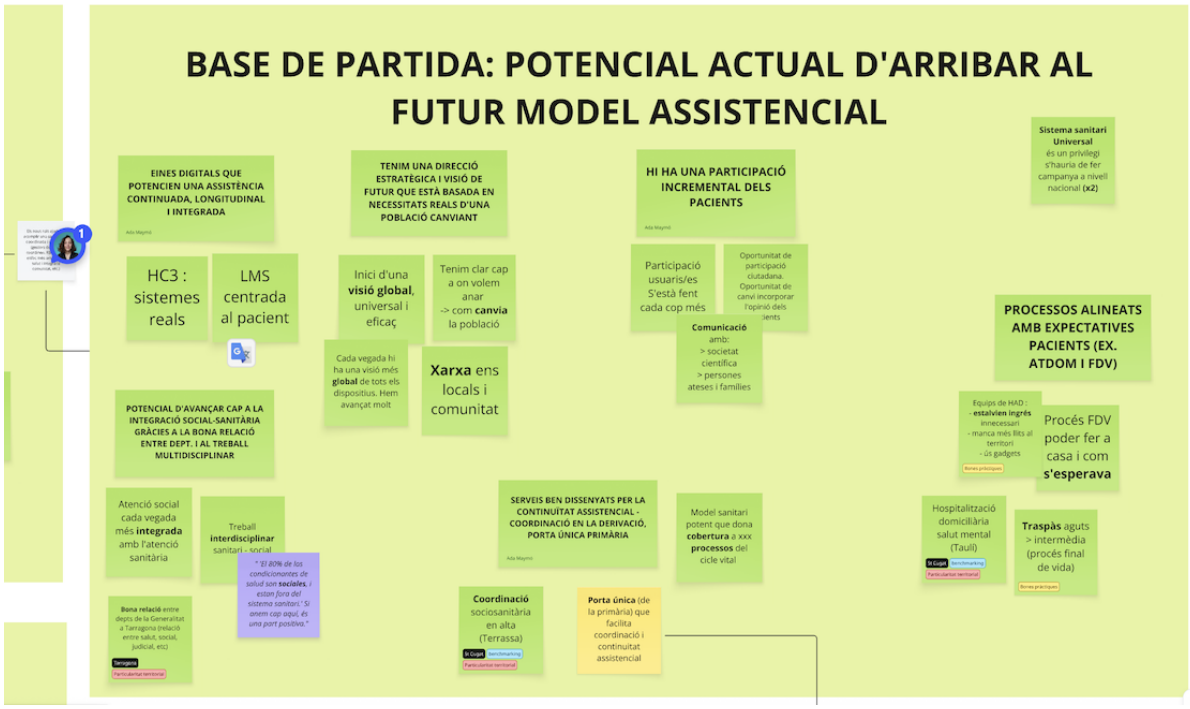

Multiple units of analysis were used. Visual outputs (colored notes and flipchart content) captured discrete participant insights and ideas. One researcher specifically documented verbatim quotes to preserve participant language. Debrief notes synthesized contextual observations about group dynamics and emergent patterns. These complementary sources enabled triangulation across data types.

Each data unit was tagged with structured metadata: geographic territory (urban, semiurban, and rural), stakeholder type (citizen, professional, manager, and expert), and workshop identifier. A systematic topic list tracked emerging themes across sessions to monitor saturation.

Analysis combined affinity mapping [53] with thematic synthesis [52]. Multiple analytical methods were integrated to ensure rigorous interpretation [54]. Initial affinity mapping clustered related insights on digital boards, grouping similar participant contributions whilst preserving their source metadata. This visual organization revealed natural patterns and relationships within the data [55]. Three researchers with complementary analytical perspectives—lived experience (citizen perspectives), operational processes (professional practice), and system architecture (organizational design)—then conducted thematic analysis. Open coding identified discrete concepts, followed by axial coding to explore relationships and create preliminary categories. Iterative team analysis sessions consolidated categories into broader themes, identifying priority challenges and opportunities.

Analytical consistency was ensured through regular synthesis sessions amongst the coding team and validation discussions with project steering board members (health information technology and clinical experts). These sessions cross-checked interpretations, identified analytical blind spots, and confirmed no significant gaps in knowledge capture.

From the Round 1 themes, we extracted foundational elements through design synthesis methods [56]. These foundations were translated into draft design principles through internal team workshops. Round 2 sessions then validated, refined, and finalized these principles through participant feedback.

XSL•FO
RenderX

Ethical Considerations

All participants provided written informed consent, and the study received approval from the Ethics Committee of Hospital Universitari de Bellvitge (reference PR123/24). Participants were informed that their participation was voluntary and could be withdrawn at any time without consequence. Data confidentiality was maintained through anonymization procedures, and all data were stored securely with access limited to research team members. Participants provided informed consent for the recording of verbatim statements and photographs during the sessions; no images were captured for participants who rejected consent. To ensure inclusive participation, transport and professional translation services were offered when necessary. In line with the PHCD approach, we organized give-back online sessions to share the findings with participants and inform them of subsequent steps, ensuring the research delivered value to all involved.

Results

Overview

A total of 265 participants across the Catalan health care ecosystem participated in a 2-round PHCD process that explored current health care delivery challenges and cocreated principles for equitable digital health transformation. The findings reveal both systemic barriers and collaborative solutions emerging from systematic engagement with diverse stakeholders across Catalonia's health care territories.

Participant Characteristics and Territorial Distribution

Participant demographics and territorial distribution are presented in Table 2, reflecting robust representation across stakeholder groups and geographic contexts throughout the study period. The sample included citizens and informal caregivers (n=106), health care professionals (n=83), health care managers and leaders (n=50), and experts in various domains of digital health innovation (n=26). Territorial representation spanned urban centers, rural areas, and intermediate zones, ensuring diverse infrastructure and resource contexts informed the findings.

XSL•FO
RenderX

Table . Demographic characteristics of participants.

Participant characteristics	Values, n (%)
Patients and caregivers (n=106)	
Gender	
Men	39 (36.8)
Women	67 (63.2)
Age (years; missing=9; n=97)	
18 - 35	2 (2.1)
36 - 50	16 (16.5)
51 - 65	38 (39.2)
>65	41 (42.3)
Profile (n=106)	
Patient	68 (64.2)
Caregiver	38 (35.8)
Territorial context	
Urban	20 (18.9)
Semiurban	49 (46.2)
Rural	37 (34.9)
Health and social professionals, experts, and leaders (n=159)	
Gender	
Male	57 (35.8)
Female	102 (64.2)
Age (years; n=159)	
18 - 35	7 (4.4)
36 - 50	100 (62.9)
51 - 65	52 (32.7)
>65	0 (0.0)
Profile (n=159)	
Family physician	4 (2.5)
Hospital and intermediate care physician	16 (10.1)
Primary care nurse	11 (6.9)
Hospital and intermediate care nurse	9 (5.7)
Psychology	4 (2.5)
Social worker	7 (4.4)
Health administration	7 (4.4)
New roles (community well-being, physiotherapist, and nutritionist)	7 (4.4)
Others (emergency service, pharmacy and care homes)	13 (8.2)
Health care managers and leaders	50 (31.4)
Experts in digital health innovation	26 (16.4)
Territorial context	
Urban	91 (57.2)
Semiurban	48 (30.2)
Rural	20 (12.6)

XSL•FO
RenderX

Through journey mapping and storytelling activities, Round 1 participants articulated 13 priority barriers to digital health equity that span technological, organizational, and sociocultural dimensions. These barriers emerged consistently across urban, semiurban, and rural contexts, though with varying emphases reflecting territorial specificities (see Figure 7 for territorial variations).

Figure 7. Summary of barriers to digital health equity identified during Round 1, contextualized per geographic area.

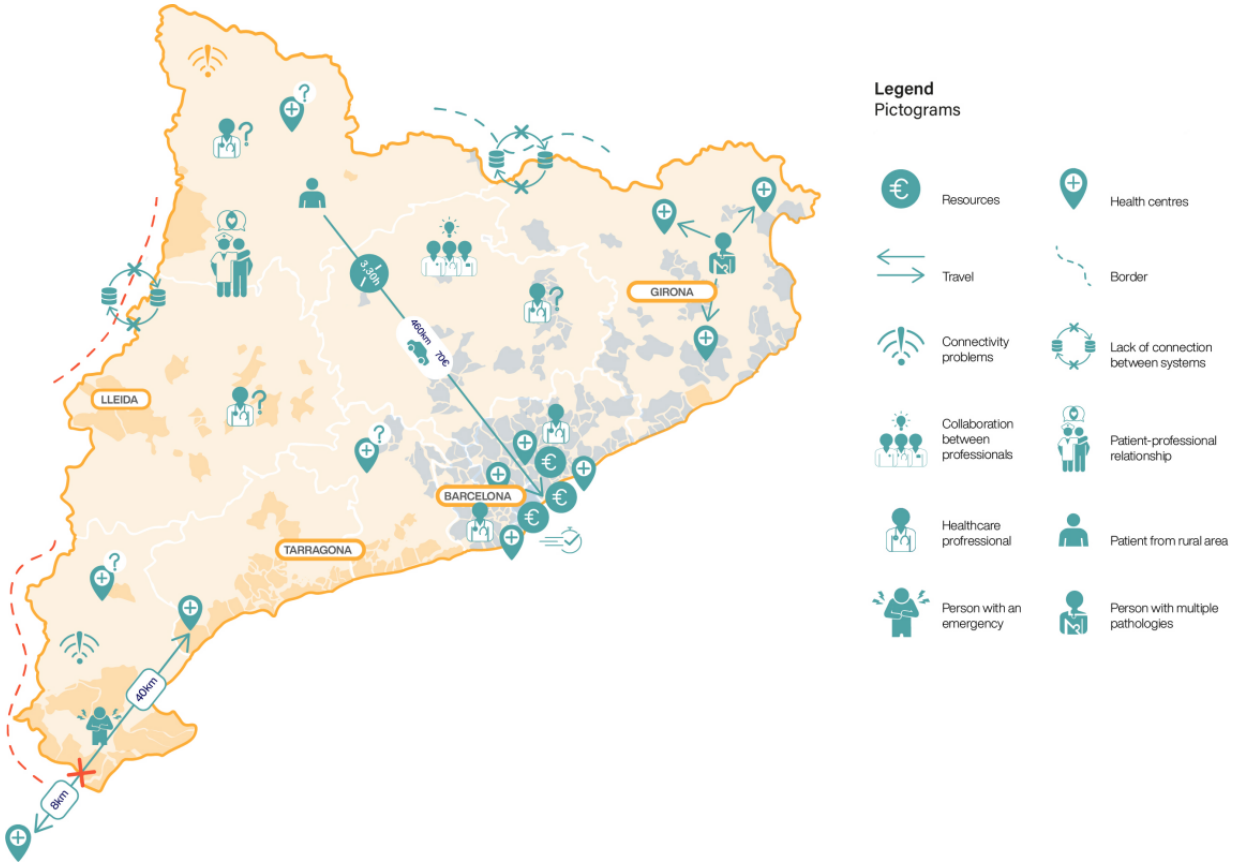

While these barriers emerged across all territories, their manifestation and intensity varied by geographic context. Rural areas emphasized connectivity and resource access challenges; urban contexts highlighted diversity and complexity issues; border regions contributed unique perspectives on administrative complexity and cross-jurisdictional care. Table 3 presents the 13 priority barriers with illustrative participant quotes demonstrating lived experiences across these diverse contexts. Detailed descriptions of all 13 barriers with extended context and additional participant quotes are available in Multimedia Appendix 2.

XSL•FO
RenderX

Table . Barriers to digital health equity identified in round 1, with illustrative participant quotes.

Barrier[a]	Description	Participant quote[b]
Fragmented health and social care integration	Services function in a fragmented manner due to hyperspecialization, generating multiple referrals that break care continuity. Primary care, under pressure, cannot assume its connecting role.	"Everything to do with the social system is a black box. I don't know what happens there. And we should know! 80% of health determinants are social." [Health care manager]
Insufficient emotional well-being support	Growing need for emotional support with a preventive, holistic, and community approach. Current system designed primarily for emergencies and physical pathologies.	"Community networks and neighbors are great prescribers of health and could be protective and informative elements." [Health care manager]
New citizen expectations are creating system pressure	Citizens have growing expectations about immediacy, flexibility, and personalization influenced by digital society, generating pressure on a system lacking resources to respond.	"We need to move towards a one-stop-shop system. It will make life easier for both the user and us." [Health care manager]
Lack of health education and prevention	Many people lack knowledge about self-care, prevention, and appropriate use of health services, leading to system misuse (eg, A&E saturation for nonurgent cases).	"Health literacy is fundamental. If people don't understand their health, they can't manage it properly." [Health care professional]
Social, language, and digital barriers	Digital gaps, medical terminology, and language barriers prevent many from accessing and navigating the system, especially older adults and migrant populations. Overall, 48% of the foreign population is at risk of poverty or social exclusion.	"Many times I don't understand the doctors' calls. If they give me a biopsy result I'm none the wiser, or I have to say, please, tell me in words I can understand. I prefer to go to the place rather than receive a call." [Service user]
Insufficient resources for aging and complexity	Progressive aging brings increased care needs for people with frailty and complexity. The current system lacks sufficient human and material resources for continuous, specialized, often home-based care.	"Resources should be moved towards the patient, rather than moving the patient towards resources. For example, prioritizing home care over residential facilities." [Health care manager]
Waiting lists with human consequences	Long waiting lists carry health risks and a strong emotional impact. Overall, 64% of patients wait >5 days to access a family physician; 150 days average for surgical intervention. Many are forced to turn to private health care, generating access inequalities.	"I'm an electrician, I'm self-employed, and they were giving me a physio appointment in 8 months. I can't wait, what will I eat? If you don't have private healthcare, you're done for." [Service user]
Resource planning based on quantitative criteria	Planning and resource allocation based on quantitative criteria (number of appointments, procedures) rather than health outcome indicators, leading to inadequate incentives.	"As a professional, you make the diagnosis, but when it comes to implementing truly, it's not possible because a specialist is missing. Something behind there should be a map of specialists that lets us zoom in and see, for example, the difference between supply and demand in areas like allergies and dermatology." [Health care professional]
Primary care unable to assume connecting role	Primary care should coordinate and ensure continuity, but pressure, lack of resources, and system fragmentation prevent it from exercising this role, generating breaks in care continuity.	"Primary Care must be the real entry and staying point for care. It shouldn't be a referral center. We need to do good work and strengthen Primary Care more." [Health care professional]
Inadequate protocols for chronicity and mental health	System excels at emergencies but lacks reinforcement in continuous preventive care and adequate support for daily management of chronic diseases. Overall, 37.8% of the adult population has chronic illness; 9.84% increase in chronic mental health patients since 2017.	"Mental health is left to God, more resources should be allocated to it. [...] Pure mental pathology is 'the great forgotten.' It works well in critical moments and acute hospitalization, but it doesn't have enough resources for chronic, continuous care within the community." [Service user and Health care professional]
High professional turnover	High turnover due to work culture changes and stressful conditions significantly affects care quality and continuity. Rotation prevents creating trust bonds essential for effective care, particularly impacting older patients and those with mental health needs.	"My psychologist has changed every other time. Each time you go you have to expose yourself again and explain what's happening to you." — Service user"I've even thought about leaving medicine and changing jobs. I work loads of hours, and you have no help from anywhere. I'm taking diazepam for anxiety." [Health care professional]

XSL•FO
RenderX

Barrier[a]	Description	Participant quote[b]
Administrative burden compromising care	Health care professionals dedicate significant time to administrative and bureaucratic tasks, reducing time for direct patient care, compromising care quality, and contributing to staff burnout.	”There should be an automatic process that allows you to spend more time with the patient. Here some technology like AI or a different way of entering information could play an important role.” [Health care manager]
Multiple noninteroperable information systems	Multiple noninteroperable systems hinder clinical information access and sharing between professionals from different care settings, generating inefficiencies, duplications, risk of errors, and coordination difficulties.	”Information should be accessible to all professionals because the information belongs to the patient, not to the professionals.” [Health care professional]

[a]Barriers synthesized from 189 participants across 19 round 1 sessions (16 territorial cocreation sessions +3 expert interviews).

[b]Quotes translated from Catalan/Spanish.

Cocreated Design Principles for Equitable Health Information Systems

Round 2 involved 8 territorial cocreation sessions, one validation session with Catalan Health Service experts, and 3 topic-specific sessions. Building on barriers identified in Round 1, participants cocreated 10 fundamental principles for equitable health information systems that address technological, organizational, and social dimensions of digital health equity (Figure 8).

Figure 8. Cocreated fundamental principles for equitable digital health information systems.

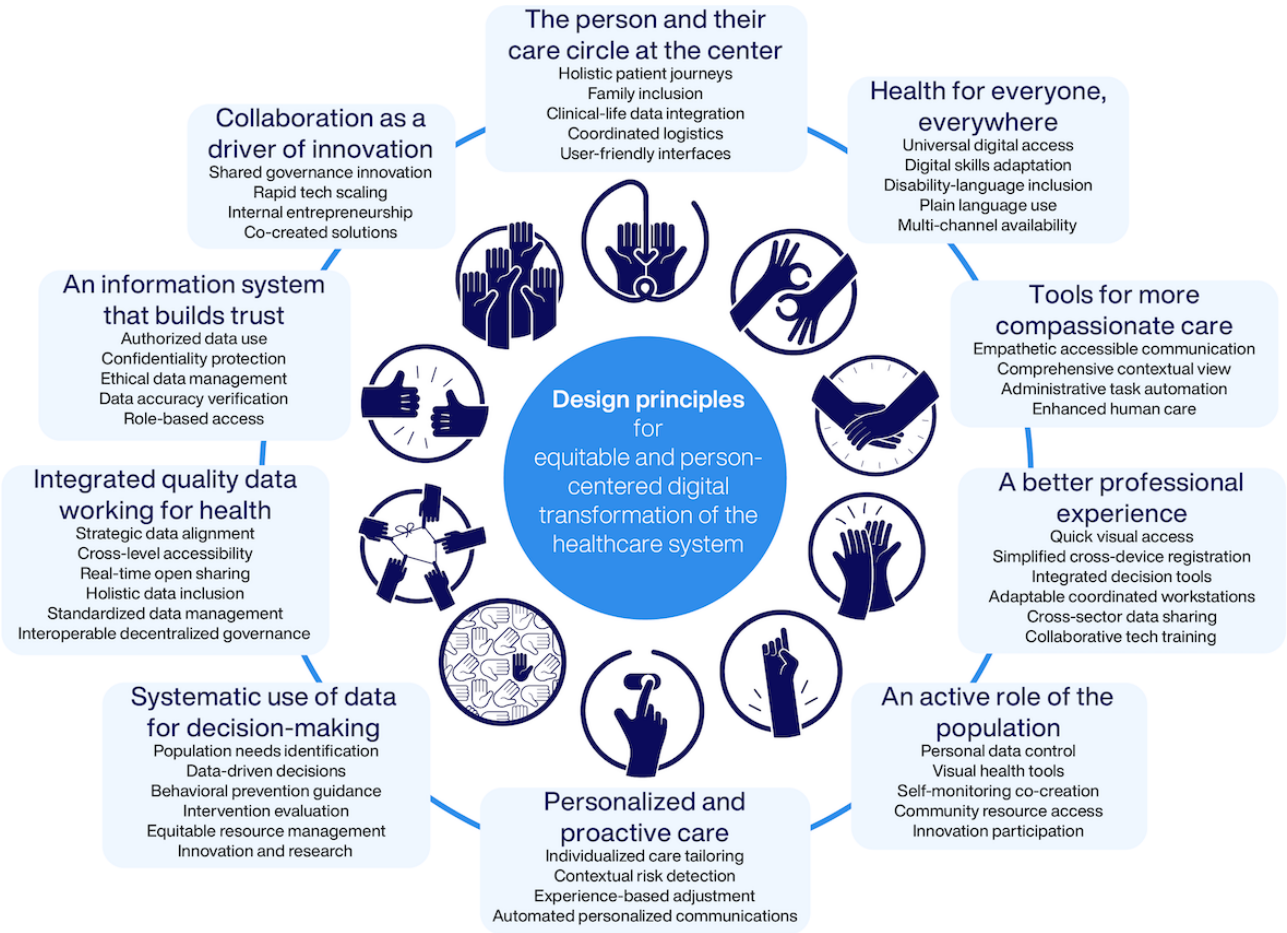

Territorial Context and Principle Development

Territorial analysis revealed how geographic and infrastructure contexts shaped both barrier experiences and the development of information system design principles. Rural areas emphasized connectivity and resource access issues, leading to a stronger emphasis on principles addressing “Health for Everyone, Everywhere” and “Collaboration as a Driver of Innovation.” Urban contexts highlighted diversity and complexity challenges, informing principles focused on cultural responsiveness and care coordination. Border regions contributed unique perspectives on administrative complexity and cross-jurisdictional care access, strengthening principles addressing system integration and user navigation support. These territorial differences informed the development of principles that explicitly address contextual variation rather than assuming standardized solutions.

XSL•FO
RenderX

Stakeholder Perspective Alignment and Divergence

While remarkable alignment emerged across stakeholder groups on fundamental principles, meaningful divergences in emphasis provided valuable insights for implementation. Health care professionals prioritized workflow integration and administrative burden reduction more strongly than other groups, directly informing the principle of "A Better Professional Experience." Patients and caregivers emphasized ease of use and personal control over health information, strengthening principles focused on user empowerment and data governance. Health care managers and leaders focused on sustainability and implementation feasibility, contributing to principles addressing systematic data use and collaborative innovation. Experts in digital health innovation highlighted technical interoperability and innovation potential, enriching principles related to data integration and system connectivity.

Although rare, certain divergences revealed conflicting views on the same topic. For instance, regarding mental health data sharing, parents expressed concerns about schools accessing their children's information, while professionals emphasized the need for cross-team coordination to optimize care. Such tensions were addressed by exploring underlying reasoning during facilitation and using divergences generatively during synthesis. Rather than forcing consensus, conflicting perspectives directly informed principles that could respond equitably to multiple stakeholder needs, such as Principle 9 on data governance, which emphasizes individual control over context-specific information sharing.

Implementation Considerations

Participants not only cocreated principles but also provided detailed insights into implementation considerations. They emphasized the need for phased implementation approaches that build confidence and capability gradually, strong change management and training programs for both citizens and health and social care professionals, and maintaining human connection even as systems become more automated. The cocreation process itself was identified as a valuable model for ongoing stakeholder engagement throughout implementation, with participants requesting continued involvement in system development and evaluation.

Participation Experience and Engagement

Participants demonstrated high levels of engagement throughout the cocreation process. Postsession evaluation surveys indicated strong satisfaction with the participatory experience, with participants reporting that they felt heard, respected, and valued in the process. The visual facilitation tools, particularly the "Pérez Family" journey mapping and storytelling, proved effective in enabling participants with varying communication styles and literacy levels to contribute meaningfully to discussions while providing a common visual framework upon which diverse experiences, ideas, and perspectives could be consolidated to create a shared mental model. The diversity of profiles facilitated rich exchanges between different perspectives, with participants frequently noting the value of hearing experiences from other stakeholder groups.

Discussion

Principal Findings

In this study, PHCD allowed us to generate a deep understanding of lived experiences and challenges faced by diverse stakeholders navigating the health care ecosystem in Catalonia, while simultaneously cocreating principles for guiding equitable digital transformation. Through engaging citizens, caregivers, professionals, managers, and different expert profiles in collective reflection and design, the process not only uncovered structural and experiential barriers but also translated these insights into actionable principles for future health information systems. Finally, the implemented and documented large-scale, iterative co-design approach provides a replicable methodology that other health systems can adapt to embed equitable care supported by health information technology as a central value in their transformation processes.

Our participatory approach yielded ten design principles that address critical gaps in current digital health development. These principles emerge from a comprehensive understanding of real-world challenges faced by patients, caregivers, health care professionals, and system leaders across diverse territorial contexts. The cocreation process revealed that digital health equity requires attention to multiple interconnected dimensions: technological accessibility, organizational readiness, professional workflow integration, and citizen empowerment in data governance.

The barriers identified (eg, digital literacy gaps, fragmented information systems, and insufficient consideration of vulnerable populations) align with the Digital Health Care Equity Framework in the United States [27], which similarly emphasizes addressing digital determinants of health through participatory design and community engagement. Our findings extend this work by showcasing how large-scale participatory processes can systematically translate these equity considerations into actionable design principles for driving digital transformations in the health care system.

The 10 principles address long-standing challenges in health IT development, particularly the gap between technological capabilities and real-world usability. Recent research continues to document usability problems, with over one-third of medication safety events in pediatric hospitals related to EHR usability issues[57].

The principle of "The Person and their Circle at the Center" reflects growing recognition that meaningful stakeholder involvement throughout design is essential [58]. This aligns with evidence from an umbrella review identifying participatory design and community collaboration as crucial for creating digital health tools serving diverse user needs [59]. Our study extends this evidence by scaling the community collaboration across an entire regional health system and applying it to define core design principles.

The emphasis on data control and citizen agency in the principle "An Information System that Builds Trust" resonates with recent research on patient preferences for health information exchange design, revealing complex attitudes toward health information

XSL•FO
RenderX

control and supporting our finding that genuine patient engagement is essential for addressing privacy and trust concerns [60].

Knowledge Contributions Beyond Design Artifacts

While producing 10 specific design principles, our research contributes to understanding digital health equity beyond these outputs through three key areas.

First, our findings illustrate how digital health equity barriers operate across multiple interconnected levels in practice. Previous frameworks have conceptualized equity barriers at individual, interpersonal, community, and societal levels [26,28]; implementation guidance has often remained siloed [22,29]. Our participatory process documented how stakeholders experience these barriers as deeply interconnected: technically accessible interfaces are undermined when organizational training support is missing; well-designed systems fail when procurement policies exclude user input; individual digital literacy interventions have limited impact when infrastructure gaps persist. In this regard, our findings suggest that equity-centered transformation may require coordinated technical, organizational, and systemic change because barriers appear to operate synergistically across these dimensions.

Second, our study documents how participatory design processes can serve as more than requirements-gathering exercises. Traditional health IT development treats stakeholder engagement as a requirements-gathering phase that feeds into expert-driven design [5,6]. Our experience indicates that participatory processes may generate equity-relevant outcomes through several mechanisms: surfacing barriers invisible to system designers [10], legitimizing experiential knowledge alongside technical expertise [18], and building stakeholder investment in transformation [59].

Finally, our research contributes to growing evidence highlighting the challenge of retrofitting equity considerations, which instead need to be embedded from the outset. The barriers identified appeared to stem not from isolated technical shortcomings but from design processes that systematically excluded diverse users [61]. Therefore, increasing equity may require restructuring design processes themselves, shifting from technology-driven implementation, where solutions are predetermined, to genuinely participatory cocreation, where diverse voices shape problem definition, solution ideation, and implementation priorities [14,15,21]. This concept aligns with that of other authors claiming that design processes themselves can either reproduce or mitigate existing health inequities [1,2,10].

Limitations and Transferability

While our study provides valuable insights into participatory design for digital health equity, the external validity of our principles and conclusions may be influenced by intrinsic features of the study design.

The geographic focus on a single area under the same health care system (ie, Catalonia) may have led to some insights not applicable to other countries and thus limiting transferability. However, we consider most of our findings applicable and valuable beyond regional boundaries through several dimensions. First, many of the identified barriers represent universal systemic challenges (workforce shortages, resource constraints, service fragmentation, and navigation complexity) documented across diverse health care contexts. Second, the methodological framework itself is transferable due to the inclusion of territorial segmentation (rural vs urban contexts) and multilevel stakeholder engagement (citizens to policymakers and health care to social care), which are categories relevant to most health systems. Third, specific findings regarding data duplication between public and private providers are particularly relevant and common in mixed-model health care systems, which are increasingly prevalent globally. Finally, the governance of the Catalan health system is regionally commissioned, like other environments, such as England's Integrated Care Boards, enhancing organizational-level transferability. Future research should examine how these principles perform across different health care contexts, particularly in low-resource settings where digital equity challenges may be more pronounced.

Our study captured stakeholder perspectives at a specific point during rapid digital transformation. The long-term sustainability and relevance of identified principles will require ongoing evaluation as digital health technologies and user needs evolve. However, we attempted to ensure future-proofing through grounding principles in fundamental human-centered design values, focusing on equity considerations that persist regardless of specific technologies, and validating principles through diverse stakeholder perspectives representing enduring health care needs.

Another feature that may influence generalizability is the participant profile involved in the sessions. Despite our effort in recruiting a diverse profile regarding roles, geographic location, and demographic characteristics, the voluntary nature of participation may have introduced selection bias towards individuals with a stronger interest in digital health transformation. This was particularly relevant for migrant and underserved communities, which were underrepresented. These populations often face the most significant barriers to digital health access. The exclusion of these voices represents a significant limitation, as these communities are frequently most affected by digital health inequities.

Future research could explore mechanisms for engaging harder-to-reach populations, with particular attention to developing culturally appropriate recruitment strategies for migrant communities and other underserved populations.

Implications for Practice and Policy

Our findings have important implications for health care organizations, technology vendors, and policymakers. For health care organizations, our study highlights the value and feasibility of investing in meaningful participatory processes before, during, and after health information system implementations. The principles provide a framework for evaluating existing systems and guiding future technology decisions with explicit attention to equity considerations.

XSL•FO
RenderX

A critical insight is the importance of innovative partnership approaches when navigating system complexity to drive large-scale engagement and co-design processes. In our experience, successful system-level transformation requires sophisticated collaboration models bringing together diverse expertise and perspectives. The partnership between public health authorities, academic institutions, and specialized design practitioners created a unique configuration enabling navigation of complex health care system dynamics whilst maintaining methodological rigor and stakeholder trust. This proved essential for accessing diverse participant networks, managing territorial complexities, and ensuring that cocreated principles would be both evidence-based and implementable within existing structures.

For technology vendors, our research highlights the importance of embedding human-centered design processes throughout product development, not merely as compliance exercises. Persistent usability problems documented in EHR research suggest current vendor approaches to stakeholder engagement are insufficient [52]. Our study provides evidence that more comprehensive participatory approaches can yield insights leading to more effective, equitable, and sustainable system designs.

For policymakers, our findings support the need for regulatory frameworks that incentivize or require meaningful stakeholder engagement in health technology development. Our study demonstrates the feasibility and value of comprehensive participatory approaches that could inform future policy development. Additionally, policymakers should consider how to support and incentivize the development of collaborative partnerships and co-design capabilities within health care systems, recognizing that these represent critical infrastructure for equitable digital transformation.

Future Directions

The 10 principles provide a foundation for future research examining their implementation and effectiveness in practice. Critical areas include extending these principles to social and community care settings, examining how they apply to artificial intelligence–enabled health technologies [62], and exploring how they can inform the development of community health information systems. Implementation research examining how organizations can operationalize these principles, including necessary capabilities, resource requirements, and implementation strategies, would provide valuable guidance for translating insights into improved health information systems.

Conclusions

This study showcases how PHCD, conducted at a large scale and grounded in rigorous methodology while engaging all relevant stakeholders, can serve as a powerful tool to identify equity barriers in health information systems and inform relevant and specific principles driving digital transformation of health care systems. The 10 principles emerging from this study provide a comprehensive framework for developing person-centered health information systems that are not only technically robust but also genuinely responsive to user needs, culturally appropriate, and accessible to all populations.

This study makes three distinctive contributions that differentiate it from existing research. First, it is the first successful application of PHCD at a regional health system scale, substantially exceeding the typical scale of participatory design studies in health care. Second, it provides empirical evidence of how equity barriers operate synergistically across technical, organizational, and systemic levels in practice, challenging the siloed implementation approaches suggested by previous conceptual frameworks. Third, it exemplifies how participatory processes can serve as transformative mechanisms for embedding equity from the outset, rather than the requirements-gathering exercises that characterize traditional health IT development. These contributions provide both a methodological blueprint for scaling participatory approaches and evidence that restructuring design processes themselves is essential for achieving digital health equity.

Acknowledgments

We wish to acknowledge the valuable contributions of those individuals who participated in the study, including patients, caregivers, health care professionals, and managers who contributed to the cocreation sessions. We are grateful to the territorial health coordinators across Catalonia who facilitated participant recruitment and supported session logistics in their respective regions. We acknowledge the invaluable contributions of The Care Lab team members who made this research possible: Andreu Coy, Andrea Barbiero, and Mercè Gamell, whose expertise in project management, facilitation, documentation, and methodological support was essential to the success of this participatory process. The authors disclose the use of generative artificial intelligence software (Grammarly) for grammar accuracy and writing consistency. No artificial intelligence bots were used to generate contents.

Funding

This study was funded by the European Union – NextGenerationEU, under the Recovery, Transformation, and Resilience Plan (PRTR) of the Spanish Government. The views expressed are those of the authors and not necessarily those of the Spanish Ministry of Health or the Funding bodies, which did not contribute to study design or results interpretation.

Data Availability

The transcription of discussions, pictures of the sticky notes as part of the participatory process, and clustering of themes (in Catalan and Spanish) are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

JP-J conceived and designed the study, developed the methodology, acquired funding, supervised the research, and wrote the original draft of the manuscript. XM, LLV, AC, LP, LH, NV, and AM contributed to study design and methodology development. JP-J, XM, LLV, AC, LP, LH, AB, AF, NV, and AM conducted data analysis and interpretation. AF, NV, and AM managed data collection and curation. XM, LP, NV, AM, and GCS contributed to manuscript writing and revision. AC provided project administration and coordination. NV, AF, and AM created visualizations and figures. All authors contributed to the interpretation of results, critically reviewed the manuscript for important intellectual content, and approved the final version for publication.

Conflicts of Interest

JP-J is Associate Editor of the *Journal of Medical Internet Research.*

Multimedia Appendix 1
Participant selection and recruitment strategy, co-design materials, and participation surveys.
[PDF File, 12001 KB - jmir_v28i1e84129_app1.pdf]

Multimedia Appendix 2
Round 1 outcomes (extended version).
[PDF File, 7802 KB - jmir_v28i1e84129_app2.pdf]

References

1. Crawford A, Serhal E. Digital health equity and COVID-19: the innovation curve cannot reinforce the social gradient of health. J Med Internet Res 2020 Jun 2;22(6):e19361. [doi: 10.2196/19361] [Medline: 32452816]
2. Rodriguez JA, Clark CR, Bates DW. Digital health equity as a necessity in the 21st century cures act era. JAMA 2020 Jun 16;323(23):2381-2382. [doi: 10.1001/jama.2020.7858] [Medline: 32463421]
3. Lyles CR, Sharma AE, Fields JD, Getachew Y, Sarkar U, Zephyrin L. Centering health equity in telemedicine. Ann Fam Med 2022;20(4):362-367. [doi: 10.1370/afm.2823] [Medline: 35879077]
4. Ramsetty A, Adams C. Impact of the digital divide in the age of COVID-19. J Am Med Inform Assoc 2020 Jul 1;27(7):1147-1148. [doi: 10.1093/jamia/ocaa078] [Medline: 32343813]
5. Martikainen S, Korpela M, Tiihonen T. User participation in healthcare IT development: a developers' viewpoint in Finland. Int J Med Inform 2014 Mar;83(3):189-200. [doi: 10.1016/j.ijmedinf.2013.12.003] [Medline: 24382475]
6. Kushniruk A, Nøhr C. Participatory design, user involvement and health IT evaluation. Stud Health Technol Inform 2016;222(139–151):139-151. [Medline: 27198099]
7. Clemensen J, Rothmann MJ, Smith AC, Caffery LJ, Danbjorg DB. Participatory design methods in telemedicine research. J Telemed Telecare 2017 Oct;23(9):780-785. [doi: 10.1177/1357633X16686747] [Medline: 28027678]
8. Arndt BG, Beasley JW, Watkinson MD, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Ann Fam Med 2017 Sep;15(5):419-426. [doi: 10.1370/afm.2121] [Medline: 28893811]
9. Irizarry T, DeVito Dabbs A, Curran CR. Patient portals and patient engagement: a state of the science review. J Med Internet Res 2015 Jun 23;17(6):e148. [doi: 10.2196/jmir.4255] [Medline: 26104044]
10. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. J Am Med Inform Assoc 2018 Aug 1;25(8):1080-1088. [doi: 10.1093/jamia/ocy052] [Medline: 29788380]
11. Boy GA, Riedel N. Participatory human-centered design: user involvement and design cross-fertilization. In: Kurosu M, editor. Human Centered Design: Springer; 2009:835-843. [doi: 10.1007/978-3-642-02806-9_96]
12. Giacomin J. What is human centred design? Des J 2014 Dec;17(4):606-623. [doi: 10.2752/175630614X14056185480186]
13. Bazzano AN, Martin J, Hicks E, Faughnan M, Murphy L. Human-centred design in global health: a scoping review of applications and contexts. In: Virgili G, editor. PLoS ONE 2017;12(11):e0186744. [doi: 10.1371/journal.pone.0186744] [Medline: 29091935]
14. Chen E, Leos C, Kowitt SD, Moracco KE. Enhancing community-based participatory research through human-centered design strategies. Health Promot Pract 2020 Jan;21(1):37-48. [doi: 10.1177/1524839919850557] [Medline: 31131633]
15. Göttgens I, Oertelt-Prigione S. The application of human-centered design aproaches in health research and innovation: a narrative review of current practices. JMIR Mhealth Uhealth 2021 Dec 6;9(12):e28102. [doi: 10.2196/28102] [Medline: 34874893]
16. Altman M, Huang TTK, Breland JY. Design thinking in health care. Prev Chronic Dis 2018 Sep 27;15:E117. [doi: 10.5888/pcd15.180128] [Medline: 30264690]
17. Holeman I, Kane D. Human-centered design for global health equity. Inf Technol Dev 2020 Jul 2;26(3):477-505. [doi: 10.1080/02681102.2019.1667289]

XSL•FO
RenderX

18. Bratteteig T, Wagner I. Disentangling power and decision-making in participatory design. 2012 Aug 12 Presented at: PDC '12; Aug 12, 2012; Roskilde, Denmark p. 41-50 URL: https://dl.acm.org/doi/proceedings/10.1145/2347635 [accessed 2025-12-18] [doi: 10.1145/2347635.2347642]
19. Merkel S, Kucharski A. Participatory design in gerontechnology: a systematic literature review. Gerontologist 2019 Jan 9;59(1):e16-e25. [doi: 10.1093/geront/gny034] [Medline: 29788319]
20. Duque E, Fonseca G, Vieira H, Gontijo G, Ishitani L. A systematic literature review on user centered design and participatory design with older people. 2019 Oct 22 Presented at: Proceedings of the 18th Brazilian Symposium on Human Factors in Computing Systems; Oct 22, 2019; Vitória Espírito Santo, Brazil p. 1-11 URL: https://dl.acm.org/doi/proceedings/10.1145/3357155 [accessed 2025-12-18] [doi: 10.1145/3357155.3358471]
21. Clemensen J, Larsen SB, Kyng M, Kirkevold M. Participatory design in health sciences: using cooperative experimental methods in developing health services and computer technology. Qual Health Res 2007 Jan;17(1):122-130. [doi: 10.1177/1049732306293664] [Medline: 17170250]
22. Evans L, Evans J, Pagliari C, Källander K. Scoping review: exploring the equity impact of current digital health design practices. Oxford Open Digital Health 2023 Jan 1;1. [doi: 10.1093/oodh/oqad006]
23. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2025-12-03]
24. Sanders EBN, Stappers PJ. Co-creation and the new landscapes of design. CoDesign 2008 Mar;4(1):5-18. [doi: 10.1080/15710880701875068]
25. Wadley G, Lederman R, Gleeson J, Alvarez-Jimenez M. Participatory design of an online therapy for youth mental health. 2013 Nov 25 Presented at: OzCHI '13; Nov 25, 2013; Adelaide Australia p. 517-526 URL: https://dl.acm.org/doi/proceedings/10.1145/2541016 [accessed 2025-12-18] [doi: 10.1145/2541016.2541030]
26. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119. [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
27. Hatef E, Hudson Scholle S, Buckley B, Weiner JP, Austin JM. Development of an evidence- and consensus-based Digital Healthcare Equity Framework. JAMIA Open 2024 Dec;7(4):ooae136. [doi: 10.1093/jamiaopen/ooae136] [Medline: 39553827]
28. Alvidrez J, Castille D, Laude-Sharp M, Rosario A, Tabor D. The National Institute on Minority Health and Health Disparities Research Framework. Am J Public Health 2019 Jan;109(S1):S16-S20. [doi: 10.2105/AJPH.2018.304883] [Medline: 30699025]
29. Bucher A, Chaudhry BM, Davis JW, et al. How to design equitable digital health tools: a narrative review of design tactics, case studies, and opportunities. In: Kuo PC, editor. PLOS Digit Health 2024 Aug;3(8):e0000591. [doi: 10.1371/journal.pdig.0000591] [Medline: 39172776]
30. Pérez Sust P, Solans O, Fajardo JC, et al. Turning the crisis into an opportunity: digital health strategies deployed during the COVID-19 outbreak. JMIR Public Health Surveill 2020 May 4;6(2):e19106. [doi: 10.2196/19106] [Medline: 32339998]
31. Solans O, Vidal-Alaball J, Roig Cabo P, et al. Characteristics of citizens and their use of teleconsultations in primary care in the Catalan public health system before and during the COVID-19 pandemic: retrospective descriptive cross-sectional study. J Med Internet Res 2021 May 27;23(5):e28629. [doi: 10.2196/28629] [Medline: 33970867]
32. Vidal-Alaball J, López Seguí F, Garcia Domingo JL, et al. Primary care professionals' acceptance of medical record-based, store and forward provider-to-provider telemedicine in Catalonia: results of a web-based survey. Int J Environ Res Public Health 2020 Jun 8;17(11):4092. [doi: 10.3390/ijerph17114092] [Medline: 32521740]
33. Moll J, Rexhepi H, Cajander Å, et al. Patients' experiences of accessing their electronic health records: national patient survey in Sweden. J Med Internet Res 2018 Nov 1;20(11):e278. [doi: 10.2196/jmir.9492] [Medline: 30389647]
34. Norman DA, Draper SW. User Centered System Design: New Perspectives on Human-Computer Interaction 1986. URL: https://www.taylorfrancis.com/books/9781482229639 [accessed 2025-12-03]
35. IDEO. Human Centered Design: Toolkit, 2nd edition: IDEO; 2011.
36. Steen M. Tensions in human-centred design. CoDesign 2011 Mar;7(1):45-60. [doi: 10.1080/15710882.2011.563314]
37. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
38. Trisha G, Seye A. The NASSS Framework – a Synthesis of Multiple Theories of Technology Implementation: Studies in Health Technology and Informatics IOS Press; 2019. [doi: 10.3233/SHTI190123]
39. Marcus HJ, Ramirez PT, Khan DZ, et al. The IDEAL framework for surgical robotics: development, comparative evaluation and long-term monitoring. Nat Med 2024 Jan;30(1):61-75. [doi: 10.1038/s41591-023-02732-7] [Medline: 38242979]
40. Rangachari P, Al Arkoubi K, Shindi R. A multi-level framework for advancing digital health equity in learning health systems: aligning practice and theory with the Quintuple Aim. Int J Equity Health 2025 Oct 7;24(1):253. [doi: 10.1186/s12939-025-02663-4] [Medline: 41057836]
41. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]

XSL•FO
RenderX

42. Marc de la participació ciutadana en salut: facilitem i potenciem la participació ciutadana; fem-ho entre tots [Article in Catalan]. Secretaria d'Atenció Sanitària i Participació. 2017. URL: https://scientiasalut.gencat.cat/bitstream/handle/11351/3503/marc_participacio_ciutadana_salut_2017.pdf?sequence=1&isAllowed=y [accessed 2025-08-11]
43. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health 2015 Sep;42(5):533-544. [doi: 10.1007/s10488-013-0528-y] [Medline: 24193818]
44. Baum FE, Ziersch AM, Zhang G, Osborne K. Do perceived neighbourhood cohesion and safety contribute to neighbourhood differences in health? Health Place 2009 Dec;15(4):925-934. [doi: 10.1016/j.healthplace.2009.02.013] [Medline: 19403326]
45. Hirko KA, Kerver JM, Ford S, et al. Telehealth in response to the COVID-19 pandemic: implications for rural health disparities. J Am Med Inform Assoc 2020 Nov 1;27(11):1816-1818. [doi: 10.1093/jamia/ocaa156] [Medline: 32589735]
46. Trebble TM, Hansi N, Hydes T, Smith MA, Baker M. Process mapping the patient journey: an introduction. BMJ 2010 Aug 13;341(aug13 1):c4078. [doi: 10.1136/bmj.c4078] [Medline: 20709715]
47. McCarthy S, O'Raghallaigh P, Woodworth S, Lim YL, Kenny LC, Adam F. An integrated patient journey mapping tool for embedding quality in healthcare service reform. Journal of Decision Systems 2016 Jun 10;25(sup1):354-368. [doi: 10.1080/12460125.2016.1187394]
48. Simonse L, Albayrak A, Starre S. Patient journey method for integrated service design. Design for Health 2019 Jan 2;3(1):82-97. [doi: 10.1080/24735132.2019.1582741]
49. Best A, Greenhalgh T, Lewis S, Saul JE, Carroll S, Bitz J. Large-system transformation in health care: a realist review. Milbank Q 2012 Sep;90(3):421-456. [doi: 10.1111/j.1468-0009.2012.00670.x] [Medline: 22985277]
50. Joseph AL, Kushniruk AW, Borycki EM. Patient journey mapping: Current practices, challenges and future opportunities in healthcare. Knowl Manag E-Learn 2020 Dec 26:387-404. [doi: 10.34105/j.kmel.2020.12.021]
51. WeCare Toolkit. National Council of Social Service. 2016. URL: https://isomer-user-content.by.gov.sg/24/c875b227-ad66-4127-b982-34412636febd/We-Care-Toolkit.pdf [accessed 2025-10-26]
52. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
53. Beyer H, Holtzblatt K. Contextual Design: Defining Customer-Centered Systems: Morgan Kaufmann; 2009.
54. Miles MB, Huberman AM, Saldaña J. Qualitative Data Analysis: A Methods Sourcebook: SAGE; 2020.
55. Martin B. Universal Methods of Design: 100 Ways to Research Complex Problems, Develop Innovative Ideas, and Design Effective Solutions: Rockport Publishers; 2012.
56. Kolko J. Exposing the Magic of Design: A Practitioner's Guide to the Methods and Theory of Synthesis First Issued as an Oxford University Press Paperback: Oxford University Press; 2015.
57. Ratwani RM, Savage E, Will A, et al. Identifying Electronic Health Record Usability And Safety Challenges In Pediatric Settings. Health Aff (Millwood) 2018 Nov;37(11):1752-1759. [doi: 10.1377/hlthaff.2018.0699]
58. Barker W, Chang W, Everson J, et al. The evolution of health information technology for enhanced patient-centric care in the United States: data-driven descriptive study. J Med Internet Res 2024 Oct 28;26:e59791. [doi: 10.2196/59791] [Medline: 39466303]
59. Kilfoy A, Hsu TCC, Stockton-Powdrell C, Whelan P, Chu CH, Jibb L. An umbrella review on how digital health intervention co-design is conducted and described. NPJ Digit Med 2024 Dec 23;7(1):374. [doi: 10.1038/s41746-024-01385-1] [Medline: 39715947]
60. Jabour AM. Putting patients at the center of health information exchange design: an exploration of patient preferences for data sharing. Health Informatics J 2024;30(3):14604582241277029. [doi: 10.1177/14604582241277029] [Medline: 39142341]
61. van Velsen L, Ludden G, Grünloh C. The limitations of user-and human-centered design in an eHealth context and how to move beyond them. J Med Internet Res 2022 Oct 5;24(10):e37341. [doi: 10.2196/37341] [Medline: 36197718]
62. Wang T, Emami E, Jafarpour D, Tolentino R, Gore G, Rahimi SA. Integrating equity, diversity, and inclusion throughout the lifecycle of artificial intelligence for healthcare: a scoping review. In: Kuo PC, editor. PLOS Digit Health 2025 Jul;4(7):e0000941. [doi: 10.1371/journal.pdig.0000941] [Medline: 40658719]

Abbreviations

EHR: electronic health record
PHCD: participatory human-centered design
SRQR: Standards for Reporting Qualitative Research
WHO: World Health Organization

XSL•FO
RenderX

Edited by S Brini; submitted 15.Sep.2025; peer-reviewed by C Xie, C Kunze; revised version received 19.Nov.2025; accepted 20.Nov.2025; published 06.Jan.2026.

Please cite as:
Piera-Jiménez J, Vilarasau Creus N, Maymó Costa A, Michelena X, Climent Fageda A, Farré A, Herczeg L, Parameswaran L, Carot-Sans G, Valle L
Cocreating Principles for Digital Health Equity: Cross-Sectional, Qualitative Study for Participatory Human-Centered Design in Catalonia
J Med Internet Res 2026;28:e84129
URL: https://www.jmir.org/2026/1/e84129
doi:10.2196/84129

© Jordi Piera-Jiménez, Núria Vilarasau Creus, Ada Maymó Costa, Xabier Michelena, Andrea Climent Fageda, Alèxia Farré, László Herczeg, Lekshmy Parameswaran, Gerard Carot-Sans, Luis Valle. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Moving Research to Practice: A Qualitative Study Exploring Patient Perspectives of the Implementation and Sustainability of a Digital Health Intervention for Chronic Kidney Disease

Courtney J Lightfoot[1,2,3], PhD; Roseanne E Billany[1,2,3], PhD; Gurneet K Sohansoha[1,2,3], BSc; Thomas J Wilkinson[3,4,5], PhD; Melanie J Davies[3,4,5], MD; Thomas Yates[3,4,5], PhD; Alice C Smith[1,2,3], PhD; Matthew PM Graham-Brown[1,2,3,6], PhD

[1]Division of Cardiovascular Sciences, University of Leicester, Leicester, England, United Kingdom

[2]Leicester Partnership for Kidney Health Research, University of Leicester, University Road, Leicester, England, United Kingdom

[3]NIHR Leicester Biomedical Research Centre, Leicester, England, United Kingdom

[4]Diabetes Research Centre, University of Leicester, Leicester, England, United Kingdom

[5]Division of Global, Lifestyle and Metabolic Health, University of Leicester, Leicester, England, United Kingdom

[6]Department of Renal Medicine, University Hospitals of Leicester NHS Trust, Leicester, England, United Kingdom

Corresponding Author:
Courtney J Lightfoot, PhD
Division of Cardiovascular Sciences, University of Leicester, Leicester, England, United Kingdom

Abstract

Background: My Kidneys & Me (MK&M) is a digital health intervention (DHI) codeveloped to provide specialist health and lifestyle education aimed at enhancing self-management among individuals with chronic kidney disease. The MK&M DHI was shown to improve self-management behaviors in a multicenter randomized controlled trial (SMILE-K), but strategies to support effective integration into routine clinical practice are not known.

Objective: This study aimed to explore patient perspectives about real-world implementation and sustainability of the MK&M DHI.

Methods: For this substudy, participants from the SMILE-K trial were invited to take part in a semistructured qualitative interview to explore their perspectives on the implementation and sustainability of the MK&M DHI in routine clinical practice. Barriers and facilitators that could potentially influence the uptake and usage of the program in real-world settings, outside of a clinical trial, were also explored. Topics included accessibility, usability, integration into routine care, and types of support required to sustain engagement over time. Data were analyzed using reflexive thematic analysis.

Results: A total of 42 interviews were conducted with 35 individuals (mean age 62.5, SD 9.9, range 38‑84 years; 23/35, 66% male; 33/35, 94% White British; estimated glomerular filtration rate: mean 33.5, SD 12.4 mL/min/1.73 m^2). Five themes were identified, encompassing a range of critical factors that should be addressed to ensure the successful implementation of MK&M into routine kidney care and its alignment with real-world clinical needs. These include ensuring that information is timely, relevant, and tailored to individual needs. Participants highlighted the importance of promotion through multimodal strategies and approaches, leveraging diverse communication channels to maximize reach and engagement. Equitable access was identified as a priority, with potential barriers, such as digital literacy and digital poverty, requiring attention. Building trust, through reassurance and endorsement from trusted health care professionals, was seen as essential to strengthen confidence in MK&M. Finally, participants stressed the need for continued engagement and sustainability, ensuring that MK&M is successfully integrated into care pathways and supported over time to maintain its impact.

Conclusions: The study findings highlight that timely, well-targeted communication using diverse strategies will be critical for the successful uptake of MK&M among people living with chronic kidney disease. The identification of perceived factors that will influence implementation of the program provides actionable insights to guide the development of tailored implementation strategies to ensure that MK&M is relevant, acceptable, and feasible in real-world practice. These findings provide a foundation for designing approaches that are not only patient-centered and inclusive but also adaptable to diverse clinical contexts. By addressing these factors proactively, implementation efforts can promote equitable access, foster sustained engagement, and support the long-term integration of DHIs, such as MK&M, into routine kidney care.

Trial Registration: ISRCTN ISRCTN18314195; https://www.isrctn.com/ISRCTN18314195

(J Med Internet Res 2026;28:e91966) doi:10.2196/91966

XSL•FO
RenderX

KEYWORDS

chronic kidney disease; digital health; implementation; implementation science; self-management; patient education; health education; awareness

Introduction

Chronic kidney disease (CKD) affects approximately 13% of the general population worldwide [1]; it is currently the 10th leading cause of death, and projected to be the fifth leading cause of premature death worldwide by 2040 [2,3]. Individuals with CKD face a significantly higher risk of morbidity and mortality, primarily due to associated cardiovascular disease [4], and live with poor health and high symptom burden [5]. Self-management plays a critical role in helping people with CKD monitor their health, adhere to treatment plans, maintain a healthy lifestyle, and make informed decisions about their care [6,7]. Engaging in effective self-management behaviors can slow progression of the disease, reduce the risks of associated complications, and improve quality of life [6,8]. Providing accessible education, practical strategies, and interactive resources can empower patients to take greater control of their condition and stay engaged in their health care.

To support people living with CKD to better self-manage their condition, we developed a digital health intervention (DHI) called “My Kidneys & Me” (MK&M) [9]. MK&M is an evidence- and theory-based DHI designed to deliver comprehensive CKD-specific self-management education, support, and guidance for individuals with mild to moderate CKD [9], and is hosted on the MyDESMOND (a diabetes digital self-management program) platform [10]. The MK&M program was codeveloped with kidney health professionals, researchers, digital health specialists, and experts in complex intervention design, in partnership with people living with CKD and their families. MK&M offers a structured, interactive, and tailored approach to improve (1) patient activation, which includes enhancing awareness, understanding of CKD, health-related knowledge, confidence, and skills for engaging in health-promoting behaviors; and (2) CKD self-management behaviors, including symptom management, lifestyle changes, and physical function [9,11,12]. The development of MK&M and its features is discussed in detail elsewhere [9,11]. The MK&M DHI has been well received and used by people with CKD [13] and has demonstrated positive effects, particularly among those with low levels of activation (knowledge, skills, and confidence in managing their own health) and among those who used the program more [12].

Having demonstrated efficacy in a large multicenter randomized controlled trial, the next critical step is to understand how best to implement MK&M in routine clinical practice to maximize its impact on patient care and health outcomes. Translating research findings into clinical practice is essential to ensure that evidence-based interventions deliver real-world benefits for patients [14]. Successful implementation requires collaboration with health care professionals, integration into existing care pathways, and consideration of patient needs, including accessibility and usability [15,16]. Actively incorporating patient perspectives on how to improve access, engagement, and use of DHIs, such as MK&M, in routine clinical practice is essential to ensure successful implementation and that the program reflects real-world needs. Understanding key barriers and facilitators can further inform the development of targeted implementation strategies that enhance the efficiency of intervention delivery and support its successful uptake and large-scale adoption. Thus, the aim of this study was to explore patients’ perspectives on the real-world implementation and sustainability of the MK&M DHI.

Methods

Study Design

This qualitative substudy comprises qualitative data collected as part of the SMILE-K trial (ISRCTN18314195). The SMILE-K study protocol [9], pilot study [16], and primary findings [10,11] are published elsewhere. As this paper explores participants’ perceptions of the implementation and sustainability of the MK&M DHI, the data presented are predominantly from those who had access to and used the program (intervention group); data from the control group are included but are limited. The COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines were followed when reporting this study to enhance rigor, transparency, and completeness [17].

SMILE-K Study Design

The SMILE-K trial was a 20-week prospective, mixed methods, multicenter, randomized controlled, and parallel-group trial recruiting patients from 26 hospital sites across England between May 2021 and December 2022. All study procedures were conducted remotely due to the COVID-19 pandemic and to facilitate pragmatic evaluation. The study used a parallel mixed methods design; quantitative and qualitative data were collected concurrently during the randomized controlled trial. Full details of study methodology are published elsewhere [11,18]. Randomization was conducted using a computer-generated sequence in a 2:1 ratio (intervention:control), stratified by age (≤63 years vs >63 years). Group allocation was not blinded to researchers at the lead site or participants; however, researchers at local sites were blinded. The intervention group received the MK&M program, while the control group continued usual care, maintaining their lifestyle and clinical care plans. Control participants did not access MK&M during the trial but were offered the program upon study completion.

MK&M Intervention

The MK&M program is a web-based application, hosted on the MyDESMOND platform [10], that was systematically developed as a theory- and evidence-based digital self-management intervention for people with nondialysis CKD; it is designed to provide holistic CKD-specific self-management education, support, and guidance for people living with mild to moderate CKD. The development of MK&M and its features is discussed in detail elsewhere [9,11].

XSL•FO
RenderX

The MK&M program delivers content and resources aimed at increasing CKD knowledge and patient activation, reducing health risks, managing symptoms, and improving physical function. Key components include educational "Learn about" sessions covering kidneys, CKD, treatment, symptoms, associated risks, and lifestyle factors (eg, diet and physical activity), alongside behavior-focused "How to..." sessions offering practical strategies for making small, sustainable health and lifestyle changes. These sessions support goal setting, self-monitoring, and action planning to help individuals apply learning. Additional features include health trackers (eg, blood pressure, symptoms, and exercise), integration with activity devices (Fitbit, Garmin, and Google Fit), goal-setting tools, and social support forums. Content is delivered in varied formats, including text, graphics, animations, and videos, alongside interactive quizzes to reinforce knowledge.

For the SMILE-K trial, new content was released weekly for 10 weeks, so 20 weeks was chosen as the primary end point to allow participants time to review the educational content, put learning into action, and explore other features on the program. As part of the MK&M program, participants received automated email reminders when new content (such as "How to..." sessions) became available. Additional reminders were sent if participants had not logged in after 7, 14, or 28 days. While the program can send push notifications, these were disabled for the study to allow participants to use the program as they wished. Participants' usability and experiences of using the program have been described elsewhere [13].

Participants and Recruitment

To participate in the main trial, participants had to be aged 18 years and older with CKD stages 3-4 (estimated glomerular filtration rate [eGFR] 15 - 59 mL/min/1.73 m^2) or diagnosed with a progressive kidney condition with an eGFR >60 mL/min/1.73 m^2 (eg, polycystic kidney disease), not receiving renal replacement therapy. Eligible individuals were approached via an invitation letter and promotional flyer detailing the SMILE-K study and MK&M DHI by a member of their clinical team or a research nurse. Those interested contacted the research team by email to receive a web-based information sheet and electronic consent form. The information sheet detailed the data collected using the MK&M program and copies of the MyDESMOND (platform on which MK&M sits) 'Terms of Use," "Privacy/Cookies Policy," and "Acceptable Use Policy." After providing informed consent, participants completed an online baseline survey that collected outcome measures (detailed elsewhere [11]) via a secure link (Jisc Online Surveys). Following survey completion, participants were randomized (2:1) to either the intervention or control group. Participants randomized to the intervention group received immediate access to the MK&M program via a direct link with a unique user log-in. The control group participants were provided access following study completion. A video orientation to the platform was available upon login to MK&M, and technical support was provided to those who requested it.

All trial participants were eligible for the qualitative substudy. Interview participants were purposively selected to capture a diverse range of participant characteristics to ensure a representative population. Potential participants were sent an invitation letter via email, with a web-based information sheet and electronic consent form. Participants were contacted via email to arrange a suitable date and time for their interview. Participants were recruited to baseline interviews between November 2021 and October 2022, with postintervention interviews conducted between February 2022 and March 2023. All participants were approached at 10 weeks; a 20-week interview was offered to those who had completed the 10-week interview and agreed to further follow-up, had missed or declined the 10-week interview but remained willing to participate later, or if inclusion at 20 weeks supported maximum variation sampling. The control group participants were offered only 1 interview; intervention group participants could complete 1 or both interviews based on availability and sampling needs. Nonresponders were recontacted once, with no more than 3 total contact attempts across both time points.

Data Collection

Individual semistructured interviews were used to explore patient participants' perceptions of the implementation and sustainability of the MK&M DHI in routine clinical practice and barriers and facilitators that could potentially influence the uptake and usage of the program in real-world settings, outside of a clinical trial (eg, SMILE-K). The interview schedule was informed by current literature and developed in consultation with patient and public involvement partners. Topics included accessibility, usability, integration into routine care, and types of support required to sustain engagement over time. A common core interview guide was used across arms and time points to ensure comparability, with intervention-specific probes (eg, intervention engagement/usability and sustainability effects). Although the control group participants had not used the MK&M DHI, they were aware of it, and their perspectives were sought to explore anticipated barriers, contextual influences, and system-level factors relevant to implementation.

For this qualitative substudy, a target sample of 25 - 35 participants (20 - 30 intervention group and 5 - 10 control group) was set based on the concept of information power [19] and resource availability. Purposive sampling was used to capture a diverse range of perspectives. The adequacy of the sample size was continuously reviewed during data collection to ensure that an in-depth understanding of the investigated phenomena was obtained.

Participants were interviewed by telephone at 10 and/or 20 weeks postrandomization, using the same topic guide. Interviews lasted approximately 60 minutes, were audio-recorded, and transcribed verbatim by a professional service. All interviews were conducted by a single researcher (CJL), who is an experienced qualitative researcher. Participants had no relationship with the researcher prior to study commencement. The researcher kept a personal reflective logbook during interview conduction and analysis.

Data Analysis

All identifiable information, including names and personal details, was removed from transcripts. Data were managed in NVivo (version 12; QRS International) and analyzed using

XSL•FO
RenderX

interpretive reflexive thematic analysis following Braun and Clarke's approach [20]. This provides flexibility and an inductive, meaning-focused analytic process, allowing for interpretation of how participants made sense of their experiences while acknowledging the researchers' active, reflexive role in theme development. Initially, 1 researcher (CJL) reviewed the full dataset to familiarize themselves with the content through a process of immersion, noting early impressions and potential ideas. They then identified segments of the data that appeared potentially interesting and relevant, generating and applying initial codes, which were refined and finalized through review. Each transcript was independently coded by CJL. A sample of transcripts was checked by another researcher (REB) and cross-referenced against a developed codebook. Codes were compiled into clusters and collated together to form candidate themes. Potential themes were identified by CJL and reviewed and refined with REB through an iterative process, assessing to ensure the fit of the themes in relation to the data, both the coded extracts and the overall dataset. The researchers clearly defined and named the themes to provide detailed descriptions and clarification about what the theme is about and how it contributes to the analytical story. CJL and REB met regularly to discuss and reflect on the analysis and consider alternative interpretations; consensus was achieved through an iterative collaborative approach. A codebook and detailed analytic diary were maintained to ensure consistency and transparency. Findings are presented as direct quotes with participant ID, sex, age, and study group (I: intervention; C: control).

Reflexivity Statement

Both CJL and REB are experienced, qualitative researchers who had no prior relationship with the participants. CJL has expertise in patients' experiences of living with and self-managing their long-term condition(s), particularly CKD, as well as in the development, evaluation, and implementation of complex interventions to better support patient self-management. REB has expertise in patients' experiences of rehabilitation and exercise-based interventions, particularly kidney transplant recipients. While CJL was a codeveloper of the MK&M DHI, the interview participants were not aware of this to reduce any social desirability bias. REB was not involved in the development of MK&M.

To mitigate any potential bias, CJL engaged in reflexive bracketing and kept a reflexive diary, detailing their reflections on their own values and experiences and how these might shape the data. They also met regularly with REB, who was not involved in the development of MK&M, during data collection and analysis to discuss any potential biases that might influence the process and how to prioritize the participants' voices. Additionally, CJL and REB met with the wider research group to discuss findings and interpretation of the data.

Ethical Considerations

This study was fully approved by the Research Ethics Committee−Leicester South on November 19, 2020 (reference: 17/EM/0357). All participants provided online informed consent. All participants were given the opportunity to ask questions throughout the consent process. The participants did not receive compensation for participation. The study was conducted in accordance with the Declaration of Helsinki. The trial is sponsored by the University of Leicester (rgosponsor@le.ac.uk). The study was prospectively registered as ISRCTN18314195 on December 18, 2020.

Results

Overview

This paper focuses specifically on participants' views on the implementation and sustainability of MK&M into routine care. Participants' perspectives and experiences of using MK&M and ongoing engagement are reported elsewhere [13]. No meaningful differences between the groups were identified; thus, findings are reported at an aggregate level to provide a more coherent and analytically robust account of participants' views.

Summary of Findings From the SMILE-K Trial

A total of 420 participants were recruited into the SMILE-K trial, of whom 280 (66.7%) participants were randomized into the intervention group and were subsequently provided access to the MK&M program. The SMILE-K trial demonstrated that the MK&M program improved patient activation, although the overall effect was nonsignificant [12]. Notably, the greatest improvements occurred in individuals with low baseline activation and in those who used the program. Furthermore, MK&M was well-received and actively used by the participants [13].

Participant Characteristics

A total of 55 (36 intervention and 19 control) people were invited to participate, of whom 43 (78%) (31 intervention and 12 control) responded, and 35 (81%) individuals (27 intervention and 8 control) were subsequently interviewed. Interview participant characteristics are displayed in Table 1; in summary, participants had a mean age of 62.5 (SD 9.9; range 38 - 84) years, 66% (23/35) were male, 94% (33/35) were White British, and had a mean eGFR of 33.5 (SD 12.4) mL/min/1.73 m^2. A total of 42 interviews were conducted (29 at 10 weeks [23 intervention and 6 control] and 13 at 20 weeks [10 intervention and 2 control]), with 7 intervention participants interviewed twice at 10 and 20 weeks. Interviews lasted an average of 58 (range 32 - 96) minutes.

XSL•FO
RenderX

Table . Participant characteristics.

Participant ID	Study group	Age (years)	Sex	eGFR[a]	Ethnicity	Education	Employment	Follow-up interviews
SS01[b]	Intervention	38	Female	45.7	White British	University	Y[c]	1 - 10 weeks
SS02	Intervention	66	Female	28.4	White British	College	R[d]	1 - 10 weeks
SS03	Intervention	71	Female	37.2	White British	Other trade/vocation qualification	Y	1 - 10 weeks
SS04	Intervention	45	Female	57.6	White British	High school	Y	1 - 10 weeks
SS05	Intervention	61	Female	58.8	White British	Other trade/vocation qualification	R	1 - 10 weeks
SS06	Intervention	55	Female	35.9	White British	College	Y	1 - 10 weeks
SS07	Intervention	55	Male	20	White British	Other trade/vocation qualification	Y	1 - 10 weeks
SS08	Intervention	54	Male	29.8	White British	University	Y	1 - 10 weeks
SS09	Intervention	69	Male	24.1	White British	University	R	1 - 20 weeks
SS10	Intervention	60	Female	59.2	Other Mixed	College	NR[e]	1 - 10 weeks
SS11	Intervention	59	Male	47.5	White British	University	Y	1 - 10 weeks
SS12	Intervention	72	Male	33	White British	University	R	1 - 10 weeks
SS13	Intervention	57	Female	25.6	White British	University	R	2 times at 10 and 20 weeks
SS14	Intervention	57	Male	35.3	Other White	High school	Y	1 - 20 weeks
SS15	Intervention	67	Male	30.1	White British	College	R	2 times at 10 and 20 weeks
SS16	Intervention	68	Male	38.2	White British	College	Y	1 - 10 weeks
SS17	Intervention	69	Female	26.9	White British	University	R	2 times at 10 and 20 weeks
SS18	Intervention	60	Male	41.4	White British	High school	R	1 - 10 weeks
SS19	Intervention	71	Male	34.5	White British	University	Y=SE[f]	2 times at 10 and 20 weeks
SS20	Intervention	71	Male	35.7	White British	University	Y	1 - 10 weeks
SS21	Intervention	71	Male	19.1	White British	University	R	1 - 10 weeks
SS22	Intervention	55	Female	23.4	White British	Other trade/vocation qualification	Y	1 - 10 weeks
SS23	Intervention	54	Female	21.9	White British	College	Y	2 times at 10 and 20 weeks
SS24	Intervention	72	Male	37.5	White British	High school	R	1 - 20 weeks
SS25	Intervention	76	Male	16.7	White British	University	R	2 times at 10 and 20 weeks
SS26	Intervention	51	Male	32.1	White British	College	Y	2 times at 10 and 20 weeks
SS27	Intervention	62	Male	25.5	White British	College	Y	1 - 20 weeks
SS28	Control	58	Male	42.5	White British	University	Y	1 - 10 weeks
SS29	Control	69	Male	33.5	White British	College	R	1 - 10 weeks
SS30	Control	74	Male	57.0	White British	High school	R	1 - 10 weeks
SS31	Control	84	Male	19.5	White British	University	R	1 - 10 weeks

XSL•FO
RenderX

Participant ID	Study group	Age (years)	Sex	eGFR[a]	Ethnicity	Education	Employment	Follow-up interviews
SS32	Control	62	Male	46.2	White British	University	Y	1 - 10 weeks
SS33	Control	64	Male	18.8	White British	University	Y	1 - 10 weeks
SS34	Control	68	Male	21.4	White British	Other trade/vocation qualification	Y	1 - 20 weeks
SS35	Control	43	Female	15.7	White British	High school	Y	1 - 20 weeks

[a]eGFR: estimated glomerular filtration rate.
[b]SS: semistructured.
[c]Y: yes.
[d]R: retired.
[e]NR: not reported.
[f]SE: self-employed.

Themes

Overview of Findings

Five overarching themes were identified, each containing subthemes. Many of the identified factors perceived to influence the implementation and sustainability of MK&M DHI could be considered as both a barrier and a facilitator, depending on the context and individual circumstances. These factors can be encapsulated within implementation frameworks as determinants that influence the outcome of implementation efforts and will need to be considered when implementation planning. The key determinants perceived to influence the implementation of MK&M into routine kidney care are shown in Figure 1. The determinants identified are conceptualized as overlapping and dynamically interrelated rather than independent, reflecting the complex ways in which factors influencing DHI implementation interact.

Figure 1. Key determinants perceived to influence implementation of My Kidneys & Me. CKD: chronic kidney disease.

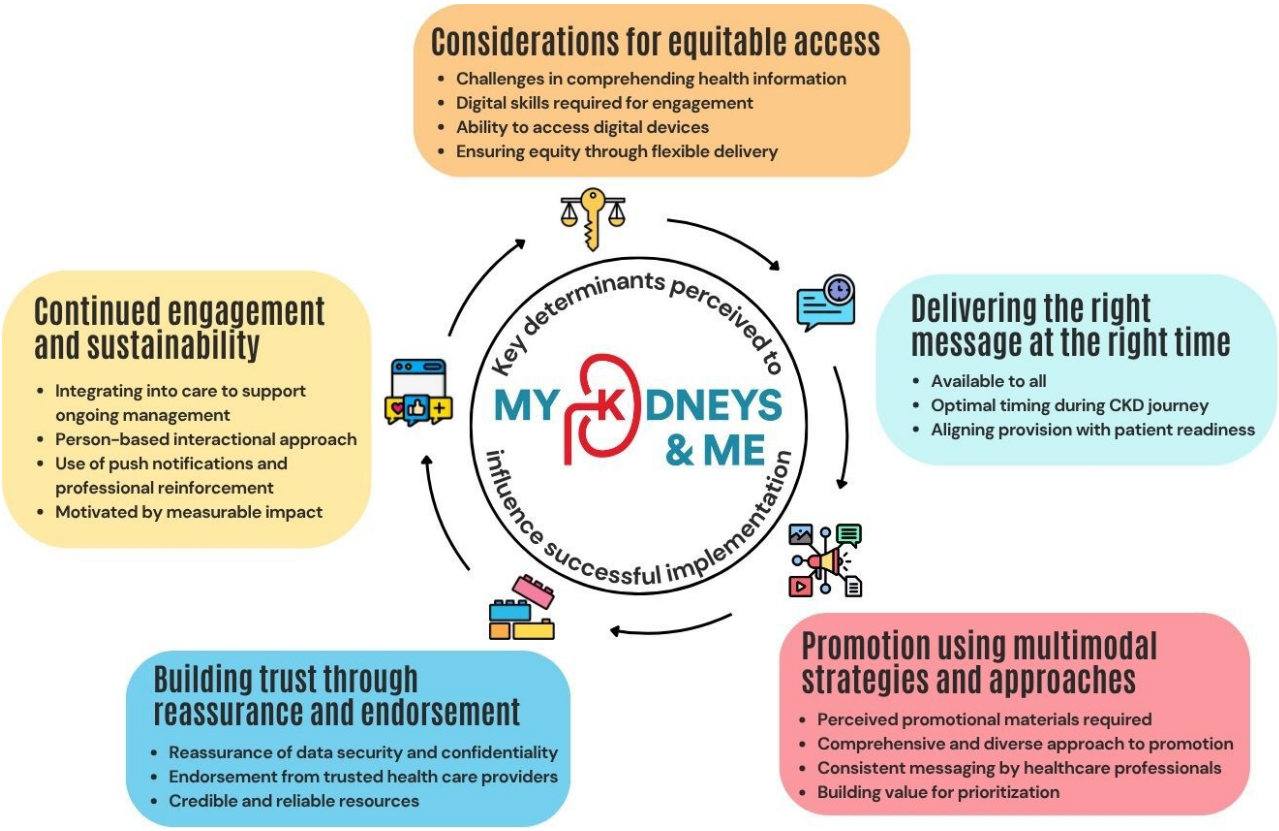

Delivering the Right Message at the Right Time

Available to All

Participants expressed that the intervention should be widely accessible and routinely available as part of standard care. They emphasized the importance of increasing general awareness and providing clear, comprehensive information to explain the program effectively. Making the resource broadly available could help empower individuals to better manage their health and support their overall well-being.

XSL•FO
RenderX

I think it would be good to have it routinely available, because I think people have got to the stage, these days especially, where they think oh I don't want to phone my doctor about that, oh I don't know if I should be doing this, and I think it would be helpful to be able to look it up and have a bit of a definite answer. [SS17 (F, 69, I)]

I think anything that's out there that people can just access online and read it, it's a good thing, as I say, there's loads of information in the programmes for people...It's going to help them to manage the condition and to hopefully slow down the progression, I mean, there's no cure for kidney disease and it is a progressive illness, but...you can slow it down. [SS18 (M, 60, I)]

Optimal Timing During CKD Journey

There were mixed views on the most appropriate timing for receiving the MK&M DHI. Some participants felt that offering it immediately at diagnosis would be most beneficial, providing clarity and support during an overwhelming period. However, other participants reported preferring a short time period between diagnosis and receiving MK&M, allowing individuals time to process and come to terms with their diagnosis before engaging with additional information. The appropriate time frame suggested for providing MK&M ranged from a few weeks after diagnosis to several months. Despite these differing perspectives, there was a shared consensus that MK&M should be provided early in a person's CKD journey, enabling individuals to understand their condition and learn how best to manage it.

It would have been a useful thing at the point of diagnosis. I think that would have really helped actually because it would have sent me away with something...so that at that point, the information about managing the condition and what you can do to potentially arrest its progression is really incredibly useful. [SS01 (F, 38, I)]

I suppose by the time you've got a consultant, you've actually got a proper diagnosis...you've gone to see them initially and then they've given you the bad news, and—I think maybe before you go see them, oh here you go, you can have a look at this—maybe it's a bit too daunting...Maybe to see them first, so you've got some rapport going. And then if you did have any issues or questions or what have you, you could talk to them after that. [SS06 (F, 55, I)]

I think it's better to be proactive really and I would think that something like the My Kidneys programme should probably be broadcast more widely...I think because there was a gap between when I was diagnosed and treatment was planned, it might be attractive to offer it at diagnosis, on the other hand of course some people might be in a bit of a state...So it might be better if it's just slightly after the diagnosis when people have had a chance to think, you know, what am I going to do. [SS30 (M, 74, C)]

Aligning Provision With Patient Readiness

The timing at which MK&M is provided and how it is communicated were considered to be crucial by participants and may require some flexibility to accommodate individual readiness. Participants suggested that there could be a most effective moment for providing MK&M, but this would depend on when an individual is most receptive to learning and motivated to engage with information about their CKD. Participants explained how clear, empathetic communication and repeated opportunities are essential to ensure that individuals receive MK&M when they are prepared to act, ultimately enhancing engagement and self-management.

It's initially quite a lot to take on. You think oh my goodness. But if people can be brought round to the idea of well it is a new way of life because that's what you want, isn't it, you do want life, you don't want to not have life. [SS17 (F, 69, I)]

I think maybe thinking about people's feelings, maybe it shouldn't be right at the minute you get diagnosed, maybe it should—you perhaps need that to sink in a little bit first—I think whenever you're given any sort of diagnosis it's something you need to come to terms with first of all. [SS18 (M, 60, I)]

Having something online like that would actually I think be very beneficial. when patients are first diagnosed that's when they want to know all the information about when they're going to help themselves and do this and do that. [SS05 (F, 61, I)]

It's about accessibility, isn't it, about context and benefits that an individual receives from that and all that will always be an individual situation, so then you can look at it in the broader sense it will always be down to the individual as to how that's perceived isn't it. [SS02 (F, 66, I)]

Promotion Using Multimodal Strategies and Approaches

Perceived Promotional Materials Required

Participants suggested practical ways to promote access to MK&M, including a leaflet with a QR code linking directly to the resource. They felt that this should ideally be provided by a clinician during an appointment to reinforce trust and encourage engagement. Posters with QR codes in waiting rooms were also recommended as a visible and convenient option. In addition, participants proposed sending information via post, email, or SMS text message, ensuring that those who do not have regular appointments can still easily receive information and guidance about how to access MK&M.

Possibly a leaflet or a poster in the waiting room type thing with a QR code or something like that. [SS08 (M, 54, I)]

I think a flier is always helpful to people, they can come away and think about it...most people have an email address so there's potential either to do that online or the piece of paper. [SS02 (F, 66, I)]

XSL•FO
RenderX

Have a leaflet. Hand it over. Put a QR code on the leaflet and tie it to the app so it's easy to download...If your clinician could pass over a leaflet at the point of an appointment then I think that would be a good thing. [SS01 (F, 38, I)]

Comprehensive and Diverse Approach to Promotion

Both general practitioner (GP) practices and hospitals were seen, by participants, as responsible for distributing promotional materials, ensuring consistent communication and availability across care settings. Many felt that the information should ideally be provided by a clinician during an appointment, followed by additional information later to act as a reminder. Multiple touchpoints were considered important because patients may not act on information the first time it is offered, and repeated exposure was felt likely to increase awareness and facilitate uptake and ongoing use.

Electronically through email...that's probably the easiest thing, or through a text even, I mean, people have got, probably got, most people have got phones haven't they. Or face to face, it doesn't matter. Actually, my preferred method would be probably through a professional! Because then I'd trust it basically. [SS32 (M, 62, C)]

I always like a face to face with a letter follow up. Because I think the face to face, well I like face to face so I can ask questions and I like the follow up letter for the things that I might have forgotten after being told whatever's wrong with me. I prefer a letter. I just like the consultant's letters, you know, it's set out just exactly what's wrong and what you've discussed. If there's any leaflets on the subject I don't mind those being included. [SS34 (M, 64, C)]

A leaflet handed to them saying if you have got access to it look at this, look at that or best practice, these are the dos and these are the don'ts. And maybe the doctor could hand one of those leaflets out...If it is not the doctor then the nephrology clinic should hand them out...I'd like it in a letter form first of all rather than a random e-mail. [SS29 (M, 69, C)]

Consistent Messaging by Health Care Professionals

Promotion of MK&M by doctors or other members of the multidisciplinary team at clinical appointments was believed to encourage engagement. Participants felt that all health care professionals, across GP practices and hospital settings, should actively advertise and promote MK&M, ensuring consistent messaging throughout the patient journey. Having different members of the multidisciplinary team reiterate the message and provide reminders of MK&M at multiple points, using the same information and consistent style of delivery, was deemed important. Repeated and uniform messaging was seen as essential for encouraging engagement and preventing the information from being overlooked. This approach was perceived to normalize MK&M as part of routine care and increase the likelihood of patients engaging with it.

I think it's important to come from a professional and I think it should be common across the professionals so they give the same information as well. And the same style of delivery if possible. [SS34 (M, 64, C)]

Nurses mentioning it or giving you leaflets or you can collect some sort of a booklet from somewhere when you go for tests, etc. Or to see a doctor...when you go for your blood pressure tests I guess the nurse could say something about it or whatever if they were more informed. [SS24 (M, 72, I)]

The obvious things through GPs and nephrology units in hospitals and medical centres,...I think they could easily offer it in much the same way that it was offered via my consultant, I mean, the GP could have said, knowing I've got this long term kidney problem, why don't you try this programme because I think it might help you. [SS30 (M, 74, C)]

Building Value for Prioritization

Participants believed that for individuals to benefit from MK&M, they need to perceive it as an opportunity to actively engage and take greater control of their health and well-being. Dedicating time to understanding the information provided by MK&M and applying these insights was believed to help people make informed decisions, manage their health more effectively, and feel empowered in their everyday lives. To ensure that MK&M becomes a priority for patients, participants considered it essential to create a personal connection to the program and provide support that encourages proactive steps toward better self-management. Participants also highlighted the need to actively promote and "sell" the value of MK&M, so that individuals clearly see its benefits and feel motivated to invest their time and effort.

Hopefully it's self-interest. I can't think of any other reason you would need, you would want to help, you should want to help yourself. [SS25 (M, 76, C)]

You need to tempt it—and it sounds ridiculous...but whatever it is you need to sell it...I think people are just so kind of perhaps naive or whatever, but it just needs to be sold to them, whatever it is, to get their interest. [SS07 (M, 55, I)]

I suppose life gets in the way, other things get in the way, you get waylaid by things...But I think the more informed you can be and I see this platform as giving me information and helping me take control, which is what I'm looking forward to really. Quite nice to have that guidance. [SS13 (F, 57, I)]

You need to be able to engender some sort of personal attachment to it really don't you...I don't know how you get people to take onboard their own authority in engaging and taking on some individual thought process!...I don't know how do you get people to start to think about themselves. You see you've got me more engaged now because I feel a sense of personal involvement to a degree. [SS17 (F, 69, I)]

XSL•FO
RenderX

Considerations for Equitable Access

Challenges in Comprehending Health Information

Health literacy was identified as a potential barrier to engaging with MK&M. Participants described how many individuals lack awareness of how best to look after their health and may not find the program useful if they do not understand the information provided. Participants, who considered themselves to be well educated, reported difficulties pronouncing medical terms, describing them as "another language" that can be challenging to comprehend. They felt that people with low levels of education would struggle, resulting in disengagement. Participants suggested that providing audio support alongside written information, particularly for complex terminology, combined with clear explanations, could significantly improve health literacy and enhance engagement with MK&M.

> *It sounds ridiculous, but trying to be less medical, OK, to put your head inside the patient who's going to know nothing about it whatsoever...but without being patronising.* [(SS08 (M, 55, I)]
>
> *Illiteracy levels. People are not aware of how they can help themselves, I think that's part of the problem.* [(SS13 (F, 57, I)]
>
> *There are certain words on it that you can't really pronounce and if you're not literate, you know what I mean, it does affect you, or if you're not... you know, if you've not got that brilliant education.* [(SS10 (F, 60, I)]

Digital Skills Required for Engagement

Participants discussed how individuals with low levels of digital literacy may face challenges in accessing and engaging with MK&M. To overcome this barrier, they suggested having "digital buddies" or "digital champions"; supportive individuals who can help participants develop essential digital skills and build confidence in using technology to promote inclusion and ensure that everyone can benefit from MK&M. It was considered that the majority of patients would be able to use MK&M, as the digital skills required are not overly complex. Participants noted that if individuals can get online or use an app, they should be able to navigate MK&M successfully.

> *You need to be someone who's either, well who's happy and confident using either a smart phone, smart device, computer, whatever. The barrier might be maybe more older people who aren't so happy using something like that. I would have thought most people my age and younger wouldn't have any great problem with it.* [SS14 (M, 57, I)]
>
> *I think modern technology is a barrier to a lot of people, particularly older people or people with learning needs. You tend to need to have a certain technical knowledge these days just to survive...a lot of people my age and above aren't quite as tech savvy as a twenty year old...As we know 90% of people could probably use the programme online.* [SS18 (M, 60, I)]
>
> *I'm sure that there are people who would struggle with it but I think the vast proportion wouldn't. I think it's pretty straightforward.* [SS25 (M, 76, I)]
>
> *I think there's always going to be a few people with difficulties. I think mainly because of it is online, but I think in this day and age, ninety something percent of people will be able to get online and get an app or something.* [SS27 (M, 62, I)]

Ability to Access Digital Devices

Digital poverty was identified as a potential barrier to engagement. Participants explained that some individuals may not own a computer or a smartphone, have access to one, or be able to afford a digital device or an internet connection. Participants perceived that there is always a risk of excluding people but believed that most individuals would have access to a digital device. Delivering information digitally was considered the most appropriate approach to ensure widespread availability. Suggestions about how to address gaps in access included creating opportunities for people to come together in group settings or encouraging the use of community resources such as libraries, where access to technology and support may be available.

> *If they've not got a phone or if you've got no credit on your phone or whatever, but you don't even need credit because it does it…you know, you can do it via your broadband, you know what I mean, you can do it on your laptop or on whatever you set it up on, so the only barrier that I would say is the individual.* [SS10 (F, 60, I)]
>
> *Certainly using it online or on an app is probably the most sensible way. There is always a danger that it excludes, there are always going to be people that haven't got phones or what have you, so there's always a danger of that, but I think realistically it probably is the most sensible way to do it.* [SS26 (M, 51, I)]
>
> *If you've got access to a computer it's good. I can use it but again we've got to think about the people who can't or can't afford it or just can't manage with computers. And I think these days I think you've got to think of both things.* [SS29 (M, 69, C)]

Ensuring Equity Through Flexible Delivery

Participants emphasized that accessibility is essential for equitable support. While online resources were seen as useful, many noted that personal circumstances, such as limited internet access, cost, or difficulty using technology, can make DHIs, such as MK&M, impractical for some individuals. They felt that alternative methods of delivering the educational content were needed for those who cannot access the online platform and should be offered in multiple formats. Participants also highlighted that health information and education should be tailored to different learning styles, using audio-visual formats alongside written materials, and supported by face-to-face sessions for clarity and personalized guidance.

> *As long as there's a mixture of - it's not just all one particular mode of learning, you know, so it's not just*

XSL•FO
RenderX

all online or in person or, you know, so a mixture of different things I think I'd prefer then it makes it more real, being able to interact but also do your own work and reflection. [SS11 (M, 59, I)]

These computer based sessions are very good if you enjoy doing computer based sessions...Personally, I get a lot more benefit from a face-to-face practical session than from a screen. It's just the way I like to learn. I get far more from the visual. [SS12 (M, 72, I)]

Maybe they would benefit from some sort of pamphlet or something, where it's just sort of like small concise bits, as opposed to maybe a whole programme. [SS06 (F, 55, I)]

If you could condense that programme with just the facts that would really help people, even if it was in a book form or a programme like you're doing...I think a booklet or something would probably benefit people that are starting out on a kidney problem. Then they could have perhaps at the end of that book a contact to go onto something further like this...because that would get them interested in it and then once they're interested in it they'll be happy to go further. [SS01 (F, 38, I)]

Building Trust Through Reassurance and Endorsement

Reassurance of Data Security and Confidentiality

Data protection was a key concern for most participants, who acknowledged that people can be cautious about sharing personal information online. Some individuals reported how they may be particularly suspicious about how their data are collected and used. Participants emphasized the importance of providing clear reassurance about data security, confidentiality, and anonymity to build trust and encourage participation in MK&M and other DHIs.

I suppose one of the views that you have that's always in the back of my mind about the data that you say, the things that you say and the information that's collected, that sort of stuff - need to be shared, so the anonymity is quite good in one sense isn't it. Because you see all the things about data breaches and all that sort of stuff. [SS11 (M, 59, I)]

People don't like to think people have got information about them and things like that, very wary in this day and age about releasing digital information, like any digital thing people aren't savvy about keeping their details locked away...You can take all the precautions you like and you can tell people that, whether they want to believe you is another story. [SS29 (M, 69, C)]

I don't want to give data out willy-nilly...I think some sceptics you won't ever reassure. [SS13 (F, 57, I)]

Endorsement From Trusted Health Care Providers

Participants emphasized that MK&M should be recommended by a known, trusted health care professional. Having clinician endorsement was seen as critical for credibility and uptake. It was believed that if health care professionals were invested in MK&M and its benefits, patients would be more likely to engage. While MK&M was perceived as potentially available nationally, participants highlighted the importance of providing relevant information locally through GP practices and hospital renal units to strengthen trust and accessibility.

If a clinician was able to go "You have chronic kidney disease. Your level of function is level 3a" or whatever. "This means that your GFR is x but the likelihood is that you're going to have this for a long time. Here's an app that would help you manage this condition"—that would have helped a lot...They're just authoritative. They're the person that you trust and if they are invested in the benefits of it you will get patient investment. [SS01 (F, 38, I)]

I would definitely use something if it was given as advice to use it I would definitely use it, especially if they came from a trusted source like a doctor or a hospital or whatever, or even somewhere on a trusted source on the internet that you trusted like a charity or something you would go to. I would go to them. [SS32 (M, 62, C)]

I feel more assured if it's come from somebody professional straight out. If it's from the doctor you tend to believe them more or the nurse practitioners...rather than get the information off a friend or a neighbour. [SS34 (M, 68, C)]

If somebody that you trust, a medical professional that you trust is recommending a programme, I think you'd be more likely to do it than if you just kind of googled something. [SS35 (F, 43, C)]

Credible and Reliable Resources

The reliability of the information provided was considered to be essential. Participants discussed previous difficulties experienced in accessing reputable resources and expressed the need for a central, trusted source where all relevant information is available. A single access platform was perceived to save people from having to search and "dig" through the internet for appropriate content. Having credible resources that avoid speculation and eliminate the risk of harmful or inaccurate information was believed to be critical to supporting patient confidence and engagement in accessing and using digital health information.

It's completely reliable, so that's one of the things that I found really difficult about kidneys in particular...I found [it] really difficult to get information on and it's just difficult, like even the kidney charities, their information is not so great. So, in terms of a resource about information and treatment and progression of the disease, it's a reliable source that I can go back to that's stable. [SS01 (F, 38, I)]

I like to go to things that would be trusted, so I want to go to a trusted source not a speculative one. The British Heart Foundation, I would trust them, the

XSL•FO
RenderX

national kidney thing, your research thing, I would trust you, you know...I'm very sceptical of some American stuff and YouTube. [SS32 (M, 62, C)]

There is a lot of bullshit on the internet so you look for sources that are OK or that you can see instantly that consider to be golden source. Certain charities yes and certain websites absolutely not—I wouldn't believe a word they say even if it happened to be the truth, it's surrounded by verbiage and selling opportunities. [SS33 (M, 64, C)]

Continued Engagement and Sustainability

Integrating Into Care to Support Ongoing Management

Participants valued the holistic nature of MK&M, as it addressed not only physical health but also psychosocial well-being. To encourage sustained engagement, they felt that it should be closely integrated with their care and used to support the ongoing management of their CKD. Participants believed that health care professionals should adopt an individualized approach when promoting MK&M, signpost to relevant aspects of the program based on health status, and be appropriately followed up.

The thing that occurred to me was you can really tie this into people's care. So there are points at which I have to monitor my blood pressure on a weekly basis. So if there was an option that you could flag that says "I'm going to monitor my blood pressure" and it reminds you to do it, that would be really helpful...it would be potentially quite empowering. [SS01 (F, 38, I)]

I guess if something happened that I wasn't quite sure about or I thought there might be some information to support me on there with it or something I hadn't taken in or I want to revisit because can't quite remember what I learned from it in the first place I absolutely would think about going back into it. I don't feel it's I've ticked all the boxes and that's it, done, I don't see it like that, I see it as an additional support trying to work my way through this 'fun' situation! [SS02 (F, 66, I)]

Person-Based Interactional Approach

Participants highlighted the value of face-to-face interaction and how it can complement the digital element of MK&M. Human connection was believed to increase engagement, whether through health care professionals or peer support that fosters a sense of shared experience. Participants felt that it was reassuring to have someone to talk to rather than feeling like communication is solely with a machine. To create a more personal and interactive experience, participants suggested using live chats and Q&A sessions. These options were also seen as a way to build confidence in using the DHI. While general health and disease information could be delivered digitally, it was recommended that personalized information for an individual should be provided in person.

I think there's still a place obviously for face to face, just depends on what the situation is, but I think as far as overall help and education is then, yeah, I'm all for online stuff. If it's just general health education I prefer it online. For something specific to me then I would prefer face to face. [SS23 (F, 54, I)]

You've got to make them human and even things like live chats, Q&A live chats or this kind of stuff might be something, just a little bit more human, because otherwise you think you're talking to a machine, which you could be in many ways, you don't bother and you definitely don't interact and you probably don't get the correct answers anyhow...it needs to be slightly more human, because you're talking about people's lives, their bodies...we all rely too heavily on computers and forget we're all humans. [SS07 (M, 55, I)]

They've got to be backed by manned, where you can actually speak to someone because the problem with the unmanned ones is that they will give you an answer but what they don't know is if, the nuance isn't there. If I want to ask about a specific thing or worry about it, they can't do that. They're alright as an initial thing but I think that they've got to be backed up by an in-person contact, because otherwise you know you're not really going to be confident about it. [SS31 (M, 84, C)]

Use of Push Notifications and Professional Reinforcement

Participants appreciated the MK&M platform reminder functions but suggested increasing their frequency. They preferred reminders delivered as push notifications rather than solely via email. Additionally, they believed that health care professionals should play an active role by reinforcing the message and reminding patients about the program. It was felt that this would serve as a memory prompt to maintain engagement, particularly over the long term.

It's just that reminder isn't it, or even just an email to say have you finished the sessions, have you recorded anything this week, have you done this—just a memory job because once you've done the reading and you come out of it and busy you just—I just literally forgot about it. So, it I had an email dropping in saying 'have you thought about logging this or have you done this or visit the chat room', or something, I think it might have probably made me log back in again. [SS23 (F, 54, I)]

Like a monthly reminder email have you visited the thing or something that would help, because the world's full of good intentions and you try to do things and if you just got a reminder once a month, you know, pretty much like if you get an email...just to say have you checked out this topic. [SS27 (M, 62, I)]

Getting the boosters to remind you to do it I think! That probably is a good reminder because I think we all, sort of forget and you think oh I've done it now because I've read it, maybe need to go back and look at it again, so that is a good idea. [SS05 (F, 61, I)]

XSL•FO
RenderX

Motivated by Measurable Impact

Observing positive changes and experiencing benefits from engaging with MK&M were reported as strong motivators for continued participation. Participants highlighted the importance of committing time to improving health and making appropriate lifestyle changes. The MK&M DHI was valued for its ability to allow users to pick up where they left off, monitor progress, and track achievements over time. Features such as goal setting were seen as particularly motivating, as successfully reaching goals encouraged individuals to set and achieve new ones, reinforcing ongoing engagement.

> *You can see how far you've got, for a start. And it's nice to know you can carry on from where you've left off. You've got these bits on the dashboard which are quite good, it tells you how far you've got, and what have you...I can see some changes or something, so maybe when I get to the booster sessions I can set some goals and what have you, then maybe if I see some changes then I'll be more—yeah, let's do this, you know, more often.* [SS06 (F, 55, I)]

> *I think if you're interested, and I am, I'm always interested in health, not as an egocentric thing, just because we want to live the best that we can. I want to make sure I get the best out of life by making sure that I'm ready to face the challenges. So I think this looks really interesting. The more I look at it the more I see that there's there, so bring it on.* [SS13 (F, 57, I)]

Discussion

Principal Findings

This qualitative substudy explored SMILE-K trial participants' perspectives about the implementation and sustainability of the MK&M DHI to understand how patients would like MK&M to be delivered in clinical practice. The study findings provide a detailed account of factors patients felt were important to consider to support the implementation of MK&M into routine patient care to ensure that it meets real-world needs. This information offers critical context and is being used to inform the development and testing of implementation strategies, ensuring that MK&M is relevant, acceptable, and feasible in real-world practice.

The perceived most appropriate timing for MK&M to be provided was highly variable, with suggestions ranging from at diagnosis to a few weeks or several months later. This reflects the need for flexibility in delivery or prescription of MK&M and should be based on individual readiness rather than a particular time point in the patient's CKD journey. While there may not be a single teachable moment in the CKD journey, identifying a moment where individuals are more receptive to health messages, which may not necessarily correspond with a health event, will increase the likelihood of a resulting behavior change [21,22]. This underscores the importance of revisiting the offer multiple times throughout the early stages of the journey, as initial discussions may not always lead to uptake. Providing repeated exposure and opportunities can increase awareness and uptake of MK&M. It aligns with patients' desire for flexibility and autonomy, ensuring that they can engage with MK&M when they feel ready and motivated to act. Considering the COM-B (Capability-Opportunity-Motivation Behavior) model [23] to determine patients' readiness and receptivity to interventions, such as MK&M, may offer opportune intervention points, and capability or opportunity barriers may preclude behavior change that might be expected based on motivational shifts [24]. Understanding an individual's specific barriers and strengths can enable tailored strategies to increase acceptance and engagement, ultimately enhancing self-management.

Key barriers identified to accessing and engaging with MK&M included low health and digital literacy, as well as digital poverty. These barriers reflect broader challenges in CKD education, where low awareness of the condition and its risks [25], limited health [26,27] and digital literacy [28], and a lack of accessible educational materials for at-risk populations [29] are well-documented obstacles to effective patient engagement. Such challenges are not unique to CKD and similarly affect the management of other long-term conditions [30], including diabetes [31] and chronic obstructive pulmonary disease [32], and musculoskeletal conditions [33]. Clear communication from health care professionals, combined with patient reinforcement and targeted education, plays a critical role in improving health literacy and promoting effective self-management [34-36]. Delivering health messages in a clear, concise manner, using simple language and avoiding medical jargon, is essential to promote understanding, especially in those with limited health literacy [36]. Our previous work identified that while patients found the information provided in MK&M to be comprehensive, its presentation made it easy to access and engage with [13]. Despite positive feedback, it is clear from this study that further refinements are required to MK&M and the delivery of the education to ensure equitable access to all people living with early-stage CKD, including those from disadvantaged and underserved groups. Offering content in multiple formats (eg, print, video, and audio) and providing tailored support options can help overcome barriers faced by individuals experiencing digital poverty or low digital literacy. Future work should focus on enhancing DHIs, such as MK&M, to improve inclusion for individuals who are potentially disadvantaged yet stand to benefit most, with efforts aimed at addressing these gaps and ensuring equitable access.

Participants felt that interventions endorsed by trusted health care professionals and delivered with consistent, unified messaging are more likely to build patient confidence, encourage engagement, and improve adherence. This highlights the importance of health care professional involvement and consistent messaging in driving patient engagement, as interventions lacking such endorsement may struggle to achieve widespread adoption. Research examining patient portals found that most interventions were not delivered by a health care provider or an administrator [37], which may limit uptake of such tools by patients who may need encouragement and support to access these resources [38]. Health care providers may need to shift from acting as intermediaries, controlling and relaying information, to serving as apomediaries (ie, people who help users navigate information independently), guiding patients toward high-quality information and services [39]. MK&M

XSL•FO
RenderX

offers a potential solution to facilitate clinical care by providing comprehensive, evidence-based, and disease-specific information and supporting ongoing disease management. It can be integrated into clinical appointments and tailored to individual patient needs, which may enhance engagement and the health care professional–patient relationship [40]—a key factor in effective CKD self-management [7].

Integration of MK&M into clinical care and using it to support the ongoing CKD management was believed to be vital to encourage sustainability. The COVID-19 pandemic accelerated the uptake of digital tools and platforms within the National Health Service [41], leading to substantial transformations in service delivery [42]. Central to the National Health Service 10-year plan is the adoption of technology-enabled solutions, transforming traditional models into digitally connected, patient-centered care [43]. Digital health ecosystems have the potential to build long-term, collaborative relationships between patients and care teams that encourage and support maintenance of health-promoting behaviors [44]. However, in order for DHIs to deliver maximum impact, they must be effectively adopted and embedded into routine practice. The introduction of Integrated Care Systems and virtual care programs for CKD, such as LUCID [45], that are designed to optimize CKD management, provides a promising solution for integrating MK&M into routine care. Embedding MK&M within these models could strengthen patient education and awareness, supporting more comprehensive and coordinated care.

Assessing the readiness of health services to integrate these technologies is critical to identify gaps, allocate resources, and enable successful implementation and sustainability. Implementation models and frameworks can be used to systematically identify and map potential barriers and facilitators, link them to appropriate strategies, and guide the development of comprehensive implementation and evaluation plans. Jassemi and colleagues [38] illustrate this through the integration of the patient-focused eHealth intervention, My Kidneys My Health, into primary care and general nephrology kidney care. Key barriers identified by health care professionals included competing priorities and patient accessibility challenges (eg, internet access, technology use, and health literacy), while enablers included the ability to tailor the intervention to supplement education and alignment with core values such as accessibility, credibility, and usefulness [38]. These factors mirror those reported by patients in our study, particularly around digital literacy, engagement challenges, and the need for organizational support. This highlights the shared importance of usability, accessibility, and perceived value in successful implementation. At the same time, our data add patient-specific insights on autonomy, personalized communication, and uncertainties about how the DHI would fit within existing care pathways, and in some areas diverge from professional assumptions (eg, differing expectations about usability and required support). Together, these complementary and contrasting insights offer a more comprehensive understanding of factors shaping DHI implementation. Ensuring that interventions meet both clinical priorities and patient needs can increase the likelihood of successful adoption and sustained engagement. Future work should focus on co-designing implementation plans and readiness assessments to strengthen alignment of perspectives and enable scalable, sustainable implementation of DHIs to support CKD self-management.

Across long-term conditions, the value of DHIs is evident; however, their effectiveness and implementation challenges differ markedly between conditions and remain dependent on successful integration into routine care workflows. Successful implementation is consistently linked to patient-centered design and attention to individual, technological, and organizational factors that shape real-world use [46,47]. DHIs are most effective when they offer personalization, clear feedback, and simplicity, with human connection (eg, coaching or guided support) playing a key role in sustaining engagement over time [48]. Although DHIs can enhance self-management and social participation, their effectiveness differs by delivery mode and the degree to which they facilitate meaningful, tailored interaction [49]. Persistent barriers, including limited digital literacy, variable user confidence, and usability or access challenges, especially among older adults and people with multimorbidity [50], continue to hinder adoption. In addition, poor integration with existing workflows and clinical record systems constrains scalability and increases workload for staff, further limiting sustained implementation in routine practice [51]. Even structured implementation support may not overcome these challenges. A cluster randomized controlled trial of an embedding package for diabetes self-management education reported limited clinical improvement and difficulties normalizing new practices in primary care [52]. Primary care teams struggled to embed self-management education into routine practice, with low referral and attendance, no meaningful clinical improvement, and persistent difficulties normalizing new workflows, despite structured support. Competing clinical demands, entrenched routines, and limited prioritization of self-management contributed to poor uptake, alongside ongoing inequities in engagement across patient groups. These findings reinforce that implementing MK&M successfully will require not only robust technological functionality but also seamless integration into routine care and targeted strategies to address digital and health literacy, ensuring that the DHI remains relevant, acceptable, and feasible in real-world practice [53].

Strengths and Limitations

This study has several strengths and limitations. A key strength of this study was the use of qualitative methodology, which enabled us to obtain rich, in-depth insights into participants' experiences and perspectives, providing a nuanced understanding of factors that will affect the implementation of MK&M. The inclusion of participants from both the intervention group and the control group allows for exploration of differences across groups. Although participants had not used the DHI themselves, their organizational and experiential insights provide an important perspective on factors that shape adoption, readiness, and feasibility. These perspectives therefore make a meaningful contribution to understanding the wider implementation context within which the DHI will operate. It also offers context for perceived benefits and challenges and provides an insight into how implementation factors (eg, usability and accessibility) influence engagement. This approach supports patient-centered learning that can inform future design

XSL•FO
RenderX

and implementation strategies. Future work will build on these insights to test implementation strategies in real-world settings, ensuring scalability and sustainability. Implementation models and frameworks were not included in this exploratory study to enable the identification of context-specific issues that may not be fully captured by existing frameworks. However, the findings will inform future DHI implementation strategies with implementation plans to be supported and adapted using an appropriate framework.

A limitation of the study was the potential bias of participants recruited, which may limit the generalizability of the findings. While efforts were made to select a variety of participants, using maximum variation sampling, there was self-selection bias as some participants who were invited to take part did not respond to the invitation. Some subthemes, such as digital poverty, reflect the experiences of underrepresented groups who were largely absent from the sample, potentially limiting the depth of insight in these areas. Despite this, most participants still recognized and discussed issues relating to digital poverty, even if they had not experienced these challenges themselves. As the views represented in our study may not fully capture the experiences or barriers faced by underrepresented or digitally marginalized populations, the generalizability of these specific themes is constrained, and our findings should be interpreted with this context in mind.

Delivering the trial and intervention virtually created barriers for individuals with poor digital literacy and rendered those without digital access ineligible. As a result, valuable insights from participants who may face the greatest barriers to engagement were missed. Future research that actively incorporates the perspectives of disadvantaged populations is essential to ensuring that interventions and policy recommendations address the full spectrum of barriers and support greater equity in digital health engagement. Their perspectives could inform strategies to improve accessibility, usability, and equity. Thus, future research should include enhanced strategies to actively include digitally underserved populations through alternative recruitment and multiformat delivery, addressing digital poverty and low literacy to ensure inclusive, scalable, and sustainable implementation.

Participant experiences were collected within the context of a trial. Although the trial took a pragmatic approach and aimed to reflect aspects of routine practice, providing less structured support than traditional trials and allowing participants to use the MK&M DHI as they wished, elements of the trial environment may still have fostered more positive views of sustainability, feasibility, and integration than might be observed in everyday clinical settings. Trial features, such as clearer processes and greater predictability, can create a more enabling context than routine care, where competing demands and resource constraints are more prominent. As such, the influence of the trial context should be considered when interpreting the transferability of these findings to real-world implementation.

Conclusions

Our study highlights that timely, well-targeted communication using diverse strategies is critical for the successful uptake of MK&M among people living with CKD. The identification of perceived factors that will affect implementation of the program provides actionable insights to guide the development of tailored implementation strategies. These findings highlight the need to align MK&M with real-world contexts to ensure that it is relevant, acceptable, and feasible in routine practice. Consideration for the multiple contextual factors required, and the equity challenges, in integrating MK&M into routine kidney care, is essential for implementation to be successful. Embedding these principles into implementation planning and addressing key contextual factors can enhance the quality, effectiveness, and sustainability of kidney care interventions, such as MK&M, ultimately supporting patient care and improving outcomes.

Acknowledgments

This is a summary of independent research funded by the Stoneygate Trust and delivered through the National Institute for Health and Care Research (NIHR) Leicester Biomedical Research Centre (BRC) (NIHR203327). The views expressed are those of the authors and not necessarily those of the Stoneygate Trust, National Health Service (NHS), NIHR, or the Department of Health and Social Care. The authors would like to acknowledge and thank those involved in the development of My Kidneys & Me, the intervention tested in this trial, but were not involved in the management and interpretation of the trial itself: Michelle Hadjiconstantinou, Jonathan Barratt, Christopher Brough, James Burton, Jenny Hainsworth, Vicki Johnson, Maria Martinez, Victoria Pursey, Sally Schreder, Lucina Wilde, and Hannah Young. The authors would also like to thank the Leicester Diabetes Centre IMPACT (Innovative Management, Prevention And Care for long Term conditions) Team and Creative Team, specifically Charlie Franklin and Mike Bonar, for their invaluable input and contribution to the development of My Kidneys & Me. The authors would like to thank participating members of their patient and public steering group for their support in the design and development of My Kidneys & Me and the SMILE-K study. The authors are grateful to all the participants who took part in the study. They are thankful for the support provided by Lucy Byrne, Ella Ford, and Noemi Vadaszy during recruitment and data collection for this study. They are also grateful to the sites that recruited participants to the SMILE-K study: Barts Health NHS Trust; Dorset County Hospital NHS Foundation Trust; East and North Hertfordshire NHS Trust; East Kent Hospitals University NHS Foundation Trust; King’s College Hospital NHS Foundation Trust; Lancashire Teaching Hospitals NHS Trust; Liverpool University Hospitals NHS Foundation Trust; Manchester University NHS Foundation Trust; Mid & South Essex Hospital Services NHS Trust; Northampton General Hospital NHS Trust; Northern Care Alliance NHS Foundation Trust; Nottingham University Hospitals NHS Trust; Portsmouth Hospitals University NHS Trust; Royal Cornwall Hospitals NHS Trust; Royal Derby Hospitals NHS Trust; Royal Devon & Exeter NHS Foundation Trust; Sheffield Teaching Hospitals NHS Foundation Trust; St George's

XSL•FO
RenderX

University Hospitals NHS Foundation Trust; University Hospitals Coventry and Warwickshire NHS Trust; University Hospitals of Leicester NHS Trust; University Hospital Southampton NHS Foundation Trust; West Suffolk NHS Foundation Trust; York & Scarborough Teaching Hospital NHS Foundation Trust. The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing. Responsibility for the final manuscript lies entirely with the authors.

Funding

This study was funded by the Stoneygate Trust and the Leicester NIHR Biomedical Research Centre. All patient and public involvement and engagement (PPIE) activities were supported by 2 grants from the Leicester Kidney Care Appeal awarded to TJW and ACS. Funders had no input into the study protocol or interpretation of findings.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: CJL, TJW, TY, MJD, ACS
Methodology – data acquisition: CJL
Formal analysis and investigation: CJL, REB
Writing – original draft preparation: CJL
Writing – review and editing: CJL, REB, GKS, TJW, MJD, TY, ACS, MPMG-B
Funding acquisition: ACS
Supervision: ACS, MPMG-B
All authors have full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. All authors gave final approval.

Conflicts of Interest

MJD has acted as a consultant/advisor and speaker for Eli Lilly, Novo Nordisk, and Sanofi; as a consultant/advisor for Kailera; has attended advisory boards for AbbVie, Amgen, AstraZeneca, Biomea Fusion, Carmot/Roche, Daewoong Pharmaceutical, Sanofi, Zealand Pharma, Regeneron, GSK, Innovent Bio, and EktaH; and has acted as a speaker for AstraZeneca, Boehringer Ingelheim, Obesity Alliance, and Zuellig Pharma. She has received grants from AstraZeneca, Boehringer Ingelheim, and Novo Nordisk. All other authors declare no conflicts of interest.

Checklist 1
COREQ checklist.
[PDF File, 433 KB - jmir_v28i1e91966_app1.pdf]

References

1. Duff R, Awofala O, Arshad MT, et al. Global health inequalities of chronic kidney disease: a meta-analysis. Nephrol Dial Transplant 2024 Sep 27;39(10):1692-1709. [doi: 10.1093/ndt/gfae048] [Medline: 38389223]
2. Reducing the burden of noncommunicable diseases through promotion of kidney health and strengthening prevention and control of kidney disease. : World Health Organization; 2025 URL: https://apps.who.int/gb/ebwha/pdf_files/EB156/B156_CONF6-en.pdf [accessed 2026-04-10]
3. Foreman KJ, Marquez N, Dolgert A, et al. Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016-40 for 195 countries and territories. Lancet 2018 Nov 10;392(10159):2052-2090. [doi: 10.1016/S0140-6736(18)31694-5] [Medline: 30340847]
4. Webster AC, Nagler EV, Morton RL, Masson P. Chronic kidney disease. Lancet 2017 Mar 25;389(10075):1238-1252. [doi: 10.1016/S0140-6736(16)32064-5] [Medline: 27887750]
5. Kalantar-Zadeh K, Jafar TH, Nitsch D, Neuen BL, Perkovic V. Chronic kidney disease. Lancet 2021 Aug 28;398(10302):786-802. [doi: 10.1016/S0140-6736(21)00519-5] [Medline: 34175022]
6. Peng S, He J, Huang J, et al. Self-management interventions for chronic kidney disease: a systematic review and meta-analysis. BMC Nephrol 2019 Apr 26;20(1):142. [doi: 10.1186/s12882-019-1309-y] [Medline: 31027481]
7. Lightfoot CJ, Nair D, Bennett PN, et al. Patient activation: the cornerstone of effective self-management in chronic kidney disease? Kidney Dial 2022 Mar;2(1):91-105. [doi: 10.3390/kidneydial2010012] [Medline: 37101653]
8. Levey AS, Coresh J. Chronic kidney disease. Lancet 2012 Jan 14;379(9811):165-180. [doi: 10.1016/S0140-6736(11)60178-5] [Medline: 21840587]

XSL•FO
RenderX

9. Lightfoot CJ, Wilkinson TJ, Hadjiconstantinou M, et al. The codevelopment of “My Kidneys & Me”: a digital self-management program for people with chronic kidney disease. J Med Internet Res 2022 Nov 14;24(11):e39657. [doi: 10.2196/39657] [Medline: 36374538]
10. Hadjiconstantinou M, Schreder S, Brough C, et al. Using intervention mapping to develop a digital self-management program for people with type 2 diabetes: tutorial on MyDESMOND. J Med Internet Res 2020;22(5):e17316. [doi: 10.2196/17316]
11. Lightfoot CJ, Wilkinson TJ, Yates T, Davies MJ, Smith AC. “Self-Management Intervention through Lifestyle Education for Kidney health” (the SMILE-K study): protocol for a single-blind longitudinal randomised controlled trial with nested pilot study. BMJ Open 2022 Nov 16;12(11):e064916. [doi: 10.1136/bmjopen-2022-064916] [Medline: 36385018]
12. Lightfoot CJ, Wilkinson TJ, Sohansoha GK, et al. The effects of a digital health intervention on patient activation in chronic kidney disease. NPJ Digit Med 2024 Nov 12;7(1):318. [doi: 10.1038/s41746-024-01296-1] [Medline: 39533053]
13. Lightfoot CJ, Wilkinson TJ, Billany RE, et al. Use, usability, and experience testing of a digital health intervention to support chronic kidney disease self-management: mixed methods study. J Med Internet Res 2025 Jun 19;27:e75845. [doi: 10.2196/75845] [Medline: 40537087]
14. Curtis K, Fry M, Shaban RZ, Considine J. Translating research findings to clinical nursing practice. J Clin Nurs 2017 Mar;26(5-6):862-872. [doi: 10.1111/jocn.13586]
15. Evans-Lacko S, Jarrett M, McCrone P, Thornicroft G. Facilitators and barriers to implementing clinical care pathways. BMC Health Serv Res 2010 Jun 28;10(1):182. [doi: 10.1186/1472-6963-10-182] [Medline: 20584273]
16. Chow JSF, Gonzalez-Arce VE, Tam CWM, Warner K, Maurya N, Mcdougall A. Creating a successful health pathway to support the integration of patient care. JICA 2019 Dec 20;28(2):171-182. [doi: 10.1108/JICA-08-2019-0039]
17. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
18. Lightfoot CJ, Wilkinson TJ, Vadaszy N, et al. Improving self-management behaviour through a digital lifestyle intervention: an internal pilot study. J Ren Care 2024 Sep;50(3):283-296. [doi: 10.1111/jorc.12488] [Medline: 38296833]
19. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
20. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Aug 8;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
21. McBride CM, Emmons KM, Lipkus IM. Understanding the potential of teachable moments: the case of smoking cessation. Health Educ Res 2003 Apr;18(2):156-170. [doi: 10.1093/her/18.2.156] [Medline: 12729175]
22. Lawson PJ, Flocke SA. Teachable moments for health behavior change: a concept analysis. Patient Educ Couns 2009 Jul;76(1):25-30. [doi: 10.1016/j.pec.2008.11.002] [Medline: 19110395]
23. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6(1):42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
24. Olander EK, Darwin ZJ, Atkinson L, Smith DM, Gardner B. Beyond the “teachable moment” - a conceptual analysis of women’s perinatal behaviour change. Women Birth 2016 Jun;29(3):e67-e71. [doi: 10.1016/j.wombi.2015.11.005] [Medline: 26626592]
25. Narva AS, Norton JM, Boulware LE. Educating patients about CKD: the path to self-management and patient-centered care. Clin J Am Soc Nephrol 2016 Apr 7;11(4):694-703. [doi: 10.2215/CJN.07680715] [Medline: 26536899]
26. Fraser SDS, Roderick PJ, Casey M, Taal MW, Yuen HM, Nutbeam D. Prevalence and associations of limited health literacy in chronic kidney disease: a systematic review. Nephrol Dial Transplant 2013 Jan;28(1):129-137. [doi: 10.1093/ndt/gfs371] [Medline: 23222414]
27. Abdel-Kader K, Dew MA, Bhatnagar M, et al. Numeracy skills in CKD: correlates and outcomes. Clin J Am Soc Nephrol 2010 Sep;5(9):1566-1573. [doi: 10.2215/CJN.08121109] [Medline: 20507954]
28. Bonner A, Gillespie K, Campbell KL, et al. Evaluating the prevalence and opportunity for technology use in chronic kidney disease patients: a cross-sectional study. BMC Nephrol 2018 Feb 2;19(1):28. [doi: 10.1186/s12882-018-0830-8] [Medline: 29394930]
29. Tuot DS, Davis E, Velasquez A, Banerjee T, Powe NR. Assessment of printed patient-educational materials for chronic kidney disease. Am J Nephrol 2013;38(3):184-194. [doi: 10.1159/000354314] [Medline: 23970127]
30. Pong C, Tseng R, Tham YC, Lum E. Current implementation of digital health in chronic disease management: scoping review. J Med Internet Res 2024 Dec 12;26:e53576. [doi: 10.2196/53576] [Medline: 39666972]
31. Lee JY, Chan CKY, Chua SS, et al. Using telemedicine to support care for people with type 2 diabetes mellitus: a qualitative analysis of patients’ perspectives. BMJ Open 2019 Oct;9(10):e026575. [doi: 10.1136/bmjopen-2018-026575]
32. Ramachandran HJ, Oh JL, Cheong YK, et al. Barriers and facilitators to the adoption of digital health interventions for COPD management: a scoping review. Heart Lung 2023;59:117-127. [doi: 10.1016/j.hrtlng.2023.02.004] [Medline: 36801546]

XSL•FO
RenderX

33. van Tilburg ML, Spin I, Pisters MF, et al. Barriers and facilitators to the implementation of digital health services for people with musculoskeletal conditions in the primary health care setting: systematic review. J Med Internet Res 2024 Aug 27;26:e49868. [doi: 10.2196/49868] [Medline: 39190440]
34. Shah JM, Ramsbotham J, Seib C, Muir R, Bonner A. A scoping review of the role of health literacy in chronic kidney disease self-management. J Ren Care 2021 Dec;47(4):221-233. [doi: 10.1111/jorc.12364] [Medline: 33533199]
35. Langham RG, Kalantar-Zadeh K, Bonner A, et al. Kidney health for all: bridging the gap in kidney health education and literacy. Nephrol Dial Transplant 2022 Mar 25;37(4):605-612. [doi: 10.1093/ndt/gfac038] [Medline: 35142861]
36. Jagodage H, McGuire A, Seib C, Bonner A. Effectiveness of teach‐back for chronic kidney disease patient education: a systematic review. J Ren Care 2024 Jun;50(2):92-103. [doi: 10.1111/jorc.12462]
37. Han HR, Gleason KT, Sun CA, et al. Using patient portals to improve patient outcomes: systematic review. JMIR Hum Factors 2019 Dec 19;6(4):e15038. [doi: 10.2196/15038] [Medline: 31855187]
38. Jassemi S, Sparkes D, Delgado M, et al. Implementation and evaluation of a patient-focused eHealth intervention, My Kidneys My Health, in primary care and general nephrology clinics: multimethods study. J Med Internet Res 2025 Aug 29;27:e71832. [doi: 10.2196/71832] [Medline: 40882216]
39. Mesko B, Győrffy Z. The rise of the empowered physician in the digital health era: viewpoint. J Med Internet Res 2019 Mar 26;21(3):e12490. [doi: 10.2196/12490] [Medline: 30912758]
40. Carini E, Villani L, Pezzullo AM, et al. The impact of digital patient portals on health outcomes, system efficiency, and patient attitudes: updated systematic literature review. J Med Internet Res 2021 Sep 8;23(9):e26189. [doi: 10.2196/26189] [Medline: 34494966]
41. Peek N, Sujan M, Scott P. Digital health and care in pandemic times: impact of COVID-19. BMJ Health Care Inform 2020 Jun;27(1):e100166. [doi: 10.1136/bmjhci-2020-100166]
42. Hutchings R. The impact of Covid-19 on the use of digital technology in the NHS. Nuffield Trust. 2020. URL: https://www.nuffieldtrust.org.uk/files/2020-08/the-impact-of-covid-19-on-the-use-of-digital-technology-in-the-nhs-web-2.pdf [accessed 2026-04-10]
43. Fit for the future: 10 year health plan for England. NHS England. 2025. URL: https://www.gov.uk/government/publications/10-year-health-plan-for-england-fit-for-the-future [accessed 2026-04-10]
44. Abernethy A, Adams L, Barrett M, et al. The Promise of Digital Health: Then, Now, and the Future. NAM Perspect 2022;2022. [doi: 10.31478/202206e] [Medline: 36177208]
45. Major RW, Lakhani N, Ahmed Y, et al. Integrated primary and secondary care optimizes the management of people with CKD-the LUCID project. Clin Kidney J 2025 Apr;18(4):sfaf049. [doi: 10.1093/ckj/sfaf049] [Medline: 40207101]
46. Moore L, Britten N, Lydahl D, Naldemirci Ö, Elam M, Wolf A. Barriers and facilitators to the implementation of person-centred care in different healthcare contexts. Scand J Caring Sci 2017 Dec;31(4):662-673. [doi: 10.1111/scs.12376] [Medline: 27859459]
47. Erku D, Khatri R, Endalamaw A, et al. Digital health interventions to improve access to and quality of primary health care services: a scoping review. Int J Environ Res Public Health 2023 Sep 28;20(19):6854. [doi: 10.3390/ijerph20196854] [Medline: 37835125]
48. Christie R, Sadler E, Sait M, et al. Digital empowerment in long-term condition management: a systematic review and narrative synthesis of the experiences and perceptions of home-based digital health coaching interventions. Digit Health 2024;10:20552076241302230. [doi: 10.1177/20552076241302230] [Medline: 39669392]
49. Deng H, Vu KQ, Franco JR, et al. Digital interventions for social participation in adults with long-term physical conditions: a systematic review. J Med Syst 2023;47(1):26. [doi: 10.1007/s10916-023-01914-7]
50. Yin R, Martinengo L, Smith HE, Subramaniam M, Griva K, Tudor Car L. The views and experiences of older adults regarding digital mental health interventions: a systematic review of qualitative studies. Lancet Healthy Longev 2024 Nov;5(11):100638. [doi: 10.1016/j.lanhl.2024.08.007] [Medline: 39491881]
51. Feenstra M, Zuidema SU, Janus SIM. Long-term care needs guidance for the implementation of digital health technologies. Age Ageing 2023 Dec 1;52(12):afad223. [doi: 10.1093/ageing/afad223] [Medline: 38037227]
52. Davies MJ, Bodicoat DH, Brennan A, et al. Uptake of self-management education programmes for people with type 2 diabetes in primary care through the embedding package: a cluster randomised control trial and ethnographic study. BMC Prim Care 2024 Apr 25;25(1):136. [doi: 10.1186/s12875-024-02372-x] [Medline: 38664727]
53. Al Mahmud A, Joachim S, Jayaraman PP, et al. Digital health interventions to support chronic disease management: systematic scoping review. JMIR Mhealth Uhealth 2026 Jan 14;14:e63742. [doi: 10.2196/63742] [Medline: 41533959]

Abbreviations

COM-B: Capability-Opportunity-Motivation Behavior
DHI: digital health intervention
eGFR: estimated glomerular filtration rate
GP: general practitioner
MK&M: My Kidneys & Me

XSL•FO
RenderX

Edited by A Stone; submitted 22.Jan.2026; peer-reviewed by FJ Yang, S Jassemi; revised version received 13.Apr.2026; accepted 13.Apr.2026; published 30.Jun.2026.

Please cite as:
Lightfoot CJ, Billany RE, Sohansoha GK, Wilkinson TJ, Davies MJ, Yates T, Smith AC, Graham-Brown MPM
Moving Research to Practice: A Qualitative Study Exploring Patient Perspectives of the Implementation and Sustainability of a Digital Health Intervention for Chronic Kidney Disease
J Med Internet Res 2026;28:e91966
URL: https://www.jmir.org/2026/1/e91966
doi:10.2196/91966

© Courtney J Lightfoot, Roseanne E Billany, Gurneet K Sohansoha, Thomas J Wilkinson, Melanie J Davies, Thomas Yates, Alice C Smith, Matthew PM Graham-Brown. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Health Intervention to Promote Lifelong Specialized Care in Adults With Congenital Heart Disease: Theory-Driven Community Co-Designed Study

Anushree Agarwal[1], MAS, MD; Joseph Valente[2,3], BS, BCPA; Karina Buenrostro[3], BA; Katelyn Macholl[3], BS; Juhi Mehta[3], BS; Keerthana Reddy[4], MD; Karina Manayan[3], NP; Parang Kim[3], MA; Aleah Sparks[5], NP; Kunyi Li[3], MS, CGC; Pranav Ahuja[3], BS; Kevin Sun[3], MS; Kimberly Payton[6], EdD; Mark D Norris[7], MS, MD; Katia Bravo-Jaimes[8], MD; Leigh Reardon[9], MD; Philip Moons[10,11,12], RN, PhD; Megumi J Okumura[13], MAS, MD; Gregory M Marcus[3], MAS, MD; Michelle Gurvitz[14], MD

[1]Adult Congenital Heart Disease Section, Division of Cardiology, Associate Professor of Medicine, University of California, San Francisco, 500 Parnassus Avenue, M-1177B, Box 0124, San Francisco, United States

[10]KU Leuven Department of Public Health and Primary Care, KU Leuven, Leuven, Belgium

[11]Sahlgrenska Academy, University of Gothenburg Centre for Person-Centered Care (GPCC), University of Gothenburg, Gothenburg, Sweden

[12]Department of Pediatrics and Child Health, University of Cape Town, Cape Town, South Africa

[13]General Internal Medicine and PRL Institute for Health Policy Studies, Departments of Pediatrics and Internal Medicine, Division of General Pediatrics, University of California, San Francisco, San Francisco, United States

[14]Department of Cardiology, Boston Children's Hospital, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States

[2]Team Uncle Joe, Katy, TX, United States

[3]Department of Medicine, Division of Cardiology, University of California, San Francisco, San Francisco, CA, United States

[4]Division of Pediatric Cardiology, University of Miami, Miami, FL, United States

[5]Golden Gate Regional Center, San Francisco, CA, United States

[6]Senior Patient Advocate and 1st Vice East County NAACP, Parent of an adult with congenital heart disease, East County NAACP, Pittsburg, CA, United States

[7]Department of Pediatrics, Division of Cardiology, University of Michigan, Ann Arbor, United States

[8]Department of Cardiovascular Medicine, Mayo Clinic in Florida, Jacksonville, United States

[9]Division of Pediatric Cardiology, University of California, Los Angeles, Los Angeles, United States

Corresponding Author:
Anushree Agarwal, MAS, MD
Adult Congenital Heart Disease Section, Division of Cardiology, Associate Professor of Medicine, University of California, San Francisco, 500 Parnassus Avenue, M-1177B, Box 0124, San Francisco, United States

Abstract

Background: Adults with congenital heart disease frequently experience gaps in lifelong adult congenital heart disease (ACHD) specialty care, leading to preventable complications, hospitalizations, and premature mortality. However, effective, scalable, accessible, and sustainable strategies to reduce these gaps are lacking. Digital health offers potential solutions but requires a rigorous scientific approach in its design to address the needs of the target population.

Objective: This study aims to describe the development of a theory-based, community co-designed digital health intervention to improve lifelong ACHD care.

Methods: We integrated theory-based behavioral frameworks, semistructured qualitative interviews with patients and clinicians, and a community-engaged approach to develop a digital health intervention for ACHD. The primary behavioral target was completing an ACHD specialist appointment. We conducted Capability, Opportunity, Motivation, and Behavior (COM-B)–guided semistructured interviews with patients with ACHD and clinicians to identify barriers to specialized care amenable to a digital intervention, and patient-centered goals for the digital tool. The community partners helped develop key intervention objectives, create a theory-driven framework, and specify how each intervention component targets specific COM-B barriers.

Results: We interviewed 54 participants (n=37 patients with ACHD, and n=17 clinicians) and engaged 21 community partners representing 4 advocacy organizations. Design objectives emphasized addressing patient loneliness, ensuring accessibility and credibility, and enabling scalability while centering patient perspectives. Participants identified 4 priorities: providing credible resources, uplifting patient voices, customizing to patient needs, and centering positivity and joy. The digital tool, named by community partners as Empower My Congenital Heart (EMCH), was designed within the web- and mobile-based, Apple- and

XSL•FO
RenderX

Android-compatible, Eureka Digital Research platform (University of California, San Francisco). Key intervention components included educational modules, peer support, appointment planning nudges, and a digital medical passport. The EMCH's theory-driven framework specifies how each intervention component targets specific COM-B barriers to specialized ACHD care.

Conclusions: The theory-driven, community co-designed EMCH digital tool provides a scalable approach to promote lifelong ACHD specialist care. Ongoing process evaluation and a planned randomized controlled trial will assess acceptability, engagement, and effectiveness in reducing care gaps. If proven effective, EMCH has the potential to prevent complications, emergency hospitalizations, and mortality affecting the majority of adults with congenital heart disease.

Trial Registration: ClinicalTrials.gov NCT06581484; https://clinicaltrials.gov/study/NCT06581484

(J Med Internet Res 2026;28:e75867) doi:10.2196/75867

KEYWORDS

congenital heart defects; adult; digital health; mobile apps; health behavior; patient-centered care; community-based participatory research; implementation science

Introduction

Adult congenital heart disease (ACHD) is rapidly growing in the adult population [1,2]. The American Heart Association and American College of Cardiology guidelines provide recommendations on the frequency and intervals at which patients with ACHD should receive lifelong care from clinicians specialized in managing ACHD [3,4]. Despite this, up to 85% of patients with ACHD experience gaps in receiving specialized care throughout their adult lives [5-8]. Those with care gaps are at a higher risk of poor outcomes, including emergent admissions, need for urgent cardiac procedures, and mortality [9-11]. Thus, there is a need to identify accessible, scalable, and sustainable strategies to reduce care gaps and improve outcomes for patients with ACHD.

Individual-, provider-, and system-level barriers contribute to gaps in lifelong ACHD care [5,12,13]. Interventions, such as education and support for navigating the health system, can reduce these gaps [14-16]. However, existing interventions are clinic-based (nurse-led education, transition clinics); limited to single-center studies of 16 - to 21-year-olds focused on transfer from pediatric to ACHD care; and are resource-intensive, inconsistently implemented, and constrained by reimbursement, clinic time, and missed visits [17-19].

Web- and mobile-based solutions offer a scalable way to overcome these barriers by engaging patients outside of clinic visits [20,21]. Although 94% of patients with ACHD use smartphones and most report openness to app-based solutions [22], existing congenital heart disease (CHD) digital interventions focus on care transition readiness in adolescents and young adults [23] or general patient monitoring and education [24]. None systematically address the establishment and maintenance of lifelong ACHD care across the full adult lifespan using theory-driven behavior change frameworks. Effective and scalable strategies to support all adults (≥18 y), including those who did not transfer from pediatrics or were later lost from care, in establishing and maintaining ACHD care remain unknown.

App-based interventions grounded in evidence-based theoretical frameworks [21,25] are more robust, adaptable, and effective in achieving lasting behavior change in real-world settings, not just in ideal research environments [26,27]. Designing interventions to change behaviors also requires an understanding of the perspectives and psychosocial contexts of the people who will use them—a concept fundamental to community-engaged research [28-30]. Combining theory- and community-driven approaches is vital to ensure that interventions are usable, acceptable, and move beyond *what* works to *how* to make it work in diverse, complex environments [27,31].

Thus, the aim of this paper is to describe a theory- and community-driven approach to designing a scalable digital health intervention to promote ACHD care. The methods and findings can inform others developing interventions to improve outcomes for ACHD and other chronic, lifelong conditions.

Methods

Intervention Development Process

Overview

We began by establishing guiding principles for intervention design (Textbox 1). We then integrated behavioral frameworks with semistructured interviews and community-engaged research to identify and refine intervention components for promoting ACHD care (Figure 1).

XSL•FO
RenderX

Textbox 1. Guiding philosophy for the design of an adult congenital heart disease (ACHD) digital health intervention.

Characteristics designed to drive evidence-based approach

- Explicitly evidence-based, using scientific rationale for behavioral theories
- Expert guidance for content creation (eg, adult congenital heart disease follow-up guidelines)
- Person- and theory-informed approach at all stages of design and evaluation to meet individual needs of the target adult congenital heart disease population
- Noncommercial, developed by the named team of medical and behavior change experts

Characteristics designed to encourage uptake and long-term engagement

- Nimble, user-friendly interface with layperson-oriented messaging, leveraging existing tested capabilities and proven methods of the digital research platform
- Simplicity, with direct, concise, and actionable messaging designed to foster participants' capabilities, opportunities, and motivation
- Proactive, repetitive content delivery using nudge principles rather than relying on patients to seek out information, enhancing its reach and impact
- Mechanisms for modeling and sharing experiences within the user community, aimed at boosting self-efficacy and promoting positive affect among participants

Figure 1. Overview of the digital health intervention planning and development process. Behavioral frameworks (Behavior Change Wheel, Digital Health Equity, Diffusion of Innovation, and Behavioral Economics) were integrated with semistructured interviews and an iterative design process with the community partners. This process optimized the development of a digital health intervention with key objective to empower patients in navigating their congenital heart disease and promote lifelong specialized care. ACHD: adult congenital heart disease.

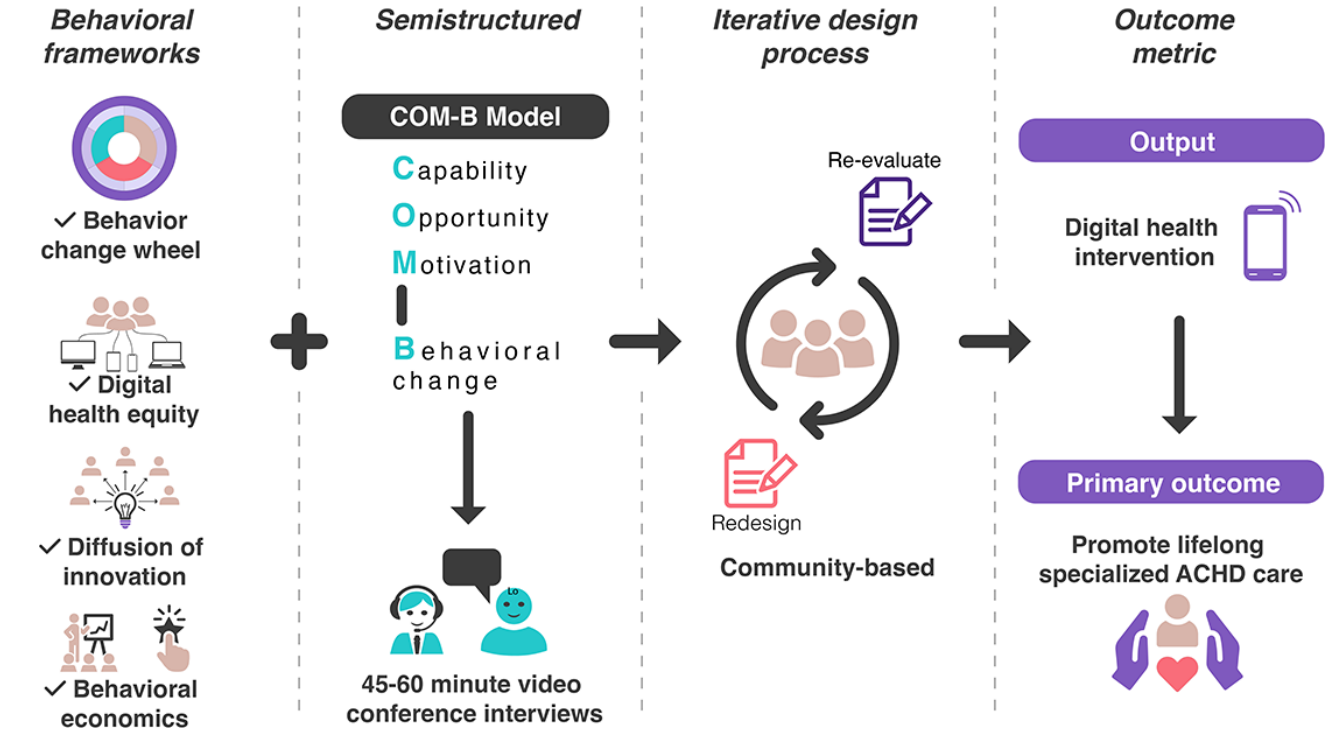

Behavioral Frameworks

The study team chose 4 evidence-based behavioral frameworks that complemented each other. First, the Behavior Change Wheel (BCW) theoretical model was used to specify the target behavior, identify the barriers and enablers of the behavior, and select the means by which the intervention can change behaviors [32,33]. To make the intervention equitable and relevant to diverse patients with ACHD, we incorporated the principles of Digital Health Equity, Diffusion of Innovation, and Behavioral Economics (Table 1) [34-36].

The BCW framework is based on multiple models of health behavior and has been shown to be effective in various cardiac and noncardiac chronic conditions [37-40] but has not yet been used in ACHD. The core driver in BCW (or hub of the wheel) is the Capability, Opportunity, Motivation, and Behavior (COM-B) model, which consists of three necessary conditions for a given "Behavior" to occur: (1) "Capability" (psychological or physical), (2) "Opportunity" (physical or social), and (3) "Motivation" (reflective or automatic). The BCW framework then supports the selection of intervention functions and policy categories. Intervention function refers to broad categories of ways an intervention can change a behavior. The 9 intervention

XSL•FO
RenderX

functions described in BCW include education, persuasion, incentivization, coercion, training, restrictions, environmental restructuring, modeling, and enablement. Policy categories comprise the final outer layer, or the wheel's rim, and help identify the types of policy categories one may wish to consider to further influence the drivers of behaviors (COM-B).

Table . Theoretical frameworks for designing ACHD[a] digital intervention.

Frameworks	Summary of the theoretical frameworks
Behavioral Change Wheel [32]	• Define health problem in behavioral terms (eg, scheduling and completing ACHD specialist visit) • Understand determinants of behavior (Capability, Opportunity, Motivation for Behavior Change model) • Link determinants of behavior to identify domains most likely to influence behavior • Select intervention functions that are means by which an intervention changes behavior (ie, education, enablement, persuasion, or training)
Digital Health Equity [34]	Apply the following concepts: • Health literacy: include relevant content: assume no background knowledge of participants, avoid cognitive burden from too much information, and use simple numbers and percentages to develop content • Readability: use plain and clear language tested with target population • Ease of use: accessible on various platforms, format conducive to comprehension; content appeals to users of different identities and backgrounds
Diffusion of Innovation [35]	• Engaging all levels of adopters (highly engaged and less engaged patients with CHD[b]) and CHD champions in the design, adoption, and dissemination of the intervention
Behavioral Economics [36]	• Nudges and default settings of the digital tool alter behavior and facilitate decision-making process

[a]ACHD: adult congenital heart disease.

[b]CHD: congenital heart disease.

Semistructured Interviews

The semistructured interviews for this study were part of a larger qualitative study that was conducted to identify barriers and enablers for lifelong specialized ACHD care, described in detail separately [12]. For this study, we used COM-B as a guide to identify barriers amenable to a digital intervention and the patient-centered goals for the digital tool. Briefly, we recruited patients with ACHD (≥18 y) who spoke English or Spanish and could provide informed consent. Initially, patients were recruited from University of California, San Francisco (UCSF) using chart review and purposefully sampled [41] to recruit those with more than 3-year gaps in ACHD specialist care during their adulthood. Once we achieved thematic saturation [42] at UCSF, we recruited non-UCSF patients with convenience and snowball sampling [41] to have representation from all regions of the United States (Northeast, Southeast, Midwest, Southwest, and West). All patients underwent an informed consent process and completed a demographic questionnaire at the end of their interview. Clinicians included adult congenital cardiology and pediatric cardiology nurses, doctors, coordinators, board-certified patient advocates, and program leaders. Clinicians were recruited from UCSF and outside UCSF using a similar approach of snowballing and thematic saturation to have representation from all regions. Separate interview guides were developed for patients and clinicians using content domains within the COM-B framework and to explore opinions toward a digital tool and its key features (Tables S1 and S2 in Multimedia Appendix 1). Interviews were conducted over Zoom, each lasting 45 to 60 minutes, audio-recorded, and professionally transcribed.

Community-Engaged Approach

The goal of the research team in recruiting community advisory board (CAB) partners was to represent diverse perspectives in ACHD care. We used convenience sampling to recruit patients from UCSF and snowball sampling to recruit patients outside UCSF, as well as clinicians and advocacy representatives. CAB partners were involved in all aspects of intervention planning, design, and implementation, including recruitment, reviewing interview findings, finalizing tool features and branding, sharing personal experiences, and iterative prototype testing. One patient partner (JV) led the creation of a design guide (Table S3 in Multimedia Appendix 1) to ensure consistency and uniformity in materials, built the study website, and developed strategies to elevate participant experience. CAB members also contributed as coauthors in scientific and community writing and dissemination. CAB members provided ongoing input through monthly hour-long meetings. Meeting discussions were documented in study logs, iteratively integrated with interview data, and analyzed by the research team to identify key themes and design recommendations.

XSL•FO
RenderX

Data Analysis

We conducted thematic analysis of all semistructured interviews using an inductive approach that was subsequently mapped to the COM-B framework. Codes emerged organically from participant narratives and were then organized according to COM-B domains (capability, opportunity, motivation). The coding team consisted of AA (ACHD cardiologist with qualitative research training), MJO (pediatric-to-adult transition specialist and qualitative researcher), 1 medical student (K Macholl), and 3 research assistants (KB, JM, PA). All team members completed structured training in the COM-B or TDF framework application and qualitative coding principles, including group practice coding of pilot transcripts to establish shared code definitions. The multicoder consensus process, overseen by AA and MJO, ensured coding rigor while providing methodological training for less experienced team members. Two team members independently coded each transcript in Microsoft Word to identify themes and subthemes related to barriers and facilitators to ACHD care and toward key aspects of the digital tool. After independent coding, coders (AA, K Macholl, KB, JM, PA) met in iterative group sessions to compare coded transcripts and discuss discrepancies. We prioritized resolving discrepancies through consensus discussion among the coding team, rather than quantitative intercoder reliability metrics. The finalized codes were compiled into a single coded transcript for each interview. We then created a structured data matrix using rapid qualitative analysis [43], an action-oriented method for qualitative data analysis conducted in Microsoft Excel, to efficiently produce results for real-world interventions. Coded data from finalized transcripts were transferred into the matrix, which organized codes across participants, allowing systematic cross-case comparison. Sample codebook entries showing how themes were mapped to COM-B and an excerpt of the rapid qualitative analysis matrix structure are provided in Tables S4 and S5 respectively in Multimedia Appendix 1. We summarized the matrix findings via iterative group meetings (AA, K Macholl, KB, JM, PA, MJO). Through ongoing, iterative reviews of CAB meeting discussions documented in study logs, we informed, planned, and refined the intervention components concurrent with their development.

Ethical Considerations

Ethical approval for this study was obtained from the UCSF institutional review board (IRB number 22‑36667). Verbal informed consent was obtained from interview participants. The National Clinical Trial number for this study is NCT06581484. All interview participants and CAB members were offered US $50 per hour in the form of an electronic gift card to compensate them for their time.

Results

Study Population

We interviewed 54 participants (n=37 patients and n=17 clinicians) (Table 2) and partnered with 21 CAB members. The CAB comprised 11 patients or family members who are racially diverse and from all regions of the United States, 6 pediatric or ACHD physicians, 1 social worker, 1 ACHD nurse, 2 nurse practitioners, and 6 advocacy representatives, some of whom represent more than 1 role. Advocacy organizations included the Adult Congenital Heart Association, Conquering CHD, Mended Hearts, and Team Uncle Joe [44-47].

XSL•FO
RenderX

Table . Characteristics of study participants.

Characteristic	Patients (n=37)	Clinicians (n=17)
Age (y), median (IQR)	32 (18 - 65)	—[a]
Female, n (%)	21 (57)	—
Race or ethnicity, n (%)		
Asian	8 (21)	—
Black or African American	6 (15)	—
Hispanic or Latino	7 (18)	—
White	13 (35)	—
Other or multiple	3 (8)	—
Education: less than or equal to some college, n (%)	14 (41)	—
Care gap ≥3 years, n (%)	13 (35)	—
Recruitment source, n (%)		
UCSF[b]	27 (73)	9 (52)
Community or snowball	10 (27)	8 (48)
Clinician roles, n (%)		
ACHD[c] or pediatric cardiologists	—	8 (47)
Nurses or NPs[d]	—	6 (35)
Other (coordinators, advocates)	—	3 (18)

[a]Not applicable.

[b]UCSF: University of California San Francisco.

[c]ACHD: adult congenital heart disease.

[d]NP: nurse practitioner.

CHD-Related Unique Challenges to Inform the Digital Intervention Design and Features

The challenges uniquely faced by patients with ACHD centered around users' perspectives, loneliness, feasibility, accessibility, credibility, and scalability (Table 3). They then helped identify design objectives and intervention features that could address these challenges. The interview participants also identified four patient-centered goals to consider during the design of the digital tool (Table S6 in Multimedia Appendix 1): (1) easy access to credible resources (How can we make it easier to connect patients with pre-existing resources?); (2) uplifting of patient voices (How can we uplift patient voices to create a personal connection?); (3) customization to patient needs (How can we cater to different needs within the community?); and (4) centering positivity and joy (How can we bring positivity and joy into our tool?).

XSL•FO
RenderX

Table . Digital tool design objectives and features for adult congenital heart disease (ACHD) specialized care.

Key challenge	Design objective	Key features	Evidence base
Perspectives: Patients may not understand the need for specialist care when feeling well; prefer to avoid thinking about their heart	Promote well-being rather than illness management	Empowering tone positioning patients as active participants; simple, concise interventions (eg, clarifying defect vs disease); user-driven engagement with personally relevant content; building motivation from first contact	Interviews + CAB[a]
Loneliness: Patients feel isolated; unsure how their condition compares to others	Build community	Peer empowerment quotes providing practical guidance; user story sharing; links to community events and peer connections	Interviews + CAB
Feasibility: Risk for the intervention to be overly complex given diverse patient needs and resource constraints	Efficient, multiphase design	Phased development targeting key care gap drivers; features addressing multiple outcomes simultaneously; balance between broad applicability and personalized relevance; continuous adaptation based on user feedback	CAB
Accessibility: Young adults with competing priorities need quick, convenient access	Enable easy, timely, nonintrusive access	Brief, actionable content (few minutes to complete); mobile-optimized display; automated nudge delivery via email, texting, or push notifications; small information bursts over time	Interviews + CAB
Credibility: Abundance of information creates overwhelm; unclear what sources are reliable	Use credible, curated sources	Partnership with advocacy organizations for content curation; easy-to-understand formatting; iterative tailoring based on user feedback	Interviews + CAB
Scalability: Variability in age, health literacy, and prior engagement creates diverse needs	Design for diverse needs	Cross-platform web or mobile access (Android or iOS); low-bandwidth content; materials addressing both activated and nonactivated patients; anonymous story sharing; community building for current and future generations	CAB, guided by interviews

[a]CAB: community advisory board.

Theory-Driven Intervention Components

Using the BCW framework, we mapped barriers to specialized care (identified through interviews and CAB discussions) to behavioral targets and intervention functions (Table 4). This systematic approach enabled us to select 6 BCW functions (education, training, environmental restructuring, modeling, persuasion, and enablement) and translate them into specific intervention components. For example, educational modules with analogies explaining CHD conditions address capability barriers (lack of knowledge), while peer narratives simultaneously address social opportunity barriers (modeling self-advocacy) and motivation barriers (building confidence and hope). We further mapped these intervention components to intervention objectives to develop a theoretical framework showing how intervention functions are hypothesized to impact outcomes (Figure 2).

XSL•FO
RenderX

Table . Capability, Opportunity, Motivation for Behavior Change (COM-B) and Behavior Change Wheel (BCW) framework for the digital intervention to promote specialized ACHD[a] care.

COM-B and component	Identified barriers (from interviews and CAB[b])	Behavioral targets (what needs to change?)	BCW intervention functions[c]	Intervention components within the digital tool
Capability				
Psychological	Lack of knowledge about CHD[d]; unable to find ACHD specialists; forget to schedule routine visits	Increase knowledge, help find ACHD specialists, reassurance	Education, Enablement, Persuasion	Educational modules; ACHD Clinic Directory link; ACHD appointment nudges
Opportunity				
Physical	Difficulty keeping all medical history centralized and easily available	Provide secure, easily accessible location	Environmental restructuring, Enablement	Digital medical passport with medical history and provider information
Social	Unable to advocate or connect with physicians; feeling "lonely" given CHD is rare	Prepared to advocate	Modeling, Enablement	Peer narratives demonstrating self-advocacy; peer connection opportunities through community event listings
Motivation				
Reflective	Belief in unmet needs (eg, family planning, travel, drugs, work/school); social stigma and low confidence related to scars	Reassurance that needs can be addressed with the ACHD team	Education, Training or Modeling	Educational content on life topics (pregnancy, career); peer success stories
Automotive	Fear of being dismissed	Foster optimism and address readiness	Persuasion, Modeling	Positive framing; peer narratives building confidence

[a]ACHD: adult congenital heart disease.

[b]CAB: community advisory board.

[c]"Functions" in the COM-B model are as follows: Education, Persuasion (induce positive feelings or stimulate action), Training (impart skills), Environmental restructuring (change the physical or social context), Modeling (examples to aspire to or imitate), and Enablement (reduce barriers).

[d]CHD: congenital heart disease.

XSL•FO
RenderX

Figure 2. Theoretical framework hypothesizing the impact of digital health intervention components on outcomes. The intervention components (educational modules, digital medical passport, peer connections, and appointment nudges), delivered via web and mobile app, function as tools for education, training, enablement, persuasion, modeling, and environmental restructuring to enhance the Capability, Opportunity, and Motivation (COM-B) domains of patients with ACHD. We hypothesize that these components will support patient activation and engagement skills and lifelong specialized ACHD care, ultimately improving patient experience and health outcomes. ACHD: adult congenital heart disease.

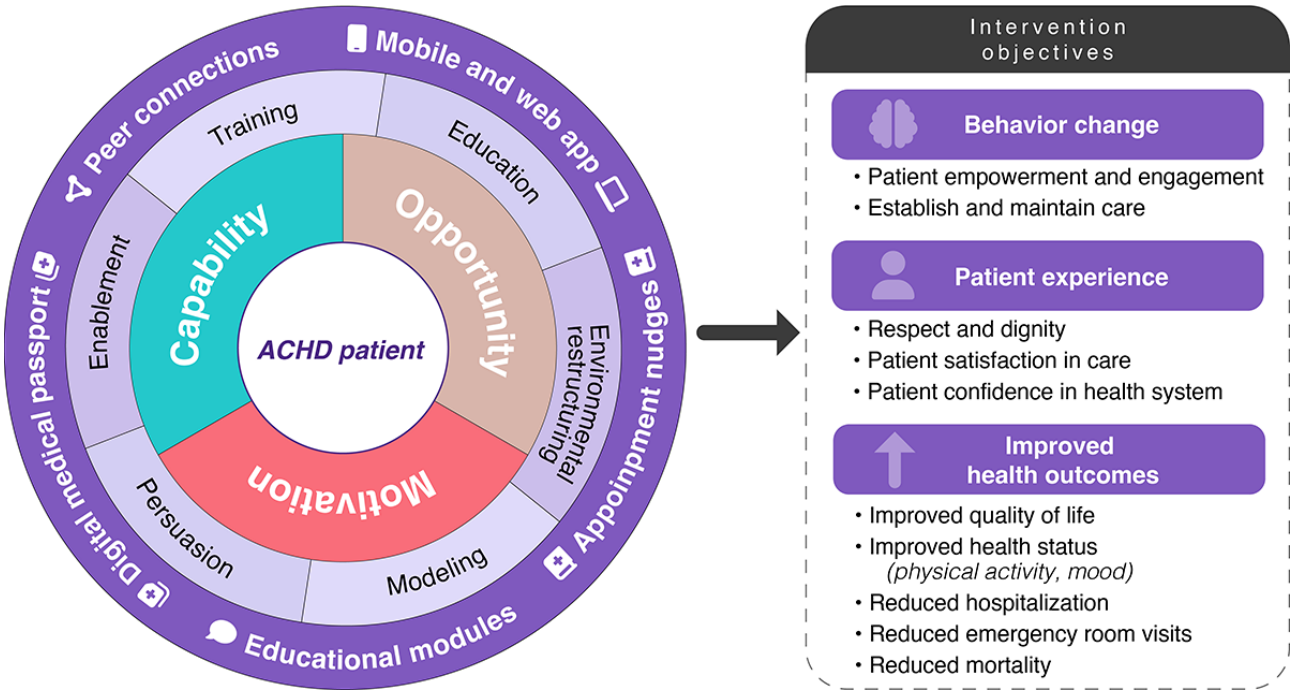

Design of the Digital Tool

Digital Research Platform

The digital tool was developed on the Eureka Research Platform (University of California, San Francisco), a Health Insurance Portability and Accountability Act–secure, National Institutes of Health (NIH)−supported infrastructure enabling integrated recruitment, consent, data collection, and intervention delivery [48]. Participants enroll remotely using QR codes or web links and consent electronically (UCSF IRB number 22‑36667; Table S7 in Multimedia Appendix 1).

Branding

The community partners discussed various options for the name and logo of the digital tool and narrowed them down to 4 potential ones: My Congenital Heart Care, Empower My Congenital Heart, My Congenital Heart Guide, and Uplift My Congenital Heart. The CAB members suggested including an arrow in the logo to reflect the primary goal of uplifting and empowering patients. Red, blue, and purple colors were chosen to reflect the diversity within CHD lesions and their uniqueness from other heart diseases. The design team developed 4 branding prototypes (Figure S1 in Multimedia Appendix 1). Through CAB voting, “Empower My Congenital Heart (EMCH)” with an integrated heart-and-arrow logo (option B-2) was selected.

Participant Flow Within EMCH

Participants can be recruited through email, clinic flyers, social media, or community outreach using QR codes. Participants can engage via smartphone or web browser and sign electronic consent. Activities are delivered bimonthly (every 2 months) and remain available for 2 months. Each activity includes surveys, educational modules, and optional wearable or portal linkages (Figure 3; Table S8 in Multimedia Appendix 1). Engagement is supported through multimodal communication (app notifications, email, texting), with up to 3 reminders per activity cycle. Bimonthly “Pokes” use variable timing and content (peer stories, study updates, CHD research opportunities) to maintain interest through intermittent reinforcement.

XSL•FO
RenderX

Figure 3. Participant flow through Empower My Congenital Heart (EMCH). EMCH is accessible via the UCSF Eureka Research mobile app or web platform. Participants first create a UCSF Eureka Research account (step 1) and then review and electronically sign the EMCH consent form (step 2). Following consent, participants receive study activities every 2 months (step 3), including surveys (participant data, feedback, and participant experiences), modules (clinician guidance and peer narratives, resources on topics related to congenital heart disease [CHD] and health system navigation), and optional connections to wearables, smartphone data, and patient portals. Participants also receive a "Poke" notification roughly every 2 months to remind them of new activities (step 4). UCSF: University of California San Francisco.

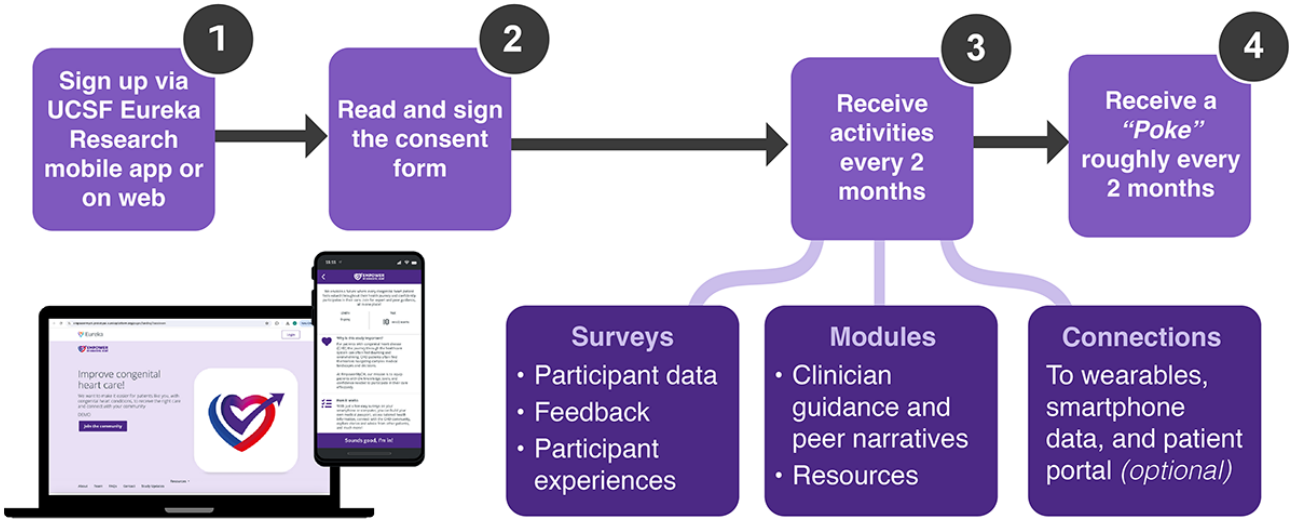

Core Intervention Components

The core intervention component includes educational modules (addressing *capability, opportunity, and motivation* barriers) that combine concise educational content with clinician guidance, "Empowerment" tips, and patient "Peer Empowerment" messages (Figure 4 and Multimedia Appendix 2; Table S9 in Multimedia Appendix 1) to build knowledge, confidence, and self-advocacy skills. Year 1 modules cover CHD education and health system navigation, while subsequent modules address psychosocial concerns, lifestyle, and other topics (Table S10 in Multimedia Appendix 1). Additional components include appointment planning nudges (addressing *capability* barriers) embedded within annual surveys to prompt scheduling ACHD specialist visits; personal digital medical passport (addressing *physical opportunity* barrier), auto-generated from baseline self-reported data (diagnosis, providers, etc) serving as a centralized information hub always accessible via smartphone (Figure 5; Table S11 in Multimedia Appendix 1); and peer connections (addressing *social opportunity and motivation* barriers) through vetted Peer Empowerment and community event listings, with all content reviewed for accuracy and safety before inclusion.

Figure 4. Representative core intervention components within Empower My Congenital Heart (EMCH). The screenshots show representative screens of the educational modules (with clinician guidance, "Empowerment," and patient-derived "Peer Empowerment" stories), appointment planning nudges (embedded within the surveys), and personal digital medical passport (a centralized hub for all medical and provider information while also linking participants to curated community events, all EMCH modules, the adult congenital heart disease [ACHD] provider directory, and a secure feedback and contact form). The peer connections component of EMCH is integrated within the modules and community event listings.

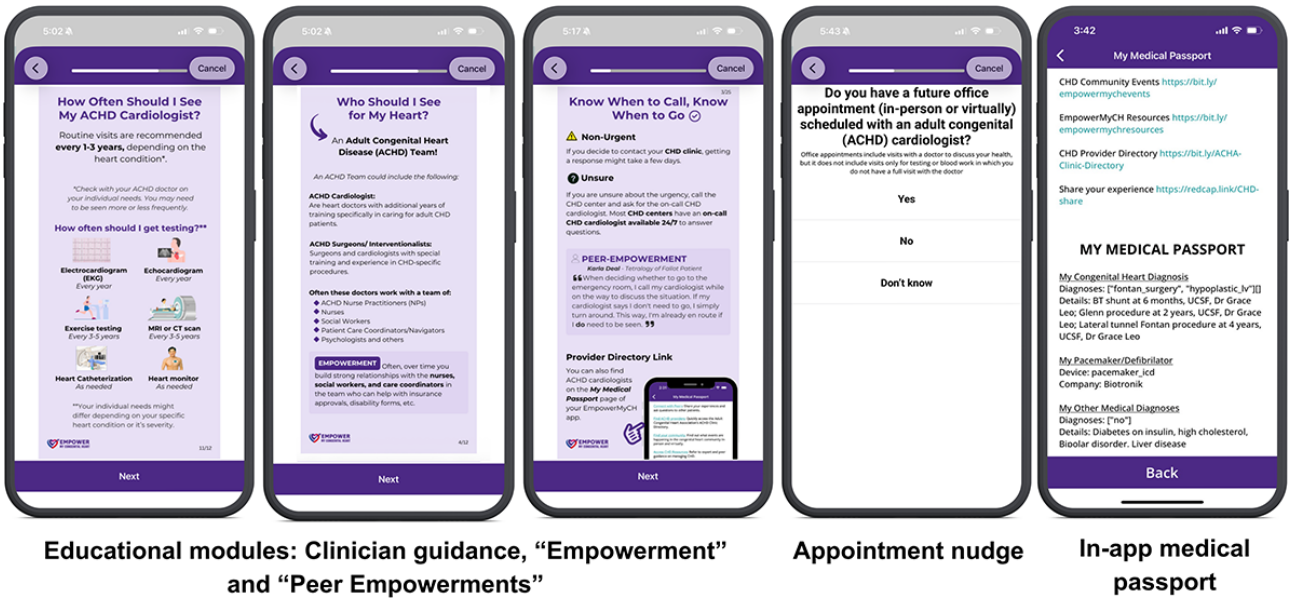

XSL•FO
RenderX

Figure 5. Digital medical passport built using self-reported data. This figure illustrates the type of health information easily accessible to participants at any time through the Empower My Congenital Heart (EMCH) mobile app. By completing a series of surveys with self-reported data, participants build their own digital medical passport, a personalized hub for their vital health and provider information.

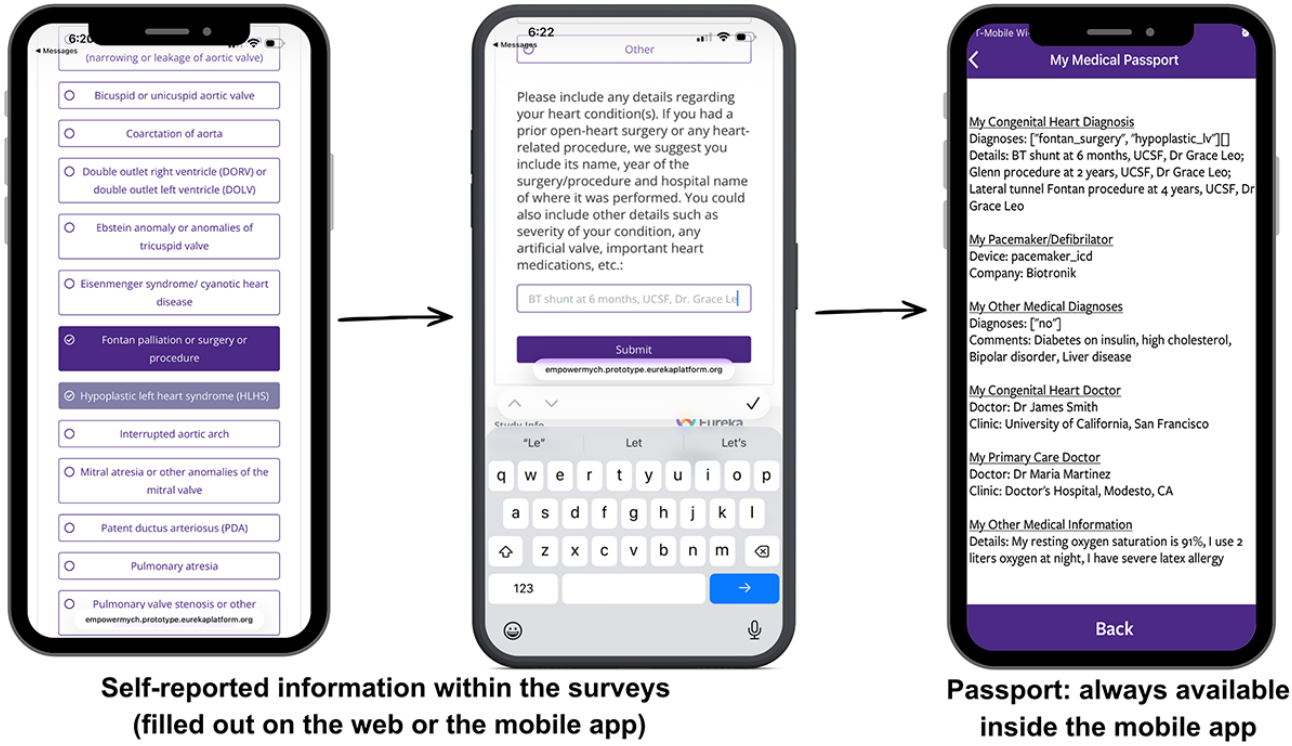

Iterative Refinements During Intervention Development

Throughout development, intervention features were refined based on pilot testing, CAB feedback, and adherence to digital health equity principles (Table 5). These refinements demonstrate the responsive, user-centered development process that shaped the final EMCH design.

XSL•FO
RenderX

Table . Iterative refinements during intervention development.

Initial design	Refinement	Rationale	Data source
Monthly delivery of the module	Modules delivered every 2 months	Reduce participant burden while maintaining engagement	CAB[a] feedback
Colorful icons or imagery	Flat, minimalist icons	Enhance readability, reduce bandwidth needs, improve accessibility	Digital health equity principles + CAB feedback
Frequent reminders (9 total: 3 email, 3 push, 3 texts per module)	Reduced to 3 total reminders per EMCH[b] activity	Minimize intrusiveness, respect participant time	User testing + CAB feedback
Reminders sent until all activities completed (survey and modules)	Reminders sent only if module incomplete; most surveys are not reprompted	Minimize intrusiveness for data collection while providing opportunities to engage with modules	CAB feedback
Digital passport surveys delivered at 2 months post enrollment	Digital passport surveys delivered at the time of enrollment	Maximize early engagement with digital passport from first access	User testing + CAB feedback
Clinician guidance and peer narratives labeled as “Tips”	“Clinician guidance,” “Empowerment” quotes, and “Peer Empowerment” quotes included as distinct color-coded categories	Emphasize empowerment, positivity, and a component of “expert and peer guidance” as primary intervention goals	CAB feedback
Module content primarily informational	Added interactive elements (trivia questions, action items) within the modules	Enhance engagement and encourage clear, actionable takeaways	CAB feedback
Digital passport to include comprehensive data (CHD[c] history, other medical history, all providers, medications, insurance)	Digital passport limited to pertinent medical and provider information	Simplify content to enhance usability and relevance; reduce data entry burden	Eureka platform team + CAB feedback

[a]CAB: community advisory board.

[b]EMCH: Empower My Congenital Heart.

[c]CHD: congenital heart disease.

Data Security and Privacy

The Eureka Platform uses 128-bit secure socket layer encryption for data transmission between the app and servers. Data are stored on Health Insurance Portability and Accountability Act–compliant servers managed by UCSF with restricted access, multifactor authentication, and role-based permissions. Survey responses and engagement metrics are stored as separate comma-separated values files using unique identifiers without personal identifiable information. Only the authorized research team can link the deidentified data to individual participants. Participants can withdraw at any time by contacting the research team. Upon withdrawal, they may request the deletion of their data or allow deidentified data to remain for research purposes. Participants control whether their stories can be shared as Peer Empowerment content (explicit opt-in required). Wearable device data (Fitbit) and patient portal connections require separate participant authorization and are governed by those platforms’ privacy policies in addition to EMCH protections. Study data will be retained per UCSF IRB and NIH data sharing policies, after which the data will be securely destroyed or permanently deidentified.

Discussion

Principal Results

We describe a theory-based, community co-designed approach to developing EMCH, a digital intervention promoting confidence and skills to enhance engagement in lifelong ACHD care. To our knowledge, this is the first cross-platform (Android and iOS) digital tool specifically designed to address ACHD-specific barriers to lifelong specialized engagement. Our approach integrated evidence-based behavioral frameworks (COM-B or BCW) with extensive qualitative research (n=54) and community-based participatory design to map barriers to engagement and develop targeted intervention components. Our approach ensured that patient and clinician perspectives shaped all design decisions. EMCH enables convenient, passive content access using user-friendly design and behavioral economics (nudge) principles while fostering peer connections and community contributions.

Clinical Relevance and Expected Impact

Existing CHD digital interventions focus narrowly on symptom monitoring, exercise support, or transition readiness in young adults and lack theoretical grounding or rigorous evaluation [23,24]. EMCH addresses the full adult lifespan (18+ years) and targets the complete spectrum of barriers to ongoing specialist care for all patients, including those who never transitioned from pediatric care or were lost to follow-up. The intervention’s systematic foundation in behavioral theory (COM-B or BCW) distinguishes it from education-only approaches by addressing psychological capability, social opportunity, and motivation simultaneously. If effective, EMCH could reduce preventable ACHD complications, emergent hospitalizations, and mortality associated with care gaps by promoting sustained engagement with specialist care.

XSL•FO
RenderX

Scalability and Sustainability

We aim to recruit patients with ACHD at high risk for care gaps (young adults, men, rural residence, lower socioeconomic status, those living far away from CHD centers, etc) and those not previously engaged in research. Mobile phones are nearly ubiquitous—96% of Americans (and 94% of patients with ACHD) in their twenties and thirties own smartphones, with particularly high reliance among young adults, non-Whites, and lower-income Americans [22,49,50]. EMCH cross-platform, low-bandwidth design, and digital health equity framework enable reach across diverse populations. EMCH was designed to augment, not replace, patient-provider interactions by building engagement skills outside clinic visits. Unlike clinic-based navigator programs requiring ongoing personnel costs, EMCH automated content delivery and peer-generated stories provide sustainable engagement with minimal marginal cost per additional user. We selected the Eureka Digital Research Platform for its sustainability (NIH-supported, UCSF-owned), integration of recruitment through intervention delivery, and cost-efficiency in supporting hundreds of concurrent studies.

How EMCH Differs From Existing Approaches

We selected BCW over alternative frameworks (eg, Intervention mapping or the Consolidated Framework for Implementation Research) [51,52] because it provides flexible, evidence-based guidance for digital behavior change interventions without being overly prescriptive [25,53]. The BCW process enabled transparent documentation of our intervention's theory of action by systematically mapping behavioral barriers to specific components using shared taxonomic language [32]. Integrating a community-engaged approach with behavioral theory proved essential for moving beyond identifying barriers [30]. While we knew many patients with ACHD lack the awareness of specialist care needs [5], interviews and CAB feedback showed that evidence-based information alone was insufficient. Participants needed peer stories and real-life scenarios to internalize the relevance. This insight led to "Empowerment" and "Peer Empowerment" quotes becoming core components, using modeling to build confidence and demonstrate care's importance when patients feel well.

The combined qualitative, theory- and community-engaged approach can be resource-intensive but adaptable. Teams with limited resources can supplement qualitative work with rapid stakeholder consultation while maintaining the framework's core strengths: preventing the reinvention of existing content, simultaneously addressing engagement and implementation barriers, and ensuring intervention components meet users' actual needs.

Limitations

EMCH addresses patient-level barriers but cannot overcome system-level challenges (ACHD specialist shortages, insurance gaps, geographic distance from centers). The intervention is most appropriate for stable patients capable of self-management with remote guidance; patients with complex acute needs or severe cognitive impairment require in-person clinical support. While 94% of patients with ACHD own smartphones [22] and EMCH is available on web and native iOS or Android platforms with low-bandwidth design, the most vulnerable patients may face barriers beyond digital intervention scope. However, by reducing routine educational burden for digitally engaged patients, EMCH may enable clinical teams to prioritize in-person resources for those with greater needs. The digital passport currently captures baseline data only; update functionality is under development, and we are exploring electronic health record integration to auto-populate verified clinical data. Although we incorporated diverse perspectives during development, we may not have captured all viewpoints. The EMCH infrastructure supports continuous adaptation, enabling rapid implementation of new components (eg, mental health modules, culturally adapted content). Currently, individuals with intellectual disabilities are excluded due to unique accommodation needs requiring caregiver-focused approaches.

Next Steps

EMCH was launched in September 2024 and has enrolled more than 500 participants. Ongoing process evaluation is assessing acceptability, usability, and preliminary engagement patterns across diverse ACHD populations. We are developing a Spanish-language adaptation to reach the 15% to 20% of patients with ACHD who are primarily Spanish-speaking.

Following process evaluation, we will conduct a randomized controlled trial evaluating EMCH's effectiveness in completing ACHD specialist visits and improving knowledge, self-efficacy, and social connectedness. If proven effective, EMCH will be scaled to reach patients with ACHD nationally through partnerships with community advocacy organizations and integration into diverse clinical settings.

Conclusions

The theory-based, community co-designed EMCH digital tool equips patients with ACHD with evidence-based resources, peer support, and practical skills to sustain engagement with lifelong specialized care. Ongoing process evaluation will identify the acceptability, usability, and engagement patterns to inform a planned randomized controlled trial. If proven effective in reducing care gaps, EMCH has the potential to reduce preventable ACHD complications, emergent hospitalizations, and mortality—outcomes currently affecting the majority of patients with ACHD. The behavioral frameworks and intervention strategies address universal barriers to care engagement and can be adapted for other chronic conditions requiring lifelong specialized care. The EMCH infrastructure also supports the rapid recruitment of diverse ACHD populations for future research and health policy initiatives.

Acknowledgments

The authors would like to acknowledge the important contributions of the patients, families, health care professionals, and advocacy representatives who partnered as our community advisory board.

XSL•FO
RenderX

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing, summarizing text, and reformatting. The GAI tool used was Claude Sonnet 4.5. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Declaration submitted by: AA

Funding

Research reported in this manuscript was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under award number K23HL151866 (AA). The content is solely the responsibility of the presenters and does not necessarily represent the official views of the National Institutes of Health. GMM discloses support for the Eureka Research Platform from National Institute of Biomedical Imaging and Bioengineering under award number 3U2CEB021881-05S1.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

AA and JV contributed to all aspects of the study design and to the preparation of the initial drafts and revisions of the paper. KB contributed to the recruitment and retention of study participants, qualitative analysis, development of study activities, and study regulatory management. K Macholl contributed to the qualitative interview process, identification of necessary design objectives, and intervention features for Empower My Congenital Heart. JM, KR, K Macholl, and PA contributed to qualitative analysis, study conduct, and critical review of the manuscript. PK contributed to the technical development and adaptation of Empower My Congenital Heart study activities on the University of California, San Francisco Eureka Research platform. KL, KP, MDN, KB-J, LR, PM, MJO, GMM,
K Manayan, and MG contributed to the overall study design and critical review of the manuscript.

Conflicts of Interest

MDN has been a consultant for American College of Cardiology since 2024. The other authors declare no conflicts of interest.

Multimedia Appendix 1
Supporting materials and tables.
[DOCX File, 737 KB - jmir_v28i1e75867_app1.docx]

Multimedia Appendix 2
Screenshots of the intervention components within Empower My Congenital Heart (EMCH).
[PDF File, 3729 KB - jmir_v28i1e75867_app2.pdf]

References

1. Gilboa SM, Devine OJ, Kucik JE, et al. Congenital heart defects in the United States: estimating the magnitude of the affected population in 2010. Circulation 2016 Jul 12;134(2):101-109. [doi: 10.1161/CIRCULATIONAHA.115.019307] [Medline: 27382105]
2. Khairy P, Ionescu-Ittu R, Mackie AS, Abrahamowicz M, Pilote L, Marelli AJ. Changing mortality in congenital heart disease. J Am Coll Cardiol 2010 Sep 28;56(14):1149-1157. [doi: 10.1016/j.jacc.2010.03.085] [Medline: 20863956]
3. Stout KK, Daniels CJ, Aboulhosn JA, et al. 2018 AHA/ACC guideline for the management of adults with congenital heart disease: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol 2019 Apr 2;73(12):1494-1563. [doi: 10.1016/j.jacc.2018.08.1028] [Medline: 30121240]
4. Gurvitz M, Krieger EV, Fuller S, et al. 2025 ACC/AHA/HRS/ISACHD/SCAI guideline for the management of adults with congenital heart disease: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. J Am Coll Cardiol 2026 Feb 24;87(7):822-976. [doi: 10.1016/j.jacc.2025.09.006] [Medline: 41411480]
5. Gurvitz M, Valente AM, Broberg C, et al. Prevalence and predictors of gaps in care among adult congenital HEART disease patients: HEART-ACHD (The Health, Education, and Access Research Trial). J Am Coll Cardiol 2013 May 28;61(21):2180-2184. [doi: 10.1016/j.jacc.2013.02.048] [Medline: 23542112]
6. Khan AM, McGrath LB, Ramsey K, Agarwal A, Slatore CG, Broberg CS. Distance to care, rural dwelling status, and patterns of care utilization in adult congenital heart disease. Pediatr Cardiol 2022 Mar;43(3):532-540. [doi: 10.1007/s00246-021-02750-7] [Medline: 34705069]

XSL•FO
RenderX

7. Mackie AS, Ionescu-Ittu R, Therrien J, Pilote L, Abrahamowicz M, Marelli AJ. Children and adults with congenital heart disease lost to follow-up: who and when? Circulation 2009 Jul 28;120(4):302-309. [doi: 10.1161/CIRCULATIONAHA.108.839464] [Medline: 19597053]
8. Bayne J, Duan R, Rudov L, et al. Understanding gaps in guideline-recommended adult congenital heart disease care: data from 12 US health care centers. Am Heart J 2026 Jan;291:53-62. [doi: 10.1016/j.ahj.2025.08.002] [Medline: 40783166]
9. Yeung E, Kay J, Roosevelt GE, Brandon M, Yetman AT. Lapse of care as a predictor for morbidity in adults with congenital heart disease. Int J Cardiol 2008 Mar 28;125(1):62-65. [doi: 10.1016/j.ijcard.2007.02.023] [Medline: 17442438]
10. Mylotte D, Pilote L, Ionescu-Ittu R, et al. Specialized adult congenital heart disease care: the impact of policy on mortality. Circulation 2014 May 6;129(18):1804-1812. [doi: 10.1161/CIRCULATIONAHA.113.005817] [Medline: 24589851]
11. Agarwal A, Duan R, Bayne J, et al. Impact of adult congenital heart disease specialist visits on emergent admissions: evidence for guidelines. JACC Adv 2025 Aug;4(8):102021. [doi: 10.1016/j.jacadv.2025.102021] [Medline: 40730048]
12. Agarwal A, Macholl K, Qian A, et al. Patient- and clinician-solutions to improve specialized ACHD care: a theory-based approach. medRxiv. Preprint posted online on Apr 3, 2025. [doi: 10.1101/2025.04.01.25325065]
13. John AS, Jackson JL, Moons P, et al. Advances in managing transition to adulthood for adolescents with congenital heart disease: a practical approach to transition program design: a scientific statement from the American Heart Association. J Am Heart Assoc 2022 Apr 5;11(7):e025278. [doi: 10.1161/JAHA.122.025278] [Medline: 35297271]
14. Bratt EL, Mora MA, Sparud-Lundin C, et al. Effectiveness of the STEPSTONES transition program for adolescents with congenital heart disease—a randomized controlled trial. J Adolesc Health 2023 Oct;73(4):655-663. [doi: 10.1016/j.jadohealth.2023.02.019] [Medline: 37032211]
15. Gaydos SS, Chowdhury SM, Judd RN, McHugh KE. A transition clinic intervention to improve follow-up rates in adolescents and young adults with congenital heart disease. Cardiol Young 2020 May;30(5):633-640. [doi: 10.1017/S1047951120000682] [Medline: 32279695]
16. Mackie AS, Rempel GR, Kovacs AH, et al. Transition intervention for adolescents with congenital heart disease. J Am Coll Cardiol 2018 Apr 24;71(16):1768-1777. [doi: 10.1016/j.jacc.2018.02.043] [Medline: 29673467]
17. Thomet C, Schwerzmann M, Budts W, et al. Transfer and transition practices in 96 European adult congenital heart disease centres. Int J Cardiol 2021 Apr 1;328:89-95. [doi: 10.1016/j.ijcard.2020.11.031] [Medline: 33276020]
18. Fernandes SM, Khairy P, Fishman L, et al. Referral patterns and perceived barriers to adult congenital heart disease care: results of a survey of U.S. pediatric cardiologists. J Am Coll Cardiol 2012 Dec 11;60(23):2411-2418. [doi: 10.1016/j.jacc.2012.09.015] [Medline: 23141490]
19. Fernandes SM, Verstappen A, Clair M, et al. Knowledge of life-long cardiac care by adolescents and young adults with congenital heart disease. Pediatr Cardiol 2019 Oct;40(7):1439-1444. [doi: 10.1007/s00246-019-02154-8] [Medline: 31367952]
20. Free C, Phillips G, Watson L, et al. The effectiveness of mobile-health technologies to improve health care service delivery processes: a systematic review and meta-analysis. PLoS Med 2013;10(1):e1001363. [doi: 10.1371/journal.pmed.1001363] [Medline: 23458994]
21. Low JK, Manias E. Use of technology-based tools to support adolescents and young adults with chronic disease: systematic review and meta-analysis. JMIR Mhealth Uhealth 2019 Jul 18;7(7):e12042. [doi: 10.2196/12042] [Medline: 31322129]
22. Lopez KN, O'Connor M, King J, et al. Improving transitions of care for young adults with congenital heart disease: mobile app development using formative research. JMIR Form Res 2018;2(2):e16. [doi: 10.2196/formative.9963] [Medline: 30574573]
23. Marelli A, Rozenblum R, Bolster-Foucault C, et al. Development of MyREADY transition BBD mobile app, a health intervention technology platform, to improve care transition for youth with brain-based disabilities: user-centered design approach. JMIR Pediatr Parent 2024 Oct 1;7:e51606. [doi: 10.2196/51606] [Medline: 39352737]
24. Nashat H, Habibi H, Heng EL, et al. Patient monitoring and education over a tailored digital application platform for congenital heart disease: a feasibility pilot study. Int J Cardiol 2022 Sep 1;362:68-73. [doi: 10.1016/j.ijcard.2022.05.002] [Medline: 35526658]
25. Klonoff DC. Behavioral theory: the missing ingredient for digital health tools to change behavior and increase adherence. J Diabetes Sci Technol 2019 Mar;13(2):276-281. [doi: 10.1177/1932296818820303] [Medline: 30678472]
26. Liang L, Bernhardsson S, Vernooij RWM, et al. Use of theory to plan or evaluate guideline implementation among physicians: a scoping review. Implement Sci 2017 Feb 27;12(1):26. [doi: 10.1186/s13012-017-0557-0] [Medline: 28241771]
27. Handley MA, Abascal Miguel L, Thompson LM, Velloza J. Bridging community-engaged research and implementation science methods to advance public health practice. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2025 Jul;68(7):797-808. [doi: 10.1007/s00103-025-04079-5] [Medline: 40471262]
28. Riccardi M, Pettinicchio V, Di Pumpo M, et al. Community-based participatory research to engage disadvantaged communities: levels of engagement reached and how to increase it. A systematic review. Health Policy 2023 Nov;137:104905. [doi: 10.1016/j.healthpol.2023.104905] [Medline: 37716190]
29. Baker TB, Gustafson DH, Shah D. How can research keep up with eHealth? Ten strategies for increasing the timeliness and usefulness of eHealth research. J Med Internet Res 2014 Feb 19;16(2):e36. [doi: 10.2196/jmir.2925] [Medline: 24554442]

XSL•FO
RenderX

30. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015 Jan 30;17(1):e30. [doi: 10.2196/jmir.4055] [Medline: 25639757]
31. Cooper C, Watson K, Alvarado F, et al. Community engagement in implementation science: the impact of community engagement activities in the DECIPHeR alliance. Ethn Dis 2024 Dec;DECIPHeR(Spec Issue):52-59. [doi: 10.18865/ed.DECIPHeR.52] [Medline: 38846725]
32. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
33. Cane J, O'Connor D, Michie S. Validation of the theoretical domains framework for use in behaviour change and implementation research. Implement Sci 2012 Apr 24;7:37. [doi: 10.1186/1748-5908-7-37] [Medline: 22530986]
34. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119. [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
35. Haider M, Kreps GL. Forty years of diffusion of innovations: utility and value in public health. J Health Commun 2004;9:3-11. [doi: 10.1080/10810730490271430] [Medline: 14960400]
36. Liang SY, Stults CD, Jones VG, et al. Effects of behavioral economics-based messaging on appointment scheduling through patient portals and appointment completion: observational study. JMIR Hum Factors 2022 Mar 30;9(1):e34090. [doi: 10.2196/34090] [Medline: 35353051]
37. Alexander KE, Brijnath B, Mazza D. Barriers and enablers to delivery of the Healthy Kids Check: an analysis informed by the Theoretical Domains Framework and COM-B model. Implement Sci 2014 May 23;9:60. [doi: 10.1186/1748-5908-9-60] [Medline: 24886520]
38. Whittal A, Störk S, Riegel B, Herber OR. Applying the COM-B behaviour model to overcome barriers to heart failure self-care: a practical application of a conceptual framework for the development of complex interventions (ACHIEVE study). Eur J Cardiovasc Nurs 2021 Mar 1;20(3):261-267. [doi: 10.1177/1474515120957292] [Medline: 33909892]
39. Jatau AI, Peterson GM, Bereznicki L, et al. Applying the Capability, Opportunity, and Motivation Behaviour model (COM-B) to guide the development of interventions to improve early detection of atrial fibrillation. Clin Med Insights Cardiol 2019;13:1179546819885134. [doi: 10.1177/1179546819885134] [Medline: 31700252]
40. Mc Sharry J, Murphy PJ, Byrne M. Implementing international sexual counselling guidelines in hospital cardiac rehabilitation: development of the CHARMS intervention using the Behaviour Change Wheel. Implement Sci 2016 Oct 10;11(1):134. [doi: 10.1186/s13012-016-0493-4] [Medline: 27724957]
41. Higginbottom GMA. Sampling issues in qualitative research. Nurse Res 2004;12(1):7-19. [doi: 10.7748/nr2004.07.12.1.7.c5927] [Medline: 15493211]
42. Saunders B, Sim J, Kingstone T, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907. [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
43. St George SM, Harkness AR, Rodriguez-Diaz CE, Weinstein ER, Pavia V, Hamilton AB. Applying rapid qualitative analysis for health equity: lessons learned using "EARS" with Latino communities. Int J Qual Methods 2023;22:16094069231164938. [doi: 10.1177/16094069231164938] [Medline: 38463016]
44. Adult Congenital Heart Association. URL: https://www.achaheart.org/ [accessed 2025-03-19]
45. Conquering Congenital Heart Disease. URL: https://www.conqueringchd.org/ [accessed 2025-03-19]
46. The Mended Hearts, Inc. 2025. URL: https://mendedhearts.org/ [accessed 2026-05-28]
47. Team Uncle Joe. 2025. URL: https://teamunclejoe.org/ [accessed 2026-05-28]
48. Beatty AL, Peyser ND, Butcher XE, et al. The COVID-19 citizen science study: protocol for a longitudinal digital health cohort study. JMIR Res Protoc 2021 Aug 30;10(8):e28169. [doi: 10.2196/28169] [Medline: 34310336]
49. Mobile fact sheet. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2026-05-28]
50. Lopez MH, Gonzales-Barrera A, Patten EM. Closing the digital divide: Latinos and technology adoption. : Pew Research Center; 2013 URL: https://assets.pewresearch.org/wp-content/uploads/sites/7/2013/03/Latinos_Social_Media_and_Mobile_Tech_03-2013_final.pdf [accessed 2020-04-22]
51. Acuña Mora M, Saarijärvi M, Sparud-Lundin C, Moons P, Bratt EL. Empowering young persons with congenital heart disease: using intervention mapping to develop a transition program—the STEPSTONES project. J Pediatr Nurs 2020;50:e8-e17. [doi: 10.1016/j.pedn.2019.09.021] [Medline: 31669495]
52. Damschroder LJ, Reardon CM, Opra Widerquist MA, Lowery J. Conceptualizing outcomes for use with the Consolidated Framework for Implementation Research (CFIR): the CFIR Outcomes Addendum. Implementation Sci 2022 Dec;17(1):7. [doi: 10.1186/s13012-021-01181-5]
53. Kuhl EA, Sears SF, Conti JB. Internet-based behavioral change and psychosocial care for patients with cardiovascular disease: a review of cardiac disease-specific applications. Heart Lung 2006;35(6):374-382. [doi: 10.1016/j.hrtlng.2006.02.004] [Medline: 17137938]

XSL•FO
RenderX

Abbreviations

ACHD: adult congenital heart disease
BCW: Behavior Change Wheel
CAB: community advisory board
CHD: congenital heart disease
COM-B: Capability, Opportunity, Motivation for Behavior Change
EMCH: Empower My Congenital Heart
IRB: institutional review board
NIH: National Institutes of Health
UCSF: University of California, San Francisco

Edited by A Stone; submitted 11.Apr.2025; peer-reviewed by A Cassidy, A Mackie, C Lauff, NR Lawrence; revised version received 08.May.2026; accepted 13.May.2026; published 23.Jun.2026.

Please cite as:
Agarwal A, Valente J, Buenrostro K, Macholl K, Mehta J, Reddy K, Manayan K, Kim P, Sparks A, Li K, Ahuja P, Sun K, Payton K, Norris MD, Bravo-Jaimes K, Reardon L, Moons P, Okumura MJ, Marcus GM, Gurvitz M
Digital Health Intervention to Promote Lifelong Specialized Care in Adults With Congenital Heart Disease: Theory-Driven Community Co-Designed Study
J Med Internet Res 2026;28:e75867
URL: https://www.jmir.org/2026/1/e75867
doi:10.2196/75867

© Anushree Agarwal, Joseph Valente, Karina Buenrostro, Katelyn Macholl, Juhi Mehta, Keerthana Reddy, Karina Manayan, Parang Kim, Aleah Sparks, Kunyi Li, Pranav Ahuja, Kevin Sun, Kimberly Payton, Mark D Norris, Katia Bravo-Jaimes, Leigh Reardon, Philip Moons, Megumi J Okumura, Gregory M Marcus, Michelle Gurvitz. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Multilingual Digital Microlearning Intervention for Oral Health in Refugee Shelters: Randomized Controlled Trial

Maxi Katharina Müller[1*]; Fabian Schulz[2*]; George Jogho[3,4]; Simone Steffens[2]; Kirstin Vach[5,6], Prof Dr; Ahmad Amro Baradee[2]; Daniel R Reissmann[7], MSc, Prof Dr; Benedikt C Spies[2], Prof Dr; Anna-Lena Hillebrecht[2]

[1]Faculty of Medicine, Department of Operative Dentistry and Periodontology, Medical Center—University of Freiburg, Freiburg, Germany, Freiburg, Germany

[2]Department of Prosthetic Dentistry, Faculty of Medicine, Medical Center—University of Freiburg, Hugstetter Str. 55, Freiburg, Germany

[3]Section of Health Care Research and Rehabilitation Research, Institute of Medical Biometry and Statistics, Medical Faculty and Medical Center - University of Freiburg, Germany, Freiburg, Germany

[4]RefuDocs Freiburg e. V, Wirthstr. 7A, Freiburg, Germany

[5]Institute of Medical Biometry and Statistics, Medical Faculty and Medical Center - University of Freiburg, Germany, Freiburg, Germany

[6]Center for Data Science, Eberswalde University for Sustainable Development, Eberswalde, Germany

[7]Department of Prosthodontics and Materials Science, Centre for Dental Medicine, Medical Centre, University of Leipzig, Leipzig, Germany, Leipzig, Germany

[*]these authors contributed equally

Corresponding Author:
Anna-Lena Hillebrecht
Department of Prosthetic Dentistry, Faculty of Medicine, Medical Center—University of Freiburg, Hugstetter Str. 55, Freiburg, Germany

Abstract

Background: Refugees frequently face language and access barriers to preventive oral health information. Brief multilingual digital interventions may help reduce such barriers in shelter settings.

Objective: This randomized controlled trial evaluated whether a multilingual digital microlearning video improved plaque control and selected self-reported oral health–related behaviors among adults living in refugee shelters.

Methods: A 2-arm, parallel-group randomized controlled trial was conducted among 86 adults living in 2 municipal refugee shelters in Germany. Participants were randomized (1:1) to receive either a multilingual 4-minute oral hygiene microlearning video or delayed access (control group). Plaque index and gingival index were assessed clinically at baseline and at 2-month follow-up. Secondary outcomes included questionnaire-based measures of oral health literacy–related cognitions and self-reported oral health behaviors. Between-group differences in change scores were analyzed using 2-sided tests; exploratory multivariable regression analyses were conducted to assess potential effect modifiers.

Results: Follow-up was completed by 83 (97%) of 86 participants. Plaque index decreased more in the intervention group than in the control group (mean change −0.21, SD 0.27 vs mean change −0.04, SD 0.17; P=.002). Gingival index decreased in both groups, but the between-group difference was not significant. Among questionnaire-based outcomes, toothbrushing frequency increased substantially, whereas the remaining oral health literacy–related items showed small numerical changes that did not reach statistical significance or remained stable. Approximately three-quarters of participants in the intervention group (32/42, 76%) reported reviewing the video at least once.

Conclusions: Brief multilingual digital microlearning improved plaque control and self-reported toothbrushing frequency in refugee shelters. Effects on broader oral health literacy–related outcomes were limited and should be interpreted cautiously. Larger, prospectively powered trials with longer follow-up periods and blinded outcome assessment are warranted.

Trial Registration: German Clinical Trials Register DRKS00032017; https://drks.de/search/en/trial/DRKS00032017

(J Med Internet Res 2026;28:e95562) doi:10.2196/95562

KEYWORDS

refugees; digital health; health equity; health literacy; oral health literacy; microlearning; randomized controlled trial

XSL•FO
RenderX

Introduction

Digital access to health information is not dependent on staff or time, is cost-effective, and can thus be made available to more individuals more promptly and with fewer barriers. Digital health information has expanded rapidly over the past decade, yet vulnerable and displaced populations continue to face profound informational inequities. Refugees often encounter linguistic, structural, and administrative barriers that limit access to preventive services and reliable health education. These information barriers contribute to avoidable disease burden and widen existing health disparities. Scalable, low-threshold digital interventions may offer a pragmatic approach to reducing such inequities, particularly in high-mobility and resource-constrained settings [1-3].

The global number of forcibly displaced people increased from approximately 65 million in 2015 to more than 82 million by the end of 2020 and has continued to rise in recent years, reaching more than 120 million worldwide [4,5]. Oral health is a key but often overlooked dimension of refugee health, with untreated dental diseases reducing quality of life and increasing demands on public health care systems [6,7]. Poor oral health can also be linked to systemic conditions such as cardiovascular disease and diabetes, and it contributes to psychosocial stress, stigma, and reduced employability [8]. These interrelations underline the importance of addressing oral health as part of overall refugee health.

In Germany, oral health has improved over recent decades, yet refugees remain an underserved population [9]. During the initial period after arrival, dental care is generally restricted under the Asylum Seekers Benefits Act to treatment required for acute illness or pain. In addition to uncertainties regarding insurance status and repeated transfers, this can lead to most carious lesions remaining untreated, with substantial consequences for individual well-being and increased health care costs [10]. Administrative access models during the initial reception period may further contribute to inequities in health care use, with voucher-based systems associated with less equitable access than electronic health card models or regular access [11]. Similar disparities have been reported across other European host countries, including Greece, Sweden, and the United Kingdom, where refugees show higher caries prevalence, lower dental service use, and greater unmet treatment needs than the resident populations [12].

Disparities are compounded by conditions in countries of origin, where oral health systems often have fewer dental professionals, limited preventive services, and a higher prevalence of caries [13-17]. In host countries, barriers such as low income, limited education, language difficulties, and differing cultural health beliefs further undermine oral health behaviors [18].

Digital and video-based approaches have shown promise for oral health promotion, particularly when adapted to the needs of populations facing language or literacy barriers [19-21]. Web-based oral health promotion programs have also been shown to improve knowledge, attitudes, self-efficacy, and oral hygiene practices in adult populations, supporting the potential of digital formats for preventive oral health education [22]. A mixed methods systematic review of digital health literacy interventions for forced migrant populations found that 83% of included studies reported positive outcomes despite cultural, linguistic, and practical barriers [23]. Recent research has shown that adults with migration experience demonstrate significantly poorer oral health and lower oral health literacy than the general population, with oral health literacy emerging as an independent predictor of clinical outcomes [24]. Educational interventions targeting refugees have also demonstrated improvements in oral health literacy, awareness, and self-reported oral hygiene behaviors following structured oral health education programs [25]. In parallel, evidence from the development of a digital prevention program indicated that tailored mobile health interventions can successfully improve oral health literacy and oral hygiene in these groups, although such approaches are often complex and resource intensive [26]. Existing oral health promotion resources are often designed for majority-language populations and may require sustained engagement with mobile apps or web-based platforms, stable internet access, and literacy levels that cannot be assumed in newly arrived displaced populations. In addition, many available tools lack cultural and linguistic tailoring to the heterogeneous contexts of refugee shelters and may therefore fail to reach those with the greatest preventive needs. These limitations informed the development of GlobeSmile as an ultrabrief, multilingual, low-threshold microlearning intervention designed for access via standard smartphones using QR codes and requiring minimal digital skills or prior knowledge.

The “GlobeSmile” intervention was designed as a highly accessible, low-threshold tool to provide refugees with essential oral health information in their own languages. The short video focused on simple, practical instructions for daily oral hygiene procedures, such as toothbrushing technique, the use of fluoride, and preventive practices, in line with the clinical practice recommendations of the *Fédération Dentaire Internationale* (commonly known as FDI World Dental Federation) for maintaining good oral health and with evidence-informed international consensus recommendations on toothbrushing and oral hygiene behaviors [27,28]. Therefore, this study aimed to evaluate the effectiveness of the GlobeSmile video intervention among refugees in Germany. It was hypothesized that participants receiving the intervention would show greater improvements in plaque index (PI) and gingival index (GI), as well as in selected oral health literacy outcomes, than participants in the control group.

Methods

Trial Design

This study was designed as a 2-arm, parallel-group, superiority randomized controlled trial with individual randomization and a 1:1 allocation ratio. The trial was conducted in 2 municipal refugee shelters in Freiburg, Germany, between November 2023 and February 2024. Outcomes and assessment time points were defined a priori, and no protocol modifications were made after trial initiation.

The study was approved by the ethics committee of the University of Freiburg prior to participant enrollment. Due to

XSL•FO
RenderX

administrative delays, registration in the German Clinical Trials Register (DRKS00032017) was completed after recruitment and follow-up had been completed; therefore, the trial is classified as retrospectively registered. No changes were made after study initiation to the primary objective, primary end point, eligibility criteria, intervention, allocation procedure, or core study design.

Participants

Eligible participants were adults (aged ≥18 years) residing in the participating shelters who were able to understand at least 1 of the languages in which the intervention was available (Arabic, Kurdish, Persian [*Dari* and *Farsi*], or Ukrainian) and had at least 10 teeth per jaw to allow reliable clinical index assessment. Individuals with acute dental emergencies or those requiring antibiotic prophylaxis were excluded. All eligible residents present during the recruitment period were invited to participate. A total of 86 participants met the inclusion criteria, provided written informed consent, and were enrolled. Participants with clinically relevant findings during examination received individual treatment recommendations and, if necessary, were referred to the university dental clinic to avoid delays in necessary care.

Randomization and Allocation

Randomization was performed after baseline assessment using a stratified allocation procedure to balance groups by age and native language. The allocation sequence was generated independently by the statistician (KV), who was not involved in recruitment, intervention delivery, or clinical outcome assessment. Allocation was concealed from the examining dentist until baseline assessments had been completed.

Intervention

The intervention was developed through an iterative participatory process involving dental professionals, translators, and individuals with refugee experience recruited through collaborating shelter networks and community contacts. Development was informed by oral health literacy principles emphasizing plain language, visual demonstration, multilingual narration, and accessibility across different literacy levels. Draft scripts and visual materials underwent repeated refinement based on feedback regarding comprehensibility and usability within refugee shelter settings.

The GlobeSmile intervention [29] is publicly accessible online via YouTube (Alphabet Inc) and consists of a short, multilingual educational video providing simple instructions on daily oral hygiene practices. The full script is provided in Multimedia Appendix 1.

The educational content focuses on key preventive behaviors, including toothbrushing technique, the use of fluoride-containing toothpaste, brushing duration and sequence, tongue cleaning, and rinsing, and was informed by international oral hygiene recommendations [27,28]. The video has a total length of approximately 4 minutes and combines simple animations with multilingual voice-over explanations to ensure accessibility across different literacy levels. The final version was made available in Arabic, Kurdish, *Dari*, *Farsi*, and Ukrainian. The control group received access to the video after completion of the follow-up to ensure ethical equipoise.

Implementation and Outcomes

The primary outcomes were changes in the PI (according to Silness and Loe [30]) and GI (according to Loe and Silness [31]). Secondary outcomes were assessed using a study-specific multilingual questionnaire developed for this field setting, addressing oral hygiene behaviors as well as oral health literacy–related cognitions.

Follow-up was scheduled approximately 2 months after baseline. This interval was chosen pragmatically, as longer retention in communal shelter settings could not be assumed, while still allowing sufficient time for the adoption of oral hygiene behaviors and the detection of short-term changes in plaque accumulation and gingival inflammation.

Clinical examinations were conducted at baseline and follow-up under standardized field conditions. At both time points, the PI and GI were assessed.

All examinations were conducted by a single trained and calibrated examiner (FS) using portable dental equipment and headlamps to ensure comparable examination conditions across time points. Prior to data collection, the examiner underwent calibration with an experienced practitioner to ensure measurement reliability (interrater reliability: PI Cohen $\kappa > 0.8$ and GI Cohen $\kappa > 0.8$).

Caries experience (Decayed, Missing, and Filled Teeth index according to Klein and Palmer [32]) was recorded at baseline as a descriptive characteristic.

Participants in the intervention group viewed the video on a tablet computer immediately after the baseline clinical examination in one of the available intervention languages that they reported understanding. They also received a printed QR code for continued access, allowing them to rewatch the video as often as desired during the 2-month study period (Figures S1 and S2 in Multimedia Appendix 1). Participants in the control group received no intervention during the trial but were provided with access to the video via a QR code after completion of the follow-up examination.

Questionnaires were administered in the participants' native languages at baseline and follow-up, with assistance from trained translators to ensure comprehension and cultural appropriateness. The survey included items on demographic information (age, sex, country of origin, native language, and education); refugee experience; oral health knowledge, attitudes, and perceived self-efficacy; oral hygiene behavior (brushing frequency and utensils); nutrition (according to Woelber et al [33]); physical activity; and tobacco and alcohol use (Tables S1 and S2 in Multimedia Appendix 1). The questionnaire was developed specifically for this multilingual field study and included behavioral, lifestyle-related, and oral health literacy–related items informed by oral health behavior literature, anti-inflammatory dietary recommendations, and the Internal-External Locus of Control Short Scale–4. Translation was supported by trained translators and reviewed for comprehensibility and cultural appropriateness. The

XSL•FO
RenderX

questionnaire was pilot-tested informally prior to implementation and was not intended as a formally validated oral health literacy instrument.

Blinding

Participant blinding was not feasible due to the nature of the intervention. At follow-up, clinical assessments were performed before participants were asked about intervention use, such that the examiner was not routinely aware of group allocation. However, due to the shared shelter setting, complete assessor blinding could not be guaranteed. Examinations were conducted by a single calibrated examiner under standardized conditions, and data analysis was performed using coded group labels.

Sample Size

No formal a priori sample size calculation was performed; the sample size was determined pragmatically by the number of eligible shelter residents who were present during the recruitment period and willing to participate. All individuals who met the inclusion criteria and provided written informed consent were enrolled. On the basis of the observed variability of the PI, which represented the primary outcome expected to change most rapidly, 40 participants per group were sufficient to detect moderate between-group differences in PI change with 90% power at a 2-sided significance level of α=.05.

Statistical Analysis

Continuous outcomes are reported as mean (SD) to aid interpretability. For univariate analyses, within-group changes in PI and GI were assessed using Wilcoxon signed-rank tests. Between-group comparisons of change scores were conducted using 2-sided *t* tests as the primary contrasts. Multivariable linear regression models were used to examine associations of age, sex, smoking status, and group allocation with changes in PI and GI and additionally served as sensitivity analyses for the nonparametric results. Analyses were performed in Stata (version 19; StataCorp), with a significance level of α=.05.

Ethical Considerations

Ethics approval was obtained from the Ethics Committee of the University Medical Center Freiburg (23‑1392-S1) prior to participant enrollment. All participants provided written informed consent before participation. Study information and consent materials were provided in relevant participant languages, with translator support where needed to ensure comprehension. Participant data were pseudonymized at the time of collection and analyzed in deidentified form to ensure confidentiality and data protection. No financial compensation was provided. Participants with clinically relevant findings during examination were informed and referred for appropriate dental care.

Results

Participant Flow and Retention

A total of 86 residents of the 2 participating shelters were assessed for eligibility, all of whom met the inclusion criteria, consented, and were enrolled in the study. Of these, 42 (48.8%) participants were randomized to the intervention group and 51.2% (44/86) to the control group. Follow-up at 2 months was completed by 83 (97%) participants. In total, 3.5% (3/86) participants were lost to follow-up (2/42, 4.8%) in the intervention group and 1 (2.3%) in the control group. No missing clinical or questionnaire data were observed among participants who completed follow-up (Figure 1).

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram of participant recruitment, randomization, allocation, follow-up, and analysis in a 2-arm parallel randomized controlled trial evaluating a multilingual oral health microlearning intervention among adults living in 2 municipal refugee shelters in Freiburg, Germany, between November 2023 and February 2024.

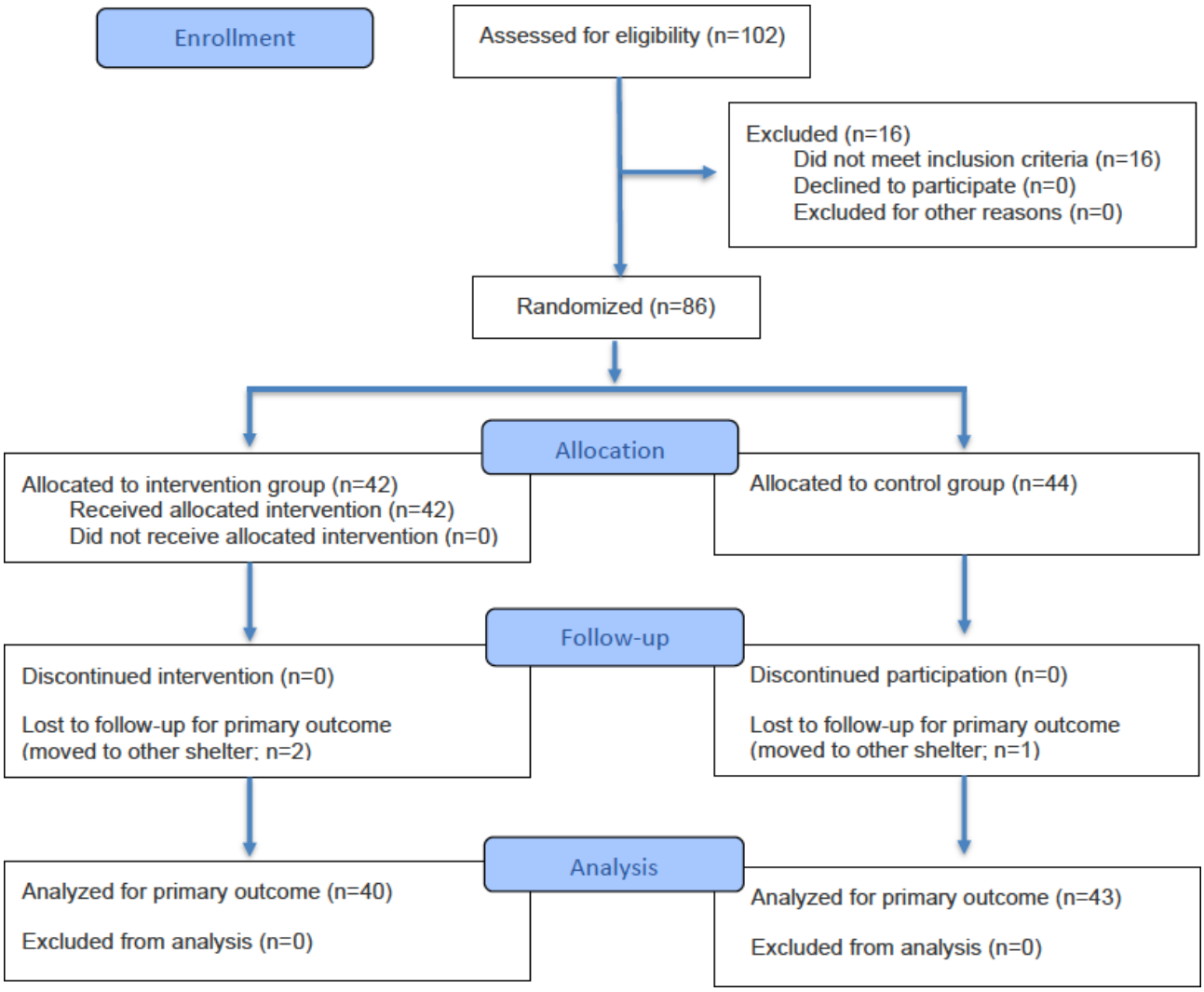

Baseline Characteristics

Baseline sociodemographic and clinical characteristics were broadly comparable between groups (Table 1). Participants had a mean age of 35.4 (SD 12.6) years, and 54.7% (47/86) of participants were female. The largest subgroups originated from Ukraine, Iraq, and Afghanistan, with native languages reflecting this distribution. All participants reported sufficient understanding of at least 1 of the intervention languages.

Educational level, lifestyle factors, and baseline clinical indices (PI; GI; and the Decayed, Missing, and Filled Teeth index) showed no relevant differences between groups.

XSL•FO
RenderX

Table . Baseline sociodemographic, behavioral, and clinical characteristics of participants randomized to the intervention or delayed-access control group in a randomized controlled trial conducted in refugee shelters in Freiburg, Germany.

Characteristics	Intervention group (n=42)	Control group (n=44)	Total participants (n=86)
Age (years), mean (SD; range)	36.0 (12.9; 18-67)	34.8 (12.4; 18-60)	35.4 (12.6; 18-67)
School education (years), mean (SD; range)	8.5 (1.9; 3-11)	8.4 (2.5; 3-12)	8.4 (2.2; 3-12)
Time since displacement (years), mean (SD; range)	6.7 (5.5; 0.2-16.0)	6.9 (5.9; 0.1-16.0)	6.8 (5.7; 0.1-16.0)
Sex, n (%)			
Male	18 (42.9)	21 (47.7)	39 (45.3)
Female	24 (57.1)	23 (52.3)	47 (54.7)
Country of origin, n (%)			
Ukraine	14 (33.3)	14 (31.8)	28 (32.6)
Iraq	11 (26.2)	11 (25)	22 (25.6)
Afghanistan	6 (14.3)	6 (13.6)	12 (14)
Kosovo	5 (11.9)	4 (9.1)	9 (10.5)
Nigeria	2 (4.8)	4 (9.1)	6 (7)
Georgia	2 (4.8)	3 (6.8)	5 (5.8)
Serbia	2 (4.8)	0 (0)	2 (2.3)
Macedonia	0 (0.0)	2 (4.5)	2 (2.3)
Native language, n (%)			
Ukrainian	14 (33.3)	14 (31.8)	28 (32.6)
Kurdish	11 (26.2)	11 (25)	22 (25.6)
Persian (*Dari* and *Farsi*)	6 (14.3)	6 (13.6)	12 (14)
Others	11 (26.2)	13 (29.5)	24 (27.9)
Oral health behaviors, n (%)			
Toothbrushing (≥2 times per day)	13 (31)	15 (34.1)	28 (32.6)
Use of interdental cleaning aids	15 (35.7)	16 (36.4)	31 (36)
Current smoker, n (%)	19 (45.2)	22 (50.0)	41 (47.7)

Primary Clinical Outcomes

Mean PI score decreased from 1.21 (SD 0.48) to 1.01 (SD 0.46) in the intervention group (mean change −0.21, SD 0.27; $P<.001$), whereas the control group showed only a small change (mean change −0.04, SD 0.17; $P=.23$). The between-group difference in PI change was significant ($P=.002$).

GI decreased in both groups (intervention group: mean change −0.11, SD 0.23; $P=.002$ and control group: mean change −0.07, SD 0.16; $P=.006$), but the between-group difference was not significant ($P=.55$). Table 2 summarizes all primary outcomes.

XSL•FO
RenderX

Table . Clinical oral health outcomes at baseline and 2-month follow-up among intervention and control participants in a randomized controlled trial of a multilingual oral health microlearning intervention[a].

Clinical outcome	Intervention group				Control group				Between-group *P* value
	T_0[b]	T_1[c]	Change (T_1–T_0)	*P* value	T_0	T_1	Change (T_1–T_0)	*P* value	
Plaque index, mean (SD)	1.21 (0.48)	1.01 (0.46)	−0.21 (0.27)	<.001	1.12 (0.39)	1.08 (0.34)	−0.04 (0.17)	.23	.002
Gingival index, mean (SD)	0.49 (0.49)	0.37 (0.44)	−0.11 (0.23)	.002	0.38 (0.35)	0.30 (0.29)	−0.07 (0.16)	.006	.55

[a]No significant differences between the sexes could be identified for plaque index or gingival index (Figure 2).

[b]T_0: baseline assessment.

[c]T_1: follow-up assessment.

XSL•FO
RenderX

Figure 2. Distribution of individual changes from baseline (T_0) to follow-up (T_1) in (**A**) plaque index and (**B**) gingival index, stratified by study group and sex among adults living in refugee shelters. Boxes indicate IQRs, horizontal lines indicate medians, whiskers indicate ranges excluding outliers, and dots represent individual observations.

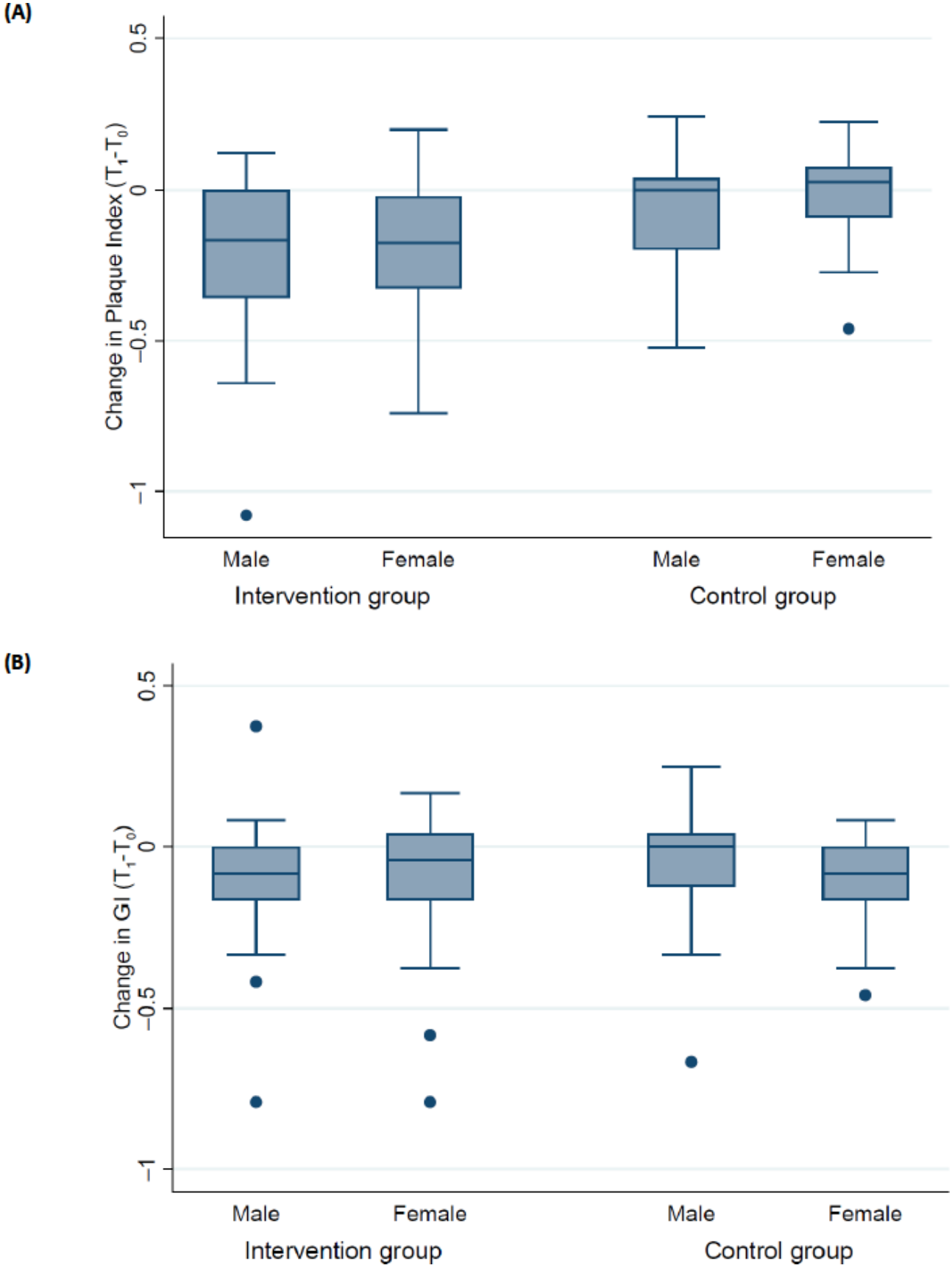

XSL•FO
RenderX

Secondary Outcomes: Oral Health Literacy Components

Questionnaire findings were mixed (Multimedia Appendix 1). Brushing frequency improved significantly, while selected cognitive items showed favorable but nonsignificant trends. Other items remained largely unchanged.

Additional Analyses and Effect Modifiers

Exploratory adjusted regression analyses were conducted to assess whether participant characteristics were associated with changes in clinical outcomes. In these models, group allocation remained significantly associated with PI change, confirming greater plaque reduction in the intervention group (β=0.1662, 95% CI 0.0697-0.2626; *P*=.001). No significant intervention effect was observed for GI change (β=0.0361, 95% CI −0.0504 to 0.1226; *P*=.41).

Age was significantly associated with greater improvements in both clinical outcomes. With each additional year of age, participants showed slightly greater reductions in PI (β=−0.0046, 95% CI −0.0087 to −0.0006; *P*=.02) and GI (β=−0.0042, 95% CI −0.0078 to −0.0005; *P*=.02). Sex and smoking status were not significantly associated with changes in PI or GI. These findings were consistent with the primary analyses and supported by nonparametric sensitivity analyses (Multimedia Appendix 1).

Intervention Adherence and Digital Engagement

Reported engagement with the video was generally high: 76.2% (32/42) of participants in the intervention group rewatched the video via the QR code at least once during the study period. Participants who reported repeated viewing tended to show larger reductions in PI, suggesting that repeated exposure to the content may be beneficial.

Discussion

Principal Findings

This randomized controlled trial evaluated whether a brief microlearning video could improve oral hygiene outcomes and oral health literacy among refugees living in communal shelters in Germany. The findings indicate that a simple digital intervention may improve plaque control and selected cognitive determinants of oral health behavior, even in settings with limited access to preventive care and substantial linguistic diversity. Oral health served as a pragmatic model for addressing health information barriers in an underserved population.

Interpretation of Findings

The intervention was associated with a greater reduction in PI than that observed in the control condition, suggesting improved oral hygiene performance independent of brushing frequency. In addition to a general Hawthorne effect, the baseline dental examination itself may have acted as a behavioral prompt, increasing oral hygiene awareness in both groups. The script recommends a brushing duration of approximately 3 minutes, which is slightly longer than commonly cited minimum recommendations. This was intentionally chosen as a behavioral strategy to encourage sufficiently thorough cleaning, acknowledging that actual brushing duration in everyday practice is often shorter than recommended. Video-based educational interventions have previously been shown to influence health behaviors across a range of contexts by facilitating observational learning and modeling of desired behaviors, supporting the plausibility of such mechanisms in this study [34]. One possible interpretation is that the intervention affected brushing technique rather than brushing frequency. Toothbrushing is often a habitual and partly automatic behavior, and previously learned routines may interfere with the acquisition of new oral hygiene skills [35]. Because PI and GI involve examiner judgment, the absence of fully blinded outcome assessment may have influenced estimates of effect size and should be considered when interpreting the findings. However, the examiner was not routinely aware of group allocation during follow-up assessments, which may have reduced the risk of observer bias.

Questionnaire findings were mixed; only brushing frequency improved significantly, while selected cognitive items showed favorable but nonsignificant trends. These findings suggest that broader oral health literacy–related effects should be interpreted cautiously. This interpretation is consistent with recent meta-analytic evidence indicating that psychologically informed interventions can improve oral hygiene behaviors, plaque levels, and self-efficacy in the short term by targeting cognitive and motivational determinants of oral health behavior [36]. These findings are consistent with established health literacy frameworks, in which increased knowledge, enhanced confidence, and a stronger internal health locus of control are key mechanisms linking educational interventions to improvements in oral health behaviors and clinical outcomes [18,24,34]. Therefore, the data support a plausible mechanism whereby culturally adapted microlearning enhances both oral health literacy and the application of brushing techniques, leading to reduced plaque accumulation.

Participants demonstrated high digital engagement. Three-quarters of the participants (32/42, 76.2%) rewatched the video at least once, and individuals who reported repeated viewing tended to show larger PI reductions. Although the frequency of repeated viewing was insufficient to support a formal dose–response analysis, the observed pattern suggests that repeated exposure may reinforce skill acquisition and message internalization. This finding is consistent with behavior change theory and supports the potential value of microdose repetition in digital health literacy interventions [23,36]. The study thereby highlights an important practical implication: even minimal recurrency of exposure may meaningfully enhance effectiveness.

Comparison With Existing Research

The results complement current international evidence documenting pronounced oral health inequities among refugees and asylum seekers, driven by structural access barriers, linguistic limitations, inconsistent entitlements to care, and fragmented service pathways [6,10,12,17]. Consistent with reports from Germany and other European countries, participants exhibited high baseline levels of untreated caries and low oral

XSL•FO
RenderX

health literacy, factors strongly associated with reduced service use and poorer health outcomes [10,18,24].

Digital health interventions have been proposed as a promising strategy to reduce inequities in forced migrant populations, particularly where conventional education is difficult to deliver due to low literacy, language heterogeneity, or limited access to health professionals. Similarly, behavioral interventions using visual cues and motivational messaging have demonstrated improvements in plaque control among schoolchildren with refugee and immigrant backgrounds, highlighting the potential of simple visual communication strategies for promoting oral self-care in underserved populations [37]. These findings are consistent with previous work suggesting that digital health literacy interventions can be beneficial in forced migrant populations [23]. These findings add experimental evidence to this literature and suggest that ultralow-threshold, multilingual microlearning tools can produce clinical improvements without requiring complex infrastructure or advanced digital skills.

Compared with multicomponent mobile health programs that demand extensive engagement, device access, and continuous coaching [26], the simplicity of the present intervention may offer greater scalability for settings such as shelters, reception centers, or transit facilities. Importantly, however, digital education alone cannot compensate for structural constraints, including the emergency-focused care entitlements that limit preventive dental treatment for refugees in Germany [10]. Therefore, the intervention should be viewed as a complementary public health strategy rather than a substitute for system-level reform.

The observed effects may in part reflect reinforcement of previously known oral health information rather than entirely novel knowledge acquisition. However, this distinction is less relevant from a public health perspective: even when basic knowledge exists, accessible, linguistically tailored microlearning may facilitate translation into daily practice and measurable clinical outcomes. Therefore, in populations facing structural and informational barriers, reinforcement and accessibility may represent key mechanisms of effective prevention.

Strengths and Limitations

Strengths of this study include its randomized controlled design, the use of calibrated clinical assessments, the high follow-up rate of 97%, and the naturalistic implementation within refugee shelters, which enhances ecological validity. The participatory development of the video, involving translators and individuals with refugee experience, likely contributed to cultural acceptability and comprehension.

Several limitations should be considered. First, the sample size was modest and determined pragmatically rather than through a formal a priori power calculation. Therefore, the study was primarily exploratory and may have been underpowered to detect smaller between-group differences, particularly for secondary outcomes and subgroup analyses. Second, the study population was heterogeneous with regard to country of origin, language, migration history, and social circumstances. Although this heterogeneity reflects the real-world composition of refugee shelters and strengthens the practical relevance of the findings, it may also have increased variability in responses to the intervention, limited the precision of effect estimates, and reduced the interpretability of subgroup patterns. Third, the 2-month follow-up period restricts conclusions regarding long-term behavioral change and the sustainability of effects. Fourth, some degree of contamination between groups cannot be entirely ruled out, given the shared living environment. Fifth, gingival outcomes are difficult to influence over a short period without professional prophylaxis, which may partly explain the absence of a significant between-group effect for GI. Complete assessor blinding could not be guaranteed, and some risk of observer bias remains. Questionnaire outcomes were assessed using a pragmatic study-specific instrument rather than a formally validated oral health literacy scale, which limits precision, reliability, and comparability with other studies. Finally, adherence was assessed by self-report, and objective use metrics for repeated video viewing were not available.

Implications for Public Health and Policy

This trial provides actionable evidence supporting the implementation of multilingual digital microlearning as part of preventive strategies for displaced populations. Although evaluated here in the context of oral health, the intervention model may be transferable to other preventive health domains in which multilingual, low-threshold access to information is limited. The intervention was feasible to deliver, well accepted, and associated with a meaningful improvement in plaque control and limited favorable changes in selected self-reported outcomes. Because it is low cost, language adaptable, and accessible via QR code, it could be readily integrated into refugee shelters, reception centers, community health programs, and nongovernmental organization–led services.

From a health equity perspective, the results indicate that well-designed tools can reduce information barriers, thereby supporting the overarching aims of Inclusion Health, which focuses on improving health outcomes and access to care for socially excluded populations. Nevertheless, digital education should be embedded within broader multisectoral approaches that address restricted care entitlements, financial and administrative barriers, and limited access to preventive dental services for refugees [6,10,18]. Without structural reforms, improvements in knowledge and technique may not translate into long-term disease reduction. Because the intervention requires minimal infrastructure and can be distributed via QR codes or standard video platforms, it represents a highly scalable digital public health strategy that could be adapted to other preventive health domains and multilingual populations.

Detailed information on previous dental attendance and prior exposure to oral health education interventions was not systematically collected. These factors may have influenced baseline behaviors and responsiveness to the intervention. Future trials should assess prior dental care use and previous preventive education exposure as potential moderators of intervention effects.

XSL•FO
RenderX

Future Research

This study illustrates how even ultrabrief digital interventions can produce measurable clinical change when delivered in linguistically adapted formats. The model may be transferable to other preventive domains where information barriers limit health equity. The GlobeSmile video was developed without external funding and relied on deliberately simplified, self-produced visual materials to ensure rapid and low-cost implementation. Although this didactic reduction enhances accessibility and scalability, future iterations could explore more professionally designed or visually enriched formats to assess whether enhanced aesthetic and pedagogical features further increase engagement and effectiveness. Future studies should evaluate long-term outcomes, assess cost-effectiveness, and determine whether combining video-based education with in-person demonstrations or support from community health workers provides additional benefits. Research across diverse refugee settings and language groups could assess generalizability and inform implementation strategies. Additionally, investigating the mechanisms linking improved oral health literacy to clinical change could support theory-based optimization of digital microlearning interventions.

Conclusions

Brief multilingual digital microlearning may represent a feasible and scalable strategy to improve plaque control and reduce health information barriers in refugee shelter settings. Larger, prospectively powered trials with longer follow-up periods are warranted.

Acknowledgments

The authors sincerely thank all participants for their time, trust, and valuable contributions to this study. The authors are grateful to the management and staff of the refugee shelters for their support during recruitment and data collection. Special thanks are dedicated by the authors to Anna Hardt for her creative illustrations that enriched the educational video and to the translators with refugee backgrounds, whose expertise and cultural sensitivity were essential for ensuring accessibility and linguistic accuracy. Their engagement and commitment greatly contributed to the success of this project.

Funding

The authors declared that no financial support was received for this work.

Data Availability

Deidentified participant data, the data dictionary, and the study protocol may be made available upon reasonable request to the corresponding author, subject to applicable ethical, legal, and data protection requirements, beginning at the time of publication and for up to 10 years thereafter. Requests will be considered on a case-by-case basis.

Authors' Contributions

Conceptualization: ALH, DRR
Data curation: FS, KV
Formal analysis: KV
Investigation: FS, MKM
Methodology (including video design and language adaptation): ALH, DRR, KV, SS
Project administration: ALH
Resources: BCS
Supervision: ALH, BCS, DRR
Validation (examiner calibration): MKM
Visualization (figures and video visuals): FS, SS, KV
Writing—original draft: FS, MKM, ALH, AAB
Writing—review and editing: ALH, FS, MKM, DRR, SS, BCS, KV, AAB, GJ

Conflicts of Interest

None declared.

Editorial Notice

This randomized study was retrospectively registered, which the authors explained as resulting from administrative delays prior to recruitment. The editor granted an exception from the International Committee of Medical Journal Editors' rules mandating prospective registration of randomized trials because the risk of bias appears low, and the design and outcomes did not change from those outlined in the original protocol approved by the ethics committee prior to recruitment. However, readers are advised to carefully assess the validity of any potential explicit or implicit claims related to primary outcomes or effectiveness, as retrospective registration does not prevent authors from changing their outcome measures retrospectively.

XSL•FO
RenderX

Multimedia Appendix 1
Supplementary materials related to the GlobeSmile multilingual oral health micro-learning intervention, including the intervention video script, oral health questionnaire, and outcome measures, QR-code access card used for participant access to the intervention.
[DOCX File, 1527 KB - jmir_v28i1e95562_app1.docx]

Checklist 1
CONSORT checklist.
[PDF File, 254 KB - jmir_v28i1e95562_app2.pdf]

References

1. Kickbusch I, Piselli D, Agrawal A, et al. The Lancet and Financial Times Commission on governing health futures 2030: growing up in a digital world. Lancet 2021 Nov;398(10312):1727-1776. [doi: 10.1016/S0140-6736(21)01824-9]
2. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2026-06-02]
3. Robertshaw L, Dhesi S, Jones LL. Challenges and facilitators for health professionals providing primary healthcare for refugees and asylum seekers in high-income countries: a systematic review and thematic synthesis of qualitative research. BMJ Open 2017 Aug 4;7(8):e015981. [doi: 10.1136/bmjopen-2017-015981] [Medline: 28780549]
4. United Nations High Commissioner for Refugees. Global Trends: Forced Displacement in 2020: United Nations; 2021. [doi: 10.18356/9789211066005]
5. Global trends. UNHCR. URL: https://www.unhcr.org/global-trends [accessed 2026-03-16]
6. Paisi M, Baines R, Burns L, et al. Barriers and facilitators to dental care access among asylum seekers and refugees in highly developed countries: a systematic review. BMC Oral Health 2020 Nov 25;20(1):337. [doi: 10.1186/s12903-020-01321-1] [Medline: 33238954]
7. Ziersch A, Due C, Walsh M, Peres MA. Oral health, quality of life and general health of people from refugee and asylum-seeking backgrounds from South Asia and the Middle East resettled in South Australia: a cross sectional survey. BMC Public Health 2025 May 15;25(1):1796. [doi: 10.1186/s12889-025-22907-w] [Medline: 40375194]
8. Sheiham A, Watt RG. The common risk factor approach: a rational basis for promoting oral health. Community Dent Oral Epidemiol 2000 Dec;28(6):399-406. [doi: 10.1034/j.1600-0528.2000.028006399.x] [Medline: 11106011]
9. Jordan AR, Meyer-Lueckel H, Kuhr K, Sasunna D, Bekes K, Schiffner U. Caries experience and care in Germany: results of the 6th German Oral Health Study (DMS • 6). Quintessence Int 2025 Mar 17;56(11):S30-S39. [doi: 10.3290/j.qi.b5986212] [Medline: 40091720]
10. Al-Ani A, Takriti M, Schmoeckel J, Alkilzy M, Splieth CH. National oral health survey on refugees in Germany 2016/2017: caries and subsequent complications. Clin Oral Investig 2021 Apr;25(4):2399-2405. [doi: 10.1007/s00784-020-03563-3] [Medline: 33011846]
11. Wenner J, Biddle L, Gottlieb N, Bozorgmehr K. Inequalities in access to healthcare by local policy model among newly arrived refugees: evidence from population-based studies in two German states. Int J Equity Health 2022 Jan 24;21(1):11. [doi: 10.1186/s12939-021-01607-y] [Medline: 35073919]
12. Bhusari S, Ilechukwu C, Elwishahy A, Horstick O, Winkler V, Antia K. Dental caries among refugees in Europe: a systematic literature review. Int J Environ Res Public Health 2020 Dec 18;17(24):9510. [doi: 10.3390/ijerph17249510] [Medline: 33353039]
13. Khoshnevisan MH, Albujeer AN, Taher A, Almahafdha A. Dental education in Iraq: issues, challenges and future. J Contemp Med Sci 2017;3(11):260-263. [doi: 10.22317/jcms.09201703]
14. Kanjevac T. Dental services funding and affordability in Serbia - decade-long perspective. Front Public Health 2015;3:145. [doi: 10.3389/fpubh.2015.00145] [Medline: 26125016]
15. Lekhan V, Rudiy V, Richardson E. Ukraine: health system review. : European Observatory on Health Systems and Policies; 2010 URL: https://iris.who.int/server/api/core/bitstreams/b9f22b61-6411-4e0e-bc80-b1ec0d39b23a/content [accessed 2026-06-02]
16. Rolke K, Walter J, Weckbecker K, Münster E, Tillmann J. Identifying gaps in healthcare: a qualitative study of Ukrainian refugee experiences in the German system, uncovering differences, information and support needs. BMC Health Serv Res 2024 May 4;24(1):585. [doi: 10.1186/s12913-024-11052-6] [Medline: 38704571]
17. Zinah E, Al-Ibrahim HM. Oral health problems facing refugees in Europe: a scoping review. BMC Public Health 2021 Jun 24;21(1):1207. [doi: 10.1186/s12889-021-11272-z] [Medline: 34162351]
18. Valdez R, Spinler K, Kofahl C, et al. Oral health literacy in migrant and ethnic minority populations: a systematic review. J Immigr Minor Health 2022 Aug;24(4):1061-1080. [doi: 10.1007/s10903-021-01266-9] [Medline: 34448993]
19. Luai AF, Malek AA, Hassan EH, Sabri BA, Radzi NA. Effectiveness of immersive technologies in oral health promotion and education: a systematic review. J Dent 2024 Nov;150:105324. [doi: 10.1016/j.jdent.2024.105324] [Medline: 39191319]
20. Gunaratnam P, Sestakova L, Smith M, Torvaldsen S. Evaluation of a multilingual oral health DVD for newly arrived refugees. Health Promot J Austr 2013 Aug;24(2):159-159. [doi: 10.1071/HE13070]

XSL•FO
RenderX

21. Carpenter SM, Greer ZM, Newman R, Murphy SA, Shetty V, Nahum-Shani I. Developing message strategies to engage racial and ethnic minority groups in digital oral self-care interventions: participatory co-design approach. JMIR Form Res 2023 Dec 11;7:e49179. [doi: 10.2196/49179] [Medline: 38079204]
22. Mariño RJ, Marwaha P, Barrow SY. Web-based oral health promotion program for older adults: development and preliminary evaluation. Int J Med Inform 2016 Jul;91:e9-e15. [doi: 10.1016/j.ijmedinf.2016.04.002] [Medline: 27117812]
23. Yameogo AR, Délétroz C, Sasseville M, et al. Effectiveness of interventions to improve digital health literacy in forced migrant populations: mixed methods systematic review. J Med Internet Res 2025 Jun 25;27:e69880. [doi: 10.2196/69880] [Medline: 40561467]
24. Lieske B, Kofahl C, Borof K, et al. Oral health literacy and oral health status of a German adult population with migration background - findings from the MuMi Study. J Migr Health 2025;11:100309. [doi: 10.1016/j.jmh.2025.100309] [Medline: 40066255]
25. Muller R, Bilich L, Jones M. Impact of an oral health education program on the oral health literacy of refugees. J Immigr Minor Health 2024 Aug;26(4):699-710. [doi: 10.1007/s10903-024-01594-6] [Medline: 38589753]
26. Weil MT, Spinler K, Lieske B, et al. An evidence-based digital prevention program to improve oral health literacy of people with a migration background: intervention mapping approach. JMIR Form Res 2023 May 11;7:e36815. [doi: 10.2196/36815] [Medline: 37166956]
27. Glenny AM, Walsh T, Iwasaki M, et al. Development of tooth brushing recommendations through professional consensus. Int Dent J 2024 Jun;74(3):526-535. [doi: 10.1016/j.identj.2023.10.018] [Medline: 38052700]
28. How to practice good oral hygiene. FDI World Dental Federation. URL: https://www.fdiworlddental.org/how-practice-good-oral-hygiene [accessed 2025-09-06]
29. GlobeSmile YouTube page. YouTube. URL: https://www.youtube.com/@globesmile7359 [accessed 2026-04-30]
30. Silness J, Loe H. Periodontal disease in pregnancy. II. Correlation between oral hygiene and periodontal condition. Acta Odontol Scand 1964 Feb;22:121-135. [doi: 10.3109/00016356408993968] [Medline: 14158464]
31. Loe H, Silness J. Periodontal disease in pregnancy. I. Prevalence and severity. Acta Odontol Scand 1963 Dec;21:533-551. [doi: 10.3109/00016356309011240] [Medline: 14121956]
32. Klein H, Palmer CE. Studies on dental caries: X. A procedure for the recording and statistical processing of dental examination findings. J Dent Res 1940;19(3):243-256. [doi: 10.1177/00220345400190030401]
33. Woelber JP, Bremer K, Vach K, et al. An oral health optimized diet can reduce gingival and periodontal inflammation in humans - a randomized controlled pilot study. BMC Oral Health 2016 Jul 26;17(1):28. [doi: 10.1186/s12903-016-0257-1] [Medline: 27460471]
34. Tuong W, Larsen ER, Armstrong AW. Videos to influence: a systematic review of effectiveness of video-based education in modifying health behaviors. J Behav Med 2014 Apr;37(2):218-233. [doi: 10.1007/s10865-012-9480-7] [Medline: 23188480]
35. Thavarajah R, Kumar M, Mohandoss AA, Vernon LT. Drilling deeper into tooth brushing skills: is proactive interference an under-recognized factor in oral hygiene behavior change? Curr Oral Health Rep 2015 Sep;2(3):123-128. [doi: 10.1007/s40496-015-0053-z] [Medline: 26457238]
36. He IL, Liu P, Wong MC, Chu CH, Lo EC. Effectiveness of psychological intervention in improving adolescents' oral health: a systematic review and meta-analysis. J Dent 2024 Nov;150:105365. [doi: 10.1016/j.jdent.2024.105365] [Medline: 39362300]
37. Banihashem Rad SA, Esteves-Oliveira M, Kazemian A, et al. Evaluating the effectiveness of the Nudge theory in improving the oral self-care of schoolchildren with refugee and immigrant backgrounds in Mashhad, Iran. Dent J (Basel) 2024;12(7):228. [doi: 10.3390/dj12070228]

Abbreviations

GI: gingival index
PI: plaque index

Edited by A Schwartz; submitted 17.Mar.2026; peer-reviewed by NC Krupa, O Owolo; revised version received 22.May.2026; accepted 25.May.2026; published 15.Jun.2026.

Please cite as:
Müller MK, Schulz F, Jogho G, Steffens S, Vach K, Baradee AA, Reissmann DR, Spies BC, Hillebrecht AL
A Multilingual Digital Microlearning Intervention for Oral Health in Refugee Shelters: Randomized Controlled Trial
J Med Internet Res 2026;28:e95562
URL: https://www.jmir.org/2026/1/e95562
doi:10.2196/95562

XSL•FO
RenderX

© Maxi Katharina Müller, Fabian Schulz, George Jogho, Simone Steffens, Kirstin Vach, Ahmad Amro Baradee, Daniel R Reissmann, Benedikt C Spies, Anna-Lena Hillebrecht. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Text Messaging Support for Patients Diagnosed With Impaired Glucose Tolerance During Pregnancy: Nonrandomized Pre-Post Implementation Study Assessing Impact on Postpartum Transitions of Care

Natasha R Kumar[1], MD, MSCE; Whitney Bender[2], MD; Celeste Durnwald[3], MD

[1]Department of Obstetrics and Gynecology, Michigan Medicine, University of Michigan, 1500 E Medical Center Drive, Ann Arbor, MI, United States
[2]Department of Obstetrics and Gynecology, Sidney Kimmel Medical College, Thomas Jefferson University Hospital, Philadelphia, PA, United States
[3]Department of Obstetrics and Gynecology, Perelman School of Medicine, University of Pennsylvania Health System, Philadelphia, PA, United States

Corresponding Author:
Natasha R Kumar, MD, MSCE
Department of Obstetrics and Gynecology, Michigan Medicine, University of Michigan, 1500 E Medical Center Drive, Ann Arbor, MI, United States

Abstract

Background: Patients with impaired glucose tolerance (IGT) identified during pregnancy who do not develop gestational diabetes mellitus (GDM) often do not receive additional interventions for their long-term metabolic risks.

Objective: This nonrandomized pre-post implementation study reports the design process and initial program evaluation for Better Follow-up of Impaired Glucose Tolerance (BRIDGE), a 12-week text-based postpartum support program promoting hemoglobin A_{1c} (HbA_{1c}) completion and primary care provider (PCP) visit scheduling for patients diagnosed with IGT during pregnancy, assessing improvement in desired postpartum transition milestones. The 19-month program was divided into 2 arms lasting 9.5 months each, BRIDGE− (SMS text messaging support alone; October 2021-July 2022) and BRIDGE+ (SMS text messaging and IGT-focused postpartum visit; July 2022-April 2023). We aimed to assess whether BRIDGE improved desired postpartum transition milestones.

Methods: Patients were eligible for BRIDGE if they received prenatal care at the study site (a northeastern US academic tertiary care center), were diagnosed with IGT during pregnancy, never developed GDM, and could receive English text messages. We performed a program evaluation using a pre/postimplementation design, comparing outcomes for the BRIDGE population to a 19-month historical population. Primary outcomes were (1) completion of HbA_{1c} testing by 1 year postpartum and (2) PCP visit scheduling by 12 weeks postpartum. A comparative analysis between BRIDGE− and BRIDGE+ was performed. Multivariable logistic regressions controlled for the history of IGT after stepwise backward elimination.

Results: In the program evaluation, 503 individuals were included (n=342 in historical population, n=82 in BRIDGE− population, and n=79 in BRIDGE+ population), with similar demographic and clinical characteristics across populations. A total of 212 individuals were screened for eligibility in BRIDGE, and 161 individuals participated in the program. BRIDGE participants had increased odds of HbA_{1c} completion by 1 year postpartum (39.8% vs 12.5%; adjusted odds ratio [aOR] 4.28, 95% CI 2.71 - 6.78) and PCP visit scheduling (31.0% vs 12.0%; aOR 9.58, 95% CI 4.39 - 20.9) compared to the historical population. BRIDGE+ patients were more likely to complete HbA_{1c} testing by 12 weeks postpartum than BRIDGE− participants. Most patients attended scheduled PCP visits, but rates of IGT counseling at PCP visits were low.

Conclusions: Individuals with IGT rarely receive targeted interventions during pregnancy or delivery hospitalization. This innovative study demonstrates that individuals with IGT have high rates of uptake for postpartum SMS text messaging support, which tripled completion rates of HbA_{1c} screening within 1 year postpartum and doubled the scheduling rate for PCP visits by 12 weeks postpartum. While attendance at scheduled PCP visits was very high, <60% of PCP visits included IGT counseling, highlighting key improvement areas in the quality of postpartum transitions to primary care. While a randomized trial is needed to ascertain definitive impact, SMS text messaging support may be an effective tool to improve postpartum transitions of care for this underserved population.

(J Med Internet Res 2026;28:e76493) doi:10.2196/76493

KEYWORDS

postpartum transitions of care; impaired glucose tolerance; obstetrics; quality improvement; mobile health interventions

XSL•FO
RenderX

Introduction

Background

According to the 2020 National Diabetes Statistics Report, approximately 1 in 3 American adults has impaired glucose tolerance (IGT) or prediabetes, defined as hemoglobin A_{1c} (HbA_{1c}) of 5.7 to 6.4 [1]. IGT is often considered an intermediate state of abnormal glucose metabolism that exists between normal glucose homeostasis and diabetes. The natural history of IGT is variable and may include normalization, persistence of impaired tolerance, or progression to diabetes [2-4]. Up to 70% of patients with IGT may ultimately develop diabetes without intervention [5]. Given the variable natural history and the ability to alter disease trajectory with behavioral modifications, individuals with IGT are ideal candidates for diabetes prevention efforts [6,7]. Pregnancy is a time when individuals enter the health care system and may be motivated to focus on their health. As a result, obstetricians can play a role in promoting overall health, impacting both future pregnancy and long-term health outcomes.

As the prevalence of preexisting and gestational diabetes mellitus (GDM) among pregnant people has increased in recent years [8,9], universal HbA_{1c} screening in early pregnancy has been proposed to identify undiagnosed pre-existing diabetes as well as to screen for early-onset GDM [10,11]. In a retrospective cohort at a single institution, 12% of patients had IGT in early pregnancy but never developed GDM [12]. Patients with IGT in pregnancy are at risk of both adverse pregnancy outcomes and long-term health outcomes, even if they never develop GDM. Although there are relatively few studies of pregnant patients with IGT and without GDM, several studies have demonstrated an increased risk of spontaneous preterm birth less than 37 weeks [13]. A singular cohort study has also shown a potentially increased risk of hypertensive disorders of pregnancy (HDP), major congenital anomalies, shoulder dystocia, large for gestational age births, and perinatal death in this population [13]. The American Diabetes Association recommends that patients with IGT complete HbA_{1c} screening for the development of diabetes every 1 to 2 years [14]. Pregnant patients diagnosed with GDM receive counseling on dietary and physical activity recommendations, antenatal glucose monitoring, counseling on long-term metabolic risks, and postpartum glucose testing. In contrast, there are no established guidelines for the prenatal or postpartum management of pregnant patients with IGT and without GDM. They may not receive dietary and exercise counseling, counseling on long-term metabolic risks, postpartum glucose testing, or a formal transition of care to a primary care physician or endocrinologist. This can create gaps in care with implications for future pregnancy and future health outcomes. In previously unpublished data from our institution, an exploratory analysis of long-term outcomes for 100 patients with IGT in pregnancy at a single institution demonstrated that 71% of patients did not have a visit with a primary care provider (PCP) in 1 to 3 years after delivery. Of the 29 patients with PCP visits, nearly half (48.3%) did not obtain repeat HbA_{1c} values to determine if there was persistence of IGT. In half of the cases, this was due to a lack of ordering by the PCP. Of the 15 patients who completed repeat HbA_{1c} testing, 93.3% remained more than 5.7% 1 to 3 years after delivery. There were also 14 repeat pregnancies in this cohort, 78.6% of which continued to demonstrate IGT during early pregnancy screening.

SMS text messaging support programs have been used effectively during the postpartum period to support transitions of care. A randomized controlled trial of an SMS text messaging support program prompting primary care transitions demonstrated significant improvement in primary care visit completion, as well as fewer postpartum readmissions and increased receipt of services, such as blood pressure assessments and screenings for mood symptoms [15]. Other SMS text messaging support programs have focused on the completion of recommended clinical assessments postpartum for patients with pregnancy complications, such as hypertensive disorders of pregnancy or gestational diabetes, with many programs reporting improved engagement with SMS text messaging support [16-20].

Based on our institutional data demonstrating clinical need, our team developed a program called Better Follow-up of Impaired Glucose Tolerance (BRIDGE) to address postpartum transitions of care for patients diagnosed with IGT during pregnancy. Here, we present a single institution nonrandomized pre-post implementation study, including (1) the design process for BRIDGE and a program evaluation, (2) a comparison of clinical outcomes for BRIDGE patients to a historical population, and (3) a comparative analysis examining differences in outcomes between 2 programmatic strategies used within BRIDGE. We hypothesize that the addition of SMS text messaging support and an IGT-focused postpartum visit will improve the completion of desired postpartum milestones for this population (ie, HbA_{1c} completion, primary care visit scheduling). Based on the findings noted in this study, we anticipate performing a future randomized trial to definitively assess the impact of this intervention.

Intervention

Overview of Program Design

Our team developed a program called BRIDGE to address postpartum transitions of care for patients diagnosed with IGT during pregnancy. The first half of BRIDGE (BRIDGE−; October 2021-July 2022) enrolled participants in biweekly text messages between 0 and 12 weeks postpartum, including educational content and reminders to complete desired postpartum transition milestones, namely (1) completing HbA_{1c} assessment and (2) scheduling a visit with a PCP. During the last half of the program (BRIDGE+; July 2022-April 2023), participants were offered an additional postpartum visit for standardized IGT education in addition to the SMS text messaging support.

Text Message Design and Content

The BRIDGE SMS text messaging program was designed in collaboration with multiple expert groups within our health care institution, as well as lived-experience experts. First, we partnered with Way2Health (W2H), a web-based platform

XSL•FO
RenderX

providing technology infrastructure for sustainable behavior change interventions to build the SMS text messaging support program [21-23]. W2H infrastructure is integrated into the electronic medical record (EMR) system at our institution, which allows providers to confirm enrollment in BRIDGE and review text message content if desired. To optimize the frequency and content of messaging in the program, we collaborated with the Nudge Unit, a behavioral design team embedded within our institution's health care system (Figure 1). Finally, we reviewed the planned messages with a lived experience expert to confirm that the language and content of the messaging were appropriate and affirming.

Figure 1. Overview of Better Follow-up of Impaired Glucose Tolerance (BRIDGE) text message content and timing. HbA_{1c}: hemoglobin A_{1c}.

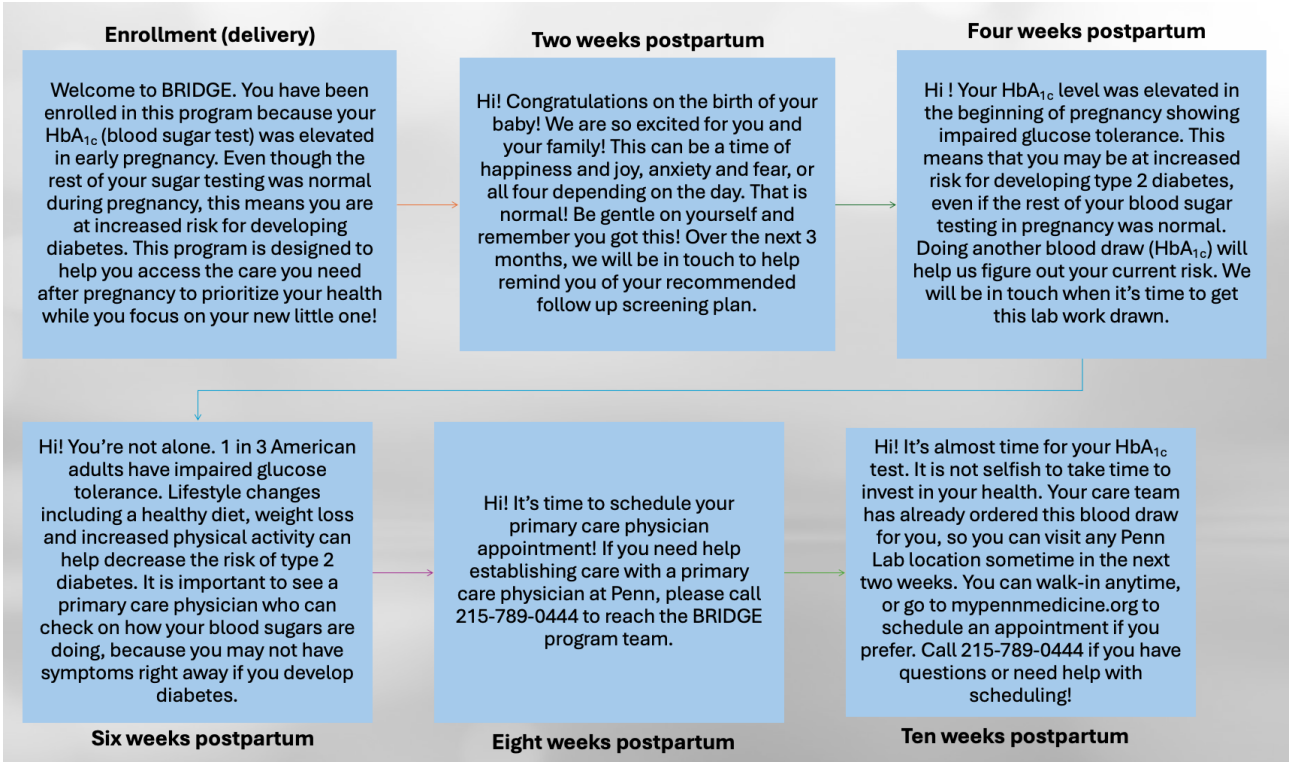

Primary Care Physician Scheduling and Follow-Up

Our study team held preimplementation planning meetings with both our prenatal care provider teams and the primary care teams to optimize workflows for patient enrollment and engagement with the program.

To facilitate participants' success with expected milestones, we developed an expedited pathway for PCP referral through the BRIDGE program within our health care system. We used a referral order with a specific indication related to BRIDGE within our EMR. These BRIDGE referrals to primary care were reviewed by the administrative team for the primary care offices on a bimonthly or monthly basis to expedite scheduling.

Monthly implementation meetings were held to review programmatic outcomes, and certain implementation strategies were added or removed to address specific barriers communicated by our prenatal care providers or PCPs, as well as patients. For example, referrals were integrated into a virtual social work support program offered by the institution to address patients' reports regarding social drivers of health that were limiting their engagement with care.

Importantly, the ongoing implementation meetings led to a significant strategic shift at the midway point of the program where an IGT-focused postpartum visit (iPPV) was added, during which prenatal providers discussed the long-term metabolic risks associated with IGT and reiterated the importance of HbA_{1c} completion and PCP visit scheduling. The addition of the iPPV was based on feedback from both patients, who desired to continue care in settings where they had established trust during pregnancy, as well as the primary care team, which reported workforce capacity and insurance challenges. This marked the transition from BRIDGE− to BRIDGE+.

Methods

Pre-Post Implementation Study Design

This is a pre-post implementation study of BRIDGE, an SMS text messaging support program promoting postpartum transitions of care for patients with IGT during pregnancy. A program evaluation was performed assessing differences in outcomes between the BRIDGE population and a historical population, as well as between 2 groups enrolled in 2 different programmatic strategies of BRIDGE, as described previously in the *Overview of Program Design* section.

Participants

Patients were eligible for the program if they had an HbA_{1c} value of 5.7 to 6.4 prior to 20 weeks of gestation, completed third-trimester glucose testing, did not receive a diagnosis of GDM, had a singleton pregnancy, delivered at our hospital, and

XSL•FO
RenderX

were English-speaking. BRIDGE was implemented at a single tertiary care institution, with approximately 4000 deliveries per year, in the northeastern United States. Opt-out enrollment of eligible patients was performed manually by the study team at the time of their delivery. Patients received a text message with a description of the program, recommended milestones, and the option to decline participation.

For our historical population, we selected all patients who met identical eligibility criteria to BRIDGE participants and delivered between June 2018 and December 2019. We selected a continuous 19-month time period (identical to the BRIDGE program duration) that preceded the COVID-19 pandemic in March 2020 (given the disruptions to typical care during that time).

Outcomes

The 2 primary outcomes for the BRIDGE program were (1) completion of an HbA_{1c} test and (2) scheduling of a visit with a PCP by 12 weeks postpartum. Given the paucity of national recommendations regarding follow-up for IGT identified in early pregnancy, these postpartum transition milestones were determined based on expert opinions from our maternal fetal medicine, primary care, and endocrinology teams. When completing the comparative analysis with the historical population, HbA_{1c} completion was assessed by 1 year postpartum rather than by 12 weeks postpartum to provide an appropriate comparison with the standard of care during the historical period. For our analysis comparing the 2 programmatic strategies in BRIDGE− and BRIDGE+, we assessed HbA_{1c} completion by 12 weeks postpartum, which is the programmatic goal, as well as by 1 year postpartum, which is aligned with national recommendations from some organizations such as the American Diabetes Association [24]. To account for potential bias caused by missing data, because the authors were unable to use retrospective chart abstraction to identify data outside of the health care system, reporting of these outcomes was stratified among patients with an identified PCP within the network versus patients with an identified PCP outside of the network. In addition, the evaluation was adjusted for demographic and clinical characteristics that differed between the study groups.

Secondary outcomes included other process outcomes, such as PCP visit attendance and the inclusion of IGT counseling during PCP visits (for in-network patients where these data were available in our EMR). We also reported descriptive data regarding iPPV scheduling and attendance for BRIDGE+ participants only. We stratified all outcomes related to PCP visit scheduling and attendance by patients who had in-network PCPs versus all patients due to the insurance and workforce capacity challenges faced by patients who did not have established PCPs within our health care system, as well as considerations regarding potential missing data.

The data were collected by the study team on participant demographics and clinical characteristics, as well as primary and secondary outcomes, through an EMR review.

Sample Size

With a sample size of 246 patients, we would be powered to detect an improvement in PCP visit scheduling from an assumed baseline of 30% based on prior chart review at our institution as well as existing literature [25] to 45% with 80% power and an α level of .05. With an annual birth volume of 4000 patients per year and baseline data at the institution demonstrating that 12% of the population has IGT in early pregnancy, the study anticipated the need to collect data for at least 7 months to accrue a sufficient sample size. All analyses were performed using Stata version 17.0 (StataCorp LLC).

For this nonrandomized study, the CONSORT (Consolidated Standards of Reporting Trials) extension for pilot and feasibility trials was used to report the findings including the abstract extension [26,27].

Statistical Methods

Race and ethnicity were self-reported. For the primary analysis comparing the BRIDGE population to a historical population, bivariate analyses comparing clinical and demographic characteristics between our BRIDGE− or BRIDGE+ and historical populations were performed using 2-tailed *t* tests for continuous variables and chi-squared tests for categorical variables. We also assessed 95% CIs for point estimates of these characteristics in each group using the Wilson score intervals. Multivariable logistic regressions comparing the odds of all assessed outcomes in the BRIDGE versus historical population included all variables with $P<.20$ in the bivariate analyses. Backward stepwise regression was used to create parsimonious models using $P>.20$ for elimination. An identical strategy was used to develop models for comparative analysis assessing differences in outcomes between our BRIDGE− and BRIDGE+ populations. For both analyses, models comparing outcomes across populations ultimately only included a history of IGT in prior pregnancy as a covariate. Multiple imputation analysis was used to address missing data at random for primary care visit scheduling and attendance.

Ethical Considerations

The project was designated as a quality improvement project and deemed exempt from review by the University of Pennsylvania Institutional Review Board, as (1) this intervention did not guide any clinical decisions regarding care and was intended solely to improve care delivery within the health care system, and (2) participants were not randomized to the intervention and were intended to obtain direct benefits from the intervention. Informed consent was not required for this quality improvement initiative. Deidentified data were collected for retrospective review and stored on secure institutional servers. Participants were not compensated for their participation in this initiative. There is no identification of any individual participants in any manuscript images or supplementary materials.

XSL•FO
RenderX

Results

Participant Recruitment

The recruitment of participants occurred between October 2021 and April 2023 and was stopped due to funding constraints, although the target sample size was not achieved. A convenience sampling of patients meeting the eligibility criteria for the intervention and presenting to prenatal clinics within the health care system over the study time period was performed. Providers had the option of directly enrolling patients from the clinic. In addition, on a weekly basis, the study team received an automated report from the health care system analytics team with all newly eligible participants. After reviewing the medical records to confirm eligibility, the study team manually enrolled all eligible patients. The team also reviewed the EMR of any participants added directly by clinicians and removed participants deemed ineligible. A participant flow diagram was included to demonstrate the number of individuals enrolled and ultimately included in the program evaluation (Figure 2).

XSL•FO
RenderX

Figure 2. Participant flow diagram for Better Follow-up of Impaired Glucose Tolerance (BRIDGE) enrollment and program evaluation. [a]Participants who were incorrectly enrolled did not meet eligibility requirements, that is, individuals for whom English was not a primary language, individuals diagnosed with gestational diabetes, or did not meet criteria for impaired glucose tolerance (IGT) in early pregnancy, and individuals who did not deliver at our hospital.

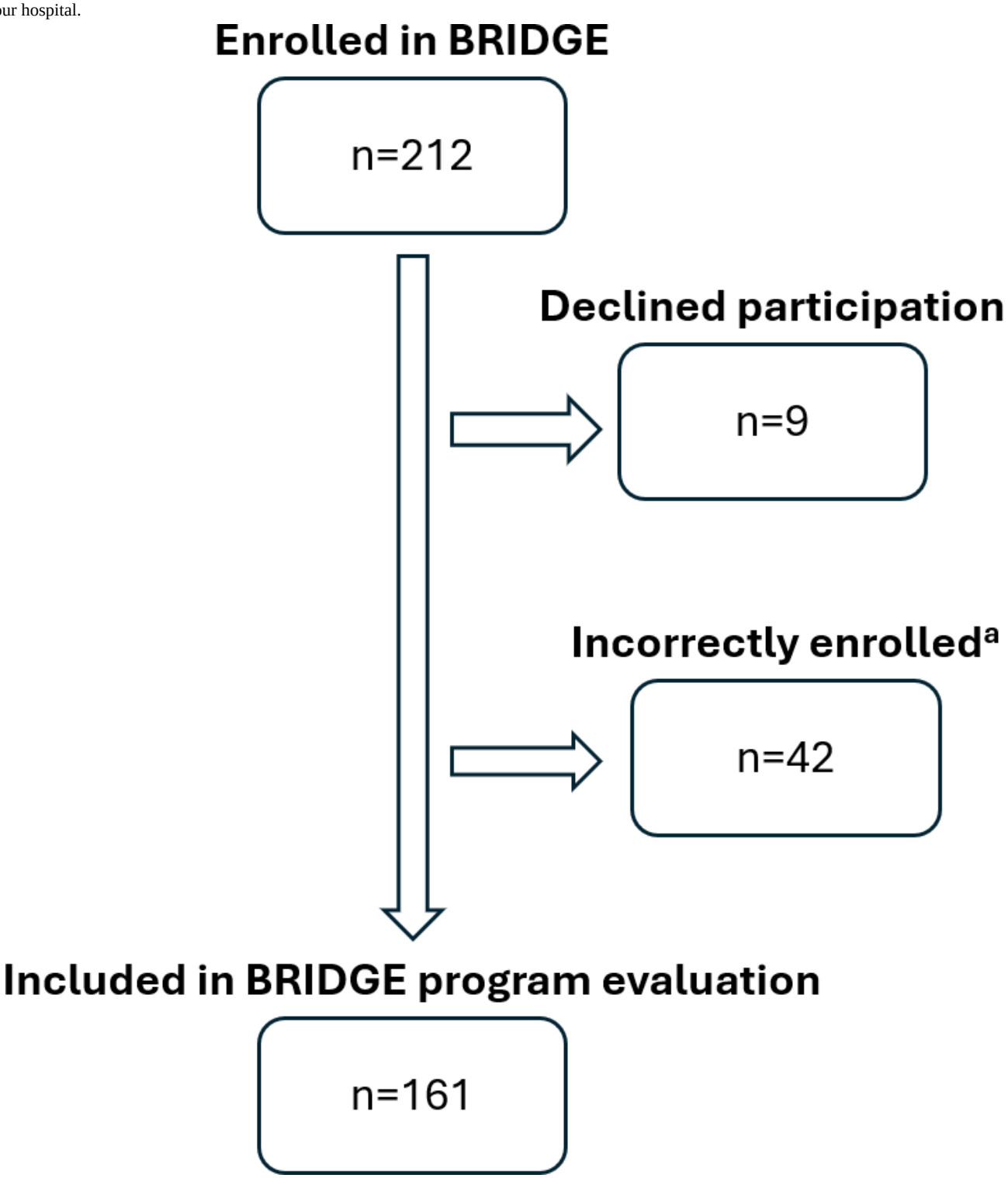

Demographic and Clinical Characteristics of the Study Populations

In total, 161 participants were enrolled in the BRIDGE program, compared to 342 participants in the historical population (Table 1). A higher proportion of the BRIDGE population had a personal history of IGT and polycystic ovarian syndrome compared to the historical population (Table 1). There were 82 participants in the BRIDGE− arm and 79 participants in the BRIDGE+ arm (Table 2). The BRIDGE− and BRIDGE+ populations had similar demographic and clinical characteristics.

XSL•FO
RenderX

Table . Demographic and clinical characteristics of patients with impaired glucose tolerance (IGT) in a historical population (June 2018-December 2019) versus those enrolled in SMS text messaging support program (BRIDGE[a]; October 2021-April 2023).

Characteristics	Historical population (n=342)	BRIDGE[b] (N=161)
Demographics		
Age (y), median (IQR[b])	31 (27-35)	32 (28-36)
Insurance, n (%; 95% CI)		
Private	159 (46.5%; 41.3%-51.8%)	65 (40.4%; 33.1%-48.1%)
Public	179 (52.3%; 47%-57.6%)	96 (59.6%; 51.9%-66.9%)
None	4 (1.2%; 0.5%-3%)	0 (0%; 0%-2.3%)
Race, n (%; 95% CI)		
White	36 (10.5%; 7.7%-14.2%)	11 (6.8%; 3.9%-11.8%)
Black	256 (74.9%; 70%-79.2%)	131 (81.4%; 74.6%-86.6%)
Asian	20 (5.8%; 3.8%-8.9%)	10 (6.2%; 3.4%-11.1%)
Other or unknown	30 (8.8%; 6.2%-12.2%)	9 (5.6%; 2.6%-10.4%)
Hispanic ethnicity, n (%; 95% CI)	16 (4.7%; 2.9%-7.5%)	7 (4.3%; 1.8%-8.8%)
Scant prenatal care (<5 visits), n (%; 95% CI)	6 (1.8%; 0.8%-3.8%)	2 (1.3%; 0.3%-4.4%)
Late prenatal care initiation (>16 weeks), n (%; 95% CI)	9 (2.6%; 1.4%-4.9%)	8 (5.0%; 2.6%-9.6%)
Nulliparous, n (%; 95% CI)	107 (31.3%; 26.6%-36.4%)	44 (27.3%; 21.0%-34.7%)
Primary care provider, n (%; 95% CI)		
In network	118 (34.5%; 30.0%-40.0%)	54 (33.5%; 26.7%-41.1%)
Out of network	118 (34.5%; 30.0%-40.0%)	46 (28.6%; 22.2%-36.0%)
None	106 (31.0%; 26.3%-36.1%)	61 (37.9%; 30.8%-45.6%)
Clinical characteristics, n (%; 95% CI)		
History of prior IGT[h]	35 (10.2%; 7.45%-13.9%)	40 (24.8%; 18.8%-32.1%)
History of GDM[d] in prior pregnancy	5 (1.5%; 0.6%-3.4%)	2 (1.3%; 0.3%-4.4%)
History of macrosomia in prior pregnancy	3 (0.8%; 0.2%-2.5%)	4 (2.5%; 0.9%-6.2%)
History of IUFD[e] >20 weeks gestation	1 (0.3%; 0.05%-1.6%)	0 (0%; 0%-2.3%)
Family history of diabetes	83 (24.3%; 20.0%-29.1%)	49 (30.4%; 23.9%-37.9%)
History of PCOS[f]	4 (1.2%; 0.4%-3.0%)	11 (6.8%; 3.9%-11.8%)
History of metformin use	7 (2.0%; 1.0%-4.2%)	7 (4.3%; 2.1%-8.7%)

[a]BRIDGE: Better Follow-up of Impaired Glucose Tolerance.

[b]Missing values: scant prenatal care (n=1 in BRIDGE population); late prenatal care (n=1 in BRIDGE population); GDM in prior pregnancy (n=1 in BRIDGE population).

[c]IGT: impaired glucose tolerance.

[d]GDM: gestational diabetes mellitus.

[e]IUFD: intrauterine fetal death.

[f]PCOS: polycystic ovarian syndrome.

XSL•FO
RenderX

Table . Demographic and clinical characteristics of patients with impaired glucose tolerance (IGT) enrolled in SMS text messaging support (BRIDGE−[a]; October 2021-July 2022) versus SMS text messaging support + IGT-focused postpartum visit (BRIDGE+; July 2022-April 2023).

Characteristics	BRIDGE−[b] (n=82)	BRIDGE+ (n=79)
Demographics		
Age (y), median (IQR[b])	32 (28-36)	33 (29-36)
Insurance, n (%; 95% CI)		
Private	32 (39.0%; 29.2%-49.8%)	33 (41.8%; 31.5%-52.8%)
Public	50 (61.0%; 50.2%-70.8%)	46 (58.2%; 47.2%-68.4%)
Race, n (%; 95% CI)		
White	3 (3.7%; 1.3%-10.2%)	8 (10.1%; 5.2%-18.7%)
Black	72 (87.8%; 79%-93.2%)	59 (74.7%; 64.1%-83%)
Asian	4 (4.9%; 1.9%-11.9%)	6 (7.6%; 3.5%-15.6%)
Other or unknown	3 (3.7%; 1.3%-10.2%)	6 (7.6%; 3.5%-15.6%)
Hispanic ethnicity, n (%; 95% CI)	2 (2.4%; 0.7%-8.5%)	5 (6.3%; 2.7%-14%)
Scant prenatal care (<5 visits), n (%; 95% CI)	0 (0%; 0%-4.5%)	2 (2.5%; 0.7%-8.9%)
Late prenatal care initiation (>16 weeks), n (%; 95% CI)	4 (4.9%; 1.9%-12%)	4 (5.1%; 2.0%-12.5%)
Nulliparous	21 (25.6%; 17.4%-36%)	23 (29.1%; 20.3%-39.9%)
Primary care provider		
In network	24 (29.3%; 20.5%-39.9%)	30 (38.0%; 28.1%-49%)
Out of network	22 (26.8%; 18.4%-37.3%)	24 (30.4%; 21.3%-41.2%)
None	36 (43.9%; 33.7%-54.7%)	25 (31.6%; 22.4%-42.5%)
Clinical characteristics, n (%; 95% CI)		
History of prior IGT	16 (19.5%; 12.4%-29.4%)	24 (30.3%; 21.3%-41.2%)
History of GDM[c] in prior pregnancy	0 (0%; 0%-4.5%)	2 (2.5%; 0.7%-8.9%)
History of macrosomia in prior pregnancy	3 (3.7%; 1.3%-10.2%)	1 (1.3%; 0.2%-6.8%)
History of IUFD[d] >20 weeks gestation	0 (0%; 0%-4.5%)	0 (0%; 0%-4.6%)
Family history of diabetes	23 (28.0%; 19.5%-38.6%)	26 (32.9%; 23.6%-43.9%)
History of PCOS[e]	5 (6.1%; 2.6%-13.5%)	6 (7.6%; 3.5%-15.6%)
History of metformin use	5 (6.1%; 2.6%-13.5%)	2 (2.5%; 0.7%-8.9%)

[a]BRIDGE: Better Follow-up of Impaired Glucose Tolerance.

[b]Missing values: scant prenatal care (n=1 in BRIDGE− population); late prenatal care (n=1 in BRIDGE− population); GDM in prior pregnancy (n=1 in BRIDGE− population).

[c]GDM: gestational diabetes.

[d]IUFD: intrauterine fetal death.

[e]PCOS: polycystic ovarian syndrome.

Program Evaluation of BRIDGE Using Historical Population for Comparison

BRIDGE participants had increased odds of HbA_{1c} completion by 1 year postpartum (39.8% vs 12.5%; adjusted odds ratio [aOR] 4.28, 95% CI 2.71 - 6.78) and PCP visit scheduling (31.0% vs 12.0%; aOR 9.58, 95% CI 4.39 - 20.9) compared to the historical population (Table 3). The difference in PCP visit scheduling remained statistically significant in the multiple imputation analysis (aOR 9.49, 95% CI 4.60 - 19.6; Multimedia Appendix 1). This was also true among the subgroup of patients with in-network PCPs (27.3% vs 12.0%; aOR 8.30, 95% CI 3.77 - 18.2). Importantly, in the historical population, every patient who scheduled a PCP visit had an established in-network PCP during pregnancy; in contrast, in the BRIDGE population, 6 out of the 50 (12%) patients who scheduled a PCP visit did not already have an established in-network PCP.

XSL•FO
RenderX

Table . Comparing postpartum transitions of care for patients with impaired glucose tolerance (IGT) in a historical population (June 2018-December 2019) versus SMS text messaging support Better Follow-up of Impaired Glucose Tolerance (BRIDGE) population (October 2021-April 2023).

Postpartum transitions	Historical population (n=342)	BRIDGE population (N=161)	aOR[a] (95% CI)
HbA_{1c}[b] completion by 1 year postpartum, n (%; 95% CI)	43 (12.5%; 9.5%-16.5%)	64 (39.8%; 32.5%-47.5%)	4.28 (2.71-6.78)
PCP[c] visit scheduling, n (%; 95% CI)			
In network PCPs[d]	41 (12.0%; 9%-15.9%)	44 (27.3%; 21%-34.7%)	8.30 (3.77-18.2)
All patients[e], n (%; 95% CI)	41 (12.0%; 9%-15.9%)	50 (31.0%; 24.4%-38.6%)	9.58 (4.39-20.9)
PCP visit attendance, n (%; 95% CI[f])			
In network PCPs	38 (92.7%; 80.6%-97.5%)	42 (95.4%; 84.9%-98.7%)	1.81 (0.28-11.7)
All patients	38 (92.7%; 80.6%-97.5%)	48 (96%; 86.5%-98.9%)	2.11 (0.33-13.6)
PCP visits with IGT counseling, n (%; 95% CI)			
In network PCPs	15 (39.5%; 25.6%-55.3%)	21 (50.0%; 35.5%-64.5%)	1.45 (0.59-3.56)
All PCPs	15 (39.5%; 25.6%-55.3%)	26 (54.2%; 40.3%-67.4%)	1.65 (0.68-3.98)

[a]aOR: adjusted odds ratios for history of prior IGT.

[b]HbA_{1c}: hemoglobin A_{1c}.

[c]PCP: primary care physician.

[d]A limited number of patients had in-network PCPs (n=119 for patients in historical population; n=54 for patients in BRIDGE).

[e]The data on PCP visit scheduling and attendance for patients with out-of-network PCPs were obtained via electronic medical record review (of postpartum notes) for both historical and BRIDGE populations. Phone interviews were also attempted for BRIDGE patients to obtain self-reported data, which were limited by a low response rate (n=10 interviews total).

[f]There were missing data for PCP visit scheduling and attendance (n=221 in historical population; n=101 in BRIDGE population).

There were remarkably high levels of attendance of PCP visits among those who scheduled PCP visits, across both the historical and BRIDGE populations (95% CI 92.7% - 96% of scheduled visits). However, rates of IGT counseling during PCP visits were relatively low in both the historical and BRIDGE populations (95% CI 39.5% - 54.2% of visits).

Comparative Analysis Between BRIDGE− and BRIDGE+ Arms

In comparing the BRIDGE− to BRIDGE+ arms of the program, the BRIDGE+ population had slightly improved odds of HbA_{1c} completion by 12 weeks postpartum (36.7% vs 19.5%; aOR 2.31, 95% CI 1.12 - 4.74), though this positive effect using the BRIDGE+ strategy did not persist by 1 year postpartum (48.1% vs 31.7%; aOR 1.88, 95% CI 0.98 - 3.61; Table 4). There was no difference in PCP visit scheduling between the BRIDGE− and BRIDGE+ program strategies (32.9% vs 29.3%; aOR 0.46, 95% CI 0.11 - 2.00), which remained consistent in multiple imputation analysis (aOR 0.49, 95% CI 0.10 - 2.25; Multimedia Appendix 1). This was similar when looking at PCP scheduling among patients with established in-network PCPs specifically (29.1% vs 25.6%; aOR 0.47, 95% CI 0.11 - 2.09). A similar proportion of patients who did not have established in-network PCPs were able to schedule a PCP visit in both the BRIDGE− and BRIDGE+ populations.

XSL•FO
RenderX

Table . Comparing the effectiveness of SMS text messaging support (BRIDGE−[a]) versus SMS text messaging support + impaired glucose tolerance (IGT)-focused postpartum visit (BRIDGE+) in completing postpartum transitions of care for patients with IGT.

Postpartum transitions	BRIDGE− (n=82)	BRIDGE+ (n=79)	aOR[b] (95% CI)
HbA_{1c}[c] completion at 12 weeks postpartum, n (%; 95% CI)	16 (19.5%; 12.4%-29.4%)	29 (36.7%; 26.9%-47.7%)	2.31 (1.12-4.74)
HbA_{1c} completion at 1 year postpartum, n (%; 95% CI)	26 (31.7%; 22.6%-42.4%)	38 (48.1%; 37.4%-58.9%)	1.88 (0.98-3.61)
PCP[d] visit scheduling[e], n (%; 95% CI)			
In network PCPs[e]	21 (25.6%; 17.4%-36%)	23 (29.1%; 20.3%-39.9%)	0.47 (0.11-2.09)
All patients[f]	24 (29.3%; 20.5%-39.9%)	26 (32.9%; 23.6%-43.9%)	0.46 (0.11-2.00)
PCP visit attendance[e], n (%; 95% CI)			
In network PCPs	19 (90.4%; 71.1%-97.3%)	23 (100%; 85%-100%)	—[g]
All patients	22 (91.6%; 74.2%-97.7%)	26 (100%; 87%-100%)	—
PCP visits with IGT counseling, n (%; 95% CI)			
In network PCPs	10 (52.6%; 31.7%-72.7%)	11 (47.8%; 29.2%-67%)	0.70 (0.23-2.73)
All PCPs	12 (54.5%; 34.7%-73.1%)	14 (53.8%; 40.8%-77.8%)	0.98 (0.30-3.20)
IGT-focused postpartum visit[h], n (%; 95% CI)			
Appointment scheduled	—	47 (59.5%; 48.5%-69.6%)	—
Appointment attended	—	23 (48.9%; 35.3%-62.8%)	—

[a]BRIDGE: Better Follow-Up of Impaired Glucose Tolerance.

[b]aOR: adjusted odds ratios for history of prior IGT.

[c]HbA_{1c}: hemoglobin A_{1c}.

[d]PCP: primary care physician.

[e]There were missing data for PCP visit scheduling and attendance (n=55 in BRIDGE− population, n=46 in BRIDGE+ population).

[f]A limited number of patients had in-network PCPs (n=119 for patients in historical population; n=24 for patients in BRIDGE−; n=30 for patients in BRIDGE+).

[g]Not applicable.

[h]The data on PCP visit scheduling and attendance for patients with out-of-network PCPs were obtained via electronic medical record review (of postpartum notes) for both historical and BRIDGE populations. Phone interviews were also attempted for BRIDGE patients to obtain self-reported data, which were limited by low response rate (n=10 interviews total).

[i]Only BRIDGE+ participants were eligible for this visit (n=79).

Once again, there were remarkably high attendance rates for scheduled PCP visits (91.6%-100%). There were relatively low levels of IGT counseling during PCP visits (53.8%-54.5%). In the BRIDGE+ population, a larger proportion of patients scheduled an iPPV than a PCP visit (59.5% vs 29.1%), but attendance rates at scheduled iPPV visits were much lower than those for PCP visits (48.9% vs 100%).

Discussion

Principal Results

Implementing BRIDGE, an SMS text messaging support program with educational and behavioral prompts focusing on postpartum transitions for patients with IGT during pregnancy, resulted in increased completion of an HbA_{1c} test within 1 year postpartum and increased scheduling of a PCP visit by 12 weeks postpartum. The addition of an iPPV to the SMS text messaging support improved the completion of HbA_{1c} testing by 6 weeks postpartum, but this effect did not persist at 1 year postpartum and did not impact PCP visit scheduling.

Comparison With Prior Work

To our knowledge, this is the first study examining the impact of an intervention on postpartum transition milestones for patients diagnosed with IGT in pregnancy, who often do not receive tailored support during pregnancy and the immediate postpartum period through specialized programming given to patients with HDP or GDM (eg, transition clinics, remote blood pressure monitoring programs, and diabetes in pregnancy programs). This is an important population to target as these patients experience higher rates of negative interpregnancy and long-term health outcomes [28-30].

While there is a paucity of literature focusing on patients with IGT, postpartum transitions for patients with GDM have been well studied. Postpartum glucose screening rates within 12 weeks postpartum for patients with GDM (primarily assessed via completion of a 2-hour glucose tolerance test rather than

XSL•FO
RenderX

HbA_{1c} completion by 6 weeks postpartum, based on guidelines) range from 7% to 59% in the literature, which are similar to the 6-week HbA_{1c} completion rate in historical and BRIDGE populations [31-34]. The PCP visit attendance rates for both the historical population and BRIDGE population were similar to published rates, which range from 22% to 60% [15,35]. Multiple systematic reviews assessing the impact of reminder interventions (eg, emails, phone calls, and mail) on postpartum glucose screening rates for patients with GDM have shown potential benefits [36-38]. A randomized controlled trial of a text message intervention targeting primary care transitions for patients with GDM and HDP improved PCP visit completion by 18.7%, which is similar to the difference in PCP visit attendance between the historical and BRIDGE populations [15]. Overall, this consistency with existing literature provides reassurance that the BRIDGE intervention may be generalizable to other settings.

This program evaluation demonstrated that the integration of iPPV increased HbA_{1c} completion in the first 12 weeks postpartum, but this improvement did not persist until 1 year postpartum and did not impact PCP visit scheduling. Patient engagement may play a role in this transient impact of iPPV integration on desired postpartum milestones. While the iPPV scheduling rate was nearly double the PCP visit scheduling rate, the number of individuals who attended the iPPV was similar to the number of individuals who attended the PCP visit. Although the authors were insufficiently powered to assess the relationship between iPPV and PCP visit attendance, it is also possible that patients who completed an iPPV perceived a decrease in the usefulness of PCP visits in the first year postpartum because they had already received IGT-counseling and screening via HbA_{1c} completion from a prenatal care provider.

This program evaluation highlights areas for improvement in the quality of the postpartum transition for patients with complications of pregnancy, as a large proportion of PCP visits in both the historical and BRIDGE populations did not include counseling regarding IGT. Per existing literature, potential contributing factors to this gap in primary care transitions may be related to communication breakdowns between the obstetric and primary care teams regarding pregnancy complications that may influence long-term counseling or management, awareness and knowledge of GDM guidelines among PCPs, and patients' recall of providers' counseling regarding long-term risks, among other factors [39,40]. Potential solutions include enhanced support for care coordination (eg, expedited referral pathways to primary care and patient navigation services), improved handoffs between obstetric and PCP teams (eg, standardized documentation of pregnancy complications), and reminder systems [41].

Our efforts to effectively implement BRIDGE highlighted many deep-rooted challenges with primary care transitions for pregnant patients. From a health care system standpoint, PCP access was limited by both health insurance specifications and reduced capacity in the primary care workforce. This limited the ability to successfully schedule PCP visits for postpartum patients without targeted efforts in this area. In addition, patients reported negative social drivers of health (eg, lack of childcare support, transportation support, or challenges with newborn care such as a city-wide formula shortage) as well as physical or mental health challenges that served as barriers or competing priorities for patients attempting to complete postpartum transition milestones. These challenges have been well characterized in both quantitative and qualitative analyses within the literature and highlight the need for a multilevel approach to optimizing postpartum transitions of care, addressing individual, health care system, and societal barriers [42-44].

Limitations and Strengths

The study had several key limitations. As a nonrandomized pre-post implementation study, our analysis cannot account for unmeasured confounders. The BRIDGE population had higher-risk metabolic comorbidities than the historical population (eg, history of IGT in prior pregnancy), but the overall cohort size was much smaller. As only 8 participants declined enrollment in BRIDGE, the study does not suspect that this difference in cohort size is reflective of selection bias but may be due to other unmeasured factors contributing to decreased rates of IGT in the contemporary population. The COVID-19 pandemic also introduced constraints to the preimplementation or postimplementation design that increase the risk of temporal bias. For example, the historical population in this study was not subject to the systemic changes in health care post–COVID-19, such as the rising use of telemedicine [45], which may have improved access to care [46]. Finally, the study was unable to recruit the target number of participants despite extending the planned enrollment period due to decreases in the prevalence of IGT in the population. The smaller sample size raises some concern regarding the precision and robustness of estimates in the primary analysis, as well as with the multiple imputation analysis. A larger, randomized trial will be required to ascertain the definitive impact of this intervention, though this study provides positive preliminary results.

There are also some scalability concerns with the BRIDGE program. Specifically, the iteration of the BRIDGE program used in this study was resource intensive, involving the manual enrollment of participants by study staff using EMR review; however, if proven effective, these resource intensive steps could be automated, as has been done with other W2H programs in our institution [47]. In addition, other institutions may not have access to the staffing capacity required to support other aspects of the program, such as social work support or IGT-focused postpartum visits. Finally, the study enrolled only English-speaking patients at a single tertiary care institution in the northeastern United States, which limits generalizability.

The study may also be prone to bias due to missing data. For patients with out-of-network PCPs, the study had extremely limited data on PCP visit scheduling and attendance. Phone interviews were attempted for BRIDGE patients to obtain self-reported data but had a low response rate (n=10). Phone interviews were deferred for the historical population because of concerns about recall bias. It is possible, therefore, that the PCP visit scheduling and attendance rates among patients with out-of-network PCPs in the historical population were under-reported. This would bias the results toward programmatic

XSL•FO
RenderX

impact. However, reassuringly, the beneficial impact of BRIDGE on PCP visit scheduling persists when looking at patients with in-network PCPs alone, where complete access to primary care data is present.

The study also had many strengths. First, the development of the BRIDGE program used robust input from interdisciplinary teams spanning behavioral health, digital health, and communications expertise [48]. In addition, the preimplementation and ongoing implementation work with prenatal care sites allowed the authors to optimize the effectiveness of the program using flexible and varied implementation strategies [49], based on an expansive understanding of the barriers faced by patients and providers attempting to achieve a successful postpartum transition. Finally, the population was racially and socioeconomically diverse with varied experiences with PCP care entering pregnancy, which may make their experiences more generalizable to other populations.

Future directions for research could include additional assessments of implementation outcomes (eg, fidelity and acceptability) to ascertain contributing factors to the difference in outcomes between BRIDGE− and BRIDGE+ [50]. Following this work, a formal implementation trial of BRIDGE where groups are randomized to the intervention, which would account for any unmeasured confounding in this study [51]. Planning for that trial should include phone or in-person surveys or other strategies to obtain complete information on primary care transitions for individuals with out-of-network PCPs. Consideration should also be given to adapting BRIDGE to be applicable to more diverse or resource-limited settings [52]; potential strategies could include the automation of participant enrollment, integration of iPPV content into routine postpartum visits (thus eliminating the need for additional staffing), the conversion of iPPV from an in-person to telehealth visit (thus eliminating the need for additional clinic capacity), and the cocreation of adapted messaging in other languages alongside community partners.

Conclusions

In this nonrandomized pre-post implementation study, SMS text messaging support improved postpartum transitions of care for patients diagnosed with IGT in pregnancy, though the quality of the transitions to primary care remained low with poor rates of IGT counseling by PCPs. SMS text messaging support with educational and behavioral prompts tripled completion rates of HbA_{1c} screening within 1 year postpartum and doubled the scheduling rate for PCP visits by 12 weeks postpartum. While attendance at scheduled PCP visits was very high, less than 60% of PCP visits included IGT counseling, highlighting key areas for improvement in the quality of postpartum transitions to primary care. This study also highlighted a complex array of individual, systemic, and societal barriers to achieving desired postpartum milestones. A larger randomized trial is needed to definitively ascertain the impact of this SMS text messaging intervention. Optimizing postpartum transitions of care for populations with longitudinal metabolic risk after pregnancy will require a multilevel approach, of which text message support could be an important component.

Acknowledgments

The authors would like to acknowledge the collaboration of their excellent prenatal care providers and primary care providers in facilitating the successful implementation of Better Follow-up of Impaired Glucose Tolerance, as well as the funding support from the Sloan Family Fund Innovation in Women's Health grant. No artificial intelligence tools were used in the writing of this study.

Funding

This project was supported by the Sloane Family Fund for Innovation in Women's Health. The funder was not involved in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The data may be available upon request from the authors after review and approval by the University of Pennsylvania Institutional Review Board, due to this project's designation as a quality improvement initiative. In light of these constraints, the authors have not placed the data in a publicly available repository.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Multiple imputation analysis for primary care visit scheduling and attendance.
[DOCX File, 17 KB - jmir_v28i1e76493_app1.docx]

References

1. Centers for Disease Control and Prevention (CDC). Awareness of prediabetes--United States, 2005-2010. MMWR Morb Mortal Wkly Rep 2013 Mar 22;62(11):209-212. [Medline: 23515058]

XSL•FO
RenderX

2. Warram JH, Sigal RJ, Martin BC, Krolewski AS, Soeldner JS. Natural history of impaired glucose tolerance: follow-up at Joslin Clinic. Diabet Med 1996 Sep;13(9 Suppl 6):S40-S45. [Medline: 8894480]
3. Yudkin JS, Alberti KG, McLarty DG, Swai AB. Impaired glucose tolerance. BMJ 1990 Sep 1;301(6749):397-402. [doi: 10.1136/bmj.301.6749.397] [Medline: 2282392]
4. King H, Zimmet P, Raper LR, Balkau B. The natural history of impaired glucose tolerance in the Micronesian population of Nauru: a six-year follow-up study. Diabetologia 1984 Jan;26(1):39-43. [doi: 10.1007/BF00252261] [Medline: 6706044]
5. Nathan DM, Davidson MB, DeFronzo RA, et al. Impaired fasting glucose and impaired glucose tolerance: implications for care. Diabetes Care 2007 Mar;30(3):753-759. [doi: 10.2337/dc07-9920] [Medline: 17327355]
6. Schäfer S, Kantartzis K, Machann J, et al. Lifestyle intervention in individuals with normal versus impaired glucose tolerance. Eur J Clin Invest 2007 Jul;37(7):535-543. [doi: 10.1111/j.1365-2362.2007.01820.x] [Medline: 17576204]
7. den Boer AT, Herraets IJT, Stegen J, et al. Prevention of the metabolic syndrome in IGT subjects in a lifestyle intervention: results from the SLIM study. Nutr Metab Cardiovasc Dis 2013 Nov;23(11):1147-1153. [doi: 10.1016/j.numecd.2012.12.005] [Medline: 23462149]
8. Hunt KJ, Schuller KL. The increasing prevalence of diabetes in pregnancy. Obstet Gynecol Clin North Am 2007 Jun;34(2):173-199. [doi: 10.1016/j.ogc.2007.03.002] [Medline: 17572266]
9. QuickStats: percentage of mothers with gestational diabetes,* by maternal age—National Vital Statistics System, United States, 2016 and 2021. MMWR Morb Mortal Wkly Rep 2023 Jan 6;72(1):16. [doi: 10.15585/mmwr.mm7201a4] [Medline: 36602935]
10. Boe B, Barbour LA, Allshouse AA, Heyborne KD. Universal early pregnancy glycosylated hemoglobin A1c as an adjunct to Carpenter-Coustan screening: an observational cohort study. Am J Obstet Gynecol MFM 2019 Mar;1(1):24-32. [doi: 10.1016/j.ajogmf.2019.02.003] [Medline: 33319754]
11. Bender W, McCarthy C, Elovitz M, Parry S, Durnwald C. Universal HbA1c screening and gestational diabetes: a comparison with clinical risk factors. J Matern Fetal Neonatal Med 2022 Dec;35(25):6430-6436. [doi: 10.1080/14767058.2021.1914578] [Medline: 34044736]
12. Bender WR, McCarthy C, Elovitz M, Parry S, Durnwald C. Adverse pregnancy outcomes in nondiabetic patients with an elevated early pregnancy HbA1c. Am J Perinatol 2022 Oct;29(14):1496-1502. [doi: 10.1055/a-1877-8696] [Medline: 35709738]
13. Hughes RCE, Moore MP, Gullam JE, Mohamed K, Rowan J. An early pregnancy HbA1c ≥5.9% (41 mmol/mol) is optimal for detecting diabetes and identifies women at increased risk of adverse pregnancy outcomes. Diabetes Care 2014 Nov;37(11):2953-2959. [doi: 10.2337/dc14-1312] [Medline: 25190675]
14. Baron AD. Impaired glucose tolerance as a disease. Am J Cardiol 2001 Sep 20;88(6A):16H-19H. [doi: 10.1016/s0002-9149(01)01832-x] [Medline: 11576521]
15. Clapp MA, Ray A, Liang P, James KE, Ganguli I, Cohen JL. Postpartum primary care engagement using default scheduling and tailored messaging: a randomized clinical trial. JAMA Netw Open 2024 Jul 1;7(7):e2422500. [doi: 10.1001/jamanetworkopen.2024.22500] [Medline: 39012630]
16. Gomez HB, Hoffman MK. Text messaging as a means to engage patients in the postpartum period. Clin Obstet Gynecol 2021 Jun 1;64(2):366-374. [doi: 10.1097/GRF.0000000000000609] [Medline: 33904842]
17. Melis M, Mastinu M, Sollai G. Effect of the rs2821557 polymorphism of the human Kv1.3 gene on olfactory function and BMI in different age groups. Nutrients 2024;16(6):821. [doi: 10.3390/nu16060821]
18. Heatley E, Middleton P, Hague W, Crowther C. The DIAMIND study: postpartum SMS reminders to women who have had gestational diabetes mellitus to test for type 2 diabetes: a randomised controlled trial—study protocol. BMC Pregnancy Childbirth 2013 Apr 12;13:1-6. [doi: 10.1186/1471-2393-13-92] [Medline: 23587090]
19. Strohbach A, Hu F, Martinez NG, Yee LM. Evaluating the use of text message communication in a postpartum patient navigation program for publicly insured women. Patient Educ Couns 2019 Apr;102(4):753-759. [doi: 10.1016/j.pec.2018.10.028] [Medline: 30448040]
20. Hirshberg A, Downes K, Srinivas S. Comparing standard office-based follow-up with text-based remote monitoring in the management of postpartum hypertension: a randomised clinical trial. BMJ Qual Saf 2018 Nov;27(11):871-877. [doi: 10.1136/bmjqs-2018-007837] [Medline: 29703800]
21. Mehta SJ, Troxel AB, Marcus N, et al. Participation rates with opt-out enrollment in a remote monitoring intervention for patients with myocardial infarction. JAMA Cardiol 2016 Oct 1;1(7):847-848. [doi: 10.1001/jamacardio.2016.2374] [Medline: 27603755]
22. Halpern SD, French B, Small DS, et al. Randomized trial of four financial-incentive programs for smoking cessation. N Engl J Med 2015 May 28;372(22):2108-2117. [doi: 10.1056/NEJMoa1414293] [Medline: 25970009]
23. Asch DA, Troxel AB, Stewart WF, et al. Effect of financial incentives to physicians, patients, or both on lipid levels: a randomized clinical trial. JAMA 2015 Nov 10;314(18):1926-1935. [doi: 10.1001/jama.2015.14850] [Medline: 26547464]
24. American Diabetes Association. Standards of care in diabetes-2023 abridged for primary care providers. Clin Diabetes 2023 Jan 2;41(1):4-31. [doi: 10.2337/cd23-as01] [Medline: 36714254]
25. Bernstein JA, Quinn E, Ameli O, et al. Follow-up after gestational diabetes: a fixable gap in women's preventive healthcare. BMJ Open Diabetes Res Care 2017;5(1):e000445. [doi: 10.1136/bmjdrc-2017-000445] [Medline: 28948028]

XSL•FO
RenderX

26. Eldridge SM, Chan CL, Campbell MJ, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016 Oct 24;355:i5239. [doi: 10.1136/bmj.i5239] [Medline: 27777223]
27. Lancaster GA, Thabane L. Guidelines for reporting non-randomised pilot and feasibility studies. Pilot Feasibility Stud 2019;5(1):114. [doi: 10.1186/s40814-019-0499-1] [Medline: 31608150]
28. Flatt SB, Pudwell J, Smith GN. Evaluation of a postpartum cardiovascular risk screening clinic: an analysis of interpregnancy and subsequent pregnancy outcomes. J Obstet Gynaecol Can 2022 Feb;44(2):157-166. [doi: 10.1016/j.jogc.2021.07.018] [Medline: 34425300]
29. Kumar NR, Hirshberg A, Srinivas SK. Best practices for managing postpartum hypertension. Curr Obstet Gynecol Rep 2022;11(3):159-168. [doi: 10.1007/s13669-022-00343-6] [Medline: 35757523]
30. Carolan-Olah MC. Educational and intervention programmes for gestational diabetes mellitus (GDM) management: an integrative review. Collegian 2016 Mar;23(1):103-114. [doi: 10.1016/j.colegn.2015.01.001] [Medline: 27188046]
31. Farid H, Blake R, Hacker MR, Erlinger AL, Modest AM. Strategies to improve postpartum glucose screening rates are needed. J Obstet Gynaecol Res 2021 Aug;47(8):2641-2645. [doi: 10.1111/jog.14868] [Medline: 34041808]
32. Linnenkamp U, Greiner GG, Haastert B, et al. Postpartum screening of women with GDM in specialised practices: data from 12,991 women in the GestDiab register. Diabet Med 2022 Jul;39(7):e14861. [doi: 10.1111/dme.14861] [Medline: 35472098]
33. Eggleston EM, LeCates RF, Zhang F, Wharam JF, Ross-Degnan D, Oken E. Variation in postpartum glycemic screening in women with a history of gestational diabetes mellitus. Obstet Gynecol 2016;128(1):159-167. [doi: 10.1097/AOG.0000000000001467] [Medline: 27275787]
34. Battarbee AN, Yee LM. Barriers to postpartum follow-up and glucose tolerance testing in women with gestational diabetes mellitus. Am J Perinatol 2018 Mar;35(4):354-360. [doi: 10.1055/s-0037-1607284] [Medline: 29020693]
35. Bennett WL, Chang HY, Levine DM, et al. Utilization of primary and obstetric care after medically complicated pregnancies: an analysis of medical claims data. J Gen Intern Med 2014 Apr;29(4):636-645. [doi: 10.1007/s11606-013-2744-2] [Medline: 24474651]
36. Jeppesen C, Kristensen JK, Ovesen P, Maindal HT. The forgotten risk? A systematic review of the effect of reminder systems for postpartum screening for type 2 diabetes in women with previous gestational diabetes. BMC Res Notes 2015 Aug 26;8:373. [doi: 10.1186/s13104-015-1334-2] [Medline: 26306499]
37. Middleton P, Crowther CA. Reminder systems for women with previous gestational diabetes mellitus to increase uptake of testing for type 2 diabetes or impaired glucose tolerance. Cochrane Database Syst Rev 2014 Mar 18;2014(3):CD009578. [doi: 10.1002/14651858.CD009578.pub2] [Medline: 24638998]
38. Nielsen JH, Melendez-Torres GJ, Rotevatn TA, Peven K, Fonager K, Overgaard C. How do reminder systems in follow-up screening for women with previous gestational diabetes work?—A realist review. BMC Health Serv Res 2021 Jun 1;21(1):535. [doi: 10.1186/s12913-021-06569-z] [Medline: 34074278]
39. Wilkinson SA, Brodribb WE, Upham S, Janamian T, Nicholson C, Jackson CL. Primary care of women after gestational diabetes mellitus: mapping the evidence-practice gap. Med J Aust 2014 Aug 4;201(3 Suppl):S74-S77. [doi: 10.5694/mja14.00264] [Medline: 25047888]
40. Kim C, McEwen LN, Kerr EA, et al. Preventive counseling among women with histories of gestational diabetes mellitus. Diabetes Care 2007 Oct;30(10):2489-2495. [doi: 10.2337/dc07-0435] [Medline: 17623826]
41. Martinez NG, Niznik CM, Yee LM. Optimizing postpartum care for the patient with gestational diabetes mellitus. Am J Obstet Gynecol 2017 Sep;217(3):314-321. [doi: 10.1016/j.ajog.2017.04.033] [Medline: 28455081]
42. Lithgow GE, Rossi J, Griffin SJ, Usher-Smith JA, Dennison RA. Barriers to postpartum diabetes screening: a qualitative synthesis of clinicians' views. Br J Gen Pract 2021 Jun;71(707):e473-e482. [doi: 10.3399/BJGP.2020.0928] [Medline: 33947667]
43. Nielsen KK, Kapur A, Damm P, de Courten M, Bygbjerg IC. From screening to postpartum follow-up—the determinants and barriers for gestational diabetes mellitus (GDM) services, a systematic review. BMC Pregnancy Childbirth 2014 Jan 22;14:1-18. [doi: 10.1186/1471-2393-14-41] [Medline: 24450389]
44. McCloskey L, Quinn E, Ameli O, et al. Interrupting the pathway from gestational diabetes mellitus to type 2 diabetes: the role of primary care. Womens Health Issues 2019;29(6):480-488. [doi: 10.1016/j.whi.2019.08.003] [Medline: 31562051]
45. Friedman AB, Gervasi S, Song H, et al. Telemedicine catches on: changes in the utilization of telemedicine services during the COVID-19 pandemic. Am J Manag Care 2022 Jan 1;28(1):e1-e6. [doi: 10.37765/ajmc.2022.88771] [Medline: 35049260]
46. Kumar NR, Arias MP, Leitner K, Wang E, Clement EG, Hamm RF. Assessing the impact of telehealth implementation on postpartum outcomes for Black birthing people. Am J Obstet Gynecol MFM 2023 Feb;5(2):100831. [doi: 10.1016/j.ajogmf.2022.100831] [Medline: 36496115]
47. Graseck A, Josephs M, Dokras A, Srinivas SK. THEA: prenatal education and remote blood pressure monitoring. NEJM Catal Innov Care Deliv 2024 Jan 17;5(2):23. [doi: 10.1056/CAT.23.0343]
48. Fouquet SD, Miranda AT. Asking the right questions-human factors considerations for telemedicine design. Curr Allergy Asthma Rep 2020 Aug 29;20(11):66. [doi: 10.1007/s11882-020-00965-x] [Medline: 32862299]

XSL•FO
RenderX

49. Powell BJ, Waltz TJ, Chinman MJ, et al. A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. Implement Sci 2015 Feb 12;10(1):21. [doi: 10.1186/s13012-015-0209-1] [Medline: 25889199]
50. Proctor E, Silmere H, Raghavan R, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011 Mar;38(2):65-76. [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
51. Curran GM, Bauer M, Mittman B, Pyne JM, Stetler C. Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. Med Care 2012 Mar;50(3):217-226. [doi: 10.1097/MLR.0b013e3182408812] [Medline: 22310560]
52. Materia FT, Smyth JM, Puoane T, et al. Implementing text-messaging to support and enhance delivery of health behavior change interventions in low- to middle-income countries: case study of the Lifestyle Africa intervention. BMC Public Health 2023 Aug 10;23(1):1526. [doi: 10.1186/s12889-023-16388-y] [Medline: 37563595]

Abbreviations

aOR: adjusted odds ratio
BRIDGE: Better Follow-up of Impaired Glucose Tolerance
CONSORT: Consolidated Standards of Reporting Trials
EMR: electronic medical record
GDM: gestational diabetes mellitus
HbA_{1c}: hemoglobin A_{1c}
HDP: hypertensive disorders of pregnancy
IGT: impaired glucose tolerance
iPPV: IGT-focused postpartum visit
PCP: primary care provider
W2H: Way2Health

Edited by S Brini; submitted 24.Apr.2025; peer-reviewed by R Appiah-Kubi, S Marschner; revised version received 21.Apr.2026; accepted 01.May.2026; published 09.Jun.2026.

Please cite as:
Kumar NR, Bender W, Durnwald C
Text Messaging Support for Patients Diagnosed With Impaired Glucose Tolerance During Pregnancy: Nonrandomized Pre-Post Implementation Study Assessing Impact on Postpartum Transitions of Care
J Med Internet Res 2026;28:e76493
URL: https://www.jmir.org/2026/1/e76493
doi:10.2196/76493

© Natasha Kumar, Whitney Bender, Celeste Durnwald. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Predictors of Outcomes of Parent Training Targeting Disruptive Behavior in Children Aged 4 Years at 6-Month Follow-Up: Results From a Large Prospective Cohort Implementation Study

Minja Westerlund[1,2], PhD; Susanna Hinkka-Yli-Salomäki[1,2], PhLic; Terja Ristkari[1,2], RN, MNSc; Malin Kinnunen[1,2], MHc; Amit Baumel[3], PhD; Andre Sourander[1,2,4], MD, PhD

[1]Research Centre for Child Psychiatry, Faculty of Medicine, University of Turku, Lemminkäisenkatu 3a, Teutori, 3rd floor, Turku, Finland

[2]INVEST Research Flagship Centre, University of Turku, Turku, Finland

[3]Department of Community Mental Health, University of Haifa, Haifa, Israel

[4]Department of Child Psychiatry, Turku University Hospital, Turku, Finland

Corresponding Author:
Andre Sourander, MD, PhD
Research Centre for Child Psychiatry, Faculty of Medicine, University of Turku, Lemminkäisenkatu 3a, Teutori, 3rd floor, Turku, Finland

Abstract

Background: Parent training interventions effectively reduce disruptive behavior in children. However, research on how participant characteristics and program factors influence the outcomes in real-world settings remains scarce.

Objective: This study aimed to identify factors predicting outcomes of the internet-based, telephone-assisted Strongest Families Parent training program.

Methods: A prospective cohort implementation study conducted within population-based screening embedded in routine health checkups targeting all children aged 4 years in Finland, to identify children with high levels of conduct problems and functional impairment. From a study population of 49,504, altogether 3911 participants completed baseline measures, 707 participants did not do so at 6 months follow-up, resulting in a sample of 3204 (1158/3186, 36.3% girl, 2028/3186, 63.7% boys). Reported duration of difficulties was 6 months in 29.57% (934/3159) of participants, 6 - 12 months in 27% (853/3159) of participants, and >12 months in 43.43% (1372/3159) of participants. Most children lived with 2 biological parents (2721/3194, 85.19%). A total of 35.24% (1121/3181) of mothers and 26.18% (797/3044) of fathers had a university degree. Data was collected via parent report. Multinomial logistic regression analyses were conducted to identify which child-, family-, and program-related factors predicted changes in the Child Behavior Checklist 1.5-5 (CBCL) externalizing subscale from baseline to 6-month follow-up. The standardized change in CBCL externalizing score was created by subtracting the mean at baseline from the individual 6-month measurement, divided by the SD at baseline. The standardized change was categorized to ±0.5 SD (no change); +0.5 to +1.5 SD (moderate improvement), >+1.5 SD (large improvement), and more than −0.5 SD (deterioration). A *P* value of <.05 was considered significant.

Results: In 77% (2468/3204) of participants, symptoms improved at 6-month follow-up. Multinomial logistic regression analyses with α-level of <0.05 showed that >12 months duration of initial problems, callous-unemotional traits, and CBCL internalizing symptoms were linked to lower likelihood of large improvement (odds ratio [OR] 0.43, 95% CI 0.33 - 0.56; *P*<.001; OR 0.64, 95% CI 0.57 - 0.73; *P*<.001; OR 0.54, 95% CI 0.47 - 0.63; *P*<.001, respectively). Definite and severe problems at baseline were linked to deterioration (OR 2.29, 95% CI 1.62 - 3.24; *P*<.001; OR 4.38, 95% CI 2.80 - 6.85; *P*<.001, respectively). Parental stress was linked to a lower likelihood of large improvement (OR 0.78, 95% CI 0.67 - 0.91; *P*=.002), and anxiety to a higher likelihood of deterioration (OR 1.20, 95% CI 1.04 - 1.39; *P*=.02).

Conclusions: Children with longer-term and more severe behavioral symptoms may require tailored intervention. Support for parents with stress may be recommended. Much of the current literature on parent training is based on randomized controlled trials, while the literature on the implementation of parenting programs and studies examining change is limited. Our study informs about predictors of treatment outcomes when interventions are implemented. These results are important clinically as they allow personalization of interventions.

(J Med Internet Res 2026;28:e79592) doi:10.2196/79592

KEYWORDS

parent training; implementation; digital interventions; psychosocial; children

XSL•FO
RenderX

Introduction

Randomized controlled trials (RCTs) have offered robust empirical support for the efficacy of parent training in reducing disruptive behavior in children [1,2]. However, as widely acknowledged by clinicians, family responses to parent training interventions, even when they are empirically supported, can vary considerably [3]. There have been calls for studies evaluating psychosocial interventions outside of structured research settings, using larger and more heterogeneous samples than are typically used in RCTs [4]. An ideal opportunity for conducting such studies would be when interventions are implemented in practice and used by substantial portions of the target population, but such large-scale implementation studies are not routinely carried out. Consequently, the literature on the implementation of parenting programs in community settings is scarce [5-7]. Much of the current evidence is based on RCTs with limited sample sizes and often clinical samples, limiting a comprehensive examination of the predictors of treatment outcomes in real-life settings. Implementation research can provide a basis for more personalized care by identifying [8] which interventions are most effective for which individuals [9].

Multiple factors may influence the efficacy of parent training programs [10]. However, much of the existing evidence is derived from RCTs with rather small sample sizes, and only a limited number of studies have been conducted in implementation settings [11,12]. Further, few have used population-based screening methods [13,14] or have assessed digital parent training interventions [15-18]. Among the various predicting characteristics of children, younger age has been associated with the likelihood of benefitting from parent training programs [19], including when such programs are delivered in digital format [20]. Higher initial severity of behavioral problems has been associated with greater reductions in problem behaviors following a digitally delivered parent training program [16]. Relatedly, another study [21] showed that callous-unemotional (callous-unemotional) traits among children (ie, individual characteristics marked by a lack of empathy, remorse, and guilt) reduced the functional gains of a parent-training program, especially when treatment was delivered digitally. Nevertheless, callous-unemotional traits were found to attenuate the program's effects on problem severity and treatment response status. Regarding parental characteristics, a meta-analysis [22] found that low socioeconomic status or lower education level did not seem to diminish the effectiveness of parent training interventions, although it was suggested that the gained effects were more difficult to maintain over longer follow-up periods [16]. Findings related to parental psychopathology, and depression in particular, have been mixed, with some studies pointing to greater improvement in child disruptive behavior and others pointing to the opposite [10,23,24]. There is also some qualitative evidence indicating that family stressors such as illness or bereavement may be linked to a deterioration of parenting skills that are taught as part of a parent group training program for child conduct problems [25]. Finally, program-related factors associated with intervention success include program completion [16] as well as formats involving direct user engagement rather than formats that do not involve direct user engagement when assessing digital interventions [20].

Based on the demonstrated efficacy of the internet-based and telephone-assisted Finnish Strongest Families Parent (FSFP) training program in reducing children's disruptive behavior both in RCT and population implementation settings [26,27], the present study aimed to identify factors that predict poor and favorable program outcomes when applied in a real-world setting. To address key gaps in the literature, we conducted this investigation within an implementation context, using population-based screening and a large sample size to assess factors potentially influencing the efficacy and adequacy of the FSFP. Specifically, we aimed to evaluate the outcomes at 6-month follow-up in families who had completed the FSFP, to assess the distribution of the outcomes, and to identify factors associated with these different outcomes.

Methods

Study Design

The participants (n=3204) were derived from an implementation study spanning over 100 child health clinics in 15 administrative regions in Finland. All families who participated in the FSFP implementation study and who provided data at 6 months follow-up were included in this study. The families were originally recruited to the FSFP program between February 8, 2015, and October 9, 2022. The target group comprised children aged 4 years attending routine medical check-ups at primary care child health clinics. Child health clinic services are free primary health care services provided to families with children under school age in Finland. At least 15 child health check-ups are offered with intervals varying from weeks in infancy to months later and about a year before school age. Before the health check-up, the families received the screening survey by post and were requested to complete it and bring it along to the health check-up. Participants who had not completed the survey were offered the opportunity to do so at the consultation. This screening survey was collected in paper format and included demographic information and the Strengths and Difficulties Questionnaire (SDQ). We requested the participants to complete the baseline (Child Behavior Checklist 1.5-5 [CBCL], Inventory of Callous-Unemotional Traits [ICU], Parenting Scale [PS], 21-Item Depression, Anxiety and Stress Scale [DASS-21]; see Measures) and follow-up CBCL measures online.

From a study population of 49,504, a total of 6426 (13.1%) families met the screening criteria and consented (Figure 1). The screening criteria were a score of ≥5 on the conduct scale of SDQ [28], and at least minor difficulties in mood, concentration, behavior, or getting along with other people over the past 6 months, according to parental report, and a willingness to be contacted by this study's team. The cutoff score of ≥5 was based on the SDQ conduct score 80th percentile cutoff point in a population-based sample of 931 Finnish children aged 4 years [26,29].

XSL•FO
RenderX

Figure 1. Flow chart of the screening procedures. CBCL: Child Behavior Checklist 1.5-5; SDQ: Strengths and Difficulties Questionnaire.

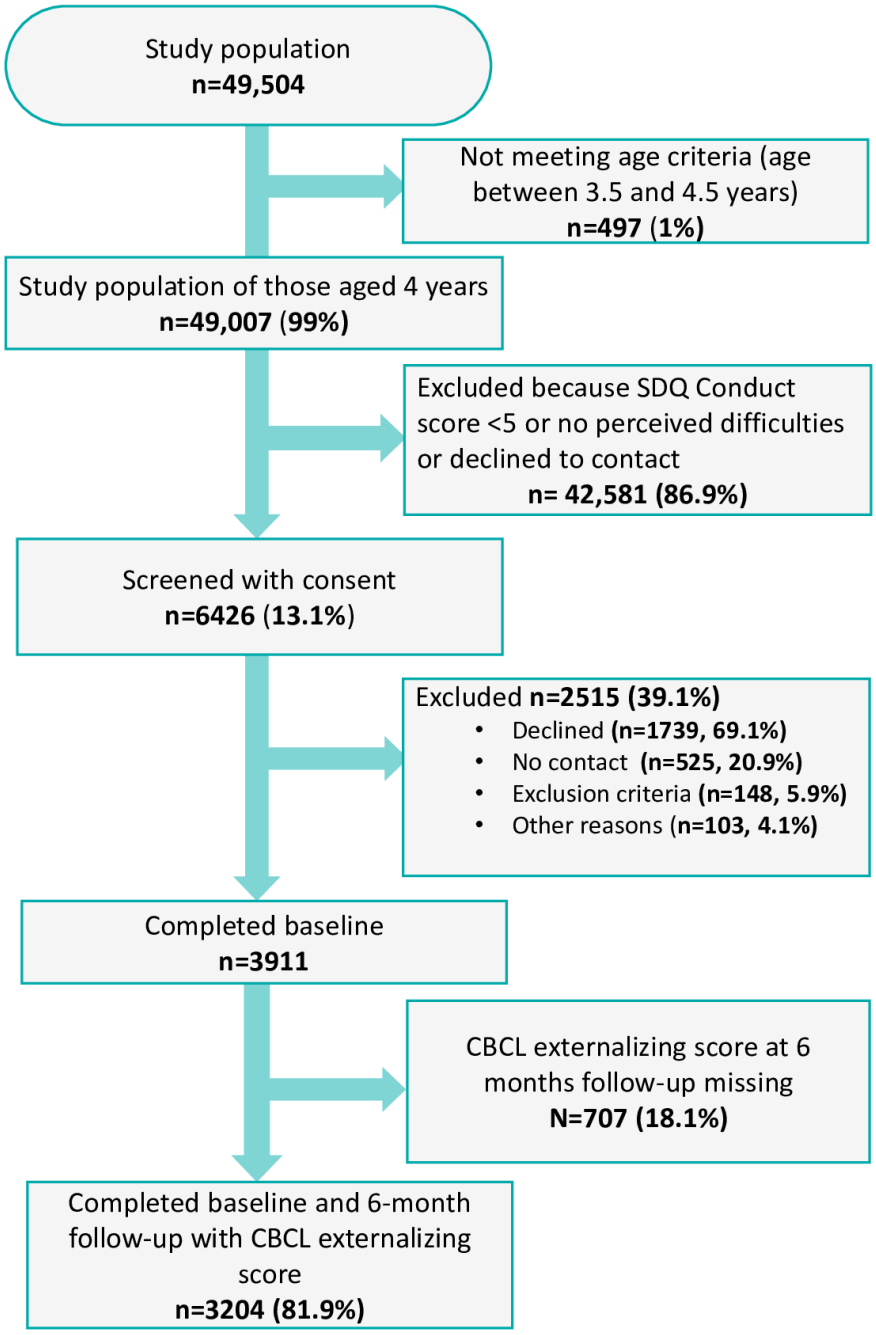

A total of 3911 participants eligible for the FSFP completed the baseline measure. Further, 707 (18.1%) participants did not complete the CBCL externalizing measure at 6 months follow-up, which resulted in a total of 3204 (81.9%) participants meeting the inclusion criteria for this study assessing predictors of FSFP outcomes at 6-month follow-up. The participants had to meet the abovementioned screening criteria; have access to a telephone and internet at home, have fluency in Finnish or Swedish by at least 1 parent, residence in one of the administrative regions participating in the program, and a completed 6-month CBCL externalizing score. Exclusion criteria included autism, Down syndrome, fetal alcohol syndrome, an intellectual disability, or severe mental disorders.

Ethical Considerations

This study received ethical approval from the Ethics Committee of the University of Turku (drno. 44/2014 and 25/2018). Study permission was given by local administrations from participating regions in Finland. Written informed consent for this study was obtained from parents participating in this study, and they were free to decline participation in this study at any time. This study's data were deidentified by replacing personal information

XSL•FO
RenderX

with a study subject code. The ID code key is kept separate from other research data. Participants did not receive any kind of compensation, monetary or otherwise. No identifying information or images of individual participants are included in this paper or its supplementary materials.

Intervention

The FSFP parent training program aims to strengthen skills to improve parent-child relationships. Important elements of the program include supporting parents in recognizing positive behaviors in their child and reinforcing those behaviors by responding to them with positive methods. The 11-week program consists of web-based materials (eg, psychoeducational material, videos, and exercises) and weekly telephone coaching by trained family coaches (licensed health care or social work professionals). Telephone coaching allows for a goal-oriented and personalized approach to solving the challenges that the family is experiencing. The family coach guides the parent in practicing skills that strengthen their parenting, such as positive interaction and rewarding, sharing attention, preempting transitions, planning everyday situations, collaborating with daycare staff, and managing restraint of the child [26,27]. Each week consists of a specific theme, which the parents complete before progressing to the following theme. The last sessions focus on solidifying the newly learned positive parenting skills. The web-based material consists of an introduction, the session content, video exercises, troubleshooting, review, and practical application of the new skills. Each telephone coaching session is scheduled at the conclusion of the preceding weekly call and is allotted a duration of 1 hour. Missed appointments are followed up by the family coach.

Implementation

The FSFP was implemented in the various regions of the Finnish health care service system starting in 2015. The implementation plan included staff recruitment, training, supervision, performance evaluation, and administrative support. Given that the fidelity of an intervention delivery may be compromised when transferring it from a controlled research context to a community setting [30], the FSFP program uses a centralized treatment provider unit, the Research Center for Child Psychiatry at the University of Turku. The coaches are trained on the digital program protocol, and about 10% of calls are randomly audited for quality. Additional training is provided when needed.

FSFP has become part of routine child health clinic services, and families meeting the inclusion criteria are eligible to participate. Child health care nurses are kept informed about client progress, including screening summaries, consent updates, and treatment outcomes. Stakeholders in child and family services are engaged, sign annual research agreements, and receive monthly progress reports. Regular training and updates support the integration into primary care, and media campaigns have raised public awareness. Centralized delivery and development of the digital platform are key administrative strategies, ensuring high implementation fidelity—adherence to the strategy and adaptation to local practices.

Measures

The demographic data included child's sex (boy or girl), family structure (2 biological parents, 1 biological parent, or 2 nonbiological parents), parental age (<26 y/26–40 y/>40 y), and parental education (basic education, secondary education, or upper degree in applied sciences or university).

The CBCL [31] is a parent-reported measure of children's emotional, behavioral, and social problems. It consists of 7 subscales: emotionally reactive, anxious or depressed, somatic complaints, withdrawn, sleep problems, attention problems, and aggressive behavior. The first 4 syndromes form an internalizing score, and the latter 2 an externalizing score. The externalizing score contains 24 items and was used to measure disruptive behavior in this study (ie, the primary outcome measure). Higher scores indicate a higher degree of problems. In the current sample, the internal consistencies were good for both the internalizing and externalizing subscales (0.86 and 0.87, respectively). For the analyses, we used the CBCL baseline and 6-month measures.

The ICU [32] is a 24-item scale used to measure callous-unemotional traits in children, where higher scores indicate a greater degree of callous-unemotional traits. The instrument has been shown to have good reliability and adequate validity for children aged 3-4 years, with a good internal consistency for the total score (0.93) [33]. In our sample, the Cronbach α for the ICU total score was 0.83. We used the ICU baseline measure in the analyses.

The PS was used to assess parenting skills [34]. This 30-item questionnaire measures 3 factors of dysfunctional parental discipline practices: laxness, overreactivity, and verbosity. Higher scores indicate a greater degree of dysfunctional parental discipline practices. The internal consistency for the total score was 0.77 in this study. We used the PS baseline measure in the analyses.

The Depression, Anxiety, and Stress Scale [35] was used to measure parents' stress, anxiety, and depression symptoms. Higher scores indicate more symptoms. The short version of the instrument consists of 21 items (DASS-21) and has previously shown good internal consistency with Cronbach α of 0.91 (depression items), 0.80 (anxiety items), and 0.84 (stress items) in a large nonclinical sample of US adults [36]. The Cronbach α in our sample was 0.87, 0.75, and 0.84 for the depression, anxiety, and stress scales, respectively. We used the DASS-21 baseline measure in the analyses.

As for program measures, completion of at least 7 of the 11 program weeks was considered program completion, as the first 7 weeks of the intervention introduce the core elements of the intervention, while the final 4 weeks focus on reinforcing the acquired skills.

The program completion periods were defined as 2015 - 2017, March 10, 2018, 2020 (ie, before the start of the COVID-19 pandemic), March 11, 2020, to December 31, 2020 (the acute COVID period, the World Health Organization declared the COVID-19 outbreak a global pandemic and many restrictions were in place), 2021 (the Finnish government stated that COVID-19 no longer constituted a state of emergency in

XSL•FO
RenderX

Finland; vaccinations started on December 27, 2020), and 2022 - 2023 (the end of restrictions and the post-COVID period).

Analytic Strategy

The categorical family and child characteristics (as defined in the Measures paragraph) are presented as numbers and percentages, while the continuous program, family, and child characteristics are reported as means and SDs. The distributions of the CBCL externalizing scores at baseline (before starting the program) and at 6-month follow-up are presented using histograms. The outcome variable was derived from the CBCL externalizing score using the following formula:

$$\text{Standardized change} = \frac{\text{Mean baseline} - \text{measurement at 6 months}}{\text{SD baseline}} = 1,3204$$

In this formula, standardized change refers to the change in the CBCL externalizing score; mean (baseline) and SD (baseline) refer to the mean and SD of the CBCL externalizing score at baseline; and measurement at 6 months refers to each participant's individual CBCL externalizing score at 6-month follow-up. Our derivation is a variant of a standard z-transformation, where our interest is to standardize each individual's 6-month change, that is, how many SDs each individual deviates from the mean baseline, instead of calculating how many SDs each individual deviates from the mean at the same time point. Although the children participating in the FSFP had all elevated levels of disruptive behavior at baseline, variation in externalizing behavior levels was still high (quartile range 17 - 27). Hence, children with more pronounced symptoms have more room for improvement, whereas for children with mild symptoms, only a small reduction of symptoms is possible, indicating a floor effect [37]. Therefore, we standardized the change in CBCL externalizing score by subtracting the mean at baseline from the 6-month measurement divided by the SD at baseline to compare individual changes more effectively. The standardized change was categorized so that a ±0.5 SD change indicated no change; +0.5 to 1.5 SD indicated a moderate improvement, and a change >1.5 SD indicated a large improvement. A negative change of more than −0.5 SD indicated a deterioration. We based our categorization on changes of 0.5 SD, which is an empirically derived criterion based on psychological theory [38,39] and established practice [40].

The associations of individual predictors of the outcome variable were analyzed with multinomial logistic regression models. Significant predictors were selected for the adjusted models based on a statistical significance level of 0.10. The point estimates are presented as odds ratios with 95% CIs and *P* values. A *P* value less than .05 was considered significant in the final model. The baseline scores of the psychometric scales were standardized before the analyses. All statistical analyses were performed with SAS (9.4 software; SAS Institute). In terms of analyses, the relationship of each predictor with the FSFP outcome was first examined; separate multinomial models of child characteristics and of parent or family characteristics were then performed to identify variables that remained associated within each domain. Finally, a combined multinomial analysis including child, parent, or family and program characteristics determined the independent predictors of the CBCL externalizing change categories. Multiple imputations were not performed as values were not missing at random.

Results

The mean CBCL at baseline was 21.6 (SD 7.3) and at 6 months 16.1 (SD 8.1). The mean change from baseline was 5.5 (SD 6.7), Cohen *d* 0.83. Altogether, 77% (2468/3204) of the participants showed a favorable change (ie, reduction of externalizing behavior) at 6-month follow-up. A total of 26% (833/3204) showed large (more than 1.5 SD) improvement, and 34.4% (1103/3204) showed moderate (0.5 - 1.5 SD) improvement in the outcome variable (ie, change in CBCL externalizing). When the magnitude of change was smaller than 0.5 SD, the change was considered minor. A total of 16.6% (532/3204) experienced a minor improvement and 10% (319/3204) a minor decline. A total of 9.3% (297/3204) and 3.8% (120/3204) showed a moderate (−0.5 to −1.5 SD) or large (more than −1.5 SD) deterioration, respectively (Figure 2). Demographic and program information is presented in Table 1.

Figure 2. Distributions of CBCL externalizing (**A**) before starting the program (baseline), (**B**) at six months after starting the program, and (**C**) after standardization, where participants' CBCL externalizing measurement at six months after starting the program was reduced from the baseline mean of the CBCL externalizing divided by its SD. CBCL: Child Behavior Checklist 1.5-5.

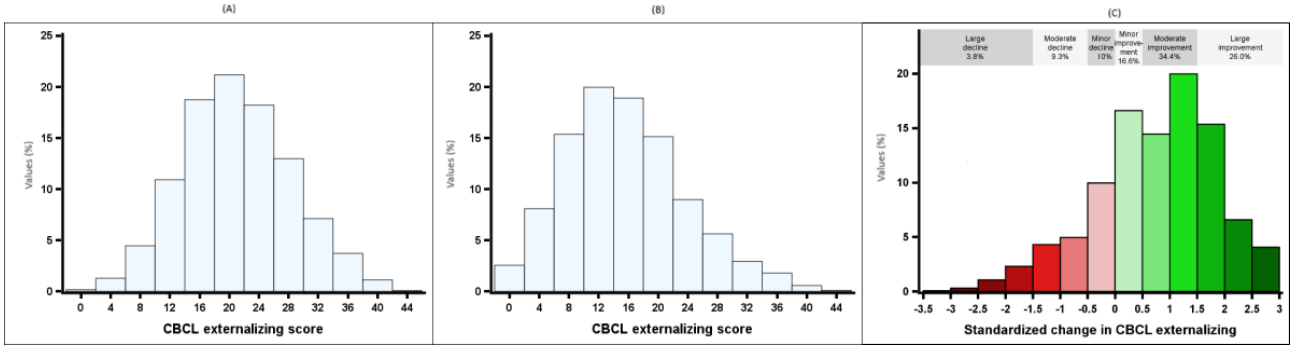

XSL•FO
RenderX

Table . Frequencies of child, parent, or family and program characteristics by the categorized standardized change in the CBCL[a] externalizing score at 6 months after a parent training intervention targeting disruptive behavior in those aged 4 years in real-world settings, 2015 - 2022. Missing observations per variable: sex (n=18), duration (n=45), adverse life events (n=239), family structure (n=10), maternal age (n=26), paternal age (n=117), maternal education (n=23), and paternal education (n=160).

Variable	All (n=3204)	Large improvement (n=833)	Moderate improvement (n=1103)	No change (n=851)	Moderate deterioration (n=297)	Large deterioration (n=120)
Child characteristics						
Sex, n (%)						
Female	1158 (36.35)	354 (11.1)	422 (13.25)	271 (8.51)	84 (2.64)	27 (0.85)
Male	2028 (63.65)	474 (14.9)	677 (21.25)	575 (18.05)	211 (6.62)	91 (2.86)
Difficulties, n (%)						
Minor	1407 (43.91)	527 (16.45)	537 (16.76)	287 (8.96)	49 (1.53)	7 (0.22)
Definite	1497 (46.72)	274 (8.55)	499 (15.57)	474 (14.79)	184 (5.74)	66 (2.06)
Severe	300 (9.36)	32 (1.00)	67 (2.09)	90 (2.81)	64 (2.00)	47 (1.47)
Duration of difficulties (months), n (%)						
6	934 (29.57)	347 (10.98)	346 (10.95)	178 (5.63)	48 (1.52)	15 (0.47)
6 - 12	853 (27.0)	236 (7.47)	295 (9.34)	226 (7.15)	81 (2.56)	15 (0.47)
>12	1372 (43.43)	237 (7.50)	447 (14.15)	433 (13.71)	166 (5.25)	89 (2.82)
Adverse life events, n (%)						
None	2013 (67.89)	553 (18.65)	715 (24.11)	515 (17.37)	163 (5.50)	67 (2.26)
1	732 (24.69)	177 (5.97)	235 (7.93)	209 (7.05)	80 (2.70)	31 (1.05)
≥2	220 (7.42)	36 (1.21)	79 (2.66)	63 (2.12)	31 (1.05)	11 (0.37)
Psychometric scales, mean (SD)						
CBCL internalizing	−0.01 (0.98)	−0.46 (0.77)	−0.11 (0.89)	0.20 (0.96)	0.61 (1.14)	0.85 (1.18)
ICU[b] score	−0.00 (0.99)	−0.40 (0.93)	−0.11 (0.90)	0.24 (0.94)	0.46 (0.94)	0.94 (1.07)
Family characteristics						
Family structure, n (%)						
2 biological parents	2721 (85.19)	729 (22.82)	952 (29.81)	714 (22.35)	239 (7.48)	87 (2.72)
One biological parent	362 (11.33)	74 (2.32)	114 (3.57)	102 (3.19)	50 (1.57)	22 (0.69)
Two nonbiological parents	111 (3.48)	28 (0.88)	35 (1.10)	30 (0.94)	8 (0.25)	10 (0.31)
Maternal age (years), n (%)						
<26	444 (13.97)	113 (3.56)	150 (4.72)	106 (3.34)	47 (1.48)	28 (0.88)
26 - 40	2657 (83.61)	693 (21.81)	925 (29.11)	711 (22.37)	242 (7.61)	86 (2.71)
>40	77 (2.42)	21 (0.66)	19 (0.60)	26 (0.82)	6 (0.19)	5 (0.16)
Paternal age (years), n (%)						
<26	261 (8.45)	64 (2.07)	95 (3.08)	52 (1.68)	35 (1.13)	15 (0.49)
26 - 40	2574 (83.38)	682 (22.09)	878 (28.44)	696 (22.55)	227 (7.35)	91 (2.95)
>40	252 (8.16)	58 (1.88)	94 (3.05)	69 (2.24)	21 (0.68)	10 (0.32)
Maternal education, n (%)						
Basic education	50 (1.57)	12 (0.38)	15 (0.47)	10 (0.31)	7 (0.22)	6 (0.19)
Secondary education	951 (29.90)	222 (6.98)	328 (10.31)	259 (8.14)	91 (2.86)	51 (1.60)

XSL•FO
RenderX

Variable	All (n=3204)	Large improvement (n=833)	Moderate improvement (n=1103)	No change (n=851)	Moderate deterioration (n=297)	Large deterioration (n=120)
Upper degree in applied sciences	1059 (33.29)	287 (9.02)	374 (11.76)	271 (8.52)	92 (2.89)	35 (1.10)
University	1121 (35.24)	308 (9.68)	379 (11.91)	304 (9.56)	104 (3.27)	26 (0.82)
Paternal education, n (%)						
Basic education	144 (4.73)	41 (1.35)	46 (1.51)	36 (1.18)	16 (0.53)	5 (0.16)
Secondary education	1346 (44.22)	321 (10.55)	472 (15.51)	359 (11.79)	123 (4.04)	71 (2.33)
Upper degree in applied sciences	757 (24.87)	206 (6.77)	251 (8.25)	204 (6.70)	71 (2.33)	25 (0.82)
University	797 (26.18)	231 (7.59)	282 (9.26)	203 (6.67)	68 (2.23)	13 (0.43)
Psychometric scales, mean (SD)						
Parenting Scale	−0.02 (0.98)	−0.22 (0.99)	−0.02 (0.95)	0.13 (0.96)	0.21 (1.06)	−0.10 (0.95)
DASS[c]-depression	−0.05 (0.97)	−0.26 (0.82)	−0.09 (0.93)	0.04 (0.97)	0.36 (1.20)	0.23 (1.15)
DASS-anxiety	−0.03 (0.95)	−0.20 (0.80)	−0.08 (0.88)	0.01 (0.92)	0.37 (1.21)	0.35 (1.44)
DASS-stress	−0.03 (0.98)	−0.33 (0.93)	-0.06 (0.93)	0.08 (0.96)	0.45 (1.07)	0.30 (1.12)
Intervention characteristics						
Program completion, n (%)						
Completed <7 themes	70 (2.18)	19 (0.59)	29 (0.91)	16 (0.50)	1 (0.03)	5 (0.16)
Completed ≥7 themes	3134 (97.82)	814 (25.41)	1074 (33.52)	835 (26.06)	296 (9.24)	115 (3.59)
Period of completion, n (%)						
2015 - 2017	482 (15.04)	135 (4.21)	183 (5.71)	114 (3.56)	33 (1.03)	17 (0.53)
2018 to March 10, 2020	975 (30.43)	240 (7.49)	334 (10.42)	260 (8.11)	108 (3.37)	33 (1.03)
March 11, 2020, to December 31, 2020	429 (13.39)	104 (3.25)	145 (4.53)	127 (3.96)	38 (1.19)	15 (0.47)
2021	531 (16.57)	131 (4.09)	172 (5.37)	154 (4.81)	49 (1.53)	25 (0.78)
2022 - 2023	662 (20.66)	187 (5.84)	221 (6.90)	169 (5.27)	61 (1.90)	24 (0.75)

[a]CBCL: Child Behavior Checklist 1.5-5.

[b]ICU: Inventory of Callous-Unemotional Traits.

[c]DASS: 21-Item Depression, Anxiety and Stress Scale.

The relationships of each of the individual predictor variables of the outcome of the FSFP program are presented in Multimedia Appendix 1. The multinomial analysis (Table 2), including only child characteristics as predictors, showed that sex, internalizing symptoms, severity, and duration of problems, as well as callous-unemotional traits, were significantly associated with the different categories of the outcome. The multinomial analysis, including only parent or family characteristics, showed that maternal age >40 years, paternal age <26 years, mother's education, parenting skills, parental anxiety, and stress were significantly associated with the outcome.

XSL•FO
RenderX

Table . Multivariable multinomial logistic regression analyses of the associations of child and family-level predictors, respectively, on the standardized change in CBCL[a] externalizing as categorized into large improvement, moderate improvement, no change (reference group, n=851), and deterioration at 6 months after a parent training intervention targeting disruptive behavior in those aged 4 years in real-world settings, 2015 - 2022. Missing observations per variable: sex (n=18), duration (n=45), adverse life events (n=239), family structure (n=10), maternal age (n=26), paternal age (n=117), maternal education (n=23), and paternal education (n=160).

Variable	Large improvement (n=833), OR[b] (95% CI)	*P* value	Moderate improvement (n=1103), OR (95% CI)	*P* value	Deterioration (n=417), OR (95% CI)	*P* value
Child characteristics						
Sex	0.65 (0.52 - 0.81)	<.001	0.80 (0.66 - 0.98)	.03	1.24 (0.94 - 1.63)	.13
CBCL internalizing (baseline)	0.51 (0.44 - 0.58)	<.001	0.80 (0.72 - 0.89)	<.001	1.28 (1.13 - 1.44)	<.001
ICU[c] total (baseline)	0.6 (0.5 - 0.7)	<.001	0.75 (0.67 - 0.83)	<.001	1.24 (1.09 - 1.41)	.001
Duration of difficulties						
6 - 12 months vs <6 months	0.75 (0.57 - 0.99)	.045	0.79 (0.61 - 1.03)	.08	0.96 (0.65 - 1.41)	.84
>12 months vs <6 months	0.45 (0.34 - 0.58)	<.001	0.66 (0.53 - 0.84)	<.001	1.14 (0.81 - 1.61)	.44
Severity at screening						
Definite versus minor	0.45 (0.36 - 0.56)	<.001	0.66 (0.54 - 0.81)	<.001	2.35 (1.67 - 3.29)	<.001
Severe versus minor	0.43 (0.27 - 0.68)	<.001	0.57 (0.40 - 0.83)	.003	4.40 (2.87 - 6.75)	<.001
Parent or family characteristics						
Family structure						
One biological parent versus 2 biological parents	0.81 (0.52 - 1.26)	.36	0.80 (0.54 - 1.19)	.27	1.16 (0.74 - 1.84)	.52
Nonbiological parents versus 2 biological parents	0.91 (0.49 - 1.70)	.78	0.92 (0.52 - 1.63)	.78	1.06 (0.52 - 2.15)	.88
Mother's age group (years)						
≤26 versus 26 - 40	1.09 (0.73 - 1.62)	.67	0.91 (0.63 - 1.30)	.59	1.02 (0.65 - 1.61)	.94
>40 versus 26 - 40	1.01 (0.49 - 2.08)	.99	0.40 (0.18 - 0.87)	.02	0.79 (0.30 - 2.11)	.64
Father's age group (years)						
≤26 versus 26 - 40	1.54 (0.94 - 2.54)	.09	1.68 (1.07 - 2.66)	.03	1.55 (0.89 - 2.72)	.12
>40 versus 26 - 40	0.86 (0.56 - 1.32)	.49	1.21 (0.84 - 1.75)	.31	0.90 (0.54 - 1.50)	.68
Mother's education level						
Basic education versus university	1.69 (0.54 - 5.34)	.37	1.60 (0.53 - 4.83)	.41	2.79 (0.87 - 8.92)	.08
Secondary education versus university	0.90 (0.66 - 1.24)	.53	1.02 (0.76 - 1.37)	.88	0.95 (0.64 - 1.40)	.78
Upper degree in applied sciences versus a university degree	1.04 (0.79 - 1.36)	.80	1.13 (0.88 - 1.45)	.36	0.95 (0.68 - 1.34)	.79

XSL•FO
RenderX

Variable	Large improvement (n=833), OR[b] (95% CI)	*P* value	Moderate improvement (n=1103), OR (95% CI)	*P* value	Deterioration (n=417), OR (95% CI)	*P* value
Father's education level						
Basic education versus university	1.07 (0.62 - 1.86)	.81	0.89 (0.52 - 1.51)	.66	1.02 (0.51 - 2.03)	.95
Secondary education versus university	0.79 (0.58 - 1.06)	.12	0.91 (0.69 - 1.20)	.49	1.12 (0.77 - 1.65)	.55
Upper degree in applied sciences versus a university degree	0.79 (0.58 - 1.07)	.12	0.82 (0.62 - 1.09)	.16	1.07 (0.72 - 1.57)	.75
Psychometric scales						
Parenting Scale total score	0.80 (0.71 - 0.90)	<.001	0.90 (0.81 - 1.01)	.07	0.86 (0.75 - 1.00)	.047
DASS[d]-depression	0.93 (0.79 - 1.09)	.35	0.98 (0.86 - 1.12)	.77	1.06 (0.90 - 1.24)	.51
DASS-anxiety	1.11 (0.95 - 1.30)	.20	1.03 (0.90 - 1.18)	.63	1.21 (1.04 - 1.41)	.01
DASS-stress	0.66 (0.56 - 0.77)	<.001	0.88 (0.76 - 1.02)	.09	1.17 (0.97 - 1.40)	.10
Adverse life events						
1 versus none	0.85 (0.65 - 1.11)	.24	0.84 (0.66 - 1.07)	.15	1.01 (0.73 - 1.38)	.98
At least 2 versus none	0.87 (0.52 - 1.44)	.58	1.14 (0.74 - 1.75)	.55	1.13 (0.68 - 1.91)	.63

[a]CBCL: Child Behavior Checklist 1.5-5.

[b]OR: odds ratio.

[c]ICU: Inventory of Callous-Unemotional Traits.

[d]DASS: 21-Item Depression, Anxiety and Stress Scale.

The final multinomial analysis of the effects of child, parent, or family, and program characteristics on the CBCL externalizing change categories is presented in Table 3. The statistically significant predictors related to child characteristics that were associated with limited improvement were the child being a male, internalizing symptoms, higher callous-unemotional traits, longer duration of difficulties, and higher severity at screening. The significant parent or family variables associated with reduced improvement included mother's age >40 years, anxiety, and stress. Higher scores in dysfunctional parenting practices (PS) were associated with a lower likelihood of large improvement. The year of program completion, 2021, was associated with a lesser likelihood of a large or moderate improvement.

XSL•FO
RenderX

Table . Final multinomial model of the associations of all the predictors on the standardized change in CBCL[a] externalizing as categorized into large improvement, moderate improvement, no change (reference group, n=851), and deterioration at 6 months after a parent training intervention targeting disruptive behavior in those aged 4 years in real-world settings, 2015 - 2022. Missing observations per variable: sex (n=18), duration (n=45), adverse life events (n=239), family structure (n=10), maternal age (n=26), paternal age (n=117), maternal education (n=23), and paternal education (n=160).

Variable	Large improvement (n=833), OR[b] (95% CI)	*P* value	Moderate improvement (n=1103), OR (95% CI)	*P* value	Deterioration (n=417), OR (95% CI)	*P* value
Child characteristics						
Boy versus girl	0.65 (0.52 - 0.82)	<.001	0.80 (0.6 - 0.98)	.03	1.26 (0.95 - 1.68)	.11
CBCL internalizing (baseline)	0.54 (0.47 - 0.63)	<.001	0.80 (0.72 - 0.90)	<.001	1.15 (1.01 - 1.30)	.04
ICU[c] total (baseline)	0.64 (0.57 - 0.73)	<.001	0.77 (0.69 - 0.86)	<.001	1.32 (1.15 - 1.51)	<.001
Duration of difficulties (months)						
6 - 12 versus <6	0.75 (0.56 - 0.99)	.04	0.80 (0.62 - 1.05)	.10	0.92 (0.62 - 1.37)	.67
>12 versus <6	0.43 (0.33 - 0.56)	<.001	0.65 (0.51 - 0.83)	<.001	1.09 (0.76 - 1.55)	.65
Severity at screening						
Definite versus minor	0.45 (0.35 - 0.56)	<.001	0.68 (0.55 - 0.83)	<.001	2.29 (1.62 - 3.24)	<.001
Severe versus minor	0.48 (0.30 - 0.78)	.003	0.62 (0.43 - 0.91)	.01	4.38 (2.80 - 6.85)	<.001
Parent or family characteristics						
Mother's age group (years)						
≤26 versus 26 - 40	1.30 (0.88 - 1.94)	.19	1.00 (0.70 - 1.43)	.98	0.90 (0.57 - 1.43)	.66
>40 versus 26 - 40	1.09 (0.51 - 2.31)	.83	0.40 (0.19 - 0.86)	.02	0.97 (0.40 - 2.36)	.95
Father's age group (years)						
≤26 versus 26 - 40	1.11 (0.67 - 1.82)	.69	1.46 (0.94 - 2.27)	.09	1.67 (0.96 - 2.91)	.07
>40 versus 26 - 40	0.82 (0.53 - 1.27)	.38	1.22 (0.85 - 1.75)	.27	0.87 (0.53 - 1.45)	.60
Mother's education level						
Basic education versus university	1.14 (0.41 - 3.20)	.80	0.94 (0.36 - 2.48)	.91	2.31 (0.85 - 6.23)	.10
Secondary education versus university	0.75 (0.56 - 1.01)	.054	0.91 (0.70 - 1.17)	.46	1.0 (0.71 - 1.40)	.99
Upper degree in applied sciences versus a university degree	0.93 (0.72 - 1.21)	.59	1.04 (0.82 - 1.31)	.75	1.10 (0.77 - 1.44)	.74
Parenting Scale total score	0.87 (0.77 - 1.00)	.03	0.94 (0.85 - 1.05)	.27	0.90 (0.78 - 1.03)	.13
DASS[d]-anxiety	1.10 (0.94 - 1.28)	.26	1.02 (0.90 - 1.17)	.74	1.20 (1.04 - 1.39)	.02
DASS-stress	0.78 (0.67 - 0.91)	.002	0.96 (0.84 - 1.09)	.50	1.10 (0.94 - 1.30)	.24
Program characteristics						
Year of completion						

XSL•FO
RenderX

Variable	Large improvement (n=833), OR[b] (95% CI)	*P* value	Moderate improvement (n=1103), OR (95% CI)	*P* value	Deterioration (n=417), OR (95% CI)	*P* value
2018 - 10 March 2020 versus 2015 - 2017	0.75 (0.54 - 1.06)	.10	0.83 (0.61 - 1.12)	.21	1.20 (0.79 - 1.84)	.39
11 MAR–31 DEC 2020 versus 2015 - 17	0.69 (0.46 - 1.03)	.07	0.75 (0.53 - 1.08)	.12	1.08 (0.65 - 1.78)	.77
2021 versus 2015 - 2017	0.66 (0.45 - 0.97)	.03	0.69 (0.49 - 0.97)	.03	1.13 (0.70 - 1.83)	.61
2022 - 2023 versus 2015 - 2017	1.00 (0.69 - 1.44)	.99	0.87 (0.62 - 1.20)	.39	1.17 (0.74 - 1.85)	.51
Not completed versus 2015 - 2017	1.18 (0.63 - 2.20)	.60	1.07 (0.61 - 1.87)	.82	0.97 (0.43 - 2.21)	.95

[a]CBCL: Child Behavior Checklist 1.5-5.

[b]OR: odds ratio.

[c]ICU: Inventory of Callous-Unemotional Traits.

[d]DASS: 21-Item Depression, Anxiety and Stress Scale.

Discussion

Principal Findings

This study, including 3204 families screened out of 49,504 families participating in annual medical check-ups for children aged 4 years, represents the largest implementation study to date that has examined the factors predicting the outcome of a parent training program. This study is unique in its use of population-based screening and its exceptionally large implementation sample. The findings apply to real-world contexts, and they have significant implications for public health when planning early interventions for children with disruptive behavior problems. First, 77% (2468/3204) of the children showed a successful program outcome at 6-month follow-up. Second, the severity and duration of the children's problems at baseline were the strongest predictors for limited improvement. Third, parental anxiety or stress at baseline predicted limited improvement. Finally, program completion during the COVID-19 pandemic was associated with a reduced likelihood of improvement.

Overall, a clear majority of children in this study showed successful program outcomes at 6 months, which is in line with earlier implementation studies of the FSFP [27]. Literature on parent training remains limited in studies conducted in an implementation context where data are analyzed based on the type of outcome [5-7]. Such research is essential for a comprehensive understanding of the mechanisms and factors underlying the changes induced by a parent training program. For example, it is possible that the overall efficacy of parent training interventions is driven by a subgroup of children who show significant improvement [41]. However, in our study, 13% of the children showed a clear deterioration at 6-month follow-up. This group may reflect intervention resistance, indicating a need for different or additional support measures. Another possibility is that the 6-month follow-up after baseline—about 3 months postintervention—was too short to capture longer-term benefits. Future research should expand beyond group-level results by assessing patterns behind different types of outcomes.

Child-Related Factors

As expected, there were a number of factors that influenced the levels of outcome, but overall, child characteristics had the strongest effect. Boys had a lower likelihood of large improvement and a higher likelihood of deterioration compared to girls, which suggests that girls responded to the program better than boys. Higher internalizing symptoms and callous-unemotional traits at baseline were associated with a lower likelihood of a favorable treatment response. Previously, callous-unemotional traits have been linked to poor treatment outcomes in family interventions for conduct problems [42]. Children with severe and definite initial behavioral difficulties had a fourfold and twofold risk of deterioration compared to those with minor difficulties, respectively. A longer duration of initial difficulties was associated with a lower likelihood of a favorable outcome, but not with a higher likelihood of a poor outcome. It is also plausible that experiencing difficulties over a prolonged period contributes to establishing unhelpful functioning and interaction patterns within the family system, which may require additional resources to change [43]. These findings highlight the importance of early identification of child disruptive behavior and tailored interventions designed to address each child's specific needs. A marked presence of psychological maladjustment, along with a longer duration and greater severity of behavioral problems, can reflect psychopathological processes that may require different or additional forms of interventions.

Parent and Family Factors

Parental stress and anxiety symptoms were also associated with outcomes. This may happen through a variety of mechanisms. For example, stress is linked to impaired, altered, and rigid memory functions [44], which may hinder parents' ability to learn new skills. Greater parental mentalization skills are linked

XSL•FO
RenderX

with lower externalizing behavior in their children [45]; however, anxiety is associated with deficits in mentalization [46]. Effective parenting requires the regulation of both internal (the parents') and external (the child's) emotional states [47,48], but psychopathological processes, again, are typically characterized by deficits in emotion regulation [49]. It may be beneficial to remain attentive to the psychosocial well-being of the parents throughout the course of a parent training program. Additionally, incorporating targeted supportive measures for parents with psychopathological symptoms at the starting phase of a parent training program may enhance treatment outcomes [50]. Such measures could reduce symptomatology, thereby making the training seem less demanding and more manageable for the parents. This, in turn, could facilitate more efficient skill acquisition and make the training more rewarding, thereby improving training outcomes and reducing dropouts [50].

The results indicated that maternal age over 40 years was linked to a decreased likelihood of moderate improvement. All other associations between parental age and outcomes were no longer significant in the final multinomial model. Although factors such as family structure and parents' education level were individually predictive, they lost significance when tested alongside other factors. This reflects earlier findings [41] and suggests that the intervention is adaptable to diverse family backgrounds, irrespective of socioeconomic factors, which is an important element of equity in service quality and accessibility.

It should be noted that our findings in this study and previous literature show some key differences. The results derived from individual participant data meta-analyses [51-55] showed that the effects of the Incredible Years parenting intervention on disruptive behavior did not differ among families with different social and socioeconomic disadvantage, families with different styles of parenting, and children with different comorbidities, and severity of problems, sibling conflict, and parental depression were linked to stronger intervention effects. Several factors may explain these differences: the Incredible Years intervention was delivered in group format and face-to-face, and the samples spanned larger age ranges (from 1 - 2 y up to 10 - 12 y) [51-55]. In the present study, we used a standardized measure of individual change to mitigate possible floor effects and to show the individual change relative to baseline to identify factors predicting poor and favorable program outcomes.

Program-Related Factors

Lastly, regarding program characteristics, completion during 2021 was linked to decreased improvement compared to completion during 2015 - 2017. This may reflect additional stressors that families were experiencing during the COVID-19 public health restrictions (eg, lockdowns). Although these public health measures were introduced early in the pandemic in 2020, they persisted throughout 2021 [56]. Such stressors may have caused cumulative strain on parents and their children, potentially affecting parents' ability to engage in positive parenting practices as the restrictions were prolonged [57,58]. It could be, for example, that at the beginning of the pandemic, the novel situation may have been perceived as novel and engaging, and families generally had sufficient resources to adapt to the new reality. Over time, however, as the restrictions persisted, everyday life may have been perceived as increasingly stressful [59]. This could have compromised parents' resources to support their children and to consistently and effectively implement the FSFP skills training during the later stages of the pandemic compared to earlier phases.

The implementation of the FSFP parent training program involves integrating an empirically supported intervention into the health care service system. The FSFP has been implemented for over a decade [27] and is aligned with the existing service structure, which supports its feasibility and sustainability. Of note, the program completion rate was relatively high. We believe the most important contributing factors are the proactive screening process that identifies and engages parents at the onset of difficulties, and the fact that the intervention is not an unguided digital intervention. The intervention includes weekly human support from a designated family coach with whom the parent can build a working alliance. The family coach helps motivate and encourage the parent, clarify the content, support with challenging situations, and support the skills training. This likely translates into visible results in the family's everyday life, which further encourages the participation. The remote delivery format is again likely to remove certain access barriers.

Limitations

Several limitations of this study warrant discussion. The grouping of participants according to the type of outcome was based on statistical differences. While the statistics-based metric of change is commonly used and based on psychological theory, it may neither accurately capture the practical or clinical significance of these differences [60], nor reflect the participants' subjective perceptions of the outcomes. Further, the outcome and child psychopathology measures are based solely on parent reports. It is possible that parental psychopathology affected their perceptions and appraisals of the child's behavior [61]. Conversely, for young children who are under school age, the availability of other informants, such as teachers, is limited, and self-reports by children may not be a feasible option. Another limitation involves the participants for whom 6-month follow-up data were unavailable. We considered conducting multiple imputation; however, a comparative analysis between the group that provided 6-month data and the group that did not (Multimedia Appendix 2) showed that the groups differed on a number of factors. The data were hence not missing at random, whereby imputation can introduce bias rather than correct for it [62]. As the number of participants who did not provide data at the 6-month follow-up was relatively low (under 20%), we believe it would not cause a profound impact on the results. Participants who did not provide 6-month data tended to have completed fewer than 7 themes of the program, be single-parent families, have a lower parental age and education level, have experienced more adverse events, and exhibit higher scores on the depression, stress, and anxiety scales. However, given the large sample size, even small differences may have reached statistical significance.

The lack of a control group prevents us from drawing definite conclusions about the real-world effectiveness and the outcomes of the implementation strategy. However, we have previously

XSL•FO
RenderX

shown in an RCT that the FSFP intervention was effective up to 24 months of follow-up [63]. In our previous study conducted in an implementation setting (n=600), we found that the 6-month follow-up results did not differ from the RCT study [27]. At the same time, we acknowledge that spontaneous improvement takes place among children exhibiting disruptive behavior [64], and this could have been a factor in some of the improvements observed in the current study.

As for the generalizability of the sample, the screening for the program is based on population-based screening. The participating child health clinics are geographically distributed across the whole country. A previous report [65] examining enrollment in the FSFP program showed that enrollment was associated with more severe and longer-term levels of disruptive behavior and comorbid emotional problems. Lower parental education and younger paternal age were associated with nonenrollment.

Conclusions

The questions of which treatment works for whom, for which difficulties, under which circumstances, and delivered by whom were first raised long ago [66]. This study's contribution to the field includes answering this question by identifying characteristics that may influence intervention performance in real-world settings, thereby allowing for more tailored interventions. An implication of our results in the real world is that children with longer-term and more severe behavioral symptoms may require more individualized approaches, and additional support may be necessary for parents experiencing elevated stress and anxiety. Such efforts could help ensure that more families can receive the optimal benefit from evidence-based parenting interventions.

Our work also broadens the methodological landscape by examining standardized individual change, which is an approach that has been used sparingly in previous parent training implementation studies. Our study further differs from much existing literature on parent training, which, to a large degree, is based on RCTs, while the literature on the implementation of parenting programs in community settings is scarcer. More studies expanding the existing RCT literature by showing factors influencing treatment outcome in real-world settings are needed to help guide parents and children to the right interventions.

Acknowledgments

This research was supported by the Inequalities, Interventions, and a New Welfare State (INVEST) Research Flagship.

Funding

This research was supported by the Research Council of Finland (decision number: 345546). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101020767; ERC Advanced, AS). The funder had no involvement in this study's design, data collection, analysis, interpretation, or the writing of this paper.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to study privacy policy and data protection regulations (including General Data Protection Regulation), but summary measures are available from the corresponding author on reasonable request. Generative artificial intelligence was not used in any portion of this paper's writing.

Conflicts of Interest

AS is the founder and director of Digifamilies, which provides evidence-based digital treatments to Finnish specialized services. Digifamilies did not provide services for the present sample.

Multimedia Appendix 1
Univariate analyses of the associations of individual predictors on the standardized change in CBCL externalizing, as categorized into large improvement, moderate improvement, no change (reference group), and decline. CBCL: Child Behavior Checklist 1.5-5.
[DOCX File, 32 KB - jmir_v28i1e79592_app1.docx]

Multimedia Appendix 2
Comparisons of child and family characteristics between those completing both baseline and 6-month follow-up (n=3204) and those completing only baseline (n=707) questionnaire.
[DOCX File, 21 KB - jmir_v28i1e79592_app2.docx]

References

1. Dretzke J, Frew E, Davenport C, et al. The effectiveness and cost-effectiveness of parent training/education programmes for the treatment of conduct disorder, including oppositional defiant disorder, in children. Health Technol Assess 2005 Dec;9(50):iii. [doi: 10.3310/hta9500] [Medline: 16336845]

XSL•FO
RenderX

2. Weber L, Kamp-Becker I, Christiansen H, Mingebach T. Treatment of child externalizing behavior problems: a comprehensive review and meta-meta-analysis on effects of parent-based interventions on parental characteristics. Eur Child Adolesc Psychiatry 2019 Aug;28(8):1025-1036. [doi: 10.1007/s00787-018-1175-3] [Medline: 29948228]
3. Scott S, Dadds MR. Practitioner review: when parent training doesn't work: theory - driven clinical strategies. Child Psychology Psychiatry 2009 Dec;50(12):1441-1450. [doi: 10.1111/j.1469-7610.2009.02161.x]
4. Tolin DF, McKay D, Forman EM, Klonsky ED, Thombs BD. Empirically supported treatment: recommendations for a new model. Clin Psychol Sci Pract 2015 Dec;22(4):317-338. [doi: 10.1111/cpsp.12122]
5. Lansford JE, Betancourt TS, Boller K, et al. The future of parenting programs: II implementation. Parent Sci Pract 2022;22(3):235-257. [doi: 10.1080/15295192.2022.2086807] [Medline: 36439707]
6. Romano M, Schnurr M. Mind the gap: strategies to bridge the research-to-practice divide in early intervention caregiver coaching practices. Topics Early Child Spec Educ 2022 May;42(1):64-76. [doi: 10.1177/0271121419899163]
7. Michelson D, Davenport C, Dretzke J, Barlow J, Day C. Do evidence-based interventions work when tested in the "real world?" A systematic review and meta-analysis of parent management training for the treatment of child disruptive behavior. Clin Child Fam Psychol Rev 2013 Mar;16(1):18-34. [doi: 10.1007/s10567-013-0128-0] [Medline: 23420407]
8. Nock MK, Kazdin AE, Hiripi E, Kessler RC. Lifetime prevalence, correlates, and persistence of oppositional defiant disorder: results from the National Comorbidity Survey Replication. J Child Psychol Psychiatry 2007 Jul;48(7):703-713. [doi: 10.1111/j.1469-7610.2007.01733.x] [Medline: 17593151]
9. DeRubeis RJ. The history, current status, and possible future of precision mental health. Behav Res Ther 2019 Dec;123:103506. [doi: 10.1016/j.brat.2019.103506] [Medline: 31706160]
10. Reyno SM, McGrath PJ. Predictors of parent training efficacy for child externalizing behavior problems--a meta-analytic review. J Child Psychol Psychiatry 2006 Jan;47(1):99-111. [doi: 10.1111/j.1469-7610.2005.01544.x] [Medline: 16405646]
11. Gardner F, Connell A, Trentacosta CJ, Shaw DS, Dishion TJ, Wilson MN. Moderators of outcome in a brief family-centered intervention for preventing early problem behavior. J Consult Clin Psychol 2009 Jun;77(3):543-553. [doi: 10.1037/a0015622] [Medline: 19485594]
12. McGilloway S, Mhaille GN, Bywater T, et al. A parenting intervention for childhood behavioral problems: a randomized controlled trial in disadvantaged community-based settings. J Consult Clin Psychol 2012 Feb;80(1):116-127. [doi: 10.1037/a0026304] [Medline: 22148879]
13. Stolk MN, Mesman J, van Zeijl J, et al. Early parenting intervention: family risk and first-time parenting related to intervention effectiveness. J Child Fam Stud 2008 Feb;17(1):55-83. [doi: 10.1007/s10826-007-9136-3]
14. Van Zeijl J, Mesman J, Van IJzendoorn MH, et al. Attachment-based intervention for enhancing sensitive discipline in mothers of 1- to 3-year-old children at risk for externalizing behavior problems: a randomized controlled trial. J Consult Clin Psychol 2006 Dec;74(6):994-1005. [doi: 10.1037/0022-006X.74.6.994] [Medline: 17154730]
15. Baker S, Sanders MR. Predictors of program use and child and parent outcomes of a brief online parenting intervention. Child Psychiatry Hum Dev 2017 Oct;48(5):807-817. [doi: 10.1007/s10578-016-0706-8] [Medline: 28035556]
16. Day JJ, Baker S, Dittman CK, et al. Predicting positive outcomes and successful completion in an online parenting program for parents of children with disruptive behavior: an integrated data analysis. Behav Res Ther 2021 Nov;146:103951. [doi: 10.1016/j.brat.2021.103951] [Medline: 34507006]
17. Engelbrektsson J, Salomonsson S, Högström J, Sorjonen K, Sundell K, Forster M. Is internet-based parent training for everyone? Predictors and moderators of outcomes in group vs. internet-based parent training for children with disruptive behavior problems. Behav Res Ther 2023 Dec;171:104426. [doi: 10.1016/j.brat.2023.104426] [Medline: 37924567]
18. Pokorna N, Palmer M, Pearson O, et al. Moderators of the effects of a digital parenting intervention on child conduct and emotional problems implemented during the COVID-19 pandemic: results from a secondary analysis of data from the supporting parents and kids through lockdown experiences (SPARKLE) randomized controlled trial. JMIR Pediatr Parent 2024 Oct 8;7:e53864. [doi: 10.2196/53864] [Medline: 39378100]
19. Shelleby EC, Shaw DS. Outcomes of parenting interventions for child conduct problems: a review of differential effectiveness. Child Psychiatry Hum Dev 2014 Oct;45(5):628-645. [doi: 10.1007/s10578-013-0431-5] [Medline: 24390592]
20. Baumel A, Pawar A, Kane JM, Correll CU. Digital parent training for children with disruptive behaviors: systematic review and meta-analysis of randomized trials. J Child Adolesc Psychopharmacol 2016 Oct;26(8):740-749. [doi: 10.1089/cap.2016.0048] [Medline: 27286325]
21. Fleming GE, Kimonis ER, Furr JM, Comer JS. Internet-delivered parent training for preschoolers with conduct problems: do callous-unemotional traits moderate efficacy and engagement? J Abnorm Child Psychol 2020 Sep;48(9):1169-1182. [doi: 10.1007/s10802-020-00660-5] [Medline: 32533295]
22. Leijten P, Raaijmakers MAJ, de Castro BO, Matthys W. Does socioeconomic status matter? A meta-analysis on parent training effectiveness for disruptive child behavior. J Clinical Child Adolesc Psychol 2013 May 1;42(3):384-392. [doi: 10.1080/15374416.2013.769169]
23. Beauchaine TP, Webster-Stratton C, Reid MJ. Mediators, moderators, and predictors of 1-year outcomes among children treated for early-onset conduct problems: a latent growth curve analysis. J Consult Clin Psychol 2005;73(3):371-388. [doi: 10.1037/0022-006X.73.3.371]

XSL•FO
RenderX

24. Gardner F, Hutchings J, Bywater T, Whitaker C. Who benefits and how does it work? Moderators and mediators of outcome in an effectiveness trial of a parenting intervention. J Clin Child Adolesc Psychol 2010;39(4):568-580. [doi: 10.1080/15374416.2010.486315] [Medline: 20589567]
25. Furlong M, McGilloway S. The longer term experiences of parent training: a qualitative analysis. Child Care Health Dev 2015 Sep;41(5):687-696. [doi: 10.1111/cch.12195] [Medline: 25256901]
26. Sourander A, McGrath PJ, Ristkari T, et al. Internet-assisted parent training intervention for disruptive behavior in 4-year-old children: a randomized clinical trial. JAMA Psychiatry 2016 Apr;73(4):378-387. [doi: 10.1001/jamapsychiatry.2015.3411] [Medline: 26913614]
27. Sourander A, Ristkari T, Kurki M, et al. Effectiveness of an internet-based and telephone-assisted training for parents of 4-year-old children with disruptive behavior: implementation research. J Med Internet Res 2022;24(4):e27900. [doi: 10.2196/27900]
28. Goodman R. Psychometric properties of the strengths and difficulties questionnaire. J Am Acad Child Adolesc Psychiatry 2001 Nov;40(11):1337-1345. [doi: 10.1097/00004583-200111000-00015] [Medline: 11699809]
29. Ilola AM, Lempinen L, Huttunen J, Ristkari T, Sourander A. Bullying and victimisation are common in four-year-old children and are associated with somatic symptoms and conduct and peer problems. Acta Paediatr 2016 May;105(5):522-528. [doi: 10.1111/apa.13327] [Medline: 26741067]
30. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. Mayo Clin Proc 2002 Apr;77(4):371-383. [doi: 10.4065/77.4.371] [Medline: 11936935]
31. Achenbach TM, Rescorla LA. Manual for the ASEBA Preschool Forms and Profiles: ASEBA Europe; 2000. URL: https://aseba-europe.org/fi-en/product/110/manual-for-the-aseba-preschool-forms-profiles [accessed 2026-03-26]
32. Essau CA, Sasagawa S, Frick PJ. Callous-unemotional traits in a community sample of adolescents. Assessment 2006 Dec;13(4):454-469. [doi: 10.1177/1073191106287354] [Medline: 17050915]
33. Ezpeleta L, de la Osa N, Granero R, Penelo E, Domènech JM. Inventory of callous-unemotional traits in a community sample of preschoolers. J Clin Child Adolesc Psychol 2013;42(1):91-105. [doi: 10.1080/15374416.2012.734221] [Medline: 23095075]
34. Arnold DS, O'Leary SG, Wolff LS, Acker MM. The parenting scale: a measure of dysfunctional parenting in discipline situations. Psychol Assess 1993;5(2):137-144. [doi: 10.1037/1040-3590.5.2.137]
35. Lovibond PF, Lovibond SH. The structure of negative emotional states: comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. Behav Res Ther 1995 Mar;33(3):335-343. [doi: 10.1016/0005-7967(94)00075-u] [Medline: 7726811]
36. Sinclair SJ, Siefert CJ, Slavin-Mulford JM, Stein MB, Renna M, Blais MA. Psychometric evaluation and normative data for the Depression, Anxiety, and Stress Scales-21 (DASS-21) in a nonclinical sample of U.S. adults. Eval Health Prof 2012 Sep;35(3):259-279. [doi: 10.1177/0163278711424282] [Medline: 22008979]
37. Leijten P, Gardner F, Melendez-Torres GJ, et al. Meta-analyses: key parenting program components for disruptive child behavior. J Am Acad Child Adolesc Psychiatry 2019 Feb;58(2):180-190. [doi: 10.1016/j.jaac.2018.07.900]
38. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 2003 May;41(5):582-592. [doi: 10.1097/01.MLR.0000062554.74615.4C] [Medline: 12719681]
39. Mouelhi Y, Jouve E, Castelli C, Gentile S. How is the minimal clinically important difference established in health-related quality of life instruments? Review of anchors and methods. Health Qual Life Outcomes 2020 May 12;18(1):136. [doi: 10.1186/s12955-020-01344-w] [Medline: 32398083]
40. Korda H. Bringing evidence-based interventions to the field: the fidelity challenge. J Public Health Manag Pract 2013;19(1):1-3. [doi: 10.1097/PHH.0b013e318249bc06] [Medline: 23169396]
41. Leijten P, Raaijmakers M, Wijngaards L, et al. Understanding who benefits from parenting interventions for children's conduct problems: an integrative data analysis. Prev Sci 2018 May;19(4):579-588. [doi: 10.1007/s11121-018-0864-y] [Medline: 29349546]
42. Hawes DJ, Price MJ, Dadds MR. Callous-unemotional traits and the treatment of conduct problems in childhood and adolescence: a comprehensive review. Clin Child Fam Psychol Rev 2014 Sep;17(3):248-267. [doi: 10.1007/s10567-014-0167-1] [Medline: 24748077]
43. Carr A. Conduct problems. In: Family Therapy: Concepts, Process and Practice, 3rd edition: John Wiley & Sons; 2012:323-345. [doi: 10.1002/9780470713051]
44. Sandi C. Stress and cognition. Wiley Interdiscip Rev Cogn Sci 2013 May;4(3):245-261. [doi: 10.1002/wcs.1222] [Medline: 26304203]
45. Nieto-Retuerto M, Torres-Gomez B, Alonso-Arbiol I. Parental mentalization and children's externalizing problems: a systematic review and meta-analysis. Dev Psychopathol 2025 Oct;37(4):1804-1820. [doi: 10.1017/S0954579424001391] [Medline: 39397703]
46. Chevalier V, Simard V, Achim J. Meta-analyses of the associations of mentalization and proxy variables with anxiety and internalizing problems. J Anxiety Disord 2023 Apr;95:102694. [doi: 10.1016/j.janxdis.2023.102694] [Medline: 36905852]

XSL•FO
RenderX

47. Casline E, Patel ZS, Timpano KR, Jensen-Doss A. Exploring the link between transdiagnostic cognitive risk factors, anxiogenic parenting behaviors, and child anxiety. Child Psychiatry Hum Dev 2021 Dec;52(6):1032-1043. [doi: 10.1007/s10578-020-01078-2]
48. Rutherford HJV, Wallace NS, Laurent HK, Mayes LC. Emotion regulation in parenthood. Dev Rev 2015 Jun 1;36:1-14. [doi: 10.1016/j.dr.2014.12.008] [Medline: 26085709]
49. Aldao A, Nolen-Hoeksema S, Schweizer S. Emotion-regulation strategies across psychopathology: a meta-analytic review. Clin Psychol Rev 2010 Mar;30(2):217-237. [doi: 10.1016/j.cpr.2009.11.004] [Medline: 20015584]
50. Maliken AC, Katz LF. Exploring the impact of parental psychopathology and emotion regulation on evidence-based parenting interventions: a transdiagnostic approach to improving treatment effectiveness. Clin Child Fam Psychol Rev 2013 Jun;16(2):173-186. [doi: 10.1007/s10567-013-0132-4] [Medline: 23595362]
51. Gardner F, Leijten P, Mann J, et al. Public Health Res 2017 Dec;5(10). [doi: 10.3310/phr05100]
52. Gardner F, Leijten P, Melendez-Torres GJ, et al. The earlier the better? Individual participant data and traditional meta-analysis of age effects of parenting interventions. Child Dev 2019 Jan;90(1):7-19. [doi: 10.1111/cdev.13138] [Medline: 30216433]
53. Gardner F, Leijten P, Harris V, et al. Equity effects of parenting interventions for child conduct problems: a pan-European individual participant data meta-analysis. Lancet Psychiatry 2019 Jun;6(6):518-527. [doi: 10.1016/S2215-0366(19)30162-2] [Medline: 31072801]
54. Leijten P, Scott S, Landau S, et al. Individual participant data meta-analysis: impact of conduct problem severity, comorbid attention-deficit/hyperactivity disorder and emotional problems, and maternal depression on parenting program effects. J Am Acad Child Adolesc Psychiatry 2020 Aug;59(8):933-943. [doi: 10.1016/j.jaac.2020.01.023] [Medline: 32084529]
55. Melendez-Torres GJ, Leijten P, Scott S, et al. Disruptive child behavior severity and parenting program session attendance: Individual participant data meta-analysis. J Consult Clin Psychol 2024 Oct;92(10):692-697. [doi: 10.1037/ccp0000893] [Medline: 38780573]
56. Government resolution STM/2021/170: government resolution on an action plan implementing the hybrid strategy to manage the COVID-19 epidemic. Finnish Government. 2021 Sep 9. URL: https://valtioneuvosto.fi/paatokset/paatos?decisionId=0900908f80754d69 [accessed 2026-03-26]
57. Chung G, Tilley JL, Netto N, Chan A, Lanier P. Parenting stress and its impact on parental and child functioning during the COVID-19 pandemic: a meta-analytical review. Int J Stress Manag 2024;31(3):238-251. [doi: 10.1037/str0000329]
58. Connell CM, Strambler MJ. Experiences with COVID-19 stressors and parents' use of neglectful, harsh, and positive parenting practices in the Northeastern United States. Child Maltreat 2021 Aug;26(3):255-266. [doi: 10.1177/10775595211006465] [Medline: 33787377]
59. Friedmann A, Buechel C, Seifert C, Eber S, Mall V, Nehring I. Easing pandemic-related restrictions, easing psychosocial stress factors in families with infants and toddlers? Cross-sectional results of the three wave CoronabaBY study from Germany. Child Adolesc Psychiatry Ment Health 2023 Jun 23;17(1):76. [doi: 10.1186/s13034-023-00618-7] [Medline: 37353851]
60. Jacobson NS, Roberts LJ, Berns SB, McGlinchey JB. Methods for defining and determining the clinical significance of treatment effects: description, application, and alternatives. J Consult Clin Psychol 1999;67(3):300-307. [doi: 10.1037/0022-006X.67.3.300]
61. Callender KA, Olson SL, Choe DE, Sameroff AJ. The effects of parental depressive symptoms, appraisals, and physical punishment on later child externalizing behavior. J Abnorm Child Psychol 2012 Apr;40(3):471-483. [doi: 10.1007/s10802-011-9572-9] [Medline: 21947616]
62. Curnow E, Tilling K, Heron JE, Cornish RP, Carpenter JR. Multiple imputation of missing data under missing at random: including a collider as an auxiliary variable in the imputation model can induce bias. Front Epidemiol 2023 Sep 15;3:1237447. [doi: 10.3389/fepid.2023.1237447] [Medline: 37974561]
63. Sourander A, McGrath PJ, Ristkari T, et al. Two-year follow-up of internet and telephone assisted parent training for disruptive behavior at age 4. J Am Acad Child Adolesc Psychiatry 2018 Sep;57(9):658-668. [doi: 10.1016/j.jaac.2018.07.001] [Medline: 30196869]
64. López-Romero L, Romero E, Andershed H. Conduct problems in childhood and adolescence: developmental trajectories, predictors and outcomes in a six-year follow up. Child Psychiatry Hum Dev 2015 Oct;46(5):762-773. [doi: 10.1007/s10578-014-0518-7] [Medline: 25354563]
65. Lintula S, Sourander A, Hinkka-Yli-Salomäki S, et al. Enrollment and completion rates of a nationwide guided digital parenting program for children with disruptive behavior before and during COVID-19. Eur Child Adolesc Psychiatry 2025 Feb;34(2):739-749. [doi: 10.1007/s00787-024-02523-6] [Medline: 39141106]
66. Hofmann SG. A network control theory of dynamic systems approach to personalize therapy. Behav Ther 2025 Jan;56(1):199-212. [doi: 10.1016/j.beth.2024.10.006] [Medline: 39814513]

Abbreviations

CBCL: Child Behavior Checklist 1.5-5

XSL•FO
RenderX

DASS-21: 21-Item Depression, Anxiety and Stress Scale
FSFP: Finnish Strongest Families Parent
ICU: Inventory of Callous-Unemotional Traits
OR: odds ratio
PS: Parenting Scale
RCT: randomized controlled trial
SDQ: Strengths and Difficulties Questionnaire

Edited by S Brini; submitted 25.Jun.2025; peer-reviewed by K Kostyrka-Allchorne, S Scott; revised version received 10.Mar.2026; accepted 10.Mar.2026; published 28.Apr.2026.

Please cite as:
Westerlund M, Hinkka-Yli-Salomäki S, Ristkari T, Kinnunen M, Baumel A, Sourander A
Predictors of Outcomes of Parent Training Targeting Disruptive Behavior in Children Aged 4 Years at 6-Month Follow-Up: Results From a Large Prospective Cohort Implementation Study
J Med Internet Res 2026;28:e79592
URL: https://www.jmir.org/2026/1/e79592
doi:10.2196/79592

© Minja Westerlund, Susanna Hinkka-Yli-Salomäki, Terja Ristkari, Malin Kinnunen, Amit Baumel, Andre Sourander. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Leveraging Text Messaging and Behavior Theory to Improve Colorectal Cancer Screening in Federally Qualified Health Centers: Cohort Study

Tracy Angelocci[1], MD; Tonghui Xu[2], PhD; Tushar Talaparthy[3], BS; Cecilia Corral[4], BS; Omolola E Adepoju[2,3], MPH, PhD

[1]Lone Star Circle of Care, Georgetown, TX, United States

[2]Humana Integrated Health Systems Science Institute, 5055 Medical Cir, Houston, TX, United States

[3]Tilman J. Fertitta Family College of Medicine, University of Houston, 5055 Medical Circle, Houston, TX, United States

[4]CareMessage, San Francisco, CA, United States

Corresponding Author:
Omolola E Adepoju, MPH, PhD
Humana Integrated Health Systems Science Institute, 5055 Medical Cir, Houston, TX, United States

Abstract

Background: Colorectal cancer (CRC) is the second leading cause of cancer-related deaths in the United States. Despite recommendations for screening to begin at the age of 45 years, significant disparities persist, particularly among medically underserved populations.

Objective: This study examines the effectiveness of SMS text messaging reminders in improving CRC screening rates across 2 large federally qualified health centers (FQHCs) serving vulnerable populations.

Methods: The study included 4822 adults aged ≥45 years, receiving care at 2 large FQHC networks in Texas and California. Participants were assigned to one of four groups: (1) control (no SMS text messages), (2) single-outreach SMS text overdue message, (3) three-week SMS overdue and reminder text messages, and (4) six-week SMS text messages that were informed by behavior theory. Data were collected from May 2023 to July 2024. The outcome measure was a binary indicator of whether the participant underwent 1 of 3 CRC tests, fecal immunochemical test, colonoscopy, and Cologuard, within 90 days of completing the SMS text messaging reminders. Independent variables included demographic, geographic, clinical, and primary care access variables. Multivariate logistic regression models were used to examine associations between CRC screening completion and the SMS text messaging reminder groups, adjusting for covariates. Adjusted odds ratios (aORs) and 95% CIs were reported.

Results: In the combined 3-test model, patients in the single-outreach SMS text message (aOR 1.22, 95% CI 1.00-1.47) and the 3-week SMS text message (aOR 1.27, 95% CI 1.05-1.53) groups had higher odds of completing the screening test compared to those in the control group. Within the fecal immunochemical test–only model, patients in the 3-week SMS text message group (aOR 1.25, 95% CI 1.00-1.56) were more likely to complete the screening test. Within the Cologuard-only model, patients in the 3-week SMS text message group (aOR 7.01, 95% CI 1.96-25.07) and the 6-week SMS text message group (aOR 5.75, 95% CI 1.53-21.61) had significantly higher odds of CRC screening completion.

Conclusions: The findings highlight that moderate-frequency SMS text messaging reminders can effectively increase CRC screening rates in FQHCs; however, critical factors include the timing and frequency of these reminders. The 3-week intervention was associated with improved screening uptake, whereas the 6-week theory-informed intervention did not demonstrate a significant advantage over the control group, potentially reflecting a ceiling effect or message fatigue associated with more frequent messaging. Additionally, the study highlights unique screening patterns that contradict previous literature, underscoring the importance of a tailored approach for vulnerable communities.

(J Med Internet Res 2026;28:e86408) doi:10.2196/86408

KEYWORDS

colorectal cancer; cancer screening; primary care; social determinants of health; federally qualified health center; FQHC

Introduction

Colorectal cancer (CRC) is a growing public health concern in the United States, with an estimated 152,810 new cases and 53,010 deaths in 2024 [1]. As a leading cause of cancer-related deaths in the United States [1], professional organizations, including the American Cancer Society, have emphasized the importance of increasing CRC screening awareness and uptake. While the US Preventive Services Task Force typically

XSL•FO
RenderX

recommends screening for CRC in all adults aged 50 to 75 years, recent US Preventive Services Task Force recommendations now suggest earlier screening starting at the age of 45 years [2] due to the rising incidence of CRC in individuals aged under 50 years [3].

Recent studies suggest that 71.8% of adults aged 50 to 75 years were up to date with CRC screening, reflecting a modest increase from previous years [4]. Despite this progress, screening disparities persist in underserved communities [5], especially among patients served by federally qualified health centers (FQHCs). FQHCs serve as a vital safety net in the US health care system, providing care to over 50 million individuals annually [6,7]. These centers predominantly serve populations that are economically and socially vulnerable. Recent estimates suggest that over 90% of patients seen at FQHCs reported household incomes below 200% of the federal poverty threshold, with nearly half enrolled in Medicaid and one-quarter uninsured. While FQHCs provide CRC screenings, utilization rates have been historically low, with work by Amboree et al [8] reporting an average screening rate of 40.2% among eligible individuals in FQHCs compared to 72.3% among the general US population. As such, CRC continues to disproportionately affect underserved populations in the United States. Among both men and women, American Indian or Alaska Native and non-Hispanic Black populations experience the highest rates of CRC incidence and mortality [9]. Notably, Black individuals experience 32% more CRC-related deaths compared to White individuals [10].

In efforts to increase screening rates, SMS text messaging reminders have become common. SMS text messaging reminders have proven effective across various health care settings, with a review by Schwebel and Larimer [11] finding that the majority of SMS text messaging reminder studies have shown improved patient appointment and medical compliance. The literature on SMS reminders for CRC screening in general settings currently suggests a limited impact. For example, Uy et al [12] in their systematic review, reported modest increases in breast and cervical cancer screening completion following SMS text messaging interventions, but only small effects for CRC screening. Considering SMS text messaging reminders have been shown to significantly boost screening rates for other cancers [12,13], the research team hypothesized that with theory of behavior change (TOBC)–informed SMS text messages that incorporate behavioral nudges, personalized messaging, and actionable prompts to reduce structural and motivational barriers may be better suited to improve CRC screening uptake in FQHC settings.

Recent evidence suggests a role for infusing established TOBC in SMS text messages [14-16]. Specifically, constructs from the health belief model [17] and the theory of planned behavior [18] emphasize key determinants of behavior that can potentially increase patient motivation, enhance confidence in completing recommended screenings, and provide cues to take actionable steps toward completing their cancer screening. This study examines the effectiveness of a TOBC-informed SMS intervention on CRC screening completion among overdue patients in 2 large FQHC networks in Texas and California.

Methods

Sample

The study sample built on prior quality improvement work [19], which enrolled 5337 patients who were overdue for CRC screening from 2 FQHCs: site 1 (56 clinics, Texas) and site 2 (2 full-time and 5 satellite clinics, California). Data were collected from May 2023 to July 2024. A total of 5337 patients were assigned to 1 of 4 study groups over a 90-day period. These included (1) control (no SMS text messages), (2) a single overdue reminder message, (3) a 3-week reminder schedule (1 message per week; total of 3 messages), and (4) a 6-week theory-informed messaging schedule (3 messages per week; total of 18 messages). The unit of randomization (ie, the level at which the intervention was assigned) was the patient, and participants were assigned in a 1:1:1:1 ratio to 1 of the 4 groups using a computer-generated randomization list. After excluding 515 (10%) participants who had no phone number, had an invalid mobile phone number, or had previously opted out of SMS text messaging communication, the final analytic sample consisted of 4822 participants: 1143 in the control group, 1193 in the single message group, 1265 in the 3-week message group, and 1221 in the 6-week SMS text message group (Figure 1).

XSL•FO
RenderX

Figure 1. Flowchart of the colorectal cancer test data collection process: (1) group 1: control group, (2) group 2: single message group, (3) group 3: three-week message group, and (4) group 4: six-week message group. FQHC: federally qualified health centers.

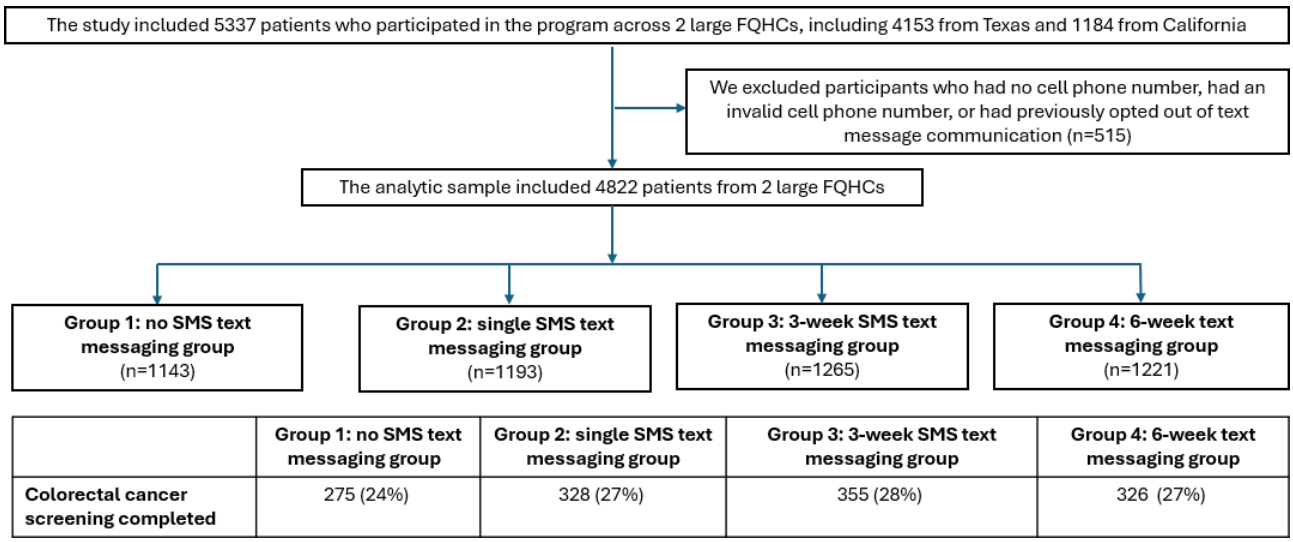

	Group 1: no SMS text messaging group	Group 2: single SMS text messaging group	Group 3: 3-week SMS text messaging group	Group 4: 6-week text messaging group
Colorectal cancer screening completed	275 (24%)	328 (27%)	355 (28%)	326 (27%)

Ethical Considerations

This study was reviewed and approved by the institutional review board of the University of Houston (STUDY00004081). All procedures were conducted in accordance with relevant guidelines and regulations. In this study, SMS text messaging reminders were implemented as part of an ongoing quality improvement initiative. As such, patient consent beyond the usual SMS text messaging opt-out procedures was not required. Patient privacy was protected through the use of deidentified data for analysis, and all data linkage between the electronic health record system and the SMS text messaging vendor platform was conducted in accordance with institutional data governance and HIPAA (Health Insurance Portability and Accountability Act)-compliant procedures.

Quality Improvement Intervention

In 2023, the FQHCs implemented an SMS text messaging reminder initiative to increase cancer screening uptake among individuals overdue for their screenings. After identifying eligible patients, the FQHC worked with academic partners to randomly assign eligible patients into 1 of 4 study groups described above. In partnership with an SMS text messaging reminder service, the FQHC sent messages to patients in the 3 interventions. Patients in group 4 were the only participants who received TOBC-informed SMS text messages. These were designed to be bidirectional SMS text messages, allowing the recipient to respond either to opt out of the message program, answer a question, or respond to a survey. Messages were delivered in both English and Spanish, and the language was matched to each patient's documented language preference. Examples of TOBC messages include "The overall rates of colon cancer have gone down as more people are getting regular screening. But colon cancer in people younger than age 50 is going up. Because of this, it is now recommended that everyone have their first screen at age 45. Watch this: https://msg.care/8Snk4. Call to get tested: ###-###-####"; "Prioritize your health and get cancer screenings done to make sure you will be there to take care of the ones you love."; "Listen as Terrance Howard talks about losing his mom to colon cancer: https://msg.care/8Snkw."

Measures

The outcome of interest was a binary measure indicating whether participants completed a CRC screening test (fecal immunochemical test [FIT], colonoscopy, or Cologuard) within 90 days of the intervention (0=no, 1=yes). Overall, 87.27% (n=4208) were in the FIT group, while 5.89% (n=284) were in the colonoscopy group and 6.84% (n=330) were in the Cologuard group. The independent variables included demographic characteristics including age, race or ethnicity, primary language, and insurance status. Other independent variables included health conditions—heart disease, diabetes mellitus, chronic lower respiratory diseases, overweight or obesity, hypertension, depression, and mood disorders.

On the basis of the patient's zip code, we included the following city- and county-level independent variables: (1) 2023 city population, (2) county-level population to primary care physician (PCP) ratio, and (3) city-level CRC death rate. The city population was based on the population estimate as of April 1, 2020, from the US census. The county-level population to PCP ratio was sourced from the County Health Rankings & Roadmaps website [20]. The city-level CRC death rate was sourced from the City Health Dashboard website [21].

Statistical Analysis

Statistical descriptions were conducted to describe patients and their demographic and health status. Four multiple logistic regression models were used to determine the relationship between independent variables and whether the participant underwent a CRC screening test within 90 days of the intervention. The primary regression model included all participants across all 3 CRC screening test interventions. Subsequently, 3 additional regression models were estimated separately for participants in each screening subgroup: FIT, Cologuard, and colonoscopy. Variance inflation factor (VIF) was used to rule out multicollinearity. All analyses were conducted using R (version 4.4.2; R Foundation for Statistical Computing). Statistical significance was defined as $P<.05$, and

XSL•FO
RenderX

results are reported as adjusted odds ratios (aORs), SEs, and 95% CIs.

Results

Table 1 summarizes the descriptive statistics for CRC screening completion, and Figure 1 illustrates the distribution of participant response messages. In the control group, 24.06% (n=275) of participants completed the screening test, compared to 27.49% (n=328) in the single message group. Completion rates were 28.06% (n=355) in cohort 3 (3-week message group) and 26.7% (n=326) in cohort 4 (6-week message group).

XSL•FO
RenderX

Table . Patient characteristics and results of the primary regression model (N=4822).

Variable category	Frequency, n (%)	aOR[a] (SE)	*P* value
Study group			
Control group	1143 (23.70)	R[b]ef	—[c]
Single message	1193 (24.74)	1.22 (0.10)	.04
3-week messages	1265 (26.23)	1.27 (0.10)	.02
6-week messages	1221 (25.32)	1.15 (0.10)	.14
Age group (years)			
45-59	2554 (52.97)	Ref	—
60-64	1078 (22.35)	0.53 (0.09)	<.001
65-69	781 (16.20)	0.52 (0.11)	<.001
>70	409 (8.48)	0.47 (0.15)	<.001
Sex			
Male	1998 (41.44)	Ref	—
Female	2824 (58.56)	1.10 (0.07)	.18
Race or ethnicity			
Asian	197 (4.09)	0.89 (0.20)	.54
Black or African American	622 (12.90)	1.02 (0.13)	.88
Non-Hispanic White	883 (18.31)	Ref	—
Other or unknown	397 (8.23)	1.05 (0.16)	.76
Hispanic	2813 (58.34)	0.98 (0.07)	.84
Primary language			
English	2520 (52.26)	Ref	—
Spanish	2016 (41.81)	1.20 (0.09)	.04
Other	286 (5.93)	1.33 (0.15)	.06
Insurance type			
Uninsured	2178 (45.17)	1.04 (0.12)	.75
Medicaid	1046 (21.69)	Ref	—
Medicare	692 (14.35)	1.24 (0.14)	.12
Private	906 (18.79)	1.10 (0.13)	.46
Diabetes			
No	3332 (69.10)	Ref	—
Yes	1490 (30.90)	0.83 (0.08)	.02
Depression			
No	3580 (74.24)	Ref	—
Yes	1242 (25.76)	0.86 (0.08)	.07
Hypertension			
No	1979 (41.04)	Ref	—
Yes	2843 (58.96)	1.06 (0.13)	.13
Obesity or overweight			
No	934 (19.37)	Ref	—
Yes	3888 (80.63)	1.39 (0.10)	<.001
Heart disease			

XSL•FO
RenderX

Variable category	Frequency, n (%)	aOR[a] (SE)	*P* value
No	2355 (48.84)	Ref	—
Yes	2467 (51.16)	0.65 (0.14)	.003
US state			
California	1136 (23.56)	1.32 (0.26)	.29
Texas	3686 (76.44)	Ref	—
Chronic lower respiratory diseases 2022			
No	4478 (92.87)	Ref	—
Yes	344 (7.13)	0.93 (0.15)	.63
2021 city-level colorectal cancer death rate[d]			
Less than the average	4122 (85.48)	Ref	—
Higher than the average	700 (14.52)	1.11 (0.20)	.53
2023 city size[e]			
Small town or city (<150,000)	577 (11.97)	Ref	—
Large city (>150,000)	4245 (88.03)	1.19 (0.27)	.66
2021 county-level population to PCP[f] ratio[g]			
Less than the average	1637 (33.95)	Ref	—
Higher than the average	3185 (66.05)	1.17 (0.25)	.63

[a]aOR: adjusted odds ratio.

[b]REF: reference group.

[c]Not applicable.

[d]The 2021 city-level colorectal cancer death rate is based on the City Health Dashboard website. The average city-level colorectal cancer death ratio is calculated by summing the colorectal cancer death rates of all sample cities and dividing by the number of cities in the sample.

[e]The city population is based on the population estimate as of April 1, 2020 (V2023), from the US census.

[f]PCP: primary care physician.

[g]The 2021 county-level population to primary care physician (PCP) ratio is sourced from the County Health Ranking & Roadmaps website. The average county-level population to PCP ratio is calculated by summing the population to PCP rates of all sample cities and dividing by the number of cities in the sample.

Table 1 shows the characteristics and combined 3-test regression results. Overall, 23.7% (n=1143) were in the control group, 24.74% (n=1193) in the single message group, 26.23% (n=1265) in the 3-message group, and 25.32% (n=1221) in the 6-week message group. The age distribution showed that the majority of patients were aged 45 to 59 years (n=2554, 52.97%), followed by those aged 60 to 64 years (n=1078, 22.35%), 65 to 69 years (n=781, 16.2%), and ≥70 years (n=409, 8.48%). Most patients were female (n=2824, 58.56%). The racial composition was as follows: 58.34% (n=2813) Hispanic, 18.31% (n=883) White, 8.23% (n=397) in other or unknown racial categories, 12.90% (n=622) Black or African American, and 4.09% (n=197) Asian. Most patients reported English as their primary language (n=2520, 52.26%), followed by Spanish (n=2016, 42.81%), and other languages (n=286, 5.93%). Most patients were from Texas (n=3686, 76.44%), and the remaining patients were from California (n=1136, 23.56%). Regarding insurance, 45.17% (n=2178) were uninsured, 18.79% (n=906) had private insurance, 14.35% (n=692) had Medicaid, and 21.69% (n=1046) had Medicare. Additionally, 80.63% (n=3888) were obese or overweight, 30.9% (n=1490) had diabetes, 7.13% (n=344) had chronic lower respiratory disease, 51.16% (n=2467) had heart disease, 58.96% (n=2843) had hypertension, and 25.76% (n=1242) had a history of depression or mood disorders. At the community level, 85.48% (n=4122) of patients lived in cities where the CRC death rate was below the sample average, and most patients (n=4245, 88.03%) were from large cities with populations of ≥150,000. Additionally, 66.05% (n=3185) lived in cities where the county-level PCP ratio was at or above the sample average.

The VIF test showed that the combined 3-test model, the FIT model, and the Cologuard model did not exhibit evidence of multicollinearity. Table 1 shows that patients in the single message group (aOR 1.22, 95% CI 1.00-1.47) and 3-week message group (aOR 1.27, 95% CI 1.05 - 1.53) were more likely to complete the screening test. Compared to patients aged 45 to 59 years, those aged 60 to 65 years (aOR 0.53, 95% CI 0.45 - 0.64), 65 to 69 years (aOR 0.52, 95% CI 0.42 - 0.55), and >70 years (aOR 0.47, 95% CI 0.35 - 0.64) were less likely to complete the test. Regarding language, Spanish-speaking patients (aOR 1.20, 95% CI 1.00 - 1.43) were more likely to complete the test. Regarding health conditions, patients with diabetes (aOR 0.83, 95% CI 0.70 - 0.97) were more likely to complete the test, whereas patients with obesity or overweight (aOR 1.39, 95% CI 1.15 - 1.68) were more likely to complete

XSL•FO
RenderX

the test. In contrast, those with heart disease (aOR 0.65, 95% CI 0.49 - 0.86) were less likely to complete it.

Tables 2 and 3 present the results for the FIT, colonoscopy, and Cologuard tests. For FIT (Table 2), the regression model shows that patients in the 3-message group (aOR 1.25, 95% CI 1.00 - 1.56) were more likely to complete the screening test. Compared to patients aged 45 to 59 years, those aged 60 to 65 years (aOR 0.64, 95% CI 0.53 - 0.79), 65 to 69 years (aOR 0.59, 95% CI 0.45 - 0.77), and >70 years (aOR 0.48, 95% CI 0.33 - 0.69) were less likely to complete the test. In terms of language, Spanish-speaking patients (aOR 1.54, 95% CI 1.24 - 1.91) and speakers of other languages (aOR 1.55, 95% CI 1.12 - 2.15) were more likely to complete the test. Compared to patients with Medicaid, those without any insurance (aOR 0.58, 95% CI 0.44 - 0.77), with Medicare (aOR 0.56, 95% CI 0.40 - 0.79), and with private insurance (aOR 0.30, 95% CI 0.22 - 0.40) were more likely to complete it. Patients with hypertension (aOR 0.72, 95% CI 0.60 - 0.87) and those with obesity or overweight (aOR 0.73, 95% CI 0.60 - 0.87) were less likely to complete the test. Patients living in cities with higher CRC death rates (aOR 0.53, 95% CI 0.31 - 0.90) were also less likely to complete it. Finally, patients residing in California (aOR 2.51, 95% CI 1.44 - 4.40) were more likely to complete it compared to those living in Texas.

XSL•FO
RenderX

Table . Colorectal cancer results of the fecal immunochemical test–only and colonoscopy–only models.

Variable category	FIT[b] (n=4208)				Colonoscopy (n=284)			
	Frequency, n (%)	aOR[c] (SE)	*P* value	95% CI	Frequency, n (%)	aOR (SE)	*P* value	95% CI
Study group								
Control group	1012 (24.05)	R[d]ef	—[e]	—	61 (21.48)	Ref	—	—
Single message	1039 (24.69)	1.14 (0.11)	.26	0.91-1.42	69 (24.30)	1.57 (0.43)	.3	0.67-3.64
3-week messages	1094 (26.00)	1.25 (0.11)	.04	1.00-1.56	79 (27.82)	1.23 (0.42)	.61	0.54-2.80
6-week messages	1063 (25.26)	1.08 (0.11)	.52	0.86-1.34	75 (26.41)	0.95 (0.41)	.91	0.42-2.13
Age group (years)								
45-59	2215 (52.64)	Ref	—	—	121 (42.61)	Ref	—	—
60-64	976 (23.19)	0.64 (0.10)	<.001	0.53-0.79	64 (22.54)	0.61 (0.38)	.19	0.29-1.29
65-69	666 (15.83)	0.59 (0.14)	<.001	0.45-0.77	67 (23.59)	0.55 (0.45)	.18	0.23-1.32
>70	351 (8.34)	0.48 (0.19)	<.001	0.33-0.69	32 (11.27)	0.69 (0.54)	.5	0.24-2.01
Gender								
Male	1178 (42.25)	Ref	—	—	102 (35.92)	Ref	—	—
Female	2430 (57.75)	1.07 (0.08)	.41	0.91-1.26	182 (64.08)	1.03 (0.30)	.91	0.57-1.87
Race or ethnicity								
Asian	172 (4.09)	0.76 (0.27)	.3	0.45-1.28	9 (3.17)	3.42 (0.83)	.14	0.67-17.17
Black or African American	530 (12.60)	1.11 (0.16)	.53	0.80-1.53	50 (17.61)	1.16 (0.46)	.76	0.47-2.86
Non-Hispanic White	717 (17.04)	Ref	—	—	71 (25.00)	Ref	—	—
Other or unknown	259 (6.15)	0.95 (0.21)	.82	0.63-1.44	21 (7.39)	1.09 (0.60)	.89	0.34-3.49
Hispanic	2530 (60.12)	1.07 (0.14)	.64	0.81-1.41	133 (46.83)	0.95 (0.42)	.9	0.42-2.16
Primary language								
English	2087 (49.60)	Ref	—	—	192 (67.61)	Ref	—	—
Spanish	1849 (43.94)	1.54 (0.11)	<.001	1.24-1.91	82 (28.87)	2.48 (0.37)	.02	1.18-5.12
Other	272 (6.46)	1.55 (0.17)	.01	1.12-2.15	10 (3.52)	2.75 (0.82)	.22	0.55-13.71
Insurance type								
Uninsured	2068 (49.14)	Ref	—	—	62 (21.83)	Ref	—	—
Medicaid	947 (22.50)	0.58 (0.14)	<.001	0.44-0.77	56 (19.72)	3.55 (0.52)	.01	1.28-9.83
Medicare	539 (12.81)	0.56 (0.17)	<.001	0.40-0.79	84 (29.58)	1.83 (0.52)	.24	0.65-5.11
Private	654 (15.54)	0.30 (0.16)	<.001	0.22-0.40	82 (28.87)	3.11 (0.50)	.02	1.16-8.32
Hypertension								
No	1187 (28.21)	Ref	—	—	85 (29.93)	Ref	—	—
Yes	3021 (71.79)	0.72 (0.09)	<.001	0.60-0.87	199 (70.07)	0.81 (0.32)	.51	0.43-1.52
Obesity or overweight								
No	341 (8.10)	Ref	—	—	21 (7.39)	Ref	—	—
Yes	3867 (91.90)	0.73 (0.15)	.03	0.54-0.97	263 (92.71)	0.65 (0.54)	.42	0.43-1.52

XSL•FO
RenderX

Variable category	FIT[b] (n=4208)				Colonoscopy (n=284)			
	Frequency, n (%)	aOR[c] (SE)	*P* value	95% CI	Frequency, n (%)	aOR (SE)	*P* value	95% CI
2021 city-level cancer death rate								
Less than average	3603 (85.62)	Ref	—	—	247 (86.97)	Ref	—	—
Higher than average	605 (14.38)	0.53 (0.28)	.02	0.31-0.90	37 (13.03)	0.60 (0.75)	.49	0.14-2.69
2023 city size								
Small town or city (<150,000)	516 (12.26)	Ref	—	—	65 (24.89)	Ref	—	—
Large city (≥150,000)	3692 (87.74)	0.77 (0.33)	.42	0.40-1.46	219 (77.11)	0.40 (1.18)	.27	0.03-3.36
US state								
California	1094 (26.00)	Ref	—	—	42 (17.36)	Ref	—	—
Texas	3114 (74.00)	2.51 (0.29)	<.001	1.44-4.40	242 (86.64)	2.67 (1.15)	.04	0.35-32.28
2021 county-level								
Population to PCP[f] ratio								
Less than the average	1535 (36.48)	Ref	—	—	65 (22.89)	Ref	—	—
Higher than the average	2673 (63.52)	0.85 (0.28)	.56	0.49-1.48	219 (77.11)	1.28 (1.17)	.23	0.13-12.90

[b]FIT: fecal immunochemical test.

[c]aOR: adjusted odds ratio.

[d]Ref: reference group.

[e]Not applicable.

[f]PCP: primary care physician.

XSL•FO
RenderX

Table . Colorectal cancer results of the Cologuard-only model (n=330).

Variable category	Frequency, n (%)	aOR[a] (SE)	*P* value	95% CI
Study group				
Control group	70 (21.21)	R[b]ef	—[c]	—
Single message	85 (25.76)	3.20 (0.60)	.05	0.98-10.37
3-week messages	92 (27.88)	7.01 (0.64)	<.001	1.96-25.07
6-week messages	93 (25.15)	5.75 (0.68)	.01	1.53-21.61
Age group (years)				
45-59	218 (66.06)	Ref	—	—
60-64	38 (11.52)	0.31 (0.82)	.27	0.50-12.38
65-69	48 (14.55)	0.30 (0.89)	.78	0.14-4.47
>70	26 (7.88)	0.40 (0.82)	.7	0.15-3.64
Gender				
Male	118 (35.76)	Ref	—	—
Female	212 (64.26)	1.17 (0.53)	.76	0.42-3.28
Race				
Asian or Black or African American or others or unknown	95 (28.79)	21.30 (1.15)	.01	2.26-201.09
Non-Hispanic White	95 (28.79)	Ref	—	—
Hispanic	140 (42.42)	1.10 (0.53)	.86	0.39-3.11
Primary language				
English	242 (73.33)	Ref	—	—
Spanish or other	88 (26.67)	2.07 (0.63)	.25	0.61-7.08
Insurance type				
Uninsured	47 (14.24)	1.32 (0.99)	.78	0.19-9.20
Medicaid	43 (13.03)	Ref	—	—
Medicare	70 (21.21)	0.86 (0.78)	.84	0.18-3.98
Private	170 (51.52)	0.62 (0.72)	.51	0.15-2.54
Diabetes				
No	223 (67.58)	Ref	—	—
Yes	107 (32.42)	1.68 (0.59)	.38	0.53-5.30
Depression				
No	241(73.03)	Ref	—	—
Yes	89 (26.97)	1.48 (0.62)	.52	0.44-4.97
Hypertension				
No	140 (42.42)	Ref	—	—
Yes	190 (57.58)	0.56 (0.58)	.31	0.61-7.08
Obesity or overweight				
No	41 (12.42)	Ref	—	—
Yes	289 (87.58)	0.95 (0.68)	.93	0.25-3.57
Chronic lower respiratory diseases				
No	304 (91.52)	Ref	—	—
Yes	26 (7.88)	3.71 (1.12)	.24	0.41-33.14

XSL•FO
RenderX

Variable category	Frequency, n (%)	aOR[a] (SE)	*P* value	95% CI
Family has any cancer				
No	302 (92.12)	Ref	—	—
Yes	28 (8.48)	1.14 (0.96)	.89	0.17-7.48
Visit count in 2022				
0	87 (26.36)	Ref	—	—
1-2	72 (21.82)	2.09 (0.77)	.34	0.46-9.46
3-4	65 (19.70)	1.46 (0.80)	.64	0.31-6.93
>5	106 (32.12)	1.07 (0.76)	.93	0.24-4.71
2022 annual wellness visit				
No	254 (76.97)	Ref	—	—
Yes	76 (23.03)	0.55 (0.61)	.33	0.17-4.97
City-level colorectal cancer death rate				
Less than average	272 (82.42)	Ref	—	—
Higher than average	58 (17.58)	1.62 (1.18)	.69	0.16-16.44
County-level population to PCP[d] ratio				
Less than the average	37 (11.21)	Ref	—	—
Higher than the average	293 (88.79)	1.25 (1.34)	.87	0.09-17.23

[a]aOR: adjusted odds ratio.

[b]REF: reference group.

[c]Not applicable.

[d]PCP: primary care physician.

For the colonoscopy test (Table 2), compared to English-speaking patients, Spanish-speaking patients (aOR 2.48, 95% CI 1.18 - 5.12) were more likely to complete the test. Compared to patients with Medicaid, uninsured patients (aOR 3.55, 95% CI 1.28 - 9.83) and those with private insurance (aOR 3.11, 95% CI 1.16 - 8.32) were more likely to complete the test. For the Cologuard model (Table 3), the city size variable was removed from the analysis because its VIF exceeded 5, indicating potential multicollinearity. Additionally, for categorical variables with very small cell counts in certain categories, we combined categories to avoid sparse data issues and improve the robustness of the estimates. Compared to the control group, the intervention 3-week message group (aOR 7.01, 95% CI 1.96 - 25.07) and 6-week message group (aOR 5.75, 95% CI 1.53 - 21.61) were more likely to complete the test. Compared to White, other races (Asian, Black or African American, others, and unknown; aOR 21.30, 95% CI 2.26 - 201.09) were more likely to complete it. In the Cologuard model, we observed wide 95% CIs for several variables, indicating imprecision in the estimates. In particular, the race or ethnicity estimate was 21.30 with a correspondingly wide CI. To investigate this, we cross-tabulated race or ethnicity by Cologuard test completion and found that among Asian patients (n=95), 94 (99%) completed the test. This near-complete separation (ie, very high completion within a subgroup) likely contributed to the inflated estimate and wide 95% CI.

Discussion

Principal Findings

In this study of TOBC-informed SMS text messaging reminders for CRC screening in an FQHC, the combined 3-test model, FIT model, and Cologuard model showed that sending 1 reminder per week for 3 weeks significantly increased screening rates. This aligns with existing work on the overall impact of SMS text messaging reminder interventions, with meta-analyses associating SMS text messaging reminders with improved general appointment and medical compliance [11]. However, results for screening settings vary, with work finding some [22-24] and no changes [5,25,26] following SMS text messaging reminders. Our findings on reminder frequency align with previous research, showing that 2 reminders are more effective than a single reminder in reducing missed appointments and same-day cancelations [27]. Additionally, work by Teo et al [28] suggests that excessive reminders can cause “reminder fatigue,” leading to redundancy and inattention, aligning with the no effect seen in the 6 weeks messaging group.

Age was found to be a significant factor, with screening rates being inversely associated with advancing age. This finding aligns with Walter et al [29], who found screening rates to decrease with increasing age in a Veterans Affairs clinic setting. While current guidelines recommend a colonoscopy every 10 years for adults aged 45 years and older [30], Patel et al [31] emphasized that screening decisions for older adults should be individualized based on prior screening history and patient

XSL•FO
RenderX

preference. However, given the heightened challenges older adults face in accessing care [32], this finding raises concern that, in practice, older adults may not receive appropriate follow-up or adhere to these tailored recommendations. Obesity or overweight was another significant predictor, as patients who identified as having obesity were associated with greater CRC screening rates compared to patients without obesity. Prior work on BMI status and screening rates has yielded mixed findings, with obesity reported as both a negative [33,34] and positive predictor for screening [35]. Notably, Kendall et al [36] found obesity status to be a positive predictor for screening among Medicare beneficiaries. Patients in FQHC settings who receive more frequent care for obesity management have increased opportunities to be assessed and recommended for screening.

Our findings also highlighted that patients with diabetes had lower CRC screening uptake. These findings are consistent with prior studies demonstrating a negative association between diabetes and CRC screening [37,38]. We believe that patients with diabetes often have multiple comorbidities, which may contribute to competing health demands and reduce the likelihood of completing preventive screening. Similarly, patients with heart disease were associated with decreased CRC screening uptake compared to patients without heart disease. This finding is consistent with Bhatia et al [39], who also found lower screening rates among patients with heart failure and cardiovascular disease, citing competing demands of more intensive chronic disease management over preventative care. However, because unfavorable cardiovascular risk factors are linked to an increased risk of CRC [40], a critical need exists to enhance access to preventive screening services for patients managing cardiac comorbidities within FQHC settings.

On the basis of the results of FIT and colonoscopy models, patients who primarily spoke Spanish or other languages were more likely to complete CRC screening, compared to patients who were English speakers. This finding is unique among current literature, as work suggests that patients with limited English-language proficiency are less likely to receive physician recommendations for a colonoscopy and have lower screening rates [41,42]. As FQHCs serve a large non-English-speaking patient population, this highlights their ability to increase awareness of and access to preventive care across the minoritized populations they serve. Additionally, FQHCs often provide culturally tailored education and language-concordant care, which can enhance trust and communication between patients and providers [43,44]. This supportive environment helps non-English-speaking patients understand the screening process and feel comfortable completing tests, thereby boosting screening rates despite language barriers.

Uninsured patients were associated with greater CRC screening rates compared to patients with Medicaid, while having Medicare and private insurance did not lead to a significant difference. These findings are unique among most work covering general clinic settings, which argues that uninsured patients [45,46] and those without nonprivate insurance are associated with lower adherence to CRC screening guidelines [47]. Given the sliding scale this setting uses [48], our observed screening uptake potentially speaks to the FQHC's ability to provide free or low-cost preventive service use among those who are uninsured. Notably, the clinics in this study provide a free FIT kit, mailing it with prepaid return envelopes to eliminate the need for clinic visits.

State of residence was a major factor influencing screening rates, with patients in California showing higher screening rates compared to those in Texas. This disparity is particularly pronounced in safety-net settings, as Texas has one of the largest uninsured populations in the nation [49]. Furthermore, CRC screening coverage was only recently included under Medicare in 2021 [50], further complicating access for underserved patients. CRC screening rates at Texas FQHCs remain below the national average, with only 35% of eligible patients up to date on screenings in 2020 [51]. A key contributor to the difference in screening rates between California and Texas is the variation in Medicaid expansion policies [52]. California's Medicaid expansion has resulted in the majority of adult FQHC patients having some form of insurance coverage, improving access to preventive services [53,54]. Texas, a nonexpansion state, leaves many FQHC patients uninsured. Despite free FIT kits, lack of consistent coverage likely lowers screening engagement, highlighting the impact of state policy on underserved populations.

There are many limitations to be considered in interpreting these findings. First, screening tests completed outside the FQHC systems may not have been captured, potentially leading to underestimation of true screening uptake. Second, the study included only 2 FQHC organizations, which may limit generalizability to non-FQHC settings. Third, the follow-up period was limited to 90 days, preventing assessment of longer-term adherence. Fourth, exact timing information (eg, weekends, holidays, or time of day) was not collected, limiting our ability to assess potential timing effects on participant response. Additionally, some variables, including community-level characteristics, may have been measured with limited precision or at a relatively crude level. Finally, for the colonoscopy and Cologuard models, given the relatively small sample sizes and the large ORs observed for some variables, the results should be interpreted with caution. Future studies should examine the effect of additional optimization variables (eg, message timing and length) to assess effectiveness.

Conclusions

In an FQHC study, 3 weekly TOBC-informed SMS text messaging reminders most effectively improved CRC screening. Screening rates also varied by age, language, insurance, obesity, heart disease, and state, underscoring the need to further explore TOBC-informed reminders to promote equitable screening.

Funding

The authors received no financial support for the research, authorship, and/or publication of this paper.

XSL•FO
RenderX

Data Availability

The data generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

Cecilia Corral is an employee of CareMessage. Others do not have conflicts to declare.

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. CA Cancer J Clin 2021 Jan;71(1):7-33. [doi: 10.3322/caac.21654] [Medline: 33433946]
2. US Preventive Services Task Force, Davidson KW, Barry MJ, et al. Screening for colorectal cancer: US Preventive Services Task Force recommendation statement. JAMA 2021 May 18;325(19):1965-1977. [doi: 10.1001/jama.2021.6238] [Medline: 34003218]
3. Siegel RL, Miller KD, Goding Sauer A, et al. Colorectal cancer statistics, 2020. CA Cancer J Clin 2020 May;70(3):145-164. [doi: 10.3322/caac.21601] [Medline: 32133645]
4. Ebner DW, Finney Rutten LJ, Miller-Wilson LA, et al. Trends in colorectal cancer screening from the National Health Interview Survey: analysis of the impact of different modalities on overall screening rates. Cancer Prev Res (Phila) 2024 Jun 4;17(6):275-280. [doi: 10.1158/1940-6207.CAPR-23-0443] [Medline: 38561018]
5. Coronado GD, Rivelli JS, Fuoco MJ, et al. Effect of reminding patients to complete fecal immunochemical testing: a comparative effectiveness study of automated and live approaches. J Gen Intern Med 2018 Jan;33(1):72-78. [doi: 10.1007/s11606-017-4184-x] [Medline: 29019046]
6. America's health centers: by the numbers. NACHC. URL: https://www.nachc.org/resource/americas-health-centers-by-the-numbers/ [accessed 2026-04-08]
7. Masselil M, Sin P. How many patients do community health centers really care for? You'll be surprised: National Association of Community Health Centers; 2025. URL: https://www.nachc.org/how-many-patients-do-community-health-centers-really-care-for-youll-be-surprised/
8. Amboree TL, Montealegre JR, Parker SL, et al. National breast, cervical, and colorectal cancer screening use in federally qualified health centers. JAMA Intern Med 2024 Jun 1;184(6):671-679. [doi: 10.1001/jamainternmed.2024.0693] [Medline: 38683574]
9. Siegel RL, Wagle NS, Cercek A, Smith RA, Jemal A. Colorectal cancer statistics, 2023. CA Cancer J Clin 2023;73(3):233-254. [doi: 10.3322/caac.21772] [Medline: 36856579]
10. Anyane-Yeboa A, Bermudez H, Fredericks M, et al. The revised colorectal cancer screening guideline and screening burden at community health centers. Sci Rep 2025 Jan 2;15(1):336. [doi: 10.1038/s41598-024-83343-1] [Medline: 39747176]
11. Schwebel FJ, Larimer ME. Using text message reminders in health care services: a narrative literature review. Internet Interv 2018;13:82-104. [doi: 10.1016/j.invent.2018.06.002] [Medline: 30206523]
12. Uy C, Lopez J, Trinh-Shevrin C, Kwon SC, Sherman SE, Liang PS. Text messaging interventions on cancer screening rates: a systematic review. J Med Internet Res 2017 Aug 24;19(8):e296. [doi: 10.2196/jmir.7893] [Medline: 28838885]
13. Kerrison RS, Shukla H, Cunningham D, Oyebode O, Friedman E. Text-message reminders increase uptake of routine breast screening appointments: a randomised controlled trial in a hard-to-reach population. Br J Cancer 2015 Mar 17;112(6):1005-1010. [doi: 10.1038/bjc.2015.36] [Medline: 25668008]
14. Sharma V, Feldman M, Sharma R. Telehealth technologies in diabetes self-management and education. J Diabetes Sci Technol 2024 Jan;18(1):148-158. [doi: 10.1177/19322968221093078] [Medline: 35485769]
15. Fishbein M, Hennessy M, Kamb M, et al. Using intervention theory to model factors influencing behavior change. Project RESPECT. Eval Health Prof 2001 Dec;24(4):363-384. [doi: 10.1177/01632780122034966] [Medline: 11817197]
16. Berset AE, Epstein JN, Hommel KA, Brinkman WB. Examining the unified theory of behavior change constructs among adolescents taking attention-deficit/hyperactivity disorder medicine: a longitudinal study. Acad Pediatr 2023;23(4):762-772. [doi: 10.1016/j.acap.2022.09.021] [Medline: 36202295]
17. Alyafei A, Easton-Carr R. The health belief model of behavior change. In: StatPearls [Internet]: StatPearls Publishing; 2026.
18. Conner M. Theory of planned behavior. In: Tenenbaum G, Eklund RC, editors. Handbook of Sport Psychology, 4th edition: John Wiley & Sons; 2020:1-18. [doi: 10.1002/9781119568124]
19. Xu T, Chavez S, Angelocci T, Adepoju O. Does text messaging increase breast cancer screening gap closure rates?: evidence from federally qualified health centers. JMIR Mhealth Uhealth. Preprint posted online on Jul 18, 2025. [doi: 10.2196/preprints.80901]
20. Kingery KL. County health rankings & roadmaps. J Youth Dev 2018;13(3):259-263. [doi: 10.5195/jyd.2018.649]
21. Gourevitch MN, Athens JK, Levine SE, Kleiman N, Thorpe LE. City-level measures of health, health determinants, and equity to foster population health improvement: the City Health Dashboard. Am J Public Health 2019 Apr;109(4):585-592. [doi: 10.2105/AJPH.2018.304903] [Medline: 30789770]

XSL•FO
RenderX

22. Muller CJ, Robinson RF, Smith JJ, et al. Text message reminders increased colorectal cancer screening in a randomized trial with Alaska Native and American Indian people. Cancer 2017 Apr 15;123(8):1382-1389. [doi: 10.1002/cncr.30499] [Medline: 28001304]
23. Weaver KE, Ellis SD, Denizard-Thompson N, Kronner D, Miller DP. Crafting appealing text messages to encourage colorectal cancer screening test completion: a qualitative study. JMIR Mhealth Uhealth 2015 Nov 4;3(4):e100. [doi: 10.2196/mhealth.4651] [Medline: 26537553]
24. Vives N, Travier N, Farre A, et al. Effectiveness and acceptability of targeted text message reminders in colorectal cancer screening: randomized controlled trial (M-TICS study). JMIR Public Health Surveill 2024 Jul 31;10:e57959. [doi: 10.2196/57959] [Medline: 39083331]
25. Vives N, Binefa G, Travier N, et al. Text messaging versus postal reminders to improve participation in a colorectal cancer screening program: randomized controlled trial. JMIR Mhealth Uhealth 2025 Jan 1;13:e64243. [doi: 10.2196/64243] [Medline: 39742454]
26. Hamad W, Purushotham A, Hughes S, Round T. Text-message reminders to increase participation in colorectal cancer screening. Br J Gen Pract 2024;74(741):152. [doi: 10.3399/bjgp24X736761] [Medline: 38538132]
27. Ulloa-Pérez E, Blasi PR, Westbrook EO, Lozano P, Coleman KF, Coley RY. Pragmatic randomized study of targeted text message reminders to reduce missed clinic visits. Perm J 2022 Apr 5;26(1):64-72. [doi: 10.7812/TPP/21.078] [Medline: 35609163]
28. Teo AR, Metcalf EE, Strange W, et al. Enhancing usability of appointment reminders: qualitative interviews of patients receiving care in the Veterans Health Administration. J Gen Intern Med 2021 Jan;36(1):121-128. [doi: 10.1007/s11606-020-06183-5] [Medline: 32909229]
29. Walter LC, Lindquist K, Nugent S, et al. Impact of age and comorbidity on colorectal cancer screening among older veterans. Ann Intern Med 2009 Apr 7;150(7):465-473. [doi: 10.7326/0003-4819-150-7-200904070-00006] [Medline: 19349631]
30. Inadomi JM, Vijan S, Janz NK, et al. Adherence to colorectal cancer screening: a randomized clinical trial of competing strategies. Arch Intern Med 2012 Apr 9;172(7):575-582. [doi: 10.1001/archinternmed.2012.332] [Medline: 22493463]
31. Patel SG, May FP, Anderson JC, et al. Updates on age to start and stop colorectal cancer screening: recommendations from the U.S. Multi-Society Task Force on colorectal cancer. Gastroenterology 2022 Jan;162(1):285-299. [doi: 10.1053/j.gastro.2021.10.007] [Medline: 34794816]
32. Fitzpatrick AL, Powe NR, Cooper LS, Ives DG, Robbins JA. Barriers to health care access among the elderly and who perceives them. Am J Public Health 2004 Oct;94(10):1788-1794. [doi: 10.2105/ajph.94.10.1788] [Medline: 15451751]
33. Heo M, Allison DB, Fontaine KR. Overweight, obesity, and colorectal cancer screening: disparity between men and women. BMC Public Health 2004 Nov 8;4:53. [doi: 10.1186/1471-2458-4-53] [Medline: 15533259]
34. Rosen AB, Schneider EC. Colorectal cancer screening disparities related to obesity and gender. J Gen Intern Med 2004 Apr;19(4):332-338. [doi: 10.1111/j.1525-1497.2004.30339.x] [Medline: 15061742]
35. Chang VW, Asch DA, Werner RM. Quality of care among obese patients. JAMA 2010 Apr 7;303(13):1274-1281. [doi: 10.1001/jama.2010.339] [Medline: 20371786]
36. Kendall KA, Lee E, Zuckerman IH, et al. Obesity status and colorectal cancer screening in the United States. J Obes 2013;2013:920270. [doi: 10.1155/2013/920270] [Medline: 23691289]
37. von Wagner C, Cadar D, Hackett RA, et al. Type 2 diabetes and colorectal cancer screening: findings from the English Longitudinal Study of Ageing. J Med Screen 2020 Mar;27(1):25-30. [doi: 10.1177/0969141319874834] [Medline: 31547753]
38. Wilkinson JE, Culpepper L. Associations between colorectal cancer screening and glycemic control in people with diabetes, Boston, Massachusetts, 2005-2010. Prev Chronic Dis 2011 Jul;8(4):A82. [Medline: 21672406]
39. Bhatia D, Sutradhar R, Tinmouth J, Singh S, Lau C, Lipscombe LL. Influence of chronic comorbidities on periodic colorectal cancer screening participation: a population-based cohort study. Prev Med 2021 Jun;147:106530. [doi: 10.1016/j.ypmed.2021.106530] [Medline: 33771564]
40. Zhang C, Cheng Y, Luo D, et al. Association between cardiovascular risk factors and colorectal cancer: a systematic review and meta-analysis of prospective cohort studies. EClinicalMedicine 2021;34:100794. [doi: 10.1016/j.eclinm.2021.100794] [Medline: 33997727]
41. Castañeda-Avila MA, Tisminetzky M, Oyinbo AG, Lapane K. Racial and ethnic disparities in use of colorectal cancer screening among adults with chronic medical conditions: BRFSS 2012-2020. Prev Chronic Dis 2024 Feb 22;21:E12. [doi: 10.5888/pcd21.230257] [Medline: 38386629]
42. Cataneo JL, Kim TD, Park JJ, Marecik S, Kochar K. Disparities in screening for colorectal cancer based on limited language proficiency. Am Surg 2022 Nov;88(11):2737-2744. [doi: 10.1177/00031348221105596] [Medline: 35642266]
43. Glaser KM, Crabtree-Ide CR, McNulty AD, et al. Improving guideline-recommended colorectal cancer screening in a federally qualified health center (FQHC): implementing a patient navigation and practice facilitation intervention to promote health equity. Int J Environ Res Public Health 2024 Jan 24;21(2):126. [doi: 10.3390/ijerph21020126] [Medline: 38397617]
44. Lin SC, McKinley D, Sripipatana A, Makaroff L. Colorectal cancer screening at US community health centers: examination of sociodemographic disparities and association with patient-provider communication. Cancer 2017 Nov 1;123(21):4185-4192. [doi: 10.1002/cncr.30855] [Medline: 28708933]

XSL•FO
RenderX

45. Bayly JE, Schonberg MA, Castro MC, Mukamal KJ. Individual and geospatial factors associated with receipt of colorectal cancer screening: a state-wide mixed-level analysis. Fam Med Community Health 2024 Jul;12(Suppl 2):e002983. [doi: 10.1136/fmch-2024-002983]
46. Freund KM, Reisinger SA, LeClair AM, et al. Insurance stability and cancer screening behaviors. Health Equity 2019;3(1):177-182. [doi: 10.1089/heq.2018.0093] [Medline: 31289777]
47. WHO housing and health guidelines. World Health Organization. 2018. URL: https://www.who.int/publications/i/item/9789241550376 [accessed 2026-04-26]
48. Conway SJ, Murphy J, Efron JE. academic medical centers and federally qualified health centers: collaboration for the care of underserved communities. J Prim Care Community Health 2024;15:21501319241266121. [doi: 10.1177/21501319241266121] [Medline: 39051652]
49. Olmstead T, Spencer JC, Kluz N, Zhan FB, Shokar NK, Pignone M. Costs and projected effect of a federally qualified health center-based mailed colorectal cancer screening program in Texas. Prev Chronic Dis 2024 May 2;21:E30. [doi: 10.5888/pcd21.230266] [Medline: 38696253]
50. Abbott takes final step to ensure affordable colorectal cancer screenings for Texans. American Cancer Society Cancer Action Network. 2021. URL: https://www.fightcancer.org/releases/abbott-takes-final-step-ensure-affordable-colorectal-cancer-screenings-texans [accessed 2026-03-10]
51. Texas Health Center Program Uniform Data System (UDS) data. Health Resources & Services Administration. URL: https://data.hrsa.gov/topics/healthcenters/uds/overview/state/TX [accessed 2026-03-11]
52. Status of state Medicaid expansion decisions. KFF. 2026. URL: https://www.kff.org/medicaid/status-of-state-medicaid-expansion-decisions/ [accessed 2026-03-10]
53. Zerhouni YA, Trinh QD, Lipsitz S, et al. Effect of Medicaid expansion on colorectal cancer screening rates. Dis Colon Rectum 2019 Jan;62(1):97-103. [doi: 10.1097/DCR.0000000000001260] [Medline: 30407931]
54. Ojinnaka CO, Suri Y. Impact of Medicaid expansion on healthcare access among individuals living with chronic diseases. Am J Prev Med 2020 Aug;59(2):149-156. [doi: 10.1016/j.amepre.2020.03.012] [Medline: 32456998]

Abbreviations

aOR: adjusted odds ratio
CRC: colorectal cancer
FIT: fecal immunochemical test
FQHC: federally qualified health center
HIPAA: Health Insurance Portability and Accountability Act
TOBC: theories of behavior change
VIF: variance inflation factor

Edited by A Schwartz, M Balcarras; submitted 23.Oct.2025; peer-reviewed by I Tosun, S Saini, S Hoover; revised version received 08.Apr.2026; accepted 08.Apr.2026; published 13.May.2026.

Please cite as:
Angelocci T, Xu T, Talaparthy T, Corral C, Adepoju OE
Leveraging Text Messaging and Behavior Theory to Improve Colorectal Cancer Screening in Federally Qualified Health Centers: Cohort Study
J Med Internet Res 2026;28:e86408
URL: https://www.jmir.org/2026/1/e86408
doi:10.2196/86408

© Tracy Angelocci, Tonghui Xu, Tushar Talaparthy, Cecilia Corral, Omolola E Adepoju. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Digital Diabetes Self-Management Education and Support Program Integrated With Continuous Glucose Monitoring for Type 2 Diabetes: Randomized Controlled Trial

Ashley Berthoumieux[1*], RD, MS; Jeanean B Naqvi[1*], PhD; Sean Zion[1], PhD; Jenna Napoleone[1], MPH, PhD; Amanda McGuill[1], MSN, CDCES, FNP-C; Christian J Cerrada[2], PhD; Hyun Jung Lee[3], PhD; Timothy C Dunn[3], PhD; David Kerr[4], MBChB; Carolyn Bradner Jasik[1], MD; Sarah Linke[1], MPH, PhD

[1]Omada Health, 500 Sansome St, #200, San Francisco, CA, United States
[2]Evidation Health, Inc., San Mateo, CA, United States
[3]Abbott Diabetes Care, Inc., Alameda, CA, United States
[4]Sutter Health Center for Health Systems Research, Santa Barbara, CA, United States
[*]these authors contributed equally

Corresponding Author:
Jenna Napoleone, MPH, PhD
Omada Health, 500 Sansome St, #200, San Francisco, CA, United States

Abstract

Background: Previous research has demonstrated that the use of continuous glucose monitoring (CGM) can improve glycemic control in people with type 2 diabetes when used regularly alongside a digital diabetes self-management education and support (DSMES) program. However, to date, there is limited evidence showing the benefits of a digitally delivered DSMES program combined with real-time CGM for adults with type 2 diabetes.

Objective: The objective of this study is to evaluate the impact of a DSMES program coupled with CGM on hemoglobin A_{1c} (HbA_{1c}) and CGM-derived glycemic measures compared to usual care for adults with type 2 diabetes over 6 months.

Methods: Participants with type 2 diabetes and HbA_{1c} of 8% or higher (64 mmol/mol) who were not using mealtime bolus insulin (aged 26 - 83 y; mean HbA_{1c} 9.6%, SD 1.4% [mean 81.2 mmol/mol, SD 15.8 mmol/mol]) were randomly assigned to a digital DSMES+CGM integrated solution (n=51) or usual care (n=49) for 6 months. The primary outcome was HbA_{1c}. The secondary outcomes were CGM-derived glycemic measures, including glucose management indicator, percent time in range 70 to 180 mg/dL, percent time above range (>180 mg/dL), percent time below range (<70 mg/dL), and mean glucose. Linear mixed effects models were used for intention-to-treat analyses.

Results: HbA_{1c} was lower among the intervention group versus the usual care group at 3 months (difference=−0.7%, 95% CI −1.4% to −0.1% or difference=−8.1 mmol/mol, 95% CI −15.5 to −0.7 mmol/mol; *P*=.03) and at 6 months (difference=−0.6%, 95% CI −1.4% to 0.2% or difference=−6.9 mmol/mol, 95% CI −15.7 to 1.9 mmol/mol; *P*=.12) but only reached statistical significance at 3 months. CGM-derived glycemic measures, including glucose management indicator (difference=−0.9%, 95% CI −1.7% to −0.1%; *P*=.03), time in range (difference=14.6%, 95% CI 1.0% to 28.2%; *P*=.04), time above range (difference=−14.9%, 95% CI −29.0% to −0.9%; *P*=.04), and mean glucose (difference=−36.4 mg/dL, 95% CI −70.0 to −2.9 mg/dL; *P*=.03), also significantly improved for the intervention group versus the usual care group at 6 months.

Conclusions: The combination of digital DSMES+CGM is effective for supporting adults with type 2 diabetes in managing their condition and has the potential to reduce the risk of long-term health complications.

Trial Registration: ClinicalTrials.gov NCT05368454; https://clinicaltrials.gov/ct2/show/NCT05368454

(J Med Internet Res 2026;28:e78321) doi:10.2196/78321

KEYWORDS

diabetes self-management education and support; diabetes education; digital health; continuous glucose monitoring; time in range; chronic disease management

XSL•FO
RenderX

Introduction

Type 2 diabetes is a chronic disease that affects the body's ability to regulate and use insulin effectively, leading to high blood glucose levels. Persistent hyperglycemia, as estimated by hemoglobin A_{1c} (HbA_{1c}) [1], increases the risk of severe diabetes complications, including but not limited to vision loss, renal failure, and cardiovascular disease [2]. As the incidence and prevalence of type 2 diabetes continue to rise nationwide, impacting more than 34 million adults in 2021 [3], there is a pressing need to develop and scale accessible, evidence-based diabetes management programs to alleviate the significant societal and individual burden of the chronic condition.

Diabetes self-management education and support (DSMES) is an effective approach for helping individuals practice diabetes self-care with intensive counseling and support [4,5]. DSMES programs are scalable using technology-enabled platforms [6], allowing people with chronic diseases to access DSMES programs from their personal mobile devices (eg, smartphones, tablets, laptops) and eliminate the need to travel to physical facilities for diabetes education sessions. This convenience potentially increases access and availability, thereby reaching more participants than traditional in-person programs [7].

The introduction of continuous glucose monitoring (CGM) devices has revolutionized diabetes management by providing real-time glucose readings without painful fingersticks. For those taking insulin or at high risk for complications, CGMs help monitor fluctuations in blood glucose, informing potential changes to their diet or medication regimens [8]. Studies have demonstrated that CGM can be a cost-effective option compared to conventional self-monitoring of blood glucose [9-11] and can improve glycemic control in people with type 2 diabetes when used regularly alongside DSMES programs [12].

However, previous research studies evaluating DSMES programs integrated with CGM have predominantly focused on in-person programs and short-term intermittent use of CGM for less than 8 weeks [13-16]. A few observational studies have explored the effects of a telemedicine lifestyle program for diabetes that included a CGM offering [17,18], but CGM use among participants was short term and the findings could not be extrapolated to show a cause-effect relationship between the virtually integrated approach and improved glycemic control.

Nevertheless, a meta-analysis of randomized controlled trials (RCTs) evaluating the clinical effectiveness of telemedicine for diabetes revealed that virtual health care services are generally more effective for diabetes care management than in-person care alone [19]. People participating in digital diabetes care programs have seen improved clinical outcomes (eg, HbA_{1c} reduction) and reduced medical expenditures [20-22], but clinical trials evaluating many of these digital programs have not reported CGM data related to their interventions for type 2 diabetes [23-25].

Despite growing observational research demonstrating the effectiveness of digital DSMES and CGM, to date, no RCTs have shown the benefits of a well-integrated solution that combines a digitally delivered DSMES program with longer-term, real-time CGM for adults with type 2 diabetes. Therefore, the aim of this RCT was to evaluate the impact of the integrated solution—defined as a digital DSMES program+CGM—on HbA_{1c}, CGM-derived glycemic measures, and other relevant diabetes management outcomes compared to usual care over 6 months.

Methods

Study Design

This single-blind, parallel, decentralized RCT was designed to assess the efficacy of a digital DSMES+CGM integrated solution on glycemic management among adults with type 2 diabetes not on mealtime bolus insulin therapy, in accordance with the intended labeling of the FreeStyle Libre (FSL) 14-Day CGM system (Abbott Diabetes Care, Inc). The primary objective of this study was to evaluate the relative improvement in HbA_{1c} from baseline to 3 and 6 months of the digital DSMES+CGM intervention condition compared with a usual care control condition.

Ethical Considerations

The protocol was approved by the WCG Institutional Review Board (IRB Tracking ID: 20220021) and registered on ClinicalTrials.gov (NCT05368454). All participants provided electronic informed consent. All research data were deidentified and coded with unique identifiers, and the data were transmitted using secure encrypted protocols and stored on encrypted disks on secure servers. Participants received up to a total of US $250 for completing all study activities.

Participants

To meet the inclusion criteria, participants had to be at least 18 years old, living in the United States, have English as their primary language, have a compatible smartphone for access to the digital DSMES program, and have a self-reported type 2 diabetes diagnosis for at least 6 months prior to enrollment. Additionally, they needed a laboratory-confirmed HbA_{1c} level of 8% or higher (64 mmol/mol) at baseline, a treatment regimen that included a combination of basal insulin, oral medications, and diet and exercise, and a report being compliant with their diabetes management plan. If on a medication regimen, they must have been on their current medication regimen for at least 3 months prior to enrollment and have been willing to adhere to their regimen throughout the study. During the initial recruitment period (June-October 2022), participants were also required to have a percent time in range (TIR; 70 - 180 mg/dL) of <60% prior to enrollment, which was eventually removed from the eligibility criteria to reduce participant burden and improve study activity workflow during the remainder of the recruitment period (October 2022-May 2023).

Exclusion criteria included type 1 diabetes or a history of diabetic ketoacidosis, cardiovascular issues within 3 months before enrollment (ie, transient ischemic attack or stroke, heart attack, hospitalization for congestive heart failure), recent (within 6 mo before enrollment) or planned cancer treatment, a self-reported condition leading to a life expectancy <12 months, renal impairment, alcohol or substance abuse issue or

XSL•FO
RenderX

dependency, bariatric surgery or organ transplant within 6 months before enrollment, and pregnancy or intentions of conceiving during the study period (female participants). Other exclusion criteria included mealtime bolus insulin therapy, premixed insulin formulations or a continuous subcutaneous infusion of insulin, routine regimen of glucocorticoids or psychotropic medications, skin conditions that could interfere with CGM device placement, diagnostic imaging (ie, X-ray, magnetic resonance imaging, computed tomography) scheduled while wearing a CGM for measurement purposes, and visual impairment that could interfere with viewing of CGM data or use of the digital DSMES program.

Recruitment

We recruited individuals to join this decentralized trial through Evidation (Evidation Health, Inc.), an online health platform through which users connect their digital health tools, such as fitness apps and wearable activity trackers, in exchange for monetary incentives. Targeted emails and platform posts were sent to existing Evidation members who had agreed to receive communications about studies that may be relevant to them based on the health and lifestyle information they had provided to the platform. We also recruited non-Evidation members, targeting individuals in geographic areas with the highest prevalence of type 2 diabetes through paid advertisements on Google Search and Facebook and forum posts on Reddit between February and May 2023.

Eligibility Screening

Those who passed an initial screening proceeded to sign an electronic informed consent form for the study and completed an online baseline survey, which collected their demographic information, health and diabetes history, and patient-reported outcomes. A secondary screening visit was performed to confirm that potential participants met the additional eligibility criteria of an HbA_{1c} level of 8% or higher (64 mmol/mol) and, prior to October 2022, a TIR (70 - 180 mg/dL) of <60%. Contracted vendors (PCM Trials/Act for Health, Inc, and Hawthorne Effect, Inc) deployed certified and trained clinicians to conduct home visits in cities and rural areas across the United States, eliminating the need for participants to travel to a physical site or clinic for clinical assessments. During the home visits, they collected up to 4 mL of blood using venipuncture, anthropometric measurements, resting blood pressure, and a list of participants' cardiometabolic medications. They also applied and activated 2 FSL Pro CGM sensors on the participants for measurement purposes, and participants were instructed to wear the FSL Pro CGM sensors for up to 14 days while blinded to the data. Afterwards, participants mailed the devices to the study team for data download and TIR analysis. The contracted clinicians mailed the blood specimens to Quest Diagnostics laboratory for HbA_{1c} analysis. Participants who had a laboratory-confirmed HbA_{1c} of 8% or higher (64 mmol/mol) and a TIR (70 - 180 mg/dL) of <60% if recruited prior to mid-October 2022 were randomized and considered enrolled in the study.

Randomization and Masking

Each eligible participant was randomized to 1 of the 2 conditions (digital DSMES+CGM or usual care) according to an Evidation-controlled algorithm that used randomly permuted blocks of fixed size. Randomization was stratified by gender to ensure nearly equal sample sizes for the 2 treatment arms for each gender.

The assessment staff members (ie, clinicians conducting home visits, Quest Diagnostics) and the Evidation-designated statistician that developed the statistical analysis plan were blinded to the intervention assignment during the study period. Adverse-event adjudicators, designated research personnel at Evidation and Omada, and Omada coaching staff were not blinded. Participants were not blinded to their assignment but were blinded to their HbA_{1c} levels and CGM blood glucose values collected by FSL Pro sensors.

Digital DSMES+CGM Integrated Solution

Participants randomized to the digital DSMES+CGM intervention arm completed an application for the Omada for Diabetes program, a digitally delivered DSMES program that has achieved accreditation from the National Committee for Quality Assurance's Population Health Program and Accreditation and the Association of Diabetes Care and Education Specialists. The program pairs asynchronous health coaching from a care team comprising certified Diabetes Prevention Program Lifestyle Coaches and Certified Diabetes Care and Education Specialists (CDCESs) with a virtual platform accessed through a website or mobile app available on web-enabled devices. Upon enrollment, members receive cellularly connected Food and Drug Administration–approved devices, such as body weight scales and CGMs, along with tailored diabetes self-management and lifestyle modification support under the guidance of their care team. The care team interacts with members asynchronously through messaging on the virtual platform, regularly monitoring participants' engagement and biometric data collected during the program and using it to provide tailored feedback on diabetes self-management through proactive outreach and responding to member questions. CGM data are continuously collected in real time, and when scanned, glucose data are sent to the program app, where they are available for members to see patterns, track progress, and get personalized recommendations from their assigned coach. Additional program features include weekly lessons with nutrition and lifestyle content, support around personalized goal setting, and online peer communities. The weekly curriculum lessons and all program features are accessible on both computers and mobile devices, enabling members to engage at their preferred times, frequency, and with the tools and resources they find most useful [26].

After submitting an application to join the digital DSMES program as research members, digital DSMES+CGM participants completed their account setup through a one-time consultation with a board-certified physician from CirrusMD, a contracted third-party telehealth provider that independently determined whether a CGM prescription was clinically appropriate for an individual. CirrusMD physicians evaluated members and prescribed 14 Abbott FSL 14-Day CGM sensors

XSL•FO
RenderX

as part of the integrated solution, which were sent directly to members' homes to cover 6 months of continuous wear while participating in the digital DSMES program. During the follow-up assessment periods (3-mo and 6-mo timepoints), participants in the intervention condition were instructed to wear their FSL Pro sensors and FSL 14-Day sensor at the same time. Unlike the FSL Pro sensors used for measurement of CGM-derived glycemic outcomes, members and their care team at Omada had full access to the CGM data collected through the FSL 14-Day sensors for 6 months.

Usual Care Control Condition

Participants randomized to usual care were advised during the informed consent process to continue the diabetes treatment plan as recommended by their health care provider from screening to the end of the study period. Usual care included oral medications, basal insulin, or nonpharmacologic therapies such as diet and exercise, in addition to ongoing medical monitoring by their health care provider. During the follow-up assessment periods (3-mo and 6-mo timepoints), participants in the usual care control condition only needed to wear their FSL Pro sensors and did not receive FSL 14-Day sensors.

Follow-Up Visits

Interim and poststudy follow-up visits occurred 3 and 6 months (±14 d) after randomization for all study participants, respectively. During these follow-up visits at participants' homes, contracted clinicians performed blood draws, took anthropometric measurements, assessed resting blood pressure, and reviewed cardiometabolic medication regimens. At each follow-up visit, clinicians applied and activated 2 FSL Pro sensors on participants for measurement and data quality purposes. Study participants were then instructed to wear the sensors for up to 14 consecutive days while blinded to the data and to mail them back to the study team using a prepaid shipping envelope.

Return follow-up visits occurred outside of the ±14-day range if a repeat blood draw was needed. This transpired if there was no usable data from the sample or FSL Pro sensors through no fault of the participant (eg, mail delays causing the blood sample to be outside the testing window, sensor malfunction). Participants received compensation for all completed study visits.

Measures

Study outcomes were compared between the digital DSMES program+CGM versus usual care at 3 and 6 months. The primary outcome was HbA_{1c} level. The secondary outcomes measured by the FSL Pro sensors included CGM 14-day % TIR, time above range (TAR; >180 mg/dL and >250 mg/dL), and time below range (TBR; <70 mg/dL). Furthermore, differences in mean blood glucose, SD of blood glucose, coefficient of variation, and glucose management indicator (GMI) were examined. Lastly, body weight, resting systolic and diastolic blood pressure, and diabetes distress [27] were examined. Engagement in the digital DSMES+CGM integrated solution was measured using 5 metrics: average number of meals tracked per week using the digital DSMES app, average number of physical activity bouts tracked per week (manually self-reported in the digital DSMES app or synced from a wearable device), average number of weigh-ins per week using the provided cellularly connected body weight scale, average number of messages sent to their care team per week, and median number of days in which participants had at least 1 CGM reading from the FSL 14-Day device (connected through the digital DSMES app).

Statistical Analysis

Power was calculated assuming a −0.5% (5 mmol/mol) mean difference in HbA_{1c} between the digital DSMES+CGM condition and usual care from baseline to 6 months. Using an estimated SD of 1.0, a 2-tailed type I error rate of 0.05, and 80% power, the sample size required for the primary analysis was 126 participants in total (63 per group) [28]. Assuming an attrition rate of 20%, the adjusted sample size was 160 participants (80 per group). Therefore, since we reduced our recruitment goal during the active study period due to challenges with recruitment, we were underpowered to detect a statistically significant difference with our estimated effect size of 0.5%.

All analyses were conducted with R (version 4.2.3) [29]. Demographics and baseline clinical indicators (ie, age, BMI, sex, cardiovascular medication use, and type 2 diabetes medication use) were examined as potential covariates by exploring condition differences in each covariate and examining associations between each covariate and outcome. Given that there were no condition differences on any of the potential covariates, no covariates were included in the analyses.

Linear mixed effects models were used to analyze both the primary and secondary outcomes using intention-to-treat (ITT) methodology, given that these models utilize all available data [30]. The Little missing completely at random (MCAR) test [31] was nonsignificant for all primary and secondary outcomes, supporting the MCAR assumption for this analysis and indicating that imputation was not necessary. For the primary outcome model, fixed effects included timepoint, condition, and a timepoint-by-condition interaction. A random slope for time and a random intercept for participant were included. For the secondary outcome models, fixed effects included timepoint, condition, and a timepoint-by-condition interaction. Given that several of the models would not converge when a random slope for time was included, only a random intercept for participant was included in the models [32]. All models utilized restricted maximum likelihood estimation with an unstructured covariance matrix. Estimated marginal means were obtained from the models, and planned contrasts were conducted to examine differences by condition at 3 and 6 months.

In addition to the main analysis, we conducted sensitivity analyses for all primary and secondary outcomes where we excluded participants (n=9) who were randomized to the intervention condition but never enrolled in the digital DSMES+CGM integrated solution (sample size n=42). These models utilized the same fixed and random effects for both the primary and secondary outcomes. Finally, engagement descriptives (mean and SD) were assessed among this group of intervention participants that enrolled in the digital solution (n=42).

XSL•FO
RenderX

Results

Baseline Characteristics

A total of 39,506 individuals were initially screened for eligibility (Figure 1 and Checklist 1). Of these, 594 were screened as eligible, and 292 individuals completed an informed consent form and the baseline survey. After participating in a secondary screening (n=227), 112 individuals were disqualified due to HbA_{1c} or TIR eligibility criteria, and 15 individuals were withdrawn or lost to follow-up. Thus, the final sample included 100 participants randomized to the digital DSMES+CGM intervention condition (n=51) or usual care (n=49). At 3 months, 90.2% (46/51) digital DSMES+CGM participants and 87.8% (43/49) usual care participants completed their follow-up. At 6 months, 84.3% (43/51) digital DSMES+CGM participants and 91.8% (45/49) usual care participants completed their follow-up. The level of missingness for the primary outcome (HbA_{1c}) was 24% at 3 months and 15% at 6 months, and the level of missingness for the secondary outcomes ranged from 12% to 14% at 3 months and from 13% to 23% at 6 months.

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram. [a]Participants were recruited from Evidation Health's online platform if they had previously agreed to receive communications about research studies when they signed up to become an Evidation member; the number reflected in this box is the number of participants who started filling out the screener survey to assess their potential eligibility for this decentralized RCT. [b]Home visit was not done, but the research staff was able to maintain contact with participants who missed a 3-month assessment. [c]Participants were either withdrawn due to ineligibility criteria or were marked lost to follow-up if the research staff was unable to reach participants after multiple attempts. CGM: continuous glucose monitoring; DSMES: diabetes self-management education and support; HbA_{1c}: hemoglobin A_{1c}; RCT: randomized controlled trial.

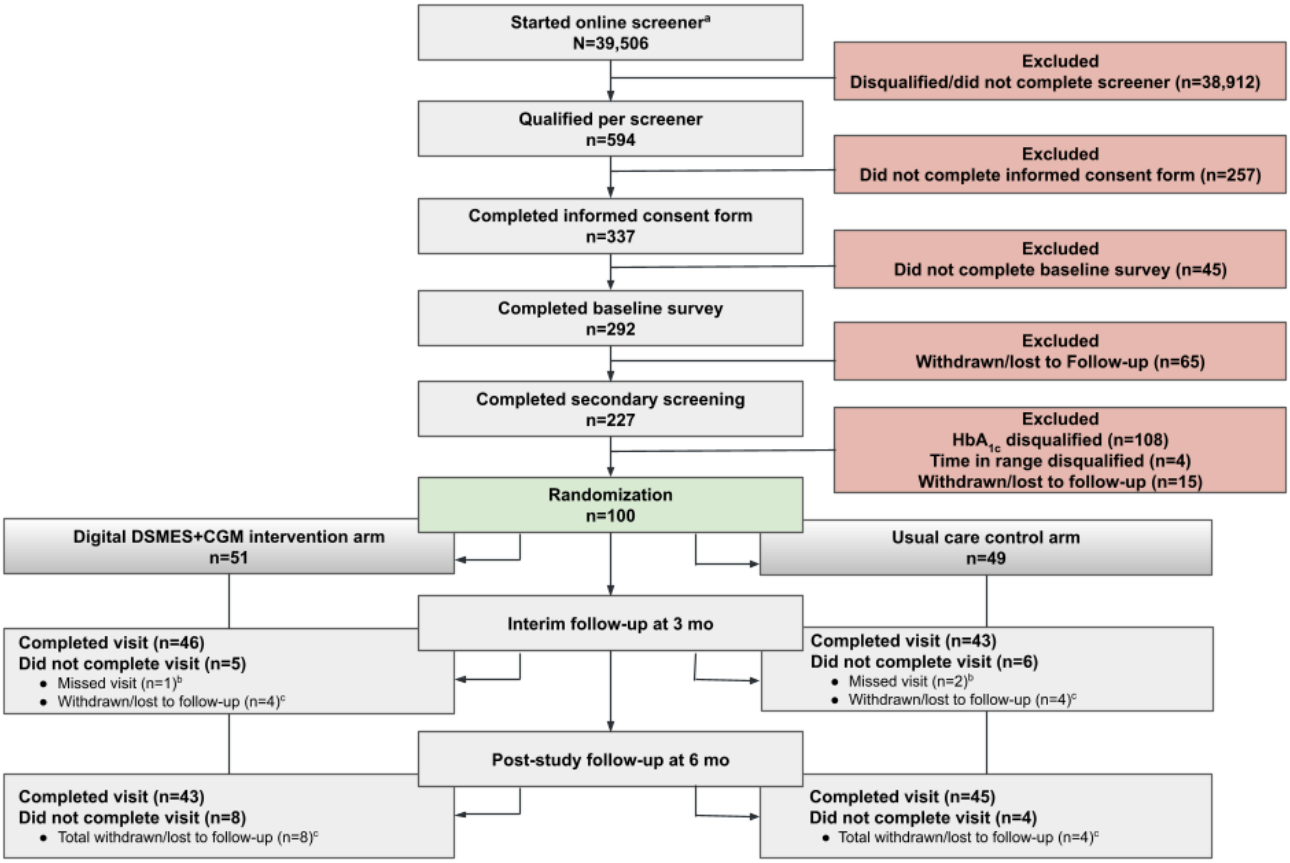

Baseline demographic and clinical characteristics of the study population are presented in Table 1. Across conditions, participants were 50.7 (SD 10.8) years of age on average, and the average weight was 215.9 (SD 45.1) pounds. Most participants were White (55/100, 55%) and female (70/100, 70%), and half had less than a 4-year college degree (50/100, 50%). The median household income was between US $50,000 and $74,999. Mean baseline HbA_{1c} was 9.6% (SD 1.4%) or 81.2 (SD 15.8) mmol/mol, and % TIR (70 - 180 mg/dL) was 33.6 (SD 28.0). Of members who reported taking diabetes medication at one or more of the 3 study visits (n=95), 86 (91%) reported a change in their medication over time. Specifically, a total of 53% (50/95) of members started a new medication and 58% (55/95) of members stopped a medication during the study period, with 35% (33/95) of members increasing the total number of medications reported and 42% (40/95) decreasing the total number of medications reported during the study period. There were no significant differences in baseline characteristics (including baseline diabetes medications) or in the number of members who started or stopped a medication, used insulin, increased or decreased the number of medications reported, or had any medication change during the study period between the 2 conditions. When looking only at members who took oral medication only at baseline (n=67), there were no differences in baseline characteristics between the 2 conditions.

XSL•FO
RenderX

Table . Baseline characteristics of study participants by condition.

Characteristic	Digital DSMES[a]+CGM[b] (n=51)	Usual care (n=49)	Overall (N=100)
Age (y), mean (SD)	51.7 (10.2)	49.5 (11.5)	50.7 (10.8)
Sex (female), n (%)	35 (68.6)	35 (71.4)	70 (70)
Race or ethnicity, n (%)			
White	29 (56.9)	26 (53.1)	55 (55)
Black/African American	11 (21.6)	9 (18.4)	20 (20)
Hispanic/Latino/Spanish	4 (7.8)	8 (16.3)	12 (12)
Another race	7 (13.7)	6 (12.2)	13 (13)
Education (<4-y college degree), n (%)	28 (54.9)	22 (44.9)	50 (50)
Household income, median	US $50,000-$74,999	US $50,000-$74,999	US $50,000-$74,999
Duration of type 2 diabetes (y), mean (SD)	10.4 (7.5)	10.5 (7.6)	10.5 (7.5)
Diabetes medication, n (%)			
GLP-1 injection	13 (25.5)	8 (16.3)	21 (21)
Oral medication only	31 (60.8)	36 (73.5)	67 (67)
Diet and exercise only	6 (11.8)	2 (4.1)	8 (8)
Basal insulin and oral medication	1 (2)	3 (6.1)	4 (4)
Weight (lb), mean (SD)	213.1 (40.5)	218.6 (49.6)	215.9 (45.1)
BMI (kg/m^2), mean (SD)	34.2 (5.8)	35.9 (7.9)	35.0 (6.9)
HbA_{1c}[c] (%), mean (SD)	9.5 (1.4)	9.7 (1.5)	9.6 (1.4)
HbA_{1c} (mmol/mol), mean (SD)	80.1 (15.5)	82.2 (16.2)	81.2 (15.8)
% TIR[d] (70 - 180 mg/dL), mean (SD)	36.0 (28.5)	31.0 (27.5)	33.6 (28.0)
% TAR[e] (>180 mg/dL), mean (SD)	63.0 (30.1)	68.3 (28.4)	65.5 (29.3)
% TBR[f] (<70 mg/dL), mean (SD)	1.0 (3.4)	0.8 (2.5)	0.9 (3.0)
Mean glucose (mg/dL), mean (SD)	219.1 (68.9)	236.1 (75.8)	227.3 (72.4)
GMI[g] (%), mean (SD)	8.6 (1.6)	9.0 (1.8)	8.8 (1.7)

[a]DSMES: diabetes self-management education and support.

[b]CGM: continuous glucose monitoring.

[c]HbA_{1c}: hemoglobin A_{1c}.

[d]TIR: time in range.

[e]TAR: time above range.

[f]TBR: time below range.

[g]GMI: glucose management indicator.

Engagement Metrics (Digital DSMES+CGM Group Only)

Of those who enrolled in the digital DSMES+CGM integrated solution (n=42), the mean number of meals tracked was 4.6 (SD 6.7) per week, and the mean number of physical activity bouts recorded was 4.3 (SD 3.1) per week within the digital DSMES app. Participants also weighed themselves an average of 3.3 (SD 2.5) times per week using the cellularly connected body weight scale as part of the digital DSMES program. Participants sent an average of 1.2 (SD 1.1) messages to their care team per week. The median number of days per week in which participants had at least 1 CGM reading from the FSL 14-Day sensor was 4.8.

Primary Outcome: HbA_{1c}

The results for HbA_{1c} are presented in Figure 2. Participants in the digital DSMES+CGM condition had significantly lower HbA_{1c} than those in usual care at 3 months (−0.7%, 95% CI −1.4% to −0.1% or −8.1 mmol/mol, 95% CI −15.5 to −0.7 mmol/mol; *P*=.03) and lower, nonsignificant reductions in

XSL•FO
RenderX

HbA_{1c} at 6 months (−0.6%, 95% CI −1.4% to 0.2% or −6.9 mmol/mol, 95% CI −15.7 to 1.9 mmol/mol; *P*=.12).

A sensitivity analysis excluding participants who never participated in the intervention showed that participants in the digital DSMES+CGM condition had significantly lower HbA_{1c} than those in usual care at both 3 months (−1%, 95% CI −1.6% to −0.3% or −10.5 mmol/mol, 95% CI −17.9 to −3.1 mmol/mol; *P*=.01) and 6 months (−0.8%, 95% CI −1.6% to −0.02% or −9.1 mmol/mol, 95% CI −18.0 to −0.2 mmol/mol; *P*=.046).

Figure 2. Change in mean hemoglobin A_{1c} over time by condition. *P* values <.05 are italicized. CGM: continuous glucose monitoring; DSMES: diabetes self-management education and support; HbA_{1c}: hemoglobin A_{1c}; M: mean.

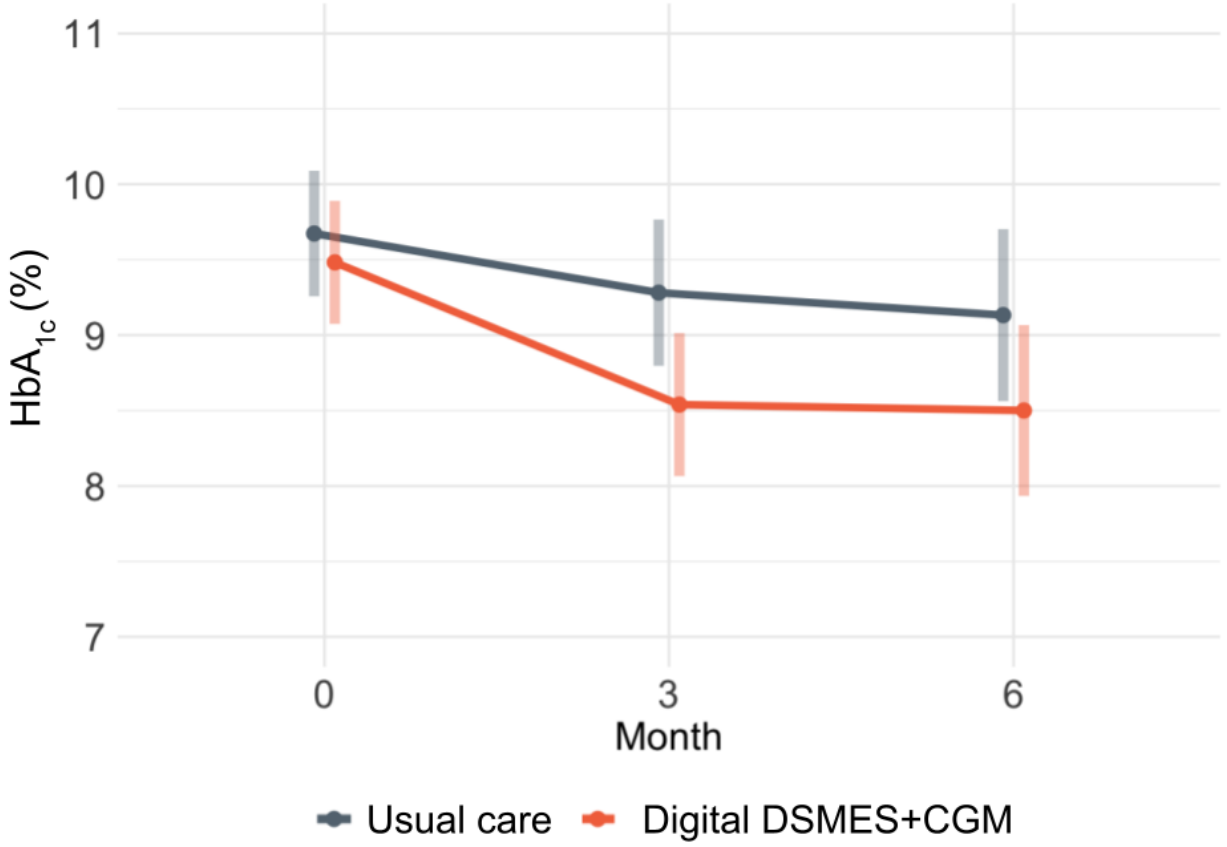

	Digital DSMES+CGM M (SE)	Usual care M (SE)	Adjusted mean difference (95% CI)	*P* value
HbA_{1c} (%): 3 mo	8.5 (0.2)	9.3 (0.3)	–0.7 (–1.4 to –0.1)	*.03*
HbA_{1c} (mmol/mol): 3 mo	69.8 (2.6)	77.9 (2.7)	–8.1 (–15.5 to –0.7)	
HbA_{1c} (%): 6 mo	8.5 (0.3)	9.1 (0.3)	–0.6 (–1.4 to 0.2)	.12
HbA_{1c} (mmol/mol): 6 mo	69.4 (3.1)	76.3 (3.1)	–6.9 (–15.7 to 1.9)	

XSL•FO
RenderX

Secondary Outcomes

The results for the secondary outcomes are presented in Table 2 (3-month outcomes), Table 3 (6-month outcomes), and Figure 3. CGM-derived glycemic measures were collected using the blinded FSL Pro sensors. At 3 months, participants in the digital DSMES+CGM condition had a 13.5% (95% CI 0.3% to 26.6%; *P*=.045) higher TIR and a 16.9% (95% CI −30.4% to −3.3%; *P*=.02) lower TAR >250 mg/dL than those in usual care. In addition, at 3 months, participants in the digital DSMES+CGM condition had 39.8 mg/dL (95% CI −72.2 to −7.4 mg/dL; *P*=.02) lower mean glucose and 0.9% (95% CI −1.7% to −0.2%; *P*=.02) lower GMI than those in usual care.

Table . Secondary clinical outcomes by condition at 3 months.

Variable	Digital DSMES[a]+CGM[b] (n=51), mean (SE)	Usual care (n=49), mean (SE)	Adjusted mean difference (95% CI)	*P* value[c]
% TIR[d] (70 - 180 mg/dL)	47.5 (4.6)	34.0 (4.8)	13.5 (0.3 to 26.6)	*.045*
% TAR[e] (>180 mg/dL)	51.7 (4.8)	65.1 (5.0)	−13.3 (−26.9 to 0.2)	.05
% TAR (>250 mg/dL)	23.7 (4.8)	40.6 (5.0)	−16.9 (−30.4 to −3.3)	*.02*
% TBR[f] (<70 mg/dL)	0.7 (0.4)	0.8 (0.4)	−0.1 (−1.2 to 0.9)	.78
Mean glucose (mg/dL)	199.5 (11.4)	239.4 (11.8)	−39.8 (−72.2 to −7.4)	*.02*
Mean SD of glucose (mg/dL)	40.0 (1.7)	44.4 (1.8)	−4.5 (−9.4 to 0.5)	.08
Mean % CV[g]	21.0 (0.9)	20.1 (0.9)	0.9 (−1.54 to 3.3)	.47
GMI[h] (%)	8.1 (0.3)	9.0 (0.3)	−0.9 (−1.7 to −0.2)	*.02*
Weight (lb)	212.4 (6.6)	214.7 (6.7)	−2.3 (−21.0 to 16.4)	.81
Systolic BP[i] (mm Hg)	129.5 (2.3)	130.6 (2.4)	−1.1 (−7.6 to 5.5)	.75
Diastolic BP (mm Hg)	83.7 (1.5)	83.5 (1.5)	0.2 (−3.9 to 4.3)	.92
Diabetes distress	0.8 (0.1)	0.9 (0.1)	−0.1 (−0.3 to 0.1)	.19

[a]DSMES: diabetes self-management education and support.

[b]CGM: continuous glucose monitoring.

[c]*P* values <.05 are italicized.

[d]TIR: time in range.

[e]TAR: time above range.

[f]TBR: time below range.

[g]CV: coefficient of variation.

[h]GMI: glucose management indicator.

[i]BP: blood pressure.

XSL•FO
RenderX

Table . Secondary clinical outcomes by condition at 6 months.

Variable	Digital DSMES[a]+CGM[b] (n=51), mean (SE)	Usual care (n=49), mean (SE)	Adjusted mean difference (95% CI)	*P* value[c]
% TIR[d] (70 - 180 mg/dL)	51.3 (4.8)	36.7 (5.0)	14.6 (1.0 to 28.2)	*.04*
% TAR[e] (>180 mg/dL)	47.7 (4.9)	62.6 (5.1)	−14.9 (−29.0 to −0.9)	*.04*
% TAR (>250 mg/dL)	19.7 (5.0)	37.6 (5.1)	−17.9 (−32.0 to −3.8)	*.01*
% TBR[f] (<70 mg/dL)	0.9 (0.4)	0.6 (0.4)	0.3 (−0.8 to 1.4)	.60
Mean glucose (mg/dL)	196.7 (11.8)	233.1 (12.2)	−36.4 (−70.0 to −2.9)	*.03*
Mean SD of glucose (mg/dL)	38.9 (1.8)	44.1 (1.9)	−5.2 (−10.3 to −0.1)	*.047*
Mean % CV[g]	21.3 (0.9)	20.6 (0.9)	0.7 (−1.8 to 3.2)	.59
GMI[h] (%)	8.0 (0.3)	8.9 (0.3)	−0.9 (−1.7 to −0.1)	*.03*
Weight (lb)	212.1 (6.6)	210.6 (6.7)	1.5 (−17.2 to 20.2)	.87
Systolic BP[i] (mm Hg)	123.7 (2.4)	125.6 (2.3)	−1.9 (−8.5 to 4.6)	.56
Diastolic BP (mm Hg)	82.3 (1.5)	85.1 (1.5)	−2.8 (−6.9 to 1.3)	.18
Diabetes distress	0.7 (0.1)	0.8 (0.1)	−0.1 (−0.3 to 0.0)	.14

[a]DSMES: diabetes self-management education and support.

[b]CGM: continuous glucose monitoring.

[c]*P* values <.05 are italicized.

[d]TIR: time in range.

[e]TAR: time above range.

[f]TBR: time below range.

[g]CV: coefficient of variation.

[h]GMI: glucose management indicator.

[i]BP: blood pressure.

XSL•FO
RenderX

Figure 3. Mean change in (A) GMI (%); (B) % TIR; 70 - 180 mg/dL, (C) % TAR; >180 mg/dL, and (D) mean glucose (mg/dL) over time by condition. CGM: continuous glucose monitoring; DSMES: diabetes self-management education and support; GMI: glucose management indicator; TAR: time above range; TIR: time in range.

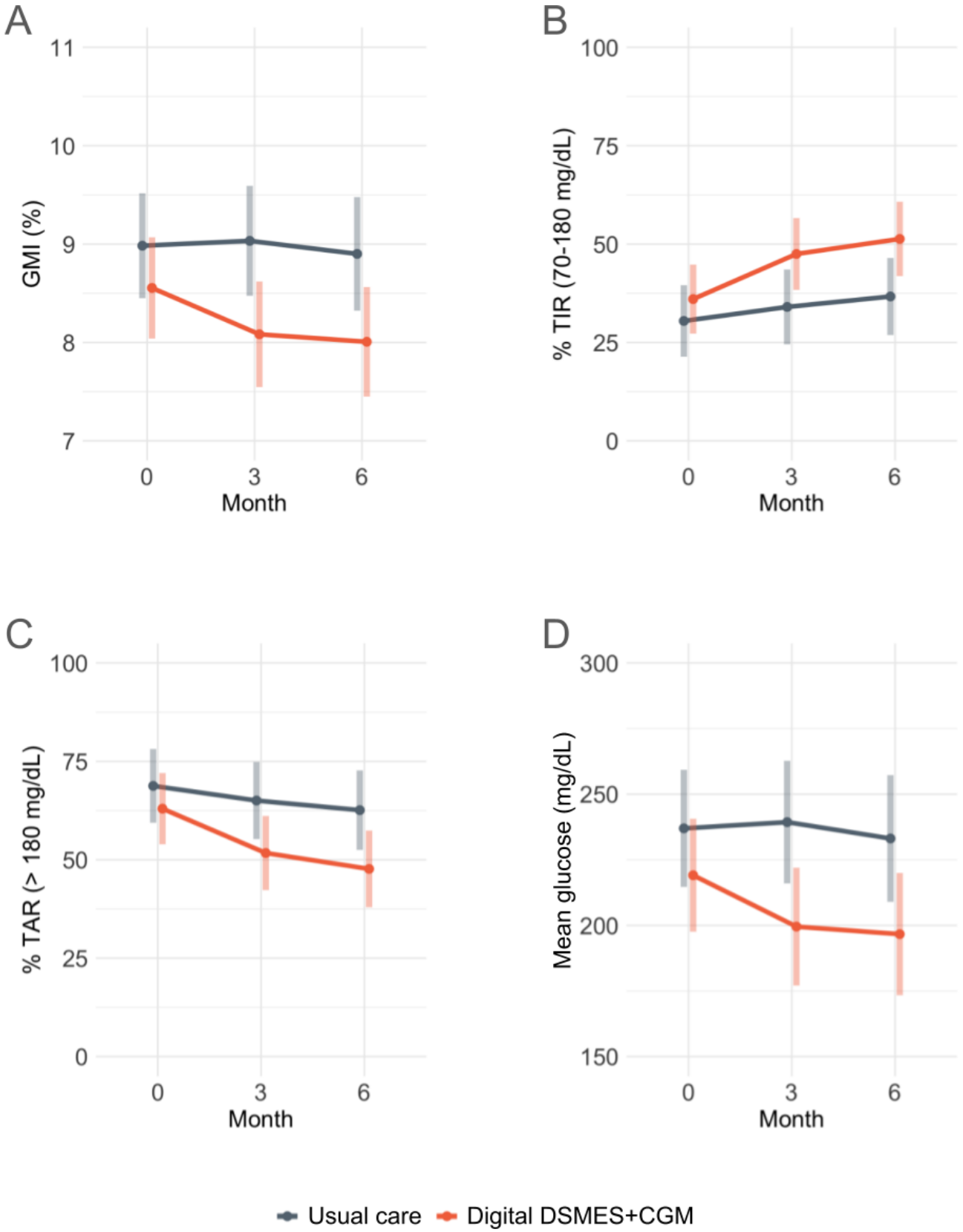

At 6 months, participants in the digital DSMES+CGM condition had a 14.6% (95% CI 1.0% to 28.2%; *P*=.04) higher TIR, a −14.9% (95% CI −29.0% to −0.9%; *P*=.04) lower TAR >180 mg/dL, and a 17.9% (95% CI −32.0% to −3.8%; *P*=.01) lower TAR >250 mg/dL than those in usual care. In addition, at 6 months, participants in the digital DSMES+CGM condition had 36.4 mg/dL (95% CI −70.0 to −2.9 mg/dL; *P*=.03) lower mean glucose, 5.2 mg/dL (95% CI −10.3 to −0.01 mg/dL; *P*=.047)

XSL•FO
RenderX

lower SD of glucose, and 0.9% (95% CI −1.7% to −0.1%; *P*=.03) lower GMI than those in usual care.

Finally, no statistically significant differences between the intervention and control groups were seen for weight, systolic or diastolic blood pressure, or diabetes distress at either 3 or 6 months.

When conducting sensitivity analyses excluding participants who never enrolled or participated in the intervention, results remained largely the same as the ITT analyses with 2 exceptions: participants in the digital DSMES+CGM condition had significantly lower TAR >180 mg/dL (−16.3%, 95% CI −30.3% to −2.3%; *P*=.02) and SD of glucose (−5.8 mg/dL, 95% CI −11.0 to −0.7 mg/dL; *P*=.03) than those in usual care at 3 months.

There were no reported serious adverse events related to the study devices or program, and the number of adverse events did not differ significantly across conditions.

Discussion

Results from this RCT evaluating an integrated digital DSMES program+CGM revealed that participants randomized to the integrated solution experienced improvements in HbA_{1c} and CGM-derived measures of glycemic control compared to usual care over 6 months. Though the difference between conditions in HbA_{1c} at 6 months was not significant within an ITT analysis, sensitivity analyses showed that participants who enrolled in the digital DSMES+CGM condition had significantly lower HbA_{1c} than those in usual care. These findings are clinically meaningful given that, on average, the intervention group saw an almost 1-point reduction in HbA_{1c} levels over the study period—a degree of change that is twice as large as the benchmark for a clinically significant change in HbA_{1c} of 0.5% [33] and is associated with a reduced risk of diabetes-related complications, including microvascular disease, myocardial infarction, heart failure, stroke, and death [34].

Given the focus of this trial on the integration of CGM with DSMES [35], CGM-derived glycemic measures also improved in the digital DSMES+CGM condition compared to usual care. From baseline to 6 months, participants in the digital DSMES+CGM condition experienced a 15% increase in percent TIR and a 15% decrease in percent TAR >180 mg/dL, which were more favorable improvements compared to usual care and at a magnitude of change 3-fold greater than what previous research has demonstrated to be clinically meaningful [36]. Additionally, participants in our integrated solution had a 0.89% lower GMI than those in usual care at 6 months, indicating that those receiving the digital DSMES program+CGM had better glycemic control and a lower risk for diabetes-related complications [37]. Together, our findings demonstrate improvement in TIR and reduced variability of blood glucose among participants who received the digital DSMES+CGM integrated solution compared to usual care.

The design of our integrated solution leveraged a combination of the benefits of CGM and DSMES programs for type 2 diabetes management. Previous research revealed that using CGM for 8 weeks helped people with diabetes significantly reduce their HbA_{1c} by an average 1-point reduction, which is consistent with our present study and demonstrates the importance of self-monitoring for diabetes management [38]. Additional research also showed that coupling CGMs with DSMES over 16 weeks led to significantly greater mean TIR, significantly less TAR, and a greater reduction in mean HbA_{1c} compared to DSMES alone [15]. However, a unique feature of our clinical trial is the integration of both real-time CGM and digital DSMES over a longer 6-month period to evaluate the consistent and sustainable aspects of the intervention.

As part of the intervention, research participants in the digital DSMES program received 1:1 support from a CDCES, which included systematic outreach at fixed program weeks in addition to continuous health coaching and tailored feedback in response to members' biometric data collected throughout the program (eg, CGM glucose values and body weight). Their strategic outreach and continuous monitoring highlights the proactive feature of the integrated solution, aligning well with past research revealing that proactive case management significantly improved glycemic control in patients with diabetes [39]. Furthermore, digital DSMES+CGM participants in our study were also highly engaged, as exhibited by frequent weighing, tracking meals, exercising, and CGM-wearing for several days per week. These behaviors are impactful, given their association with improved glycemic control in the digital DSMES program [26,40].

Our study has several strengths, with a notable advantage being the rigor of its single-blind RCT design. The sensitivity analysis also allowed us to more accurately estimate the impact of the integrated solution on clinical outcomes among participants who had the opportunity to engage with it. The convergence between the ITT and sensitivity results demonstrates the beneficial impact of the digital DSMES program+CGM on clinical outcomes. Furthermore, by leveraging a decentralized recruitment process, we were able to include a racially, socioeconomically, and geographically diverse group of participants. This inclusive, patient-centered approach enhanced convenience and minimized dropout by bringing clinicians to participants' homes for assessments [41]. Moreover, our trial's postrandomization attrition rate was 12%, despite having a longitudinal study design that required several follow-up assessments and biospecimen collection. Notably, this attrition rate is lower than the approximate 30% dropout rate typically seen across site-based clinical trials [42], underscoring the benefits of decentralized operations.

We also acknowledge several limitations with the current study. First, the complexities of remote study operations, such as costs, shipping delays, scheduling conflicts, and recruitment difficulties, caused us to fall short of our intended sample size and highlighted the challenges of implementing a decentralized RCT. However, the clinically and statistically significant results observed with a smaller sample size underscore the effectiveness of the digital DSMES+CGM integrated solution. Second, 9 participants randomized to the intervention did not participate in the digital DSMES program and, therefore, did not experience the integrated solution. Reasons contributing to treatment nonadherence included screen failure due to disqualifying TIR

XSL•FO
RenderX

or previous exposure to the DSMES program, failure to apply to the DSMES program, and eventual loss to follow-up. That said, all 9 participants had at least 1 follow-up assessment, so we were still able to evaluate their clinical outcomes. Finally, as part of the integrated solution, participants were expected to wear an FSL 14-Day sensor continuously for the duration of the study (along with 2 FSL Pro sensors at baseline, 3 mo, and 6 mo for measurement purposes), which may have increased participant burden and impacted adherence over the 6-month study period. However, these results likely reflect real-world compliance with CGM usage on a long-term, continuous basis.

In conclusion, the statistically significant and clinically meaningful improvement in HbA_{1c} and CGM outcomes among participants in the digital DSMES+CGM group demonstrates that combining CGMs with a digital DSMES program can have a beneficial impact on glucose management among adults with type 2 diabetes. Nevertheless, longer and larger-scale trials are warranted to assess the reliability of our findings and the benefit of our integrated solution on potentially reducing the risk of long-term health complications. Ultimately, increasing access to digital DSMES programs with integrated CGM capabilities may be a promising approach to help adults with type 2 diabetes manage their condition, improve health outcomes, and support broader efforts to enhance the quality and effectiveness of diabetes care.

Acknowledgments

The authors thank Melinda Merry, Bailey Peterka, and Carolyn Salter at Omada Health, Inc, for providing research members in the digital diabetes self-management education and support program with one-on-one health coaching and care team support as Certified Diabetes Care and Education Specialists. They also thank Stephanie Salwen at Evidation Health, Inc, and Hyein Cho, who provided extensive project management and data monitoring throughout the full study operational period, as well as Hope Chang and Lisa Stanton for assistance with additional medication analyses.

Funding

This study was funded by Omada Health, Inc, and Abbott Diabetes Care, Inc—the sponsors of this study. The authors receive no commissions or other compensation that are tied to the results of this research or its publication. The sponsors had no role in determining how the data were analyzed or how the results were interpreted beyond what is described in the manuscript.

Data Availability

The datasets generated during and/or analyzed during the current study are proprietary and cannot be made publicly available but may be available from the corresponding author upon reasonable request.

Authors' Contributions

TCD, SL, CBJ, DK, CJC, and AB were involved in the conception, design, and conduct of the study. JBN, SZ, and JN conducted the analysis, and all authors were involved in the interpretation of the results. AB and JBN wrote the first draft of the manuscript, and all authors edited, reviewed, and approved the final version of the manuscript.

Conflicts of Interest

AB, JBN, SZ, JN, AM, CBJ, and SL are or were salaried employees of Omada Health, Inc, at the time this manuscript was written, with AB, JBN, JN, AM, CBJ, and SL receiving stock options and/or equity compensation. HJL and TCD are employees of Abbott Diabetes Care, Inc, and receive salary and stock. CJC is a salaried employee of Evidation Health, Inc, and receives stock options. DK has received research support from Abbott Diabetes Care, Inc, and received payment as a contractor for Evidation Health, Inc, to perform clinical oversight of adverse event reporting throughout this study.

Checklist 1
CONSORT-EHEALTH (V 1.6.1) checklist.
[PDF File, 1192 KB - jmir_v28i1e78321_app1.pdf]

References

1. American Diabetes Association Professional Practice Committee. 2. Diagnosis and classification of diabetes: standards of care in diabetes-2024. Diabetes Care 2024 Jan 1;47(Suppl 1):S20-S42. [doi: 10.2337/dc24-S002] [Medline: 38078589]
2. Farmaki P, Damaskos C, Garmpis N, Garmpi A, Savvanis S, Diamantis E. Complications of the type 2 diabetes mellitus. Curr Cardiol Rev 2020;16(4):249-251. [doi: 10.2174/1573403X1604201229115531] [Medline: 33407062]
3. Type 2 diabetes. Centers for Disease Control and Prevention (CDC). 2023. URL: https://www.cdc.gov/diabetes/about/about-type-2-diabetes.html [accessed 2024-04-24]

XSL•FO
RenderX

4. Chrvala CA, Sherr D, Lipman RD. Diabetes self-management education for adults with type 2 diabetes mellitus: a systematic review of the effect on glycemic control. Patient Educ Couns 2016 Jun;99(6):926-943. [doi: 10.1016/j.pec.2015.11.003] [Medline: 26658704]
5. Davis J, Fischl AH, Beck J, et al. 2022 National standards for diabetes self-management education and support. Diabetes Care 2022 Feb 1;45(2):484-494. [doi: 10.2337/dc21-2396] [Medline: 35050365]
6. Veazie S, Winchell K, Gilbert J, et al. Mobile health applications for self-management of diabetes. : Agency for Healthcare Research and Quality (US); 2018 URL: https://effectivehealthcare.ahrq.gov/sites/default/files/pdf/technical-brief-31_mobile-health-applications-for-self-management-of-diabetes.pdf [accessed 2026-04-24]
7. Improving access to diabetes education. Centers for Disease Control and Prevention (CDC). 2022. URL: https://www.cdc.gov/diabetes/health-equity/improving-access-education.html [accessed 2024-04-19]
8. Miller EM. Using continuous glucose monitoring in clinical practice. Clin Diabetes 2020 Dec;38(5):429-438. [doi: 10.2337/cd20-0043] [Medline: 33384468]
9. Fonda SJ, Graham C, Munakata J, Powers JM, Price D, Vigersky RA. The cost-effectiveness of real-time continuous glucose monitoring (RT-CGM) in type 2 diabetes. J Diabetes Sci Technol 2016 Jul;10(4):898-904. [doi: 10.1177/1932296816628547] [Medline: 26843480]
10. Bahia L, Mello KF, Lemos LLP, Costa NL, Mulinari E, Malerbi DA. Cost-effectiveness of continuous glucose monitoring with FreeStyle Libre® in Brazilian insulin-treated patients with types 1 and 2 diabetes mellitus. Diabetol Metab Syndr 2023 Nov 25;15(1):242. [doi: 10.1186/s13098-023-01208-5] [Medline: 38001509]
11. Jiao Y, Lin R, Hua X, et al. A systematic review: cost-effectiveness of continuous glucose monitoring compared to self-monitoring of blood glucose in type 1 diabetes. Endocrinol Diabetes Metab 2022 Nov;5(6):e369. [doi: 10.1002/edm2.369] [Medline: 36112608]
12. American Diabetes Association Professional Practice Committee. 7. Diabetes technology: standards of medical care in diabetes-2022. Diabetes Care 2022 Jan 1;45(Supplement_1):S97-S112. [doi: 10.2337/dc22-S007] [Medline: 34964871]
13. Allen NA, Fain JA, Braun B, Chipkin SR. Continuous glucose monitoring counseling improves physical activity behaviors of individuals with type 2 diabetes: a randomized clinical trial. Diabetes Res Clin Pract 2008 Jun;80(3):371-379. [doi: 10.1016/j.diabres.2008.01.006] [Medline: 18304674]
14. Allen NA, Jacelon CS, Chipkin SR. Feasibility and acceptability of continuous glucose monitoring and accelerometer technology in exercising individuals with type 2 diabetes. J Clin Nurs 2009 Feb;18(3):373-383. [doi: 10.1111/j.1365-2702.2008.02533.x] [Medline: 19191984]
15. Aronson R, Brown RE, Chu L, et al. IMpact of flash glucose Monitoring in pEople with type 2 diabetes inadequately controlled with non-insulin antihyperglycaemic thErapy (IMMEDIATE): a randomized controlled trial. Diabetes Obes Metab 2023 Apr;25(4):1024-1031. [doi: 10.1111/dom.14949] [Medline: 36546594]
16. Polonsky WH, Fortmann AL, Soriano EC, Guzman SJ, Funnell MM. The AH-HA! Project: transforming group diabetes self-management education through the addition of flash glucose monitoring. Diabetes Technol Ther 2023 Mar;25(3):194-200. [doi: 10.1089/dia.2022.0419] [Medline: 36409486]
17. Dixon RF, Zisser H, Layne JE, et al. A virtual type 2 diabetes clinic using continuous glucose monitoring and endocrinology visits. J Diabetes Sci Technol 2020 Sep;14(5):908-911. [doi: 10.1177/1932296819888662] [Medline: 31762302]
18. Bergenstal RM, Layne JE, Zisser H, et al. Remote application and use of real-time continuous glucose monitoring by adults with type 2 diabetes in a virtual diabetes clinic. Diabetes Technol Ther 2021 Feb;23(2):128-132. [doi: 10.1089/dia.2020.0396] [Medline: 33026839]
19. Tchero H, Kangambega P, Briatte C, Brunet-Houdard S, Retali GR, Rusch E. Clinical effectiveness of telemedicine in diabetes mellitus: a meta-analysis of 42 randomized controlled trials. Telemed J E Health 2019 Jul;25(7):569-583. [doi: 10.1089/tmj.2018.0128] [Medline: 30124394]
20. Whaley CM, Bollyky JB, Lu W, et al. Reduced medical spending associated with increased use of a remote diabetes management program and lower mean blood glucose values. J Med Econ 2019 Sep;22(9):869-877. [doi: 10.1080/13696998.2019.1609483] [Medline: 31012392]
21. Turner RM, Ma Q, Lorig K, Greenberg J, DeVries AR. Evaluation of a diabetes self-management program: claims analysis on comorbid illnesses, health care utilization, and cost. J Med Internet Res 2018 Jun 22;20(6):e207. [doi: 10.2196/jmir.9225] [Medline: 29934284]
22. Phillip M, Bergenstal RM, Close KL, et al. The digital/virtual diabetes clinic: the future is now-recommendations from an international panel on diabetes digital technologies introduction. Diabetes Technol Ther 2021 Feb;23(2):146-154. [doi: 10.1089/dia.2020.0375] [Medline: 32905711]
23. Ku EJ, Park JI, Jeon HJ, Oh T, Choi HJ. Clinical efficacy and plausibility of a smartphone-based integrated online real-time diabetes care system via glucose and diet data management: a pilot study. Intern Med J 2020 Dec;50(12):1524-1532. [doi: 10.1111/imj.14738] [Medline: 31904890]
24. Hallberg SJ, McKenzie AL, Williams PT, et al. Effectiveness and safety of a novel care model for the management of type 2 diabetes at 1 year: an open-label, non-randomized, controlled study. Diabetes Ther 2018 Apr;9(2):583-612. [doi: 10.1007/s13300-018-0373-9] [Medline: 29417495]

XSL•FO
RenderX

25. Athinarayanan SJ, Adams RN, Hallberg SJ, et al. Long-term effects of a novel continuous remote care intervention including nutritional ketosis for the management of type 2 diabetes: a 2-year non-randomized clinical trial. Front Endocrinol (Lausanne) 2019;10:348. [doi: 10.3389/fendo.2019.00348] [Medline: 31231311]
26. Berthoumieux A, Linke S, Merry M, Megliola A, Juusola J, Napoleone J. Long-term results of a digital diabetes self-management and education support program among adults with type 2 diabetes: a retrospective cohort study. Sci Diabetes Self Manag Care 2024 Feb;50(1):19-31. [doi: 10.1177/26350106231221456] [Medline: 38240247]
27. Polonsky WH, Fisher L, Earles J, et al. Assessing psychosocial distress in diabetes: development of the diabetes distress scale. Diabetes Care 2005 Mar;28(3):626-631. [doi: 10.2337/diacare.28.3.626] [Medline: 15735199]
28. Leelarathna L, Evans ML, Neupane S, et al. Intermittently scanned continuous glucose monitoring for type 1 diabetes. N Engl J Med 2022 Oct 20;387(16):1477-1487. [doi: 10.1056/NEJMoa2205650] [Medline: 36198143]
29. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2023. URL: https://www.R-project.org/ [accessed 2026-04-24]
30. Chakraborty H, Gu H. A mixed model approach for intent-to-treat analysis in longitudinal clinical trials with missing values. : RTI Press; 2009 URL: https://www.rti.org/rti-press-publication/mixed-model-approach-intent-treat-analysis-longitudinal-clinical-trials-missing-values/fulltext.pdf [accessed 2026-04-24]
31. Little RJA. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 1988 Dec;83(404):1198-1202. [doi: 10.1080/01621459.1988.10478722]
32. Barr DJ, Levy R, Scheepers C, Tily HJ. Random effects structure for confirmatory hypothesis testing: keep it maximal. J Mem Lang 2013 Apr;68(3):255-278. [doi: 10.1016/j.jml.2012.11.001] [Medline: 24403724]
33. Little RR, Rohlfing CL, Sacks DB, National Glycohemoglobin Standardization Program (NGSP) Steering Committee. Status of hemoglobin A1c measurement and goals for improvement: from chaos to order for improving diabetes care. Clin Chem 2011 Feb;57(2):205-214. [doi: 10.1373/clinchem.2010.148841] [Medline: 21148304]
34. Stratton IM, Adler AI, Neil HA, et al. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. BMJ 2000 Aug 12;321(7258):405-412. [doi: 10.1136/bmj.321.7258.405] [Medline: 10938048]
35. Beck RW, Bergenstal RM, Riddlesworth TD, et al. Validation of time in range as an outcome measure for diabetes clinical trials. Diabetes Care 2019 Mar;42(3):400-405. [doi: 10.2337/dc18-1444] [Medline: 30352896]
36. Dovc K, Battelino T. Time in range centered diabetes care. Clin Pediatr Endocrinol 2021;30(1):1-10. [doi: 10.1297/cpe.30.1] [Medline: 33446946]
37. Yapanis M, James S, Craig ME, O'Neal D, Ekinci EI. Complications of diabetes and metrics of glycemic management derived from continuous glucose monitoring. J Clin Endocrinol Metab 2022 May 17;107(6):e2221-e2236. [doi: 10.1210/clinem/dgac034] [Medline: 35094087]
38. Ehrhardt NM, Chellappa M, Walker MS, Fonda SJ, Vigersky RA. The effect of real-time continuous glucose monitoring on glycemic control in patients with type 2 diabetes mellitus. J Diabetes Sci Technol 2011 May 1;5(3):668-675. [doi: 10.1177/193229681100500320] [Medline: 21722581]
39. Choe HM, Mitrovich S, Dubay D, Hayward RA, Krein SL, Vijan S. Proactive case management of high-risk patients with type 2 diabetes mellitus by a clinical pharmacist: a randomized controlled trial. Am J Manag Care 2005 Apr;11(4):253-260. [Medline: 15839185]
40. Wilson-Anumudu F, Quan R, Castro Sweet C, et al. Early insights from a digitally enhanced diabetes self-management education and support program: single-arm nonrandomized trial. JMIR Diabetes 2021 Feb 22;6(1):e25295. [doi: 10.2196/25295] [Medline: 33616533]
41. Sinha SD, Chary Sriramadasu S, Raphael R, Roy S. Decentralisation in clinical trials and patient centricity: benefits and challenges. Pharmaceut Med 2024 Mar;38(2):109-120. [doi: 10.1007/s40290-024-00518-x] [Medline: 38453755]
42. Alexander W. The uphill path to successful clinical trials: keeping patients enrolled. P T 2013 Apr;38(4):225-227. [Medline: 23785228]

Abbreviations

CDCES: Certified Diabetes Care and Education Specialist
CGM: continuous glucose monitoring
DSMES: diabetes self-management education and support
FSL: FreeStyle Libre
GMI: glucose management indicator
HbA_{1c}: hemoglobin A_{1c}
ITT: intention-to-treat
MCAR: missing completely at random
RCT: randomized controlled trial
TAR: time above range
TBR: time below range

XSL•FO
RenderX

TIR: time in range

Edited by G Eysenbach, T Leung; submitted 16.Jun.2025; peer-reviewed by AN Ali, M Lei; revised version received 09.Apr.2026; accepted 12.Apr.2026; published 14.May.2026.

Please cite as:
Berthoumieux A, Naqvi JB, Zion S, Napoleone J, McGuill A, Cerrada CJ, Lee HJ, Dunn TC, Kerr D, Jasik CB, Linke S
A Digital Diabetes Self-Management Education and Support Program Integrated With Continuous Glucose Monitoring for Type 2 Diabetes: Randomized Controlled Trial
J Med Internet Res 2026;28:e78321
URL: https://www.jmir.org/2026/1/e78321
doi:10.2196/78321

© Ashley Berthoumieux, Jeanean B Naqvi, Sean Zion, Jenna Napoleone, Amanda McGuill, Christian J Cerrada, Hyun Jung Lee, Timothy C Dunn, David Kerr, Carolyn Bradner Jasik, Sarah Linke. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Nicotine Dependence and Quit Self-Confidence in a Smoking Cessation Program Using a Group-Based Digital Peer-Supported App and Cigarette Consumption–Adjusted Nicotine Aids Among Japanese Workers: Retrospective Cohort Study

Shota Yoshihara[1,2], OTR, MSc; Kayoko Takahashi[1,3], OTR, ScD; Chiaki Uemura[2], RD; Shin Murakami[2], PhD; Satoru Amano[1,3], OTR, DMSc; Daichi Harada[2], BA; Ying Jiang[4], PhD; Hiroshi Yamato[4], MD, PhD

[1]Department of Rehabilitation Sciences, Graduate School of Medical Sciences, Kitasato University, 1-15-1, Kitasato, Sagamihara, Kanagawa, Japan

[2]A10 Lab Inc, Chuo-ku, Tokyo, Japan

[3]School of Allied Health Science, Kitasato University, Sagamihara, Kanagawa, Japan

[4]Department of Health Development, Institute of Industrial Ecological Sciences, University of Occupational and Environmental Health, Japan, Kitakyushu, Fukuoka, Japan

Corresponding Author:
Shota Yoshihara, OTR, MSc
Department of Rehabilitation Sciences, Graduate School of Medical Sciences, Kitasato University, 1-15-1, Kitasato, Sagamihara, Kanagawa, Japan

Abstract

Background: Completion rates for smoking cessation treatments under Japan's national health insurance system remain suboptimal. A workplace cessation program, combining nicotine gum or patches with a group-based digital peer-supported app, has reported high cessation success rates. Although nicotine dependence is generally associated with lower cessation success, and self-confidence is generally associated with higher success, these associations may differ by tobacco product type. Evidence on these relationships in app-based cessation programs remains limited.

Objective: This study aimed to examine the independent and combined associations of nicotine dependence and self-confidence in quitting with smoking cessation success among cigarette-only smokers, heated tobacco product–only users, and dual users.

Methods: This retrospective cohort study used data from a workplace cessation program in Japan. Participants were eligible if they were employed, owned a smartphone, and self-enrolled in the program. Recruitment was conducted through workplace promotion and individual outreach, primarily via email from companies. The program combined a digital peer-support app with nicotine gum or patches, adjusted according to cigarette consumption. The app included anonymous peer-support group chats of up to 5 participants, where participants shared progress, photos, and comments. Nicotine dependence was assessed by the time to the first cigarette after waking (high: ≤30 min; low: >30 min). Self-confidence for quitting was rated on a 0 - 10 scale and dichotomized at the median. A 4-level variable combined nicotine dependence and self-confidence for quitting. Logistic regression analyses were conducted by tobacco product type, and odds ratios (ORs) and 95% CIs were estimated.

Results: A total of 2143 participants were included in the analysis. Their mean age was 46.5 (SD 10.9) years, and approximately 90% were men. Overall cessation success was 53.8% (1152/2143). Participants with high self-confidence had a higher cessation success rate than those with low self-confidence (529/834, 63.4% vs 623/1309, 47.6%), with an OR of 1.81 (95% CI 1.55 - 2.12). Nicotine dependence was significantly associated with cessation success only among cigarette-only smokers; those with low nicotine dependence had a higher OR than those with high dependence (OR 1.59, 95% CI 1.07 - 2.37). Across all tobacco product types, the subgroup with low nicotine dependence and high self-confidence showed the highest cessation success, followed by the subgroup with high nicotine dependence but high self-confidence.

Conclusions: By distinguishing cigarette-only smokers, heated tobacco product–only users, and dual users in a workplace cessation program, this study provides novel, product-specific evidence regarding nicotine dependence and self-confidence related to cessation success. This study extends prior research beyond cigarette-only smokers by examining tobacco user groups and suggests that self-confidence may predict cessation across tobacco product types, whereas the role of nicotine dependence may be product-specific. These findings may inform tailored, scalable smoking-cessation support in workplaces; however, they should be interpreted with caution because cessation outcomes were self-reported.

(J Med Internet Res 2026;28:e84792) doi:10.2196/84792

XSL•FO
RenderX

KEYWORDS

smoking cessation; nicotine dependence; self-confidence; heated tobacco products; digital health; workplace health promotion

Introduction

Smoking is the leading cause of mortality from noncommunicable diseases in Japan and contributes to the incidence of ischemic heart disease, stroke, chronic obstructive pulmonary disease, and cancer [1,2]. As the societal burden of smoking increases, smoking cessation has become a global imperative [3]. Despite global and national efforts to strengthen tobacco control policies, the desire to quit smoking among current smokers in Japan remains low at 20.7% [4]. Recent national survey data indicate that conventional cigarette-only smokers, heated tobacco product (HTP)-only users, and dual users account for 60.5%, 29.2%, and 9.2% of current male smokers and 56.2%, 35.3%, and 7% of current female smokers, respectively [4].

In Japan, a standard outpatient smoking cessation program is available for smokers with nicotine dependence, including HTP users [5]. This program includes a 12-week course of nicotine replacement therapy using nicotine patches, delivered over 5 sessions, along with counseling provided by physicians either in person or via telemedicine. However, the completion rate of this smoking cessation program remains low.

Several systematic reviews have reported that digital interventions, such as smartphone apps, can aid smoking cessation [6,7]. Moreover, a meta-analysis indicated that group-based interventions incorporating peers enhanced smoking cessation success rates [8]. Our previous study revealed a significantly higher cessation success rate among current smokers using a digital peer-supported app together with nicotine aids in the workplace (59.2%) than among those using only nicotine aids (38.7%) [9]. Furthermore, the cessation success rate was significantly higher among employed smokers using only HTPs (63.3%) than among those using only cigarettes (52.8%) [10].

Multiple factors identified by previous studies as being associated with smoking cessation include age [11,12], educational level [12-14], socioeconomic status and occupation [15], marital or cohabitation status [12,13,15], number of friends who smoke [14], housing conditions [11], and social or family support [13,15]. Among these, nicotine dependence has been consistently recognized as an important factor in smoking cessation [11,16-20]. Additionally, self-confidence for quitting, defined as a smoker's confidence in their ability to quit, is considered a key psychological factor for smoking cessation. A meta-analysis of 54 prospective studies showed that individuals who remained abstinent at follow-up had higher self-confidence for quitting, such as self-abstinence self-efficacy, than those who continued smoking, and that greater self-efficacy was associated with an increased likelihood of cessation success [21]. Moreover, previous reports have suggested a negative association between self-confidence for quitting and nicotine dependence [22,23].

Nonetheless, most previous studies have focused solely on cigarette smokers, and little is known about whether nicotine dependence and self-confidence for quitting are associated with smoking cessation among HTP or dual users, or whether these associations vary by tobacco product type. This gap in the literature is important because studies have increasingly suggested that smoking cessation success may differ by tobacco product type, indicating the need for product-specific evaluation of factors associated with cessation [24-26]. HTPs are marketed in over 40 countries [27], with Japan being the largest market [28]. HTP users heat tobacco to generate aerosols (vapors) rather than smoke [29] and are marketed as substitutes for cigarettes [30]. Although HTPs emit lower concentrations of combustion-related toxicants, the aerosols produced by HTPs still contain several harmful constituents, including those related to additives such as glycerol [31]. Moreover, the proportion of smokers intending to quit in Japan has declined since 2014, in parallel with increasing HTP use [4], suggesting that the promotion of cessation among HTP users represents a public health issue [32].

To address these gaps, in the present study, we investigated the following research questions in a company-implemented smoking cessation program combining a digital peer-support app with nicotine gum or patches:

- Are nicotine dependence and self-confidence for quitting independently associated with smoking cessation success?
- Do these associations vary by tobacco product type (cigarette-only smokers, HTP-only users, and dual users)?
- What is the joint association of nicotine dependence and self-confidence for quitting with smoking cessation success across these tobacco product groups?

Methods

Study Design

This retrospective cohort study evaluated differences in smoking cessation success rates according to nicotine dependence and self-confidence in quitting among users of a group-based digital peer-supported app and nicotine aids (nicotine patches or gum). The study was conducted among current smokers working at 38 companies from May 2024 to March 2025.

Sampling Procedures

Participants were recruited for a company-implemented, workplace-based smoking cessation program using a convenience sampling approach based on self-selection through individual outreach and workplace promotion over approximately 3 weeks, beginning about 2 weeks before the program started. Participants were invited primarily via email from their companies and union officials associated with their health insurance societies. In addition, posters and other promotional materials were displayed on company bulletin boards to raise awareness. Participants enrolled in the smoking cessation program by independently submitting the provided application form (eg, Google Forms).

XSL•FO
RenderX

Participant Inclusion and Exclusion Criteria

Current smokers enrolled in the smoking cessation program who used the digital peer-supported app were included in this retrospective analysis. Eligible participants were individuals who owned either an iOS or Android smartphone and were enrolled in their company's corporate health insurance program. Smokers were defined as participants who reported smoking at least 1 cigarette per day, including novel tobacco products such as HTPs and e-cigarettes. Participants who never used the digital peer-supported app, or nicotine gum or patches, as well as those with missing outcome data, were excluded from the primary analysis.

Smoking Cessation Program

Our smoking cessation program, which lasted for 12 weeks, was conducted remotely by each company (Figure 1). In the preprogram phase, the program's structure was explained during a real-time web-based briefing session, which included educational content led by a physician who had quit smoking. The session covered topics such as the health effects of smoking, the benefits of quitting smoking, coping with withdrawal, creating a smoke-free environment, and adopting alternative behaviors. The digital peer-supported app was also introduced, and the proper use of nicotine patches and gum was demonstrated. For those who were unable to attend a live session, the same content was made available in an e-learning course.

Figure 1. Timeline of study procedures.

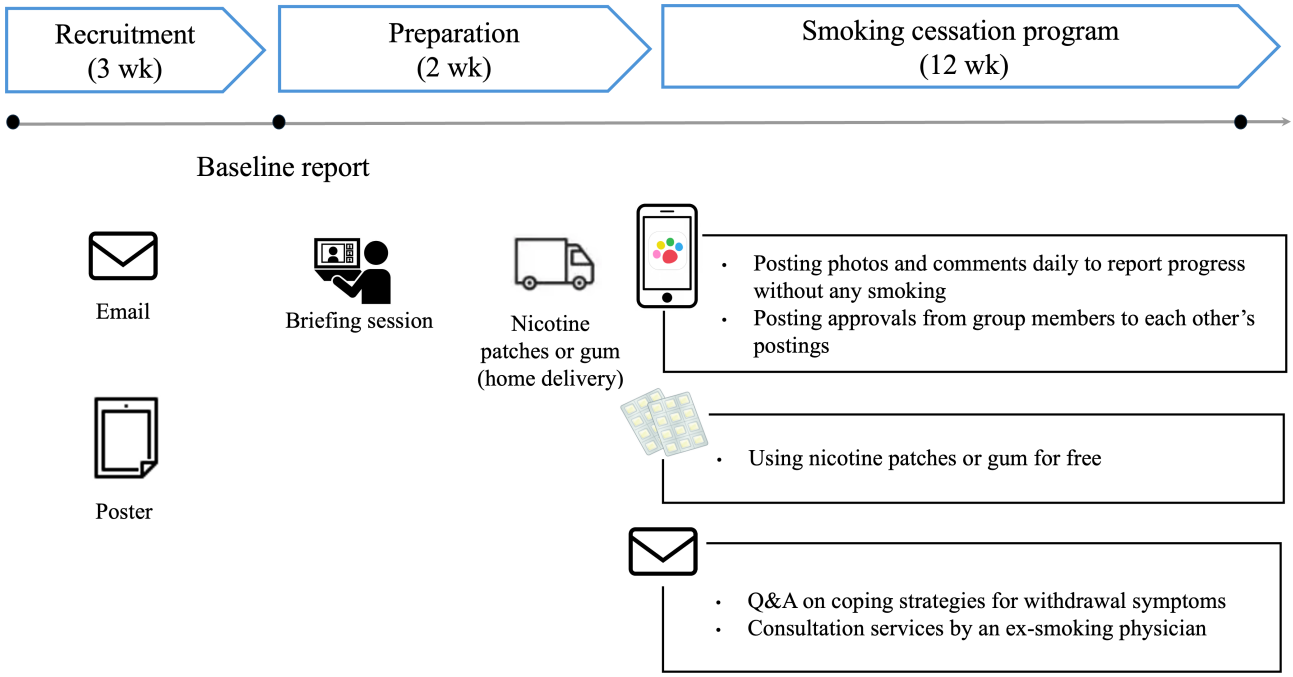

As summarized in Table 1, nicotine patches or gum were provided according to the participants' daily cigarette consumption (1 - 4, 5 - 10, or ≥11 cigarettes per day), workplace environment, and preference regarding gum use.

After the web-based briefing session, the participants received complimentary nicotine patches or gum, which were available over the counter, and began their smoking cessation efforts using the digital peer-supported app. Additional support, such as "Q&A on Coping Strategies for Withdrawal Symptoms," was provided through a dedicated email inquiry service, and an ex-smoking physician specializing in smoking cessation offered consultation services.

Table . Provision patterns of nicotine gum or patches.

Number of cigarettes/day	Gum allowed at the workplace	Gum not allowed at the workplace
≥11	Pattern A: • 7 sheets of 20 cm² patches • +96 pieces of gum	Pattern C: • 14 sheets of 20 cm² patches • +14 sheets of 10 cm² patches
5 - 10	Pattern B: • 14 sheets of 10 cm² patches • +48 pieces of gum	Pattern D: • 20 sheets of 10 cm² patches
1-4	Pattern E[a]: • 96 pieces of gum alone	Pattern E[a]: • 96 pieces of gum alone

[a]Based on physician consultation, a 10 cm² patch was considered too strong for individuals who smoke 1 - 4 cigarettes/day; hence, the gum-only regimen was applied uniformly to this group, irrespective of workplace gum restrictions.

XSL•FO
RenderX

Digital Peer-Supported App

A commercially available digital peer-supported app ("Minchalle"; A10 Lab Inc), accessible on both iOS and Android smartphones, was used in this study [9,10,33-36]. This app was originally developed to help users build desirable habits, such as exercising, dieting, and learning English conversations, and it has been downloaded more than 1.6 million times. In this study, the app was adapted to support smoking cessation [9,10].

The participants were automatically assigned to teams comprising 5 individuals based on their registration order in the app. This app facilitated group chats among team members sharing a common goal (in this case, smoking cessation) and promoted anonymous interactions by allowing users to post messages and photos about smoking abstinence (Figure 2). All app interactions were designed to be anonymous, and the participants could use the app freely. In addition to self-monitoring abstinence, the app supported peer interaction through team-based participation, daily posting, approvals from group members, and progress tracking.

Within the app, the participants could (1) post photos and comments daily to report their progress without smoking, (2) post approvals from group members on each other's posts, (3) receive feedback on the team's collective performance toward its goals, and (4) be automatically removed from the group if they exceeded a specified inactivity period of 15 consecutive days. The participants posted comments or photographs multiple times daily and interacted with others as they wished, although participation was not mandatory. The user interface, including its layout and interactive features, has been described in detail in previous studies [9,10].

Figure 2. Examples of app screens. (A) Selecting a group consisting of 5 members, (B) uploading a photo of the day with a comment related to abstinence, (C) displaying the content of posts in the group, (D) responding to posts by group members and pressing the approval button, and (E) reviewing the records of abstinence through photos. These functions are designed to support self-monitoring of abstinence, peer interaction through team-based participation and approvals from group members, and progress tracking within the app.

Data Collection Procedures

Baseline participant characteristics were collected using a web-based survey before the start of the program. App engagement data were automatically extracted from the digital peer-supported app database during the 12-week program. Smoking cessation outcomes were collected at the end of the program using a web-based form, where participants reported their last smoking date.

Participant Characteristics and Baseline Measures

Baseline data, including age (continuous), sex (male or female), and smoking-related characteristics (smoking duration, number of cigarettes smoked per day, number of previous smoking cessation attempts, importance of smoking cessation, and type of cigarettes smoked), were collected via a web-based survey.

Smoking duration (years) was assessed by asking the question, "How many years have you been smoking?" The participants were instructed to select one of the following options: 0 - 2, 3 - 4, 5 - 9, 10 - 19, 20 - 29, or ≥30 years. Smoking duration was determined without distinguishing between tobacco product types. Given that HTPs have been available for less than 10 years, HTP-only users with a smoking duration of more than 10 years were assumed to have previously used combustible cigarettes but were currently using only HTPs. Cigarette consumption was assessed by asking the question, "How many cigarettes do you currently smoke per day?" The participants provided continuous responses. The number of previous smoking cessation attempts was recorded as a categorical variable (0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or ≥11), indicating the number of prior quit attempts before the current program. The importance of smoking cessation was assessed with the item, "How important is it to you that you can quit smoking?" rated on a self-reported 0 - 10 scale.

App Engagement Measures

The digital peer-supported app log data automatically recorded in the database were obtained to assess the participants' engagement with the app. Two data indicators were analyzed: (1) posting frequency during the 12-week program and (2) the

XSL•FO
RenderX

number of approval button presses from group members. Posting frequency was defined as the total number of posts (including messages, challenge posts, normal photos, and stamps) made by each participant throughout the program. The number of approval button presses from group members in the app was extracted from log data and regarded as a continuous variable.

Tobacco Product Type

At baseline, cigarette smoking status was assessed by asking the question, "Which type of cigarette do you currently use?" The participants selected from multiple options: (1) combustible cigarettes, (2) HTPs (eg, IQOS, Ploom, glo, PULZE, and WEECKE), or (3) e-cigarettes. The use of e-cigarettes was not considered in this study because the prevalence of e-cigarette use is low in Japan [37-39]; moreover, e-cigarettes in Japan do not contain nicotine. Based on their responses, the participants were categorized into cigarette-only smokers, HTP-only users, and dual users [10].

Measures of Nicotine Dependence and Self-Confidence for Quitting

Nicotine dependence was assessed by asking the question, "How soon after waking do you smoke your first cigarette?" The participants selected from the following options: within 5 minutes, 6 to 30 minutes, 31 to 60 minutes, or after 60 minutes. Based on previous studies and clinical guidelines [40-42], the participants who responded "within 5 minutes" or "6 to 30 minutes" were categorized into the high nicotine dependence group, whereas the participants who answered "31 to 60 minutes" or "after 60 minutes" were assigned to the low nicotine dependence group.

Self-confidence for quitting was assessed with the item, "How confident are you that you can quit smoking?" rated on a self-reported 0‐10 scale [43-45]. Participants were dichotomized at the median into low and high self-confidence groups.

In this study, a composite 4-level variable was created by combining nicotine dependence (low or high) with self-confidence (low or high): (1) low nicotine dependence and low self-confidence, (2) low nicotine dependence and high self-confidence, (3) high nicotine dependence and low self-confidence, and (4) high nicotine dependence and high self-confidence.

Outcome Measure: Smoking Cessation Success Rates

Smoking cessation success rates were calculated based on the participants' self-reported abstinence from smoking for at least 4 consecutive weeks. At the end of the 12-week program, the participants submitted the date when they had last smoked via a web-based form. The participants were considered to have successfully quit smoking if the number of days between their self-reported "last smoking date" and the final report date was at least 4 weeks. Those who reported complete abstinence during this period without smoking a single cigarette were classified as "quitters."

Sample Size, Power, and Precision

This retrospective cohort study used data from a company-implemented smoking cessation program. No formal sample size calculation was performed a priori, as the sample size was determined by the number of eligible participants available during the study period. Similarly, no formal power analysis was conducted for the same reason. Precision was assessed by reporting odds ratios (ORs) with 95% CIs.

Statistical Analysis and Covariates

A series of logistic regression analyses was conducted to examine the associations between nicotine dependence and self-confidence in quitting with smoking cessation success. The outcome variable, "cessation success," was binary (1="quit" and 0="not quit"), and ORs with 95% CIs were calculated to estimate the strength of the associations. These analyses were conducted both overall and separately according to tobacco product types (cigarette-only smokers, HTP-only users, and dual users) to explore potential differences in associations across tobacco use types.

First, participants in the high-dependence group were used as the reference to analyze the relationship between nicotine dependence and smoking cessation. The model evaluated whether those in the low-dependence group had higher ORs for achieving cessation. Subsequently, the relationship between self-confidence for quitting and smoking cessation rates was evaluated. Using participants in the low self-confidence group as the reference, the ORs for the high self-confidence group were estimated. Finally, the combined effects of nicotine dependence and self-confidence were analyzed, with the groups characterized by high and low self-confidence serving as the reference.

Three levels of adjustment were applied. Model 1 was adjusted for age (continuous) and sex (male or female). Model 2 was additionally adjusted for smoking duration (0‐2, 3‐4, 5‐9, 10‐19, 20‐29, or ≥30 y) and the number of previous smoking cessation attempts (0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or ≥11). Model 3 was further adjusted for nicotine dependence (low or high) and self-confidence for quitting (low or high) to account for potential mutual confounding factors. In all models, companies were incorporated as clusters. The conceptual framework is shown in Figure S1 in Multimedia Appendix 1.

All statistical analyses were performed using Stata version 19.0 (StataCorp). This study was reported in accordance with the American Psychological Association (APA) Journal Article Reporting Standards for Quantitative Research (JARS-Quant) [46].

Additional Analyses for Missing Data

To assess the robustness of our findings, we conducted additional analyses. First, the Little test was used to assess whether the missing outcome data were missing completely at random, using all covariates included in the primary adjusted analyses (χ^2_6=41.77; P<.001). We also performed multiple imputations for missing outcome data using a logistic imputation model, with the outcome variable imputed and baseline covariates treated as regular variables. The imputation model included the same covariates as in the primary adjusted analyses, and 20 imputed datasets were generated. To improve model stability, the number of previous smoking cessation attempts

XSL•FO
RenderX

was collapsed into 5 categories (0, 1, 2, 3, and ≥4) for the imputation procedure. Estimates were combined across imputations using the Rubin rules. Second, as a sensitivity analysis, all participants with missing outcome data were assumed to have failed smoking cessation, and the analyses were repeated using the same covariates as in the primary adjusted models.

Ethical Considerations

This study was approved by the Research Ethics Committee of the Institutional Review Board of Kitasato University School of Allied Health Sciences (approval number: 2024-008-3). This retrospective cohort study was a secondary analysis of data from a company-implemented smoking cessation program in which the amount of nicotine patches or gum provided was adjusted according to the participants' cigarette consumption. The program content, delivery, and operation were not modified for the purpose of this research, and no additional data were collected specifically for this study. In accordance with the ethics committee–approved protocol, the requirement for additional individual informed consent for this secondary analysis was waived, and an opt-out procedure was used. Information about the study was publicly disclosed on the institutional website, and participants were given the opportunity to decline participation or withdraw their data at any time. All data were anonymized or deidentified prior to analysis, and no personally identifiable information was included in the analytic dataset. Participants received no financial compensation or other incentives. We confirm that individual participants or users cannot be identified in any images included in the paper or Multimedia Appendix 1.

Results

The participant flowchart is presented in Figure 3. Overall, 3018 employed smokers were enrolled in the smoking cessation program, which combined the use of the digital peer-supported app and the use of nicotine gum or patches. After excluding 781 individuals who had never used the app, nicotine gum, or patches, and 93 individuals who had missing outcome data on smoking cessation at the end of the program, the final analytic sample included 2143 participants who adhered to the intervention and provided complete outcome data.

Participant characteristics, total and stratified by cessation status, are shown in Table 2. Compared with the cessation failure group, the cessation success group was older, had a higher proportion of HTP-only users, and demonstrated higher self-confidence for quitting, higher posting frequency in the digital peer-supported application, and more approvals from group members. In addition, the cessation success group tended to have a smoking history of 20 years or more, smoke fewer cigarettes per day, rate the importance of smoking cessation more highly, and show lower nicotine dependence. Characteristics stratified by nicotine dependence, self-confidence for quitting, and their combination are shown in Tables S1-S3 in Multimedia Appendix 1, respectively.

XSL•FO
RenderX

Figure 3. Flowchart showing participant eligibility, exclusions, data screening, and the final analytic sample in this retrospective cohort study using data from a company-implemented smoking cessation program conducted between May 2024 and March 2025. HTP: heated tobacco product.

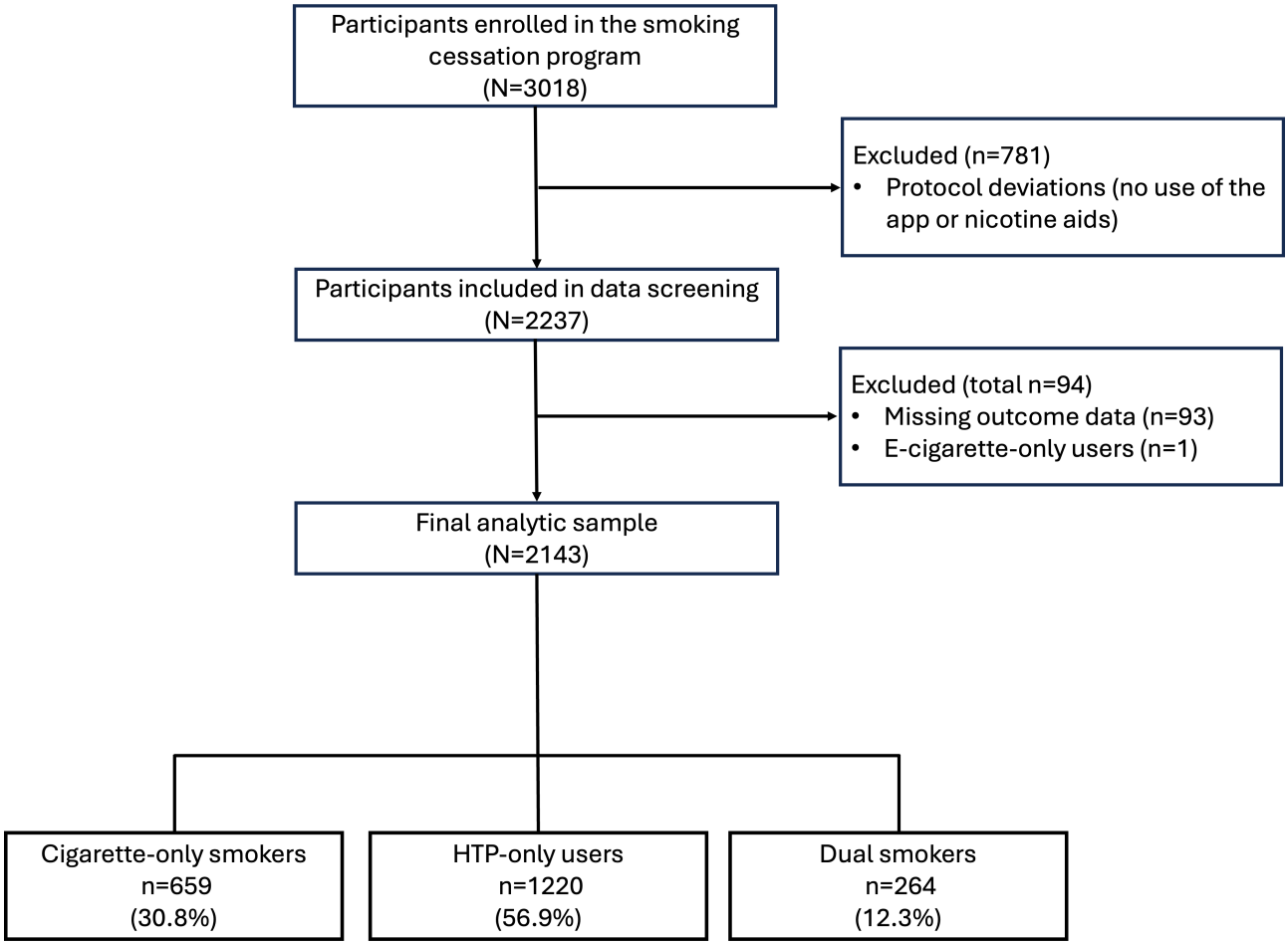

XSL•FO
RenderX

Table . Baseline characteristics of current smokers enrolled in a company-implemented smoking cessation program in Japan, overall and stratified by smoking cessation status—a retrospective cohort study of employees from 38 companies conducted from May 2024 to March 2025 (N=2143)[a].

Characteristics	Total (N=2143)	Cessation success (n=1152)	Cessation failure (n=991)	*P* value
Age (y), mean (SD)	46.5 (10.9)	47.7 (10.4)	45.0 (11.3)	<.001
Sex, n (%)				.61
Male	1903 (88.8)	1027 (89.1)	876 (88.4)	
Female	209 (9.8)	111 (9.6)	98 (9.9)	
Others	31 (1.4)	14 (1.2)	17 (1.7)	
Type of cigarettes used, n (%)				<.001
Cigarette-only smokers	659 (30.8)	332 (28.8)	327 (33)	
Heated tobacco product–only users	1220 (56.9)	696 (60.4)	524 (52.9)	
Dual users	264 (12.3)	124 (10.8)	140 (14.1)	
Smoking duration (y), n (%)				.05
0 - 2	51 (2.4)	24 (2.1)	27 (2.7)	
3 - 4	68 (3.2)	31 (2.7)	37 (3.7)	
5 - 9	190 (8.9)	97 (8.4)	93 (9.4)	
10 - 19	391 (18.2)	191 (16.6)	200 (20.2)	
20 - 29	673 (31.4)	369 (32)	304 (30.7)	
≥30	770 (35.9)	440 (38.2)	330 (33.3)	
Cigarettes/day, mean (SD)	17.8 (8.4)	17.4 (8.5)	18.1 (8.2)	.06
Number of prior smoking cessation attempts, n (%)				.85
0	666 (31.1)	361 (31.3)	305 (30.8)	
1	645 (30.1)	347 (30.1)	298 (30.1)	
2	355 (16.6)	197 (17.1)	158 (15.9)	
3	257 (12.0)	135 (11.7)	122 (12.3)	
≥4	220 (10.3)	112 (9.7)	108 (10.9)	
Provision patterns of nicotine gum or patches, n (%)				.32
A: 7 sheets of 20 cm^2 patches + 96 pieces of gum	1584 (73.9)	834 (72.4)	750 (75.7)	
B: 14 sheets of 10 cm^2 patches + 48 pieces of gum	491 (22.9)	284 (24.7)	207 (20.9)	
C: 14 sheets of 20 cm^2 patches + 14 sheets of 10 mg patches	32 (1.5)	15 (1.3)	17 (1.7)	
D: 20 sheets of 10 cm^2 patches	8 (0.4)	4 (0.3)	4 (0.4)	
E: 96 pieces of gum alone	28 (1.3)	15 (1.3)	13 (1.3)	
Importance of smoking cessation (score: 0 - 10), mean (SD)	7 (2.4)	7.2 (2.3)	6.7 (2.4)	<.001
Nicotine dependence, n (%)				.07
High (≤30 min after waking)	1653 (77.1)	871 (75.6)	782 (78.9)	
Low (>30 min after waking)	490 (22.9)	281 (24.4)	209 (21.1)	
Self-confidence for quitting, n (%)				<.001
High (scores of 5 - 10)	834 (38.9)	529 (45.9)	305 (30.8)	

XSL•FO
RenderX

Characteristics	Total (N=2143)	Cessation success (n=1152)	Cessation failure (n=991)	*P* value
Low (scores of 0 - 4)	1309 (61.1)	623 (54.1)	686 (69.2)	
Posting frequency in the digital peer-supported app, mean (SD)	32.6 (78.4)	51.5 (98.7)	10.7 (33.1)	<.001
Posting approvals from group members, mean (SD)	54.8 (100.6)	88.1 (121.2)	16.0 (44.9)	<.001

[a]Data are presented as means and SDs for continuous variables and as numbers (percentages) for categorical variables. The Pearson chi-square test was used for categorical variables, and the Mann-Whitney *U* test was used for continuous variables, as appropriate. For presentation purposes, the number of previous smoking cessation attempts was collapsed into 5 categories (0, 1, 2, 3, and ≥4) in this table. In all regression analyses, the original detailed categories were used.

The association between nicotine dependence and smoking cessation success among all participants, according to tobacco product type, is presented in Table 3. Among all participants, smokers with low nicotine dependence showed a nonsignificantly higher likelihood of smoking cessation success compared to those with high dependence (281/490, 57.3% vs 871/1753, 52.7%, respectively). When stratified by tobacco product type, the association between lower nicotine dependence and cessation success remained significant among cigarette-only smokers (96/166, 57.8% vs 236/493, 47.9%, respectively). The adjusted OR was 1.59 (95% CI 1.07 - 2.37) in model 3. However, no significant associations were observed among HTP-only and dual users.

XSL•FO
RenderX

Table . Odds ratios (ORs) for smoking cessation success according to nicotine dependence among current smokers enrolled in a company-implemented smoking cessation program in Japan—a retrospective cohort study of employees from 38 companies conducted from May 2024 to March 2025[a].

Characteristics	Nicotine dependence	
	High (≤30 min after waking)	Low (>30 min after waking)
Overall		
Number of participants	1653	490
Smoking cessation success, n (%)	871 (52.7)	281 (57.3)
Model 1, OR (95% CI)[b]	1 (reference)	1.25 (1.05 - 1.49)
Model 2, OR (95% CI)[c]	1 (reference)	1.22 (1.02 - 1.46)
Model 3, OR (95% CI)[d]	1 (reference)	1.14 (0.96 - 1.36)
Cigarette-only smokers		
Number of participants	493	166
Smoking cessation success, n (%)	236 (47.9)	96 (57.8)
Model 1, OR (95% CI)[b]	1 (reference)	1.66 (1.11 - 2.47)
Model 2, OR (95% CI)[c]	1 (reference)	1.67 (1.11 - 2.51)
Model 3, OR (95% CI)[d]	1 (reference)	1.59 (1.07 - 2.37)
HTP[e]-only users		
Number of participants	952	268
Smoking cessation success, n (%)	541 (56.8)	155 (57.8)
Model 1, OR (95% CI)[b]	1 (reference)	1.07 (0.84 - 1.36)
Model 2, OR (95% CI)[c]	1 (reference)	1.05 (0.81 - 1.36)
Model 3, OR (95% CI)[d]	1 (reference)	0.97 (0.74 - 1.26)
Dual users		
Number of participants	208	56
Smoking cessation success, n (%)	94 (45.2)	30 (53.6)
Model 1, OR (95% CI)[b]	1 (reference)	1.41 (0.74 - 2.71)
Model 2, OR (95% CI)[c]	1 (reference)	1.38 (0.70 - 2.70)
Model 3, OR (95% CI)[d]	1 (reference)	1.27 (0.68 - 2.37)

[a]ORs and 95% CIs were estimated using logistic regression analyses. In all models, companies were incorporated as clusters.

[b]Adjusted for age and sex.

[c]Adjusted for age, sex, smoking duration, and the number of previous smoking cessation attempts.

[d]Adjusted for age, sex, smoking duration, the number of previous smoking cessation attempts, and self-confidence in quitting.

[e]HTP: heated tobacco product.

The association between self-confidence for quitting and cessation success among all participants and tobacco product types is presented in Table 4. Among all participants, the higher self-confidence group was associated with higher smoking cessation success rates: 47.6% (623/1309) in participants with low self-confidence and 63.4% (529/834) in those with high self-confidence. Compared with the low self-confidence group, the adjusted OR for cessation success in the high self-confidence group was 1.81 (95% CI 1.55 - 2.12) in model 3. When stratified by tobacco product types, this positive association remained significant across all groups. Among cigarette-only smokers, the cessation success rate increased from 44.7% (177/396; low) to 58.9% (155/263; high), with an adjusted OR of 1.71 (95% CI 1.26 - 2.31) for the high group in model 3. Similarly, among HTP-only users, the success rates were 51.4% (380/740; low) and 65.8% (316/480; high), and the adjusted OR was 1.78 (95% CI 1.44 - 2.21) in model 3. A similar association was observed in dual users, with adjusted ORs of 2.95 (95% CI 1.58 - 5.53) for the high group in model 3.

XSL•FO
RenderX

Table . Odds ratios (ORs) for smoking cessation success according to self-confidence in quitting among current smokers enrolled in a company-implemented smoking cessation program in Japan: a retrospective cohort study of employees from 38 companies conducted from May 2024 to March 2025[a].

Characteristics	Self-confidence for quitting	
	Low (scores of 0-4)	High (scores of 5 - 10)
Overall		
Number of participants	1309	834
Smoking cessation success, n (%)	623 (47.6)	529 (63.4)
Model 1, OR (95% CI)[b]	1 (reference)	1.84 (1.56 - 2.18)
Model 2, OR (95% CI)[c]	1 (reference)	1.83 (1.57 - 2.15)
Model 3, OR (95% CI)[d]	1 (reference)	1.81 (1.55 - 2.12)
Cigarette-only smokers		
Number of participants	396	263
Smoking cessation success, n (%)	177 (44.7)	155 (58.9)
Model 1, OR (95% CI)[b]	1 (reference)	1.73 (1.28 - 2.35)
Model 2, OR (95% CI)[c]	1 (reference)	1.76 (1.30 - 2.40)
Model 3, OR (95% CI)[d]	1 (reference)	1.71 (1.26 - 2.31)
HTP[e]-only users		
Number of participants	740	480
Smoking cessation success, n (%)	380 (51.4)	316 (65.8)
Model 1, OR (95% CI)[b]	1 (reference)	1.76 (1.41 - 2.19)
Model 2, OR (95% CI)[c]	1 (reference)	1.78 (1.43 - 2.21)
Model 3, OR (95% CI)[d]	1 (reference)	1.78 (1.44 - 2.21)
Dual users		
Number of participants	173	91
Smoking cessation success, n (%)	66 (38.2)	58 (63.7)
Model 1, OR (95% CI)[b]	1 (reference)	2.58 (1.49 - 4.49)
Model 2, OR (95% CI)[c]	1 (reference)	3.00 (1.60 - 5.61)
Model 3, OR (95% CI)[d]	1 (reference)	2.95 (1.58 - 5.53)

[a]ORs and 95% CIs were estimated using logistic regression analyses. In all models, companies were incorporated as clusters.

[b]Adjusted for age and sex.

[c]Adjusted for age, sex, smoking duration, and the number of previous smoking cessation attempts.

[d]Adjusted for age, sex, smoking duration, the number of previous smoking cessation attempts, and nicotine dependence.

[e]HTP: heated tobacco product.

The combined association of nicotine dependence and self-confidence for quitting with cessation success among all participants, according to tobacco product types, is shown in Table 5 and Figure 4. Among all participants, the smoking cessation success rate was highest in participants with low nicotine dependence and high self-confidence (159/241, 66%), followed by those with high dependence and high self-confidence (370/593, 62.4%), those with low dependence and low self-confidence (122/249, 49%), and those with high dependence and low self-confidence (501/1060, 47.3%). Using the high dependence and low self-confidence group as the reference, the adjusted ORs for cessation success were 1.78 (95% CI 1.53 - 2.06) in the high dependence and high self-confidence group and 2.12 (95% CI 1.55 - 2.91) in the low dependence and high self-confidence group in model 2. A similar pattern was observed when stratified by tobacco product types. Among cigarette-only smokers, those with low dependence and high self-confidence had the highest cessation success rate (52/79, 65.8%), with an adjusted OR of 2.85 (95% CI 1.48 - 5.49), whereas those with high dependence and high self-confidence achieved a success rate of 56% (103/184), with an adjusted OR of 1.65 (95% CI 1.14 - 2.38). Among HTP-only

XSL•FO
RenderX

users, cessation success was also highest in the low dependence and high self-confidence group (91/137, 66.4%; OR 1.78, 95% CI 1.19 - 2.66), followed by the high dependence and high self-confidence group (225/343, 65.6%; OR 1.74, 95% CI 1.41 - 2.14). Among dual users, the same pattern was observed, with cessation rates of 64% (16/25) and 63.6% (42/66) and adjusted ORs of 3.21 (95% CI 1.37 - 7.53) and 3.26 (95% CI 1.59 - 6.68) for the low dependence and high self-confidence group and high dependence and high self-confidence group, respectively.

Among the 2237 participants eligible for outcome assessment after excluding those who never used the intervention components, 93 (4.2%) had missing outcome data. Little's test was significant, indicating that the missing outcome data were not missing completely at random. We, therefore, conducted additional analyses using multiple imputation (Tables S4-S6 in Multimedia Appendix 1) and a sensitivity analysis in which all participants with missing outcome data were assumed to have failed smoking cessation (Tables S7-S9 in Multimedia Appendix 1). In both additional analyses, the overall pattern of results was consistent with that of the primary complete-case analysis, with no material changes in the point estimates or 95% CIs (Tables S4-S9 in Multimedia Appendix 1). Baseline characteristics under the assumption that participants with missing outcome data had failed smoking cessation and those according to inclusion in the analytic sample or exclusion due to missing outcome data are summarized in Tables S10 and S11 in Multimedia Appendix 1, respectively.

Table . Odds ratios (ORs) for smoking cessation success according to combined categories of nicotine dependence and self-confidence in quitting among current smokers enrolled in a company-implemented smoking cessation program in Japan—a retrospective cohort study of employees from 38 companies conducted from May 2024 to March 2025[a].

Characteristics	Combination of nicotine dependence and self-confidence for quitting			
	High dependence and low self-confidence	Low dependence and low self-confidence	High dependence and high self-confidence	Low dependence and high self-confidence
Overall				
Number of participants	1060	249	593	241
Smoking cessation success, n (%)	501 (47.3)	122 (49)	370 (62.4)	159 (66)
Model 1, OR (95% CI)[b]	1 (reference)	1.10 (0.86 - 1.40)	1.77 (1.51 - 2.08)	2.17 (1.57 - 3.01)
Model 2, OR (95% CI)[c]	1 (reference)	1.10 (0.85 - 1.42)	1.78 (1.53 - 2.06)	2.12 (1.55 - 2.91)
Cigarette-only smokers				
Number of participants	309	87	184	79
Smoking cessation success, n (%)	133 (43)	44 (50.6)	103 (56)	52 (65.8)
Model 1, OR (95% CI)[b]	1 (reference)	1.46 (0.94 - 2.29)	1.60 (1.14 - 2.25)	2.78 (1.43 - 5.39)
Model 2, OR (95% CI)[c]	1 (reference)	1.50 (0.96 - 2.34)	1.65 (1.14 - 2.38)	2.85 (1.48 - 5.49)
HTP[d]-only users				
Number of participants	609	131	343	137
Smoking cessation success, n (%)	316 (51.9)	64 (48.9)	225 (65.6)	91 (66.4)
Model 1, OR (95% CI)[b]	1 (reference)	0.92 (0.67 - 1.25)	1.70 (1.40 - 2.07)	1.81 (1.22 - 2.69)
Model 2, OR (95% CI)[c]	1 (reference)	0.92 (0.67 - 1.25)	1.74 (1.41 - 2.14)	1.78 (1.19 - 2.66)
Dual users				
Number of participants	142	31	66	25
Smoking cessation success, n (%)	52 (36.6)	14 (45.2)	42 (63.6)	16 (64)
Model 1, OR (95% CI)[b]	1 (reference)	1.47 (0.66 - 3.25)	2.75 (1.47 - 5.13)	2.83 (1.27 - 6.35)
Model 2, OR (95% CI)[c]	1 (reference)	1.50 (0.64 - 3.48)	3.26 (1.59 - 6.68)	3.21 (1.37 - 7.53)

[a]ORs and 95% CIs were estimated using logistic regression analyses. In all models, companies were incorporated as clusters.

[b]Adjusted for age and sex.

[c]Adjusted for age, sex, smoking duration, and the number of previous smoking cessation attempts.

[d]HTP: heated tobacco product.

XSL•FO
RenderX

Figure 4. Adjusted odds ratios for smoking cessation success according to combined categories of nicotine dependence (ND) and self-confidence for quitting (SC) among current smokers enrolled in a company-implemented smoking cessation program in Japan—a retrospective cohort study of employees from 38 companies conducted from May 2024 to March 2025. Markers indicate adjusted odds ratios, and error bars indicate 95% CIs. Estimates are shown for all participants, cigarette-only smokers, heated tobacco product (HTP)-only users, and dual smokers. The reference group consisted of participants with high ND and low self-confidence for quitting.

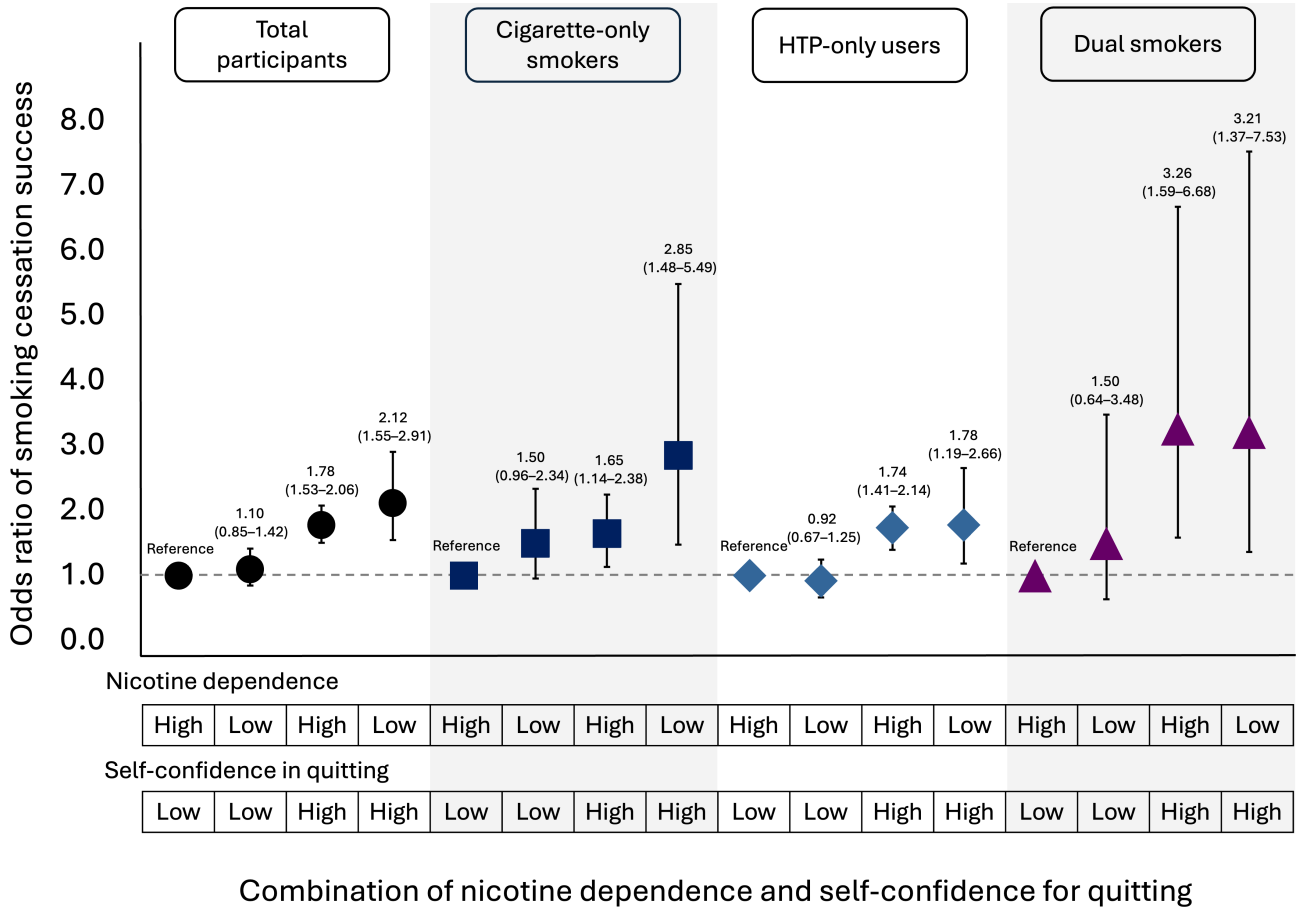

Discussion

Principal Results

This study examined the independent and combined associations of nicotine dependence and self-confidence in quitting with smoking cessation success across cigarette-only smokers, HTP-only users, and dual users in a company-implemented smoking cessation program using a digital peer-supported app and nicotine aids. Overall, higher self-confidence in quitting was consistently associated with greater smoking cessation success across tobacco product types, whereas lower nicotine dependence was associated with greater cessation success only among cigarette-only smokers. In addition, when nicotine dependence and self-confidence were considered jointly, groups with high self-confidence showed the most favorable cessation outcomes regardless of nicotine dependence level.

Interpretation and Comparison With Previous Studies

Our findings align with previous studies on cigarette-only smokers, which have consistently reported an association between nicotine dependence and smoking cessation [11,16-20]. However, no statistically significant association was observed among HTP-only and dual users in the present study. Although some app-based intervention studies have examined cessation outcomes according to tobacco product type, the association between nicotine dependence and smoking cessation success by tobacco product type has received limited attention [10,24,25]. In this context, our study extends previous research conducted among cigarette-only smokers [11,16-20] by stratifying participants according to tobacco product type and estimating the association between nicotine dependence and smoking cessation success in an app-based intervention.

Furthermore, our findings that higher self-confidence for quitting contributed to successful quitting are consistent with a meta-analysis [21], which highlights self-confidence for quitting as a key psychological determinant of smoking cessation. In addition, in the context of a digital mobile health app, a previous study examined psychological empowerment [47], whereas our study focused on self-confidence for quitting, which is a related but conceptually distinct construct.

One possible explanation for the nonsignificant association between nicotine dependence and cessation success among HTP-only and dual users may be differences in underlying motivations and perceptions compared with cigarette-only smokers [48,49]. For example, a study in the United States using cross-sectional data from 2470 young adults identified “fashion” as a notable perceived attribute of HTPs [48]. Another qualitative study involving adults in Switzerland and Japan found that many perceived the packaging of HTPs as appealing

XSL•FO
RenderX

and described HTPs as sophisticated, high-tech, or aspirational [49].

Furthermore, the digital peer-supported app may have contributed to self-confidence in quitting through its team-based, interactive features. Consistent with this interpretation, a pilot study using the same digital peer-supported app found that participants regarded the team system and challenge-tracking features as particularly useful and reported that the app provided emotional support and motivation through social comparison [50].

The tobacco product–specific findings in this study, particularly among dual users, should be interpreted cautiously because the corresponding estimates were less precise than those in the overall sample. Nevertheless, the overall findings suggest that both nicotine dependence and self-confidence for quitting, alongside tobacco product type, may be useful considerations in tailoring smoking cessation support in similar workplace app-based settings.

Limitations

This study has some limitations. First, the lack of a control group limited the ability to attribute the observed smoking cessation success solely to the app. However, in our proof-of-concept trial, the group using both smoking cessation aids and the digital peer-supported app achieved a higher smoking cessation success rate than the group using only smoking cessation aids [9]. Second, the outcomes were self-reported and not biochemically verified, which may have affected the accuracy of smoking cessation status. This may have led to an overestimation of the absolute quit rates, as some participants may have reported successful cessation due to social desirability or a desire to present themselves favorably. In addition, participants with higher self-confidence for quitting may have been more likely to report success, which could have particularly affected the magnitude of the observed quit rates among those with high self-confidence. However, although the lack of biochemical verification may have influenced the absolute success rates, it is less likely to have substantially altered the relative differences across tobacco product groups or nicotine dependence levels. Although expert consensus indicates that biochemical verification is often impractical and unnecessary in similar studies [51], future research should consider incorporating such methods to enhance the validity of the findings. Third, the sample size was determined by the number of eligible participants available during the study period rather than by an a priori sample size calculation. Although precision was assessed using 95% CIs, the study may have had limited statistical power, especially for subgroup and additional analyses. Fourth, we did not account for potential confounding variables, including sociodemographic characteristics (eg, occupation, education, and income), smartphone usage patterns, adverse events, continued nicotine product use, or a history of smoking cessation [13]. Additionally, other potentially relevant factors such as alcohol or substance use, mental health conditions (eg, depression or anxiety), chronic diseases (eg, asthma or chronic obstructive pulmonary disease), and behavioral stages of change (eg, precontemplation, contemplation, preparation, and action) were not measured. Furthermore, nicotine dependence and self-confidence for quitting were assessed only at baseline and may have changed during follow-up; therefore, the analyses could not account for time-varying changes in these factors.

Fifth, activity-related indicators within the digital peer-supported app were limited to basic posting frequency and duration. Future evaluations should incorporate detailed in-app behavioral metrics (eg, message content, peer engagement quality, and feedback responsiveness) to better assess the mechanisms of behavior change. Finally, this study sample consisted primarily of male employees from large Japanese companies, which limits the generalizability of our findings to other populations, including women and nonworking individuals. Moreover, because the participants were employees enrolled in a company-based smoking cessation program, they may have had greater organizational support for cessation than the general population. They may also have been more motivated to quit than smokers not enrolled in such programs, which may have introduced selection bias. These factors may also have contributed to the different distribution of cigarette-only smokers and HTP-only users observed in our sample compared with national estimates, thereby limiting generalizability.

Conclusions

By distinguishing cigarette-only smokers, HTP-only users, and dual users in a workplace cessation program, this study provides novel, product-specific evidence on how nicotine dependence and self-confidence relate to cessation success. By examining tobacco user groups, this study extends prior research beyond cigarette-only smokers and suggests that self-confidence may predict cessation across tobacco product types, whereas the role of nicotine dependence may be product-specific. These findings may inform tailored, scalable smoking-cessation support in workplaces; however, they should be interpreted with caution because cessation outcomes were self-reported and not biochemically verified. Further research, including randomized controlled trials with objective outcome verification, is needed to confirm these findings.

Acknowledgments

The authors would like to thank Editage for English-language editing. The authors declare the use of generative artificial intelligence (AI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to generative AI tools under full human supervision. The scientific content, interpretation, and conclusions were developed by the authors. All AI-assisted revisions were reviewed and approved by the authors, and the authors take full responsibility for the final manuscript.

XSL•FO
RenderX

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets generated and/or analyzed in this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: SY, KT, SA
Formal analysis: SY
Investigation: CU, SM, DH
Methodology: SY
Supervision: YJ, HY
Writing—original draft: SY
Writing—review and editing: all authors.

Conflicts of Interest

SY, CU, SM, and DH were employees of A10 Lab Inc at the time of the research. HY received compensation from A10 Lab Inc for supervising the program. Moreover, A10 Lab Inc funded the article processing charge for this manuscript.

Multimedia Appendix 1
Baseline characteristics and additional analyses.
[DOCX File, 161 KB - jmir_v28i1e84792_app1.docx]

References

1. Ikeda N, Saito E, Kondo N, et al. What has made the population of Japan healthy? Lancet 2011 Sep 17;378(9796):1094-1105. [doi: 10.1016/S0140-6736(11)61055-6] [Medline: 21885105]
2. Okui T. An age-period-cohort analysis of the difference in smoking prevalence between urban and non-urban areas in Japan (2004-2019). Epidemiol Health 2020;42:e2020072. [doi: 10.4178/epih.e2020072] [Medline: 33285059]
3. Reitsma MB, Fullman N, Ng M, et al. Smoking prevalence and attributable disease burden in 195 countries and territories, 1990–2015: a systematic analysis from the Global Burden of Disease Study 2015. The Lancet 2017 May;389(10082):1885-1906. [doi: 10.1016/S0140-6736(17)30819-X]
4. Summary of the results of the national health and nutrition survey in Japan [Report in Japanese]. : Ministry of Health, Labour and Welfare; 2023 URL: https://www.mhlw.go.jp/content/10900000/001338334.pdf [accessed 2025-09-02]
5. Special survey on the effectiveness of smoking cessation treatment under the nicotine dependence management fee: results from the 2017 survey related to the FY2016 revision of medical service fees [Report in Japanese]. : Central Social Insurance Medical Council, Ministry of Health, Labour and Welfare; 2018 URL: https://www.mhlw.go.jp/file/05-Shingikai-12404000-Hokenkyoku-Iryouka/0000192293.pdf [accessed 2025-09-02]
6. Haskins BL, Lesperance D, Gibbons P, Boudreaux ED. A systematic review of smartphone applications for smoking cessation. Transl Behav Med 2017 Jun;7(2):292-299. [doi: 10.1007/s13142-017-0492-2] [Medline: 28527027]
7. Ghorai K, Akter S, Khatun F, Ray P. mHealth for smoking cessation programs: a systematic review. J Pers Med 2014 Jul 18;4(3):412-423. [doi: 10.3390/jpm4030412] [Medline: 25563359]
8. Yuan P, Westmaas JL, Thrul J, Toussaert S, Hilton JF, White JS. Effectiveness of peer-support interventions for smoking cessation: a systematic review and meta-analysis. Nicotine Tob Res 2023 Aug 19;25(9):1515-1524. [doi: 10.1093/ntr/ntad059] [Medline: 37042206]
9. Yoshihara S, Takahashi K, Uemura C, Murakami S, Harada D, Yamato H. Effectiveness of a digital peer-supported app intervention in promoting smoking cessations: nonrandomized controlled trial. JMIR Mhealth Uhealth 2025 Aug 19;13:e68638. [doi: 10.2196/68638] [Medline: 40829152]
10. Yoshihara S, Takahashi K, Uemura C, et al. Smoking cessation strategies for different types of cigarette users using a digital peer-supported app and nicotine aids: prospective study. J Med Internet Res 2025 Aug 5;27:e75876. [doi: 10.2196/75876] [Medline: 40764052]
11. Monsó E, Campbell J, Tønnesen P, Gustavsson G, Morera J. Sociodemographic predictors of success in smoking intervention. Tob Control 2001 Jun;10(2):165-169. [doi: 10.1136/tc.10.2.165] [Medline: 11387538]
12. Lee CW, Kahende J. Factors associated with successful smoking cessation in the United States, 2000. Am J Public Health 2007 Aug;97(8):1503-1509. [doi: 10.2105/AJPH.2005.083527] [Medline: 17600268]
13. Caponnetto P, Polosa R. Common predictors of smoking cessation in clinical practice. Respir Med 2008 Aug;102(8):1182-1192. [doi: 10.1016/j.rmed.2008.02.017] [Medline: 18586479]

XSL•FO
RenderX

14. Rose JS, Chassin L, Presson CC, Sherman SJ. Prospective predictors of quit attempts and smoking cessation in young adults. Health Psychol 1996 Jul;15(4):261-268. [doi: 10.1037//0278-6133.15.4.261] [Medline: 8818672]
15. Chandola T, Head J, Bartley M. Socio-demographic predictors of quitting smoking: How important are household factors? Addiction 2004 Jun;99(6):770-777. [doi: 10.1111/j.1360-0443.2004.00756.x] [Medline: 15139875]
16. Hyland A, Li Q, Bauer JE, Giovino GA, Steger C, Cummings KM. Predictors of cessation in a cohort of current and former smokers followed over 13 years. Nicotine Tob Res 2004 Dec;6(suppl 3):S363-S369. [doi: 10.1080/14622200412331320761] [Medline: 15799599]
17. Hyland A, Borland R, Li Q, et al. Individual-level predictors of cessation behaviours among participants in the International Tobacco Control (ITC) Four Country Survey. Tob Control 2006 Jun;15(Suppl 3):iii83-iii94. [doi: 10.1136/tc.2005.013516] [Medline: 16754952]
18. Harris KJ, Okuyemi KS, Catley D, Mayo MS, Ge B, Ahluwalia JS. Predictors of smoking cessation among African-Americans enrolled in a randomized controlled trial of bupropion. Prev Med 2004 Apr;38(4):498-502. [doi: 10.1016/j.ypmed.2003.12.008] [Medline: 15020185]
19. Hymowitz N, Cummings KM, Hyland A, Lynn WR, Pechacek TF, Hartwell TD. Predictors of smoking cessation in a cohort of adult smokers followed for five years. Tob Control 1997;6(Suppl 2):S57-S62. [doi: 10.1136/tc.6.suppl_2.s57] [Medline: 9583654]
20. Siahpush M, McNeill A, Borland R, Fong GT. Socioeconomic variations in nicotine dependence, self-efficacy, and intention to quit across four countries: findings from the International Tobacco Control (ITC) Four Country Survey. Tob Control 2006 Jun;15(Suppl 3):iii71-iii75. [doi: 10.1136/tc.2004.008763] [Medline: 16754950]
21. Gwaltney CJ, Metrik J, Kahler CW, Shiffman S. Self-efficacy and smoking cessation: a meta-analysis. Psychol Addict Behav 2009 Mar;23(1):56-66. [doi: 10.1037/a0013529] [Medline: 19290690]
22. John U, Meyer C, Rumpf HJ, Hapke U. Self-efficacy to refrain from smoking predicted by major depression and nicotine dependence. Addict Behav 2004 Jul;29(5):857-866. [doi: 10.1016/j.addbeh.2004.02.053] [Medline: 15219330]
23. Jalli VV, Sreepradha C, Halemani S, et al. Impact of nicotine dependence on self-efficacy and readiness to quit among tobacco consumers in rural areas. J Family Med Prim Care 2024 Mar;13(3):1079-1084. [doi: 10.4103/jfmpc.jfmpc_1624_23] [Medline: 38736831]
24. Noda Y, So R, Sonoda M, Tabuchi T, Nomura A. The usefulness of a smartphone app-based smoking cessation program for conventional cigarette users, heated tobacco product users, and dual users: retrospective study. J Med Internet Res 2023 Mar 17;25:e42776. [doi: 10.2196/42776] [Medline: 36930197]
25. Nomura A, Ikeda T, Fujimoto T, et al. Outcomes of a telemedicine smoking cessation programme for heated tobacco product users in Japan: a retrospective cohort study. BMJ Open 2022 Dec 20;12(12):e063489. [doi: 10.1136/bmjopen-2022-063489] [Medline: 36600419]
26. Sutanto E, Miller C, Smith DM, et al. Concurrent daily and non-daily use of heated tobacco products with combustible cigarettes: findings from the 2018 ITC Japan Survey. Int J Environ Res Public Health 2020 Mar 22;17(6):2098. [doi: 10.3390/ijerph17062098] [Medline: 32235702]
27. Heated tobacco product: information sheet. : World Health Organization; 2020 URL: https://iris.who.int/server/api/core/bitstreams/140f2987-2562-4432-b46d-7adea62880a6/content [accessed 2025-09-02]
28. Sugiyama T, Tabuchi T. Use of multiple tobacco and tobacco-like products including heated tobacco and e-cigarettes in Japan: a cross-sectional assessment of the 2017 JASTIS study. Int J Environ Res Public Health 2020 Mar 24;17(6):2161. [doi: 10.3390/ijerph17062161] [Medline: 32213924]
29. Ratajczak A, Jankowski P, Strus P, Feleszko W. Heat not burn tobacco product—a new global trend: impact of heat-not-burn tobacco products on public health, a systematic review. Int J Environ Res Public Health 2020 Jan 8;17(2):409. [doi: 10.3390/ijerph17020409] [Medline: 31936252]
30. Bialous SA, Glantz SA. Heated tobacco products: another tobacco industry global strategy to slow progress in tobacco control. Tob Control 2018 Nov;27(Suppl 1):s111-s117. [doi: 10.1136/tobaccocontrol-2018-054340] [Medline: 30209207]
31. Uguna CN, Snape CE. Should IQOS emissions be considered as smoke and harmful to health? A review of the chemical evidence. ACS Omega 2022 Jul 5;7(26):22111-22124. [doi: 10.1021/acsomega.2c01527] [Medline: 35811880]
32. Simonavicius E, McNeill A, Shahab L, Brose LS. Heat-not-burn tobacco products: a systematic literature review. Tob Control 2019 Sep;28(5):582-594. [doi: 10.1136/tobaccocontrol-2018-054419] [Medline: 30181382]
33. Yoshihara S, Takahashi K, Kawaguchi H, Harai N, Tsuchiya K. Effects of using a digital peer-supported app on glycated hemoglobin changes among patients with type 2 diabetes: prospective single-arm pilot study. JMIR Form Res 2025 May 20;9:e72659. [doi: 10.2196/72659] [Medline: 40393047]
34. Tabira K, Oguma Y, Yoshihara S, et al. Digital peer-supported app intervention to promote physical activity among community-dwelling older adults: nonrandomized controlled trial. JMIR Aging 2024 May 30;7:e56184. [doi: 10.2196/56184] [Medline: 38814686]
35. Tabira K, Oguma Y, Yoshihara S, et al. A 12-month digital peer-supported app intervention to promote physical activity among community-dwelling older adults: follow-up study of a nonrandomized controlled trial. JMIR Aging 2025 May 28;8:e66610. [doi: 10.2196/66610] [Medline: 40435519]

XSL•FO
RenderX

36. Okamoto M, Saito Y, Nakamura S, et al. Smartphone-based digital peer support for a walking intervention among public officers in Kanagawa Prefecture: single-arm pre- and postintervention evaluation. JMIR Form Res 2024 Sep 24;8:e53759. [doi: 10.2196/53759] [Medline: 39316793]
37. Hu H, Nakagawa T, Honda T, et al. Heated tobacco products and circulating high-density lipoprotein cholesterol concentrations. Sci Rep 2022;12(1):17385. [doi: 10.1038/s41598-022-22337-3]
38. Lau YK, Okawa S, Meza R, Katanoda K, Tabuchi T. Nicotine dependence of cigarette and heated tobacco users in Japan, 2019: a cross-sectional analysis of the JASTIS Study. Tob Control 2022 Aug;31(e1):e50-e56. [doi: 10.1136/tobaccocontrol-2020-056237] [Medline: 33741741]
39. Nakama C, Tabuchi T. Use of heated tobacco products by people with chronic diseases: the 2019 JASTIS study. PLOS One 2021;16(11):e0260154. [doi: 10.1371/journal.pone.0260154] [Medline: 34793517]
40. Muscat JE, Stellman SD, Caraballo RS, Richie JP. Time to first cigarette after waking predicts cotinine levels. Cancer Epidemiol Biomarkers Prev 2009 Dec;18(12):3415-3420. [doi: 10.1158/1055-9965.EPI-09-0737] [Medline: 19959690]
41. Transdisciplinary Tobacco Use Research Center (TTURC) Tobacco Dependence, Baker TB, Piper ME, McCarthy DE, et al. Time to first cigarette in the morning as an index of ability to quit smoking: Implications for nicotine dependence. Nicotine Tob Res 2007 Nov;9(Suppl 4):S555-S570. [doi: 10.1080/14622200701673480] [Medline: 18067032]
42. JCS Joint Working Group; Japanese Society for Oral Health; Japanese Society of Oral and Maxillofacial Surgeons; Japanese Society of Public Health; Japanese Respiratory Society; Japan Society of Obstetrics and Gynecology; Japanese Circulation Society; Japan Pediatric Society; Japanese College of Cardiology; Japan Lung Cancer Society. Guidelines for smoking cessation (JCS 2010)—digest version. Circ J 2012;76(4):1024-1043. [doi: 10.1253/circj.CJ-88-0021]
43. Papadakis S, Tulloch HE, Gharib M, Pipe AL. Profile of tobacco users identified in primary care practice and predictors of readiness to quit: a cross-sectional survey. CMAJ Open 2016;4(1):E41-E47. [doi: 10.9778/cmajo.20150055] [Medline: 27280113]
44. Hoeppner BB, Kahler CW, Gwaltney CJ. Relationship between momentary affect states and self-efficacy in adolescent smokers. Health Psychol 2014 Dec;33(12):1507-1517. [doi: 10.1037/hea0000075] [Medline: 25020151]
45. Boudreaux ED, Sullivan A, Abar B, Bernstein SL, Ginde AA, Camargo CA. Motivation rulers for smoking cessation: a prospective observational examination of construct and predictive validity. Addict Sci Clin Pract 2012 Jun 8;7(1):8. [doi: 10.1186/1940-0640-7-8] [Medline: 23186265]
46. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
47. Lin Y, Tudor-Sfetea C, Siddiqui S, Sherwani Y, Ahmed M, Eisingerich AB. Effective behavioral changes through a digital mHealth app: exploring the impact of hedonic well-being, psychological empowerment and inspiration. JMIR Mhealth Uhealth 2018 Jun 15;6(6):e10024. [doi: 10.2196/10024] [Medline: 29907557]
48. Duan Z, Wysota CN, Romm KF, et al. Correlates of perceptions, use, and intention to use heated tobacco products among US young adults in 2020. Nicotine Tob Res 2022 Nov 12;24(12):1968-1977. [doi: 10.1093/ntr/ntac185] [Medline: 35901840]
49. Hair EC, Bennett M, Sheen E, et al. Examining perceptions about IQOS heated tobacco product: consumer studies in Japan and Switzerland. Tob Control 2018 Nov;27(Suppl 1):s70-s73. [doi: 10.1136/tobaccocontrol-2018-054322] [Medline: 29764957]
50. Inagaki S, Kato K, Matsuda T, et al. Experience with a team-based gamification health app for behavior change adapted to people with diabetes: a pilot study. Technol Health Care 2025 Sep;33(5):2220-2231. [doi: 10.1177/09287329251332454] [Medline: 40302556]
51. SRNT Subcommittee on Biochemical Verification. Biochemical verification of tobacco use and cessation. Nicotine Tob Res 2002 May 1;4(2):149-159. [doi: 10.1080/14622200210123581]

Abbreviations

APA: American Psychological Association
HTP: heated tobacco product
JARS-Quant: Journal Article Reporting Standards for Quantitative Research
OR: odds ratio

XSL•FO
RenderX

Edited by S Brini; submitted 25.Sep.2025; peer-reviewed by A Eisingerich, K Perkins; revised version received 13.Apr.2026; accepted 14.Apr.2026; published 19.May.2026.

Please cite as:
Yoshihara S, Takahashi K, Uemura C, Murakami S, Amano S, Harada D, Jiang Y, Yamato H
Nicotine Dependence and Quit Self-Confidence in a Smoking Cessation Program Using a Group-Based Digital Peer-Supported App and Cigarette Consumption–Adjusted Nicotine Aids Among Japanese Workers: Retrospective Cohort Study
J Med Internet Res 2026;28:e84792
URL: https://www.jmir.org/2026/1/e84792
doi:10.2196/84792

© Shota Yoshihara, Kayoko Takahashi, Chiaki Uemura, Shin Murakami, Satoru Amano, Daichi Harada, Ying Jiang, Hiroshi Yamato. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Digital Toolkit for Weight Loss Maintenance in European Adults (NoHoW): 2×2 Factorial Randomized Controlled Trial

R James Stubbs[1], PhD; Cristiana Duarte[2], PhD; Clarissa Dakin[1], PhD; António L Palmeira[3], PhD; Falko F Sniehotta[4], PhD; Graham W Horgan[5], PhD; Sofus C Larsen[6], PhD; Marta M Marques[7], PhD; Jorge Encantado[8], PhD; Elizabeth H Evans[9], PhD; Jake Turicchi[1], PhD; Ruari O'Driscoll[1], PhD; Sarah E Scott[1], PhD; Beth Pearson[1], PhD; Lauren Ramsey[1], PhD; Marie-Louise Mikkelsen[5], PhD; Inês Santos[3,10,11], PhD; Marcela Matos[12], PhD; Pedro J Teixeira[8], PhD; Berit L Heitmann[6,13], PhD

[1]School of Psychology, Faculty of Medicine and Health, University of Leeds, University Rd, Woodhouse, Leeds, United Kingdom

[10]Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal

[11]Instituto de Saúde Ambiental (ISAMB), Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal

[12]Center for Research in Neuropsychology and Cognitive and Behavioural Intervention (CINEICC), Universidade de Coimbra, Coimbra, Portugal

[13]Section for General Medicine, The Department of Public Health, University of Copenhagen, Copenhagen, Denmark

[2]School of Education, Language and Psychology, York St John University, York, United Kingdom

[3]Centro de Investigação em Desporto Educação Física, Exercício e Saúde (CIDEFES), Universidade Lusófona and Centro de Investigação, Formação, Inovação e Intervenção em Desporto (CIFI2D), Universidade do Porto, Porto, Portugal

[4]Division of Public Health, Social and Preventive Medicine, Centre for Preventive Medicine and Digital Health (CPD), Medical Faculty Mannheim, University of Heidelberg, Mannheim, Germany

[5]Biomathematics and Statistics Scotland, Aberdeen, United Kingdom

[6]Research Unit for Diet and Health at The Parker Institute, Bispebjerg and Frederiksberg Hospital, The Capital Region, Copenhagen, Denmark

[7]NOVA National School of Public Health, Public Health Research Centre, Comprehensive Health Research Center, CHRC, REAL, CCAL, NOVA University Lisbon, Lisbon, Portugal

[8]Interdisciplinary Center for the Study of Human Performance (CIPER), Faculdade de Motricidade Humana, Universidade de Lisboa, Cruz Quebrada, Portugal

[9]Department of Psychology, Faculty of Science, Durham University, Durham, United Kingdom

Corresponding Author:
R James Stubbs, PhD
School of Psychology, Faculty of Medicine and Health, University of Leeds, University Rd, Woodhouse, Leeds, United Kingdom

Abstract

Background: Digital approaches to weight management have the potential to produce cost-effective and scalable weight management solutions. Effective behavior change interventions typically promote self-regulation of energy balance behaviors, which may be enhanced by incorporating emotion regulation strategies.

Objective: This study aimed to evaluate the effectiveness of a digital behavior change toolkit for weight loss maintenance in European adults who had achieved ≥5% weight loss in the previous 12 months. We hypothesized that a combined intervention targeting self-regulation or motivation and emotion regulation would be more effective than either component alone, and that each would outperform an active control.

Methods: The Navigating to a Healthier Weight (NoHoW) trial was a 2×2 factorial randomized, single-blind, controlled trial involving 1627 adults who had achieved ≥5% weight loss in the previous 12 months (initial BMI ≥25 kg/m^2) across 3 European centers (the United Kingdom, Denmark, and Portugal). The trial evaluated a digital toolkit for weight management subsequent to an initial ≥5% weight loss in the prior 12 months. Participants were assigned using adaptive stratified sampling to one of four groups: (1) self-regulation or motivation (n=403), (2) emotion regulation (n=416), (3) combined motivation and emotion (n=408), or (4) active control (generic content, regular self-weighing, and Fitbit use, n=400). The primary outcome was weight change from baseline to 12 months. Prespecified secondary outcomes included cardiometabolic markers. Linear models adjusted for recruitment center, sex, age group, BMI group, and pretrial weight loss. Subgroup analyses were conducted by sex.

Results: At 12 months, 76% (364/1627) of participants remained in the study. In the primary ITT analysis in all participants, none of the intervention arms (motivation, emotion, or combined) differed significantly from the active control for weight change at 12 months. Completer and per-protocol analyses produced similar patterns and did not change the overall interpretation. In the per-protocol sample, men regained 0.14 kg, and women regained 0.54 kg of their pretrial weight loss. Subgroup analyses indicated a small effect of the motivation intervention in men, but this was not clinically meaningful and did not alter the primary null

XSL•FO
RenderX

findings. Nearly half of ITT participants regained weight, and no significant intervention effects were observed for cardiometabolic secondary outcomes.

Conclusions: The NoHoW trial was the first large-scale, multicountry 2×2 factorial randomized controlled trial to evaluate a digital-only toolkit based on self-regulation or motivation and emotion regulation techniques for weight loss maintenance. NoHoW found no evidence in the primary ITT analysis that digital interventions targeting self-regulation or emotion regulation improved weight loss maintenance compared with the active control. A small subgroup effect in men should be interpreted cautiously and does not change this conclusion. The trial provides evidence on both the limitations and potential of digital behavior change interventions for long-term weight outcomes. Future digital interventions may benefit from enhanced engagement and tailored content to improve long-term weight outcomes.

Trial Registration: ISRCTN Registry ISRCTN88405328; https://www.isrctn.com/ISRCTN88405328

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2019-029425

(J Med Internet Res 2026;28:e69634) doi:10.2196/69634

KEYWORDS

emotion regulation; information and communication technologies; motivation; obesity; self-regulation; weight loss maintenance

Introduction

Background

The societal challenges of overweight and obesity affect more than half of the adult population [1]. The health care consequences and associated economic costs are well documented [2,3]. More than 40% of Western adults report making at least 1 weight control attempt per year [4], most of which do not involve evidence-based behavior change approaches [4]. Existing community-based programs support initial weight loss but are subject to high attrition and weight regain, limiting longer-term effectiveness [5-7]. The obesogenic environment and asymmetry of human energy balance regulation facilitate weight gain, while society stigmatizes people experiencing overweight and obesity, leading to stress and negative emotions, which can undermine weight loss attempts [8-10]. The majority of weight loss attempts are followed by some degree of weight regain [11], emphasizing the need to develop more effective weight loss maintenance solutions. Long-term weight loss maintenance remains a significant challenge, with weight regain highly prevalent. These recurring patterns emphasize the classification of obesity as a chronic, relapsing disease [12,13]. Effective support for weight loss maintenance needs to address sustained changes in energy balance behaviors (diet, physical activity, and weight control) and emotional or psychosocial challenges that may undermine longer-term self-management of those behaviors [14]. Pathways of planned behavior change involving self-management of energy balance behaviors are difficult to implement unless they become practiced and habitual. Dual process models of behavior change suggest that such practice takes time. Existing habits, preferences, urges, desires, and responsiveness to environmental cues often undermine the implementation of new behavior changes [15]. Models of behavior change increasingly include both reflective and reactive elements, which, if not aligned, can deplete psychological resources and undermine maintenance of behavior change [16]. Greaves et al [14] describe longer-term weight management as generating a tension between new (weight management) behaviors and the fulfillment of preexisting psychological needs from existing patterns of behavior. This tension could be managed through self-regulation, renewed motivation, and managing external influences to change habits, finding nonobesogenic approaches to meet psychological needs, and changing self-concept [14]. Some of the factors that undermine longer-term weight loss, such as changes in energy balance physiology affecting energy expenditure, food reward-based processes, or energy intake, may be outside of conscious recognition and control [17]. Aspects of self-regulation and motivation may improve the odds of changing energy balance behaviors, and if those changes become habitual, potentially help prevent weight regain [18-20]. Changing habits can take 2 - 5 years [21]. However, reactive processes (emotions, desires, impulses resulting from associative learning, and physiological resistance to weight loss) are powerful forces that can also undermine relatively transient and fragile attempts at changing energy balance behaviors during weight loss.

Self-Regulation and Motivation

Research identifying and linking specific behavior change approaches to mechanisms of action of behavior change interventions is still a developing field [22,23]. Core features of effective weight loss maintenance interventions include (1) self-monitoring of weight and behavior, (2) goal-setting: agreement of clear weight targets or trigger points for weight control efforts, (3) feedback on behavior and weight, (4) action plans for weight control through dietary and physical activity behaviors, and (5) plans to cope with risk factors for weight regain and relapse prevention (eg, problem-solving) [18-20,24,25]. Many people also experience behavioral lapses and relapses as more pronounced situational or momentary events. Avoiding weight regain requires behavioral strategies, in which relapse coping and weight loss maintenance become learned skills of self-regulation, autonomy, and motivation as part of a longer-term process [12,26-28].

Physical activity and dietary weight loss maintenance interventions based on current behavior change theories characteristically achieve relatively modest effects of ~1.6 kg difference in weight regain compared to unsupported controls over 12 months [18,29,30]. Additional psychological processes, such as emotion regulation, potentially have additional beneficial

XSL•FO
RenderX

effects on the behavioral changes that promote weight loss maintenance [14].

Emotion Regulation

Reactive processes (emotions, desires, habits resulting from associative learning, and physiological states) may have a large impact on behavior change. These relatively rapid, automatic, impulsive (less conscious), and habitual processes (compared to the slow, deliberative processes of motivation and self-regulation) [15,31] may undermine initial self-regulation of energy balance (particularly eating) behaviors in the face of a physiological system that resists longer-term weight loss [32]. Automatic components of self-regulation may also promote longer-term behavior change if they are engaged and developed [33,34]. People with overweight and obesity commonly experience stigma, which enhances psychosocial stress and impacts physical and mental well-being [35-37]. Stigma operates via shame, self-criticism, and unfavorable social comparisons, creating feelings of inferiority and inadequacy in relation to others [38]. The relationship between stress, emotion, and food intake can derail strategies of planned behavior and promote further weight gain [39]. Limited evidence suggests that acceptance, self-compassion, and mindfulness-based approaches could potentially help some people in coping with obesity-related eating behaviors [40,41]. We hypothesized that self-monitoring, self-regulation, and autonomous motivation in weight loss maintenance could be supported favorably by strategies that promote emotion regulation and that may thereby increase coping with stress.

Digital Weight Loss Maintenance Interventions

Evidence of the effectiveness of long-term digital interventions and strategies to support weight loss maintenance is limited [42]. Digital solutions, such as smartphone apps and wearables, have the potential to be effective in supporting weight loss maintenance if they are evidence-based and offer a choice of behavior change techniques to encourage the use of self-regulatory techniques (eg, self-monitoring). Interventions with human contact are more effective than those that are fully automated [43]. Digital solutions are potentially cost-effective and scalable to large populations, which could engage citizens in health care innovations that are convenient and effective for weight management in the face of limited public budgets [44]. Recent systematic reviews indicate that weight management apps may have positive effects on weight-related outcomes, although the methodological quality of many studies is low and effect sizes are modest [45-47]. It is important to conduct randomized trials of digital technologies for weight loss maintenance and to try to understand the mechanisms by which they may influence weight and health outcomes.

Therefore, effective interventions and commercial programs for weight loss are widely available, but most people regain their lost weight. Currently, few comprehensive solutions exist to help European people manage weight loss maintenance. Current research suggests that the most promising evidence-based behavior change techniques for weight loss maintenance are self-monitoring, goal-setting, action control, building self-efficacy, and intrinsic motivation. Recent research also suggests that stress management and emotion regulation skills are key enablers of relapse prevention and weight regain, but these components have not been formally evaluated through a prospective controlled trial. The Navigating to a Healthier Weight (NoHoW) project tested whether digital delivery of the most promising evidence-based behavior change techniques is effective for weight loss maintenance in a large-scale international 3-center trial of digital tools that implement the most up-to-date behavioral science research. The trial aimed to establish the effectiveness of these tools in supporting weight loss maintenance.

Study Objectives and Hypotheses

The study's primary objectives were to evaluate whether using a new digital toolkit was effective for weight loss maintenance by improving (1) self-regulation and motivation, (2) contextual behavioral aspects of emotion regulation, or (3) these factors in combination (motivation and emotion), compared to an active control (generic toolkit content, regular self-weighing, and wearable tracking device [Fitbit] use) in 1627 participants total, at 3 European research centers. The 4-month digital toolkit targeted weight, physical activity, and dietary behaviors through 34 sessions delivered across 15 modules. We hypothesized that the combined intervention would be more effective for weight loss maintenance compared to the self-regulatory or emotion regulatory interventions alone, which individually would also be more effective than the active control group.

Secondary objectives were to determine how the intervention affected health markers (eg, levels of hemoglobin A_{1c} [HbA_{1c}]), cholesterol and cortisol levels, and body composition. The study also examined (1) the intervention impact on physical activity, sleep, self-reported dietary intake, depression, anxiety, stress, quality of life, and well-being; (2) potential mediators of weight loss maintenance, such as self-regulation (eg, planning capacity), motivation (eg, autonomous motivation), and emotion regulation processes (eg, self-compassion); (3) quantitative and qualitative assessment of user experience, acceptability, engagement, and dropout; and (4) intervention cost-effectiveness. Secondary outcome analyses (2-4) are being reported in separate publications [48,49].

Methods

Study Design

The NoHoW trial was a 3-center (University of Leeds [the United Kingdom], Frederiksberg and Bispebjerg University Hospital [Denmark], and University of Lisbon [Portugal]) 2×2 factorial, randomized, single-blind, controlled trial testing the proof-of-concept of a digital toolkit for weight loss maintenance. A complete description of the trial design can be found elsewhere [50]. In brief, a total of 1627 adults who had achieved ≥5% weight loss in the previous 12 months (initial BMI ≥25 kg/m^2) were randomized using a 2×2 factorial, single-blind design into four arms: (1) active control (self-monitoring only, including Fitbit activity tracker and smart scale with generic content), (2) self-regulation and motivation (modules targeting goal-setting, action planning, and autonomous motivation), (3) emotion regulation (contextual behavioral strategies such as mindfulness, acceptance, and self-compassion), and (4)

XSL•FO
RenderX

combined self-regulation or motivation and emotion regulation. All participants received access to a mobile-enabled digital toolkit integrated with Fitbit Charge 2 and Aria scales. The active intervention lasted 18 weeks during the first 6 months, followed by discretionary use for 12 months. Participants attended clinical investigation days (CIDs) at baseline, 6, 12, and 18 months. The primary outcome was change in weight (kilograms) at 12 months from baseline, with secondary outcomes including body composition, biomarkers (HbA_{1c}, lipids, blood pressure, and hair cortisol), physical activity, sleep, dietary intake, psychological mediators, engagement, and cost-effectiveness. Analyses adhered to the prespecified protocol, using intention-to-treat (ITT) principles with multiple imputation for missing data and linear regression models for continuous outcomes.

The toolkit included a set of theoretically informed, evidence-based web-located tools and inputs from digital tracking devices (smart scales and activity trackers) and modules that targeted weight, physical activity, and dietary behaviors. The final toolkit comprised 34 sessions that were distributed through 15 modules and provided active content over a 4-month (18-week) period. The motivation and self-regulation arm consisted of 8 theory-driven modules (17 sessions) targeting self-regulatory behavior change techniques, the emotion regulation arm was presented with 7 theory-driven modules (17 sessions) targeting techniques related to contextual emotion regulation, and the combined arm received the full toolkit (15 modules; 34 sessions). The sessions included a range of implementations, such as videos, testimonies, and questionnaires. Furthermore, the toolkit contained 5 specific data tiles for monitoring weight, steps, healthy eating, mood, and sleep. Full details of the toolkit, its design, and content are described by Marques et al [51]. The toolkit included a mobile-enabled website and tracking technologies: an activity and sleep tracker (Fitbit Charge 2) and a smart wireless body weight scale (Fitbit Aria). The toolkit development followed the Medical Research Council guidance for complex interventions [52], with specific intervention logic models and theory-driven behavior change techniques [51]. The study was conducted between March 2017 and September 2019 in the 3 academic research institutions. The study duration was 18 months with a follow-up at 6, 12 (primary end point, reported in this study), and 18 months after the baseline [50]. Changes in the primary outcome (weight in kilograms) and secondary outcomes (health biomarkers, physical activity, sleep and dietary intake, and self-reported psychological processes) are reported at 12 months from baseline.

In total, 1627 participants (536 - 555 per center) were enrolled and randomized into 1 of the 4 intervention arms in a factorial structure. All arms included self-weighing and activity trackers: (1) active control arm (consisting of generic toolkit content), (2) self-regulation and motivation arm, (3) emotion regulation arm, and (4) combined self-regulation or motivation and emotion regulation (motivation and emotion) arm (Figure 1). Participants were allocated to 1 of the 4 treatment arms by adaptive stratified sampling using minimization, where differences in age, weight lost, and current BMI were minimized between treatment arms, which were also balanced for sex. An algorithm was composed and implemented on the trial management website. Staff recruiting participants accessed it there after each individual had been recruited. They were unable to know or predict what allocation the algorithm would select. The nature of the intervention meant that they could not be blinded. The protocol was harmonized across trial centers using Good Clinical Practice guidance, research-grade translation or backtranslation of trial materials, 2 training workshops, and weekly trial management meetings. The final protocol (V2.1 20/09/2017) was approved by the Trial Steering Committee and adhered to the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) guidelines [53]. This study was reported using the CONSORT (Consolidated Standards of Reporting Trials) 2025 statement (Checklist 1) [54].

XSL•FO
RenderX

Figure 1. The 2×2 factorial design of the NoHoW trial. CID: clinical investigation day; NoHoW: Navigating to a Healthier Weight; RCT: randomized controlled trial.

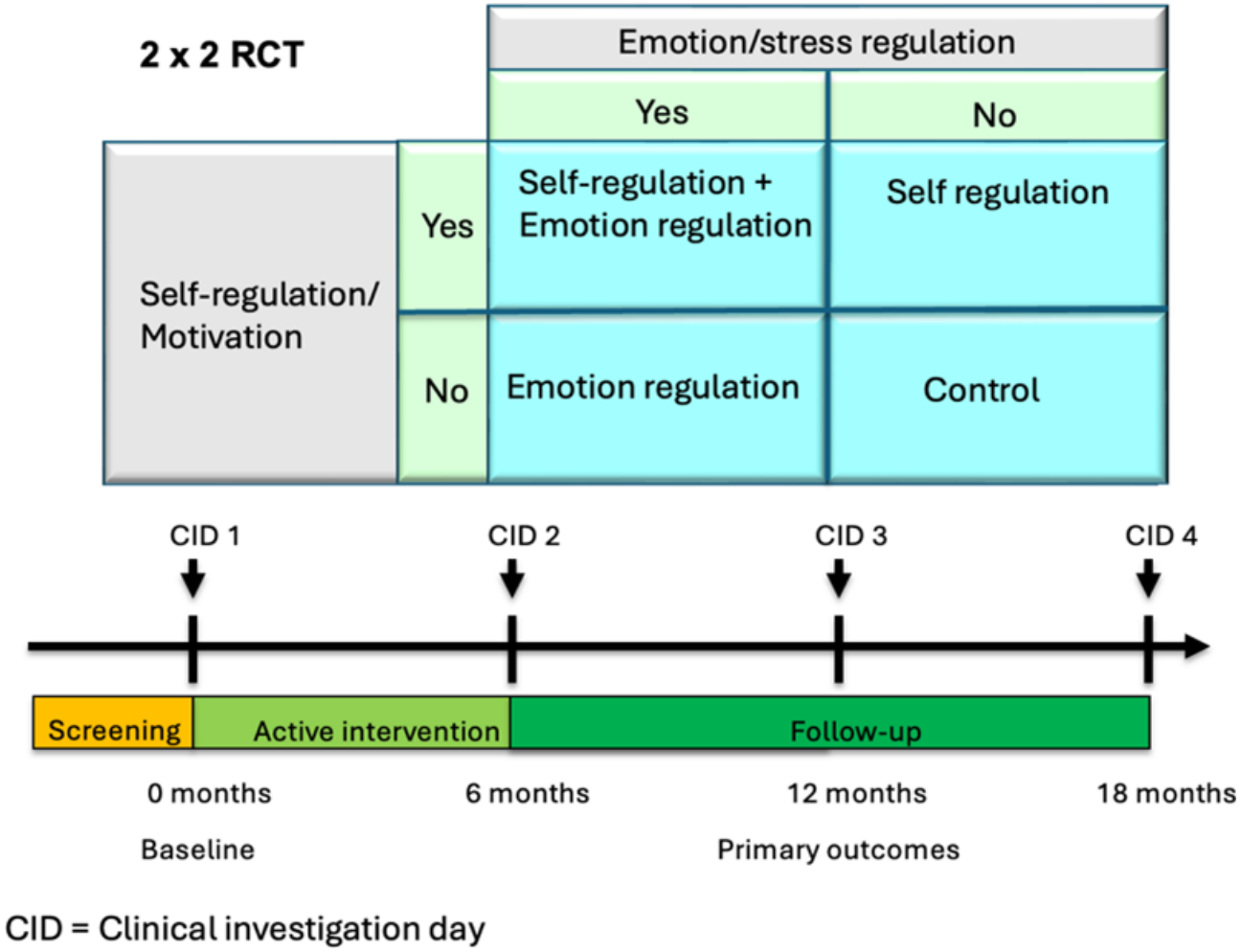

Ethical Considerations

Ethics approval (ISRCTN88405328) was granted by institutional ethics committees at the Universities of Leeds (17 - 0082; February 27, 2017), Lisbon (17/2016; February 20, 2017), and Capital Region of Denmark (H-16030495, March 8, 2017). The protocol was prospectively registered and complied with relevant European Union legislation, international conventions, and declarations relating to ethical research practices 94. All participants were required to sign informed consent. Identifiable information, including name, contact details, and consent form, was obtained as part of the screening process. The identifiable information was pseudonymized via the use of a unique participant study ID. The code link for pseudonymized data is stored for 20 years as a password-protected document accessible only to the research staff. The anonymized data will be kept indefinitely. Participants were compensated for their participation with vouchers (Love to Shop) in the form of a US $20.30 monetary voucher at months 12 and 18. Participants were also allowed to keep their study devices (Fitbit Charge 2 activity tracker and Fitbit Aria scales).

Study Population

The trial examined weight loss maintenance in those who had lost ≥5% of their weight during the previous 12 months. Participants were adults (≥18 years) with a pre–weight-loss BMI ≥25 kg/m^2, who had achieved and maintained ≥5% weight loss within the past 12 months. Eligibility required access to a smartphone or computer with internet, ability to use standing scales (≤150 kg), and provision of informed consent. Exclusion criteria included weight loss due to illness or surgery, pregnancy or breastfeeding, eating disorders, certain unstable medical conditions, recent type 1 diabetes diagnosis, inability to communicate in English, Danish, or Portuguese (at respective centers), participation in conflicting studies, extensive planned travel, or cohabitation with another trial participant. Prior to randomization, potential participants were asked to provide documented verification (by a health professional, weight loss counselor or friend, weight loss program record booklet, diary or smartphone app, or before or after photographs) that they had achieved a clinically significant weight loss of >5% during the previous 12 months. The complete inclusion or exclusion criteria for the trial, recruitment, interventions, assessment, and study procedures are described in a previously published protocol paper [50]. There were no deviations from the main details stated in the protocol. There were minor deviations for a few participants (eg, hair not being cut correctly), which were all recorded and not considered to have any significant impact on the results.

Center-specific recruitment strategies were adopted for 12 months (March 2017-March 2018) and included commercial

XSL•FO
RenderX

weight loss programs (the United Kingdom, Slimming World), the Copenhagen Municipality weight management services, dieticians from the Danish Association for Dieticians and commercial slimming companies, registered clinical dieticians or nutritionists who provide weight management services in Lisbon, leisure centers, and local or national media coverage and advertisements. All participants were directed to country-specific recruitment websites and completed an online eligibility screener using Qualtrics. The eligibility screener categorized respondents as eligible, potentially eligible (waitlist), and ineligible. Eligible individuals were contacted for a telephone screening interview, provided with study information, medical history questions, and the Physical Activity Readiness Questionnaire [55]. Eligible participants were invited to a CID, where informed consent was obtained by research staff before randomization (Multimedia Appendix 1).

Each participant attended 4 CIDs at baseline, 6, 12, and 18 months. The purpose of these visits was to conduct standardized assessments and ensure protocol adherence. At the baseline CID, informed consent was obtained, eligibility confirmed, and participants were randomized and trained on the use of the digital toolkit and Fitbit devices. Subsequent CIDs included reconsent, collection of primary and secondary outcome measures (eg, weight, body composition, blood pressure, biomarkers, and psychometric questionnaires), and troubleshooting any technical issues. Participants were reconsented before each subsequent visit and excluded if they became ineligible.

Measurements

A full description of trial measures and associated references (including mediators and moderators of any trial effects) is described in a previous publication [50]. Prespecified cardiometabolic secondary outcomes (waist circumference, fat mass, fat-free mass, HbA_{1c}, blood lipids, blood pressure, resting heart rate, and hair cortisol) are analyzed and reported in this paper. All other secondary outcomes (physical activity, sleep, dietary intake, eating behavior, well-being, and psychosocial variables) results are available in Tables S3 and S4 in Multimedia Appendix 1. All measures were taken at the intervention centers by the trial team. The research team was not blinded to the assignment.

Data were collected in a variety of ways and stored centrally in a data hub. Measurements on each CID were entered in commercial trial management software (Easytrial; EasyTrial ApS) and subsequently uploaded to the data hub. User engagement, activity, and comments were recorded by the toolkit and stored on the data hub. Dietary intakes were recorded using Intake24 [56] and periodically downloaded from their servers and uploaded to the data hub. A similar process was followed with questionnaire data, which were recorded online by Qualtrics. Wearable devices and weighing scales streamed data to Fitbit servers, and an application programming interface running on the data hub pulled data once every day from the Fitbit data storage server and added it to tables in the database. These tables were accessible to trial staff who could thereby monitor compliance.

Primary Outcome

Body weight (±0.1 kg) was measured using a Seca 704s instrument (SECA) in participants wearing light clothing. Participants were asked to self-weigh themselves at least twice weekly, in the morning after voiding and before eating using the provided Fitbit Aria scales for 18 months.

Secondary Outcomes

Body Composition and Height

Multifrequency whole bioimpedance spectroscopy was measured by ImpediMed SFB7, which measures impedance over a spectrum of frequencies for the estimation of body composition. Hanai mixture theory equations and standard resistivity constants [57] were used to calculate fluid volumes to estimate total body water and hence fat and fat-free mass. Height (±0.1 cm) was measured with participants barefoot, using a Seca 704s instrument (SECA).

Waist and Hip Measurements

Waist and hip measurements were taken according to the World Health Organization's guidance. A tape measure was used to record the hip and waist circumference to the nearest 0.1 cm [58].

Biomarkers

HbA_{1c} and Cholesterol

Fasted capillary blood samples were collected to determine HbA_{1c} (mmol/mol, %), estimate average glucose (mol/L), and calculate full lipid profiles, including total cholesterol, low-density lipoprotein, high-density lipoprotein (HDL), triglycerides, non-HDL, and cholesterol or HDL (mmol/L) assayed using a bench-top analyzer (Alere Afinion AS100 Analyzer) [59].

Systolic and Diastolic Blood Pressure and Resting Heart Rate

Systolic and diastolic blood pressure and resting heart rate were measured with the participant at rest in the sitting position (Microlife BP A2 Basic, Gentle Technology). Values were taken as the average of 3 measurements.

Cortisol

Hair cortisol was included as a measure of chronic stress using hair samples following a previously described protocol [60]. Hair samples were cut from the posterior vertex as close to the scalp as possible. Between 10 and 30 mg of the hair from 2 cm closest to the scalp was cut into small pieces and dissolved in 1 mL methanol and incubated at ultrasound sonication for 30 minutes, followed by 18 hours at 52 °C in a shaking incubator (300 rpm). Hair samples were not collected among participants with less than 2 cm of hair. The methanol was transferred to a new tube and evaporated to dryness under a stream of nitrogen at 45 °C. Dried samples were stored at −20 °C until analysis. Before analysis, samples were redissolved in 500 µL phosphate-buffered saline at pH 8 and centrifuged at 2000 rpm for 2 minutes. Reconstituted samples were analyzed with a cortisol enzyme-linked immunosorbent assay (Alpco).

XSL•FO
RenderX

Physical Activity

Minute-by-minute physical activity data and heart rate were measured by the Fitbit Charge 2 for the study duration. Data were used from 2 weeks after randomization for baseline and for the 8 weeks prior to the 12-month point for the trial outcome. Physical activity was also self-reported at 0, 6, 12, and 18 months (8-week intervals) using the International Physical Activity Questionnaire [61] and the Activity Choice Index [62].

Sleep Quality and Quantity

The Fitbit Charge 2 estimates sleep quantity (hours/minutes) and quality (stages of sleep) using proprietary algorithms. Data were collected daily throughout the trial.

Dietary Intake and Eating Behavior

Four consecutive 24-hour web-based dietary recalls, including at least 1 weekend day, were collected within 7 days of each CID visit using Intake24 [56], and the average daily intake of energy and macronutrients was calculated. Intake24 is an open-source self-completed computerized dietary recall system based on a multiple-pass 24-hour recall. The online system offers similar data quality to interviewer-led recalls at a significantly lower cost. Dietary information was collected at baseline and the 12-month follow-up using a 4-day, 24-hour dietary recall via the online platform Intake24, which included at least 1 weekend day [50,56]. Not all participants completed all four 24-hour recalls [63]. The recorded weights of foods and beverages were converted into energy and nutrient intakes for each item and then summed to yield total 24-hour energy and nutrient intakes for each participant on each recall day. Daily macronutrient intakes were estimated using a country-specific food composition database [50]. The primary exposures were the proportions of total energy intake derived from protein, fat, and carbohydrate at baseline, analyzed as continuous variables (percentage of total energy). Questionnaires were completed by the participants at home using Qualtrics. The following measures were collected at the time of each CID at 0, 6, 12, and 18 months:

- Three-Factor Eating Questionnaire-51: The Three-Factor Eating Questionnaire includes 51 items measuring eating disinhibition, susceptibility to hunger, and dietary restraint [64].
- Binge Eating Scale [65]: This includes 16 items and assesses the behavioral, cognitive, and emotional dimensions of binge eating symptomatology.
- Intuitive Eating Scale-21: This is a 21-item scale that assesses 3 subscales of intuitive eating: eating for physical rather than emotional reasons, unconditional permission to eat, and reliance on hunger and satiety cues [66].
- Well-Being and Quality of Life (EQ-5D-5L): EQ-5D-5L is a measure of generic health-related quality of life including mobility, self-care, usual activities, pain or discomfort, and anxiety or depression [67].
- Warwick-Edinburgh Well-Being Scale: This is a 14-item single-factor scale, which assesses mental well-being [68].

Sample Size Calculation

Power calculations were based on the primary outcome (weight change) in previous trials [69] at 12 months and also conducted for the secondary outcome change in HbA_{1c}. To detect a difference between treatment arms of >1.5 kg body weight with a Cohen *d* value of 0.25 for 80% power, comparing more than 2 groups required a sample size of 250 per trial arm. To detect an effect size of 0.25 SD units for HbA_{1c}, 245 participants in each trial arm gave 80% power at 5% significance. Assuming a 38% dropout, a sample size of 1600 (533 per center) was needed to achieve a sample of 1002 (334 per center, ~250 per trial arm) participants at 12 months. These choices were informed by previous trials in which the investigators had been involved [69], and we estimated dropout rates from these, though we note that dropout in NoHoW was less than we predicted. A difference of 1.5 kg was our experienced judgment on the magnitude of difference in effect that would be relevant to health.

Statistical Analysis

Primary Analysis (ITT)

The prespecified primary analysis evaluated intervention effects on 12-month weight change (kilograms) using an ITT framework. For the primary outcome, we also reported the results on 2 other study subsets, the completers (those who had measurements at 12 months) and the per-protocol sample (those who had measurements at 12 months and who also completed at least 80% of the toolkit modules for their assigned trial arm).

Baseline data were summarized by group and sex, and the significance of these factors was assessed by a linear model, which also included terms for center, age group, and pretrial weight loss group. Statistical analyses were carried out using linear models, where the response was the change in outcome value from baseline to 12 months. Explanatory terms were center, sex, motivation intervention, emotion intervention, and 2-way interactions between the terms motivation and emotion. Age group, BMI group, and pretrial weight loss group were included as additional covariates. Significance was assessed using type II sums of squares to maximize power to detect main effects in the absence of significant interactions in an ANOVA of the linear model. Means presented are marginal means estimating the effects in a population, in which other factors are equally balanced between trial arms.

ITT requires imputation of outcome data for those who had dropped out. An assumption of missingness at random did not appear valid (Little [70] test of whether data are missing completely at random gave $P<.001$), and so standard multiple imputation was unsuitable. Carrying forward, the last observation would assume no weight regain in those who drop out, which appeared likely to treat dropouts as having an unreasonably optimistic outcome. We therefore used the most conservative assumption that dropouts had regained all weight lost before the trial, that is, we carried forward the highest weight in the 12 months before CID 1 (baseline observation carried forward). Weight lost before the trial was available, but other outcomes were not. As using CID 1 measures for baseline carried forward would appear optimistic for these outcomes in dropouts, and they were not available before the pretrial weight loss, we imputed values by predicting them from the participant characteristics and their imputed weight regain. Multiple

XSL•FO
RenderX

imputation, using the R library "mice," was used to account for the uncertainty in this prediction. Data analysis was carried out using R (version 3.5; R Foundation for Statistical Computing).

Subgroup Analyses (Prespecified: Men and Women)

Prespecified sex subgroup analyses were conducted by fitting the same models separately in men and women to assess whether intervention effects differed by sex. These analyses are reported after the primary ITT results, and tables and figures are placed in Multimedia Appendix 1.

Sensitivity Analyses (Completers and Per-Protocol)

Two sensitivity analyses were conducted to examine the robustness of the primary ITT findings. This includes a completer (participants with observed 12-month weight) and per-protocol (participants with observed 12-month weight and ≥80% completion of assigned toolkit modules in their randomized arm) analysis. These analyses were prespecified in the published protocol [50].

Exploratory Analyses (Post Hoc: Engagement or Compliance)

Exploratory post hoc regressions assessed associations between engagement metrics and weight change. Predictors included toolkit compliance (percentage of assigned modules completed within the intervention arm), Fitbit compliance (number of days with >20 hours of wear recorded), and self-weighing compliance (number of weeks with ≥2 weighings on the provided smart scale). These exploratory analyses were not part of the prespecified primary analysis.

Results

Participants

The sample consisted of 1117 (68.7%) women and 510 (31.3%) men (Figure 2). In the 12 months prior to randomization, the women lost ~11% (10.8 kg), and the men lost ~12% (10.7 kg) of their weight (P<.001). The mean age was 44.5 (SD 11.9) years, weight was 84.7 (SD 17.2), and BMI was 29.7 (SD 5.5) kg/m^2. At the 12-month follow-up (CID3), 364 of 1627 (22.4%) of participants had missing data.

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) diagram showing participant flow through the NoHoW 2×2 factorial randomized controlled trial of a digital toolkit for weight loss maintenance. NoHoW: Navigating to a Healthier Weight.

XSL•FO
RenderX

At baseline, the majority of participants had a BMI of either 25‑29.9 kg/m^2 (689/1627, 42.3%) or 30‑34.9 kg/m^2 (398/1627, 24.5%). In total, 1594 participants provided data on employment. Most (n=1089, 68.3%) were employed full-time, 8.3% (n=132) were employed part-time, 8.7% (n=139) were retired, 6.3% (n=100) were students, 2.4% (n=38) were unemployed, 0.9% (n=14) were self-employed, and 0.5% (n=8) were caregivers. Most (496/555, 89.4%) UK participants described themselves as White British; 97.2% (521/536) participants from Denmark described themselves as Danish; and 69.78% (374/536) participants from Portugal described themselves as Portuguese. The majority (n=1384, 85.1%) of the sample were educated beyond secondary-level education (eg, degree level). In total, 43% (n=700) of participants were married. The median household income for the Danish and UK samples fell into the same category as their national average (range of €22,400-€53,713 [US $25,847.99-$61,980.95]). For the Portuguese sample, median household income fell below their national average (€12,000‑€18,000 [US $13,847.14-$20,770.71]). Mean baseline values of the study variables for all participants and per center are reported in Table 1.

XSL•FO
RenderX

Table . Baseline characteristics of participants in the NoHoW[a] trial, a 2×2 factorial randomized controlled trial of a digital weight loss maintenance intervention in European adults[b].

		All participants				All	*P* value
		Control	Motivation	Emotion	Motivation + emotion		
Center, n (%)							.98
	Copenhagen	131 (32.8)	137 (34)	133 (32)	135 (33.1)	536 (32.9)	
	Lisbon	133 (33.2)	136 (33.7)	134 (32.2)	133 (32.6)	536 (32.9)	
	Leeds	136 (34)	130 (32.3)	149 (35.8)	140 (34.3)	555 (34.1)	
Sex, n (%)							>.99
	Female	276 (69)	276 (68.5)	285 (68.5)	280 (68.6)	1117 (68.7)	
	Male	124 (31)	127 (31.5)	131 (31.5)	128 (31.4)	510 (31.3)	
Age group (years), mean (SD)							.47
	18 - 29	54 (13.5)	46 (11.4)	58 (13.9)	41 (10)	199 (12.2)	
	30 - 49	219 (54.8)	240 (59.6)	226 (54.3)	233 (57.1)	918 (56.4)	
	50 - 85	127 (31.8)	117 (29)	132 (31.7)	134 (32.8)	510 (31.3)	
Pretrial weight loss (kg), mean (SD)		9.6 (7.95)	10.68 (9.92)	10.07 (8.9)	9.84 (8.12)	10.05 (8.75)	.44
Body composition (kg), mean (SD)							
	Weight (kg)	84.96 (17.46)	84.62 (16.89)	84.18 (17.06)	85.45 (17.66)	84.8 (17.26)	.75
	Fat-free mass (kg)	57.44 (11.51)	57.19 (11.61)	56.85 (11.4)	57.49 (11.31)	57.24 (11.45)	.85
	Fat mass (kg)	27.3 (10.7)	27.27 (10.07)	27.07 (10.15)	27.86 (10.93)	27.37 (10.46)	.74
Waist circumference (cm), mean (SD)		93.97 (14.89)	93.6 (13.67)	93.94 (13.88)	95.38 (14.69)	94.22 (14.29)	.30
Biomarkers, mean (SD)							
	HbA_{1c}[c] (mmol mol)	33.04 (5.71)	33.09 (6.47)	32.87 (5.84)	33.04 (4)	33.01 (5.58)	.95
	HbA_{1c} (%)	5.19 (0.52)	5.2 (0.59)	5.18 (0.53)	5.19 (0.37)	5.19 (0.51)	.96
	Total cholesterol (mmol/L)	4.78 (0.99)	4.9 (1.05)	4.9 (1)	4.99 (1.09)	4.89 (1.04)	.06
	LDL[d] cholesterol (mmol/L)	2.65 (0.84)	2.74 (0.87)	2.76 (0.85)	2.84 (0.92)	2.75 (0.87)	.03
	HDL[e] cholesterol (mmol/L)	1.58 (0.41)	1.58 (0.42)	1.57 (0.41)	1.57 (0.4)	1.57 (0.41)	>.99
	Triglycerides (mmol/L)	1.28 (0.87)	1.35 (0.93)	1.27 (0.81)	1.34 (1.01)	1.31 (0.91)	.54
	Systolic blood pressure	120.87 (14.34)	120.68 (14.74)	122.49 (15.75)	122.69 (14.04)	121.69 (14.75)	.10
	Diastolic blood pressure	75.95 (8.93)	75.81 (8.96)	76.62 (9.34)	77.51 (9.31)	76.48 (9.16)	.03
	Heart rate (bpm)	65.81 (10.08)	65.91 (11.01)	65.58 (10.54)	66.76 (11.06)	66.01 (10.68)	.41
	Hair cortisol (log)	3.98 (1.14)	3.92 (1.2)	3.99 (1.2)	3.99 (1.25)	3.97 (1.2)	.90

[a]NoHoW: Navigating to a Healthier Weight.

[b]Characteristics include demographics, anthropometry, and clinical markers at baseline across the 4 intervention groups (control, motivation, emotion, and motivation+emotion). Study population: adults with ≥5% weight loss in the previous 12 months, recruited in the United Kingdom, Denmark, and Portugal.

[c]HbA_{1c}: hemoglobin A_{1c}.

XSL•FO
RenderX

[d]LDL: low-density lipoprotein.

[e]HDL: high-density lipoprotein.

Primary Analysis (ITT)

There was an overall tendency for pretrial weight loss to be regained on average (Table 2). Participants were assigned to the following weight change categories: >−3% weight loss (weight loser), between −2.99% and +2.99% (weight maintainer), and >+3% weight gain (weight gainer) [71]. The results showed that 19% (302/1627) of participants continued to lose weight, 32% (302/1627) maintained weight, and 49% (811/1627) regained some weight (at ±3% weight change). When we investigated weaker assumptions, such as regain of a proportion of this weight loss, the main assessment of the intervention effects was similar, although absolute estimates of weight regain were lower in all arms (Table 2). Table 2 shows body composition and cardiometabolic biomarkers at 12 months by intervention arm for the whole sample. The motivation arm, the emotion arm, and the combined emotion and motivation arms did not significantly affect weight outcomes compared to the active control.

XSL•FO
RenderX

Table . Primary and prespecified secondary cardiometabolic outcomes at 12 months in the NoHoW[a] 2×2 factorial randomized controlled trial[b].

		Control, mean (95% CI)	Motivation, mean (95% CI)	Emotion, mean (95% CI)	Motivation + emotion	Motivation, *P* value	Emotion, *P* value	Motivation × emotion, *P* value
Body composition								
	Weight change (kg)	4.58 (3.41 to 5.75)	4.01 (2.82 to 5.2)	5.09 (3.92 to 6.26)	4.57 (3.38 to 5.76)	.49	.38	.95
	Fat-free mass (kg)	4.58 (3.42 to 5.75)	4.47 (3.43 to 5.51)	4.62 (3.61 to 5.63)	4.69 (3.53 to 5.84)	.86	.81	.73
	Fat mass (kg)	2.88 (1.7 to 4.07)	2.58 (1.57 to 3.58)	3.32 (2.31 to 4.33)	2.88 (1.76 to 4)	.29	.79	.78
	Waist circumference (cm)	4.03 (2.97 to 5.09)	3.41 (2.33 to 4.49)	4.24 (3.04 to 5.45)	3.89 (2.73 to 5.04)	.39	.62	.80
Biomarkers								
	HbA_{1c}[c] (mmol mol)	1.46 (0.96 to 1.96)	1.3 (0.84 to 1.77)	1.54 (1.07 to 2.01)	1.4 (0.91 to 1.88)	.43	.28	.83
	HbA_{1c} (%)	0.15 (0.12 to 0.18)	0.13 (0.1 to 0.17)	0.15 (0.11 to 0.18)	0.13 (0.1 to 0.17)	.31	.69	.91
	Total cholesterol (mmol/L)	0.19 (0.08 to 0.29)	0.23 (0.13 to 0.34)	0.26 (0.14 to 0.38)	0.29 (0.17 to 0.42)	.37	.12	.96
	LDL[d] cholesterol (mmol/L)	−0.05 (−0.2 to 0.11)	−0.01 (−0.12 to 0.1)	0.03 (−0.09 to 0.14)	0.05 (−0.06 to 0.16)	.67	.06	.72
	HDL[e] cholesterol (mmol/L)	0.1 (0.04 to 0.16)	0.11 (0.06 to 0.15)	0.08 (0.04 to 0.13)	0.11 (0.06 to 0.16)	.66	.92	.55
	Triglycerides (mmol/L)	0.3 (0.14 to 0.47)	0.25 (0.11 to 0.4)	0.29 (0.14 to 0.45)	0.23 (0.08 to 0.38)	.47	.66	>.99
	Systolic blood pressure (mm Hg)	1.97 (0.4 to 3.55)	1.44 (−0.3 to 3.19)	1.58 (−0.11 to 3.27)	2.55 (0.95 to 4.14)	.58	.27	.15
	Diastolic blood pressure (mm Hg)	1.48 (0.17 to 2.79)	0.54 (−0.75 to 1.83)	1.11 (−0.11 to 2.33)	1.06 (−0.07 to 2.2)	.55	>.99	.26
	Heart rate (bpm)	2.05 (0.62 to 3.48)	2.13 (0.75 to 3.5)	3.06 (1.62 to 4.51)	2.15 (0.52 to 3.79)	.46	.31	.39
	Hair cortisol (log)	0.22 (0.01 to 0.43)	0.24 (0.02 to 0.46)	0.26 (0.04 to 0.48)	0.22 (0.02 to 0.42)	.70	.59	.77

[a]NoHoW: Navigating to a Healthier Weight.

[b]Values are mean (95% CI) change from baseline to 12 months for waist circumference, fat mass, fat-free mass, HbA_{1c}, blood lipids, blood pressure, resting heart rate, and hair cortisol across intervention arms. *P* values reflect factorial main effects for the motivation component, emotion regulation component, and their interaction (motivation×emotion). Population: adults enrolled in the NoHoW digital weight-loss-maintenance randomized controlled trial across 3 European centers.

[c]HbA_{1c}: hemoglobin A_{1c}.

[d]LDL: low-density lipoprotein.

[e]HDL: high-density lipoprotein.

Subgroup Analyses (Prespecified: Men and Women)

Table S1 in Multimedia Appendix 1 shows sex-stratified body composition and cardiometabolic biomarkers at 12 months by intervention arm. Sex differences showed that a small effect of the motivation arm was significant in men only for the ITT, per-protocol, and completer populations. The analysis was factorial, looking at the effect of motivation and the effect of emotion and interactions. The effect of motivation was slightly significant, but we could not conclude that the interaction was significant; that is, we could not conclude that we had enough evidence to say motivation and emotion were more effective than motivation.

Figure S1 in Multimedia Appendix 1 shows longitudinal changes in body weight by trial arm (control, self-regulation and motivation, emotion regulation, and self-regulation and motivation plus emotion regulation combined) for men and women. Figure S2 in Multimedia Appendix 1 shows the percentage of the ITT population gaining, losing, or maintaining

XSL•FO
RenderX

pretrial weight loss by trial arm (control, motivation, motivation and emotion, and emotion). Percentage weight change was calculated as (CID3 kg–CID1 kg)/CID1 kg×100. The weight category was then calculated from the percentage weight change.

Sensitivity Analyses (Completers and Per-Protocol)

There was an overall tendency for pretrial weight loss to be regained on average (Table 2). For those who completed measures (ie, the per-protocol sample) at 12 months, men regained 0.14 kg (0.27%) of the weight they lost, and women regained 0.54 kg (0.85%) of the weight they lost. However, there was substantial variation between participants. For the per-protocol population, 24% (302/1263) continued to lose weight, 41% (514/1263) maintained weight, and 35% (447/1263) regained weight (Table 3). See Table S2 in Multimedia Appendix 1 for the sex-stratified sensitivity analyses of 12-month weight change (kilograms) for the ITT, per-protocol, and completer populations by intervention arm.

Table . Sensitivity analyses of 12-month weight change (kilograms) in the NoHoW[a] trial, including intention-to-treat (ITT), completer, and per-protocol (PP) populations[b].

	Control, mean (95% CI)	Motivation, mean (95% CI)	Emotion, mean (95% CI)	Motivation+emotion, mean (95% CI)	Motivation, *P* value	Emotion, *P* value	Motivation×emotion, *P* value
ITT	4.58 (3.41 to 5.75)	4.01 (2.82 to 5.2)	5.09 (3.92 to 6.26)	4.57 (3.38 to 5.76)	.49	.38	.95
PP	0.46 (−0.85 to 1.78)	−0.74 (−2.24 to 0.76)	−0.56 (−2.12 to 1)	−1.63 (−3.37 to 0.12)	.05	.39	.89
Completer	0.85 (−0.25 to 1.95)	0.39 (−0.72 to 1.51)	0.94 (−0.17 to 2.04)	0.46 (−0.67 to 1.58)	.44	.94	.97

[a]NoHoW: Navigating to a Healthier Weight.

[b]Values are mean (95% CI) change in body weight from baseline to 12 months for each intervention arm. *P* values reflect factorial main effects for motivation, emotion regulation, and their interaction (motivation×emotion). The PP population includes participants completing ≥80% of their assigned intervention modules.

Figure 3 shows that retention of participants in the trial was high and greater than assumed for the power calculations. By the 12-month primary outcome, over 77% (1263/1627) of the ITT sample was still engaged in the intervention, and by 18 months, over 71% of the ITT sample was still engaged in the intervention for all trial arms.

Figure 3. Retention of participants over time across intervention arms in the NoHoW 2×2 factorial randomized controlled trial. The figure displays cumulative dropout from baseline to 12 months and to 18 months for each intervention group (control, motivation, emotion, and motivation+emotion). Values represent the proportion retained at each follow-up time point. The purple dot at month 18 indicates an assumed dropout of 38% estimated from previous randomized controlled trials. NoHoW: Navigating to a Healthier Weight.

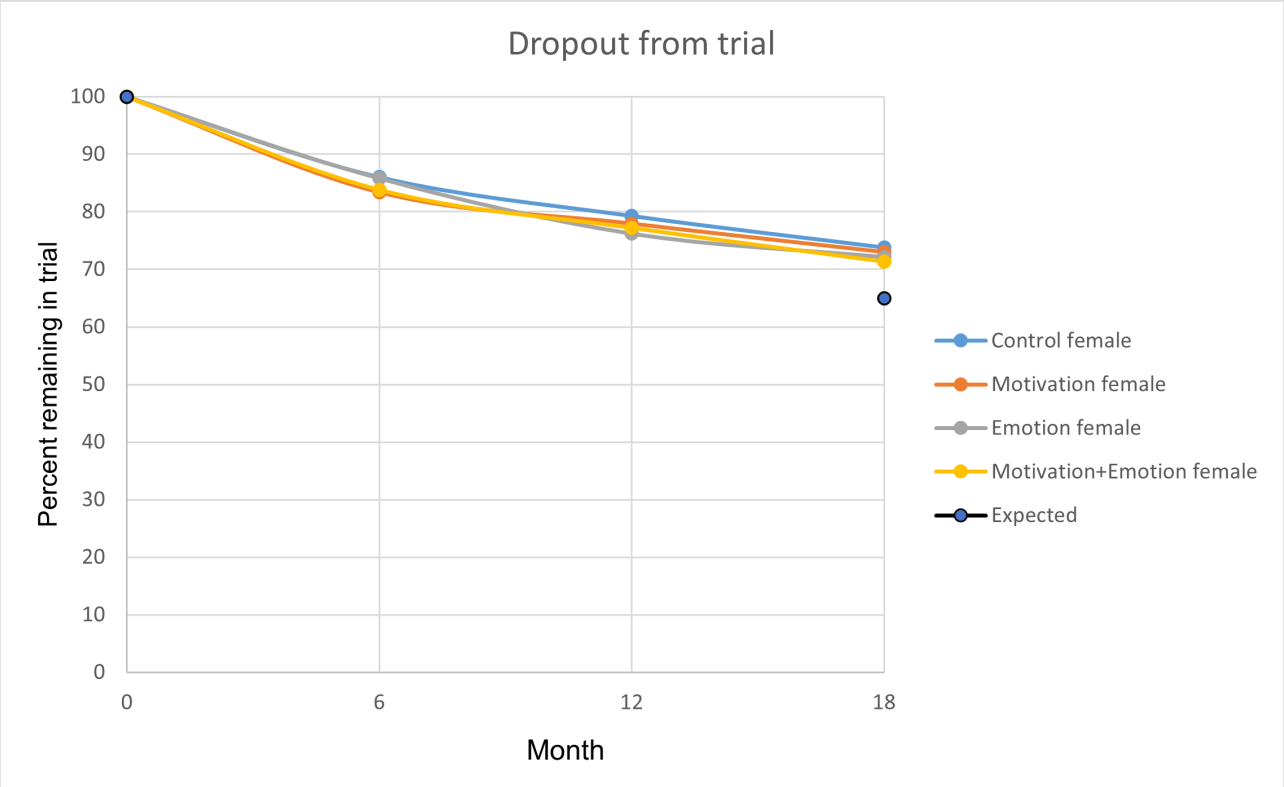

XSL•FO
RenderX

Exploratory Analyses (Post Hoc: Engagement or Compliance)

Table 4 shows univariate and multivariate regression analysis using toolkit compliance, Fitbit compliance, and self-weighing as predictors of weight outcomes. The univariate models for toolkit use, Fitbit compliance, and self-weighing all showed small predictive effects on weight loss, but in multivariate models, only self-weighing remained as a significant predictor independently of the other measures. However, the proportion of variance explained remains small.

Table . Exploratory associations between digital engagement metrics and 12-month weight change in the NoHoW[a] randomized controlled trial[b].

Predictor	Univariate models			Multivariate models		
	Coefficient	SE	*P* value	Coefficient	SE	*P* value
Toolkit use compliance (percentage)	−0.0259	0.0076	.001	−0.0113	0.0080	.16
Fitbit use compliance (number of days>20 hours use)	−0.0048	0.0014	.001	−0.0004	0.0017	.80
Self-weighing compliance (number of weeks with ≥2 weighings)	−0.0905	0.0109	<.0001	−0.0860	0.0135	<.0001

[a]NoHoW: Navigating to a Healthier Weight.

[b]Univariate and multivariable linear regression models examined associations between toolkit compliance, Fitbit wear-time compliance, and self-weighing frequency with weight change from baseline to 12 months. These analyses were post hoc and not prespecified.

Discussion

Principal Findings

NoHoW was the first project to develop and formally evaluate (via a randomized controlled trial [RCT]) a digital toolkit combining continuous tracking of energy balance behaviors and body weight with theoretically informed, evidence-based digital interventions targeting self-regulation and motivation, and emotion regulation in a 2×2 factorial design, to target long-term weight management.

This study was designed to test three hypotheses that (1) motivation and self-regulation of energy balance behaviors improve longer-term weight outcomes, (2) emotion regulation strategies help prevent weight relapse, and (3) there is an additive effect between emotion regulation and self-regulation or motivation. In the primary ITT analysis for all participants, none of the intervention arms differed significantly from the active control at 12 months. Exploratory subgroup analyses suggested a small effect of the motivation intervention among men, but this effect was not clinically meaningful and did not alter the overall null findings of the primary analysis. This intervention arm was also the arm in which intervention completion (but not necessarily the number of visits to the toolkit) had the highest correlation with weight loss [48]. The 2×2 factorial design was well-powered to detect intervention effects; yet, the interventions showed no detectable effect on the primary outcome and no meaningful effect on prespecified cardiometabolic secondary outcomes.

Although subgroup analyses indicated a small effect in men, these analyses were exploratory and should be interpreted with caution. These findings were not consistent with engagement data and, importantly, do not outweigh the primary ITT result, showing no intervention effect in the overall sample. Men may have been more engaged with the process-based intervention content, but user metrics for the digital intervention do not support this notion. Indeed, focusing on toolkit engagement (compliance) alone, there was a negative correlation between use metrics and weight change (all $P<.05$). However, multiple comparison adjustment would weaken these statistics. Measured use of the toolkit components suggested that men tended to use the toolkit less than women and that use was lower in the motivation arm (in which weight loss was greatest in men only), which does not support the hypothesis that compliance or engagement with the motivation arm of the toolkit explains the slightly greater weight loss in men.

Explaining the Negligible Effect of the Digital Intervention for Weight Loss Maintenance

Possible explanations for the lack of intervention effect may relate to the fact that the intervention was digital, that the toolkit content and dose per se (ie, the logic models) were not as effective as hypothesized (ie, those mechanisms are just not very effective and the dose was not sufficient), or that participants did not engage with the intervention (ie, this implies that the toolkit's design may not have adequately supported the delivery of its proposed mechanisms of action, which, if implemented more effectively, might have yielded the expected effects). Evidence-based behavior change interventions tend to show small effect sizes anyway [18-20,25,72,73]. In fact, the evidence tends to point in the direction that the most salient component of this trial associated with weight change was self-weighing.

The participant user engagement and experience evaluation of the content of the toolkit itself was revealing [48]. In this study, the actual use of the toolkit during 18 weeks of the active intervention was relatively high and comparable to other digital health interventions, and as with similar interventions, tended to decay with time [48]. The use declined rapidly during the period of discretionary use. Weight outcomes were influenced

XSL•FO
RenderX

by the degree of engagement with the study, as has been shown previously [74]. Weight outcomes were also weakly correlated with the use of the digital intervention content, but correlations between weight change from baseline to 12 months and use metrics were small and not consistent with the small differences in weight outcomes across trial arms (Table 2). While participants had some reservations about the content and ease of use of the toolkit, it does not appear that the degree of engagement, as measured by actual use of the toolkit, is a viable explanation for the general lack of between-arm effects of the intervention (Table 4).

Usability was rated as satisfactory, whereas acceptability was modest (between 3 and 4 on a 5-point scale). Both declined across all trial arms over the course of the study. Qualitative data showed that these perceptions of the toolkit were similar across centers and intervention arms [48]. Participants appreciated support for self-monitoring of weight, sleep, and activity, all of which were derived from the Fitbit devices. Participants also appreciated the content of the toolkit based on self-regulation of behavior and emotion regulation. They associated their decreased toolkit use with frustrations in design that hindered its use, for example, the log-in process as well as time commitment and frustrations using the toolkit on mobile phones [48]. Participants also tended to compare the design aspects of the toolkit unfavorably to the Fitbit app and other commercial offerings [48]. Such offerings are also regularly updated, while the NoHoW toolkit was static for 18 months to ensure that the trial was testing the same intervention in all participants. Thus, the user acceptability and experience data offer important insights into the limited impact of the digital intervention on weight outcomes. The intervention appears to have been less preferred than the Fitbit app, and intervention effects were very small. It does not matter how evidence-based or research-informed an intervention is if it is deemed to be difficult or even mediocre to use compared to other offerings available.

Thus, while it is important that digital behavior change interventions need to be developed with an interdisciplinary team and that the content, delivery, and structure are informed by theory and evidence [51], there are other important design aspects that could be addressed to improve the effect sizes of interventions such as the NoHoW toolkit. The NoHoW toolkit was meticulously developed with respect to (1) state-of-the-art theories (self-regulation theory, self-determination theory, and aspects of emotional regulation theory), (2) integration of consumer physical activity and weight tracking devices, (3) web-design expertise, (4) some user testing, and (5) preliminary mixed methods research. However, the delivery of the digital intervention may have benefited from greater involvement from human-computer interaction experts and co-design with prospective users together with commercial partners. There is some evidence that combining digital and human support may improve effective engagement with digital behavior change interventions, although more evidence is needed to configure the most effective combinations [75,76]. There is some evidence for similar effects in relation to mental health [77]. Such partners may have offered resources and infrastructure to produce an offering that motivates engagement and reinforces the practice of behavior change techniques using refreshed content, which may have improved outcomes. Due to resource constraints, these design aspects were limited in this trial. Involvement of health care professionals and engagement with social networks as a part of the intervention may improve some outcomes [78,79]. Using optimization designs for iterative development and testing of smaller components (eg, clusters or individual behavior change techniques) could also be beneficial to identify the combination of components that could work better together [80]. The NoHoW trial deliberately excluded interactions with health care professionals and social networks in order to test the effect of a digital-only intervention. A constraint of an RCT design is the fixed nature of the intervention during the trial. This approach precludes iterative development, testing, and ongoing content and design updates, the lack of which disappointed participants who are used to such features of commercial apps [48]. These factors should be taken into account in developing future digital interventions for weight management.

Is There Evidence That Fitbit Use Predicted Weight Loss Maintenance?

A possible factor (point 2 in the design specifications listed earlier) that may have limited the effect sizes observed in this study is the use of an active control, in which participants each received and used a Fitbit Charge 2 activity tracker and a set of Fitbit Aria wireless scales. All participants also had access to the Fitbit smartphone app. The use of trackers and scales was high [81,82]; therefore, it is likely that these represented an "intervention within an intervention." Ideally, it would have been better to conduct a 2×2 plus 1 trial, in which there was a genuine "no intervention" control, but resources did not permit this option. In regression models, weight change was significantly predicted by all of toolkit compliance, self-weighing compliance, and Fitbit use compliance. However, when all 3 were included in a model, only self-weighing compliance was significant; that is, there was little evidence that the toolkit had any effect over and above using the Fitbit devices alone.

Comparison to Other Trials Using Behavior Change Approaches for Longer-Term Weight Management

At present, systematic reviews and meta-analyses show the extent to which behavior change interventions for weight loss maintenance in adult populations are effective [18]. Generally, per-protocol results show greater weight loss than ITT analyses. A number of large trials focusing on evidence-based approaches to weight loss maintenance have demonstrated effects on weight-related outcomes, generally not exceeding 2 kg by trial end, over time periods ranging between 6 and 12 months [42]. These include the weight loss maintenance RCT [25], DiOGenes [83], PREVIEW [84], NuLevel [85], and NoHoW [50] trials. The Look AHEAD trial produced clinically significant weight loss (≥5%) after 8 years of intensive lifestyle intervention in 50% of 2570 adults with type 2 diabetes, a patient population with a strong clinical reason for trying to achieve weight loss maintenance [86].

In this study, imputing the missing data differently tended to have an effect on absolute but not relative weight outcomes

XSL•FO
RenderX

between trial arms. The more conservative our assumptions about dropout, the more favorable the resulting weight outcomes appeared; however, trial arm comparisons remained the same. When the ITT model was used with a baseline observation carried forward imputation, participants regained 40% - 50% of the weight they had lost prior to the intervention. This is consistent with similar studies in the magnitude and direction of weight regain over 12 months in response to dietary or behavior change interventions for weight management [25,83,85,86].

Strengths and Limitations of This Study

Key strengths of this study were the 2×2 factorial design of the trial, the sample size at 3 European centers across Northern, Western, and Southern Europe, standardization and harmonization of standard operating procedures across centers, the use of minimization in randomizing participants, the duration of follow-up, the relatively low dropout rate, and the fact that weight losses prior to the intervention were verified. Limitations included the lack of a genuine baseline due to the expediency of recruiting participants who had already lost weight in the previous 12 months, the lack of balance in content type and duration between trial arms, limitations to the design of the toolkit itself that were apparent in the user acceptability and experience data, the absence of a genuine "no intervention" control, and likely contamination of the intervention itself with the content of the Fitbit app. Several of these limitations may well have impacted weight outcomes and are discussed earlier. In addition, the sample was biased toward women (1117/1627, 68.7%) and so may not be as generalizable to men who are making weight management attempts. Therefore, the small effect observed in men should be interpreted cautiously and does not alter the overall null effect in the primary ITT analysis. Finally, this trial was conducted between 2017 and 2019, with sample processing and data analysis completed in 2020. There has been a significant delay in publication due to COVID-19, organizational changes in some universities, and staff relocations among some of the key authors. However, to our knowledge, this study is still the first to use 2×2 factorial RCT in different countries to test a digital toolkit for weight loss maintenance after a period of clinically significant weight loss using self-regulation and emotion regulation; therefore, it provides a unique and timely contribution to the literature on digital interventions for weight loss maintenance, addressing a critical gap in evidence-based strategies.

Future Directions

The NoHoW trial has demonstrated the immense benefits of using cloud-connected digital trackers to help users track their weight, physical activity, and sleep. It is notable that the salient predictor of successful weight management was engagement with the Fitbit digital weighing scales. Multicomponent behavior change interventions are, by nature, complex [87], as are physiological and behavioral responses to weight loss. It is well-established that weight loss leads to some compensation of energy expenditure and energy intake and that some behavioral changes in these components of energy balance are not always under conscious control [42]. The limited evidence from systematic reviews and meta-analyses to identify mediators of longer-term weight loss suggests that navigating from initial weight loss to weight loss maintenance requires long-term self-management of energy balance behaviors in the face of physiological resistance to weight loss [14,18-20].

The Capability, Opportunity, Motivation, Behavior model provides an overarching theoretical framework to understand the barriers and facilitators of behavior change. Specifically, the model suggests that behavior change requires capability (physiological or physical ability), motivation (reflective and automatic processes that activate or inhibit behavior), and opportunity (physical and social environment to enable behavior) [88]. Kwasnicka et al [16] have systematically reviewed theoretical explanations for the maintenance of behavior change and identified 5 overarching theoretical explanations for the maintenance of behavior change representing motives, self-regulation, psychological and physical resources, habits, and environmental or social influences on behavior. A key question is how such frameworks for reflective and automatic mechanisms of behavior change interface with the physiology of energy balance compensation in response to attempted or imposed energy deficits. Tracking compensatory changes in energy balance behaviors over time may improve the prevention of weight regain through the provision of behavioral navigation solutions, personalization of weight management interventions, and by offering a quantitative behavioral context in which psychological moderators and mediators of energy balance behaviors and body weight can be tracked [42]. Similarly, using metadata from digital interventions to track user engagement offers a powerful quantitative tool to understand the behavior of participants during the course of longer-term interventions [48]. Thus, while the NoHoW trial produced very modest results in terms of the hypothesized mechanisms of action of digital interventions, it provided unique insights into the limitations of digital intervention design and how they can be addressed and incorporated into future weight management provision.

Our understanding of factors that promote or undermine self-management of eating and physical activity is still limited but includes physiological adaptation and behavioral compensation in eating and physical activity, reactive processes related to emotions, stress, rewards, and desires that meet psychological needs. Optimizing and piloting evidence-based intervention content to the needs of individuals may improve outcomes. Objective longitudinal tracking of weight and physical activity and through mathematical modeling of energy balance components over time would provide a quantitative framework to understand the dynamics and mechanisms of action of behavior change interventions [42]. Such tracking also offers users navigational solutions. These can be combined with tracking of user engagement with intervention components to potentially improve weight management intervention design and evaluation.

Conclusions

The NoHoW trial was a well-designed trial, adequately powered to detect changes in weight and health outcomes. In the primary ITT analysis in all participants, we found no evidence that any of the digital interventions improved weight loss maintenance compared with the active control. A small effect observed in

XSL•FO
RenderX

men in exploratory subgroup analyses was not clinically meaningful and did not change the overall interpretation of the primary results.

This study is innovative in being the first large-scale, multicountry RCT to evaluate a digital-only toolkit for weight loss maintenance that integrates self-regulation, motivation, and emotion regulation within a 2×2 factorial design. NoHoW differs from prior trials by focusing on maintenance after clinically significant weight loss and by testing the independent and additive effects of digitally delivered techniques for behavioral self-regulation and emotion regulation components. The trial provides evidence on both the limitations and potential of digital behavior change interventions for long-term weight outcomes. The trial is also unique in demonstrating the potential for combined digital tracking of energy balance metrics (body weight and physical activity) and user engagement with intervention components. For health care providers, developers, and policymakers, the findings suggest that digital interventions alone may be insufficient for sustained weight management unless combined with interactive user-based feedback enabling personalization and, where feasible, human or social support. The findings emphasize the importance of routine day-to-day user experience and acceptability of digital interventions for health-related behavior change in addition to theoretical and methodological rigor.

Despite being well designed from a theoretical and evidence-based perspective, the NoHoW trial showed some limitations in consumer or user acceptability, and this study highlights some possible design aspects of the digital behavior change intervention that could be improved to enhance intervention effectiveness. These include development by an interdisciplinary team including specialists in app design and marketing or delivery; content, delivery, and structure informed by theory and evidence [51]; greater involvement from human-computer interaction experts; using optimization designs for interactive development and testing of smaller components (eg, clusters or individual behavior change techniques) to identify the combination of components that could work better together [80]; greater co-design with prospective users [89,90]; commercial partners to make digital intervention appearance and content comparable to those currently available; involvement of health care professionals; and embedding interventions and social networks to provide an ecosystem of support. These aspects of intervention design do not lend themselves well to the design and conduct of RCTs in a traditional grant-funded environment, which are constrained in time and budgetary allocation. We therefore recommend that future digital interventions for weight management are developed through academic and commercial or health care partnerships, extensively tested and refined by users and providers by the time they are evaluated through randomized trials. Novel designs, including using optimization designs and modifiable intervention content during trial designs, may also enhance effectiveness in the future.

Acknowledgments

Generative artificial intelligence was not used in any portion of the manuscript writing.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation program (grant 643309). The material presented and views expressed here are the responsibility of the authors only. The European Union Commission takes no responsibility for any use made of the information set out. MMM was funded by the Marie-Sklodowska-Curie (EDGE) Fellowship program (grant 713567).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

RJS, BLH, PJT, FFS, ALP, and GWH conceived the study. RJS is the principal investigator, BLH is the grant coordinator, and SES and C Duarte are the trial managers. SCL, ALP, and SES are site coordinators. GWH is the trial statistician and leads data management. ALP, MMM, MM, and C Duarte developed the content of the arms of the digital toolkit. PJT, JE, IS, BLH, M-LM, SCL, RJS, SES, BP, and LR conducted the trial activities at each respective trial site. EHE and FFS were involved in the development of some of the trial outcome measures. RJS, C Duarte, and C Dakin drafted the manuscript, and all authors revised and approved the final version. C Dakin submitted the manuscript for publication. All authors made substantial contributions to the conceptualization of the study design and conduct of the protocol.

Conflicts of Interest

RJS consults for Slimming World through the University of Leeds. MMM and GWH also conducted consultancy work with Slimming World.

Multimedia Appendix 1
Additional tables and figures.

XSL•FO
RenderX

[DOCX File, 157 KB - jmir_v28i1e69634_app1.docx]

Checklist 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File, 2902 KB - jmir_v28i1e69634_app2.pdf]

References

1. WHO European regional obesity report 2022. World Health Organization. 2022. URL: https://www.who.int/europe/publications/i/item/9789289057738 [accessed 2026-03-13]
2. Müller-Riemenschneider F, Reinhold T, Berghöfer A, Willich SN. Health-economic burden of obesity in Europe. Eur J Epidemiol 2008;23(8):499-509. [doi: 10.1007/s10654-008-9239-1] [Medline: 18509729]
3. Wang YC, McPherson K, Marsh T, Gortmaker SL, Brown M. Health and economic burden of the projected obesity trends in the USA and the UK. Lancet 2011 Aug 27;378(9793):815-825. [doi: 10.1016/S0140-6736(11)60814-3] [Medline: 21872750]
4. Santos I, Sniehotta FF, Marques MM, Carraça EV, Teixeira PJ. Prevalence of personal weight control attempts in adults: a systematic review and meta-analysis. Obes Rev 2017 Jan;18(1):32-50. [doi: 10.1111/obr.12466] [Medline: 27653242]
5. Gudzune KA, Doshi RS, Mehta AK, et al. Efficacy of commercial weight-loss programs: an updated systematic review. Ann Intern Med 2015 Apr 7;162(7):501-512. [doi: 10.7326/M14-2238] [Medline: 25844997]
6. Vakil RM, Chaudhry ZW, Doshi RS, Clark JM, Gudzune KA. Commercial programs' online weight-loss claims compared to results from randomized controlled trials. Obesity (Silver Spring) 2017 Nov;25(11):1885-1893. [doi: 10.1002/oby.21959] [Medline: 28865085]
7. Vakil RM, Doshi RS, Mehta AK, et al. Direct comparisons of commercial weight-loss programs on weight, waist circumference, and blood pressure: a systematic review. BMC Public Health 2016 Dec;16(1):460. [doi: 10.1186/s12889-016-3112-z]
8. Carr D, Friedman MA. Is obesity stigmatizing? Body weight, perceived discrimination, and psychological well-being in the United States. J Health Soc Behav 2005 Sep;46(3):244-259. [doi: 10.1177/002214650504600303] [Medline: 16259147]
9. O'Brien KS, Latner JD, Puhl RM, et al. The relationship between weight stigma and eating behavior is explained by weight bias internalization and psychological distress. Appetite 2016 Jul 1;102:70-76. [doi: 10.1016/j.appet.2016.02.032] [Medline: 26898319]
10. Puhl RM, Heuer CA. The stigma of obesity: a review and update. Obesity (Silver Spring) 2009 May;17(5):941-964. [doi: 10.1038/oby.2008.636] [Medline: 19165161]
11. Hartmann-Boyce J, Cobiac LJ, Theodoulou A, et al. Weight regain after behavioural weight management programmes and its impact on quality of life and cost effectiveness: evidence synthesis and health economic analyses. Diabetes Obes Metab 2023 Feb;25(2):526-535. [doi: 10.1111/dom.14895] [Medline: 36239137]
12. Rössner S. Relapse management. Int J Obes 2008;31:S10.
13. Stubbs RJ, Lavin JH. The challenges of implementing behaviour changes that lead to sustained weight management. Nutr Bull 2013 Mar;38(1):5-22. [doi: 10.1111/nbu.12002]
14. Greaves CJ, Sheppard KE, Abraham C, et al. Systematic review of reviews of intervention components associated with increased effectiveness in dietary and physical activity interventions. BMC Public Health 2011 Feb 18;11(1):119. [doi: 10.1186/1471-2458-11-119] [Medline: 21333011]
15. Hagger MS. Non-conscious processes and dual-process theories in health psychology. Health Psychol Rev 2016 Dec;10(4):375-380. [doi: 10.1080/17437199.2016.1244647] [Medline: 27718880]
16. Kwasnicka D, Dombrowski SU, White M, Sniehotta F. Theoretical explanations for maintenance of behaviour change: a systematic review of behaviour theories. Health Psychol Rev 2016 Sep;10(3):277-296. [doi: 10.1080/17437199.2016.1151372] [Medline: 26854092]
17. James Stubbs R, Horgan G, Robinson E, Hopkins M, Dakin C, Finlayson G. Diet composition and energy intake in humans. Phil Trans R Soc B 2023 Oct 23;378(1888):20220449. [doi: 10.1098/rstb.2022.0449]
18. Dombrowski SU, Knittle K, Avenell A, Araújo-Soares V, Sniehotta FF. Long term maintenance of weight loss with non-surgical interventions in obese adults: systematic review and meta-analyses of randomised controlled trials. BMJ 2014 May 14;348:g2646. [doi: 10.1136/bmj.g2646] [Medline: 25134100]
19. Teixeira PJ, Carraça EV, Marques MM, et al. Successful behavior change in obesity interventions in adults: a systematic review of self-regulation mediators. BMC Med 2015 Apr 16;13(1):84. [doi: 10.1186/s12916-015-0323-6] [Medline: 25907778]
20. Varkevisser RDM, van Stralen MM, Kroeze W, Ket JCF, Steenhuis IHM. Determinants of weight loss maintenance: a systematic review. Obes Rev 2019 Feb;20(2):171-211. [doi: 10.1111/obr.12772] [Medline: 30324651]
21. Phelan S, Hill JO, Lang W, Dibello JR, Wing RR. Recovery from relapse among successful weight maintainers. Am J Clin Nutr 2003 Dec;78(6):1079-1084. [doi: 10.1093/ajcn/78.6.1079]

XSL•FO
RenderX

22. Carey RN, Connell LE, Johnston M, et al. Behavior change techniques and their mechanisms of action: a synthesis of links described in published intervention literature. Ann Behav Med 2019 Jul 17;53(8):693-707. [doi: 10.1093/abm/kay078] [Medline: 30304386]
23. Connell LE, Carey RN, de Bruin M, et al. Links between behavior change techniques and mechanisms of action: an expert consensus study. Ann Behav Med 2019 Jul 17;53(8):708-720. [doi: 10.1093/abm/kay082] [Medline: 30452535]
24. Maes S, Karoly P. Self‐regulation assessment and intervention in physical health and illness: a review. Appl Psychol 2005 Apr;54(2):267-299. [doi: 10.1111/j.1464-0597.2005.00210.x]
25. Svetkey LP, Stevens VJ, Brantley PJ, et al. Comparison of strategies for sustaining weight loss: the weight loss maintenance randomized controlled trial. JAMA 2008 Mar 12;299(10):1139-1148. [doi: 10.1001/jama.299.10.1139] [Medline: 18334689]
26. Larimer ME, Palmer RS, Marlatt GA. Relapse prevention: an overview of Marlatt's cognitive-behavioral model. In: Psychosocial Treatments: Routledge; 2004.
27. Silva MN, Markland D, Carraça EV, et al. Exercise autonomous motivation predicts 3-yr weight loss in women. Med Sci Sports Exerc 2011 Apr;43(4):728-737. [doi: 10.1249/MSS.0b013e3181f3818f] [Medline: 20689448]
28. Teixeira PJ, Silva MN, Mata J, Palmeira AL, Markland D. Motivation, self-determination, and long-term weight control. Int J Behav Nutr Phys Act 2012 Mar 2;9(1):22. [doi: 10.1186/1479-5868-9-22] [Medline: 22385818]
29. Hennessy EA, Johnson BT, Acabchuk RL, McCloskey K, Stewart-James J. Self-regulation mechanisms in health behavior change: a systematic meta-review of meta-analyses, 2006-2017. Health Psychol Rev 2020 Mar;14(1):6-42. [doi: 10.1080/17437199.2019.1679654] [Medline: 31662031]
30. Samdal GB, Eide GE, Barth T, Williams G, Meland E. Effective behaviour change techniques for physical activity and healthy eating in overweight and obese adults; systematic review and meta-regression analyses. Int J Behav Nutr Phys Act 2017 Mar 28;14(1):42. [doi: 10.1186/s12966-017-0494-y] [Medline: 28351367]
31. Gibbons FX, Houlihan AE, Gerrard M. Reason and reaction: the utility of a dual-focus, dual-processing perspective on promotion and prevention of adolescent health risk behaviour. Br J Health Psychol 2009 May;14(Pt 2):231-248. [doi: 10.1348/135910708X376640] [Medline: 19026095]
32. Stubbs RJ, Duarte C, O'Driscoll R, Turicchi J, Michalowska J. Developing evidence-based behavioural strategies to overcome physiological resistance to weight loss in the general population. Proc Nutr Soc 2019 Nov;78(4):576-589. [doi: 10.1017/S0029665119001083] [Medline: 31670628]
33. Bargh JA, Chartrand TL. The unbearable automaticity of being. Am Psychol 1999;54(7):462-479. [doi: 10.1037/0003-066X.54.7.462]
34. Gardner B, Rebar AL, Lally P. A matter of habit: recognizing the multiple roles of habit in health behaviour. Br J Health Psychol 2019 May;24(2):241-249. [doi: 10.1111/bjhp.12369] [Medline: 30945793]
35. Latner JD, O'Brien KS, Durso LE, Brinkman LA, MacDonald T. Weighing obesity stigma: the relative strength of different forms of bias. Int J Obes 2008 Jul;32(7):1145-1152. [doi: 10.1038/ijo.2008.53]
36. Schafer MH. The stigma of obesity: does weight discrimination affect identity and physical health? Soc Psychol Q 2011;74:76-97. [doi: 10.1177/0190272511398197]
37. Sikorski C, Luppa M, Luck T, Riedel-Heller SG. Weight stigma "gets under the skin"—evidence for an adapted psychological mediation framework—a systematic review. Obesity (Silver Spring) 2015 Feb;23(2):266-276. [doi: 10.1002/oby.20952] [Medline: 25627624]
38. Carter A, Hoang N, Gilbert P, Kirby JN. Body weight perception outweighs body weight when predicting shame, criticism, depression and anxiety for lower BMI range and higher BMI range adults. J Health Psychol 2022 Sep;27(10):2276-2290. [doi: 10.1177/13591053211027641] [Medline: 34240637]
39. Haynes A, Kersbergen I, Sutin A, Daly M, Robinson E. A systematic review of the relationship between weight status perceptions and weight loss attempts, strategies, behaviours and outcomes. Obes Rev 2018 Mar;19(3):347-363. [doi: 10.1111/obr.12634] [Medline: 29266851]
40. Carrière K, Khoury B, Günak MM, Knäuper B. Mindfulness-based interventions for weight loss: a systematic review and meta-analysis. Obes Rev 2018 Feb;19(2):164-177. [doi: 10.1111/obr.12623] [Medline: 29076610]
41. Forman EM, Butryn ML. A new look at the science of weight control: how acceptance and commitment strategies can address the challenge of self-regulation. Appetite 2015 Jan;84:171-180. [doi: 10.1016/j.appet.2014.10.004] [Medline: 25445199]
42. Stubbs RJ, Duarte C, O'Driscoll R, et al. The H2020 "NoHoW Project": a position statement on behavioural approaches to longer-term weight management. Obes Facts 2021;14(2):246-258. [doi: 10.1159/000513042] [Medline: 33662958]
43. Kupila SKE, Joki A, Suojanen LU, Pietiläinen KH. The effectiveness of eHealth interventions for weight loss and weight loss maintenance in adults with overweight or obesity: a systematic review of systematic reviews. Curr Obes Rep 2023 Sep;12(3):371-394. [doi: 10.1007/s13679-023-00515-2] [Medline: 37354334]
44. eHealth Action Plan 2012-2020. European Commission. 2012. URL: https://health.ec.europa.eu/publications/ehealth-action-plan-2012-2020_en [accessed 2026-03-13]
45. Bardus M, Smith JR, Samaha L, Abraham C. Mobile and Web 2.0 interventions for weight management: an overview of review evidence and its methodological quality. Eur J Public Health 2016 Aug;26(4):602-610. [doi: 10.1093/eurpub/ckw090]

XSL•FO
RenderX

46. Dounavi K, Tsoumani O. Mobile health applications in weight management: a systematic literature review. Am J Prev Med 2019 Jun;56(6):894-903. [doi: 10.1016/j.amepre.2018.12.005] [Medline: 31003801]
47. Villinger K, Wahl DR, Boeing H, Schupp HT, Renner B. The effectiveness of app-based mobile interventions on nutrition behaviours and nutrition-related health outcomes: a systematic review and meta-analysis. Obes Rev 2019 Oct;20(10):1465-1484. [doi: 10.1111/obr.12903] [Medline: 31353783]
48. Mattila E, Hansen S, Bundgaard L, et al. Users' experiences with the NoHoW web-based toolkit with weight and activity tracking in weight loss maintenance: long-term randomized controlled trial. J Med Internet Res 2022 Jan 10;24(1):e29302. [doi: 10.2196/29302] [Medline: 35006081]
49. Palmeira AL, Marques MM, Sánchez-Oliva D, et al. Are motivational and self-regulation factors associated with 12 months' weight regain prevention in the NoHoW study? An analysis of European adults. Int J Behav Nutr Phys Act 2023;20(1):128. [doi: 10.1186/s12966-023-01529-8]
50. Scott SE, Duarte C, Encantado J, et al. The NoHoW protocol: a multicentre 2×2 factorial randomised controlled trial investigating an evidence-based digital toolkit for weight loss maintenance in European adults. BMJ Open 2019 Sep 30;9(9):e029425. [doi: 10.1136/bmjopen-2019-029425] [Medline: 31575569]
51. Marques MM, Matos M, Mattila E, et al. A theory- and evidence-based digital intervention tool for weight loss maintenance (NoHoW toolkit): systematic development and refinement study. J Med Internet Res 2021 Dec 3;23(12):e25305. [doi: 10.2196/25305] [Medline: 34870602]
52. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ 2008:a1655. [doi: 10.1136/bmj.a1655]
53. Moher D, Chan AW. SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials). In: Guidelines for Reporting Health Research: A User's Manual: John Wiley & Sons, Ltd; 2014:56-67. [doi: 10.1002/9781118715598]
54. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. Lancet 2025 May;405(10489):1633-1640. [doi: 10.1016/S0140-6736(25)00672-5]
55. Jamnik VK, Warburton DER, Makarski J, et al. Enhancing the effectiveness of clearance for physical activity participation: background and overall process. Appl Physiol Nutr Metab 2011 Jul;36(S1):S3-S13. [doi: 10.1139/h11-044]
56. Simpson E, Bradley J, Poliakov I, et al. Iterative development of an online dietary recall tool: INTAKE24. Nutrients 2017 Feb 9;9(2):118. [doi: 10.3390/nu9020118] [Medline: 28208763]
57. De Lorenzo A, Andreoli A, Matthie J, Withers P. Predicting body cell mass with bioimpedance by using theoretical methods: a technological review. J Appl Physiol (1985) 1997 May;82(5):1542-1558. [doi: 10.1152/jappl.1997.82.5.1542] [Medline: 9134904]
58. Guide to physical measurements secondary guide to physical measurements. World Health Organization. 2008. URL: https://cdn.who.int/media/docs/default-source/ncds/ncd-surveillance/steps/part3-section5.pdf [accessed 2026-03-13]
59. Jain A, Rao N, Sharifi M, et al. Evaluation of the point of care Afinion AS100 analyser in a community setting. Ann Clin Biochem 2017 May;54(3):331-341. [doi: 10.1177/0004563216661737] [Medline: 27422136]
60. Van Uum SHM, Sauvé B, Fraser LA, Morley-Forster P, Paul TL, Koren G. Elevated content of cortisol in hair of patients with severe chronic pain: a novel biomarker for stress. Stress 2008 Nov;11(6):483-488. [doi: 10.1080/10253890801887388] [Medline: 18609301]
61. Craig CL, Marshall AL, Sjöström M, et al. International Physical Activity Questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 2003 Aug;35(8):1381-1395. [doi: 10.1249/01.MSS.0000078924.61453.FB]
62. Mullen SP, Silva MN, Sardinha LB, Teixeira PJ. Initial validation of the activity choice index among overweight women. Res Q Exerc Sport 2016 Jun;87(2):174-181. [doi: 10.1080/02701367.2016.1152349] [Medline: 27030291]
63. Vasileiou A, Searle D, Larsen SC, et al. Comparing self-reported energy intake using an online dietary tool with energy expenditure by an activity tracker. Nutrition 2024 Feb;118:112258. [doi: 10.1016/j.nut.2023.112258] [Medline: 38007995]
64. Stunkard AJ, Messick S. The three-factor eating questionnaire to measure dietary restraint, disinhibition and hunger. J Psychosom Res 1985;29(1):71-83. [doi: 10.1016/0022-3999(85)90010-8] [Medline: 3981480]
65. Gormally J, Black S, Daston S, Rardin D. The assessment of binge eating severity among obese persons. Addict Behav 1982;7(1):47-55. [doi: 10.1016/0306-4603(82)90024-7] [Medline: 7080884]
66. Tylka TL, Kroon Van Diest AM. The Intuitive Eating Scale-2: item refinement and psychometric evaluation with college women and men. J Couns Psychol 2013 Jan;60(1):137-153. [doi: 10.1037/a0030893] [Medline: 23356469]
67. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011 Dec;20(10):1727-1736. [doi: 10.1007/s11136-011-9903-x] [Medline: 21479777]
68. Tennant R, Hiller L, Fishwick R, et al. The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. Health Qual Life Outcomes 2007 Nov 27;5(1):1-13. [doi: 10.1186/1477-7525-5-63] [Medline: 18042300]
69. Larsen TM, Dalskov S, van Baak M, et al. The Diet, Obesity and Genes (Diogenes) Dietary Study in eight European countries—a comprehensive design for long-term intervention. Obes Rev 2010 Jan;11(1):76-91. [doi: 10.1111/j.1467-789X.2009.00603.x] [Medline: 19470086]
70. Little RJA. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 1988 Dec;83(404):1198-1202. [doi: 10.1080/01621459.1988.10478722]

XSL•FO
RenderX

71. Stevens J, Truesdale KP, McClain JE, Cai J. The definition of weight maintenance. Int J Obes 2006 Mar;30(3):391-399. [doi: 10.1038/sj.ijo.0803175]
72. Dombrowski SU, Sniehotta FF, Avenell A, Johnston M, MacLennan G, Araújo-Soares V. Identifying active ingredients in complex behavioural interventions for obese adults with obesity-related co-morbidities or additional risk factors for co-morbidities: a systematic review. Health Psychol Rev 2012 Mar;6(1):7-32. [doi: 10.1080/17437199.2010.513298]
73. Dombrowski SU, Avenell A, Sniehott FF. Behavioural interventions for obese adults with additional risk factors for morbidity: systematic review of effects on behaviour, weight and disease risk factors. Obes Facts 2010 Dec;3(6):377-396. [doi: 10.1159/000323076] [Medline: 21196792]
74. Jebb SA, Ahern AL, Olson AD, et al. Primary care referral to a commercial provider for weight loss treatment versus standard care: a randomised controlled trial. The Lancet 2011 Oct;378(9801):1485-1492. [doi: 10.1016/S0140-6736(11)61344-5]
75. Yardley L, Spring BJ, Riper H, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
76. Mair JL, Salamanca-Sanabria A, Augsburger M, et al. Effective behavior change techniques in digital health interventions for the prevention or management of noncommunicable diseases: an umbrella review. Ann Behav Med 2023 Sep 13;57(10):817-835. [doi: 10.1093/abm/kaad041] [Medline: 37625030]
77. Werntz A, Amado S, Jasman M, Ervin A, Rhodes JE. Providing human support for the use of digital mental health interventions: systematic meta-review. J Med Internet Res 2023 Feb 6;25:e42864. [doi: 10.2196/42864] [Medline: 36745497]
78. Chopra S, Malhotra A, Ranjan P, et al. Predictors of successful weight loss outcomes amongst individuals with obesity undergoing lifestyle interventions: a systematic review. Obes Rev 2021 Mar;22(3):e13148. [doi: 10.1111/obr.13148] [Medline: 33200547]
79. Jensen MT, Nielsen SS, Jessen-Winge C, et al. The effectiveness of social-support-based weight-loss interventions—a systematic review and meta-analysis. Int J Obes 2024 May;48(5):599-611. [doi: 10.1038/s41366-024-01468-9]
80. Marques MM, Guastaferro K. The development of effective and tailored digital behavior change interventions: an introduction to the multiphase optimization strategy (MOST). Eur Health Psychol 2022;22(4):793-800.
81. Turicchi J, O'Driscoll R, Horgan G, et al. Weekly, seasonal and holiday body weight fluctuation patterns among individuals engaged in a European multi-centre behavioural weight loss maintenance intervention. PLoS ONE 2020;15(4):e0232152. [doi: 10.1371/journal.pone.0232152]
82. Turicchi J, O'Driscoll R, Lowe M, et al. The impact of early body-weight variability on long-term weight maintenance: exploratory results from the NoHoW weight-loss maintenance intervention. Int J Obes 2021 Mar;45(3):525-534. [doi: 10.1038/s41366-020-00706-0]
83. Larsen TM, Dalskov SM, van Baak M, et al. Diets with high or low protein content and glycemic index for weight-loss maintenance. N Engl J Med 2010 Nov 25;363(22):2102-2113. [doi: 10.1056/NEJMoa1007137] [Medline: 21105792]
84. Huttunen-Lenz M, Hansen S, Larsen TM, et al. The PREVIEW study. Eur J Health Psychol 2019 May;26(1):10-20. [doi: 10.1027/2512-8442/a000026]
85. Sniehotta FF, Evans EH, Sainsbury K, et al. Behavioural intervention for weight loss maintenance versus standard weight advice in adults with obesity: a randomised controlled trial in the UK (NULevel Trial). PLoS Med 2019 May;16(5):e1002793. [doi: 10.1371/journal.pmed.1002793] [Medline: 31063507]
86. The Look AHEAD Research Group. Eight‐year weight losses with an intensive lifestyle intervention: the Look AHEAD study. Obesity (Silver Spring) 2014 Jan;22(1):5-13. [doi: 10.1002/oby.20662]
87. Michie S, Johnston M. Theories and techniques of behaviour change: developing a cumulative science of behaviour change. Health Psychol Rev 2012 Mar;6(1):1-6. [doi: 10.1080/17437199.2012.654964]
88. West R, Michie S. A brief introduction to the COM-B model of behaviour and the PRIME Theory of motivation. Qeios 2020. [doi: 10.32388/WW04E6.2]
89. Leung MW, Yen IH, Minkler M. Community based participatory research: a promising approach for increasing epidemiology's relevance in the 21st century. Int J Epidemiol 2004 Jun;33(3):499-506. [doi: 10.1093/ije/dyh010] [Medline: 15155709]
90. Rice K, Te Hiwi B, Zwarenstein M, et al. Best practices for the prevention and management of diabetes and obesity-related chronic disease among Indigenous peoples in Canada: a review. Can J Diabetes 2016 Jun;40(3):216-225. [doi: 10.1016/j.jcjd.2015.10.007] [Medline: 27066857]

Abbreviations

CID: clinical investigation day
CONSORT: Consolidated Standards of Reporting Trials
HbA_{1c}: hemoglobin A_{1c}
HDL: high-density lipoprotein
ITT: intention-to-treat
NoHoW: Navigating to a Healthier Weight

XSL•FO
RenderX

RCT: randomized controlled trial
SPIRIT: Standard Protocol Items: Recommendations for Interventional Trials

Edited by N Cahill, S Brini; submitted 05.Dec.2024; peer-reviewed by H Ranjani, S Kondapally, S Babatope; revised version received 15.Jan.2026; accepted 16.Jan.2026; published 20.May.2026.

Please cite as:
Stubbs RJ, Duarte C, Dakin C, Palmeira AL, Sniehotta FF, Horgan GW, Larsen SC, Marques MM, Encantado J, Evans EH, Turicchi J, O'Driscoll R, Scott SE, Pearson B, Ramsey L, Mikkelsen ML, Santos I, Matos M, Teixeira PJ, Heitmann BL
A Digital Toolkit for Weight Loss Maintenance in European Adults (NoHoW): 2×2 Factorial Randomized Controlled Trial
J Med Internet Res 2026;28:e69634
URL: https://www.jmir.org/2026/1/e69634
doi:10.2196/69634

© R James Stubbs, Cristiana Duarte, Clarissa Dakin, António L Palmeira, Falko F Sniehotta, Graham W Horgan, Sofus Larsen, Marta M Marques, Jorge Encantado, Elizabeth H Evans, Jake Turicchi, Ruari O'Driscoll, Sarah E Scott, Beth Pearson, Lauren Ramsey, Marie-Louise Mikkelsen, Inês Santos, Marcela Matos, Pedro J Teixeira, Berit L Heitmann. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effect of Integrated Internet-Based Acceptance and Commitment Therapy and Behavioral Activation Among Ethnic Minority Young Adults With Alcohol Use Disorder in Hong Kong: Pilot Randomized Controlled Trial

Getaneh Mulualem Belay[1,2], BN, MSc, RN; Katherine Ka Wai Lam[1], BN, MPhil, PhD, RN; Qi Liu[1], BN, MPhil, PhD, RN; Ting Mao[1], BN, MPhil, RN; Yim Wah Mak[1], BN, MPhil, PhD, RN; Ka Yan Ho[1], BN, MPhil, PhD, RN

[1]School of Nursing, The Hong Kong Polytechnic University, 11 Yuk Choi Road, Hung Hom, Kowloon, Hong Kong, China

[2]School of Nursing, College of Medicine and Health Sciences, University of Gondar, Gondar, Ethiopia

Corresponding Author:
Ka Yan Ho, BN, MPhil, PhD, RN
School of Nursing, The Hong Kong Polytechnic University, 11 Yuk Choi Road, Hung Hom, Kowloon, Hong Kong, China

Abstract

Background: Alcohol use disorder (AUD) profoundly affected ethnic minority young adults due to unique stressors, such as acculturation challenges and discrimination. Our prior studies found that experiential avoidance behavior and discrimination significantly contributed to AUD. These findings warranted third-wave interventions, such as acceptance and commitment therapy (ACT). Nonetheless, barriers to engagement with value-based activities may affect ACT efficacy. Behavioral activation (BA) may be an alternative for this shortcoming. Integrating these interventions has shown promise. However, no studies evaluated its effectiveness in this group. Therefore, evaluating the preliminary effectiveness and feasibility of this intervention was needed.

Objective: This study aimed to evaluate the feasibility measures. The secondary objective aimed to examine its effectiveness on cumulative abstinence duration, drinking days, drinks per drinking day, heavy drinking days, alcohol abstinence self-efficacy, readiness to change, psychological flexibility (Acceptance and Action Questionnaire—version II), and everyday discrimination score.

Methods: A pilot parallel randomized controlled trial was conducted. Forty young ethnic minorities who can speak English and have AUD based on *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]) criteria were recruited and randomly assigned to either Acceptance and Commitment Therapy with Behavioral Activation (ACT-BA; n=20) or Treatment-as-Usual (TAU; n=20) using a computer-generated random number; assessors were blinded to group assignment. Outcomes were collected at baseline (T0) and postintervention (T1). The analysis included descriptive statistics, independent samples *t* test, chi-square test, and generalized estimating equations with multiple imputations. To further supplement findings, a qualitative interview was conducted.

Results: A total of 148 ethnic minority young adults were screened, with an eligibility rate of 38.5% (57/148). Of these, 85.9% (49/57) consented, and among them, 81.6% (40/49) were randomized to either the ACT-BA or TAU. The retention rate was 82.5% (33/40), of which 75% (30/40) completed the postintervention assessment (ACT-BA: 16/20, 80%; TAU: 14/20, 70%). The adherence rate was 81.7% (4.9/6 sessions), and participants reported no adverse effects. Finally, 40 participants (20 for each group) were analyzed. The intervention group showed a promising improvement in drinking days (B=−4.12, 95% CI −8.10 to −0.13; P=.04, d=−0.57), drinks per drinking day (B=−1.56, 95% CI −3.06 to −0.07; P=.04, d=−1.89), alcohol abstinence self-efficacy (B=11.95, 95% CI 0.10-23.81; P=.048, d=0.81), and Acceptance and Action Questionnaire—version II (B=−6.41, 95% CI −12.77 to −0.06; P=.04, d=−0.65).

Conclusions: This study, unlike existing evidence, presents an innovative integration of ACT and BA delivered via an internet-based self-help format. The findings contribute to the field by providing preliminary evidence that this integrated intervention is feasible and promising for AUD. The main implication in the real world is to conduct a fully powered randomized controlled trial to further examine its effectiveness with longer follow-up to serve as a stand-alone treatment option for ethnic minorities.

Trial Registration: ClinicalTrials.gov NCT06779006; https://clinicaltrials.gov/study/NCT06779006

(J Med Internet Res 2026;28:e83896) doi:10.2196/83896

XSL•FO
RenderX

KEYWORDS

acceptance and commitment therapy; alcohol addiction; alcohol use disorder; behavioral activation; ethnic minority young adults

Introduction

Background

Alcohol use disorder (AUD) is a severe medical problem characterized by an inability to control alcohol consumption despite adverse physical, psychological, and social consequences [1-6]. Early alcohol consumption accelerates AUD development, leading to academic decline, injuries, and premature mortality among young adults aged 20 - 39 years, accounting for 13.5% of global alcohol-related deaths in this age group [7-15].

AUD remains a significant public health issue, disproportionately affecting ethnic minority young adults compared with the general population [16-18]. In Hong Kong, ethnic minorities constitute about 8.4% of the population [19,20]. However, unlike those with the majority ethnicity, that is, Hong Kong and/or Chinese populations, ethnic minorities encounter various challenges, including stigma, discrimination, acculturation stress, low self-esteem, social integration difficulties, and loneliness, contributing to unhealthy drinking behaviors and AUD [7,21-27]. Previous studies showed that over 50% of ethnic minorities have less access to education and employment than the Chinese population, and about 60% of Chinese respondents believed that ethnic minorities should be limited to low-skilled occupations [28]. Our previous empirical evidence further identifies cultural factors, family influences, hedonistic motives, curiosity, low-risk perception, coping motives, social pressures, and subjective cravings as reasons behind AUD development among young adult ethnic minorities [29].

In this regard, existing research has evaluated several psychosocial interventions for AUD [30-36], including motivational interviewing [30,37-39], integrated family and cognitive behavioral therapy [40], combined motivational enhancement therapy and cognitive behavioral therapy [41], common elements treatment approaches [42], guided self-change [35], and home-based ecologically based family therapy [36]. However, our systematic review and meta-analysis revealed limited efficacy of these approaches, highlighting the need for other, more effective treatments [43]. Consistent with this knowledge gap, our qualitative findings demonstrated that ethnic minority adults with AUD frequently engaged in experiential avoidance, using alcohol to cope with acculturation stress, anxiety, depression, and discrimination. This pattern is further corroborated by our cross-sectional study, which identified discrimination experiences and avoidance coping behaviors as significant contributors to AUD. Together, these maladaptive behaviors underscore the challenge in achieving a sustained behavioral change. Consequently, our findings indicate a need for third-wave therapies that promote acceptance of such negative internal experiences [44-47]. Third-wave therapies are a kind of psychosocial intervention focusing on accepting rather than changing negative thoughts and emotions [48].

Acceptance and commitment therapy (ACT) has emerged as a promising intervention for substance use and mental health disorders, enhancing psychological flexibility [45,49]. Grounded in Relational Frame Theory [48,50,51], ACT enhances conscious behavioral adaptation aligned with personal values while accepting negative thoughts and emotions [52]. Unlike traditional therapies, ACT focuses on behavioral change over symptom reduction, making it particularly suitable for avoidance-driven conditions such as AUD [48,51,53]. Literature review suggests that ACT may be a viable treatment option for ethnic minority young adults with AUD [47,54,55].

However, our review identified methodological limitations in existing studies on the effectiveness of ACT on AUD and substantial challenges in engagement and adherence to ACT [56-58]. First, many studies lacked appropriate control groups and failed to directly assess alcohol-related outcomes (eg, drinking days, drinks per drinking day, heavy drinking days, and abstinence duration). Second, for the adherence issues, participants showed low engagement with value-based activities for guiding committed action toward a meaningful life [56,57]. This is potentially attributed to difficulty in identifying enjoyable, rewarding alcohol-free activities, and overemphasis on internal experiences (motions or thoughts) without sufficient attention to overt behaviors. Finally, ACT does not contain any daily activity for practice, which is a critical challenge for ethnic minorities with AUD to identify pleasurable alternatives to replace drinking, leading to relapse despite quitting attempts. Consequently, these limitations warrant integrating novel strategies into ACT to enhance engagement and adherence.

Behavioral activation (BA) may be a possible method to overcome engagement and adherence problems in ACT. BA is a concept from behaviorism [59,60] and is developed from a concept that behavior can directly affect emotions for better or worse [59]. Hence, individuals can put themselves in a situation that is likely to foster positive emotions. For example, a positive event such as listening to uplifting music can lead to happiness and vice versa. BA is a widely recognized and theoretically sound approach that has been applied to treat depression [61,62]. Its application has been extended to substance use disorders (SUDs) via providing rewarding experiences in daily life, different from substance use, thus achieving abstinence and preventing relapse [63-65]. Evidence suggests that BA is a promising intervention for SUD as a stand-alone treatment or being integrated with other therapeutic modalities. The emphasis on rewarding experiences in BA exactly addresses the low engagement and adherence of ACT due to the failure to identify enjoyable, rewarding, and positively reinforcing activities in the intervention process. Because of the distinctive benefits of BA, it appears to be a treatment component complementary to the shortcomings of ACT, thus enhancing the intervention effectiveness if they both are integrated into the treatment protocol. Also, BA is a brief intervention that can be delivered within 4 weeks, which will not significantly prolong the duration of each intervention session, ensuring the participants' adherence and engagement. In sum, ACT simply encourages an individual to ACT, which includes *A*ccepting the reactions and being present, *C*hoosing a valued direction, and *T*ake actions, while

XSL•FO
RenderX

BA motivates an individual to take real actions. Two studies have integrated BA into ACT to manage different conditions in various populations, including an internet-based BA and acceptance-based treatment for depression [66], and an open trial of a new acceptance-based behavioral treatment for major depression and psychotic features [64]. The results were promising, which supported that the integration is not only theoretically sound but also feasible and, importantly, can translate into observable benefits.

Regarding the mode of intervention delivery, our qualitative findings revealed significant treatment barriers among ethnic minority young adults, in which no participant had accessed treatments for AUD due to their engagement in work. Apart from work, their time and financial constraints also hindered them from seeking hospital-based care [26]. Internet-based interventions may be a viable option to address these challenges by enhancing accessibility [67-69], reducing stigma [70], minimizing work disruption [71], and improving cost-effectiveness [68,72,73]. The ACT's efficacy for depression further supports this modality [72,74]. However, no studies have evaluated internet-based self-help interventions integrating ACT and BA for these populations. Therefore, this study examined the preliminary effectiveness and feasibility of an internet-based self-help–integrated ACT and BA intervention in improving alcohol use abstinence among ethnic minority young adults with AUD.

Theoretical Framework

The integration of ACT and BA is grounded in principles of functional contextualism [75]. ACT derives from Relational Frame Theory [76,77]. ACT enhances psychological flexibility through the 6 ACT core processes (ie, acceptance, cognitive defusion, being present, self as context, values, and committed action). These processes represent 6 facets of 1 diamond, and the diamond itself is psychological flexibility. Psychological flexibility is defined as the capacity to persist or change one's behavior to be open and accepting of personal thoughts and feelings, to appreciate and adapt to different situations, and respond to situations that facilitate one's values and goals [51,76]. Each process contributes to psychological flexibility for AUD treatment as follows: (1) *Being present*: Individuals with AUD often drink alcohol in response to cravings and emotional triggers. By learning the present moment, they can mindfully notice these triggers and create a space for acceptance rather than avoidance. (2) *Acceptance*: This involves opening and making room for such painful feelings, sensations, urges, and emotions that trigger drinking. Instead of struggling with these experiences, individuals learn to allow them to come and go, just like cars driving past outside our house, reducing the urge to escape through drinking. (3) *Cognitive defusion*: Individuals with AUD may be hooked on thoughts like "I need a drink to cope." Defusion helps them step back and watch thoughts for what they are—nothing more or less than words or pictures. (4) *Self-as-context*: This process assists individuals in recognizing that they are more than their thoughts and feelings and makes individuals aware of whatever they are thinking, feeling, and sensing. (5) *Values*: AUD leads individuals away from what is important to them. Clarifying values helps identify what matters to them in their lives. (6) *Committed action*: This is taking effective actions, guided by our values. For people with AUD, it refers to committing to value-congruent activities despite pain and discomfort.

BA is derived from the behavioral theory of substance use [78]. According to this theory, substance use persists partly because of the limited access to competing rewarding activities and a lack of positive reinforcement from behaviors that do not involve substance use. To tackle AUD by BA, individuals have to identify adaptive alternatives to drinking that provide comparable rewards (eg, exercise for stress relief) without adverse consequences. The integration of ACT and BA provides a synergetic efficacy. ACT encourages patients to change their relationship with triggering factors rather than avoiding them, and creates a metacognitive space to tolerate distress, for example, acceptance of cravings [77], while BA complements ACT by increasing engagement in value-congruent activities that compete with drinking [75] and provides a structured method (eg, scheduling exercises and social events) to modify external behavior, that is, alcohol use [79]. For example, a client who values health and connection might replace bar visits with an alcohol-free gathering. This integrated approach targets both internal and external factors that drive AUD [75]. ACT equips individuals with skills to accept negative emotions, whereas BA provides a concrete strategy to disrupt reinforcement cycles and build healthy habits. Together, they provide a sustained psychological flexibility [75] (Figure 1,1).

XSL•FO
RenderX

Figure 1. Theoretical framework for integrated acceptance and commitment therapy and behavioral activation interventions. ACT: acceptance and commitment therapy; BA: behavioral activation.

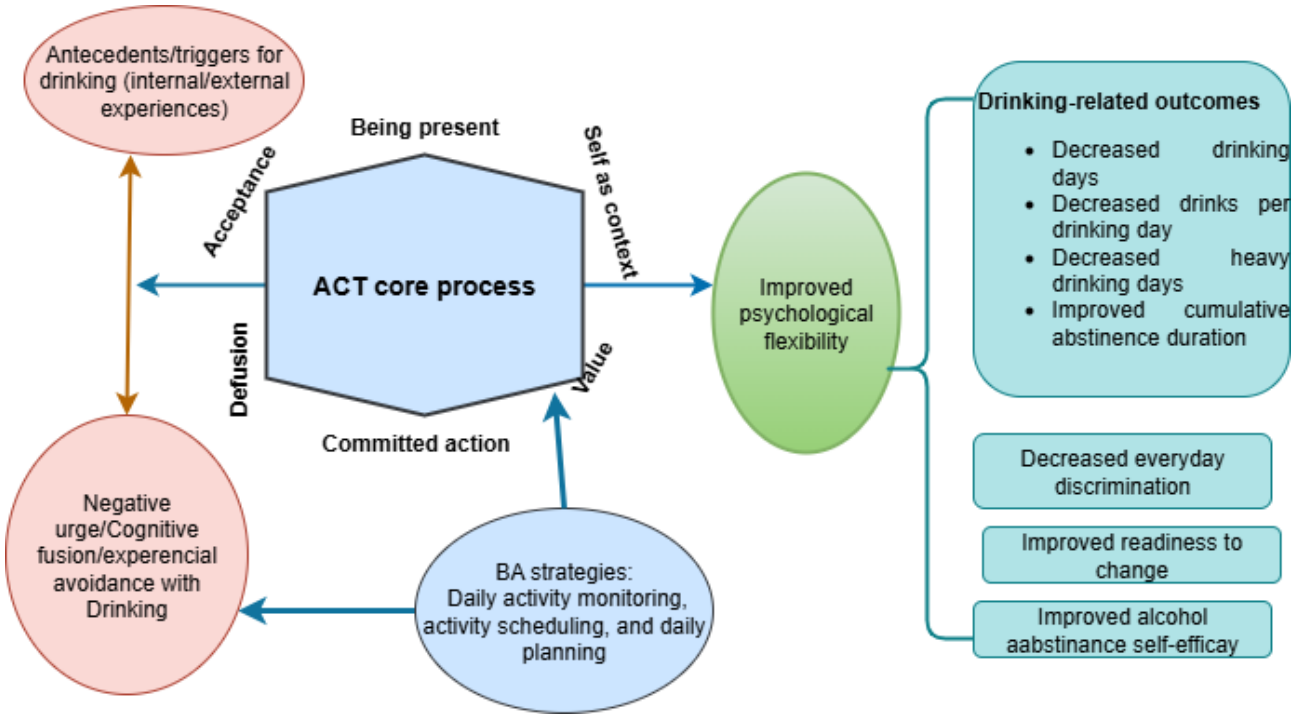

Objectives

This study examined the preliminary effectiveness and feasibility of an internet-based self-help program integrating ACT and BA on alcohol abstinence among Hong Kong ethnic minority young adults with AUD. The primary objective of this study was to evaluate the intervention's feasibility in terms of various feasibility outcomes under CONSORT (Consolidated Standards of Reporting Trials) 2025, including rates of eligibility, consent, randomization, adherence, retention, completion, missing data, and adverse events. The secondary objective was (1) to evaluate the preliminary effectiveness of this internet-based self-help program integrating ACT and BA on several outcomes, including drinking days, drinks per drinking day, heavy drinking days, cumulative abstinence duration (CAD), alcohol abstinence self-efficacy (AASE), psychological flexibility, readiness to change, and everyday discrimination in this population; and (2) to estimate the effect size of the Acceptance and Commitment Therapy with Behavioral Activation (ACT-BA) intervention on these same outcomes.

Methods

Patient and Public Involvement

Due to limited time and the online mode of self-help intervention, patients were not involved in designing and developing the research question, nor in reporting the findings. However, experts participated in developing and reviewing the intervention protocol.

Trial Design

A single-blinded randomized controlled trial (RCT) was conducted among Hong Kong ethnic minority young adults with AUD from September 15, 2024, to May 15, 2025. This study was used according to the protocol. There were no changes to the trial design, participants, analysis, and outcomes after the study commenced.

Trial Setting

This study was conducted in Hong Kong, a special Chinese administrative region located east of the Pearl River Delta. In Hong Kong, approximately 7.5 million people live on a land area of 1104 square kilometers [80], of which about 619,552 (8.4%) are ethnic minorities [19].

Study Participants

Eligibility Criteria

Participants who met the following criteria were considered for this study: (1) young adults (aged 18-35 years) from ethnic minorities in Hong Kong, (2) proficient in English (read, write, and speak), (3) voluntary participants, (4) possessing internet-enabled smartphones or electronic devices, and (5) meeting the *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]) criteria for AUD, defined as exhibiting at least 2 of the 11 symptoms specified in the DSM-5 [81]. Ethnic minorities in this age range were chosen as it aligns with the US Census Bureau classification for young adults [82], and this age group is the population group with the highest vulnerability for SUDs [83].

Exclusion Criteria

Participants were excluded if they met any of the following criteria: (1) they were currently experiencing other psychiatric and severe mental health problems, or actively involved in suicide attempts; (2) planned to leave Hong Kong during the intervention period; (3) had no and/or lack of access to a smartphone and/or other electronic devices; and (4) had prior engagement in ACT and/or BA for AUD.

XSL•FO
RenderX

Sample Size

For this pilot study, a sample size of 40 participants (20 per group) was determined through methodological and practical considerations. This is consistent with guidelines suggesting 12 - 30 participants per group as adequate for generating meaningful preliminary data [84]. Previous evidence from comparable studies demonstrates feasibility and provides preliminary effect size estimation with similar sample sizes [57,85]. As it was a pilot pre-post study, the final analysis was performed once the full sample size was reached; an interim analysis was not planned.

Randomization

Stratified block randomization was implemented by a trained research assistant (RA) to allocate participants into intervention and control groups (20 per group). To ensure baseline balance of prognostic factors, particularly ethnicity, sex, and disease severity, which could confound the intervention effects if unevenly distributed, participants were first stratified by these factors. Within each stratum, block randomization (1:1 ratio, block size of 4) was performed using a computer-based random number generator.

Allocation Concealment and Blinding

Allocation sequences were concealed using a computer-generated random number. Due to the nature of the intervention and self-administered questionnaires, blinding of participants was not feasible. However, the outcome assessor was blinded to the participants' group assignments throughout the study.

Recruitment and Data Collection

An RA was responsible for participant recruitment. The RA received a 1-day training regarding study details and data collection procedures from the principal investigator. Participants were recruited in the community through face-to-face approaches at bars, communal recreational areas, schools, and gardens where ethnic minorities usually gather. The RA approached potential ethnic minorities, introduced the study, and gave them a poster to read the details, objectives, and eligibility criteria of the study. Interested ethnic minorities were instructed to scan the QR code and complete a screening in a Google form to determine their eligibility. A brief session was conducted to guide the participants to access the intervention using their smart devices during participant recruitment. Eligible ethnic minorities who agreed to participate gave written consent. The participants' contact addresses, including email, WhatsApp number, WeChat number, and/or Telegram account, were taken for further connections.

Participants were then provided a personal link to access the Qualtrics website. They were provided instructions to complete baseline data. Consequently, participants were randomly assigned to either the ACT-BA group or the Treatment-as-Usual (TAU) group. Quantitative data were collected at 2 time points, that is, baseline and postintervention. Then, a semistructured interview was conducted among the intervention group after the postintervention assessment.

Semistructured Interview

The interviewer contacted all participants in the intervention group for the semistructured interview. However, only 13 participants consented and were invited to the interview immediately after the postintervention assessment. Data saturation was achieved upon interviewing the 13th participant; this was the point at which no additional themes were identified. A male research student conducted interviews face-to-face and/or by phone using a semistructured interview guide on four areas: (1) general experience, (2) intervention content, (3) delivery mode and duration, and (4) perceived usefulness. The interview took an average of 15 minutes. Findings might provide pertinent and supplementary data to enhance the understanding of the intervention feasibility and effectiveness.

Intervention

The intervention group received an internet-based self-help program that integrated both ACT and BA, using an online platform in Qualtrics, through a personalized link. The intervention contained a total of 6 weekly individual sessions structured sequentially. An RA in the team sent text reminders to the participants via different information communication systems, for example, WhatsApp and WeChat, 1 day before the intervention session to ensure their attendance. For session 1, it aimed to establish the therapeutic relationship with the participants and assess the participants' experiential avoidance behaviors. In this session, participants learned through multimedia resources what avoidance behavior was; how it impacted their life in terms of different perspectives, including family, health, work, social life, and the community; and withdrawal symptoms. They also identified alcohol-free rewarding activities and initiated daily monitoring through rating enjoyment or importance (from 1 to 10) in a workbook. High-scoring activities were considered for further value-driven planning. Session 2 focused on value identification, in which participants learned and identified core values across life domains and developed corresponding value-driven activity plans using a structured worksheet. They scheduled at least 1 value-congruent activity per domain. Session 3 introduced the concept of acceptance and cognitive defusion. Participants learned and practiced various acceptance and diffusion skills and mindfulness exercises to manage experiential avoidance, which was usually managed by drinking. They learned different techniques of acceptance by (1) letting feelings, urges, and thoughts occur without a desire to act on them; (2) noticing their strengths while acknowledging their shortcomings; and (3) facing the difficulty without running away from or avoiding it. They also learned defusion skills by considering their thoughts as thoughts only that could come and go, and sometimes their impulsive thoughts about drinking were not facts. Session 4 focused on present moment awareness via mindful practices. Participants learned to observe internal experiences (such as cravings, negative emotions, and thoughts) related to past or future events that led to AUD. Observing self-skills was also developed through metaphors. Particularly, the participants learned how to consider themselves and hence understood that there was a part of them who could think, feel, notice, and observe their cravings and feelings toward drinking. Session 5 reinforced value-directed BA, with participants identifying new

XSL•FO
RenderX

value-congruent alcohol-free activities and integrating them into actionable daily plans and taking committed action toward their life goals. Session 6 emphasized the posttreatment plan and relapse prevention strategies. Previous sessions were reviewed, and another long-term goal congruent with their values was developed. However, participants have not received any additional concomitant care or intervention outside of the study protocol from the researcher.

Control Group

The control group, referred to as TAU, received standard or usual community care without any active interventions from the study. Particularly, TAU did not contain any standardized, structured psychoeducation, counseling, or concomitant care from the researcher. However, participants continued to receive routine health care, self-help, or any community services that they had already engaged in, and they were asked to report the received services to the research team during the study period. Participants were also provided referrals to specialist care when needed. In previous studies, such standards have been used as a benchmark for comparison with ACT [86-88].

Intervention Integrity

The intervention integrity was maintained in several ways. First, the same RA was responsible for updating, reminding, and sending the intervention link to the participants. Second, the RA was trained for 7 days on how to manage online intervention using Qualtrics and follow the intervention process. Third, the intervention content was adapted from the empirically validated ACT manual and BA strategies, with expert validation done by a panel that contained 2 associate professors, experienced in research and delivery of different psychosocial interventions, particularly ACT, an assistant professor with experience in research of addiction and substance abuse, and 3 social workers in an NGO providing services for ethnic minorities in Hong Kong. Fourth, regular supervision and process monitoring were conducted to ensure the smoothness of the implementation. Fifth, all procedures of implementation were systematically documented for transparency.

Contamination

Study participants were recruited at various points, and the intervention was provided online through each participant's account. Participants should not know each other to exchange information. Therefore, the possibility of information sharing seemed unlikely in this study.

Outcomes

Feasibility Outcomes

The feasibility of this study was assessed by evaluating eligibility rate (number of eligible participants divided by number of screened ethnic minorities), consent rate (number of participants consented divided by number of eligible participants), randomization rate (number of participants who were randomized divided by number of participants who consented), adherence rate (number of participants in the intervention group who completed all intervention sessions divided by the number of participants randomized into the intervention group), retention rate (number of participants who remained in the study divided by the number of participants who were randomized), completion rate (number of participants who answered the questionnaires divided by the number of questionnaires being distributed), adverse events that were defined as unfavorable and unintended events that were absent from baseline or appeared to worsen from baseline during the study period, and missing data (the percentage of missing in the dataset).

Effectiveness Outcomes

Drinking-Related Outcomes

The Timeline Follow-Back Questionnaire was used to assess the following drinking-related outcomes, including drinking days, drinks per drinking day, heavy drinking days, and CAD. Every participant was asked to report the number of standard drinks (1 unit contains 12 grams of pure alcohol) consumed within 30 days pre- and postintervention. Then, the average number of drinks per drinking day, numbers of drinking days and heavy drinking days, and cumulative abstinence were computed. Heavy drinking day is defined as having 4 or more drinks for females, and 5 or more drinks on any day for males [89]. Abstinence refers to the total number of days a participant abstains from alcohol use postintervention [87]. The CAD was calculated by dividing the number of abstinent days by 30 days.

Alcohol Abstinence Self-Efficacy

The self-efficacy of participants in alcohol abstinence was measured using the AASE Scale. This scale contains 20 items rated on a 5-point Likert scale, ranging from 0 to 80. Higher scores indicate high self-efficacy to abstain from alcohol use. In this study, the Cronbach α value of this scale was 0.91.

Readiness to Change

The participant's readiness to change in drinking was measured using the Readiness to Change Questionnaire, a 12-item instrument [90]. The internal consistency of the scale was good.

Psychological Flexibility

The Acceptance and Action Questionnaire—version II was used to measure psychological flexibility. This questionnaire contains 7 items rated on a 7-point rating scale, from 1 (never true) to 7 (always true) [91]. Higher scores indicate greater levels of psychological inflexibility. This scale showed a good internal consistency (Cronbach α=0.779) in our previous studies conducted among ethnic minority young adults in Hong Kong [92].

Everyday Discrimination

The participants' everyday discrimination was measured using the short version of the Everyday Discrimination Scale [93]. The scale comprises 5 items, rated on a 6-point scale, ranging from 5 to 36 (1=almost never, 6=almost always). They also identified perceived causes of discrimination from options including religion, ethnicity, age, gender, sexuality, disability, or an open-ended category, that is "other." A higher score on the Everyday Discrimination Scale indicates a frequent experience of discrimination. This scale revealed an excellent internal consistency in our previous cross-sectional study, with a Cronbach α of 0.963 [92].

XSL•FO
RenderX

Statistical Methods

The data were analyzed using the IBM SPSS (version 23; IBM Corp). Descriptive statistics were used to summarize the social-demographic characteristics of participants and the feasibility measures, including eligibility rate, consent rate, randomization rate, adherence rate, retention rate, completion rate, percentage of missing data, and adverse events. The baseline and posttest outcome variables were checked for normality assumption and the homogeneity of variance using the Shapiro-Wilk and Levene tests, respectively. Mean and SD were used for continuous variables with a normal distribution, while medians and IQRs were used for non–normal distributions. Comparability of participants between groups was examined by independent samples *t* test or Mann-Whitney *U* test for continuous variables, and chi-square test or Fisher exact test for categorical variables. Additionally, paired *t* test (for normally distributed data) and Wilcoxon signed-rank test (non–normally distributed data) were used to examine within-participant mean differences.

Missing data were handled using intention-to-treat (ITT) analysis, considering all participants enrolled and randomized to either 1 of the 2 groups. Little Missing Completely at Random (MCAR) test was used to determine whether the missingness was completely at random or not. Sensitive analyses were also conducted to compare results from per-protocol (PP) (complete-case data) and ITT analyses.

To compare pre-and posttest outcomes, we used the Generalized Estimating Equations (GEE) under the assumption that data missing were completely at random. Additionally, an independent samples *t* test was performed as a supplementary analysis to evaluate a pre-post change in outcomes. However, to preserve statistical power and reduce bias under the assumption that data were missing at random, GEE with multiple imputations (MIs) was used due to the presence of substantial missing data (>5% missingness in posttest measures). The pooled estimates of MIs were used to compute the analysis. If the data were missing at random, a weighted GEE analysis was used, and the effect of time × group assignment interaction on outcome variables was determined.

Parameter estimations for signs of instability were calculated using the quasi-likelihood estimation method integrated within the GEE procedure of IBM SPSS. Quasi-Information Criterion values were used to select the appropriate correlation structure. Model convergence was examined across all imputations and found that they converged successfully without warnings or errors. Intervention effects were estimated based on the difference in pre-post difference using the estimation-based approach for multiple comparisons, that is, we emphasized 95% CI and effect sizes rather than relying solely on *P* values to assess the precision and clinical significance of the intervention effects. The effect size was calculated to assess the clinical significance of the outcomes for between-group comparisons of change scores and was interpreted using Cohen *d*, as follows: small=0.2, moderate=0.5, and large=0.8. Finally, the findings were reported following the CONSORT 2025 guideline [94].

Qualitative Data Analysis

Regarding the qualitative data, a thematic method of analysis was used. First, all audio recordings were transcribed into English verbatim. A descriptive phenomenological design was chosen to explore the experiences of participants regarding the intervention. Two researchers (GMB and KYH) carefully read transcripts and listened to audio multiple times to understand the participants' experiences. They independently highlighted the key quotes and labeled them with codes. Any disagreement between coders was resolved through discussion until a consensus was reached. Senior researchers (KYH and YWM) were consulted if a consensus was not reached. Similar codes were grouped into subthemes and eventually themes, and all 13 participants reviewed the findings to ensure that the findings matched with their experiences. Finally, the results were reported following the COREQ (Consolidated Criteria for Reporting Qualitative Research) [95].

Ethical Considerations

Ethical approval was obtained from the Human Subjects Ethics Subcommittee of the Hong Kong Polytechnic University (reference number HSEARS20240712003). The protocol of this study was submitted on August 28, 2024, and was subsequently published on January 15, 2025, on ClinicalTrials.gov with National Clinical Trial number NCT06779006. This record is available online at ClinicalTrials.gov. Informed written consent was obtained from each participant before the intervention. To ensure confidentiality and privacy, all participants' data were kept confidential, anonymous, and stored on a secure password-protected server; no personally identifiable information was used in the study. Each participant has been compensated for costs to their time and internet usage; that is, each of them received a 300 Hong Kong dollars (approximately US $38.46) gift card with an exchange rate of 1 US $=7.8 Hong Kong dollars. Finally, there were no supplementary materials or images containing identifiable information for participants in the manuscript. They were informed of their right to decline or withdraw from the study at any time.

Results

Participant Characteristics

The study participants' sociodemographic characteristics are presented in Table 1. Participants had a mean age of 32.43 (SD 2.57) years, with 80% (32/40) female and 87% (35/40) as Filipino; 85% (34/40) were Catholic, 52.5% (21/40) were single, and 85% (34/40) were employed. Of those being employed, 87.5% (35/40) were domestic helpers. Approximately 60% (24/40) attained primary or secondary education, and 32.5% (13/40) cohabited with partners. The median duration of residence in Hong Kong was 5 years, with an IQR of 2 - 6. Regarding the AUD severity, 47.5% (19/40) had moderate AUD, 32.5% (13/40) had severe AUD, and 20% (8/40) had mild AUD. Baseline comparisons between groups revealed no significant differences in demographics and the AUD severity.

XSL•FO
RenderX

Table . Baseline sociodemographic characteristics of ethnic minority young adults with AUD[a] in Hong Kong (N=40): a pilot randomized controlled trial.

Variables	ACT-BA[b] (n=20)	TAU[c] (n=20)	Total (N=40)	*P* value
Age (years) mean (SD)	32.6 (1.98)	32.25 (3.09)	32.43 (2.57)	.68[d]
Sex, n (%)				
Female	16 (40)	16 (40)	32 (80)	≥.99[e]
Male	4 (10)	4 (10)	8 (20)	
Ethnicity, n (%)				
Filipino	17 (42.5)	18 (45)	35 (87.5)	≥.99[e]
Whites	1 (2.5)	1 (2.5)	2 (5)	
African	2 (5)	1 (2.5)	3 (7.5)	
Religion, n (%)				
Catholic	15 (37.5)	19 (47.5)	34 (85)	.25[e]
Orthodox Christians	3 (7.5)	1 (2.5)	4 (10)	
Others	2 (5)	0 (0)	2 (5)	
Marital status, n (%)				
Single	12 (30)	9 (22.5)	21 (52.5)	.40[e]
Married/cohabiting	4 (10)	8 (20)	12 (30)	
Separated/widowed	4 (10)	3 (7.5)	7 (17.5)	
Educational status, n (%)				
≤12 grades (below senior secondary)	11 (27.5)	13 (32.5)	24 (60)	.52[f]
College and above	9 (22.5)	7 (17.5)	16 (40)	
Employment, n (%)				
Unemployed	3 (7.5)	3 (7.5)	6 (15)	≥.99[e]
Employed	17 (42.5)	17 (42.5)	34 (85)	
Occupation, n (%)				
Domestic helper	17 (42.5)	18 (45)	35 (87.5)	≥.99[e]
Daily laborer/others	3 (7.5)	2 (5)	5 (12.5)	
Living arrangements, n (%)				
Living alone	5 (12.5)	1 (2.5)	6 (15)	.25[e]
Living with parents/relatives	4 (10)	5 (12.5)	9 (22.5)	
Living with nonrelatives	7 (17.5)	5 (12.5)	12 (30)	
Living with partners	4 (10)	9 (22.5)	13 (32.5)	
Duration of residence in Hong Kong				
Years, median (IQR)	4.5 (4)	5 (5)	5 (4)	.50[g]
DSM-5[h] criteria for AUD, n (%)				
Mild AUD	5 (12.5)	3 (7.5)	8 (20)	.75[e]
Moderate AUD	8 (20)	11(27.5)	19 (47.5)	
Severe AUD	7 (17.5)	6 (15)	13 (32.5)	

XSL•FO
RenderX

[a]AUD: alcohol use disorder.
[b]ACT-BA: Acceptance and Commitment Therapy with Behavioral Activation.
[c]TAU: Treatment-as-Usual.
[d]Independent samples *t* test.
[e]Fisher exact test.
[f]Chi-square test.
[g]Mann-Whitney *U* test.
[h]*DSM-5*: *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition).

Baseline Data

Independent samples *t* test was used to compare the baseline outcomes between the 2 groups (Shapiro-Wilk test, *P*>.05; Levene test, *P*>.05). There was no significant difference between the intervention and control groups regarding the baseline outcome measures (Table 2).

Table . Baseline outcome measure of ethnic minority young adults with AUD[a] in Hong Kong: a pilot randomized controlled trial between groups, ACT-BA[b] (n=20) vs TAU[c] (n=20).

Baseline outcome	Total (N=40), mean (SD)	ACT-BA (n=20), mean (SD)	TAU (n=20), mean (SD)	Mean difference (95% CI)	*P* value
AASE[d]	47 (14.78)	44.45 (10.79)	49.55 (17.85)	−5.1 (−14.54 to 4.34)	.28
AAQ-II[e]	25.27 (9.83)	24.45 (9.01)	26.1 (10.76)	−1.65 (−8.00 to 4.70)	.60
DDD[f]	5.09 (0.71)	5.16 (0.75)	5.03 (0.68)	.13 (−0.33 to 0.59)	.57
EDS[g]	20.93 (8.50)	19.5 (7.75)	22.35 (9.15)	2.38 (−6.37 to 3.95)	.30
RTCQ[h]	35.88 (9.69)	37.55(8.26)	34.20 (10.90)	3.35 (−2.84 to 9.54)	.46
DD[i]	9.73 (7.20)	9.35 (6.93)	10.10 (7.62)	−0.75 (−5.41 to 3.91)	.67
HDD[j]	6.28 (4.19)	6.55 (5.38)	6.00(2.64)	0.55 (−2.16 to 3.26)	.35
CAD[k]	20.27 (7.20)	20.65 (6.93)	19.90 (7.62)	0.75 (−3.91 to 5.41)	.67

[a]AUD: alcohol use disorder.
[b]ACT-BA: Acceptance and Commitment Therapy with Behavioral Activation.
[c]TAU: Treatment-as-Usual.
[d]AASE: Alcohol Abstinence Efficacy Scale.
[e]AAQ-II: Acceptance and Action Questionnaire—version II.
[f]DDD: drinks per drinking day in the past 30 days.
[g]EDS: Everyday Discrimination Scale.
[h]RTCQ: Readiness to Change Questionnaires.
[i]DD: drinks per day in the past 30 days.
[j]HDD: heavy drinking days in the past 30 days.
[k]CAD: cumulative abstinence durations.

Feasibility of Intervention

Figure 2 presents the CONSORT 2025 flow diagram. A total of 148 ethnic minority young adults who drank alcohol were approached from the community. After screening, 38.5% (57/148) met the eligibility criteria. Of them, 85.96% (49/57) consented to join the study. Among these, 81.6% (40/49) were randomly assigned either to the ACT-BA (n=20) or TAU (n=20) group. The retention rate was 82.5% (33/40), with 85% (17/20) in the ACT-BA group and 80% (16/20) in the TAU group. The attrition rate was 15% (3/20) for ACT-BA and 20% (4/20) for TAU, but there was no significant difference in attrition between groups (*P*=.47). Of those who remained in the study, 30 completed the postintervention test, providing a completion rate of 75% (30/40), with 80% (16/20) in the ACT-BA group and 70% (14/20) in the TAU group (Table 3). Participants in the ACT-BA group completed an average of 4.9 out of 6 sessions (adherence rate: 81.7%). No adverse events were reported at postintervention. The trial was not ended or stopped early. The median session duration was 30 (IQR 35) minutes. The median interval between sessions was 7 (IQR 2) days.

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) 2025 participants flow diagram. The process through the phases of a randomized trial of 2 groups (ie, enrollment, intervention allocation, follow-up, and data analysis).

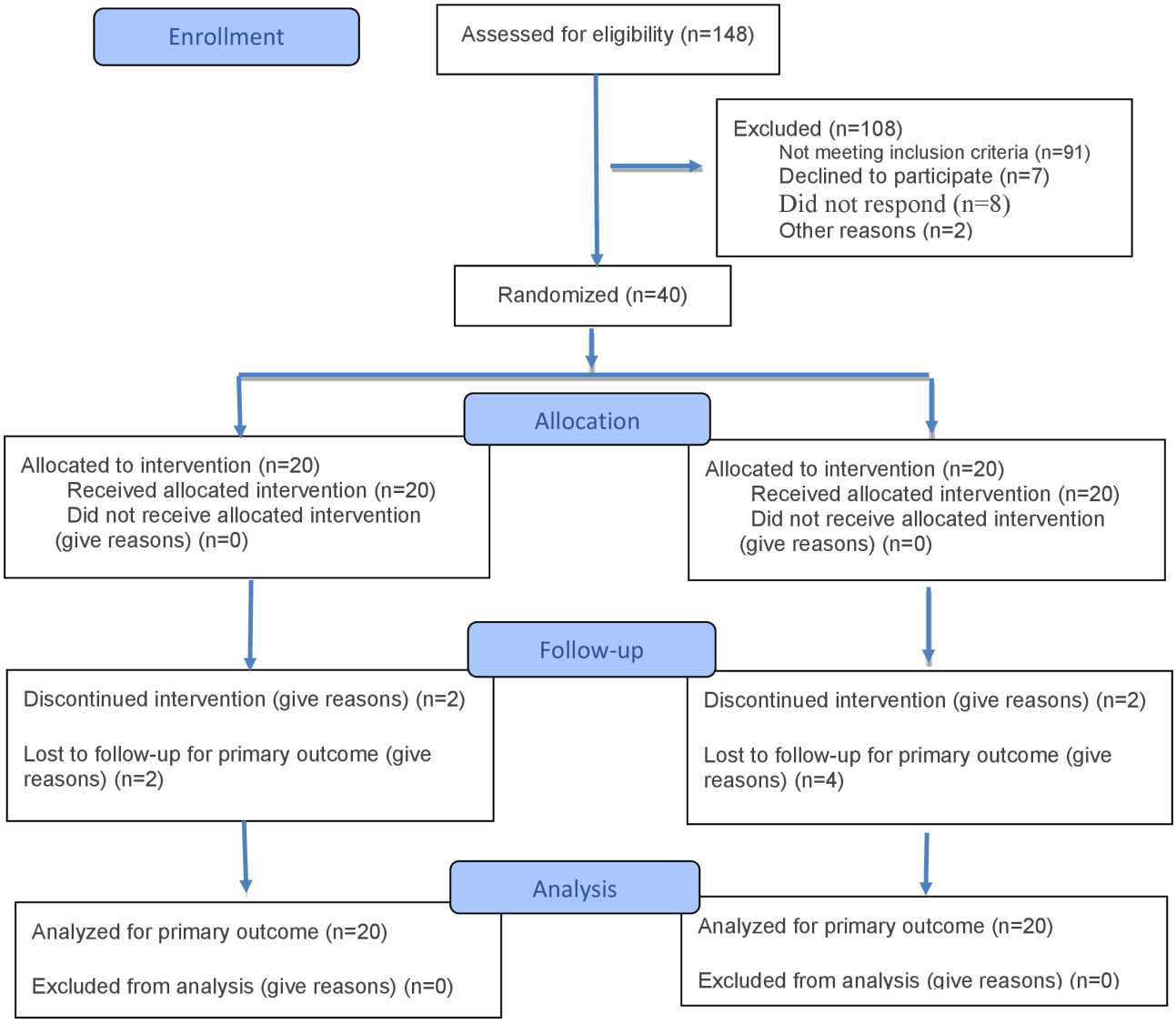

Table . Adherence to intervention among ethnic minority young adults with alcohol use disorder in Hong Kong: a pilot randomized controlled trial.

Sessions	ACT-BA[a] group (n=20)		TAU[b] group (n=20)	
	Access session, n (%)	Completed session, n (%)	Access session, n (%)	Completed session, n (%)
Week 1	20 (100)	18 (90)	20 (100)	17 (85)
Week 2	18 (90)	17 (85)	16 (80)	15 (75)
Week 3	16 (80)	15 (75)	17 (85)	13 (65)
Week 4	17 (85)	17 (85)	16 (80)	15 (75)
Week 5	17 (85)	15 (75)	14 (70)	14 (70)
Week 6	17 (80)	16 (80)	15 (75)	14 (70)

[a]ACT-BA: Acceptance and Commitment Therapy with Behavioral activation; average session completion rate=81.70%.

[b]TAU: Treatment-as-Usual; average session completion rate=73.33%.

Missing Data Handling

This study had a total of 5.6% missing values. Little's MCAR test indicated that the pattern of missing data followed the MCAR test ($\chi^{2}{}^{10}$=0.000; P≥.99). However, due to substantial missingness (>5% missingness in posttest measures), MIs were used to preserve the statistical power and reduce bias, assuming that data were missing at random. Using SPSS, 10 imputed datasets were generated to calculate the pooled estimates. The results from the ITT analysis were compared with the PP analysis.

XSL•FO
RenderX

Effectiveness of Intervention

Effectiveness of ACT-BA Intervention on Drinking-Related Outcomes

The ACT-BA intervention showed a significant group-by-time interaction effect on various drinking outcomes. Compared with the TAU group, participants in the intervention group showed a statistically significant reduction in drinking days (ITT: Group × time, B=−4.12, 95% CI −8.10 to −0.13, Wald χ^2_4=4.224, P=.04; PP: Group × time, B=−4.02, 95% CI −7.50 to −0.54, Wald χ^2_1=5.122, P=.02) and drinks per drinking day (ITT: Group × time, B=−1.56, 95% CI −3.06 to −0.07, Wald χ^2_4=4.298, P=.04; PP: Group × time, B=−1.31, 95% CI −2.60 to −0.02, Wald χ^2_1=3.956, P=.047) at postintervention. The ACT-BA intervention provided a moderate effect size on drinking days (Cohen d=−0.57) and a large effect size on drinks per drinking day (Cohen d=−1.89; Table 4). Although a statistically significant difference was not reached for CAD (ITT: Group × time, B=2.34, 95% CI −3.68 to 8.36, Wald χ^2_4=0.634, P=.43; PP: Group × time, B=2.36, 95% CI −2.45 to 7.17, Wald χ^2_1=.925, P=.34) and heavy drinking days (ITT: Group × time, B=−3.01, 95% CI −6.37 to −0.36, Wald χ^2_4=3.135, P=.08; PP: Group × time, B=−3.58, 95% CI −7.19 to 0.02, Wald χ^2_1=3.794, P=.05) at postintervention assessment (Table 5), the ACT-BA group showed improvement with small to moderate (Cohen d=0.32) and moderate to large (Cohen d=−0.71) effect sizes on these outcomes, respectively.

XSL•FO
RenderX

Table . Mean score and difference of the outcome variables for ACT-BA[a] and TAU[b] among ethnic minority young adults with AUD[c] in Hong Kong across time points (pre and post; N=40): a pilot randomized controlled trial.

Outcomes	ACT-BA (n=20)			TAU (n=20)			Effect size
	Pretest, mean (SD)	Posttest, mean (SD)	Mean difference (SE)	Pretest, mean (SD)	Posttest, mean (SD)	Mean difference (SE)	Cohen *d*
AAES[d]							
PP[e]	44.45 (10.79)	58.69 (12.67)	14.24 (3.91)	49.55 (17.85)	51.11 (10.97)	1.56 (5.38)	0.83
ITT[f]	44.45 (10.79)	55.97 (17.39)	11.52 (4.58)	49.55 (17.85)	49.12 (13.95)	−0.43 (5.19)	0.81
AAQ-II[g]							
PP	24.45 (9.01)	14.31 (4.89)	−10.13 (2.51)	26.10 (10.76)	22.93 (5.33)	−3.17 (3.12)	−0.7
ITT	24.45 (9.01)	15.38 (6.31)	−9.07 (2.46)	26.10 (10.76)	23.44 (9.03)	−2.66 (3.15)	−0.65
RTCQ[h]							
PP	37.55 (8.26)	42.88 (6.55)	5.33 (2.53)	34.20 (10.90)	36.86 (12.62)	2.66 (4.05)	0.28
ITT	37.55 (8.26)	42.00 (7.69)	4.45 (2.52)	34.20 (10.90)	35.94 (13.64)	1.74 (3.90)	0.28
EDS[i]							
PP	19.50 (7.76)	11.31 (7.47)	−8.19 (2.56)	22.35 (9.16)	18.79 (10.34)	−3.56 (3.36)	−0.55
ITT	19.50 (7.76)	12.84 (9.03)	−6.66 (2.66)	22.35 (9.16)	19.73 (10.96)	−2.62 (3.30)	−0.48
DD[j]							
PP	9.35 (6.93)	5.56 (4.40)	−3.79 (1.99)	10.10 (7.62)	11.14 (7.81)	1.02 (2.68)	−0.66
ITT	9.35 (6.93)	5.98 (7.78)	−3.37 (2.14)	10.10 (7.62)	10.84 (9.44)	0.75 (2.71)	−0.57
DDD[k]							
PP	5.16 (0.75)	2.13 (1.45)	−3.03 (0.37)	5.03 (0.68)	3.32 (1.90)	−1.71 (0.46)	−1.84
ITT	5.16 (0.75)	2.45 (1.70)	−2.71 (0.42)	5.03 (0.68)	3.88 (2.72)	−1.15 (0.63)	−1.89
HDD[l]							
PP	6.55 (5.38)	2.88 (5.39)	−3.68 (1.81)	6.00 (2.64)	5.93 (6.31)	−0.71 (1.57)	−0.70
ITT	6.55 (5.38)	3.12 (5.99)	−3.43 (1.80)	6.00 (2.64)	5.57 (7.47)	−0.43 (1.77)	−0.71
CAD[m]							
PP	20.65 (6.93)	22.56 (7.30)	1.91 (2.38)	19.90 (7.62)	18.86 (7.81)	−1.04 (2.68)	0.41
ITT	20.65 (6.93)	22.52 (8.93)	1.87 (2.53)	19.90 (7.62)	19.43 (9.12)	−0.46 (2.65)	0.32

[a]ACT-BA: Acceptance and Commitment Therapy with Behavioral activation.
[b]TAU: Treatment-as-Usual.
[c]AUD: alcohol use disorder.
[d]AAES: Alcohol Abstinence Efficacy Scale.
[e]PP: per-protocol
[f]ITT: intention-to-treat.
[g]AAQ-II: Acceptance and Action Questionnaire—version II.
[h]RTCQ: Readiness to Change Questionnaires.
[i]EDS: Everyday Discrimination Scale.
[j]DD: drinks per day in the past 30 days.
[k]DDD: drinks per drinking day in the past 30 days.
[l]HDD: heavy drinking days in the past 30 days.
[m]CAD: cumulative abstinence durations in the past 30 days.

XSL•FO
RenderX

Table . Generalized Estimating Equations models with multiple imputation for the comparison of outcomes between the control and intervention groups of ethnic minority young adults with alcohol use disorder in Hong Kong across time points (pre and post; n=40): a pilot randomized controlled trial.

Outcomes and time points	ACT-BA[a]	TAU[b]	Group			Time			Group × time interaction		
	Mean (SE)	Mean (SE)	*B* (95% CI)	*P* value	Wald χ^2 (*df*)	*B* (95% CI)	*P* value	Wald χ^2 (*df*)	*B* (95% CI)	*P* value	Wald χ^2 (*df*)
AAES[c]											
GEE[d]											
Pre	44.45 (2.35)	49.55 (3.89)	1	≥.99	N/A[e]	1	N/A	N/A	1	≥.99	N/A
Post	58.73 (3.03)	51.11 (2.88)	−5.10 (−14.01 to 3.81)	.26	1.259 (1)	1.56 (−6.30 to 9.43)	.70	0.152 (1)	12.71 (2.79 to 22.64)	. .01	6.306 (1)
MI-GEE[f]											
Post	55.97 (3.83)	49.12 (3.26)	−0.43 (−9.69 to 8.82)	.93	1.259 (4)	−5.1 (−14.01 to 3.81)	.26	0.010 (4)	11.95 (0.10 to 23.81)	.05	3.960 (4)
RTCQ[g]										.01	
GEE											
Pre	37.55 (1.80)	34.20 (2.38)	1	≥.99		1	≥.99		1	≥.99	N/A
Post	43.49 (1.63)	37.65 (3.20)	3.35 (−2.49 to 9.19)	.26	1.264 (1)	3.45 (−5.47 to 12.36)	.45	0.574 (1)	2.50 (−7.79 to 12.79)	.23	0.227 (1)
MI-GEE											
Post	42.00 (1.68)	35.94 (2.99)	3.35 (−2.49 to 9.19)	.26	1.264 (4)	1.74 (−6.30 to 9.78)	.67	0.182 (4)	2.71 (−6.80 to 12.23)	.57	0.317 (4)
AAQ-II[h]											
GEE											
Pre	24.45 (1.96)	26.10 (2.35)	N/A	≥.99		1	≥.99		1	≥.99	N/A
Post	15.03 (1.25)	21.20 (1.84)	−1.65 (−7.65 to 4.34)	.59	0.291 (1)	−3.48 (−7.44 to −0.48)	.08	2.966 (1)	−6.93 (−11.95 to −1.92)	.001	7.333 (1)
MI-GEE											
Post	15.38 (1.38)	24.42 (1.62)	−1.65 (−7.65 to 4.35)	.59	0.291 (4)	−2.66 (−7.91 to 2.60)	.59	1.026 (4)	−6.41 (−12.77 to −0.06)	.048	3.960 (4)
EDS[i]											
GEE											
Pre	19.5 (1.69)	22.35 (2.00)	1	≥.99		1	≥.99	N/A	N/A	N/A	N/A
Post	11.6 (1.86)	19.72 (2.53)	−2.85 (−7.98 to 2.28)	.28	1.188 (1)	−2.62 (−7.04 to 1.79)	.22	1.360 (1)	−5.26 (−11.84 to 1.32)	.12	2.458 (1)

XSL•FO
RenderX

Outcomes and time points	ACT-BA[a]	TAU[b]	Group			Time			Group × time interaction		
	Mean (SE)	Mean (SE)	*B* (95% CI)	*P* value	Wald χ^2 (*df*)	*B* (95% CI)	*P* value	Wald χ^2 (*df*)	*B* (95% CI)	*P* value	Wald χ^2 (*df*)
MI-GEE											
Post	12.84 (1.97)	19.73 (2.40)	−2.85 (−7.98 to 2.28)	.28	1.188 (4)	−2.62 (−7.28 to 2.04)	.27	1.229 (4)	−4.04 (−10.43 to 2.35)	.22	1.533 (4)
DD[j]											
GEE											
Pre	9.35 (1.51)	10.10 (1.66)	1	N/A	N/A	1	≥.99	N/A	N/A	N/A	N/A
Post	5.30 (0.74)	10.07 (1.64)	−0.75 (−5.15 to 3.65)	.74	0.112 (1)	−0.03 (−3.06 to 2.99)	.98	0.000 (1)	−4.02 (−7.50 to −0.54)	.02	5.122 (1)
MI-GEE											
Post	5.98 (1.45)	10.84 (2.08)	−0.75 (−5.15 to 3.65)	.74	0.112 (4)	0.75 (−3.21 to 4.70)	.70	0.153 (4)	−4.12 (−8.10 to −0.13)	.04	4.224 (4)
DDD[k]											
GEE											
Pre	5.16 (1.64)	5.03 (0.15)	1	≥.99	N/A	1	≥.99	N/A	1	≥.99	1
Post	2.14 (0.35)	3.33 (0.49)	0.13 (−0.31 to 0.56)	.56	0.339 (1)	−1.71 (−2.75 to −0.66)	.001	10.31 (1)	−1.31 (−2.60 to −0.02)	.05	3.956 (1)
MI-GEE											
Post	2.45 (0.37)	3.88 (0.59)	0.13 (−0.31 to 0.56)	.56	0.339 (4)	−1.15 (−2.43 to 0.12)	.08	3.336 (4)	−1.56 (−3.06 to −0.07)	.04	4.298 (4)
HDD[l]											
GEE											
Pre	6.55 (1.17)	6.00 (0.57)	1	≥.99	N/A	1	≥.99	N/A	1	≥.99	1
Post	2.77 (1.23)	5.69 (1.63)	0.55 (−2.01 to 3.11)	.67	0.177 (1)	−0.32 (−3.70 to 3.06)	.85	0.034 (1)	−3.58 (−7.19 to 0.02)	.05	3.794 (1)
MI-GEE											
Post	3.12 (1.32)	5.57 (1.65)	0.55 (−2.01 to 3.11)	.67	0.177 (4)	−0.43 (−3.90 to 3.05)	.80	0.064 (4)	−3.01 (−6.37 to −0.36)	.08	3.135 (4)
CAD[m]											
GEE											
Pre	20.65 (1.51)	19.9 (1.66)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

XSL•FO
RenderX

Outcomes and time points	ACT-BA[a]	TAU[b]	Group			Time			Group × time interaction		
	Mean (SE)	Mean (SE)	*B* (95% CI)	*P* value	Wald χ^2 (*df*)	*B* (95% CI)	*P* value	Wald χ^2 (*df*)	*B* (95% CI)	*P* value	Wald χ^2 (*df*)
Post	22.78 (1.68)	19.65 (1.71)	0.75 (−3.65 to 5.15)	.74	0.112 (1)	−0.251 (−3.14 to 2.63)	.86	0.030 (1)	2.36 (−2.45 to 7.17)	.34	0.925 (1)
MI-GEE											
Post	22.52 (1.96)	19.43 (2.00)	0.75 (−3.65 to 5.15)	.74	0.112 (4)	−0.47 (−4.16 to 3.23)	.80	0.067 (4)	2.34 (−3.68 to 8.36)	.43	0.634 (4)

[a]ACT-BA: Acceptance and Commitment Therapy with Behavioral activation.
[b]TAU: Treatment-as-Usual.
[c]AAES: Alcohol Abstinence Efficacy Scale.
[d]GEE: Generalized Estimating Equations.
[e]N/A: not applicable.
[f]MI-GEE: Multiple Imputation with Generalized Estimating Equation.
[g]RTCQ: Readiness to Change Questionnaires.
[h]AAQ-II: Acceptance and Action Questionnaire—version II.
[i]EDS: Everyday Discrimination Scale.
[j]DD: drinks per day in the past 30 days.
[k]DDD: drinks per drinking day in the past 30 days.
[l]HDD: heavy drinking days in the past 30 days.
[m]CAD: cumulative abstinence durations in the past 30 days.

Effectiveness of ACT-BA on AASE

A significant improvement in AASE was observed in the intervention group compared with those in the control group at postintervention. Participants who received the ACT-BA intervention showed a statistically significant increase in the mean AAES score from baseline to postintervention in both ITT and PP analyses (ITT: Group × time, *B*=11.95, 95% CI 0.10-23.81, Wald χ^2_4=3.960,*P*=.048; PP: Group × time, *B*=12.71, 95% CI 2.79-22.64, Wald χ^2_1=6.306, *P*=.01; Table 5). The effect size of the intervention on AASE was also large in both ITT (Cohen *d*=0.81) and PP (Cohen *d*=0.83) analyses (Table 4).

Effectiveness of ACT-BA on Psychological Flexibility

Participants in the intervention group revealed a statistically significant reduction in the mean score of Acceptance and Action Questionnaire—version II at postintervention in both analyses (ITT: Group × time, *B*=−6.41, 95% CI −12.77 to −0.06, Wald χ^2_4=3.960,*P*=.04; PP: Group × time, *B*=−6.93, 95% CI −11.95 to −1.92, Wald χ^2_1=7.333,*P*=.001). The effect size on psychological flexibility was moderate (ITT: Cohen *d*=−0.65; PP: Cohen *d*=−0.7; Table 4).

Effectiveness of ACT-BA on Readiness to Change

A mild improvement was observed in the mean Readiness to Change Questionnaire score in the intervention group in both analyses (ITT: Group × time, *B*=2.71, 95% CI −6.80 to 12.23, Wald χ^2_4=0.317, *P*=.57; PP: Group × time, *B*=2.50, 95% CI −7.79 to 12.79, Wald χ^2_1=0.227, *P*=.23; Table 5), with a small effect size in both ITT and PP analyses (Cohen *d*=0.28; Table 4).

Effectiveness of ACT-BA on Everyday Discrimination

As presented in Table 5, a greater improvement in the mean score of everyday discrimination was observed in the ACT-BA group compared with the TAU group at postintervention (ITT: Group × time, *B*=−4.04, 95% CI −10.43 to 2.35, Wald χ^2_4=1.533, *P*=.22; PP: Group × time, *B*=−5.26, 95% CI −11.84 to 1.32, Wald χ^2_1=2.458, *P*=.12) with a small effect size (ITT: Cohen *d*=−0.48; PP: Cohen *d*=−0.55).

Participants' Feelings, Experiences, and Confidence

Participants showed high satisfaction with the ACT-BA intervention for its guidance to manage drinking, living, thoughts, and emotions. For instance, in the semistructured interview, a participant said, "I find it very helpful for everyday living," while others expressed its benefits as "I am happy as it helps people overcome drinking," "a guide on how to live," and a tool to "learn many things about drinking." Another participant stated that the mindfulness exercise "go to nature, observe, and relax" reduced her stress. Participants also reported self-efficacy in engaging with the program due to its clear instructions, simple intervention content, perceived health benefits, and easy website navigation. For instance, a participant said, "The website is easy without assistance, I just followed the directions, 1 do not need assistance because I do understand the context. So, I did it on my own," and "I'm confident in doing it. I just do it by myself."

XSL•FO
RenderX

Intervention Content, Delivery, and Duration

Most participants (n=11) expressed that the content was simple, with clear instructions. For instance, a participant described the content as "it's well explained, easy to understand" and instructions as "it is easy because I tried to follow the instructions." However, 2 of them noted challenges with unfamiliar terminology, particularly "need to search what words mean," and task difficulty as "some parts are easy, some are hard." All participants preferred the online delivery mode due to its flexibility and accessibility, and one said, "online is better because you can do it anywhere, anytime." Despite the majority of participants enjoying the interventions, a few faced challenges related to the content, duration, and access to the intervention. Regarding duration, while most found the duration suitable, 2 participants suggested shortening it and said, "a little bit longer, but the interface is beautiful. Every content is important." Additionally, one participant expressed difficulty in understanding some of the materials. In the semistructured interview, a participant said, "It's quite difficult for me to understand some of the content." Furthermore, 2 participants reported difficulties in accessing the intervention due to poor internet connection and stated that "sometimes there is an error when submitting, and another time there are errors in connection and login."

Perceived Usefulness of Interventions

As presented in Multimedia Appendix 1, the participants identified 5 themes regarding the benefits of the intervention. Participants reported increased awareness of their health and its consequences. For example, a participant said, "I got sick because of drinking alcohol… we are not getting younger." Another participant also noted, "I read all the consequences...it helps me realize now what will happen if I keep on drinking...how would it affect me if I didn't control it." By realizing the consequences, participants applied the BA strategies such as replacing drinking with alcohol-free activities. As reported by one participant, "I concentrate on other activities and don't drink as much as I can. I can do another thing instead of drinking because I know that drinking is not giving me anything more." Some participants also avoided triggers by "limiting time in the triggering environment." Besides, some participants learned to accept urges and negative emotions. One participant said, "I go to nature and just think and observe, and the switch can ease my mind, I can relax and reduce some of the stress." Some also learned cognitive defusion and understood that "drinking too much can bring themselves out of their mind," and such thinking helped them further detach from their urges. Overall, participants noted that this integrated intervention was a very good program and could be applied to their and others' lives in the future. For instance, one of the participants said, "Oh, I think my feedback is that your program is very good. It can help other people...Honestly, I was a drinker before. Now, I realize that our body is not healthy because of alcohol. Even though I drink once a week, but this is not good. I suffered for almost 2 years, I got sick because of drinking alcohol." Despite its usefulness, some participants reported difficulties in managing the withdrawal symptoms, for example, agitation and increased heart and respiratory rates, although they were motivated to quit alcohol use after the intervention.

Discussion

Principal Findings

This study was conducted among Hong Kong ethnic minority young adults with AUD, a population that often encounters barriers to health care access and has low treatment engagement and intervention adherence due to low-risk perception or poor health literacy [29,96]. To address this problem, we designed and evaluated a novel internet-based self-help program integrating ACT and BA among Hong Kong ethnic minority young adults with AUD.

This study provides preliminary evidence that our proposed intervention statistically significantly improved psychological flexibility, with a moderate effect size among ethnic minority young adults with AUD. This finding is consistent with previous studies on ACT for AUD [88,97]. However, our effect size on psychological flexibility is larger than that in previous studies, ranging from small to moderate [98]. The larger effect size might be due to the integration of ACT and BA [64,66]. Particularly, ACT emphasizes enhancing psychological flexibility in which participants learned different skills to allow difficult thoughts, feelings, and sensations, thus reducing experiential avoidance, which might trigger drinking as a coping behavior [50]. BA, as an additional component, further boosted the participants' psychological flexibility by guiding them to identify a list of alternative value-congruent activities to replace drinking [99].

Our previous cross-sectional study and qualitative interviews identified that experiences of discrimination were a key factor leading to the development and maintenance of AUD among ethnic minority young adults [29]. Our quantitative findings showed a moderate effect on discrimination, revealing its potential efficacy in reducing the mean score of discrimination experiences, although a statistically significant mean difference was not reached between groups at postintervention. The insignificant difference between groups could be that our intervention did not aim to reduce the discrimination encountered by the participants [100]. Instead, the core focus of the intervention was to enhance the participants' psychological flexibility to cope with the discrimination [50]. Another possible explanation for the insignificant difference is that the sample size of this study was small, resulting in a type 2 error in the analysis [101].

The ACT-BA intervention successfully enhanced participants' AASE in the intervention group compared with the TAU group, with a large effect size at postintervention. AASE is defined as personal confidence in one's own ability to abstain from alcohol [102]. Bandura [103] (1977) pointed out that self-efficacy is an important construct in determining whether a person will make attempts and take actions for behavioral changes. Although our intervention did not contain any strategy that could directly improve self-efficacy in alcohol abstinence among our participants, it can be indirectly and gradually built up in our intervention through participants' successful experience in alcohol reduction when they tried to cut down on drinking [104]. Based on this preliminary finding, we expect that the intervention effect on AASE would be stronger at a later time

XSL•FO
RenderX

because of more successful experience in alcohol reduction [105]. This postulation shall be further confirmed by a full-scale RCT with longer follow-up.

In this study, the ACT-BA was demonstrated to have statistically significant effects on some drinking outcomes. Particularly, compared with the TAU group, participants in the intervention group reported an average reduction of 4.12 drinking days per month and 1.15 drinks per drinking day at postintervention, with −0.57 and −1.89 effect sizes, respectively. These reductions are perhaps a reflection of the synergetic efficacy of the integrated intervention on the participants' psychological flexibility, which emphasized noticing negative and unfavorable feelings without judgment, while adapting alternative behaviors to replace drinking in view of pursuing life goals [106]. Notwithstanding a statistically significant mean difference on the number of drinking days and number of drinks per drinking day, we did not find any statistically significant difference on the number of heavy drinking days and cumulative alcohol abstinence, but still the intervention group showed improvement on those outcomes. The insignificant mean difference for these 2 outcomes was likely because complete abstinence from alcohol could trigger withdrawal symptoms, which were not heavily emphasized in our intervention, and these withdrawal symptoms were hard to overcome by only improving their psychological flexibility and self-efficacy in alcohol abstinence [107,108]. This explanation is also supported by our qualitative findings that some participants in the experimental group reported difficulties in managing withdrawal symptoms, for example, agitation and increased heart and respiratory rates, although they were motivated to quit alcohol after the intervention. In response to this issue, this intervention can be strengthened, on one hand, by increasing the content about the management of withdrawal symptoms and, on the other hand, by referring the participants who may require medications for withdrawal symptoms to health care professionals alongside the intervention [109]. Another possible explanation for the insignificant findings was that reduction in heavy drinking and alcohol abstinence are more distal outcomes, which could only be captured and detected by a longer follow-up, which is 6 months, a widely accepted international standard for evaluation of behavioral changes [110]. Thus, a follow-up duration of 6 months or more is recommended for future full-scale RCTs.

Although no significant difference was observed between groups regarding readiness to change, mainly due to the limited power of the study, the intervention group showed improvement in such outcome at postintervention in ITT analysis (B=2.71). Again, this was likely because our proposed intervention focused on enhancing psychological flexibility and value-driven action skills rather than the shift in stages of readiness to change [77]. Additionally, the nonsignificant group-by-time interaction could be partly attributed to a ceiling effect, as most participants reported a high score of readiness to change at baseline, resulting from the convenience sampling technique [111].

The study revealed that the ACT-BA program was feasible for ethnic minority young adults with AUD, evidenced by a high intervention completion of 80% and a low attrition rate of 15%. This completion rate is higher than those reported in existing studies on various interventions, including ACT for AUD [57,98]. The variation could be attributable to 2 innovations of this study. First, the incorporation of BA strategies into ACT interventions may enhance engagement through value-aligned activity substitution. Second, when compared with the traditional face-to-face method to deliver ACT, our intervention was delivered using an online platform with no fixed schedule, allowing the participants to access the intervention at their convenience throughout the week, minimizing work disruptions [112]. While there was variation in time between sessions (2 median intervals of 7 days; IQR days), ethnic minority young adults confirmed that a 6-week duration was appropriate, consistent with existing literature [56].

Our postintervention semistructured interview revealed an overall high satisfaction among the participants in the experimental group, with all endorsing the intervention as effective in guiding them to reduce alcohol consumption, improve emotional regulation, and enhance life management. The participants particularly highlighted that the intervention has clear instructions, simplified content, and easy website navigation, enhancing their interest in engaging with the intervention. In their interview, most participants described the intervention sessions as "simple, well explained, and easy to understand." All participants preferred the online delivery format due to its flexibility and accessibility [112]. Importantly, the participants reported various degrees in behavior change in their alcohol use and expressed intent to practice and sustain the learned intervention skills in the long term. These positive responses from the participants provide compelling evidence to support the feasibility of the ACT-BA, apart from the feasibility indicators.

Limitations

This RCT has several limitations. First, the small sample size may reduce the statistical power, potentially affecting the true effect size of the interventions. Second, the self-help intervention limited direct measurement of the participants' engagement with exercises, metaphors, and practices of each module. Nevertheless, engagement was assessed indirectly through online tracking of the module completion time and daily exercise logs. Third, the preliminary effectiveness was assessed immediately postintervention. Although psychosocial interventions usually bring behavioral changes over time following the intervention, the absence of follow-up assessment may limit the evaluation of the long-term effect of the ACT-BA. Fourth, this study relied solely on participants' self-reported data rather than objective assessments, which might introduce bias and potentially lead to an over- or underestimation of the effect size of the outcome variables. Fifth, although the ACT-BA showed significant effects across multiple outcomes, how the effects varied across different ethnic minority groups was not examined. Sixth, although almost all ethnic minority participants were able to speak English, excluding individuals who do not speak English may have introduced selection bias, potentially limiting the generalizability of findings to all ethnic minority groups.

Future Implications

This study addresses the literature gaps by providing evidence on the preliminary effectiveness and feasibility of the ACT-BA on different drinking-related outcomes among ethnic minorities

XSL•FO
RenderX

with AUD. Therefore, a key implication for future study was to conduct a fully powered RCT to further examine the effectiveness of such intervention, with longer follow-up to capture the long-term effect on different drinking outcomes and incorporating objective measures to assess the drinking outcomes to minimize the social desirability bias. Given that the intervention can be self-administered in an online platform with only minimal involvement from interventionalists in a flexible schedule for working groups and at a low cost, it is expected that this intervention, if proven to be effective in a large-scale similar RCT, can be further delivered to a large group of ethnic minority young adults in the community.

Conclusions

This study presents an innovative integrated ACT and BA intervention, which has been delivered through an internet-based self-help format for ethnic minority young adults with AUD. Unlike the existing evidence that mainly focuses on a single intervention, either ACT or BA, this study integrated those 2 interventions to enhance improvement in abstinence. The findings contribute to the field by providing preliminary evidence for both the feasibility and the potential efficacy of this novel, self-help, flexible, and low-intensity integrated intervention. Feasibility was supported by the high completion rate and low attrition rate, while preliminary efficacy was reflected in reductions in drinks per drinking day and drinking days, alongside improvements in AASE and psychological flexibility among ethnic minority young adults with AUD. The main implication of this study is that this intervention would serve as a key treatment option for ethnic minorities with AUD who encounter challenges in face-to-face therapy. Given these promising findings, the effectiveness of the ACT-BA should be further confirmed in a fully powered RCT before its implementation for this population group in the community.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: KYH (lead), GMB (equal), YWM (equal)
Data curation: GMB (equal), YWM (equal), KYH (equal)
Formal analysis: GMB (equal), YWM (equal), KYH (equal)
Investigation: GMB (equal), YWM (equal), KYH (equal)
Methodology: GMB (lead), YWM (equal), KYH (equal), KKWL (supporting), QL (supporting), TM (supporting)
Project administration: KYH (lead), GMB (equal), YWM (equal)
Resources: KYH (lead), YWM (equal)
Software: GMB (equal), KYH (equal)
Validation: GMB (lead), YWM (equal), KYH (equal), KKWL (supporting)
Visualization: GMB (lead), YWM (equal), KYH (equal), KKWL (supporting), QL (supporting)
Supervision: YWM (equal), KYH (equal)
Writing – original draft: GMB (lead), YWM (supporting), KYH (supporting)
Writing – review & editing: GMB (lead), YWM (equal), KYH (equal), KKWL (supporting), QL (supporting), TM (supporting)
All authors have read, reviewed, and approved the final version of the manuscript, and agreed to the order of authorship and the assignment of CRediT roles.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Themes, codes, sample quotes, and descriptions for the qualitative findings.
[DOCX File, 25 KB - jmir_v28i1e83896_app1.docx]

Checklist 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File, 3428 KB - jmir_v28i1e83896_app2.pdf]

Checklist 2
CONSORT 2025 checklist for reporting a randomized trial.
[PDF File, 252 KB - jmir_v28i1e83896_app3.pdf]

XSL•FO
RenderX

References

1. Understanding alcohol use disorder. National Institute on Alcohol Abuse and Alcoholism. 2022. URL: https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/understanding-alcohol-use-disorder [accessed 2026-04-29]
2. Substance use in adolescents. MSD MANUAL. 2022. URL: https://www.msdmanuals.com/home/children-s-health-issues/problems-in adolescents/substance-use-in-adolescents [accessed 2026-04-29]
3. Naar S. The transition from adolescence to adulthood. Psychology Today. 2023. URL: https://www.psychologytoday.com/us/blog/metamorphosis/202107/the-transition-adolescence-adulthood [accessed 2026-04-29]
4. What increases the risk for alcohol use disorder? National Institute on Alcohol Abuse and Alcoholism. 2025. URL: https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/understanding-alcohol-use-disorder [accessed 2026-01-01]
5. Dawson DA, Goldstein RB, Chou SP, Ruan WJ, Grant BF. Age at first drink and the first incidence of adult-onset DSM-IV alcohol use disorders. Alcohol Clin Exp Res 2008 Dec;32(12):2149-2160. [doi: 10.1111/j.1530-0277.2008.00806.x] [Medline: 18828796]
6. Squeglia LM. Alcohol and the developing adolescent brain. World Psychiatry 2020 Oct;19(3):393-394. [doi: 10.1002/wps.20786] [Medline: 32931114]
7. The global action for measurement of adolescent health (GAMA). World Health Organization. 2022. URL: https://www.who.int/groups/the-global-action-for-measurement-of-adolescent-health [accessed 2026-04-29]
8. Sullivan M, Risler E. Understanding college alcohol abuse and academic performance: selecting appropriate intervention strategies. J of College Counseling 2002 Sep;5(2):114-123. [doi: 10.1002/j.2161-1882.2002.tb00213.x]
9. Castaño-Perez GA, Calderon-Vallejo GA. Problems associated with alcohol consumption by university students. Rev Lat Am Enfermagem 2014 Oct;22(5):739-746. [doi: 10.1590/0104-1169.3579.2475] [Medline: 25493668]
10. Oluwafemi A. Alcohol consumption as anticipator of academic performance among undergraduate students (a case study of Ondo State University of Science and Technology Okitipupa, Ondo State Nigeria). IFE PsychologIA Int J 2020;28(1):159-165 [FREE Full text]
11. Sindelar HA, Barnett NP, Spirito A. Adolescent alcohol use and injury. A summary and critical review of the literature. Minerva Pediatr 2004 Jun;56(3):291-309. [Medline: 15252378]
12. Hingson RW, Zha W. Age of drinking onset, alcohol use disorders, frequent heavy drinking, and unintentionally injuring oneself and others after drinking. Pediatrics 2009 Jun;123(6):1477-1484. [doi: 10.1542/peds.2008-2176] [Medline: 19482757]
13. Callaghan RC, Gatley JM, Veldhuizen S, Lev-Ran S, Mann R, Asbridge M. Alcohol- or drug-use disorders and motor vehicle accident mortality: a retrospective cohort study. Accident Analysis & Prevention 2013 Apr;53:149-155. [doi: 10.1016/j.aap.2013.01.008]
14. Holcomb RL. Alcohol in relation to traffic accidents. JAMA 1938 Sep 17;111(12):1076. [doi: 10.1001/jama.1938.02790380018006]
15. Alcohol, fact sheet. World Health Organization. 2022. URL: https://www.who.int/news-room/fact-sheets/detail/alcohol [accessed 2026-04-29]
16. Cheung YW. Ethnic identification and alcohol use among Canadian-born and foreign-born high school students in Toronto. Int J Addict 1993 Sep;28(11):1095-1109. [doi: 10.3109/10826089309056245] [Medline: 8407030]
17. Edwards RW, Thurman PJ, Beauvais F. Patterns of alcohol use among ethnic minority adolescent women. Recent Dev Alcohol 1995;12:369-386. [doi: 10.1007/0-306-47138-8_23] [Medline: 7624553]
18. Statistics CfBH, Quality. Racial/ethnic differences in substance use, substance use disorders, and substance use treatment utilization among people aged 12 or older (2015-2019). : Substance Abuse and Mental Health Services Administration; 2021.
19. The demographics: ethnic groups. Relation Units. 2021. URL: https://www.had.gov.hk/rru/english/info/demographics.htm [accessed 2026-04-29]
20. Discrimination and mental health—a guide for young people. Cool Minds. 2022. URL: https://kely.org/our-work/mental-health-wellbeing/coolminds# [accessed 2022-07-20]
21. Dominic L, Michelle L. Being young and a minority: a study of the challenges encountered by young ethnic minorities in Hong Kong. J Youth Stud 2011;14(2) [FREE Full text]
22. Rohde P, Lewinsohn PM, Kahler CW, Seeley JR, Brown RA. Natural course of alcohol use disorders from adolescence to young adulthood. J Am Acad Child Adolesc Psychiatry 2001 Jan;40(1):83-90. [doi: 10.1097/00004583-200101000-00020] [Medline: 11195569]
23. Chou T, Asnaani A, Hofmann SG. Perception of racial discrimination and psychopathology across three U.S. ethnic minority groups. Cultur Divers Ethnic Minor Psychol 2012 Jan;18(1):74-81. [doi: 10.1037/a0025432] [Medline: 21967527]
24. Demombynes G. Why is ethnic minority poverty persistent in Vietnam. : World Bank Blogs; 2013 URL: https://blogs worldbank org/eastasiapacific/why-ethnic-minority-poverty-persistent-vietnam [accessed 2026-05-08]
25. Liu MA, Prestigiacomo CJ, Karim MFA, Ashburn-Nardo L, Cyders MA. Psychological outcomes and culturally relevant moderators associated with events of discrimination among Asian American adults. Cultur Divers Ethnic Minor Psychol 2024 Apr;30(2):363-373. [doi: 10.1037/cdp0000568] [Medline: 36355695]
26. Vaeth PAC, Wang-Schweig M, Caetano R. Drinking, alcohol use disorder, and treatment access and utilization among U.S. racial/ethnic groups. Alcohol Clin Exp Res 2017 Jan;41(1):6-19. [doi: 10.1111/acer.13285] [Medline: 28019654]

XSL•FO
RenderX

27. Vandan N, Wong JYH, Lee JJJ, Yip PSF, Fong DYT. Challenges of healthcare professionals in providing care to South Asian ethnic minority patients in Hong Kong: a qualitative study. Health Soc Care Community 2020 Mar;28(2):591-601. [doi: 10.1111/hsc.12892] [Medline: 31750578]
28. Research on social inclusion: employment situations of ethnic minority in Hong Kong. TREATS. 2021. URL: https://treats.org.hk/en/research-on-social-inclusion/ [accessed 2026-04-29]
29. Belay GM, Wah MY, Lam KKW, et al. Reasons and impacts of alcohol use disorder among ethnic minority young adults: a descriptive phenomenological study. Front Public Health 2025;13:1560092. [doi: 10.3389/fpubh.2025.1560092] [Medline: 40612563]
30. Bernstein J, Heeren T, Edward E, et al. A brief motivational interview in a pediatric emergency department, plus 10-day telephone follow-up, increases attempts to quit drinking among youth and young adults who screen positive for problematic drinking. Acad Emerg Med 2010 Aug;17(8):890-902. [doi: 10.1111/j.1553-2712.2010.00818.x] [Medline: 20670329]
31. Borsari B, Short EE, Mastroleo NR, et al. Phone-delivered brief motivational interventions for mandated college students delivered during the summer months. J Subst Abuse Treat 2014;46(5):592-596. [doi: 10.1016/j.jsat.2014.01.001] [Medline: 24512944]
32. Brown RA, Abrantes AM, Minami H, et al. Motivational interviewing to reduce substance use in adolescents with psychiatric comorbidity. J Subst Abuse Treat 2015 Dec;59:20-29. [doi: 10.1016/j.jsat.2015.06.016] [Medline: 26362000]
33. Newton AS, Dow N, Dong K, et al. A randomised controlled pilot trial evaluating feasibility and acceptability of a computer-based tool to identify and reduce harmful and hazardous drinking among adolescents with alcohol-related presentations in Canadian pediatric emergency departments. BMJ Open 2017 Aug 11;7(8):e015423. [doi: 10.1136/bmjopen-2016-015423] [Medline: 28801399]
34. Segatto ML, Andreoni S, de Souza e Silva R, Diehl A, Pinsky I. Brief motivational interview and educational brochure in emergency room settings for adolescents and young adults with alcohol-related problems: a randomized single-blind clinical trial. Braz J Psychiatry 2011 Sep;33(3):225-233. [doi: 10.1590/s1516-44462011000300004] [Medline: 21971774]
35. Wagner EF, Hospital MM, Graziano JN, Morris SL, Gil AG. A randomized controlled trial of guided self-change with minority adolescents. J Consult Clin Psychol 2014 Dec;82(6):1128-1139. [doi: 10.1037/a0036939] [Medline: 24841864]
36. Slesnick N, Prestopnik JL. Comparison of family therapy outcome with alcohol-abusing, runaway adolescents. J Marital Fam Ther 2009 Jul;35(3):255-277. [doi: 10.1111/j.1752-0606.2009.00121.x] [Medline: 19522781]
37. Arnaud N, Diestelkamp S, Wartberg L, Sack PM, Daubmann A, Thomasius R. Short- to midterm effectiveness of a brief motivational intervention to reduce alcohol use and related problems for alcohol intoxicated children and adolescents in pediatric emergency departments: a randomized controlled trial. Acad Emerg Med 2017 Feb;24(2):186-200. [doi: 10.1111/acem.13126] [Medline: 27801991]
38. Martín-Pérez C, Navas JF, Perales JC, et al. Brief group-delivered motivational interviewing is equally effective as brief group-delivered cognitive-behavioral therapy at reducing alcohol use in risky college drinkers. PLoS One 2019;14(12):e0226271. [doi: 10.1371/journal.pone.0226271] [Medline: 31821350]
39. Spirito A, Monti PM, Barnett NP, et al. A randomized clinical trial of a brief motivational intervention for alcohol-positive adolescents treated in an emergency department. J Pediatr 2004 Sep;145(3):396-402. [doi: 10.1016/j.jpeds.2004.04.057] [Medline: 15343198]
40. Latimer WW, Winters KC, D'Zurilla T, Nichols M. Integrated family and cognitive-behavioral therapy for adolescent substance abusers: a stage I efficacy study. Drug Alcohol Depend 2003 Sep 10;71(3):303-317. [doi: 10.1016/s0376-8716(03)00171-6] [Medline: 12957348]
41. Kaminer Y, Burleson JA, Burke RH. Efficacy of outpatient aftercare for adolescents with alcohol use disorders: a randomized controlled study. J Am Acad Child Adolesc Psychiatry 2008 Dec;47(12):1405-1412. [doi: 10.1097/CHI.0b013e318189147c] [Medline: 18978635]
42. Murray LK, Kane JC, Glass N, et al. Effectiveness of the Common Elements Treatment Approach (CETA) in reducing intimate partner violence and hazardous alcohol use in Zambia (VATU): a randomized controlled trial. PLoS Med 2020 Apr;17(4):e1003056. [doi: 10.1371/journal.pmed.1003056] [Medline: 32302308]
43. Belay GM, Mak YW, Wong FKY, et al. Psychosocial treatment options for adolescents and young adults with alcohol use disorder: systematic review and meta-analysis. Front Public Health 2024;12:1371497. [doi: 10.3389/fpubh.2024.1371497] [Medline: 39114525]
44. Perkins AM, Meiser-Stedman R, Spaul SW, Bowers G, Perkins AG, Pass L. The effectiveness of third wave cognitive behavioural therapies for children and adolescents: a systematic review and meta-analysis. Br J Clin Psychol 2023 Mar;62(1):209-227. [doi: 10.1111/bjc.12404] [Medline: 36443910]
45. Kahl KG, Winter L, Schweiger U. The third wave of cognitive behavioural therapies: what is new and what is effective? Curr Opin Psychiatry 2012 Nov;25(6):522-528. [doi: 10.1097/YCO.0b013e328358e531] [Medline: 22992547]
46. Dela Cruz GA, Johnstone S, Kim HS, Castle DJ. Review of third-wave therapies for substance use disorders in people of color and collectivist cultures: current evidence and future directions. Psychol Addict Behav 2023 Aug;37(5):681-694. [doi: 10.1037/adb0000883] [Medline: 36107641]

XSL•FO
RenderX

47. Vujanovic AA, Meyer TD, Heads AM, Stotts AL, Villarreal YR, Schmitz JM. Cognitive-behavioral therapies for depression and substance use disorders: an overview of traditional, third-wave, and transdiagnostic approaches. Am J Drug Alcohol Abuse 2017 Jul;43(4):402-415. [doi: 10.1080/00952990.2016.1199697] [Medline: 27494547]
48. Hayes SC. Acceptance and commitment therapy, relational frame theory, and the third wave of behavioral and cognitive therapies. Behav Ther 2004 ;35(4):639-665. [doi: 10.1016/S0005-7894(04)80013-3]
49. Stotts AL, Northrup TF. The promise of third-wave behavioral therapies in the treatment of substance use disorders. Curr Opin Psychol 2015 Apr 1;2:75-81. [doi: 10.1016/j.copsyc.2014.12.028] [Medline: 26693170]
50. Hayes SC, Luoma JB, Bond FW, Masuda A, Lillis J. Acceptance and commitment therapy: model, processes and outcomes. Behav Res Ther 2006 Jan;44(1):1-25. [doi: 10.1016/j.brat.2005.06.006] [Medline: 16300724]
51. Hayes SC, Strosahl KD. A Practical Guide to Acceptance and Commitment Therapy: Springer Science & Business Media; 2004. URL: https://psycnet.apa.org/record/2005-02327-000 [accessed 2026-04-29]
52. Forman EM, Juarascio AS, Martin LM, Herbert JD. Acceptance and commitment therapy (ACT). The Encyclopedia of Clinical Psychology. 2014. [doi: 10.1002/9781118625392]
53. Sevinçok D, Memiş Ç, Çakaloz B. Acceptance and commitment therapy experience in an adolescent patient with social anxiety disorder, panic disorder, and major depression. Psychiatry Clin Psychopharmacol 2018;28:138-139 [FREE Full text]
54. Lee EB, An W, Levin ME, Twohig MP. An initial meta-analysis of Acceptance and Commitment Therapy for treating substance use disorders. Drug Alcohol Depend 2015 Oct 1;155:1-7. [doi: 10.1016/j.drugalcdep.2015.08.004] [Medline: 26298552]
55. Osaji J, Ojimba C, Ahmed S. The use of acceptance and commitment therapy in substance use disorders: a review of literature. J Clin Med Res 2020 Oct;12(10):629-633. [doi: 10.14740/jocmr4311] [Medline: 33029268]
56. Levin ME, Haeger JA, Pierce BG, Twohig MP. Web-based acceptance and commitment therapy for mental health problems in college students: a randomized controlled trial. Behav Modif 2017 Jan 1;41(1):141-162. [doi: 10.1177/0145445516659645] [Medline: 27440189]
57. Khandelwal N, Das K, Sharma R, Ghosh A. Testing the waters: a pilot trial of acceptance and commitment therapy (ACT) for alcohol use disorder. Indian J Psychiatry 2024 Feb;66(2):191-194. [doi: 10.4103/indianjpsychiatry.indianjpsychiatry_820_23] [Medline: 38523767]
58. Brown M, Hooper N, James P, Scott D, Bodger O, John A. A web-delivered acceptance and commitment therapy intervention with email reminders to enhance subjective well-being and encourage engagement with lifestyle behavior change in health care staff: randomized cluster feasibility stud. JMIR Form Res 2020 Aug 7;4(8):e18586. [doi: 10.2196/18586] [Medline: 32763887]
59. Clark KR. Learning theories: Behaviorism. Radiol Technol 2018 Nov;90(2):172-175. [Medline: 30420574]
60. Martell CR, Dimidjian S, Herman-Dunn R. Behavioral activation for depression. In: Barlow DH, editor. Clinical Handbook of Psychological Disorders: A Step-by-Step Treatment Manual, 6th edition: Guilford Press:339-380 URL: https://psycnet.apa.org/record/2021-54173-009 [accessed 2026-04-29]
61. Kanter JW, Manos RC, Bowe WM, Baruch DE, Busch AM, Rusch LC. What is behavioral activation? A review of the empirical literature. Clin Psychol Rev 2010 Aug;30(6):608-620. [doi: 10.1016/j.cpr.2010.04.001] [Medline: 20677369]
62. Santos MM, Puspitasari AJ, Nagy GA, Kanter JW. Behavioral activation. In: Wenzel A, editor. Handbook of Cognitive Behavioral Therapy: Overview and Approaches: American Psychological Association; 2021:235-273. [doi: 10.1037/0000218-009]
63. Fernández-Rodríguez C, Coto-Lesmes R, Martínez-Loredo V, González-Fernández S, Cuesta M. Is activation the active ingredient of transdiagnostic therapies? A randomized clinical trial of behavioral activation, acceptance and commitment therapy, and transdiagnostic cognitive-behavioral therapy for emotional disorders. Behav Modif 2023 Jan;47(1):3-45. [doi: 10.1177/01454455221083309] [Medline: 35426318]
64. Gaudiano BA, Nowlan K, Brown LA, Epstein-Lubow G, Miller IW. An open trial of a new acceptance-based behavioral treatment for major depression with psychotic features. Behav Modif 2013 May;37(3):324-355. [doi: 10.1177/0145445512465173] [Medline: 23223385]
65. Daughters SB, Magidson JF, Lejuez CW, Chen Y. LETS ACT: a behavioral activation treatment for substance use and depression. Adv Dual Diagn 2016 Aug 15;9(2/3). [doi: 10.1108/ADD-02-2016-0006]
66. Carlbring P, Hägglund M, Luthström A, et al. Internet-based behavioral activation and acceptance-based treatment for depression: a randomized controlled trial. J Affect Disord 2013 Jun;148(2-3):331-337. [doi: 10.1016/j.jad.2012.12.020] [Medline: 23357657]
67. Moock J. Support from the internet for individuals with mental disorders: advantages and disadvantages of e-mental health service delivery. Front Public Health 2014;2:65. [doi: 10.3389/fpubh.2014.00065] [Medline: 24967221]
68. Taylor CB, Luce KH. Computer- and internet-based psychotherapy interventions. Curr Dir Psychol Sci 2003 Feb;12(1):18-22. [doi: 10.1111/1467-8721.01214]
69. Andersson G, Titov N. Advantages and limitations of Internet-based interventions for common mental disorders. World Psychiatry 2014 Feb;13(1):4-11. [doi: 10.1002/wps.20083] [Medline: 24497236]

XSL•FO
RenderX

70. Thomas N, McLeod B, Jones N, Abbott JA. Developing Internet interventions to target the individual impact of stigma in health conditions. Internet Interv 2015 Sep;2(3):351-358. [doi: 10.1016/j.invent.2015.01.003]
71. Cunningham JA, Wild TC, Cordingley J, van Mierlo T, Humphreys K. A randomized controlled trial of an internet-based intervention for alcohol abusers. Addiction 2009 Dec;104(12):2023-2032. [doi: 10.1111/j.1360-0443.2009.02726.x] [Medline: 19922569]
72. Lappalainen P, Granlund A, Siltanen S, et al. ACT internet-based vs face-to-face? A randomized controlled trial of two ways to deliver Acceptance and Commitment Therapy for depressive symptoms: an 18-month follow-up. Behav Res Ther 2014 Oct;61:43-54. [doi: 10.1016/j.brat.2014.07.006] [Medline: 25127179]
73. Blankers M, Nabitz U, Smit F, Koeter MWJ, Schippers GM. Economic evaluation of internet-based interventions for harmful alcohol use alongside a pragmatic randomized controlled trial. J Med Internet Res 2012 Oct 29;14(5):e134. [doi: 10.2196/jmir.2052] [Medline: 23103771]
74. Pots WTM, Fledderus M, Meulenbeek PAM, ten Klooster PM, Schreurs KMG, Bohlmeijer ET. Acceptance and commitment therapy as a web-based intervention for depressive symptoms: randomised controlled trial. Br J Psychiatry 2016 Jan;208(1):69-77. [doi: 10.1192/bjp.bp.114.146068] [Medline: 26250745]
75. Zettle R. ACT for Depression: A Clinician's Guide to Using Acceptance and Commitment Therapy in Treating Depression: New Harbinger Publications; 2007. URL: https://books.google.com.hk/books/about/ACT_for_Depression.html?id=4VUp7wlwXH8C&redir_esc=y [accessed 2026-04-29]
76. Hayes SC, Strosahl KD, Wilson KG. Acceptance and Commitment Therapy: The Process and Practice of Mindful Change: Guilford Press; 1999. URL: https://www.guilford.com/excerpts/hayes_ch3.pdf?t=1 [accessed 2026-04-29]
77. Hayes SC, Pierson H. Acceptance and commitment therapy. In: Encyclopedia of Cognitive Behavior Therapy: Springer US; 1999:1-4. [doi: 10.1007/0-306-48581-8_1]
78. Vuchinich RE, Tucker JA. Contributions from behavioral theories of choice to an analysis of alcohol abuse. J Abnorm Psychol 1988 May;97(2):181-195. [doi: 10.1037//0021-843x.97.2.181] [Medline: 3133403]
79. Daughters SB, Magidson JF, Anand D, Seitz-Brown CJ, Chen Y, Baker S. The effect of a behavioral activation treatment for substance use on post-treatment abstinence: a randomized controlled trial. Addiction 2018 Mar;113(3):535-544. [doi: 10.1111/add.14049] [Medline: 28963853]
80. Population, total—Hong Kong SAR. World Bank. 2022. URL: https://data.worldbank.org/indicator/SP.POP.TOTL?locations=HK [accessed 2026-04-29]
81. Alcohol use disorder: a comparison between DSM–IV and DSM–5 on alcohol abuse or alcoholism. National Institute on Alcohol Abuse and Alcoholism. 2021. URL: https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/alcohol-use-disorder-comparison-between-dsm [accessed 2026-04-29]
82. Vespa J. The changing economics and demographics of young adulthood: 1975-2016. : US Department of Commerce, Economics and Statistics Administration, US Census Bureau; 2017 Apr 7 URL: https://hispanicad.com/wp-content/uploads/2017/04/p20-579.pdf [accessed 2026-04-29]
83. Liu T, Gietel-Basten S. The demography of drug abuse in Hong Kong. China J Soc Work 2019 Sep 2;12(3):254-272. [doi: 10.1080/17525098.2019.1661073]
84. Julious SA. Sample size of 12 per group rule of thumb for a pilot study. Pharm Stat 2005 Oct;4(4):287-291. [doi: 10.1002/pst.185]
85. Johns EK, Kokokyi S, Neufeld D, Krysanski V, Meek BP. Taking ACTion for anxiety and depression: a pilot study of a brief virtual acceptance and commitment therapy group in primary care. J Contextual Behav Sci 2024 Jul;33:100812. [doi: 10.1016/j.jcbs.2024.100812]
86. Petersen CL, Zettle RD. Treating inpatients with comorbid depression and alcohol use disorders: a comparison of acceptance and commitment therapy versus treatment as usual. Psychol Rec 2009 Oct;59(4):521-536. [doi: 10.1007/BF03395679]
87. Thekiso TB, Murphy P, Milnes J, Lambe K, Curtin A, Farren CK. Acceptance and commitment therapy in the treatment of alcohol use disorder and comorbid affective disorder: a pilot matched control trial. Behav Ther 2015 Nov;46(6):717-728. [doi: 10.1016/j.beth.2015.05.005] [Medline: 26520216]
88. Weststrate TR, Briggs CA, Miller A, Shuster AE, Gaynor ST. Brief acceptance and commitment therapy added to medication management during acute alcohol detoxification: A pilot randomized controlled effectiveness trial. J Contextual Behav Sci 2023 Jul;29:76-85. [doi: 10.1016/j.jcbs.2023.06.002]
89. Understanding alcohol drinking patterns. National Institute on Alcohol Abuse and Alcoholism. 2025. URL: https://www.niaaa.nih.gov/alcohols-effects-health/alcohol-drinking-patterns [accessed 2026-04-29]
90. Heather N, Rollnick S. Readiness to Change Questionnaire: user's manual. : National Drug and Alcohol Research Centre; 1993 URL: https://ndarc.med.unsw.edu.au/sites/default/files/ndarc/resources/TR.019.pdf [accessed 2026-04-29]
91. Bond FW, Hayes SC, Baer RA, et al. Preliminary psychometric properties of the Acceptance and Action Questionnaire-II: a revised measure of psychological inflexibility and experiential avoidance. Behav Ther 2011 Dec;42(4):676-688. [doi: 10.1016/j.beth.2011.03.007] [Medline: 22035996]
92. Belay GM, Wai KLK, Mak YW, et al. Factors associated with alcohol use disorder among young adults from ethnic minorities: a cross-sectional study. BMC Psychol 2026 Mar 20:s40359-s40026. [doi: 10.1186/s40359-026-04273-3] [Medline: 41857666]

XSL•FO
RenderX

93. Sternthal MJ, Slopen N, Williams DR. Racial disparities in health: how much does stress really matter? Du Bois Rev 2011;8(1):95-113. [doi: 10.1017/S1742058X11000087] [Medline: 29887911]
94. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. Lancet 2025 Apr 14;405(10489):1633-1640. [doi: 10.1016/S0140-6736(25)00672-5] [Medline: 40245901]
95. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
96. Vandan N, Wong JYH, Fong DYT. Accessing health care: experiences of South Asian ethnic minority women in Hong Kong. Nurs Health Sci 2019 Mar;21(1):93-101. [doi: 10.1111/nhs.12564] [Medline: 30156370]
97. Svanberg G, Munck I, Levander M. Acceptance and commitment therapy for clients institutionalized for severe substance-use disorder: a pilot study. Subst Abuse Rehabil 2017;8:45-51. [doi: 10.2147/SAR.S132255] [Medline: 28794670]
98. Meyer EC, Walser R, Hermann B, et al. Acceptance and commitment therapy for co-occurring posttraumatic stress disorder and alcohol use disorders in veterans: pilot treatment outcomes. J Trauma Stress 2018 Oct;31(5):781-789. [doi: 10.1002/jts.22322] [Medline: 30338561]
99. Ghodrati S, Vaziri Nekoo R. The effectiveness of behavioral activation (BA) on psychological well-being and psychological flexibility in female students. J Med Counc Iran 2018;36(4):233-240 [FREE Full text]
100. Harris R. ACT made simple: an easy-to-read primer on acceptance and commitment therapy. : New Harbinger Publications; 2019 May 1 URL: https://www.actmindfully.com.au/upimages/ACT_Made_Simple_Introduction_and_first_two_chapters.pdf [accessed 2026-04-29]
101. Akobeng AK. Understanding type I and type II errors, statistical power and sample size. Acta Paediatr 2016 Jun;105(6):605-609. [doi: 10.1111/apa.13384] [Medline: 26935977]
102. DiClemente CC, Carbonari JP, Montgomery RP, Hughes SO. The Alcohol Abstinence Self-Efficacy Scale. J Stud Alcohol 1994 Mar;55(2):141-148. [doi: 10.15288/jsa.1994.55.141] [Medline: 8189734]
103. Bandura A. Self-efficacy: toward a unifying theory of behavioral change. Psychol Rev 1977 Mar;84(2):191-215. [doi: 10.1037//0033-295x.84.2.191] [Medline: 847061]
104. Kadden RM, Litt MD. The role of self-efficacy in the treatment of substance use disorders. Addict Behav 2011 Dec;36(12):1120-1126. [doi: 10.1016/j.addbeh.2011.07.032] [Medline: 21849232]
105. Rollnick S, Heather N. The application of Bandura's self-efficacy theory to abstinence-oriented alcoholism treatment. Addict Behav 1982;7(3):243-250. [doi: 10.1016/0306-4603(82)90051-x] [Medline: 7180618]
106. Reynolds EK, Macpherson L, Tull MT, Baruch DE, Lejuez CW. Integration of the brief behavioral activation treatment for depression (BATD) into a college orientation program: depression and alcohol outcomes. J Couns Psychol 2011 Oct;58(4):555-564. [doi: 10.1037/a0024634] [Medline: 21787070]
107. Hall W, Zador D. The alcohol withdrawal syndrome. Lancet 1997 Jun 28;349(9069):1897-1900. [doi: 10.1016/S0140-6736(97)04572-8] [Medline: 9217770]
108. Becker HC. Alcohol dependence, withdrawal, and relapse. Alcohol Res Health 2008;31(4):348-361. [Medline: 23584009]
109. Kattimani S, Bharadwaj B. Clinical management of alcohol withdrawal: a systematic review. Ind Psychiatry J 2013 Jul;22(2):100-108. [doi: 10.4103/0972-6748.132914] [Medline: 25013309]
110. Wayne W. The transtheoretical model (stages of change). Encyclopedia. 2022. URL: https://encyclopedia.pub/entry/30695 [accessed 2026-04-29]
111. Andrade C. The ceiling effect, the floor effect, and the importance of active and placebo control arms in randomized controlled trials of an investigational drug. Indian J Psychol Med 2021 Jul;43(4):360-361. [doi: 10.1177/02537176211021280] [Medline: 34385732]
112. Rafieifar M, Hanbidge AS, Lorenzini SB, Macgowan MJ. Comparative efficacy of online vs. face-to-face group interventions: a systematic review. Res Soc Work Pract 2025 Jul;35(5):524-542. [doi: 10.1177/10497315241236966]

Abbreviations

AASE: alcohol abstinence self-efficacy
ACT: acceptance and commitment therapy
ACT-BA: Acceptance and Commitment Therapy with Behavioral Activation
AUD: alcohol use disorder
BA: behavioral activation
CAD: cumulative abstinence duration
CONSORT: Consolidated Standards of Reporting Trials
COREQ: Consolidated Criteria for Reporting Qualitative Research
DSM-5: *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition)
GEE: Generalized Estimating Equations
ITT: intention-to-treat
MCAR: Missing Completely at Random

XSL•FO
RenderX

MI: multiple imputation
PP: per-protocol
RA: research assistant
RCT: randomized controlled trial
SUD: substance use disorder
TAU: Treatment-as-Usual

Edited by S Brini; submitted 20.Sep.2025; peer-reviewed by MW Huang, TB Xu; revised version received 24.Mar.2026; accepted 25.Mar.2026; published 21.May.2026.

Please cite as:
Belay GM, Lam KKW, Liu Q, Mao T, Mak YW, Ho KY
Effect of Integrated Internet-Based Acceptance and Commitment Therapy and Behavioral Activation Among Ethnic Minority Young Adults With Alcohol Use Disorder in Hong Kong: Pilot Randomized Controlled Trial
J Med Internet Res 2026;28:e83896
URL: https://www.jmir.org/2026/1/e83896
doi:10.2196/83896

© Getaneh Mulualem Belay, Katherine Ka Wai Lam, Qi Liu, Ting Mao, Yim Wah Mak, Ka Yan Ho. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Cardiac Arrest Survivors' Perspectives to Inform the Co-Design of a Web-Based Support and Learning Platform: Qualitative Content Analysis

Annette Waldemar[1,2], RN, PhD; Anders Bremer[3,4], RN, PhD; Johan Israelsson[3,4], RN, PhD; Katarina Heimburg[5,6], FT, PhD; Per Nordberg[7,8], MD, PhD; Erik Blennow Nordström[5,9], PhD; Kristofer Årestedt[1,4], RN, PhD; Ingela Thylén[1,10], RN, PhD

[1]Department of Health, Medicine and Caring Sciences, Linköping University, Linköping, Sweden
[10]Clinical Department of Cardiology, Linköping University Hospital, Linköping, Region Östergötland, Sweden
[2]Clinical Department of Cardiology, Vrinnevi Hospital, Norrköping, Region Östergötland, Sweden
[3]Department of Health and Caring Sciences, Faculty of Health and Life Sciences, Linnaeus University, Växjö, Sweden
[4]Department of Internal Medicine, Kalmar County Hospital, Kalmar, Region Kalmar, Sweden
[5]Neurology, Department of Clinical Sciences, Lund University, Lund, Sweden
[6]Department of Cardiology, Skåne University Hospital, Lund, Sweden
[7]Department of Physiology and Pharmacology, Karolinska Institutet, Stockholm, Sweden
[8]Department of Clinical Sciences and Education, Center for Resuscitation Science, Karolinska Institutet, Stockholm, Sweden
[9]Department of Rehabilitation Medicine, Skåne University Hospital, Lund, Sweden

Corresponding Author:
Annette Waldemar, RN, PhD
Department of Health, Medicine and Caring Sciences, Linköping University, Linköping, Sweden

Abstract

Background: Survivors of cardiac arrest often face multifaceted challenges—cognitive, emotional, physical, and existential—that extend beyond clinical recovery. Despite these long-term consequences, follow-up care is often insufficient, and access to reliable information and support remains limited. Broader initiatives to address post–cardiac arrest care are still lacking. This qualitative study represents the initial phase of a multiphase development process to cocreate, design, and later evaluate a web-based support and learning platform for cardiac arrest survivors. The platform is intended to complement existing health care services and support survivors in managing life after cardiac arrest.

Objective: This study aimed to explore survivors' perspectives on digital support and identify relevant content and delivery formats for a web-based support and learning platform.

Methods: Eight women and 12 men (aged 44‑80 years) were recruited via a moderated peer support network for cardiac arrest survivors. Time since cardiac arrest ranged from 3 months to 19 years. Data were collected between November 2024 and February 2025 through 3 individual and 4 focus group interviews, analyzed using qualitative content analysis.

Results: Three main categories—(1) *digital communication and guided health care navigation*, (2) *digital opportunities to support recovery and address unmet needs*, and (3) *digital and interpersonal pathways to safe social contexts*—were identified as key design requirements for digital support. Substantial gaps in post–cardiac arrest care emerged, including fragmented and sometimes contradictory information, regional disparities, and limited psychosocial follow-up, underscoring the value of a national, accessible, trustworthy web-based program that complements standard care throughout recovery. Flexible formats—such as short videos, read-aloud functions, and information available both as concise and more in-depth versions—were considered essential to accommodate fatigue and cognitive difficulties. A digital platform was further identified as uniquely suited to gather relevant information in one place, provide expert-based explanations and links to further resources, and offer practical tools that could be accessed at home. Across categories, several unmet needs emerged as particularly suited to digital delivery, including guidance on health and everyday decisions, support for managing emotional and physical aftermath, resources to navigate altered social relations, intimacy and personality changes, and dedicated support for family members, who often lack tailored and continuous follow-up.

Conclusions: Findings underscore the need for a tailored digital support program that extends beyond clinical encounters and offers structured, accessible, and personalized guidance across the recovery trajectory. By addressing long-term cognitive, physical, emotional, and relational needs, a contextually adapted digital program has the potential to bridge existing gaps in post–cardiac

XSL•FO
RenderX

arrest care and strengthen survivors' recovery. These user-driven insights provide a foundation for the cocreation and iterative development of a clinically grounded and adaptable digital support platform.

(J Med Internet Res 2026;28:e84432) doi:10.2196/84432

KEYWORDS

cardiac arrest; co-design; cocreation; digital health intervention; focus groups; internet-based intervention; patient participation; patient preference; qualitative research; self-management support; user-centered design

Introduction

Many survivors of cardiac arrest experience cognitive, emotional, physical, and existential challenges that affect their everyday lives [1-4]. These difficulties often manifest as symptoms of anxiety, depression, posttraumatic stress, fatigue [5], sleep disturbances [6,7], and kinesiophobia [8], with fatigue and sleep problems frequently persisting long after the cardiac arrest event [9]. A recent state-of-the-art review highlights that psychological distress after cardiac arrest may also involve heightened vigilance to bodily signals, behavioral avoidance, and physiological hyperarousal, all of which can hinder recovery and increase long-term health risks [10].

The transition from hospital to home is often described as abrupt and disorienting [11,12], and many survivors report a lack of structured follow-up and adequate support to manage life after the event [13]. A poor understanding of the arrest and its consequences can contribute to uncertainty and distress [14]. Evidence from broader cardiac populations suggests that patients who feel unprepared for discharge may be at higher risk of readmission [15], although similar data specific to cardiac arrest survivors are lacking. Conversely, patients who feel well informed report an increased sense of control and a reduced need for emergency care [11]. Despite survivors' desire to discuss the event and receive an early follow-up at cardiac clinics [11,16], many are left without such opportunities [17].

International and national postresuscitation care guidelines emphasize the importance of follow-up within 3 months, including patient education and psychosocial screening [18,19]. However, care after cardiac arrest is often fragmented and unevenly distributed across health care levels and regions [20]. Survivors—particularly those with non–cardiac arrest etiology—may fall outside the scope of specialist clinics and are often left to seek support through primary care or digital platforms [17]. As a result, many turn to online and community-based resources to find knowledge, connection, and tools for self-management [21].

Recent work has highlighted that community-based and nongovernmental organizations play an important role in addressing the diverse informational, emotional, social, and practical needs of cardiac arrest survivors—needs that often extend beyond formal health care services. A recent exploratory survey conducted by the European Resuscitation Council identified only a small number of organizations worldwide that provide structured and continuous support for survivors, including peer support networks and information platforms such as Sudden Cardiac Arrest UK and Heartsight in the United States [21]. These organizations vary widely in their activities and in the extent to which they are connected to national health care systems, with many reporting limited or no formal integration. Although some centers refer patients to these organizations at discharge, their overall integration into formal health care pathways remains limited, and they have not been scientifically evaluated as structured support interventions. Moreover, they are not available in Swedish or tailored to the Swedish health care context, where follow-up practices and access to services differ substantially. There remains a need for evidence-informed, contextually adapted digital support that is developed through a systematic exploration of survivors' needs and cocreated with stakeholders within the national health care system.

Digital support interventions have been developed and evaluated in several other patient groups, including cancer, stroke, and heart failure, where web-based programs have shown potential to improve self-management, psychological well-being, and patient engagement [22-24]. Systematic reviews in these areas highlight the value of accessible, person-centered digital tools that combine education, peer support, and behavioral strategies [25,26]. Together, these findings indicate that digital support can be effective across a range of chronic and acute conditions, but they also highlight the importance of tailoring content and delivery to the specific needs of each survivor group. However, the needs of cardiac arrest survivors differ in important ways. Cognitive impairment, fatigue, memory difficulties, and posttraumatic stress are common [1,5] and may affect how survivors process information and engage with digital content. Moreover, unlike cancer or stroke, there is no established rehabilitation pathway or structured follow-up model for cardiac arrest survivors, and their access to care is often fragmented. These differences underscore the importance of developing support that is specifically tailored to this population and grounded in a systematic exploration of their lived experiences.

Digital solutions, grounded in a person-centered approach, could serve as an accessible means of delivering evidence-based education and psychosocial support regardless of diagnosis, geography, or health care structure [27]. However, to ensure that such interventions are meaningful and relevant, they must be cocreated with input from those who have experienced cardiac arrest.

There is currently limited knowledge about what survivors themselves perceive as relevant in terms of content, structure, and delivery of digital support. Understanding these preferences is a crucial first step in cocreating effective and tailored interventions. This qualitative study therefore forms part of a broader initiative to cocreate and evaluate a digital support program for cardiac arrest survivors, intended as a complement to existing health care services. The aim was to explore

XSL•FO
RenderX

survivors' perspectives on digital support and identify relevant content and delivery formats for a web-based support and learning platform.

Methods

This study adheres to the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist [28].

Study Design

A qualitative design was used, integrating focus group interviews and individual interviews to gain an in-depth understanding of survivors' experiences and preferences regarding the content and delivery of potential support components. The work represents the initial phase of a multiphase development process aimed at cocreating, designing, and later evaluating a web-based support and learning platform for cardiac arrest survivors. The qualitative findings will directly inform the subsequent design and early development of the digital platform.

Setting

The study was conducted in Sweden between November 2024 and February 2025. Participants were recruited via advertisements in the National Peer Support Network for Cardiac Arrest Survivors and Their Family Members. Additional advertisements were placed on the Swedish Heart and Lung Association's national website and local bulletin boards. Survivors who expressed interest to participate contacted the research team via a dedicated study email address. Potential participants were then telephoned by a research team member to receive verbal information about the study with the opportunity to ask questions. If interest persisted, written participant information, a consent form, and a self-administered sociodemographic questionnaire (covering age, gender, employment, geographic location, time since cardiac arrest, and use of digital devices and internet habits) were mailed to the home address. Completed documents were returned in a prepaid envelope. Once consent was confirmed, participants were contacted to indicate their preference for partaking in a focus group or an individual interview. Offering both focus groups and individual interviews ensured that survivors who wished to share their experiences but were uncomfortable participating in a group setting could still take part. This flexible approach enhanced inclusivity and enabled the collection of complementary forms of data.

Participants

Purposive sampling was used to ensure that participants possessed characteristics and experiences relevant to the study aim. Cardiac arrest survivors were recruited through the National Peer Support Network for Cardiac Arrest Survivors and Their Family Members. To obtain a broad range of perspectives, efforts were made to include variation in age, gender, time since the cardiac arrest event, arrest location (in- or out-of-hospital), and geographical region.

Inclusion criteria were (1)participants aged 18 years and older, (2) ability to speak and write Swedish, and (3) survival of a cardiac arrest at least 1 month prior to the interview. These criteria reflected the intended target group for the future program (adults and Swedish-speaking users) and ensured that participants were no longer in the most acute postevent phase. While participants were included according to these criteria, diversity in background characteristics was actively sought.

Data Collection

Data were collected through 4 focus group interviews (4‑5 participants per group) and 3 individual interviews. The sample size was guided by the principle of information power to ensure that it was large and varied enough to elucidate the study aim [29]. The total number of interview sessions and participants was determined in a stepwise manner. Sufficient information power was considered to have been achieved, given the study's narrow aim, the high specificity of the sample, the strong quality of the interview dialogue, and a combined analytical strategy partly focused on in-depth experiential exploration.

The focus groups were conducted in accordance with Krueger and Casey's [30] recommendations regarding group composition, question design, and moderation techniques, while the individual interviews followed Kvale and Brinkmann's [31] principles for qualitative interviewing, emphasizing openness, depth, and reflective listening. The face-to-face focus groups took place in a hospital consultation room and at the Swedish Heart and Lung Association, and the virtual groups were conducted via Zoom (Zoom Communications, Inc). The individual interviews were held at locations chosen by the participants, either in their home or in a hospital consultation room.

Before each interview, the interviewer clarified that the purpose was to explore survivors' perspectives on digital support as a complement to existing health care services, and participants were given the opportunity to ask questions. Both focus groups and individual interviews were guided by a semistructured, pilot-tested interview guide containing open-ended exploratory questions, followed by reflective prompts when appropriate (Textbox 1).

XSL•FO
RenderX

Textbox 1. Interview guide.

Questions:

Before the interview began, the interviewer explained their reasons for, and interest in, the research topic.

Introductory questions

- Could you start by describing your experience following the cardiac arrest and how you managed daily life (eg, what was most challenging or helpful)?
- What kind of support did you receive from those around you (eg, family, friends, and social networks)?

Support from health care

- Can you describe the support you received from health care services after the cardiac arrest?
- How do you perceive the need for support from health care in the aftermath?
- What forms of support do you think would be appropriate (eg, phone calls and in-person meetings—group or individual)?

Educational needs

- What information or guidance did you receive from health care after the cardiac arrest?
- How do you view the need for further information or education in this context?
- In what format could such education be delivered?

Future digital support and education program

- What is your view on the potential for a digital support program, delivered via a website, to complement existing health care services?
- If relevant, what kind of content should such a program include?
- How should the content ideally be presented?

Closing reflections

- Finally, is there anything we haven't discussed that you believe is important to highlight?

Follow-up prompts

- Probing questions such as "Could you tell me more?" "What do you think about that?" or "Have I understood you correctly?" were used to deepen or clarify responses.

The focus groups were moderated by the second author (AB), an ambulance nurse and researcher with extensive experience of qualitative data collection and clinical experience in supporting survivors and their families. The first author (AW), who is a cardiac specialist nurse and researcher with expertise in post–cardiac arrest care, attended as an observer, responsible for taking notes, documenting observations, and posing supplementary questions when appropriate. The individual interviews were conducted by AW using the same semistructured guide.

Participants were encouraged to speak freely, share recovery experiences and perceptions in their own words, and reflect on what aspects of support and content a complementary digital format could provide beyond existing information sources. An open, dialogic atmosphere was fostered through active listening. After participants had first reflected freely on the potential for a digital support and education program and discussed what such a program might include, they were shown a preliminary paper-based schematic overview illustrating suggested content areas and structural components (Figure 1). The overview, developed by the research team and informed by a review of existing peer-driven platforms, such as Heartsight and Sudden Cardiac Arrest UK, was presented as a provisional discussion tool. Participants were informed that its purpose was to elicit feedback rather than to propose a fixed solution. They were invited to comment on the content, suggest additions or modifications, and reflect on the overall structure, including the use of overarching headings with clickable subtopics.

XSL•FO
RenderX

Figure 1. Preliminary paper-based schematic overview used to facilitate discussion about the content and structure of a future co-designed web-based support and learning platform.

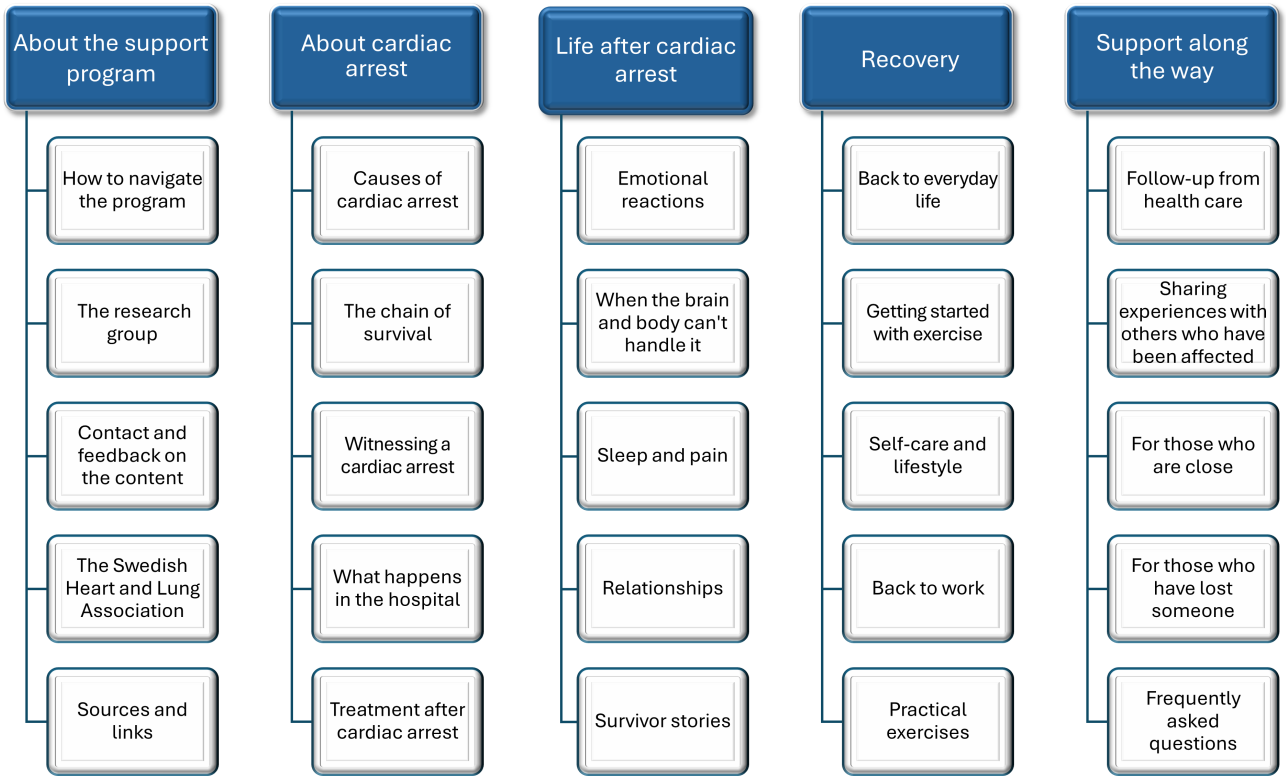

All interviews were audio-recorded and transcribed verbatim by AW. The focus group interviews lasted 96 - 113 minutes and the individual interviews lasted 27 - 44 minutes. In total, the dataset comprised 55,901 words, 835 citations, and 98 single-spaced pages.

Data Analysis

Qualitative content analysis was used to systematically summarize the data while preserving their informational value and contextual depth [32]. The same analytic approach was applied to both focus group and individual interview data. This enabled the identification of recurring categories, patterns, and contrasts that contributed to a nuanced understanding of participants' perceptions and proposals.

According to Krippendorff [33], qualitative content analysis is a systematic method for describing and interpreting verbal communication, with emphasis on contextual meaning rather than on quantification. While some approaches of content analysis may include attention to frequencies, Krippendorff highlights that the primary aim is to generate meaningful interpretations through a transparent and consistent analytic process. This involves systematically developing and applying a coding scheme across the dataset so that the analytic decisions can be followed and scrutinized by other researchers, thereby supporting reliability and validity. Meaning is understood in relation to its broader context, requiring the researcher to attend to both content and context when interpreting the data. The coding process is iterative, with categories being developed, refined, and sometimes merged or split as the analysis progresses, ultimately revealing patterns and interpretations that help explain the data [33].

The initial coding and analysis were conducted by the first author (AW), following a thorough and structured process described by Graneheim and Lundman [32]. First, all transcripts were read repeatedly to obtain a sense of the whole, with attention paid to both diversity and commonalities in participants' experiences, perceptions, and proposals. The text was then divided into coding units, which were subsequently condensed while retaining their core meaning. These units were further labeled with codes. Finally, the codes were grouped into subcategories and main categories to represent the manifest content (Table 1). Emerging interpretations (latent content), subcategories, and main categories were then iteratively discussed with the coauthors to ensure credibility and agreement between authors.

XSL•FO
RenderX

Table . Example of the analysis.

Quote	Coding units	Code	Subcategory	Main category
"Not too long texts on what you see first, then you should be able to delve deeper, I think. But long, really long factual texts are really hard for me. It's hard to read. Brief information should come first. It will make it easier for those with acquired brain damage if they have aphasia or something like that and have it read out to them."	• Not too long texts. • Be able to delve deeper. • Very long factual texts are hard to read. • Brief information should come first. • Make it easier for those with acquired brain damage if they have aphasia to have it read to them.	• Short texts are preferable and should come first. • Longer factual texts should be available for those who want to read more. • Read-out texts help with cognitive impairment.	• Short, easy-to-read texts with audio option.	• Clear and tailored digital communication.
"I also think it would be smart if the support program was in an app, since more and more people are using mobile phones and tablets instead of computers. It doesn't have to be an advanced app, but it would be important that the function is available there."	• Smart if the support program was in an app. • People are using mobile phones and tablets instead of computers. • It doesn't have to be an advanced app.	• The support program must be available as an app. • People use phones and tablets rather than computers.	• The support program should be offered as an app.	• Easier access by an app.

Reflexivity and Positionality

As this study constituted the initial step in a broader co-design process, the researchers were actively engaged throughout data collection and analysis. This involvement, together with the team's clinical and academic experience in post–cardiac arrest care, provided valuable contextual understanding but also introduced a potential risk of preunderstanding shaping the interpretation. To address this, the researchers engaged in continuous reflexive dialogue, critically considering how their prior preunderstanding and commitment to the development process might influence analytic decisions. Coding and emerging interpretations were discussed iteratively to ensure that the analysis remained grounded in participants' accounts rather than researcher assumptions.

Ethical Considerations

The study complies with the principles outlined in the Declaration of Helsinki and received approval from the Swedish Ethical Review Authority (reference number 2024-01840-01). Prior to the focus group sessions, all participants were asked to agree to a confidentiality agreement, ensuring that all discussions remained within the group and that everyone would have an equal opportunity to share their experiences. Participants received both written and verbal information about the study's purpose, the voluntary nature of participation, and their right to withdraw at any time. Written informed consent was obtained from all participants prior to participation in interviews or focus groups. Participant confidentiality and privacy were strictly maintained throughout the study. All participants were assigned pseudonyms, and all interview data were anonymized during transcription. Any potentially identifying information was removed or altered to protect confidentiality. Audio recordings and transcripts were stored securely on password-protected computers accessible only to the research team. The study included participants who had experienced cardiac arrest over an 18-year period, which further reduces the risk of individual identification. Nevertheless, care was taken to ensure that no information presented in the findings could reasonably lead to identification of participants. Data on computer, mobile phone, and internet use, as well as cardiac arrest–specific variables including years since cardiac arrest, age, implantable cardioverter-defibrillator (ICD), location of cardiac arrest, etiology, and post–cardiac arrest follow-up, were collected individually using written questionnaires before the interviews. All questionnaires were coded, and the code list was stored separately in a locked cabinet at the research unit. Access to both the data and the code list was restricted to a single member of the research team in order to ensure participant privacy and confidentiality. No financial compensation was provided. The research team recognized that the interviews could evoke distressing memories or suppressed emotions related to the cardiac arrest event. To support participants, contact details for a designated member of the research team were provided, in case further support was needed. The research team has extensive clinical and research experience working with this patient population, which contributed to ethical sensitivity throughout the study process.

Results

Participant Characteristics

A total of 20 participants took part in the study. Participants ranged in age from 44 to 80 years and included 8 women and 12 men. Time since cardiac arrest varied from 3 months to 19 years. Twelve participants had received an ICD following their cardiac arrest. The sample represented broad geographical diversity, with participants residing in southern, central, and northern Sweden. Myocardial infarction was the most commonly reported underlying cause of the cardiac arrest, particularly

XSL•FO
RenderX

among men (Table 2). All participants reported daily use of both a computer and a smartphone, and 19 participants used the internet daily. Six participants also used a tablet.

Table . Background characteristics.

Characteristics	Women (n=8)	Men (n=12)	Total (N=20)
Years since cardiac arrest, median (range)	2.5 (0 - 8)	7.0 (3-19)	3.5 (0 - 19)
Age (years), median (range)	53.0 (44-63)	67.0 (46-80)	62.5 (44-80)
ICD[a] recipient (yes), n (%)	4 (50)	8 (67)	12 (60)
Location of cardiac arrest, n (%)			
At home	4 (50)	3 (25)	7 (35)
In a public place	3 (37.5)	9 (75)	12 (60)
In hospital	1 (12.5)	__[b]	1 (5)
Etiology of cardiac arrest, n (%)			
Myocardial infarction	2 (25)	10 (83)	12 (60)
Arrhythmia	1 (12.5)	1 (8.5)	2 (10)
Cardiomyopathy	1 (12.5)	—	1 (5)
Pulmonary embolism	1 (12.5)	—	1 (5)
Other	3 (37.5)	1 (8.5)	4 (20)
Received hospital follow-up (yes), n (%)	6 (75)	11 (92)	17 (85)

[a]ICD: implantable cardioverter-defibrillator.

[b]Not available.

Categories and Subcategories

The results illustrate survivors' preferences regarding the design, content, and delivery of user-friendly digital support. These preferences were grounded in their lived experiences of recovery, which shaped their expectations of what a digital program should provide. In describing these experiences, participants also reflected on how a digital support resource could address the needs they identified. The findings are presented in 3 main categories and 6 subcategories (Figure 2 [11]).

XSL•FO
RenderX

Figure 2. Conceptual overview of 3 core unmet needs identified from cardiac arrest survivors' interviews (Waldemar et al [11]). The categorization reflects survivors' views of digital support as a complement to existing care structures. The figure illustrates survivors' calls for (1) digital communication and guided health care navigation, (2) digital opportunities to support recovery and address unmet needs, and (3) digital and interpersonal pathways to safe social contexts.

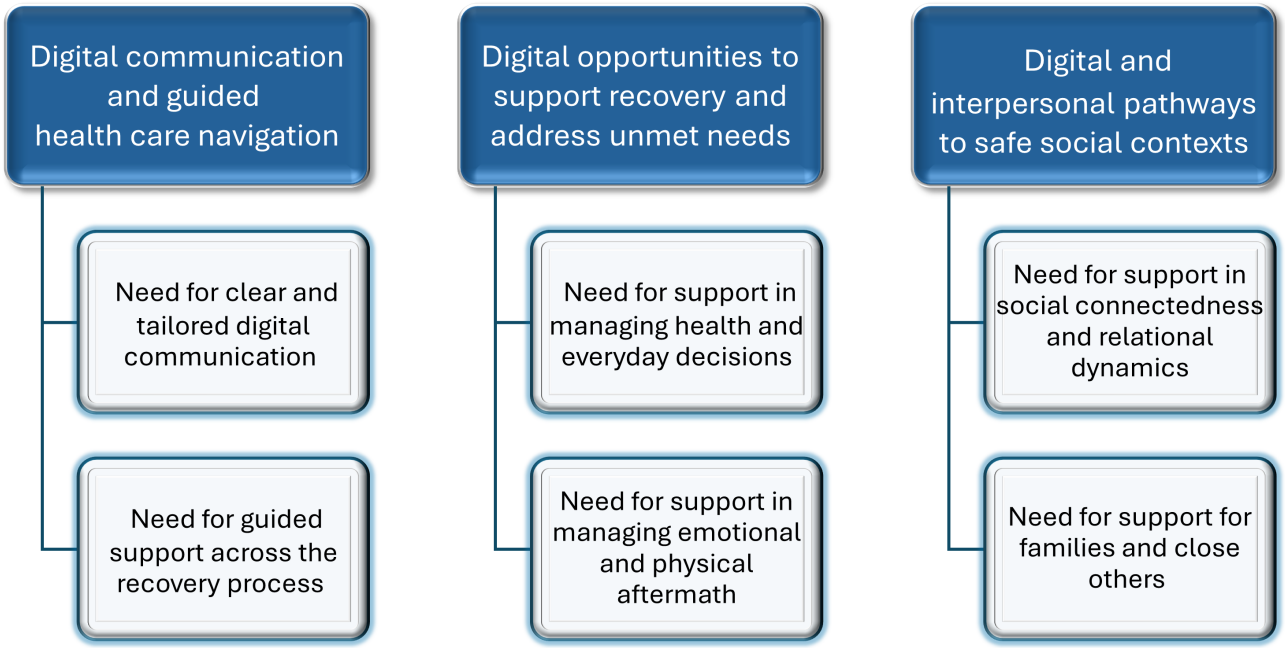

Digital Communication and Guided Health Care Navigation

Dispersed and sometimes contradictory information from different health sources created a need for a digital program that integrated, clarified, and tailored guidance across the care continuum. Effective digital accessibility was understood to depend not only on technical design but also on usability, linguistic clarity, and cognitive manageability. Digital communication that was easy to navigate, easy to process, and available in multiple formats—such as links and videos—enhanced inclusivity by accommodating diverse user needs and capabilities. The centralization of credible and trustworthy resources within a single, intuitive interface was regarded as essential for reducing exposure to misinformation. Trust in such systems was shaped by the degree of transparency, clarity of information sources, and the presence of professional endorsement. Structured digital guidance was considered necessary across all phases of care—from the cardiac arrest event and prehospital interventions to acute hospital treatment, discharge, and long-term adjustment.

Need for Clear and Tailored Digital Communication

A central aspect of participants' vision for a digital support program was the need for communication that was cognitively accessible, emotionally attuned, and practically useful. Survivors described how difficulties with memory, concentration, and information processing made it challenging to engage with conventional health information, particularly when written in complex language or presented in fragmented ways. This often led to confusion and disengagement: "It all gets muddled in my head. I end up just reading headlines in the news—anything more and I lose focus."

To address these challenges in a digital format, participants advocated for layered communication: short, plain language summaries with optional links to more detailed content. Concise, easily accessible, and multimodal information was consistently emphasized. Short texts were preferred, as extensive reading was perceived as overwhelming, and participants valued the option to access more in-depth material through linked pages. Detailed explanatory texts, including glossaries, were requested: "Hospital language is something you can't keep up with...nothing makes sense. So, a glossary would really help." These features were seen not only as tools for understanding but also as signs of respect for the cognitive realities many face after cardiac arrest.

Multimodal formats were considered essential. Participants highlighted the importance of text-to-speech functionality and translations into widely spoken languages, such as English and Arabic, to enhance accessibility: "It would be great if the text could be read out loud as you go." Visual communication was also emphasized, particularly for those experiencing cognitive fatigue. Participants appreciated having alternative formats, such as the option to engage with content through text or video. Short videos or read-aloud text were described as easier to interact with than traditional written material.

The inclusion of lived experience quotations was suggested as a way to foster recognition and emotional connection within the digital environment. Visual elements were viewed as important for supporting understanding, with a clear preference for photographs over stylized illustrations: "They used illustrated images at first and I couldn't make sense of them. But when they switched to real photos, I understood immediately."

Participants also envisioned the platform as a space for connection with trusted networks. Survivor narratives were considered valuable for reducing feelings of isolation and

XSL•FO
RenderX

fostering recognition: “I just wanted to talk to someone who’d been through the same. Someone who gets it.” The inclusion of family members’ perspectives was also requested: “Families’ stories are important too—so they feel seen and recognised.”

Need for Guided Support Across the Recovery Process

Participants emphasized the importance of including structured guidance throughout all stages of the recovery in the support program. A recurring wish was for step-by-step information that could clarify what had happened, explain what to expect, and provide guidance moving forward. Many survivors described the acute phase—particularly the hospital stay following an out-of-hospital cardiac arrest—as fragmented or entirely absent from memory. To support understanding of this phase, participants proposed a dedicated section or timeline within the program, tentatively titled “What happened at the hospital,” outlining the care trajectory from ambulance arrival to discharge. Emergency interventions such as cardiopulmonary resuscitation (CPR), defibrillation, intraosseous access, and mechanical chest compressions were frequently mentioned as sources of both physical pain and emotional confusion.

Experiences from the intensive care unit (ICU) were similarly described as surreal or frightening. Participants stressed the importance of including information in the digital program that could explain these experiences and normalize common reactions. “I was hallucinating after I woke up...Later I was told this is common after being in the ICU. But at the time, I thought I was going mad.”

The transition from hospital to home was identified as a particularly vulnerable phase, and participants called for this to be addressed explicitly in the program. While some described receiving adequate support, others reported being discharged without sufficient preparation or follow-up. “It’s incredibly valuable when someone from the hospital calls and asks, ‘Is there anything you need?’ Rather than just disconnecting you from the telemetry and leaving you to figure it out on your own.”

Follow-up care was experienced as inconsistent and dependent on regional practices. Participants emphasized that existing national guidelines for post–cardiac arrest care are not uniformly implemented and suggested that these should be clearly presented within the program. Access to cardiac rehabilitation was described as essential for regaining both confidence and physical strength. “When I was transferred to primary care, all care just stopped...I didn’t know who to turn to.” To address these gaps, participants proposed that the digital program should include a clearly structured pathway, outlining key stages and responsibilities across the recovery timeline. These suggestions reflected participants’ desire for a coherent and continuous source of guidance across the entire recovery journey.

Digital Opportunities to Support Recovery and Address Unmet Needs

Challenges in managing the physical, cognitive, and emotional consequences of cardiac arrest highlighted the need for digital resources that provided comprehensive, long-term recovery support extending beyond standard medical follow-up. Recovery was characterized as a multifaceted process involving both clinical management and ongoing emotional adjustment. Digital support was required to improve understanding of causes, treatments, and medications, as well as to address managing cognitive and physical difficulties such as fatigue, pain, and memory impairment. Unmet needs related to psychological and existential concerns were evident, and limited access to counseling or psychological interventions underscored the importance of incorporating dedicated modules addressing these aspects of recovery. Clear, accessible, and adaptable guidance was required to support survivors in navigating the often-overlooked phase of postdischarge life in ways that current information pathways do not consistently provide. This need was best supported by a flexible, modular design capable of adjusting to varying needs, preferences, and recovery trajectories.

Need for Support in Managing Health and Everyday Decisions

Participants were asked to reflect on how challenges could be supported digitally. They highlighted a range of topics that they believed should be addressed in a digital program to help navigate everyday decisions and uncertainties after cardiac arrest. Existing written and oral patient information was often described as insufficient, fragmented, or difficult to revisit, whereas a digital format could offer structured explanations, tailored guidance, and practical tools such as video-based exercises to support long-term recovery. Participants appreciated the idea of having all relevant information gathered in one place, presented in an accessible way, and available both as concise and more in-depth text:

> *During the first year, I searched for an enormous amount of information. I wanted to hear other people’s stories, see their experiences. I listened to some British podcast that I thought was good at the time, and I read a lot of scientific papers—but that’s a language you just can’t follow. And these general brochures for us ordinary folks...eventually you’ve read all of them, and then you end up with the medical papers instead. And you’re really interested and curious because it feels spot-on for you, but you don’t understand a thing.*

A central concern was the lack of clarity regarding the cause of the arrest. This uncertainty was described as difficult to live with and generated a strong desire for accessible on-demand information in a consolidated space with explanations of possible underlying mechanisms. “I wish it had been something with my heart they could fix, or that the ICD made me feel safe. But they don’t even know why I had the arrest.” Some participants described confusing cardiac arrest with myocardial infarction and emphasized the importance of having key medical concepts explained in clear, plain language within the program.

Understanding prescribed medication was another area requiring targeted support centralized within 1 trustworthy digital resource. Survivors often felt underinformed about the purpose of their medications and were hesitant to seek clarification online due to fear of misinformation or unnecessary anxiety. “Why am I taking beta-blockers? I still don’t know...I don’t want to Google too much because then you think you have every side effect.” Survivors also described discontinuing medications

XSL•FO
RenderX

because of side effects or unclear instructions. In response, participants proposed that the digital program should offer personalized, trustworthy information about medications and their role in secondary prevention.

Living with an ICD was another example where participants expressed a strong desire for digital structured information and emotional guidance, since existing health resources were experienced as insufficient, diffuse, and sometimes contradictory. Similarly, survivors who had undergone interventions following myocardial infarction requested clearer explanations of what had been done and what the procedures meant for their long-term health.

Cognitive impairments were described as a prominent and often unexpected consequence, with longer-lasting effects than many survivors had anticipated. Acceptance of these changes was reported to take time. Participants therefore recommended that the program should include practical cognitive rehabilitation exercises, such as memory training tools, occupational therapy, and guidance on reintegration into work and everyday activities. “In the beginning, it was really hard to accept help...But the rehab has been incredibly helpful.”

Preventing recurrence and regaining a sense of control were described as unmet needs that were particularly suited to digital delivery, especially when expert-based information with links for further reading and practical exercises could be accessed from home in a calm and familiar setting. Participants requested practical, evidence-based advice within the program on everyday concerns, including diet, alcohol, physical activity, sexual health, and sleep. Exercise, in particular, was described as anxiety-provoking, with many survivors uncertain about what levels of exertion were safe. “My brain doesn’t want to push too hard...I asked for an extra exercise test...Now I know I can get my heart rate up to 165 without it being dangerous.”

Need for Support in Managing Emotional and Physical Aftermath

In addition to medical and practical concerns, participants highlighted the need for the digital support program to address the emotional and embodied consequences of surviving a cardiac arrest. Many described persistent physical sensations—such as chest discomfort, palpitations, and breathlessness—that triggered fear and uncertainty. As one participant asked: “If I get pressure on my chest, should I go to the emergency department? I’ve no idea.” To meet this need, participants expressed a strong desire for access to digital resources and interventions to support symptom management and receive guidance on interpreting bodily signals and recognition when medical attention is warranted. Pain from CPR-related injuries, such as rib fractures, further compounded this distress. One participant recalled: “I was in so much pain, and everyone just laughed. Just a smile—‘at least you’re alive.’”

Sleep disturbances emerged as another significant and underaddressed issue in existing written information. Participants described racing thoughts at bedtime, fear of dying in their sleep, panic attacks during the night, and recurring nightmares. These experiences highlighted a need for video-based exercises and strategies to improve sleep, manage nocturnal anxiety, and recognize trauma-related symptoms.

Cognitive impairments, including memory problems, fatigue, and what participants referred to as “brain fog,” were among the most disruptive aspects of recovery. Many reported struggling with concentration, “zooming out,” or falling asleep during conversations or while sitting upright. Participants stressed that the program should explain the neurological effects of cardiac arrest and provide realistic information about recovery trajectories, including the slow and often unpredictable nature of cognitive rehabilitation. To accommodate difficulties with concentration and fatigue, participants emphasized the importance of offering this material not only in written form but also as short, video-based content for days when reading felt too demanding. As one participant noted: “I thought I’d be back at work by December...there’s no way. I’m exhausted.”

Existential fear and emotional upheaval were also prominent themes, which participants felt warranted a dedicated module within the program. Survivors described profound shifts in priorities, relationships, and identity. While some expressed gratitude for survival, others lived with a persistent fear of recurrence that shaped everyday life. One participant reflected: “My child now says ‘goodbye’ instead of ‘goodnight.’ That’s how scared we are.” Several participants noted that these fears were often minimized or dismissed by health care providers. “My doctor said, ‘But it’s been quite a while since your arrest,’ even though I’d had a second one two years later.”

Participants emphasized that the program should provide clear information about available psychological support services and offer guidance on how to access appropriate help when needed. The need to talk openly about these emotional and existential experiences was described as both urgent and unmet. Participants expressed frustration that family members, friends, and even clinicians often avoided such conversations. Psychological interventions and cardiac-specific counseling were highlighted as crucial yet inconsistently available resources, with most participants reporting that they had received no psychological follow-up.

Digital and Interpersonal Pathways to Safe Social Contexts

Relational and emotional dynamics following cardiac arrest highlighted the need for a digital program that normalized common reactions, supported family communication, and provided guidance on intimacy, safety, and social reintegration. Cardiac arrest often disrupted established relationship patterns and generated fears related to solitude and vulnerability, underscoring the importance of connection, closeness, and relational stability. The broader social context required mechanisms that addressed intimacy, everyday interactions, and the reestablishment of a sense of interpersonal safety.

Family members emerged as essential sources of support, yet they were also placed in vulnerable positions and were frequently overlooked within formal recovery structures. These dynamics indicated a need for a digital resource that offered counseling, practical guidance, and tools to help families cope with emotional strain, monitor health-related changes, and

XSL•FO
RenderX

navigate the challenges of daily life. A digital support needed to incorporate clear information, practical caregiving strategies, and validation of family experiences, ensuring that family involvement was recognized as an essential component of survivor recovery.

Need for Support in Social Connectedness and Relational Dynamics

Participants discussed how a digital resource could help normalize relational experiences and provide guidance that could help navigate the social consequences of cardiac arrest that many described. Participants emphasized that the digital support program should address the profound social and relational consequences of surviving a cardiac arrest. The event was described as a turning point that disrupted identity and altered relationships, both with the self and with others. Many reported feeling like a different person after the arrest, often struggling with changes in personality or self-image: “It’s really strange in my head. Who am I now? I’m not myself anymore—such a huge part of my personality has been taken from me.” These internal changes were frequently reflected in how survivors were perceived and treated by others. Well-meaning but overprotective responses from family and friends could feel diminishing: “It feels like everyone looks down on me, even my children—who actually need their mother. That part is quite hard.”

Participants stressed that the program should include information about how cardiac arrest can affect social roles and offer practical strategies for navigating altered relational dynamics. Parent-child relationships were often reshaped, with children—both young and adult—assuming new responsibilities or roles. The program was seen as a potential digital space for sharing strategies to support families in understanding and negotiating each other’s ways of coping.

A recurring concern was the fear of being alone, both physically and emotionally. Survivors described intense feelings of isolation, exacerbated by the rarity of their experience and uncertainty about recurrence: “My first thought when I woke up was—I can’t be alone in this. But I felt very alone.” Participants suggested that the program could help normalize these reactions and provide guidance on how to seek emotional and practical support when needed. In addition to helping survivors cope with these stressors, digital peer support was noted as a means of supporting social connectedness for both survivors and close family.

Many also described social withdrawal or distancing by friends and colleagues, which they attributed to others’ uncertainty about how to act or respond. This was experienced as deeply painful and, in some cases, led to a profound sense of abandonment. Several participants also reported the breakdown of romantic relationships following the arrest, frequently related to emotional strain or unresolved trauma: “My boyfriend couldn’t handle it—it ended there. So now I feel very alone.”

Sexuality and intimacy were additional areas participants felt must be explicitly addressed in the program since these were lacking in existing written information. Survivors described how cardiovascular medications negatively affected sexual desire and function, issues rarely acknowledged in clinical encounters: “Beta-blockers and blood pressure meds are no party boosters...It would’ve been much easier if someone had just said: Things might not work so well in bed.” Participants emphasized that existing support was fragmented, difficult to access, or entirely absent, whereas a digital format could offer normalizing information, family-oriented guidance, and practical tools to strengthen communication, intimacy, and social reintegration. In this way, participants viewed a digital program as a means to create safer social contexts and to support the relational dimensions of recovery that are often overlooked in standard care.

Many worried that sexual activity might trigger another cardiac arrest, particularly in relation to arrhythmias, but felt too embarrassed or uncertain to raise these concerns. When such topics were eventually discussed, responses were often reassuring, underscoring the need for the program to provide clear, proactive, and destigmatizing information about sexual health and intimacy after cardiac arrest. Overall, participants underscored that rebuilding meaningful social connections and addressing intimacy concerns were essential components of recovery and should be treated as such within the digital support program.

Need for Support for Families and Close Others

Participants highlighted the need for a dedicated section in the digital support program for family members, describing them as both central to the cardiac arrest experience and profoundly affected by it. Despite their crucial role, family members were often perceived as overlooked by health care professionals: “There’s a lot of focus on the one who survived...but someone else was there too, watching over you in those first two days.” Because family members rarely receive formal follow-up or information tailored to their specific needs, participants viewed a digital format as particularly suitable for providing accessible, reliable, and continuous support.

Survivors recounted how traumatizing the cardiac arrest had been for their loved ones—sometimes more than for themselves. Many family members received abrupt phone calls, arriving at the hospital to find the survivor sedated and unresponsive: “We said goodbye in the morning, and then they get a phone call saying I’ve had a cardiac arrest.” Witnessing CPR, waiting through prolonged transfers, or being separated during emergencies contributed to a pervasive sense of helplessness and lasting trauma. Participants therefore viewed a program module explaining early care trajectories and common reactions among family members as particularly valuable during the acute phase. As such content is largely absent from existing written materials, a digital platform available 24/7 was considered an important way to provide structured explanations and reassurance during this initial period.

Children and adolescents were described as especially vulnerable yet rarely offered appropriate support. Some had performed CPR or witnessed the arrest and subsequently struggled with long-term fear and anxiety: “My husband and son, who were fourteen at the time, did CPR. The kids didn’t cope very well...they were frightened for a long time.” Attempts to access psychological support for children were often unsuccessful.

XSL•FO
RenderX

Children and grandchildren were deeply affected by the event, sometimes developing separation anxiety or persistent worry about the survivor's safety. Participants reported that they did not fully recognize the extent of their loved ones' trauma until much later and therefore emphasized the need for age-appropriate guidance and tools for supporting children and young people. A digital program was viewed as a practical way to provide families with accessible support when professional services were unavailable.

Partners and close family members were also described as being physically affected by the emotional toll, developing stress-related illness or severe exhaustion. Participants clearly articulated that when the needs of family members are overlooked, this has a direct impact on their own recovery: "Support for the family is also support for us, the survivors." As family members often carry this responsibility without preparation or guidance, participants saw digital support as a way to provide timely information and reassurance in situations where health care contact was limited.

Following hospital discharge, family members frequently assumed extensive caregiving responsibilities without preparation or guidance. They were expected to monitor symptoms, interpret medical information, and make decisions amid uncertainty. Vigilant behaviors—such as checking breathing at night or using apps to track the survivor's location, activity, or heart rhythm—were common and driven by the family's own fear of recurrence. Cognitive impairments, fatigue, and emotional changes in the survivor further intensified this burden.

Friends and bystanders who had performed CPR or witnessed the event were likewise described as emotionally affected but received no follow-up, debriefing, or support. Digital peer support was also viewed as uniquely suited to these groups, as informal networks are otherwise difficult to access and no formal follow-up is routinely offered: "It must have been terrifying. They should have been offered help, but nobody contacted them."

Discussion

Principal Findings

This paper explored cardiac arrest survivors' perspectives on the essential content and structure of a web-based support and learning platform. Survivors described a need for digital resources that are tailored, cognitively accessible, emotionally attuned, and integrated with existing health care services. While many of the cardiac arrest experiences identified have been reported previously, our findings extend this knowledge by demonstrating how these experiences translate into concrete design requirements for a digital support program. These insights represent the initial phase of a multiphase development process to cocreate, design, and later evaluate a web-based support and learning platform for cardiac arrest survivors.

Existing community-based and online resources for cardiac arrest survivors demonstrate the value of peer-driven information and shared lived experiences. Recent international surveys also show that such organizations play an important role in addressing unmet informational and emotional needs and yet remain only loosely connected to formal health care systems. Participants in our study described needs that extend beyond what these initiatives currently offer, including structured, clinically endorsed, and contextually adapted digital support. Survivors highlighted the importance of trustworthy information, guidance across care transitions, and content aligned with their cognitive, emotional, and practical recovery needs. Our findings therefore contribute to user-driven design requirements that can inform the development of a more comprehensive, evidence-informed digital support program tailored to the Swedish health care context.

Comparison With Prior Work

Participants' difficulties with memory, concentration, and information processing align with previous research on health literacy and digital engagement among individuals with cognitive impairments [34,35]. Their preferences for layered, plain language communication are consistent with principles of universal design [36]. Requests for visual aids, text-to-speech functionality, and multilingual options echo recommendations in the WHO Global Strategy on Digital Health 2020 - 2025 [37] and reinforce the importance of inclusive digital tools that accommodate cognitive diversity [38,39].

Survivors' memory gaps during the acute phase, particularly in the ICU, are well documented [3,40,41]. Our findings extend this work by showing how these gaps translate into a specific digital need: a clear, structured timeline explaining what happened; why certain procedures were performed; and what to expect next. Such features may support meaning-making, reduce distress, and enhance trust in the recovery process.

Participants' desire for step-by-step guidance across the recovery trajectory resonates with the call for rigorous interventional studies addressing survivor- and family-focused support by Douma et al [42]. Transfers between care settings were described as emotionally significant, echoing earlier work showing that families often experience the ICU as both traumatic and reassuring, with distress increasing during transitions [43]. Digital support may help bridge these gaps by providing consistent information, clarifying roles, and outlining available services.

The cognitive and emotional aftermath of cardiac arrest—including fatigue, memory loss, anxiety, and existential distress—has been widely reported [44-46]. Cognitive impairment, depression, and reduced mobility have been linked to lower societal participation [47]. Our findings add to this literature by identifying digital features survivors believe would support recovery, such as modules explaining neurological recovery, strategies for managing fatigue, and guidance on when to seek medical attention. Participants' reports of feeling dismissed by health care providers underscore the need for digital content that normalizes common symptoms, sets realistic expectations, and offers psychological coping strategies.

Survey data from the Danish Cardiac Arrest Survivors study further highlight the long-term nature of these challenges, with many survivors reporting persistent symptoms and incomplete recovery up to 5 years postevent [48]. Similarly, Mion et al [7]

XSL•FO
RenderX

found that survivors and families prefer early follow-up within the first month. These findings reinforce the need for digital tools that provide timely, realistic, and ongoing support.

Survivors' descriptions of altered identity, family role shifts, and challenges related to intimacy and social reintegration align with research on traumatic brain injury [49,50] and studies showing that living with an ICD can contribute to marital strain [43]. Our study extends this work by demonstrating how these relational challenges translate into digital needs, including content on communication within families, guidance on intimacy and sexuality, and peer stories that normalize emotional reactions. Southern et al [43] similarly describe the postarrest period as one of existential reevaluation and renegotiation of roles, suggesting that digital support may help survivors and families navigate these transitions through psychoeducation and coping strategies.

Participants frequently described their family members as overlooked in the recovery process, echoing Haywood and Dainty's [51] description of family members as "forgotten patients." Evidence that survivors and their family members can influence each other's health outcomes and recovery trajectories [52] and calls for family-centered care in cardiac arrest survivorship [53] further support the need for dedicated family-oriented content. A recent scoping review by Douma et al [42] identified 5 key domains of family support: attention during the cardiac arrest, collaboration with the resuscitation team, contextual understanding, postresuscitation needs, and implementation of family-centered care. These domains align closely with the needs identified by participants in our study. Emerging digital psychoeducational interventions targeting survivors and families [54] highlight growing interest in this area. Our findings suggest that family-centered content should be integrated as a core component of digital support programs rather than a supplementary component.

Implications for Digital Intervention Development

Taken together, our findings indicate that digital support tools for cardiac arrest survivors should not only deliver medical information but also foster emotional connection, peer support, and family inclusion across care transitions. Grounding design in survivors' experiences and preferences can help ensure that digital interventions are responsive and accessible. In this way, cocreating a digital support tool is not merely a technical endeavor but reflects survivors' call for more consistent, equitable, and person-centered care after cardiac arrest. Such tools may complement emerging psychoeducational efforts and address persistent gaps in current care pathways, including the limited integration of existing community-based support organizations into formal health care systems. In doing so, this study addresses a critical gap in current survivorship research by specifying the user-driven requirements necessary for developing effective, scalable digital support.

Strengths and Limitations

A key strength of this study lies in its grounding in the experiences of end users. By actively incorporating survivors' perspectives in a cocreative process, the research aligns with person-centered principles and enhances the relevance and applicability of its findings to real-world care and support systems.

Purposive sampling enabled the inclusion of participants with diverse experiences of cardiac arrest, capturing variation in underlying etiologies, age, time since the event, and geographic location. Participants had been hospitalized in different health care settings and received varying degrees of follow-up, which strengthened the contextual richness and transferability of the findings. However, recruitment through peer networks may have favored individuals with higher engagement or digital literacy, potentially limiting the diversity of perspectives.

Focus group interviews generated rich data through participant interaction, complemented by individual interviews that offered flexibility and inclusiveness. Smaller focus groups facilitated open discussions, including existential reflections, and digital interviews enabled participation from across Sweden. This combination provided both breadth and depth of insight [55].

The research team's clinical and academic background in post–cardiac arrest care may have shaped interpretive sensitivity. Although reflexive practice was used to minimize bias, researcher preunderstanding may still have influenced interpretation. Although generalizability is not the goal of qualitative research, the findings are likely transferable to similar populations outside Sweden. While individual recovery experiences and support needs are inherently personal, the study contributes knowledge that can guide the development of tailored interventions.

Conclusions

This study demonstrates how cardiac arrest survivors' multifaceted recovery experiences can be translated into actionable user-driven design requirements for a web-based support program. The findings indicate a clear need for digital support that extends beyond clinical encounters by addressing persistent gaps in follow-up, providing accessible and cognitively manageable information, offering dedicated resources for family members, and delivering stepwise guidance across care transitions. A coherent digital format—combining concise and in-depth text, short videos, read-aloud functions, and practical tools—emerged as central to accommodating fatigue, cognitive difficulties, and the need to revisit information throughout recovery. These insights outline key priorities for a clinically grounded and contextually adaptable web-based support program positioned to strengthen long-term recovery and everyday functioning for cardiac arrest survivors, forming the foundation for cocreation and iterative development within this multiphase project.

Acknowledgments

The authors would like to thank all the participants who gave of their valuable time to participate in the study. The authors would also like to express their gratitude to the Swedish Network for Cardiac Arrest Survivors and their Relatives and the Swedish Heart

XSL•FO
RenderX

and Lung Association for their valuable support in carrying out the study. No generative artificial intelligence tools were used at any stage in the preparation of this manuscript.

Funding

This study was funded by FORSS (FORSS-994496).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: AW, AB, JI, KÅ, IT
Data curation: IT
Formal analysis: AW, IT
Funding acquisition: IT
Investigation: AW, AB
Methodology: AW, AB, JI, KÅ, IT
Project administration: AW
Supervision: IT
Validation: AB, JI, KÅ, IT
Visualization: AW
Writing – original draft: AW
Writing – review & editing: AB, IT, JI, KÅ, KH, PN, EBN

Conflicts of Interest

None declared.

Checklist 1
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[PDF File, 495 KB - jmir_v28i1e84432_app1.pdf]

References

1. Zook N, Voss S, Blennow Nordström E, et al. Neurocognitive function following out-of-hospital cardiac arrest: a systematic review. Resuscitation 2022 Jan;170:238-246. [doi: 10.1016/j.resuscitation.2021.10.005] [Medline: 34648921]
2. Chen X, Li D, He L, et al. The prevalence of anxiety and depression in cardiac arrest survivors: a systematic review and meta-analysis. Gen Hosp Psychiatry 2023;83:8-19. [doi: 10.1016/j.genhosppsych.2023.03.013] [Medline: 37028095]
3. Moulaert VRM, van Heugten CM, Gorgels TPM, Wade DT, Verbunt JA. Long-term outcome after survival of a cardiac arrest: a prospective longitudinal cohort study. Neurorehabil Neural Repair 2017 Jun;31(6):530-539. [doi: 10.1177/1545968317697032] [Medline: 28506147]
4. Heimburg K, Blennow Nordström E, Dankiewicz J, et al. Low physical activity level in out-of-hospital cardiac arrest survivors with obesity, mobility problems and cognitive impairment: results from the TTM2 trial. Resuscitation 2024 Nov;204:110407. [doi: 10.1016/j.resuscitation.2024.110407] [Medline: 39368797]
5. Yaow CYL, Teoh SE, Lim WS, et al. Prevalence of anxiety, depression, and post-traumatic stress disorder after cardiac arrest: a systematic review and meta-analysis. Resuscitation 2022 Jan;170:82-91. [doi: 10.1016/j.resuscitation.2021.11.023] [Medline: 34826580]
6. Hellström A, Bremer A, Gunnarsson LL, Hjelm C. Sleep in cardiac arrest survivors. Nurs Crit Care 2023 Nov;28(6):870-877. [doi: 10.1111/nicc.12843] [Medline: 36168674]
7. Mion M, Case R, Smith K, et al. Follow-up care after out-of-hospital cardiac arrest: a pilot study of survivors and families' experiences and recommendations. Resusc Plus 2021 Sep;7:100154. [doi: 10.1016/j.resplu.2021.100154] [Medline: 34386781]
8. Heimburg K, Nordström EB, Friberg H, et al. Comparison of self-reported physical activity between survivors of out-of-hospital cardiac arrest and patients with myocardial infarction without cardiac arrest: a case-control study. Eur J Cardiovasc Nurs 2025 Jul 21;24(5):700-709. [doi: 10.1093/eurjcn/zvaf032] [Medline: 39999023]
9. Evald L, Skrifvars MB, Virta JJ, et al. Long-term functional outcome and quality of life 5-8 years after out of hospital cardiac arrest. Resuscitation 2025 Oct;215:110672. [doi: 10.1016/j.resuscitation.2025.110672] [Medline: 40499675]
10. Agarwal S, Wagner MK, Mion M. Psychological and behavioral dimensions in cardiac arrest survivors and their families: a state-of-the-art review. Neurotherapeutics 2025 Jan;22(1):e00509. [doi: 10.1016/j.neurot.2024.e00509] [Medline: 39709245]

XSL•FO
RenderX

11. Waldemar A, Strömberg A, Thylén I, Bremer A. Experiences of family-witnessed cardiopulmonary resuscitation in hospital and its impact on life: an interview study with cardiac arrest survivors and their family members. J Clin Nurs 2023 Oct;32(19-20):7412-7424. [doi: 10.1111/jocn.16788] [Medline: 37300340]
12. Rojas DA, DeForge CE, Abukhadra SL, Farrell L, George M, Agarwal S. Family experiences and health outcomes following a loved ones' hospital discharge or death after cardiac arrest: a scoping review. Resusc Plus 2023 Jun;14:100370. [doi: 10.1016/j.resplu.2023.100370] [Medline: 36909925]
13. Jensen AN, Bonnén KB, Kristiansen M. "We don't talk about his heart": narrative sense-making and long-term readjustment among older out-of-hospital cardiac arrest survivors and their spouses. Resusc Plus 2020 Sep;3:100024. [doi: 10.1016/j.resplu.2020.100024] [Medline: 34223307]
14. Presciutti A, Siry-Bove B, Newman MM, et al. Qualitative study of long-term cardiac arrest survivors' challenges and recommendations for improving survivorship. J Am Heart Assoc 2022 Jul 19;11(14):e025713. [doi: 10.1161/JAHA.121.025713] [Medline: 35861822]
15. Crawshaw J, Bartoli-Abdou JK, Weinman J, et al. The transition from hospital to home following acute coronary syndrome: an exploratory qualitative study of patient perceptions and early experiences in two countries. Int J Pharm Pract 2021 Feb 10;29(1):61-69. [doi: 10.1093/ijpp/riaa009] [Medline: 33793821]
16. Dichman C, Wagner MK, Joshi VL, Bernild C. Feeling responsible but unsupported: how relatives of out-of-hospital cardiac arrest survivors experience the transition from hospital to daily life—a focus group study. Nurs Open 2021 Sep;8(5):2520-2527. [doi: 10.1002/nop2.779] [Medline: 33638939]
17. Israelsson J, Lilja G, Bremer A, Stevenson-Ågren J, Årestedt K. Post cardiac arrest care and follow-up in Sweden—a national web-survey. BMC Nurs 2016;15:1. [doi: 10.1186/s12912-016-0123-0] [Medline: 26752975]
18. Svenska riktlinjer för uppföljning efter hjärtstopp [Swedish guidelines for follow-up after cardiac arrest]. Swedish Resuscitation Council. URL: https://www.hlr.nu/wp-content/uploads/2021/11/Svenska.riktlinjer.2021.uppfoljning.pdf [accessed 2026-04-17]
19. Nolan JP, Sandroni C, Cariou A, et al. European Resuscitation Council and European Society of Intensive Care Medicine guidelines 2025: post-resuscitation care. Intensive Care Med 2025 Dec;51(12):2213-2288. [doi: 10.1007/s00134-025-08117-3] [Medline: 41123621]
20. Holmér S, Nedlund AC, Thomas K, Krevers B. How health care professionals handle limited resources in primary care—an interview study. BMC Health Serv Res 2023 Jan 3;23(1):6. [doi: 10.1186/s12913-022-08996-y] [Medline: 36597086]
21. Gamberini L, Haywood KL, Schnaubelt S, et al. Organisations supporting cardiac arrest survivors: an exploratory survey of organisational structures and activities. Resusc Plus 2025 Jul;24:100986. [doi: 10.1016/j.resplu.2025.100986] [Medline: 40520569]
22. Elbadawi A, Tan BEX, Assaf Y, et al. Digital health intervention in patients with recent hospitalization for acute heart failure: a systematic review and meta-analysis of randomized trials. Int J Cardiol 2022 Jul 15;359:46-53. [doi: 10.1016/j.ijcard.2022.04.039] [Medline: 35436558]
23. Shaffer KM, Turner KL, Siwik C, et al. Digital health and telehealth in cancer care: a scoping review of reviews. Lancet Digit Health 2023 May;5(5):e316-e327. [doi: 10.1016/S2589-7500(23)00049-3] [Medline: 37100545]
24. Suresh S, Bhardwaj S, Rodrigues HC, et al. Smartphone-based digital health interventions: a comprehensive systematic review of efficacy for cardiovascular and cerebrovascular outcomes. J Med Syst 2025 Aug 26;49(1):109. [doi: 10.1007/s10916-025-02236-6] [Medline: 40856892]
25. Al Mahmud A, Joachim S, Jayaraman PP, et al. Digital health interventions to support chronic disease management: systematic scoping review. JMIR Mhealth Uhealth 2026 Jan 14;14:e63742. [doi: 10.2196/63742] [Medline: 41533959]
26. Marcolino MS, Oliveira JAQ, D'Agostino M, Ribeiro AL, Alkmim MBM, Novillo-Ortiz D. The impact of mHealth interventions: systematic review of systematic reviews. JMIR Mhealth Uhealth 2018 Jan 17;6(1):e23. [doi: 10.2196/mhealth.8873] [Medline: 29343463]
27. Johnson AM, Brimhall AS, Johnson ET, et al. A systematic review of the effectiveness of patient education through patient portals. JAMIA Open 2023 Apr;6(1):ooac085. [doi: 10.1093/jamiaopen/ooac085] [Medline: 36686972]
28. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
29. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
30. Krueger RA, Casey MA. Focus Groups: A Practical Guide for Applied Research, 5th edition: SAGE Publications, Inc; 2015:280.
31. Kvale S, Brinkmann S. InterViews: Learning the Craft of Qualitative Research Interviewing, 3rd edition: SAGE Publications, Inc; 2015.
32. Graneheim UH, Lundman B. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse Educ Today 2004 Feb;24(2):105-112. [doi: 10.1016/j.nedt.2003.10.001] [Medline: 14769454]
33. Krippendorff K. Content Analysis: An Introduction to Its Methodology, 4th edition: SAGE Publications, Inc; 2019. [doi: 10.4135/9781071878781]

XSL•FO
RenderX

34. Baker DW, Wolf MS, Feinglass J, Thompson JA, Gazmararian JA, Huang J. Health literacy and mortality among elderly persons. Arch Intern Med 2007 Jul 23;167(14):1503-1509. [doi: 10.1001/archinte.167.14.1503] [Medline: 17646604]
35. McCaffery KJ, Morony S, Muscat DM, et al. Evaluation of an Australian health literacy program delivered in adult education settings. Health Lit Res Pract 2019 Jul;3(3 Suppl):S42-S57. [doi: 10.3928/24748307-20190402-01] [Medline: 31687657]
36. Treviranus J. The Three Dimensions of Inclusive Design: A Design Framework for a Digitally Transformed and Complexly Interconnected Society: University College Dublin; 2018. URL: https://openresearch.ocadu.ca/id/eprint/2745/1/TreviranusThesisVolume1%262_v5_July%204_2018.pdf [accessed 2026-04-17]
37. World Health Organization. Global Strategy on Digital Health 2020-2025: World Health Organization. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2026-04-17]
38. Brewer LC, Fortuna KL, Jones C, et al. Back to the future: achieving health equity through health informatics and digital health. JMIR Mhealth Uhealth 2020 Jan 14;8(1):e14512. [doi: 10.2196/14512] [Medline: 31934874]
39. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. J Am Med Inform Assoc 2018 Aug 1;25(8):1080-1088. [doi: 10.1093/jamia/ocy052] [Medline: 29788380]
40. Parker AM, Sricharoenchai T, Raparla S, Schneck KW, Bienvenu OJ, Needham DM. Posttraumatic stress disorder in critical illness survivors: a metaanalysis. Crit Care Med 2015 May;43(5):1121-1129. [doi: 10.1097/CCM.0000000000000882] [Medline: 25654178]
41. Steinbusch CVM, van Heugten CM, Rasquin SMC, Verbunt JA, Moulaert VRM. Cognitive impairments and subjective cognitive complaints after survival of cardiac arrest: a prospective longitudinal cohort study. Resuscitation 2017 Nov;120:132-137. [doi: 10.1016/j.resuscitation.2017.08.007] [Medline: 28818523]
42. Douma MJ, Graham TAD, Ali S, et al. What are the care needs of families experiencing cardiac arrest?: A survivor and family led scoping review. Resuscitation 2021 Nov;168:119-141. [doi: 10.1016/j.resuscitation.2021.09.019] [Medline: 34592400]
43. Southern C, Tutton E, Dainty KN, et al. The experiences of cardiac arrest survivors and their key supporters following cardiac arrest: a systematic review and meta-ethnography. Resuscitation 2024 May;198:110188. [doi: 10.1016/j.resuscitation.2024.110188] [Medline: 38548009]
44. Cronberg T, Lilja G, Horn J, et al. Neurologic function and health-related quality of life in patients following targeted temperature management at 33°C vs 36°C after out-of-hospital cardiac arrest: a randomized clinical trial. JAMA Neurol 2015 Jun;72(6):634-641. [doi: 10.1001/jamaneurol.2015.0169] [Medline: 25844993]
45. Blennow Nordström E, Vestberg S, Evald L, et al. Neuropsychological outcome after cardiac arrest: results from a sub-study of the targeted hypothermia versus targeted normothermia after out-of-hospital cardiac arrest (TTM2) trial. Crit Care 2023 Aug 26;27(1):328. [doi: 10.1186/s13054-023-04617-0] [Medline: 37633944]
46. Whitehead L, Tierney S, Biggerstaff D, Perkins GD, Haywood KL. Trapped in a disrupted normality: survivors' and partners' experiences of life after a sudden cardiac arrest. Resuscitation 2020 Feb 1;147:81-87. [doi: 10.1016/j.resuscitation.2019.12.017] [Medline: 31887365]
47. Lilja G, Nielsen N, Bro-Jeppesen J, et al. Return to work and participation in society after out-of-hospital cardiac arrest. Circ Cardiovasc Qual Outcomes 2018 Jan;11(1):e003566. [doi: 10.1161/CIRCOUTCOMES.117.003566] [Medline: 29326145]
48. Joshi VL, Tang LH, Mikkelsen TB, et al. Does time heal fatigue, psychological, cognitive and disability problems in people who experience an out-of-hospital cardiac arrest? Results from the DANCAS survey study. Resuscitation 2023 Jan;182:109639. [doi: 10.1016/j.resuscitation.2022.11.005] [Medline: 36455704]
49. van den Broek B, Rijnen S, Stiekema A, van Heugten C, Bus B. Factors related to the quality and stability of partner relationships after traumatic brain injury: a systematic literature review. Arch Phys Med Rehabil 2022 Nov;103(11):2219-2231. [doi: 10.1016/j.apmr.2022.02.021] [Medline: 35395254]
50. Knobloch LK, Pool K. Interpersonal dynamics of traumatic brain injury within romantic relationships described in online forums. J Soc Pers Relat 2025 Mar;42(3):901-921. [doi: 10.1177/02654075241257457]
51. Haywood K, Dainty KN. Life after cardiac arrest: the importance of engaging with the "forgotten patient". Resuscitation 2018 Jul;128:A1-A2. [doi: 10.1016/j.resuscitation.2018.04.034] [Medline: 29727707]
52. Israelsson J, Persson C, Bremer A, Strömberg A, Årestedt K. Dyadic effects of type D personality and perceived control on health-related quality of life in cardiac arrest survivors and their spouses using the actor-partner interdependence model. Eur J Cardiovasc Nurs 2020 Apr;19(4):351-358. [doi: 10.1177/1474515119890466] [Medline: 31752502]
53. Sawyer KN, Camp-Rogers TR, Kotini-Shah P, et al. Sudden cardiac arrest survivorship: a scientific statement from the American Heart Association. Circulation 2020 Mar 24;141(12):e654-e685. [doi: 10.1161/CIR.0000000000000747] [Medline: 32078390]
54. Mion M, Lilja G, Bohm M, et al. V-CARE (Virtual CARE After REsuscitation): protocol for a randomized feasibility study of a virtual psychoeducational intervention after cardiac arrest—a STEPCARE sub-study. J Clin Med 2025 Jun 22;14(13):4429. [doi: 10.3390/jcm14134429] [Medline: 40648804]
55. Fusch P, Fusch GE, Hall JA, Walker NA, Booker JM. How to conduct a focus group interview: tips, strategies, and examples for novice researchers. Educ Inf 2022;38(2):171-187. [doi: 10.3233/EFI-211520]

XSL•FO
RenderX

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
CPR: cardiopulmonary resuscitation
ICD: implantable cardioverter-defibrillator
ICU: intensive care unit

Edited by A Stone; submitted 19.Sep.2025; peer-reviewed by A Presciutti, K Dainty; revised version received 04.Mar.2026; accepted 30.Mar.2026; published 22.May.2026.

Please cite as:
Waldemar A, Bremer A, Israelsson J, Heimburg K, Nordberg P, Blennow Nordström E, Årestedt K, Thylén I
Cardiac Arrest Survivors' Perspectives to Inform the Co-Design of a Web-Based Support and Learning Platform: Qualitative Content Analysis
J Med Internet Res 2026;28:e84432
URL: https://www.jmir.org/2026/1/e84432
doi:10.2196/84432

© Annette Waldemar, Anders Bremer, Johan Israelsson, Katarina Heimburg, Per Nordberg, Erik Blennow Nordström, Kristofer Årestedt, Ingela Thylén. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effectiveness of a Blended Intervention to Promote Physical Activity Among Office Employees: Randomized Controlled Trial

Yan Sun[1], PhD; Yang Gao[1], PhD; Alison Y T Ou[1], PhD; Xueying Wang[1], MSc; Yaojie Xie[2], PhD; Xiang-yan Chen[3], PhD; Shirley S M Fong[4], PhD; Wendy Yajun Huang[1], PhD; Chun-Qing Zhang[5], PhD; Julien S Baker[1], PhD; Martin S Hagger[6,7,8,9], PhD

[1]Department of Sports and Health Sciences, Academy of Wellness and Human Development, Hong Kong Baptist University, Academic and Administration Building, Baptist University Road Campus, Hong Kong, China (Hong Kong)

[2]School of Nursing, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)

[3]Division of Science, Engineering, and Health Studies, College of Professional and Continuing Education, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)

[4]Department of Health and Physical Education, The Education University of Hong Kong, Hong Kong, China (Hong Kong)

[5]Department of Psychology, Sun Yat-sen University, Guang Zhou, China

[6]Department of Psychological Science, University of California, Merced, CA, United States

[7]Health Sciences Research Institute, University of California, Merced, CA, United States

[8]School of Applied Psychology, Griffith University, Brisbane, Australia

[9]Faculty of Sport and Health Sciences, University of Jyväskylä, Jyväskylä, Finland

Corresponding Author:
Yang Gao, PhD
Department of Sports and Health Sciences, Academy of Wellness and Human Development, Hong Kong Baptist University, Academic and Administration Building, Baptist University Road Campus, Hong Kong, China (Hong Kong)

Abstract

Background: Regular moderate-to-vigorous physical activity (MVPA) reduces the risk of noncommunicable diseases, yet one-third of adults globally fail to meet MVPA recommendations. Office employees, among the least active groups, face heightened risks due to prolonged sedentary behavior and barriers like lack of time, fatigue, and low motivation. Although scalable, web-based interventions frequently face challenges, such as low engagement, high attrition, and limited personalization. Blended interventions, combining digital tools with interactive components, show promise but lack robust evidence among office employees.

Objective: This study aimed to evaluate the effectiveness of a theory- and evidence-based blended intervention in increasing MVPA among office employees.

Methods: This 24-week, 3-arm randomized controlled trial included 141 office employees, randomized equally into blended (web-based+electronic workshops [e-workshops]), web-based only, and control groups. The intervention was developed using the intervention mapping framework, ensuring a systematic, evidence- and theory-based design tailored to the needs of the target population. All participants accessed a study website; the control group used the "Library" module with general MVPA information, while intervention groups also accessed a tailored, theory-driven "Intervention" module. The blended group additionally attended 3 live e-workshops (weeks 2, 4, and 8). MVPA was objectively measured using hip-worn accelerometers at baseline (T1), postintervention (12 wk, T2), and follow-up (24 wk, T3). Retention, engagement, and acceptability were assessed. Analyses followed the intention-to-treat principle, and generalized linear mixed models were used to evaluate between-group differences at T2 and T3, adjusted for baseline MVPA, sex, and ActiGraph LEAP (Ametris) wear time.

Results: A total of 141 participants (mean age 32.1, SD 9.2 y; 75/141, 53.2% female) were randomized evenly across groups with comparable baseline characteristics, except for sex (P=.002). The blended group showed significantly greater MVPA increases than the control at T2 (β coefficient=0.247, 95% CI 0.013 - 0.480; P=.04) and T3 (β coefficient=0.373, 95% CI 0.139 - 0.607; P=.002), and greater increases than the web-based at T2 (β coefficient=0.287, 95% CI 0.062 - 0.512; P=.01) and T3 (β coefficient=0.368, 95% CI 0.138 - 0.597; P=.002). Sensitivity analyses confirmed the robustness of these findings. Retention and engagement rates were 83% and 56%, respectively. Participants reported high acceptability across all domains.

Conclusions: A blended intervention integrating tailored web content with interactive e-workshops significantly increased MVPA at T2 and T3 compared with both web-based and control groups. This success stems from the rigorous intervention mapping framework, which systematically identified effective behavior change and habit formation theories, while the blended approach addressed key limitations of traditional and web-based methods. These findings underscore the potential of evidence-

XSL•FO
RenderX

and theory-based blended approaches to address common barriers to MVPA and promote sustainable MVPA in workplace settings, offering promising opportunities for broader scalability and personalization.

Trial Registration: ClinicalTrials.gov NCT04391270; https://clinicaltrials.gov/study/NCT04391270

(J Med Internet Res 2026;28:e80249) doi:10.2196/80249

KEYWORDS

physical inactivity; moderate-to-vigorous physical activity; RCT; randomized controlled trial; health promotion

Introduction

Noncommunicable diseases (NCDs) pose a significant global health crisis, accounting for 41 million deaths annually, or 74% of all deaths worldwide [1]. Cardiovascular diseases, cancers, chronic respiratory diseases, and diabetes are the leading contributors, driven by behavioral risk factors such as smoking, physical inactivity, unhealthy diets, and excessive alcohol consumption [2]. Among these, physical inactivity is a major contributor to NCDs and premature mortality. Regular engagement in moderate-to-vigorous physical activity (MVPA) significantly reduces the risk of cardiovascular diseases, cancers, diabetes, obesity, and mental health disorders [3]. However, nearly one-third of adults globally (1.8 billion) fail to meet the World Health Organization's (WHO) MVPA recommendations, making insufficient MVPA the fourth leading risk factor for global mortality [2].

This issue is particularly pronounced among office employees, who spend over two-thirds of their workday in sedentary activities and often remain inactive during leisure time [4-6]. Consequently, office employees are among the least active occupational groups, facing heightened risks of NCDs and premature mortality [4]. In Hong Kong, only 46.2% (2944/6368) of adults (age: 17-79) meet global MVPA recommendations—significantly below the global average of 75% [7]. Common barriers, including lack of time, fatigue, and low motivation, further prevent office employees from incorporating MVPA into their routines [8].

The COVID-19 pandemic exacerbated physical inactivity globally. Movement restrictions, lockdowns, and the transition to remote work dramatically reduced opportunities for MVPA [9,10]. In Hong Kong, 70.4% (444/631) of adults did not meet MVPA guidelines during the pandemic [11]. Office employees, in particular, saw activity levels decline further, as remote work reduced MVPA levels in domains such as leisure time and transportation [12-14].

Web-based interventions have gained attention for their scalability and cost-effectiveness [15], with evidence showing positive effects on MVPA levels among working adults. However, their long-term effectiveness remains inconclusive due to small effect sizes, challenges such as low engagement, high attrition, and a lack of interactive or personalized components [16]. These challenges suggest that web-based interventions alone are insufficient for long-term behavior change, particularly in real-world settings. Interactive elements, such as group sessions and one-on-one interviews, have been shown to enhance engagement and effectiveness, but their integration into workplace interventions remains limited [17].

Blended interventions, which integrate digital tools (eg, web-based self-monitoring and feedback systems) with interactive elements, present a promising alternative by providing personalized support, real-time feedback, and social interaction [17]. Nevertheless, systematic reviews, such as Hohberg et al [17], highlight significant methodological heterogeneity in blended interventions, which can be attributed to variations in duration, intensity, theoretical foundations, and component integration. Such inconsistencies limit definitive conclusions about their efficacy, especially for specific populations, such as office employees. Therefore, there is an urgent call for systematically designed interventions that address these limitations by using robust, theory-based, and evidence-based frameworks.

To address these gaps, this study evaluates the effectiveness of a novel blended intervention designed to increase MVPA, improve health outcomes, and enhance work productivity among office employees. The intervention was systematically developed using the intervention mapping (IM) framework, which integrates evidence-based practices and behavior change theories to guide the design of interventions that are specifically tailored to the needs and barriers of the target population [18]. By combining web-based self-monitoring tools with interactive electronic workshops (e-workshops), the intervention aims to overcome key limitations of both web-based and traditional approaches. Importantly, the use of the IM framework ensures a structured and comprehensive approach to intervention design, addressing the methodological heterogeneity observed in previous studies. The specific objectives of this study are (1) to assess changes in actigraphy-measured MVPA levels among participants, and (2) to examine differences in engagement and retention rates across intervention groups.

Methods

Study Design

This study adhered to the CONSORT (Consolidated Standards of Reporting Trials) 2010 guidelines for randomized controlled trials (RCT) [19,20]. The original study was designed as a cluster-RCT [21]. However, disruptions caused by the COVID-19 pandemic and recruitment challenges necessitated its adaptation to a standard RCT with convenience sampling. Key changes included adjusting the estimated sample size from 495 (cluster-RCT design) to 135 (RCT design), replacing face-to-face workshops with e-workshops conducted via Zoom (Zoom Communications), and omitting workplace-level environmental changes.

The intervention was developed using the IM framework, a systematic and iterative approach for designing evidence-based

XSL•FO
RenderX

health promotion programs. The IM framework comprises 6 steps, with the first 4 steps focusing on the development of the intervention. These steps provided a structured, theory-based, and evidence-based process to create a tailored intervention addressing the needs of the target population. Figure 1 summarizes the IM framework and the tasks involved in each step.

Step 1 identified the health problem, its determinants, and the needs of the target population. Step 2 defined the intervention's expected outcomes, specified performance objectives, and developed a matrix of change objectives. In step 3, evidence-based behavior change methods were selected and translated into practical applications to address these objectives. Finally, step 4 focused on preparing and pilot-testing intervention materials to ensure feasibility, acceptability, and usability. This process enabled the development of a comprehensive and adaptable intervention, even under the constraints posed by the COVID-19 pandemic.

This study was a 24-week, 3-arm RCT. Participants were randomly allocated to one of the 3 groups—a web-based intervention group, a blended intervention group (combining web-based with e-workshops), or a control group. Assessments were conducted at baseline (T1), postintervention (T2, 12 wk), and follow-up (T3, 24 wk). All participants accessed the same website, which consisted of 2 main modules, a "Library" module with generic MVPA information and an "Intervention" module with personalized, theory-driven content (accessible only to the web-based and blended intervention groups). Additionally, participants in the blended intervention group attended 3 e-workshops at weeks 2, 4, and 8. All participants received fortnightly reminders to engage with the website.

Figure 1. Overview of intervention mapping steps and corresponding tasks.

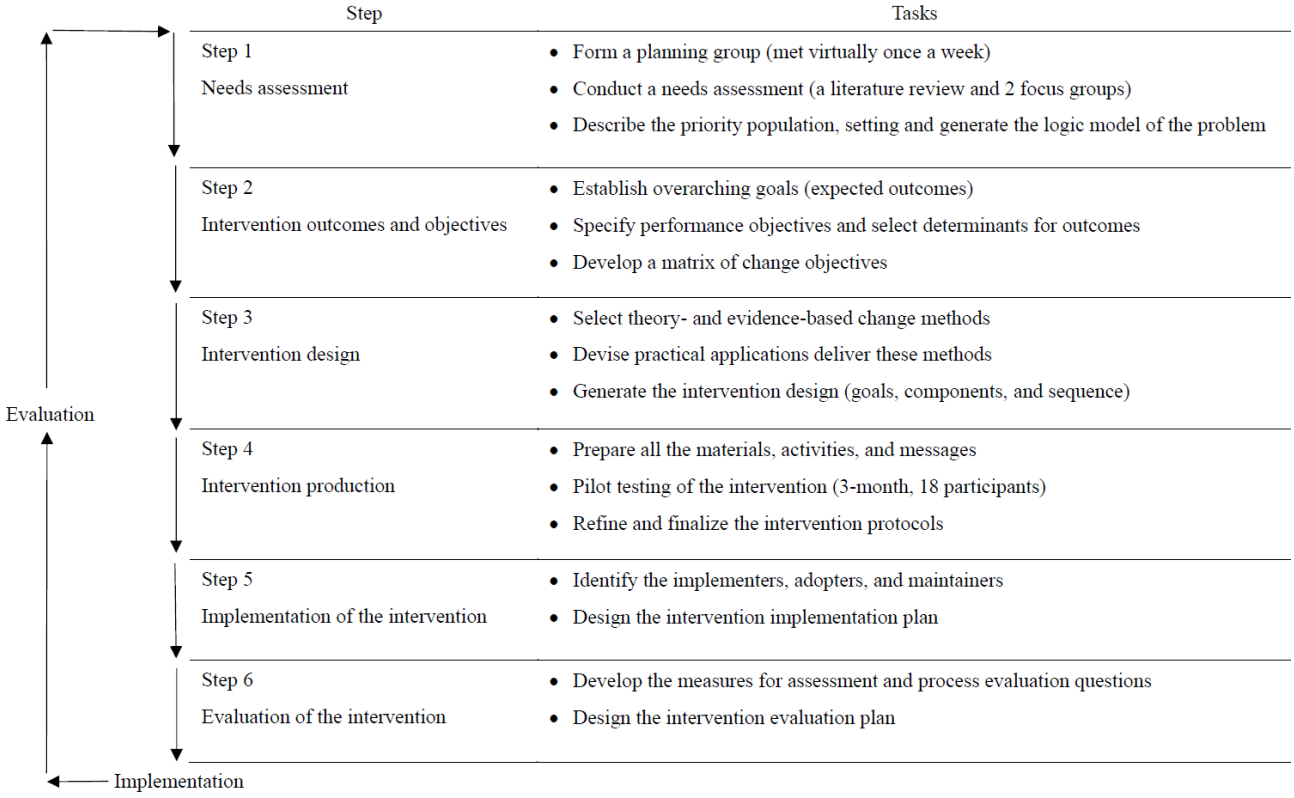

Ethical Considerations

This study was approved by the Research Ethics Committee of Hong Kong Baptist University (approval HASC/17-18/0529). All participants provided electronic informed consent via the QuestionPro platform after being fully informed of the study's objectives, procedures, potential risks, and benefits. The informed consent process also granted permission for secondary analysis of the collected data without requiring additional consent.

Identifiable personal data will be destroyed after the study is completed, while deidentified datasets will be retained for research purposes and can be requested from the corresponding author (YG) upon reasonable request. Participants were compensated with HK $50 (approximately US $6.40) coupons for each completed evaluation session. Additionally, those who completed all study requirements were entered into a lottery to win a HK $1000 (approximately US $128; quota: 5) coupon. As part of the study benefits, participants also received a personalized physical activity and health tracking report summarizing their progress and providing actionable insights. No identifiable images or information about participants are included in the manuscript or supplementary materials.

Eligibility Criteria and Recruitment Process

Eligible participants were office employees in Hong Kong aged 18 years or older, engaging in less than 150 minutes of MVPA per week according to WHO guidelines, assessed using the short-form International Physical Activity Questionnaire [22]. To ensure safety, participants completed the Physical Activity Readiness Questionnaire, and those who answered "yes" to any

XSL•FO
RenderX

question were required to obtain medical clearance before participation. Individuals enrolled in other MVPA promotion programs or with conditions preventing physical activity were excluded.

Recruitment was conducted remotely during the pandemic of COVID-19 (September-December 2022) when many employees were working from home. A prerandomized list of 440 companies, representing 20% of eligible Hong Kong companies with more than 100 employees, was obtained from the Hong Kong Census and Statistics Department. Due to pandemic-related concerns, most companies declined participation or did not respond, with only 4 agreeing to disseminate study materials to their office employees. Social media platforms, such as Facebook (Meta), Instagram (Meta), and WeChat (Tencent Holdings Limited), were also used, resulting in the recruitment of 88 participants. Referrals were encouraged through unique tracking links shared by interested individuals with colleagues, friends, or family members who met the eligibility criteria, which led to the recruitment of an additional 27 participants. In total, 172 individuals were assessed for eligibility, and 141 participants (82%) completed baseline enrollment.

Randomization and Allocation

Participants were randomly assigned to the 3 study groups using block randomization with a fixed block size of 6 to ensure balanced allocation throughout enrollment. The randomization sequence was generated using a computer-based random number generator and administered by an independent research team member who was not involved in the intervention or outcome assessment. Allocation concealment was ensured, and outcome assessors remained blinded to group assignments.

Sample Size Estimation

The sample size was calculated using repeated measures ANOVA within-between interaction approach (G*Power, version 3.1.9.6), where a power set at 0.80 and an α at .05, with 3 groups and 3 measurement waves. An effect size of 0.30 (Cohen *d*) was chosen based on a similar study in terms of design and intervention components [23], and on a median effect size of 0.30 (IQR 0.16-0.38), calculated from 19 studies reporting Cohen *d* in a recent systematic review of blended interventions [17]. To account for an expected attrition rate of 30%, the required sample size was estimated at 135 participants, with 45 participants allocated to each group [24].

Intervention

Control Group

Participants in the control group accessed only the “Library” module of the website, which provided general, nontailored information on MVPA, health, and work productivity. This content consisted of 18 short essays, delivered every 2 weeks across 6 sessions, covering topics such as recommended MVPA levels, the benefits of MVPA, and the risks of inactivity.

In total, 3 behavior change techniques (BCTs) were embedded in the materials, focusing on the consequences of health, social, and emotional outcomes (codes 5.1, 5.3, and 5.6, respectively) [25]. For instance, 1 essay (#3 in session 1) addressing the consequences of health outcomes (code 5.1) explained the risks of inactivity (eg, heart disease, diabetes, and cancer), and outlined benefits such as improved cardiovascular health and better sleep, supported by evidence-based data. Another essay (#4 in session 2) focused on the social benefits of physical activity (code 5.3), such as meeting new people, fostering teamwork, and building support systems through group activities like sports or fitness classes. It described how exercising with others can create a sense of belonging and improve teamwork skills. Additionally, an essay (#5 in session 2) explored emotional benefits (code 5.6), explaining how physical activity can boost mood and reduce stress, depression, and anxiety by releasing endorphins and serotonin, with examples like running, yoga, or brisk walking.

After the study concluded, control group participants were granted access to the “Intervention” section for exposure to the web-based intervention content.

Web-Based Intervention Group

In addition to accessing the “Library,” participants in the web-based intervention group engaged with the “Intervention” module of the website, following the same schedule as the control group. Within this module, participants answered questions and received individually tailored feedback on their MVPA levels, behavioral determinants, and social and environmental factors. The feedback was based on a theoretical framework that integrates motivational, volitional, and habit-formation processes, drawing from dual-process theories and habit formation theories to support both behavior change and maintenance [26-28].

A total of 31 BCTs were selected and integrated based on these 3 processes [25,29]. For example, in session 3**,** participants were guided to stay motivated by identifying their main barriers to MVPA and developing tailored coping strategies. Suggestions included trying lighter activities, such as yoga or walking, to address fatigue, or scheduling short, manageable bouts of activity (eg, 10-min brisk walks) for busy days. Participants were also reminded of the mental and emotional benefits of exercise, such as improved mood and reduced stress, to reinforce motivation. Further session-specific details, including questions, feedback, and BCTs, are provided in Table S1 in Multimedia Appendix 1.

Blended Intervention Group

In addition to accessing 2 modules on the website, participants in the blended intervention group attended 3 live e-workshops via Zoom, scheduled for weeks 2, 4, and 8. Each e-workshop lasted 30 to 45 minutes and involved a facilitator, a moderator, and 4 to 10 participants. These workshops were designed to provide direct and immediate individualized feedback, facilitate the sharing of experiences among participants, and encourage adherence to the intervention [15].

To enhance the effectiveness of the intervention, 4 additional BCTs were integrated into the workshops: problem solving (code 1.2), instruction on how to perform the behavior (code 4.1), reattribution (code 4.3), and mental rehearsal of successful performance (code 15.2) [25]. For example, during the third workshop, participants applied problem solving (code 1.2) by

XSL•FO
RenderX

identifying barriers to MVPA (eg, bad weather or lack of time) and collaboratively developing coping strategies using the "If…then…" framework (eg, "If the weather is bad, then I will do a 20-min yoga session at home"). The facilitator provided real-time feedback to ensure practicality, while the group setting encouraged idea sharing and accountability, enhancing engagement compared with online modules.

Outcome Measures

MVPA was objectively measured by a hip-worn accelerometer (ActiGraph, Models wGT3X-BT) at T1-T3, and data were analyzed by using ActiLife software v6.13.4 (ActiGraph LLC). The participants were instructed to wear an accelerometer on the right hip during waking hours for 7 consecutive days, except when bathing, performing water-based activities, and sleeping [30]. Freedson cutoff points were used to classify VPA, MPA, and light-intensity physical activity [31]. Data of at least 10 hours per day for a minimum of 3 days (including at least 1 weekend day) were regarded as valid and included in data analysis [32,33]. Then, the total weekly MVPA was calculated by multiplying the average daily MVPA levels by 7.

The retention rate (%) was calculated as participants at the end of the intervention divided by the participants at baseline.

The engagement rate (%) was calculated by the number of intervention sessions attended by participants divided by the total number of sessions provided to them. In the blended group, engagement rates were calculated for web-based sessions and e-workshop sessions separately and in combination.

The acceptability of the intervention was evaluated using a 5-point Likert scale (from "1=strongly dissatisfied" to "5=strongly satisfied") across several dimensions, such as ease of use, understandability, usefulness, enjoyability, time acceptability, and overall satisfaction [34,35]. Statements included, for example, "The content of the intervention is understandable" and "The time taken to complete the content of the intervention was acceptable." For the evaluation of the e-workshops, similar statements were adapted. Responses were grouped, combining "agree" and "strongly agree" into one percentage, and "disagree" and "strongly disagree" into another, to reflect participant agreement or disagreement.

Data Analysis

MVPA (min/wk) was summarized as mean and SD at each assessment time point based on valid ActiGraph data. Primary intervention effects at T2 and T3 were evaluated under an intention-to-treat (ITT) framework using generalized linear mixed models with a γ distribution and a log link. Participants were analyzed as randomized. The generalized linear mixed models included fixed effects for group, time, and the group×time interaction, with participant treated as a random effect. Models were adjusted for sex, ActiGraph wear time, and baseline MVPA. Between-group intervention effects were estimated as differences in change from baseline using the group×time interaction (difference-in-differences); pairwise contrasts were obtained by changing reference groups where needed. For completeness, estimated within-group changes from baseline are also presented.

At baseline, MVPA values were unavailable for 15 participants because ActiGraph recordings did not meet the prespecified wear-time validity criteria. These baseline values were replaced using within-group mean imputation for the primary analyses [36]. Sensitivity analyses used multiple imputation (MI) for MVPA across all 3 time points, generating 5 imputations [37,38]. The MI model incorporated baseline values and other variables associated with MVPA to improve estimation accuracy.

Retention and engagement rates were compared across groups using chi-square tests. Statistical significance was set at $P<.05$, with 95% CIs reported. All analyses were performed in IBM SPSS Statistics (version 29.0.1).

Results

A total of 172 office employees were assessed for eligibility, and 141 of them completed the baseline assessment (mean age 32.05, SD 9.22 y; 75/141, 53.2% female), with each group consisting of 47 participants. The participant flow is illustrated in Figure 2. The baseline sociodemographic characteristics are presented in Table 1. There were no demographic differences between participants who completed 3 assessments and those who were lost to follow-up.

XSL•FO
RenderX

Figure 2. Flow diagram of the study. MVPA: moderate-to-vigorous physical activity; WHO: World Health Organization.

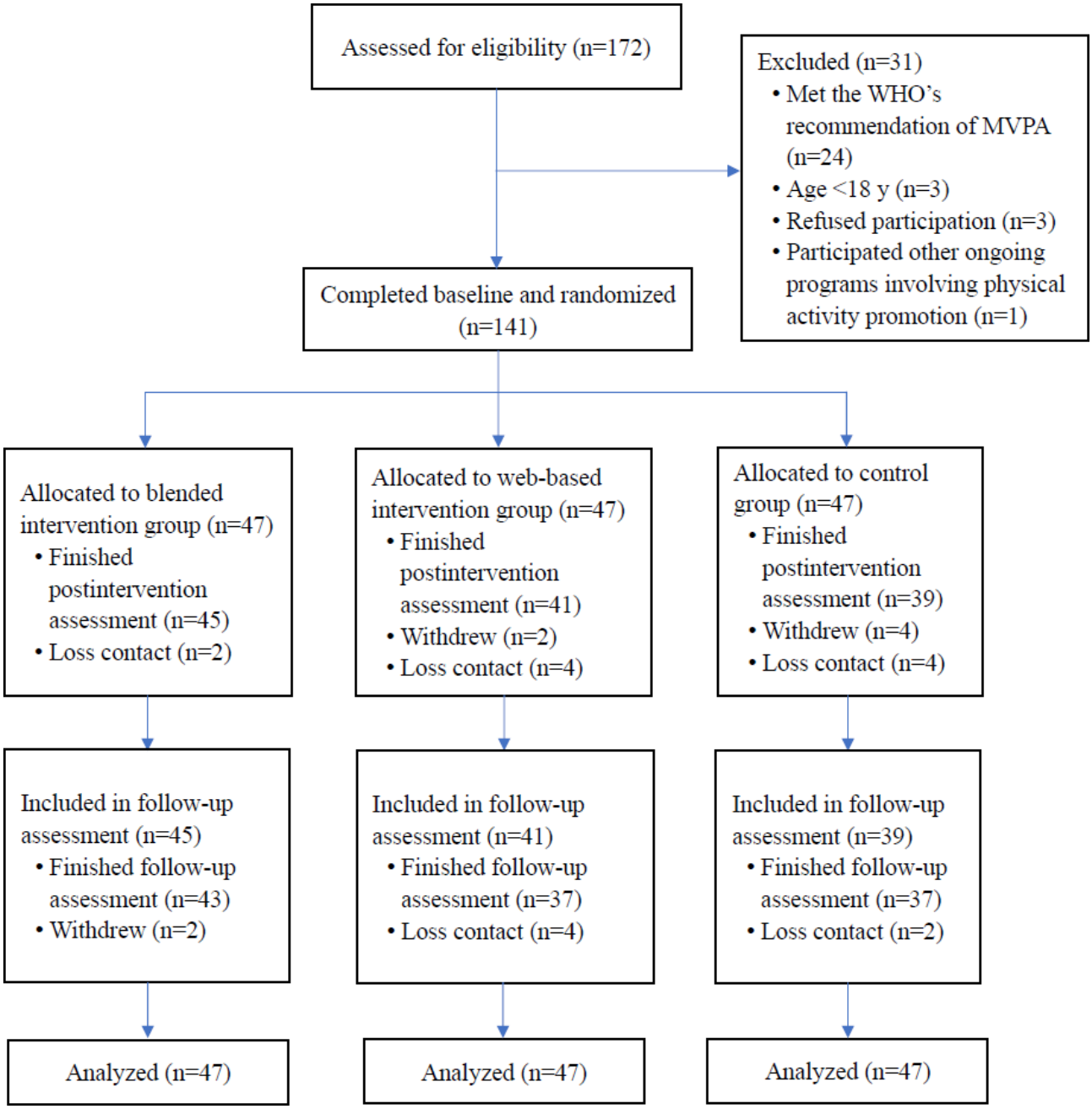

XSL•FO
RenderX

Table . Sociodemographic characteristics of the participants at baseline.

Characteristic	Total (n=141)	Blended (n=47)	Web-based (n=47)	Control (n=47)
Age (y), mean (SD)	32.05 (9.22)	33.15 (7.77)	32.14 (10.27)	30.57 (9.38)
Sex, n (%)				
Male	66 (46.8)	19 (40.4)	27 (57.4)	20 (42.6)
Female	75 (53.2)	28 (59.6)	20 (42.6)	27 (57.4)
Education level, n (%)				
Diploma or below	28 (19.9)	10 (21.3)	9 (19.1)	9 (19.1)
Bachelor or above	113 (80.1)	37 (78.7)	38 (80.9)	38 (80.9)
Job category, n (%)				
Administrators or associate professionals	84 (59.6)	28 (59.6)	33 (70.2)	23 (48.9)
Clerical or services related	57 (40.4)	19 (40.4)	14 (29.8)	24 (51.1)
Marital status, n (%)				
Single	88 (62.4)	27 (57.4)	29 (61.7)	32 (68.1)
Not single	53 (37.6)	20 (42.6)	18 (38.3)	15 (31.9)
Monthly salary (HKD), n (%)				
<20,000 (<US $2560)	52 (36.9)	13 (27.7)	16 (34)	23 (48.9)
≥20,000 (≥US $2560)	89 (63.1)	34 (72.3)	31 (66)	24 (51.1)

Table 2 presents means and SDs for actigraphy-measured MVPA (min/wk) and ActiGraph wear time (min/d) by group at three time points. Descriptive statistics are based on valid recordings only (complete cases for each time point). Sample sizes differ across time points because some recordings did not meet the prespecified wear-time validity criteria and some assessments were missing (including dropout).

Table . Mean (SD) for actigraphy-measured MVPA[a] (min/wk) and ActiGraph wear time (min/d) by group at 3 time points. Values are presented as mean (SD) based on valid ActiGraph data. Sample sizes (n) vary across time points due to invalid recordings and missing assessments (including dropout).

Time point	n	MVPA (min/wk), mean (SD)	Wear time (min/d), mean (SD)
Baseline			
Blended	43	163.17 (80.39)	729.95 (77.04)
Web-based	45	163.47 (76.72)	740.92 (104.53)
Control	38	159.19 (79.79)	699.03 (70.55)
Postintervention			
Blended	39	244.80 (115.37)	725.50 (71.44)
Web-based	37	185.78 (60.95)	727.84 (77.53)
Control	34	183.71 (62.22)	717.78 (104.35)
Follow-up			
Blended	40	268.00 (103.20)	754.67 (77.14)
Web-based	34	197.70 (73.24)	693.63 (60.56)
Control	33	194.56 (83.86)	697.33 (66.96)

[a]MVPA: moderate-to-vigorous physical activity.

Table 3 presents within-group changes and between-group differences in changes in actigraphy-measured MVPA at T2 and T3 relative to T1, using both ITT and sensitivity analyses. In the ITT analysis, MVPA increased significantly from T1 to T2 and T3 within all 3 groups (all $P<.01$). For between-group differences in change, the blended group showed significantly greater improvements than the control group at T2 (β coefficient=0.247, 95% CI 0.013 - 0.480; $P=.04$) and than the

XSL•FO
RenderX

web-based group at T2 (β coefficient=0.287, 95% CI 0.062 - 0.512; *P*=.01) at T2. These effects were maintained at T3, with the blended group continuing to demonstrate greater improvements compared with both the control group (β coefficient=0.373, 95% CI 0.139 - 0.607; *P*=.002) and the web-based group (β coefficient=0.368, 95% CI 0.138 - 0.597; *P*=.002). No significant differences in changes were observed between the web-based and control groups at either time point.

Sensitivity analyses using MI displayed a similar pattern, with the blended group demonstrating greater improvements than both the control and web-based groups at both T2 and T3 (all *P*<.001). These findings are consistent with the primary ITT analysis and support the robustness of the conclusions.

XSL•FO
RenderX

Table . GLMM[a] estimates for actigraphy-measured MVPA[b] (within-group changes from T1 to T2 and T3 and between-group differences in change [ITT[c] {primary} and sensitivity analyses[d]]).

Effect	Standardized β coefficient (95% CI)	SE	*t* test (*df*)	*P* value
ITT estimates				
Baseline to postintervention				
Within blended	0.500 (0.342 to 0.658)	0.080	6.228 (327)	<.001
Within web-based	0.213 (0.053 to 0.373)	0.081	2.617 (327)	.009
Within control	0.253 (0.064 to 0.442)	0.087	2.901 (327)	.006
Blended versus control	0.247 (0.013 to 0.480)	0.119	2.080 (327)	.04
Web-based versus control	−0.040 (−0.275 to 0.194)	0.119	−0.339 (327)	.74
Blended versus web-based	0.287 (0.062 to 0.512)	0.114	2.515 (327)	.01
Baseline to follow-up				
Within blended	0.633 (0.452 to 0.814)	0.080	7.868 (327)	<.001
Within web-based	0.265 (0.077 to 0.453)	0.084	3.162 (327)	.003
Within control	0.260 (0.063 to 0.456)	0.088	2.965 (327)	.006
Blended versus control	0.373 (0.139 to 0.607)	0.119	3.137 (327)	.002
Web-based versus control	0.005 (−0.233 to 0.244)	0.121	0.045 (327)	.96
Blended versus web-based	0.368 (0.138 to 0.597)	0.117	3.149 (327)	.002
Sensitivity estimates				
Baseline to postintervention				
Within blended	0.500 (0.436 to 0.563)	0.032	15.452 (2022)	<.001
Within web-based	0.213 (0.148 to 0.277)	0.033	6.494 (2022)	<.001
Within control	0.253 (0.184 to 0.322)	0.035	7.197 (2022)	<.001
Blended versus control	0.247 (0.153 to 0.340)	0.048	5.160 (2022)	<.001
Web-based versus control	−0.040 (−0.135 to 0.054)	0.048	−0.839 (2022)	.40
Blended versus web-based	0.287 (0.197 to 0.377)	0.046	6.237 (2022)	<.001
Baseline to follow-up				
Within blended	0.633 (0.560 to 0.705)	0.032	19.566 (2022)	<.001
Within web-based	0.265 (0.189 to 0.340)	0.034	7.858 (2022)	<.001
Within control	0.260 (0.181 to 0.338)	0.035	7.370 (2022)	<.001
Blended versus control	0.373 (0.279 to 0.467)	0.048	7.804 (2022)	<.001
Web-based versus control	0.005 (−0.090 to 0.101)	0.049	0.109 (2022)	.91
Blended versus web-based	0.368 (0.276 to 0.460)	0.047	7.834 (2022)	<.001

[a]GLMM: generalized linear mixed models with a Gamma distribution and log link. Models included fixed effects for group, time, and their interaction (group×time), adjusted for sex, wear time, and baseline moderate-to-vigorous physical activity. Between-group rows report coefficients for the group×time interaction only. Main effects (group and time), covariates (baseline MVPA, wear time, and sex), and other interaction terms were included in the models but are not displayed.

[b]MVPA: moderate-to-vigorous physical activity.

[c]ITT: intention-to-treat. Primary analyses followed the intention-to-treat principle using all available repeated measures.

[d]Sensitivity analyses used multiple imputation with pooled estimates across imputations.

XSL•FO
RenderX

Retention and engagement rates across groups are presented in Table 4. The overall retention and engagement rates were 83% and 56%, respectively, with no significant differences observed between groups. In the blended group, the overall engagement rate was 56%, consisting of 61% for web-based sessions and 46.1% for e-workshops.

Participants across all groups reported high levels of acceptability for the web-based sessions, with strong agreement observed for ease of use, understandability, usefulness, enjoyability, time acceptability, and satisfaction (Table S2 in Multimedia Appendix 2). Ease of use and understandability were rated highly, with over 90% agreement in all groups. Usefulness and satisfaction also showed high agreement, exceeding 88% across groups. Enjoyability had slightly lower agreement rates but remained above 77% for all groups. Time acceptability was rated positively by most participants, with agreement rates above 88% in all groups.

No significant harms or unintended effects were reported in any of the intervention groups. Participants did not report any adverse events related to the intervention, and all activities were conducted within safe and acceptable parameters.

Table . Retention and engagement rates of the intervention by group.

Outcomes	Total, n/N (%)	Blended, n/N (%)	Web-based, n/N (%)	Control, n/N (%)	Chi-square (*df*)	*P* value
Retention	117/141 (83)	43/47 (91.5)	37/47 (78.7)	37/47 (78.7)	3.6 (2)	.16
Engagement	553/987 (56)	237/423 (56)	156/282 (55.3)	160/282 (56.7)	0.1 (2)	.94

Discussion

Principal Findings

This study evaluated the effectiveness of a blended intervention, combining web-based content with interactive e-workshops, in promoting MVPA among physically inactive office employees. The findings demonstrated that the blended intervention significantly increased MVPA levels at both T2 and T3, outperforming the web-based intervention and the control group. Retention rates were high across all groups, with the blended group achieving the highest rate, further supporting the feasibility and acceptability of the intervention. These results underscore the potential of blended approaches in addressing physical inactivity among office employees, a population particularly vulnerable to sedentary behaviors.

The significant improvements in MVPA levels observed in the blended group at both T2 and T3 underscore the effectiveness of integrating web-based tools with interactive e-workshops. Compared with the web-based and control groups, the blended intervention achieved the most substantial increases in MVPA, highlighting its potential to surpass web-only or traditional approaches. In contrast, the web-based group showed modest improvements, while the control group exhibited minimal changes, with both groups showing significant changes at T3.

The success of the blended intervention can be attributed to its comprehensive, theory-based design, which incorporated a total of 34 BCTs targeting motivation, self-regulation, and habit formation. Previous meta-analyses have demonstrated that interventions using multiple BCTs grounded in behavioral theories tend to produce stronger effects on physical activity [39,40]. While the web-based intervention also used 31 BCTs, the blended intervention expanded upon these by incorporating 3 additional BCTs unique to the e-workshops. These additional BCTs further enhanced the intervention's overall effectiveness by promoting interpersonal interaction, fostering peer support, and providing real-time feedback and accountability.

The group-based nature, combined with the digital person-to-person component of the e-workshops, played a pivotal role in the success of the blended intervention. The group-based format fostered interpersonal interaction and collaboration, encouraging participants to share experiences, brainstorm solutions to common barriers, and build self-efficacy [41]. Peer support within the group likely strengthened participants' motivation and commitment, as they could learn from others' strategies and successes, while also feeling a sense of collective accountability. At the same time, the digital person-to-person component added a human support that is often missing in purely digital interventions [42]. Unlike the impersonal nature of human-computer interactions, face-to-face communication offered warmth, empathy, and dynamic support. Facilitators were able to provide real-time feedback and adaptive guidance, promptly addressing participants' individual challenges. Together, these features fostered a supportive and interactive environment that not only promoted encouragement, belonging, and accountability but also strengthened the application of behavior change theories targeting motivation, volition, and habit formation. By reinforcing these theoretical foundations, the blended approach amplified and consolidated the impact of the web-based content, ensuring a more robust and sustainable behavior change process.

In contrast, the web-based and control groups lacked these collaborative and interactive elements, which may explain the superior effectiveness of the blended intervention. While the web-based intervention provided valuable tools for self-regulation and goal-setting, it did not offer the same level of social reinforcement, dynamic problem-solving, or accountability as the blended approach. These findings align with previous research emphasizing the superiority of blended interventions, which effectively combine the flexibility of digital tools with the social and interactive benefits of face-to-face components [17].

The long-term improvements in MVPA levels observed at T3 suggest that the blended intervention may have facilitated long-term behavior change. This aligns with habit formation theory, which posits that consistent behaviors are more likely to be maintained when reinforced in stable contexts and supported by environmental cues [43]. Periodic e-workshops

XSL•FO
RenderX

likely served as motivational "boosters," helping participants overcome lapses and maintain engagement over time. This is particularly critical for workplace interventions, where competing demands and time constraints often challenge long-term adherence.

However, the web-based intervention alone demonstrated limited impact on MVPA levels compared with the control group, despite offering tailored feedback based on behavioral determinants. This is consistent with previous studies reporting mixed results for digital-only interventions, especially those lacking interactive or social components [36,44]. The absence of interpersonal interaction and real-time feedback in the web-based group may have reduced participants' sense of accountability and support—both key drivers of long-term behavior change [16]. Without the reinforcement provided by workshops, participants may have struggled to translate digital content into consistent physical activity behaviors. These findings highlight the critical role of interactive components, such as group-based e-workshops, in enhancing the effectiveness of digital interventions.

The findings highlight the feasibility and acceptability of the intervention, as evidenced by high retention rates and positive participant feedback. Notably, the retention rates in this study were relatively high compared with similar interventions, which typically report rates ranging from 59% to 82% [36,45-47]. This can be attributed to the systematic and evidence-based design of the intervention using IM, which ensured the intervention was theory-driven and tailored to meet the needs of the target population [48]. Furthermore, participant retention was supported by strategies such as providing MVPA and health progress reports, as well as incentives like supermarket coupons and a grand prize for completing all outcome assessments. These measures created a structured and supportive environment, encouraging long-term participation.

Directly comparing engagement across studies is challenging due to differences in how engagement is defined and measured. While our study calculated engagement as a rate, other studies have used metrics like mean session completions or participant distributions, complicating direct comparisons. Despite this, previous studies have shown that blended interventions often achieve better engagement than platform-only or control groups. For example, 1 study found that adding coaching significantly improved session completion rates [49], while another highlighted the value of interactive components in enhancing adherence [50]. In our study, the lack of significant differences in engagement may be explained by limited participation in e-workshops, likely due to scheduling challenges and competing responsibilities during personal time, which diminished the potential advantages of the blended format. Future research could explore strategies such as offering more flexible time slots to overcome these barriers and improve engagement in blended interventions.

Strengths and Limitations

The strengths of this study lie in its rigorous design, guided by the IM framework, which ensured a systematic and theory-driven approach to intervention development and evaluation. The use of objective measures, such as actigraphy, provided robust and reliable data on MVPA outcomes. Additionally, the high retention and engagement rates observed in the blended group demonstrate the intervention's feasibility and acceptability in a workplace setting.

However, several limitations should be acknowledged. First, the ActiGraph device used to measure MVPA did not capture water-based activities, potentially underestimating MVPA levels among participants who engaged in swimming or similar activities. Second, the sample was skewed toward younger, highly educated, and digitally literate individuals, limiting the generalizability of the findings to other populations. Third, baseline MVPA levels in this study were relatively high, possibly due to discrepancies between self-reported activity during recruitment and objective measurements at baseline. This might have resulted in a ceiling effect, underestimating the intervention's true potential. Finally, the study was conducted exclusively among office employees in Hong Kong, a dense urban setting with unique cultural and environmental factors, which may limit the applicability of the findings to other contexts.

Conclusions

This study provides strong evidence for the effectiveness of a blended intervention in promoting MVPA among physically inactive office employees. By integrating web-based platforms with interactive e-workshops, the intervention effectively addressed key barriers to MVPA and achieved significant improvements in engagement, retention, and behavioral outcomes. These findings highlight the potential of blended approaches to address physical inactivity in workplace settings and beyond.

Future research should build on these findings by exploring the long-term sustainability of behavioral changes achieved through blended interventions and their scalability across diverse populations and settings. Additionally, leveraging emerging technologies, such as artificial intelligence, to deliver personalized feedback and adaptive interventions could further enhance engagement and effectiveness. Addressing these questions will be critical for advancing workplace health promotion and tackling the global physical inactivity crisis.

Acknowledgments

We sincerely thank all the participants who took part in this study for their time and commitment, which made this research possible. Special thanks to Miss Xiong Zhennan for her invaluable efforts in data collection. We also extend our gratitude to the department for providing the necessary equipment and consistent support throughout the study.

XSL•FO
RenderX

Funding

This study was supported by the General Research Fund of the Research Grants Council of Hong Kong (grant 12609919). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author (YG) on reasonable request.

Authors' Contributions

YS was responsible for methodology, data curation, formal analysis, writing the original draft, and reviewing and editing this paper. YG was responsible for conceptualization, project administration, supervision, funding acquisition, and reviewing and editing this paper. AYTO contributed to methodology and data curation. XW, YX, XyC, SSMF, WYH, CQZ, JSB, and MSH were responsible for reviewing and editing. All authors have read and approved the published version of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Timeline, questions, personally-tailored feedback, and behavior change techniques involved in 6 web-based sessions in 2 intervention groups.
[DOCX File, 25 KB - jmir_v28i1e80249_app1.docx]

Multimedia Appendix 2
Acceptability of the intervention between groups.
[DOCX File, 22 KB - jmir_v28i1e80249_app2.docx]

Checklist 1
CONSORT checklist: moderate-to-vigorous physical activity.
[PDF File, 184 KB - jmir_v28i1e80249_app3.pdf]

References

1. Noncommunicable diseases. World Health Organization. URL: https://www.who.int/en/news-room/fact-sheets/detail/noncommunicable-diseases [accessed 2026-04-15]
2. Physical activity. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/physical-activity [accessed 2026-04-15]
3. Lim SS, Vos T, Flaxman AD, et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 2012 Dec 15;380(9859):2224-2260. [doi: 10.1016/S0140-6736(12)61766-8] [Medline: 23245609]
4. Prince SA, Elliott CG, Scott K, Visintini S, Reed JL. Device-measured physical activity, sedentary behaviour and cardiometabolic health and fitness across occupational groups: a systematic review and meta-analysis. Int J Behav Nutr Phys Act 2019 Apr 2;16(1):30. [doi: 10.1186/s12966-019-0790-9] [Medline: 30940176]
5. Thorp AA, Healy GN, Winkler E, et al. Prolonged sedentary time and physical activity in workplace and non-work contexts: a cross-sectional study of office, customer service and call centre employees. Int J Behav Nutr Phys Act 2012 Oct 26;9(1):1-9. [doi: 10.1186/1479-5868-9-128] [Medline: 23101767]
6. Clemes SA, O'Connell SE, Edwardson CL. Office workers' objectively measured sedentary behavior and physical activity during and outside working hours. J Occup Environ Med 2014 Mar;56(3):298-303. [doi: 10.1097/JOM.0000000000000101] [Medline: 24603203]
7. Leisure and Cultural Services Department. Territory-wide physical fitness survey for the community: research report. 2023 URL: https://www.lcsd.gov.hk/en/healthy/physical_fitness_test_2021/common/images/PFS%20Research%20Report_ENG.pdf [accessed 2026-04-15]
8. Consultancy study on sport for all: participation patterns of Hong Kong people in physical activities | summary. : Community Sports Committee Leisure and Cultural Services Department; 2009 URL: https://www.lcsd.gov.hk/en/cscommittee/common/form/paper_csc_0805_annex_e.pdf [accessed 2026-04-15]
9. Caputo EL, Reichert FF. Studies of physical activity and COVID-19 during the pandemic: a scoping review. J Phys Act Health 2020 Nov 3;17(12):1275-1284. [doi: 10.1123/jpah.2020-0406] [Medline: 33152693]
10. Stockwell S, Trott M, Tully M, et al. Changes in physical activity and sedentary behaviours from before to during the COVID-19 pandemic lockdown: a systematic review. BMJ Open Sport Exerc Med 2021;7(1):e000960. [doi: 10.1136/bmjsem-2020-000960] [Medline: 34192010]

XSL•FO
RenderX

11. Zheng C, Huang WY, Sheridan S, Sit CHP, Chen XK, Wong SHS. COVID-19 pandemic brings a sedentary lifestyle in young adults: a cross-sectional and longitudinal study. Int J Environ Res Public Health 2020 Aug 19;17(17):6035. [doi: 10.3390/ijerph17176035] [Medline: 32825092]
12. Prince SA, Rasmussen CL, Biswas A, et al. The effect of leisure time physical activity and sedentary behaviour on the health of workers with different occupational physical activity demands: a systematic review. Int J Behav Nutr Phys Act 2021 Jul 20;18(1):100. [doi: 10.1186/s12966-021-01166-z] [Medline: 34284795]
13. Ráthonyi G, Kósa K, Bács Z, et al. Changes in workers' physical activity and sedentary behavior during the COVID-19 pandemic. Sustainability 2021;13(17):9524. [doi: 10.3390/su13179524]
14. Fukushima N, Machida M, Kikuchi H, et al. Associations of working from home with occupational physical activity and sedentary behavior under the COVID-19 pandemic. J Occup Health 2021 Jan;63(1):e12212. [doi: 10.1002/1348-9585.12212] [Medline: 33683779]
15. Erbe D, Eichert HC, Riper H, Ebert DD. Blending face-to-face and internet-based interventions for the treatment of mental disorders in adults: systematic review. J Med Internet Res 2017 Sep 15;19(9):e306. [doi: 10.2196/jmir.6588] [Medline: 28916506]
16. Mouton A, Cloes M. Efficacy of a web-based, center-based or combined physical activity intervention among older adults. Health Educ Res 2015 Jun;30(3):422-435. [doi: 10.1093/her/cyv012] [Medline: 25772974]
17. Hohberg V, Fuchs R, Gerber M, Künzler D, Paganini S, Faude O. Blended care interventions to promote physical activity: a systematic review of randomized controlled trials. Sports Med Open 2022 Jul 30;8(1):100. [doi: 10.1186/s40798-022-00489-w] [Medline: 35907158]
18. Eldredge LKB, Markham CM, Ruiter RA, Fernandez ME, Kok G, Parcel G. Planning Health Promotion Programs: An Intervention Mapping Approach, 4th edition: Jossey Bass; 2016.
19. Cuschieri S. The CONSORT statement. Saudi J Anaesth 2019 Apr;13(Suppl 1):S27-S30. [doi: 10.4103/sja.SJA_559_18] [Medline: 30930716]
20. Juszczak E, Altman DG, Hopewell S, Schulz K. Reporting of multi-arm parallel-group randomized trials: extension of the CONSORT 2010 statement. JAMA 2019 Apr 23;321(16):1610-1620. [doi: 10.1001/jama.2019.3087] [Medline: 31012939]
21. Sun Y, Wang A, Yu S, et al. A blended intervention to promote physical activity, health and work productivity among office employees using intervention mapping: a study protocol for a cluster-randomized controlled trial. BMC Public Health 2020 Jun 25;20(1):994. [doi: 10.1186/s12889-020-09128-z] [Medline: 32586293]
22. Booth M. Assessment of physical activity: an international perspective. Res Q Exerc Sport 2000 Jun;71 Suppl 2(sup2):114-120. [doi: 10.1080/02701367.2000.11082794] [Medline: 25680021]
23. van der Weegen S, Verwey R, Spreeuwenberg M, Tange H, van der Weijden T, de Witte L. It's LiFe! Mobile and web-based monitoring and feedback tool embedded in primary care increases physical activity: a cluster randomized controlled trial. J Med Internet Res 2015 Jul 24;17(7):e184. [doi: 10.2196/jmir.4579] [Medline: 26209025]
24. Davies CA, Spence JC, Vandelanotte C, Caperchione CM, Mummery WK. Meta-analysis of internet-delivered interventions to increase physical activity levels. Int J Behav Nutr Phys Act 2012 Apr 30;9(1):1-13. [doi: 10.1186/1479-5868-9-52] [Medline: 22546283]
25. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
26. Bargh JA. The four horsemen of automaticity: intention, awareness, efficiency, and control as separate issues. In: Wyer RS, Srull TK, editors. Handbook of Social Cognition: Applications: Psychology Press; 1994:1-40.
27. Chaiken S, Trope Y, editors. Dual-Process Theories in Social Psychology, 1st edition: Guilford Press; 1999.
28. Evans JSB, Frankish KE. In Two Minds: Dual Processes and Beyond 2009.
29. Kwasnicka D, Vandelanotte C, Rebar A, et al. Comparing motivational, self-regulatory and habitual processes in a computer-tailored physical activity intervention in hospital employees - protocol for the PATHS randomised controlled trial. BMC Public Health 2017 May 26;17(1):518. [doi: 10.1186/s12889-017-4415-4] [Medline: 28549476]
30. Boudet G, Chausse P, Thivel D, et al. How to measure sedentary behavior at work? Front Public Health 2019;7:167. [doi: 10.3389/fpubh.2019.00167] [Medline: 31355172]
31. Freedson PS, Melanson E, Sirard J. Calibration of the Computer Science and Applications, Inc. accelerometer. Med Sci Sports Exerc 1998 May;30(5):777-781. [doi: 10.1097/00005768-199805000-00021] [Medline: 9588623]
32. Troiano RP, Berrigan D, Dodd KW, Mâsse LC, Tilert T, McDowell M. Physical activity in the United States measured by accelerometer. Med Sci Sports Exerc 2008 Jan;40(1):181-188. [doi: 10.1249/mss.0b013e31815a51b3] [Medline: 18091006]
33. Trost SG, McIver KL, Pate RR. Conducting accelerometer-based activity assessments in field-based research. Med Sci Sports Exerc 2005 Nov;37(11 Suppl):S531-S543. [doi: 10.1249/01.mss.0000185657.86065.98] [Medline: 16294116]
34. Schoeppe S, Duncan MJ, Plotnikoff RC, et al. Acceptability, usefulness, and satisfaction with a web-based video-tailored physical activity intervention: The TaylorActive randomized controlled trial. J Sport Health Sci 2022 Mar;11(2):133-144. [doi: 10.1016/j.jshs.2021.09.001] [Medline: 34487910]

XSL•FO
RenderX

35. Tariman JD, Berry DL, Halpenny B, Wolpin S, Schepp K. Validation and testing of the Acceptability E-scale for web-based patient-reported outcomes in cancer care. Appl Nurs Res 2011 Feb;24(1):53-58. [doi: 10.1016/j.apnr.2009.04.003] [Medline: 20974066]
36. Vandelanotte C, Short CE, Plotnikoff RC, et al. Are web-based personally tailored physical activity videos more effective than personally tailored text-based interventions? Results from the three-arm randomised controlled TaylorActive trial. Br J Sports Med 2021 Mar;55(6):336-343. [doi: 10.1136/bjsports-2020-102521] [Medline: 33144346]
37. White IR, Horton NJ, Carpenter J, Pocock SJ. Strategy for intention to treat analysis in randomised trials with missing outcome data. BMJ 2011 Feb 7;342:d40. [doi: 10.1136/bmj.d40] [Medline: 21300711]
38. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 2001 Apr 14;357(9263):1191-1194. [doi: 10.1016/S0140-6736(00)04337-3] [Medline: 11323066]
39. McEwan D, Beauchamp MR, Kouvousis C, Ray CM, Wyrough A, Rhodes RE. Examining the active ingredients of physical activity interventions underpinned by theory versus no stated theory: a meta-analysis. Health Psychol Rev 2019 Mar;13(1):1-17. [doi: 10.1080/17437199.2018.1547120] [Medline: 30412685]
40. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010 Feb 17;12(1):e4. [doi: 10.2196/jmir.1376] [Medline: 20164043]
41. Burke SM, Carron AV, Eys MA, Ntoumanis N, Estabrooks PA. Group versus individual approach? A meta-analysis of the effectiveness of interventions to promote physical activity. bpssepr 2006 Feb;2(1):13-29. [doi: 10.53841/bpssepr.2006.2.1.13]
42. Santarossa S, Kane D, Senn CY, Woodruff SJ. Exploring the role of in-person components for online health behavior change interventions: can a digital person-to-person component suffice? J Med Internet Res 2018 Apr 11;20(4):e144. [doi: 10.2196/jmir.8480] [Medline: 29643048]
43. Neal DT, Wood W, Quinn JM. Habits—a repeat performance. Curr Dir Psychol Sci 2006 Aug;15(4):198-202. [doi: 10.1111/j.1467-8721.2006.00435.x]
44. Muellmann S, Buck C, Voelcker-Rehage C, et al. Effects of two web-based interventions promoting physical activity among older adults compared to a delayed intervention control group in Northwestern Germany: results of the PROMOTE community-based intervention trial. Prev Med Rep 2019 Sep;15:100958. [doi: 10.1016/j.pmedr.2019.100958] [Medline: 31410347]
45. Kelders SM, Van Gemert-Pijnen J, Werkman A, Nijland N, Seydel ER. Effectiveness of a web-based intervention aimed at healthy dietary and physical activity behavior: a randomized controlled trial about users and usage. J Med Internet Res 2011 Apr 14;13(2):e32. [doi: 10.2196/jmir.1624] [Medline: 21493191]
46. Hansen AW, Grønbæk M, Helge JW, Severin M, Curtis T, Tolstrup JS. Effect of a web-based intervention to promote physical activity and improve health among physically inactive adults: a population-based randomized controlled trial. J Med Internet Res 2012 Oct 30;14(5):e145. [doi: 10.2196/jmir.2109] [Medline: 23111127]
47. Irvine AB, Gelatt VA, Seeley JR, Macfarlane P, Gau JM. Web-based intervention to promote physical activity by sedentary older adults: randomized controlled trial. J Med Internet Res 2013 Feb 5;15(2):e19. [doi: 10.2196/jmir.2158] [Medline: 23470322]
48. Bishop FL, Fenge-Davies AL, Kirby S, Geraghty AWA. Context effects and behaviour change techniques in randomised trials: a systematic review using the example of trials to increase adherence to physical activity in musculoskeletal pain. Psychol Health 2015 Jan;30(1):104-121. [doi: 10.1080/08870446.2014.953529] [Medline: 25109300]
49. Beleigoli A, Andrade AQ, Diniz MDF, Ribeiro AL. Personalized web-based weight loss behavior change program with and without dietitian online coaching for adults with overweight and obesity: randomized controlled trial. J Med Internet Res 2020 Nov 5;22(11):e17494. [doi: 10.2196/17494] [Medline: 33151151]
50. Althammer SE, Wöhrmann AM, Michel A. Comparing web-based and blended training for coping with challenges of flexible work designs: randomized controlled trial. J Med Internet Res 2023 Dec 19;25:e42510. [doi: 10.2196/42510] [Medline: 38113084]

Abbreviations

BCT: behavior change technique
CONSORT: Consolidated Standards of Reporting Trials
e-workshop: electronic workshop
IM: intervention mapping
ITT: intention-to-treat
MI: multiple imputation
MVPA: moderate-to-vigorous physical activity
NCDs: noncommunicable diseases
RCT: randomized controlled trial
WHO: World Health Organization

XSL•FO
RenderX

Edited by A Stone; submitted 10.Jul.2025; peer-reviewed by A Adepoju, J Fogel, V Sekhon; revised version received 02.Mar.2026; accepted 02.Mar.2026; published 22.May.2026.

Please cite as:
Sun Y, Gao Y, Ou AYT, Wang X, Xie Y, Chen XY, Fong SSM, Huang WY, Zhang CQ, Baker JS, Hagger MS
Effectiveness of a Blended Intervention to Promote Physical Activity Among Office Employees: Randomized Controlled Trial
J Med Internet Res 2026;28:e80249
URL: https://www.jmir.org/2026/1/e80249
doi:10.2196/80249

© Yan Sun, Yang Gao, Alison Y T Ou, Xueying Wang, Yaojie Xie, Xiang-yan Chen, Shirley S M Fong, Wendy Yajun Huang, Chun-Qing Zhang, Julien S Baker, Martin S Hagger. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Self-Guided Internet-Based Mindfulness-Informed Stress Management for Generalized Anxiety Disorder: Randomized Controlled Trial With Longitudinal Network Analysis

Ziwei Wang[1,2], MD; Hui Qi Tong[3], PhD; Jianrong Yue[1,2], BSN; Siyan Wu[1,2], MM; Ye Xia[1,2], MD; Han Zhang[1,2*], MD; Yuan Yang[1,2*], MD

[1]Department of Neurology and Psychiatry, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, 1095 Jiefang Avenue, Wuhan, Hubei, China

[2]Hubei Key Laboratory of Neural Injury and Functional Reconstruction, Huazhong University of Science and Technology, Wuhan, Hubei, China

[3]Department of Psychiatry and Behavioral Sciences, Stanford University, Palo Alto, CA, United States

[*]these authors contributed equally

Corresponding Author:
Yuan Yang, MD
Department of Neurology and Psychiatry, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, 1095 Jiefang Avenue, Wuhan, Hubei, China

Abstract

Background: Generalized anxiety disorder (GAD) is characterized by stress-anxiety reinforcement cycles. Evidence for brief, self-guided, internet-based stress management as an adjunct to pharmacotherapy remains limited.

Objective: This study aimed to evaluate the efficacy of an 8-week self-guided, internet-based stress management program (iSM) incorporating mindfulness techniques as an adjunct to treatment as usual (TAU) in adults with GAD and to examine individual predictors of response and symptom dynamics.

Methods: A single-blind, parallel-group, superiority randomized controlled trial was conducted at Tongji Hospital, Wuhan, China. A total of 140 adults with GAD were randomly assigned to iSM+TAU (n=73) or TAU (n=67). Outcome assessors were blinded to group allocation. The iSM intervention consisted of 8 weekly self-guided online modules integrating mindfulness-based training and Baduanjin-based stretching exercises. TAU comprised routine pharmacotherapy. Primary outcomes were posttreatment changes in anxiety and depressive symptoms. Secondary outcomes included sleep quality, somatic symptoms, social functioning, mindfulness, rumination, and perceived stress. Exploratory cross-lagged panel network (CLPN) and random-intercept cross-lagged panel modeling (RI-CLPM) analyses were used to examine temporal symptom dynamics.

Results: All 140 randomized participants were included in the intention-to-treat analysis. Compared with TAU, iSM+TAU showed greater reductions at posttreatment in anxiety (Cohen d=−0.277, 95% CI −0.521 to −0.033) and depressive symptoms (Cohen d=−0.309, 95% CI −0.592 to −0.026). Significant between-group differences were also observed in somatic symptoms (Cohen d=−0.340, 95% CI −0.604 to −0.077), state anxiety (Cohen d=−0.537, 95% CI −0.849 to −0.224), mindfulness (Cohen d=0.666, 95% CI 0.327-1.006), rumination (Cohen d=−0.344, 95% CI −0.626 to −0.062), and perceived stress (Cohen d=−0.429, 95% CI −0.725 to −0.133), but not in sleep quality or social functioning. No serious adverse events were reported. Median session completion was 7 of 8. In exploratory analyses, higher baseline acting with awareness predicted greater treatment response (standardized β=0.167; 95% CI 0.031-0.335), whereas higher trait anxiety predicted poorer outcomes (standardized β=−0.150; 95% CI −0.234 to −0.002). CLPN and RI-CLPM identified 2 key within-person pathways, a bidirectional association between perceived stress (mean β=0.219; P<.001) and state anxiety (mean β=0.165; P=.02), and a unidirectional effect of mindfulness on subsequent anxiety reduction (mean β=–0.285; P<.001).

Conclusions: To our knowledge, this is the first trial to evaluate a brief, self-guided, culturally adapted digital mindfulness intervention as an adjunct to pharmacotherapy in Chinese adults with GAD. The intervention showed clinically meaningful benefits and practical potential as a scalable, resource-efficient approach. Unlike prior studies focused mainly on symptom outcomes alone, this trial combined a randomized design with CLPN and RI-CLPM analyses to provide preliminary insight into symptom change processes over time, adding a more process-oriented analytic perspective that may inform future intervention refinement.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2300078470; https://www.chictr.org.cn/showproj.html?proj=200673

(J Med Internet Res 2026;28:e91751) doi:10.2196/91751

XSL•FO
RenderX

KEYWORDS

generalized anxiety disorder; internet-based intervention; mindfulness; stress management; randomized controlled trial; longitudinal network analysis

Introduction

Generalized anxiety disorder (GAD) is a prevalent psychiatric condition affecting approximately 3.7% of the global population [1]. Beyond its hallmark symptoms of persistent worry and tension, GAD is increasingly conceptualized as involving central impairments in stress regulation mechanisms [2,3]. Patients with GAD exhibit heightened physiological and psychological stress responses, impaired stress recovery, and maladaptive stress appraisal patterns [3,4]. This chronic stress reactivity not only triggers anxiety episodes but also maintains the disorder through self-reinforcing stress-anxiety cycles [3,5]. Understanding GAD through this stress-centered lens suggests that interventions targeting stress management may address core maintaining mechanisms rather than merely alleviating downstream symptoms [6].

At the time of this writing, first-line treatments for GAD include pharmacotherapy and cognitive behavioral therapy (CBT) [7]. While pharmacological interventions provide symptomatic relief, they primarily target downstream manifestations rather than upstream stress processes, and adverse effects, including sedation and withdrawal, often limit long-term adherence [8,9]. Relapse rates of 56% following medication discontinuation further suggest that pharmacotherapy alone does not adequately address the disorder [10]. CBT, which includes cognitive restructuring as well as behavioral and physiological regulation techniques, is a well-established first-line treatment for GAD [7,11]. However, its uptake in routine care remains limited by constraints in training and supervision, underdeveloped service infrastructure, and practical barriers to patient participation [12,13]. Scalable adjunctive interventions that further strengthen stress management and coping capacity in daily life may therefore still have important clinical value. Stress management interventions incorporating mindfulness techniques have demonstrated promise for anxiety disorders [14]. Mindfulness-based stress reduction (MBSR), a well-established program that helps patients cope with chronic stress, teaches skills for recognizing and adaptively responding to stress rather than reacting automatically [15]. Meta-analyses support small-to-moderate effects on anxiety symptoms [14,16,17]. Mindfulness-based programs, including some web-based formats, also appear generally safe and well tolerated [18]. However, traditional face-to-face programs remain underused due to barriers including limited therapist availability, geographic constraints, and scheduling inflexibility [19]. Cultural factors further influence treatment engagement; for instance, Chinese individuals may be less accustomed to group-based psychological interventions involving open discussion of internal experiences. In resource-constrained settings like China, where mental health professionals number only 2.19 per 100,000 population, scalable stress management solutions are urgently needed [20,21].

Internet-delivered self-guided interventions offer a promising solution for expanding access to stress management training [22,23]. Digital formats eliminate therapist requirements, enabling immediate scalability without staffing constraints. Mindfulness-based interventions are theoretically well suited to target core maintaining mechanisms of GAD, including attentional dysregulation, worry reactivity, and experiential avoidance [16,24]. Core skills such as sustained attention and decentering from perseverative thinking may be strengthened through structured self-guided practice, as standardized sequential delivery can facilitate consolidation in daily life without continuous therapist support [15,25]. Several randomized controlled trials (RCTs) support efficacy for internet-based programs targeting individuals with GAD, although most evaluated stand-alone interventions [26-28]. Evidence supporting brief, self-guided, internet-based stress management as an adjunct to first-line pharmacotherapy for GAD remains limited. To address this gap, we developed an 8-week internet-delivered stress management program integrating mindfulness-based techniques with culturally adapted mind-body exercises (Baduanjin), designed specifically for self-guided use by Chinese adults with GAD.

Despite growing interest in digital stress management interventions, several critical gaps remain. First, most studies prioritize symptom reduction while overlooking broader functional outcomes such as sleep quality and social functioning [29], and evidence from non-Western clinical populations receiving pharmacological treatment remains scarce. Second, limited understanding of individual predictors constrains efforts toward personalized intervention [30]. Third, although theoretical models suggest that stress reduction mediates anxiety improvement, the temporal dynamics of stress-anxiety interactions during treatment remain underexplored [31]. Clarifying how these constructs influence each other over time—and how interventions disrupt maladaptive patterns—may uncover core mechanisms and guide optimization. Emerging longitudinal network approaches offer promising tools in this regard [32,33]. Specifically, cross-lagged panel network (CLPN) analysis can be used to examine directional associations among symptoms over time [34,35], whereas the random intercept cross-lagged panel model (RI-CLPM) more explicitly accounts for stable between-person differences and thereby provides a more stringent test of whether key associations are consistent with within-person temporal dynamics [36,37]. Together, these methods may help generate hypotheses about temporal pathways of symptom change and identify clinically relevant processes that may inform future intervention refinement and personalization.

In this study, we conducted an RCT evaluating an 8-week internet-based, self-guided stress management program as an adjunct to treatment as usual (TAU) among adults with GAD. We hypothesized that the intervention would reduce anxiety symptoms and that perceived stress and mindfulness would serve as key processes underlying symptom change. We also used regression analyses to identify predictors of treatment

XSL•FO
RenderX

response and combined CLPN with RI-CLPM to explore temporal dynamics of the hypothesized stress-anxiety pathway.

Methods

Trial Design and Setting

This study was a single-blind, parallel-group, superiority RCT evaluating an 8-week online self-help mindfulness-based stress management program as an adjunct to pharmacological treatment for patients with GAD. Participants were randomized at the individual level. The trial was prospectively registered at the Chinese Clinical Trial Registry (registration number: ChiCTR2300078470) on December 8, 2023, prior to enrollment of the first participant on December 20, 2023. This trial is reported in accordance with the CONSORT (Consolidated Standards of Reporting Trials) 2025 statement and the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) extension for web-based and mobile health interventions [38,39]. The completed CONSORT 2025 and CONSORT-EHEALTH checklists are provided as Checklist 1 and Checklist 2, respectively. This was a single-center trial conducted at Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China. Participants were consecutively recruited from the outpatient anxiety and depression clinics across 3 campuses of Tongji Hospital (Main Campus, Optical Valley Branch, and Sino-French New City Branch) between December 2023 and September 2024. Follow-up for the final participant, including outcome and adverse event assessment, was completed in December 2024. The full study protocol was not publicly posted or separately preregistered before trial completion, but it is available from the corresponding author upon reasonable request.

Ethical Considerations

The study was approved by the Ethics Committee of Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology (approval number: TJ-IRB20230617) and was conducted in accordance with the Declaration of Helsinki and relevant ethical guidelines. In April 2024, an approved protocol amendment (TJ-IRB202404034) added electroencephalography and magnetic resonance imaging assessments for mechanistic exploration; these measures were not part of the original protocol and are not reported here. Written informed consent was obtained from all participants after they received a full explanation of the study purpose, procedures, and potential risks. Study data were deidentified prior to analysis and were accessible only to authorized research personnel. No financial compensation was provided for participation. The screenshot from the course interface shown in Figure S1 in Multimedia Appendix 1 was included with the written informed consent of the individual depicted. No other identifiable participant information is included in the manuscript or supplementary materials.

Eligibility Criteria

Participants were screened for eligibility according to the following inclusion and exclusion criteria. Inclusion criteria were as follows: (1) a primary diagnosis of GAD according to the *DSM-5 (Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]); (2) a Hamilton Anxiety Rating Scale (HAMA) score of ≥7; (3) age ≥18 years; (4) access to a smartphone or computer with a stable internet connection; and (5) sufficient cognitive and behavioral ability to follow the intervention procedures and complete study assessments. Exclusion criteria were as follows: (1) a diagnosis of schizophrenia, bipolar disorder, major depressive disorder with suicide risk, or substance use disorder (eg, alcohol or drug abuse) within the past 6 months; (2) receipt of formal psychotherapy (eg, CBT, psychodynamic therapy) or structured mindfulness training within the past 6 months; (3) severe visual, auditory, or cognitive impairments; (4) use of psychotropic medication (eg, antidepressants, antipsychotics, and benzodiazepines) within the past 3 months prior to enrollment; (5) pregnant, planned pregnancy, or breastfeeding; (6) concurrent participation in another clinical trial or enrollment of a first-degree relative in this study; (7) inability to cooperate in the intervention or data collection process; and (8) any other condition that, in the opinion of the investigators, would interfere with participation or data integrity. The primary diagnosis of GAD was established by the treating psychiatrist according to *DSM-5* criteria during routine psychiatric evaluation. Inclusion criteria were used for eligibility screening. Major psychiatric exclusion criteria were assessed during the same evaluation based on clinical interview, clinical history, and available medical records. Detailed procedures are provided in Multimedia Appendix 1.

Procedures

At baseline, all participants underwent a comprehensive assessment, including demographic information, clinical history, and psychological evaluations. Eligible participants were randomly assigned to either the internet-based stress management program (iSM) plus TAU group or the TAU group. The iSM+TAU group received internet-based mindfulness stress management combined with routine pharmacological treatment, while the TAU group received only pharmacological treatment. Assessments were conducted at baseline (T0), midintervention (T1, Week 4), postintervention (T2, Week 8), and at 3-month follow-up (T3) to evaluate efficacy and safety, including psychological assessments, medication adjustments, adherence, and adverse event monitoring.

Intervention

TAU Intervention

TAU was selected as the comparator to reflect real-world clinical practice and to evaluate the incremental benefit of the digital stress management program when added to standard pharmacological care. Participants in the TAU group received pharmacological treatment primarily based on selective serotonin reuptake inhibitors. Short-term benzodiazepines (≤4 wk) were permitted when clinically indicated, under physician supervision. Medication selection and dosage adjustments were made by attending psychiatrists in accordance with standard clinical practice, based on each patient's clinical presentation, treatment needs, and tolerability. To support engagement and ensure comparable contact exposure across groups, the research

XSL•FO
RenderX

team maintained participant communication through a designated WeChat (Tencent) group. Interactions with TAU participants were strictly limited to data collection and follow-up reminders, with no psychological support or behavioral guidance provided. To ensure ethical fairness, participants in the TAU group were offered access to the iSM after the study concluded. However, any data generated from this posttrial intervention were outside the scope of this analysis.

Mindfulness-Based Stress Management Program

Participants in the iSM+TAU group received an 8-week internet-delivered mindfulness-based stress management program in addition to the same pharmacological treatment provided to the TAU group. The intervention was delivered through a dedicated online learning platform, which provided structured weekly modules comprising instructional videos, audio-guided meditations, and daily mindfulness practice assignments. Participants were required to complete a 40‑60-minute core session each week, which served as a prerequisite for accessing the following session. The course emphasized progressive skill acquisition, covering core elements of mindfulness such as nonjudgmental present-moment awareness, breath-focused attention, and open monitoring. The weekly content followed a progressive thematic structure, and details of the weekly curriculum are provided in Table S1 of Multimedia Appendix 1. The first 4 sessions encompassed the core components of mindfulness training. Therefore, completion of at least 4 sessions was prespecified as the minimum threshold for adequate intervention exposure. All participants in the iSM group received the same prerecorded modules and standardized weekly content, while a separate WeChat group provided reminders and asynchronous technical support only, without individualized mindfulness guidance, live coaching, or psychotherapy. Individual progress was continuously monitored through the learning management system, which captured module completion status, video viewing duration, number of completed sessions, and cumulative practice time as indicators of intervention engagement (Figure S2 in Multimedia Appendix 1). Apart from the structured intervention modules, communication was limited to administrative and technical matters and kept consistent between study arms.

The mindfulness-based stress management program was adapted from the MBSR course originally developed by Kabat-Zinn [40] and was developed, reviewed, and adapted with input from a certified MBSR instructor. The prerecorded sessions were delivered by an experienced mindfulness therapist who had received standardized training in mindfulness-based practice. Compared to traditional face-to-face mindfulness programs, this intervention was shortened in duration to improve feasibility and adherence, making it more suitable for a self-paced, internet-based format. To enhance cultural relevance and accessibility, stretching exercises were adapted from Baduanjin (commonly known as the Eight Section Brocade), a traditional Chinese Qigong practice consisting of 8 gentle movements aimed at supporting flexibility, balance, and physical well-being instead of using conventional yoga-based stretching. All components of the intervention were delivered in a fully self-guided format without real-time therapist involvement. No structured psychotherapy, formal mindfulness training, or other concomitant interventions that could affect study outcomes were permitted during the trial.

Outcomes

Baseline demographic and clinical data included age, sex, BMI, education, residence type, marital status, childhood parental separation, employment, living situation, and medication use.

The primary outcomes were improvement in anxiety and depression from baseline to posttreatment (8 wk), measured with the Hamilton Anxiety Rating Scale [41] (HAMA; overall severity of anxiety symptoms) and the 21-item Hamilton Depression Rating Scale [42] (HAMD-21; overall severity of depressive symptoms), with additional assessments at midtreatment (4 wk) and 3-month follow-up to evaluate treatment trajectory and durability of effects.

Secondary outcomes were assessed at all time points, including state anxiety (State Anxiety Inventory [S-AI]) [43], sleep quality (Pittsburgh Sleep Quality Index [PSQI]) [44], mindfulness (Five Facet Mindfulness Questionnaire [FFMQ]) [45], somatic symptoms (the 15-item Patient Health Questionnaire [PHQ-15]) [46], social functioning (Social Disability Screening Schedule [SDSS]) [47], rumination (Ruminative Responses Scale [RRS]) [48], and perceived stress (the 10-item Perceived Stress Scale [PSS-10]) [49]. The public trial registration listed the primary outcomes and selected secondary outcomes but did not fully reflect the complete questionnaire-based longitudinal assessment battery used in the study. Several registered objective assessments were supplementary rather than core longitudinal outcomes and are not reported here because data completeness was insufficient for reliable analysis.

Several psychological trait measures were collected at baseline only, including trait anxiety (Trait Anxiety Inventory [T-AI]) [43], perceived security (Security Questionnaire [SQ]) [50], perceived social support (Perceived Social Support Scale [PSSS]) [51], loneliness (UCLA Loneliness Scale [UCLA-LS]) [52], interpersonal trust (Interpersonal Trust Scale [ITS]) [53], and social avoidance and distress (Social Avoidance and Distress Scale [SAD]) [54]. Detailed descriptions of each scale, scoring procedures, and psychometric properties are provided in the Multimedia Appendix 1.

Harms

Adverse events were assessed at each scheduled study assessment during the intervention and follow-up periods through participant self-report and routine clinical inquiry. Participants were asked about any unintended negative experiences during the study, including symptom worsening, physical discomfort, or other clinically relevant concerns. Medication selection and adjustment, including switching medications because of tolerability or clinical response, were allowed within TAU and were managed by the treating psychiatrists according to routine clinical practice. Reported adverse events were reviewed by the study psychiatrist, who assessed their severity and their likely relation to the internet-based intervention or concurrent pharmacotherapy on the basis of clinical judgment. Serious adverse events were defined as clinically significant deterioration requiring emergency psychiatric or medical evaluation or hospitalization,

XSL•FO
RenderX

including suicidality or self-harm, severe worsening of anxiety or depressive symptoms, psychotic symptoms, or other acute conditions requiring immediate clinical intervention. Serious adverse events and withdrawals related to adverse events were also specifically recorded throughout the trial.

Randomization and Blinding

Participants were randomly assigned in a 1:1 ratio using simple randomization with a computer-generated sequence in R (R Core Team). The randomization sequence was generated and held by an independent researcher who was not involved in participant recruitment, outcome assessment, or intervention delivery. Eligible participants were enrolled by study investigators. The allocation sequence was not accessible to recruiting staff before assignment and was released only after eligibility had been confirmed and baseline assessment had been completed. Outcome assessors were not involved in intervention delivery and did not have access to the allocation sequence. The trial adopted a single-blind design; participants were necessarily aware of their allocation due to the nature of the intervention, but outcome assessors were blinded. To maintain blinding during analysis, datasets were anonymized and coded prior to statistical evaluation.

Patient and Public Involvement

Patients and members of the public were not involved in the design, conduct, reporting, or dissemination planning of this trial.

Sample Size and Statistical Analysis

Sample Size Calculation

The sample size calculation was based on the primary outcome of anxiety symptom improvement. Using G*Power 3.1 (Heinrich Heine University, Düsseldorf) and informed by a previous study [55], we conservatively assumed a medium effect size (Cohen d=0.50). With a 2-tailed α=.05 and 80% power, a minimum of 128 participants (64 per group) was required. Allowing for a 20% dropout rate, the study was preregistered with a recruitment target of 160 participants. Due to funding and time constraints, 140 participants were ultimately randomized (iSM+TAU: n=73; TAU: n=67). This sample size exceeded the minimum required and provided approximately 84% power to detect an effect size of d=0.50 under unequal group sizes (2-tailed α=.05).

Statistical Analysis

All statistical analyses were conducted using SPSS (version 26.0; IBM Corp) and R (version 4.5.1; R Core Team). The primary analysis followed the intention-to-treat (ITT) principle, including all randomized participants analyzed in their assigned groups. Missing data patterns were assessed using Little's missing completely at random test [56]. Missing data were then handled using multiple imputation under the assumption that the data were missing at random. Per-protocol (PP) analyses were conducted as sensitivity analyses among participants who completed the study according to protocol. In the iSM+TAU group, the PP sample included participants who completed at least 4 intervention sessions and provided posttreatment primary outcome data. In the TAU group, the PP sample included participants who completed the posttreatment primary outcome assessment. No interim analyses were planned or conducted, and no stopping guidelines were prespecified. All tests were 2-sided with a significance level of α=.05, and effect estimates were reported with 95% CIs where appropriate. Adverse events were summarized descriptively.

Intervention Effects

Linear mixed effects modeling was used to account for the hierarchical structure of the data, with repeated measurements (time points) nested within individuals, using the lmerTest package developed by Kuznetsova, Brockhoff, and Christensen. The models included Group (iSM+TAU vs TAU), Time (T0, T1, T2, and T3), and the group×time interaction as fixed effects, with a random intercept for participants. Effect sizes were estimated using Cohen d derived from the estimated marginal means.

Predictors of Responses

Pearson or Spearman correlation analyses were performed to examine associations between baseline variables and primary outcomes. Variables showing significant associations (P<.05) were entered into multiple linear regression models, adjusting for demographic characteristics. These analyses were conducted on an exploratory basis.

Network and Temporal Dynamics Analysis

A 2-step exploratory analytic strategy was used to examine temporal symptom associations and to further evaluate selected pathways under a model that more explicitly accounts for stable between-person differences. All analyses were performed in R version 4.5.1 using the *bootnet*, *qgraph*, and *glmnet packages* for network estimation and the *lavaan* and *semPlot* packages for structural equation modeling.

First, CLPNs were estimated across 3 intervals (T0-T1, T1-T2, and T2-T3) using least absolute shrinkage and selection operator (LASSO) regression to explore directional associations among 6 core variables (HAMA, HAMD-21, FFMQ, PSQI, S-AI, and PSS-10). These variables were selected a priori to represent the hypothesized stress-management pathway and key GAD-related symptom domains. Confidence intervals for edge estimates were obtained using nonparametric bootstrap resampling. Network stability was evaluated using 3000 case-dropping bootstraps, and edges present in >90% of samples were considered stable.

Second, RI-CLPMs were constructed to further examine selected pathways identified in the CLPN analysis in a framework that more explicitly accounts for stable between-person differences through random intercepts. The models applied stationarity constraints and were estimated using robust maximum likelihood estimation. Model adequacy was evaluated using conventional fit indices, comparative fit index (CFI>0.90), root-mean-square error of approximation (RMSEA<0.08), and standardized root-mean-square residual (SRMR<0.08). Detailed model specifications and estimation procedures are provided in the Multimedia Appendix 1.

XSL•FO
RenderX

Results

Sample Characteristics

Of 272 participants screened, 140 eligible participants were randomized to iSM+TAU (n=73) or TAU (n=67; Figure 1). Following randomization, 7 participants in the iSM+TAU group and 4 in the TAU group withdrew before initiating the intervention. At posttreatment, 59 participants in the iSM+TAU group and 49 in the TAU group completed the primary outcome assessment. By the 3-month follow-up, 54 participants in the iSM+TAU group and 41 in the TAU group had completed all assessments. Of the 73 participants randomized to the iSM+TAU group, 74% (54/73) completed at least 4 sessions, and the median number of sessions completed in the full intervention sample was 7 (IQR 3‐8). Missing data rates were 15.7% (22/140) at midintervention (4 wk, T1), 22.9% (32/140) at posttreatment (8 wk, T2), and 32.1% (45/140) at the 3-month follow-up (T3). Across all postbaseline assessments, 99 of 420 expected observations were missing, corresponding to an overall missingness rate of 23.6%. Little missing completely at random test was nonsignificant (χ^2_{172}=184.2; P=.25), indicating no significant departure from missing at random.

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram of participants with generalized anxiety disorder (GAD) in the randomized controlled trial conducted at Tongji Hospital. Participants were screened for eligibility, randomized to the internet-based stress management+treatment as usual (iSM+TAU) group or the TAU group, and followed through the intervention period, posttreatment assessment, 3-month follow-up, and inclusion in the final analyses, with reasons for not receiving the allocated intervention and loss to follow-up. GAD: generalized anxiety disorder; iSM: internet-based stress management program; ITT: intention-to-treat; TAU: treatment as usual.

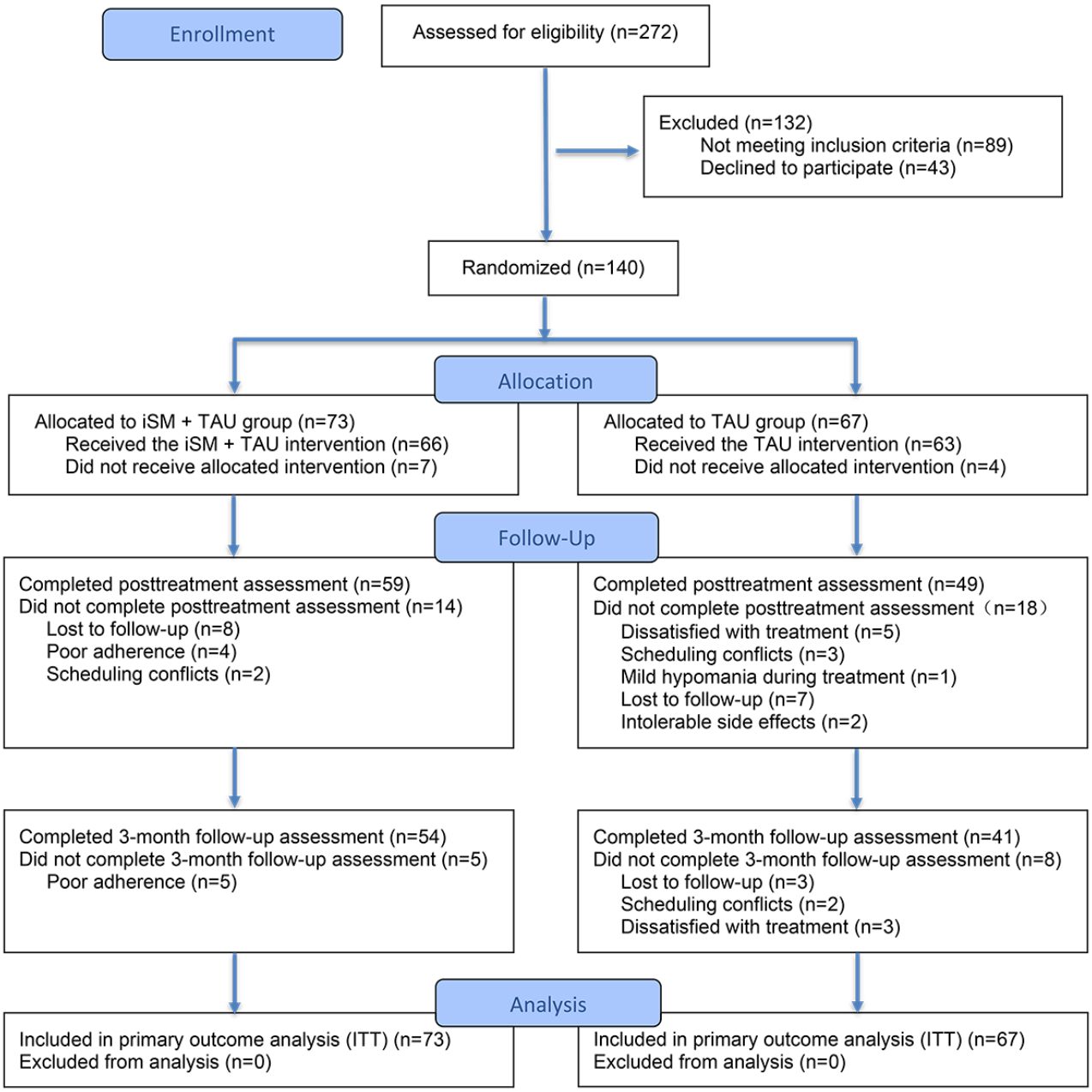

XSL•FO
RenderX

No significant differences were observed in any baseline characteristics between the iSM+TAU and TAU groups (all *P*>.05; Table S2 in Multimedia Appendix 1). The median age of all participants was 33 (IQR 26 - 39) years, and 66.4% (93/140) were female. The sample was predominantly urban and relatively well educated, 81.4% (114/140) were from urban areas, and 77.9% (109/140) had a college education or above. Medication use was similar between the 2 groups (*P*>.05), with selective serotonin reuptake inhibitor monotherapy being the most common regimen, accounting for 74% (54/73) in the iSM+TAU group and 70.1% (47/67) in the TAU group.

Primary Outcome Analysis

All randomized participants (n=140) were included in the primary ITT analyses. Linear mixed-effects models showed significant group×time interactions for both anxiety and depressive symptoms (*P*<.001). At posttreatment (T2), the iSM+TAU group demonstrated significantly greater reductions in anxiety symptoms (HAMA: mean difference [MD]=−1.994, 95% CI −3.758 to −0.230; *P*=.03; Cohen *d*=−0.277, 95% CI −0.521 to −0.033) and depressive symptoms (HAMD-21: MD=−1.649, 95% CI −3.165 to −0.133; *P*=.03; Cohen *d*=−0.309, 95% CI −0.592 to −0.026) compared with the TAU group. These between-group differences further increased at the 3-month follow-up, reaching moderate effect sizes for both HAMA (Cohen *d*=−0.511, 95% CI −0.755 to −0.267) and HAMD-21 (Cohen *d*=−0.637, 95% CI −0.920 to −0.355), as illustrated in Figures 2A and 2B. Detailed estimates are provided in Table S3 in Multimedia Appendix 1, and consistent findings were observed in the PP sensitivity analysis (Table S4 in Multimedia Appendix 1).

Figure 2. Changes in primary and secondary outcome measures over time among participants with generalized anxiety disorder (GAD) in the intention-to-treat (ITT) analysis of the randomized controlled trial at Tongji Hospital (N=140). Panels show longitudinal changes in (A) anxiety, (B) depression, (C) somatic symptoms, (D) state anxiety, (E) mindfulness, (F) perceived stress, (G) sleep quality, (H) rumination, and (I) functional impairment from baseline (T0) to midintervention (T1), posttreatment (T2), and 3-month follow-up (T3). Error bars represent standard errors of the mean. FFMQ: Five Facet Mindfulness Questionnaire; HAMA: Hamilton Anxiety Rating Scale; HAMD-21: 21-item Hamilton Depression Rating Scale; ITT: intention-to-treat; iSM: internet-based stress management program; PHQ-15: 15-item Patient Health Questionnaire; PSQI: Pittsburgh Sleep Quality Index; PSS-10: 10-item Perceived Stress Scale; RRS: Ruminative Responses Scale; SDSS: Social Disability Screening Schedule; S-AI: State Anxiety Inventory; TAU: treatment as usual; *: *P*<.05; **: *P*<.01; ***: *P*<.001.

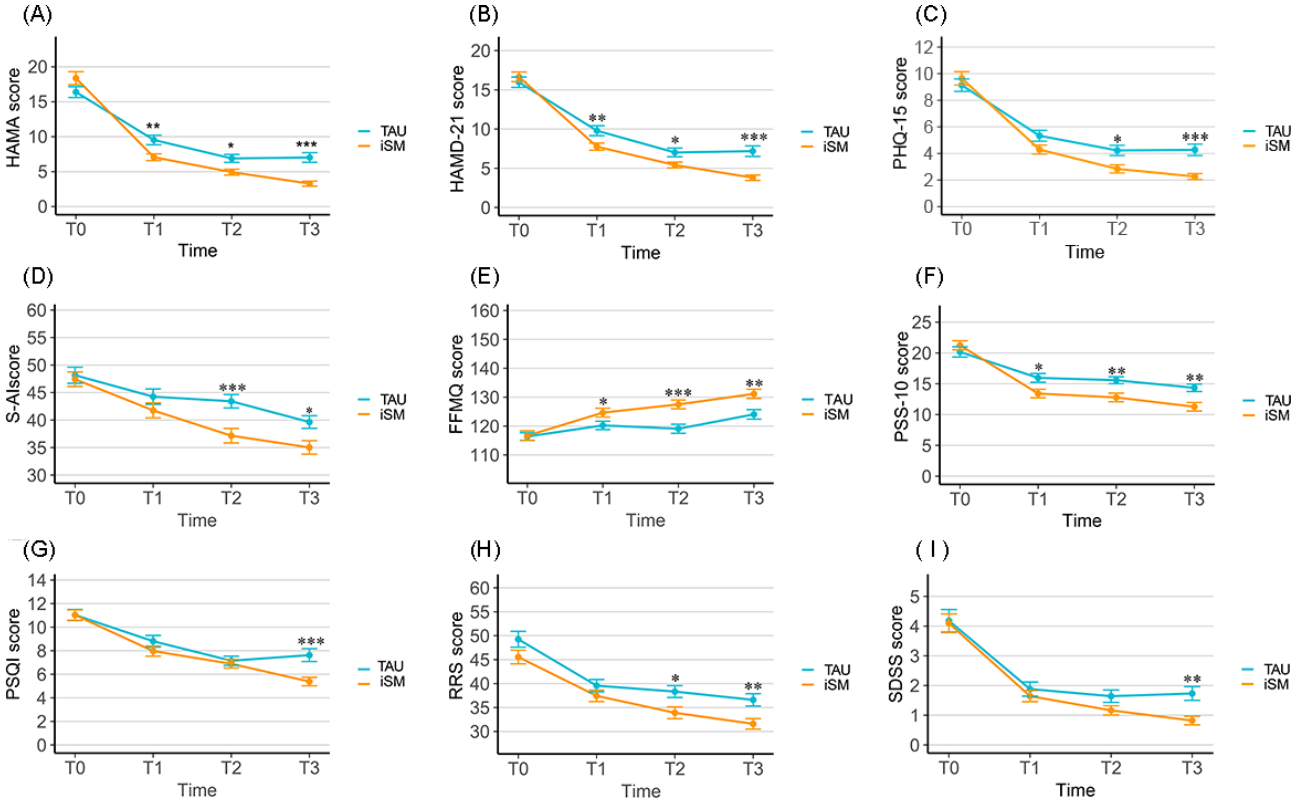

Secondary Outcome Analysis

In the ITT sample, significant group×time interactions were found for PHQ-15, PSQI, S-AI, FFMQ, and PSS-10 (Table S3 in Multimedia Appendix 1). At posttreatment, the iSM+TAU group showed significant improvements in somatic symptoms (PHQ-15: MD=−1.388, 95% CI −2.465 to −0.312; *P*=.01; Cohen *d*=−0.340, 95% CI −0.604 to −0.077; Figure 2C), state anxiety (S-AI: MD=−6.254, 95% CI −9.905 to −2.602; *P*<.001; Cohen *d*=−0.537, 95% CI −0.849 to −0.224; Figure 2D), mindfulness level (FFMQ: MD=8.433, 95% CI 4.119-12.748; *P*<.001; Cohen *d*=0.666, 95% CI 0.327-1.006; Figure 2E), and perceived stress (PSS-10: MD=−2.801, 95% CI −4.740 to −0.863; *P*=.005; Cohen *d*=−0.429, 95% CI −0.725 to −0.133; Figure 2F). Sleep quality (PSQI: MD=−0.259, 95% CI −1.479 to 0.962; *P*=.68; Cohen *d*=−0.068, 95% CI −0.386 to 0.251) showed no significant between-group difference at posttreatment but demonstrated a significant effect at follow-up (MD=−2.243, 95% CI −3.464 to −1.023, *P*<.001; Cohen *d*=−0.588, 95% CI −0.906 to −0.269; Figure 2G).

XSL•FO
RenderX

Rumination (RRS: MD=−4.424, 95% CI −8.069 to −0.780; *P*=.02; Cohen *d*=−0.344, 95% CI −0.626 to −0.062) showed significant posttreatment reductions, although the overall group×time interaction was not significant (*P*=.38; Figure 2H). Social functioning measured by the SDSS showed no significant between-group difference at posttreatment (MD=−0.424, 95% CI −1.091 to 0.243; *P*=.21; Cohen *d*=−0.146, 95% CI −0.376 to 0.083), with significant effects emerging only at follow-up (MD=−0.947, 95% CI −1.614 to −0.279; *P*=.006; Cohen *d*=−0.327, 95% CI −0.557 to −0.097; Figure 2I). The PP sensitivity analysis revealed consistent patterns of change across all secondary outcomes (Table S4 in Multimedia Appendix 1).

Harms

No serious adverse events were reported during the trial. A limited number of participants (15/140, 10.7%) reported mild transient dizziness or nausea, which was considered more likely related to concurrent pharmacotherapy than to the internet-based behavioral intervention. No adverse events led to study withdrawal.

Dose-Response Relationship and Treatment Adherence

Of the 73 participants randomized to the iSM+TAU group, 7 (9.6%) did not complete a full session, 5 (6.8%) completed 1 session, 4 (5.5%) completed 2 sessions, 3 (4.1%) completed 3 sessions, 8 (11%) completed 4 sessions, 5 (6.8%) completed 5 sessions, 4 (5.5%) completed 6 sessions, 11 (15.1%) completed 7 sessions, and 26 (35.6%) completed all 8 sessions (Table S5 in Multimedia Appendix 1). The mean number of sessions completed in the full intervention sample was 5.3 (SD 2.9), and the median number of sessions completed was 7 (IQR 3 - 8).

We then conducted exploratory analyses to examine whether greater intervention exposure was associated with better clinical outcomes. Among participants who completed all follow-up assessments, those who completed ≥4 sessions showed significantly greater anxiety reduction than those completing <4 sessions (HAMA: $F_{1,47}$=5.13; *P*=.03), but no significant difference was observed for depressive symptoms (HAMD-21: $F_{1,47}$=2.86; *P*=.10; Table 1). In addition, compared with the TAU group, participants in the iSM+TAU subgroup who did not complete the intervention showed no significant differences in either anxiety symptoms (HAMA: $F_{1,62}$=0.31; *P*=.58) or depressive symptoms (HAMD-21: $F_{1,62}$=0.01; *P*=.91; Table 1), suggesting that limited intervention exposure may have attenuated the potential treatment benefit.

Table . Comparison of primary outcomes according to intervention completion status among participants with generalized anxiety disorder (GAD) who completed the 3-month follow-up assessment (T3) at Tongji Hospital (n=95). Observed posttreatment scores are presented as mean (SD). Adjusted scores are presented as adjusted mean (SE), estimated using analysis of covariance for each pairwise comparison, controlling for age, sex, BMI, education level, marital status, and employment status.

Outcome measure	Completed versus noncompleted		Noncompleted versus TAU[a]	
	Completed (n=26)	Noncompleted (n=28)	Noncompleted (n=28)	TAU (n=41)
HAMA[b]				
Observed mean (SD)	3.46 (2.52)	5.61 (3.61)	5.61 (3.61)	6.05 (3.81)
Adjusted mean (SE)	3.56 (0.60)	5.51 (0.58)	5.53 (0.75)	6.10 (0.61)
F test (*df*)	5.13 (1, 47)	—[c]	0.31 (1, 62)	—
P value	.03	—	.58	—
HAMD-21[d]				
Observed mean (SD)	4.12 (2.93)	6.00 (3.39)	6.00 (3.39)	6.41 (4.49)
Adjusted mean (SE)	4.38 (0.56)	5.75 (0.54)	6.18 (0.81)	6.29 (0.66)
F test (*df*)	2.86 (1, 47)	—	0.01 (1, 62)	—
P value	.10	—	.91	—

[a]TAU: treatment as usual.

[b] HAMA: Hamilton Anxiety Rating Scale.

[c]Not applicable.

[d]HAMD-21: 21-item Hamilton Depression Rating Scale.

Predictors of Treatment Response in the iSM+TAU Group

Among iSM completers (n=54), several adjusted regression models were used to examine baseline predictors of anxiety reduction (HAMA change from baseline to posttreatment). After univariate screening (*P*<.05) and adjustment for covariates including baseline HAMA scores, age, sex, marital status, education level, residential status, employment status, and childhood parental separation, separate covariate-adjusted linear regression models showed that higher trait anxiety was significantly associated with smaller reductions in anxiety symptoms (standardized β=–0.150, 95% CI –0.234 to –0.002; *P*=.046), while greater acting with awareness predicted larger improvements (standardized β=0.167, 95% CI 0.031-0.335; *P*=.02; Table 2). These findings suggest that individuals with lower trait anxiety and higher baseline mindful awareness may

XSL•FO
RenderX

be more suitable for self-guided digital mindfulness interventions.

Table . Multiple linear regression analysis of baseline psychological traits predicting Hamilton Anxiety Rating Scale (HAMA) reduction among participants with generalized anxiety disorder (GAD) in the internet-based stress management+treatment as usual (iSM+TAU) group who completed at least 4 intervention sessions at Tongji Hospital (n=54). Models were adjusted for age, sex, baseline HAMA score, marital status, educational attainment, residential setting, employment status, and childhood parental separation.

Predictors	Standardized β coefficient (95% CI)	*P* value	ΔR^2[a]	Model R^2[b]
Trait anxiety (T-AI)[c]	–0.150 (–0.234 to –0.002)	.046	0.013	0.872
Acting with awareness (FFMQ)[d]	0.167 (0.031 to 0.335)	.02	0.017	0.876
Symptom rumination (RRS)[e]	–0.099 (–0.281 to 0.075)	.25	0.004	0.863
Perceived helplessness (PSS-10)[f]	–0.040 (−0.316 to 0.189)	.61	0.001	0.860
Lack of self-efficacy (PSS-10)	–0.121 (–0.708 to 0.022)	.07	0.011	0.870
Certainty control (SQ)[g]	0.044 (–0.143 to 0.254)	.57	0.001	0.860

[a]ΔR^2 indicates the incremental variance explained by each predictor.

[b]Model R^2 represents the total variance explained by each separately fitted adjusted model.

[c]T-AI: Trait Anxiety Inventory.

[d]FFMQ: Five Facet Mindfulness Questionnaire.

[e]RRS: Ruminative Responses Scale.

[f]PSS-10: 10-item Perceived Stress Scale.

[g]SQ: Security Questionnaire.

CLPN Analysis

Across the 3 intervals (T0-T1, T1-T2, and T2-T3), the temporal network structure revealed distinct dynamic patterns among the 6 retained variables: HAMA, HAMD-21, FFMQ, PSQI, S-AI, and PSS (Figure 3). In these temporal networks, nodes represent psychological variables, and directed edges indicate temporal predictive associations from time *t* to time *t*+1. Red edges denote positive associations, whereas blue edges denote negative associations. Solid lines indicate edges with at least 90% bootstrap support, whereas dashed lines indicate edges with less than 90% bootstrap support.

XSL•FO
RenderX

Figure 3. Temporal network dynamics and random-intercept cross-lagged panel models of psychological variables across 4 assessment time points in participants with generalized anxiety disorder (GAD) in this randomized controlled trial at Tongji Hospital. Panels show temporal cross-lagged panel networks from (A) T0 to T1, (B) T1 to T2, and (C) T2 to T3, and constrained random-intercept cross-lagged panel models for (D) Model 1, examining perceived stress and state anxiety, and (E) Model 2, examining mindfulness and state anxiety. FFMQ: Five Facet Mindfulness Questionnaire; GAD: generalized anxiety disorder; HAMA: Hamilton Anxiety Rating Scale; HAMD: Hamilton Depression Rating Scale; PSQI: Pittsburgh Sleep Quality Index; PSS: Perceived Stress Scale; S-AI: State Anxiety Inventory; T0: baseline; T1: midintervention; T2: postintervention; T3: 3-month follow-up.

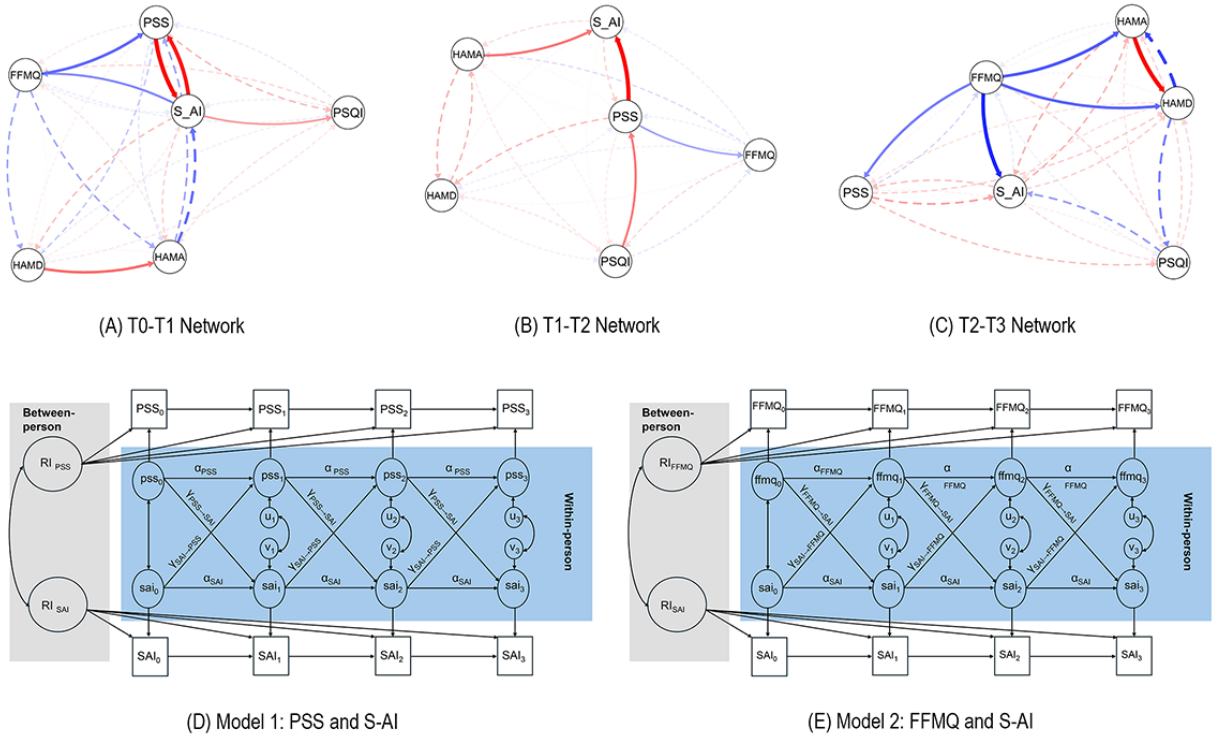

During the early treatment phase (T0-T1; Figure 3A), the strongest cross-lagged associations were observed between perceived stress and state anxiety, with PSS predicting subsequent S-AI (B=0.389, 95% CI 0.066-0.672) and S-AI, in turn, predicting PSS (B=0.351, 95% CI 0.084-0.570), suggesting a reciprocal temporal association between perceived stress and state anxiety. Higher baseline mindfulness (FFMQ) predicted lower subsequent perceived stress (B=−0.226, 95% CI −0.419 to −0.017), whereas higher state anxiety predicted reductions in mindfulness (B=−0.182, 95% CI −0.352 to −0.015). In addition, depressive symptoms predicted higher subsequent anxiety (HAMD → HAMA: B=0.237, 95% CI 0.058-0.501), and elevated state anxiety was associated with poorer subsequent sleep quality (S-AI → PSQI: B=0.164, 95% CI 0.093-0.323). In this interval, PSS showed the highest in-predictability, whereas S-AI exhibited the highest out-predictability (Figure S3 in Multimedia Appendix 1).

During midtreatment (T1-T2; Figure 3B), the PSS → S-AI pathway remained the strongest cross-lagged edge and further strengthened (B=0.442, 95% CI 0.214-0.621). Anxiety symptoms (HAMA) also showed a positive predictive effect on subsequent state anxiety (B=0.227, 95% CI 0.017-0.472). In parallel, poorer sleep quality predicted higher perceived stress (PSQI → PSS: B=0.205, 95% CI 0.003-0.408), while perceived stress predicted lower mindfulness (PSS → FFMQ: B=−0.151, 95% CI −0.322 to −0.044). State anxiety demonstrated the highest in- and out-predictability during this phase (Figure S3 in Multimedia Appendix 1).

During the consolidation phase (T2-T3; Figure 3C), the strongest cross-lagged association emerged between anxiety and depressive symptoms, with HAMA predicting higher subsequent HAMD-21 scores (B=0.311, 95% CI 0.086-0.481). Notably, mindfulness was negatively associated with subsequent state anxiety (B=−0.270, 95% CI −0.397 to −0.088), anxiety (B=−0.202, 95% CI −0.398 to −0.004), depression (B=−0.202, 95% CI −0.341 to −0.018), and perceived stress (B=−0.150, 95% CI −0.352 to −0.018). In this interval, depressive symptoms (HAMD-21) exhibited the highest in-predictability and out-predictability, suggesting that depressive symptoms were more strongly connected with subsequent symptom changes during the late phase (Figure S3 in Multimedia Appendix 1). Together, these findings suggest that higher mindfulness was associated with lower subsequent symptoms across multiple domains. Over time, the stress-anxiety pattern appeared to weaken, whereas the anxiety-depression linkage became more pronounced in the late phase.

The correlation stability coefficients for edge weights were 0.284 for T0-T1, 0.358 for T1-T2, and 0.358 for T2-T3, all exceeding the recommended minimum threshold of 0.25, indicating modest but acceptable stability of edge estimates across intervals. In contrast, correlation stability coefficients for in-strength and out-strength centrality were below 0.25 in most intervals, and these centrality metrics were therefore not interpreted.

XSL•FO
RenderX

RI-CLPM Analyses

To reduce the risk of spurious causal inferences arising from stable between-person differences in conventional cross-lagged models, RI-CLPM analyses were conducted to isolate within-person lagged associations. Random intercepts capture stable between-person differences (trait-like components) across time. Squares and uppercase letters indicate observed variables, whereas ovals and lowercase letters denote latent within-person components. Autoregressive (α) paths represent temporal stability within constructs, and cross-lagged (γ) paths represent directional within-person effects between constructs across adjacent time points. Residual terms (u_1 -u_3 and v_1 -v_3) represent time-specific unexplained within-person variance. All autoregressive and cross-lagged paths are constrained to be equal across time points, estimating lagged effects from time *t* to time *t*+1 across 4 measurement waves. Only 2 pathways showed significant within-person effects: the bidirectional pathway between PSS and S-AI (Figure 3D) and the unidirectional pathway from FFMQ to S-AI (Figure 3E). Both models demonstrated good fit with appropriate variance partitioning between stable traits and temporal fluctuations (see Multimedia Appendix 1).

Model 1: Perceived Stress and State Anxiety

Cross-lagged paths revealed robust bidirectional within-person associations (Figure 3D): greater perceived stress predicted higher subsequent state anxiety (b=0.363, 95% CI 0.163-0.563, *P*<.001; mean β=0.219), and higher state anxiety predicted greater subsequent stress (b=0.062, 95% CI 0.012-0.111, *P*=.02; mean β=0.165). Both effects exceeded the large effect size threshold (β≥0.12), supporting a reciprocal reinforcement mechanism at the individual level (Table 3).

Table . Regression parameter estimates from random-intercept cross-lagged panel models of perceived stress, mindfulness, and state anxiety among participants with generalized anxiety disorder (GAD) at Tongji Hospital.

Model and regression parameters	b[a] (95% CI)	SE	Standardized β (range)	*P* value
Model 1:[b] PSS[c] and S-AI[d]				
Autoregressive effects				
PSS → PSS at t+1	0.018 (–0.098 to 0.133)	0.059	0.017 to 0.031	.77
S-AI → S-AI at t+1	–0.129 (–0.272 to 0.013)	0.073	–0.160 to –0.142	.08
Cross-lagged effects				
PSS → S-AI at *t*+1	0.363 (0.163 to 0.563)	0.102	0.164 to 0.288	<.001
S-AI → PSS at *t*+1	0.062 (0.012 to 0.111)	0.025	0.148 to 0.194	.02
Model 2:[b] FFMQ and S-AI				
Autoregressive effects				
FFMQ[e] → FFMQ at t+1	0.362 (0.193 to 0.531)	0.086	0.498 to 0.567	<.001
S-AI → S-AI at t+1	–0.124 (–0.266 to 0.019)	0.073	–0.160 to –0.129	.09
Cross-lagged effects				
FFMQ → S-AI at t+1	–0.210 (–0.326 to –0.093)	0.059	–0.312 to –0.262	<.001
S-AI → FFMQ at t+1	–0.067 (–0.164 to 0.031)	0.050	–0.108 to –0.064	.18

[a]b: unstandardized regression coefficient (constrained across time); β range, range of standardized coefficients across time points.

[b]Model 1 examined perceived stress and state anxiety. Model 2 examined mindfulness and state anxiety. Both models included random intercepts to account for between-person differences.

[c]PSS: Perceived Stress Scale.

[d]S-AI: State Anxiety Inventory.

[e]FFMQ: Five Facet Mindfulness Questionnaire.

Model 2: Mindfulness and State Anxiety

A significant unidirectional within-person effect was observed (Figure 3E): higher mindfulness predicted lower subsequent state anxiety (b=−0.210, 95% CI −0.326 to −0.093; *P*<.001; mean β=−0.285), whereas state anxiety did not significantly predict later mindfulness (b=−0.067, 95% CI −0.164 to 0.031; *P*=.18). These findings suggest that increases in mindfulness precede reductions in state anxiety over time (Table 3).

Discussion

Principal Findings

This study evaluated a brief, self-guided, internet-delivered mindfulness-informed stress management program as an adjunct to TAU for patients with GAD. Overall, the intervention was associated with improvement in anxiety, depression, and several functional outcomes, while also showing acceptable adherence and feasibility in a real-world clinical setting. In addition, the

XSL•FO
RenderX

longitudinal network-based analyses provided exploratory evidence suggesting that stress- and mindfulness-related symptom processes may be involved in how improvement unfolded over time.

The present study demonstrated significant anxiety reduction in the iSM+TAU group compared with TAU, with effects that were sustained and larger at the 3-month follow-up. This pattern of sustained and growing improvement suggests that skills acquired during training continue to consolidate after formal intervention ends [57]. Although the effect size was lower than that found in a comparable RCT with more intensive interventions in a therapist-guided online mindfulness intervention [55], it was larger than the small effects reported in a meta-analysis of online mindfulness-based interventions [22] and consistent with recent work suggesting that fully self-guided digital mindfulness approaches can be acceptable and beneficial for adults with GAD [58].

Beyond symptom reduction in anxiety and depression, iSM demonstrated broad functional improvements, including sleep quality, somatic symptoms, and social functioning. These multidimensional benefits are particularly important in real-world clinical settings where functional recovery often lags behind symptom improvement [59-62]. More broadly, mindfulness-based interventions may have effects across multiple domains rather than on emotional symptoms alone, with meta-analytic evidence suggesting benefits for global cognition and several cognitive control–related processes [63]. Although these domains were not directly examined in the present study, they may provide a broader framework for understanding symptom improvement. Notably, perceived stress, a central focus of the intervention, was significantly reduced by midtreatment, suggesting that stress reduction may be an early and clinically relevant component of improvement.

The retention rate of 77% (108/140) and median completion of 7 sessions demonstrate excellent feasibility for a self-directed intervention. Participant retention in the present trial was comparable to that reported by Li et al [64] and notably higher than rates documented in other large-scale RCTs of online mindfulness-based interventions, where attrition has reached approximately 38.7% [65]. Importantly, most dropouts occurred before or during the initial sessions, suggesting that early engagement is critical for retention [66]. The relatively high adherence to this self-guided format underscores the feasibility of iSM, particularly in settings with limited therapeutic resources [20,21]. Treatment adherence was closely associated with outcomes; participants who completed more sessions showed greater anxiety reduction, whereas those with incomplete participation derived little benefit beyond TAU. Dose-response analyses further indicated that completing at least 4 sessions, covering core mindfulness components, was associated with greater improvement and may represent a pragmatic minimum exposure target in self-guided settings [67].

Individual psychological characteristics appeared to significantly influence the effectiveness of the intervention. Higher baseline acting with awareness—a mindfulness facet reflecting the tendency to be attentive to present-moment activities—predicted greater improvements, consistent with previous research suggesting that existing mindfulness capacity facilitates stress management skill acquisition [68]. This may reflect better ability to engage with training exercises or greater openness to experiential approaches. Conversely, elevated trait anxiety was associated with attenuated benefits, potentially reflecting impaired attentional control that limits effective engagement with stress management techniques [69]. Although these predictor effects were modest, they may still be informative as partial contributors to treatment response variability. From a clinical perspective, these findings suggest that individuals with higher baseline mindfulness capacity may be well-suited for self-guided digital programs as adjuncts to standard care, whereas those with very elevated trait anxiety may require more intensive or therapist-supported interventions. Such stratification could maximize both clinical outcomes and health care efficiency.

While traditional RCT analyses established the overall efficacy of the intervention, network analyses were used to explore how symptom associations changed over time and to generate hypotheses about possible temporal processes underlying change. CLPN analyses suggested stage-dependent changes in the pattern of temporal symptom associations across the study period. In the initial phase, perceived stress and anxiety showed a relatively prominent bidirectional association, a pattern broadly consistent with network theories emphasizing persistent interconnections among symptoms despite reductions in overall symptom levels [70]. This pattern appeared to weaken over time, whereas anxiety remained highly connected and stress and sleep continued to show temporal associations with other variables. In later phases, depressive symptoms appeared more strongly embedded in the temporal network and were more often linked to subsequent changes in other variables. Meanwhile, higher mindfulness was generally associated with lower subsequent levels of stress, anxiety, and depression across phases. These findings are broadly consistent with nonlinear models of psychological change, which propose that symptom improvement may occur through stage-dependent reorganization [71].

However, not all temporal associations identified in the CLPN analyses remained evident under the RI-CLPM framework. RI-CLPM, which more explicitly accounts for stable trait-like between-person differences, suggested that 2 pathways remained relatively robust under this more stringent model and were broadly compatible with within-person temporal dynamics. First, the stress-anxiety relationship was bidirectional: perceived stress predicted subsequent anxiety, and anxiety predicted subsequent stress, a pattern broadly consistent with reciprocal processes described in transactional stress models [72]. Second, mindfulness showed a unidirectional protective effect on anxiety, whereas the reverse path was not supported in the model, suggesting that mindfulness may function as a protective regulatory process, although this interpretation remains inferential and model-based, consistent with attentional control theory [73]. Other associations identified in the CLPN analyses were not retained in the RI-CLPM framework, suggesting that they may be less robust as within-person temporal pathways. Taken together, these findings are broadly consistent with theoretical models emphasizing stress reactivity, attentional

XSL•FO
RenderX

regulation, and stage-dependent symptom change and suggest that mindfulness-related processes may contribute to symptom improvement over time [31,72,73].

The present study provides new evidence for a brief, self-guided, culturally adapted internet-based mindfulness-informed stress management program evaluated as a feasible and clinically useful adjunct to pharmacological treatment for GAD in a Chinese clinical setting. Unlike prior work focused mainly on symptom outcomes alone, it also incorporated exploratory longitudinal analyses to provide a more process-oriented perspective on how symptoms may change over time, suggesting that perceived stress and mindfulness may represent potential processes involved in treatment response and warrant further investigation. The iSM may also help expand access to mental health care in settings with limited psychotherapy resources. As a brief digital program requiring only basic internet access and no therapist scheduling, iSM may offer a practical and relatively low-cost option for wider implementation, especially within stepped-care models and among underserved populations [74,75]. The integration of culturally relevant elements further enhances engagement and acceptability, supporting iSM's potential as an equitable adjunct to traditional services within evolving mental health systems.

Limitations

This study has several limitations. First, although participants were recruited across 3 hospital campuses, this study was conducted within a single hospital system with a relatively homogeneous sample, limiting generalizability. The sample profile may partly reflect self-selection into a self-guided digital intervention, which may be more readily adopted by younger and more educated participants. These factors highlight the need for future adaptations and testing in broader populations. In addition, the achieved sample size was slightly below the pre-registered recruitment target due to funding and time constraints. Future studies should expand the sample size and include multiple centers with more diverse populations across different regions, while also considering adaptations that may improve accessibility for broader clinical groups and enhance external validity. Additionally, the CLPN and RI-CLPM analyses were exploratory and should be interpreted cautiously. Future studies with larger samples are needed to verify the robustness and reproducibility of these longitudinal findings. Second, the follow-up period was limited to 3 months after the intervention, making it difficult to evaluate long-term effectiveness and adherence. Future research should extend the observation period and incorporate strategies such as booster sessions and individualized guidance to support sustained engagement and outcomes. Third, this study primarily relied on standardized scales for assessment, which may be subject to expectancy effects. Future research should incorporate objective physiological and neuroimaging measures to improve the precision of mechanism evaluation. Fourth, the network analysis was restricted to 6 theory-driven variables capturing the hypothesized stress-management pathway and key GAD-related symptom domains, which helped preserve model interpretability and stability given the sample size and repeated-measures design, but other potentially relevant mechanisms (eg, emotion regulation and self-efficacy) were not examined. Future studies with larger samples and broader assessments should investigate more comprehensive mechanistic networks. Fifth, the use of TAU rather than an active control also limits the ability to distinguish intervention-specific effects from nonspecific influences such as attention, expectancy, or engagement with a digital platform, although this design was consistent with the pragmatic aim of evaluating added benefit beyond usual care. Finally, while RI-CLPM strengthens within-person temporal inference by controlling for stable individual differences, causal conclusions still require experimental manipulation of specific mechanisms.

Conclusion

To our knowledge, this is the first RCT to evaluate a brief, self-guided, culturally adapted digital mindfulness intervention as an adjunct to pharmacotherapy in Chinese adults with GAD. The intervention yielded clinically meaningful benefits across anxiety, depression, and several secondary outcomes related to stress, mindfulness, and associated clinical domains. Its fully self-guided format, culturally adapted content, and minimal infrastructure requirements highlight its practical potential as a scalable, resource-efficient approach in real-world settings where psychotherapy availability is limited. Unlike prior studies focused mainly on symptom outcomes alone, this trial combined a randomized design with exploratory CLPN and RI-CLPM analyses to provide preliminary insight into symptom change processes over time. The longitudinal findings suggested temporally distinct dynamics, with stress-anxiety coupling more prominent early in treatment and mindfulness-related pathways becoming more apparent in later phases. These findings contribute not only preliminary evidence of adjunctive clinical utility but also add a more process-oriented analytic perspective that may inform future intervention refinement.

Acknowledgments

The authors would like to thank all participants for their time and commitment to this study. We also gratefully acknowledge the engineers and technical staff who contributed to the development and maintenance of the online intervention platform. The authors declare the use of generative artificial intelligence (AI) in the research and writing process. According to the GAIDeT taxonomy (2025) [76], the following tasks were delegated to generative AI tools under full human supervision: proofreading and editing. The generative AI tool used was ChatGPT (OpenAI). Responsibility for the final manuscript lies entirely with the authors. Generative AI tools are not listed as authors and do not bear responsibility for the final outcomes. This declaration is submitted on behalf of all authors (collective responsibility).

XSL•FO
RenderX

Funding

This work was supported by the National Natural Science Foundation of China (Grant No.82090034). The funding agency had no role in the design, conduct, analysis, or reporting of the trial.

Data Availability

Deidentified participant data underlying the findings of this study, along with the data dictionary and statistical analysis code, are available from the corresponding author upon reasonable request. Access to the data will be granted following review and approval of a methodologically sound proposal and completion of a data use agreement, in accordance with institutional and ethical regulations.

Authors' Contributions

Study conception and design: YY, HZ, ZW
Intervention development: HQT
Intervention delivery and standardization: JY
Participant recruitment, randomization, and enrollment: ZW, SW, YX
Patient management, follow-up, and adherence monitoring: ZW, SW, YX
Data management and curation: YX, SW
Statistical analysis: ZW
Project administration and coordination: HZ, YY
Resources: YY
Funding acquisition: YY
Supervision: YY, HZ
Manuscript writing: ZW drafted the original manuscript; HZ, YY revised the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
This multimedia appendix provides detailed supplementary methods, including a comprehensive description of the internet-based mindfulness stress management intervention, full descriptions of all psychological measurement instruments and scoring procedures, and detailed analytic methods for the cross-lagged panel network and random-intercept cross-lagged panel modeling analyses. Supplementary figures and tables supporting the main results are also included.
[DOCX File, 1029 KB - jmir_v28i1e91751_app1.docx]

Checklist 1
CONSORT 2025 reporting checklist.
[PDF File, 278 KB - jmir_v28i1e91751_app2.pdf]

Checklist 2
CONSORT-EHEALTH reporting checklist.
[PDF File, 1310 KB - jmir_v28i1e91751_app3.pdf]

References

1. Ruscio AM, Hallion LS, Lim CCW, et al. Cross-sectional comparison of the epidemiology of DSM-5 generalized anxiety disorder across the globe. JAMA Psychiatry 2017 May 1;74(5):465-475. [doi: 10.1001/jamapsychiatry.2017.0056] [Medline: 28297020]
2. Newman MG, Llera SJ, Erickson TM, Przeworski A, Castonguay LG. Worry and generalized anxiety disorder: a review and theoretical synthesis of evidence on nature, etiology, mechanisms, and treatment. Annu Rev Clin Psychol 2013;9(1):275-297. [doi: 10.1146/annurev-clinpsy-050212-185544] [Medline: 23537486]
3. Patriquin MA, Mathew SJ. The neurobiological mechanisms of generalized anxiety disorder and chronic stress. Chronic Stress (Thousand Oaks) 2017;1:2470547017703993. [doi: 10.1177/2470547017703993] [Medline: 29503978]
4. Gkintoni E, Ortiz PS. Neuropsychology of generalized anxiety disorder in clinical setting: a systematic evaluation. Healthcare (Basel) 2023 Aug 31;11(17):2446. [doi: 10.3390/healthcare11172446] [Medline: 37685479]
5. Daviu N, Bruchas MR, Moghaddam B, Sandi C, Beyeler A. Neurobiological links between stress and anxiety. Neurobiol Stress 2019 Nov;11:100191. [doi: 10.1016/j.ynstr.2019.100191] [Medline: 31467945]

XSL•FO
RenderX

6. Curtiss JE, Levine DS, Ander I, Baker AW. Cognitive-behavioral treatments for anxiety and stress-related disorders. Focus (Am Psychiatr Publ) 2021 Jun;19(2):184-189. [doi: 10.1176/appi.focus.20200045] [Medline: 34690581]
7. Bandelow B, Boerner J R, Kasper S, Linden M, Wittchen HU, Möller HJ. The diagnosis and treatment of generalized anxiety disorder. Dtsch Arztebl Int 2013 Apr;110(17):300-309. [doi: 10.3238/arztebl.2013.0300] [Medline: 23671484]
8. Moncrieff J. Persistent adverse effects of antidepressants. Epidemiol Psychiatr Sci 2019 Sep 23;29:e56. [doi: 10.1017/S2045796019000520] [Medline: 31543093]
9. Niarchou E, Roberts LH, Naughton BD. What is the impact of antidepressant side effects on medication adherence among adult patients diagnosed with depressive disorder: a systematic review. J Psychopharmacol 2024 Feb;38(2):127-136. [doi: 10.1177/02698811231224171] [Medline: 38344912]
10. Lewis G, Marston L, Duffy L, et al. Maintenance or discontinuation of antidepressants in primary care. N Engl J Med 2021 Sep 30;385(14):1257-1267. [doi: 10.1056/NEJMoa2106356] [Medline: 34587384]
11. Baker AW, Skolnik AM, Park JM, Sprich SE, Wilhelm S. Basic principles and practice of cognitive-behavioral therapy. In: Sprich SE, Petersen T, Wilhelm S, editors. The Massachusetts General Hospital Handbook of Cognitive Behavioral Therapy: Springer International Publishing; 2023:7-17. [doi: 10.1007/978-3-031-29368-9_2]
12. Wolitzky-Taylor K, Fenwick K, Lengnick-Hall R, et al. A preliminary exploration of the barriers to delivering (and receiving) exposure-based cognitive behavioral therapy for anxiety disorders in adult community mental health settings. Community Ment Health J 2018 Oct;54(7):899-911. [doi: 10.1007/s10597-018-0252-x] [Medline: 29524078]
13. Man J, Yan R, Yang K, et al. Cognitive behavioral therapy in China: practices and exploration. J Cogn Ther 2024;17(2):231-250. [doi: 10.1007/s41811-024-00203-6]
14. Hofmann SG, Sawyer AT, Witt AA, Oh D. The effect of mindfulness-based therapy on anxiety and depression: a meta-analytic review. J Consult Clin Psychol 2010 Apr;78(2):169-183. [doi: 10.1037/a0018555] [Medline: 20350028]
15. Kabat-Zinn J. Mindfulness-based interventions in context: past, present, and future. Clin Psychol Sci Pract 2003;10(2):144-156. [doi: 10.1093/clipsy.bpg016]
16. Reangsing C, Trakooltorwong P, Maneekunwong K, Thepsaw J, Oerther S. Effects of online mindfulness-based interventions (MBIs) on anxiety symptoms in adults: a systematic review and meta-analysis. BMC Complement Med Ther 2023 Jul 28;23(1):269. [doi: 10.1186/s12906-023-04102-9] [Medline: 37507747]
17. Blanck P, Perleth S, Heidenreich T, et al. Effects of mindfulness exercises as stand-alone intervention on symptoms of anxiety and depression: systematic review and meta-analysis. Behav Res Ther 2018 Mar;102:25-35. [doi: 10.1016/j.brat.2017.12.002] [Medline: 29291584]
18. Binda DD, Greco CM, Morone NE. What are adverse events in mindfulness meditation? Glob Adv Health Med 2022;11:2164957X221096640. [doi: 10.1177/2164957X221096640] [Medline: 35464906]
19. Norsworthy KL. Mindful activism: Embracing the complexities of international border crossings. Am Psychol 2017 Dec;72(9):1035-1043. [doi: 10.1037/amp0000262] [Medline: 29283668]
20. Que J, Lu L, Shi L. Development and challenges of mental health in China. Gen Psychiatr 2019;32(1):e100053. [doi: 10.1136/gpsych-2019-100053] [Medline: 31179426]
21. Shi C, Ma N, Wang L, et al. Study of the mental health resources in China. Chin J health policy 2019;12(2):51-57 [FREE Full text] [doi: 10.3969/j.issn.1674-2982.2019.02.008]
22. Spijkerman MPJ, Pots WTM, Bohlmeijer ET. Effectiveness of online mindfulness-based interventions in improving mental health: a review and meta-analysis of randomised controlled trials. Clin Psychol Rev 2016 Apr;45:102-114. [doi: 10.1016/j.cpr.2016.03.009] [Medline: 27111302]
23. Mrazek AJ, Mrazek MD, Cherolini CM, et al. The future of mindfulness training is digital, and the future is now. Curr Opin Psychol 2019 Aug;28:81-86. [doi: 10.1016/j.copsyc.2018.11.012] [Medline: 30529975]
24. Roemer L, Salters K, Raffa SD, Orsillo SM. Fear and avoidance of internal experiences in GAD: preliminary tests of a conceptual model. Cogn Ther Res 2005 Feb;29(1):71-88. [doi: 10.1007/s10608-005-1650-2]
25. Taylor H, Strauss C, Cavanagh K. Can a little bit of mindfulness do you good? A systematic review and meta-analyses of unguided mindfulness-based self-help interventions. Clin Psychol Rev 2021 Nov;89:102078. [doi: 10.1016/j.cpr.2021.102078] [Medline: 34537665]
26. Dahlin M, Andersson G, Magnusson K, et al. Internet-delivered acceptance-based behaviour therapy for generalized anxiety disorder: a randomized controlled trial. Behav Res Ther 2016 Feb;77:86-95. [doi: 10.1016/j.brat.2015.12.007] [Medline: 26731173]
27. Boettcher J, Aström V, Påhlsson D, Schenström O, Andersson G, Carlbring P. Internet-based mindfulness treatment for anxiety disorders: a randomized controlled trial. Behav Ther 2014 Mar;45(2):241-253. [doi: 10.1016/j.beth.2013.11.003] [Medline: 24491199]
28. Kladnitski N, Smith J, Allen A, Andrews G, Newby JM. Online mindfulness-enhanced cognitive behavioural therapy for anxiety and depression: outcomes of a pilot trial. Internet Interv 2018 Sep;13:41-50. [doi: 10.1016/j.invent.2018.06.003] [Medline: 30206518]
29. Pan Y, Li F, Liang H, et al. Effectiveness of mindfulness-based stress reduction on mental health and psychological quality of life among university students: a GRADE-assessed systematic review. Evid Based Complement Alternat Med 2024;2024:8872685. [doi: 10.1155/2024/8872685] [Medline: 38414520]

XSL•FO
RenderX

30. Tang R, Braver TS. Towards an individual differences perspective in mindfulness training research: theoretical and empirical considerations. Front Psychol 2020;11:818. [doi: 10.3389/fpsyg.2020.00818] [Medline: 32508702]
31. Pruessner L, Timm C, Kalmar J, Bents H, Barnow S, Mander J. Emotion regulation as a mechanism of mindfulness in individual cognitive-behavioral therapy for depression and anxiety disorders. Depress Anxiety 2024;2024:9081139. [doi: 10.1155/2024/9081139] [Medline: 40226724]
32. Epskamp S, van Borkulo CD, van der Veen DC, et al. Personalized network modeling in psychopathology: the importance of contemporaneous and temporal connections. Clin Psychol Sci 2018 May;6(3):416-427. [doi: 10.1177/2167702617744325] [Medline: 29805918]
33. Borsboom D, Deserno MK, Rhemtulla M, et al. Network analysis of multivariate data in psychological science. Nat Rev Methods Primers 2021;1(1):58. [doi: 10.1038/s43586-021-00055-w]
34. Wysocki A, McCarthy I, van Bork R, Cramer AOJ, Rhemtulla M. Cross-lagged panel networks. advin/psych 2025;2(1):e739621. [doi: 10.56296/aip00037]
35. Bernstein EE, Kleiman EM, van Bork R, et al. Unique and predictive relationships between components of cognitive vulnerability and symptoms of depression. Depress Anxiety 2019 Oct;36(10):950-959. [doi: 10.1002/da.22935] [Medline: 31332887]
36. Hamaker EL, Kuiper RM, Grasman R. A critique of the cross-lagged panel model. Psychol Methods 2015 Mar;20(1):102-116. [doi: 10.1037/a0038889] [Medline: 25822208]
37. Mulder JD, Hamaker EL. Three extensions of the random intercept cross-lagged panel model. Struct Equ Modeling 2021 Jul 4;28(4):638-648. [doi: 10.1080/10705511.2020.1784738]
38. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ 2025 Apr 14;389:e081123. [doi: 10.1136/bmj-2024-081123] [Medline: 40228833]
39. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
40. Kabat-Zinn J. An outpatient program in behavioral medicine for chronic pain patients based on the practice of mindfulness meditation: theoretical considerations and preliminary results. Gen Hosp Psychiatry 1982 Apr;4(1):33-47. [doi: 10.1016/0163-8343(82)90026-3] [Medline: 7042457]
41. HAMILTON M. The assessment of anxiety states by rating. Br J Med Psychol 1959;32(1):50-55. [doi: 10.1111/j.2044-8341.1959.tb00467.x] [Medline: 13638508]
42. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry 1960 Feb;23(1):56-62. [doi: 10.1136/jnnp.23.1.56] [Medline: 14399272]
43. Spielberger CD, Gonzalez-Reigosa F, Martinez-Urrutia A, Natalicio LF, Natalicio DS. The state-trait anxiety inventory. Interam J Psychol 1971;5 [FREE Full text] [doi: 10.30849/rip/ijp.v5i3&4.620]
44. Buysse DJ, Reynolds CF III, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. Psychiatry Res 1989 May;28(2):193-213. [doi: 10.1016/0165-1781(89)90047-4] [Medline: 2748771]
45. Deng YQ, Liu XH, Rodriguez MA, Xia CY. The Five Facet Mindfulness Questionnaire: psychometric properties of the Chinese version. Mindfulness (N Y) 2011 Jun;2(2):123-128. [doi: 10.1007/s12671-011-0050-9]
46. Kroenke K, Spitzer RL, Williams JBW. The PHQ-15: validity of a new measure for evaluating the severity of somatic symptoms. Psychosom Med 2002;64(2):258-266. [doi: 10.1097/00006842-200203000-00008] [Medline: 11914441]
47. WHO psychiatric disability assessment schedule (WHO/DAS:with a guide to its use. World Health Organization. 1988. URL: https://iris.who.int/items/17e147a5-c46f-4dd1-b3aa-617bd6337a26 [accessed 2026-05-27]
48. Treynor W, Gonzalez R, Nolen-Hoeksema S. Rumination reconsidered: a psychometric analysis. Cognit Ther Res 2003 Jun;27(3):247-259. [doi: 10.1023/A:1023910315561]
49. Lu W, Bian Q, Wang W, Wu X, Wang Z, Zhao M. Chinese version of the Perceived Stress Scale-10: a psychometric study in Chinese university students. PLoS ONE 2017;12(12):e0189543. [doi: 10.1371/journal.pone.0189543] [Medline: 29252989]
50. Zhong C, Lijuan A. Developing of security questionnaire and its reliability and validity. Chin Ment Health J 2004;18(2):97-99 [FREE Full text] [doi: 10.3321/j.issn:1000-6729.2004.02.010]
51. Zimet GD, Powell SS, Farley GK, Werkman S, Berkoff KA. Psychometric characteristics of the multidimensional scale of perceived social support. J Pers Assess 1990;55(3-4):610-617. [doi: 10.1080/00223891.1990.9674095] [Medline: 2280326]
52. Russell DW. UCLA Loneliness Scale (Version 3): reliability, validity, and factor structure. J Pers Assess 1996 Feb;66(1):20-40. [doi: 10.1207/s15327752jpa6601_2] [Medline: 8576833]
53. Rotter JB. A new scale for the measurement of interpersonal trust. J Pers 1967 Dec;35(4):651-665. [doi: 10.1111/j.1467-6494.1967.tb01454.x] [Medline: 4865583]
54. Watson D, Friend R. Measurement of social-evaluative anxiety. J Consult Clin Psychol 1969 Aug;33(4):448-457. [doi: 10.1037/h0027806] [Medline: 5810590]

XSL•FO
RenderX

55. Roy A, Hoge EA, Abrante P, Druker S, Liu T, Brewer JA. Clinical efficacy and psychological mechanisms of an app-based digital therapeutic for generalized anxiety disorder: randomized controlled trial. J Med Internet Res 2021 Dec 2;23(12):e26987. [doi: 10.2196/26987] [Medline: 34860673]
56. Little RJA. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 1988 Dec;83(404):1198-1202. [doi: 10.1080/01621459.1988.10478722]
57. Yavuz Sercekman M. Exploring the sustained impact of the mindfulness-based stress reduction program: a thematic analysis. Front Psychol 2024;15:1347336. [doi: 10.3389/fpsyg.2024.1347336] [Medline: 39100567]
58. Zainal NH, Newman MG. Which client with generalized anxiety disorder benefits from a mindfulness ecological momentary intervention versus a self-monitoring app? Developing a multivariable machine learning predictive model. J Anxiety Disord 2024 Mar;102:102825. [doi: 10.1016/j.janxdis.2024.102825] [Medline: 38245961]
59. Simms LJ, Prisciandaro JJ, Krueger RF, Goldberg DP. The structure of depression, anxiety and somatic symptoms in primary care. Psychol Med 2012 Jan;42(1):15-28. [doi: 10.1017/S0033291711000985] [Medline: 21682948]
60. Kyle SD, Morgan K, Espie CA. Insomnia and health-related quality of life. Sleep Med Rev 2010 Feb;14(1):69-82. [doi: 10.1016/j.smrv.2009.07.004] [Medline: 19962922]
61. Liao SC, Ma HM, Lin YL, Huang WL. Functioning and quality of life in patients with somatic symptom disorder: the association with comorbid depression. Compr Psychiatry 2019 Apr;90:88-94. [doi: 10.1016/j.comppsych.2019.02.004] [Medline: 30818088]
62. Sarsour K, Kalsekar A, Swindle R, Foley K, Walsh JK. The association between insomnia severity and healthcare and productivity costs in a health plan sample. Sleep 2011 Apr 1;34(4):443-450. [doi: 10.1093/sleep/34.4.443] [Medline: 21461322]
63. Zainal NH, Newman MG. Mindfulness enhances cognitive functioning: a meta-analysis of 111 randomized controlled trials. Health Psychol Rev 2024 Jun;18(2):369-395. [doi: 10.1080/17437199.2023.2248222] [Medline: 37578065]
64. Li Y, Zhang Y, Wang C, et al. Supported mindfulness-based self-help intervention as an adjunctive treatment for rapid symptom change in emotional disorders: a practice-oriented multicenter randomized controlled trial. Psychother Psychosom 2025;94(2):119-129. [doi: 10.1159/000542937] [Medline: 39809242]
65. Linardon J. Rates of attrition and engagement in randomized controlled trials of mindfulness apps: systematic review and meta-analysis. Behav Res Ther 2023 Nov;170:104421. [doi: 10.1016/j.brat.2023.104421] [Medline: 37862854]
66. Winter N, Russell L, Ugalde A, White V, Livingston P. Engagement strategies to improve adherence and retention in web-based mindfulness programs: systematic review. J Med Internet Res 2022 Jan 12;24(1):e30026. [doi: 10.2196/30026] [Medline: 35019851]
67. Strohmaier S, Jones FW, Cane JE. Effects of length of mindfulness practice on mindfulness, depression, anxiety, and stress: a randomized controlled experiment. Mindfulness (N Y) 2021 Jan;12(1):198-214. [doi: 10.1007/s12671-020-01512-5]
68. Mitsea E, Drigas A, Skianis C. Digitally assisted mindfulness in training self-regulation skills for sustainable mental health: a systematic review. Behav Sci (Basel) 2023 Dec 10;13(12):1008. [doi: 10.3390/bs13121008] [Medline: 38131865]
69. Eysenck MW, Moser JS, Derakshan N, Hepsomali P, Allen P. A neurocognitive account of attentional control theory: how does trait anxiety affect the brain's attentional networks? Cogn Emot 2023 Mar;37(2):220-237. [doi: 10.1080/02699931.2022.2159936] [Medline: 36583855]
70. Borsboom D. A network theory of mental disorders. World Psychiatry 2017 Feb;16(1):5-13. [doi: 10.1002/wps.20375] [Medline: 28127906]
71. Hayes AM, Laurenceau JP, Feldman G, Strauss JL, Cardaciotto L. Change is not always linear: the study of nonlinear and discontinuous patterns of change in psychotherapy. Clin Psychol Rev 2007 Jul;27(6):715-723. [doi: 10.1016/j.cpr.2007.01.008] [Medline: 17316941]
72. Lazarus RS, Folkman S. Stress, Appraisal, and Coping: Springer Publishing Company; 1984:1-445 URL: https://search.worldcat.org/title/Stress-appraisal-and-coping/oclc/1164611259 [accessed 2026-05-27]
73. Eysenck MW, Derakshan N, Santos R, Calvo MG. Anxiety and cognitive performance: attentional control theory. Emotion 2007 May;7(2):336-353. [doi: 10.1037/1528-3542.7.2.336] [Medline: 17516812]
74. Luangapichart P, Saisavoey N, Viravan N. Efficacy and feasibility of the minimal therapist-guided four-week online audio-based mindfulness program "Mindful Senses" for burnout and stress reduction in medical personnel: a randomized controlled trial. Healthcare (Basel) 2022 Dec 14;10(12):2532. [doi: 10.3390/healthcare10122532] [Medline: 36554056]
75. Buntrock C. Cost-effectiveness of digital interventions for mental health: current evidence, common misconceptions, and future directions. Front Digit Health 2024;6:1486728. [doi: 10.3389/fdgth.2024.1486728] [Medline: 39498103]
76. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (Generative AI Delegation Taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res 2026 Apr;33(3):2544331. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]

Abbreviations

CBT: cognitive behavioral therapy
CFI: comparative fit index

XSL•FO
RenderX

CLPN: cross-lagged panel network
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
***DSM-5*:** *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition)
FFMQ: Five Facet Mindfulness Questionnaire
GAD: generalized anxiety disorder
HAMA: Hamilton Anxiety Rating Scale
HAMD-21: 21-item Hamilton Depression Rating Scale
iSM: internet-based stress management program
ITS: Interpersonal Trust Scale
ITT: intention-to-treat
LASSO: least absolute shrinkage and selection operator
MBSR: mindfulness-based stress reduction
MD: mean difference
PHQ-15: 15-item Patient Health Questionnaire
PP: per-protocol
PSQI: Pittsburgh Sleep Quality Index
PSS-10: 10-item Perceived Stress Scale
PSSS: Perceived Social Support Scale
RCT: randomized controlled trial
RI-CLPM: random-intercept cross-lagged panel model
RMSEA: root-mean-square error of approximation
RRS: Ruminative Responses Scale
S-AI: State Anxiety Inventory
SAD: Social Avoidance and Distress Scale
SDSS: Social Disability Screening Schedule
SQ: Security Questionnaire
SRMR: standardized root-mean-square residual
T-AI: Trait Anxiety Inventory
TAU: treatment as usual
UCLA-LS: UCLA Loneliness Scale

Edited by S Brini; submitted 19.Jan.2026; peer-reviewed by L Sun, NH Zainal; revised version received 21.Apr.2026; accepted 01.May.2026; published 05.Jun.2026.

Please cite as:
Wang Z, Tong HQ, Yue J, Wu S, Xia Y, Zhang H, Yang Y
Self-Guided Internet-Based Mindfulness-Informed Stress Management for Generalized Anxiety Disorder: Randomized Controlled Trial With Longitudinal Network Analysis
J Med Internet Res 2026;28:e91751
URL: https://www.jmir.org/2026/1/e91751
doi:10.2196/91751

© Ziwei Wang, Hui Qi Tong, Jianrong Yue, Siyan Wu, Ye Xia, Han Zhang, Yuan Yang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Social Media Intervention Based on the Information-Motivation-Behavioral Skills Model Promotes HIV Testing and Reduces High-Risk Behaviors Among Men Who Have Sex With Men in Resource-Limited Settings in China: Randomized Controlled Trial

Huanzhuo Mai[1,2,3*], MPH; Jianyuan Liu[1,2,3*], BSc; Liping Cheng[1,2,3*], BSc; Hao Liang[1,2,4], PhD; Ping Cui[2,4], PhD; Qin Yao[5], MSc; Qin Cao[1,2,3], BSc; Heng Zhang[1,2,3], BSc; Wenping Liao[1,2,3], BSc; Shiran He[1,2,3], BSc; Xing Yang[6*], PhD; Ping Cen[7*], PhD; Jiegang Huang[1,2,3*], PhD

[1]School of Public Health, Guangxi Medical University, 22, Shuangyong Road, Nanning, Guangxi, China

[2]Guangxi Key Laboratory of AIDS Prevention and Treatment, Guangxi Medical University, Nanning, Guangxi, China

[3]Key Laboratory of Prevention and Control of Highly Prevalent Diseases (Guangxi Medical University), Education Department of Guangxi Zhuang Autonomous Region, Nanning, Guangxi, China

[4]Life Sciences Institute, Guangxi Medical University, Nanning, Guangxi, China

[5]Nursing College, Guangxi Medical University, Nanning, Guangxi, China

[6]Health Management Research Institute, People's Hospital of Guangxi Zhuang Autonomous Region and Guangxi Academy of Medical Sciences, Nanning, Guangxi, China

[7]Infectious Disease Laboratory, Guangxi AIDS Clinical Treatment Centre (Nanning), The Fourth People's Hospital of Nanning, Nanning, Guangxi, China

[*]these authors contributed equally

Corresponding Author:
Jiegang Huang, PhD
School of Public Health, Guangxi Medical University, 22, Shuangyong Road, Nanning, Guangxi, China

Abstract

Background: Social media intervention may enhance HIV prevention among men who have sex with men, but the effect of this intervention in resource-limited settings remains unclear.

Objective: This randomized controlled trial evaluated whether a social media intervention grounded in the information-motivation-behavioral skills (IMB) model could be beneficial for HIV prevention among men who have sex with men in resource-limited settings.

Methods: Participants were recruited in Nanning, China, between April 2023 and April 2024. Eligible participants were randomly assigned to either the social media intervention group or the routine HIV prevention services control group. Participants in the intervention group received a 3-month social media intervention, which included completing video-based tasks. Baseline surveys were conducted, followed by follow-up surveys every 3 months, for a total of 2 follow-ups. Outcomes included HIV testing uptake, high-risk behavior, AIDS-related knowledge, safe sex self-efficacy, and attitude.

Results: A total of 180 eligible men who have sex with men were enrolled (90 per group). Follow-up rates were 97.8% (88/90) and 95.5% (86/90) for the intervention and control groups, respectively. At the follow-ups, the intervention group demonstrated significantly higher uptake of HIV testing, a lower proportion of participants reporting high-risk sexual behaviors, and higher condom use self-efficacy compared to the control group (all $P<.05$). After controlling for sociodemographic variables, generalized estimating equations analysis revealed that the intervention group had significantly higher odds of HIV testing (risk ratio [RR] 1.739, 95% CI 1.110 - 2.730), HIV self-testing (RR 2.306, 95% CI 1.593 - 3.340), and consistent condom use (RR 2.457, 95% CI 1.636 - 3.690) than the control group. Cochran-Armitage trend tests within the intervention group revealed that with increasing intervention duration, both HIV testing and HIV self-testing significantly increased, while high-risk sexual behaviors significantly decreased (all $P<.05$).

Conclusions: The social media intervention guided by the IMB model demonstrated a positive effect on expanding HIV testing coverage, reducing high-risk behavior, enhancing AIDS-related knowledge, and improving safer sex self-efficacy among men who have sex with men in resource-limited settings. These findings provide valuable guidance for future HIV prevention and control efforts targeting this population.

XSL•FO
RenderX

Trial Registration: Chinese Clinical Trial Registry ChiCTR2500095176; https://www.chictr.org.cn/hvshowproject.html?id=267136

(J Med Internet Res 2026;28:e84279) doi:10.2196/84279

KEYWORDS

men who have sex with men; social media intervention; information-motivation-behavioral skills model; HIV testing; high-risk behaviors

Introduction

Men who have sex with men represent a high-risk group for HIV infection. As of 2024, men who have sex with men account for 7.6% of global HIV infections among individuals aged 15 - 49 years [1]. In China, the HIV infection rate among men who have sex with men is approximately 5.7%, with rates as high as 10.7% in the southwestern region, and infection rates continue to rise [2,3]. A significant proportion of men who have sex with men remain unaware of their HIV status. By 2020, only 62.2% of men who have sex with men in China had ever undergone HIV testing [4,5], which hinders progress toward achieving the first “95” target of “95-95-95” set by the Joint United Nations Programme on HIV/AIDS (UNAIDS)—aiming for 95% of people living with HIV to know their status [6]. Men who have sex with men often engage in behaviors associated with high HIV risk, including unprotected sex, having multiple sexual partners, and recreational drug use [7]. For example, in one study conducted in China, about 37% of men who have sex with men reported engaging in unprotected anal intercourse in the past 6 months [8]. These behaviors increase the likelihood of sexual transmission and exacerbate the epidemic. Reducing such high-risk sexual behaviors can effectively lower the risk of HIV transmission among the men who have sex with men population [9]. Furthermore, they often face social discrimination and stigma [10], creating significant barriers and difficulties to seeking health services and information. These structural and social factors facilitate the spread of HIV within the population. Therefore, implementing effective strategies to improve HIV testing uptake and reduce high-risk behaviors among men who have sex with men is an urgent priority for current HIV prevention and control efforts.

In recent years, social media–based interventions have garnered prominence and have been increasingly adopted for HIV prevention among men who have sex with men [11-13]. Social media refers to various internet-based communication channels, such as social networking sites, blogs, online social media networks, and other forms of communication technology [14]—essentially, digital technology-based platforms using networks for information dissemination. With the rapid advancement of the internet and social media, mobile devices have become ubiquitous personal items. This gives social media interventions the advantage of offering fragmented learning opportunities, addressing limitations of traditional offline interventions [15]. To date, existing studies of social media–based HIV prevention interventions for men who have sex with men in China have largely focused on economically advanced metropolitan regions, including Beijing [16], Anhui [17], and Guangdong [18], which are characterized by relatively high GDP (gross domestic product), well-developed infrastructure, and strong regional influence. In contrast, there is a scarcity of research exploring the effectiveness of such interventions in resource-limited settings within China. Although Nanning serves as the capital of Guangxi, it remains relatively underresourced compared with China’s more developed cities, facing structural and resource constraints in economic development, infrastructure, and research capacity amid ongoing rapid urbanization [19,20]. Therefore, investigating the impact of social media interventions on improving HIV testing uptake and reducing high-risk behaviors among men who have sex with men in these underserved areas is crucial for enhancing local HIV prevention and control efforts.

To maximize the effectiveness of social media interventions, integrating them with established behavioral intervention theories is essential. Compared to other primary behavioral intervention theories [21,22], the information-motivation-behavioral skills (IMB) model, proposed by Fisher, comprehensively considers the influence of psychosocial factors on behavior. It categorizes factors influencing behavior change into 3 components: information, motivation, and behavioral skills. The model posits and validates that possessing disease-prevention information and motivation to engage in prevention can trigger preventive behaviors through behavioral skills to initiate and sustain action [23]. Currently, interventions based on the IMB model are widely applied in HIV behavioral interventions [24,25]. Furthermore, with the rapid development of new media technologies, integrating the IMB model into social media interventions has emerged as a growing trend in public health. Therefore, exploring the feasibility of combining social media interventions with the IMB model holds significant importance for developing more effective HIV prevention and control strategies.

Previous studies in China have demonstrated that social media–based interventions can improve HIV testing uptake among men who have sex with men, but evaluations have primarily focused on economically developed metropolitan areas [16-18]. Although this approach has been shown to increase HIV testing, evidence regarding its effectiveness in reducing high-risk behaviors remains limited. Moreover, in prior studies, intervention effects have mostly been limited to increasing HIV testing willingness or behavior, while effects on promoting broader key behavioral changes such as condom use and pre-exposure prophylaxis (PrEP) adoption remain unclear; most studies only conducted online interventions and data collection, making it difficult to ensure the quality of online information; interventions often lacked grounding in established behavioral change theories; furthermore, outcomes were often measured only at the end of the intervention period without

XSL•FO
RenderX

follow-up assessments to determine whether the effects were sustained over time [16-18]. Guangxi, a resource-limited province in southern China, has one of the highest HIV burdens nationwide, particularly among men who have sex with men [26,27]. To address these gaps, we conducted a 2-arm, parallel-group randomized controlled trial (RCT) in Guangxi, with additional emphasis on outcomes related to promoting broader key behavioral changes such as condom use and PrEP adoption, incorporated offline data collection methods to ensure quality, and the application of the IMB model to guide the intervention process, in order to evaluate whether a social media–based intervention could both increase HIV testing and reduce high-risk behaviors. Furthermore, a continued follow-up assessment was incorporated after the intervention to evaluate the durability of the intervention effects over time. This study aimed to provide context-specific evidence to evaluate a social media–based intervention's efficacy and sustainability in increasing HIV testing and reducing high-risk behaviors among men who have sex with men, thereby guiding more targeted and scalable HIV prevention strategies for men who have sex with men, supporting optimal allocation of limited health resources in comparable settings, and contributing to the achievement of the 95-95-95 targets.

Methods

Trial Design and Participants

This study used a 2-armed, parallel-group RCT design. Recruitment of participants among men who have sex with men was conducted using a combination of convenience sampling and peer-driven sampling methods through Guangxi Caihong, a gay-led community-based organization (CBO) based in Nanning, China. On-site peer educators posted online recruitment notices and registration questionnaires on social platforms such as WeChat Moments, WeChat Official Accounts, WeChat/QQ groups, and Blued. Individuals who completed online registration underwent an initial screening by staff, and the results were notified via SMS text message. Those who passed the initial screening were required to arrive at the CBO site to proceed with the next steps of the survey. Participants who were selected and completed the on-site survey received a cash compensation of RMB 20 (approximately US $3), regardless of whether they were eventually enrolled in the study. Additionally, men who have sex with men who had registered, participated, or were subsequently enrolled were encouraged to refer friends or partners who met the inclusion criteria to join the study.

The entire social media intervention procedure was conducted online via WeChat, the predominant social media platform in China [28]. Other research procedures, including informed consent, baseline assessment, and follow-up, were carried out offline. The 2-armed study design was (1) social media intervention group and (2) routine HIV prevention (control) services group. All participants completed assessments at 3 time points: baseline (preintervention), 3 months postbaseline (postintervention), and 6 months postbaseline (3 months after postintervention; Figure 1).

Figure 1. Trial design diagram. The study used a 2-armed, parallel-group randomized controlled trial design. Participants were recruited from among men who have sex with men through the Guangxi Caihong organization in Nanning, China. After recruitment and baseline survey, a 3-month intervention period commenced, with follow-up assessments conducted at the 3rd and 6th months (focusing on the 6 assessment categories shown under "follow-up endpoint" in the figure). Both the intervention and control groups participated in HIV prevention activities over 3 months, while the intervention group additionally received a social media–based intervention. PEP: post-exposure prophylaxis; PrEP: pre-exposure prophylaxis.

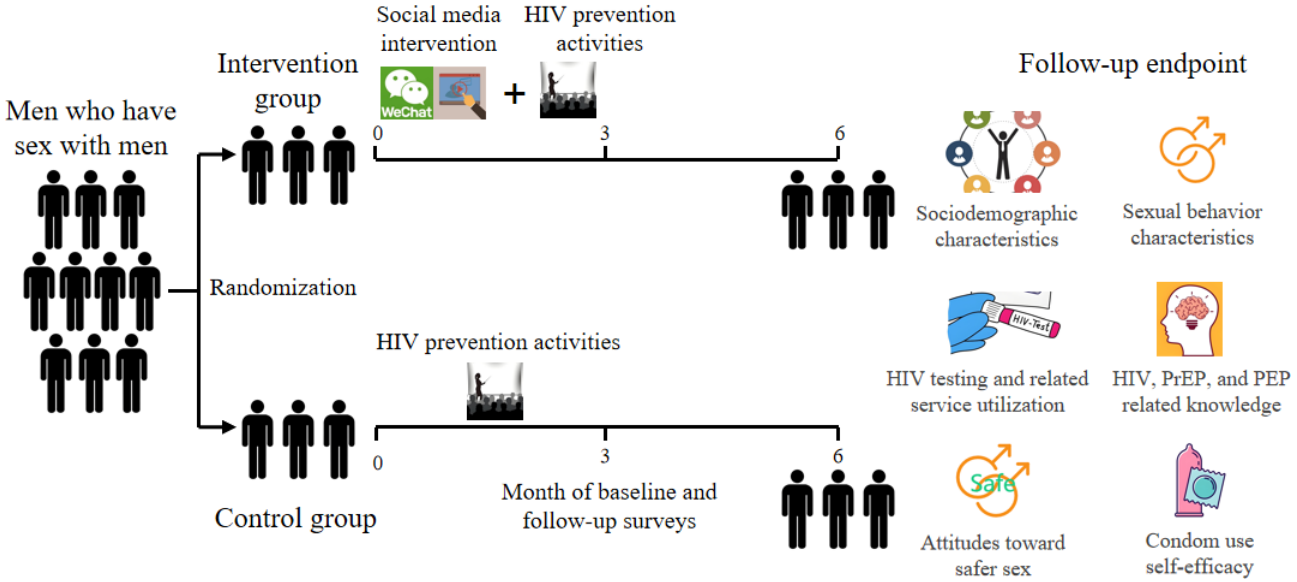

Participants were eligible if they met the following criteria: (1) born biologically male, (2) aged 18‑45 years, (3) current Nanning residents planning to reside here for most of the next year, (4) self-reported male-to-male sex in the past 6 months, (5) proficient in using WeChat and willing to connect with the study account, (6) tested HIV-negative within the past 3 months (with test results provided by participants or verified on-site by CBO), and (7) demonstrated sufficient comprehension to complete study procedures. Individuals who had significant psychological, psychiatric, and intellectual disabilities preventing survey participation, or who were currently participating in other related intervention studies, were excluded.

XSL•FO
RenderX

We followed the standard guidelines for reporting parallel group RCTs of CONSORT (Consolidated Standards of Reporting Trials) 2010 [29]. Further details regarding study design and intervention development are provided in Multimedia Appendix 1.

Sample Size

The sample size was calculated for this on-site RCT based on previous similar studies. The calculation formula is as follows:

$$N=Z\alpha 2\bar{P}(1-\bar{P})+Z\beta P1(1-P1)+P2(1-P2)(P1-P2)2$$

The study has 2 primary outcomes: HIV testing rate and the incidence of high-risk sexual behaviors. Previous studies have shown that digital media interventions significantly increase HIV testing rates or HIV self-testing rates (by 48.7% - 63%) [16,17]. Meanwhile, prior research has reported a relatively small expected difference in the incidence of high-risk sexual behaviors (17% - 35%) [9,30]. According to the RCT sample size formula, calculating based on the incidence of high-risk sexual behaviors yields a larger required sample size. To ensure sufficient study power and reflect the most conservative estimate, this study's sample size was determined based on the expected difference in the incidence of high-risk sexual behaviors between the two men who have sex with men groups.

Previous studies had reported that 25% - 42% of men who have sex with men engaged in high-risk sexual behaviors in the past 6 months [8,30]. Based on this evidence, we assumed a baseline prevalence (ρ) of 35%. Prior interventions had demonstrated reductions of 17% - 35% in such behaviors [9,31]. To ensure a conservative estimate, we assumed a 15% reduction, setting the expected prevalence in the intervention group (P_1) at 20% and in the control group (P_2) at 35%. With a significance level (α) of .05 (Z_α=1.96) and power (1 – β) of 80% (Z_β=0.84), the required sample size per group was calculated as n=78. Accounting for an estimated 10% loss to follow-up, we planned to recruit 90 participants per group, yielding a total of 180 men who have sex with men.

Randomization and Blinding

To ensure the privacy of study participants, each was assigned a unique identification number for all subsequent follow-ups. For randomization, the allocation sequence was generated by an independent researcher, who was not involved in participant recruitment or any subsequent intervention procedures, using the "randomizr" package in R software (R Foundation for Statistical Computing). Participants were then randomly allocated in a 1:1 ratio to either the intervention or control group based on their assigned identification numbers. Participants were blinded to their group assignment and instructed to maintain confidentiality regarding their intervention status. Throughout the recruitment, baseline assessment, and all follow-up stages, the research staff responsible for these tasks remained unaware of the group allocations, thereby ensuring allocation concealment. While research staff administering the interventions were necessarily unblinded, the statisticians conducting data analysis remained blinded to group allocations throughout the study to ensure objective outcome assessment.

Intervention Exposure

All participants added the study team's WeChat work account for information notifications. The control arm received routine HIV prevention services provided by Guangxi Caihong. These services, conducted offline, included HIV/AIDS education and awareness campaigns, regular HIV testing services, and structured peer education sessions, held at least once a month.

Participants in the intervention group received the same services as the control group, with the addition of a social media–based intervention. This intervention was delivered via "Rainbow Home," a WeChat mini-program developed by the research team. The research team assisted participants in the intervention group with guided registration, provided a detailed explanation of the mini-program's features, and conducted weekly check-in tasks (Multimedia Appendix 2). To minimize intervention contamination between groups, researchers provided both verbal and written instructions to all participants in the intervention group before the intervention began, requesting that they refrain from sharing the mini-program or any task-related content with others until the end of the study period. This request did not impose any coercion or restriction on the participants. Furthermore, the WeChat mini-program used in this study was configured with restricted access and anonymous log-in. Access privileges were granted exclusively to the intervention group participants by the researchers via dynamic invitation codes, ensuring that unauthorized individuals could not use the mini-program. Participants in the intervention group were unable to view any information about other members within their group. Additionally, service schedules of the intervention group differed from the control group, while all other elements—including activity content and study personnel—remained consistent across both groups.

The intervention based on the IMB model was structured around 4 core components: information, motivation, behavioral skills, and behavior. Content mainly contained comprehensive HIV/AIDS knowledge, real-world case studies highlighting HIV infection risks and prevention narratives, practical guidance on safer sexual practices and correct use of prevention tools, and provision of free HIV self-testing kits. See Table 1 for detailed intervention content.

XSL•FO
RenderX

Table . The detailed intervention content based on the information-motivation-behavioral skills (IMB) model[a].

Component	Content
Information	Delivered knowledge via animated videos on HIV/AIDS, sexually transmitted infections, correct HIV testing and self-testing, proper use of PrEP[b] and PEP[c], correct condom use, recreational drug side effects, and HIV case studies with treatment.
Motivation	Raised awareness of safer sex and self-protection through sharing HIV infection case studies and provided one-on-one psychological counseling to participants in need.
Behavioral skills	Weekly video-based check-in tasks were sent via mini-program, with staff available for consultation and tailored advice. Free HIV self-test kits with usage guidance were provided, and partner testing was encouraged.
Behavior	Change attitudes toward safer sex, improve self-efficacy in condom use, promote intentions to engage in safer sexual practices, increase awareness of HIV testing, and reduce high-risk sexual behaviors.

[a]Guided by the IMB model, the intervention content was structured around the 4 core components of the model: information, motivation, behavioral skills, and behavior.

[b]PrEP: pre-exposure prophylaxis.

[c]PEP: post-exposure prophylaxis.

The intervention was delivered primarily through video viewing; for example, procedural knowledge (such as PrEP usage guidelines and HIV self-testing procedures) was provided through animated videos. Research staff released 1‑2 video-based tasks per week (a total of 12 tasks) via the mini-program backend. The main content, video duration of each task, and its alignment with the 4 components based on the IMB model are provided in Multimedia Appendix 3. Tasks were pushed every Sunday at 7:30 PM and remained accessible for 2 weeks to allow flexibility in completion. Automated reminder notifications were sent through WeChat, and participants completing all tasks within the 3-month intervention period received financial incentives. In parallel, trained staff provided real-time consultation through WeChat, addressing participant inquiries and encouraging feedback on platform usability. At the 3- and 6-month follow-ups, intervention participants were encouraged to conduct HIV testing on-site, request free HIV self-testing kits by mail or on-site pickup, and invite their sexual partners to undergo testing.

During the follow-up process, for participants suspected of being HIV-positive, the research team members provided them with corresponding psychological support and medical referral services, assisting them in visiting the local Centers for Disease Control or designated hospitals for further confirmatory testing and subsequent follow-up management. Once a participant was confirmed positive, it was considered an endpoint event, and further follow-up was discontinued from the date of diagnosis.

The intervention in this study was conducted in accordance with the research protocol. After the intervention materials were developed, they were reviewed by 3 public health experts and peer educators before being finalized and disseminated. All research staff received standardized training prior to the intervention, and they documented and reported the implementation of interventions in real time. Dedicated staff monitored each participant's intervention engagement and task completion rate via the system backend, and backend monitoring data were exported and evaluated every 2 weeks.

After the conclusion of this study, participants in the control group will be granted the same access to the WeChat mini-program and all intervention content as the intervention group. However, the program still requires further optimization and is not currently intended for public release.

Outcomes, Data Collection, and Management

The primary outcome was the uptake of HIV testing and HIV self-testing (verified through testing at the CBO and photo-submitted self-test results to the staff at the 3-month and 6-month follow-ups after enrollment, respectively), along with self-reported high-risk behaviors among men who have sex with men in the past 3 months (defined as engaging in anal intercourse with a male partner without consistent condom use throughout the act or experiencing condom breakage during intercourse). Secondary outcomes included the uptake of syphilis testing (verified through testing at the CBO), as well as knowledge of AIDS, PrEP, and post-exposure prophylaxis (PEP), self-efficacy for condom use, and safe sex attitudes. Validated scales were used to measure knowledge, self-efficacy, and safe sex attitudes. Definitions of outcomes and methods of outcome assessment were detailed in Multimedia Appendix 4.

All self-reported outcomes were evaluated via follow-up surveys administered at enrollment and at 3 and 6 months postenrollment, conducted offline at the CBO site through one-on-one interviews by research staff for data collection.

All original questionnaires and backend-exported data were independently double-entered by two researchers after quality control checks. Personal information was removed, and the data were stored on an encrypted hard drive accessible only to the principal investigator of the research team. The data used during the analysis phase were anonymized. Both the original paper questionnaires and backup electronic questionnaire files are stored in the research team's encrypted hard drive.

XSL•FO
RenderX

Statistical Analysis

Descriptive statistics were used to summarize the baseline characteristics. Continuous variables are presented as mean (SD) and were compared using *t* tests or ANOVA. Categorical variables are presented as frequency (percentage) and were compared using chi-square tests or Fisher exact test. The Cochran-Armitage trend test was used for repeated categorical measures within the intervention group. To investigate the effect of the intervention, generalized estimating equations (GEE) were used to analyze correlations for repeated measures data (eg, HIV testing and condom use) across baseline and follow-up time points and between intervention and control groups. Bonferroni correction was applied for post hoc pairwise comparisons. All tests were 2-tailed, with a significance level defined as α=.05; *P* values <.05 were considered statistically significant. All data analyses were completed using R Studio (version 4.2.1; Posit Software) or SPSS (version 26.0; IBM Inc).

Ethical Considerations

This RCT was retrospectively registered at the Chinese Clinical Trial Registry (ChiCTR2500095176). Ethical approval was obtained from the Medical Ethics Committee of Nanning Center for Disease Control and Prevention (2023014) and the Medical Ethics Committee of Guangxi Medical University (2024-KY0148).

Following ethics committee approval, the study was submitted to the Chinese Clinical Trial Registry. However, due to the platform's review process and timeline, participant recruitment began under the ethics-approved protocol before the official trial registration number was issued, resulting in a retrospective registration. Throughout the study, all procedures adhered strictly to the original ethics-approved protocol, with no changes to the study content or primary outcomes before or after registration.

Participation in this study was voluntary, and written informed consent was provided by all participants. All analytical data were processed anonymously to ensure participant confidentiality. All research staff committed to not disclosing any participant information. At baseline and each follow-up, participants received a cash compensation of RMB 20 (approximately US $3) upon completing the survey. Additionally, those who completed all tasks during the intervention period were provided an extra cash compensation of RMB 20 (approximately US $3).

Results

Participant Flow and Intervention Adherence

Recruitment took place between April 2023 and August 2023, during which 392 men who have sex with men were initially screened. After applying the inclusion and exclusion criteria, 180 eligible participants were enrolled and randomly allocated to either the social media intervention group (n=90) or the routine HIV prevention control group (n=90). The formal intervention began in September 2023, and all follow-ups were completed by April 2024. All participants were retained at the 3-month follow-up assessment. At the 6-month follow-up, loss to follow-up occurred in 2 participants (2/90, 2.2%) from the intervention group and 4 participants (4/90, 4.5%) from the control group, yielding follow-up rates of 97.8% (88/90) and 95.5% (86/90), and an overall follow-up rate of 96.7% (174/180). The participant recruitment, intervention, and follow-up flowchart is detailed in Figure 2. During the 3-month intervention period, 12 video check-in tasks were delivered to the intervention group. The average duration per video task was 112 seconds. Participants completed a total of 861 check-ins, representing a task completion rate of 79.7% (861/1080). Due to the low proportion of missing data, no data imputation was performed in this study. The analysis of this study was based on complete case analysis, including only participants with complete outcome data (ie, those who completed all follow-ups).

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flow diagram. Inclusion criteria: (1) born biologically male, (2) aged 18-45 years, (3) current Nanning residents planning to reside here for most of the next year, (4) self-reported male-to-male sex in the past 6 months, (5) proficient in using WeChat and willing to connect with the study account, (6) tested HIV-negative within the past 3 months, and (7) demonstrated sufficient comprehension to complete study procedures.

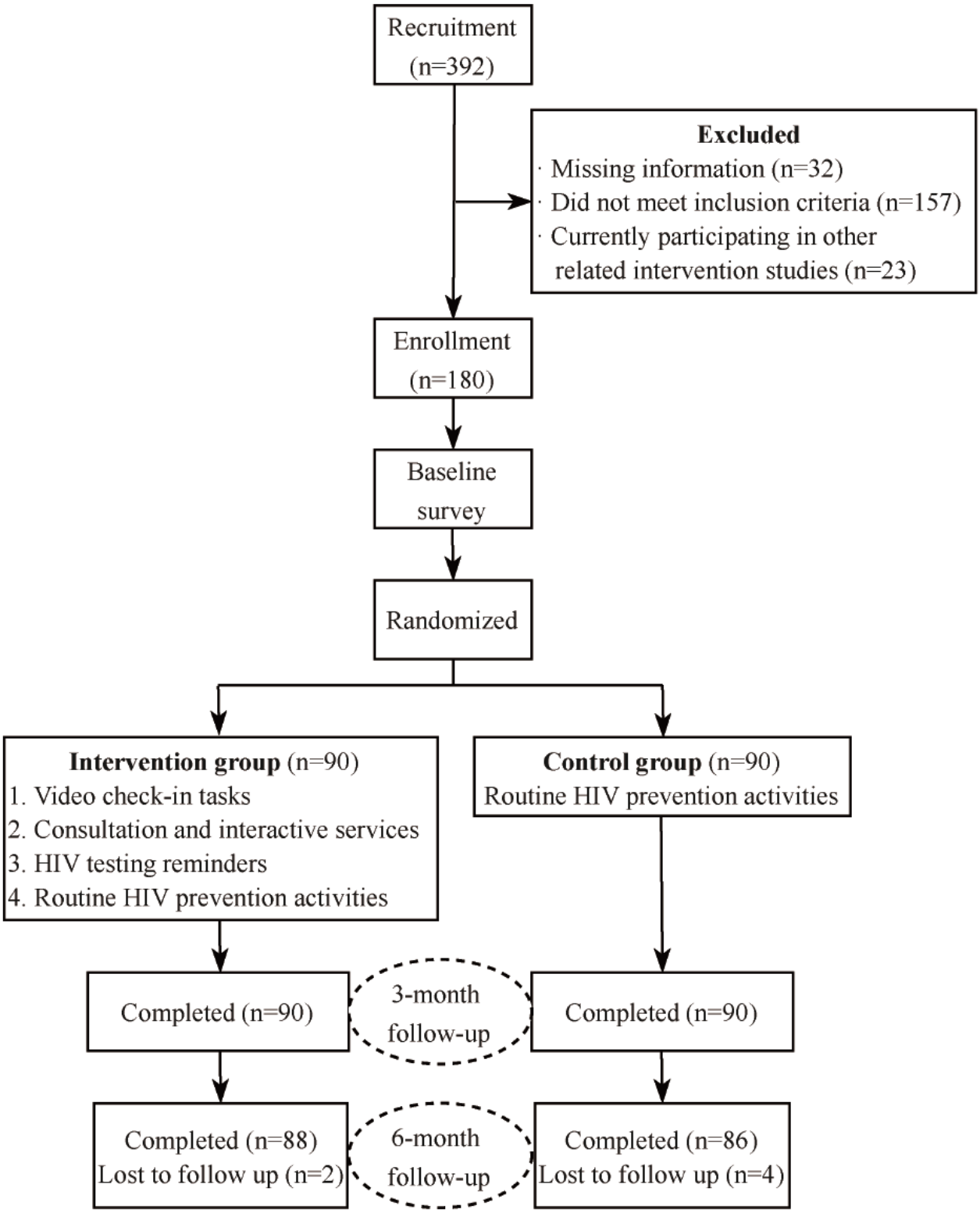

Baseline Characteristics of Participants

The mean age of participants was 28.5 (SD 6.9) years. The majority (143/180, 79.4%) had an educational level of college/university or above. Most participants were employed in government institutions or as office workers (91/180, 50.6%). Regarding sexual behavior characteristics, 68.3% (123/180) identified as homosexual in orientation. More than half (104/180, 57.8%) reported engaging in high-risk sexual behaviors. Of the 180 participants, 106 (58.9%) reported having

XSL•FO
RenderX

performed HIV self-testing in the past 3 months. After randomization, there were no statistically significant differences in demographic or behavior characteristics between the two groups, indicating comparability (Table 2).

XSL•FO
RenderX

Table . Sociodemographic and behavioral characteristics of study participants.

Characteristics	Total (n=180), n (%)	Control group (n=90), n (%)	Intervention group (n=90), n (%)	Chi-square (*df*)	*P* value
Age (year)				0.7 (2)	.70
≤24	51 (28.3)	23 (25.6)	28 (31.1)		
25-34	87 (48.3)	45 (50.0)	42 (46.7)		
≥35	42 (23.3)	22 (24.4)	20 (22.2)		
Household registration				0.4 (2)	.82
Nanning City	84 (46.7)	40 (44.4)	44 (48.9)		
Other cities within the province	80 (44.4)	42 (46.7)	38 (42.2)		
Other cities outside the province	16 (8.9)	8 (8.9)	8 (8.9)		
Ethnicity				2.1 (2)[a]	.34
Han	100 (55.6)	47 (52.2)	53 (58.9)		
Zhuang	75 (41.7)	39 (43.3)	36 (40.0)		
Other	5 (2.8)	4 (4.4)	1 (1.1)		
Education level				2.3 (2)[a]	.34
Junior high school or below	9 (5.0)	6 (6.7)	3 (3.3)		
Vocational school/high school	28 (15.6)	11 (12.2)	17 (18.9)		
College/bachelor's degree or higher	143 (79.4)	73 (81.1)	70 (77.8)		
Occupation				4.1 (4)[a]	.38
Student	35 (19.4)	15 (16.7)	20 (22.2)		
Public institution/staff	91 (50.6)	48 (53.4)	43 (47.8)		
Farmer/worker	11 (6.1)	3 (3.3)	8 (8.9)		
Self-employed	24 (13.3)	12 (13.3)	12 (13.3)		
Unemployed	19 (10.6)	12 (13.3)	7 (7.8)		
Monthly income (CNY)				0.7 (2)	.70
<3000 (US $435)	63 (35.0)	32 (35.6)	31 (34.4)		
3000-5000 (US $435-$725)	67 (37.2)	31 (34.4)	36 (40.0)		
>5000 (US $725)	50 (27.8)	27 (30.0)	23 (25.6)		
Marital status				1.2 (2)	.55
Unmarried	160 (88.8)	78 (86.7)	82 (91.1)		
Married	10 (5.6)	7 (7.8)	3 (3.3)		
Divorced	10 (5.6)	5 (5.5)	5 (5.6)		
Living arrangement				0.6 (2)	.75
Living alone	95 (52.8)	47 (52.2)	48 (53.3)		
Living with family/partner	42 (23.3)	23 (25.6)	19 (21.1)		
Dormitory/shared rental	43 (23.9)	20 (22.2)	23 (25.6)		

XSL•FO
RenderX

Characteristics	Total (n=180), n (%)	Control group (n=90), n (%)	Intervention group (n=90), n (%)	Chi-square (*df*)	*P* value
Sexual orientation				1.1 (2)[a]	.59
Homosexual	123 (68.3)	60 (66.7)	63 (70.0)		
Bisexual	49 (27.2)	27 (30.0)	22 (24.4)		
Uncertain	8 (4.4)	3 (3.3)	5 (5.6)		
Sexual role				0.1 (2)	.96
Insertive role	48 (26.7)	23 (25.6)	25 (27.8)		
Receptive role	65 (36.1)	33 (36.7)	32 (35.6)		
Versatile	67 (37.2)	34 (37.8)	33 (36.7)		
Age at first anal sex (years)				0.8 (1)	.45
≤18	74 (41.1)	34 (37.8)	40 (44.4)		
>18	106 (58.9)	56 (62.2)	50 (55.6)		
Number of regular partners				0.5 (2)	.85
0	50 (27.8)	27 (30.0)	23 (25.6)		
1-2	120 (66.7)	58 (64.4)	62 (68.9)		
≥3	10 (5.6)	5 (5.6)	5 (5.6)		
Number of casual partners				2.0 (2)	.38
0	77 (42.8)	34 (37.8)	43 (47.8)		
1-2	46 (25.6)	24 (26.7)	22 (24.4)		
≥3	57 (31.7)	32 (35.6)	25 (27.8)		
Awareness of partner's HIV status				2.3 (1)	.13
No	78 (43.3)	44 (48.9)	34 (37.8)		
Yes	102 (56.7)	46 (51.1)	56 (62.2)		
Sexting				1.1 (1)	.39
No	46 (25.6)	26 (28.9)	20 (22.2)		
Yes	134 (74.4)	64 (71.1)	70 (77.8)		
Popper/alkyl nitrite inhalant use				0.3 (1)	.62
No	131 (72.8)	64 (71.1)	67 (74.4)		
Yes	49 (27.2)	26 (28.9)	23 (25.5)		
Consistent condom use				3.3 (1)	.10
Yes	76 (42.2)	32 (35.6)	44 (48.9)		
No	104 (57.8)	58 (64.4)	46 (51.1)		
HIV self-testing				0.4 (1)	.65
No	74 (41.1)	35 (38.9)	39 (43.3)		
Yes	106 (58.9)	55 (61.1)	51 (56.7)		
Syphilis test result				1.6 (1)	.31
Negative	163 (90.6)	84 (93.3)	79 (87.8)		
Positive	17 (9.4)	6 (6.7)	11 (12.2)		
HIV testing				1.5 (1)	.15
No	43 (23.9)	18 (20.0)	25 (27.8)		

XSL•FO
RenderX

Characteristics	Total (n=180), n (%)	Control group (n=90), n (%)	Intervention group (n=90), n (%)	Chi-square (*df*)	*P* value
Yes	137 (76.1)	72 (80.0)	65 (72.2)		

[a]Fisher exact test.

Social Media Interventions Drive Increases Testing and Reductions in High-Risk Behavior Among Men Who Have Sex With Men

At the 3-month follow-up, the HIV self-testing rate was significantly higher in the intervention group than in the control group ($P<.001$), and the proportion of participants reporting high-risk sexual behaviors was significantly lower ($P<.05$). The number of casual partners and the rate of popper/alkyl nitrite inhalant use in the intervention group were significantly lower than in the control group ($P<.05$). The intervention group had a significantly higher rate of awareness of their sexual partners' HIV status compared to the control group ($P<.05$; Table 3).

XSL•FO
RenderX

Table . Comparison of sexual behavior characteristics and HIV testing status between the intervention group and the control group among study participants after 6-month follow-up.

Variables and groups	Baseline (intervention: n=90; control: n=90), n (%)	3-month follow-up (intervention: n=90; control: n=90), n (%)	6-month follow-up (intervention: n=88; control: n=86), n (%)
HIV testing			
Intervention	65 (72.2)	74 (82.2)	81 (92.0)[a,b]
Control	72 (80.0)	66 (73.3)	56 (65.1)
HIV self-testing			
Intervention	51 (56.7)	63 (70.0)[a]	63 (71.6)[a,c]
Control	55 (61.1)	31 (34.4)	32 (37.2)
Syphilis test result (positive)			
Intervention	11 (12.2)	3 (3.3)	4 (4.5)
Control	6 (6.7)	13 (14.4)	7 (8.1)
Number of regular partners[d]			
0			
Intervention	23 (25.6)	24 (26.7)	25 (28.4)
Control	27 (30.0)	33 (36.7)	21 (26.7)
1 - 2			
Intervention	62 (68.9)	69 (76.7)	60 (68.2)
Control	58 (64.4)	52 (57.8)	61 (76.7)
≥3			
Intervention	5 (5.6)	5 (5.6)	3 (3.4)
Control	5 (5.6)	5 (5.6)	4 (5.5)
Number of casual partners[e]			
0			
Intervention	43 (47.8)	54 (60.0)	38 (43.2)
Control	34 (37.8)	46 (51.1)	47 (60.0)
1 - 2			
Intervention	22 (24.4)	31 (34.4)	32 (36.4)
Control	24 (26.7)	30 (33.3)	31 (34.4)
≥3			
Intervention	25 (27.8)	11 (12.2)	13 (14.8)
Control	32 (35.6)	14 (15.5)	8 (12.2)
Awareness of partner's HIV status[f]			
Intervention	56 (62.2)	69 (76.7)[d]	62 (74.7)[a]
Control	46 (51.1)	50 (62.5)	40 (49.4)
High-risk sexual behavior[f]			
Intervention	46 (51.1)	33 (36.7)[d]	26 (31.3)[g]
Control	58 (64.4)	45 (55.5)	36 (44.4)
Sexting			
Intervention	70 (77.8)	63 (70.0)	56 (63.6)[c]
Control	64 (71.1)	57 (63.3)	58 (67.4)
Popper/alkyl nitrite inhalant use			

XSL•FO
RenderX

Variables and groups	Baseline (intervention: n=90; control: n=90), n (%)	3-month follow-up (intervention: n=90; control: n=90), n (%)	6-month follow-up (intervention: n=88; control: n=86), n (%)
Intervention	23 (25.5)	16 (17.8)[d]	12 (13.6)[c,e]
Control	26 (28.9)	31 (34.4)	26 (30.2)

[a]*P*<.001 compared to the control group.

[b]*P*<.001 compared to the baseline.

[c]*P*<.05 compared to the baseline.

[d]*P*<.05 compared to the control group.

[e]*P*<.01 compared to the control group.

[f]There are missing values as some study participants had no sexual partners in the last 3 months and were therefore excluded from the analysis.

[g]*P*<.01 compared to the baseline.

At the 6-month follow-up, both HIV testing rate and HIV self-testing rate were significantly higher in the intervention group compared to the control group (*P*<.001). The proportions of participants reporting high-risk sexual behaviors and popper/alkyl nitrite inhalant use were lower in the intervention group than in the control group (*P*<.05). Additionally, participants in the intervention group were significantly more likely to be aware of their sexual partners' HIV status than those in the control group (*P*<.001). Regarding sexually transmitted infections, 16 cases of syphilis were detected at the 3-month follow-up (3 in the intervention group and 13 in the control group), and 11 cases were identified at the 6-month follow-up (4 in the intervention group and 7 in the control group). No participants tested positive for HIV at either follow-up assessment (Table 3).

Cochran-Armitage trend tests within the intervention group revealed significant linear trends over time for all measured variables except number of regular sexual partners, number of casual sexual partners, and awareness of partner HIV status (*P*<.05). These variables included HIV testing, HIV self-testing, high-risk sexual behaviors, sexting behavior, and popper/alkyl nitrite inhalant use. Specifically, the positive effects observed in the intervention group at month 3 (the end of the intervention)—such as the promotion of HIV testing and HIV self-testing, reduction of high-risk sexual behaviors, and reduction of Rush popper use—were maintained and even strengthened at the 6-month follow-up assessment (3 months after the intervention ended; Table 3).

After controlling for sociodemographic variables, repeated measures ANOVA (with time as the within-subject factor and group as the between-subject factor) revealed: The intervention group had significantly higher odds of HIV testing (risk ratio [RR] 1.739, 95% CI 1.110 - 2.730), HIV self-testing (RR 2.306, 95% CI 1.593 - 3.340), and consistent condom use (RR 2.457, 95% CI 1.636 - 3.690) compared to the control group. The intervention group had significantly lower odds of popper/alkyl nitrite inhalant use compared to the control group (RR 0.554, 95% CI 0.365 - 0.841). Compared to the control group, the intervention group showed significantly increased odds of HIV self-testing at the 3-month follow-up (RR 5.651, 95% CI 2.321 - 13.761); also, significantly increased odds of both HIV testing (RR 10.473, 95% CI 3.276 - 33.491) and HIV self-testing (RR 5.651, 95% CI 2.321 - 13.761) were observed at the 6-month follow-up (Table 4).

XSL•FO
RenderX

Table . Analysis of the impact of new media intervention on high-risk sexual behaviors and HIV testing before and after intervention[a].

Variable	β value	RR[b] (95% CI)	*P* value[c]
HIV testing (no=0, yes=1)			
Group (intervention vs control)	0.553	1.739 (1.110 - 2.730)	*.02*
Time1 (3-month follow-up)	0.208	1.187 (0.709 - 1.990)	.51
Time2 (6-month follow-up)	0.172	0.475 (0.732 - 2.070)	.43
Group * Time1	0.893	2.443 (0.876 - 6.808)	.09
Group * Time2	2.349	10.473 (3.276 - 33.491)	*<.001*
HIV self-testing (no=0, yes=1)			
Group (intervention vs control)	0.835	2.306 (1.593 - 3.340)	*<.001*
Time1 (3-month follow-up)	−0.225	0.798 (0.512 - 1.250)	.32
Time2 (6-month follow-up)	−0.151	0.860 (0.551 - 1.340)	.51
Group * Time1	1.732	5.651 (2.321 - 13.761)	*<.001*
Group * Time2	1.686	5.395 (2.221 - 13.117)	*<.001*
Consistent condom use[d] (no=0, yes=1)			
Group (intervention vs control)	0.899	2.457 (1.638 - 3.687)	*<.001*
Time1 (3-month follow-up)	0.699	2.011 (1.252 - 3.233)	*.004*
Time2 (6-month follow-up)	0.873	2.395 (1.475 - 3.885)	*<.001*
Group * Time1	−0.222	0.801 (0.310 - 2.068)	.65
Group * Time2	0.013	1.013 (0.382 - 2.683)	.98
Sexting (no=0, yes=1)			
Group (intervention vs control)	0.242	0.785 (0.530 - 1.163)	.23
Time1 (3-month follow-up)	−0.509	0.636 (0.375 - 0.964)	*.04*
Time2 (6-month follow-up)	−0.564	0.748 (0.351 - 0.923)	*.02*
Group * Time1	−0.125	0.882 (0.340 - 2.290)	.80
Group * Time2	−0.560	0.571 (0.216 - 1.510)	.26
Popper/alkyl nitrite inhalant use (no=0, yes=1)			
Group (intervention vs control)	−0.591	0.554 (0.365 - 0.841)	*.006*
Time1 (3-month follow-up)	−0.031	0.970 (0.595 - 1.579)	.90
Time2 (6-month follow-up)	−0.327	1.060 (0.436 - 1.193)	.21
Group * Time1	−0.808	0.446 (0.167 - 1.190)	.11
Group * Time2	−0.883	0.414 (0.147 - 1.160)	.09

[a]The generalized estimating equations model was adjusted for age, household registration, ethnicity, education level, occupation, and monthly income.

[b]RR: risk ratio.

[c]Italic values represent statistically significant at $P<.05$.

[d]There are missing values as some study participants had no sexual partners within the last 3 months and were therefore excluded from the analysis.

Social Media Intervention Significantly Increases AIDS-Related Knowledge

Knowledge scores for HIV/AIDS, PrEP, and PEP at baseline, 3 months, and 6 months were presented in Table 5 (mean scores shown). Comparison results indicated that compared to the control group, the intervention group showed significantly higher PrEP knowledge scores at both the 3-month and 6-month follow-ups (both $P<.05$). Furthermore, within the intervention group, knowledge scores for HIV/AIDS, PrEP, and PEP were significantly higher at both follow-up time points compared to baseline (all $P<.05$; Table 5).

XSL•FO
RenderX

Table . Comparison of awareness levels of relevant health knowledge between the intervention group and the control group before and after intervention.

Variables and groups	Baseline, mean (SD)	3-month follow-up, mean (SD)	6-month follow-up, mean (SD)
AIDS awareness level			
Intervention	13.10 (3.14)	15.10 (2.28)[a,b]	14.50 (2.02)[a,b]
Control	13.10 (3.14)	13.08 (2.32)	13.09 (2.75)
PrEP[c] awareness level			
Intervention	3.31 (1.56)	5.91 (1.39)[a,b]	6.01 (1.69)[a,b]
Control	3.14 (1.41)	4.87 (2.17)[a]	4.79 (2.38)[a]
PEP[d] awareness level			
Intervention	5.71 (1.83)	6.71 (1.60)[a,b]	6.68 (1.48)[a,b]
Control	5.58 (1.75)	5.48 (2.17)	5.42 (2.36)

[a]$P<.05$ compared to the baseline.
[b]$P<.05$ compared to the control group.
[c]PrEP: pre-exposure prophylaxis.
[d]PEP: post-exposure prophylaxis.

Social Media Intervention Improves Condom Use Self-Efficacy and Modifies Sex Attitudes in Men Who Have Sex With Men

Scores for condom use self-efficacy and safer sex attitudes at baseline, 3 months, and 6 months are presented in Table 6 (mean scores shown). Results demonstrated that compared to the control group, condom use self-efficacy was significantly higher in the intervention group at the 3-month follow-up ($P<.05$). Within the intervention group, condom use self-efficacy scores increased significantly from baseline to both the 3-month and 6-month follow-ups (both $P<.05$), with scores at 6 months being significantly higher than those at 3 months ($P<.05$; Table 6).

Table . Comparison of safe behavior attitudes and self-efficacy in condom use between the intervention group and the control group before and after intervention.

Variables and groups	Baseline, mean (SD)	3-month follow-up, mean (SD)	6-month follow-up, mean (SD)
Condom use self-efficacy			
Intervention	30.18 (6.27)	32.29 (5.38)[a,b]	33.21 (5.06)[a,c]
Control	29.03 (6.24)	30.82 (4.83)[a]	32.89 (5.83)[a,c]
Safe behavior attitudes			
Intervention	45.95 (8.65)	47.74 (7.82)	47.65 (7.20)
Control	45.39 (8.06)	46.73 (7.57)	46.87 (7.38)

[a]$P<.05$ compared to the control group.
[b]$P<.05$ compared to the baseline.
[c]$P<.05$ compared to the 3-month follow-up.

Discussion

Concerning the impact of the social media intervention on promoting HIV testing among men who have sex with men, this study found a higher uptake of HIV testing in the intervention group compared to the control group. This finding aligns with the results of some previous similar studies, such as two WeChat-based interventions conducted in Anhui, China, and Guangdong, China, where the HIV testing rates in the intervention groups were significantly higher than those in the control groups [17,18]. Our research further confirms the effectiveness of WeChat-based social media interventions in promoting HIV testing among the men who have sex with men population, indicating a positive effect of the social media–based strategy in increasing HIV testing coverage. Particularly for young, internet-savvy men who have sex with men, social media interventions could more effectively convey HIV prevention knowledge and enhance their testing awareness and capacity [32].

Regarding the effect of the social media intervention on reducing high-risk behaviors among men who have sex with men, our results demonstrated that the intervention group exhibited significantly lower frequencies of high-risk sexual behaviors and popper/alkyl nitrite inhalant use compared to the control group. This indicated that the social media intervention effectively reduces high-risk sexual behaviors among men who have sex with men, aligning with findings from multiple similar studies [9,31,33]. It demonstrates the broad application potential

XSL•FO
RenderX

of social media interventions in reducing high-risk behaviors and promoting HIV prevention. However, in China, research exploring the impact of WeChat-based social media interventions on reducing high-risk sexual behaviors remains relatively scarce. A WeChat-based intervention study conducted in Anhui Province, China, showed no significant difference in condom use behavior between the intervention and control groups, which may be related to the measures implemented in the control group [17]. This indicates that further and more in-depth research is still needed regarding the effects of WeChat-based social media interventions on reducing high-risk sexual behaviors in China. Furthermore, within the intervention group, after receiving the social media intervention, significant reductions were observed in the frequency of high-risk sexual behaviors, sexting behavior, and popper/alkyl nitrite inhalant use compared to baseline. GEE model analysis confirmed the positive effect of the intervention on reducing high-risk behaviors.

Furthermore, our study found that the positive effects observed in the intervention group at month 3 (the end of the intervention)—such as the promotion of HIV testing and HIV self-testing, reduction of high-risk sexual behaviors, and reduction of Rush popper use compared to the control group—were maintained and even strengthened at the 6-month follow-up assessment (3 months after the intervention ended). This indicates that the social media intervention in this study had sustained effects in promoting risk behavior change among men who have sex with men, continuing to positively influence their health behaviors even after the intervention concluded. This highlights the potential long-term benefits of social media as a health education platform for men who have sex with men, and the important role of social media as a platform for information dissemination and health education.

Our results showed that AIDS-related knowledge levels significantly increased from baseline to postintervention in both groups, with a more pronounced increase in the intervention group. This suggested that, compared to conventional methods, the social media intervention was more effective in enhancing relevant knowledge levels among men who have sex with men. Due to societal discrimination and stigma, especially in resource-limited settings, traditional sexual education and counseling services are often inaccessible or ineffective for men who have sex with men [34,35]. Emerging internet-based social media interventions, popular in men who have sex with men research, use engaging graphics, animations, and short videos to capture attention, foster interactivity and participation, and create a positive atmosphere for interaction. This strengthened men who have sex with men's understanding and retention of health information, further promoting the dissemination and depth of HIV knowledge delivery. Moreover, PrEP has been proven effective in reducing HIV risk [3]. As a novel and effective educational tool, social media interventions played a crucial role in improving AIDS knowledge, particularly awareness and willingness to use PrEP and PEP, among men who have sex with men. Therefore, future interventions should continue to leverage internet platforms and the strengths of social media to provide more personalized and acceptable services for men who have sex with men.

Using GEE models, this study also evaluated the intervention's impact on safer sex attitudes and condom use self-efficacy. For condom use self-efficacy, results showed a significant improvement in the intervention group compared to the control group. Within the intervention group, scores significantly increased at both 3 and 6 months compared to baseline. This indicated the positive role of the new media intervention in promoting condom use self-efficacy among men who have sex with men. Bogale et al [36] found significantly increased condom use intention among rural Ethiopian women following an audio-based HIV intervention. Hirshfield and Hennessy et al [37,38] found that television- and radio-based media interventions can increase the intention to use condoms among African American adolescents. Our study further confirms that media-based interventions can enhance individuals' confidence and capability to use condoms during sex practices. However, no significant differences were found in safer sex attitudes between the groups. This might indicate that men who have sex with men have relatively fixed attitudes toward sexual behavior, shaped by long-term sociocultural influences and personal experiences, which were difficult to change through short-term interventions. Future research could further extend the intervention and follow-up period to confirm its effects on attitudes toward safe sexual behavior.

Our study demonstrates that, compared to routine HIV prevention services alone, a social media–based intervention via WeChat significantly increased HIV testing rates, reduced high-risk sexual behaviors, and enhanced HIV-related knowledge among men who have sex with men. As a leading social media platform in China, WeChat features Mini Programs that require no additional installation, offering low user accessibility barriers [28,39]. This makes it particularly suitable for use in resource-limited settings such as Guangxi. Our research provides novel insights and a reference for integrating social media interventions with routine HIV prevention services to develop diversified strategies in resource-limited regions.

The IMB model categorizes factors influencing behavior change into 3 components: information, motivation, and behavioral skills. It proposes and validates that possessing disease prevention information and having the motivation to engage in prevention can, through behavioral skills, trigger preventive actions to initiate and sustain behavioral change [23]. This study found that interventions based on the IMB model effectively promoted the adoption of HIV preventive behaviors, following the theoretically hypothesized progressive pathway: First, procedural knowledge (such as PrEP usage guidelines and HIV self-testing procedures) provided through animated videos established a solid informational foundation (eg, individuals learned and accumulated knowledge related to HIV testing), while case studies and psychological counseling simultaneously enhanced personal risk perception and awareness of social support, collectively addressing the issues of "knowledge" and "motivation" (eg, individuals recognized the harms of not undergoing HIV testing through case studies, developed a sense of urgency, and strengthened their motivation and support for HIV testing through psychological counseling). Subsequently, skill-building components—such as weekly video check-ins, personalized consultations, and the distribution of free

XSL•FO
RenderX

self-testing kits—facilitated the practical mastery of operational skills and improved self-efficacy by creating a low-risk practice environment (eg, individuals learned and mastered the skills for HIV testing). These efforts successfully translated the accumulated information and motivation into actionable behavioral capabilities. Ultimately, this multicomponent synergistic effect drove significant behavioral improvements (eg, individuals ultimately decided to undergo HIV testing). In this study, this was specifically evidenced by increased HIV testing rates, enhanced self-efficacy in condom use, and reduced frequency of high-risk sexual behaviors. Thus, the study validates the utility of the IMB model in constructing effective social media interventions for men who have sex with men populations.

This study holds significant research and policy implications. First, by adopting an RCT design and integrating social media interventions with an established behavior change theory—the IMB model—we were able to more comprehensively account for factors influencing individual behavior change. Second, conducting the research in Guangxi, China—a relatively resource-limited region with a high HIV burden—expands the demographic and geographical diversity of studies on social media–based HIV prevention interventions among men who have sex with men. Third, by using multiple outcome measures to evaluate the intervention's effectiveness from various dimensions, we demonstrated that our approach effectively promoted HIV testing and reduced high-risk behaviors among men who have sex with men. Fourth, a 3-month postintervention follow-up was conducted, which confirmed the sustained effects of the social media intervention in promoting HIV testing and reducing high-risk behaviors. These findings provide new insights and methodologies for HIV prevention behavioral interventions.

Several limitations should be acknowledged. First, the participants were recruited from one site in China, which may limit the generalizability of the findings due to potential homogeneity within the recruited sample. Second, the survey relied on staff-administered one-on-one interviews, which might introduce discrepancies, especially for sensitive information like sexual behaviors. During follow-up, participants might have been influenced by social desirability bias, potentially underreporting high-risk behaviors and leading to report bias. Social desirability bias can distort the authenticity of data, weaken or even reverse the true relationships between variables. Third, the 6-month duration of this study limited the assessment of long-term effects postintervention. Fourth, although we aimed to reduce the risk of contamination by instructing participants to maintain confidentiality and differentiating service schedules, the use of respondent-driven sampling means that participants in the intervention and control groups may have social connections. Therefore, the possibility of intervention information spreading (contamination) cannot be entirely ruled out theoretically. Fifth, due to the nature of the intervention, complete blinding of participants was challenging, which may have allowed them to deduce their group assignment, potentially introducing expectation bias and reporting bias. Future studies should adopt a combination of strategies—such as implementing "anonymous" designs, using specialized measurement methods like the "randomized response technique," and incorporating social desirability scales for statistical control during data analysis—to systematically identify and mitigate the impact of social desirability bias, while also expanding recruitment to more diverse geographic regions to verify the effectiveness of social media–based strategies in resource-limited settings, extending the postintervention period of follow-up to examine their sustained impact on HIV prevention behaviors, considering designs such as cluster randomization or randomization based on social network communities to more effectively avoid contamination risks, and implementing a structured active control design to maintain participant blinding.

In conclusion, social media intervention based on the IMB model represented a promising strategy for HIV prevention in resource-limited settings, aimed at men who have sex with men. It significantly contributed to expanding HIV testing coverage, reducing high-risk behavior, enhancing AIDS-related knowledge, and improving safer sex self-efficacy within this high-risk population. Moreover, the positive effects became more pronounced with increasing the intervention and follow-up period. It demonstrated considerable potential and promising application prospects in the broader fight against HIV/AIDS. To achieve the 95-95-95 targets, adapting the IMB model–based social media intervention in low-resource settings and other key populations could prove highly beneficial.

Acknowledgments

We are grateful to the Nanning Center for Disease Control and Prevention (CDC), the Guangxi Caihong, and their peer educators for their invaluable support and assistance throughout this study. No generative artificial intelligence tools were used for manuscript preparation.

Funding

This work was supported by funds from the National Natural Science Foundation of China (NSFC, grant 82273694), the Natural Science Foundation of Guangxi Zhuang Autonomous Region (grants 2023GXNSFAA026287 and 2023GXNSFDA026036), and the Open Research Project from the Key Laboratory of Prevention and Control of Highly Prevalent Diseases (KLPCHPD202501). The funding source played a role in the design and conduct of the study and the preparation of the manuscript.

Data Availability

The deidentified data used and/or analyzed during the study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: HM, JL, LC
Data curation: HL, PC, QY, QC, HZ, WL, SH
Methodology: HM, JL, LC
Supervision: XY, PC, JH
Writing – original draft: HM, JL, LC
Writing – review & editing: XY, PC, JH

Conflicts of Interest

None declared.

Editorial Notice

This randomized study was only retrospectively registered, explained by authors due to delays with the registration platform's review process. The editor granted an exception from ICMJE rules mandating prospective registration of randomized trials, because the risk of bias appears low along with the authors' assertion that there were no deviations from the registered protocol once recruitment began. However, readers are advised to carefully assess the validity of any potential explicit or implicit claims related to primary outcomes or effectiveness, as retrospective registration does not prevent authors from changing their outcome measures retrospectively.

Multimedia Appendix 1
Intervention development and pilot study.
[PDF File, 1405 KB - jmir_v28i1e84279_app1.pdf]

Multimedia Appendix 2
Operational process of WeChat mini-program.
[PDF File, 1209 KB - jmir_v28i1e84279_app2.pdf]

Multimedia Appendix 3
Video tasks of the intervention plan based on the information-motivation-behavioral skills model.
[PDF File, 88 KB - jmir_v28i1e84279_app3.pdf]

Multimedia Appendix 4
Definitions of outcomes and methods of outcome assessment.
[PDF File, 155 KB - jmir_v28i1e84279_app4.pdf]

Checklist 1
CONSORT 2010 checklist.
[PDF File, 219 KB - jmir_v28i1e84279_app5.pdf]

References

1. Latest global and regional statistics on the status of the AIDS epidemic. : UNAIDS; 2025 URL: https://www.unaids.org/sites/default/files/2025-07/2025_Global_HIV_Factsheet_en.pdf [accessed 2026-03-05]
2. Dong MJ, Peng B, Liu ZF, et al. The prevalence of HIV among MSM in China: a large-scale systematic analysis. BMC Infect Dis 2019 Nov 27;19(1):1000. [doi: 10.1186/s12879-019-4559-1] [Medline: 31775654]
3. Gao Y, Hu Q, Leuba SI, et al. Medication non-adherence and condomless anal intercourse increased substantially during the COVID-19 pandemic among MSM PrEP users: a retrospective cohort study in four Chinese metropolises. Front Med (Lausanne) 2022;9:738541. [doi: 10.3389/fmed.2022.738541] [Medline: 35573017]
4. He N. Research progress in the epidemiology of HIV/AIDS in China. China CDC Wkly 2021 Nov 26;3(48):1022-1030. [doi: 10.46234/ccdcw2021.249] [Medline: 34888119]
5. Zhou Y, Lu Y, Ni Y, et al. Monetary incentives and peer referral in promoting secondary distribution of HIV self-testing among men who have sex with men in China: a randomized controlled trial. PLOS Med 2022 Feb;19(2):e1003928. [doi: 10.1371/journal.pmed.1003928] [Medline: 35157727]
6. Frescura L, Godfrey-Faussett P, Feizzadeh A A, et al. Achieving the 95 95 95 targets for all: a pathway to ending AIDS. PLOS One 2022;17(8):e0272405. [doi: 10.1371/journal.pone.0272405] [Medline: 35925943]
7. Shoptaw S. HIV prevention for people who use substances: evidence-based strategies. J Food Drug Anal 2013 Dec;21(4):S91-S94. [doi: 10.1016/j.jfda.2013.09.042] [Medline: 25214751]

XSL•FO
RenderX

8. Luo R, Xie Z, Silenzio VMB, Kuang Y, Luo D. Gay app use, sexuality traits, and high-risk sexual behaviors among men who have sex with men in China: mediation analysis. J Med Internet Res 2023 Nov 1;25:e49137. [doi: 10.2196/49137] [Medline: 37910154]
9. Johnson WD, Diaz RM, Flanders WD, et al. Behavioral interventions to reduce risk for sexual transmission of HIV among men who have sex with men. Cochrane Database Syst Rev 2008 Jul 16(3):CD001230. [doi: 10.1002/14651858.CD001230.pub2] [Medline: 18646068]
10. Romijnders K, de Groot L, Vervoort S, et al. The perceived impact of an HIV cure by people living with HIV and key populations vulnerable to HIV in the Netherlands: a qualitative study. J Virus Erad 2022 Mar;8(1):100066. [doi: 10.1016/j.jve.2022.100066] [Medline: 35280938]
11. Cao B, Gupta S, Wang J, et al. Social media interventions to promote HIV testing, linkage, adherence, and retention: systematic review and meta-analysis. J Med Internet Res 2017 Nov 24;19(11):e394. [doi: 10.2196/jmir.7997] [Medline: 29175811]
12. Young SD, Cumberland WG, Nianogo R, Menacho LA, Galea JT, Coates T. The HOPE social media intervention for global HIV prevention in Peru: a cluster randomised controlled trial. Lancet HIV 2015 Jan;2(1):e27-e32. [doi: 10.1016/S2352-3018(14)00006-X] [Medline: 26236767]
13. Pagkas-Bather J, Young LE, Chen YT, Schneider JA. Social network interventions for HIV transmission elimination. Curr HIV/AIDS Rep 2020 Oct;17(5):450-457. [doi: 10.1007/s11904-020-00524-z] [Medline: 32720253]
14. Yujie Z, Al Imran Yasin M, Alsagoff S, Hoon AL. The mediating role of new media engagement in this digital age. Front Public Health 2022;10:879530. [doi: 10.3389/fpubh.2022.879530] [Medline: 35586011]
15. Muessig KE, Bien CH, Wei C, et al. A mixed-methods study on the acceptability of using eHealth for HIV prevention and sexual health care among men who have sex with men in China. J Med Internet Res 2015 Apr 21;17(4):e100. [doi: 10.2196/jmir.3370] [Medline: 25900881]
16. Wang C, Ong JJ, Zhao P, et al. Expanding syphilis test uptake using rapid dual self-testing for syphilis and HIV among men who have sex with men in China: a multiarm randomized controlled trial. PLOS Med 2022 Mar;19(3):e1003930. [doi: 10.1371/journal.pmed.1003930] [Medline: 35235573]
17. Zhu X, Zhang W, Operario D, et al. Effects of a mobile health intervention to promote HIV self-testing with MSM in China: a randomized controlled trial. AIDS Behav 2019 Nov;23(11):3129-3139. [doi: 10.1007/s10461-019-02452-5] [Medline: 30852728]
18. Li JS, Gu YZ, Hou FS, et al. Delivery of WeChat-based HIV result e-reports in social networks for recruitment of high-risk population: baseline data from a cluster randomized controlled trial. J Med Internet Res 2023 Jun 15;25:e46793. [doi: 10.2196/46793] [Medline: 37318850]
19. Zou Y, Deng M, Wang Q, Zhang Q, Rong Y. Evolution characteristics of new urbanization in the provincial capital cities of Western China. Front Environ Sci 2022;10. [doi: 10.3389/fenvs.2022.926444]
20. Ya-Feng Z, Min D, Ya-Jing L, Yao R. Evolution characteristics and policy implications of new urbanization in provincial capital cities in Western China. PLOS One 2020;15(5):e0233555. [doi: 10.1371/journal.pone.0233555] [Medline: 32453772]
21. Prabawanti C, Dijkstra A, Riono P, Hartana G. A survey on HIV-related health-seeking behaviors among transgender individuals in Jakarta, based on the theory of planned behavior. BMC Public Health 2015 Nov 16;15:1138. [doi: 10.1186/s12889-015-2480-0] [Medline: 26573618]
22. Fishbein M. The role of theory in HIV prevention. AIDS Care 2000 Jun;12(3):273-278. [doi: 10.1080/09540120050042918] [Medline: 10928203]
23. Kalichman SC, Picciano JF, Roffman RA. Motivation to reduce HIV risk behaviors in the context of the information, motivation and behavioral skills (IMB) model of HIV prevention. J Health Psychol 2008 Jul;13(5):680-689. [doi: 10.1177/1359105307082456] [Medline: 18519441]
24. Krishnan A, Ranjit YS, Zhou X, Altice FL. Predicting antiretroviral medication adherence among substance-using people with HIV: test and extension of the information-motivation-behavioral skills model. J Behav Med 2025 Jun;48(3):403-413. [doi: 10.1007/s10865-025-00557-y] [Medline: 39924604]
25. Wan Mohamad Darani WNS, Mat Ruzlin AN, Azhar ZI, Chen XW. Development and validation of the information-motivation-behavioural skills model-based human immunodeficiency virus education kit for adolescents in Malaysia. Sci Rep 2024 Mar 22;14(1):6890. [doi: 10.1038/s41598-024-57593-y] [Medline: 38519534]
26. Chen H, Luo L, Pan SW, et al. HIV epidemiology and prevention in Southwestern China: trends from 1996-2017. Curr HIV Res 2019;17(2):85-93. [doi: 10.2174/1570162X17666190703163838] [Medline: 31269884]
27. Pang X, Wei H, Huang J, et al. Patterns and risk of HIV-1 transmission network among men who have sex with men in Guangxi, China. Sci Rep 2021 Jan 12;11(1):513. [doi: 10.1038/s41598-020-79951-2] [Medline: 33436843]
28. Zhang X, Wen D, Liang J, Lei J. How the public uses social media wechat to obtain health information in China: A survey study. BMC Med Inform Decis Mak 2017 Jul 5;17(Suppl 2):66. [doi: 10.1186/s12911-017-0470-0] [Medline: 28699549]
29. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c332. [doi: 10.1136/bmj.c332] [Medline: 20332509]
30. Liu L, Liu JH, Ruan YH, et al. Rush poppers use and associated factors among men who have sex with men in Guangxi, China. Chin J AIDS STD 2024;30(1):23-28. [doi: 10.13419/j.cnki.aids.2024.01.05]

XSL•FO
RenderX

31. Knight R, Karamouzian M, Salway T, Gilbert M, Shoveller J. Online interventions to address HIV and other sexually transmitted and blood-borne infections among young gay, bisexual and other men who have sex with men: a systematic review. J Int AIDS Soc 2017 Nov;20(3). [doi: 10.1002/jia2.25017] [Medline: 29091340]
32. Moreno MA, D'Angelo J. Social media intervention design: applying an affordances framework. J Med Internet Res 2019 Mar 26;21(3):e11014. [doi: 10.2196/11014] [Medline: 30912754]
33. Carpenter KM, Stoner SA, Mikko AN, Dhanak LP, Parsons JT. Efficacy of a web-based intervention to reduce sexual risk in men who have sex with men. AIDS Behav 2010 Jun;14(3):549-557. [doi: 10.1007/s10461-009-9578-2] [Medline: 19499321]
34. Nelson KM, Carey MP. Media literacy is an essential component of HIV prevention for young men who have sex with men. Arch Sex Behav 2016 May;45(4):787-788. [doi: 10.1007/s10508-016-0712-8] [Medline: 26895006]
35. Gyamerah AO, Taylor KD, Atuahene K, et al. Stigma, discrimination, violence, and HIV testing among men who have sex with men in four major cities in Ghana. AIDS Care 2020 Aug;32(8):1036-1044. [doi: 10.1080/09540121.2020.1757020] [Medline: 32362131]
36. Bogale GW, Boer H, Seydel ER. Effects of a theory-based audio HIV/AIDS intervention for illiterate rural females in Amhara, Ethiopia. AIDS Educ Prev 2011 Feb;23(1):25-37. [doi: 10.1521/aeap.2011.23.1.25] [Medline: 21341958]
37. Hirshfield S, Chiasson MA, Joseph H, et al. An online randomized controlled trial evaluating HIV prevention digital media interventions for men who have sex with men. PLOS One 2012;7(10):e46252. [doi: 10.1371/journal.pone.0046252] [Medline: 23071551]
38. Hennessy M, Romer D, Valois RF, et al. Safer sex media messages and adolescent sexual behavior: 3-year follow-up results from project iMPPACS. Am J Public Health 2013 Jan;103(1):134-140. [doi: 10.2105/AJPH.2012.300856] [Medline: 23153149]
39. Cheng A, Ren G, Hong T, Nam K, Koo C. An exploratory analysis of travel-related WeChat mini program usage: an affordance theory perspective. In: Information and Communication Technologies in Tourism 2019: Springer; 2019:333-343. [doi: 10.1007/978-3-030-05940-8_26]

Abbreviations

CBO: community-based organization
CONSORT: Consolidated Standards of Reporting Trials
GDP: gross domestic product
GEE: generalized estimating equation
IMB: information-motivation-behavioral skills
PEP: post-exposure prophylaxis
PrEP: pre-exposure prophylaxis
RCT: randomized controlled trial
UNAIDS: United Nations Programme on HIV/AIDS

Edited by A Mavragani; submitted 17.Sep.2025; peer-reviewed by E Ocean, X Li; accepted 30.Jan.2026; published 07.Apr.2026.

Please cite as:
Mai H, Liu J, Cheng L, Liang H, Cui P, Yao Q, Cao Q, Zhang H, Liao W, He S, Yang X, Cen P, Huang J
Social Media Intervention Based on the Information-Motivation-Behavioral Skills Model Promotes HIV Testing and Reduces High-Risk Behaviors Among Men Who Have Sex With Men in Resource-Limited Settings in China: Randomized Controlled Trial
J Med Internet Res 2026;28:e84279
URL: https://www.jmir.org/2026/1/e84279
doi:10.2196/84279

© Huanzhuo Mai, Jianyuan Liu, Liping Cheng, Hao Liang, Ping Cui, Qin Yao, Qin Cao, Heng Zhang, Wenping Liao, Shiran He, Xing Yang, Ping Cen, Jiegang Huang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Usability of iSupport Swiss, a World Health Organization Digital Intervention for Caregivers of People With Dementia: Mixed Methods Study

Anna Messina*, PhD; Rebecca Amati*, PhD; Anna Maria Annoni*, MSc; Emiliano Albanese*, Prof Dr Med, PhD; Maddalena Fiordelli*, PhD

Institute of Public Health, Faculty of Biomedical Sciences, Università della Svizzera italiana, Via Buffi 13, Lugano, Switzerland

*all authors contributed equally

Corresponding Author:
Anna Messina, PhD
Institute of Public Health, Faculty of Biomedical Sciences, Università della Svizzera italiana, Via Buffi 13, Lugano, Switzerland

Abstract

Background: Dementia presents a pressing public health challenge, with informal caregivers (ICs) playing a pivotal role in supporting people with dementia. Digital interventions, such as the World Health Organization (WHO) iSupport, offer scalable, self-guided psychosocial and educational support for caregivers. However, effective implementation relies on strong usability, particularly for older adults with varying levels of digital literacy.

Objective: This study aimed to evaluate the usability of the desktop version of iSupport Swiss, a WHO culturally adapted digital intervention for ICs of people with dementia in the Italian-speaking region of Switzerland.

Methods: We conducted a mixed methods usability study at the Università della Svizzera italiana. Participants (formal caregivers and ICs) completed a pretest questionnaire to collect sociodemographic data and digital health literacy measures via the Italian version of the eHealth Literacy Scale, 11 structured usability tasks, a posttest System Usability Scale, and a semistructured interview. Task performance was observed and analyzed, including completion rates and qualitative feedback obtained through the think-aloud technique. Interviews were analyzed using thematic analysis.

Results: Twelve caregivers (10 ICs and 2 formal caregivers) took part in the study. Most participants showed high digital proficiency (Italian version of the eHealth Literacy Scale: mean 31.17, SD 5.84) and completed most tasks successfully or with minimal support. The mean System Usability Scale score was 71.5 (SD 12.27), indicating good usability. Website features were rated positively, especially information quality (mean 4.58/5, SD 0.51) and quantity (mean 4.42/5, SD 0.51), though ease of navigation scored lower (mean 3.42/5, SD 1.00). Qualitative interviews revealed 5 core dimensions: content quality and relevance, credibility of the source, navigation and usability, interactivity, and emotional impact. Suggestions included clearer language switching, improved navigation cues, and more personalized feedback in exercises.

Conclusions: iSupport Swiss demonstrated overall good usability among caregivers, supporting its further implementation. Key areas for refinement include navigation clarity and user-tailored feedback. Participants appreciated the emotional resonance and credibility of the program, underlining its potential to enhance caregiver self-efficacy and early help-seeking. Usability testing proved essential in identifying both functional and affective dimensions of user experience. Future work should extend to broader and more diverse populations, incorporate real-world usage settings, and examine long-term engagement to optimize adoption and impact. Successful use and further development of iSupport, beyond the specific Swiss context, could contribute to expanding the availability of accessible and appropriate resources for ICs, ultimately helping to reduce caregiver burden and alleviate its impact on social and health care systems globally.

(J Med Internet Res 2026;28:e81247) doi:10.2196/81247

KEYWORDS

eHealth; dementia; caregiver; psychosocial interventions; usability

Introduction

Background

Dementia poses a significant and growing public health challenge globally [1], and Switzerland is no exception. The burden of dementia is substantial and complex, with marked implications at the family and societal levels. With over 150,000 individuals currently living with dementia [2] and projections estimating a doubling of this number by 2050, the need for effective support systems is urgent, particularly for family caregivers, also known as “informal caregivers” (ICs), who

XSL•FO
RenderX

serve as the primary source of care for people living with dementia [3].

Over the past 2 decades, digital health (eHealth) interventions have emerged as scalable and accessible means to provide psychosocial and educational support to ICs [4,5]. These solutions are considered particularly well-suited for caregivers, offering flexibility in terms of time and location, which is critical given the time and organizational constraints of dementia care and its complexity [6]. Despite the potential of eHealth interventions to alleviate the burden and improve the quality of life for both caregivers and people with dementia, several challenges may hamper their successful implementation and uptake. A scoping review of 32 studies reported an average retention rate of approximately 70%, indicating that nearly one-third of participants discontinued their involvement [7]. More recently, a systematic review including 18 studies involving more than 2000 caregivers found an overall pooled dropout rate of 18.4% [8]. In the context of the World Health Organization (WHO) iSupport adaptation in Spain [9], the authors highlighted the risk that eHealth interventions may primarily reach a narrow segment of digitally engaged users rather than the broader population of caregivers for whom they are intended. Overall, evidence on actual adoption and reach is extremely scarce, and properly designed implementation studies that test the combined effectiveness, scalability, and sustainability of eHealth interventions for caregivers are still lacking. Existing literature provides more consistent insights into the factors associated with attrition during experimental and feasibility stages. These factors can be broadly categorized into content-related challenges (eg, a gap between caregivers' needs and the solution or competing care priorities) [8] or technology-related challenges (eg, interface design, digital readiness, and ease of navigation) [10,11]. The latter may be particularly pronounced among older caregivers, who may have relatively low levels of digital literacy and technological proficiency [12].

Usability studies play a crucial role in ensuring that the interactive system is tailored to users' needs and skills and play a key role in informing implementation. According to Bastien [13], usability assessment focuses on 3 main dimensions: effectiveness, the degree to which the system achieves its intended functions; efficiency, the resources, such as time and effort, required to accomplish specific tasks; and user satisfaction, the extent to which users experience positive attitudes and responses while using the system. Additionally, usability studies are instrumental in uncovering design problems and errors, as well as in suggesting potential refinements before an effectiveness evaluation [14].

This study aims to evaluate the usability of the desktop version of iSupport Swiss [15], which is presented in detail in the "Overview of iSupport Swiss" section.

Overview of iSupport Swiss

iSupport Swiss is a culturally and linguistically adapted version of the original iSupport program developed by the WHO to provide support and education to ICs of people living with dementia [16], tailored to the Swiss context and culture [15]. The adaptation process was based on a community-based participatory approach [17], incorporating feedback from family caregivers, health care professionals, and relevant stakeholders. The active engagement of all stakeholders enabled the modification of the original content to better reflect the linguistic preferences, cultural values and preferences, informational needs, and caregiving practices specific to the Swiss population. Accordingly, adaptations were made across various domains, including language, resources, content, visual design, and functional features, to ensure cultural relevance. A detailed description of the adaptation procedures and the specific modifications implemented can be found in prior publications [15,18].

Since 2023, iSupport Swiss has been available in both online and print formats. The digital version is freely accessible as a mobile app (for smartphones and tablets) and as a desktop website. Consistent with the original WHO iSupport tool, the program consists of 22 chapters organized into 5 core modules, each addressing key aspects of caregiving. These include both practical support for care recipients (eg, assistance with toileting, personal hygiene, and nutrition) and strategies for caregiver well-being (eg, stress management and mobilizing support networks). Each chapter combines theoretical background with practical guidance, illustrated through real-life case scenarios and interactive components, such as multiple-choice and open-ended questions. Upon accessing the homepage, users are prompted to register or log in using an email address and password. Following authentication, users are asked to complete 2 brief multiple-choice questions regarding their caregiving situation and their reasons for accessing the program. Based on their responses, a personalized selection of chapters is suggested. By default, users are encouraged to begin with the introductory guide (chapter 0). They can then either proceed with the recommended content or freely browse the full list of available chapters. Screenshots from the program are available in Multimedia Appendix 1.

Methods

Ethical Considerations

Prior to starting the study, the project (2020 - 02030/CE 3731) was submitted to the Cantonal Ethics Committee, which reviewed and evaluated the study protocol. Following their assessment, the committee concluded that the project does not qualify as research involving human participants under Article 3 of the Swiss Human Research Act (Loi relative à la recherche sur l'être humain) [19], as it did not involve the collection of sensitive health-related data or clinical interventions. The committee also confirmed that no ethical concerns were identified in relation to the study procedures. All participants received written information about the study, were informed about confidentiality and data protection, and provided written informed consent prior to participation. No financial compensation or reimbursement was provided to participants for their involvement in the study.

Study Procedure and Sample

The study was conducted in the Italian-speaking region of southern Switzerland, namely the Canton of Ticino. We conducted this study in accordance with the WHO iSupport

XSL•FO
RenderX

adaptation guidelines [16], which also recommend locally testing the intervention before its final implementation. The procedures described in this study represent a direct continuation of the cultural adaptation process of iSupport in Switzerland, the methodology of which has already been detailed elsewhere [15].

Before conducting the usability study, we pretested the usability tasks with the members of the iSupport Community Advisory Board (CAB). Members of the CAB included caregivers of people with dementia, representatives of the project's funding agencies, and other national collaborating partners who had participated in the previous cultural adaptation process of iSupport Swiss.

For this study, we recruited both formal caregivers (FCs) and ICs: the former as prospective target users of iSupport and the latter as potential facilitators and communicators of the program. In January 2022, we drafted an invitation letter and distributed it to the members of the CAB and their associations and institutions, as well as to participants from previous phases of iSupport adaptation or other ongoing projects who had consented to receive information about further research activities. Eligibility criteria for ICs included: (1) having (at present or in the past) experience in caring for a person with dementia, (2) being fluent in Italian, and (3) being familiar with digital technology. Inclusion criteria for FCs included: (1) having experience in dementia care, (2) being fluent in Italian, and (3) being familiar with technology. We sent the informed consent forms to participants who contacted us and met the inclusion criteria. Once we received the signed consent forms, we called them to arrange their sessions.

Data Collection

All participants' sessions were carried out in person in a laboratory setting at the Università della Svizzera italiana between January 28, 2022, and February 10, 2022. The sessions were conducted by 1 facilitator with expertise in iSupport (AM) and 1 observer experienced in qualitative research methods (IF). We provided participants with a personal desktop computer with internet access.

The usability sessions consisted of four distinct steps: (1) pretest questionnaire, (2) usability tasks, (3) posttest questionnaire, and (4) posttest interview. The methodology combined quantitative and qualitative approaches, implemented sequentially across the usability sessions.

Quantitative methods were used during the first 3 phases to collect sociodemographic information and to assess participants' interaction with the system during task performance, including task completion, need for assistance, and usability ratings.

Qualitative methods were applied both during and after the usability tasks. During task performance, qualitative data were collected through the think-aloud technique and observational notes to capture immediate, in-task feedback ("real-time" reflections) on users' interactions with the system.

Following task completion, qualitative methods were further applied in the final phase through 4 open-ended questions, allowing an in-depth exploration of participants' subjective experiences, perceptions, and emotional responses.

During the pretest questionnaire, we asked participants to fill out an online questionnaire via REDCap (Research Electronic Data Capture; Vanderbilt University) [20]. Through this survey, we collected sociodemographic characteristics (age, sex, occupational status, and type of caregiver), caregiving data (living situation of people with dementia and the time spent caregiving), and measured the frequency of average internet use (more than once a day, at least once a week, or less than once a week), the type of devices used (smartphone, tablet, and computer), and health digital literacy. The latter was measured through the validated Italian version of the eHealth Literacy Scale (IT-eHEALS) [21], an 8-item standardized scale developed to measure users' skills and knowledge in finding, evaluating, and using eHealth information to address health issues [22].

For the usability tasks, we based our procedures on the testing protocol of the iSupport-Portuguese version [23] and on previous studies and recommendations in co-designing health interventions [13,24,25].

As part of the usability test material, we provided participants with a hard copy guide that described 11 tasks (Multimedia Appendix 2) to be performed using the iSupport website. The tasks ranged from easier tasks (eg, entering the website and completing the login procedure) to more complex activities (eg, navigating a chapter and completing exercises).

Participants were instructed to try to perform the task autonomously, with the possibility to ask for support from the facilitator (AM), and to verbalize their navigation behavior by sharing impressions and comments, according to the "think-aloud technique" [26]. Assistance was not provided unless explicitly requested by the participant, in order to replicate behavior in a more realistic and naturalistic context. In parallel, the observer (IF) filled a structured observation sheet to collect the following data: task completion (categorized as concluded, partially concluded, concluded with support, or not concluded), task completion time (minutes taken to complete the task), main difficulties encountered, and user verbalization. We recorded participants' navigation behavior and commentary using audio recorders and screen-capturing software (Panopto).

After completing the tasks, participants were asked to fill out a posttest questionnaire on REDCap. We measured usability with an Italian back-translated version of the System Usability Scale (SUS) [27], a widely used 10-item Likert scale that measures the subjective assessment of a system's usability. The final SUS score ranges from 0 to 100, with higher values representing better usability, and a standard cutoff of 68 for defining a usability score as above or below average [27].

Based on literature recommendations [26,28], we also provided participants with a list of 34 adjectives that users could choose from to best describe the website, along with a rating scale (1-5) for website features (colors, font type and size, quantity and quality of information, ease of navigation and reading, and design and graphics). The choice of adjectives was informed by the User Experience Questionnaire by Schrepp [29].

The final phase included 4 open-ended questions aimed at gathering participants' perspectives on the usability and

XSL•FO
RenderX

usefulness of the website. More specifically, participants were asked what they had liked and what they had not liked about the platform, whether they would use it and why, and if they would recommend it to others. They were also asked how using the platform had made them feel and whether they had any additional comments. Given the targeted scope of the interview and the limited sample, interviews were conducted with all participants included in the study.

During the tasks, the observer asked follow-up questions to further investigate and clarify specific observations noted throughout the sessions. This final qualitative phase allowed us to complement and deepen the interpretation of the usability findings, providing insights into observed task performance patterns and usability scores within participants' subjective perspectives.

All interviews were audio recorded, transcribed verbatim, and analyzed using an inductive thematic analysis approach [30].

Following transcription, 2 researchers (AM and AMA) independently familiarized themselves with the data, generated initial codes, and iteratively grouped them into candidate themes. Themes were then reviewed, refined, and agreed upon through discussion to reach consensus. A high degree of convergence in participants' responses was observed across interviews, suggesting that the collected qualitative data were sufficient to inform the usability evaluation and complement quantitative findings. Integration of qualitative and quantitative data occurred during the interpretation phase, by comparing task performance metrics, usability scores, and website ratings with qualitative findings from think-aloud comments and interviews.

Results

Characteristics of Participants and Context of Care

Twelve participants were involved in the study: 8 women and 4 men, with a mean age of 56.5 (SD 9.64) years. Most participants (n=10) were ICs, providing care for a family member with dementia. Two participants were FCs, more specifically home nurses. Seven cared for an older adult living independently, while in the other cases, the person with dementia lived in the caregiver's home (n=2) or was in a care facility (n=1). Seven participants had between 1 and 2 years of caregiving experience, while 2 had over 10 years of experience. Most caregivers accessed the internet more than once a day (n=10), using either a computer (n=6), a tablet (n=4), or a smartphone (n=2).

The mean score on the IT-eHEALS was 31.17 (SD 5.84; range 23 - 40), indicating a high level of proficiency in locating, evaluating, and using electronic health information for decision-making. Table 1 summarizes the main characteristics of participants, including scores from the IT-eHEALS.

XSL•FO
RenderX

Table . Characteristics of formal and informal caregivers of people with dementia in the usability study (N=12).

Characteristics	Value
Age (y), mean (SD)	56.5 (9.9)
Female, n (%)	8 (66.7)
Occupational status, n (%)	
Employed	8 (66.7)
Housewives	2 (16.7)
Retired	1 (8.3)
Missing	1 (8.3)
Caregiver, n (%)	
Formal	2 (16.7)
Informal	10 (83.3)
People with dementia alive, n (%)	7 (58.3)
People with dementia living situation, n (%)	
Community-dwelling	7 (58.3)
Caregiver's home	2 (16.7)
Nursing home	1 (8.3)
Missing	2 (16.7)
Caregiving experience (y), n (%)	
1 - 2	7 (58.3)
3 - 5	2 (16.7)
6 - 10	1 (8.3)
>10	2 (16.7)
Internet use, n (%)	
More than once a day	10 (83.3)
At least once a week	2 (16.7)
Device, n (%)	
Smartphone	2 (16.7)
Tablet	4 (33.3)
Computer	6 (50)
IT-eHEALS[a], mean (SD)	31.17 (5.84)

[a]IT-eHEALS: Italian version of the eHealth Literacy Scale.

Task Performance

Both FCs and ICs successfully or partially completed most tasks, with a success rate exceeding 70%. Participants who required assistance to complete a task were classified as having partially completed the task or completed it with support. Partial or no success was mainly related to task 2 (changing language options after the login procedure), task 5 (going to the homepage using a shortcut), task 6 (distinguishing the difference between the modules and chapters and clicking on the respective links), and task 10 (returning to the previous chapter using a shortcut). Task performance is summarized in Table 2.

XSL•FO
RenderX

Table . Task performance of formal and informal caregivers using the iSupport Swiss website.

Task	Completed, n (%)	Partially completed, n (%)	Noncompleted, n (%)	Task completion time in minutes	Type of errors
1. Go to the login page	12 (100)	__[a]	—	<1	• N/A[b]
2. Login procedure and change language	9 (75)	3 (25)	—	2	• Unable to find the language settings
3. Fill the navigation survey	11 (92)	1 (8)	—	2	• Unable to provide only one answer • Difficulty to see the button "send the survey"
4. Browse the list of suggested chapters	12 (100)	—	—	6	• N/A
5. Going to the homepage	7 (58)	4 (33)	1 (8)	4	• Unable to go take the shortcut from the opened chapter to the homepage
6. Select a module and a course	4 (33)	4 (33)	4 (33)	6	• Unable to understand the difference between modules and chapters and to click on the respective links
7. Open and read a chapter	10 (83)	2 (17)	—	6	• N/A
8. Change chapter	8 (67)	3 (25)	3 (25)	1	• N/A
9. Fill an exercise or activity	11 (92)	1 (8)	—	9	• Unclear whether the answers are read by someone or not • Exercises are funny and useful as positive reinforcement • The chance to have feedback on the exercises is appreciated but not so visible
10. Return to the previous chapter	6 (50)	6 (50)	—	7	• Unable to take the shortcut
11. Logout procedure	8 (67)	4 (33)	—	<1	• N/A

[a]Not applicable.

[b]Not available.

The think-aloud procedure allowed us to collect impressions and real-time feedback during the task performances. For instance, 1 participant commented on their first impression while landing on the login page (task 1): "Yellow gives you the strength to act; the color is balanced" (ID4, IC).

On the same task, 2 participants suggested adding a sentence detailing the topic of the program:

The colors are neutral; a welcome message is important [ID5, FC]

There is no initial explanation of what it is, who it is for, I suggest putting in large font "for family caregivers…." [ID2, IC]

On task 2, few participants took the time to change languages:

XSL•FO
RenderX

> *I expected to find macro themes on this page,…the language icon should be standardized.* [ID3, IC]
>
> *I don't understand how to proceed, it is not clear how to change the language.* [ID6, IC]

Task 3 required participants to answer 2 multiple closed-ended questions to suggest a thematic path. While some participants found them useful for focusing on their situation, others found them not enough and expected more:

> *The questions are helpful, they help me understand where I am.* [ID7, IC]
>
> *The second question already requires a level of self-reflection that caregivers might not have, I was expecting more guiding questions....* [ID9, FC]

Task 4 required participants to browse the list of suggested chapters based on the answers provided during the previous task. While first impressions were generally positive, some participants did not realize that this was a suggested thematic path and not a comprehensive list of all chapters. Comments included, "Great, let's get straight to the point!" (ID4, IC) and "It's not very clear that this is a selection" (ID12, IC).

Consequently, some caregivers became confused when they were asked to go to the homepage (task 5):

> *Ah, I didn't realize this was the homepage. It should be made clearer, perhaps with an overlay.* [ID4, IC]
>
> *A person who is usually tired, like a caregiver, should be able to use the program easily.* [ID2, IC]

In addition, some participants did not understand the difference between a module and a chapter (task 6):

> *I didn't really understand the difference between "Module" and "Chapter." The buttons to enter should be larger and clearer.* [ID10, IC]

However, when participants landed on the selected chapter (task 6), first impressions were generally very positive, even though the quantity of text emerged as a potential issue:

> *Wow, it seems like everything is there, though it's a bit too much written.* [ID7, IC]

Using shortcuts to return to previous chapters (task 10) or to move to others (task 7) was also perceived as complicated by some participants:

> *It should be clearer how to go back.* [ID5, FC]
>
> *It should be more intuitive.* [ID2, IC]

The exercises and activities were found easy to be carried out (task 9). While some participants appreciated the fact that automatic answers were provided, others would have appreciated more personalized feedback, especially for the open-ended questions:

> *That's fun! It makes me want to try all the answers too!* [ID7, IC]
>
> *It's nice that no one is judging you, because it's just you and your computer.* [ID6, IC]
>
> *Who is interested in what I'm writing? It would be nice if it could link to something else or to a specific answer.* [ID1, IC]

System Usability Scale

The average usability score using the SUS was 71.50 (SD 12.27), showing a good perception of iSupport Swiss usability. The most appreciated characteristic of the program was the perceived learnability of the website (mean score 4.00, SD 0.74; item 7: "I would imagine that most people would learn to use this system very quickly"), while the least appreciated was the perceived consistency and integration of functionalities (mean score 2.67, SD 0.07; item 6: "I thought there was too much inconsistency in this system").

Website Characteristics

Participants positively rated the website characteristics, with a mean score of 4.17 (SD 0.49) out of 5. The most appreciated features were related to the quality and quantity of information provided in the program with mean (SD) scores of 4.58 (0.51) and 4.42 (0.51), respectively. The least appreciated features were related to ease of navigation (mean score 3.42, SD 1.00) and ease of reading (mean score 4.08, SD 0.79). Table 3 summarizes scores from website characteristics and the SUS.

Table . Website characteristics and System Usability Scale (SUS) scores after using iSupport Swiss.

Characteristics	Value, mean (SD)
SUS[a]	71.50 (12.27)
Website characteristics[b]	31.17 (5.84)
Quality of information	4.58 (0.51)
Quantity of information	4.42 (0.51)
Design and graphics	4.33 (0.65)
Ease of reading	4.08 (0.79)
Ease of navigation	3.42 (1.00)

[a]Scores range from 0 to 100.

[b]Scores range from 1 to 5.

XSL•FO
RenderX

List of Adjectives

Adjectives selected by participants to describe their first impressions about the program conveyed a general positive experience.

Out of a total of 50 impressions, the most selected attributes were “interesting” (75%) and “accessible” (50%), followed by “clear” (33%), “comprehensible” (33%), “engaging” (33%), and “structured” (33%). Only 1 participant selected negative adjectives such as “ambiguous,” “boring,” and “frustrating.”

Open-Ended Questions

General Findings

Five main areas were identified through the analysis of postusability interviews: quality and relevance of the content, credibility of the source, navigation and usability, interactivity, and emotional impact. Table 4 summarizes the profiles of caregivers.

Table . Profile of caregivers by sex, age, type, and years of experience.

ID	Sex	Age (y)	Type of caregiver	Caregiver experience (y)
1	Female	58	Informal	6-10
2	Female	59	Informal	3-5
3	Female	75	Informal	6-10
4	Female	56	Informal	1-2
5	Male	32	Formal	6-10
6	Female	52	Informal	1-2
7	Male	57	Informal	1-2
8	Female	56	Informal	1-2
9	Male	55	Formal	>10
10	Male	59	Informal	6-10
11	Female	56	Informal	6-10
12	Female	63	Informal	1-2

Quality and Relevance of the Content

Both ICs and FCs particularly appreciated the purpose of the website, describing it as useful and innovative compared to existing resources.

All participants referred that they would use (or would have used) iSupport in the past, especially at the beginning of their caregiving journey:

> *I would navigate it for a while…I would have used it at the initial stage; it would have been much more interesting because you are not prepared.* [ID1, IC]

More specifically, the selection of topics and content was perceived as comprehensive, accessible, and closely aligned with their experiences, as highlighted by several participants:

> *It is simple and reassuring because it provides advanced answers but presents them in a clear and accessible way, supporting those who want to learn more…we need this kind of program.* [ID10, IC]
>
> *I like the purpose…there is a need that may remain silent in family caregivers.* [ID5, FC]
>
> *There is a great selection, the chapters are not too overwhelming, and you have covered the essential points.* [ID3, IC]

All participants would recommend iSupport Swiss to other caregivers, especially younger ones, noting that older caregivers might need assistance navigating the website:

> *Certainly, I would recommend it to everyone, especially younger individuals, such as children, dealing with dementia, or even older ones, but with an introduction to guide them through.* [ID8, IC]

Credibility of the Source

Additionally, 2 participants highlighted the credibility of the source as a key strength, noting that the fact that the program was created by the WHO and developed by the university contributed to a sense of trust and reassurance:

> *Knowing that there is a credible website coming from the university to seek information is reassuring.* [ID11, IC]

Navigation and Usability

Participants generally found the website functional and accessible; however, several usability issues emerged. Critical feedback mainly concerned navigation issues, as highlighted by participants during task performance. Some participants noted that it was not intuitive to switch from one module to another or to understand the differences between modules and chapters:

> *Some navigation steps are a bit cumbersome…but, like many websites, with trial and error, you can find the information.* [ID9, FC]
>
> *Understanding how the chapters and subchapters work can be challenging; you can’t expect to arrive and immediately find an answer….* [ID2, IC]

XSL•FO
RenderX

The same participant also suggested including a navigation guide to facilitate the navigation across the website:

> *It's a bit difficult to jump back and forth between the chapters; I would suggest a navigation guide.* [ID2, IC]

Interactivity

The inclusion of closed-ended and open-ended questions was generally appreciated by most participants for the level of engagement and feedback regarding their roles as caregivers. For example, a user highlighted the potential of including case scenarios to rethink her own experiences:

> *I really like the questions about situations that happen in daily life and receiving feedback…it's important to reflect on oneself.* [ID6, IC]

However, some participants expressed a preference for more personalized responses, particularly to open-ended questions:

> *It would be helpful if someone provided a response in the exercises or if there were an option to share one's own experience.* [ID4, IC]

In addition, both FCs who took part in the study expressed their belief that they could assist and guide other caregivers at home in accessing the platform. However, they emphasized the need for additional time to do so effectively, as noted by a participant:

> *I could provide more input on complex situations and support the caregiver to help them understand how it works, as a consultant…the problem is always time, which is barely enough to do what needs to be done.* [ID9, FC]

Emotional Impact

In general, the overall user experience was positive: several users emphasized the potential of iSupport to provide relief and enhance self-efficacy as caregivers, while also facilitating a reframing of past experiences, as well as being documented by 3 participants:

> *It was a bit like going back in time and, in the end, telling myself, "Many things could have been better, but it didn't turn out that bad."* [ID12, IC]

> *It made me feel good…based on the responses I received, I got confirmation that I'm on the right track, I give myself a pat on the back.* [ID1, IC]

> *It was relaxing…the images and feedback give you the feeling that there is always a solution.* [ID6, IC]

Integrated Synthesis of Quantitative and Qualitative Findings

The integration of quantitative and qualitative findings provides a comprehensive understanding of the usability of iSupport Swiss by linking observed task performance and usability scores with users' lived experiences and perceptions.

Overall, quantitative task performance data indicated that most participants were able to complete the majority of tasks successfully or with minimal support, with success rates exceeding 70% for most activities. Qualitative data from think-aloud comments and interviews helped explain these results, revealing that participants generally perceived the platform as useful, credible, and reassuring, despite encountering specific navigation challenges.

In particular, tasks associated with lower completion rates or a higher need for assistance, such as changing language settings, navigating between modules and chapters, and using shortcuts, were reflected in qualitative feedback describing confusion related to information hierarchy, labeling, and navigation cues.

These qualitative insights contextualize the lower quantitative scores observed for ease of navigation and system consistency, highlighting concrete interface elements that limited smooth interaction with the program.

Conversely, high completion rates for tasks related to reading content and completing exercises were aligned with qualitative findings emphasizing the clarity, relevance, and emotional resonance of the content. Participants' comments illustrated how interactive exercises and case-based examples supported reflection and self-efficacy, complementing quantitative indicators of good learnability and information quality.

This integrated analysis demonstrates how quantitative usability metrics identified where usability issues occurred, while qualitative findings explained why these issues emerged and how they affected users' experiences.

Discussion

Principal Findings

This study aimed to assess the usability of iSupport Swiss (desktop version), a culturally adapted digital intervention designed to support ICs of people living with dementia. Overall, our findings indicate that iSupport Swiss is a user-friendly and valuable resource, with clear, credible content tailored to caregivers' needs. Both FCs and ICs reported that the platform could complement existing support options, although several usability challenges, particularly related to navigation and personalization, were identified.

Usability scores indicated that the platform meets core usability criteria [27] and that participants found iSupport Swiss to be user-friendly and acceptable overall [27]. Moreover, task performance rates were above 70% for most tasks, suggesting an overall positive navigation experience. These results are consistent with findings on the usability of iSupport from Portugal and Brazil [23,31].

Nevertheless, several usability issues emerged during the task performance test, particularly regarding the ease of navigation and the content presentation. Specific tasks, such as modifying language settings, differentiating between modules and chapters, and using shortcuts to navigate back or forth, were associated with lower success rates or required support from the researchers. These difficulties were reflected in both behavioral observations and verbal feedback, pointing to challenges related to interface design and information hierarchy. At first glance, this may appear counterintuitive given that the digital literacy levels in our sample were relatively high, indicating that most participants possessed adequate skills to search, evaluate, and

XSL•FO
RenderX

apply online health information. However, individual differences in eHealth literacy may help explain some of the usability challenges observed, particularly among those who required assistance during navigation tasks. Previous studies have similarly reported that even individuals with moderate-to-high digital literacy may encounter difficulties when engaging with novel or content-rich eHealth platforms, especially when the information structure or navigation paths are complex [22,32]. These findings suggest that future implementation strategies should consider providing introductory guidance or brief digital literacy reinforcement, particularly for older or less experienced caregivers, to ensure equitable access and sustained engagement with the program.

The think-aloud technique proved its anticipated complementarity with the other approaches we used. Caregivers commented freely on challenges encountered and offered suggestions to overcome navigation issues. Some of the most frequently reported issues were addressed following this study modality. For example, a welcome message was added to the login page to provide clearer orientation, and visual distinctions between modules and chapters were improved to enhance content clarity.

The posttest interviews complemented the qualitative findings by integrating qualitative insights into users' experiences. More specifically, we collected diverging preferences regarding the interactive features of iSupport Swiss, highlighting the need for user-tailored integrations and personalization. Some participants appreciated the self-paced nature of the program and the lack of judgment or external pressure. Others expressed the desire for more personalized feedback and interaction with professionals, a finding that had already emerged during the cultural adaptation of iSupport Swiss [15].

Personalization is a well-discussed topic in the context of psychoeducational self-help programs. Some authors have argued that incorporating social components, such as human-based interaction, may improve the adherence of caregivers of people with dementia in using online interventions by mimicking analogical, humanized intervention-to-user interaction modalities [8].

However, iSupport was originally conceived by the WHO and was intended to function as a self-help tool that users can access at their own convenience and that does not require any additional support from external sources, under evidence-informed posits of greater flexibility of access and use, practicality, sustainability, and reduced stigma and inequality. Introducing personalized or human-mediated support would require substantial modifications to the program's core structure and purpose, as well as significant investments in both technical and financial resources. Considering these constraints, with the goal of preserving the fidelity of the original program and in keeping with its assumed different potential reach through an eHealth platform, we decided to implement 2 measures. First, a disclaimer was added at the end of each exercise, clarifying that the automatic responses provided may not fully reflect the user's individual situation. Second, updated and expanded links to external support resources were implemented, enabling users to seek more tailored assistance if needed.

Our study provides corroboration for the future implementation of iSupport. Participants highlighted the program's clear and accessible language, comprehensive coverage of caregiver-relevant topics, and its overall credibility due to its institutional affiliations. These elements were perceived as reassuring and valuable, especially in contrast to existing resources for caregivers, which are often fragmented.

Importantly, the emotional dimension of usability also emerged as a key theme. Participants described feelings of reassurance, relief, and validation in their roles as caregivers when engaging with the content. The design and tone of the intervention appeared to foster a sense of self-efficacy and support, aligning with prior research emphasizing the role of affective responses in technology acceptance and sustained engagement [33].

While the digital format was generally well received, some participants emphasized the need for a multimodal approach to program dissemination, such as including a navigation guide or receiving support through health care professionals. In this regard, the engagement of health care professionals as facilitators of the program may represent a promising implementation strategy, though constraints related to time and workload must be addressed, as expressed by both FCs who took part in our study.

Finally, participants recommended iSupport Swiss, especially to caregivers who are caring for someone in the initial stage of dementia. This finding closely aligns with those described in the usability study of iSupport in Portugal [23] and during the adaptation work of iSupport Swiss [18].

This finding contrasts with existing literature showing that caregivers often tend to seek help only at later stages of the disease, as dementia progresses [34]. This discrepancy highlights an important opportunity: by promoting the dissemination of iSupport Swiss and encouraging its use earlier in the caregiving process, it may be possible to foster timely help-seeking behaviors and prevent the escalation of caregiver burden over time. Adjusting iSupport to the needs of caregivers at the early stages and facilitating access through the engagement of key stakeholders, such as family physicians and memory clinics, could therefore enhance its preventive impact and long-term effectiveness.

Limitations

This study is not without limitations. First, the small sample size and the convenience sampling approach may limit the generalizability of the findings to the broader population of caregivers.

However, the sample size is consistent with those used in other usability studies of iSupport and eHealth interventions more broadly [13,23]. Moreover, the reduced number of participants allowed for an in-depth examination of each user's interaction with the platform, enabling us to capture nuanced insights into the usability experience that might be overlooked in larger-scale evaluations.

Second, only 2 FCs were included in the study, which further limits the generalizability of the findings to this specific group. However, it is important to note that iSupport Swiss is primarily

XSL•FO
RenderX

designed for ICs, who represent the main target audience of the intervention. In this sense, the limited number of FCs reflects the study's focus. Nonetheless, their perspectives were particularly valuable, as they provided insights into how professional caregivers could play a supportive role in promoting, communicating, and integrating the use of iSupport among family caregivers.

Third, all participants were Italian-speaking and based in Ticino, a single region of Switzerland, which may not reflect the experiences of caregivers in other linguistic or cultural contexts. Nevertheless, it is important to note that this study was part of a broader project focused on the adaptation of iSupport to the Swiss context. As such, it serves as a foundational step to inform the development of a culturally relevant and linguistically accessible national intervention.

Fourth, regarding the measurement tools, it should be acknowledged that the use of self-reported questionnaires, such as the SUS and IT-eHEALS, may be subject to response bias or social desirability effects. Similarly, while the thematic analysis was conducted rigorously and independently by multiple researchers, qualitative interpretation inevitably involves some degree of subjectivity. Future work could combine these methods with more objective measures of user engagement, such as log-data analytics or longitudinal follow-ups, to further strengthen the validity of the findings.

Finally, the laboratory setting may not fully replicate the real-world conditions under which users engage with the platform, potentially influencing task performance and user feedback. At the same time, the controlled setting offered important methodological advantages: it enabled us to observe participants' interactions in real time, ensure consistency across sessions, and promptly identify usability issues such as interface bugs or navigation errors. This allowed for a more accurate comprehension of the user experience and ensured that user difficulties were attributable to design features rather than external distractions or contextual variability.

Conclusions and Future Directions

Our study contributes to and enriches the growing body of international research on iSupport, offering specific insights from the Swiss Italian-speaking context that complement findings from other implementation efforts in countries such as Spain [35], Portugal [23], Brazil [31], Australia [36], China [37], and the United Kingdom [38].

Our findings provide preliminary evidence supporting the usability and acceptability of iSupport Swiss among its target users. Several actionable insights were identified, particularly related to navigation, content clarity, and personalization. Future iterations of the platform should address these areas to ensure successful scaling and implementation. Additional research that involves more diverse samples, is conducted in naturalistic settings, and includes longitudinal assessments is needed to explore the sustained use and long-term impact of the intervention.

Beyond the iSupport framework, these findings have broader implications for the design, adaptation, and implementation of digital health interventions targeting caregivers and other vulnerable populations. They highlight the importance of combining rigorous usability testing with participatory and context-sensitive approaches to ensure that eHealth tools are not only technically functional but also culturally relevant, emotionally supportive, and accessible to users with diverse digital competencies. Our results suggest that successful digital interventions for caregivers must balance autonomy and guidance, offering self-paced learning while providing clear orientation and pathways for personalized support when needed. Moreover, the involvement of professional caregivers and health care providers as potential facilitators underscores the value of interdisciplinary collaboration in digital health innovation.

Ultimately, usability testing represents not only a methodological step in technology development but also a strategic investment in digital equity, ensuring that technological innovation translates into meaningful, inclusive, and sustainable support for caregivers and the communities they serve.

Acknowledgments

The authors would like to express their sincere gratitude to all the caregivers who generously dedicated their time and shared their experiences to participate in this study. The authors also thank IF for her crucial support in organizing and conducting the usability sessions. Generative artificial intelligence (ChatGPT, OpenAI) was used to improve the fluency and readability of the English text. The authors reviewed and verified all artificial intelligence–assisted edits, and all intellectual and scientific content was fully conceived, written, and approved by the authors.

Funding

The project is funded by the Ticino Canton (Divisione dell'azione sociale e delle famiglie), Pro Senectute, and Università della Svizzera italiana.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to privacy and confidentiality restrictions, as the qualitative data contain potentially identifiable information. However, anonymized excerpts supporting the study findings are included in the manuscript and its appendices. Additional deidentified materials (eg, coding framework and summary task data) are available from the corresponding author upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: MF, RA, AM
Formal analysis: AM, AMA
Investigation: AM
Methodology: MF, AM
Supervision: MF, RA, AMA
Writing – original draft: AM
Writing – review & editing: MF, RA, AMA, EA

Conflicts of Interest

None declared.

Multimedia Appendix 1
Screenshots from iSupport Swiss website version.
[DOCX File, 1979 KB - jmir_v28i1e81247_app1.docx]

Multimedia Appendix 2
User guide during the usability session.
[PDF File, 259 KB - jmir_v28i1e81247_app2.pdf]

References

1. World Alzheimer Report 2022—life after diagnosis: navigating treatment, care and support. : Alzheimer's Disease International; 2022 URL: https://www.alzint.org/u/World-Alzheimer-Report-2022.pdf [accessed 2026-03-24]
2. La demenza in svizzera [Webpage in Italian]. Alzheimer Schweiz. URL: https://www.alzheimer-schweiz.ch/it/sulla-demenza/articolo/la-demenza-in-svizzera [accessed 2026-03-24]
3. Shah R, Salek MS, Ali FM, et al. Dementia and its profound impact on family members and partners: a large UK cross-sectional study. Alzheimer Dis Assoc Disord 2024;38(4):338-343. [doi: 10.1097/WAD.0000000000000647] [Medline: 39506214]
4. Hopwood J, Walker N, McDonagh L, et al. Internet-based interventions aimed at supporting family caregivers of people with dementia: systematic review. J Med Internet Res 2018 Jun 12;20(6):e216. [doi: 10.2196/jmir.9548] [Medline: 29895512]
5. Zhai S, Chu F, Tan M, Chi NC, Ward T, Yuwen W. Digital health interventions to support family caregivers: an updated systematic review. Digit Health 2023;9:20552076231171967. [doi: 10.1177/20552076231171967] [Medline: 37223775]
6. Leng M, Zhao Y, Xiao H, Li C, Wang Z. Internet-based supportive interventions for family caregivers of people with dementia: systematic review and meta-analysis. J Med Internet Res 2020 Sep 9;22(9):e19468. [doi: 10.2196/19468] [Medline: 32902388]
7. Wu KC, Su Y, Chu F, Chen AT, Zaslavsky O. Behavioral change factors and retention in web-based interventions for informal caregivers of people living with dementia: scoping review. J Med Internet Res 2022 Jul 7;24(7):e38595. [doi: 10.2196/38595] [Medline: 35797100]
8. Atefi GL, Koh WQ, Kohl G, et al. Adherence to online interventions for family caregivers of people with dementia: a meta-analysis and systematic review. Am J Geriatr Psychiatry 2024 Oct;32(10):1271-1291. [doi: 10.1016/j.jagp.2024.04.008] [Medline: 38735829]
9. Teles S, Paúl C, Viana J, et al. iSupport-Portugal: challenges and Insights in designing a web platform for intervention and research on informal dementia caregivers. Presented at: 11th International Conference on Information and Communication Technologies for Ageing Well and e-Health; Apr 6-8, 2025 URL: http://www.scitepress.org/DigitalLibrary/ProceedingLink.aspx?ID=1901 [accessed 2026-03-24] [doi: 10.5220/0013328100003938]
10. Christie HL, Schichel MCP, Tange HJ, Veenstra MY, Verhey FRJ, de Vugt ME. Perspectives from municipality officials on the adoption, dissemination, and implementation of electronic health interventions to support caregivers of people with dementia: inductive thematic analysis. JMIR Aging 2020 May 13;3(1):e17255. [doi: 10.2196/17255] [Medline: 32401217]
11. Zwingmann I, Dreier-Wolfgramm A, Esser A, et al. Why do family dementia caregivers reject caregiver support services? Analyzing types of rejection and associated health-impairments in a cluster–randomized controlled intervention trial. BMC Health Serv Res 2020 Feb 14;20(1):121. [doi: 10.1186/s12913-020-4970-8] [Medline: 32059724]
12. Brodaty H, Donkin M. Family caregivers of people with dementia. Dialogues Clin Neurosci 2009;11(2):217-228. [doi: 10.31887/DCNS.2009.11.2/hbrodaty] [Medline: 19585957]
13. Bastien JMC. Usability testing: a review of some methodological and technical aspects of the method. Int J Med Inform 2010 Apr;79(4):e18-e23. [doi: 10.1016/j.ijmedinf.2008.12.004] [Medline: 19345139]
14. Rubin J, Chisnell D. Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests, 2nd edition: Wiley; 2008.

XSL•FO
RenderX

15. Messina A, Amati R, Annoni AM, Bano B, Albanese E, Fiordelli M. Culturally adapting the World Health Organization digital intervention for family caregivers of people with dementia (iSupport): community-based participatory approach. JMIR Form Res 2024 Jan 24;8:e46941. [doi: 10.2196/46941] [Medline: 38265857]
16. Pot AM, Gallagher-Thompson D, Xiao LD, et al. iSupport: a WHO global online intervention for informal caregivers of people with dementia. World Psychiatry 2019 Oct;18(3):365-366. [doi: 10.1002/wps.20684] [Medline: 31496096]
17. Newman SD, Andrews JO, Magwood GS, Jenkins C, Cox MJ, Williamson DC. Community advisory boards in community-based participatory research: a synthesis of best processes. Prev Chronic Dis 2011 May;8(3):A70. [Medline: 21477510]
18. Messina A, Amati R, Albanese E, Fiordelli M. Help-seeking in informal family caregivers of people with dementia: a qualitative study with iSupport as a case in point. Int J Environ Res Public Health 2022 Jun 19;19(12):7504. [doi: 10.3390/ijerph19127504] [Medline: 35742751]
19. SR 810.30 - Federal Act of 30 September 2011 on Research Involving Human Beings (Human Research Act, HRA). Fedlex. URL: https://www.fedlex.admin.ch/eli/cc/2013/617/en [accessed 2026-04-04]
20. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
21. De Caro W, Corvo E, Marucci AR, Mitello L, Lancia L, Sansoni J. eHealth Literacy Scale: an nursing analisys and Italian validation. Stud Health Technol Inform 2016;225:949. [Medline: 27332425]
22. Norman CD, Skinner HA. eHEALS: the eHealth Literacy Scale. J Med Internet Res 2006 Nov 14;8(4):e27. [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
23. Teles S, Paúl C, Ferreira A, Center for Health Technology and Services Research (CINTESIS). Acceptability of an online training program for dementia caregivers: results from a mixed-methods pilot randomized controlled trial. Alzheimers Dement 2021 Dec;17(S11). [doi: 10.1002/alz.052309]
24. Amann J, Fiordelli M, Scheel-Sailer A, Brach M, Rubinelli S. Opportunities and challenges of a self-management app to support people with spinal cord injury in the prevention of pressure injuries: qualitative study. JMIR Mhealth Uhealth 2020 Dec 9;8(12):e22452. [doi: 10.2196/22452] [Medline: 33295876]
25. Cho H, Yen PY, Dowding D, Merrill JA, Schnall R. A multi-level usability evaluation of mobile health applications: a case study. J Biomed Inform 2018 Oct;86:79-89. [doi: 10.1016/j.jbi.2018.08.012] [Medline: 30145317]
26. Nielsen J. Usability Engineering: Morgan Kaufmann; 1994.
27. Brooke J. SUS: a “quick and dirty” usability scale. In: Jordan PW, Thomas B, McClelland IJ, Weerdmeester B, editors. Usability Evaluation in Industry: CRC Press; 1995.
28. Tullis T, Albert B. Chapter 11 - ten keys to success. In: Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics, 2nd edition: Morgan Kaufmann; 2013:279-288. [doi: 10.1016/B978-0-12-415781-1.00011-X]
29. Schrepp M. User Experience Questionnaire Handbook. User Experience Questionnaire. 2015. URL: https://www.ueq-online.org/Material/Handbook.pdf [accessed 2026-04-01]
30. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
31. Ottaviani AC, Monteiro DQ, Ferreira Campos CR, et al. iSupport-Brasil: preliminary results of the usability and acceptability assessment by caregivers of people who have dementia. Front Med (Lausanne) 2022;9:981748. [doi: 10.3389/fmed.2022.981748] [Medline: 36059826]
32. Stellefson M, Hanik B, Chaney B, Chaney D, Tennant B, Chavarria EA. eHealth literacy among college students: a systematic review with implications for eHealth education. J Med Internet Res 2011 Dec 1;13(4):e102. [doi: 10.2196/jmir.1703] [Medline: 22155629]
33. Hoong ALS, Thi LS, Lin MH. Affective technology acceptance model: extending technology acceptance model with positive and negative affect. In: Knowledge Management Strategies and Applications: IntechOpen Limited; 2017. [doi: 10.5772/intechopen.70351]
34. Parker M, Barlow S, Hoe J, Aitken L. Persistent barriers and facilitators to seeking help for a dementia diagnosis: a systematic review of 30 years of the perspectives of carers and people with dementia. Int Psychogeriatr 2020 May;32(5):611-634. [doi: 10.1017/S1041610219002229] [Medline: 32024558]
35. Molinari-Ulate M, Guirado-Sánchez Y, Platón L, van der Roest HG, Bahillo A, Franco-Martín M. Cultural adaptation of the iSupport online training and support programme for caregivers of people with dementia in Castilla y León, Spain. Dementia (London) 2023 Jul;22(5):1010-1026. [doi: 10.1177/14713012231165578] [Medline: 36942726]
36. Xiao LD, McKechnie S, Jeffers L, et al. Stakeholders’ perspectives on adapting the World Health Organization iSupport for Dementia in Australia. Dementia (London) 2021 Jul;20(5):1536-1552. [doi: 10.1177/1471301220954675] [Medline: 32864991]
37. Xiao LD, Ye M, Zhou Y, et al. Cultural adaptation of World Health Organization iSupport for Dementia program for Chinese-Australian caregivers. Dementia (London) 2022 Aug;21(6):2035-2052. [doi: 10.1177/14713012221110003] [Medline: 35724375]

XSL•FO
RenderX

38. Windle G, Flynn G, Hoare Z, et al. Evaluating the effects of the World Health Organization's online intervention "iSupport" to reduce depression and distress in dementia carers: a multi-centre six-month randomised controlled trial in the UK. Lancet Reg Health Eur 2025 Jan;48:101125. [doi: 10.1016/j.lanepe.2024.101125] [Medline: 40046086]

Abbreviations

CAB: Community Advisory Board
FC: formal caregiver
IC: informal caregiver
IT-eHEALS: Italian version of the eHealth Literacy Scale
REDCap: Research Electronic Data Capture
SUS: System Usability Scale
WHO: World Health Organization

Edited by A Stone; submitted 24.Jul.2025; peer-reviewed by B Davis, M Gagnon-Roy; accepted 28.Jan.2026; published 15.Apr.2026.

Please cite as:
Messina A, Amati R, Annoni AM, Albanese E, Fiordelli M
Usability of iSupport Swiss, a World Health Organization Digital Intervention for Caregivers of People With Dementia: Mixed Methods Study
J Med Internet Res 2026;28:e81247
URL: https://www.jmir.org/2026/1/e81247
doi:10.2196/81247

© Anna Messina, Rebecca Amati, Anna Maria Annoni, Emiliano Albanese, Maddalena Fiordelli. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effectiveness of mHealth Interventions for Improving eHealth Literacy Among Patients With Chronic Diseases: Meta-Analysis and Systematic Review

Xueqin Yang[1], MSN; Yingping Ma[1,2], MSN; Yueting Wang[1], MSN; Mengting Liu[1,3], MSN; Xingyu Liu[1], MSN; Xueping Jiao[1], MSN; Fanghong Yan[1], MD; Yuxia Ma[1], PhD; Lin Han[1,4], PhD; Yanan Zhang[1], PhD

[1]School of Nursing, Lanzhou University, No. 28, Yanxi Road, Chengguan District, Lanzhou, Gansu Province, China

[2]Department of Neurosurgery, The Second Affiliated Hospital of Lanzhou University, Lan Zhou, Gansu Province, Gansu, China

[3]Department of Neurosurgery, Gansu Provincial People's Hospital, Lan Zhou, Gansu, China

[4]Gansu Provincial People's Hospital, Lan Zhou, Gansu, China

Corresponding Author:
Yanan Zhang, PhD
School of Nursing, Lanzhou University, No. 28, Yanxi Road, Chengguan District, Lanzhou, Gansu Province, China

Abstract

Background: With the widespread use of the internet and mobile devices, eHealth literacy promotion is critical for medical equity. Mobile health (mHealth) serves as a pivotal tool for enhancing eHealth literacy by providing accessible, interactive platforms for health information engagement. However, the evidence regarding the effectiveness of mHealth interventions on eHealth literacy among patients with chronic diseases remains inconclusive.

Objective: This study aimed to evaluate the effectiveness of mHealth interventions on eHealth literacy among patients with chronic diseases based on randomized controlled trials (RCTs) and summarize supportive evidence from quasi-experimental and qualitative studies.

Methods: A comprehensive search strategy was developed, and 8 electronic databases were systematically searched for studies published up to February 12, 2026. Patients with chronic diseases were included based on predefined inclusion criteria. The Cochrane risk of bias 2 tool for RCTs and the ROBINS-I tool for quasi-experimental studies were used to assess the risk of bias. Given the anticipated substantial heterogeneity among the studies included, we used a random-effects model based on the Hartung-Knapp-Sidik-Jonkman method to pool effect sizes. A narrative and quantitative synthesis of the findings was provided where appropriate.

Results: A total of 15 studies were included in this review, including 6 RCTs, 5 quasi-experimental studies, and 4 qualitative studies, involving a total of 2884 patients with chronic diseases. Meta-analyses of RCTs suggested that mHealth interventions could improve eHealth literacy, with a pooled mean effect size of standardized mean difference (SMD)=1. 19 (95% CI 0.14-2.23; P=.03; I^2=97.75%; PI [prediction interval]=−2.68 to 5.05). Subgroup analyses by intervention targets showed that interventions on targets with specific disease produced larger mean effects (SMD=1.61; 95% CI 0.16-3.06; PI=−5.40 to 8.63), while interventions targeting the population with general chronic diseases produced smaller effects (SMD=0.36; 95% CI 0-0. 73; PI=−0. 21 to 0. 94). Analysis by intervention duration subgroup showed that the combined effect of studies with intervention duration <3 months was statistically significant (SMD=0.61; 95% CI 0.09-1.13; I^2=88.04%; PI=−5.72 to 6.95); while the combined effect of studies with intervention duration ≥3 months was not statistically significant. Taking into account bias and the risk of GRADE (Grading of Recommendations, Assessment, Development, and Evaluation), the certainty of RCT evidence was moderate, and the certainty of quasi-experimental evidence was low.

Conclusions: mHealth interventions could improve eHealth literacy among patients with chronic diseases on average. By using prediction intervals, this study reveals that the effectiveness of mHealth interventions is highly context-dependent and closely linked to implementation factors. Advancing beyond prior work, this study centers on eHealth literacy as a core outcome and integrates multiple types of evidence. Meanwhile, this finding emphasizes the need for evidence-based intervention programs and more rigorous implementation of intervention designs in future research.

Trial Registration: PROSPERO CRD 42024622807; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024622807

(J Med Internet Res 2026;28:e82004) doi:10.2196/82004

KEYWORDS

eHealth literacy; chronic diseases; mHealth intervention; meta-analysis; mobile health

XSL•FO
RenderX

Introduction

Chronic diseases pose a threat to population health across the life course, particularly affecting older adults, and have emerged as a major challenge in global public health. According to the World Health Organization, chronic diseases, including cardiovascular diseases, cancers, chronic respiratory diseases, and diabetes, lead to approximately 41 million deaths globally each year, accounting for 74% of all-cause deaths [1], with type 2 diabetes mellitus, cardiovascular diseases, cancers, and respiratory diseases accounting for 80% of premature deaths [2]. Moreover, factors such as population aging, lifestyle changes such as smoking, physical inactivity, and environmental pollution increase the incidence of patients with chronic diseases [3]. To address the rising prevalence of patients with chronic diseases, effective strategies for managing chronic diseases are urgently needed.

eHealth literacy is an extension of health literacy, underpinned by digital literacy as its technical foundation. It refers to an individual's comprehensive ability to use electronic resources to obtain, process, evaluate, and apply health information. eHealth literacy is defined as the ability of individuals to seek, find, understand, and appraise health information from electronic sources and apply the knowledge gained to address or solve a health problem [4], whose development relies upon the synergistic interplay of cognition, skills, motivation, and self-efficacy, rather than being a mere outcome variable. Studies have reported that enhanced eHealth literacy empowers patients to leverage online resources, telemedicine, and mobile health (mHealth) apps, thereby promoting treatment adherence and health outcomes [5]. Several studies showed eHealth literacy was insufficient among patients with chronic diseases. Zhenxiang et al [6] reported that only 5% of hospitalized patients with stroke in tertiary hospitals achieved adequate health literacy. Similarly, Sadeghi et al [7] reported that only 38.5% of patients with type 2 diabetes mellitus had adequate eHealth literacy, with 61.5% exhibiting low literacy. Moreover, studies reported that inadequate health literacy among patients with chronic diseases was associated with increased mortality and readmission rates [8,9], decreased quality of life [10], and increased disease-related health care costs [11]. Studies collectively underscore that inadequate health literacy impedes self-management and preventive care adherence in chronic disease populations [12,13]. Despite the potential of digital tools, many patients exhibit low eHealth literacy, limiting their capacity to effectively navigate and use these technologies. Consequently, improving eHealth literacy among patients with chronic diseases has become an urgent priority [14]. Improving eHealth literacy is key to achieving medical equity. It bridges the digital divide for vulnerable patients, ensures a fairer distribution of mHealth benefits, and reduces disparities in chronic disease outcomes.

mHealth is defined as "medical and public health practice supported by mobile devices, such as mobile phones, patient monitoring devices, personal digital assistants, and other wireless devices" [15]. Compared with traditional education, mHealth can offer multimodal learning materials (text, audio, video, and images), interactive functions (reminders, quizzes, and goal tracking), and repeated practice opportunities. The Lily model proposed by Norman posits that the abilities to obtain, communicate, and engage are core skills. And mHealth could promote eHealth literacy through convenient access to health care resources, facilitating doctor-patient communication and interaction, and enabling personalized feedback and monitoring [16,17]. The Integrative Model of eHealth Use theory posits that eHealth literacy serves as a crucial mediating variable for health outcomes, including behavioral and prognostic outcomes [18]. For instance, Qiu et al [19] research reported that mHealth apps could enhance the eHealth literacy and thereby improve the quality of life for older adult populations. Melholt et al [20] developed an interactive cardiac telerehabilitation tool within the dedicated online platform "Vital Heart," and found that rehabilitation interventions delivered via text, video, and images significantly enhance eHealth literacy among patients with chronic diseases. However, He et al [21] conducted a pilot study in Hong Kong where trained university students provided home-based, personalized mHealth apps for older adult participants and found that eHealth literacy was not significantly improved, suggesting that mHealth might limit the effectiveness in enhancing electronic health literacy. Given the great advantages of mHealth interventions in enhancing eHealth literacy, it is urgently needed to explore their effectiveness on eHealth literacy among patients with chronic diseases.

Hence, despite the potential of mHealth tools to improve eHealth literacy and health outcomes, few studies have examined the effectiveness of mHealth interventions on enhancing eHealth literacy among patients with chronic diseases. Although many researchers have explored strategies to improve eHealth literacy among patients with chronic diseases, existing studies remain limited by heterogeneous methodologies, inconsistent outcomes, and a lack of theoretical grounding [22,23]. Hence, there is an urgent need for rigorous, standardized approaches to evaluate intervention effectiveness, with randomized controlled trials (RCTs) providing the primary basis for causal inference, and to optimize eHealth literacy interventions for patients with chronic diseases. Therefore, we conducted this study to evaluate the effectiveness of mHealth interventions on eHealth literacy among patients with chronic diseases based on RCT evidence, while synthesizing quasi-experimental and qualitative studies as supportive and contextual evidence. Our findings will provide evidence for designing targeted, scalable mHealth strategies.

Methods

Data Sources and Search Strategy

We performed this systematic review and meta-analysis in accordance with the PICOS (population, intervention, comparator, outcome, and study design) framework and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search) guidelines [24] (Checklist 1). The study protocol was registered with PROSPERO (CRD42024622807). All databases were searched individually on their native platforms (no multidatabase platform searching was used). Two researchers (XY and YW) independently conducted a thorough search of all relevant studies published in the following databases: PubMed, Embase, Web of Science,

XSL•FO
RenderX

Cochrane Library, WanFang Database, China National Knowledge Infrastructure (CNKI), CQVIP, and Chinese Biomedical Literature (CBM) from inception to February 2026 to ensure no recent studies had been published since the initial search. No study registries (eg, ClinicalTrials.gov and World Health Organization International Clinical Trials Registry Platform) were searched. No additional online resources or print sources (eg, conference proceedings and websites) were purposefully searched or browsed. Medical subject headings terms and free words were used in combination to search the relevant studies. No published search filters were used in any of the database searches. The search strategies were developed de novo by the authors based on the PICOS framework and were not adapted from prior reviews. To maximize sensitivity for the population concept, the chronic disease component was searched using both generic terms (eg, chronic disease or illness, noncommunicable disease, and multimorbidity or comorbidity) and comprehensive controlled vocabulary or keywords covering major chronic conditions (cardiovascular diseases, cancers or neoplasms, chronic respiratory diseases, diabetes, hypertension, stroke or cerebrovascular disease, chronic kidney disease, and chronic liver disease). In addition, the language of the studies was set as English and Chinese. The references of pertinent previous systematic reviews were also screened as supplementary sources. We did not contact authors, experts, or manufacturers to seek additional studies or data. No additional search methods beyond those described were used. Following the completion of the initial draft, a supplementary literature search was conducted in February 2026 to identify any newly published studies (as noted in Deviations from the Registered Protocol). No automatic email alerts were used. The search strategies were developed by 2 reviewers (XY and YW) and reviewed by a third reviewer (XL) for completeness and accuracy; however, no formal peer review of the search (eg, using PRESS [Peer Review of Electronic Search Strategies]) was conducted. The full search strategies were shown in Tables S1 to S8 (Multimedia Appendix 1).

Eligibility Criteria

Eligibility criteria were set based on the PICOS framework:

1. Population: chronic disease has various definitions, which generally include the concepts that the condition is long-lasting, daily life is impacted, and ongoing management is required [25]. In our retrieval strategy, we used common chronic disease terminology, people with chronic illnesses (diabetes, high blood pressure, heart disease, lung disease, kidney disease, cancer, stroke, or any combination of these conditions that meet the diagnostic criteria for clinical diseases).
2. Interventions: mHealth-based interventions aimed at improving eHealth literacy in patients with chronic diseases. Eligible interventions included: mHealth apps (eg, disease management apps and medication reminders), SMS or text message–based health education, telehealth consultations via mobile devices, interactive mobile learning modules, and wearable device–integrated education programs. All interventions required primary delivery through mobile platforms (smartphones or tablets) with at least one eHealth literacy outcome measure.
3. Control: the study subjects in the control group received either no intervention at all or conventional interventions (eg, health education or general nursing practices).
4. Outcome: eHealth literacy score. The primary outcome measure of this study is the change in patients' eHealth literacy scores, assessed using validated scales such as eHealth Literacy Scale (eHEALS). This indicator was selected as the core basis for evaluating effectiveness (primarily within RCTs). In line with best practice, meta-analyses were conducted separately by study design (like-for-like pooling): RCTs were pooled only with RCTs, and quasi-experimental studies were pooled only with quasi-experimental studies; different designs were not combined within the same meta-analysis. The primary outcome was the eHealth literacy score. This outcome was selected because eHealth literacy represents both a key prerequisite capability and a direct outcome for patients to effectively use mHealth and engage in self-management. Given the anticipated clinical and methodological diversity, we used a random-effects model for all quantitative syntheses. Because eHealth literacy represents both a key prerequisite capability and a direct outcome for patients to effectively use mHealth and engage in self-management. It serves as a fundamental and widely recognized surrogate measure for determining whether interventions are effective.
5. Research design: RCTs and quasi-experimental studies were included in this systematic review and meta-analysis. In the former, the experimental group receives mHealth intervention, while the latter group either receives no intervention or traditional interventions (mostly written or spoken health education materials). Nonexperimental trials, review or discussion papers, non-Chinese and non-English papers, studies with no full text available, and studies with inadequate data were excluded.

Study Selection

A thorough screening procedure was conducted in compliance with PRISMA 2020 and 2 reviewers (XY and YW) screened the included studies independently. Among the initial 4057 records imported into EndNote X9 software, a total of 1142 duplicate publications were identified through a combination of EndNote software and manual verification. Subsequently, studies were excluded based on inclusion and exclusion criteria by reviewing titles and abstracts. Finally, full-text articles were reviewed for eligibility. Discussions or negotiations with a third reviewer (XL) were used to settle disagreements.

Data Extraction and Quality Assessment

XY and YM extracted data from each study separately. Study authors, publication date, country, population (population and sample size), study design (description of numerical interventions, characteristics of intervention groups, and duration), theoretical framework, intervention site outcome indicators (measurement tools), and primary research outcome were extracted. The author team engaged in discussions and reexamined the original study in cases where there were disagreements over data extraction until an agreement was reached. The quality of evidence was assessed using the GRADE (Grading of Recommendations, Assessment, Development, and

XSL•FO
RenderX

Evaluation) professional guidance development tool. Evidence quality was determined by evaluating five aspects: (1) risk of bias, (2) inconsistency, (3) indirectness, (4) imprecision, and (5) small-study effects [26].

Risk of Bias Assessment

This study used the risk-of-bias assessment tools recommended by the Cochrane Collaboration, evaluating studies according to their specific design types. For RCTs, the Cochrane Risk of Bias 2 (RoB 2) tool was used [27]. This tool evaluates six domains: random sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome data, selective reporting, and other bias. Each domain is rated as "Low risk of bias," "High risk of bias," or "Unclear risk of bias." Bias in nonrandomized studies (eg, quasi-experimental designs) was appraised using the ROBINS-I tool specifically designed for such intervention studies [28]. The 7 domains of bias addressed in the ROBINS-I assessment tool include bias due to confounding, bias arising from measurement of the exposure, bias in the selection of participants into the study, bias due to post-exposure interventions, bias due to missing data, bias in measurement of the outcome, and bias in selection of the reported result. Each domain is addressed using a series of signaling questions that aim to gather relevant information about the study and analysis being assessed. Most signaling questions offer five response options: "Yes," "Probably yes," "Probably no," "No," and "No information." Responses of "Yes" and "Probably yes" are treated equivalently in risk-of-bias judgments, as are "No" and "Probably no." The categories for risk of bias judgments are "No information," "Low risk," "Moderate risk," "Serious risk," and "Critical risk" of bias. All assessments were conducted independently by 2 reviewers (XY and YT). Disagreements were resolved through joint discussion or consultation with a third reviewer (XL).

Small-Study Effects

As the funnel plot requires at least 10 original studies, this study used regression analysis with Egger test to assess small-study effects. The results indicated that the intercept term β_1 was not statistically significant (β_1 =2.22, SE=1.83; *P*=.22), suggesting no significant small-study effects were detected in the current analysis, as shown in Table 1. For details on testing small-study effects in quasi-experimental studies, see Table S9 (Multimedia Appendix 1).

Table . Egger test for small-study effects in randomized controlled trial (n=6)[a].

Terms	Values
β_1	2.22
SE	1.827
z	1.21
P	.22

[a]The regression-based Egger test was performed under a random-effects model (Hartung–Knapp–Sidik–Jonkman estimator). H : β_1 =0 (no small-study effects). The non-significant *P* value suggests no evidence of small-study effects was detected in the present set of studies. However, it is important to note that the statistical power of the Egger test is limited when the number of included studies is small (k=6 in this analysis). Therefore, this result should be interpreted with caution, and the possibility of publication bias cannot be definitively ruled out.

Data Synthesis and Statistical Analysis

Although most included studies used Norman's eHealth Literacy Scale [29], some used other validated instruments to measure eHealth literacy. Given the variability in assessment tools, the standardized mean difference (SMD) and its 95% CI were calculated to enable pooling of continuous outcome data. In line with best practice, meta-analyses were conducted separately by study design (like-for-like pooling): RCTs were pooled only with RCTs, and quasi-experimental studies were pooled only with quasi-experimental studies; different designs were not combined within the same meta-analysis. Given anticipated clinical and methodological diversity, we used a random-effects model for all quantitative syntheses. The Hartung-Knapp-Sidik-Jonkman method was used to construct the random-effects model for effect estimation and pooling. This method is strongly recommended in meta-analyses as it enables more reliable uncertainty estimation while reducing the rate of false positives [30]. All analyses were conducted in Stata 17 using the meta suite with the Hartung–Knapp–Sidik–Jonkman method. To investigate potential sources of heterogeneity, we conducted pre-specified subgroup analyses based on target of intervention (specific chronic disease vs chronic disease in general) and intervention duration (≥3 mo vs <3 mo). Studies were sequentially excluded in sensitivity analyses. Egger tests were used to assess small-study effects for meta-analyses that involved more than 10 studies [31]. *P*<.05 was set as statistically significant.

Prediction Interval Calculation

To quantify the real-world impact of heterogeneity, we computed the 95% prediction interval (PI) for the pooled overall estimate. Due to the limited number of included studies (k=6) and substantial between-study heterogeneity, we applied the confidence distribution method proposed by Nagashima et al [32], which is recommended for small-sample meta-analyses. PIs were presented for meta-analyses with a sufficient number of studies (≥3). When the number of studies in a subgroup was small, PI estimates were interpreted cautiously.

Deviations From the Registered Protocol

During the implementation of this study, several adjustments were made to the protocol pre-registered on the PROSPERO platform (CRD42024622807). Following the completion of the initial draft, a supplementary literature search was conducted in February 2026, resulting in the inclusion of 15 studies [16,20,33-45] that met the eligibility criteria. Given the limited number of included studies, small-study effects were assessed

XSL•FO
RenderX

using the Egger test instead of the initially planned funnel plot analysis. For data synthesis, the Hartung-Knapp-Sidik-Jonkman random-effects model was used to pool effect sizes. The literature search was restricted to publications in Chinese and English. All adjustments were documented throughout the research process and were deemed unlikely to substantially affect the validity of the study conclusions.

Results

Selection of the Studies

The screening procedure for the subsequent subphases (title or abstract screening, full text assessment) was strictly followed by the PRISMA framework to ensure transparency and reproducibility. Initially, 4057 articles were searched, including PubMed (n=652), Embase (n=110 1), Web of Science (n=955), Cochrane Library (n=875), the Wan Fang Database (n=65), the CBM (n=144), the CQVIP (n=55), and the CNKI (n=210). Using EndNote software, 1142 duplicate references were excluded. The remaining 2915 papers were screened by title and abstract. A total of 2840 papers were excluded due to noncompliance with inclusion criteria or irrelevance to the research topic. After full-text review of the remaining 84 studies, 69 were excluded for reasons detailed in the supplementary file. (Multimedia Appendix 2). The final analysis incorporated 15 eligible studies [16,20,33-45], comprising 6 RCTs [33,38-40,44,45] for the primary quantitative synthesis, 5 quasi-experimental studies [34-37,43] for supplementary quantitative synthesis, and 4 qualitative studies [16,20,41,42] were included and synthesized narratively. Figure 1 shows the comprehensive flowchart of the literature search process.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the literature selection process for the systematic review on the effectiveness of mobile health interventions on eHealth literacy among patients with chronic diseases (studies published between 2016 and 2026).

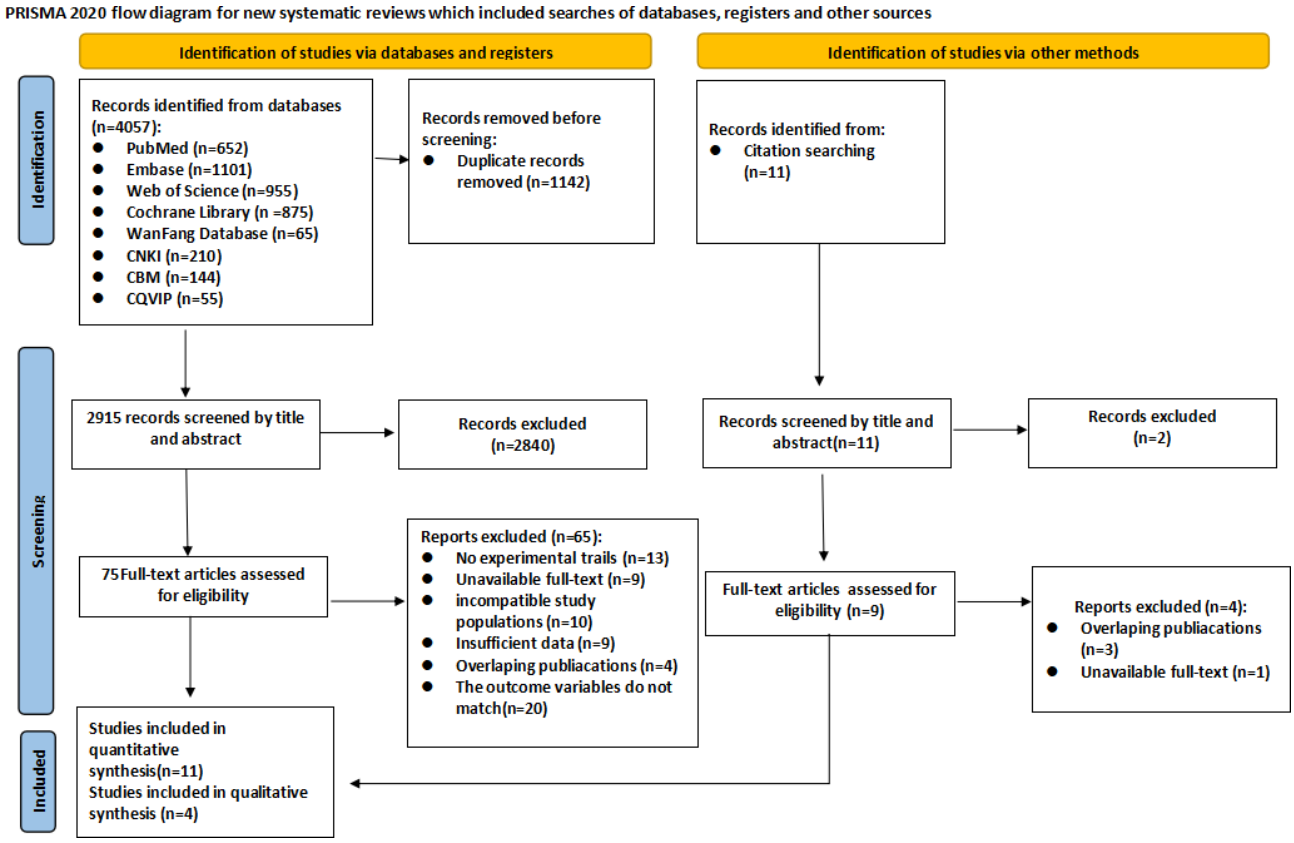

Study Characteristics

15 clinical studies [16,20,33-45] were reviewed and 2884 participants with various chronic conditions were included. Geographically, the studies were predominantly conducted in China (n=7) [33-39], Australia (n=2) [40,41], Canada (n=1) [42], Denmark (n=2) [16,20], South Korea (n=1) [43], and the United States [44,45]. All studies were published between 2016 and 2026. Two of the papers were published in Chinese, while thirteen were published in English. Regarding intervention formats, the 15 included studies predominantly used blended online-offline models (n=8) [33-36,38-40,45], followed by purely online interventions (n=4) [20,41,42,44], and face-to-face interventions (n=3) [16,37,43]. Regarding intervention content, the primary focus encompassed eHealth literacy skills training (n=6) [33,34,36-39], disease self-management and monitoring (n=5) [16,20,35,39,40], health knowledge dissemination (n=4) [16,20,35,41], and social interaction and support (n=3) [16,39,44]. Regarding intervention methodologies, approximately half of the studies were designed based on explicit theoretical frameworks (n=9) [16,33,35,37-39,42-44], with widespread adoption of personalized feedback (n=4) [16,35,40,43], interactive learning (n=5) [16,33,35,36,43], and multimedia instruction (n=5) [16,20,36,37,44], to enhance patient engagement and learning outcomes. The key features of the 15 studies are presented in Table 2. A summary of

XSL•FO
RenderX

intervention implementation is provided in Table S10 (Multimedia Appendix 1).

XSL•FO
RenderX

Table . Characteristics of included randomized controlled trials and quasi-experimental studies on mobile health (mHealth) interventions for improving eHealth literacy in patients with chronic diseases (n=15).

Study design	Author (year)	Country	Total (intervention/control)	Men/ women	Age (years), mean (SD)	Intervention type	Intervention content	Intervention population	Theory	Duration of intervention
RCT[a]	Yu [33] (2022)	China	58 (29/29)	19/39	51.03 (7.25)	Theory-based digital program (TMeHL)	Four-stage eHealth literacy training (positioning, communication, evaluation, and application).	Young or middle-aged stroke inpatients (18 - 59 years)	Transactional Model of eHealth Literacy	12 weeks
RCT	Nahm et al [44] (2019)	America	272 (138/134)	81/191	70 (8.5)	Online education program (T-PeP)	Structured learning modules + moderated discussions + virtual library.	Community-dwelling older adults (50+ years) with chronic conditions	Self-Efficacy Theory	3 weeks
QED[b]	Hu [34] (2024)	China	73 (38/35)	48/25	—	Blended online/offline training	Health information searching/evaluation/application practice.	Older patients with stroke (≥60 years) with internet experience	—	1 - 2 weeks
QED	Jiang [35] (2024)	China	60 (30/30)	31/29	—	WeChat-based mini-program	Health education + blood pressure monitoring + offline event coordination.	Hypertensive adults (18+ years) with smartphone proficiency	Health consciousness theory	3 months
QED	Guo et al [36] (2023)	China	132 （96/36）	93/39	41.22 (8.42)	Mobile eHealth education	Diabetes app navigation + reliable website identification.	Adults with type 2 diabetes (20 - 65 years)	—	3 months
RCT	Parker et al [40] (2022)	Australia	215 （120/95）	123/92	58.9 (8.8)	Nurse-led multifaceted intervention	Health checks + lifestyle app (mysnapp) + telephone coaching.	Adults with overweight or obesity (40 - 74 years) with low health literacy	—	6 months
QED	Chiu et al [37] (2016)	Taiwan, China	39 （39/20）	14/25	69.5 (6.9)	Touchscreen device training	Health/entertainment/social app tutorials.	Older adults (50+ years) with limited internet experience	Technology Acceptance Model	8 weeks
RCT	Redfern et al [41] (2020)	Australia	934 (486/448)	716/218	67.6 (8.1)	EHR[c]-integrated web application	CVD[d] risk assessment + goal-setting + social features.	Adults at high CVD risk (≥18 years)	—	Follow-up 12 months

XSL•FO
RenderX

Study design	Author (year)	Country	Total (intervention/control)	Men/women	Age (years), mean (SD)	Intervention type	Intervention content	Intervention population	Theory	Duration of intervention
RCT	Kastner et al [42] (2021)	Canada	440 （220/220)	—	—	Web-based self-management tool (Keep Well)	Personalized lifestyle advice + multicondition tracking.	Older adults (≥65 years) with multimorbidity	Knowledge-to-Action (KTA) model	6 months
QED	Melholt et al [20] (2017)	Denmark	109 (49/49)	40/9	60.64 (10.75)	Cardiac telerehabilitation portal	Multimedia disease management resources + self-monitoring.	Adult patients with cardiac issues (≥18 years) post-surgery or with heart failure	—	3 month
RCT	Cheng et al [38] (2024)	Taiwan, China	92 (46/46)	33/59	62.38 (12.9)	eHealth care experiential learning program	experiential learning program based on eHealth literacy framework (eHLF) and experiential learning theory (ELT), covering use of eHealth tools for diabetes self-management.	Patients with type 2 diabetes, aged ≥20 years, owning mobile phone or tablet with internet	eHLF and ELT	3 months
QED	Son et al [43] (2023)	South Korea	100 (50/50)	83/17	58.78 (8.83)	Interactive text message-based mHealth intervention (WithUs)	Weekly goal setting + interactive text messaging (chatbot) for heart failure self-care.	Patients with heart failure aged ≥40 years, owning Android smartphone	Information-Motivation-Behavioral skills model (IMB)	6 months
RCT	Spindler et al [16] (2022)	Denmark	137 (67/70)	105/32	61.5 (11.1)	Telerehabilitation program (HeartPortal)	Digital toolbox with educational materials, self-tracking (blood pressure, weight, and steps), communication with health professionals, goal setting, and patient-reported outcomes monitoring.	Patients with heart failure (NYHA I-IV), aged ≥18 years	Self-Determination Theory (SDT)	12 months

XSL•FO
RenderX

Study design	Author (year)	Country	Total (intervention/control)	Men/women	Age (years), mean (SD)	Intervention type	Intervention content	Intervention population	Theory	Duration of intervention
RCT	Gao et al [39] (2026)	China	130 (65/65)	108/22	73.92 (5.14)	Theory-guided eHealth literacy program (SRL-SEe PR)[e]	4-stage eHealth literacy course Self-learning via manuals/QR codes+ reflective journals+ remote WeChat support.	Older adults (≥65 years) with COPD[f]	Self-Efficacy theory	8 weeks
RCT	Lyles et al [45] (2019)	United States	93 (44/49)	45/48	54.3 (13)	Portal training program	In-person: research assistant–guided access to 11 video tutorials on patient portal use. Take-home: handout with link to self-access same 11 videos.	Vulnerable English-speaking adults (≥18 years) with chronic diseases and an email address	—	One-time session

[a]RCT: randomized controlled trial.

[b]QED: quasi-experiment design.

[c]EHR: electronic health record.

[d]CVD: cardiovascular disease.

[e]SRL-SEe PR: self-regulated learning model and self-efficacy theory.

[f]COPD: chronic obstructive pulmonary disease

Risk of Bias Assessment

A total of 15 studies [16,20,33-45] were included in this review. Of these, 11 studies contributed quantitative data to the meta-analyses (the primary meta-analysis pooled 6 RCTs [33,38-40,44,45], and the supplementary meta-analysis pooled 5 quasi-experimental studies [34-37,43], while the remaining 4 qualitative studies [16,20,41,42] were synthesized narratively. Importantly, risk-of-bias assessments were conducted for all included quantitative studies, regardless of whether they were ultimately pooled in the meta-analyses. Specifically, we used the Cochrane risk of bias assessment tool to evaluate the methodological quality of the 9 included RCTs [16,33,38-42,44,45,] (Figure 2A). Five studies [16,33,38,40,41] were judged to be at high risk of bias due to deviations from intended interventions or implementation issues. Four studies [16,33,38,44] had some concerns related to the randomization processor selection-bias–related information. Two studies [16,45] presented an unclear risk of reporting bias. Overall, the primary limitations of the RCT evidence centered on inadequate blinding and implementation, whereas other domains were relatively consistent, leading to an acceptable overall risk-of-bias profile. For the 6 quasi-experimental studies [20,34-37,43] included, we used the ROBINS-I tool for bias risk assessment (Figure 2B). The assessment revealed that 5 studies [20,34,35,37,43] exhibited a core issue of serious risk of confounding bias, and 2 studies [20,37] presented a high risk of bias due to missing data.

XSL•FO
RenderX

Figure 2. Summary of risk of bias assessment for included studies. (A) Risk of bias in randomized controlled trials (assessed using the Cochrane RoB 2 tool). (B) Risk of bias in non-randomized controlled trials (assessed using the ROBINS-I tool) [16,20,33-45].

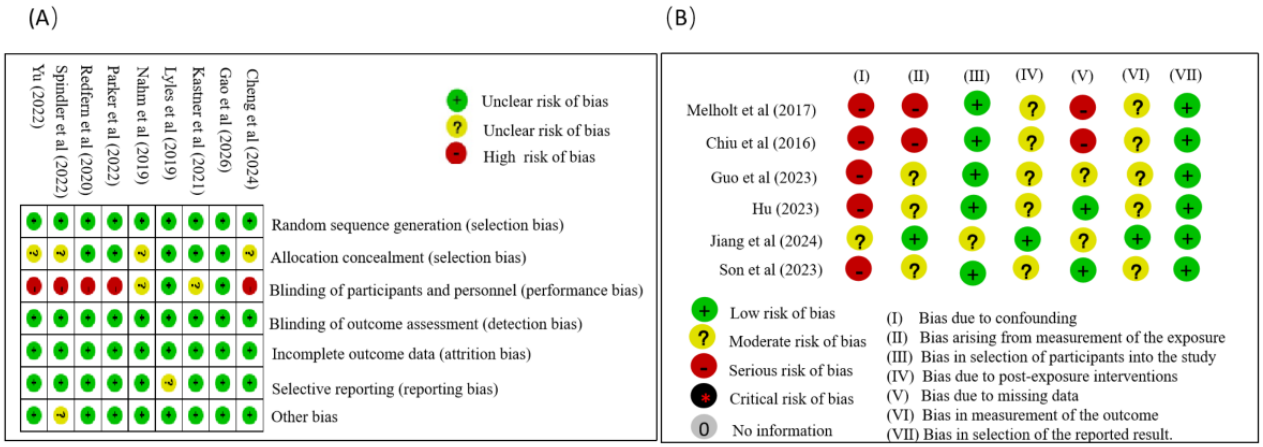

Certainty of Evidence

The initial quality of evidence in this study was rated as high due to the inclusion of RCTs. However, the certainty of evidence was downgraded by one level owing to substantial heterogeneity between studies that could not be fully explained. Consequently, the overall quality of evidence supporting the conclusion that mHealth interventions can improve eHealth literacy among patients with chronic diseases is moderate**.** The GRADE evidence level for the pooled estimate from quasi-experimental studies was, as expected, rated as low due to the very serious risk of confounding bias inherent in their design. (Multimedia Appendix 3.)

Overall Effectiveness of RCTs mHealth Interventions on eHealth Literacy for Patients With Chronic Diseases

The effectiveness of eHealth literacy interventions in patients with chronic diseases was systematically assessed in this review, with effectiveness primarily assessed using RCT evidence. As explained below, specific data could be found in Table S11 (Multimedia Appendix 1). Quantitative syntheses followed a like-for-like approach by design: the primary meta-analysis pooled RCTs only (Figure 3 [33,38-40,44,45]), and a separate supplementary meta-analysis pooled quasi-experimental studies only (Figure 4 [34-37,43]).

Figure 3. Forest plot of the meta-analysis of randomized controlled trials: effect of mHealth interventions on eHealth literacy in patients with chronic diseases. mHealth: mobile health; SMD: standardized mean difference [33,38-40,44,45].

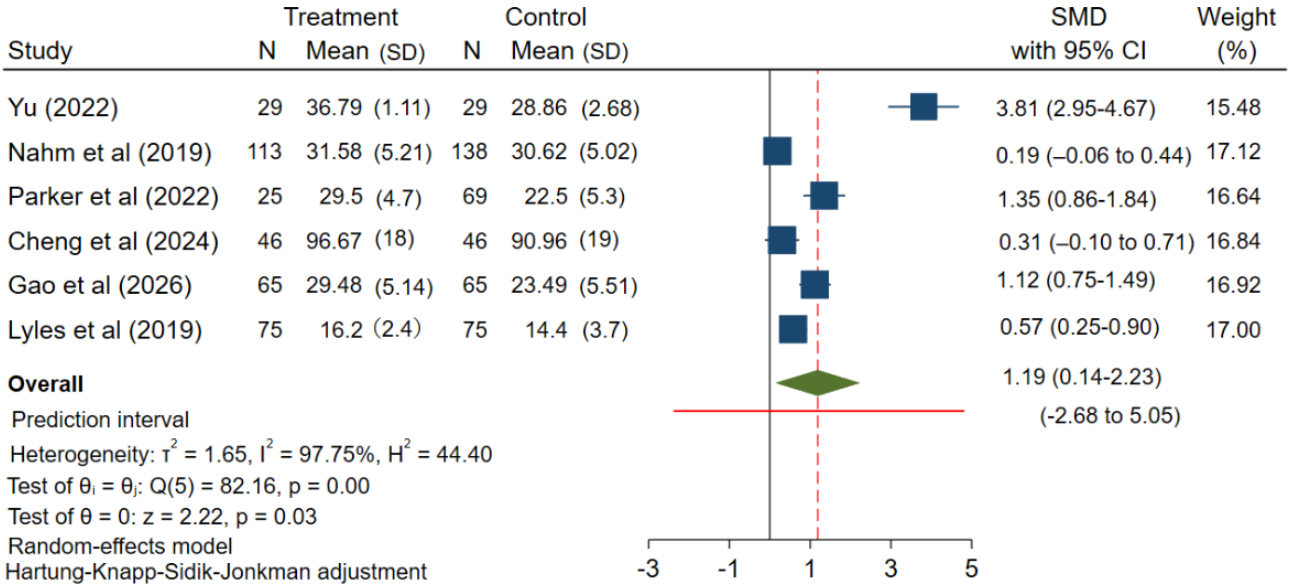

XSL•FO
RenderX

Figure 4. Forest plot of the meta-analysis of quasi-experimental studies: effect of mHealth interventions on eHealth literacy in patients with chronic diseases. mHealth: mobile health; SMD: standardized mean difference [34-37,43].

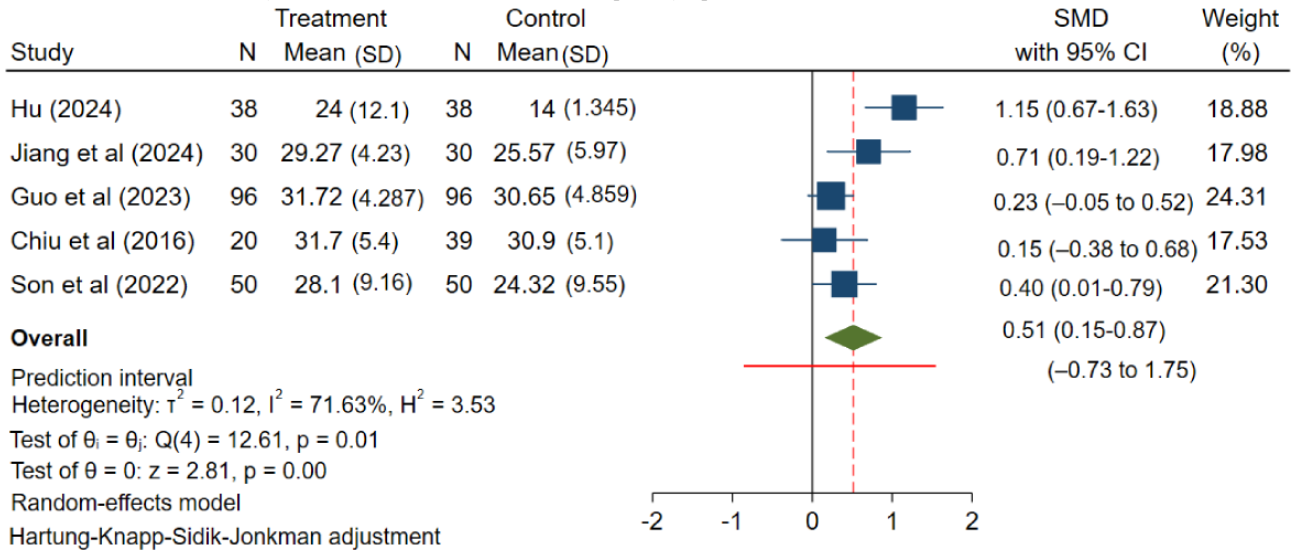

Primary Analysis: RCTs

The meta-analysis of 6 RCTs [33,38-40,44,45] showed that mHealth interventions can enhance eHealth literacy among patients with chronic diseases. The pooled average effect was positive (SMD=1.19; 95% CI 0.14-2.23; *P*=.03; Figure 3). The 95% CI quantifies uncertainty around this pooled average effect. Between-study heterogeneity was substantial; while I^2 is commonly reported (I^2=97.75%, *P*<.001), it has limited pragmatic utility for understanding how much true effects vary across settings. Therefore, to describe the expected distribution of true effects across different populations and settings, we calculated the 95% PI, which was extremely wide (−2.68 to 5.05; Figure 3). Thus, although the average effect is beneficial, a future implementation could plausibly observe little or no benefit or even harm, or conversely, a very large benefit, depending on context and implementation. This interpretation is consistent with the risk-of-bias assessment (several trials had limitations related to blinding and implementation) and with the GRADE rating of moderate certainty for the RCT evidence, downgraded for unexplained inconsistency.

Supplementary Analysis: Quasi-Experimental Studies

Overview

A separate meta-analysis of the 5 quasi-experimental studies [34-37,43] also indicated a positive pooled average effect (SMD=0.51; 95% CI 0.15 - 0.87; *P*<.001; Figure 4). Between-study heterogeneity was notable; although I^2 was moderate to high (I^2=71.63%), the 95% PI was also broad (−0.73 to 1.75), indicating that true effects may range from little or no benefit to meaningful benefit across settings. Given the serious risk of bias (primarily confounding) and low-grade certainty associated with these studies, these results should be considered supportive but less robust than the primary RCT findings.

Analysis of Subgroups

We conducted subgroup analyses based on prespecified subgroups (including intervention targets and intervention duration), as detailed below, to more specifically explore potential effect modifiers in the RCT-based estimates of intervention effectiveness on eHealth literacy.

The Effect of mHealth Intervention on eHealth Literacy Varied by Intervention Targets

After grouping by intervention target, specific disease interventions yielded a positive and substantial pooled average effect (SMD=1.61, 95% CI 0.16 - 3.06; Figure 5). However, the 95% PI was very wide (PI −5.40 to 8.63), indicating that true effects may vary considerably across implementation contexts, ranging from apparent ineffectiveness or even negative outcomes to highly significant benefits. In contrast, interventions targeting the general chronic disease population yielded smaller average effects (SMD=0.36, 95% CI 0.00 - 0.73; Figure 5 [33,38-40,44,45]), with a PI of −0.21 to 0.94. This implies that in certain contexts, the intervention may approach ineffectiveness, while in others, there remains a modest possibility of achieving some degree of positive effect. For completeness, conventional I^2 statistics are reported in Figure 5 (specific disease: I^2=97.25%; general chronic disease population: I^2=68.46%), but the PI provides the more clinically interpretable description of between-setting variability. The test for between-group differences failed to reach statistical significance (*P*=.10). Consequently, this grouping outcome should be regarded more as a suggestive trend than as a definitive effect modifier.

XSL•FO
RenderX

Figure 5. Subgroup analysis by intervention target: effect of mHealth interventions on eHealth literacy in patients with chronic diseases; SMD: standardized mean difference (Group 1: interventions targeting a specific disease [33,38-40]; Group 2: interventions targeting general chronic diseases population [44,45]).

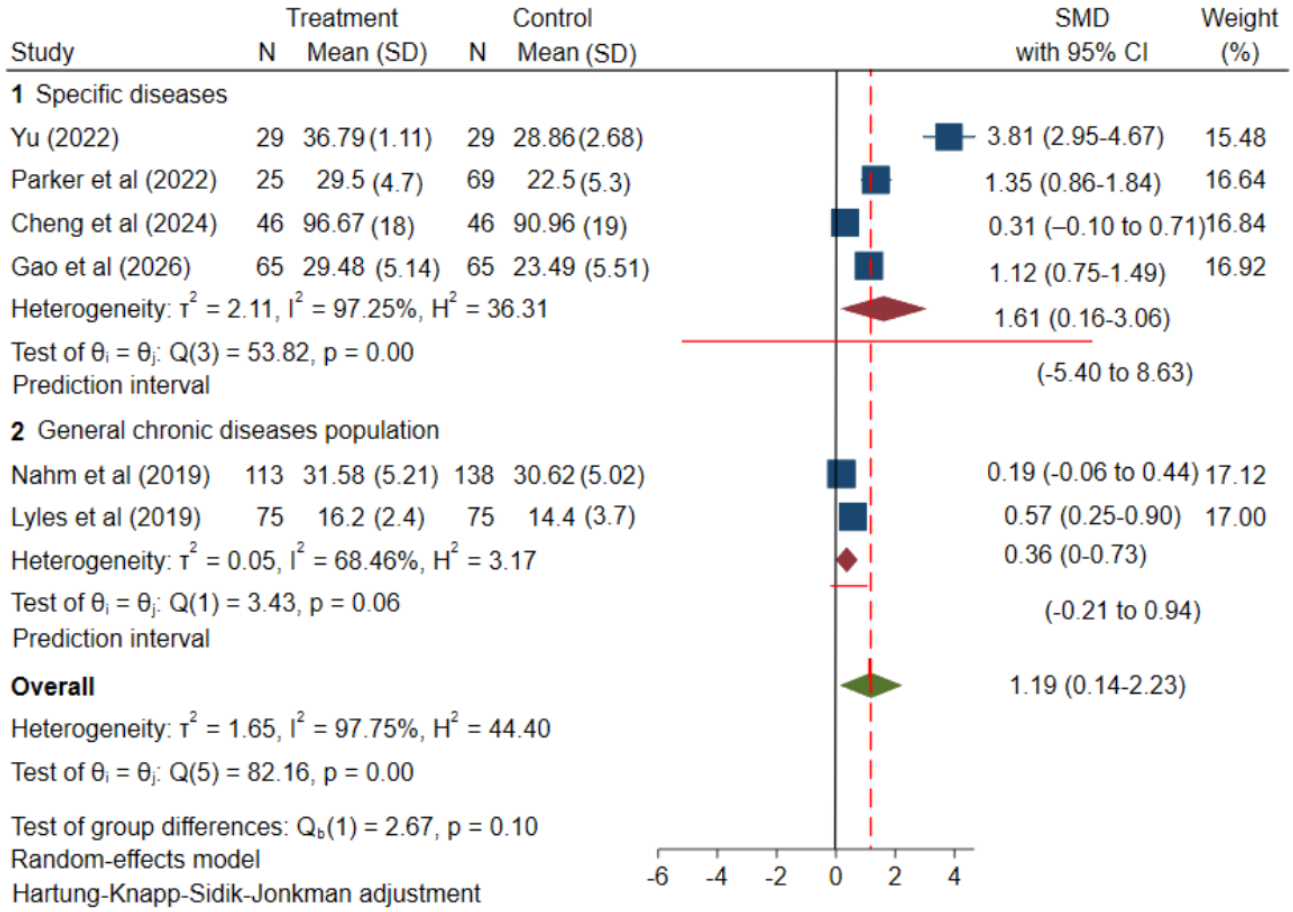

The Effect of mHealth Intervention on eHealth Literacy Varied by Duration of Intervention

After grouping by intervention duration, results indicated that the mean effect in the ≥3-month group was positive, yet its 95% CI encompassed zero (SMD=1.79, 95% CI −0.22 to 3.80; Figure 6 [33,38-40,44,45]). The corresponding PI was extremely wide (95% PI −23.98 to 27.56), suggesting that true effects may vary dramatically across studies and implementation contexts. In contrast, the group with an intervention duration of <3 months also exhibited a positive mean effect (SMD=0.61, 95% CI 0.09 - 1.13; Figure 6) with a broad PI (PI −5.72 to 6.95). Although conventional I^2 values were reported in Figure 6 (≥3 mo: I^2=97.52%;<3 mo: I^2=88.04%), the PIs better quantify the likely range of effects in new settings. This suggests that short-term interventions may yield benefits in some contexts, yet the possibility of near-ineffectiveness or even negative outcomes persists.

XSL•FO
RenderX

Figure 6. Subgroup analysis by intervention duration: effect of mHealth interventions on eHealth literacy in patients with chronic diseases; SMD: standardized mean difference (Group 1: intervention duration ≥3 months [33,38,40]; Group 2: intervention duration <3 months [39,44,45]).

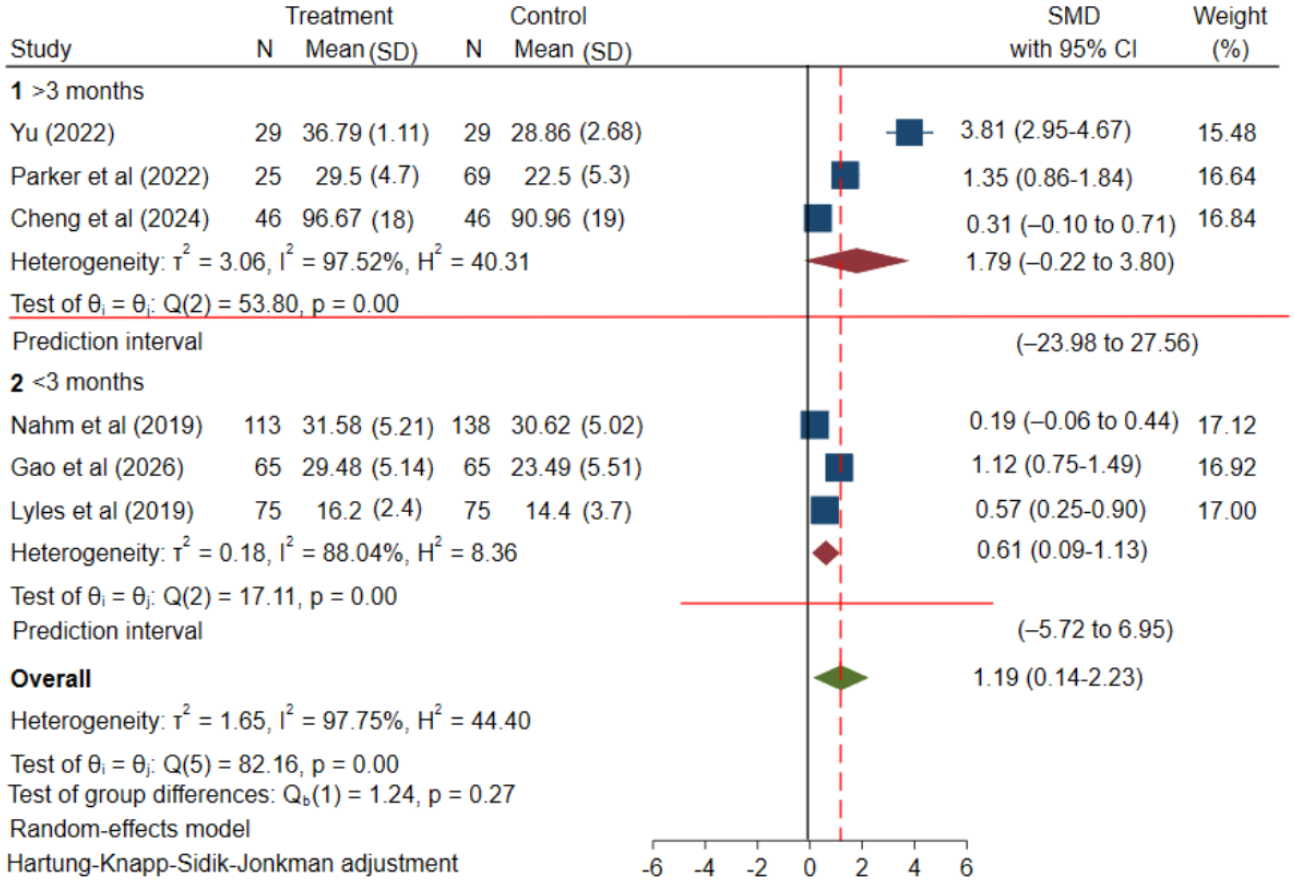

Sensitivity Analyses

The leave-one-out sensitivity analysis (Figure 7) demonstrated that no single study was solely responsible for the overall positive pooled effect. The point estimate remained positive and all CIs overlapped after the sequential exclusion of each study, indicating that the overall conclusion of an average positive effect is relatively robust to the influence of any one study.

XSL•FO
RenderX

Figure 7. Leave-one-out sensitivity analysis of the pooled effect size for the effect of mHealth interventions on eHealth literacy in patients with chronic diseases [33,38-40,44,45].

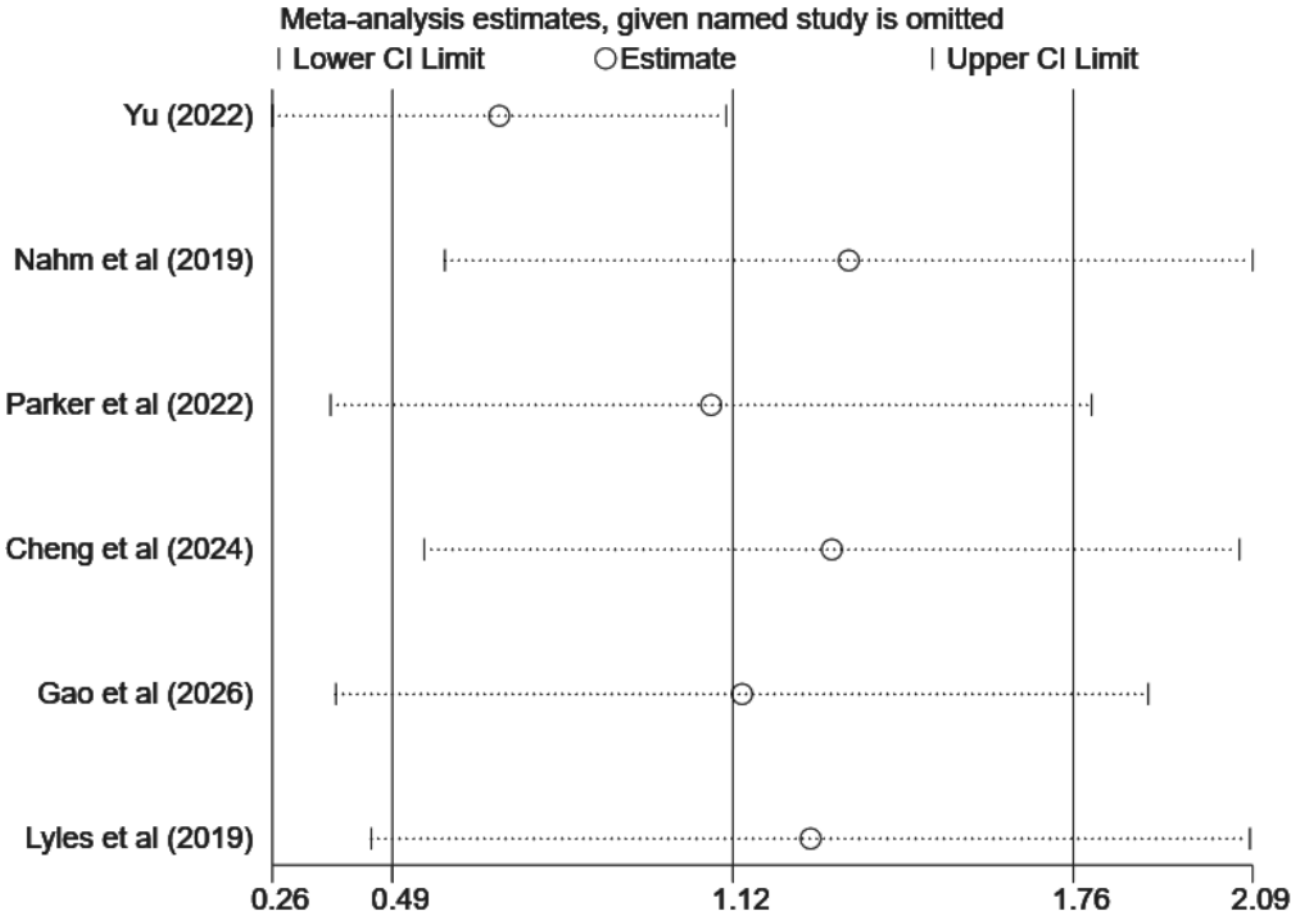

Discussion

Principal Findings

This review aimed to understand whether mHealth tools could improve eHealth literacy among patients with chronic diseases. We used the results of RCTs as primary evidence. A combined analysis of 6 RCTs showed that the intervention had a positive effect on average. This study is consistent with previous research findings [46,47]. However, there are large differences between the studies. The 95% PI was wide. The wide pPI suggests that the effectiveness of the same intervention may differ substantially across implementation contexts, with true effects potentially ranging from negligible to clinically meaningful. Due to this large uncertainty and some problems with the study (such as the lack of blinding), we rated the overall quality of evidence from RCTs as moderate using the GRADE system [48,49]. We also synthesized 5 quasi-experimental studies, and the results of the quasi-experiment supported the results of RCT with large PI.

Effectiveness of mHealth Interventions

The findings indicated that mHealth interventions are effective in enhancing eHealth literacy among patients with chronic diseases. The effect was more robust in RCTs, while quasi-experimental studies showed more modest effects accompanied by higher uncertainty. This result is generally consistent with previous reviews, which have shown that mHealth interventions are generally beneficial for health literacy-related outcomes, but the effects vary widely [50]. The observed differences may stem from heterogeneity in patient characteristics, intervention design, and outcome measures [51]. Lyles et al [45], in a study of underserved populations, found that portal use among individuals with limited health literacy was substantially lower than among those with adequate health literacy. Nahm et al [44] reported more favorable outcomes in older adults with higher educational attainment, suggesting that socioeconomic status and access to resources are critical determinants of intervention effectiveness. Intervention intensity and theoretical grounding also appear to be pivotal [52,53]. Gao et al [39] demonstrated that an intervention based on the self-regulated learning model was more effective than one relying on a single theoretical framework. Cheng et al [38] showed that experiential learning significantly improved eHealth literacy. By contrast, the findings of Parker et al [40] did not achieve meaningful improvements in clinical endpoints, likely owing to insufficient intervention intensity and limited tailoring. Future research should prioritize the development of stratified, adaptive, and theory-informed multimodal interventions [54]. In particular, greater attention should be directed toward designing and validating more personalized and supportive strategies for vulnerable populations with limited health literacy and inadequate social resources [55,56].

Intervention Targets

This study conducted subgroup analyses by intervention target. The results showed that interventions designed for one specific

XSL•FO
RenderX

disease (eg, diabetes alone or heart failure alone) appeared to work better than interventions designed for patients with all chronic conditions. This finding is consistent with previous research [57]. Interventions on patients with specific diseases showed greater effects on average. mHealth intervention programs targeting the general chronic disease population have smaller average effects, and their PIs suggest that they may not always be beneficial. When an intervention focuses on a specific disease, its content, tasks, and feedback closely match what patients need to do every day [58]. This may make the tool more useful to patients and keep them engaged [59]. In the field of mHealth, higher engagement generally leads to better outcomes [60,61]. For example, apps that help manage diabetes or blood pressure often include self-management tasks and provide direct feedback, which may explain why they work so well [62]. In contrast, trying to cover multiple conditions can make it difficult to provide in-depth, personalized support for each condition [51]. Recent research on tools for patients with multiple conditions showed that combining several disease-specific tools could feel burdensome and did not always provide useful, actionable recommendations [63,64]. Therefore, it may be wiser to start with small, sophisticated tools for a single disease and consider expanding to more diseases later.

Duration of Intervention

We further compared studies with an intervention duration of less than 3 months versus an intervention duration of more than 3 months. Subgroup analysis showed a statistically significant combined effect of the less than 3-month intervention, which is generally consistent with the findings of Sevda and Mechelle [65,66]. However, the PI remained notably wide, suggesting considerable uncertainty in the effectiveness of the intervention when applied across diverse implementation contexts. In contrast, the subgroup with intervention for more than 3 months had larger point estimates, but its combined effect did not reach statistical significance, and the PI was extremely wide. Given that the test for subgroup differences did not reach statistical significance, it is unwarranted to conclude that long-term interventions are inherently less effective than short-term ones. A more plausible interpretation is that the effectiveness of longer-duration mHealth interventions may be more susceptible to contextual factors and the fidelity of implementation processes [67]. The possible reason is that the benefits of mHealth interventions depend not only on the duration of the intervention itself but also on whether patient engagement can be sustained over time [68]. Existing evidence suggests that user engagement in mHealth interventions often declines as time progresses, while higher levels of engagement are generally associated with better adherence and improved self-management outcomes [52]. For longer-duration interventions, factors such as treatment burden, reduced interaction, or implementation barriers may weaken sustained patient participation [69]. Therefore, future research should not only focus on extending the duration of interventions but also prioritize optimizing intervention content, feedback frequency, personalized support, and engagement strategies to enhance long-term adherence and participation, thereby maximizing the long-term effectiveness of mHealth interventions.

Limitations

Four major limitations of this study should be mentioned. First, the restriction to English and Chinese language publications may introduce selection bias and limit the cross-cultural applicability of our findings. Second, there was considerable methodological heterogeneity among the studies included, particularly in terms of intervention protocols and outcome measures. Third, this study acknowledges certain limitations in the current analysis, primarily reflected in the relatively small sample size. Fourth, this study used eHealth literacy scores as the primary indicator of intervention effectiveness. While this metric directly reflects improvements in digital health capabilities, it struggles to comprehensively capture the intervention's overall benefits regarding clinical symptoms, quality of life, and health care costs. Future research should adopt multidimensional indicators to conduct more comprehensive assessments.

Conclusions

Overall, current evidence suggests that mHealth interventions can improve eHealth literacy among patients with chronic diseases on average. However, the wide PIs indicate substantial between-study variability, suggesting that intervention effects are highly context-dependent and closely related to implementation factors. By focusing on eHealth literacy as a core capability outcome and integrating evidence from RCTs, quasi-experimental studies, and qualitative research, this review extends previous work in the field. These findings highlight the need for more rigorously designed and better-implemented mHealth interventions in future research.

Acknowledgments

The authors sincerely thanked all participants and investigators involved in the primary studies included in this systematic review. We sincerely thanked Gansu Provincial People's Hospital for support of this study.

Funding

This work was supported by the Gansu Provincial University Faculty Innovation Fund [No. 2025B-016].

Disclaimer

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision:

XSL•FO
RenderX

- Idea generation

Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes. During the preparation of this manuscript, the authors used AI tools (ChatGPT-5.4 and DeepSeek) solely for language polishing and grammar checking. The AI tools were not used for manuscript writing, data analysis, figure preparation, or drawing scientific conclusions. The authors take full responsibility for the content and originality of this article.

Data Availability

All data used in this systematic review and meta-analysis were derived from publicly available research literature. Data extracted for this study may be accessed in the supplementary information files.

Authors' Contributions

YZ and XY conceptualized the study. YW and XL developed the methodology. XY and YW performed formal analysis. XY, YW, ML, and FY conducted the investigation. XY and YW managed data curation. XY wrote the original draft. YZ, Yingping M, Yuxia M, and FY contributed to writing review and editing. YZ provided supervision and acquired funding. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary data.
[DOCX File, 1718 KB - jmir_v28i1e82004_app1.docx]

Multimedia Appendix 2
Studies excluded during full-text screening and reasons.
[XLSX File, 159 KB - jmir_v28i1e82004_app2.xlsx]

Multimedia Appendix 3
GRADE (Grading of Recommendations, Assessment, Development, and Evaluations) for evidence.
[PDF File, 96 KB - jmir_v28i1e82004_app3.pdf]

Checklist 1
PRISMA-S checklist.
[PDF File, 87 KB - jmir_v28i1e82004_app4.pdf]

Checklist 2
PRISMA 2020 checklist.
[PDF File, 103 KB - jmir_v28i1e82004_app5.pdf]

References

1. Wang M, Ma H, Qin C, et al. Burden of diet-related chronic diseases in Chinese and Japanese adults attributable to dietary risk factors from 1990 to 2021: a systematic analysis of the global burden of disease study 2021. Front Nutr 2024;11:1472451. [doi: 10.3389/fnut.2024.1472451] [Medline: 39925968]
2. Noncommunicable diseases. World Health Organization. 2025. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases [accessed 2026-03-30]
3. Devaux M, Lerouge A, Giuffre G, et al. How will the main risk factors contribute to the burden of non-communicable diseases under different scenarios by 2050? A modelling study. PLoS ONE 2020;15(4):e0231725. [doi: 10.1371/journal.pone.0231725] [Medline: 32348352]
4. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9. [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
5. Lu X, Zhang R. Association between eHealth literacy in online health communities and patient adherence: cross-sectional questionnaire study. J Med Internet Res 2021 Sep 13;23(9):e14908. [doi: 10.2196/14908] [Medline: 34515638]
6. Zhenxiang Z, Hui R, Zhiguang P, Yunfei G. Status and influencing factors of eHealth literacy in stroke patients [Article in Chinese]. Chin Gen Pract 2021;24(22):2850-2854. [doi: 10.12114/j.issn.1007-9572.2021.00.209]

XSL•FO
RenderX

7. Sadeghi S, Borzu ZA, Shirzadi S, Jahanfar S, Abadi YG, Taziki M. Investigating e-health literacy and its relationship with self-care behaviors among Iranian middle-aged type 2 diabetes: a cross-sectional study. BMC Public Health 2025 Apr 12;25(1):1376. [doi: 10.1186/s12889-025-22648-w] [Medline: 40221762]
8. Estrella ML, Allen-Meares P, Ricardo AC, et al. Prospective associations of health literacy with clinical outcomes in adults with CKD: findings from the CRIC study. Nephrol Dial Transplant 2023 Mar 31;38(4):904-912. [doi: 10.1093/ndt/gfac201] [Medline: 35746879]
9. Kanejima Y, Shimogai T, Kitamura M, Ishihara K, Izawa KP. Impact of health literacy in patients with cardiovascular diseases: a systematic review and meta-analysis. Patient Educ Couns 2022 Jul;105(7):1793-1800. [doi: 10.1016/j.pec.2021.11.021] [Medline: 34862114]
10. Elisabeth Stømer U, Klopstad Wahl A, Gunnar Gøransson L, Hjorthaug Urstad K. Health literacy in kidney disease: associations with quality of life and adherence. J Ren Care 2020 Jun;46(2):85-94. [doi: 10.1111/jorc.12314] [Medline: 31950601]
11. Palumbo R. Examining the impacts of health literacy on healthcare costs. An evidence synthesis. Health Serv Manage Res 2017 Nov;30(4):197-212. [doi: 10.1177/0951484817733366] [Medline: 29034727]
12. Shiferaw KB, Tilahun BC, Endehabtu BF, Gullslett MK, Mengiste SA. E-health literacy and associated factors among chronic patients in a low-income country: a cross-sectional survey. BMC Med Inform Decis Mak 2020 Aug 6;20(1):181. [doi: 10.1186/s12911-020-01202-1] [Medline: 32762745]
13. Jiang X, Wang L, Leng Y, et al. The level of electronic health literacy among older adults: a systematic review and meta-analysis. Arch Public Health 2024 Nov 7;82(1):204. [doi: 10.1186/s13690-024-01428-9] [Medline: 39511667]
14. van der Gaag M, Heijmans M, Spoiala C, Rademakers J. The importance of health literacy for self-management: a scoping review of reviews. Chronic Illn 2022 Jun;18(2):234-254. [doi: 10.1177/17423953211035472] [Medline: 34402309]
15. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009 Jul 21;339(jul21 1):b2535. [doi: 10.1136/bmj.b2535] [Medline: 19622551]
16. Spindler H, Dyrvig AK, Schacksen CS, et al. Increased motivation for and use of digital services in heart failure patients participating in a telerehabilitation program: a randomized controlled trial. Mhealth 2022;8:25. [doi: 10.21037/mhealth-21-56] [Medline: 35928510]
17. Choi S. Comparison of self-tracking health practices, eHealth literacy, and subjective well-being between college students with and without disabilities: cross-sectional survey. JMIR Form Res 2024 Apr 10;8:e48783. [doi: 10.2196/48783] [Medline: 38598285]
18. Hsu W, Chiang C, Yang S. The effect of individual factors on health behaviors among college students: the mediating effects of eHealth literacy. J Med Internet Res 2014 Dec 12;16(12):e287. [doi: 10.2196/jmir.3542] [Medline: 25499086]
19. Qiu SS, Ye JF, You F, Liu M, Zhao X. How does mHealth benefit older Chinese adults' quality of life? Examining the roles of eHealth literacy, health motivation, and patient activation. Digit Health 2025;11:20552076241313160. [doi: 10.1177/20552076241313160] [Medline: 39807427]
20. Melholt C, Joensson K, Spindler H, et al. Cardiac patients' experiences with a telerehabilitation web portal: implications for eHealth literacy. Patient Educ Couns 2018 May;101(5):854-861. [doi: 10.1016/j.pec.2017.12.017] [Medline: 29305064]
21. He AWJ, Yuan R, Luk TT, Wang KMP, Chan SSC. Boosting digital health engagement among older adults in Hong Kong: pilot pre-post study of the generations connect project. JMIR Form Res 2025 May 8;9:e69611. [doi: 10.2196/69611] [Medline: 40340793]
22. Yun S, Enjuanes C, Calero-Molina E, et al. Usefulness of telemedicine-based heart failure monitoring according to "eHealth literacy" domains: insights from the iCOR randomized controlled trial. Eur J Intern Med 2022 Jul;101:56-67. [doi: 10.1016/j.ejim.2022.04.008] [Medline: 35483994]
23. Shi YX, Ma DH, Zhang J, Chen BW. In the digital age: a systematic literature review of the e-health literacy and influencing factors among Chinese older adults. Z Gesundh Wiss 2023;31(5):679-687. [doi: 10.1007/s10389-021-01604-z] [Medline: 34104627]
24. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
25. Dale MT, Elkins MR. Chronic disease. J Physiother 2021 Apr;67(2):84-86. [doi: 10.1016/j.jphys.2021.02.001] [Medline: 33622651]
26. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-926. [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
27. Higgins JPT, Altman DG, Gøtzsche PC, et al. The Cochrane collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011 Oct 18;343(oct18 2):d5928. [doi: 10.1136/bmj.d5928] [Medline: 22008217]
28. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016 Oct 12;355:i4919. [doi: 10.1136/bmj.i4919] [Medline: 27733354]
29. Norman CD, Skinner HA. eHEALS: the eHealth Literacy Scale. J Med Internet Res 2006 Nov 14;8(4):e27. [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]

XSL•FO
RenderX

30. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
31. Dalton JE, Bolen SD, Mascha EJ. Publication bias: the elephant in the review. Anesth Analg 2016 Oct;123(4):812-813. [doi: 10.1213/ANE.0000000000001596] [Medline: 27636569]
32. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019 Jun;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]
33. Yu J. Construction of electronic health literacy intervention program for young and middle-aged stroke patients and empirical research. J Nurs Sci 2023;38(24). [doi: 10.27115/d.cnki.glygc.2023.001083]
34. Hu Y. Construction and Empirical Study of Electronic Health Literacy Intervention Program for Elderly Patients with Chronic Diseases-A Case Study of Stroke Patients: Nanjing University of Chinese Medicine; 2021. [doi: 10.27253/d.cnki.gnjzu.2024.000231]
35. Jiang Y. Impact of Community Proactive Health Management Application on Electronic Health Literacy and Self-Management of Hypertensive Patients. Public Health Nurs 2024;41(6):1436-1445. [doi: 10.1111/phn.13424] [Medline: 39297673]
36. Guo SHM, Lin JL, Hsing HC, Lee CC, Chuang SM. The Effect of mobile eHealth education to improve knowledge, skills, self-care, and mobile eHealth literacies among patients with diabetes: development and evaluation study. J Med Internet Res 2023 Dec 6;25:e42497. [doi: 10.2196/42497] [Medline: 38055321]
37. Chiu CJ, Hu YH, Lin DC, Chang FY, Chang CS, Lai CF. The attitudes, impact, and learning needs of older adults using apps on touchscreen mobile devices: results from a pilot study. Comput Human Behav 2016 Oct;63:189-197. [doi: 10.1016/j.chb.2016.05.020]
38. Cheng YS, Lin CP, Chen LYA, Hwang WR, Lin YC, Chen YC. Short-term effects of an eHealth care experiential learning program among patients with type 2 diabetes: randomized controlled trial. J Med Internet Res 2024 Aug 16;26:e53509. [doi: 10.2196/53509] [Medline: 39150761]
39. Gao J, Cui S, Jiang Y, et al. Theory-guided eHealth literacy support for tele-pulmonary rehabilitation in older adults with chronic obstructive pulmonary disease: a randomized controlled trial. Digit Health 2026;12:20552076261425499. [doi: 10.1177/20552076261425499] [Medline: 41710759]
40. Parker SM, Barr M, Stocks N, et al. Preventing chronic disease in overweight and obese patients with low health literacy using eHealth and teamwork in primary healthcare (HeLP-GP): a cluster randomised controlled trial. BMJ Open 2022 Nov 30;12(11):e060393. [doi: 10.1136/bmjopen-2021-060393] [Medline: 36450426]
41. Redfern J, Coorey G, Mulley J, et al. A digital health intervention for cardiovascular disease management in primary care (CONNECT) randomized controlled trial. NPJ Digit Med 2020;3(1):117. [doi: 10.1038/s41746-020-00325-z] [Medline: 32964140]
42. Kastner M, Makarski J, Hayden L, et al. Effectiveness of an eHealth self-management tool for older adults with multimorbidity (KeepWell): protocol for a hybrid effectiveness-implementation randomised controlled trial. BMJ Open 2021 Feb 17;11(2):e048350. [doi: 10.1136/bmjopen-2020-048350] [Medline: 33597147]
43. Son YJ, Kim SW, Lee WS, et al. The effects of a 24-week interactive text message-based mobile health intervention for enhancing self-care behaviours of patients with heart failure: a quasi-experimental study. Nurs Open 2023 Sep;10(9):6309-6319. [doi: 10.1002/nop2.1878] [Medline: 37313589]
44. Nahm ES, Zhu S, Bellantoni M, et al. The effects of a theory-based patient portal e-Learning program for older adults with chronic illnesses. Telemed J E Health 2019 Oct;25(10):940-951. [doi: 10.1089/tmj.2018.0184] [Medline: 30431393]
45. Lyles CR, Tieu L, Sarkar U, et al. A randomized trial to train vulnerable primary care patients to use a patient portal. J Am Board Fam Med 2019;32(2):248-258. [doi: 10.3122/jabfm.2019.02.180263] [Medline: 30850461]
46. Barbati C, Maranesi E, Giammarchi C, et al. Effectiveness of eHealth literacy interventions: a systematic review and meta-analysis of experimental studies. BMC Public Health 2025 Jan 23;25(1):288. [doi: 10.1186/s12889-025-21354-x] [Medline: 39849354]
47. Dong Q, Liu T, Liu R, Yang H, Liu C. Effectiveness of digital health literacy interventions in older adults: single-arm meta-analysis. J Med Internet Res 2023 Jun 28;25:e48166. [doi: 10.2196/48166] [Medline: 37379077]
48. Guyatt G, Schandelmaier S, Brignardello-Petersen R, et al. Core GRADE 3: rating certainty of evidence-assessing inconsistency. BMJ 2025 May 6;389:e081905. [doi: 10.1136/bmj-2024-081905] [Medline: 40328467]
49. Guyatt G, Zhao Y, Mayer M, et al. GRADE guidance 36: updates to GRADE's approach to addressing inconsistency. J Clin Epidemiol 2023 Jun;158:70-83. [doi: 10.1016/j.jclinepi.2023.03.003] [Medline: 36898507]
50. Yang Y, Yao X, Lu D, et al. Improving the eHealth literacy of older adults: a scoping review. Geriatr Nurs 2024;60:128-136. [doi: 10.1016/j.gerinurse.2024.07.028] [Medline: 39244798]
51. Ambrosi E, Mezzalira E, Canzan F, et al. Effectiveness of digital health interventions for chronic conditions management in European primary care settings: systematic review and meta-analysis. Int J Med Inform 2025 Apr;196:105820. [doi: 10.1016/j.ijmedinf.2025.105820] [Medline: 39922064]
52. Lu T, Lin Q, Yu B, Hu J. A systematic review of strategies in digital technologies for motivating adherence to chronic illness self-care. npj Health Syst 2025;2(1):13. [doi: 10.1038/s44401-025-00017-4]

XSL•FO
RenderX

53. Zhang K, Jiang Y, Xi K, Sun N. Application of persuasive system design in mobile health interventions for chronic disease management: a mini review. Front Public Health 2025;13:1718371. [doi: 10.3389/fpubh.2025.1718371] [Medline: 41358248]
54. Hovingh JW, Elderson-van Duin C, Kuipers DA, et al. Tailoring for health literacy in the design and development of eHealth interventions: systematic review. JMIR Hum Factors 2025 Sep 2;12:e76172. [doi: 10.2196/76172] [Medline: 40896843]
55. Wilson S, Tolley C, Mc Ardle R, et al. Recommendations to advance digital health equity: a systematic review of qualitative studies. NPJ Digit Med 2024 Jun 29;7(1):173. [doi: 10.1038/s41746-024-01177-7] [Medline: 38951666]
56. Badr J, Motulsky A, Denis JL. Digital health technologies and inequalities: a scoping review of potential impacts and policy recommendations. Health Policy 2024 Aug;146:105122. [doi: 10.1016/j.healthpol.2024.105122] [Medline: 38986333]
57. Kraef C, van der Meirschen M, Free C. Digital telemedicine interventions for patients with multimorbidity: a systematic review and meta-analysis. BMJ Open 2020 Oct 13;10(10):e036904. [doi: 10.1136/bmjopen-2020-036904] [Medline: 33051232]
58. Hietbrink EAG, Lansink C, Laverman GD, Vollenbroek-Hutten MMR, Middelweerd A, Tabak M. Systematic review of dynamically tailored eHealth interventions targeting physical activity and healthy diet in chronic disease. NPJ Digit Med 2025 Nov 19;8(1):696. [doi: 10.1038/s41746-025-02054-7] [Medline: 41261160]
59. Hepburn J, Williams L, McCann L. Barriers to and facilitators of digital health technology adoption among older adults with chronic diseases: updated systematic review. JMIR Aging 2025 Sep 11;8:e80000. [doi: 10.2196/80000] [Medline: 40934502]
60. Grady A, Pearson N, Lamont H, et al. The effectiveness of strategies to improve user engagement with digital health interventions targeting nutrition, physical activity, and overweight and obesity: systematic review and meta-analysis. J Med Internet Res 2023 Dec 19;25:e47987. [doi: 10.2196/47987] [Medline: 38113062]
61. Elkes J, Cro S, Batchelor R, et al. User engagement in clinical trials of digital mental health interventions: a systematic review. BMC Med Res Methodol 2024 Aug 24;24(1):184. [doi: 10.1186/s12874-024-02308-0] [Medline: 39182064]
62. Liu K, Xie Z, Or CK. Effectiveness of mobile app-assisted self-care interventions for improving patient outcomes in type 2 diabetes and/or hypertension: systematic review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2020 Aug 4;8(8):e15779. [doi: 10.2196/15779] [Medline: 32459654]
63. Phi NTT, Montori VM, Kunneman M, Ravaud P, Tran VT. Cumulative burden of digital health technologies for patients with multimorbidity: a systematic review. JAMA Netw Open 2025 Apr 1;8(4):e257288. [doi: 10.1001/jamanetworkopen.2025.7288] [Medline: 40279126]
64. Smith L, Simpson G, Holt S, Dambha-Miller H. Digital applications to support self-management of multimorbidity: a scoping review. Int J Med Inform 2025 Oct;202:105988. [doi: 10.1016/j.ijmedinf.2025.105988] [Medline: 40424867]
65. Sanders M, Tobin JN, Cassells A, et al. Can a brief peer-led group training intervention improve health literacy in persons living with HIV? Results from a randomized controlled trial. Patient Educ Couns 2021 May;104(5):1176-1182. [doi: 10.1016/j.pec.2020.10.031] [Medline: 33221117]
66. Kizilkilic SE, Xu L, Akinci B, et al. Educational methods to improve digital health literacy: a systematic review and meta-analysis for the EAPC Opti(MI)se initiative. Eur J Prev Cardiol 2025 Jul 15:zwaf354. [doi: 10.1093/eurjpc/zwaf354] [Medline: 40659537]
67. Chong SOK, Pedron S, Abdelmalak N, Laxy M, Stephan AJ. An umbrella review of effectiveness and efficacy trials for app-based health interventions. NPJ Digit Med 2023 Dec 16;6(1):233. [doi: 10.1038/s41746-023-00981-x] [Medline: 38104213]
68. Eaton C, Vallejo N, McDonald X, et al. User engagement with mhealth interventions to promote treatment adherence and self-management in people with chronic health conditions: systematic review. J Med Internet Res 2024 Sep 24;26:e50508. [doi: 10.2196/50508] [Medline: 39316431]
69. Pong C, Tseng R, Tham YC, Lum E. Current implementation of digital health in chronic disease management: scoping review. J Med Internet Res 2024 Dec 12;26:e53576. [doi: 10.2196/53576] [Medline: 39666972]

Abbreviations

CBM: Chinese Biomedical Literature
CNKI: China National Knowledge Infrastructure
eHEALS: eHealth Literacy Scale
GRADE: Grading of Recommendations, Assessment, Development, and Evaluation
mHealth: mobile health
PI: prediction interval
PICOS: population, intervention, comparator, outcome, and study design
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search
RCT: randomized controlled trial
SMD: standardized mean difference

XSL•FO
RenderX

Edited by S Brini; submitted 07.Aug.2025; peer-reviewed by YC Chen, Y Luo; revised version received 09.Mar.2026; accepted 09.Mar.2026; published 17.Apr.2026.

Please cite as:
Yang X, Ma Y, Wang Y, Liu M, Liu X, Jiao X, Yan F, Ma Y, Han L, Zhang Y
Effectiveness of mHealth Interventions for Improving eHealth Literacy Among Patients With Chronic Diseases: Meta-Analysis and Systematic Review
J Med Internet Res 2026;28:e82004
URL: https://www.jmir.org/2026/1/e82004
doi:10.2196/82004

© Xueqin Yang, Yingping Ma, Yueting Wang, Mengting Liu, Xingyu Liu, Xueping Jiao, Fanghong Yan, Yuxia Ma, Lin Han, Yanan Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Determinants of the Uptake and Frequency of Use of a Web Portal Digital Health Intervention in Patients With Type 2 Diabetes and/or Coronary Heart Disease: Secondary Analysis of a Randomized Controlled Trial

Maximilian Scholl[1], MA; Claas Lendt[1,2], MSc; Sebastian Appelbaum[3,4], MSc; Bianca Biallas[1], PhD; Katja Brenk-Franz[5], PhD; Chloé Chermette[6], PhD; Friederike Frank[7,8], MSc; Angeli Gawlik[6], PhD; Lisa Giesen[9], PhD; Martina Heßbrügge[10], MD; Lucas Küppers[11], MD; Larisa Pilic[12,13], MSc; Marcus Redaèlli[9], MD; Lara Schneider[14], BA; Frank Vitinius[14,15], MD; Stefan Wilm[16], MD; Uwe Konerding[3,4], PhD

[1]Institute of Movement Therapy and Movement-oriented Prevention and Rehabilitation, Section I, German Sports University Cologne, Cologne, Germany

[10]Institute of General Practice, University of Duisburg-Essen, Essen, Germany

[11]Institute of Family Medicine and General Practice, University of Bonn, Bonn, Germany

[12]Institute of General Practice, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany

[13]Department of Child and Adolescent Psychiatry, Psychosomatic Medicine and Psychotherapy, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany

[14]Department of Psychosomatics and Psychotherapy, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany

[15]Department of Psychosomatic Medicine, Robert Bosch Hospital, Stuttgart, Germany

[16]Institute of General Practice, Centre for Health and Society, Medical Faculty and University Hospital, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[2]Department of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany

[3]Department of Psychology and Psychotherapy, Witten/Herdecke University, Witten, Germany

[4]Trimberg Research Academy, University of Bamberg, An der Weberei 5, Bamberg, Germany

[5]Institute of Psychosocial Medicine, Psychotherapy and Psychooncology, Jena University Hospital, Jena, Germany

[6]Institute of Psychology, Section I: Health and Social Psychology, German Sport University Cologne, Cologne, Germany

[7]Institute for Digitalization and General Medicine, Medical Faculty, RWTH Aachen University, Aachen, Germany

[8]Center for Rare Diseases Aachen (ZSEA), Medical Faculty, RWTH Aachen University, Aachen, Germany

[9]Institute of Health Economics and Clinical Epidemiology, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany

Corresponding Author:
Uwe Konerding, PhD
Department of Psychology and Psychotherapy, Witten/Herdecke University, Witten, Germany

Abstract

Background: The targeted application and design of digital health interventions (DHIs) require an understanding of usage determinants. Usage includes uptake (initial use) and frequency (extent of use), but it is unclear whether both components are driven by the same determinants.

Objective: This study aimed to examine the determinants of uptake and frequency of use and assess whether they differ.

Methods: The investigated DHI was a web portal provided in an intervention for improving disease-related self-management. This study is a secondary analysis of intervention group data from a parallel-group randomized controlled trial. Eligibility criteria were being an adult and being diagnosed with type 2 diabetes and/or coronary heart disease. Sociodemographic, psychological, and health-related variables were examined as determinants. Determinants were analyzed using simple and multiple regression models. Uptake was analyzed using logistic regression, and frequency was analyzed using negative binomial regression with robust SEs. Frequency was analyzed for those who used the DHI at least once. Except for sociodemographic variables, all other variables were standardized to a range from 0 to 1. For simple regression, inflation of the α error due to multiple testing was controlled via the approach of Benjamini and Hochberg, and for multiple regression, it was controlled via the significance of the complete multiple regression model.

Results: Of 462 intervention group members, 199 (43.1%) used the web portal at least once. After controlling for inflation of the α error, simple regression for uptake yielded significant effects for higher education (B=0.56, 95% CI 0.18-0.95; *P*=.004),

XSL•FO
RenderX

openness (B=1.08, 95% CI 0.33-1.83; *P*=.005), intention regarding physical activity (B=2.28, 95% CI 1.30-3.26; *P*<.001), and intention regarding healthy nutrition (B=2.30, 95% CI 1.30-3.31; *P*<.001). The multiple regression model for uptake was highly significant (*P*<.001), with significant positive associations for intentions regarding physical activity (B=1.86, 95% CI 0.74-2.97; *P*=.001) and healthy nutrition (B=2.22, 95% CI 1.00-3.44; *P*<.001), as well as a significant negative association for patient activation (B=−3.20, 95% CI −4.95 to −1.46; *P*<.001). After controlling for inflation of the α error, simple regression for frequency yielded no statistically significant effect, and the multiple regression model for frequency was not significant (*P*=.07).

Conclusions: This study is innovative in jointly examining determinants of the uptake and frequency of use of the same DHI within a single context and sample. By demonstrating that factors driving uptake do not necessarily increase the frequency of use, it advances existing research. The study contributes to a more differentiated understanding of DHI use and shows that distinct strategies are required to promote adoption versus sustained engagement. Applying this approach to other DHIs and settings may support more targeted and equitable digital health implementation in real-world contexts, thereby optimizing digital health deployment strategies overall.

Trial Registration: German Clinical Trials Registry (DRKS) DRKS00020592; https://drks.de/search/de/trial/DRKS00020592

International Registered Report Identifier (IRRID): RR2-10.1186/s13063-021-05636-4

(J Med Internet Res 2026;28:e80895) doi:10.2196/80895

KEYWORDS

digital health interventions; eHealth; type 2 diabetes; coronary heart disease; predictors of use; chronic disease self-management

Introduction

Chronic noncommunicable diseases, such as type 2 diabetes mellitus (T2DM) and coronary heart disease (CHD), place an enormous burden on patients and health care providers worldwide [1,2]. Improving patient self-management could help to reduce this burden [3,4]. The use of digital health interventions (DHIs), that is, interventions meant to influence health via digital tools, is a promising approach that might serve this purpose. Some examples are patient portals [5], mobile health apps [6], and web-based interventions [7]. DHIs make it possible to provide personalized, location-independent, and cost-effective support to patients [8,9]. Accordingly, DHIs have already been used in rehabilitation [10], medication adherence [11], and chronic disease management [12-15], as well as in the promotion of healthy lifestyles [16,17]. Recent systematic reviews and meta-analyses have indicated that DHIs help to improve hemoglobin A_{1c} levels [13,18,19], CHD status [14], and health-promoting behaviors, such as physical activity and healthy nutrition intake [14,20].

DHIs can only be effective if they are actually used. In some cases, repeated use might be necessary to have an impact [21,22]. Accordingly, purposeful design of interventions requires knowledge regarding the determinants of usage, that is, knowledge regarding conditions under which the DHI in question is used more and knowledge regarding the conditions under which the DHI is used less. Knowledge regarding the conditions under which the DHI is used less can be helpful in two regards: (1) it might show how use could be increased by adapting the DHI itself and/or its mode of presentation, and (2) it might show where usage is limited, that is, when neither the condition can be changed nor the DHI can be adapted. In this case, alternative interventions for the purpose addressed by the DHI in question are needed, and these might very well be nondigital interventions.

Previous research regarding potential determinants of DHI usage primarily focused on sociodemographic factors, especially age, gender, and education. Findings regarding age [23-26] and gender [24-27] were inconsistent across studies, whereas education showed a more consistent positive association with DHI usage [28-30]. Further, previous research regarding the potential determinants of DHI referred to psychological variables, such as the “big five” [29]; loneliness, depression, and anxiety [31]; health literacy [31,32]; patient activation [33,34]; and health-related behavioral intentions [26,35]. Among the “big five,” openness appeared to be positively associated with usage, whereas extraversion appeared to be negatively associated [29]. There were mixed results for mental health factors [31,35] and behavioral intentions [26,35], whereas health literacy appeared to be positively associated with usage [31,32]. Moreover, mixed results were reported for patient activation, with some studies reporting no association [36] and others reporting a positive association [37].

A major problem in research regarding the determinants of DHI usage is the operationalization of usage itself [38]. In previous research, the following two different operationalizations were mainly applied: (1) uptake, that is, the act of using the DHI at least once [26,30,39], and (2) frequency, that is, the amount of use thereafter [23-25]. Uptake should, at best, be investigated in all persons who have used the DHI, and frequency should be investigated only in those who have used it at least once. Both operationalizations constitute different components of usage. These 2 components might be determined by completely different factors. Therefore, in research regarding the determinants of DHI use, both components should be distinguished from each other. This is not only important for the sake of scientific precision but also relevant with regard to deriving implications for the targeted design of the DHI and the corresponding communication and outreach strategies. Insights into the determinants of uptake are important for the design of communication and outreach strategies, whereas insights into the determinants of frequency are relevant for the design of the DHIs themselves.

XSL•FO
RenderX

The best approach for identifying which variables affect uptake and which affect frequency is to investigate both effects for the same DHI in the same context within the same sample. There is hardly any previous research on this approach. The objective of this study is to demonstrate this approach, that is, to investigate the determinants of uptake and the frequency of use for the same DHI in the same context within the same sample.

Methods

Study Design

This study is a secondary analysis of a randomized controlled trial (RCT) involving a complex intervention called the Personalized Self-management Support Program (P-SUP; 2019 - 2024) [40]. The P-SUP consists of the following four modules: (1) peer support groups, which met for weekly group-based exercise sessions and regularly scheduled digital expert education classes; (2) telephone coaching, which is provided to a subgroup of patients with low health literacy and low patient activation [41]; (3) feedback reports, which are provided to patients and contain routine data on medical parameters that can be used as a basis for discussing their state of health with their general practitioners during routine check-ups; and (4) a browser-based web portal, which provides practical and theoretical content on exercise, nutrition, and motivation. The last component is the DHI investigated in this study.

The P-SUP was implemented and evaluated within disease management programs (DMPs) [42] for T2DM and CHD in North Rhine-Westphalia, Germany. DMPs are structured treatment programs that address individuals with chronic conditions and that are added to standard primary care. The trial followed a parallel-group randomized controlled design with a delayed intervention for the control group and was conducted consecutively (ie, participants started sequentially rather than simultaneously). Participants were randomized to the intervention group and control group with an allocation ratio of 6:7. In the first 18 months, the intervention group received the intervention, whereas the control group received regular DMP care. After this interval, the control group was offered the intervention for 9 months. The primary outcome of the original trial was the number of hospital days during the year preceding the end of the intervention. Secondary outcomes covered three domains: (1) self-reported physical health status, (2) self-reported health-related behaviors, and (3) self-reported thoughts and feelings (ie, psychological variables).

Patient and Public Involvement

No formal patient or public involvement took place in the design, conduct, or reporting of the trial.

Changes to the Trial Protocol

In some aspects, the conduction of the trial deviated from the protocol. Not all peer support group meetings could be guided by members of the groups because not enough participants were willing to take on this role. Moreover, some peer support group meetings had to be held online because of the COVID-19 pandemic, and the feedback report only started with a delay due to organizational problems.

Harms

In accordance with the trial protocol, harms and potential adverse events were monitored and documented during peer support group meetings and telephone coaching sessions. No adverse events or harms were reported.

Sample Size

The sample size of the original trial was determined based on the primary outcome (ie, the number of hospital days in the last year of the intervention interval) and was powered to detect small effect sizes (Cohen d=0.2) with a significance level of .05 and a power of 0.80, resulting in a required sample size of 1001 participants, as detailed in the trial protocol [40]. No separate power calculation was conducted for this secondary analysis, which focused on a specific intervention component and on the intervention group alone.

Investigated DHI

The browser-based web portal provided as part of the complete intervention (ie, the DHI investigated in this study) was developed based on the self-determination theory (SDT) [43]. Among other things, the SDT posits that individuals strive to fulfill the following 3 basic psychological needs: autonomy (eg, possession of choice and action scope), competence (eg, personal growth and skill development), and social relatedness (eg, relationship building and integration). The DHI was developed within this conceptual framework, with a particular focus on supporting autonomy and enhancing competence. The development was conducted in three stages: (1) a workshop, (2) content design by core teams, and (3) consolidation of content and posting on the web portal.

The aim of the workshop was to define the content, communication style, and visual design appropriate for the target group. This workshop was guided by an external digital agency based in Cologne. Participants of the workshop included members of the interdisciplinary project team (sports science, health science, psychology, and medicine), a practicing physician, and a potential study participant. Based on the workshop results, content and functions for the web portal were developed. Subsequently, the following four core teams were established (each for a different topic): (1) medicine, (2) physical activity, (3) nutrition, and (4) motivation. These core teams systematically developed the intervention content. In the final step, the developed content was consolidated and successively integrated into the DHI by a specific project team member.

The web portal included the following: (1) evidence-based information regarding T2DM and CHD, physical activity, nutrition, and self-motivation; (2) exercise videos conducted by a sports therapist focusing on strength, endurance, coordination, and flexibility; (3) recipes for healthy food; (4) a motivational 12-week coaching program; and (5) a download area for nutritional help tools (Figure 1). Usability and visual design of the complete web portal were tested formatively during the development process with a cohort comparable to the final study sample. The 12-week coaching program was developed iteratively according to the Integrate, Design, Assess, and Share (IDEAS) framework [44-46]. For further details on the content and design of the web portal, see Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. Screen capture images from the P-SUP (Personalized Self-management Support Program) web portal.

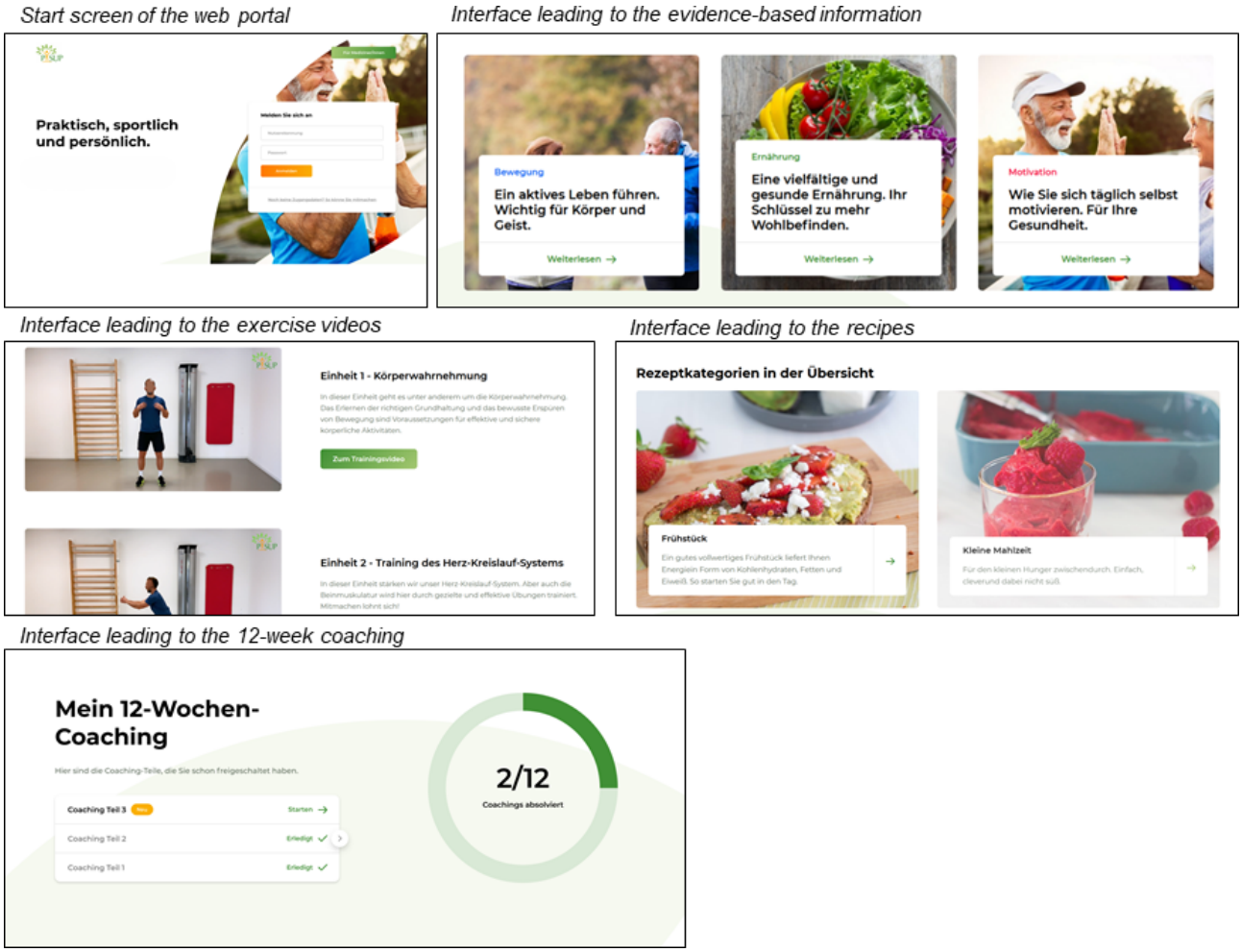

The final web portal was made accessible to all participants of the intervention group by sending them a log-in identification number and a password by mail prior to the start of the study. In addition to the log-in credentials, they received a letter that included an explanation of the web portal contents, data protection information (eg, data storage, server location, and privacy), a step-by-step guide for the registration process, and contact details (email and telephone) for technical support. To maintain participant interest in the application, the web portal was updated with new content at irregular intervals. To accompany these updates and support user engagement, a total of 10 newsletters were distributed via email at regular intervals throughout the intervention period. These newsletters explained the different components of the P-SUP intervention and provided direct links to newly published content.

Study Participants

The study participants of the complete RCT were recruited between February and October 2021 from 6 different regions in North Rhine-Westphalia, Germany, including Aachen, Bonn, Cologne, Düsseldorf, Essen, and Bünde, via general practitioners, newspapers, and online advertisements. General practitioners informed eligible patients about the intervention during routine DMP appointments and enrolled them directly. Individuals who responded to newspaper or online advertisements were enrolled in the study by the study team if they met the inclusion criteria. The inclusion criteria for patients were as follows: (1) being ≥18 years of age, (2) being enrolled in a DMP for T2DM and/or CHD [42], and (3) being a resident of North Rhine-Westphalia. The exclusion criteria were as follows: (1) insufficient German language skills, (2) severe cognitive or physical impairments, and (3) severe comorbidities. Those persons who were assigned to the intervention group and who were finally included in the analyses for examining the effectiveness of the complete intervention were also included in the secondary analysis presented in this article.

Recruiting Sites and Personnel

Recruitment was conducted through general practitioners participating in the DMPs for T2DM and/or CHD. The eligibility for physician recruitment was limited to standard DMP accreditation. No additional eligibility criteria, training, or intervention delivery responsibilities were applicable, and no additional personnel involved in delivering or supporting the DHI were investigated in this secondary analysis.

Data Collection and Measures

DHI Usage

DHI usage was determined based on server data from the web portal, covering the entire 18-month intervention period. Because of the consecutive nature of the study, these periods began in the period from April 2021 to February 2022. The

XSL•FO
RenderX

collected data included information on log-ins and viewed content.

However, for some participants, the documentation of web portal usage did not start with an identified log-in. This indicated technical problems with the documentation. As a result, the log-ins, which were meant to be the basic units of DHI usage, could not be determined via the log-in data. Instead, log-ins were partly reconstructed based on timestamps of viewed content. A new log-in was assumed whenever there was a gap of at least 2 hours between 2 consecutive content views. This threshold was deliberately set longer than the 30 or 60 minutes recommended in previous literature [47,48], because certain types of content on the web portal, such as exercise videos, lasted around 30 minutes. Thus, if a participant logged into the web portal and subsequently watched 1-2 videos, the activity following this should still be attributed to the same session.

Uptake was operationalized as the incidence of at least one log-in, whereas frequency was operationalized as the total number of log-ins minus the first one. The first log-in was subtracted because the frequencies were analyzed to obtain information about processes that occur after the first log-in. As a sensitivity analysis, all analyses referring to frequencies were also performed with log-ins defined using 30- and 60-minute intervals.

All variables determined for operationalizing usage only reflected direct use of the DHI. Indirect use was possible by downloading exercise videos or healthy food recipes and using them offline. This shifted some web portal activities from tracked online sessions to unobserved offline use and thus likely led to an underestimation of usage frequency.

Possible Determinants of DHI Usage

Possible determinants of DHI usage were selected from two sources: (1) register data from the organization unit coordinating the whole project, and (2) data from the baseline survey conducted prior to the start of the intervention. The register data were collected and updated from January 2021 to April 2024. The baseline survey was conducted from February to December 2021. The criterion for selecting variables from these 2 sources was that these variables had already been considered as possible determinants in previous research regarding the determinants of DHI usage.

Sociodemographic Variables

Sociodemographic characteristics included age, gender, and educational attainment. Age and gender were assessed via both the register data and the baseline survey. Educational attainment was only assessed via the baseline survey. In this survey, participants were asked whether they had the qualification required to attend a higher education institution, such as a university of applied sciences or an ordinary university. Those who answered “yes” were classified as having a higher level of educational attainment, while those who answered “no” were classified as having a lower level. Data for age and gender were available for all participants. Missing data for education were multiply imputed by producing 50 different datasets using the multiple imputation procedure provided as standard by SPSS (IBM Corp).

Psychological Variables

Psychological variables were solely assessed via the baseline survey. These variables can be assigned to two different categories: (1) general psychological variables, and (2) health-related psychological variables.

Five general psychological variables included the “big five” (ie, openness, conscientiousness, extraversion, agreeableness, and neuroticism). Three further general psychological variables included loneliness, depression, and anxiety. The “big five” variables were assessed using the Big Five Inventory-10 items (BFI-10) [49], which consists of 10 items, with each dimension addressed by 2 items. Loneliness was assessed using the 6-item short version of the revised University of California Los Angeles (UCLA) Loneliness Scale [50,51]. Depression and anxiety were assessed using the Patient Health Questionnaire-4 items (PHQ-4). The PHQ-4 is a 4-item self-report questionnaire that measures depression using the first 2 items and anxiety using the last 2 items [52].

The health-related psychological variables were health literacy, patient activation, and intentions regarding physical activity or healthy nutrition. Health literacy was assessed using the 5-item Brief Health Literacy Screen (BHLS) [53], and patient activation was evaluated using the Patient Activation Measure-13 (PAM13-D) [54,55]. Each of the 2 intention variables was assessed using 1 item with a 7-category scale. Physical activity was operationalized as engaging in at least 150 minutes of moderate or vigorous physical activity per week (“Do you intend to engage in at least two-and-a-half hours of moderate or vigorous physical activity per week in the coming months?”), and healthy nutrition was operationalized as performing at least five of nine items of healthy nutrition behavior (“Do you intend to follow at least five of the dietary behaviors described above in the coming months?”) listed in a self-constructed questionnaire (Multimedia Appendix 2).

Missing data for psychological variables were multiply imputed by producing 50 different datasets using the multiple imputation procedure provided as standard by SPSS. Except for the PAM13-D, all variables based on more than one item were constructed by summing up the scale values assigned to the different answer categories, with all answer categories coded with equal distances between all adjacent categories. For the PAM13-D, the sums determined in this way were additionally transformed into scale values obtained by Rasch scaling [56] (Multimedia Appendix 3).

All psychological variables were standardized via a linear transformation, with 0 being the lowest and 1 the highest value that could be obtained with the respective measurement instrument. This standardization was performed because all these variables possess at best an interval scale level, and thus, the corresponding measurement values have no absolute meaning but only a meaning relative to 2 reference points. For all variables, the boundaries of the corresponding measurement instrument constitute the best available reference points, and coding these reference points with 0 and 1 makes it possible to compare the magnitudes of the effects of these variables on DHI usage. All variables coded this way were pooled according to the denomination applied for the respective variable (ie, 1 is

XSL•FO
RenderX

the highest level of extraversion, agreeableness, openness, conscientiousness, neuroticism, loneliness, depression, anxiety, patient activation, health literacy, intention regarding physical activity, or intention regarding healthy nutrition).

Statistical Analysis

Patient characteristics were documented based on complete data. This was put into practice separately for nonusers and users, as well as for all investigated persons together. Gender and education were described using absolute and relative frequencies; all other variables were described using means and SDs. To provide a clear overview of overall usage of the web portal, the 18-month intervention period (78 weeks) was divided into 13 intervals of 6 weeks each. This segmentation was chosen because 78 is divisible by 13, allowing for consistent time intervals across the entire period. For each 6-week interval, user activity was described using relative and absolute frequencies. Absolute frequencies included (1) the absolute frequency of log-ins per interval; (2) the absolute frequency of users per interval; and (3) the absolute frequency of first log-ins per interval. Relative frequencies included (1) the relative frequency of log-ins per interval, expressed as a percentage of all log-ins across all intervals; (2) the relative frequency of users per interval, expressed as a percentage of all user occurrences across all intervals; and (3) the relative frequency of first log-ins per interval, expressed as a percentage of all first log-ins across all intervals.

All further analyses were conducted with missing values being multiply imputed. In response to comments raised during the review process, a test for data missing completely at random (MCAR) was performed [57]. However, the results of this test have no implications for the interpretation of the results of the further analyses, because statistical analyses based on multiply imputed data are unbiased as long as the much less restrictive condition of MAR (ie, data missing at random) is fulfilled [58]. However, for this condition, no feasible statistical test presently exists.

The analyses regarding uptake and frequency were conducted with different samples. For uptake, these were all members of the intervention group, while for frequency, these were only those who had used the web portal at least once (ie, users). In both cases, three different analyses were performed: (1) an analysis of the correlations between the variables investigated as possible determinants; (2) simple regression analyses for all possible determinants, with either uptake or frequency as a criterion; and (3) a multiple regression analysis with all possible determinants entered as predictors in a common model, with either uptake or frequency as a criterion. If the usage variable was uptake, logistic regression was applied, while if the usage variable was frequency, negative binomial regression with data-based (robust) SEs was applied. The latter procedure was chosen because negative binomial regression is specifically designed for count variables and because data-based SEs are not based on a specific model, so that inference statistical tests based on these errors cannot be compromised by invalidity of the presupposed model [59].

Several different statistics were computed in connection with the regression analyses. These were regression coefficients, corresponding 95% CIs, and corresponding *P* values for all regression analyses. In logistic regression, Nagelkerke pseudo R^2 was used to assess how well the determinants explain the actual values. In negative binomial regression, the proportion of explained variance was used for this purpose. For negative binomial regression, the dispersion parameter was also computed. This parameter refers to the variance of the actual values that are predicted by the same predicted value. If this variance is equal to or lower than the predicted value, the dispersion parameter is 0. Otherwise, this parameter is larger than 0, with larger dispersion parameters indicating larger variances. To investigate how the different predictor variables interfere with each other when regression coefficients in multiple regression are estimated, variance inflation factors were computed for both multiple regressions. These factors are equal to 1 if there is no interference at all, and they increase with increasing interference. With a variance inflation factor larger than 5, the estimates of the regression coefficients are not very trustworthy.

Inflation of α errors due to multiple testing was controlled differently for simple and multiple regression. For simple regression, the approach of Benjamini and Hochberg was applied [60]. In this approach, a specific comparator is computed for each *P* value. For this purpose, the *P* values for the different tests are arranged in ascending order. The comparator for each *P* value is as follows:

$$\text{comparator}_i = \frac{i}{m}\alpha \quad (1)$$

where *i* is the test number in ascending order, *m* is the number of tests, and α is the significance level applied for an individual test. The largest *i* for which the *P* value is smaller than the comparator is identified, and all tests with an index smaller than or equal to this *i* are considered to be significant. In multiple regression, the final decision regarding the significance was based on the result of the significance test of the complete regression model. If this test yielded a significant result, all significant results of tests regarding the individual regression coefficients were considered as actually significant; otherwise, none of the individual test results were considered as significant.

To compare the regression models for uptake and frequency, the rank correlation between the *P* values was computed.

All data were analyzed using IBM SPSS Statistics version 29.

Ethical Considerations

Ethical approval for the complete RCT was obtained from the Ethics Committee of the Medical Faculty of the University Hospital Cologne in July 2020 (20-1155). All patients provided written informed consent prior to participation in the RCT. This consent also covered the use of pseudonymized data for secondary analyses without requiring additional consent. Patient data were stored within three separate organizational units at the University Hospital Cologne: (1) Trust Center, which held address data linked to study identification numbers; (2) Data Warehouse, which held evaluation data linked to study identification numbers; and (3) Multiple Pseudonym Assignment Unit, which maintained the assignment list linking personal and study identification numbers. These units were strictly separated both digitally and in terms of personnel, ensuring confidentiality

XSL•FO
RenderX

and preventing reidentification. Participants received no financial incentives.

No images in the article or supplementary materials allow for the identification of individual participants. Any visible human faces have been deliberately blurred to ensure anonymity, or the faces originate from commercially available stock images and do not depict study participants. No consent from identifiable individuals was therefore required.

Results

General Information

A total of 1227 patients were enrolled in the RCT, with 566 patients allocated to the intervention group and 661 patients allocated to the control group. Ultimately, 1002 patients were included in the RCT analyses (intervention group: n=462; control group: n=540; Figure 2). The analyses presented here are restricted to the 462 participants allocated to the intervention group with available baseline questionnaire data and register data at the start of the intervention. The flow diagram in Figure 2 follows the structure recommended by CONSORT (Consolidated Standards of Reporting Trials) 2025 [61]; however, only the elements relevant to this secondary analysis are addressed in the flow diagram. The CONSORT checklist is presented in Checklist 1. Follow-up assessments and analyses of primary and secondary outcomes were conducted as part of the trial but are not relevant to the analyses reported here and will be presented in the effectiveness evaluation in a separate publication.

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flow diagram. DHI: digital health intervention; DMP: disease management program; RCT: randomized controlled trial.

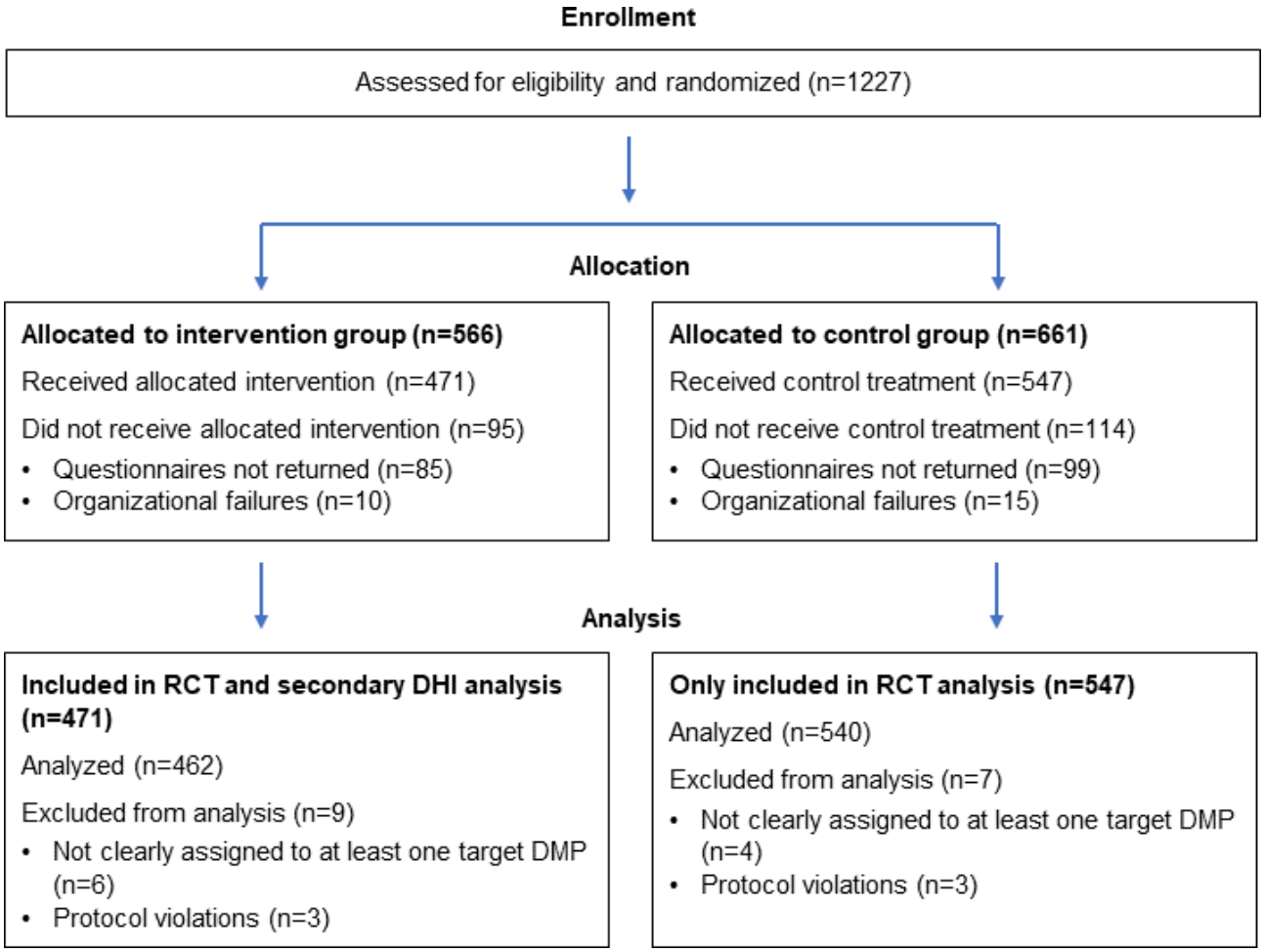

Among the data included in the secondary analysis, an average of 2.2% (range 0.7%-4.8%) of data were missing. MCAR was not fulfilled (P=.008). The mean participant age at the time of the survey was 64.8 (SD 10.2) years. Moreover, 49.8% (230/462) of participants were female, 38.6% (170/462) had a university entrance qualification, and 43.1% (199/462) had at least one log-in to the web portal (Table 1). Participants in the last group were classified as users.

XSL•FO
RenderX

Table . Patient characteristics[a].

Variable	Nonusers		Users		Total	
	n[b]	Statistics	n[b]	Statistics	n[b]	Statistics
Sociodemographic variable						
Age (years), mean (SD)	263	65.4 (11.0)	199	64.0 (9.2)	462	64.8 (10.2)
Female gender[c], n (%)	263	125 (47.5)	199	92 (46.2)	462	230 (49.8)
Higher education, n (%)	248	81 (32.7)	192	89 (46.4)	440	170 (38.6)
Psychological variables[d]						
General psychological variables, mean (SD)						
Extraversion	254	0.60 (0.24)	195	0.60 (0.26)	449	0.60 (0.25)
Agreeableness	258	0.56 (0.21)	195	0.56 (0.19)	453	0.56 (0.20)
Openness	259	0.55 (0.26)	195	0.62 (0.24)	454	0.58 (0.25)
Conscientiousness	254	0.70 (0.21)	195	0.69 (0.20)	449	0.70 (0.21)
Neuroticism	261	0.49 (0.23)	195	0.48 (0.23)	456	0.49 (0.23)
Loneliness	254	0.32 (0.28)	194	0.35 (0.27)	448	0.33 (0.28)
Depression	255	0.24 (0.23)	198	0.24 (0.24)	453	0.24 (0.24)
Anxiety	255	0.23 (0.25)	198	0.22 (0.23)	453	0.23 (0.24)
Health-related psychological variables, mean (SD)						
Patient activation	254	0.67 (0.16)	197	0.65 (0.13)	451	0.66 (0.15)
Health literacy	250	0.70 (0.21)	196	0.73 (0.18)	446	0.71 (0.20)
Intention–physical activity	257	0.79 (0.27)	198	0.89 (0.16)	455	0.84 (0.23)
Intention–healthy nutrition	261	0.78 (0.25)	198	0.88 (0.15)	459	0.82 (0.22)

[a]Patients are members of the intervention group of the randomized controlled trial conducted within the P-SUP (Personalized Self-Management Support Program) project. They were enrolled in disease management programs for type 2 diabetes and/or coronary heart disease and were residing in North Rhine-Westphalia, Germany. Data were collected during the baseline assessment of the P-SUP project.

[b]Number of persons with complete data for the respective variable.

[c]None of the participants reported their gender as diverse.

[d]For better comparability, all psychological variables were scaled to a range from 0 to 1. Higher values indicate a higher expression of the construct.

Among users, 86.4% (172/199) had their first log-in within the first 6 weeks of the complete intervention period. The final first log-in was recorded in week 39. The total number of log-ins was 637 within these first 6 weeks. In the following 6-week intervals, both the number of persons using the web portal and the total number of log-ins decreased almost continuously (Figure 3). With the first log-in included, participants classified as users visited the web portal on average 7.4 (SD 8.5; range 1 - 50) times, if log-ins were determined using an interval of 120 minutes between 2 subsequent activities on the web portal. For intervals of 60 and 30 minutes, the corresponding values were 7.7 (SD 8.8; range 1 - 52) times and 8.2 (SD 9.5; range 1 - 59) times, respectively.

XSL•FO
RenderX

Figure 3. User activity.

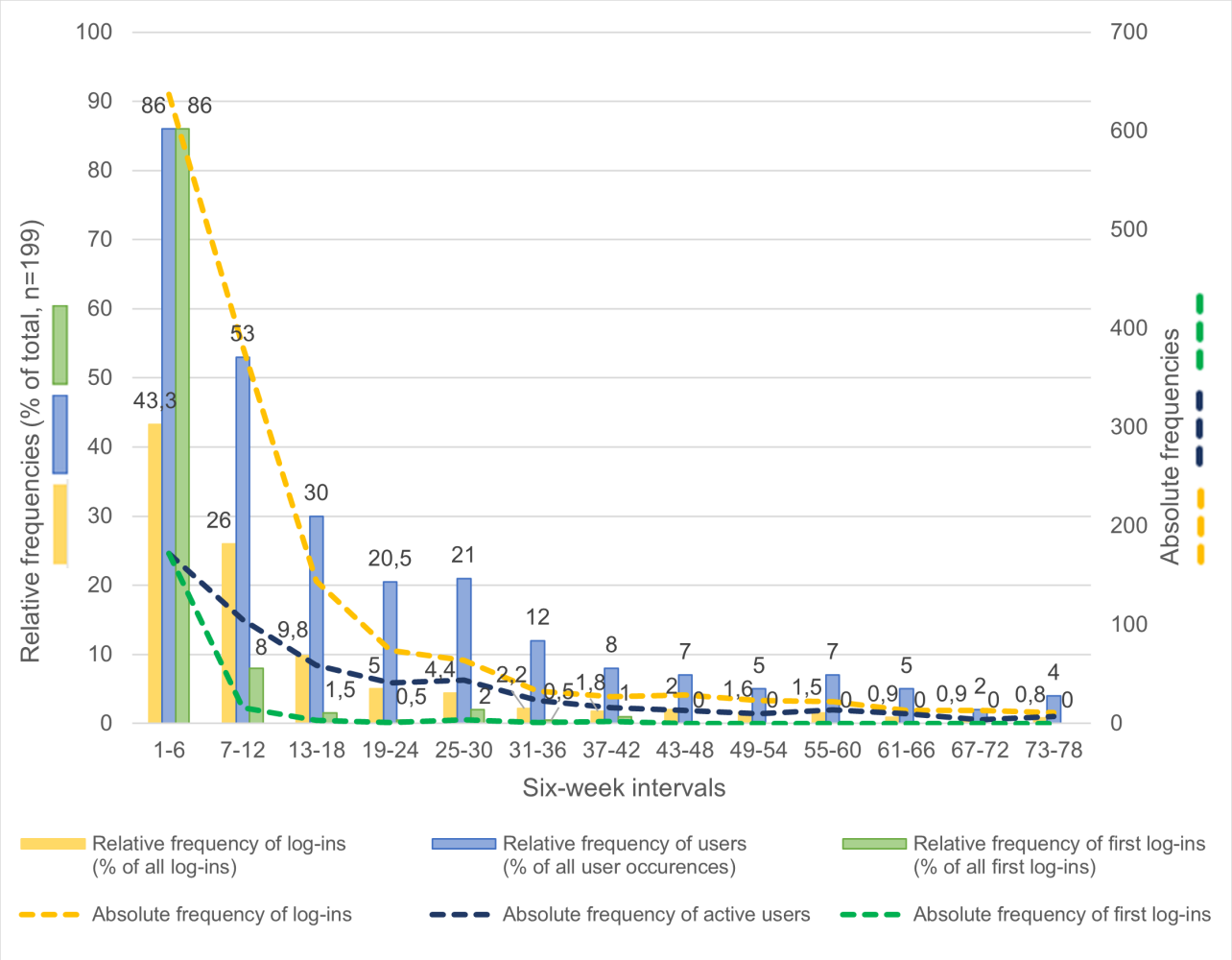

Uptake

Many of the variables taken as possible determinants for uptake correlated moderately with each other. This applied especially to loneliness, depression, and anxiety, as well as to intentions regarding physical activity and healthy nutrition (Multimedia Appendix 4). Nevertheless, the largest inflation factor was smaller than 3 (Table 2). Accordingly, all possible determinants could be included in the multiple regression.

XSL•FO
RenderX

Table . Simple and multiple regression analyses for predictors of uptake[a].

Variable	Simple regression analysis[b]			Multiple regression analysis[b]		
	B[c] (95% CI)	*P* value	NK[d]	B[c] (95% CI)	*P* value	VIF[e]
Sociodemographic variables						
Age in years	−0.01 (−0.03 to 0.00)	.12	0.006	−0.00 (−0.02 to 0.02)	.90	1.14
Female gender	−0.25 (−0.62 to 0.12)	.18	0.005	−0.30 (−0.73 to 0.12)	.16	1.15
Higher education	0.56 (0.18 to 0.95)	.004[f]	0.024	0.29 (−0.15 to 0.73)	.20	1.14
Psychological variables[g]						
General psychological variables						
Extraversion	−0.05 (−0.81 to 0.70)	.89	0.000	−0.08 (−1.00 to 0.84)	.87	1.27
Agreeableness	−0.18 (−1.09 to 0.73)	.70	0.000	0.02 (−1.01 to 1.06)	.97	1.09
Openness	1.08 (0.33 to 1.83)	.005[f]	0.024	0.81 (−0.06 to 1.67)	.07	1.18
Conscientiousness	−0.31 (−1.21 to 0.58)	.49	0.001	−0.03 (−1.11 to 1.04)	.95	1.23
Neuroticism	−0.14 (−0.94 to 0.67)	.74	0.000	−0.05 (−1.06 to 0.96)	.92	1.35
Loneliness	0.34 (−0.33 to 1.01)	.32	0.003	0.54 (−0.43 to 1.51)	.27	1.76
Depression	0.07 (−0.72 to 0.85)	.87	0.000	0.08 (−1.33 to 1.48)	.92	2.64
Anxiety	−0.08 (−0.85 to 0.68)	.83	0.000	−0.38 (−1.83 to 1.07)	.61	2.96
Health-related psychological variables						
Patient activation	−1.17 (−2.48 to 0.14)	.08	0.010	−3.20 (−4.95 to −1.46)	<.001[f]	1.36
Health literacy	1.05 (0.08 to 2.03)	.03[f]	0.014	1.01 (−0.26 to 2.29)	.12	1.43
Intention–physical activity	2.28 (1.30 to 3.26)	<.001[f]	0.071	1.86 (0.74 to 2.97)	.001[f]	1.23
Intention–healthy nutrition	2.30 (1.30 to 3.31)	<.001[f]	0.067	2.22 (1.00 to 3.44)	<.001[f]	1.34

[a]Patients are members of the intervention group of the randomized controlled trial conducted within the P-SUP (Personalized Self-Management Support Program) project. They were enrolled in disease management programs for type 2 diabetes and/or coronary heart disease and were residing in North Rhine-Westphalia, Germany. Data were collected during the baseline assessment of the P-SUP project.

[b]Simple and multiple regression analyses performed on data after multiple imputation (n=462).

[c]B: regression coefficient B.

[d]NK: Nagelkerke pseudo R^2.

[e]VIF: variance inflation factor.

[f]Significant value.

[g]For better comparability, all psychological variables were scaled to a range from 0 to 1. Higher values indicate a higher expression of the construct.

In the simple regression analyses, education, openness, health literacy, intention regarding physical activity, and intention regarding healthy nutrition had statistically significant effects on uptake (education: *P*=.004; openness: *P*=.005; health literacy: *P*=.03; intention-physical activity: *P*<.001; intention-healthy nutrition: *P*<.001; Table 2). In all 5 cases, these effects were positive, with the effects for the 2 intention variables being the largest. After controlling for multiple testing according to Benjamini and Hochberg [60], the effect for health literacy disappeared, but the effects for the other 4 variables with significant effects in the individual tests remained. In the multiple regression analysis, the pattern of statistically significant effects changed. Education (*P*=.20), openness (*P*=.07), and health literacy (*P*=.12) no longer had statistically significant effects. Instead, patient activation had a strong statistically significant negative effect on uptake (*P*<.001), that

XSL•FO
RenderX

is, persons with high patient activation were less likely to use the web portal. The intentions regarding physical activity and healthy nutrition remained statistically significant positive predictors (intention-physical activity: P=.001; intention-healthy nutrition: P<.001). However, the positive effects of these 2 variables were smaller than the negative effect of patient activation. The complete model used in the multiple regression analysis had high statistical significance (P<.001; Table 2). This indicates that the statistically significant effects found for 3 of the 15 variables cannot simply be explained by a higher chance of α errors due to multiple testing. Nagelkerke pseudo R^2 for the complete model was 0.18.

Frequency

If the correlations between the possible determinants are only computed for participants who used the web portal at least once, the pattern of correlations would be very similar to the pattern of correlations for all investigated persons. Many of the variables analyzed were moderately correlated with each other. Again, the highest correlations were among loneliness, depression, and anxiety. However, the correlation between the intentions regarding physical activity and healthy nutrition decreased (Multimedia Appendix 5). Notwithstanding the moderate correlations, the largest inflation factor was smaller than 3, similar to the finding in the analyses for uptake (Table 3). Accordingly, all possible determinants could be included in the multiple regression.

XSL•FO
RenderX

Table . Simple and multiple regression analyses for predictors of frequency[a].

Variable	Simple regression analysis[b]				Multiple regression analysis[b]		
	B[c] (95% CI)	*P* value	DP[d]	PEV[e]	B[c] (95% CI)	*P* value	VIF[f]
Sociodemographic variables							
Age in years	0.03 (0.01 to 0.05)	.005[g]	1.66	0.024	0.02 (0.00 to 0.04)	.03[g]	1.19
Female gender	0.12 (−0.25 to 0.49)	.53	1.72	0.002	0.25 (−0.09 to 0.59)	.15	1.25
Higher education	0.01 (−0.36 to 0.39)	.94	1.73	0.000	−0.02 (−0.36 to 0.33)	.93	1.15
Psychological variables[h]							
General psychological variables							
Extraversion	−0.41 (−1.16 to 0.33)	.28	1.71	0.004	−0.63 (−1.37 to 0.12)	.10	1.43
Agreeableness	−0.42 (−1.35 to 0.51)	.37	1.72	0.004	−0.65 (−1.43 to 0.14)	.11	1.21
Openness	−0.14 (−0.88 to 0.60)	.72	1.72	0.000	0.04 (−0.64 to 0.72)	.91	1.15
Conscientiousness	1.09 (0.25 to 1.93)	.01[g]	1.66	0.039	1.00 (0.16 to 1.84)	.02[g]	1.34
Neuroticism	0.37 (−0.40 to 1.14)	.35	1.72	0.003	0.50 (−0.32 to 1.33)	.23	1.53
Loneliness	−0.62 (−1.26 to 0.01)	.06	1.69	0.020	−0.64 (−1.42 to 0.15)	.11	1.94
Depression	−0.89 (−1.58 to −0.19)	.01[g]	1.68	0.015	−0.74 (−1.64 to 0.17)	.11	2.58
Anxiety	−0.24 (−1.05 to 0.58)	.57	1.72	0.001	0.60 (−0.52 to 1.72)	.29	2.88
Health-related psychological variables							
Patient activation	1.25 (−0.12 to 2.62)	.07	1.69	0.018	0.52 (−0.95 to 1.98)	.49	1.57
Health literacy	0.35 (−0.78 to 1.48)	.55	1.72	0.002	0.09 (−1.08 to 1.25)	.89	1.56
Intention–physical activity	−0.19 (−1.12 to 0.75)	.69	1.72	0.001	−0.63 (−1.49 to 0.22)	.15	1.17
Intention–healthy nutrition	0.49 (−0.82 to 1.79)	.47	1.72	0.002	0.48 (−0.61 to 1.57)	.39	1.19

[a]Patients are members of the intervention group of the randomized controlled trial conducted within the P-SUP (Personalized Self-Management Support Program) project. They were enrolled in disease management programs for type 2 diabetes and/or coronary heart disease and were residing in North Rhine-Westphalia, Germany. Data were collected during the baseline assessment of the P-SUP project.

[b]Simple and multiple regression analyses performed on data after multiple imputation (n=199).

[c]B: regression coefficient B.

[d]DP: dispersion parameter.

[e]PEV: proportion of explained variance.

[f]VIF: variance inflation factor.

[g]Significant value.

[h]For better comparability, all psychological variables were scaled to a range from 0 to 1. Higher values indicate a higher expression of the construct.

The results of the regression analyses were virtually the same for all 3 intervals applied for identifying new log-ins. Therefore, only the results for the 120-minute interval have been reported. These results differed strongly from those for uptake (Table 3). In the simple regression analyses, only 3 of the 15 variables had statistically significant effects (ie, age: *P*=.005; conscientiousness: *P*=.01; and depression: *P*=.01). The first 2 variables related positively to frequency, whereas there was a

XSL•FO
RenderX

negative effect for depression. However, after controlling for multiple testing according to Benjamini and Hochberg [60], all 3 effects disappeared. In the multiple regression analysis, there were effects for age (P=.03) and conscientiousness (P=.02), whereas the effect for depression was no longer observed (P=.11; Table 3). However, in contrast to the complete multiple regression model for predicting uptake, the complete multiple regression model for predicting frequency was not statistically significant (P=.07). Accordingly, whether age and conscientiousness actually affect frequency is questionable. The dispersion parameter for the complete model was 1.48, and the proportion of explained variance was 0.16.

Comparison Between Uptake and Frequency

The comparison of the multiple regression models for uptake and frequency showed that the rank order of the P values differed markedly between uptake and frequency. The correlation between the rank orders was −0.50, with a statistically significant deviation from 0 (P=.009).

Discussion

Principal Findings

The main finding of this study is that uptake and frequency of use are influenced in different ways by the variables considered as potential determinants. In simple regression analyses, uptake of the web portal was positively associated with higher education, openness, health literacy, and intentions regarding physical activity and healthy nutrition; however, the effect for health literacy did not remain significant after controlling for multiple testing. In the multiple regression model, uptake was primarily explained by strong intentions regarding physical activity and healthy nutrition, alongside a significant negative association with patient activation. In contrast, frequency of use showed a different pattern. In simple regression analyses, frequency was positively associated with age and conscientiousness and negatively associated with depression; however, in the multiple regression model, only age and conscientiousness remained associated. After controlling for inflation of the α error, none of the associations for frequency of use remained statistically significant, and the multiple regression model for frequency was not significant.

The observation that the multivariate regression model has a strong effect for uptake but no statistically significant effect for frequency may be partly attributable to the smaller sample size available for the frequency analyses. As higher values of Nagelkerke pseudo R^2 are more difficult to obtain than higher values of proportions of explained variance for continuous variables [62], the comparison of these 2 statistics suggests that the set of variables investigated in this study has a stronger impact on uptake than on frequency of use. A further difference between the 2 analyses concerns the rank order of the P values for the 2 multiple regressions, which deviated markedly between uptake and frequency. This strongly indicates that the considered variables affect uptake and frequency in very different ways. This demonstrates that both components of usage are distinct and should be investigated and interpreted separately from each other.

Uptake

In this study, intentions had a very strong positive association with uptake. This finding is supported by research on behavior change, which emphasizes the central role of intention in bridging the gap between attitudes and actual behavior [63]. According to the Theory of Planned Behavior [64], intention is viewed as the proximal determinant of behavior. According to this, people are more likely to perform a behavior if they strongly intend to do so. In a previous study that most closely resembles this study, Voncken-Brewster et al [26] examined a web-based self-management intervention for individuals with or at high risk for chronic obstructive pulmonary disease (COPD). In a common multiple regression model, they considered age, gender, education, and intention to change behavior alongside clinical factors such as COPD status, smoking status, physical activity level, and dyspnea status and found no statistically significant effects. In contrast, this study identified a strong positive association between intention and uptake. The difference between the 2 results might be explained by the fact that both studies refer to different kinds of intentions. In the study by Voncken-Brewster et al [26], intention referred to a change in a specific behavior, while in this study, intention referred to the performance of a specific behavior. The latter seems to be a strong motivating factor for uptake, whereas the former does not show the same effect.

In this study, patient activation showed a strong negative association with uptake. This contrasts with previous findings, where either a positive association [37] or no effect was reported [36]. A central assumption is that higher patient activation is associated with greater engagement in health-promoting behaviors [65]. However, DHIs are not necessarily health-promoting, and the DHI investigated in this study may provide little additional value for individuals who already engage in the targeted behaviors. Highly activated individuals may therefore be less inclined to use such interventions. In contrast, less activated individuals in this study may have been more likely to perceive the DHI as beneficial, given that individuals with lower activation levels are more likely to experience poorer clinical outcomes and higher rates of hospitalization [65], and thus have greater potential to benefit from additional self-management support. When initial use does not require immediate behavioral change, such tools may represent an accessible entry point into health-related activation. This suggests that the relationship between patient activation and DHI uptake may depend on perceived relevance and expected benefit. Differences in the added value of the DHIs investigated across studies may therefore explain the divergent findings. Whether this mechanism accounts for the negative association observed in this study remains a question for future research.

In this study, the effect of openness was only seen in the simple analysis and not in the multiple regression analysis. This finding might be unexpected, as individuals with high openness are generally more willing to explore unfamiliar tools and information sources [66]. Moreover, in a previous study by Su et al [29], where a smartphone-based app for improving self-management in people with T2DM was examined, openness had a statistically significant positive effect on uptake. One

XSL•FO
RenderX

possible explanation for this discrepancy is that Su et al [29] did not include an intention variable in their multiple regression analysis. Given the strong and proximal role of intention in predicting behavior, intention may represent a more proximal and stronger determinant of uptake, thereby attenuating the independent effect of more distal personality traits, such as openness, when both variables are considered simultaneously. This interpretation is further supported by evidence suggesting that openness primarily influences behavior by moderating the intention-behavior relationship rather than exerting a strong independent effect. Openness has been shown to enhance the predictive strength of intention for behavior, while its direct association diminishes once intention is accounted for [67].

Taken together, the findings for uptake suggest that individuals who either have low intention to engage in health-promoting behaviors or consider their current levels of physical activity and healthy nutrition to be sufficient may not be reached by such interventions. To increase uptake among individuals with low intention with regard to the targeted behaviors, this intention may first need to be developed. This could be achieved through targeted outreach strategies, motivational inputs that emphasize the relevance and potential benefits of the intervention, or integration into routine care settings such as general practice consultations [68]. In particular, endorsement by trusted health care professionals may help legitimize the intervention and lower initial barriers to engagement [69]. Highly activated individuals might benefit from DHIs that go beyond structured guidance and motivational support by providing tools that support the optimization of routines, monitoring of progress, and maintenance of long-term health-promoting behaviors. In contrast, less activated patients may benefit from low-threshold DHIs that provide consolidated, easily accessible information and basic guidance. Such interventions may reduce initial barriers to engagement by requiring little effort or commitment at the outset and may serve as an accessible starting point for health-related activation.

Frequency

Without controlling for multiple testing, both simple and multiple regression analyses showed statistically significant positive effects of age and conscientiousness on frequency of use. After controlling for multiple testing, none of these effects remained statistically significant in any of the regression analyses. The following considerations regarding age and conscientiousness are therefore exploratory and are intended to contextualize observed tendencies only.

The tendency regarding age found in this study is consistent with findings reported by Goyal et al [24], who examined a smartphone-based app for the assessment and prevention of heart disease. In that study, a core feature of the app was the proposal of health-related activities, referred to as challenges, and the frequency of use was operationalized as the number of challenges completed within the first 30 days of app use. An explanation of this effect or tendency might be that health-related goals tend to increase over a person's life course [70,71] because health becomes a more central concern due to increasing functional limitations and health issues [72]. Furthermore, older individuals may particularly appreciate DHIs that bundle relevant information in an easily accessible and user-friendly platform and allow self-paced engagement [73]. The lower uptake tendency among younger individuals may reflect different usage preferences rather than a lack of interest in health topics. Younger individuals may be more likely to access health-related information through multiple digital channels and may prefer more flexible or interactive formats due to their higher digital health literacy [74]. Further studies should investigate whether different types of DHIs are preferentially used by different age groups, which could inform more age-sensitive design and implementation strategies.

There was no gender effect in this study. This contrasts with findings from previous studies showing that women are generally more likely than men to engage in health-promoting behavior [75-77] and that gender influences DHI use [24,78]. However, the direction of this gender effect was not specified in the study by Goyal et al [24], whereas Rising et al [78] reported higher DHI use among women. One explanation for the absence of a gender effect in this study is that prior studies used substantially larger sample sizes (n=69,952 and n=6789), which increased the likelihood of detecting gender differences. A further reason might be the measurement of usage. Goyal et al [24] assessed usage based on the completion of real-life health challenges, such as engaging in physical activity or preparing healthy meals, while this study focused on direct interaction with the DHI. Rising et al [78] relied on self-reported patterns of digital health use derived from questionnaire data, distinguishing users based on ownership and subjective use of smartphones, apps, and tracking devices rather than on actual interaction with a specific DHI. These differences in how usage is operationalized may lead to divergent findings regarding gender effects.

As no statistically significant effects on frequency of use were observed after controlling for multiple testing, no practical implications were derived from these findings.

Limitations

The study has 5 main limitations regarding the interpretability of the results. First, the selection of potential determinants was limited to variables available in the register and survey data. As the survey did not assess participants' access to digital devices or their level of digital health literacy [79], the results will presumably not cover all possible determinants of usage. Second, the operationalization of the frequency of use was a limitation. The participants had the option to download the exercise videos provided within the DHI. Since neither the number of downloads nor the frequency of offline usage could be tracked, actual usage may have been underestimated, potentially affecting the results. Third, the study was embedded in a complex, multicomponent intervention. The other components may have interacted with the effect of the web portal on DHI usage. Fourth, just as in all other studies addressing the relationship between person-related variables and DHI use, the variables examined might correlate with other variables that are not examined, and these variables, in turn, might be the actual determinants of DHI usage. Fifth, the investigations presented here refer to a very specific DHI and target group. All participants were enrolled in a German DMP

XSL•FO
RenderX

and recruited in a specific region in Germany, where digital infrastructure, care pathways, and culture may differ from those in other regions or countries. Thus, usage patterns and determinants of usage may not be generalizable to other DHIs, other persons, or other settings.

Conclusion

This study differs from previous studies in jointly examining the determinants of uptake and frequency of use for the same DHI within the same context and sample. It contributes to a more differentiated understanding of DHI use by demonstrating that uptake and frequency of use are influenced by distinct patterns of determinants. Uptake was primarily driven by strong behavioral intentions, whereas associations with frequency of use were less robust after adjustment for multiple testing. These findings suggest that initial adoption and sustained engagement represent related but analytically and practically distinct processes. Consequently, strategies aimed at promoting uptake may need to focus on strengthening health-related intentions, while efforts to foster continued use may require different, potentially more context-sensitive approaches. Applying the approach of this study to other DHIs and settings may support more targeted, effective, and equitable digital health development and implementation.

Acknowledgments

We would like to thank August-Wilhelm Bödecker, Cosima Bootz, Suzan Botzenhardt, Iris Dapper, Katja Dehnen, Anja Engels, Christian Funke, Ekaterini Giagkou, Anette Graute, Larisa Pilic, Jasmin Marikar, Mareike Michely, Inés van der Arend, and Birgitta Weltermann for their cooperation in recruiting study participants; Matthias Funken and Michael Frank for coordinating work; and Pete Bereza for his linguistic advice.

The authors declare the use of generative artificial intelligence (AI) in the research and writing process. According to the GAIDeT taxonomy (2025) [80], the following tasks were delegated to generative AI tools under full human supervision: proofreading, editing, and translation. The generative AI tool used was ChatGPT 5.2 (OpenAI). Responsibility for the final manuscript lies entirely with the authors. Generative AI tools are not listed as authors and do not bear responsibility for the final outcomes. The declaration has been submitted by the first author (MS).

Funding

The study was funded by the Innovation Fund of the Federal Joint Committee (funding number: 01NVF18033).

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the publication agreement between the project consortium partners but are available from the corresponding author upon reasonable request.

Authors' Contributions

MS contributed to developing the investigated digital health intervention (DHI), developed essential parts of the conception, wrote essential parts of the manuscript, performed some of the statistical analyses, and did the literature research. UK contributed essentially to developing the conception and to writing the manuscript, advised MS in the statistical analyses, and performed some of the statistical analyses. CL consolidated the different components of the DHI and integrated them into the final DHI. SA advised on the statistical analyses. MR advised on the conception. KB-F advised on the Patient Activation Measure. BB, CC, AG, LG, and LS contributed to developing the DHI. FF, MH, LK, LP, FV, and SW contributed to recruiting the study participants.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Web portal design and content.
[DOC File, 3458 KB - jmir_v28i1e80895_app1.doc]

Multimedia Appendix 2
Nutritional behavior questionnaire
[DOCX File, 15 KB - jmir_v28i1e80895_app2.docx]

Multimedia Appendix 3
Rasch scaling.
[DOC File, 49 KB - jmir_v28i1e80895_app3.doc]

Multimedia Appendix 4

XSL•FO
RenderX

Correlation analysis for uptake.
[DOC File, 112 KB - jmir_v28i1e80895_app4.doc]

Multimedia Appendix 5
Correlation analysis for frequency.
[DOC File, 112 KB - jmir_v28i1e80895_app5.doc]

Checklist 1
CONSORT checklist
[PDF File, 141 KB - jmir_v28i1e80895_app6.pdf]

References

1. Hajat C, Stein E. The global burden of multiple chronic conditions: a narrative review. Prev Med Rep 2018 Dec;12:284-293. [doi: 10.1016/j.pmedr.2018.10.008] [Medline: 30406006]
2. Joseph P, Leong D, McKee M, et al. Reducing the global burden of cardiovascular disease, part 1: the epidemiology and risk factors. Circ Res 2017 Sep 1;121(6):677-694. [doi: 10.1161/CIRCRESAHA.117.308903] [Medline: 28860318]
3. Bekele BB, Negash S, Bogale B, et al. Effect of diabetes self-management education (DSME) on glycated hemoglobin (HbA1c) level among patients with T2DM: systematic review and meta-analysis of randomized controlled trials. Diabetes Metab Syndr 2021;15(1):177-185. [doi: 10.1016/j.dsx.2020.12.030] [Medline: 33360516]
4. Reynolds R, Dennis S, Hasan I, et al. A systematic review of chronic disease management interventions in primary care. BMC Fam Pract 2018 Jan 9;19(1):11. [doi: 10.1186/s12875-017-0692-3] [Medline: 29316889]
5. Scheckel B, Schmidt K, Stock S, Redaèlli M. Patient portals as facilitators of engagement in patients with diabetes and chronic heart disease: scoping review of usage and usability. J Med Internet Res 2023 Aug 25;25:e38447. [doi: 10.2196/38447] [Medline: 37624629]
6. Cheah KJ, Abdul Manaf Z, Fitri Mat Ludin A, Razalli NH, Mohd Mokhtar N, Md Ali SH. Mobile apps for common noncommunicable disease management: systematic search in app stores and evaluation using the mobile app rating scale. JMIR Mhealth Uhealth 2024 Mar 12;12(1):e49055. [doi: 10.2196/49055] [Medline: 38532298]
7. Lavoie A, Dubé V. Web-based interventions to promote healthy lifestyles for older adults: scoping review. Interact J Med Res 2022 Aug 23;11(2):e37315. [doi: 10.2196/37315] [Medline: 35998024]
8. Gentili A, Failla G, Melnyk A, et al. The cost-effectiveness of digital health interventions: a systematic review of the literature. Front Public Health 2022;10:787135. [doi: 10.3389/fpubh.2022.787135] [Medline: 36033812]
9. Kvedar J, Coye MJ, Everett W. Connected health: a review of technologies and strategies to improve patient care with telemedicine and telehealth. Health Aff (Millwood) 2014 Feb;33(2):194-199. [doi: 10.1377/hlthaff.2013.0992] [Medline: 24493760]
10. Rawstorn JC, Gant N, Direito A, Beckmann C, Maddison R. Telehealth exercise-based cardiac rehabilitation: a systematic review and meta-analysis. Heart 2016 Aug 1;102(15):1183-1192. [doi: 10.1136/heartjnl-2015-308966] [Medline: 26936337]
11. Mistry N, Keepanasseril A, Wilczynski NL, et al. Technology-mediated interventions for enhancing medication adherence. J Am Med Inform Assoc 2015 Apr;22(e1):e177-e193. [doi: 10.1093/jamia/ocu047] [Medline: 25726568]
12. Captieux M, Pearce G, Parke HL, et al. Supported self-management for people with type 2 diabetes: a meta-review of quantitative systematic reviews. BMJ Open 2018 Dec 14;8(12):e024262. [doi: 10.1136/bmjopen-2018-024262] [Medline: 30552277]
13. Greenwood DA, Gee PM, Fatkin KJ, Peeples M. A systematic review of reviews evaluating technology-enabled diabetes self-management education and support. J Diabetes Sci Technol 2017 Sep;11(5):1015-1027. [doi: 10.1177/1932296817713506] [Medline: 28560898]
14. Palacios J, Lee GA, Duaso M, et al. Internet-delivered self-management support for improving coronary heart disease and self-management-related outcomes: a systematic review. J Cardiovasc Nurs 2017;32(4):E9-E23. [doi: 10.1097/JCN.0000000000000392] [Medline: 28107251]
15. Su D, Zhou J, Kelley MS, et al. Does telemedicine improve treatment outcomes for diabetes? A meta-analysis of results from 55 randomized controlled trials. Diabetes Res Clin Pract 2016 Jun;116:136-148. [doi: 10.1016/j.diabres.2016.04.019] [Medline: 27321329]
16. Flores Mateo G, Granado-Font E, Ferré-Grau C, Montaña-Carreras X. Mobile phone apps to promote weight loss and increase physical activity: a systematic review and meta-analysis. J Med Internet Res 2015 Nov 10;17(11):e253. [doi: 10.2196/jmir.4836] [Medline: 26554314]
17. Middelweerd A, Mollee JS, van der Wal CN, Brug J, Te Velde SJ. Apps to promote physical activity among adults: a review and content analysis. Int J Behav Nutr Phys Act 2014 Jul 25;11:97. [doi: 10.1186/s12966-014-0097-9] [Medline: 25059981]

XSL•FO
RenderX

18. Hou C, Carter B, Hewitt J, Francisa T, Mayor S. Do mobile phone applications improve glycemic control (HbA1c) in the self-management of diabetes? A systematic review, meta-analysis, and GRADE of 14 randomized trials. Diabetes Care 2016 Nov;39(11):2089-2095. [doi: 10.2337/dc16-0346] [Medline: 27926892]
19. Pal K, Eastwood SV, Michie S, et al. Computer-based diabetes self-management interventions for adults with type 2 diabetes mellitus. Cochrane Database Syst Rev 2013 Mar 28;2013(3):CD008776. [doi: 10.1002/14651858.CD008776.pub2] [Medline: 23543567]
20. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010 Feb 17;12(1):e4. [doi: 10.2196/jmir.1376] [Medline: 20164043]
21. Donkin L, Christensen H, Naismith SL, Neal B, Hickie IB, Glozier N. A systematic review of the impact of adherence on the effectiveness of e-therapies. J Med Internet Res 2011 Aug 5;13(3):e52. [doi: 10.2196/jmir.1772] [Medline: 21821503]
22. Schoeppe S, Alley S, Van Lippevelde W, et al. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review. Int J Behav Nutr Phys Act 2016 Dec 7;13(1):127. [doi: 10.1186/s12966-016-0454-y] [Medline: 27927218]
23. Böhm AK, Jensen ML, Sørensen MR, Stargardt T. Real-world evidence of user engagement with mobile health for diabetes management: longitudinal observational study. JMIR Mhealth Uhealth 2020 Nov 6;8(11):e22212. [doi: 10.2196/22212] [Medline: 32975198]
24. Goyal S, Morita PP, Picton P, Seto E, Zbib A, Cafazzo JA. Uptake of a consumer-focused mHealth application for the assessment and prevention of heart disease: the <30 days study. JMIR Mhealth Uhealth 2016 Mar 24;4(1):e32. [doi: 10.2196/mhealth.4730] [Medline: 27012937]
25. Poduval S, Ahmed S, Marston L, Hamilton F, Murray E. Crossing the digital divide in online self-management support: analysis of usage data from HeLP-Diabetes. JMIR Diabetes 2018 Dec 6;3(4):e10925. [doi: 10.2196/10925] [Medline: 30522988]
26. Voncken-Brewster V, Amoureus M, de Vries H, et al. The impact of participant characteristics on use and satisfaction of a web-based computer-tailored chronic obstructive pulmonary disease self-management intervention: a process evaluation. JMIR Form Res 2017 Aug 21;1(1):e1. [doi: 10.2196//formative.6585] [Medline: 30684399]
27. Signal V, McLeod M, Stanley J, et al. A mobile- and web-based health intervention program for diabetes and prediabetes self-management (BetaMe/Melon): process evaluation following a randomized controlled trial. J Med Internet Res 2020 Dec 1;22(12):e19150. [doi: 10.2196/19150] [Medline: 33258776]
28. Coorey G, Peiris D, Scaria A, et al. An internet-based intervention for cardiovascular disease management integrated with primary care electronic health records: mixed methods evaluation of implementation fidelity and user engagement. J Med Internet Res 2021 Apr 26;23(4):e25333. [doi: 10.2196/25333] [Medline: 33900204]
29. Su J, Dugas M, Guo X, Gao GG. Influence of personality on mHealth use in patients with diabetes: prospective pilot study. JMIR Mhealth Uhealth 2020 Aug 10;8(8):e17709. [doi: 10.2196/17709] [Medline: 32773382]
30. Ruotsalainen I, Valtanen M, Kärsämä R, et al. Influencing factors in digital health intervention uptake: the interplay of education, lifestyle, and digital literacy. In: Särestöniemi M, Keikhosrokiani P, Singh D, Harjula E, Tiulpin A, Jansson M, editors. Digital Health and Wireless Solutions NCDHWS 2024 Communications in Computer and Information Science, Vol 2084: Springer; 2024:322-332. [doi: 10.1007/978-3-031-59091-7_21]
31. Yan M, Sun W, Tan C, Wu Y, Liu Y. Analysis of factors influencing the willingness of Chinese older adults to use mHealth devices: nationwide cross-sectional survey study. J Med Internet Res 2025 Mar 4;27(1):e66804. [doi: 10.2196/66804] [Medline: 40053781]
32. Glasgow RE, Strycker LA, King DK, Toobert DJ. Understanding who benefits at each step in an internet-based diabetes self-management program: application of a recursive partitioning approach. Med Decis Making 2014 Feb;34(2):180-191. [doi: 10.1177/0272989X13498156] [Medline: 23913917]
33. Brenk-Franz K, Leonhardt LJ, Strauß B. Attachment and patient activation as predictors of the interest and use of telemedical health applications -results of an observational study in primary health care. BMC Prim Care 2022 Apr 29;23(1):96. [doi: 10.1186/s12875-022-01711-0] [Medline: 35488203]
34. Dolničar V, Petrovčič A, Škafar M, Laznik J, Prevodnik K, Hvalič-Touzery S. Determinants of the intention to use mHealth in the future: evidence from an intervention study of patients with chronic diseases in Slovenia. Int J Med Inform 2024 Oct;190:105537. [doi: 10.1016/j.ijmedinf.2024.105537] [Medline: 39002206]
35. Ochnik D, Cholewa-Wiktor M, Jakubiak M, Pataj M. eHealth tools use and mental health: a cross-sectional network analysis in a representative sample. Sci Rep 2024 Mar 2;14(1):5173. [doi: 10.1038/s41598-024-55910-z] [Medline: 38431653]
36. Ancker JS, Osorio SN, Cheriff A, Cole CL, Silver M, Kaushal R. Patient activation and use of an electronic patient portal. Inform Health Soc Care 2015;40(3):254-266. [doi: 10.3109/17538157.2014.908200] [Medline: 24786648]
37. Hibbard JH, Greene J. Who are we reaching through the patient portal: engaging the already engaged? Int J Pers Cent Med 2011;1(4):788-793 [FREE Full text]
38. Short CE, DeSmet A, Woods C, et al. Measuring engagement in eHealth and mHealth behavior change interventions: viewpoint of methodologies. J Med Internet Res 2018 Nov 16;20(11):e292. [doi: 10.2196/jmir.9397] [Medline: 30446482]

XSL•FO
RenderX

39. Engelen M, van Gaal B, Vermeulen H, Zuidema R, Bredie S, van Dulmen S. The influencing contexts and potential mechanisms behind the use of web-based self-management support interventions: realistic evaluation. JMIR Hum Factors 2022 Jul 1;9(3):e34925. [doi: 10.2196/34925] [Medline: 35776437]
40. Konerding U, Redaèlli M, Ackermann K, et al. A pragmatic randomised controlled trial referring to a Personalised Self-management SUPport Programme (P-SUP) for persons enrolled in a disease management programme for type 2 diabetes mellitus and/or for coronary heart disease. Trials 2021 Sep 27;22(1):659. [doi: 10.1186/s13063-021-05636-4] [Medline: 34579783]
41. Gawlik A, Nacak Y, Kleinert J, Konerding U, Vitinius F. Theoretical derivation of a telephone-based health coaching intervention for promoting physical activity and healthy nutrition. Int J Environ Res Public Health 2023 Jun 30;20(13):6271. [doi: 10.3390/ijerph20136271] [Medline: 37444117]
42. Busse R. Disease management programs in Germany's statutory health insurance system. Health Aff (Millwood) 2004;23(3):56-67. [doi: 10.1377/hlthaff.23.3.56] [Medline: 15160803]
43. Ryan RM, Deci EL. Overview of self-determination theory: an organismic-dialectical perspective. In: Handbook of Self-Determination Research: University of Rochester Press; 2002:3-33.
44. Mummah SA, Robinson TN, King AC, Gardner CD, Sutton S. IDEAS (integrate, design, assess, and share): a framework and toolkit of strategies for the development of more effective digital interventions to change health behavior. J Med Internet Res 2016 Dec 16;18(12):e317. [doi: 10.2196/jmir.5927] [Medline: 27986647]
45. Lendt C, Gawlik A, Chermette C, Biallas B, Kleinert J. Entwicklung einer webbasierten intervention zur aktivitätsförderung bei patient:innen mit typ-2 diabetes mellitus und koronarer herzkrankheit [Article in German]. Presented at: 31 Rehabilitationswissenschaftliches Kolloquium: Deutscher Kongress für Rehabilitationsforschung Rehabilitation: Neue Wege, neue Chancen; Mar 7-9, 2022 URL: https://fis.dshs-koeln.de/de/publications/entwicklung-einer-webbasierten-intervention-zur-aktivit%C3%A4tsf%C3%B6rderu/ [accessed 2026-03-08]
46. Reck C, Lendt C, Biallas B. User engagement and usability of a digital behavior change intervention to promote physical activity in older patients with type-2 diabetes and coronary heart disease. Presented at: 26th Annual Congress of the European College of Sport Science; Sep 8-10, 2021 URL: https://fis.dshs-koeln.de/en/publications/user-engagement-and-usability-of-a-digital-behavior-change-interv/ [accessed 2026-03-08]
47. Steitz BD, Cronin RM, Davis SE, Yan E, Jackson GP. Long-term patterns of patient portal use for pediatric patients at an academic medical center. Appl Clin Inform 2017 Aug 2;8(3):779-793. [doi: 10.4338/ACI-2017-01-RA-0005] [Medline: 28765865]
48. Halfaker A, Keyes O, Kluver D, et al. User session identification based on strong regularities in inter-activity time. arXiv. Preprint posted online on Nov 11, 2014. [doi: 10.48550/arXiv.1411.2878]
49. Rammstedt B, Kemper CJ, Klein MC, Beierlein C, Kovaleva A. Big Five Inventory (BFI-10). GESIS – Leibniz Institute for the Social Sciences - Zusammenstellung Sozialwissenschaftlicher Items Und Skalen (ZIS). 2014. URL: https://search.gesis.org/instruments_tools/zis76 [accessed 2026-03-08]
50. Russell D, Peplau LA, Cutrona CE. The revised UCLA loneliness scale: concurrent and discriminant validity evidence. J Pers Soc Psychol 1980 Sep;39(3):472-480. [doi: 10.1037//0022-3514.39.3.472] [Medline: 7431205]
51. Wongpakaran N, Wongpakaran T, Pinyopornpanish M, et al. Development and validation of a 6-item revised UCLA loneliness scale (RULS-6) using Rasch analysis. Br J Health Psychol 2020 May;25(2):233-256. [doi: 10.1111/bjhp.12404] [Medline: 31999891]
52. Löwe B, Wahl I, Rose M, et al. A 4-item measure of depression and anxiety: validation and standardization of the Patient Health Questionnaire-4 (PHQ-4) in the general population. J Affect Disord 2010 Apr;122(1-2):86-95. [doi: 10.1016/j.jad.2009.06.019] [Medline: 19616305]
53. Wallston KA, Cawthon C, McNaughton CD, Rothman RL, Osborn CY, Kripalani S. Psychometric properties of the brief health literacy screen in clinical practice. J Gen Intern Med 2014 Jan;29(1):119-126. [doi: 10.1007/s11606-013-2568-0] [Medline: 23918160]
54. Brenk-Franz K, Hibbard JH, Herrmann WJ, et al. Validation of the German version of the patient activation measure 13 (PAM13-D) in an international multicentre study of primary care patients. PLoS ONE 2013;8(9):e74786. [doi: 10.1371/journal.pone.0074786] [Medline: 24098669]
55. Hibbard JH, Mahoney ER, Stockard J, Tusler M. Development and testing of a short form of the patient activation measure. Health Serv Res 2005 Dec;40(6 Pt 1):1918-1930. [doi: 10.1111/j.1475-6773.2005.00438.x] [Medline: 16336556]
56. Brenk-Franz K, Hibbard J, Tiesler F, Gensichen J. Manual Zur Deutschen Version Des PAM 13 [Book in German]: Institut für Allgemeinmedizin, Friedrich-Schiller-Universität Jena; 2014.
57. Little RJA. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 1988 Dec;83(404):1198-1202. [doi: 10.1080/01621459.1988.10478722]
58. van Buuren S. Flexible Imputation of Missing Data, 2nd edition: Chapman and Hall/CRC; 2018. [doi: 10.1201/9780429492259]
59. Hilbe JM. Modeling Count Data: Cambridge University Press; 2014. [doi: 10.1017/cbo9781139236065]
60. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 1995 Jan 1;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]

XSL•FO
RenderX

61. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ 2025 Apr 14;389:e081123. [doi: 10.1136/bmj-2024-081123] [Medline: 40228833]
62. Smith TJ, McKenna CM. A comparison of logistic regression pseudo R2 indices. Multiple Linear Regression Viewpoints 2013;39(2):17-26 [FREE Full text]
63. Admassu W, Gorems K. Analyzing health service employees' intention to use e-health systems in southwest Ethiopia: using UTAUT-2 model. BMC Health Serv Res 2024 Sep 27;24(1):1136. [doi: 10.1186/s12913-024-11567-y] [Medline: 39334209]
64. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
65. Janamian T, Greco M, Cosgriff D, Baker L, Dawda P. Activating people to partner in health and self-care: use of the Patient Activation Measure. Med J Aust 2022 Jun 6;216 Suppl 10(Suppl 10):S5-S8. [doi: 10.5694/mja2.51535] [Medline: 35665937]
66. Zhu Y, Jiang H, Zhou Z. Information adoption behavior in online healthcare communities from the perspective of personality traits. Front Psychol 2022;13:973522. [doi: 10.3389/fpsyg.2022.973522] [Medline: 36329756]
67. Pérez-Fuentes MDC, Molero Jurado MDM, Gázquez Linares JJ, Oropesa Ruiz NF, Simón Márquez MDM, Saracostti M. Self-expressive creativity in the adolescent digital domain: personality, self-esteem, and emotions. Int J Environ Res Public Health 2019 Nov 15;16(22):4527. [doi: 10.3390/ijerph16224527] [Medline: 31731778]
68. Lau WY, Lian J, Yap M. The effectiveness of interventions used to improve general health check uptake by the older adult population: a systematic review and meta-analysis. PLOS Glob Public Health 2025;5(3):e0004362. [doi: 10.1371/journal.pgph.0004362] [Medline: 40163543]
69. Catapan SDC, Sazon H, Zheng S, et al. A systematic review of consumers' and healthcare professionals' trust in digital healthcare. NPJ Digit Med 2025 Feb 21;8(1):115. [doi: 10.1038/s41746-025-01510-8] [Medline: 39984678]
70. Heckhausen J. Developmental regulation across adulthood: primary and secondary control of age-related challenges. Dev Psychol 1997;33(1):176-187. [doi: 10.1037/0012-1649.33.1.176]
71. Nurmi JE, Pulliainen H, Salmela-Aro K. Age differences in adults' control beliefs related to life goals and concerns. Psychol Aging 1992 Jun;7(2):194-196. [doi: 10.1037//0882-7974.7.2.194] [Medline: 1610507]
72. Baltes PB, Baltes MM. Psychological perspectives on successful aging: the model of selective optimization with compensation. In: Baltes PB, Baltes MM, editors. Successful Aging: Perspectives from the Behavioral Sciences: Cambridge University Press; 1990:1-34. [doi: 10.1017/CBO9780511665684.003]
73. Yin R, Martinengo L, Smith HE, Subramaniam M, Griva K, Tudor Car L. The views and experiences of older adults regarding digital mental health interventions: a systematic review of qualitative studies. Lancet Healthy Longev 2024 Nov;5(11):100638. [doi: 10.1016/j.lanhl.2024.08.007] [Medline: 39491881]
74. Estrela M, Semedo G, Roque F, Ferreira PL, Herdeiro MT. Sociodemographic determinants of digital health literacy: a systematic review and meta-analysis. Int J Med Inform 2023 Sep;177:105124. [doi: 10.1016/j.ijmedinf.2023.105124] [Medline: 37329766]
75. Baker P. Who self - cares wins: an updated perspective on men and self - care. Trends Urol & Men's Health 2019 May;10(3):19-22. [doi: 10.1002/tre.694]
76. Höhn A, Gampe J, Lindahl-Jacobsen R, Christensen K, Oksuyzan A. Do men avoid seeking medical advice? A register-based analysis of gender-specific changes in primary healthcare use after first hospitalisation at ages 60+ in Denmark. J Epidemiol Community Health 2020 Jun;74(7):573-579. [doi: 10.1136/jech-2019-213435] [Medline: 32303595]
77. Seidler ZE, Dawes AJ, Rice SM, Oliffe JL, Dhillon HM. The role of masculinity in men's help-seeking for depression: a systematic review. Clin Psychol Rev 2016 Nov;49:106-118. [doi: 10.1016/j.cpr.2016.09.002] [Medline: 27664823]
78. Rising CJ, Jensen RE, Moser RP, Oh A. Characterizing the US population by patterns of mobile health use for health and behavioral tracking: analysis of the National Cancer Institute's Health Information National Trends Survey data. J Med Internet Res 2020 May 14;22(5):e16299. [doi: 10.2196/16299] [Medline: 32406865]
79. Kayser L, Karnoe A, Furstrand D, et al. A multidimensional tool based on the eHealth literacy framework: development and initial validity testing of the eHealth Literacy Questionnaire (eHLQ). J Med Internet Res 2018 Feb 12;20(2):e36. [doi: 10.2196/jmir.8371] [Medline: 29434011]
80. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (Generative AI Delegation Taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res 2026 Apr;33(3):2544331. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]

Abbreviations

BFI-10: Big Five Inventory-10 items
BHLS: Brief Health Literacy Screen
CHD: coronary heart disease
CONSORT: Consolidated Standards of Reporting Trials
COPD: chronic obstructive pulmonary disease
DHI: digital health intervention

XSL•FO
RenderX

DMP: disease management program
IDEAS: Integrate, Design, Assess, and Share
MCAR: missing completely at random
P-SUP: Personalized Self-management Support Program
PAM13-D: Patient Activation Measure-13
PHQ-4: Patient Health Questionnaire-4 items
RCT: randomized controlled trial
SDT: self-determination theory
T2DM: type 2 diabetes mellitus
UCLA: University of California Los Angeles

Edited by S Brini; submitted 18.Jul.2025; peer-reviewed by A Ojaruega, MH Rahman, O Oluwole; accepted 14.Feb.2026; published 25.Mar.2026.

Please cite as:
Scholl M, Lendt C, Appelbaum S, Biallas B, Brenk-Franz K, Chermette C, Frank F, Gawlik A, Giesen L, Heßbrügge M, Küppers L, Pilic L, Redaèlli M, Schneider L, Vitinius F, Wilm S, Konerding U
Determinants of the Uptake and Frequency of Use of a Web Portal Digital Health Intervention in Patients With Type 2 Diabetes and/or Coronary Heart Disease: Secondary Analysis of a Randomized Controlled Trial
J Med Internet Res 2026;28:e80895
URL: https://www.jmir.org/2026/1/e80895
doi:10.2196/80895

© Maximilian Scholl, Claas Lendt, Sebastian Appelbaum, Bianca Biallas, Katja Brenk-Franz, Chloé Chermette, Friederike Frank, Angeli Gawlik, Lisa Giesen, Martina Heßbrügge, Lucas Küppers, Larisa Pilic, Marcus Redaèlli, Lara Schneider, Frank Vitinius, Stefan Wilm, Uwe Konerding. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Internet-Based Problem Management Plus Intervention for Antenatal Depression: Randomized Controlled Trial

Shuanghui Zheng[1*], MSc; Shiyu Wang[1*], BSc; Xiaoyu Shui[1*], BSc; Jing Zhao[2], MSN; Ying Lin[2], BSc; Qian Huang[1], BSc; Siqi Li[1], BSc; Yang Chen[1], BSc; Zhijie Zou[1], PhD; Jin Zheng[3], MEng; Xiaoli Chen[1], PhD

[1]School of Nursing, Wuhan University, Building 2, School of Medicine, Wuhan University, 4th Fl, No. 115, Donghu Road, Wuhan, China

[2]Department of Nursing, Maternal and Child Health Hospital of Hubei Province, Wuhan, China

[3]Qianjin 4th Road Campus, Hospital of Stomatology, Wuhan University, Wuhan, China

[*]these authors contributed equally

Corresponding Author:
Zhijie Zou, PhD
School of Nursing, Wuhan University, Building 2, School of Medicine, Wuhan University, 4th Fl, No. 115, Donghu Road, Wuhan, China

Abstract

Background: The identification and management of depression during pregnancy is an important public health issue. Although many existing psychological intervention programs are effective, their implementation is plagued by issues, such as insufficient professional resources and lengthy intervention cycles. Studies have suggested that internet-based problem management plus (IPM+) can effectively address the aforementioned challenges in the management of general depression. However, its application in the pregnant population remains to be verified.

Objective: The objective of our study was to explore the effect of IPM+ on depressive symptoms in pregnant women.

Methods: From April to October 2024, 80 pregnant women were recruited from the Hubei General Hospital and randomly assigned to the intervention group (n=40) and the control group (n=40). The control group received routine care, while the intervention group received a 5-week IPM+ intervention (1.5 hours per week) in addition to routine care. The primary outcome measure is depressive mood, and the secondary outcome measures include anxiety symptoms, perceived stress, and sleep quality. The data were collected using the Edinburgh Postnatal Depression Scale, Pregnancy Anxiety Questionnaire, 4-item Perceived Stress Scale, and Insomnia Severity Index at baseline (T1), immediately after the intervention (T2), and at 3 months (T3).

Results: Compared with the control group, the intervention group showed statistically greater improvements in depressive symptoms (between-group effect size: Hedges g=0.74, 95% CI 0.23-1.24 at T2; Hedges g=0.76, 95% CI 0.25-1.28 at T3), anxiety symptoms (between-group effect size: Hedges g=0.51, 95% CI 0.03-0.98 at T2; Hedges g=0.52, 95% CI 0.04-0.99 at T3), and perceived stress (between-group effect size: Hedges g=0.71, 95% CI 0.23-1.18 at T2; Hedges g=0.62, 95% CI 0.15-1.11 at T3) at both T2 and T3. In terms of sleep quality, the intervention group demonstrated statistically greater improvement compared with the control group only at T3 (between-group effect size: Hedges g=0.68, 95% CI 0.19-1.17).

Conclusions: IPM+ effectively alleviates depression and anxiety symptoms, reduces perceived stress, and improves sleep quality in pregnant women with antenatal depression. The intervention effects are sustained for up to 3 months postintervention.

Trial Registration: Chictr.org.cn ChiCTR2400083324; https://www.chictr.org.cn/showprojEN.html?proj=222789

(J Med Internet Res 2026;28:e81998) doi:10.2196/81998

KEYWORDS

antenatal depression; problem management plus; PM+; randomized controlled trial; digital intervention; mental health

Introduction

Antenatal depression (AND) refers to depressive episodes occurring during pregnancy (from conception to before delivery) and is considered a clinical subtype of perinatal depression [1]. The global prevalence of AND is reported to be approximately 20.7%, with significant regional variations: the prevalence in low- and middle-income countries (30.3%) is substantially higher than in high-income countries (18.1%) [2]. The prevalence of prenatal depression among women in China (19.7%) has exceeded the average level in high-income countries [3]. If left untreated, it can result in a dual crisis for maternal and infant health. Mothers may experience obstetric complications, such as hyperemesis gravidarum, preeclampsia, miscarriage, preterm birth, and postpartum depression [4]. In severe cases, suicidal ideation or infanticide may occur [5]. For the offspring, long-term negative effects, including cognitive developmental delays and emotional regulation disorders, are

XSL•FO
RenderX

possible, and there is a risk of intergenerational transmission of mental health problems [6-9].

Psychotherapy has now become the first-line treatment for AND [10], including evidence-based methods, such as cognitive behavioral therapy (CBT), interpersonal psychotherapy (IPT), and mindfulness-based interventions [11]. Existing evidence suggests that these therapies significantly reduce depressive symptoms and concurrently improve anxiety, sleep quality, and social support levels [12-15]. However, traditional psychotherapy is confronted with 3 core dilemmas: one lies in the barrier of human resources, as its implementation highly relies on highly qualified professionals, involving lengthy training periods and exorbitant labor costs; another manifests as the intervention is excessively time-consuming for participants, which is hardly compatible with the fragmented and mobile schedules of perinatal women, directly resulting in a high dropout rate [16]; and a deeper obstacle stems from sociocultural factors, where the prevalence of stigma leads to only 28.5% of patients with postpartum depression actively seeking professional support [17].

Internet-based problem management plus (IPM+) may become a breakthrough strategy to improve the accessibility of psychological service. Problem management plus (PM+), developed by the World Health Organization (WHO) [18], is a low-intensity intervention that integrates CBT and problem-solving therapy. It not only features transdiagnostic applicability, short duration, and implementability by nonprofessionals (such as community nurses and volunteers) but also consists of 4 modules: stress management, problem-solving–oriented action planning, continuous behavioral activation, and interpersonal relationship enhancement [19]. Several randomized controlled trials have demonstrated that PM+ has significant effects on depression, posttraumatic stress disorder, and anxiety in the general population, with effects lasting up to 3 months postintervention [20-23]. At the same time, compared to traditional offline methods, online psychological interventions offer 3 main advantages: overcoming spatial limitations by relying on various digital platforms, allowing perinatal women to receive interventions anywhere; time flexibility, with asynchronous courses (eg, prerecorded videos, artificial intelligence question and answer) that fit into fragmented schedules; and stigma reduction, as the anonymous environment encourages help-seeking behavior [24]. IPM+ not only boasts the advantages of internet interventions, such as geographical reach, time flexibility, and reduction of stigma, but also retains the strengths of PM+ as an integrated multimodule intervention. However, the application of PM+ in pregnant populations remains in the exploratory stage, since only 1 cultural adaptation study conducted in Pakistan worldwide has involved this [23].

To explore the intervention effect of IPM+ on AND, this study developed a randomized controlled trial, which includes a 5-week intervention period and a 3-month follow-up period. From the perspective of empirical research, it aims to answer the scientific question of “whether IPM+ can improve AND.”

Methods

Study Design

This study adopts a single-blind, 1:1 parallel-group randomized controlled trial design. The study was conducted at Hubei General Hospital, a large Grade A tertiary hospital located in Wuhan, China, and strictly followed the CONSORT (Consolidated Standards of Reporting Trials; Checklist 1) guidelines.

Ethical Considerations

The research was approved by the Ethics Committee of the Medical School of Wuhan University (WHU-LFMD-IRB2024006) and registered with the Chinese Clinical Trial Registry (ChiCTR2400083324). All participants provided written informed consent before completing the baseline assessment. To ensure participant privacy and confidentiality, all collected data were deidentified and stored on a secure, password-protected server, with access strictly restricted to the research team. Regarding compensation, no financial or material incentives were provided to the participants for completing the intervention or the assessment sessions.

Participants

Participants were recruited through convenience sampling from the Obstetrics Department of Hubei General Hospital between April 2024 and October 2024. Eligible patients who met the inclusion criteria were invited to participate in the study. The inclusion criteria were as follows: (1) age 18 years and above, (2) gestational age of 28 weeks or less, (3) singleton pregnancy, (4) no plans for elective termination of pregnancy, (5) Edinburgh Postnatal Depression Scale (EPDS) score of 10 or higher (if EPDS score of 13 or higher, further assessment was required to determine suitability for enrollment), and (6) ability to understand the study questionnaire. The exclusion criteria were as follows: (1) presence of suicidal thoughts or behaviors, (2) serious organic diseases or functional disorders, (3) severe mental or cognitive disorders (eg, severe intellectual disability, dementia), (4) substance or alcohol dependence or abuse behaviors within the past 6 months, and (5) previous experience with PM+ treatment or simultaneous participation in other psychological interventions.

Blinding

Non-researchers generated a random number sequence using a random number table, which was then imported into the Excel software and arranged in ascending order. Based on the sorted random number sequence, the first 50% (n=40) of the participants were assigned to the intervention group, and the remaining 50% (n=40) were assigned to the control group. To ensure allocation concealment, the random allocation scheme was sealed in opaque envelopes, numbered sequentially. These envelopes were opened by research assistants, who were not involved in the study evaluation, after the participants completed the baseline data collection to execute the final group assignments. The study adopted a single-blind design, with all participants unaware of their group allocation. Throughout the recruitment, informed consent, and the entire study process, we consistently utilized the unified project title “Perinatal Mental

XSL•FO
RenderX

Health Online Support Project." All participants were informed that the project involved 2 different forms of online guidance programs, both of which were beneficial for pregnancy health. Furthermore, we ensured that both the control group and the intervention group included weekly one-on-one meetings with an online facilitator to maintain consistency. Due to the nature of the intervention, the implementers were not blinded.

Intervention Group

On the basis of routine care, the intervention group received a 5-week IPM+ intervention (1.5 hours per week). The intervention plan was compiled based on the "PM+" manual of the Chinese version of the WHO [25] and tailored to address the specific characteristics of perinatal depression. Key adaptations included pregnancy-specific case examples and stressors, the incorporation of safe and appropriate activities for the antenatal period, and an emphasis on social support derived from peer interactions among pregnant women. The adapted protocol was reviewed and validated by a multidisciplinary panel of experts (including a clinical psychologist, an obstetrician, a PM+ supervisor, and a senior midwife) for cultural relevance, safety, and feasibility. Furthermore, a usability pretest was conducted with 4 pregnant women (who met the inclusion criteria but did not participate in the main trial) to preliminarily assess its acceptability and comprehensibility. The intervention was conducted in an individualized online format, with structured 1.5-hour sessions every week via the WeChat platform. The intervention was delivered by 2 nonspecialist providers, both of whom were nursing graduate students. Prior to study initiation, they completed a standardized 40-hour training program based on the Chinese version of the WHO PM+ manual [25]. Competency was subsequently assessed through role-play–based simulations.

The intervention content adopted a modular design, focusing on the development of 4 core psychological abilities for pregnant and postpartum women: stress regulation, problem-solving ability, behavioral activation, and social support ability. The intervention process followed a gradual approach and was implemented in 5 stages: the first stage focused on building a good therapeutic relationship, helping participants understand their responses to adversity and master basic relaxation techniques; the second stage systematically trained problem-solving skills, using methods such as brainstorming to find feasible solutions; the third stage used behavioral experiments to verify the effectiveness of the solutions and established a regular activity pattern; the fourth stage strengthened the social support network and improved the quality of interpersonal interactions; and the final stage involved an overall review and planning for future applications, ensuring the long-term maintenance of intervention effects. After each intervention, targeted homework assignments were given, and progress was monitored using standardized assessment tools. Detailed content is presented in Multimedia Appendix 1. Detailed problem management steps are presented in Multimedia Appendix 2.

To ensure the standardization and consistency of the intervention, the research team established a comprehensive quality control system, including audio recording analysis of 10% (n=17) of the intervention sessions, biweekly supervision meetings focusing on adherence and competence, and random quality checks. Participants who exhibited severe deterioration in mental health or were identified as having a risk of self-harm or suicide during the intervention period were immediately referred to a psychiatrist for specialized care and were withdrawn from the study.

Control Group

The participants in the control group received a 5-week regimen of standard prenatal care, which was administered by obstetric nursing staff who underwent uniform training through a WeChat platform. The standard care program included weekly systematic health follow-ups, during which the nursing staff utilized standardized communication scripts to assess the physical and mental health status of the pregnant women. Additionally, professional prenatal health guidance was provided, covering basic health education services, such as prenatal examination precautions, key aspects of perinatal mental health maintenance, and newborn care knowledge.

Outcomes

The primary outcome measure was depressive symptoms, assessed using the EPDS [26,27]. This scale consists of 10 items, with a 4-point Likert scale for scoring, where the frequency of symptoms is rated from "never" to "always," corresponding to scores of 0-3. The Chinese version of the EPDS [28] demonstrates good internal reliability (Cronbach α=0.89). The clinical diagnostic criteria are as follows: a total score of less than 9 indicates a normal state, 9 - 12 points suggest possible depression, and a total score of 13 or higher or a score greater than 0 on item 10 (self-harm thoughts) indicates a depressive state.

The following secondary outcome measures were also assessed: pregnancy-related anxiety: Pregnancy Anxiety Questionnaire (PAQ) [29]; perceived stress via the 4-item Perceived Stress Scale (PSS) [30,31]; and sleep quality via the Insomnia Severity Index (ISI) [32].

Additionally, baseline general information was collected, including age, gestational age, last menstrual period, height, current or prepregnancy weight, education level, marital status, occupation, residence, per capita family monthly income, living arrangements with elderly family members, knowledge of pregnancy-related information, and a history of prepregnancy depression. The data were collected at baseline (T1), immediately postintervention (T2), and 3 months postintervention (T3). The 3-month follow-up was selected, as it represents a standard medium-term evaluation period in behavioral intervention trials and aligns with prior WHO PM+ research, thereby facilitating cross-study comparison.

Sample Size

The study employs the sample size estimation method for comparing the means of 2 samples, with depression scores as the calculation indicator, and uses the following formula to calculate the sample size:

$$(1) n1=n2=2\times[(\mu\alpha+\mu\beta)/d]2$$

XSL•FO
RenderX

A 2-tailed test was adopted with α=.05 and β=.10. The anticipated effect size for this study was informed by the meta-analysis conducted by Ansaari et al [33] on online psychological interventions for perinatal depression, from which a Hedges *g* of 0.86 was derived for the primary outcome. Accounting for an estimated 20% loss to follow-up, the calculated sample size was approximately 37 participants per group, resulting in a total required sample of 74 participants. This attrition estimate is based on the typical range (15% - 25%) observed in comparable trials of PM+ [34].

Statistical Analysis

Descriptive statistical analysis was used for baseline data: unordered categorical variables were expressed as percentages, and the chi-square test was used for between-group comparisons. When the expected count was less than 5, the Fisher exact test was applied. Ordinal categorical variables were described as median (IQR), and the Mann-Whitney *U* test was used for between-group comparisons. Normally distributed continuous variables were presented as mean (SD), with the independent samples *t*-test used for between-group comparisons; non-normally distributed continuous variables were described as median (IQR), and the Mann-Whitney *U* test was also used for between-group comparisons.

In this study, intention-to-treat analysis was adopted for primary and secondary outcomes. Based on the data after multiple imputation, linear mixed-effects models were used for analysis. The covariates included in the model were baseline variables with between-group differences (whether depression was present before pregnancy), and the random effect was set as the individual random intercept.

Standardized mean differences at time points T2 and T3 were calculated to represent between-group effect sizes, while standardized mean differences between T2, T3, and T1 within the same group served as within-group effect sizes. Both were quantified using Hedges *g* values and 95% CI. SPSS 26.0 (IBM Corp) was used for data analysis in this study, with the significance level set at $P<.05$.

Results

Overview

A total of 515 participants were screened for eligibility assessment between April 2024 and October 2024, and finally, 80 eligible participants were included in this study, who were randomly assigned to the intervention group (n=40) or the control group (n=40). In the intervention group, 34 (85.0%) participants and 32 (80%) participants completed the assessments at T2 and T3, respectively. In the control group, 32 (80%) participants and 28 (70.0%) participants completed the assessments at T2 and T3, respectively. In accordance with the intention-to-treat principle, all 80 randomized participants (40 per group) were included in the final analyses after missing data were handled using multiple imputation. The flowchart of participant recruitment and grouping is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Flowchart of participants. ITT: intention-to-treat.

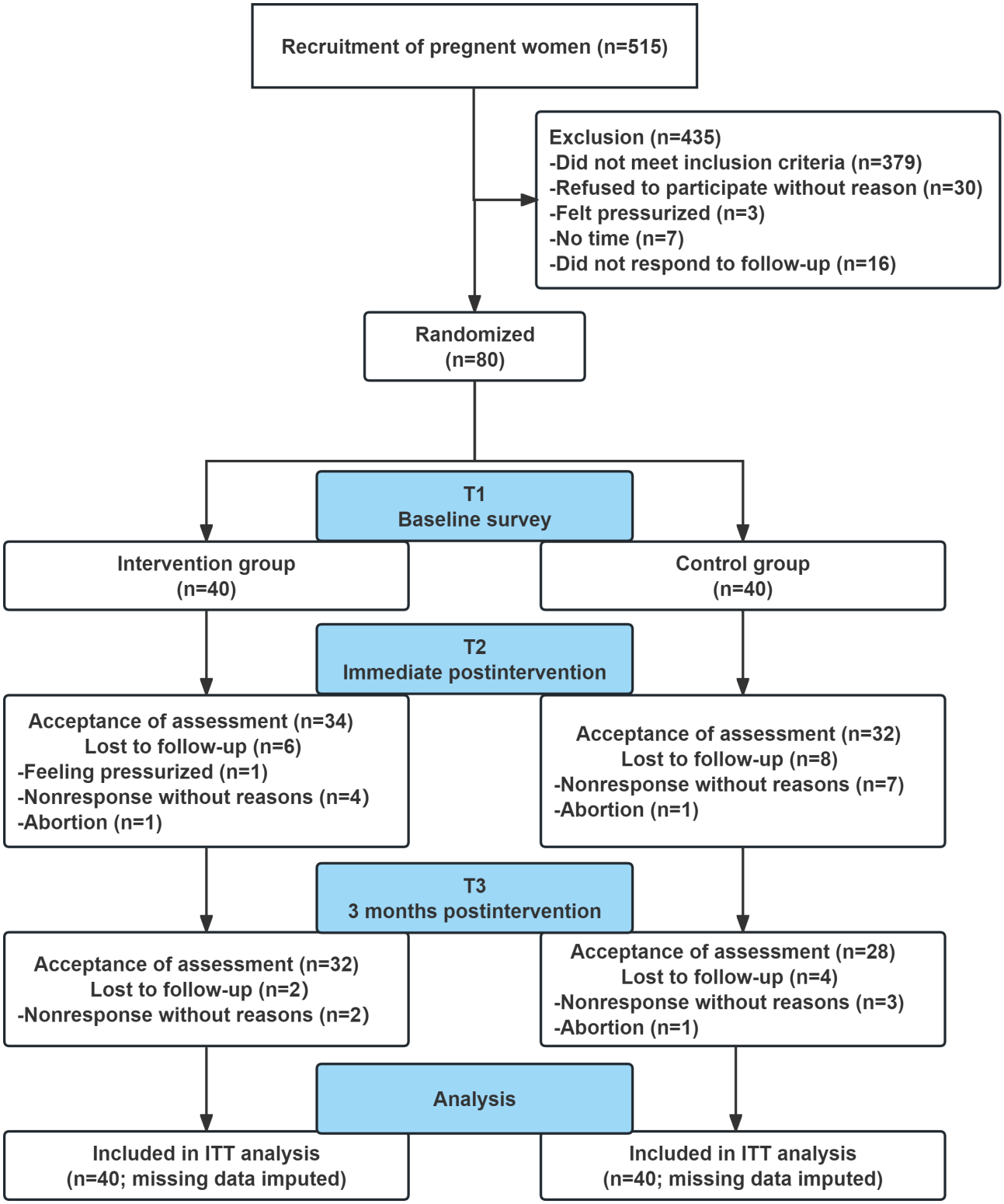

Baseline Assessment

The baseline data of both the intervention group and the control group are presented in Table 1. A comparison of the 2 groups in terms of demographic characteristics revealed a statistically significant difference in the history of depression before pregnancy (P=.005), with the intervention group showing a higher proportion of individuals who had experienced depression prior to pregnancy. Regarding the scores of the EPDS, PAQ, PSS, and ISI, the differences between the 2 groups did not reach statistical significance.

XSL•FO
RenderX

Table . Baseline characteristics of participants from the intervention and control groups (n=80).

General information indicators	Intervention group (n=40)	Control group (n=40)	*P* value
Age (years), mean (SD)	30.63 (4.78)	30.43 (4.29)	.84
Gestational age, median (IQR)	13.5 (12.0-23.7)	14.5 (11.0-22.0)	.43
Number of pregnancies, n (%)			>.99
First pregnancy	24 (60)	24 (60)	
Nonfirst pregnancy	16 (40)	16 (40)	
Number of deliveries, n (%)			.45
Primiparous	31 (77.5)	28 (70.0)	
Multiparous	9 (22.5)	12 (30.0)	
Prepregnancy BMI (kg/m²), median (IQR)	22.0 (20.0-24.0)	21.0 (20.0-25.0)	.78
Baseline BMI (kg/m²), median (IQR)	23.0 (20.0-25.0)	22.5 (12.0-27.0)	.54
Employment status, n (%)			.45
Employed	28 (70)	31 (77.5)	
Other	12 (30)	9 (22.5)	
Place of residence, n (%)			.53
Rural	5 (12.5)	7 (17.5)	
Urban or city	35 (87.5)	33 (82.5)	
Whether living with elderly family members, n (%)			.50
Yes	20 (50)	23 (57.5)	
No	20 (50)	17 (42.5)	
History of depression before pregnancy, n (%)			.005
Yes	8 (20)	0 (0)	
No	32 (80)	40 (100)	
Education level, n (%)			.05
Elementary school or below	0 (0)	0 (0)	
Middle school	1 (2.5)	5 (12.5)	
High school or vocational school	8 (20)	10 (25.0)	
Bachelor's degree	22 (55.0)	21 (52.5)	
Master's degree or higher	9 (22.5)	4 (10)	
Per capita monthly family income (yuan [US $]/month), n (%)			.52
3000-5000 [432-720]	7 (17.5)	7 (17.5)	
5001-10,000 [720-1440]	17 (42.5)	19 (47.5)	
>10,000 [>1440]	14 (35)	11 (27.5)	
Knowledge level of pregnancy, n (%)			.75
Not at all knowledgeable	3 (7.5)	3 (7.5)	
Somewhat knowledgeable	35 (87.5)	36 (90)	
Very knowledgeable	2 (5.0)	1 (2.5)	
Depressive symptoms (EPDS[a]), mean (SD)	12.32 (1.62)	11.85 (2.14)	.07

XSL•FO
RenderX

General information indicators	Intervention group (n=40)	Control group (n=40)	*P* value
Anxiety symptoms (PAQ[b]), mean (SD)	25.85 (2.92)	26.00 (4.25)	.89
Stress level (PSS[c]), mean (SD)	7.40 (1.94)	7.97 (1.34)	.47
Sleep quality (ISI[d]), mean (SD)	10.60 (5.65)	11.32 (4.89)	.66

[a]EPDS: Edinburgh Postnatal Depression Scale.
[b]PAQ: Pregnancy Anxiety Questionnaire.
[c]PSS: 4-item Perceived Stress Scale.
[d]ISI: Insomnia Severity Index.

Outcome Measures

The linear mixed model uses the total score of the EPDS, the score of the PAQ, the score of the PSS, and the score of the ISI as dependent variables, with the control group and baseline time point (T1) as reference levels. It includes group (intervention group and control group), time (T1 or T2 or T3), and group×time interaction term as fixed effects, and controls for the covariate of a history of depression before pregnancy.

Depressive Mood

Table 2 shows that the between-group effect ($F_{1,95.747}$=13.564; P<.001), time effect ($F_{2,97.600}$=116.563; P<.001), and group×time interaction effect ($F_{2,97.600}$=24.203; P<.001) of depressive mood were all statistically significant.

Table . Results of fixed effects tests for each dependent variable.[a]

Effect type		*F* test (*df*)	*P* value
EPDS[b]			
	Between-group effect	13.564 (1, 95.747)	<.001
	Time effect	116.563 (2, 97.600)	<.001
	Interaction effect (time × between-group)	24.203 (2, 97.600)	<.001
PAQ[c]			
	Between-group effect	8.388 (1, 69.106)	.005
	Time effect	30.762 (2, 100.407)	<.001
	Interaction effect (time × between-group)	7.122 (2, 100.407)	.001
PSS[d]			
	Between-group effect	20.444 (1, 75.596)	<.001
	Time effect	29.707 (2, 90.567)	<.001
	Interaction effect (time × between-group)	7.255 (2, 90.567)	.001
ISI[e]			
	Between-group effect	13.458 (1, 73.646)	<.001
	Time effect	37.581 (2, 97.521)	<.001
	Interaction effect (time × between-group)	5.566 (2, 97.521)	.005

[a]The control group, T1 (baseline), and a history of depression before pregnancy were used as the reference categories.
[b]EPDS: Edinburgh Postnatal Depression Scale.
[c]PAQ: Pregnancy Anxiety Questionnaire.
[d]PSS: 4-item Perceived Stress Scale.
[e]ISI: Insomnia Severity Index.

The fixed-effects model (Table 3) analysis showed that the baseline effect in the intervention group was not statistically significant (β=.55; P=.24). The time effects were statistically significant at both T2 (β=−2.04; P<.001, 95% CI −2.92 to −1.16) and T3 (β=−2.34; P<.001, 95% CI −3.51 to −1.18) versus T1. We observed significant time×group interactions at T2

XSL•FO
RenderX

(β=−3.26; *P*<.001, 95% CI −4.51 to −2.01) and T3 (β=−3.75; *P*<.001, 95% CI −5.34 to −2.17) versus the control group T1. No history of depression before pregnancy showed an effect coefficient of 0.33 (*P*=.68; 95% CI −1.21 to 1.87), with no statistical significance. We reanalyzed the participants without a history of depression (32 participants in the intervention group and 40 participants in the control group). The results showed that the impact of the intervention on the primary outcomes (EPDS) of T2 (β=−3.40; *P*<.001) and T3 (β=−3.76; *P*<.001) remained very significant.

Table . Fixed effect estimation results of depression.[a]

Parameter	Coefficient (β)	SE	*t* test (*df*)	*P* value	95% CI	
					Lower bound	Upper bound
Intercept	12.00	0.91	13.15 (73.2)	<.001	10.21	13.79
Intervention group	0.55	0.47	1.17 (69.6)	.24	−0.37	1.46
History of depression before pregnancy						
No	0.33	0.79	0.42 (72.2)	.68	−1.21	1.87
Time						
T2	−2.04	0.45	−4.57 (87.8)	<.001	−2.92	−1.16
T3	−2.34	0.59	−3.99 (85.2)	<.001	−3.51	−1.18
Time×group interaction						
T2	−3.26	0.64	−5.12 (87.8)	<.001	−4.51	−2.01
T3	−3.75	0.80	−4.68 (85.2)	<.001	−5.34	−2.17

[a]The intervention group was referenced to the control group. Time points T2 and T3 were referenced to time point T1. The group without a history of depression before pregnancy was referenced to the group with a history of depression before pregnancy.

The trend of mean EPDS total scores over the 3 time points in the intervention group and the control group is shown in Figure 2. The score trajectories of the 2 groups diverged immediately after the intervention (T2). The scores in the intervention group decreased sharply at T2 and remained stable at T3, whereas the scores in the control group showed only a modest decline and remained consistently higher than those in the intervention group.

XSL•FO
RenderX

Figure 2. Line chart of the mean scores of Edinburgh Postnatal Depression Scale (EPDS) total scores in the 2 groups at 3 time points.

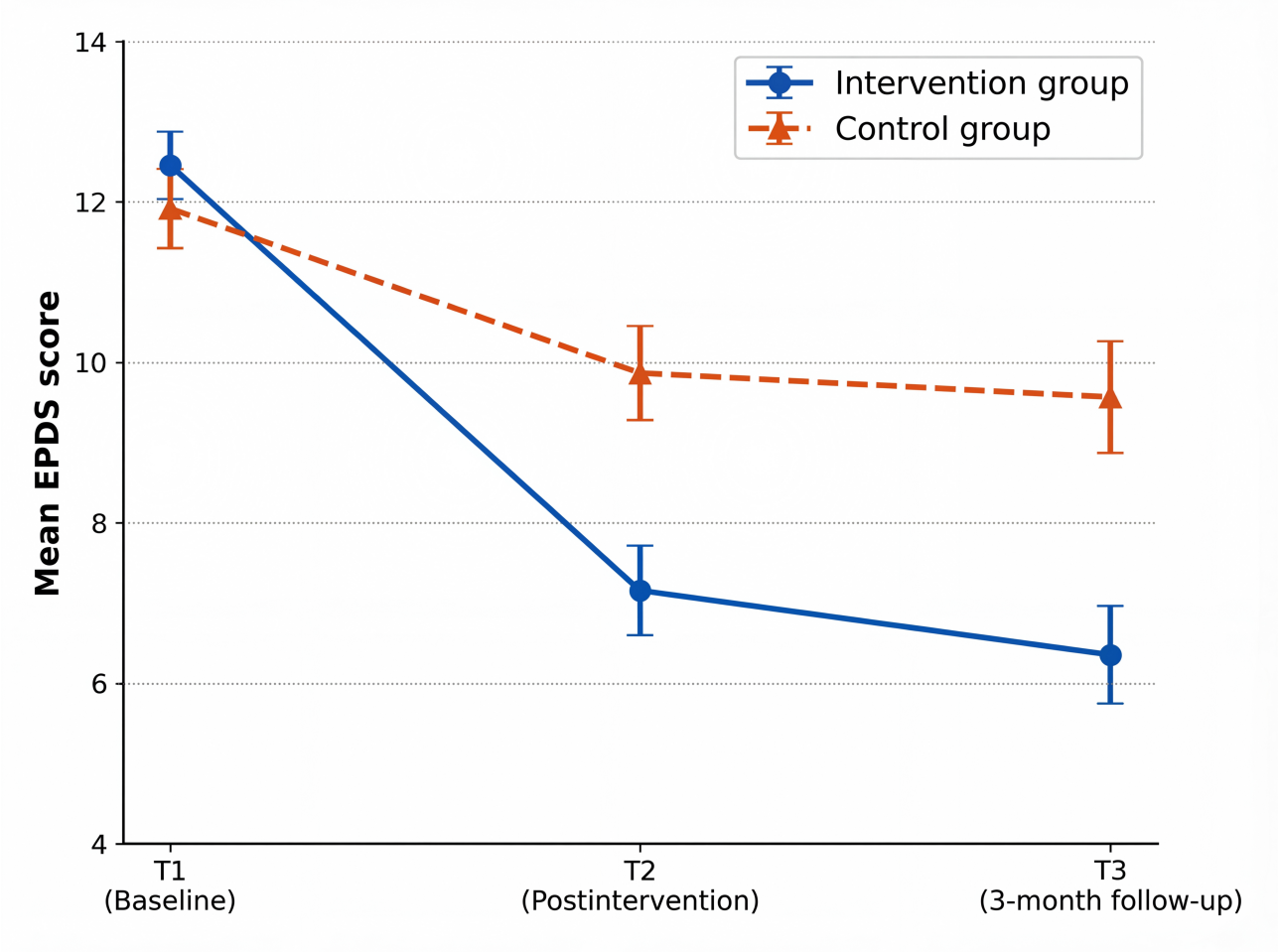

Further intra- and intergroup effect comparisons were conducted. The analysis of intragroup effect sizes (Table 4) showed that the intervention group exhibited significant reductions in depressive scores at T2 (Hedges g=1.69, 95% CI 1.18-2.20) and T3 (Hedges g=1.75, 95% CI 1.23-2.26), with large effect sizes. In contrast, the reduction in the control group was smaller (T2: Hedges g=0.58, 95% CI 0.13-1.02; T3: Hedges g=0.59, 95% CI 0.13-1.03). In intergroup comparisons, the intervention group scored significantly lower than the control group in terms of depression scores at T2 (Hedges g=0.74, 95% CI 0.23-1.24) and T3 (Hedges g=0.76, 95% CI 0.25-1.28; Table 4, Figure 2).

XSL•FO
RenderX

Table . Between-group and within-group effect sizes.[a]

	Intervention group (n=40)			Control group (n=40)			Between-group	
	Mean (SD)	Hedges *g* (95% CI)	*P* value	Mean (SD)	Hedges *g* (95% CI)	*P* value	Hedges *g* (95% CI)	*P* value
EPDS[b]								
T1	12.46 (2.67)	—[c]	—	11.92 (3.11)	—	—	—	—
T2	7.16 (3.51)	1.69 (1.18 to 2.20)	<.001	9.87 (3.71)	0.58 (0.13 to 1.02)	.003	0.74 (0.23 to 1.24)	.001
T3	6.36 (3.85)	1.75 (1.23 to 2.26)	<.001	9.57 (4.40)	0.59 (0.13 to 1.03)	<.001	0.76 (0.25 to 1.28)	<.001
PAQ[d]								
T1	24.72 (5.34)	—	—	24.89 (6.07)	—	—	—	—
T2	19.82 (5.06)	0.91 (0.44 to 1.37)	<.001	22.70 (6.13)	0.35 (0.06 to 0.62)	.008	0.51 (0.03 to 0.98)	.03
T3	20.54 (5.27)	0.77 (0.31 to 1.23)	<.001	23.56 (6.05)	0.21 (−0.03 to 0.45)	.08	0.52 (0.04 to 0.99)	.02
PSS[e]								
T1	7.65 (2.25)	—	—	8.33 (2.61)	—	—	—	—
T2	5.85 (2.51)	0.74 (0.27 to 1.22)	<.001	7.83 (2.96)	0.17 (−0.14 to 0.48)	.20	0.71 (0.23 to 1.18)	.002
T3	5.50 (2.70)	0.82 (0.34 to 1.31)	<.001	7.34 (3.01)	0.34 (−0.07 to 0.75)	.09	0.62 (0.15 to 1.11)	.005
ISI[f]								
T1	9.83 (4.36)	—	—	10.54 (4.91)	—	—	—	—
T2	5.06 (4.71)	1.01 (0.46 to 1.57)	<.001	7.26 (5.44)	0.61 (0.13 to 1.09)	.002	0.42 (−0.08 to 0.94)	.06
T3	4.88 (4.86)	1.03 (0.48 to 1.59)	<.001	8.52 (5.51)	0.37 (−0.03 to 0.79)	.079	0.68 (0.19 to 1.17)	.002

[a]Small, medium, and large effect sizes are defined as 0.2, 0.5, and 0.8, respectively. Lower scores indicate lower levels of depression. Statistical significance is determined when the 95% CI of Hedges *g* does not include "0." T1 is used as the reference.

[b]EPDS: Edinburgh Postnatal Depression Scale.

[c]Not applicable.

[d]PAQ: Pregnancy Anxiety Questionnaire.

[e]PSS: 4-item Perceived Stress Scale.

[f]ISI: Insomnia Severity Index.

Secondary Outcome

Among the secondary outcome indicators, the between-group effect ($F_{1,69.106}$=8.388; *P*=.005), time effect ($F_{2,100.407}$=30.762; *P*<.001), and interaction effect ($F_{2,100.407}$=7.122; *P*=.001) of anxiety symptoms; the between-group effect ($F_{1,75.596}$=20.444; *P*<.001), time effect ($F_{2,90.567}$=29.707; *P*<.001), and interaction effect ($F_{2,90.567}$=7.255; *P*=.001) of perceived stress; and the between-group effect ($F_{1,73.646}$=13.458; *P*<.001), time effect ($F_{2,97.521}$=37.581; *P*<.001), and interaction effect ($F_{2,97.521}$=5.566; *P*=.005) of sleep quality were all statistically significant (Table 2).

The results of fixed-effect estimates (Multimedia Appendix 3) revealed the dynamic change characteristics of secondary outcomes. For anxiety symptoms (Table S3 in Multimedia Appendix 3), the baseline levels were comparable between the 2 groups (β=−0.17, $t_{71.2}$=−0.17; *P*=.87). The time effect was significant at T2 (β=−2.19, $t_{107.5}$=−3.18; *P*<.001) and marginally significant at T3 (β=−1.33, $t_{101.0}$=−1.95; *P*=.05). Additionally, the group×time interaction effects were significant at both T2 (β=−2.72, $t_{107.5}$=−2.78; *P*=.006) and T3 (β=−2.86, $t_{101.0}$=−2.99; *P*=.003), suggesting that the intervention group had more prominent improvements in anxiety symptoms.

Regarding perceived stress (Table S4 in Multimedia Appendix 3), there was no significant difference in baseline levels between the 2 groups (β=−.68, $t_{74.8}$=−1.67; *P*=.096). The time effect was significant at T3 (β=−.98, $t_{88.4}$=−2.73; *P*=.007) but not at T2 (β=−0.49, $t_{100.7}$=−1.52; *P*=.14). The interaction effect showed that the reduction in the PSS in the intervention group during period T2 was significantly greater than that in the control group

XSL•FO
RenderX

(β=−1.30, $t_{100.7}$=−3.17; *P*=.002); the between-group difference persisted during period T3 (β=−1.16, $t_{88.4}$=−2.19; *P*=.03), but the effect intensity was slightly weaker than that at the T2 time point.

The fixed-effect estimates for sleep quality (Table S5 in Multimedia Appendix 3) indicated that the baseline levels were comparable between the 2 groups (β=−.71, $t_{54.2}$=−0.87, *P*=.38). The time effects were significant at both T2 (β=−3.27, $t_{98.3}$=−4.01; *P*<.001) and T3 (β=−2.01, $t_{118.1}$=−2.66; *P*=.008), showing a steady improvement in sleep quality over time. Moreover, the group×time interaction effect was significant only at T3 (β=−2.93, $t_{118.1}$=−2.72; *P*=.007).

Line plots (Figures 3-5) depict the trajectories of anxiety, perceived stress, and sleep quality across the 3 assessment time points. In the intervention group, anxiety decreased markedly after the intervention (T2) and showed a slight rebound at follow-up (T3) while remaining well below baseline levels, whereas the control group exhibited only minor fluctuations over time. Perceived stress in the intervention group declined substantially at T2 and was maintained at follow-up, indicating sustained effects, while stress levels in the control group remained largely stable throughout the study. Differences in sleep quality between the groups were the smallest immediately postintervention (T2); however, a clear divergence emerged at follow-up (T3), with continued improvement observed in the intervention group and a rebound trend in the control group.

Figure 3. Line chart of the mean scores of Pregnancy Anxiety Questionnaire (PAQ) total scores in the 2 groups at 3 time points.

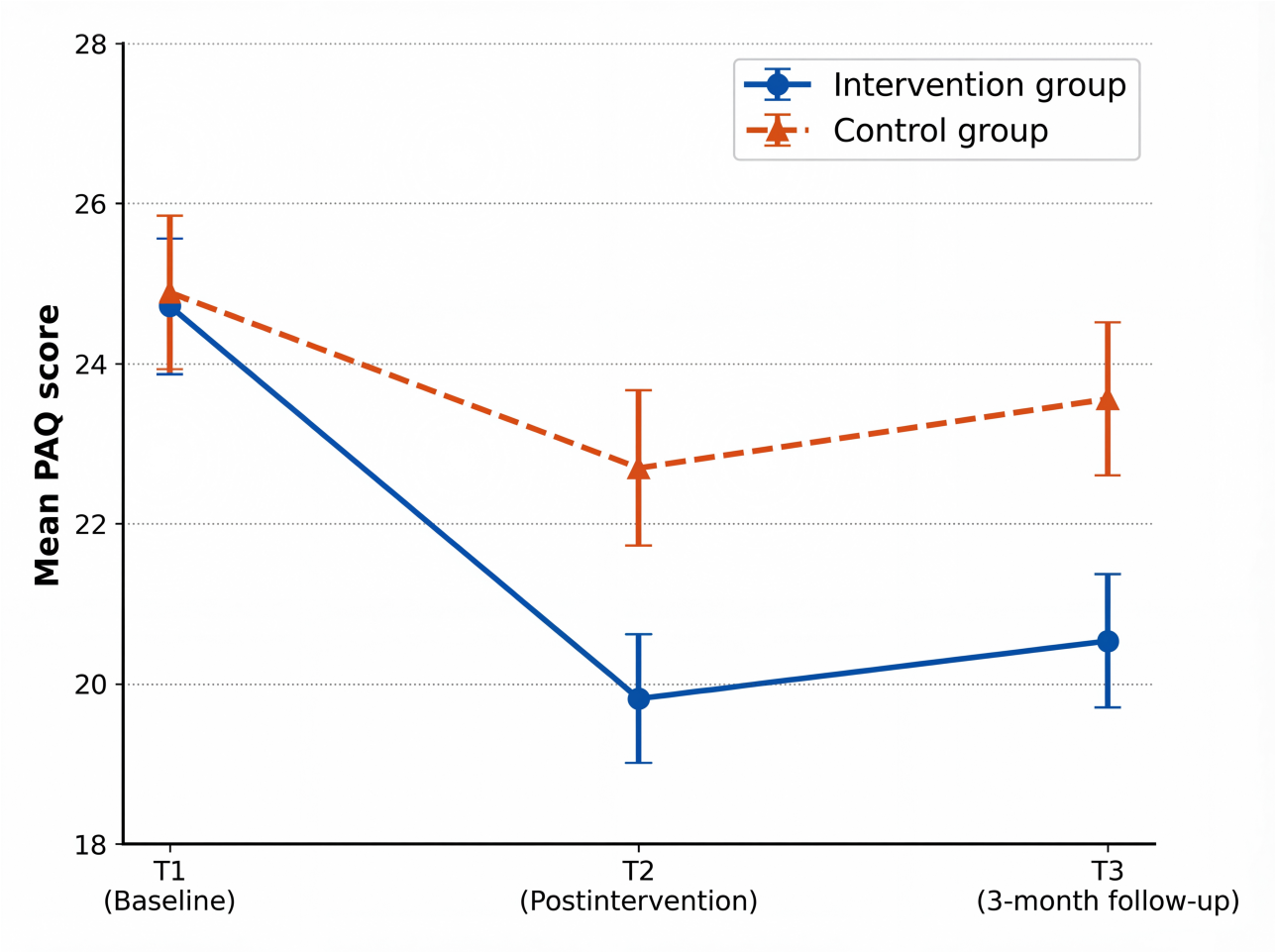

XSL•FO
RenderX

Figure 4. Line chart of the mean scores of the 4-item Perceived Stress Scale (PSS) total scores in the 2 groups at 3 time points.

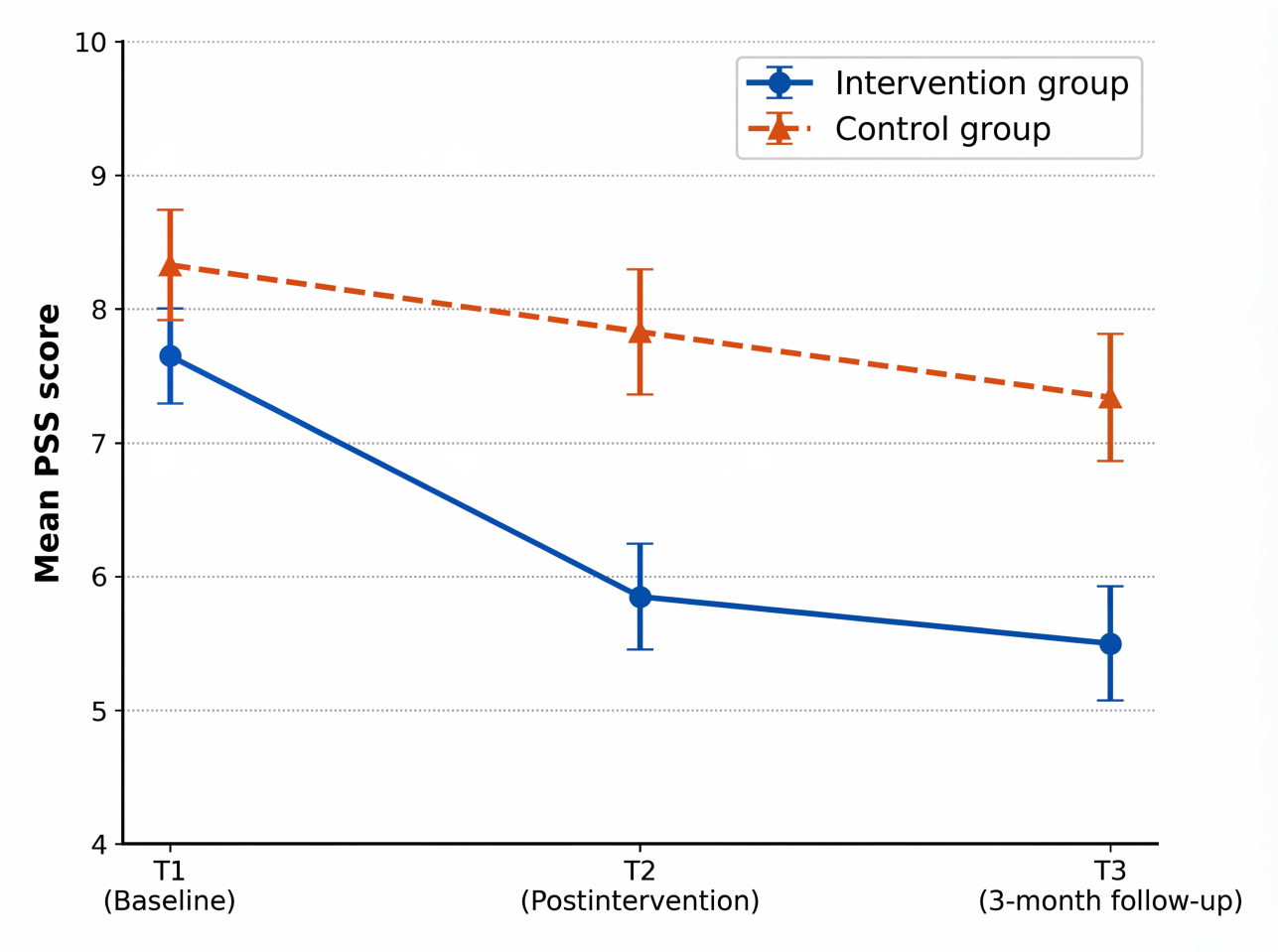

XSL•FO
RenderX

Figure 5. Line chart of the mean scores of Insomnia Severity Index (ISI) total scores in the 2 groups at 3 time points.

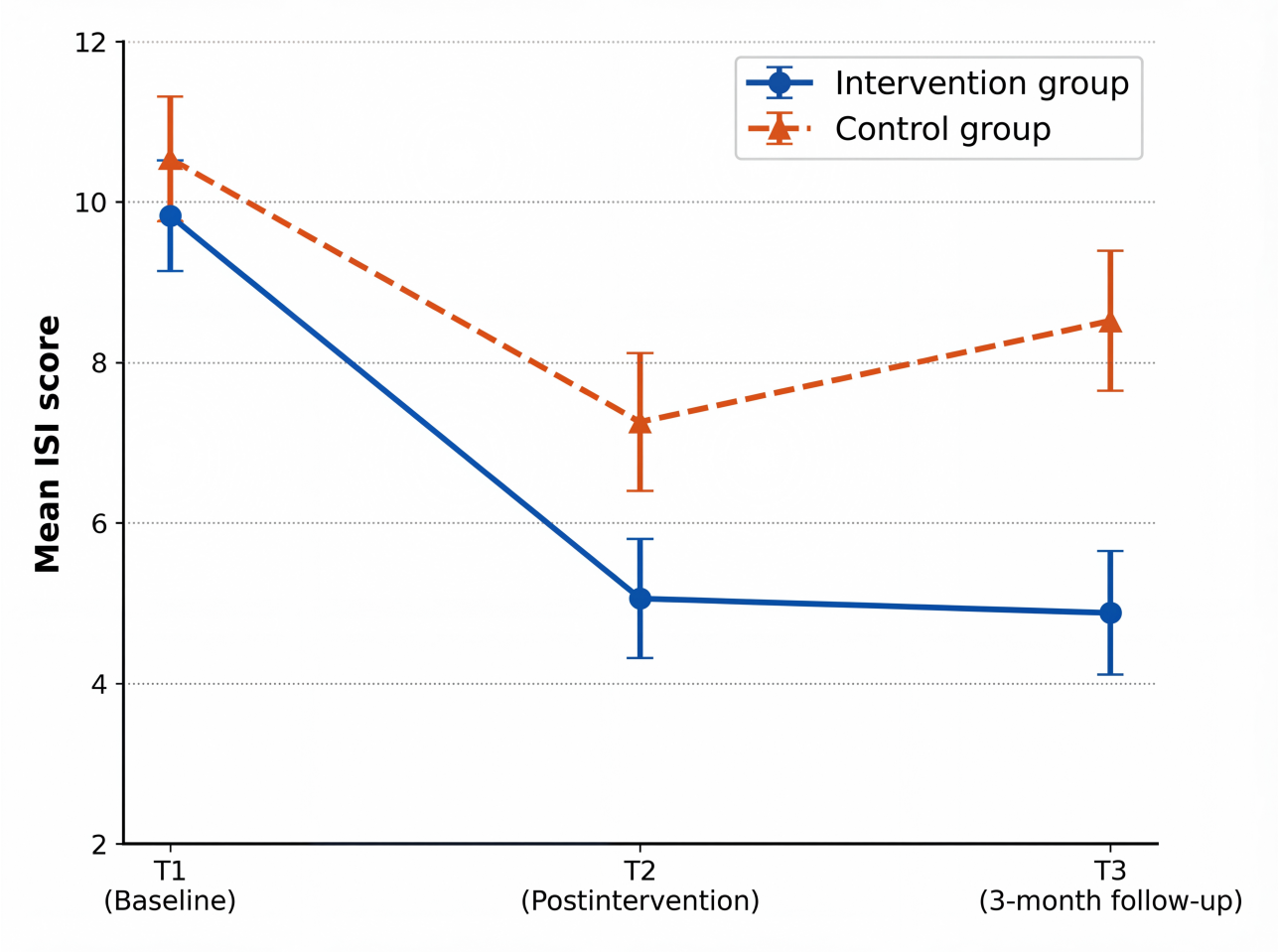

In terms of between-group effect sizes (Table 4), for anxiety symptoms, the scores of the intervention group at T2 (Hedges *g*=0.51, 95% CI 0.03-0.98) and T3 (Hedges *g*=0.52, 95% CI 0.04-0.99) were significantly lower than those of the control group, showing a medium effect. For perceived stress, the scores of the intervention group at T2 (Hedges *g*=0.71, 95% CI 0.23-1.18) and T3 (Hedges *g*=0.62, 95% CI 0.15-1.11) were significantly lower than those of the control group, presenting a medium effect. For sleep quality, the score of the intervention group at T3 (Hedges *g*=0.68, 95% CI 0.19-1.17) was significantly lower than that of the control group, with a medium effect, while the effect size at T2 was 0.42 (95% CI –0.08 to 0.94), which was not statistically significant.

Sensitivity Analysis

To assess the potential impact of missing data on the results, this study conducted a linear mixed-effects model analysis on the original un-imputed data and compared the effect sizes (Hedges *g*), direction, magnitude, confidence intervals (95% CI), and significance levels (*P*-values) before and after multiple imputation. The results showed that the missing data did not alter the significance or direction of the core outcome measures, supporting the robustness of the study's conclusions. The detailed results are presented in Table S6 in Multimedia Appendix 3.

Safety and Adverse Events

No serious adverse events related to the intervention were reported throughout the trial. In the intervention group, 1 participant withdrew at T2 due to feeling stressed by the pace of the program, which was recorded as a nonserious intervention–related adverse event. Another participant withdrew because of a miscarriage; this event was determined to be unrelated to the intervention. No participants required referral to psychiatric services due to acute suicidal risk or severe deterioration in mental health.

Discussion

Principal Findings

To our knowledge, this study represents the first randomized controlled trial in a Chinese cultural context to examine the effectiveness of an IPM+ intervention for antenatal depression. The results demonstrate that IPM+ was associated with significant and sustained improvements in depressive symptoms, anxiety, and perceived stress up to 3 months after the intervention. Notably, improvements in sleep quality exhibited a delayed pattern, with clearer benefits observed at follow-up (T3).

XSL•FO
RenderX

The Intervention Effect of IPM+ on Prenatal Depressive Mood

The results of this study demonstrate that the IPM+ intervention effectively alleviated AND, with effects sustained up to 3 months postintervention, consistent with the findings by Fink et al [35] showing prolonged benefits of telephone-delivered PM+ on maternal psychological distress. The significant efficacy observed in this study may stem from the synergistic intervention of the IPM+ modular design on the multidimensional causes of perinatal depression, including stress management techniques that regulate physiological stress responses, problem-solving strategies that enhance perceived control and self-efficacy, behavioral activation that counters low mood and social withdrawal, and social support facilitation that buffers stress and isolation. By simultaneously targeting the physiological, cognitive, behavioral, and social dimensions of depression, IPM+ may achieve sustained benefits within a relatively brief intervention period of 5 weeks. Notably, despite a higher proportion of participants with a history of prepregnancy depression in the intervention group, analyses indicated that prepregnancy depressive history did not significantly influence improvements in depressive symptoms, suggesting that IPM+ may be broadly applicable across perinatal women with varying levels of depression risk.

Compared with internet-based cognitive behavioral therapy (iCBT), IPM+ appears to be more time-efficient. A meta-analysis by Pan et al [36] indicated that interventions lasting fewer than 9 weeks did not produce significant improvements in postnatal depression (standardized mean difference=−0.36; P=.06), whereas the interventions of 9 weeks or longer demonstrated large and significant effects (standardized mean difference=−0.85; P<.01). In contrast, the IPM+ intervention achieved significant reductions in depressive symptoms within only 5 weeks, thereby reducing participant time burden and potentially enhancing the accessibility of psychological interventions for perinatal women. From a cost perspective, the IPM+ allows for delivery by nonspecialists, offering a scalable and economical alternative to therapist-intensive models, such as iCBT. This is particularly vital for implementation in resource-constrained settings.

Relative to existing internet-based mindfulness-based interventions, IPM+ also demonstrated comparatively larger effect sizes. For example, a randomized controlled trial by Kim et al. [37] reported moderate effects of a mobile-based mindfulness intervention on depressive severity (partial η^2=0.06), whereas this study observed a large effect size for perinatal depression. This disparity likely stems from the divergent mechanisms underlying each intervention. While internet-based mindfulness-based interventions primarily foster the awareness and acceptance of emotional distress, IPM+ is anchored in structured problem-solving—a framework that directly targets the concrete, real-world challenges inherent to the perinatal period, such as role transitions and shifting relationship dynamics. By equipping women to actively resolve these tangible stressors, IPM+ appears mechanistically better aligned with the needs of this population, ultimately driving the larger effect sizes observed in our study.

Furthermore, although IPT is well established as an effective treatment, IPM+ offers notable advantages in scalability. A systematic review by Couch et al [38] showed that IPT relies almost exclusively on specialist providers and typically requires 8‑14 sessions, limiting its large-scale implementation in resource-constrained settings. By contrast, the structured 5-week IPM+ program can be delivered by nonspecialist providers, substantially lowering the threshold for dissemination of depression interventions.

The Intervention Effect of IPM+ on Anxiety, Perceived Stress, and Sleep

This study further demonstrates that the IPM+ intervention exerted significant and sustained positive effects on anxiety and perceived stress among pregnant women, supporting its effectiveness as a transdiagnostic intervention recommended by the WHO. The intervention group showed sharp reductions in anxiety (PAQ) and stress scores immediately after the intervention (T2), with effects maintained at a moderate magnitude at the 3-month follow-up (T3). This pattern of rapid onset and durable improvement is highly consistent with the findings from face-to-face PM+ trials reported by Dawson et al [39] and Jordans et al [40], which showed superiority over usual care, providing strong evidence that the therapeutic effects of PM+ remain robust when extended to digital delivery formats. The marked improvements in anxiety and stress are likely attributable to the synergistic effects of the stress management and problem-solving modules, which may confer advantages over mindfulness-based interventions that primarily emphasize emotional awareness [41] or iCBT approaches that focus on cognitive restructuring [42] when addressing the unique stressors of the perinatal period. The stress management module offers relaxation techniques aimed at regulating heightened physiological arousal, consistent with established principles of relaxation training for anxiety management [11]. Importantly, the problem-solving module guides participants through structured steps to transform diffuse anxiety-provoking worries—such as excessive concerns about fetal health—into concrete, manageable problems with actionable plans. This process may directly target core cognitive maintenance factors of anxiety and stress, including intolerance of uncertainty and avoidance. The sustained reduction in stress observed at follow-up may further reflect the continued application of problem-solving skills in daily life, enhancing perceived control, alongside the facilitative role of the social support module in mobilizing available resources.

Sleep quality exhibited a distinct pattern of delayed improvement in our study, which contrasts with the findings from a CBT for insomnia study on perinatal insomnia by Quin et al [43], where interventions directly targeting sleep-related behaviors typically led to immediate postintervention improvements. The therapeutic pathway of PM+ may involve initial reductions in daytime emotional burden and cognitive rumination through the stress management and problem-solving modules, leading to decreased nocturnal psychophysiological arousal. Building on this foundation, the behavioral activation module may contribute to more regular daily routines and increased daytime activity, thereby supporting circadian rhythm stabilization, while the social support module may alleviate

XSL•FO
RenderX

feelings of loneliness and burden during pregnancy by strengthening social connectedness. Together, these changes form a gradual cycle of "reduced daytime emotional distress → enhanced nighttime calmness → progressive establishment of healthier sleep patterns." This cycle likely requires several weeks to consolidate into measurable sleep improvements, providing a plausible explanation for why sleep-related effects were not fully evident immediately postintervention (T2) but became significant at follow-up (T3). This finding aligns with emerging evidence highlighting emotion regulation as a key mechanism underlying improvements in comorbid insomnia [44].

Limitations and Suggestions

This study has several limitations that should be acknowledged. First, the sample size is relatively small (n=80) and of limited representativeness. Participants were recruited from a single tertiary hospital in Wuhan, where the levels of education and access to health care are generally high. Therefore, the generalizability of the study findings to the maternal population in rural areas or primary health care institutions may be limited. Future studies should adopt multicenter designs with larger, more diverse samples that include women from different geographic regions, health care tiers, and socioeconomic backgrounds to better evaluate the generalizability of IPM+. Second, we did not assess participants' blinding success after the intervention or examine whether group guessing influenced self-reported outcomes. Future studies could systematically evaluate participants' group guesses after the intervention and include this as a covariate in analyses to control for potential expectation effects or reporting bias. Third, symptom outcomes were primarily measured using self-report measures without confirmation via structured clinical interviews based on diagnostic criteria. Combining self-report scales with standardized diagnostic interviews in future work would help mitigate subjective bias and strengthen clinical validity. Finally, the 3-month follow-up period limits conclusions regarding the longer-term impact of IPM+ on postnatal mental health and mother-infant outcomes. Future studies with extended follow-up windows are needed and should include more ecologically meaningful endpoints, such as maternal-infant attachment quality and infant emotional-behavioral development.

This study presents a scalable framework for integrating perinatal mental health support into China's existing health care infrastructure. In the future, the core modules of IPM+ could be integrated into hospital-based official online service platforms as a universal and preventive psychological support tool for all pregnant women, facilitating early identification and timely intervention. In addition, the task-shifting potential of PM+ could be further leveraged through standardized training and supervision of maternal and child health providers in community health centers, enabling the delivery of preliminary yet fidelity-consistent psychological support at the primary care level. Such an approach may help alleviate the imbalance between the high demand for perinatal mental health services and the limited availability of specialist providers.

Conclusion

The findings from this randomized controlled trial indicate that the IPM+ intervention significantly alleviated depressive and anxiety symptoms, reduced perceived stress, and improved sleep quality among pregnant women, with benefits sustained at the 3-month follow-up. By employing structured modules that simultaneously address perinatal-specific stressors and internal emotional distress, IPM+ establishes a targeted and practical support framework. These results support IPM+ as an effective, sustainable, and scalable approach to promoting perinatal mental health. Future implementation should focus on integrating IPM+ into routine antenatal care pathways to enhance accessibility and impact.

Acknowledgments

The authors sincerely appreciate the participants and the support from the medical staff of Renmin Hospital of Wuhan University.

Funding

This study was supported by the National Social Science Foundation of China (Project No. 21BSH073).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

SZ, SW, and XS contributed to conceptualization, methodology, and investigation, with SZ and SW drafting the manuscript and XS verifying results. J Zhao and YL were responsible for data curation. SZ prepared figures and tables. QH, SL, and YC conducted formal analysis and visualization. The study was supervised by ZZ, XC, and J Zheng. Specifically, ZZ as the methodological lead led the intervention adaptation and data analysis; XC as the clinical lead oversaw trial design, ethics, execution, and funding acquisition; and J Zheng contributed to methodology design and provided critical intellectual revision. All authors reviewed and approved the final manuscript. ZZ and XC contributed equally as the corresponding authors of this manuscript. XC can be reached via email at chenxl72@whu.edu.cn.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
The problem management plus (PM+) group intervention program.
[DOCX File, 20 KB - jmir_v28i1e81998_app1.docx]

Multimedia Appendix 2
Problem management steps.
[DOC File, 23 KB - jmir_v28i1e81998_app2.doc]

Multimedia Appendix 3
Fixed effect estimation results of anxiety, perceived stress, and sleep quality; Comparison of analysis results before and after multiple imputation.
[DOCX File, 26 KB - jmir_v28i1e81998_app3.docx]

Checklist 1
CONSORT checklist.
[DOCX File, 38 KB - jmir_v28i1e81998_app4.docx]

References

1. Dagher RK, Bruckheim HE, Colpe LJ, Edwards E, White DB. Perinatal depression: challenges and opportunities. J Womens Health (Larchmt) 2021 Feb;30(2):154-159. [doi: 10.1089/jwh.2020.8862] [Medline: 33156730]
2. Yin X, Sun N, Jiang N, et al. Prevalence and associated factors of antenatal depression: systematic reviews and meta-analyses. Clin Psychol Rev 2021 Feb;83:101932. [doi: 10.1016/j.cpr.2020.101932] [Medline: 33176244]
3. Nisar A, Yin J, Waqas A, et al. Prevalence of perinatal depression and its determinants in mainland China: a systematic review and meta-analysis. J Affect Disord 2020 Dec 1;277:1022-1037. [doi: 10.1016/j.jad.2020.07.046] [Medline: 33065811]
4. Bränn E, Shen Q, Lu D. Perinatal depression and its health impact. BMJ 2024 Jan 9;384:2777. [doi: 10.1136/bmj.p2777] [Medline: 38195147]
5. Yu H, Shen Q, Bränn E, et al. Perinatal depression and risk of suicidal behavior. JAMA Netw Open 2024 Jan 2;7(1):e2350897. [doi: 10.1001/jamanetworkopen.2023.50897] [Medline: 38194232]
6. Slomian J, Honvo G, Emonts P, Reginster JY, Bruyère O. Consequences of maternal postpartum depression: a systematic review of maternal and infant outcomes. Womens Health (Lond Engl) 2019 Jan;15:1745506519844044. [doi: 10.1177/1745506519844044]
7. Ghimire U, Papabathini SS, Kawuki J, Obore N, Musa TH. Depression during pregnancy and the risk of low birth weight, preterm birth and intrauterine growth restriction—an updated meta-analysis. Early Hum Dev 2021 Jan;152:105243. [doi: 10.1016/j.earlhumdev.2020.105243] [Medline: 33190020]
8. Simonovich SD, Nidey NL, Gavin AR, et al. Meta-analysis of antenatal depression and adverse birth outcomes In US populations, 2010-20. Health Aff (Millwood) 2021 Oct;40(10):1560-1565. [doi: 10.1377/hlthaff.2021.00801] [Medline: 34606360]
9. Adeoye IA, Sogbesan A, Esan O. Prevalence, associated factors and perinatal outcomes of antepartum depression in Ibadan Nigeria. BMC Pregnancy Childbirth 2022 Mar 18;22(1):219. [doi: 10.1186/s12884-022-04549-7] [Medline: 35303804]
10. Molenaar NM, Kamperman AM, Boyce P, Bergink V. Guidelines on treatment of perinatal depression with antidepressants: an international review. Aust N Z J Psychiatry 2018 Apr;52(4):320-327. [doi: 10.1177/0004867418762057] [Medline: 29506399]
11. Domínguez-Solís E, Lima-Serrano M, Lima-Rodríguez JS. Non-pharmacological interventions to reduce anxiety in pregnancy, labour and postpartum: a systematic review. Midwifery 2021 Nov;102:103126. [doi: 10.1016/j.midw.2021.103126] [Medline: 34464836]
12. Cuijpers P, Franco P, Ciharova M, et al. Psychological treatment of perinatal depression: a meta-analysis. Psychol Med 2023 Apr;53(6):2596-2608. [doi: 10.1017/S0033291721004529] [Medline: 37310303]
13. Liu H, Yang Y. Effects of a psychological nursing intervention on prevention of anxiety and depression in the postpartum period: a randomized controlled trial. Ann Gen Psychiatry 2021 Jan 4;20(1):2. [doi: 10.1186/s12991-020-00320-4] [Medline: 33397393]
14. Pan WL, Lin LC, Kuo LY, Chiu MJ, Ling PY. Effects of a prenatal mindfulness program on longitudinal changes in stress, anxiety, depression, and mother-infant bonding of women with a tendency to perinatal mood and anxiety disorder: a randomized controlled trial. BMC Pregnancy Childbirth 2023 Jul 31;23(1):547. [doi: 10.1186/s12884-023-05873-2] [Medline: 37525110]
15. Pettman D, O'Mahen H, Blomberg O, Svanberg AS, von Essen L, Woodford J. Effectiveness of cognitive behavioural therapy-based interventions for maternal perinatal depression: a systematic review and meta-analysis. BMC Psychiatry 2023 Mar 29;23(1):208. [doi: 10.1186/s12888-023-04547-9] [Medline: 36991389]

XSL•FO
RenderX

16. Boele FW, Klein M, Verdonck-de Leeuw IM, et al. Internet-based guided self-help for glioma patients with depressive symptoms: a randomized controlled trial. J Neurooncol 2018 Mar;137(1):191-203. [doi: 10.1007/s11060-017-2712-5] [Medline: 29236238]
17. Cacciola E, Psouni E. Insecure attachment and other help-seeking barriers among women depressed postpartum. Int J Environ Res Public Health 2020 May 30;17(11):3887. [doi: 10.3390/ijerph17113887] [Medline: 32486285]
18. Dawson KS, Bryant RA, Harper M, et al. Problem Management Plus (PM+): a WHO transdiagnostic psychological intervention for common mental health problems. World Psychiatry 2015 Oct;14(3):354-357. [doi: 10.1002/wps.20255] [Medline: 26407793]
19. Problem management plus (PM+): individual psychological help for adults impaired by distress in communities exposed to adversity. : World Health Organization; 2016 URL: https://iris.who.int/server/api/core/bitstreams/0d493975-285d-4d0b-86b1-2365cfe45748/content [accessed 2026-01-23]
20. Hamdani SU, Huma ZE, Masood A, et al. Effect of adding a psychological intervention to routine care of common mental disorders in a specialized mental healthcare facility in Pakistan: a randomized controlled trial. Int J Ment Health Syst 2021 Jan 19;15(1):11. [doi: 10.1186/s13033-020-00434-y] [Medline: 33468192]
21. Zhang H, Zhang D, Lin H, Huang L. Problem management plus in the treatment of mental disorders in patients with multiple myeloma. Support Care Cancer 2020 Oct;28(10):4721-4727. [doi: 10.1007/s00520-019-05289-9] [Medline: 31970512]
22. de Graaff AM, Cuijpers P, Elsawy M, et al. The effectiveness of Problem Management Plus at 1-year follow-up for Syrian refugees in a high-income setting. Epidemiol Psychiatr Sci 2024 Oct 25;33:e50. [doi: 10.1017/S2045796024000519] [Medline: 39449610]
23. Coleman SF, Mukasakindi H, Rose AL, et al. Adapting problem management plus for implementation: lessons learned from public sector settings across Rwanda, Peru, Mexico and Malawi. Intervention (Amstelveen) 2021;19(1):58-66. [Medline: 34642580]
24. Danaher BG, Seeley JR, Silver RK, et al. Trial of a patient-directed eHealth program to ameliorate perinatal depression: the MomMoodBooster2 practical effectiveness study. Am J Obstet Gynecol 2023 Apr;228(4):453. [doi: 10.1016/j.ajog.2022.09.027] [Medline: 36174746]
25. Problem management plus (PM+): individual psychological help for adults impaired by distress in communities exposed to adversity [Report in Chinese]. : CITIC Press Group; 2022 URL: https://iris.who.int/server/api/core/bitstreams/54198a49-f09e-4307-ac75-ff2353bf42ff/content [accessed 2026-02-05]
26. Cox JL, Holden JM, Sagovsky R. Detection of postnatal depression: development of the 10-item Edinburgh Postnatal Depression Scale. Br J Psychiatry 1987 Jun;150:782-786. [doi: 10.1192/bjp.150.6.782] [Medline: 3651732]
27. Lee DT, Yip SK, Chiu HF, et al. Detecting postnatal depression in Chinese women: validation of the Chinese version of the Edinburgh Postnatal Depression Scale. Br J Psychiatry 1998 May;172:433-437. [doi: 10.1192/bjp.172.5.433] [Medline: 9747407]
28. Cox J. Thirty years with the Edinburgh Postnatal Depression Scale: voices from the past and recommendations for the future. Br J Psychiatry 2019 Mar;214(3):127-129. [doi: 10.1192/bjp.2018.245] [Medline: 30774059]
29. Xiao L. Development of the pregnancy-specific anxiety questionnaire and research on pregnancy-specific anxiety in early pregnancy [Master's thesis in Chinese]. : Southern Medical University; 2010 URL: https://d.wanfangdata.com.cn/thesis/Ch1UaGVzaXNOZXdTb2xyOVMyMDI2MDExNzA4NTkxNhIHRDEyODgyNBoIcnEyMzd6amg%25253D [accessed 2026-02-05]
30. Leung DY, Lam TH, Chan SS. Three versions of Perceived Stress Scale: validation in a sample of Chinese cardiac patients who smoke. BMC Public Health 2010 Aug 25;10:513. [doi: 10.1186/1471-2458-10-513] [Medline: 20735860]
31. Staneva A, Bogossian F, Pritchard M, Wittkowski A. The effects of maternal depression, anxiety, and perceived stress during pregnancy on preterm birth: a systematic review. Women Birth 2015 Sep;28(3):179-193. [doi: 10.1016/j.wombi.2015.02.003] [Medline: 25765470]
32. Morin CM. Insomnia: Psychological Assessment and Management: Guilford Press; 1996. URL: https://search.worldcat.org/zh-cn/title/27385461 [accessed 2026-02-05]
33. Ansaari N, Rajan SK, Kuruveettissery S. Efficacy of in-person versus digital mental health interventions for postpartum depression: meta-analysis of randomized controlled trials. J Reprod Infant Psychol 2025 Jul;43(4):1067-1087. [doi: 10.1080/02646838.2024.2303470] [Medline: 38221725]
34. Rahman A, Khan MN, Hamdani SU, et al. Effectiveness of a brief group psychological intervention for women in a post-conflict setting in Pakistan: a single-blind, cluster, randomised controlled trial. The Lancet 2019 Apr;393(10182):1733-1744. [doi: 10.1016/S0140-6736(18)32343-2]
35. Fink G, Melero-Dominguez M, Chembe M, et al. Feasibility and acceptability of the Problem Management for Moms programme for improving maternal mental health in Zambia: an open-label trial. Lancet Psychiatry 2024 Dec;11(12):965-974. [doi: 10.1016/S2215-0366(24)00256-6] [Medline: 39515361]
36. Pan J, Luo W, Zhang H, et al. The effects of online cognitive behavioral therapy on postpartum depression: a systematic review and meta-analysis. Healthcare (Basel) 2025 Mar 21;13(7):696. [doi: 10.3390/healthcare13070696] [Medline: 40217993]

XSL•FO
RenderX

37. Kim M, Park JY, Park S, et al. Effectiveness of a mindfulness-based mobile intervention for improving perinatal mental health and reducing depression during pregnancy: randomized controlled trial. J Med Internet Res 2025 Oct 2;27:e75630. [doi: 10.2196/75630] [Medline: 41037777]
38. Couch E, Mai HJ, Kanaan G, et al. Effectiveness of psychological therapies for depression during the perinatal period: a systematic review and meta-analysis. Ann Intern Med 2025 Dec;178(12):1752-1762. [doi: 10.7326/ANNALS-24-03520] [Medline: 41183344]
39. Dawson KS, Schafer A, Anjuri D, et al. Feasibility trial of a scalable psychological intervention for women affected by urban adversity and gender-based violence in Nairobi. BMC Psychiatry 2016 Nov 18;16(1):410. [doi: 10.1186/s12888-016-1117-x] [Medline: 27863515]
40. Jordans MJD, Kohrt BA, Sangraula M, et al. Effectiveness of Group Problem Management Plus, a brief psychological intervention for adults affected by humanitarian disasters in Nepal: a cluster randomized controlled trial. PLoS Med 2021 Jun;18(6):e1003621. [doi: 10.1371/journal.pmed.1003621] [Medline: 34138875]
41. Mao F, Sun Y, Li Y, Cui N, Cao F. Internet-delivered mindfulness-based interventions for mental health outcomes among perinatal women: a systematic review. Asian J Psychiatr 2023 Feb;80:103321. [doi: 10.1016/j.ajp.2022.103321] [Medline: 36423436]
42. Mahoney A, Shiner CT, Grierson AB, et al. Online cognitive behaviour therapy for maternal antenatal and postnatal anxiety and depression in routine care. J Affect Disord 2023 Oct 1;338:121-128. [doi: 10.1016/j.jad.2023.06.008] [Medline: 37295653]
43. Quin N, Tikotzky L, Astbury L, et al. Preventing postpartum insomnia: findings from a three-arm randomized-controlled trial of cognitive behavioral therapy for insomnia, a responsive bassinet, and sleep hygiene. Sleep 2024 Aug 14;47(8):zsae106. [doi: 10.1093/sleep/zsae106] [Medline: 38736364]
44. Meneo D, Samea F, Tahmasian M, Baglioni C. The emotional component of insomnia disorder: a focus on emotion regulation and affect dynamics in relation to sleep quality and insomnia. J Sleep Res 2023 Dec;32(6):e13983. [doi: 10.1111/jsr.13983] [Medline: 37394234]

Abbreviations

AND: antenatal depression
CBT: cognitive behavioral therapy
CONSORT: Consolidated Standards of Reporting Trials
EPDS: Edinburgh Postnatal Depression Scale
iCBT: internet-based cognitive behavioral therapy
IPM+: internet-based problem management plus
IPT: interpersonal psychotherapy
ISI: Insomnia Severity Index
PAQ: Pregnancy Anxiety Questionnaire
PM+: problem management plus
PSS: 4-item Perceived Stress Scale
WHO: World Health Organization

Edited by A Schwartz, M Balcarras; submitted 07.Aug.2025; peer-reviewed by S Alves, X Xiao; accepted 08.Jan.2026; published 27.Mar.2026.

Please cite as:
Zheng S, Wang S, Shui X, Zhao J, Lin Y, Huang Q, Li S, Chen Y, Zou Z, Zheng J, Chen X
Internet-Based Problem Management Plus Intervention for Antenatal Depression: Randomized Controlled Trial
J Med Internet Res 2026;28:e81998
URL: https://www.jmir.org/2026/1/e81998
doi:10.2196/81998

© Shuanghui Zheng, Shiyu Wang, Xiaoyu Shui, Jing Zhao, Ying Lin, Qian Huang, Siqi Li, Yang Chen, Zhijie Zou, Jin Zheng, Xiaoli Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Intervention (MiVacunaLA 2.0) to Promote COVID-19 Vaccine Acceptance Among Hispanic Children: Community-Based Randomized Controlled Trial

Luisa R Blanco[1], MBA, PhD; Alexandra Klomhaus[2], MS, PhD; Rebecca Dudovitz[3], MSHS, MD; Denise Marquez[4], MPH; Ray Lopez-Chang[5], MUP; Evelyn Aleman[6], MPP; Bas Weerman[5]; Michael Moldoff[5], MA; Keith Norris[2], MD, PhD; Yelba Castellon-Lopez[7], MS, MD

[1]School of Public Policy, Pepperdine University, 24255 Pacific Coast Highway, Malibu, CA, United States

[2]Department of Medicine, David Geffen School of Medicine, Division of General Internal Medicine and Health Services Research, University of California, Los Angeles, Los Angeles, CA, United States

[3]Department of Pediatrics, David Geffen School of Medicine, Division of General Pediatrics, University of California, Los Angeles, Los Angeles, CA, United States

[4]Self-Sufficiency Services, Office of Immigrant and Refugee Affairs, County of San Diego, Health and Human Services Agency, San Diego, CA, United States

[5]Center for Economic and Social Research, USC Dornsife, University of Southern California, Los Angeles, CA, United States

[6]Our Voice: Communities for Quality Education (Nuestra Voz), Los Angeles, CA, United States

[7]Cancer Research Center for Health Equity, Samuel Oschin Comprehensive Cancer Institute, Department of Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, CA, United States

Corresponding Author:
Luisa R Blanco, MBA, PhD
School of Public Policy, Pepperdine University, 24255 Pacific Coast Highway, Malibu, CA, United States

Abstract

Background: Early in the children's COVID-19 rollout in the United States, racial and ethnic vaccination rate disparities were evident. Based on COVID-19 communication literature and qualitative interviews with Hispanic parents, we developed a mobile phone–delivered digital intervention to address factors associated with low vaccine confidence.

Objective: We conducted a community-based randomized controlled trial of a digital intervention called MiVacunaLA/MyShotLA to increase COVID-19 vaccine uptake among Hispanic children. The fully automated digital intervention was designed in collaboration with community organizations and linguistically and culturally tailored to meet the informational needs of Hispanic caregivers. The intervention focused on families with unvaccinated children 5 to 11 years old but was offered to families with any unvaccinated children 17 years or younger.

Methods: Participants were recruited with community organization partners and trained parent ambassadors via an open online screener. The 4-week intervention consisted of 3 SMS text messages with culturally and linguistically tailored educational information weekly. Intervention materials were delivered digitally through a closed online platform. Study team members were blinded. We used a difference-in-difference model with an intention-to-treat approach. The primary outcome was self-reported COVID-19 vaccine uptake among household children collected via online questionnaires. Secondary outcomes included COVID-19 vaccine knowledge, vaccine trust, and measures of participant engagement. We conducted a sensitivity analysis using the treatment-on-the-treated approach.

Results: In total, 254 participants completed the baseline survey (128 control and 126 intervention). The average participant age was 34 (SD 6.3) years with an average of 1.7 (SD 0.8) minors in the household, and among households, 62.2% (n=158) reported having children aged 5 to 11 years old. Most participants (n=207, 81.5%) reported English as their primary language. We found a statistically significant difference of 13.3% (95% CI 0.3%-26.4%; *P*=.04) points in self-reported vaccine uptake between intervention and control groups among caregivers of Hispanic children aged 5 to 11 years old. We also found a statistically significant point difference of 14.3% (95% CI 0% - 23.7%; *P*=.003) between intervention and control groups in trust of governmental approval processes for the children's COVID-19 vaccine. Most participants reported that the weekly digital videos and educational information were "very" (892/1031, 86.5%) or "extremely" (888/1019, 87.1%) useful.

Conclusions: MiVacunaLA demonstrates that a culturally tailored, community-based, mobile phone–delivered vaccine educational intervention can increase COVID-19 vaccine uptake among Hispanic children and improve caregivers' trust in governmental vaccine processes. MiVacunaLA is innovative in its integration of community-informed design with a fully automated, mobile phone–centric format and builds on prior literature by prospectively evaluating a culturally tailored, SMS text messaging–linked

XSL•FO
RenderX

web curriculum in a community setting. Findings provide evidence that scalable, low-cost, digital strategies can measurably improve trust and uptake in a population facing persistent vaccination gaps. Real-world implications include the portability and adaptability of this approach across diverse communities and settings to support timely, community-engaged vaccination efforts for broader applicability and scalability in public health.

Trial Registration: ClinicalTrials.gov NCT05234372; https://clinicaltrials.gov/study/NCT05234372

(J Med Internet Res 2026;28:e78103) doi:10.2196/78103

KEYWORDS

children digital education; Latino/Hispanic children; health equity; COVID-19 vaccines; mobile-based interventions

Introduction

Background

During the initial rollout of the COVID-19 vaccine for children of ages 5 to 11 years in the United States, children identifying as part of a racial and ethnic minority group were underrepresented among first-dose recipients [1-5]. This was due, in part, to lack of access to accurate information, concerns about vaccine safety and efficacy, exposure to disinformation and misinformation, and lack of trust in governmental vaccine processes [5-8]. Studies show that parental and caregiver acceptance of COVID-19 vaccination for their children is positively associated with beliefs about vaccination benefits [9]. Improving parental and caregiver acceptance is a key public health focus, given the trends in childhood vaccine uptake after the COVID-19 pandemic.

Following the pandemic, vaccination rates among US kindergartners across other childhood vaccinations declined compared to prior years [10,11]. Meanwhile, the United States has seen the resurgence of previously eliminated or highly controlled vaccine-preventable diseases, such as measles-mumps-rubella (MMR) [12,13]. While prepandemic vaccination rate disparities existed among racial and ethnic subgroups [14], these trends were further exacerbated by the pandemic [15]. As such, novel interventions are needed to effectively combat disinformation and misinformation, rebuild trust in vaccination benefits and safety, and engage undervaccinated communities.

In the early stages of the COVID-19 vaccine rollout, leveraging community-engaged strategies aided in promoting vaccine confidence and addressing barriers such as language accessibility, addressing disinformation and misinformation, and building trust in vaccination among the Hispanic population [16,17]. Incorporating community voices and perspectives into vaccine promotion and distribution strategies can have more meaningful impact with community members as opposed to traditional communication methods [18,19]. Digital interventions have also increased uptake of other vaccines, such as human papillomavirus vaccination, suggesting that community-based digital interventions might be particularly effective at increasing vaccine confidence and uptake to address recent declines in children vaccination rates [20,21]. However, few randomized controlled trial (RCT) studies exist on the use of community-based digital interventions to improve COVID-19 vaccination rates among Hispanic children. Testing and identifying such interventions can yield critical public health and primary care approaches and lessons for responding to both current vaccine uptake disparities and future public health emergencies.

Study Objectives

This study describes efforts to increase COVID-19 vaccination rates among Hispanic children of ages 5 to 11 years in the Los Angeles County area through a community-informed culturally and linguistically appropriate RCT. We evaluated the efficacy of a digital intervention, called MiVacunaLA/MyShotLA (MVLA) 2.0, for parents or caregivers of Hispanic children to increase COVID-19 vaccination uptake for children. We also evaluated the intervention's efficacy in increasing knowledge of COVID-19 vaccines as well as trust in the governmental approval process for the vaccine.

Methods

Intervention Content and Approach

MVLA 2.0 builds on MVLA 1.0, which was created in response to the US Food and Drug Administration–approved COVID-19 vaccines for children of ages 12 to 17 years and COVID-19 communication literature [22]. The MVLA 1.0 design and implementation are described in detail elsewhere [22]. Briefly, MVLA 1.0 used a community-partnered approach to develop and distribute COVID-19 vaccine education via SMS text messaging, emails, and brief videos multiple times a week for 4 weeks. MVLA 2.0, the focus of this study, began recruitment between June and August of 2022, with the intervention implemented between August and November 2022.

For MVLA 2.0, we updated and refined the digital materials to focus on vaccination for children aged 5 to 11 years based on parental and community partner feedback. We conducted focus groups with parents and caregivers of children between the ages of 5 and 11 years from MVLA 1.0 to tailor MVLA 2.0 for the 5- to 11-year-old age group. Detailed focus group results are reported in a separate paper [23] and revealed the need to adapt the intervention's educational content to address emerging COVID-19 vaccine concerns specific to children, as well as myths and misinformation [23]. Next, we collaborated with community-based organizations to conduct community sessions with parent advocates from these organizations. Participants suggested several new features to MVLA that included (1) infographics to visually summarize weekly content, (2) voice-overs in English and Spanish of text to address literacy needs, (3) a discussion board where participants could post questions about the vaccine and have a physician respond, and

XSL•FO
RenderX

(4) parent testimonials from those who had vaccinated their children. In addition, community partners identified the need for new content to address questions about emerging COVID strains. These suggestions were all incorporated into the MVLA 2.0 intervention.

For the MVLA 2.0 intervention, participants received 3 SMS text messaging or emails per week (Monday, Wednesday, and Thursday) inviting them to complete a specific activity. On Mondays, participants were invited to watch a relevant informational video for that week. On Wednesdays, participants were invited to visit the platform to read additional relevant educational information. On Thursdays, participants received a summary infographic of the week's material. Figures S1 and S2 in Multimedia Appendix 1 provide examples of the infographics we used for week 3 (English and Spanish), and Table S1 in Multimedia Appendix 1 provides an outline of the intervention material and activities.

To deliver MVLA 2.0, we utilized the digital infrastructure of the Understanding America Study (UAS) platform. We created a participant-only website to provide access to weekly study content. The online platform for MVLA 2.0 included all study information, a discussion board, and downloadable infographics summarizing each week's COVID-19 topic. Participants could select their preferred language (English or Spanish) to engage with the website and all study materials. Table S2 in Multimedia Appendix 1 shows a detailed description of the MVLA 2.0 intervention components (visual summaries, audio content, discussion board, and testimonial videos).

Community Engagement, Recruitment, Timeline, and Other Intervention Logistics

We applied community-partnered research methods throughout the study. Recruitment involved community organization leaders and a group of parent ambassadors—parents who completed MVLA 1.0 and expressed interest in continuing to raise awareness by aiding with recruitment for MVLA 2.0. Recruitment was conducted through public school serving organizations located in East and South Los Angeles, communities with a high proportion of Hispanic families and disproportionately lower COVID-19 vaccination rates in children and youth based on Los Angeles County Department of Public Health Data [24]. We recruited participants using word-of-mouth and community partner social media sites to share the project flyer within their networks. The following organizations supported recruitment efforts by sharing our study flyer: Best Start LA, Coalition for Humane Immigrant Rights of Los Angeles, Communities in School, Families in School, Inner City Struggle, Great Public Schools Now, KIPP Public Charter Schools, Latino Equality Alliance, Mexican American Legal Defense and Education Fund, Mexican American Opportunity Foundation, Nuestra Voz (Our Voice), and Young Men's Christian Association of Metropolitan Los Angeles.

We planned to start recruitment when vaccines for children of ages 5 to 11 were newly released. Therefore, we tailored the intervention to address informational needs of Hispanic parents or caregivers of children in this age group. While the intervention information was relevant across all age groups, recruitment and intervention materials focused on households with children of ages 5 to 11 years as we anticipated the greatest impact among this age group. COVID-19 vaccines for children younger than 5 years were released in June 2022, and therefore, we updated intervention content to reflect up-to-date vaccine approval for the younger group (see Figures S1 and S2 in Multimedia Appendix 1).

We recruited parents and caregivers of Hispanic minors below 18 years of age who had access to a computer or cell phone to receive the intervention (inclusion criteria) via an online access open screener. Recruitment efforts were targeted not only to reach families with unvaccinated children of ages 5 to 11 years but also allowed interested families with at least one child below 18 years to participate in the intervention. While recruitment was focused among Los Angeles County parents, participants were not excluded if they resided in other neighboring areas or were from other racial or ethnic groups (no exclusion criteria).

We ended recruitment 2 weeks prior to starting the intervention to allow time to randomize all participants and enter their information for receiving SMS text messaging and emails into the digital platform. Figure S3 in Multimedia Appendix 1 shows the intervention implementation timeline. Participants first filled out an online screening survey to determine eligibility. Qualifying participants received an SMS text messaging and email inviting them to participate and directing them to the electronic consent. Upon consenting, participants were directed to the baseline survey. We did not establish specific eligibility criteria for sites or for individuals delivering the intervention, as the intervention was delivered digitally.

We used separate block randomization for cohorts, conducted by one UAS staff using Stata software, version 17 (StataCorp LLC). Randomized participants received MVLA 2.0 either at month 1 (intervention) or at month 2 (control). We used a 1:1 allocation ratio for randomization to the intervention and control groups. All participants were invited to complete a baseline and a 1-month follow-up survey. In the first month, the intervention group received educational material for 4 weeks, while the control group received a biweekly message reminding them of the number of days left until the beginning of month 2 when they were scheduled to start MiVacunaLA. After the intervention group completed the 4-week educational intervention, we sent all participants reminders twice a week for 2 weeks to complete the 1-month follow-up survey. The control group received the intervention after all groups completed the follow-up survey.

All intervention materials were delivered digitally. Participants completed the intervention activities in order. If a participant did not complete an intervention activity by the end of the week, they still received content for the following week. However, when participants attempted to access the next activity, they were prompted to first complete the missed activity. Participants had the flexibility to complete all intervention activities by the end of the month, with an additional 1-week grace period before the 1-month follow-up survey became available. Once the survey was available, participants could not complete additional intervention activities.

XSL•FO
RenderX

Ethical Considerations

MiVacunaLA 2.0 was registered and approved by the University of California, Los Angeles Institutional Review Board (21‐000857). Participants completed an online informed consent form in their preferred language. Consent information was also available to participants as a PDF on the intervention platform for their reference. Participants' contact information was collected in the screening survey. A study principal investigator worked with UAS staff to create identifiers so that all data downloaded from the UAS platform, and subsequently used in data analysis, were deidentified. Thus, no identification of individual participants in any images of the paper or supplementary material is possible. Only 1 principal investigator and 1 UAS staff member had access to participant identifiers, which were password-protected in an encrypted file, as well as access to pretrial randomization information. Study team members were blinded. After randomization, intervention and control participants started the intervention immediately (intervention arm) or in 2 months (waitlist control arm). Participants received a US $40 gift card sent electronically after completing all intervention activities.

Given this was a behavioral intervention and no concomitant care was provided during the intervention, there was minimal risk, and we did not collect data regarding harms. There were no significant changes to the trial after commencement or interim analysis during the data collection phase. We added information about US Food and Drug Administration–approved COVID-19 vaccines for children of ages 6 months to 4 years to the material during the design phase. During consent, participants were informed that they could decline participation at any time and could opt out of any intervention activity they did not feel comfortable completing. In addition to the detailed information provided on the curriculum and digital approach, those interested in accessing all the educational materials and survey documents to replicate this RCT may contact the corresponding author.

Primary Outcomes

The primary outcome was self-reported vaccination status of household children in 3 age groups based on COVID-19 vaccine availability pre- and postintervention. The age groups used were (1) 6 months to 4 years, (2) 5 to 11 years, and (3) 12 to 17 years. The primary outcome was analyzed at the household level with at least 1 child in the specified age groups. The baseline survey asked for vaccination status for all household children. When participants confirmed a specific child's vaccination, only the status of the household's remaining nonvaccinated children was queried in the follow-up survey. We imputed the same vaccination status for the household expressed at baseline and the 1-month follow-up survey. For those with missing data, we assumed that if a household vaccinated 1 child in a specific age group, then the household's children in that same age group were vaccinated at follow-up.

Table S3 in Multimedia Appendix 1 shows the vaccination status of qualifying households with children by age groups at baseline and follow-up for all participants, regardless of loss to follow-up. For those participants who did not complete the follow-up survey, we assumed that the vaccination status remained unchanged from baseline.

Secondary Outcomes

Secondary outcomes included perceived change in COVID-19 vaccine knowledge and trust from baseline and follow-up. For vaccine knowledge, we asked participants to indicate whether the following statement was true or false: "After being fully vaccinated with the COVID-19 vaccine, your chances of being hospitalized or dying from COVID-19 are reduced at least by 90% if you contract the virus." We coded the variable as equal to 1 if participants answered true, and 0 if they answered false or unsure. Second, we asked participants, "How would you rate your knowledge about the COVID-19 vaccine?" This variable was coded equal to 1 if participants answered adequate or superior knowledge, and equal to 0 if the answer was basic, minimal, or no knowledge.

We measured vaccine trust via 2 separate questions: "How much do you trust the governmental approval process to ensure the COVID-19 vaccine is safe for (1) the public and (2) children?" We coded this variable as equal to 1 for those who answered they trust government fully, mostly, or somewhat, and equal to 0 if they answered they do not trust the government for each question. We also measured participant engagement with the intervention regardless of the study arm. This analysis included all participants (intervention and waitlist control groups together) who interacted with the material on the digital platform and completed at least 1 activity. At the end of each week, we collected information on the perceived usefulness of the weekly video and in-platform information and participants' familiarity with the information. We also tracked participants' in-platform click-through rates of a link to the Los Angeles County Department of Public Health appointment portal for COVID-19 vaccination; whether participants downloaded infographics to share; and whether participants utilized audio files for written information and infographics sent midweek.

Statistical Analysis

We evaluated efficacy using an intention-to-treat (ITT) approach. For the ITT analysis, we imputed the vaccination status for those who did not complete the follow-up survey with the same value entered at baseline. This conservative approach assumes no behavioral change among participants who were lost to follow-up. We used the same approach for the secondary outcomes.

We used a treatment-on-the-treated (TOT) approach as a sensitivity analysis. For the TOT analysis, we only included participants who completed the follow-up survey and whose vaccination status was reported at follow-up. We used a difference-in-difference (DID) model estimation for ITT and TOT approaches. Table S4 in Multimedia Appendix 1 shows the description of the DID model specification used for the statistical analysis.

To fit these models, we used a generalized linear model, using maximum likelihood estimation, with a binomial distribution and an identity link function to fit vaccination status as the dependent variable, with main fixed effects for both month (0=baseline and 1=follow-up) and intervention arm (0=control

XSL•FO
RenderX

and 1=intervention), and an interaction term between month and intervention. The interaction term, defined specifically as Month×Intervention, represents the primary variable of interest and reflects the DID in the outcome variable. The models also accounted for repeated measures within the same household and utilized an unstructured correlation structure.

We ran Little's Missing Completely at Random (MCAR) test to evaluate whether our data were missing completely at random. Subsequently, for the TOT analysis, we evaluated whether statistically significant differences exist between those who completed follow-up surveys and those who did not. We then modeled the TOT data using estimated weights based on significantly different characteristics between the 2 groups. This approach was used to help ensure we were not overestimating efficacy based on the characteristics of those who completed follow-up surveys. We estimated a Probit model where we regressed the attrition-related variable on those statistically significant variables between the 2 groups and created weights inverting these probabilities to estimate the intervention's efficacy on COVID-19 vaccine uptake among household children within a specific age group. This RCT was reported using the CONSORT 2025 (Consolidated Standards of Reporting Trials) checklist (Checklist 1) [25].

Results

Demographic Characteristics for Overall, Control, and Intervention Samples

Figure 1 presents the CONSORT (Consolidated Standards of Reporting Trials) flow diagram for the intervention. Based on screening survey completion and study criteria, we invited 512 qualifying participants. Among those, 310 (60.5%) people consented to participate, completed the baseline survey, and were randomized to intervention or control groups. From those who consented and completed the initial baseline survey, we excluded 56 participants who reported no household minors in their baseline responses, which conflicted with their screener responses. Of the remaining 254 parents and caregivers in the baseline sample (126 and 128 participants in intervention and control groups, respectively), we had 216 participants who completed the 1-month follow-up survey (99 and 117 participants in intervention and control groups, respectively) with a retention rate of 85%. We had 17% missing data for the primary outcome or vaccine uptake. Of those, we imputed 37 for a total analytical sample of 247. We had 4% missing data for secondary outcomes or vaccine knowledge and trust. Our total analytic sample for secondary outcomes was 243.

Table 1 provides baseline demographic characteristics for the full sample and for control and intervention groups separately. The sample includes 254 participants with a household minor recruited at baseline. Most participants were Hispanic (n=230, 90.5%) and completed the intervention in English (n=207, 81.5%). Most participants had some college education or more (n=196, 77.2%), were employed (n=202, 79.5%), and over half (n=150, 59.1%) had an annual household income greater than US $50,000. Just over half had children aged 5 to 11 years (n=158, 62.2%), and about one-third had children aged 12 to 17 years (n=92, 36.2%). The average age of parents or caregivers was 33.8 (SD 6.3) years, and the average number of children in the household was 1.7 (SD 0.8). We evaluated whether differences in participant characteristics between study arms were statistically significant using a *t* test.

XSL•FO
RenderX

Figure 1. The CONSORT (Consolidated Standards of Reporting Trials) flow diagram for the intervention.

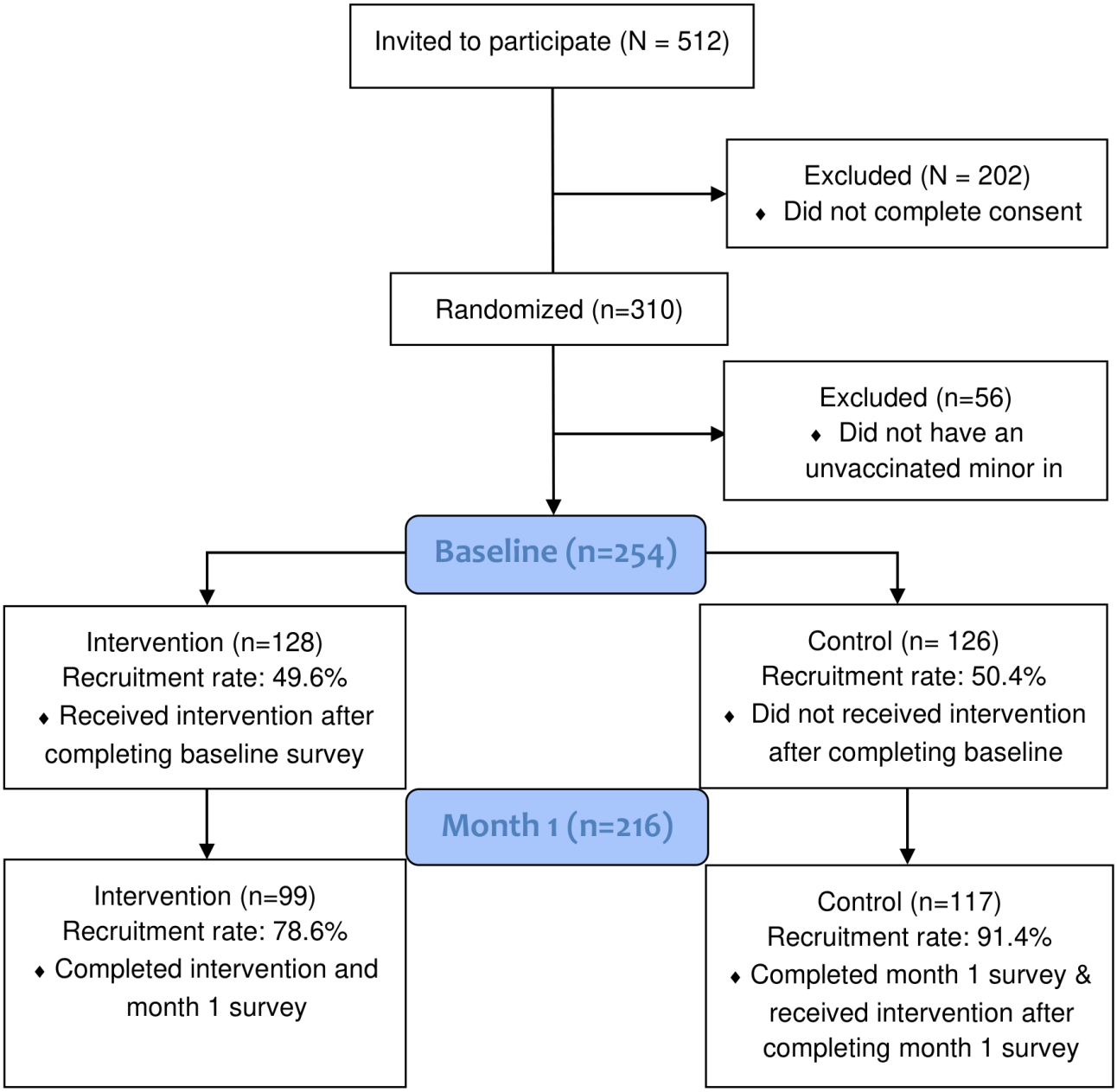

XSL•FO
RenderX

Table . Baseline demographic characteristics by overall, control, and intervention samples.

Characteristic	Overall (N=254)[a]	Control (n=128)	Intervention (n=126)	*P* value[b]
Age of parent, mean (SD)[c]	33.8 (6.3)	33.7 (6.2)	33.9 (6.4)	.68
Number of minors in household, mean (SD)	1.7 (0.8)	1.7 (0.9)	1.6 (0.7)	.82
Language, n (%)				.45
English	207 (81.5)	102 (79.7)	105 (83.3)	
Spanish	47 (18.5)	26 (20.3)	21 (16.7)	
Parent COVID-19 vaccination status, n (%)				.31
Vaccinated	95 (37.4)	51 (39.8)	44 (34.9)	
Not vaccinated	155 (61)	76 (59.4)	79 (62.7)	
Unsure/missing	4 (1.6)	1 (0.8)	3 (2.4)	
Ethnicity, n (%)				.89
Not Hispanic/Spanish origin or missing	24 (9.4)	11 (8.6)	13 (10.3)	
Mexican/Mexican American/Chicano	152 (59.8)	79 (61.7)	73 (57.9)	
Other Hispanic/Spanish origin[d]	78 (30.7)	38 (29.7)	40 (31.8)	
Born in the United States, n (%)				.18
Yes	187 (73.6)	99 (77.3)	88 (69.8)	
No	52 (20.5)	25 (19.5)	27 (21.4)	
Prefer not to respond/did not respond	15 (5.9)	4 (3.1)	11 (8.7)	
Highest education attained, n (%)				.99
Some high school or less	22 (8.7)	11 (8.6)	11 (8.7)	
High school graduate/GED[e]	30 (11.8)	15 (11.7)	15 (11.9)	
Some college or more	196 (77.2)	100 (78.1)	96 (76.2)	
Missing	6 (2.4)	2 (1.6)	4 (3.2)	
Employment status, n (%)				.10
Employed	202 (79.5)	107 (83.6)	95 (75.4)	
Unemployed	8 (3.2)	3 (2.3)	5 (4)	
Other[f]	44 (17.3)	18 (14.1)	26 (20.6)	
Household income, n (%)				.64
<$25,000	51 (20.1)	23 (18)	28 (22.2)	
$25,000-$49,000	46 (18.1)	25 (19.5)	21 (16.7)	
>$50,000	150 (59.1)	77 (60.2)	73 (57.9)	
Missing	7 (2.8)	3 (2.3)	4 (3.2)	
Health insurance status, n (%)				.24
Insured[g]	227 (89.4)	115 (89.8)	112 (88.9)	
Not insured/don't know	27 (10.6)	13 (10.2)	14 (11.1)	

XSL•FO
RenderX

Characteristic	Overall (N=254)[a]	Control (n=128)	Intervention (n=126)	*P* value[b]
Marital status, n (%)				.40
Currently married	203 (79.9)	105 (82)	98 (77.8)	
Widowed/divorced/separated	20 (7.9)	9 (7)	11 (8.7)	
Never married	20 (7.9)	8 (6.3)	12 (9.5)	
Other[h]	11 (4.3)	6 (4.7)	5 (4)	
Type of household, n (%)				.61
Married with children	207 (81.5)	107 (83.6)	100 (79.4)	
Single/married without children	7 (2.8)	2 (1.6)	5 (4)	
Single with children	16 (6.3)	9 (7)	7 (5.6)	
Other[i]	24 (9.4)	10 (7.8)	14 (11.1)	
Any minors in household under 6 months of age, n (%)				
Yes[j]	8 (3.1)	5 (3.9)	3 (2.4)	.72
Any minors in household from 6 months to 4 years, n (%)				
Yes[j]	105 (41.3)	51 (39.8)	54 (42.9)	.63
Any minors in household from 5 to 11 years, n (%)				
Yes[j]	158 (62.2)	87 (68)	71 (56.4)	.06
Any minors in household from 12 to 17 years, n (%)				
Yes[j]	92 (36.2)	44 (34.4)	48 (38.1)	.54

[a]Excludes n=56 who reported no minors in their household at baseline.

[b]*P* value from χ^2 tests (or Fisher exact tests, when appropriate) for categorical variables and Wilcoxon tests for continuous variables.

[c]Control group: n=125 parents had nonmissing age; intervention group: n=117 had nonmissing age.

[d]Includes Puerto Rican, Cuban, multiple ethnicities, and "other".

[e]GED: general education development.

[f]Includes housekeeper, retired, disabled, temporary employment, student, and "other".

[g]Includes government insurance, insurance through the Veteran Affairs, private insurance, and Medicare.

[h]Includes missing and cohabitation (common law marriage).

[i]Includes don't know, prefer not to respond, missing, and "other".

[j]All other participants indicated "no".

There were no statistically significant differences between intervention and control groups in demographic characteristics. Table S3 in Multimedia Appendix 1 shows more intervention versus control households indicated at least 1 minor aged 5 to 11 years in their household (22.5% vs 19.5%) was vaccinated at baseline, while fewer intervention versus control households reported having at least 1 minor vaccinated at baseline in other age groups (5.6% vs 13.7% for minors 6 mo to 4 y; 27.1% vs 34.1% for minors 12 to 17 years old).

Table S5 in Multimedia Appendix 1 shows demographic characteristics at baseline for those completing the follow-up survey (216/254, 85%) and those who did not (38/254, 15%). Most significant demographic characteristics were associated with follow-up survey completion. Those who completed the follow-up survey were more likely to be unvaccinated, Hispanic, employed, US born, currently married or cohabitating, younger, with a smaller number of children at baseline. They were also more likely to have higher educational attainment and a household income greater than US $50,000. Table S5 also discusses weights considered for TOT estimation related to significant differences between follow-up completers and noncompleters.

DID Model With the ITT Approach

Estimates from the DID model using the ITT approach for the primary outcomes related to vaccination are shown in Table 2 for households with children in each age group individually and all together. Table 2 shows vaccination status estimates for control and intervention groups at baseline and at follow-up in the regression analysis setting, alongside estimates of the difference within and between groups at pre- and postintervention. The control and intervention groups had overall baseline vaccination rates of 25.7% and 15.7%, respectively. At follow-up, the overall vaccination rates were 44.2% and 48.4%, respectively, among these groups. The estimate of interest in Table 2 is the DID estimate shown as "difference."

XSL•FO
RenderX

We found a statistically significant difference between intervention and control groups of 13.3% points (95% CI 0.3%-26.4%; *P*=.04) among children aged 5 to 11 years in change from pre- to postintervention. We did not find a significant DID for other age groups. Given ITT as the conservative estimate of intervention efficacy, we concluded that the intervention had a positive effect on COVID-19 vaccine uptake among the primary target group for this intervention, children aged 5 to 11 years.

Table . Difference-in-difference (DID) regression estimates for primary outcomes with the intention-to-treat (ITT) approach: vaccination status (N=247)[a].

	Baseline, %	1-month follow-up, %	Change Δ (95% CI)	*P* value
6 months-4 years[b]				
Control	13.8	31.4	17.6 (7.0 to 28.2)	.001
Intervention	5.5	27.8	22.3 (11.1 to 33.5)	<.001
Difference	N/A[c]	N/A	4.7 (–10.7 to 20.1)	.55
5 - 11 years[b]				
Control	19.5	34.5	14.9 (7.5 to 22.4)	<.001
Intervention	22.4	50.7	28.3 (17.6 to 38.9)	<.001
Difference	N/A	N/A	13.3 (0.3 to 26.4)	.04
12 - 17 years[b]				
Control	34.1	50	15.9 (5.1 to 26.7)	.004
Intervention	27.1	47.9	20.8 (9.3 to 32.3)	<.001
Difference	N/A	N/A	4.9 (–10.9 to 20.7)	.54
All ages[b]				
Control	26.6	43.6	16.9 (10.3 to 23.6)	<.001
Intervention	21.9	46.3	24.4 (16.8 to 32.1)	<.001
Difference	N/A	N/A	7.5 (–2.6 to 17.7)	.15

[a]Simple (unadjusted) DIDs of vaccination of minors in household. Rates of vaccination are among those participants who indicated having at least 1 minor in the household within that age range and assume that anyone missing a 1-month follow-up had minors who remained unvaccinated. For those who reported a minor was vaccinated at baseline (and thus were not asked the vaccination question at 1 mo), we imputed the "yes" value to the 1-month follow-up, given that the question is not asked again at follow-up for this group. Models are estimated using a generalized linear model and maximum likelihood estimates.

[b]Sample sizes for each age group are as follows: n=105 households for 6 months-4 years; n=158 households for 5-11 years; n=92 households for 12 - 17 years.

[c]N/A: not applicable.

Estimates of the DID model with the ITT approach for secondary outcomes related to vaccine knowledge and trust are shown in Table 3. We found a statistically significant difference between the intervention and control groups of 14.3% points (95% CI 5.0% - 23.7%; *P*=.003) preintervention to postintervention for trust in the governmental approval process to ensure the COVID-19 vaccine is safe for children. We did not find significant effects of the intervention on other secondary outcomes.

XSL•FO
RenderX

Table . Difference-in-difference (DID) regression estimates for secondary outcomes with the intention-to-treat (ITT) approach: vaccine knowledge and trust (N=243)[a].

	Baseline, %	1-month follow-up, %	Change Δ (95% CI)	*P* value
...chances of being hospitalized or dying from COVID-19...[b]				
Control	52.9	65.1	12.2 (3.6 to 20.8)	.005
Intervention	50	65.1	15.0 (5.9 to 24.1)	.001
Difference	N/A[c]	N/A	2.8 (−9.8 to 15.4)	.66
...rate knowledge about COVID-19 vaccine...[d]				
Control	25.2	43.1	17.9 (8.9 to 26.9)	.001
Intervention	30.8	56.7	25.8 (15.9 to 35.8)	<.001
Difference	N/A	N/A	8.0 (–5.5 to 21.4)	.25
...trust governmental approval process... for public...[e]				
Control	81.3	89.4	8.1 (1.9 to 14.3)	.01
Intervention	74.2	90	15.8 (9.3 to 22.4)	<.001
Difference	N/A	N/A	7.7 (–1.3 to 16.7)	.09
...trust governmental approval process... for children...[e]				
Control	78.9	85.4	6.5 (0.7 to 12.4)	.03
Intervention	66.7	87.5	20.8 (13.6 to 28.1)	<.001
Difference	N/A	N/A	14.3 (5.0 to 23.7)	.003

[a]Simple (unadjusted) DID of knowledge and trust. Rates are among those participants who indicated having at least 1 minor in the household and assume that anyone missing a 1-month follow-up did not change their trust and knowledge from baseline. Models are estimated using a generalized linear model and maximum likelihood estimates.

[b]Variable coded equal to 1 if true, equal to 0 if false/unsure.

[c]N/A: not applicable.

[d]Variable coded equal to 1 if adequate/superior knowledge, equal to 0 if no/minimal/basic knowledge.

[e]Variable coded equal to 1 if fully/mostly/somewhat trust, equal to zero if do not trust.

Participant Engagement

Table S6 in Multimedia Appendix 1 shows most participants (228/270, 84% to 228/257, 89%) found the video (provided on Monday) and educational information (provided on Wednesday) very/extremely useful. Materials in weeks 1 and 4 had the highest percentages of participants (236/266 and 220/247, 89%) expressing material was very or extremely useful. When asked about familiarity with that week's provided information, only about half of the participants (108/266, 41% to 132/247, 53%) responded they were very/extremely familiar, with week 1 information showing the lowest percentage of participants (108/266, 41%) responding being very/extremely familiar with information.

Table S7 in Multimedia Appendix 1 summarizes participant engagement data. More than half of the participants (124/265, 47% to 147/252, 58%) clicked the specific link for making a COVID-19 vaccination appointment each week. Among those who completed that week's activity, 89% (219/246) to 90% (227/252) of participants clicked to download the summary infographic to save/share. A large number of participants listened to the audio files from the written information sent on Wednesdays (681/1037, 66%) and the summary infographic sent on Thursdays (562/1037, 54%). Please note that percentages presented in Tables S6 and S7 are based on all participant information collected in our platform, regardless of whether the participant was included in our main analysis of primary and secondary outcomes.

Sensitivity Analysis

A TOT sensitivity analysis with DID model estimation was used. Table S8 in Multimedia Appendix 1 presents estimates from the DID model for the primary vaccination outcome, both for individual age groups and collectively. The analysis revealed an 18.6 percentage point difference (95% CI) in vaccination uptake for children aged 5 to 11 years in the intervention group compared to controls (P=.02), and a 12.1 percentage point increase when considering all age groups (P=.04). Little's MCAR test found evidence that data were not MCAR (P<.01). Weighted estimates from the DID model, which account for statistically significant differences between participants who completed the follow-up survey and those who did not, are provided in Table S9 in Multimedia Appendix 1. These weighted TOT estimates for children aged 5 to 11 years and the overall sample closely align with the unweighted estimates, showing increases of 20.1 percentage points (P=.01) and 12.2 percentage points (P=.04), respectively.

XSL•FO
RenderX

Similarly, TOT analyses of secondary outcomes related to vaccine knowledge and trust (Tables S10 and S11 in Multimedia Appendix 1) are consistent with the ITT results. Specifically, we found a larger increase in trust in the governmental approval process for children post intervention, with TOT unweighted and weighted estimates showing differences of 18.8 and 20.1 percentage points (*P*=.01), respectively. In addition, these tables indicate a significant difference of 10.8 and 10.6 percentage points (*P*=.05) in government approval processes for the public for unweighted and weighted TOT analyses, respectively. The detailed estimates using the TOT approach with the DID model are presented in Tables S8 to S11 in Multimedia Appendix 1.

Discussion

The Role of Community-Based Digital Interventions in Addressing Vaccine Hesitancy

In this study, we examined whether a community-based digital intervention could increase COVID-19 vaccination rates and trust in the governmental approval process to ensure vaccine safety in children aged 5 to 11 years. MVLA 2.0 significantly increased COVID-19 vaccine uptake in children aged 5 to 11 years and trust in the COVID-19 vaccine governmental approval processes for children. The MVLA 2.0 digital intervention demonstrated an effect on vaccine uptake comparable to the magnitude of effect observed in another such COVID-19-specific digital intervention [26], which found an adjusted absolute difference of 11.9 (95% CI 4.5‑19.3) percentage points in their 1-week intervention conducted during emergency room visits. Electronic health record approaches typically used in broader immunization efforts have demonstrated effect sizes between 4.2% (adjusted) and 7% (*P*=.05) [27,28]. These findings offer new insights into the role of community-informed and culturally tailored digital strategies, such as MVLA 2.0, in increasing trust in vaccines and improving vaccination rates. Furthermore, this digital intervention demonstrated the potential for scaling across diverse communities to address other pediatric vaccination gaps.

MVLA 2.0 intervention impact appears to reflect a combination of delivery, cultural tailoring, and user-centered design. Although we cannot rule out selection or reporting bias as contributors to the observed effects, interviews from MVLA 1.0 participants suggest that several components were likely impactful [23]. These included digitally accessible SMS text messaging delivering a web-based curriculum, with written educational content and videos reinforcing weekly material content over time. We also incorporated culturally tailored, real-time feedback to address parental concerns about COVID-19 vaccinations in children of ages 5 to 11 years and counter misinformation, disinformation, and myths prevalent in the Hispanic community. Finally, we applied user-centered design features such as a discussion board and voice-overs for written content to enhance engagement, accessibility, and adherence.

Although some progress for COVID-19 vaccination was made during the pandemic within the Hispanic community, trends in overall vaccinations remain low. Furthermore, national estimates on intent to get a COVID-19 vaccine among Hispanic groups have dropped from 18.6% to 11.4% between September 2024 and January 2025 [29]. Perhaps more concerning is the postpandemic decline in uptake of other childhood vaccines [2]. In California, where this intervention was conducted, statewide estimates on unvaccinated Hispanics have increased nearly 5% in the 4-month period between the end of September 2024 (15.7%) and the end of January 2025 (19.2%) alone [29]. A January 2025 article received media coverage over its analysis of 2019‑2022 vaccination data showing an overall decline in MMR and diphtheria, tetanus, and acellular pertussis vaccination rates across 11 states and, more worrisome, that three of those states had MMR statewide rates below herd immunity and broadly vulnerable to an outbreak [30]. Just weeks after these data were published, also widely reported, Texas experienced its largest measles outbreak in 3 decades, with the first US mortality from measles in a decade and the first pediatric mortality in 2 decades [31].

In addition, there are concerns that many of the impediments to COVID vaccination are similarly involved with the more recent downward trends in vaccination elsewhere, and some initial evidence to suggest these concerns may be warranted [32]. Because the MVLA 2.0 intervention was designed to be digital and culturally tailored, this model may have the potential for scaling across diverse communities to address other pediatric vaccination gaps. Future research is needed to explore the use of digital culturally tailored interventions such as MVLA 2.0 more broadly, as a cost-effective strategy to improve vaccination rates and reduce disparities in undervaccination for children.

Both intervention and control groups showed an increase in knowledge postintervention. Although the magnitude of this increase was not significantly different between groups for most outcomes, this study showed a significant intervention effect for increasing trust regarding the governmental approval process of the COVID-19 vaccine for children. A recent systematic review reported that health concerns, vaccine attributes, and mistrust were primary drivers of COVID-19 vaccine hesitancy, whereas trust and confidence, community and social factors, and demographics and identity were key factors associated with vaccine uptake [33]. Altogether, these broader trends in vaccine hesitancy and the associated causes reinforce the need for additional research into broad-based, easily disseminated multifactorial approaches to improve vaccine confidence.

Approaches, such as the mobile phone–centric format used in MVLA 2.0, may help promote vaccine acceptance, especially given the widespread cellphone and smartphone ownership. According to the Pew Research Center, 98% of Americans own some kind of cellphone and 91% some kind of smartphone. The penetration of ownership/type of usage is relatively high regardless of income (95% and 84%, respectively, those making under $30,000 per year) or ethnicity (97%/87% Black and 99%/93% Hispanic) and even age (94%/79% for those above 65 years of age) [34]. Furthermore, considerable support can be found in the literature for adopting digital, mobile-based interventions to change health behaviors, including immunization among adolescents [35], influenza vaccination among pediatric Hispanic patients in low-income urban areas [36], increasing human papillomavirus vaccine uptake among underserved high-risk populations [37], as well as culturally

XSL•FO
RenderX

specific vaccine education to combat misinformation and improve vaccination behaviors [36,38-46]. Prior studies have recognized the ease of implementation as a benefit of mobile phone–delivered text-based interventions [36,38-53].

SMS text messaging interventions were successful in increasing influenza vaccination rates among Hispanic patients in low-income urban areas [36] and in increasing human papillomavirus vaccine uptake among underserved high-risk populations [37], as well as culturally specific vaccine education to combat misinformation [36,38-46]. Prior studies have recognized the ease of implementation as a benefit of mobile phone–delivered text-based interventions [36,38-53].

Study Limitations and Suggested Future Work

This study has several limitations. We used self-reported vaccination status, which is subject to over- or under-reporting. Parents may have over-reported vaccination due to social desirability bias, particularly after participating in a provaccine educational intervention. The lack of objective verification through medical records or vaccination databases reduces confidence in the primary findings. However, the target population for this intervention may be less likely to provide official records due to concerns with immigration and documentation status. Thus, building trust to ensure the collection of official vaccination records for future digital vaccine acceptance interventions is key.

Lower baseline vaccination rates in the intervention group may overestimate efficacy. Due to a 15% attrition rate and systematic differences between completers and noncompleters, the ITT analysis may be compromised, despite attempts to adjust for these differences using weighting in the TOT analysis. The assumption that vaccinating 1 child in an age group translates to all children in that household being vaccinated for those with missing data could lead to inflated effect sizes. In this study, individual children in a household did not have unique IDs; therefore, we were unable to link responses between baseline and follow-up. Thus, future work should aim to analyze vaccination uptake at the individual level. Furthermore, CIs in our results were wide in relation to our primary (vaccination status of children aged 5-11 years) and secondary outcomes (trust in the government process for the approval of COVID-19 vaccines for children). This variability may suggest that estimates of intervention effect lack precision, likely explained by the small sample size. Thus, future studies warrant a larger sample size to provide a more precise estimate of the efficacy of a digital intervention on vaccine uptake and trust.

Although a strength of this study included using a community-based approach to recruit populations typically “under-represented” in research, this also may limit the generalizability of our findings. First, we primarily recruited participants who had already engaged with community organizations and were willing to participate in vaccine research, which could create selection bias. This population could differ systematically from the broader Hispanic community in terms of health engagement, technology access, and baseline vaccine attitudes. Second, most participants completed the intervention in English, so results may not generalize to households who prefer to communicate in Spanish. Interestingly, during the MVLA 1.0 intervention, 79% of participants completed the intervention in Spanish [22]. This shift in language predominance might be explained by differences in the community partner organizations engaged in MVLA 2.0 and the populations they serve. Third, there may have been a bias in participants willing to be screened for eligibility to participate. Future digital intervention studies should work closely with community organizations to incorporate a screening method where the community organization can confirm whether participants are affiliated with their organization to improve recruitment tracking.

Future directions for vaccine interventions should emphasize the development of culturally tailored community-based digital strategies to build trust in vaccines. This includes actively engaging with community leaders and advocates to ensure that interventions are relevant and resonate with the target population. Addressing misinformation could involve using mobile platforms to provide accurate, accessible information and counteract false narratives in the same modality in which they are being disseminated. In addition, improving vaccination behaviors might be achieved by integrating behavioral insights to identify and overcome specific barriers to vaccination within the community. Scaling these strategies to diverse populations is critical to address the challenge of decreasing vaccination rates.

Conclusions

The findings from this RCT evaluating MVLA 2.0 (a culturally tailored, community-based, mobile phone–delivered intervention) demonstrate higher COVID-19 vaccine uptake among Hispanic children and increased parental trust in the governmental vaccine approval process compared with the control group. This study is innovative in combining community-informed content, real-time culturally tailored feedback, and a fully automated, mobile phone–centric delivery model. It adds to existing literature by testing a culturally tailored digital curriculum in a community setting and showing improvements in both uptake and trust. These findings provide rigorous evidence that scalable, low-cost digital strategies can address vaccine confidence and uptake in populations with persistent gaps in vaccine acceptance. Real-world implications include the portability and adaptability of this approach across diverse communities and settings—urban or rural—and its feasibility and potential for broader applicability and scalability to support public health vaccination campaigns.

Acknowledgments

University of California, Los Angeles, and Pepperdine University provided in-kind support. The authors are thankful to Gael Perez for his help supporting the submission of the manuscript.

XSL•FO
RenderX

Funding

Funding for this project came through the VaxUp Challenge, in which the selection process was conducted jointly by the Los Angeles Department of Public Health and Children's Hospital of Los Angeles. The Possibility Labs through the Digital Transformation Fund also provided supplementary funding. This research was also, in part, funded by the National Institutes of Health (NIH) Agreement OT2HL158287, NIH/National Institute on Aging (NIA) Grant P30-AG021684, and NIH/National Center for Advancing Translational Sciences (NCATS) Grant Number UL1TR001881. The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated and analyzed during this study are not publicly available to protect patient confidentiality. Deidentified data, study protocol, and statistical analysis plan may be accessed upon request to the corresponding author and institutional review board approval.

Authors' Contributions

Conceptualization: LRB, YC-L
Methodology: LRB, AK, RL-C, EA, BW, MM, YC-L
Software: LRB, AK, BW, MM, YC-L
Validation: LRB, AK, RD, KN, YC-L
Writing—original draft preparation: LRB, AK, RD, DM, RL-C, EA, BW, MM, KN, YC-L
Writing—review and editing: LRB, AK, RD, DM, RL-C, EA, BW, MM, KN, YC-L
Visualization: LRB, AK, YC-L
Supervision: LRB, YC-L
Project administration: LRB, DM, BW, MM, YC-L
Funding acquisition: LRB, RD, RL-C, EA, KN, YC-L
All authors have read and agreed to the published version of the manuscript.
We certify that no artificial intelligence tools were used in the writing, editing, or preparation of any portion of this manuscript. All content, including text, figures, and tables, was created by the authors without the use of artificial intelligence.

Conflicts of Interest

YCL reports consulting fees for advisory work from Eli Lilly and Company and Exact Sciences for work unrelated to this study. Other authors have no conflicts to declare.

Multimedia Appendix 1
Examples of the culturally tailored infographics used in the MiVacunaLA 2.0 digital intervention.
[PDF File, 579 KB - jmir_v28i1e78103_app1.pdf]

Checklist 1
CONSORT checklist.
[PDF File, 313 KB - jmir_v28i1e78103_app2.pdf]

References

1. Murthy NC, Zell E, Fast HE, et al. Disparities in first dose COVID-19 vaccination coverage among children 5-11 years of age, United States. Emerg Infect Dis 2022 May;28(5):986-989. [doi: 10.3201/eid2805.220166] [Medline: 35226801]
2. COVID-19 vaccination coverage and intent for vaccination, children 6 months through 17 years, United States. Centers for Disease Control and Prevention (CDC). URL: https://www.cdc.gov/covidvaxview/weekly-dashboard/child-coverage-vaccination.html [accessed 2026-03-11]
3. COVID-19 vaccination data - vaccination status by age and race and ethnicity. California Department of Public Health. URL: https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-Vaccine-Data.aspx#Age [accessed 2026-03-04]
4. Kim L, Whitaker M, O'Halloran A, et al. Hospitalization rates and characteristics of children aged <18 years hospitalized with laboratory-confirmed COVID-19 - COVID-NET, 14 states, March 1-July 25, 2020. MMWR Morb Mortal Wkly Rep 2020 Aug 14;69(32):1081-1088. [doi: 10.15585/mmwr.mm6932e3] [Medline: 32790664]
5. Gonzalez SL, Sarik DA, Dean-Olmsted E, Salyakina D. COVID-19 vaccination knowledge, attitudes, and practices within a majority Hispanic/Latino pediatric healthcare system. J Pediatr Nurs 2025;82:e65-e74. [doi: 10.1016/j.pedn.2025.03.020] [Medline: 40246617]
6. Bateman LB, Hall AG, Anderson WA, et al. Exploring COVID-19 vaccine hesitancy among stakeholders in African American and Latinx communities in the Deep South through the lens of the health belief model. Am J Health Promot 2022 Feb;36(2):288-295. [doi: 10.1177/08901171211045038] [Medline: 34719985]

XSL•FO
RenderX

7. Milo RB, Aguayo C, Chico AR, et al. Understanding the barriers to COVID-19 vaccine among Hispanic/Latinx communities. J Prim Care Community Health 2023;14:21501319231174810. [doi: 10.1177/21501319231174810] [Medline: 37306326]
8. Haro-Ramos AY, Sanchez GR, Barreto MA. Health care discrimination and immigration fears: unpacking COVID-19 vaccine hesitancy in Latino adults. Am J Public Health 2024 Jul;114(S6):S505-S509. [doi: 10.2105/AJPH.2024.307668] [Medline: 39083732]
9. Hammershaimb EA, Cole LD, Liang Y, et al. COVID-19 vaccine acceptance among US parents: a nationally representative survey. J Pediatric Infect Dis Soc 2022 Aug 30;11(8):361-370. [doi: 10.1093/jpids/piac049] [Medline: 35748047]
10. Ataç Ö, Peterson LE, Waters TM. The impact of the COVID-19 pandemic on vaccinations in United States primary care practices. PLoS One 2025;20(6):e0325934. [doi: 10.1371/journal.pone.0325934] [Medline: 40493658]
11. Vaccination coverage and exemptions among kindergartners. Centers for Disease Control and Prevention (CDC). 2025. URL: https://www.cdc.gov/schoolvaxview/data/index.html#cdc_data_surveillance_section_2-new-findings-on-vaccination-coverage-and-exemptions-among-kindergartners-2024-2025-school-year [accessed 2025-09-16]
12. Seither R, Yusuf OB, Dramann D, Calhoun K, Mugerwa-Kasujja A, Knighton CL. Coverage with selected vaccines and exemption from school vaccine requirements among children in kindergarten - United States, 2022-23 school year. MMWR Morb Mortal Wkly Rep 2023 Nov 10;72(45):1217-1224. [doi: 10.15585/mmwr.mm7245a2] [Medline: 37943705]
13. Dong E, Saiyed S, Nearchou A, Okura Y, Gardner LM. Trends in county-level MMR vaccination coverage in children in the United States. JAMA 2025 Aug 26;334(8):730-732. [doi: 10.1001/jama.2025.8952] [Medline: 40455620]
14. Hill HA, Yankey D, Elam-Evans LD, Singleton JA, Sterrett N. Vaccination coverage by age 24 months among children born in 2017 and 2018 - National Immunization Survey-Child, United States, 2018-2020. MMWR Morb Mortal Wkly Rep 2021 Oct 15;70(41):1435-1440. [doi: 10.15585/mmwr.mm7041a1] [Medline: 34648486]
15. Cunniff L, Alyanak E, Fix A, et al. The impact of the COVID-19 pandemic on vaccination uptake in the United States and strategies to recover and improve vaccination rates: a review. Hum Vaccin Immunother 2023 Aug 1;19(2):2246502. [doi: 10.1080/21645515.2023.2246502] [Medline: 37671468]
16. Abdul-Mutakabbir JC, Granillo C, Peteet B, et al. Rapid implementation of a community-academic partnership model to promote COVID-19 vaccine equity within racially and ethnically minoritized communities. Vaccines (Basel) 2022 Aug 20;10(8):1364. [doi: 10.3390/vaccines10081364] [Medline: 36016251]
17. Munoz-Lavanderos C, Oluyomi A, Rosales O, Hernandez N, Mensah-Bonsu N, Badr H. Development, implementation, and evaluation of three outreach events to improve COVID-19 vaccine uptake among racial and ethnic minority communities in Houston, Texas, 2022. Public Health Rep 2024;139(1_suppl):71S-80S. [doi: 10.1177/00333549231213848] [Medline: 38140821]
18. Demeke J, McFadden SM, Dada D, et al. Strategies that promote equity in COVID-19 vaccine uptake for undocumented immigrants: a review. J Community Health 2022 Jun;47(3):554-562. [doi: 10.1007/s10900-022-01063-x] [Medline: 35084639]
19. AuYoung M, Rodriguez Espinosa P, Chen WT, et al. Addressing racial/ethnic inequities in vaccine hesitancy and uptake: lessons learned from the California alliance against COVID-19. J Behav Med 2023 Apr;46(1-2):153-166. [doi: 10.1007/s10865-022-00284-8] [Medline: 35066696]
20. Riggle-van Schagen C, Andrade EL, Chandarana S, et al. Assessing changes in COVID-19 vaccine uptake and intentions among the Brigada Digital Latino social media audience: a repeated cross-sectional study. BMC Digit Health 2024;2(1):72. [doi: 10.1186/s44247-024-00131-7]
21. Chandeying N, Thongseiratch T. Systematic review and meta-analysis comparing educational and reminder digital interventions for promoting HPV vaccination uptake. NPJ Digit Med 2023 Aug 29;6(1):162. [doi: 10.1038/s41746-023-00912-w] [Medline: 37644090]
22. Castellon-Lopez YM, Klomhaus AM, Garcia C, et al. MivacunaLA (MyShotLA): a community-partnered mobile phone intervention to improve COVID-19 vaccination behaviors among low-income, Spanish-speaking, and immigrant Latino parents or caregivers. Vaccines (Basel) 2024 May 8;12(5):511. [doi: 10.3390/vaccines12050511] [Medline: 38793762]
23. Panameno M, Blanco LR, Hernandez AM, et al. Using digital technology to build COVID-19 vaccine confidence: a qualitative study among Latinx parents of children aged 5-11 in under-resourced communities across Los Angeles County. Vaccines (Basel) 2023 May 30;11(6):1042. [doi: 10.3390/vaccines11061042] [Medline: 37376431]
24. LA county COVID-19 vaccine dashboard. County of Los Angeles Public Health. 2025. URL: http://publichealth.lacounty.gov/media/coronavirus/vaccine/vaccine-dashboard.htm [accessed 2026-03-04]
25. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ 2025 Apr 14;389:e081123. [doi: 10.1136/bmj-2024-081123] [Medline: 40228833]
26. Rodriguez RM, Nichol G, Eucker SA, et al. Effect of COVID-19 vaccine messaging platforms in emergency departments on vaccine acceptance and uptake: a cluster randomized clinical trial. JAMA Intern Med 2023 Feb 1;183(2):115-123. [doi: 10.1001/jamainternmed.2022.5909] [Medline: 36574256]
27. Fiks AG, Hunter KF, Localio AR, et al. Impact of electronic health record-based alerts on influenza vaccination for children with asthma. Pediatrics 2009 Jul;124(1):159-169. [doi: 10.1542/peds.2008-2823] [Medline: 19564296]

XSL•FO
RenderX

28. Wijesundara JG, Ito Fukunaga M, Ogarek J, et al. Electronic health record portal messages and interactive voice response calls to improve rates of early season influenza vaccination: randomized controlled trial. J Med Internet Res 2020 Sep 25;22(9):e16373. [doi: 10.2196/16373] [Medline: 32975529]
29. Vaccination uptake, intent, and confidence. Centers for Disease Control and Prevention (CDC). 2025. URL: https://www.cdc.gov/respvaxview/dashboards/vaccination-behavioral-social-drivers.html [accessed 2026-03-04]
30. Moghtaderi A, Callaghan T, Luo Q, et al. Evidence on trends in uptake of childhood vaccines and association with COVID-19 vaccination rates. Vaccine 2025 Jan 25;45:126631. [doi: 10.1016/j.vaccine.2024.126631] [Medline: 39700909]
31. Shastri D, Seitz A. A Texas child who was not vaccinated has died of measles, a first for the US in a decade. AP News. 2025. URL: https://apnews.com/article/measles-outbreak-west-texas-death-rfk-41adc66641e4a56ce2b2677480031ab9 [accessed 2025-02-26]
32. Melani D. Child dies of measles: here's a look at the country's re-emerging disease. University of Colorado Anschutz. 2025. URL: https://news.cuanschutz.edu/news-stories/child-dies-of-measles-heres-a-look-at-the-countrys-re-emerging-disease [accessed 2025-03-03]
33. Nwachukwu G, Rihan A, Nwachukwu E, Uduma N, Elliott KS, Tiruneh YM. Understanding COVID-19 vaccine hesitancy in the United States: a systematic review. Vaccines (Basel) 2024 Jul 6;12(7):747. [doi: 10.3390/vaccines12070747] [Medline: 39066385]
34. Mobile fact sheet. Pew Research Center. URL: https://www.pewresearch.org/internet/fact-sheet/mobile [accessed 2025-02-28]
35. Stockwell MS, Kharbanda EO, Martinez RA, Vargas CY, Vawdrey DK, Camargo S. Effect of a text messaging intervention on influenza vaccination in an urban, low-income pediatric and adolescent population: a randomized controlled trial. JAMA 2012 Apr 25;307(16):1702-1708. [doi: 10.1001/jama.2012.502] [Medline: 22535855]
36. Stockwell MS, Hofstetter AM, DuRivage N, et al. Text message reminders for second dose of influenza vaccine: a randomized controlled trial. Pediatrics 2015 Jan;135(1):e83-e91. [doi: 10.1542/peds.2014-2475] [Medline: 25548329]
37. Sandi YDL, Yang LY, Andarini E, Maryam D, Wu LM. Effectiveness of digital education on human papillomavirus knowledge, vaccination intent and completion rates in adolescents and young adults: a meta-analysis. J Adv Nurs 2025 Apr;81(4):2199-2213. [doi: 10.1111/jan.16575] [Medline: 39465620]
38. Marquez C, Kerkhoff AD, Naso J, et al. A multi-component, community-based strategy to facilitate COVID-19 vaccine uptake among Latinx populations: from theory to practice. PLoS One 2021;16(9):e0257111. [doi: 10.1371/journal.pone.0257111] [Medline: 34543291]
39. Balasuriya L, Santilli A, Morone J, et al. COVID-19 vaccine acceptance and access among Black and Latinx communities. JAMA Netw Open 2021 Oct 1;4(10):e2128575. [doi: 10.1001/jamanetworkopen.2021.28575] [Medline: 34643719]
40. Castellon-Lopez YM, Carson SL, Mansfield L, et al. "The System Doesn't Let Us in"-a call for inclusive COVID-19 vaccine outreach rooted in Los Angeles Latinos' experience of pandemic hardships and inequities. Int J Environ Res Public Health 2022 May 10;19(10):5785. [doi: 10.3390/ijerph19105785] [Medline: 35627322]
41. Cialdini RB, Goldstein NJ. Social influence: compliance and conformity. Annu Rev Psychol 2004;55:591-621. [doi: 10.1146/annurev.psych.55.090902.142015] [Medline: 14744228]
42. Garcia J, Vargas N, de la Torre C, Magana Alvarez M, Clark JL. Engaging Latino families about COVID-19 vaccines: a qualitative study conducted in Oregon, USA. Health Educ Behav 2021 Dec;48(6):747-757. [doi: 10.1177/10901981211045937] [Medline: 34596462]
43. Kapteyn A, Angrisani M, Bennett D, Bruin W, Darling J, Gutsche T, et al. Tracking the effect of the COVID-19 pandemic on the lives of American households. Surv Res Methods 2020;14(2):179-186. [doi: 10.18148/srm/2020.v14i2.7737]
44. Lustria MLA, Noar SM, Cortese J, Van Stee SK, Glueckauf RL, Lee J. A meta-analysis of web-delivered tailored health behavior change interventions. J Health Commun 2013;18(9):1039-1069. [doi: 10.1080/10810730.2013.768727] [Medline: 23750972]
45. Webb Hooper M, Nápoles AM, Pérez-Stable EJ. No populations left behind: vaccine hesitancy and equitable diffusion of effective COVID-19 vaccines. J Gen Intern Med 2021 Jul;36(7):2130-2133. [doi: 10.1007/s11606-021-06698-5] [Medline: 33754319]
46. Benefits of getting vaccinated. Centers for Disease Control and Prevention. 2025. URL: https://www.cdc.gov/covid/vaccines/benefits.html [accessed 2026-03-04]
47. Cáceres NA, Shirazipour CH, Herrera E, Figueiredo JC, Salvy SJ. Exploring Latino Promotores/a de Salud (Community Health Workers) knowledge, attitudes, and perceptions of COVID-19 vaccines. SSM Qual Res Health 2022 Dec;2:100033. [doi: 10.1016/j.ssmqr.2021.100033] [Medline: 34904136]
48. Carson SL, Casillas A, Castellon-Lopez Y, et al. COVID-19 vaccine decision-making factors in racial and ethnic minority communities in Los Angeles, California. JAMA Netw Open 2021 Sep 1;4(9):e2127582. [doi: 10.1001/jamanetworkopen.2021.27582] [Medline: 34591103]
49. Gargano LM, Herbert NL, Painter JE, et al. Impact of a physician recommendation and parental immunization attitudes on receipt or intention to receive adolescent vaccines. Hum Vaccin Immunother 2013 Dec;9(12):2627-2633. [doi: 10.4161/hv.25823] [Medline: 23883781]
50. Lee HY, Koopmeiners JS, McHugh J, Raveis VH, Ahluwalia JS. mHealth pilot study: text messaging intervention to promote HPV vaccination. Am J Health Behav 2016 Jan;40(1):67-76. [doi: 10.5993/AJHB.40.1.8] [Medline: 26685815]

XSL•FO
RenderX

51. McAboy K. UCLA-led coalition works with minority communities on COVID-19 and vaccine education. FOX 11 Los Angeles. 2021. URL: https://www.foxla.com/news/ucla-led-coalition-works-with-minority-communities-on-covid-19-and-vaccine-education [accessed 2026-03-04]
52. Silverman-Lloyd LG, Dominguez Cortez J, Godage SK, et al. Immigrant Latino parents demonstrated high interactivity with pediatric primary care text messaging intervention. Mhealth 2020;6:45. [doi: 10.21037/mhealth.2020.01.06] [Medline: 33209916]
53. Szilagyi PG, Thomas K, Shah MD, et al. The role of trust in the likelihood of receiving a COVID-19 vaccine: results from a national survey. Prev Med 2021 Dec;153:106727. [doi: 10.1016/j.ypmed.2021.106727] [Medline: 34280405]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
DID: difference-in-difference
HPV: human papillomavirus
ITT: intention-to-treat
MCAR: Missing Completely at Random
MMR: measles-mumps-rubella
MVLA: MiVacunaLA/MyShotLA
RCT: randomized controlled trial
TOT: treatment-on-the-treated
UAS: Understanding America Survey
UCLA: University of California, Los Angeles

Edited by S Brini; submitted 26.May.2025; peer-reviewed by L Jain, SW Lee; accepted 10.Dec.2025; published 30.Mar.2026.

Please cite as:
Blanco LR, Klomhaus A, Dudovitz R, Marquez D, Lopez-Chang R, Aleman E, Weerman B, Moldoff M, Norris K, Castellon-Lopez Y Digital Intervention (MiVacunaLA 2.0) to Promote COVID-19 Vaccine Acceptance Among Hispanic Children: Community-Based Randomized Controlled Trial
J Med Internet Res 2026;28:e78103
URL: https://www.jmir.org/2026/1/e78103
doi:10.2196/78103

© Luisa R Blanco, Alexandra Klomhaus, Rebecca Dudovitz, Denise Marquez, Ray Lopez-Chang, Evelyn Aleman, Bas Weerman, Michael Moldoff, Keith Norris, Yelba Castellon-Lopez. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Designing a Carbohydrate Counting App for Young Adults With Type 1 Diabetes: Usability Testing Interview Study

Asmaa Housni[1], RD, PhD; Aidan Shulkin[2], MD; Alexandra Katz[3], MSc, MD; Giuliana Giannini[1], RD; Amélie Roy-Fleming[1], RD, CDE, MSc; Meranda Nakhla[4], MSc, MD; Courtney South[1], RD, MSc; Anne-Sophie Brazeau[1], RD, PhD

[1]School of Human Nutrition, McGill University, 21,111 Lakeshore drive, Montreal, QC, Canada

[2]Faculté de médecine, Université de Montréal, Montreal, QC, Canada

[3]Internal Medicine, McGill University, Montreal, QC, Canada

[4]Division of Endocrinology, Montreal Children's Hospital, Montreal, QC, Canada

Corresponding Author:
Anne-Sophie Brazeau, RD, PhD
School of Human Nutrition, McGill University, 21,111 Lakeshore drive, Montreal, QC, Canada

Abstract

Background: Carbohydrate counting (CC) assists people with type 1 diabetes (T1D) adjust mealtime insulin doses; however, it is often burdensome. Mobile apps can simplify this process by automating carbohydrate estimation and insulin calculations, yet no comprehensive solution currently combines photo-based carbohydrate recognition with an integrated bolus calculator.

Objective: This study aimed to identify user-informed design principles from usability testing interviews to optimize a novel app supporting young adults with T1D in CC and insulin dosing.

Methods: We conducted 4 iterative rounds of usability testing interviews, each with 3 to 5 participants, using a think-aloud protocol to evaluate how easily and effectively users interacted with the app and to identify areas for improvement. Interviews were analyzed qualitatively to derive main design principles, and findings from each round informed the refinement of the app prior to subsequent testing.

Results: A total of 18 participants completed the usability testing (median age of 23, IQR 19-24 y and diabetes duration of 9, IQR 6-12 y; n=12, 66.7% young women). Thematic analysis highlighted that a person-centered design that prioritizes the lived experiences of youth with T1D was essential to position the app as a self-management support system, beyond a clinical tool. Personalization was central, including customizable treatment profiles, tailored dashboard metrics, esthetic preferences, and artificial intelligence–driven recommendations based on personal trends. Early usability barriers revealed the need for intuitive navigation, streamlined multistep processes, and clear guidance for data entry and interpretation. Participants valued culturally inclusive content and familiar terminology to enhance accessibility and engagement. Users perceived strong potential for the app to centralize diabetes management tasks, integrate contextual factors (eg, exercise, diet, and timing of insulin) with glucose data, generate sharable reports to facilitate patient-practitioner communication, and strengthen self-efficacy through personalized trend analysis. Concerns about over-reliance on automation underscored the necessity of transparent data verification and user override options to maintain trust in insulin dosing decisions.

Conclusions: Iterative usability testing highlighted the importance of balancing automation with user control, personalization, and contextual understanding of personal trends, as key design principles to enhance engagement and the apps' relevance as a self-management tool. Incorporating these features into a CC and insulin-dosing app could improve self-efficacy in youth living with T1D.

(J Med Internet Res 2026;28:e86024) doi:10.2196/86024

KEYWORDS

type 1 diabetes; carbohydrate counting; mobile apps; youth; usability testing

Introduction

Type 1 diabetes (T1D) is an autoimmune condition resulting in pancreatic β cell destruction [1]. People with T1D need daily administration of insulin to regulate glucose levels and prevent acute events and long-term health risks [2,3]. Intensive insulin therapy through carbohydrate counting (CC) is the recommended treatment for individuals living with T1D due to its association with lowered glycated hemoglobin (HbA_{1c}) levels and the prevention of diabetes-associated complications [4-6]. Moreover, CC offers dietary flexibility by allowing individuals to tailor their mealtime insulin based on the carbohydrate content of

XSL•FO
RenderX

their meals [7]. However, CC adherence can be burdensome and challenging as it requires tracking intake, reading and interpreting nutrition labels, and weighing portions [8,9]. Additionally, calculations are multifactorial and prone to error, due to the complexities of combining insulin sensitivity factor, insulin-to-carbohydrate ratio, and glucose targets [10].

The transition from childhood to adulthood introduces an additional barrier to CC, namely due to social and developmental changes commonly observed in this age group [11-13], and CC tasks transferred from parent to child. A study by Gurmani et al [14] reported that 43.6% of adolescents inaccurately calculated carbohydrates by 10 to 20 g from the actual amount. Incorrectly identified meals either lacked nutrition labels or required fiber subtraction to calculate carbohydrates accurately, underscoring the difficulty young adults face estimating carbohydrates in unpackaged foods [14]. This contributes to the broader challenge of achieving glycemic stability during adolescence and young adulthood, as evidenced by the higher HbA_{1c} levels observed in these age groups compared to others [13]. Given the various benefits of CC, strategies are needed to promote its proper use among youth.

A promising solution lies in mobile health (mHealth), a branch of telemedicine that uses technology such as mobile devices and tablets to improve health care services for both patients and health care providers [15]. In the context of diabetes management, mHealth can automate CC as a novel care approach for individuals living with T1D. In fact, the efficacy of mHealth, that is, the use of smartphone or tablet apps in improving diabetes self-management among adults, is well established [16-18]. Despite the large variety of apps able to offer photo recognition powered by artificial intelligence (AI) to quantify meal composition, most are not designed for diabetes management, few have been tested in a clinical setting, and even fewer limit their search to validated food databases only [19]. Additionally, while there are separate apps available for bolus insulin calculations, few have undergone usability testing or been developed using user-centered design approach, raising concerns about the safety of insulin calculation apps [20]. A needs assessment survey was conducted across Canada to identify the barriers and needs related to CC in people with T1D, including youth. The survey emphasized that existing mobile apps currently used for CC often fail to meet the needs of this population. A novel app incorporating features such as photo recognition, reliable nutrient values, and personalized bolus calculations could help reduce the burden of CC [21].

Scientific literature has demonstrated a need for a comprehensive app tailored for individuals living with T1D that integrates accurate CC and bolus insulin calculations. Thus, we collaborated with Ikigai Development, Inc, a company specializing in app development, to design the Petit Cactus app, which involved users in the development process to promote a user-centered design [22]. The Petit Cactus app was designed to include four main features: (1) the ability to calculate the number of carbohydrates in meals either through AI-powered photo recognition or through manual selection of food items, with data derived from validated Canadian food databases; (2) bolus insulin dose calculation that incorporates multiple variables (glucose levels, insulin-to-carbohydrates ratio, and correction factors), and integrates continuous glucose monitoring (CGM) data to help users visualize how logged events (ie, meals, stress, sick days, physical activities) affect glucose levels; (3) an electronic food journal for reviewing logged meals and event history; and (4) a connecting portal to a health care provider interface.

A primary objective was to identify user-informed design principles that could guide the development of this and future digital tools for supporting self-management in adolescents and young adults with T1D. Building on these principles, a broader objective was to collaboratively optimize the app's usability to ensure it effectively meets the needs of this population.

Methods

Study Design

The study adopts a user-centered design framework, using iterative usability testing with young adults living with T1D [23]. User feedback was systematically collected and integrated into successive stages of app development to inform design refinements and enhance overall usability. We conducted usability testing to evaluate how users interacted with the Petit Cactus app and to identify usability issues, preferences, and unmet needs. Participants completed task-based interactions with the prototype while verbalizing their thoughts using a think-aloud approach, following a semistructured interview. Usability testing with young adults involved 4 iterative rounds, each with a new group of 3 to 5 participants testing an updated version of the Petit Cactus app refined based on feedback from previous rounds. The usability testing interviews were followed by 1 week of independent use and completion of a feedback questionnaire to capture additional insights (Figure 1). Data from observations and questionnaires were analyzed using inductive qualitative thematic analysis to derive user needs and design insights.

XSL•FO
RenderX

Figure 1. Methodology for conducting iterative usability testing. T1D: type 1 diabetes.

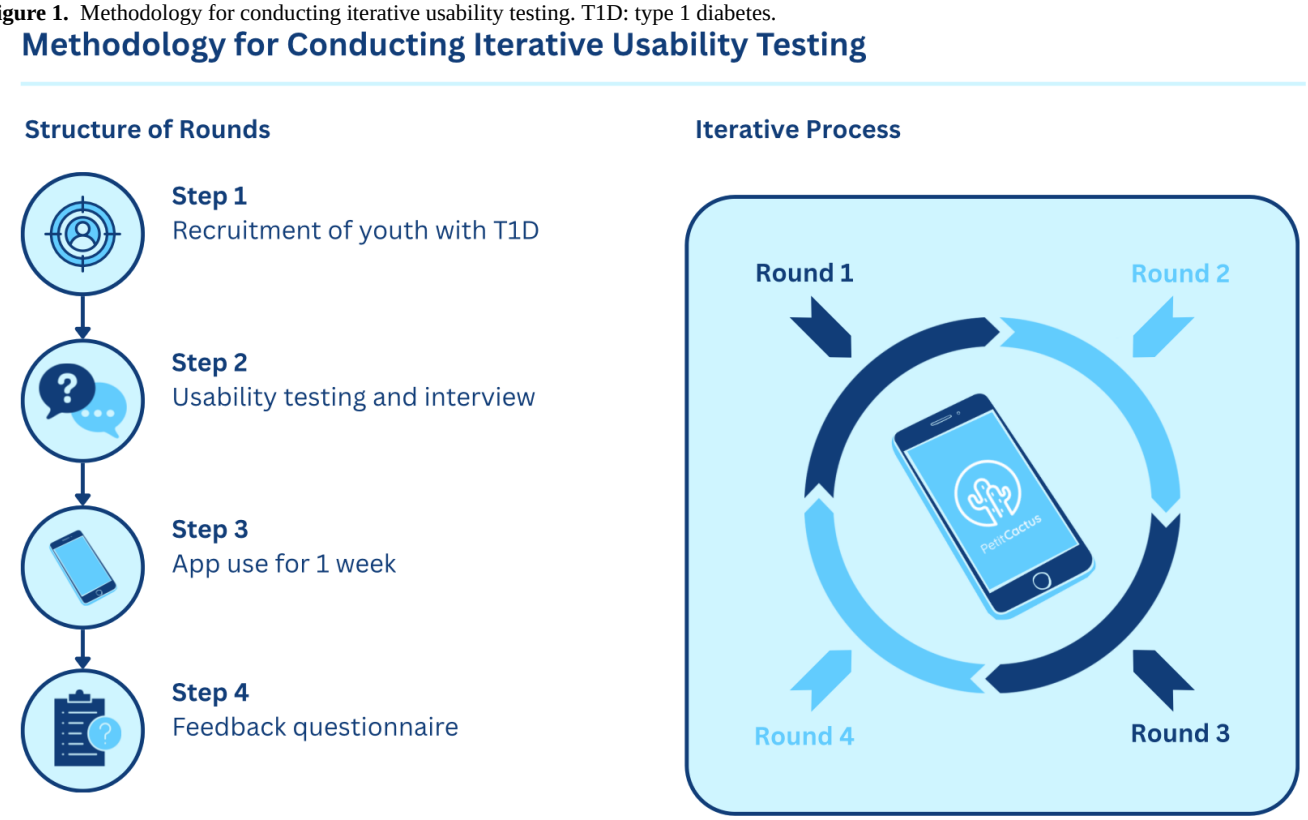

Eligibility Criteria

Young adults aged 14 to 24 years, with a T1D diagnosis of more than 1 year, using intensive insulin therapy (ie, basal-bolus injections or insulin pump therapy), and residing in Canada, were eligible to participate in the usability testing.

Participant Recruitment and Consent

Usability testing typically requires an average of 5 participants, as it was found to be generally sufficient in identifying significant issues for each version [23]. Recruitment was conducted using convenience sampling from BETTER (Behaviors, Therapies, Technologies and Hypoglycemic Risk in Type 1 Diabetes Registry), a volunteer registry of individuals living with T1D in Canada [24], Canadian diabetes organizations, social media, and word-of-mouth. Prior to enrollment in the study, each participant provided an informed consent and completed a demographic and background questionnaire on LimeSurvey (LimeSurvey GmbH) hosted on secure servers.

Ethical Considerations

The study was approved by McGill University Research Ethics Board (#23-04-018). Written informed consent was obtained before participation. Adults aged 18 years or older provided their own consent, and parents or legal guardians provided consent for participants aged 14 to 17 years. All data collected were deidentified to ensure privacy and confidentiality and were hosted on secure McGill servers. Upon completion of the usability testing interview and questionnaire, participants were provided a CAD $50 (US $36.44) gift card of their choice from a selection of retailers as a compensation for their time and participation in the study. The study was conducted in accordance with the principles of the Declaration of Helsinki for research involving humans and/or human data.

Procedures

Following consent, participants were contacted by email to schedule a usability interview. Subsequently, a link to a Microsoft Teams meeting was sent. One or two days preceding the interview, participants received a user guide to the app with access and log-in information. Interviews were 60 to 90 minutes in length and included the participant, the interviewer, and an observer who took notes to complement verbatim transcripts, when necessary. Prior to commencing the interview, participants were reminded of the objective of the usability testing and their role in the co-design process: to provide feedback on their impression of the app and necessary changes required to improve its functionality. Each session consisted of two components: (1) task-based usability testing and (2) a semistructured interview. During the usability testing component, participants were asked to interact with the app in a sequential order, starting with completing their user profile and dashboard, and then later trialing core features (ie, meal entry and the bolus insulin calculator; Multimedia Appendix 1). Using a think-aloud approach, participants verbalized their thoughts while completing predefined tasks. Probing questions from the usability guide were used as needed to clarify navigation challenges, functionality, design, and ease of use. Following task completion, a semistructured interview was conducted to explore participants' broader perceptions of the app, expectations, and experiences using the app. Following the interview, participants used the app for one week and were then requested to complete a final questionnaire including 8 open-ended questions. The questionnaire was developed by the research team to better understand potential barriers to app use,

XSL•FO
RenderX

the usability of core features (ie, the food journal, meal entry and CC, readability of glucose graphs, desired support from health care providers), motivations for trying the app, and suggestions to enhance engagement (Multimedia Appendix 2). The open-ended format was selected to allow participants to elaborate on their experiences and identify unanticipated issues.

Data Collection

The interviews were recorded and transcribed using Microsoft Teams. Three members of the research team (AH, AS, and GG) verified the accuracy of the transcripts and anonymized all identifiable information. The transcripts were then translated from French to English using a forward-backward translation method [25], when necessary by AH and verified by GG and AS, all of whom are fully bilingual.

Data Analysis

The transcripts were open coded independently by 3 researchers (AH, AS, and GG), using MAXQDA (version 24; VERBI GmbH), to develop a preliminary code book [26]. Interviews were conducted in 4 rounds, with thematic saturation reached within each round, defined as the point at which no new ideas emerged and no new codes were identified, and overall saturation achieved by the fourth round. Two researchers then independently coded all transcripts using the finalized codebook. To ensure consistency, 2 transcripts were initially co-coded, and discrepancies were discussed until consensus was reached, leading to refinement of code definitions. The remaining transcripts were coded independently, with regular meetings held to compare coding decisions and resolve any discrepancies through discussion. An inductive thematic analysis was conducted to interpret the data [27]. After transcripts were first coded, codes were then iteratively grouped and refined to identify themes based on patterns and similarities across the dataset.

Results

Participants

Twenty participants were recruited and provided consent. However, 2 participants were excluded due to not meeting the inclusion criteria for age and T1D diagnosis, despite self-reporting eligibility in the initial online survey. A total of 18 eligible participants completed the interview and the questionnaire. About two-thirds of the participants identified as women (n=12, 66.7%), and Caucasian (n=11, 61.1%). The median (IQR) for age was 23 (19-24) years and for diabetes duration was 9 (6-12) years. Almost all participants (n=17, 94.4%) used CGM systems. The majority used an insulin pump to administer insulin (n=14, 77.8%), and 12 (66.7%) used an automated insulin delivery (AID) system. Participants' characteristics are outlined in Table 1.

XSL•FO
RenderX

Table . Participant characteristics.

Characteristics	Value
Demographics (N=18)	
Age (y), median (IQR)	23 (19-24)
Gender (women), n (%)	12 (66.7)
Country of birth (Canada), n (%)	15 (83.3)
Ethnicity[a], n (%)	
Arab	1 (5.6)
Black (African, Afro-American, and Caribbean)	3 (16.7)
Indigenous (First Nation)	1 (5.6)
White or Caucasian	11 (61.1)
Currently studying[b], n (%)	13 (72.2)
Highest level of education acquired, n(%)	
High school level	7 (38.9)
CEGEP[d], vocational or community college	6 (33.3)
University level	5 (27.8)
Medical history and diabetes management (N=18)	
Diabetes duration (y), median (IQR)	9 (6-12)
Had at least 1 diabetes-related consultation in the last year with, n (%)	
Medical specialist (endocrinologist, pediatrician, or internist)	17 (94.4)
Family physician or general practitioner	15 (83.3)
Registered dietitian	12 (66.7)
Registered nurse	12 (66.7)
Use continuous glucose monitoring systems (N=18), n (%)	17 (94.4)
Dexcom G6	10 (55.6)
Freestyle Libre	1 (5.6)
Freestyle Libre 2	1 (5.6)
Guardian Sensor 3	2 (11.1)
Guardian Sensor 4	3 (16.7)
Use insulin pens or syringes exclusively (N=18), n (%)	4 (22.2)
Use an insulin pump (N=18), n (%)	14 (77.8)
Use of pump features in the last week[c] (n=14), n (%)	
Use an integrated bolus wizard or calculator function	14 (100)
Use an automated insulin pump	12 (86)

[a]Two participants do not know their ethnicities.

[b]One participant preferred not to answer.

[c]Proportion among the 14 pump users only.

[d]CEGEP: Collège d'Enseignement Général et Professionnel (general and professional college).

Critical Issues and User Feedback Across Rounds: Thematic Analysis

Overview

Data from the usability testing of the CC app with young adults with T1D was analyzed using thematic analysis through which usability issues and user preferences identified were coded and iteratively grouped based on shared patterns across 4 rounds. This process resulted in 7 themes that represent recurring usability priorities and corresponding design principles. The codes and emerging themes, representing distinct patterns across all rounds of the dataset, are presented in Figure 2. A summary table of all themes, associated codes, and illustrative participant quotes is provided in Multimedia Appendix 3. Evaluations and

XSL•FO
RenderX

suggestions for the app's features were compiled to illustrate how the features evolved across the usability testing rounds (Table 2).

Figure 2. Themes and codes identified through usability testing of a type 1 diabetes (T1D) management app with youth.

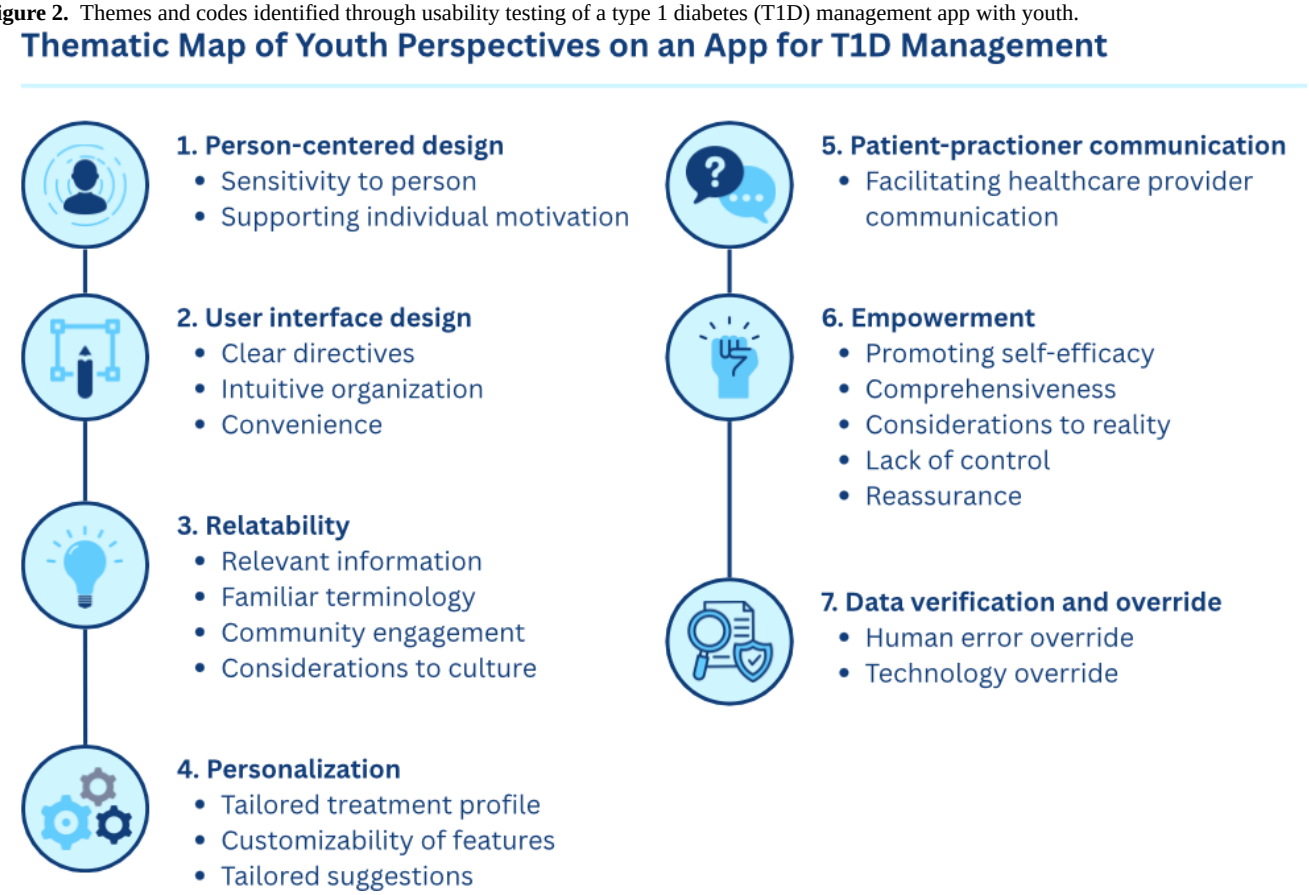

XSL•FO
RenderX

Table . Feature development through usability testing rounds.

Features	Round 1	Round 2	Round 3	Round 4
Photo recognition for carbohydrate suggestions	The photo capture feature for carbohydrate estimations was presented unclearly, leading some participants to mistake it for a tool to simply add pictures of meals.	User feedback was incorporated to improve the presentation of the photo recognition feature, which was clarified by adding prompts to explain its function. Photo uploading remained slow.	Ease of use was enhanced by adding simple options to edit food items incorrectly recognized by AI[a].	The algorithm was optimized for accuracy, with faster photo uploading.
Meal logging	Meal logging was designed with step-by-step navigation instead of showing everything on one page. However, it was overly complex, unintuitive, and too lengthy. The food database also had issues with irrelevant items and inaccurate search results.	The food database was subsequently improved by adding a better search function that presents key food items first, while retaining a large database.	The next iteration addressed this by displaying everything on the same page with streamlined manual entry and dropdown features. The food database was refined to prioritize commonly used items, but estimating quantities remained challenging.	The final iteration enhanced functionality by allowing users to copy or edit previously logged meals, offering more units of measure and options to determine food quantities. It also included options to add notes or assign meal tags (eg, high fat, high protein, pre or post-workout, or injection timing).
Event logging	The feature was not included in this round, but users suggested it to track events other than meals that could impact glucose levels.	Initial concept presented.	The full feature was added, allowing users to track various events beyond meals, such as glucose and insulin (independent of meals), and physical activity.	More events were added, such as the menstrual cycle, medication, and a customizable event, to enable more comprehensive event tracking.
Insulin dosing calculator	Basic model for dosing recommendations.	No changes were made.	The feature now presents an added safety measure. Users are shown their glucose levels and ratios and must confirm entries before calculations.	No changes were made.
Dashboard	The dashboard featured a basic design with limited functionality, presenting statistics on the number of meals, units of insulin, and glucose levels, all compiled over the course of the day.	User feedback was incorporated, resulting in a clearer visual presentation and the addition of adjustable timeframes for compiling statistics.	The final design included a personalized data display with the addition of a graph to show glucose levels alongside logged events (meals, activity, stress, and others).	When the CGM[b] device is connected, glucose levels are displayed in real time, and the cactus avatar dynamically changes to reflect these levels.
Profile settings	Basic settings included name, age, date of birth, gender, target glucose levels, insulin modality, glucose monitoring preferences, and app language.	Settings were expanded to enable users to enter multiple correction factors, input their basal insulin regimen, as well as the option for users to remove information irrelevant to their diabetes management.	Further refinements were made, including the option to add a profile picture or an avatar.	No changes were made.
Connection to a health care provider portal	Not available. Participants recommended a direct method to share data with the health care team.	Initial concept of provider connection presented.	The prototype was tested with basic integration, including an option to link the account with a health care team portal.	No changes were made.
AI chatbot	The AI chatbot was presented unclearly, often mistaken for a help center.	Prompts and examples of queries were added to clarify its use.	In the final iteration, the AI was integrated to provide personalized feedback based on previous discussions.	No changes were made.

[a]AI: artificial intelligence.

[b]CGM: continuous glucose monitoring.

XSL•FO
RenderX

Theme 1: Person-Centered Design

This theme reflects the importance of designing the app with youth living with T1D in mind, prioritizing their perspectives and lived experiences rather than focusing solely on what may be deemed important from a clinician-centered perspective. Participants consistently emphasized the need for the app to acknowledge the person living with T1D, not only the condition it aims to support. This theme evolved across the 4 rounds to reflect a progression and growing interest in emotional well-being and sensitivity to stigma within the app's design.

Initially, participants highlighted the importance of prioritizing personal profile settings (eg, name, gender, and date of birth) before treatment-related parameters during onboarding. This approach was appreciated for making the app feel more personal.

In later rounds, participants further distinguished information relevant to health care providers from that intended for users with T1D. While clinical parameters such as blood glucose targets were considered essential during medical appointments, participants expressed a desire not to have these metrics displayed at all times. Suggestions included the ability to customize dashboard displays, such as hiding A_{1c} levels, which some found stress-inducing. They also recommended avoiding design elements that could unintentionally reinforce negative perceptions (eg, a thumbs down for late insulin administration) and replacing red and green indicators with neutral colors to classify food choices:

> *Especially when you're a young teenager and you see your A_{1c}'s high, it can stress you out. Maybe it could be an option to hide it on your dashboard.* [Woman, 20 years, 13 years living with T1D]
>
> *Say for example I had a bad few days I wouldn't want to remind myself of it all the time.* [Man, 24 years, 9 years living with T1D]

The final round added features to support individual motivation. For instance, participants suggested integrating challenges and objectives that align with users' preferences for managing their diabetes, focusing not just on what they need to do but on how they want to experience and engage with the process. This also included offering peer support and the option to send encouragement messages to others within the app, if they consent to it.

Theme 2: User Interface Design

During the first round, users identified foundational problems. They reported significant challenges with navigation, struggling to locate key elements such as the photo recognition feature, the dashboard, the settings and treatment profile, and the help menus. There was a lack of clear directives, leaving users uncertain about what needed to be entered and unclear about the descriptions of what was displayed. Prompts, descriptors, and visual cues were suggested to optimize the user's comprehension of entry requirements.

By round 2, navigation improvements reduced these challenges; however, multistep processes remained cumbersome. For instance, meal logging required navigating through multiple pages (first to confirm food items recognized by the AI algorithm, then confirm the detected quantities, and finally to confirm bolus calculations), whereas users suggested a streamlined approach using a list or drop-down menu in the same page. Another important aspect is intuitive organization, ensuring that the order of information display is practical. For example, users instinctively clicked on their profile picture to navigate back to their profile.

In later rounds, critical navigation issues were resolved, and only minor refinements were suggested to improve convenience, such as faster photo upload and quicker analysis of the photos. Participants also highlighted the value of automatic logging of meals in the food journal, when a photo is used to estimate carbohydrate content, along with glucose trends and insulin dosing. They further requested simple data extraction for the food journal and the ability to filter and sort information efficiently. This was further reinforced in the feedback questionnaire (Multimedia Appendix 4).

Theme 3: Relatability

This theme focused on enhancing relatability by providing relevant and accessible information while minimizing complexity. In round 1, participants emphasized the need to streamline the food journal feature by avoiding overly detailed descriptions in the database and instead prioritizing a simple, comprehensive list of commonly used items. The importance of using familiar terminology was highlighted to improve user comprehension and align with user preferences (eg, correction factor vs insulin sensitivity factor, 1 cup vs 250 ml).

In later rounds, fostering community engagement became a key focus. Participants proposed adding an in-app peer messaging feature, and a *Frequently Asked Questions* section to promote peer feedback and provide actionable insights. Additionally, participants stressed the importance of cultural inclusivity by ensuring that the app's content, such as food items in the database, is relevant and accessible to users from diverse backgrounds:

> *I think that the different cultures and cultural foods might be something to take into consideration. [...] We really need to make sure there are many different people and many different cultures represented.* [Woman, 22 years, 11 years living with T1D]

Theme 4: Personalization

In earlier rounds, participants emphasized the importance of tailoring the treatment profile to individual needs, including options for CGM connection, weight tracking, injection reminders, and specifying insulin treatment types (eg, pen or injections vs pump). This was particularly relevant when discussing the dashboard and the types of statistics users wanted displayed, depending on their insulin delivery methods and glucose level monitoring preferences. Beyond diabetes management, customizability of the app itself is important. Participants expressed a desire to personalize its esthetics, such as colors and design, as well as practical features like glucose units, custom tags, data visualization, and the organization of the features to align with their individual routines and preferences.

XSL•FO
RenderX

By later rounds, participants suggested the incorporation of AI technology to provide tailored diabetes management recommendations based on user trends and status:

> *If, say you've had too many [hyperglycemic episodes] and it's dinner time. [...] Maybe the AI could give suggestions based on your statistics for the day.* [Man, 18 years, 12 years living with T1D]

Theme 5: Patient-Practitioner Communication

Participants highlighted the potential for the app to enhance patient-practitioner communication, facilitating more effective diabetes management. Initially, participants reported that these features could empower users to take charge of their management during appointments with their health care team, fostering more productive discussions:

> *I went to see the dietitian, and I had to provide a dietary recall, but trying to remember everything I'd eaten, it's definitely more tedious. But to see that [the food journal] is accessible, to have it on hand all the time, and to be able to enter it as I go along, I find it really practical.* [Woman, 19 years, 10 years living with T1D]

By later rounds, the emphasis shifted to refining and expanding these features. Participants suggested generating downloadable reports that combine dietary data, automatically generated from photos taken with the app, with glucose levels and insulin doses, to facilitate sharing with health care professionals and supporting more informed decisions.

Theme 6: Empowerment

The theme of empowerment focuses on how the app can enhance users' confidence and ability to manage T1D effectively, while addressing concerns about over-reliance on technology. In earlier rounds, participants emphasized how a CC app could empower young adults with T1D by promoting self-efficacy. The app's ability to centralize tools and information into a single platform was highlighted as essential for simplifying diabetes management. Features like long-term tracking and trend analysis allowed users to understand factors influencing glycemia, such as meals, activity, stress, and timing of insulin injections:

> *If I look at my graph later and see that one day I was trending high all day, but then I look back at the food journal and remember what I ate, I'll understand better.* [Woman, 19 years, 10 years since diagnosis]

Additionally, the app reassured users by validating daily behaviors, even those deviating from guidelines, providing a realistic and supportive approach to condition management:

> *I really liked [the feature of inputting insulin before or after a meal]. I sometimes forget to give myself insulin before I eat. I think it's important to see the trend.* [Man, 18 years, 12 years living with T1D]

In later rounds, participants expressed appreciation for these empowering features but raised concerns about the potential trade-off between reliance on technology and their sense of control. Some users noted skepticism regarding the accuracy of the app's recommendations, which they felt could undermine their self-efficacy.

Discussions emphasized the importance of building user trust to address these concerns. Participants suggested that transparent mechanisms to verify the app's outputs and ensure accuracy would help mitigate doubts. This evolution across rounds demonstrates the balance needed between empowering users and maintaining their confidence in the app as a reliable support tool for diabetes management. This is further highlighted in the feedback questionnaire, where participants expressed reluctance to rely on AI-driven photo recognition if the carbohydrate content of meals is already known (Multimedia Appendix 4).

Theme 7: Data Verification and Override

Participants highlighted the importance of data verification and override capabilities within the app to ensure accurate diabetes management. In earlier rounds, the focus was on addressing critical issues, such as implementing mechanisms to override technological errors and the ability to accept or reject automated suggestions made by the app. They also stressed the importance of addressing human error by incorporating prompts that encourage users to verify and correct data entries to enhance reliability.

By later rounds, the emphasis shifted to refining these features, including streamlining the verification process to avoid unnecessary disruptions during app use. Participants suggested simplifying error notifications to ensure they were clear but not intrusive and incorporating an option to review and confirm CGM readings. These minor improvements aimed to strike a balance between maintaining user control and creating a seamless experience.

Discussion

Principal Results

The usability testing revealed 7 major themes highlighting key design principles. These themes focused on young adults' empowerment, enhancing patient-practitioner communication, data verification and override for accuracy in diabetes management, an intuitive user interface design, and the relatability of the information provided. While both the person-centered design and personalization themes reflect participants' desire for individual relevance, they differ in scope and intent. Person-centered design represents a broader guiding principle emphasizing empathy and how the app acknowledges the person behind the condition. In contrast, personalization refers to the practical and technological implementation of this principle ie, how the app's content, settings, and recommendations adapt to the user's specific needs and data.

Comparison With Prior Work

Clinical guidelines recommend adjusting meal-time insulin dose to manage glucose variability due to different meal compositions [28,29]. However, there is no consensus as to which insulin strategies to use. Significant interindividual differences in insulin dose requirements exist, necessitating personalized advice based on postprandial glucose monitoring for several hours afterwards [30]. This is consistent with a major emerging theme: empowering users to gain control over their T1D management by understanding personal trends and factors affecting glycemia. While users appreciated receiving personalized AI-generated

XSL•FO
RenderX

recommendations based on their trends, the sense of empowerment mainly came from accessing personal statistics and tracking various factors.

The interest to learn about personal trends aligned with participants' desire to remain in control and avoid full reliance on the app, which could hinder self-efficacy. Some users were skeptical about the accuracy of AI-generated information, emphasizing the need for safety mechanisms for data verification and the ability to override when necessary. This concern aligns with previous research on the lag between interstitial and capillary glucose levels, which led to data distrust and prompted some CGM users to rely on capillary glucose to clarify high and low readings [31]. Similarly, a recent study examining patient perspectives on AI in health care highlighted varying levels of comfort depending on the role of AI in care, with higher acceptance of apps that play an assistive role (eg, digital scribing or radiology review), and lower comfort when AI systems directly influenced decision-making [32]. Together, these findings highlight the tension between empowerment and over-reliance on technology in contexts where trust in automated systems is still evolving.

This sense of empowerment was further emphasized in the context of patient-practitioner communication. Participants reported that access to their own data enabled them to take a more active role in diabetes management by sharing relevant information and tracking topics they wished to discuss during medical appointments. This form of empowerment is particularly important during adolescence, a time when young people are striving to gain independence and shape their personal and social identities [33]. It is especially relevant given evidence that young adults with T1D often report health care needs and priorities that significantly differ from those perceived by primary caregivers and health care professionals [34]. Consistent with literature on shared medical appointments, which emphasize patient leadership and active participation in care, greater access to and understanding of personal health data has been shown to enhance patients' ability to engage more confidently in discussions with clinicians, supporting more collaborative and patient-led care [35].

Importantly, access to information must be implemented through a person-centered lens to balance what information youth want to access and *how* they want to engage with it. For example, participants did not want glucose data displayed on their dashboard each time they opened the app. It is well documented in the literature that living with and managing T1D is associated with significant psychological burden, with young people at increased risk of psychiatric disorders compared to peers without diabetes [36]. More specifically, data overload and constant exposure to CGM and insulin pump data have been identified as sources of stress [37]. Notably, while CGM use has been associated with improved well-being among young people, only a minority report regularly engaging in detailed data analysis [38]. These findings highlight the importance of selectively presenting information in digital tools intended to support diabetes management. At the individual level, participants voiced different needs: some valued a comprehensive app that consolidates all information necessary for diabetes management and reduces the need to juggle multiple sources, while others felt that repeatedly displaying data such as glucose readings or insulin doses was redundant, given its availability through CGM and insulin pump reports. Filtering information, by suppressing, deleting, or selecting irrelevant information, has been identified as a helpful strategy for mitigating information overload [39]. This aligns with personalization as a key design principle, enabling users to determine which factors to track, how data are visualized, and which trends are emphasized.

Participants also expressed a desire to access consolidated data, with logged meals alongside corresponding glucose values, to better understand personal patterns. The multifactorial nature of glycemic management is well established, as insulin requirements are influenced by factors beyond carbohydrate intake, including physical activity, stress, and illness [40]. This complexity highlights the importance of tools that can help support decision-making in diabetes management. Consistent with this, findings from a needs assessment survey emphasized the importance of tracking factors beyond dietary intake, such as hormonal changes, timing of insulin delivery, and sleep, alongside glucose data [21].

These considerations are particularly important given clinical appointments often prioritize glucose data and glycemic management, which, although essential, may leave limited space to address broader psychosocial, behavioral, and contextual factors that shape day-to-day management, largely due to time constraints [41]. Designing the app around youths' lived experiences, rather than solely around clinical priorities, helps bridge this gap and ensures greater relevance and usability. Incorporating features that reflect real-world practices, such as tracking whether insulin was administered before or after meals despite guidelines recommending injecting before meals [42], acknowledges daily realities and avoids setting unrealistic expectations. Consistent with this, a study on the co-design of a digital mental health platform for young adults highlighted the importance of relatability, with participants emphasizing that the platform should be one they can identify with to support meaningful engagement [43].

Strengths and Limitations

These results should be interpreted in light of the study's limitations. A major limitation is that most participants were users of AID, relying on systems that automate insulin delivery and integrate CGM data. Consequently, feedback may under-represent the perceived value of features aimed at supporting manual bolus calculations, while overrepresenting preferences for features that support retrospective reflection and learning from personal trends rather than real-time decision support. In addition, AID users may represent a subgroup with higher baseline engagement with diabetes technologies, greater digital literacy, and increased comfort with data interpretation. While some app functions may overlap with AID features, the app serves a complementary rather than redundant role. AID systems focus on insulin automation but provide limited insight into the behavioral and contextual factors influencing glucose stability. In contrast, the proposed app supports self-awareness, education, and behavioral reflection by linking glucose trends with meal composition, stress, physical activity, and other lifestyle factors. It also facilitates communication with health

XSL•FO
RenderX

care providers through integrated reports that combine these contextual data. Finally, its relevance extends to individuals on basal-bolus injection therapy, who continue to make daily dosing decisions manually, and to those without access to AID systems due to cost, eligibility, or access limitations. Other limitations include the potential selection bias arising from the nonrandom sample, which may affect generalizability. However, the users represented various ethnicities, genders, education levels, and insulin treatment modalities. Researcher bias was mitigated by using the same interviewer throughout, ensuring consistency in questioning style, tone, and data collection procedures across rounds. In addition to having an observer present during the interviews, subjectivity in data analysis was addressed by having 3 coders reach consensus through discussions. Observer bias was minimized by adhering to a preset questionnaire and having no prior relationships with participants. Due to the potential for authority and response bias, the interviewer mitigated their impact by establishing rapport with participants to create a comfortable and nonjudgmental environment, used neutral language and tone to avoid perceived authority, and encouraged openness by reassuring participants that their honest opinions were valued.

A key strength of the study is that end user needs directly informed the app's development, with participant input integrated iteratively across all design rounds. This co-design approach ensured that the resulting features were grounded in lived experience and reflected the priorities of young adults with T1D. By continuously refining the app based on user feedback, the design process enhanced relevance and usability, laying a strong foundation for future pilot testing and real-world engagement.

Conclusions

Petit Cactus is a mobile app that enables carbohydrate estimations through photo recognition and supports automated insulin dosing within a single, user-informed platform. Through its co-design process, the study identified design principles that extend beyond this specific application, emphasizing person-centered, intuitive, and emotionally sensitive interfaces that align with users' daily routines and language. Participants also underscored the importance of personalization, data accuracy, and effective communication tools to support meaningful interaction with health care providers. These insights could inform the design of digital health tools aimed at supporting diabetes self-management. Usability testing has refined Petit Cactus accordingly, and a pilot trial will next evaluate its impact on engagement and glycemic outcomes among young adults with T1D.

Funding

This work was supported by Mitacs (grant IT31370 ACC). AH received a Mitacs Accelerate scholarship. GG received a Natural Sciences and Engineering Research Council of Canada summer scholarship. AK received a Canadian Graduate Scholarship–Master's Program. CS received a Fonds de Recherche du Québec en Santé–Doctoral scholarship. ASB is a Fonds de Recherche du Québec en Santé research scholar and Mitacs grant recipient. All other authors received no financial support for the research, authorship, and/or publication of this article.

Data Availability

The datasets used and analyzed during the study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: AH, ASB
Formal analysis: AH
Interview facilitation: AH
Interview observation: AS, GG, CS
Methodology: AH
Participation recruitment: AH
Protocol – original draft: AH
Protocol – review & editing: AH, ASB
Project administration: AH
Supervision: ASB
Visualization: AH
Writing – original draft: AH, AS, AK, GG
Writing – review & editing: AH, AS, AK, GG, AR-F, CS, MN, ASB

Multimedia Appendix 1
Interview guide and questions.
[DOCX File, 18 KB - jmir_v28i1e86024_app1.docx]

Multimedia Appendix 2
Feedback questionnaire.

XSL•FO
RenderX

[DOCX File, 18 KB - jmir_v28i1e86024_app2.docx]

Multimedia Appendix 3
Themes and codes description
[DOCX File, 27 KB - jmir_v28i1e86024_app3.docx]

Multimedia Appendix 4
Appreciations and suggestions from the feedback questionnaire after 1 week of app use.
[DOCX File, 19 KB - jmir_v28i1e86024_app4.docx]

References

1. Ilonen J, Lempainen J, Veijola R. The heterogeneous pathogenesis of type 1 diabetes mellitus. Nat Rev Endocrinol 2019 Nov;15(11):635-650. [doi: 10.1038/s41574-019-0254-y] [Medline: 31534209]
2. Diabetes in Canada: Backgrounder. : Diabetes Canada; 2021 URL: https://www.diabetes.ca/DiabetesCanadaWebsite/media/Advocacy-and-Policy/Backgrounder/2021_Backgrounder_Canada_English_FINAL.pdf [accessed 2026-03-13]
3. Diabetes Canada Clinical Practice Guidelines Expert Committee, McGibbon A, Adams L, Ingersoll K, Kader T, Tugwell B. Glycemic Management in Adults With Type 1 Diabetes. Can J Diabetes 2018 Apr;42 Suppl 1:S80-S87. [doi: 10.1016/j.jcjd.2017.10.012] [Medline: 29650115]
4. Diabetes Control and Complications Trial Research Group, Nathan DM, Genuth S, et al. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. N Engl J Med 1993 Sep 30;329(14):977-986. [doi: 10.1056/NEJM199309303291401] [Medline: 8366922]
5. Diabetes Control and Complications Trial Research Group. Effect of intensive diabetes treatment on the development and progression of long-term complications in adolescents with insulin-dependent diabetes mellitus: diabetes control and complications trial. J Pediatr 1994 Aug;125(2):177-188. [doi: 10.1016/S0022-3476(94)70190-3]
6. Gonzalez C, Picón MJ, Tomé M, Pujol I, Fernández-García JC, Chico A. Expert study: Utility of an automated bolus advisor system in patients with type 1 diabetes treated with multiple daily injections of insulin-A crossover study. Diabetes Technol Ther 2016 May;18(5):282-287. [doi: 10.1089/dia.2015.0383] [Medline: 26886163]
7. Tascini G, Berioli MG, Cerquiglini L, et al. Carbohydrate counting in children and adolescents with type 1 diabetes. Nutrients 2018 Jan 22;10(1):109. [doi: 10.3390/nu10010109] [Medline: 29361766]
8. Spiegel G, Bortsov A, Bishop FK, et al. Randomized nutrition education intervention to improve carbohydrate counting in adolescents with type 1 diabetes study: Is more intensive education needed? J Acad Nutr Diet 2012 Nov;112(11):1736-1746. [doi: 10.1016/j.jand.2012.06.001] [Medline: 22975086]
9. Singla R, Bindra J, Singla A, Gupta G, Gupta Y, Aggarwal S. Automation of insulin bolus dose calculation in type 1 diabetes: a feasibility study. Int J Diabetes Dev Ctries 2023 Feb;43(1):66-71. [doi: 10.1007/s13410-022-01054-7]
10. Sussman A, Taylor EJ, Patel M, et al. Performance of a glucose meter with a built-in automated bolus calculator versus manual bolus calculation in insulin-using subjects. J Diabetes Sci Technol 2012 Mar 1;6(2):339-344. [doi: 10.1177/193229681200600218] [Medline: 22538144]
11. Borus JS, Laffel L. Adherence challenges in the management of type 1 diabetes in adolescents: prevention and intervention. Curr Opin Pediatr 2010 Aug;22(4):405-411. [doi: 10.1097/MOP.0b013e32833a46a7] [Medline: 20489639]
12. Shah BR, Lipscombe LL, Booth GL. Glycemic control among people with diabetes in Ontario: a population-based cross-sectional study. Can J Diabetes 2021 Jun;45(4):313-318. [doi: 10.1016/j.jcjd.2020.09.017] [Medline: 33221136]
13. Foster NC, Beck RW, Miller KM, et al. State of type 1 diabetes management and outcomes from the T1D exchange in 2016-2018. Diabetes Technol Ther 2019 Feb;21(2):66-72. [doi: 10.1089/dia.2018.0384] [Medline: 30657336]
14. Gurnani M, Pais V, Cordeiro K, Steele S, Chen S, Hamilton JK. One potato, two potato,… assessing carbohydrate counting accuracy in adolescents with type 1 diabetes. Pediatr Diabetes 2018 Nov;19(7):1302-1308. [doi: 10.1111/pedi.12717] [Medline: 29999219]
15. Klonoff DC. The current status of mHealth for diabetes: Will it be the next big thing? J Diabetes Sci Technol 2013 May 1;7(3):749-758. [doi: 10.1177/193229681300700321] [Medline: 23759409]
16. Liu NF, Brown AS, Folias AE, et al. Stigma in people with type 1 or type 2 diabetes. Clin Diabetes 2017 Jan;35(1):27-34. [doi: 10.2337/cd16-0020] [Medline: 28144043]
17. Whitehead L, Seaton P. The effectiveness of self-management mobile phone and tablet apps in long-term condition management: a systematic review. J Med Internet Res 2016 May 16;18(5):e97. [doi: 10.2196/jmir.4883] [Medline: 27185295]
18. Liang X, Wang Q, Yang X, et al. Effect of mobile phone intervention for diabetes on glycaemic control: a meta-analysis. Diabet Med 2011 Apr;28(4):455-463. [doi: 10.1111/j.1464-5491.2010.03180.x] [Medline: 21392066]
19. Vasiloglou MF, Marcano I, Lizama S, Papathanail I, Spanakis EK, Mougiakakou S. Multimedia data-based mobile applications for dietary assessment. J Diabetes Sci Technol 2023 Jul;17(4):1056-1065. [doi: 10.1177/19322968221085026] [Medline: 35348398]

XSL•FO
RenderX

20. Eiland L, McLarney M, Thangavelu T, Drincic A. App-based insulin calculators: current and future state. Curr Diab Rep 2018 Oct 4;18(11):123. [doi: 10.1007/s11892-018-1097-y] [Medline: 30284645]
21. Housni A, Katz A, Bergeron LJ, et al. Bridging the gap in carbohydrate counting with a mobile app: needs assessment survey. J Med Internet Res 2025 Mar 28;27:e63278. [doi: 10.2196/63278] [Medline: 40153793]
22. Gulliksen J, Göransson B, Boivie I, Blomkvist S, Persson J, Cajander Å. Key principles for user-centred systems design. Behav Inf Technol 2003 Nov;22(6):397-409. [doi: 10.1080/01449290310001624329]
23. Alwashmi MF, Hawboldt J, Davis E, Fetters MD. The iterative convergent design for mobile health usability testing: mixed methods approach. JMIR Mhealth Uhealth 2019 Apr 26;7(4):e11656. [doi: 10.2196/11656] [Medline: 31025951]
24. Brazeau AS, Messier V, Talbo MK, et al. Self-reported severe and nonsevere hypoglycemia in type 1 diabetes: population surveillance through the BETTER Patient Engagement Registry: development and baseline characteristics. Can J Diabetes 2022 Dec;46(8):813-821. [doi: 10.1016/j.jcjd.2022.05.010] [Medline: 35835670]
25. Pan Y, de la Puente M. Census bureau guideline for the translation of data collection instruments and supporting materials: documentation on how the guideline was developed. : Statistical Research Division, US Bureau of the Census; 2005 URL: https://www.census.gov/content/dam/Census/library/working-papers/2005/adrm/rsm2005-06.pdf [accessed 2026-03-13]
26. Zeynali-Tazehkandi M. Analyzing qualitative data with MAXQDA: a book review. Inf Discov Deliv 2022;50 [FREE Full text]
27. Kiger ME, Varpio L. Thematic analysis of qualitative data: AMEE Guide No. 131. Med Teach 2020 Aug;42(8):846-854. [doi: 10.1080/0142159X.2020.1755030] [Medline: 32356468]
28. Pańkowska E, Błazik M, Groele L. Does the fat-protein meal increase postprandial glucose level in type 1 diabetes patients on insulin pump: the conclusion of a randomized study. Diabetes Technol Ther 2012 Jan;14(1):16-22. [doi: 10.1089/dia.2011.0083] [Medline: 22013887]
29. Smart CE, Annan F, Higgins LA, Jelleryd E, Lopez M, Acerini CL. ISPAD Clinical practice consensus guidelines 2018: Nutritional management in children and adolescents with diabetes. Pediatr Diabetes 2018 Oct;27:136-154. [doi: 10.1111/pedi.12738] [Medline: 30062718]
30. Bell KJ, Toschi E, Steil GM, Wolpert HA. Optimized mealtime insulin dosing for fat and protein in type 1 diabetes: application of a model-based approach to derive insulin doses for open-loop diabetes management. Diabetes Care 2016 Sep;39(9):1631-1634. [doi: 10.2337/dc15-2855] [Medline: 27388474]
31. Lawton J, Blackburn M, Allen J, et al. Patients' and caregivers' experiences of using continuous glucose monitoring to support diabetes self-management: qualitative study. BMC Endocr Disord 2018 Feb 20;18(1):12. [doi: 10.1186/s12902-018-0239-1] [Medline: 29458348]
32. Foresman G, Biro J, Tran A, et al. Patient perspectives on artificial intelligence in health care: focus group study for diagnostic communication and tool implementation. J Particip Med 2025 Jul 24;17:e69564. [doi: 10.2196/69564] [Medline: 40705399]
33. King KM, King PJ, Nayar R, Wilkes S. Perceptions of adolescent patients of the "lived experience" of type 1 diabetes. Diabetes Spectr 2017 Feb;30(1):23-35. [doi: 10.2337/ds15-0041] [Medline: 28270712]
34. Chiang Y, Tsay P, Chen C, et al. A Delphi study on the healthcare needs of patients with type 1 diabetes during the transition from adolescence to adulthood: consensus among patients, primary caregivers, and healthcare providers. Int J Environ Res Public Health 2021 Jul 4;18(13):7149. [doi: 10.3390/ijerph18137149] [Medline: 34281086]
35. Wadsworth KH, Archibald TG, Payne AE, Cleary AK, Haney BL, Hoverman AS. Shared medical appointments and patient-centered experience: a mixed-methods systematic review. BMC Fam Pract 2019 Jul 8;20(1):97. [doi: 10.1186/s12875-019-0972-1] [Medline: 31286876]
36. Dybdal D, Tolstrup JS, Sildorf SM, et al. Increasing risk of psychiatric morbidity after childhood onset type 1 diabetes: a population-based cohort study. Diabetologia 2018 Apr;61(4):831-838. [doi: 10.1007/s00125-017-4517-7] [Medline: 29242985]
37. Kruger S, Deacon E, van Rensburg E, Segal D. Adjustment experiences of adolescents living with well-controlled type 1 diabetes using closed-loop technology. Front Clin Diabetes Healthc 2024;5:1445972. [doi: 10.3389/fcdhc.2024.1445972] [Medline: 39483147]
38. Huhn F, Lange K, Jördening M, Ernst G. Real-world use of continuous glucose monitoring systems among adolescents and young adults with type 1 diabetes: reduced burden, but little interest in data analysis. J Diabetes Sci Technol 2023 Jul;17(4):943-950. [doi: 10.1177/19322968221081216] [Medline: 35255729]
39. Kluge A, Antoni CH, Ellwart T. Digitalization as the problem of and the solution to vast amounts of data in future work—challenges for individuals, teams, and organizations. Z Arb Organisationspsychol A&O 2020 Jan;64(1):1-5. [doi: 10.1026/0932-4089/a000317]
40. DiMeglio LA, Evans-Molina C, Oram RA. Type 1 diabetes. Lancet 2018 Jun 16;391(10138):2449-2462. [doi: 10.1016/S0140-6736(18)31320-5] [Medline: 29916386]
41. Irving G, Neves AL, Dambha-Miller H, et al. International variations in primary care physician consultation time: a systematic review of 67 countries. BMJ Open 2017 Nov 8;7(10):e017902. [doi: 10.1136/bmjopen-2017-017902] [Medline: 29118053]

XSL•FO
RenderX

42. Jovanovic L, Giammattei J, Acquistapace M, Bornstein K, Sommermann E, Pettitt DJ. Efficacy comparison between preprandial and postprandial insulin aspart administration with dose adjustment for unpredictable meal size. Clin Ther 2004 Sep;26(9):1492-1497. [doi: 10.1016/j.clinthera.2004.09.001] [Medline: 15531011]
43. Ludlow K, Russell JK, Ryan B, et al. Co-designing a digital mental health platform, “Momentum”, with young people aged 7-17: a qualitative study. Digit Health 2023;9:20552076231216410. [doi: 10.1177/20552076231216410] [Medline: 38033517]

Abbreviations:

AI: artificial intelligence
AID: automated insulin delivery
BETTER: Behaviors, Therapies, Technologies and Hypoglycemic Risk in Type 1 Diabetes Registry
CC: carbohydrate counting
CGM: continuous glucose monitoring
HbA_{1c}: glycated hemoglobin
mHealth: mobile health
T1D: type 1 diabetes

Edited by A Stone; submitted 17.Oct.2025; peer-reviewed by M Ekwueme, O Ajayi, SM Youngblood; revised version received 25.Feb.2026; accepted 25.Feb.2026; published 31.Mar.2026.

Please cite as:
Housni A, Shulkin A, Katz A, Giannini G, Roy-Fleming A, Nakhla M, South C, Brazeau AS
Designing a Carbohydrate Counting App for Young Adults With Type 1 Diabetes: Usability Testing Interview Study
J Med Internet Res 2026;28:e86024
URL: https://www.jmir.org/2026/1/e86024
doi:10.2196/86024

© Asmaa Housni, Aidan Shulkin, Alexandra Katz, Giuliana Giannini, Amélie Roy-Fleming, Meranda Nakhla, Courtney South, Anne-Sophie Brazeau. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effectiveness of Postdischarge Telephone Calls in Reducing Hospital Utilization: Quasi-Randomized Controlled Trial

Sandie Du[1], MHA; Abigail Anada[1], MM; Maria Montenegro[1], MSc; Maurie Maitland[1], MSN; Sean Chuan[1], MSc; Sonia Jose[1], RN, BSN; Harveer Sihota[1], BSN, MSN; Delia Cooper[1], MSc; Joanna Rivera[1], MPH; Dianne Burditt[1], SEA; Minetaro Naruki-van Velzen[1], MSc; Megan MacPherson[2], PhD

[1]Digital Patient and Provider Experience, Fraser Health, Surrey, BC, Canada

[2]Department of Occupation Science & Occupational Therapy, Faculty of Medicine, University of British Columbia, 13737 96th Avenue Office #249, Surrey, BC, Canada

Corresponding Author:
Megan MacPherson, PhD
Department of Occupation Science & Occupational Therapy, Faculty of Medicine, University of British Columbia, 13737 96th Avenue Office #249, Surrey, BC, Canada

Abstract

Background: Unplanned emergency department (ED) visits and hospital readmissions following discharge contribute to patient distress, increased health care costs, and system inefficiencies. Early postdischarge follow-up can improve care transitions, yet evidence on the effectiveness of telephone-based interventions remains mixed. Telephone calls, a low-barrier form of digital health, may enhance equity and accessibility by reaching patients who face challenges with in-person or higher-technology follow-up.

Objective: This study evaluated the impact of a nurse-led postdischarge telephone intervention delivered by Fraser Health Virtual Care on short-term ED visits and hospital readmissions among recently discharged high-risk patients. Secondary objectives included examining patient experiences with the service and identifying care gaps addressed during follow-up calls.

Methods: A pragmatic quasi-randomized trial was conducted (May 2022-September 2022). Participants were eligible if they were aged 18 years or older and classified as high-risk for readmission using the LACE (Length of stay, Acuity of admission, Comorbidities, and Emergency department use) index (≥10 or <9 and ≥45 y). Participants were allocated to either a postdischarge telephone intervention group or a standard care control group based on daily nurse availability. Intervention participants received a structured nurse-led call 48 hours after discharge assessing understanding of discharge instructions, medication management, follow-up appointments, and home supports. Primary outcomes were ED visits within 7 and 30 days post call; secondary outcomes were hospital readmissions and patient experience. Negative binomial regression models were used to calculate adjusted incident rate ratios (IRRs).

Results: A total of 7091 participants were included (intervention: n=3911, of whom 1752 completed the call; control: n=3180). Postdischarge calls significantly reduced ED visits at both 7 days (adjusted IRR 0.719, 95% CI 0.617 - 0.837; $P<.001$) and 30 days (IRR 0.878, 95% CI 0.783 - 0.983; $P=.02$). No statistically significant reductions were observed in hospital readmissions at either 7 days (IRR 0.809; $P=.13$) or 30 days (IRR 0.942; $P=.54$). Forty percent of completed calls (n=701) identified at least 1 gap in discharge understanding or follow-up care. Most participants found the calls helpful and reported increased confidence in managing their care.

Conclusions: Structured nurse-led postdischarge telephone calls significantly reduced short-term ED utilization but did not impact readmission rates. These findings support the role of telephone-based virtual care as a scalable, low-barrier strategy to improve care transitions and reduce avoidable ED visits. Additional or ongoing interventions may be required to influence hospital readmission outcomes among high-risk patients.

Trial Registration: International Clinical Trials Registry Platform ISRCTN26233632; https://www.isrctn.com/ISRCTN26233632

(J Med Internet Res 2026;28:e80529) doi:10.2196/80529

KEYWORDS

patient readmission; hospitalization; emergency room; continuity of care; digital health

XSL•FO
RenderX

Introduction

Background

High volumes of emergency department (ED) visits and unplanned hospital readmissions following discharge have significant implications for patient outcomes, health care costs, and overall system efficiency. Readmissions within 30 days, which are often considered unplanned and avoidable, are linked to poor patient satisfaction [1], increased mortality risk [2], and substantial financial burden, accounting for over US $2.3 billion annually in Canada [3]. While not all ED visits and readmissions can be prevented, strengthening postdischarge follow-up and care coordination can help reduce these rates and improve patient experience.

The postdischarge period is a particularly vulnerable time for patients as they transition from hospital to home and assume responsibility for their own care. Patients often leave the hospital with new medications, recommended lifestyle changes, or with ongoing care needs, but many may not fully understand their discharge instructions [4]. Limited awareness of their symptoms, difficulty accessing follow-up care, and a lack of support at home can further increase their risk of complications [5]. Without proper guidance and resources, patients are more likely to experience preventable adverse events, leading to ED visits or hospital readmissions [6]. Proactive care coordination during this transition period has the potential to reduce hospital utilization, including ED [7] and hospital readmission rates [8,9]. One promising intervention is postdischarge telephone calls.

Telephone Calls as Digital Health Interventions

While digital health is often associated with apps, wearables, and video consultations, it is important to recognize that telephone-based care is a legitimate and impactful digital health tool, particularly in real-world contexts where access and equity are paramount. Digital health can be broadly defined as the delivery of care at a distance, the digital storage and sharing of health information, and the use of health-related data to improve services and systems. This broad definition allows inclusion of “low-tech” modalities, such as telephone or SMS text messaging–based interventions, alongside more “high-tech” solutions like app-based programs or algorithm-driven self-management tools [10].

Telephones are widely accessible and require no broadband internet, which makes them particularly valuable for reaching populations who may be marginalized by more complex digital platforms (eg, older adults, those with lower incomes, and those with limited English proficiency) [11]. Unlike video visits and other forms of digital care which demand stable internet and technical know-how [12], telephone calls lower the barrier to entry, enabling broader participation in virtual care [13].

From an equity standpoint, telephone-based care helps mitigate the digital divide [14,15]. Research has shown that disparities in mobile phone ownership are significantly less pronounced than those in internet access or smartphone usage [16]. This positions telephone calls as a uniquely inclusive tool in digital health strategies.

Postdischarge Calls

Postdischarge calls aim to bridge information gaps, enhance adherence to care plans, and offer emotional support to patients during their transition from hospital to home [17]. Although postdischarge calls have been implemented across various patient populations [18], the evidence supporting their effectiveness remains mixed. Studies have found that postdischarge calls are feasible [19], can improve patients’ experience [20], and may improve follow-up with primary care providers [21]. However, findings on their effectiveness in reducing ED visits and hospital readmissions are inconclusive, partly due to methodological limitations such as small sample size or variability in intervention protocols [22]. Despite these mixed results, the potential for postdischarge calls as part of a comprehensive care intervention continues to be explored, with some studies suggesting their effectiveness in reducing readmissions [23-25].

This study aimed to evaluate the effectiveness of nurse-led postdischarge telephone calls provided by Fraser Health Virtual Care (FHVC) for high-risk patients. Specifically, we examined the impact of a telephone call made 48 hours after hospital discharge on ED visits and hospital readmission within 7 and 30 days following the call. By assessing this intervention, we aim to provide further insight into the role of structured postdischarge follow-up calls in reducing hospital service use and improving patient outcomes.

FHVC Service

Fraser Health is 1 of 5 health authorities in British Columbia, Canada, and provides care to more than 2 million people across 20 diverse communities through 12 acute care hospitals, outpatient and surgery centers, and various community care sites [26].

In response to the COVID-19 pandemic and the growing need for digitally enabled care, Fraser Health launched FHVC in April 2020, a nurse-led virtual health service that delivers clinical assessment and navigation support through telephone, video, and secure web chat. FHVC services include general health advice and information, personalized clinical assessments, and referrals to appropriate Fraser Health programs and services [26].

FHVC is integrated within the health authority’s digital infrastructure: clinicians document each encounter directly into the electronic health record (EHR), access real-time clinical data, and coordinate follow-up or referrals electronically. The service also supports medical interpreter access, ensuring linguistic inclusivity and equitable reach across the region.

In March 2021, FHVC expanded to include proactive postdischarge outreach calls to patients within 48 hours of hospital discharge. This initiative was introduced as part of a regional digital quality improvement strategy to strengthen care transitions and reduce avoidable ED visits and readmissions. Calls are initiated automatically from daily EHR-generated discharge lists, and nurses follow a standardized digital protocol addressing symptoms, medications, follow-up appointments, and home supports.

XSL•FO
RenderX

This postdischarge component represents a digitally enabled care coordination model that leverages FHVC's integrated data systems, digital workflows, and virtual communication tools. Unlike usual care, which relies on patients or community providers to initiate follow-up, FHVC's model provides proactive, system-triggered outreach, a scalable approach to virtual care that blends clinical expertise with health-system data infrastructure.

Methods

Study Design

This study was a multicenter, pragmatic quasi-randomized trial conducted across all 12 acute care hospitals in the Fraser Health region (British Columbia, Canada) from May 16, 2022, to September 21, 2022. The trial evaluated the effect of a routine, operationally delivered, postdischarge follow-up phone call on ED visits and hospital readmissions among high-risk patients. The design was pragmatic in nature, embedded within routine care delivery and implemented by the existing FHVC nursing staff, to reflect real-world practice conditions. The trial was retrospectively registered with the ISRCTN registry (registration submission date: September 3, 2025).

The primary objective was to determine the impact of a 1-time follow-up phone call 48 hours after hospital discharge on 30-day ED visits for participants at high risk of readmission. Secondary objectives were to assess 7-day ED visits, 7- and 30-day readmissions, and participant experiences with the service. This trial is reported in accordance with the CONSORT guidelines (Checklist 1).

Participants and Setting

Participants were adults (≥18 y) discharged from any of Fraser Health's 12 acute care hospitals. The inclusion and exclusion criteria reflected routine FHVC operational parameters, designed to identify typical patients most likely to benefit from postdischarge follow-up.

Eligibility was informed by the LACE index [9,27,28], a validated measure of 30-day readmission risk incorporating Length of stay (LOS), Acuity of admission, Comorbidities, and Emergency department use. Participants were eligible if they had a LACE score ≥10, or a score <9, and were aged 45 years or older. These thresholds were derived from internal data showing an elevated risk of unplanned readmission.

Exclusion criteria included patients discharged from psychiatric units, to hospice or community nursing programs (where follow-up is routine), those who received additional postdischarge support, or were readmitted or deceased prior to follow-up.

Intervention

The intervention was a one-time, proactive follow-up call from an FHVC registered nurse approximately 48 hours after discharge. The intervention required no additional infrastructure, using existing FHVC systems, staff, and workflows. Nurses used automated discharge lists generated daily from the EHR to identify eligible participants.

During the calls, nurses followed a semistructured clinical assessment tool covering as follows:

- Understanding of discharge diagnosis and instructions,
- Medication adherence and reconciliation,
- Symptoms or concerns postdischarge, and
- Follow-up appointments and home supports

Calls were tailored to each patient's discharge plan and medical record. If issues were identified, nurses provided clarification, reinforced education, or facilitated referrals to community or primary care services. Interpreter services were available when required. Nurses used an assessment form to track this information as shown in Appendix S2 in Multimedia Appendix 1.

The intervention reflected usual care processes augmented by proactive outreach and was intentionally flexible to accommodate real-world nursing workflow and patient needs.

Comparator (Usual Care)

Participants in the control group received standard care only, which typically includes written discharge instructions and follow-up arranged by hospital units or primary care providers at their discretion. No proactive FHVC calls were made to these participants.

Recruitment and Allocation

Eligible participants were identified through the use of a custom iTracker report that automatically retrieved electronic medical record (EMR) discharge data across Fraser Health hospitals. Each morning, a research assistant applied the eligibility criteria and randomly selected up to 70 participants (using Microsoft Excel's RAND function) for the intervention group, based on daily nurse capacity. The remaining eligible participants were assigned to the control group.

Because the assignment depended on operational capacity, the allocation was quasi-random and not concealed. On days when the total number of eligible discharges was fewer than 70, all available participants were assigned to the intervention group to maintain operational continuity. Conversely, on days with limited nurse availability, more participants were assigned to the control group. This pragmatic approach preserved ecological validity by embedding group assignment within real-world service delivery constraints.

Registered nurses were blinded to the control group list and conducted intervention calls as part of their routine FHVC workload. Participants were unaware of the research component, ensuring naturalistic behavior.

Outcomes

Our primary outcome measure was the number of ED visits within 30 days of receiving the intervention, defined as any visit in which a participant registered in any of the 12 EDs within the Fraser Health region, regardless of whether they were ultimately seen by a clinician.

Secondary outcome measures included the following:

1. ED visits within 7 days of the intervention;

XSL•FO
RenderX

2. Hospital readmissions, at 7 and 30 days postintervention, defined as any admission to a hospital postdischarge; and
3. Participant experience, measured via a survey.

Sample Size Considerations

The sample size calculation was based on a retrospective analysis of EMRs from June 1, 2020, to May 31, 2021. This analysis estimated the average ED visits within 30 days of discharge before and after FHVC's postdischarge calls were implemented (0.73 and 0.59 ED visits, respectively). To detect a statistically significant difference with 80% power and α=.05, a required sample size of 6930 participants (3465 per group) was calculated for 2 independent groups.

Procedures and Data Collection

FHVC nurses received standardized training to conduct postdischarge follow-up calls from a Clinical Nurse Educator, ensuring consistency in care delivery. Calls were conducted using standard FHVC telehealth systems, and data were recorded in a secure Excel file linked to the EHR.

During each call, FHVC nurses recorded participant information, including name, age, sex, phone number, LACE score, and discharging hospital, into a Microsoft Excel file stored on a secure network drive. Additional data included any identified knowledge gaps during the postdischarge call. A research assistant monitored data daily to ensure completeness and accuracy of data collection.

If participants were unreachable, the nurse left a voicemail message with the FHVC phone number, inviting the participant to call back if they needed support. Data on ED visits, hospital readmissions, LOS of initial hospitalization, city of residence, and the day of discharge were extracted from the EMR using unique identifiers.

All participants were invited via a telephone call to complete a patient experience survey 30 days after the intervention until 80 participants were reached for each of the intervention and control groups. Surveys were administered by a research assistant and a patient partner. The survey instruments differed by study arm to reflect the experience received (see Appendix S3 in Multimedia Appendix 1 for survey questions). Participants who received the postdischarge call from FHVC answered questions about the call itself, including its helpfulness in clarifying discharge instructions, medications, follow-up appointments, and confidence in managing care at home. Participants who did not receive a postdischarge call were asked parallel questions about their discharge experience, including understanding of instructions, medications, follow-up appointments, and confidence in self-management.

Statistical Analysis

The analysis was completed by a Fraser Health statistician using SPSS (Statistical Package for the Social Sciences) version 21.0 (IBM Corp). The primary analyses followed an intention-to-treat approach. Differences in demographic and outcome variables between groups were analyzed using chi-square tests for categorical variables and the Welch 2-tailed *t* tests for continuous variables. Mann-Whitney *U* tests were used where data were not normally distributed. Average levels of risk factors were compared between groups using relative risks, 95% CIs, and 2-tailed *P* values. To reduce the risk of type I error due to multiple comparisons, a Bonferroni correction was applied to the *P* values from unadjusted univariate tests conducted on demographic and clinical risk variables. This included both between-group comparisons (intervention vs control) and exploratory within-group comparisons (reached vs not reached in the intervention arm). Adjusted *P* values were reported where appropriate.

Negative binomial regression models were used to examine the relationship between the dependent variables (ie, the number of ED visits and count of readmissions at 7 and 30 d post call) and relevant predictors. This modeling approach was selected to account for the overdispersion in the data due to the high occurrence of zero values in the count data for ED visits and hospital admissions. Independent variables included intervention group, age, sex, LACE score, LOS of initial hospitalization, day of discharge of initial hospitalization, and geographical location of ED of initial hospitalization. Regression models were not corrected for multiple comparisons, as they represented prespecified primary analyses. Unlike univariate tests, regression coefficients are estimated jointly, inherently accounting for covariation among predictors. Therefore, corrections for multiple comparisons were not applied.

Incident rate ratios (IRRs) were reported to describe the association between the independent variables and the relative frequency of count of ED visits and count of readmissions at 7 and 30 days. A sensitivity analysis was conducted in which an outlier participant with 23 ED visits was recoded to the next-highest value observed in the dataset (13 visits). The results remained robust following this adjustment. Due to missing values in the initial LOS variable, 32 observations (9 in the control group and 22 in the intervention group) were excluded from the final regression models.

In addition to the primary and secondary outcomes, the number needed to treat (NNT) was calculated to further assess the potential clinical significance of the intervention [29]. NNT is a measure used to determine how many participants need to be treated with the intervention to prevent one additional adverse outcome, such as an ED visit or hospital readmission. The NNT was calculated for the primary outcome of ED visits and the secondary outcome of hospital readmissions at both 7 and 30 days after the follow-up call.

The formula used for calculating NNT is:

$$NNT=1/|CER - EER|$$

where CER is the control event rate (ie, the proportion of participants in the control group who experience the adverse outcomes), and EER is the experimental event rate (ie, the proportion of participants in the intervention group who experienced the adverse outcome). The calculated NNT values provide a clinical perspective on the effectiveness of the intervention, helping to assess how many participants need to receive the postdischarge follow-up call to prevent 1 additional adverse outcome.

Descriptive statistics for the participant experience survey were calculated using Microsoft Excel. No inferential tests were

XSL•FO
RenderX

conducted because the survey questions differed between groups. The intervention group was asked whether the postdischarge call improved their understanding, and the control group was asked about their understanding of the initial discharge instructions; therefore, direct between-group comparisons were not possible.

Ethical Considerations

This study was reviewed and approved by the Fraser Health Research Ethics Board (number 2022192). Given that the study evaluated a quality improvement initiative embedded within routine operations and used deidentified secondary administrative data, the requirement for individual informed consent was waived by the ethics committees in accordance with institutional policy. All data were deidentified before analysis. No patient-identifiable information was shared outside the health authority. Data were stored on secure, password-protected servers in compliance with provincial privacy legislation. No financial or material compensation was provided to participants, as all activities represented standard care or evaluation of routine service delivery.

Results

Study Population

A total of 7091 participants were included in the study, with 3911 in the intervention group and 3180 in the control group (Figure 1). The overall sample had a mean age of 68.1 (SD 13.74) years (95% CI 67.8 - 68.4), and 47.2% (n=3349) were female.

Among the participants in the intervention group, 1752 fully completed the postdischarge call, while nurses were unable to reach the remainder directly. Of those who completed the call, 40% (n=701) had at least 1 care gap identified.

Figure 1. Flow diagram of participant inclusion and study design. This flowchart outlines participant selection for the retrospective cohort study evaluating the impact of nurse-led postdischarge telephone calls provided by Fraser Health Virtual Care (FHVC) on emergency department (ED) visits and hospital readmissions. The study included adult patients (≥18 y) discharged from acute care hospitals within Fraser Health, British Columbia, Canada, between January 2022 and December 2023.

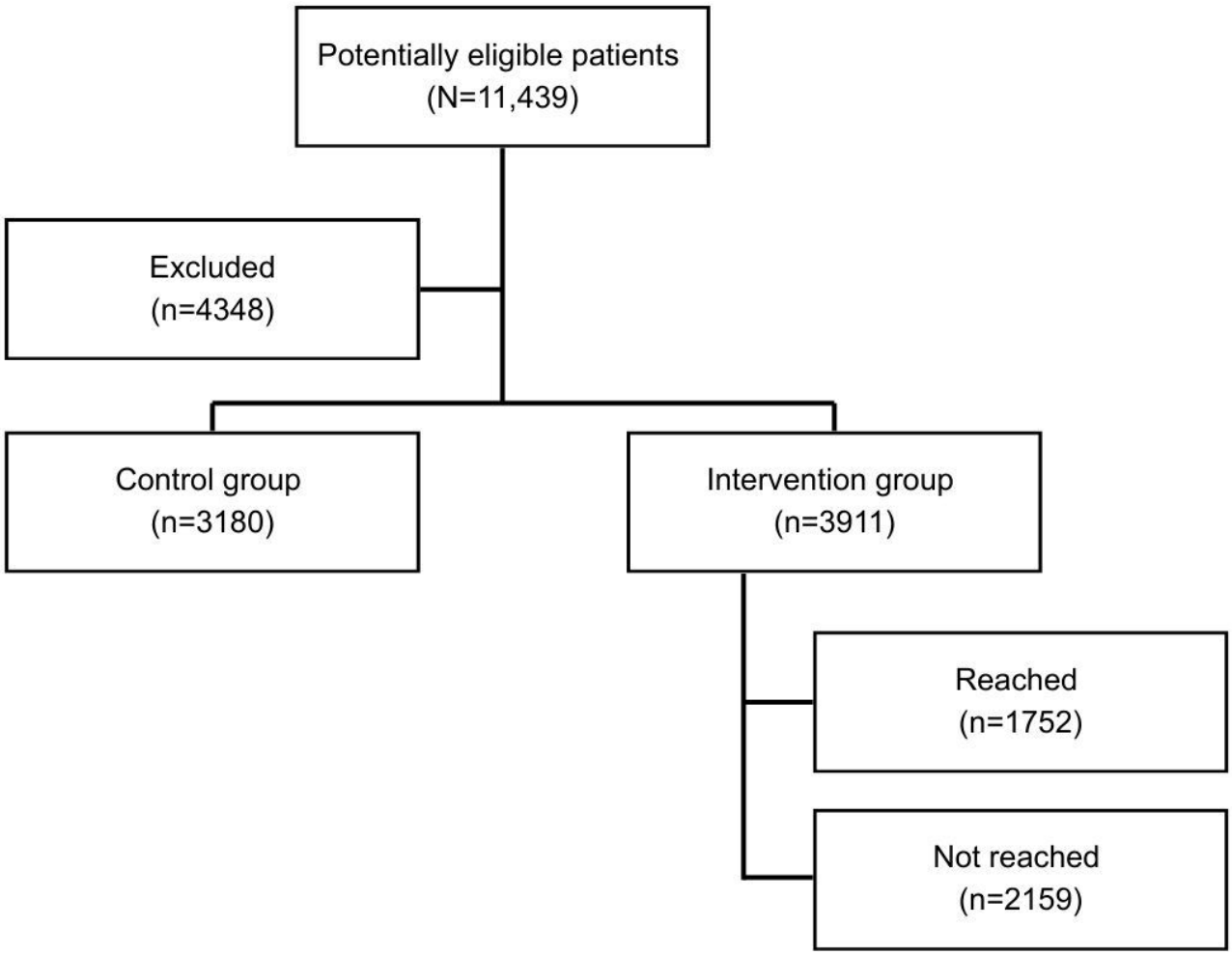

Baseline Characteristics

Baseline characteristics for the intervention and control groups are presented in Table 1. After applying a Bonferroni correction for multiple comparisons (adjusted α=.0042 and adjusted P<.012), the only statistically significant difference was a higher proportion of weekend discharges in the intervention group (33.8% vs 21.5%; adjusted P<.012). No meaningful differences were observed in age, sex, LACE score, LOS, or hospital location. Within the intervention arm, baseline characteristics were similar between participants who completed the call and those who did not (Appendix S4 in Multimedia Appendix 1).

XSL•FO
RenderX

Table . Baseline characteristics of study sample. Characteristics of patients included in the retrospective cohort study assessing Fraser Health Virtual Care (FHVC)'s postdischarge telephone calls.[a]

	Intervention (n=3911)	Control (n=3180)	Difference statistics	Unadjusted *P* value	Bonferroni adjusted *P* value
Age, mean (SD)	67.7 (13.5)	68.6 (14.1)	6001487.5[b]	.01	.132
Sex, n (%)			χ^2=7.4[c]	.006	.072
Female	1790 (45.8)	1559 (49)			
Male	2121 (54.2)	1621 (51)			
LACE[d] score, mean (SD)	8.56 (4.42)	8.29 (4.53)	5943871[b]	.001	.012
LOS[e] days of initial hospitalization, mean (SD)	5.49 (6.94)	6.57 (9.70)	6001849[b]	.05	.6
Discharge timing, n (%)			χ^2=131.2[c]	<.001	<.012
Discharged on weekend	1323 (33.8)	684 (21.5)			
Discharged on weekday	2588 (66.2)	2496 (78.5)			
Hospital location, n (%)			χ^2=3.4[c]	.06	.78
Rural or urban hospital	2278 (58.2)	1783 (56.1)			
Metro hospital	1633 (41.8)	1397 (43.9)			

[a]Data include demographics, clinical risk factors (eg, LACE score and LOS), discharge timing, and hospital location. Comparisons are presented between the intervention group (received nurse-led call within 48 h postdischarge) and matched control group (no call), with difference statistics and *P* values.

[b]Mann-Whitney *U* test.

[c]Chi-square test. All degrees of freedom were 1.

[d]LACE: Length of stay, Acuity of admission, Comorbidities, and Emergency department use

[e]LOS: length of stay.

Primary Outcome: ED Visits

Participants in the intervention group had significantly fewer ED visits at both 7-day and 30-day postdischarge call compared to those in the control group. Within 7 days of the discharge call, 9.33% (365/3911) of intervention participants visited an ED, compared to 11.86% (377/3180) of control participants (Table 2).

The NNT to prevent 1 ED visit within 7 days post call was 40 (95% CI 2774). For ED visits within 30 days, NNT=86; however, the 95% CI crossed zero, indicating nonsignificance.

Table . Trial outcomes with unadjusted 7-day and 30-day emergency department (ED) visits and hospital readmissions following discharge, and counts and proportions of patients with ED use or hospital readmission after discharge among those who received a postdischarge telephone call (intervention) and those who did not (control) in Fraser Health, British Columbia, Canada (2022–2023).

Postdischarge call	Intervention (n=3911)	Control (n=3180)
Within 7 days of call		
ED visits, n	427	472
ED use, n (%)	365 (9.33)	377 (11.9)
Hospital admissions, n	109	115
Hospital use, n (%)	106 (2.7)	109 (3.4)
Within 30 days of call		
ED visits, n	1072	973
ED use, n	759 (19.4)	654 (20.6)
Hospital admissions, n	292	253
Hospital use, n (%)	257 (6.6)	214 (6.7)

XSL•FO
RenderX

Adjusted analyses confirmed that the intervention was associated with a significant reduction in 7-day ED visits (IRR 0.72, 95% CI 0.62 - 0.84; *P*<.001) and a modest reduction at 30 days (IRR 0.88, 95% CI 0.78 - 0.98; *P*=.02). Higher LACE scores were consistently associated with increased ED use. No significant associations were observed for age, sex, LOS, day of discharge, or hospital location (Tables 3 and 4).

Table . Negative binomial regression results for 7-day emergency department (ED) visits following discharge[a].

	IRR[b] (95% CI)	*P* value
Intervention	0.719 (0.617-0.837)	<.001
Age (y)	0.996 (0.990-1.001)	.11
Female (vs male)	0.886 (0.762-1.030)	.12
LACE[c] score	1.051 (1.034-1.068)	<.001
LOS[d] at initial hospitalization	0.992 (0.982-1.002)	.12
Weekday discharge (vs weekend)	0.982 (0.827-1.164)	.84
Metro hospital (vs rural or urban)	0.897 (0.769-1.045)	.16

[a]Adjusted IRR and 95% CIs for factors associated with ED visits within 7 days of hospital discharge among adult patients in Fraser Health, British Columbia (2022–2023). The model includes intervention exposure and demographic and clinical covariates.

[b]IRR: incident rate ratio.

[c]LACE: Length of stay, Acuity of admission, Comorbidities, and Emergency department use.

[d]LOS: length of stay.

Table . Negative binomial regression results for 30-day emergency department (ED) visits following discharge[a].

	IRR[b] (95% CI)	*P* value
Intervention	0.878 (0.783 - 0.983)	.02
Age	0.991 (0.987 - 0.995)	<.001
Female (vs male)	0.941 (0.841 - 1.053)	.29
LACE[c] score	1.082 (1.069 - 1.095)	<.001
LOS[d] at initial hospitalization	1.000 (0.993 - 1.006)	.96
Weekday discharge (vs weekend)	0.957 (0.843 - 1.086)	.50
Metro hospital (vs rural or urban)	0.897 (0.801 - 1.006)	.06

[a]Adjusted IRR for ED visits within 30 days of hospital discharge among adult patients in Fraser Health, British Columbia (2022–2023), comparing those who received a postdischarge Fraser Health Virtual Care (FHVC) nurse call versus matched controls.

[b]IRR: incident rate ratio.

[c]LACE: Length of stay, Acuity of admission, Comorbidities, and Emergency department use.

[d]LOS: length of stay.

Secondary Outcome: Hospital Readmissions

There was no statistically significant reduction in hospital readmissions at 7 or 30 days. Within 7 days, 2.7% of intervention participants were readmitted versus 3.4% of controls (IRR 0.81, 95% CI 0.62 - 1.06; *P*=.13). At 30 days, readmissions were 6.6% versus 6.7% (IRR 0.94, 95% CI 0.78 - 1.14; *P*=.54). LACE score and LOS were significant predictors of readmission, while other covariates were not. NNT estimates to prevent one hospital readmission were 139 (7 days) and 625 (30 days), with 95% CIs crossing zero (Table 2; Appendix S5 in Multimedia Appendix 1).

Patient Experience Survey Results

Between June 14, 2022, and August 29, 2022, a total of 457 participants (control group: n=223; intervention group: n=264). The study design specified a target sample size of 80 participants per group, and recruitment was concluded once this threshold was reached, resulting in 160 participants (n=80 per group), yielding response rates of 36% in the control group and 30% in the intervention group. Among control group respondents, 67 (83.8%) participants were between the ages of 50 and 70 years, 43 (53.8%) were male, and 51 (63.8%) identified as White. Among intervention group respondents, 56 (70%) participants were between the ages of 50 and 70 years, 45 (56.3%) were male, and 57 (71.3%) identified as White.

Participants in the intervention group reported that the postdischarge call improved their understanding of hospital discharge instructions (n=46, 57.5% agreed) and 40% (n=32) stated that the call provided additional information they did not receive at the time of discharge. Most participants found the

XSL•FO
RenderX

postdischarge call helpful in answering their questions (n=33; 41.3% strongly agreed) and 46.3% (n=36) felt more confident managing their own health care after discharge. Nearly all participants viewed the postdischarge call as a valuable service (n=40, 50.6% strongly agreed; n=36, 45.6% agreed). Overall, 71.6% of (n=53) participants in the intervention group visited a primary care provider for follow-up care after discharge.

Among participants in the control group, 47.5% (n=38) strongly agreed that they clearly understood the hospital discharge instructions, and 46.3% (n=37) agreed that they were satisfied with the information provided. Additionally, 51.3% (n=41) felt confident managing their health at home. In total, 60.8% (n=45) of participants in the control group visited a primary care provider for follow-up care after discharge. See Appendix S3 in Multimedia Appendix 1 for full results from patient experience surveys.

Discussion

Overview

This study evaluated a 1-time, nurse-led postdischarge follow-up call delivered through an integrated virtual care program. The intervention was associated with modest improvements in early postdischarge outcomes, with participants in the intervention group experiencing significantly fewer ED visits at both 7 and 30 days. However, hospital readmission rates were not significantly affected.

The observed 28% relative reduction in 7-day ED visits and 12% reduction in 30-day ED visits highlight the clinical relevance of the intervention, supported by the NNT estimates which showed that 39 participants needed to receive the intervention to prevent one 7-day ED visit, and 83 patients to prevent one 30-day ED visit. These are clinically meaningful numbers, especially in the context of high hospital discharge volumes, where even modest improvements can translate into substantial system-wide benefits. While hospital readmission reductions were numerically favorable, they did not reach statistical significance, potentially due to lower event rates and limited power. Alternatively, while a postdischarge phone call can clarify instructions, reduce anxiety, and address basic care gaps, a single call may be insufficient to address the complex medical and social factors that contribute to hospital readmissions [30,31] and patients with elevated LACE scores may require more comprehensive or repeated interventions to meaningfully reduce readmissions.

Patient Experience

Participants receiving postdischarge phone calls reported overall satisfaction with the intervention and found that the calls provided them with additional information, reduced their concerns, and gave them more confidence to manage their own health care at home (Appendix S3 in Multimedia Appendix 1). Interestingly, participants in the control group also reported satisfaction with the discharge information they received, suggesting that unrecognized gaps may remain without structured follow-up. During the postdischarge calls, FHVC nurses asked participants to explain their discharge plan, which often revealed missing or misunderstood information. Nurses followed a standard script which covered essential postdischarge topics. Through this process, 40% of participants in the intervention group were identified as having at least 1 care gap. It is plausible that some participants in the control group had similar gaps, but without prompting, they were not aware of them at the time of the survey, and therefore felt they could manage adequately post discharge. Additionally, 41% of control participants expressed that they would have appreciated FHVC contact information sooner, highlighting the potential value of proactive outreach.

Potential Mechanisms of Action

Follow-up calls likely operate through multiple mechanisms: reinforcing patient education, supporting medication adherence, addressing emotional concerns, and signaling coordinated care across the health system. The intervention may serve both as a direct support tool and as an organizational signal promoting better discharge and follow-up practices. While single-touch calls improved ED utilization, they appear insufficient to address complex factors driving hospital readmissions, underscoring the need for more intensive or repeated interventions for high-risk populations.

Several studies have demonstrated that patient education [32] and medication adherence [33,34] are critical factors in reducing avoidable health care utilization, which may explain the observed reduction in ED visits in our study. Prior research has demonstrated that inadequate understanding of discharge instructions is a major driver of postdischarge complications and unplanned care use [35,36]. By reinforcing these instructions, postdischarge follow-up calls may have improved patient comprehension of their recovery plan, reduced knowledge gaps, and empowered patients to manage symptoms at home, thereby lowering the need for ED visits. In addition, the calls may have alleviated anxiety and promoted confidence in self-management [37] by providing reassurance and tailored guidance, thereby reducing the likelihood of seeking avoidable emergency care.

Medication adherence is another well-documented challenge following hospital discharge, with nonadherence associated with increased ED visits and readmissions [38]. The follow-up calls offered a valuable opportunity to clarify prescriptions, address concerns, and reinforce adherence strategies. This may have helped prevent adverse events related to medication errors. Similar findings have been observed in previous studies where medication reconciliation and adherence support were associated with reductions in acute care use [32]. These results suggest that postdischarge calls extend beyond simple follow-up; they actively address gaps in understanding, medication adherence, and emotional well-being, contributing to their observed effectiveness in reducing ED visits.

Importantly, within the intervention arm, completing the call did not meaningfully change ED or readmission outcomes compared with participants who were not reached, suggesting that the benefits observed may reflect system-level effects of the outreach model rather than call completion alone. For example, the intervention may have functioned as a signal of coordinated care, reinforcing follow-up workflows, prompting timely triage, or improving discharge planning processes across

XSL•FO
RenderX

the health system. Alternatively, exposure misclassification, where a participant assigned to the intervention group did not actually receive a call, may have diluted the measured effect of individual calls, meaning the observed reductions in ED visits reflect the broader impact of the intervention's integration into routine care rather than direct participant-level engagement alone. Recognizing these mechanisms underscores that single-touch outreach can influence outcomes through organizational processes as well as patient-level interactions.

Clinical Implications

Postdischarge calls have been shown to enhance patient engagement with primary care providers, improve medication adherence, and contribute to better overall health outcomes [21]. Given the increasing difficulty in accessing primary care [39,40], these calls may serve as a low-barrier intervention to support care transitions and fill gaps in follow-up care.

Discharge calls can facilitate smoother transitions from hospital to home by proactively identifying and addressing common barriers to recovery. By intervening early, the discharge calls can reduce the risk of preventable complications and improve the overall patient experiences [38] while also relieving pressures on overstretched health care systems [41].

Reducing ED visits has important implications for health care systems, as ED overcrowding remains a critical issue [42]. A reduction in ED visits suggests that patients may be managing their conditions more effectively in outpatient settings, potentially leading to cost savings and improved resource allocation. Future research should explore whether postdischarge calls lead to increased engagement with primary care providers, potentially shifting health care utilization patterns.

Strengths and Limitations

Previous reviews, including 1 Cochrane review [18] and 2 systematic reviews [21,25] have noted that many trials lacked sufficient statistical power, clearly defined sample size calculations, rigorous methodology, or standardized interventions. In contrast, our study addressed these limitations by using a larger sample size, informed by Fraser Health historical data. Additionally, we implemented a structured intervention that included standardized nurse training and call templates, ensuring consistency in care delivery across settings.

A major strength of this trial was its integration into an existing digital health care service. FHVC was already operational before the study, meaning that nurses conducting the calls had prior experience and received training from a clinical nurse educator. This level of standardization may have contributed to the observed effectiveness of the intervention, differentiating our study from others that implemented newly introduced follow-up call programs. Conducted within an existing operational health care service, the study reflected real-world variations in patient care, staffing constraints, and service delivery. As a result, our findings represent an intervention that health care organizations could realistically implement to support patients' transitions from hospital to home.

The broad eligibility criteria further enhance the generalizability of our findings. Although the study focused on high-risk patients, it included patients with a range of health conditions, demographics, and socioeconomic backgrounds. The standardized discharge call protocol was designed to be adaptable across clinical contexts, and a detailed version is provided for reference (see Appendix S2 in Multimedia Appendix 1).

A key limitation is the potential exclusion of patients without reliable phone access. Certain socioeconomic groups may have faced barriers to participating, including a lack of access to phones or data plans. Although FHVC implemented strategies to enhance accessibility, such as using plain language and integrating interpreters to support patients who do not speak English, these barriers remain a challenge for equitable access to virtual care.

Another limitation of this study was the low intervention exposure rate within the intervention group. While the analysis followed an intention-to-treat approach, a substantial proportion of participants did not answer the postdischarge call and therefore were not actually exposed to the intervention. This may have diluted the observed effect and may have underestimated the true impact of the intervention had more participants engaged. Future efforts should focus on improving patient engagement, such as further increasing awareness of the follow-up service at the point of discharge, to enhance reach and effectiveness.

Despite being guided by a sample size calculation, the study faced logistical challenges due to fluctuating discharge volumes and staffing availability. While the goal was to collect data for 58 days, recruitment extended to 84 days to achieve the desired sample size. Nurse capacity constraints also required adjustments to the intervention group sizes, leading to a control group that was smaller than the initially planned sample size (3180 instead of 3465) and an intervention group that was larger (3911 vs 3465). These adjustments reflect the realities of implementing research within an operational care model and underscore the need to balance methodological rigor with service delivery. These deviations from the original allocation plan may impact statistical power and group comparability and should be considered when interpreting results.

Future Directions

Overview

While this study provides strong evidence for the effectiveness of postdischarge follow-up calls, several areas require further exploration to optimize implementation, improve equity, and maximize impact. Future research should focus on leveraging emerging technologies, understanding the mechanisms driving intervention success, addressing disparities in access to virtual care, and evaluating cost-effectiveness. Additionally, qualitative and subgroup analyses may help further refine the intervention's design and delivery to better meet the needs of diverse patient populations.

Leveraging Emerging Technologies for Enhanced Support

Artificial intelligence (AI) has the potential to improve the scalability and efficiency of postdischarge interventions by

XSL•FO
RenderX

augmenting human-led care with data-driven insights. AI-driven decision support systems could help identify patients most likely to benefit from follow-up calls, thereby optimizing resource allocation and ensuring that high-risk individuals receive timely support. By using predictive analytics to prioritize outreach, health care systems could enhance the effectiveness of follow-up programs while also managing staff workload more efficiently.

Additionally, natural language processing technologies could be integrated into call workflows to assist nurses in real-time. These tools could provide structured prompts to ensure comprehensive patient assessments, summarize patient concerns for documentation, or generate automated follow-up messages tailored to each patient's needs. For example, natural language processing could automatically flag patients with high LACE scores who have missed follow-ups, prompting timely outreach and helping prioritize those at greater risk of admission. Future studies should explore how AI can support, rather than replace, clinical decision-making—maintaining the essential human connection that characterizes high-quality transitional care. A hybrid model where AI improves workflow efficiency while nurses remain the primary point of engagement may offer an optimal balance between scalability and personalized care.

Understanding Mechanisms of Action and Patient-Centered Benefits

While this study demonstrated a reduction in ED visits, future research should explore the mechanisms underlying this effect. It remains unclear which components of the intervention, such as patient education, medication adherence support, emotional reassurance, or care coordination, or the simple act of receiving a call, contribute most to the observed improvements. Identifying the most influential elements would support refinement of the intervention to maximize its impact and cost-effectiveness. Involving patients in this refinement process through co-design of the call content or format can further enhance the relevance and effectiveness of the intervention, aligning it more closely with patient needs and preferences.

Beyond clinical outcomes, patient-reported experiences offer valuable insights into the intervention's effectiveness. Future qualitative research should examine how follow-up calls influence patient confidence in managing their recovery, their perceived quality of care, and any remaining unmet needs following discharge. Additionally, understanding the perspectives of nurses and other health care providers delivering follow-up calls can help identify workflow challenges and opportunities for improvement. Research in this area could inform best practices for scaling and sustaining postdischarge calls while also minimizing provider burden and enhancing patient engagement.

Promoting Equity in Postdischarge Support

Ensuring equitable access to follow-up care remains a critical challenge. Patients from lower socioeconomic backgrounds, those with limited English proficiency, and individuals without reliable phone access are at risk of being disproportionately excluded from telephone-based interventions. As a result, those who are most vulnerable to hospital readmission may not receive the support they need. To effectively reach and support diverse populations, follow-up care must also be culturally appropriate and safe. Embedding cultural safety into the design and delivery of interventions can help build trust and improve engagement across communities.

Future research should explore alternative and complementary strategies to reach underserved populations. These may include text-based follow-ups, video consultations, or partnerships with community health workers who can provide in-person support. Additionally, health systems should investigate more seamless integration of interpreter services into virtual care workflows and evaluate the potential of multilingual AI-driven chatbots to help bridge communications barriers. Addressing these disparities is essential to ensure that all patients, regardless of socioeconomic status, digital access, or language, can benefit equitably from postdischarge support.

Economic Analysis of Postdischarge Follow-Up Calls

A critical next step is to conduct a cost-effectiveness analysis to determine whether the observed reduction in ED visits results in net cost savings for the health care system. By evaluating factors such as reduced hospital readmissions, decreased ED use, and the cost of nursing time invested in delivering the intervention, an economic evaluation can guide policy decisions and resource allocation for broader implementation.

In addition to understanding overall cost-effectiveness, future studies should assess differential outcomes across patient subgroups. Certain populations may derive more substantial benefit from follow-up calls, informing a more targeted and efficient use of resources. Conducting subgroup analyses based on factors such as age, comorbidities, hospital discharge diagnosis, and social determinants of health could help refine intervention targeting. For example, patients with chronic conditions such as heart failure or chronic obstructive pulmonary disease may benefit more from medication adherence support, while those with lower health literacy may require additional educational reinforcement. Identifying which subgroups experience the greatest reduction in ED visits or readmissions will enable a more tailored approach, ensuring that resources are directed toward those who need them most.

Conclusions

This study suggests that proactive, digitally enabled outreach may support safer care transitions and reduce early ED utilization, even when delivered as a pragmatic, single-touch intervention within routine clinical operations. The findings highlight the potential for this low-cost, scalable intervention to support safer care transitions, particularly in healthcare systems where timely access to primary care remains a challenge. While the intervention did not significantly reduce hospital readmissions, its impact on ED use underscores its value in improving short-term outcomes and alleviating pressure on acute care services. Future studies should investigate whether certain patient subgroups derive greater benefit and explore ways to improve accessibility for populations with limited phone or digital access.

XSL•FO
RenderX

Acknowledgments

The authors thank the patients and the Fraser Health Virtual Care nurses that participated in this study.

Funding

This work was supported by the Fraser Health Strategic Priorities Grant and the Health Research Foundation of Innovative Medicines Canada Team Grant in Virtual Care. The funders had no role in the study design, data collection, analysis, interpretation, writing of the manuscript, or the decision to submit the paper for publication.

Data Availability

Due to conditions of the informed consent obtained from participants, the institutional and Ministry of Health ethical requirements do not permit us to share participant data from this study.

Authors' Contributions

SD contributed to conceptualization, design, methodology, data curation, writing of the manuscript, and project administration. AA contributed to data curation, project administration, investigation, data acquisition, data analysis, and writing of the manuscript. M Montenegro contributed to methodology, data curation, data analysis, software, and writing of the manuscript. M Maitland contributed to conceptualization, design, methodology, and funding acquisition. SC contributed to software and data analysis. SJ contributed to design, methodology, and project administration. HS contributed to design and methodology. DC contributed to investigation and review and editing of the manuscript. JMR contributed to writing. M MacPherson contributed to writing of the manuscript. DB contributed to the review and editing of the manuscript. MN-vV contributed to conceptualization, design, methodology, and supervision. All authors read and approved the final manuscript.

Conflicts of Interest

Two authors (MN-vV and SJ) are employed by Fraser Health Virtual Care, the program responsible for implementing the intervention examined in this study, and therefore may be perceived as having an interest in positive outcomes. All other authors declare no financial or nonfinancial competing interests.

Multimedia Appendix 1
Appendix S1: PRECIS-2 assessment of the post-discharge follow-up call trial. Appendix S2: Discharge Call Protocol Checklist. Appendix S3: Patient Experience Survey. Appendix S4: Intervention Group Characteristics (those Reached vs Not Reached). Appendix S5: Negative Binomial regression results for Hospital Readmission,
[DOCX File, 85 KB - jmir_v28i1e80529_app1.docx]

Checklist 1
CONSORT checklist.
[PDF File, 177 KB - jmir_v28i1e80529_app2.pdf]

References

1. Howard-Anderson J, Lonowski S, Vangala S, Tseng CH, Busuttil A, Afsar-Manesh N. Readmissions in the era of patient engagement. JAMA Intern Med 2014 Nov;174(11):1870-1872. [doi: 10.1001/jamainternmed.2014.4782] [Medline: 25264930]
2. Staples JA, Thiruchelvam D, Redelmeier DA. Site of hospital readmission and mortality: a population-based retrospective cohort study. CMAJ Open 2014 Apr;2(2):E77-E85. [doi: 10.9778/cmajo.20130053] [Medline: 25077133]
3. Sinden S. Is the path to better outcomes paved in gold? Physician incentives and hospital readmissions. Centre for Advancing Health Outcomes. 2024. URL: https://www.advancinghealth.ubc.ca/research-in-action/is-the-path-to-better-outcomes-paved-in-gold-physician-incentives-and-hospital-readmissions/ [accessed 2026-02-04]
4. Horwitz LI, Moriarty JP, Chen C, et al. Quality of discharge practices and patient understanding at an academic medical center. JAMA Intern Med 2013 Oct 14;173(18):1715-1722. [doi: 10.1001/jamainternmed.2013.9318] [Medline: 23958851]
5. Feigenbaum P, Neuwirth E, Trowbridge L, et al. Factors contributing to all-cause 30-day readmissions: a structured case series across 18 hospitals. Med Care 2012 Jul;50(7):599-605. [doi: 10.1097/MLR.0b013e318249ce72] [Medline: 22354212]
6. Shahid R, Shoker M, Chu LM, Frehlick R, Ward H, Pahwa P. Impact of low health literacy on patients' health outcomes: a multicenter cohort study. BMC Health Serv Res 2022 Sep 12;22(1):1148. [doi: 10.1186/s12913-022-08527-9] [Medline: 36096793]

XSL•FO
RenderX

7. Enard KR, Ganelin DM. Reducing preventable emergency department utilization and costs by using community health workers as patient navigators. J Healthc Manag 2013;58(6):412-427. [doi: 10.1097/00115514-201311000-00007] [Medline: 24400457]
8. Bamforth RJ, Chhibba R, Ferguson TW, et al. Strategies to prevent hospital readmission and death in patients with chronic heart failure, chronic obstructive pulmonary disease, and chronic kidney disease: a systematic review and meta-analysis. PLoS One 2021;16(4):e0249542. [doi: 10.1371/journal.pone.0249542] [Medline: 33886582]
9. Miller DA, Schaper AM. Implementation of a follow-up telephone call process for patients at high risk for readmission. J Nurs Care Qual 2015;30(1):63-70. [doi: 10.1097/NCQ.0000000000000069] [Medline: 24844916]
10. Shaw J, Abejirinde IOO, Agarwal P, Shahid S, Martin D. Digital health and equitable access to care. PLOS Digit Health 2024 Sep;3(9):e0000573. [doi: 10.1371/journal.pdig.0000573] [Medline: 39321176]
11. Tong L, George B, Crotty BH, et al. Telemedicine and health disparities: association between patient characteristics and telemedicine, in-person, telephone and message-based care during the COVID-19 pandemic. IPEM Transl 2022;3:100010. [doi: 10.1016/j.ipemt.2022.100010] [Medline: 36340828]
12. Kaihlanen AM, Virtanen L, Buchert U, et al. Towards digital health equity—a qualitative study of the challenges experienced by vulnerable groups in using digital health services in the COVID-19 era. BMC Health Serv Res 2022 Feb 12;22(1):188. [doi: 10.1186/s12913-022-07584-4] [Medline: 35151302]
13. Rodriguez JA, Betancourt JR, Sequist TD, Ganguli I. Differences in the use of telephone and video telemedicine visits during the COVID-19 pandemic. Am J Manag Care 2021 Jan;27(1):21-26. [doi: 10.37765/ajmc.2021.88573] [Medline: 33471458]
14. Badr J, Motulsky A, Denis JL. Digital health technologies and inequalities: a scoping review of potential impacts and policy recommendations. Health Policy 2024 Aug;146:105122. [doi: 10.1016/j.healthpol.2024.105122] [Medline: 38986333]
15. Thomas-Jacques T, Jamieson T, Shaw J. Telephone, video, equity and access in virtual care. NPJ Digit Med 2021 Nov 18;4(1):159. [doi: 10.1038/s41746-021-00528-y] [Medline: 34795356]
16. García E, Vincent BC, Thakur S, et al. Variables influencing telemedicine utilization via telephone appointments among rural patients. BMC Health Serv Res 2025 Jan 2;25(1):1. [doi: 10.1186/s12913-024-12122-5] [Medline: 39748425]
17. Harrison PL, Hara PA, Pope JE, Young MC, Rula EY. The impact of postdischarge telephonic follow-up on hospital readmissions. Popul Health Manag 2011 Feb;14(1):27-32. [doi: 10.1089/pop.2009.0076] [Medline: 21090991]
18. Mistiaen P, Poot E. Telephone follow-up, initiated by a hospital-based health professional, for postdischarge problems in patients discharged from hospital to home. Cochrane Database Syst Rev 2006 Oct 18;2006(4):CD004510. [doi: 10.1002/14651858.CD004510.pub3] [Medline: 17054207]
19. Hodalova S, Moore S, Dowds J, Murphy N, Martin-Loeches I, Broderick J. Feasibility of telephone follow-up after critical care discharge. Med Sci (Basel) 2020 Mar 14;8(1):16. [doi: 10.3390/medsci8010016] [Medline: 32183263]
20. Dudas V, Bookwalter T, Kerr KM, Pantilat SZ. The impact of follow-up telephone calls to patients after hospitalization. Dis Mon 2002 Apr;48(4):239-248. [doi: 10.1016/s0011-5029(02)90031-3] [Medline: 12021756]
21. Crocker JB, Crocker JT, Greenwald JL. Telephone follow-up as a primary care intervention for postdischarge outcomes improvement: a systematic review. Am J Med 2012 Sep;125(9):915-921. [doi: 10.1016/j.amjmed.2012.01.035] [Medline: 22938927]
22. Woods CE, Jones R, O'Shea E, Grist E, Wiggers J, Usher K. Nurse-led postdischarge telephone follow-up calls: a mixed study systematic review. J Clin Nurs 2019 Oct;28(19-20):3386-3399. [doi: 10.1111/jocn.14951] [Medline: 31162748]
23. Leppin AL, Gionfriddo MR, Kessler M, et al. Preventing 30-day hospital readmissions: a systematic review and meta-analysis of randomized trials. JAMA Intern Med 2014 Jul;174(7):1095-1107. [doi: 10.1001/jamainternmed.2014.1608] [Medline: 24820131]
24. Finlayson K, Chang AM, Courtney MD, et al. Transitional care interventions reduce unplanned hospital readmissions in high-risk older adults. BMC Health Serv Res 2018 Dec 12;18(1):956. [doi: 10.1186/s12913-018-3771-9] [Medline: 30541530]
25. Jayakody A, Bryant J, Carey M, Hobden B, Dodd N, Sanson-Fisher R. Effectiveness of interventions utilising telephone follow up in reducing hospital readmission within 30 days for individuals with chronic disease: a systematic review. BMC Health Serv Res 2016 Aug 18;16(1):403. [doi: 10.1186/s12913-016-1650-9] [Medline: 27538884]
26. Fraser Health Virtual Care. Fraser Health. URL: https://www.fraserhealth.ca/patients-and-visitors/fraser-health-virtual-care [accessed 2026-02-04]
27. van Walraven C, Dhalla IA, Bell C, et al. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. CMAJ 2010 Apr 6;182(6):551-557. [doi: 10.1503/cmaj.091117] [Medline: 20194559]
28. Gruneir A, Dhalla IA, van Walraven C, et al. Unplanned readmissions after hospital discharge among patients identified as being at high risk for readmission using a validated predictive algorithm. Open Med 2011;5(2):e104-e111. [Medline: 21915234]
29. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. BMJ 1995 Feb 18;310(6977):452-454. [doi: 10.1136/bmj.310.6977.452] [Medline: 7873954]
30. Kaya S, Sain Guven G, Aydan S, Toka O. Predictors of hospital readmissions in internal medicine patients: application of Andersen's model. Int J Health Plann Manage 2019 Jan;34(1):370-383. [doi: 10.1002/hpm.2648] [Medline: 30221793]

XSL•FO
RenderX

31. Kroch E, Duan M, Martin J, Bankowitz RA. Patient factors predictive of hospital readmissions within 30 days. J Healthc Qual 2016;38(2):106-115. [doi: 10.1097/JHQ.0000000000000003] [Medline: 26042742]
32. Wilhelmsen NC, Eriksson T. Medication adherence interventions and outcomes: an overview of systematic reviews. Eur J Hosp Pharm 2019 Jul;26(4):187-192. [doi: 10.1136/ejhpharm-2018-001725] [Medline: 31338165]
33. Allison K. Adherence through education: a call to clinicians to educate all patients on medication use. Ment Health Clin 2012;2(4):83-85. [doi: 10.9740/mhc.n117761]
34. Rosen OZ, Fridman R, Rosen BT, Shane R, Pevnick JM. Medication adherence as a predictor of 30-day hospital readmissions. Patient Prefer Adherence 2017;11:801-810. [doi: 10.2147/PPA.S125672] [Medline: 28461742]
35. Sheikh H, Brezar A, Dzwonek A, Yau L, Calder LA. Patient understanding of discharge instructions in the emergency department: do different patients need different approaches? Int J Emerg Med 2018 Feb 8;11(1):5. [doi: 10.1186/s12245-018-0164-0] [Medline: 29423767]
36. Hastings SN, Barrett A, Weinberger M, et al. Older patients' understanding of emergency department discharge information and its relationship with adverse outcomes. J Patient Saf 2011 Mar;7(1):19-25. [doi: 10.1097/PTS.0b013e31820c7678] [Medline: 21921863]
37. van Loon-van Gaalen M, van Winsen B, van der Linden MC, Gussekloo J, van der Mast RC. The effect of a telephone follow-up call for older patients, discharged home from the emergency department on health-related outcomes: a systematic review of controlled studies. Int J Emerg Med 2021 Feb 18;14(1):13. [doi: 10.1186/s12245-021-00336-x] [Medline: 33602115]
38. Aremu TO, Oluwole OE, Adeyinka KO, Schommer JC. Medication adherence and compliance: recipe for improving patient outcomes. Pharmacy (Basel) 2022 Aug 28;10(5):106. [doi: 10.3390/pharmacy10050106] [Medline: 36136839]
39. Starfield B, Shi L, Macinko J. Contribution of primary care to health systems and health. Milbank Q 2005;83(3):457-502. [doi: 10.1111/j.1468-0009.2005.00409.x] [Medline: 16202000]
40. Duong D, Vogel L. National survey highlights worsening primary care access. CMAJ 2023 Apr 24;195(16):E592-E593. [doi: 10.1503/cmaj.1096049] [Medline: 37094873]
41. Sawesi S, Rashrash M, Phalakornkule K, Carpenter JS, Jones JF. The impact of information technology on patient engagement and health behavior change: a systematic review of the literature. JMIR Med Inform 2016 Jan 21;4(1):e1. [doi: 10.2196/medinform.4514] [Medline: 26795082]
42. Pines JM, Hilton JA, Weber EJ, et al. International perspectives on emergency department crowding. Acad Emerg Med 2011 Dec;18(12):1358-1370. [doi: 10.1111/j.1553-2712.2011.01235.x] [Medline: 22168200]

Abbreviations

AI: artificial intelligence
CER: control event rate
ED: emergency department
EER: experimental event rate
EHR: electronic health record
EMR: electronic medical record
FHVC: Fraser Health Virtual Care
IRR: incident rate ratio
LACE: Length of stay, Acuity of admission, Comorbidities, and Emergency department use
LOS: length of stay
NNT: number needed to treat
SPSS: Statistical Package for the Social Sciences

Edited by A Stone, N Cahill; submitted 11.Jul.2025; peer-reviewed by LS Topaz, NC Favour; revised version received 19.Dec.2025; accepted 24.Dec.2025; published 17.Mar.2026.

Please cite as:
Du S, Anada A, Montenegro M, Maitland M, Chuan S, Jose S, Sihota H, Cooper D, Rivera J, Burditt D, Naruki-van Velzen M, MacPherson M
Effectiveness of Postdischarge Telephone Calls in Reducing Hospital Utilization: Quasi-Randomized Controlled Trial
J Med Internet Res 2026;28:e80529
URL: https://www.jmir.org/2026/1/e80529
doi:10.2196/80529

XSL•FO
RenderX

© Sandie Du, Abigail Anada, Maria Montenegro, Maurie Maitland, Sean Chuan, Sonia Jose, Harveer Sihota, Delia Cooper, Joanna Rivera, Dianne Burditt, Minetaro Naruki-van Velzen, Megan MacPherson. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effects of Self-Compassion and Mindfulness Interventions on Mental Health and Work-Related Outcomes Among Japanese Workers: Randomized Controlled Trial

Takumu Kurosawa*, PhD; Koichiro Adachi, MA; Ryu Takizawa*, MD, PhD

Department of Clinical Psychology, Graduate School of Education, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, Japan

*these authors contributed equally

Corresponding Author:
Ryu Takizawa, MD, PhD
Department of Clinical Psychology, Graduate School of Education, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, Japan

Abstract

Background: Mental health problems among workers are a significant global concern, leading to substantial economic losses, particularly due to presenteeism. While mindfulness and self-compassion practices have shown promise in improving psychological well-being and occupational outcomes, traditional interventions often require a high time commitment. Low-intensity digital interventions remain underexplored, especially in the context of occupational health.

Objective: This study aimed to develop and evaluate the effectiveness of a smartphone-based self-care app that enables independent practice of mindfulness and self-compassion without facilitator support. Specifically, it assessed the effects of self-compassion meditation (SCM) and mindfulness meditation (MM) on mental health and work-related outcomes among nonclinical Japanese workers.

Methods: This open-label, 3-arm randomized controlled trial recruited 300 working adults in Japan, who were randomly assigned to SCM (n=101), MM (n=100), or a waitlist control group (n=99). Participants in the SCM and MM groups engaged in daily guided meditation via a custom-built smartphone app over 4 weeks. Primary outcomes included psychological distress (Kessler Psychological Distress Scale-6) and work performance (World Health Organization Health and Work Performance Questionnaire and the Stanford Presenteeism Scale), while secondary outcomes encompassed cognitive flexibility, self-compassion, perceived stress, work engagement, psychological safety, and self-perceived creativity. Assessments were conducted at baseline, postintervention (4 wk), and 1-month follow-up (8 wk). Recruitment occurred in 2 waves (November 2022-December 2022 and June 2023), with follow-up assessments completed by September 2023.

Results: A total of 300 participants (mean age of 35.44, SD 9.14 y; n=180, 60% female) were randomized. Adherence was high in both intervention groups (mean completed days: SCM 23.30, SD 5.58; MM 22.95, SD 6.25), with an overall dropout rate of 21.4%. No significant group × time interactions were detected for most outcomes, although significant main effects of time were observed for several measures, including work performance and self-compassion, with small effect sizes. Within-group analyses suggested broader improvements in the SCM group, whereas changes in the MM group were more limited. In sensitivity analyses using linear mixed-effects models, a significant group × time interaction was detected for work performance, with a larger pre-post improvement in SCM.

Conclusions: This low-intensity, fully self-guided, smartphone-based meditation program demonstrated good feasibility, as reflected by high adherence and relatively low attrition. Between-group effects were limited; therefore, the findings should be interpreted as preliminary and do not support strong causal claims of intervention superiority. Clinical significance remains unclear in this nonclinical sample and should be evaluated in future studies.

Trial Registration: UMIN-CTR UMIN000049466; https://center6.umin.ac.jp/cgi-open-bin/ctr_e/ctr_view.cgi?recptno=R000056324

International Registered Report Identifier (IRRID): RR2-10.2196/53541

(J Med Internet Res 2026;28:e79991) doi:10.2196/79991

KEYWORDS

self-compassion; mindfulness; mobile apps; digital health; occupational health; mental health; randomized controlled trial; RCT; work performance

XSL•FO
RenderX

Introduction

Mental health problems among workers pose a critical challenge with far-reaching implications for the global economy. According to the World Health Organization (WHO), depression and anxiety disorders account for the loss of approximately 12 billion workdays annually, resulting in an annual economic burden of approximately US $1 trillion [1]. A survey of 10,000 workers in Japan identified mental illnesses, stiff shoulders, and lower back pain as health problems affecting job performance. Among these, presenteeism—employees being present at work but unable to perform at full capacity due to mental illness—causes the most significant economic losses [2]. These findings underscore the need for initiatives that improve workers' mental health and productivity.

Mindfulness and self-compassion are effective approaches for supporting workers' mental health. These approaches are often categorized within the "third wave" of cognitive-behavioral therapies, which emphasize context- and acceptance-based strategies aimed at altering one's relationship with internal experiences rather than primarily modifying cognitive content [3]. Mindfulness is the practice of maintaining nonjudgmental awareness of one's thoughts, emotions, bodily sensations, and surrounding environment in the present moment [4], while self-compassion entails being warm and understanding toward ourselves when we suffer, fail, or feel inadequate, rather than ignoring our pain or flagellating ourselves with self-criticism [5]. Mindfulness practice has been shown to enhance focus at work, self-fulfillment, and sleep quality, while reducing emotional exhaustion, occupational stress, and anxiety [6-8]. It has also been reported to improve job performance, workplace relationships, and creativity [9,10]. Similarly, self-compassion has been found to alleviate stress and burnout and to improve well-being and resilience. However, further research is needed to clarify its specific effects on work-related outcomes [11].

Mindfulness and self-compassion both use meditation as a core intervention strategy and share the goal of fostering nonjudgmental acceptance of one's experiences. However, self-compassion places additional emphasis on responding to difficulty with self-kindness and a sense of common humanity [12]. Furthermore, mindfulness-based interventions may cultivate self-compassion more implicitly than self-compassion–focused programs, which teach these elements explicitly [13]. Elevated levels of self-compassion have been correlated with more adaptive emotion regulation and enhanced well-being [14], and interventions aimed at increasing self-compassion have proven effective in mitigating self-criticism [15]. Within the employee samples, self-compassion has been linked to job-related outcomes, including performance indicators [16], and its related capacities may contribute to the development of psychologically safer environments and promote learning-oriented behaviors within teams [17]. When implementing short-term, low-intensity interventions for busy workers, it is crucial to select the most effective components. Accordingly, directly comparing mindfulness meditation (MM) and self-compassion meditation (SCM) may help clarify which elements are particularly relevant for improving work-related outcomes.

Traditional mindfulness and self-compassion programs typically consist of 8 weeks of in-person sessions with daily practice, which has been identified as a barrier to participation for busy workers [4,5,18]. In contrast, digital interventions using websites or smartphone apps have been shown to be effective in reducing stress and improving well-being, while requiring less time and cost [19,20]. These low-intensity interventions provide a more accessible alternative for users. Digital interventions exhibit varying levels of human support, ranging from fully self-administered programs to guided formats that incorporate contact with a coach or therapist [21]. Although unguided interventions offer high scalability and low cost, they tend to show lower adherence and reduced efficacy compared to guided approaches [21]. In the context of occupational digital mental health interventions, additional support and increased intensity may improve engagement and outcomes, albeit at the expense of scalability and increased implementation demands [22]. Consequently, engagement with mindfulness-based mobile apps varies widely, underscoring the importance of evaluating feasibility and reporting engagement transparently, especially in fully self-guided delivery [23]. However, despite their growing popularity, these low-intensity digital interventions remain underrepresented in randomized controlled trials (RCTs), underscoring the need for further empirical research. Additionally, the COVID-19 pandemic has accelerated the adoption of online psychological support, emphasizing the importance of developing accessible digital tools to assist workers, especially as remote work becomes increasingly prevalent [24]. In response to this need, this study aims to develop a self-care app that facilitates independent practice of mindfulness and self-compassion and to evaluate its effectiveness using mental health and work-related outcomes through an RCT design.

Methods

Study Design

This open-label, 3-arm RCT compared the effects of 2 intervention groups—the SCM and MM groups—with a waitlist control group (WCG). Both intervention groups engaged in guided meditation sessions delivered through a mobile app developed specifically for the study. Further details are provided in the study protocol [25].

Participants

The inclusion criteria were employees who were (1) working more than 20 hours per week, (2) aged between 18 and 54 years, (3) not on a leave of absence, (4) not business owners or students, and (5) not currently diagnosed with a mental disorder and scoring below 13 on the Kessler Psychological Distress Scale-6 [26]. The smartphone app used in this study was only available on iOS, restricting participation to iPhone users. The study involved 300 workers in Japan via open calls from websites. Participants were required to have basic digital literacy to install and use the smartphone app and to complete web-based questionnaires. The participants were not informed of their group allocation.

XSL•FO
RenderX

Ethical Considerations

This study was approved by the ethics committee of the University of Tokyo (22-326). The study was registered with the University Hospital Medical Information Network Clinical Trials Registry (UMIN000049466), and the study protocol was published previously [25]. Electronic informed consent was obtained via a web-based form prior to eligibility screening and randomization, and participants could withdraw at any time without penalty. Survey responses and app log data were managed and analyzed using study-generated participant IDs. Identifying information was stored separately from study data; access to the data was restricted to the research team, and the linkage file was stored offline in a secure location. Participants received a gift card worth up to JP ¥10,000 (approximately US $63; JP ¥158=US $1 as of January 11, 2026), depending on completion of study procedures; compensation was not contingent on adherence to the intervention (eg, the number of meditation days completed).

Procedure

Participants were recruited via open calls during 2 recruitment waves (November 2022-December 2022 and June 2023) through a psychology experiment recruitment website and our laboratory's website, due to insufficient enrollment in the first wave. Study details and contact information were available on website posters. Interested individuals completed a web form with a screening survey (Kessler Psychological Distress Scale-6) and were screened based on the inclusion criteria. The baseline assessment (T0), the posttreatment assessment at 4 weeks (T1), and the follow-up assessment at 8 weeks (T2) were assessed using self-report questionnaires. Questionnaires at each time point were administered independently of the intervention through an external web-based form, rather than within the app, with survey links distributed to participants via email. For the SCM and MM groups, T2 (8 wk) represents the 1-month follow-up after the 4-week intervention period. As assessments were scheduled relative to each participant's intervention timeline, the exact calendar dates varied by participant; however, all follow-up (T2) assessments were completed by September 2023. No significant methodological changes were implemented after trial commencement; any deviations from the protocol regarding outcome measures are described in the "Methods" section.

Randomization and Blinding

After obtaining informed consent and screening participants, those who met the eligibility criteria were randomly assigned to 1 of the 3 groups (1:1:1) via a computerized blocked randomization scheme (block size=15). This randomization was conducted independently by research staff, and participants were immediately assigned to either the intervention (SCM or MM) or WCG. The intervention commenced approximately 1 week after allocation, while the waitlist group received no specific instructions and waited for 4 weeks. Participants received group-specific instructions via automated email. As the participants followed a meditation guide, blinding was not possible, thus making this an open-label RCT. The instructions were sent via automated emails, and although the participants could infer their group, they were not explicitly informed of their assignment.

Interventions

The SCM and MM courses were delivered through a smartphone app developed by the research team. Participants assigned to the SCM and MM groups used the app to engage in guided meditation once daily for 4 weeks. Each course comprised 3 modules that integrated guided self-help meditation and psychoeducation. The guided meditations in weeks 1 and 2 were identical across both courses: a 7-minute breathing meditation in week 1 and a 6-minute short body scan in week 2. In weeks 3 and 4, SCM participants practiced a 12-minute loving-kindness meditation focused on self-kindness and common humanity, while MM participants engaged in a 12-minute breath, sound, and body meditation. The guided meditations were adapted from scripts developed by the University of California, Los Angeles, Mindful Awareness Research Center and modified for a Japanese audience. Psychoeducation covered topics such as stress responses, mindfulness, and self-compassion and was provided as the studies were accessible via the app at any time. Weekly email reminders encouraged participation, though no minimum time commitment was required. Meditation practice was systematically recorded by the app through data logging, and engagement metrics were derived from these logs. The intervention was fully automated and self-directed, with no support from a human instructor or therapist. Participants could view their completed practice days and total practice time within the app.

We did not instruct participants to use the app in a specific location and did not collect data on the settings where sessions were completed (eg, home vs workplace); participants were advised to practice in a quiet, private environment without interruptions. Participants in the WCG received no intervention during the 4-week period; afterward, they were allocated to either the SCM or MM course for an additional 4 weeks of guided meditation. Because WCG participants were offered access to an intervention after the 4-week waiting period for ethical reasons, they did not remain in a controlled condition during weeks 4 to 8 and were therefore not included in the 8-week (T2) follow-up analyses.

Outcome Measures

This study aimed to investigate the efficacy of a brief, self-administered smartphone-based meditation program, focusing on self-compassion and mindfulness, in enhancing workers' mental health and work-related functioning. Therefore, the primary outcomes were defined as (1) psychological distress, which is a direct target of the intervention; and (2) work performance (presenteeism), a pragmatic occupational outcome. As secondary outcomes, we assessed potentially modifiable psychological processes and related indicators (eg, cognitive flexibility, self-compassion, perceived stress, and work engagement) to complement the evaluation of changes in mental health and work-related outcomes. Although daily physiological measures (including heart rate and heart rate variability) were planned in the protocol, they are not reported in this paper because this report focuses on psychological and work-related

XSL•FO
RenderX

outcomes; these physiological outcomes will be reported separately.

Primary Outcomes

The primary outcomes were psychological distress and work performance, assessed at 3 time points: baseline (T0), posttreatment at 4 weeks (T1), and follow-up at 8 weeks (T2). Psychological distress was measured using the Japanese version of the Kessler Psychological Distress Scale-6 [26,27], a 6-item instrument designed to measure nonspecific psychological distress (depressive mood and anxiety) over the preceding 30 days (total score range: 0 - 24; higher scores indicate greater distress), with good internal consistency in this study (Cronbach α at T0=0.85). A cutoff score of 12/13 on the Kessler Psychological Distress Scale-6 has been proposed for screening severe depressive symptoms [28], and thus, an ethical exclusion criterion of Kessler Psychological Distress Scale-6 ≥13 was established. Work performance was evaluated using one presenteeism item from the Japanese version of the World Health Organization Health and Work Performance Questionnaire (WHO-HPQ) [29,30], which assesses overall work performance over the past 28 days (absolute presenteeism score range: 0 - 100; higher scores indicate better performance). Moreover, work inefficiency was assessed using the work impairment score from the Stanford Presenteeism Scale (SPS) [31,32], which captures the extent to which health-related cognitive, emotional, and behavioral difficulties impede task completion and focus (10 items; 1="always" to 5="never"; scored such that higher scores indicate greater impairment). This scale exhibited acceptable internal consistency (Cronbach α=0.72) in this study. While the protocol specified presenteeism assessed using the WHO-HPQ, this report additionally includes the SPS work impairment score to capture health-related work impairment.

Secondary Outcomes

The following self-reported questionnaires were administered as secondary outcome measures at each time point: the Cognitive Flexibility Scale (Cronbach α=0.86) [33,34], which assesses the ability to generate alternative responses and adapt behavior to changing circumstances (12 items; total score range: 12 - 72; higher scores indicate greater cognitive flexibility); the Self-Compassion Scale (SCS; Cronbach α=0.94) [35,36], which assesses compassionate self-responding toward oneself across 6 facets (26 items; total score range: 26 - 130; higher scores indicate greater self-compassion), including its subscales: self-kindness (Cronbach α=0.92), common humanity (Cronbach α=0.80), mindfulness (Cronbach α=0.79), self-judgment (Cronbach α=0.89), isolation (Cronbach α=0.84), and overidentification (Cronbach α=0.81); the Perceived Stress Scale-14 (Cronbach α=0.89) [37-39], which assesses the extent to which situations in the past month are perceived as unpredictable, uncontrollable, and overloaded (total score range: 0 - 56; higher scores indicate greater perceived stress); and the Utrecht Work Engagement Scale-9 (Cronbach α=0.95) [40-42], which assesses vigor, dedication, and absorption (total score range: 0 - 54; higher scores indicate higher engagement), including its subscales: vigor (Cronbach α=0.89), dedication (Cronbach α=0.84), and absorption (Cronbach α=0.88). Additionally, psychological safety was assessed using the Psychological Safety Scale (7 items; Cronbach α=0.80) [43,44], which reflects the extent to which individuals can express themselves at work without fear of negative consequences; items are rated from 1 ("not at all") to 7 ("strongly agree"), yielding total scores of 7 to 49. Creativity (idea generation at work) was assessed using the Self-Perceived Creativity Scale (3 items; Cronbach α=0.91) [45], rated from 1 ("do not agree") to 7 ("agree"), yielding total scores of 3 to 21. For further details on these outcome measures, refer to the study protocol [25].

Other Measures

Demographic variables were assessed using a questionnaire at baseline (T0). The collected variables included age, gender, employment status, marital status, educational background, and prior meditation experience, including mindfulness practices.

Sample Size

In the protocol, the sample size was planned based on detecting a small group-by-time interaction effect (Cohen f=0.15) for the primary pre-post comparison across the 3 arms, assuming a correlation of 0.50 between repeated measures, with 80% power at a 2-sided α=.05 (G*Power; Heinrich Heine University Düsseldorf) [25]. This yielded a required sample size of 37 participants per arm. The protocol, therefore, initially planned to randomize 200 participants to account for expected attrition. During the trial conduct, we observed that a proportion of participants did not complete the baseline assessment following randomization; therefore, recruitment was extended and a total of 300 participants were randomized to ensure an adequate evaluable sample size at key assessment time points and to improve the precision of effect estimates. This target sample size was reflected in the trial registration (UMIN000049466). The increase in sample size was an operational decision made without reference to intervention effects, and no interim efficacy analyses were conducted.

Statistical Analyses

Missing Data and Sensitivity Analysis

Analyses were conducted using available cases at each assessment point. Participants with missing data for a specific analysis were excluded from that analysis, resulting in varying sample sizes across time points. As a sensitivity analysis, linear mixed-effects models (LMMs) were fitted using all available data to assess the robustness of the findings in relation to attrition.

Baseline Analyses

Demographic variables and baseline psychological indices were summarized descriptively (mean and SD for continuous variables; n and % for categorical variables) for each group. In accordance with CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) guidance (Checklist 1), we did not perform significance testing for baseline group comparisons.

XSL•FO
RenderX

Analysis of Preintervention and Postintervention Effects

Changes before and after the intervention or waitlist periods were assessed using 2-tailed paired *t* tests within each intervention group. Effect sizes were calculated as Cohen *d* and interpreted using cutoff values of 0.20, 0.50, and 0.80 for small, medium, and large effects, respectively [38].

Analysis of Effect Size in Group Comparisons

To examine preintervention and postintervention changes and pre-post differences between groups, assessment questionnaire results were analyzed using a 2 (time: pre and post) × 3 (group) repeated-measures ANOVA. Effect sizes for ANOVA were reported as generalized η^2, with cutoff values of 0.01, 0.06, and 0.14 for small, medium, and large effects, respectively [38,46].

Analysis of 1-Month Follow-Up

As the waitlist group transitioned to receive an intervention after T1, all T2 follow-up analyses were restricted to participants initially allocated to SCM or MM. The 1-month follow-up scores were analyzed using a series of one-way within-subjects ANOVAs to compare scores within each intervention group across three time points: (1) preintervention, (2) postintervention, and (3) 1-month follow-up. For measures that showed significant time effects, Bonferroni-adjusted pairwise comparisons were conducted to examine changes within each intervention group.

To assess the maintenance of treatment effects and differences between the intervention groups, a 3 (time: pre, post, and follow-up) × 2 (group: SCM and MM) repeated-measures ANOVA was performed. For measures that showed significant time effects, Bonferroni-adjusted pairwise comparisons were conducted to investigate changes across the intervention groups.

Sensitivity Analyses Using LMMs

Primary inference in this trial was derived from between-group comparisons (group × time). Within-group tests in the main analyses were conducted as exploratory analyses to describe within-group patterns and should not be interpreted as evidence of causal superiority. As a sensitivity analysis, we reanalyzed the primary and secondary outcomes using LMMs to assess the robustness of the findings. These models included fixed effects for group, time, and their interaction, with a random intercept for participants to account for within-subject correlations. Additionally, variables showing notable baseline imbalances were included as covariates in sensitivity analyses to assess the robustness of the findings. The models were fitted using all available data under a missing-at-random assumption. For the pre-post comparison (T0-T1), analyses included all randomized groups (SCM, MM, and WCG), whereas follow-up analyses (T0-T1-T2) were restricted to SCM and MM because the WCG crossed over after T1.

Results

Participant Flow and Sample Characteristics

In total, 300 participants were randomly allocated to SCM (n=101), MM (n=100), and WCG (n=99; Figure 1). Of these 300 randomized participants, 251 completed the baseline assessment (T0; 83.7%), including 86 (85.1%) in SCM, 85 (85%) in MM, and 80 (80.8%) in WCG, and were included in Table 1. At posttreatment (T1), outcome data were available for 234 participants (78%; SCM: n=79, 78.2%; MM: n=79, 79%; WCG: n=76, 76.8%). At follow-up (T2), outcome data were available for 136 (67.7% of the SCM and MM) participants in the intervention arms (SCM: n=70, 69.7%; MM: n=66, 66%). No important harms or unintended effects (including major technical problems or privacy incidents) were reported to the study team during the trial. Among those who completed the baseline assessment (Table 1), the sample consisted of 60% (n=180) female participants, with a high proportion of regular employees (n=57, 67.1%). The mean age was 35.44 (SD 9.14) years. The groups were generally similar at baseline; however, age and prior meditation experience showed some imbalance across groups. Therefore, age (grand-mean centered) and prior meditation experience were included as covariates in the sensitivity analyses using LMMs (Figure 1).

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram of participant enrollment, allocation, and follow-up assessments. MM: mindfulness meditation; SCM: self-compassion meditation; WCG: waitlist control group.

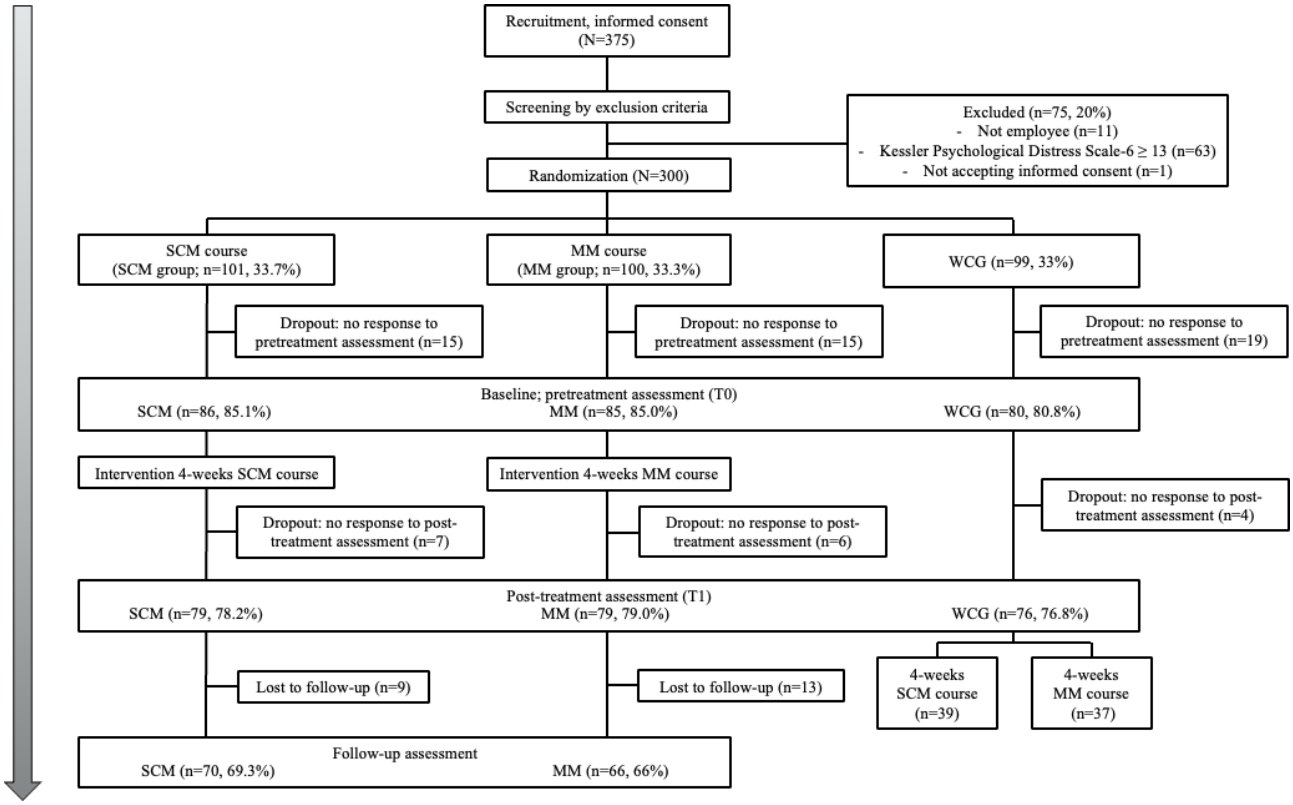

XSL•FO
RenderX

Table . Baseline demographic characteristics of participants by group.

Variable	SCM[a] group (n=86)	MM[b] group (n=85)	Waitlist group (n=80)
Age (y), mean (SD)	33.96 (8.84)	35.02 (9.65)	37.45 (8.65)
Sex, n (%)			
Female	51 (59.3)	47 (55.3)	52 (65)
Male	34 (39.5)	38 (44.7)	28 (35)
Employment status, n (%)			
Regular	59 (68.6)	57 (67.1)	52 (65)
Nonregular	27 (31.4)	27 (31.8)	28 (35)
Marital status, n (%)			
Single	50 (58.1)	52 (61.2)	35 (43.8)
Married	35 (40.7)	32 (37.6)	45 (56.2)
Educational attainment, mean (SD)	15.70 (1.67)	15.88 (1.72)	15.79 (1.63)
Previous meditation experience, n (%)			
Yes	30 (34.9)	18 (21.2)	15 (18.8)
No	56 (65.1)	66 (77.6)	64 (80)
Psychological distress, mean (SD)	7.05 (4.51)	6.23 (4.96)	6.62 (4.92)
Work performance, mean (SD)	62.79 (18.32)	59.88 (18.16)	60.25 (20.68)
Work inefficiency, mean (SD)	26.46 (5.21)	26.89 (6.02)	26.13 (4.92)
Cognitive flexibility, mean (SD)	47.28 (8.43)	46.01 (8.46)	47.00 (9.11)
Self-compassion, mean (SD)	76.39 (17.43)	74.14 (19.46)	79.85 (18.73)
Self-kindness, mean (SD)	15.74 (4.55)	15.02 (4.89)	16.16 (4.81)
Self-judgment, mean (SD)	15.80 (4.82)	15.79 (4.76)	14.83 (5.26)
Common humanity, mean (SD)	10.89 (3.72)	10.85 (3.76)	11.49 (3.80)
Isolation, mean (SD)	10.81 (4.05)	11.74 (4.04)	10.74 (4.44)
Mindfulness, mean (SD)	12.47 (3.07)	11.88 (3.39)	12.81 (3.18)
Overidentification, mean (SD)	13.52 (3.55)	13.75 (3.86)	13.46 (3.72)
Perceived stress, mean (SD)	28.50 (9.58)	28.75 (10.81)	28.50 (9.47)
Work engagement, mean (SD)	28.41 (13.51)	26.14 (12.30)	27.69 (12.20)
Vigor, mean (SD)	8.64 (4.72)	7.79 (4.51)	8.66 (4.52)
Dedication, mean (SD)	10.62 (4.53)	9.98 (4.10)	10.28 (4.14)
Absorption, mean (SD)	9.29 (4.95)	8.38 (4.41)	8.57 (4.47)
Psychological safety, mean (SD)	35.30 (6.65)	34.29 (7.44)	33.60 (7.93)
Creativity, mean (SD)	12.19 (4.37)	11.67 (4.35)	11.36 (4.09)

[a]SCM: self-compassion meditation.

[b]MM: mindfulness meditation.

The rate of participants who dropped out before posttreatment assessment did not differ among groups (χ^2_2=0.66; *P*=.72). On average, the participants practiced for 23.30 (SD 5.58) days in the SCM group and 22.95 (SD 6.25) days in the MM group during the intervention period. The 2-tailed *t* tests showed no significant differences in the number of practice days between the 2 groups (t_{154}=0.38; *P*=.71).

Pre-Post Differences Between the Groups

Table 2 summarizes the mean and SD for each group, along with preintervention and postintervention comparisons using 2-tailed paired *t* tests within each intervention group. In the SCM group, paired *t* tests revealed significant improvements in work performance, cognitive flexibility, self-compassion, common humanity (as measured by the SCS), and psychological safety. Significant reductions were observed in work inefficiency, perceived stress, self-judgment, and overidentification (as measured by the SCS). In the MM group,

XSL•FO
RenderX

2-tailed paired *t* tests showed a significant improvement in common humanity (as measured by the SCS). Table 3 presents the results of between-group comparisons using ANOVA. No significant interaction effect between group and time was found.

XSL•FO
RenderX

Table . Mean (SD) scores for intervention and control groups at preintervention and postintervention and changes in scores in each intervention group.

	Preintervention, mean (SD)	Postintervention, mean (SD)	*t* test (*df*)	*P* value	Cohen *d*
Psychological distress					
SCM[a] group	6.68 (4.27)	6.90 (4.40)	0.49 (77)	.62	0.06
MM[b] group	5.94 (4.92)	6.38 (5.08)	1.05 (77)	.30	0.09
Waitlist group	6.75 (4.76)	7.12 (5.44)	1.14 (73)	.26	0.09
Work performance					
SCM group	63.29 (16.93)	70.00 (15.85)	4.28 (78)	<.001	0.41
MM group	59.37 (17.27)	60.25 (18.81)	0.49 (78)	.62	0.05
Waitlist group	59.87 (21.07)	62.93 (20.58)	1.56 (74)	.12	0.15
Work inefficiency					
SCM group	26.22 (5.04)	24.85 (5.52)	–2.21 (75)	.03	–0.27
MM group	26.78 (5.99)	25.83 (5.81)	–1.34 (72)	.18	–0.14
Waitlist group	26.15 (5.02)	25.81 (6.00)	–1.11 (71)	.27	–0.09
Cognitive flexibility					
SCM group	47.69 (8.18)	49.04 (8.73)	2.27 (71)	.03	0.18
MM group	45.81 (8.49)	45.86 (9.09)	0.11 (76)	.91	0.01
Waitlist group	46.84 (9.31)	46.85 (9.32)	0.26 (73)	.80	0.01
Self-compassion					
SCM group	76.30 (17.94)	81.17 (18.06)	2.86 (65)	.006	0.24
MM group	74.55 (19.27)	77.32 (20.11)	1.91 (71)	.06	0.10
Waitlist group	79.42 (19.11)	79.70 (19.71)	1.33 (70)	.19	0.07
Self-kindness					
SCM group	15.79 (4.66)	16.56 (4.77)	1.68 (73)	.10	0.15
MM group	15.12 (4.91)	15.48 (4.99)	0.80 (77)	.42	0.05
Waitlist group	16.28 (4.90)	16.33 (5.01)	0.15 (75)	.88	0.01
Self-judgment					
SCM group	15.70 (4.96)	14.77 (4.38)	–2.36 (74)	.02	–0.21
MM group	15.68 (4.74)	15.42 (4.92)	–0.83 (77)	.41	–0.05
Waitlist group	15.13 (5.22)	14.69 (4.92)	–1.29 (74)	.20	–0.09
Common humanity					
SCM group	10.67 (3.78)	11.47 (3.98)	2.39 (75)	.02	0.19
MM group	10.97 (3.73)	11.48 (3.75)	2.19 (75)	.03	0.14
Waitlist group	11.56 (3.85)	11.63 (3.81)	0.21 (74)	.83	0.02
Isolation					
SCM group	10.64 (4.10)	10.14 (3.84)	–1.65 (74)	.10	–0.14
MM group	11.51 (3.93)	11.23 (3.81)	–0.85 (78)	.40	–0.07
Waitlist group	10.92 (4.47)	10.91 (4.27)	–0.04 (75)	.97	0.00
Mindfulness					
SCM group	12.38 (3.13)	12.78 (3.53)	1.19 (76)	.24	0.11
MM group	11.94 (3.34)	12.14 (3.55)	0.58 (77)	.56	0.05
Waitlist group	12.80 (3.24)	13.01 (3.16)	0.90 (74)	.37	0.07
Overidentification					

XSL•FO
RenderX

	Preintervention, mean (SD)	Postintervention, mean (SD)	*t* test (*df*)	*P* value	Cohen *d*
SCM group	13.42 (3.57)	12.74 (3.73)	–2.27 (76)	.03	–0.20
MM group	13.68 (3.89)	13.37 (3.86)	–1.19 (76)	.24	–0.08
Waitlist group	13.57 (3.73)	13.49 (4.12)	–0.31 (73)	.76	–0.02
Perceived stress					
SCM group	28.31 (9.86)	25.19 (8.52)	–2.65 (75)	.01	–0.31
MM group	28.46 (10.69)	27.05 (9.65)	–1.27 (77)	.21	–0.13
Waitlist group	28.99 (9.35)	27.61 (9.54)	–1.39 (67)	.17	–0.11
Work engagement					
SCM group	28.21 (13.55)	29.51 (13.29)	1.32 (73)	.19	0.07
MM group	25.58 (12.04)	26.03 (12.72)	0.52 (78)	.60	0.04
Waitlist group	27.61 (12.07)	28.08 (13.36)	0.79 (73)	.43	0.05
Vigor					
SCM group	8.54 (4.73)	8.96 (4.60)	1.17 (77)	.25	0.08
MM group	7.54 (4.41)	8.04 (4.54)	1.44 (78)	.15	0.11
Waitlist group	8.71 (4.46)	8.84 (4.97)	0.67 (74)	.51	0.05
Dedication					
SCM group	10.59 (4.58)	10.87 (4.55)	1.31 (77)	.19	0.07
MM group	9.84 (4.03)	9.87 (4.30)	0.13 (78)	.90	0.01
Waitlist group	10.28 (4.10)	10.50 (4.43)	0.82 (75)	.42	0.05
Absorption					
SCM group	9.22 (4.95)	9.72 (4.77)	1.27 (75)	.21	0.09
MM group	8.20 (4.35)	8.11 (4.62)	–0.25 (78)	.80	–0.02
Waitlist group	8.60 (4.43)	8.74 (4.78)	0.31 (74)	.76	0.02
Psychological safety					
SCM group	35.40 (6.83)	36.71 (6.73)	2.12 (75)	.04	0.17
MM group	34.34 (7.38)	34.23 (7.79)	–0.18 (78)	.86	–0.01
Waitlist group	33.57 (8.01)	33.80 (9.10)	0.48 (74)	.64	0.04
Creativity					
SCM group	12.17 (4.33)	12.45 (4.46)	1.09 (73)	.28	0.09
MM group	11.55 (4.38)	11.08 (4.26)	–1.18 (77)	.24	–0.10
Waitlist group	11.41 (4.18)	11.65 (4.73)	0.76 (74)	.45	0.06

[a]SCM: self-compassion meditation.

[b]MM: mindfulness meditation.

XSL•FO
RenderX

Table . Group × time interaction effect using ANOVA.

Variable	*F* test (*df*)	*P* value	General η^2
Psychological distress	0.09 (2, 227)	.91	0.000
Work performance	2.67 (2, 230)	.07	0.004
Work inefficiency	0.61 (2, 218)	.54	0.001
Cognitive flexibility	1.70 (2, 220)	.19	0.001
Self-compassion	1.73 (2, 206)	.18	0.001
Self-kindness	0.85 (2, 225)	.43	0.001
Self-judgment	1.19 (2, 225)	.31	0.001
Common humanity	1.36 (2, 224)	.26	0.001
Isolation	0.68 (2, 227)	.51	0.001
Mindfulness	0.15 (2, 227)	.86	0.000
OverIdentification	1.14 (2, 225)	.32	0.001
Perceived stress	1.07 (2, 219)	.34	0.002
Work engagement	0.11 (2, 224)	.90	0.000
Vigor	0.16 (2, 229)	.85	0.000
Dedication	0.28 (2, 230)	.76	0.000
Absorption	0.54 (2, 227)	.58	0.000
Psychological safety	0.99 (2, 227)	.37	0.001
Creativity	1.61 (2, 224)	.20	0.002

As a supplementary analysis, the results of the LMM for repeated measures are shown in Table S1 in Multimedia Appendix 1. A significant group × time interaction was detected for work performance ($F_{2,233}$=3.32; *P*=.04; partial η^2=0.028). A simple slope analysis revealed that participants in the SCM group showed a significant change (t_{237}=3.99; *P*<.001), whereas no significant difference was observed in the MM group (t_{237}=0.39; *P*=.70) and the WLC group (t_{237}=1.63; *P*=.11).

After adjusting for age and previous experience with meditation, significant group × time interactions were detected for work performance ($F_{2,230}$=3.22; *P*=.04; partial η^2=0.027) and cognitive flexibility ($F_{2,223}$=3.13; *P*=.046; partial η^2=0.027). Using simple slope analyses, participants in the SCM group showed a significant change in work performance (t_{234} = 3.93; *P*<.001) and cognitive flexibility (t_{224} = 2.87; *P*=.004), whereas no significant difference was observed in the MM and WLC groups.

Follow-Up Analysis

Table 4 summarizes the means and SDs of preintervention, postintervention, and follow-up measures in each intervention group. In the SCM group, significant main effects of time using ANOVA were observed for work performance ($F_{2,138}$=6.42; *P* = .002; η^2G=0.029), self-compassion ($F_{2,108}$=7.38; *P* <.001; η^2G=0.020), self-judgment ($F_{2,130}$=6.19; *P*=.003; η^2G=0.016), common humanity ($F_{2,134}$=5.72; *P*=.004; η^2G=0.012), overidentification ($F_{2,130}$=3.38; *P*=.04; η^2G=0.008), perceived stress ($F_{2,116}$=4.43; *P*=.01; η^2G=0.021), and psychological safety ($F_{2,118}$=4.39; *P*=.01; η^2G=0.013). In the SCM group, Bonferroni-adjusted simple comparisons revealed a significant increase in work performance from preintervention to postintervention (*P*<.001) and a significant decrease from postintervention to follow-up (*P*=.03). Significant increases were also observed in self-compassion from preintervention to postintervention (*P*=.03) and from preintervention to follow-up (*P*=.002), common humanity from preintervention to follow-up (*P*=.006), and psychological safety from postintervention to follow-up (*P*=.002). Significant decreases were found in self-judgment from preintervention to follow-up (*P*=.001), and perceived stress from preintervention to follow-up (*P*=.03). In the MM group, significant main effects of time using ANOVA were observed for self-compassion ($F_{2,108}$=3.35; *P*=.04; η^2G=0.004) and overidentification ($F_{2,126}$=4.28; *P*=.02; η^2G=.007). Bonferroni-adjusted simple comparisons indicated a significant decrease in overidentification from preintervention to follow-up (*P*=.01).

XSL•FO
RenderX

Table . Mean (SD) at follow-up and repeated-measures analyses across pre, post, and follow-up within the intervention groups.

	Follow-up assessment	Repeated measure ANOVA		
	Mean (SD)	*F* test (*df*)	*P* value	General η^2
Psychological distress				
SCM[a] group	5.93 (4.03)	1.98 (2, 134)	.14	0.008
MM[b] group	6.15 (5.21)	0.75 (2, 130)	.47	0.002
Work performance				
SCM group	66.71 (17.00)	6.42 (2, 138)	.002	0.029
MM group	62.42 (18.98)	0.38 (2, 130)	.68	0.002
Work inefficiency				
SCM group	26.21 (4.17)	1.68 (2, 116)	.19	0.013
MM group	25.52 (5.50)	1.02 (2, 114)	.36	0.007
Cognitive flexibility				
SCM group	48.57 (9.70)	2.00 (2, 108)	.14	0.006
MM group	46.86 (8.70)	1.97 (2, 126)	.14	0.004
Self-compassion				
SCM group	81.42 (18.05)	7.38 (2, 108)	<.001	0.020
MM group	77.35 (21.94)	3.35 (2, 108)	.04	0.004
Self-kindness				
SCM group	16.19 (5.12)	1.42 (2, 124)	.24	0.004
MM group	15.69 (5.21)	2.70 (2, 126)	.07	0.005
Self-judgment				
SCM group	14.30 (4.68)	6.19 (2, 130)	.003	0.016
MM group	15.18 (5.06)	2.45 (2, 128)	.09	0.004
Common humanity				
SCM group	11.77 (3.87)	5.72 (2, 134)	.004	0.012
MM group	15.69 (5.21)	2.37 (2, 124)	.10	0.006
Isolation				
SCM group	9.74 (3.69)	2.78 (2, 130)	.07	0.009
MM group	10.83 (3.97)	3.02 (2, 128)	.052	0.007
Mindfulness				
SCM group	13.11 (3.47)	2.70 (2, 134)	.07	0.007
MM group	11.84 (3.53)	0.22 (2, 124)	.80	0.001
Overidentification				
SCM group	12.79 (3.81)	3.38 (2, 130)	.04	0.008
MM group	13.17 (4.14)	4.28 (2, 126)	.02	0.007
Perceived stress				
SCM group	25.03 (9.88)	4.43 (2, 116)	.01	0.021
MM group	26.98 (10.42)	0.56 (2, 126)	.57	0.002
Work engagement				
SCM group	28.91 (13.99)	1.29 (2, 128)	.28	0.002
MM group	25.58 (13.63)	0.45 (2, 128)	.64	0.001
Vigor				

XSL•FO
RenderX

	Follow-up assessment	Repeated measure ANOVA		
	Mean (SD)	*F* test (*df*)	*P* value	General η^2
SCM group	8.60 (4.94)	1.44 (2, 136)	.24	0.003
MM group	7.83 (4.65)	0.64 (2, 128)	.53	0.001
Dedication				
SCM group	10.71 (4.63)	1.60 (2, 136)	.21	0.002
MM group	9.73 (4.49)	0.79 (2, 130)	.46	0.002
Absorption				
SCM group	9.60 (5.04)	1.48 (2, 132)	.23	0.002
MM group	8.14 (5.02)	0.26 (2, 130)	.77	0.001
Psychological safety				
SCM group	35.86 (7.37)	4.39 (2, 118)	.01	0.013
MM group	33.79 (7.91)	0.50 (2, 130)	.61	0.001
Creativity				
SCM group	12.19 (4.57)	0.71 (2, 130)	.49	0.001
MM group	11.44 (4.37)	1.23 (2, 128)	.29	0.003

[a]SCM: self-compassion meditation.

[b]MM: mindfulness meditation.

In the SCM group, significant main effects of time using LMM were observed for work performance ($F_{2,147}$=7.96; P<.001; partial η^2=0.098), work inefficiency ($F_{2,141}$=3.19; P=.04; partial η^2=0.043), cognitive flexibility ($F_{2,135}$=3.27; P=.04; partial η^2=0.046), self-compassion ($F_{2,134}$=8.66; P<.001; partial η^2=0.114), self-judgment ($F_{2,148}$=6.66; P=.002; partial η^2=0.082), common humanity ($F_{2,149}$=6.14; P=.003; partial η^2=0.076), isolation ($F_{2,148}$=3.74; P=.03; partial η^2=0.048), overidentification ($F_{2,147}$=4.08; P=.02; partial η^2=0.053), perceived stress ($F_{2,145}$=6.48; P=.002; partial η^2=0.082), and psychological safety ($F_{2,143}$=3.18; P=.04; partial η^2=0.043; Table S2 in Multimedia Appendix 2). In the SCM group, Bonferroni-adjusted simple comparisons revealed a significant increase in work performance from preintervention to postintervention (P<.001). Significant increases were also observed in self-compassion from preintervention to postintervention (P=.002) and to follow-up (P<.001), and in common humanity from preintervention to postintervention (P=.04) and to follow-up (P=.003). Significant decreases were found in self-judgment from preintervention to postintervention (P=.02) and to follow-up (P=.002), isolation from preintervention to follow-up (P=.02), overidentification from preintervention to follow-up (P=.02), and perceived stress from preintervention to postintervention (P=.008) and to follow-up (P=.008).

In the MM group, significant main effects of time using LMM were observed for self-compassion ($F_{2,133}$=4.22; P=.02; partial η^2=0.060), self-kindness ($F_{2,144}$=3.13; P=.047; partial η^2=0.042), common humanity ($F_{2,143}$=3.07; P=.049; partial η^2=0.041), isolation ($F_{2,145}$=3.62; P=.03; partial η^2=0.047), and overidentification ($F_{2,144}$=3.96; P=.02; partial η^2=0.052). Bonferroni-adjusted simple comparisons indicated a significant increase in self-compassion from preintervention to follow-up (P=.02) and self-kindness from preintervention to follow-up (P=.04). Significant decreases were found in isolation from preintervention to follow-up (P=.02) and overidentification from preintervention to follow-up (P=.02; Table S2 in Multimedia Appendix 2).

Table 5 summarizes the main effects for time and the interaction effects for group and time. Across all intervention groups, significant main effects for time were found for work performance ($F_{2,268}$=3.80; P=.02; η^2G=0.009), self-compassion ($F_{2,216}$=10.44; P<.001; η^2G=0.010), common humanity ($F_{2,258}$=7.54; P<.001; η^2G=0.008), self-judgment ($F_{2,258}$=8.35; P<.001; η^2G=0.009), overidentification (in the SCS; $F_{2,256}$=6.62; P=.001; η^2G=0.007), and perceived stress ($F_{2,242}$=4.26; P=.02; η^2G=0.008). Bonferroni-adjusted simple comparisons showed that in the intervention group, a significant increase was observed in work performance from preintervention to postintervention (P=.01), as well as significant increases in self-compassion from preintervention to postintervention (P=.006) and to follow-up (P<.001), and in common humanity from preintervention to postintervention (P=.02) and to follow-up (P=.002). Significant decreases were observed in self-judgment from preintervention to postintervention (P=.03) and to follow-up (P<.001), in isolation from preintervention to follow-up (P=.002), in overidentification from preintervention to postintervention (P=.02) and to follow-up (P=.002), and in perceived stress from preintervention to follow-up (P=.02).

XSL•FO
RenderX

Table . Main effects for time and interaction effects for group × time using ANOVA.

	Time			Interaction effects		
	F test (*df*)	*P* value	General η^2	*F* test (*df*)	*P* value	General η^2
Psychological distress	2.44 (2, 264)	.09	0.004	0.23 (2, 264)	.80	0.000
Work performance	3.80 (2, 268)	.02	0.009	2.32 (2, 268)	.10	0.005
Work inefficiency	2.37 (2, 230)	.10	0.008	0.26 (2, 230)	.77	0.001
Cognitive flexibility	2.48 (2, 234)	.09	0.003	1.53 (2, 234)	.22	0.002
Self-compassion	10.44 (2, 216)	<.001	0.010	0.92 (2, 216)	.40	0.001
Self-kindness	2.22 (2, 250)	.11	0.003	1.68 (2, 250)	.19	0.002
Self-judgment	8.35 (2, 258)	<.001	0.009	1.02 (2, 258)	.36	0.001
Common humanity	7.54 (2, 258)	<.001	0.008	0.32 (2, 258)	.73	0.000
Isolation	5.58 (2, 258)	.004	0.007	0.18 (2, 258)	.84	0.000
Mindfulness	1.57 (2, 258)	.21	0.002	1.12 (2, 258)	.33	0.002
Overidentification	6.62 (2, 256)	.002	0.007	0.95 (2, 256)	.39	0.001
Perceived stress	4.26 (2, 242)	.02	0.008	1.12 (2, 242)	.33	0.002
Work engagement	1.24 (2, 256)	.29	0.001	0.33 (2, 256)	.72	0.000
Vigor	1.94 (2, 264)	.15	0.002	0.10 (2, 264)	.90	0.000
Dedication	1.81 (2, 266)	.16	0.001	0.39 (2, 266)	.67	0.000
Absorption	0.53 (2, 262)	.59	0.001	0.86 (2, 262)	.42	0.001
Psychological safety	2.72 (2, 248)	.07	0.004	1.75 (2, 248)	.18	0.002
Creativity	0.07 (2, 258)	.94	0.000	1.93 (2, 258)	.15	0.002

As a supplementary analysis, the main effects for time and the interaction effects for group × time were examined using LMM (Table S3 in Multimedia Appendix 3). Across all intervention groups, significant main effects for time were found for work performance ($F_{2,292}$=5.52; *P*=.004; partial η^2=0.036), self-compassion ($F_{2,267}$=3.93; *P*=.02; partial η^2=0.029), self-judgment ($F_{2,293}$=3.84; *P*=.02; partial η^2=0.025), and perceived stress ($F_{2,290}$=3.35; *P*=.04; partial η^2=0.023). Bonferroni-adjusted simple comparisons showed that in the intervention group, a significant increase was observed in work performance from preintervention to postintervention (*P*=.008), and in self-compassion from preintervention to postintervention (*P*<.001) and follow-up (*P*<.001). Significant decreases were observed in self-judgment from preintervention to postintervention (*P*<.001) and follow-up (*P*=.02), and in perceived stress from preintervention to postintervention (*P*=.006) and follow-up (*P*=.003). A significant group × time interaction was detected for work performance ($F_{2,293}$=3.23; *P*=.04; partial η^2=0.022). A simple slope analysis revealed that participants in the SCM group showed a significant increase from preintervention to postintervention (t_{301}= 3.88; *P*<.001), whereas no other significant differences were observed.

After adjusting for age and previous experience with meditation, all outcomes that were statistically significant in the unadjusted analyses remained significant with respect to the main effect of time. In addition, cognitive flexibility also showed a significant main effect of time ($F_{2,277}$=3.09; *P*=.047; partial η^2=0.022). Bonferroni-adjusted simple comparisons showed that in the intervention group, a significant increase was also observed in cognitive flexibility from preintervention to follow-up (*P*=.046). A significant group × time interaction remained for work performance ($F_{2,291}$=3.16; *P*=.04; partial η^2=0.021). A simple slope analysis revealed that participants in the SCM group showed a significant increase from preintervention to postintervention (t_{298}=3.83; *P*<.001), whereas no other significant differences were observed.

Discussion

Principal Findings

This study aimed to develop a self-care smartphone app that enables individuals to independently practice mindfulness and self-compassion without the support of a facilitator, and to evaluate its effectiveness using psychological and work-related outcomes. Participants were randomized to SCM, MM, or WCG, and intervention effects were evaluated in an RCT.

SCM participants demonstrated improvements in work performance, cognitive flexibility, self-compassion (especially common humanity), and psychological safety, along with reductions in perceived stress, work inefficiency, self-judgment, and overidentification. In contrast, MM participants only showed

XSL•FO
RenderX

improvement in common humanity. No significant group × time interaction was observed in psychological outcomes. However, sensitivity analyses using LMMs revealed a significant group × time interaction for work performance, indicating a relatively greater improvement in the SCM group. Longitudinal analyses revealed that the SCM group maintained several improvements at follow-up, particularly in self-compassion, common humanity, and stress reduction. The MM group showed limited effects, mainly in overidentification. Taken together, within-group analyses suggested improvements in several psychological and work-related outcomes in the SCM group, while changes in the MM group were more limited. Significant reductions were also found in perceived stress, work inefficiency, self-judgment, and overidentification. The next section discusses these findings in detail, including their strengths, limitations, and implications for future research and applied settings.

A comparison of psychological indicators before and after the intervention in the intervention groups revealed that, in the SCM group, work performance, cognitive flexibility, self-compassion, self-esteem, and psychological safety improved compared to preintervention levels, while perceived stress, work inefficiency, self-judgment, and overidentification significantly decreased. In the MM group, a significant increase was observed only in the common humanity subscale of self-compassion. These findings are consistent with previous studies on self-compassion interventions via smartphone apps. For example, a 7-day intervention targeting young people in New Zealand also reported improvements in self-compassion [47]. Similarly, Beshai et al [48] reported that a 4-week guided self-help program (Mind-OP), which combined mindfulness, self-compassion, and goal-setting exercises and included video and audio guidance, resulted in a significant reduction in perceived stress compared to an active control group. This study implemented a 4-week, easy-to-follow guided meditation program without the assistance of a facilitator, targeting Japanese workers, and similarly demonstrated an improvement in self-compassion. Few studies on psychological support for workers have reported intervention effects that extend beyond mental health [11]. This study is one of the few to show that even a short-term self-help intervention among workers can have positive effects on workplace-related outcomes, such as improved work performance, enhanced psychological safety, and reduced productivity loss.

This study used an RCT design to examine the effects of guided meditation practiced individually, using a rigorous approach that compared 2 intervention groups—SCM and MM groups—with a control group. However, when comparing the intervention groups with the control group, no significant interaction effects between intervention and time were observed. One possible reason for the lack of intervention effects relative to the control group is the low intensity of the intervention, which may have resulted in only modest effects. Digital interventions targeting anxiety and depressive symptoms are generally known to yield small to moderate effect sizes [49]. Additionally, the shorter 4-week duration of this fully self-guided intervention, relative to the typical 8-week format of established programs, may account for the limited between-group effects.

Spijkerman et al [50] also pointed out a limitation of psychological interventions in populations with relatively high baseline mental health, similar to participants in this study. Specifically, they noted that individuals without clinical levels of depression or anxiety tend to have lower baseline scores on psychological symptom measures, resulting in limited scope for improvement compared to clinical populations. This may result in smaller effect sizes due to floor effects. In particular, previous studies on online self-compassion interventions have reported small effects on anxiety symptoms and no significant effects on depressive symptoms when compared to active control groups [48].

Moreover, previous research has suggested that therapist-guided interventions are more effective for treating depression than fully self-guided programs [51]. Given that this study used a completely self-directed format without any therapist contact, participants may have found it difficult to independently acquire the necessary skills. This could partly account for the limited improvement, especially in depressive symptoms. In this intervention, loving-kindness meditation was implemented as a form of meditation that incorporates elements of self-compassion. Loving-kindness meditation is also introduced as an optional practice during the day-long retreat in the Mindfulness-Based Stress Reduction program and is recognized as part of the broader mindfulness framework [52]. This type of meditation is designed to cultivate feelings of warmth and benevolence toward oneself and others, including all beings.

Neff and Dahm [53] argued that self-compassion can be more effectively cultivated through explicit rather than implicit methods. The inclusion of loving-kindness meditation, which explicitly integrates components of self-compassion, may have contributed to the effects observed in this intervention over standard mindfulness practices. This may be particularly relevant for workers who are highly self-critical, as the practice of self-compassion could have helped improve their perceived work performance.

Despite the notable findings, this study has several limitations. Specifically, no significant differences were observed in the magnitude of change between the intervention and control groups from preintervention to postintervention. This limitation may be attributed to the relatively low intensity of our intervention program, which may have produced smaller effect sizes insufficient to demonstrate statistically significant group differences.

In terms of feasibility, the study demonstrated better retention and adherence than previous research reporting dropout rates exceeding 50% [54]. Several factors may have contributed to this, including a brief and clearly structured app-based intervention that lasted approximately 6 to 12 minutes per session and allowed participants to complete it at their convenience in terms of time and location. Additionally, weekly email reminders were used. Recruitment via an online crowdsourcing platform—where participants may have approached the intervention as a structured “task”—and the provision of compensation (which was not contingent on the number of meditation days) may have further encouraged engagement. We did not collect or report detailed app usage

XSL•FO
RenderX

analytics, such as time-of-day, changes in use intensity over time, or module-level engagement; consequently, total practice time could not be summarized. Given the 1-session-per-day design with a fixed session type and duration (except for interruptions), variability in dose was primarily reflected in the number of completed practice days. Future studies should prospectively capture richer log-based metrics, such as continuity patterns and total practice time, to better characterize feasibility and inform intervention optimization. To enhance intervention effectiveness while maintaining feasibility, future studies could extend the practice period beyond 4 weeks without changing daily practice duration or combine meditation with other intervention techniques (modules). Given the typical trade-off between intervention intensity and accessibility, future research should focus on developing interventions that remain accessible and user-friendly for working adults while improving therapeutic outcomes.

Separately, work performance was assessed using brief self-report measures (a single WHO-HPQ presenteeism item and the SPS work impairment score) to minimize participant burden in this fully self-guided digital trial. Nevertheless, single-item and self-reported measures may be less sensitive to change and may be susceptible to reporting biases, potentially diminishing our ability to detect effects. Future studies should consider incorporating more comprehensive and/or objective indicators of productivity, such as supervisor-rated performance or administrative and behavioral metrics, where feasible.

Furthermore, we could not confirm improvement in depressive symptoms, which was one of our main outcome measures for mental health. It has been noted that unresolved trauma memories may emerge during self-compassion practice, representing a potential risk of such training. For participants with trauma memories practicing at home, it is recommended to limit sessions to about 10 to 15 minutes per day [12]. Therefore, to reduce risks in our fully self-care intervention format, we implemented a low-intensity intervention that could be completed in a short time and targeted only workers with relatively high baseline health status. However, as previous research has indicated, preventive interventions targeting healthy individuals tend to show smaller effects [50]. Future research could incorporate risk reduction strategies, such as regular follow-ups by facilitators based on participants' circumstances. This approach would potentially allow participation from workers experiencing moderate stress levels, enabling investigations not only into prevention effects for relatively healthy workers but also into supportive effects for workers currently experiencing mental health difficulties.

Finally, the mechanisms by which the intervention improved work performance remain insufficiently understood. Currently, few studies have examined the detailed mechanisms of the relationship between self-compassion practice and work performance. In previous research on mindfulness, a similar psychological intervention, and work performance, mindfulness practice has been shown to enhance attention functions and positively affect major domains of human functioning, including cognition, emotion, behavior, and physiology, thereby improving work performance [10]. Good et al [10] suggest that improved cognitive function and cognitive flexibility may allow sustained attention to work even in environments where attention is depleted or work is performed discontinuously. Additionally, Kirk et al [55], in a study examining emotional regulation in meditation practitioners using economic decision-making games, reported that meditation practitioners activate different brain regions during emotional regulation compared to nonmeditators, suggesting the possibility of rational engagement with others even in emotional situations. Furthermore, higher levels of self-compassion are associated with reduced fear of failure and an increased tendency to try again after failure [56,57]. Based on these findings, future research could enhance understanding of the intervention's mechanisms by incorporating measures of cognitive function, brain function measurements, and psychological constructs related to failure and goal setting.

Conclusions

The intervention program was designed to enable nonclinical working adults to practice self-care meditation using a smartphone app without facilitator involvement. The study findings support the feasibility of delivering a brief, low-burden, fully self-guided meditation program incorporating self-compassion elements in occupational settings, with participants able to practice at times and locations of their choosing. However, given the limited between-group effects observed across most outcomes, the improvements noted should be considered preliminary and do not substantiate strong causal claims regarding the superiority of the intervention. Clinical significance remains unclear in this nonclinical sample and should be evaluated in future research. These findings inform the development of scalable, low-burden digital self-care tools for occupational mental health and highlight the need for further studies to evaluate clinical significance.

Acknowledgments

The authors would like to thank all participants in this study. The authors also thank Kei Minoura for his technical assistance with data collection via smartphone apps. The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: publication support. The GAI tool used was ChatGPT 5.2. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by TK.

XSL•FO
RenderX

Funding

This work was supported by the Japan Society for the Promotion of Science Grant-in-Aid for Scientific Research (grant JP16H05653, JP19K03278, 22H01091, 22K18582, 23K22362, and 25K21957 to RT, 24K22798 to TK), the Japan Science and Technology Agency (grant JPMJSP2108 for KA), the Royal Society and British Academy (grant AL150003 to RT), and University of Tokyo Social Cooperation Program "Fulfillment through Work" (to RT). The funders had no role in data collection or analysis, decision to publish, or preparation of the manuscript.

Data Availability

The datasets generated and analyzed during this study are not publicly available because data sharing in publicly available repositories was not included in the informed consent. However, the datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

TK and RT conceptualized the study. TK and KA performed the formal analysis, investigation, and visualization. KA was responsible for data curation and resources. RT secured funding for the project and provided supervision. TK managed the project administration and performed the validation. TK wrote the original draft, and RT contributed to the review and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Group × time interaction effect using linear mixed-effects models (LMMs).
[DOCX File, 26 KB - jmir_v28i1e79991_app1.docx]

Multimedia Appendix 2
Mean (SD) at follow-up and repeated-measures analyses across pre, post, and follow-up within the intervention groups.
[DOCX File, 30 KB - jmir_v28i1e79991_app2.docx]

Multimedia Appendix 3
Main effects for time and interaction effects for group × time using linear mixed-effects models (LMMs).
[DOCX File, 28 KB - jmir_v28i1e79991_app3.docx]

Checklist 1
CONSORT-EHEALTH checklist [58].
[PDF File, 453 KB - jmir_v28i1e79991_app4.pdf]

References

1. Mental health at work. World Health Organization. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/mental-health-at-work [accessed 2024-12-26]
2. Yoshimoto T, Oka H, Fujii T, Nagata T, Matsudaira K. The economic burden of lost productivity due to presenteeism caused by health conditions among workers in Japan. J Occup Environ Med 2020 Oct;62(10):883-888. [doi: 10.1097/JOM.0000000000002001] [Medline: 32826548]
3. Hayes SC. Acceptance and commitment therapy, relational frame theory, and the third wave of behavioral and cognitive therapies. Behav Ther 2004;35(4):639-665. [doi: 10.1016/S0005-7894(04)80013-3]
4. Creswell JD. Mindfulness interventions. Annu Rev Psychol 2017 Jan 3;68(1):491-516. [doi: 10.1146/annurev-psych-042716-051139] [Medline: 27687118]
5. Neff KD, Germer CK. A pilot study and randomized controlled trial of the mindful self-compassion program. J Clin Psychol 2013 Jan;69(1):28-44. [doi: 10.1002/jclp.21923] [Medline: 23070875]
6. Bégin C, Berthod J, Martinez LZ, Truchon M. Use of mobile apps and online programs of mindfulness and self-compassion training in workers: a scoping review. J Technol Behav Sci 2022;7(4):477-515. [doi: 10.1007/s41347-022-00267-1] [Medline: 36091081]
7. Michel A, Groß C, Hoppe A, González‐Morales MG, Steidle A, O'Shea D. Mindfulness and positive activities at work: intervention effects on motivation-related constructs, sleep quality, and fatigue. J Occupat & Organ Psyc 2021 Jun;94(2):309-337. [doi: 10.1111/joop.12345]

XSL•FO
RenderX

8. Janssen M, Heerkens Y, Kuijer W, van der Heijden B, Engels J. Effects of mindfulness-based stress reduction on employees' mental health: a systematic review. PLoS One 2018;13(1):e0191332. [doi: 10.1371/journal.pone.0191332] [Medline: 29364935]
9. Coo C, Salanova M. Mindfulness can make you happy-and-productive: a mindfulness controlled trial and its effects on happiness, work engagement and performance. J Happiness Stud 2018 Aug;19(6):1691-1711. [doi: 10.1007/s10902-017-9892-8]
10. Good DJ, Lyddy CJ, Glomb TM, et al. Contemplating mindfulness at work: an integrative review. J Manag 2016 Jan;42(1):114-142. [doi: 10.1177/0149206315617003]
11. Kotera Y, Van Gordon W. Effects of self-compassion training on work-related well-being: a systematic review. Front Psychol 2021;12:630798. [doi: 10.3389/fpsyg.2021.630798] [Medline: 33967896]
12. Germer C, Neff K. Teaching the Mindful Self-Compassion Program: A Guide for Professionals: Guilford Press; 2019.
13. Wasson RS, Barratt C, O'Brien WH. Effects of mindfulness-based interventions on self-compassion in health care professionals: a meta-analysis. Mindfulness (N Y) 2020;11(8):1914-1934. [doi: 10.1007/s12671-020-01342-5] [Medline: 32421083]
14. Inwood E, Ferrari M. Mechanisms of change in the relationship between self-compassion, emotion regulation, and mental health: a systematic review. Appl Psychol Health Well Being 2018 Jul;10(2):215-235. [doi: 10.1111/aphw.12127] [Medline: 29673093]
15. Wakelin KE, Perman G, Simonds LM. Effectiveness of self-compassion-related interventions for reducing self-criticism: a systematic review and meta-analysis. Clin Psychol Psychother 2022 Jan;29(1):1-25. [doi: 10.1002/cpp.2586] [Medline: 33749936]
16. Reizer A. Bringing self-kindness into the workplace: Exploring the mediating role of self-compassion in the associations between attachment and organizational outcomes. Front Psychol 2019;10:1148. [doi: 10.3389/fpsyg.2019.01148] [Medline: 31164857]
17. Wouters-Soomers L, Van Ruysseveldt J, Bos AER, Jacobs N. An individual perspective on psychological safety: the role of basic need satisfaction and self-compassion. Front Psychol 2022;13:920908. [doi: 10.3389/fpsyg.2022.920908] [Medline: 36059778]
18. Li Y, Hu Y, Yang W, Wang Y. Daily interventions and assessments: the effect of online self-compassion meditation on psychological health. Applied Psych Health & Well 2021 Nov;13(4):906-921. [doi: 10.1111/aphw.12278] [Medline: 33890708]
19. Mitsea E, Drigas A, Skianis C. Mindfulness for anxiety management and happiness: the role of VR, metacognition, and hormones. Technium BioChemMed 2022 Sep;3(3):37-52. [doi: 10.47577/biochemmed.v3i3.7343]
20. Linardon J. Can acceptance, mindfulness, and self-compassion be learned by smartphone apps? A systematic and meta-analytic review of randomized controlled trials. Behav Ther 2020 Jul;51(4):646-658. [doi: 10.1016/j.beth.2019.10.002] [Medline: 32586436]
21. Musiat P, Johnson C, Atkinson M, Wilksch S, Wade T. Impact of guidance on intervention adherence in computerised interventions for mental health problems: a meta-analysis. Psychol Med 2022 Jan;52(2):229-240. [doi: 10.1017/S0033291721004621] [Medline: 34802474]
22. Cameron G, Mulvenna M, Ennis E, et al. Effectiveness of digital mental health interventions in the workplace: umbrella review of systematic reviews. JMIR Ment Health 2025 Jan 24;12:e67785. [doi: 10.2196/67785] [Medline: 39854722]
23. Schwartz K, Ganster FM, Tran US. Mindfulness-based mobile apps and their impact on well-being in nonclinical populations: systematic review of randomized controlled trials. J Med Internet Res 2023 Aug 4;25:e44638. [doi: 10.2196/44638] [Medline: 37540550]
24. Psychologists embrace telehealth to prevent the spread of COVID-19: a survey of APA members reveals how the pandemic has impacted clinicians and their practices. American Psychological Association Services, INC. URL: https://www.apaservices.org/practice/legal/technology/psychologists-embrace-telehealth [accessed 2024-12-26]
25. Kurosawa T, Adachi K, Takizawa R. Mindful self-compassion smartphone intervention for worker mental health in Japan: protocol for a randomized controlled trial. JMIR Res Protoc 2024 Jul 15;13:e53541. [doi: 10.2196/53541] [Medline: 39008345]
26. Kessler RC, Andrews G, Colpe LJ, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002 Aug;32(6):959-976. [doi: 10.1017/s0033291702006074] [Medline: 12214795]
27. Furukawa TA, Kawakami N, Saitoh M, et al. The performance of the Japanese version of the K6 and K10 in the World Mental Health Survey Japan. Int J Methods Psychiatr Res 2008;17(3):152-158. [doi: 10.1002/mpr.257] [Medline: 18763695]
28. Sakurai K, Nishi A, Kondo K, Yanagida K, Kawakami N. Screening performance of K6/K10 and other screening instruments for mood and anxiety disorders in Japan. Psychiatry Clin Neurosci 2011 Aug;65(5):434-441. [doi: 10.1111/j.1440-1819.2011.02236.x] [Medline: 21851452]
29. Kessler RC, Barber C, Beck A, et al. The World Health Organization Health and Work Performance Questionnaire (HPQ). J Occup Environ Med 2003 Feb;45(2):156-174. [doi: 10.1097/01.jom.0000052967.43131.51] [Medline: 12625231]

XSL•FO
RenderX

30. Kawakami N, Inoue A, Tsuchiya M, et al. Construct validity and test-retest reliability of the World Mental Health Japan version of the World Health Organization Health and Work Performance Questionnaire short version: a preliminary study. Ind Health 2020 Aug 7;58(4):375-387. [doi: 10.2486/indhealth.2019-0090] [Medline: 32173661]
31. Koopman C, Pelletier KR, Murray JF, et al. Stanford Presenteeism Scale: health status and employee productivity. J Occup Environ Med 2002 Jan;44(1):14-20. [doi: 10.1097/00043764-200201000-00004] [Medline: 11802460]
32. Wada K, Moriyama M, Narai R, et al. The effect of chronic health conditions on work performance in Japanese companies. J Occup Health 2007 May;49(3):103-109. [doi: 10.1539/sangyoeisei.49.103] [Medline: 17575410]
33. Martin MM, Rubin RB. A new measure of cognitive flexibility. Psychol Rep 1995 Apr;76(2):623-626. [doi: 10.2466/pr0.1995.76.2.623]
34. Oshiro K, Nagaoka S, Shimizu E. Development and validation of the Japanese version of Cognitive Flexibility Scale. BMC Res Notes 2016 May 17;9(1):275. [doi: 10.1186/s13104-016-2070-y] [Medline: 27188498]
35. Neff KD. The development and validation of a scale to measure self-compassion. Self Identity 2003 Jul;2(3):223-250. [doi: 10.1080/15298860309027]
36. Arimitsu K. Development and validation of the Japanese version of the Self-Compassion Scale. Shinrigaku Kenkyu 2014 Apr;85(1):50-59. [doi: 10.4992/jjpsy.85.50] [Medline: 24804430]
37. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [doi: 10.2307/2136404] [Medline: 6668417]
38. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Routledge Academic; 1988.
39. Sumi K. Reliability and validity of the Japanese version of the Perceived Stress Scale. Jpn J Health Psychol 2006;19(2):44-53. [doi: 10.11560/jahp.19.2_44]
40. Schaufeli WB, Salanova M, González-romá V, Bakker AB. The measurement of engagement and burnout: a two sample confirmatory factor analytic approach. J Happiness Stud 2002 Mar;3(1):71-92. [doi: 10.1023/A:1015630930326]
41. Schaufeli WB, Bakker AB, Salanova M. The measurement of work engagement with a short questionnaire: a cross-national study. Educ Psychol Meas 2006 Aug;66(4):701-716. [doi: 10.1177/0013164405282471]
42. Shimazu A, Schaufeli WB, Kosugi S, et al. Work engagement in Japan: validation of the Japanese version of the Utrecht Work Engagement Scale. Appl Psychol 2008 Jul;57(3):510-523. [doi: 10.1111/j.1464-0597.2008.00333.x]
43. Edmondson A. Psychological safety and learning behavior in work teams. Adm Sci Q 1999 Jun;44(2):350-383. [doi: 10.2307/2666999]
44. Matsuo A, Tsujita M, Kita K, Ayaya S, Kumagaya S. Developing and validating Japanese versions of Psychological Safety Scale, Knowledge Sharing Scale and Expressed Humility Scale. Manag Labour Stud 2024 Aug;49(3):375-388. [doi: 10.1177/0258042X231191871]
45. George JM, Zhou J. When openness to experience and conscientiousness are related to creative behavior: an interactional approach. J Appl Psychol 2001 Jun;86(3):513-524. [doi: 10.1037/0021-9010.86.3.513] [Medline: 11419810]
46. Olejnik S, Algina J. Generalized eta and omega squared statistics: measures of effect size for some common research designs. Psychol Methods 2003 Dec;8(4):434-447. [doi: 10.1037/1082-989X.8.4.434] [Medline: 14664681]
47. Thabrew H, Boggiss AL, Lim D, et al. Well-being app to support young people during the COVID-19 pandemic: randomised controlled trial. BMJ Open 2022 May 19;12(5):e058144. [doi: 10.1136/bmjopen-2021-058144] [Medline: 35589362]
48. Beshai S, Bueno C, Yu M, Feeney JR, Pitariu A. Examining the effectiveness of an online program to cultivate mindfulness and self-compassion skills (Mind-OP): randomized controlled trial on Amazon's Mechanical Turk. Behav Res Ther 2020 Nov;134:103724. [doi: 10.1016/j.brat.2020.103724] [Medline: 32942203]
49. Rauschenberg C, Schick A, Hirjak D, et al. Evidence synthesis of digital interventions to mitigate the negative impact of the COVID-19 pandemic on public mental health: rapid meta-review. J Med Internet Res 2021 Mar 10;23(3):e23365. [doi: 10.2196/23365] [Medline: 33606657]
50. Spijkerman MPJ, Pots WTM, Bohlmeijer ET. Effectiveness of online mindfulness-based interventions in improving mental health: a review and meta-analysis of randomised controlled trials. Clin Psychol Rev 2016 Apr;45:102-114. [doi: 10.1016/j.cpr.2016.03.009] [Medline: 27111302]
51. Gellatly J, Bower P, Hennessy S, Richards D, Gilbody S, Lovell K. What makes self-help interventions effective in the management of depressive symptoms? Meta-analysis and meta-regression. Psychol Med 2007 Sep;37(9):1217-1228. [doi: 10.1017/S0033291707000062] [Medline: 17306044]
52. Santorelli S, Meleo-Meyer F, Koerbel L, Kabat-Zinn J. Mindfulness-Based Stress Reduction (MBSR) Authorized Curriculum Guide.: University of Massachusetts Medical School, Center for Mindfulness in Medicine, Health Care, and Society; 2017. URL: https://www.rightdecisions.scot.nhs.uk/m/2105/mbsr-curriculum-guide-2017.pdf [accessed 2025-06-12]
53. Neff KD, Dahm KA. Self-compassion: what it is, what it does, and how it relates to mindfulness. In: Ostafin BD, Robinson MD, Meier BP, editors. Handbook of Mindfulness and Self-Regulation: Springer; 2015:121-137.
54. Carvalho SA, Skvarc D, Barbosa R, Tavares T, Santos D, Trindade IA. A pilot randomized controlled trial of online acceptance and commitment therapy versus compassion-focused therapy for chronic illness. Clin Psychol Psychother 2022 Mar;29(2):524-541. [doi: 10.1002/cpp.2643] [Medline: 34269493]
55. Kirk U, Downar J, Montague PR. Interoception drives increased rational decision-making in meditators playing the ultimatum game. Front Neurosci 2011;5:49. [doi: 10.3389/fnins.2011.00049] [Medline: 21559066]

XSL•FO
RenderX

56. Neff KD, Hsieh YP, Dejitterat K. Self-compassion, achievement goals, and coping with academic failure. Self Identity 2005 Jul;4(3):263-287. [doi: 10.1080/13576500444000317]
57. Neely ME, Schallert DL, Mohammed SS, Roberts RM, Chen YJ. Self-kindness when facing stress: the role of self-compassion, goal regulation, and support in college students' well-being. Motiv Emot 2009 Mar;33(1):88-97. [doi: 10.1007/s11031-008-9119-8]
58. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: Improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011;13(4):e126. [doi: 10.2196/jmir.1923]

Abbreviations

CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
LMM: linear mixed-effects model
MM: mindfulness meditation
RCT: randomized controlled trial
SCM: self-compassion meditation
SCS: Self-Compassion Scale
SPS: Stanford Presenteeism Scale
WCG: waitlist control group
WHO: World Health Organization
WHO-HPQ: World Health Organization Health and Work Performance Questionnaire

Edited by A Stone; submitted 05.Jul.2025; peer-reviewed by A Perez-Aranda, B Olaya, N Sripo; revised version received 08.Feb.2026; accepted 10.Feb.2026; published 17.Mar.2026.

Please cite as:
Kurosawa T, Adachi K, Takizawa R
Effects of Self-Compassion and Mindfulness Interventions on Mental Health and Work-Related Outcomes Among Japanese Workers: Randomized Controlled Trial
J Med Internet Res 2026;28:e79991
URL: https://www.jmir.org/2026/1/e79991
doi:10.2196/79991

© Takumu Kurosawa, Koichiro Adachi, Ryu Takizawa. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effects of Internet-Based Dementia Risk Reduction Education on Risk and Protective Factor Knowledge, Intentions, and Health Behaviors: Randomized Controlled Trial

Anthony J Levinson[1,2*], BA, MA, MSc, MD; Stephanie Ayers[1*], BA, BEd; Sandra Clark[1*], BSc, MSc; Rebekah Woodburn[1*], BA; Amy Schneeberg[1], PhD; Dima Hadid[1], MPH; Nick Kates[1], MD; Gillian Strudwick[3], RN, PhD; Roland Grad[4], MSc, MD; Alexandra Papaioannou[1,2], BScN, MSc, MD; Maureen Dobbins[1,5], RN, PhD; Henry Siu[1], MSc, MD; Dante Duarte[1,2], MD, PhD; Karen Saperson[1], MBChB; Sharon Marr[6], BSc, MEd, MD; Doug Oliver[1], BSc, MSc, MD; Sarah Neil-Sztramko[1,7], PhD

[1]Faculty of Health Sciences, McMaster University, 1280 Main Street West, Hamilton, ON, Canada

[2]Faculty of Health Sciences, Geras Centre for Aging Research, 88 Maplewood Avenue, Hamilton, ON, Canada

[3]Centre for Addiction and Mental Health, Toronto, Canada

[4]Department of Family Medicine, McGill University, Montreal, Canada

[5]Faculty of Health Sciences, National Collaborating Centre for Methods and Tools, Hamilton, Canada

[6]Department of Medicine, Unity Health Toronto, Toronto, Canada

[7]Faculty of Health Sciences, National Collaborating Centre for Methods and Tools, Hamilton, ON, Canada

[*]these authors contributed equally

Corresponding Author:
Anthony J Levinson, BA, MA, MSc, MD
Faculty of Health Sciences, McMaster University, 1280 Main Street West, Hamilton, ON, Canada

Abstract

Background: Dementia prevention through the reduction of modifiable risk factors is gaining attention as a public health strategy. However, public knowledge of dementia risk and protective factors remains low. Web-based education offers a potential solution to raise awareness and promote risk-reduction behaviors.

Objective: This randomized controlled trial evaluated the effectiveness of DementiaRisk.ca, an internet-based multimedia educational intervention, in increasing knowledge of dementia risk factors, intentions to engage in risk reduction behaviors, and changes in health behaviors.

Methods: A 2-arm randomized controlled trial was conducted with 510 participants (265 in the intervention group and 245 in the control group). Participants were randomized to receive either the e-learning about dementia risk and promoting brain health, which included a multimedia lesson and microlearning emails, or a control intervention focused on mild cognitive impairment. Outcomes included knowledge of dementia risk factors, intentions to engage in risk reduction, and health behaviors, measured at baseline (T1), 4 weeks (T2), and 2 months postintervention (T3). Outcomes were analyzed using linear mixed effects models with fixed effects for group, time, and their interaction, and a random intercept for participants.

Results: Of the 510 randomized participants, 405 (79.4%) completed all intervention components. Participants were predominantly female (n=309, 60.6%) and aged 55 years or older (n=284, 55.7%). Baseline mean dementia knowledge scores were 17.0 (SD 5.5) in the intervention group and 17.4 (SD 6.0) in the control group. At T2, scores increased to 25.8 (SD 4.5) and 23.6 (SD 5.1), respectively, yielding a between-group difference of 2.2 points (95% CI 1.2 - 3.2; $P<.001$), which was sustained at T3. Both groups showed significant improvements in knowledge, intentions, and health behaviors over time, with larger knowledge gains in the intervention group and particularly among participants with lower educational attainment. Intentions to engage in dementia risk reduction improved in both groups at T2 (intervention: +1.0, 95% CI 0.2 - 1.8; control: +1.4, 95% CI 0.5 - 2.3), with no significant between-group difference. Self-reported physical activity increased from 31.7 (SD 25.0) to 38.6 (SD 27.5) in the intervention group and from 29.9 (SD 23.5) to 32.5 (SD 26.6) in the control group, with a between-group difference of 5.4 points at T2 (95% CI 0.3 - 10.5; $P=.04$). No significant between-group differences were observed for diet, alcohol use, or other health behaviors. Qualitative findings indicated that participants valued the intervention for improving awareness of dementia risk factors, motivating proactive lifestyle changes, and enhancing confidence in applying prevention information.

Conclusions: This internet-based dementia risk reduction e-learning program improved dementia-related knowledge and increased self-reported physical activity, particularly among participants with lower educational attainment. Effects on other

XSL•FO
RenderX

health behaviors were limited. These findings support the use of well-designed e-learning as a scalable public health strategy to strengthen dementia risk reduction literacy and encourage selected healthy behaviors.

Trial Registration: ClinicalTrials.gov NCT05383118; https://clinicaltrials.gov/study/NCT05383118

International Registered Report Identifier (IRRID): RR2-10.2196/64718

(J Med Internet Res 2026;28:e79405) doi:10.2196/79405

KEYWORDS

web-based intervention; internet; eHealth; dementia risk; dementia prevention; Alzheimer disease; education and training; clinical trial; public health; digital health interventions

Introduction

Background

As our population ages, the prevalence of dementia is expected to continue to rise. Globally, dementia affects over 55 million people, with annual costs to health care, long-term care, and informal care partners exceeding US $1.3 trillion [1-4]. Without effective prevention strategies, the number of dementia cases is projected to triple by 2050, representing a significant global health and economic burden. In Canada, it is estimated that the number of people living with dementia will triple over the next 30 years [4]. However, increasing research highlights that there are many actions individuals can take to reduce their risk of dementia. According to the recent Lancet Commission report, up to 45% of dementias could be prevented through a variety of modifiable factors [4]. Similar estimates have been reported from the Canadian Longitudinal Study on Aging, for the impact of 12 risk factors on dementia cases, with physical inactivity, depression, and hypertension being particularly important in the Canadian context [5]. Other longitudinal studies have found that engaging in up to 4 - 5 healthy lifestyle behaviors may reduce the risk of developing Alzheimer dementia by 37% to 60% [6,7]. The Landmark Study, which developed evidence-based data modeling to demonstrate the impacts of improving risk reduction efforts to delay onset across the population, suggests that a delay of developing dementia by as little as one year could result in nearly 500,000 fewer cases of dementia by 2050 [8]. Additionally, if risk reduction efforts can delay the onset of dementia by 10 years, over 4 million new cases of dementia could be avoided by 2050 [8]. Initiatives to help prevent dementia may also prove to be cost-effective at the societal level while simultaneously allowing individuals to help themselves and for care partners to support their loved ones [9].

Gaps in Knowledge

Despite growing evidence about modifiable risk factors, public understanding of dementia prevention remains limited. A national public opinion survey conducted by EKOS Research Associates Inc for the Public Health Agency of Canada revealed that only 37% of Canadians recognized chronic conditions such as hypertension, heart disease, and diabetes as risk factors for dementia [10]. Awareness of other significant factors, such as hearing loss, was even lower, with only over 10% identifying it as a risk. Meanwhile, misconceptions, such as overestimating the role of toxic chemicals, for which there is limited evidence, persist [10].

Educational Interventions

Educational interventions have the potential to improve knowledge of dementia risk and promote health behavior change [11,12]. While much of the existing literature has focused on training health care providers, a smaller number of studies have examined public education initiatives, which typically involve synchronous, one-on-one delivery by professionals. Most existing resources are predominantly text-based and often lack interactivity or adherence to principles of instructional design [13].

Web-based education offers a promising alternative due to its scalability, flexibility, and accessibility. Evidence from other health domains shows that e-learning tools that incorporate multimedia elements, such as audio narration, visuals, and interactive quizzes, are more engaging and improve long-term knowledge retention compared to static, text-only formats [14-16]. However, most of the web-based content about modifying dementia risk factors is text-based, such as static web pages or online PDF pamphlets or booklets, often requires a substantial time commitment, is frequently associated with a commercial entity or product for sale, or requires real-time participation [14,17,18]. Moreover, few, if any, evidence-based multimedia interventions on this topic have been rigorously studied using randomized controlled trials (RCTs), particularly those freely accessible to the public in multiple languages.

Objectives

The objectives of this study were to: (1) evaluate whether exposure to an e-learning intervention (DementiaRisk.ca) changes knowledge of dementia risk factors, intention to engage in risk reduction activities, and health behaviors related to dementia risk reduction. The primary outcome was the change in these variables from baseline (T1) to 4 weeks postintervention (T2), with secondary analyses examining whether changes were sustained at 2 months postintervention (T3); and (2) explore qualitative aspects such as participants' engagement and satisfaction with the intervention, as well as barriers and facilitators to use and dissemination of the information provided by the intervention, in adults without cognitive impairment.

Methods

Study Design

This study used a sequential explanatory mixed methods design, consisting of a 2-arm RCT, followed by a qualitative thematic content analysis to explore the findings from the RCT in greater depth. Full details of the methods have been published [19]. No

XSL•FO
RenderX

changes were made from the published protocol. A mixed methods design was chosen to incorporate both quantitative and qualitative analysis, which allowed for a more comprehensive and nuanced understanding of the relationship between the intervention and its impacts.

Participants

Participants were recruited from March 2023 to April 2023 using AskingCanadians, a paid panel service to allow for the recruitment of a representative and diverse sample of Canadians. Eligible participants were English-speaking adults with access to email and the internet, aged 18 years or older who had never been diagnosed with dementia. Participants were provided with the study information and eligibility criteria through an online link. After meeting eligibility criteria, participants were redirected to our online research study platform to complete informed consent and baseline measures. Informed consent was obtained digitally, with participants required to review an online information page before consenting to participate. We attempted to recruit participants across a range of ages, as risk factors may impact people throughout their life course, and many risk factors often begin to accumulate in midlife, many years before cognitive signs or symptoms emerge [4]. Ethical approval was provided by the Hamilton Integrated Research Ethics Board (project ID 14886).

Study Procedures

Following completion of online informed consent, participants were randomized and directed to their assigned intervention or control group learning path. Participants were randomized in a 1:1 ratio via a computer-generated sequence through our research platform using a permuted block stratified randomization, with education and age as the stratification variables.

To the best of our ability, efforts were made to blind participants to the study hypothesis and allocation group by referring to the study in broad terms, such as “e-learning about cognitive impairment and promoting brain health” (rather than explicitly “dementia prevention,” for example). The principal investigator (AJL) and data analysts (DH and AS) remained blinded to group allocation throughout the data collection and analysis phases.

Intervention

The intervention group received e-learning about dementia risk reduction and promoting brain health consisting of the following components: (1) one 30-minute multimedia e-learning lesson (DementiaRisk.ca) about dementia risk and protective factors (Figure 1), and (2) a series of 12 prescheduled microlearning emails. The emails included small, focused segments of content taken directly from the e-learning lesson designed to reinforce key information from the e-learning lesson through spaced repetition.

Participants had 4 weeks to complete the e-learning lesson and received 3 emails per week during that period.

Content for the intervention was based predominantly on the World Health Organization (WHO) guidelines on risk reduction for cognitive decline and dementia and the Lancet Commission’s 2020 report on dementia prevention, intervention, and care [4,20]. Instructional design was guided by the Mayer Cognitive Theory of Multimedia Learning, ensuring that materials were delivered in manageable segments, using a combination of visuals, audio narration, and other evidence-based principles essential to effective e-learning that have been shown to enhance knowledge retention [13,14,16,17].

Other theoretical frameworks also informed the overall design of the intervention and its evaluation. The theory of planned behavior (TPB) [21,22] was used due to its extensive application with a number of dementia prevention risk factors (eg, smoking, diet, alcohol consumption, and physical activity) [23-27]. TPB suggests that intentions to engage in or change specific behaviors depend on motivation and ability across 6 constructs that collectively represent a person’s actual control over their behavior. DementiaRisk aimed to change participants’ attitudes and behavioral intentions by increasing knowledge related to dementia prevention and modifiable risk factors. The intervention and its evaluation were also informed by the Acquisition Cognition Application–Level of Outcomes model, which helps explain the value of the information from the user perspective, looking at the domains of situational relevance, cognitive impact, use of the information, and if used, any perceived health benefits [28].

XSL•FO
RenderX

Figure 1. Screenshot of the DementiaRisk e-learning intervention.

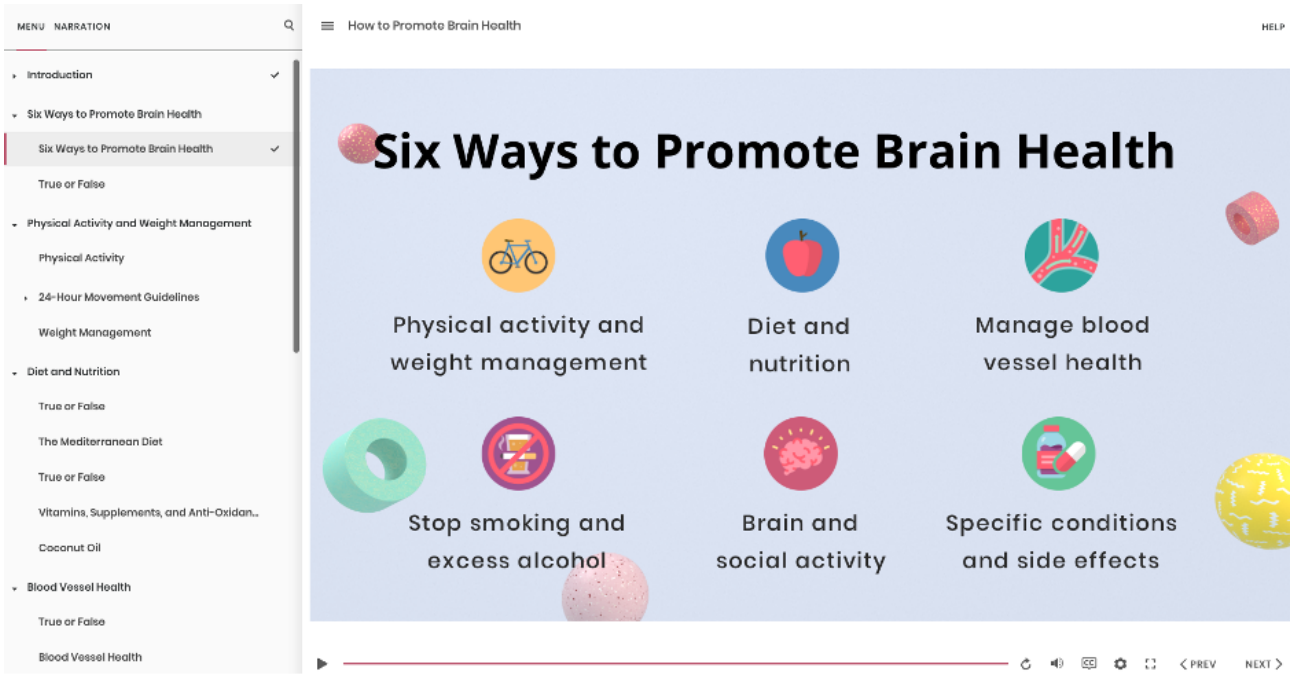

Control Condition

The control group participants received the following components: (1) 30-minute lesson on mild cognitive impairment (Figure 2), and (2) a series of 12 microlearning emails with small segments of content to reinforce material from the lesson. The design of the control group e-learning was also informed by similar principles and theoretical frameworks.

Figure 2. Screenshot from the control group e-learning.

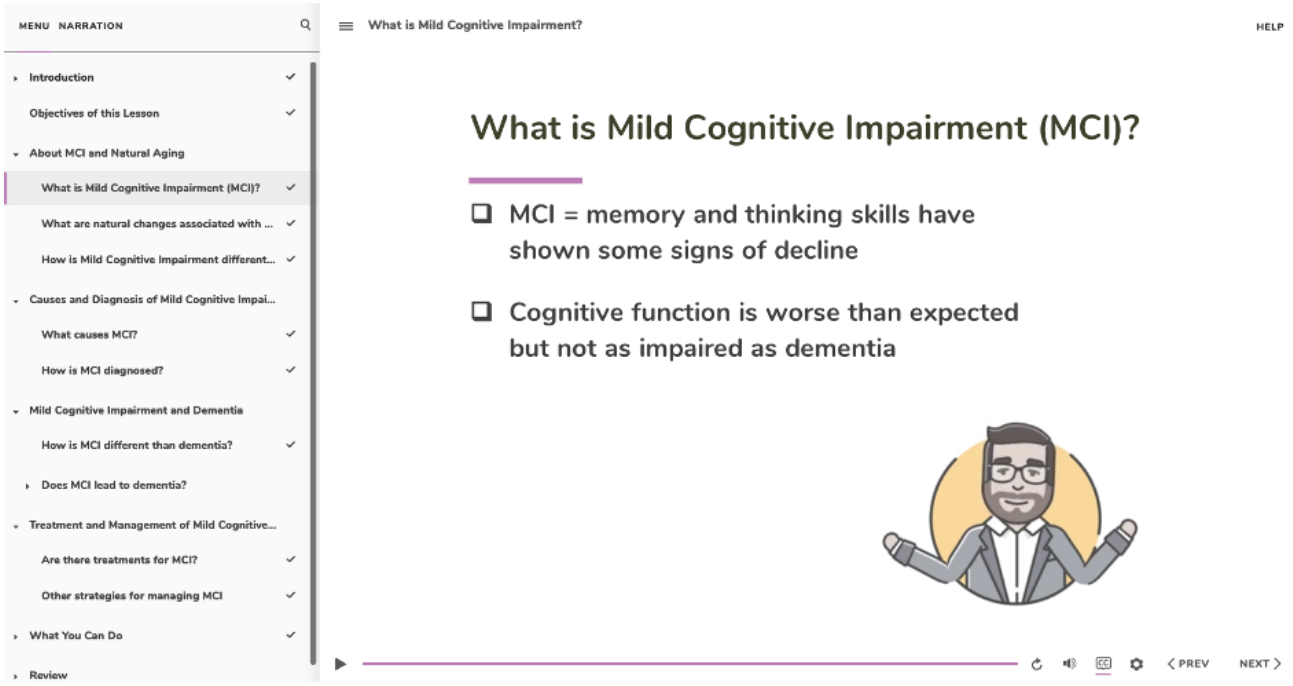

Outcome Measures

Quantitative data were collected through online questionnaires at baseline (T1), 4 weeks (T2), and at 2 months postintervention (T3). Knowledge of dementia risk factors was collected using a 15-item multiple choice assessment aligned with the content of the intervention as well as the Risks and Health Promotion subdomain of the Dementia Knowledge Assessment Scale [29]. The total maximum score for this assessment is 30 (maximum 2 points per question), with higher scores indicating higher levels of knowledge related to dementia risk factors. Intentions to engage in health behaviors were measured using a 10-item assessment. The total maximum score for this assessment is 60, with a higher score indicating higher levels of intention to engage. Current participant health behaviors related to dementia risk factors were assessed using the Godin-Shephard Leisure-Time Physical Activity Questionnaire for physical activity [30] along with additional questions related to diet,

XSL•FO
RenderX

smoking, alcohol consumption, social activity, traumatic brain injury, blood pressure, depression, air pollution, diabetes, sleep habits, brain activity, and hearing. Typical interpretation of total leisure activity scores uses a range of "insufficiently active or sedentary" (less than 14 points), "moderately active" (14 - 23 points), and "active" (24 points or more).

Data related to engagement with the e-learning programs were collected during the active study phase to gain a clearer understanding of how the use of web-based learning throughout the study (or fidelity of delivery) might have impacted changes in knowledge, intentions, or health behaviors. The following metrics were used: number of e-learning starts and completions, e-learning completion rates, time spent on lessons, and microlearning email open rates.

All data were collected through secure password-protected online platforms, which automatically recorded responses and stored data. Participant responses were anonymized prior to analysis. Missing data were handled using linear mixed effects models, which incorporate all available data under the assumption of missing at random (MAR). Given the structure of the data and the analytic approach, multiple imputation was not performed, as there is evidence that linear mixed models (LMMs) provide unbiased estimates without imputation when using full information maximum likelihood estimation [31].

Qualitative Data Collection

Qualitative data were collected from participants at T2 through both structured and open-ended questions using the Information Assessment Method for Patients and Consumers (IAM4all), a validated questionnaire that assesses outcomes of online consumer health information [19,32]. Qualitative questions explored participant satisfaction with the intervention, changes in attitudes, beliefs, and behaviors during the study, and feedback on proposed future dissemination methods. Data collection continued until no new themes emerged.

Data Analysis

Quantitative Data

RStudio (version 4.4.1; R Foundation) was used for quantitative data analysis [33-36]. LMMs were used to examine the relationship between group, time point, and each of the 3 outcomes.

All models included outcome scores from all 3 time points: baseline (T1), 4 weeks (T2), and 2 months post-intervention (T3) as the dependent variable. An interaction between time, group, and participant ID was included as a random intercept. Random slopes were not included due to limited time points and sample size constraints. The point estimates and 95% CIs comparing groups at T2 were used to address the primary objective. Exploratory analysis of whether changes observed at T2 were sustained at T3 was also addressed through the inclusion of the group by time point interaction.

Sex, age group (<45, 45 - 64, and ≥65 y), and education level (high school equivalent or less, some college or university, and college or university or graduate degree) were included as covariates in all models. Additionally, exploratory subgroup analyses of these were performed through testing interaction terms of group assignment and the stratification terms, as well as between time point and stratification terms. Nested models were compared using likelihood ratio tests ($P<.05$). For knowledge score only, having a family history of dementia or having been a care partner was combined into a single variable and tested as a potential effect modifier.

Missing data were explored across all 3 outcome variables: knowledge of dementia risk factors, intentions to engage in dementia risk reduction behaviors, and intentions to access dementia risk reduction information. For each outcome, participants with missing versus observed scores were compared at each timepoint across key demographic and exposure variables. Descriptive summaries and formal comparisons using chi-square or Fisher exact tests were conducted to assess potential differential attrition. Linear mixed effects models were used to incorporate all available data under the assumption of MAR, and multiple imputation was not performed.

All models were run for both intention-to-treat (ITT) and per-protocol (PP) group allotment. Effect sizes for group comparisons were calculated using the *emmeans* package in RStudio version 2023.12.1.402 [33-36]. Data analysts were blinded to participant grouping.

Using a conservative estimate of a small effect size (0.16, from a meta-analysis of internet health behavior change interventions), with a power of 0.80 and α of 0.05, we aimed to recruit a total of 388 participants in the study [37,38]. The sample size was increased by 25% to account for anticipated dropout, which brought the target recruitment number to 485 individuals [37,38]. These recruitment numbers and strategies have been used successfully in similar knowledge translation intervention studies and were a feasible target given the use of the paid panel [39].

Qualitative Data

For qualitative data, a conventional inductive content analysis approach was used [40]. All qualitative data were entered into NVivo (version 14; Lumivero) [41,42]. Data were systematically examined [19]. Codes and themes were developed based on the content, with no reliance on preconceived theories. The primary analyst (DH) conducted the initial coding and generated preliminary themes to ensure consistency and depth in the interpretation of the data. To strengthen rigor and minimize bias, the primary coder (DH) met weekly with the principal investigator (AJL) and other members of the research team to discuss developing codes and themes, confirm interpretations, and ensure analytic dependability. This approach has been shown to enhance analytic rigor through peer debriefing and triangulation [31]. The analysis focused on each theme separately. An iterative process of confirmation and reflection was used to analyze the data collaboratively. An audit trail was maintained, including memos from each stage of data collection and analysis. Findings of themes were triangulated between analysts. We ensured the study's trustworthiness by considering validity standards that ensure rigor, confirmability, credibility, dependability, and transferability.

XSL•FO
RenderX

Data Integration

Although the quantitative and qualitative components were analyzed separately, integration occurred during interpretation. The quantitative findings demonstrated changes in dementia risk–related knowledge, behavioral intentions, and selected health behaviors over time, while the qualitative findings helped contextualize these outcomes by elucidating participants' perceptions of the intervention's relevance, understandability, and anticipated benefits. In particular, qualitative data provided insight into how increased awareness translated into motivation for short-term lifestyle changes and highlighted perceived facilitators and barriers to applying dementia risk–reduction information. This sequential explanatory mixed methods approach enabled a more comprehensive understanding of the intervention's impact than either quantitative or qualitative findings alone.

Ethical Considerations

This study gained Hamilton Integrated Research Ethics Board approval on August 24, 2022 (project ID 14886). Participants completed all study components entirely online. Participants were required to provide informed consent and were informed of the length of time of the e-learning, surveys, and email campaign as well as informed about details surrounding data collection, storage, and investigator identities. Participants' identities and confidentiality were maintained throughout the research study. All participant data were deidentified, and all findings will be nonidentifiable. There was no known risk or harm to participating in this study or publicizing its results or findings. Participants were provided with AskingCanadians points that could be redeemed for various gift cards as compensation for participation in the study.

Results

Participant Characteristics

A total of 1319 individuals responded to the initial online recruitment message and 879 eligible participants were invited to complete informed consent. 534 individuals consented and were randomly assigned to the intervention (n=272) or control group (n=262). A total of 510 participants began the study (intervention n=265; control n=245). Retention at T3 was 79.4% (intervention n=211; control n=194). Participant flow is detailed in Figure 3. Participants were able to access all assessments at various time points, regardless of whether they had completed previous components and were therefore not considered lost to follow-up (see Checklist 1).

XSL•FO
RenderX

Figure 3. CONSORT (Consolidated Standards of Reporting Trials) study flow diagram. Note that participants were able to access all assessments at various time points, regardless of whether or not they completed previous components and were therefore not considered lost to follow-up. MCI: mild cognitive impairment.

XSL•FO
RenderX

Baseline characteristics were similar across all groups (all $P>.05$). Table 1 presents the demographic details of participants in each group. Participants were predominantly older (55.7%, aged ≥55 y), female (60.6%), and in good, very good, or excellent health (81.2%). The majority of participants identified as Caucasian (55.3%) followed by Chinese (10.4%) and South Asian or multiracial (each at 6.1%). Distribution patterns were relatively consistent across groups, though the control group had a slightly higher proportion of multiracial and Indigenous or First Nations participants. When asked about a family history of dementia, 31.4% of participants reported having a family history, 51.6% reported no family history, and 16.1% were unsure. The responses were consistent between groups. Finally, approximately one-quarter of participants (25.1%) identified as being a dementia care partner, with a higher proportion in the intervention group (27.9%) compared to the control group (22%).

XSL•FO
RenderX

Table . Participant demographics (N=510).

Characteristics	Intervention (n=265), n (%)	Control (n=245), n (%)	Overall (N=510), n (%)
Sex			
Male	107 (40.4)	89 (36.3)	196 (38.4)
Female	155 (58.5)	154 (62.9)	309 (60.6)
Missing	3 (1.1)	2 (0.8)	5 (1)
Age (y)			
18-24	1 (0.4)	3 (1.2)	4 (0.8)
25-34	26 (9.8)	23 (9.4)	49 (9.6)
35-44	41 (15.5)	37 (15.1)	78 (15.3)
45-54	51 (19.2)	44 (18)	95 (18.6)
55-64	53 (20)	58 (23.7)	111 (21.8)
65-74	59 (22.3)	60 (24.5)	119 (23.3)
≥75	31 (11.7)	18 (7.3)	49 (9.6)
Missing	3 (1.1)	2 (0.8)	5 (1)
Race and ethnicity			
Asian	7 (2.6)	10 (4.1)	17 (3.3)
Asian, East Asian	1 (0.4)	2 (0.8)	3 (0.6)
Black or African	8 (3)	6 (2.4)	14 (2.7)
Canadian	11 (4.2)	11 (4.5)	22 (4.3)
Caribbean	0 (0)	1 (0.4)	1 (0.2)
Chinese	28 (10.6)	25 (10.2)	53 (10.4)
East Asian	1 (0.4)	1 (0.4)	2 (0.4)
Eastern European	1 (0.4)	0 (0)	1 (0.2)
European	4 (1.5)	3 (1.2)	7 (1.4)
Filipino	1 (0.4)	2 (0.8)	3 (0.6)
Indigenous or First Nations	1 (0.4)	6 (2.4)	7 (1.4)
Jewish	3 (1.1)	3 (1.2)	6 (1.2)
Latino	2 (0.8)	0 (0)	2 (0.4)
Middle Eastern	1 (0.4)	0 (0)	1 (0.2)
Multiracial	11 (4.2)	20 (8.2)	31 (6.1)
South Asian	20 (7.5)	11 (4.5)	31 (6.1)
Western European	12 (4.5)	10 (4.1)	22 (4.3)
White or Caucasian	150 (56.6)	132 (53.9)	282 (55.3)
Missing	3 (1.1)	2 (0.8)	5 (1)
Health status			
Excellent	16 (6)	14 (5.7)	30 (5.9)
Very good	83 (31.3)	68 (27.8)	151 (29.6)
Good	124 (46.8)	109 (44.5)	233 (45.7)
Fair	38 (14.3)	48 (19.6)	86 (16.9)
Poor	1 (0.4)	4 (1.6)	5 (1)
Missing	3 (1.1)	2 (0.8)	5 (1)
Family history of dementia			
I am not sure	44 (16.6)	38 (15.5)	82 (16.1)

XSL•FO
RenderX

Characteristics	Intervention (n=265), n (%)	Control (n=245), n (%)	Overall (N=510), n (%)
No	136 (51.3)	127 (51.8)	263 (51.6)
Yes	82 (30.9)	78 (31.8)	160 (31.4)
Missing	3 (1.1)	2 (0.8)	5 (1)
Dementia care partner			
No	188 (70.9)	189 (77.1)	377 (73.9)
Yes	74 (27.9)	54 (22)	128 (25.1)
Missing	3 (1.1)	2 (0.8)	5 (1)

e-Learning Program Engagement

Metrics related to participant engagement with the e-learning programs are presented in Table 2. Overall, participants had a high e-learning lesson completion rate of 93% for both the intervention and control groups. Participants spent roughly the expected time on the e-learning lessons: an average of 31 minutes and 52 seconds (SD 17 minutes and 47 seconds) on the intervention lesson (estimated time to complete was 30 - 35 min), and an average of 26 minutes and 11 seconds (SD 10 minutes 37 seconds) on the control lesson (estimated time to complete was 25 - 30 min). The microlearning email campaign open rates were also strong (80.5% in the intervention and 78.6% in the control).

Table . e-Learning metrics.

e-Learning metric	Intervention (n=265)	Control (n=245)
Lesson starts	235	223
Lesson completions	218	207
Lesson completion rate (%)[a]	93	93
Average time spent on lesson	31 min, 52 s[b]	26 min 11 s
Micro-learning email open rate (%)	80.5	78.6

[a]Lesson completion rate is defined as lesson completions divided by lesson starts.

[b]Participants who spent over 1.5 hours on the module (n=20) who may have left their computer on after completing the lesson were removed from this average.

Changes in Knowledge, Intentions, and Health Behaviors

The results of the linear mixed effects model revealed no significant difference between groups in baseline knowledge, intentions to engage in health behaviors, or health behaviors' scores. The results of all models can be found in Multimedia Appendix 1.

Knowledge of Dementia Risk Factors

Knowledge scores did not differ between groups at baseline (estimate −0.39, 95% CI −1.29 to 0.51; P=.40). Both groups improved significantly from T1 to T2, but the intervention group showed a statistically significantly greater increase. The difference at T2 was 2.66 points in favor of the intervention (95% CI 1.70 - 3.63; P<.001). Knowledge gains were maintained at T3, with the intervention group continuing to exhibit significantly greater gain (estimate 1.60, 95% CI 0.65 - 2.55; P=.001). Crude scores are shown in Table 3 and marginal means in Figure 4.

Table . Crude knowledge scores across time points by group (N=510).

	Intervention (n=265)			Control (n=245)			Overall (N=510)		
	T1	T2	T3	T1	T2	T3	T1	T2	T3
Mean (SD)	17.0 (5.51)	25.8 (4.54)	25.6 (4.68)	17.4 (5.98)	23.6 (5.13)	24.2 (5.06)	17.2 (5.74)	24.8 (4.95)	24.9 (4.91)
Median (IQR)	17.0 (14.0-20.0)	28.0 (24.0-29.0)	27.0 (23.0-29.0)	17.0 (13.0-22.0)	25.0 (21.0-28.0)	25.0 (21.0-28.0)	17.0 (12.0-21.0)	26.0 (22.0-28.0)	26.0 (22.0-29.0)
Missing, n (%)	3 (1.1)	56 (21.1)	54 (20.4)	3 (1.2)	57 (23.3)	51 (20.8)	6 (1.2)	113 (22.2)	105 (20.6)

XSL•FO
RenderX

Figure 4. Marginal means from linear mixed model (LMM) with 95% CI for knowledge score by: (A) group, (B) level of education, and (C) dementia history or care partner.

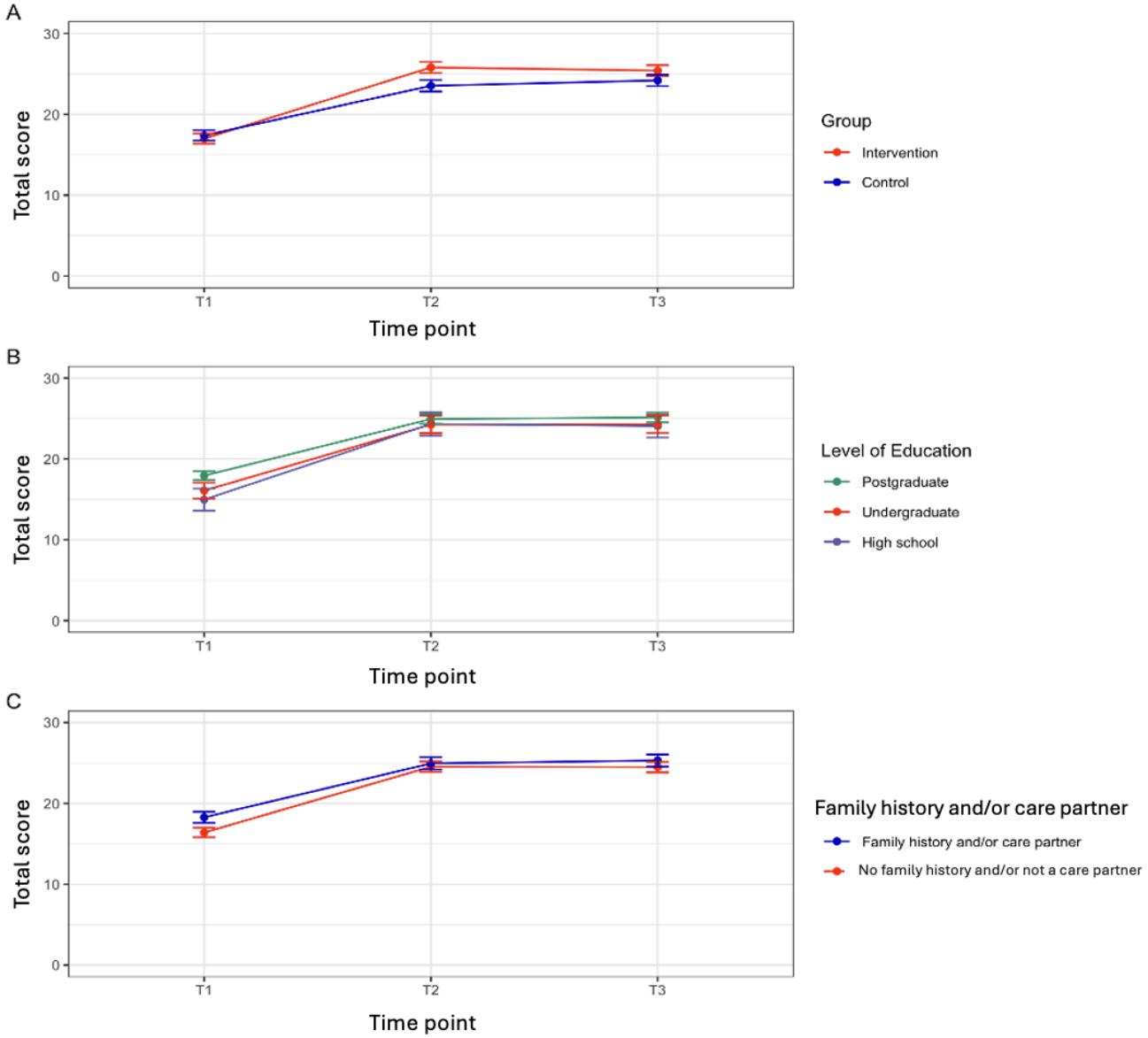

Subgroup analysis indicated that participants with lower levels of education achieved the greatest gains at both T2 (estimate 2.36, 95% CI 0.83 - 3.88; *P*=.003) and T3 (estimate 1.90, 95% CI 0.41 - 3.39; *P*=.01). Those with a family history of dementia or caregiving experience began with higher baseline knowledge (estimate 1.88, 95% CI 0.96 - 2.79; *P*<.001) and improved less over time, suggesting a ceiling effect. No significant interactions between group assignment and stratification variables were observed. Missing data analyses supported the assumption that data were missing at random.

Intentions to Engage in Risk Reduction

Intention scores differed at baseline, with higher scores in the intervention group. Both groups had a modest increase in their intentions from T1 to T2; the control group's improvement was significant (estimate 1.42, 95% CI 0.18 - 2.68; *P*=.03). However, there were no significant between-group differences at T2 (estimate −0.31, 95% CI −1.34 to 0.72; *P*=.55) or T3 (Multimedia Appendix 1; Figure 5). Table 4 provides crude intentions scores. Subgroup analyses did not identify significant interactions, and missingness followed an MAR pattern.

XSL•FO
RenderX

Figure 5. Intentions to engage in risk reduction score by level of education and group.

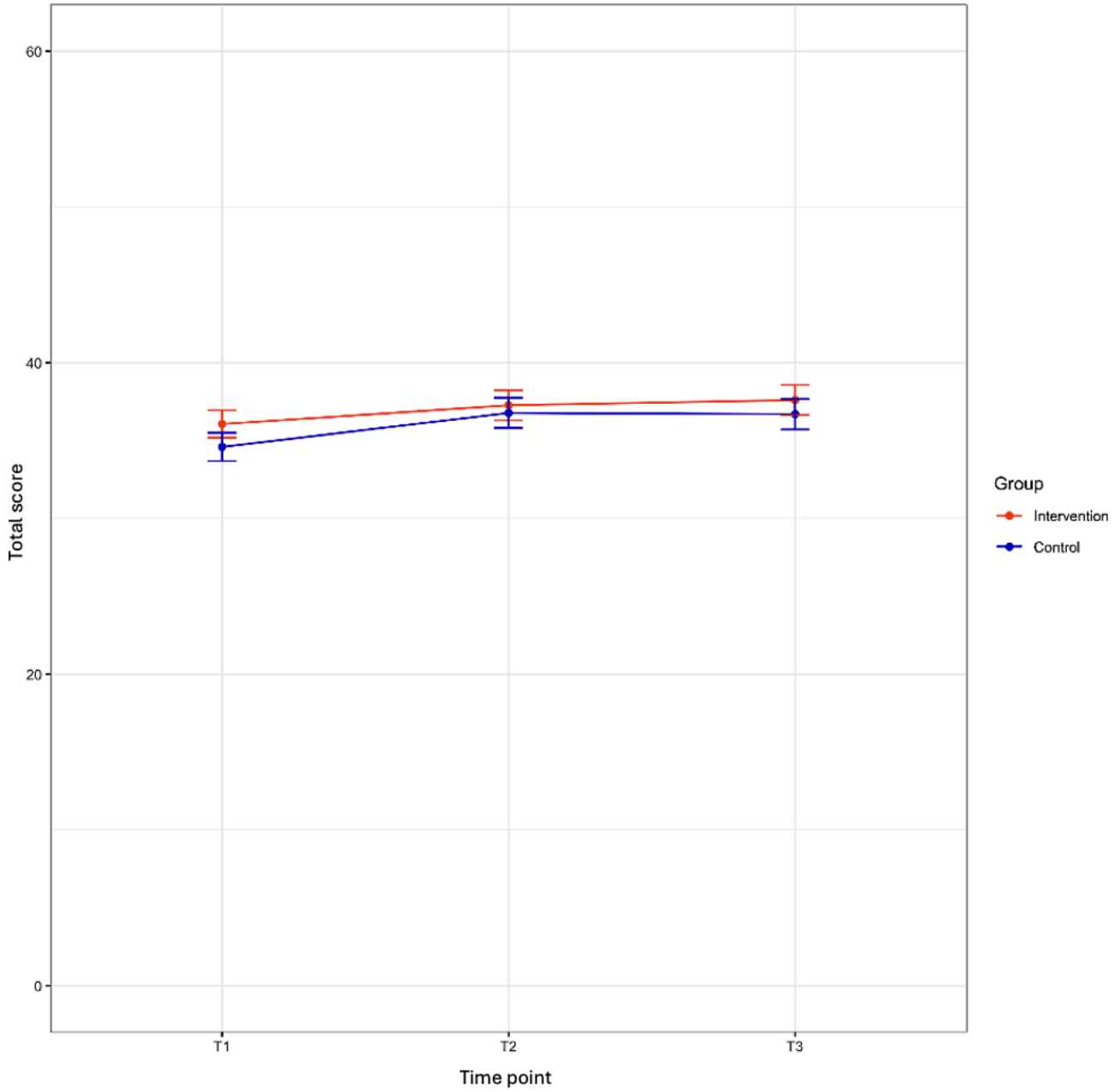

Table . Crude intentions scores across time points by group (N=510).

	Intervention (n=265)			Control (n=245)			Overall (N=510)		
	T1	T2	T3	T1	T2	T3	T1	T2	T3
Mean (SD)	37.5 (6.41)	39.2 (6.53)	39.3 (6.13)	36.2 (7.76)	37.6 (7.19)	38.1 (7.62)	36.9 (7.12)	38.4 (6.89)	38.7 (6.92)
Median (IQR)	38.0 (34.0-42.0)	40.0 (35.0-44.0)	40.0 (36.0-43.0)	37.0 (31.5-42.0)	39.0 (34.0-43.0)	40.0 (34-44.0)	38.0 (32.0-42.0)	40.0 (34.0-44.0)	40.0 (35.0-44.0)
Missing, n (%)	35 (13.2)	77 (29.1)	78 (29.4)	26 (10.6)	74 (30.2)	66 (26.9)	61 (12)	151 (29.6)	144 (28.2)

Health Behavior Changes: Godin-Shephard Leisure-Time Physical Activity Questionnaire

Self-reported physical activity levels did not differ between groups at baseline (estimate 1.87, 95% CI −2.52 to 6.25; *P*=.40). Both groups increased activity over time; the control group's activity rose significantly at T2 (estimate 3.23, 95% CI 0.81 - 5.64; *P*=.009) and was maintained at T3 (estimate 2.47, 95% CI 0.09 - 4.86; *P*=.04). Although the intervention group showed numerically greater gains, the group× time interactions were not statistically significant at either T2 (estimate 2.65, 95% CI −0.69 to 5.99; *P*=.12) or T3 (estimate 2.59, 95% CI −0.72 to 5.89; *P*=.13). Crude values are shown in Table 5, with

XSL•FO
RenderX

modeled estimates in Figure 6. Results of all models are shown in Multimedia Appendix 1.

Table . Crude Godin-Shephard Leisure-Time Physical Activity scores across time points by group (N=510).

	Intervention (n=265)			Control (n=245)			Overall (N=510)		
	T1	T2	T3	T1	T2	T3	T1	T2	T3
Mean (SD)	31.7 (25.0)	38.6 (27.5)	37.7 (27.1)	29.9 (23.5)	32.5 (26.6)	32.5 (26.5)	30.8 (24.3)	35.7 (27.2)	35.2 (26.9)
Median (IQR)	24.0 (14.0-46.8)	32.0 (19.0-56.0)	33.5 (17.0-50.8)	26.0 (14.0-40.0)	27.0 (12.0-44.2)	26.0 (15.0-42.0)	24.5 (14.0-44.0)	30.0 (15.0-49.0)	28.0 (15.8-48.0)
Missing, n (%)	3 (1.1)	56 (21.1)	55 (20.8)	3 (1.2)	57 (23.3)	51 (20.8)	6 (1.2)	113 (22.2)	106 (20.8)

Figure 6. Linear mixed model (LMM) estimates and 95% CI for Godin-Shephard Leisure-Time Physical Activity Questionnaire score by group.

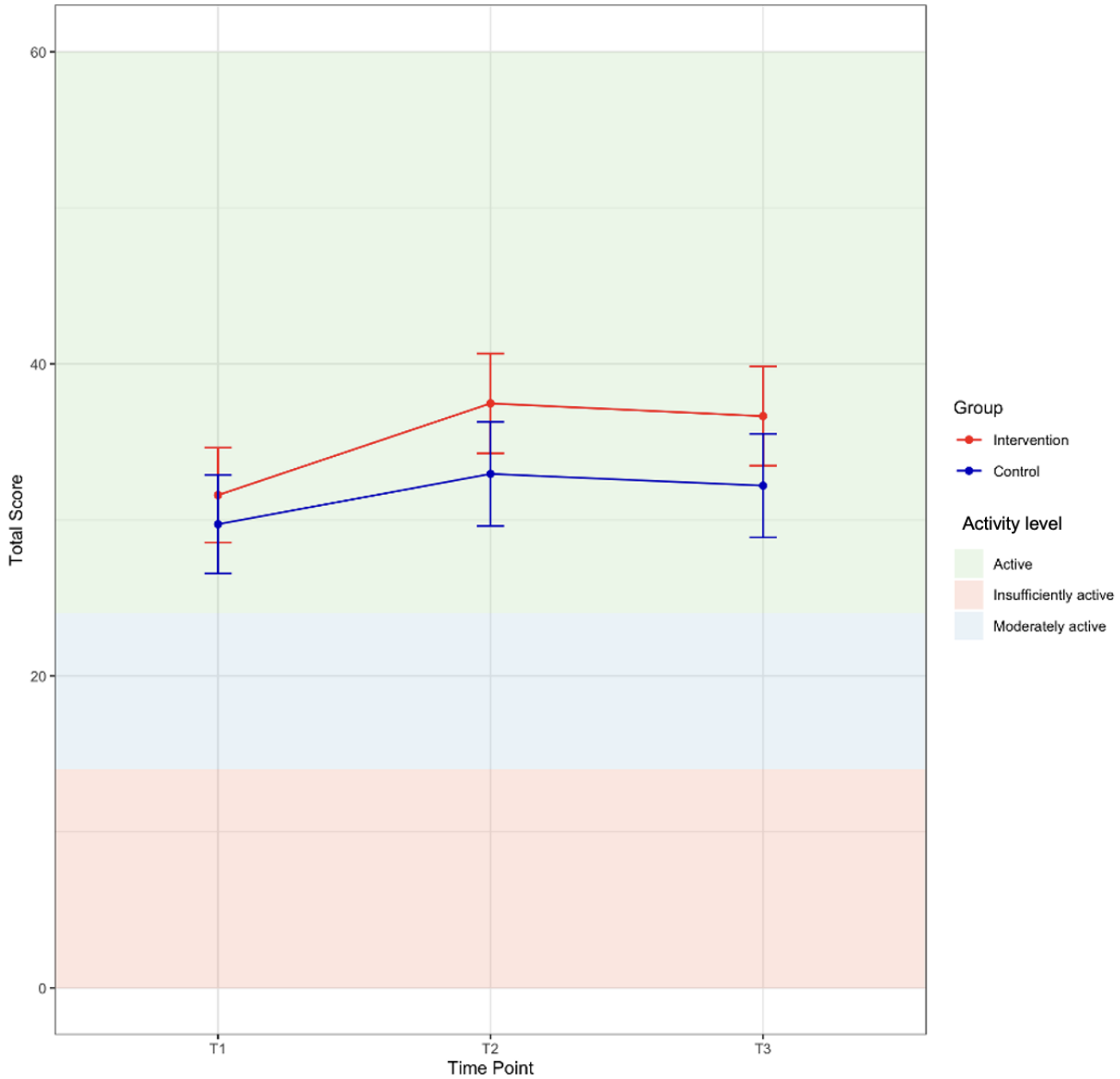

Other health behavior measures, including diet, alcohol use, sleep, and social activity, showed no significant between-group differences (Multimedia Appendix 2). PP analyses aligned with the ITT results.

IAM4all Outcomes

Using the IAM4all questionnaire at T2, participants in both groups rated the information as relevant or very relevant (intervention 94.1%, n=209; control 91.5%, n=194), easy to understand (intervention 100%, n=222; control 100%, n=212),

XSL•FO
RenderX

and beneficial (intervention 95.9%, n=213; control 91%, n=193). Nearly all participants indicated that they intend to use what they have learned (intervention 98.2%, n=218; control 97.6%, n=207), and anticipate using the information to better understand dementia prevention (intervention 60.3%, n=134; control 81.6%, n=173), discuss with someone else (intervention 63.5%, n=141; control 55.6%, n=118), or do things differently (intervention 62.1%, n=138; control 35.3%, n=75). These quantitative findings help to contextualize the qualitative themes reported in the "Qualitative Findings" section.

Qualitative Findings

Qualitative analysis was conducted inductively to generate themes and subthemes. A full list of qualitative themes and subthemes is provided in Multimedia Appendix 3. Overall, qualitative responses converged with data from the IAM4all, indicating that participants perceived the content as relevant, understandable, and actionable, with anticipated benefits for brain health and dementia risk reduction.

Relevance and Understandability

Participants commonly described the information as personally meaningful and focused on gaining a clearer understanding of dementia, including symptoms, early signs, and modifiable risk factors. Many emphasized the value of being able to recognize dementia symptoms and respond appropriately.

> *Recognize symptoms of dementia and learn the best ways to cope* [ID 551]

Participants also demonstrated a strong interest in recognizing the early signs of dementia to facilitate timely intervention. They displayed significant concern about the condition's risk factors and root causes, aiming to make informed lifestyle choices to mitigate or delay its onset.

> *Learn how to reduce the risk of dementia, strategies to delay the onset and minimize the symptoms of, dementia. Be able to recognize the signs of the onset of dementia.* [ID 298]

Intended Use and Action

Responses frequently demonstrated a proactive stance towards health, particularly in adopting lifestyle changes that could potentially delay or prevent the onset of dementia.

> *I will try to implement changes in my life based on what I learned in this.* [ID 248]

They were highly motivated to integrate new knowledge into their daily routines, making informed decisions related to diet, exercise, and mental health practices.

> *Motivates me to try and do more healthy things.* [ID 163]

> *Get motivated to do more walking, with the dogs. Maybe playing more cards with friends. I already play darts 2x a week and keep score in all games; adding and subtracting in my head while standing by the boards. Possibly learn how to play Sudoku, instead of just doing the crossword puzzles.* [ID 332]

Participants demonstrated a strong commitment to applying what they learned, with many initiating medical check-ups, adjusting their diets, and incorporating regular physical activity as direct responses to the information provided.

> *As a result of the information about the high correlation between hearing loss and dementia (higher than many other risk factors), I have made an appointment for ear wax removal, and will speak to the professional about hearing loss. Everyone should read the excellent information provided in the email campaign and surveys.* [ID 68]

> *Will go back on healthy diet, reduce sugar intake, get more exercise, more social activity.* [ID 328]

Participants frequently described increased awareness, motivation, and intentions to make lifestyle changes, and some reported early self-initiated actions.

> *This module did teach me somewhat more of what I need to do to keep myself and control my mental health, as well as physical health as I age. At 68, I do admit that I need to follow a better diet to help control my diabetes. As Emergency Medical Technicians, we were taught that we treated all causes of Dementia the same, as we could not differentiate between Diabetic, Alzheimer's or anything else...they were all forms of some mental health. Having seen the decline of my former Spouse into the Diabetic Dementia, and his refusal to even try to modify any of his routine: I know that I must make some changes to my lifestyle.* [ID 332]

Perceived Benefit

Beyond prevention, some participants valued learning about broader health connections, as well as current and emerging treatments. They appreciated the educational value of the study, which enhanced their understanding of brain health and its connections to other health issues to support more informed health and caregiving decisions.

> *I would never have believed that hearing loss might be a sign of dementia.* [ID 446]

Suggestions for Improvement

Participants also identified a desire for more comprehensive content, including practical steps for prevention and management, indicating areas for improving user experience and engagement in future studies.

> *I think that some of the suggestions for potential action could use more direction as to sources to engage in action - websites. Other information to enable people to follow up actively.* [ID 51]

The analysis further highlighted sex-specific interests, with female participants showing interest in caregiving strategies while male participants focused more on genetic risks. Additionally, across different age groups, there was a shift from a general focus on prevention in younger participants to practical management and caregiving in older adults, with a consistent appreciation for the informative impact of the study across all demographics.

XSL•FO
RenderX

Discussion

Principal Findings

The findings from this RCT demonstrate the effectiveness of the DementiaRisk internet-based education program in enhancing participants' knowledge of dementia risk and protective factors. These results were statistically significant at both the 4-week and 2-month postintervention follow-up assessments, reflecting sustained improvements post-intervention.

While both groups improved, the intervention group demonstrated significantly greater gains. This highlights the value of health education in general, but underscores the importance of providing well-designed dementia risk reduction-specific content.

The intervention improved knowledge particularly among participants with lower education levels, suggesting that e-learning may help bridge knowledge gaps among populations with potentially less prior exposure to dementia-related information. Participants who identified as having a family history of dementia or as a care partner had higher levels of baseline knowledge of dementia. Gains among those without such experience indicate that the intervention may help inform individuals who have not been previously motivated by personal exposure to dementia, thereby expanding its potential reach and impact on the general population.

Age Differences in Engagement With Preventative Measures

Older adults in our sample demonstrated greater intentions to engage in risk reduction behaviors compared to younger participants, which is consistent with prior evidence that perceived risk may increase motivation to adopt protective behaviors [43,44]. In contrast, younger adults may have lower perceived susceptibility to dementia, which can reduce their motivation to adopt lifestyle changes early in life [43,44].

Public health education efforts face unique challenges in engaging younger adults in reducing their risk of dementia. The concept of dementia as a 'disease of aging' contributes to a lack of urgency among younger populations [45]. Additionally, dementia prevention requires long-term behavior changes that may not provide immediate, tangible benefits, making it difficult to incentivize engagement in risk-reduction behaviors. Strategies such as emphasizing the broader benefits of lifestyle modifications (eg, improved cardiovascular health, mental well-being, and cognitive function) may help increase engagement among younger populations [44-47]. Framing dementia prevention within a broader "brain health" context, as suggested by the TPB, may also improve receptivity among younger individuals [48].

Interpreting Intentions to Engage in Risk Reduction

The increase in intentions observed across both groups indicates that exposure to health-related e-learning, even when not specific to dementia-risk reduction, may positively influence motivation to adopt healthier behaviors. However, measuring health behavior intentions remains a challenge due to the lack of a widely accepted, validated tool specifically designed for assessing intentions related to dementia risk reduction. The TPB suggests that intentions are strong predictors of behavior, but this relationship can be influenced by external factors such as perceived barriers or social norms [22,49,50]. Given the absence of a standardized instrument for measuring intentions in this context, our study relied on a custom-developed measure. This introduces potential limitations in comparability with other studies. Furthermore, it is important to acknowledge the "intention-behaviour gap," where individuals express a willingness to change but do not always follow through with action [50,51].

Health Behavior Change and Practical Implications

An interesting finding was the increase in self-reported physical activity levels among the participants, which suggests that the educational content may have prompted short-term behavior change. The sustained increase in physical activity observed at both follow-up points (4 wk and 2 mo postintervention) aligns with previous research suggesting that e-learning platforms, when well-designed, can encourage behavior changes beyond mere knowledge acquisition [52]. Given that physical activity is a key modifiable risk factor for dementia, this result has promising implications for the broader public health goal of dementia prevention. While the increase in physical activity levels among intervention participants is noteworthy, it should be interpreted within the context of a short-term study. Many health behaviors, particularly those related to physical activity, require sustained effort over time to produce meaningful long-term changes [53]. Some risk factors such as traumatic brain injury or hearing loss are not easily modifiable within the time frame of a 3-month study or may require additional resources or support from health care providers [54]. This distinction highlights the importance of differentiating between modifiable and nonmodifiable risk factors when designing and interpreting public health interventions.

Our study's findings align with existing research suggesting that short-term interventions are most effective for behaviors that require relatively low effort to change, such as increasing physical activity or making dietary improvements [55]. However, for more complex or chronic risk factors (eg, hypertension management and smoking cessation), longer-term interventions with ongoing reinforcement may be necessary to sustain behavior change.

Together, these results highlight that even brief, web-based interventions can produce measurable improvements in knowledge and short-term behavior, particularly when designed using evidence-based principles of instructional design. However, the practical significance and broader population impact of these changes warrant careful interpretation.

Interpretation of Findings

Across outcomes, the impact and practical relevance of observed changes varied. Knowledge of dementia risk factors increased by approximately 30% from baseline. Although modest, this gain likely reflects a meaningful increase in dementia literacy—particularly because small, scalable improvements in public awareness can produce significant population-level

XSL•FO
RenderX

benefits when applied broadly. While knowledge gains do not guarantee sustained behavior change, participants' qualitative comments indicate that many viewed the information as personally important and, in some cases, actionable in the short term, supporting the practical relevance of the knowledge gain as a precursor to informed risk-reduction decisions.

Behavioral intentions to engage in dementia risk-reduction activities also increased throughout the study, but the absolute change was small (approximately 2% - 3%) and occurred in both groups. The qualitative findings help explain these results by demonstrating that participants primarily described changes in awareness, motivation, and short-term action, particularly for behaviors perceived as immediately feasible, such as getting a hearing test, walking, or dietary adjustments, while expressing uncertainty about how to implement more complex or long-term risk-reduction strategies. This pattern suggests that exposure to health-related online education may positively influence motivation to adopt healthier behaviors, though the clinical importance of this change remains uncertain. The use of a novel intentions measure further limits comparability across studies, highlighting the need for validated scales tailored to dementia prevention contexts. In contrast, self-reported leisure-time physical activity increased by roughly 20% among participants, which may represent a meaningful short-term gain, even if participants largely remained within the same physical activity category. These effects may have been smaller in magnitude because participants were already relatively active and health conscious at baseline.

Instructional Design

One factor that may have contributed to the intervention's effectiveness is the use of evidence-based principles of instructional design. These principles, based on the Mayer Cognitive Theory of Multimedia Learning, have been shown to positively impact learning outcomes and retention [14,17]. Applying multimedia design principles such as segmenting, personalization, self-testing with feedback, and worked examples may strengthen comprehension and engagement in e-learning [15,56]. While the majority of research has examined these principles in health professionals, our findings extend this work to other general public populations.

The micro-learning email campaign also incorporated the principles of spaced education and repetition. These principles have also been shown to be effective in enhancing knowledge retention in health care professionals, but do not appear to have been systematically studied with respect to health education for the public [57,58].

Limitations and Future Directions

This study has several limitations. Although the sample was recruited from a representative panel of the Canadian population, it may not reflect populations with lower digital access or literacy. This could have introduced an element of selection bias, leading to higher baseline knowledge levels than anticipated. The sample also appeared relatively health-conscious, which may have limited the extent of behavior change. Additionally, all data were self-reported, which could introduce bias, especially in terms of behavior changes like physical activity. Finally, although a single coder performed the primary qualitative analysis, ongoing team-based verification and peer debriefing reduced potential bias and supported analytic consistency.

Conclusions

The findings from this randomized controlled trial indicate that a multimedia, instructionally designed, internet-based dementia-risk program can significantly improve dementia-related knowledge, particularly among participants with lower levels of education. We also observed modest, short-term increases in self-reported intention to engage in dementia risk–reducing behaviors and in physical activity levels in the intervention group, although effects on other behaviors were limited. Given its scalability and accessibility, this type of e-learning intervention represents a promising public health strategy to strengthen dementia risk-reduction literacy and support selected healthy behaviors at scale. Future research should broaden the program's reach and evaluate the durability of effects on knowledge, intentions, and health behaviors over time.

Acknowledgments

This work was supported by the research team at the Division of e-Learning Innovation at McMaster University. The authors would like to thank all who participated in this project for their time.

Funding

Funding was provided by the Public Health Agency of Canada Dementia Strategic Fund. The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript. AJL receives support through the John R Evans Chair in Health Sciences Educational Research and Instructional Development, Faculty of Health Sciences, McMaster University.

Data Availability

The datasets generated and analyzed during this study can be made available from the corresponding author upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

AJL and SA drafted the original study protocol. AJL, SC, and RW oversaw data collection, analysis, and interpretation. AS completed the primary quantitative analysis. DH completed the primary qualitative analysis. SA drafted the manuscript. All authors have reviewed the manuscript and approved the final version.

Conflicts of Interest

AJL, RS, and McMaster University are co-owners of DementiaRisk. The other authors declare no conflicts of interest.

Multimedia Appendix 1
Results of all models.
[XLSX File, 12 KB - jmir_v28i1e79405_app1.xlsx]

Multimedia Appendix 2
Health behavior data.
[XLSX File, 21 KB - jmir_v28i1e79405_app2.xlsx]

Multimedia Appendix 3
Qualitative themes and subthemes.
[DOCX File, 19 KB - jmir_v28i1e79405_app3.docx]

Checklist 1
CONSORT-EHEALTH checklist.
[PDF File, 1169 KB - jmir_v28i1e79405_app4.pdf]

References

1. Global status report on the public health response to dementia. : World Health Organization; 2021 URL: https://iris.who.int/server/api/core/bitstreams/9e8aab50-ca34-45e0-9fb0-26e675746f65/content [accessed 2026-02-14]
2. Prince M, Ali GC, Guerchet M, Prina AM, Albanese E, Wu YT. Recent global trends in the prevalence and incidence of dementia, and survival with dementia. Alzheimers Res Ther 2016 Jul 30;8(1):23. [doi: 10.1186/s13195-016-0188-8] [Medline: 27473681]
3. World Alzheimer Report 2021: Journey through the diagnosis of dementia. Alzheimer's Disease International. 2021 Sep. URL: https://www.alzint.org/resource/world-alzheimer-report-2021/ [accessed 2024-10-27]
4. Livingston G, Huntley J, Liu KY, et al. Dementia prevention, intervention, and care: 2024 report of the Lancet Standing Commission. The Lancet 2024 Aug;404(10452):572-628. [doi: 10.1016/S0140-6736(24)01296-0]
5. Dolatshahi Y, Mayhew A, O'Connell ME, et al. Prevalence and population attributable fractions of potentially modifiable risk factors for dementia in Canada: a cross-sectional analysis of the Canadian Longitudinal Study on Aging. Can J Public Health 2024 Dec;115(6):953-963. [doi: 10.17269/s41997-024-00920-7] [Medline: 39048849]
6. Dhana K, Evans DA, Rajan KB, Bennett DA, Morris MC. Healthy lifestyle and the risk of Alzheimer dementia: findings from 2 longitudinal studies. Neurology (ECronicon) 2020 Jul 28;95(4):e374-e383. [doi: 10.1212/WNL.0000000000009816] [Medline: 32554763]
7. Dhana K, Agarwal P, James BD, et al. Healthy lifestyle and cognition in older adults with common neuropathologies of dementia. JAMA Neurol 2024 Mar 1;81(3):233-239. [doi: 10.1001/jamaneurol.2023.5491] [Medline: 38315471]
8. The Landmark Study Report 1: Path to progress—navigating the path forward in dementia. : Alzheimer Society of Canada; 2022 URL: https://alzheimer.ca/sites/default/files/documents/Landmark-Study-Report-1-Path_Alzheimer-Society-Canada_0.pdf [accessed 2026-02-14]
9. Wimo A, Handels R, Antikainen R, et al. Dementia prevention: the potential long‐term cost‐effectiveness of the FINGER prevention program. Alzheimer's & Dementia 2023 Mar;19(3):999-1008. [doi: 10.1002/alz.12698]
10. Survey of Canadians regarding dementia prevention. Public Health Agency of Canada. 2022. URL: https://publications.gc.ca/collections/collection_2022/sc-hc/H14-414-2022-1-eng.pdf [accessed 2026-02-14]
11. Korthauer LE, Rosen RK, Tremont G, Davis JD. Intervention development for tailored education for aging and cognitive health (TEACH) for dementia prevention in midlife adults: protocol for a randomized controlled trial. JMIR Res Protoc 2024 Oct 16;13:e60395. [doi: 10.2196/60395]
12. Gong D, Peng Y, Liu X, et al. Dose health education on dementia prevention have more effects on community residents when a community physician/nurse leads it? A cross-sectional study. Front Public Health 2023;11:1101913. [doi: 10.3389/fpubh.2023.1101913] [Medline: 37206874]

XSL•FO
RenderX

13. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Instructional design variations in internet-based learning for health professions education: a systematic review and meta-analysis. Acad Med 2010 May;85(5):909-922. [doi: 10.1097/ACM.0b013e3181d6c319] [Medline: 20520049]
14. Clark R, Mayer R. e‐Learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning, 5th edition: Wiley; 2023. [doi: 10.1002/9781119239086]
15. Issa N, Schuller M, Santacaterina S, et al. Applying multimedia design principles enhances learning in medical education. Med Educ 2011 Aug;45(8):818-826. [doi: 10.1111/j.1365-2923.2011.03988.x] [Medline: 21752078]
16. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Internet-based learning in the health professions: a meta-analysis. JAMA 2008 Sep 10;300(10):1181-1196. [doi: 10.1001/jama.300.10.1181] [Medline: 18780847]
17. Mayer RE. Multimedia Learning, 3rd edition: Cambridge University Press; 2020.
18. Belleville S, Anderson ND, Bherer L, et al. Brain health PRO/Santé cerveau PRO: the development of a web-based program for dementia literacy and risk factor reduction. J Prev Alzheimers Dis 2025 Jun;12(6):100134. [doi: 10.1016/j.tjpad.2025.100134] [Medline: 40090784]
19. Levinson AJ, Ayers S, Clark S, et al. Internet-based dementia prevention intervention (DementiaRisk): protocol for a randomized controlled trial and knowledge translation. JMIR Res Protoc 2025 Jan 27;14:e64718. [doi: 10.2196/64718] [Medline: 39869903]
20. Risk reduction of cognitive decline and dementia WHO guidelines. : World Health Organization; 2019 URL: https://iris.who.int/server/api/core/bitstreams/8fc586d7-b864-42eb-9274-95462496b328/content [accessed 2026-02-14]
21. Ajzen I. From intentions to actions: a theory of planned behavior. In: Action Control: Springer; 1985:11-39. [doi: 10.1007/978-3-642-69746-3_2]
22. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
23. Rise J, Kovac V, Kraft P, Moan IS. Predicting the intention to quit smoking and quitting behaviour: extending the theory of planned behaviour. Br J Health Psychol 2008 May;13(2):291-310. [doi: 10.1348/135910707X187245]
24. Conner M, Norman P, Bell R. The theory of planned behavior and healthy eating. Health Psychol 2002 Mar;21(2):194-201. [doi: 10.1037/0278-6133.21.2.194] [Medline: 11950110]
25. Cooke R, Dahdah M, Norman P, French DP. How well does the theory of planned behaviour predict alcohol consumption? A systematic review and meta-analysis. Health Psychol Rev 2016 Jun;10(2):148-167. [doi: 10.1080/17437199.2014.947547] [Medline: 25089611]
26. Chevance G, Caudroit J, Romain AJ, Boiché J. The adoption of physical activity and eating behaviors among persons with obesity and in the general population: the role of implicit attitudes within the theory of planned behavior. Psychol Health Med 2017 Mar;22(3):319-324. [doi: 10.1080/13548506.2016.1159705] [Medline: 26952746]
27. Blue CL. The predictive capacity of the theory of reasoned action and the theory of planned behavior in exercise research: an integrated literature review. Res Nurs Health 1995 Apr;18(2):105-121. [doi: 10.1002/nur.4770180205] [Medline: 7899566]
28. Pluye P, Grad R, Repchinsky C, et al. Four levels of outcomes of information‐seeking: a mixed methods study in primary health care. J Am Soc Inf Sci Tec 2013 Jan;64(1):108-125. [doi: 10.1002/asi.22793]
29. Annear MJ, Toye C, Elliott KEJ, McInerney F, Eccleston C, Robinson A. Dementia Knowledge Assessment Scale (DKAS): Confirmatory factor analysis and comparative subscale scores among an international cohort. BMC Geriatr 2017 Jul 31;17(1):168. [doi: 10.1186/s12877-017-0552-y] [Medline: 28760154]
30. Godin G, Shephard RJ. A simple method to assess exercise behavior in the community. Can J Appl Sport Sci 1985 Sep;10(3):141-146. [Medline: 4053261]
31. Pluye P, Granikov V, Bartlett G, et al. Development and content validation of the Information Assessment Method for Patients and Consumers. JMIR Res Protoc 2014 Feb 18;3(1):e7. [doi: 10.2196/resprot.2908] [Medline: 24550180]
32. RStudio integrated development environment (IDE). Posit. 2020. URL: https://posit.co/products/open-source/rstudio/?sid=1 [accessed 2026-03-02]
33. Bates D, Mächler M, Bolker BM, Walker SC. Fitting linear mixed-effects models using lme4. J Stat Softw 2015 Oct 7;67(1):1-48. [doi: 10.18637/jss.v067.i01]
34. Lenth RV. emmeans: estimated marginal means, aka least-squares means. The Comprehensive R Archive Network. 2024. URL: https://cran.r-project.org/web/packages/emmeans/index.html [accessed 2026-02-14]
35. Wickham H. Ggplot2 Elegant Graphics for Data Analysis, 2nd edition: Springer International Publishing; 2016. [doi: 10.1007/978-3-319-24277-4]
36. Norman G, Monteiro S, Salama S. Sample size calculations: should the emperor's clothes be off the peg or made to measure? BMJ 2012 Aug 23;345:e5278. [doi: 10.1136/bmj.e5278] [Medline: 22918496]
37. Lehr R. Sixteen S-squared over D-squared: a relation for crude sample size estimates. Stat Med 1992 Jun 15;11(8):1099-1102. [doi: 10.1002/sim.4780110811] [Medline: 1496197]
38. Neil-Sztramko SE, Belita E, Levinson AJ, Boyko J, Dobbins M. Evaluation of an online knowledge translation intervention to promote cancer risk reduction behaviours: findings from a randomized controlled trial. BMC Cancer 2019 Nov 21;19(1):1138. [doi: 10.1186/s12885-019-6361-2] [Medline: 31752751]

XSL•FO
RenderX

39. Vears DF, Gillam L. Inductive content analysis: a guide for beginning qualitative researchers. Focus Health Prof Educ 2022;23(1):111-127. [doi: 10.11157/fohpe.v23i1.544]
40. NVivo (version 14). Lumivero. 2023. URL: https://lumivero.com/ [accessed 2026-02-14]
41. Jackson K, Bazeley P. Qualitative Data Analysis with NVivo, 3rd edition: Sage Publications; 2019.
42. Leventhal H, Phillips LA, Burns E. The common-sense model of self-regulation (CSM): a dynamic framework for understanding illness self-management. J Behav Med 2016 Dec;39(6):935-946. [doi: 10.1007/s10865-016-9782-2] [Medline: 27515801]
43. Hagger MS, Orbell S. The common sense model of illness self-regulation: a conceptual review and proposed extended model. Health Psychol Rev 2022 Sep;16(3):347-377. [doi: 10.1080/17437199.2021.1878050] [Medline: 33461402]
44. Cations M, Radisic G, Crotty M, Laver KE. What does the general public understand about prevention and treatment of dementia? A systematic review of population-based surveys. PLoS One 2018;13(4):e0196085. [doi: 10.1371/journal.pone.0196085] [Medline: 29672559]
45. Norton S, Matthews FE, Barnes DE, Yaffe K, Brayne C. Potential for primary prevention of Alzheimer's disease: an analysis of population-based data. Lancet Neurol 2014 Aug;13(8):788-794. [doi: 10.1016/S1474-4422(14)70136-X] [Medline: 25030513]
46. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res 2021 May 12;23(5):e17917. [doi: 10.2196/17917] [Medline: 33978589]
47. Farina FR, Booi L, Occhipinti JA, et al. Young adult brain capital: a new opportunity for dementia prevention. J Alzheimers Dis 2023;94(2):415-423. [doi: 10.3233/JAD-230260] [Medline: 37302036]
48. Ajzen I. The theory of planned behaviour: reactions and reflections. Psychol Health 2011 Sep;26(9):1113-1127. [doi: 10.1080/08870446.2011.613995] [Medline: 21929476]
49. Sheeran P, Webb TL. The intention-behavior gap. Soc Pers Psychol 2016 Sep;10(9):503-518. [doi: 10.1111/spc3.12265]
50. Rhodes RE, de Bruijn G. How big is the physical activity intention–behaviour gap? A meta - analysis using the action control framework. Br J Health Psychol 2013 May;18(2):296-309. [doi: 10.1111/bjhp.12032]
51. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010 Feb 17;12(1):e4. [doi: 10.2196/jmir.1376] [Medline: 20164043]
52. Bauman AE, Reis RS, Sallis JF, et al. Correlates of physical activity: why are some people physically active and others not? Lancet 2012 Jul 21;380(9838):258-271. [doi: 10.1016/S0140-6736(12)60735-1] [Medline: 22818938]
53. Whiteneck GG, Cuthbert JP, Corrigan JD, Bogner JA. Prevalence of self-reported lifetime history of traumatic brain injury and associated disability: a statewide population-based survey. J Head Trauma Rehabil 2016;31(1):E55-E62. [doi: 10.1097/HTR.0000000000000140] [Medline: 25931187]
54. Kwasnicka D, Dombrowski SU, White M, Sniehotta F. Theoretical explanations for maintenance of behaviour change: a systematic review of behaviour theories. Health Psychol Rev 2016 Sep;10(3):277-296. [doi: 10.1080/17437199.2016.1151372] [Medline: 26854092]
55. Issa N, Mayer RE, Schuller M, Wang E, Shapiro MB, DaRosa DA. Teaching for understanding in medical classrooms using multimedia design principles. Med Educ 2013 Apr;47(4):388-396. [doi: 10.1111/medu.12127] [Medline: 23488758]
56. Matos J, Petri CR, Mukamal KJ, Vanka A. Spaced education in medical residents: an electronic intervention to improve competency and retention of medical knowledge. PLoS ONE 2017;12(7):e0181418. [doi: 10.1371/journal.pone.0181418] [Medline: 28759606]
57. Phillips JL, Heneka N, Bhattarai P, Fraser C, Shaw T. Effectiveness of the spaced education pedagogy for clinicians' continuing professional development: a systematic review. Med Educ 2019 Sep;53(9):886-902. [doi: 10.1111/medu.13895] [Medline: 31144348]
58. A dementia strategy for Canada: Together We Aspire. Public Health Agency of Canada. 2019. URL: https://www.canada.ca/en/public-health/services/publications/diseases-conditions/dementia-strategy.html [accessed 2026-02-14]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
CONSORT-eHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
IAM4all : Information Assessment Method for Patients and Consumers
ITT: intention-to-treat
LMM: linear mixed model
MAR: missing at random
PP: per protocol
RCT: randomized controlled trial
TPB: theory of planned behavior
WHO: World Health Organization

XSL•FO
RenderX

Edited by A Stone; submitted 20.Jun.2025; peer-reviewed by N Coley, R Newton; revised version received 28.Jan.2026; accepted 28.Jan.2026; published 10.Mar.2026.

Please cite as:
Levinson AJ, Ayers S, Clark S, Woodburn R, Schneeberg A, Hadid D, Kates N, Strudwick G, Grad R, Papaioannou A, Dobbins M, Siu H, Duarte D, Saperson K, Marr S, Oliver D, Neil-Sztramko S
Effects of Internet-Based Dementia Risk Reduction Education on Risk and Protective Factor Knowledge, Intentions, and Health Behaviors: Randomized Controlled Trial
J Med Internet Res 2026;28:e79405
URL: https://www.jmir.org/2026/1/e79405
doi:10.2196/79405

© Anthony J Levinson, Stephanie Ayers, Sandra Clark, Rebekah Woodburn, Amy Schneeberg, Dima Hadid, Nick Kates, Gillian Strudwick, Roland Grad, Alexandra Papaioannou, Maureen Dobbins, Henry Siu, Dante Duarte, Karen Saperson, Sharon Marr, Doug Oliver, Sarah Neil-Sztramko. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Enhancing LGBTQ+ Inclusivity in an AI-Powered Sexual Health Chatbot: User-Centered Design Approach Through a Nonprofit and Academic Partnership

William Wibowo Liem[1], MSW; Elizabeth Casline[1], PhD; Julianna Lorenzo[1], BS; Jacob D Gordon[2], MSW, PhD; Andrés Alvarado Avila[1], MPH; Attia Taylor[3], BA; Nicole Levitz[3], MPH; Michael C O'Keefe[3], MA; Karen Shum[3], BA; Kathryn Macapagal[1,4], PhD

[1]Impact Institute, Northwestern University, Chicago, IL, United States

[2]College of Nursing, University of Cincinnati, Cincinnati, OH, United States

[3]Planned Parenthood Federation of America, New York, NY, United States

[4]Department of Medical Social Sciences, Impact Institute, Northwestern University, 625 North Michigan Avenue, Floor 14, Chicago, IL, United States

Corresponding Author:
Kathryn Macapagal, PhD
Impact Institute, Northwestern University, Chicago, IL, United States

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2026/1/e96639

Abstract

Background: Despite the growing use of digital platforms for sexual health education, many tools fail to meet the needs of LGBTQ+ (lesbian, gay, bisexual, transgender, and queer) teenagers, who often lack access to inclusive, affirming resources. Artificial intelligence (AI)–enabled chatbots have emerged as promising tools to address these gaps, but concerns remain around bias, usability, and trustworthiness—particularly for queer and transgender teenagers. Participatory design approaches centered around marginalized teenagers are critical to ensuring these tools are relevant, trustworthy, and equitable; yet, few studies have systematically engaged LGBTQ+ teenagers in the co-design of AI-powered sexual health interventions.

Objective: This paper examines LGBTQ+ teenagers' perceptions of Roo, Planned Parenthood Federation of America's (PPFA) AI-powered sexual education chatbot to identify opportunities and challenges in delivering LGBTQ+-inclusive, affirming sexual health information.

Methods: Embedded within Sharing Health Education Resources, a hybrid effectiveness implementation trial of a digital HIV prevention intervention for LGBTQ+ teenagers, we collaborated with PPFA to create a customized instance of Roo for integration into this study. We engaged a Youth Advisory Council comprising 15 LGBTQ+ teenagers to independently explore and interact with Roo, then gathered feedback through a week-long asynchronous discussion on a private Discord (Discord Inc) server. The research team posed open-ended questions prompting participants to reflect on Roo's inclusivity, usability, and content priorities. We used rapid qualitative analysis organized around our research questions.

Results: Participants expressed both skepticism and curiosity about AI's role in delivering sexual health information, offering critical insights on the chatbot's language, trustworthiness, and relevance. Teenagers identified key limitations in Roo's inclusivity, tone, and interface, particularly around transgender-specific content, conversational depth, and stigma reduction. These findings informed targeted content updates, interface refinements, and transparency improvements, implemented by PPFA to enhance Roo for broader use. Specific changes included expanding LGBTQ+ affirming content, revising language to eliminate gendered assumptions, incorporating concrete statistics and contextualized examples to reduce stigma, and adding clearer disclosures around Roo's AI capabilities and limitations.

Conclusions: Academic and nonprofit collaborations can leverage participatory methods to enhance digital health tools in real-world contexts. LGBTQ+ teenagers served not only as testers but as co-designers, shaping the chatbot's evolution and surfacing broader lessons about trust, AI literacy, and health equity. This study demonstrates that marginalized teenagers possess the critical insights needed to meaningfully shape AI-enabled health interventions when provided with structured opportunities for engagement. This partnership offers a scalable model for integrating community voice into the development, evaluation, and implementation of inclusive, AI-enabled health technologies.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e78621) doi:10.2196/78621

KEYWORDS

sexual health; sex education; adolescents; sexual and gender minorities; artificial intelligence; stakeholder participation; digital health; AI; LGBTQ+; inclusivity; lesbian, gay, bisexual, transgender, and queer

Introduction

LGBTQ+ (lesbian, gay, bisexual, transgender, and queer) teenagers experience significant sexual and reproductive health disparities compared to their cisgender, heterosexual peers. They face higher rates of sexually transmitted infections (STIs) and HIV, with young men who have sex with men accounting for 81% of new HIV diagnoses among teenagers aged 13-19 years in the United States [1]. Sexual minority individuals who can become pregnant are also at increased risk for unintended pregnancies due to higher rates of sexual victimization and coerced sex [2]. Additionally, LGBTQ+ teenagers are more likely to experience intimate partner violence than their heterosexual peers, with transgender and nonbinary teenagers reporting particularly high rates of victimization [3]. Without access to inclusive, affirming sexual health resources designed for their needs, these disparities risk worsening. Accordingly, inclusivity, usability, and perceived helpfulness emerge as key criteria for evaluating digital sexual health tools—determining whether LGBTQ+ teenagers feel represented, can navigate the platform, and find information meaningful for decision-making.

Access to comprehensive, inclusive sexual health education plays a critical role in equipping teenagers with the knowledge and skills to navigate consent, boundaries, and healthy relationships. Yet, school-based sexual health education in the United States remains overwhelmingly focused on cisgender, heterosexual experiences, often omitting or stigmatizing LGBTQ+ identities [4,5]. As a result, many LGBTQ+ teenagers turn to the internet to fill these educational gaps [6], seeking inclusive and affirming resources that traditional channels fail to provide [7-9]. In the United States, digital interventions have emerged as powerful tools for delivering inclusive sexual health education to teenagers. Platforms ranging from mobile apps to interactive websites offer accessible, affirming alternatives [10]. For LGBTQ+ teenagers in particular, online platforms can serve as safer spaces to access accurate information, connect with peers navigating similar challenges, and build critical knowledge around consent, relationships, and HIV prevention [11]. Evidence from interventions such as TechStep (designed for transgender teenagers) and MyPEEPS Mobile (tailored to same-sex attracted teenage boys) demonstrates how well-designed, targeted content can successfully reduce condomless sex practices [12,13]. The success of these programs underscores the importance of thoughtful, community-informed design; features such as user-friendly interfaces, engaging and culturally relevant messaging, and interactive components (eg, reminder texts or skill-building exercises) can significantly enhance user engagement and educational outcomes [14]. As digital tools continue to evolve, these insights offer a critical foundation for the next generation of technology-enabled interventions, including those powered by artificial intelligence (AI).

Incorporating AI in personalized health care is emerging as a promising avenue to further enhance accessibility and engagement with users [15]. However, the adoption of AI-driven tools in sexual health education has sparked both optimism and skepticism, particularly among LGBTQ+ teenagers. While chatbots can effectively deliver health interventions when designed appropriately with LGBTQ+ individuals, significant concerns remain about information quality, reading level appropriateness, and inconsistent accuracy [16,17]. These concerns are particularly acute for Black LGBTQ+ youth facing health care access barriers. A scoping review of generative conversational AI apps for the LGBTQ community found that while initial deployments were feasible and well-received, substantial improvements are needed in content quality and engagement [18]. These findings underscore both AI's potential to address health disparities and the critical need for further development, formal evaluation, and community-informed design.

The "myth of AI," as described by McAra-Hunter [19], creates unrealistic expectations that AI can universally solve complex problems, often ignoring the nuanced needs of marginalized groups. This disconnect poses risks, as AI tools are frequently developed without sufficient input from LGBTQ+ communities, leading to unrealistic expectations and potentially harmful outcomes. For example, ChatGPT (OpenAI) has been shown to use outdated or exclusionary language—assuming people with vaginas are women, and those with penises are men—and to lack sociopolitical sensitivity in its responses [20]. These findings echo broader concerns raised by Tomasev et al [21], who caution that AI systems trained on large-scale datasets without community-informed oversight risk encoding and amplifying societal biases, particularly those affecting gender and sexual minorities. Moreover, a study by Hopelab found that LGBTQ+ teenagers are more likely than their cisgender and heterosexual peers to avoid AI tools due to concerns about inaccuracies and biases [22]. Integrating teenagers' input into the design process can enhance trust, usability, and the overall experience of AI-powered interventions, ensuring they align more closely with the needs of the communities they aim to serve. However, queer youth are rarely positioned as designers of AI health tools, despite their unique expertise in navigating heteronormative and cisnormative health care systems [23].

Despite the promise of digital and AI-powered tools, research-driven innovations often face a persistent research-to-adoption gap, where effective interventions fail to achieve sustained public health impact. It is estimated that it can take an average of 17 years for research findings to be translated into real-world practice, with only a small fraction ever implemented at scale [24]. To address this gap, it is critical to embed considerations of implementation and sustainability early in the design process, especially for teenager-centered tools. This is especially important for LGBTQ+ teenagers who

XSL•FO
RenderX

often rely on digital tools to access affirming information and support that may be unavailable in their offline environments.

Participatory research approaches are particularly well-suited for engaging marginalized populations such as LGBTQ+ teenagers, as they center community expertise, address power imbalances inherent in research, and ensure interventions reflect the lived realities of those most affected [25]. By involving teenagers as co-designers rather than passive recipients, participatory methods not only improve intervention relevance and acceptability but also build trust with communities that have historically experienced harm or exclusion from medical institutions [26]. A key implementation is developing nonprofit and academic partnerships that bridge the research-to-adoption gap while facilitating rigorous co-design and evaluation methods. Evidence shows that involving teenagers in intervention planning leads to better reflection of their needs and perspectives while addressing power imbalances in decision-making [27]. Coproducing interventions with teenagers increases their sense of ownership and supports adoption, with 1 review reporting more than a 2-fold increase in self-reported use of sexual and reproductive health services [28]. For these partnerships to maximize lasting public health benefit, there is a high need to prioritize cocreation, effectively manage expectations early, and integrate implementation planning from the outset.

To illustrate how aligning missions across sectors can accelerate impact, we describe the nonprofit and academic collaboration between Planned Parenthood Federation of America (PPFA) and Northwestern University aimed to enhance Roo (Figure 1), an existing, publicly available PPFA-operated AI-powered platform, through a week-long asynchronous discussion with LGBTQ+ teenagers on Discord.

Figure 1. Roo chatbot interface (Planned Parenthood, accessed May 31, 2025). Roo, Planned Parenthood Federation of America's free, web-based AI-powered sexual education chatbot. The interface features example prompts (eg, "What happens during puberty?" and "How long after unprotected sex can I know if I'm really pregnant?"), a privacy assurance message, and an open text input field for user-generated questions. Roo's database covers topics including abortion, birth control, relationships, LGBTQ+ health, and symptoms. AI: artificial intelligence; BTW: by the way; LGBTQ+: lesbian, gay, bisexual, transgender, and queer.

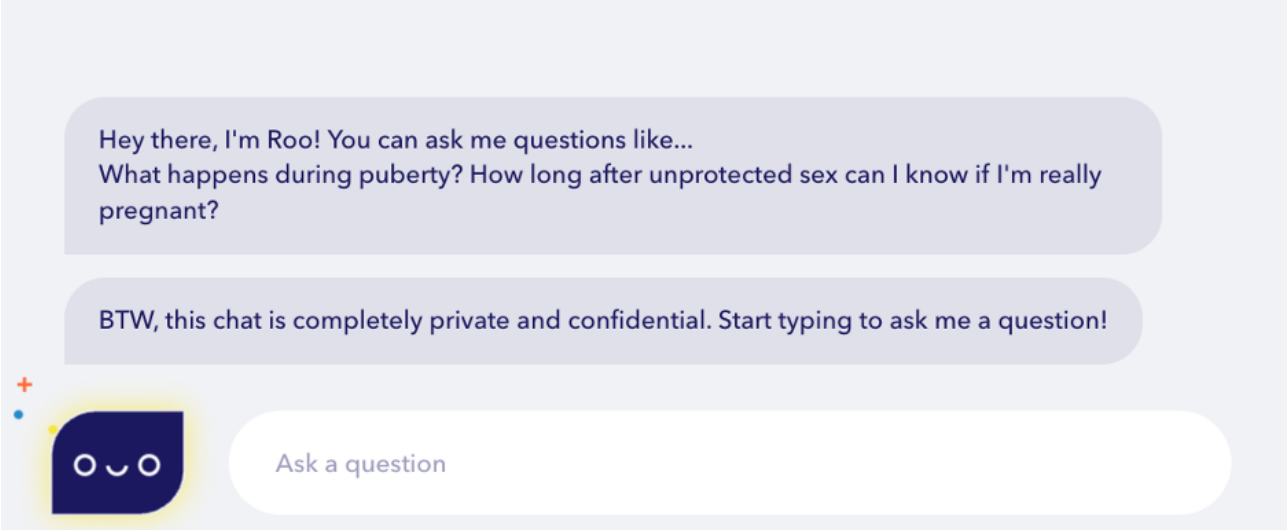

Methods

Study Design

This study was conducted as part of a hybrid type 1 effectiveness implementation randomized controlled trial examining the effectiveness of a digital text-based HIV prevention intervention called Sharing Health Education Resources (SHER) compared to an information-only control among LGBTQ+ teenagers. SHER builds on an earlier text-messaging intervention called Guy2Guy (G2G), a text messaging-based HIV prevention program to deliver educational and skill-oriented information to male teenagers who identify as gay, bisexual, and/or queer [29]. Within SHER, Roo, Planned Parenthood's AI-powered sexual education chatbot, was adapted and evaluated as a component intervention. For the purposes of this paper, we focus specifically on Roo and report findings from Youth Advisory Council (YAC) feedback about its inclusivity, usability, and trustworthiness.

The original G2G intervention featured an on-demand texting function, G2Genie, that allowed participants to request additional information by sending keywords (eg, texting "G2Genie relationships" would prompt a curated set of messages about relationship skills). Building on this concept—and recognizing the growing emphasis on personalization and rapid delivery in digital health interventions—SHER planned to adapt this feature by developing a bespoke sexual education chatbot feature tailored to the specific needs of LGBTQ+ teenagers [30,31]. The goal of the chatbot was to enable participants to ask follow-up questions and receive individually tailored responses across a broader range of sexual health topics than those included in the SHER text messages (Figure 2). By collaborating with a nonprofit resource already integrated into public health infrastructure, Northwestern University was able to accelerate translation, broaden reach among diverse teenage audiences, and leverage existing dissemination pathways to enhance both impact and equity. Rather than building a proprietary chatbot accessible only to study participants, our

XSL•FO
RenderX

team strategically partnered with an existing, publicly available platform, PPFA's Roo.

Figure 2. Integration of the Roo chatbot into the SHER digital HIV prevention intervention for LGBTQ+ teenagers. The screenshot demonstrates how participants who enrolled in the SHER text messaging intervention (left panel) received a prompt directing them to Roo, Planned Parenthood's AI-powered sexual education chatbot (right panel). The customized Northwestern University instance of Roo was accessible via a dedicated microsite and provided on-demand sexual health information, complementing SHER's structured intervention content. AI: artificial intelligence; LGBTQ+: lesbian, gay, bisexual, transgender, and queer; PEP: post-exposure prophylaxis; PrEP: pre-exposure prophylaxis; SHER: Sharing Health Education Resources; TasP: Treatment as Prevention.

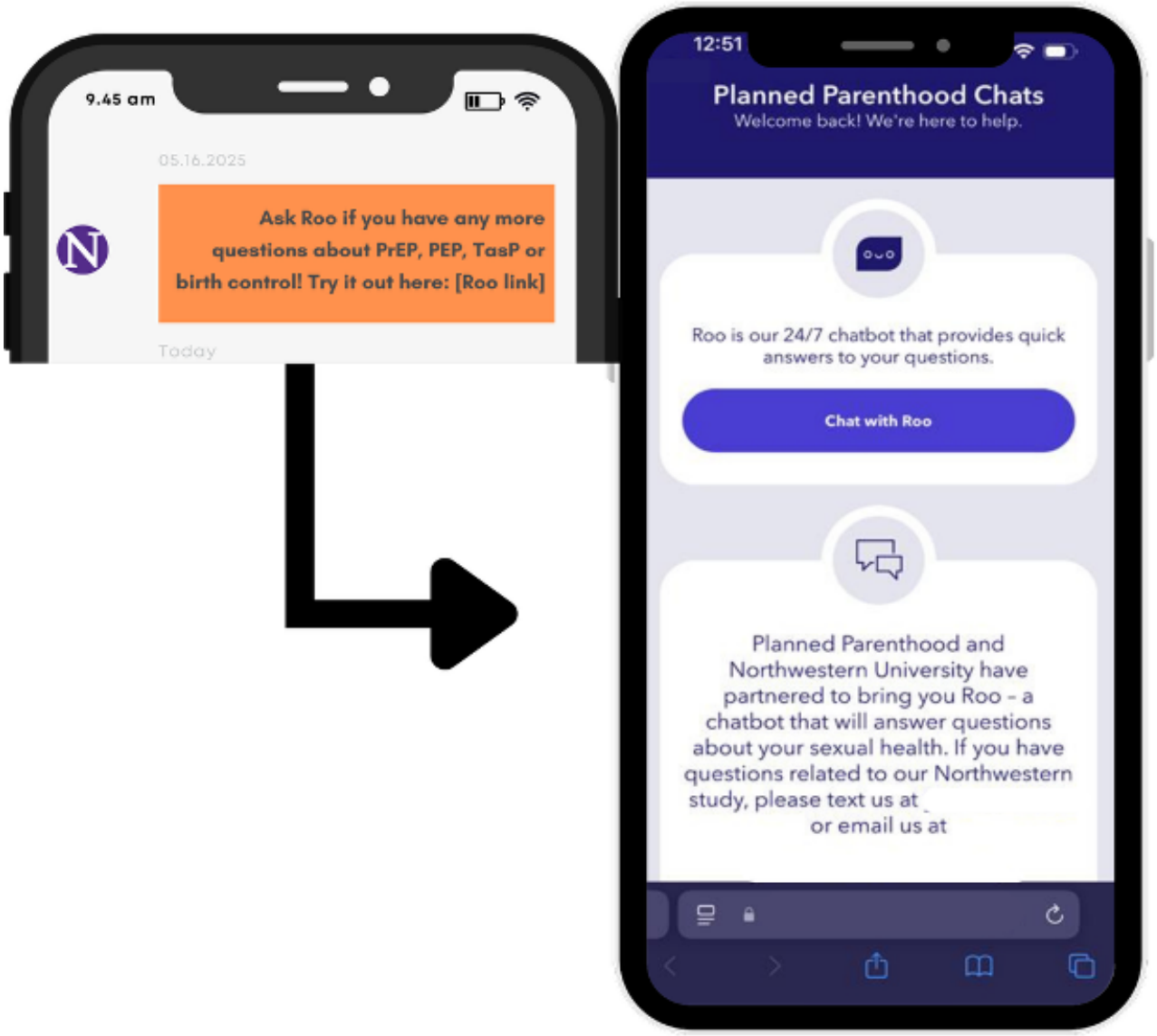

Roo is an AI-powered chatbot designed to provide confidential, accessible, and medically accurate sexual health information. Roo's responses are developed by Planned Parenthood's content and education teams to maintain a friendly, nonjudgmental tone through intentional strategies such as using conversational language (eg, "You deserve fun, comfortable, and pleasurable sex"), avoiding clinical jargon, affirming diverse experiences as normal, and framing sensitive topics without moral judgment. For instance, rather than emphasizing risk, Roo normalizes STI testing as routine care (eg, "Getting tested is often as easy as peeing in a cup at a health center") and highlights that "most of the time, STDs don't show symptoms, so the only way to know for sure is to get tested." Accessible via the PPFA website [32], Roo allows users to ask free-text questions and responds immediately with a nonjudgmental answer and a list of additional resources for the user to engage with.

Roo's development is grounded in the Unified Theory of Behavior, which integrates multiple behavioral theories to explain and influence health behaviors through attitudes, normative beliefs, self-efficacy, and behavioral intention [33]. Roo's content and education teams crafted responses to both provide information while supporting users in setting behavioral intentions. This application of the theory is operationalized through several strategies: presenting pros and cons to support informed decision-making, acknowledging emotions users may

XSL•FO
RenderX

experience, connecting users to health care resources, and providing referrals to partner organizations. For example, when discussing contraception options, Roo provides balanced information about effectiveness and side effects while prompting users to consider their personal priorities and next steps, such as scheduling a clinic appointment. All educational content on Planned Parenthood's website, including Roo, reflects this framework by helping users identify action plans and set intentions for next steps in their sexual health journey. To our knowledge, Roo is the only sexual and reproductive health chatbot that explicitly applies the Unified Theory of Behavior in its design and content strategy, distinguishing it from other AI-powered tools focused primarily on information delivery rather than behavioral support.

Roo uses natural language processing to scan each question for keywords to trigger the correct answer or send the user to another resource that can answer their question. Natural language processing uses machine learning and deep learning models to recognize patterns in the human language. Roo is a Planned Parenthood proprietary product that draws exclusively from a curated library of responses designed specifically for Roo users. All content is crafted and consistently iterated upon by Planned Parenthood's technology and sex education leadership teams, with all responses conforming to the federation's medical standards and guidelines. This human-curated approach ensures that responses maintain medical accuracy, age-appropriateness, and alignment with PPFA's mission of providing inclusive, affirming sexual health education. Topics in Roo cover a wide range of sexual and reproductive health topics. The questions that Roo can answer are limited by the database of content that Roo is able to pull from. The content team at Planned Parenthood uses data from Roo, other digital products, and user experience research to add common and frequently asked questions and their responses into Roo. If Roo cannot answer a user's question, it will let them know and offer alternatives, including PPFA's anonymous live chat with real health educators, as well as links to PPFA's blog and other educational resources.

In 2023, Roo facilitated approximately 640,000 sessions [34]. Roo offers a complementary approach to SHER's content and structure. Whereas SHER delivers structured, sequential educational messages with preplanned daily content, Roo is designed to offer on-demand, conversational access to medically accurate, affirming, and developmentally appropriate sexual health education. Roo's database covers a broad range of topics relevant to teenagers, including safe sex practices, consent, healthy relationships, LGBTQ+ health, and navigating health care systems.

This complementary functionality made Roo a promising addition to SHER, enhancing the program's ability to support LGBTQ+ teenagers' diverse information needs without duplicating content delivery. For PPFA, Roo has provided an opportunity to align its mission with innovative delivery models; however, implementing an AI-enabled tool that adequately meets the needs of LGBTQ+ teenagers has required technical and evaluation infrastructure beyond PPFA's in-house capacity. By collaborating with an academic institution, PPFA was able to benefit from Northwestern University's methodological expertise to co-design an intervention evaluation to receive actionable recommendations for iteratively refining Roo's responsiveness to LGBTQ+ teenagers' needs.

Academic and Nonprofit Partnership Collaborative Development Process

We sought to develop an academic and nonprofit partnership that advances both scientific rigor and real-world impact in the development of inclusive digital health innovations. Below, we describe the development of this academic and industry partnership using Proctor et al [35] domains for systematically specifying implementation strategies.

Speaking of actors, the partnership was initiated by the Northwestern University research team, who became aware of PPFA's Roo chatbot and were connected through a mutual colleague. Key actors included academic investigators, digital health researchers, product managers, and health education specialists from both institutions.

For action, we developed a collaborative relationship with PPFA that required iterative conversations to align organizational goals, resources, and capacities. Academic and industry collaborations in digital health often face challenges around aligning timelines, balancing rigor and innovation, and navigating competing priorities [30]. Early meetings focused on relationship-building, clarifying organizational priorities, and identifying areas of mutual benefit. Specific actions included assessing the fit between Roo and the SHER intervention, mapping resource needs, identifying technical constraints, establishing intervention goals across both teams, and drafting and negotiating formal agreements, including a nondisclosure agreement and data use agreement to facilitate ethical collaboration and data sharing. Given the reciprocal benefits identified, both parties agreed that there would be no financial exchange.

Subsequent meetings explored integration pathways, ultimately deciding to create a standalone customized microsite version of Roo for SHER to facilitate tracking participant engagement and allow key modifications to maintain intervention integrity. These changes included removing the option to talk to a live health educator and directing users who needed help with Roo to call the SHER study team rather than chat with PPFA. This instance was built and maintained by PPFA's product team, with all feedback integrated into both the SHER study instance and the publicly facing version of Roo.

Once the partnership was established, Northwestern University engaged its YAC—a diverse group of LGBTQ+ teenagers to test Roo and provide structured feedback on its content, user experience, and inclusivity.

Of temporality and dose, partnership development occurred over 6 months (February to July 2023) with monthly formal meetings supplemented by asynchronous collaboration, with higher-frequency early communication for trust-building and later focus on YAC feedback and platform customization.

Of target and justification, the overarching target was to integrate Roo into SHER in a way that enhanced accessibility, inclusivity, and personalization of sexual health resources for

XSL•FO
RenderX

LGBTQ+ teenagers. PPFA aimed to expand its reach to LGBTQ+ teenagers, while Northwestern University aimed to apply participatory design principles centering LGBTQ+ youth voices. Partnering with an existing, trusted platform embedded in public health infrastructure allowed the project to accelerate translation and broaden access beyond this study's sample.

Ethical Considerations

Concerning ethics approval, all study procedures were approved by the Northwestern University Institutional Review Board (study number: STU00217358) with a waiver of parental consent, given the potential risks of requiring guardian permission for LGBTQ+ minors.

Of informed consent, for minors, informed assent and a capacity-to-consent assessment were required. Eligibility was confirmed through screening forms and one-on-one video calls with study staff, which also served to review participation expectations and privacy protocols.

Of privacy and confidentiality, the YAC operates on a private, invite-only Discord server, where participants can use pseudonyms to protect their identity. Study staff moderated the server to ensure a safe, respectful environment. Participants could withdraw from this study at any time without penalty, and all data were stored securely in compliance with institutional data protection standards. The principal investigator is a licensed clinical psychologist with expertise in LGBTQ+ adolescent health, ensuring clinical oversight of all activities involving sensitive sexual health discussions.

Concerning compensation, participants received US $30 in digital gift cards per month for active participation, with opportunities to earn up to US $200 annually for optional activities.

Of the use of images, no identifiable participant images are included in this manuscript or supplementary materials.

Recruitment

As part of the SHER effectiveness implementation trial, LGBTQ+ teenagers were recruited through paid Meta (Facebook or Instagram) advertisements and an institutional participant registry. Interested teenagers were directed to a secure REDCap (Research Electronic Data Capture; Vanderbilt University) screening survey to complete eligibility questions and provide contact information. To reduce impostor enrollment, identity verification included a brief video call assessing personal information and comprehension of study details [36-38]. Following a virtual orientation session conducted via Zoom (Zoom Communications, Inc), eligible participants were invited to a private Discord server, an instant messaging platform that supports multimedia messaging services, voice, video, and multimedia communication.

Eligible participants were aged 13‑18 years, identified as sexual and/or gender minority, lived in the United States or US territories, and were able to read in English at an eighth-grade level. Recruitment sought to oversample transgender or gender-diverse and racial or ethnic minority teenagers to ensure diverse feedback. Recruitment occurred between January 2023 and July 2023 and was conducted entirely online with no in-person involvement. The YAC ultimately included 15 LGBTQ+ teenagers aged 15‑18 (mean 16.9) years (see Table 1 for demographics). For additional details on the YAC methods, see the study by Gordon et al [39].

XSL•FO
RenderX

Table . Demographic characteristics of the YAC[a] participants, including age, race or ethnicity, gender identity, and sexual orientation. The YAC comprised 15 LGBTQ+[b] adolescents (aged 15-18, mean 16.9 years) recruited from across the United States between January and July 2023 via paid social media advertisements and an institutional participant registry. The sample included 9 (60%) gender-diverse teenagers and 6 (40%) teenagers of color. All participants identified as sexual minorities, with the majority identifying as bisexual, pansexual, or gay.

Category	Count (n=15)
Age (years), n (%)	
15	1 (7)
16	3 (20)
17	8 (54)
18	3 (20)
Race or ethnicity, n (%)	
White	9 (60)
American Indian or Alaska Native	2 (13)
Black or African American	1 (7)
Asian	1 (7)
Mestizo Latin American	1 (7)
Not disclosed	1 (7)
Hispanic or Latino	3 (20)
Gender identity, n (%)	
Cisgender man or boy	6 (40)
Transgender man or boy	3 (20)
Transgender woman or girl	4 (27)
Gender nonconforming	1 (7)
Questioning gender identity	1 (7)
Sexual orientation, n (%)	
Bisexual or pansexual	7 (47)
Gay	5 (35)
Queer	2 (14)
Questioning	1 (7)

[a]YAC: Youth Advisory Council.

[b]LGBTQ+: lesbian, gay, bisexual, transgender, and queer.

Participatory Feedback Process With LGBTQ+ Teenagers

The YAC is a standing, Discord-based advisory group of LGBTQ+ teenagers who provide structured input across multiple digital health projects. Before the Roo evaluation, participants reviewed intervention materials and were advised on website features, familiarizing them with research goals and developing confidence as co-designers. The Roo evaluation was therefore embedded within a broader program of participatory engagement, with inclusivity, usability, and perceived helpfulness as central evaluation criteria.

The participatory feedback process was informed by a research-through-design orientation study, which treats interventions as iterative, generative artifacts shaped through ongoing engagement and inquiry [40]. Rather than approaching Roo as a static digital tool to be evaluated in isolation, we conceptualized it as a dynamic prototype whose development was guided by LGBTQ+ teenagers' feedback and contextualized within their lived experiences. This orientation supported creativity by encouraging teenagers to propose new content areas (eg, gender-affirming care and sex toy safety) and interface redesigns (eg, category-based navigation and dark mode), which moved beyond minor adjustments to reimagining Roo's functionality. It also enhanced responsiveness by enabling rapid incorporation of feedback into both the SHER-specific and public-facing versions of Roo, ensuring that participants saw their input translated into visible changes. Finally, research through design positioned social and structural factors (eg, stigma in school-based sex education, barriers to gender-affirming care, and mistrust of AI systems) as central design considerations, shaping not only what content was added but also how information was framed to reduce harm and foster trust.

XSL•FO
RenderX

To prepare for the use of Roo for the intervention, the Northwestern University team and PPFA were engaged over a period of 5 months to create a Northwestern University–specific instance of Roo that would be used in the intervention. This included establishing a data use agreement, removing the health educator feature, which could be considered contamination with RCT outcomes given its behavioral component, and adding study contact information. The Northwestern University instance of Roo was shared with the YAC.

The YAC provided feedback on all components of the SHER intervention design and implementation over approximately 18 months and was focused on providing Roo feedback over 1 week. As part of this evaluation, our analysis was guided by three research questions: (1) What do LGBTQ+ teenagers perceive as gaps and/or limitations in Roo's current content, particularly regarding inclusivity and relevance to their lived experiences? (2) How do LGBTQ+ teenagers evaluate the usability, tone, and format of Roo's chatbot interface? (3) Under what conditions do LGBTQ+ teenagers find chatbots such as Roo helpful or unhelpful for accessing sexual health information?

To address these questions, participants were invited to interact with Roo independently and provide feedback on its usability, inclusivity, and perceived helpfulness.

Over the course of 1 week, the research team posted 3 open-ended questions in the Discord server to guide feedback. These questions prompted participants to (1) explore Roo and share initial impressions; (2) reflect on inclusivity, especially for LGBTQ+ and transgender and gender-diverse users, and suggest improvements to Roo's interface and structure; and (3) identify additional topics Roo should cover more effectively, such as sexual assault, miscarriage, sex toys, and gender inclusivity. Participants had 7 days to respond to any or all questions at their own pace.

Each question was posted in the main research channel and was accompanied by a dedicated discussion thread to organize all responses. Participants provided written replies, reacted to peers' comments using emojis, and engaged in follow-up exchanges within threads. Emoji reactions were used as indicators of consensus or shared resonance across YAC members.

Data for this study consisted of text responses, emoji reactions, and threaded follow-up conversations. After data collection, raw Discord content was organized into summaries by question, capturing recurring themes and illustrative quotes. In line with member-checking practices, summaries were shared with YAC members for review before being provided to the Roo team [41].

Data Analysis

We used rapid qualitative analysis (RQA) to conduct content analysis to synthesize participant insights across our 3 research questions. RQA is an increasingly adopted approach in applied health research contexts, emphasizing structured data summarization over line-by-line coding to generate timely, actionable findings [42]. RQA is particularly well-suited for participatory projects that seek to elevate community voices while maintaining momentum in design iteration cycles [43]. Consistent with the content analysis approach, our RQA focused on a close reading of the text to identify substantive feedback that could inform actionable design improvements rather than developing emergent themes through open coding [44]. As such, our code development process prioritized capturing concrete, implementable recommendations within a time-bounded participatory design framework, an approach suitable for applied health intervention research where timely translation is essential.

The analysis was conducted collaboratively between members of the Northwestern University research team and PPFA's digital content strategy team to ensure findings were both academically rigorous and actionable for implementation. The analytic team included the lead author (WWL, MSW, and a doctoral student with training in qualitative methods and LGBTQ+ health), a research assistant professor with expertise in implementation science and LGBTQ+ adolescent health (EC, PhD), and a research project manager with experience in community-based participatory research (AAA, MPH).

The lead author conducted the initial analysis, organizing participant feedback into a matrix where each row represented a unique participant and each column corresponded to 1 of the 3 research questions. Feedback was organized to identify substantive codes such as specific critiques, suggestions, and design recommendations that could inform actionable improvements. Responses that addressed multiple research questions were mapped across domains to reflect the interrelated nature of the data. The lead author then conducted a systematic review of the coded content to identify recurring patterns, frequently mentioned issues, and high-priority recommendations, revisiting the data for each research question domain and identifying the most actionable codes grounded in participant language. This process emphasized iterative refinement, prioritizing content with the strongest support and clearest implementation pathways and was conducted until no new substantive actionable content codes were identified. Particular attention was given to points of consensus—where feedback was echoed or affirmed across participants—as well as divergence, where responses reflected differing or underrepresented perspectives. Agreement indicators such as "likes" and emoji reactions were systematically documented as they signaled areas of shared resonance and helped prioritize implementation efforts.

The preliminary analysis was reviewed collaboratively by the Northwestern University analytic team (WWL, EC, and AAA) to ensure accuracy, comprehensiveness, and alignment with the research questions. Following this internal review, the refined findings were shared with PPFA's Associate Director of Digital Content Strategy (AT), who brought expertise in sexual health education, digital product development, and youth engagement. The cross-organizational team then conducted a collaborative review to (1) validate themes and ensure they accurately reflected user feedback and (2) translate findings into concrete, implementable content and design changes. PPFA's implementation process systematically prioritized changes based on frequency of feedback, level of participant endorsement (indicated by likes or reactions), feasibility, and alignment with organizational capacity for rapid implementation. This

XSL•FO
RenderX

partnership approach directly informed real-world product improvements.

Results

Overview

The YAC provided feedback across three domains mapping to our research questions: (1) content gaps and relevance, (2) design and usability, and (3) trust and AI transparency.

RQ1: Content Gaps, Depth, and Relevance

Content Depth and Specificity

YAC participants consistently emphasized the need for Roo to go beyond foundational sexual health content and engage more meaningfully with topics specific to LGBTQ+ teenagers. While Roo provided accurate general information, participants found its responses insufficiently nuanced for addressing the kinds of "medium-depth" questions they were most likely to ask. These questions, while not highly clinical, often reflected real-life scenarios central to their identities and relationships. Several participants specifically flagged the absence of content on sexual assault, miscarriages, the safe use of sex toys, and queer-specific sexual practices. One highly endorsed comment by 5 other council members summarized this concern:

> *Why doesn't Roo have any clickable question buttons about sexual assault, sex toys, and miscarriages.* [P1, 5 likes]

This reflects a broader expectation among teenagers that a credible sexual health tool should anticipate sensitive topics and proactively make them accessible, rather than placing the burden on teenagers to generate the correct wording. Others echoed this sentiment, highlighting that Roo did not adequately respond to moderately detailed prompts related to LGBTQ+ health, with 1 participant noting, "For a lot of medium-depth questions even related [to sexual health], it didn't have a reply" (P4, 4 likes).

Gender-Affirming Care and Transgender Inclusion

A particularly salient gap concerned Roo's limited content on transgender and nonbinary health needs. Participants described the absence of information on gender-affirming care (eg, hormone therapy, safe binding practices, and navigating health care as a transgender person) as both alienating and harmful. As P9 explained,

> *It's supposed to be for transgender people like myself… but it's not so inclusive, which would deter me from reading further most times.* [P9]

This sense of exclusion was compounded by the broader difficulty teenagers described in accessing accurate, affirming sexual health information elsewhere. As 1 participant put it, "It's so hard to get accurate queer transgender sex ed" (P1, 6 likes), highlighting that Roo's shortcomings do not exist in isolation but within an already sparse landscape of transgender-inclusive health resources.

Taken together, these responses reveal that transgender-specific content is not a secondary feature but a core component of trust-building for this population. Its absence signals that the platform may not fully see or support them. Including gender-affirming information is therefore not only about comprehensiveness, but it is also about ensuring that transgender teenagers can see their realities reflected in the resource and feel affirmed and empowered to engage with it for their sexual health needs.

Reducing Stigma in Health Information

Participants emphasized the importance of presenting sexual health information in ways that actively reduce stigma, particularly around STIs and developmental norms. Teenagers called for content that normalizes diverse experiences rather than pathologizing them. As one participant noted, "We need info that gets rid of stigma on STIs" (P1, 2 likes), underscoring the need for nonjudgmental, affirming messaging.

At the same time, participants highlighted how even well-intentioned information could inadvertently reinforce stigma when framed without sufficient context. P3 reflected:

> *It just completely negates the ' oh it's okay you can take your time' message… it may also pressure someone to have [sex] by 17, which is awful.* [P3, 3 likes]

Rather than reassuring teenagers that there is no single timeline for sexual experiences, content that references normative milestones without careful framing was seen as inadvertently pressuring or alienating. Participants also expressed frustration with overly vague or evasive responses to personal questions. As P4 explained, "If you're gonna say something like 'everybody is different,' then I'd rather just not get an answer… I'd like to see statistics… instead of just being told it depends" (7 likes).

Such comments highlight the risk of vague responses not only being unhelpful but also eroding trust in the platform's authority. These insights suggest that effective stigma reduction requires more than inclusive language. It demands concrete, contextualized information that normalizes a range of sexual health experiences and empowers teenagers to feel validated, rather than judged, in their individual journeys.

RQ2: Design and Usability

Conversational Functionality and Interface

Participants had mixed reactions to Roo's chat-based format. While some appreciated its friendly, conversational tone, others found it cumbersome for retrieving information efficiently. Some teenagers likened the chatbot to a "complicated FAQ" (P8, 3 likes) where the conversational structure introduced unnecessary layers of interaction to access straightforward answers. As 1 participant explained:

> *It doesn't really feel like a chatbot if you're limited to pre-written questions.* [P4]

To address these concerns, participants proposed alternative interface designs that would offer greater flexibility, speed, and user control. Several teenagers suggested repositioning the chat input box to the bottom of the screen, aligning with familiar web design conventions and streamlining user interactions. Others advocated for adding features such as dark mode to

XSL•FO
RenderX

enhance visual accessibility, particularly for users who prefer reduced screen brightness during extended sessions. The idea of categorizing topics into a dropdown menu structure (rather than relying solely on "popular questions") was seen as critical to improving navigation and reducing the frustration associated with multiple conversational turns.

One participant (P9) proposed a comprehensive redesign that combined esthetic appeal with functionality, illustrating their vision through a self-created layout (Figure 3). Their design emphasized the importance of reducing friction in the user experience, advocating for organized topic categories (eg, "Sex>Vaginal sex>Resources"), easy-to-access FAQs, and a visually appealing, dark mode interface. Beyond esthetics, they stressed the need for Roo to maintain credibility by linking responses to trusted sources and providing external resources for users who wished to learn more. P9 explains:

> *People don't come to the AI like a social media platform—they come to ask questions about sexual health… a category dropdown that leads into a list of questions gives a more balanced system where everyone gets the same answer.* [P9]

XSL•FO
RenderX

Figure 3. Alternative interface mockup created by P9 for Roo. The proposed redesign features a hierarchical dropdown menu structure (eg, “sex”→“vaginal sex”→“resources”) to improve navigation efficiency, dark mode for visual accessibility, a FAQ section, and an “ask Roo” chat function. This design reflects teenagers’ critiques of Roo’s original chat-based interface as cumbersome for information retrieval and their desire for categorized topic organization rather than conversational navigation. AI: artificial intelligence; FAQ: frequently asked question.

In addition to these interface suggestions, P9 proposed backend enhancements to strengthen Roo’s reliability and responsiveness. These included using a GPT-4 model trained specifically on sexual health content, a “report AI” button for flagging problematic outputs, and an “ask an agent” feature to allow users to escalate complex or sensitive questions to trained human health educators. Collectively, these ideas reflect a broader, teenager-centered vision for Roo—not merely as a chatbot, but as an integrated, trustworthy health information platform designed to meet the expectations of digitally savvy teenagers. This emphasis on transparency, accessibility, and user agency echoed broader themes across YAC feedback, highlighting a shared desire for efficient, affirming, and user-friendly sexual health resources.

XSL•FO
RenderX

Inclusive and Accurate Language Use

Language emerged as a significant concern, particularly with some responses defaulting to cisgender or heteronormative assumptions. For instance, a phrase such as "Does it hurt for girls to have sex for the first time?" was flagged as exclusionary by participants, who noted that such wording could alienate gender-diverse users. Teenagers highlighted the importance of using gender-neutral language to ensure that Roo feels inclusive and affirming for all users, regardless of gender identity.

> *I think it's important it be able to answer questions even if they're phrased in a gender essentialist way–a young cis girl is probably going to ask something like that rather than ' does it hurt to have sex for the first time if you have a vagina.'* [P6, 4 likes]

As P6 noted, this included rephrasing responses to be both inclusive and accurate, such as changing the example to "Is it painful for people with vaginas to have sex for the first time?" Incorporating such thoughtful language choices was seen as a key step toward making the tool more accessible and welcoming for LGBTQ+ teenagers, while still providing accessible answers for cisgender heterosexual teenagers.

RQ3: Transparency, Trust, and AI Literacy

Transparency and Limitations of Roo's AI Capabilities

Participants expressed mixed perceptions of Roo's trustworthiness as an AI-driven sexual health tool. While many teenagers appreciated Roo's friendly, nonjudgmental tone—describing it as more approachable than other online sources—concerns emerged about its perceived lack of adaptability and clarity about its underlying technology. Some participants found responses too generic or static, raising doubts about Roo's ability to handle complex or highly individualized questions. Others were unsure whether they were interacting with an AI, a scripted FAQ, or a live human, which undermined their confidence in the chatbot. As 1 participant explained:

> *AI learning models are still young and may not always have what you're looking for, so I think that a live agent would be a good addition for more complex questions. They would be trained to provide accurate factual information about more complex questions that Roo could not.* [P9]

Teenagers emphasized the need for Roo to be more explicit about its AI nature and its limitations, advocating for clear disclosures, transparent sourcing, and pathways to escalate questions beyond the chatbot when needed. The ambiguity surrounding Roo's status, paired with inconsistent depth in responses, left some users hesitant to trust the platform fully.

AI Literacy and Critical Perspectives on AI Use

Beyond their critiques of Roo specifically, participants demonstrated a striking degree of AI literacy and sophistication in evaluating the role of conversational AI in sexual health education. Teenagers engaged in nuanced discussions comparing Roo to general-purpose large language models (LLMs) such as ChatGPT. Some participants expressed enthusiasm about ChatGPT's conversational style, with 1 participant remarking, "Yeah I think almost everyone has a positive attitude to ChatGPT. I like it so much. It's been really helpful" (P10, 2 likes). Others, however, voiced caution about LLMs' well-documented tendency to produce plausible-sounding but inaccurate information. As 1 participant warned, "It does literally just lie sometimes… if the AI gives out misleading information, it's going to have a lot of bad effects on the people who use the service" (P8, 6 likes). Rather than passively consuming information, participants positioned themselves as critical evaluators of AI outputs, demonstrating awareness of risks and suggesting safeguards. Several teenagers even articulated concrete design pathways for improving Roo, imagining how a model fine-tuned on credible sources could balance depth with accuracy.

> *If trained specifically on trusted sources… it could be, while not perfect, essentially very in-depth GPT—but rather than general questions, it could be way more in-depth on sexual health.* [P9]

These discussions revealed a nuanced tension: while teenagers were receptive to the idea of more dynamic, conversational AI, they also demanded safeguards such as expert oversight, transparent sourcing, and clear escalation pathways to human support. Their critical perspectives suggest that LGBTQ+ teenagers are not only capable of evaluating emerging technologies but are poised to play a vital role in shaping safer, more accountable AI-driven health resources.

These critiques not only highlighted teenagers' ability to interrogate the promises and risks of AI but also offered a roadmap for actionable improvements. In particular, their emphasis on the need for clear disclosures, sourcing, and escalation pathways informed Roo's subsequent revisions around transparency and live educator integration. Teenagers' calls for AI trained on trusted content also underscored the importance of expanding Roo's database and linking responses to vetted resources, changes that Planned Parenthood prioritized in the next iteration of the platform.

Implemented Changes

In response to YAC feedback, PPFA implemented targeted modifications to Roo between May 2023 and October 2024. Changes fell into 3 primary categories: content expansion, language and tone revision, and transparency and interface improvements.

Content Expansion

For transgender and gender-affirming care, PPFA added 15 new responses covering testosterone therapy, chest binding, and navigating health care as a transgender person. These included practical guidance (eg, "Taking testosterone can make your period lighter, less frequent, or make them totally go away") and access information (eg, "Contact your nearest Planned Parenthood health center to see if they offer gender-affirming hormone therapy").

For queer-specific sexual practices, new content addressed anal health and safety, clitoral anatomy and function, and sexual practices such as scissoring, providing both pleasure-focused information and safety guidance. Additional entries covered topics that teenagers flagged as absent or insufficient, including sex toy safety, miscarriage, and sexual assault resources.

XSL•FO
RenderX

To address concerns about vague or overly general responses, PPFA incorporated more concrete statistics and contextualized examples throughout Roo's responses. This shift responded directly to participants' calls for information that normalizes diverse sexual health experiences while providing actionable, specific guidance rather than platitudes such as "everybody is different."

Language and Tone Revisions

PPFA systematically revised Roo's language to eliminate gendered assumptions and heteronormative framing while emphasizing affirmation and stigma reduction. Gender-essentialist prompts were rewritten to use gender-neutral phrasing (eg, changing "Does it hurt for girls to have sex for the first time?" to "Is it painful for people with vaginas to have sex for the first time?"). This approach ensured that responses remained accessible to cisgender users while affirming gender-diverse teenagers.

STI-related content was updated to explicitly include diverse sexual practices and partnership types, moving beyond heteronormative assumptions. Tone modifications shifted from clinical risk language to affirming, nonjudgmental statements. For example, anal sex information was reframed to state "Anal sex is healthy when done safely. Using lube, going slow, and avoiding anal sex when you have hemorrhoids can help," positioning pleasure and safety as compatible rather than opposing concerns. Similarly, STI testing was normalized as routine self-care: "Getting tested is often as easy as peeing in a cup at a health center."

These revisions addressed participants' concerns that even well-intentioned content could inadvertently reinforce stigma through framing that emphasized risk, pathology, or normative developmental timelines without adequate context.

Transparency and Interface Improvements

To address concerns about Roo's ambiguous identity and perceived limitations, PPFA implemented several transparency-focused changes. Roo's introductory text was updated to more clearly identify it as an AI-powered tool with specific capabilities and constraints, directly responding to participants' confusion about whether they were interacting with a live person, a scripted FAQ, or an AI system.

Recognizing teenagers' desire for human escalation pathways, PPFA added alternative prompts (eg, "Can I talk to a real person?" and "I want to chat to a human") to facilitate seamless connection to live health educators through Planned Parenthood's chat services. While this live chat feature has always been part of Roo's broader infrastructure, it was temporarily removed from the version used during the Northwestern University study due to research constraints. Nevertheless, participants' strong interest in live support, evidenced by proposals for "ask an agent" features, reaffirms its value as a critical complement to AI-driven answers for handling complex or sensitive questions.

Further efforts were made to increase Roo's credibility and transparency through the inclusion of "learn more" links within responses, guiding users to vetted Planned Parenthood materials or trustworthy external sites. This addressed teenagers' concerns about insufficient sourcing and their desire to verify information or explore topics in greater depth.

While structural changes such as adding a full category-based dropdown system, integrating dark mode, or upgrading to more advanced backend AI models (as suggested by P9) remain under consideration for future development phases, PPFA prioritized user experience improvements that could be implemented rapidly. Conversational pathways were streamlined to reduce the number of steps needed to access frequently asked questions, and the language in suggested prompts was reviewed to avoid heteronormative assumptions.

These enhancements reflect Roo's continued evolution into a transparent and user-centered tool, one that better communicates its capabilities and limitations while expanding teenagers' access to reliable, affirming health information. Table 2 summarizes key areas of participant feedback alongside corresponding updates made to the platform, illustrating how LGBTQ+ teenager perspectives directly informed and improved this AI-driven health tool.

XSL•FO
RenderX

Table . Summary of YAC[a] feedback and corresponding updates to Roo. The table maps participant-identified limitations across 7 domains (content depth, transgender inclusion, stigma reduction, chatbot functionality, language inclusivity, AI[b] transparency, and AI literacy) to specific design and content changes implemented by PPFA[c].

Gap identified	Participant feedback	Implemented changes
Content depth and specificity	Need for more nuanced content on medium-depth questions (eg, sexual assault, sex toys, miscarriages, and queer-specific practices).	Added new entries on these topics to improve coverage of LGBTQ+[d] experiences.
Gender-affirming care and transgender inclusion	Lack of content on hormone therapy, binding, and transgender health care alienates transgender users.	Introduced content on testosterone, binding, and accessing affirming care.
Reducing stigma in health information	Content felt pathologizing or vague; need for nonjudgmental framing and clear statistics.	Revised tone to reduce stigma, included more examples or statistics to contextualize experiences.
Chatbot functionality	Chat felt like a complicated FAQ[e]; suggested dropdowns, dark mode, GPT-4 backend, and escalation to humans.	Streamlined conversational flow; future consideration for dropdown interface and backend upgrades.
Inclusivity and language use	Gendered or heteronormative language alienated users. Requested gender-neutral rephrasings.	Reviewed or revised language in prompts and answers to be gender-neutral and inclusive.
Transparency of AI identity	Confusion about whether Roo was AI or human reduced trust.	Updated introduction text to clarify Roo's AI identity and limitations.
AI literacy and oversight	Teenagers raised concerns about hallucinations and lack of sourcing; requested live agent escalation and better links.	Emphasized pathways to live health educators and added "learn more" links to vetted sources.

[a]YAC: Youth Advisory Council.

[b]AI: artificial intelligence.

[c]PPFA: Planned Parenthood Federation of America.

[d]LGBTQ+: lesbian, gay, bisexual, transgender, and queer.

[e]FAQ: Frequently Asked Questions.

These updates illustrate how teenager-led feedback can shape not just what information is delivered, but how it is structured, presented, and framed. Importantly, many of the changes implemented (eg, improvements to inclusive language, clearer sourcing, expanded topic coverage, etc) speak to both informational and relational aspects of trust-building in AI-based health communication.

Discussion

Principal Results

This study demonstrates how participatory design with LGBTQ+ teenagers can meaningfully shape the development of AI-powered sexual health tools. By partnering with PPFA to gather structured feedback on Roo, an AI-enabled chatbot, we engaged a YAC of LGBTQ+ teenagers in identifying gaps, proposing redesigns, and surfacing concerns about trust, inclusivity, and usability. Participants highlighted several key areas for improvement: the need for more in-depth and affirming content (particularly around transgender health, queer-specific sexual practices, and stigma reduction), more flexible and intuitive interface design, and greater transparency about Roo's limitations as an AI-driven system.

Teenagers also offered critical insight into the epistemic and ethical challenges posed by rule-based AI systems in health contexts, comparing Roo's limitations to the conversational fluidity and risks of LLMs such as ChatGPT. Rather than rejecting AI altogether, participants articulated a desire for hybrid systems that combined the responsiveness of AI with safeguards such as vetted content, human escalation pathways, and transparent sourcing. Their contributions not only informed direct improvements to Roo's content, interface, and tone but also pointed toward broader design principles for teenager-centered, accountable AI in health education.

Reimagining Trust and the Role of AI in Sexual Health Education

Teenagers' critiques of Roo centered on its perceived limitations as a rule-based system—an AI approach that relies on preprogrammed responses to anticipated queries rather than dynamically generating answers based on context. While many appreciated Roo's nonjudgmental tone, they described the platform as rigid, generic, and limited in depth, particularly when responding to nuanced or medium-complexity questions. Participants noted that while rule-based systems can offer predictability and safety, they often struggle to meet users' evolving or individualized information needs, especially in sensitive areas such as LGBTQ+ sexual health.

Several teenagers compared Roo to more dynamic tools such as ChatGPT, noting that while Roo felt safe and reliable, it sometimes lacked the adaptability or richness of a conversational AI. These critiques mirror broader limitations of digital health resources, which often lack the contextual responsiveness needed to address users' diverse needs, especially in areas such

XSL•FO
RenderX

as mental health and identity. Research has shown that many digital tools fail to provide accessible, personalized, or comprehensible information, limiting their effectiveness in real-world contexts [45]. Although static resources offer advantages in terms of accuracy and control, they often struggle to serve teenagers navigating sensitive, intersectional, or highly personalized sexual health concerns [46,47].

Yet, it is important to balance these concerns with evidence demonstrating AI's potential promise in health care. Systematic reviews show that conversational agents can effectively deliver health interventions when designed appropriately, with research indicating that chatbots can improve patient engagement, provide timely information access, and demonstrate acceptable accuracy [48]. Studies of user acceptance show that when people believe AI technology will be useful, this can serve as a strong predictor of adoption among LGBTQ+ individuals [49]. In contrast, dynamic and interactive tools (eg, chatbots or AI-based conversational agents) offer real-time responsiveness and perceived personalization, which may foster deeper engagement and affirmation, particularly for LGBTQ+ teenagers seeking information not readily reflected in mainstream health education [50].

Our findings extend this literature by demonstrating how trust is co-constructed through participatory design processes that enable users to contest and refine system outputs. While participants raised concerns about Roo's limitations, research demonstrates that conversational agents can be highly effective in health care when designed with transparency and accountability in mind [48,51]. The framework by Shneiderman [52] for trustworthy AI emphasizes the importance of human control and oversight, principles echoed in teenagers' proposals for "report AI" buttons and live agent escalation. Our findings extend this framework by demonstrating how trust emerges not from system perfection but from honest acknowledgment of constraints coupled with meaningful avenues for user input and recourse.

These expectations for responsiveness were closely tied to broader notions of trust. Existing research suggests that teenagers are more likely to trust AI health tools when systems clearly communicate their limitations, provide opportunities for human escalation, and demonstrate credibility through transparent sourcing [53]. Liao and Sundar [54] argue that designing for responsible trust in AI systems requires clear communication strategies that help users understand system capabilities and limitations. Trust was not predicated simply on whether a system used AI, but on how well it acknowledged uncertainty, supported user agency, and created channels for recourse when information felt insufficient or inaccurate. In this sense, trust and accountability were deeply intertwined: participants wanted tools they could rely on, not because they were infallible, but because they made space for users to push back, seek clarification, and surface unmet needs.

At the same time, participants demonstrated striking AI literacy. Teenagers critically discussed the strengths and risks of integrating LLMs such as ChatGPT into health tools, raising concerns about hallucinated facts, citation gaps, and the overconfidence of generative outputs. While some valued the conversational range offered by ChatGPT, others emphasized the potential harm of inaccurate information in high-stakes health contexts. Rather than rejecting AI altogether, participants proposed pragmatic solutions, such as fine-tuning models on verified health content, linking outputs to reputable sources, and integrating moderation or "report" features that would allow users to flag problematic responses and request additional review. These reflections signal that LGBTQ+ teenagers are not only capable of identifying epistemic risks but are actively imagining systems of accountability. Their feedback points toward a broader vision of participatory AI governance, one in which teenagers are positioned not as passive users, but as costewards of tools designed for their well-being.

Centering Teenage Voices to Advance Equity in AI-Enabled Health Tools

Roo is one of the few publicly available AI-powered chatbots designed to provide on-demand sexual health information to teenagers. Unlike static websites, Roo offers scalable, interactive engagement across a broad spectrum of questions, positioning it as a unique resource within an otherwise sparse landscape of LGBTQ+ inclusive digital health tools. Roo was developed in part to address the information marginalization faced by LGBTQ+ teenagers whose needs are often excluded from mainstream sexual health education. Through this project, Roo incorporated content specifically tailored to queer and transgender teenagers, including information on gender-affirming care, STI stigma, and mental health in sexual contexts. Yet despite these intentions, participants identified ongoing limitations in Roo's ability to represent and respond to the full scope of their lived experiences. Feedback highlighted not just content gaps, but structural constraints, such as rigid dialog flows and static language, that limited Roo's responsiveness, particularly for users with intersectional or stigmatized identities. These observations reinforced a critical lesson: content inclusivity is necessary, but not sufficient, for digital equity. The form and functionality of the system must also reflect the realities and needs of those it intends to serve.

To that end, Roo's participatory development process offers an important model for advancing equity in AI-based health interventions. Our engagement with the YAC gestures toward what Constanza-Chock [55] terms "design justice," in its effort to foreground the experiences of those most affected by the technology. LGBTQ+ teenagers were treated as co-designers whose insights materially shaped both the chatbot's interface and content. Their feedback drove refinements to language, layout, and tone, ensuring that Roo felt accessible, affirming, and useful to those it was built for. Participants' sophisticated critiques align with the observation by Yang et al [56] that human-AI interaction presents unique design challenges that require iterative, user-centered approaches. By positioning teenagers as epistemic contributors, we demonstrate how participatory methods can address the guidelines for human-AI interactions proposed by Amershi et al [57], particularly around transparency, user control, and system adaptability. These design contributions align with research showing that participatory methods, particularly when embedded early and iteratively, can produce tools that are not only more usable but also more trusted and ethically aligned [55]. Beyond improving Roo itself, the

XSL•FO
RenderX

process also demonstrated how co-design with marginalized teenagers can serve as a form of epistemic justice: making space for community knowledge to shape what counts as valid, visible, and valuable in digital systems.

Importantly, this process also cultivated a form of applied AI literacy among participants. Teenagers evaluated its limitations, questioned its logic, and proposed sophisticated alternatives, such as integrating GPT-4 with guardrails or flagging outputs for review. Engaging teenagers in the design process encouraged them to question the intelligence and trustworthiness of AI, equipping them with the skills to navigate the complexities of digital health technologies. This aligns with the growing recognition that AI literacy is essential for empowering marginalized populations to not only use AI tools effectively but also challenge their limitations [58]. By blending participatory design with transparent and inclusive AI practices, this study highlights a replicable model for fostering critical engagement with AI—one that enables LGBTQ+ teenagers to transition from passive users to informed, empowered stakeholders.

Privacy and Confidentiality Considerations

Interestingly, while participants raised concerns about content accuracy and trustworthiness, explicit concerns about data privacy were not prominent in their feedback. This finding contrasts with research suggesting that LGBTQ+ teenagers often prioritize anonymity and privacy in online health-seeking [59]. The relative absence of privacy concerns may reflect Roo's design as a web-based tool requiring no account creation or personal information, or it may suggest that content quality and inclusivity were more salient concerns than data practices. Roo's design incorporates privacy-protective features to encourage honest engagement: no account creation, login, or personal information is required, and users access the platform anonymously via web browser. However, this study did not directly assess whether teenagers experienced self-censorship despite these protections. Future research should examine whether privacy concerns shape the types of questions teenagers feel comfortable asking AI health tools.

However, this should not be interpreted as evidence that privacy is unimportant to LGBTQ+ teenagers accessing sexual health information. The broader literature on digital health emphasizes that privacy concerns intersect with identity safety and institutional trust, particularly for populations experiencing health inequities [60]. For LGBTQ+ teenagers who may face family rejection or school-based discrimination, privacy in health-seeking is not merely a preference but a safety imperative. Future research should explicitly probe privacy expectations and concerns among LGBTQ+ teenagers using AI health tools, particularly as these systems become more personalized and data-intensive.

Implications for Human-Centered AI, Health Informatics, and Public Health

This work contributes to human-computer interaction scholarship by demonstrating how participatory design with marginalized teenagers can surface unique requirements for AI systems that serve vulnerable populations. While guidelines for human-AI interaction emphasize transparency and user control, our findings reveal additional considerations specific to LGBTQ+ teenagers: the need for affirming language, intersectional content, and explicit acknowledgment of AI limitations [57]. These insights extend existing human-computer interaction frameworks and suggest that design guidelines must be iteratively refined through engagement with diverse user communities.

From a health informatics perspective, our findings corroborate observations that health chatbots require improvements in content quality and engagement depth, while extending this to show that effectiveness depends on attending to relational and cultural dimensions of care [61]. The emphasis participants placed on inclusive language, transgender-specific content, and stigma-reducing framing highlights that health information systems must address both informational accuracy and the social contexts in which health information is sought and used.

For public health practice, this study underscores AI's potential to address health disparities when development processes are actively centered around marginalized communities. However, realizing this potential requires sustained participatory engagement rather than 1-time consultation, as demonstrated by Madaio et al [62]. Co-designing checklists and interventions with affected communities helps organizations understand and address bias in AI systems while building trust directly reflected in our partnership model.

Lessons Learned From Academic and Nonprofit Partnership

This project exemplifies the value of cross-organizational collaboration in advancing equitable digital health tools, while also surfacing important lessons for future partnerships.

Complementary Organizational Capacities Accelerate Impact

Universities and nonprofits bring distinct but synergistic strengths. Academic institutions offer methodological expertise and research infrastructure; nonprofits bring public health infrastructure and capacity for rapid implementation. This partnership demonstrated how combining these capacities collapsed the research-to-practice timeline, with teenage feedback translated into platform updates within months rather than years.

Early Alignment on Values and Decision-Making Prevents Conflict

Establishing clear agreements about decision-making authority was essential. Northwestern University maintained control over research design and analysis; PPFA retained authority over content implementation. This division preserved both scientific integrity and organizational autonomy but required extensive upfront communication and shared commitment to health equity values. Organizations considering similar partnerships should assess both structural agreements and value compatibility early.

XSL•FO
RenderX

Partnership Infrastructure Requires Dedicated Resources and Sustained Engagement

While this model offers significant advantages, it demands substantial coordination that is often unfunded or underestimated. Faculty time for relationship-building, legal negotiations, and ongoing communication represents real costs, as does nonprofit capacity for research collaboration that may compete with direct service priorities.

Implications for Practice and Future Directions

This study highlights both the promise and complexity of developing AI-enabled health education tools for LGBTQ+ teenagers. Roo demonstrates how chatbots can expand access to affirming, immediate, and nonjudgmental information, particularly for teenagers navigating sensitive, stigmatized, or nuanced questions. For sexual health education in particular, questions often defy simple rule-based answers and demand context, empathy, and nuance. Our findings suggest that participatory design should not end once a tool is launched; instead, teenagers must be meaningfully engaged in ongoing feedback, evaluation, and refinement cycles. Designing for accountability means enabling mechanisms for teenagers to contest, correct, or expand system responses, and treating these interventions not as critiques to be managed, but as vital contributions to more inclusive, responsive AI systems.

This study also reinforces the value of embedding participatory design as a continuous, rather than episodic, element of tool development. LGBTQ+ teenagers brought critical perspectives that shaped Roo's language, tone, content depth, and usability, insights that could not have emerged without their sustained involvement. Importantly, the co-design process also cultivated AI literacy, empowering teenagers to assess the risks and capabilities of emerging technologies. As generative AI tools proliferate in health and education settings, there is an urgent need for implementation models that support participatory governance, foster critical engagement, and include marginalized teenagers not only as users but as epistemic contributors. Future efforts might explore how such participatory frameworks can be institutionalized across digital health product life cycles, from ideation to deployment to postlaunch evaluation.

Scalability will also require stronger cross-sector partnerships. This project benefited from the combined expertise of a public health nonprofit and a university-based research team, enabling agile iteration while preserving scientific rigor. Finally, more research is needed to evaluate the long-term impact of AI-based sexual health tools on teenage health outcomes, particularly among queer and transgender teenagers. Future studies should explore not only what works, but for whom, under what conditions, and with what ethical guardrails in place.

Limitations

While our YAC included participants across gender identities and races or ethnicities, the sample was limited. Of 15 participants, 9 (60%) identified as White, suggesting recruitment methods may have inadvertently excluded some racial and ethnic minority teenagers whose perspectives may differ regarding institutional trust and health care bias. The sample size was relatively small, and participants self-selected into this study, introducing potential bias toward those already motivated to engage with sexual health tools. While we examined how gender identity shaped experiences with Roo, we did not systematically analyze how histories of gender-based discrimination—including transphobia and transmisogyny—influenced teenagers' trust in AI health tools. Future research should intentionally oversample LGBTQ+ teenagers of color using culturally responsive methods and explicitly examine how systemic marginalization informs interactions with AI platforms.

The customized instance of Roo used during this study differed from the publicly available version—most notably by removing the live chat feature due to research-related monitoring constraints. This may have influenced perceptions of trust, usability, and comprehensiveness. Additionally, this study did not capture or analyze individual conversation logs between participants and Roo, instead relying on participants' reflections posted in Discord discussions. While this preserves privacy and reduces data burden, future studies might benefit from analyzing actual conversation data (with consent) to identify specific interaction patterns or content gaps not fully captured through retrospective self-report.

Although our YAC discussions did not capture which specific sources teenagers consider trustworthy, prior work suggests that teenagers often place greater confidence in well-established health organizations, including the Centers for Disease Control and Prevention and the World Health Organization, when information is perceived as transparent, relevant, and easy to understand [63]. At the same time, marginalized teenagers may be more likely to trust organizations with reputations for inclusive and affirming care, such as Planned Parenthood, which align more closely with their lived experiences and needs [64]. Future research should directly examine which health information sources LGBTQ+ teenagers consider credible, to better inform how chatbots such as Roo can direct users to external resources.

Finally, while our academic and nonprofit partnership enabled rapid feedback translation and impact, it also surfaced challenges around legal, operational, and infrastructural coordination. Negotiating data use, content ownership, and design authority required extensive collaboration, and the scalability of this model may depend on the capacity and alignment of future partnering institutions. Nonetheless, this case offers a valuable proof of concept for how cross-sector partnerships can meaningfully center community voices in AI health tool development.

Conclusions

This project illustrates the importance of designing AI-enabled health tools not just for marginalized teenagers, but with them—grounding development in participatory design, cross-sector collaboration, and iterative accountability. Roo's evolution through feedback from LGBTQ+ teenagers highlights the critical need to center those most affected by information gaps in shaping both the content and infrastructure of digital health tools. More than just expanding access to stigma-free sexual health information, Roo's development process surfaced deeper questions about how AI systems earn trust, respond to critique, and evolve alongside community needs. By treating

XSL•FO
RenderX

teenagers as epistemic contributors, this work offers a model for how AI can be governed more ethically, transparently, and equitably. The collaboration between Northwestern University and Planned Parenthood demonstrates how academic and public partnerships can meaningfully integrate community voices into scalable interventions that are not only technically sound but socially accountable.

Acknowledgments

We gratefully acknowledge our Teen Health Lab Youth Advisory Council members for their thoughtful insights and continued engagement with us.

Funding

Research reported here was supported by the National Institute of Mental Health under award numbers R01MH129207 (principal investigator: Macapagal). JDG's time was supported by T32MH130325 (principal investigator: Newcomb). The funding agencies had no role in the research or manuscript preparation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

Data from this study may be shared with qualified researchers upon reasonable request to the corresponding author, contingent upon execution of appropriate data sharing agreements that maintain participant privacy and confidentiality protections.

Authors' Contributions

KS reviewed the manuscript and provided approval of the final work.

Conflicts of Interest

None declared.

References

1. CDC. Estimated HIV incidence and prevalence. HIV Data. 2025 Feb 7. URL: https://www.cdc.gov/hiv-data/nhss/estimated-hiv-incidence-and-prevalence.html [accessed 2026-01-30]
2. Charlton BM, Everett BG, Light A, et al. Sexual orientation differences in pregnancy and abortion across the lifecourse. Womens Health Issues 2020 Mar;30(2):65-72. [doi: 10.1016/j.whi.2019.10.007]
3. Langenderfer-Magruder L, Walls NE, Whitfield DL, Brown SM, Barrett CM. Partner violence victimization among lesbian, gay, bisexual, transgender, and queer youth: associations among risk factors. Child Adolesc Soc Work J 2016 Feb;33(1):55-68. [doi: 10.1007/s10560-015-0402-8]
4. Gowen LK, Winges-Yanez N. Lesbian, gay, bisexual, transgender, queer, and questioning youths' perspectives of inclusive school-based sexuality education. J Sex Res 2014;51(7):788-800. [doi: 10.1080/00224499.2013.806648] [Medline: 24003908]
5. Goldfarb ES, Lieberman LD. Three decades of research: the case for comprehensive sex education. J Adolesc Health 2021 Jan;68(1):13-27. [doi: 10.1016/j.jadohealth.2020.07.036] [Medline: 33059958]
6. Guse K, Levine D, Martins S, et al. Interventions using new digital media to improve adolescent sexual health: a systematic review. J Adolesc Health 2012 Dec;51(6):535-543. [doi: 10.1016/j.jadohealth.2012.03.014] [Medline: 23174462]
7. Proulx CN, Coulter RW, Egan JE, Matthews DD, Mair C. Associations of LGBTQ-inclusive sex education with mental health outcomes and school-based victimization in US high school students. J Adolesc Health 2019;64(5):608-614. [doi: 10.1016/j.jadohealth.2018.11.012]
8. Delmonaco D, Harper G, Li S, Paneda C, et al. Information needs for an online resource for LGBTQ+ young people: mental health, sexual health, and navigating services. Currents: J Diversity Scholarship Soc Change 2022;2(1). [doi: 10.3998/ncidcurrents.1774]
9. Hobaica S, Kwon P. "This Is how you hetero:" sexual minorities in heteronormative sex education. Am J Sex Educ 2017 Oct 2;12(4):423-450. [doi: 10.1080/15546128.2017.1399491]
10. Wadham E, Green C, Debattista J, Somerset S, Sav A. New digital media interventions for sexual health promotion among young people: a systematic review. Sex Health 2019 Apr;16(2):101-123. [doi: 10.1071/SH18127] [Medline: 30819326]
11. Steinke J, Root-Bowman M, Estabrook S, Levine DS, Kantor LM. Meeting the needs of sexual and gender minority youth: formative research on potential digital health interventions. J Adolesc Health 2017 May;60(5):541-548. [doi: 10.1016/j.jadohealth.2016.11.023] [Medline: 28108088]
12. Benkeser D, Horvath K, Reback CJ, Rusow J, Hudgens M. Design and analysis considerations for a sequentially randomized HIV prevention trial. Stat Biosci 2020 Dec;12(3):446-467. [doi: 10.1007/s12561-020-09274-3] [Medline: 33767798]

XSL•FO
RenderX

13. Schnall R, Kuhns LM, Pearson C, et al. Efficacy of MyPEEPS mobile, an HIV prevention intervention using mobile technology, on reducing sexual risk among same-sex attracted adolescent males: a randomized clinical trial. JAMA Netw Open 2022 Sep 1;5(9):e2231853. [doi: 10.1001/jamanetworkopen.2022.31853] [Medline: 36129712]
14. Martin P, Alberti C, Gottot S, Bourmaud A, de La Rochebrochard E. Young people's proposals for a web-based intervention for sexual health promotion: a French qualitative study. BMC Public Health 2023 Jul 19;23(1):1389. [doi: 10.1186/s12889-023-16257-8] [Medline: 37468863]
15. Hengstler M, Enkel E, Duelli S. Applied artificial intelligence and trust—the case of autonomous vehicles and medical assistance devices. Technol Forecast Soc Change 2016 Apr;105:105-120. [doi: 10.1016/j.techfore.2015.12.014]
16. de Cock C, Milne-Ives M, van Velthoven MH, Alturkistani A, Lam C, Meinert E. Effectiveness of conversational agents (virtual assistants) in health care: protocol for a systematic review. JMIR Res Protoc 2020 Mar 9;9(3):e16934. [doi: 10.2196/16934] [Medline: 32149717]
17. Darien K, Lee S, Knowles K, et al. Health information from web search engines and virtual assistants about pre-exposure prophylaxis for HIV prevention in adolescents and young adults: content analysis. JMIR Pediatr Parent 2023 Jul 18;6:e41806. [doi: 10.2196/41806] [Medline: 37463044]
18. Bragazzi NL, Crapanzano A, Converti M, Zerbetto R, Khamisy-Farah R. The impact of generative conversational artificial intelligence on the lesbian, gay, bisexual, transgender, and queer community: scoping review. J Med Internet Res 2023 Dec 6;25:e52091. [doi: 10.2196/52091] [Medline: 37864350]
19. McAra-Hunter D. How AI hype impacts the LGBTQ + community. AI Ethics 2024 Aug;4(3):771-790. [doi: 10.1007/s43681-024-00423-8]
20. De Vito A, Colpani A, Moi G, et al. Assessing ChatGPT's potential in HIV prevention communication: a comprehensive evaluation of accuracy, completeness, and inclusivity. AIDS Behav 2024 Aug;28(8):2746-2754. [doi: 10.1007/s10461-024-04391-2] [Medline: 38836986]
21. Tomasev N, McKee KR, Kay J, Mohamed S. Fairness for unobserved characteristics: insights from technological impacts on queer communities. arXiv. Preprint posted online on Apr 28, 2021. [doi: 10.48550/ARXIV.2102.04257]
22. Teen and young adult perspectives on generative AI. : Hopelab; 2024 URL: https://assets.hopelab.org/wp-content/uploads/2024/05/Teen-and-Young-Adult-Perspectives-on-Generative-AI.pdf [accessed 2026-01-30]
23. Liem W, Berry A, Macapagal K. Reclaiming power over AI: equipping queer teens as AI designers for HIV prevention. arXiv. Preprint posted online on Jun 18, 2024. [doi: 10.48550/arXiv.2406.13018]
24. Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. J R Soc Med 2011 Dec;104(12):510-520. [doi: 10.1258/jrsm.2011.110180] [Medline: 22179294]
25. Vaughn LM, Whetstone C, Boards A, Busch MD, Magnusson M, Määttä S. Partnering with insiders: a review of peer models across community-engaged research, education and social care. Health Soc Care Community 2018 Nov;26(6):769-786. [doi: 10.1111/hsc.12562] [Medline: 29512217]
26. Flicker S. Who benefits from community-based participatory research? A case study of The Positive Youth Project. Health Educ Behav 2008 Feb;35(1):70-86. [doi: 10.1177/1090198105285927]
27. Freire K, Pope R, Jeffrey K, Andrews K, Nott M, Bowman T. Engaging with children and adolescents: a systematic review of participatory methods and approaches in research informing the development of health resources and interventions. Adolescent Res Rev 2022 Sep;7(3):335-354. [doi: 10.1007/s40894-022-00181-w]
28. Gilbey D, Morgan H, Lin A, Perry Y. Effectiveness, acceptability, and feasibility of digital health interventions for LGBTIQ+ young people: systematic review. J Med Internet Res 2020 Dec 3;22(12):e20158. [doi: 10.2196/20158] [Medline: 33270039]
29. Ybarra ML, Prescott TL, Philips GL, Bull SS, Parsons JT, Mustanski B. Iteratively developing an mHealth HIV prevention program for sexual minority adolescent men. AIDS Behav 2016 Jun;20(6):1157-1172. [doi: 10.1007/s10461-015-1146-3] [Medline: 26238038]
30. Hingle M, Patrick H, Sacher PM, Sweet CC. The intersection of behavioral science and digital health: the case for academic-industry partnerships. Health Educ Behav 2019 Feb;46(1):5-9. [doi: 10.1177/1090198118788600] [Medline: 30041556]
31. Austin D, May J, Andrade J, Jones R. Delivering digital health: the barriers and facilitators to university-industry collaboration. Health Policy Technol 2021 Mar;10(1):104-110. [doi: 10.1016/j.hlpt.2020.10.003]
32. Planned Parenthood Chat. URL: https://roo.plannedparenthood.org/ [accessed 2026-01-30]
33. Jaccard J. The prevention of problem behaviors in adolescents and young adults: perspectives on theory and practice. J Soc Social Work Res 2016 Dec 1;7(4):585-613. [doi: 10.1086/689354]
34. 2024 PPFA annual report. : Planned Parenthood Federation of America; 2024 URL: https://www.plannedparenthood.org/uploads/filer_public/bf/a3/bfa35e39-14e5-43c3-8192-de42ed9309f1/2024-ppfa-annualreport-c3-digital.pdf [accessed 2026-01-30]
35. Proctor EK, Powell BJ, McMillen JC. Implementation strategies: recommendations for specifying and reporting. Implement Sci 2013 Dec 1;8(1):139. [doi: 10.1186/1748-5908-8-139] [Medline: 24289295]
36. Bauermeister JA, Pingel E, Zimmerman M, Couper M, Carballo-Diéguez A, Strecher VJ. Data quality in HIV/AIDS web-based surveys: handling invalid and suspicious data. Field Methods 2012;24(3):272-291. [doi: 10.1177/1525822X12443097]

XSL•FO
RenderX

37. Grey JA, Konstan J, Iantaffi A, Wilkerson JM, Galos D, Rosser BRS. An updated protocol to detect invalid entries in an online survey of men who have sex with men (MSM): how do valid and invalid submissions compare? AIDS Behav 2015 Oct;19(10):1928-1937. [doi: 10.1007/s10461-015-1033-y] [Medline: 25805443]
38. Teitcher JEF, Bockting WO, Bauermeister JA, Hoefer CJ, Miner MH, Klitzman RL. Detecting, preventing, and responding to “fraudsters” in internet research: ethics and tradeoffs. J Law Med Ethics 2015;43(1):116-133. [doi: 10.1111/jlme.12200] [Medline: 25846043]
39. Gordon J, Lorenzo J, Avila AA, et al. Using discord to engage LGBTQ+ youth in human-centered design of a digital mental health intervention: a methods case study (preprint). JMIR Formative Res. Preprint posted online on Jul 18, 2025. [doi: 10.2196/preprints.80852]
40. Zimmerman J, Forlizzi J, Evenson S. Research through design as a method for interaction design research in HCI. Proc SIGCHI Conf Hum Factors Comput Syst 2007 Apr 29:493-502. [doi: 10.1145/1240624.1240704]
41. Birt L, Scott S, Cavers D, Campbell C, Walter F. Member checking: a tool to enhance trustworthiness or merely a nod to validation? Qual Health Res 2016 Nov;26(13):1802-1811. [doi: 10.1177/1049732316654870] [Medline: 27340178]
42. Gale RC, Wu J, Erhardt T, et al. Comparison of rapid vs in-depth qualitative analytic methods from a process evaluation of academic detailing in the Veterans Health Administration. Implement Sci 2019 Feb 1;14(1):11. [doi: 10.1186/s13012-019-0853-y] [Medline: 30709368]
43. Vindrola-Padros C, Johnson GA. Rapid techniques in qualitative research: a critical review of the literature. Qual Health Res 2020 Aug;30(10):1596-1604. [doi: 10.1177/1049732320921835] [Medline: 32667277]
44. Bengtsson M. How to plan and perform a qualitative study using content analysis. NursingPlus Open 2016;2:8-14. [doi: 10.1016/j.npls.2016.01.001]
45. Robillard JM, Feng TL, Sporn AB, et al. Availability, readability, and content of privacy policies and terms of agreements of mental health apps. Internet Interv 2019 Sep;17:100243. [doi: 10.1016/j.invent.2019.100243] [Medline: 30949436]
46. Yonker LM, Zan S, Scirica CV, Jethwani K, Kinane TB. “Friending” teens: systematic review of social media in adolescent and young adult health care. J Med Internet Res 2015 Jan 5;17(1):e4. [doi: 10.2196/jmir.3692] [Medline: 25560751]
47. Widman L, Nesi J, Kamke K, Choukas-Bradley S, Stewart JL. Technology-based interventions to reduce sexually transmitted infections and unintended pregnancy among youth. J Adolesc Health 2018 Jun;62(6):651-660. [doi: 10.1016/j.jadohealth.2018.02.007] [Medline: 29784112]
48. Laranjo L, Dunn AG, Tong HL, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018 Sep 1;25(9):1248-1258. [doi: 10.1093/jamia/ocy072] [Medline: 30010941]
49. Henkel T, Linn AJ, van der Goot MJ. Understanding the intention to use mental health chatbots among LGBTQIA+ individuals: testing and extending the UTAUT. Chatbot research and design. In: Følstad A, Araujo T, Papadopoulos S, Law ELC, editors. Chatbot Research and Design: 6th International Workshop, CONVERSATIONS 2022, Amsterdam, The Netherlands, November 22–23, 2022 Revised Selected Papers; Lecture Notes in Computer Science: Springer; 2023:83-100. [doi: 10.1007/978-3-031-25581-6_6]
50. Das RK, Drolet BC. Assessment of artificial intelligence chatbot attitudes toward LGBTQ+ individuals. J Adolesc Health 2024 Jun;74(6):1264-1266. [doi: 10.1016/j.jadohealth.2024.02.030] [Medline: 38762249]
51. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study. Digit Health 2019;5:2055207619871808. [doi: 10.1177/2055207619871808] [Medline: 31467682]
52. Shneiderman B. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Trans Interact Intell Syst 2020;10(4):26. [doi: 10.1145/3419764]
53. Park JK, Singh VK, Wisniewski P. Current landscape and future directions for mental health conversational agents for youth: scoping review. JMIR Med Inform 2025 Feb 28;13:e62758. [doi: 10.2196/62758] [Medline: 40053735]
54. Liao QV, Sundar SS. Designing for responsible trust in AI systems: a communication perspective. Proc 2022 ACM Conf Fairness, Accountability, Transparency 2022 Jun 20:1257-1268. [doi: 10.1145/3531146.3533182]
55. Costanza-Chock S. Design Justice: Community-Led Practices to Build the Worlds We Need: The MIT Press; 2020.
56. Yang Q, Steinfeld A, Rosé C, Zimmerman J. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. Proc 2020 CHI Conf Hum Factors Comput Syst 2020 Apr 23:1-13. [doi: 10.1145/3313831.3376301]
57. Amershi S, Weld D, Vorvoreanu M, et al. Guidelines for human-AI interaction. Proc 2019 CHI Conf Hum Factors Comput Syst 2019 May 2:1-13. [doi: 10.1145/3290605.3300233]
58. Long D, Magerko B. What is AI literacy? Competencies and design considerations. Proc 2020 CHI Conf Hum Factors Comput Syst 2020 Apr 23:1-16. [doi: 10.1145/3313831.3376727]
59. Ybarra ML, Mitchell KJ. A national study of lesbian, gay, bisexual (LGB), and non-LGB youth sexual behavior online and in-person. Arch Sex Behav 2016 Aug;45(6):1357-1372. [doi: 10.1007/s10508-015-0491-7] [Medline: 25894645]
60. Baum F, Newman L, Biedrzycki K. Vicious cycles: digital technologies and determinants of health in Australia. Health Promot Int 2014 Jun;29(2):349-360. [doi: 10.1093/heapro/das062] [Medline: 23144236]
61. Abd-alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M. An overview of the features of chatbots in mental health: a scoping review. Int J Med Inform 2019 Dec;132:103978. [doi: 10.1016/j.ijmedinf.2019.103978]
62. Madaio MA, Stark L, Wortman Vaughan J, Wallach H. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. CHI ’20 2020 Apr 23:1-14. [doi: 10.1145/3313831.3376445]

XSL•FO
RenderX

63. Kowitt SD, Schmidt AM, Hannan A, Goldstein AO. Awareness and trust of the FDA and CDC: results from a national sample of US adults and adolescents. PLOS ONE 2017;12(5):e0177546. [doi: 10.1371/journal.pone.0177546] [Medline: 28520750]
64. Magee JC, Bigelow L, Dehaan S, Mustanski BS. Sexual health information seeking online: a mixed-methods study among lesbian, gay, bisexual, and transgender young people. Health Educ Behav 2012 Jun;39(3):276-289. [doi: 10.1177/1090198111401384] [Medline: 21490310]

Abbreviations

AI: artificial intelligence
G2G: Guy2Guy
LGBTQ+: lesbian, gay, bisexual, transgender, and queer
LLM: large language model
PPFA: Planned Parenthood Federation of America
REDCap: Research Electronic Data Capture
RQA: rapid qualitative analysis
SHER: Sharing Health Education Resources
STI: sexually transmitted infection
YAC: Youth Advisory Council

Edited by A Mavragani; submitted 06.Jun.2025; peer-reviewed by A Adepoju, CL Coleman, GH Koo, JK Park, N Sun, R Seretlo; revised version received 02.Dec.2025; accepted 03.Dec.2025; published 19.Feb.2026.

Please cite as:
Liem WW, Casline E, Lorenzo J, Gordon JD, Avila AA, Taylor A, Levitz N, O'Keefe MC, Shum K, Macapagal K
Enhancing LGBTQ+ Inclusivity in an AI-Powered Sexual Health Chatbot: User-Centered Design Approach Through a Nonprofit and Academic Partnership
J Med Internet Res 2026;28:e78621
URL: https://www.jmir.org/2026/1/e78621
doi:10.2196/78621

© William Wibowo Liem, Elizabeth Casline, Julianna Lorenzo, Jacob D Gordon, Andrés Alvarado Avila, Attia Taylor, Nicole Levitz, Michael C O'Keefe, Karen Shum, Kathryn Macapagal. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Impact of a Health Coaching App on the Subjective Well-Being of Individuals With Multimorbidity: Mixed Methods Study

Isabelle Symes[1,2], MSc; Alexandra Burton[3], PhD; Daniela Mercado[4], PhD; Feifei Bu[1], PhD

[1]Department of Behavioural Science and Health, Institute of Epidemiology and Health Care, University College London, 1-19 Torrington Place, London, United Kingdom

[2]Centre for Preventive Neurology, Wolfson Institute of Population Health, Queen Mary University of London, London, United Kingdom

[3]Centre for Psychiatry and Mental Health, Wolfson Institute of Population Health, Queen Mary University of London, London, United Kingdom

[4]Holly Health, London, United Kingdom

Corresponding Author:
Feifei Bu, PhD
Department of Behavioural Science and Health, Institute of Epidemiology and Health Care, University College London, 1-19 Torrington Place, London, United Kingdom

Abstract

Background: Multimorbidity, the coexistence of 2 or more chronic conditions, is associated with poor well-being. Health coaching apps offer cost-effective and accessible support. However, there is a lack of evidence of the impact of health coaching apps on individuals with multimorbidity.

Objective: This study aimed to assess the impact and acceptability of a health coaching app (the Holly Health [HH] app) on the subjective well-being (SWB) of adults with multimorbidity.

Methods: This study used an explanatory-sequential mixed methods design, with quantitative secondary data analysis in the first phase and qualitative interviews in the second phase. In the quantitative phase (n=565), pre- and post-SWB (Office for National Statistics' 4 personal well-being questions [ONS4]) scores from existing app users with multimorbidity were analyzed using Bayesian growth curve modeling to assess the impact of HH. In the qualitative phase (n=22), data were collected via semistructured interviews and analyzed using reflexive thematic analysis. Mechanisms of action that supported SWB were categorized using the Multi-Level Leisure Mechanisms Framework.

Results: There was a significant increase in life satisfaction (Coef.=0.71, 95% highest density interval [HDI] 0.52 - 0.89), worthwhileness (Coef.=0.62, 95% HDI 0.43 - 0.81), and happiness (Coef.=0.74, 95% HDI 0.54 - 0.92) and a decrease in anxiety (Coef.=−0.50, 95% HDI −0.74 to −0.25) before and after using the HH app. Overall, 8 acceptable app features activated 5 mechanisms of action, including behavioral, psychological, and social mechanisms. Three additional factors influenced the acceptability of the health coaching app: type of chronic condition, availability of time, and the use of other support tools.

Conclusions: The study demonstrates that health coaching apps could be effective and acceptable support tools for individuals with multimorbidity. This study contributes to understanding why health coaching apps support SWB and could be used to inform the development of future digital health interventions in multimorbidity.

(J Med Internet Res 2026;28:e78738) doi:10.2196/78738

KEYWORDS

multimorbidity; health coaching app; digital health intervention; subjective well-being; mixed methods; mechanisms of action

Introduction

By 2035, it is projected that two-thirds of adults in England aged 65 years and older will have multiple chronic conditions, with a 50% increase in people developing 4 or more chronic conditions [1]. Multimorbidity is the coexistence of 2 or more long-term physical or mental health conditions [2]. It has been associated with key disease clusters including hypertension, diabetes, asthma, cancer, depression, and anxiety [3]. Multimorbidity is more prevalent among older adults, females, and those living in lower-income areas [4,5]. It is also influenced by socioeconomic (eg, education), behavioral (eg, smoking), psychosocial (eg, loneliness), and biological factors [6].

Due to the complex nature of multimorbidity, consequences range in severity and scope. Associated outcomes include premature mortality [7], disability [8], elevated frailty risk [9], and hospitalization [10]. Multimorbidity also contributes to economic burden, including increased health care use and costs [11]. The complex health care needs associated with multimorbidity management cannot be met sufficiently by current health care systems with a single disease focus [12]. There is a general lack of holistic support options, especially

XSL•FO
RenderX

an inattention to well-being in primary care consultations [13]. Emphasis on patient-centered approaches, where individuals can establish and work toward their own health and well-being goals, is essential to the effective management of multimorbidity [14]. Patient-centered care is a key component of the National Institute for Health and Care Excellence (NICE) guidance on multimorbidity [15].

Subjective well-being (SWB) is defined as an individual's thinking and feeling toward life as desirable [16]. It encompasses 3 dimensions: evaluative, eudaemonic, and affective [17,18]. Evaluative well-being refers to the overall assessment of satisfaction with life, eudaemonic well-being links to accounts of the meaning of life, and affective well-being reflects feelings and moods experienced every day [19]. Previous research has consistently shown the association between multimorbidity and poor well-being [6]. This is concerning, given SWB could play an important role in health maintenance and management. For example, increased SWB has been found to predict greater engagement in health-promoting behaviors [20], which has significant implications for multimorbidity management [21].

In the last few decades, there has been an increasing recognition of engagement in leisure activities, such as physical activity, as a protective factor against multimorbidity [22]. The Multi-Level Leisure Mechanisms Framework [23] suggests that engaging in leisure activities elicits various health benefits via psychological (eg, improved well-being), biological (eg, increased physical fitness), social (eg, increased social support), and behavioral (eg, increased motivation) mechanisms. This is supported by empirical research showing that regular exercise (behavioral mechanism) and strong social networks (social mechanism) are associated with a lower risk of multimorbidity [24], while increased social support (social mechanism) and lifestyle changes (behavioral mechanism) are associated with slower illness progression [25] and improved life expectancy [26].

Digital solutions can be used to promote these positive behavioral changes, such as increased engagement in physical activities, which can then subsequently enhance subjective and physical well-being, with health coaching apps offering cost-effective and accessible support options [27]. Health coaching apps are a form of digital health intervention (DHI), which are tools that use information and communication technologies for the improvement of health management, monitoring, prevention, treatment, and lifestyle [28]. Various studies have demonstrated the feasibility, acceptability, and effectiveness of the use of health coaching apps to support the self-management of common chronic conditions such as depression [29], diabetes [30], and hypertension [31]. However, there is a paucity of evidence exploring the mechanisms of their impacts. Moreover, evidence of their impacts on individuals with multimorbidity remains preliminary and limited, particularly with interventions that explicitly target SWB as a primary outcome [32]. To our knowledge, there are currently no studies specifically focusing on the impact of digital interventions on the SWB of people with multimorbidity. Therefore, this study focused on well-being, given its established association with multimorbidity and as it remains underexplored compared to more commonly studied outcomes (eg, quality of life).

This study aimed to assess the impact of a health coaching app on the SWB of people with multimorbidity. More specifically, it addressed the following research questions: (1) Does using a health coaching app have a positive impact on the SWB of people with multimorbidity? (2) If a positive impact is observed, which app features contribute to improved SWB for people experiencing multimorbidity and how do they support SWB? (3) Which factors influence the perceived acceptability of the app?

Methods

Holly Health App

This study is partnered with Holly Health (HH), a health coaching app supported by iOS, Android, and web [33]. Key features include daily chatbot support, nudges and reminders, mood, stress, and energy tracking, habit tracking, health reviews, tailored educational content, and short-term challenges (see Multimedia Appendix 1 for app images). These challenges typically involve 1-week goals (eg, mindfulness or physical activity challenges) that are ranked from easy to hard. The app targets well-being, sleep, exercise, and eating to promote healthy aging and support chronic condition management. Engagement with the app is user-directed, allowing individuals to interact with different features according to their needs and preferences rather than through a prescribed pathway. HH is supported by the UK National Health Service (NHS) Innovation Accelerator, which aims to support high-impact innovations that address NHS priorities [34]. Currently, the app is partnered with more than 200 general practitioner (GP) practices in the United Kingdom. Users typically access the app through GP practices, where they are invited via an SMS text message to download and use the app free of charge. HH is grounded in cognitive behavioral therapy (CBT) [35], acceptance and commitment therapy [36], mindfulness, and the small habits approach, based on frameworks such as the Capability-Opportunity-Motivation Behavior (COM-B) model [37]. While HH incorporates features to support chronic condition management, it was not specifically developed for people with multimorbidity.

Study Design

This study used an explanatory sequential mixed methods design [38], comprising a quantitative phase followed by the qualitative phase (see Figure 1). First, the quantitative phase involved secondary data analysis of repeated survey data (n=565). This was followed by the qualitative phase, where semistructured interviews were conducted with 22 app users. The two phases were conducted and reported sequentially, with integration occurring during the interpretation of findings to explain and expand on the quantitative results. This study defines multimorbidity as 2 or more chronic conditions. Guidance for secondary data analysis (STROSA: Standardized Reporting of Secondary Data Analyses [39]) and qualitative research (COREQ: Consolidated Criteria for Reporting Qualitative Research [40]) were followed.

XSL•FO
RenderX

Figure 1. Study design flowchart.

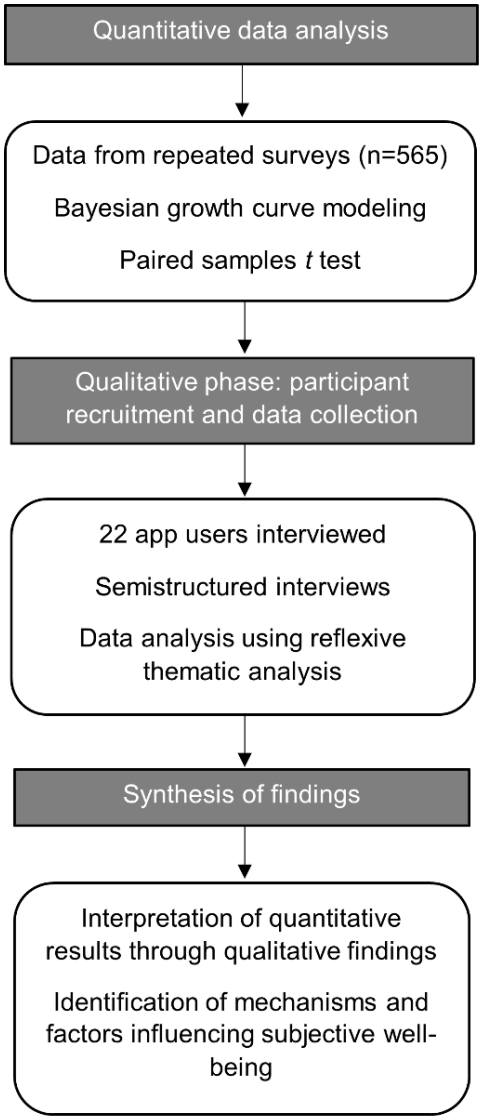

Ethical Considerations

The University College London (UCL) Research Ethics Committee approved this study (26737/001). All participants provided informed consent prior to participation, and all procedures were conducted in accordance with the ethical standards and institutional regulations. Secondary quantitative data were fully anonymized by HH before being transferred to the researchers for analysis. Data were stored on a secure research environment provided by UCL, accessible only to the research team. The interview recordings and transcripts from the qualitative phase were transferred to a secure research environment, ready for checking and deidentification. HH provided £5 (US $6.75) Amazon vouchers to those who completed an interview.

Participants

Quantitative Phase

Secondary data were obtained from health surveys of app users conducted by HH between March 2023 and March 2024. The surveys were originally collected outside the app but were later integrated within the app platform. Eligibility criteria for this study required participants to (1) report 2 or more chronic conditions during the onboarding health survey, (2) be aged 18 years or more, and (3) complete the baseline and follow-up (≥8 weeks) SWB questionnaires (n=565). Completion of the baseline questionnaire was compulsory; however, demographic information questions (eg, age, gender, and ethnicity) and the follow-up questionnaire were optional. Before applying the eligibility criteria, 882 app users had completed the baseline and follow-up questionnaire. Of this, 565 app users were eligible (64.1% of the available data). The sample size meets the minimum sample size to yield an alpha of .05, a statistical power of 0.8, and a small effect size (d=0.2) [41] in SWB outcomes. The assumption of a small effect size was based on previous literature that investigated the influence of a DHI on the well-being of individuals with a single chronic condition [42]. Demographics such as age, gender, and ethnicity were captured through the app, which were only available for 55.2% (312/565) of participants. Missingness occurred because the SWB survey was not initially integrated into the app; therefore, it cannot be linked to demographic information.

XSL•FO
RenderX

Qualitative Phase

A criterion-based purposive sampling technique was used to identify eligible app users with multimorbidity who had used the app for at least 2 weeks. Recruitment was conducted via an email campaign distributed by the HH team to eligible app users, and participants self-selected to take part by scheduling an interview through a Calendly link. The first email was sent to 2190 app users. A total of 30 app users replied, and interviews were booked; however, 9 did not go ahead due to participants' scheduling conflicts and technical difficulties. The email was then sent to 40 new app users who met the criteria, resulting in 1 more interview and 22 interviews in total. Information power was referenced during the planning and data collection phases to evaluate the sufficiency of the sample size [43]. This process was iterative, with no predefined number of interviews. Given the broad aim of the study and the use of cross-case analysis alongside the application of an established theory and a specific sample, a moderate sample size was deemed appropriate.

Materials and Procedure

Quantitative Phase

The primary outcome was measured by the Office for National Statistics' 4 personal well-being questions (ONS4) [44]. The 4 questions cover life satisfaction (evaluative), worthwhileness (eudemonic), happiness (affective), and anxiety (affective), each measured on an 11-point Likert scale (0 - 10) (see Multimedia Appendix 2 for questions and scoring). The ONS4 data were collected at baseline and the 8-week follow-up with no time limit for completion. The median number of days between baseline and follow-up completion was 84.5 (IQR 59-160) (~12 weeks). Other relevant data included age (18 - 34, 35 - 54, 55 - 64, and 65+ years), gender (male or female), ethnicity (White and other), number of daily habits completed (eg, 10 minutes of mindfulness=1 habit), number of medical conditions, and whether the app user felt they had developed automatic habits without relying on the app (yes or no). Chronic conditions were recorded through self-report using a predefined checklist. If a condition was not represented, app users had the option to select "something else." The checklist was developed based on publicly available NHS data on condition prevalence and the most commonly diagnosed chronic conditions identified among the first 20,000 HH users. They were not able to provide any additional free-text responses.

Qualitative Phase

Potential participants were sent an information sheet and a Calendly link to an interview booking page by HH via email. Participants could book 45-minute time slots for the interview. To book the time slot, participants were required to read the information sheet and check all the consent boxes, providing consent to take part in the study. The interviews were recorded and transcribed via Microsoft Teams. Due to confidentiality, we cannot confirm whether the interview participants were also part of the survey dataset. All interviews were conducted by the lead researcher (IS). The interviews were semistructured, following a topic guide (see Multimedia Appendix 3) discussing participants' multimorbidity and positive and negative experiences of using HH, with the flexibility of follow-up questions and prompts. The topic guide was developed by the lead researcher (IS), with supervision from 2 coauthors (FB and AB). It was informed by the study aims and the conceptual underpinnings of HH. The topic guide was piloted, with one question subsequently divided into two to improve clarity. This interview was included in the analysis. Demographic information, including self-reported chronic conditions, was collected during the interviews. The mean duration of interviews was 24.68 minutes (range 12.31-52.34 minutes). A distress protocol was in place due to potentially sensitive topics being disclosed, informed by the Qualitative Research Distress Protocol tool [45].

A reflective log was used by the lead researcher (IS) to guide self-reflection and transparency during the research process [46]. IS is a White female researcher (MSc) with experience in conducting research with people with long-term conditions and semistructured interviewing. Independence from HH was clarified to participants before the interview to ensure transparency and reduce potential bias. IS's values of respect, transparency, and evidence-based practice fostered rapport and shaped expectations of the app. IS's positionality, shared with many participants who were predominantly White and female, required reflexive attention to how shared identity could influence assumptions. Reflexive journaling and supervision supported critical reflection on how personal characteristics, interviewing style, and emotional responses influenced both data collection and analysis.

Analyses

Quantitative Phase

Descriptive analyses were conducted to summarize characteristics, including frequencies, percentages, means, and SDs. To examine changes in SWB before and after the intervention, quantitative data were analyzed using Bayesian growth curve modeling. It allowed us to examine person-specific and average changes between 2 time points, with random intercept and slope. For the main analysis, we fitted an unconditional linear growth model to the full sample (n=565) separately for each of the ONS4 SWB measures, using noninformative priors, 2000 iterations, 6 chains, a burn-in of 1000, and a thinning of 5. Please see Multimedia Appendix 4 for technical details. A paired samples *t* test was conducted as sensitivity analysis, using both the full sample and a conditional sample restricted to complete demographic data (n=312) to assess the potential influence of missing data. Sensitivity analyses were also conducted to examine if changes in SWB differed by individual characteristics including age, gender, ethnicity, number of medical conditions, automatic habits, and number of habit completions. This was done by fitting separate conditional growth models for each demographic covariate using data from a reduced sample with valid data for that specific covariate (n=292 to 312). The Bayesian growth curve models were fitted in R 4.3.2 (R Foundation for Statistical Computing) and using Markov Chain Monte Carlo algorithms, implemented in JAGS. Model convergence was assessed using the Gelman-Rubin statistic and visual inspection of the posterior distributions. Paired sample *t* tests were conducted in IBM SPSS Statistics 28.0 for Microsoft Windows.

XSL•FO
RenderX

Qualitative Phase

Transcripts were analyzed using NVivo 12 (Lumivero). Reflexive thematic analysis was used, a theoretically flexible methodology used to understand experiences and behaviors [47,48]. The analysis adopted an essentialist, experiential, inductive approach to capture and reflect participants' direct experiences and perspectives [49]. Once the themes were fully developed inductively, they were mapped onto the Multi-Level Leisure Mechanisms Framework [23] top-level categories and were interpreted in the context of participants' reported experiences with HH. This framework was selected as it offers a clear structure for understanding (eg, psychological, social, and behavioral mechanisms of change), which aligns with the study's aim of identifying which app features may support SWB and the factors linked to acceptability. The framework did not shape the coding or theme development process; rather, it was used post hoc to organize the themes within established mechanisms. This supported subsequent interpretation of findings, enabling connections to be drawn between themes and app features.

The lead researcher (IS) coded the transcripts twice, followed by discussion with two coauthors (FB and AB), who each read 2 transcripts. The reflective log was referred to throughout the analysis, with potential preconceptions about the familiarity of using technology considered.

Synthesis of Findings

This study followed an explanatory sequential mixed methods design, in which synthesis of findings occurs during the interpretive stage. In the quantitative phase, the analysis identified both the direction and significance of change in SWB. The qualitative phase was undertaken to provide explanatory depth and contextualization, focusing on potential mechanisms and specific app features that may be associated with the observed findings. The interpretation of the quantitative findings through consideration of the qualitative evidence is presented in the Discussion section.

Results

Quantitative Results

Sample Characteristics

Among app users with valid demographic information (n=312), 35.9% (n=112) were aged 65+ years, forming the largest age category (see Table 1). Over three-quarters (n=245, 78.5%) were female, and 94% (n=280) were White.

Table . Quantitative characteristics.

Variables	Values
Sex (n=312), n (%)	
Female	245 (78.5)
Male	67 (21.5)
Age in years (n=312), n (%)	
18 - 34	14 (4.5)
35 - 54	84 (26.9)
55 - 64	102 (32.7)
65+	112 (35.9)
Ethnicity (n=298), n (%)	
White	280 (94)
Other (Asian, Black, Mixed, or Other)	18 (6)
Automatic habits[a] (n=292), n (%)	
Yes	219 (75)
No	73 (25)
Number of medical conditions (n=312), mean (SD)	4.64 (3.36)
Habit completions[b] (n=312), mean (SD)	161.86 (229.51)

[a]Have you started to do any of your habits automatically? That is, without relying on the app to remind you?

[b]Number of habits recorded as completed.

The mean number of chronic conditions recorded was 4.64 (SD 3.36; Table 1). As shown in Table 2, the most common conditions were anxiety (126/312, 40.4%), hypertension (122/312, 39.1%), depression (108/312, 34.6%), and arthritis (96/312, 30.8%).

XSL•FO
RenderX

Table . Chronic condition frequency (n=312).

Chronic condition	Values, n (%)[a]
Mental health	
Anxiety	126 (40.4)
Depression	108 (34.6)
Cardiometabolic	
Hypertension	122 (39.1)
High cholesterol	85 (27.2)
Type 2 diabetes	43 (13.8)
Prediabetes	38 (12.2)
Heart disease	25 (8.0)
Musculoskeletal or pain	
Arthritis	96 (30.8)
Joint mobility issues	83 (26.6)
Chronic pain	41 (13.1)
Fibromyalgia	16 (7.1)
Osteoporosis	20 (6.4)
Severe mobility impairments	13 (4.2)
Respiratory	
Asthma	58 (18.6)
Chronic obstructive pulmonary disease	18 (5.8)
Gastrointestinal	
Irritable bowel syndrome	28 (17.2)
Irritable bowel syndrome or inflammatory bowel disease	19 (7.2)
Other	
Something else	104 (33.3)
Insomnia	52 (19.6)
Fatty liver	14 (9.2)
Stroke	15 (4.8)
Long COVID	12 (3.8)
Kidney disease	9 (2.9)

[a]Valid percentages calculated based on the number of app users who responded to each condition item. App users could report multiple chronic conditions.

Bayesian Growth Curve Model

The results from the unconditional Bayesian growth curve models are shown in Figure 2. On average, life satisfaction increased by 0.71 (95% highest density interval [HDI] 0.52-0.89) after using the app for 8 or more weeks, worthwhileness increased by 0.62 (95% HDI 0.43-0.81), and happiness increased by 0.74 (95% HDI 0.54-0.92). Alongside this, anxiety decreased by 0.50 (95% HDI −0.74 to −0.25).

XSL•FO
RenderX

Figure 2. Bayesian growth curve model mean trajectories.

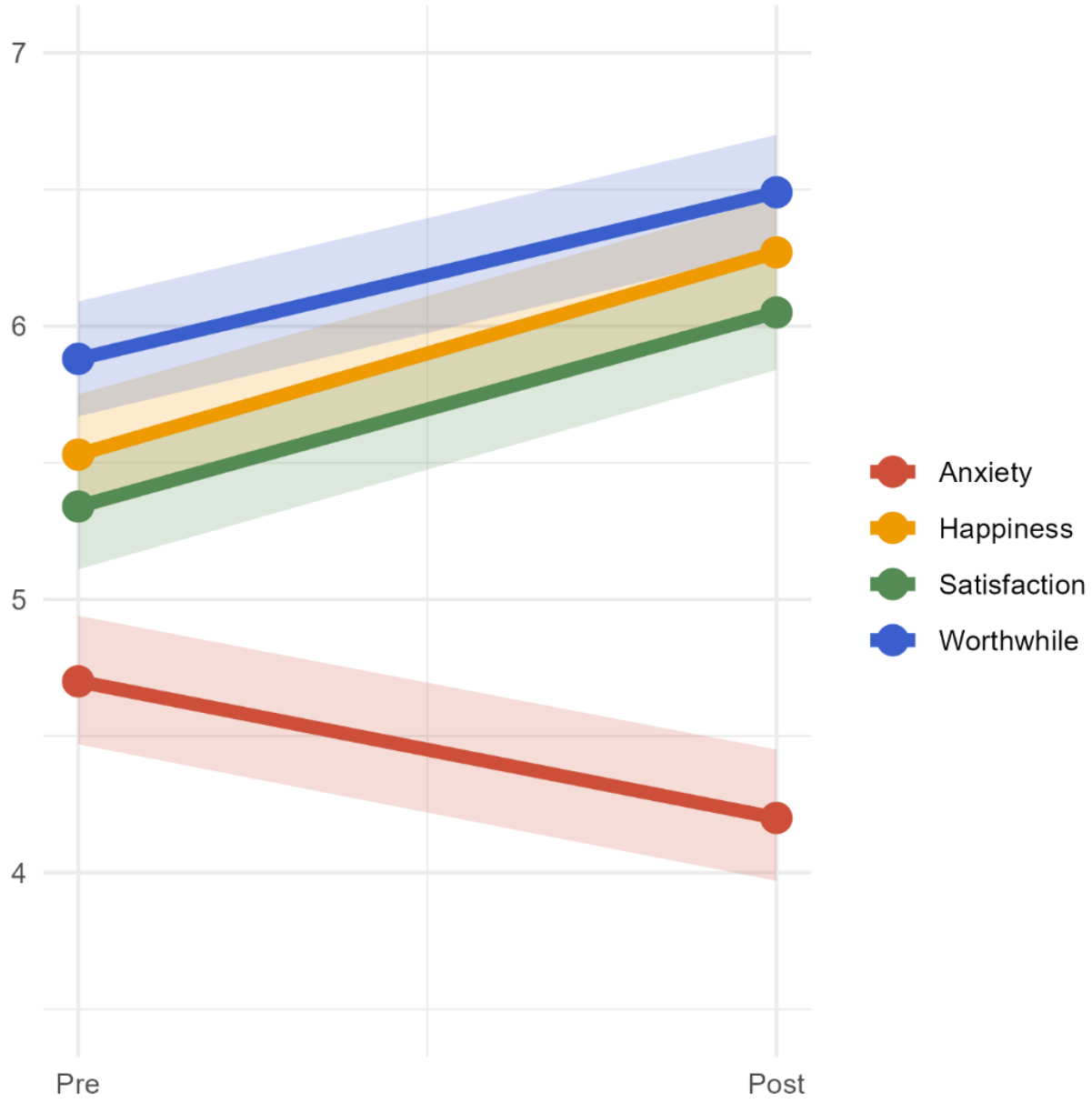

Sensitivity Analyses

As shown in Table 3, paired samples *t* tests revealed a significant difference in life satisfaction (t_{564}=−7.65, *P*<.001), worthwhileness (t_{564}=−6.58, *P*<.001), happiness (t_{564}=−7.46, *P*<.001), and anxiety scores (t_{564}=3.87, *P*<.001). This finding aligns with the Bayesian growth curve model, again suggesting a significant improvement in all ONS4 domains.

XSL•FO
RenderX

Table . Paired samples *t* test results (n=565).

ONS4[a] domains	Mean (SD)	*t* test[b]	*P* value	Cohen *d*[c]
Life satisfaction[d]		–7.65	<.001	–0.32
Prescore	5.34 (2.60)			
Postscore	6.06 (2.42)			
Worthwhileness[d]		–6.58	<.001	–0.28
Prescore	5.88 (2.55)			
Postscore	6.50 (2.46)			
Happiness[d]		–7.46	<.001	–0.31
Prescore	5.54 (2.64)			
Postscore	6.28 (2.53)			
Anxiety[e]		3.87	<.001	0.16
Prescore	4.70 (2.89)			
Postscore	4.20 (2.93)			

[a]ONS4: Office for National Statistics' 4 personal well-being questions.

[b]Two-tailed *t* test (*df*=564).

[c]Cohen *d* (1988) effect sizes: small (*d*=0.2), medium (*d*=0.5), and large (*d*=0.8).

[d]Life satisfaction, worthwhileness, and happiness thresholds: low (0-4), medium (5-6), high (7-8), and very high (9-10).

[e]Anxiety threshold: very low (0-1), low (2-3), medium (4-5), and high (6-10).

In the conditional Bayesian growth curve models (n=292 to 312), group effects were tested (see Multimedia Appendix 5). We found some evidence that age, automatic habits, and number of habit completions were associated with SWB outcomes at baseline. However, there was little evidence that these variables were related to the rate of change for any of the SWB measures.

Qualitative Results

Sample Characteristics

Participants in the qualitative sample (n=22) were aged 29 - 73 years (mean age 55.6 years). Most were female (n=18, 81.8%) and White (n=19, 86.4%). The most common chronic conditions reported were high blood pressure (n=7), anxiety (n=6), and type 2 diabetes (n=6) (see Table 4). Participants described varying levels of engagement with HH, from ongoing active use to reduced or discontinued use.

XSL•FO
RenderX

Table . Qualitative sample characteristics (n=22).

Characteristics	Values
Sex, n (%)	
Female	18 (81.8)
Male	4 (18.2)
Age (years), n (%)	
18 - 34	1 (4.6)
35 - 54	6 (27.3)
55 - 64	10 (45.5)
65+	5 (22.7)
Ethnicity, n (%)	
White	19 (86.4)
Other (Asian, Black, or prefer not to say)	3 (13.6)
Chronic conditions, n	
High blood pressure	7
Anxiety	6
Type 2 diabetes	6
Depression	4
High cholesterol	4
Fibromyalgia	3
Asthma	2
Chronic fatigue syndrome	3
Diverticulitis	2
Irritable bowel syndrome	2
Osteoarthritis	2
Prediabetes	2
Sciatica	2
Stroke	2
Other[a]	39

[a]Atrial fibrillation, attention-deficit/hyperactivity disorder, borderline osteoporosis, bowel cancer, bowel problems, chronic pain, cochlear hydrops, complex posttraumatic stress disorder, congenital heart disease, coronary heart disease, diabetic neuropathy, Graves' disease, gout, hard of hearing, heart disease, heart valve problems, hepatitis, hypermobility spectrum disorder, hypermobility syndrome, hypertension, inflammatory bowel disease, long COVID, low iron, mobility problems, multiple myeloma, musculoskeletal problems, perimenopausal disorder, peripheral neuropathy, rheumatoid arthritis, sleep apnea, tinnitus, thrombocytopenia, and unstable bladder.

Mechanisms of Action

Summary

Participants described 8 app features that activated 5 mechanisms of action leading to perceived improvements in SWB. App features included (1) those promoting action, such as "checking off daily goals," "setting small goals," "reminders and nudges," and "tracking and progress reports"; (2) features fostering reflective practice, including "mindfulness and affirmations" and "positive reinforcement"; and (3) interactive features such as "AI interface" (AI: artificial intelligence) and "constant support availability." Behavioral mechanisms included "maintenance of healthy behaviors" and "increased sense of achievement." The psychological mechanism was "improved self-compassion," and the social mechanisms included "increased feelings of social support" and "reduced feelings of loneliness." The acceptability of the app was impacted by "type of chronic condition," "availability of time," and "the use of other support tools." These relationships are presented in Figure 3. Participant quotes are used to illustrate the relationship between the app features and the activated mechanisms of action.

XSL•FO
RenderX

Figure 3. Thematic associations diagram. This figure illustrates the relationship between app features and the mechanisms of action identified in the study. The figure reflects both overarching categories from the Multi-Level Leisure Mechanisms Framework and participants' reported experiences. Dark pink denotes overarching mechanisms, while light pink indicates specific submechanisms. Similarly, dark blue denotes specific features within these groups. The black dotted arrows indicate the key associations between app features and mechanisms. Factors influencing app acceptability are shown in green, surrounding both features and mechanisms. AI: artificial intelligence.

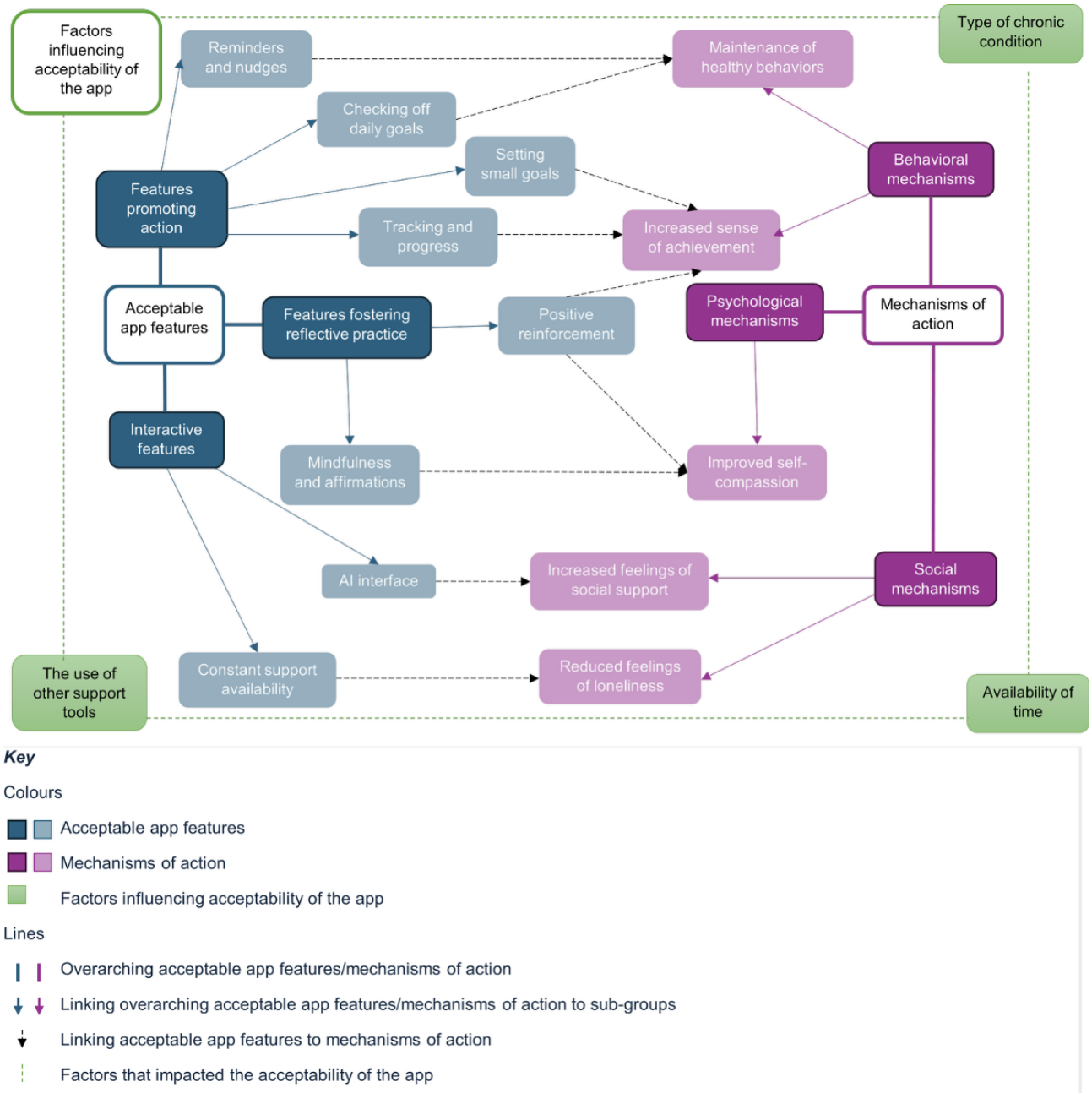

Behavioral Mechanisms

Maintenance of Healthy Behaviors

To maximize the benefits of engaging in healthy habits, including improved SWB, participants needed to reach a point of maintaining them. This process was supported by the app features "reminders and nudges" and "checking off daily goals." Participants felt that they had achieved maintenance when they no longer relied on the app to fulfill their daily goals:

> *The meditation, I just go and do that myself. Now I don't even [use the app] because I'm in the habit of doing it.* [Participant 13]

The app worked in the background to support participants to stay "on track" while maintaining their habits. Some participants reached a point where they felt confident enough to start introducing new habits due to the strong maintenance process:

> *I guess I really should be changing them, the ones that are now coming naturally to me.* [Participant 11]

Participants acknowledged the challenge of maintenance, whether that be due to forgetfulness, needing further nudges, or the nature of the habit itself. The success of the maintenance process varied, with some habits proving more difficult to maintain than others:

XSL•FO
RenderX

> *I try and respond to it daily. I don't always remember, but I try to. I would say that I've been less successful with the exercise things that I've set.* [Participant 11]

Increased Sense of Achievement

Often, participants attributed their progress to an increased sense of achievement, facilitated by the features "positive reinforcement," "tracking and progress reports," and "setting small goals." This "sense of accomplishment" served as a powerful motivator, encouraging them to continue progressing with their goals:

> *I do like the little confetti you get when you've done your habit. You know, I know it's just psychological, but I quite like the confetti. Yeah, there is a certain satisfaction.* [Participant 3]

Moreover, highlighting progress milestones provided further gratification, promoting positive reflection and motivation:

> *It's more encouraging. And I also like when it comes up with a kind of big explosion, that's 100 hours now that you've done in the brain studio.* [Participant 12]

Psychological Mechanisms: Increased Self-Compassion

Participants emphasized the importance of self-compassion in fostering a more benevolent outlook. This mechanism was linked to the features of "positive reinforcement" and "mindfulness and affirmations." Many participants described feeling discouraged about incomplete goals and restrictions on what they could achieve due to multimorbidity. This led some participants to "talk poorly" to themselves, using HH to help challenge these negative beliefs and "rationalize." This was especially effective in prompting participants to acknowledge their progress and establish realistic expectations:

> *Since I've had Holly Health, I've been able to put a lot of things into perspective. Because it's saying it's okay to have down days, it's okay not to feel your best. But it's also inspiring because it gives you the thought that you can do it another day when you feel better.* [Participant 6]

By adopting this mindset, HH helped participants cultivate greater self-compassion by not framing experience in absolute terms:

> *I think the problem is with my tick charts or where I've got a notebook and I'm checking a box or something, I can get down on myself if I've bitten off more than I can chew. I'm a bit of a perfectionist I think - avoiding that black and white, you either did it or you didn't, you were great, or you were rubbish kind of thing.* [Participant 4]

Social Mechanisms

Increased Feelings of Social Support

Participants described the human-like qualities of HH, which contributed to improving their SWB. This was directly attributed to the "AI interface" feature. Such support was particularly valuable when considering multimorbidity, as interactions with health care professionals are often related to a deterioration in health. The chatbot enabled participants to seek out and experience more preventative and proactive holistic support:

> *It almost feels like somebody's looking out for you, but not in a purely medical context. Because you end most of what you do with the healthcare professional is usually when something goes wrong. So, it's not when something's gone wrong, it's actually slightly helping things not to go wrong.* [Participant 3]

The impact of the support was especially strong when it was tailored; the "more holistic touch" of HH made the support feel "personal." This introduced feelings of having friend-like, reliable support:

> *It helps with depression because it can talk to you through times of no stimulation and depressive thoughts. It's like a friend.* [Participant 6]

Reduced Feelings of Loneliness

Living with multimorbidity often led participants to experience feelings of loneliness, with peers sometimes struggling to understand the associated challenges. Feeling listened to and acknowledged was crucial to participants, with some encountering dismissive interactions elsewhere. The "constant support availability" and "AI interface" of HH helped participants to feel less lonely:

> *I have chats with them maybe once a week. It does help knowing that somebody is at the other side, even though it's just a computer, and that somebody's listening.* [Participant 16]

Participants expressed that enduring support options were crucial for their well-being, particularly when they found it challenging to actively seek support:

> *I do feel Holly (Health) is what's carried me through that really tough spell. Where I genuinely didn't really have any support from anyone, and it is much because I couldn't ask for it.* [Participant 10]

Factors Influencing Acceptability of the App

Three factors that affected the acceptability of the app were identified: "type of chronic condition," "availability of time," and "using other support tools alongside HH."

Type of Chronic Condition

Many participants viewed HH as a "holistic" support option, capable of addressing a wide range of chronic conditions simultaneously:

> *It's quite universal though, I think there's so much in there that there'll be something that would work for anybody's condition.* [Participant 11]

However, others noted HH seemed more beneficial for certain conditions, with more advice and support available for general chronic conditions:

> *It seems to be designed for quite common conditions, anxiety, depression. I think there are more specific problems I had around rumination […] I don't think it was specifically good for that, I think it's good for general wellbeing.* [Participant 5]

XSL•FO
RenderX

Furthermore, participants with chronic fatigue syndrome (CFS) noted that the app provided limited information and advice on managing their condition, feeling it "needs something extra." Some felt their conditions were too complex or unique to be fully addressed by HH:

> *I don't really feel that I fit into an easy category. Particularly people with fibromyalgia. Your experience is very individual.* [Participant 21]

Availability of Time

Having time to fully engage in the app and complete the goals was a key factor in how suitable the app was for improving SWB. Many participants valued having the designated time to focus on their well-being goals daily:

> *There's notifications to give you that reminder, because let's be honest, well life gets in the way of everything, but sometimes you have to take that five-minutes for yourself.* [Participant 17]

Another participant described that "it's time for me," with the app allowing participants to adapt goals to fit within schedules. This was especially apparent in participants who were retired or semiretired, having more time to engage with the app and explore the activities:

> *But that's one of the things about being semiretired and having more time is trying these different things.* [Participant 9]

The Use of Other Support Tools

Many participants used a combination of support to ensure comprehensive care for their mental and physical health, with HH being one component of a broader support regimen:

> *Another tool in your general armoury of things. You're trying to look after yourself as best you can.* [Participant 15]

Some participants mentioned using multiple apps together, selecting the most suitable based on their current needs:

> *I have quite a few different apps on my phone. I suppose I'm always seeking which one feels right in the moment but also will give me longer-term support.* [Participant 8]

Others combined HH with more intensive forms of support, such as health coaching, adapting techniques from both face-to-face and digital formats to enhance their support plan and maximize the benefits:

> *I'm having counselling and stuff anyway, so obviously some techniques from there which I already knew anyway, but it opened my eyes to wider things.* [Participant 20]

HH was especially valuable when other support tools were not readily available, serving as an interim solution, particularly when waiting for support:

> *I was on the wait list again for CBT, so I thought some of the same techniques might be accessible in the app while I'm waiting.* [Participant 5]

However, some participants found using multiple support tools simultaneously challenging. This limited the extent of the progress made, with some participants perceiving the required commitment as excessive, especially when using multiple health apps together:

> *It just felt (it was) difficult to keep doing it because it's a big commitment to do any one of those apps for a long period.* [Participant 5]

Discussion

Principal Results

This study aimed to examine the impact of a health coaching app on the SWB of individuals with multimorbidity and understand how and why the app supported SWB. Our quantitative analyses revealed that using the app for 8 weeks or longer was associated with increased SWB, including life satisfaction, worthwhileness, happiness, and a reduction in anxiety. There was little evidence that these changes differed by individual characteristics.

Through qualitative analyses, 8 app features were identified as acceptable, categorized by features promoting action, reflective practice, and interaction. These features facilitated a positive impact on SWB through activating behavioral, psychological, and social mechanisms of action. Wider contexts that impact the app acceptability were also identified, including chronic condition type, having time to engage, and the availability of other support tools alongside HH. These findings suggest that a health coaching app can serve as an acceptable support tool within these contexts.

Comparison With Prior Work

These findings provide evidence for the SWB benefits of a health coaching app for individuals with multimorbidity after 8 weeks or longer. This aligns with previous research demonstrating the effectiveness of digital solutions in enhancing well-being and self-management of single chronic conditions [42,50] alongside previous internal reports from HH, indicating that app usage boosts confidence in managing multimorbidity [51]. Additionally, this is consistent with prior evidence indicating that behavioral interventions can produce measurable well-being improvements within as little as 4 weeks in a range of populations, including people with chronic conditions [52]. Although the estimated changes in SWB scores were modest (<1 point on the Likert scales), such differences are often regarded as meaningful at the population level. For example, the UK HM Treasury Green Book Wellbeing Guidance [53] equates a 1-point increase in life satisfaction with a monetary value of approximately £13,000 (US $17,556.65) per person per year. Furthermore, the observed effect sizes in this study align with those reported in comparable digital behavior change and well-being interventions [54]. While research on digital support tools for multimorbidity is sparse, these findings extend insights from single-condition literature [55], suggesting that this health coaching app may similarly promote self-management and SWB among people with multimorbidity.

The mapping of perceived mechanisms of action to the Multi-Level Leisure Mechanisms Framework [23] enabled the

XSL•FO
RenderX

connection between app use and improved SWB to be explored in-depth. These mechanisms often intersect, with behavioral, psychological, and social mechanisms reinforcing one another to support SWB outcomes. Identified behavioral mechanisms of action included maintenance of healthy behaviors and increased sense of achievement. These mechanisms were observed to be supported by reminders and nudges, checking off daily goals, setting small goals, and tracking progress reports. The process involved in checking off daily goals was highlighted as particularly impactful in promoting action, reflected in app users reaching, maintaining, and setting new goals. Positive reinforcement also played a key role in the sense of achievement, encouraging app users to accomplish new goals. This finding is consistent with previous research that suggests goal setting, action planning, and reinforcement are crucial for sustained user engagement and promoting behavior change [56]. Although evidence of the effectiveness of reminders and nudges has previously been inconsistent, the context-specific design of reminders (eg, those tailored to specific habits) was found to be effective in this study. This finding aligns with previous research showing that context-specific reminders are crucial in enabling habit formation [57,58].

Positive reinforcement was also reported to be linked to the psychological mechanism of improved self-compassion. Goal setting similarly facilitated improved self-compassion; app users were able to establish realistic expectations and recognize their progress. This mechanism of action was additionally facilitated by the mindfulness and affirmations content within the DHI, fostering reflective practice. Social mechanisms of action were identified as increased feelings of social support, observed to be activated by the AI interface feature, and reduced feelings of loneliness, reported to be activated by the constant support availability within the DHI. The AI interface was highly valued by app users, and the live chat supported app users through a dependable, personalized approach. These findings are consistent with previous research, with several features identified as valuable in DHIs, including relaxation, personalization, and live support [59]. Distinctly, previous research has recognized increased social support as a key mechanism of improved self-management in multimorbidity interventions [60]. This emphasizes the value of providing features that facilitate this mechanism of action, especially in multimorbidity interventions.

The effect of the DHI can also be strengthened by using evidence-based approaches, such as CBT [35] and the COM-B model [37]. Incorporating these approaches has been found to improve the effectiveness, engagement, perceived quality, and credibility of apps [61-63]. Findings in this study suggest that using app-based support alongside other support tools (eg, health coaching) enhances effectiveness and acceptability. This aligns with previous literature [64], emphasizing the importance of complementary principles to ensure consistency of support.

Novel findings from the qualitative interviews suggest that the complexity of multimorbidity plays a crucial role in the impact and acceptability of the health coaching app as a support tool. While the health coaching app demonstrated a positive impact on SWB, indicating its potential to enhance SWB in people with multimorbidity to some extent, qualitative findings identified that chronic condition type influenced the acceptability of the health coaching app, with lack of information and education on specific conditions consistently a key barrier to support [65]. CFS and fatigue-based symptoms were particularly less accommodated by the app content, which primarily targeted more general conditions such as anxiety. Previous research has emphasized the significance of fear-avoidance beliefs and determining personal thresholds in the efficacy of supporting CFS [66,67]. Therefore, acknowledging these factors is crucial to the acceptability of app-based support for individuals with a CFS comorbidity. Conversely, conditions such as depression and anxiety, which have the highest health care service use in individuals with multimorbidity [24], appear to benefit from the health coaching app as an effective support tool for improving SWB.

The value of DHIs in older adults has been recognized in previous literature, particularly given the high prevalence of long-term conditions in this demographic group [68]. However, there have been concerns that the digital exclusivity of these interventions may lead to higher attrition rates, along with decreased engagement and effectiveness [68]. Notably, our quantitative findings did not identify age as a barrier to SWB improvement, indicating the potential effectiveness of the health coaching app across age groups, including older adults. In fact, our qualitative findings identified that retirement appeared to facilitate greater app engagement, providing individuals with more time to achieve daily goals. This observation aligns with previous research that associates retirement with increased leisure activities, including personal growth and physical exercise [69].

Strengths and Limitations

The main strength of this study lies in its explanatory sequential mixed methods design, which allowed for a comprehensive understanding of the insights gained from the quantitative phase through qualitative work. This allowed an in-depth exploration of a complex, under-researched area. However, several limitations must be acknowledged. First, the sample in both phases predominantly consisted of White females aged 55 years and older. While this partly reflects the prevalence of multimorbidity in older adults and females [4,5], the lack of diversity in the sample raises concerns about the generalizability of the results to the wider population [70]. Second, confidentiality measures prevented verifying if the same participants were involved in both data collection phases, potentially introducing participant variability and affecting the consistency of results. Third, as participation in the qualitative phase was based on self-selection, the sample may have predominantly consisted of motivated or engaged app users. This self-selection could introduce volunteer bias and limit the generalizability of the findings. However, participants did report varying levels of engagement with the app and spoke about various challenges of use. Similarly, in the quantitative phase, completion of the follow-up questionnaire was optional, which may have resulted in overrepresentation of engaged app users. Additionally, the absence of a control group makes it challenging to establish a causal effect as there is no comparison group against the observed effects, ultimately reducing internal

XSL•FO
RenderX

validity. Due to these limitations, caution is advised when interpreting the study findings.

Conclusions

Our study provides empirical evidence that a health coaching app can be an effective and acceptable support tool to improve the SWB of individuals with multimorbidity. These effects were driven by specific app features promoting action, reflective practice, and interaction. These features led to improved SWB through the activation of reported behavioral, psychological, and social mechanisms. However, the magnitude of these effects could be affected by contextual factors, including users' time availability for engagement, specific chronic condition profiles, and concurrent use of other support tools. By elucidating the mechanisms and contextual nuances underlying app efficacy, this study provides critical insights to inform the refinement of existing interventions and the design of future DHIs tailored to the complex needs of individuals with multimorbidity.

Acknowledgments

The authors would like to thank Holly Health, particularly Liliana Chow, for their involvement and work toward the partnered research project, and for providing access to their data. The authors also thank the participants for their time and contribution to the study.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

Data will be shared upon reasonable request from the corresponding author.

Authors' Contributions

Conceptualization: FB, AB, IS
Data curation: IS, FB
Formal analysis: IS, FB
Methodology: IS, FB, AB
Supervision: FB, AB
Writing – original draft: IS, FB
Writing – review and editing: IS, FB, AB, DM

Conflicts of Interest

DM is employed by Holly Health. She was not involved in conducting the interviews or the data analysis. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Multimedia Appendix 1
Images of the Holly Health app.
[DOCX File, 941 KB - jmir_v28i1e78738_app1.docx]

Multimedia Appendix 2
ONS4 questions and thresholds. ONS4: Office for National Statistics' 4 personal well-being questions.
[DOCX File, 19 KB - jmir_v28i1e78738_app2.docx]

Multimedia Appendix 3
Interview topic guide.
[DOCX File, 17 KB - jmir_v28i1e78738_app3.docx]

Multimedia Appendix 4
Bayesian unconditional growth curve model technical specifications.
[DOCX File, 15 KB - jmir_v28i1e78738_app4.docx]

Multimedia Appendix 5
Bayesian growth curve model: group effects results.
[DOCX File, 22 KB - jmir_v28i1e78738_app5.docx]

XSL•FO
RenderX

Checklist 1
STROSA and COREQ checklists.
[DOCX File, 31 KB - jmir_v28i1e78738_app6.docx]

References

1. Kingston A, Robinson L, Booth H, Knapp M, Jagger C, MODEM project. Projections of multi-morbidity in the older population in England to 2035: estimates from the Population Ageing and Care Simulation (PACSim) model. Age Ageing 2018 May 1;47(3):374-380. [doi: 10.1093/ageing/afx201] [Medline: 29370339]
2. Johnston MC, Crilly M, Black C, Prescott GJ, Mercer SW. Defining and measuring multimorbidity: a systematic review of systematic reviews. Eur J Public Health 2019 Feb 1;29(1):182-189. [doi: 10.1093/eurpub/cky098] [Medline: 29878097]
3. Zemedikun DT, Gray LJ, Khunti K, Davies MJ, Dhalwani NN. Patterns of multimorbidity in middle-aged and older adults: an analysis of the UK Biobank data. Mayo Clin Proc 2018 Jul;93(7):857-866. [doi: 10.1016/j.mayocp.2018.02.012]
4. Agborsangaya CB, Lau D, Lahtinen M, Cooke T, Johnson JA. Multimorbidity prevalence and patterns across socioeconomic determinants: a cross-sectional survey. BMC Public Health 2012 Mar 19;12(1):201. [doi: 10.1186/1471-2458-12-201] [Medline: 22429338]
5. Kone AP, Mondor L, Maxwell C, Kabir US, Rosella LC, Wodchis WP. Rising burden of multimorbidity and related socio-demographic factors: a repeated cross-sectional study of Ontarians. Can J Public Health 2021 Aug;112(4):737-747. [doi: 10.17269/s41997-021-00474-y] [Medline: 33847995]
6. Skou ST, Mair FS, Fortin M, et al. Multimorbidity. Nat Rev Dis Primers 2022 Jul 14;8(1):1-22. [doi: 10.1038/s41572-022-00376-4]
7. Johnston MC, Black C, Mercer SW, Prescott GJ, Crilly MA. Prevalence of secondary care multimorbidity in mid-life and its association with premature mortality in a large longitudinal cohort study. BMJ Open 2020 May;10(5):e033622. [doi: 10.1136/bmjopen-2019-033622]
8. Rizzuto D, Melis RJF, Angleman S, Qiu C, Marengoni A. Effect of chronic diseases and multimorbidity on survival and functioning in elderly adults. J Am Geriatr Soc 2017 May;65(5):1056-1060. [doi: 10.1111/jgs.14868] [Medline: 28306158]
9. Lujic S, Randall DA, Simpson JM, Falster MO, Jorm LR. Interaction effects of multimorbidity and frailty on adverse health outcomes in elderly hospitalised patients. Sci Rep 2022 Aug 19;12(1):14139. [doi: 10.1038/s41598-022-18346-x]
10. Buja A, Rivera M, De Battisti E, et al. Multimorbidity and hospital admissions in high-need, high-cost elderly patients. J Aging Health 2020 Jun;32(5-6):259-268. [doi: 10.1177/0898264318817091]
11. Soley-Bori M, Ashworth M, Bisquera A, et al. Impact of multimorbidity on healthcare costs and utilisation: a systematic review of the UK literature. Br J Gen Pract 2021 Jan;71(702):e39-e46. [doi: 10.3399/bjgp20X713897]
12. Moffat K, Mercer SW. Challenges of managing people with multimorbidity in today's healthcare systems. BMC Fam Pract 2015 Oct 14;16(1):129. [doi: 10.1186/s12875-015-0344-4] [Medline: 26462820]
13. Rimmelzwaan LM, Bogerd MJL, Schumacher BMA, Slottje P, Van Hout HPJ, Reinders ME. Multimorbidity in general practice: unmet care needs from a patient perspective. Front Med 2020 Dec 22;7. [doi: 10.3389/fmed.2020.530085]
14. Wilkinson I, Preston J. Managing patients with multimorbidity. Clinics in Integrated Care 2021 Apr;5:100045. [doi: 10.1016/j.intcar.2021.100045]
15. Multimorbidity: clinical assessment and management. : NICE; 2016 Sep. URL: https://www.nice.org.uk/guidance/ng56/chapter/Recommendations#general-principles [accessed 2025-12-23]
16. Diener E, editor. Assessing Well-Being: Springer Netherlands; 2009. [doi: 10.1007/978-90-481-2354-4]
17. Hicks S, Tinkler L, Allin P. Measuring subjective well-being and its potential role in policy: perspectives from the UK Office for National Statistics. Soc Indic Res 2013 Oct;114(1):73-86. [doi: 10.1007/s11205-013-0384-x]
18. Personal well-being user guidance. Office for National Statistics. 2025 Jan. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/methodologies/personalwellbeingsurveyuserguide [accessed 2025-12-23]
19. Steptoe A, Deaton A, Stone AA. Subjective wellbeing, health, and ageing. Lancet 2015 Feb 14;385(9968):640-648. [doi: 10.1016/S0140-6736(13)61489-0] [Medline: 25468152]
20. Kushlev K, Drummond DM, Diener E. Subjective well-being and health behaviors in 2.5 million Americans. Appl Psychol Health Well Being 2020 Mar;12(1):166-187. [doi: 10.1111/aphw.12178] [Medline: 31482675]
21. Multimorbidity | Health topics A to Z. CKS | NICE. URL: https://cks.nice.org.uk/topics/multimorbidity/ [accessed 2025-12-23]
22. Feter N, Leite JS, Umpierre D, Caputo EL, Rombaldi AJ. Multimorbidity and leisure-time physical activity over the life course: a population-based birth cohort study. BMC Public Health 2021 Apr 9;21(1):700. [doi: 10.1186/s12889-021-10719-7] [Medline: 33836716]
23. Fancourt D, Aughterson H, Finn S, Walker E, Steptoe A. How leisure activities affect health: a narrative review and multi-level theoretical framework of mechanisms of action. Lancet Psychiatry 2021 Apr;8(4):329-339. [doi: 10.1016/S2215-0366(20)30384-9]
24. Multiple long-term conditions (multimorbidity): making sense of the evidence. : NIHR Evidence; 2021 Mar. [doi: 10.3310/collection_45881]

XSL•FO
RenderX

25. Schäfer I, Hansen H, Kaduszkiewicz H, et al. Health behaviour, social support, socio-economic status and the 5-year progression of multimorbidity: results from the MultiCare Cohort Study. J Comorb 2019;9:2235042X19883560. [doi: 10.1177/2235042X19883560] [Medline: 35174099]
26. Chudasama YV, Khunti K, Gillies CL, et al. Healthy lifestyle and life expectancy in people with multimorbidity in the UK Biobank: a longitudinal cohort study. PLOS Med 2020;17(9):e1003332. [doi: 10.1371/journal.pmed.1003332]
27. Eisenstadt M, Liverpool S, Infanti E, Ciuvat RM, Carlsson C. Mobile apps that promote emotion regulation, positive mental health, and well-being in the general population: systematic review and meta-analysis. JMIR Ment Health 2021 Nov 8;8(11):e31170. [doi: 10.2196/31170] [Medline: 34747713]
28. Digital health and care. European Commission. 2025. URL: https://health.ec.europa.eu/ehealth-digital-health-and-care/digital-health-and-care_en [accessed 2025-12-23]
29. Deady M, Glozier N, Calvo R, et al. Preventing depression using a smartphone app: a randomized controlled trial. Psychol Med 2022 Feb;52(3):457-466. [doi: 10.1017/S0033291720002081] [Medline: 32624013]
30. Bailey DP, Mugridge LH, Dong F, Zhang X, Chater AM. Randomised controlled feasibility study of the MyHealthAvatar-Diabetes smartphone app for reducing prolonged sitting time in type 2 diabetes mellitus.. Int J Environ Res Public Health 2020 Jun 19;17(12):4414. [doi: 10.3390/ijerph17124414] [Medline: 32575482]
31. Alessa T, Abdi S, Hawley MS, de Witte L. Mobile apps to support the self-management of hypertension: systematic review of effectiveness, usability, and user satisfaction. JMIR Mhealth Uhealth 2018 Jul 23;6(7):e10723. [doi: 10.2196/10723] [Medline: 30037787]
32. Irfan Khan A, Gill A, Cott C, Hans PK, Steele Gray C. mHealth tools for the self-management of patients with multimorbidity in primary care settings: pilot study to explore user experience. JMIR Mhealth Uhealth 2018 Aug 28;6(8):e171. [doi: 10.2196/mhealth.8593] [Medline: 30021707]
33. Holly Health | Home. URL: https://www.hollyhealth.io/ [accessed 2025-12-23]
34. Holly Health – digital support for women's health and wellbeing. NHS Innovation Accelerator. URL: https://nhsaccelerator.com/innovations/holly-health/ [accessed 2025-12-23]
35. Beck AT. Cognitive Therapy and the Emotional Disorders: Penguin; 1979.
36. Hayes SC, Pierson H. Acceptance and commitment therapy. In: Freeman A, Felgoise SH, Nezu CM, Nezu AM, Reinecke MA, editors. Encyclopedia of Cognitive Behavior Therapy: Springer US; 2005:1-4. [doi: 10.1007/0-306-48581-8_1]
37. Michie S, Atkins L, West R. The Behaviour Change Wheel: A Guide to Designing Interventions: Silverback Publishing; 2024. URL: https://www.behaviourchangewheel.com/ [accessed 2025-12-23]
38. Ivankova NV, Creswell JW, Stick SL. Using mixed-methods sequential explanatory design: from theory to practice. Field methods 2006 Feb;18(1):3-20. [doi: 10.1177/1525822X05282260]
39. Swart E, Schmitt J. Standardized Reporting of Secondary Data Analyses (STROSA)—a recommendation. Z Evid Fortbild Qual Gesundhwes 2014;108(8-9):511-516. [doi: 10.1016/j.zefq.2014.08.022] [Medline: 25523850]
40. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
41. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
42. Parks AC, Williams AL, Kackloudis GM, Stafford JL, Boucher EM, Honomichl RD. The effects of a digital well-being intervention on patients with chronic conditions: observational study. J Med Internet Res 2020 Jan 10;22(1):e16211. [doi: 10.2196/16211] [Medline: 31922491]
43. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
44. Surveys using our four personal well-being questions. Office for National Statistics. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/methodologies/surveysusingthe4officefornationalstatisticspersonalwellbeingquestions [accessed 2025-12-23]
45. Whitney C, Evered JA. The qualitative research distress protocol: a participant-centered tool for navigating distress during data collection. Int J Qual Methods 2022 Apr;21. [doi: 10.1177/16094069221110317]
46. Darawsheh W. Reflexivity in research: promoting rigour, reliability and validity in qualitative research. Int J Ther Rehabil 2014 Dec 2;21(12):560-568. [doi: 10.12968/ijtr.2014.21.12.560]
47. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
48. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qualitative Research in Sport, Exercise and Health 2019 Aug 8;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
49. Byrne D. A worked example of Braun and Clarke's approach to reflexive thematic analysis. Qual Quant 2022 Jun;56(3):1391-1412. [doi: 10.1007/s11135-021-01182-y]
50. Bricca A, Pellegrini A, Zangger G, Ahler J, Jäger M, Skou ST. The quality of health apps and their potential to promote behavior change in patients with a chronic condition or multimorbidity: systematic search in App Store and Google Play. JMIR Mhealth Uhealth 2022 Feb 4;10(2):e33168. [doi: 10.2196/33168] [Medline: 35119367]

XSL•FO
RenderX

51. Long-term condition management across Holly Health users: a retrospective survey. Holly Health. 2023 Nov 30. URL: https://www.hollyhealth.io/articles/long-term-condition-management [accessed 2025-12-23]
52. Weiss LA, Westerhof GJ, Bohlmeijer ET. Can we increase psychological well-being? The effects of interventions on psychological well-being: a meta-analysis of randomized controlled trials. PLoS ONE 2016;11(6):e0158092. [doi: 10.1371/journal.pone.0158092] [Medline: 27328124]
53. Wellbeing guidance for appraisal: supplementary green book guidance. : HM Treasury; 2021 Jul URL: https://assets.publishing.service.gov.uk/media/60fa9169d3bf7f0448719daf/Wellbeing_guidance_for_appraisal_-_supplementary_Green_Book_guidance.pdf [accessed 2025-12-23]
54. Tiley K, Crellin R, Domun T, Harkness F, Blodgett JM. Effectiveness of 234 interventions to improve life satisfaction: a rapid systematic review. Soc Sci Med 2025;366:117662. [doi: 10.1016/j.socscimed.2024.117662]
55. Whitehead L, Seaton P. The effectiveness of self-management mobile phone and tablet apps in long-term condition management: a systematic review. J Med Internet Res 2016 May 16;18(5):e97. [doi: 10.2196/jmir.4883] [Medline: 27185295]
56. Szinay D, Perski O, Jones A, Chadborn T, Brown J, Naughton F. Perceptions of factors influencing engagement with health and well-being apps in the United Kingdom: qualitative interview study. JMIR Mhealth Uhealth 2021 Dec 16;9(12):e29098. [doi: 10.2196/29098] [Medline: 34927597]
57. Auf H, Dagman J, Renström S, Chaplin J. Gamification and nudging techniques for improving user engagement in mental health and well-being apps. Proc Des Soc 2021 Aug;1:1647-1656. [doi: 10.1017/pds.2021.426]
58. Pinder C, Vermeulen J, Cowan BR, Beale R. Digital behaviour change interventions to break and form habits. ACM Trans Comput-Hum Interact 2018 Jun 30;25(3):1-66. [doi: 10.1145/3196830]
59. Alqahtani F, Winn A, Orji R. Co-designing a mobile app to improve mental health and well-being: focus group study. JMIR Form Res 2021 Feb 26;5(2):e18172. [doi: 10.2196/18172] [Medline: 33635281]
60. Miller JJ, Pozehl BJ, Alonso W, Schmaderer M, Eisenhauer C. Intervention components targeting self-management in individuals with multiple chronic conditions: an integrative review. West J Nurs Res 2020 Nov;42(11):948-962. [doi: 10.1177/0193945920902146] [Medline: 32075541]
61. Rathbone AL, Clarry L, Prescott J. Assessing the efficacy of mobile health apps using the basic principles of cognitive behavioral therapy: systematic review. J Med Internet Res 2017 Nov 28;19(11):e399. [doi: 10.2196/jmir.8598] [Medline: 29187342]
62. Fitzgerald M, McClelland T. What makes a mobile app successful in supporting health behaviour change? Health Educ J 2017 Apr;76(3):373-381. [doi: 10.1177/0017896916681179]
63. Roberts AL, Potts HW, Koutoukidis DA, Smith L, Fisher A. Breast, prostate, and colorectal cancer survivors' experiences of using publicly available physical activity mobile apps: qualitative study. JMIR Mhealth Uhealth 2019 Jan 4;7(1):e10918. [doi: 10.2196/10918] [Medline: 30609982]
64. Obro LF, Heiselberg K, Krogh PG, et al. Combining mHealth and health-coaching for improving self-management in chronic care: a scoping review. Patient Educ Couns 2021 Apr;104(4):680-688. [doi: 10.1016/j.pec.2020.10.026] [Medline: 33143907]
65. Doyle J, Murphy E, Kuiper J, et al. Managing multimorbidity: identifying design requirements for a digital self-management tool to support older adults with multiple chronic conditions. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems: Association for Computing Machinery; 2019:1-14. [doi: 10.1145/3290605.3300629]
66. Stahl D, Rimes KA, Chalder T. Mechanisms of change underlying the efficacy of cognitive behaviour therapy for chronic fatigue syndrome in a specialist clinic: a mediation analysis. Psychol Med 2014 Apr;44(6):1331-1344. [doi: 10.1017/S0033291713002006]
67. Davies T, Jones SL, Kelly RM. Patient perspectives on self-management technologies for chronic fatigue syndrome. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems: Association for Computing Machinery; 2019:1-13. [doi: 10.1145/3290605.3300452]
68. Gulliford M, Alageel S. Digital health intervention at older ages. Lancet Digit Health 2019 Dec;1(8):e382-e383. [doi: 10.1016/S2589-7500(19)30194-3] [Medline: 33323216]
69. Tunney O, Henkens K, van Solinge H. A life of leisure? Investigating the differential impact of retirement on leisure activity.. J Gerontol 2023 Oct 9;78(10):1775-1784. [doi: 10.1093/geronb/gbad097]
70. Tindle R. Improving the global reach of psychological research. Discov Psychol 2021 Dec;1(1):5. [doi: 10.1007/s44202-021-00004-4]

Abbreviations

AI: artificial intelligence
CBT: cognitive behavioral therapy
CFS: chronic fatigue syndrome
COM-B: Capability-Opportunity-Motivation Behavior Model
COREQ: Consolidated Criteria for Reporting Qualitative Research
DHI: digital health intervention

XSL•FO
RenderX

GP: general practitioner
HDI: highest density interval
HH: Holly Health
NHS: UK National Health Service
NICE: National Institute for Health and Care Excellence
ONS4: Office for National Statistics' 4 personal well-being questions
STROSA: Standardized Reporting of Secondary Data Analyses
SWB: subjective well-being
UCL: University College London

Edited by N Cahill, TDA Cardoso; submitted 10.Jun.2025; peer-reviewed by C Yang, N O'Brien; revised version received 08.Dec.2025; accepted 10.Dec.2025; published 04.Feb.2026.

Please cite as:
Symes I, Burton A, Mercado D, Bu F
The Impact of a Health Coaching App on the Subjective Well-Being of Individuals With Multimorbidity: Mixed Methods Study
J Med Internet Res 2026;28:e78738
URL: https://www.jmir.org/2026/1/e78738
doi:10.2196/78738

© Isabelle Symes, Alexandra Burton, Daniela Mercado, Feifei Bu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 4.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Problems and Barriers Regarding the Admission, Financing, and Service Provision of Digital Health Apps: Qualitative Stakeholder Survey

Felix Plescher[1]; Godwin Denk Giebel[1], Dr rer medic; Carina Abels[1], Dr rer medic; Silke Neusser[1], DrPH; Bonnie Kampka[2]; Kirstin Börchers[2], Prof Dr Med; Jürgen Wasem[1], Prof Dr; Nikola Blase[1], Dr med

[1]Faculty of Economics, Institute for Healthcare Management and Research, University Duisburg-Essen, Thea-Leymann-Street 9, Essen, Germany
[2]QM Börchers Consulting +, Herne, Germany

Corresponding Author:
Felix Plescher
Faculty of Economics, Institute for Healthcare Management and Research, University Duisburg-Essen, Thea-Leymann-Street 9, Essen, Germany

Abstract

Background: Since their introduction with the Digital Care Act in 2019, selected digital health apps (DiGA) have been a part of the German statutory health care system. In order to become a DiGA, digital health apps have to complete a certification process covering both technical and evidence-related aspects. After completion, DiGA are added to the DiGA directory, containing a list of all reimbursable DiGA within German statutory health insurance. The first apps were added at the end of 2020, with the number steadily increasing. The novelty of this digital health care service and the fast implementation led to problems and barriers to optimal use along the way, which are studied from different stakeholder perspectives in this paper.

Objective: The aim of this survey was to identify problems and barriers in the context of the admission, financing, and service provision of DiGA in the statutory health care system in Germany.

Methods: We used semistructured expert interviews to evaluate the perspective of stakeholders of the German health care system on DiGA. The interview guide was developed according to Helfferich. The interviews were transcribed and analyzed using the qualitative content approach by Mayring, with the adjustments by Kuckartz. We conducted web-based interviews with stakeholders between February and June 2022. The stakeholder collective consisted of DiGA, statutory health insurance, physician, patient, technological, and quality assurance representatives.

Results: To identify problems from stakeholder perspectives regarding the admission, financing, and service distribution of DiGA, 21 interviews were conducted. The interviewed stakeholders reported problems with the authorization of DiGA and the corresponding process, for example, the duration of the DiGA Fast Track process. DiGA prices and the different negotiation positions were criticized, and financial challenges for smaller DiGA manufacturers were noted. Another problem was seen in the reimbursement of DiGA, independent of actual use by the patients. Within service provision, the participants reported superordinate aspects, for example, the negative public perception of DiGA and negative statements from stakeholders. In relation to the direct care process, technical problems (eg, with activation codes or software surrounding DiGA prescription) and insufficient knowledge and skills on the side of the patients, as well as the medical providers, were mentioned.

Conclusions: Digital health apps have the potential to improve health care by addressing health problems in new, innovative ways. Since the evidence-based and regulated use of this technology is relatively new, problems and barriers limiting the optimized, patient-centered use arose throughout the first years. This study provides an overview of problems and barriers in the context of DiGA in Germany from different stakeholder perspectives. Nevertheless, with these problems being continuously addressed, digital health apps are trending toward becoming a contributing factor to health care in Germany. Since other countries showed interest in implementing a federally regulated approach similar to Germany, valuable implications can be drawn from this survey.

(J Med Internet Res 2026;28:e73332) doi:10.2196/73332

KEYWORDS

mobile health apps; mHealth; stakeholder; DiGA; digitale Gesundheitsanwendungen; digital health apps; challenges; digital health care; health innovation; regulation; mobile health

XSL•FO
RenderX

Introduction

The inclusion and adoption of digitalized processes, products, and the integration and development of digital health care infrastructure is an ongoing process throughout health care systems around the world. Within the German health care system, the transition from analog to digital, connected structures is facing difficulties and still needs improvement (eg, [1,2]). For example, Germany is lagging behind other countries in the process of implementing an electronic health record [2]. Some solutions, for example, video consultation for statutory health care providers, were used frequently after implementation, though this might be partly due to contact restrictions during the COVID-19 pandemic. Other initiatives, like the above-mentioned electronic health record (ePA), are still in evolving stages [3].

German health legislation emphasized the focus on a more digitalized and connected health care system with the introduction of the Digital Healthcare Act. As part of this legislative agenda, selected digital health apps (DiGA) were added as health care services for statutorily insured people. Since the introduction in 2019, DiGA are reimbursable within the German statutory health insurance (SHI) system. As DiGA can be both mobile apps and web-based interventions, the term digital health apps also incorporates both these options. In general, after the completion of a certification process (DiGA Fast-Track process), DiGA are added to a directory, which lists all reimbursable apps. Once DiGA are listed, they are eligible for use within the German SHI system [4]. Over a year after the legislation was passed, the first DiGA were added to the DiGA directory in September 2020 [5]. There are a number of requirements for eligibility for reimbursement within the German SHI system. All listed applications need to be certified as medical devices according to the European Medical Device Regulation [6]. Furthermore, the listing requires the additional completion of a certification process, which is carried out by a federal institute (Federal Institute for Drugs and Medical Devices [BfArM: Bundesinstitut für Arzneimittel und Medizinprodukte]) covering both general criteria and a positive medical effect for the patients. The general requirements include safety, functionality, quality, as well as data security and protection [7]. The positive medical effect can either be proven with medical benefit, for example, improvement of the health status, or patient-relevant improvement of structures and processes, for example, adherence [8].

There are 2 different alternatives when applying for a listing in the DiGA directory, which have to be declared when starting the application process. All apps applying for a listing have to fulfill the general requirements, as described above. If the positive care effect has been proven in studies complying with the study requirements (§ 10 Digital Health Applications Regulation), apps can apply for a permanent listing. Nevertheless, if there is not enough evidence yet to prove the positive care effect, a preliminary listing for a period of up to 12 months (with a possible extension of 12 months) can be obtained. Systematic data evaluations indicating positive (medical) effects and an evaluation concept have to be submitted. The applying DiGA manufacturer has to prove the medical benefit within a period of up to 12 months, with a subsequent listing if successful [8].

Starting with the implementation of the DiGA directory, a total number of 71 DiGA have been listed (as of July 2025) so far. A total of 14 DiGA have been removed from the DiGA directory, leaving 57 currently listed. About 20% (13 of 57 currently listed) of DiGA are preliminarily added to the DiGA directory, with the final proof of a positive care effect still pending. The remaining 44 have proven the positive care effect through scientific studies and are therefore permanently listed. DiGA for mental diseases currently account for the largest share with almost 50% (28/57, 49%). Endocrine, nutritional, and metabolic diseases, diseases of the musculoskeletal systems, and diseases of the nervous system are other examples of indications that currently have multiple DiGA listed [9].

After successful completion of the certification process, manufacturers can set a price they consider suitable for the first 12 months. If DiGA for comparable indications are already added to the DiGA directory, the price cannot exceed the group price. The group price is calculated for DiGA of similar indications in order to ensure equal pricing [10]. After this, a price is negotiated between the National Association of Statutory Health Insurance Funds (GKV-SV: GKV-Spitzenverband) and the DiGA manufacturer, with repayments if the negotiated price is lower than initially set [11].

In order to have the cost reimbursed by the SHI funds, patients have 2 options. They can receive the application if it is prescribed by a physician or directly from the SHI fund without prior prescription (documented diagnosis, eg, in patients' electronic health record within last 6 mo [12]). Both options are fully reimbursed through SHI [13]. Giebel et al [14] described DiGA care and the corresponding processes in detail.

Patients obtain access to the respective apps through activation codes. About 81% of all activation codes have actually been used. The number of used activation codes increased from 50,122 DiGA in 2021 to 374,377 DiGA in total through 2023 to 861,053 at the end of 2024. But not all DiGA are used at the same rate. The top 6 DiGA account for almost two-thirds of the entered activation codes (approximately 500,000). This includes DiGA for metabolic diseases, diseases of the musculoskeletal system, diseases of the ear, as well as mental and behavioral diseases [15,16]. There are also differences when looking at the distribution of DiGA across indications. DiGA for mental diseases accounts for the largest share, with about 259,000 prescriptions. Metabolic diseases make up about 241,000 prescriptions, and diseases of the musculoskeletal system around 137,000 [16].

The introduction and corresponding processes of DiGA were a new and challenging task for the German health care system. As part of a larger research project (Refer to Acknowledgements), which is funded by the Federal Joint Committee (Gemeinsamer Bundesausschuss; Funding code: 01VSF20007), the presented qualitative analysis identified problems and barriers of DiGA regarding the regulation and organization, as well as the use in everyday practice. Even with increasing prescription numbers since the introduction of DiGA, skeptical public statements by different stakeholders remain

XSL•FO
RenderX

[17,18] (see Results for more), showing there is still disagreement about the disruption and success of DiGA. This poses the question of which problems are still persisting concerning DiGA and their corresponding processes.

With DiGA being evaluated for a lot of aspects prior to admission, we adjusted the aim of the study from DiGA itself—for example, problems with the technical components of DiGA—to certification, financing, and use of the same. Furthermore, since just a few countries have actually implemented digital health apps (similar to DiGA) in their health care system, we decided to focus on problems and barriers from the German infrastructural health care system perspective.

Prior research by Giebel et al [1] described problems with implementation of digital health apps "because of a lack of infrastructure […], [l]ack of access and organizational barriers, such as lack of preparedness of health care systems and reimbursement structures." A review investigating mHealth (mobile health) interventions during the COVID-19 pandemic identified barriers such as insufficient interaction, time, or privacy, but also the nonacademic nature of some apps [19]. This shows that there are studies investigating problems surrounding digital health apps themselves, yet they often include a wide range of interventions, potentially limiting the transferability of the generated results. At the same time—to the best of our knowledge—the perspective of a large collective of stakeholders concerning the process and infrastructure around federally regulated digital health apps has not had an exclusive focus in the existing literature. This is particularly interesting in the context of German DiGA, as different stakeholder groups with competing interests (eg, [20,21]) are involved with the implementation of this new federal health care service. Previous qualitative research in Germany covering DiGA has, for example, focused on certain aspects in a specific medical field (eg, [therapeutic] suitability in rheumatology) [22] or the perspective of a single stakeholder group (eg, physicians, patients) [23,24]. In addition, an article addressing a similar stakeholder group [25] has used a broader focus covering a wide range of aspects of DiGA (eg, potentials of DiGA for patients, suggestions for the improvement of DiGA care).

Our study is adding to current literature by focusing on digital health apps from an infrastructural standpoint. As described above, other studies have also mentioned infrastructural problems and barriers. However, these studies, for example, include external mHealth apps and therefore issues with the implementation into a health care system [1,26] or a broader scope [25]. Our approach covers mHealth apps, which have been implemented as a health care service in a statutory health care system, and the corresponding infrastructural issues from a stakeholder perspective. These include DiGA, SHI, physician, patient, technological, and quality assurance representatives, as well as experts from federal institutions. The stakeholder perspective is also important, as other research has suggested that more stakeholder collaboration is necessary in order to realize the full potential of mHealth apps [26].

This survey is being conducted shortly after the first DiGA were implemented; therefore, it will possibly identify first-hand problems during the implementation process. To provide perspective on how the DiGA market has evolved, the following paper also lays out what adaptations have been made to solve them and what problems may still persist. This study contributes to digital health app research by giving valuable insights for health care policy makers aiming to implement a system similar to the German DiGA approach.

Methods

Overview

The qualitative stakeholder survey was planned, conducted, and the results were reported in accordance with the Consolidated Criteria for Reporting Qualitative Studies (COREQ) checklist (refer to Checklist 1) [27]. We identified the perspective of the stakeholders using semistructured interviews. The study consisted of 3 main stages: In the first stage, the interviews were prepared by recruiting participants and developing an interview guide. The second stage included the actual interviews. During the third stage, interviews were transcribed, and data were analyzed and reported as presented below.

Ethical Considerations

The study was presented to and approved to be in accordance with the current Ethics Regulation by the Ethics Committee of the Medical Faculty of the University Duisburg-Essen (no conventional ethics approval vote was necessary). Additionally, the suitability of the data for publication was approved by the Ethics Committee of the Faculty of Economics of the University Duisburg-Essen.

An agreement covering the terms and conditions of participation and data protection was sent to potential candidates beforehand. All stakeholders had to sign the agreement before participating in the study (refer to the Recruitment section).

The interviews were recorded and pseudonymized during the subsequent transcription process. This ensured that transcripts did not contain any identifying information and the statements made cannot be traced back to the respective person (refer to the Interviews and Transcription sections).

Participants did not receive any compensation for participation in the study.

Recruitment

An initial list of stakeholders was assembled during the application for project funding in 2020 (see Acknowledgments for more). Before starting the recruitment process, this list was reviewed, and necessary adaptations were made. These adaptations were based on the DiGA legislation, which was not available during project application [7,8,13,15,16]. The documents were checked for additional stakeholder groups, and new, not yet included ones were added. The final list included the association of SHI physicians, representatives of patient organizations, the SHI funds, digital health app companies, and experts in the field of quality management, data security and protection, as well as medical device regulation. The recruitment process started in December 2021 and continued until May 2022. A total of 23 organizations were contacted. After each interview, the research team decided if sufficient information for the stakeholder group (in reliance on their involvement in

XSL•FO
RenderX

DiGA processes) was collected. Once saturation was reached for the stakeholder group, no further requests were sent out. However, if stakeholders, who were contacted before, decided to participate in the study, they were also included. Patients (representatives) were only represented by a small subsample, even though they are of particular importance in the context of DiGA, because we did an additional interview and focus group study prior to this study, exclusively covering patients' perspectives [28].

We approached the representatives of the organizations by direct phone call or e-mail. We requested the participation of the employee responsible for quality assurance, if available. If participation in the study was agreed, additional information about the topics of the discussion and the terms and conditions of participation and data protection was transmitted. We provided information on DiGA and the research project in order to ensure all participants had the same base knowledge before taking part in the interviews. Nonresponders to the telephone call were followed up by e-mail (directly after the phone call) containing the aforementioned information and additional general information about the study. An additional follow-up was sent out one week after the initial phone call. Participants had to sign the agreement about participation and data protection before the start of the interview.

Interview Guide

The interview guide was developed according to Helfferich et al [29]. We used a semistructured interview guide to provide guidance and cover the relevant topics, but also leave room to address aspects not included in the interview guide. As the interviews are part of a larger research project, the interview guide was based on results generated earlier in the project. A scoping review [1] covering general problems and barriers related to digital health apps provided the evidence base for the main topics. This was extended using results from a—as mentioned above—qualitative survey with DiGA users and nonusers [28]. As the previous study modules focused on the apps themselves, the interview guide aimed at identifying problems and barriers concerning the infrastructure and process surrounding DiGA. Open questions were used to identify aspects without influencing the interview participants. More precise follow-up questions were asked to inquire more deeply about certain aspects. The interview guide was not available to the study participants prior to the interviews. Only the topics of the interview guide were communicated. We carried out a pretest interview and made adaptations to the interview guide in order to improve comprehensibility and feasibility. As the following interviews were conducted within the planned time frame and the interview partners did not have problems answering the question, the adapted interview guide was used throughout the remaining study.

Interviews

The interviews were conducted by 4 of the authors of this article (NB, CA, GDG, and FP) between February and June 2022. Two of the interviewers are female and 2 male, with 2 senior researchers (PhD: Dr med/Dr rer medic) and 2 junior researchers (MSc/MA). All interviewers are part of the QuaSiApps project team (NB is project lead), with prior experience in qualitative research projects. During the interviews, only the interview study participants and part of the research team were present. There were no prior relationships between the researchers and any of the study participants. Participants did not receive any compensation for participation in the study.

The interviews were carried out using the videoconference platform ZOOM (Zoom Video Communications Inc). One of the researchers moderated the conversation with at least one additional researcher taking notes on the main talking points and overseeing the technical infrastructure. The interview sessions were recorded in accordance with the signed agreement using the integrated recording function of ZOOM. An additional disclaimer was stated by the moderator before starting the recording. The duration ranged from about 50 minutes to 1 hour and 45 minutes.

Transcription

After the conclusion, the recorded interviews were transcribed following the transcription rules by Claussen et al [30]. The transcript was revised by an independent reviewer from the research team, not involved in the transcription. After discussing the remarks and correcting mistakes, the transcript was finalized. All individuals taking part in the interviews, as well as organizations and institutions, were pseudonymized. Therefore, no connections can be made between the statements made and the participants.

Analysis

We analyzed the interviews in accordance with the qualitative content analysis approach by Mayring and the extensions by Kuckartz (content-structuring qualitative analysis) [31,32]. The analysis was carried out as described below using MAXQDA (VERBI), a software tool for qualitative research [33]. The transcripts were imported into MAXQDA, and a coding system containing deductive and inductive codes served as a tool for content extraction. The main topics of the interview guide provided the deductive codes. These included the codes "Problems: Admission," "Problems: Financing" and "Problems: Service Provision," which represent the primary outcomes of our study. We also added deductive codes covering problems regarding the main stakeholders in the context of DiGA, for example, DiGA manufacturers, SHI funds, or patients. Throughout the first coding process, inductive codes were added to the code system. The full coding scheme was reviewed, discussed, and finalized by the research team (Multimedia Appendix 1). After completion of the first coding process and the finalization of the coding scheme, all interview transcripts were coded again using the final code system, which included all inductive and deductive codes. This process was also carried out by 2 researchers independently, with a third researcher reviewing coded transcripts and deciding in case of disagreement. The codes were further grouped into the 3 categories: admission, financing, and service provision. This step was reviewed, discussed, and necessary adaptations were made by the research team. The results of the analysis and corresponding quotes were sent to the study participants for review. We did not receive any additional comments or requests for corrections.

XSL•FO
RenderX

Results

Interview Sample

Of the 23 contacted stakeholders, 2 declined and 21 agreed to participate in the study. None of the 21 stakeholders dropped out after further information about the study and the terms and conditions were sent out. The interviews were conducted with all 21 stakeholders. None of the stakeholders decided to withdraw after the interviews were completed. The characteristics of the 21 included participants are presented in Table 1.

Table . Surveyed stakeholder groups.

Stakeholder group	Participants, n (%)
DiGA representatives[a]	7 (33.3)
SHI representatives[b]	6 (28.6)
Physician representatives	2 (9.5)
Patient representatives	1 (4.8)
Representative from the Federal Joint Committee	1 (4.8)
Technological experts	2 (9.5)
Quality assurance experts	2 (9.5)

[a]DiGA : Digitale Gesundheitsanwendung (digital health application).

[b]SHI: statutory health insurance.

Interview Results

The results are presented for each of the 3 subtopics: admission, financing, and service provision. Anonymous quotes from the interviews are used to illustrate the mentioned aspects. In order to guarantee that, only the stakeholder group (eg, DiGA manufacturer) is mentioned. If not already described in the introduction, we provided additional context throughout the Results section about the specifications of the German DiGA systems in order to ensure a more comprehensive understanding of the presented problems and barriers.

As we aimed to investigate problems and barriers, we also wanted to mention that a large share of stakeholders expressed a generally positive attitude toward DiGA and DiGA processes during the early implementation phase. Better access to health care services for patients in rural regions (patient representative), bridging long waiting times (DiGA representative), and supporting general care processes (SHI representatives) are a few examples of potential benefits of DiGA, which were suggested during the interviews.

Admission

Even though the fast process for digital health apps to become DiGA was generally seen as a good innovation, different stakeholders mentioned problems in the way the DiGA Fast-Track is structured and the corresponding responsibilities. Especially the DiGA manufacturers expressed that there is not enough expertise to evaluate—in particular—the technical, nonevidence-related aspects.

> *So, the BfArM review is very focused on the whole topic of clinical evidence, e.g., study planning. That's their main task, you don't have to talk about it. Yes, so I think in the area of usability, user-friendliness, that could be-, so that's not their topic either. But I don't think it would do it any harm to go into it in more depth […].* [DiGA manufacturer]

Other problems were seen in the high volume of applications for the BfArM and the low level of flexibility (DiGA manufacturer), which was partly credited to the new and challenging nature of implementing a new digital health solution (SHI representative).

There was also criticism about the duration of the DiGA Fast-Track. Upon application by the DiGA manufacturer, a decision has to be made by the BfArM within 3 months [34]. The shift away from the regular admission process of medical products into SHI (quality assurance stakeholder) and the relatively short certification process raised concerns about the quality of the assessment (quality assurance stakeholder, SHI representative, physician representatives).

> *Because the assessment process has been rammed out of the ground by the legislator in a relatively short time, because it all wanted to happen very quickly and that is also - the terminology speaks for itself, fast-track procedure, eg, the fastest possible market entry is in our view a bit in the foreground and then naturally the proof of benefit comes a bit short. So that would be a structural problem or a very fundamental problem, which we still see.* [Physician representative]

A similar aspect with a shifted focus was mentioned by DiGA manufacturers. They criticized that the 3-month assessment period is not put on hold when, after all initial paperwork is turned in, the BfArM requires additional paperwork from the applying DiGA company. This led to very short deadlines to process, produce, and turn in the required documents (DiGA manufacturer).

A diverse range of perspectives was reported surrounding the authorization requirements for both general and evidence-related aspects. On the one hand, the requirements were seen as too low (SHI representative); on the other hand, the steadily

XSL•FO
RenderX

increasing requirements were described as limiting for the flexibility of innovation (DiGA manufacturers).

Since DiGA must be medical products and therefore are/were subject to external certification when obtaining the CE certification, problems surrounding the notified bodies (German: "Benannte Stellen," eg, Technical Inspection Association) were also mentioned. Especially in the early days of DiGA certification, the insufficient levels of knowledge about DiGA within these companies and the low availability of appointments for certification were described (certification companies, DiGA manufacturer):

> *And there are extreme waiting times and a kind of undersupply. So many people have a problem getting appointments at the designated places [certification companies]. So, there's already a big bottleneck and you have to look at how realistic it is to implement it.* [DiGA manufacturer]

There was also criticism concerning the innovation of the preliminary approved DiGA. Physician representatives admonished that the possibility of preliminary approved DiGA yields apps with low evidence to be reimbursed by SHI funds. The equivalent handling of preliminary and permanent approval may also distort the level of evidence between different DiGA (physician representative, SHI representative, DiGA manufacturer).

> *Yes, what is also problematic, for example, is the fact that there are two stages. On the one hand, there is the stage where you are finally included in the BfArM directory and on the other hand, there is the stage where you are temporarily included and then have to provide evidence.* [SHI representative]

> *What is viewed negatively, at least from my point of view in terms of quality, is that we have provisional approval that is not based on RCTs, which has a certain risk because it suggests that these products without evidence have the same efficacy as products with evidence and that is of course a problem, especially for the medical profession, which we have been trying to push into evidence-based medicine for 20 years.* [DiGA manufacturer]

A problem described by DiGA manufacturers regarding the conclusion of the Fast-Track process was the inconsistent handling of the market access. One manufacturer reported that after successfully becoming a DiGA, the first prescriptions are possible the next day, but in their case, no listing in the prescription software led to no prescriptions within the first days. A different manufacturer reported the same problem with no possibility of prescriptions due to the unavailability of a necessary interface port to the prescription software (DiGA manufacturer).

Financing

The interviewed stakeholders mostly talked about the price of DiGA, the process of negotiation between the SHI funds and the DiGA manufacturer, and the price-related policy. DiGA manufacturers said the price expectations of health insurance funds indirectly show low willingness to negotiate and lead to high initial prices (DiGA manufacturer).

> *The offers that the GKV-SV distributes here in the market mean that manufacturers anticipate that they will take the highest possible price at the time when it is possible, so that they can somehow reduce some of the debt that they have accumulated.* [DiGA manufacturer]

They also criticized the superior negotiation position of the GKV-SV—the head organization of all SHI funds in Germany, since they get to deal with multiple DiGA manufacturers, but a single DiGA manufacturer only gets to deal with the GKV-SV once (DiGA manufacturer). One health insurance fund representative mentioned that—even with this negotiation position—they have difficulties achieving favorable SHI positions in price negotiations (SHI representative).

Multiple stakeholders took a look at the finance aspect from the DiGA manufacturers' view and laid out that the financing of the app and corresponding activities (eg, evaluation studies, personnel) are challenging, especially for small companies (DiGA manufacturer, technological expert, patient representative).

> *However, due to the hurdles posed by the legislator, it is not always easy for start-ups to raise the basic capital to gain a foothold in the market, especially when it comes to financing.* [Technological expert]

One SHI representative mentioned that the refund of excess payments after price negotiations might lead to a market shrinkage, due to the superior financial resources of large companies compared with smaller DiGA manufacturers. Additionally, besides general criticism, a separate regulation for preliminary and permanently listed DiGA was called for, as the described regulation above is currently in place regardless of admission status. The necessity of usage and quality-related pricing was also pointed out (SHI representative).

As of now, DiGA are being reimbursed as soon as the patient downloads the app and enters the DiGA activation code, provided by the health insurance fund. Therefore, reimbursement is independent of usage. There were concerns among SHI representatives about the financing of DiGA without knowing if the app is actually used by the patient.

> *Problems are certainly, as an initial survey of insured persons by the health insurance fund [name anonymized] has also shown, problems are certainly that you don't know whether and how the insured person uses this DiGA in the end.* [SHI representatives]

Service Provision

In the early days of DiGA, there was disagreement among the questioned stakeholders about how well the communication between different stakeholders was going. Especially, the manufacturers emphasized that negative public reporting does not help DiGA flourish. According to them, different organizations made negative remarks about DiGA or even gave out "warnings" (DiGA manufacturer). General skepticism

XSL•FO
RenderX

among stakeholders when it comes to digitalization may also have an impact on how well DiGA are perceived. Multiple DiGA manufacturers added that this was accompanied by negative public statements and comments by SHI representatives about DiGA, which further hindered the development.

> *I often see very sweeping arguments being used when DiGA are criticized. Often the basics are not so well known, the legal and that is of course-, so DiGA are sometimes also seen as a symbol for digitization per se. And then perhaps no distinction is made as to whether it is a project such as an e-prescription or a DiGA, but it is simply a blanket term for digitization and there is a great deal of skepticism* [DiGA manufacturer]

The DiGA directory and the lack of (technical) information presented in it were also criticized among stakeholders. SHI and physician representatives both reported that they did not, or only in a limited fashion, receive access to selected DiGA (SHI representative, physician representative). The DiGA directory could not fill this knowledge gap sufficiently either, as especially SHI representatives pointed out that information about DiGA was too complex and no information on the quality of the applications was provided:

> *We would say that the issue of transparency, or rather the lack of transparency, is a major problem from our point of view. Also, in terms of the lack of comparison options for the apps. We are of the opinion - at the moment, the DiGA directory, the BfArM directory is basically the source that you can look at. […] In other words, to prepare it differently for patients and physicians. We still see a lot of room for improvement, especially as it is currently the case that - when you look at the DiGA directory, you can't really find anything about quality at first glance.* [SHI representatives]

Low cooperation was also seen in the work of medical software companies (German: “Praxisverwaltungssoftware [PVS] – Hersteller,” PVS manufacturer) responsible for technical solutions in SHI medical practices. This showed in slow adoption of existing practice software (Technological expert), establishment of own structures for app prescriptions (DiGA manufacturer), and little interest and knowledge in new DiGA policies:

> *And then I call a medium-sized PVS manufacturer and they say: "No, we don't make apps. “[…] They just didn't realize it and of course, the legislator doesn't have to do the work for us now and inform all the PVS manufacturers and message them to implement all this, but it’s important for those who want to make the DiGA project a success to understand: "The PVS manufacturers are sometimes to often not at all helpful to harmful."* [DiGA manufacturer]

Stepping away from the superordinate perspective and looking at problems and barriers to better use of DiGA within the direct care context, problems surrounding DiGA prescribing physicians, patients, and their interactions were reported. Especially in the early days of DiGA, the interviewed stakeholders reported insufficient knowledge about DiGA, their content, and the corresponding regulations among physicians (DiGA manufacturer, SHI representative, patient representative, physician representative). This was generally acknowledged by physicians while emphasizing the limited capacity for DiGA within a regular workday:

> *Today I had another Monday where [I] didn't even know where to stack the patients. And then when some people come in who have problems with their app and need technical support, I find it difficult in some cases how to accommodate that in an open consultation. That might be a bit out of touch with reality.* [Physician representative]

They also reported heterogeneous interests and skepticism among colleagues and concerns of (partially) being replaced by DiGA at some point (physician representative).

Different stakeholders agreed about insufficient awareness of the information services and content of the DiGA among patients. Patients did not know what DiGA were or how to claim them. The process of going through the SHI fund to receive the activation code hindered patients from using DiGA (DiGA manufacturer, patient representative). Patients also showed low compliance levels according to an SHI representative, which they found out through a company internal survey. There was also concern among patients that the “normal” treatment option may be withheld from them if they opt in to using a DiGA (patient representative).

> *And another problem for patients is that they sometimes have the feeling that if they use a DiGA, they will be deprived of another therapy, such as behavioral therapy. So, if you have the feeling that you have to choose either or, then in my experience most patients tend to go down the usual treatment routes.* [Patient representative]

Stakeholders also expressed similar opinions regarding problems with the prescription process. In some cases, low prescription quality (eg, missing diagnosis) was registered by SHI funds when processing DiGA prescriptions (SHI representative). As mentioned above, the prescription process of DiGA was also criticized, and a digitalized adaptation was called for (DiGA manufacturer). Closely interrelated with the prescription process, the production and distribution of DiGA activation codes were described as problematic in the early days of DiGA. Incorrect codes were sent out, which led to patients not being able to use the DiGA (DiGA manufacturer, SHI representative).

Two stakeholders criticized the missing combination of DiGA use and therapy by physicians. They reported that, especially for risk-associated DiGA (eg, DiGA for mental diseases), monitoring and guided app use might be necessary (SHI representative, patient representative).

> *There are also DiGA that are associated with risks, especially when I think of DiGA for depression. This prescription and use should at least be accompanied by a physician. So, I see problems here if the physician basically just prescribes and doesn't stay*

XSL•FO
RenderX

> *in conversation with the insured person about how the DiGA actually works.* [SHI representative]

At the time of the study, DiGA were meant to be used mainly by the patients themselves. Active involvement of physicians was possible, but only a few DiGA incorporated such tools [9]. DiGA manufacturers described the integration of physicians with existing legal regulations as difficult and added that this may cut back on the potential of DiGA (DiGA manufacturer).

Discussion

Principal Findings

The aim of this study was to identify problems and barriers regarding the admission, financing, and use of DiGA. As other studies have focused on the apps themselves, we wanted to determine which problems concerning the infrastructure and processes influence optimal use of this new digital health care service. The findings show that some problems can be attributed to the early days of DiGA, while other issues are more prevalent for long-term success. Regarding the admission, the DiGA Fast-Track itself was criticized because of insufficient levels of expertise among the responsible institutions and the duration of the process. From a financial perspective, problems regarding the price negotiations and reimbursement were most frequently discussed. The superior negotiation position of the GKV-SV was pointed out, as well as criticism that reimbursement is independent from DiGA usage. DiGA service provision was criticized on a superordinate level because of, for example, the public perception of these new health care services and negative statements among involved stakeholders. As aforementioned, insufficient knowledge and skills surrounding DiGA were also a problem among patients and health care professionals, along with early issues of the prescription process and a lack of integration of physicians into DiGA use.

Admission

As mentioned before, some of the identified problems and barriers can be attributed to the short existence of DiGA at the time the survey was conducted (first DiGA: 09/2020, Interviews: 02-06/2022). DiGA manufacturers pointed out the challenges associated with a 3-month certification process, including review and additional paperwork. To this day (July 2025), only 25 applications were rejected after the DiGA Fast-Track process. Notably, 124 applications have been withdrawn from the certification process [34]. As we cannot view all reasons of withdrawal, we can assume that, at least to some extent, it may have been due to deadlines, which were not or could not be met. This leaves the question of whether just the quality of applications of digital health apps is not good enough, or if process adjustments may have the chance of allowing more qualitatively high digital health apps access to the DiGA directory.

In comparison to the 57 currently listed DiGA, over twice as many applications and digital health apps have not been included in the directory. In addition to withdrawals (n=124) and negative decisions (n=25), 14 DiGA have been deleted from the directory after originally being included (total number of DiGA not included: n=163). Again, as the reasons for unsuccessful applications cannot be viewed, this may also suggest that the DiGA Fast-Track with its requirements has the potential to ensure quality and reliability of DiGA [34].

This may be a small indicator that concerns about the DiGA Fast-Track allowing digital health apps to easily enter the German statutory health care market should be examined in more detail. The length of the process was not the only argument for this position. We identified a disagreement about the strength of the requirements of the DiGA Fast-Track as well. While DiGA manufacturers highlighted a high (but bearable) burden for successful completion of the DiGA Fast-Track, especially SHI-related stakeholders mentioned concerns about the (missing) strength of the requirements. Already in the early days of DiGA, the DiGA Fast-Track and its framework (eg, preliminary approved DiGA) were subject to SHI criticism [35]. As laid out above, the high number of withdrawn and negatively decided DiGA applications indicates that existing requirements cannot be met by a large share of applying companies. In addition, the underlying legislation (Digital Health Applications Regulation) has been updated frequently since its introduction in 2021 [36]. It has to be kept in mind that up until the “Act to accelerate the Digitization of the Healthcare System (Digital Act–DigiG; effective since March 2024)”, only DiGA of low-risk classes were reimbursable within SHI. The expansion to higher-risk classes within the act may influence the appropriateness of the DiGA Fast-Track requirements. For example, the act includes the requirement for DiGA of higher risk classes to prove the medical benefit through a prospective comparative study [37].

The preliminary approval of DiGA caused some controversy, as it presents a new and never-before-seen approach to introducing health care services into the SHI system. We like to point out that even if no final results are necessary to obtain preliminary approval, there are still requirements tied to this status. In addition to meeting all “general” requirements (eg, security, functionality, quality, data security), a preliminary systematic data evaluation as well as an evaluation concept is necessary, plausibly showing that the medical benefit can be proven within the preliminary approval period [4]. As this still bears the possibility for the evaluation studies to show the corresponding DiGA does not provide a medical benefit, it opens the door for an early market entry of new, innovative, potentially disruptive technologies. Nevertheless, DiGA are used by patients with diseases. The worst case for any health care services would be harming the patient. To ensure health care services are safe and beneficial to the patient, studies providing evidence of such effects with final, statistically significant information are conducted. These studies must also fulfill high-quality standards. Existing literature shows that some DiGA have deficiencies regarding the quality of the studies showing the medical benefit (for example, lack of blinding, inhomogeneous dropout rates) [38]. Therefore, in the case of DiGA, the balancing of innovation and (final) evidence may currently be prioritizing innovation.

The implementation of admission processes for DiGA-equivalent digital health apps in other European countries (eg, France and Belgium) shows that there is an agreement about DiGA being a necessary part of health care systems in the future.

XSL•FO
RenderX

This is underlined by ongoing discussions about the implementation of admission processes in additional European countries, for example, Spain, Austria, and Italy. However, some countries do not prioritize this topic as of now, which cannot be neglected [39].

Financing

Finally, one of the most talked-about topics in the public discourse surrounding DiGA is their price and the corresponding processes, for example, negotiations between DiGA manufacturers and the National Association of SHI funds. Some of the interviewed experts (eg, physician representatives, SHI representatives) criticized the possibility for DiGA providers to set their own price for the first 12 months of approval. To keep DiGA prices within a reasonable and comparable range, the legislator made adjustments regarding the pricing processes. A collective agreement between the National Association of SHI funds and leading DiGA manufacturer associations was put in place, introducing maximum prices for DiGA, setting the range for the price of the DiGA in relation to other DiGA treating the same or a comparable disease [40].

The opposing views by SHI representatives and DiGA manufacturers regarding the price of DiGA may be due to their competing interests. As the SHI funds are paying for the use of DiGA by their insured persons, lower and criteria-based (eg, quality, usage) prices can lead to lower costs for the corresponding SHI fund. At the same time, DiGA manufacturers are financing (to a significant extent) their company through the income generated from patients using their app, therefore having an interest in higher prices.

Service Provision

A problem medical providers mentioned was the current prescription process, calling for a more digitalized prescription process (“Digital health solution needs digital prescription” – DiGA manufacturer). This was shared by patients, as some of them did not want to use the DiGA because of the complicated prescription process. With the mandatory introduction of the e-prescription in Germany starting in January 2024, digital prescriptions are now the standard within German health care. They are currently only used for prescribable drugs, but are planned to be extended to DiGA in 2025 [41]. This has the potential of further boosting DiGA as the prescription process and, therefore, the access to DiGA can improve once the e-prescription is established [42]. With problems in the early days of e-prescription now solved, they are mainly seen as an improvement among representatives within the German SHI system [43] and may therefore contribute to more DiGA use when extended. However, even if they are not the main target group of DiGA, the preferences of people with low digital literacy and non–digital natives also need to be incorporated to create an option of low-barrier access to DiGA. This may be of particular relevance for indications that have a substantial number of patients in higher age groups.

A problem was also presented in the low knowledge and skills around DiGA by both the prescribing (medical providers) and the using parties (patients). One tool to tackle this challenge was the DiGA directory, providing a range of information separately presented for health care professionals and patients [9]. Our study showed that for a share of the involved stakeholders, this goal has not been met yet. As growing prescription and usage numbers indicate (see SHI Reports [15,16] for more), we assume the lack of knowledge was—at least to some part—due to the nature of the new digital health service and the introduction during the COVID-19 pandemic. The importance of knowledge and skills regarding digital health care services has also been emphasized by the World Health Organization. Borges do Nascimento et al [44] identified digital health services as possibly “a real asset to the health and care workforce.” To realize these potentials, improving the digital health literacy among health care professionals is a fundamental step.

To further boost DiGA and enable it to thrive, more integration of medical providers in the use of DiGA might be a solution. International Studies have shown that this also has the potential to increase the adherence of the patients [45,46]—which was criticized by SHI representatives—assuring that reimbursed DiGA are being used as good and intended as possible to create a medical benefit for the patient. A large German health insurance fund reported that almost 40% of insured people with prescribed DiGA did not use the app over the intended time span. Of the people who discontinued using the DiGA, only about 20% reported it was due to improved health [47]. To improve adherence and guide the patient, some DiGA have integrated modules to include medical providers [9]. Monitoring of patient data and consultation of the patient at follow-up visits are some examples of these tasks. This appeal results in a call for reimbursement of the corresponding tasks (eg, monitoring of patient data, consulting services) carried out by the medical provider. Up to today, only a few DiGA have negotiated reimbursement for their DiGA-supporting tasks [48]. Applying similar structures to more DiGA may improve adherence among patients and, therefore, the conditions for DiGA to provide a medical benefit to the patient.

However, more integration of health care providers cannot be seen as an isolated option to improve DiGA use. One fundamental component of the long-term success of health care services is the acceptance among medical providers and patients. Since the evaluation reports [15,16] have shown that over 90% of DiGA are prescribed by medical providers, they play a key role as they are (for a large share) the gatekeepers in the distribution of this health care service. At the same time, acceptance of patients is also important, since they are the ones using DiGA. A web-based survey [49] covering potential users in Germany showed that there is a generally high willingness (76%) to use digital health apps. Still, the authors conclude that “targeted communication around effortless usage of mHealth services across age groups and demographics” is necessary to improve digital health app adoption. Another study [50] (covering mental illnesses) adds to this conclusion, for example, outlining that there is unequal access to digital health technologies for some patient groups, and this might lead to the worsening of health inequalities. This emphasizes the importance of addressing the acceptance of medical providers and patients in order to improve digital health app use.

XSL•FO
RenderX

Many of the presented problems show a strong reliance on the German health care system. Nevertheless, the results present an overview of relevant problems for researchers and policymakers from different health care system approaches when considering the implementation of digital health apps. The early days of DiGA contained a wide range of first impressions and statements from stakeholders. As we found out within this study, there was disagreement among involved parties about different aspects of DiGA within the German health care market. There are various reasons why stakeholders potentially disagreed, such as competing interests or system logic. On the one hand, DiGA are brought into the market by DiGA manufacturers, who get reimbursed for their service. On the other hand, as DiGA are part of the statutory health care system in Germany, they are reimbursed by SHI funds. These 2 stakeholders have competing interests, and this showed in opposing statements. One example is the authorization requirements for DiGA, where SHI representatives suggested the criteria are too low, when at the same time DiGA manufacturers described the increasing requirements as challenging (refer to Results: Admission). Therefore, it has to be kept in mind that some of these contradictory statements are also influenced by interests and system logic as described above.

Nevertheless, DiGA made their start in 2020 and have been growing since. As laid out throughout this article, this was accompanied by challenges along the way, not unexpectedly due to the iterative nature of the legislative process. On the side of the stakeholders involved in the actual distribution of DiGA, many challenges can be attributed to the early days. These aspects, therefore, should be addressed early on when implementing a DiGA-equivalent health care service in other countries.

Limitations

First, we interviewed stakeholders from relevant stakeholder groups. Nevertheless, within each stakeholder group, we could not completely ensure representativeness; for example, we did not interview a DiGA manufacturer for each disease in the DiGA directory. The sample is also not representative of factors like the number of insured people with the SHI funds or the share of the corresponding DiGA indication among all listed DiGA. In order to still cover a large share of the involved stakeholders, we included central associations of the corresponding stakeholder groups (eg, GKV-SV, DiGA-Spitzenverband [DiGA-SV]). Even though we tried to cover all relevant stakeholder groups by sending out multiple requests, some declined participation or opted for later involvement in the research project, for example, BfArM. We also only included one patient representative, which may lead readers to think this important stakeholder group is underrepresented in our study. We conducted a focus group and interview study exclusively covering the patient's perspective prior to this study and therefore refer to the corresponding article for more insights into patients' views [28].

We want to point out that all statements made are subjective and only represent the view of the interviewed stakeholder and may or may not translate to DiGA in general. The roles of the participants or the stakeholder group, which they represented, and the corresponding statements may have been influenced by personal interests or a political agenda. Statements were checked with existing evidence but remain subjective to some extent. Since this research paper aims to report problems surrounding DiGA from different stakeholder perspectives, this limitation has to be accepted in order to reach the research goal.

When conducting an interview study, the role and position of the involved researchers is an aspect that can generally influence the results of the study (eg, personal preconceptions regarding the research topic, prior collaboration between researchers and interview participants). To minimize the influence of this limitation, all steps of the study were generally carried out by at least 2 researchers. The generated data was coded and analyzed by 2 researchers independently, with subsequent cross-checking by a third experienced senior researcher. In addition, all transcripts and codes were quality assured by a researcher not primarily involved in the corresponding interview in order to identify any analytical results and conclusions influenced by preconceptions.

Another limitation is the period of time during which we conducted the interviews. Since we started conducting the interviews in February 2022, there have been only a few DiGA inducted into the DiGA directory. This ultimately led to a few prescribed DiGA, which decreased the representativeness of the generated evidence. To account for this limitation, we laid out which of the presented problems have already been solved and how processes and the general perception of DiGA may have improved since then. Still, we cannot fully guarantee that some of the aspects we described may be transitional or only related to early implementation efforts and may have been resolved after the study period. At the same time, we assume that some of the problems, for example, the imbalance between stakeholders during price negotiations or the different perceptions toward the preliminary approval of DiGA, may not be transitional and still have an influence on the current DiGA landscape.

The online format (interview via videocall) may have also impacted the quality of the generated data, as it was not possible to control, for example, for additional people in the room or technical problems. However, since the interviews were conducted almost 2 years after the COVID-19 pandemic started, online video tools were already common. There were also no major technical problems during the interviews. There is research indicating that there are no significant differences between online and face-to-face interviews [51,52].

Conclusions

With the introduction of DiGA to the German health care system, a step toward more digitization was taken, with the potential of improving care for patients. Nevertheless, there was also skepticism and risks to how well the new technology would improve and fit into the German health care system. The identified problems can be seen as "necessary" steps in the development and evolution of DiGA, for example, early problems with activation codes, insufficient knowledge about DiGA among involved actors, and different perceptions about the prices of DiGA. Some of the barriers have already been addressed along the way, for example, different adaptions with

XSL•FO
RenderX

pricing of DiGA. This goes along with new health policies further expanding DiGA to higher-risk classes, which opens up opportunities for new DiGA and, therefore, more patients to benefit from it. This also shows in a higher number of DiGA being prescribed and more DiGA being added to the DiGA directory.

Therefore, we are concluding that problems and barriers have been and are still impacting DiGA on the way to becoming a contributing component of patient care. To identify the current state of DiGA and which problems may have been solved and additionally arisen through the years, new research investigating the perspective of a similar group of stakeholders may be necessary. We also want to point out that although we aimed at identifying problems and barriers in the early days of German DiGA use, one result of our study was that a large share of the questioned stakeholders had a positive attitude toward the potential of DiGA and the processes in the early implementation phase. Yet, as our study has shown, there is still room for improvement so that DiGA can reach its proclaimed potential.

Finally, the identified problems and barriers can help other countries when implementing DiGA-equivalent digital health apps. With conclusions drawn from the presented challenges, an optimized implementation process may lead to an early realization of positive effects.

Acknowledgments

Generative artificial intelligence was not used for any part of this research article.

Funding

This qualitative stakeholder survey is part of the research project called “Continuous quality assurance of health apps used in SHI standard care (QuaSiApps)” funded by the German Federal Joint Committee (Gemeinsamer Bundesausschuss, G-BA; funding code: 01VSF20007). We acknowledge support from the Open Access Publication fund of the University of Duisburg-Essen.

Data Availability

Information and data generated or analyzed during the research project (QuaSiApps) are available from the corresponding author on reasonable request.

Authors' Contributions

All authors were involved in the conception and design of the study. Participants were recruited by FP and CA. The interview guide was constructed and redefined by FP, CA, GDG, and NB. FP, NB, CA, and GDG conducted the interviews. The transcription and qualitative analysis were performed by FP. FP wrote the first draft and the revision of the manuscript, which was critically revised by NB, CA, GDG, SN, BK, KB, and JW. The final version of the manuscript was read and approved by all authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Coding scheme.
[DOCX File, 19 KB - jmir_v28i1e73332_app1.docx]

Checklist 1
COREQ checklist.
[PDF File, 424 KB - jmir_v28i1e73332_app2.pdf]

References

1. Giebel GD, Speckemeier C, Abels C, et al. Problems and barriers related to the use of digital health applications: scoping review. J Med Internet Res 2023 May 12;25:e43808. [doi: 10.2196/43808] [Medline: 37171838]
2. Digitalisierung im Gesundheitswesen dringend weiterdenken [Urgent need to further develop digitization in health care]. Deutsches Ärzteblatt [German Medical Journal]. 2022. URL: https://www.aerzteblatt.de/nachrichten/135229/Digitalisierung-im-Gesundheitswesen-dringend-weiterdenken [accessed 2024-07-22]
3. Die elektronische Patientenakte (ePA) [The electronic health record]. Bundesministerium für Gesundheit [Federal Ministry of Health].: Federal Ministry of Health; 2024. URL: https://www.bundesgesundheitsministerium.de/elektronische-patientenakte.html [accessed 2025-01-06]
4. The Fast-Track-Process for digital health applications (DiGA) according to § 139e SGB V.: Federal Institute for Drugs and Medical Devices; 2023. URL: https://www.bfarm.de/SharedDocs/Downloads/DE/Medizinprodukte/diga_leitfaden.pdf?__blob=publicationFile [accessed 2025-12-27]

XSL•FO
RenderX

5. BfArM nimmt erste "Apps auf Rezept“ ins Verzeichnis digitaler Gesundheitsanwendungen (DiGA) auf [BfArM adds first “apps on prescription” to its directory of digital health applications (DiGA)].: Federal Institute for Drugs and Medical Devices; 2020. URL: https://www.bfarm.de/SharedDocs/Pressemitteilungen/DE/2020/pm4-2020.html [accessed 2024-07-22]
6. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. EUR-Lex. 2017. URL: http://data.europa.eu/eli/reg/2017/745/oj [accessed 2026-01-07]
7. § 139e Directory for Digital Health Applications. Social Security Code V. In: Statutory Health Insurance: Federal Law Gazette; 1988:2477.
8. Digital health applications regulation of April 8, 2020 (Federal Law Gazette). : BGBl; 2024 URL: https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/Gesetze_und_Verordnungen/GuV/D/DiGAV_RefE.pdf [accessed 2026-01-06]
9. Federal institute for drugs and medical devices. DiGA-directory. 2024. URL: https://diga.bfarm.de/de [accessed 2024-07-22]
10. Group prices and thresholds. National Association of Statutory Health Insurance Funds. 2024. URL: https://www.gkv-spitzenverband.de/krankenversicherung/digitalisierung/kv_diga/hoechstbetraege_und_schwellenwerte/hb_sw.jsp [accessed 2025-01-06]
11. § 134 Agreement between the National Association of Statutory Health Insurance Funds and the Manufacturers of Digital Health Applications on Remuneration Amounts; Regulation Authorization.: Statutory Health Insurance; 1988. URL: https://www.gesetze-im-internet.de/sgb_5/__134.html [accessed 2026-01-07]
12. Eine DiGA beantragen [Appyling for a DiGA].: Techniker Krankenkasse [Techniker health insurance fund]; 2024. URL: https://www.tk.de/techniker/krankheit-und-behandlungen/praxisbesuch-und-klinikaufenthalt/kompetent-als-patient/digitale-gesundheitsangebote-verstehen/digitale-gesundheitsanwendungen-antrag-stellen-2116854 [accessed 2025-08-20]
13. § 33a Digital Health Applications. Bundesamt für Justiz.: Statutory Health Insurance; 1988. URL: https://www.gesetze-im-internet.de/sgb_5/__33a.html [accessed 2026-01-19]
14. Giebel GD, Abels C, Börchers K, et al. Integration of digital health applications into the German healthcare system: development of “The DiGA-Care Path”. Front Health Serv 2024;4:1372522. [doi: 10.3389/frhs.2024.1372522] [Medline: 38545382]
15. Report by the National Association of Statutory Health Insurance Funds on the utilization and development of care with digital health applications (DiGA-Bericht). : National Association of Statutory Health Insurance Funds; 2023 URL: https://www.gkv-spitzenverband.de/media/dokumente/krankenversicherung_1/telematik/digitales/2023_DiGA_Bericht_GKV-Spitzenverband.pdf [accessed 2026-01-07]
16. Report by the National Association of Statutory Health Insurance Funds on the utilization and development of care with digital health applications (DiGA-Bericht). : National Association of Statutory Health Insurance Funds; 2024 URL: https://www.gkv-spitzenverband.de/media/dokumente/krankenversicherung_1/telematik/digitales/2024_DiGA-Bericht_final.pdf [accessed 2026-01-19]
17. Digitalisierung stößt bei Ärzteschaft auf große Skepsis [digitization met with great skepticism among doctors]. Deutsches Ärzteblatt [German Medical Journal]. 2022. URL: https://www.aerzteblatt.de/nachrichten/131027/Digitalisierung-stoesst-bei-Aerzteschaft-auf-grosse-Skepsis [accessed 2024-07-22]
18. KVB-Vorstand warnt vor Gesundheits-Apps [KVB board warns against health apps]. Ärzte Zeitung [The German Doctors Journal]. 2020. URL: https://www.aerztezeitung.de/Wirtschaft/KVB-Vorstand-warnt-vor-Gesundheits-Apps-413035.html [accessed 2024-07-22]
19. Amiri P, Nadri H, Bahaadinbeigy K. Facilitators and barriers of mHealth interventions during the Covid-19 pandemic: systematic review. BMC Health Serv Res 2023 Oct 28;23(1):1176. [doi: 10.1186/s12913-023-10171-w] [Medline: 37898755]
20. Stellungnahme des GKV-Spitzenverbandes vom 08.04.2021 zum Entwurf eines Gesetzes zur digitalen Modernisierung von Versorgung und Pflege [Statement by the GKV-spitzenverband (National Association of Statutory Health Insurance Funds) dated April 8, 2021, on the draft law on the digital modernization of health care and nursing care]. : National Association of Statutory Health Insurance Funds; 2021 URL: https://www.gkv-spitzenverband.de/media/dokumente/presse/p_stellungnahmen/2021_2023/210409_GKV-SV_SN_DVPMG.pdf [accessed 2026-01-07]
21. Geier AS. Digitale Gesundheitsanwendungen (DiGA) auf dem Weg zum Erfolg – die Perspektive des Spitzenverbandes Digitale Gesundheitsversorgung [Digital health applications (DiGA) on the road to success – the perspective of the German Digital Healthcare Association]. Bundesgesundheitsbl 2021 Oct;64(10):1228-1231. [doi: 10.1007/s00103-021-03419-5]
22. Boy K, May S, Labinsky H, et al. Back on track - digital health applications to treat back pain of rheumatic patients? Results of a qualitative interview study. Rheumatol Int 2024 Dec;44(12):2977-2987. [doi: 10.1007/s00296-024-05726-x] [Medline: 39340666]
23. Zugaj MR, Germeyer A, Kranz K, Züger A, Keßler J. Experiences of patients with endometriosis with a digital health application: a qualitative analysis. Arch Gynecol Obstet 2024 Oct;310(4):2253-2263. [doi: 10.1007/s00404-024-07651-7] [Medline: 39066952]
24. Schroeder T, Seaman K, Nguyen A, Gewald H, Georgiou A. Enablers and inhibitors to the adoption of mHealth apps by patients - a qualitative analysis of German doctors’ perspectives. Patient Educ Couns 2023 Sep;114:107865. [doi: 10.1016/j.pec.2023.107865] [Medline: 37356116]

XSL•FO
RenderX

25. Heidel A, Hagist C, Spinler S, Schoeneberger M. Removing dust from the German health care system by introducing health apps into standard care: semistructured interview study. JMIR Hum Factors 2023 May 4;10:e42186. [doi: 10.2196/42186] [Medline: 37140980]
26. Kao CK, Moawad A, Bhargava A. Chapter 4 - mobile health apps: current state, barriers, and future directions. In: The Digital Doctor: Academic Press; 2025. [doi: 10.1016/B978-0-443-15728-8.00019-7]
27. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
28. Giebel GD, Abels C, Plescher F, et al. Problems and barriers related to the use of mHealth apps from the perspective of patients: focus group and interview study. J Med Internet Res 2024 Apr 23;26:e49982. [doi: 10.2196/49982] [Medline: 38652508]
29. Helfferich C. Leitfaden- und Experteninterviews [Guideline and expert interviews]. In: Baur N, Blasius J, editors. Handbuch Methoden Der Empirischen Sozialforschung [Methods of Empirical Social Research] 2014. [doi: 10.1007/978-3-531-18939-0_39]
30. Claussen J, Jankowski D, Dawid F. Aufnehmen, Abtippen, Analysieren: Wegweiser zur Durchführung von Interview und Transkription [Recording, Typing, Analyzing: Guide to Conducting an Interview and Transcription]: Hannover; 2020.
31. Mayring P. Qualitative Inhaltsanalyse: Grundlagen und Techniken [Qualitative content analysis: basics and techniques]. In: Handbuch Methoden Der Empirischen Sozialforschung [Handbook of Methods in Empirical Social Research]: Beltz Publishing; 2015. [doi: 10.1007/978-3-658-21308-4_42]
32. Kuckartz U. Qualitative Inhaltsanalyse Methoden, Praxis, Umsetzung Mit Software Und Künstlicher Intelligenz [Qualitative Content Analysis: Methods, Practice, Computer Support]: Beltz Juventa; 2018.
33. MAXQDA. URL: https://www.maxqda.com/de/produkte/maxqda [accessed 2024-07-22]
34. Wissenswertes zu DiGA [Interesting facts about DiGA]. Bundesinstitut für Arzneimittel und Medizinprodukte [Federal Institute for Drugs and Medical Devices]. 2024. URL: https://www.bfarm.de/DE/Medizinprodukte/Aufgaben/DiGA-und-DiPA/DiGA/Wissenswertes/_node.html [accessed 2025-08-20]
35. Urbanek M. Kassen-Kritik and DiGA: Zentrale Anforderung ist medizinischer Nutzen [Health insurance criticism of DiGA: key requirement is medical benefit].: Ärzte Zeitung [The German Doctors Journal]; 2021. URL: https://www.aerztezeitung.de/Wirtschaft/Kassen-Kritik-an-DiGA-Zentrale-Anforderung-ist-medizinischer-Nutzen-416271.html [accessed 2024-07-22]
36. Regulation on the procedure and requirements for assessing the eligibility for reimbursement of digital health applications in statutory health insurance (Digital Health Applications Regulation - DiGAV). Bundesinstitut für Arzneimittel und Medizinprodukte [Federal Institute for Drugs and Medical Devices].: Federal Law Gazette I; 2020. URL: https://www.gesetze-im-internet.de/digav/BJNR076800020.html [accessed 2024-07-22]
37. Act to accelerate the digitization of the healthcare system (Digital Act – DigiG). Bundesinstitut für Arzneimittel und Medizinprodukte [Federal Institute for Drugs and Medical Devices].: Federal Law Gazette I; 2024. URL: https://www.recht.bund.de/bgbl/1/2024/101/VO.html [accessed 2025-01-06]
38. Kolominsky-Rabas PL, Tauscher M, Gerlach R, Perleth M, Dietzel N. Wie belastbar sind Studien der aktuell dauerhaft aufgenommenen digitalen Gesundheitsanwendungen (DiGA)? Methodische Qualität der Studien zum Nachweis positiver Versorgungseffekte von DiGA [How reliable are studies of currently approved digital health applications (DiGA)? Methodological quality of studies demonstrating positive healthcare effects of DiGA]. Z Evid Fortbild Qual Gesundhwes 2022 Dec;175:1-16. [doi: 10.1016/j.zefq.2022.09.008]
39. Bucksch M. Der DiGA-Fast-Track – bald auch europaweit? Zulassung von DiGA in der EU (2024): Überblick zum aktuellen Stand in allen Ländern der Europäischen Union [The DiGA fast track – soon to be available across Europe? Approval of DiGA in the EU (2024): overview of the current status in all countries of the European Union]. Quickbird Medical. 2023. URL: https://quickbirdmedical.com/diga-eu-europa-schweiz/ [accessed 2024-07-22]
40. Agreement according to § 134 Abs. 4 und 5 SGB V. National Association of Statutory Health Insurance Funds. 2023. URL: https://www.gkv-spitzenverband.de/media/dokumente/krankenversicherung_1/telematik/digitales/DiGA_Rahmenvereinbarung_nach_134_Abs_4_und_5_SGB_V_Stand_27.02.2024.pdf [accessed 2025-01-06]
41. Lau T. Gematik-Konzept: DiGA mit der E-Rezept-App verordnen [Gematik concept: prescribing DiGA with the e-prescription app].: Deutsches Ärzteblatt [German Medical Journal]; 2024 Aug. URL: https://www.aerzteblatt.de/archiv/240645/gematik-Konzept-DiGA-mit-der-E-Rezept-App-verordnen [accessed 2025-08-20]
42. Elektronisches Rezept (E-Rezept) [The electronic prescription]. Bundesministerium für Gesundheit [Federal Ministry of Health]. 2024. URL: https://www.bundesgesundheitsministerium.de/e-rezept [accessed 2025-08-20]
43. E-Rezept: Verbände ziehen überwiegend positive Bilanz [E-prescriptions: associations give a largely positive assessment].: Deutsches Ärzteblatt [German Medical Journal]; 2024. URL: https://www.aerzteblatt.de/nachrichten/152606/E-Rezept-Verbaende-ziehen-ueberwiegend-positive-Bilanz [accessed 2025-08-20]
44. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023 Sep 18;6(1):161. [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]

XSL•FO
RenderX

45. Kreyenbuhl J, Record EJ, Himelhoch S, et al. Development and feasibility testing of a smartphone intervention to improve adherence to antipsychotic medications. Clin Schizophr Relat Psychoses 2019;12(4):152-167. [doi: 10.3371/CSRP.KRRE.070816] [Medline: 27454213]
46. Nguyen AD, Frensham LJ, Baysari MT, Carland JE, Day RO. Patients' use of mobile health applications: what general practitioners think. Fam Pract 2019 Mar 20;36(2):214-218. [doi: 10.1093/fampra/cmy052] [Medline: 29873708]
47. Barmer Arztreport 2024 [Barmer physician report 2024]. Barmer. 2024. URL: https://www.barmer.de/resource/blob/1255894/d58c1bd7c1a99c78a54d410c490920c8/dl-arztreport-2024-data.pdf [accessed 2025-01-06]
48. Einheitlicher Bewertungsmaßstab (EBM). : National Association of Statutory Health Insurance Physicians; 2024 URL: https://www.kbv.de/documents/praxis/abrechnung/ebm/archiv/2024-4-ebm.pdf [accessed 2025-01-06]
49. Uncovska M, Freitag B, Meister S, Fehring L. Patient acceptance of prescribed and fully reimbursed mhealth apps in Germany: an UTAUT2-based online survey study. J Med Syst 2023 Jan 27;47(1):14. [doi: 10.1007/s10916-023-01910-x] [Medline: 36705853]
50. Machleid F, Jansky B, Wild V, Wiegmann C, Kaminski J, Schreiter S. Mobile Gesundheitstechnologien für eine gerechte Versorgung bei psychischen Erkrankungen [Mobile health technologies for equitable care in mental illness]. Nervenheilkunde [Ger J Neurol] 2024 Nov;43(12):688-700. [doi: 10.1055/a-2415-8433]
51. Peasgood T, Bourke M, Devlin N, Rowen D, Yang Y, Dalziel K. Randomised comparison of online interviews versus face-to-face interviews to value health states. Soc Sci Med 2023 Apr;323:115818. [doi: 10.1016/j.socscimed.2023.115818] [Medline: 36940582]
52. Guest G, Name E. Comparing Interview and Focus Group Data Collected in Person and Online [Internet]: Patient-Centered Outcomes Research Institute (PCORI); 2020. [doi: 10.25302/05.2020.ME.1403117064] [Medline: 36701499]

Abbreviations

BfArM: Bundesinstitut für Arzneimittel und Medizinprodukte (Federal Institute for Drugs and Medical Devices)
COREQ: Consolidated Criteria for Reporting Qualitative Studies
DiGA : Digitale Gesundheitsanwendung (digital health application)
GKV-SV: GK-Spitzenverband (National Association of Statutory Health Insurance Funds)
mHealth: mobile health
SHI: statutory health insurance

Edited by J Sarvestan; submitted 02.Mar.2025; peer-reviewed by M Guth, S Scheibe; revised version received 25.Nov.2025; accepted 25.Nov.2025; published 09.Feb.2026.

Please cite as:
Plescher F, Giebel GD, Abels C, Neusser S, Kampka B, Börchers K, Wasem J, Blase N
Problems and Barriers Regarding the Admission, Financing, and Service Provision of Digital Health Apps: Qualitative Stakeholder Survey
J Med Internet Res 2026;28:e73332
URL: https://www.jmir.org/2026/1/e73332
doi:10.2196/73332

© Felix Plescher, Godwin Denk Giebel, Carina Abels, Silke Neusser, Bonnie Kampka, Kirstin Börchers, Jürgen Wasem, Nikola Blase. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effects of Using a Smartphone App Combined With Behavior Change Techniques on the Level of Physical Activity Among Adults and Older Adults: Sequential Multiple Assignment Randomized Trial

Maria do Socorro Morais Pereira Simoes[1], PT, PhD; Neli Leite Proença[1], PT, PhD; Vinícius Tonon Lauria[1], BPE, PhD; Matheus Bibian do Nascimento[1], BPE; Ricardo da Costa Padovani[2], PsyD, PhD; Victor Zuniga Dourado[1,3], PT, BEng, PhD

[1]Departamento de Ciências do Movimento Humano, Instituto de Saúde e Sociedade, Universidade Federal de São Paulo, 136 Silva Jardim St, Room 338, Santos, Brazil

[2]Departamento de Saúde, Educação e Sociedade, Instituto de Saúde e Sociedade, Universidade Federal de São Paulo, Santos, Brazil

[3]Department of Global Health and Population, Bernard Lown Scholars in Cardiovascular Health Program, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, United States

Corresponding Author:
Victor Zuniga Dourado, PT, BEng, PhD
Departamento de Ciências do Movimento Humano, Instituto de Saúde e Sociedade, Universidade Federal de São Paulo, 136 Silva Jardim St, Room 338, Santos, Brazil

Abstract

Background: The use of tools, such as smartphone apps, to increase the level of physical activity (PA) decreases over time. Adaptive intervention trials have been recommended to test technology-based interventions owing to the possibility of adapting interventions based on individual responses.

Objective: This study aimed to investigate the effects of using a smartphone app combined with behavior change techniques on the PA level in adults and older adults (assessed using the step count). Moreover, the study investigated the time spent in sedentary behavior and time spent in moderate-to-vigorous PA (MVPA).

Methods: In this single-blinded, sequential multiple assignment randomized trial, participants were randomized into 3 groups during a 24-week intervention (group 1: app with tailored messages; group 2: app with tailored messages plus gamification I; and control group: educational information). In the sixth week, participants from groups 1 and 2 were classified as responders and nonresponders according to their average daily step count. Nonresponders were rerandomized among the other groups, adding a second type of gamification (group 3: app with tailored messages plus gamification II). After another 6 weeks, participants were reassessed and advised to keep monitoring their step count with the app, but without interference from the researchers. Face-to-face assessments were conducted. The behavior change techniques included app features (goal setting, auto-monitoring, ranking, and virtual badges) and researcher-provided resources (tailored messages and in-person sessions of PA). The intervention effects were analyzed using linear mixed models.

Results: The study included 53 participants (control group: n=17, group 1: n=17, group 2: n=19; mean age 44.0, SD 12.7 years). Groups 1 and 2 had 63% (10/16) and 47% (7/15) responders, respectively (P=.38). Regarding the PA level, participants from group 1 showed increases in the average daily step count at all assessments (final vs initial: B=797.2 steps/day, 95% CI 475.3-1119.1; P<.001; follow-up vs initial: B=2097.6 steps/day, 95% CI 1577.2-2618.1; P<.001). All participants showed a reduction in the time spent in sedentary behavior at the final assessment compared with the initial assessment (B=−70.8 min/week, 95% CI −88.8 to −52.9; P<.001), without differences among groups. The time spent in MVPA varied across time among all participants. Regardless of the initial group and allocation in the second randomization, responders from groups 1 and 2 showed a constant increase in the average daily step count (week 6 vs week 1: B=1548.0 steps/day, 95% CI 1407.4-1688.6; P<.001; week 12 vs week 1: B=1720.3 steps/day, 95% CI 1568.8-1871.7; P<.001; week 12 vs week 6: B=172.3, 95% CI 20.8-323.8; P=.03).

Conclusions: The adaptive intervention protocol using a smartphone app with behavior change techniques increased participants' PA levels. Stepping up behavior change techniques and progressively offering new stimuli may contribute to a change in behavior regarding PA.

Trial Registration: Brazilian Registry of Clinical Trials RBR-8xtc9c UTN: U1111-1218-1092; https://ensaiosclinicos.gov.br/rg/RBR-8xtc9c/

International Registered Report Identifier (IRRID): RR2-10.1186/s13063-019-3879-1

XSL•FO
RenderX

(J Med Internet Res 2026;28:e73388) doi:10.2196/73388

KEYWORDS

physical activity; adaptive clinical trial; transtheoretical model of behavior change; smartphone; app

Introduction

Despite the well-known benefits of physical activity (PA) in the prevention and treatment of several health conditions [1,2], a significant proportion of the global population does not meet its minimal recommended level [3]. The World Health Organization aims to reduce physical inactivity by 15% by 2030 [3]. Recent studies emphasize the complex interaction of individual [1,4,5], social [1,4,5], environmental [1,4,5], and political [4,5] factors influencing the practice of PA, suggesting that interventions should consider them [4,5]. Specifically, regarding individual and social factors, popular technologies, such as mobile devices and apps, have been used to increase the engagement of insufficiently active individuals.

Using smartphones as part of intervention programs can facilitate the auto-monitoring of PA, contributing to the behavior change process [6,7]. Duncan et al [6] compared auto-monitoring interventions for PA and eating habits using a website and printed material. They showed that both delivery methods improved the behavior, with no differences between groups. Evidence with moderate effects suggests that app-based interventions and pedometers may be effective [8]. Still, the association between app use and behavior change techniques may be more effective than app-based interventions alone [7].

Among the most commonly used behavior change techniques to promote PA are tailored messages, health education, gamification (the use of game elements in contexts other than games), and social support [7]. Regarding apps for PA and health, few are based on behavior change techniques. Among at least 25 behavior change techniques, only 1-8 techniques are offered by apps [9-12].

Despite the promising use of technologies, such as websites and smartphone apps, to deliver interventions targeted to increase the level of PA, studies show that the use of such devices decreases over time [6,7]. Interestingly, behavior change usually presents with the same pattern: individuals begin the program more engaged and active but do not maintain the new behavior for a long time [13]. That said, adaptive intervention trials, such as the sequential multiple assignment randomized trial (SMART), have been recommended to test technology-based interventions instead of classical clinical trials [14,15] due to the possibility of adapting interventions over time based on individuals' responses. It would be reasonable to extrapolate this rationale to interventions targeted to promote PA. Clinical trials with a SMART design have been shown to be feasible [16,17], but so far, this design has been barely used for PA interventions [17].

Our primary aim was to investigate the effects of using a smartphone app combined with behavior change techniques on the level of PA among adults and older adults, assessed by the average number of daily steps. Our secondary aim was to investigate the effects of the intervention on time spent in sedentary behavior and time spent in moderate-to-vigorous PA (MVPA), assessed by a triaxial accelerometer.

We hypothesized that the smartphone app combined with behavior change techniques offered in a SMART adaptive intervention design would be an effective tool to increase the level of PA among adults and older adults.

Methods

Study Design

This study had a single-blinded SMART design, in which the assessor was blinded regarding group allocations, with a 1:1 allocation ratio. The protocol was registered at the Brazilian Register of Clinical Trials (RBR-8xtc9c), and the study has been presented according to the CONSORT (Consolidated Standards of Reporting Trials) 2025 [18] and CONSORT-EHEALTH (V.1.6.1) [19] recommendations (Checklist 1). We enrolled participants continuously from November 2018 to February 2020. The study protocol was previously published in detail [20].

The primary outcome of the study was the average number of daily steps, and the secondary outcomes were the time spent in sedentary behavior and the time spent in MVPA, which were all assessed by a triaxial accelerometer.

Ethical Considerations

The project was approved by the Local Ethics Committee of the Federal University of São Paulo (CAAE 89112418.8.0000.5505), and all volunteers signed an informed consent form in person before participation. The informed consent form contained detailed information about the assessments and the group allocation, including a detailed description of the control and intervention groups. Participants did not receive any compensation before, during, or after the study. Only the principal researchers had access to participants' personal information. No identifiable information and/or images were or will be published.

Participants and Setting

We recruited participants through social media posts, distribution of printed material in different neighborhoods, and recommendations from other participants. The content of the recruitment material included the question "Do you want to increase your level of physical activity? We invite volunteers aged from 20 years, users of smartphones, that desire to move more. During 6 months, we will follow you to promote physical activity. Contact us!" or "Do you consider yourself a not very active person? Do you want to increase your level of physical activity? The EPIMOV Laboratory is looking for volunteers aged 20 years and older, able to walk without assistance from another person, and free of cardiac or pulmonary diseases. Contact us!" Those interested in participating were required to

XSL•FO
RenderX

call or text a number to schedule the assessment. Participants did not receive any type of incentive to cooperate in the study.

The inclusion criteria were as follows: (1) age 20 years or older; (2) absence of diagnosed cardiopulmonary, locomotor, or other conditions that could preclude the safe unsupervised performance of PA; and (3) having a smartphone and being familiar with its use, which was assessed by researcher observation and judgment during the enrollment phase.

The exclusion criteria were as follows: (1) basal level of PA of ≥10,000 average steps per day; (2) self-reported recent respiratory infection; (3) abnormalities on the cardiorespiratory fitness test that preclude the safe performance of unsupervised PA; and (4) refusal to participate by signing the informed consent. Although Tudor-Lock et al [21] suggested that interventions to increase the level of PA should mainly target individuals who perform less than 5000 steps per day, we decided not to restrict the sample to sedentary individuals and set the limit of 10,000 steps per day, which indicates a considerable amount of PA. This criterion was assessed using a triaxial accelerometer, worn by participants for 7 consecutive days. The procedure is described in detail in the Description of the Assessment Procedure section.

The assessments and in-person meetings for decision-making regarding the protocol took place at the Epidemiology and Human Movement (EPIMOV) Laboratory at the Federal University of São Paulo in Santos, Brazil. The target population was the residents in the metropolitan area of Santos, São Paulo, Brazil. The city of Santos is mainly flat and offers a gardened beachfront with an extension of approximately 7 km [22]. Cycle lanes and public equipment widely cover the city to practice PA. It is common to watch people exercising in different spaces of the city. Moreover, it is ranked 5th in Brazil in terms of quality of life [22].

Randomization Procedures

The randomization sequence (initial randomization and rerandomization) was generated in blocks of 6 participants, using a website [23], and was sealed in opaque envelopes numbered in the opening sequence. The envelopes were prepared by a third person who was not involved in any other study phase. The first envelope was opened at the time of enrollment after participants watched the animated video about the benefits of practicing PA. The second envelope, containing the new group allocation, was kept sealed until the sixth week of the intervention. We opened the second envelope only among nonresponder participants in their presence. After the onset of the COVID-19 pandemic and the suspension of in-person activities at the university, a research team member opened the envelopes and informed the nonresponder participants about the new group intervention, using the same text messaging app.

Description of Interventions

The adaptive intervention protocol was designed to present different stimuli in a stepwise manner, which could contribute to optimizing the cost-effectiveness of interventions delivered in health services. Based on the ecological model proposed by Bauman et al [4] and Sallis et al [5], the intervention was initially focused on individual (demographic and psychological) and social (social support) factors. By the end of the SMART protocol, it would be possible to analyze if there is an ideal sequence of stimuli to deliver, and thus, strategies in public health could be planned in a more cost-effective sequence, starting with those less demanding in financial and human resources and progressing to those more demanding in financial and human resources.

The intervention lasted a total of 24 weeks and was delivered in 2 phases. Initially, we randomized participants into 3 groups (group 1: app plus tailored messages; group 2: app plus tailored messages plus gamification; and control group: advice) (Figure 1). The first intervention phase lasted 12 weeks, during which the average number of participants' daily steps was registered weekly in groups 1 and 2. By the sixth week of the intervention, we assessed participants' performance in terms of daily steps and classified them into responders and nonresponders. We then rerandomized nonresponders, adding a third group (group 3: app plus tailored messages plus gamification II) (Figure 1). After completing 12 weeks, we advised participants to continue using the app and monitor daily steps for over 12 weeks. From that moment on, they did not receive any direct intervention from the researchers. We invited participants from all groups for reassessments at weeks 12 and 24.

XSL•FO
RenderX

Figure 1. Flowchart of the study protocol. R: randomization.

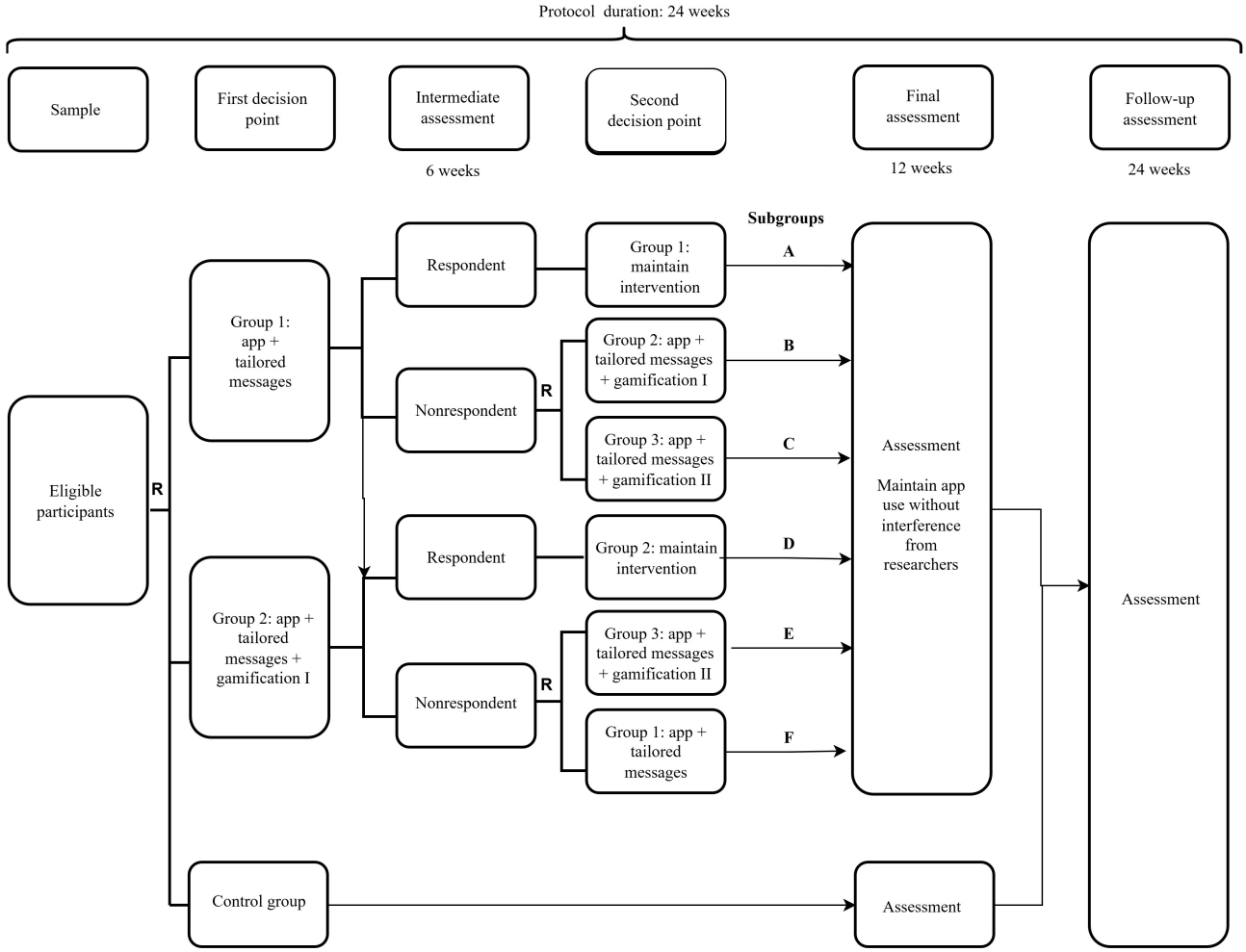

Initially, all participants watched an animated video about the benefits of practicing PA and received a printed booklet with the same content as the video. Moreover, we sent them a link so that they could watch the video again as often as they wanted. After this procedure, we opened the first randomization envelope containing the initial group assignment and explained the detailed intervention according to the group allocation.

A detailed description of the intervention has been published elsewhere [20]. Briefly, the approaches used in each group were as follows:

- Group 1: Participants used the app to self-monitor their daily steps. We sent tailored messages weekly using a text messaging app, according to their performance in the last week, and motivated them to increase their step count during the following week.
- Group 2: In addition to the self-monitoring and tailored messages, we instructed participants on how to access the gamification features the app offers, such as ranking, virtual challenges, virtual badges, and social media interaction.
- Control group: We advised participants to increase their PA levels, based on the information provided by the animated video and printed booklet. We did not inform them about using apps. Moreover, since these participants would not be rerandomized, we did not invite them to the intermediate assessment at week 6.
- Group 3: We included the second gamification phase at the intermediate assessment (week 6). Gamification II offered the same features as gamification I plus opportunities for socialization among participants and the research team. We planned in-person meetings every 2 weeks to practice PA in groups and created a group in a text messaging app to facilitate interaction.

Before beginning the study, researchers from the EPIMOV team used smartphone apps to track PA. After a period of use and comparing the features, easiness, and difficulties of the apps, they decided that Pacer (Pacer Health Inc) was the most suitable to use in the trial. It registers the step count and offers behavior change techniques (self-monitoring, goal setting, progress bars, competition, ranking, virtual badges, and social interaction). In addition, we tested the feasibility of the SMART design using Pacer [17].

We used the free version of the Pacer app in this study, providing the behavior change techniques according to group allocation. The participants downloaded the app on their phones with the aid of researchers. In addition to in-person instructions on using the app, we provided a printed booklet with the same instructions. In addition, at any time, participants could contact researchers to ask for help regarding the use of the app. There were no changes to the protocol after the trial commenced, including no changes in the version of the app.

XSL•FO
RenderX

The content of the weekly tailored text messages was developed by one of the researchers supervised by a sports psychologist, according to the stage of behavior change, and is available elsewhere [17]. We created a sequence of messages for precontemplation and contemplation, and another sequence for participants in preparation, action, and maintenance. For each sequence, we had messages for occasions when participants achieved or overcame their goals and messages for occasions when they did not reach their goals. Other than the tailored messages, participants did not receive reminders or prompts to use the app.

The opportunities for socialization offered to participants reallocated to group 3 occurred biweekly as in-person meetings. During these meetings, we practiced PA in groups (participants and researchers), which was planned and conducted by an experienced professional in physical education. All meetings took place on Saturdays at the beachfront, and at the end of the PA sessions, researchers remained available to interact with participants. The sessions lasted, on average, 1 hour and 30 minutes (exercise and postexercise interaction) and included a warm-up period (around 5 minutes) and planned walking or running exercises targeted (each session) at step count, activity duration, or pace. In addition, participants had the option to join a group in the text messaging app so that they could interact with each other and plan the PA meetings.

Goal Setting and Intermediate Assessment

We set individualized goals for each participant. We instructed them to maintain their routines during the first week of app use. Then, we registered the average number of daily steps collected by the app, and this step count was considered as the participants' goal. The goal-setting process adopted in this study has been published in another study [20]. Finally, we sent a text message informing participants that their goal was to overcome the initial step count from that day on.

We used the average number of daily steps collected by the app to determine responders and nonresponders at the intermediate assessment. We registered the weekly performance of participants and plotted it in individualized sheets. Then, we used linear regression lines along the performance data during the first 6 weeks of the intervention. Finally, responders were those participants with a positive slope on the regression line, while nonresponders were those with a zero or negative slope on the regression line. We set this assessment to consider the individualized performance of participants over time, attempting to set feasible goals for our sample. If we faced errors when extracting the step count from the app interface, such as a lack of performance updates or operational errors, participants remained in the same group for initial allocation.

After the intermediate assessment, we invited participants for an in-person meeting at the laboratory. In this meeting, we explained their performance during the first 6 weeks of the intervention, and nonresponders were rerandomized (we opened the second sealed envelope that remained stored since the first randomization). Moreover, we provided information about the different components of the intervention (Figure 1). After the onset of the COVID-19 pandemic and the restrictive measures, we conducted the intermediate assessment remotely (one of the researchers opened the second envelope and communicated with participants via text messages).

Study Measurements

We assessed participants at the beginning of the protocol (initial assessment) and reassessed them at weeks 12 (final assessment) and 24 (follow-up assessment). All of the assessors were blinded regarding the group allocation. The assessment protocol included demographic and socioeconomic information, general health condition, stages of behavior change for PA, level of PA, and anthropometric measures.

The assessors completed the protocol on 2 different days, 7 days apart. On the first day, participants collected their demographic and socioeconomic information, stages of behavior change for PA, and general health status. At the end of day 1, they received an accelerometer to wear during the subsequent 7 days when they returned to complete the assessments. On day 2, participants returned the accelerometer, and we assessed their level of PA and anthropometric measures.

By week 12 (the end of the intervention protocol), we reassessed participants and advised them to keep using the app to track their daily steps, but from this moment on, they did not receive tailored messages or any other direct intervention from the researchers. Finally, at week 24, we invited participants to return for the follow-up assessment, irrespective of whether they had presented for the 12-week evaluation.

At each assessment point, we presented the results of the previous evaluation to participants. At week 12, they received the results from the initial assessment, and at week 24, they received the results from the final evaluation. After the study ended at week 24, the results were provided according to their wish and availability. There were no consequences for participants who did not withdraw their results. The results were stored at a safe location and were available to participants at any time.

Description of the Assessment Procedure

The tests and interviews were standardized and conducted by trained personnel. Moreover, the equipment was periodically checked and calibrated according to the manufacturer's instructions. We invited all participants to the follow-up assessment regardless of whether they had presented for the final evaluation.

We collected the following demographic and socioeconomic information: age (years), sex, marital status, education, and occupational status. We used the *Critério Brasil* questionnaire to assess socioeconomic status [24]. This questionnaire contains 15 questions, divided into 3 categories: household items, educational level of the family's chief, and access to services. According to the total score, participants were classified into 6 socioeconomic strata, from A to D, where A represents the highest socioeconomic status and D represents the lowest socioeconomic status.

For assessing general health conditions, we registered participants' self-reported personal and family health precedents and medication use. Moreover, we registered the following self-reported cardiovascular risk factors: age 45 years or older

XSL•FO
RenderX

for men and 55 years or older for women, family history of cardiovascular diseases, diagnosis of arterial blood hypertension, diagnosis of hyperglycemia or diabetes, diagnosis of hypercholesterolemia or dyslipidemia, and smoking status. We complemented the assessment of cardiovascular risk factors with the evaluation of obesity and physical inactivity.

For evaluating the stages of behavior change for PA, we used the Brazilian version of the transtheoretical model [25], which considers that individuals may, according to the circumstances, transit nonlinearly among different behavior change stages [25]. According to the status of the current practice of PA, the willingness to change the behavior, and the fact that PA is essential to health, participants were classified into one of the following stages: precontemplation, contemplation, preparation, and maintenance [25].

For assessing the level of PA, participants received a triaxial accelerometer (Actigraph) at the end of the first day of assessments, which was required to be used for 7 consecutive days and worn over the hip according to the self-reported dominant side of the body. Data were considered valid when available for at least 4 days, including 1 weekend day, and for at least 10 hours per day [26]. When a participant presented with invalid data for the first time, we returned the accelerometer and asked them to use it for more than 7 days. We extracted accelerometer data using the manufacturer's software and registered the following information: average number of steps per day, time spent in sedentary behavior (0 to 99 step counts/min) [27], and time spent in MVPA (≥1952 step counts/min) [27].

With regard to anthropometric measures, we registered the height (m) and body mass (kg) of participants using a digital scale with a stadiometer (Toledo). We positioned participants standing barefoot over the equipment and asked them to keep looking forward during the measurement. For height measurement, participants kept their arms crossed over the trunk while sustaining maximal inspiration, and for body mass measurement, they kept their arms along the body.

Sample Size Calculation

Considering a fixed *k* probability, at least *m* participants would be allocated to subgroups B and E, as demonstrated in Figure 1. Moreover, based on a nonresponse rate *q* of 35% - 65% in SMART studies, we estimated that 3 participants per group would be sufficient to ensure familiarity with the protocol, allow identification, and deal with potential difficulties that could occur during the intervention period. Thus, for a probability *k* of 90% and a nonresponse rate *q* of 50% by the end of the sixth week of the protocol, the study should include at least 42 participants [28]. Due to the lack of a similar study to use data for estimating the sample size, after 42 participants completed the protocol, we planned to perform a new sample size calculation and continue the study until the sample size target was achieved. However, this approach was not possible because the protocol had to be interrupted after the onset of the COVID-19 pandemic. Therefore, we completed the study after including 53 participants.

Data Analysis

Participants' characteristics are presented using descriptive statistics, such as mean and SD for continuous variables, and absolute number and proportion for categorical variables. We compared the characteristics of participants among groups using 1-way ANOVA for continuous variables and the chi-square test for categorical variables.

We checked for participant performance in groups 1 and 2, using graphs with linear regression lines. For each group and participant, we plotted 3 graphs using the average number of daily steps: 1 with data from the 12 weeks of the intervention, 1 with data from the first 6 weeks (between the first and intermediate assessments), and 1 with data from the last 6 weeks (between the intermediate and final assessments).

To analyze the effects of the intervention on primary and secondary outcomes, we used repeated measures linear mixed models, considering the initial group allocation and time as fixed effects and the individual as a random effect. Moreover, we grouped participants from groups 1 and 2 according to their response to the intervention (responders and nonresponders) and compared their performance using repeated measures linear mixed models, considering response to the intervention and time as fixed effects and the individual as a random effect.

We managed missing data using intention-to-treat analysis by multivariate data imputation, adjusted by sex and age. The graphs for weekly performance were generated using Microsoft Excel (Microsoft Corporation), and data analysis was conducted using Stata version 14 (StataCorp LLC).

Results

Of 62 potential individuals who volunteered to participate, 53 met the inclusion criteria and were included in the protocol. Some participants were lost to follow-up, especially at the 24th-week assessment, mainly due to the COVID-19 pandemic. For the same reason, we had to interrupt the protocol and could not achieve a larger sample size. As shown in Figure 2, according to the intention-to-treat analysis, we included all participants in the data analysis.

XSL•FO
RenderX

Figure 2. Flowchart of participants in the study.

XSL•FO
RenderX

As we suddenly had to interrupt the inclusion of new participants, there were differences in the number of participants between group 2 and the other groups. Furthermore, our plan to conduct a second trial phase, using intermediate data to estimate sample size, could not be executed.

Protocol deviations were continuously checked, such as accessing app features that were not advised. Only 1 participant from group 1 explored the app and joined its gamification features, such as ranking and challenges of walking distance. There were no harmful events or unintended effects during the study. According to the taxonomy proposed by Michie et al [29], the behavior change techniques in our study included rewards (virtual badges provided by the app), incentives (tailored messages), graded tasks (overcoming the goal), feedback and monitoring (self-monitoring using the app and tailored messages), goals and planning (individual goal setting), social support (practical and general; provided by the app via teaming up with other participants, and for group 3, provided via in-person meetings and text messaging), comparison of behavior (social comparison provided by the ranking in the app), and instructions on how to perform a behavior (provided by the educational material).

In general, participants were middle-aged adults (mean age 44.0, SD 12.7 years), were mainly female (30/53, 57%), and were overweight or obese (mean BMI 29.8, SD 6.5 kg/m^2). Regarding the stage of behavior change for PA, almost half of the sample (24/53, 45%) was in the preparation stage. As shown in Table 1, there were no differences in participant characteristics, except for sex (P=.03) and the stage of behavior change (P=.047).

XSL•FO
RenderX

Table . Characteristics of participants at baseline.

Variable	All participants (n=53)	Control group (n=17)	Group 1 (n=17)	Group 2 (n=19)	*P* value[a]
Age (years), mean (SD)	44.0 (12.7)	45.2 (13.4)	47.7 (11.2)	39.6 (12.5)	.15
Sex (female), n (%)	30 (57)	9 (53)	6 (35)[b]	15 (79)	.03
Body mass (kg), mean (SD)	83.3 (21.9)	89.1 (28.8)	85.1 (14.7)	76.5 (19.3)	.21
Stature (m), mean (SD)	1.66 (0.08)	1.67 (0.08)	1.68 (0.09)	1.65 (0.07)	.49
BMI (kg/m^2), mean (SD)	29.8 (6.5)	31.5 (8.5)	30.2 (4.7)	28.0 (5.8)	.26
Race, n (%)					.15
White	33 (62)	12 (71)	7 (41)	14 (74)	
Black	6 (11)	0 (0)	4 (24)	2 (11)	
Brown	12 (23)	4 (24)	6 (35)	2 (11)	
Do not know/not reported	2 (4)	1 (6)	0 (0)	1 (5)	
Education, n (%)					.32
Middle school, incomplete	2 (4)	0 (0)	1 (6)	1 (5)	
Middle school, complete	2 (4)	1 (6)	1 (6)	0 (0)	
High school, complete	28 (53)	13 (77)	7 (41)	8 (42)	
Higher education, complete	13 (25)	3 (18)	4 (24)	6 (32)	
Postgraduation, complete	8 (15)	0 (0)	4 (24)	4 (21)	
Employed (yes), n (%)	39 (74)	11 (65)	14 (82)	14 (74)	.62
Socioeconomic classification[c], n (%)					.20
A	7 (13)	2 (12)	0 (0)	5 (26)	
B1	7 (13)	1 (6)	4 (24)	2 (11)	
B2	18 (34)	5 (29)	6 (35)	7 (37)	
C1	9 (17)	6 (35)	2 (12)	1 (5)	
C2	10 (19)	3 (18)	4 (24)	3 (16)	
D	2 (4)	0 (0)	1 (6)	1 (5)	
Stage of behavior change, n (%)					.047
Contemplation	1 (2)	0 (0)	0 (0)	1 (5)	
Preparation	24 (45)	6 (35)	9 (53)	9 (47)[d]	
Action	13 (25)	4 (24)	1 (6)[b]	8 (42)[d]	
Maintenance	15 (28)	7 (41)	7 (41)[b]	1 (5)[d]	

[a]*P* values refer to comparisons among groups.

[b]Significant difference between group 1 and group 2.

[c]For socioeconomic classification, A represents the highest socioeconomic status and D represents the lowest socioeconomic status.

[d]Significant difference between the control group and group 2.

All groups were similar when considering the outcome measures. The nonresponse rate was 53% (28/53), without differences in the proportion of respondents at week 6 (*P*=.38) (Table 2).

XSL•FO
RenderX

Table . Baseline assessment and comparison among groups.

Variable	All participants (N=53)	Control group (n=17)	Group 1 (n=17)	Group 2 (n=19)	*P* value[a]
Average daily step count, mean (SD)	6134 (1932)	6616 (2069)	5651 (1810)	6170 (1942)	.44
Average time spent in moderate-to-vigorous physical activity (min/day), mean (SD)	33 (16)	39 (14)	27 (18)	33 (13)	.17
Average time spent in sedentary behavior (min/day), mean (SD)	671 (130)	669 (77)	678 (154)	664 (150)	.96
Respondent at week 6 (yes), n (%)	—[b]	—	10[c] (63)	7[d] (47)	.38
Group allocation after rerandomization, n (%)					.003
Group 1	—	0 (0)	14 (82)	5 (26)	
Group 2	—	0 (0)	3 (18)	10 (53)	
Group 3	—	0 (0)	0 (0)	4 (21)	

[a]*P* values refer to comparisons among groups.

[b]Not applicable.

[c]Out of a total of 16.

[d]Out of a total of 15.

Table 3 presents the outcome measures in each group at each assessment (initial, final, and follow-up). We managed missing data using multivariate multiple imputation adjusted for sex and age.

Table . Effects of the intervention within and between groups.

Variable	Control group[a]	Group 1[a]	Group 2[a]
Average daily step count, mean (SD)			
Initial (baseline)	6606 (1989)[b]	5643 (1740)[b,c]	6180 (1887)[c]
Final (week 12)	7063 (2215)	7617 (1873)[d]	7524 (2256)[d]
Follow-up (week 24)	6271 (1574)	8480 (2387)	6350 (2413)
Average time spent in moderate-to-vigorous physical activity (min/day), mean (SD)			
Initial (baseline)	39 (14)[c]	27 (18)[b,e]	33 (13)
Final (week 12)	36 (20)[d]	35 (21)[d]	36 (14)
Follow-up (week 24)	41 (19)	44 (15)	33 (17)
Average time spent in sedentary behavior (min/day), mean (SD)			
Initial (baseline)	670 (74)[c]	678 (148)[c]	663 (144)[c]
Final (week 12)	569 (129)[d]	644 (137)[d]	613 (102)[d]
Follow-up (week 24)	694 (137)	677 (117)	663 (132)

[a]Missing data were treated using multivariate multiple imputation adjusted by sex and age.

[b]Significant difference within groups (initial vs follow-up assessment).

[c]Significant difference within groups (initial vs final assessment).

[d]Significant difference within groups (final vs follow-up assessment).

[e]Significant difference between groups (control group vs group 1).

For the primary outcome (the average count of daily steps), we observed differences in the factor *time* and at the individual level (Table 4). While participants from the control group showed a reduction in the step count over time (follow-up vs initial assessment: β=−575.8; *P*<.05) and participants from group 2 showed an increase in the step count between the initial and final assessments (β=897.4; *P*<.001), participants from group 1 showed an increase in the step count for all assessments (final vs initial assessment: β=650.2; *P*<.001; follow-up vs initial assessment: β=1521.9; *P*<.001).

XSL•FO
RenderX

Table . Average daily step count considering group, time, and group and time.

Variable[a]	Coefficient, value (95% CI)	*P* value
Group (Ref[b]: control group)		
Group 1	−895.2 (−2179.3 to 388.8)	.17
Group 2	−903.2 (−2158.7 to 352.2)	.16
Time (Ref: I[c])		
Fi[d]	−147.0 (−380.2 to 86.1)	.22
Fo[e]	−575.8 (−1045.3 to −106.2)	.02
Group 1 and time (Ref: control group and I)		
Fi	797.2 (475.3 to 1119.1)	<.001
Fo	2097.6 (1577.2 to 2618.1)	<.001
Group 2 and time (Ref: control group and I)		
Fi	1044.4 (747.0 to 1341.8)	<.001
Fo	982.8 (332.9 to 1632.6)	<.001
Contrasts, control group		
Fi and I	−147.0 (−380.2 to 86.1)	.22
Fo and I	−575.8 (−1045.3 to −106.2)	.02
Fo and Fi	−428.7 (−920.3 to 63.1)	.09
Contrasts, group 1		
Fi and I	650.2 (428.2 to 872.1)	<.001
Fo and I	1521.9 (1297.4 to 1746.4)	<.001
Fo and Fi	871.7 (640.4 to 1103.0)	<.001
Contrasts, group 2		
Fi and I	897.4 (712.8 to 1082.0)	<.001
Fo and I	407.0 (−42.2 to 856.2)	.08
Fo and Fi	−490.3 (−949.3 to −31.4)	.04

[a]Contrasts within groups over time (linear mixed model analysis). Fixed effects: initial group allocation and time of assessment; random effect: participant.
[b]Ref: reference.
[c]I: initial assessment.
[d]Fi: final assessment.
[e]Fo: follow-up assessment.

Results from the analysis of the time spent in sedentary behavior are presented in Table 5. We observed differences in the factor *time* in all groups (final vs initial assessment: β=−70.8; *P*<.001) and at the individual level (*P*<.001), without differences among groups.

XSL•FO
RenderX

Table . Time spent in sedentary behavior considering group, time, and group and time.

Variable[a]	Coefficient, value (95% CI)	*P* value
Group (Ref[b]: control group)		
Group 1	−6.9 (−81.9 to 68.1)	.86
Group 2	−20.3 (−93.8 to 53.3)	.59
Time (Ref: I[c])		
Fi[d]	−70.8 (−88.8 to −52.9)	<.001
Fo[e]	19.7 (−16.4 to 55.9)	.29
Group 1 and time (Ref: control group and I)		
Fi	26.0 (1.2 to 50.7)	.04
Fo	−23.8 (−63.9 to 16.2)	.24
Group 2 and time (Ref: control group and I)		
Fi	41.9 (19.0 to 64.8)	<.001
Fo	−10.3 (−60.3 to 39.7)	.69
Contrasts, control group and time		
Fi and I	−70.8 (−88.8 to −52.9)	<.001
Fo and I	19.7 (−16.4 to 55.9)	.29
Fo and Fi	90.5 (−88.8 to −52.9)	<.001
Contrasts, group 1 and time		
Fi and I	−44.9 (−61.9 to −27.8)	<.001
Fo and I	−4.1 (−21.4 to 13.2)	.64
Fo and Fi		
Contrasts, group 2 and time		
Fi and I	−29.0 (−43.2 to −14.7)	<.001
Fo and I	9.4 (−25.2 to 43.9)	.59
Fo and Fi	38.3 (3.04 to 73.6)	.03

[a]Contrasts within groups over time (linear mixed model analysis). Fixed effects: initial group allocation and time of assessment; random effect: participant.
[b]Ref: reference.
[c]I: initial assessment.
[d]Fi: final assessment.
[e]Fo: follow-up assessment.

We also observed differences at the individual level for the time spent in MVPA (Table 6). For the factors *group* and *time*, participants from the control group showed a reduction in the time spent in MVPA from the initial assessment to the final assessment (β=−7.9; *P*<.001) and then showed an increase in the time by the follow-up assessment (β=9.0; *P*<.001). On the other hand, participants from group 1 showed an increase in the time spent in MVPA at the follow-up assessment (β=7.8; *P*<.001).

XSL•FO
RenderX

Table . Time spent in moderate-to-vigorous physical activity considering group, time, and group and time.

Variable[a]	Coefficient, value (95% CI)	*P* value
Group (Ref[b]: control group)		
Group 1	−13.0 (−22.4 to −3.5)	.007
Group 2	−5.5 (−14.8 to 3.8)	.24
Time (Ref: I[c])		
Fi[d]	−7.9 (−9.9 to −5.9)	<.001
Fo[e]	1.1 (−2.9 to 5.2)	.59
Group 1 and time (Ref: control group and I)		
Fi	9.2 (6.4 to 11.9)	<.001
Fo	7.9 (3.4 to 12.4)	.001
Group 2 and time (Ref: control group and I)		
Fi	8.8 (6.2 to 11.4)	<.001
Fo	−2.6 (−8.2 to 3.0)	.37
Contrasts, control group and time		
Fi and I	−7.9 (−9.9 to −5.9)	<.001
Fo and I	1.1 (−2.9 to 5.2)	.59
Fo and Fi	9.0 (4.8 to 13.3)	<.001
Contrasts, group 1 and time		
Fi and I	1.2 (−0.7 to 3.2)	.20
Fo and I	9.1 (7.1 to 11.0)	<.001
Fo and Fi	7.8 (5.8 to 9.8)	<.001
Contrasts, group 2 and time		
Fi and I	0.9 (−0.7 to 2.5)	.27
Fo and I	−1.5 (−5.4 to 2.4)	.46
Fo and Fi	−2.4 (−6.3 to 1.6)	.25

[a]Contrasts within groups over time (linear mixed model analysis). Fixed effects: initial group allocation and time of assessment; random effect: participant.
[b]Ref: reference.
[c]I: initial assessment.
[d]Fi: final assessment.
[e]Fo: follow-up assessment.

Figures 3 and 4 present the performance of participants from groups 1 and 2 during the 12 weeks of the protocol and from weeks 1 to 6 and weeks 7 to 12 (before and after rerandomization, respectively).

XSL•FO
RenderX

Figure 3. Weekly performance of participants from group 1. (A) Performance from week 1 to week 12; (B) performance from week 1 to week 6 (before rerandomization); (C) performance from week 7 to week 12 (after rerandomization).

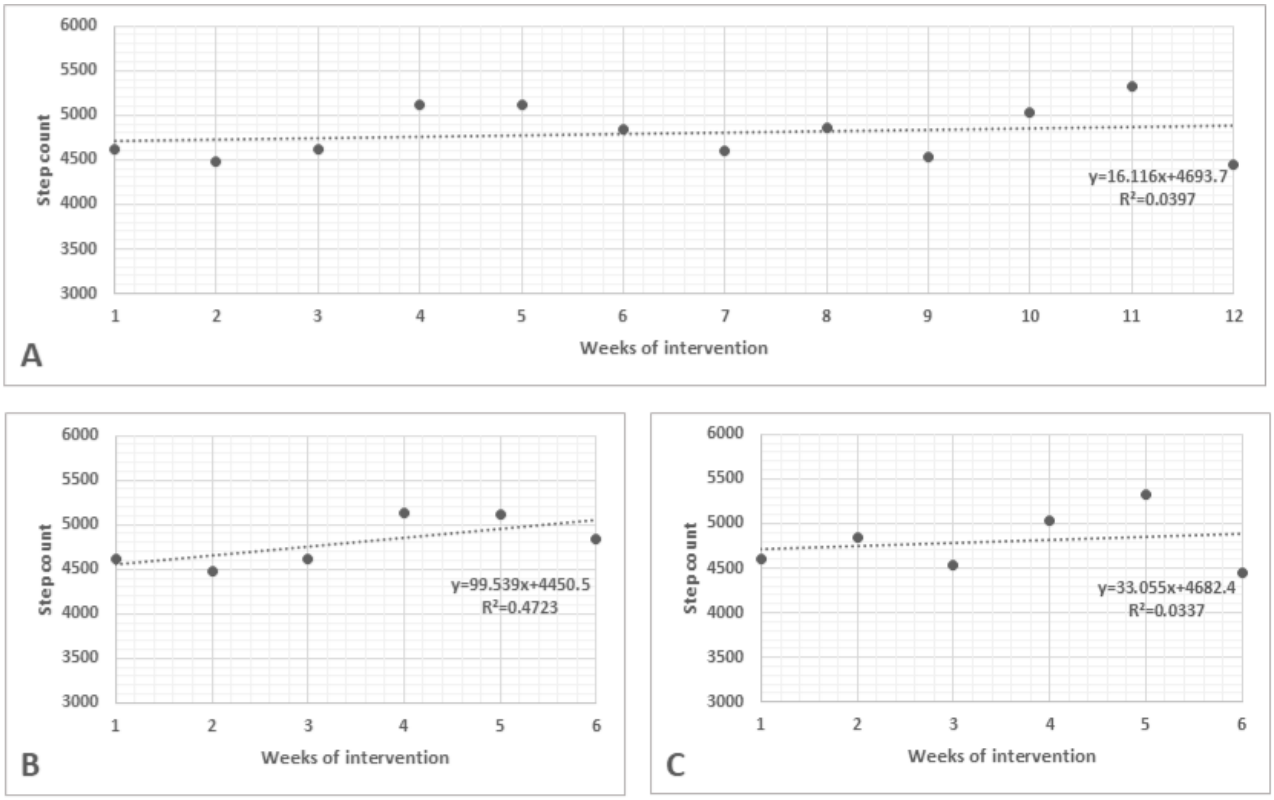

Figure 4. Weekly performance of participants from group 2. (A) Performance from week 1 to week 12; (B) performance from week 1 to week 6 (before rerandomization); (C) performance from week 7 to week 12 (after rerandomization).

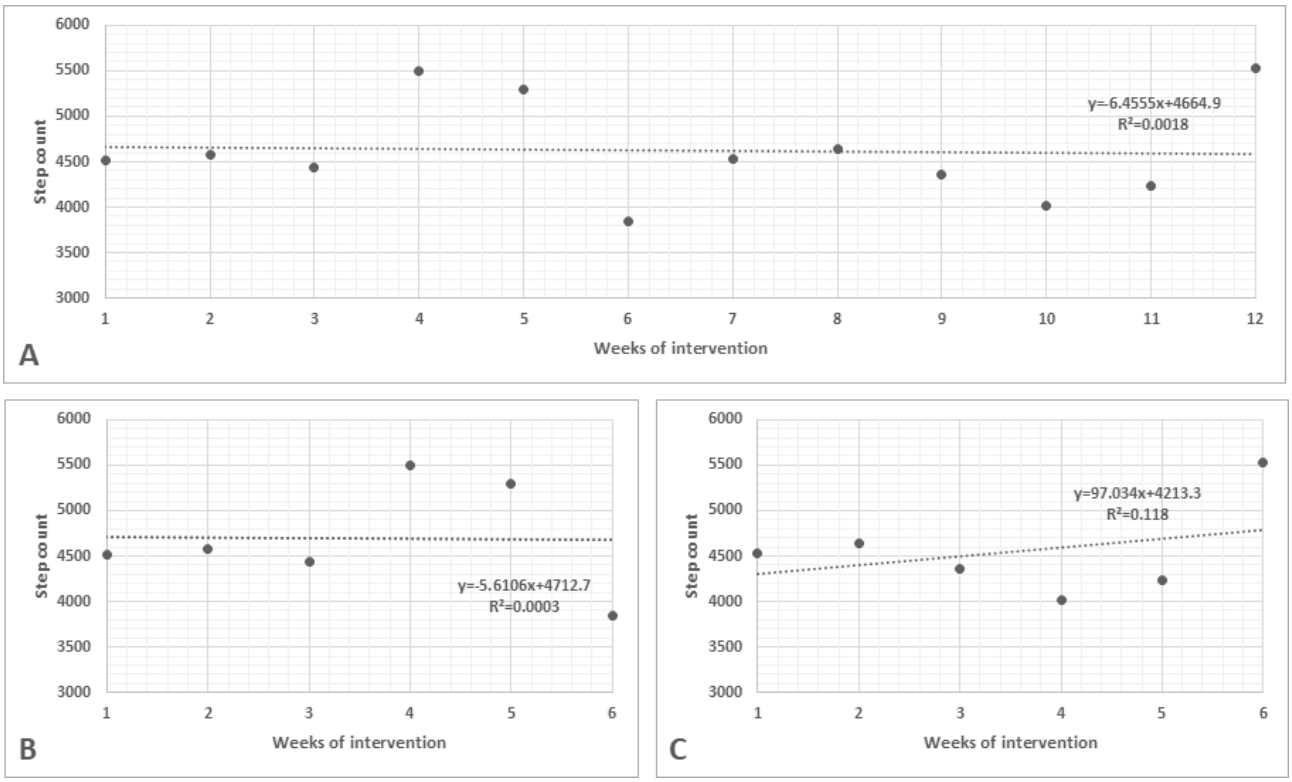

In Table 7, we present the analysis of the average daily step count of responders and nonresponders, regardless of group allocation. Using the Pacer app, we collected these data in weeks 1, 6, and 12. We observed differences at the individual level, and although all participants showed reductions in the average daily step count over time (week 6 vs week 1: β=−496.7; P<.001; week 12 vs week 1: β=−579.3; P<.001), among

XSL•FO
RenderX

responders, the step count constantly increased (week 6 vs week 1: β=2044.7; *P*<.001; week 12 vs week 1: β=2299.6; *P*<.001).

Table . Pacer app's average daily step count from responders and nonresponders, regardless of group allocation, according to the response to the intervention and time.

Variable[a]	Coefficient, value (95% CI)	*P* value
Response to the intervention (Ref[b]: no)		
Yes	−1588.8 (−3252.4 to 74.9)	.06
Time (Ref: week 1)		
Week 6	−496.7 (−651.6 to −341.8)	<.001
Week 12	−579.3 (−743.6 to −414.9)	<.001
Response and time (Ref: no and week 1)		
Yes and week 6	2044.7 (1835.5 to 2253.9)	<.001
Yes and week 12	2299.6 (2076.1 to 2523.0)	<.001
Contrasts, nonresponders and time		
Week 6 and week 1	−496.7 (−651.6 to −341.8)	<.001
Week 12 and week 1	−579.3 (−743.6 to −414.9)	<.001
Week 12 and week 6	−82.6 (−246.9 to 81.7)	.33
Contrasts, responders and time		
Week 6 and week 1	1548.0 (1407.4 to 1688.6)	<.001
Week 12 and week 1	1720.3 (1568.8 to 1871.7)	<.001
Week 12 and week 6	172.3 (20.8 to 323.8)	.03

[a]Contrasts between response and time (linear mixed model analysis). Fixed effects: response to the intervention and time of assessment; random effect: participant.

[b]Ref: reference.

Discussion

This study aimed to investigate the effects of using a smartphone app combined with behavior change techniques on the level of PA among adults and older adults. Due to the sudden interruption of the protocol after the onset of the COVID-19 pandemic, we completed the study with a small sample size. Overall, all participants presented with improvements in outcomes over time. Specifically for PA, all participants from group 1 showed a consistent increase in the average daily step count. According to the benchmarks for assessing the effectiveness of interventions to promote PA proposed by Wright et al [30], considering the digital intervention setting, participants from the control group presented a performance below the 25th percentile for step count, while participants from groups 1 and 2 performed greater than the 75th percentile and between the 50th and 75th percentiles, respectively.

Only participants from group 1 continued to increase the step count even after the end of the intervention. Despite the nonstatistical difference in the proportion of respondents and nonrespondents between groups, which may have occurred due to the small sample size and consequently low statistical power, it is possible that the stepwise offering of intervention components (initiation with less complex strategies and gradual increases according to participants' responses) may be more effective in supporting behavior change than offering more complicated strategies. Furthermore, gamification may be more effective for specific individual profiles and may not be as attractive to others. We support this hypothesis because participants from group 1 had positive regression lines over the 12 weeks of the intervention.

Interestingly, the time spent in sedentary behavior and the time spent in MVPA did not present the same trajectory as step count. The increase in the average daily step count was not accompanied by a reduction in the time spent in sedentary behavior or an increase in the time spent in MVPA, which can be explained by the outcome we measured. Participants may have altered their time for physical activities that we did not assess. Although the step count and the time spent in sedentary behavior are related [21], a systematic review and meta-analysis also found increases in step count not accompanied by reductions in sedentary behavior [31].

Our results are similar to those of Vandelanotte et al [32], who did not observe differences in the level of PA among 3 groups (1 control and 2 intervention groups). Despite differences in the study design and intervention protocol between the studies, both similarly associated the use of technology with behavior change techniques. The clinical trial conducted by Plotnikoff et al [33] found benefits in PA and general health among participants with type 2 diabetes, presenting a considerably higher effect size (Cohen *d*=0.67 for step count). Their protocol associated using a smartphone app with in-person meetings for group therapy and outdoor physical exercises [33]. It may be possible that, beyond the regular in-person meetings, a population

XSL•FO
RenderX

diagnosed with chronic conditions, such as diabetes, may be more motivated to become more active.

Likewise, other clinical trials aiming to increase the level of PA found better results in populations with clinical conditions than studies involving a healthy population [31]. In a systematic review and meta-analysis, Ringeval et al [31] showed that almost 60% of clinical trials that used an electronic device to monitor the level of PA included individuals at risk or already presenting with a clinical condition, and only 1 study included healthy individuals as the target population.

The nonresponse rate observed in our study (28/53, 53%) was lower than the expected rate (65%). This information may contribute to estimating the nonresponse rate in the planning of future clinical trials aiming to increase the level of PA in adaptive intervention designs, as well as the planning of adaptive intervention protocols with more decision points for rerandomization and changing the stimuli to favor behavior change in the practice of PA.

Respondents in week 6, independently of the initial or intermediate randomized allocation, constantly increased their average daily step count, while nonrespondents showed negative performance. Since our protocol did not include more decision points, it is impossible to know if these participants would benefit from more rerandomizations. These findings led us to reflect on the response at the intermediate assessment, which could be a prognostic indicator of the practice of PA. In this sense, Shang et al [34] pointed out the importance of deepening the investigation of fluctuating behavior in PA.

It is important for future studies to investigate if nonrespondents at the first phase of the intervention would benefit from new stimuli. In addition, a study with a larger sample size could investigate if nonrespondents present characteristics that could help health professionals and services to identify the population at higher risk for physical inactivity and thus design more personal approaches [34]. When planning interventions to increase PA, planning beyond the activity itself is necessary. As mentioned earlier, psychosocial, economic, and environmental factors are involved in this challenge [4]. Furthermore, there is a need to consider the dynamic characteristics of behavior change, which may include relapse periods and difficulties in maintaining the new behavior [34,35].

The loss to follow-up in our study, except for the control group, was consistent with the expectation for an interventional study to increase PA, which varies from 20% to 35% [32,33]. We faced difficulties following participants from the control group even before the onset of the COVID-19 pandemic. Our hypotheses for these losses are as follows: (1) it is possible that researchers and participants from groups 1 and 2 developed a bond due to the weekly text message interaction, which may have contributed to keeping the appointments for reassessments, and (2) it is possible that some of the participants may have volunteered for our study with the expectation of using an app for PA, which was not met for participants from the control group. For future studies, as a strategy to increase adherence among control participants, it may be helpful to offer an app without associating other behavior change techniques or even contact them periodically to develop a bond.

Many individuals have difficulty establishing effective strategies to change their behavior regarding PA. As pointed out by Warburton and Bredin [36], there is enough knowledge about the effects and benefits of the practice of PA, and the current challenge lies in translating knowledge into strategies to increase PA at a populational level. In our study, we tried to encompass different determinants of PA [4,5] at the individual level (eg, text messages and self-monitoring could support motivation) and the interpersonal level (eg, social support from the research team, in-person activities in group 3, and establishment of cultural norms among participants); provide a social or cultural environment through app features, such as ranking and groups/teams; and build an information environment through counseling and education.

After analyzing the performance of participants in this study, we suggest that intervention protocols aiming to increase the level of PA in research, clinical, and collective contexts should consider the adaptive intervention design as a viable and potentially effective alternative, as well as the association with environmental and institutional interventions [1]. The recommendations of the World Health Organization Global Action Plan [3] and ecological models [4,5] reinforce the complexity of this behavior change.

Finally, it is essential to highlight that recommendations about the ideal quantity of PA are being constantly revised. Considering the current recommendation of the World Health Organization for the adult population that PA of any amount and intensity is better than no PA [37] and the findings from Warburton and Bredin [36] that there are benefits of transitioning from a lower to a higher level of PA independently of the amount, we believe that our protocol was effective, as all participants presented some increase in the average daily step count.

As strengths of our study, we highlight the adaptive intervention design, which is still not explored in PA, and its association with behavior change techniques and complex technology. Moreover, despite the small sample size, we found relevant information that may contribute to advancing knowledge for PA promotion.

Our study had some limitations. Although we completed the first phase of the study, the second phase was not possible due to the onset of the COVID-19 pandemic. Data analysis was limited due to the small sample size and losses to follow-up, especially after the onset of the pandemic, which led to the sudden interruption of the protocol. Although we managed this limitation using adequate statistical treatment of the data, the analysis may be underpowered. Another limitation regarding the data analysis was the intention-to-treat approach adopted in our study, which could underestimate the impact of other factors influencing participation in the study and adherence to the protocol [38]. Finally, the characteristics of the sample, such as a higher level of education and having access to technologies, may limit the generalization of the results.

As observed in other PA studies, most participants completed at least high school [32,39,40]. Epidemiological studies have indicated that a higher education level and income are associated with higher practice of leisure-time PA [4]. One of the

XSL•FO
RenderX

challenges in designing programs for PA is engaging individuals who are less likely to participate in leisure-time PA due to the perception of limited time, a limited understanding of the effects of the practice of PA, or fatigue resulting from physically demanding work [40]. The socioeconomic characteristics of our sample may have contributed to our results.

In conclusion, our adaptive intervention protocol using a smartphone app combined with behavior change techniques led to increases in the level of PA over time, especially among participants from group 1. Participants from group 1 showed increases in PA levels at all assessments, while those from group 2 showed increases only at the final assessment when compared with the initial assessment. Moreover, there was a time effect for the increase in the time spent in MVPA. It is possible that offering stepwise new stimuli and behavior change techniques may contribute positively to the process of behavior change regarding PA. We acknowledge that the small sample size may limit the interpretation of the results.

Acknowledgments

We are thankful to all participants who dedicated their time to this study. We also appreciate the commitment of each member of the Epidemiology and Human Movement team during all phases of the study.

Funding

This study was funded by the Sao Paulo Research Foundation (grant 2016/50249–3) and by the Sao Paulo Research Foundation and the Brazilian Federal Agency for Support and Evaluation (grant 2018/21536).

Data Availability

Data may be available upon reasonable request to the corresponding author.

Authors' Contributions

Conceptualization: MSMPS, NLP, VTL, MBN, RCP, VZD
Data curation: MSMPS, VZD
Formal analysis: MSMPS, VZD
Funding acquisition: VZD
Investigation: MSMPS, NLP, VTL, MBN
Methodology: MSMPS, NLP, VTL, MBN, RCP, VZD
Project administration: MSMPS, VZD
Resources: VZD
Supervision: RCP, VZD
Visualization: MSMPS, NLP, VTL, MBN, RCP, VZD
Writing – original draft: MSMPS
Writing – review & editing: NLP, VTL, MBN, RCP, VZD

Conflicts of Interest

None declared.

Checklist 1
CONSORT‐EHEALTH checklist (V 1.6.1).
[PDF File, 10034 KB - jmir_v28i1e73388_app1.pdf]

References

1. Heath GW, Parra DC, Sarmiento OL, et al. Evidence-based intervention in physical activity: lessons from around the world. The Lancet 2012 Jul;380(9838):272-281. [doi: 10.1016/S0140-6736(12)60816-2]
2. Salvador EP, Ribeiro EH, Garcia LM, et al. Interventions for physical activity promotion applied to the primary healthcare settings for people living in regions of low socioeconomic level: study protocol for a non-randomized controlled trial. Arch Public Health 2014 Mar 13;72(1):8. [doi: 10.1186/2049-3258-72-8] [Medline: 24624930]
3. Global action plan on physical activity 2018–2030: more active people for a healthier world. World Health Organization. 2018. URL: https://www.who.int/publications/i/item/9789241514187 [accessed 2025-11-21]
4. Bauman AE, Reis RS, Sallis JF, Wells JC, Loos RJ, Martin BW. Correlates of physical activity: why are some people physically active and others not? The Lancet 2012 Jul;380(9838):258-271. [doi: 10.1016/S0140-6736(12)60735-1]
5. Sallis JF, Cervero RB, Ascher W, Henderson KA, Kraft MK, Kerr J. An ecological approach to creating active living communities. Annu Rev Public Health 2006;27:297-322. [doi: 10.1146/annurev.publhealth.27.021405.102100] [Medline: 16533119]

XSL•FO
RenderX

6. Duncan M, Vandelanotte C, Kolt GS, et al. Effectiveness of a web- and mobile phone-based intervention to promote physical activity and healthy eating in middle-aged males: randomized controlled trial of the ManUp study. J Med Internet Res 2014 Jun 12;16(6):e136. [doi: 10.2196/jmir.3107] [Medline: 24927299]
7. Schoeppe S, Alley S, Van Lippevelde W, et al. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review. Int J Behav Nutr Phys Act 2016 Dec 7;13(1):127. [doi: 10.1186/s12966-016-0454-y] [Medline: 27927218]
8. Fanning J, Mullen SP, McAuley E. Increasing physical activity with mobile devices: a meta-analysis. J Med Internet Res 2012 Nov 21;14(6):e161. [doi: 10.2196/jmir.2171] [Medline: 23171838]
9. Schoeppe S, Alley S, Rebar AL, et al. Apps to improve diet, physical activity and sedentary behaviour in children and adolescents: a review of quality, features and behaviour change techniques. Int J Behav Nutr Phys Act 2017 Jun 24;14(1):83. [doi: 10.1186/s12966-017-0538-3] [Medline: 28646889]
10. Sun L, Wang Y, Greene B, et al. Facilitators and barriers to using physical activity smartphone apps among Chinese patients with chronic diseases. BMC Med Inform Decis Mak 2017 Apr 19;17(1):44. [doi: 10.1186/s12911-017-0446-0] [Medline: 28420355]
11. Middelweerd A, Mollee JS, van der Wal CN, Brug J, Te Velde SJ. Apps to promote physical activity among adults: a review and content analysis. Int J Behav Nutr Phys Act 2014 Jul 25;11:97. [doi: 10.1186/s12966-014-0097-9] [Medline: 25059981]
12. Vandelanotte C, Müller AM, Short CE, et al. Past, present, and future of eHealth and mHealth research to improve physical activity and dietary behaviors. J Nutr Educ Behav 2016 Mar;48(3):219-228. [doi: 10.1016/j.jneb.2015.12.006] [Medline: 26965100]
13. van Stralen MM, Lechner L, Mudde AN, de Vries H, Bolman C. Determinants of awareness, initiation and maintenance of physical activity among the over-fifties: a Delphi study. Health Educ Res 2010 Apr;25(2):233-247. [doi: 10.1093/her/cyn045] [Medline: 18927443]
14. Kumar S, Nilsen WJ, Abernethy A, et al. Mobile health technology evaluation: the mHealth evidence workshop. Am J Prev Med 2013 Aug;45(2):228-236. [doi: 10.1016/j.amepre.2013.03.017] [Medline: 23867031]
15. Collins LM, Murphy SA, Strecher V. The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): new methods for more potent eHealth interventions. Am J Prev Med 2007 May;32(5 Suppl):S112-S118. [doi: 10.1016/j.amepre.2007.01.022] [Medline: 17466815]
16. Martin SS, Feldman DI, Blumenthal RS, et al. mActive: a randomized clinical trial of an automated mHealth intervention for physical activity promotion. J Am Heart Assoc 2015 Nov 9;4(11):e002239. [doi: 10.1161/JAHA.115.002239] [Medline: 26553211]
17. Gonze BDB, Padovani RDC, Simoes MDS, et al. Use of a smartphone app to increase physical activity levels in insufficiently active adults: feasibility sequential multiple assignment randomized trial (SMART). JMIR Res Protoc 2020 Oct 23;9(10):e14322. [doi: 10.2196/14322] [Medline: 33094733]
18. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. Lancet 2025 Apr 14:S0140-6736(25)00672-5. [doi: 10.1016/S0140-6736(25)00672-5] [Medline: 40245901]
19. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
20. Morais Pereira Simões MDS, de Barros Gonze B, Leite Proença N, et al. Use of a smartphone app combined with gamification to increase the level of physical activity of adults and older adults: protocol of a sequential multiple assignment randomized trial. Trials 2019 Dec 27;20(1):780. [doi: 10.1186/s13063-019-3879-1] [Medline: 31881987]
21. Tudor-Locke C, Craig CL, Thyfault JP, Spence JC. A step-defined sedentary lifestyle index: <5000 steps/day. Appl Physiol Nutr Metab 2013 Feb;38(2):100-114. [doi: 10.1139/apnm-2012-0235] [Medline: 23438219]
22. Conheça Santos [Article in Portuguese]. Prefeitura de Santos. URL: https://www.santos.sp.gov.br/?q=hotsite/conheca-santos [accessed 2022-03-23]
23. Dallal GE. Randomization. 2018. URL: http://www.randomization.com [accessed 2020-12-22]
24. Critério Brasil [Article in Portuguese]. Associação Brasileira de Empresas de Pesquisa. URL: https://abep.org/criterio-brasil/ [accessed 2025-11-21]
25. Guedes DP, Anira dos Santos C, Lopes CC. Estágios de mudança de comportamento e prática habitual de atividade física em universitários [Article in Portuguese]. Rev Bras Cineantropom Desempenho Hum 2006;8(4):5-15 [FREE Full text]
26. Troiano RP, Berrigan D, Dodd KW, Mâsse LC, Tilert T, McDowell M. Physical activity in the United States measured by accelerometer. Med Sci Sports Exerc 2008 Jan;40(1):181-188. [doi: 10.1249/mss.0b013e31815a51b3] [Medline: 18091006]
27. Freedson PS, Melanson E, Sirard J. Calibration of the Computer Science and Applications, Inc. accelerometer. Medicine & Science in Sports & Exercise 1998 May;30(5):777-781. [doi: 10.1097/00005768-199805000-00021]
28. Almirall D, Compton SN, Gunlicks-Stoessel M, Duan N, Murphy SA. Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy. Stat Med 2012 Jul 30;31(17):1887-1902. [doi: 10.1002/sim.4512] [Medline: 22438190]

XSL•FO
RenderX

29. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
30. Wright CE, Rhodes RE, Ruggiero EW, Sheeran P. Benchmarking the effectiveness of interventions to promote physical activity: a metasynthesis. Health Psychol 2021 Nov;40(11):811-821. [doi: 10.1037/hea0001118] [Medline: 34914485]
31. Ringeval M, Wagner G, Denford J, Paré G, Kitsiou S. Fitbit-based interventions for healthy lifestyle outcomes: systematic review and meta-analysis. J Med Internet Res 2020 Oct 12;22(10):e23954. [doi: 10.2196/23954] [Medline: 33044175]
32. Vandelanotte C, Short CE, Plotnikoff RC, et al. Are web-based personally tailored physical activity videos more effective than personally tailored text-based interventions? Results from the three-arm randomised controlled TaylorActive trial. Br J Sports Med 2021 Mar;55(6):336-343. [doi: 10.1136/bjsports-2020-102521] [Medline: 33144346]
33. Plotnikoff RC, Wilczynska M, Cohen KE, Smith JJ, Lubans DR. Integrating smartphone technology, social support and the outdoor physical environment to improve fitness among adults at risk of, or diagnosed with, Type 2 Diabetes: findings from the "eCoFit" randomized controlled trial. Prev Med 2017 Dec;105:404-411. [doi: 10.1016/j.ypmed.2017.08.027] [Medline: 28887192]
34. Shang B, Duan Y, Huang WY, Brehm W. Fluctuation - a common but neglected pattern of physical activity behaviour: an exploratory review of studies in recent 20 years. Eur J Sport Sci 2018 Mar;18(2):266-278. [doi: 10.1080/17461391.2017.1417486] [Medline: 29334317]
35. Kwasnicka D, Dombrowski SU, White M, Sniehotta F. Theoretical explanations for maintenance of behaviour change: a systematic review of behaviour theories. Health Psychol Rev 2016 Sep;10(3):277-296. [doi: 10.1080/17437199.2016.1151372] [Medline: 26854092]
36. Warburton DER, Bredin SSD. Health benefits of physical activity: a systematic review of current systematic reviews. Curr Opin Cardiol 2017 Sep;32(5):541-556. [doi: 10.1097/HCO.0000000000000437] [Medline: 28708630]
37. WHO guidelines on physical activity and sedentary behaviour: at a glance. World Health Organization. 2020. URL: https://www.who.int/publications/i/item/9789240014886 [accessed 2025-11-21]
38. Nich C, Carroll KM. Intention-to-treat meets missing data: implications of alternate strategies for analyzing clinical trials data. Drug Alcohol Depend 2002 Oct 1;68(2):121-130. [doi: 10.1016/s0376-8716(02)00111-4] [Medline: 12234641]
39. Lee CF, Ho JWC, Fong DYT, et al. Dietary and physical activity interventions for colorectal cancer survivors: a randomized controlled trial. Sci Rep 2018 Apr 10;8(1):5731. [doi: 10.1038/s41598-018-24042-6] [Medline: 29636539]
40. Biswas A, Dobson KG, Gignac MAM, de Oliveira C, Smith PM. Changes in work factors and concurrent changes in leisure time physical activity: a 12-year longitudinal analysis. Occup Environ Med 2020 May;77(5):309-315. [doi: 10.1136/oemed-2019-106158] [Medline: 32107318]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
EPIMOV: Epidemiology and Human Movement
MVPA: moderate-to-vigorous physical activity
PA: physical activity
SMART: sequential multiple assignment randomized trial

Edited by J Sarvestan; submitted 03.Mar.2025; peer-reviewed by A Tabaczynski, H Namba; revised version received 14.Jul.2025; accepted 16.Jul.2025; published 30.Jan.2026.

Please cite as:
Simoes MDSMP, Proença NL, Lauria VT, do Nascimento MB, Padovani RDC, Dourado VZ
Effects of Using a Smartphone App Combined With Behavior Change Techniques on the Level of Physical Activity Among Adults and Older Adults: Sequential Multiple Assignment Randomized Trial
J Med Internet Res 2026;28:e73388
URL: https://www.jmir.org/2026/1/e73388
doi:10.2196/73388

© Maria do Socorro Morais Pereira Simoes, Neli Leite Proença, Vinícius Tonon Lauria, Matheus Bibian do Nascimento, Ricardo da Costa Padovani, Victor Zuniga Dourado. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effect of Lung Cancer Screening, Smoking Cessation, and Cessation Smartphone App to Health-Related Quality of Life Among Heavy Smokers: Randomized Controlled Trial

Antti Kurtti[1,2,3], BMed; Sanna Iivanainen[1,2,3], MD, PhD; Riitta Kaarteenaho[3,4], MD, PhD; Heidi Andersen[5,6], MD, PhD; Antti Jekunen[5,6], MD, PhD; Tuula Vasankari[6,7], MD, PhD; Jussi Koivunen[1,2,3], MD, PhD

[1]Cancer Center, Oulu University Hospital, Kajaanintie 50, Oulu, Finland
[2]Faculty of Medicine, University of Oulu, Oulu, Finland
[3]Medical Research Center Oulu, Oulu, Finland
[4]Research Unit of Biomedicine and Internal Medicine, University of Oulu, Oulu, Finland
[5]Department of Oncology and Radiotherapy, Vaasa Central Hospital, Vaasa, Finland
[6]Faculty of Medicine, University of Turku, Turku, Southwest Finland, Finland
[7]FILHA, Helsinki, Finland

Corresponding Author:
Jussi Koivunen, MD, PhD
Cancer Center, Oulu University Hospital, Kajaanintie 50, Oulu, Finland

Abstract

Background: Lung cancer screening with low-dose computed tomography (LDCT) among heavy smokers can decrease lung cancer mortality. Smoking cessation intervention is recommended within the screening program; however, the methods for smoking cessation in the LDCT screening context are not well established. We have previously shown that a novel smartphone app can increase the chance for smoking cessation along with lung cancer screening. The effects of lung cancer screening, smoking cessation, and the use of smartphone apps on health-related quality of life (HRQoL) are widely unknown.

Objective: This study aims to investigate the effect of lung cancer screening, smoking cessation, and the use of smoking cessation app on HRQoL, an exploratory end point of the low-dose computed tomography screening for lung cancer combined to different smoking cessation methods in Finland (LDCT-SC-FI) study.

Methods: This study was conducted as a part of the LDCT-SC-FI (NCT05630950), which was a randomized controlled trial investigating different smoking cessation methods in participants undergoing lung cancer screening with LDCT. The main inclusion criteria included an age of 50 - 74 years, a marked smoking history (smoked ≥15 cigarettes per day for ≥25 years or smoked ≥10 cigarettes per day for ≥30 years), an active smoking status, and access to a smartphone. The recruitment was carried out by newspaper and internet advertisements and informing relevant health care units at hospital districts. The study participants (n=200), all at Oulu University Hospital, were randomized in 1:1 fashion to a yearly LDCT with standard smoking cessation (written material) or a stand-alone smartphone app–based cessation. HRQoL, an exploratory study end point, was assessed at baseline and at 1 year with Quality of Life Questionnaire Core 30 (QLQ-C30) and EQ-5D.

Results: In total, 199 and 186 individuals had both questionnaires completed at baseline and at 1 year, respectively. We did not detect a change in HRQoL between the time points using QLQ-C30 global health status score or EQ-5D index score. Smoking cessation at 1-year time did not affect QLQ-C30 global health status or EQ-5D. We observed improved quality of life scores by EQ-5D at 1 year (control: mean 0.720, SD 0.197 vs app: mean 0.799, SD 0.197; improved in 17/93, 18% of controls vs 29/93, 31% in app arm), while there was no difference in means at baseline. Smartphone app arm reported reduced pain (EQ-5D effect size [ES] 0.049, 95% CI 0.006 - 0.12; P=.01; adjusted ES 0.026; P=.007; QLQ-C30 ES 0.076, 95% CI 0.02 - 0.16; P<.001; adjusted ES 0.05; P=.02) and increased mobility (EQ-5D ES 0.031, 95% CI 0.01 - 0.09; P=.02; adjusted ES 0.037; P=.008) at 1 year. The number of completed questionnaires in the app was associated with improved HRQoL by EQ-5D (ES 0.073, 95% CI 0.00 - 0.180; P=.04; adjusted ES 0.071; P=.04).

Conclusions: This is the first study to test a smoking cessation smartphone app in the context of lung cancer screening. The use of the developed app correlated with improved HRQoL, mainly by decreased pain and fatigue. To conclude, the studied app provides a feasible and effective cessation intervention that is readily implementable in population-based lung cancer screening programs, with enhanced health benefits beyond smoking cessation.

Trial Registration: ClinicalTrial.gov NCT05630950; https://clinicaltrials.gov/study/NCT05630950

XSL•FO
RenderX

(J Med Internet Res 2026;28:e81687) doi:10.2196/81687

KEYWORDS

lung cancer screening; smoking cessation; mobile app; low-dose computed tomography; health-related quality of life

Introduction

The premier cause of cancer mortality in the Western world is lung cancer, of which smoking is the single most important risk factor [1]. Lung cancer is typically diagnosed at an advanced stage, preventing curative intent treatments. Using low-dose computed tomography (LDCT) for lung cancer screening in individuals with substantial smoking history can decrease lung cancer mortality [2,3]. Concordantly, smoking cessation interventions are recommended in lung cancer screening programs and may come with enhanced efficiency, regardless of screening results [4-7]. Approximately 7%‑23% of participants in LDCT programs achieve smoking cessation [8]. Still, optimal smoking cessation methods in the context of lung cancer screening are not well established.

Possible health-related quality of life (HRQoL) effects and losses are important to evaluate in cancer screening trials [9]. There is moderate evidence that, compared with no screening, persons receiving LDCT screening do not have worse general HRQoL or distress over 2 years of follow-up, and anxiety may even be lowered. However, consequences might differ based on screening results, at least in the short term [10-12]. Observed negative psychological effects diminished over time, and there were no known negative long-term effects on HRQoL [13,14].

Opportunities for using digital tools to access smoking cessation treatment are evolving rapidly due to the expansion in the proportion of the global population with access to a mobile phone. Data imply that apps that provide personalized and adaptive as well as interactive support may be more effective in promoting engagement [15-17]. Smartphone apps have reached abstinence odds ratios (ORs) of 1.25‑1.51 (95% CI 0.99‑1.56, 1.24‑1.84) in meta-analyses [18]. Their efficacy can be enhanced with pharmacotherapy and physical participant recruitment [19]. We have recently shown that a novel smartphone app can increase smoking cessation with an OR of ~3 at 3 and 6 months when applied within LDCT screening for lung cancer [20].

Several studies have indicated that smoking is related to lower HRQoL and mental well-being [21-23]. However, studies on the impact of smoking cessation interventions on HRQoL have provided mixed results, some concluding that it may improve, yet others showing negative or no changes [24-26]. Smoking cessation may have a positive effect on physical and general health but no significant effects on mental aspects [24]. Nevertheless, it might take years of abstinence for smokers' HRQoL to be equal to nonsmokers' [24,27].

Low-dose computed tomography screening for lung cancer combined with different smoking cessation methods in Finland (LDCT-SC-FI) is a randomized controlled trial investigating smoking cessation with a smartphone app compared to written cessation materials in individuals participating in LDCT screening for lung cancer. The core concepts behind the developed app include cognitive behavioral (enhancing self-awareness, problem-solving skills, goal setting, and coping with cravings) and social cognitive theories and acceptance and commitment therapy as well as mindfulness. Therefore, the effects of smartphone apps could extend beyond smoking cessation.

This study aims to investigate changes in HRQoL, an exploratory end point of the LDCT-SCI-FI trial. We investigated the changes in HRQoL using 3 different patient-reported outcome measures (PROMs). Since the LDCT-SC-FI study focused on 2 main themes (feasibility of LDCT screening and smoking cessation with a smartphone app), the results were analyzed to ascertain what effects participation in screening, smoking cessation, and use of the interventional app might have on HRQoL outcomes.

Methods

Study Design

This study was conducted as a part of the LDCT-SC-FI (NCT05630950), which is a randomized controlled trial investigating different smoking cessation methods in participants undergoing lung cancer screening with LDCT. The study participants were randomized in a 1:1 fashion to a yearly LDCT with standard smoking cessation (written material) or the same LDCT screening approach with a smartphone app–based smoking cessation (experimental). The study was powered (80%) with 156 participants to detect a 15% difference with α of .1 (75% vs 90%) in the number of active smokers at 3 and 6 months after inclusion using an online calculator. With the expected dropout rate, the sample size was adjusted to 200.

Ethical Considerations

The study was approved by the ethics committee of Northern Ostrobothnia Hospital District (EETTKM 21/2022) and was prospectively registered at ClinicalTrials.gov (NCT05630950). All the participants signed an informed consent before any study procedures. The study participants were not compensated for their participation. Of note, LDCT lung cancer screening is not among the publicly funded cancer screenings in Finland. The study was conducted in accordance with the Declaration of Helsinki [28] and Good Clinical Practice guidelines [29]. Participant data were deidentified prior to analysis. Unique study codes were assigned to each participant, while identifying information was removed. All results are reported in aggregate form to protect participant confidentiality. Consent for publication has been granted by identifiable individuals.

Participants

Eligibility followed closely to the Nederlands-Leuvens Longkanker Screenings Onderzoek (NELSON) lung cancer screening trial criteria [2]. The main inclusion criteria included an age of 50‑74 years, a marked smoking history (smoked ≥15 cigarettes per day for ≥25 years or smoked ≥10 cigarettes per

XSL•FO
RenderX

day for ≥30 years), an active smoking status (smoking during the last 2 weeks including regular [daily smoking] and occasional [nondaily smoking] habits), and access to a smartphone (iPhone or Android). The main exclusion criteria, as in the NELSON trial, included a moderate or bad self-reported health; current or past melanoma, lung, renal, or breast cancer; and a chest computed tomography (CT) examination within 1 year.

Recruitment and Randomization

All the participants were recruited at the Oulu University Hospital (from November 18, 2022, to April 14, 2023; the final study visit occurred on March 20, 2024). The recruitment was carried out by newspaper and internet advertisements and informing relevant health care units at the hospital district. A physical screening visit was performed at the site where participating individuals signed the informed consent. Eligibility was verified by a study nurse according to a checklist. Study participants did not receive any compensation for participation, and all the study procedures were free of charge. Eligible participants were randomized by study nurses with block method (sequentially numbered containers with a block size of 10, study arm written on a paper in an opaque, sealed envelope) in 1:1 fashion with stratification according to pack years (<30 or ≥30 pack years) and age (<65 or ≥65 years) to smartphone-based smoking cessation and control (written smoking cessation material) arms. The stratification factors were selected based on the assumption that bias could be generated by (1) adoption of smartphone use in older people and (2) higher pack years to be associated with a lesser likelihood of smoking cessation. The random allocation sequence (randomization envelopes and numbered blocks) was generated by the investigators (JK and SI) to ensure concealment. Because of an inability to blind the study participants from the intervention, as well as self-reported smoking cessation being the primary end point of the study, the study personnel were not blinded.

Outcomes

Data collected at baseline included standard demographics and detailed smoking history. Self-reported smoking cessation status was verified by phone at 3 and 6 months, and as a part of physical visit at 1 year. Another physical visit and LDCT screening investigation took place at 1 year. Quality of life (QoL) questionnaires were collected after randomization at baseline and at 1 year.

The primary outcomes of the study were self-reported smoking cessation at 3 and 6 months (±1 month), which has been recently reported, and study details are available in this publication [20]. Furthermore, secondary outcomes of the study included efficiency of different smoking cessation methods in the reduction of smoking, sensitivity and positive predictive value of CT screening, and costs related to CT screening, which are reported in previous publications. HRQoL assessed with different PROMs was an exploratory end point of the study and was not specified in detail in the protocol. Since the study focused on 2 main themes (smoking cessation with a smartphone app and feasibility of LDCT lung cancer screening), we planned to investigate HRQoL in the context of smoking cessation, study arm, and temporally. All the analysis was carried out according to specific instructions by the HRQoL questionnaire. ED-5D index scores and QLQ-C30 global health status (GHS) have standardized population-based cutoffs (5% for EQ-5D and 10% for the QLQ-C30), which were used in the main analysis of HRQoL. If any of the main analyses showed statistically significant differences temporally, or by study arm, or smoking status at 1 year, detailed analysis of specific subscales or scores would be carried out.

Procedures

The developed smoking cessation app called Suunta supports smokers in the cessation process and aids them to retain a smoking-free lifestyle (Multimedia Appendix 1). The theoretical and functional concept was created by the study team members, and the technical execution was done under subcontract by a company specialized in mobile app development (Techinspire). The stand-alone app is Android- and iPhone-compatible with cloud-based database back-up. The Suunta app was downloaded on participants' smartphones on the randomization visit, which was assisted by the study nurse. The Suunta smartphone app was offered for the individuals in the control arm after 6 months of follow-up.

The written materials used for smoking cessation are based on the Finnish Current Care Guideline for Prevention and Treatment of Smoking and Nicotine Addiction. A printed version of the patient guide was handed out to all the study participants. At the 6-month smoking status call, participants in the control arm were offered the possibility to start using the smoking cessation app. No other counseling for smoking cessation was provided to the participants during the study visits regardless of the study arm.

The LDCT-SC-FI study protocol for LDCT (Multimedia Appendix 2) follows the NELSON study protocol [2]. In brief, all the study participants undergo LDCT screening within 6 weeks from the randomization, and the next LDCT is scheduled for 1 year.

HRQoL Measurements

Study participants were assessed for HRQoL with European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire Core 30 (QLQ-C30)+Quality of Life Questionnaire Lung Cancer 13 (QLQ-LC13) and EuroQol EQ-5D-3L at baseline and at 1 year. The participants completed the questionnaires at a physical visit using paper format. HRQoL and changes therein were collected as the QLQ-C30 GHS and EQ-5D index scores at baseline and 1 year in all participants as well as according to smoking status at 1 year and study arms. HRQoL-related secondary outcomes included the means and changes in the QLQ-C30 functional scales as well as in the specific symptom scores of QLQ-C30, QLQ-LC13, and EQ-5D. These were carried out only if the overall score was found to be significant at a 1-year time point.

The EORTC QLQ-C30 (version 3.0) and QLQ-LC13 questionnaires were scored according to official EORTC scoring manuals. In brief, all the scales and single-item measures undergo linear transformation to range from 0 to 100, so that a high score for the functional and GHS scale represents a high level of functioning or QoL, while the score for the symptom

XSL•FO
RenderX

scales and single items represents a high level of symptomatology or problems. A 10-point change in scores is considered meaningful, so a 10% cutoff was selected for analysis [30].

With EQ-5D-3L questionnaires, single scores were reported by 3-grade answers representing no to severe problems. For the EQ-5D index score, all the scores were transformed to a single raw score between 0 and 1 according to the official scoring manual. Raw scores were transformed to a country-specific index score (Finland) using a Microsoft Excel-based calculator. A 5% change in EQ-5D index score is considered significant in Finland, and this was selected as a primary measure [31,32].

The use of the app and its association with HRQoL changes observed in the QLQ-C30 GHS and EQ-5D index scores were studied in the experimental arm. The smartphone app included weekly symptom questionnaires, and the frequency of app use was investigated by analyzing the number of completed symptom questionnaires by 24 weeks.

Statistical Analysis

Data analysis was carried out using SPSS (version 29.0.1; IBM Corp). All the statistical analyses were carried out blinded to the group allocation. Reliability of HRQoL data was evaluated with Cronbach α, with values of <0.6 considered poor or unacceptable. Paired 2-tailed *t* test was used to compare means of 2 related groups (change over time) and estimate *P* value and effect size (ES) with 95% CIs. For categorized variables, Pearson chi-square test was used to estimate *P* values, and 1-way ANOVA for ES with 95% CIs. For continuous variables, 1-way ANOVA was applied to estimate the means, *P* values, and ES with 95% CIs. One-way ANOVA was also applied in the analysis of variables with more than 2 categories to estimate *P* values and ES with 95% CIs. Adjusted analysis for *P* values and ES was carried out using analysis of covariance corrected for relevant baseline factors. Since the statistical software used does not provide CIs for adjusted analysis of covariance, these are not provided. Ordinal regression analysis with the proportional odds model was used to estimate ORs, 95% CIs, and *P* values. The generated model was evaluated with model fitting, goodness-of-fit, and test-of-parallel-lines test. *P* values of <.05 were considered statistically significant. For the ES, partial η^2 (small: 0.01; medium: 0.06; large: 0.14) or Cohen *d* (small: 0.2; medium: 0.5; large: 0.8) were used to estimate the magnitude of the effect. Missing data were not replaced.

Results

Demographics

The recruitment was initiated on November 18, 2022, the last participant was included on April 14, 2023, and the final study visit occurred on March 20, 2024. The median age was 60 (IQR 56-66) years, and 51% (102/201) of them were female. In smoking-related demographics, the median number of pack years was 31 (IQR 24-40), and the number of smoked cigarettes per day was 15. The detailed demographics are presented in Table S1 in Multimedia Appendix 3 and the study flowchart in Figure 1.

XSL•FO
RenderX

Figure 1. Study flowchart of the LDCT-SC-FI, which was a randomized controlled trial investigating different smoking cessation methods in participants undergoing lung cancer screening with low-dose computed tomography. LDCT-SC-FI: low-dose computed tomography screening for lung cancer combined with different smoking cessation methods in Finland.

XSL•FO
RenderX

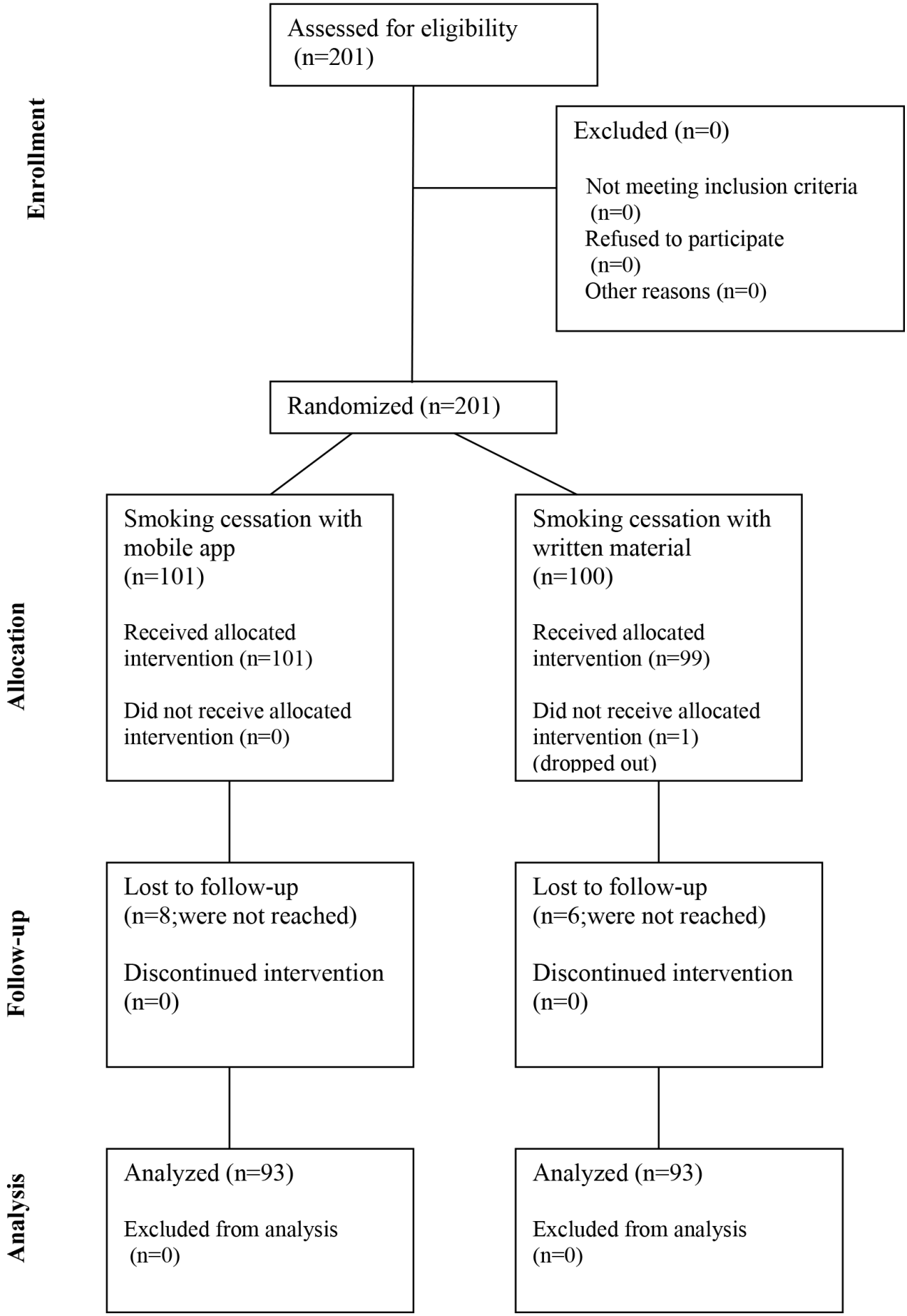

XSL•FO
RenderX

Temporal Changes in HRQoL During Screening

All the study participants (n=200) underwent LDCT screening at baseline and 183 of the eligible 196 (93%) at 1 year. At baseline, all the study participants (n=200) had 3 QoL (EQ-5D and QLQ-CO30+LC13) questionnaires filled, and all (n=186) came to the 1-year study visit (Figure 1). There were 14 dropouts for the 1-year visit (experimental: n=8; control: n=6), 4 by protocol-defined exclusion and 10 by self-willed withdrawal. The overall HRQoL scores and their changes between baseline and 1 year were analyzed. Data reliability was considered good or acceptable for overall HRQoL scores (Table S2 in Multimedia Appendix 3). Among all study participants, there was no statistically significant difference between the means of either score at baseline or at 1 year. However, there was a slight decrease in GHS over time. Most participants experienced no changes in either of their scores (GHS: 120/186, 65%, EQ-5D index score: 90/186, 48%). The subgroup of participants experiencing either improved or decreased HRQOL scores was quite balanced (Table 1). Furthermore, the association of EQ-5D index score change (5%) and GHS change (10%) was evaluated, and they showed a good correlation (Table S3 in Multimedia Appendix 3).

Table . Study participants were assessed for HRQoL[a] with European Organisation for Research and Treatment of Cancer QLQ-C30[b] and EuroQol EQ-5D-3L at baseline and at 1 year at a physical visit[c].

	Baseline	At 1 year	Effect size[d] (95% CI)	*P* value[e]
QLQ-C30 GHS[f], mean (SD)	73.21 (15.83)	71.64 (19.71)	0.09 (−0.6 to 0.23)	.23
QLQ-C30 GHS (10% change), n (%)	—[g]	186 (100)	—	—
Improved		33 (18)		
No change		120 (65)		
Decrease		33 (18)		
EQ-5D index score, mean (SD)	0.763 (0.195)	0.760 (0.201)	0.02 (−0.13 to 0.16)	.80
EQ-5D index score (5% change), n (%)	—	186 (100)	—	—
Improved		46 (25)		
No change		90 (48)		
Decrease		50 (27)		

[a]HRQoL: health-related quality of life.

[b]QLQ-C30: Quality of Life Questionnaire Core 30.

[c]HRQoL and changes therein were collected as the QLQ-C30 GHS and EQ-5D index scores. A 10-point change in scores of QLQ-C30 is considered meaningful; thus, a 10% cutoff was selected, while a 5% change in EQ-5D index score is considered significant in Finland.

[d]Cohen *d*.

[e]Paired samples *t* test.

[f]GHS: global health status.

[g]Not applicable.

HRQoL According to Smoking Status

Since smoking cessation was the primary end point of the study, we wanted to analyze whether smoking cessation would have an impact on HRQoL outcomes. Smoking status was available for all study participants (n=186) coming to the study visit at 1 year (Figure 1). We assessed the association of smoking cessation to GHS and EQ-5D scores. Smoking cessation status showed no statistically significant difference in either of the HRQoL measures at baseline or at 1 year. Furthermore, we detected no difference according to smoking status in participants whose scores improved, decreased, or remained the same (Table 2).

XSL•FO
RenderX

Table . Changes in quality of life based on QLQ-C30[a] and EQ-5D between baseline and at 1 year according to smoking status (at 1 year) irrespective of study arm.

	Smoker	Nonsmoker	Effect size (95% CI)[b]	*P* value[c]
QLQ-C30 GHS[d], mean (SD)				
Baseline	72.78 (15.97)	74.14 (16.72)	0.001 (0.00-0.003)	.70
1 year	71.44 (19.48)	72.70 (21.24)	0.001 (0.00-0.03)	.75
Mean change from baseline	−1.592 (17.09)	−1.437 (22.28)	0.00 (0.00-0.003)	.97
QLQ-C30 GHS (10% change), n (%)	157 (100)	29 (100)	0.006 (0.00-0.05)	.57
Improved	26 (17)	7 (24)		
No change	102 (65)	18 (62.1)		
Declined	29 (19)	4 (14)		
EQ-5D index score, mean (SD)				
Baseline	0.759 (0.196)	0.767 (0.204)	0.00 (0.00-0.02)	.86
1 year	0.760 (0.201)	0.758 (0.202)	0.00 (0.00-0.005)	.96
Mean change from baseline	−0.002 (0.17)	−0.009 (0.17)	0.00 (0.00-0.02)	.85
EQ-5D index score (5% change), n (%)	157 (100)	29 (100)	0.002 (0.00-0.04)	.80
Improved	40 (26)	6 (21)		
No change	76 (48)	14 (48)		
Declined	41 (26)	9 (31)		

[a]QLQ-C30: Quality of Life Questionnaire Core 30.

[b]Partial eta-square.

[c]ANOVA or Pearson chi-square test.

[d]GHS: global health status.

HRQoL According to Randomization Arm

Next, we assessed HRQoL outcomes according to study arm. There was no difference either in the GHS or EQ-5D index score at baseline. At 1 year, we observed a difference between the study arms in the means of both the GHS (68.46 vs 74.82; partial η^2=0.026, 95% CI 0.00 - 0.09; *P*=.03) and the EQ-5D index scores (0.720 vs 0.799; partial η^2=0.039, 95% CI 0.003 - 0.11; *P*=.007), with individuals randomized to the app arm showing higher scores. However, the statistical significance was lost for the GHS after adjusting for baseline scores. Using the 3-class classification of scores (improved, unchanged, or declined), there was a statistically significant difference between the study arms in the portion of participants whose scores improved and declined by EQ-5D (17/186, 18% vs 29/186, 31%; 31/186, 33% vs 19/186, 20%; partial η^2=0.032, 95% CI 0.001 - 0.10; *P*=.049). A similar trend was observed with GHS (improved: 11/186, 12% vs 22/186, 24%), but this did not reach statistical significance (Table 3).

XSL•FO
RenderX

Table . Changes in overall quality of life based on QLQ-C30[a] and EQ-5D questionnaires between randomization arms[b].

	Control[c]	App[d]	*P* value[e]	Effect size[f] (95% CI)	Adjusted *P* value[g]	Adjusted effect size[g] (95% CI)
QLQ-C30 GHS[h], mean (SD)						
Baseline	70.29 (15.85)	74.17 (16.31)	.09	0.015 (0.00-0.06)	—[i]	—
1 year	68.46 (19.77)	74.82 (19.23)	.03	0.026 (0.00-0.09)	.18	0.01
Mean change from baseline	−2.330 (16.41)	−0.807 (19.39)	.56	0.002 (0.00-0.03)	—	—
QLQ-C30 GHS (10% change), n (%)	93 (100)	93 (100)			—	—
Improved	11 (12)	22 (24)	.11	0.016 (0.00-0.07)		
No change	64 (69)	56 (60)	—	—		
Declined	18 (19)	15 (16)	—	—		
EQ-5D index score, mean (SD)						
Baseline	0.752 (0.197)	0.759 (0.206)	.82	0.000 (0.00-0.02)	—	—
1 year	0.720 (0.197)	0.799 (0.197)	.007	0.039 (0.003-0.11)	.003	0.047
Mean change from baseline	−0.033 (0.16)	0.027 (0.17)	.01	0.032 (0.001-0.10)	—	—
EQ-5D index score (5% change), n (%)	93 (100)	93 (100)			—	—
Improved	17 (18)	29 (31)	.049	0.032 (0.001-0.10)		
No change	45 (48)	45 (48)	—	—		
Declined	31 (33)	19 (20)	—	—		

[a]QLQ-C30: Quality of Life Questionnaire Core 30.

[b]Comparison is carried out based on means at individual time points, changes from baseline, means adjusted to baseline values, and by using clinically significant threshold changes (QLQ-C30 10% change; EQ-5D 5% change).

[c]Control: smoking cessation with written material.

[d]App: smartphone app–based smoking cessation.

[e]ANOVA or Pearson chi-square test.

[f]Partial eta-square.

[g]Analysis of covariance test adjusted for baseline QLQ-C30 GHS or EQ-5D index score.

[h]GHS: global health status.

[i]Not applicable.

We also carried out an ordinal regression analysis of EQ-5D index change (5%) with baseline factors and smoking status at 1 year. The model performance was poor with high *P* values for model fit, while goodness of fit and test of parallel lines showed both *P* values over >.5. Of the analyzed factors, only randomization arm (OR −0.668, 95% CI −1.227 to −0.109; *P*=.02) significantly predicted EQ-5D change (Table 4).

Since Suunta app was also offered to the control arm after 6 months, we wanted to assess whether this would affect the observed results. The app use was assessed at the 1-year study visit, and 14 participants (15% of the controls) reported having used the app. Removing these individuals from the analysis did not markedly alter the results compared to the intent-to-treat analysis. In brief, EQ-5D mean index scores and QLQ-C30 GHS means were similar at baseline, while there was a statistically significant difference at 1 year. Furthermore, according to 3-class classification of EQ-5D or QLQ-C30, very similar results favoring the app were observed while this was only statistically significant with QLQ-C30 (partial η^2=0.027, 95% CI 0.00 - 0.09; *P*=.09; partial η^2=0.023, 95% CI 0.00 - 0.08; *P*=.04, respectively; Table S4 in Multimedia Appendix 3).

XSL•FO
RenderX

Table . Ordinal regression analysis for clinically significant EQ-5D index change (5%) at 1 year.

	OR[a] (95% CI)	*P* value[b]
Age	0.034 (−0.018 to 0.087)	.20
Sex		
Female	−0.001 (−0.633 to 0.630)	.10
Male	0	__[c]
Relationship status		
Single	−0.082 (−0.713 to 0.548)	.83
In a relationship	0	—
ICT[d] skills		
Novice	−0.962 (−2.099 to 0.174)	.10
Average	−0.124 (−0.700 to 0.452)	.67
Experienced	0	—
How many cigarettes per day	0.004 (−0.072 to 0.080)	.92
Pack years	−0.003 (−0.041 to 0.035)	.88
Fagerström test	−0.23 (−0.337 to 0.291)	.89
Smoking status at 1 year		
Nonsmokers	−0.363 (−1.130 to 0.405)	.35
Smoker	0	—
Randomization arm		
Control	−0.668 (−1.227 to −0.109)	.02
App	0	—

[a]OR: odds ratio.

[b]Ordinal regression analysis with proportional odds model; model fitting: *P*=.36; goodness-of-fit: *P*=.38; test of parallel lines: *P*=.06.

[c]Not applicable.

[d]ICT: information and communication technology.

Data reliability of QLQ-C30-, LC13-, and EQ-5D–defined functioning or symptom subscales was considered good to questionable, and only a few symptoms (hemoptysis, sore mouth, and alopecia) produced poor or unacceptable scores (Table S2 in Multimedia Appendix 3). In the analysis for the QLQ-C30 functional scales, we observed no difference at baseline. At 1 year, there were statistically significant differences between study arms in role (mean 81.70, SD 25.20 vs mean 89.13, SD 19.83; partial η^2=0.026, 95% CI 0.00 - 0.09; *P*=.03) and social (mean 86.02, SD 19.24 vs mean 93.55, SD 15.74; partial η^2=0.04, 95% CI 0.01 - 0.11; *P*=.004) functioning scales, with higher means detected in the app arm. After adjusting for baseline scores, only social functioning scales retained the statistical significance (partial η^2=0.025; *P*=.03). Furthermore, improvements in means over time occurred only in the app arm (Table 5).

XSL•FO
RenderX

Table . QLQ-C30[a] functional scales of quality of life according to randomization arm[b].

	Control[c], mean (SD)	App[d], mean (SD)	*P* value[e]	Effect size[f] (95% CI)	Adjusted *P* value[g]	Adjusted effect size[f]
Physical functioning					—[h]	—
Baseline	80.49 (16.90)	82.60 (16.52)	.38	0.004 (0.00 to 0.04)		
1 year	79.20 (18.06)	81.54 (17.86)	.38	0.004 (0.00 to 0.04)		
Mean change	−2.097 (12.78)	−2.148 (13.40)	.98	0.00 (0.00 to 0.00)		
Role functioning						
Baseline	86.03 (21.65)	89.11 (18.62)	.28	0.006 (−0.00 to 0.04)	—	—
1 year	81.70 (25.20)	89.13 (19.83)	.03	0.026 (0.00 to 0.09)	.11	0.014
Mean change	−4.348 (20.36)	−1.268 (17.69)	.28	0.007 (0.00 to 0.05)	—	—
Emotional functioning					—	—
Baseline	83.93 (15.29)	85.81 (14.84)	.38	0.004 (0.00 to 0.04)		
1 year	82.97 (19.27)	87.72 (14.04)	.06	0.02 (0.00 to 0.08)		
Mean change	−0.815 (16.53)	0.807 (12.17)	.45	0.003 (0.00 to 0.04)		
Cognitive functioning					—	—
Baseline	87.71 (16.08)	88.78 (15.11)	.63	0.001 (0.00 to 0.03)		
1 year	85.90 (19.07)	90.50 (16.00)	.08	0.017 (0.00 to 0.07)		
Mean change	−1.648 (14.50)	0.000 (13.23)	.42	0.004 (0.00 to 0.04)		
Social functioning						
Baseline	90.57 (18.16)	94.72 (12.00)	.06	0.018 (0.00 to 0.07)	—	—
1 year	86.02 (19.24)	93.55 (15.74)	.004	0.044 (0.01 to 0.11)	.03	0.025
Mean change	−4.480 (19.06)	−1.971 (17.18)	.35	0.005 (0.00 to 0.04)	—	—

[a]QLQ-C30: Quality of Life Questionnaire Core 30.

[b]Comparison was carried out based on means at individual time points, changes from baseline, and means adjusted to baseline values.

[c]Control: smoking cessation with written material.

[d]App: smartphone app–based smoking cessation.

[e]ANOVA.

[f]Partial eta-square.

[g]Analysis of covariance test adjusted for baseline QLQ-C30 functional scales.

[h]Not applicable.

Of the specific symptom scores, fatigue, pain, insomnia, and financial difficulties were calculated based on the QLQ-C30, while dyspnea at rest and pain in other parts were based on the QLQ-LC13. In the analysis, only insomnia and dyspnea at rest showed differences at baseline, but differences were subtle. At 1 year, fatigue, pain, insomnia, financial difficulties, and pain in other parts showed lower means and symptom burden in the app arm. After adjusting for baseline scores, only pain and final difficulties retained statistical difference (partial η^2=0.05; *P*=.02; partial η^2=0.03; *P*=.02). Only the pain score showed a statistically significant difference in mean change (7.065 vs −0.538; partial η^2=0.026, 95% CI 0.00 - 0.09; *P*=.03) between the study arms (Table 6).

We also analyzed specific symptoms from the EQ-5D questionnaire. In the analysis, there was no difference at baseline according to study arm (not shown). At 1 year, we observed improvements in both mobility (partial η^2=0.031, 95% CI 0.001 - 0.09; *P*=.02) and pain or discomfort (partial η^2=0.049, 95% CI 0.006 - 0.12; *P*=.01) with individuals randomized to the app arm, and these retained in the adjusted analysis (partial η^2=0.037; *P*=.008; partial η^2=0.026; *P*=.007; Table 7).

XSL•FO
RenderX

Table . QLQ-C30[a]+QLQ-LC13[b]–specific symptom scores showing significant changes according to randomization arm[c].

	Control[d], mean (SD)	App[e], mean (SD)	*P* value[f]	Effect size[g] (95% CI)	Adjusted *P* value[h]	Adjusted effect size[g]
Fatigue						
Baseline	24.02 (18.08)	21.56 (19.93)	.36	0.004 (0.00-0.04)	—[i]	—
1 year	26.05 (21.27)	19.59 (17.75)	.03	0.027 (0.00-0.09)	.10	0.015
Mean change	2.031 (15.62)	−0.239 (17.91)	.36	0.005 (0.00-0.04)	—	—
Pain						
Baseline	24.32 (27.62)	18.65 (22.15)	.11	0.013 (0.00-0.06)	—	—
1 year	32.26 (32.68)	16.49 (21.35)	<.001	0.076 (0.02-0.16)	.02	0.05
Mean change	7.065 (26.29)	−0.538 (20.03)	.03	0.026 (0.00-0.09)	—	—
Insomnia						
Baseline	25.25 (27.80)	17.82 (25.19)	.049	0.019 (0.00-0.07)	—	—
1 year	22.58 (27.44)	15.05 (22.26)	.04	0.022 (0.00-0.08)	.25	0.07
Mean change	−1.792 (25.34)	−1.792 (23.24)	>.99	0.00 (0.00-0.00)	—	—
Financial difficulties						
Baseline	10.77 (19.53)	5.94 (17.25)	.06	0.017 (0.00-0.07)	—	—
1 year	13.26 (22.60)	3.94 (14.62)	.001	0.057 (0.01-0.13)	.02	0.03
Mean change	2.151 (18.26)	−0.717 (14.73)	.24	0.007 (0.00-0.05)	—	—
Dyspnea in rest					—	—
Baseline	1.35 (6.60)	4.29 (11.22)	.02	0.025 (0.00-0.008)		
1 year	3.23 (9.91)	2.51 (12.27)	.66	0.001 (0.00-0.03)		
Mean change	1.792 (9.02)	−1.075 (14.29)	.10	0.014 (0.00-0.07)		
Pain in other parts						
Baseline	33.33 (32.97)	26.53 (30.26)	.14	0.012 (0.00-0.06)	—	—
1 year	37.12 (35.53)	24.73 (30.26)	.01	0.035 (0.002-0.10)	.06	0.02
Mean change	2.811 (30.45)	−1.482 (35.30)	.40	0.004 (0.00-0.04)	—	—

[a]QLQ-C30: Quality of Life Questionnaire Core 30.
[b]QLQ-LC13: Quality of Life Questionnaire Lung Cancer 13.
[c]Comparison was carried out based on means at individual time points, changes from baseline, and means adjusted to baseline values.
[d]Control: smoking cessation with written material.
[e]App: smartphone app–based smoking cessation.
[f]ANOVA.
[g]Partial eta-square.
[h]Analysis of covariance adjusted for baseline QLQ-C30 symptom scores.
[i]Not applicable.

XSL•FO
RenderX

Table . EQ-5D–specific symptom scores and their change between baseline and 1 year according to prespecified classification of the questionnaire according to randomization arm[a].

	Control[b], n (%)	App[c], n (%)	*P* value[d]	Effect size[e] (95% CI)	Adjusted *P* value[f]	Adjusted effect size[f]
Mobility						
Baseline			.86	0.00 (0.00-0.017)	—[g]	—
No problems	61 (62)	61 (60)				
Some problems	38 (38)	40 (40)				
1 year			.02	0.031 (0.001-0.09)	.008	0.037
No problems	48 (52)	64 (69)				
Some problems	45 (48)	29 (31)				
Pain or discomfort						
Baseline			.46	0.006 (0.00-0.04)	—	—
No pain or discomfort	45 (46)	52 (52)				
Moderate pain or discomfort	51 (52)	58 (48)				
Extreme pain or discomfort	3 (3)	1 (1)				
1 year			.01	0.049 (0.006-0.12)	.007	0.026
No pain or discomfort	36 (39)	55 (59)				
Moderate pain or discomfort	50 (54)	36 (39)				
Extreme pain or discomfort	7 (8)	2 (2)				

[a]Comparison was carried out based on symptom scales.
[b]Control: smoking cessation with written material.
[c]App: smartphone app–based smoking cessation.
[d]Pearson chi-square test.
[e]ANOVA.
[f]Analysis of covariance adjusted for baseline EQ-5D symptom scores.
[g]Not applicable.

Frequency of App Use and Its Association to HRQoL

To further support that our observed improvements result from app use, we analyzed the connection between app use and changes in GHS and EQ-5D index scores. We assessed app use by the number of questionnaires the user had completed in the app over the course of the first 24 weeks. There was a statistically significant difference between the mean number of filled questionnaires by those whose EQ-5D index score showed no change or improved (partial η^2=0.073, 95% CI 0.00 - 0.18; *P*=.04; partial η^2=0.071; adjusted *P*=.04). In the analysis of GHS and app use, nonsignificant trends similar to those of the EQ-5D were observed (Table 8).

XSL•FO
RenderX

Table . The extent of app use in relation to health-related quality of life[a].

	Improved	No change	Declined	*P* value[b]	Effect size[b]	Adjusted *P* value[c]	Adjusted effect size[c]
QLQ-C30[d] GHS[e] (10% change), n (%)	21 (23)	55 (60)	15 (16.5)	—[f]	—	—	—
Completed questionnaires, mean (SD)	11.14 (14)	6.22 (7)	5.73 (6)	.08	0.056 (0.00-0.16)	—	—
EQ-5D index score (5% change), n (%)	29 (32)	43 (47)	19 (21)	—	—	—	—
Completed questionnaires, mean (SD)	9.86 (11)	7.40 (8)	3.05 (4)	.04	0.073(0.00-0.18)	.04	0.071

[a]An analysis of EQ-5D index score (5%) or QLQ-C30 (10%) change in the experimental arm. The use of the app was assessed by the number of questionnaires the user had completed in the smoking cessation app over the course of the first 24 weeks.

[b]One-way ANOVA.

[c]Analysis of covariance adjusted for age, sex, pack years.

[d]QLQ-C30: Quality of Life Questionnaire Core 30.

[e]GHS: global health status.

[f]Not applicable.

Discussion

Principal Findings

To our knowledge, LDCT-SC-FI was the first to report the results of a smartphone app–based smoking cessation in a randomized controlled trial setting among individuals participating in LDCT lung cancer screening. In this study, we investigated the changes in HRQoL in participants undergoing LDCT screening for lung cancer over a 1-year study period using 3 different PROMs. Additionally, we analyzed the effects of smoking cessation as well as the use of our interventional smoking cessation app on HRQoL outcomes. According to our results, neither LDCT screening for lung cancer nor smoking cessation was associated with changes in HRQoL, while the individuals randomized to the Suunta smoking cessation smartphone app reported improved HRQoL mainly by decreased pain and mobility. In general, observed changes in mean values of HRQoL are small, but these, however, fulfill the population norm for significant change (±5%) using the EQ-5D measure. Our results suggest that the positive effects of smoking cessation apps can be comprehensive and may not be limited to cessation only.

In addition to early detection of lung cancer, LDCT screening can have various health effects, which need to be considered when initiating screening programs [2,3,9,11,12]. There is some evidence that LDCT screening may induce negative effects on participants' HRQoL, especially psychosocial in nature. In most studies, these effects have been short-term and reversed over time [11,13,14]. Our results are in line with these findings, suggesting that lung cancer screening does not negatively impact HRQoL. Since our study had a limited number of participants, we were not able to study the correlation between screening results and HRQoL.

Smoking is the leading risk factor for lung cancer, and smoking cessation clearly prevents the disease as well as improves outcomes [1]. Lung cancer screening programs offer a superb opportunity for smoking cessation interventions and enhance their efficacy [4-6,8]. We have recently shown that smoking cessation can be enhanced 3-fold, and it is of interest if this would translate to alter HRQoL [20]. However, our results showed that smoking cessation does not have an impact on HRQoL or specific scales or symptoms, even though we had hypothesized the contrary. Reflecting on the previous literature, our negative results are not surprising since positive HRQoL effects of smoking cessation might require years to develop [24,27].

Comparison to Prior Work

Currently, smartphones are ubiquitous and provide an accessible platform for health interventions. In various medical conditions, including cancer, such interventions have shown beneficial effects on HRQoL among other study outcomes [33-35]. In recent years, various smartphone-based methods for smoking cessation have been developed, and some have proven to be effective [18,19]. To our knowledge, these studies have not assessed HRQoL, so our study brings novel information on this topic. Although the Suunta app was effective in smoking cessation and surpassed meta-analysis–based ORs for smoking cessation with mobile apps, we observed no association between cessation at 1 year and HRQoL [18,20]. Interestingly, our study showed that individuals randomized to the Suunta app arm presented with improved overall HRQoL as well as specific scales and symptoms. Improvements in HRQoL seemed to be driven by improvements in social function, mobility, as well as pain. Our findings suggest that the improved HRQoL is a result of a direct effect of Suunta app use, rather than this resulting due to smoking cessation. The effects of mobile health apps on

XSL•FO
RenderX

HRQoL have seldom been investigated, and it is possible that other apps would bear similar effects. Furthermore, HRQoL, or happiness, might also be measured with nonvalidated instruments [36]. These are interesting areas of future research.

The theoretical background of the Suunta app includes cognitive behavioral and social cognitive theories and commitment therapy as well as mindfulness. The participants may use the app for goal setting, decision-making, and personal empowerment in smoking cessation as well as for overall management of their health. Functionalities include a weekly symptom questionnaire with personalized feedback, mindfulness practices, self-reflection, and a guided smoking cessation plan. Observed improvements in functional scales might relate to these features, which aim to enhance self-reflection capabilities and induce empowerment.

It is unclear how the Suunta app relays its favorable effects on the pain symptom scores, observed with all 3 PROMs used in this study. To be precise, the pain symptom scores decreased in the app arm over the 1-year study period, whereas in the control arm, they increased. Since the sensation of pain is a complex phenomenon with a psychosocial component, features of the app, specifically mindfulness meditation–based exercises, might alter the overall experience of pain. There are evolving data that mindfulness meditation–based analgesia reduces pain through multiple, unique neural mechanisms and may lead to long-term effects [37]. Our results suggested that the app favorably affected generalized pain rather than regional or nociceptive pain, for example, chest pain and angina, which likely reflects its effect on the psychosocial, and maybe even neurophysiological components of pain.

We also analyzed the extent of app use in relation to HRQoL, and the more frequent use was associated with improved overall HRQoL. This favors the explanation that the features of and adherence to the app are involved in the changes observed. In addition, we have previously shown that adherence to app use was associated with increased smoking cessation [20]. Therefore, the means to increase app adherence could enhance the effects on both smoking cessation and HRQoL. Our results indicate that a mobile app can assist in achieving smoking cessation and retaining or even improving HRQoL during the LDCT screening for lung cancer.

Strengths and Limitations

Our study has some limitations. The study recruitment occurred only in a single center, and the number of participants was moderate, which might limit the generalizability of the results to a wider screening population. Smoking was assessed by self-reporting without biochemical verification, which may overestimate the number of nonsmokers. Yet, it has previously been shown that self-reported smoking cessation follows closely to biochemically verified data [38]. HRQoL was an exploratory end point of the trial, and analysis was not prespecified in detail, and the results should be interpreted with caution. In addition, HRQoL was analyzed only at 2 time points, which limits the time scope of the analysis. Nevertheless, in the context of the study population, the frequency of HRQoL assessments is reasonable. The credibility of our results is supported by using 3 different PROMs assessing HRQoL, all of which produced similar results. It has been suggested that combining generic HRQoL measures with more condition-specific ones may increase their overall usefulness [39]. In addition, the coverage of HRQoL questionnaires was very high because of good participant adherence. The smartphone app was offered to the controls at 6 months, which might have had an impact on the observed results. However, only 15% (14/93) of the controls reported app use at 1 year, and the observed HRQoL results were very similar to intention-to-treat when excluding these from analysis. As in all clinical trials, there were dropouts in our study. The dropout rate was 7.5%, which is very low compared to smoking cessation studies that often report rates as high as 50%. Furthermore, the dropout rates were very similar in both arms, and it is likely that dropouts have only a marginal impact on observed results.

Conclusions

We investigated the effect of LDCT lung cancer screening, smoking cessation, and a smoking cessation app on HRQoL. The LDCT-SC-FI trial was the first to investigate a smoking cessation mobile app in the context of LDCT; thus, the presented HRQoL results are unique. We were able to show that randomization to the smartphone app arm, as well as the frequency of app use, was associated with improved HRQoL, while participation in screening or smoking cessation had no effect. According to the study results, the beneficial impact of the app is driven by enhanced social functioning, mobility, and pain compared to controls. Our study suggests that a smartphone app aiming at smoking cessation is an effective intervention in the context of LDCT lung cancer screening since it can both enhance smoking cessation and improve HRQoL.

Acknowledgments

Generative artificial intelligence was not used in any portion of the manuscript.

Funding

This study was supported by AstraZeneca, Roche, and Cancer Foundation Finland. The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

Data collected for the study will be made available on a reasonable request to the corresponding author.

XSL•FO
RenderX

Authors' Contributions

JK, SI, HA, AJ, RK, and TV designed the study. JK, SI, and AK collected the data. JK, SI, and AK analyzed data and drafted the manuscript. All the authors read and approved the final version of the manuscript.

Conflicts of Interest

AK, HA, and AJ declare no conflict of interest. SI reports institutional grants from AstraZeneca and Roche for the conduct of this study. SI reports personal fees from MSD, Roche, BMS, AstraZeneca, Novartis, Takeda, and Eisai and lecture fees from Siemens Healthineers, all outside the submitted work. RK reports consulting, lecture, and advisory board fees from Boehringer Ingelheim and an advisory board fee from MSD outside the submitted work. TV reports advisory board fees from NordicInfu Care, MSD, and AstraZeneca outside the submitted work. JK reports institutional grants from AstraZeneca and Roche for the conduct of this study. JK reports a personal grant for the conduct of the study from Cancer Foundation Finland. JK reports personal fees from Roche, AstraZeneca, Janssen, BMS, Merck, Amgen, Novartis, Merck KgA, Sanofi, and Pfizer and lecturing fees from Siemens Healthineers, all outside of the submitted work. JK is a former part-time employee of Faron Pharmaceuticals.

Multimedia Appendix 1
Snapshots of the app.
[PDF File, 1108 KB - jmir_v28i1e81687_app1.pdf]

Multimedia Appendix 2
Study protocol.
[PDF File, 1516 KB - jmir_v28i1e81687_app2.pdf]

Multimedia Appendix 3
LDCT-SC-FI study demographics, and the integrity of the HRQoL questionnaire data.
[DOCX File, 39 KB - jmir_v28i1e81687_app3.docx]

Checklist 1
CONSORT-eHealth checklist (V 1.6.1).
[PDF File, 361 KB - jmir_v28i1e81687_app4.pdf]

References

1. Tesfaw LM, Dessie ZG, Mekonnen Fenta H. Lung cancer mortality and associated predictors: systematic review using 32 scientific research findings. Front Oncol 2023;13:1308897. [doi: 10.3389/fonc.2023.1308897] [Medline: 38156114]
2. de Koning HJ, van der Aalst CM, de Jong PA, et al. Reduced lung-cancer mortality with volume CT screening in a randomized trial. N Engl J Med 2020 Feb 6;382(6):503-513. [doi: 10.1056/NEJMoa1911793] [Medline: 31995683]
3. National Lung Screening Trial Research Team, Aberle DR, Adams AM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011 Aug 4;365(5):395-409. [doi: 10.1056/NEJMoa1102873] [Medline: 21714641]
4. Brain K, Carter B, Lifford KJ, et al. Impact of low-dose CT screening on smoking cessation among high-risk participants in the UK Lung Cancer Screening Trial. Thorax 2017 Oct;72(10):912-918. [doi: 10.1136/thoraxjnl-2016-209690] [Medline: 28710339]
5. Kauczor HU, Baird AM, Blum TG, et al. ESR/ERS statement paper on lung cancer screening. Eur Radiol 2020 Jun;30(6):3277-3294. [doi: 10.1007/s00330-020-06727-7] [Medline: 32052170]
6. van der Aalst CM, van Klaveren RJ, van den Bergh KAM, Willemsen MC, de Koning HJ. The impact of a lung cancer computed tomography screening result on smoking abstinence. Eur Respir J 2011 Jun;37(6):1466-1473. [doi: 10.1183/09031936.00035410] [Medline: 21148233]
7. Wood DE, Kazerooni EA, Aberle D, et al. NCCN Guidelines® Insights: Lung Cancer Screening, Version 1.2022. J Natl Compr Canc Netw 2022 Jul;20(7):754-764. [doi: 10.6004/jnccn.2022.0036] [Medline: 35830884]
8. Moldovanu D, de Koning HJ, van der Aalst CM. Lung cancer screening and smoking cessation efforts. Transl Lung Cancer Res 2021 Feb;10(2):1099-1109. [doi: 10.21037/tlcr-20-899] [Medline: 33718048]
9. Ratushnyak S, Hoogendoorn M, van Baal PHM. Cost-effectiveness of cancer screening: health and costs in life years gained. Am J Prev Med 2019 Dec;57(6):792-799. [doi: 10.1016/j.amepre.2019.07.027] [Medline: 31753260]
10. Bonney A, Malouf R, Marchal C, et al. Impact of low-dose computed tomography (LDCT) screening on lung cancer-related mortality. Cochrane Database Syst Rev 2022 Aug 3;8(8):CD013829. [doi: 10.1002/14651858.CD013829.pub2] [Medline: 35921047]

XSL•FO
RenderX

11. Gareen IF, Duan F, Greco EM, et al. Impact of lung cancer screening results on participant health-related quality of life and state anxiety in the National Lung Screening Trial. Cancer 2014 Nov 1;120(21):3401-3409. [doi: 10.1002/cncr.28833] [Medline: 25065710]
12. Jonas DE, Reuland DS, Reddy SM, et al. Screening for lung cancer with low-dose computed tomography: updated evidence report and systematic review for the US Preventive Services Task Force. JAMA 2021 Mar 9;325(10):971-987. [doi: 10.1001/jama.2021.0377] [Medline: 33687468]
13. van den Bergh KAM, Essink-Bot ML, Borsboom G, Scholten ET, van Klaveren RJ, de Koning HJ. Long-term effects of lung cancer computed tomography screening on health-related quality of life: the NELSON trial. Eur Respir J 2011 Jul;38(1):154-161. [doi: 10.1183/09031936.00123410] [Medline: 21148229]
14. Wu GX, Raz DJ, Brown L, Sun V. Psychological burden associated with lung cancer screening: a systematic review. Clin Lung Cancer 2016 Sep;17(5):315-324. [doi: 10.1016/j.cllc.2016.03.007] [Medline: 27130469]
15. Naughton F, Nadkarni A. The contribution of digital treatment to efforts to reduce global tobacco use. N Engl J Med 2025 Oct 16;393(15):1449-1452. [doi: 10.1056/NEJMp2500683] [Medline: 41085052]
16. Perski O, Hébert ET, Naughton F, Hekler EB, Brown J, Businelle MS. Technology-mediated just-in-time adaptive interventions (JITAIs) to reduce harmful substance use: a systematic review. Addiction 2022 May;117(5):1220-1241. [doi: 10.1111/add.15687] [Medline: 34514668]
17. Szinay D, Cameron RA, Jones A, et al. Eliciting preferences for the uptake of smoking cessation apps: discrete choice experiment. J Med Internet Res 2025 Jan 14;27:e37083. [doi: 10.2196/37083] [Medline: 39808479]
18. Sha L, Yang X, Deng R, et al. Automated digital interventions and smoking cessation: systematic review and meta-analysis relating efficiency to a psychological theory of intervention perspective. J Med Internet Res 2022 Nov 16;24(11):e38206. [doi: 10.2196/38206] [Medline: 36383408]
19. Guo YQ, Chen Y, Dabbs AD, Wu Y. The effectiveness of smartphone app-based interventions for assisting smoking cessation: systematic review and meta-analysis. J Med Internet Res 2023 Apr 20;25:e43242. [doi: 10.2196/43242] [Medline: 37079352]
20. Iivanainen S, Kurtti A, Wichmann V, et al. Smartphone application versus written material for smoking reduction and cessation in individuals undergoing low-dose computed tomography (LDCT) screening for lung cancer: a phase II open-label randomised controlled trial. Lancet Reg Health Eur 2024 Jul;42:100946. [doi: 10.1016/j.lanepe.2024.100946] [Medline: 39070744]
21. Becoña E, Vázquez MI, Míguez MDC, et al. Smoking habit profile and health-related quality of life. Psicothema 2013;25(4):421-426. [doi: 10.7334/psicothema2013.73] [Medline: 24124772]
22. Coste J, Quinquis L, D'Almeida S, Audureau E. Smoking and health-related quality of life in the general population. Independent relationships and large differences according to patterns and quantity of smoking and to gender. PLoS ONE 2014;9(3):e91562. [doi: 10.1371/journal.pone.0091562] [Medline: 24637739]
23. Mesquita R, Gonçalves CG, Hayashi D, et al. Smoking status and its relationship with exercise capacity, physical activity in daily life and quality of life in physically independent, elderly individuals. Physiotherapy 2015 Mar;101(1):55-61. [doi: 10.1016/j.physio.2014.04.008] [Medline: 25108641]
24. Moayeri F, Hsueh YSA, Dunt D, Clarke P. Smoking cessation and quality of life: insights from analysis of longitudinal Australian data, an application for economic evaluations. Value Health 2021 May;24(5):724-732. [doi: 10.1016/j.jval.2020.11.022] [Medline: 33933242]
25. Taylor G, McNeill A, Girling A, Farley A, Lindson-Hawley N, Aveyard P. Change in mental health after smoking cessation: systematic review and meta-analysis. BMJ 2014 Feb 13;348(feb13 1):g1151. [doi: 10.1136/bmj.g1151] [Medline: 24524926]
26. Tomioka H, Sekiya R, Nishio C, Ishimoto G. Impact of smoking cessation therapy on health-related quality of life. BMJ Open Respir Res 2014;1(1):e000047. [doi: 10.1136/bmjresp-2014-000047] [Medline: 25478191]
27. Shields M, Garner RE, Wilkins K. Dynamics of smoking cessation and health-related quality of life among Canadians. Health Rep 2013 Feb;24(2):3-11. [Medline: 24257905]
28. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
29. ICH E6 (R3) guideline for good clinical practice (GCP). EMA/CHMP/ICH/135/1995. European Medicines Agency. URL: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e6-r3-guideline-good-clinical-practice-gcp-step-5_en.pdf [accessed 2026-01-13]
30. Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. J Clin Oncol 1998 Jan;16(1):139-144. [doi: 10.1200/JCO.1998.16.1.139] [Medline: 9440735]
31. Koskinen S, Lundqvist A, Ristiluoma N. Terveys, toimintakyky ja hyvinvointi suomessa 2011 [Article in Finnish]. : THL; 2012 68/2012 URL: https://www.julkari.fi/handle/10024/90832 [accessed 2025-01-13]
32. Ohinmaa A, Sintonen H. Inconsistencies and modelling of the Finnish EuroQol (EQ-5D) preference values. : EuroQol Plenary Meeting, Health Economics and Health System Research, University of Hannover; 1998 Jan 10 URL: https://eq-5dpublications.euroqol.org/download?id=0_53500&fileId=53923 [accessed 2025-01-13]

XSL•FO
RenderX

33. Chung C, Lee JW, Lee SW, Jo MW. Clinical efficacy of mobile app-based, self-directed pulmonary rehabilitation for patients with chronic obstructive pulmonary disease: systematic review and meta-analysis. JMIR Mhealth Uhealth 2024 Jan 4;12:e41753. [doi: 10.2196/41753] [Medline: 38179689]
34. Öztürk ES, Kutlutürkan S. The effect of the mobile application-based symptom monitoring process on the symptom control and quality of life in breast cancer patients. Semin Oncol Nurs 2021 Jun;37(3):151161. [doi: 10.1016/j.soncn.2021.151161] [Medline: 34088557]
35. Thiengwittayaporn S, Wattanapreechanon P, Sakon P, et al. Development of a mobile application to improve exercise accuracy and quality of life in knee osteoarthritis patients: a randomized controlled trial. Arch Orthop Trauma Surg 2023 Feb;143(2):729-738. [doi: 10.1007/s00402-021-04149-8] [Medline: 34453570]
36. Arigayota A, Duffek B, Hou C, Eisingerich AB. Effects of The Legend of Zelda: Breath of the Wild and Studio Ghibli Films on young people's sense of exploration, calm, mastery and skill, purpose and meaning, and overall happiness in life: exploratory randomized controlled study. JMIR Serious Games 2025 Aug 1;13(1):e76522. [doi: 10.2196/76522] [Medline: 40750097]
37. Eich W, Diezemann-Prößdorf A, Hasenbring M, et al. Psychosocial factors in pain and pain management: a statement. Schmerz 2023 Jun;37(3):159-167. [doi: 10.1007/s00482-022-00633-1] [Medline: 35303149]
38. Auer R, Schoeni A, Humair JP, et al. Electronic nicotine-delivery systems for smoking cessation. N Engl J Med 2024 Feb 15;390(7):601-610. [doi: 10.1056/NEJMoa2308815] [Medline: 38354139]
39. Payakachat N, Ali MM, Tilford JM. Can the EQ-5D detect meaningful change? A systematic review. Pharmacoeconomics 2015 Nov;33(11):1137-1154. [doi: 10.1007/s40273-015-0295-6] [Medline: 26040242]

Abbreviations

CT: computed tomography
EORTC: European Organisation for Research and Treatment of Cancer
ES: effect size
GHS: global health status
HRQoL: health-related quality of life
LDCT: low-dose computed tomography
LDCT-SC-FI: low-dose computed tomography screening for lung cancer combined to different smoking cessation methods in Finland
NELSON: Nederlands-Leuvens Longkanker Screenings Onderzoek
OR: odds ratio
PROM: patient-reported outcome measure
QLQ-C30: Quality of Life Questionnaire Core 30
QLQ-LC13: Quality of Life Questionnaire Lung Cancer 13
QoL: quality of life

Edited by S Brini; submitted 01.Aug.2025; peer-reviewed by A Eisingerich, W Li; revised version received 07.Nov.2025; accepted 11.Nov.2025; published 20.Jan.2026.

Please cite as:
Kurtti A, Iivanainen S, Kaarteenaho R, Andersen H, Jekunen A, Vasankari T, Koivunen J
Effect of Lung Cancer Screening, Smoking Cessation, and Cessation Smartphone App to Health-Related Quality of Life Among Heavy Smokers: Randomized Controlled Trial
J Med Internet Res 2026;28:e81687
URL: https://www.jmir.org/2026/1/e81687
doi:10.2196/81687

© Antti Kurtti, Sanna Iivanainen, Riitta Kaarteenaho, Heidi Andersen, Antti Jekunen, Tuula Vasankari, Jussi Koivunen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Patient and Care Team Perspectives of Barriers to and Facilitators for the Implementation of a Digital Health Program for Depression in Primary Care: Qualitative Study

Andrea Nederveld[1], MPH, MD; Elise A Robertson[2,3], MA; Angela M Lanigan[2], MPA, RD, LD; Elisabeth F Callen[2,3], PhD; Tarin L Clay[2,3], BA; Ben Fehnert[4], MA; Lambros Chrones[5], MD; Michael L Martin[5], MBA, MD; Margaret McCue[5], MS, RD; Christina M Hester[2,3], MPH, PhD; Melissa K Filippi[2,6,7], MPH, PhD

[1]University of Colorado Anschutz Medical Campus, Aurora, CO, United States

[2]National Research Network, American Academy of Family Physicians, Leawood, KS, United States

[3]DARTNet Institute, 12635 E Montview Blvd Ste 129, Aurora, CO, United States

[4]Ctrl Group/Fora Health, London, United Kingdom

[5]Takeda Pharmaceuticals U.S.A., Inc., Lexington, MA, United States

[6]Robert Graham Center, American Academy of Family Physicians, Washington, DC, United States

[7]Department of Family Medicine, Georgetown University, Washington, DC, United States

Corresponding Author:
Christina M Hester, MPH, PhD
National Research Network, American Academy of Family Physicians, Leawood, KS, United States

Abstract

Background: Depression is pervasive, and rates are rising in the United States. Most people with depression receive care from primary care clinicians, but gaps in the quality of care exist. Team-based approaches to depression care have been shown to aid in treatment and management; yet, challenges exist in implementation. Digital health apps have been shown to be effective in improving depression symptoms and enhancing patient engagement in some populations. Many, however, do not share data with clinical care teams.

Objective: This study aimed to understand the barriers to and facilitators for implementation of a digital health program that supports coordinated use by clinical care teams and patients, via a mobile app and care team–facing web interface, for depression in primary care.

Methods: This study was part of a larger intervention study that included 4 primary care practices: 2 intervention and 2 control sites. The intervention sites used a patient-facing mobile app and a care team–facing web interface, and the control sites continued usual care. The study team conducted interviews from May to October 2021. Patient and care team participants were recruited toward the end of their study involvement. Separate semistructured interview guides were developed for patient and care team participants. Interviews were recorded and transcribed. Data were coded using Atlas.ti.9, and data analysis was completed using a grounded theory approach.

Results: Interviews with patient (n=8) and care team (n=8) participants revealed 3 main topics for program implementation: app/interface usability, tracking, and program recommendations. For app/interface usability, overall, navigation for both patient and care team participants was simple and straightforward. Although app content was relevant, patient participants desired additional educational resources and information to aid in their depression treatment and management. In terms of tracking, care team participants indicated that data obtained via the interface enabled them to monitor patients in between visits; and in some circumstances, these data facilitated conversations with patients about treatment plans. Tracking medication adherence differed among patient participants due to established routines of taking medications consistently, lack of motivation to track, or lack of interest in tracking. Care team participants reported the ability to respond more quickly to side effects. Patients commented on tracking difficulties: confusing response options, insufficient goal attainment response options, not being able to provide details or write notes, and no ability to change goals. Some patient and care team participants perceived that tracking encouraged communication with one another; others perceived tracking as having no impact on shared decision-making.

Conclusions: Results suggest implementation of a digital health program for depression treatment and management in primary care practices could impact patient medication adherence, produce faster turnaround time for medication optimization, encourage goal setting, and foster communication between patients and care team members. Program enhancements could optimize patient and care team member engagement.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e72003) doi:10.2196/72003

KEYWORDS

depression treatment; digital health; family physicians; goal setting; depression; major depressive disorder; MDD; mobile health; primary care; qualitative; shared decision-making

Introduction

Primary care clinical teams play a critical role in depression diagnosis, treatment, and management. Rates of depression in the United States have risen since the beginning of the COVID-19 pandemic [1]. Up to 70% of patients being treated for behavioral and mental health conditions, including depression, receive care within primary care [2], and approximately 13% of primary care visits involved a patient with a depression diagnosis between 2010 and 2018 [3]. Clinician comfort with treating depression varies [4], and gaps in quality of care exist; notably in areas such as screening, diagnosis, patient engagement and education, clinician training, and clinician follow-up [5-10]. There is evidence that team-based care and care management between clinic visits, as well as patient self-management [11], can improve depression outcomes [12,13]. However, these strategies can be time and resource intensive for primary care practices, requiring regular check-ins with patients and providing education on self-management. Encouraging patients to use digital health apps is one strategy to assist patients in better self-management.

Health care providers can play an integral role in patient initiation of use of a digital or mobile app [14]. Digital health apps are acceptable to patients, easy to use, and potentially effective in supporting behavior change and self-management, particularly in improving depression and anxiety symptoms in various populations [15-19]. However, there is a lack of standardization in evaluating health apps [20,21], and many behavioral health apps are not able to share data with clinical care teams [20].

This pilot study examined a digital health program that included a patient-facing mobile app and care team web interface that support depression treatment and symptom management in relation to shared decision-making and goal setting in primary care practices. Here, we describe patient and care team participant perspectives regarding barriers to and facilitators for using the digital health program in their practices.

Methods

Study Design, Setting, and Recruitment

This study was part of a larger intervention study (currently under review) that included 4 primary care practices, located in Michigan (n=1), Florida (n=1), and New York (n=2), all with some level of behavioral health integration. The intervention arm (n=2 practices, 1 residency and 1 private practice) piloted the use of the Primary Care Path, a program with a patient-facing mobile app and accompanying care team–facing web interface, which was developed collaboratively by Takeda and American Academy of Family Physicians National Research Network (AAFP NRN) and powered by Fora Health (Ctrl Group Limited UK). The control arm (n=2 practices; 1 Community Health Center and 1 private practice) continued usual care for depression. Participating practices were recruited from the AAFP NRN. Patient interview participants were recruited from the patients enrolled in the intervention arm via email invitation in the final weeks of their study involvement. Practice champion clinicians, study coordinators, and clinicians who cared for patients enrolled in the study were invited to participate postintervention.

Ethical Considerations

This study was approved by the American Academy of Family Physicians Institutional Review Board (Protocol 20 - 389). All participants consented to be in the study. However, those participants who took part in the intervention were sent an email asking if they would be willing to participate in an interview. Those who responded received an invitation for a scheduled interview. Prior to the start of the interview, a study team member thoroughly described participation: expected interview length, questions asked, permission to record, and participant rights (voluntary participation, can terminate participation at any time without penalty, remuneration, and possible risks). Risks, such as discomfort answering certain questions and the potential loss of anonymity and confidentiality, were explained. Verbal consent was then obtained. On completion of the interview, participants received compensation; a US $100 e-gift card was sent to their email. All identifying information was removed from interviews and study data to protect participant anonymity.

Intervention

The Fora Health app supports goal setting and tracks daily medication adherence and side effects, weekly goal attainment, and behavioral health participation; assesses the 9-item Patient Health Questionnaire (PHQ-9) [22,23], 13-item Patient Activation Measure [24,25], World Health Organization-Five Well-Being Index [26,27], and Perceived Deficits Questionnaire [28,29] on a biweekly basis; and provides short educational tutorials about depression management on topics such as talking with your physician and goal setting. Patient participants began using the Primary Care Path app (powered by Fora Health) after meeting with the study coordinator to enroll and set goals. The Primary Care Path web interface allows staff and clinicians to see patient-reported data. The expectation was that the care team would use the information before visits or review between visits.

Analysis

Separate semistructured interview guides were developed by the AAFP NRN study team for patient and care team participants, according to the RE-AIM (reach, effectiveness, adoption, implementation, and maintenance) evaluation framework [30-32]. The interview guides were reviewed and revised by the broader study team. Interviews were conducted from May to October 2021 by a family physician (AN) and 2 research project managers (AML and EAR), each with a

XSL•FO
RenderX

master's degree and training and experience in qualitative methods. Interviews were conducted online and lasted approximately 30 minutes, and were audio recorded and transcribed. Four research team members (MKF, AN, AML, and EAR) met frequently over 13 weeks to develop the codebook, using a priori and emergent codes. Meaning, some codes were generated based on the discussion guide, and others were developed through open coding, where codes emerge from the data. After the research team members agreed on the codes and their definitions, each transcript was coded once (MKF and EAR) using Atlas.ti.9 (Berlin and Germany), and an audit was conducted (AML) to ensure intercoder reliability. This inductive approach to data analysis was completed by 4 members of the research team (MKF, AN, AML, and EAR), using a combination of grounded theory and constant comparative methods [33,34]. Themes emerged through repeated reading of the data. Themes were discussed and agreed upon by the broader research team.

Results

Comparing Patient and Care Team Perspectives

The study team interviewed intervention patients (n=8) and intervention care team participants (lead clinicians, study coordinators, and other clinicians; n=8, shown in Tables 1 and 2). The sample size was determined based on the 2 participant categories of the sampling frame and the sample's sufficiency to provide a meaningful perspective and nuanced detail [35]. Data comparisons between the patient and care team participant strata showed similarities and differences for topics of app/interface usability (navigation and content and usability), tracking (medication adherence, side effects, goals, suggestions for additional tracking features, and communication), and program recommendations. Saturation was reached within the patient stratum as there was consistency across all patient data. Even though much of the data indicates overall agreement, additional explanations are provided where discordance is present. Example quotations are displayed at the end of each section.

Table . Patient demographics.

Patient participant	Sex	Race	Ethnicity	Year of birth	Employment status	Living situation	Treatments used for depression
1	Female	White	Non-Hispanic or Latinx	1997	Part time	With spouse or other family	Antidepressant medication and behavioral health/talk therapy
2	Female	White	Non-Hispanic or Latinx	1986	Full time	Alone	Antidepressant medication, behavioral health/talk therapy, and self-help strategies (books, apps, or other tools)
3	Female	White	Non-Hispanic or Latinx	1982	Part time	With spouse or other family	Antidepressant medication
4	Female	White	Non-Hispanic or Latinx	1988	Part time	With friends and/or roommates	Antidepressant medication
5	Female	White	Non-Hispanic or Latinx	1957	Part time	Alone	Self-help strategies (books, apps, or other tools)
6	Male	No response	No response	1964	Retired	With spouse or other family	Antidepressant medication and self-help strategies (books, apps, or other tools)
7	Male	White	Hispanic or Latinx	1969	Full time	With spouse or other family	Antidepressant medication
8	Male	White	Non-Hispanic or Latinx	1962	Part time	Alone	No response

XSL•FO
RenderX

Table . Care team demographics.

Care team participant	Sex	Race	Ethnicity	Birth year	Years in practice	Role/credential
1	Female	White	Non-Hispanic or Latinx	1993	1 - 5	BH[a] care team
2	Female	White	Non-Hispanic or Latinx	1970	16 - 20	MD[b]
3	Female	White	Non-Hispanic or Latinx	1984	1 - 5	BH care team, PhD[c]
4	Female	White	Non-Hispanic or Latinx	1991	1 - 5	DO[d]
5	Male	White	Non-Hispanic or Latinx	1984	6 - 10	MD
6	Male	Missing	Missing	Missing	Missing	Coordinator, Practice Manager
7	Male	White	Non-Hispanic or Latinx	1963	≥21	MD
8	Male	White	Non-Hispanic or Latinx	1989	6 - 10	APRN[e]

[a]BH: behavioral health.

[b]MD: doctor of medicine.

[c]PhD: doctor of philosophy.

[d]DO: doctor of osteopathic medicine.

[e]APRN: advanced practice registered nurse.

App/Interface Usability

Navigation

Overall, both patient and care team participants reported that the app/interface was simple and easy to use. Patient participants said entering data did not consume much time, making it practical to use in daily and weekly routines. They also liked the convenience of being able to enter data at any time. Care team participants stated that reporting, monitoring, viewing, and inputting information (eg, joint goals) in the interface was straightforward. Particularly, care team participants liked the red flag next to the patient ID for the suicidality alert for patients who responded positively to the PHQ-9 question 9, making it easy to identify which patients needed follow-up.

> *I think how it emailed us about the [suicidality] alerts helped. Of course, as you would log on to the app, you would see that there was an alert on a specific patient's chart and then being able to dismiss it once that concern was addressed.* [Care team participant]

> *Oh, it was definitely easy. It was very well designed, very clear. It never crashed. I never had an issue with it on my phone.* [Patient participant]

Content and Usability

In terms of app content, patient participants thought the app was a dependable source of information. However, patient participants overwhelmingly stated they desired additional content such as videos and guides. They said the app was static, repetitive, and compliance-driven, with too much focus on medication adherence. They reported a desire for personalized insights regarding their own depression and targeted information on interpreting their results over time, as well as content options that evolve as their treatment changes. They also suggested displaying how much time an action takes to complete (watching videos, reading guides, etc). Care team participants who spoke with their patients specifically about the app responded similarly to patient participant impressions—that content was appropriate, but they wanted more substance. Care team participants also suggested expanding resources for patients beyond depression care and management to include topics such as substance use disorders and additional information for care team members, such as education on prescribing medications and using goal-setting techniques with patients. Some care team members liked that the program tenets of shared decision-making and goal setting were incorporated in the app/interface.

Some care team participants stated that the data and insights were helpful as they could monitor patients between visits via the interface. They said that seeing patterns and trends helped or could help facilitate conversations, particularly if a change was needed, such as medication, therapy, or goals. The residency practice specifically used the data in daily huddles to better prepare for appointments. At this practice, a few point people regularly logged into the interface, but many care team participants only reviewed printed reports. For those who did use the interface, a feature they appreciated was the ability to view the level of patient involvement and their activities, allowing insight into how patient participants were using and applying the tenets of the program. Some care team participants stated that they believed the content helped patients pay attention and be more mindful of their depression care and management; yet, they were also unsure if the data they viewed really

XSL•FO
RenderX

contributed to better patient outcomes, treatment, or symptom management.

Care team participants appreciated seeing the results of the PHQ-9, which initiated a specific alert to a designated care team member if the patient responded positively to question 9 (suicidal ideation), particularly if this was not a concern at the previous visit. However, they thought patient outreach could be more efficient. For example, PHQ-9 information was not in the medical chart, so they had to create a note and send it to other care team members. Even though care team participants thought the PHQ-9 alerts were appropriate, they were concerned about patient truthfulness, thinking patient participants may not want to be contacted about their answers in the future, which could impact how they answered questions.

> *My only thing that I couldn't figure out how to fix was every once in a while, there would be the small four-minute guides about setting goals with your caregiver, those ones. The last month and a half I was using it, it prompted me with the same guide every two days for a month and a half.* [Patient participant]

> *When it comes to features, about the only thing that I think would be useful is if there were any relevant articles, studies, videos or anything like that, that could be useful, that could be recommended or something like that. Just different ways to engage the user and to put more resources, basically.* [Patient participant]

Tracking

Medication Adherence

Patient participants who had never taken medication before stated that the daily medication question was particularly helpful, reporting that the daily check-in helped them think about goals and well-being more frequently. They reported being more mindful of their depression treatment and care throughout the week. However, for patient participants who already had a routine for taking medication, the reminders were not particularly helpful. Care team participants reiterated that for patients who need help establishing routines, reminders seemed to help with medication adherence.

> *I think it's really helpful if you're not good at taking your medication. I think that that is the number—I take it every night before bed, so I'm pretty strict in that routine. I can see that if it's something you're not used to doing, it being really helpful for that. It was really simple to use. If there were older users that weren't comfortable with technology, it would be really easy to navigate for them.* [Patient participant]

Overall, patient participants liked the option for tracking medication use, as it showed day-to-day adherence. Some patient participants viewed even small changes in their depression symptoms as motivation to continue taking medication. However, some patient participants did not track their medication use, or they only tracked it for a short period of time because they either already took medications consistently, lacked motivation to track, or were not interested in tracking. Patient and care team participants thought the act of tracking medication use contributed to a sense of accountability in depression treatment and care. However, patient participants mentioned that when they switched medications or changed dosage, it was difficult to account for those changes in the app. They also wanted questions streamlined to save time and avoid repetition. Finally, some care team participants questioned whether this approach changed medication adherence.

> *I think the biggest way it [tracking] probably affected their treatment was just actually (a) holding them a bit more accountable for what they were trying to do on their end, but (b) just seeing them check in daily with themselves or however often they were doing it...I think being able to be a little more retrospective more frequently helped them gain more insight too into how they were doing.* [Care team participant]

> *No, it [the program] didn't look like it [changed medication adherence]. It looked like we saw the same things. Some people were consistently taking their meds, some people who had some side effects stopped without telling us and would show up for the next visit. That didn't look like it changed much.* [Care team participant]

Side Effects

Tracking side effects was reported to be particularly helpful as patient participants often were not confident in differentiating between medication side effects and depression-related symptoms. For some patient participants, tracking side effects led to new discussions with their care team or therapist. Some care team participants stated they responded or could respond more quickly to side effects because patients were tracking them in the app, stating this was the most useful data collected because of in-depth, near real-time information, which allowed for timely adjustment of medications. Furthermore, in the residency practice, the data were used to help resident physicians make care decisions.

> *They really like the fact that if they were experiencing side effects that they could note it in the app, and their care team would be able to see it sooner before their next visit. It also allowed us to be aware of what's going on and make necessary changes if we needed to, if the side effects were that uncomfortable, or just to see if their PHQ was increasing that a change needed to be made.* [Care team participant]

> *It'd make you cognizant of the fact that it's something that you need to monitor, need to stay on top of. I think what it really helps is when it comes—'cause it makes you think about any side effects you my have had, and you may be having side effects but not necessarily attribute 'em to what you're takin' medication for, so it's a way of makes you think, and it's like, you know what? I did have this today, or I did go do that. Did that have something to do with it? At very least, it makes you think about.* [Patient participant]

XSL•FO
RenderX

Goals

In general, both patient and care team participants found value in setting and tracking goals. Patient participants stated it was helpful to set simple and reachable goals, and they could apply goal setting to other areas in their depression self-management beyond medication adherence (eg, exercising, spending time with family, reading, etc). In addition, patient participants appreciated seeing their progress displayed concretely in the app. Care team participants also thought that the tracking helped facilitate patient participants' depression self-management. Some care team participants reported that they had used goal setting as part of their normal care prior to study involvement, but the app/interface further facilitated the process of setting and tracking goals. They perceived the program, via the app/interface, as providing structure and consistency to ensure that goal setting was incorporated into discussions with patients. One care team participant stated that when patient participants used the app, it made conversations about setting goals easier, and it helped patient and care team participants to be on the same page.

Patient and care team participants also noted some hurdles with the goal-setting features. Patient participants discussed the difficulty of tracking goals within the app, including confusing responses (eg, goal benchmarks that did not make sense due to care team user error) or insufficient goal attainment responses. They wanted the ability to modify goals or remove outdated goals. Some patient participants reported feeling worse when reminded of a goal they were not meeting or that was no longer aligned with their current plan. Care team participants agreed that patient participants had difficulty managing goals within the app regarding frequency, milestones, and attainment selection.

> *Like I said, my favorite part of it is that you guys have those goals on there, which is typically something that I really like to do with patients and engaging them in shared decision-making, problem-solving with them, and goal setting to help accomplish their goals.* [Care team participant]

> *Yeah, one of our patient's weight was a big thing tied to her depression. She felt like the more she gained weight, the more depressed she felt. One of the things that really helped her was making goals in terms of exercising. She said having the goal of exercise by walking outside three times a week definitely really helped her become more engaged. She did lose, I believe, about 10 pounds, 15 pounds, something around there, which also improved her mood as well 'cause she was seeing that she was making these different changes.* [Care team participant]

> *That [questions about goals] was a little bit difficult to answer because my goal was to read twice a week for 15 minutes. The language of the app was, "Did you complete it once or did you complete it twice?" I wasn't sure because I was doing it twice already, should I say I did it once or should I say I did it twice? Sometimes just due to the language of the app, it didn't fit exactly what I was trying to do. Maybe a better way to see your progress with goals, rather than just checking off that you did it, maybe something like that.* [Patient participant]

Suggestions for Additional Tracking Features

Patient participants suggested streamlining questions to save time and avoid repetition. They also wanted to add some additional tracking features (eg, other self-management behavior that did not involve medication, such as physical activity, emotions, or well-being, with a place to write notes). Some desired a way to record more details on their progress. Care team participants said it was difficult to interpret the accuracy of the log (eg, medication compliance) and incomplete information (eg, missing data could indicate noncompliance or forgetfulness).

Communication

Patient and care team participants expressed that tracking encouraged or had the potential to encourage communication with one another. Patient participants liked the possibility of checking in with care team participants, especially without an appointment. Some felt that inputting data prior to appointments made the time together more efficient. Some care team participants also thought the appointments were easier and more effective when patients had tracked medication and goals. Care team participants stated that tracking enabled them to reference data before or during visits, highlight potential concerns, reassure alignment with the patient, and use information in follow-up calls. Some patient participants thought that the app had no impact on shared decision-making for depression treatment and management. Other patient participants stated it had made a difference, although some were unsure if a care team member viewed their information beforehand. Some patient and care team participants thought tracking impacted their relationship with one another; tracking reinforced feelings that the care team was invested in patient well-being, and the patient was invested in their own care. Care team participants thought that patient participants were more involved and proactive in their own care, and some believed patient engagement improved in the process.

> *I don't know that it cured the symptoms or anything like that. I just think that it was helpful and that I can keep that communication going with my team, so they knew what was going on.* [Patient participant]

> *This provided data points in between visits, especially while we're titrating medications because often we're starting a new [medication], changing something, changing a dosing, and this allowed communication points in between those visits or ways to track that data outside of these. We often will have them come back in six weeks after a medication change has been made. This can allow for interim information to help making a decision moving forward too.* [Care team participant]

App Use Recommendations

Patient and care team participants believed using the app could benefit several groups, namely those who are new to taking or switching medications, have trouble remembering to take

XSL•FO
RenderX

medications, are proactive in their care, want to make a behavioral change, need reminders, need accountability, and are seeking increased adherence. Some care team participants recommended targeting patients who needed closer follow-up and thought the app was more beneficial for those who were more diligent about their depression care and management, as the increased engagement gave them better insight into their health. Some care team participants indicated that using the app interface did not impact or change setting goals, reaching goals, making shared decisions, or patient outcomes. Reasons provided were either that their practice had implemented a patient-centered approach previously or that they needed data to compare before and after study participation to make a true assessment.

Care team participants thought the workflow around the interface could be improved. They overwhelmingly stated the need for electronic health record (EHR) integration because logging into the interface interrupted workflow, making it impractical during a patient appointment. Some care team participants admitted they did not access patient participant information: logging in was not well integrated into their workflow, they did not remember to log in, or they did not know how to log in. For example, one care team participant in the residency practice commented that it would have been helpful for the attending physician to see the data, and one patient participant stated that they wanted to share data with a psychiatrist outside of their primary care practice. Furthermore, some care team participants spoke about the need for app support within the practice. One care team participant stated that the practice had a dedicated behavioral health care specialist who carried out nonmedication interventions, and much of the program fell under their purview (eg, providing information to physicians and following up with patients). It was stressed that behavioral health is a team effort, and office staff are needed for integration.

Overall, patient participants who tracked and care team participants who used data to inform conversations saw value in shared decision-making and setting goals and thought these tenets could be applied to other conditions, such as anxiety, weight management, diabetes, high blood pressure, attention deficit disorder or attention-deficit/hyperactivity disorder, and medication compliance.

> *It really is about the EMR [electronic medical record], just because I'm spending so much time on it every day. It's always gonna be hard for me to have to login into something else outside of that when everything else is integrated.* [Care team participant]

Discussion

Principal Findings

Through this study, we identified barriers to and facilitators for implementation of a pilot program that incorporates a patient-facing mobile app and care team web interface to support depression treatment and management in primary care practices. These results on patient and care team participant perspectives of program implementation contribute to the rapidly expanding field of the use of digital health for depression care [17,36-45]. While some results between patient and care team participant strata were rather uniform, others differed. Our findings indicate that tracking medication adherence, side effects, and goals could allow both patient and care team participants greater insight into depression treatment and management. However, though the app may have helped some patient participants compile information for tracking, which may have led to more focused visits, not all care team participants thought app use improved patient outcomes. This may be more a matter of perception—perhaps patient participants who engaged regularly with the app thought they had better outcomes, possibly related to the perception of having more connectivity with their care team. Care team participants reported they could not definitively determine whether the app and program impacted patient outcomes, patient relationships, shared decision-making, goal setting, or goal attainment without tangible pre- and postintervention data. Nonetheless, medication adherence and goal setting are often related to depression outcomes and can be challenging [46-48]; this program could be another tool in their toolbox. Practices that already have patient-centered approaches to care may find aspects of this program duplicative; yet practices that do not prioritize approaches to goal setting and shared decision-making may experience more benefits if implementing this program, or a similar one.

Data and Action

App data in and of itself may not impact patient outcomes; it may be that the increased follow-up and touchpoints promote improvement in depression symptoms, as is evident in the literature [49,50]. The program was designed to connect patient participant data to the care team, allowing the care team to get real-time insight into patient symptoms and functioning between appointments. Thus, care team members who receive and review app data may be better able to recommend follow-up, suggest care changes, inform conversations, and shape next steps versus those who do not use additional data points collected outside of a care visit. The novel contribution of this study is that care team members who viewed near real-time information had an opportunity to make more timely decisions, especially for medication changes. Therefore, data insights provided to the care team via the interface may allow treatment pivots based on more accurate data on what is or is not working. This finding suggests that those care team members who accessed patient information in near real-time could make more prompt evidence-based decisions, which is particularly important for a patient with depression symptoms that impair day-to-day functioning.

Program Implementation

Even though care team participants appreciated aspects of the program, they agreed that successful implementation would need to be more integrated into existing workflows. It is unclear to what extent care team participants, especially physicians, accessed data prior to and during visits, and they often said that accessing a separate website was too time-consuming during a busy clinical day. This is consistent with other literature that demonstrates that app data integration into clinical workflows is often a requirement for use due to heavy workloads [51-55]. EHR integration would allow all care team members access, a

XSL•FO
RenderX

seamless transition for viewing data, and a comprehensive view of patient data. For example, if a patient scheduled an appointment for something unrelated to depression, a care team member would not necessarily review their depression data on a separate website. It is well documented that EHRs can help clinicians prepare for appointments, making them more efficient [56,57].

For uptake of this program, other workflow processes may need to be established depending on the practice, such as the level of behavioral health integration or resource support (eg, access to behavioral health professionals or referrals to community resources) [58-61]. Another consideration is designing the best approach for using data in conversations with patients [17]. Future iterations of this program could include additional education for care team members regarding roles, recommendations, and patient interactions.

Limitations

This study has limitations. First, our study sample was small and was racially and ethnically homogeneous, limiting the generalizability of the findings. Second, data saturation was reached in the patient stratum but not the care team member stratum. The care team contained diverse roles. We interviewed 8 primary care physicians and clinicians; their experiences varied. For example, a behavioral health professional or care coordinator may have had a touchpoint with the patient, and physicians may or may not have used the interface. Therefore, it is inconclusive as to whether physicians accessed and used the data before or during visits. To help remedy this situation, care team participants recommended embedding the interface data into the EHR. Third, to some extent, these data were influenced by patient and care team member engagement with the program. For example, if a patient used the app for a limited time span, they may not have tracked their behaviors long enough to see trends or for care team participants to make recommendations. Finally, a patient's depression cycle may have affected app use and program involvement. If depression symptoms were more prominent, patient participants may not have had the motivation to track medication adherence, side effects, or goals.

Conclusions

Implementation of a digital health program for depression treatment and management in primary care practices may help support patient medication adherence, facilitate timely medication changes, encourage goal setting, and foster communication between patients and care team members. While patient and care team participants valued program tenets, enhancements such as minimizing workflow interruptions, integrating data into the EHR, providing education and best practices for patient interactions, augmenting content with helpful resources, and adding personalized content that evolves alongside patient progression could increase program engagement.

Acknowledgments

The authors would like to acknowledge the collaboration of Takeda Pharmaceutical Company Limited with the primary care path developer Ctrl Group/Fora Health. They also want to acknowledge Takeda Pharmaceutical Company Limited for their collaborative study partnership, study design, and protocol development. The authors would also like to thank the practice team and patient participants who made this study possible. Generative artificial intelligence was not used in generating the study design, data collection, data analysis, or manuscript content.

Funding

Supported by funding from The Takeda Pharmaceutical Company Limited. This information or content and conclusions are those of the authors independent of the funder.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: LC, MM, CMH, MKF
Formal analysis: AN, EAR, AML, MKF
Funding acquisition: CMH
Investigation: EFC, TLC
Methodology: EFC, CMH, MKF
Project administration: AML
Supervision: CMH
Writing – original draft: AN, AML, MKF
Writing – review & editing: AN, EAR, EFC, TLC, BF, LC, MLM, MM, CMH, MKF

Conflicts of Interest

MM, MLM, and LC are or were employees of Takeda Pharmaceuticals at the time of the study. BF is an employee of Ctrl Group.

XSL•FO
RenderX

References

1. Ettman CK, Abdalla SM, Cohen GH, Sampson L, Vivier PM, Galea S. Prevalence of depression symptoms in US adults before and during the COVID-19 pandemic. JAMA Netw Open 2020 Sep 1;3(9):e2019686. [doi: 10.1001/jamanetworkopen.2020.19686] [Medline: 32876685]
2. Hunter CL, Goodie JL, Oordt MS, Dobmeyer AC. Integrated Behavioral Health in Primary Care: Step-by-Step Guidance for Assessment and Intervention: American Psychological Association; 2009. [doi: 10.1037/11871-000]
3. Jackson JL, Kuriyama A, Bernstein J, Demchuk C. Depression in primary care, 2010-2018. Am J Med 2022 Dec;135(12):1505-1508. [doi: 10.1016/j.amjmed.2022.06.022] [Medline: 35878693]
4. Chang ET, Magnabosco JL, Chaney E, et al. Predictors of primary care management of depression in the Veterans Affairs healthcare system. J Gen Intern Med 2014 Jul;29(7):1017-1025. [doi: 10.1007/s11606-014-2807-z] [Medline: 24567200]
5. Chew-Graham CA, Mullin S, May CR, Hedley S, Cole H. Managing depression in primary care: another example of the inverse care law? Fam Pract 2002 Dec;19(6):632-637. [doi: 10.1093/fampra/19.6.632] [Medline: 12429666]
6. Egede LE. Failure to recognize depression in primary care: issues and challenges. J Gen Intern Med 2007 May;22(5):701-703. [doi: 10.1007/s11606-007-0170-z] [Medline: 17370030]
7. Blackstone SR, Sebring AN, Allen C, Tan JS, Compton R. Improving depression screening in primary care: a quality improvement initiative. J Community Health 2022 Jun;47(3):400-407. [doi: 10.1007/s10900-022-01068-6] [Medline: 35076803]
8. Garcia ME, Hinton L, Neuhaus J, Feldman M, Livaudais-Toman J, Karliner LS. Equitability of depression screening after implementation of general adult screening in primary care. JAMA Netw Open 2022 Aug 1;5(8):e2227658. [doi: 10.1001/jamanetworkopen.2022.27658] [Medline: 35980633]
9. Kyanko KA, A Curry L, E Keene D, Sutherland R, Naik K, Busch SH. Does primary care fill the gap in access to specialty mental health care? a mixed methods study. J Gen Intern Med 2022 May;37(7):1641-1647. [doi: 10.1007/s11606-021-07260-z] [Medline: 34993864]
10. Samples H, Stuart EA, Saloner B, Barry CL, Mojtabai R. The role of screening in depression diagnosis and treatment in a representative sample of US primary care visits. J Gen Intern Med 2020 Jan;35(1):12-20. [doi: 10.1007/s11606-019-05192-3] [Medline: 31388917]
11. Dineen-Griffin S, Garcia-Cardenas V, Williams K, Benrimoj SI. Helping patients help themselves: a systematic review of self-management support strategies in primary health care practice. PLoS ONE 2019;14(8):e0220116. [doi: 10.1371/journal.pone.0220116] [Medline: 31369582]
12. Kappelin C, Carlsson AC, Wachtler C. Specific content for collaborative care: a systematic review of collaborative care interventions for patients with multimorbidity involving depression and/or anxiety in primary care. Fam Pract 2022 Jul 19;39(4):725-734. [doi: 10.1093/fampra/cmab079] [Medline: 34546354]
13. Menear M, Duhoux A, Roberge P, Fournier L. Primary care practice characteristics associated with the quality of care received by patients with depression and comorbid chronic conditions. Gen Hosp Psychiatry 2014;36(3):302-309. [doi: 10.1016/j.genhosppsych.2014.01.013] [Medline: 24629824]
14. Pung A, Fletcher SL, Gunn JM. Mobile app use by primary care patients to manage their depressive symptoms: qualitative study. J Med Internet Res 2018 Sep 27;20(9):e10035. [doi: 10.2196/10035] [Medline: 30262449]
15. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019 Jul 22;21(7):e12869. [doi: 10.2196/12869] [Medline: 31333198]
16. Wang K, Varma DS, Prosperi M. A systematic review of the effectiveness of mobile apps for monitoring and management of mental health symptoms or disorders. J Psychiatr Res 2018 Dec;107:73-78. [doi: 10.1016/j.jpsychires.2018.10.006] [Medline: 30347316]
17. McCue M, Blair C, Fehnert B, et al. Mobile app to enhance patient activation and patient-provider communication in major depressive disorder management: collaborative, randomized controlled pilot study. JMIR Form Res 2022 Oct 27;6(10):e34923. [doi: 10.2196/34923] [Medline: 36301599]
18. Rathbone AL, Prescott J. The use of mobile apps and SMS messaging as physical and mental health interventions: systematic review. J Med Internet Res 2017 Aug 24;19(8):e295. [doi: 10.2196/jmir.7740] [Medline: 28838887]
19. Milne-Ives M, Lam C, De Cock C, Van Velthoven MH, Meinert E. Mobile apps for health behavior change in physical activity, diet, drug and alcohol use, and mental health: systematic review. JMIR Mhealth Uhealth 2020 Mar 18;8(3):e17046. [doi: 10.2196/17046] [Medline: 32186518]
20. Carlo AD, Hosseini Ghomi R, Renn BN, Areán PA. By the numbers: ratings and utilization of behavioral health mobile applications. NPJ Digit Med 2019;2:54. [doi: 10.1038/s41746-019-0129-6] [Medline: 31304400]
21. Torous J, Andersson G, Bertagnoli A, et al. Towards a consensus around standards for smartphone apps and digital mental health. World Psychiatry 2019 Feb;18(1):97-98. [doi: 10.1002/wps.20592] [Medline: 30600619]
22. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
23. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatr Ann 2002 Sep;32(9):509-515. [doi: 10.3928/0048-5713-20020901-06]

XSL•FO
RenderX

24. Hibbard JH, Mahoney ER, Stockard J, Tusler M. Development and testing of a short form of the patient activation measure. Health Serv Res 2005 Dec;40(6 Pt 1):1918-1930. [doi: 10.1111/j.1475-6773.2005.00438.x] [Medline: 16336556]
25. Patient Activation Measure® (PAM®): learn more about the leading assessment of patient activation. Insignia Health. 2021. URL: https://www.insigniahealth.com/pam/ [accessed 2026-01-07]
26. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 Well-Being Index: a systematic review of the literature. Psychother Psychosom 2015;84(3):167-176. [doi: 10.1159/000376585] [Medline: 25831962]
27. Wellbeing measures in primary health care/the depcare project. : World Health Organization Regional Office for Europe; 1998 URL: https://iris.who.int/server/api/core/bitstreams/8af98105-30ec-4da4-9fe9-097f7459e6da/content [accessed 2026-01-07]
28. Sullivan MJ, Edgley K, Dehoux E. A survey of multiple sclerosis: I. Perceived cognitive problems and compensatory strategy use. Can J Rehabil 1990;4(2):99-105 [FREE Full text]
29. Lam RW, Lamy FX, Danchenko N, et al. Psychometric validation of the Perceived Deficits Questionnaire-Depression (PDQ-D) instrument in US and UK respondents with major depressive disorder. Neuropsychiatr Dis Treat 2018;14:2861-2877. [doi: 10.2147/NDT.S175188] [Medline: 30464471]
30. Klesges LM, Estabrooks PA, Dzewaltowski DA, Bull SS, Glasgow RE. Beginning with the application in mind: designing and planning health behavior change interventions to enhance dissemination. Ann Behav Med 2005 Apr;29 Suppl:66-75. [doi: 10.1207/s15324796abm2902s_10] [Medline: 15921491]
31. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
32. Gaglio B, Shoup JA, Glasgow RE. The RE-AIM framework: a systematic review of use over time. Am J Public Health 2013 Jun;103(6):e38-e46. [doi: 10.2105/AJPH.2013.301299] [Medline: 23597377]
33. Singer M, Baer H. Introducing Medical Anthropology: A Discipline in Action, 2nd edition: Lanham: AltaMira Press; 2012. URL: https://search.worldcat.org/title/introducing-medical-anthropology-a-discipline-in-action/oclc/800937988 [accessed 2026-01-07]
34. Patton MQ. Qualitative Evaluation and Research Methods, 2nd edition: Sage Publications; 1990. URL: https://openlibrary.org/books/OL2202510M/Qualitative_evaluation_and_research_methods [accessed 2026-01-07]
35. Corbin JM, Strauss AL. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory, 4th edition: SAGE Publications; 2014. URL: https://collegepublishing.sagepub.com/products/basics-of-qualitative-research-4-235578 [accessed 2026-01-07]
36. Serrano-Ripoll MJ, Zamanillo-Campos R, Fiol-DeRoque MA, Castro A, Ricci-Cabello I. Impact of smartphone app-based psychological interventions for reducing depressive symptoms in people with depression: systematic literature review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2022 Jan 27;10(1):e29621. [doi: 10.2196/29621] [Medline: 35084346]
37. Graham AK, Greene CJ, Kwasny MJ, et al. Coached mobile app platform for the treatment of depression and anxiety among primary care patients: a randomized clinical trial. JAMA Psychiatry 2020 Sep 1;77(9):906-914. [doi: 10.1001/jamapsychiatry.2020.1011] [Medline: 32432695]
38. Leong QY, Sridhar S, Blasiak A, et al. Characteristics of mobile health platforms for depression and anxiety: content analysis through a systematic review of the literature and systematic search of two app stores. J Med Internet Res 2022 Feb 4;24(2):e27388. [doi: 10.2196/27388] [Medline: 35119370]
39. McCloud T, Jones R, Lewis G, Bell V, Tsakanikos E. Effectiveness of a mobile app intervention for anxiety and depression symptoms in university students: randomized controlled trial. JMIR Mhealth Uhealth 2020 Jul 31;8(7):e15418. [doi: 10.2196/15418] [Medline: 32735221]
40. Deady M, Glozier N, Calvo R, et al. Preventing depression using a smartphone app: a randomized controlled trial. Psychol Med 2022 Feb;52(3):457-466. [doi: 10.1017/S0033291720002081] [Medline: 32624013]
41. Broglia E, Millings A, Barkham M. Counseling with guided use of a mobile well-being app for students experiencing anxiety or depression: clinical outcomes of a feasibility trial embedded in a student counseling service. JMIR Mhealth Uhealth 2019 Aug 15;7(8):e14318. [doi: 10.2196/14318] [Medline: 31418424]
42. Teepe GW, Da Fonseca A, Kleim B, et al. Just-in-time adaptive mechanisms of popular mobile apps for individuals with depression: systematic app search and literature review. J Med Internet Res 2021 Sep 28;23(9):e29412. [doi: 10.2196/29412] [Medline: 34309569]
43. Myers A, Chesebrough L, Hu R, Turchioe MR, Pathak J, Creber RM. Evaluating commercially available mobile apps for depression self-management. AMIA Annu Symp Proc 2020;2020:906-914. [Medline: 33936466]
44. Lu SC, Xu M, Wang M, et al. Effectiveness and minimum effective dose of app-based mobile health interventions for anxiety and depression symptom reduction: systematic review and meta-analysis. JMIR Ment Health 2022 Sep 7;9(9):e39454. [doi: 10.2196/39454] [Medline: 36069841]
45. Zhang R, Nicholas J, Knapp AA, et al. Clinically meaningful use of mental health apps and its effects on depression: mixed methods study. J Med Internet Res 2019 Dec 20;21(12):e15644. [doi: 10.2196/15644] [Medline: 31859682]

XSL•FO
RenderX

46. Bosworth HB, Voils CI, Potter GG, Steffens DC. The effects of antidepressant medication adherence as well as psychosocial and clinical factors on depression outcome among older adults. Int J Geriatr Psychiatry 2008 Feb;23(2):129-134. [doi: 10.1002/gps.1852] [Medline: 17563920]
47. Coote HMJ, MacLeod AK. A self-help, positive goal-focused intervention to increase well-being in people with depression. Clin Psychol Psychother 2012;19(4):305-315. [doi: 10.1002/cpp.1797] [Medline: 22610936]
48. Jacob J, Stankovic M, Spuerck I, Shokraneh F. Goal setting with young people for anxiety and depression: what works for whom in therapeutic relationships? A literature review and insight analysis. BMC Psychol 2022 Jul 13;10(1):171. [doi: 10.1186/s40359-022-00879-5] [Medline: 35831897]
49. Simon GE, Ralston JD, Savarino J, Pabiniak C, Wentzel C, Operskalski BH. Randomized trial of depression follow-up care by online messaging. J Gen Intern Med 2011 Jul;26(7):698-704. [doi: 10.1007/s11606-011-1679-8] [Medline: 21384219]
50. Solberg LI, Trangle MA, Wineman AP. Follow-up and follow-through of depressed patients in primary care: the critical missing components of quality care. J Am Board Fam Pract 2005;18(6):520-527. [doi: 10.3122/jabfm.18.6.520] [Medline: 16322414]
51. Huckvale K, Nicholas J, Torous J, Larsen ME. Smartphone apps for the treatment of mental health conditions: status and considerations. Curr Opin Psychol 2020 Dec;36(65-70):65-70. [doi: 10.1016/j.copsyc.2020.04.008] [Medline: 32553848]
52. Dinkel D, Harsh Caspari J, Fok L, et al. A qualitative exploration of the feasibility of incorporating depression apps into integrated primary care clinics. Transl Behav Med 2021 Sep 15;11(9):1708-1716. [doi: 10.1093/tbm/ibab075] [Medline: 34231855]
53. Cohen DJ, Keller SR, Hayes GR, Dorr DA, Ash JS, Sittig DF. Integrating patient-generated health data into clinical care settings or clinical decision-making: lessons learned from project HealthDesign. JMIR Hum Factors 2016 Oct 19;3(2):e26. [doi: 10.2196/humanfactors.5919] [Medline: 27760726]
54. Ye J. The impact of electronic health record-integrated patient-generated health data on clinician burnout. J Am Med Inform Assoc 2021 Apr 23;28(5):1051-1056. [doi: 10.1093/jamia/ocab017] [Medline: 33822095]
55. Gordon WJ, Landman A, Zhang H, Bates DW. Beyond validation: getting health apps into clinical practice. NPJ Digit Med 2020;3(14):14. [doi: 10.1038/s41746-019-0212-z] [Medline: 32047860]
56. O'Malley AS, Draper K, Gourevitch R, Cross DA, Scholle SH. Electronic health records and support for primary care teamwork. J Am Med Inform Assoc 2015 Mar;22(2):426-434. [doi: 10.1093/jamia/ocu029] [Medline: 25627278]
57. Sinsky CA, Sinsky TA, Rajcevich E. Putting pre-visit planning into practice. Fam Pract Manag 2015;22(6):34-38. [Medline: 26761083]
58. Moon K, Sobolev M, Kane JM. Digital and mobile health technology in collaborative behavioral health care: scoping review. JMIR Ment Health 2022 Feb 16;9(2):e30810. [doi: 10.2196/30810] [Medline: 35171105]
59. Moon KC, Sobolev M, Grella M, Alvarado G, Sapra M, Ball T. An mHealth platform for augmenting behavioral health in primary care: longitudinal feasibility study. JMIR Form Res 2022 Jul 1;6(7):e36021. [doi: 10.2196/36021] [Medline: 35776491]
60. Hoffman L, Benedetto E, Huang H, et al. Augmenting mental health in primary care: a 1-year study of deploying smartphone apps in a multi-site primary care/behavioral health integration program. Front Psychiatry 2019;10:94. [doi: 10.3389/fpsyt.2019.00094] [Medline: 30873053]
61. Carleton KE, Patel UB, Stein D, Mou D, Mallow A, Blackmore MA. Enhancing the scalability of the collaborative care model for depression using mobile technology. Transl Behav Med 2020 Aug 7;10(3):573-579. [doi: 10.1093/tbm/ibz146] [Medline: 32766866]

Abbreviations

AAFP NRN: American Academy of Family Physicians National Research Network
EHR: electronic health record
PHQ-9: 9-item Patient Health Questionnaire
RE-AIM: reach, effectiveness, adoption, implementation, and maintenance

Edited by A Stone; submitted 31.Jan.2025; peer-reviewed by AJ Saxon, HV Marwijk; accepted 15.Dec.2025; published 29.Jan.2026.

Please cite as:
Nederveld A, Robertson EA, Lanigan AM, Callen EF, Clay TL, Fehnert B, Chrones L, Martin ML, McCue M, Hester CM, Filippi MK Patient and Care Team Perspectives of Barriers to and Facilitators for the Implementation of a Digital Health Program for Depression in Primary Care: Qualitative Study
J Med Internet Res 2026;28:e72003
URL: https://www.jmir.org/2026/1/e72003
doi:10.2196/72003

XSL•FO
RenderX

© Andrea Nederveld, Elise A Robertson, Angela M Lanigan, Elisabeth F Callen, Tarin L Clay, Ben Fehnert, Lambros Chrones, Michael L Martin, Margaret McCue, Christina M Hester, Melissa K Filippi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Internet Health Care Service Use Behavioral Pattern Among Older Adults and the Role of the Technology Acceptance and Social Ecological Theory Model: Cross-Sectional Survey

Rui Li[1*], MPH; Xinyu Xu[2,3*], MPH; Qingsong Li[2,3*], MPH; Haobiao Liu[2,3], PhD; Ting Ting Zhou[1], MPH; Abebe Feyissa Amhare[2,3], PhD; Peiyu Liu[2,3], MPH; Jing Tang[2,3], MPH; Wei Wang[4], MPH; Fuju Zheng[4], PhD; Jing Han[3,5], PhD

[1]Shandong Center for Disease Control and Prevention, Ji'nan, Shandong, China

[2]Key Laboratory of Environment and Genes Related to Disease, School of Public Health, Health Science Center, Xi'an Jiaotong University, Xi'an, Shaanxi, China

[3]Department of Occupational and Environmental Health, School of Public Health, Health Science Center, Xi'an Jiaotong University, No. 76, Yanta West Road, Xi'an, Shaanxi, China

[4]Department of General Dentistry, Jinan Stomatological Hospital, Jinan, China

[5]Department of Health Science Center Global Health Institute, Health Science Center, Xi'an Jiaotong University, Xi'an, Shaanxi, China

[*]these authors contributed equally

Corresponding Author:
Jing Han, PhD
Department of Occupational and Environmental Health, School of Public Health, Health Science Center, Xi'an Jiaotong University, No. 76, Yanta West Road, Xi'an, Shaanxi, China

Abstract

Background: The rapid growth of internet health care (IH) offers older adults convenient medical services like remote consultations and health monitoring. However, its adoption among this group remains low, highlighting a significant digital divide. Understanding the behavioral patterns and determinants of IH use in the older population is crucial for optimizing digital health design and improving service accessibility.

Objective: This study aimed to analyze the multidimensional influencing factors of Chinese older adults' use of IH services based on the integrated framework of the technology acceptance model and social ecological model, and explore their behavioral patterns and key driving factors.

Methods: A cross-sectional study design was adopted to conduct a multistage stratified cluster random sampling survey in 3 cities in Shandong Province from May 2024 to July 2024, with a total of 1828 older adults aged 60 to 75 years included. The study uses latent category analysis to classify the use of IH service behaviors and employs multiple logistic regression, decision tree models, and structural equation modeling to analyze influencing factors and mediating pathways.

Results: Five distinct user groups were identified: nonusers (n=911), registration-dominant users (n=286), low-activity users (n=320), moderate comprehensive users (n=288), and full-service users (n=23). Multinomial logistic regression with nonusers as the reference group identified key determinants: individuals with below primary education had 96% lower odds of membership (odds ratios [OR] 0.039, 95% CI 0.012 - 0.084) compared to the reference group with junior college education or above in moderate comprehensive users, while male participants had higher odds of being full-service (OR 1.980, 95% CI 1.126 - 3.514) or moderate comprehensive (OR 1.310, 95% CI 1.012 - 1.705) users. Older age was consistently associated with lower adoption across all classes. Full-service users exhibited exceptionally high social support (OR 4.502, 95% CI 3.601 - 5.627), while moderate comprehensive users showed the highest technology acceptance (OR 2.803, 95% CI 2.355 - 3.342). The decision tree model (area under the curve of 0.94) found the optimal path: sufficient social support (≥2), good health status (>5), and high technical acceptance (≥30) yield the highest use probability (92%→96%). Mediation analysis indicated that social support influences usage willingness through both direct and indirect pathways. The direct effect was 0.712 (95% CI 0.552 - 0.972; P<.001). Among indirect pathways, technology availability and practicality accounted for the largest proportion of mediation (19.7%, 95% CI 16.8% - 22.6%), followed by technology acceptance (13.7%, 95% CI 11.1% - 16.3%) and social influence (8.9%, 95% CI 6.9% - 10.9%).

Conclusions: Optimizing age-friendly design, strengthening social support networks, and improving technological usability are keys to increasing the adoption of IH services among the older population. Future policies should develop targeted intervention strategies for different user groups to narrow the digital health divide.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e78037) doi:10.2196/78037

KEYWORDS

aged; internet medicine; mediation effect; social ecological model; technology acceptance model

Introduction

The rapid development of internet health care (IH) has revolutionized health care delivery, particularly for aging populations [1,2]. IH enhances access to teleconsultations for chronic disease management and enables real-time remote monitoring of vital signs through wearable devices [3-5]. These innovations streamline access to everyday health care tasks, such as teleappointments, online prescription refills, and contactless payment, while enhancing medication adherence through user-friendly digital reminders, reducing the logistical barriers to care for older adults with complex needs [6,7]. However, older adults' adoption of IH services remains suboptimal, with usage rates and digital literacy gaps perpetuating health inequities [8].

Due to the decline of the older adults' physiological function, the decline of cognitive flexibility, the lack of digital literacy, and the transformation of social roles, their internet use research has a unique feature of multidimensional interaction and intergenerational cultural conflict [9-12]. These characteristics present particular challenges during the transition to digitization. A study by Yang [11] revealed that perceived usefulness, perceived ease of use, and social influence positively influence older adults' behavioral intention toward health care services. This finding is consistent with recent extensions of the technology acceptance model (TAM) in digital health contexts, such as the work by Mouloudj et al [12], which validated the core TAM constructs (perceived usefulness and ease of use) in predicting intentions to use digital dental health services, while also highlighting the additional roles of trust and social influence. Meraya's [13] findings indicated that the primary barriers to telemedicine use were disinclination towards the technology and difficulties in scheduling appointments. Tan's [14] study underscores the significance of addressing older adults' subjective well-being and enhancing the accessibility of IH services. While existing studies have significantly advanced our understanding of individual determinants in IH adoption, the complex interplay of technological, psychological, and social factors underscores the need for integrated frameworks to fully capture older adults-specific digital disparities. This multidimensional complexity requires a theoretical framework that simultaneously addresses individual technology perception and macro background influences. The TAM operates on an individual level of adoption through dimensions such as perceived technology usefulness and ease of use, capturing cognitive assessments of technology utility [15].

However, the adoption of digital health technologies among older adults is a process shaped by a complex interplay of individual cognition and multilayered environmental influences. While the TAM effectively explains individual-level cognitive mechanisms through perceived usefulness and ease of use, its limited consideration of broader social contexts necessitates integration with the social ecological model (SEM). The SEM systematically categorizes environmental determinants across interpersonal, community, and policy levels, and this integration creates a conceptually synergistic framework wherein factors at these outer levels actively shape the core TAM constructs. This is particularly salient for older adults who often face unique systemic and interpersonal barriers to technology adoption. Specifically, at the interpersonal level, social support and family recommendations not only enhance perceived usefulness through endorsement but also directly help overcome practical usage barriers, thereby reinforcing perceived ease of use [16,17]. At the community level, resources such as age-friendly training programs and accessible digital infrastructure systematically reduce technical and accessibility barriers, transforming potential obstacles into facilitators of technology acceptance [18]. This combined TAM-SEM framework thus unifies macro-level environmental embeddedness with microlevel cognitive processes, providing a comprehensive theoretical lens to understand and optimize IH accessibility for the older population.

This study synthesizes the TAM and the SEM into a unified theoretical framework, establishing a multidimensional analytical structure to systematically investigate intelligent health technology adoption among older adults. The proposed model systematically integrates individual-level determinants such as electronic health literacy and self-efficacy, interpersonal dynamics exemplified by social support networks, and environmental dimensions encompassing perceived security risks and service accessibility. This tripartite integration enables a stratified exploration of adoption mechanisms, effectively delineating how cognitive-behavioral predispositions, relational influences, and systemic enablers collectively shape technology adoption trajectories within geriatric populations.

The primary aim of this study was to characterize patterns of IH adoption among older adults and to determine the extent to which these patterns can be explained by an integrated framework combining constructs from the SEM and the TAM. We hypothesized that older adults would exhibit distinct, empirically identifiable adoption trajectories, ranging from nonuse to comprehensive engagement. The secondary aims were to examine how a combined SEM–TAM framework predicts IH adoption behavior among older adults. We hypothesized that integrating SEM-based contextual determinants (such as social support and health status) with TAM-based cognitive determinants (such as technology acceptance and perceived usability) would provide a coherent basis for understanding and predicting IH adoption patterns within this population.

XSL•FO
RenderX

Methods

Study Population

This study employed a multistage stratified cluster random sampling design to investigate IH service use among older adults aged 60 to 75 in Shandong Province, China. To ensure regional representation, 3 prefecture-level cities representing medium gross domestic product rankings were selected. From each city, we randomly selected 2 counties (representing rural areas) and 2 districts (representing urban areas), achieving a 20% sampling rate at this stage. Within each selected county or district, 2 villages or communities were further chosen through simple random sampling.

Potential participants were identified through local household registry systems. The inclusion criteria required participants to be (1) aged 60 to 75 years, (2) Chinese nationals, and (3) capable of reading and writing. Exclusion criteria included (1) transient residents and (2) individuals with cognitive impairments. Specifically, we excluded those who reported a physician diagnosis of dementia. Additionally, to ensure data quality, we implemented an objective cognitive assessment protocol adapted from standard cognitive screening principles. Exclusion was applied when participants demonstrated (a) failure to comprehend the study purpose and informed consent after 3 structured explanations; (b) inability to answer basic orientation questions (current age, season, or city); or (c) evident disorientation to time or place as evaluated by trained interviewers [19,20].

Between May and July 2024, trained investigators from the Shandong Provincial Center for Disease Control and Prevention conducted face-to-face household interviews. The investigation team consisted of public health professionals with backgrounds in epidemiology, preventive medicine, and health statistics, all of whom had received standardized training on the survey protocol, interview techniques, and ethical considerations specific to this study. A validated questionnaire integrating the TAM and SEM constructs was used. The questionnaire assessed IH service use patterns, perceived ease of use, social support, and environmental barriers. From 1950 eligible individuals approached, 1828 completed valid questionnaires, yielding a response rate of 93.7% (n=1828). The final sample consisted of 48.6% (890/1828) female participants and 52.7% (964/1828) rural residents.

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki and was reviewed and approved by the Medical Ethics Review Committee of the Shandong Provincial Center for Disease Control and Prevention (approval SDJK (K) 2024-046-01). Written or verbal informed consent was obtained from all participants prior to their involvement in the study. As a token of appreciation for their time, participants were provided with a small financial compensation of 10 Chinese Yuan (US $ 1.4). To protect participants' privacy and ensure confidentiality, all collected data were anonymized and stored securely, with access restricted to the research team. The reporting of this cross-sectional research followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guideline [21] (Table S1 and Figure S1 in Multimedia Appendix 1).

Outcome

The primary outcome of IH services use was assessed through a multidimensional approach capturing different aspects of service adoption. Initial screening determined any prior IH use through a binary (yes or no) response to the question "have you used IH services before?" Further use was captured through 2 complementary measures: service breadth, represented by the count of different IH service types used from a comprehensive list of 10 specific services; and user typology, derived via latent class analysis of the specific service use patterns.

The secondary outcome, willingness to use IH in the future, was measured as an indicator of behavioral intention, an established precursor to technology adoption in acceptance theories. This construct was assessed using a 5-point Likert scale in response to the question, "Would you be willing to consider using IH services at your next medical visit?" This approach enables the examination of psychological precursors to adoption alongside actual use behavior within the observational study context.

Exposure Factors and Covariates

In this study, basic demographic information was collected from participants through a questionnaire. The participants' educational attainment was categorized as follows: below primary school, primary school, secondary school, high school or technical secondary school, junior college, or above. The marital status of the participants was also documented, with categories including married and unmarried. The participants' place of residence was categorized as either urban or rural. Key constructs, including usability, self-efficacy, eHealth literacy, perceived risks, social influence, and technology acceptance, were measured using validated scales adapted to the IH context. Usability was assessed using a 6-item adapted System Usability Scale, while self-efficacy was assessed using a 3-item adapted General Self-Efficacy Scale. The eHealth literacy construct was measured using the eHealth scale, and perceived risks were evaluated through a 4-item scale based on the Health Belief Model. Technology acceptance was captured using an 8-item scale derived from the TAM. Social influence was assessed as the degree to which people around me believe I should use IH services. It was measured using the Social Impact Scale [22]. All scales used 5-point Likert-type response formats. Complete measurement details, including all scale items, reliability metrics, and validity evidence, are provided in Table S2 in Multimedia Appendix 1.

Conceptual Framework

Our research used a conceptual framework that systematically integrated the TAM and the SEM, with variable selection directly informed by the theoretical constructs of both models. This integrated framework specified that technology adoption (the core outcome in TAM) was influenced not only by individual-level factors but also by multilevel social ecological determinants, thus providing the theoretical rationale for our variable selection. The chosen variables mapped directly onto specific theoretical constructs: usability and technology

XSL•FO
RenderX

acceptance represented core TAM dimensions capturing perceived ease of use and behavioral intention; self-efficacy and eHealth literacy constituted individual-level factors in the SEM that influenced technology adoption capabilities; perceived risks extended the TAM framework by incorporating threat appraisal mechanisms; social impact and social support operationalized the interpersonal level of the SEM, reflecting how social norms and relational resources affect adoption decisions; and health condition represented an individual-level factor in the SEM that shaped technology adoption capacity and motivation. Notably, "peer influence," a specific manifestation of social impact, theoretically operated at the interpersonal level of the SEM while simultaneously shaping the core TAM constructs of perceived usefulness and ease of use through social validation and practical assistance. This theoretical mapping explained why social factors demonstrated both direct effects on adoption and indirect effects through technology acceptance pathways in our analyses. The SEM's levels were intertwined and interactive, with interaction between the older adults and new medical technology occurring in societal (interpersonal) collaboration, all contextualized within health care system environments (Figure S2 in Multimedia Appendix 1) [15,23,24]. This theoretical grounding ensures our variable selection comprehensively captured the multidimensional nature of technology adoption across individual, interpersonal, and community levels.

Statistical Analysis

Latent class analysis (LCA) implemented in *R* classified participants into 5 IH service utilization profiles, optimized via Bayesian information criterion and Akaike information criterion [25]. The resulting classes were subsequently labeled according to their dominant use patterns for intuitive interpretation: class 1: "low-activity testers" (limited exploration of basic functions), class 2: "full-service users" (high probability of using all service types), class 3: "registration-dominant users" (highly focused on appointment services), class 4: "nonusers" (negligible use across all services), and class 5: "moderate comprehensive users" (balanced use of multiple core services). The 5-class solution was retained despite the small sample size (n=23) in class 2 ("full-service users") due to its strong theoretical relevance as a unique profile of early adopters, characterized by high education, income, and technological proficiency. This group is critical for understanding the full spectrum of digital health adoption. These descriptive labels were used in all subsequent tables and figures to replace numerical codes, enhancing the readability and direct interpretability of the findings. The characteristic use patterns of each class were visualized in the radar chart (Figure S3 in Multimedia Appendix 1). Multinomial logistic regression with elastic net regularization (λ=0.1, α=.5) identified predictors of class membership, contrasting nonusers (class 4) against other classes (classes 1 - 3, 5) while adjusting for age, health status, and behavioral covariates. A classification and regression tree algorithm with 10-fold cross-validation, trained on an 80:20 stratified split, predicted IH adoption using feature-selected predictors. Moderated mediation pathways were tested via structural equation modeling with robust maximum likelihood estimation, evaluating indirect effects and moderation by perceived risk. Analyses were supported by nonparametric bootstrap resampling (5000 iterations) for bias-corrected confidence intervals. Analyses used *R* packages tidy (LCA), caret classification and regression tree, and lavaan (structural equation modeling), with model diagnostics confirming adequacy (comparative fit index as 0.921, root mean square error of approximation as 0.043, and standardized root mean square residual as 0.049) [26] (Figure 1). All statistical analyses were conducted using R software (version 4.3.1).

Figure 1. Hypothesized pathways between high technical acceptance, mediating effects, and the health literacy group. (A) Conceptual diagram and (B) statistical diagram.

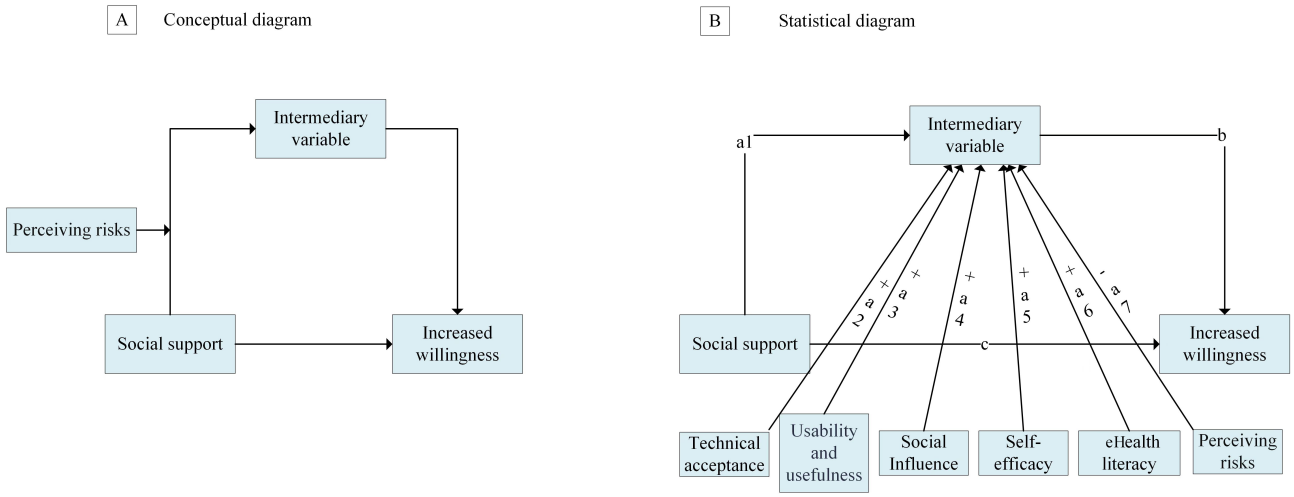

Results

Patterns and Characteristics of IH Service Adoption Among Older Adults

Among the 1828 respondents, 42.2% (n=772) reported using online appointment registration, 29.7% (n=543) used online payment, and 22.4% (n=410) checked test results online, whereas fewer engaged in online consultations (n=271, 14.8%), test scheduling (n=343, 18.8%), or advanced services such as psychological counseling (n=33, 1.8%), medication delivery (n=39, 2.1%), vaccination appointments (n=109, 6.0%), and in-home care (n=28, 1.5%).

XSL•FO
RenderX

Characteristics of Crowd Classification

We used LCA models, stratifying participants into 5 IH use classes (Tables S3 and S4 in Multimedia Appendix 1). Radar/Sankey plots visualized use patterns across clusters (Figures S3 and S4 in Multimedia Appendix 1). Low-activity users (class 1, n=320) demonstrated moderate education (n=161, 50.31%) and high technical acceptance (mean 33.72, SD 3.95), primarily engaging in basic functions like online payment. Compared to nonusers (class 4, n=911), they were younger (age: 65.9 y vs 69.0 y). Full-service users (class 2, n=23) featured high education (n=10, 43.48%), urban residence, male predominance (n=15, 65.22%), superior technical adaptability (mean 38.34, SD 4.77), service availability (mean 24.57, SD 3.25), and the highest median income. Registration-dominant users (class 3, n=286) focused on online appointments (>80%), showed moderate technical acceptance (mean 31.22, SD 3.64), lowest perceived risk (mean 10.98, SD 3.32), and male predominance (n=172, 60.14%). Moderate comprehensive users (class 5, N=288) displayed balanced multiservice engagement, with peak technology acceptance (mean 36.17, SD 5.89) and health literacy (mean 11.83, SD 1.67). Key intergroup distinctions involved technology acceptance, eHealth literacy, and urban-rural distribution. Nonusers (class 4, n=911) were predominantly rural (n=566, 62.13%), older (mean 69.0 y), and less educated (n=230, 25.25% ≤primary), and had elevated risk behaviors (n=107, 11.75% smoking every day), indicating a need for targeted interventions (Table 1).

XSL•FO
RenderX

Table . Baseline characteristics of the study population.

	Total (N=1828)	Class 1 (n=320)[a]	Class 2 (n=23)[a]	Class 3 (n=286)[a]	Class 4 (n=911)[a]	Class 5 (n=288)[a]	*P* value[b]	Relative rank across classes[c]
Sex, n (%)							<.001	
Male	938 (51.31)	158 (49.38)	15 (65.22)	172 (60.14)	426 (46.76)	167 (57.99)		4<1<5<3<2
Female	890 (48.69)	162 (50.63)	8 (34.78)	114 (39.86)	485 (53.24)	121 (42.01)		2<3<5<1<4
Age (years), average (range)	67.9 (60.0 - 75.0)	65.9 (60.0 - 75.0)	67.2 (60.0 - 73.0)	68.2 (60.0 - 75.0)	69.0 (60.0 - 65.0)	66.6 (60.0 - 74.0)	<.001	1<5<2<3<4
Education, n (%)							<.001	
Below primary school	303 (16.58)	22 (6.88)	2 (8.70)	38 (13.29)	230 (25.25)	11 (3.82)		5<1<2<3<4
Primary school	643 (35.18)	63 (19.69)	1 (4.35)	107 (37.41)	355 (38.97)	108 (37.5)		2<1<3<5<4
Secondary school	565 (30.91)	161 (50.31)	5 (21.74)	89 (31.12)	225 (24.7)	85 (29.51)		2<4<5<3<1
High school or technical secondary school	250 (13.68)	57 (17.81)	5 (21.74)	44 (15.38)	85 (9.33)	59 (20.49)		4<3<1<5<2
Junior college or above	67 (3.67)	17 (5.31)	10 (43.48)	8 (2.8)	16 (1.76)	25 (8.68)		4<3<1<5<2
Residency, n (%)							<.001	
Urban	864 (47.26)	165 (51.56)	23 (100)	120 (41.96)	345 (37.87)	211 (73.26)		4<3<1<5<2
Rural	964 (52.74)	155 (48.44)	0 (0)	166 (59.04)	566 (62.13)	77 (26.74)		2<5<1<3<4
Marriage, n (%)							<.001	
No	212 (11.60)	25 (7.81)	2 (8.70)	42 (14.69)	117 (12.84)	26 (9.03)		1<2<5<4<3
Yes	1616 (88.40)	295 (92.19)	21 (91.30)	244 (85.31)	794 (87.16)	262 (90.97)		3<4<5<2<1
Incomes[d]	4.34 (0.67 - 26.82)	5.89 (1.42 - 22.15)	8.73 (4.24 - 24.36)	4.96 (2.17 - 15.83)	2.87 (0.67 - 9.29)	6.92 (2.94 - 18.65)	.16	__[e]
Smoke, n (%)							<.001	
Everyday	196 (10.72)	48 (15.00)	2 (8.70)	18 (6.30)	107 (11.75)	21 (7.29)		3<5<2<4<1
Not everyday	58(3.17)	18 (5.63)	1 (4.35)	22 (7.69)	13 (1.42)	4 (1.39)		5<4<2<1<3
Never	1574 (86.11)	254 (79.37)	20 (86.95)	246 (86.01)	791 (86.83)	263 (91.32)		1<3<4<2<5
Drink, n (%)							<.001	
No	1333 (72.92)	222 (69.38)	20 (86.95)	207 (72.38)	662 (72.67)	222 (77.08)		1<3<4<5<2
Quit drinking	218 (11.93)	38 (11.86)	1 (4.35)	51 (17.82)	108 (11.86)	20 (6.95)		2<5<1<4<3
Yes	277 (15.15)	60 (18.76)	2 (8.70)	28 (9.80)	141 (15.47)	46 (15.97)		2<3<4<5<1
Usability and usefulness, mean (SD)	20.30 (2.47)	22.58 (1.83)	24.57 (3.25)	21.55 (1.98)	17.25 (2.61)	25.83 (2.34)	<.001	4<3<1<2<5
Self-efficacy, mean (SD)	9.53 (1.72)	11.14 (1.56)	11.61 (3.14)	9.67 (1.37)	7.92 (1.89)	12.51 (1.78)	<.001	4<3<1<2<5
Perceiving risks, mean (SD)	12.55 (3.01)	12.54 (2.98)	13.09 (4.56)	10.98 (3.32)	13.41 (3.17)	13.14 (2.89)	<.001	3<2<5<1<4
Electronic health literacy, mean (SD)	9.71 (1.65)	11.20 (1.79)	10.78 (3.82)	10.53 (1.46)	8.23 (2.15)	11.83 (1.67)	<.001	4<3<2<1<5
Social impact, mean (SD)	9.71 (1.88)	11.13 (1.92)	11.48 (3.63)	10.16 (1.54)	8.19 (2.27)	12.34 (1.81)	<.001	4<3<1<2<5

XSL•FO
RenderX

	Total (N=1828)	Class 1 (n=320)[a]	Class 2 (n=23)[a]	Class 3 (n=286)[a]	Class 4 (n=911)[a]	Class 5 (n=288)[a]	*P* value[b]	Relative rank across classes[c]
Social support, mean (SD)	2.94 (0.87)	2.35 (0.65)	6.07 (2.43)	2.91 (0.92)	2.00 (0.58)	2.03 (0.69)	<.001	4<5<1<3<2
Health status, mean (SD)	8.70 (1.34)	8.88 (1.17)	9.91 (2.75)	8.65 (1.28)	8.41 (1.43)	9.39 (1.56)	<.001	4<3<1<5<2
eHealth literacy, mean (SD)	9.92 (3.51)	10.78 (2.39)	11.83 (2.28)	10.53 (4.30)	8.23 (3.11)	11.20 (2.86)	<.001	4<3<1<2<5
Technology acceptance, mean (SD)	29.83 (4.12)	33.72 (3.95)	36.17 (5.89)	31.22 (3.64)	25.17 (3.38)	38.34 (4.77)	<.001	4<3<1<2<5

[a]A total of 5 user classes were identified: class 1 (low-activity tasters, n=320; minimal engagement with select internet health care services), class 2 (full-service users, n=23; high-frequency multifunctional platform use), class 3 (registration-dominant users, n=286; predominant appointment bookings), class 4 (nonusers, n=911; no internet health care adoption), and class 5 (moderate comprehensive users, n=288; intermediate engagement with integrated services, at lower intensity than class 2).

[b]*P* values were derived in accordance with established methodology, encompassing the chi-square test for categorical variables (eg, sex, education) and the Kruskal-Wallis test for continuous variables (eg, age, income). The significance level was set at .05.

[c]Classes ranked from lowest to highest value based on descriptive statistics; only shown for variables with *P*<.05.

[d]The data regarding income are expressed in RMB and the unit is 10,000 yuan (US $1428.6), the value is the median (range), rounded to 2 decimal places. Data processed by Winsorize (truncated by 1% extreme values at the beginning and end).

[e]Not available.

As shown in Table 2, the factors associated with membership in the distinct IH user typologies were derived from a multinomial logistic regression analysis with class 4 (nonusers) as the reference group. The analysis revealed a clear social gradient in IH adoption. This was most evident in class 5 (moderate comprehensive users), where individuals with below primary education had 96% lower odds of membership (OR 0.039, 95% CI 0.012 - 0.084) compared to the reference group with junior college education or above. Male gender was a significant predictor of membership in the more engaged classes, specifically class 2 (full-service users; OR 1.980, 95% CI 1.126 - 3.514) and class 5 (OR 1.310, 95% CI 1.012 - 1.705). Conversely, older age was consistently associated with significantly diminished odds of being in any user class relative to nonusers.

The psychographic and social determinants further delineated the user classes. Class 1 (low-activity tasters) was reported as having higher social impact (OR 1.221, 95% CI 1.125 - 1.325) and marginally positive technology acceptance (OR 1.951, 95% CI 1.630 - 2.332); their risk perception was not significantly different from nonusers, though they demonstrated moderate eHealth literacy (OR 1.905, 95% CI 1.597 - 2.279). Class 2 (full-service users) was distinguished by higher levels of social support (OR 4.502, 95% CI 3.601 - 5.627) and technology acceptance (OR 2.302, 95% CI 1.937 - 2.759), alongside elevated eHealth literacy (OR 1.729, 95% CI 1.443 - 2.060). Class 3 (registration-dominant users) also exhibited significant advantages in social support (OR 1.658, 95% CI 1.381 - 1.979) and technology acceptance (OR 1.652, 95% CI 1.314 - 1.908) compared to nonusers, complemented by higher eHealth literacy (OR 1.651, 95% CI 1.383 - 1.974). Class 5 (moderate comprehensive users) demonstrated the highest odds ratios (ORs) in eHealth literacy (OR 2.202, 95% CI 1.849 - 2.633), social impact (OR 1.320, 95% CI 1.194 - 1.459), and technology acceptance (OR 2.803, 95% CI 2.355 - 3.342).

XSL•FO
RenderX

Table . Correlates of user typology membership from multivariable logistic regression.[a]

Covariates	Class 1: low-activity tasters, OR[b] (95% CI)	Class 2: full-service users, OR (95% CI)	Class 3: registration-dominant users, OR (95% CI)	Class 5: moderate comprehensive users, OR (95% CI)
Sex				
Female (reference)	1.0	1.0	1.0	1.0
Male	0.696 (0.474 - 1.023)	1.980[c] (1.126 - 3.514)	1.463[c] (1.110 - 1.932)	1.310[c] (1.012 - 1.705)
Age	0.879[c] (0.846 - 0.919)	0.892[c] (0.790 - 0.982)	0.937[c] (0.903 - 0.973)	0.888[c] (0.846 - 0.931)
Residency				
Urban	1.404 (0.963 - 2.045)	__[d]	1.182 (0.897 - 1.574)	2.663[c] (1.724 - 4.113)
Rural (reference)	1.0	1.0	1.0	1.0
Education				
Below the primary school	0.111[c] (0.047 - 0.191)	0.073[c] (0.022 - 0.142)	0.282[c] (0.121 - 0.667)	0.039[c] (0.012 - 0.084)
Primary school	0.271[c] (0.151 - 0.497)	0.018[c] (0.002 - 0.059)	0.804 (0.367 - 1.778)	0.331[c] (0.202 - 0.552)
Secondary school	1.109 (0.611 - 3.728)	0.152[c] (0.095 - 0.231)	0.668 (0.297 - 1.499)	0.255[c] (0.157 - 0.424)
High school or technical secondary school	0.662 (0.364 - 1.211)	0.152[c] (0.095 - 0.231)	0.348[c] (0.159 - 0.786)	0.180[c] (0.108 - 0.329)
Junior college or above (reference)	1.0	1.0	1.0	1.0
Marriage				
No	0.584[c] (0.372 - 0.899)	0.453[c] (0.211 - 0.959)	1.508[c] (1.108 - 2.050)	0.651[c] (0.430 - 0.997)
Yes (reference)	1.0	1.0	1.0	1.0
Smoke				
Everyday	1.402 (0.973 - 2.020)	0.745 (0.167 - 3.416)	0.457[c] (0.242 - 0.861)	0.642 (0.393 - 1.057)
Not everyday	3.803 (2.051 - 7.061)	0.393 (0.055 - 3.113)	1.866[c] (1.169 - 8.042)	0.120[c] (0.004 - 0.346)
Never (reference)	1.0	1.0	1.0	1.0
Drink				
No	0.529[c] (0.315 - 0.889)	2.621 (0.566 - 12.300)	1.338 (0.772 - 2.320)	1.275 (0.853 - 1.894)
Quit drinking	0.687 (0.351 - 1.354)	0.314 (0.114 - 7.430)	1.809 (0.934 - 3.503)	0.654 (0.293 - 1.459)
Yes (reference)	1.0	1.0	1.0	1.0
Usability and usefulness	1.041 (0.994 - 1.090)	1.553[c] (1.107 - 2.184)	1.084[c] (1.038 - 1.132)	1.136[c] (1.077 - 1.199)
Self-efficacy	1.103 (0.853 - 1.388)	1.811[c] (1.253 - 2.601)	1.012 (0.934 - 1.097)	1.971[c] (1.372 - 2.879)
Perceiving risks	0.975 (0.931 - 1.022)	0.827[c] (0.619 - 0.981)	0.806[c] (0.730 - 0.904)	0.955[c] (0.913 - 0.999)
Electronic health literacy	1.905[c] (1.597 - 2.279)	1.729[c] (1.443 - 2.060)	1.651[c] (1.383 - 1.974)	2.202[c] (1.849 - 2.633)
Social impact	1.221[c] (1.125 - 1.325)	1.248[c] (1.071 - 1.603)	1.164[c] (1.078 - 1.258)	1.320[c] (1.194 - 1.459)
Social support	1.257[c] (1.055 - 1.496)	4.502[c] (3.601 - 5.627)	1.658[c] (1.381 - 1.979)	1.027 (0.857 - 1.229)
Health status	1.201[c] (1.014 - 1.430)	1.853[c] (1.555 - 2.214)	1.108 (0.927 - 1.319)	1.451[c] (1.224 - 1.727)
Technology acceptance	1.951[c] (1.630 - 2.332)	2.302[c] (1.937 - 2.759)	1.652[c] (1.314 - 1.908)	2.803[c] (2.355 - 3.342)

[a]With class 4 (nonusers) as a reference OR value (ratio), OR represents the probability ratio of a certain type of user possessing a certain feature relative to nonusers. OR>1: this type of user is more likely to possess the feature, OR<1: this type of user is less likely to possess the feature. The analysis used multinomial logistic regression with Bonferroni correction for multiple comparisons. Adjusted covariates included sex, age, education level, residency, marital status, income, smoking status, and drinking status. Five user classes were identified: class 1 (low-activity tasters, n=320; minimal engagement with select internet health care services), class 2 (full-service users, n=23; high-frequency multifunctional platform use), class 3 (registration-dominant users, n=286; predominant appointment bookings), class 4 (nonusers, n=911; no internet health care adoption), and class 5 (moderate comprehensive

XSL•FO
RenderX

users, n=288; intermediate engagement with integrated services, at lower intensity than class 2).

[b]OR: odds ratio.

[c]Statistically significant at the .05 level.

[d]OR and CI could not be calculated because the model did not converge, likely due to a lack of variation in residency (urban/rural) within class 2.

Predictive Factors

The model showed high accuracy in classifying future internet use intention, with area under the curve values of 0.94 (training set) and 0.93 (test set) (Figure 2B). Figure 2C and 2D showed the decision tree model's confusion matrices for the validation and training sets, respectively. Table S5 in Multimedia Appendix 1 shows high sensitivity in both test and training sets, with higher specificity, accuracy, precision, F1 score, lower false positive rate, and good overall test-set performance. The decision tree model revealed the recursive dichotomy of the multivariate feature space affecting IH service use probability. The variable importance order is: social support, health status, usability and usefulness, technology acceptance, perceiving risks, and eHealth literacy. In the left branch, future use probability dropped sharply to 23%. Further division showed that a technology usability and practicality score ≥21 (21% subsample) could partially offset insufficient social support, increasing probability to 55%; but with technology acceptance ≤29 (13% subsample), probability drops to 41%. When perceived health risk >13 (40% subsample), usage probability dropped to 7%, forming the lowest usage group. In the right branch, usage probability jumped to 90%, with health status >5 (38% subsample), further pushing it to 92%, forming a core high-usage group. For those with technology acceptance ≥30 and eHealth literacy ≥12 (34% subsample), use probability reached 96%, reflecting technology and literacy's multiplier effect. The optimal path: sufficient social support (≥2), good health status (>5), and high technical acceptance (≥30) yielded the highest use probability (92% →96%). Insufficient social support (<2) with high-risk perception (>13) led to extremely low use probability (7%), an 89% point difference from the optimal path (Figure 2).

The direct effect showed that for every 1 unit increase in social support, the willingness to use increased by 0.712 (95% CI 0.552 - 0.872; P<.001). Among the indirect effects, the mediating contribution of technology usability and practicality is the largest, accounting for 19.7% (95% CI 16.8% - 22.6%) of the total effect, followed by technology acceptance (13.7%, 95% CI 11.1% - 16.3%) and social influence (8.9%, 95% CI: 6.9% - 10.9%). Although self-efficacy (3.2%) and eHealth literacy (2.7%, 95% CI 0.9% - 4.5%) were significant, their contributions are relatively small, while perceived risk was weakly negatively mediated (−0.6%, β=−.007) (Table 3).

XSL•FO
RenderX

Figure 2. A decision tree model for predicting internet health care service use among older adults. AUC: area under the curve; ROC: receiver operating characteristic.

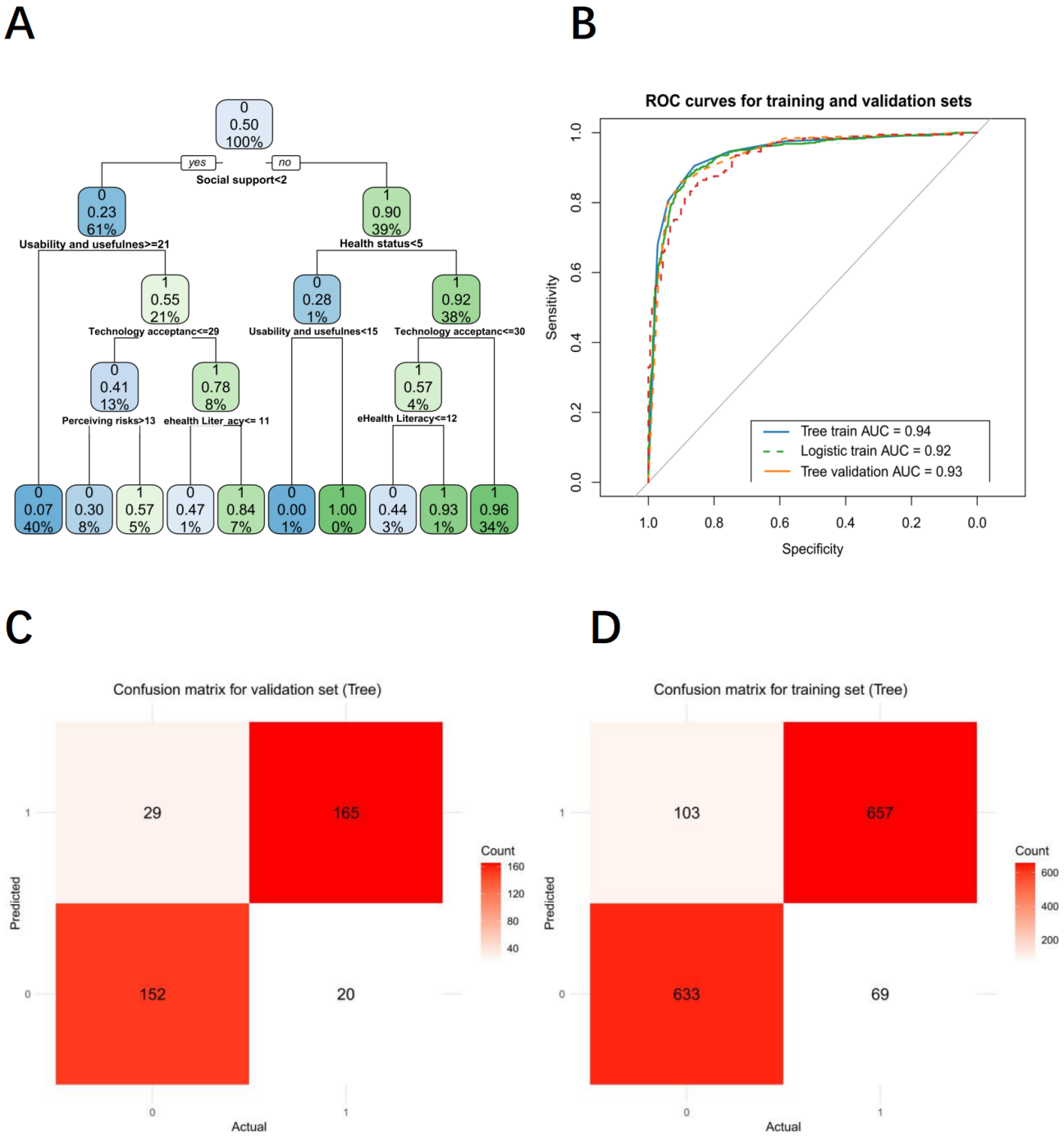

XSL•FO
RenderX

Table . Mediation analysis of the effect of social support on willingness for future internet health care use[a].

Pathway	β (95% CI)	*P* value	Mediation proportion[b], % (95% CI)
Total effect (social support → willingness to use)	1.198 (1.036 to 1.360)	<.001	100
Direct effect (social support → willingness to use)	0.712[c] (0.552 to 0.872)	<.001	59.4 (54.1 to 64.7)
Total indirect effect	0.486[c] (0.420 to 0.552)	<.001	40.6 (35.3 to 45.9)
Mediated pathways (social support → mediator → willingness)			
Usability and usefulness	0.236[c] (0.205 to 0.267)	<.001	19.7 (16.8 to 22.6)
Technology acceptance	0.164[c] (0.134 to 0.194)	<.001	13.7 (11.1 to 16.3)
Social influence	0.107[c] (0.085 to 0.129)	<.001	8.9 (6.9 to 10.9)
Self-efficacy	0.038[c] (0.021 to 0.055)	<.001	3.2 (1.7 to 4.7)
eHealth literacy	0.032[c] (0.011 to 0.053)	.004	2.7 (0.9 to 4.5)
Perceived risk	−0.007[c] (−0.010 to −0.004)	<.001	−0.6 (−0.9 to - 0.3)
Health status	0.009 (−0.002 to 0.020)	.12	__[d]

[a]Model adjusted variables: all analyses adjusted for sex, age, education level, place of residence (urban-rural), marital status, income level, and chronic disease status. Mediation effect calculation: the bias-corrected bootstrap method (5000 repeated samples) is used to estimate the indirect effect. Path explanation: Direct effect: independent impact of social support on usage intention (without mediating variables). Indirect effect: The chain effect of social support on usage intention through mediating variables.

[b]The formula for calculating the mediation ratio is: indirect effect/(direct effect + indirect effect) × 100%.

[c]*P*<.001; a 95% CI that is non-zero is considered significant.

[d]Not applicable.

Discussion

Principal Findings

This study divides the use of IH services by older adults into 5 groups through cluster analysis and constructs an integrated analysis framework based on ecological and technological models, providing a multidimensional perspective for understanding the complex mechanisms of older adults' use of IH services. Research has found that the overall adoption rate of IH services among the elderly population is low, with nearly half of the older adults never using related services and only a very small number becoming regular users. Multivariate analysis shows that advanced age, psychosocial factors, and ease of use of technology are key predictive factors affecting IH participation. Age growth and daily material use are negatively correlated with IH use, while technology usability awareness and electronic health literacy significantly promote usage behavior. Social support drives usage behavior through both direct effects (increasing willingness to use) and indirect effects (mediating effects of technology availability, technology acceptance, and social impact), with the mediating contribution of technology availability being the most prominent. The study emphasizes the need to restructure the design and policy framework of IH services to address the social and technological needs of the older population, in order to bridge the intergenerational gap in the use of health technology.

Theoretical Interpretation of User Patterns Through TAM-SEM Integration

The 5 distinct user patterns emerging from our analysis can be effectively interpreted through our integrated theoretical framework, with the SEM elucidating how macro-level social structures regulate microlevel behavioral patterns. Consistent with prior research on technology adoption barriers [11,27], the low-activity testers (class 1), despite demonstrating moderate technical acceptance, exhibited limited engagement, reflecting how high perceived risk (a TAM extension) and insufficient reinforcing factors at the interpersonal level of the SEM can inhibit adoption even when basic ease of use is acknowledged. This pattern aligns with studies highlighting the critical role of privacy concerns in limiting technology engagement among vulnerable populations [28]. Conversely, the full-service users (class 2), though small in number, exemplify the synergistic effect postulated by the SEM. However, the exceptionally low proportion of comprehensive users in our Chinese sample contrasts sharply with reports from high-income countries [29], suggesting contextual moderators, including structural barriers and cultural preferences for traditional care.

The registration-dominant users (class 3) present a compelling case where a specific, high-use function drives adoption despite moderate overall technical acceptance, echoing previous observations of "feature-specific adoption" in health technology literature [30]. This pattern provides empirical support for the "minimum viable product" approach advocated in implementation science, while simultaneously revealing its

XSL•FO
RenderX

limitations in fostering broader platform engagement. Meanwhile, the moderate comprehensive users (class 5) demonstrate how optimized perceptions across multiple TAM dimensions facilitate broader adoption, reinforcing established technology acceptance theories while extending them to the geriatric context [31]. The nonusers (class 4) represent a population where deficiencies across multiple SEM levels create compounded barriers, a finding that resonates with digital divide research across socioeconomic strata. This multilevel barrier pattern aligns with Bronfenbrenner's ecological systems theory while highlighting its specific manifestations in digital health contexts [32].

TAM-SEM Predictors

The introduction of technical models further reveals the data-driven prediction mechanism. The decision tree model identifies 6 core predictors of older adults' future willingness to use IH services. While perceived risk demonstrates a significant moderating effect that weakens the impact of technology usability on use frequency, usability and usefulness emerge as equally crucial factors that directly facilitate adoption through enhancing perceived ease of use. This finding extends beyond the risk-focused narrative prevalent in digital health literature, revealing a more balanced interplay between barriers and facilitators [33]. Cross-cultural comparisons further illuminate these relationships. Our findings regarding the strong predictive power of usability align with Möller et al's [34] European-Japanese study documenting how interface design fundamentally shapes older adults' digital engagement patterns. However, whereas their research emphasized technological coaching systems, our model reveals how usability interacts with locally specific factors like family support networks in the Chinese context. Similarly, Morishita-Suzuki et al [35] identified depression and leisure activities as crucial determinants of eHealth literacy across European Union and Japanese older adults, while our study extends this understanding by quantifying how eHealth literacy functions within a broader predictive framework alongside usability and social support. The identified predictor hierarchy—social support, health status, usability and usefulness, technology acceptance, eHealth literacy, and perceived risks—represents a significant departure from the conventional TAM that typically prioritizes cognitive perceptions over contextual factors. This corresponds to the "perceived usefulness → usage behavior" pathway in the TAM framework, where perceived usability directly affects technology adoption, while social support indirectly enhances usage intention through a mediating pathway that alleviates anxiety. The prominent role of usability in our model underscores its fundamental importance in gerontechnology design, particularly for populations with limited digital experience. This dual-framework approach reveals that social support not only directly encourages adoption but also indirectly facilitates use through enhancing technology acceptance and reducing anxiety, while usability serves as the foundational element that makes initial engagement possible for older users. This suggests that technology design needs to prioritize addressing pain points such as privacy protection and operational transparency to reduce decision-making barriers for older users [36-38]. This dual-framework approach reveals that social support not only directly encourages adoption but also indirectly facilitates usage through enhancing technology acceptance and reducing anxiety.

Digital Inequality and Policy Implications

This study also underscores the role of digital inequality and broader social determinants. Limited access to smartphones, poor connectivity in rural communities, and lower education levels all constrained adoption, reflecting entrenched disparities in digital health equity. Such inequalities suggest that IH adoption cannot be understood solely as an individual decision but rather as an outcome shaped by structural opportunities and constraints. Addressing these challenges requires coordinated policies that integrate infrastructure investment, targeted subsidies, and the inclusion of IH capacity indicators in routine health assessments [37,38].

The cross-analysis of the 2 models shows that moderate comprehensive service users (class 5) are more sensitive to the usability of the technical interface, while registered dominant users (class 3) are limited to the singular use of service functions, reflecting the gap between the functional integration of the IH platform and the matching needs of user segmentation. From the perspective of policy practice, the inspiration of the ecological model lies in constructing a 3-dimensional intervention system of "individual capability enhancement, social support strengthening, technological environment adaptation" [39,40]: for nonservice users (class 4), it is necessary to reduce perceived risks through community digital literacy training; for registered dominant users, the flow design of appointment services and other functions should be optimized to promote the transformation from a single service to comprehensive use. The successful cases of full-service users provide empirical evidence for the "technology feedback" family model, which amplifies the positive effects of social support through intergenerational collaboration mechanisms. The application value of technical models lies in accurately identifying high-risk groups, such as daily smokers and the older population, who account for a higher proportion of non–service users. Hierarchical intervention can be carried out through the risk scores output by decision tree models, combined with simulation training and privacy protection technology, gradually building an inclusive digital health ecosystem [41,42].

To implement these insights, we propose a further intervention strategy. At the microlevel, develop TAM-driven "cognitive scaffolding" interfaces, such as AI voice assistants with gerontologically optimized dialogue flow, to reduce navigation complexity. Enforce the IH platform to implement general data protection regulation privacy dashboards using intuitive icons, directly addressing the perceived risks we identified in risk literacy balance. At the macro level: standardize algorithm governance through the "elderly friction" policy—forcing IH platforms to disable autoplay functionality and insert reflective prompts (such as "You browsed for 10 min—take a break!") to prevent forced use [43]. Incorporate IH capability indicators into the national elderly health assessment, prioritizing infrastructure upgrades in areas with a lower prevalence of level 2 users [44]. Establish "Digital Health Hub" Peer-Learning Models: community centers (eg, libraries and senior clubs) should host regular peer-led training sessions, facilitated by

XSL•FO
RenderX

trained "Digital Health Champions" from among tech-savvy seniors [45].

The combined application of ecological and technological models also highlights the originality of this study. While ecological models emphasize external environments and TAM focuses on perceptions of usability, their integration offers a multidimensional explanation of adoption. This dual perspective enhances the explanatory power of our findings and demonstrates that interventions must simultaneously target cognitive, relational, and systemic barriers to promote digital health use among older adults.

Limitations

There are several limitations that should be acknowledged in this study. First, a cross-sectional design limits causal reasoning. Although the relationship between the identified factors (such as social support and technology acceptance) and the use of IH services has been observed, a longitudinal study is needed to establish a time relationship and determine whether the improvement of social support and technology acceptance will lead to an increase in the IH adoption rate over time. Second, due to regional sampling, its generalizability is limited. This regional focus may limit the generalizability of research results to other socio-economic backgrounds or geographical regions. Third, the screening of participants for cognitive impairment was based on prior research and clinical judgment rather than a standardized instrument like the Mini-Mental State Examination. While this approach was pragmatic for our large-scale study, it may have resulted in some residual confounding. Fourth, our scale was adapted from a standardized instrument based on the actual conditions of older adults in China, which may affect its external generalizability. These limitations highlight important directions for future research, including longitudinal design, broader geographic sampling, inclusion of objective use data, and a more comprehensive assessment of background factors influencing the adoption of IH in the older population.

Conclusions

This study uses an integrated TAM and SEM framework to examine IH service use among older adults in China. The identification of 5 distinct user profiles demonstrates substantial diversity in adoption patterns, ranging from complete nonuse to comprehensive adoption. Findings reveal that technology adoption is shaped by a multifaceted interplay of cognitive perceptions, relational resources, and systemic enablers. Willingness to use serves as a crucial psychological precursor to actual adoption, influenced by both technological factors and social-environmental contexts. Usability and perceived risk represent significant technological determinants, while social support facilitates adoption through multiple pathways. The integrated TAM-SEM framework provides theoretical value by demonstrating how factors across ecological levels collectively shape adoption behaviors. For practical application, technology design should prioritize aging-friendly interfaces and privacy protections, while community programs should strengthen social support through tailored digital literacy initiatives. Policy efforts must address structural barriers in underserved areas. Future research should explore the longitudinal evolution of user profiles across diverse contexts. By developing targeted strategies for different user segments, we can foster an inclusive digital health ecosystem that benefits all older adults regardless of technological proficiency or social resources.

Acknowledgments

The authors extend their gratitude to all the participants in this study. The authors also thanked the staff from the Shandong CDC and the local community workers for their invaluable support in data collection. In addition, FZ (email: 2778201941@qq.com) and WW (email: 2562109293@qq.com) acted as common corresponding authors. The authors thank them for their expert guidance, which has helped improve the research design and analysis framework.

Funding

This work was funded by Shandong Province Medical and Health Science and Technology Development Plan Project (202312071011) and Shandong Province Humanities and Social Sciences Research Projects (2023-JKZX-11). The funders had no involvement in data collection and analysis or the preparation of this article. The analysis and interpretation of the evidence was done by the authors themselves, not by the funders.

Data Availability

The datasets obtained during the current study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: RL, XX, QL, JH
Formal analysis: RL, XX, QL
Investigation: HL, TTZ, AFA, PL, JT, WW, FZ
Visualization: RL, XX, QL, JH
Writing – original draft: RL, XX, QL
Writing – review & editing: HL, TTZ, AFA, PL, JT, WW, FZ
All authors read and approved the final version.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Study design, measures, and analytic visuals related to internet use among older adults.
[DOCX File, 958 KB - jmir_v28i1e78037_app1.docx]

Checklist 1
STROBE checklist.
[DOCX File, 31 KB - jmir_v28i1e78037_app2.docx]

References

1. Yang F, Shu H, Zhang X. Understanding “internet plus healthcare” in China: policy text analysis. J Med Internet Res 2021 Jul 26;23(7):e23779. [doi: 10.2196/23779] [Medline: 34309581]
2. Yu J, Meng S. Impacts of the internet on health inequality and healthcare access: a cross-country study. Front Public Health 2022;10:935608. [doi: 10.3389/fpubh.2022.935608]
3. Olmedo-Aguirre JO, Reyes-Campos J, Alor-Hernández G, Machorro-Cano I, Rodríguez-Mazahua L, Sánchez-Cervantes JL. Remote healthcare for elderly people using wearables: a review. Biosensors (Basel) 2022 Jan 27;12(2):73. [doi: 10.3390/bios12020073] [Medline: 35200334]
4. Gill SK, Barsky A, Guan X, et al. Consumer wearable devices for evaluation of heart rate control using digoxin versus beta-blockers: the RATE-AF randomized trial. Nat Med 2024 Jul;30(7):2030-2036. [doi: 10.1038/s41591-024-03094-4] [Medline: 39009776]
5. Farsi D. Social media and health care, part I: literature review of social media use by health care providers. J Med Internet Res 2021 Apr 5;23(4):e23205. [doi: 10.2196/23205] [Medline: 33664014]
6. Slabodkin G. Text reminders help seniors improve medication adherence. Fierce Healthcare. 2013 May 20. URL: https://www.fiercehealthcare.com/mobile/text-reminders-help-seniors-improve-medication-adherence [accessed 2025-04-21]
7. Mansour F. Here are 6 ways mobile payments simplify transactions for seniors. Yves Brooks. 2024 Jan. URL: https://yves-brooks.com/here-are-6-ways-mobile-payments-simplify-transactions-for-seniors/ [accessed 2025-04-21]
8. Ausserhofer D, Piccoliori G, Engl A, et al. Community-dwelling older adults' readiness for adopting digital health technologies: cross-sectional survey study. JMIR Form Res 2024 Apr 30;8:e54120. [doi: 10.2196/54120] [Medline: 38687989]
9. Estrela M, Semedo G, Roque F, Ferreira PL, Herdeiro MT. Sociodemographic determinants of digital health literacy: a systematic review and meta-analysis. Int J Med Inform 2023 Sep;177:105124. [doi: 10.1016/j.ijmedinf.2023.105124] [Medline: 37329766]
10. Lin YC, Chung CP, Lee PL, et al. The flexibility of physio-cognitive decline syndrome: a longitudinal cohort study. Front Public Health 2022;10:820383. [doi: 10.3389/fpubh.2022.820383] [Medline: 35734760]
11. Yang HJ, Lee JH, Lee W. Factors influencing health care technology acceptance in older adults based on the technology acceptance model and the unified theory of acceptance and use of technology: meta-analysis. J Med Internet Res 2025 Mar 28;27:e65269. [doi: 10.2196/65269] [Medline: 40153796]
12. Mouloudj K, Bouarar AC, Martínez Asanza DA, Dimitrova M. Understanding customers’ intentions to use digital dental health services: an expanded technology acceptance model. In: Martínez Asanza D, editor. Advances in Medical Technologies and Clinical Practice IGI Global: IGI Global Scientific Publishing; 2024:189-210. [doi: 10.4018/979-8-3693-7165-7.ch007]
13. Meraya AM, Khardali A, Ahmad S, et al. Telehealth perceptions and associated factors among older adults with chronic conditions in Saudi Arabia: a comparative study of users and non-users. Front Public Health 2025;13:1542974. [doi: 10.3389/fpubh.2025.1542974] [Medline: 40144973]
14. Tan SH, Yap YY, Tan SK, Wong CK. Determinants of telehealth adoption among older adults: cross-sectional survey study. JMIR Aging 2025 Mar 24;8:e60936. [doi: 10.2196/60936] [Medline: 40126531]
15. Nadal C, Sas C, Doherty G. Technology acceptance in mobile health: scoping review of definitions, models, and measurement. J Med Internet Res 2020 Jul 6;22(7):e17256. [doi: 10.2196/17256] [Medline: 32628122]
16. De la Peña-López IJ, Acosta-Gonzaga E. Adoption of technology in older adults in Mexico City: an approach from the technology acceptance model. Brain Sci 2025 Jun 12;15(6):632. [doi: 10.3390/brainsci15060632] [Medline: 40563802]
17. Freeman S, Marston HR, Olynick J, et al. Intergenerational effects on the impacts of technology use in later life: insights from an international, multi-site study. Int J Environ Res Public Health 2020 Aug 7;17(16):5711. [doi: 10.3390/ijerph17165711] [Medline: 32784651]
18. Arthanat S. Promoting information communication technology adoption and acceptance for aging-in-place: a randomized controlled trial. J Appl Gerontol 2021 May;40(5):471-480. [doi: 10.1177/0733464819891045] [Medline: 31782347]
19. First MB. Diagnostic and statistical manual of mental disorders, 5th edition, and clinical utility. J Nerv Ment Dis 2013 Sep;201(9):727-729. [doi: 10.1097/NMD.0b013e3182a2168a] [Medline: 23995026]

XSL•FO
RenderX

20. Stephan BCM, Minett T, Pagett E, Siervo M, Brayne C, McKeith IG. Diagnosing mild cognitive impairment (MCI) in clinical trials: a systematic review. BMJ Open 2013;3(2):e001909. [doi: 10.1136/bmjopen-2012-001909] [Medline: 23386579]
21. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
22. Sifat MS, Saperstein SL, Tasnim N, Green KM. Motivations toward using digital health and exploring the possibility of using digital health for mental health in Bangladesh university students: cross-sectional questionnaire study. JMIR Form Res 2022 Mar 4;6(3):e34901. [doi: 10.2196/34901] [Medline: 35254267]
23. Holden RJ, Karsh BT. The technology acceptance model: its past and its future in health care. J Biomed Inform 2010 Feb;43(1):159-172. [doi: 10.1016/j.jbi.2009.07.002] [Medline: 19615467]
24. Davidson P, Rushton CH, Kurtz M, et al. A social-ecological framework: a model for addressing ethical practice in nursing. J Clin Nurs 2018 Mar;27(5-6):e1233-e1241. [doi: 10.1111/jocn.14158] [Medline: 29119653]
25. Grant RW, McCloskey J, Hatfield M, et al. Use of latent class analysis and k-means clustering to identify complex patient profiles. JAMA Netw Open 2020 Dec 1;3(12):e2029068. [doi: 10.1001/jamanetworkopen.2020.29068] [Medline: 33306116]
26. Choi DH, Ro YS, Park JH, et al. Evaluation of socioeconomic position and survival after out-of-hospital cardiac arrest in Korea using structural equation modeling. JAMA Netw Open 2023 May 1;6(5):e2312722. [doi: 10.1001/jamanetworkopen.2023.12722] [Medline: 37163262]
27. An J, Zhu X, Wan K, et al. Older adults' self-perception, technology anxiety, and intention to use digital public services. BMC Public Health 2024 Dec 19;24(1):3533. [doi: 10.1186/s12889-024-21088-2] [Medline: 39702092]
28. Liu H, Zhou Q, Liang S. Digital inclusion in public services for vulnerable groups: a systematic review for research themes and goal-action framework from the lens of public service ecosystem theory. Gov Inf Q 2025 Jun;42(2):102019. [doi: 10.1016/j.giq.2025.102019]
29. Tavares AI. Self-assessed health among older people in Europe and internet use. Int J Med Inform 2020 Sep;141:104240. [doi: 10.1016/j.ijmedinf.2020.104240] [Medline: 32739610]
30. Zhang Y, Wu P. Continuous adoption of online healthcare platforms: an extension to the expectation confirmation model and network externalities. BMC Public Health 2024 Sep 27;24(1):2630. [doi: 10.1186/s12889-024-20072-0]
31. Karahanna E, Straub DW, Chervany NL. Information technology adoption across time: a cross-sectional comparison of pre-adoption and post-adoption beliefs. MIS Q 1999 Jun;23(2):183-213. [doi: 10.2307/249751]
32. Ucar I, Gramaglia M, Fiore M, Smoreda Z, Moro E. News or social media? Socio-economic divide of mobile service consumption. J R Soc Interface 2021 Dec;18(185):20210350. [doi: 10.1098/rsif.2021.0350] [Medline: 34847793]
33. Tustin JL, Crowcroft NS, Gesink D, Johnson I, Keelan J. Internet exposure associated with Canadian parents' perception of risk on childhood immunization: cross-sectional study. JMIR Public Health Surveill 2018 Jan 19;4(1):e7. [doi: 10.2196/publichealth.8921] [Medline: 29351896]
34. Möller J, Stara V, Amabili G, et al. Toward innovation in healthcare: An analysis of the digital behavior of older people in Europe and Japan for the introduction of a technological coaching system. Healthcare (Basel) 2024 Jan 8;12(2):143. [doi: 10.3390/healthcare12020143] [Medline: 38255032]
35. Morishita-Suzuki K, Ogawa T, Bevilacqua R, et al. The impact of depression and leisure activities on e-health literacy among older adults: a cross-cultural study in the EU and Japan. Int J Environ Res Public Health 2025 Mar 10;22(3):403. [doi: 10.3390/ijerph22030403] [Medline: 40238518]
36. Wutz M, Hermes M, Winter V, Köberlein-Neu J. Factors influencing the acceptability, acceptance, and adoption of conversational agents in health care: integrative review. J Med Internet Res 2023 Sep 26;25:e46548. [doi: 10.2196/46548] [Medline: 37751279]
37. Zhang J, Gallifant J, Pierce RL, et al. Quantifying digital health inequality across a national healthcare system. BMJ Health Care Inform 2023 Nov 24;30(1):e100809. [doi: 10.1136/bmjhci-2023-100809] [Medline: 38007224]
38. Jahnel T, Dassow HH, Gerhardus A, Schüz B. The digital rainbow: digital determinants of health inequities. Digit Health 2022;8:20552076221129093. [doi: 10.1177/20552076221129093] [Medline: 36204706]
39. Ae-Ri J, Kowoon L, Eun-A P. Development and evaluation of the information and communication technology-based loneliness alleviation program for community-dwelling older adults: a pilot study and randomized controlled trial. Geriatr Nurs 2023;53:204-211. [doi: 10.1016/j.gerinurse.2023.07.011] [Medline: 37544264]
40. Sullivan J, Kosuth E. Technology use, barriers, and future needs among community-dwelling older adults. J Gerontol Nurs 2024 Feb;50(2):26-31. [doi: 10.3928/00989134-20240110-04] [Medline: 38290100]
41. Golinelli D, Pecoraro V, Tedesco D, et al. Population risk stratification tools and interventions for chronic disease management in primary care: a systematic literature review. BMC Health Serv Res 2025 Apr 10;25(1):526. [doi: 10.1186/s12913-025-12690-0] [Medline: 40205373]
42. Mishra US, Yadav S, Joe W. The Ayushman Bharat digital mission of India: an assessment. Health Syst Reform 2024 Dec 17;10(2):2392290. [doi: 10.1080/23288604.2024.2392290] [Medline: 39437234]
43. Liu M, Wang C, Hu J. Older adults' intention to use voice assistants: usability and emotional needs. Heliyon 2023 Nov;9(11):e21932. [doi: 10.1016/j.heliyon.2023.e21932]

XSL•FO
RenderX

44. Murdoch B. Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Med Ethics 2021 Sep 15;22(1):122. [doi: 10.1186/s12910-021-00687-3] [Medline: 34525993]
45. Lim GP, Appalasamy JR, Ahmad B, Quek KF, Shaharuddin S, Ramadas A. PEER-led digital health lifestyle intervention for a low-income community at risk for cardiovascular diseases (MYCardio-PEER): a quasi-experimental study protocol. Prim Health Care Res Dev 2025 Mar 3;26:e20. [doi: 10.1017/S1463423625000192] [Medline: 40025749]

Abbreviations

IH: internet health care
LCA: latent class analysis
OR: odds ratio
SEM: social ecological model
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TAM: technology acceptance model

Edited by A Stone; submitted 25.May.2025; peer-reviewed by G Amabili, J Kaner, K Mouloudj; accepted 26.Nov.2025; published 15.Jan.2026.

Please cite as:
Li R, Xu X, Li Q, Liu H, Zhou TT, Amhare AF, Liu P, Tang J, Wang W, Zheng F, Han J
Internet Health Care Service Use Behavioral Pattern Among Older Adults and the Role of the Technology Acceptance and Social Ecological Theory Model: Cross-Sectional Survey
J Med Internet Res 2026;28:e78037
URL: https://www.jmir.org/2026/1/e78037
doi:10.2196/78037

© Rui Li, Xinyu Xu, Qingsong Li, Haobiao Liu, Ting Zhou, Abebe Feyissa Amhare, Peiyu Liu, Jing Tang, Wei Wang, Fuju Zheng, Jing Han. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Therapeutic Effects of a WeChat Mini-Program on Metabolic Dysfunction–Associated Fatty Liver Disease: Randomized Controlled Trial

Chao Sun[1], MD, PhD; Guangyu Chen[2], MD, PhD; Cuicui Shi[1], MD, PhD; Haixia Cao[1], MD, PhD; Ruixu Yang[1], MD, PhD; Jing Zeng[1], MD, PhD; Xiaoyan Duan[1], MD, PhD; Xin Sun[2], MD; Jian-Gao Fan[1], MD, PhD

[1]Center for Fatty Liver Disease, Department of Gastroenterology, Xinhua Hospital, Shanghai Jiao Tong University School of Medicine, 1665 Kong Jiang Road, Shanghai, China

[2]Clinical Research and Innovation Unit, Xinhua Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Corresponding Author:
Jian-Gao Fan, MD, PhD
Center for Fatty Liver Disease, Department of Gastroenterology, Xinhua Hospital, Shanghai Jiao Tong University School of Medicine, 1665 Kong Jiang Road, Shanghai, China

Abstract

Background: For patients with metabolic dysfunction–associated fatty liver disease (MAFLD), weight loss is advised but challenging in practice. In China, there is a pronounced shortage of tailored digital lifestyle interventions for this population.

Objective: This study aimed to assess the effects of a WeChat mini-program-delivered lifestyle intervention on weight loss and hepatic steatosis among individuals with MAFLD who were overweight or obese.

Methods: Adults who are overweight or obese and have clinically diagnosed MAFLD with transient elastography examination were enrolled in this prospective randomized controlled trial. Patients were randomly assigned to receive either WeChat mini-program management (intervention group) or standard care (control group) at a 1:1 ratio. The intervention was structured around the development and implementation of personalized diet and exercise plans, supplemented by guided exercise video courses and reinforced through continuous monitoring and informational support. Body weight and clinical parameters were assessed at baseline and then at 6 months.

Results: A total of 89 patients met the inclusion criteria and were randomly assigned to the intervention group (n=45) or control group (n=44). Among the 89 patients with MAFLD, 60% (27/45) of them achieved a weight loss of ≥5%, and 24.4% (11/45) of them had a weight loss of ≥10% in the intervention group, which was greater than those in the control group (27/45 vs 7/44; relative risk [RR] 3.771, 95% CI 1.836 - 7.748; *P*<.001; 11/45 vs 3/44, RR 3.585, 95% CI 1.072 - 11.988; *P*=.02). Importantly, patients receiving the intervention were significantly more likely to achieve a ≥10% reduction or normalization of controlled attenuation parameter (CAP) than those in the control group (26/45 vs 14/44; RR 1.816, 95% CI 1.102 - 2.992; *P*=.01). After adjusting for key baseline covariates, multivariate analysis confirmed the intervention's positive effect on achieving a weight loss of ≥5% (OR [odds ratio] 8.380, 95% CI 2.886 - 24.331; *P*<.001) of ≥10% (OR 4.612, 95% CI 1.138 - 18.686; *P*=.03), as well as on CAP reduction of ≥10 % or normalization (OR 2.853, 95% CI 1.092 - 7.456; *P*=.03). In parallel, the intervention group presented greater reductions in liver enzymes (alanine aminotransferase, aspartate aminotransferase, and γ-glutamyl transpeptidase) and metabolic parameters (fasting insulin, hemoglobin A_{1c}, and triglyceride) than the control group (all *P*<.05). According to the fibrosis assessment, only the FibroScan-aspartate aminotransferase score decreased more in the intervention group than in the control group (median difference −0.06, 95% CI −0.13 to −0.01; *P*=.02), as compared to other non-invasive indicators.

Conclusions: Readily scalable in primary care and varied-resource settings, our WeChat mini-program-based intervention extends beyond weight loss to reduce hepatic steatosis and improve metabolic parameters, thereby addressing the critical gap in targeted MAFLD management in China with a low-cost model for high-burden populations. Nevertheless, larger future studies are needed to confirm these findings with greater precision and assess long-term sustainability.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2500100197; https://www.chictr.org.cn/showprojEN.html?proj=247294

(J Med Internet Res 2026;28:e76204) doi:10.2196/76204

XSL•FO
RenderX

KEYWORDS

metabolic dysfunction-associated fatty liver disease; WeChat mini-program; lifestyle intervention; weight loss; metabolic parameters; randomized controlled trial

Introduction

Background

Metabolic dysfunction–associated fatty liver disease (MAFLD) represents a significant and growing public health burden worldwide, paralleling the epidemics of obesity and type 2 diabetes mellitus [1]. It is a gradually progressive disease that evolves through a spectrum that begins with simple hepatic steatosis and advances to metabolic dysfunction-associated steatohepatitis (MASH), ultimately resulting in liver fibrosis, cirrhosis, and even hepatocellular carcinoma [1,2]. Despite its severe clinical sequelae, there are no pharmacologic therapies approved in China to reverse the histological progression of MAFLD, underscoring the critical role of nonpharmacological management [3].

Lifestyle intervention involving dietary adjustments and physical exercise is still the cornerstone of MAFLD management. Clinical guidelines confirm that a body weight reduction of 3%-5% can ameliorate hepatic steatosis, a 7%-10% loss can improve MASH, and over 10% may even lead to fibrosis regression [3]. However, the real-world implementation of these interventions faces significant challenges [4,5].

Traditional lifestyle interventions rely on face-to-face guidance from clinical doctors in hospital settings, which often results in poor execution and low compliance, limiting the coverage and sustainability of treatment and seriously affecting patient outcomes [6]. A cross-sectional survey from China revealed suboptimal self-management of lifestyle behaviors among adults with MAFLD [7]. Lack of motivation and real-time supervision are key contributing factors to the inadequate effectiveness of lifestyle interventions. In view of this, there is an urgent need for more convenient and effective measures to enhance the implementation and adherence of lifestyle changes for patients with MAFLD, thereby improving the prognosis.

With the popularization of the internet and the rapid development of artificial intelligence, digital therapeutics (DTx), such as mobile health apps or programs, have emerged as a promising solution to bridge this care gap [8,9]. By delivering personalized health information and professional guidance, DTx enables convenient, home-based self-management, which can correct unhealthy lifestyle habits and improve treatment adherence [10-12]. A meta-analysis of 8 studies demonstrated that DTx interventions lasting 4‑24 months achieved a clinically significant weight loss at a rate of 33%, accompanied by improvements in liver enzymes and reduced liver fat [13]. In China, the near-ubiquitous use of WeChat offers a unique platform for such interventions [14,15]. Its widespread adoption eliminates the need for additional app downloads, fosters user trust through a familiar interface, and facilitates seamless communication between health care providers and patients, creating an ideal ecosystem for closed-loop chronic disease management [16,17].

Despite this potential, there is a notable scarcity of WeChat mini-programs specifically designed and rigorously evaluated for structured lifestyle intervention in Chinese patients with MAFLD. To address this gap, we developed an innovative interactive WeChat mini-program, named the therapeutic lifestyle changes (TLCs) program, which is designed to assist patients in implementing sustainable lifestyle changes for the management of MAFLD. The program offers a suite of features, including weight loss goal setting, diet records and recommendations, and instructional exercise videos.

Objectives

Therefore, the aim of this study was to evaluate the efficacy of this novel TLC program in managing MAFLD. We hypothesized that, compared to participants receiving standard care, those engaged in the TLC intervention would demonstrate significantly greater improvements in reduction of liver fat content and body weight, as well as glucose and lipid metabolic parameters.

Methods

Study Design and Setting

A 6-month parallel, randomized controlled trial was conducted at Xinhua Hospital, affiliated with Shanghai Jiao Tong University School of Medicine, a tertiary-care hospital in China. The study period for participant recruitment and data collection spanned from August 2022 to May 2024 at our hospital. All eligible participants were randomly assigned to either the TLC program management group (intervention group) or the standard care group (control group) at a 1:1 ratio. The protocol (which was retrospectively registered) with no deviations was approved by the Institutional Review Board (IRB) prospectively. The reporting of this trial adhered to the CONSORT (Consolidated Standards of Reporting Trials) statement [18], with the checklist presented in Checklist 1.

Participants

The participants with clinically diagnosed MAFLD were consecutively recruited by clinicians from our hepatology clinic. Those who were interested provided written informed consent. All participants underwent an initial, face-to-face eligibility assessment at Xinhua Hospital according to the study criteria. The inclusion criteria were as follows: (1) aged 18‑65 years, (2) BMI between 24 and 35 kg/m^2, (3) clinically diagnosed with MAFLD by radiological features with ultrasound confirmation and a controlled attenuation parameter (CAP) value of ≥248 dB/m, and (4) able to learn and operate a mobile-based program. Patients were excluded if they had excessive alcohol consumption (weekly ethanol intake of ≥210 g for males and ≥140 g for females weekly), etiological evidence of an alternative liver disease, such as chronic viral hepatitis, autoimmune liver diseases, drug-induced hepatic steatosis, hepatic decompensation, severe comorbidities (severe cardiopulmonary disease, uncontrolled diabetes, chronic kidney

XSL•FO
RenderX

disease, psychiatric disease, and malignant tumors), recent weight loss, or weight loss medication intake.

Randomization

After the completion of baseline assessments, eligible participants were registered in the study. They were then automatically and randomly allocated to each study group at a 1:1 ratio by a computer-generated random system. The allocation was concealed until the moment of assignment, as the outcome was only revealed by the system after the participant's formal registration. Given the nature of the digital intervention, participants and health care providers were not blinded to group assignment, but outcome assessors were blinded during data analysis.

Lifestyle Intervention

"Design of the TLC program"

The TLC program-based lifestyle intervention was collaboratively designed by physicians from various departments, including Gastroenterology, Sport Rehabilitation, and Clinical Nutrition. Individuals in the control group were provided with standard care in the clinic according to the established treatment guidelines [19]. The TLC program intervention spanned 24 weeks and was divided into 4 distinct modules, which are described below.

Development of Diet and Exercise Plans

Individualized weight loss and exercise goals were set based on the baseline body weight, existing metabolic disorders, and severity of fatty liver disease. The target for weight reduction was established as 5%-10% of the initial body weight.

Dietary Intervention Plan

Based on preset weight loss goals, the program calculated daily dietary energy requirements, providing a recommended meal plan with the corresponding energy intake. If patients did not prefer the recipes suggested by the program, they could use the food exchange list to select alternative food with equivalent energy values. The program required patients to log and upload photos of their daily food intake, after which a human nutritionist analyzed the diet log and provided feedback via WeChat.

Exercise Video Courses

Participants were provided with a heart rate armband (HW702, YESOUL) to track heart rate during exercise. During each exercise session, data regarding heart rate were automatically uploaded by the armband. Each course lasted 30 minutes, and patients were to choose the training course at least 5 times per week. Additionally, patients could select other forms of exercise and record the duration of exercise, such as fast walking, running, cycling, and swimming.

Monitoring and Information Support

Patients were expected to complete their diet and exercise tasks through a check-in system every day. The research team sent tailored feedback messages weekly according to the status of task completion and offered regular suggestions through WeChat. Moreover, the system intelligently provided relevant educational articles, which provided professional information and support tailored to the patient's interests and concerns.

Measurements

Clinical Characteristics

Clinical characteristics regarding anthropometric and laboratory data were obtained at our hospital using a standard protocol at baseline and 6 months. Body composition was assessed via a bioimpedance analyzer (MC-980MA, TANITA). CAP and liver stiffness measurement (LSM) were performed using the FibroScan-502 device (Echosens) with an M or XL probe, following the manufacturer's guidelines. A reliable examination was defined as at least 10 valid measurements and an IQR to median ratio of LSM below 30%. Based on an individual patient data meta-analysis, CAP thresholds of 268 and 280 dB/m were identified as optimal cutoffs for diagnosing moderate and severe steatosis, respectively [20].

Dietary Nutrition and Physical Activity Assessment

Dietary nutrition was assessed through a face-to-face interview with a questionnaire developed by the Chinese Society of Health Management and the Chinese Nutrition Society, which consists of 24 questions with a total score of 100 [21]. A score below 60 indicates a risk of dietary nutrition issues; a score between 60 and 75 suggests a potential risk; and a score above 75 indicates no dietary nutritional risk. Physical activity levels were evaluated via the International Physical Activity Questionnaire-Short Form (IPAQ-SF), which captures data on activity intensity, frequency, and daily duration in a typical week. Energy expenditure was quantified by the metabolic equivalent of task (MET) [22]. The weekly physical activity level for a specific intensity was calculated as the MET value of the activity× weekly frequency (d/wk)× daily duration (min/d).

Diagnostic Criteria

Obesity was considered a BMI≥28 kg/m^2 [3]. The definition of MetS adhered to previously published criteria [23]. The homeostasis model assessment-insulin resistance (HOMA-IR) score was calculated as fasting insulin (mU/L)× fasting plasma glucose (FPG, mmol/L)/22.5. Insulin resistance was characterized by a HOMA-IR score of 2.5 or higher [1]. The fatty liver index was computed to indirectly assess the degree of hepatic steatosis using an established formula [24]. The FibroScan-aspartate aminotransferase (FAST) score was used to identify patients with MASH with significant fibrosis [25], whereas the aspartate aminotransferase-to-platelet ratio index (APRI), fibrosis-4 (FIB-4) index, nonalcoholic fatty liver disease fibrosis score (NFS), and Agile 3+ score were used to assess the risk of advanced fibrosis [26-29].

Outcome Measurements

The analyses included all participants who were randomized. The intention-to-treat (ITT) population for efficacy evaluation in this trial consisted of all cases that were randomized and had efficacy baseline records. The per-protocol (PP) population included participants who completed the entire planned observation period and had no major protocol deviations. The primary efficacy endpoint was the reduction of body weight by

XSL•FO
RenderX

at least 5%. The secondary endpoints included a CAP value of ≥10% reduction or normalization and changes in the levels of liver enzymes and lipid and glucose parameters.

Sample Size Calculation

The sample size calculation was based on the primary study outcome, which is the proportion of patients achieving a weight loss of ≥5% from baseline. It was postulated that the intervention would produce a 5-fold increase in the success rate from a baseline of 8% in the control group [30]. To detect this effect with 80% power at a 5% significance level and accounting for a 20% dropout rate, a total sample size of 84 participants was required.

Data Analysis

Statistical analyses were performed using SPSS 23.0 software (IBM Corp). Missing data were due to patient dropouts and imputed by the last observation carried forward. Regarding the primary endpoint, this approach was chosen as it provides a conservative estimate of the treatment effect by assuming no further improvement in their weight status due to the discontinuation of active intervention, thereby avoiding overoptimistic extrapolation [6,31,32]. Descriptive statistics were conducted for all variables, with means (95% CIs) or medians (IQRs) reported for continuous variables and percentages for categorical variables. Categorical variables were analyzed using the chi-square test or Fisher exact test between groups (reporting relative risk and 95% CI) and the marginal homogeneity test within each group. Continuous variables were assessed using 2-tailed paired *t* tests or Wilcoxon signed-rank tests within each group, whereas unpaired *t* tests (reporting mean difference and 95% CI) or Mann-Whitney *U* tests (reporting Hodges-Lehmann median difference and 95% CI) were used to compare the changes from the baseline between groups. Binary logistic regression was used to evaluate the effect of the intervention on weight and CAP changes. The results are presented as odds ratios (ORs) with 95% CIs. A 2-tailed *P* value of .05 was considered significant.

Ethical Considerations

This research was approved by the IRB of Xinhua Hospital (XHEC-C-2021-076-2) and was retrospectively registered with the Chinese Clinical Trial Registry (ChiCTR2500100197). This trial was registered retrospectively owing to disruptions in the research team's workflow and uncertainties about the study's continuity during the COVID-19 pandemic in China. We confirm that all data collection for this study was initiated subsequent to receiving approval from the IRB of Xinhua Hospital, with the IRB approval document provided in Multimedia Appendix 1. The study start was delayed due to COVID-19, and there were no other deviations or changes from the protocol after IRB approval and trial commencement. All eligible participants provided written informed consent. Participants were informed of their voluntary participation and their right to withdraw from the study at any time. It was clarified that all study procedures were offered at no cost, but no financial compensation would be made upon study completion. All personally identifiable information was coded with a unique study ID, and the anonymized data were securely stored in a password-protected cabinet.

Results

Participants and Baseline Characteristics

Figure 1 presents the study flowchart in accordance with the CONSORT guidelines. A total of 89 patients met the inclusion criteria and were randomly assigned to the intervention group (n=45) or control group (n=44); these patients comprised the ITT population. At the 6-month follow-up point, 8 participants were lost to follow-up in the intervention group and 8 in the control group. A total of 73/89 (82%) patients successfully completed the study and comprised the PP population, while the remaining 16/89 (18%) patients dropped out (Figure 1). The baseline characteristics of the patients are shown in Table 1. The mean age of the whole study population was 40.0 (SD 9.8) years. Around 61.8% (55/89) were males and 55.1% (49/89) were obese (BMI≥28 kg/m^2). Overall, 34.8% (31/89) had hypertension, 70.8% (63/89) had dyslipidemia, and 6.7% (6/89) had type 2 diabetes mellitus. There were no significant differences between the 2 groups in terms of age, sex, weight, presence of comorbidities, biochemical variables, diet, or activity-related parameters (*P*>.05). Additionally, no difference was found in the CAP or LSM between the groups at baseline (*P*>.05).

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram.

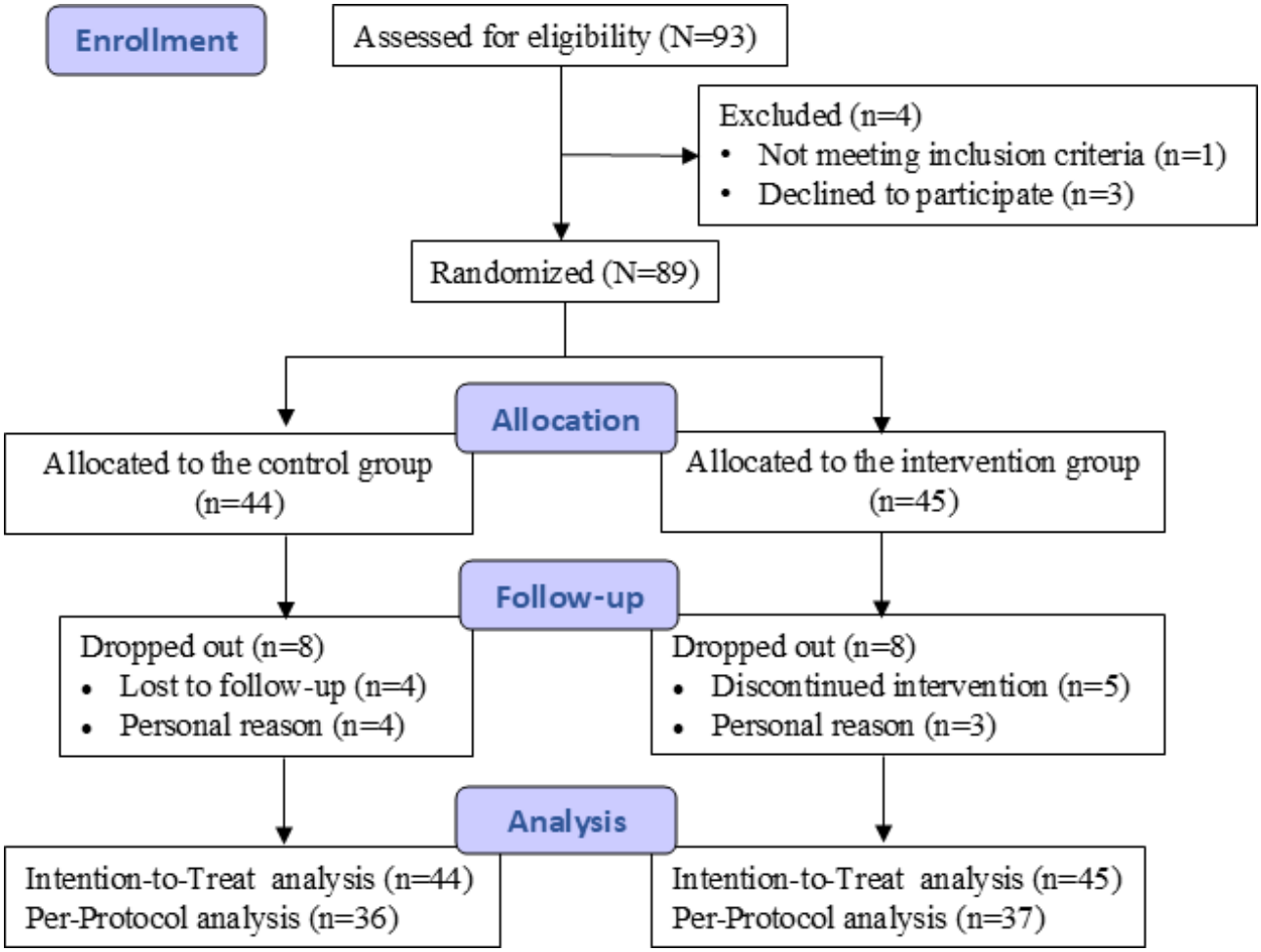

XSL•FO
RenderX

Table . Baseline characteristics of the total study population and the randomized allocation groups in intention-to-treat population.

Characteristic	Total (N=89)	Control (n=44)	Intervention (n=45)	*P* value[a]
Sex (male/female), n(%)	55 (61.8)/ 34(38.2)	31 (70.5)/ 13(29.5)	24 (53.3)/ 21(46.7)	.10
Age (years), mean (SD; 95% CI)	40.0 (9.8; 37.9-42.1)	40.7 (10.5; 37.5-43.9)	39.3 (9.1; 36.6-42.1)	.84
Weight (kg), mean (SD; 95% CI)	82.5 (11.6; 80.1-85.0)	82.3 (11.2; 78.9-85.7)	82.7 (12.1; 79.1-86.4	.92
BMI (kg/m^2), mean (SD; 95% CI)	28.6 (3.0; 27.9-29.2	28.2 (3.0; 27.2-29.1)	29.0 (2.9; 28.1-29.8)	.25
BMI≥28 (kg/m^2), n (%)	49 (55.1)	21 (47.7)	28 (50.9)	.75
WC[b] (cm), mean (SD; 95% CI)	92.8 (8.9; 90.9-94.7)	92.6 (8.9; 89.9-95.3)	93.0 (9.0; 90.3-95.7)	.97
Waist to hip ratio, mean (SD; 95% CI)	0.90 (0.06; 0.89-0.92)	0.91 (0.06; 0.89-0.93)	0.90 (0.07; 0.88-0.92)	.38
Waist to height ratio, mean (SD; 95% CI)	0.55 (0.05; 0.54-0.56)	0.54 (0.05; 0.53-0.56)	0.55 (0.04; 0.54-0.56)	.45
Systolic blood pressure (mmHg), mean (SD; 95% CI)	123.8 (6.8; 122.4-125.3)	124.4 (6.7; 122.4-126.5)	123.3 (7.0; 121.2-125.4)	.24
Diastolic blood pressure (mmHg), mean (SD; 95% CI)	79.3 (5.7; 78.1-80.5)	79.6 (5.9; 77.9-81.4)	79.0 (5.6; 77.3-80.7)	.35
Hypertension, n (%)	31 (34.8)	17 (38.6)	14 (31.1)	.46
Type 2 diabetes mellitus, n (%)	6 (6.7)	4 (9.1)	2 (4.4)	.43
Dyslipidemia, n (%)	63 (70.8)	33 (75)	30 (66.7)	.39
Hyperuricemia, n (%)	34 (38.2)	18 (40.9)	16 (35.6)	.60
Metabolic syndrome, n (%)	37 (41.6)	19 (43.2)	18 (40)	.76
Platelet count (10^9/L), mean (SD; 95% CI)	256.6 (51.7; 245.8-267.5)	254.5 (51.9; 238.7-270.3)	258.8 (51.9; 243.2-274.4)	.58
ALT[c] (U/L), median (IQR)	36.0 (26.0-59.0)	38.0 (24.5-57.3)	34.0 (26.5-61.0)	.90
AST[d] (U/L), median (IQR)	26.0 (21.0-36.0)	26.0 (21.3-35.5)	25.0 (19.5-36.0)	.68
GGT[e] (U/L), median (IQR)	35.0 (23.0-53.5)	39.0 (25.3-59.8)	34.0 (22.0-50.0)	.16
Albumin (g/L), mean (SD; 95% CI)	46.3 (2.4; 45.8-46.8)	46.6 (2.0; 46.0-47.2)	45.9 (2.6; 45.1-46.7)	.20
Urea (mmol/L), mean (SD; 95% CI)	4.9 (1.1; 4.7-5.1)	5.0 (1.1; 4.6-5.3)	4.9 (1.0; 4.6-5.2)	.75
Serum creatinine (μmol/L), mean (SD; 95% CI)	66.8 (14.4; 63.8-70.0)	68.9 (14.0; 64.6-73.1)	64.9 (14.6; 60.5-69.3)	.18
Uric acid (μmol/L), mean (SD; 95% CI)	410.7 (109.9; 387.5-433.8)	416.1 (112.5; 381.9-450.2)	405.4 (108.3; 372.9-437.9)	.62
Fasting plasma glucose (mmol/L), mean (SD; 95% CI)	5.3 (0.6; 5.2-5.5)	5.4 (0.6; 5.2-5.5)	5.3 (0.6; 5.1-5.5)	.54
Fasting insulin (pmol/L), mean (SD; 95% CI)	78.0 (39.3; 69.7-86.2)	80.1 (45.6; 66.2-94.0)	76.0 (32.4; 66.2-85.7)	.93
HOMA-IR[f], median (IQR)	2.4 (1.7-3.3)	2.4 (1.6-3.4)	2.4 (1.7-3.2)	.96
HOMA-IR≥2.5, n (%)	40 (44.9)	19 (43.2)	21 (46.7)	.74

XSL•FO
RenderX

Characteristic	Total (N=89)	Control (n=44)	Intervention (n=45)	*P* value[a]
Hemoglobin A_{1c} (%), median (IQR)	5.6 (5.4-5.8)	5.6 (5.2-5.8)	5.6 (5.4-5.9)	.36
TG[g] (mmol/L), median (IQR)	1.6 (1.0-2.0)	1.6 (1.0-2.2)	1.5 (1.0-1.8)	.43
TC[h] (mmol/L), median (IQR)	4.9 (4.2-5.3)	4.9 (4.2-5.4)	5.0 (4.3-5.4)	.78
HDL-C[i] (mmol/L), median (IQR)	1.2 (1.0-1.3)	1.1 (1.0-1.3)	1.2 (1.0-1.4)	.26
LDL-C[j] (mmol/L), median (IQR)	3.2 (2.6-3.7)	3.3 (2.6-3.8)	3.2 (2.5-3.6)	.68
CAP[k] (dB/m), median (IQR)	334 (295-365)	336 (288-362)	334 (297-369)	.25
Liver stiffness measurement (kPa), median (IQR)	5.1 (4.4-6.5)	5.2 (4.4-6.4)	4.9 (4.4-6.7)	.56
Diet quality score, median (IQR)	57 (50-66)	56 (49-66)	58 (51-66)	.64
Total MET[l]-minutes, median (IQR)	1116 (544-2057)	1118 (620-2068)	1102 (465-2088)	.59

[a]Continuous variables were assessed using unpaired t tests (normally distributed data) and Mann-Whitney *U* tests (non-normally distributed data) between groups. Categorical variables were analyzed using the chi-square test or Fisher exact test between groups.

[b]WC: waist circumference.

[c]ALT: alanine aminotransferase.

[d]AST: aspartate aminotransferase.

[e]GGT: γ-glutamyl transpeptidase.

[f]HOMA-IR: homeostasis model assessment-insulin resistance.

[g]TG: triglyceride.

[h]TC: total cholesterol.

[i]HDL-C: high-density lipoprotein-cholesterol.

[j]LDL-C: low-density lipoprotein-cholesterol.

[k]CAP: controlled attenuation parameter.

[l]MET: metabolic equivalent of task.

Changes in Clinical Parameters

Table 2 and Figure 2 present the clinical parameters at baseline and 24 weeks and their changes between groups in intention-to-treat population. There was a significant reduction in weight, waist circumference, waist-to-height ratio, and BMI from baseline within each group (*P*<.05). The reductions in these factors were also significantly greater in the intervention group than in the control group (*P*<.05). Patients in the TLC program intervention group had lower alanine aminotransferase (ALT), aspartate aminotransferase (AST), and γ-glutamyl transpeptidase (GGT) levels at 6 months than at baseline, and these liver enzyme levels decreased more in the intervention group than in the control group (*P*<.05). At 6 months, reductions in the levels of glucose metabolism–related factors, including fasting insulin, hemoglobin A_{1c} (HbA_{1c}), and HOMA-IR, were observed only within the intervention group (*P*<.05), and the levels of these factors showed significantly greater reductions in the intervention group than in the control group (*P*<.05). However, there was no between-group difference in FPG (*P*=.22). Patients in the intervention group had greater reductions in triglyceride (median difference −0.5, 95% CI −0.7 to −0.2; *P*=.001) and increases in high-density lipoprotein-cholesterol (HDL-C) levels than did those in the control group (median difference 0.1, 95% CI 0 - 0.1; *P*=.02). The diet quality score improved more in the TLC program group than in the control group (median difference 6, 95% CI 0 - 10; *P*=.04), whereas MET did not significantly differ between the groups (*P*=.92).

XSL•FO
RenderX

Table . Clinical parameters between the two study groups at baseline and 24 weeks in the intention-to-treat population.

Parameter	Control (n=44)			Intervention (n=45)			Mean or Hodges-Lehmann median difference (95% CI)[a]	*P* value[a]
	Baseline	24 weeks	*P* value[b]	Baseline	24 weeks	*P* value[b]		
Weight (kg), mean (SD; 95% CI)	82.3 (11.2; 78.9-85.7)	81.4 (12.2 ;77.7-85.0)	.02	82.7 (12.1; 79.1-86.4)	77.2 (11.7; 73.7-80.8)	<.001	–3.7 (–5.5 to –1.7)	<.001
BMI (kg/m^2), mean (SD; 95% CI)	28.2 (3.0; 27.2-29.1)	27.5 (3.4; 26.5-28.6)	.004	29.0 (2.9; 28.1-29.8)	26.9 (2.9; 26.0-27.8)	<.001	–1.4 (–2.1 to –0.7)	<.001
WC[c](cm), mean (SD; 95% CI)	92.6 (8.9; 89.9-95.3)	90.6 (8.6; 88.0-93.2)	.003	93.0 (9.0; 90.3-95.7)	85.3 (9.5; 82.7-88.4)	<.001	–5.0 (–7.0 to –2.0)	<.001
Waist to hip ratio, mean (SD; 95% CI)	0.91 (0.06; 0.89-0.93)	0.90 (0.06; 0.89-0.92)	.21	0.90 (0.07; 0.88-0.92)	0.87 (0.07; 0.85-0.89)	<.001	–0.02 (–0.04 to 0.01)	.06
Waist to height ratio, mean (SD; 95% CI)	0.54 (0.05; 0.53-0.56)	0.53 (0.05; 0.52-0.55)	.003	0.55 (0.04; 0.54-0.56)	0.50 (0.05; 0.49-0.52)	<.001	–0.03 (–0.04 to –0.02)	<.001
Systolic blood pressure (mmHg), mean (SD; 95% CI)	124.4 (6.7; 122.4-126.5)	125.6 (9.9; 122.6-128.6)	.98	123.3 (7.0; 121.2-125.4)	123.4 (11.3; 120.0-126.8)	.17	–1.0 (–6.0 to 4.0)	.87
Diastolic blood pressure (mmHg), mean (SD; 95% CI)	79.6 (5.9; 77.9-81.4)	81.1 (5.9; 79.3-82.9)	.18	79.0 (5.6; 77.3-80.7)	79.6 (5.8; 77.8-81.3)	.47	–0.9 (–3.7 to 1.9)	.62
Platelet count (10^9/L), mean (SD; 95% CI)	254.5 (51.9; 238.7-270.3)	256.0 (66.5; 235.7-276.2)	.92	258.8 (51.9; 243.2-274.4)	250.0 (44.3; 237.0-262.0)	.03	–12.8 (–26.7 to 1.1)	.07
ALT[d] (U/L), median (IQR)	38.0 (24.5-57.3)	30.0 (22.3-44.3)	.01	34.0 (26.5-61.0)	23.0 (17.0-33.5)	<.001	–8.0 (–15.0 to –1.0)	.02
AST[e] (U/L), median (IQR)	26.0 (21.3-35.5)	22.0 (19.0-27.8)	.003	25.0 (19.5-36.0)	19.0 (15.0-23.0)	<.001	–3.0 (–6.0 to 0)	.04
GGT[f] (U/L), median (IQR)	39.0 (25.3-59.8)	41.0 (25.3-57.5)	.76	34.0 (22.0-50.0)	23.0 (16.5-34.5)	<.001	–8.0 (–14.0 to –3.0)	<.001
Albumin (g/L), mean (SD; 95% CI)	46.6 (2.0; 46.0-47.2)	45.7 (2.1; 44.8-46.0)	.001	45.9 (2.6; 45.1-46.7)	45.1 (4.0; 43.9-46.3)	.15	0.4 (–0.8 to 1.6)	.06
Urea (mmol/L), mean (SD; 95% CI)	5.0 (1.1; 4.6-5.3)	5.0 (1.2; 4.6-5.4)	.40	4.9 (1.0; 4.6-5.2)	5.0 (1.1; 4.7-5.3)	.61	0.1 (–0.3 to 0.4)	.95
Serum creatinine (μmol/L), mean (SD; 95% CI)	68.9 (14.0; 64.6-73.1)	68.7 (13.8; 64.5-72.9)	.85	64.9 (14.6; 60.5-69.3)	64.5 (14.0; 60.3-68.7)	.78	–0.2 (–2.2 to 1.8)	.79

XSL•FO
RenderX

Parameter	Control (n=44)			Intervention (n=45)			Mean or Hodges-Lehmann median difference (95% CI)[a]	*P* value[a]
	Baseline	24 weeks	*P* value[b]	Baseline	24 weeks	*P* value[b]		
Uric acid (μmol/L), mean (SD; 95% CI)	416.1 (112.5; 381.9-450.2)	404.1 (75.2; 380.4-427.2)	.57	405.4 (108.3; 372.9-437.9)	386.8 (105.6; 355.0-418.5)	.008	–6.6 (–36.1 to 22.8)	.16
Fasting plasma glucose (mmol/L), mean (SD; 95% CI)	5.4 (0.6; 5.2-5.5)	5.5 (0.9; 5.3-5.8)	.39	5.3 (0.6; 5.1-5.5)	5.3 (0.5; 5.1-5.4)	.24	–0.1 (–0.5 to 0.3)	.22
Fasting insulin (pmol/L), mean (SD; 95% CI)	80.1 (45.6; 66.2-94.0)	81.0 (40.8; 68.6-93.4)	.35	76.0 (32.4; 66.2-85.7)	60.1 (34.8; 49.6-70.6)	.001	–16.8 (–30.0 to –3.5)	.002
HOMA-IR[g], median (IQR)	2.4 (1.6-3.4)	2.7 (1.9-3.7)	.25	2.4 (1.7-3.2)	1.7 (1.1-2.8)	.001	–0.7 (–1.1 to –0.2)	.002
HbA_{1c}[h] (%), median (IQR)	5.6 (5.2-5.8)	5.5 (5.3-5.7)	.58	5.6 (5.4-5.9)	5.4 (5.3-5.7)	.003	–0.1 (–0.2 to 0)	.047
TG[i] (mmol/L), median (IQR)	1.6 (1.0-2.2)	1.9 (1.2-2.4)	.09	1.5 (1.0-1.8)	1.0 (0.8-1.5)	.001	–0.5 (–0.7 to –0.2)	.001
TC[j] (mmol/L), median (IQR)	4.9 (4.2-5.4)	4.6 (4.0-5.2)	.15	5.0 (4.3-5.4)	4.5 (3.9-5.1)	.009	–0.1 (–0.4 to 0.1)	.32
HDL-C[k] (mmol/L), median (IQR)	1.1 (1.0-1.3)	1.1 (0.9-1.3)	.21	1.2 (1.0-1.4)	1.3 (1.0-1.5)	.07	0.1 (0-0.1)	.02
LDL-C[l] (mmol/L), median (IQR)	3.3 (2.6-3.8)	3.1 (2.6-3.6)	.06	3.2 (2.5-3.6)	3.1 (2.3-3.5)	.04	–0.1 (–0.2 to 0.1)	.75
CAP[m] (dB/m), median (IQR)	336 (288-362)	310 (288-342)	.04	334 (297-369)	291 (239-317)	<.001	–38 (–62 to –16)	.001
CAP<268, n (%)	4 (9.1)	6 (13.6)	.55	1 (2.2)	16 (35.6)	<.001	—[n]	—
268≤CAP<280, n (%)	4 (9.1)	3 (6.8)	—	1 (2.2)	4 (8.9)	—	—	—
CAP≥280, n (%)	36 (81.8)	35 (79.5)	—	43 (95.6)	25 (55.6)	—	—	—
Fatty liver index, mean (SD; 95% CI)	58.4 (22.3; 51.6-65.2)	57.1 (22.2; 50.3-63.8)	.38	56.7 (21.8; 50.2-63.3)	34.2 (22.3; 27.5-40.9)	<.001	–21.1 (–28.4 to –13.9)	<.001
Liver stiffness measurement (kPa), median (IQR)	5.2 (4.4-6.4)	5.1 (4.1-6.8)	.07	4.9 (4.4-6.7)	5.1 (4.0-6.0)	.16	–0.1 (–0.6 to 0.5)	.94
AST-to-platelet ratio index, median (IQR)	0.28 (0.20-0.37)	0.24 (0.18-0.30)	.008	0.27 (0.19-0.35)	0.21 (0.16-0.27)	<.001	–0.03 (–0.06 to 0.01)	.13
Fibrosis-4 index, median (IQR)	0.68 (0.50-1.00)	0.66 (0.46-0.99)	.20	0.64 (0.49-0.90)	0.62 (0.49-0.76)	.15	–0.05 (–0.12 to 0.03)	.85

XSL•FO
RenderX

Parameter	Control (n=44)			Intervention (n=45)			Mean or Hodges-Lehmann median difference (95% CI)[a]	*P* value[a]
	Baseline	24 weeks	*P* value[b]	Baseline	24 weeks	*P* value[b]		
NAFLD[o] fibrosis score, mean (SD; 95% CI)	–2.9 (1.2; –3.2 to –2.5)	–2.6 (1.2; –3.0 to –2.3)	.09	–2.9 (1.0; –3.2 to –2.6)	–2.7 (1.0; –3.0 to –2.4)	.15	–0.1 (–0.4 to 0.3)	.96
FAST[p], median (IQR)	0.20 (0.11-0.42)	0.11 (0.07-0.22)	.001	0.22 (0.09-0.43)	0.06 (0.03-0.15)	<.001	–0.06 (–0.13 to –0.01)	.02
Agile3+, median (IQR)	0.06 (0.02-0.15)	0.05 (0.02-0.15)	.96	0.05 (0.02-0.12)	0.05 (0.03-0.12)	.63	0.01 (–0.02 to 0.03)	.68
Diet quality score, median (IQR)	56 (49-66)	64 (55-71)	.01	58 (51-66)	66 (61-76)	<.001	6 (0-10)	.04
Total MET[q]-minutes, median (IQR)	1118 (620-2068)	1442 (620-2608)	.29	1102 (465-2088)	1102 (647-1725)	.44	–10 (–204 to 192)	.92

[a]Continuous variables were assessed using 2-tailed unpaired *t* tests (reporting mean difference and 95% CI) for normally distributed data and Mann-Whitney *U* tests (reporting Hodges-Lehmann median difference and 95% CI) for non-normally distributed data to compare the changes between the two groups.

[b]Continuous variables were assessed using paired *t* tests (normally distributed data) or Wilcoxon signed-rank tests (non-normally distributed data) within each group. Categorical variables were analyzed using marginal homogeneity tests within each group.

[c]WC: waist circumference.

[d]ALT: alanine aminotransferase.

[e]AST: aspartate aminotransferase.

[f]GGT: γ-glutamyl transpeptidase.

[g]HOMA-IR: homeostasis model assessment-insulin resistance.

[h]HbA_{1c}: hemoglobin A_{1c}.

[i]TG: triglyceride.

[j]TC: total cholesterol.

[k]HDL-C: high-density lipoprotein-cholesterol.

[l]LDL-C: low-density lipoprotein-cholesterol.

[m]CAP: controlled attenuation parameter.

[n]Not applicable.

[o]NAFLD: nonalcoholic fatty liver disease.

[p]FAST: FibroScan-aspartate aminotransferase.

[q]MET: metabolic equivalent of task.

XSL•FO
RenderX

Figure 2. Change in clinical parameters from baseline to 24 weeks between control group and intervention group. (A) Median change in weight between groups (–0.3 vs –4.4; median difference –3.7, 95% CI –5.5 to –1.7; *P*<.001). (B) Median change in waist circumference between groups (–1.5 vs –7.0; median difference –5.0, 95% CI –7.0 to –2.0; *P*<.001). (C) Median change in alanine aminotransferase values between groups (–2.5 vs –12.0; median difference –8.0, 95% CI –15.0 to –1.0; *P*=.02). (D) Median change in aspartate aminotransferase values between groups (–2.0 vs –6.0; median difference –3.0, 95% CI –6.0 to 0; *P*=.04). (E) Median change in hemoglobin A_{1c} between groups (0 vs –0.1; median difference –0.1, 95% CI –0.2 to 0; *P*=.047). (F) Median change in homeostasis model assessment-insulin resistance between groups (0 vs –0.6; median difference –0.7, 95% CI –1.1 to –0.2; *P*=.002). (G) Median change in triglyceride values between groups (0 vs –0.3; median difference –0.5, 95% CI –0.7 to –0.2; *P*=.001). (H) Median change in high-density lipoprotein-cholesterol values between groups (0 vs 0.1; median difference 0.1, 95% CI 0-0.1; *P* =.02). (I) Median change in controlled attenuation parameter values between groups (–4.0 vs –56.0; median difference –38, 95% CI –62 to –16; *P*=.001). (J) Median change in FibroScan-aspartate aminotransferase scores between groups (–0.04 vs –0.11; median difference –0.06, 95% CI –0.13 to –0.01; *P* =.02). ALT: alanine aminotransferase; AST: aspartate aminotransferase; CAP: controlled attenuation parameter; FAST: FibroScan–aspartate aminotransferase; HbA_{1c}: hemoglobin A_{1c}; HDL-C: high-density lipoprotein cholesterol; HOMA-IR: homeostasis model assessment–insulin resistance; TG: triglyceride; WC: waist circumference.

XSL•FO
RenderX

Changes in Noninvasive Factors of Hepatic Steatosis and Liver Fibrosis

At 6 months, both groups presented a reduction in the CAP value. The intervention group experienced a decrease to 291(239-317) dB/m, whereas the control group had a median CAP of 310 (288-342)dB/m (Table 2, *P*=.001). CAP values were significantly lower in the intervention group than in the control group (Figure 2, median difference −38, 95% CI −62 to −16; *P*=.001). In the subgroup analysis, the percentage of patients with severe steatosis decreased significantly in the intervention group from baseline at 6 months (*P*<.001). However, the LSM values were not significantly different between groups or within each group (*P*>.05). Based on the noninvasive assessment, patients in the TLC program intervention group had lower FAST scores than did those in the control group (median difference −0.06, 95% CI −0.13 to −0.01; *P*=.02). In the comparison of these 2 groups, no significant differences were found in noninvasive scores, including the APRI, FIB-4, NFS, and Agile 3+ score (*P*>.05).

Changes in Body Composition

According to the body composition analysis (Table 3), each group presented a decrease in total body fat mass, total body fat percentage, and visceral fat level, and these variables decreased more in the intervention group than in the control group across the 6-month period (*P*<.05). The basal metabolic rate underwent a greater reduction in the intervention group than in the control group (mean difference −42, 95% CI −62 to −22; *P*<.001). The intervention group also presented more pronounced decreases in fat-free mass, total lean body mass, and appendicular skeletal muscle mass (ASM; *P*<.05). The ratio of ASM to weight increased in the intervention group, but no significant difference was observed in this ratio between the groups (*P*>.05).

XSL•FO
RenderX

Table . Body composition analysis between the two study groups at baseline and 24 weeks in the intention-to-treat population.

Body composition	Control n=44			Intervention n=45			Mean or Hodges-Lehmann median difference (95% CI)[a]	*P* value[a]
	Baseline	24weeks	*P* value[b]	Baseline	24 weeks	*P* value[b]		
Total body fat mass (kg), mean (SD; 95% CI)	25.3 (6.8; 23.3-27.4)	24.3 (7.0; 22.2-26.4)	.006	28.3 (7.3; 26.2-30.6)	24.6 (7.3; 22.4-26.7)	<.001	–2.8 (–4.4 to –1.2)	<.001
Total body fat percentage (%), mean (SD; 95% CI)	30.7 (6.9; 28.6-32.8)	29.9 (6.8; 27.9-32.0)	.002	34.3 (7.8; 32.0-36.7)	30.9 (8.5; 28.3-33.5)	<.001	–2.6 (–4.1 to –1.2)	<.001
Visceral fat level, median (IQR)	13 (11-16)	12 (10-14)	.001	12 (10-15)	10 (8-13)	<.001	–1 (–2 to 0)	<.001
Fat-free mass (kg), mean (SD; 95% CI)	57.0 (9.1; 54.2-59.7)	57.1 (9.7; 54.1-60.0)	.44	54.3 (10.6; 51.1-57.5)	52.7 (10.1; 49.6-55.7)	<.001	–1.1 (–1.9 to –0.2)	.01
Total lean body mass (kg), mean (SD; 95% CI)	53.4 (9.2; 50.6-56.1)	52.7 (9.3; 49.9-55.5)	.29	51.0 (10.0; 48.0-54.0)	49.9 (9.9; 47.0-52.9)	<.001	–0.7 (–1.2 to –0.2)	.002
ASM[c] (kg), mean (SD; 95% CI)	25.5 (5.4; 23.8-27.1)	25.0 (5.8; 23.2-26.8)	.70	24.8 (5.9; 23.0-26.6)	23.9 (5.7; 22.2-25.6)	<.001	–0.6 (–1.0 to –0.3)	.001
ASM to weight ratio (%), mean (SD; 95% CI)	30.7 (3.8; 29.6-31.9)	30.8 (5.3; 29.1-32.3)	.009	29.8 (4.4; 28.5-31.1)	30.8 (4.5; 29.5-32.1)	<.001	0.5 (–0.1 to 1.1)	.08
Trunk muscle mass (kg), mean (SD; 95% CI)	27.9 (4.1; 26.6-29.1)	27.7 (4.1; 26.5-29.0)	.31	26.2 (4.5; 24.9-27.6)	26.0 (4.4; 24.7-27.3)	.06	–0.1 (–0.5 to 0.1)	.27
Total body water (kg), mean (SD; 95% CI)	38.1 (5.3; 36.5-39.7)	38.2 (5.4; 36.6-39.8)	.54	37.7 (5.6; 36.1-39.5)	36.9 (5.6; 35.2-38.6)	.001	–0.9 (–1.5 to –0.4)	.003
Total body water percentage (%), mean (SD; 95% CI)	47.0 (3.9; 45.8-48.1)	47.1 (5.1; 45.5-48.6)	.07	46.0 (4.0; 44.9-47.3)	47.8 (5.2; 46.3-49.4)	<.001	1.6 (0.2-3.0)	.001
Extracellular water percentage (%), mean (SD; 95% CI)	41.2 (2.1; 40.5-41.8)	40.9 (2.0; 40.3-41.5)	.003	41.4 (2.2; 40.7-42.0)	40.8 (2.0; 40.2-41.4)	.001	–0.2 (–0.6 to 0.1)	.10
BMC[d] (kcal), mean (SD; 95% CI)	1631 (252; 1555-1708)	1621 (251; 1545-1698)	.27	1597 (273; 1515-1679)	1545 (274; 1463-1627)	<.001	–42 (–62 to –22)	<.001

[a]Continuous variables were assessed using unpaired *t* tests (reporting mean difference and 95% CI) for normally distributed data and Mann-Whitney *U* tests (reporting Hodges-Lehmann median difference and 95% CI) for non-normally distributed data to compare the changes between the two groups.

[b]Continuous variables were assessed using paired *t* tests (normally distributed data) or Wilcoxon signed-rank tests (non-normally distributed data) within each group.

[c]ASM: appendicular skeletal muscle mass.

[d]BMC: basal metabolic rate.

XSL•FO
RenderX

Analysis of Primary and Secondary Efficacy Outcomes

As shown in Figure 3, the primary and secondary efficacy outcomes were compared between the control and intervention groups. According to the ITT analysis Figure 3, 60% (27/45) of patients achieved weight loss of ≥5% (27/45 vs 7/44; RR 3.771, 95% CI 1.836 - 7.748; *P*<.001), and 24.4 % (11/45) of patients attained weight loss of ≥10% among patients in the TLC program-delivered intervention group, which was greater than those in the control group (11/45 vs 3/44; RR 3.585, 95% CI 1.072 - 11.988; *P*=.02). A CAP value of ≥10% reduction or normalization was achieved in more patients who received the intervention than in those who received standard care (26/45 vs 14/44; RR 1.816, 95% CI 1.102 - 2.992; *P*=.01). Furthermore, patients in the intervention group were more likely than those in the control group to achieve an ALT reduction of ≥50% or normalization (30/45 vs 16/44; RR 1.833, 95% CI 1.178 - 2.853; *P*=.004), as well as triglyceride normalization (37/45 vs 20/44; RR 1.809, 95% CI 1.273 - 2.570; *P*<.001). Similarly, these results from the PP analysis were in line with those from the ITT analysis (Figure 3).

The effects of the TLC program on weight loss and the CAP value were also analyzed by multivariate regression (Table 4). Among the ITT population or PP population, the intervention had a positive effect on weight loss of ≥5% or ≥10% after adjusting for age, sex, and baseline weight (*P*<.05). Additionally, patients in the intervention group were more likely to achieve a CAP reduction of ≥10% or normalization after adjusting for age, sex, and baseline CAP in both the ITT population (OR 2.853, 95% CI 1.092 - 7.456; *P*=.03) and the PP population (OR 3.319, 95% CI 1.117 - 9.860; *P*=.03).

Figure 3. Analysis of primary and secondary efficacy outcome between control group and intervention group. (A) The percentage of patients with weight loss ≥5% between groups in the intention-to-treat population (27/45 vs 7/44, RR 3.771, 95% CI 1.836-7.748; *P*<.001) and the per-protocol population (26/37 vs 5/36, RR 5.059, 95% CI 2.184-11.719; *P* <.001) ; (B) The percentage of patients with weight loss ≥10% between groups in the intention-to-treat population (11/45 vs 3/44, RR 3.585, 95% CI 1.072-11.988; *P*=.02) and the per-protocol population (10/37 vs 3/36, RR 3.243, 95% CI 1.066-10.360; *P*=.04) ; (C) The percentage of patients with controlled attenuation parameter ≥10% reduction or normalization between groups in the intention-to-treat population (26/45 vs 14/44, RR 1.816, 95% CI 1.102-2.992; *P*=.01) and the per-protocol population (25/37 vs 13/36, RR 1.871, 95% CI: 1.148-3.050; *P*=.007; (D) The percentage of patients with alanine aminotransferase ≥50% reduction or normalization between groups in the intention-to-treat population (30/45 vs 16/44, RR 1.833, 95% CI 1.178-2.853; *P*=.004) and the per-protocol population (25/37 vs 16/36, RR 1.520, 95% CI 1.006-2.391; *P*=.047; (E) The percentage of patients with triglyceride normalization between groups in the intention-to-treat population (37/45 vs 20/44, RR 1.809, 95% CI 1.273-2.570; *P*<.001) and the per-protocol population (32/37 vs 17/36, RR 1.831, 95% CI 1.267-2.646; *P*<.001). ALT: aspartate aminotransferase; CAP: controlled attenuation parameter; ITT: intention-to-treat; PP: per-protocol; TG: triglyceride.

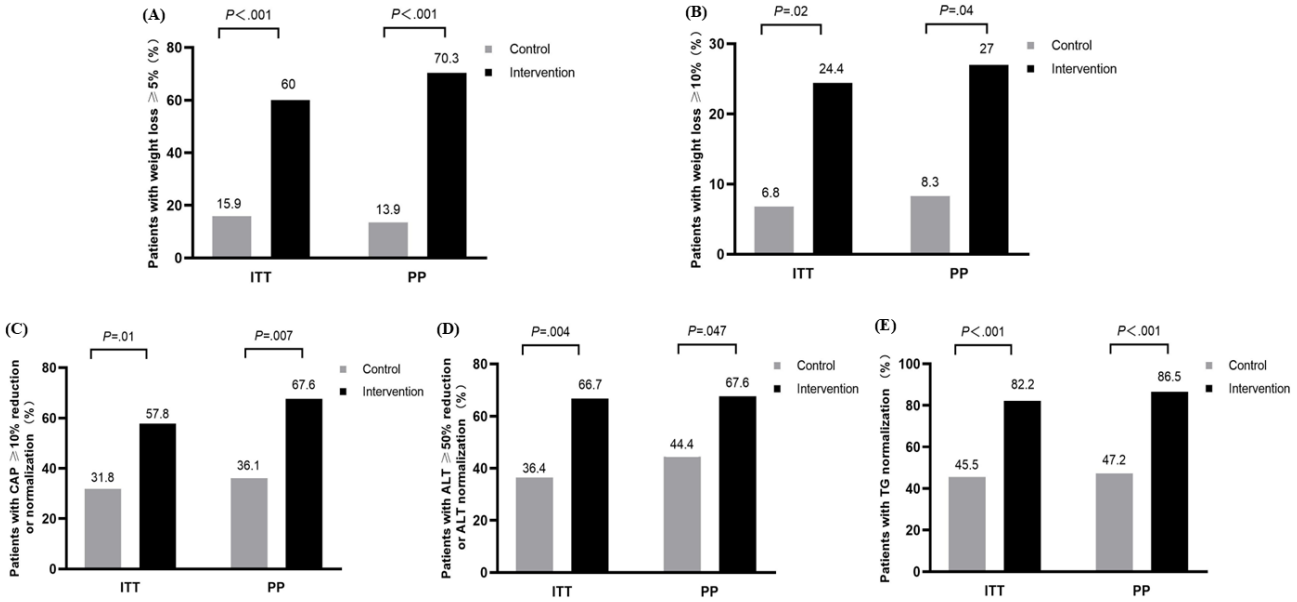

XSL•FO
RenderX

Table . Multivariate logistic analysis model of reduction in weight and controlled attenuation parameter value in intention-to-treat and per-protocol population.

Parameter	Unadjusted	Model 1[a]	Model 2[b]	Unadjusted	Model 1	Model 2
	ITT[c] population (N=89)			PP[d] population (N=73)		
Weight loss≥5%						
Control	reference	reference	reference	reference	reference	reference
Intervention, OR (95% CI)	7.929 (2.905-21.641)	8.537 (2.947-24.730)	8.380 (2.886-24.331)	14.655 (4.509-47.626)	14.534 (4.267-49.499)	15.231 (4.380-52.969)
P value	＜.001	＜.001	＜.001	＜.001	＜.001	＜.001
Weight loss ≥10%						
Control	reference	reference	reference	reference	reference	reference
Intervention, OR[e] (95% CI)	4.422 (1.140 - 17.144)	4.677 (1.160 - 18.855)	4.612 (1.138 - 18.686)	4.074 (1.018 - 16.305)	4.604 (1.093 - 19.386)	4.571 (1.078 - 19.387)
P value	.03	.03	.03	.047	.04	.04
CAP[f]≥10% reduction or normalization						
Control	reference	reference	reference	reference	reference	reference
Intervention, OR (95% CI)	2.932 (1.232-6.981)	2.843 (1.161-6.965)	2.853 (1.092-7.456)	3.686 (1.401-9.700)	3.399 (1.259-9.175)	3.319 (1.117-9.860)
P value	.02	.02	.03	.008	.02	.03

[a]Model 1: Adjusted for age, sex.

[b]Model 2: Adjusted for model 1 and baseline weight.

[c]ITT: intention-to-treat.

[d]PP: per-protocol population.

[e]OR: odds ratio.

[f]Model 2：Adjusted for model 1 and baseline controlled attenuation parameter value; CAP: controlled attenuation parameter.

Safety

During the study period, the intervention was safe in this cohort of patients with MAFLD, with no adverse events reported during the study period. The heart rate monitoring during the video exercises effectively protected participants from the risk of overexertion.

Discussion

Principal Findings

MAFLD is a public health concern and is associated with unhealthy lifestyles. Effective weight control is a crucial measure for reversing MASH and attenuating metabolic dysfunction [11,33]. While digital health tools offer a scalable solution, there is a notable scarcity of interventions specifically designed for the population of individuals with MAFLD in China. Therefore, we conducted a longitudinal study to evaluate an innovative WeChat mini-program-based lifestyle modification in patients with MAFLD. Our study revealed that the TLC program intervention was more effective for weight loss among patients with MAFLD who were overweight or obese than standard care was. Importantly, CAP values decreased significantly more in the intervention group than in the control group, suggesting that the TLC program intervention ameliorated hepatic steatosis. Along with weight reduction, the intervention group presented significant decreases in liver enzymes, HbA_{1c}, and triglyceride levels, indicating improved liver function and metabolic-related indicators. To investigate the changes in body composition, we found that the intervention group had a lower total body fat percentage and visceral fat level than did the control group, with parallel improvement in hepatic steatosis. In terms of liver fibrosis assessment, only the FAST score showed improvement in MASH with significant fibrosis after the intervention, whereas other fibrosis indicators, such as LSM and noninvasive scores (APRI, FIB-4, NFS, Agile 3+) showed no treatment effect.

Based on established evidence, DTx lifestyle interventions lead to significant weight loss in patients with MAFLD [6,30]. According to a meta-analysis of 8 cohort studies on DTx, 33% of patients with MAFLD experienced significant weight loss of ≥5% [13]. In our study, more than half (27/45, 60%) of the patients with MAFLD achieved a weight loss of ≥5% , demonstrating the efficacy of weight management over 6 months. The significant improvement in diet quality scores observed in the intervention group implies a potential key role of dietary modification within the TLC program. With the reduction in body weight, there was a significant improvement in liver chemistry variables, including ALT, AST, and GGT, which was consistent with published data [30,34]. In particular, approximately two-thirds of the patients with TLC intervention experienced an ALT reduction of more than 50% or

XSL•FO
RenderX

normalization, suggesting that this digital intervention improved liver function remarkably.

Importantly, 57.8% (26/45) of the patients in the intervention group presented a CAP value of more than a 10% reduction or normalization, indicating that the severity of hepatic steatosis was highly reduced among individuals after intervention. A similar study investigated a mobile technology–based intervention program for patients with MAFLD and demonstrated that 42.4% (14/33) of participants exhibited a reduction in CAP values, with 21.2% (7/33) achieving a significant decrease of ≥10% in the CAP value [35]. Based on our study, the reduction in body weight and visceral fat led to a marked decline in hepatic fat content and subsequent improvement in liver function.

Among patients with MAFLD, the severity of fibrosis is a major disease modifier and affects hepatic outcomes [36,37]. Gradual weight loss of more than 10% may ameliorate fibrosis in patients with MAFLD [38]. In our study, although the proportion of patients with weight loss greater than 10% reached 24.4% after intervention, the LSM value was not significantly different between the groups. Similarly, no significant difference was found in noninvasive scores, such as the APRI, FIB-4 score, NFS, and Agile 3+ score, except for the FAST score. With respect to noninvasive assessment of liver fibrosis, current evidence from different studies has shown inconsistencies. A German web-based exercise program intervention for patients with MAFLD reported improved APRI, FIB-4 score, and LSM, whereas a US mobile-based program intervention for patients with biopsy-confirmed MAFLD showed no changes in NFS or FIB-4 score [39,40]. The observed discrepancies across studies may stem from variations in intervention protocols, differential durations of implementation, and heterogeneity in patient adherence levels. Based on our data, the overall baseline LSM values of the patients enrolled seem to be remarkably low, with a median value of only 5.1 (IQR 4.4- 6.5) kPa. Furthermore, the duration of the intervention was only 6 months, which is too short for observing changes in fibrosis.

FAST scores reflect MASH with significant fibrosis and are calculated via the CAP, LSM, and AST. In our study, the observed reduction in the FAST score in the intervention group may be associated with the significant decreases in CAP and AST levels, suggesting potential alleviation of MASH to some extent by using the TLC program. Undoubtedly, the definitive evidence of MASH alleviation is contingent upon further study with patients who have undergone liver biopsies.

MAFLD results in impaired metabolic profiles in multiple organ systems, which not only promotes the progression of liver disease but also increases the risk of extrahepatic complications [41,42]. In the current investigation, we assessed key metrics of glucose metabolism, observing substantially greater reductions in fasting insulin, HbA_{1c}, and HOMA-IR levels among intervention group participants, which aligns with established findings in the literature [35,39]. Nevertheless, there was no significant difference in the FPG level between these two groups. It is important to emphasize that our cohort consisted predominantly of nondiabetic individuals, with an overall low baseline glucose (5.3 mmol/L), which limited the potential for a large absolute reduction. Interestingly, some studies concerning DTx-mediated lifestyle modifications in MAFLD management have reported similar results [34,40]. Indeed, insulin resistance is the key pathogenic link between obesity and associated metabolic disorders. Moreover, it is the primary driver in the development of MAFLD and cardiovascular disease. A significant reduction in body weight has favorable effects on increasing insulin sensitivity and improving insulin resistance, leading to a reduction in fasting insulin and glycemic fluctuations, substantially reducing cardiovascular events in patients with MAFLD [43]. A reduction in the HbA_{1c} level, reflecting good glycemic control over the past 2‑3 months, may lead to decreased hepatic fat accumulation, reduced liver fat synthesis, and increased fat breakdown [44].

For the lipid profile, decreased TG or elevated HDL-C levels were found in the intervention group, which was in line with previously published literature [35,45,46]. Favorable modifications in TG or HDL profiles have been correlated with the amelioration of insulin resistance, which not only upregulates lipoprotein lipase expression and accelerates TG hydrolysis but also downregulates hepatic lipase activity, reduces HDL-C degradation, and promotes HDL-C synthesis [41,47]. Overall, TLC digital intervention markedly improved metabolic indices, such as glucose and lipid levels, which are pivotal determinants of cardiovascular morbidity and mortality in patients with MAFLD.

To evaluate the quality of weight loss, we explored parameters involved in body composition. Compared with the control group, the intervention group presented greater reductions in total body fat percentage and visceral fat, as well as the alleviation of hepatic steatosis. Although the intervention group demonstrated significant reductions in total lean body mass and absolute ASM, the ratio of ASM to weight increased within the intervention group, suggesting that weight loss was driven primarily by fat reduction. As is well known, healthy weight loss involves fat loss with muscle gain. Furthermore, no significant difference in MET levels was detected between the groups in our study. To build muscle and improve body composition, future improvements to the TLC program should integrate progressive strength training, such as bodyweight exercises, resistance bands, and dumbbells. Monitoring changes in muscle mass using body fat scales can further support the personalization and optimization of these interventions.

Strengths and Limitations

This study has several strengths. Its innovation lies in being the first tailored digital tool to address the specific self-management challenges of this population in China, leveraging a widely accessible platform. Additionally, this mobile-based program offers customized dietary plans based on each patient's health status. Notably, this smartphone-delivered lifestyle program aids in weight control and liver fat reduction by facilitating healthier behaviors, which could be implemented in home settings. This high accessibility removes common barriers to care, such as transportation and time constraints, while its timely, personalized feedback promotes greater adherence and sustained habit formation. Together, these advantages make it

XSL•FO
RenderX

a scalable and cost-effective strategy for the long-term management of MAFLD.

This study has several limitations. First, all the patients enrolled were from a single tertiary hepatology clinic, and more motivated individuals tended to self-select into the program. This selection bias was inevitable. Second, hepatic steatosis and fibrosis were evaluated by FibroScan or noninvasive tests rather than liver biopsy. On the one hand, FibroScan is a widely used and noninvasive tool to detect hepatic steatosis and screen for fibrosis. On the other hand, it is extremely challenging to obtain liver biopsy specimens in clinical practice. Third, no postintervention follow-up was conducted to assess the maintenance of weight loss. The lack of long-term follow-up data limits the conclusions regarding the sustainability of the WeChat-based intervention effects. Future research will include a follow-up phase where the research team will formulate personalized weight-management plans and encourage the involvement of family or friends for support and supervision to facilitate long-term weight maintenance for participants. Additionally, the sample size herein was limited, and the dropout rate was relatively high (18%). The application of last observation carried forward for handling missing data may have introduced bias into the treatment effect estimate. These factors, collectively, may have contributed to the imprecision in our effect estimates, as evidenced by the wide CIs. Subsequent study should incorporate larger, multicenter cohorts and advanced modeling approaches. Furthermore, the mini-program itself will be further developed to integrate incentives, gamification, and social support. This comprehensive strategy aims to enhance user adherence, boost engagement, and reduce dropout rates.

Conclusions

Our study introduces a WeChat mini-program that delivers a tailored lifestyle intervention designed for patients with MAFLD in China, addressing a gap in scalable management. Our intervention is characterized by its ability to not only reduce weight but also to significantly improve hepatic steatosis and metabolic measures. These findings contribute to the field by establishing a feasible, low-cost model for managing MAFLD in high-burden populations, with strong potential for implementation in primary care and resource-limited settings. Given the imprecision of effect estimates, further larger-scale studies are essential to confirm these preliminary findings, obtain more precise estimates, and evaluate sustained long-term clinical outcomes with long-term follow-up.

Acknowledgments

This study was supported by grants from Noncommunicable Chronic Diseases-National Science and Technology Major Project (2023ZD0508704), the Construction Project of the “Discipline Peak-Climbing Plan” from Xinhua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine (XKPF2024B400). The authors thank all the participants and clinicians who participated in our study. We would like to thank Hangzhou Jianhai Technology Company for providing supportive digital technology.

Artificial intelligence (AI) was not used in any portion of the manuscript writing.

Funding

The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated and analyzed in the current study are available from the corresponding author upon reasonable request.

Authors' Contributions

SC and FJG contributed to the conceptualization of the study. SC, SCC, CHX, YRX, ZJ, and DXY were responsible for the investigation and data curation. CGY and SX guided the methodology. FJG and SX provided resources and secured funding. SC, SX, and FJG were responsible for the program development, project administration, and supervision. SC and FJG contributed to formal analysis and writing-original draft. SC, CGY, SCC, CHX, YRX, ZJ, DXY SX, and FJG carried out validation, writing-review, and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Institutional Review Board approval document.
[PDF File, 2513 KB - jmir_v28i1e76204_app1.pdf]

Checklist 1
CONSORT eHealth checklist (V1.6.1).
[PDF File, 37271 KB - jmir_v28i1e76204_app2.pdf]

XSL•FO
RenderX

References

1. Eslam M, Fan JG, Yu ML, et al. The Asian Pacific association for the study of the liver clinical practice guidelines for the diagnosis and management of metabolic dysfunction-associated fatty liver disease. Hepatol Int 2025 Apr;19(2):261-301. [doi: 10.1007/s12072-024-10774-3] [Medline: 40016576]
2. European Association for the Study of the Liver (EASL), European Association for the Study of Diabetes (EASD), European Association for the Study of Obesity (EASO). EASL-EASD-EASO Clinical Practice Guidelines on the Management of Metabolic Dysfunction-Associated Steatotic Liver Disease (MASLD). Obes Facts 2024;17(4):374-444. [doi: 10.1159/000539371] [Medline: 38852583]
3. Fan JG, Xu XY, Yang RX, et al. Guideline for the Prevention and Treatment of Metabolic Dysfunction-associated Fatty Liver Disease (Version 2024). J Clin Transl Hepatol 2024 Nov 28;12(11):955-974. [doi: 10.14218/JCTH.2024.00311] [Medline: 39544247]
4. Seeberg KA, Borgeraas H, Hofsø D, et al. Gastric bypass versus sleeve gastrectomy in type 2 diabetes: effects on hepatic steatosis and fibrosis: a randomized controlled trial. Ann Intern Med 2022 Jan;175(1):74-83. [doi: 10.7326/M21-1962] [Medline: 34843380]
5. Rinella ME, Neuschwander-Tetri BA, Siddiqui MS, et al. AASLD Practice Guidance on the clinical assessment and management of nonalcoholic fatty liver disease. Hepatology 2023 May 1;77(5):1797-1835. [doi: 10.1097/HEP.0000000000000323] [Medline: 36727674]
6. Pfirrmann D, Huber Y, Schattenberg JM, Simon P. Web-based exercise as an effective complementary treatment for patients with nonalcoholic fatty liver disease: intervention study. J Med Internet Res 2019 Jan 2;21(1):e11250. [doi: 10.2196/11250] [Medline: 30602434]
7. Zhou R, Zhang B, Zhang W, et al. Self-management behaviours in adults with non-alcoholic fatty liver disease: a cross-sectional survey from China. BMJ Open 2024 Feb 22;14(2):e078333. [doi: 10.1136/bmjopen-2023-078333] [Medline: 38388505]
8. Tincopa MA, Patel N, Shahab A, Asefa H, Lok AS. Implementation of a randomized mobile-technology lifestyle program in individuals with nonalcoholic fatty liver disease. Sci Rep 2024 Mar 28;14(1):7452. [doi: 10.1038/s41598-024-57722-7] [Medline: 38548875]
9. Zhou R, Gu Y, Zhang B, et al. Digital therapeutics: emerging new therapy for nonalcoholic fatty liver disease. Clin Transl Gastroenterol 2023 Apr 1;14(4):e00575. [doi: 10.14309/ctg.0000000000000575] [Medline: 36854062]
10. Kwon OY, Lee MK, Lee HW, Kim H, Lee JS, Jang Y. Mobile app-based lifestyle coaching intervention for patients with nonalcoholic fatty liver disease: randomized controlled trial. J Med Internet Res 2024 Feb 15;26:e49839. [doi: 10.2196/49839] [Medline: 38358794]
11. Sato M, Akamatsu M, Shima T, et al. Impact of a novel digital therapeutics system on nonalcoholic steatohepatitis: the NASH App clinical trial. Am J Gastroenterol 2023 Aug 1;118(8):1365-1372. [doi: 10.14309/ajg.0000000000002143] [Medline: 36656974]
12. Björnsdottir S, Ulfsdottir H, Gudmundsson EF, et al. User engagement, acceptability, and clinical markers in a digital health program for nonalcoholic fatty liver disease: prospective, single-arm feasibility study. JMIR Cardio 2024 Feb 15;8:e52576. [doi: 10.2196/52576] [Medline: 38152892]
13. Albhaisi S, Tondt J, Cyrus J, Chinchilli VM, Conroy DE, Stine JG. Digital therapeutics lead to clinically significant body weight loss in patients with metabolic dysfunction-associated steatotic liver disease: a systematic review and meta-analysis. Hepatol Commun 2024 Aug 1;8(8):e0499. [doi: 10.1097/HC9.0000000000000499] [Medline: 39082956]
14. Chen D, Shao J, Zhang H, et al. Development of an individualized WeChat mini program-based intervention to increase adherence to dietary recommendations applying the behaviour change wheel among individuals with metabolic syndrome. Ann Med 2023;55(2):2267587. [doi: 10.1080/07853890.2023.2267587] [Medline: 37898907]
15. Duan Y, Li X, Guo L, Liang W, Shang B, Lippke S. A WeChat mini program-based intervention for physical activity, fruit and vegetable consumption among Chinese cardiovascular patients in home-based rehabilitation: a study protocol. Front Public Health 2022;10:739100. [doi: 10.3389/fpubh.2022.739100] [Medline: 35392478]
16. Duan Y, Liang W, Guo L, et al. Effectiveness of a wechat mini program-based intervention on promoting multiple health behavior changes among Chinese patients with cardiovascular diseases in home-based rehabilitation: randomized controlled trial. J Med Internet Res 2025 Jun 3;27:e66249. [doi: 10.2196/66249] [Medline: 40460318]
17. Pan Y, Tang J, Lu B, et al. Effects of cognitive behavioral therapy for diet on postprandial glucose and pregnancy outcomes in gestational diabetes mellitus: multicenter randomized controlled trial. J Med Internet Res 2025 Jul 29;27:e71075. [doi: 10.2196/71075] [Medline: 40729762]
18. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
19. National Workshop on Fatty Liver and Alcoholic Liver Disease, Chinese Society of Hepatology, Chinese Medical Association, Fatty Liver Expert Committee, Chinese Medical Doctor Association. Guidelines of prevention and treatment for nonalcoholic fatty liver disease: a 2018 update. Zhonghua Gan Zang Bing Za Zhi 2018 Mar 20;26(3):195-203. [doi: 10.3760/cma.j.issn.1007-3418.2018.03.008] [Medline: 29804393]

XSL•FO
RenderX

20. Karlas T, Petroff D, Sasso M, et al. Individual patient data meta-analysis of controlled attenuation parameter (CAP) technology for assessing steatosis. J Hepatol 2017 May;66(5):1022-1030. [doi: 10.1016/j.jhep.2016.12.022] [Medline: 28039099]
21. Chinese Society of Health Management, Chinese Nutrition Society, Reproductive Medicine Branch of China International Exchange and Promotion Association for Medicine and Healthcare, China Health Promotion Foundation, Zhejiang Provincial Clinical Nutrition Center. Expert consensus & standard on weight management for overweight or obese people. Chin J Health Manage 2018;12(3):200-207. [doi: 10.3760/cma.j.issn.1674-0815.2018.03.003]
22. Craig CL, Marshall AL, Sjöström M, et al. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 2003 Aug;35(8):1381-1395. [doi: 10.1249/01.MSS.0000078924.61453.FB] [Medline: 12900694]
23. Sun C, Goh GBB, Chow WC, et al. Prevalence and risk factors for impaired renal function among Asian patients with nonalcoholic fatty liver disease. Hepatobiliary Pancreat Dis Int 2024 Jun;23(3):241-248. [doi: 10.1016/j.hbpd.2023.08.004] [Medline: 37620227]
24. Bedogni G, Bellentani S, Miglioli L, et al. The Fatty Liver Index: a simple and accurate predictor of hepatic steatosis in the general population. BMC Gastroenterol 2006 Nov 2;6:33. [doi: 10.1186/1471-230X-6-33] [Medline: 17081293]
25. Newsome PN, Sasso M, Deeks JJ, et al. FibroScan-AST (FAST) score for the non-invasive identification of patients with non-alcoholic steatohepatitis with significant activity and fibrosis: a prospective derivation and global validation study. Lancet Gastroenterol Hepatol 2020 Apr;5(4):362-373. [doi: 10.1016/S2468-1253(19)30383-8] [Medline: 32027858]
26. Kruger FC, Daniels CR, Kidd M, et al. APRI: a simple bedside marker for advanced fibrosis that can avoid liver biopsy in patients with NAFLD/NASH. S Afr Med J 2011 Jun 27;101(7):477-480. [doi: 10.10520/EJC67626] [Medline: 21920102]
27. Angulo P, Hui JM, Marchesini G, et al. The NAFLD fibrosis score: a noninvasive system that identifies liver fibrosis in patients with NAFLD. Hepatology 2007 Apr;45(4):846-854. [doi: 10.1002/hep.21496] [Medline: 17393509]
28. Sterling RK, Lissen E, Clumeck N, et al. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. Hepatology 2006 Jun;43(6):1317-1325. [doi: 10.1002/hep.21178] [Medline: 16729309]
29. Sanyal AJ, Foucquier J, Younossi ZM, et al. Enhanced diagnosis of advanced fibrosis and cirrhosis in individuals with NAFLD using FibroScan-based Agile scores. J Hepatol 2023 Feb;78(2):247-259. [doi: 10.1016/j.jhep.2022.10.034] [Medline: 36375686]
30. Lim SL, Johal J, Ong KW, et al. Lifestyle intervention enabled by mobile technology on weight loss in patients with nonalcoholic fatty liver disease: randomized controlled trial. JMIR Mhealth Uhealth 2020 Apr 13;8(4):e14802. [doi: 10.2196/14802] [Medline: 32281943]
31. Björnsdottir S, Ulfsdottir H, Gudmundsson EF, et al. Long-term feasibility and outcomes of a digital health program to improve liver fat and cardiometabolic markers in individuals with nonalcoholic fatty liver disease: prospective single-arm feasibility study. JMIR Cardio 2025 Sep 12;9:e72074. [doi: 10.2196/72074] [Medline: 40939135]
32. Harrison SA, Ruane PJ, Freilich B, et al. A randomized, double-blind, placebo-controlled phase IIa trial of efruxifermin for patients with compensated NASH cirrhosis. JHEP Rep 2023 Jan;5(1):100563. [doi: 10.1016/j.jhepr.2022.100563] [Medline: 36644237]
33. Armandi A, Bugianesi E. Dietary and pharmacological treatment in patients with metabolic-dysfunction associated steatotic liver disease. Eur J Intern Med 2024 Apr;122:20-27. [doi: 10.1016/j.ejim.2024.01.005] [Medline: 38262842]
34. Mazzotti A, Caletti MT, Brodosi L, et al. An internet-based approach for lifestyle changes in patients with NAFLD: two-year effects on weight loss and surrogate markers. J Hepatol 2018 Nov;69(5):1155-1163. [doi: 10.1016/j.jhep.2018.07.013] [Medline: 30290973]
35. Tincopa MA, Lyden A, Wong J, Jackson EA, Richardson C, Lok AS. Impact of a pilot structured mobile technology based lifestyle intervention for patients with nonalcoholic fatty liver disease. Dig Dis Sci 2022 Feb;67(2):481-491. [doi: 10.1007/s10620-021-06922-6] [Medline: 33939147]
36. Lin H, Lee HW, Yip TCF, et al. Vibration-controlled transient elastography scores to predict liver-related Events in steatotic liver disease. JAMA 2024 Apr 16;331(15):1287-1297. [doi: 10.1001/jama.2024.1447] [Medline: 38512249]
37. Lee J, Vali Y, Boursier J, et al. Prognostic accuracy of FIB-4, NAFLD fibrosis score and APRI for NAFLD-related events: a systematic review. Liver Int 2021 Feb;41(2):261-270. [doi: 10.1111/liv.14669] [Medline: 32946642]
38. Romero-Gómez M, Zelber-Sagi S, Trenell M. Treatment of NAFLD with diet, physical activity and exercise. J Hepatol 2017 Oct;67(4):829-846. [doi: 10.1016/j.jhep.2017.05.016] [Medline: 28545937]
39. Huber Y, Pfirrmann D, Gebhardt I, et al. Improvement of non-invasive markers of NAFLD from an individualised, web-based exercise program. Aliment Pharmacol Ther 2019 Oct;50(8):930-939. [doi: 10.1111/apt.15427] [Medline: 31342533]
40. Stine JG, Rivas G, Hummer B, et al. Mobile health lifestyle intervention program leads to clinically significant loss of body weight in patients with NASH. Hepatol Commun 2023 Apr 1;7(4):e0052. [doi: 10.1097/HC9.0000000000000052] [Medline: 36930864]
41. Chew NWS, Mehta A, Goh RSJ, et al. Cardiovascular-liver-metabolic health: recommendations in screening, diagnosis, and management of metabolic dysfunction-associated steatotic liver disease in Cardiovascular disease via modified Delphi approach. Circulation 2025 Jan 7;151(1):98-119. [doi: 10.1161/CIRCULATIONAHA.124.070535] [Medline: 39723980]
42. Chan WK, Wong VWS, Adams LA, Nguyen MH. MAFLD in adults: non-invasive tests for diagnosis and monitoring of MAFLD. Hepatol Int 2024 Oct;18(Suppl 2):909-921. [doi: 10.1007/s12072-024-10661-x] [Medline: 38913148]

XSL•FO
RenderX

43. Li M, Cui M, Li G, et al. The Pathophysiological associations between obesity, NAFLD, and atherosclerotic cardiovascular diseases. Horm Metab Res 2024 Oct;56(10):683-696. [doi: 10.1055/a-2266-1503] [Medline: 38471571]
44. He S, Lu S, Yu C, et al. The newly proposed plasma-glycosylated hemoglobin A1c/High-Density lipoprotein cholesterol ratio serves as a simple and practical indicator for screening metabolic associated fatty liver disease: an observational study based on a physical examination population. BMC Gastroenterol 2024 Aug 19;24(1):274. [doi: 10.1186/s12876-024-03362-0] [Medline: 39160462]
45. Axley P, Kodali S, Kuo YF, et al. Text messaging approach improves weight loss in patients with nonalcoholic fatty liver disease: a randomized study. Liver Int 2018 May;38(5):924-931. [doi: 10.1111/liv.13622] [Medline: 29117472]
46. Motz V, Faust A, Dahmus J, Stern B, Soriano C, Stine JG. Utilization of a directly supervised telehealth-based exercise training program in patients with nonalcoholic steatohepatitis: feasibility study. JMIR Form Res 2021 Aug 17;5(8):e30239. [doi: 10.2196/30239] [Medline: 34402795]
47. Colantoni A, Bucci T, Cocomello N, et al. Lipid-based insulin-resistance markers predict cardiovascular events in metabolic dysfunction associated steatotic liver disease. Cardiovasc Diabetol 2024 May 20;23(1):175. [doi: 10.1186/s12933-024-02263-6] [Medline: 38769519]

Abbreviations

ALT: alanine aminotransferase
APRI: aspartate aminotransferase-to-platelet ratio index
ASM: appendicular skeletal muscle mass
AST: aspartate aminotransferase
CAP: controlled attenuation parameter
CONSORT: Consolidated Standards of Reporting Trials
DTx: digital therapeutics
FAST: FibroScan-aspartate aminotransferase
FIB-4: fibrosis-4
FPG: fasting plasma glucose
GGT: γ-glutamyl transpeptidase
HbA_{1c}: hemoglobin A_{1c}
HDL-C: high-density lipoprotein-cholesterol
HOMA-IR: homeostasis model assessment-insulin resistance
IPAQ-SF: International Physical Activity Questionnaire-Short Form
ITT: intention-to-treat
LSM: liver stiffness measurement
MAFLD: metabolic dysfunction–associated fatty liver disease
MASH: metabolic dysfunction-associated steatohepatitis
MET: metabolic equivalent of task
OR: odds ratio
PP: per-protocol
RR: relative risk
TLC: therapeutic lifestyle changes

Edited by S Brini; submitted 18.Apr.2025; peer-reviewed by J Zhang, N Maye; revised version received 26.Nov.2025; accepted 02.Dec.2025; published 27.Jan.2026.

Please cite as:
Sun C, Chen G, Shi C, Cao H, Yang R, Zeng J, Duan X, Sun X, Fan JG
Therapeutic Effects of a WeChat Mini-Program on Metabolic Dysfunction–Associated Fatty Liver Disease: Randomized Controlled Trial
J Med Internet Res 2026;28:e76204
URL: https://www.jmir.org/2026/1/e76204
doi:10.2196/76204

© Chao Sun, Guangyu Chen, Cuicui Shi, Haixia Cao, Ruixu Yang, Jing Zeng, Xiaoyan Duan, Xin Sun, Jian-Gao Fan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of

XSL•FO
RenderX

Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Cognitive-Behavioral Therapy−Based Massed Brief Psychoeducational Group via Videoconference for Social Anxiety: Randomized Controlled Trial

Lele Feng[1], MA; Wei Liu[1], MA; Liechuan Cui[1], MA; Deborah Dobson[2], PhD; Xinfeng Tang[1], PhD

[1]Department of Psychology, Renmin University of China, Huixian Building, 9th, 59 Zhongguancun St., Haidian District, Beijing, China

[2]Department of Psychology, University of Calgary, Calgary, AB, Canada

Corresponding Author:
Xinfeng Tang, PhD
Department of Psychology, Renmin University of China, Huixian Building, 9th, 59 Zhongguancun St., Haidian District, Beijing, China

Abstract

Background: Group cognitive-behavioral therapy (CBT), delivered through weekly videoconference sessions, has been shown to effectively reduce social anxiety. However, no studies have evaluated CBT delivered via videoconference in a 2-day massed brief psychoeducational group format.

Objective: This randomized controlled trial aimed to evaluate the efficacy of a videoconferencing CBT−based massed brief psychoeducational group among Chinese university students with social anxiety.

Methods: University students with social anxiety were recruited online and randomly assigned to an intervention group or a waitlist control group. Participants in the intervention group attended a 2-day workshop via videoconference. Assessments were conducted at baseline (T1), posttest (T2), 1-month follow-up (T3), and 3-month follow-up (T4), using the Social Phobia Inventory, Brief Fear of Negative Evaluation Scale, Depression Anxiety Stress Scales–Short Form, Social Anxiety Knowledge Test, Social Anxiety Stigma Inventory, and Attitudes Toward Seeking Professional Psychological Help Scale–Short Form.

Results: The intervention group showed significant reductions in Social Phobia Inventory scores (β=−4.00, 95% bootstrap CI −6.55 to −1.22; $d_{T2\text{-}4}$=−0.97 to −0.81) and Brief Fear of Negative Evaluation Scale scores (β=−1.37, 95% bootstrap CI −2.64 to −0.08; d_{T3}=−0.56), as well as significant increases in Social Anxiety Knowledge Test scores (β=.62, 95% bootstrap CI 0.05-1.17; $d_{T2\text{-}4}$=0.86-1.53). No significant changes were observed in Depression Anxiety Stress Scales–Short Form, Social Anxiety Stigma Inventory, or Attitudes Toward Seeking Professional Psychological Help Scale–Short Form scores.

Conclusions: The findings indicate that videoconferencing CBT−based massed brief psychoeducational group was effective in reducing social anxiety among university students. Future research with larger and more diverse samples is recommended to validate the efficacy and assess the scalability of this intervention format.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2400093444; www.chictr.org.cn/showprojEN.html?proj=235703

(J Med Internet Res 2026;28:e79825) doi:10.2196/79825

KEYWORDS

cognitive behavioral therapy; massed intervention; psychoeducational group; randomized controlled trial; social anxiety; workshop intervention; low-intensity intervention

Introduction

Social anxiety refers to the intense fear, worry, or avoidance of social interactions and situations in which one may be evaluated by others [1]. When social anxiety reaches a severity that impairs daily functioning, it is diagnosed as social anxiety disorder or social phobia [2]. A meta-analysis reported that the prevalence of social anxiety symptoms among Chinese adolescents and young adults aged 15 - 25 years is 29.8% [3]. Social anxiety disorder can lead to functional impairments across multiple aspects, including work, academics, social functioning, and cognitive processes [4]. Even subclinical symptoms, which do not meet diagnostic criteria, can negatively impact various aspects of life [5] and may progress into a chronic and debilitating condition if left unaddressed [6]. For example, social anxiety may hinder college students' learning, decrease their well-being [7], and even elevate their risk of suicidal thoughts or behaviors [8]. Early intervention is, therefore, critical to alleviating symptom severity and the likelihood of developing social anxiety disorder.

Cognitive-behavioral therapy (CBT) is widely recognized as the gold standard for treating social anxiety [9]. Meta-analytic

XSL•FO
RenderX

findings indicate that compared to pharmacological and alternative therapeutic approaches, CBT demonstrates superior efficacy, greater safety, and lower relapse rates in both the short and long term, making it the most highly recommended treatment [10,11]. CBT is typically delivered in individual or group formats, both of which have shown positive effects on social anxiety. However, group CBT has its unique advantages by allowing participants to engage with unfamiliar peers in socially relevant contexts while benefiting from mutual support. It also enables more individuals to receive treatment within a given timeframe [12]. Therefore, group interventions are also frequently used in both research and clinical practice.

In terms of therapist-guided delivery medium, psychological interventions can be categorized into face-to-face CBT and remote CBT. Traditional group CBT typically involves therapists working face-to-face with participants; however, limited mental health resources often prevent individuals from accessing CBT or other mental health services in their daily lives [13]. Moreover, traditional group CBT may entail several temporal and locational constraints, limiting the delivery of timely clinical interventions [14]. In this context, internet-based remote CBT offers a viable alternative [15]. Among remote modalities, *videoconferencing* is commonly used, as it enables synchronous, real-time communication through audio and video and allows therapists and participants to interact directly from different locations. Compared to other digital formats, such as message-based or app-based interventions, videoconferencing CBT maintains therapist visibility, facilitating greater interaction and engagement between participants and therapists [16]. Studies have shown that videoconferencing CBT yields significant therapeutic effects for social anxiety and social anxiety disorder, with symptom improvements maintained for up to 3 months postintervention [17,18] and effect sizes ranging from medium to large [19]. Moreover, compared to traditional CBT, videoconferencing CBT may alleviate the resistance to therapy triggered by face-to-face interaction in individuals with social anxiety [20].

Low-intensity interventions have become a common approach in the delivery of psychological interventions. The National Institute for Health and Care Excellence recommends that group CBT for social anxiety be delivered in weekly sessions, for a total of 8 - 12 sessions [21]. This implementation of full protocol requires substantial time and financial resources; as such, low-intensity CBT has emerged to meet the growing demand for mental health services while ensuring treatment effectiveness [22]. Within health care systems, low-intensity interventions are commonly employed in primary care settings for adults experiencing symptoms of depression or specific phobias [23]. A meta-analysis revealed that low-intensity CBT produced large effect sizes in the treatment of anxiety disorders (d=1.06) [24].

A common form of low-intensity intervention involves employing only select core therapeutic techniques rather than the full treatment protocol; this approach is often referred to as a *brief* intervention. For example, in social anxiety interventions, Clark's full protocol typically includes components, such as attention training, video feedback, behavioral experiments, and discrimination training [25]. However, Heimberg's intervention full protocol incorporates multiple sessions of cognitive restructuring and graduated exposure. Brief interventions for social anxiety based on these 2 protocols have also demonstrated promising outcomes [26]. A group CBT for children with social anxiety, delivered in three weekly 3-hour sessions, included psychoeducation, cognitive strategies, and behavioral exposure. Significant reductions in participants' social anxiety scores were observed at both posttest and 3-week follow-up [27]. A recent 7-day internet−based CBT program for social anxiety disorder also demonstrated substantial outcomes (Hedges gs=1.26 - 1.9). This program consisted of 6 online lessons accompanied by practice tasks. Participants were required to complete lessons and corresponding exercises, which included exposure tasks, cognitive challenges, and communication skills practice [28]. More recently, a 1-day CBT−based workshop consisting of cognitive restructuring and assertiveness training for secondary vocational students demonstrated moderate effect sizes in reducing social anxiety symptoms [29].

Another common type of low-intensity intervention is the psychoeducational group. According to Gladding's classification, group interventions are generally divided into *psychoeducational* groups and *counseling* groups [30]. Compared to counseling groups, psychoeducational groups are lower in intensity and emphasize using educational methods to acquire information and develop related meaning and skills [31]. These groups integrate both knowledge acquisition and skill development, efficiently delivering information about psychological disorders while also providing opportunities for practice and experiential learning. Accordingly, 1 major aim of psychoeducational groups is to enhance participants' mental health literacy, which is referred to as the ability to recognize a disorder, understand it, reduce stigma, and seek appropriate psychological help [32]. In addition, the skills training component can help alleviate psychological distress. Evidence indicates that psychoeducational interventions are effective in alleviating anxiety symptoms, including social anxiety [33,34]. A further advantage of psychoeducational groups is their ability to accommodate a larger number of participants. Counseling groups for social anxiety typically have limited membership, with recommended group sizes ranging from 6 to 12 participants [26,35]. In contrast, psychoeducational groups are often larger in scale, commonly ranging from 15 to 40 participants [30].

In terms of intervention frequency, the conventional model involves *spaced* interventions (eg, 1 session per week). Alternatively, *massed* intervention compresses the treatment timeline by delivering multiple sessions over a relatively short timeframe. Meta-analytic evidence indicates that for anxiety disorders in young adults, massed CBT can achieve intervention outcomes comparable to those of full-protocol CBT [36]. For example, Deacon and Abramowitz [37] conducted a 2-day CBT for panic disorder, condensing a 12-session protocol into 2 sessions of 6 and 3 hours, respectively. The intervention included psychoeducation, therapist-assisted exposure, fear hierarchy construction, and in vivo exposure. Following the treatment, all 10 participants showed significant reductions on the Panic Disorder Severity Scale, with large effect sizes (d_s=0.96 - 3.29). To date, spaced brief CBT has been shown to effectively reduce social anxiety and related negative beliefs

XSL•FO
RenderX

[38,39]; however, no empirical evidence currently exists on the effects of massed brief CBT for social anxiety.

The vast majority of existing remote interventions for social anxiety employ spaced, full-protocol treatment programs, typically consisting of 8 or 12 weekly sessions. These interventions are generally conducted in groups of 5 - 12 participants [40-42]. While this model offers several advantages, it also presents challenges. For example, participants may find it difficult to sustain long-term engagement remotely, which can result in them dropping out. A meta-analysis reported that the dropout rate for group CBT interventions targeting anxiety disorders is 24.6% [43]. To alleviate social anxiety more rapidly, efficiently, and conveniently, we developed a low-intensity intervention format—the massed brief psychoeducational group (MBPG)—delivered via videoconference. This study aimed to evaluate the effects of this delivery model on university students with social anxiety through a randomized controlled trial. We hypothesized that compared to the waitlist control group, participants in the intervention group would show significant reductions in social anxiety, fear of negative evaluation, and depressive symptoms following treatment. Additionally, we expected improvements in social anxiety literacy, including increased knowledge of social anxiety, reduced social anxiety stigma, and more positive attitudes toward seeking professional psychological help.

Methods

Participants

According to a meta-analysis by Mayo-Wilson et al [10], interventions with shortened sessions based on the Clark and Wells model for social anxiety have demonstrated large effect sizes. Based on these findings, the expected effect size for this study was conservatively set at 0.80. A priori power analysis using the G*Power 3.1 by Faul et al [44] indicated that a total sample size of 52 would be required. Accounting for an anticipated dropout rate of 20%, the target sample size was set at 65. Accordingly, a minimum of 65 participants were targeted.

Participants were recruited online and were required to meet the following eligibility criteria: (1) current enrollment in a university program, with no restrictions on degree level; (2) availability to attend a 2-day remote workshop and willingness to be randomly assigned; (3) no suicidal ideation and no formal diagnosis of any psychiatric disorder; (4) no psychotropic medication use within the past year and no history of psychological treatment or counseling; and (5) no specific cutoff on social anxiety scores was required for inclusion. As a psychoeducational group, our primary goal was to reach individuals who experienced social anxiety−related difficulties and were motivated to make changes, regardless of their symptom severity. Therefore, although a cutoff value of the Social Phobia Inventory (SPIN) is commonly used to indicate clinically significant symptoms, we did not adopt this cutoff during recruitment. Individuals were eligible for this workshop as long as they perceived themselves affected by social anxiety. Despite the absence of a formal cutoff, participants' baseline levels of social anxiety were relatively high. Among those enrolled, 62 (90%) of the 69 participants scored above the clinical threshold of 19 on the SPIN, with a mean score of 37.91 (SD=13.67).

Participants were randomly assigned to either the intervention group (n=39) or the waitlist control group (n=39). Prior to the start of the intervention, 7 participants from the intervention group and 2 from the control group withdrew. A total of 69 participants ultimately took part in the study (n=32 in the intervention group; n=37 in the control group).

Ethical Considerations

This study received approval from the Ethical Review Committee of the Department of Psychology at Renmin University of China (IRB-24 - 041). All participants provided written informed consent before participating. Individuals who completed all 4 assessments received both monetary compensation of 40 RMB (US $ 5.70) in total (10 RMB [US $1.42] per assessment) and a commemorative gift. To ensure the protection of participants' privacy, the raw data were accessible only to the research team and numerical IDs were assigned in lieu of personally identifiable information during data handling. All reported results were fully anonymized, and no identifiable personal features are presented in the manuscript.

Procedure

Simple randomization was employed to assign participants to groups. Random sequences were generated using Microsoft Excel, with half of the participants allocated to the intervention group and the other half to the waitlist control group.

Before the intervention, an online questionnaire was administered to all participants to collect baseline data (T1). The posttest (T2) was conducted 1 week after the intervention, with follow-up assessments at 1 month (T3) and 3 months (T4) postintervention. Participants in the waitlist control group received the same 2-day online workshop intervention as those in the intervention group after all data collection was completed.

Intervention

This study implemented a CBT-based MBPG intervention, delivered via videoconferencing. The intervention was delivered synchronously via Tencent Meeting, an online platform similar to Zoom. All needed materials were presented in the form of slides shared on screen during the online sessions. The intervention program was named *Joymaster Workshop* to engage participants' interest, reduce associated stigma, and symbolize mastery of social enjoyment. For CBT-based counseling groups targeting social anxiety, a recommended group size is approximately 6 participants [45], whereas psychoeducational groups are typically larger [30]. To avoid reductions in interaction, experiential engagement, and opportunities for practice associated with overly large groups, we set the size of the psychoeducational group at 15 - 20 participants. Therefore, we divided the 32 participants in the intervention condition into 2 workshops. In allocating participants, we considered their time availability and aimed to maintain balance across the 2 groups in terms of size and gender distribution to minimize potential biases arising from group composition. The 2 workshops ultimately included 17 and 15 participants, respectively. Both workshops were delivered by the same group

XSL•FO
RenderX

leaders on 2 consecutive weekends, with identical content to ensure treatment fidelity. The timing of the post-intervention and follow-up assessments was adjusted according to the specific workshop each participant attended.

The term *massed* refers to the delivery format, which consisted of 2 consecutive days, a total duration of 12 hours. As a *brief* intervention, the program selectively incorporated effective techniques from full-protocol CBT, focusing on cognitive restructuring and behavioral experiments. Contemporary CBT interventions for social anxiety typically follow either the Heimberg protocol or the Clark protocol [21]. Given the massed format and the aim of achieving effective outcomes within a short timeframe, this study's program prioritized cognitive change over habituation through exposure, as the latter generally requires more extended treatment. Therefore, we adapted the cognitive restructuring technique from Hope and Heimberg's model, which involves challenging automatic thoughts using supporting and opposing evidence [45].

Behavioral experiments were based on the approach by Leigh and Clark [46] as well as Hofmann and Otto [47], in which participants design and engage in social tasks to test and challenge their automatic thoughts. Behavioral experiments also enhanced the interactivity of the massed intervention and promoted participant engagement. A summary of the main intervention modules is presented in Table 1. Among these, the out-of-session behavioral experiments required participants to complete a selected task during a 2-hour lunch break, based on plans developed at the end of the morning session. Examples included asking a passerby to take a photo of them with a trash can or initiating a brief conversation with someone new in the campus canteen. All participants completed a behavioral experiment.

This intervention was conducted as a *psychoeducational* group rather than a counseling group. First, we employed online slide presentation as visual aids, combining didactic instruction for efficient knowledge delivery with opportunities for experiential practice of therapeutic techniques. Second, the intervention accommodated a larger number of participants than is common in counseling groups, with each workshop hosting 15 - 20 individuals.

Table . Intervention modules.

Module	Techniques	Contents
Day 1—Morning	Psychoeducation	1. Introduce the concept of social anxiety: definition, prevalence, diagnostic criteria, typical age of onset, and negative impacts 2. Provide an overview of cognitive models and contributing factors 3. Introduce treatment approaches for social anxiety
Day 1—Afternoon	Cognitive restructuring (Part 1)	1. Practice the first 2 steps of cognitive restructuring: • Identify automatic thoughts • Identify cognitive biases
Day 2—Morning	Behavioral experiments	1. Explain the purpose and steps of behavioral experiments 2. Conduct in-session behavioral experiments 3. Design out-of-session behavioral experiments for lunch break
Day 2—Afternoon	Cognitive restructuring (Part 2)	1. Reflect on behavioral experiments conducted during lunch break 2. Practice the final 2 steps of cognitive restructuring • Challenge automatic thoughts • Develop realistic, rational thoughts 3. Develop an action plan and conclude the session

Treatment Fidelity

The intervention was delivered by a leader and an assistant. The leader was a PhD-level counselor intensively trained in CBT, and the assistant was a master's student in psychology. Psychoeducational content was presented primarily through PowerPoint slides, with all procedures structured around the slide content. This ensured a highly standardized delivery and strong adherence to the intervention protocol.

Measurements

Primary Outcome—Social Anxiety

The Chinese version of the 17-item SPIN, developed by Connor et al [48] and revised by Xiao et al [49], was used to assess the severity of social anxiety. The scale comprises 3 subscales: fear (6 items; eg, fear of embarrassment or rejection), avoidance (7 items; eg, avoiding social situations because of fear), and physiological symptoms (4 items; eg, blushing or sweating in

XSL•FO
RenderX

social settings)—covering the core clinical manifestations of social anxiety. It uses a 5-point Likert rating scale from 0 (*not at all*) to 4 (*extremely*), with higher scores indicating greater levels of social anxiety. In this study, the SPIN had a Cronbach α of 0.94 at baseline.

Secondary Outcomes

Fear of Negative Evaluation

The Chinese version of the 12-item Brief Fear of Negative Evaluation Scale (BFNES), developed by Leary [50] and revised by Chen [51], was used to assess individuals' concerns about being negatively evaluated by others. It uses a 5-point Likert scale from 1 (*not at all characteristic of me*) to 5 (*extremely characteristic of me*), with higher scores indicating greater fear of negative evaluation. In this study, the BFNES had a Cronbach α of 0.95 at baseline.

Depression, Anxiety, and Stress

The Chinese version of the 21-item Depression Anxiety Stress Scales–Short Form (DASS-21), developed by Lovibond and Lovibond [52] and revised by Gong et al [53], was used to assess the severity of depressive symptoms, anxiety or autonomic arousal, and stress. It uses a 4-point Likert scale from 0 (*did not apply to me at all*) to 3 (*applied to me very much or most of the time*), with higher scores indicating greater symptom severity. In this study, the DASS-21 had a Cronbach α of 0.93 at baseline.

Social Anxiety Knowledge

The self-developed Social Anxiety Knowledge Test (SAKT) was used to assess participants' knowledge of the definition, symptoms, treatment, and related aspects of social anxiety. Based on a literature review, the content was categorized into 8 domains: basic knowledge, symptoms, prevalence, age of onset, gender differences, cultural differences, risk factors, and treatment. An initial pool of 23 single-choice items (with 4 answer options each) was rated and reviewed by 7 experts for content validity, and semistructured cognitive interviews were conducted with 5 nonspecialists. After revisions based on the expert ratings and cognitive interview results, 21 items were retained. Responses were scored as correct (1 point) or incorrect (0 points), with higher total scores (maximum=21) indicating better social anxiety literacy. The SAKT has been adopted in several interventions and shown good validity to test social anxiety−related knowledge [29,54].

Stigmatizing Attitudes Toward Social Anxiety

The self-developed Social Anxiety Stigma Inventory (SASI) was used to assess individuals' stigmatizing attitudes toward social anxiety. The 10-item questionnaire uses a 5-point Likert scale from 0 (*strongly disagree*) to 4 (*strongly agree*), with higher scores indicating stronger stigma. The scale demonstrated good structural validity (χ^{232}= 80.4, root mean square error of approximation=0.061, comparative fit index=0.933, Tucker-Lewis index=0.906, root mean square residual=0.048) and an internal consistency reliability of 0.75. In this study, the SASI had a Cronbach α of 0.86 at baseline. The SASI has been adopted in several interventions and shown good reliability and validity to assess the stigma toward social anxiety [29,54].

Attitudes Toward Seeking Professional Psychological Help

The Chinese version of the 10-item Attitudes Toward Seeking Professional Psychological Help Scale–Short Form (ATSPPH-SF), developed by Fischer and Farina [55] and revised by Fang et al [56], was used to assess individuals' attitudes toward psychological help-seeking. This instrument employs a 4-point Likert scale from 0 (*strongly disagree*) to 3 (*strongly agree*), with higher scores indicating a stronger willingness to seek help. In this study, the ATSPPH-SF had an acceptable Cronbach α of 0.73 at baseline.

Data Analysis

The data were analyzed following the intent-to-treat principle, including all randomized participants. The Shapiro-Wilk test was employed to assess residual normality for each scale across groups and time points. Intervention effects over time were examined using linear mixed models, which are less sensitive to missing data. Given that residuals for certain variables at specific time points were skewed and the sample size was limited, a bootstrap method with 1000 resamples was applied to estimate the linear mixed model fixed-effect parameters and their 95% CIs, thereby enhancing the robustness of statistical inferences and avoiding strict distributional assumptions [57].

Cohen *d* was calculated to determine both between-group and within-group effect sizes. Owing to baseline differences between the intervention and waitlist control groups, Morris' baseline-adjusted formula was applied to adjust the between-group effect sizes at posttest and follow-up [58]. The effect size was based on mean pre-post change ($M_{post,\ T} - M_{pre,\ T}$) in the intervention group minus the mean pre-post change ($M_{post,\ C} - M_{pre,\ C}$), divided by the pooled pretest standard deviation (SD_{pre}). All analyses were conducted using RStudio (version 4.5.0).

d=(Mpost, T−Mpre, T)−(Mpost, C−Mpre, C)SDpre

where the pooled standard deviation is defined as

SDpre=(nT−1)SDpre,T2+(nC−1)SDpre,C2nT+nC−2

Results

Participants' Demographic Characteristics and Baseline Scores

The detailed flow of participants' recruitment, allocation, and analysis is presented as Figure 1. No significant differences were found between the intervention and control groups in demographic variables or baseline outcome scores (see Table 2).

XSL•FO
RenderX

Figure 1. Participant flow.

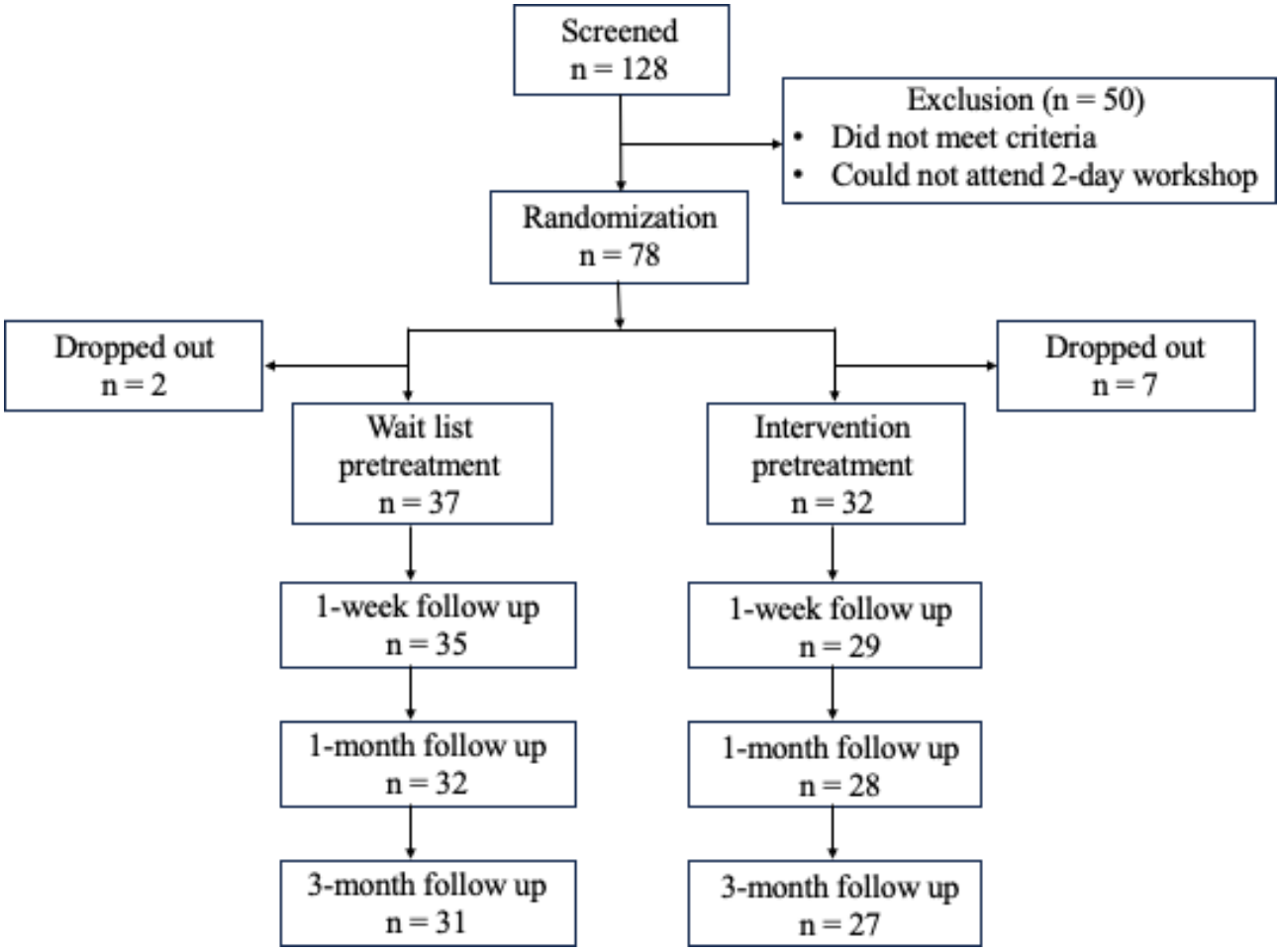

XSL•FO
RenderX

Table . Demographic characteristics and the baseline of participants.

Variable	All	Group		*t* test (*df*)	Chi-square (*df*)	*P* value
		Intervention (n=32)	Control (n=37)			
Age, mean (SD) (y)	23.16 (3.89)	22.47 (2.99)	23.76 (4.48)	−1.39 (67)	__[a]	.17
Gender, n (%)				—	0.39 (1)	.53
Man	17 (24.6)	9 (28.1)	8 (21.6)			
Woman	52 (75.4)	23 (71.9)	29 (78.4)			
Residence, n (%)				—	2.41 (2)	.30
Rural area	24 (34.8)	14 (43.8)	10 (27)			
Township	12 (17.4)	4 (12.5)	8 (21.6)			
City	33 (47.8)	14 (43.8)	19 (51.4)			
Education, n (%)				—	1.41 (2)	.495
Bachelor	40 (58)	18 (56.3)	22 (59.5)			
Master	25 (36.2)	11 (34.4)	14 (37.8)			
Doctor	4 (5.8)	3 (9.4)	1 (2.7)			
Baseline scores, mean (SD)						
SPIN[b]	—	38.97 (15.78)	37.00 (11.69)	−0.58 (67)	—	.56
BFNES[c]	—	46.19 (7.41)	45.51 (7.76)	−0.37 (67)	—	.71
DASS-21[d]	—	20.84 (10.16)	18.54 (10.53)	−0.92 (67)	—	.36
SAKT[e]	—	10.34 (2.67)	11.22 (2.18)	1.47 (67)	—	.14
SASI[f]	—	17.41 (6.71)	17.49 (6.06)	0.05 (67)	—	.96
ATSPPH-SF[g]	—	19.19 (4.64)	19.72 (4.15)	0.51 (67)	—	.61

[a]Not applicable.

[b]SPIN: Social Phobia Inventory.

[c]BFNES: Brief Fear of Negative Evaluation Scale.

[d]DASS-21: Depression Anxiety Stress Scales–Short Form.

[e]SAKT: Social Anxiety Knowledge Test.

[f]SASI: Social Anxiety Stigma Inventory.

[g]ATSPPH-SF: Attitudes Toward Seeking Professional Psychological Help Scale–Short Form.

Shapiro-Wilk Normality Test

Shapiro-Wilk normality tests were conducted separately for the intervention and waitlist control groups at each time point for all outcome variables. The results showed that for both groups, the residuals of the 5 outcome variables across 4 time points were generally normally distributed (W=0.93 - 0.98, P's=.052-.902, n=35). Exceptions were observed in a few cases: the BFNES in the waitlist group at T1 (W=0.93, P<.05) and T4 (W=0.89, P=.004) and the DASS-21 in the intervention group at T2 (W=0.91, P<.05) and T4 (W=0.90, P<.05) and in the waitlist control group at T4 (W=0.89, P=.003). Overall, these findings suggest that the residual distributions did not exhibit significant skewness.

Primary Outcomes

According to the results (Table 3), the main effects of group and time on SPIN scores were not significant. However, the group×time interaction was significant (β=−4.00, 95% bootstrap CI −6.55 to −1.22), indicating that compared to the waitlist control group, the intervention group experienced a significant reduction in social anxiety over time (see Figure 2 for the trend). Moreover, the intervention group exhibited significant reductions with large effect sizes at T2 (d=−0.81, 95% CI −1.28 to −0.33), T3 (d=−0.97, 95% CI −1.44 to −0.49), and T4 (d=−0.81, 95% CI −1.28 to −0.33; Table 4).

XSL•FO
RenderX

Table . Results of linear mixed model analysis.

Variable	Group		Time		Group×time	
	β	95% boot CI	β	95% boot CI	β	95% boot CI
SPIN[a]	2.76	−4.48 to 9.19	3.26	−1.71 to 6.80	−4.00	−6.55 to −1.22
BFNES[b]	1.22	−2.50 to 4.60	0.65	−1.31 to 2.64	−1.37	−2.64 to −0.08
DASS-21[c]	2.89	−2.18 to 7.37	1.20	−1.81 to 4.21	−1.77	−3.42 to 0.17
SAKT[d]	−0.42	−1.93 to 1.10	−0.43	−1.14 to 0.38	0.62	0.05 to 1.17
SASI[e]	0.46	−2.09 to 3.43	−0.85	−2.41 to 0.62	−0.43	−1.43 to 0.62
ATSPPH-SF[f]	−0.75	−2.87 to 0.94	0.02	−1.04, 1.03	0.32	−0.39 to 1.09

[a]SPIN: Social Phobia Inventory.

[b]BFNES: Brief Fear of Negative Evaluation Scale.

[c]DASS-21: Depression Anxiety Stress Scales–Short Form.

[d]SAKT: Social Anxiety Knowledge Test.

[e]SASI: Social Anxiety Stigma Inventory.

[f]ATSPPH-SF: Attitudes Toward Seeking Professional Psychological Help Scale–Short Form.

XSL•FO
RenderX

Figure 2. Change of outcomes. ATSPPH-SF: Attitudes Toward Seeking Professional Psychological Help Scale–Short Form; BFNES: Brief Fear of Negative Evaluation Scale; DASS-21: Depression Anxiety Stress Scales–Short Form; SAKT: Social Anxiety Knowledge Test; SASI: Social Anxiety Stigma Inventory; SPIN: Social Phobia Inventory.

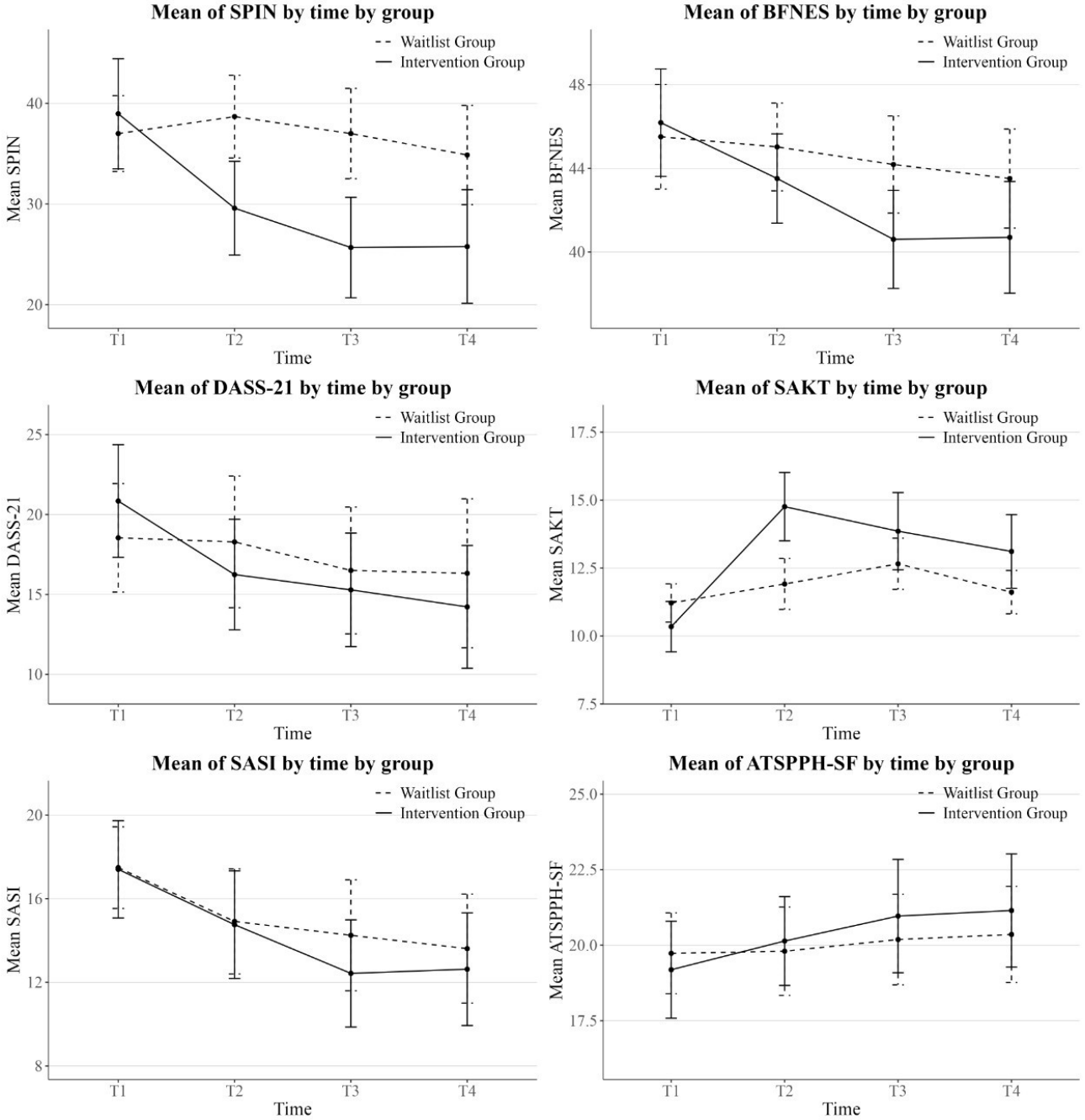

XSL•FO
RenderX

Table . Means, Standard deviations, and Cohen *d* for outcome variables by group and time[a].

Variable	Intervention group		Control group		$d_{b\text{-}g}$[c] (95% CI)
	Mean (SD)	$d_{w\text{-}g}$[b] (95% CI)	Mean (SD)	$d_{w\text{-}g}$	
SPIN[d]					
T1	38.97 (15.78)	__[e]	37.00 (11.69)	—	0.14 (–0.33 to 0.62)
T2	29.31 (13.16)	0.71 (–1.07 to –0.31)	37.59 (12.98)	0.10 (–0.27 to 0.41)	–0.81 (–1.28 to –0.33)
T3	26.03 (13.44)	–0.91 (–1.27 to –0.51)	36.46 (12.71)	–0.08 (–0.45 to 0.29)	–0.97 (–1.44 to –0.49)
T4	25.78 (14.96)	–0.89 (–1.24 to –0.53)	34.87 (14.00)	–0.11 (–0.48 to 0.25)	–0.81 (–1.28 to –0.33)
BFNES[f]					
T1	46.19 (7.41)	—	45.51 (7.76)	—	0.009 (–0.39 to 0.56)
T2	42.81 (6.19)	–0.52 (–0.81 to –0.17)	44.57 (6.48)	–0.01 (–0.44 to 0.32)	–0.29 (–0.76 to 0.19)
T3	40.16 (6.61)	–0.74 (–1.09 to –0.40)	43.00 (7.00)	–0.12 (–0.53 to 0.28)	–0.56 (–1.03 to –0.09)
T4	40.7 (7.08)	–0.76 (–1.04 to –0.42)	43.52 (6.72)	–0.20 (–0.52 to 0.24)	–0.46 (–0.93 to 0.01)
DASS-21[g]					
T1	20.84 (10.16)	—	18.54 (10.53)	—	0.22 (–0.25 to 0.70)
T2	16.66 (9.24)	–0.63 (–1.01 to –0.26)	17.97 (12.17)	0.02 (–0.35 to 0.32)	–0.42 (–0.89 to 0.05)
T3	15.41 (9.85)	–0.71 (–1.17 to –0.31)	16.14 (10.82)	–0.22 (–0.55 to 0.15)	–0.34 (–0.81 to 0.13)
T4	14.22 (10.18)	–0.78 (–1.19 to –0.36)	16.32 (13.23)	–0.07 (–0.41 to 0.31)	–0.43 (–0.90 to 0.05)
SAKT[h]					
T1	10.34 (2.67)	—	11.22 (2.19)	—	–0.36 (–0.84 to 0.12)
T2	14.19 (3.82)	1.21 (0.67 to 1.70)	11.65 (2.98)	0.23 (–0.12 to 0.50)	1.53 (1.06 to 2.01)
T3	13.31 (3.92)	0.98 (0.48 to 1.48)	12.46 (2.86)	0.40 (0.01 to 0.76)	0.86 (0.38 to 1.33)
T4	13.11 (3.59)	0.77 (0.37 to 1.18)	11.61 (2.26)	0.11 (–0.30 to 0.47)	0.98 (0.51 to 1.45)
SASI[i]					
T1	17.41 (6.71)	—	17.49 (6.06)	—	0.01 (–0.49 to 0.46)
T2	14.84 (6.73)	–0.44 (–0.85 to –0.05)	14.51 (7.76)	–0.55 (–0.86 to –0.16)	–0.01 (–0.49 to 0.46)
T3	12.88 (7.09)	–0.78 (–1.19 to –0.38)	14.27 (8.15)	–0.50 (–0.85 to –0.06)	0.27 (–0.75 to 0.20)
T4	12.63 (7.14)	–0.68 (–0.96 to –0.23)	13.61 (7.41)	–0.74 (–1.15 to –0.19)	0.14 (–0.62 to 0.33)
ATSPPH-SF[j]					
T1	19.19 (4.64)	—	19.73 (4.15)	—	0.12 (–0.60 to 0.35)
T2	19.84 (3.96)	0.19 (–0.19 to 0.56)	20.03 (4.41)	0.15 (–0.25 0.42)	0.20 (–0.27 to 0.67)
T3	20.69 (4.86)	0.34 (–0.10 to 0.68)	20.32 (4.36)	0.14 (–0.20, 0.51)	0.30 (–0.17 to 0.77)
T4	21.15 (4.97)	0.39 (0.01 to 0.75)	20.35 (4.52)	0.31 (–0.09, 0.70)	0.31 (–0.17 to 0.78)

[a]Negative values reflect reductions from the baseline. Bootstrapped 95% CIs are reported in parentheses.
[b]Within-group Cohen *d*.
[c]Between-group Cohen *d*.
[d]SPIN: Social Phobia Inventory.
[e]Not available.
[f]BFNES: Brief Fear of Negative Evaluation Scale.
[g]DASS-21: Depression Anxiety Stress Scales–Short Form.
[h]SAKT: Social Anxiety Knowledge Test.
[i]SASI: Social Anxiety Stigma Inventory.
[j]ATSPPH-SF: Attitudes Toward Seeking Professional Psychological Help Scale–Short Form.

XSL•FO
RenderX

Secondary Outcomes

The main effects of group and time on BFNES scores were not significant. However, the group×time interaction effect was significant (β=−1.37, 95% bootstrap CI −2.64 to −0.08), indicating that compared to the waitlist control group, the intervention group experienced a significant reduction in fear of negative evaluation over time. However, this reduction was statistically significant only at T3, with a moderate effect size (*d*=−0.56, 95% CI −1.03 to −0.09).

For DASS-21 scores, neither the main effects of group and time nor the group×time interaction effect was significant (β=-1.77, 95% bootstrap CI −3.42 to 0.17). This indicates that no significant differences existed between the groups in terms of depression, anxiety, or stress symptom trajectories over time.

The main effects of group and time on SAKT scores were not significant. However, the group×time interaction effect was significant (β=.62, 95% bootstrap CI 0.05 to 1.17), indicating that compared to the waitlist control group, the intervention group showed a significant increase in social anxiety knowledge over time. Large effect sizes were observed at T2 (*d*=1.53, 95% CI 1.06-2.00), T3 (*d*=0.86, 95% CI 0.38-1.33), and T4 (*d*=0.98, 95% CI 0.51-1.45).

For SASI scores, the main effects of group and time, as well as the group×time interaction (β=−.43, 95% bootstrap CI −1.43 to 0.62), were not significant. This suggests that there were no significant differences between the groups in terms of social anxiety stigma over time. Similarly, no significant effects were found for ATSPPH-SF scores. The pattern of results mirrored that of the SASI, indicating no significant between-group differences in attitudes toward seeking professional psychological help over time.

Discussion

General Findings

This study examined the effectiveness of a CBT-based MBPG delivered via videoconference for addressing social anxiety in university students. The results showed that the intervention led to significant improvements in social anxiety, including reductions in participants' fear of negative evaluation over a certain period and a marked increase in social anxiety knowledge. However, no significant improvements were observed in emotional distress (ie, depression, anxiety, and stress as measured by the DASS-21), social anxiety stigma, or attitudes toward seeking professional psychological help.

First, compared to the waitlist control group, participants in the intervention group reported significant reductions in social anxiety, with large effect sizes observed in both the short term (1-week posttest and 1-month follow-up) and longer term (3-month follow-up). Moreover, fear of negative evaluation significantly decreased 1 month after the intervention, consistent with the study's hypotheses. A recent meta-analysis reported that various forms of remote CBT yield medium-to-large effect sizes in treating social anxiety [19]. While this study did not directly compare videoconference delivery to other remote CBT modalities, the observed effect sizes align with those reported in prior research on remote CBT for social anxiety, with relatively durable effects. This further suggests that core CBT techniques, particularly cognitive restructuring and behavioral experiments targeting automatic thoughts, can be effectively delivered in both in-person and remote formats, likely without significant differences in efficacy.

Second, this study employed a massed brief intervention, delivering 12 hours of intervention intensively over 2 consecutive days. High-frequency delivery of this kind has been shown to sustain participants' motivation and engagement [36]. The selected intervention components in this study are well-established techniques for reducing social anxiety, and the massed format enables tighter integration across modules, allowing therapeutic gains to accumulate within a short period [59]. Therefore, despite involving fewer sessions than full-protocol interventions, this brief intervention yielded robust therapeutic outcomes. Given that fear of negative evaluation is a core cognitive mechanism in the development and maintenance of social anxiety [60], its reduction further supports the intervention's efficacy. However, in this study, improvement in fear of negative evaluation was significant only at the 1-month follow-up. This suggests that the observed reduction in social anxiety may involve additional mediating mechanisms beyond cognitive change alone, which warrants further investigation in future research.

However, the intervention did not have significant effects on anxiety, depression, or stress, which was inconsistent with the study's hypotheses. Cognitive restructuring, a core component of this intervention, is a widely applicable CBT technique and is theoretically effective for treating anxious or depressive symptoms as well [61]. Moreover, prior studies using videoconferencing group CBT for social anxiety have reported significant reductions in DASS-21 scores [40,41]. A 1-day CBT-based workshop for secondary vocational students with social anxiety also significantly reduced DASS-21 scores, with moderate effect sizes [29]. However, not all prior findings are consistent: an in-person group CBT intervention for social anxiety using a full treatment protocol found no improvements in depressive symptoms [62]. Taken together, the absence of significant change in DASS-21 scores in this study may suggest that the MBPG via videoconference has limitations in the intervention's depth. Compared to full-protocol CBT, the brief, massed format involves fewer techniques and a considerably shortened duration. Specifically, spaced CBT interventions commonly provide opportunities for participants to practice newly learned skills between weekly sessions through homework assignments, thereby consolidating gains and facilitating symptom improvement [63]. Although the MBPG also allowed participants to practice during sessions and lunch breaks, these chances were still limited, which in turn constrained the extent to which negative beliefs and cognitive biases could be corrected. This may inherently limit its capacity to address broader emotional symptom domains beyond social anxiety or reduce the extent of therapeutic change achievable within participants. In addition, the DASS-21 primarily assesses physiological arousal and subjective emotional experiences, rather than cognitive content. As this intervention targeted participants' cognitive biases, improvements in social anxiety were likely driven by cognition-focused changes. Such changes,

XSL•FO
RenderX

however, are not readily captured by the DASS-21. Although the DASS-21 is commonly used in social anxiety interventions, it may be less suitable for detecting the specific types of cognitive changes targeted in this study.

The outcomes related to mental health literacy on social anxiety were somewhat complex, with the corresponding hypotheses only partially supported. Participants' knowledge of social anxiety significantly improved, reflecting the most direct impact of the psychoeducational component and aligning with the study's hypotheses. However, no significant changes were observed in social anxiety stigma or in attitudes toward seeking professional help. Prior research has shown that when individuals' mental health literacy regarding social anxiety improves, they tend to adopt a more objective view of the condition, experience reduced stigma, and show a greater willingness to seek professional help [64]. Furthermore, Cui et al [54] conducted a 6-week, in-person psychoeducation group with the same measurements and found significant improvements in knowledge, stigma, and help-seeking attitudes. We believe several factors may account for the discrepancy between their findings and ours. First, disclosure among people with mental illness is an effective strategy to manage stigma [65], yet opportunities for interpersonal contact were highly constrained in the online format compared with in-person groups. This limitation likely hindered meaningful changes in stigma. Second, help-seeking attitudes are influenced by multiple factors, including mental health literacy and perceived service accessibility [66], and improving these attitudes requires shifts in participants' deeper beliefs and motivations [64]. However, participants in the online workshop were recruited from diverse regions across the country, resulting in variability in access to mental health resources. For participants from low-resource backgrounds, severely limited perceived accessibility of services may have constrained improvements in help-seeking attitudes. Therefore, no significant change in help-seeking attitudes was observed at the group level. Nonetheless, these findings tentatively suggest that improvements in social anxiety symptoms may occur independently of changes in stigmatizing attitudes toward social anxiety.

Implications, Limitations, and Future Research

This study has several notable strengths. First, it employs a videoconference platform to deliver a new format of intervention, that is, the MBPG, which offers numerous advantages. The videoconferencing delivery format overcomes geographical barriers, increasing access to evidence-based interventions. At the same time, however, online delivery inevitably reduces opportunities for in-person exposure. It must be carefully considered when selecting intervention techniques, as the effectiveness of exposure-based strategies may be more substantially limited. In this intervention, both the cognitive restructuring and behavioral experiments were cognition-targeted techniques that do not rely on exposure in vivo, thereby minimizing the potential impact of the videoconferencing format on treatment effects. The massed format—delivered over a short, tightly scheduled period—may improve efficiency and reduce dropout. Moreover, the psychoeducational group format accommodates more participants per session, and its highly structured, slide-based delivery reduces the demands on group leaders while ensuring treatment fidelity. Finally, the use of a randomized controlled trial design provides preliminary yet promising evidence for the efficacy of this intervention model.

The study also has some limitations. First, the sample size was relatively small, and we conducted our intervention only in the nonclinical sample. Future research should expand upon these findings by conducting large-scale, multicenter randomized controlled trials in both clinical and community samples to further evaluate the effects of the MBPG-based program for social anxiety. Second, videoconferencing-based CBT has inherent constraints compared to the face-to-face format. These include increased exposure to distractions, reduced attentional engagement, and limited ability for leaders to observe and monitor participant cues [67], which further reduce opportunities for direct contact and spontaneous communication both between the leaders and participants and among participants themselves. Such limitations may reduce intervention efficacy, warranting improved technological solutions and content organization to address these challenges. In addition, compared to traditional treatments, the massed brief intervention format offers limited opportunities for participants to practice newly acquired cognitive skills. As a result, these skills may not be sufficiently reinforced, which in turn constrains the depth of therapeutic change. Future iterations of the intervention protocol should incorporate additional strategies to address this limitation. Finally, the follow-up period was limited to 3 months. Longer-term follow-up (eg, 6 months, 1 year, or beyond) is necessary to assess the maintenance of treatment gains over time.

We finally derived several implications from this intervention for future research. First, although online delivery is convenient and lowers access barriers, it is important to select a platform that supports essential interactive features, particularly breakout rooms and spaces that allow participants to communicate more freely, which may partially mitigate the limitations associated with reduced in-person exposure. Second, because the content of this workshop was delivered in a highly structured, slide-based format, future dissemination could broaden not only the reach of the program but also the range of potential leaders. Beyond mental health professionals, trained paraprofessionals or nonspecialists (eg, peer counselors or students majoring in psychology or related fields) could deliver the intervention using standardized materials, such as guidance menus and slide decks, thereby enhancing the accessibility of mental health care.

Conclusion

This study demonstrates that a videoconferencing CBT-based massed brief psychoeducational group can produce meaningful and sustained improvements in social anxiety among university students. Future research should focus on scaling up this model through larger randomized controlled trials, adapting it for more diverse populations, and exploring the feasibility of delivery by nonspecialist leaders.

XSL•FO
RenderX

Funding

This research was funded by the Fundamental Research Funds for the Central Universities,and the Research Funds of Renmin University of China (22XNKJ27).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Data curation: LF, WL
Formal analysis: LF, WL
Methodology: LF
Investigation: WL, XT
Funding acquisition: XT
Conceptualization: DD
Visualization: LC, DD
Resources: XT
Supervision: XT
Writing-review & editing: LF, LC, XT
Writing-original drafting: LF

Conflicts of Interest

None declared.

Checklist 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File, 1282 KB - jmir_v28i1e79825_app1.pdf]

References

1. Diagnostic and Statistical Manual of Mental Disorders (5th Ed): American Psychiatric Association; 2013. [doi: 10.1176/appi.books.9780890425596]
2. Morrison AS, Heimberg RG. Social anxiety and social anxiety disorder. Annu Rev Clin Psychol 2013;9(1):249-274. [doi: 10.1146/annurev-clinpsy-050212-185631] [Medline: 23537485]
3. Tang X, Liu Q, Cai F, Tian H, Shi X, Tang S. Prevalence of social anxiety disorder and symptoms among Chinese children, adolescents and young adults: a systematic review and meta-analysis. Front Psychol 2022;13:792356. [doi: 10.3389/fpsyg.2022.792356] [Medline: 36072051]
4. Aderka IM, Hofmann SG, Nickerson A, Hermesh H, Gilboa-Schechtman E, Marom S. Functional impairment in social anxiety disorder. J Anxiety Disord 2012 Apr;26(3):393-400. [doi: 10.1016/j.janxdis.2012.01.003] [Medline: 22306132]
5. Tonge NA, Lim MH, Piccirillo ML, Fernandez KC, Langer JK, Rodebaugh TL. Interpersonal problems in social anxiety disorder across different relational contexts. J Anxiety Disord 2020 Oct;75:102275. [doi: 10.1016/j.janxdis.2020.102275] [Medline: 32891027]
6. Hayward C, Wilson KA, Lagle K, Kraemer HC, Killen JD, Taylor CB. The developmental psychopathology of social anxiety in adolescents. Depress Anxiety 2008;25(3):200-206. [doi: 10.1002/da.20289] [Medline: 17348001]
7. Russell G, Topham P. The impact of social anxiety on student learning and well-being in higher education. J Ment Health 2012 Aug;21(4):375-385. [doi: 10.3109/09638237.2012.694505] [Medline: 22823093]
8. Seo EH, Yang HJ, Kim SG, Yoon HJ. Ego-resiliency moderates the risk of depression and social anxiety symptoms on suicidal ideation in medical students. Ann Gen Psychiatry 2022 Jun 18;21(1):19. [doi: 10.1186/s12991-022-00399-x] [Medline: 35717375]
9. Barkowski S, Schwartze D, Strauss B, Burlingame GM, Barth J, Rosendahl J. Efficacy of group psychotherapy for social anxiety disorder: a meta-analysis of randomized-controlled trials. J Anxiety Disord 2016 Apr;39:44-64. [doi: 10.1016/j.janxdis.2016.02.005] [Medline: 26953823]
10. Mayo-Wilson E, Dias S, Mavranezouli I, et al. Psychological and pharmacological interventions for social anxiety disorder in adults: a systematic review and network meta-analysis. Lancet Psychiatry 2014 Oct;1(5):368-376. [doi: 10.1016/S2215-0366(14)70329-3] [Medline: 26361000]
11. Sun L, Dai X, Zhu S, Liu Z, Zhongming Z. Psychotherapies for social anxiety disorder in adults: a systematic review and Bayesian network meta-analysis. J Affect Disord 2025 Jun 1;378:301-319. [doi: 10.1016/j.jad.2025.02.092] [Medline: 40023260]

XSL•FO
RenderX

12. Pilling S, Mayo-Wilson E, Mavranezouli I, et al. Recognition, assessment and treatment of social anxiety disorder: summary of NICE guidance. BMJ 2013 May 22;346:f2541. [doi: 10.1136/bmj.f2541] [Medline: 23697669]
13. Firth J, Torous J, Nicholas J, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry 2017 Oct;16(3):287-298. [doi: 10.1002/wps.20472] [Medline: 28941113]
14. Hedman E, Andersson G, Ljótsson B, et al. Internet-based cognitive behavior therapy vs. cognitive behavioral group therapy for social anxiety disorder: a randomized controlled non-inferiority trial. PLoS One 2011 Mar 25;6(3):e18001. [doi: 10.1371/journal.pone.0018001] [Medline: 21483704]
15. Leichsenring F, Leweke F. Social anxiety disorder. N Engl J Med 2017 Jun 8;376(23):2255-2264. [doi: 10.1056/NEJMcp1614701] [Medline: 28591542]
16. Orsolini L, Pompili S, Salvi V, Volpe U. A systematic review on telemental health in youth mental health: focus on anxiety, depression and obsessive-compulsive disorder. Medicina (Kaunas) 2021 Jul 31;57(8):793. [doi: 10.3390/medicina57080793] [Medline: 34440999]
17. Ganho-Ávila A, Vieira Figueiredo D, Vagos P. Online cognitive therapy for social anxiety disorder in adolescence: a clinical case study using the CT@TeenSAD. Clin Case Stud 2022 Dec;21(6):533-551. [doi: 10.1177/15346501221091519]
18. Peros OM, Webb L, Fox S, Bernstein A, Hoffman L. Conducting exposure-based groups via telehealth for adolescents and young adults with social anxiety disorder. Cogn Behav Pract 2021 Nov;28(4):679-689. [doi: 10.1016/j.cbpra.2021.04.001] [Medline: 34690482]
19. Winter HR, Norton AR, Burley JL, Wootton BM. Remote cognitive behaviour therapy for social anxiety disorder: a meta-analysis. J Anxiety Disord 2023 Dec;100:102787. [doi: 10.1016/j.janxdis.2023.102787] [Medline: 37890219]
20. Vu BP, Dang HM, Andersen PN. Does videoconferencing-based cognitive behavioral therapy for anxious youth work? A systematic review of the literature. Curr Treat Options Psych 2023 Sep 2;10(4):511-533. [doi: 10.1007/s40501-023-00302-9]
21. Social anxiety disorder: recognition, assessment and treatment [clinical guideline no. CG159]. : National Institute for Health and Care Excellence; 2013 URL: https://www.nice.org.uk/guidance/cg159 [accessed 2025-12-19]
22. Bennett S, Myles-Hooton P, Schleider J, Shafran R, editors. Oxford Guide to Brief and Low Intensity Interventions for Children and Young People: Oxford University Press; 2022. [doi: 10.1093/med-psych/9780198867791.001.0001]
23. Shafran R, Myles-Hooton P, Bennett S, Öst LG. The concept and definition of low intensity cognitive behaviour therapy. Behav Res Ther 2021 Mar;138:103803. [doi: 10.1016/j.brat.2021.103803] [Medline: 33540242]
24. Cape J, Whittington C, Buszewicz M, Wallace P, Underwood L. Brief psychological therapies for anxiety and depression in primary care: meta-analysis and meta-regression. BMC Med 2010 Jun 25;8(1):38. [doi: 10.1186/1741-7015-8-38] [Medline: 20579335]
25. Heimberg RG, Liebowitz MR, Hope DA, Schneier FR, editors. Social Phobia: Diagnosis, Assessment, and Treatment: Guilford Press; 1995. [doi: 10.1192/S0007125000147543]
26. Heimberg RG, Becker RE. Cognitive-Behavioral Group Therapy for Social Phobia: Basic Mechanisms and Clinical Strategies: Guilford Press; 2002.
27. Gallagher HM, Rabian BA, McCloskey MS. A brief group cognitive-behavioral intervention for social phobia in childhood. J Anxiety Disord 2004;18(4):459-479. [doi: 10.1016/S0887-6185(03)00027-6] [Medline: 15149708]
28. Jain N, Stech E, Grierson AB, et al. A pilot study of intensive 7-day internet-based cognitive behavioral therapy for social anxiety disorder. J Anxiety Disord 2021 Dec;84:102473. [doi: 10.1016/j.janxdis.2021.102473] [Medline: 34534800]
29. Li H, Tang X. Effects of a cognitive behavioral therapy-based workshop intervention on social anxiety among secondary vocational students: a randomized controlled trial. Behav Ther 2025 Nov;56(6):1118-1132. [doi: 10.1016/j.beth.2025.05.005] [Medline: 41139107]
30. Gladding ST. Groups: A Counseling Specialty: Pearson; 2016.
31. Brown BM. Psychoeducation group work. Couns Human Dev 1997;29(7):1-14.
32. Kutcher S, Wei Y, Coniglio C. Mental health literacy: past, present, and future. Can J Psychiatry 2016 Mar;61(3):154-158. [doi: 10.1177/0706743715616609] [Medline: 27254090]
33. Houghton S, Saxon D. An evaluation of large group CBT psycho-education for anxiety disorders delivered in routine practice. Patient Educ Couns 2007 Sep;68(1):107-110. [doi: 10.1016/j.pec.2007.05.010] [Medline: 17582724]
34. Öst LG, Cederlund R, Reuterskiöld L. Behavioral treatment of social phobia in youth: does parent education training improve the outcome? Behav Res Ther 2015 Apr;67:19-29. [doi: 10.1016/j.brat.2015.02.001] [Medline: 25727679]
35. Yalom ID, Leszcz M. The Theory and Practice of Group Psychotherapy, 6th edition: Basic Books; 2020. [doi: 10.1176/appi.psychotherapy.20210007]
36. Öst LG, Ollendick TH. Brief, intensive and concentrated cognitive behavioral treatments for anxiety disorders in children: a systematic review and meta-analysis. Behav Res Ther 2017 Oct;97:134-145. [doi: 10.1016/j.brat.2017.07.008] [Medline: 28772195]
37. Deacon B, Abramowitz J. A pilot study of two-day cognitive-behavioral therapy for panic disorder. Behav Res Ther 2006 Jun;44(6):807-817. [doi: 10.1016/j.brat.2005.05.008] [Medline: 16084488]

XSL•FO
RenderX

38. Avramchuk O, Nizdran-Fedorovych O, Blozva P, Plevachuk O. Internet-delivered low-intensity CBT for people with social anxiety disorder in a period of COVID-19: results of pilot research. Wiad Lek 2022;75(12):3109-3114. [doi: 10.36740/WLek202212136] [Medline: 36723335]
39. Noda S, Shirotsuki K, Nakao M. Low-intensity mindfulness and cognitive-behavioral therapy for social anxiety: a pilot randomized controlled trial. BMC Psychiatry 2024 Mar 7;24(1):190. [doi: 10.1186/s12888-024-05651-0] [Medline: 38454396]
40. Nauphal M, Swetlitz C, Smith L, Rosellini AJ. A preliminary examination of the acceptability, feasibility, and effectiveness of a telehealth cognitive-behavioral therapy group for social anxiety disorder. Cogn Behav Pract 2021 Nov;28(4):730-742. [doi: 10.1016/j.cbpra.2021.04.011]
41. Shapiro IR, Boyd JE, McCabe RE, Rowa K. Lost connection? Comparing group cohesion and treatment outcomes between videoconference and in-person cognitive behavioural group therapy for social anxiety disorder and other anxiety disorders. Behav Cogn Psychother 2025 Jun;53(2):159-173. [doi: 10.1017/S1352465825000013] [Medline: 40077887]
42. Wang Y, Chen X, Chen X, et al. Intervention effects of Internet-based cognitive behavior therapy on social anxiety among medical students. Chin J Sch Health 2013;34(2):139-141. [doi: 10.16835/j.cnki.1000-9817.2013.02.006]
43. Fernandez E, Salem D, Swift JK, Ramtahal N. Meta-analysis of dropout from cognitive behavioral therapy: magnitude, timing, and moderators. J Consult Clin Psychol 2015 Dec;83(6):1108-1122. [doi: 10.1037/ccp0000044] [Medline: 26302248]
44. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
45. Hope DA, Heimberg RG, Turk CL. Managing Social Anxiety, Therapist Guide: A Cognitive-Behavioral Therapy Approach, 3rd edition: Oxford University Press; 2019. [doi: 10.1093/med-psych/9780190247591.001.0001]
46. Leigh E, Clark DM. Understanding social anxiety disorder in adolescents and improving treatment outcomes: applying the cognitive model of Clark and Wells (1995). Clin Child Fam Psychol Rev 2018 Sep;21(3):388-414. [doi: 10.1007/s10567-018-0258-5] [Medline: 29654442]
47. Hofmann SG, Otto MW. Cognitive Behavioral Therapy for Social Anxiety Disorder: Evidence-Based and Disorder-Specific Treatment Techniques: Routledge; 2018. [doi: 10.4324/9781315617039]
48. Connor KM, Davidson JR, Churchill LE, Sherwood A, Foa E, Weisler RH. Psychometric properties of the Social Phobia Inventory (SPIN). New self-rating scale. Br J Psychiatry 2000 Apr;176(4):379-386. [doi: 10.1192/bjp.176.4.379] [Medline: 10827888]
49. Xiao R, Wu W, Zhang W. The reliability and validity of the Chinese version of social phobia inventory. West China Med J 2007(3):477-479 [FREE Full text]
50. Leary MR. A brief version of the Fear of Negative Evaluation Scale. Pers Soc Psychol Bull 1983 Sep;9(3):371-375. [doi: 10.1177/0146167283093007]
51. Chen Z. Fear of negative evaluation and test anxiety in middle school students. Chinese Mental Health J 2002(12):855-857 [FREE Full text]
52. Lovibond PF, Lovibond SH. The structure of negative emotional states: comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. Behav Res Ther 1995 Mar;33(3):335-343. [doi: 10.1016/0005-7967(94)00075-u] [Medline: 7726811]
53. Gong X, Xie X, Xu R, Luo Y. Psychometric properties of the Chinese versions of DASS-21 in Chinese college students. Chinese J Clin Psychol 2010;18(4):443-446. [doi: 10.16128/j.cnki.1005-3611.2010.04.020]
54. Cui L, Zhang J, Hu H, Jia S, Tang X. Effectiveness and feasibility of a cognitive-behavioral therapy-based brief psychoeducational group intervention for social anxiety in university students. Res Soc Work Pract 2025 Nov 28. [doi: 10.1177/10497315251399894]
55. Fischer EH, Farina A. Attitudes toward seeking professional psychological help: a shortened form and considerations for research. J Coll Stud Dev 1995;36(4):368-373 [FREE Full text]
56. Fang S, Yang B, Yang F, Zhou Y. Study on reliability and validity of the Chinese version of the Attitudes toward Seeking Professional Psychological Help Scale-Short Form for community population in China. Chinese Nurs Res 2019;33(14):2410-2414. [doi: 10.12102/j.issn.1009-6493.2019.14.009]
57. Luke SG. Evaluating significance in linear mixed-effects models in R. Behav Res Methods 2017 Aug;49(4):1494-1502. [doi: 10.3758/s13428-016-0809-y] [Medline: 27620283]
58. Morris SB. Estimating effect sizes from pretest-posttest-control group designs. Organ Res Methods 2008 Apr;11(2):364-386. [doi: 10.1177/1094428106291059]
59. Pincus DB, Elkins RM, Hardway CL. Intensive treatments for adolescents with panic disorder and agoraphobia: helping youth move beyond avoidance. Psychopathol Rev 2014 Jul;a1(1):189-194. [doi: 10.5127/pr.033313]
60. Hofmann SG. Cognitive factors that maintain social anxiety disorder: a comprehensive model and its treatment implications. Cogn Behav Ther 2007;36(4):193-209. [doi: 10.1080/16506070701421313] [Medline: 18049945]
61. Santos B, Pinho L, Nogueira MJ, Pires R, Sequeira C, Montesó-Curto P. Cognitive restructuring during depressive symptoms: a scoping review. Healthcare (Basel) 2024 Jun 28;12(13):1292. [doi: 10.3390/healthcare12131292] [Medline: 38998827]

XSL•FO
RenderX

62. Rapee RM, McLellan LF, Carl T, et al. Comparison of transdiagnostic treatment and specialized social anxiety treatment for children and adolescents with social anxiety disorder: a randomized controlled trial. J Am Acad Child Adolesc Psychiatry 2023 Jun;62(6):646-655. [doi: 10.1016/j.jaac.2022.08.003] [Medline: 35987298]
63. McEvoy PM, Johnson AR, Kazantzis N, Egan SJ. Predictors of homework engagement in group CBT for social anxiety: client beliefs about homework, its consequences, group cohesion, and working alliance. Psychother Res 2024 Jan;34(1):68-80. [doi: 10.1080/10503307.2023.2286993] [Medline: 38109521]
64. Coles ME, Ravid A, Gibb B, George-Denn D, Bronstein LR, McLeod S. Adolescent mental health literacy: young people's knowledge of depression and social anxiety disorder. J Adolesc Health 2016 Jan;58(1):57-62. [doi: 10.1016/j.jadohealth.2015.09.017] [Medline: 26707229]
65. Corrigan PW, Larson JE, Michaels PJ, et al. Diminishing the self-stigma of mental illness by coming out proud. Psychiatry Res 2015 Sep 30;229(1-2):148-154. [doi: 10.1016/j.psychres.2015.07.053] [Medline: 26213379]
66. Hom MA, Stanley IH, Joiner TE. Evaluating factors and interventions that influence help-seeking and mental health service utilization among suicidal individuals: a review of the literature. Clin Psychol Rev 2015 Aug;40:28-39. [doi: 10.1016/j.cpr.2015.05.006] [Medline: 26048165]
67. Luo C, Sanger N, Singhal N, et al. A comparison of electronically-delivered and face to face cognitive behavioural therapies in depressive disorders: a systematic review and meta-analysis. EClinicalMedicine 2020 Jul;24:100442. [doi: 10.1016/j.eclinm.2020.100442] [Medline: 32775969]

Abbreviations

ATSPPH-SF: Attitudes Toward Seeking Professional Psychological Help Scale−Short Form
BFNES: Brief Fear of Negative Evaluation Scale
CBT: cognitive-behavioral therapy
DASS-21: Depression Anxiety Stress Scales
MBPG: Massed Brief Psychoeducational Group
SAKT: Social Anxiety Knowledge Test
SASI: Social Anxiety Stigma Inventory
SPIN: Social Phobia Inventory

Edited by A Schwartz, M Balcarras; submitted 30.Jun.2025; peer-reviewed by G Alcolado, S Chen; revised version received 15.Dec.2025; accepted 16.Dec.2025; published 06.Jan.2026.

Please cite as:
Feng L, Liu W, Cui L, Dobson D, Tang X
Cognitive-Behavioral Therapy−Based Massed Brief Psychoeducational Group via Videoconference for Social Anxiety: Randomized Controlled Trial
J Med Internet Res 2026;28:e79825
URL: https://www.jmir.org/2026/1/e79825
doi:10.2196/79825

© Lele Feng, Wei Liu, Liechuan Cui, Deborah Dobson, Xinfeng Tang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Smartphone-Based Digital Eczema Education Program for Atopic Dermatitis in Children Aged 0 to 6 Years: Multicenter, Randomized, Parallel Controlled Clinical Study

Huan Yang[1*], MD, PhD; Hong Shu[2*], MD; Liu-hui Wang[3*], MD; Ping Li[4*], MD, PhD; Yun-ling Li[5], MD; Qin-feng Li[6], MD; Xiu-ping Han[7], MD; Jing Tian[8], MD; Jing Chang[9], MD; Hua Qian[10], MD, PhD; Jing-ping Chen[11], MM; Xin-qiang Ding[12], MD; Pan-qian Wu[2], MD; Li-min Dou[3], MD; Zhen Luo[4], MD; Wei Li[5], MM; Yang-yang Lin[6], MD; Lin Li[7], MD; Shu-zhen Yue[9], MD; Yang Gu[10], MM; Li Yang[11], MD; Xiao-hong Sun[12], MD; Xiao-yan Luo[1*], MD, PhD; Lin Ma[8*], MD, PhD; Hua Wang[1*], MD, PhD

[1]Department of Dermatology, Children's Hospital of Chongqing Medical University, National Clinical Research Center for Child Health and Disorders, Ministry of Education Key Laboratory of Child Development and Disorders, NO.136 Zhongshan 2nd Road, Yuzhong District, Chongqing, China

[10]Department of Dermatology, Children's Hospital of Soochow University, Suzhou, China

[11]Department of Dermatology, Guangzhou Woman and Children's Medical Center, Guangzhou, China

[12]Department of Dermatology, Xi'an Children's Hospital, Xi'an, China

[2]Department of Dermatology, Kunming Children's Hospital, Kunming, China

[3]Department of Dermatology, Children's Hospital of Fudan University, Shanghai, China

[4]Department of Dermatology, Shenzhen Children's Hospital, Shenzhen, China

[5]Department of Dermatology, The Children's Hospital, Zhejiang University School of Medicine, National Clinical Research Center of Child Health, Hangzhou, China

[6]Department of Dermatology, Tianjin Children's Hospital, Tianjin, China

[7]Department of Dermatology, Shengjing Hospital of China Medical University, Shenyang, China

[8]Department of Dermatology, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing, China

[9]Department of Dermatology, Hunan Children's Hospital, Changsha, China

[*]these authors contributed equally

Corresponding Author:
Hua Wang, MD, PhD
Department of Dermatology, Children's Hospital of Chongqing Medical University, National Clinical Research Center for Child Health and Disorders, Ministry of Education Key Laboratory of Child Development and Disorders, NO.136 Zhongshan 2nd Road, Yuzhong District, Chongqing, China

Abstract

Background: Atopic dermatitis (AD) is a chronic, relapsing inflammatory skin condition that affects approximately 10% to 20% of children, imposing substantial health and economic burdens. Although education for patients and caregivers is acknowledged as a crucial element in the management of AD, conventional approaches, such as workshops or in-person consultations, are often resource intensive and face challenges related to scalability, personalization, and relapse prevention. Digital tools present promising alternatives; however, empirical evidence supporting their effectiveness in young children is currently limited.

Objective: This study aimed to evaluate whether a smartphone-based patient-caregiver educational program could reduce relapse rates in children aged 0 to 6 years with moderate-to-severe AD, compared with conventional outpatient consultation alone.

Methods: In this multicenter, randomized, parallel-controlled trial, 615 children were enrolled across 12 tertiary pediatric dermatology centers in China and randomized (1:1) to receive either a smartphone-based digital education program with standard care (intervention group) or conventional face-to-face consultation only (control group). The 12-week digital program, delivered via the WeChat-based *Skin Care E-Station* platform, included structured multimedia modules, interactive educational materials, and a dynamic electronic action plan tailored to the child's age and disease stage. The primary endpoint was the 12-week relapse rate after the acute treatment phase. The secondary endpoints included changes in disease severity (Scoring Atopic Dermatitis, Peak Pruritus Numerical Rating Scale, and Patient-Oriented Eczema Measure) and quality of life (Children's Dermatology Life Quality Index or Infant's Dermatitis Quality of Life Index and Dermatitis Family Impact) up to 52 weeks.

Results: Among 615 randomized participants (mean age 3.3, SD 1.7 y; n=317, 51.5% male), relapse at 12 weeks occurred significantly less frequently in the digital education group than in the control group (16.6% vs 24.0%; relative risk 0.69, 95% CI 0.50 - 0.96; *P*=.02). Kaplan-Meier analysis showed superior relapse-free survival over the first 100 days (hazard ratio 0.688,

XSL•FO
RenderX

95% CI 0.490 - 0.966; *P*=.03). Differences in relapse rates beyond 12 weeks and in secondary outcomes were not statistically significant. Engagement tracking indicated high adherence to the intervention, with 58.0% of caregivers maintaining regular weekly use of the digital platform.

Conclusions: A structured smartphone-based patient-caregiver educational intervention significantly reduced short-term relapse risk among young children with moderate-to-severe AD, likely through improved caregiver recognition and early management of disease flares. Although effects diminished beyond 12 weeks, this approach demonstrates that scalable digital education is a feasible and effective adjunct to standard care in pediatric AD. Future research should focus on sustaining engagement, optimizing long-term reinforcement, and assessing cost-effectiveness in diverse caregiver populations.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2000031474; https://www.chictr.org.cn/showproj.html?proj=32400

(J Med Internet Res 2026;28:e79559) doi:10.2196/79559

KEYWORDS

atopic dermatitis; digital education; smartphone-based intervention; caregiver engagement; relapse prevention; randomized controlled trial; pediatric dermatology; telemedicine portal; self-management; multicenter study

Introduction

Atopic dermatitis (AD) is a chronic inflammatory skin condition characterized by intense pruritus and eczematous lesions, accompanied by episodic exacerbations and persistent skin symptoms [1]. AD is recognized as a global health concern, affecting up to 20% of children and approximately 3% of adults worldwide [2]. In China, the prevalence of AD stands at approximately 30.48% among infants aged <1 year [3] and 12.94% among children aged 1 to 7 years [4], imposing substantial economic and public health burdens [5,6]. The recurrent nature of AD presents unique challenges, as patients often experience periods of exacerbation interspersed with phases of relative remission [1]. This cyclical pattern highlights the necessity for effective management strategies that not only address current symptoms but also equip patients and their caregivers with the knowledge and skills to prevent future recurrences [7].

Patient-caregiver education has emerged as a pivotal component of AD management, as highlighted in many international guidelines [8-12]. It is important to note that for pediatric AD populations, the primary focus of education is often on parents and caregivers, especially for infants and young children. As children grow older and their understanding develops, education can be progressively tailored and delivered directly to the patients themselves, in an age-appropriate manner. Providing patients and their caregivers with comprehensive information about the condition, its triggers, management strategies, and the importance of adherence to treatment regimens can empower them to take an active role in their own care [13]. Studies have shown that patients and families who receive structured educational programs are more likely to engage in treatment and preventive measures, leading to better overall outcomes and reduced disease severity, as evidenced by metrics such as Scoring Atopic Dermatitis (SCORAD) and Children's Dermatology Life Quality Index (CDLQI) [14]. Furthermore, patient-caregiver education may contribute to interrupting the "atopic march" toward comorbidities such as allergic rhinitis [15]. Despite the recognized importance of education, gaps remain in the current approaches used to educate patients with AD and their caregivers. Existing educational models, including workshops [16-19], eczema schools [20-23], and printed materials [19,24-26], are resource intensive, lack personalization, or fail to sufficiently address the specific concerns and experiences of patients and caregivers. As a result, patients and their families may feel overwhelmed or unsupported in managing their condition, leading to suboptimal adherence and an increased risk of recurrent flares [27,28]. In addition, prior investigations have predominantly focused on the short-term efficacy of educational interventions, whereas the impact of education on the prevention of AD relapse remains considerably underexplored [29].

The COVID-19 pandemic has instigated a paradigm shift toward digital health [30], resulting in an increased willingness among patients with AD and caregivers to use digital tools. Mobile platforms that integrate multimedia content, such as videos, interactive texts, and illustrated narratives, offer advantages in accessibility, engagement, and personalized learning experiences. These platforms theoretically address traditional limitations by providing scalable and adaptive interventions. However, empirical evidence supporting such innovations remains limited, particularly in children with AD.

In this multicenter, randomized controlled trial (RCT), we aim to assess the efficacy of a smartphone-based digital educational program in comparison to conventional outpatient consultation alone on the relapse rates of AD. By leveraging high smartphone penetration, this trial seeks to develop a more effective educational intervention that will ultimately reduce the burden of AD.

Methods

Study Design

This multicenter, parallel RCT (ChiCTR2000031474) was conducted within the pediatric dermatology departments of 12 tertiary public hospitals across China, including locations in Chongqing, Beijing, Liaoning, Shanghai, Shenzhen, Zhejiang, Hunan, Xi'an, Suzhou, Tianjin, Kunming, and Guangdong (Multimedia Appendix 1). The trial adhered to the CONSORT (Consolidated Standards of Reporting Trials) reporting guidelines (Checklist 1).

XSL•FO
RenderX

Participants

Children aged 0 to 6 years diagnosed with AD according to the American Academy of Dermatology Consensus criteria were eligible for inclusion if they presented with moderate-to-severe AD (SCORAD ≥25; Investigator's Global Assessment [IGA] ≥3) and had caregivers proficient in reading Chinese characters and using a smartphone. All participants were recruited through in-person clinical visits. Exclusion criteria included children with severe infections, psychiatric disorders, primary or secondary immune deficiencies, malignancies, or other medical conditions that significantly impair quality of life. Additionally, caregivers with mental disorders or cognitive impairment, prior participation in any patient-caregiver education program, and severe AD requiring systemic treatment were excluded.

Most caregivers were the patients' parents, and a minority were grandparents. All caregivers lived in the same household as the child, as families of "left-behind children," defined as children who remain in their hometowns in China for more than 6 months under the care of grandparents or other relatives while both parents migrate to urban areas for work, were excluded to avoid potential confounding due to social and environmental separation. Caregiver demographic details (eg, age and education level) were not collected, as the original study design focused on patient-level outcomes.

Randomization and Blinding

Eligible patients were randomly assigned in a 1:1 ratio to either the intervention group (smartphone-based digital education program plus standard care) or the control group (conventional outpatient consultation only). Randomization was conducted using a computer-generated number table with a block size of 4. The digital educational intervention protocol was developed and administered through a secure cloud-based platform operated by an independent clinical research organization (CRO) not involved in trial execution. This third-party CRO was responsible for maintaining the randomization database, delivering standardized digital educational content (including interactive modules, video tutorials, and self-assessment tools), and monitoring intervention adherence by automated adherence reminders (SMS text messaging or telephone) and digital footprint tracking (platform logins and content interaction time). Crucially, the CRO had no involvement in participant recruitment, clinical management, or outcome evaluation processes to ensure blinding integrity.

Intervention and Procedures

Overview

Before trial initiation, all participating dermatologists completed a 4-hour standardized training program on evidence-based patient-caregiver education.

At baseline enrollment, all patients were provided with a 1-page education leaflet (Multimedia Appendix 2) and completed disease severity assessments along with standardized questionnaires. Subsequently, participants entered a 2-week acute-phase treatment period, conducted strictly according to established clinical guidelines [10-12]. The acute-phase regimen included once-daily topical pharmacotherapy, 0.05% desonide or 0.1% hydrocortisone butyrate for those aged 0 to 2 years and 0.1% mometasone furoate for children older than 2 years, combined with nonpharmacological care. The nonpharmacological regimen consisted of twice-daily application of a ceramide-dominant emollient and daily bathing restricted to fewer than 10 minutes at water temperatures below 38°C. Patient status was reassessed every 14 (SD 2) days using the IGA. Patients achieving an IGA score ≤2 were advanced to the maintenance phase.

Digital Education Intervention

Upon entering the maintenance phase, participants in the intervention group were instructed to receive the digital health education program via the WeChat (Tencent Holdings Ltd)–based online platform "Skin Care E-Station" (Figure 1). This digital portal contained all educational components and could be accessed by scanning a QR code, without requiring any additional software installation. A procedural overview of patient enrollment and identity verification within the platform is provided in Multimedia Appendix 3. As WeChat is a widely used multifunctional communication, social, and payment application in China, this ensured high accessibility and user engagement. The platform consisted of three core components: (1) structured multimedia educational modules, (2) automated follow-up reminders, and (3) emergency response protocols for disease exacerbations.

XSL•FO
RenderX

Figure 1. Skin Care E-Station workflow. Participants in the digital education group accessed the clinical research organization–managed smartphone platform, receiving thrice-weekly education action plans (Mondays/Wednesdays/Fridays) for 12 wk alongside 54 structured multimedia modules (text/images/video/animations) distributed evenly throughout the 52-week program. The content covered AD pathophysiology, care routines, treatments, lifestyle, and mental health support, with real-time clinician access via an encrypted telemedicine portal (<8 h response) during flares. The illustration was created on the FIGDRAW 2.0 platform. AD: atopic dermatitis.

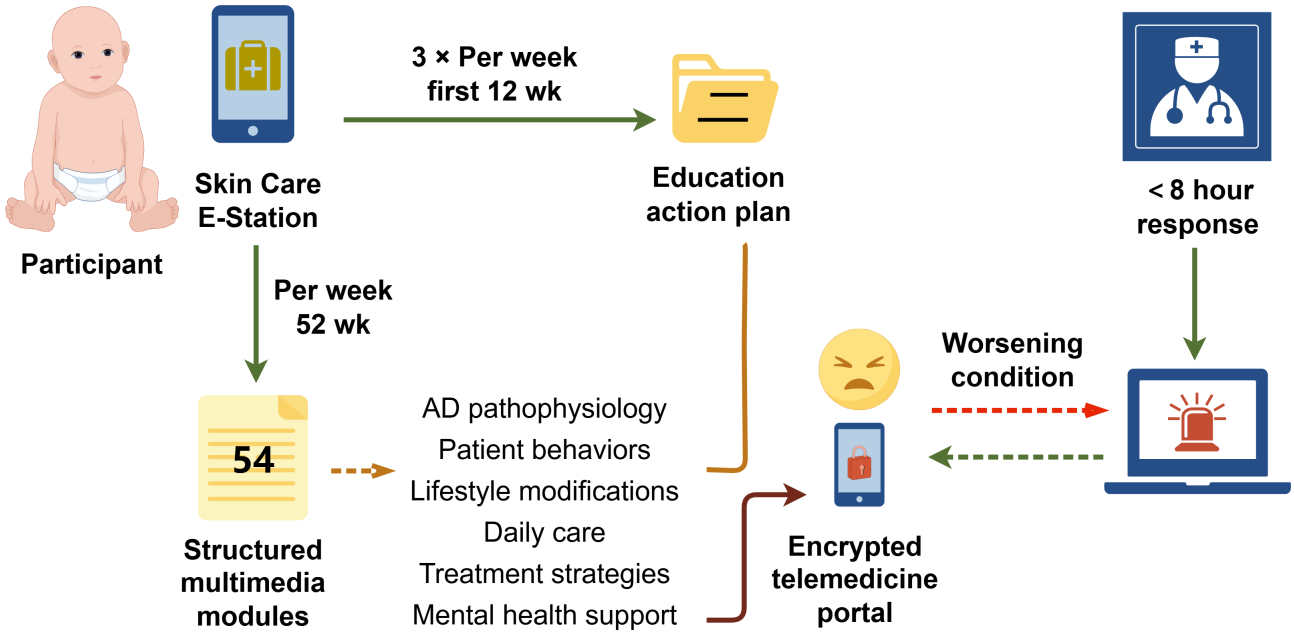

The digital education comprised 2 major elements. The first element was an electronic education action plan (EAP). Each participant in the intervention group received a digital, dynamically tailored EAP (Multimedia Appendix 4), customized according to the child's age group (1‑2 y vs 2‑6 y) and disease phase (acute flare vs maintenance). The EAP provided phase-specific guidance on bathing routines, emollient application, medication use, self-assessment of disease severity, and access to emergency consultations. The EAP was automatically delivered 3 times per week (Mondays, Wednesdays, and Fridays) for 12 consecutive weeks. The second element was structured multimedia educational modules. The core educational curriculum was collaboratively developed by the 12 participating centers, integrating an analysis of the most frequent caregiver concerns and key clinician recommendations from routine practice, supplemented by insights and experiences from other research teams [17-24]. This content was organized into 6 domains (AD pathophysiology, patient behaviors, lifestyle modifications, daily care, treatment strategies, and mental health support) and subsequently produced by an independent third-party company into 54 structured multimedia modules. These modules were delivered in formats tailored for children aged 0 to 6 years and their caregivers, including illustrated text (sample in Multimedia Appendix 5), images (sample in Multimedia Appendix 6), videos (sample in Multimedia Appendix 7), and animated stories (sample in Multimedia Appendix 8). The intervention group did not receive any additional systematic face-to-face patient education during scheduled follow-up visits or at unscheduled clinical encounters.

Control Group

Participants in the control group received only conventional education, consisting of a 15-minute face-to-face counseling session conducted at each scheduled follow-up visit. They were also provided with access to the "Skin Care E-Station" platform for noneducational purposes, including data collection and automated follow-up reminders, but did not receive any educational content, such as the EAP or structured multimedia modules.

Maintenance-Phase Treatment Regimen

Both the intervention and control groups received identical pharmacological treatments during the maintenance phase. In the 0 to 2 years age group, maintenance therapy comprised twice-weekly application of 0.05% desonide or 0.1% hydrocortisone butyrate cream, together with daily emollient use. In children aged >2 years, maintenance consisted of twice-weekly application of 0.03% tacrolimus ointment, alongside regular emollient therapy.

Assessments and Outcomes

Assessments were conducted at 4, 8, 12, 24, 36, and 52 weeks following randomization. Disease severity was evaluated using SCORAD, IGA, Peak Pruritus Numerical Rating Scale (PP-NRS) for pruritus, and the Patient-Oriented Eczema Measure (POEM). Quality of life was assessed using the Dermatitis Family Impact (DFI), the CDLQI, and the Infant's Dermatitis Quality of Life Index (IDQOL). The assessments were completed as follows: If the child was able to attend an in-person visit, the evaluations were performed by physicians not involved in the educational intervention; if not, caregivers uploaded photos and completed the PP-NRS, POEM, Quality of Life (QoL) questionnaires, and SCORAD self-assessment (Multimedia Appendix 9) on the platform. Subsequently, physicians not involved in the educational intervention performed the IGA and SCORAD assessments based on the uploaded photos and the SCORAD self-assessment form.

In addition, the timing of each relapse was meticulously recorded. Relapse was defined as an increase of ≥10 points in

XSL•FO
RenderX

the SCORAD score relative to the value recorded at the conclusion of the 2-week acute-phase treatment. The primary endpoint was the relapse rate at 12 weeks. The secondary endpoints included the changes in disease severity scores (SCORAD, PP-NRS, and POEM) and quality of life scores (CDLQI/IDQOL and DFI) from week 2 (end of the acute-phase treatment) to weeks 4, 8, 12, 24, 36, and 52 in both groups.

Statistical Analyses

Statistical analyses were performed using R (version 4.0.5; R Foundation for Statistical Computing). Sample size estimation was based on an expected 12-week relapse rate of 30% in the digital education group and 45% in the control group, which were derived from a multicenter randomized controlled clinical study on long-term intermittent maintenance therapy for children with AD conducted in China [31]. Assuming a 1:1 allocation ratio and a 30% expected dropout rate, an empirically supported adjustment commonly recommended in sample size estimation to maintain statistical power for both completer and intention-to-treat (ITT) analyses [32], a minimum of 229 patients per group was required to achieve 80% power with an α level of .05.

All analyses followed the ITT principle, which includes all randomized participants in the groups to which they were originally assigned. This approach maintains the integrity of randomization and provides a conservative estimate of treatment effectiveness in real-world clinical settings. To address missing data, multiple imputation by chained equations (MICEs) was performed under the assumption that the data were missing at random. Missing at random implies that the probability of missingness depends only on observed data and not on unobserved data. MICEs generate multiple plausible values for missing data based on observed variables, and we used 5 imputations (m=10) to account for uncertainty. The results from the imputed datasets were combined according to Rubin's rules. Normality of continuous variables was evaluated using the Kolmogorov-Smirnov test, a nonparametric test that compares the empirical distribution of the data with a theoretical normal distribution to assess deviations from normality. Data with normal distributions were presented as mean (SD) and compared using independent *t* tests. Nonnormally distributed data were presented as median (IQR) and analyzed using the Mann-Whitney test.

For the primary outcome, the 12-week relapse rate was compared between groups using the chi-square test. Time to relapse over 100 days was visualized using Kaplan-Meier curves and compared using the log-rank test. Cox proportional hazards regression was used to estimate hazard ratios (HRs) and 95% CIs, both unadjusted and adjusted for age and sex.

For continuous outcomes repeatedly measured over time (SCORAD, POEM, PP-NRS, IDQOL/CDLQI, and DFI), population-averaged generalized estimating equations were used with a Gaussian family, identity link, exchangeable correlation structure, and robust SEs to account for within-subject correlations. Each model was adjusted for baseline score, age, and sex. Missing data in these repeated measures were handled using MICEs (10 imputations), with final estimates combined using Rubin's rules.

A 2-sided *P* value of .04 was considered statistically significant.

Ethical Considerations

This study was reviewed and approved by the Medical Research Ethics Committee of the Children's Hospital of Chongqing Medical University (approval No. 2019-44-1). Written informed consent was obtained from all participants and their guardians before their involvement in the study. All collected data were deidentified to protect participant privacy and confidentiality. No compensation was provided to the participants.

Results

Participant Flow and Cohort Characteristics

The multicenter trial was conducted across 12 tertiary dermatology centers in China from July 2020 to March 2021, with a detailed CONSORT flow presented in Figure 2. Of 980 screened children with moderate-to-severe AD, 615 (62.8%) met inclusion criteria and were subsequently randomized into the digital education group (n=307, 49.9%) and the control group (n=308, 50.1%). Significant differences in attrition rates were observed at 12 weeks, with 20.2% (62/307) in the digital education group and 27.6% (85/308) in the control group (95% CI 0.66 - 0.98; χ^2_1=4.1; *P*=.04). Ultimately, 48.62% (299/615) completed the 52-week follow-up, comprising the per-protocol population: digital education (n=148, 24.1%) and control (n=151, 24.6%).

Baseline characteristics demonstrated adequate randomization balance (Table 1), with no significant differences observed in gender distribution (men: 52.1% vs 50.8%; *P*=.74), mean age (3.2, SD 1.8, vs 3.4, SD 1.7 y; *P*=.21), or disease severity measures, including SCORAD (49 [40, 59] vs 47.5 [37, 55]; *P*=.29), PP-NRS (6 [6, 8] vs 6 [5, 8]; *P*=.31) and POEM (15 [11, 20] vs 15 [11, 19]; *P*=.50). Similarly, no significant differences were noted in quality-of-life measures, including DFI (9 [4, 16] vs 10 [5,16]; *P*=.77) and QoL scores (13 [11, 15] vs 14 [11, 16]; *P*=.11). Quality of life was assessed using the CDLQI for children and IDQOL for infants.

XSL•FO
RenderX

Figure 2. CONSORT diagram of the randomization and follow-up of the study participants. AD: atopic dermatitis.

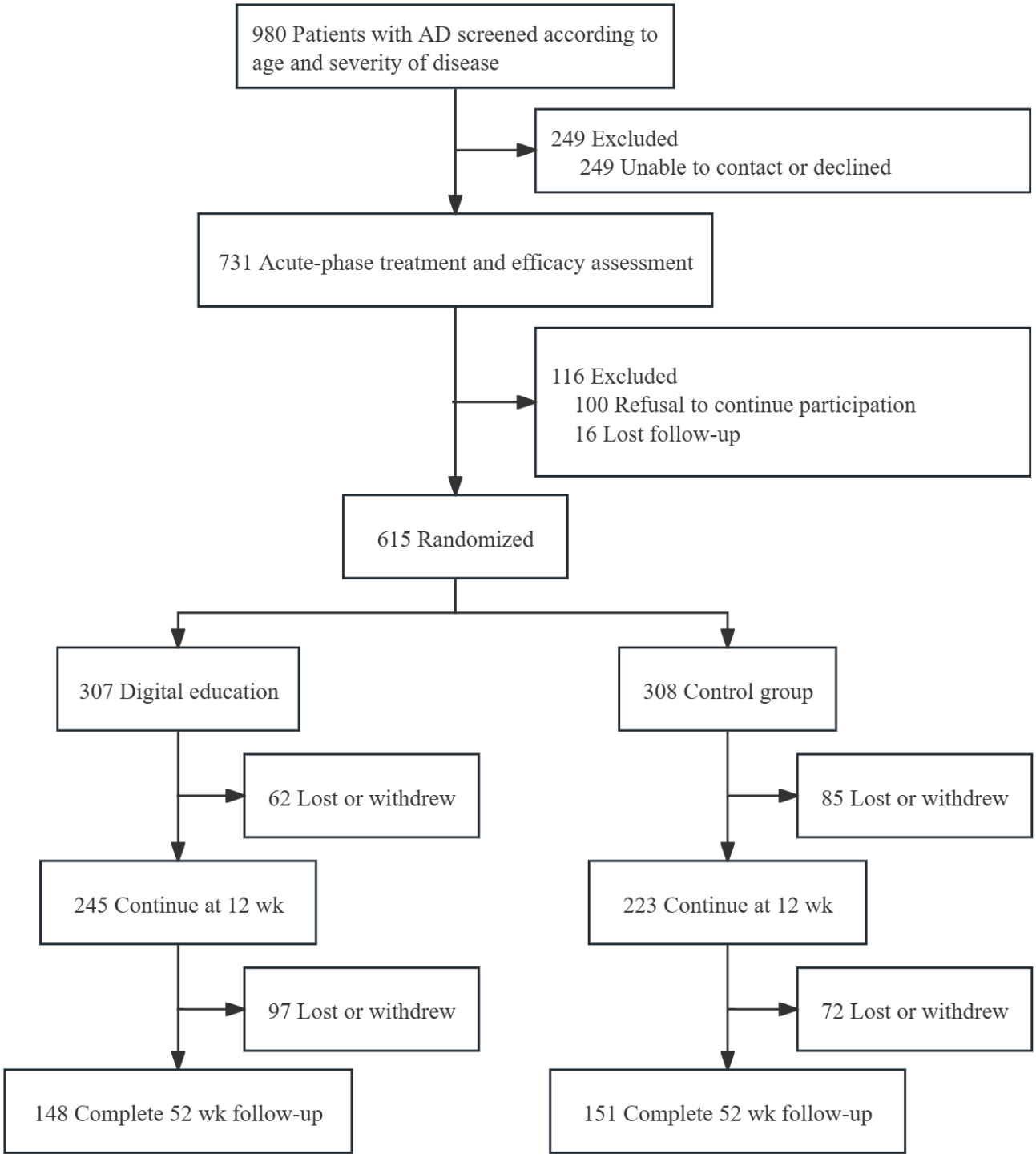

XSL•FO
RenderX

Table . The details of clinical characteristics of randomized patients. QoL was assessed using the Children's Dermatology Life Quality Index for children and the Infants' Dermatitis Quality of Life Index for infants.

Characteristic	Digital education (n=307)	Control (n=308)	*P* value
Gender, male (%)	52.1	50.8	.74
Age (y), mean (SD)	3.2 (1.8)	3.4 (1.7)	.21
SCORAD[a], median (Q1[b] ,Q3[c])	49 (40, 59)	47.5 (37, 55)	.29
PP-NRS[d], median (Q1, Q3)	6 (6, 8)	6 (5, 8)	.31
DFI[e], median (Q1, Q3)	9 (4, 16)	10 (5, 16)	.77
POEM[f], median (Q1, Q3)	15 (11, 20)	15 (11, 19)	.50
QoL[g], median (Q1, Q3)	13 (11, 15)	14 (11, 16)	.11

[a]SCORAD: Scoring Atopic Dermatitis.

[b]Q1: the first quartile.

[c]Q3: the third quartile.

[d]PP-NRS: Peak-Pruritus Numerical Rating Scale.

[e]DFI: Dermatitis Family Impact.

[f]POEM: Patient-Oriented Eczema Measure.

[g]QoL: quality of life.

Intervention Adherence and Participant Engagement

Intervention adherence, as measured by digital footprint tracking, demonstrated that 58.0% (178/307) of participants in the digital education group maintained regular engagement, averaging 1.64 (SD 0.38) sessions per week, and 26.7% (82/307) accessed the expedited physician consultation portals during disease flares. In the control arm, clinic attendance at the 52-week follow-up was 49.0 % (151/308). The 9.0% absolute difference in adherence between groups was statistically significant (Z=2.3, 95% CI 1.15%-16.85%; *P*=.03).

Primary Outcome: Relapse Prevention

The ITT analysis revealed a significant reduction in relapse rates in the digital education group at 12 weeks (16.6% [51/307] vs 24.0% [74/308]; relative risk [RR] 0.69, 95% CI 0.50 - 0.96; χ^2_1=5.1; *P*=.02). However, there was no statistically significant difference between the 2 groups at other time points: 4 weeks (8.79% vs 12.99%; RR 0.68, 95% CI 0.42 - 1.09; χ^2_1=2.8; *P*=.10), 8 weeks (15.64% vs 21.43%; RR 0.73, 95% CI 0.52 - 1.02; χ^2_1=3.4; *P*=.07), 24 weeks (21.82% vs 27.60%; RR 0.79, 95% CI 0.60 - 1.04; χ^2_1=2.8; *P*=.10), 36 weeks (23.13% vs 29.22%; RR 0.79, 95% CI 0.60 - 1.03; χ^2_1=3.0; *P*=.09), and 52 weeks (24.76% vs 31.17%; RR 0.79, 95% CI 0.62 - 1.03, χ^2_1=3.1; *P*=.08).

Kaplan-Meier analysis (100-d follow-up) further supported the 12-week finding, showing a significant difference in relapse risk between groups (HR] 0.688, 95% CI 0.490 - 0.966; χ^2_1=4.7; *P*=.03; Figure 3). In a Cox proportional hazards model adjusting for age and sex, the HR remained significant (adjusted HR 0.665, 95% CI 0.469 - 0.943; *P*=.02).

XSL•FO
RenderX

Figure 3. Kaplan-Meier estimates of relapse. Hundred-day relapse-free survival curves in the intention-to-treat set analysis.

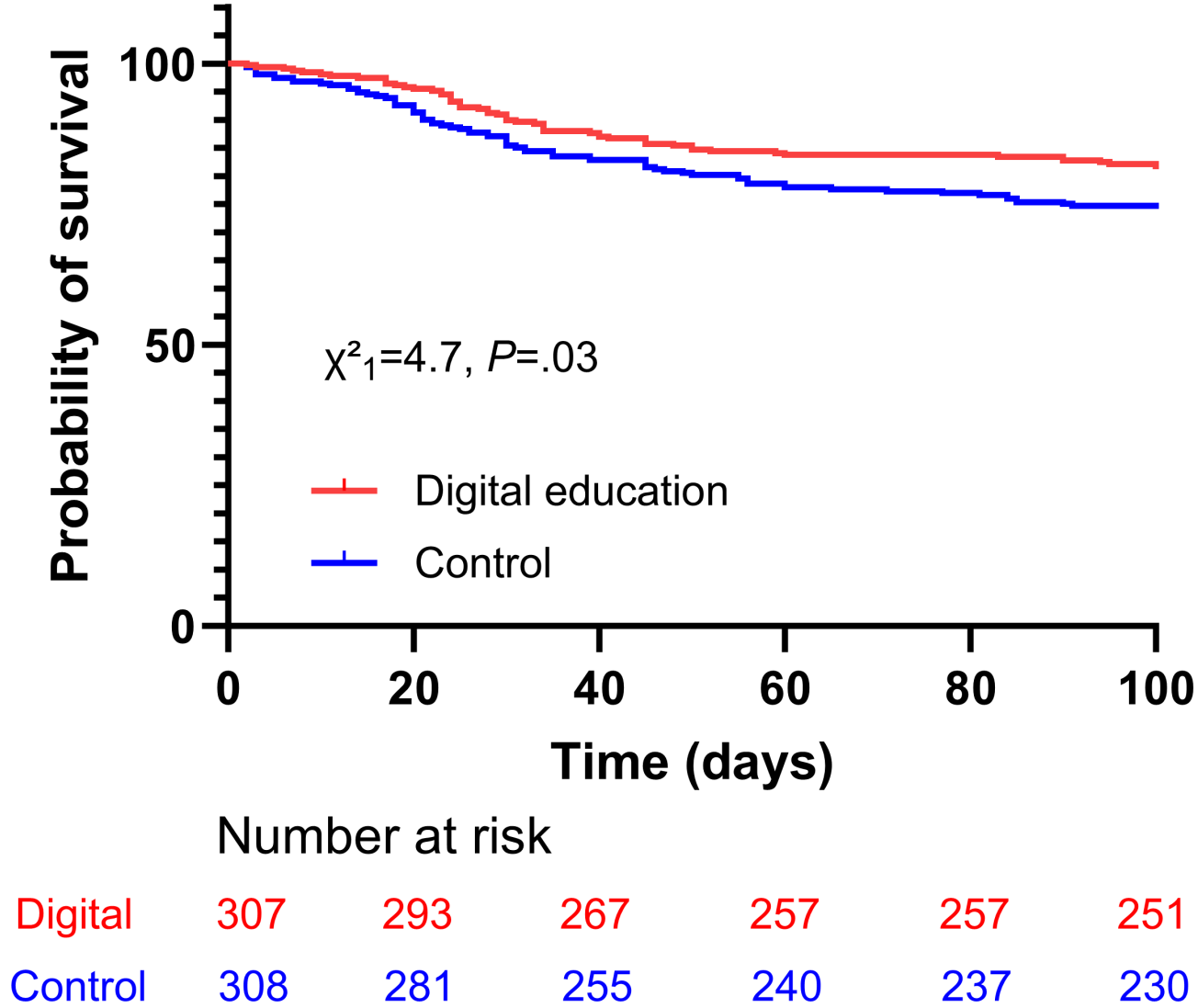

Secondary Outcomes: Disease Severity Trajectories

Changes in disease severity scores (including SCORAD, PP-NRS, and POEM) and QoL scores (IDQOL/CDLQI and DFI) from week 2 (the end of the acute treatment phase) were compared between the 2 groups at weeks 4, 8, 12, 24, 36, and 52. Using population-averaged GEE models adjusted for baseline score, age, and sex, the results demonstrated no statistically significant differences between the groups in the magnitude of score changes at any of the predefined follow-up time points (Multimedia Appendix 10).

Discussion

Principal Findings

This multicenter RCT demonstrated that a structured, smartphone-based multimodal digital education program for caregivers, when integrated with standard care, significantly reduced the substantial proportion of early relapses in children aged 0 to 6 years with moderate-to-severe AD, compared to conventional outpatient consultation alone. This early benefit was further supported by Kaplan-Meier survival analysis showing a clearer separation of relapse-free curves during the early follow-up period. Consistent with the study's primary objective, these results confirm that the digital educational intervention effectively achieved short-term relapse prevention. However, differences in relapse rates appeared to diminish over time, and no notable differences were observed in secondary end points assessing disease severity or quality of life across the 52-week follow-up.

The observed reduction in early relapse aligns with the core principle of AD management, which emphasizes the critical role of effective patient and caregiver education in improving clinical outcomes [8-12]. The success of our digital program likely stems from its multimodal design, which combined frequent, brief multimedia modules with a "Real-Time Clinician Access" pathway ("Green Channel"). This combination probably enhanced caregivers' ability to detect early signs of worsening and initiate prompt interventions, thereby preventing progression to full-blown relapse, even in the absence of measurable changes in objective severity indices. The platform's adherence rate suggested better engagement compared with what is often observed in traditional formats [33], indicating improved feasibility and user engagement.

XSL•FO
RenderX

Our findings are consistent with prior evidence demonstrating that structured education improves short-term management and adherence in pediatric AD [16-23,34-37]. However, they also highlight a common challenge: achieving long-term disease modulation with brief interventions is difficult. The diminishing effect on relapse prevention after the initial 12 weeks may be attributed to several factors. First, the structured program builds knowledge and habits, but its impact may wane without ongoing reinforcement [38]. Second, the natural relapsing-remitting course of AD may eventually overshadow the effects of a time-limited intervention. Furthermore, despite good initial adherence, participant engagement with digital health interventions often declines over time [39].

A critical component contributing to the early success was the integrated "Real-Time Clinician Access," used by a notable portion of participants during flares or uncertainty. This feature provided timely guidance, effectively replicating the benefit of rapid in-person intervention during exacerbations without the associated logistical burdens. This mechanism, improved early recognition and intervention facilitated by accessible education and on-demand clinical support, provides a plausible explanation for the significant reduction in early relapse rates, underscoring that the primary value of the intervention lies in optimizing the behavioral and self-management aspects of AD care.

The absence of significant differences in disease severity and QoL scores suggests that while the intervention improved caregivers' short-term flare management, it did not fundamentally alter the underlying core inflammatory process within the assessed follow-up period. This distinction has practical implications: digital education may be most valuable as a tool to reduce short-term exacerbations and health care utilization, complementing rather than replacing ongoing medical and pharmacological strategies needed to modify underlying disease activity.

Traditional educational formats (face-to-face counseling [34-37], workshops [16-19], and eczema schools [20-23] are effective but face significant scalability challenges due to their resource-intensive nature, geographical barriers, and time constraints for families and clinicians [35]. Passive educational materials, such as leaflets and videos, are limited by a lack of interactivity and personalization [26,36,37]. Our study addresses these limitations by leveraging widespread smartphone access to deliver a scalable, multimodal intervention. The platform's use of brief, engaging modules (animated stories, videos, and illustrated texts) delivered 3 times weekly made essential educational topics accessible to caregivers with varying literacy levels and busy schedules.

Limitations

Several limitations must be acknowledged. First, smartphone proficiency was an enrollment requirement, potentially excluding digitally underserved groups (eg, from rural or socioeconomically disadvantaged backgrounds) and limiting the generalizability of our findings to urban or digitally connected families. Future implementations should consider hybrid delivery models combining digital modules with printed materials or in-person sessions to enhance inclusivity. Second, understanding of the educational materials was not formally assessed, which may influence outcomes. Third, although early retention differed between groups, a considerable number of participants did not complete long-term follow-up, potentially affecting the accuracy of long-term outcome estimates. Fourth, although outcome assessors were blinded, caregivers and clinicians were not, creating potential for performance or reporting bias. Fifth, caregiver demographic variables (eg, age, education, and occupation) were not collected, limiting our ability to explore effect modification by caregiver characteristics. While the exclusion of 'left-behind children' and the predominance of parent caregivers likely reduced heterogeneity in living arrangements and direct caregiving availability, unmeasured socioeconomic or educational differences could still influence the uptake and use of digital education. Finally, while the control group also had access to the platform for noneducational purposes (eg, data entry, random allocation, and scheduling visits), they did not receive the structured educational modules or 'Green Channel' clinician access. Therefore, the observed effect likely reflects the combined impact of the educational content and the interactive digital delivery modality, rather than platform access alone. Future trials should include comparative arms using alternative digital formats (eg, app-based platforms and interactive chatbots) to disentangle which specific features most strongly influence adherence and relapse outcomes.

Conclusions and Broader Implications

In conclusion, this large multicenter RCT demonstrates that a structured, smartphone-based digital education program for caregivers, featuring interactive modules and real-time clinician support, can significantly reduce early relapse rates in young children with moderate-to-severe AD. This finding indicates that scalable digital education can strengthen short-term relapse prevention by optimizing caregiver empowerment and flare management, a strategy particularly relevant in regions with high smartphone penetration.

The broader implications of our study are 3-fold. First, given the high burden of AD and the scalability of digital tools, such interventions hold promise for improving access to quality education and potentially reducing health care disparities. Second, the lower dropout rate in the intervention group suggests that well-designed digital platforms may improve retention in long-term pediatric dermatology research and care. Finally, the transient nature of the benefit underscores that digital education should be viewed as a complementary adjunct to, not a replacement for, ongoing medical management.

Future research should prioritize (1) developing hybrid models that integrate digital education with targeted in-person support to enhance emotional connection and engagement; (2) designing strategies for sustained engagement, such as "booster" modules and adaptive content; (3) ensuring digital accessibility for underserved populations to address health equity; (4) conducting rigorous economic evaluations; and (5) conducting head-to-head comparisons of different digital features and mechanistic studies to identify the active ingredients of digital education.

XSL•FO
RenderX

Acknowledgments

The authors sincerely thank Mr. Lei Yan and Ms. Ping Yang for their kind support and assistance in the statistical analysis of the data. The authors declare the use of generative AI in the research and writing process. According to the GAIDeT taxonomy (2025), translation was delegated to GAI tools under full human supervision. The GAI tool used was ChatGPT 5.

Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by Huan Yang.

Funding

This study was funded by Beijing Medical Award Foundation (2019‐1115), the Joint Foundation of Chongqing Municipal Health Commission and Municipal Science and Technology Bureau (2021MSXM203), CQMU Program for Youth Innovation in Future Medicine (W0177), and Children's Hospital Affiliated to Chongqing Medical University (RC05036). The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Data Availability

The anonymized datasets generated and analyzed during this study are not publicly available due to participant privacy and institutional data protection policy but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: HY, HW, LM, XL
Methodology: HY, HW, LM, XL
Investigation: HY, HS, LW, PL, YL, QL, XH, JT, JC, HQ, JP Chen, XD, PW, LD, ZL, WL, YL, LL, SY, YG, LY, XS
Data Curation: HY, HS, LW, PL
Formal Analysis: HY, HS, LW, PL
Resources: HW, XL, HY
Software: HY
Validation: HY, HS, LW, PL
Funding Acquisition: HW, XL, HY
Project Administration: HW, LM, XL
Supervision: HW, LM, XL
Writing-Original Draft: HY
Writing-Review & Editing: HY, HS, LW
Visualization: HY, HS, LW, PL
All authors reviewed and approved the final manuscript and agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of participating hospitals.
[PDF File, 66 KB - jmir_v28i1e79559_app1.pdf]

Multimedia Appendix 2
Atopic dermatitis awareness material.
[PDF File, 78 KB - jmir_v28i1e79559_app2.pdf]

Multimedia Appendix 3
Procedural overview illustrating the standardized workflow for patient enrollment and identity verification within the "Skin Care E-Station" platform, with representative interface screenshots supporting each operational step.
[PDF File, 294 KB - jmir_v28i1e79559_app3.pdf]

Multimedia Appendix 4
Education action plan.
[PDF File, 4038 KB - jmir_v28i1e79559_app4.pdf]

Multimedia Appendix 5

XSL•FO
RenderX

Sample illustrated educational text.
[PNG File, 537 KB - jmir_v28i1e79559_app5.png]

Multimedia Appendix 6
Sample caregiver-facing informational images.
[PNG File, 294 KB - jmir_v28i1e79559_app6.png]

Multimedia Appendix 7
Sample video.
[MP4 File, 15585 KB - jmir_v28i1e79559_app7.mp4]

Multimedia Appendix 8
Animated stories.
[PDF File, 121 KB - jmir_v28i1e79559_app8.pdf]

Multimedia Appendix 9
SCORAD self-assessment.
[PDF File, 121 KB - jmir_v28i1e79559_app9.pdf]

Multimedia Appendix 10
Comparison of changes in disease severity and quality of life scores between 2 groups over time.
[DOCX File, 16 KB - jmir_v28i1e79559_app10.docx]

Checklist 1
CONSORT checklist.
[PDF File, 1121 KB - jmir_v28i1e79559_app11.pdf]

References

1. Langan SM, Irvine AD, Weidinger S. Atopic dermatitis. Lancet 2020 Aug 1;396(10247):345-360. [doi: 10.1016/S0140-6736(20)31286-1] [Medline: 32738956]
2. Nutten S. Atopic dermatitis: global epidemiology and risk factors. Ann Nutr Metab 2015;66 Suppl 1:8-16. [doi: 10.1159/000370220] [Medline: 25925336]
3. Guo Y, Zhang H, Liu Q, et al. Phenotypic analysis of atopic dermatitis in children aged 1-12 months: elaboration of novel diagnostic criteria for infants in China and estimation of prevalence. J Eur Acad Dermatol Venereol 2019 Aug;33(8):1569-1576. [doi: 10.1111/jdv.15618] [Medline: 30989708]
4. Guo Y, Li P, Tang J, et al. Prevalence of atopic dermatitis in Chinese children aged 1–7 ys. Sci Rep 2016;6(1):29751. [doi: 10.1038/srep29751]
5. Silverberg JI. Public health burden and epidemiology of atopic dermatitis. Dermatol Clin 2017 Jul;35(3):283-289. [doi: 10.1016/j.det.2017.02.002] [Medline: 28577797]
6. Nicolescu AC, Strilciuc S, Lăpădatu R, Grad DA, Vlădescu C, Olteanu R. The economic burden of atopic dermatitis in Romania: a broad perspective. Front Public Health 2024;12:1531042. [doi: 10.3389/fpubh.2024.1531042] [Medline: 39850862]
7. Girolomoni G, Busà VM. Flare management in atopic dermatitis: from definition to treatment. Ther Adv Chronic Dis 2022 Jan;13. [doi: 10.1177/20406223211066728]
8. Lapeere H, Speeckaert R, Baeck M, et al. Belgian atopic dermatitis guidelines. Acta Clin Belg 2024 Feb;79(1):62-74. [doi: 10.1080/17843286.2023.2285576] [Medline: 37997950]
9. Wollenberg A, Kinberger M, Arents B, et al. European guideline (EuroGuiDerm) on atopic eczema - part II: non-systemic treatments and treatment recommendations for special AE patient populations. J Eur Acad Dermatol Venereol 2022 Nov;36(11):1904-1926. [doi: 10.1111/jdv.18429] [Medline: 36056736]
10. Saeki H, Ohya Y, Furuta J, et al. English version of clinical practice guidelines for the management of atopic dermatitis 2021. J Dermatol 2022 Oct;49(10):e315-e375. [doi: 10.1111/1346-8138.16527] [Medline: 35996152]
11. Panel A, Chu DK, Schneider L, et al. Atopic dermatitis (eczema) guidelines: 2023 American Academy of Allergy, Asthma and Immunology/American College of Allergy, Asthma and Immunology Joint Task Force on Practice Parameters GRADE- and Institute of Medicine-based recommendations. Ann Allergy Asthma Immunol Mar 2024;132(3):274-312. [doi: 10.1016/j.anai.2023.11.009]
12. Xu Y, Zhi-Qiang S, Wei L, et al. Guidelines for diagnosis and treatment of atopic dermatitis in China. Int J Derm Venereol 2021;4(1):1-9. [doi: 10.1097/JD9.0000000000000143]

XSL•FO
RenderX

13. Greenwell K, Sivyer K, Howells L, et al. “Eczema shouldn’t control you; you should control eczema”: qualitative process evaluation of online behavioural interventions to support young people and parents/carers of children with eczema. Br J Dermatol 2023 Mar 30;188(4):506-513. [doi: 10.1093/bjd/ljac115] [Medline: 36745562]
14. Andrade LF, Abdi P, Mashoudy KD, et al. Effectiveness of atopic dermatitis patient education programs - a systematic review and meta-analysis. Arch Dermatol Res 2024 Apr 25;316(5):135. [doi: 10.1007/s00403-024-02871-y] [Medline: 38662127]
15. Gabryszewski SJ, Hill DA. One march, many paths: insights into allergic march trajectories. Ann Allergy Asthma Immunol 2021 Sep;127(3):293-300. [doi: 10.1016/j.anai.2021.04.036] [Medline: 33971364]
16. Moore EJ, Williams A, Manias E, Varigos G, Donath S. Eczema workshops reduce severity of childhood atopic eczema. Australas J Dermatol 2009 May;50(2):100-106. [doi: 10.1111/j.1440-0960.2009.00515.x] [Medline: 19397561]
17. Heratizadeh A, Werfel T, Wollenberg A, et al. Effects of structured patient education in adults with atopic dermatitis: multicenter randomized controlled trial. J Allergy Clin Immunol 2017 Sep;140(3):845-853. [doi: 10.1016/j.jaci.2017.01.029] [Medline: 28242304]
18. Kupfer J, Gieler U, Diepgen TL, et al. Structured education program improves the coping with atopic dermatitis in children and their parents-a multicenter, randomized controlled trial. J Psychosom Res 2010 Apr;68(4):353-358. [doi: 10.1016/j.jpsychores.2009.04.014] [Medline: 20307702]
19. Liang Y, Tian J, Shen CP, et al. Therapeutic patient education in children with moderate to severe atopic dermatitis: a multicenter randomized controlled trial in China. Pediatr Dermatol 2018 Jan;35(1):70-75. [doi: 10.1111/pde.13362] [Medline: 29243849]
20. Grossman SK, Schut C, Kupfer J, Valdes-Rodriguez R, Gieler U, Yosipovitch G. Experiences with the first eczema school in the United States. Clin Dermatol 2018;36(5):662-667. [doi: 10.1016/j.clindermatol.2018.05.006] [Medline: 30217280]
21. Staab D, Diepgen TL, Fartasch M, et al. Age related, structured educational programmes for the management of atopic dermatitis in children and adolescents: multicentre, randomised controlled trial. BMJ 2006 Apr 22;332(7547):933-938. [doi: 10.1136/bmj.332.7547.933] [Medline: 16627509]
22. Broberg A, Kalimo K, Lindblad B, Swanbeck G. Parental education in the treatment of childhood atopic eczema. Acta Derm Venereol 1990;70(6):495-499. [Medline: 1981422]
23. Futamura M, Ito K, Otsuji K, et al. Effects of “Skin Care School,” a parental education program on childhood atopic dermatitis conducted during short hospitalization stays. Arerugi 2009 Dec;58(12):1610-1618. [Medline: 20220302]
24. LeBovidge JS, Timmons K, Delano S, et al. Improving patient education for atopic dermatitis: a randomized controlled trial of a caregiver handbook. Pediatr Dermatol 2021 Mar;38(2):396-404. [doi: 10.1111/pde.14519] [Medline: 33486817]
25. Park GY, Park HS, Cho S, Yoon HS. The effectiveness of tailored education on the usage of moisturizers in atopic dermatitis: a pilot study. Ann Dermatol 2017;29(3):360. [doi: 10.5021/ad.2017.29.3.360]
26. Armstrong AW, Kim RH, Idriss NZ, Larsen LN, Lio PA. Online video improves clinical outcomes in adults with atopic dermatitis: a randomized controlled trial. J Am Acad Dermatol 2011 Mar;64(3):502-507. [doi: 10.1016/j.jaad.2010.01.051] [Medline: 21236514]
27. Tier HL, Balogh EA, Bashyam AM, et al. Tolerability of and adherence to topical treatments in atopic dermatitis: a narrative review. Dermatol Ther (Heidelb) 2021 Apr;11(2):415-431. [doi: 10.1007/s13555-021-00500-4] [Medline: 33599887]
28. Patel N, Feldman SR. Adherence in atopic dermatitis. Adv Exp Med Biol 2017;1027:139-159. [doi: 10.1007/978-3-319-64804-0_12] [Medline: 29063437]
29. Wilken B, Zaman M, Asai Y. Patient education in atopic dermatitis: a scoping review. Allergy Asthma Clin Immunol 2023 Oct 13;19(1):89. [doi: 10.1186/s13223-023-00844-w] [Medline: 37833754]
30. Golinelli D, Boetto E, Carullo G, Nuzzolese AG, Landini MP, Fantini MP. Adoption of digital technologies in health care during the COVID-19 pandemic: systematic review of early scientific literature. J Med Internet Res 2020 Nov 6;22(11):e22280. [doi: 10.2196/22280] [Medline: 33079693]
31. Yuan L, Lingling L, Shan W, et al. Efficacy and safety of 0.03% tacrolimus ointment in the long-term intermittent maintenance treatment of atopic dermatitis in children: a multicenter randomized controlled clinical trial. Chin J Dermatol 2019;52(8):519-524. [doi: 10.3760/cma.j.issn.0412-4030.2019.08.001] [Medline: 30656725]
32. Overall JE, Shobaki G, Shivakumar C, Steele J. Adjusting sample size for anticipated dropouts in clinical trials. Psychopharmacol Bull 1998;34(1):25-33. [Medline: 9564195]
33. Bass AM, Anderson KL, Feldman SR. Interventions to increase treatment adherence in pediatric atopic dermatitis: a systematic review. J Clin Med 2015 Jan 27;4(2):231-242. [doi: 10.3390/jcm4020231] [Medline: 26239125]
34. Barbarot S, Aubert H, Vibet MA, et al. Effectiveness of a nurse-led one-to-one education programme in addition to standard care in children with atopic dermatitis: a multicentre randomized control trial. Br J Dermatol 2024 Jul 16;191(2):177-186. [doi: 10.1093/bjd/ljae111] [Medline: 38863109]
35. Cheong JYV, Hie SL, Koh EW, de Souza NNA, Koh MJA. Impact of pharmacists’ counseling on caregiver’s knowledge in the management of pediatric atopic dermatitis. Pediatr Dermatol 2019 Jan;36(1):105-109. [doi: 10.1111/pde.13708] [Medline: 30408232]

XSL•FO
RenderX

36. Chinn DJ, Poyner T, Sibley G. Randomized controlled trial of a single dermatology nurse consultation in primary care on the quality of life of children with atopic eczema. Br J Dermatol 2002 Mar;146(3):432-439. [doi: 10.1046/j.1365-2133.2002.04603.x] [Medline: 11952543]
37. Rolinck-Werninghaus C, Trentmann M, Reich A, Lehmann C, Staab D. Improved management of childhood atopic dermatitis after individually tailored nurse consultations: a pilot study. Pediatr Allergy Immunol 2015 Dec;26(8):805-810. [doi: 10.1111/pai.12338] [Medline: 25643831]
38. Boswell JF, Kraus DR, Miller SD, Lambert MJ. Implementing routine outcome monitoring in clinical practice: benefits, challenges, and solutions. Psychother Res 2015;25(1):6-19. [doi: 10.1080/10503307.2013.817696] [Medline: 23885809]
39. Cross SP, Alvarez-Jimenez M. The digital cumulative complexity model: a framework for improving engagement in digital mental health interventions. Front Psychiatry 2024;15:1382726. [doi: 10.3389/fpsyt.2024.1382726] [Medline: 39290300]

Abbreviations

AD: atopic dermatitis
CDLQI: Children's Dermatology Life Quality Index
CONSORT: Consolidated Standards of Reporting Trials
CRO: Clinical Research Organization
DFI: Dermatitis Family Impact
EAP: education action plan
HR: hazard ratio
IDQOL: Infant's Dermatitis Quality of Life Index
IGA: Investigator's Global Assessment
ITT: intention-to-treat
MICE: multiple imputation by chained equation
POEM: Patient-Oriented Eczema Measure
PP-NRS: Peak Pruritus Numerical Rating Scale
QoL: quality of life
RCT: randomized controlled trial
RR: relative risk
SCORAD: Scoring Atopic Dermatitis

Edited by N Cahill; submitted 24.Jun.2025; peer-reviewed by X Liang, Y Zhang; revised version received 14.Nov.2025; accepted 18.Nov.2025; published 07.Jan.2026.

Please cite as:
Yang H, Shu H, Wang LH, Li P, Li YL, Li QF, Han XP, Tian J, Chang J, Qian H, Chen JP, Ding XQ, Wu PQ, Dou LM, Luo Z, Li W, Lin YY, Li L, Yue SZ, Gu Y, Yang L, Sun XH, Luo XY, Ma L, Wang H
Smartphone-Based Digital Eczema Education Program for Atopic Dermatitis in Children Aged 0 to 6 Years: Multicenter, Randomized, Parallel Controlled Clinical Study
J Med Internet Res 2026;28:e79559
URL: https://www.jmir.org/2026/1/e79559
doi:10.2196/79559

© Huan Yang, Hong Shu, Liu-hui Wang, Ping Li, Yun-ling Li, Qin-feng Li, Xiu-ping Han, Jing Tian, Jing Chang, Hua Qian, Jing-ping Chen, Xin-qiang Ding, Pan-qian Wu, Li-min Dou, Zhen Luo, Wei Li, Yang-yang Lin, Lin Li, Shu-zhen Yue, Yang Gu, Li Yang, Xiao-hong Sun, Xiao-yan Luo, Lin Ma, Hua Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Bayesian-Based Pharmacokinetic Framework Integrated with Therapeutic Drug Monitoring for Assessing Adherence to Antiseizure Medications: A Clinical Trial Simulation Study

Xiao-Qin Liu[1], PhD; Zi-Ran Li[2], PhD; Wei-Wei Lin[3], PhD; Juan Wang[1], BSc; Fu-Qing Gu[1], BSc; Jun-Jie Ding[4], PhD; Zheng Jiao[1], PhD

[1]Department of Pharmacy, Shanghai Chest Hospital, Shanghai Jiao Tong University School of Medicine, 241 West Huaihai Road, Shanghai, China
[2]Department of Bioengineering & Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, United States
[3]Department of Pharmacy, The First Affiliated Hospital of Fujian Medical University, Fuzhou, Fujian, China
[4]Center for Tropical Disease and Global Health, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, United Kingdom

Corresponding Author:
Zheng Jiao, PhD
Department of Pharmacy, Shanghai Chest Hospital, Shanghai Jiao Tong University School of Medicine, 241 West Huaihai Road, Shanghai, China

Abstract

Background: Adherence to antiseizure medications (ASMs) is a cornerstone of effective epilepsy management. However, current consensus guidelines for assessing medication adherence via therapeutic drug monitoring (TDM) may neglect individual patient characteristics, thereby compromising the accuracy of adherence assessments.

Objective: This study proposed an innovative Bayesian–based pharmacokinetic (PK) framework integrated with TDM data to address the above limitations, with a focus on 14 widely prescribed ASMs, including brivaracetam, carbamazepine, clobazam, eslicarbazepine acetate, lacosamide, lamotrigine, levetiracetam, oxcarbazepine, perampanel, phenobarbital, topiramate, valproic acid, vigabatrin, and zonisamide.

Methods: Comprehensive clinical trial simulations were conducted to investigate the PK of ASMs in patients with epilepsy under conditions of full adherence and various nonadherent dosing behaviors, including omission of the last dose and consecutive missed doses. Bayesian posterior probabilities of these dosing behaviors were derived by integrating validated population PK models, individual patient demographics (eg, age, weight, creatinine clearance), dosing history, prior adherence probabilities and TDM measurements. Additionally, the influence of covariates on assessment outcomes was systematically evaluated.

Results: The Bayesian-based PK approach demonstrated robust discriminative ability. Under idealized simulation conditions with minimized variabilities, the approach achieved accurate retrodiction of the last 1 or 2 doses across all 14 ASMs and partial retrodiction of extended nonadherence trajectories for 6 ASMs. Concentration thresholds for adherence classification varied significantly across drugs and are influenced by patient-specific factors, comedications, formulation, sampling time, and prior probability. To translate these insights into practice, an adaptable web-based dashboard was developed using the *shiny* package in R software to enable precise and real-time assessments of medication adherence.

Conclusions: This study establishes a Bayesian-based PK approach to enhance the assessment of ASMs adherence. This approach facilitates a paradigm shift from population-based management to patient-specific adherence profiling, offering a practical methodology for the precise evaluation of medication-taking behaviors.

(J Med Internet Res 2026;28:e77917) doi:10.2196/77917

KEYWORDS

antiseizure medications; medication adherence; therapeutic drug monitoring; Bayesian theory; population pharmacokinetics

Introduction

Epilepsy is the second most common neurological disease globally. Antiseizure medications (ASMs) represent the cornerstone of treatment for epilepsy [1,2], with long-term medication adherence being critical to achieving successful therapeutic outcomes [3]. However, adherence to ASMs among people with epilepsy is often suboptimal [4-6], which is strongly associated with a range of adverse clinical outcomes, including increased mortality, heightened morbidity, greater health care utilization, and substantial economic burden [7,8]. Therefore, when evaluating treatment failures, it is imperative for health care providers to comprehensively assess patients' adherence,

XSL•FO
RenderX

to identify underlying issues and provide tailored support to improve seizure control and treatment efficacy.

In clinical practice, self-reported adherence is inherently subjective and prone to bias [9], in contrast to therapeutic drug monitoring (TDM), which offers a more objective measure of recent medication-taking behaviors [10,11]. Accurate TDM interpretation is straightforward in some situations, such as a notably low drug concentration suggesting nonadherence. However, it becomes much more complex in other cases due to various intrinsic and extrinsic confounders that affect drug concentrations, including organ function, drug-drug interactions and dosing intervals.

The Consensus Guidelines for TDM in Neuropsycho-pharmacology: 2017 update [12] (hereafter referred to as the 2017 Guidelines) provide reference ranges for commonly used ASMs, which are specifically tailored for steady-state trough concentrations (C_0) in adult patients undergoing monotherapy. The reference ranges for each ASM were determined by multiplying the daily dose by dose-related concentration factors, and then could be used to help identify nonadherence [12]. However, the reference ranges were based on average pharmacokinetic (PK) parameters from an adult population, and do not account for key subpopulations, such as pediatric, geriatric and pregnant patients, who exhibit clinically significant PK differences [13-15].

PK modeling and simulation approaches have been successfully used to evaluate the impact of medication nonadherence [16,17], and to design remedial dosing strategies for missed or delayed doses [16,18]. When combined with Bayesian principle, this methodology offers a powerful framework for integrating individual TDM data with population PK models [19,20]. By leveraging this approach, it becomes possible to infer posterior probabilities of different dosing patterns, thereby enabling a more refined and quantitative assessment of medication-taking behavior.

In light of the above, this study aims to characterize medication adherence patterns to ASMs using TDM measurements and a Bayesian-based PK approach. Additionally, a user-friendly dashboard is developed to offer health care providers an intuitive, practical tool for assessing individual adherence levels, thereby optimizing ASM therapy and improving the treatment outcomes of ASMs.

Method

Ethical Considerations

As this study exclusively used computational modeling and simulation techniques without involving direct human subject participation or personal data collection, it is exempt from institutional review board approval requirements in accordance with international ethical guidelines.

Rationale

When patients fully adhere to their medication regimens, the drug concentration fluctuates in a predictable manner. However, if patients miss any of their doses, the drug concentration will gradually decline to a suboptimal level, which may ultimately result in treatment failure. The differences in the probability distribution of drug concentration provides a valuable reference for differentiating between adherence to the prescribed medication and nonadherence.

In this study, the Bayesian-based PK approach, calculating the posterior probability of special dosing events, was used to assess medication adherence. The principle of the Bayesian approach is as follows [19]: given the probability of the occurrence of a specific scenario (ie, the prior probability, P(ωj)) and the probability of a particular drug concentration at a given scenario ωj (ie, conditional probability, P(C|ωj)) , the probability of the scenario at a given drug concentration (ie, posterior probability,P(ωjC) can be estimated, as presented in Equation 1.

$$(1)P(\omega j \mid C)=P(\omega j)\times P(C \mid \omega j)P(C)$$

Where P(C) is the full probability and could be calculated with Equation 2.

$$(2)P(C)=\sum jP(\omega j)\times P(C \mid \omega j)$$

The prior probability Pωj refers to the pre-existing or baseline probability estimate of a patient's likelihood to adhere to a prescribed medication regimen before any new, specific data related to that individual patient's adherence behavior in the current treatment course is considered. The conditional probability P(C|ωj) is calculated using Monte Carlo simulations based on population PK. Once these probabilities have been obtained, the posterior probability P(ωjC of each individual scenario is calculated using Equations 1; 2. The scenario with the highest posterior probability is considered the most likely to occur, while the one with the lowest posterior probability is deemed the least probable.

In our study, the dosing event scenario is defined by whether patients adhere to or miss their scheduled doses. There are 2^n possible scenarios when considering the last *n* dosing events prior to sampling. For instance, as depicted in Figure 1A, there are two scenarios (ω0 and ω1) when considering the last dosing instance. This can be expanded to four scenarios (ω00, ω01, ω10, and ω11) when the last two dosing instances are considered (Figure 1B), and eight scenarios (ω000, ω010, ω100, ω001, ω110, ω011, ω101, ω111) when the last three dosing events are taken into account (Figure 1C).

XSL•FO
RenderX

Figure 1. The dosing scenarios when the most recent one (A), two (B) or three (C) dosing events are considered. ω: medication-taking behavior, where the first, second and third digit after ω indicates the most recent one, two and three medication-taking events prior to sampling, respectively, where 1 indicates dose taken and 0 indicates dose missed.

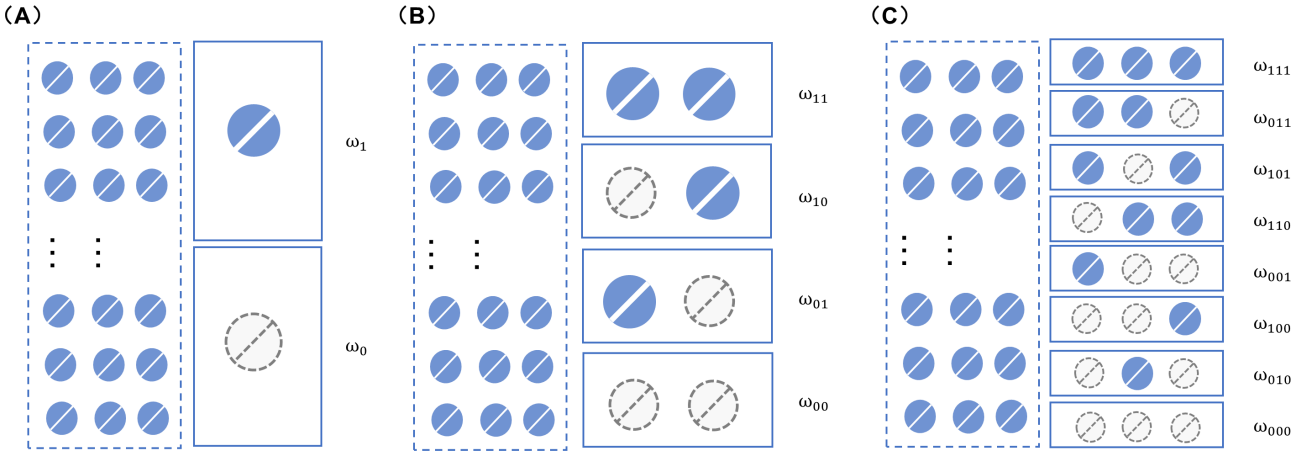

The workflow of adherence assessment is graphically represented with Figure 2, using the example of a 70 kg adult patient receiving oxcarbazepine 300 mg every 12 hours (q12h) and reached steady state. C_0 of oxcarbazepine was used to infer the patient's dosing behavior over the last two dosing intervals. When the C_0 approaches zero, the posterior probability of at least one missed dose is high. As C_0 increases, this probability decreases, while the posterior probability of complete adherence rises correspondingly, eventually approaching 100%.

XSL•FO
RenderX

Figure 2. The workflow of adherence assessment by therapeutic drug monitoring with Bayesian-based pharmacokinetic approach.

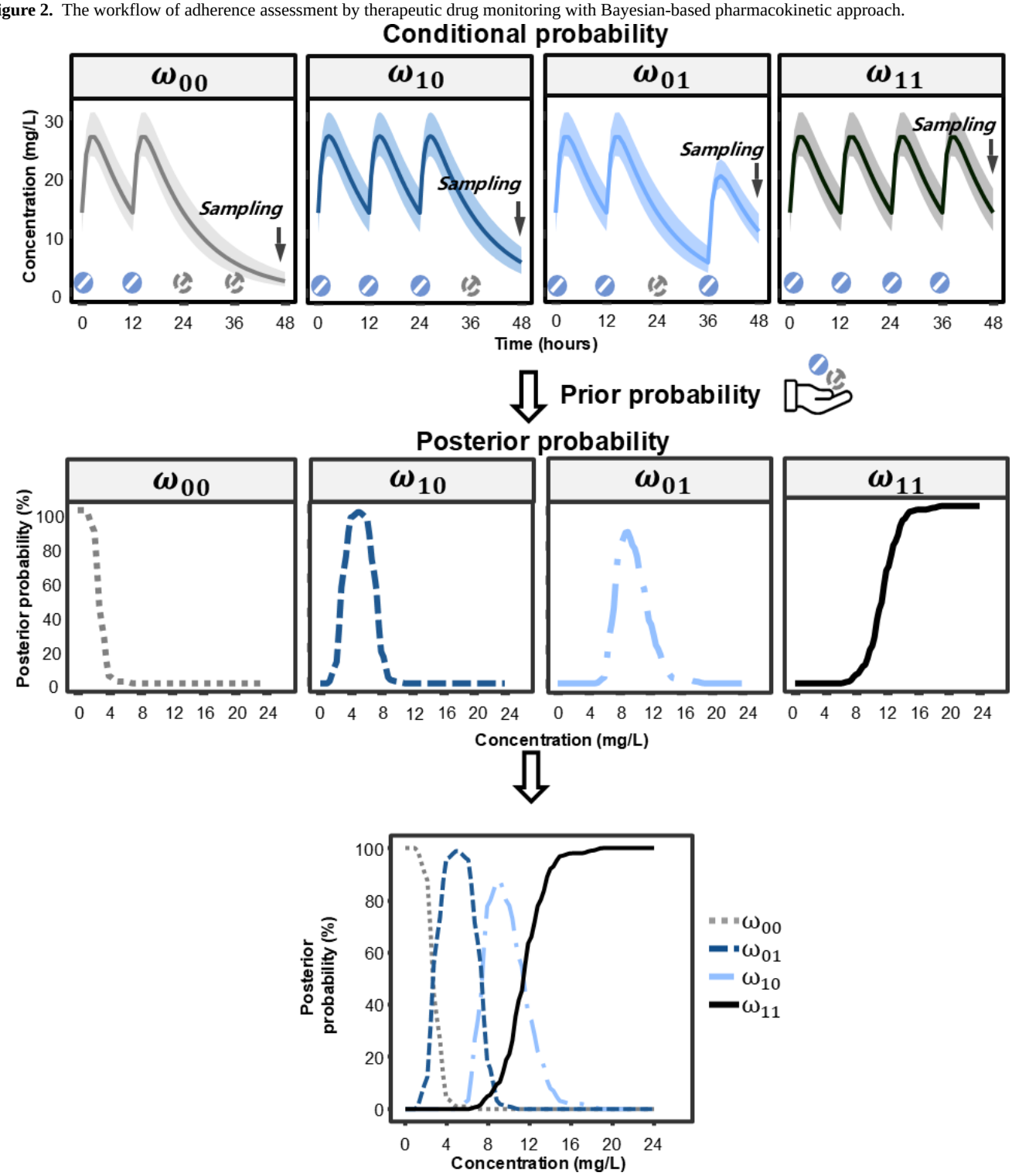

The C_0 value at which the posterior probabilities of two distinct dosing scenarios are equal, can be derived from Equations 1; 2, serving as a threshold to discriminate between these scenarios. As depicted in Figure 2, the posterior probabilities of missing two doses (ω00) and missing only the last dose while having taken the second last dose (ω01) are equal when the C_0 is approximately 3 mg/L. When C_0 is less than 3 mg/L, the probability of ω00 is the highest. Similarly, the C_0 range maps to the most probable scenario as follows: 3 - 7.5 mg/L to ω01, 7.5 - 11.5 mg/L to missing the second-last dose but taking the last dose (ω10) , and levels above 11.5 mg/L to taking both of the last two doses (ω11).

In this case, all dosing events could have a maximum posterior probability exceeding 80%, which was defined as complete retrodiction (Figure 3A). If only one or none dosing events had a maximum posterior probability exceeding 80%, it was defined as no retrodiction (Figure 3C). Other cases were defined as partial retrodiction (Figure 3B).

XSL•FO
RenderX

Figure 3. Illustration of retrodiction types based on posterior probabilities of dosing events prior to sampling. (A) complete retrodiction: the maximum posterior probabilities of all dosing events are ≥80%; (B) partial retrodiction: only events which are fully adherent and fully nonadherent have a maximum posterior probability ≥80%; and (C) no retrodiction: only one or no dosing events have a maximum posterior probability ≥80%.

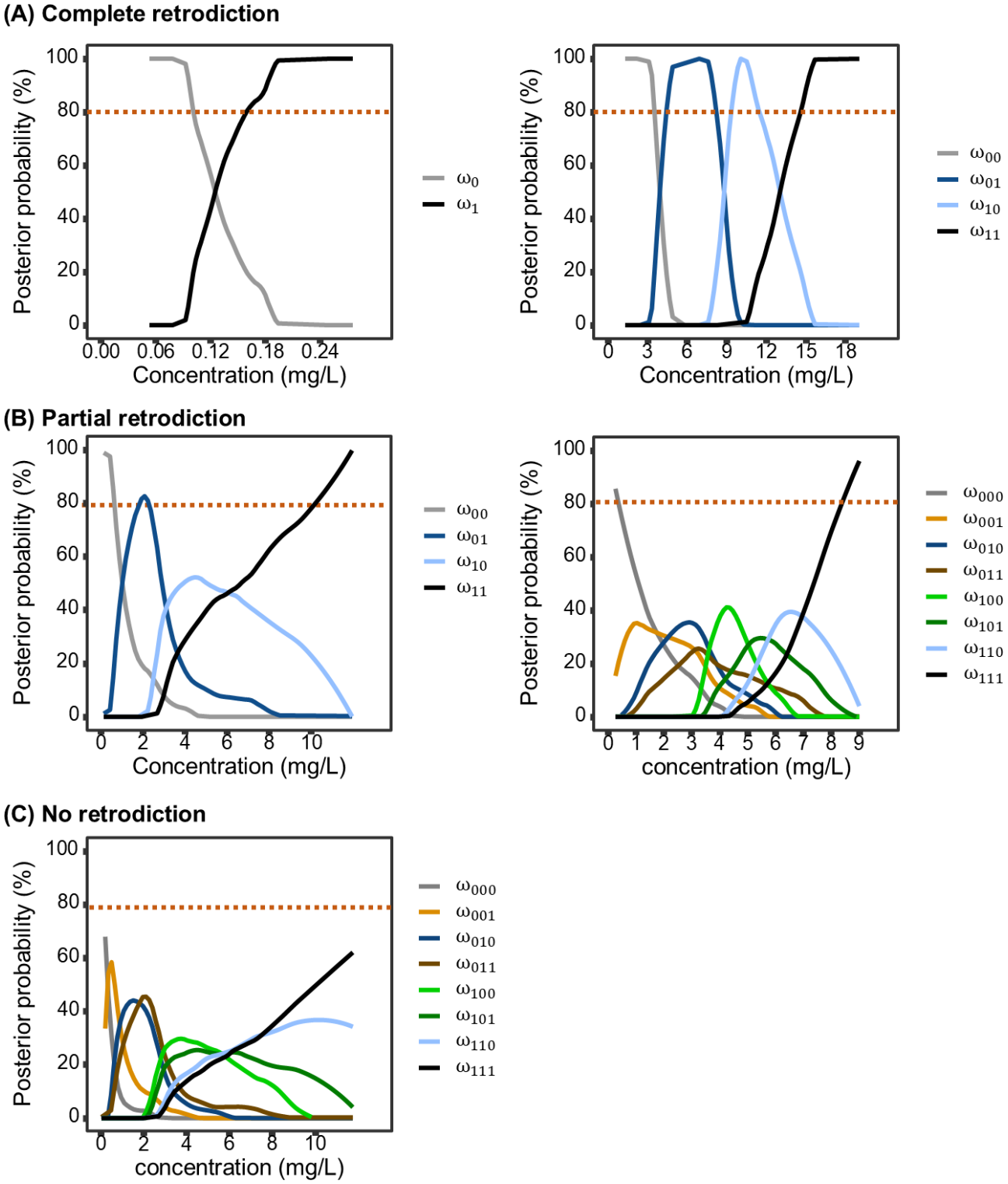

Population Pharmacokinetic Characteristics

To characterize the conditional probability associated with each dosing scenario, a systematic literature search was conducted in PubMed and Embase to collate available population PK parameters for ASMs across various formulations, including conventional tablets, oral solutions, suspensions, syrups, and extended-release (ER) formulations. In our previous study [18], population PK models for 10 commonly used ASMs were identified up to March 31, 2022, including carbamazepine, clobazam, eslicarbazepine acetate, lamotrigine, levetiracetam, oxcarbazepine, phenobarbital, topiramate, valproic acid, and zonisamide. An update search was then performed up to November 30, 2024. Additionally, the population PK characteristics of brivaracetam, lacosamide, perampanel and vigabatrin were also incorporated in the study. Details of the literature review were summarized in Multimedia Appendix 1.

XSL•FO
RenderX

Assessment of Adherence

During the adherence assessment, pediatric patients (aged 8 y, weighing 25 kg, and measuring 127 cm), adult patients (aged 40 y, weighing 70 kg, and measuring 180 cm), and pregnant women (aged 25 y, weighing 70 kg, measuring 160 cm, 30 wk pregnant) taking conventional tablet of ASMs were selected as typical patients. All typical patients had normal renal and liver function and were not taking any concomitant medications.

The evaluated dosing behaviors included the administration of the last one, two, and three doses prior to sampling at steady state. Sampling was conducted immediately before the subsequent dose. As for prior probability, it was assumed that each scenario had an equal chance of occurring. Specifically, this implies a probability of 50% for each scenario when considering only the last dosing behavior, 25% when considering the last two dosing behaviors, and 12.5% when considering the last three dosing behaviors.

For the Monte Carlo simulations, the parameters were fixed according to the final reported values, except for the residual unexplained variability (RUV) to obtain the "true" concentration-time profiles under various nonadherence scenarios. Consequently, RUV was set to negligible levels [21]—specifically 0.01 mg/L for additive error and 0.1% for proportional error—to minimize noise in the study. A total of 40,000 virtual patients were generated for each scenario. The Monte Carlo simulations were conducted using R programming (version 4.2.2; R Foundation for Statistical Computing) with the *rxode2* package (version 2.1.2). The results were plotted using the *ggplot2* package (version 3.5.1).

Evaluation of Critical Factors Affecting Adherence Assessment

The factors reported to significantly influence the PK of ASMs were investigated for their impact on adherence assessment, including renal function (estimated glomerular filtration rate, eGFR: 30, 60 and 90 mL/min/1.73m^2) and concomitant medications. Additionally, the effect of formulation (extending dosing interval to every 24 h for ER formulation), sampling time (2 h earlier or later), and prior probabilities (10%, 30%, 50%, 70%, and 90%) on medication adherence was also tested. The impact of these factors was assessed from two perspectives: the first was the ability to retrodict the number of the last scheduled doses, and the second was their influence on the concentration threshold used to distinguish between nonadherence patterns.

Development of Web-Based Dashboard

To facilitate quick calculation, an interactive online dashboard was developed to assess ASMs' medication adherence, informed by TDM results, and individual characteristics that were determined as significant factors on PK parameters in the included models. This tool was built using *rxode2* (version 2.1.2), *ggplot2* (version 3.5.1), and *shiny* (version 1.8.1.1) within the R framework (version 4.2.2; R Foundation for Statistical Computing).

Results

Population Pharmacokinetic Characteristics

A total of 23 population PK models encompassing 14 ASMs were ultimately included in the analysis [22-44]. Among these, models for adult [22-25,27-29,31,33,35,36,38,40,42] and pediatric patients [22,23,25-30,32,34,36,37,39,44] were available, while models specific to pregnant women were only available for lamotrigine [41] and levetiracetam [43]. Since age was consistently identified as a significant covariate for PK parameters in adults, elderly patients were thus grouped with the adult population. Models characterizing multiple formulations were identified for eslicarbazepine acetate [34], lamotrigine [36] and valproic acid [30,42]. Details of the identification of literature, included studies, and the final parameter estimates used in the analysis were comprehensively summarized in Figures S1-S9 in Multimedia Appendix 1 and Tables S1-S2 in Multimedia Appendix 1.

Assessment of Adherence

The posterior probabilities of various dosing behaviors when considering the last one, two, and three dosing behaviors for each ASM are detailed in Figure S10-S23 in Multimedia Appendix 1. Figure 4 demonstrates the ability to retrodict the number of the last scheduled doses for each ASM in typical patients under conditions of minimized RUV. Results indicated that when investigating the most recent dosing behavior, all ASMs can be fully retrodicted. Regarding the scenarios involving the last two dosing behaviors, complete retrodiction was achievable only for oxcarbazepine in pediatric patients, whereas other ASMs were partially retrodicted. When extending to the last three dosing behaviors, no ASMs can be fully retrodicted, and only carbamazepine, clobazam, eslicarbazepine acetate, oxcarbazepine, phenobarbital and zonisamide could be partially retrodicted across all investigated population. From a pharmacokinetic perspective, ASMs with higher clearance are eliminated rapidly, thereby diminishing the concentration "signal" necessary to distinguish earlier dosing events. Furthermore, the traceability may vary in clinical scenarios where patient characteristics significantly deviate from the typical population.

XSL•FO
RenderX

Figure 4. The ability to retrodict the last one, two and three dosing behaviors prior to sampling for typical patients in theoretical condition when minimizing residual unexplained variabilities. D1: the last dosing behavior; D2: the last two dosing behaviors; D3: the last three dosing behaviors; CR, complete retrodiction, which is defined as when the maximum posterior probabilities of all dosing events are ≥80%; PR, partial retrodiction, which is defined as when only events which are fully adherent and fully nonadherent (have a maximum posterior probability ≥80%; NR: no retrodiction, which is defined as when only one or no dosing events have a maximum posterior probability ≥80%. Adults: aged 40 y, weighing 70 kg, and measuring 180 cm; children: aged 8 y, weighing 25 kg, and measuring 127 cm; pregnant women: aged 25 y, weighing 70 kg, measuring 160 cm, and being 30 weeks pregnant. Residual unexplained variabilities were minimized by defining additive error as 0.01 mg/L and proportional error as 0.1%.

Antiseizure medication	Population	D1	D2	D3
Brivaracetam	Adults	CR	PR	NR
	Children	CR	PR	NR
Carbamazepine	Adults	CR	PR	PR
	Children	CR	PR	PR
Clobazam	Adults	CR	PR	PR
	Children	CR	PR	PR
Eslicarbazepine acetate	Adults	CR	PR	PR
	Children (tablet)	CR	PR	PR
	Children (oral suspension)	CR	PR	PR
Lacosamide	Adults	CR	PR	NR
	Children	CR	PR	NR
Lamotrigine	Adults (immediate-release/ extended-release formulation)	CR	PR	PR
	Children (immediate-release formulation)	CR	PR	NR
	Children (extended-release formulation)	CR	PR	PR
	Pregnant women	CR	PR	NR
Levetiracetam	Adults	CR	PR	NR
	Children	CR	PR	NR
	Pregnant women	CR	PR	NR
Oxcarbazepine	Adults	CR	PR	PR
	Children	CR	CR	PR
Perampanel	Adults	CR	PR	NR
	Children	CR	PR	PR
Phenobarbital	Adults	CR	PR	PR
	Children	CR	PR	PR
Topiramate	Adults	CR	PR	NR
	Children	CR	PR	NR
Valproic acid	Adults (tablet/oral solution/modified-release coated tablet)	CR	PR	NR
	Children (tablet/syrup/sustained-release)	CR	PR	NR
Vigabatrin	Adults	CR	PR	NR
	Children	CR	PR	NR
Zonisamide	Adults	CR	PR	PR
	Children	CR	PR	PR

XSL•FO
RenderX

Impact of Critical Factors on Adherence Assessment

It has been reported that renal function affects the apparent clearance (CL/F) of eslicarbazepine acetate [33], levetiracetam [24], oxcarbazepine [38] and vigabatrin [28]. As renal function decreases, there is no significant effect on the identification of nonadherence patterns. The effect of renal function on levetiracetam is illustrated in Figure S24 in Multimedia Appendix 1, with levetiracetam, oxcarbazepine, and vigabatrin demonstrating similar trends.

Pregnancy enhances the clearance of lamotrigine and levetiracetam, leading to lower systemic drug exposure and, consequently, may decrease the concentration thresholds (Figure S15-S16 in Multimedia Appendix 1). Similarly, pediatric patients show higher clearance per body weight compared with adults, leading to the decreased concentration thresholds (Figure S10-S23 in Multimedia Appendix 1).

The effects of concomitant inducers and inhibitors were also evaluated. Administration of inducers or inhibitors did not alter the fundamental distinguishability of nonadherence patterns but shifted the concentration threshold required for their discrimination. Specifically, the threshold was lowered by inducers and raised by inhibitors. The magnitude of these adjustments varied substantially across ASMs (Figure S25 in Multimedia Appendix 1). For instance, in a typical adult patient taking lamotrigine, coadministration with enzyme inducers (eg, carbamazepine, phenobarbital) lowered the threshold by approximately 67%, whereas the inhibitors valproic acid elevated it by approximately 83%. The effect on topiramate was less pronounced.

The impact of formulation on adherence assessment was evaluated for eslicarbazepine acetate, lamotrigine and valproic acid. At equivalent total daily doses, ER formulations with prolonged dosing intervals enhanced the discriminative capacity for dosing behaviors compared to immediate-release (IR) or other oral formulations requiring more frequent administration (Figure S15, S21 in Multimedia Appendix 1). In contrast, formulations such as oral suspensions and syrups exhibited minimal impact on the assessment (Figure S13, S21 in Multimedia Appendix 1).

Sampling time also influences adherence assessment (Figure S26 in Multimedia Appendix 1). Sampling 2 hours earlier or later than the scheduled time does not significantly influence the distinguishability of nonadherence patterns. However, compared to sampling just before administration, the concentration threshold for distinguishing nonadherence patterns increases when sampling is done earlier and decreases when sampling is done later. The magnitude of this change varies among different ASMs.

The impacts of prior probabilities on adherence assessment were also evaluated. The results indicated that prior probabilities could not only significantly affect the distinguishability of the nonadherence dosing scenarios, but also notably alter the concentration threshold for distinguishability (Figure S27 in Multimedia Appendix 1). The magnitude of the concentration threshold change was found to be dependent on the type of ASMs.

Application of Web-Based Dashboard

A web-based dashboard for assessing medication adherence has been developed and is freely accessible online [45]. After inputting the type of ASMs, patient characteristics (age, body weight, height, gender), scheduled dosing regimens, sampling time, TDM data, and prior probabilities for each scenario, the system estimates the posterior probabilities of each dosing scenario and plots them against the drug concentration. RUV are initialized with literature-reported values (as listed in Table S2 in Multimedia Appendix 1) when requiring consideration, but remain user-adjustable to accommodate specific clinical situations, thereby enabling the precise identification of medication adherence patterns.

Figure 5A presents a case of a 75-year-old male (70 kg) with epilepsy and impaired renal function (eGFR 40 mL/min/1.73m^2) who had remained seizure-free for over 3 years on oxcarbazepine 300 mg q12h. Following a recent increase in seizure frequency, TDM was performed to assess potential nonadherence. The measured C_0 of oxcarbazepine was 12 mg/L, which lies within the conventional therapeutic range of 10 - 35 mg/L and aligns with the recommended range of 6 - 24 mg/L as per the 2017 guidelines [12]. However, model-based estimates from the dashboard indicated a nearly negligible probability of full adherence and a high probability of having missed at least one dose. The posterior probability of full adherence remained at 0%, regardless of whether the prior probability was set as low as 1% (suggesting poor adherence) or as high as 99% (denoting high adherence) (Figure S28 in Multimedia Appendix 1), demonstrating the minimal influence of the prior in this case. By contrast, increasing the RUV from 0.1% to 30% increased the posterior probability of full adherence from 0% to 56% (Figure S29 in Multimedia Appendix 1), underscoring the substantial impact of RUV on the adherence assessment in this case.

Figure 5B illustrates another case of a 10-year-old (30 kg) boy with epilepsy and normal renal/hepatic function, treated with valproic acid tablet 500 mg and carbamazepine 150 mg q12h. His measured C_0 of valproic acid was 40 mg/L, below both the conventional therapeutic range (50 - 100 mg/L) and the 2017 guideline-recommended range (62.2 - 134.8 mg/L). Although subtherapeutic concentrations initially raised suspicion of nonadherence, model-based analysis estimated a probability of full adherence exceeding 80%, suggesting that the low valproic acid concentration likely resulted from carbamazepine-induced metabolic induction rather than missed doses. Varying the prior probability of adherence from 1% to 99% had minimal impact on this conclusion (Figure S30 in Multimedia Appendix 1). Similarly, increasing RUV from 0.1% to 30% altered the posterior probability of full adherence by only 5% (Figure S31 in Multimedia Appendix 1), demonstrating the robustness of the assessment against RUV variation in this clinical scenario.

XSL•FO
RenderX

Figure 5. Screenshot of the dashboard for adherence assessment. (A) Elderly patient, 75 years old, weighing 70 kg, measuring 180 cm, eGFR 40 mL/min/1.73m^2, taking oxcarbazepine 300 mg q12h; (B) pediatric patient, 10 years old, weighing 30 kg, measuring 130 cm, taking valproic acid 500 mg q12h and carbamazepine 100 mg q12h. $_{00}$: missing two continuous doses before sampling; $_{01}$, missing the second-to-last dose but taking the last dose; $_{10}$, missing the last dose but taking the second-to-last dose; $_{11}$, taking all doses.

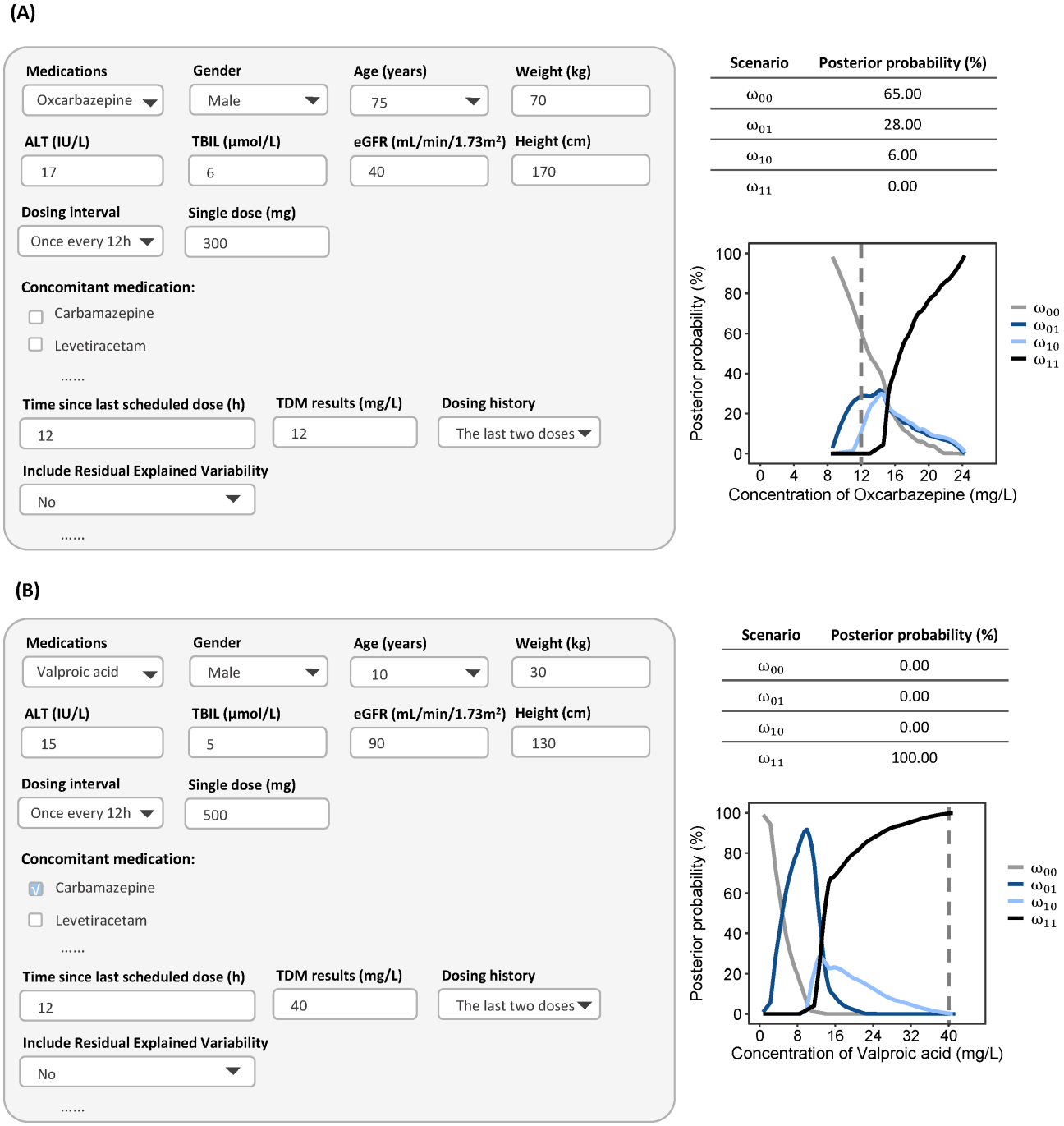

Discussion

Principal Findings

This study is the first to introduce a clinical framework to investigate the role of TDM in assessing medication adherence for 14 commonly used ASMs among diverse patients. By integrating Bayesian theory with population PK, we demonstrated that routine TDM, when combined with clinical factors, enables quantitative retrodiction of recent medication-taking behaviors for all investigated ASMs. The Bayesian-based PK framework can also be easily available through the open-access dashboard developed in this study.

Comparison to Prior Work

The 2017 Guideline established reference ranges encompassing approximately 66% of patients for commonly used ASMs, primarily derived from PK data obtained in adult patients receiving monotherapy [12]. While clinically useful, these population-derived thresholds have limited generalizability to special populations with distinct PK profiles. In contrast, the model-informed algorithm developed in this study enables a more personalized assessment of medication adherence. This

XSL•FO
RenderX

approach explicitly accounts for patient-specific factors including age (eg, pediatric and geriatric populations), pregnancy status, renal and hepatic function, concomitant use of enzyme inducers or inhibitors, formulation characteristics. It thereby provides a refined framework for evaluating medication-taking behavior across diverse clinical scenarios.

Interpretation of the Findings

We identified multiple critical factors that influence the adherence assessment, including intrinsic factors (physiological differences, concomitant medication, renal function, formulation , etc) and extrinsic factors (prior probability, RUV, etc).

The influence of intrinsic factors on adherence evaluation is primarily mediated through alterations in PK parameters, most notably systemic drug clearance. Enhanced drug clearance, commonly observed in pediatric and pregnant patients, as well as in those receiving enzyme inducers, reduces both the ability to differentiate adherence patterns and the corresponding concentration thresholds. In pediatric populations, the higher clearance per body weight results from ongoing organ maturation, larger organ size-to-body weight ratios, and increased metabolic enzyme activity [15,36,46]. In pregnant women, elevated clearance arises from increased cardiac output, enhanced renal blood flow, and hormonally mediated induction of metabolic pathways [14,41,47,48]. Conversely, reduced clearance, frequently encountered in patients with renal impairment or those receiving enzyme inhibitors, may improve differentiation ability or raise the concentration thresholds required for pattern discrimination.

Despite the increasing use of ER formulations of ASMs, conventional formulations continue to account for a substantial proportion of prescriptions due to their lower cost and wider availability [49,50]. Consequently, population PK studies have more frequently characterized conventional formulations. Based on the limited population PK data available for ER ASMs, our findings suggest that ER formulations may improve the differentiation of adherence behaviors, a finding attributable to the extended dosing interval (eg, from 12 to 24 h) and the resulting concentration-time fluctuation.

Prior probability is essential for estimating the posterior probability of dosing behaviors. In this study, we adopted an equiprobable prior probability to reflect a state of maximum uncertainty before considering the evidence (TDM measurements). This represents a conventional and conservative strategy in Bayesian modeling when reliable, specific prior knowledge is unavailable [19,51]. In real clinical settings, the prior probability can be informed by pharmacy refill data or population-average adherence estimates. When individual-level data are absent, population-based priors derived from patients with comparable covariates (eg, age, comorbidities, and socioeconomic status) may be applied. Although the impact of the prior was limited in the cases illustrated in Figure 5, its influence on adherence assessment can be substantial and depends on both the specific ASM and TDM measurements. Consequently, the ability for user-defined priors implemented in the dashboard remains highly valuable.

RUV in population PK analysis captures unexplained stochastic variations, including assay error, sampling inaccuracies, and model misspecification. As these elements may confound medication adherence assessments, RUV was intentionally minimized in the present analysis to reduce setting-specific noise and facilitate clearer characterization of covariate effects. To enhance real-world applicability, the accompanying dashboard allows users to adjust the RUV level based on reported values from source population PK studies (Table S2 in Multimedia Appendix 1), known assay variability, or clinical experience.

Limitations

The study has several limitations. First, there are numerous patterns of nonadherence and we only considered the scenario of missing doses. Other types of nonadherence, such as delayed doses, missed partial doses, and inadvertent overdoses, were not considered. Second, due to the lack of population-PK studies, specifically in pediatric patients, pregnant women, and for ER formulation, we did not include all these scenarios in our analysis. However, our dashboard can be readily extended to include these populations, or novel formulations once their population PK parameters become available. In addition, it is important to note that while we provide accurate estimation of probabilities for recent medication events, the clinical judgment should not be solely based on it. Comprehensive assessment must be performed to incorporate the patient's overall condition, medication history, and relevant information.

Future Directions

As epilepsy pharmacotherapy evolves, the dashboard will be updated to incorporate emerging population PK models for novel ASMs when available. Concurrently, it will extend to pharmacodynamic models to bridge the gap between PK and clinical outcomes, thereby quantifying the risks of breakthrough seizures from nonadherence trajectories. Leveraging prior work on remedial dosing regimens on ASMs [18], the tool can be expanded to provide remedial dosing strategies for clinicians to safely restore therapeutic concentrations after missed doses. Finally, the integration with multidimensional data, such as digital biomarkers and electronic health records, will be explored. The comprehensive strategy will ultimately facilitate the realization of precise, patient-specific life-cycle management in epilepsy treatment.

Conclusion

In conclusion, this study establishes a Bayesian-based PK approach to enhance the objective assessment of ASM adherence. By leveraging TDM data, the approach showed large improvement in assessing nonadherence patterns compared to the previous guidelines. In addition, to bridge the methodology with clinical practice, we developed an interactive dashboard that translates PK principles into visual and interpretable outputs. The work demonstrates the feasibility of transitioning from traditional population-based monitoring to individual-specific management for patients with epilepsy.

XSL•FO
RenderX

Funding

XQL was supported by the Elite Development Supporting Program from Shanghai Chest Hospital, Shanghai Jiao Tong University School of Medicine (RC-202407-076).

Data Availability

Authors can confirm that all relevant data are included in the article and its supplementary information files. The code used for data analysis is available from the corresponding author upon reasonable request

Authors' Contributions

XQL: Conceptualization, Software, Investigation, Formal analysis, Data Curation, Writing-Original Draft, Writing-Review & Editing, Funding acquisition; ZRL: Methodology, Formal analysis, Software, Writing-Review & Editing; WWL: Methodology, Software, Writing-Review & Editing; JW: Visualization, Writing-Review & Editing; FQG: Visualization, Writing-Review & Editing; JJD: Conceptualization, Validation, Writing-Review & Editing; ZJ: Conceptualization, Supervision, Project Administration, Writing-Original Draft, Writing-Review & Editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Literature identification, the posterior probabilities-concentration curves, sensitivity analysis results, and the summary of identified population pharmacokinetic models.
[DOCX File, 7346 KB - jmir_v28i1e77917_app1.docx]

References

1. Glauser T, Ben-Menachem E, Bourgeois B, et al. Updated ILAE evidence review of antiepileptic drug efficacy and effectiveness as initial monotherapy for epileptic seizures and syndromes. Epilepsia 2013 Mar;54(3):551-563. [doi: 10.1111/epi.12074] [Medline: 23350722]
2. Nevitt SJ, Sudell M, Cividini S, Marson AG, Tudur Smith C. Antiepileptic drug monotherapy for epilepsy: a network meta-analysis of individual participant data. Cochrane Database Syst Rev 2022 Apr 1;4(4):CD011412. [doi: 10.1002/14651858.CD011412.pub4] [Medline: 35363878]
3. Al-Aqeel S. Nonadherence to antiseizure medications: what have we learned and what can be done next? Expert Rev Pharmacoecon Outcomes Res 2024 Sep;24(7):791-798. [doi: 10.1080/14737167.2024.2349191] [Medline: 38676921]
4. Malek N, Heath CA, Greene J. A review of medication adherence in people with epilepsy. Acta Neurol Scand 2017 May;135(5):507-515. [doi: 10.1111/ane.12703] [Medline: 27781263]
5. Niriayo YL, Mamo A, Gidey K, Demoz GT. Medication belief and adherence among patients with epilepsy. Behav Neurol 2019 Apr 23;2019:2806341. [doi: 10.1155/2019/2806341] [Medline: 31178940]
6. Adisu MA, Zemariam AB, Derso YA, et al. Global patterns and predictors of anti-seizure medication adherence in pediatric epilepsy: A systematic review and meta-analysis. Epilepsy Behav 2025 Nov;172:110544. [doi: 10.1016/j.yebeh.2025.110544] [Medline: 40499262]
7. Faught E, Duh MS, Weiner JR, Guérin A, Cunnington MC. Nonadherence to antiepileptic drugs and increased mortality: findings from the RANSOM Study. Neurology 2008 Nov 11;71(20):1572-1578. [doi: 10.1212/01.wnl.0000319693.10338.b9] [Medline: 18565827]
8. Faught RE, Weiner JR, Guérin A, Cunnington MC, Duh MS. Impact of nonadherence to antiepileptic drugs on health care utilization and costs: findings from the RANSOM study. Epilepsia 2009 Mar;50(3):501-509. [doi: 10.1111/j.1528-1167.2008.01794.x] [Medline: 19183224]
9. Chan AHY, Wright DFB. Medication adherence-Everybody's problem but nobody's responsibility? Br J Clin Pharmacol 2025 Mar;91(3):681-683. [doi: 10.1111/bcp.16384] [Medline: 39734279]
10. Patsalos PN, Spencer EP, Berry DJ. Therapeutic drug monitoring of antiepileptic drugs in epilepsy: a 2018 update. Ther Drug Monit 2018 Oct;40(5):526-548. [doi: 10.1097/FTD.0000000000000546] [Medline: 29957667]
11. Stepanova D, Beran RG. The benefits of antiepileptic drug (AED) blood level monitoring to complement clinical management of people with epilepsy. Epilepsy Behav 2015 Jan;42:7-9. [doi: 10.1016/j.yebeh.2014.09.069] [Medline: 25499154]
12. Hiemke C, Bergemann N, Clement HW, et al. Consensus Guidelines for Therapeutic Drug Monitoring in Neuropsychopharmacology: Update 2017. Pharmacopsychiatry 2018 Jan;51(1-2):9-62. [doi: 10.1055/s-0043-116492] [Medline: 28910830]
13. Roberti R, Palleria C, Nesci V, et al. Pharmacokinetic considerations about antiseizure medications in the elderly. Expert Opin Drug Metab Toxicol 2020 Oct;16(10):983-995. [doi: 10.1080/17425255.2020.1806236] [Medline: 32757857]

XSL•FO
RenderX

14. Arfman IJ, Wammes-van der Heijden EA, Ter Horst PGJ, Lambrechts DA, Wegner I, Touw DJ. Therapeutic drug monitoring of antiepileptic drugs in women with epilepsy before, during, and after pregnancy. Clin Pharmacokinet 2020 Apr;59(4):427-445. [doi: 10.1007/s40262-019-00845-2] [Medline: 31912315]
15. Maglalang PD, Wen J, Hornik CP, Gonzalez D. Sources of pharmacokinetic and pharmacodynamic variability and clinical pharmacology studies of antiseizure medications in the pediatric population. Clin Transl Sci 2024 Apr;17(4):e13793. [doi: 10.1111/cts.13793] [Medline: 38618871]
16. Clark ED, Lawley SD. Should patients skip late doses of medication? A pharmacokinetic perspective. J Pharmacokinet Pharmacodyn 2022 Aug;49(4):429-444. [doi: 10.1007/s10928-022-09812-0] [Medline: 35726046]
17. Clark ED, Lawley SD. How drug onset rate and duration of action affect drug forgiveness. J Pharmacokinet Pharmacodyn 2024 Jun;51(3):213-226. [doi: 10.1007/s10928-023-09897-1] [Medline: 38198076]
18. Li ZR, Wang CY, Lin WW, Chen YT, Liu XQ, Jiao Z. Handling delayed or missed dose of antiseizure medications: a model-informed individual remedial dosing. Neurology 2023 Feb 28;100(9):e921-e931. [doi: 10.1212/WNL.0000000000201604] [Medline: 36450606]
19. Barrière O, Li J, Nekka F. A Bayesian approach for the estimation of patient compliance based on the last sampling information. J Pharmacokinet Pharmacodyn 2011 Jun;38(3):333-351. [doi: 10.1007/s10928-011-9196-2] [Medline: 21445612]
20. Ding J, Coldiron ME, Assao B, et al. Adherence and population pharmacokinetic properties of amodiaquine when used for seasonal malaria chemoprevention in african children. Clin Pharmacol Ther 2020 May;107(5):1179-1188. [doi: 10.1002/cpt.1707] [Medline: 31652336]
21. Nick H. Simulations with/without residual error. The Mail Archive. 2009. URL: https://www.mail-archive.com/nmusers@globomaxnm.com/msg01817.html [accessed 2025-11-30]
22. Jiao Z, Shi XJ, Zhao ZG, Zhong MK. Population pharmacokinetic modeling of steady state clearance of carbamazepine and its epoxide metabolite from sparse routine clinical data. J Clin Pharm Ther 2004 Jun;29(3):247-256. [doi: 10.1111/j.1365-2710.2004.00557.x] [Medline: 15153086]
23. Goto S, Seo T, Murata T, et al. Population estimation of the effects of cytochrome P450 2C9 and 2C19 polymorphisms on phenobarbital clearance in Japanese. Ther Drug Monit 2007 Feb;29(1):118-121. [doi: 10.1097/FTD.0b013e318030def0] [Medline: 17304159]
24. Pigeolet E, Jacqmin P, Sargentini-Maier ML, Stockis A. Population pharmacokinetics of levetiracetam in Japanese and Western adults. Clin Pharmacokinet 2007;46(6):503-512. [doi: 10.2165/00003088-200746060-00004] [Medline: 17518509]
25. Okada Y, Seo T, Ishitsu T, et al. Population estimation regarding the effects of cytochrome P450 2C19 and 3A5 polymorphisms on zonisamide clearance. Ther Drug Monit 2008 Aug;30(4):540-543. [doi: 10.1097/FTD.0b013e31817d842a] [Medline: 18641551]
26. Chhun S, Jullien V, Rey E, Dulac O, Chiron C, Pons G. Population pharmacokinetics of levetiracetam and dosing recommendation in children with epilepsy. Epilepsia 2009 May;50(5):1150-1157. [doi: 10.1111/j.1528-1167.2008.01974.x] [Medline: 19175400]
27. Girgis IG, Nandy P, Nye JS, et al. Pharmacokinetic-pharmacodynamic assessment of topiramate dosing regimens for children with epilepsy 2 to <10 years of age. Epilepsia 2010 Oct;51(10):1954-1962. [doi: 10.1111/j.1528-1167.2010.02598.x] [Medline: 20880232]
28. Nielsen JC, Kowalski KG, Karim A, Patel M, Wesche DL, Tolbert D. Population pharmacokinetics analysis of vigabatrin in adults and children with epilepsy and children with infantile spasms. Clin Pharmacokinet 2014 Nov;53(11):1019-1031. [doi: 10.1007/s40262-014-0172-z] [Medline: 25172554]
29. Saruwatari J, Ogusu N, Shimomasuda M, et al. Effects of CYP2C19 and P450 oxidoreductase polymorphisms on the population pharmacokinetics of clobazam and N-desmethylclobazam in japanese patients with epilepsy. Ther Drug Monit 2014 Jun;36(3):302-309. [doi: 10.1097/FTD.0000000000000015] [Medline: 24345815]
30. Ding J, Wang Y, Lin W, et al. A population pharmacokinetic model of valproic acid in pediatric patients with epilepsy: a non-linear pharmacokinetic model based on protein-binding saturation. Clin Pharmacokinet 2015 Mar;54(3):305-317. [doi: 10.1007/s40262-014-0212-8] [Medline: 25388986]
31. Schoemaker R, Wade JR, Stockis A. Brivaracetam population pharmacokinetics and exposure‐response modeling in adult subjects with partial‐onset seizures. J Clin Pharmacol 2016 Dec;56(12):1591-1602. [doi: 10.1002/jcph.761]
32. Schoemaker R, Wade JR, Stockis A. Brivaracetam population pharmacokinetics in children with epilepsy aged 1 month to 16 years. Eur J Clin Pharmacol 2017 Jun;73(6):727-733. [doi: 10.1007/s00228-017-2230-6] [Medline: 28280887]
33. Gidal BE, Jacobson MP, Ben-Menachem E, et al. Exposure-safety and efficacy response relationships and population pharmacokinetics of eslicarbazepine acetate. Acta Neurol Scand 2018 Sep;138(3):203-211. [doi: 10.1111/ane.12950] [Medline: 29732549]
34. Sunkaraneni S, Ludwig E, Fiedler-Kelly J, Hopkins S, Galluppi G, Blum D. Modeling and simulations to support dose selection for eslicarbazepine acetate therapy in pediatric patients with partial-onset seizures. J Pharmacokinet Pharmacodyn 2018 Aug;45(4):649-658. [doi: 10.1007/s10928-018-9596-7] [Medline: 29948795]
35. Takenaka O, Ferry J, Saeki K, Laurenza A. Pharmacokinetic/pharmacodynamic analysis of adjunctive perampanel in subjects with partial-onset seizures. Acta Neurol Scand 2018 Apr;137(4):400-408. [doi: 10.1111/ane.12874] [Medline: 29171002]

XSL•FO
RenderX

36. van Dijkman SC, de Jager NCB, Rauwé WM, Danhof M, Della Pasqua O. Effect of age-related factors on the pharmacokinetics of lamotrigine and potential implications for maintenance dose optimisation in future clinical trials. Clin Pharmacokinet 2018 Aug;57(8):1039-1053. [doi: 10.1007/s40262-017-0614-5] [Medline: 29363050]
37. Lin WW, Li XW, Jiao Z, et al. Population pharmacokinetics of oxcarbazepine active metabolite in Chinese paediatric epilepsy patients and its application in individualised dosage regimens. Eur J Clin Pharmacol 2019 Mar;75(3):381-392. [doi: 10.1007/s00228-018-2600-8] [Medline: 30456415]
38. Lin WW, Wang CL, Jiao Z, et al. Glomerular filtration rate is a major predictor of clearance of oxcarbazepine active metabolite in adult Chinese epileptic patients: a population pharmacokinetic analysis. Ther Drug Monit 2019 Oct;41(5):665-673. [doi: 10.1097/FTD.0000000000000644] [Medline: 31033858]
39. Winkler J, Schoemaker R, Stockis A. Population pharmacokinetics of adjunctive lacosamide in pediatric patients with epilepsy. J Clin Pharmacol 2019 Apr;59(4):541-547. [doi: 10.1002/jcph.1340] [Medline: 30427550]
40. Winkler J, Schoemaker R, Stockis A. Modeling and simulation for the evaluation of dose adaptation rules of intravenous lacosamide in children. Epilepsy Res 2019 Jan;149:13-16. [doi: 10.1016/j.eplepsyres.2018.10.011] [Medline: 30415109]
41. Wang ML, Tao YY, Sun XY, et al. Estrogen profile- and pharmacogenetics-based lamotrigine dosing regimen optimization: Recommendations for pregnant women with epilepsy. Pharmacol Res 2021 Jul;169:105610. [doi: 10.1016/j.phrs.2021.105610] [Medline: 33857625]
42. Teixeira-da-Silva P, Pérez-Blanco JS, Santos-Buelga D, Otero MJ, García MJ. Population Pharmacokinetics of Valproic Acid in Pediatric and Adult Caucasian Patients. Pharmaceutics 2022 Apr 7;14(4):811. [doi: 10.3390/pharmaceutics14040811] [Medline: 35456645]
43. Li Y, Wang ML, Guo Y, Cao YF, Zhao MM, Zhao LM. Population pharmacokinetics and dosing regimen optimization of levetiracetam in epilepsy during pregnancy. Br J Clin Pharmacol 2023 Mar;89(3):1152-1161. [doi: 10.1111/bcp.15572] [Medline: 36260320]
44. Li S, Yi J, Tuo Y, et al. Population pharmacokinetics and dosing optimization of perampanel in children with epilepsy: A real-world study. Epilepsia 2024 Jun;65(6):1687-1697. [doi: 10.1111/epi.17954] [Medline: 38572689]
45. Pharmacokinetic-based bayesian approach to assess medication adherence. MIPD. URL: https://mipd.shinyapps.io/adherence_ASM/ [accessed 2025-11-30]
46. Krekels EHJ, Calvier EAM, van der Graaf PH, Knibbe CAJ. Children are not small adults, but can we treat them as such? CPT Pharmacometrics Syst Pharmacol 2019 Jan;8(1):34-38. [doi: 10.1002/psp4.12366] [Medline: 30689298]
47. Pennell PB, Karanam A, Meador KJ, et al. Antiseizure medication concentrations during pregnancy: results from the Maternal Outcomes and Neurodevelopmental Effects of Antiepileptic Drugs (MONEAD) Study. JAMA Neurol 2022 Apr 1;79(4):370-379. [doi: 10.1001/jamaneurol.2021.5487] [Medline: 35157004]
48. Yin X, Liu Y, Guo Y, Zhao L, Li G, Tan X. Pharmacokinetic changes for newer antiepileptic drugs and seizure control during pregnancy. CNS Neurosci Ther 2022 May;28(5):658-666. [doi: 10.1111/cns.13796] [Medline: 35037389]
49. Gidal BE, Ferry J, Reyderman L, Piña-Garza JE. Use of extended-release and immediate-release anti-seizure medications with a long half-life to improve adherence in epilepsy: a guide for clinicians. Epilepsy Behav 2021 Jul;120:107993. [doi: 10.1016/j.yebeh.2021.107993] [Medline: 33971390]
50. Javarayee P, Mtchedlidze T, Snell W, et al. Pricing dynamics of anti-seizure medications in the U.S. Seizure 2024 Nov;122:26-33. [doi: 10.1016/j.seizure.2024.09.010] [Medline: 39306895]
51. Barrière O, Li J, Nekka F. Compliance spectrum as a drug fingerprint of drug intake and drug disposition. J Pharmacokinet Pharmacodyn 2013 Feb;40(1):41-52. [doi: 10.1007/s10928-012-9285-x] [Medline: 23250805]

Abbreviations

ASM: antiseizure medications
C_0: steady-state trough concentration
PK: pharmacokinetic
PPK: population pharmacokinetic
TDM: therapeutic drug monitoring

Edited by A Coristine; submitted 26.May.2025; peer-reviewed by J Liao, M Guignet, S Lawley, Y Wang; revised version received 09.Dec.2025; accepted 09.Dec.2025; published 02.Jan.2026.

Please cite as:
Liu XQ, Li ZR, Lin WW, Wang J, Gu FQ, Ding JJ, Jiao Z
Bayesian-Based Pharmacokinetic Framework Integrated with Therapeutic Drug Monitoring for Assessing Adherence to Antiseizure Medications: A Clinical Trial Simulation Study
J Med Internet Res 2026;28:e77917
URL: https://www.jmir.org/2026/1/e77917
doi:10.2196/77917

XSL•FO
RenderX

© Xiao-Qin Liu, Zi-Ran Li, Wei-Wei Lin, Juan Wang, Fu-Qing Gu, Jun-Jie Ding, Zheng Jiao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Real-Life Digital Intervention for Personalized Nutrition in Adults With Overweight or Obesity: Remote Randomized Controlled Trial

Jelle CBC de Jong[1*], PhD; Femke PM Hoevenaars[1*], PhD; Lotte GP Peters[1,2], MSc; Charlotte MM Berendsen[1], MSc; Wilrike J Pasman[1], PhD; Martien PM Caspers[1], BSc; Remon Dulos[1], BSc; Suzan Wopereis[1,2], PhD

[1]Department of Microbiology and Systems Biology, Netherlands Organisation for Applied Scientific Research, Leiden, The Netherlands

[2]Department of Biomedical Signals and Systems, University of Twente, Enschede, The Netherlands

[*]these authors contributed equally

Corresponding Author:
Suzan Wopereis, PhD
Department of Microbiology and Systems Biology
Netherlands Organisation for Applied Scientific Research
Sylviusweg 71
Leiden, 2333 BE
The Netherlands
Phone: 31 88 866 75 00
Email: suzan.wopereis@tno.nl

Abstract

Background: A digital-first strategy is increasingly implemented to reduce participant burden, accelerate recruitment, collect real-world data, and increase the diversity of the study population. However, fully remote studies lack face-to-face interaction, which may affect motivation, particularly in the delivery of personalized nutritional advice. Additionally, self-reported data may vary in terms of standardization and completeness.

Objective: The study's primary objective is to evaluate the feasibility of conducting a fully remote, fully digital randomized controlled nutritional intervention, including participant experience and the capability to perform do-it-yourself anthropometric measurements at home. Secondary objectives are to determine whether self-collected data could detect changes in body weight and other anthropometric outcomes, and to compare the effectiveness of generic versus personalized nutrition advice, with and without personalized food boxes.

Methods: We conducted a fully online, 3-arm randomized controlled trial including adults with overweight or obesity who were motivated to lose weight. Participants were assigned to a control group that received generic advice (n=43), a personalized intervention group that received personalized advice only (n=40), or a personalized intervention plus group that received personalized advice plus personalized food boxes (n=39). The 6-week intervention was delivered entirely digitally, and all anthropometric measurements, questionnaires, and dietary data were self-collected at home. Feasibility was assessed using adherence metrics, completion of self-measurements, and a user-experience questionnaire. Secondary analyses evaluated weight loss, changes in anthropometry, and exploratory associations, including sex differences.

Results: Feasibility was high—102 out of 122 (83.6%) participants found the self-measured anthropometric assessments easy to perform, and 112 (91.8%) participants reported that completing questionnaires from home was easy. For secondary outcomes, participants receiving personalized, but not generic, nutritional advice significantly lost body weight (–1.0 kg; P=.002). Participants receiving personalized food boxes in addition to personalized nutritional advice lost significantly more body weight than the other 2 groups (–2.5 kg; P=.001) and also showed a decrease in hip circumference (–2.9 cm; P=.01). Personalized advice was not easier or more enjoyable to implement than generic nutritional advice, whereas the addition of personalized food boxes improved the ease of implementing personalized nutritional advice (P<.001). All participants, irrespective of the intervention arm, reduced intake of unhealthy food groups, including ready-made meals (113.6 g vs 78.5 g, –30.9%); sauces and gravy (18.8 g vs 10.0 g, –46.8%); sweet snacks (84.8 g vs 64.1 g, –24.4%); savory snacks (50.5 g vs 40.0 g, –20.1%); bread, pasta, rice, and wraps (nutritional quality score of 1.9 vs 1.7, –10.5%); and vegetables (129.0 g vs 118.7 g, –8.0%); and replaced coffee with tea.

Conclusions: This study demonstrates that fully remote, participant-led nutritional intervention studies are feasible, with participants able to independently perform anthropometric measurements and self-report data of sufficient quality to detect meaningful effects. Personalized nutritional advice resulted in greater weight loss than generic advice, and the addition of

XSL•FO
RenderX

personalized food boxes further enhanced the beneficial anthropometric effects of the intervention. We conclude that such nutritional intervention studies can be conducted fully online, resulting in measurable anthropometric effects after 6 weeks.

Trial Registration: ClinicalTrials.gov NCT06547983; https://clinicaltrials.gov/ct2/show/NCT06547983

(J Med Internet Res 2026;28:e73367) doi:10.2196/73367

KEYWORDS

personalized nutrition; digital; dietary advice; behavior; feedback; digital first; real-world study; remote; do it yourself; DIY

Introduction

Decentralization of clinical trials is on the rise, meaning that participant inclusion and data collection are performed by the participants themselves, for example, from their homes. This may help attenuate the burden of trial participation, improve participant recruitment, and decrease study dropout [1]. In this way, decentralization helps to avoid insufficient recruitment and underpowered clinical trials [2,3], or complete trial failure [4,5]. Online recruitment and study inclusion may also create opportunities to reach a more diverse population, that is, a better representation of the complete population [6]. In addition, data collected through decentralized trials may represent real-world data, revealing whether interventions work when implemented in real-world settings.

Nutrition studies often require a large sample size, with possibly multiple follow-up measurements and multiple (control) groups. Therefore, decentralization is of interest in this field; however, this may also introduce novel caveats, since protocol adherence and, subsequently, data quality could be jeopardized, possibly leading to false conclusions. A limited number of nutritional studies have indicated that similar results can be produced at home compared with a controlled laboratory setting [7-9]. However, this topic remains understudied, and research is needed to assess participant experience and capability to perform simple measurements from home and fill in questionnaires.

Concomitantly, the paradigm of nutritional intervention studies aiming to improve dietary habits is shifting toward more personalized approaches. Personalized nutrition interventions are tailored for specific individuals based on their individual data, such as phenotype, nutritional habits, and biomarkers [10]. Ideally, personalized nutrition approaches also consider socioeconomic, behavioral, and cultural factors, as well as the food environment [11]. Decentralized studies conducted in real-world settings offer an opportunity to assess external validity and to explore how these contextual factors, such as the local food environment and broader food systems [12], influence intervention effectiveness. Moreover, they provide a unique opportunity to investigate behavioral patterns within the food environment and the food systems framework.

Studies have shown that such personalized nutritional advice is more effective than a generic one-size-fits-all approach in improving dietary habits [10,13] and in promoting weight loss in individuals with obesity [14,15]. However, more research is needed in this field to understand the contexts in which personalized nutrition is most effective [14]. Some studies have suggested that personalized nutrition is more effective than generic advice due to its face-to-face aspect and its perception as personally relevant [16,17]. Consequently, it can be questioned whether personalized nutrition interventions are as effective when delivered through the internet, possibly failing to provide recipients with the same level of motivation. Therefore, we chose to test the effectiveness of personalized versus generic nutritional interventions in this real-life, fully remote study.

In summary, the primary objective of this study was to evaluate the feasibility of conducting a fully remote, fully digital nutritional intervention, including participants' experience and their ability to perform do-it-yourself anthropometric measurements from home. In addition, feasibility was evaluated by examining the reliability of self-reported anthropometric data through analyses of correlations among changes in body weight, waist circumference, and hip circumference. Feasibility was defined as the capability of participants to independently perform all study procedures, including self-measured anthropometry and digital questionnaire completion, in a fully remote setting. Secondary objectives were to assess whether significant changes in body weight and other anthropometric outcomes could be detected using self-collected data, and to compare the effectiveness of generic versus personalized nutritional advice delivered digitally. In addition, we explored whether the provision of personalized food boxes enhanced adherence and led to greater weight loss, and we conducted exploratory analyses of sex differences.

Methods

Ethics Statement

This study was reviewed and approved by the independent Internal Ethical Review Board of The Netherlands Organization for Applied Scientific Research (TNO), registration number #2023-083, in September 2023. The study was registered at ClinicalTrials.gov (NCT06547983). All participants provided written informed consent before inclusion in the study. The study was conducted in accordance with the Declaration of Helsinki. All analyses presented in this paper were conducted in accordance with the primary approved study protocol. All data were anonymized. Each participant received a Christmas food box (valued at €90 [US $106]) as compensation for their participation in the trial. Additionally, depending on group allocation, some participants received free daily meals during the trial.

Study Population

Participants were 25-60 years of age and had a BMI between 25 and 40 kg/m^2. Participants were included only if they were motivated to lose weight; had the skills to complete digital

XSL•FO
RenderX

questionnaires; and had a computer and a smartphone, a weighing scale, and a measuring tape to perform the anthropometric measurements, as confirmed in the screening questionnaire (Multimedia Appendix 1). Participants were excluded if they had a food allergy, followed specific diets (eg, a keto or vegan diet), suffered from a chronic disease that influences food intake (eg, inflammatory bowel disease), or participated in another intervention study. Females in menopause transition, as defined through the screening questionnaire, were excluded as well. A formal power calculation was not feasible because reliable estimates of effect size and variance for this specific intervention and outcomes were not available at the time of study design. Instead, the sample size was based on the results of previous studies [13,18,19], in which 37, 59, or 82 participants were included and exposed to personalized nutritional advice for 16, 9, or 10 weeks, respectively. Consequently, a sample size of 150 at the start of the intervention was considered acceptable for this study. This number was higher than those in the cited previous studies due to the shorter intervention period (6 weeks), which was expected to require a larger sample size. To compensate for an anticipated 10%-20% dropout rate during the intervention, a total sample size of 180 participants was selected.

Participant recruitment was organized via advertisements on social media, as well as through the Dutch participant database of Norstat [20], focusing on zip codes to which food boxes could be delivered (Uitgekookt Meal Service). Only individuals who indicated that their motivation to participate in the study was to lose weight were selected (participants could select 1 of 4 predefined reasons). These participants were randomly allocated to 1 of 3 intervention groups, while taking into account the covariates biological sex, age, and BMI. Allocation was performed by calculating the Euclidean distance for the variable sex and the Kullback-Leibler divergence for the variables age and BMI for each permutation separately. The permutation with the best overall score (ie, the most balanced distribution among the groups) was selected. This allocation process was automated and blinded to both participants and investigators involved in data collection and analysis.

Study Design

The primary objective of this study was to evaluate the feasibility of conducting a fully remote, fully digital nutritional intervention trial, including participants' ability to perform do-it-yourself anthropometric measurements, adherence to the protocol, and user experience. To further evaluate feasibility, the reliability of self-reported anthropometric data was examined by analyzing correlations among changes in body weight, waist circumference, and hip circumference. Secondary objectives were to assess whether self-collected data could detect significant changes in body weight and other anthropometric measures, and to compare the effectiveness of generic versus personalized nutritional advice, with and without personalized food boxes. Exploratory analyses included examining sex differences to assess the reliability of the self-assessed anthropometric data. The study was a randomized controlled trial with 1 control intervention arm and 2 personalized intervention arms (Figure 1). The study duration was 6 weeks. Participants allocated to the control group received generic nutritional advice (see example in Multimedia Appendix 2), which was based on the Dutch national guidelines for healthy nutrition, assembled by the Dutch Health Council (Gezondheidsraad), an independent scientific advisory body for the Dutch government.

Figure 1. Overview of the study design. Self-measured anthropometric data, including body weight, and assessments of fiber intake and physical activity (via questionnaires) were collected through the online How Am I app at weeks 0, 2, 4, and 6. All intervention groups received feedback on their self-assessments at weeks 0, 2, 4, and 6. Nutritional advice provided to intervention groups P and PP was updated based on participants' self-assessments. Qualitative assessments of study experience were collected at weeks 2, 4, and 6. Habitual food intake was assessed at the beginning and end of the trial using the Nutri+ module. P-group: personalized intervention group; PP-group: personalized plus intervention group.

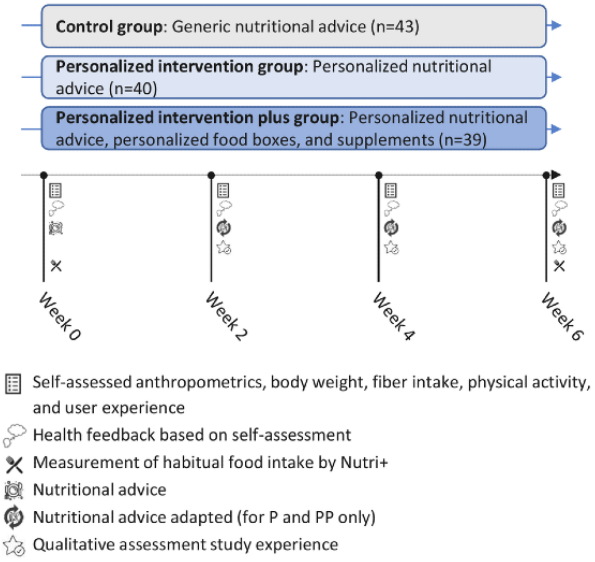

XSL•FO
RenderX

Study Intervention

Participants allocated to the personalized intervention group (P-group) received personalized advice instead of generic advice. The personalized advice entailed information on the specific number of calories to consume on a daily basis to meet their personal goal of weight loss. Furthermore, the proportion of macronutrients the participant should consume to reach their weight loss goal was also made explicit in the nutritional advice (see example in Multimedia Appendix 2). The prescribed caloric intake provided a daily energy deficit of 400 kcal below the daily energy requirement, as calculated using the Ten Haaf Formula [21] at the start of the trial, based on personal information including age, sex, and physical activity level (PAL). In addition, instructions were provided for the implementation of the Eetmeter application (Voedingscentrum [22]), a freely available digital application that participants were advised to use to check whether their daily meals matched their personalized nutritional advice according to the Dutch dietary guidelines.

Participants allocated to the personalized plus intervention group (PP-group) received the same personalized advice as the P-group; however, they also received personalized food boxes containing meals that were composed to match the personalized nutritional advice in terms of calories and macronutrients. These food boxes contained 3 main meals (breakfast, lunch, and dinner) and 3 snacks to be consumed between the main meals (Uitgekookt Meal Service, IJsselmuiden). The food boxes provided a daily energy deficit of 200-600 kcal below the daily energy requirement. The food boxes were provided on 5 out of 7 days per week. On the remaining 2 days, participants were instructed to use the online Eetmeter application to select and follow their diet based on the provided personalized advice.

Participants allocated to groups receiving personalized interventions received updated nutritional advice based on the outcomes of the self-measurements in terms of calories and macronutrients, whereas the control group received the same generic advice at weeks 2, 4, and 6.

Remote Data Collection

The study was fully digital, without any face-to-face contact between participants and researchers. To this end, participants were required to install the online research application, the How Am I app (TNO), through which all instructions, feedback, nutritional advice, questionnaires, and reminders were provided, and which served as the data collection platform [23]. At weeks 0, 2, 4, and 6, participants submitted data via the How Am I app on feasibility-related user experience items, self-measured anthropometric measurements, body weight, and brief questions related to fiber intake and physical activity (Multimedia Appendix 1). PAL and fiber intake were calculated according to validated methodologies, as described by Healey et al [24] and in the Food and Agriculture Organization (FAO)/World Health Organization (WHO)/United Nations University (UNU) Expert Consultation Report [25]. At weeks 2, 4, and 6, all participants received a reminder via the TNO How Am I app to complete all anthropometric measurements and questions and were provided with feedback on their measurements (see Multimedia Appendix 2). Researchers monitored data entry, issuing a reminder to participants who did not complete their questionnaires within 48 hours and instructing them to respond within an additional 24-hour period. Participants who remained noncompliant after this interval were excluded from the study.

Feasibility Outcomes

The primary outcome of this study was the feasibility of conducting a fully remote, fully digital nutritional intervention. Feasibility was assessed at weeks 2, 4, and 6 through participants' reported user experience, including ease of performing anthropometric measurements, clarity of instructions, and perceived burden, measured using multiple-choice questions, Likert scales, and visual analog scales (see Multimedia Appendix 1). Feasibility was further evaluated by examining the reliability of self-reported anthropometric data, based on correlations among changes in body weight, waist circumference, and hip circumference.

Self-Measured Anthropometrics and Body Weight

Secondary outcomes related to anthropometrics and body weight were also evaluated. At baseline and at weeks 2, 4, and 6, all participants were requested to self-measure body weight (kg). Using a measuring tape, participants were asked to measure height (cm), waist circumference (cm), and hip circumference (cm) according to standard operating procedures [26]. Detailed instructions on performing the anthropometric measurements were provided via the TNO How Am I app through videos containing step-by-step guidance. Body weight and height were used to calculate BMI (kg/m^2).

Measurement of Habitual Dietary Intake

At the start and end of the trial, all participants completed the Nutri+ module (TNO), a web-based questionnaire consisting of 89 questions that measures the dietary intake of 21 food groups for each participant and thereby estimates habitual food intake. Consequently, the average food intake for certain food groups (eg, fruit, meat, dairy) or the average frequency of consumption per week (eg, bread, potatoes, cooking fats) was calculated.

Statistical Analysis

All analyses were conducted using GraphPad Prism 10 (GraphPad Software) or IBM SPSS Statistics 29. Outlier analysis was performed, and data points greater than 3× the IQR were excluded. As the study aimed to evaluate intervention efficacy based on complete longitudinal data, analyses were performed per protocol. Participants with incomplete datasets (eg, due to dropping out) were not included in the analysis, as missing repeated-measures data would have violated model assumptions and precluded valid within-participant comparisons over time. Normality was confirmed using the D'Agostino and Pearson test. All data are reported as means (SD). Two-tailed P values were used, and P values <.05 were considered statistically significant.

User experience data collected via Likert scales, multiple-choice questions, and visual analog scales were analyzed descriptively. Group differences in categorical or ordinal feasibility responses were evaluated using Pearson chi-square tests, with Bonferroni-adjusted post hoc tests applied when appropriate.

XSL•FO
RenderX

The internal validity of self-reported anthropometric measurements was examined through Pearson correlations among changes in body weight, waist circumference, and hip circumference.

Differences between groups at baseline were tested using one-way ANOVA with Tukey post hoc tests. Anthropometric outcomes collected across multiple time points (weeks 0, 2, 4, and 6) were analyzed using repeated-measures ANOVA, with "time" as a within-participant factor and "group" as a between-participant factor. The Tukey post hoc tests were performed when main effects or interactions were detected. Categorical variables were compared using Pearson chi-square tests with Bonferroni corrections. Sex-stratified analyses were also performed to explore potential sex differences in anthropometric changes by testing for significant interaction effects with sex. Correlations between variables were assessed using the Pearson correlation coefficient, and statistical significance was defined as $P<.05$.

Results

Baseline Study Population Characteristics

A total of 315 participants were screened, and of these, 91 were excluded from participation for various reasons (Figure 2). The remaining 224 participants were invited to participate in the study, of whom 180 responded positively and were included in the study and randomly allocated to 1 of the 3 groups. Of the 180 recruited participants, 43 did not respond to the questionnaires and were consequently excluded from the study (Figure 2). Data from the remaining 137 participants were included in the analysis. An additional 15 participants were excluded based on outlier analysis, resulting in a final sample size of 122 participants. Results are reported in accordance with CONSORT (Consolidated Standards of Reporting Trials)-EHEALTH recommendations (Multimedia Appendix 3).

Figure 2. A flowchart of the recruitment and selection process, randomization, and group allocation of included participants.

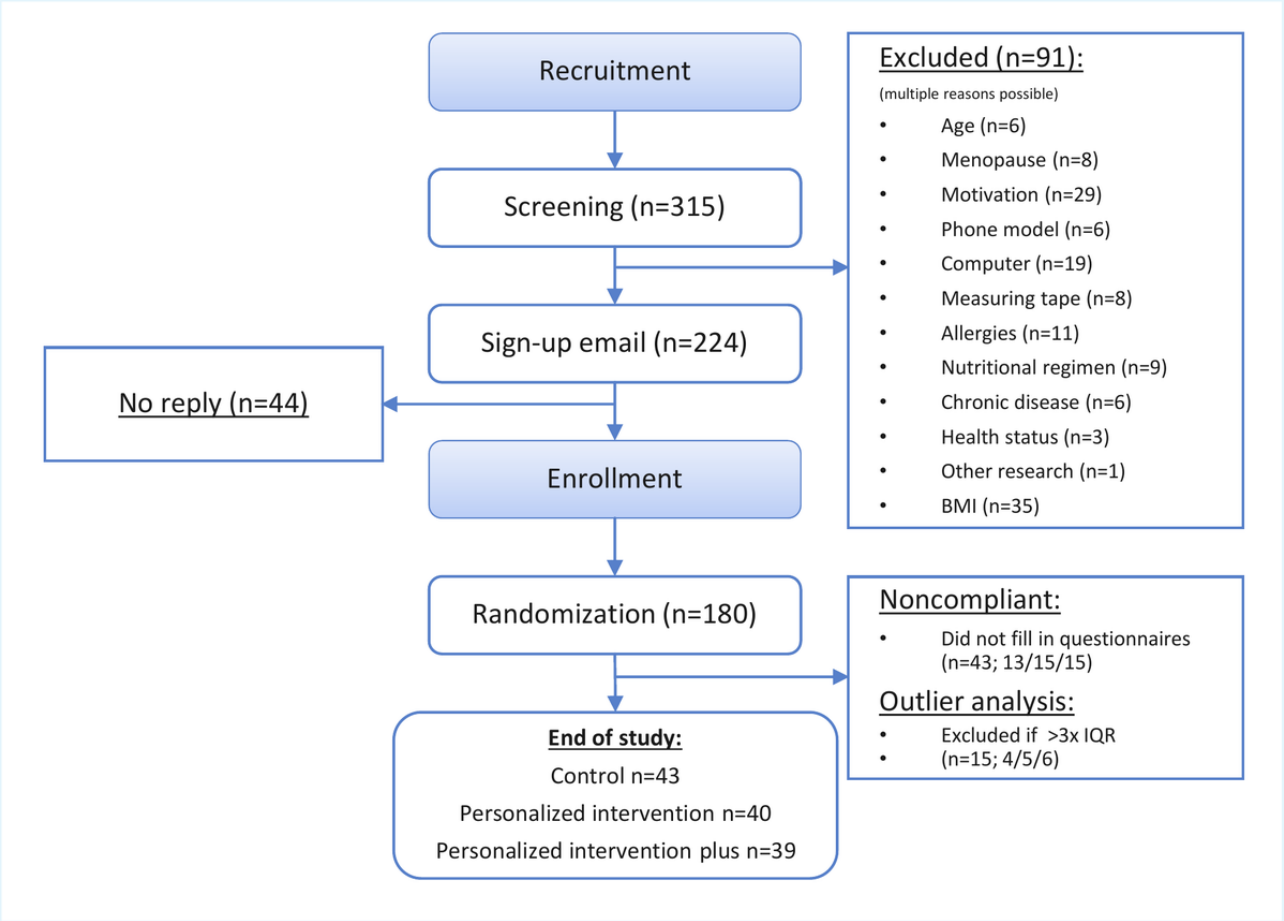

No statistically significant differences in any baseline characteristics were detected among the 3 intervention groups (Table 1). Habitual food intake was also measured for each food group before the intervention (Table 2). No statistically significant differences in habitual food intake were detected among the 3 intervention groups for any food group, except for sauces and gravy. Similarly, no statistically significant differences were found in baseline PAL among the 3 groups (Table 3; $P=.15$).

XSL•FO
RenderX

Table 1. Baseline characteristics of the study population.

Characteristics	Control (n=43)	Personalized intervention (n=40)	Personalized plus intervention (n=39)	*P* value
Age (years), mean (SD)	39.9 (6.1)	41.0 (6.0)	40.9 (5.9)	.60
Female sex, n (%)	27 (63)	27 (68)	19 (49)	.18
Weight (kg), mean (SD)	89.9 (12.0)	89.5 (13.3)	92.4 (14.1)	.65
Height (m), mean (SD)	1.75 (0.09)	1.75 (0.11)	1.75 (0.10)	.54
BMI (kg/m^2), mean (SD)	30.4 (4.1)	30.2 (4.5)	31.2 (4.8)	.81
Waist circumference (cm), mean (SD)	99.0 (10.1)	101.3 (10.0)	103.0 (12.4)	.26
Hip circumference (cm), mean (SD)	108.3 (10.1)	108.3 (7.5)	110.8 (8.9)	.36
Self-perceived health (out of 5 points, 5 being the highest score), mean (SD)	3.0 (0.6)	3.2 (0.7)	3.0 (0.8)	.48
Physical activity level, mean (SD)	1.40 (0.2)	1.44 (0.1)	1.40 (0.1)	.15
Fiber intake (g), mean (SD)	11.3 (5.2)	12.4 (8.4)	12.5 (10.6)	.99

Table 2. Baseline habitual food intake (based on the Nutri+ module) per food group of the study population.[a]

Food group	Control (n=43), mean (SD)	Personalized intervention (n=40), mean (SD)	Personalized plus intervention (n=39), mean (SD)	*P* value
Fruit (g/day)	141.6 (110.3)	136.7 (130.6)	113.3 (95.4)	.53
Vegetables (g/day)	122.3 (47.6)	132.6 (56.7)	125.2 (63.4)	.69
Legumes (g/week)	172.3 (284.0)	148.1 (202.1)	151.2 (161.6)	.91
Bread (nutritional quality score)	1.8 (0.6)	1.8 (0.7)	2.1 (0.7)	.09
Pasta, rice, and wraps (nutritional quality score)	1.8 (0.9)	1.9 (0.9)	2.1 (0.7)	.23
Potatoes (nutritional quality score)	2.4 (1.5)	2.7 (1.7)	2.2 (1.3)	.36
Meat (g/week)	2367.9 (1737.8)	2422.8 (1336.4)	2584.9 (1650.3)	.79
Fish (g/week)	348.8 (701.5)	223.1 (242.1)	218.9 (232.9)	.40
Eggs (g/week)	209.5 (305.1)	263.8 (210.6)	233.8 (163.3)	.60
Dairy (g/day)	314.6 (573.5)	308.3 (301.5)	197.3 (193.3)	.39
Nuts (g/day)	15.1 (22.7)	12.9 (16.4)	8.3 (10.3)	.24
Spreadable fats (nutritional quality score)	1.1 (0.9)	1.2 (1.0)	1.1 (0.8)	.94
Cooking fats (nutritional quality score)	1.5 (0.6)	1.3 (0.7)	1.4 (0.8)	.47
Sweet snacks (g/week)	77.6 (83.3)	83.4 (83.9)	94.3 (68.6)	.64
Savory snacks (g/week)	55.6 (74.0)	41.8 (37.4)	54.3 (49.4)	.39
Ready-made meals (g/week)	127.3 (218.5)	112.9 (85.9)	98.8 (97.2)	.73
Sauces and gravy (g/week)	15.0 (20.6)[b]	24.5 (22.8)[b]	17.0 (17.1)[b]	.02
Alcohol (units drank/day)	0.7 (1.1)	0.7 (1.2)	0.5 (0.8)	.82
Soft drinks and fruit juice (g/day)	209.4 (374.2)	176.1 (272.0)	278.3 (489.4)	.49
Tea (g/day)	453.1 (448.1)	360.7 (390.7)	400.2 (399.0)	.63
Coffee (g/day)	424.7 (476.0)	338.8 (378.1)	403.5 (451.5)	.67

[a]Data represent the average habitual food intake. Nutritional quality scores reflect factors such as whether whole wheat versus multigrain products were consumed or whether olive oil versus butter is used for frying. A lower score represents a healthier food choice. Extensive descriptions of these nutritional quality scores are attached as Multimedia Appendix 4.

[b]Control versus personalized intervention, *P*=.02; control versus personalized plus intervention, *P*=.54; personalized intervention versus personalized plus intervention, *P*=.23.

XSL•FO
RenderX

Table 3. Physical activity levels during the study period.

Physical activity level	Week 0	Week 2	Week 4	Week 6
Control group, mean (SD)	1.40 (0.2)	1.40 (0.2)	1.42 (0.2)	1.41 (0.2)
Personalized intervention, mean (SD)	1.44 (0.1)	1.45 (0.1)	1.44 (0.1)	1.45 (0.1)
Personalized plus intervention, mean (SD)	1.40 (0.1)	1.41 (0.1)	1.40 (0.1)	1.40 (0.1)

Participant-Reported Usability and Reliability of Measurements

The ease of collecting anthropometric data by participants themselves and completing the questionnaires digitally via the How Am I app was reported (Multimedia Appendix 1). The majority of participants indicated that they could perform the anthropometric measurements very easily or easily (102/122, 83.6%, participants across all groups; Figure 3A), with no differences among groups. In addition, most participants reported that they could understand the questionnaires (112/122, 91.8%, participants across all groups; Figure 3B). The quality of anthropometric data was cross-validated, as significant correlations were detected between changes in body weight, waist circumference, and hip circumference (week 6 to week 0; all $P<.001$, $r>0.34$; Figure 3C-3E).

Figure 3. Ease and ability to perform anthropometric measurements at home and to complete questionnaires. (A) Ease of independently performing anthropometric measurements. (B) Clarity of the questionnaires. (C) Correlation between Δ body weight and Δ waist circumference. (D) Correlation between Δ body weight and Δ hip circumference. (E) Correlation between Δ waist and Δ hip circumference. In all correlation analyses, Δ values were calculated as the difference between measurements at the beginning and end of the 6-week trial. P-group: personalized intervention group; PP-group: personalized plus intervention group.

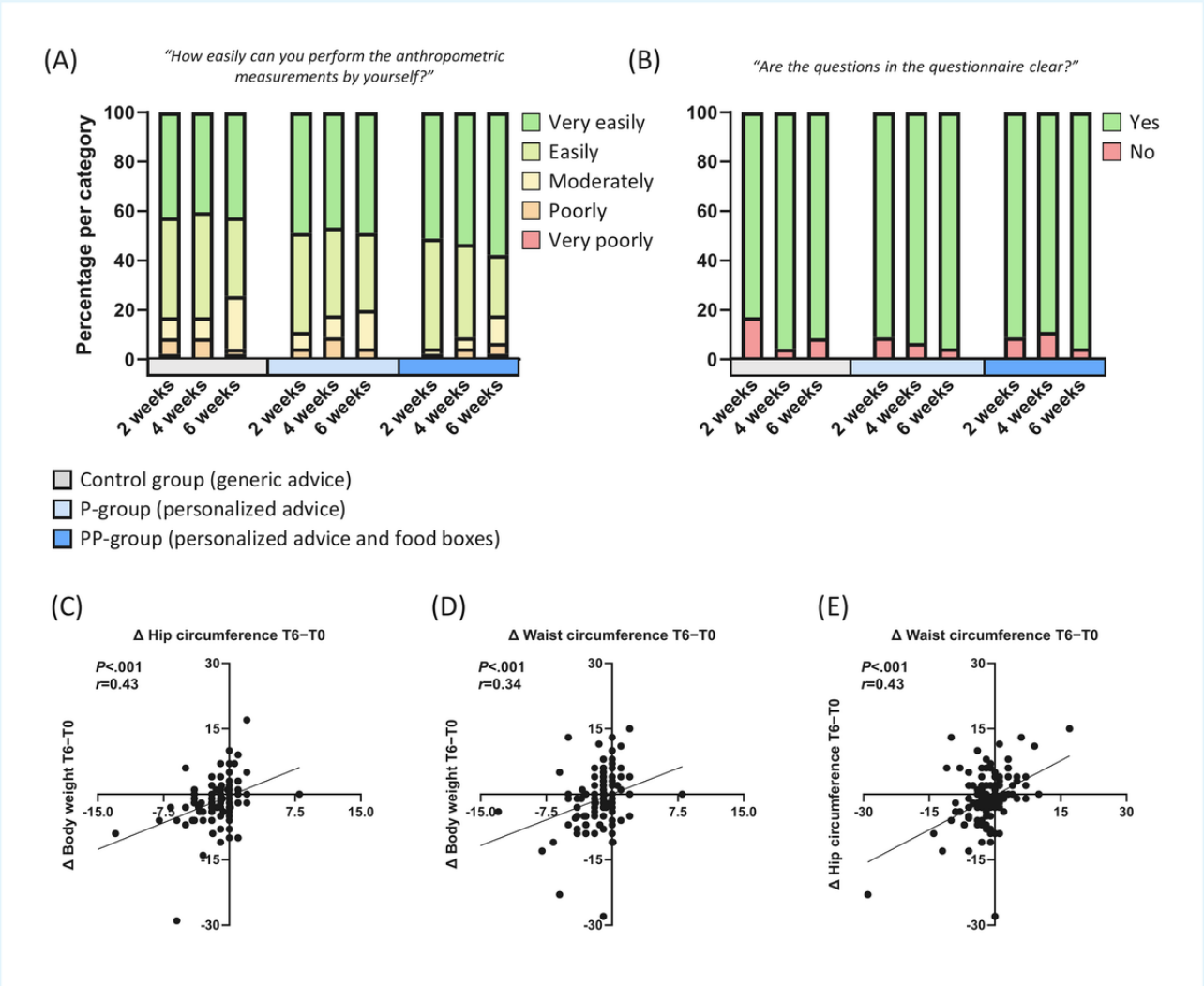

Trial Experience and Self-Reported Dietary Changes

No significant differences were detected in the experienced ease of implementing generic versus personalized nutritional advice during the trial ($P=.43$; Figure 4A). However, participants in the PP-group, who received personalized food boxes in addition to the personalized nutritional advice, reported that they could implement the nutritional advice more easily ($P<.001$; Figure 4A). At week 4 only, participants in the PP-group reported experiencing a greater amount of pleasure while implementing

XSL•FO
RenderX

their nutritional advice (*P*<.001 at week 4; Figure 4B), whereas no differences were observed between the groups receiving generic versus personalized nutritional advice. No differences in self-reported changes in diet were detected between the control group and the P-group (Figure 4C), while the PP-group consistently reported significant changes in their diet during the trial (*P*<.001; Figure 4C). Similarly, the PP-group reported significantly more often that they assessed their diet as being healthier at week 2 (*P*=.01; Figure 4D).

Figure 4. Participants' experience of participating in the fully digital trial. (A) Ease of implementing the nutritional advice during the trial. (B) Enjoyment of implementing the nutritional advice during the trial. (C) Self-reported changes in participants' diets during the trial. (D) Self-reported improvements in dietary healthfulness during the trial. **P*<.05, ***P*<.01, ****P*<.001. P-group: personalized intervention group; PP-group: personalized plus intervention group.

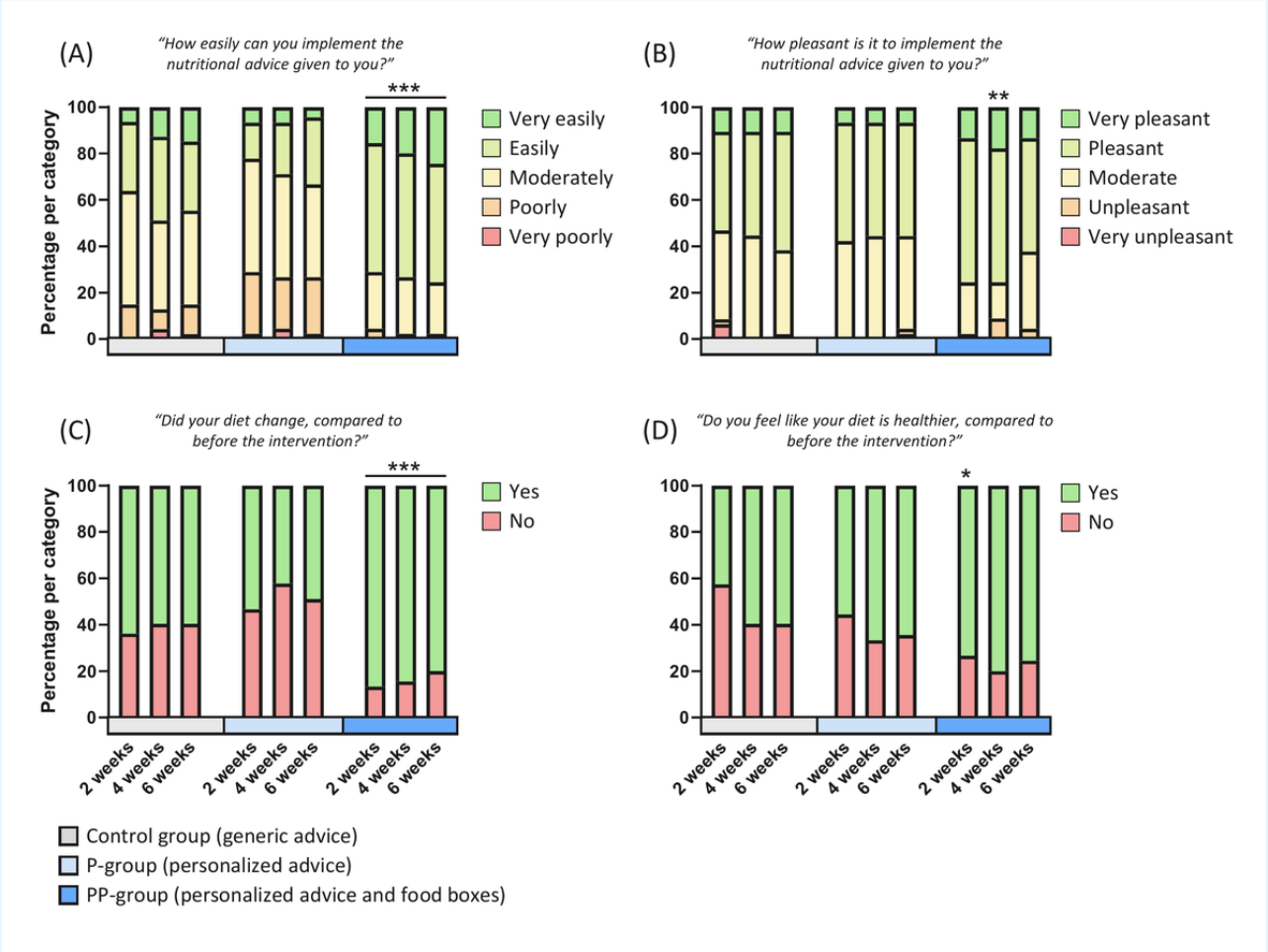

Changes in Body Weight, BMI, and Anthropometric Measures

After 6 weeks, the average body weight of the control group did not change compared with the start of the trial (*P*=.06; Figure 5A). In the P-group, body weight was significantly lower at week 4 (–0.8 kg; *P*=.04) and week 6 (–1 kg; *P*=.002) compared with baseline (Figure 5A). In the PP-group, body weight was significantly lower at week 2 (–0.9 kg; *P*=.004), week 4 (–1.6 kg; *P*=.002), and week 6 (–2.3 kg; *P*=.001) compared with baseline (Figure 5A). With respect to BMI, similar differences between the groups were observed (Figure 5B). BMI decreased significantly more in the PP-group (–0.8 kg/m^2, *P*=.001) compared with the P-group (–0.3 kg/m^2; *P*=.002) and the control group (–0.2 kg/m^2; *P*=.06) at week 6 (Figure 5B).

XSL•FO
RenderX

Figure 5. Effects of the different interventions on anthropometric measurements. (A) Body weight, (B) BMI, (C) hip circumference, and (D) waist circumference during the trial. Data are expressed as means (SDs). **P*<.05 and ***P*<.01. The $ symbols indicate significant differences attributable to a statistically significant time effect in the absence of an interaction effect. P-group: personalized intervention group; PP-group: personalized plus intervention group.

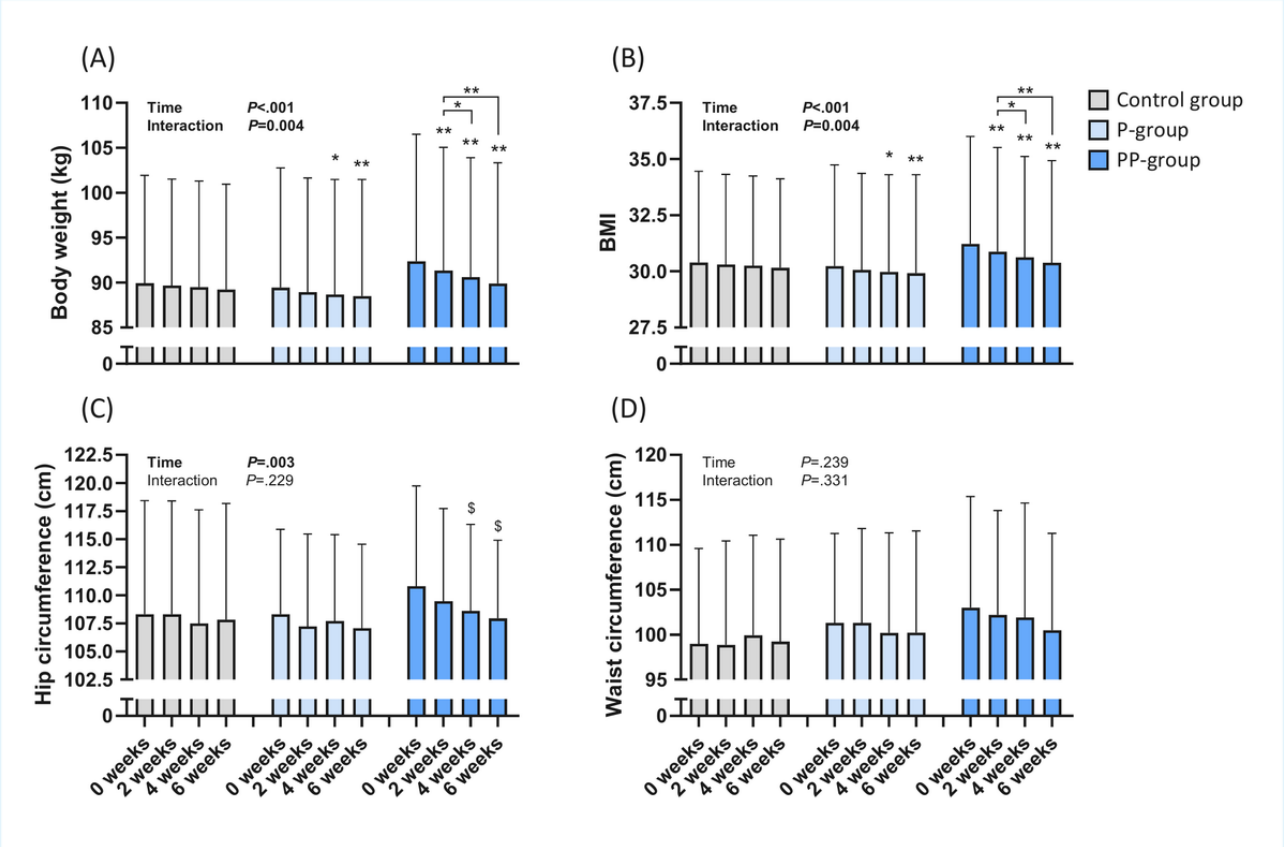

Hip circumference decreased significantly over time (*P*=.01), but no significant interaction between time and group was detected (*P*=.22; Figure 5C). Post hoc analysis revealed that the time effect was mainly driven by a significant loss of hip circumference in the PP-group, which decreased at week 4 (–2.2 cm; *P*=.04) and week 6 (–2.9 cm; *P*=.01) compared with baseline (Figure 5C). No significant effects were observed on waist circumference ($P_{interaction}$ =.33, P_{time}=.24; Figure 5D).

Changes in Habitual Food Intake and Physical Activity

Changes in intake of the food groups vegetables, bread, pasta/rice/wraps, sweet snacks, savory snacks, ready-made meals, sauces and gravy, tea, and coffee were observed (Table 4). The intake of all these food groups was lower after 6 weeks compared with baseline (vegetables, –9.0 g; bread, –0.2 times eaten per week; pasta/rice/wraps, –0.2 times eaten per week; sweet snacks, –21.1 g; savory snacks, –10.7 g; ready-made meals, –60.3 g; sauces and gravy, –8.7 g; coffee, –62.1 g), except tea, which significantly increased (+64.5 g; Table 4). No changes were detected for the food groups fruit, legumes, potatoes, meat, fish, eggs, dairy, nuts, spreadable and cooking fats, alcohol, or soft drinks and fruit juice. Only time effects were significant (see Table 4), without interaction effects, implying that changes in food intake occurred across all intervention groups, but no group-specific effects were observed. Only the food group legumes nearly reached statistical significance for both personalized interventions compared with the control group (*P*=.051), with increased intake postintervention for both the P- and PP-groups, whereas the control group, on average, showed decreased intake postintervention. Fiber intake was measured every 2 weeks, but no changes were detected (Multimedia Appendix 5). In addition, PALs did not change during the study (Table 3).

XSL•FO
RenderX

Table 4. Habitual food intake per food group before and after the intervention.a

Food group	Control (preintervention), mean (SD)	Control (postintervention), mean (SD)	Personalized intervention (preintervention), mean (SD)	Personalized intervention (postintervention), mean (SD)	Personalized plus intervention (preintervention), mean (SD)	Personalized plus intervention (postintervention), mean (SD)	Time effect (*P* value)	Interaction effect (*P* value)
Fruit (g/day)	141.6 (110.3)	142.5 (101.4)	136.7 (130.6)	128.8 (136.5)	113.3 (95.4)	105.5 (82.4)	.64	.92
Vegetables (g/day)	122.3 (47.6)	113.4 (44.0)	132.6 (56.7)	116 (52.0)	125.2 (63.4)	123.6 (57.6)	.03[b]	.35
Legumes (g/week)	172.3 (284.0)	121.7 (166.8)	148.1 (202.1)	172.0 (266.3)	151.2 (161.6)	183.6 (225.3)	.90	.051
Bread (nutritional quality score)	1.8 (0.6)	1.6 (0.7)	1.8 (0.7)	1.7 (0.7)	2.1 (0.7)	1.9 (0.6)	.02[b]	.37
Pasta, rice, and wraps (nutritional quality score)	1.8 (0.9)	1.6 (0.8)	1.9 (0.9)	1.5 (0.9)	2.1 (0.7)	1.9 (0.8)	<.01[b]	.57
Potatoes (nutritional quality score)	2.4 (1.5)	1.6 (1.6)	2.7 (1.7)	1.5 (1.5)	2.2 (1.3)	1.9 (1.1)	.71	.97
Meat (g/week)	2367.9 (1737.8)	2428.8 (1809.2)	2422.8 (1336.4)	2473.0 (1605.4)	2584.9 (1650.3)	2280.3 (1300.9)	.65	.50
Fish (g/week)	348.8 (701.5)	337.5 (710.7)	223.1 (242.1)	229.4 (267.8)	218.9 (232.9)	257.4 (237.1)	.61	.64
Eggs (g/week)	209.5 (305.1)	211.9 (265.7)	263.8 (210.6)	248.8 (250.0)	233.8 (163.6)	212.2 (200.1)	.40	.76
Dairy (g/day)	314.6 (573.5)	327.0 (800.3)	308.3 (301.5)	284.0 (383.7)	197.3 (193.3)	159.1 (172.0)	.49	.68
Nuts (g/day)	15.1 (22.7)	16.2 (19.9)	12.9 (16.4)	15.3 (18.3)	8.3 (10.3)	11.7 (13.1)	.11	.80
Spreadable fats (nutritional quality score)	1.1 (0.9)	1.1 (0.8)	1.2 (1.0)	1.1 (0.8)	1.1 (0.8)	1.0 (0.8)	.65	.91
Cooking fats (nutritional quality score)	1.5 (0.6)	1.2 (0.7)	1.3 (0.7)	1.3 (0.8)	1.4 (0.8)	1.3 (0.8)	.32	.41
Sweet snacks (g/week)	77.6 (83.3)	64.5 (89.6)	83.4 (83.9)	65.4 (56.4)	94.3 (68.6)	62.2 (60.6)	<.01[b]	.59
Savory snacks (g/week)	55.6 (74.0)	48.0 (67.6)	41.8 (37.4)	35.0 (37.0)	54.3 (49.4)	36.5 (47.8)	<.01[b]	.47
Ready-made meals (g/week)	127.3 (218.5)	77.4 (114.7)	112.9 (85.9)	78.2 (71.2)	98.8 (97.2)	80.0 (88.3)	.01[b]	.66
Sauces and gravy (g/week)	15.0 (20.6)	6.7 (9.1)	24.5 (22.8)	13.8 (18.2)	17.0 (17.1)	9.9 (14.3)	<.01[b]	.70
Alcohol (U/day)	0.7 (1.1)	0.7 (1.2)	0.7 (1.2)	0.5 (0.9)	0.5 (0.8)	0.5 (0.7)	.66	.66
Soft drinks and fruit juice (g/day)	209.4 (374.2)	212.1 (345.9)	176.1 (272.0)	204.8 (341.9)	278.3 (489.4)	218.3 (329.9)	.77	.66
Tea (g/day)	453.1 (448.1)	536.6 (504.6)	360.7 (390.7)	403.7 (433.0)	400.2 (399.0)	467.2 (464.2)	.03[b]	.95
Coffee (g/day)	424.7 (476.0)	396.3 (449.2)	338.8 (378.1)	303.6 (384.7)	403.5 (451.5)	280.9 (277.5)	.02[b]	.20

[a]Nutritional quality scores reflect factors such as whether whole wheat versus multigrain products were consumed or whether olive oil versus butter is used for frying. A lower score represents a healthier food choice. Extensive descriptions of these nutritional quality scores are attached as Multimedia Appendix 4.

[b]Significant values (ie, <.05).

XSL•FO
RenderX

Sex Differences and Correlations in Body Composition Changes

Exploratory analyses revealed that the waist-to-hip ratio was higher in males compared with females (Figure 6A). Furthermore, weight loss correlated similarly with loss in hip circumference in females (r=0.48, P<.001) and males (r=0.44, P<.001), as no significant interaction with biological sex was detected (P=.15; Figure 6B). Body weight loss correlated differently with loss in waist circumference in females (r=0.26, P=.02) and males (r=0.44, P<.001), as a significant interaction with biological sex was detected (P=.01; Figure 6C).

Figure 6. Sex differences in anthropometric measurements and fat mass loss. (A) Sex differences in waist-to-hip ratio at week 0, assessed fully online; data are expressed as means (SDs). Sex-specific correlations between Δ body weight and Δ waist circumference (B) and Δ hip circumference (C). In all correlation analyses, Δ values were calculated as the difference between measurements at the start and end of the 6-week trial. *P<.05.

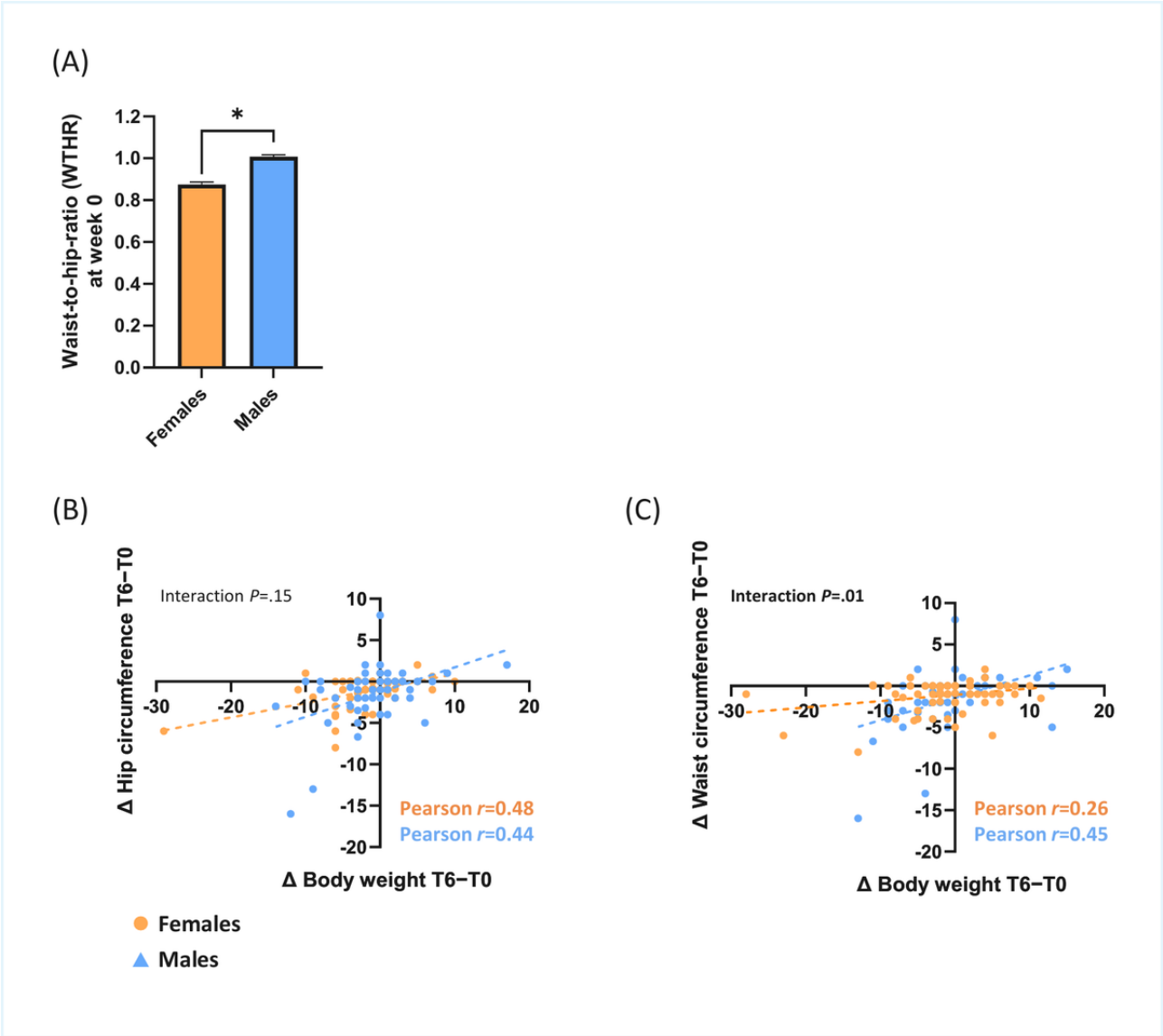

Discussion

Principal Findings

The primary aim of this study was to evaluate the feasibility of conducting a fully remote, fully digital nutritional intervention in adults with overweight or obesity. Our findings demonstrate that such a digital-first approach is feasible: most participants reported that anthropometric measurements were easy to perform at home, the digital questionnaires were clear and understandable, and the reliability of self-reported measurements was supported by consistent correlations among changes in body weight, waist circumference, and hip circumference (Figure 3). Together, these results indicate that participants were able to self-collect anthropometric data of sufficient quality for use in a randomized controlled trial.

The secondary aim of the study was to assess whether meaningful anthropometric changes could be detected using self-collected data. Participants who received personalized nutritional advice lost more body weight than those receiving generic advice (Figure 5), confirming the greater effectiveness of personalized advice even in a fully digital format. This effect was not explained by differences in the ease of implementing the advice, which increased only when personalized food boxes were provided in addition to the advice (Figure 4). The addition of personalized food boxes further increased weight loss and reduced hip circumference more than either form of advice

XSL•FO
RenderX

alone. Taken together, these results indicate that nutrition studies can be effectively conducted in a real-life, fully remote setting, achieving measurable intervention effects while maintaining ease of participation and high-quality data collection.

Exploratory analyses provided additional insights into sex differences in anthropometric responses. As expected from prior literature, males had higher waist-to-hip ratios and showed stronger correlations between weight loss and reductions in waist circumference compared with females. These findings support the biological plausibility of the self-measured data and highlight potential sex-specific patterns in responsiveness to nutritional interventions.

Comparison With Prior Work

An important factor underlying the success of personalized nutrition is the tailoring of the information provided to the recipient [27], which makes the advice perceived as personally relevant, for which face-to-face contact could be an important contributor [16,17]. Digital delivery of advice lacks this potentially important component, which may reduce its effectiveness. Importantly, in this study, we show that although all advice was provided digitally without any face-to-face contact, personalized nutritional advice still led to a significant loss of body weight, whereas generic nutritional advice did not produce a statistically significant effect on body weight or other anthropometric measurements (Figure 5A). This finding is consistent with other studies comparing digitally provided generic versus personalized advice in populations with obesity [28-31]. Together, the findings of these studies and our study provide evidence for the added value of personalized nutritional advice compared with generic advice, even when delivered online. Yet, it remains a topic of discussion whether the same personalized nutritional advice would have been more effective if delivered in person, rather than online. Interestingly, one study compared the effect of delivering personalized nutritional advice online with or without a personal coach [32]. Participants receiving similar advice through an online coach demonstrated greater engagement with the program and had an increased likelihood of achieving clinically relevant weight loss. This finding suggests that receiving feedback from an actual human can enhance protocol adherence, possibly due to an increased sense of accountability, motivation, and perception of receiving care tailored to an individual's specific needs. Similarly, another study found that offline nutritional advice was more effective than online advice in a small sample of rugby athletes [25], highlighting the potential added benefit of in-person delivery [33]. However, conventional face-to-face counseling also has its downsides, as it is more expensive, less flexible in terms of time and location, and ultimately less scalable compared with online methods [34]. Consequently, both online and in-person approaches have their own strengths, and one may be more appropriate than the other depending on an individual's personal preferences and needs.

Interestingly, participants in the PP-group, who received personalized food boxes based on their personalized advice, were able to lose more body weight compared with the other two groups. This effect was not due to differences in physical activity, as PAL did not change in any of the groups (Table 3). This outcome may reflect the difficulty of translating personalized nutritional advice into an actual diet over a prolonged period. Indeed, these participants reported greater ease in following the personalized nutritional advice, as well as a greater perceived dietary change, compared with participants in the control and P-groups (Figure 4A and 4C). This novel finding suggests that future interventions may benefit from placing particular emphasis on helping participants translate their advice into practice, for example, by simplifying meal planning and reducing decision fatigue. These data also revealed that participants in the P-group did not find it easier or more enjoyable to implement their personalized advice compared with participants in the control group receiving generic advice (Figure 4A and 4B). However, participants in the P-group did achieve a significant loss of body weight, which was not observed in the control group (Figure 5A). These findings suggest that the success of personalized nutritional advice is not necessarily due to greater ease or enjoyment of implementation, which would be expected to increase protocol adherence. Instead, a better fit and higher quality of the advice are likely the main contributors to the enhanced effectiveness of personalized nutritional advice.

Habitual food intake was also measured, which indicated that all intervention groups changed their dietary habits in a similar way. All groups reduced the intake of ready-made meals (113.6 g vs 78.5 g, –30.9%); sauces and gravy (18.8 g vs 10.0 g, –46.8%); sweet snacks (84.8 g vs 64.1 g, –24.4%); savory snacks (50.5 g vs 40.0 g, –20.1%); bread, pasta, rice, and wraps (nutritional quality score of 1.9 vs 1.7, –10.5%); and vegetables (129.0 g vs 118.7 g, –8.0%); and replaced coffee with tea (Table 4). These results indicate that the intake of unhealthy food groups was particularly reduced, while the intake of healthy food groups such as fruit, legumes, nuts, and fish was maintained. Only vegetable intake decreased, but to a lesser extent (129.0 g vs 118.7 g, –8.0%). Fiber intake was also maintained and did not change (Multimedia Appendix 5). This suggests that all 3 intervention groups shifted their dietary intake toward a healthier habitual pattern, in line with the nutritional advice provided. No significant time × group interaction effects were found in the habitual food intake data. This is noteworthy, as the diet of the PP-group changed substantially—they received personalized food boxes for all meals throughout the study—while the control group and P-group did not receive such food boxes. Based on this approach, we hypothesized that the PP-group would show greater changes in habitual food intake compared with the other groups; however, these changes were not detected. This suggests that participants in the PP-group may have had difficulty accurately recalling the foods they consumed. In contrast to the habitual food intake data, the anthropometric data revealed group-specific effects of the nutritional interventions. For these reasons, we suspect that the collected habitual food intake data were not accurate enough to detect group-specific effects. Therefore, future studies may benefit from incorporating new technology-based dietary assessment tools that combine web-based programs with mobile apps and wearable devices [35]. Interestingly, legume intake showed a divergent pattern between groups, with a decrease in the control group and an increase in both intervention groups. Although this difference did not reach conventional statistical

XSL•FO
RenderX

significance (interaction P=.051), the trend suggests a potential effect of the personalized interventions. Participants in the P- and PP-groups received targeted advice regarding protein and macronutrient intake, which may have encouraged higher legume consumption, whereas the control group received only general dietary guidance. Moreover, although the between-person variability in habitual food intake was considerable, this is not unexpected in nutritional intervention studies. Such variability likely reflects differences in individual dietary habits, adherence, and reporting accuracy. This observation emphasizes the importance of personalized dietary assessment and highlights the inherent heterogeneity in dietary behavior, even within relatively homogeneous study populations.

Sex differences in fat tissue distribution and fat loss across the waist and hip regions were observed in this fully remote study. Consistent with the literature, females had a lower waist-to-hip ratio compared with males (Figure 6A) [36]. Notably, body weight loss correlated more strongly with waist circumference loss in males (r=0.45) than in females (r=0.26, interaction P=.01; Figure 6B), which is also consistent with previous findings [37]. This sexual dimorphism is likely explained by the fact that males store more fat in the waist region compared with females, making the waist region more responsive to weight loss [37]. Together, these findings provide evidence for the accuracy of remotely collected data and demonstrate its potential to study sex differences in responses to nutritional interventions.

Strengths and Limitations

Strengths of the study included its 3-arm design, with a total of 122 participants who were randomized and completed the collection of real-world study data, providing an ecological sample. This design allowed for the evaluation of the effectiveness of the nutritional interventions, albeit in a setting that was less controlled compared with a classical nutritional intervention study. However, there were some limitations associated with this study. First, no comparison was made with nondigital nutritional interventions. Such a comparison would have been valuable, as it could reveal potential differences in the direct effect size of online versus offline personalized nutrition advice. Future studies could include a nondigital control group to address this. Second, the study duration was 6 weeks, reflecting only short-term effects. The long-term impact of the interventions remains unknown, and future research with longer follow-up is needed to assess sustained effects. Third, due to dropouts, the PP-group included a relatively higher proportion of males (Table 1). This likely resulted in a slightly higher baseline body weight in the PP-group compared with the other groups, although the difference was not statistically significant. Randomization and statistical adjustments were applied to mitigate this imbalance, but future studies could aim for larger sample sizes to avoid similar discrepancies. Fourth, although this study demonstrated short-term beneficial effects of personalized nutritional interventions, these results do not guarantee long-term benefits, such as sustained weight loss. Previous studies have shown that maintaining weight loss is challenging, with many individuals regaining weight within 2-5 years [38,39]. In addition, long-term success is influenced not only by individual behavior change but also by broader factors, such as the food environment and supportive public health policies [40], which were beyond the scope of this study. Future research should therefore investigate whether personalized digital interventions can contribute to sustainable behavior change when combined with strategies addressing environmental and policy-level determinants. Fifth, a potential bias in studies relying on self-reported outcomes is participants' tendency to report behaviors that align with perceived study goals, such as weight loss. To minimize this potential bias, we provided standardized instructions for anthropometric measurements, used validated dietary and activity questionnaires, and emphasized the importance of accurate reporting. The randomized design helped distribute any residual bias across study arms. In the food-provision arm, reporting bias may have been slightly higher due to participants' awareness of the intervention; however, randomization and the use of multiple anthropometric measures (body weight, waist, and hip circumference) enhanced the robustness of the results. Cross-validation analyses (Figure 3) and sex-specific effects (Figure 6), consistent with prior literature, further support the reliability of our findings. Sixth, a sample size of 150 at the start of the intervention was considered sufficient to demonstrate the effectiveness of the personalized nutritional interventions. However, due to a relatively high number of dropouts (n=43 due to noncompliance and an additional n=15 due to outlier exclusion), only 122 participants were included in the final data analysis. This reduction in sample size may have limited the statistical power to detect the intended effects. While a dropout rate of 10%-20% was anticipated, the actual dropout due to noncompliance was 23.9% (43/180 participants), with an additional 8.3% (15/180 participants) excluded as outliers. This relatively high attrition rate may be attributed to the remote digital study design without any face-to-face contact and may also reflect a relatively high burden of participation. Data from noncompliant participants were excluded from the analysis, and it is important to acknowledge that the exclusion of participants who did not complete the procedures may have led to an overestimation of feasibility outcomes. Moreover, the high number of outliers could reflect increased misreporting, possibly resulting from unclear instructions. In future studies, real-time plausibility checks could be helpful to flag implausible values and prompt participants to reenter their measurements. Additionally, digital-first studies should account for potentially higher dropout rates and provide clear participant instructions. Despite the attrition, the final sample remained balanced across the 3 study arms, with 39-43 participants per group, allowing for meaningful comparisons of intervention effects. Seventh, because 91 individuals were excluded before participation, the feasibility results reflect only those who entered and completed the study. This selective inclusion may limit generalizability and could lead to an overestimation of feasibility in the broader target population.

Future Directions

While personalized nutrition approaches hold promise in tailoring dietary advice to individual characteristics and improving adherence, it is important to recognize that individual choices are embedded within a broader socioecological context. Dietary behavior is shaped not only by biological and psychological factors but also by the surrounding food

XSL•FO
RenderX

environment, which includes the availability, affordability, accessibility, and cultural norms of food consumption. A growing body of public health literature emphasizes the importance of considering food systems and food environments when designing interventions, as these structural factors often constrain or facilitate the adoption of individualized recommendations [12]. Furthermore, diets are a key driver of environmental change, linking nutrition with planetary health challenges, such as climate change, biodiversity loss, and unsustainable land and water use [41]. Integrating personalized approaches with strategies that improve local food environments and align with sustainable food system goals may therefore offer a more comprehensive and equitable pathway for improving dietary behavior and long-term health outcomes. Moreover, factors such as socioeconomic status, language barriers, and digital literacy can influence both access to and adherence with personalized digital nutrition advice. For example, populations living in socioeconomically deprived neighborhoods often face cumulative disadvantages, including unhealthy food environments, financial constraints, and limited digital access, which can undermine the effectiveness of digital health tools [42]. These considerations suggest that personalized nutrition approaches are likely to be most effective when coupled with efforts to reduce structural barriers and enhance inclusivity, such as tailoring content to different languages, improving accessibility for individuals with lower digital skills, and embedding interventions within supportive community and policy contexts.

Conclusions

We conclude that it is feasible to conduct a fully remote, fully digital nutritional intervention study. Participants were able to independently perform anthropometric measurements at home, reported positive user experiences, and generated self-collected data of sufficient internal validity to detect meaningful changes over time. In addition, the study demonstrated that personalized nutritional advice led to greater weight loss than generic advice, even in the absence of face-to-face contact. The provision of personalized food boxes further facilitated the translation of advice into daily dietary behavior, resulting in the largest reductions in body weight and hip circumference. Taken together, our findings indicate that fully online nutritional intervention studies can be successfully implemented and offer a scalable approach for reaching broader populations while still producing reliable and actionable outcomes.

Acknowledgments

We thank all the participants who participated in this study.

Data Availability

Data are available from the corresponding author (SW) upon reasonable request.

Authors' Contributions

Conceptualization: SW
Data analysis: JCBCdJ
Data curation: MPMC, RD
Data interpretation: FPMH, WJP, LGPP, JCBCdJ, SW
Investigation: CMMB, FPMH, WJP
Methodology: FPMH, SW, WJP
Visualization: JCBCdJ, SW. All authors have read and agreed to the published version of the manuscript.
Writing – original draft: JCBCdJ
Writing– review & editing: all authors

Conflicts of Interest

This study was funded by MixMasters B.V., and Uitgekookt Meal Service provided the personalized food boxes to the participants. Study design, data collection, analyses, interpretation, and manuscript writing were conducted independently by the authors of this paper.

Multimedia Appendix 1
Questionnaires used in this study to (1) screen participants for inclusion and exclusion criteria, (2) collect baseline data, and (3) obtain feedback on the online intervention.
[XLSX File (Microsoft Excel File), 30 KB - jmir_v28i1e73367_app1.xlsx]

Multimedia Appendix 2
Examples of participant feedback, along with the generic and personalized nutritional advice provided.
[DOCX File , 40 KB - jmir_v28i1e73367_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 10789 KB - jmir_v28i1e73367_app3.pdf]

Multimedia Appendix 4
Overview of the calculation of nutritional quality scores.
[DOCX File , 31 KB - jmir_v28i1e73367_app4.docx]

Multimedia Appendix 5
Overall fiber intake during the study.
[DOCX File , 30 KB - jmir_v28i1e73367_app5.docx]

References

1. de Jong AJ, van Rijssel TI, Zuidgeest MGP, van Thiel GJMW, Askin S, Fons-Martínez J, Trials@Home Consortium. Opportunities and challenges for decentralized clinical trials: European regulators' perspective. Clin Pharmacol Ther 2022 Aug 17;112(2):344-352 [FREE Full text] [doi: 10.1002/cpt.2628] [Medline: 35488483]
2. Gul RB, Ali PA. Clinical trials: the challenge of recruitment and retention of participants. Journal of Clinical Nursing 2009 Dec 17;19(1-2):227-233. [doi: 10.1111/j.1365-2702.2009.03041.x]
3. Carlisle B, Kimmelman J, Ramsay T, MacKinnon N. Unsuccessful trial accrual and human subjects protections: An empirical analysis of recently closed trials. Clinical Trials 2014 Dec 04;12(1):77-83. [doi: 10.1177/1740774514558307] [Medline: 25475878]
4. Amstutz A, Schandelmaier S, Frei R, Surina J, Agarwal A, Olu KK, et al. Discontinuation and non-publication of randomised clinical trials supported by the main public funding body in Switzerland: a retrospective cohort study. BMJ Open 2017 Aug 01;7(7):e016216. [doi: 10.1136/bmjopen-2017-016216]
5. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. Contemporary Clinical Trials Communications 2018 Sep;11:156-164. [doi: 10.1016/j.conctc.2018.08.001]
6. Goodson N, Wicks P, Morgan J, Hashem L, Callinan S, Reites J. Opportunities and counterintuitive challenges for decentralized clinical trials to broaden participant inclusion. npj Digit. Med 2022 May 05;5(1):58. [doi: 10.1038/s41746-022-00603-y]
7. Pasman WJ, Hendriks HF, Minekus MM, de Ligt RA, Scholtes-Timmerman MJ, Clabbers ND, et al. Subjective feelings of appetite of wholegrain breakfasts evaluated under controlled, laboratory and 'at home' conditions. Physiology & Behavior 2018 Oct;194:285-291. [doi: 10.1016/j.physbeh.2018.06.024]
8. Pasman WJ, Boessen R, Donner Y, Clabbers N, Boorsma A. Effect of Caffeine on Attention and Alertness Measured in a Home-Setting, Using Web-Based Cognition Tests. JMIR Res Protoc 2017 Sep 07;6(9):e169. [doi: 10.2196/resprot.6727]
9. Dorhout BG, Wezenbeek N, de Groot LCPGM, Grootswagers P. Web-Based Exercise and Nutrition Intervention to Improve Leg Muscle Strength and Physical Functioning in Older Adults: Pre-Post Pilot Study. JMIR Form Res 2025 Jan 30;9:e54392-e54392. [doi: 10.2196/54392]
10. Jinnette R, Narita A, Manning B, McNaughton SA, Mathers JC, Livingstone KM. Does Personalized Nutrition Advice Improve Dietary Intake in Healthy Adults? A Systematic Review of Randomized Controlled Trials. Advances in Nutrition 2021 May;12(3):657-669. [doi: 10.1093/advances/nmaa144]
11. Livingstone KM, Love P, Mathers JC, Kirkpatrick SI, Olstad DL. Cultural adaptations and tailoring of public health nutrition interventions in Indigenous peoples and ethnic minority groups: opportunities for personalised and precision nutrition. Proc. Nutr. Soc 2023 Jun 19;82(4):478-486. [doi: 10.1017/s002966512300304x]
12. Fanzo J, Bellows AL, Spiker ML, Thorne-Lyman AL, Bloem MW. The importance of food systems and the environment for nutrition. The American Journal of Clinical Nutrition 2021 Jan;113(1):7-16. [doi: 10.1093/ajcn/nqaa313]
13. Doets EL, de Hoogh IM, Holthuysen N, Wopereis S, Verain MC, van den Puttelaar J, et al. Beneficial effect of personalized lifestyle advice compared to generic advice on wellbeing among Dutch seniors - An explorative study. Physiol Behav 2019 Oct 15;210:112642 [FREE Full text] [doi: 10.1016/j.physbeh.2019.112642] [Medline: 31394106]
14. Lau Y, Wong SH, Chee DGH, Ng BSP, Ang WW, Han CY, et al. Technology‐delivered personalized nutrition intervention on dietary outcomes among adults with overweight and obesity: A systematic review, meta‐analysis, and meta‐regression. Obesity Reviews 2024 Jan 31;25(5):e13699. [doi: 10.1111/obr.13699]
15. Kan J, Ni J, Xue K, Wang F, Zheng J, Cheng J, et al. Personalized Nutrition Intervention Improves Health Status in Overweight/Obese Chinese Adults: A Randomized Controlled Trial. Front. Nutr 2022 Jun 22;9:919882. [doi: 10.3389/fnut.2022.919882]
16. Al-Awadhi B, Fallaize R, Zenun Franco R, Hwang F, Lovegrove JA. Insights Into the Delivery of Personalized Nutrition: Evidence From Face-To-Face and Web-Based Dietary Interventions. Front. Nutr 2021 Jan 27;7:570531. [doi: 10.3389/fnut.2020.570531]

XSL•FO
RenderX

17. Griffiths F, Lindenmeyer A, Powell J, Lowe P, Thorogood M. Why Are Health Care Interventions Delivered Over the Internet? A Systematic Review of the Published Literature. J Med Internet Res 2006 Jun 23;8(2):e498. [doi: 10.2196/jmir.8.2.e10]
18. de Hoogh IM, Winters BL, Nieman KM, Bijlsma S, Krone T, van den Broek TJ, et al. A Novel Personalized Systems Nutrition Program Improves Dietary Patterns, Lifestyle Behaviors and Health-Related Outcomes: Results from the Habit Study. Nutrients 2021 May 22;13(6):1763. [doi: 10.3390/nu13061763]
19. van der Haar S, Hoevenaars FPM, van den Brink WJ, van den Broek T, Timmer M, Boorsma A, et al. Exploring the Potential of Personalized Dietary Advice for Health Improvement in Motivated Individuals With Premetabolic Syndrome: Pretest-Posttest Study. JMIR Form Res 2021 Jun 24;5(6):e25043. [doi: 10.2196/25043]
20. Norstat. URL: https://www.norstat.co [accessed 2025-12-22]
21. ten Haaf T, Weijs PJM. Resting Energy Expenditure Prediction in Recreational Athletes of 18–35 Years: Confirmation of Cunningham Equation and an Improved Weight-Based Alternative. PLoS ONE 2014 Oct 2;9(10):e108460. [doi: 10.1371/journal.pone.0108460]
22. Mijn Eetmeter. URL: https://mijn.voedingscentrum.nl/nl/eetmeter/ [accessed 2025-12-22]
23. van den Brink WJ, van den Broek TJ, Palmisano S, Wopereis S, de Hoogh IM. Digital Biomarkers for Personalized Nutrition: Predicting Meal Moments and Interstitial Glucose with Non-Invasive, Wearable Technologies. Nutrients 2022 Oct 24;14(21):4465. [doi: 10.3390/nu14214465]
24. Healey G, Brough L, Murphy R, Hedderley D, Butts C, Coad J. Validity and Reproducibility of a Habitual Dietary Fibre Intake Short Food Frequency Questionnaire. Nutrients 2016 Sep 10;8(9):558. [doi: 10.3390/nu8090558]
25. FAO/WHO/UNU Expert Consultation. Human energy requirements: report of a joint FAO/WHO/UNU Expert Consultation. In: FAO Food and Nutrition Technical Report Series. Rome: Food and Agriculture Organization of the United Nations; 2004:1-96.
26. World Health Organization. Waist circumference and waist-hip ratio: report of a WHO expert consultation. In: World Health Organization meeting report. Geneva: World Health Organization; 2011:1-39.
27. Ryan K, Dockray S, Linehan C. A systematic review of tailored eHealth interventions for weight loss. DIGITAL HEALTH 2019 Feb 05;5:e1. [doi: 10.1177/2055207619826685]
28. Cheng J, Costacou T, Sereika SM, Conroy MB, Parmanto B, Rockette-Wagner B, et al. Effect of an mHealth weight loss intervention on Healthy Eating Index diet quality: the SMARTER randomised controlled trial. Br J Nutr 2023 Jun 07;130(11):2013-2021. [doi: 10.1017/s0007114523001137]
29. Ambeba E, Ye L, Sereika S, Styn M, Acharya S, Sevick M, et al. The use of mHealth to deliver tailored messages reduces reported energy and fat intake. Journal of Cardiovascular Nursing 2015;30:35-43. [doi: 10.1097/jcn.0000000000000120]
30. Gonzalez-Ramirez M, Sanchez-Carrera R, Cejudo-Lopez A, Lozano-Navarrete M, Salamero Sánchez-Gabriel E, Torres-Bengoa MA, et al. Short-Term Pilot Study to Evaluate the Impact of Salbi Educa Nutrition App in Macronutrients Intake and Adherence to the Mediterranean Diet: Randomized Controlled Trial. Nutrients 2022 May 14;14(10):2061. [doi: 10.3390/nu14102061]
31. Kohl J, Brame J, Centner C, Wurst R, Fuchs R, Sehlbrede M, et al. Effects of a Web-Based Lifestyle Intervention on Weight Loss and Cardiometabolic Risk Factors in Adults With Overweight and Obesity: Randomized Controlled Clinical Trial. Journal of medical Internet research 2023;25:e43426. [doi: 10.2196/43426]
32. Beleigoli A, Andrade AQ, Diniz MDF, Ribeiro AL. Personalized Web-Based Weight Loss Behavior Change Program With and Without Dietitian Online Coaching for Adults With Overweight and Obesity: Randomized Controlled Trial. J Med Internet Res 2020 Nov 5;22(11):e17494. [doi: 10.2196/17494]
33. Junaidi, Apriyanto T, Laily I, Rizki P. The comparison of offline and online nutrition education on body mass index in rugby athletes during the Covid-19 Pandemic (The Body Mass Index profile of Jakarta athletes during Covid-19 Pandemic). Journal of Physical Education and Sport 2021 Aug 15:2295-2301.
34. Lau Y, Chee DGH, Chow XP, Cheng LJ, Wong SN. Personalised eHealth interventions in adults with overweight and obesity: A systematic review and meta-analysis of randomised controlled trials. Preventive Medicine 2020 Mar;132:106001. [doi: 10.1016/j.ypmed.2020.106001]
35. Eldridge AL, Piernas C, Illner A, Gibney MJ, Gurinović MA, De Vries JH, et al. Evaluation of New Technology-Based Tools for Dietary Intake Assessment—An ILSI Europe Dietary Intake and Exposure Task Force Evaluation. Nutrients 2018 Dec 28;11(1):55. [doi: 10.3390/nu11010055]
36. Karastergiou K, Smith SR, Greenberg AS, Fried SK. Sex differences in human adipose tissues – the biology of pear shape. Biol Sex Differ 2012 May 31;3(1):1-12. [doi: 10.1186/2042-6410-3-13]
37. Kuk JL, Ross R. Influence of sex on total and regional fat loss in overweight and obese men and women. Int J Obes 2009 Mar 10;33(6):629-634. [doi: 10.1038/ijo.2009.48]
38. Weiss EC, Galuska DA, Kettel Khan L, Gillespie C, Serdula MK. Weight Regain in U.S. Adults Who Experienced Substantial Weight Loss, 1999–2002. American Journal of Preventive Medicine 2007 Jul;33(1):34-40. [doi: 10.1016/j.amepre.2007.02.040]
39. Lowe MR, Kral TVE, Miller-Kovach K. Weight-loss maintenance 1, 2 and 5 years after successful completion of a weight-loss programme. Br J Nutr 2007 Nov 28;99(4):925-930. [doi: 10.1017/s0007114507862416]

XSL•FO
RenderX

40. Tewahade S, Berrigan D, Slotman B, Stinchcomb DG, Sayer RD, Catenacci VA, et al. Impact of the built, social, and food environment on long‐term weight loss within a behavioral weight loss intervention. Obesity Science & Practice 2022 Nov 03;9(3):261-273. [doi: 10.1002/osp4.645]
41. Willett W, Rockström J, Loken B, Springmann M, Lang T, Vermeulen S, et al. Food in the Anthropocene: the EAT–Lancet Commission on healthy diets from sustainable food systems. The Lancet 2019 Feb;393(10170):447-492. [doi: 10.1016/s0140-6736(18)31788-4]
42. Ter Ellen F, Oude Groeniger J, Stronks K, Hagenaars L, Kamphuis C, Mackenbach J, et al. Understanding the dynamics driving obesity in socioeconomically deprived urban neighbourhoods: an expert-based systems map. BMC Med 2025 Jan 07;23(1):2 [FREE Full text] [doi: 10.1186/s12916-024-03798-x] [Medline: 39762884]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
FAO: Food and Agriculture Organization
PAL: physical activity level
P-group: personalized intervention group
PP-group: personalized plus intervention group
TNO: Netherlands Organization for Applied Scientific Research
UNU: United Nations University
WHO: World Health Organization

Edited by N Cahill; submitted 04.Mar.2025; peer-reviewed by N Holliday, J Nielsen; comments to author 11.Sep.2025; revised version received 26.Nov.2025; accepted 26.Nov.2025; published 05.Jan.2026.

Please cite as:
de Jong JCBC, Hoevenaars FPM, Peters LGP, Berendsen CMM, Pasman WJ, Caspers MPM, Dulos R, Wopereis S
A Real-Life Digital Intervention for Personalized Nutrition in Adults With Overweight or Obesity: Remote Randomized Controlled Trial
J Med Internet Res 2026;28:e73367
URL: https://www.jmir.org/2026/1/e73367
doi:10.2196/73367
PMID:

©Jelle CBC de Jong, Femke PM Hoevenaars, Lotte GP Peters, Charlotte MM Berendsen, Wilrike J Pasman, Martien PM Caspers, Remon Dulos, Suzan Wopereis. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Web-Based Cancer Prevention Intervention for Rural Emerging Adults: Mixed Methods Development and Pilot-Testing Study

Echo L Warner[1], MPH, PhD; Alishia Kinsey[2], BS; Barbara J Walkosz[2], PhD; Julia Berteletti[2], MSW; Kayla Nuss[2], PhD; Annelise Small[2], BA; W Gill Woodall[2], PhD; Deanna Kepka[3], MPH, PhD; Douglas Taren[4], PhD; Meghan B Skiba[5], PhD; Dolores D Guest[6], PhD; Cindy K Blair[6], PhD; Judith S Gordon[5], PhD; David W Wetter[7], PhD; Evelinn A Borrayo[8], PhD; Kimberly L Henry[9], PhD; Andrew L Sussman[6*], PhD; David B Buller[2*], PhD

[1]College of Nursing, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, United States
[2]Klein Buendel (United States), Golden, United States
[3]Circle of Hope, College of Nursing; Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, United States
[4]Department of Pediatrics Nutrition Center, School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, United States
[5]College of Nursing, University of Arizona, Tucson, AZ, United States
[6]Comprehensive Cancer Care Center, University of New Mexico, Albuquerque, United States
[7]Department of Population Health Sciences, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, United States
[8]University of Colorado Cancer Center, Aurora, United States
[9]Department of Psychology, Colorado State University, Fort Collins, CO, United States
[*]these authors contributed equally

Corresponding Author:
Echo L Warner, MPH, PhD
College of Nursing, Huntsman Cancer Institute
University of Utah
10 S 2000 E
Salt Lake City, UT, 84112
United States
Phone: 1 8012447040
Email: echo.warner@nurs.utah.edu

Abstract

Background: The rapid growth of user-generated web-based health information increases the complexity of cancer information seeking. One promising strategy for promoting high-quality cancer information consumption is through targeted interventions that are intentionally designed to reach individuals in the web-based spaces they occupy. However, there is a paucity of evidence-based information on the best strategies for designing and implementing web-based health behavior change interventions to improve individuals' cancer-related knowledge and prevent cancer.

Objective: This study aimed to develop and pilot test a theory-based intervention via the web to reduce 6 cancer risk factors among rural emerging adults (EAs) through community-engaged research.

Methods: This mixed methods evaluation describes the development of a web-based cancer prevention intervention aimed at rural EAs aged 18-26 years in the United States and delivered in Facebook private groups. The intervention was guided by behavior change theory and cocreated with EA and Stakeholder Organization Advisory Boards to ensure relevance, accessibility, and appropriateness. We report on 3 formative surveys, a pilot intervention, protocol development, and the community-engaged process for intervention development. Descriptive statistics were applied to the surveys and pilot intervention baseline results to produce means and SDs using R.

Results: We developed posts (n=400) for a Facebook feed aimed at reducing 6 cancer risk behaviors (unhealthy diet, lack of physical activity, tobacco use, alcohol use, sun exposure, and human papillomavirus infection) with iterative input from the EA and stakeholder advisory boards. Formative surveys with rural EAs (n=297) and a pilot study of the intervention with this population (n=26) were conducted. In the pilot study, the intervention reached participants across rural counties, with sustained engagement (post views=1060, reactions=346, comments=72) over a one-month period. Key modifications to the intervention content and design emerged from both advisory boards, the formative surveys, and the pilot intervention, focusing on using perceived reliable sources and direct links to source material.

XSL•FO
RenderX

Conclusions: This web-based cancer prevention intervention is scalable and delivers engaging, evidence-informed health information to rural EAs. We offer key insights into the design and implementation of web-based cancer prevention interventions for EAs by describing the resources, timelines, and expertise needed to design and implement the intervention. Considerations for fully engaging EA and community stakeholder partners are presented, and we discuss how their involvement resulted in modifications that strengthened the intervention. Finally, we highlight the importance of theory-based health-behavior messaging, digital messaging skillsets, and platform-tailored dissemination strategies for maximizing web-based intervention acceptability.

Trial Registration: ClinicalTrials.gov NCT05618158; https://classic.clinicaltrials.gov/ct2/show/NCT05618158

International Registered Report Identifier (IRRID): RR2-10.2196/50392

(J Med Internet Res 2026;28:e80803) doi:10.2196/80803

KEYWORDS

emerging adult; web-based; intervention development; cancer; prevention; rural; community engagement

Introduction

Emerging adults (EAs) in rural and remote areas of the United States face unique challenges that increase their vulnerability to cancer. Emerging adulthood, a critical life stage from ages 18 to 26 years, is marked by increased autonomy and transitions in financial, residential, and employment responsibilities [1]. However, many EAs establish unhealthy lifestyle patterns during this period, including reduced physical activity, poor dietary habits, nicotine and tobacco use, binge drinking, sporadic sun safety practices, and skipping human papillomavirus (HPV) vaccination [1,2]. Living in a rural setting can heighten feelings of isolation and limit access to resources that support healthy behaviors [3]. These modifiable risk factors contribute to premature cancer morbidity and mortality, making rural EAs a priority population for cancer prevention efforts.

Given its popularity among EAs [4], using a social web-based intervention offers a promising avenue for addressing cancer prevention among EAs in rural areas. Over 80% of individuals in this age group access the internet several times per day [4]. Despite known flaws in information quality, web-based interventions that are strategically and theoretically designed are promising strategies for providing high-quality health information from trusted voices [5]. Web-based educational interventions can also disseminate timely and relevant public health messages, leverage user-generated content to personalize information, engage audiences in 2-way communication, and be used to detect and respond to emerging trends [6,7]. The success of web-based interventions hinges on their ability to address the broader social determinants of health, community and cultural perceptions, and built environments in which individuals live. However, web-based media can also perpetuate harmful and misleading health information [8-11], and interventions that relate accurate and truthful cancer prevention strategies are needed [12-16]. Studies of web-based interventions have often lacked theoretical grounding, rigorous design and evaluation procedures, and guidelines on the development of content, limiting the rigor and reproducibility of this intervention approach.

Rural communities often face unique barriers to engaging in healthy lifestyles, such as limited health care access, socioeconomic challenges, and geographic isolation [17-19]. EAs have limited interaction with traditional community channels like schools, workplaces, and health care settings, so web-based interventions may provide unique opportunities for tailored, real-time engagement with EAs. Use of community engagement strategies in the development of both the content and structure of social media interventions can increase relevance to the specific needs of rural EAs [20,21]. A community-engaged approach to web-based intervention design should increase the relevance of interventions to the specific needs of rural EAs.

In this context, we developed a theory-based, web intervention to reduce 6 cancer risk factors common among rural EAs, including physical inactivity, unhealthy diets, nicotine/tobacco use, binge drinking, unprotected ultraviolet exposure, and preventing HPV infection. It was designed and pilot tested using community-engaged methods. Herein, we describe the development, pilot-testing, and refinement of our intervention prior to its launch in a randomized quasi-experimental trial. This developmental research project involved community-engaged research methods with a dynamic group of EAs and community stakeholders, a theory-informed process for creating content tailored to rural EAs, and formative research to refine and pilot test the intervention.

Methods

Setting and Study Design

The overall aim of the study was to create a web-based intervention, named PEAK Wellness Chat, and evaluate its effectiveness in a sample of EAs in a stepped wedge randomized quasi-experimental trial design (NCT05618158) [22]. The private-group function in Facebook was the platform for intervention delivery. Facebook is used by a large number of American adults, regardless of race/ethnicity, including EAs (67% of adults use it and 70% of rural adults) [23]. The private-group function cultivates the privacy of group members, which is essential to control experimental exposure to the intervention and avoid experimental contamination. Facebook also allows a variety of content delivery by length and type (eg, images and videos as well as links to other sites and sharing of Facebook posts from other organizations) and allows posts to remain in the group in perpetuity. Facebook also records engagement of each participant with posts (ie, reactions and comments) and tracks retention via group membership, to enable testing of intervention dose effects. Finally, the Facebook

XSL•FO
RenderX

algorithm promotes private-group posts in participants' feeds in terms of frequency and prominence when they engage more with group posts.

Ethical Considerations

This study was approved by the WCG Institutional Review Board (IRB20223673 and IRB20223673). Informed consent was obtained and documented electronically prior to enrollment in the formative surveys and pilot study described below. Privacy and confidentiality were maintained by storing all research subjects' data in password-protected and encrypted drives. No identifying information is presented herein. Participants were provided with a modest incentive for completing the formative survey (US $40) or pilot study (US $50).

Community-Engaged Strategy for Developing PEAK Wellness Chat

A library of Facebook posts for the PEAK Wellness Chat intervention was created by a core team of investigators and staff with expertise in cancer prevention, public health promotion, health communication, and emerging adulthood. The library was iteratively developed with structured community input from an Emerging Adult Advisory Board (EAAB) and Stakeholder Organization Advisory Board (SOAB), responses by EAs to formative surveys, and guidance by experts in the 6 cancer risk factors. Intervention management by the study's administrative team and engagement by EAs were assessed in a pilot study (Figure 1). The administrative team supported the overall study procedures, meeting facilitation, coordination and deployment of surveys, recruitment efforts, and software acquisition for post illustrations. Additionally, the administrative team produced and maintained a cancer information audit and a resource website. The cancer information audit was updated weekly with emergent cancer information and news about health and wellness, alongside any emerging internet trends and current events that should be incorporated into posts to promote relevance and relatability of post content. The resource website contained additional information about the cancer risk factors, community and national resources (eg, quit smoking hotline and food banks), and links to health education resources.

Figure 1. Community-engaged strategy for designing a Facebook post library. EA: emerging adult.

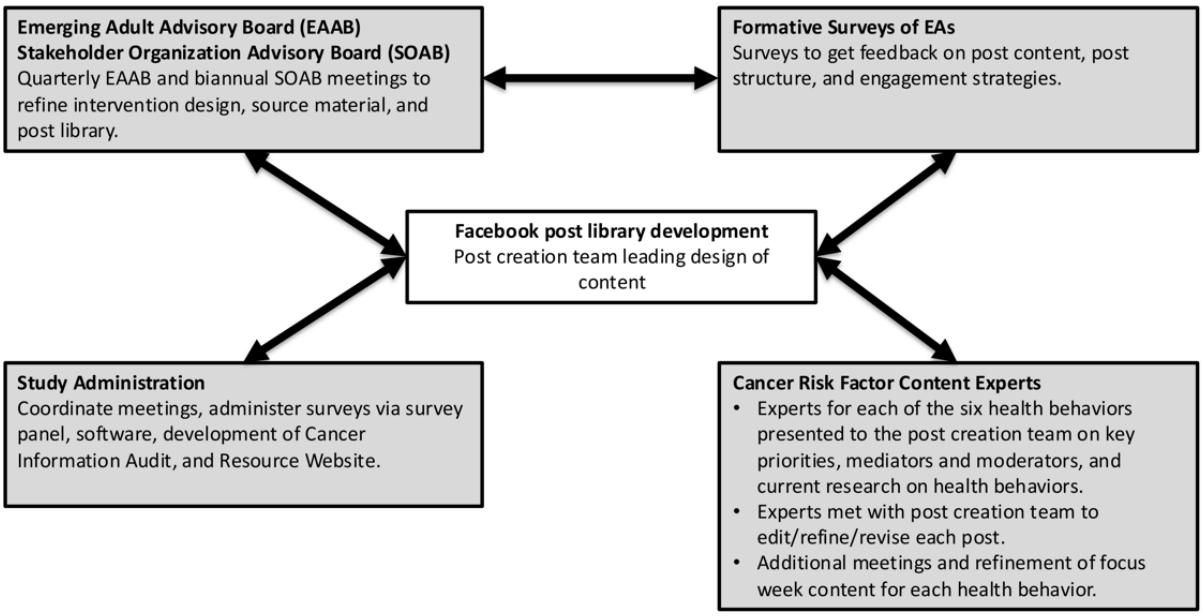

EAABs and SOABs

We used partnership processes aligned with the community-based participatory research model of Wallerstein et al [20] as a framework for the participatory approach. Through regular advisory board meetings, we addressed: (1) knowledge of rural "Contexts" that inform catchment area needs; (2) culturally informed "Partnership Processes" that facilitated a regular and iterative feedback cycle between the advisory boards, content experts, and post creation committee to refine the design features, language, and content of the Facebook posts; (3) development of responsive "Intervention & Research" protocols for rural conditions; and (4) participatory "Outcomes" that were iteratively disseminated to community partners through the intervention development and advisory board meetings (and are continuing during the intervention implementation in the trial).

EAAB and SOAB Composition

A total of 15 EAs, representing the composition of participating rural communities, agreed to serve on the EAAB. EAAB members were recruited through word-of-mouth referrals from the research teams' community networks. The EAAB held quarterly meetings to give input on the lived experiences of EAs, review the web-based intervention plans and protocols (eg, engagement strategies in posts), consider appropriateness, relevance, and engagement of proposed posts, and advise on recruiting EAs for the trial. Project investigators and staff at each participating Cancer Center also led direct engagement efforts with stakeholder organizations that serve rural EAs to convene members of the SOAB, which had 14 members. The

XSL•FO
RenderX

SOAB complemented the EAAB by identifying local resources for cancer risk reduction, reviewing intervention plans, protocols, and posts, and supporting recruitment and dissemination efforts in biannual meetings. Topics for discussion in the EAAB and SOAB meetings were jointly decided based on project needs, PEAK Wellness Chat content, and priorities of the board members.

Creation of the PEAK Wellness Post Library

The library of Facebook posts contained messages educating EAs about strategies to improve the 6 cancer risk behaviors. Messages were also created to increase EAs' media education about cancer and cancer prevention. Finally, posts were produced to enhance EAs' skills for communicating with family and friends because they reported: (1) living in a variety of arrangements, (2) relying on family and friends for acquiring and preparing food and paying for health care, and (3) living with others who engaged in the cancer risk behaviors. A focus week of up to 8 posts for each cancer risk factor was designed to emphasize messages on one important principle for improving the cancer risk factor. A focus week of posts on each of the 6 risk factors is being delivered during each 3-month period during the intervention.

Post Development Process

Each post was designed to address key features of health behavior change theories (self-determination theory, social cognitive theory, and diffusion of innovation theory [24-26]) to ensure the posts are conceptually driven to impact EAs' health behaviors. Specific principles incorporated into the posts included: intrinsic motivation, extrinsic motivation, social support, relatedness, personal goals, ability to control, self-efficacy, response efficacy, injunctive norms, descriptive norms, cancer risk perceptions, response cost, compatibility with values, observable benefits, autonomy, and simplicity. Table S1 in Multimedia Appendix 1 provides examples of posts incorporating each principle. Principles from self-determination theory, social cognitive theory, and diffusion of innovation were operationalized using messaging design and engagement strategies to ensure messages were not only informative but also encouraged motivation and engagement. Message design strategies included polls, narratives, calls for sharing, testimonials from EAs, source credibility, behavioral skills, and referrals to community resources. Engagement techniques included content from near peers, invitations to comment, cultural barriers and facilitators, images of ethnically diverse EAs, stories, videos or other visuals, question and answer formatting, and behavioral change techniques (ie, self-improvement and freedom to act), and high-interest, useful, and current event content. The framework also included message strategies to promote engagement (eg, visuals, current event content, cultural barriers/facilitators, polls, requests for comments, story requests, knowledge test, picture request). These features will be tracked and evaluated to ensure a variety of engagement strategies and best practices for health behavior change are integrated throughout the campaign. The post library was created and maintained in Microsoft Excel, where each post was coded for key theoretical concepts and engagement features (Table 1).

For this phase of intervention development, the EAAB and SOAB provided timely feedback on the content and design of posts, novel engagement strategies (ie, videos, use of polls), trustworthy and credible source material, and incorporation of local resources where possible. Along with the post library, the authors developed protocols for disseminating the intervention, moderating the Facebook group, and verifying the legitimacy of a participant's Facebook account. These included procedures for scheduling posts, responding to questions, and auditing the cancer prevention communication environment to identify emergent topics that should be included in the intervention to make messages relatable, relevant, and engaging. For example, the protocols included specific guidance for the moderator. Per the protocol, the moderator reacted (ie, like, love, and care) to every participant's comment left on posts. If a participant's comment resonated with the moderator, they would respond with their own comment to foster an authentic sense of community. Moderators will also respond to any direct Facebook messages participants send regarding the study (ie, questions about certain topics the participant does not want to share publicly). The protocol also includes boundaries for the moderator to model feasible and replicable interactions with the groups (eg, not to like comments that are misinformation).

XSL•FO
RenderX

Table 1. PEAK wellness post library.

Post features	Definition	Example
Post #	Indicates the order of the post and record number	1, 2, and 3
Date and time posted	Calendar date, time posted	26 February, 2025 9:00 AM
Day of week	Differentiate weekdays from weekends	Wednesday and Saturday
Topic	Identifies one of the 6 cancer prevention risk factors, media literacy, or family communication	Sun safety, HPV[a] prevention strategies, healthy diet
Message	Main content shared in the post	Knowing how to correctly use sunscreen to protect your skin is important. Here are some general guidelines.
Link	Source content link to be shared in the comments with viewers	https://www.cdc.gov/tobacco/campaign/tips/quit-smoking/index.html
Post visual suggestions	Indicates the type of visual content that should be included in the post	Team created/external content creator, TikTok video, YouTube video, infographic, or photo
Facebook post link	Link to the post on Facebook for tracking and review purposes	URL
Gender presentation	Indicate the visual gender presentation of people in the posts	Woman, man, N/A[b] if none
Character features	Indicate the body type, presenting race/ethnicity, and conventionally attractive features	Athletic build, Asian, conventionally attractive
Colors used in graphics	2-3 main colors	Red, blue, green
Engagement post	Indicator of whether or not the post is designed to promote engagement	Poll, “Tell us in the comments,” share a story, knowledge test, picture request
Primary/secondary outcome addressed	Health behavior addressed with the post, also includes key topics addressed in the post	Cancer risk, autonomy, cooking skills, increased physical activity, preventing drunk driving, vaping knowledge, winter sun protection
Theoretic mediators	Key behavior mediators designed to promote behavior and health belief change	Intrinsic motivation, extrinsic motivation, social support, relatedness, personal goals, ability to control, self-efficacy, response efficacy, injunctive norms, descriptive norms, cancer risk perceptions, response cost, compatibility with values, observable benefits, simplicity
Message design features	Key design features based on health communication practices	Moderator instructions, referral to community resources, testimonials from EAs[c]
Engagement techniques	Strategies used in the post to promote engagement	Visuals, current event content, cultural barriers/facilitators, poll, “Tell us in the comments,” share a story, knowledge test, picture request

[a]HPV: human papillomavirus.

[b]N/A: not applicable.

[c]EA: emerging adult.

Formative Surveys With Rural EAs

Overview

During the development of the post library, we administered 3 web-based surveys with rural EAs to evaluate initial reactions to the posts (Table 2). The surveys also collected information on EAs lived experiences, use of Facebook and other web-based channels, and perceived credibility of potential information sources used in the posts.

XSL•FO
RenderX

Table 2. PEAK Wellness Chat intervention development surveys.

Survey name	Purpose	Sample size (N)	Fielded dates
Formative survey 1	Health topics of interest, popular influencers or content creators, and credibility of health information sources.	100	February 2024
Formative survey 2	Discussing health behaviors with others, cancer risk behaviors by household, Facebook group name, and dietary experiences	100	July 2024
Formative survey 3	EAs'[a] engagement in physically active jobs and recreational activity, access to and perceptions of healthy and affordable food	97	January 2025
Pilot pretest survey	Baseline test of health behaviors, mediators, other covariates, and demographics	26	March 2024
Pilot posttest survey	Baseline test of health behaviors, mediators, other covariates, and demographics	23	April 2024

[a]EA: emerging adult.

Participants

Eligible individuals were ages 18-26 years, living in counties designated as Rural Urban Continuum Codes (RUCC) 4-9 [27] in Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming, and Rural Urban Commuting Area Codes 4-10, which classifies rurality by census tract [28] in Arizona where RUCC codes excluded several key rural communities. To facilitate adequate recruitment for the full trial, we expanded our recruitment to states beyond the original 4 corners states, and this modification was reflected in the formative survey participants. They also needed to be able to access a web-based survey and to use Facebook once or more times per week. Participants were identified through a survey panel company, Dynata [29]. The overall survey response rate was 302/1080 (28.0%), and the overall survey completion rate was 297/302 (98.3%).

Measures and Analysis

Posts were purposefully selected for inclusion in the surveys, focusing on issues that arose during post development and discussions with the EAAB, SOAB, and content experts. For example, the credibility of sources cited in the posts was a concern, so we selected posts with local and national source material (eg, Centers for Disease Control, newspapers, and local health departments). Respondents were asked to (1) rate each post for appropriateness, relevance, believability, amount of text, and trustworthiness on 5-point Likert scales; (2) indicate their potential engagement with the post through reading, scrolling past, reacting to, commenting on, and clicking on links (no, yes maybe, yes definitely); and (3) suggest ways for improving the post (open ended).

Surveys also contained questions on EAs lived experiences that informed postdevelopment. In the first survey (February 2024), participants were asked about topics of interest, popular influencers/content creators, and the credibility of information sources. In the second survey (July 2024), they were reported on discussing health behaviors with parents, siblings, partners, and friends, cancer risk behaviors by household members (use of nicotine products, consumption of alcohol, and intentional suntanning), preferences for the Facebook group name, and dietary experiences including sources of food, dependence on others for food purchasing, and sources of free food. In the third survey (January 2025), questions inquired about EAs' engagement in physically active jobs and recreational physical activity outside work, access to healthy and affordable food, and perceptions about healthy eating. Formative surveys are available upon reasonable request from the authors.

Analysis of EA Survey Responses

Summary statistics of participants' sociodemographics, lived experiences, and post feedback were calculated for each survey, using RStudio (Posit Software, PBC), 2024. Two open-ended questions ("What would you comment on this post?" and "How would you improve this post?") were coded to identify emergent themes. All responses were categorized into an individual theme. From the Comment question, 134 responses were analyzed, and 53 responses were excluded due to being out of context, incomplete statements, or illegible responses. The 81 responses coded resulted in the following themes: message-specific reactions and affirmative comments. From the Improvement question, 506 responses were analyzed, and 73 responses were excluded due to being out of context, incomplete statements, or illegible responses. The remaining 433 responses were coded by 2 coders in the following themes: suggestions on image, dislike or wrong audience, more information requested, reinforce source/suggestions on source, suggestion on content, too much information, unclear message, nothing to change, unsure.

Survey and open-ended feedback were iteratively reviewed by the full study team to improve the content and format of the posts in the intervention feed and ensure that posts were responsive, timely, and engaging for rural EAs. Summary reports of formative survey results were discussed in EAAB and SOAB meetings to review thematic findings and explore contextual and social considerations.

PEAK Wellness Chat Pilot Study

Overview

A 4-week pilot test of the intervention was conducted with a sample of rural EAs to refine procedures for the full trial pertaining to recruitment, baseline and posttest surveys, retention, and intervention protocols. Data were also obtained to confirm that posts were engaging for EAs. The pretest survey is available as a Multimedia Appendix 2.

XSL•FO
RenderX

Participants

Participants were recruited by Verasight [30], a research services company that recruits participants for studies through a web-based panel and advertising. Eligible participants for the pilot study were ages 18-26 years, living in counties designated as RUCC 4-9 in Arizona, Colorado, New Mexico, and Utah, and able to access the web-based survey. Pregnant participants were excluded because they may depart from their normal dietary and activity levels and consumption of alcohol and nicotine due to the pregnancy. Participants were screened for having an existing Facebook account for at least 1 year with weekly activity, with 2 EAs deemed ineligible. In addition, 3 EAs were excluded because they did not friend the Group Moderator and could not be joined to the Facebook private group for the pilot feed.

Pilot Study Procedures

Participants were enrolled in a single group, pretest-posttest design. Initially, they completed a baseline survey in REDCap (Research Electronic Data Capture; Vanderbilt University). The intervention was presented in a Facebook private group, with 2 posts made per day. During a 7-day period in the month, 8 posts in a focus week on physical activity were posted. This focus week strategy was designed to provide an in-depth intervention on improving self- and response-efficacy, increasing perceived risk associated with not being physically active [31-34], and linking cancer prevention to personal physical activity goals. During the 4-week campaign, 62 posts were made in the private group feed across domains related to physical activity (n=13), diet (n=7), sun safety (n=7), tobacco cessation (n=6), alcohol reduction (n=7), HPV prevention (n=7), media education (n=3), and family communication (n=2). Media literacy and family communication content were more general and not specifically related to the 6 health behaviors. At the end of the 4-week intervention period, participants were invited to complete a posttest survey in REDCap.

Measures and Analysis

The pretest and posttest surveys included questions about the 6 cancer risk factors: physical activity (Global Physical Activity Questionnaire [35]), diet (dietary screener; meal behaviors, food insecurity) [36,37], alcohol intake (Alcohol Use Disorders Identification Test–Consumption [38]), nicotine product use (30-day and 7-day smoking or vaping, quit ladder) [39,40], HPV prevention (initiation and completion of multi-shot vaccine series) [41], and ultraviolet protection (use of personal sun protection practices and sunburn) [42-45]. In addition, questions assessed basic needs, self-efficacy for cancer risk-reduction behaviors, cancer information overload, digital media use, health insurance coverage, last routine check-up, personal and family cancer history, and demographics.

Behavioral and experiential engagement with posts in the Facebook private group feed was measured. Behavioral engagement was assessed in two ways: (1) project staff kept a log of posts published to the Facebook private group that tracked date and time posted, topic, and post visual (graphic, gif, image, link, poll, or video); and (2) project staff extracted Facebook metadata reported in the private group platform, which included obtaining the total number of views, likes (eg, like and sad), and comments (comments from participants, comments from moderator) [46]. Facebook's reporting function permitted likes and comments but not views to be associated with specific users. The posttest survey contained questions assessing participants' experiential engagement with the intervention (ie, frequency of reading posts and sharing of content with others). Given the very small sample, the analysis involved descriptive statistics. All analyses were conducted in RStudio (version 4.4.2; Posit Software, PBC). The STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist is provided in Multimedia Appendix 3.

Results

Facebook Post Library

We produced a library of 400 Facebook posts on reducing the 6 cancer risk behaviors, media education, and family communication, using a real-time agile process to ensure posts engaged EAs. The development process followed an iterative feedback loop for each new post created. Content experts identified the key topics and information to address in the posts and essential mediators of behavior change for the cancer risk behaviors. The content from these presentations was adapted by a group of investigators and project staff into posts following the theoretical framework shown in Table 1. Content experts reviewed and approved the posts for inclusion in the final PEAK Wellness library, suggesting revisions that were incorporated to finalize the posts. Selected posts were pretested by having them reviewed by both the EAAB and SOAB, and included in the formative surveys for feedback as needed before finalization. Figure 2 illustrates the evolution of one post (about binge drinking) that was created using a photo visual and a local news source.

XSL•FO
RenderX

Figure 2. Evolution of a post through an iterative community-engaged development process. EA: emerging adult; EAAB: Emerging Adult Advisory Board; SOAB: Stakeholder Organization Advisory Board.

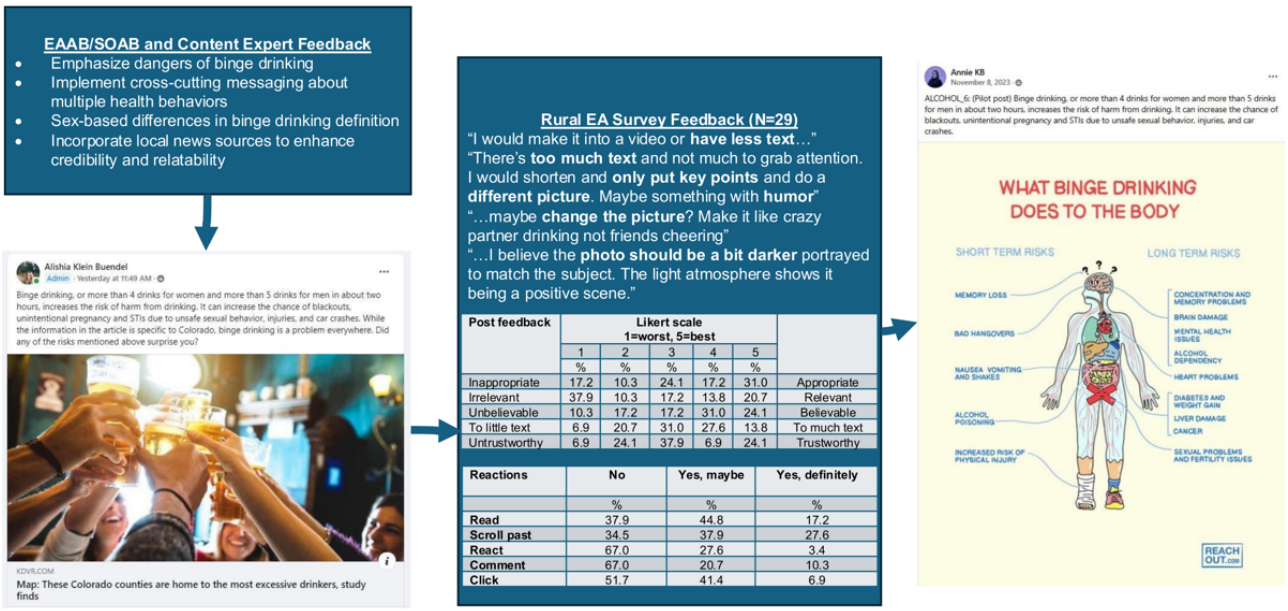

Formative Rural EA Surveys

Formative surveys with rural EAs provided insights about developing, revising, and finalizing posts, health behavior topics of importance, and contextual information to include to improve the posts. The 3 formative surveys with EAs obtained responses from 99, 98, and 100 EAs, for a total sample of 297 rural EA respondents. As shown in Table 3, the average age of respondents was 22.5 (SD 2.49) years. The sample of participants was primarily identified as 72.4% (n=215) female, 24.2% (n=72) Hispanic, 70.7% (n=210) White, 54.5% (n=54) high school educated or less, 43.2% (n=82) unemployed, 46.2% (n=90) living with a spouse/partner, and 34.3% (n=99) living with other relatives besides parents. The majority, 63.6% (n=63), had health insurance, and many EAs involved others in their health decisions, primarily parents (56/99, 55.6%) and a spouse/partner (40/99, 40.4%). Respondents resided in 8 states, with subsamples ranging from Nevada (26/297, 8.8%) to Idaho (61/297, 20.5%). Across the 3 formative surveys, respondents provided feedback about 61 posts (survey 1=21 posts, survey 2=20 posts, survey 3=20 posts).

Content analysis of the 2 open-ended questions asking about how participants would engage with the posts and improve the posts resulted in 2 themes for post comments and 9 themes for improving posts (Table 4). Content analysis was applied by one research team member, who met with 2 other researchers to discuss definitions, interpretations, and the application of codes. Discrepancies were resolved through consensus, after which coding was finalized. In terms of comments, themes focused on respondents' initial reaction to the post content and affirmations or approvals of the post content, creator, or message. Most posts received overwhelmingly positive feedback, but there were suggestions for improvements on visual appeal and less text across many posts. Due to this feedback, the post campaign library was revised to incorporate more video and photo content and eliminate text-only posts. Source material was a primary area of feedback that also resulted in changes to the campaign library, to provide greater transparency in primary source material and emphasize trusted organizations (local newspapers had lower levels of trust, while national organizations were seen as more trustworthy). Another unique theme related to the relevancy of the content, with sizable numbers of rural EAs reporting that they did not drink alcohol and thus felt alcohol-related posts were irrelevant. This reinforced the decision to focus the intervention on all 6 cancer risk behaviors rather than just one, so participants would likely receive at least some posts relevant to them. A final salient theme was the need for relatable characters in the posts. This request led us to partner with EAAB members and content experts to identify EAs who volunteered to record videos in which they talked about their experiences with each health behavior. These videos were developed following best practices for health communication via microvideos [47-49], with scripting designed to concisely convey the most important and relevant information and be emotionally engaging and authentic. Revised posts were reviewed by the health behavior content experts to ensure information accuracy.

XSL•FO
RenderX

Table 3. Formative rural emerging adult surveys respondent demographics.

Category	Total (N=297)
Age (years) (n=297), mean (SD)	22.5 (2.49)
State of residence (n=297), n (%)	
Arizona	29 (9.8)
Colorado	34 (11.4)
Idaho	61 (20.5)
Montana	37 (12.5)
New Mexico	48 (16.2)
Nevada	26 (8.8)
Utah	32 (10.8)
Wyoming	30 (10.1)
Gender identity (n=297), n (%)	
Women	215 (72.4)
Men	70 (23.6)
Transgender/gender fluid/nonbinary	11 (3.7)
Ethnicity (n=297), n (%)	
Not Hispanic/Latino	216 (72.7)
Hispanic/Latino	72 (24.2)
Race[a] (n=297), n (%)	
American Indian/Alaska Native	26 (8.8)
Asian	4 (1.3)
Black/African American	17 (5.7)
White	210 (70.7)
More than one group	19 (6.4)
Employment[a,b] (n=197), n (%)	
Full-time	66 (38.1)
Part-time	49 (24.1)
Looking for work	37 (13.1)
Unemployed, caregiver, student	82 (43.2)
Living situation[b] (n=197), n (%)	
Alone	20 (13.0)
With parents	57 (22.2)
With roommates/friends	27 (14.1)
With spouse/partner	90 (46.2)
With other relatives	99 (34.3)
Education [c] (n=99), n (%)	
High school or less	54 (54.5)
Some college/trade school	17 (17.2)
Four-year college degree	20 (20.2)
Graduate degree	6 (6.1)
Health insurance[c] (n=99), n (%)	
Yes	63 (63.6)

XSL•FO
RenderX

Category	Total (N=297)
No	15 (15.2)
Who helps with health decisions?[a,c] (n=99), n (%)	
I make them alone	14 (14.1)
Parents	56 (55.6)
Roommates/friends	10 (10.1)
Spouse/partner	40 (40.4)
Others	15 (15.2)

[a]Multiple responses could be selected, so totals may not equal 100%.

[b]Question not asked in the third round of surveys.

[c]Question not asked in the second or third round of surveys.

XSL•FO
RenderX

Table 4. Thematic categories of feedback with reactions and recommendations for posts by rural emerging adults (EAs) responding to the formative surveys (N=297), Emerging Adult Advisory Board (EAAB) members (N=17), and Stakeholder Organization Advisory Board (SOAB) members (N=16).

Theme	Definition	Example quote on how to improve posts	Resultant modifications
Comment			
Message-specific reaction	Unique specific reactions, opinions, or inquiries about the message or content.	• What product do you use to clear up skin? • Groceries are very expensive nowadays, you can just go shopping and take whatever you want. You need to think twice… • Wow, I didn't know this! • Maybe I should try this. • How urgently should I get it [HPV vaccine]?	Specific unclear messages were revised; additional information was provided
Affirmative comment	Positive receptivity or relatability, makes a direct affirmative comment, associates a positive reflection, or indicates a positive personal resonance on the message, content, or creator.	• That this statement is true • That i like the message • Thank you for trying to help society • How helpful this ad would be • This is great…how things like this can help people • I appreciate the vote [of] confidence • How I can relate	Prioritization of posts that received ample affirmations
Improvement			
Suggestions on image	Improving visual elements that would enhance the overall impact and reach of the posts. Selecting images that are, but not limited to, authentic, relevant to the topic, visually appealing, and tailored to look like and engage the intended audience effectively.	• Using a real image as opposed to clip-art/AI art style image will help, as I would immediately see this as an ad and scroll past it. Also, the tips at the top make the block of text too large, leading to less people reading it. • Change the stock photo to a clearer image, this one is a bit blurry • Since it is geared towards young adults, I would change the picture to be of a young adult rather than an older woman.	Inclusion of more realistic images, improvements in image resolution, and emphasis on younger appearing characters in posts
Dislike or wrong audience	The content, although potentially valuable or well-constructed, fails to engage them personally; misalignment with their interests; direct or indirect discontent with the post; lack of connection with the message or intent.	• I'm sure it's helpful to many people but for me it isn't • It's a great post just not for me because I don't drink. • It is well made, just irrelevant to me. • I don't really keep up with influencer content, I don't know why I should stop and listen or care about this person's experience with quitting	Revision of posts deemed irrelevant
More information requested	Content, while engaging, lacks adequate detail or depth.	• I would just add more facts of why you overeat when you're watching a show because it can happen • More facts to back up the information • Give a description about what HPV is for people who do not know and how one might contract this	Revision of posts to include additional detail where possible; additional posts created when more in-depth information was requested
Reinforce source or suggestions on source	Demands for greater transparency and verification concerning the origins or credibility of the information presented in post; any suggestion about the source materials or the need to further elaborate on them; avoid certain sources due to local political climate, testimonials must seem trustworthy, such as those from partnering with influencers.	• Make sure any links provided and information listed is factual and from a trusted source. • The "doctor" needs to have credentials posted so he is more believable. Anyone can say they are a doctor. • I would add sources that you could click on so you could read on more about it • Maybe linking sites to help people quit • We are just a very politically charged town, I know some will see CNN and just avoid it	Inclusion of source material from various outlets and improved transparency of source material across posts

XSL•FO
RenderX

Theme	Definition	Example quote on how to improve posts	Resultant modifications
Suggestions on content	Direct and postspecific suggestions that reflect a collection of feedback aimed at improving the effectiveness or engagement; recommendations on, but not limited to, enriching the informational depth, avoid use of scare tactics, shocking words, or condescension, keep tone positive, optimizing visual appeal, adding interactive content, and content's relevance and accuracy.	• Add subtitles • I would tell a story to help build common ground with people who are addicted • Maybe get a more attractive guy have him lose the hat, people stop scrolling for people pleasing to the eye • Explain why quitting nicotine is hard and not to beat yourself up if you can't quit immediately • If she was wearing reflective clothing it'd get the point across better. • I would show examples of what could happen, perhaps provide stories from real life people as well • I'd be more likely to click on a poll post to also see how other people responded	Subtitles added to video posts, emphasis on images that align with main messages, increased use of interactive posts like polls, addition of posts with real-world rural EAs for relatability
Too much information presented	Concerns about the overload of text or information, highlights the need for simplification and streamlining in content presentation, emphasizing that concise and clear communication often resonates better where attention spans are short. Brief posts and videos are preferred.	• I would make it a little more less detailed • Maybe make the video a little shorter, or have graphic on the screen with the overarching message (like the power of new habits or something) • This post is direct and straight to the point. It does have a lot of text, so I would find a way to shorten the text to make it more enticing to read. • Limit what is said before the post. I've recently looked up the information it is talking about but it says a bit too much before having to click the link.	Revision of posts to reduce text, shorten videos (limited to under 2 minutes), and abbreviate post captions
Unclear message	Content's intended message is not comprehensible or straightforward, leading to confusion and diminished engagement.	• I couldn't really tell if it was about limiting screen time or food intake, and didn't really get how it was connecting them, so maybe make the overall message/goal clearer • Use less biased language, as it is already painting these sponsorships in a bad light, which will affect who clicks on the link more than a neutral factual title • Starting with something beside the question, or maybe a different question. It just didn't feel particularly intriguing. • Make the message more clear. I do not really understand this video	Unclear messages were revised and reviewed with EAAB and SOAB members and content experts to improve clarity in messaging

EAAB and SOAB Feedback and Resultant Modifications

Finally, feedback from the EAAB and SOAB largely focused on using widely recognized sources instead of smaller, more local sources for posts due to the fact that EAAB members stated they would feel weary of information provided by a source they did not recognize and providing links to every source used in the messaging as EAAB members felt the need to vet the information for themselves in order to trust it (Table 4).

PEAK Wellness Pilot Study

There were 26 EAs who completed the pretest survey and took part in the pilot study. Three participants were lost to follow-up at posttest, with 23 (89%) participants completing posttests. Participants were 85% female and 35% Hispanic, with a median age of 23 years (SD 2.32). At pretest, many EAs reported having elevated cancer risk factors: 53% engaged in <150 minutes of moderate-to-vigorous physical activity weekly, 85% had low daily intake of fruits and vegetables, 35% used nicotine products, 58% had binged alcoholic beverages in the past 30 days, 65% were sunburned in the past 3 months, and 38% had not received the HPV vaccination.

Among all participants there were 1060 post views (mean per post: 16.6, mean per participant: 40.8), 346 reactions (mean per post: 5.4, mean per participant: 13.3), and 72 comments (mean per post: 1.2, mean per participant: 3.0). The participants viewed most posts and reacted to some posts (Table 5). Comments from participants on the posts included requests for more information, and the intervention manager posted 6 responses. EAs viewed posts on all topics a similar number of times, with most posts receiving 16-17 views. Posts on nicotine product use, alcohol consumption, diet, and physical activity received the most reactions (Table 6). The higher viewership of the physical activity posts was due in part to being the topic for the one focus week of posts in the pilot feed. Media education posts received the most comments. Most participants reported reading Facebook posts from private groups in general at least once or more per day, increasing from 58% at pretest to 77% at posttest. Feedback about the Facebook group was largely positive: at posttest 83% strongly agreed or agreed they would like to use the group, thought it was easy to use (91%), felt confident using the group (82%), and thought it was user-friendly (91.7% said it was good, excellent, or best imaginable, Table S2 in Multimedia Appendix 1).

XSL•FO
RenderX

Table 5. Aggregated feedback on posts reviewed in the formative surveys (64 posts, across 26 participants).

Engagement	Total	Mean per post	Mean per participant
Views	1060	16.6	40.8
Reactions (eg, likes)	346	5.4	13.3
Comments	72	1.2	3.0

Table 6. Aggregated engagement on posts reviewed in the formative surveys (64 posts, across 26 participants).

Post topic	Number of posts	Views, N (mean/post)	Reactions, N (mean/post)	Comments, N (mean/post)
Physical activity	13	216 (16.6)	77 (5.9)	14 (1.1)
Diet	7	118 (16.9)	44 (6.3)	10 (1.4)
Sun safety	7	115 (16.4)	30 (4.3)	8 (1.1)
Tobacco cessation	6	99 (16.5)	56 (9.3)	2 (0.3)
Alcohol reduction	7	125 (17.9)	52 (7.4)	5 (0.7)
HPV[a] vaccination	7	115 (16.4)	26 (3.7)	4 (0.6)
Media education	3	52 (17.3)	15 (5.0)	8 (2.7)
Family communication	2	28 (14.0)	2 (1.0)	0 (0.0)

[a]HPV: human papillomavirus.

Discussion

Principal Findings

The purpose of this paper was to illustrate the systematic, iterative process for developing and pilot-testing a theory-based, web intervention that engages rural EAs and promotes healthy behaviors for cancer prevention in the United States. The process described herein fills a gap in the existing literature on realistic timelines, resources, and staffing required to rigorously design a theory-based web intervention by including practical examples, clear iterative steps in developing the intervention content, and community engagement strategies. We demonstrated how to combine health behavior theory, feedback from EAAB and SOAB advisory boards, and web design to develop a robust health intervention. Below, we describe keys to success, lessons learned (Textbox 1), and limitations of this approach to add to the literature on how to rigorously design reproducible web-based interventions about healthy behaviors.

Textbox 1. Lessons learned in developing a social media–based health education intervention.

- Staffing and stakeholders: Coalescing and maintaining stakeholders required cultivating networks via dedicated personnel effort, and the strongest stakeholder engagement resulted from continuous quarterly engagement.
- Post creation: Developing theory-informed, community-engaged content for posts required an iterative design and revision process that included the community with stakeholders, content experts, and media experts.
- Protocols: In addition to post creation, which largely occurred before launching the intervention, protocols for handling questions and comments, promoting audience engagement, and scheduling posts were useful for delegating tasks and enhancing the reproducibility of dynamic interventions.

Designing a successful web-based intervention required an iterative and responsive approach to the development and testing of post content, delivery processes, and engagement strategies. Our process involved continuous refinement of Facebook posts, ensuring that the posts were relevant, engaging, and theoretically grounded. We built, tested, and revised messages based on feedback from participants, allowing for an agile development process that incorporated 4 groups of experts: a core intervention post development team, health behavior content experts, emerging adult and stakeholder organization advisory boards, and an administrative core. Organizing these expert stakeholders had implications for staffing and resources because it required defining roles, maintaining iterative communication, providing leadership, and delegating. Describing these components of the community-engaged research process expands existing literature by demonstrating realistic timelines, resources, and personnel requirements. Our multi-round, expert-informed development process reflects an agile, person-based approach consistent with recommendations from Yardley et al [50], who emphasize iterative refinement grounded in end user perspectives. Unlike traditional static digital interventions, our approach involved continuous testing and revision across multiple stakeholder groups to increase the ability of participants to enroll in and engage with our content, aligning with emerging literature calling for more engaging, adaptive, and co-designed digital health tools [51]. Our iterative process enabled us to balance the needs and priorities of the communities we aimed to engage—considering factors such as content sources, design elements, and stylistic choices—alongside insights from health behavior content experts and best practices for Facebook

XSL•FO
RenderX

engagement. While prior social media–based cancer prevention interventions have largely targeted older adult or urban populations [52,53], our focus on rural EAs (ages 18-26 years) is rare in digital intervention research. This population-specific focus expands the literature by demonstrating how social media messaging must be contextually adapted to audience characteristics, including rural identity and platform use patterns.

This intervention was not only about delivering information but also about motivating audiences to critically engage with web-based content, underscoring the importance of strategic engagement strategies. Given the widespread desire for reliable, expert-driven web-based health content, our strategy involved providing evidence-based information in the digital space. Unlike most interventions that provide one-time educational modules on media literacy [54-57], ours provides media literacy education via real-time sourcing and question-and-answer opportunities, improving responsiveness through timely information correction and availability of content experts to address information gaps. A unique contribution of our project is insight into the digital skillsets required for creating a web-based intervention (eg, web-based platform format, digital editing, algorithm literacy). The intervention was fundamentally community-driven, shaped by ongoing formative research, engagement with stakeholders, and direct interactions with EAs. Recognizing that there is no singular "perfect" message or approach, we focused on key considerations to guide content selection and refinement (eg, source material, visual appeal, representativeness of rurality). While refining our messaging, we also applied a system to assess message acceptability and audience reception through multiple surveys with the target population, whereas most social media-based health messaging studies lack sufficient survey pretesting with small samples [58,59]. This was of utmost importance because it helped us understand which messages were and were not engaging to rural EAs, beyond our EAAB.

Two additional strategies for web-based intervention development emerged. The first is to include information sources that feel reliable to participants, and here we acknowledge that participant perceptions of reliability are critical, particularly given that rural audiences likely differ from urban audiences in their trust patterns in health information sources [60]. A key component of our intervention was training users to assess the quality of web-based information and equipping them with the skills to evaluate content critically, which we refer to as media education. Ideally, this media education will foster the perception of our feed as a credible and dependable resource, as well as help EAs make informed decisions about information from competing sources. Second, including content experts in user-generated engagement videos was seen as an important strategy to enhance the credibility of the research team and the intervention health messages. User-generated engagement videos are a known strategy in advertising and web-based engagement to build trust and authenticity [61], but are scarcely accepted in academic research, which often values professionally designed video content featuring influencers or clinical experts [62].

Designing posts that are theoretically informed and community-engaged required an iterative process of review and revision. In our case, it took two and a half years to develop 400 theoretically informed posts by a team that included 7 media experts, 11 health behavior content experts, 29 community advisors, 26 pilot study participants, and 297 formative survey research participants. The extent of this effort is a likely reason most social media–based interventions fail to incorporate theory [63-66]. To balance structure with responsiveness, we adopted an agile development process, starting with an initial library of posts while continuously refining and expanding content in response to EAAB and SOAB discussions, emerging health information trends, and relevant current events. A yearlong intervention was launched in a randomized trial in March 2025 with twice-daily posts. At the time of the trial launch, nearly half of the posts were developed, with the remaining portion being produced in near real-time to maintain relevance and responsiveness to current events, and reactions and comments from EA users. Our description of the web-based intervention process provides rationale and justification for the inclusion of stakeholder groups, professional networks, and research personnel's effort to iteratively design a robust web-based intervention.

The pilot test provided valuable insights into the potential acceptability of our messaging and the demographics of our audience. A key focus was examining the prevalence of cancer risk factors and understanding how our messages resonated with a group of EAs by how they engaged with the messages. One of the key lessons learned during the pilot study was the necessity of engaging with audience comments in real-time—a departure from traditional public health communication approaches that historically emerge from larger organizations and government, with little back-and-forth interaction with individuals. This required the development of a structured moderator protocol, including guidelines for managing post schedules and crafting thoughtful reactions to audience engagement, including responding to questions. We relied on this dynamic process to uncover trends in user engagement and tailor our final posts and schedules accordingly. This approach should ensure that our intervention remains relevant, engaging, and impactful.

Interestingly, none of the comments we received during the pilot study actively disputed our messages. Instead, they reflected EAs' genuine desire for more information, highlighting the growing complexity of the web-based information ecosystem and the challenges EAs face in navigating it. Determining credible information has become increasingly challenging for the public, underscoring the need for clear, accessible, and responsive health communication strategies. Pretesting participants' views of the credibility of national and local sources was an essential early step in intervention development that should repeated as views of national and local health sources may evolve over time. Adaptability in web-based public health interventions was also essential to ensure that messaging remained both informative and engaging in an ever-evolving digital landscape. For example, health information from a local source may be viewed as credible because of its familiarity to participants or ability to relate it to a local context while information from a national source may be seen as credible because of name recognition and large resources to access key information, demonstrating the nuance of how participants

XSL•FO
RenderX

perceive credibility and the need to determine and respond to participants' perceptions.

Limitations

Several limitations to the generalizability of this analysis of creating web-based health communication should be acknowledged. Our study targeted EAs, and our analysis may be limited to this age group, which is highly media-savvy and among the heaviest users of digital platforms. Our development process may also apply to interventions for middle-aged and older adults who are also active on the web. However, it may not fully translate as these older groups tend to have different web-based engagement patterns (eg, higher use of different platforms). Additionally, this formative research focused on the Mountain West region, a geographic area with a highly rural demographic composition compared with the rest of the United States, which may additionally limit the generalizability of our findings. Furthermore, the intervention is currently limited to the Facebook platform for its wide use (even among EAs) and private group function that enables us to exercise experimental control in the randomized study design. However, many Americans use other platforms for different media formats and topics. For instance, YouTube and TikTok offer primarily video-based content that may appeal to different demographics or enhance engagement better than Facebook alone can achieve. Future research could explore multi-platform strategies to extend the reach and impact of web-based interventions across EAs and other audiences.

The methods also had 2 limitations. The sample size for the pilot intervention was generally small, as is common in intervention development research [67,68], limiting the representativeness of the EA population and the power of the statistical tests. While the measures used in the pilot study are well-established with good psychometric properties, they rely on self-report, which may produce social desirability biases or recall bias. The larger trial is designed to conduct validation methods of these measures as well. However, we did include observational measures (ie, Facebook metrics) of engagement with the intervention feed in the pilot study.

Conclusions and Implications

We described a theory-informed community-engaged process for developing a web-based health behavior change intervention, PEAK Wellness. The methods and results provide a novel roadmap for developing community-driven web-based interventions. Community-engaged strategies improve the relevance, feasibility, and cultural fit of health behavior interventions by incorporating community priorities and contextual knowledge into design and pilot-testing. Such approaches can increase recruitment, retention, and implementation potential, while promoting equity and shared ownership, which is especially important in social media interventions that are implemented in geographically diverse areas. These benefits, however, come with added demands for time, partnership building, and iterative adaptation. The resources and effort required to appropriately engage stakeholders, researchers, and the target population, rural EAs, were substantial but essential for developing an engaging and acceptable intervention. This study provides insight that can be used to enhance the rigor and reproducibility of social media intervention designs and facilitate realistic planning of procedures for future web-based health behavior interventions. Testing of the PEAK Wellness intervention has recently commenced in a rigorous pragmatic trial, using a randomized stepped-wedge design, enrolling EAs aged 18-26 years in rural counties in the Western United States. Our dynamic, iterative intervention development process will allow us to remain responsive to the evolving digital landscape while maintaining a foundation of evidence-based communication to promote cancer risk reduction among our sample of rural EAs.

Acknowledgments

We acknowledge the support of former faculty and staff doctors Cynthia A Thompson and Jenna Hatcher, as well as staff Christopher F Jones, Kaila Christini, John A Torres, Emilia Yessenya Barraza Perez, and Tatiana Gerena for their contributions to this project. The authors declare the use of generative artificial intelligence in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: Literature search to find and identify articles that are likely relevant to the research reported in the manuscript. The GAI tool used was ChatGPT 5.1. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This study was supported by grants from the National Institutes of Health National Cancer Institute (R01CA268037, P30CA118100, P30CA046934, P30CA023074, and P30CA042014). The opinions are those of the authors. The funding agency had no input on the contents of this paper.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to participant privacy and confidentiality, but are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

DBB, ALS, CAT, DK, DT, KLH, ELW, BJW, WGW, KN, CKB, DDG, EAB, JSG, JH, and DW conceptualized the study, designed methods, and secured extramural funding. DBB, ALS, JSG, DK, DT, and MS are supervising project activities. AK, DDG, AKY, TG, KC, JB, and AS are managing day-to-day study activities. All authors reviewed and approved the manuscript before submission.

Conflicts of Interest

DBB, BJW, WGW, KN, JB, AS, and AK receive a salary from Klein Buendel, Inc. DBB's spouse is an owner of Klein Buendel, Inc. DK is a consultant for Merck and has received 2 Merck Investigator Studies Program research awards. DT receives compensation from the International Life Sciences Institute (ILSI) as the Editor-in-Chief for Nutrition Reviews. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Supplemental tables.
[DOCX File , 34 KB - jmir_v28i1e80803_app1.docx]

Multimedia Appendix 2
PEAK Wellness Chat Pilot Pretest Survey.
[PDF File (Adobe PDF File), 376 KB - jmir_v28i1e80803_app2.pdf]

Multimedia Appendix 3
STROBE checklist.
[DOC File , 83 KB - jmir_v28i1e80803_app3.doc]

References

1. Arnett JJ. Emerging adulthood: a theory of development from the late teens through the twenties. Am Psychol 2000;55(5):469-480. [doi: 10.1037/0003-066x.55.5.469]
2. Daw J, Margolis R, Wright L. Emerging adulthood, emergent health lifestyles: sociodemographic determinants of trajectories of smoking, binge drinking, obesity, and sedentary behavior. J Health Soc Behav 2017;58(2):181-197 [FREE Full text] [doi: 10.1177/0022146517702421] [Medline: 28661779]
3. Chen X, Orom H, Hay JL, Waters EA, Schofield E, Li Y, et al. Differences in rural and urban health information access and use. J Rural Health 2019;35(3):405-417 [FREE Full text] [doi: 10.1111/jrh.12335] [Medline: 30444935]
4. Social Media Fact Sheet. Pew Research Center. 2025. URL: https://www.pewresearch.org/internet/fact-sheet/social-media/ [accessed 2025-03-04]
5. Sarkar U, Le GM, Lyles CR, Ramo D, Linos E, Bibbins-Domingo K. Using social media to target cancer prevention in young adults: viewpoint. J Med Internet Res 2018;20(6):e203 [FREE Full text] [doi: 10.2196/jmir.8882] [Medline: 29871850]
6. Veil SR, Buehner T, Palenchar MJ. A work - in - process literature review: incorporating social media in risk and crisis communication. J Conting Crisis Manage 2011;19(2):110-122. [doi: 10.1111/j.1468-5973.2011.00639.x]
7. Breland JY, Quintiliani LM, Schneider KL, May CN, Pagoto S. Social media as a tool to increase the impact of public health research. Am J Public Health 2017;107(12):1890-1891. [doi: 10.2105/AJPH.2017.304098] [Medline: 29116846]
8. Steffens MS, Dunn AG, Leask J, Wiley KE. Using social media for vaccination promotion: practices and challenges. Digit Health 2020;6:2055207620970785 [FREE Full text] [doi: 10.1177/2055207620970785] [Medline: 35173976]
9. Kata A. Anti-vaccine activists, web 2.0, and the postmodern paradigm--an overview of tactics and tropes used online by the anti-vaccination movement. Vaccine 2012;30(25):3778-3789. [doi: 10.1016/j.vaccine.2011.11.112] [Medline: 22172504]
10. Warner EL, Basen-Engquist KM, Badger TA, Crane TE, Raber-Ramsey M. The online cancer nutrition misinformation: a framework of behavior change based on exposure to cancer nutrition misinformation. Cancer 2022;128(13):2540-2548 [FREE Full text] [doi: 10.1002/cncr.34218] [Medline: 35383913]
11. Broniatowski DA, Jamison AM, Qi S, AlKulaib L, Chen T, Benton A, et al. Weaponized health communication: twitter bots and russian trolls amplify the vaccine debate. Am J Public Health 2018;108(10):1378-1384. [doi: 10.2105/AJPH.2018.304567] [Medline: 30138075]
12. Chou WS, Prestin A, Lyons C, Wen K. Web 2.0 for health promotion: reviewing the current evidence. Am J Public Health 2013;103(1):e9-e18. [doi: 10.2105/AJPH.2012.301071] [Medline: 23153164]
13. Vaccine safety communication in the digital age. World Health Organization. 2018. URL: https://apps.who.int/iris/handle/10665/311961 [accessed 2025-03-04]
14. Chou WS, Oh A, Klein WMP. Addressing health-related misinformation on social media. JAMA 2018;320(23):2417-2418. [doi: 10.1001/jama.2018.16865] [Medline: 30428002]

XSL•FO
RenderX

15. Yang YT, Broniatowski DA, Reiss DR. Government role in regulating vaccine misinformation on social media platforms. JAMA Pediatr 2019;173(11):1011-1012. [doi: 10.1001/jamapediatrics.2019.2838] [Medline: 31479099]
16. Sylvia Chou W, Gaysynsky A, Cappella JN. Where we go from here: health misinformation on social media. Am J Public Health 2020;110(S3):S273-S275. [doi: 10.2105/AJPH.2020.305905] [Medline: 33001722]
17. Probst J, Eberth JM, Crouch E. Structural urbanism contributes to poorer health outcomes for rural America. Health Aff (Millwood) 2019;38(12):1976-1984. [doi: 10.1377/hlthaff.2019.00914] [Medline: 31794301]
18. Cosby AG, McDoom-Echebiri MM, James W, Khandekar H, Brown W, Hanna HL. Growth and persistence of place-based mortality in the United States: the rural mortality penalty. Am J Public Health 2019;109(1):155-162. [doi: 10.2105/AJPH.2018.304787] [Medline: 30496008]
19. About rural health. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/rural-health/php/about/index.html [accessed 2025-07-02]
20. Wallerstein N, Duran B, Oetzel JG, Minkler M. Community-Based Participatory Research for Health: Advancing Social and Health Equity. Hoboken ,New Jersey: John Wiley & Sons; 2017.
21. Israel BA, Schulz AJ, Parker EA, Becker AB. Review of community-based research: assessing partnership approaches to improve public health. Annu Rev Public Health 1998;19:173-202. [doi: 10.1146/annurev.publhealth.19.1.173] [Medline: 9611617]
22. Buller DB, Sussman AL, Thomson CA, Kepka D, Taren D, Henry KL, et al. #4Corners4Health social media cancer prevention campaign for emerging adults: protocol for a randomized stepped-wedge trial. JMIR Res Protoc 2024;13:e50392 [FREE Full text] [doi: 10.2196/50392] [Medline: 38386396]
23. Gottfried J. Americans' social media use. Pew Research Center. URL: https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/ [accessed 2025-03-04]
24. Bandura A. Social Foundations of Thought and Action: A Social Cognitive Theory. Englewood Cliffs, NJ, US: Prentice-Hall, Inc; 1986:617.
25. Deci EL, Ryan RM. Self-determination theory: a macrotheory of human motivation, development, and health. Can Psychol 2008;49(3):182-185. [doi: 10.1037/a0012801]
26. Rogers EM. Diffusion of Innovations. New York City: Simon and Schuster; 2003.
27. Rural-urban continuum codes. Economic Research Service. URL: https://www.ers.usda.gov/data-products/rural-urban-continuum-codes [accessed 2025-07-02]
28. Rural-urban commuting area codes. Economic Research Service. URL: https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes [accessed 2025-05-21]
29. Dynata. 2025. URL: https://www.dynata.com/ [accessed 2025-03-04]
30. Verasight. 2025. URL: https://www.verasight.io/ [accessed 2025-03-04]
31. Fisher WR. Narration as a human communication paradigm: the case of public moral argument. Commun Monogr 2009;51(1):1-22. [doi: 10.1080/03637758409390180]
32. Baker M. Preventing preventing skin cancer in adolescent girls through intervention with their mothers. East Tennessee State University. 2013. URL: https://dc.etsu.edu/etd/1163 [accessed 2025-12-05]
33. Pagoto SL, Schneider KL, Oleski J, Bodenlos JS, Ma Y. The sunless study: a beach randomized trial of a skin cancer prevention intervention promoting sunless tanning. Arch Dermatol 2010;146(9):979-984 [FREE Full text] [doi: 10.1001/archdermatol.2010.203] [Medline: 20855696]
34. Hillhouse JJ, Turrisi R. Examination of the efficacy of an appearance-focused intervention to reduce UV exposure. J Behav Med 2002;25(4):395-409. [doi: 10.1023/a:1015870516460] [Medline: 12136499]
35. Global physical activity questionnaire (GPAQ). World Health Organization. URL: https://www.who.int/publications/m/item/global-physical-activity-questionnaire [accessed 2025-06-03]
36. Thompson FE, Midthune D, Subar AF, Kahle LL, Schatzkin A, Kipnis V. Performance of a short tool to assess dietary intakes of fruits and vegetables, percentage energy from fat and fibre. Public Health Nutr 2004;7(8):1097-1105. [doi: 10.1079/PHN2004642] [Medline: 15548349]
37. Hewawitharana SC, Thompson FE, Loria CM, Strauss W, Nagaraja J, Ritchie L, et al. Comparison of the NHANES dietary screener questionnaire to the automated self-administered 24-hour recall for children in the healthy communities study. Nutr J 2018;17(1):111 [FREE Full text] [doi: 10.1186/s12937-018-0415-1] [Medline: 30482218]
38. Viral hepatitis and liver disease. United States Department of Veterans Affairs. URL: https://www.hepatitis.va.gov/alcohol/treatment/audit-c.asp [accessed 2025-06-03]
39. Hughes J, Keely J, Niaura R, Ossip-Klein D, Richmond R, Swan G. Measures of abstinence in clinical trials: issues and recommendations. Nicotine Tob Res 2003;5(1):13-25. [Medline: 12745503]
40. Velicer WF, Prochaska JO. A comparison of four self-report smoking cessation outcome measures. Addict Behav 2004;29(1):51-60. [doi: 10.1016/s0306-4603(03)00084-4] [Medline: 14667420]
41. View HINTS questions. National Cancer Institute. URL: https://hints.cancer.gov/view-questions/all-hints-questions.aspx [accessed 2025-07-02]
42. Shoveller JA, Lovato CY. Measuring self-reported sunburn: challenges and recommendations. Chronic Dis Can 2001;22(3-4):83-98. [Medline: 11779421]

XSL•FO
RenderX

43. Lovato C, Shoveller J, Mills C. Canadian national workshop on measurement of sun-related behaviours [workshop report]. Chronic Dis Can 1999;20(2):96-100. [Medline: 10455042]
44. de Waal AC, van Rossum MM, Kiemeney L, Aben K. Reproducibility of self-reported melanoma risk factors in melanoma patients. Melanoma Res 2014;24(6):592-601. [doi: 10.1097/CMR.0000000000000089] [Medline: 24892955]
45. Hillhouse J, Turrisi R, Jaccard J, Robinson J. Accuracy of self-reported sun exposure and sun protection behavior. Prev Sci 2012;13(5):519-531 [FREE Full text] [doi: 10.1007/s11121-012-0278-1] [Medline: 22855253]
46. Moreno MA, D'Angelo J. Social media intervention design: applying an affordances framework. J Med Internet Res 2019;21(3):e11014 [FREE Full text] [doi: 10.2196/11014] [Medline: 30912754]
47. Zhu Z, Liu S, Zhang R. Examining the persuasive effects of health communication in short videos: systematic review. J Med Internet Res 2023;25:e48508 [FREE Full text] [doi: 10.2196/48508] [Medline: 37831488]
48. Emilia. The ultimate guide to microvideos – use cases, tools and tips for SaaS. Userpilot Blog. URL: https://userpilot.com/blog/microvideos-guide/ [accessed 2024-06-12]
49. Ester P, Morales I, Herrero L. Micro-videos as a learning tool for professional practice during the post-COVID era: an educational experience. Sustainability 2023;15(6):5596. [doi: 10.3390/su15065596]
50. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015;17(1):e30 [FREE Full text] [doi: 10.2196/jmir.4055] [Medline: 25639757]
51. O'Connor S, Hanlon P, O'Donnell CA, Garcia S, Glanville J, Mair FS. Barriers and facilitators to patient and public engagement and recruitment to digital health interventions: protocol of a systematic review of qualitative studies. BMJ Open 2016;6(9):e010895 [FREE Full text] [doi: 10.1136/bmjopen-2015-010895] [Medline: 27591017]
52. Valle CG, Tate DF, Mayer DK, Allicock M, Cai J. A randomized trial of a Facebook-based physical activity intervention for young adult cancer survivors. J Cancer Surviv 2013;7(3):355-368 [FREE Full text] [doi: 10.1007/s11764-013-0279-5] [Medline: 23532799]
53. Han CJ, Lee YJ, Demiris G. Interventions using social media for cancer prevention and management: a systematic review. Cancer Nurs 2018;41(6):E19-E31 [FREE Full text] [doi: 10.1097/NCC.0000000000000534] [Medline: 28753192]
54. Housten AJ, Gunn CM, Paasche-Orlow MK, Basen-Engquist KM. Health literacy interventions in cancer: a systematic review. J Cancer Educ 2021;36(2):240-252 [FREE Full text] [doi: 10.1007/s13187-020-01915-x] [Medline: 33155097]
55. Peipert JD, Lad T, Khosla PG, Garcia SF, Hahn EA. A low literacy, multimedia health information technology intervention to enhance patient-centered cancer care in safety net settings increased cancer knowledge in a randomized controlled trial. Cancer Control 2021;28:10732748211036783 [FREE Full text] [doi: 10.1177/10732748211036783] [Medline: 34565193]
56. Lazard AJ, Queen TL, Pulido M, Lake S, Nicolla S, Tan H, et al. Social media prompts to encourage intervening with cancer treatment misinformation. Soc Sci Med 2025;372:117950. [doi: 10.1016/j.socscimed.2025.117950] [Medline: 40096813]
57. Lyons B, King AJ, Kaphingst KA. A health media literacy intervention increases skepticism of both inaccurate and accurate cancer news among U.S. adults. Ann Behav Med 2024;58(12):820-831. [doi: 10.1093/abm/kaae054] [Medline: 39417815]
58. Tappin BM, Hewitt LB. Using survey experiment pretesting to support future pandemic response. PNAS Nexus 2024;3(11):pgae469 [FREE Full text] [doi: 10.1093/pnasnexus/pgae469] [Medline: 39484044]
59. Ricci-Cabello I, Bobrow K, Islam SMS, Chow CK, Maddison R, Whittaker R, et al. Examining development processes for text messaging interventions to prevent cardiovascular disease: systematic literature review. JMIR Mhealth Uhealth 2019;7(3):e12191 [FREE Full text] [doi: 10.2196/12191] [Medline: 30924790]
60. Lee W, Kim EM, Nemirovski EA, Kamprath S, Masel MC, Patel DI. Public trust in different sources of information: gaps in rural residents and cancer patients. Healthcare (Basel) 2025;13(6):640 [FREE Full text] [doi: 10.3390/healthcare13060640] [Medline: 40150490]
61. Ismail S, Abdul Latif R. Authenticity issues of social media: credibility, quality, and reality. World Acad Sci Eng Technol 2013;74:254-261 [FREE Full text]
62. Kaňková J, Binder A, Matthes J. Helpful or harmful? navigating the impact of social media influencers' health advice: insights from health expert content creators. BMC Public Health 2024 Dec 18;24(1):3511 [FREE Full text] [doi: 10.1186/s12889-024-21095-3] [Medline: 39696170]
63. Paul B, Headley-Johnson S. The impact of social media on health behaviors, a systematic review. Healthcare (Basel) 2025;13(21):2763 [FREE Full text] [doi: 10.3390/healthcare13212763] [Medline: 41228133]
64. Maher CA, Lewis LK, Ferrar K, Marshall S, De Bourdeaudhuij I, Vandelanotte C. Are health behavior change interventions that use online social networks effective? a systematic review. J Med Internet Res 2014;16(2):e40 [FREE Full text] [doi: 10.2196/jmir.2952] [Medline: 24550083]
65. Laranjo L, Arguel A, Neves AL, Gallagher AM, Kaplan R, Mortimer N, et al. The influence of social networking sites on health behavior change: a systematic review and meta-analysis. J Am Med Inform Assoc 2015;22(1):243-256 [FREE Full text] [doi: 10.1136/amiajnl-2014-002841] [Medline: 25005606]
66. Seiler J, Libby TE, Jackson E, Lingappa J, Evans W. Social media-based interventions for health behavior change in low- and middle-income countries: systematic review. J Med Internet Res 2022;24(4):e31889 [FREE Full text] [doi: 10.2196/31889] [Medline: 35436220]

XSL•FO
RenderX

67. Leon AC, Davis LL, Kraemer HC. The role and interpretation of pilot studies in clinical research. J Psychiatr Res 2011;45(5):626-629 [FREE Full text] [doi: 10.1016/j.jpsychires.2010.10.008] [Medline: 21035130]
68. Hertzog MA. Considerations in determining sample size for pilot studies. Res Nurs Health 2008;31(2):180-191. [doi: 10.1002/nur.20247] [Medline: 18183564]

Abbreviations

EA: emerging adult
EAAB: Emerging Adult Advisory Board
HPV: human papillomavirus
REDCap: Research Electronic Data Capture
RUCC: Rural Urban Continuum Codes
SOAB: Stakeholder Organization Advisory Board
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Stone; submitted 17.Jul.2025; peer-reviewed by B Uko, S Narayan, M Chukwuka, D Gyimah; comments to author 23.Sep.2025; revised version received 26.Nov.2025; accepted 02.Dec.2025; published 08.Jan.2026.

Please cite as:
Warner EL, Kinsey A, Walkosz BJ, Berteletti J, Nuss K, Small A, Woodall WG, Kepka D, Taren D, Skiba MB, Guest DD, Blair CK, Gordon JS, Wetter DW, Borrayo EA, Henry KL, Sussman AL, Buller DB
A Web-Based Cancer Prevention Intervention for Rural Emerging Adults: Mixed Methods Development and Pilot-Testing Study
J Med Internet Res 2026;28:e80803
URL: https://www.jmir.org/2026/1/e80803
doi:10.2196/80803
PMID:

©Echo L Warner, Alishia Kinsey, Barbara J Walkosz, Julia Berteletti, Kayla Nuss, Annelise Small, W Gill Woodall, Deanna Kepka, Douglas Taren, Meghan B Skiba, Dolores D Guest, Cindy K Blair, Judith S Gordon, David W Wetter, Evelinn A Borrayo, Kimberly L Henry, Andrew L Sussman, David B Buller. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effects of Artificial Intelligence Recognition–Based Telerehabilitation on Exercise Capacity in Patients With Hypertension: Randomized Controlled Trial

Qiuru Yao[1,2,3,4], MS; Baizhi Qiu[1,2,3,4], MS; Longlong He[5], MS; Qin Wang[1,2,3,4], MS; Jihua Zou[1,3,4,6,7], PhD; Donghui Liang[8], PhD; Shuyang Wen[1,2,3,4], MS; Yingchao Liu[1,3,4], MM; Gege Li[1,3,4,6], MS; Jinjing Hu[1,3,4,6], MS; Huan Ma[9*], PhD; Guozhi Huang[1,3,4,6*], PhD; Qing Zeng[1,3,4,6*], PhD

[1]Center of Rehabilitation Medicine, Zhujiang Hospital, Guangzhou, China

[2]School of Nursing, Southern Medical University, Guangzhou, China

[3]GuangDong Engineering Technology Research Center of Brain Function Assessment and Neuroregulation Rehabilitation, Guangzhou, China

[4]Institute of Exercise and Rehabilitation Science, Zhujiang Hospital, Guangzhou, China

[5]Department of Clinical Medicine, Xiamen Medical College, Xiamen, China

[6]School of Rehabilitation Sciences, Southern Medical University, Guangzhou, China

[7]Department of Rehabilitation Sciences, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)

[8]Department of Traditional Chinese Medicine, Zhujiang Hospital, Guangzhou, China

[9]Department of Cardiology, Guangdong Cardiovascular Institute, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Southern Medical University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Qing Zeng, PhD
Center of Rehabilitation Medicine
Zhujiang Hospital
NO.253, Industrial Avenue Middle
Guangzhou, 510280
China
Phone: 86 15989189468
Email: zengqingyang203@126.com

Abstract

Background: Hypertension remains a major global health challenge, significantly increasing cardiovascular and all-cause mortality risks. While exercise therapy is effective, conventional approaches face limitations in accessibility and personalization, compromising adherence. Artificial intelligence (AI)–assisted remote rehabilitation enables real-time monitoring and personalized guidance, offering a promising alternative. Nevertheless, its clinical benefits and applicability require further systematic validation.

Objective: This study aimed to evaluate the efficacy of an 8-week AI-assisted telerehabilitation program on improving exercise capacity and related health outcomes in patients with hypertension.

Methods: This prospective, dual-arm, parallel, open-label, randomized controlled trial enrolled 62 patients with hypertension recruited via convenience sampling. Participants were adults aged between 18 and 75 years with a confirmed hypertension diagnosis who were excluded for severe cardiac complications, recent myocardial infarction, unstable angina, or physical disabilities preventing exercise. The participants were randomly assigned (1:1) to an intervention group that received AI-assisted remote rehabilitation plus routine health education, or a control group that received health education and conventional offline exercise guidance. The supervised exercise program included warm-up, cardiorespiratory endurance, strength resistance, balance, and flexibility training, followed by a cooldown. Sessions lasted between 30 and 50 minutes and were performed at least 3 times weekly for 8 weeks. Assessments at baseline and 8 weeks included the 6-minute walk test (6MWT), cardiopulmonary exercise testing (CPET), International Physical Activity Questionnaire (IPAQ), Short-Form Health Survey 12 (SF-12), Patient Health Questionnaire-9 (PHQ-9), Generalized Anxiety Disorder-7 (GAD-7), exercise self-efficacy, blood pressure (BP), body weight, handgrip strength, and other health-related indicators. The primary outcome was the change in 6-minute walk distance (6MWD). Data were analyzed according to the intention-to-treat principle.

XSL•FO
RenderX

Results: Throughout the 8-week intervention period, no serious adverse events related to the AI-assisted telerehabilitation intervention occurred. After 8 weeks, the intervention group demonstrated significantly greater improvements than the control group in 6-minute walk distance (6MWD; adjusted mean difference 62.77, 95% CI 26.33-99.22; *P*=.002), systolic BP reduction (adjusted mean difference 4.11, 95% CI 0.11-8.28; *P*=.046), IPAQ score (adjusted mean difference 658.96, 95% CI 159.23-1158.69; *P*=.011), exercise self-efficacy score (adjusted mean difference 21.71, 95% CI 13.59-29.82; *P*<.001), total exercise time (adjusted mean difference 98.24, 95% CI 49.39-147.08; *P*=.001) peak oxygen uptake (peak VO_2) (adjusted mean difference 3.39, 95% CI 0.49-6.29; *P*=.026), and peak oxygen uptake percent predicted (peak VO_2%pred) (adjusted mean difference 11.58, 95% CI 2.06-21.10; *P*=.021).

Conclusions: Compared with conventional exercise rehabilitation, AI-assisted remote rehabilitation was found to improve exercise capacity, boost regular physical activity and exercise self-efficacy, and aid in systolic BP control among patients with hypertension. This study positioned AI-assisted rehabilitation as a scalable and effective strategy for real-world hypertension management. It further contributes actionable guidance for developing effective home-based exercise strategies tailored to populations with hypertension.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2300076451; https://www.chictr.org.cn/showproj.html?proj=208353

(J Med Internet Res 2026;28:e81400) doi:10.2196/81400

KEYWORDS

hypertension; artificial intelligence; telerehabilitation; lifestyle change; exercise habit formation; randomized controlled trial

Introduction

Hypertension is a major public health concern with an enormous economic and social burden [1]. The World Health Statistics 2023 report by the World Health Organization (WHO) states that the worldwide prevalence of hypertension reached 33%, and it remains an upward trend in the context of global aging [2]. Given the high prevalence and severity of hypertension, there is an urgent need to implement effective, widely available, and sustainable strategies for its management.

Current interventions for hypertension are divided into pharmacological and nonpharmacological treatments. While pharmacological therapy effectively lowers blood pressure (BP) and remains the clinical mainstay [3,4], its benefits are often limited by poor adherence and the risk of BP rebound upon discontinuation, which elevates the risk of major cardiovascular events [5,6]. In contrast, nonpharmacological management focuses on lifestyle modifications, such as regular physical activity, dietary changes (eg, salt intake restriction), and weight management. The health benefits of exercise are well-established—regular moderate-to-high intensity aerobic activity can reduce BP by an average of 11/5 mm Hg [7]. Furthermore, its efficacy in lowering BP may be comparable to that of pharmacological interventions [8].

Regular exercise, a key nonpharmacological intervention, significantly reduces hypertension risk, improves BP control, and plays a vital role in its prevention and management [9]. Regular exercise not only significantly lowers BP but also helps regulate weight, improve lipid and blood sugar levels, and promote overall metabolic function. Engaging in moderate physical exercise before the onset of hypertension, combined with a healthy lifestyle, can reduce the incidence of primary hypertension by 54% [10]. Moderate exercise can help delay the progression of hypertension, while excessive exercise may cause BP to rise, thereby increasing the risk of cardiovascular and cerebrovascular diseases. However, the current state of physical activity in patients with hypertension is not encouraging. Patients with hypertension report being less physically active compared with individuals without hypertension, and they are less likely to meet physical activity recommendations [11]. Furthermore, they often encounter numerous obstacles when it comes to maintaining regular physical activity, such as a lack of self-efficacy, social support, and adequate supervision [12]. Therefore, selecting appropriate exercise methods and training approaches for patients with hypertension is of great significance for their BP management [13,14].

The rapid development of digital therapeutics in recent years has the potential to be an important part of BP management in the future [15]. It is an emerging, promising field of medicine that aims to implement lifestyle changes and ultimately facilitate disease management using software programs such as smartphone applications and device algorithms [16,17]. Digital therapeutics are gaining ground in the fields of medicine and health care, and the HERB Digital Hypertension 1 (HERB-DH1) pivotal trial conducted in Japan is one such success case [18]. Although there are many mobile technologies available for improving BP management, unfortunately, only a few of them have been developed with the involvement of health care professionals or medical organizations, and the evidence of their long-term benefits in terms of BP management is still scarce [19,20]. Meanwhile, existing digital therapeutics (such as the HERB-DH1 behavioral algorithm) face limitations due to the absence of real-time motion feedback mechanisms, making it challenging to effectively monitor and ensure patient adherence to exercise prescriptions. This technological gap highlights the medical potential of 3D skeletal pose recognition technology. By capturing and analyzing human motion data, this technology enables real-time corrective guidance and dynamically adaptive interventions, serving as an intelligent monitor that safeguards exercise safety and efficacy.

Among numerous artificial intelligence (AI) technologies, posture recognition, as a technology capable of understanding and interpreting human behavior, is gradually becoming a hot

XSL•FO
RenderX

topic of research. As an important branch of AI, posture recognition technology enables machines to better understand human behavior by capturing, analyzing, and interpreting human movements, thereby achieving a leap forward in human-machine interaction. Posture recognition technology detects the position and orientation of a person or object from a video or image and can significantly improve the accuracy and quality of telerehabilitation services [21]. Human posture estimation techniques use input data, such as images, to help with reconstructing human representations to assist in measuring changes in patient movement and assessing limb or joint function [22] and to provide feedback during the treatment process to help patients understand and correct postural errors. Research has shown [23] that an AI-based telerehabilitation system, combined with 3D posture recognition technology, can be a viable alternative to exercise intervention for patients with sarcopenia. In our previous work in the field of low back pain using AI-assisted telerehabilitation technology, we found that AI-assisted multimodal movement-based telerehabilitation could significantly improve patients' lower back pain and provide better therapeutic outcomes than traditional movement-based telerehabilitation [24]. This technology significantly improves the accuracy of training guidance through high-precision motion capture and real-time feedback mechanisms, and its intervention effect is equivalent to that of traditional face-to-face training and conventional telerehabilitation programs, while optimizing the quality of standardized execution of rehabilitation training. However, there is insufficient evidence in the field of hypertension.

Considering the above, we utilized a novel smartphone app, developed under the guidance of health care professionals, to address this need. This app was designed to help patients with hypertension develop a habit of regular physical activity and improve their overall quality of life. Its advantage lies in providing online, personalized exercise prescriptions, thereby enabling a scientific, effective, and targeted remote training intervention. Accordingly, this study aimed to examine the effects of this AI-assisted remote rehabilitation system on exercise capacity, self-efficacy, psychological well-being, and related health indicators in patients with hypertension. We hypothesized that this intervention would positively influence these treatment outcomes compared to usual care.

Methods

Study Design

This study was a multicenter, open-label randomized controlled trial conducted in Guangdong Provincial People's Hospital (Guangzhou, China) and Zhujiang Hospital, Southern Medical University (Guangzhou, China), 2 large tertiary care hospitals in southern China. The reporting of this trial conformed to the CONSORT (Consolidated Standards of Reporting Trials) guidelines for randomized controlled trials, and the study protocol adhered to the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) checklist (Multimedia Appendix 1) [25,26]. The study was prospectively registered in the Chinese Clinical Trial Registry (ChiCTR2300076451) on October 9, 2023.

Ethical Considerations

Overview

This study was conducted in accordance with the principles of the Declaration of Helsinki and relevant Chinese regulations governing clinical research. The study protocol was reviewed and approved by the ethics committee of Zhujiang Hospital, Southern Medical University (2023-KY-190-01). The study was also filed in the Medical Research Registration Information System of the National Health Commission of China. All participants provided written informed consent prior to enrollment. The consent form detailed the study objectives, procedures, potential risks and benefits, and participants' rights. For participants who were unable to continue the study due to health or personal reasons, data were retained and analyzed in accordance with the intention-to-treat principle, as approved by the ethics committee.

Privacy and Confidentiality

All participant data were deidentified prior to analysis. Personal information and medical records were stored securely and accessible only to authorized research personnel. Specifically, for the AI-assisted telerehabilitation, the technology platform was designed to safeguard privacy by not recording any personal identifiers (eg, real names or ID numbers). Unless users explicitly opted in for video recording, the system did not capture or store any recognizable video footage. Instead, it processed the video stream in real time to extract only the coordinate data of 17 skeletal key points, using this anonymized information to compute exercise completion and quality scores. Data anonymity was maintained throughout the study and in all publications.

Compensation

Participants did not receive financial compensation for their involvement in the study. However, they were provided with free health education materials and access to the AI-based telerehabilitation application during the intervention period.

Use of Images and Identifiable Information

The images included in this paper (eg, app interfaces, training scenarios) do not contain any identifiable images of the participants. All screenshots and illustrations are system display interfaces and have been obtained with the consent of the system owner and the person who took the photos. Therefore, no additional consent for image publication was required.

Participant Selection and Recruitment

Recruitment was conducted from November 2023 to October 2024. Given the pragmatic nature of this trial and the need to consecutively enroll eligible patients, a convenience sampling method was employed. Participants were recruited from the cardiology outpatient clinic and via community advertisements. Participants were primarily recruited from the cardiology outpatient clinics of Guangdong Provincial People's Hospital (Guangzhou, China) and Zhujiang Hospital, Southern Medical University (Guangzhou, China), and through community advertisements in the surrounding areas.

XSL•FO
RenderX

Recruitment information was disseminated through a combination of online and traditional channels. Online strategies included utilizing social media platforms, such as WeChat, to publish recruitment announcements and conduct preliminary online screenings. Traditional methods involved on-site activities (eg, setting up recruitment booths in hospitals and communities to provide briefings and answer questions directly), physician referrals from established partnerships, and the distribution of printed leaflets and posters.

Inclusion and Exclusion Criteria

Inclusion Criteria

The inclusion criteria were as follows: age 18 to 75 years; diagnosis of essential hypertension, defined as an office systolic blood pressure (SBP) ≥140 mmHg and/or diastolic blood pressure (DBP) ≥90 mmHg; no cognitive impairment or mental illness; passed the physical activity readiness questionnaire screening or approved by a physician prior to participation; ability to communicate normally and cooperate with the questionnaire survey; not participating in other research involving healthy lifestyle promotion, particularly physical exercise; no prior exercise habit, defined as exercising less than 3 times per week for under 30 minutes per session; and ownership of a smartphone with access to WeChat and the internet.

Exclusion Criteria

The exclusion criteria were as follows: patients with severe or ineffectively controlled hypertension (SBP ≥180 mmHg and/or DBP ≥110 mmHg at rest); patients with pulmonary hypertension, severe hypertension, hypertensive crisis, unstable stage III hypertension, hypertensive encephalopathy, acute hypertension, or other serious complications such as severe arrhythmia, tachycardia, heart failure, unstable angina pectoris, or obvious adverse reactions to antihypertensive drugs with failure to control; patients with malignant tumors or hypertension combined with left ventricular ejection fraction <55%, aortic, mitral, tricuspid, or pulmonary stenosis, severe mitral, tricuspid, or pulmonary valve insufficiency, hypertrophic cardiomyopathy, acute infection, fundus hemorrhage, diabetic acidosis, gangrene of the lower limbs, severe hypothyroidism, or renal insufficiency; patients with motor organ injury who underwent bone or joint surgery within the past year, such as hip or knee replacement or spine surgery; patients unable to understand or answer questions or complete questionnaires due to subjective or objective reasons; and participants considered unsuitable for clinical trials by the researchers.

Sample Size

In this study, the sample size was calculated using the 6-minute walk test (6MWT) distance as the primary outcome indicator. According to the relevant literature [27,28], telemedicine has been shown to improve exercise capacity and physical function in patients with cardiovascular disease. Setting the minimal clinical difference significance (>33 m), 2-sided α=.05, and the test power 1-β 0.90. Calculations were performed using PASS software (version 15.0; NCSS LLC). Considering the loss and refusal rate of 20%, a minimum of 29 participants for each group was required, for a total of at least 58 participants.

Randomization and Blinding

Block randomization was performed using SAS software (version 9.3; SAS Institute Inc) to ensure balanced group sizes and prevent the potential for imbalance inherent in simple randomization. The randomization process allocated participants to either the intervention or control group using a block size of 6, across a total of 10 blocks, to generate the random sequence and obtain group allocations. The allocation sequence was concealed using opaque, sealed envelopes. All researchers responsible for patient recruitment and data collection were blinded to the sequence generation and were not involved in preparing the concealment envelopes.

Interventions

Remote Exercise Rehabilitation

Based on hypertension health education and exercise guidance, the intervention group used the “rehabilitation exercise prescription” App (RaphAI Health Technologies Limited) to carry out AI-assisted remote rehabilitation exercise intervention (Figure 1). The specific procedures included are detailed as follows.

XSL•FO
RenderX

Figure 1. Interface demonstration of the "Rehabilitation Exercise Prescription" mobile app used for artificial intelligence (AI)–assisted telerehabilitation. A: main landing page of the app; B: interface showing a personalized exercise prescription generated for a participant; C: screen displayed upon completion of an exercise session, summarizing performance metrics.

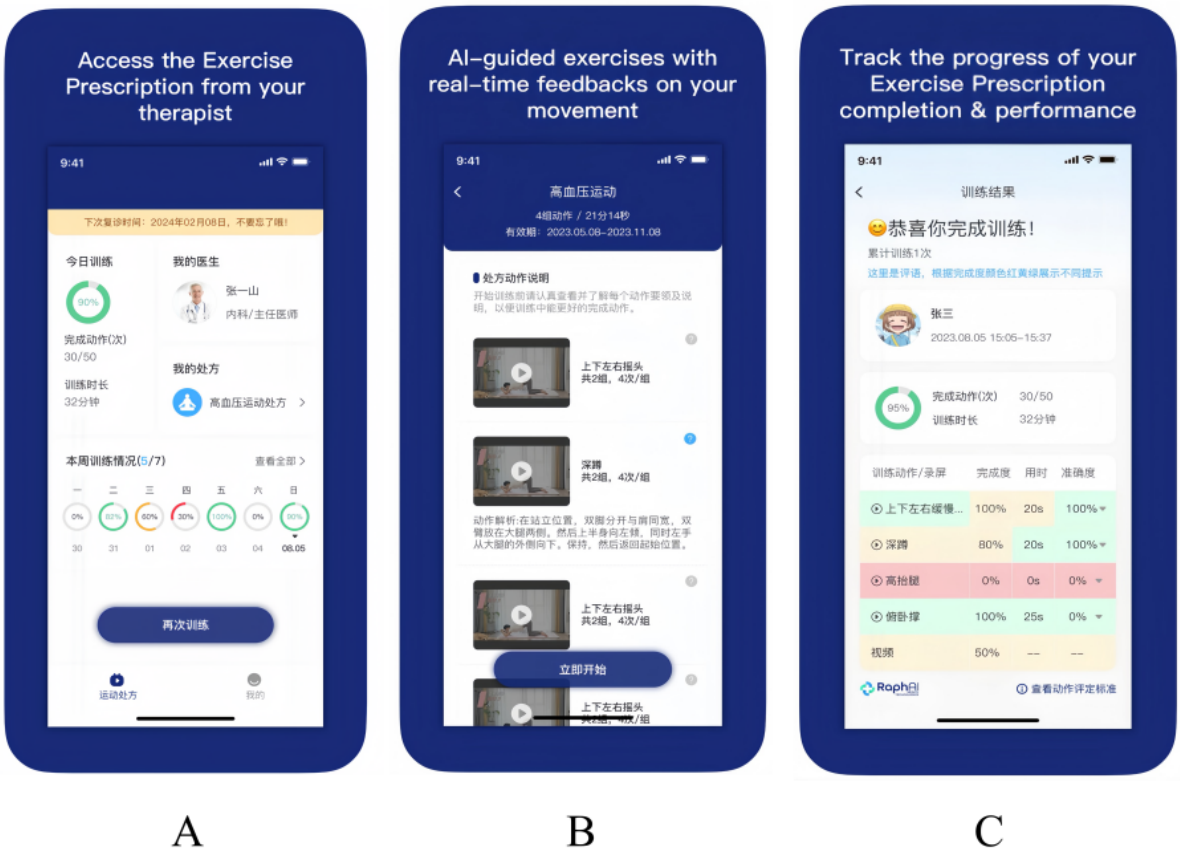

Assessment

Medical history was collected using the online assessment questionnaire, and each patient's general condition was evaluated through a combination of physical examination and questionnaire results. The assessment included (1) general information, including basic information (name, gender, age, contact information), behavioral data (smoking, drinking habits, eating habits), and medical history (past medical history, history of hypertension treatment, family history of hypertension, BP measurement); (2) current physical activity level (regular exercise, type and frequency of exercise); (3) anthropometric data (height, weight, waist circumference, and hip circumference); (4) the American Heart Association (AHA)/American College of Sports Medicine (ACSM) Health/Physical Fitness Preexercise Screening Questionnaire; (5) physical fitness assessments (cardiorespiratory endurance, strength, coordination, flexibility and balance assessment); and the (6) Physical Activity Readiness questionnaire (PAR-Q).

Exercise Prescription Generation and Delivery

For the prescription generation, 4 personalized cardiopulmonary rehabilitation prescriptions were customized according to the standards of the ACSM. After the assessment, the system determined the patient's risk stratification according to the self-assessment and offline assessment results and automatically generated exercise prescriptions.

The prescription was pushed to the patient 5 times a week for 30 to 50 minutes each time on a regular basis and was required to be completed at least 3 times a week. The content included warm-up, cardiopulmonary endurance training, strength resistance training, balance training, flexibility training, and cooldown.

AI Technology and Motion Capture Framework

The "Rehabilitation Exercise Prescription" app was built upon an AI engine that integrated TensorFlow, an open-source machine learning library, with the OpenPose architecture for real-time human pose estimation. This system employed a bottom-up approach using Part Affinity Fields to detect 17 key body joints (including eyes, nose, ears, shoulders, elbows, wrists, hips, knees, and ankles) and assembled them into a full-body skeleton without the need for physical markers. By processing video frames from the smartphone camera, the system tracked the coordinate changes of these skeletal key points across consecutive frames to quantify and analyze patient movement in real time.

Rehabilitation Training

The patients were trained according to the exercise prescription pushed by the app, and the action explanation was provided during the training. Computer AI visual recognition technology was used to identify 17 key bone points in the camera in a markerless way to monitor the training data in real time,

XSL•FO
RenderX

calculate the angle and displacement of different parts of the body in real time, and provide intelligent feedback and corrective guidance (Figure 2).

Figure 2. Examples of an artificial intelligence (AI)–assisted telerehabilitation process using the "Rehabilitation Exercise Prescription" app in patients with hypertension. A: scene during participant training; B: mobile interface showing real-time video and AI movement guidance.

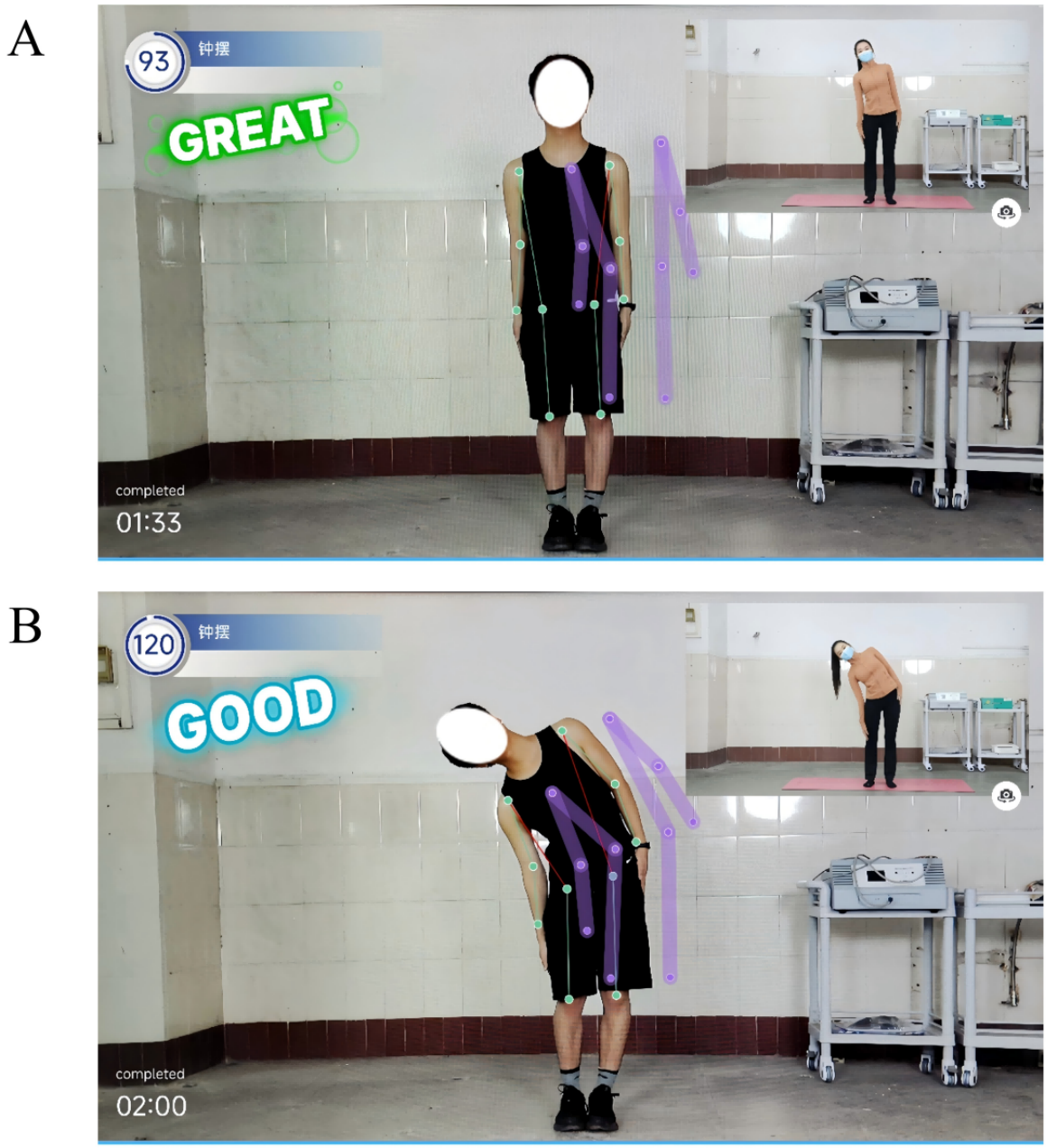

Follow-Up and Optimization

After each training, the system automatically sent the perceptual experience score and the training report. The report included the overall score (three grades: A, B, and C), completion degree, accuracy, and time for each action, and generated an exercise diary. To enhance adherence, the system also incorporated an automated reminder function that pushed notifications to prompt users to initiate their scheduled training sessions (Figure 3).

XSL•FO
RenderX

Figure 3. Postexercise evaluation and records via the "Rehabilitation Exercise Prescription" app. A: automatically generated assessment reports at the end of exercise; B: automated exercise diary recording interface.

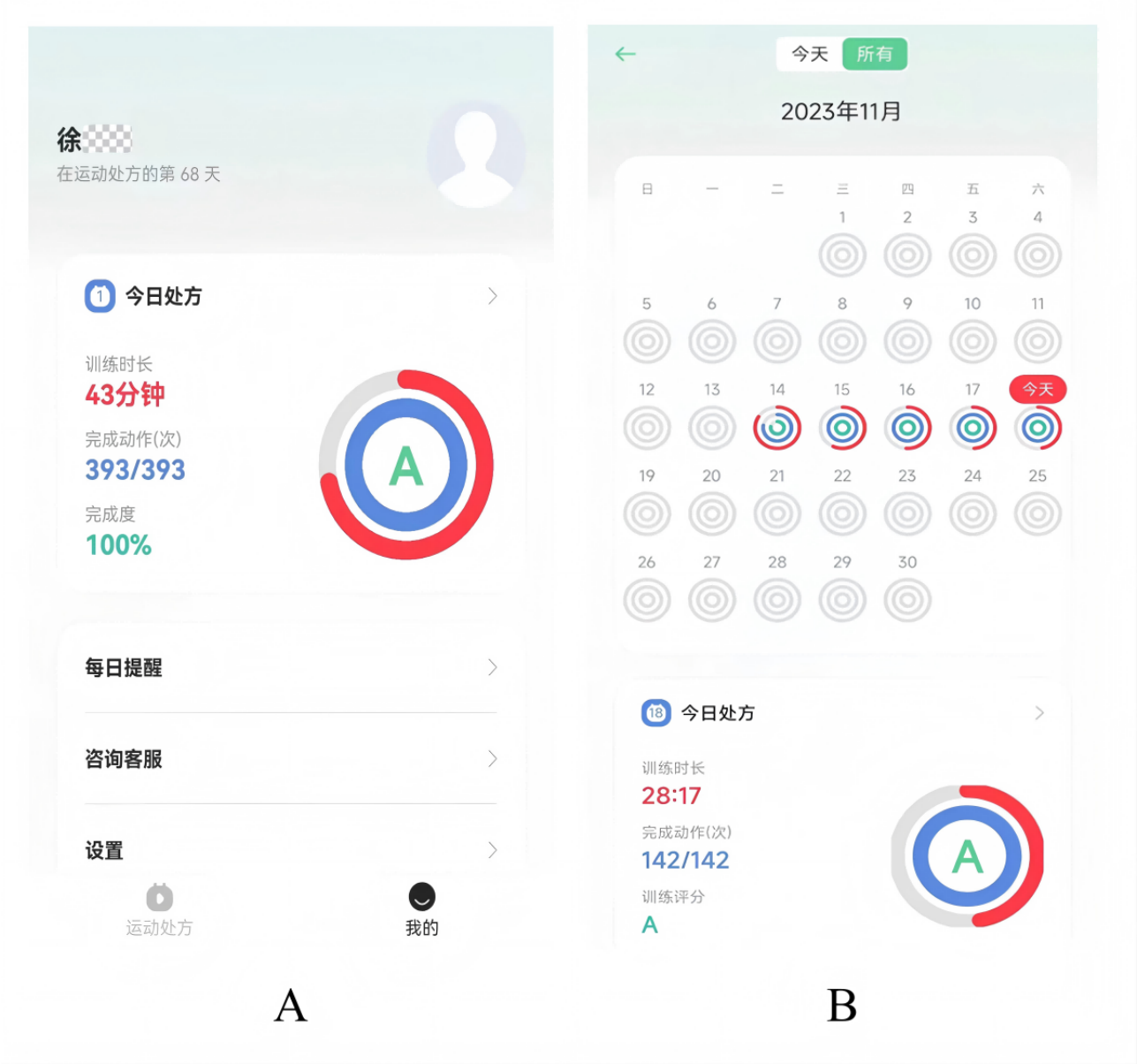

We provided immediate consultation and guidance to participants through online dialogue platforms such as WeChat or telephone follow-up, answered participants' questions during the exercise process, and adjusted the training program to suit each patient's actual situation. In addition to in-program exercise training, the suitability of higher intensity exercise for participants depended on patient feedback, remotely monitored exercise performance data, and the results of weekly online and telephone consultation feedback from the participants. The exercise program could also be adjusted, and specific exercises could be changed for individual participants if the current exercise intensity could not be achieved.

We conducted weekly online consultations or telephone follow-ups to quantify follow-up data and optimize exercise prescriptions. If the participants were unable to adapt to the current intensity, the exercise program was adjusted and replaced with a specific exercise.

Control Group

Participants in the control group received routine verbal health education on hypertension based on the current guidelines. This covered lifestyle guidance for hypertension, including dietary education emphasizing low-salt diets, balanced nutrition, alcohol restriction, adequate hydration, weight management, healthy food choices, psychological adjustment, smoking cessation guidance, sleep management, and advice on reducing disease risk and understanding medication. They also received exercise prescriptions, including warm-up, aerobics, strength, balance, and flexibility training types, intensity, and frequency, from the same movement library as the trial group. The prescription was 30 to 50 minutes per session, and patients were advised to complete it 5 times per week and required to complete it at least 3 times per week. Upon enrollment, the research team distributed the Hypertension Rehabilitation Handbook to all participants. This handbook was developed based on relevant guidelines [29-33] and covers the following topics: the definition and

XSL•FO
RenderX

classification of hypertension, common misconceptions, prevalence statistics, explanations of hazards, risk factor analysis, identification of susceptible populations, methods of early detection and diagnosis, techniques for self-BP measurement, prevention and treatment guidelines, recommendations for lifestyle adjustments, guidance on exercise rehabilitation, and guidance on medication. During the study period, control group participants only received necessary BP measurements and safety assessments at follow-up points, with no active exercise supervision or feedback on execution quality.

Statistical Analyses

All analyses were performed using SPSS statistical software (version 26.0; IBM Corp). Analyses were conducted according to the intention-to-treat principle and according to the intervention group to which participants were originally assigned, regardless of adherence to the intervention. Continuous variables are described as mean (SD) or median (IQR), and categorical variables are described as numbers (percentages). The baseline characteristics between groups were compared using independent *t* tests, Mann-Whitney U tests, or chi-square tests, as appropriate.

Missing data were handled using multiple imputations. The mechanism of missingness was evaluated using the missing completely at random (MCAR) test. The imputation procedure was performed using the fully conditional specification method, and we generated 50 imputed data sets to ensure efficiency and stability. The imputation model included all primary and secondary outcome variables, the treatment group, and auxiliary variables (eg, age and sex).

Between-group differences in primary and secondary outcomes were assessed using analysis of covariance (ANCOVA), with the 8-week measurement as the dependent variable and the baseline value as a covariate. Primary analysis was conducted on each of the 50 imputed data sets. The results were subsequently pooled to generate the final estimates, along with their 95% CIs and *P* values. The significance level was set at $P<.05$ for all statistical tests.

Results

Study Population

A total of 62 participants were included in this study, 31 each in the intervention and control groups. During the follow-up period, 2 people in the intervention group discontinued exercise intervention due to sudden illness, 1 person was unable to exercise for personal reasons, 1 person in the control group was unable to exercise for physical reasons, and 3 people were unable to adhere to exercise for personal reasons. Finally, according to the principle of intention-to-treat analysis, 31 people in the intervention group and 31 people in the control group were included (Figure 4).

Figure 4. Flowchart of the study process.

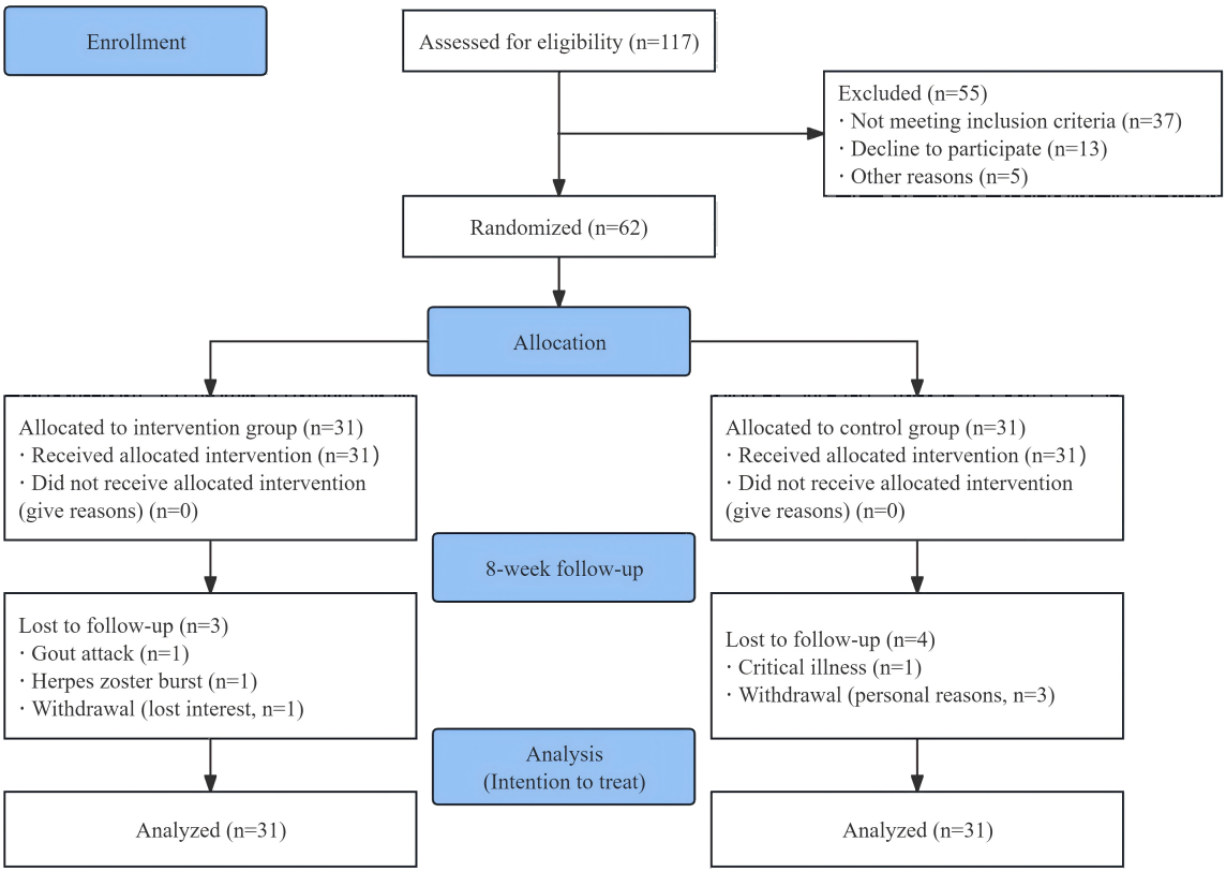

XSL•FO
RenderX

The average age of the participants overall was 52.95 (SD 11.46) years old. The average age of the intervention group was 50.48 (SD 9.44) years, and the average age of the control group was 55.42 (SD 12.86) years; there was no statistically significant difference between the 2 groups. Males accounted for 58% (n=36) and females accounted for 41.9% (n=26) of the participants. The average SBP was 128.87 (SD 12) mmHg, and the average DBP was 85.02 (SD 8.82) mmHg. The mean duration of hypertension was 7.14 (SD 7.5) years. The classification of hypertension was mainly grade 1 hypertension, accounting for 88.7% (n=55) of cases. The proportions of low-intensity, moderate-intensity, and high-intensity physical activity were 25.8% (n=16), 61.3% (n=38), and 12.9% (n=8), respectively. The results showed that the baseline data of the 2 groups were balanced, and there were no statistically significant differences (Table 1).

XSL•FO
RenderX

Table 1. Baseline characteristics of patients with hypertension in the artificial intelligence (AI)-assisted telerehabilitation randomized controlled trial (N=62).

Characteristic	All participants (n=62)	Control group (n=31)	Intervention group (n=31)	*P* value
Age (years), mean (SD)	52.95 (11.46)	55.42 (12.86)	50.48 (9.44)	.09
Gender, n (%)				.61
Male	36 (58)	17 (54.8)	19 (61.3)	
Female	26 (41.9)	14 (45.1)	12 (38.7)	
Height (m), mean (SD)	1.64 (0.07)	1.64 (0.09)	1.64 (0.06)	.77
Weight (kg), mean (SD)	65.42 (13.06)	66.32 (14.54)	64.53 (11.57)	.59
BMI (kg/m^2), mean (SD)	24.03 (3.35)	24.21 (3.37)	23.86 (3.40)	.68
Education, n (%)				.82
Primary and below	4 (6.5)	2 (6.5)	2 (6.5)	
Junior high school	9 (14.5)	3 (9.7)	6 (19.4)	
High school/technical secondary school	13 (21)	8 (25.8)	5 (16.1)	
College/bachelor's degree or above	36 (58.1)	18 (58.1)	18 (58.1)	
Smoking history, n (%)				.72
Yes	9 (14.5)	5 (16.1)	4 (12.9)	
No	53 (85.5)	26 (83.9)	27 (87.1)	
Drinking history, n (%)				.74
Yes	11 (17.7)	5 (16.1)	6 (19.4)	
No	51 (82.3)	26 (83.9)	25 (80.6)	
Operation history, n (%)				.59
Yes	42 (67.7)	22 (71)	20 (64.5)	
No	20 (32.3)	9 (29)	11 (35.5)	
Systolic BP[a] (mm/Hg), mean (SD)	128.87 (12)	129.48 (14.32)	128.26 (9.33)	.69
Diastolic BP (mm/Hg), mean (SD)	85.02 (8.82)	853 (9.60)	85 (8.13)	.99
Amount of medication, mean (SD)	2.73 (1.57)	2.62 (1.66)	2.85 (1.51)	.73
Number of comorbidities, mean (SD)	1.98 (0.96)	1.93 (0.92)	2.04 (1.02)	.74
Duration of hypertension (years), mean (SD)	7.14 (7.50)	6.57 (7.51)	7.80 (7.63)	.52
Grading of hypertension, n (%)				.14
1	55 (88.7)	28 (90.3)	27 (87.1)	
2	5 (8)	1 (3.2)	4 (12.9)	
3	2 (3.2)	2 (6.5)	0 (0)	
IPAQ[b] physical activity level, n (%)				.20
Low	16 (25.8)	7 (22.6)	9 (29)	
Middle	38 (61.3)	22 (71)	16 (51.6)	
High	8 (12.9)	2 (6.5)	6 (19.4)	

[b]BP: blood pressure.

[c]IPAQ: International Physical Activity Questionnaire.

Data Availability and Handling of Missing Data

The extent of missing data at the 8-week follow-up was detailed in Tables S1 and S2 in Multimedia Appendix 2. The proportion of missingness for the primary outcome, 6-minute walk distance (6MWD), was 6.5% (4/62). For secondary outcomes, the missingness ranged from 0% to 14.5%. A dedicated subsample

XSL•FO
RenderX

of 24 participants underwent cardiopulmonary exercise testing (CPET), among whom the missing rate for all CPET indexes was 12.5% (3/24; Table S2 in Multimedia Appendix 2). The MCAR test was performed on the data set for the entire cohort (χ^2_{596} = 557.85; *P*=.87) and separately on the CPET (χ^2_{12}=0; *P*>.99). Both results were statistically nonsignificant, which meant that the data were missing completely at random. Therefore, multiple imputation was deemed appropriate and employed to handle missing values in the primary analyses.

Comparison of Primary and Secondary Outcomes

There were no significant differences in waist circumference, hip circumference, grip strength, 6MWD, anxiety and depression, quality of life, and exercise self-efficacy scores between the 2 groups at baseline, indicating that the baseline levels of the indicators were comparable between the 2 groups before intervention.

The difference results showed that after the intervention, there were statistically significant differences in 6MWD (adjusted mean difference 62.77, 95% CI 26.33-99.22; *P*=.002), SBP reduction (adjusted mean difference 4.11, 95% CI 0.11-8.28, *P*=.046), IPAQ physical activity level (adjusted mean difference 658.96, 95% CI 159.23-1158.69; *P*=.011), exercise self-efficacy score (adjusted mean difference 21.71, 95% CI 13.59 to 29.82; *P*<.001), between the 2 groups (Table 2).

XSL•FO
RenderX

Table 2. Comparison of primary and secondary health outcomes before and after the 8-week intervention in patients with hypertension.

Outcome	Intervention group (n=31)		Control group (n=31)		Between-group difference at 8 weeks	*P* value[a]
	Before (week 0)	After (week 8)	Before (week 0)	After (week 8)	Adjusted mean difference (95% CI)	
6MWD[b] (m), mean (SD)	472.05 (79.58)	528 (69.54)	448.57 (138.53)	449.43 (94.60)	62.77 (26.33-99.22)	*.002* [c]
IPAQ[d] (MET[e]-min/week), median (IQR)	1386 (510-2190)	1983 (1597.25-3037.50)	897 (636-1386)	1464 (1016.7-51924.50)	658.96 (159.23-1158.69)	.01
PHQ-9[f], mean (SD)	6.03 (3.91)	2.72 (3.31)	5.13 (4.49)	2.75 (1.82)	–0.14 (–1.64 to 1.36)	.85
GAD-7[g], mean (SD)	6.23 (4.38)	2.41 (2.20)	4.48 (4.21)	2.25 (1.68)	–0.23 (–1.26 to 0.80)	.65
SF-12[h], mean (SD)						
PCS[i], mean (SD)	41.38 (6.86)	47.51 (6.12)	42.64 (9.06)	47.26 (6.65)	0.66 (–2.64 to 3.96)	.69
MCS[j], mean (SD)	46.28 (10.17)	54.37 (5.20)	46.34 (10.16)	54.18 (5.04)	0.19 (–2.68 to 3.07)	.89
Exercise self-efficacy, mean (SD)	52.96 (30.02)	84.23 (15.84)	60.93 (28.37)	62.85 (12.67)	21.71 (13.59-29.82)	<.001
Weight (kg), mean (SD)	64.53 (11.57)	63.92 (11.04)	66.32 (14.54)	66.66 (14.27)	–1.02 (–1.96 to 0.08)	.053
Girth (cm)	86.68 (9.76)	86.84 (9.80)	88.03 (9.31)	88 (9.40)	0.12 (–1.54 to 1.78)	.88
Hipline (cm)	97 (8.03)	96.48 (7.61)	98.45 (8.16)	97.71 (8.43)	–0.06 (–2.49 to 2.37)	.96
Right grip strength (kg), mean (SD)	32.87 (10.68)	35.89 (11.23)	26.49 (6.25)	27.70 (8.81)	4.39 (–4.67 to 13.44)	.33
Left grip strength (kg), mean (SD)	29.68 (10.16)	34 (11.24)	25.71 (6.27)	26.73 (9.71)	4.85 (–3.68 to 13.38)	.25
Systolic BP[k] (mm/Hg)	128.30 (9.28)	116.63 (7.64)	129.96 (15.21)	120.92 (7.15)	4.11 (0.11-8.28)	.046
Diastolic BP (mm/Hg)	84.52 (8.24)	75.56 (8.79)	85.25 (9.29)	77.71 (9.16)	1.81 (–2.72 to 6.35)	.43

[a]The significance level was set at *P*<.05.

[b]6MWD: 6-minute walk distance.

[c]The italicized values are the *P* value of the primary outcome.

[d]IPAQ: International Physical Activity Questionnaire.

[e]MET: metabolic equivalent of task.

[f]PHQ-9: Patient Health Questionnaire-9.

[g]GAD-7: Generalized Anxiety Disorder-7.

[h]SF-12: Short-Form Health Survey 12.

[i]PCS: Physical Component Summary.

[j]MCS: Mental Component Summary.

[k]BP: blood pressure.

After the intervention, the comparison of cardiopulmonary exercise test indexes between the 2 groups, total exercise time (adjusted mean difference 98.24, 95% CI 49.39-147.08; *P*=.001), peak oxygen uptake (peak VO2) (adjusted mean difference 3.39, 95% CI 0.49-6.29; *P*=.026) and peak oxygen uptake percent predicted (peak VO2%pred) (adjusted mean difference 11.58, 95% CI 2.06-21.10; *P*=.021) were statistically significant, and the exercise time of the intervention group was significantly longer than that of the control group. There was no significant difference in Watt, respiratory exchange rate, anaerobic threshold, anaerobic threshold %pred, HR rest, and HR peak between the two groups (Table 3).

XSL•FO
RenderX

Table 3. Comparison of cardiopulmonary exercise test (CPET) indices at baseline and after the 8-week intervention for both study subsample groups (N=24).

Outcome	Intervention group (n=12)		Control group (n=12)		Between-group difference at 8 weeks	*P* value[a]
	Before week 0), mean (SD)	After (week 8), mean (SD)	Before (week 0), mean (SD)	After (week 8), mean (SD)	Adjusted mean difference (95% CI)	
Exercise time (s)	472.17 (89.47)	564.29 (71.38)	481.67 (110.11)	457.63 (37.37)	98.24 (49.39-147.08)	.001
Maximum load (Watt)	112.17 (35.51)	131.71 (34.63)	97.33 (35.71)	93.35 (45.76)	17.08 (–18.29 to 52.44)	.31
Maximum load %pred	74.50 (11.29)	88.57 (8.42)	78.17 (19.72)	81.75 (12.43)	3.75 (–3.77 to 11.27)	.30
RER[b]	1.20 (0.13)	1.22 (0.09)	1.18 (0.10)	1.22 (0.11)	0.00 (–0.12 to 0.12)	.98
Peak VO_2[c] (ml/min/kg)	21.50 (4.80)	23.80 (2.89)	19.98 (4.44)	19.56 (4.35)	3.39 (0.49-6.29)	.03
Peak VO_2%pred[d]	66.75 (9.78)	82.57 (5.80)	76.25 (15.47)	73.25 (14.96)	11.58 (2.06-21.10)	.02
Anaerobic threshold (ml/min/kg)	13.63 (3.01)	14.13 (2.03)	12.70 (2.48)	12.60 (3.31)	1.54 (–1.08 to 4.16)	.23
Anaerobic thresh-old%pred	42.08 (7.35)	49.86 (7.90)	48.83 (10.24)	46.88 (8.92)	5.05 (–4.29 to 14.38)	.26
HR[e] at rest (bpm)	90.58 (17.90)	89.43 (13.45)	77.67 (15.71)	84.63 (17.35)	0.69 (–14.43 to 15.80)	.92
HR at peak (bpm)	155.50 (22.18)	154.29 (21.01)	134.42 (27.54)	138.13 (27.61)	–0.98 (–16.04 to 14.09)	.89

[a]The significance level was set at $P<.05$.

[b]RER: respiratory exchange rate.

[c]Peak VO_2: peak oxygen uptake.

[d]Peak VO_2%pred: peak oxygen uptake percent predicted.

[e]HR: heart rate.

The CPET data in this study were obtained from only a subset of participants. As shown in Table S3 in Multimedia Appendix 2, no statistically significant differences were observed between the 2 groups across all key baseline characteristics. This indicated that although only a subset of participants completed the CPET, they constituted a representative random sample from the overall population rather than a group with systematic bias. Throughout the 8-week intervention period, no serious adverse events related to the AI-assisted telerehabilitation intervention were reported.

Discussion

Principal Findings

The therapeutic effects of AI-assisted remote rehabilitation training on patients with hypertension were examined in this study. This study demonstrated that AI-assisted training based on 3D skeletal pose recognition significantly improved exercise capacity and BP and increased exercise self-efficacy in patients with hypertension, superior to usual care.

The Effect of AI-Assisted Remote Exercise Rehabilitation on Participants With Hypertension

Exercise rehabilitation for patients with hypertension is of great significance. It can not only effectively reduce BP level but also significantly improve cardiovascular function, thereby improving the quality of life for patients. However, there are some limitations in the current rehabilitation treatment of hypertension, such as the fact that patients frequently lack objective supervision and quality assurance during exercise implementation. This indicates that factors such as the specific execution methods of prescribed regimens, dosage standards, and the risk of muscle injury from nonstandardized exercises in home environments cannot be effectively safeguarded. This phenomenon not only significantly diminishes the effectiveness of interventions but also exacerbates issues related to individual variability. AI-based intelligent exercise rehabilitation is expected to solve these challenges.

In this study, the intervention group performed significantly better than the control group in terms of 6MWD, total exercise time, and peak VO_2 for cardiopulmonary exercise testing, indicating that AI-based remote rehabilitation has a significant effect on improving exercise endurance and cardiopulmonary function. Peak VO_2 is the core index of cardiopulmonary exercise testing, which reflects the maximum oxygen uptake capacity of patients during exercise and is an important standard to evaluate cardiopulmonary function. Although the intervention group showed significant improvement in 6MWD, the increase in peak VO_2 , although statistically significant, did not reach the minimal clinically important difference, suggesting the need for longer intervention duration or a more precise exercise prescription. For objective equipment reasons, cardiopulmonary exercise testing was performed in only 24 of 62 participants before and after the intervention. CPET data were obtained from only a subset of participants. Although baseline comparisons revealed no selection bias, future research should validate these findings across the entire sample. Nevertheless, some remarkable results were obtained. This study was similar to previous studies

XSL•FO
RenderX

on telerehabilitation in several ways. McDonagh et al [34,35] found that telecardiac rehabilitation and traditional rehabilitation had the same effect in restoring motor function, improving exercise capacity, and increasing physical activity; moreover, patients in the telecardiac rehabilitation group expressed higher satisfaction levels. There is some evidence to support greater adherence to telerehabilitation programs, which is particularly important in the context of health care crises (eg, due to pandemics) and in patients living in hard-to-reach, remote, and low- and middle-income areas. That said, the broader applicability of telerehabilitation in chronic disease management and remote health care still faces notable barriers—most prominently, difficulties in smartphone use among older adults [36]. As a key demographic for conditions like hypertension or heart disease, many older adult patients struggle with smartphone operation (eg, navigating apps, interpreting data feedback) due to limited digital literacy, sensory impairments, or tech unfamiliarity, which hinders engagement and risks widening health disparities. Therefore, future research should focus on assessing digital literacy's influence on telerehabilitation response and testing tailored strategies to mitigate these barriers, ensuring remote care benefits diverse older adult populations.

Lowering BP is a core objective in hypertension clinical research. Exercise therapy, a key adjunct to pharmacotherapy, is widely proven effective for BP reduction, with the specific magnitude dependent on the exercise program design [37-40]. In this study, the remote rehabilitation (intervention) group showed a significant SBP decrease (mean reduction >12 mmHg), while no significant change was observed in DBP, a finding that is consistent with previous studies and guidelines. The ACSM and AHA recommend moderate-intensity aerobic exercise for patients with hypertension, and existing research confirms that aerobic exercise and personalized prescriptions exert a more prominent regulatory effect on SBP [41]. The lack of significant DBP reduction may relate to patients' baseline BP, medication use, or exercise intensity. Physiologically, the significant SBP decrease can be attributed to regular exercise—particularly the "endurance and resistance" training used here—which regulates BP by improving endothelial function, reducing vascular stiffness, and mitigating oxidative stress. This aligns with findings by Guirado [42] and Kokkinos et al [43], who noted greater BP reductions with combined training. Additionally, no baseline intergroup BP differences existed, but postintervention SBP differed significantly, further validating the value of remote rehabilitation, as supported by prior research [44]. Clinically, it is meaningful that each 3 mmHg SBP reduction lowers coronary insufficiency risk by 5% to 9%, stroke risk by 8% to 14%, and all-cause mortality by 4% [45]. Notably, the impact of exercise interventions on SBP in populations with hypertension remains insufficiently understood. While exercise-induced SBP reduction is moderate but consistent, it is less pronounced than that from pharmacotherapy—although comparable to the effects of common antihypertensive drugs [37]. Future research should further explore the SBP-regulating mechanisms of exercise and assess the generalizability of such findings in real-world clinical settings.

Significant improvements in exercise self-efficacy and intervention adherence were observed with the AI-assisted remote rehabilitation program among patients with hypertension. Exercise self-efficacy is defined as an individual's confidence in their ability to successfully perform a specific physical activity. Enhanced exercise self-efficacy was achieved through multiple mechanisms. First, real-time monitoring of patient exercise data by the AI system enabled the generation of personalized feedback, allowing individuals to better understand their physiological responses and form realistic expectations about exercise outcomes. Second, a comprehensive collection of exercise demonstration videos and animated guides provided clear visual support, helping patients learn proper techniques and postures, which, in turn, boosted both skill proficiency and confidence. Third, exercise programs were dynamically adjusted using intelligent algorithms that modified intensity and difficulty based on individual health status and performance data, ensuring the intervention remained aligned with each patient's needs and enhancing overall engagement and experience. Compared with traditional face-to-face rehabilitation programs, a remote approach demonstrates higher adherence, as patients are not required to travel to hospitals or rehabilitation centers, significantly reducing time and transportation burdens and thereby increasing participation willingness. The AI system actively prompts patients to complete daily exercise tasks, establishing a closed-loop management process of "monitoring-feedback-prompting," which supports long-term engagement in physical activity and facilitates habit formation.

To our knowledge, this is the first randomized controlled trial to implement and evaluate an AI-assisted telerehabilitation system for patients with hypertension that utilizes real-time, AI-guided training based on 3D skeletal pose recognition. Our findings demonstrate that this technology can simultaneously and significantly improve objective exercise capacity, enhance psychological determinants of behavior such as self-efficacy, and aid in SBP control. This synergistic effect provides a new level of evidence, moving beyond remote monitoring to demonstrate the efficacy of interactive, form-correcting, and personalized exercise prescription. These results underscore the transformative potential of intelligent exercise rehabilitation for optimizing chronic disease management. Consequently, this study provides a foundation for developing individualized home-based rehabilitation programs, promoting sustained health behaviors, and advancing mobile health management models for hypertension and potentially other chronic conditions.

Limitations

Despite the positive results of this study, there are still some limitations. First, the study sample size was relatively small and focused on the Guangdong region in China, which may limit the generalizability of the results. Future studies could expand the sample size and validate it in different regions and populations to further confirm the effectiveness of the AI-assisted telerehabilitation program. Second, the relatively short 8-week intervention period may not capture the long-term effects of AI-driven telerehabilitation on exercise capacity and cardiovascular health. Moreover, the relatively short follow-up period in this study was limited to short-term postintervention outcome assessment. Future studies could extend the follow-up

XSL•FO
RenderX

period to assess the long-term effects and sustainability of the program. Finally, the lack of objective monitoring tools for participants' exercise intensity, such as using wearable devices to monitor heart rate to ensure training intensity and exercise safety, will be prioritized in future studies.

Conclusions

This study systematically evaluated the effects of AI-based telerehabilitation on exercise ability and related health indicators of patients with hypertension through a randomized controlled trial. The results showed that the intervention had certain effects in reducing SBP and improving exercise ability and exercise self-efficacy. Moreover, the patients' compliance was high, and the intervention process was safe and feasible. As a low-cost and accessible rehabilitation training method, the AI telerehabilitation model is expected to become a convenient and effective alternative in the future.

Acknowledgments

We would like to thank RaphAI Health Technologies Limited (Guangzhou, China) and Beijing Kangtang Medical Technology Co Ltd (Beijing, China) for their technical support. This study was supported by the National Natural Science Foundation of China (82472588), the Guangdong Natural Science Foundation of China (2025A1515012331 and 2025A1515012782), and the 2025 Guangdong Disability Cause Theory and Practice Research Project Grant. No generative AI was used in any part of the manuscript.

Data Availability

The raw data supporting the findings of this study will be made available by the corresponding author, without undue reservation.

Authors' Contributions

QZ, GH, and HM designed the study. QY, BQ, and LH drafted the manuscript. QY, LH, SW, and CL recruited participants. QW was responsible for the random assignment of participants. GL and JH were responsible for the intervention training. JZ collected the data, and DL analyzed the data. All authors agreed to the final version of the manuscript. QY, BQ, and LH contributed equally and share first authorship. HM, GH, and QZ are corresponding authors and contributed equally to this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 999 KB - jmir_v28i1e81400_app1.pdf]

Multimedia Appendix 2
Supplementary Tables.
[DOCX File , 25 KB - jmir_v28i1e81400_app2.docx]

References

1. Kearney P, Whelton M, Reynolds K, Muntner P, Whelton P, He J. Global burden of hypertension: analysis of worldwide data. Lancet 2005 Jan;365(9455):217-223 [FREE Full text] [doi: 10.1016/s0140-6736(05)17741-1]
2. Forouzanfar M, Liu P, Roth G, Ng M, Biryukov S, Marczak L, et al. Global Burden of hypertension and systolic blood pressure of at least 110 to 115 mm hg, 1990-2015. JAMA 2017 Jan 10;317(2):165-182 [FREE Full text] [doi: 10.1001/jama.2016.19043] [Medline: 28097354]
3. Kiviniemi A, Lepojärvi E, Tulppo P, Piira OP, Kenttä TV, Perkiömäki JS, et al. Prediabetes and risk for cardiac death among patients with coronary artery disease: the ARTEMIS study. Diabetes Care 2019 Jul;42(7):1319-1325 [FREE Full text] [doi: 10.2337/dc18-2549] [Medline: 31076416]
4. Lee S, Hwang S, Kang D, Yang H. Brain education-based meditation for patients with hypertension and/or type 2 diabetes: A pilot randomized controlled trial. Medicine (Baltimore) 2019 May;98(19):e15574 [FREE Full text] [doi: 10.1097/MD.0000000000015574] [Medline: 31083232]
5. Ettehad D, Emdin C, Kiran A, Anderson S, Callender T, Emberson J, et al. Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis. Lancet 2016 Mar;387(10022):957-967 [FREE Full text] [doi: 10.1016/s0140-6736(15)01225-8]
6. Burnier M. Drug adherence in hypertension. Pharmacol Res 2017 Nov;125(Pt B):142-149 [FREE Full text] [doi: 10.1016/j.phrs.2017.08.015] [Medline: 28870498]

XSL•FO
RenderX

7. Börjesson M, Onerup A, Lundqvist S, Dahlöf B. Physical activity and exercise lower blood pressure in individuals with hypertension: narrative review of 27 RCTs. Br J Sports Med 2016 Mar;50(6):356-361 [FREE Full text] [doi: 10.1136/bjsports-2015-095786] [Medline: 26787705]
8. Noone C, Leahy J, Morrissey E, Newell J, Newell M, Dwyer CP, et al. Comparative efficacy of exercise and anti-hypertensive pharmacological interventions in reducing blood pressure in people with hypertension: A network meta-analysis. Eur J Prev Cardiol 2020 Feb;27(3):247-255 [FREE Full text] [doi: 10.1177/2047487319879786] [Medline: 31615283]
9. Diaz K, Booth J, Seals S, Abdalla M, Dubbert PM, Sims M, et al. Physical activity and incident hypertension in African Americans. Hypertension 2017 Mar;69(3):421-427 [FREE Full text] [doi: 10.1161/hypertensionaha.116.08398]
10. Zhou K, Wang W, An J, Li M, Li J, Li X. Effects of progressive upper limb exercises and muscle relaxation training on upper limb function and health-related quality of life following surgery in women with breast cancer: a clinical randomized controlled trial. Ann Surg Oncol 2019 Jul;26(7):2156-2165 [FREE Full text] [doi: 10.1245/s10434-019-07305-y] [Medline: 30972655]
11. Churilla J, Ford E. Comparing physical activity patterns of hypertensive and nonhypertensive US adults. Am J Hypertens 2010 Sep;23(9):987-993 [FREE Full text] [doi: 10.1038/ajh.2010.88] [Medline: 20431526]
12. Williams N, Hendry M, France B, Lewis R, Wilkinson C. Effectiveness of exercise-referral schemes to promote physical activity in adults: systematic review. Br J Gen Pract 2007 Dec;57(545):979-986 [FREE Full text] [doi: 10.3399/096016407782604866] [Medline: 18252074]
13. Wang A, Zhang H. Efficacy of microsurgery for patients with cerebral hemorrhage secondary to gestational hypertension: a systematic review protocol of randomized controlled trial. Medicine 2019 [FREE Full text] [doi: 10.1097/md.0000000000017558]
14. Charidimou A, Turc G, Oppenheim C, Yan S, Scheitz JF, Erdur H, et al. Microbleeds, cerebral hemorrhage, and functional outcome after stroke thrombolysis. Stroke 2017 Aug;48(8):2084-2090. [doi: 10.1161/STROKEAHA.116.012992] [Medline: 28720659]
15. Ruilope L, Valenzuela P, Lucia A. Digital therapeutics and lifestyle: the start of a new era in the management of arterial hypertension? Eur Heart J 2021 Oct 21;42(40):4123-4125 [FREE Full text] [doi: 10.1093/eurheartj/ehab694] [Medline: 34571529]
16. Kario K. Digital hypertension towards to the anticipation medicine. Hypertens Res 2023 Nov;46(11):2503-2512 [FREE Full text] [doi: 10.1038/s41440-023-01409-5] [Medline: 37612370]
17. Nomura A, Tanigawa T, Kario K, Igarashi A. Cost-effectiveness of digital therapeutics for essential hypertension. Hypertens Res 2022 Oct;45(10):1538-1548 [FREE Full text] [doi: 10.1038/s41440-022-00952-x] [Medline: 35726085]
18. Kario K, Nomura A, Harada N, Okura A, Nakagawa K, Tanigawa T, et al. Efficacy of a digital therapeutics system in the management of essential hypertension: the HERB-DH1 pivotal trial. Eur Heart J 2021 Oct 21;42(40):4111-4122 [FREE Full text] [doi: 10.1093/eurheartj/ehab559] [Medline: 34455443]
19. Alessa T, Hawley M, Hock E, de Witte L. Smartphone apps to support self-management of hypertension: review and content analysis. JMIR Mhealth Uhealth 2019 May 28;7(5):e13645 [FREE Full text] [doi: 10.2196/13645] [Medline: 31140434]
20. McLean G, Band R, Saunderson K, Hanlon P, Murray E, Little P, et al. Digital interventions to promote self-management in adults with hypertension systematic review and meta-analysis. J Hypertens 2016:600-612 [FREE Full text] [doi: 10.1097/hjh.0000000000000859]
21. Maskeliūnas R, Kulikajevas A, Damaševičius R, Griškevičius J, Adomavičienė A. Biomac3D: 2D-to-3D human pose analysis model for tele-rehabilitation based on pareto optimized deep-learning architecture. Applied Sciences 2023 Jan 13;13(2):1116 [FREE Full text] [doi: 10.3390/app13021116]
22. Yang G, He S, Meng D, Wei M, Cao J, Guo H, et al. Body landmarks and genetic algorithm-based approach for non-contact detection of head forward posture among Chinese adolescents: revitalizing machine learning in medicine. BMC Med Inform Decis Mak 2023 Sep 11;23(1):179 [FREE Full text] [doi: 10.1186/s12911-023-02285-2] [Medline: 37697312]
23. Wei M, Meng D, He S, Lv Z, Guo H, Yang G, et al. Investigating the efficacy of AI-enhanced telerehabilitation in sarcopenic older individuals. Eur Geriatr Med 2025 Feb;16(1):115-123 [FREE Full text] [doi: 10.1007/s41999-024-01082-y] [Medline: 39453567]
24. Xiao C, Zhao Y, Li G, Zhang Z, Liu S, Fan W, et al. Clinical efficacy of multimodal exercise telerehabilitation based on AI for chronic nonspecific low back pain: randomized controlled trial. JMIR Mhealth Uhealth 2025 May 22;13:e56176 [FREE Full text] [doi: 10.2196/56176] [Medline: 40402551]
25. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010 Mar 23;340(mar23 1):c869-c869 [FREE Full text] [doi: 10.1136/bmj.c869] [Medline: 20332511]
26. Chan A, Tetzlaff J, Altman D, Laupacis A, Gøtzsche PC, Krleža-Jerić K, et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. Ann Intern Med 2013 Feb 05;158(3):200-207 [FREE Full text] [doi: 10.7326/0003-4819-158-3-201302050-00583] [Medline: 23295957]
27. Piotrowicz E, Zieliński T, Bodalski R, Rywik T, Dobraszkiewicz-Wasilewska B, Sobieszczańska-Małek M, et al. Home-based telemonitored Nordic walking training is well accepted, safe, effective and has high adherence among heart failure patients,

XSL•FO
RenderX

including those with cardiovascular implantable electronic devices: a randomised controlled study. Eur J Prev Cardiol 2015 Nov;22(11):1368-1377 [FREE Full text] [doi: 10.1177/2047487314551537] [Medline: 25261268]
28. Butāne L, Spilva-Ekerte L, Šablinskis M, Skride A, Šmite D. Individually tailored home-based physiotherapy program makes sustainable improvement in exercise capacity and daily physical activity in patients with pulmonary arterial hypertension. Ther Adv Respir Dis 2022;16:17534666221132477 [FREE Full text] [doi: 10.1177/17534666221132477] [Medline: 36314474]
29. World Health Organization. Global Recommendations on Physical Activity for Health. Geneva: World Health Organization; 2010.
30. Thompson P, Arena R, Riebe D, Pescatello L. ACSM's new preparticipation health screening recommendations from ACSM's guidelines for exercise testing and prescription, ninth edition. Curr Sports Med Rep 2013;12(4):215-217 [FREE Full text] [doi: 10.1249/jsr.0b013e31829a68cf]
31. Thompson P, Buchner D, Pin˜a I, Balady G, Williams M, Marcus B, et al. Exercise and physical activity in the prevention and treatment of atherosclerotic cardiovascular disease. Circulation 2003 Jun 24;107(24):3109-3116 [FREE Full text] [doi: 10.1161/01.cir.0000075572.40158.77]
32. Balady G, Williams M, Ades P, Bittner V, Comoss P, Foody J, et al. Core components of cardiac rehabilitation/secondary prevention programs: 2007 update. Circulation 2007 May 22;115(20):2675-2682 [FREE Full text] [doi: 10.1161/circulationaha.106.180945]
33. Bairey Merz CN, Alberts M, Balady G, Ballantyne CM, Berra K, Black HR, et al. ACCF/AHA/ACP 2009 Competence and Training statement: a curriculum on prevention of cardiovascular disease. Circulation 2009 Sep 29;120(13):A [FREE Full text] [doi: 10.1161/circulationaha.109.192640]
34. McDonagh S, Dalal H, Moore S, Clark C, Dean S, Jolly K, et al. Home-based versus centre-based cardiac rehabilitation. Cochrane Db Syst Rev 2023;10(10) [FREE Full text] [doi: 10.1002/14651858.cd007130.pub5]
35. McDonagh S, Dalal H, Moore S, Clark C, Taylor R. Cochrane corner: centre versus telemedicine approaches to cardiac rehabilitation. Heart 2023 Dec 15;110(1):7-10 [FREE Full text] [doi: 10.1136/heartjnl-2023-322640] [Medline: 37487697]
36. Shibata S, Hoshide S. Current situation of telemedicine research for cardiovascular risk in Japan. Hypertens Res 2023 May;46(5):1171-1180 [FREE Full text] [doi: 10.1038/s41440-023-01224-y] [Medline: 36849580]
37. Naci H, Salcher-Konrad M, Dias S, Blum M, Sahoo S, Nunan D, et al. How does exercise treatment compare with antihypertensive medications? A network meta-analysis of 391 randomised controlled trials assessing exercise and medication effects on systolic blood pressure. Br J Sports Med 2019 Jul;53(14):859-869 [FREE Full text] [doi: 10.1136/bjsports-2018-099921] [Medline: 30563873]
38. Oliveira-Dantas F, Brasileiro-Santos M, Thomas S, Silva A, Silva D, Browne R, et al. Short-term resistance training improves cardiac autonomic modulationblood pressure in hypertensive older women: a randomized controlled trial. J Strength Cond Res 2020 [FREE Full text] [doi: 10.1519/jsc.0000000000003182]
39. Ruangthai R, Phoemsapthawee J. Combined exercise training improves blood pressure and antioxidant capacity in elderly individuals with hypertension. J Exerc Sci Fit 2019 Jan 20;17(2) [FREE Full text] [doi: 10.1016/j.jesf.2019.03.001] [Medline: 30949214]
40. Saco‐Ledo G, Valenzuela P, Ruiz‐Hurtado G, Ruilope L, Lucia A. Exercise reduces ambulatory blood pressure in patients with hypertension: a systematic review and meta‐analysis of randomized controlled trials. JAHA 2020 Dec 15;9(24):A [FREE Full text] [doi: 10.1161/jaha.120.018487]
41. Herrod P, Lund J, Phillips B. Time-efficient physical activity interventions to reduce blood pressure in older adults: a randomised controlled trial. Age Ageing 2021 May 05;50(3):980-984 [FREE Full text] [doi: 10.1093/ageing/afaa211] [Medline: 33068100]
42. Guirado GN, Damatto RL, Matsubara BB, Roscani MG, Fusco DR, Cicchetto LAF, et al. Combined exercise training in asymptomatic elderly with controlled hypertension: effects on functional capacity and cardiac diastolic function. Med Sci Monit 2012 Jul;18(7):CR461-CR465 [FREE Full text] [doi: 10.12659/msm.883215] [Medline: 22739737]
43. Kokkinos P, Narayan P, Colleran J, Pittaras A, Notargiacomo A, Reda D, et al. Effects of regular exercise on blood pressure and left ventricular hypertrophy in African American men with severe hypertension. N Engl J Med 1995 Nov 30;333(22):1462-1467 [FREE Full text] [doi: 10.1056/nejm199511303332204]
44. Rodrigues-da-Silva A, Suassuna J, Monteiro E, Borges de Lima I, Santos A, Brasileiro-Santos M. Effects of telerehabilitation on cardiac remodeling and hemodynamics parameters in hypertensive older adults: A randomized controlled trial. J Telemed Telecare 2024 Mar 14;31(7):1005-1013 [FREE Full text] [doi: 10.1177/1357633x241236572]
45. Pescatello L, Franklin B, Fagard R, Farquhar W, Kelley G, Ray C, American College of Sports Medicine. American College of Sports Medicine position stand. Exercise and hypertension. Med Sci Sports Exerc 2004 Mar;36(3):533-553 [FREE Full text] [doi: 10.1249/01.mss.0000115224.88514.3a] [Medline: 15076798]

Abbreviations

6MWD: 6-minute walk distance
6MWT: 6-minute walk test

XSL•FO
RenderX

ACSM: American College of Sports Medicine
AHA: American Heart Association
ANCOVA: analysis of covariance
BP: blood pressure
CONSORT: Consolidated Standards of Reporting Trials
CPET: cardiopulmonary exercise testing
DBP: diastolic blood pressure
HERB-DH1: HERB Digital Hypertension 1
MCAR: missing completely at random
PAR-Q: Physical Activity Readiness Questionnaire
peak VO2%pred: peak oxygen uptake percent predicted
peak VO2: peak oxygen uptake
SBP: systolic blood pressure
SPIRIT: Standard Protocol Items: Recommendations for Interventional Trials
WHO: World Health Organization

Edited by S Brini; submitted 28.Jul.2025; peer-reviewed by L Zhang, A Higaki; comments to author 21.Sep.2025; accepted 11.Nov.2025; published 13.Jan.2026.

Please cite as:
Yao Q, Qiu B, He L, Wang Q, Zou J, Liang D, Wen S, Liu Y, Li G, Hu J, Ma H, Huang G, Zeng Q
Effects of Artificial Intelligence Recognition–Based Telerehabilitation on Exercise Capacity in Patients With Hypertension: Randomized Controlled Trial
J Med Internet Res 2026;28:e81400
URL: https://www.jmir.org/2026/1/e81400
doi:10.2196/81400
PMID:

©Qiuru Yao, Baizhi Qiu, Longlong He, Qin Wang, Jihua Zou, Donghui Liang, Shuyang Wen, Yingchao Liu, Gege Li, Jinjing Hu, Huan Ma, Guozhi Huang, Qing Zeng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Engagement Significantly Enhances Weight Loss Outcomes in Adults With Obesity Treated With Tirzepatide: Retrospective Cohort Study of a Digital Weight Loss Service

Hans Johnson[1,2,3], MSc, MRes, MBChB; Ashley Kieran Clift[2,4], MA, MBBS, DPhil; Daniel Reisel[2,5], MBBS, DPhil, MRCOG; David Huang[2], BSc, MBBS

[1]Engineering and Physical Sciences Research Council Centre for Doctorate Training in Digital Health and Care, University of Bristol, Bristol, United Kingdom

[2]Department of Clinical Innovation and Research, Voy (T/a Menwell LTD), London, United Kingdom

[3]Department of Education, University of Oxford, Oxford, United Kingdom

[4]Department of Surgery and Cancer, Imperial College London, London, United Kingdom

[5]Elizabeth Garrett Anderson Institute for Women's Health, University College London, London, United Kingdom

Corresponding Author:
Hans Johnson, MSc, MRes, MBChB
Engineering and Physical Sciences Research Council Centre for Doctorate Training in Digital Health and Care
University of Bristol
1 Cathedral Square, Trinity Street
Bristol, BS1 5DD
United Kingdom
Phone: 44 7916948134
Fax: 44 117 42 82343
Email: hj14789@bristol.ac.uk

Abstract

Background: The advent of tirzepatide has transformed obesity care; yet, real-world weight loss outcomes necessarily depend on patient engagement with behavioral support. Digital platforms offering coaching, self-monitoring, and automated feedback have the potential to further augment pharmacological efficacy.

Objective: The aim of the study is to examine associations between digital engagement and weight loss outcomes among adults prescribed tirzepatide in routine care over 12 months and to identify baseline correlates of engagement.

Methods: In this retrospective cohort study, we included adults (18-75 years; BMI ≥30 or ≥27.5 kg/m^2 with comorbidities) who initiated tirzepatide between February 2024 and August 2025 via a UK digital weight loss service. Engagement was defined by all 3: attendance at ≥1 coaching session AND ≥1 weekly weight log AND ≥1 app login over 12 months. Percent weight loss was analyzed at months 2, 4, 6, 8, 10, and 12 using a mixed model repeated measures adjusted for age, sex, baseline BMI, and comorbidities. Time-to-event analyses (Kaplan-Meier) assessed attainment of ≥5%, ≥10%, ≥15%, and ≥20% weight loss thresholds. Multivariable logistic regression identified correlates of engagement, reporting odds ratios (ORs) per decade of age and per 5 kg/m^2 BMI.

Results: Among 126,553 participants, 6746 (5.3%) were maximally engaged. Cohort demographics were a mean age of 42.3 (SD 12.4) years, 78.9% (99,905/126,553) female, and a mean BMI of 35.3 (SD 6.2) kg/m^2. Engaged users achieved greater adjusted weight loss at month 12 (–22.9%, 95% CI –23.2 to –22.6) versus nonengaged users (–17.5%, 95% CI –17.7 to –17.4), an absolute difference of 5.3 percentage points (P<.001; Cohen d=0.54). Differences emerged by month 2 (–7.4% vs –6.4%; P<.001) and widened steadily. Engaged participants reached all clinically significant weight loss thresholds faster (5%-20%; log-rank P<.001), and engaged participants were nearly 3 times more likely to achieve ≥20% weight loss compared to nonengaged participants (1079/6746, 16% vs 6710/119,807, 5.6%; risk ratio 2.88; P<.001). Older age (OR 1.18 per decade, 95% CI 1.15-1.20; P<.001), higher BMI (OR 1.14 per 5 kg/m^2, 95% CI 1.12-1.16; P<.001), and the presence of polycystic ovary syndrome (OR 1.59, 95% CI 1.45-1.74; P<.001) or fatty liver disease (OR 1.52, 95% CI 1.32-1.76; P<.001) correlated with engagement. Male sex (OR 0.86, 95% CI 0.81-0.92; P<.001) and diabetes (OR 0.83, 95% CI 0.73-0.95; P=.009) were associated with lower engagement.

XSL•FO
RenderX

Conclusions: Digital engagement was associated with substantially greater tirzepatide-associated weight loss in real-world practice. Integrating structured digital support with pharmacotherapy represents a promising strategy for optimizing obesity management.

(J Med Internet Res 2026;28:e83718) doi:10.2196/83718

KEYWORDS

obesity; weight loss; tirzepatide; digital health; engagement; behavior; coaching; retrospective study; GIP/GLP-1 RA; glucose-dependent insulinotropic polypeptide and glucagon-like peptide-1 receptor agonist

Introduction

Background

The global prevalence of obesity has tripled since 1975, and the economic burden of obesity and overweight is projected to reach 3.3% of global gross domestic product by 2060 [1]. Beyond economic considerations, obesity substantially increases risks for numerous chronic conditions including type 2 diabetes, cardiovascular disease, certain cancers, as well as overall mortality. Despite recognition of obesity as a complex chronic disease requiring multifaceted interventions, historically available treatments have shown limited long-term efficacy [2].

Recent advancements in pharmacotherapy have transformed the obesity treatment paradigm. Tirzepatide (Mounjaro) is a novel dual glucose-dependent insulinotropic polypeptide (GIP) and glucagon-like peptide-1 (GLP-1) receptor agonist (RA) that has demonstrated remarkable efficacy for weight management in randomized controlled trials (RCTs) [3]. The SURMOUNT-1 trial reported mean weight reductions of 15% to 20.9% at 72 weeks, significantly surpassing previous pharmacotherapies [4]. These results have generated substantial clinical interest; yet, questions remain regarding their generalization to real-world settings and the scope for adjuncts to support or optimize outcomes.

Digital weight loss services (DWLSs) combine pharmacotherapy with app-based tools, regular weight tracking, and professional health education and tailor nutrition-driven coaching. These services may address critical barriers to obesity treatment, including limited health care provider time, accessibility challenges, and inadequate behavioral support [5-7]. Understanding how digital engagement impacts pharmacotherapy outcomes could critically inform clinical recommendations and health service design. Furthermore, digital health interventions represent a promising complement to pharmacotherapy. Recent systematic reviews indicate that technology-based approaches can enhance weight loss outcomes by improving self-monitoring, accountability, personalized feedback, and remote support [8-10]. Studies of remotely delivered weight loss services have shown the efficacy of such digital intervention [11]. However, apart from a few recent studies [5], there is limited evidence regarding how digital engagement influences outcomes specifically in patients using dual GLP-1 and GIP RA for weight management.

Objectives

This study conducted a retrospective evaluation of the Voy DWLS, which is available throughout the United Kingdom. Our objectives were to characterize the real-world effectiveness of tirzepatide on weight loss, explore the associations between digital engagement and outcomes, and identify factors that correlate with engagement.

Methods

Study Design and Setting

This retrospective study used an open cohort approach. The study period spanned early February 2024 to early August 2025, with patients contributing follow-up from the date of first prescription to the earliest date of last weight measurement logged or reached 12 months of follow-up. The Voy digital health platform, a commercial telehealth DWLS for obesity management, was developed by a multidisciplinary team of clinicians, behavioral scientists, and software developers to provide remote behavioral support through live group video coaching sessions, text-based in-app support, dynamic educational content, and the direct supply of tirzepatide for weight management. Drawing upon established frameworks in behavior change and self-management support, the platform combined core digital tools (eg, real-time weight monitoring, medication adherence tracking, and personalized coaching sessions) into a single interface accessible via smartphone or web browser.

Procedure

Participants became aware of the Voy program through multiple channels: targeted social media campaigns, clinician referrals, word-of-mouth recommendations, and general web searches. They self-enrolled and self-paid via the Voy website, where they completed an online screening questionnaire covering medical history, BMI, and lifestyle factors. Thereafter, the participants interacted with the Voy digital health platform's weight management program, which integrates dual GLP-1 and GIP RA pharmacotherapies with digital behavioral support to enhance weight loss outcomes. Subsequently, participants self-enrolled via the Voy website, where they completed an online screening questionnaire covering medical history, BMI, and a range of lifestyle factors. Upon enrollment, participants underwent an initial assessment to confirm eligibility, including verification of age, BMI, and the absence of exclusion criteria (Figure 1). All participants underwent comprehensive clinical screening before treatment initiation. Structured safeguards ensured accuracy and safe prescribing: photo identification and full-body photographs for identity and BMI verification; detailed clinical questionnaires capturing medical history, contraindications, and current medications; and individual review by qualified licensed clinical prescribers with cross-checking for clinical red flags. The service operates under

XSL•FO
RenderX

Care Quality Commission registration with internal clinical audit processes, escalation protocols for safety concerns, and regular safety reviews. This approach adheres to General Medical Council and General Pharmaceutical Council standards for remote prescribing in the United Kingdom, with clinician-led prescribing decisions and independent verification of eligibility criteria to ensure patient safety.

The initial monthly cost of enrollment for tirzepatide was US $273.72. Participants receive tirzepatide (Mounjaro) via multidose prefilled KwikPen containing 4 weekly doses. Administration is a subcutaneous injection once weekly, with doses ranging from 2.5 to 15 mg. The starting dose is 2.5 mg, increased every 4 weeks in 2.5 mg increments as tolerability permits. New tirzepatide pens are prescribed every 28 days. Monthly costs cover medication, clinical support from qualified prescribing clinicians, personalized coaching from registered dietitians and nutritionists, and access to digital tracking tools and educational content (see Multimedia Appendix 1 for detailed program description). Although online weight tracking was mandatory for continued prescriptions, every other component of the Voy platform was optional.

Participants attended group onboarding sessions and were offered fortnightly coaching to enhance engagement and adherence. Coaches were trained based on principles from social cognitive theory, self-determination theory, the transtheoretical model, and the theory of planned behavior [12,13]. These techniques focused on fostering intrinsic motivation, goal setting, and problem-solving to promote sustainable lifestyle changes tailored to participants' individual progress.

Figure 1. Flowchart illustrating the critical steps of the cohort retrospective analysis conducted for the Voy digital weight management program. GIP: glucose-dependent insulinotropic polypeptide; GLP-1: glucagon-like peptide-1; RA: receptor agonist.

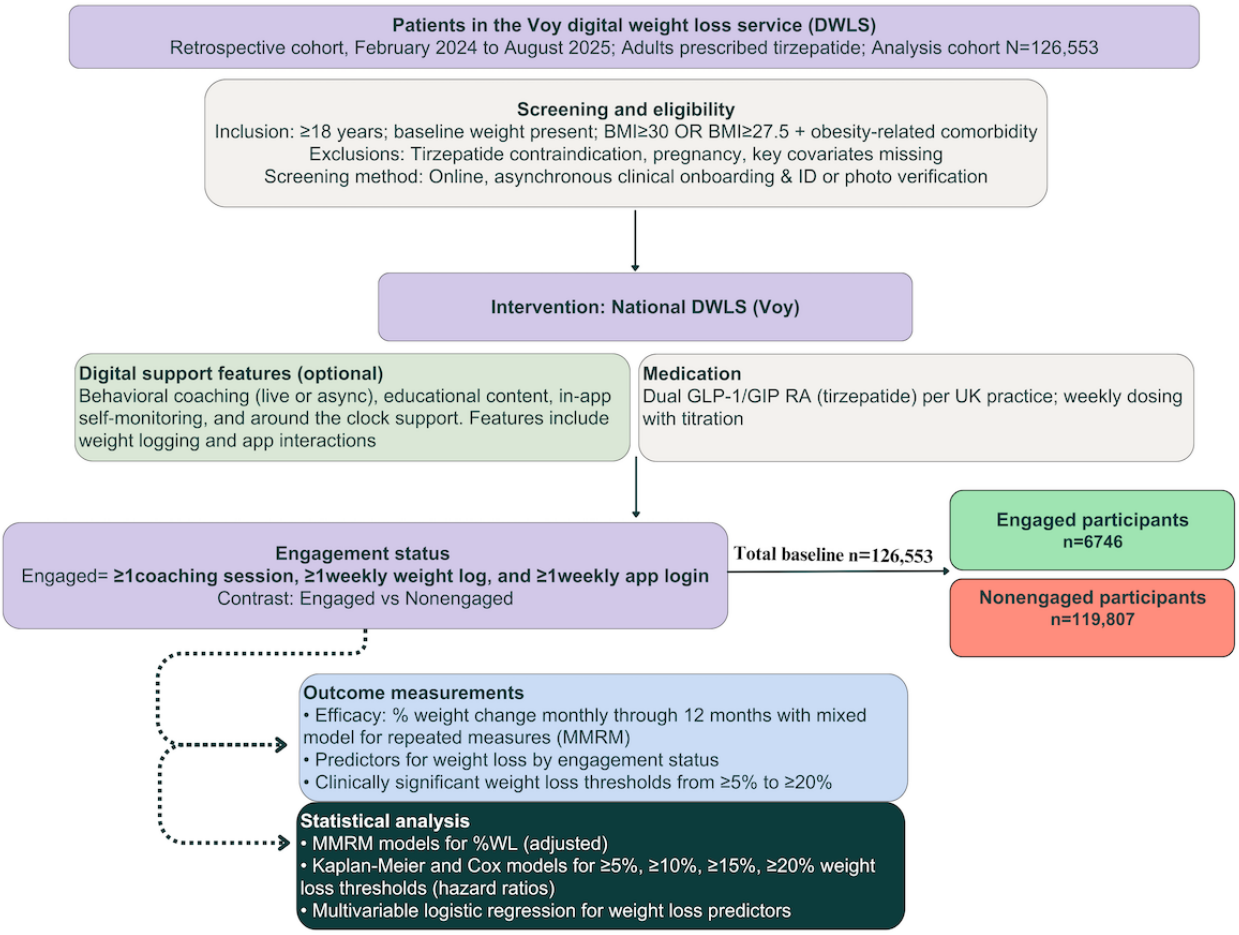

Participants

Participants were adult residents of the United Kingdom, aged between 18 and 75 years, with a BMI of 30 kg/m^2 or higher or >27.5 kg/m^2 with obesity-related comorbidities who were enrolled into the Voy digital weight loss service and were prescribed tirzepatide.

Eligibility Criteria

Eligibility required access to a smartphone or tablet. Exclusion criteria included history of self-reported eating disorders (eg, anorexia nervosa and bulimia nervosa), pregnancy or patients seeking fertility, known allergies or hypersensitivity to any components of tirzepatide, and severe medical conditions with established contraindications for tirzepatide as per Multimedia Appendix 2.

Data Completeness

Engagement metrics and weight loss outcomes were available for the entire sample of participants who used the digital platform during the study period at baseline. Complete demographic information (eg, age, sex, and comorbidities) was

XSL•FO
RenderX

available for the entire sample participants who enrolled during the initial phase of the service.

Defining Engagement and Outcome

Primary Outcome

The primary outcome was percentage weight loss from baseline. Weight measurements were entered directly in the app by participants.

Secondary Outcomes of Interest Included: Digital Engagement as an Exposure Variable

Engagement was defined based on a patient engaging in all 3 key behaviors identified as a priority by the Voy clinical and research teams as likely associated with weight loss outcomes. These behaviors are as follows:

- Coaching session attendance: participation in group or individual coaching sessions (video, audio, or text-based).
- Weight-tracking frequency: regular logging of body weight in the app.
- App use and logins: logging into the platform to view educational content, track health metrics, or interact with coaches.

As in our prior work [5], participants were classified as "engaged" if they met all of the following criteria: attended ≥1 coaching session AND tracked weight ≥1 per week AND logged into the app at least once during the study period. Those who did not fulfill all 3 of these criteria as a compound were classified as "nonengaged."

Statistical Analysis

Longitudinal Weight Loss Analysis

Our primary analysis used a mixed model for repeated measures (MMRM) approach, informed by the statistical methodology of the SURMOUNT-1 RCT study [4]. This approach accounted for repeated weight measurements for individuals, varying follow-up duration, and implicitly handled missing data without imputation under the missing at random assumption. The model incorporated fixed effects for digital engagement status, months, and the interaction between engagement and month. To adjust for potential confounding variables, we included sex, age, baseline BMI, and comorbid conditions as covariates (see Multimedia Appendix 3 for detailed methodology).

Sensitivity Analyses for Follow-Up Duration

To address the rolling enrollment design, we conducted sensitivity analyses restricted to subcohorts reaching specific follow-up milestones (3, 6, 9, and 12 months). Linear regression models were fitted for each subcohort with percentage weight change from baseline as the dependent variable, adjusted for the same confounders as the MMRM analyses. Marginal estimates (least squares means) were calculated from these models, with contrast *P* values comparing engaged versus nonengaged groups. These analyses assessed whether associations between engagement and weight loss remained consistent across different follow-up durations.

Postmodel Analyses and Effect Size Calculation

For postmodel analyses, we calculated adjusted means and conducted pairwise comparisons between engagement groups using the *emmeans* package. Significance testing of the engagement effect at each time point used 2-sided tests with α=.05, without adjustment for multiple comparisons, consistent with standard approaches for longitudinal repeated measures designs.

To quantify the magnitude and clinical relevance of the engagement effect, we calculated effect sizes using multiple approaches. Cohen *d* values were computed using the estimated difference divided by the pooled SD. Additionally, we computed both absolute differences (in percentage points) and relative differences (as a percentage of the nonengaged group's weight loss) between engaged and nonengaged tirzepatide users. To examine whether the impact of engagement on weight loss varied over time, we conducted a formal test of the engagement by time interaction.

Data Processing and Descriptive Statistics

All data processing and analyses were conducted using R software (version 4.3.1; R Foundation for Statistical Computing). Baseline characteristics were summarized using means and SDs for continuous variables, while frequencies and percentages were calculated for categorical variables. These statistics were calculated separately for engaged and nonengaged groups.

Baseline Correlates of Digital Engagement

Baseline characteristics associated with digital engagement were identified using multivariable logistic regression, where the dependent variable was digital engagement status (engaged vs not engaged, defined as mentioned earlier). Independent variables included demographic characteristics (age and sex), baseline BMI, and documented comorbidities (diabetes, hypertension, hypercholesterolemia, polycystic ovary syndrome [PCOS], and fatty liver disease). Age was modeled as a continuous variable, with additional transformations to express odds ratios (ORs) per decade increase. Similarly, BMI was analyzed with secondary calculations to express ORs per 5 kg/m^2 increase.

Achievement of Clinically Significant Weight Loss

Kaplan-Meier methods were used to estimate the proportions of users attaining ≥5%, ≥10%, ≥15%, and ≥20% weight loss during the study period with 95% CIs. This was performed for the cohort overall, then comparing engaged and nonengaged groups. Log-rank tests were used to compare cumulative incidence curves. Hazard ratios (HRs) with 95% CIs were calculated to quantify the relative rate of achieving weight loss thresholds between engagement groups, and risk ratios (RRs) were computed to assess the relative probability of threshold achievement.

Sample Size

A power analysis determined that a minimum of 118 participants per engagement group would provide 80% power to detect a 15% difference in the proportion of participants achieving ≥10% weight loss at a 5% significance level.

XSL•FO
RenderX

Bias and Missing Data

Self-reported weight measurements could introduce reporting bias. To mitigate this, participants were encouraged to provide accurate measurements through regular reminders and had the option to upload progress photographs, enhancing data validity. Additionally, data validation checks were performed internally and by statisticians to identify and address implausible values. Selection bias was minimized by including all eligible participants who initiated the program within the study period, ensuring the sample was representative of the population using the DWLS.

Missing data were addressed through the MMRM approach, which uses all available observations without requiring complete cases. This method operates under the missing at random assumption, whereby missingness may depend on observed covariates and previous measurements but not on unobserved values conditional on observed data. MMRM provides unbiased parameter estimates and maintains statistical efficiency in the presence of intermittent missing data, consistent with statistical methodology for landmark obesity trials [4]. For time-to-event analyses, participants who remained on treatment were censored at their last prescription date or study end, with censoring appropriately incorporated into Kaplan-Meier and Cox regression models.

Ethical Considerations

This retrospective open cohort study was an analysis of deidentified data collected during the routine clinical care of adults treated by Voy. The study was approved by the University College London Research Ethics Committee (Project ID 2025-0906-775). The study adhered to the principles outlined in the Declaration of Helsinki. Participants provided informed consent for their anonymized data to be used for ethically approved research and service improvement purposes upon enrollment in the program. To protect participant privacy, all data were deidentified prior to analysis, with direct identifiers (including names, contact details, and unique patient identifiers) removed and replaced with study-specific codes. Data were stored on secure servers with access restricted to authorized research personnel only. Participants did not receive any compensation for their participation in this study, as the analysis was conducted retrospectively using data collected during routine clinical care.

Adherence to STROBE Guidelines

This study adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [14] (Multimedia Appendix 4).

Results

Participant Characteristics

Among 126,553 participants who initiated tirzepatide, 6746 (5.3%) met criteria for maximally digital engagement. Baseline characteristics stratified by engagement status are presented in Table 1. The overall cohort had a mean age of 42.3 (SD 12.4) years, was predominantly female (99,905/126,553, 78.9%), and had a mean baseline BMI of 35.3 (SD 6.2) kg/m^2. Compared to nonengaged participants, digitally engaged individuals were significantly older (engaged: mean 44.9, SD 12.0 years vs nonengaged: mean 42.2, SD 12.5 years; $P<.001$), more likely to be female (engaged: 5486/6746, 81.3% vs nonengaged: 94,419/119,807, 78.8%; $P<.001$), and had higher baseline BMI (engaged: mean 36.4, SD 6.7 kg/m^2 vs nonengaged: mean 35.2, SD 6.2 kg/m^2; $P<.001$). Engaged participants also reported higher prevalence of comorbidities, including type 2 diabetes (engaged: 250/6746, 3.7% vs nonengaged: 3681/119,807, 3.1%; $P<.001$), hypertension (engaged: 978/6746, 14.5% vs nonengaged: 11,592/119,807, 9.7%; $P<.001$), hypercholesterolemia (engaged: 707/6746, 10.5% vs nonengaged: 7738/119,807, 6.5%; $P<.001$), and PCOS (engaged: 591/6746, 8.8% vs nonengaged: 7285/119,807, 6.1% of female participants; $P<.001$).

XSL•FO
RenderX

Table 1. Baseline demographic and clinical characteristics of participants by digital engagement status.

Characteristic	Overall (N=126,553)	Digitally engaged (n=6746)	Not digitally engaged (n=119,807)	*P* value[a]
Age (years)				<.001
Mean (SD)	42.3 (12.4)	44.9 (12.0)	42.2 (12.5)	
Age group (years), n (%)				<.001
18-24	6892 (5.4)	120 (1.8)	6772 (5.7)	
25-34	32,037 (25.3)	1399 (20.7)	30,638 (25.6)	
35-44	36,334 (28.7)	1977 (29.3)	34,357 (28.7)	
45-54	27,269 (21.5)	1643 (24.4)	25,626 (21.4)	
55+	24,016 (19)	1607 (23.8)	22,409 (18.7)	
Sex, n (%)				<.001
Female	99,905 (78.9)	5486 (81.3)	94,419 (78.8)	
Male	26,648 (21.1)	1260 (18.7)	25,388 (21.2)	
Anthropometric measures, mean (SD)				
Weight (kg)	98.3 (20.4)	101.9 (21.5)	98.1 (20.4)	<.001
Height (cm)	166.7 (9.1)	167.1 (8.7)	166.7 (9.2)	.003
BMI (kg/m^2)	35.3 (6.2)	36.4 (6.7)	35.2 (6.2)	<.001
BMI category, n (%)				<.001
Overweight (25-29.9 kg/m^2)	13,084 (10.4)	841 (12.5)	12,243 (10.2)	
Obese (≥30 kg/m^2)	113,246 (89.6)	5871 (87.5)	107,375 (89.8)	
Comorbidities, n (%)				
Diabetes mellitus	3931 (3.1)	250 (3.7)	3681 (3.1)	.009
Hypertension	12,570 (9.9)	978 (14.5)	11,592 (9.7)	<.001
Hypercholesterolemia	8445 (6.7)	707 (10.5)	7738 (6.5)	<.001
PCOS[b]	7876 (6.2)[c]	591 (10.8)[c]	7285 (6.1)[c]	<.001
Nonalcoholic fatty liver	2391 (1.9)	229 (3.4)	2162 (1.8)	<.001
Digital engagement metrics, n (%)				
Coaching sessions attended	37,224 (29.4)	6746 (100)	30,478 (25.4)	<.001
Weight tracking ≥1 per week readings	18,992 (15)	6746 (100)	12,246 (10.2)	<.001
Uses weight loss app[d]	93,627 (74)	6746 (100)	86,881 (72.5)	<.001

[a]*P* values calculated using independent 2-tailed *t* tests for continuous variables and chi-square tests for categorical variables.

[b]PCOS: polycystic ovary syndrome.

[c]Percentage calculated among female participants only.

[d]Defined as having logged into the app at least once during enrollment and used app features with the exception of weight measurement.

Engagement Characteristics

Among engaged participants (n=6746), all met the 3 required criteria by definition: coaching session attendance (mean 1.00, SD 0.00), app use (mean 1.00, SD 0.00), and weekly weight tracking (mean 1.00, SD 0.00). Among nonengaged participants (n=119,807), partial engagement was common: 86,881 (72.5%) used the app (mean 0.73, SD 0.45), 18,864 (15.7%) attended coaching (mean 0.16, SD 0.36), and tracking adherence (mean 0.09, SD 0.27) was lower. These patterns indicate that while most participants used some platform features, especially 93,627 of 126,553 (74%) of the overall cohort using the digital app, simultaneous engagement across all 3 modalities was achieved by 6,746 of 126,553 (5.3%) of users.

Weight Loss Outcomes

Overview

Figure 2 displays the MMRM model with compound symmetry covariance adjusted for confounders, representing percentage weight loss over 12 months by engagement status. Both digitally engaged and nonengaged groups demonstrated progressive weight loss throughout the observation period, but differences between groups emerged early and increased over time. Table 2 presents detailed weight loss outcomes at each time point.

XSL•FO
RenderX

Figure 2. Mean adjusted mixed model repeated measure derived and confounder-adjusted percentage weight loss from baseline among tirzepatide users stratified by digital engagement status.

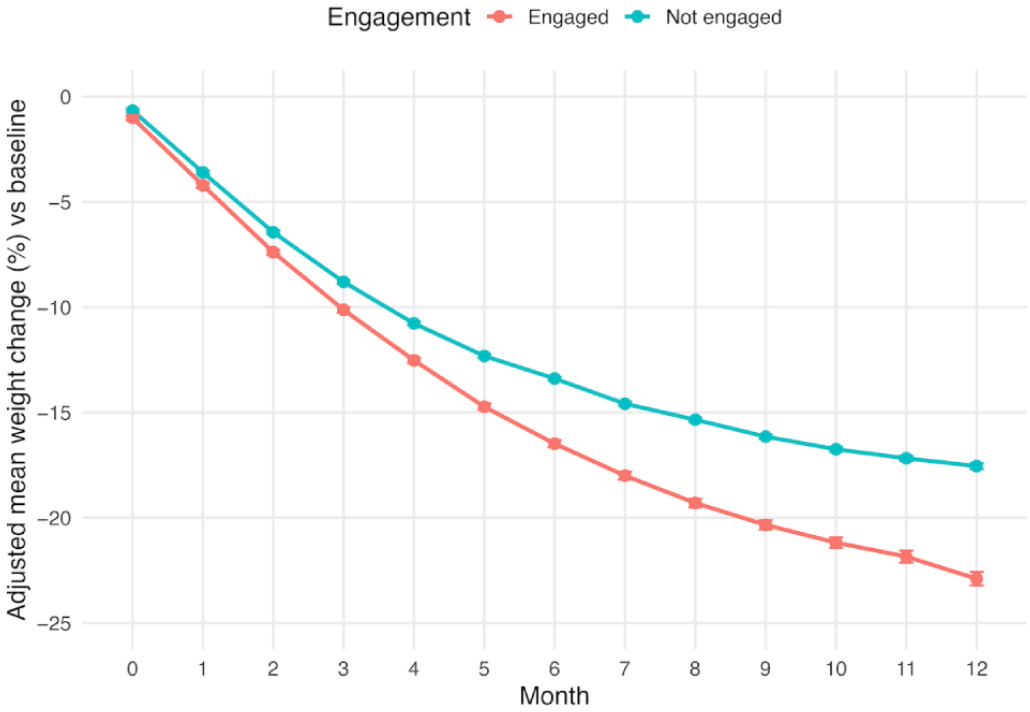

XSL•FO
RenderX

Table 2. Percentage weight loss by digital engagement status in participants taking tirzepatide (months 0-12), estimated using mixed model repeated measure (MMRM).

Month and engagement status	Participants, n	Adjusted mean % weight loss[a] (95% CI)	Absolute difference[b] (percentage points)	Relative difference[c] (%)	Effect size (Cohen *d*)	*P* value[d]
Month 0						
Not engaged	119,807	0	—[e]	—	—	—
Engaged	6746	0	—	—	—	—
Month 2						
Not engaged	40,582	–6.44 (–6.53 to –6.36)	0.95	14.8	0.083	<.001
Engaged	5751	–7.39 (–7.51 to –7.27)	—	—	—	—
Month 4						
Not engaged	21,762	–10.77 (–10.86 to –10.68)	1.75	16.2	0.207	<.001
Engaged	3420	–12.52 (–12.65 to –12.39)	—	—	—	—
Month 6						
Not engaged	13,559	–13.39 (–13.48 to –13.30)	3.09	23.1	0.422	<.001
Engaged	1724	–16.48 (–16.64 to –16.32)	—	—	—	—
Month 8						
Not engaged	6704	–15.34 (–15.45 to –15.24)	3.96	25.8	0.464	<.001
Engaged	896	–19.30 (–19.51 to –19.09)	—	—	—	—
Month 10						
Not engaged	3918	–16.75 (–16.87 to –16.63)	4.43	26.4	0.467	<.001
Engaged	567	–21.18 (–21.42 to –20.93)	—	—	—	—
Month 12						
Not engaged	2343	–17.55 (–17.69 to –17.41)	5.34	30.4	0.539	<.001
Engaged	310	–22.89 (–23.22 to –22.57)	—	—	—	—

[a]Adjusted means derived from MMRM with compound symmetry covariance structure, controlling for age, sex, baseline BMI, diabetes, hypertension, high cholesterol, polycystic ovary syndrome, and fatty liver disease.

[b]Absolute difference=engaged group mean–not engaged group mean (percentage points).

[c]Relative difference=(absolute difference/not engaged group mean)×100%.

[d]*P* values from the model time×engagement interaction terms.

[e]Not applicable.

Weight Loss by Engagement Status

By month 12, after adjusting for potential confounders (age, sex, baseline BMI, and comorbidities) in MMRM analysis, digitally engaged participants achieved a mean weight loss of 22.9% (95% CI 22.6-23.2) compared to 17.6% (95% CI 17.4-17.7) among nonengaged participants, representing an absolute difference of 5.3 percentage points ($P<.001$). The effect size (Cohen $d=0.539$) indicated a medium clinical effect of digital engagement on weight loss outcomes.

The divergence in adjusted weight loss trajectories between groups became statistically significant by month 2, with engaged participants achieving 7.4% weight loss compared to 6.4% among nonengaged participants (absolute difference 0.95 percentage points; $P<.001$). This difference progressively increased throughout the observation period. By month 6, the absolute difference reached 3.1 percentage points (16.5% vs 13.4%; $P<.001$), and by month 8, the difference widened to 4.0 percentage points (19.3% vs 15.3%; $P<.001$).

The MMRM analysis, which controlled for demographic factors and comorbidities while accounting for the correlation structure of repeated measurements, confirmed that digital engagement was consistently associated with enhanced weight loss outcomes at all time points. This statistical approach demonstrated that the engagement effect was not attributable to baseline differences or confounding factors but represented greater weight loss outcomes associated with higher engagement.

Sensitivity Analyses: Weight Loss by Follow-Up Duration

Among the cohort, 10,296 (8.1%) participants reached 3 months, 5349 (4.2%) reached 6 months, 2241 (1.8%) reached 9 months, and 2653 (2.1%) reached 12 months (see Multimedia Appendix 5 for discontinuation and censoring patterns).

Linear regression analyses adjusted for baseline weight and confounders demonstrated consistent associations between

XSL•FO
RenderX

engagement status and weight loss across all follow-up durations. At 3 months, nonengaged participants achieved 9.6% weight loss (95% CI 9.6-9.5) versus 10.3% (95% CI 10.5-10.2) for engaged participants (absolute difference 0.75 percentage points; *P*<.001). At 6 months, differences were 14.7% (95% CI 14.8-14.6) versus 17.6% (95% CI 17.9-17.3), respectively (absolute difference 2.9 percentage points; *P*<.001). At 9 months, nonengaged participants lost 17.8% (95% CI 18.1-17.6) compared to 21.8% (95% CI 22.4-21.2) for engaged participants (absolute difference 4.0 percentage points; *P*<.001). At 12 months, differences were 19.3% (95% CI 19.7-18.9) versus 24.2% (95% CI 25.3-23.2), respectively (absolute difference 4.9 percentage points; *P*<.001). These analyses confirmed that positive associations between engagement and weight loss were maintained regardless of the follow-up duration achieved.

Clinically Significant Weight Loss Thresholds

Kaplan-Meier survival analysis demonstrated that tirzepatide users achieved substantial weight loss milestones rapidly, with overall cumulative probabilities reaching high levels for clinically meaningful thresholds across all participants. When stratified by engagement status, digitally engaged participants achieved these clinically significant thresholds more rapidly and at higher rates than nonengaged participants (Figure 3, log-rank *P*<.001 for all comparisons). The association between engagement and increased weight loss was evident across all thresholds but became increasingly pronounced for more substantial weight loss targets.

Figure 3. Kaplan-Meier survival curves showing the probability of achieving clinically significant weight loss thresholds over time among tirzepatide users by digital engagement status. (A) ≥5% weight loss, (B) ≥10% weight loss, (C) ≥15% weight loss, and (D) ≥20% weight loss. Red lines represent digitally engaged participants; teal lines represent nonengaged participants. Shaded areas represent 95% CIs. *P* values determined by the log-rank test. χ is derived from the chi-square test comparing engaged versus nonengaged.

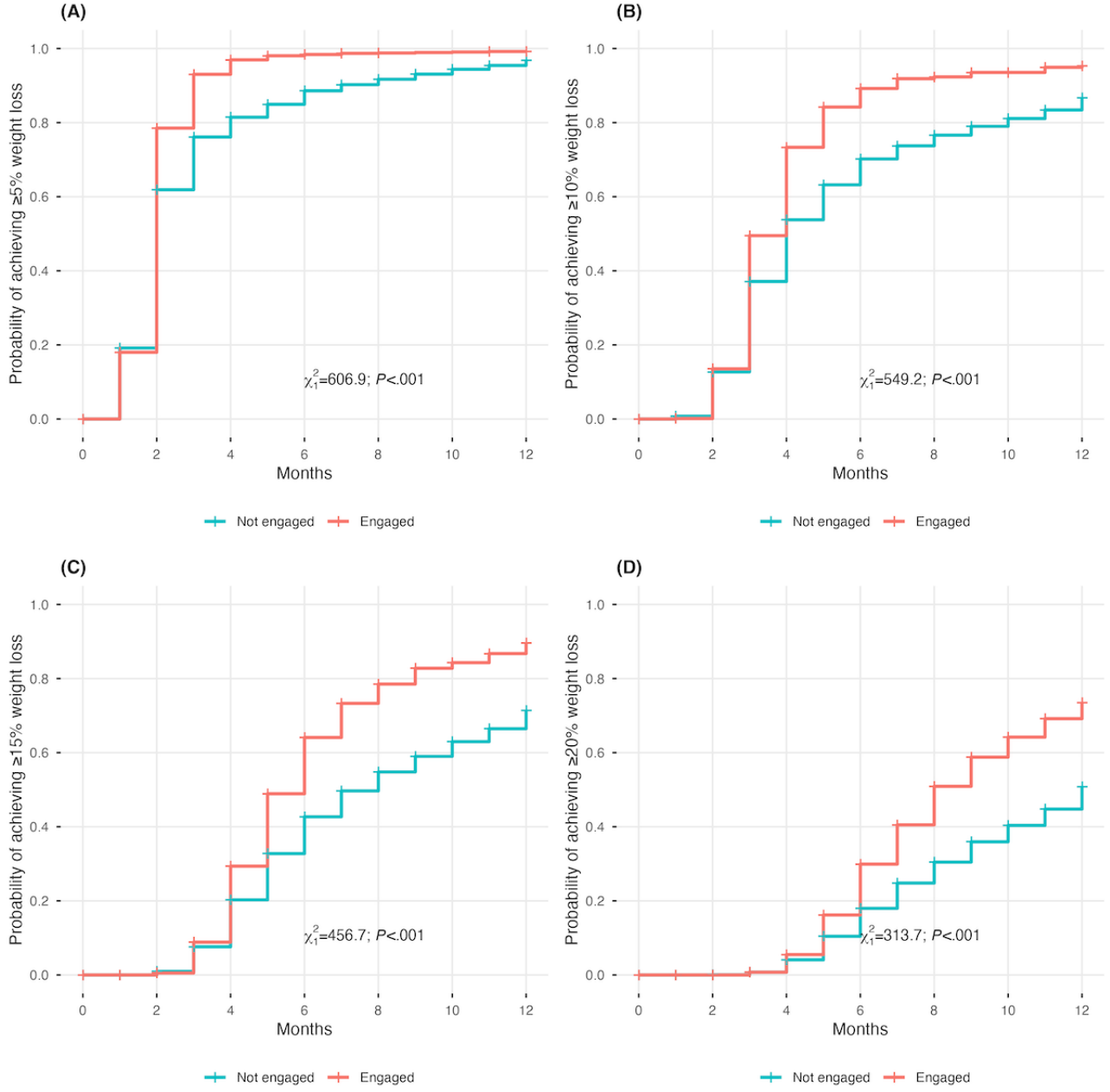

XSL•FO
RenderX

The cumulative probability of achieving ≥5% weight loss was substantially higher among engaged participants (5445/6746, 80.7% vs 44,329/119,807, 37%), demonstrating more than a 2-fold difference (RR 2.18; HR 1.49, 95% CI 1.45-1.53). This engagement advantage became even more pronounced for higher weight loss thresholds: ≥10% weight loss (3623/6746, 53.7% vs 26,240/119,807, 21.9%; RR 2.45; HR 1.55, 95% CI 1.50-1.60), ≥15% weight loss (2091/6746, 31% vs 14,017/119,807, 11.7%; RR 2.65; HR 1.67, 95% CI 1.60-1.75), and ≥20% weight loss (1079/6746, 16% vs 6710/119,807, 5.6%; RR 2.88; HR 1.79, 95% CI 1.68-1.91).

Longitudinal Participation Patterns

As expected, the number of participants with data at each time point decreased over the observation period, primarily reflecting the service's continuous enrollment pattern rather than attrition. Given the rolling nature of the service, participants joined at various time points throughout the evaluation period, with many not yet having the opportunity to reach later time points. The mean follow-up duration was 2.6 (SD 3.0) months, with substantial variation across participants reflecting the rolling enrollment design. Among the cohort, 2653 (2.1%) participants completed 12 months of follow-up. Of these, 310 (11.7%) were maximally engaged, and 2343 (88.3%) were nonengaged. Given the nature of the open cohort study, loss to follow-up could be due to discontinuation (stopped taking medication) or censoring (was still using medication at the cohort end date; Multimedia Appendix 5).

Baseline Correlates of Digital Engagement

Baseline characteristics associated with digital engagement are presented in Table 3. Male participants had significantly lower odds of engagement than female participants (adjusted OR for male: 0.86, 95% CI 0.81-0.92; $P<.001$). Each 10-year increase in age was associated with 18% higher odds of engagement (adjusted OR per decade 1.18, 95% CI 1.15-1.20; $P<.001$). Every 5 kg/m^2 increase in baseline BMI corresponded to 14% higher odds of engagement (adjusted OR 1.14, 95% CI 1.12-1.16; $P<.001$). Among comorbidities, participants with PCOS had 59% higher odds of engaging (adjusted OR 1.59, 95% CI 1.45-1.74; $P<.001$), and those with fatty liver disease had 52% higher odds (adjusted OR 1.52, 95% CI 1.32-1.76; $P<.001$). Hypercholesterolemia was associated with substantially higher engagement odds (adjusted OR 1.34, 95% CI 1.23-1.47; $P<.001$), while hypertension showed more modest associations (adjusted OR 1.16, 95% CI 1.07-1.25; $P<.001$). Notably, diabetes was associated with significantly lower odds of engagement (adjusted OR 0.83, 95% CI 0.73-0.95; $P=.009$).

Table 3. Baseline characteristics associated with digital engagement among tirzepatide users through multivariable logistic regression analysis[a].

Characteristic	Adjusted OR[b] (95% CI)	*P* value
Demographics		
Female sex	Reference (—[c])	—
Male sex	0.86 (0.81-0.92)	<.001
Age (per year)	1.02 (1.01-1.02)	<.001
Age (per decade)[d]	1.18 (1.15-1.20)	<.001
BMI (per unit)	1.03 (1.02-1.03)	<.001
BMI (per 5 kg/m^2)[d]	1.14 (1.12-1.16)	<.001
Comorbidities		
Polycystic ovary syndrome	1.59 (1.45-1.74)	<.001
Fatty liver disease	1.52 (1.32-1.76)	<.001
Hypertension	1.16 (1.07-1.25)	<.001
Hypercholesterolemia	1.34 (1.23-1.47)	<.001
Diabetes	0.83 (0.73-0.95)	.009

[a]The dependent variable was digital engagement status (engaged vs not engaged), defined as active use of digital coaching, tracking tools, or a mobile app. All characteristics were assessed at baseline (month 0) prior to tirzepatide initiation. Multivariable logistic regression model adjusted for all variables shown.

[b]OR: odds ratio.

[c]Not applicable.

[d]Derived values to express effect sizes for clinically meaningful increments.

Discussion

Principal Findings

This cohort study demonstrated the notable real-world effectiveness of tirzepatide in a UK-based DWLS and highlighted an association between higher digital engagement and greater weight loss. After controlling for demographic factors and comorbidities, digitally engaged participants achieved an absolute 5.3 percentage points greater weight loss at month 12 compared to nonengaged users ($P<.001$). This engagement effect emerged gradually, with trajectories

XSL•FO
RenderX

beginning to diverge around month 2 and differences progressively widening thereafter. By month 12, digitally engaged participants achieved a mean weight loss of 22.9% compared to 17.6% for nonengaged participants.

Nearly all tirzepatide users eventually achieved ≥5% weight loss regardless of engagement status, but more substantial differences were observed for the engaged group for higher weight loss thresholds. Notably, engaged participants demonstrated substantially higher cumulative probabilities of achieving ≥15% weight loss (2091/6746, 31% vs 14,017/119,807, 11.7%; HR 1.67, 95% CI 1.60-1.75) and ≥20% weight loss (1079/6746, 16% vs 6710/119,807, 5.6%; HR 1.79, 95% CI 1.68-1.91) by study end. The temporal pattern of weight loss differences is meaningful. While statistically significant differences emerged by month 2, the magnitude of difference progressively increased throughout the observation period. In other studies, a 21% weight loss end point was typically reached at week 72 as opposed to at week 52 (month 12) as evidenced in our findings [15].

The temporal differences in weight loss associated with engagement status were further highlighted in our analysis, which revealed significantly accelerated achievement of clinically meaningful weight loss thresholds. While engagement conferred substantial advantages for ≥5% weight loss achievement (5445/6746, 80.7% vs 44,329/119,807, 37%; RR 2.18; HR 1.49, 95% CI 1.45-1.53), more substantial differences emerged for more ambitious thresholds. Notably, engaged participants demonstrated substantially higher cumulative probabilities of achieving ≥10% weight loss (3623/6746, 53.7% vs 26,240/119,807, 21.9%; RR 2.45; HR 1.55, 95% CI 1.50-1.60), ≥15% weight loss (2091/6746, 31% vs 14,017/119,807, 11.7%; RR 2.65; HR 1.67, 95% CI 1.60-1.75), and ≥20% weight loss (1079/6746, 16% vs 6710/119,807, 5.6%; RR 2.88; HR 1.79, 95% CI 1.68-1.91).

Our analysis of baseline characteristics identified that female sex, older age, and higher BMI were associated with increased engagement likelihood, with male participants having 14% lower odds (OR 0.86, 95% CI 0.81-0.92), each decade of age conferring 18% higher odds (OR 1.18, 95% CI 1.15-1.20), and every 5 kg/m^2 BMI increase associated with 14% higher odds (OR 1.14, 95% CI 1.12-1.16; all P<.001). Among comorbidities, PCOS (59% higher odds; OR 1.59, 95% CI 1.45-1.74), fatty liver disease (52% higher odds; OR 1.52, 95% CI 1.32-1.76), and hypercholesterolemia (34% higher odds; OR 1.34, 95% CI 1.23-1.47) were positively associated with engagement, while diabetes was associated with 17% lower engagement odds (OR 0.83, 95% CI 0.73-0.95;; P=.009). These findings highlight important demographic factors influencing digital health use that may inform future program design and implementation.

Strengths and Limitations

This study has several strengths, including its large sample size, real-world setting, and comprehensive assessment of weight trajectories over time. The 126,000-participant cohort provides robust power at early time points (n=40,582 at 2 months); while 12-month completers (n=2653) match other digital studies, the real-world implementation enhances generalizability and ecological validity. Additionally, the operational definition of digital engagement captured meaningful interaction with service components while maintaining practical relevance.

While baseline differences between engaged and nonengaged participants reached statistical significance due to the large sample size (n=126,553), the absolute magnitude of these differences was modest and unlikely to be clinically meaningful. The mean age difference was less than 1 year (mean 42.3, SD 12.5 vs mean 44.9, SD 12.0 years), and the BMI difference was approximately 1 kg/m^2 (mean 35.2, SD 6.2 vs mean 36.4, SD 6.7 kg/m^2). These small baseline differences are insufficient to account for the substantial treatment effects observed, particularly the 7.4% absolute difference in weight loss at 2 months and the near 3-fold difference in treatment persistence. Furthermore, the consistency association between engagement and outcomes across all time points, combined with our multivariable adjustment for baseline characteristics, supports the conclusion that digital engagement genuinely enhances pharmacological efficacy rather than these findings being attributable to baseline confounding.

Several limitations warrant consideration. First, as a retrospective service evaluation, participants were not randomly assigned to engagement conditions, introducing potential selection bias, and we cannot ascertain the causality of this association. While we adjusted for observed confounders, unmeasured factors (eg, motivation and socioeconomic status) may influence both engagement and outcomes. Second, weight measurements were self-reported via home scales, potentially introducing measurement and reporting error. However, this limitation applied to both engaged and nonengaged groups, likely minimizing differential bias. We did not systematically capture concurrent participation in external nonpharmacological weight management programs; we note this as a limitation and as a priority for future data collection. Additionally, ethnicity data were not systematically collected during the study period, limiting our ability to examine potential disparities in engagement or outcomes across ethnic groups. Future implementations of DWLS should prioritize collecting comprehensive demographic data, including ethnicity and socioeconomic indicators, to ensure equity in access and outcomes and to identify populations that may benefit from targeted engagement strategies. The monthly cost for the service in this study service represents a significant financial consideration that may limit access primarily to individuals with higher socioeconomic status. This raises important considerations regarding health inequalities, as individuals who might benefit most from weight management interventions may face financial barriers to accessing digital health services that integrate pharmacotherapy with behavioral support.

Moreover, the rolling enrollment design means that participants were at different stages of their treatment journey during the evaluation period. While this reflects real-world implementation, it complicates the interpretation of longitudinal patterns. Our approach of examining both absolute outcomes at each time point and conducting subgroup analyses of participants with sufficient follow-up time helps address this limitation. Finally, our operational definition of digital engagement, while

XSL•FO
RenderX

evidence-informed, represents one approach among many possible definitions. Future research should explore alternative engagement metrics and potential dose-response relationships between engagement intensity and outcomes.

Comparison With Prior Work

The magnitude of additional weight loss associated with digital engagement was substantial in the context of standard obesity treatment. Previous research indicates that each 5% reduction in body weight confers significant cardiometabolic benefits, including improvements in glycemic control, blood pressure, and lipid profiles [16,17]. Our findings suggest that digital engagement may help individuals reach more substantial weight loss thresholds, potentially amplifying the health benefits of tirzepatide treatment. This suggests that digital engagement may have cumulative benefits, potentially by supporting medication adherence, dietary modifications, and behavioral changes that enhance long-term outcomes [18,19]. Similar patterns have been observed in other behavioral interventions for chronic disease management, where ongoing support provides incremental benefits over time [20].

The identified correlates of digital engagement offer insights for service optimization. The higher likelihood of engagement among female participants, older adults, and those with higher baseline BMI suggests that these demographic groups may find digital support particularly valuable. Conversely, younger male participants with lower BMI appear less likely to engage, potentially benefiting from targeted engagement strategies; yet, digital health engagement literature has shown that male participants benefit more than female participants [21]. Interestingly, studies have proposed that people with diabetes have a slower velocity of weight loss in both the GLP-1 [22] and GLP and GIP-1 [23] agonist exposure groups (albeit, faster velocity in tirzepatide vs semaglutide).

Furthermore, our study confirmed that diabetes status was significantly associated with a lower likelihood of digital engagement (adjusted OR 0.83, 95% CI 0.73-0.95; P=.009). This may reflect complex physiological and psychological mechanisms influencing weight loss velocity in people with diabetes. To address this, DWLS should consider implementing more intensive engagement and coaching interventions tailored for patients with diabetes. We propose that digital services pilot such enhanced approaches to test whether outcomes can be improved. By contrast, the positive association between comorbidities, particularly PCOS [24] and engagement, suggests that individuals with obesity-related health concerns may perceive greater value in comprehensive support services.

Several mechanisms may explain the enhanced outcomes associated with digital engagement. Health coaching provides accountability, problem-solving support, and personalized guidance that can address common barriers to weight management [25]. Regular weight tracking enhances self-monitoring, a behavioral strategy consistently associated with improved weight outcomes [8]. Additionally, digital platforms may facilitate greater medication adherence through reminders, side effect management, and addressing various concerns, particularly important for injectable medications like tirzepatide that require consistent administration [26].

Our findings align with previous digital weight management studies while revealing important distinctions. Xu et al [27] (n=153) defined engagement as any daily food tracking, identifying thresholds of 28.5%-39.4% of days for ≥3%-5% weight loss. W8Buddy used a liberal definition (any platform activity within 14 days), achieving 83.1% engagement with 0.74 kg per month additional weight loss. Our study used the strictest definition (coaching, weekly weight tracking, and app login combined), resulting in only 5.3% meeting criteria, yet demonstrating the largest effect (5.3 percentage points at 12 months) [28]. This contrast, where stricter criteria with fewer engaged participants yielding greater outcomes, suggests that multimodal engagement combining behavioral support, self-monitoring, and clinical oversight produces beneficial effects beyond single-modality interventions. Notably, partial engagement remained common (app use: 93,627/126,553, 74% and coaching: 37,224/126,553, 29.4%), and even nonengaged participants achieved substantial weight loss (17.6%), indicating that platform features provided benefit across engagement levels. These findings demonstrate that while engagement definitions substantially influence observed rates, the consistent association between digital engagement and enhanced outcomes persists across platforms and intervention designs.

The integration of digital services with pharmacotherapy represents an evolving care model addressing limitations of traditional obesity treatment approaches. By providing remote support, behavioral tools, and clinical monitoring, DWLSs may bridge critical gaps in obesity care while reducing barriers related to geography, provider availability, and stigma [29,30]. Our findings suggest that such integrated approaches may optimize the effectiveness of novel obesity pharmacotherapies.

Implications and Future Directions

Our findings have important implications for clinical practice, health service design, and research. Clinically, practitioners should consider recommending digital support services alongside tirzepatide prescriptions, particularly for patients who may benefit from additional accountability and behavioral guidance. From a service design perspective, our results support investment in research and development for digital infrastructure that facilitates engagement, particularly during early treatment phases when engagement patterns appear to be established.

Some self-funded DWLS can exacerbate health inequalities; therefore, future research should examine the cost-effectiveness of such integrated DWLS compared to medication-only approaches and explore alternative funding models, including potential National Health Service commissioning pathways, that could improve accessibility across socioeconomic groups while maintaining service quality and clinical outcomes [31]. Additionally, qualitative studies exploring patient experiences and clinician perspectives could identify specific digital components that contribute most to engagement and outcomes. Finally, RCTs comparing different digital support intensities would address causality questions and optimize resource allocation.

XSL•FO
RenderX

Conclusions

This study demonstrates that digital engagement was significantly associated with enhanced weight loss outcomes among individuals using tirzepatide for obesity management. By month 12, engaged participants achieved –22.9% weight loss and an absolute difference of –5.3% compared to nonengaged participants, with differences emerging early and increasing over time. These findings highlight the potential value of integrated care models combining pharmacotherapy with digital support services. As novel obesity medications continue to transform treatment possibilities, optimizing their implementation through complementary digital strategies represents a promising approach to addressing the global obesity epidemic.

Acknowledgments

The authors confirm that no generative artificial intelligence tools were used to generate the manuscript text, analyses, figures, or references.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the nature of the clinical data collected and the consent provided by patients; the individual participant data used in this study are not publicly available. The statistical code used in the analyses can be made available to researchers upon request to the corresponding author. Requests for access to aggregated high-level anonymized data for noncommercial projects should be directed to the corresponding author, will require appropriate ethics and legal approval, and are available on reasonable request.

Authors' Contributions

Conceptualization: HJ, AKC, DH
Methodology: HJ (lead), AKC (supporting)
Formal analysis: HJ (lead), AKC (supporting)
Validation: HJ, AKC, DR, DH
Investigation: HJ
Data curation: HJ (lead), AKC (supporting)
Visualization: HJ
Supervision: AKC, DH
Project administration: HJ
Writing—original draft: HJ (lead), AKC (supporting)
Writing—review and editing: HJ (lead), AKC (supporting), DR (supporting), DH (supporting)

Conflicts of Interest

The authors HJ, DH, AKC, and DR are members within the organization Voy, Menwell; HJ is the clinical researcher, AKC is the head of clinical research, DH is the innovation director, and DR is the research advisor.

Multimedia Appendix 1
Detailed program description.
[DOCX File , 24 KB - jmir_v28i1e83718_app1.docx]

Multimedia Appendix 2
Eligibility criteria and cohort derivation of the study.
[DOCX File , 17 KB - jmir_v28i1e83718_app2.docx]

Multimedia Appendix 3
Statistical methods and model specifications to run on R.
[DOCX File , 17 KB - jmir_v28i1e83718_app3.docx]

Multimedia Appendix 4
STROBE checklist.

XSL•FO
RenderX

[PDF File (Adobe PDF File), 184 KB - jmir_v28i1e83718_app4.pdf]

Multimedia Appendix 5
Participant flow diagram showing retention, discontinuation, and censoring patterns over 12 months of follow-up.
[DOCX File , 75 KB - jmir_v28i1e83718_app5.docx]

References

1. Okunogbe A, Nugent R, Spencer G, Ralston J, Wilding J. Economic impacts of overweight and obesity: current and future estimates for eight countries. BMJ Glob Health 2021;6(10):e006351 [FREE Full text] [doi: 10.1136/bmjgh-2021-006351] [Medline: 34737167]
2. Hajjaj AB, Guijarro PM, Khan KS, Bueno-Cavanillas A, Cano-Ibáñez N. Author correction: a systematic review and meta-analysis of weight loss in control group participants of lifestyle randomized trials. Sci Rep 2022;12(1):14444 [FREE Full text] [doi: 10.1038/s41598-022-18828-y] [Medline: 36002563]
3. Syed YY. Tirzepatide: first approval. Drugs 2022;82(11):1213-1220. [doi: 10.1007/s40265-022-01746-8] [Medline: 35830001]
4. Jastreboff AM, Aronne LJ, Ahmad NN, Wharton S, Connery L, Alves B, et al. Tirzepatide once weekly for the treatment of obesity. N Engl J Med 2022;387(3):205-216. [doi: 10.1056/NEJMoa2206038] [Medline: 35658024]
5. Johnson H, Huang D, Liu V, Ammouri MA, Jacobs C, El-Osta A. Impact of digital engagement on weight loss outcomes in obesity management among individuals using GLP-1 and dual GLP-1/GIP receptor agonist therapy: retrospective cohort service evaluation study. J Med Internet Res 2025;27:e69466 [FREE Full text] [doi: 10.2196/69466] [Medline: 40164173]
6. Tronieri JS, Wadden TA, Chao AM, Tsai AG. Primary care interventions for obesity: review of the evidence. Curr Obes Rep 2019;8(2):128-136 [FREE Full text] [doi: 10.1007/s13679-019-00341-5] [Medline: 30888632]
7. Richards R, Wren G, Whitman M. The potential of a digital weight management program to support specialist weight management services in the UK National Health Service: retrospective analysis. JMIR Diabetes 2024;9:e52987 [FREE Full text] [doi: 10.2196/52987] [Medline: 38265852]
8. Burke LE, Wang J, Sevick MA. Self-monitoring in weight loss: a systematic review of the literature. J Am Diet Assoc 2011;111(1):92-102 [FREE Full text] [doi: 10.1016/j.jada.2010.10.008] [Medline: 21185970]
9. Lau Y, Chee DGH, Chow XP, Cheng LJ, Wong SN. Personalised eHealth interventions in adults with overweight and obesity: a systematic review and meta-analysis of randomised controlled trials. Prev Med 2020;132:106001. [doi: 10.1016/j.ypmed.2020.106001] [Medline: 31991155]
10. Kouvari M, Karipidou M, Tsiampalis T, Mamalaki E, Poulimeneas D, Bathrellou E, et al. Digital health interventions for weight management in children and adolescents: systematic review and meta-analysis. J Med Internet Res 2022;24(2):e30675 [FREE Full text] [doi: 10.2196/30675] [Medline: 35156934]
11. Combet E, Haag L, Richardson J, Haig CE, Cunningham Y, Fraser HL, et al. Remotely delivered weight management for people with long COVID and overweight: the randomized wait-list-controlled ReDIRECT trial. Nat Med 2025;31(1):258-266. [doi: 10.1038/s41591-024-03384-x] [Medline: 39779922]
12. Cheung KL, Eggers SM, de Vries H. Combining the integrated-change model with self-determination theory: application in physical activity. Int J Environ Res Public Health 2020;18(1):28 [FREE Full text] [doi: 10.3390/ijerph18010028] [Medline: 33374522]
13. Farmanbar R, Niknami S, Lubans DR, Hidarnia A. Predicting exercise behaviour in Iranian college students: utility of an integrated model of health behaviour based on the transtheoretical model and self-determination theory. Health Educ J 2012;72(1):56-69. [doi: 10.1177/0017896911430549]
14. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007;335(7624):806-808 [FREE Full text] [doi: 10.1136/bmj.39335.541782.AD] [Medline: 17947786]
15. Tan B, Pan X, Chew HSJ, Goh RSJ, Lin C, Anand VV, et al. Efficacy and safety of tirzepatide for treatment of overweight or obesity. A systematic review and meta-analysis. Int J Obes (Lond) 2023;47(8):677-685. [doi: 10.1038/s41366-023-01321-5] [Medline: 37253796]
16. Ryan DH, Yockey SR. Weight loss and improvement in comorbidity: differences at 5%, 10%, 15%, and over. Curr Obes Rep 2017;6(2):187-194 [FREE Full text] [doi: 10.1007/s13679-017-0262-y] [Medline: 28455679]
17. Wing RR, Lang W, Wadden TA, Safford M, Knowler WC, Bertoni AG, et al. Benefits of modest weight loss in improving cardiovascular risk factors in overweight and obese individuals with type 2 diabetes. Diabetes Care 2011;34(7):1481-1486 [FREE Full text] [doi: 10.2337/dc10-2415] [Medline: 21593294]
18. Boucher EM, Raiker JS. Engagement and retention in digital mental health interventions: a narrative review. BMC Digit Health 2024;2(1):52. [doi: 10.1186/s44247-024-00105-9]
19. Lehmann M, Jones L, Schirmann F. App engagement as a predictor of weight loss in blended-care interventions: retrospective observational study using large-scale real-world data. J Med Internet Res 2024;26:e45469 [FREE Full text] [doi: 10.2196/45469] [Medline: 38848556]

XSL•FO
RenderX

20. Wadden TA, Tronieri JS, Butryn ML. Lifestyle modification approaches for the treatment of obesity in adults. Am Psychol 2020;75(2):235-251 [FREE Full text] [doi: 10.1037/amp0000517] [Medline: 32052997]
21. König LM, Western MJ, Denton AH, Krukowski RA. Umbrella review of social inequality in digital interventions targeting dietary and physical activity behaviors. NPJ Digit Med 2025;8(1):11 [FREE Full text] [doi: 10.1038/s41746-024-01405-0] [Medline: 39762352]
22. Jensterle M, Rizzo M, Haluzík M, Janež A. Efficacy of GLP-1 RA approved for weight management in patients with or without diabetes: a narrative review. Adv Ther 2022;39(6):2452-2467 [FREE Full text] [doi: 10.1007/s12325-022-02153-x] [Medline: 35503498]
23. Viljoen A, Pantalone KM, Galindo RJ, Cui X, Huh R, Hemmingway A, et al. Time to reach glycaemic and body weight loss thresholds with tirzepatide in patients with type 2 diabetes: a pre-planned exploratory analysis of SURPASS-2 and SURPASS-3. Diabetes Ther 2023;14(5):925-936 [FREE Full text] [doi: 10.1007/s13300-023-01398-1] [Medline: 37000390]
24. Monney M, Mavromati M, Leboulleux S, Gariani K. Endocrine and metabolic effects of GLP-1 receptor agonists on women with PCOS, a narrative review. Endocr Connect 2025;14(5):6975 [FREE Full text] [doi: 10.1530/EC-24-0529] [Medline: 40066975]
25. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
26. Oakley-Girvan I, Yunis R, Longmire M, Ouillon JS. What works best to engage participants in mobile app interventions and e-Health: a scoping review. Telemed J E Health 2022;28(6):768-780 [FREE Full text] [doi: 10.1089/tmj.2021.0176] [Medline: 34637651]
27. Xu R, Bannor R, Cardel MI, Foster GD, Pagoto S. How much food tracking during a digital weight-management program is enough to produce clinically significant weight loss? Obesity (Silver Spring) 2023 Jul;31(7):1779-1786 [FREE Full text] [doi: 10.1002/oby.23795] [Medline: 37271576]
28. Hanson P, Abdelhameed F, Sahir M, Parsons N, Panesar A, de la Fosse M, et al. Evaluation of the digital support tool Gro Health W8Buddy as part of tier 3 weight management service: observational study. J Med Internet Res 2025 May 16;27:e62661 [FREE Full text] [doi: 10.2196/62661] [Medline: 40378402]
29. Talay L, Alvi O. Digital healthcare solutions to better achieve the weight loss outcomes expected by payors and patients. Diabetes Obes Metab 2024;26(6):2521-2523. [doi: 10.1111/dom.15513] [Medline: 38379435]
30. Barron E, Bradley D, Safazadeh S, McGough B, Bakhai C, Young B, et al. Effectiveness of digital and remote provision of the Healthier You: NHS Diabetes Prevention Programme during the COVID-19 pandemic. Diabet Med 2023;40(5):e15028 [FREE Full text] [doi: 10.1111/dme.15028] [Medline: 36524707]
31. Thomsen RW, Mailhac A, Løhde JB, Pottegård A. Real-world evidence on the utilization, clinical and comparative effectiveness, and adverse effects of newer GLP-1RA-based weight-loss therapies. Diabetes Obes Metab 2025;27(Suppl 2):66-88. [doi: 10.1111/dom.16364] [Medline: 40196933]

Abbreviations

DWLS: digital weight loss service
GIP: glucose-dependent insulinotropic polypeptide
GLP-1: glucagon-like peptide-1
HR: hazard ratio
MMRM: mixed model for repeated measures
OR: odds ratio
PCOS: polycystic ovary syndrome
RA: receptor agonist
RCT: randomized controlled trial
RR: risk ratio
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by N Cahill; submitted 07.Sep.2025; peer-reviewed by G Foster, P Hanson; comments to author 27.Oct.2025; revised version received 11.Nov.2025; accepted 26.Nov.2025; published 15.Jan.2026.

Please cite as:
Johnson H, Clift AK, Reisel D, Huang D
Digital Engagement Significantly Enhances Weight Loss Outcomes in Adults With Obesity Treated With Tirzepatide: Retrospective Cohort Study of a Digital Weight Loss Service
J Med Internet Res 2026;28:e83718
URL: https://www.jmir.org/2026/1/e83718
doi:10.2196/83718
PMID:

XSL•FO
RenderX

©Hans Johnson, Ashley Kieran Clift, Daniel Reisel, David Huang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Relationship Between Physician Self-Disclosure and Patient Acquisition in Digital Health Markets: Cross-Sectional Study

Quanchen Liu[1], PhD; Pengqing Yin[1], MSc; Jing Fan[1], PhD

International Business School, Beijing Foreign Studies University, Beijing, China

Corresponding Author:
Jing Fan, PhD
International Business School
Beijing Foreign Studies University
No. 19 Xisanhuan North Road
Beijing, 100089
China
Phone: 86 (010) 88818121
Email: fanjing@bfsu.edu.cn

Abstract

Background: Online health communities have evolved into digital marketplaces where physicians have to compete for patients. Existing research examines physician-patient dynamics through a patient-centric lens, treating physicians as passive recipients of ratings and reviews, while the strategic role of physician self-disclosure remains unexamined. This gap constrains a comprehensive understanding of how physicians can actively shape patient decisions, making the investigation of strategic self-disclosure imperative.

Objective: This study aims to investigate the relationship between physician self-disclosure breadth (scope of information) and depth (detailed expertise) and patient decision-making, as well as whether regional digital health care level (DHL) moderates these relationships.

Methods: We conducted a cross-sectional analysis of observational data to test these relationships. Data were collected from China's online health care platform Haodf from September to December 2024. Self-disclosure breadth (including clinical performance, academic experience, and social reputation), self-disclosure depth (including expertise coverage, richness, and granularity), and patient decision-making (total visits) were captured through manual content coding and quantitative measurement. We used structured content analysis to extract the disclosure components, informational scope, and descriptive details of each profile. Then, using validated operational formulas, we calculated the composite indices for disclosure breadth and depth based on the coded dimensions. The study generated 1798 final physician samples with complete data across 14 focal variables. The hypotheses were tested using an ordinary least squares regression model, and 4 robustness checks were conducted, including variable substitution and different resampling techniques.

Results: In the primary ordinary least squares regression models, self-disclosure breadth was significantly and positively associated with patient visits (β=0.255, 95% CI 0.054-0.456; *P*=.01), as was self-disclosure depth (β=0.098, 95% CI 0.030-0.167; *P*=.005). The breadth×DHL interaction was positive and significant (β=0.261, 95% CI 0.061-0.461; *P*=.01). Similarly, the depth×DHL interaction was positive and significant (β=0.070, 95% CI 0.002-0.138; *P*=.045). It should be noted that the association for self-disclosure breadth was stronger than that of self-disclosure depth. DHL strengthened the relationship between the disclosure strategies with patient visits. This contextual amplification indicates that DHL serves as a critical boundary condition, determining the degree to which physician self-disclosure strategies translate into patient acquisition outcomes.

Conclusions: This study reconceptualizes physicians as strategic agents shaping patient decision-making through purposeful self-disclosure. Different from existing studies treating physicians as passive recipients of ratings and reviews, our research demonstrates that physicians can strategically shape patient acquisition through self-disclosure breadth and depth. This study brings new insights to digital health markets by demonstrating that self-disclosure operates as a viable patient acquisition mechanism, wherein the DHL acts as a critical boundary condition. The findings have real-world implications: (1) physicians can leverage evidence-based disclosure strategies, (2) platforms should implement context-adaptive features, and (3) policymakers should prioritize digital infrastructure investments to enhance physicians' competitive capabilities and patient decision-making quality.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e84963) doi:10.2196/84963

KEYWORDS

online health communities; physician online profile; physician self-disclosure; patient decision-making; physician-patient interaction; digital health care level

Introduction

Background

The health care landscape is experiencing an unprecedented digital transformation, with online health communities (OHCs) emerging as powerful intermediaries that fundamentally reshape patient-physician interactions. OHCs democratize medical information access and empower patients to actively evaluate health care providers before making consultation decisions [1]. Platforms, such as HealthTap [2] and China's Haodf, now serve millions of users globally, with the OHCs market projected to expand from US $13.3 billion in 2022 to US $42.9 billion by 2030 [3]. This revolution forces physicians to build compelling online presences beyond clinical excellence to compete effectively in an increasingly crowded digital marketplace.

Review of Relevant Scholarship

Existing research on OHCs has predominantly examined physician-patient dynamics through a patient-centric lens, treating physicians as passive recipients of online ratings and reviews rather than strategic actors capable of influencing patient decisions. Current studies focus extensively on how patients leverage physician profiles, ratings, and accumulated reviews to screen health care providers [4], effectively positioning physicians as static entities whose past digital footprints predetermine patient selection outcomes. This perspective fundamentally overlooks physicians' potential for active agency in patient acquisition. Moreover, while scholars acknowledge regional variations in digital health care level (DHL [5,6]), encompassing the sophistication of technological infrastructure, digital literacy capabilities, and information accessibility within specific health care environments), little attention has been paid to how these contextual differences might alter physicians' strategic opportunities and patients' information processing capabilities, creating a significant theoretical blind spot in understanding physician behavior within digitally heterogeneous health care environments.

In an increasingly crowded OHC landscape where patients can choose from hundreds of providers, physicians must now strategically differentiate themselves through deliberate self-presentation beyond clinical excellence and accumulated reviews. Within this context, self-disclosure theory from social psychology offers a promising framework, suggesting that strategic information disclosure across breadth (scope of information) and depth (level of detail) dimensions can enhance credibility, build trust, and reduce decision-making uncertainty [7], thereby influencing differential patient choices. However, the effectiveness of physician self-disclosure cannot be understood in isolation from regional digital health care contexts. In technologically advanced regions, physicians access sophisticated multimedia tools that enhance disclosure opportunities, while patients possess higher digital literacy, enabling effective interpretation of complex professional information [8]. Conversely, less digitally developed areas present technological constraints limiting physicians' ability to communicate expertise effectively, while patients may lack sufficient digital literacy to process and verify disclosed information [9]. This contextual complexity suggests that physician self-disclosure effectiveness may vary across digital health care environments, as strategies proving highly effective in advanced contexts might yield diminished returns in regions with limited infrastructure and lower digital literacy.

Aims, Objectives, and Hypotheses

Overview

Based on these research gaps and theoretical considerations, our research aims to contribute to the existing literature on physician strategic self-disclosure behavior in OHCs by addressing 2 primary research objectives. First, we aim to investigate how physician self-disclosure breadth and depth are associated with patient decision-making in OHCs. Second, we seek to determine whether and how the regional DHL moderates the relationship between these self-disclosure strategies and patient choices.

Physician Self-Disclosure and Patient Decision-Making

Literature review in Section 1 in Multimedia Appendix 1 highlights that physicians' self-disclosure significantly shapes patients' credibility and trust toward decision-making by providing multifaceted professional information. Within OHCs, both breadth and depth dimensions of physician self-disclosure systematically activate these cognitive evaluations, ultimately shaping patient consultation decisions.

Self-disclosure breadth enhances patient decision-making by providing comprehensive professional signals that directly build credibility. When physicians disclose extensive information across multiple professional dimensions, they create a rich tapestry of verifiable cues that patients can cross-reference and validate. This comprehensive presentation first enhances perceived source credibility, as patients can observe concrete evidence of qualifications across diverse professional domains, reducing concerns about physician competence. The breadth of disclosure subsequently fosters interpersonal trust by signaling transparency and professional openness, suggesting that physicians have "nothing to hide" and are confident in their professional standing. Finally, this extensive information scope significantly increases perceived diagnostic value by providing patients with sufficient data points to make informed assessments about physician-patient compatibility. Patients can evaluate whether the physician's experience, training, and achievements align with their specific medical needs and preferences, thereby reducing decision-making uncertainty and increasing consultation likelihood.

XSL•FO
RenderX

Self-disclosure depth is associated with patient decision-making through intensive information quality that demonstrates specialized expertise and professional communication. When physicians provide detailed expertise descriptions, they signal profound clinical knowledge and commitment to patient understanding. This depth reinforces perceived source credibility by showcasing mastery within specific medical domains, as detailed explanations indicate genuine expertise rather than superficial knowledge. It also then builds interpersonal trust by demonstrating physicians' investment in clear communication and patient education, suggesting benevolent intentions and professional dedication. Most critically, depth maximizes perceived diagnostic value by enabling patients to precisely evaluate treatment fit—detailed specialty descriptions allow patients to determine whether their specific conditions fall within the physician's demonstrated areas of expertise. This granular matching capability reduces ambiguity about treatment appropriateness and increases patients' confidence in scheduling consultations with physicians.

DHL as the Moderator

DHL systematically shapes the mechanisms through which self-disclosure breadth and depth operate, fundamentally altering how the same self-disclosure content is produced, transmitted, and interpreted.

Advanced DHL amplifies the credibility-building effects of self-disclosure breadth through enhanced verification mechanisms and seamless information processing. When DHL is high, physicians can populate comprehensive profile fields with verifiable credentials, embed direct links to official registries, and present information through user-friendly interfaces that facilitate patient navigation. Patients in these contexts possess the digital literacy to efficiently cross-validate credentials through integrated databases and verification systems, creating a low-friction pathway for establishing source credibility. This enhanced verification capability strengthens the breadth-credibility relationship posited by self-disclosure theory, thereby accelerating patients' cognitive progression from enhanced source credibility to ultimate consultation decisions. In contrast, regions with limited digital infrastructure constrain verification processes, weakening the credibility signals that breadth disclosure would otherwise provide.

Similarly, advanced DHL intensifies the trust-building effects of self-disclosure depth by enabling rich multimedia presentations and sophisticated patient interpretation capabilities. High-level digital health care environments allow physicians to create comprehensive and well-structured disclosure experiences. Patients with elevated digital literacy can effectively parse these complex multimedia presentations, interpreting granular clinical details and structured expertise descriptions as authentic signals of both professional competence and patient-centered communication. This enhanced processing capability amplifies the depth-trust relationship, as patients can fully appreciate the nuanced expertise demonstrations that deep disclosure provides. Conversely, in regions with poor connectivity and limited digital literacy, deep disclosure content may fail to render properly or overwhelm patients' interpretive capacities, potentially undermining rather than enhancing the intended trust-building effects. On the basis of the preceding discussion, we advance the following two hypotheses:

- Hypothesis 1: Physician self-disclosure breadth and depth are positively associated with patient decision-making.
- Hypothesis 2: DHL positively moderates the relationship between self-disclosure depth, breadth, and patient decision-making.

Methods

Sample Size, Power, and Precision

We conducted a cross-sectional study that was designed and reported in accordance with the JARS (Journal Article Reporting Standards) guidelines [10] to examine the relationship between physician self-disclosure and patient acquisition in digital health markets. Our cross-sectional secondary analysis aimed to estimate the association between physicians' online self-disclosure and patient acquisitions with an absolute precision of ±0.5 percentage points at a 95% CI. The required number of physician profiles was calculated with the single-proportion formula sample size (n)=$[(Z_{(1-\alpha/2)})^2P(1-P)]/d^2$ [11-13], where $Z_{(1-\alpha/2)}$=the critical value with a corresponding standard level of confidence (1.96 at 95% CI), P was the conservative prevalence of 50% (as this value maximizes variance when the true prevalence is unknown and therefore yields the largest required sample size, ensuring adequate power and precision [14]), d=5% allowable margin of error or desired precision, indicating a minimum sample size of 384 physicians. To guard against unforeseen data-quality issues, we set a conservative target of at least 1000 evaluable records. Besides, this study is observational and contains no experimental aims, power calculations for between-group comparisons were unnecessary.

Data Collection

Haodf (established in 2006) is China's largest online health care platform, covering more than 10,000 hospitals and 900,000 physicians nationwide as of July 2023. Physician participation is exceptionally high, with 280,000 physicians registered under verified real names to deliver online consultation services, rendering the platform a unique and well-suited context for investigating physician–patient interactions [15].

Guided by the platform's interface, we deployed a crawler to systematically extract information from the public profiles of 2050 physicians from September to December 2024. After rigorous data cleaning and outlier removal, we followed the coding protocol established by Herzenstein et al [16] to quantify physicians' self-disclosure. Four trained research assistants independently coded fourteen dichotomous disclosure variables. The dataset was split into 2 equal batches, with each assigned to a distinct pair of coders who worked in parallel. Only observations with unanimous agreement were retained, and all cases of coder disagreement were excluded. This intercoder validation yielded 1798 reliable observations for subsequent empirical analysis, with no missing values in the final analytic dataset—comfortably exceeding our preregistered target of 1000 evaluable records.

XSL•FO
RenderX

Variable Measurements

Dependent Variable

We used "Total visits"—the cumulative number of consultations, calls, and bookings shown on each physician's profile—as the dependent variable. This variable, therefore, provides a comprehensive foundation for examining the relationship between physician self-disclosure and patient decision-making. All variable measurement details can be found in Table S1 in Section 2 of Multimedia Appendix 1.

Independent Variable

Self-disclosure breadth was indexed by entropic weighting of 3 public signals, namely clinical performance (cp), academic experience (ae), and social reputation (sr). This composite measure captures the comprehensiveness of physicians' self-presentation strategies by integrating these fundamental aspects of credible identity. Self-disclosure depth was also indexed by entropic weighting of 3 important cues, such as expertise coverage (ec), expertise richness (er), and expertise granularity (eg). These 3 dimensions collectively capture the multifaceted nature of disclosure depth, as physicians may vary in how extensively they elaborate (coverage), how comprehensively they describe (richness), and how specifically they detail their expertise (granularity). Specifically, self-disclosure breadth and depth were calculated using the following formulas:

$$\text{Self-disclosure breadth} = \log(\alpha_1 * cp + \alpha_2 * ae + \alpha_3 * sr + 1)$$

$$\text{Self-disclosure depth} = \log(\alpha_4 * ec + \alpha_5 * er + \alpha_6 * eg + 1)$$

Moderating Variable

DHL measures the extent of digital technology integration within a city's health care infrastructure, reflecting the adoption of telemedicine platforms and mobile clinical tools in routine medical practice [17]. We operationalized this variable using the China Urban Digital Economy Index (Medical Chapter) [18], which is the most up-to-date DHL-relevant index dataset we can find, as a comprehensive assessment jointly published by the School of Management at Zhejiang University and the Digital Economy Research Centre of New H3C Group. Based on this index, each city's DHL is assigned a score from 1 to 5, with higher numbers representing higher levels of digital integration and technological advancement in digital health care. We assigned each physician the DHL score corresponding to their practice location, thereby capturing the digital maturity of their local digital health care environment, as detailed in Table S2 in Section 2 of Multimedia Appendix 1.

Controls

To separate the associations of physician self-disclosure from other factors related to patient selection, we included three control variables—professional title (title), physician popularity (popularity), and gift (gift)—to account for alternative explanations of patient decision-making. A brief summary of variable measurements is presented in Table 1.

Table 1. Brief summary of measurement of core study variables.

Variables	Measurements
Dependent variable	
Total visits	The cumulative count of all patient-initiated interactions with each physician across all service channels (online consultations, telephone consultations, and appointment bookings), as recorded on the platform.
Independent variable	
Self-disclosure breadth	Coded as 1 if at least one clinical component (clinical experience, clinical effectiveness, and clinical manner) was disclosed, and 0 otherwise.
Clinical performance	Coded as 1 if at least one academic component (research productivity, international training, and educational credentials) was disclosed, and 0 otherwise.
Academic experience	Coded as 1 if at least one social component (part-time positions, honors and awards) was disclosed, and 0 otherwise.
Self-disclosure depth	
Expertise coverage	The total length of description text in vocabularies.
Expertise richness	The number of disease types explicitly mentioned as areas of expertise in the physician's profile.
Expertise granularity	Coded as 1 if the description uses delimiters to distinguish specialties, and 0 otherwise.
Moderator	
Digital health care level	Each physician is assigned a score from 1 to 5 corresponding to their practice location.
Controls	
Title	Chief physicians receive 4, deputy chief physicians 3, attending physicians 2, and resident physicians 1.
Popularity	A composite recommendation score (0-5 continuous scale) generated by the platform.
Gift	Patients' real payment to the physician after receiving services. Take the records on the platform.

XSL•FO
RenderX

Modeling and Statistical Analysis

Research Model

Figure 1 depicts our research model. This study examines the relationship between physicians' self-disclosure breadth and depth and patient decision-making in digital health markets (H1) and the moderating role of DHL in these relationships (H2).

Figure 1. Research model.

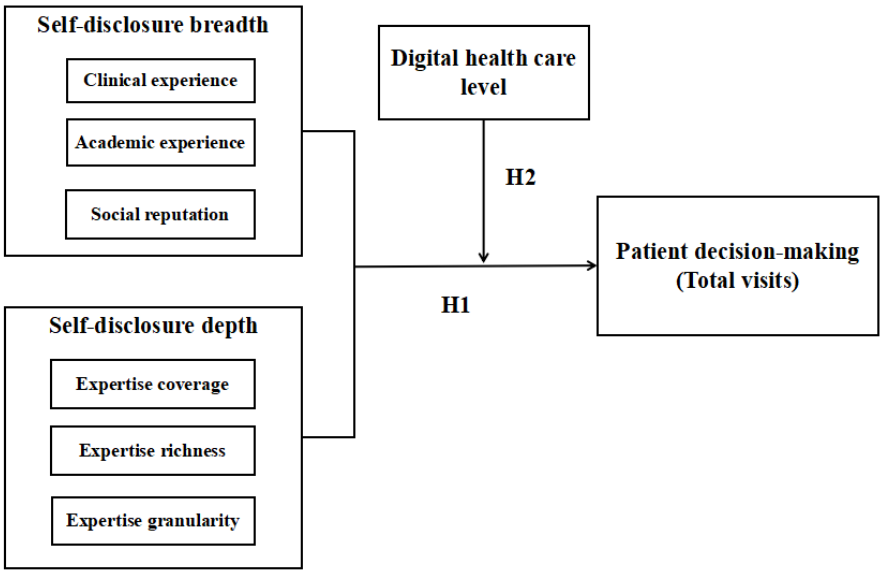

Ordinary Least Squares Model

Before testing the main and moderating effects, we first performed a descriptive statistical analysis to summarize the key characteristics of physicians and their self-disclosure behaviors. Subsequently, we conducted correlation analysis to assess the associations between key variables of interest. We then tested the main effects, focusing on how self-disclosure breadth and depth shape patient decision-making. Following this, we evaluated the moderating role of DHL, investigating whether regional digital health care conditions shape the effectiveness of these self-disclosure strategies. The empirical models pertaining to these tests were as follows:

- Model 1: Ln(Total visits) = β0 + β1Self-disclosure breadth + β2Title + β3Popularity + β4 Gift + ε
- Model 2: Ln(Total visits) = β0 + β1Self-disclosure depth + β2Title + β3Popularity + β4Gift + ε
- Model 3: Ln(Total visits) = β0 + β1Self-disclosure breadth + β2DHL + β3Self-disclosure breadth× DHL + β4Title + β5Popularity + β6Gift + ε
- Model 4: Ln(Total visits) = β0 + β1Self-disclosure depth + β2DHL + β3Self-disclosure depth × DHL + β4 Title + β5 Popularity + β6 Gift + ε

Our models include 3 control variables, including physician title, popularity metrics, and gift reception status, with ε representing the error term. We control for title because prior studies indicate that professional credentials significantly affect patient trust and provider selection [19]. Popularity serves as an important platform-based signal that may independently drive patient choices [20]. Gift reception reflects patient satisfaction from previous interactions, potentially influencing patients' future decision-making patterns [21].

We used ordinary least squares regression to estimate all models, with data analysis conducted in Stata (StataCorp LLC). To ensure comparability across variables with different scales, all variables were standardized before regression analysis.

Ethical Considerations

The Institutional Review Board of the International Business School, Beijing Foreign Studies University (001-2025-12-02) approved this study and granted an exemption from full human-subjects review. This study was conducted in accordance with the national ethical guidelines for research involving information data. Our use of legally obtained, fully anonymized public data, which contained no sensitive or commercial elements, qualified for an exemption from full ethics review as stipulated by the National Health Commission of the People's Republic of China [22]. No further approvals were required.

Researchers also confirmed that the original data-collection practices of Haodf are governed by the platform's user agreement and privacy notice, and private fields, such as physician names, clinical records, were anonymized and deidentified from the analytic file. No additional recruitment with physicians or patients occurred, no payments or incentives were offered to any physician or patient, and no identification of individual participants in any images of the manuscript or supplementary material is possible.

Results

Descriptive Statistics

Table 2 summarizes the central tendencies and dispersion of the key variables drawn from 1798 physician profiles on the Haodf platform. Min-Max denotes the actual minimum and maximum values in the sample. The IQR is calculated as the difference between the 75th and 25th percentiles and is reported

XSL•FO
RenderX

alongside the median (50th percentile). The results showed a skewed distribution of total visits per physician (mean 1471.83, SD 2486.38). In the subsequent data-processing pipeline, we applied appropriate normalization steps to mitigate the impact of extreme values and ensure the analytic dataset is well-behaved. Some physicians claim competence in clinical experience (mean 0.46, SD 0.50) and effectiveness (mean 0.32, SD 0.47), but only 4% (72/1798) explicitly mention clinical manner (mean 0.04, SD 0.20).

Table 2. Descriptive statistics of focal variables.

Variable	Mean (SD)	Median	Range
Total visits	1471.83 (2486.38)	680.5	1-27,352
Self-disclosure depth			
Expertise richness	3.76 (1.66)	4	0-8
Expertise coverage	8.16 (5.72)	7	0-58
Expertise granularity	0.51 (0.50)	1	0-1
Self-disclosure breadth			
Clinical experience	0.46 (0.50)	0	0-1
Clinical effectiveness	0.32 (0.47)	0	0-1
Clinical manner	0.04 (0.20)	0	0-1
Research productivity	0.75 (0.43)	1	0-1
International training	0.48 (0.50)	0	0-1
Educational credentials	0.60 (0.49)	1	0-1
Part-time positions	0.67 (0.47)	1	0-1
Honors and awards	0.44 (0.50)	0	0-1
Digital health care level	4.09 (1.01)	4	1-5
Title	3.35 (0.72)	3	1-4
Gift	128.65 (303.06)	39	0-4878
Popularity	4.14 (0.310)	4.1	3.4-5

Research productivity is reported by 75% (1349/1798) throughout the sample (mean 0.75, SD 0.43), whereas 48% (863/1798) list international training (mean 0.48, SD 0.50) and 60% (1079/1798) cite elite educational credentials (mean 0.60, SD 0.50). Roughly two-thirds hold part-time positions (mean 0.67, SD 0.47) and 44% (791/1798) have earned honors or awards (mean 0.44, SD 0.50). The average DHL is 4.09 (SD 1.01), and the mean title rank is 3.35 (SD 0.72), both approaching the upper end of their respective scales. Finally, popularity—an index computed by the platform—clusters tightly around 4.14 (SD 0.31), implying limited variance once the algorithmic score is normalized.

Pearson Correlation Analysis and Collinearity Testing

Table S3 in Section 2 of Multimedia Appendix 1 reports the Pearson correlations for 6 focal variables entering regression. Both self-disclosure depth and breadth are positively related to total visits (r=0.098; P<.001 and r=0.109; P<.001, respectively). The moderator DHL also positively relates to the dependent variable (r=0.119; P<.001). The VIF (variance inflation factor) analysis shows, afterwards in Table S4 in Section 2 of Multimedia Appendix 1, the largest value is 2.82 for gift, followed by 1.67 for popularity, whereas the remaining VIFs range only from 1.06 to 1.11. All VIF values are far below the conventional threshold of 5 (or 10) [23]. Taken together, the correlation matrix and the inflation factors jointly indicate that multicollinearity should not be a main concern for the stability or estimation of regression results.

Hypothesis Testing

We estimated 4 ordinary least squares models with heteroskedasticity-robust standard errors. The dependent variable, Total visits, was log-transformed to reduce skewness. All core continuous predictors, including self-disclosure breadth, self-disclosure depth, and DHL, were standardized (mean 0, SD 1) to facilitate coefficient comparability and to avert multicollinearity when interaction terms were introduced. Across all 4 models, the coefficients show the relative change in total visits associated with a one-standard-deviation shift in the focal variable. The results are reported in Table 3.

XSL•FO
RenderX

Table 3. Ordinary least squares regression results (Models 1-4) examining the relationship between physician self-disclosure breadth, depth, digital health care level, and patient decision-making.

Variable	Model 1	Model 2	Model 3	Model 4
Self-disclosure breadth, β (95% CI)	0.255[a] (0.054-0.456)	0.249[a] (0.047-0.450)	—[b]	—
Self-disclosure depth, β (95% CI)	—	—	0.098[c] (0.030-0.167)	0.092[c] (0.023-0.160)
DHL[d], β (95% CI)	—	–0.036 (–0.09 to 0.018)	—	–0.031 (–0.085 to 0.023)
Self-disclosure breadth×DHL, β (95% CI)	—	0.261[a] (0.061-0.461)	—	—
Self-disclosure depth×DHL, β (95% CI)	—	—	—	0.070[a] (0.002-0.138)
Title, β (95% CI)	0.248[c] (0.172-0.324)	0.242[c] (0.166-0.317)	0.276[c] (0.201-0.350)	0.275[c] (0.201-0.349)
Popularity, β (95% CI)	1.720[c] (1.50-1.94)	1.751[c] (1.532-1.970)	1.678[c] (1.460-1.900)	1.700[c] (1.480-1.921)
Gift, β (95% CI)	0.001[c] (0.001-0.001)	0.001[c] (0.001-0.001)	0.001[c] (0.001-0.001)	0.001[c] (0.001-001)
Constant, β (95% CI)	–1.825[c] (–2.720 to –0.930)	–1.938[c] (–2.843 to –1.031)	–1.748[c] (–2.648 to –0.847)	–1.838[c] (–2.749 to –0.926)
F test (*df*)	251.3 (6)	181.4 (8)	251.9 (6)	181.1 (8)
R^2	0.412	0.415	0.413	0.415

[a]The correlation is significant at a significance level of .05 (2-tailed).

[b]Not applicable.

[c]The correlation is significant at a significance level of .01 (2-tailed).

[d]DHL: digital health care level.

Model 1 establishes the baseline relationship for self-disclosure breadth, revealing a positive and statistically significant coefficient (β=0.255, 95% CI 0.054-0.456; *P*=.01), thereby confirming Hypothesis 1 that broader physician self-disclosure increases patient decision-making volume. Model 3 demonstrates a parallel finding for self-disclosure depth, with results showing a significant positive effect (β=0.098, 95% CI 0.030-0.167; *P*=.005), providing support for Hypothesis 1 that physician self-disclosure depth is positively associated with patient decision-making. Both findings confirm that comprehensive information disclosure, whether through diverse disclosure topics or detailed expertise presentation, enhances physician attractiveness to patients.

The interaction analyses reveal that DHL significantly amplifies self-disclosure effectiveness. Model 2 introduces the breadth×DHL interaction term, yielding a positive and significant coefficient (β=0.261, 95% CI 0.061-0.461; *P*=.01), which supports Hypothesis 2 that DHL strengthens the relationship between self-disclosure breadth and patient decision-making. Similarly, Model 4 demonstrates that the depth×DHL interaction is positive and significant (β=0.070, 95% CI 0.002-0.138; *P*=.045), corroborating Hypothesis 2 that DHL enhances the effectiveness of disclosure depth. Notably, the breadth interaction effect is substantially larger than the depth interaction effect, suggesting that DHL provides greater amplification benefits for diverse disclosure strategies compared to detailed expertise presentation.

Post Hoc Analysis

To better understand the nuanced mechanisms underlying the moderating effects of DHL, we conducted post hoc analysis using interaction plots. While our main regression results demonstrate statistically significant moderation effects, visualizing these interactions provides deeper insights into how DHL moderates the strength and nature of the relationships between physician self-disclosure strategies and patient decision-making. The visualized graphs are presented in Figures 2 and 3, where high and low DHL correspond to values of 1 SD above and below the mean (+1 and –1).

XSL•FO
RenderX

Figure 2. The moderating effect of digital health care level on the association between physician self-disclosure breadth and patient decision-making. DHL: digital health care level.

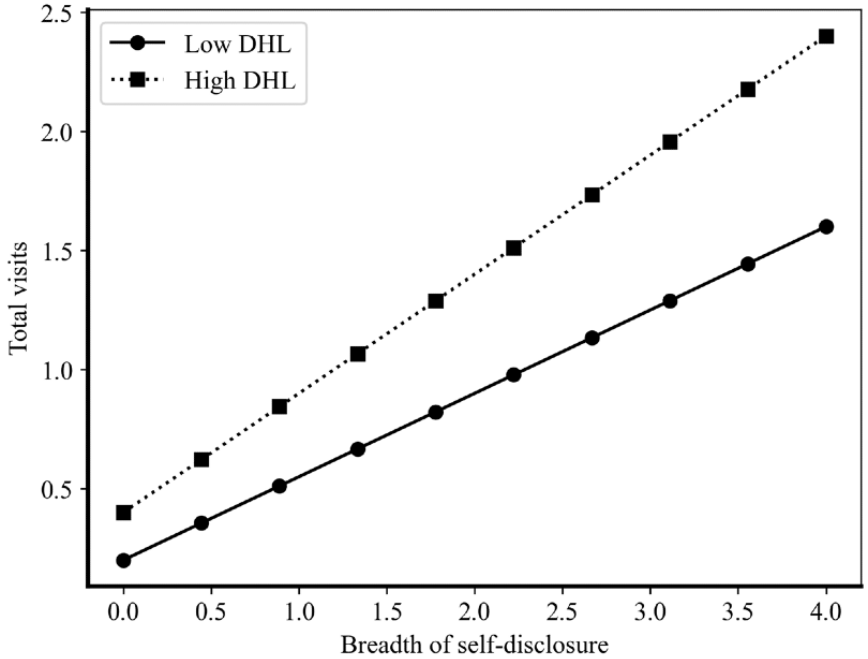

Figure 3. The moderating effect of digital health care level on the association between physician self-disclosure depth and patient decision-making. DHL: digital hhealth care level.

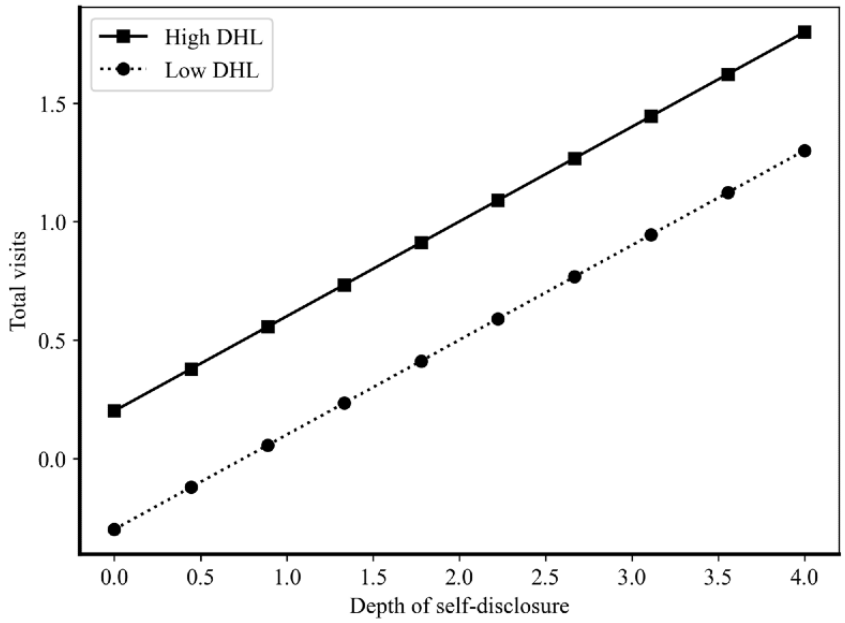

Figure 2 reveals particularly prominent insights about breadth moderation. Most notably, the gap between high and low DHL conditions grows increasingly larger as self-disclosure breadth increases, creating a pronounced divergence pattern that suggests breadth disclosure may be more sensitive to DHL than depth disclosure. In high DHL environments, patients appear exceptionally responsive to broad professional presentations. This amplified responsiveness may reflect enhanced information processing capabilities and greater appreciation for diverse professional information. Conversely, in low DHL environments, increasing breadth yields relatively restricted returns, potentially due to limited digital literacy or infrastructure constraints that prevent effective usage of comprehensive professional profiles. The steeper moderation gap for breadth compared to depth suggests that while diverse disclosure topics (breadth) become disproportionately valuable when supported by higher DHL, detailed expertise presentations (depth) maintain more consistent effectiveness across different DHL contexts. This creates a “digital divide” effect where technological advancement in health care becomes a critical prerequisite for breadth disclosure effectiveness.

Figure 3 provides compelling visual evidence for the DHL moderation effect on self-disclosure depth, revealing a striking divergence in self-disclosure effectiveness across DHL levels. In high DHL environments, the relationship between

XSL•FO
RenderX

self-disclosure depth and total visits exhibits a pronounced higher slope, demonstrating that each incremental increase in depth disclosure generates greater patient engagement gains. Conversely, in low DHL contexts, the relationship remains relatively lower, suggesting that detailed expertise presentations yield fewer additional benefits when digital health care is underdeveloped. This pattern indicates that physicians practicing in technologically advanced health care environments can leverage detailed self-disclosure strategies—spanning specialized medical capabilities, in-depth professional competencies, and granular clinical expertise—to achieve disproportionately greater patient attraction, while those in less digitized environments may find such detailed disclosure strategies less rewarding.

Robustness Checks

We used various methods for robustness checks to ensure the reliability and consistency of our research findings, including variable substitution, bootstrap resampling, subsampling, and the winsorizing technique. The core findings are summarized in Table 4 below. The procedures and interpretations for each approach are detailed below.

To ensure the reliability of our findings, we first conducted a robustness check by replacing our categorical DHL measurement with continuous digital health care scores, which were also presented in the report published by China Urban Digital Economy Index (Medical Chapter) [18], as shown in Table S5 of Section 3 in Multimedia Appendix 1. The detailed results of this first variable substitution check are shown immediately afterwards in Table S6, demonstrating consistent patterns across all model specifications. Self-disclosure breadth continues to be positively associated with total visits across Models 1 and 2 (β=0.255, 95% CI 0.054-0.456; *P*=.01 and β=0.258, 95% CI 0.057-0.459; *P*=.01, respectively), while self-disclosure depth demonstrates similar consistency in Models 3 and 4 (β=0.098, 95% CI 0.030-0.167; *P*=.005 and β=0.096, 95% CI 0.027-0.164; *P*=.006, respectively). The preservation of both effect magnitudes and significance levels indicates that our core findings are not artifacts of the initial categorical operationalization.

To rule out the possibility that physicians attract more consultations simply because they are more active or popular, we further re-estimated all models by replacing the dependent variable with Ln (total visits-per-popularity, ie, total visits divided by popularity), where Popularity is a platform-computed index that combines physician activity, patient recommendations, and review ratings to reflect overall physician popularity. Table S7 shows that the core disclosure variables remain positively significant across the 4 specifications, with Models 1 and 2 (β=0.494, 95% CI 0.290-0.699; *P*<.001 and β=0.482, 95% CI 0.277-0.687; *P*<.001, respectively); and Models 3 and 4 (β=0.157, 95% CI 0.086-0.228; *P*<.001 and β=0.151, 95% CI 0.080-0.222; *P*<.001, respectively). The persistent significance of self-disclosure after controlling for popularity corroborates our main results and demonstrates that, even after adjusting for physicians' baseline visibility and popularity, patients show a preference for physicians who provide greater breadth and depth of professional information.

To validate that our significance tests are not dependent on distributional assumptions, we re-estimated all models using bootstrap resampling with 1000 replications. Table S8 presents the bootstrapped results, demonstrating the robustness of our statistical inferences. First, self-disclosure breadth maintains its positive and significant relationship with total visits in both Models 1 and 2 (β=0.255, 95% CI 0.050-0.460; *P*=.02 and β=0.249, 95% CI 0.049-0.448; *P*=.02, respectively), while self-disclosure depth demonstrates identical significance patterns in Models 3 and 4 (β=0.098, 95% CI 0.294-0.167; *P*=.005 and β=0.092, 95% CI 0.026-0.258; *P*=.006, respectively). Besides, the moderation effects remain significant under bootstrap estimation. The breadth×DHL interaction retains its positive and significant coefficient (β=0.261, 95% CI 0.054-0.467; *P*=.01) in Model 2, while the depth×DHL interaction similarly maintains significance (β=0.070, 95% CI 0.007-0.133; *P*=.03) in Model 4.

To address potential concerns that our main findings might be driven by senior physicians who possess inherently greater credibility and resources, we conduct a robustness check by restricting our analysis to non–chief physicians only. As shown in Table S9, the detailed results of subsample analysis (non–chief physicians) show consistent effects across the restricted sample, confirming the robustness of our main findings. Self-disclosure breadth maintains its positive and significant relationship with total visits in both Models 1 and 2 (β=0.345, 95% CI 0.074-0.615; *P*=.01 and β=0.345, 95% CI 0.075-0.614; *P*=.01 respectively), while self-disclosure depth similarly shows robust positive effects in Models 3 and 4 (β=0.125, 95% CI 0.032-0.219; *P*=.009 and β=0.124, 95% CI 0.029-0.215; *P*=.01, respectively). Notably, the coefficient magnitudes are actually larger in this subsample compared to the full sample, suggesting that self-disclosure strategies may be even more crucial for physicians with lower hierarchies.

To address potential concerns about extreme values influencing our results, we re-estimated all models after winsorizing the top and bottom 10% of each variable. Table S10 presents the detailed results from this outlier-treatment approach. Self-disclosure breadth maintains its positive and significant effects across Models 1 and 2 (β=0.250, 95% CI 0.030-0.470; *P*=.03 and β=0.248, 95% CI 0.028-0.468; *P*=.03, respectively), while self-disclosure depth similarly preserves its significant positive relationship in Models 3 and 4 (β=0.106, 95% CI 0.032-0.180; *P*=.005 and β=0.101, 95% CI 0.027-0.175; *P*=.008, respectively). The coefficient magnitudes remain virtually identical to our original estimates, confirming that extreme values do not drive the main effect conclusions.

XSL•FO
RenderX

Table 4. A brief summary of the focal results of our 5 robustness tests.

Robust check	Self-disclosure breadth, β (95% CI)	Self-disclosure depth, β (95% CI)	Self-disclosure breadth×DHL[a], β (95% CI)	Self-disclosure depth×DHL, β (95% CI)
R1[b]				
Model 1	0.255[c] (0.054-0.456)	—[d]	—	—
Model 2	0.258[c] (0.057-0.459)	—	0.020[e] (0.006-0.034)	—
Model 3	—[c]	0.098[e] (0.030-0.167)	—	—
Model 4	—	0.096[e] (0.027-0.164)	—	0.006[c] (0.001-0.010)
R2[f]				
Model 1	0.494[e] (0.290-0.699)	—	—	—
Model 2	0.482[e] (0.277-0.687)	—	0.030[g] (0.002-0.063)	—
Model 3	—	0.157[e] (0.086-0.228)	—	—
Model 4	—	0.151[e] (0.080-0.222)	—	0.071[g] (0.001-0.142)
R3[h]				
Model 1	0.255[c] (0.050-0.460)	—	—	—
Model 2	0.249[c] (0.049-0.448)	—	0.261[c] (0.054-0.467)	—
Model 3	—	0.098[e] (0.029-0.167)	—	—
Model 4	—	0.092[e] (0.026-0.258)	—	0.070[c] (0.007-0.133)
R4[i]				
Model 1	0.345[c] (0.074-0.615)	—	—	—
Model 2	0.345[c] (0.075-0.614)	—	0.259[c] (0.001-0.518)	—
Model 3	—	0.125[e] (0.032-0.219)	—	—
Model 4	—	0.124[c] (0.029-0.215)	—	0.086[c] (0.001-0.173)
R5[j]				
Model 1	0.250[c] (0.030-0.470)	—	—	—

XSL•FO
RenderX

Robust check	Self-disclosure breadth, β (95% CI)	Self-disclosure depth, β (95% CI)	Self-disclosure breadth×DHL[a], β (95% CI)	Self-disclosure depth×DHL, β (95% CI)
Model 2	0.248[c] (0.028-0.468)	—	0.285[c] (0.066-0.504)	—
Model 3	—	0.106[e] (0.032-0.180)	—	—
Model 4	—	0.101[e] (0.027-0.175)	—	0.076[c] (0.003-0.149)

[a]DHL: digital health care level.

[b]Replace categorical DHL measurement with continuous digital health care scores.

[c]The correlation is significant at a significance level of .05 (2-tailed).

[d]Not applicable.

[e]The correlation is significant at a significance level of .01 (2-tailed).

[f]Replace the dependent variable with Log_e (total visits-per-popularity).

[g]The correlation is significant at a significance level of .1 (2-tailed).

[h]Re-estimated all models using bootstrap resampling with 1000 replications.

[i]Restricted our analysis to non–chief physicians only.

[j]Re-estimated all models after winsorizing the top and bottom 10% of each variable.

Discussion

Principal Findings

This study demonstrates that physicians can strategically shape patient decision-making through purposeful self-disclosure behaviors within OHCs. Our empirical analysis from China's leading online health platform reveals that both self-disclosure breadth and depth significantly increase patient visits, challenging the prevailing view of physicians as passive recipients of online reviews. Most critically, the regional DHL fundamentally moderates these relationships. In cities with advanced digital development, both breadth and depth effects are substantially amplified, while regions with limited digital development show diminished returns for the same disclosure strategies. These findings reconceptualize physician agency within digital health care platforms, demonstrating that strategic self-disclosure represents a viable patient acquisition mechanism rather than passive information provision.

Theoretical Implications

The study advances theoretical understanding of physician self-disclosure in digital health markets by embedding its 3 key contributions within—and explicitly contrasting them against—the extant literatures on physician agency, self-disclosure theory, and OHCs.

First, contrary to prior studies that predominantly frame physicians as passive recipients of online feedback [4,24,25], our reconceptualization of physician agency aligns with a growing but still underdeveloped stream of research that emphasizes the proactive role of service providers in digital platforms. For instance, recent work by Ouyang and Wang [26] and Lu and Wu [27] suggests that physicians can influence patient perceptions through profile customization and online engagement. This challenges the prevailing passive-physician paradigm in medical marketing research and establishes physicians as active agents capable of patient acquisition through evidence-based self-presentation strategies [28,29]. However, extant research has predominantly conceptualized online reputation management as a reactive endeavor—centering on prompt responses to negative feedback [30], remediation of service failures [31], or post hoc optimization of profile completeness [32]—while overlooking proactive self-disclosure breadth and depth as ex ante instruments that physicians can strategically deploy to attract new patients before any review is written. Our findings go beyond these accounts by revealing that physicians can deploy proactive, ex ante self-disclosure—systematically foregrounding personal credentials, institutional affiliations, and succinct expertise narratives—to pre-emptively sculpt patient trust and choice before any review is written or any service failure occurs. This reorients the theoretical lens from "damage control" to "impression engineering," recasting self-disclosure as a forward-looking signaling mechanism that anticipates patient heuristics rather than remedying prior dissatisfaction. By establishing strategic disclosure as a feasible alternative to reactive reputation management, we push the conceptual boundary of physician behavior in digital environments beyond traditional service-recovery frameworks and toward a predictive, marketing-as-signal paradigm.

Second, we extend self-disclosure theory beyond its traditional interpersonal communication context into health care settings while establishing critical boundary conditions [7,33,34]. We introduce the regional DHL as a critical boundary condition—a factor largely overlooked in prior self-disclosure research. While earlier studies have examined individual-level moderators, such as gender [35] or cultural orientation [36], our findings reveal that macrolevel technological development significantly moderates the disclosure-outcome relationship. This aligns with recent macrosociological perspectives on digital inequality [37], which argue that the same online behavior can yield divergent outcomes depending on the technological context in which it

XSL•FO
RenderX

is embedded. Thus, our study not only extends self-disclosure theory into a new domain but also redefines its boundary conditions by incorporating sociotechnological contingencies. Our findings confirm that self-disclosure breadth and depth operate in professional contexts, yet their effectiveness is contingent on the regional DHL. By demonstrating this macrolevel technological moderation, we extend the theory's scope and expose context boundaries previously overlooked [38,39].

Third, the prevailing view in the OHC literature maintains that these platforms mainly mitigate physician-patient information asymmetry by aggregating ratings, reviews, and outcome data [24,40,41]. We extend this perspective by demonstrating that physicians can actively reconfigure the information environment through strategic self-disclosure. Especially, our findings show that, within Chinese OHCs, breadth of disclosure markedly outweighs depth, as an expansive array of credential signals consistently exhibits a greater positive association with patient engagement than rich narrative detail does. Patients appear to follow a hierarchical signaling model in which credential heuristics operate as an initial, low-cognitive gatekeeper, with hospital tier, academic rank, and prestigious awards are rapidly recoded into a binary "pass-fail" filter that determines inclusion in the consideration set. Only after clearing this threshold do patients allocate scarce attentional resources to elaborately process narrative depth—articulations of treatment philosophy, detailed case histories, or other discursive evidence of clinical expertise. Empirically, the stronger main effect of self-disclosure breadth compared to depth evidences that, in the context of Chinese OHCs, patients may initially rely more heavily on credential heuristics—such as titles, affiliations, and awards [42,43]—as efficient signals of quality, before engaging with the more cognitively demanding narratives of detailed expertise. This insight complements and extends recent work by Wang et al [44], who find that physicians with more comprehensive profiles receive more appointment requests, thereby offering a mechanistic account of why breadth outperforms depth in Chinese OHCs.

Practical Implications

This study also provides several practical implications for multiple stakeholder groups seeking to enhance physician self-disclosure effectiveness and patient decision-making within OHCs.

For physicians, our findings emphasize a context-dependent approach to online self-disclosure [45-47]. Those in high-DHL regions should leverage comprehensive breadth—showcasing credentials across clinical, academic, and social domains—and enrich their profiles with multimedia depth, such as video introductions. In low-DHL regions, the priority shifts to maximizing clarity. Physicians should focus on core breadth elements like clinical experience and use concise, text-based depth with scannable lists of expertise to ensure accessibility.

For OHC platform designers, our results highlight the limitation of a uniform profile design and advocate for context-aware systems [48,49]. Actionable recommendations include implementing structured disclosure templates to guide physicians in highlighting decision-critical information, introducing visual verification badges for credentials to enhance trust, and developing tiered interface modes—a feature-rich version for high-DHL users and a streamlined, text-optimized version for regions with limited DHL.

For policymakers, our evidence on the moderating role of DHL underscores that digital infrastructure is a social determinant of health access. Specific interventions should prioritize closing the digital divide by investing in high-speed internet infrastructure in underserved areas, launching public health campaigns to improve patients' digital health literacy, and creating financial incentives for clinics and physicians in low-DHL regions to adopt and master digital consultation tools.

Limitations

This study acknowledges certain limitations. First, the cross-sectional design limits causal inference, leaving the temporal dynamics of these relationships unclear. Second, the DHL measurement uses a city-level categorization rather than granular technological indicators, which prevents identifying the specific infrastructure components that most critically moderate disclosure effectiveness. Third, owing to the absence of more recent publicly available datasets, the DHL indicators used in this study remain those published in 2021. The 3-year lag may introduce slight discrepancies with present-day infrastructure levels. Updated figures can further corroborate our findings. Finally, focus on China's online health platform may limit generalizability across different cultural and regulatory contexts. The findings may reflect sociocultural norms specific to Chinese health care markets, such as a pronounced hierarchy in physician-patient dynamics or a strong preference for credential-based trust signals.

Conclusions

This study is innovative in reconceptualizing physicians as strategic agents capable of actively shaping patient decision-making through purposeful self-disclosure in digital health markets. Different from existing studies that treat physicians as passive recipients of online ratings and reviews, our research demonstrates that physicians can strategically shape patient acquisition through deliberate self-disclosure breadth and depth. This study thus brings new insights to digital health markets by demonstrating that self-disclosure operates as a viable patient acquisition mechanism in professional health care contexts, wherein DHL acts as a critical boundary condition that fundamentally moderates the breadth-depth relationships. The findings have significant implications in the real world: (1) physicians can leverage evidence-based disclosure strategies for patient acquisition, (2) platform designers should implement context-adaptive features optimizing effectiveness across heterogeneous digital environments, and (3) policymakers should prioritize digital infrastructure investments to systematically enhance physicians' competitive capabilities and patient decision-making quality.

XSL•FO
RenderX

Acknowledgments

The authors gratefully acknowledge the guidance received from the editor and anonymous reviewers.

Funding

This study was supported by the Humanities and Social Science Fund of Ministry of Education of China (grant 22YJA630018); the Fundamental Research Funds for the Central Universities (grants 2022JJ007 and 2025ZZ022).

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: QL
Methodology: QL
Data curation: PY
Formal analysis: PY
Software: PY
Visualization: PY
Writing–original draft: QL
Writing–review & editing: QL, PY
Funding acquisition: QL, JF
Resources: JF
Supervision: JF

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details on literature review, variable measurements and robustness check results.
[DOCX File , 114 KB - jmir_v28i1e84963_app1.docx]

References

1. El-Tallawy SN, Pergolizzi JV, Vasiliu-Feltes I, Ahmed RS, LeQuang JK, Alzahrani T, et al. Innovative applications of telemedicine and other digital health solutions in pain management: a literature review. Pain Ther 2024;13(4):791-812. [doi: 10.1007/s40122-024-00620-7] [Medline: 38869690]
2. Digital health industry data book: Telehealth care, mHealth, health care analytics, and digital health systems—market size, share, trend analysis, and segment forecasts, 2022–2030. Grand View Research. 2022. URL: https://www.grandviewresearch.com/sector-report/digital-health-industry-data-book [accessed 2025-04-21]
3. Peng Y, Yin P, Deng Z, Wang R. Patient-physician interaction and trust in online health community: the role of perceived usefulness of health information and services. Int J Environ Res Public Health 2019;17(1):139 [FREE Full text] [doi: 10.3390/ijerph17010139] [Medline: 31878145]
4. Wu DC, Zhao X, Wu J. Online physician-patient interaction and patient satisfaction: empirical study of the internet hospital service. J Med Internet Res 2023;25:e39089 [FREE Full text] [doi: 10.2196/39089] [Medline: 37616031]
5. Fitzpatrick PJ. Improving health literacy using the power of digital communications to achieve better health outcomes for patients and practitioners. Front Digit Health 2023;5:1264780 [FREE Full text] [doi: 10.3389/fdgth.2023.1264780] [Medline: 38046643]
6. Ruangkanjanases A, Sivarak O, Jong D, Zhou Y. The effect of self-disclosure on mass trust through TikTok: an empirical study of short video streaming application users. Front Psychol 2022;13:968558 [FREE Full text] [doi: 10.3389/fpsyg.2022.968558] [Medline: 36059732]
7. Alrabiah S. Urrgh.. Will They Stop Going on About Their Relationships?: An Examination of Self-Disclosure by Travel Influencers on Consumer Outcomes. Edinburgh, Scotland: The University of Edinburgh; 2021.
8. Sundell E, Wångdahl J, Grauman Å. Health literacy and digital health information-seeking behavior - a cross-sectional study among highly educated Swedes. BMC Public Health 2022;22(1):2278 [FREE Full text] [doi: 10.1186/s12889-022-14751-z] [Medline: 36471284]
9. Saeed SA, Masters RM. Disparities in health care and the digital divide. Curr Psychiatry Rep 2021;23(9):61 [FREE Full text] [doi: 10.1007/s11920-021-01274-4] [Medline: 34297202]

XSL•FO
RenderX

10. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: the APA publications and communications board task force report. Am Psychol 2018;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
11. Sadiq IZ, Usman A, Muhammad A, Ahmad KH. Sample size calculation in biomedical, clinical and biological sciences research. J.Umm Al-Qura Univ. Appll. Sci 2024;11(1):133-141. [doi: 10.1007/s43994-024-00153-x]
12. Yu M, Shi X, Zou B, An S. [Comparison of 7 methods for sample size determination based on confidence interval estimation for a single proportion]. Nan Fang Yi Ke Da Xue Xue Bao 2023;43(1):105-110 [FREE Full text] [doi: 10.12122/j.issn.1673-4254.2023.01.14] [Medline: 36856217]
13. Li Y, Cen J, Wu J, Tang M, Guo J, Hang J, et al. The degree of anxiety and depression in patients with cardiovascular diseases as assessed using a mobile app: cross-sectional study. J Med Internet Res 2023;25:e48750 [FREE Full text] [doi: 10.2196/48750] [Medline: 37792455]
14. Zhang C, Zhou Y, Ke Y, Zhang W, Qin M, Alias H, et al. Fertility intentions in the era of the new three-child policy in China: a cross-sectional survey of married adults of reproductive age. Front Public Health 2025;13:1674687 [FREE Full text] [doi: 10.3389/fpubh.2025.1674687] [Medline: 41323596]
15. Li Y, Yan X, Song X. Provision of paid web-based medical consultation in China: cross-sectional analysis of data from a medical consultation website. J Med Internet Res 2019;21(6):e12126 [FREE Full text] [doi: 10.2196/12126] [Medline: 31162129]
16. Herzenstein M, Dholakia UM, Andrews RL. Strategic herding behavior in peer-to-peer loan auctions. J Interact Mark 2011;25(1):27-36. [doi: 10.1016/j.intmar.2010.07.001]
17. Yao L, Li Q, Li Q, Wang T, Peng S, Fu X, et al. Factors influencing the adoption of telemedicine services among middle-aged and older patients with chronic conditions in rural China: a multicentre cross-sectional study. BMC Health Serv Res 2025;25(1):775 [FREE Full text] [doi: 10.1186/s12913-025-12931-2] [Medline: 40448164]
18. Zhejiang University, New H3C Group. China Urban Digital Economy Index (Medical Chapter). URL: https://roadshow.h3c.com/zl/pdf/ylbps.pdf [accessed 2025-11-15]
19. Yu X, Wang H, Chen Z. The role of user-generated content in the sustainable development of online healthcare communities: exploring the moderating influence of signals. Sustainability 2024;16(9):3739. [doi: 10.3390/su16093739]
20. Yan J, Liang C, Gu D, Zhu K, Zhou P. Understanding patients' doctor choice behavior: elaboration likelihood perspective. Inf Dev 2024:02666669241259126. [doi: 10.1177/02666669241259126]
21. Wang Y, Wu H, Xia C, Lu N. Impact of the price of gifts from patients on physicians? service quality in online consultations: empirical study based on social exchange theory. J Med Internet Res 2019;22(5):e15685. [doi: 10.2196/15685]
22. Notice on issuing the measures for ethical review of life science and medical research involving human subjects. National Health Commission. 2025. URL: https://www.nhc.gov.cn/qjjys/c100016/202302/6b6e447b3edc4338856c9a652a85f44b.shtml [accessed 2025-12-11]
23. Kong M, Wang Y, Li M, Yao Z. Mechanism assessment of physician discourse strategies and patient consultation behaviors on online health platforms: mixed methods study. J Med Internet Res 2025;27:e54516 [FREE Full text] [doi: 10.2196/54516] [Medline: 40106798]
24. Song M, Elson J, Bastola D. Digital age transformation in patient-physician communication: 25-year narrative review (1999-2023). J Med Internet Res 2025;27:e60512 [FREE Full text] [doi: 10.2196/60512] [Medline: 39819592]
25. Guetz B, Bidmon S. The credibility of physician rating websites: a systematic literature review. Health Policy 2023;132:104821 [FREE Full text] [doi: 10.1016/j.healthpol.2023.104821] [Medline: 37084700]
26. Ouyang P, Wang J. Physician's online image and patient's choice in the online health community. Internet Res 2022;32(6):1952-1977. [doi: 10.1108/intr-04-2021-0251]
27. Lu W, Wu H. How online reviews and services affect physician outpatient visits: content analysis of evidence from two online health care communities. JMIR Med Inform 2019;7(4):e16185 [FREE Full text] [doi: 10.2196/16185] [Medline: 31789597]
28. Ouyang P, Wang J, Jasmine Chang A. Patients need emotional support: mmanaging physician disclosure information to attract more patients. Int J Med Inform 2022;158:104674. [doi: 10.1016/j.ijmedinf.2021.104674] [Medline: 34968960]
29. Duan Y, Gong S. Active medicine and passive medicine. Res Sq 2025:1-36. [doi: 10.13140/RG.2.2.22021.87527]
30. Han X, Lin Y, Han W, Liao K, Mei K. Effect of negative online reviews and physician responses on health consumers' choice: experimental study. J Med Internet Res 2024;26:e46713 [FREE Full text] [doi: 10.2196/46713] [Medline: 38470465]
31. van der Schyff EL, Ridout B, Amon KL, Forsyth R, Campbell AJ. Providing self-led mental health support through an artificial intelligence-powered chat bot (leora) to meet the demand of mental health care. J Med Internet Res 2023;25:e46448 [FREE Full text] [doi: 10.2196/46448] [Medline: 37335608]
32. Burke LE, Shiffman S, Music E, Styn MA, Kriska A, Smailagic A, et al. Ecological momentary assessment in behavioral research: addressing technological and human participant challenges. J Med Internet Res 2017;19(3):e77 [FREE Full text] [doi: 10.2196/jmir.7138] [Medline: 28298264]
33. Jo E, Jeong Y, Park S, Epstein D, Kim Y. Understanding the Impact of Long-Term Memory on Self-Disclosure with Large Language Model-Driven Chatbots for Public Health Intervention. 2024 Presented at: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems; May 11-16, 2024; New York, NY p. 1-21. [doi: 10.1145/3613904.3642420]

XSL•FO
RenderX

34. Kadji K, Schmid Mast M. The effect of physician self-disclosure on patient self-disclosure and patient perceptions of the physician. Patient Educ Couns 2021;104(9):2224-2231 [FREE Full text] [doi: 10.1016/j.pec.2021.02.030] [Medline: 33775499]
35. Nowell C, Pfeifer JH, Enticott P, Silk T, Vijayakumar N. Value of self-disclosure to parents and peers during adolescence. J Res Adolesc 2023;33(1):289-301. [doi: 10.1111/jora.12803] [Medline: 36166491]
36. Koch A, Rabins M, Messina J, Brennan-Cook J. Exploring the challenges of sexual orientation disclosure among lesbian, gay, bisexual, transgender, queer individuals. J Nurse Pract 2023;19(10):104765. [doi: 10.1016/j.nurpra.2023.104765]
37. Wu M, Xue Y, Ma C. The association between the digital divide and health inequalities among older adults in China: nationally representative cross-sectional survey. J Med Internet Res 2025;27:e62645 [FREE Full text] [doi: 10.2196/62645] [Medline: 39813666]
38. Liu J, He J, He S, Li C, Yu C, Li Q. Patients' self-disclosure positively influences the establishment of patients' trust in physicians: an empirical study of computer-mediated communication in an online health community. Front Public Health 2022;10:823692 [FREE Full text] [doi: 10.3389/fpubh.2022.823692] [Medline: 35145943]
39. Steele Gray C, Ramachandran M, Brinton C, Forte M, Loganathan M, Walsh R, et al. Digitally mediated relationships: how social representation in technology influences the therapeutic relationship in primary care. Soc Sci Med 2024;353:116962 [FREE Full text] [doi: 10.1016/j.socscimed.2024.116962] [Medline: 38908092]
40. Li H. Asymmetric information in the field of healthcare. Adv Econ Manag Polit Sci 2024;100(1):28-34. [doi: 10.54254/2754-1169/100/20241089]
41. Guo S, Wang K, Yang L, Dang Y. Extending signaling theory in online health communities to address medical information asymmetry: systematic review with narrative synthesis. J Med Internet Res 2025;27:e73208 [FREE Full text] [doi: 10.2196/73208] [Medline: 40749179]
42. Daraz L, Dogu C, Houde V, Bouseh S, Morshed KG. Assessing credibility: quality criteria for patients, caregivers, and the public in online health information-a qualitative study. J Patient Exp 2024;11:23743735241259440 [FREE Full text] [doi: 10.1177/23743735241259440] [Medline: 38827225]
43. van Velsen L, Flierman I, Tabak M. The formation of patient trust and its transference to online health services: the case of a Dutch online patient portal for rehabilitation care. BMC Med Inform Decis Mak 2021;21(1):188 [FREE Full text] [doi: 10.1186/s12911-021-01552-4] [Medline: 34118919]
44. Wang H, Jin J, Li L, Liu J, Wang D. Driving online healthcare growth amid the digital divide: how trust in professional signals from doctor biographies shapes patient decisions. Healthcare (Basel) 2025;13(12):1418 [FREE Full text] [doi: 10.3390/healthcare13121418] [Medline: 40565445]
45. Yuchao W, Ying Z, Liao Z. Health privacy information self-disclosure in online health community. Front Public Health 2020;8:602792 [FREE Full text] [doi: 10.3389/fpubh.2020.602792] [Medline: 33614566]
46. González-Pérez A, Matey-Sanz M, Granell C, Díaz-Sanahuja L, Bretón-López J, Casteleyn S. AwarNS: A framework for developing context-aware reactive mobile applications for health and mental health. J Biomed Inform 2023;141:104359 [FREE Full text] [doi: 10.1016/j.jbi.2023.104359] [Medline: 37044134]
47. Wyant K, Moshontz H, Ward SB, Fronk GE, Curtin JJ. Acceptability of personal sensing among people with alcohol use disorder: observational study. JMIR Mhealth Uhealth 2023;11:e41833 [FREE Full text] [doi: 10.2196/41833] [Medline: 37639300]
48. Tai-Seale M, Cheung M, Vaida F, Ruo B, Walker A, Rosen RL, et al. Patient-clinician communication interventions across multiple primary care sites: a cluster randomized clinical trial. JAMA Health Forum 2024;5(12):e244436. [doi: 10.1001/jamahealthforum.2024.4436] [Medline: 39671203]
49. Mostafapour M, Fortier JH, Garber G. Exploring the dynamics of physician-patient relationships: factors affecting patient satisfaction and complaints. J Healthc Risk Manag 2024;43(4):16-25. [doi: 10.1002/jhrm.21567] [Medline: 38706117]

Abbreviations

DHL: digital health care level
JARS: Journal Article Reporting Standards
OHC: online health community
VIF: variance inflation factor

XSL•FO
RenderX

Edited by S Brini; submitted 04.Oct.2025; peer-reviewed by Y Wan, Y Li; comments to author 30.Oct.2025; accepted 28.Dec.2025; published 29.Jan.2026.

Please cite as:
Liu Q, Yin P, Fan J
The Relationship Between Physician Self-Disclosure and Patient Acquisition in Digital Health Markets: Cross-Sectional Study
J Med Internet Res 2026;28:e84963
URL: https://www.jmir.org/2026/1/e84963
doi:10.2196/84963
PMID:

©Quanchen Liu, Pengqing Yin, Jing Fan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Website Use and Associations With Behavior Change and Weight Loss in Cancer Survivors and Their Partners: Secondary Analysis of a Randomized Controlled Trial

Harleen Kaur[1], PhD; Dori Pekmezi[2], PhD; Tracy E Crane[3], PhD, RD; David Farrell[4], MPH; Laura Q Rogers[5], MD, MPH; Wendy Demark-Wahnefried[6], PhD, RD

[1]Department of Psychology and Medicine, Division of Medical Oncology, Miller School of Medicine, University of Miami, Miami, FL, United States

[2]Department of Health Behavior, School of Public Health, University of Alabama at Birmingham, Birmingham, AL, United States

[3]Department of Medicine, Division of Medical Oncology, Miller School of Medicine, University of Miami, Miami, FL, United States

[4]People Designs (United States), Durham, NC, United States

[5]Department of Medicine, Division of General Internal Medicine and Population Science, School of Medicine, University of Alabama at Birmingham, Birmingham, AL, United States

[6]Department of Nutrition Sciences, School of Health Professions, University of Alabama at Birmingham, Birmingham, AL, United States

Corresponding Author:
Harleen Kaur, PhD
Department of Psychology and Medicine, Division of Medical Oncology
Miller School of Medicine
University of Miami
1425 NW Tenth Ave
Miami, FL, 33136
United States
Phone: 1 8323169470
Email: hxk629@miami.edu

Abstract

Background: Web-based lifestyle interventions to promote healthy diet and physical activity among cancer survivors and their partners are recent developments; therefore, few studies have reported patterns of website use or associations with behavior change.

Objective: The primary aim was to describe website use in the DUET (Daughters, Dudes, Mothers, and Others Together) trial and examine the associations between website use and changes in diet quality, moderate to vigorous physical activity (MVPA), and body weight.

Methods: This secondary analysis used data from 28 survivor-partner dyads (BMI ≥25 kg/m^2) randomized to the 6-month DUET web-based weight loss intervention, which released weekly e-learning sessions on diet and exercise. Website use was quantified as weeks of access, time spent, and frequency of page views. Diet quality was assessed using 2-day dietary recalls; MVPA was measured by the Godin Leisure-Time Exercise Questionnaire and accelerometry. Weight was measured on a scale. Website use was summarized descriptively, and associations were examined using Spearman partial correlations.

Results: Participants had a mean age of 58 (SD 12.5) years; 78.6% (44/56) identified as female, 66.1% (37/56) were non-Hispanic White, and 86% (24/28) were breast cancer survivors. On average, participants viewed 11.2 (SD 7.4) weeks of the 24-week intervention, or a total of 312.9 (SD 255.7) minutes per participant. *Sessions* (n=2736), *Home Page* (n=975), and *Tools* (n=967) features showed the highest activity (5885 total page views). Website use was higher among adults aged 65 years and older than younger participants, showcased by duration of use (mean 14.4, SD 7.4 weeks vs mean 9.2, SD 6.8 weeks; P=.009), time spent per week (mean 17.0, SD 9.7 minutes vs mean 10.5, SD 10.6 minutes; P=.01), and total number of page views (mean 135.7, SD 90 vs mean 85.3, SD 111.9; P=.008); higher website use was also reported among women versus men in terms of duration of use (mean 12.8, SD 7.1 weeks vs mean 5.6, SD 5.9 weeks; P=.003), time spent per week (mean 14.6, SD 10.3 minutes vs mean 7.4, SD 10.3 minutes; P=.02), and total number of page views (mean 120, SD 110.2 vs mean 50.3, SD 64.4; P=.01). Diet quality was positively associated with website use (weeks: r=0.50; P<.001; time: r=0.45; P<.001; total page views: r=0.46; P<.001; and sessions page views: r=0.39; P=.005). Self-reported MVPA was also positively associated with website use (weeks r=0.37;

XSL•FO
RenderX

P=.007; time: *r*=0.36; *P*=.009; total page views: *r*=0.36; *P*=.01; and sessions page views: *r*=0.35; *P*=.01). No significant associations were detected for accelerometry-measured MVPA or weight.

Conclusions: Cancer survivors and their partners engaged with the DUET web-based platform to support diet and physical activity (with use particularly high among older adults and females). However, larger, more diverse dyadic web-based lifestyle interventions are needed to confirm these findings.

Trial Registration: ClinicalTrials.gov NCT04132219; https://clinicaltrials.gov/study/NCT04132219

(J Med Internet Res 2026;28:e86908) doi:10.2196/86908

KEYWORDS

cancer survivors; diet; digital health; dyads; partners; physical activity; website

Introduction

Background

The number of cancer survivors in the United States is increasing, with over 18 million currently living with a history of cancer diagnosis [1,2]. While cure rates are encouraging, cancer survivors represent a population with a high burden of comorbidity, including cardiovascular disease, diabetes, and obesity [3,4]. Furthermore, most cancer survivors do not meet the nutrition and physical activity guidelines recommended by the World Cancer Research Fund/American Institute for Cancer Research (WCRF/AICR) and the American Cancer Society (ACS), often reporting low intake of vegetables and fruits (V&Fs) and insufficient physical activity [5-8]. These poor lifestyle practices can further exacerbate existing comorbidities, compounding their impact on long-term survivorship. Despite these challenges, cancer survivors frequently express a strong interest in improving their lifestyle behaviors [9]. Thus, lifestyle interventions that target diet, physical activity, and weight management have been implemented to promote healthier behaviors, support weight loss, and enhance the quality of life of cancer survivors [10-13].

Over the past decade, the landscape of lifestyle interventions for cancer survivors has evolved, with an increasing shift toward accessible and scalable digital delivery methods [14-16]. Among these, web-based lifestyle interventions have emerged as a promising approach to disseminate diet and physical activity guidance to cancer survivors [14,15]. Websites offer several advantages, such as scalability, personalization, cost-effectiveness, and the flexibility for participants to engage with content at their own pace [14,15,17]. To harness these advantages, the SurvivorSHINE pilot study, a 3-week, single-arm web-based lifestyle intervention, was implemented through an interactive website to promote diet and physical activity guidelines among 41 cancer survivors [18,19]. The website incorporated common strategies such as evidence-based diet and exercise knowledge, tools to facilitate behavior change, and resources for self-monitoring and goal setting [18]. The study reported that cancer survivors perceived the SurvivorSHINE website as a user-friendly platform that provided trustworthy information on diet and exercise, and reported improvements in knowledge related to diet and exercise [18].

Building upon the frameworks and strategies of the SurvivorSHINE intervention, a more refined website was developed that included 24 weekly serialized, interactive sessions, as well as additional tools to support behavior change not only among cancer survivors interested in cancer control, but also among their family members and friends for the purposes of cancer prevention. This newly developed website was then evaluated for feasibility and its impact on various health outcomes in a randomized controlled trial, known as DUET (Daughters, Dudes, Mothers, and Others Together; NCT04132219) that targeted cancer survivors and their chosen partners [20,21].

The DUET trial was 6 months in duration and evaluated the web-based weight loss intervention against a waitlist control among 112 participants (56 dyads, each consisting of a cancer survivor and a chosen partner) [20,21]. Results showed that dyads in the intervention arm achieved significant weight loss, along with improvements in diet quality and physical activity levels compared to the control arm [21]. A mediation analysis further revealed that reductions in perceived dietary barriers significantly contributed to the observed weight loss, highlighting the effectiveness of the intervention's strategies, which were based on social cognitive theory [22-24]. These findings underscore the potential of minimal-touch, web-based lifestyle interventions to facilitate meaningful behavior change and weight loss among cancer survivors and their partners. However, to optimize delivery and behavioral outcomes, it is important to understand how cancer survivors and their partners used the DUET website. While many web-based interventions focus on evaluating behavioral outcomes, relatively few have described website use or examined the association between website use and changes in diet, physical activity, and weight loss. Exploring patterns of use within the DUET website may provide insight into participant interaction with web-based platforms and inform refinements for future web-based lifestyle interventions for cancer survivors and their partners.

Aim and Objective

This secondary analysis aims to describe website use in the DUET trial, and also examine the associations between website use (as measured by number of weeks of website use, average time spent, total page views, and session page views) and changes in diet quality, moderate to vigorous physical activity (MVPA), and weight.

XSL•FO
RenderX

Methods

Study Design

The DUET study was a 2-arm single-blinded randomized controlled trial that enrolled 112 cancer survivors and their chosen partners (56 dyads). Dyads were randomized to either the 6-month web-based weight loss intervention or a waitlist control arm. This secondary analysis focuses on the subset of 56 participants (28 dyads) assigned to the DUET intervention arm, as website usability data were only collected for this group. Full details of the trial methods, primary outcomes, and mediation analysis have been published elsewhere [20-22].

Study Participants

Cancer survivors were identified through multiple recruitment strategies, including cancer registries, self-referrals, and curated lists of individuals who had previously expressed interest in lifestyle interventions. Recruitment efforts focused on cancer survivors who had completed treatment for obesity-related cancers with a 5-year survival rate of ≥70% (eg, localized renal, locoregional ovarian, colorectal, prostate, endometrial, or female breast cancer). Potential cancer survivors were contacted by mail with telephone follow-up, and, if interested, were screened by study staff for the following inclusion criteria: (1) BMI ≥25 kg/m²; (2) V&F intake <2.5 cups/day; (3) engaging in <150 minutes/week of MVPA; and (4) routine access to the internet via a computer, tablet, or mobile phone. Eligible cancer survivors were asked to identify a partner who lived nearby (within 10 minutes by car) and interacted with them at least biweekly; partners were screened using the same criteria, excluding a cancer diagnosis. However, partners with a history of cancer were eligible.

Study Protocol

Eligible dyads were provided with an overview of the study protocol, and written informed consent was obtained electronically using Adobe Sign [25]. After completion of baseline assessments, dyads were randomized to either the 6-month web-based weight loss intervention or a waitlist control arm. Those assigned to the intervention arm were granted access to the DUET website via an electronic link and instructed to create a secure profile using a username and password unique for each survivor and partner [26]. Dyads were encouraged to log in weekly via text messages and engage with the key features of the DUET website throughout the 6-month intervention period. Follow-up assessments were conducted at 6 months, after which waitlist control dyads were provided access to the DUET intervention. Additional details on the study procedures have been published previously [20].

DUET Intervention Website

The DUET intervention was theoretically grounded in social cognitive theory and incorporated elements from interdependence theory and the theory of communal coping to support both individual and dyadic behavior change by targeting diet and exercise barriers, enhancing social support, and building self-efficacy [23,24,27,28]. DUET was adapted from 2 prior evidence-based lifestyle interventions, Daughters and Mothers Against Breast Cancer and SurvivorSHINE, and delivered via a secure, interactive website that served as the primary web-based platform for intervention delivery [18,26,29]. Details on the intervention development have been published previously [20].

The DUET website included a total of 9 key features: *Home Page, My Profile, Sessions, Healthy Weight, Healthy Eating, Exercise, Tools, News You Can Use, and Team Support.* Upon account creation, participants accessed the *My Profile* feature to enter demographic and lifestyle data, including their current weight, height, and responses to 1-item questions that assessed their dietary intake (eg, consumption of V&Fs, whole grains, red and processed meats, added sugars, and alcohol), and frequency of snacking and physical activity (both endurance and resistance exercise). Cancer survivors also provided details on diagnosis and treatment, which informed tailored feedback on overcoming treatment-related challenges and generating dietary, physical activity, and weight management goals, as described previously [20-22]. The *Home Page* displayed a "Tip of the Day" designed to encourage ongoing engagement with the website, along with visual indicators showing participants' completed weekly sessions and the next upcoming session. It also included direct links to other key features of the website (ie, *Healthy Weight, Healthy Eating, Exercise, Tools, News You Can Use, and Team Support).*

The *Sessions* feature included 24 weekly interactive e-learning modules (~15 minutes each), designed using Articulate Storyline software [30]. Sessions were released sequentially each week over the 24 weeks and introduced via a Monday SMS "push" text message, with additional text messages sent on Wednesdays and Fridays to reinforce continued engagement. These interactive e-learning sessions were designed to provide information on the WCRF/AICR and ACS diet and physical activity guidelines and equip participants with practical and actionable strategies to support behavior change and weight management [5,6]. Dietary recommendations were supported by sessions focused on promoting the consumption of V&Fs, whole grains, and legumes, and limiting red and processed meat, added sugar, alcohol, and snacking, with additional sessions focused on portion control, grocery shopping, and food preparation to support healthy eating habits. Physical activity recommendations were supported by sessions focused on aerobic, resistance, balance, and flexibility exercises, with emphasis on goal setting and problem solving to help participants gradually achieve the goal of 150 minutes of MVPA per week. Details on weekly topics have been published previously [20].

The *Healthy Weight* feature was designed to support self-monitoring of weight by providing an interactive bar graph that tracked participants' current weight and healthy weight. Accompanying educational materials helped contextualize these values by explaining the concept of a healthy weight and its relevance to cancer prevention and survivorship. The website delivered tailored guidance to help participants progress toward a healthy weight by supporting caloric restriction and strategies to promote gradual weight loss of approximately 0.5 kg/week [31].

XSL•FO
RenderX

The *Healthy Eating* feature supported self-monitoring of dietary goals aligned with the WCRF/AICR and ACS guidelines. Participants could record their current intake of key dietary components, that is, V&Fs (≥5 servings/day), whole grains (≥50% of total grain intake), added sugars (≤6 teaspoons/day), avoid snacks, red and processed meats (≤18 ounces/week), and alcohol (≤1 drink/day), through an interactive bar graph that visually displayed their reported intake alongside goal targets. To further support dietary changes, the feature provided educational resources on each dietary component.

The *Exercise* feature supported self-monitoring of physical activity through tailored recommendations based on participants' self-reported activity levels. Physical activity goals were aligned with WCRF/AICR and ACS guidelines, encouraging participants to achieve at least 150 minutes of MVPA and engage in strength training 2-3 times per week.

The *Tools* feature provided a centralized hub for participants to access downloadable materials that supported both dietary and physical activity behaviors aimed at achieving a healthy weight. This feature included 11 distinct resources: (1) BMI calculator, (2) calorie calculator, (3) sample meal plans, (4) food exchange lists, (5) SMART goal templates, (6) serving size guides, (7) fast food guide, (8) grocery lists and shopping tips, (9) calorie-burning guide, (10) exercise logs, and (11) tools for tracking Fitbit data (note: all DUET participants regardless of randomization status, received a Fitbit Inspire and were encouraged to use it during the study period; however, Fitbit data were not integrated into the DUET website) [32]. These resources were designed to offer participants additional support, practical strategies, and accessible tools to enhance self-efficacy throughout the intervention.

The *News You Can Use* feature provided brief, evidence-based summaries on current research related to cancer survivorship, diet, and exercise. The *Team Support* section provided strategies to strengthen dyadic communication, foster mutual goal setting, and enhance social support; it also allowed participants to directly connect with study staff for additional guidance and support.

Measures

Demographics

Cancer type and time since diagnosis were obtained from cancer registries or verified by treating physicians for self-referred participants. Demographic information, including age, sex, race, residence, educational status, employment, and income, was self-reported via electronic surveys completed at baseline. Cohabitation was assessed by comparing mailing addresses; dyads with the same address (0 miles) were classified as cohabitating, and those with different addresses (>0 miles) as noncohabitating.

Diet Quality, MVPA, and Weight

Dietary intake was assessed at baseline and 6 months via two 24-hour dietary recalls (1 weekday and 1 weekend day) conducted by a registered dietitian over the telephone. The Automated Self-Administered 24-hour Dietary Assessment Tool was used to capture dietary intake data. Diet quality was evaluated using the Healthy Eating Index 2015 [33,34].

MVPA was assessed both objectively and subjectively at baseline and 6 months. Participants wore ActiGraph accelerometers for 7 consecutive days. Data were then processed using ActiLife software following standardized procedures to calculate average weekly minutes of MVPA [35,36]. Self-reported MVPA was captured using the Godin Leisure-Time Exercise Questionnaire, a validated tool frequently used in cancer survivorship research [37].

Weight was measured remotely at baseline and 6 months. Each survivor-partner dyad completed the virtual assessment together via Zoom with study staff [38]. Participants used a digital bathroom scale to report their weight; a scale was provided with the assessment materials for those who did not own one. During the virtual assessment, trained staff instructed participants to wear light clothing and remove their shoes, and partners assisted in holding the camera and angling it so that study staff could verify the weight displayed on the digital scale. Additional details on the remote assessment protocol and its validity have been described previously [39].

Website Use Metrics

Website use was assessed using tracking data logged by the DUET website platform. Each participant was assigned a unique website username and password, which was linked to their website ID and further connected to their study ID, allowing for the tracking of individual-level website activity over the 24-week intervention period. Time-stamped data recorded the days, times, and pages accessed by participants (eg, *Home Page, My Profile, Sessions, Healthy Weight, Healthy Eating, Exercise, Tools, News You Can Use, and Team Support*). From these data, key website use metrics were derived, including (1) the total number of weeks participants accessed the website, defined as the number of distinct weeks during the 24-week intervention period in which any website activity was recorded; (2) the average time spent on the website per day of use, calculated using time-stamped activity logs that captured first and last activity on a given day; (3) the total number of page views, defined as the cumulative number of times participants navigated to individual pages within each of the website's key features; and (4) the total number of session page views, defined as the number of times participants accessed the e-learning *Sessions* feature.

Statistical Analysis

Website usability was analyzed using descriptive statistics, including means, SDs, and ranges for continuous variables (eg, number of weeks accessed, time spent on the website, and page views). Website usability metrics were described for the total sample and stratified by dyad member (cancer survivors or partners), clinical (ie, cancer type and time since diagnosis) and sociodemographic factors (ie, age, sex, race, residence, educational status, employment, and income) and cohabitation status (cohabitate or did not cohabitate), with all stratification variables dichotomized for analysis. To examine differences in website use between dyad members, clinical and sociodemographic factors, and cohabitation status, assumptions

XSL•FO
RenderX

of normality were assessed using Shapiro-Wilk tests and visual inspection of histograms and Q-Q plots. Given that usability metrics were not normally distributed, Wilcoxon rank-sum tests were conducted to compare median differences in website use between dyad members, clinical and sociodemographic factors, and cohabitation status. Assumptions of normality and linearity were assessed for the independent (website use metrics) and dependent variables (diet quality, MVPA, and weight). Given evidence of nonnormal distributions, likely influenced by the modest sample size and potential nonlinear relationships, nonparametric methods, such as bivariate Spearman partial rank correlation analyses, were conducted to examine associations between website use and diet quality, MVPA, and weight. Correlation coefficients were generated using 6-month outcome values as dependent variables, adjusting for baseline values of the respective outcome, as well as age, sex, and race to account for potential confounding and assess change over time. Given this was a secondary analysis and not prospectively powered for these aims, we conducted a post hoc power calculation to aid interpretation of the correlation analyses. With a total sample size of 56 participants, the study had ≥80% power to detect correlations of approximately $r \geq 0.40$. Missing data were handled using complete-case analysis. One participant had missing diet quality data at 6-month follow-up, and 13 participants had missing accelerometer-measured MVPA data at baseline (n=8) and 6-month follow-up (n=5); these cases were excluded from their respective models. No adjustments for multiple comparisons were made, given the exploratory nature of this secondary analysis. However, to address the increased type I error risk from multiple correlations, 95% CIs were reported with P values to assist in interpreting the precision of the coefficients. All analyses were conducted using SAS (version 9.4, SAS Institute Inc), and statistical significance was set at $P<.05$ [40].

Ethical Considerations

The DUET study was approved by the Institutional Review Board at the University of Alabama at Birmingham (IRB# 300003882) and was registered with ClinicalTrials.gov (NCT04132219). All participants provided written informed consent, and study procedures were conducted in accordance with the ethical standards of the Declaration of Helsinki to maintain participant confidentiality. No compensation was provided to study participants.

Results

Sample Characteristics

The average age of the sample was 58 (SD 12.5) years, with survivors averaging 60 (SD 11.2) years and partners 56 (SD 13.7) years. The majority of cancer survivors (24/28, 85.7%) had a breast cancer diagnosis, with an average time since diagnosis of approximately 71 (SD 80.4) months (6 years). Most participants identified as female (44/56, 78.6%), non-Hispanic White (37/56, 66.1%), and residents of urban areas (53/56, 94.6%). Employment status was evenly split between employed (30/56, 53.6%) and retired (26/56, 46.4%), and most (45/56, 80.4%) reported an annual household income above US $50,000 per year (Table 1).

XSL•FO
RenderX

Table 1. Characteristics of 56 cancer survivors and their chosen partners randomized to the intervention arm, stratified by dyad status.

Characteristics	Total sample	Survivor	Partner	*P* value[a]
Age (years), mean (SD; range)	58.1 (12.5; 23-78)	60.0 (11.2; 32-78)	56.3 (13.7; 23-74)	.28
Months from diagnosis, mean (SD; range)	71 (80.4; 10-303)	71 (80.4; 10-303)	__[b]	__[b]
BMI (kg/m^2), mean (SD; range)	31.4 (4.9; 25-45)	32.0 (5.4; 25-44)	30.9 (4.4; 25-45)	.41
Diet quality (HEI[c]), mean (SD; range)	53.1 (12.8; 29-87)	53.9 (13.7; 30-87)	52.2 (12; 29-81)	.62
MVPA[d] (min/week), mean (SD; range)	43.8 (60.5; 0-280)	48.5 (67.8; 0-280)	39.1 (52.9; 0-225)	.57
Cancer type[e], n (%)				<.001
Breast	25 (44.6)	24 (85.7)	1 (3.6)	
Other[f]	7 (12.5)	4 (14.3)	3 (10.7)	
Sex, n (%)				.05
Male	12 (21.4)	3 (10.7)	9 (32.1)	
Female	44 (78.6)	25 (89.3)	19 (67.9)	
Race, n (%)				.78
Non-Hispanic White	37 (66.1)	19 (67.9)	18 (64.3)	
Non-Hispanic Black or other[g]	19 (33.9)	9 (32.1)	10 (35.7)	
Residence, n (%)				.55
Urban	53 (94.6)	27 (96.4)	26 (92.9)	
Rural	3 (5.4)	1 (3.6)	2 (7.1)	
Educational status, n (%)				.13
High school or less	8 (14.3)	2 (7.1)	6 (21.4)	
Some college or more	48 (85.7)	26 (92.9)	22 (78.6)	
Employment, n (%)				>.99
Employed	30 (53.6)	15 (53.6)	15 (53.6)	
Retired or other[h]	26 (46.4)	13 (46.4)	13 (46.4)	
Income, n (%)				.09
Less than US $50,000/year	11 (19.6)	8 (28.6)	3 (10.7)	
More than US $50,000/year	45 (80.4)	20 (71.4)	25 (89.3)	

[a]*P* values were calculated using independent samples *t* tests for continuous variables (based on equal or unequal variances as appropriate) and chi-square tests for categorical variables, comparing survivors vs partners, significance set at *P*<.05.

[b]Data on months since diagnosis is unavailable for partners.

[c]HEI: Healthy Eating Index 2015.

[d]MVPA: moderate to vigorous physical activity.

[e]Total percentage does not sum to 100% due to the inclusion of nonsurvivors who were not diagnosed with cancer.

[f]Other cancer diagnoses include prostate, colorectal, gynecologic, and renal.

[g]Other race includes Hispanic ethnicity, accounting for 1%.

[h]Other employment includes student and disabled.

DUET Website Use

On average, participants accessed the DUET website for 11.2 (SD 7.4; range 0-24) weeks and spent a total of 312.9 (SD 255.7; range 0-1119) minutes on the platform, or about 13 (SD 10.7; range 0-46.6) minutes per week over the 24-week intervention period. A total of 5885 page views were recorded, with an average of 105.1 (SD 105.7; range 0-545) page views per participant. Cancer survivors used the website more frequently than their partners, accessing it for a longer duration (mean 13.0, SD 7.2 weeks vs mean 9.5, SD 7.4 weeks; *P*=.08), spending more time per week (mean 15.6, SD 11.4 minutes vs mean 10.5, SD 9.4 minutes; *P*=.07), and recording a higher average number of page views (mean 124.2, SD 114.2 vs mean 86.0, SD 94.5; *P*=.07); however, these differences were not statistically significant (Table 2).

XSL•FO
RenderX

Table 2. Daughters, Dudes, Mothers, and Others Together website usability over the 24-week intervention period for the total sample (n=56) and stratified by dyad status.

Engagement metrics	Total sample, mean (SD; range)	Survivor, mean (SD; range)	Partner, mean (SD; range)	*P* value[a]
Weeks participants accessed website	11.2 (7.4; 0-24)	13.0 (7.2; 0-24)	9.5 (7.4; 0-23)	.08
Total time spent on website (min)	312.9 (255.7; 0-1119)	373.8 (273.3; 0-1119)	252.0 (225.5; 0-670)	.07
Time spent on website per week (min)	13.0 (10.7; 0-46.6)	15.6 (11.4; 0-46.6)	10.5 (9.4; 0-27.9)	.07
Total page views per user (n=5885)	105.1 (105.7; 0-545)	124.2 (114.2; 0-545)	86.0 (94.5; 0-352)	.07

[a]*P* values represent comparisons between survivors and partners using Wilcoxon rank-sum tests due to the nonnormal distribution of website usability metrics, significance set at $P<.05$.

Website use differed significantly by age, time since diagnosis, and sex. Older adults aged ≥65 years demonstrated higher website use compared to younger participants (<65 years), accessing it for a longer duration (mean 14.4, SD 7.4 weeks vs 9.2, SD 6.8 weeks; *P*=.009), spending more time per week (mean 17.0, SD 9.7 minutes vs mean 10.5, SD 10.6 minutes; *P*=.01), and recording a higher average number of total page views (mean 135.7, SD 90 vs mean 85.3, SD 111.9; *P*=.008). Survivors who were 5 or more years post diagnosis also accessed the website for longer duration than those more recently diagnosed (<5 years; mean 18.4, SD 6.3 weeks vs mean 10.8, SD 6.4 weeks; *P*=.009). Additionally, females used the website significantly more than males, accessing it for a longer duration (mean 12.8, SD 7.1 weeks vs mean 5.6, SD 5.9 weeks; *P*=.003), spending more time per week (mean 14.6, SD 10.3 minutes vs mean 7.4, SD 10.3 minutes; *P*=.02), and recording a higher average number of total page views (mean 120, SD 110.2 vs mean 50.3, SD 64.4; *P*=.01; Table 3).

XSL•FO
RenderX

Table 3. Website use stratified by clinical characteristics, sociodemographic factors, and cohabitation status.

Measures	Weeks			Average time			Total page views			Session page views		
	Mean (SD)	Test statistic[a]	*P* value	Mean (SD)	Test statistic	*P* value	Mean (SD)	Test statistic	*P* value	Mean (SD)	Test statistic	*P* value
Age		6.83	.009		6.55	.01		7.03	.008		6.34	.01
<65 years	9.17 (6.8)			10.5 (10.6)			85.3 (111.9)			38.4 (35.2)		
≥65 years	14.4 (7.4)			17 (9.7)			135.7 (90)			65 (38.5)		
Time since diagnosis		6.88	.009		1.62	.20		2.57	.11		0.622	.43
<5 years	10.8 (6.4)			14 (11.9)			108.4 (117.8)			52.5 (35.1)		
≥5 years	18.4 (6.3)			19.4 (9.6)			163.6 (100.9)			65.3 (32.3)		
Cancer type		0.64	.42		0.03	.87		0.53	.47		1.15	.28
Breast	13.6 (7.2)			15.3 (11.3)			122.9 (119.5)			54.8 (37.1)		
Other	11.6 (4.9)			15.9 (9.7)			129.6 (81)			27 (36.7)		
Sex		8.82	.003		5.70	.02		6.69	.01		6.39	.01
Female	12.8 (7.1)			14.6 (10.3)			120 (110.2)			54.8 (37.1)		
Male	5.6 (5.9)			7.4 (10.3)			50.3 (64.4)			27 (36.7)		
Race		0.02	.89		0.08	.78		0.11	.75		0.01	.91
Non-Hispanic White	11.4 (6.8)			12.9 (9.6)			100.2 (86.2)			46.4 (33.2)		
Non-Hispanic Black	10.9 (8.6)			13.3 (12.5)			113.9 (136)			53.4 (47.1)		
Residence		0.01	.91		0.41	.52		0.28	.60		0.90	.34
Urban	11.2 (7.3)			12.7 (10.3)			104 (106.8)			47.2 (36.3)		
Rural	12.3 (11)			18.4 (17.1)			123.7 (98.1)			77.7 (70.3)		
Education		0.95	.33		0.01	.92		0.08	.78		0.00	.98
High school or less	8.9 (6.3)			12 (10.5)			116 (131.7)			45.9 (37.8)		
Some college or above	11.6 (7.6)			13.2 (10.8)			103.3 (102.2)			49.4 (38.9)		
Employment		1.29	.26		0.14	.71		0.35	.55		0.14	.71
Employed	10.4 (6.3)			12.9 (11.7)			101.8 (117.1)			47.6 (39.7)		
Retired or other	12.2 (8.5)			13.3 (9.5)			108.9 (92.8)			50.3 (37.7)		
Income		0.08	.78		0.45	.50		0.18	.67		0.81	.37
Less than US $50k/year	10.6 (5.9)			15.9 (13.9)			132 (152.8)			59.7 (44.9)		
More than US $50k/year	11.4 (7.8)			12.3 (9.8)			98.5 (91.7)			46.2 (36.8)		

XSL•FO
RenderX

Measures	Weeks			Average time			Total page views			Session page views		
	Mean (SD)	Test statistic[a]	*P* value	Mean (SD)	Test statistic	*P* value	Mean (SD)	Test statistic	*P* value	Mean (SD)	Test statistic	*P* value
Cohabitation		2.23	.14		1.54	.21		0.40	.53		2.01	.16
Cohabitate	9.5 (8.1)			11.1 (10.3)			93.1 (91)			42.2 (39.8)		
Do not co-habitate	12.5 (6.7)			14.5 (10.8)			114.1 (116)			53.8 (37.2)		

[a]Differences between groups were assessed using the Mann-Whitney *U* test, implemented via the Wilcoxon rank-sum approach. A chi-square approximation was used to derive the test statistics, with statistical significance defined as $P<.05$.

Website use was generally highest during the initial 9 weeks of the intervention, although some fluctuations were observed, particularly in weeks 4, 6, and 8. Website use gradually declined over time, with a slight increase observed in the final week of the intervention. Cancer survivors spent more time on the website each week compared to their partners, with notable peaks observed among cancer survivors at weeks 1, 2, 6, 8, 13, 16, 19, and 24 (Figure 1).

Figure 1. Average weekly time spent on the Daughters, Dudes, Mothers, and Others Together (DUET) website during the 24-week intervention period by the total sample, highlighting high and low engagement weeks stratified by dyad status. All weekly DUET sessions were designed to take 15 minutes or less, except for the initial onboarding session.

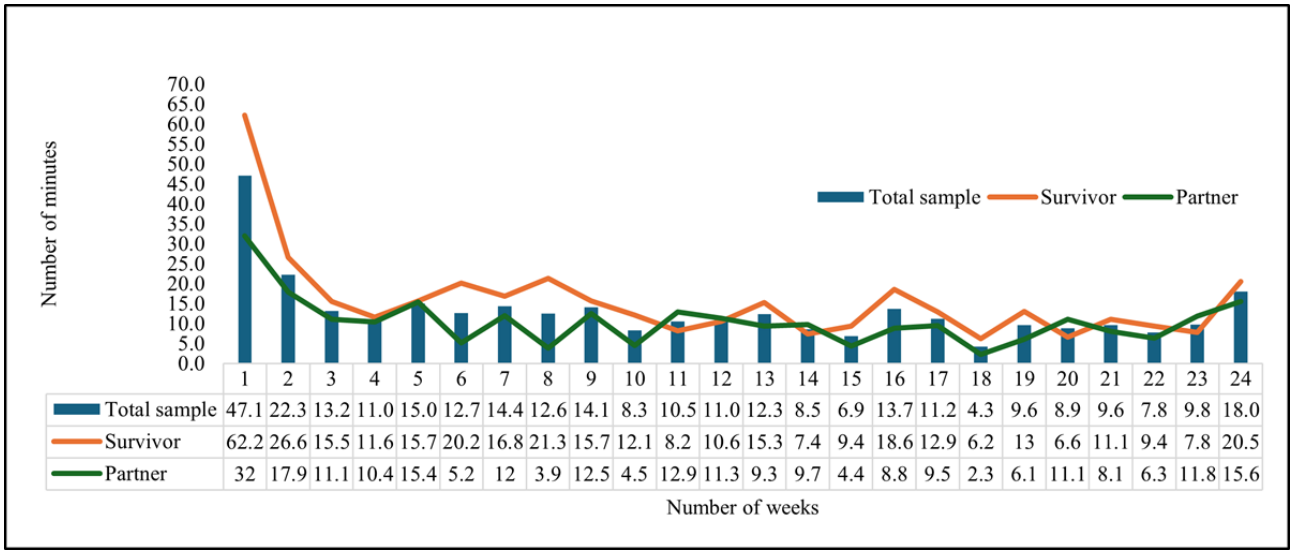

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Total sample	47.1	22.3	13.2	11.0	15.0	12.7	14.4	12.6	14.1	8.3	10.5	11.0	12.3	8.5	6.9	13.7	11.2	4.3	9.6	8.9	9.6	7.8	9.8	18.0
Survivor	62.2	26.6	15.5	11.6	15.7	20.2	16.8	21.3	15.7	12.1	8.2	10.6	15.3	7.4	9.4	18.6	12.9	6.2	13	6.6	11.1	9.4	7.8	20.5
Partner	32	17.9	11.1	10.4	15.4	5.2	12	3.9	12.5	4.5	12.9	11.3	9.3	9.7	4.4	8.8	9.5	2.3	6.1	11.1	8.1	6.3	11.8	15.6

A total of 5885 page views were recorded throughout the intervention period across the 9 key website features. The highest number of page views was observed for the weekly interactive e-learning *Sessions* feature (n=2736), followed by the *Home Page* (n=975) and *Tools* feature (n=967). The remaining features, *My Profile* (n=400), *Healthy Weight* (n=248), *Healthy Eating* (n=234), *Exercise* (n=162), *News You Can Use* (n=104), and *Team Support* (n=59), had comparatively fewer page views (Figure 2).

XSL•FO
RenderX

Figure 2. Page view distribution across key features of the Daughters, Dudes, Mothers, and Others Together website stratified by dyad status (n=5885 total page views).

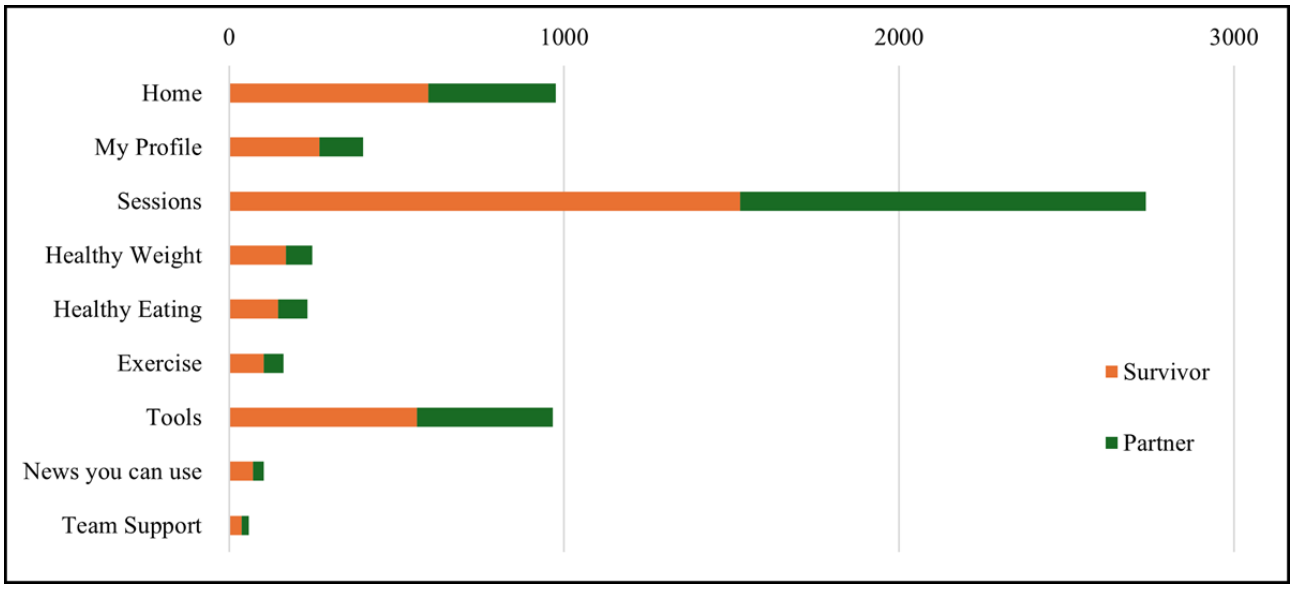

Figure 3 further illustrates the distribution of 2736 page views across the weekly e-learning *Sessions* feature over the 24-week intervention period. Page views for the weekly sessions were highest during the early weeks of the intervention, particularly sessions 1 through 8, and gradually declined over time. The onboarding session received the greatest number of views, and the three most frequently viewed sessions were (1) Get on Track for Success (session 2), (2) Moving Towards Better Health (session 4), and (3) Been Resisting "Resistance" Exercises (session 8). Conversely, the three least viewed sessions were (1) Want to Join the Party Without Blowing Your Diet? (session 20), (2) Why Am I Hungry All the Time? (session 22), (3) Are Supplements Really Good for You? (session 23).

Figure 3. Page view distribution of weekly released diet and exercise sessions accessed by cancer survivors and partners (n=2736 total session page views). Top 3 most viewed sessions for the total sample: Get on Track for Success (session 2), Moving Towards Better Health (session 4), and Been Resisting "Resistance" Exercises (session 8). Bottom 3 least viewed sessions for the total sample: Want to Join the Party Without Blowing Your Diet (session 20), Why Am I Hungry All the Time (session 22), Are Supplements Really Good for You (session 23).

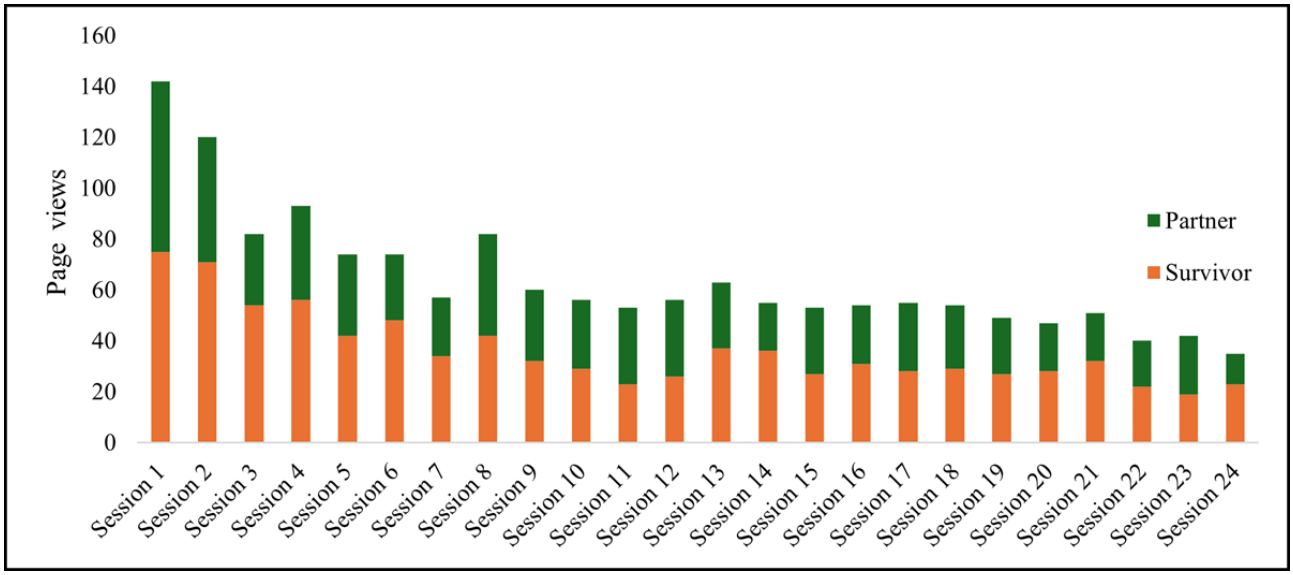

The *Tools* feature accounted for the third highest number of page views (n=967), following the *Home page* and *Sessions*, with participants most frequently viewing tools such as Sample Meal Plans (n=121 page views), Tracking with Fitbit (n=119), Exercise Logs (n=65), BMI Calculator (n=65), Fast Food Menu Maven (n=64), Calorie Calculator (n=62) and Calorie Burning Guide (n=60; Multimedia Appendix 1). The *Healthy Weight, Healthy Eating, and Exercise* features were also accessed throughout the intervention period. Within these features, the most commonly viewed content included "Common Questions About Weight Management" (n=20 page views) under the Healthy Weight category, "Increasing V&F Intake" (n=10 page views) under Healthy Eating, and "Leg Strengthening Exercises" (n=19 page views) under the Exercise category (Multimedia Appendix 2)

XSL•FO
RenderX

Website Use and Behavioral Associations

Diet quality was positively associated with website use, weeks (r=0.50; P<.001), time (r=0.45; P<.001), total page views (r=0.46; P<.001), and sessions page views (r=0.39; P=.005). Self-reported MVPA was also positively associated with website use, weeks (r=0.37; P=.007), time (r=0.36; P=.009), total page views (r=0.36; P=.01), and sessions page views (r=0.35; P=.01). However, no statistically significant associations were detected for accelerometry-measured MVPA or weight (P>.05; Table 4).

Table 4. Bivariate associations between website usability and Healthy Eating Index 2015 diet quality, moderate to vigorous physical activity (MVPA), and weight at 6-months. Bivariate associations were performed using Spearman partial correlations rank analysis for nonnormally distributed data. Adjusted for age, sex, race, and baseline outcome variables.

Measures	Weeks		Average time		Total page views		Session page views	
	r (95% CI)	*P* value	*r* (95% CI)	*P* value	*r* (95% CI)	*P* value	*r* (95% CI)	*P* value
Diet quality	0.50 (0.25 to 0.68)	<.001	0.45 (0.20 to 0.65)	<.001	0.46 (0.21 to 0.65)	<.001	0.39 (0.12 to 0.60)	.005
Self-reported MVPA	0.37 (0.10 to 0.58)	.007	0.36 (0.09 to 0.57)	.009	0.36 (0.09 to 0.57)	.01	0.35 (0.08 to 0.57)	.01
Accelerometer MVPA	0.15 (–0.16 to 0.44)	.33	0.06 (–0.25 to 0.37)	.68	0.03 (–0.28 to 0.34)	.82	0.04 (–0.27 to 0.35)	.79
Weight	–0.17 (–0.42 to 0.11)	.24	–0.06 (–0.32 to 0.22)	.67	–0.13 (–0.39 to 0.15)	.35	–0.11 (–0.38 to 0.16)	.41

Discussion

Primary Findings

There are very few web-based lifestyle interventions for cancer survivors and their support partners that use evidence-based theoretical constructs to promote healthful diet and physical activity behaviors [14]. As a result, no studies to date have described patterns of website use to inform the design and delivery of future dyadic, web-based lifestyle interventions. This study is among the first to address this gap by providing a detailed analysis of website use patterns among cancer survivors and their chosen partners randomized to the DUET intervention. Our findings suggest that cancer survivors and their partners used the DUET website to learn about healthy lifestyle guidelines, as reflected by website use metrics: on average, participants logged into roughly half of the 24-week content, interacting for a total of 313 minutes with the program, for which the highest page view activity was observed for *Sessions* (n=2736). Older adults and females engaged with the website to a significantly greater degree compared to younger participants and males, as reflected by weekly website activity (frequency, user time, and page views). Moreover, higher levels of website use were significantly associated with improvements in diet quality and self-reported MVPA. However, no significant associations were observed between website use and accelerometer-measured MVPA and weight.

Comparison With Previous Literature

Due to variability in how website usability metrics are reported, direct comparisons of website use across studies are challenging. However, findings from 3 pilot web-based lifestyle interventions implemented among cancer survivors generally align with our data that cancer survivors engage with a lifestyle website approximately once per week, particularly during the initial phase of the study. For example, the SurvivorSHINE study reported an average of 1.5 log ins per week over a 2-week period [41]. Similarly, in the A Lifestyle Intervention Via Email study, breast cancer survivors visited the website for an average of 9.6 out of 12 weeks in the physical activity arm and 10.7 out of 12 weeks in the diet arm [42]. A study by Blarigan and colleagues [43] reported that colorectal cancer survivors in the intervention arm accessed the website on a median of 13 out of 84 days. Despite the 24-week DUET intervention being roughly twice as long as web-based interventions used in previous studies, we observed comparable weekly website use during the initial 12 weeks of the program, suggesting sustained and moderate website use among dyads during the initial phases of the intervention. Our findings also showed that dyads spent a total of 313 minutes on the website, which differs from the SurvivorSHINE study (94 minutes) but aligns closely with the Breast Cancer eHealth Self-Management study (337 minutes) [41,44]. The difference in website use observed in our study compared to SurvivorSHINE is likely due to the longer intervention duration and the inclusion of weekly serialized e-learning sessions and videos, which were not part of the SurvivorSHINE's 2-week intervention.

This study also examined page view activity to assess participant usability with various DUET website features. The most frequently accessed feature was the interactive e-learning "*Sessions*," followed by the "*Home Page*" feature. Prior systematic reviews have emphasized the importance of interactive educational modules to promote real-time engagement with behavior change content [45-47]. The DUET "*Sessions*" were developed with this framework in mind, incorporating brief, skill-building activities designed to help both survivors and their support partners apply evidence-based knowledge to everyday challenges. These sessions encouraged teamwork and joint problem-solving to support healthier behaviors, a strategy supported by the theory of communal coping and prior studies [27,48,49]. Participants also received 3 weekly text messages reminding them to visit the website and engage with the weekly content, which likely contributed to the high usability observed with the "*Sessions*" feature. The brief, focused, and interactive nature of the sessions may have further

XSL•FO
RenderX

enhanced their appeal. Similar to findings from the SurvivorSHINE study, the "*Home Page*" was also commonly viewed, likely because it automatically loaded each time participants accessed the website, allowing them to engage with core behavior change constructs such as self-monitoring, goal setting, and motivation [41].

Consistent with prior research conducted in samples without a history of cancer, our findings indicate that website use was significantly higher among older adults (aged ≥65 years) and female participants. For example, Graham et al [50] reported that older adults using the commercially available Lark Health digital platform logged more meals (174 vs 89) and used more self-monitoring devices (39 vs 28) compared to younger adults. Relatedly, a scoping review found that among adults aged 50 years and older, structured, tailored digital programs were well-accepted [51]. Similarly, a growing body of literature has consistently shown that women are more likely to seek health-related information, engage with online health platforms, and participate in digital lifestyle interventions compared to men [14,52,53]. In our study, this trend may be amplified by the DUET program's structured, self-paced design and tailored content, which likely resonated with older and female participants who may prefer individualized guidance and flexibility with web-based programs [52,54].

Bivariate analyses from this study showed that website use was significantly associated with improvements in diet quality and self-reported MVPA at 6 months. The evidence on whether website use improves diet quality remains unclear, largely due to heterogeneity in how diet quality is measured across studies. For instance, our findings differ from those of a systematic review and meta-analysis of 29 studies in adults with chronic conditions, which found no evidence for website use to improve overall diet quality [55]. However, most studies in this review were short in duration (<3 months) and relied on self-reported measures; in contrast, our study used interviewer-administered 24-hour dietary recalls, the gold standard for diet quality assessment, which may be more sensitive to detecting change. Our findings are more consistent with systematic reviews and several studies that have reported moderate to strong associations between website use and increased physical activity using self-reported measures [41,55-57]. One possible reason the DUET website led to improvements in diet quality and self-reported MVPA is its inclusion of a range of relevant content for cancer survivors and their partners, such as guidance on V&F intake, reducing processed foods, portion control, setting SMART goals, and providing tools for tracking and self-monitoring diet and physical activity behaviors. However, it is important to note that diet quality and self-reported MVPA models were modestly significant, and may not withstand correction for multiple testing and should be interpreted as hypothesis-generating.

Despite these positive associations, our analysis did not find a significant association between website use and accelerometer-measured MVPA and weight. Our findings for not detecting significant associations with subjectively measured physical activity aligns with findings from previous studies [55,56], but, they contrast with findings from the Commonwealth Scientific and Industrial Research Organisation (CSIRO) Total Wellbeing Diet Online program and the Weight Loss Maintenance (WLM) trial, which found that website use was associated with greater weight loss and reduced weight regain, respectively [58,59]. However, the CSIRO program was commercially delivered and relied on participant self-reported weights within the platform, while the WLM trial focused on weight maintenance rather than weight loss [58]. Several methodological, individual, and behavioral level considerations may help explain the lack of association between website use and accelerometer-measured MVPA and weight [60]. Our power calculation confirmed that with 56 participants, the study was powered only to detect correlations of approximately $r \geq 0.40$. Given that the association between website use and accelerometer-measured MVPA and weight observed in our data was substantially smaller ($r < 0.20$), the study was likely underpowered to detect this relationship. Additionally, differences in how outcomes were measured may further clarify the findings. Website use, diet quality, and self-reported MVPA rely on participant interaction and reporting, which may naturally relate to one another and explain shared method variance. In contrast, accelerometer-measured MVPA and weight are objective measures that do not rely on participant reporting, and therefore their associations with website use may be smaller and more difficult to detect. Beyond methodological explanations, individual and behavioral factors may also play a role. For instance, website use may support behavior change but not be sufficient on its own to produce weight loss, as some participants may have relied less on the website once new habits were established. Additionally, weight loss may be influenced more by theoretical and behavioral mechanisms (ie, reduced perceived barriers, self-monitoring, calorie restriction, increased accountability from participating in the study, or lifestyle changes occurring outside the platform). These possibilities suggest that website use alone may not fully reflect the processes that contributed to weight loss in the parent trial [21].

Strengths and Limitations

This study had several notable strengths. It examined website use among cancer survivors and their partners, an area with limited prior research, as few lifestyle interventions have studied website use among dyads. Our analysis also provided detailed page view analytics across the entire DUET website, offering insight into which website features were most frequently used. Furthermore, the study used validated measures to assess both diet quality and MVPA, enhancing the accuracy of the findings. However, like all studies, there were limitations. Most importantly, website data were only available for the 56 participants in the intervention arm, resulting in a relatively small analytic sample. A post hoc power calculation indicated that with this sample size, the study was powered (≥80%) only to detect correlations of approximately $r \geq 0.40$, representing a moderate to large effect size. As a result, smaller associations, particularly for accelerometer-measured MVPA and weight change, may not have been detectable, and the null findings for these outcomes should be interpreted with caution. Future web-based trials should consider recruiting larger samples to ensure adequate power to detect smaller associations between website use and clinically relevant outcomes such as weight change. Additionally, accelerometer-measured MVPA had 20%

XSL•FO
RenderX

data missingness. As a result, complete-case analysis may introduce bias if participants with missing MVPA data differed meaningfully from those with complete data. Importantly, because multiple correlations were examined, there is an increased risk of type I error, and some associations may reflect spurious findings; therefore, the results should be interpreted within the exploratory, hypothesis-generating context of this secondary analysis. Similarly, it was not possible to determine whether participants viewed the full content of the weekly sessions they accessed, as data only represented whether participants accessed the sessions. Thus, it is likely that participants who briefly clicked into a session could therefore be coded similarly to those who reviewed the entire content, and time-stamp data may overestimate engagement if browser windows were left open. To partially address this limitation, we examined multiple measures of website use rather than relying on a single metric. Nevertheless, this measurement constraint may have led to imprecise estimates of website use and may help explain the modest strength of associations observed. Additionally, given the dyadic nature of the study, there may have been instances where dyads shared a single account and viewed website content together (as was documented in at least 1 case), potentially accounting for the difference between survivor and partner usage and underestimating individual-level usability. Another limitation is the demographic homogeneity of the sample, which primarily included female, non-Hispanic White breast cancer survivors residing in urban areas with higher socioeconomic status. As a result, these findings may not generalize to male survivors, racial and ethnic minority groups, individuals with lower income levels, or those living in rural or medically underserved settings.

Conclusions

Minimal-touch lifestyle interventions delivered through web-based platforms are being implemented to promote diet and physical activity behaviors for weight management among cancer survivors. However, patterns of website use are often understudied. Findings from our study suggest that cancer survivors and their partners (especially older adults and females) actively used the DUET website, particularly the interactive e-learning sessions, and higher levels of website use were significantly associated with improvements in diet quality and self-reported MVPA. However, no statistically significant improvements were detected for accelerometry-measured MVPA or weight. These results highlight that web-based platforms may serve as a promising, scalable approach for delivering diet and physical activity guidelines and promoting healthy behaviors for long-term older cancer survivors and their partners. However, larger, more diverse dyadic web-based lifestyle interventions, including male survivors, racial and ethnic minority populations, individuals with lower income, and survivors in rural or underserved areas, are needed to confirm and expand upon these findings. Importantly, future web-based lifestyle interventions should also use more detailed engagement tracking (ie, content completion indicators, differentiation between brief access and full interaction with embedded activities, and minimum and maximum time thresholds) to distinguish brief access from full content consumption to strengthen the validity of engagement measures.

Acknowledgments

First and foremost, we sincerely thank the cancer survivors and their partners who participated in the DUET intervention. We also gratefully acknowledge our funding sources: the American Institute for Cancer Research (585363); the American Cancer Society (CRP-19–175-06-COUN); the National Cancer Institute (P01 CA229997); O'Neal Comprehensive Cancer Center (P30 CA013148); the Cancer Prevention and Control Training Program (T32 CA047888); the Training to Reduce Burden across the Cancer Control Continuum multidisciplinary program (T32 CA251064-5); and the University of Arizona's Comprehensive Cancer Center (P30 CA023074). Lastly, we extend our appreciation to the clinical staff and students (Drs Teri Hoenemeyer and Amber Kinsey, J Ryan Buckman, Grey Freylersythe, Lauren King, Doctorre McDade, Christopher Reid, Abigail Sims, and Fariha Tariq) for their dedication and efforts in the design, implementation, and successful completion of the study. No generative artificial intelligence was used in any part of the manuscript creation.

Funding

This work was funded by the American Institute for Cancer Research (grant 585363); the American Cancer Society (grant CRP-19–175-06-COUN); the National Cancer Institute (grant P01 CA229997); O'Neal Comprehensive Cancer Center (grant P30 CA013148); the Cancer Prevention and Control Training Program (grant T32 CA047888); the Training to Reduce Burden across the Cancer Control Continuum multidisciplinary program (grant T32 CA251064-5); and the University of Arizona's Comprehensive Cancer Center (grant P30 CA023074). The funder was not involved in the conduct of the research, the decision to publish, or the approval of the publication.

Data Availability

The data reported in this study are available on request from the corresponding author.

Authors' Contributions

Conceptualization: HK (lead), WDW
Data curation: HK (lead), WDW

XSL•FO
RenderX

Formal analysis: HK (lead), WDW
Funding acquisition: WDW (lead)
Investigation: HK (lead), WDW (equal)
Methodology: WDW (lead), DF, DP, TEC, LQR, HK
Project administration: WDW (lead), DF, DP, HK
Resources: WDW (lead)
Supervision: WDW (lead)
Validation: WDW (lead)
Visualization: HK (lead), WDW
Writing – original draft: HK (lead)
Writing – review & editing: HK (lead), WDW (lead), DF, DP, LQR, TEC

Conflicts of Interest

None declared.

Multimedia Appendix 1
Distribution of DUET website page view activity for specific tools accessed by cancer survivors and their supportive partners, based on 967 total page views.
[PNG File , 55 KB - jmir_v28i1e86908_app1.png]

Multimedia Appendix 2
Distribution of DUET website page view activity for healthy eating, healthy weight, and exercise features accessed by cancer survivors and their chosen partners, based on 644 total page views.
[PNG File , 76 KB - jmir_v28i1e86908_app2.png]

References

1. Bluethmann SM, Mariotto AB, Rowland JH. Anticipating the "Silver Tsunami": prevalence trajectories and comorbidity burden among older cancer survivors in the United States. Cancer Epidemiol Biomarkers Prev 2016;25(7):1029-1036 [FREE Full text] [doi: 10.1158/1055-9965.EPI-16-0133] [Medline: 27371756]
2. Wagle NS, Nogueira L, Devasia TP, Mariotto AB, Yabroff KR, Islami F, et al. Cancer treatment and survivorship statistics, 2025. CA Cancer J Clin 2025;75(4):308-340 [FREE Full text] [doi: 10.3322/caac.70011] [Medline: 40445120]
3. Shahrokni A, Wu AJ, Carter J, Lichtman SM. Long-term toxicity of cancer treatment in older patients. Clin Geriatr Med 2016;32(1):63-80 [FREE Full text] [doi: 10.1016/j.cger.2015.08.005] [Medline: 26614861]
4. Heo J, Chun M, Oh Y, Noh OK, Kim L. Metabolic comorbidities and medical institution utilization among breast cancer survivors: a national population-based study. Korean J Intern Med 2020;35(2):421-428 [FREE Full text] [doi: 10.3904/kjim.2018.172] [Medline: 31480826]
5. Rock CL, Thomson CA, Sullivan KR, Howe CL, Kushi LH, Caan BJ, et al. American Cancer Society nutrition and physical activity guideline for cancer survivors. CA Cancer J Clin 2022;72(3):230-262 [FREE Full text] [doi: 10.3322/caac.21719] [Medline: 35294043]
6. Clinton SK, Giovannucci EL, Hursting SD. The world cancer research fund/american institute for cancer research third expert report on diet, nutrition, physical activity, and cancer: impact and future directions. J Nutr 2020;150(4):663-671 [FREE Full text] [doi: 10.1093/jn/nxz268] [Medline: 31758189]
7. Baughman C, Norman K, Mukamal K. Adherence to american cancer society nutrition and physical activity guidelines among cancer survivors. JAMA Oncol 2024;10(6):789-792. [doi: 10.1001/jamaoncol.2024.0470] [Medline: 38635238]
8. Kaur H, Pisu M, Pekmezi DW, Rogers LQ, Martin MY, Fontaine KR, et al. How healthy are the diets of cancer survivors? Characteristics of those most at risk and opportunities for improvement. J Natl Compr Canc Netw 2025;23(6):e257012. [doi: 10.6004/jnccn.2025.7012] [Medline: 40359987]
9. Demark-Wahnefried W, Aziz NM, Rowland JH, Pinto BM. Riding the crest of the teachable moment: promoting long-term health after the diagnosis of cancer. J Clin Oncol 2005;23(24):5814-5830 [FREE Full text] [doi: 10.1200/JCO.2005.01.230] [Medline: 16043830]
10. Xu J, Hoover RL, Woodard N, Leeman J, Hirschey R. A systematic review of dietary interventions for cancer survivors and their families or caregivers. Nutrients Dec 2023;16(1):56 [FREE Full text] [doi: 10.3390/nu16010056] [Medline: 38201886]
11. Sremanakova J, Sowerbutts AM, Todd C, Cooke R, Burden S. Systematic review of behaviour change theories implementation in dietary interventions for people who have survived cancer. Nutrients 2021;13(2):612. [doi: 10.3390/nu13020612]
12. Jung Y, Chung J, Son H. Physical activity interventions for colorectal cancer survivors: a systematic review and meta-analysis of randomized controlled trials. Cancer Nurs 2021;44(6):E414-E428 [FREE Full text] [doi: 10.1097/NCC.0000000000000888] [Medline: 34694086]

XSL•FO
RenderX

13. Ficarra S, Thomas E, Bianco A, Gentile A, Thaller P, Grassadonio F, et al. Impact of exercise interventions on physical fitness in breast cancer patients and survivors: a systematic review. Breast Cancer 2022;29(3):402-418 [FREE Full text] [doi: 10.1007/s12282-022-01347-z] [Medline: 35278203]
14. Williams V, Brown N, Becks A, Pekmezi D, Demark-Wahnefried W. Narrative review of web-based healthy lifestyle interventions for cancer survivors. Ann Rev Res 2020;5(4):555670 [FREE Full text] [doi: 10.19080/arr.2020.05.555670] [Medline: 33294850]
15. Lavoie A, Dubé V. Web-based interventions to promote healthy lifestyles for older adults: scoping review. Interact J Med Res 2022;11(2):e37315 [FREE Full text] [doi: 10.2196/37315] [Medline: 35998024]
16. Dee EC, Muralidhar V, Butler SS, Yu Z, Sha ST, Mahal BA, et al. General and health-related internet use among cancer survivors in the United States: a 2013-2018 cross-sectional analysis. J Natl Compr Canc Netw 2020;18(11):1468-1475. [doi: 10.6004/jnccn.2020.7591] [Medline: 33152707]
17. Holmes MM. Why people living with and beyond cancer use the internet. Integr Cancer Ther 2019;18:1534735419829830 [FREE Full text] [doi: 10.1177/1534735419829830] [Medline: 30741026]
18. Williams VA, Brown NI, Johnson R, Ainsworth MC, Farrell D, Barnes M, et al. A Web-based lifestyle intervention for cancer survivors: feasibility and acceptability of survivorSHINE. J Cancer Educ 2022;37(6):1773-1781 [FREE Full text] [doi: 10.1007/s13187-021-02026-x] [Medline: 34061334]
19. People Designs, Inc.. SURVIVORSHINE. URL: https://survivorshine.org/ [accessed 2025-05-29]
20. Pekmezi D, Crane T, Oster R, Rogers L, Hoenemeyer T, Farrell D, et al. Rationale and methods for a randomized controlled trial of a dyadic, web-based, weight loss intervention among cancer survivors and Partners: the DUET study. Nutrients 2021;13(10):3141 [FREE Full text] [doi: 10.3390/nu13103472] [Medline: 34684474]
21. Demark-Wahnefried W, Oster RA, Crane TE, Rogers LQ, Cole WW, Kaur H, et al. Results of DUET: a web-based weight loss randomized controlled feasibility trial among cancer survivors and their chosen partners. Cancers (Basel) 2023;15(5):1577 [FREE Full text] [doi: 10.3390/cancers15051577] [Medline: 36900368]
22. Kaur H, Pavela G, Pekmezi DW, Rogers LQ, Cole WW, Parrish KB, et al. Dietary barriers appear to influence the effects of a dyadic web-based lifestyle intervention on caloric intake and adiposity: a mediation analysis of the DUET trial. Nutrients 2023;15(23):4918 [FREE Full text] [doi: 10.3390/nu15234918] [Medline: 38068776]
23. Health Behavior and Health Education: Theory, Research, and Practice. 2nd ed. San Francisco: Jossey-Bass; 1997.
24. Bandura A. Health promotion from the perspective of social cognitive theory. Psychol Health 1998;13(4):623-649. [doi: 10.1080/08870449808407422]
25. Adobe Acrobat Sign computer software. Version 6.0. Adobe Inc. 2020. URL: https://acrobat.adobe.com/us/en/sign.html [accessed 2026-01-23]
26. People Designs, Inc. duet4health. URL: https://duet4health.org [accessed 2025-05-29]
27. Kelley HH. The "stimulus field" for interpersonal phenomena: the source of language and thought about interpersonal events. Pers Soc Psychol Rev 1997;1(2):140-169. [doi: 10.1207/s15327957pspr0102_3] [Medline: 15647123]
28. Lewis MA, McBride CM, Pollak KI, Puleo E, Butterfield RM, Emmons KM. Understanding health behavior change among couples: an interdependence and communal coping approach. Soc Sci Med 2006;62(6):1369-1380. [doi: 10.1016/j.socscimed.2005.08.006] [Medline: 16146666]
29. Demark-Wahnefried W, Jones LW, Snyder DC, Sloane RJ, Kimmick GG, Hughes DC, et al. Daughters and Mothers Against Breast Cancer (DAMES): main outcomes of a randomized controlled trial of weight loss in overweight mothers with breast cancer and their overweight daughters. Cancer 2014;120(16):2522-2534 [FREE Full text] [doi: 10.1002/cncr.28761] [Medline: 24804802]
30. Articulate G. Articulate Storyline [computer software]. Version 360. Articulate Global, LLC. 2017. URL: https://www.articulate.com [accessed 2026-01-14]
31. Frankenfield DC, Rowe WA, Smith J, Cooney R. Validation of several established equations for resting metabolic rate in obese and nonobese people. J Am Diet Assoc 2003;103(9):1152-1159. [doi: 10.1016/s0002-8223(03)00982-9] [Medline: 12963943]
32. Fitbit Inspire 2 [wearable activity tracker]. Google LLC. URL: https://www.fitbit.com/inspire [accessed 2026-01-14]
33. Krebs-Smith SM, Pannucci TE, Subar AF, Kirkpatrick SI, Lerman JL, Tooze JA, et al. Update of the healthy eating index: HEI-2015. J Acad Nutr Diet 2018;118(9):1591-1602 [FREE Full text] [doi: 10.1016/j.jand.2018.05.021] [Medline: 30146071]
34. Automated self-administered 24-hour dietary assessment tool. National Cancer Institute. 2025. URL: https://epi.grants.cancer.gov/asa24/ [accessed 2026-01-23]
35. Rogers LQ, McAuley E, Anton PM, Courneya KS, Vicari S, Hopkins-Price P, et al. Better exercise adherence after treatment for cancer (BEAT Cancer) study: rationale, design, and methods. Contemp Clin Trials 2012;33(1):124-137 [FREE Full text] [doi: 10.1016/j.cct.2011.09.004] [Medline: 21983625]
36. ActiGraph wGT3X-BT Activity Monitor. ActiGraph LLC. URL: https://theactigraph.com/actigraph-wgt3x-bt [accessed 2025-05-29]
37. Amireault S, Godin G, Lacombe J, Sabiston CM. Validation of the Godin-Shephard Leisure-Time Physical Activity Questionnaire classification coding system using accelerometer assessment among breast cancer survivors. J Cancer Surviv 2015;9(3):532-540. [doi: 10.1007/s11764-015-0430-6] [Medline: 25666749]

XSL•FO
RenderX

38. Inc. Zoom. Version 5.15.5. Zoom Video Communications, Inc. 2025. URL: https://zoom.us/ [accessed 2026-01-23]
39. Hoenemeyer TW, Cole WW, Oster RA, Pekmezi DW, Pye A, Demark-Wahnefried W. Test/retest reliability and validity of remote vs. in-person anthropometric and physical performance assessments in cancer survivors and supportive partners. Cancers (Basel) 2022;14(4):1075 [FREE Full text] [doi: 10.3390/cancers14041075] [Medline: 35205823]
40. SAS Software, Version 9.4. SAS Institute Inc. 2025. URL: https://www.sas.com/ [accessed 2025-05-29]
41. Williams V, Brown N, Moore JX, Farrell D, Perumean-Chaney S, Schleicher E, et al. Web-Based lifestyle interventions for survivors of cancer: usability study. JMIR Form Res 2022;6(2):e30974 [FREE Full text] [doi: 10.2196/30974] [Medline: 35188468]
42. Paxton RJ, Hajek R, Newcomb P, Dobhal M, Borra S, Taylor WC, et al. A lifestyle intervention via email in minority breast cancer survivors: randomized parallel-group feasibility study. JMIR Cancer 2017;3(2):e13 [FREE Full text] [doi: 10.2196/cancer.7495] [Medline: 28935620]
43. Van Blarigan EL, Kenfield S, Chan J, Van Loon K, Paciorek A, Zhang L, et al. Feasibility and acceptability of a web-based dietary intervention with text messages for colorectal cancer: a randomized pilot trial. Cancer Epidemiol Biomarkers Prev 2020;29(4):752-760 [FREE Full text] [doi: 10.1158/1055-9965.EPI-19-0840] [Medline: 31941707]
44. van den Berg SW, Peters EJ, Kraaijeveld JF, Gielissen MF, Prins JB. Usage of a generic web-based self-management intervention for breast cancer survivors: substudy analysis of the BREATH trial. J Med Internet Res 2013;15(8):e170 [FREE Full text] [doi: 10.2196/jmir.2566] [Medline: 23958584]
45. Fredericks S, Martorella G, Catallo C. A systematic review of web-based educational interventions. Clin Nurs Res 2015;24(1):91-113. [doi: 10.1177/1054773814522829] [Medline: 24571963]
46. Chatterjee A, Prinz A, Gerdes M, Martinez S. Digital interventions on healthy lifestyle management: systematic review. J Med Internet Res 2021;23(11):e26931. [doi: 10.2196/26931]
47. Shi Y, Wakaba K, Kiyohara K, Hayashi F, Tsushita K, Nakata Y. Effectiveness and components of web-based interventions on weight changes in adults who were overweight and obese: a systematic review with meta-analyses. Nutrients 2022;15(1):179 [FREE Full text] [doi: 10.3390/nu15010179] [Medline: 36615836]
48. Carr RM, Prestwich A, Kwasnicka D, Thøgersen-Ntoumani C, Gucciardi DF, Quested E, et al. Dyadic interventions to promote physical activity and reduce sedentary behaviour: systematic review and meta-analysis. Health Psychol Rev 2019;13(1):91-109 [FREE Full text] [doi: 10.1080/17437199.2018.1532312] [Medline: 30284501]
49. John JC, Ho J, Raber M, Basen-Engquist K, Jacobson L, Strong LL. Dyad and group-based interventions in physical activity, diet, and weight loss: a systematic review of the evidence. J Behav Med 2024;47(3):355-373. [doi: 10.1007/s10865-023-00457-z] [Medline: 38017250]
50. Graham SA, Stein N, Shemaj F, Branch OH, Paruthi J, Kanick SC. Older adults engage with personalized digital coaching programs at rates that exceed those of younger adults. Front Digit Health 2021;3:642818 [FREE Full text] [doi: 10.3389/fdgth.2021.642818] [Medline: 34713112]
51. De Santis KK, Mergenthal L, Christianson L, Busskamp A, Vonstein C, Zeeb H. Digital technologies for health promotion and disease prevention in older people: scoping review. J Med Internet Res 2023;25:e43542 [FREE Full text] [doi: 10.2196/43542] [Medline: 36951896]
52. Bidmon S, Terlutter R. Gender differences in searching for health information on the internet and the virtual patient-physician relationship in Germany: exploratory results on how Men and women differ and why. J Med Internet Res 2015;17(6):e156 [FREE Full text] [doi: 10.2196/jmir.4127] [Medline: 26099325]
53. Manierre MJ. Gaps in knowledge: tracking and explaining gender differences in health information seeking. Soc Sci Med 2015;128:151-158. [doi: 10.1016/j.socscimed.2015.01.028] [Medline: 25618604]
54. Kebede AS, Ozolins L, Holst H, Galvin K. Digital engagement of older adults: scoping review. J Med Internet Res 2022;24(12):e40192. [doi: 10.2196/40192]
55. Pogrebnoy D, Dennett AM, Simpson DB, MacDonald-Wicks L, Patterson AJ, English C. Effects of using websites on physical activity and diet quality for adults living with chronic health conditions: systematic review and meta-analysis. J Med Internet Res 2023;25:e49357 [FREE Full text] [doi: 10.2196/49357] [Medline: 37856187]
56. Linke SE, Dunsiger SI, Gans KM, Hartman SJ, Pekmezi D, Larsen BA, et al. Association between physical activity intervention website use and physical activity levels among Spanish-speaking Latinas: randomized controlled trial. J Med Internet Res 2019;21(7):e13063 [FREE Full text] [doi: 10.2196/13063] [Medline: 31342902]
57. Hachiya J, Oliveira R. The online community role in physical activity interventions. Procedia Computer Science 2024;239:781-789. [doi: 10.1016/j.procs.2024.06.236]
58. Funk KL, Stevens VJ, Appel LJ, Bauck A, Brantley PJ, Champagne CM, et al. Associations of internet website use with weight change in a long-term weight loss maintenance program. J Med Internet Res 2010;12(3):e29 [FREE Full text] [doi: 10.2196/jmir.1504] [Medline: 20663751]
59. Hendrie GA, Baird DL, Brindal E, Williams G, Brand-Miller J, Muhlhausler B. Weight loss and usage of an online commercial weight loss program (the CSIRO Total Wellbeing Diet Online) delivered in an everyday context: five-year evaluation in a community cohort. J Med Internet Res 2021;23(6):e20981 [FREE Full text] [doi: 10.2196/20981] [Medline: 34096869]

XSL•FO
RenderX

60. Dent R, McPherson R, Harper M. Factors affecting weight loss variability in obesity. Metabolism 2020;113:154388. [doi: 10.1016/j.metabol.2020.154388] [Medline: 33035570]

Abbreviations

ACS: American Cancer Society
CSIRO: Commonwealth Scientific and Industrial Research Organisation
DUET: Daughters, Dudes, Mothers, and Others Together
MVPA: moderate to vigorous physical activity
V&F: vegetable and fruit
WCRF/AICR: World Cancer Research Fund/American Institute for Cancer Research
WLM: Weight Loss Maintenance

Edited by A Mavragani; submitted 31.Oct.2025; peer-reviewed by O Ibikunle, E Madondo, A Famotire; comments to author 25.Nov.2025; revised version received 07.Dec.2025; accepted 23.Dec.2025; published 30.Jan.2026.

Please cite as:
Kaur H, Pekmezi D, E Crane T, Farrell D, Q Rogers L, Demark-Wahnefried W
Website Use and Associations With Behavior Change and Weight Loss in Cancer Survivors and Their Partners: Secondary Analysis of a Randomized Controlled Trial
J Med Internet Res 2026;28:e86908
URL: https://www.jmir.org/2026/1/e86908
doi:10.2196/86908
PMID:

©Harleen Kaur, Dori Pekmezi, Tracy E Crane, David Farrell, Laura Q Rogers, Wendy Demark-Wahnefried. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Traditional Rehabilitation Experiences, Unmet Needs, and Perspectives on Virtual Reality–Based Rehabilitation Among Patients With Stroke in China: Qualitative Thematic Analysis and Semistructured Interview Study

Xite Zheng[1], MS; Lu Xing[2], MSc; Haitao Lu[3], MD; Shimeng Hao[2*], PhD; Fen Liu[1*], PhD

[1]Department of Epidemiology and Health Statistics, School of Public Health, Beijing Key Laboratory of Environment and Aging, Capital Medical University, Beijing, China

[2]School of Architecture and Urban Planning, Beijing University of Civil Engineering and Architecture, Beijing, China

[3]Department of Neurorehabilitation, Beijing Bo'ai Hospital, Rehabilitation Research Center, School of Rehabilitation Medicine, Capital Medical University, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Fen Liu, PhD
Department of Epidemiology and Health Statistics
School of Public Health, Beijing Key Laboratory of Environment and Aging
Capital Medical University
No. 10 Xitoutiao, Youanmenwai Street
Fengtai District
Beijing, 100069
China
Phone: 86 010 83911497
Email: liufen05@ccmu.edu.cn

Abstract

Background: Traditional stroke rehabilitation is facing challenges, and virtual reality (VR)–based rehabilitation is a promising solution. However, results from studies focusing on VR-based stroke rehabilitation remain inconsistent, largely due to the use of noncustomized interventions in previous trials.

Objective: To enhance rehabilitation services and inform the development of patient-centered VR rehabilitation systems, this study aimed to (1) explore the experiences and unmet needs of survivors of stroke during current hospital rehabilitation, and (2) examine their perspectives on the use of VR technology in poststroke rehabilitation.

Methods: We conducted a qualitative thematic analysis based on descriptive phenomenology between January and July 2025 at the China Rehabilitation Research Center. Adult patients with a clinical diagnosis of stroke within the past 18 months were eligible. A total of 21 survivors of stroke (mean age, 52.7, SD 17.3 y; men, n=17) were included. Data were collected through face-to-face semistructured interviews, complemented by a short questionnaire on sociodemographic, clinical, and technology-use characteristics. All interviews were audio-recorded, transcribed verbatim, and analyzed using a thematic approach, with thematic saturation used to determine the sample size.

Results: After a stroke, patients experience significant physical and psychological changes. On the one hand, the sudden loss of abilities alters their perceived roles within the family and society; on the other hand, the sharp contrast between their desire for recovery and their current recovery limitations creates substantial psychological pressure. Accepting their condition and rebuilding confidence is a long-term process. Traditional rehabilitation is commonly described as burdensome, monotonous, and lacking continuity after discharge. Although patients desire a better rehabilitation approach and improved outcomes, attitudes toward VR-based rehabilitation vary. Some view VR as a convenient tool, while others express no interest or perceived need for technology-based rehabilitation. Patients indicated that serious games should be diversified to meet different individual and training needs, and should incorporate clearer feedback mechanisms, appropriate scoring systems, adjustable difficulty levels, and progressive game chapters. Functional expectations for VR systems included family involvement, access to personal rehabilitation data, telerehabilitation support, safety monitoring, and technical support.

XSL•FO
RenderX

Conclusions: Stroke rehabilitation services in China require improvement in the appeal of rehabilitation content, patient self-management, and continuity of care. Although patients desire better rehabilitation approaches and outcomes, the effective integration of VR technology must account for factors, such as personal characteristics and preferences, as well as socioeconomic status. Unlike previous studies that primarily examined user experiences with digital technologies or compared rehabilitation outcomes, our research contributes to the literature by linking the challenges and patient needs in conventional rehabilitation with concrete directions for the future design of VR rehabilitation. These insights deepen current understanding of how VR technologies can be meaningfully integrated into stroke care and provide a roadmap for developing patient-centered and culturally responsive VR solutions.

(J Med Internet Res 2026;28:e84532) doi:10.2196/84532

KEYWORDS

stroke patients; rehabilitation; digital health technology; virtual reality; qualitative study; user-requirements; patient-centered design

Introduction

Stroke is a leading cause of death and disability [1]. In 2021, there were an estimated 93.8 million survivors of stroke and 11.9 million new cases globally, including 26.3 million survivors and 4.1 million new cases in China [2]. Half of the survivors with stroke are left disabled, with one-third requiring assistance in daily activities [3]. In addition to physical impairments, cognitive and emotional disturbances, such as memory deficits, aphasia, and depression, are common and further reduce quality of life and hinder social reintegration [4]. Therefore, effective rehabilitation remains a critical component of stroke care in the long term.

While the rehabilitation research is moving toward the exploration of digital health technologies, current stroke rehabilitation in China primarily consists of physical therapy, occupational therapy, and speech therapy, with treatment plans tailored to individual conditions [5]. The demand for stroke rehabilitation services is increasing; however, substantial advances are yet to be made in stroke rehabilitation practice to meet this demand, including improving the effectiveness of current rehabilitation, enhancing adherence, and addressing limited rehabilitation resources, especially in home-based settings [6]. A deeper understanding of the real-world experiences and unmet needs of survivors of stroke throughout the rehabilitation process is critical for informing future service improvement and intervention development.

Virtual reality (VR)–based rehabilitation uses motion tracking, stereoscopic display, and real-time feedback to create immersive, interactive training environments, offering promising solutions for stroke rehabilitation [7]. Previous studies have shown that gamified training improves patient engagement and adherence [8]. Furthermore, the portability of the equipment and internet connectivity can improve access to treatment, promoting telerehabilitation, fostering equity, and supporting patient self-management [9,10]. VR-based rehabilitation has significant implications for improving current stroke care. Nevertheless, findings on its effectiveness remain mixed [11,12]. One major challenge is that many existing VR interventions rely on noncustomized content, such as commercial games, which fail to address the diverse functional needs, abilities, and preferences of survivors of stroke, ultimately limiting therapeutic effect [13]. Moreover, age, socioeconomic status, and educational differences influence individuals' access to and familiarity with VR technology, further shaping intervention acceptability and outcomes [14]. The success of VR-based rehabilitation depends not only on technological efficacy but also on its alignment with the lived acceptability, preferences, and perceived needs of survivors of stroke. Therefore, listening to patients' voices and developing patient-centered VR rehabilitation systems is crucial for addressing the current challenges [15].

Quantitative studies have provided important evidence on the effectiveness and feasibility of stroke rehabilitation interventions; however, they are limited in their ability to capture the rehabilitation experiences, unmet needs, and subjective perspectives of rehabilitation services among survivors of stroke. Qualitative approach and semistructured interviews were widely used to understand human phenomena, including patients' thoughts and experiences [16,17]. Such information is essential for informing the development of patient-centered and culturally appropriate VR-based rehabilitation interventions.

Therefore, this study used qualitative methods to (1) explore the experiences and unmet needs of Chinese survivors of stroke during current hospital rehabilitation and (2) examine their perspectives on the use of VR technology in poststroke rehabilitation. The overarching goal is to enhance rehabilitation services and inform the development of patient-centered VR rehabilitation systems.

Methods

Research Design Overview

This study adopted a qualitative thematic analysis based on descriptive phenomenology to explore the rehabilitation experiences, unmet needs, and perspectives of survivors of stroke on VR-based rehabilitation [17,18]. Data were collected through semistructured interviews and analyzed using thematic analysis. This study was conducted at the China Rehabilitation Research Center (Beijing Boai Hospital), as part of a larger project entitled Development of a Motor and Cognitive Rehabilitation System for Stroke Patients Based on Multisensory Virtual Reality Technology.

Researcher Description

This study was conducted through a collaboration among Capital Medical University, Beijing University of Civil Engineering

XSL•FO
RenderX

and Architecture, and the China Rehabilitation Research Center. The research team consists of university professors, chief physicians, and doctoral and master's students, all of whom have received professional training and possess extensive experience in stroke rehabilitation, qualitative research, and digital health. The primary interviewer and data analyst (XZ and LX) had received formal training in qualitative interviewing and thematic analysis, and possessed previous knowledge of stroke rehabilitation practices and the research landscape in China. XZ also had previous experience conducting and analyzing interviews with public health practitioners and primary and secondary school students. Reflexivity was maintained through regular team discussions, during which researchers reflected on their disciplinary backgrounds and potential assumptions to minimize interpretive bias. These experiences and foundation enabled the research team to successfully conduct this study and ensured the rigor of the methodology.

Participants

The final sample included 21 participants (4 women and 17 men) with a mean age of 52.7 (SD 17.3) years. For most participants (19/21, 90%), the time since stroke onset was less than 6 months. The majority (14/21, 67%) reported daily internet use, and nearly all (19/21, 90%) owned a smartphone or tablet. Detailed participant characteristics are presented in Table 1.

XSL•FO
RenderX

Table 1. Demographic characteristics of patients with stroke at the Department of Neurorehabilitation, China Rehabilitation Research Center (January to July 2025, N=21, qualitative thematic analysis based on descriptive phenomenology).

Patient	Sex	Age (y)	Education level	Marital status	Time since stroke (months)	Stroke type	Perceived health	Smartphone or tablet	Use of the internet
P01	Female	62	Senior high or vocational school	Divorced	0-3	Ischemic	Good	Smartphone	Every day
P02	Male	90	Senior high or vocational school	Widowed	12-18	Ischemic	Fair	No	No
P03	Male	76	Junior high school	Married	0-3	Ischemic	Good	Smartphone	Every day
P04	Female	60	Junior high school	Married	0-3	Ischemic	Poor	No	No
P05	Male	46	Senior high or vocational school	Married	3-6	Ischemic	Fair	Smartphone	Multiple times a week
P06	Male	38	College, University, or above	Divorced	3-6	Ischemic	Fair	Both	Every day
P07	Male	58	Junior high school	Married	3-6	Hemorrhagic	Fair	Smartphone	Multiple times a week
P08	Female	69	Primary school or below	Married	0-3	Ischemic	Fair	Smartphone	Every day
P09	Male	55	College, University, or above	Married	0-3	Ischemic	Good	Smartphone	Every day
P10	Male	46	College, University, or above	Single	3-6	Hemorrhagic	Fair	Smartphone	Every day
P11	Male	30	Senior high or vocational school	Single	3-6	Hemorrhagic	Fair	Both	Multiple times a week
P12	Male	71	Senior high or vocational school	Married	3-6	Hemorrhagic	Good	Both	Multiple times a week
P13	Male	40	College, University, or above	Married	0-3	Ischemic	Fair	Both	Every day
P14	Female	27	College, University, or above	Married	6-12	Hemorrhagic	Fair	Both	Every day
P15	Male	62	Junior high school	Married	0-3	Hemorrhagic	Fair	Smartphone	Every day
P16	Male	38	College, University, or above	Married	3-6	Hemorrhagic	Fair	Smartphone	Multiple times a month
P17	Male	62	Junior high school	Married	3-6	Hemorrhagic	Fair	Both	Every day
P18	Male	32	College, University, or above	Single	0-3	Hemorrhagic	Good	Smartphone	Every day
P19	Male	32	College, University, or above	Married	0-3	Hemorrhagic	Fair	Both	Every day
P20	Male	70	College, University, or above	Married	0-3	Ischemic	Fair	Smartphone	Every day
P21	Male	43	College, University, or above	Married	3-6	Hemorrhagic	Fair	Smartphone	Every day

Researcher–Participant Relationship

The researchers and participants had no previous relationship. At the scheduled interview time, a nurse introduced the researchers to the participants, and the participants felt at ease throughout the process.

Recruitment Process

Physician-researchers screened medical records to identify eligible participants. Patients meeting the inclusion criteria were approached during hospitalization, and those who provided verbal consent were later contacted by trained interviewers to schedule a face-to-face interview. The sample size was determined by the principle of data saturation, defined as the point at which no new themes or insights emerged during analysis [19]. A total of 26 survivors of stroke were interviewed, of whom 5 were excluded due to poor communication ability, leaving 21 participants in the final analysis.

XSL•FO
RenderX

Participant Selection

Participants were recruited through purposive sampling from the Department of Neurorehabilitation at the China Rehabilitation Research Center. Inclusion criteria were (1) age≥18 years, (2) first-ever stroke diagnosed within the past 18 months, and (3) ability to provide written informed consent. The 18-month time frame was chosen to capture both early and later recovery experiences and to examine the durability of these effects [20], while ensuring that participants were physically able to engage with the VR devices safely. To minimize unnecessary exclusion, patients with mild cognitive or communication impairments were permitted to receive support from informal caregivers when communication was slow. Exclusion criteria included inability to speak Chinese or insufficient communication capacity, as determined at the time of interview. Data collection took place between January and July 2025.

Data Collection

Data were collected through face-to-face semistructured interviews following a predefined sequence. At the scheduled interview time, a nurse led 2 researchers (XZ and LX), both trained in qualitative research, to the patient's bedside to confirm the appointment. After confirmation, the nurse left the room. The researchers then introduced themselves to the participant and explained the purpose and significance of the study. The participant then read and voluntarily signed the written informed consent. Participants were subsequently invited to complete 2 rounds of the Fruit Ninja (Halfbrick) game using a Pico 4 Enterprise headset (ByteDance; 1 round with visual access to the surrounding room environment and 1 without). Following the VR experience, participants completed a brief questionnaire collecting demographic and clinical information, including sex, age, marital status, education level, time since stroke, stroke type, self-perceived health, and use of mobile devices and the internet. Interviews were conducted after completion of the VR task and questionnaire, and took place either in the patient's room or a doctor's office to ensure a quiet, clean, and private environment. An interview guide (Multimedia Appendix 1), developed by the research team based on the study objectives and a review of relevant literature and refined through pilot interviews with 2 patients (data from the pilot interviews were not included in the final analysis), was used to facilitate the interviews. Interviews began with questions about stroke history and rehabilitation experiences, followed by exploration of perceived challenges, unmet needs, and suggestions for improving poststroke care. Participants were then asked to reflect on their VR experience, including attitudes, preferences, and perceptions of usability. Follow-up questions probed barriers to VR use and elicited recommendations for improvement. Open-ended prompts ("Is there anything else you would like to add regarding today's topic?") were used to encourage additional insights. Interviews lasted between 16 and 40 minutes (25 minutes on average), were audio-recorded, and transcribed verbatim.

Data-Analytic Strategies

Data were analyzed using thematic analysis following Braun and Clarke's 6-phase framework [21,22]. All interviews were transcribed within 24 hours by 2 researchers (XZ and LX). The first author completed a verbatim transcription, and the second author checked each transcript against the audio recordings to ensure accuracy and enhance familiarity with the dataset. Furthermore, the 2 researchers (XZ and LX) independently read and coded all of the transcripts using an inductive, data-driven approach, focusing on participants' rehabilitation experiences, needs, and perspectives on VR technology. After independent coding, the 2 researchers compared their coding and discussed similarities and discrepancies. Discrepancies were resolved through in-depth discussion, during which the researcher explained their coding decisions with reference to the original transcripts. When consensus could not be reached initially, the relevant data segments were re-examined, and codes were refined or merged as appropriate until agreement was achieved. Codes were then organized into potential themes and subthemes, which were reviewed for consistency and relevance across transcripts. Recoding was performed when necessary. Themes were iteratively refined and finalized in consultation with the broader research team. All qualitative data were managed and analyzed using NVivo software (version 1.2; Lumivero). Selected quotations were translated into English using a forward-backward translation process.

Methodological Integrity

Methodological integrity was ensured by aligning the study design, data collection, and analytic approach with the research aims. The study was conducted at the China Rehabilitation Research Center in Beijing, China, a leading institution with more than 30 years of experience in rehabilitation medicine and research, serving patients from across the country. A qualitative design informed by descriptive phenomenology was used to capture the rehabilitation experiences, unmet needs, and perspectives of survivors of stroke on VR-based rehabilitation. Semistructured interviews allowed participants to express their experiences, needs, and perspectives in their own words while ensuring coverage of key topics relevant to the research questions. Data collection continued until thematic saturation was reached, defined as the point at which no new themes emerged from successive interviews. Interviews were audio-recorded, transcribed verbatim, and analyzed using an inductive thematic analysis approach. Furthermore, 2 researchers (XZ and LX) independently coded the transcripts and engaged in iterative discussions to resolve discrepancies and refine the coding framework. Strategies to enhance methodological rigor, including reflexivity, trustworthiness, and ethical conduct, were integrated throughout the research process. These aspects are described in detail in the following sections: Researcher Description, Trustworthiness, and Ethical Considerations. This study adhered to the American Psychological Association's reporting standards for qualitative research [23] and the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist (Multimedia Appendix 2) [24].

Ethical Considerations

This study was reviewed and approved by the Medical Ethics Committee of Capital Medical University, Beijing, China (approval Z2025SY006). All procedures complied with the principles of the Declaration of Helsinki. Participants were

XSL•FO
RenderX

informed of the study's purpose, procedures, potential risks, and their right to withdraw at any time. Written informed consent was obtained before participation. To ensure privacy and confidentiality, data were deidentified through pseudonymization. Contact information was stored separately from research data. No images or personally identifiable information were included in the manuscript or supplementary materials. Participants received a small token of appreciation in the form of daily necessities valued at 20-30 Chinese yuan, equivalent to US $3-US $4.

Trustworthiness

Trustworthiness was established following Guba and Lincoln's criteria [25], encompassing credibility, transferability, dependability, and confirmability. Credibility was enhanced through timely verbatim transcription and cross-checking within 24 hours to ensure accurate and faithful representation of participants' narratives. All interviews and analyses were conducted by 2 researchers (XZ and LX), facilitating deep engagement with the data. Transferability was supported by providing detailed descriptions of the study setting, recruitment context, and participant characteristics, enabling readers to assess the relevance of the findings to other contexts. Dependability and confirmability were strengthened through transparent documentation of data collection and analysis procedures. Data were independently coded by 2 researchers, with discrepancies resolved through discussion and consensus in consultation with the wider research team. Thematic analysis followed Braun and Clarke's 6-phase framework, reducing potential researcher bias and enhancing analytic reliability.

Results

Overview

Thematic analysis generated six overarching themes: (1) changes following stroke and self-reconstruction, (2) effective yet challenging traditional rehabilitation, (3) unmet needs in the rehabilitation journey, (4) attitudes toward VR-based rehabilitation, (5) recommendations for serious game design, and (6) suggested features of VR systems (Figure 1). A summary of themes and representative quotations is provided in Multimedia Appendix 3.

XSL•FO
RenderX

Figure 1. Themes and subthemes identified during data analysis of patients with stroke at the Department of Neurorehabilitation, China Rehabilitation Research Center (January and July 2025, N=21, qualitative thematic analysis based on descriptive phenomenology). VR: virtual reality.

Changes Following Stroke and Self-Reconstruction

Shifts in Personal and Social Roles

All participants experienced varying degrees of physical impairment following stroke. Some reported relatively mild deficits, such as limb weakness or fatigue, whereas others described severe hemiplegia accompanied by cognitive impairments, including confusion and aphasia. These functional losses led to substantial disruptions in their independence and resulted in major shifts in their roles within both personal and social contexts.

> *Overnight, my entire left side became paralyzed. I can't do any work now. […] I've brought a lot of inconvenience to my family and close friends. That's what pains me the most.* [P09]
>
> *Half of my body doesn't move, so I rely on others for everything. At first, I couldn't even speak or recognize people. I couldn't do anything I was supposed to do. I became a burden to my family and brought misfortune to them.* [P13]

Inner Psychological Conflicts

The sudden loss of physical abilities and the associated role changes had profound psychological consequences. Participants described intense feelings of uselessness and a strong desire to recover bodily functions, which collectively created significant emotional pressure. Many reported experiencing anxiety, guilt, low self-esteem, irritability, and even despair.

> *After I got sick, I was constantly worried and sad. […] I felt irritable and depressed all the time.* [P08]
>
> *Right after it happened, I felt useless and even had thoughts of ending my life. But you cannot die even if you want to, because you cannot move. I couldn't*

XSL•FO
RenderX

even pick up a knife to hurt myself. It was a feeling of utter despair. [P19]

Coming to Terms With Changes

Despite these challenges, participants described conscious efforts to adopt a more positive mindset and gradually accept their condition. They highlighted the importance of patience, emotional stability, and realistic expectations to support long-term rehabilitation. Family encouragement and professional guidance played key roles in fostering acceptance and maintaining motivation.

Don't rush it, take your time. Since you already have this illness, you need to stay calm. Whether it takes one or two years, even four, it is still the same condition. Recovery is slow, so you need patience. Being too anxious is not good for your recovery; once your mindset becomes tense, nothing goes well. [P18]

Later on, with my family's encouragement and the doctors' support, I gained much more confidence in my recovery. You see, now I'm doing quite well. [P19]

Effective Yet Challenging Traditional Rehabilitation

Engagement Barriers

Many participants described current hospital rehabilitation as densely scheduled, repetitive, and lacking engagement. The monotony and high intensity of daily sessions often resulted in insufficient rest, leading to physical exhaustion and psychological strain. Over time, these factors diminished motivation, reduced adherence, and required substantial willpower to maintain participation.

The training is truly monotonous. I think it really depends on one's willpower and endurance. [P05]

My current rehabilitation includes OT, PT, robotic arm training, acupuncture, and massage. Except for massage, all the training is boring, just repeating the same movements every day. [P06]

I have trained every weekday, only resting on weekends. My daughter had to push me in a wheelchair, and we are always rushing to get there on time; otherwise, I won't be allowed to join the session. […] It's exhausting. Sometimes my legs are swollen by the time I return to the ward. [P08]

Lack of Continuity

Most participants reported performing little or no structured rehabilitation after hospital discharge. The absence of professional supervision, appropriate equipment, and adequate space at home posed major barriers to continuing rehabilitation outside the clinical environment.

At home, I can only go out for a walk and treat that as exercise. [P01]

After I went home, I didn't train at all. I think many people are like this. How can you train at home? There's no equipment, no one to guide you, and no conditions for proper training. [P03]

Perceived Benefits

Despite the challenges encountered, most participants expressed overall satisfaction with the effectiveness of rehabilitation. They also highlighted the emotional support provided by health care professionals, which contributed to their sense of comfort, motivation, and hope.

The therapists here are very experienced. The environment in the rehabilitation department is alsoheartwarming. Someone will hold my hand or greet me kindly with comforting words, this really touches me. [P02]

It's been almost a month, and I'll be discharged in a few days. My leg has improved a lot. [P03]

During the rehabilitation process, the doctors constantly encouraged me, saying things like 'You're making progress' or 'You're much better than a few days ago.' [P04]

Unmet Needs in the Rehabilitation Journey

Desire for Better Approaches

Several participants described transferring to the current hospital specifically to access more advanced technologies and effective rehabilitation interventions. Their accounts reflected a clear recognition of the limitations of previous treatments and a strong willingness to explore innovative strategies to optimize recovery.

This is already the third hospital I have been to. I started at the best hospital in my province, then went to the best local rehabilitation center, and now I am here at what they say is the top rehabilitation center in the country. I don't really know what treatment methods they use here, so I feel the need to try and find out. [P09]

Need for Rehabilitation Knowledge

Given the complexity and long duration of stroke recovery, many participants reported a lack of essential knowledge about their condition and appropriate rehabilitation strategies. They noted that their limited understanding of stroke mechanisms and treatment options hindered communication with health care providers and made it difficult to actively participate in decision-making.

We are not professionals, we don't understand how this illness occurs or what the whole process involves. […] We truly lack knowledge. Now they are giving us these treatments, but we don't know which ones are suitable for us. Even when doctors explain things, we still struggle to understand. [P10]

Demand for Social Respect

Participants emphasized that social acceptance and respect are crucial for psychological resilience, well-being, and successful social integration. They noted that the support and treatment they receive vary greatly both within rehabilitation facilities and in society at large, and expressed a desire to receive the same level of respect outside of rehabilitation institutions.

When I came to this hospital, I noticed that some of the staff members are also people with disabilities

XSL•FO
RenderX

like us. It shows we can still work and support ourselves even if we are disabled. This encourages me not to give up on life or my future. This is unlike the negative comments made by some people in society. The doctors and nurses are all very kind; they never show any prejudice because of our condition. Their respectful communication really helps with psychological healing. [P19]

Attitudes Toward VR-Based Rehabilitation

Praised and Positive Perceptions

Most participants generally expressed a positive attitude toward VR-based rehabilitation. They highlighted that the gamified and immersive nature of VR could make rehabilitation more engaging, increase motivation, and potentially enhance adherence compared with conventional, repetitive training. Participants also noted that VR could help address resource limitations after discharge and improve access by enabling rehabilitation outside the hospital environment.

Once this device is fully developed and widely promoted, it could be placed at the nurses' station for patients to use in rotation. Patients could even purchase one to use at home. [P03]

I think your VR device is very interesting, it's completely virtual. My arm still feels a bit tired afterward, but the experience was enjoyable. […] Compared with the usual hospital treatments like cycling or traditional therapy, which are quite dull, this was much more engaging, though of course I can't complain too much about hospital treatment either. [P20]

Skepticism and Lack of Interest

In contrast, a subset of participants expressed hesitation or disinterest in using VR for rehabilitation. Many preferred conventional, face-to-face therapy due to its direct physical interaction and perceived reliability. Age-related limitations, physical discomfort, and economic concerns were also mentioned as barriers to adoption.

At my age, using computers is difficult. Sometimes I even feel dizzy. [P01]

It depends on how much your device costs. I just tried it and my first impression was good, but if the price is too high, not everyone will be able to afford it, right? [P10]

Honestly, I think what you are doing is great, but I just don't think it's necessary. No matter how good it is, I personally don't need it. [P15]

I still prefer traditional rehabilitation. It feels more intuitive—you can see it and touch it. [P17]

Recommendations for Serious Game Design

Diversified Game Types

Some participants suggested expanding the range of game genres, as people with different characteristics have different preferences. They emphasized that serious games should be tailored to users' age, preferences, and rehabilitation stages. Incorporating culturally familiar or age-appropriate activities, such as calligraphy, chess, or dancing, was viewed as essential to ensure wider appeal and encourage maintained engagement.

I think you should design more types of games. I prefer games set in natural scenery and am not interested in this sword-based game. [P16]

I think age differences matter. For people in their fifties or sixties, slicing fruit may not interest them. You could include chess or similar games that are more suitable for that age group. [P19]

There are gender differences. Women might prefer dancing games. [P20]

Customized Training Content

Although enjoyment was appreciated, participants emphasized that rehabilitation should remain the primary focus of serious games. They recommended developing targeted training modules that correspond to specific functional impairments and rehabilitation goals, such as upper limb strength, hand dexterity, or fine-motor tasks. They suggest creating a library of customizable exercises that clinicians could assign based on individual needs, ensuring both personalization and clinical relevance.

To me, muscle and strength training are most important because I want to walk independently. The game is great, and the virtual environment feels real, but I hope for more specialized training. For example, if I want to practice my hand, then exercises targeting finger movement or using chopsticks should be available. You could develop a game library with many options, and clinicians can select the ones that suit us. [P05]

Desired Game Functions

Participants also proposed functional enhancements to improve the usability and therapeutic value of serious games. Adjustable difficulty levels and sequential game chapters were recommended to accommodate different impairment levels and to maintain long-term engagement. Scoring systems were considered valuable for monitoring progress and promoting self-motivation. Additionally, immersive environments and clear feedback, through audiovisual prompts, tactile feedback, or visual cues, were highlighted as necessary features, especially for individuals with sensory limitations.

Is the vibration on the controller too weak? I didn't feel any vibration. You could add prompts or voice cues like 'please touch' or use arrows such as 'move the controller here.' That would be very helpful. [P04]

I like a fully virtual environment because the scene feels realistic, and it gives you a sense of immersion. You're not distracted by the outside environment, so you stay more focused during training. [P05]

If you add a scoring system—like getting 1000 points today and 1500 tomorrow—I would feel like I'm making progress. By the third day, I'd aim for an even

XSL•FO
RenderX

> *higher score. That kind of thing motivates players.* [P10]
>
> *Gradually increasing the difficulty, such as by speeding up or offering continuous stages, helps people adapt and stay interested in continuing.* [P11]

Suggested Features of VR Systems

Family Involvement

Participants noted that VR rehabilitation devices resembled home gaming consoles, which made them especially appealing to family members, particularly children. Participants expressed a desire for family-connected gameplay, suggesting that allowing relatives to join the games could enhance enjoyment and strengthen emotional bonds during the recovery process.

> *This looks like a children's game, [...] can my grandson play it?* [P01]
>
> *This is similar to the game consoles we have at home. Doesn't Xbox have a game like this? […] I go home mainly to spend time with my daughter. She loves these kinds of games, so if we could play together, or if she could play with me in some way, that would be wonderful.* [P21]

Access to Personal Rehabilitation Data

Participants showed strong interest in accessing personal rehabilitation data directly through the VR system. They emphasized that personal medical and training records are crucial for understanding their treatment history, clarifying rehabilitation goals, and monitoring functional changes. Such transparency was viewed as essential for enhancing engagement, supporting self-management, and giving patients a sense of control over their rehabilitation journey.

> *Can you create a rehabilitation profile or progress tracking system? It looks like a game on the surface, but it's actually rehabilitation training. You need to know what you're training while playing and what you should do next.* [P05]

Telerehabilitation Support

Participants recommended integrating educational materials and professional guidance into the VR system to support rehabilitation outside the hospital. They recognized the potential of VR technology in providing remote access to medical services and ensuring continuity of rehabilitation.

> *If, during the rehabilitation process, doctors could see my progress and send me some rehabilitation-related suggestions, that would definitely be meaningful.* [P02]
>
> *The equipment in hospitals is very expensive and too large to use at home. If VR can be applied to home-based rehabilitation, I think it's an excellent method. It's lightweight, you can use it at home, and it still provides effective rehabilitation. That's very good.* [P19]

Safety Monitoring

Safety emerged as a significant concern. Participants recommended incorporating physiological monitoring, such as blood pressure or fatigue indicators, and automated alerts to prevent overexertion or adverse events. They suggested implementing time limits, real-time feedback on training intensity, and emergency warnings. These proposals emphasize the need for built-in safeguards to protect vulnerable users, particularly older adults and those with cardiovascular conditions.

> *You have to limit the game time. What if someone plays all night? That would be dangerous.* [P12]
>
> *You could add some features focused on health monitoring. Since we mainly track blood pressure, you could include blood pressure measurements and provide specific values. That would make it medically useful and safer. Otherwise, someone might get too excited while playing and suddenly feel dizzy or faint. You could set an alert when blood pressure reaches a certain level, reminding the player to stop.* [P19]

Technical Support

Although participants found the current VR device easy to operate, many expressed concerns about potential technical issues or malfunctions, especially when used at home. They emphasized the need for accessible customer support and reliable maintenance to ensure confidence in long-term use. Suggested options included 24-hour service hotlines, troubleshooting through widely used communication platforms such as WeChat (Tencent Holdings Limited), and guaranteed repair.

> *If I buy it and take it home, I need help when the device has problems, something like a customer service number or a WeChat account.* [P03]
>
> *If I am going to use it long-term, it needs to have warranty services. If it breaks, someone must be responsible for repairing it.* [P12]

Discussion

Principal Findings

This qualitative study explored the experiences and unmet needs of survivors of stroke during current hospital rehabilitation, as well as their perspectives on integrating VR technology into stroke rehabilitation. Thematic analysis revealed six overarching themes: (1) changes following stroke and self-reconstruction, (2) effective yet challenging traditional rehabilitation, (3) unmet needs in the rehabilitation journey, (4) attitudes toward VR-based rehabilitation, (5) recommendations for serious game design, and (6) suggested features of VR systems. Unlike previous studies that primarily examined user experiences with digital technologies or compared rehabilitation outcomes, our research contributes to the literature by linking the challenges and patient needs in conventional rehabilitation with concrete directions for the future design of VR-based rehabilitation.

Survivors of stroke experience profound shifts in their roles within families and society, and the stark contrast between their

XSL•FO
RenderX

desire for recovery and the harsh realities of impairment generates substantial psychological pressure. Support from family members and health care professionals plays a critical role in helping patients regain confidence and rebuild their lives. Traditional rehabilitation programs are often described as burdensome and monotonous, and the near absence of home rehabilitation presents another major challenge. Although patients desire more engaging rehabilitation approaches and better functional outcomes, future VR rehabilitation devices must account for differences in personal characteristics and preferences, as well as socioeconomic factors. Moreover, to maximize both effectiveness and acceptability, VR rehabilitation systems should incorporate serious games tailored to diverse personal and training needs, and integrate functions such as family interaction, progress tracking, telehealth support, safety monitoring, and technical assistance.

Comparison With Previous Work

The level of exercise rehabilitation among survivors of stroke remains far below guideline recommendations, often characterized by the "three lows": low initiative, poor adherence, and reduced willingness to participate [26]. From a Chinese cultural perspective, rehabilitation often relies on rest or passive treatments, such as the concept of "Deqi" in traditional Chinese medicine [27], which may weaken motivation for active, repetitive training. Furthermore, access to rehabilitation services remains uneven across regions, and high out-of-pocket costs further reduce opportunities for maintained participation [28]. In our study, traditional rehabilitation exercises were frequently described as burdensome, monotonous, and lacking continuity. Previous studies have shown that repetitive movements without feedback can diminish motivation among patients with stroke [29]. After hospital discharge, many participants experienced a sharp decline in available resources and limited professional guidance [30]. In contrast, VR-based rehabilitation is reported to be more engaging than traditional face-to-face therapy with a therapist [31]. The portability of VR devices and the gamified approach to intervention can enhance motivation and extend rehabilitation into the community and home environments, thereby improving continuity of care [32].

Consistent with previous studies, survivors of stroke in our sample expressed varying attitudes toward VR-based rehabilitation [33]. Particularly in the context of traditional Chinese culture, many individuals still view games as purely for entertainment, which may lead to skepticism about game-based rehabilitation. This underscores the importance of individual readiness, digital literacy, and personal preferences in shaping acceptance and effectiveness of VR-based interventions. Notably, even participants who were initially skeptical of VR acknowledged the value of advanced hospitals and innovative rehabilitation technologies and expressed a desire for improved rehabilitation programs. Previous research indicates that patients are more likely to adopt digital health tools when they are simple, intuitive, and user-friendly [34], whereas poorly designed equipment can lead to negative experiences and reduced adherence [35]. Importantly, firsthand experience with the benefits of digital health technologies plays a key role in promoting acceptance and maintained use [36]. Therefore, when implementing VR in clinical practice, it is essential not only to clearly communicate the practical benefits of VR-based rehabilitation but also to address patients' concerns related to usability, safety, and reliability.

Serious games have been increasingly evaluated in clinical rehabilitation protocols, with studies reporting improved patient engagement and functional outcomes [37]. However, understanding how patients with diverse characteristics perceive different game elements is essential for maximizing acceptance, adherence, and therapeutic effectiveness. In our study, participants emphasized the importance of offering game types aligned with personal interests and selecting training content tailored to individual needs. In addition, due to decreased reaction times and slower cognitive processing, serious games should incorporate clearer feedback mechanisms, appropriate scoring systems to enhance motivation, and adjustable difficulty levels with progressively structured game chapters to accommodate varying motor abilities.

Survivors of stroke in our study also identified several desirable features for VR-based rehabilitation systems, including family involvement, access to personal rehabilitation data, telerehabilitation support, safety monitoring, and technical assistance. Family engagement has long been shown to enhance motivation, emotional support, and adherence in stroke rehabilitation [38]. Participants highlighted that family support from family members is critical for psychological healing, yet complex emotional challenges are often overlooked despite their substantial influence on rehabilitation outcomes [39]. A VR gaming platform that allows family members to participate together can, while providing entertainment, strengthen the bonds between family members and even promote intergenerational communication. Additionally, progress tracking can enhance self-efficacy, support goal-setting, and promote active self-management [40]. In our study, many participants expressed strong enthusiasm about VR's potential to increase engagement, maintain motivation, and improve accessibility, particularly for home-based rehabilitation. In a home setting, patients can independently complete structured therapeutic exercises through VR tutorials delivered via head-mounted displays and controllers [41,42]. Remote therapist supervision can be supported through real-time data transmission, video consultations, or automated progress reports generated by the system [43]. This potential became particularly evident during the pandemic, when remote technologies played an essential role in maintaining continuity of care. Built-in safety features, such as heart rate and blood pressure monitoring, usage timers, and fall detection alarms, can further enhance user motivation and reduce risks during home use [44]. Finally, reliable technical support is crucial for enhancing patient trust and ensuring long-term adherence.

The successful integration of VR technology into stroke rehabilitation, as well as equitable access to such innovations, requires not only technological advancement but also governmental support and a cultural shift within clinical practice toward embracing digital rehabilitation tools. Achieving this goal calls for interdisciplinary collaboration among rehabilitation clinicians, neuroscientists, engineers, designers, and patients to develop evidence-based systems and clinical protocols that ensure both effectiveness and usability [45]. Special attention

XSL•FO
RenderX

should be given to adapting VR interventions for diverse populations, particularly older adults with limited digital literacy and individuals living in resource-constrained environments. Critical challenges, including digital literacy, cost, security, and accessibility in rural areas [14,46], may be addressed through 2 complementary pathways: designing simple, intuitive systems supported by clinical validation, and reducing production costs through governmental funding while implementing pilot programs in community and rural settings. Additionally, VR systems must account for the evolving priorities of survivors of stroke throughout their recovery [47]. A phased, holistic intervention approach is needed to support the restoration of physical, psychological, and social functions. Using Maslow's hierarchy of needs as a conceptual framework [48,49], VR system design should adapt its focus at different recovery stages. In the early phase, rehabilitation focuses on stabilization, preventing complications, and gently initiating motor function while offering psychological support. In the subacute phase, it shifts to intensive task-specific training and cognitive and motivational interventions. In the long-term community or home setting, rehabilitation emphasizes daily-life reintegration, supported by family involvement, peer support, and continued physical and psychosocial training.

Strengths and Limitations

The strengths of this study include allowing participants to experience VR rehabilitation equipment before each interview, which enhanced their understanding and enabled them to provide more informed feedback. Additionally, we implemented rigorous qualitative methods appropriate for an exploratory study, allowing patients to express their rehabilitation experiences, unmet needs, and perspectives on VR freely and comprehensively. However, several limitations should be acknowledged. First, the sample was drawn from a single rehabilitation hospital and was predominantly male, which may limit generalizability. Second, all interviews were conducted in Chinese and subsequently translated into English, which may have introduced subtle linguistic or interpretive biases. Third, most participants were within 6 months of stroke onset, which may restrict insights into long-term or home-based rehabilitation experiences. Fourth, the relatively short interview duration and insufficient depth in probing questions may have constrained the richness of the data. Finally, the absence of longitudinal follow-up prevented examination of how patients' needs, self-management behaviors, and acceptance of VR may change over time. Future research should incorporate multicenter longitudinal studies with more diverse samples to explore how patients' needs and attitudes evolve across recovery stages.

Conclusions

This study explored the rehabilitation experiences and unmet needs of survivors of stroke, as well as their perspectives of VR-based rehabilitation. Survivors of stroke undergo significant changes in their family and social roles due to the sudden loss of physical abilities. Their strong desire for recovery stands in stark contrast to the slow progress they experience in reality, which causes them immense psychological stress. Traditional rehabilitation is described as arduous and monotonous, with little to no support for home-based rehabilitation. Consequently, stroke rehabilitation services in China require improvements in the attractiveness of rehabilitation content, patient self-management, and continuity of care. Although patients desire better rehabilitation approaches and outcomes, the effective integration of VR technology into stroke care must take into account factors, such as personal characteristics and preferences, as well as socioeconomic status. Future VR rehabilitation systems should incorporate serious games that address diverse personal and training needs and integrate functions, such as family involvement, progress tracking, telemedicine support, safety protection, and technical assistance. These findings deepen understanding of how VR technologies can be meaningfully embedded into stroke rehabilitation and offer a roadmap for developing patient-centered and culturally responsive VR solutions.

Acknowledgments

The authors gratefully acknowledge the support of all the participants for sharing their experiences and perspectives in this study.

Funding

This study was sponsored by Beijing Nova Program (20240484500). The funders had no role in the study design, data analysis or interpretation, manuscript drafting, or the decision to submit for publication.

Data Availability

The data supporting this study are available upon reasonable request and subject to approval by the corresponding authors. Requests should be directed to SH and FL.

Authors' Contributions

Conceptualization: SH, FL, XZ
Data curation: XZ, LX.
Formal analysis: XZ, LX
Funding acquisition: SH, FL
Investigation: XZ, LX, HL

XSL•FO
RenderX

Methodology: SH, FL, XZ
Project administration: SH, FL, XZ
Supervision: SH, FL
Validation: XZ
Writing – original draft: XZ
Writing – review & editing: XZ, FL
The co-corresponding author, SH, can be contacted at the School of Architecture and Urban Planning, Beijing University of Civil Engineering and Architecture, No. 1 Zhanlanguan Road, Xicheng District, Beijing 100044, China (Email: haoshimeng@bucea.edu.cn).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sample Interview Guide.
[DOCX File , 16 KB - jmir_v28i1e84532_app1.docx]

Multimedia Appendix 2
COREQ checklist.
[DOCX File , 206 KB - jmir_v28i1e84532_app2.docx]

Multimedia Appendix 3
Summary of Themes and Representative Quotations.
[DOCX File , 23 KB - jmir_v28i1e84532_app3.docx]

References

1. GBD 2023 DiseaseInjuryRisk Factor Collaborators. Burden of 375 diseases and injuries, risk-attributable burden of 88 risk factors, and healthy life expectancy in 204 countries and territories, including 660 subnational locations, 1990-2023: a systematic analysis for the Global Burden of Disease Study 2023. Lancet 2025;406(10513):1873-1922 [FREE Full text] [doi: 10.1016/S0140-6736(25)01637-X] [Medline: 41092926]
2. GBD 2021 Stroke Risk Factor Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021. The Lancet Neurology 2024;23(10):973-1003 [FREE Full text] [doi: 10.1016/S1474-4422(24)00369-7] [Medline: 39304265]
3. Markus HS. Reducing disability after stroke. International journal of stroke 2022;17(3):249-250. [doi: 10.1177/17474930221080904] [Medline: 35191348]
4. Pinter D, Fandler-Höfler S, Fruhwirth V, Berger L, Bachmaier G, Horner S, et al. Relevance of cognition and emotion for patient-reported quality of life after stroke in working age: an observational cohort study. Front Neurol 2022;13:869550 [FREE Full text] [doi: 10.3389/fneur.2022.869550] [Medline: 35547373]
5. Liu L. Chinese Stroke Association. Clinical Management Guidelines for Cerebrovascular Diseases in China. 2 ed. In: The People's Health Press. Jan Swasthya Abhiyan: People's Health Press; 2020.
6. Stinear CM, Lang CE, Zeiler S, Byblow WD. Advances and challenges in stroke rehabilitation. The Lancet Neurology 2020;19(4):348-360. [doi: 10.1016/S1474-4422(19)30415-6] [Medline: 32004440]
7. Lu W, Shi M, Liu L, Wang S, Deng W, Ma Y, et al. Effect of virtual reality-based therapies on lower limb functional recovery in stroke survivors: systematic review and meta-analysis. J Med Internet Res 2025;27:e72364 [FREE Full text] [doi: 10.2196/72364] [Medline: 40737641]
8. Dawson J, Nee R, Ramirez C, Reyes S, Sanchez D, Sukhadia T, et al. Gamification in mHealth apps for rehabilitation: protocol for a scoping review. JMIR Res Protoc 2025;14:e63600 [FREE Full text] [doi: 10.2196/63600] [Medline: 40294403]
9. Verma A, Towfighi A, Brown A, Abhat A, Casillas A. Moving towards equity with digital health innovations for stroke care. Stroke 2022;53(3):689-697 [FREE Full text] [doi: 10.1161/STROKEAHA.121.035307] [Medline: 35124973]
10. Nikolaev VA, Nikolaev AA. Recent trends in telerehabilitation of stroke patients: A narrative review. NeuroRehabilitation 2022;51(1):1-22. [doi: 10.3233/NRE-210330] [Medline: 35527574]
11. Chen J, Or CK, Chen T. Effectiveness of using virtual reality-supported exercise therapy for upper extremity motor rehabilitation in patients with stroke: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2022;24(6):e24111 [FREE Full text] [doi: 10.2196/24111] [Medline: 35723907]
12. Laver KE, Lange B, George S, Deutsch JE, Saposnik G, Chapman M, et al. Virtual reality for stroke rehabilitation. Cochrane Database Syst Rev 2025;6(6):CD008349. [doi: 10.1002/14651858.CD008349.pub5] [Medline: 40537150]

XSL•FO
RenderX

13. Cucinella SL, de Winter JCF, Grauwmeijer E, Evers M, Marchal-Crespo L. Towards personalized immersive virtual reality neurorehabilitation: a human-centered design. J Neuroeng Rehabil 2025;22(1):7 [FREE Full text] [doi: 10.1186/s12984-024-01489-5] [Medline: 39833912]
14. National Institute for Health and Care Excellence: Clinical Guidelines. In: Stroke rehabilitation in adults. London: National Institute for Health and Care; 2023.
15. Ghosh R, Khan N, Migovich M, Tate JA, Maxwell CA, Newhouse PA, et al. Engaging older adults and staff in the Co-Design and evaluation of socially assistive robot and virtual reality activities for Long-Term Care: user-centered study. JMIR Aging 2025;8:e75288 [FREE Full text] [doi: 10.2196/75288] [Medline: 41330571]
16. PLoS Medicine Editors T. Qualitative research: understanding patients' needs and experiences. PLoS Med 2007;4(8):e258 [FREE Full text] [doi: 10.1371/journal.pmed.0040258] [Medline: 17760496]
17. Sundler AJ, Lindberg E, Nilsson C, Palmér L. Qualitative thematic analysis based on descriptive phenomenology. Nurs Open 2019;6(3):733-739 [FREE Full text] [doi: 10.1002/nop2.275] [Medline: 31367394]
18. Bradshaw C, Atkinson S, Doody O. Employing a qualitative description approach in health care research. Glob Qual Nurs Res 2017;4:2333393617742282 [FREE Full text] [doi: 10.1177/2333393617742282] [Medline: 29204457]
19. Fontanella BJB, Ricas J, Turato ER. [Saturation sampling in qualitative health research: theoretical contributions]. Cad Saude Publica 2008;24(1):17-27 [FREE Full text] [doi: 10.1590/s0102-311x2008000100003] [Medline: 18209831]
20. Tu WJ, Wang LD. China stroke surveillance report 2021. Military Medical Research 2023;10(1):33 [FREE Full text] [doi: 10.1186/s40779-023-00463-x] [Medline: 37468952]
21. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006;56(8):77-101. [doi: 10.1007/s10597-020-00591-x] [Medline: 32100154]
22. Braun V, Clarke V. Successful Qualitative Research. London: SAGE Publications; 2013.
23. Levitt HM, Bamberg M, Creswell JW, Frost DM, Josselson R, Suárez-Orozco C. Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA Publications and Communications Board task force report. Am Psychol 2018;73(1):26-46 [FREE Full text] [doi: 10.1037/amp0000151] [Medline: 29345485]
24. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
25. Guba EG, Lincoln Y. Competing paradigms in qualitative research. Handbook of qualitative research 1994;2:105. [doi: 10.2307/3121684]
26. Sun X, Shi Y, Liu C, Wang S, Li D, Zhu X, et al. Experience and needs of stroke patients in physical rehabilitation: a systematic review and meta-synthesis. BMC Health Serv Res 2025;25(1):1062. [doi: 10.1186/s12913-025-13213-7] [Medline: 40790477]
27. Hui KK, Sporko TN, Vangel MG, Li M, Fang J, Lao L. Perception of deqi by chinese and american acupuncturists: a pilot survey. Chin Med 2011;6(1):2 [FREE Full text] [doi: 10.1186/1749-8546-6-2] [Medline: 21251312]
28. Liang Y, Lu P. Medical insurance policy organized by Chinese government and the health inequity of the elderly: longitudinal comparison based on effect of New Cooperative Medical Scheme on health of rural elderly in 22 provinces and cities. Int J Equity Health 2014;13:37 [FREE Full text] [doi: 10.1186/1475-9276-13-37] [Medline: 24884944]
29. Ham Y, Shin J. Efficiency and usability of a modified pegboard incorporating computerized technology for upper limb rehabilitation in patients with stroke. Top Stroke Rehabil 2023;30(4):333-341. [doi: 10.1080/10749357.2022.2058293] [Medline: 35348031]
30. Lin S, Wang C, Wang Q, Xie S, Tu Q, Zhang H, et al. The experience of stroke survivors and caregivers during hospital-to-home transitional care: A qualitative longitudinal study. Int J Nurs Stud 2022;130:104213. [doi: 10.1016/j.ijnurstu.2022.104213] [Medline: 35378465]
31. Anwar N, Karimi H, Ahmad A, Gilani SA, Khalid K, Aslam AS, et al. Virtual reality training using nintendo wii games for patients with stroke: randomized controlled trial. JMIR Serious Games 2022;10(2):e29830 [FREE Full text] [doi: 10.2196/29830] [Medline: 35699989]
32. Everard G, Declerck L, Detrembleur C, Leonard S, Bower G, Dehem S. New technologies promoting active upper limb rehabilitation after stroke: an overview and network meta-analysis. Eur J Phys Rehabil Med 2022;58(4):530-548. [doi: 10.23736/s1973-9087.22.07404-4]
33. Chen J, Or CK, Li Z, Yeung EHK, Chen T. Perceptions of patients with stroke regarding an immersive virtual reality-based exercise system for upper limb rehabilitation: questionnaire and interview study. JMIR Serious Games 2025;13:e49847 [FREE Full text] [doi: 10.2196/49847] [Medline: 39742513]
34. Bally ELS, Cheng D, van Grieken A, Ferri Sanz M, Zanutto O, Carroll A, et al. Patients' perspectives regarding digital health technology to support self-management and improve integrated stroke care: qualitative interview study. J Med Internet Res 2023;25:e42556 [FREE Full text] [doi: 10.2196/42556] [Medline: 37014677]
35. Nedergård H, Sandlund M, Häger CK, Palmcrantz S. Users' experiences of intensive robotic-assisted gait training post-stroke - "a push forward or feeling pushed around?". Disabil Rehabil 2023;45(23):3861-3868 [FREE Full text] [doi: 10.1080/09638288.2022.2140848] [Medline: 36342771]

XSL•FO
RenderX

36. de Veer AJE, Peeters JM, Brabers AEM, Schellevis FG, Rademakers JJDJM, Francke AL. Determinants of the intention to use e-Health by community dwelling older people. BMC Health Serv Res 2015;15:103 [FREE Full text] [doi: 10.1186/s12913-015-0765-8] [Medline: 25889884]
37. Vieira C, Ferreira da Silva Pais-Vieira C, Novais J, Perrotta A. Serious game design and clinical improvement in physical rehabilitation: systematic review. JMIR Serious Games 2021;9(3):e20066 [FREE Full text] [doi: 10.2196/20066] [Medline: 34554102]
38. Eriksson G, Söderhielm K, Erneby M, Guidetti S. Family members' experiences of a person-centered information and communication technology-supported intervention for stroke rehabilitation (F@ce 2.0): qualitative analysis. JMIR Rehabil Assist Technol 2025;12:e69878 [FREE Full text] [doi: 10.2196/69878] [Medline: 40315427]
39. Kim JS. Post-stroke mood and emotional disturbances: pharmacological therapy based on mechanisms. J Stroke 2016;18(3):244-255 [FREE Full text] [doi: 10.5853/jos.2016.01144] [Medline: 27733031]
40. Chen Y, Li K, Lin C, Hung P, Lai H, Wu C. The effect of sequential combination of mirror therapy and robot-assisted therapy on motor function, daily function, and self-efficacy after stroke. Sci Rep 2023;13(1):16841 [FREE Full text] [doi: 10.1038/s41598-023-43981-3] [Medline: 37803096]
41. Toh SFM, Fong KNK, Gonzalez PC, Tang YM. Application of home-based wearable technologies in physical rehabilitation for stroke: a scoping review. IEEE Trans. Neural Syst. Rehabil. Eng 2023;31:1614-1623. [doi: 10.1109/tnsre.2023.3252880]
42. Laver KE, Adey-Wakeling Z, Crotty M, Lannin NA, George S, Sherrington C. Telerehabilitation services for stroke. Cochrane Database Syst Rev 2020;1(1):CD010255 [FREE Full text] [doi: 10.1002/14651858.CD010255.pub3] [Medline: 32002991]
43. Benadduci M, Franceschetti C, Marziali RA, Frese S, Sándor PS, Tombolesi V, et al. An integrated virtual reality-based telerehabilitation platform to support recovery and maintenance of functional abilities among older adults: protocol for a usability and acceptability study. JMIR Res Protoc 2025;14:e68358 [FREE Full text] [doi: 10.2196/68358] [Medline: 40729692]
44. Bargeri S, Scalea S, Agosta F, Banfi G, Corbetta D, Filippi M, et al. Effectiveness and safety of virtual reality rehabilitation after stroke: an overview of systematic reviews. EClinicalMedicine 2023;64:102220 [FREE Full text] [doi: 10.1016/j.eclinm.2023.102220] [Medline: 37745019]
45. Wankhede NL, Koppula S, Ballal S, Doshi H, Kumawat R, Raju S, et al. Virtual reality modulating dynamics of neuroplasticity: innovations in neuro-motor rehabilitation. Neuroscience 2025;566:97-111. [doi: 10.1016/j.neuroscience.2024.12.040] [Medline: 39722287]
46. Terp R, Kayser L, Lindhardt T. Older patients' competence, preferences, and attitudes toward digital technology use: explorative study. JMIR Hum Factors 2021;8(2):e27005 [FREE Full text] [doi: 10.2196/27005] [Medline: 33988512]
47. Purton J, Sim J, Hunter SM. Stroke survivors' views on their priorities for upper-limb recovery and the availability of therapy services after stroke: a longitudinal, phenomenological study. Disabil Rehabil 2023;45(19):3059-3069 [FREE Full text] [doi: 10.1080/09638288.2022.2120097] [Medline: 36111388]
48. Carpenito-Moyet LJ. Maslow's Hierarchy of Needs--revisited. Nurs Forum 2003;38(2):3-4. [doi: 10.1111/j.1744-6198.2003.tb01204.x] [Medline: 12894625]
49. Alam Z. Making international trainees feel welcome: don't forget Maslow's hierarchy of needs. BMJ (Clinical research ed) 2023;382:2087. [doi: 10.1136/bmj.p2087] [Medline: 37699640]

Abbreviations

VR: virtual reality
COREQ: Consolidated Criteria for Reporting Qualitative Research

Edited by S Brini; submitted 21.Sep.2025; peer-reviewed by S Wang, L Bulle, A Kaunnil; comments to author 25.Nov.2025; revised version received 08.Jan.2026; accepted 09.Jan.2026; published 02.Feb.2026.

Please cite as:
Zheng X, Xing L, Lu H, Hao S, Liu F
Traditional Rehabilitation Experiences, Unmet Needs, and Perspectives on Virtual Reality–Based Rehabilitation Among Patients With Stroke in China: Qualitative Thematic Analysis and Semistructured Interview Study
J Med Internet Res 2026;28:e84532
URL: https://www.jmir.org/2026/1/e84532
doi:10.2196/84532
PMID:

XSL•FO
RenderX

©Xite Zheng, Lu Xing, Haitao Lu, Shimeng Hao, Fen Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of the Mobile-Based Diabetes Little Helper Video Intervention on Medication Adherence Among Older Adults Living With Type 2 Diabetes Mellitus, Henan, China: Randomized Controlled Trial

Meng Wang[1], PhD; Khuan Lee[1], PhD; Hui Zhu Thew[2], MMed; Siti Noorkhairina Sowtali[3], PhD; Qiuhuan Jiang[4], MPH; Zhengyan Zhao[5], BSN; Yu Liu[6], MNS; Poh Ying Lim[7], PhD

[1]Department of Nursing, Faculty of Medicine & Health Sciences, Universiti Putra Malaysia, Selangor, Malaysia
[2]Department of Family Medicine, Faculty of Medicine & Health Sciences, Universiti Putra Malaysia, Selangor, Malaysia
[3]Department of Professional Nursing Studies, Kulliyyah of Nursing, International Islamic University Malaysia, Pahang, Malaysia
[4]Henan Provincial People's Hospital, Zhengzhou, China
[5]Zhengzhou Seventh People's Hospital, Zhengzhou, China
[6]The Second Mobile Contingent Hospital of the Chinese People's Armed Police Forces, Wuxi, China
[7]Department of Community Health, Faculty of Medicine & Health Sciences, Universiti Putra Malaysia, Selangor, Serdang, Malaysia

Corresponding Author:
Poh Ying Lim, PhD
Department of Community Health, Faculty of Medicine & Health Sciences
Universiti Putra Malaysia
Jalan Universiti 1
Selangor, Serdang, 43400
Malaysia
Phone: 60 397692950
Email: pohying_my@upm.edu.my

Abstract

Background: Medication adherence is vital for older adults living with type 2 diabetes mellitus (T2DM), but it remains low and needs improvement. Current interventions have limited effectiveness, while video-based interventions show promising potential for enhancing adherence.

Objective: To evaluate the impact of the “Diabetes Little Helper” video intervention, developed based on the information–motivation–behavioral skills model, on improving medication adherence in older adults living with T2DM in Henan.

Methods: This parallel-group randomized controlled trial was conducted in 2 hospitals in Zhengzhou, involving 68 patients from each hospital. The intervention group (IG) received standard care plus the video intervention for one month, while the control group (CG) received only standard care. The primary outcome was medication adherence, and secondary outcomes included medication knowledge, attitude, behavior, belief, and social support. Data were collected at baseline, postintervention, and at 3-month follow-up. Intention-to-treat analysis and the last observation carried forward method were applied for missing data, with the generalized estimating equation model used for effect assessment (P<.05 considered statistically significant).

Results: The average age of participants in the IG was 67.5 (SD 8.0) years, while in the CG, the average age was 66.0 (SD 7.0) years. Sex distribution was nearly identical, with 51.5% (n=35) of participants in the IG and 50.0% (n=34) in the CG being male. After the intervention, the IG and CG had retention rates of 95.6% (n=65) and 97.1% (n=66), respectively. At the 3-month follow-up, the retention rates for the IG and CG were 92.6% (n=63) and 92.2% (n=62), respectively. Both postintervention (β=4.956, 95% CI 3.702-6.210, P<.001) and at the 3-month follow-up (β=3.691, 95% CI 2.379-5.003, P<.001), medication adherence score for the IG was significantly higher than that for the CG. In addition, the IG showed significant improvements in secondary outcome, with medication knowledge (β=11.592, 95% CI 6.923-16.260, P<.001), attitude (β=5.467, 95% CI 4.531-6.763, P<.001), behavior (β=4.176, 95% CI 3.220-5.133, P<.001), and belief (β=2.882, 95% CI 1.990-3.775, P<.001) compared with the CG postintervention. However, there was no statistically significant difference in the secondary outcome of social support (β=0.008, 95% CI –1.834 to 2.011, P=.928).

XSL•FO
RenderX

Conclusions: The Diabetes Little Helper video intervention effectively improved medication adherence in older adults living with T2DM in Henan, highlighting the potential of digital health tools for managing chronic diseases in older adult populations.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2400083282; https://www.chictr.org.cn/showprojEN.html?proj=214847

(J Med Internet Res 2026;28:e78731) doi:10.2196/78731

KEYWORDS

older adults; video; type 2 diabetes mellitus; medication adherence; randomized controlled trial

Introduction

With the continuous improvement of economic levels, changes in lifestyle, and the accelerated process of population aging, the prevalence of diabetes mellitus (DM) has been steadily increasing, becoming a significant global public health challenge [1,2]. In 2021, the number of people aged 20-79 years with DM worldwide was approximately 536.6 million, and it is projected to rise to 783.2 million by 2045 [3]. The older adults, due to factors such as physiological decline, metabolic imbalance, and the coexistence of chronic diseases, represent a high-risk group for DM. In 2019, the number of people aged 65 years and older with DM worldwide reached 135.6 million, and it is expected to increase to 276.2 million by 2045 [4].

As one of the countries with the highest burden of DM, the trend of DM prevalence in China is consistent with the global pattern. The number of people aged 20-79 years with DM in China increased from 88.7 million in 2020 to over 108.4 million [5]. Among these, type 2 diabetes mellitus (T2DM) is the predominant form, accounting for 90%-95% of all DM cases, with the older adults being the primary affected group [6-9]. T2DM significantly shortens the lifespan of patients and can lead to severe complications, including cardiovascular diseases, kidney disease, retinopathy, and neuropathy, potentially resulting in organ failure [10-15]. Due to the long duration of the disease and the presence of multiple complications, older adults living with T2DM are at higher risk of developing complications such as diabetic foot, stroke, and nephropathy [16,17]. Medication treatment is crucial in the management of T2DM, and its effectiveness is highly dependent on patient adherence to medication, that is, whether patients follow the prescribed treatment regimen consistently [18,19]. Studies have shown that patients with high medication adherence not only experience more stable disease control but also significantly reduce mortality, hospitalization rates, and health care costs [20-22]. However, in China, more than half of older adults living with T2DM experience poor medication adherence [23,24], which severely limits the long-term effectiveness of DM management and urgently calls for effective intervention strategies. Research indicates that various factors contribute to poor medication adherence, including age [25], economic burden [26], and multiple medications [27]. To improve medication adherence among older adults, various intervention strategies have been attempted both domestically and internationally, including health management based on multidisciplinary collaborative models [28], face-to-face health education [29], mobile app support [30], and motivational interviewing [31]. However, these approaches generally face challenges such as high implementation costs, complex operations, and a strong reliance on health care resources. Furthermore, most lack systematic theoretical models, making them difficult to adapt to older adults with lower educational levels [32,33]. The information–motivation–behavioral skills (IMB) model provides a theoretical foundation for explaining and predicting individual health behaviors and has been widely applied in chronic disease self-management research [34,35]. Existing studies have confirmed that the IMB model can effectively predict medication adherence among older adults living with DM [36,37], offering scientific guidance for the development of intervention strategies. In recent years, video intervention has emerged as a new method of health communication. Existing studies have shown that videos not only effectively capture attention but also facilitate the understanding and retention of information [38]. Particularly within older adult populations, multimedia interventions have been demonstrated to overcome the limitations of traditional educational methods, improving the acquisition of health knowledge and enhancing patients' self-management capabilities [39,40]. Videos, delivered through convenient devices such as mobile phones, enable patients to view them anytime and anywhere, further reinforcing learning outcomes and improving medication adherence [41]. However, there is currently no research using video interventions to improve medication adherence among older adults living with T2DM. Therefore, exploring the application of IMB-based video interventions in this population and their potential to enhance medication adherence holds significant theoretical and practical importance. Based on this, we hypothesize that the Diabetes Little Helper video intervention will significantly improve medication adherence compared with standard care.

Methods

Ethical Considerations

This study received ethical approval from the Institutional Review Board of Universiti Putra Malaysia (JKEUPM-2023-1279) and Henan Provincial People's Hospital (20240107), and was registered with the Chinese Clinical Trial Registry (ChiCTR2400083282). Before data collection, the researcher obtained approval from the hospital and coordinated with relevant departments to ensure informed consent from participants and guide them to scan the QR code to follow the "Diabetes Little Helper" WeChat public account. All participant information was kept strictly confidential, and the questionnaire data were anonymized after collection to ensure that participants' identities could not be directly or indirectly identified. This study does not involve any monetary compensation or incentives for participants.

XSL•FO
RenderX

Study Design and Setting

This study was a parallel-group, cluster-randomized controlled trial (RCT) recruited older adults from 2 hospitals in Zhengzhou, Henan Province, China (Henan Provincial People's Hospital and Zhengzhou Central Hospital). Both hospitals are large, tertiary-level public general hospitals with an annual outpatient volume exceeding 1.1 million visits. Each hospital has over 70 clinical departments, including endocrinology, respiratory medicine, and cardiology, offering comprehensive medical services and a sufficient patient base for recruitment. The study spanned from April 2024 to June 2025, encompassing participant recruitment (from April to June 2024), intervention delivery, data collection, analysis, and manuscript preparation. The study was conducted and reported in accordance with the CONSORT (Consolidated Standards of Reporting Trials; checklist provided in Multimedia Appendix 1) guidelines. Further details of the study design and methodology are available in the published protocol in *BMJ Open* [42]. This study followed all specified procedures without any deviations from the protocol.

Participants

This study involved older adults diagnosed with T2DM who attended endocrinology outpatient clinics at 2 selected hospitals. Eligible individuals were identified from the outpatient population of each hospital and enrolled through simple random sampling. Participants must meet the following inclusion criteria: aged 60 years or older, diagnosed with T2DM [43], Chinese nationality, and able to understand Chinese, currently receiving at least one oral antidiabetic medication with a hemoglobin HbA_{1c} level of ≥7.0%, no cognitive impairment, and either the participant or their family member owns a smartphone and is able to use WeChat. Exclusion criteria include having severe comorbidities affecting major organs such as the heart, lungs, or kidneys, impairment in communication abilities, or participating in another study.

This cluster RCT study involves 2 hospitals, which were randomly assigned in a 1:1 allocation ratio to the control group (CG) and intervention group (IG) using a random number generator. Both hospitals used simple random sampling and a parallel design, recruiting 68 older adults living with T2DM. Randomization was conducted by the research coordinator using a computer-generated random number sequence, with concealed allocation to prevent selection bias. The CG received standard nursing care, while the IG received additional video intervention. Due to the visible nature of the video intervention, the study was conducted as an open-label trial, meaning blinding was not feasible for the participants or executors. However, evaluators and statistical analysts remained blind to minimize bias. To minimize contamination, participants were selected from separate hospitals, and communication between the groups was restricted. After a 3-month follow-up, the CG received the same video intervention as the IG to ensure equal treatment.

The sample size for this study was calculated using the 2-mean sample size formula [44], based on the mean and SD of medication adherence for both CG and IG at postintervention and 3-month follow-up (29.15, SD 3.10; 31.27, SD 3.35; and 26.25, SD 2.01; 30.19, SD 2.41) [45,46]. Assuming a 2-sided test with a type I error rate (α) of 0.05, a power (1-β) of 80%, an 80% retention rate, and a design effect of 1.5, the required total sample size was calculated to be 68 participants per group.

Diabetes Little Helper Intervention Development

A health education intervention program based on the IMB theory model was developed for this study. The research team conducted 3 rounds of expert consensus meetings involving a multidisciplinary expert panel, which included 2 nursing experts with extensive clinical experience and expertise in video production, a biostatistician with a background in statistics for health interventions, and a family medicine specialist with expertise in DM management. The team collaboratively established a comprehensive educational framework covering 8 core topics related to DM medication management. These topics include proper medication storage and transportation, methods for identifying medication quality, disease management strategies, the importance of medication adherence, key medication precautions, common misconceptions about medications, self-management empowerment for older adults living with DM, and guidelines for blood glucose monitoring. For each topic, a 2-5 minute educational animation was developed using the "WanCai Animation Master" tool to visualize professional medical knowledge into easy-to-understand dynamic images [47]. Table 1 shows the Diabetes Little Helper video intervention topic and plan.

Table 1. Topics and intervention plan of the Diabetes Little Helper video program for older adults living with type 2 diabetes mellitus in Henan, China.

Number	Week	Day of the week	IMB[a]	Topic	Video time
1	One	Tuesday	I	Storage & Traveling with Hypoglycemic Medications	3 min 31 s
2	One	Friday	I	How to Check Your Medications	3 min 20 s
3	Two	Tuesday	I	What to Do During Sick Days	2 min 57 s
4	Two	Friday	M	Importance of On-Time Medication	2 min 15 s
5	Three	Tuesday	M	Precautions for Taking Hypoglycemic Medication	1 min 53 s
6	Three	Friday	M	Correct Understanding of Hypoglycemic Medications	2 min
7	Four	Tuesday	B	Self-Empowerment & Support for Older Adults Living with DM	4 min 30 s
8	Four	Friday	B	Regular Blood Glucose Monitoring	3 min 6 s

[a]IMB: information–motivation–behavioral skills.

XSL•FO
RenderX

Five nursing experts were invited to assess the content validity of the program based on the standards developed by Siti Noorkhairina et al [48]. These standards were specifically designed for evaluating video-based educational materials and cover various aspects of content quality, including media characteristics (title relevance and audio clarity), educational design (plot logic and age-appropriate language), and technical specifications (appropriate duration and information density). The Siti Noorkhairina et al standards provide a comprehensive framework for assessing the effectiveness of educational videos, ensuring that the content is clear, relevant, and engaging for the target audience. Experts rated 24 evaluation items based on these criteria to determine the content validity of the program. A 4-point Likert scale was used for the ratings. The Item-level Content Validity Index (I-CVI) for each of the 24 items was calculated by determining the proportion of experts who rated the item as relevant, with scores of 3 or 4 indicating relevance. For calculation, scores of 3 and 4 were recoded as "1.00" (indicating relevant), and scores of 1 and 2 were recoded as "0" (indicating non-relevant). The I-CVI for each item was then calculated as the proportion of experts who rated the item as "relevant". A value of 1.00 for the I-CVI indicates unanimous agreement among the experts that each item was deemed relevant. The Scale-level Content Validity Index (S-CVI) was calculated by averaging the I-CVI scores across all items. A value of 1.00 for the S-CVI indicates that the overall scale was deemed fully valid by all experts. These results suggest that the content of the program was considered highly relevant and appropriate by all expert raters. Ten older adults living with T2DM were invited to participate in a face validity test, which focused on information reception indicators such as language comprehensibility, visual appeal, and pacing control. Based on patient feedback, the audiovisual elements of the videos were optimized.

Diabetes Little Helper Intervention Implementation

The intervention was implemented using a "standard care + digital health education" model, delivering content through the WeChat public account "Diabetes Little Helper." During the 4-week intervention period, educational videos were scheduled to be released at 8 AM every Tuesday and Friday, with a total of 8 episodes. Patients could watch the videos at their convenience.

To ensure that patients successfully watched the intervention videos and to improve intervention quality, the research team took the following measures: (1) real-time monitoring of video viewership to track whether patients watched the videos on time, ensuring timely completion and optimizing the intervention strategy, (2) weekly WeChat reminders for patients to watch the new videos, along with contact information for researchers to provide immediate assistance, ensuring patients received the necessary information and encouraging participation, (3) weekly phone follow-ups to check patients' video-watching progress, particularly focusing on those who missed videos, researchers would inquire about reasons for nonviewing and provide personalized help to ensure patient involvement, (4) personalized support for patients with special circumstances or frequent video nonviewing, including simple viewing guides and remote technical support to resolve technical issues, ensuring patients could watch the videos without difficulty.

Development of a Questionnaire to Evaluate Medication Adherence, Knowledge, Attitude, Behavior, Belief, and Social Support

A self-developed questionnaire was used to collect data from participants, covering 8 aspects. A detailed overview of all outcome measures used in this study can be found in the study protocol [42]. The measures analyzed in this manuscript are described below.

Sociodemographic and Disease-Related Characteristics

This study collected sociodemographic characteristics (such as age, sex, and marital status) as well as disease-related variables (such as the duration of T2DM, family history, and adverse medication reaction).

Medication Adherence

Medication adherence is assessed using the Self-Efficacy for Appropriate Medication Use Scale [49], which includes 13 items, rated on a 3-point Likert scale, with a total score range of 13 to 39. Higher scores indicate better medication adherence. The Cronbach α value of the scale is 0.934 [49,50].

Medication Knowledge, Attitude, and Behavior

The medication knowledge, attitude, and behavior questionnaires are used to measure patients' knowledge, attitudes, and behaviors regarding medication. The test-retest reliability for each dimension ranges from 0.854 to 0.920, with an overall test-retest reliability of 0.868 [51]. The knowledge questionnaire consists of 17 items, with a total score range of 0 to 17, where higher scores indicate a greater level of medication knowledge. The medication attitude questionnaire includes 10 items, with a total score range of 10 to 50, where higher scores reflect a more positive attitude toward safe medication use. The medication behavior questionnaire consists of 13 items, with scores ranging from 9 to 40, where higher scores indicate safer medication behaviors.

Medication Belief

The medication belief questionnaire is used to assess the personal medication beliefs of older adults living with DM [52]. The questionnaire consists of 10 items, divided into 2 dimensions, medication necessity and medication concern, with scores ranging from 1 to 5. Higher scores indicate stronger medication beliefs. The Cronbach α value for the questionnaire is 0.724, with Cronbach α values of 0.749 and 0.796 for the 2 dimensions, respectively.

Social Support

The social support questionnaire is used to assess the support patients receive across multiple dimensions [53]. This scale includes 4 dimensions, with a total of 19 items, and the total score ranges from 19 to 95. Higher scores indicate higher levels of social support among the older adults. The questionnaire demonstrates good reliability and validity, with a Cronbach α value of 0.803 and a content validity index of 0.958 [53].

XSL•FO
RenderX

To ensure the content validity of the questionnaire assessing medication adherence, knowledge, attitude, behavior, belief, and social support, 5 experts reviewed it to confirm that the content aligned with the research objectives and that the questions were relevant and representative of the study. Based on expert feedback, some questions were adjusted to ensure they accurately measured the intended content. The I-CVI and the S-CVI of the questionnaire were both calculated to be 1, indicating strong content validity. Additionally, to assess the face validity of the questionnaire, it was presented to 10 older adults living with T2DM, who provided feedback on the clarity and appropriateness of the language used in the questions. Based on their feedback, the wording of some questions was revised for better clarity and suitability. The revised questionnaire was found to be relevant, user-friendly, and suitable for this population.

Data Collection

Data collection occurred at 3 time points: baseline, postintervention, and 3-month follow-up. The questionnaire was available in both paper and electronic formats. Baseline data were collected between April and June 2024. Researchers conducted on-site checks to ensure the quality of questionnaire completion, confirming that all questions were fully answered and met the required standards. For older adult participants who had difficulty filling out or understanding the questions, standardized instructions were provided to help them interpret the content and complete the questionnaire correctly. During the postintervention and follow-up periods, the electronic questionnaire was sent to participants via a WeChat official account, with researchers assisting through WeChat or phone calls to ensure a high response rate and data quality. After collecting the questionnaires, researchers reviewed and processed any inconsistent or missing data to ensure completeness. Data were entered by 2 research assistants and imported into SPSS (version 29.0; IBM Corp) software for analysis. To reduce entry errors, all data were proofread twice after input to ensure consistency with the original questionnaires. Data from the electronic questionnaires were directly imported into SPSS.

Data Analysis

Data analysis was performed using SPSS Statistics, following the intention-to-treat principle. Missing data were handled using the last observation carried forward (LOCF) method. A significance level was set at $P<.05$. Histograms with normal curves were used to assess the normality of continuous variables. For normally distributed variables, data are expressed as mean (SD), while not normally distributed variables are presented as median and IQR. Categorical variables are reported as frequencies and percentages (%).

For baseline group comparisons, the Pearson chi-square test and chi-square test with continuity correction were applied for categorical variables. Independent samples *t* test and Mann-Whitney *U* test were used for continuous variables. To assess the intervention effects, generalized estimating equation models with an unstructured working correlation matrix were used to analyze changes in medication adherence, knowledge, attitude, behavior, belief, and social support scores. The models compared the changes between the IG and CG at baseline, postintervention, and at the 3-month follow-up, with time, group, and their interaction as independent variables. The interaction term was used to assess whether there were different changes between the groups at different time points. Significant differences between the groups were adjusted for as covariates in the final model.

Results

Participant Flow

The IG consisted of 68 participants randomly selected from a pool of 110 eligible patients in one hospital, while the CG included 68 participants randomly selected from a pool of 100 eligible patients in the other hospital. Compared with the web-based questionnaire, 26 (29.0%) participants in the IG and 20 (38.0%) participants in the CG chose to complete the paper questionnaire. Both groups had an initial retention rate of 100% (n=68). After the intervention, 3 participants dropped out of the IG, leading to a retention rate of 95.6% (n=65). In the CG, one participant did not participate due to health issues, and another withdrew, resulting in a retention rate of 97.1% (n=66). Following a 3-month follow-up, 2 participants left the IG, giving a retention rate of 92.6% (n=63). In the CG, 3 participants withdrew, and one could not be contacted, leading to a retention rate of 92.2% (n=62). Figure 1 shows the data collection flow according to the CONSORT guidelines.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram for a cluster randomized controlled trial on the effectiveness of the Diabetes Little Helper intervention video among older adults living with type 2 diabetes mellitus.

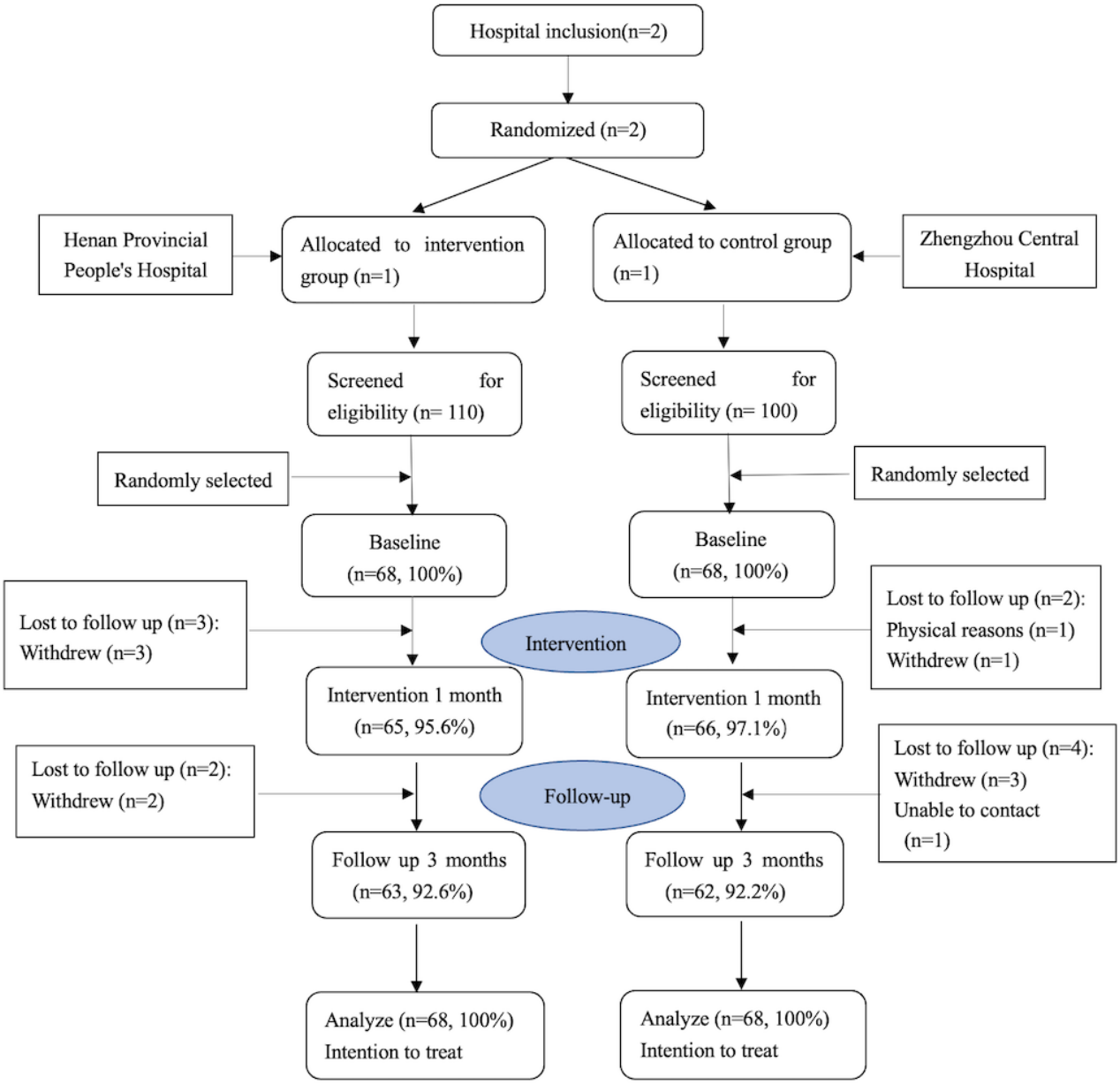

Participant Characteristics

Sociodemographic and Disease-Related Characteristics Between the IG and CG at Baseline

At baseline, a comparison of sociodemographic and disease-related characteristics between the IG and CG showed no statistically significant differences (P>.05). The sex distribution was nearly identical, with 51.5% (n=35) in the IG and 50.0% (n=34) in the CG being male. Most participants lived in urban areas (IG: 38/68, 55.9%; CG: 42/68, 61.8%) and resided with family members (IG: 62/68, 91.2%; CG: 60/68, 88.2%). A large portion of participants were married (IG: 50/68, 73.5%; CG: 52/68, 76.5%). All participants had children, and education levels were primarily at the junior high school (IG: 22/68, 32.4%; CG: 10/68, 14.7%) and high school (IG: 21/68, 30.9%; CG: 22/68, 32.4%) levels. Additionally, a significant portion of participants did not use smartphones independently (IG: 15/68, 22.1%; CG: 10/68, 14.7%). The mean and SD for BMI and the duration of T2DM were as follows: IG 24.04 (SD 3.475), CG 24.08 (SD 3.428); IG 14.76 (SD 8.381), CG 14.19 (SD 7.658). Most participants used oral medications (IG: 37/68, 54.4%; CG: 35/68, 53.3%), and the proportion of family members without diabetes was similar (IG: 26/68, 33.3%; CG: 23/68, 29.5%). The most common adverse reactions were hypoglycemia (IG: 21/68, 30.9%; CG: 16/68, 26.5%) and digestive system issues (IG: 13/68, 10.1%; CG: 10/68, 14.7%). Details are provided in Table 2.

XSL•FO
RenderX

Table 2. Comparison of sociodemographic characteristics between the intervention group (IG) and control group (CG) at baseline among older adults living with type 2 diabetes mellitus (T2DM) in Henan, China (N=136).

Variable	IG (n=68)	CG (n=68)	Chi-square (*df*)[a,b]	*t* test (*df*)[c]	Z[d]	*P* value
Age (years), median (IQR)	67.5 (63.3-71)	66.0 (69.8-63)	—[e]	—	–0.901[d]	.37
Sex, n (%)			0.029 (1)[a]	—	—	.86
Male	35 (51.5)	34 (50.0)				
Female	33 (48.5)	34 (50.0)				
Residence, n (%)			0.486 (1)[a]	—	—	.49
Countryside	30 (44.1)	26 (38.2)				
City	38 (55.9)	42 (61.8)				
Marital status, n (%)			2.039 (2)[a]	—	—	.36
Married	50 (73.5)	52 (76.5)				
Divorced	3 (4.4)	6 (8.8)				
Widowed	15 (22.1)	10 (14.7)				
Living arrangement, n (%)			0.319 (1)[a]	—	—	.57
Living alone	6 (8.8)	8 (11.8)				
Living with family	62 (91.2)	60 (88.2)				
Education level, n (%)			7.344 (5)[a]	—	—	.20
Illiterate	4 (5.9)	6 (8.8)				
Primary school	9 (13.2)	14 (20.6)				
Junior high school	22 (32.4)	10 (14.7)				
High school	21 (30.9)	22 (32.4)				
Bachelor degree	12 (17.6)	15 (22.1)				
Postgraduate and above	0 (0)	1 (1.5)				
Have children, n (%)			—	—	—	—
Yes	68 (100)	68 (100)				
No	0 (0)	0 (0)				
Smoke, n (%)			4.946 (3) [a]	—	—	.18
Yes, smoker	11 (16.2)	19 (27.9)				
No, never smoke	38 (55.9)	39 (57.4)				
Used smoke, quite now	13 (19.1)	7 (10.3)				
Occasional smoker	6 (8.8)	3 (4.4)				
Smoke year, median (IQR)	0 (0-30)	0 (0-34.3)	—	—	–0.024[d]	.98
Drink, n (%)			2.079 (3)[a]	—	—	.56
Yes, drinker	8 (11.8)	14 (20.6)				
No, never drink	35 (51.5)	33 (48.5)				
Used drink, quite now	9 (13.2)	7 (10.3)				
Occasional drinker	16 (23.5)	14 (20.6)				
Drink year, median (IQR)	0 (0-39.3)	15 (0-40)	—	—	–0.398[d]	.69
Personal income, n (%)			0.256 (1)[a]	—	—	.61
Have income	58 (85.3)	60 (88.2)				
No income	10 (14.7)	8 (11.8)				

XSL•FO
RenderX

Variable	IG (n=68)	CG (n=68)	Chi-square (*df*)[a,b]	*t* test (*df*)[c]	Z[d]	*P* value
Personal income (RMB)[f,g], median (IQR)	4000 (2800-5375)	4000 (2850-5375)	—	—	–0.264[d]	.79
Household income, n (%)			0.000 (1)[b]	—	—	>.99
Have income	66 (97.1)	66 (97.1)				
No income	2 (2.9)	2 (2.9)				
Household income (RMB)[f,g], median (IQR)	8000 (7000-10,750)	8000 (6000-10,000)	—	—	–0.389[d]	.70
Occupational status, n (%)			1.578 (3)[a]	—	—	.66
Employed	9 (13.2)	10 (14.7)				
Unemployed	6 (8.8)	8 (11.8)				
Retired	40 (58.8)	42 (61.8)				
Others	13 (19.1)	8 (11.8)				
Medical payment method, n (%)			0.342 (3)[a]	—	—	.93
Self-paid	1 (1.5)	2 (2.9)				
Medical insurance	61 (89.7)	60 (88.2)				
Commercial insurance	5 (7.4)	5 (7.4)				
Others	1 (1.5)	1 (1.5)				
BMI (kg/m^2), mean (SD)	24.04 (3.475)	24.08 (3.428)	—	–0.066 (134)[c]	—	.65
<18.5	4 (5.9)	4 (5.9)				
18.5-23.9	30 (44.1)	27 (39.7)				
24-27.9	27 (39.7)	29 (42.6)				
≥28	7 (10.3)	8 (11.8)				
Smartphone usage, n (%)			1.225 (1)[a]	—	—	.27
Yes	53 (77.9)	58 (85.3)				
No	15 (22.1)	10 (14.7)				
T2DM duration, mean (SD)	14.76 (8.381)	14.19 (7.658)	—	–0.063 (134)[c]	—	.55
Family history, n (%)			0.353 (2)[a]	—	—	.84
Yes	23 (33.8)	26 (38.2)				
No	28 (41.2)	25 (36.8)				
Do not know	17 (25.0)	17 (25.0)				
Family member has diabetes mellitus, n (%)			6.543 (6)[a]	—	—	.37
No	26 (33.3)	23 (29.5)				
Father	9 (11.5)	17 (21.8)				
Mother	10 (12.8)	10 (12.8)				
Sibling	11 (14.1)	5 (6.4)				
Grandparents	2 (2.6)	5 (6.4)				
Do not know	18 (23.1)	17 (21.8)				
Other relative	2 (2.6)	1 (1.3)				
Combination disease						
Hypertension, n (%)			0.034 (1)[a]	—	—	.85
No	21 (30.9)	22 (32.4)				
Yes	47 (69.1)	46 (67.6)				

XSL•FO
RenderX

Variable	IG (n=68)	CG (n=68)	Chi-square (*df*)[a,b]	*t* test (*df*)[c]	Z[d]	*P* value
Coronary heart disease, n (%)			0.873 (1)[a]	—	—	.35
No	50 (73.5)	45 (66.2)				
Yes	18 (26.5)	23 (33.8)				
Chronic heart failure, n (%)			0.831 (1)[b]	—	—	.36
No	64 (94.1)	67 (98.5)				
Yes	4 (5.0)	1 (1.5)				
Stroke, n (%)			0.133 (1)[b]	—	—	.71
No	63 (92.6)	65 (95.6)				
Yes	5 (7.4)	3 (4.4)				
Kidney disease, n (%)			0.831 (1)[b]	—	—	.36
No	64 (94.1)	67 (98.5)				
Yes	4 (5.9)	1 (1.5)				
Other diseases, n (%)			0.067 (1)[a]	—	—	.80
No	60 (88.2)	59 (86.8)				
Yes	8 (11.8)	9 (13.2)				
Type of medication, n (%)			0.216 (2)[a]	—	—	.90
Oral	37 (54.4)	35 (53.3)				
Insulin	0 (0)	0 (0)				
Both	28 (41.2)	29 (42.6)				
Others	3 (4.4)	4 (5.9)				
Kind of medication, median (IQR)	3.0 (3-4.8)	3.0 (2-5)	—	—	–0.401[d]	.69
Adverse reactions						
Hypoglycemia, n (%)			0.324 (1)[a]	—	—	.57
No	47 (69.1)	50 (73.5)				
Yes	21 (30.9)	16 (26.5)				
Digestive system, n (%)			0.471 (1)[a]	—	—	.49
No	55 (80.9)	58 (85.3)				
Yes	13 (10.1)	10 (14.7)				
Chronic heart failure, n (%)			0.133 (1)[b]	—	—	.72
No	63 (92.6)	65 (95.6)				
Yes	5 (7.4)	3 (4.4)				
Allergic reaction, n (%)			0.000 (1)[b]	—	—	>.99
No	66 (97.1)	65 (97.1)				
Yes	2 (2.9)	2 (2.9)				
Liver damage, n (%)			1.681 (1)[a]	—	—	.20
No	62 (83.8)	69 (92.0)				
Yes	12 (16.2)	6 (8.0)				
Kidney damage, n (%)			0.319 (1)[a]	—	—	.57
No	60 (88.2)	62 (91.2)				
Yes	8 (11.8)	6 (8.8)				

XSL•FO
RenderX

Variable	IG (n=68)	CG (n=68)	Chi-square (*df*)[a,b]	*t* test (*df*)[c]	Z[d]	*P* value
Others diseases, n (%)			0.000 (1)[b]	—	—	>.99
No	66 (97.1)	65 (95.6)				
Yes	2 (2.9)	3 (4.4)				

[a]Pearson chi-square.
[b]Continuity correction.
[c]Independent sample *t* test.
[d]Mann-Whitney *U* test.
[e]Not applicable.
[f]1 RMB ≈ RM 0.61 (November 08, 2024).
[g]1 RMB ≈ US $0.14 (January 07, 2026).

Medication Adherence, Medication Knowledge, Medication Attitude, Medication Behavior, Medication Belief, and Social Support Between the IG and CG at Baseline

There were no significant differences between the IG and CG of medication adherence (23.71, SD 5.725 vs 24.01, SD 5.700; t_{134}=–0.315; *P*=.820), medication knowledge (63.41, SD 19.145 vs 62.37, SD 20.202; t_{134}=0.308; *P*=.547), medication attitude (35.76, SD 3.801 vs 36.15, SD 4.129; t_{134}=–0.562; *P*=.568), medication behavior (29.62, SD 4.381 vs 29.13, SD 4.029; t_{134}=0.672; *P*=.468), medication belief (3.00, SD 3.515 vs 2.94, SD 3.777; t_{134}=0.094; *P*=.520) and social support (69.62, SD 11.402 vs 68.22, SD 10.935; t_{134}=0.729; *P*=.805) at baseline (Table 3).

Table 3. Total score of medication adherence, knowledge, attitude, behavior, belief, and social support between the intervention group (IG) and control group (CG) at baseline among older adults living with type 2 diabetes mellitus in Henan, China.

Outcomes (range)[a]	IG, mean (SD)	CG, mean (SD)	*t* test(*df*)	*P* value[b]
Medication adherence (13 to 39)	23.71 (5.725)	24.01 (5.700)	–0.315 (134)	.82
Medication knowledge (0 to 100)	63.41 (19.145)	62.37 (20.202)	0.308 (134)	.55
Medication attitude (10 to 50)	35.76 (3.801)	36.15 (4.129)	–0.562 (134)	.57
Medication behavior (9 to 40)	29.62 (4.381)	29.13 (4.029)	0.672 (134)	.47
Medication belief (–25 to +25)	3.00 (3.515)	2.94 (3.777)	0.094 (134)	.52
Social support (19 to 95)	69.62 (11.402)	68.22 (10.935)	0.729 (134)	.81

[a]Higher scores of outcomes indicate better medication adherence, knowledge, attitude, behavior, belief, and social support.
[b]*P* value with *P*>.05 indicates that there are no significant differences between the IG and CG for the measured outcomes.

Outcomes

Overview

This study includes 6 outcome variables, including medication adherence, medication knowledge, medication attitude, medication behavior, medication beliefs, and social support. Given the absence of significant effects, no covariates were incorporated into the final model. Generalized estimating equation was used to investigate the changes in outcomes between IG and CG across 3 time points. The results are shown in Table 4.

XSL•FO
RenderX

Table 4. Comparison of medication adherence, medication knowledge, medication attitude, medication behavior, medication belief, and social support scores between the intervention group (IG) and control group (CG) across time points using generalized estimating equations among older adults living with type 2 diabetes mellitus in Henan, China.

Variable	Medication knowledge[a]		Medication attitude[b]		Social support[c]		Medication belief[d]		Medication behavior[e]		Medication adherence[f]	
	β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value
Group effect												
CG	Ref[g]	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
IG	7.887 (3.428 to 12.345)	.001[h]	2.465 (1.466 to 3.464)	<.001[h]	1.277 (–4.771 to 7.352)	.679	1.845 (1.078 to 2.613)	<.001[h]	2.731 (1.639 to 3.823)	<.001[h]	2.288 (0.681 to 3.894)	.005[h]
Time effect												
T0[i]	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
T1[j]	7.093 (4.564 to 9.623)	<.001[h]	3.103 (2.370 to 3.836)	<.001[h]	0.397 (–0.564 to 1.358)	.418	1.632 (1.125 to 2.140)	<.001[h]	2.441 (1.848 to 3.034)	<.001[h]	2.601 (1.858 to 3.363)	<.001[h]
T2[k]	4.844 (2.462 to 7.227)	<.001[h]	1.662 (1.059 to 2.264)	<.001[h]	–0.574 (–1.491 to 0.344)	.221	0.588 (0.062 to 1.114)	.028[h]	1.125 (0.586 to 1.664)	<.001[h]	1.684 (0.958 to 2.409)	<.001[h]
Timepoint×Group effect												
T0 × CG	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
T1 × IG	11.592 (6.923 to 16.260)	<.001[h]	5.467 (4.531 to 6.763)	<.001[h]	0.008 (–1.834 to 2.011)	.928	2.882 (1.990 to 3.775)	<.001[h]	4.176 (3.220 to 5.133)	<.001[h]	4.956 (3.702 to 6.210)	<.001[h]
T2× IG	6.055 (1.401 to 10.710)	.011[h]	3.059 (1.969 to 4.149)	<.001[h]	–0.147 (–1.982 to 1.688)	.875	1.882 (0.880 to 2.885)	<.001[h]	2.426 (1.429 to 3.424)	<.001[h]	3.691 (2.379 to 5.003)	<.001[h]

[a]Group: QIC=105,178.155, QICC=105,173.944, time: QIC=106,219.108, QICC=106,219.108, time × group: QIC=99,054.337, QICC=99,054.337.

[b]Group: QIC=5783.910, QICC=5780.419, time: QIC=5787.824, QICC=5787.824, time × group: QIC=4587.706, QICC=4687.706.

[c]Group: QIC=37,306.144, QICC=36,267.594, time: QIC=37,378.507, QICC=37,378.597, time × group: QIC=37,198.191, QICC=37,198.191.

[d]Group: QIC=3561.609, QICC=3558.406 time: QIC=3643.971, QICC=3643.971, time × group: QIC=3227.618, QICC=3227.618.

[e]Group: QIC=6009.207, QICC=6004.590, time: QIC=6301.456, QICC=6301.456, time × group: QIC=5304.176, QICC=5304.176.

[f]Group: QIC=11,557.678, QICC=11,551.984 time: QIC=11,731.963, QICC=11,731.963, time × group: QIC=10,617.505, QICC=10,617.505.

[g]Ref: Reference group.

[h]Significant at *P*<.05.

[i]T0: baseline.

[j]T1: postintervention.

[k]T2: 3-month follow up.

Primary Outcome

The IG exhibited a significantly higher medication adherence score compared with the CG, with a mean difference of 2.288 points (β=2.288, 95% CI 0.681-3.894, *P*=.005). Participants demonstrated significant improvements in medication adherence both immediately after the intervention (β=2.610, 95% CI 1.858-3.363, *P*<.001) and at the 3-month follow-up (β=1.684, 95% CI 0.958-2.409, *P*<.001) when compared with baseline. The IG showed a significantly higher medication adherence score postintervention, with a difference of 4.956 points compared with the CG at baseline (β=4.956, 95% CI 3.702-6.210, *P*<.001). At the 3-month follow-up, the IG maintained a significantly higher score than the CG at baseline, with a difference of 3.691 points (β=3.691, 95% CI 2.379-5.003, *P*<.001).

Secondary Outcomes

Compared with the CG, the IG showed a significant improvement in medication knowledge score, with a postintervention score higher by 7.887 points (β=7.887, 95% CI 3.428-12.345, *P*=.001). Compared with baseline, the

XSL•FO
RenderX

medication knowledge score of the participants increased by 7.093 points after the intervention and by 4.844 points at the 3-month follow-up (β=7.093, 95% CI 4.564-9.623, *P*<.001; β=4.844, 95% CI 2.462–7.227, *P*<.001). After the intervention, the medication knowledge score of the IG group was significantly higher than the CG group at baseline (β=11.592, 95% CI 6.923-16.260, *P*<.001). At the 3-month follow-up, the medication knowledge score of the IG group was 6.055 points higher than that of the CG group at baseline (β=6.055, 95% CI 1.401-10.710, *P*=.011).

The medication attitude score of the IG was significantly higher than that of CG, with a difference of 2.465 points (β=2.465, 95% CI 1.466-3.464, *P*<.001). Compared with the baseline, the medication attitude score of the participants significantly increased by 3.103 points after the intervention and by 1.662 points at the 3-month follow-up, respectively (β=3.103, 95% CI 2.370-3.836, *P*<.001; β=1.662, 95% CI 1.059-2.264, *P*<.001). After the intervention, the medication attitude score of the IG group was significantly higher than that of the CG group at baseline, with a difference of 5.467 points (β=5.467, 95% CI 4.531-6.763, *P*<.001). At the 3-month follow-up, the medication attitude score of the IG group was 3.059 points higher than that of the CG group at baseline (β=3.059, 95% CI 1.969-4.149, *P*<.001).

The IG showed a significantly higher medication behavior score compared with the CG, with a mean difference of 2.731 points (β=2.731, 95% CI 1.639-3.823, *P*<.001). Participants demonstrated significant improvements in medication behavior relative to baseline, both immediately after the intervention (β=2.441, 95% CI 1.848-3.034, *P*<.001) and at the 3-month follow-up (β=1.125, 95% CI 0.586-1.664, *P*<.001). Postintervention, the IG scored 4.176 points higher than the CG at baseline (β=4.176, 95% CI 3.220-5.133, *P*<.001). At the 3-month follow-up, the IG also outperformed the CG at baseline, with a difference of 2.426 points (β=2.426, 95% CI 1.429-3.424, *P*<.001).

The IG demonstrated a significantly higher medication belief score than the CG, with an estimated difference of 1.845 points (β=1.845, 95% CI 1.078-2.613, *P*<.001). Compared with baseline, participants showed a significant increase in medication belief score both immediately after the intervention (β=1.632, 95% CI 1.125-2.140, *P*<.001) and at the 3-month follow-up (β=0.588, 95% CI 0.062-1.114, *P*=.028). Furthermore, in the analysis of the interaction between time and group, the IG at postintervention scored significantly higher than the CG at baseline, with a difference of 2.882 points (β=2.882, 95% CI 1.990-3.775, *P*<.001). Similarly, at the 3-month follow-up, the IG maintained a significantly higher score than the CG at baseline (β=1.882, 95% CI 0.880-2.885, *P*<.001).

However, social support did not reach statistical significance in terms of time, group, and the interaction between time and group, in either the IG or CG group (*P*>.05).

Discussion

Principal Findings

Research has shown that the IMB model can effectively explain, predict, and improve medication adherence among older adults living with chronic diseases [54,55]. Based on previous research findings, this study combined the IMB model to develop an animated intervention video called “Diabetes Little Helper” which was disseminated through the WeChat platform, further validating the model’s applicability in older adults. As with the results of this study, a quasi-experimental study by Mirzaei-Alavijeh et al [56], based on the IMB model, significantly improved medication adherence in older adults living with T2DM through 6 face-to-face educational sessions, further supporting the effectiveness of the IMB model. This study, through the WeChat platform, disseminates educational information in video format, providing new evidence and intervention pathways for improving medication adherence among older adults living with T2DM.

A notable proportion of participants reported not using smartphones independently. This finding is significant, as it could influence the overall effectiveness of the intervention. However, it is important to note that all participants in this study either owned a smartphone and used WeChat or received assistance from family members to participate. Nonetheless, the involvement of family members could be seen as a limitation, as it may affect the autonomy of older adults and potentially influence their overall engagement with the intervention. Additionally, reliance on family assistance raises concerns about the generalizability and long-term feasibility of the intervention, particularly for older adults who may not have family members available to offer support. To address this, future research could explore alternative strategies to enhance technological support for older adults, such as simplified user interfaces or targeted training programs. These approaches could ensure broader accessibility and improve the effectiveness of similar interventions in diverse populations.

Participants’ medication adherence score showed significant improvement following the intervention, a result that mirrors outcomes seen in other studies using different methods. For instance, mobile apps were used to provide educational information to family members, who in turn helped older adults living with DM complete self-management tasks, pharmacy students conducted motivational interviewing through phone calls to enhance medication adherence, while pharmacists used apps, WeChat public accounts, WeChat groups, and both web-based and offline lectures to improve medication adherence among older adults living with DM [30,31,33]. These interventions effectively increased medication adherence in older adults living with DM, though they also presented certain challenges. Relying on family members to convey information poses the risk of information distortion and places psychological pressure on the family. Dependence on pharmacy students for motivational interviewing might lead to insufficient mastery of communication and problem-solving skills, potentially affecting the interview’s effectiveness. Although pharmacists use a combination of web-based and offline platforms to provide

XSL•FO
RenderX

comprehensive services, the use of multiple channels may confuse older adults, especially those less familiar with digital technologies. In contrast, this study directly engaged patients via WeChat, eliminating hierarchical errors in information transmission and avoiding family involvement in the medical process. The professionally designed animated videos ensured the accuracy and coherence of the information, minimizing human bias. Moreover, the approach did not require the development of complex programs or offline services, thus saving both manpower and financial resources.

This study demonstrated that the Diabetes Little Helper video significantly improved participants' medication knowledge score, which is consistent with previous research findings [57]. Compared with the one-on-one face-to-face educational model used by Yang et al [57], this study overcame the limitations of time and location while also saving human resources. Xiong and Wan [58] also improved older adult participants' medication knowledge score by sending text and images via the WeChat platform. In comparison, the animated video used in this study not only incorporated the advantages of text and images but also effectively captured patients' attention through its vivid and engaging format, helping viewers better understand concepts and stimulate cognitive responses. The use of animated videos has been shown to be a more attractive and dynamic method, aiding viewers in gaining a deeper understanding of health education content [59]. Future intervention research could explore how to combine face-to-face education with digital video interventions, leveraging the strengths of both traditional and modern educational methods to further enhance patients' medication knowledge scores.

Participants' medication attitudes were effectively improved, as agreed with Chen et al [60], who used a multidimensional, multiscenario intervention approach, which also successfully enhanced medication attitudes. However, their method requires substantial human resources and high costs, making it challenging to implement in regions with limited health care resources. Zhou [61] conducted an intervention through community health service centers, regularly monitoring patients' medication levels, metabolism, and relevant physiological indicators to guide medication adjustments, which improved medication attitudes in older adults. In contrast, the intervention model in this study overcomes regional and resource limitations through digital means, making it simple and feasible. Each intervention method has its unique advantages and suitable contexts. Future research should flexibly select intervention strategies based on the specific characteristics of hospitals and patients, and integrate multiple approaches to more effectively improve medication attitudes in older adults.

The intervention also had a positive impact on improving participants' medication behavior score, in line with the findings of Wang et al [62], whose study increased medication adherence through personalized plans and continuous optimization. However, their approach relied heavily on frequent face-to-face reminders, which may have provoked resistance from some patients. In this study, patients were able to choose their learning time and frequency, which reduced resistance and improved psychological acceptance. Moreover, the intervention program by Wan et al [63] required 50 participants to provide more tailored health education and medication management. Although it also led to some improvement in medication behavior, the management process was relatively complex and added additional costs. Moreover, staff turnover, absenteeism, and communication difficulties could negatively impact the effectiveness of the intervention.

The results of this study also indicate a significant improvement in participants' medication beliefs, aligning with findings from previous research [64,65]. These studies used interviews or multidisciplinary collaboration (involving family doctors, pharmacists, and nurses) to integrate and guide medication management. However, these approaches are typically limited to stable patients who can attend interviews and are not applicable to older adults living at home. Additionally, team collaboration demands more human resources. However, this study focused on broadening medication knowledge and promoting self-management, encouraging older adults to independently adjust their medication attitudes.

This study did not show a significant effect on improving participants' social support scores. This finding is consistent with the results of Yang et al [57], who also did not observe significant improvements in social support after using traditional intervention methods, such as face-to-face education, interviews, phone communication, and record-keeping. One possible reason for this is that the development of social support requires sustained intervention over more than 6 months [66]. Additionally, some studies have shown that including education for patients' family members and facilitating web-based interactive communication effectively improved social support among older adults [67,68]. Future research could integrate interactive modules, such as involving family members in the intervention plan, promoting web-based communication among patients, and combining community nursing, to further strengthen the patient's social support network and enhance the overall intervention effectiveness.

This study is the first to develop an animated video intervention to improve medication adherence in older adults living with T2DM, expanding the application of this model in medication management for older adults living with DM and providing new insights for chronic disease management. Moreover, this study innovatively used the WeChat public platform to deliver animated video intervention content to older adults living with DM. This approach not only overcomes the time and space limitations of traditional health education but also leverages the widespread use and interactivity of the WeChat platform, enabling continuous health interventions. As a result, it significantly increased the accessibility of intervention content and enhanced patient acceptance. Patients could repeatedly watch the intervention videos in their daily lives, which helped reinforce memory and promote behavior change.

Limitations and Future Research

While the findings of this study provide valuable insights into the effectiveness of the intervention, several limitations must be acknowledged. First, the reliance on self-reported adherence introduces potential biases, such as over-reporting, underreporting, and recall bias. These biases can compromise the accuracy of adherence measurements and may not fully

XSL•FO
RenderX

reflect the participants' true behaviors. Future studies could incorporate objective measures of adherence, such as electronic monitoring or biomarkers, to reduce these biases and provide a more accurate picture of adherence. Second, while the 3-month follow-up period allowed us to assess the short-term effects of the intervention, it was insufficient to evaluate the long-term sustainability of the intervention's impact. Future studies should aim for extended follow-up periods to gain a deeper understanding of the long-term benefits and challenges of such interventions. Furthermore, the inability to blind participants due to the visible nature of the intervention video may have introduced bias, as participants were aware of their intervention condition, which could have influenced their responses or behaviors. To mitigate this in future studies, an RCT design with blinding could be implemented, or a different form of intervention that is less visible could be explored.

Adherence monitoring measures, though designed to ensure compliance, may have inadvertently influenced engagement and adherence, thereby introducing additional intervention effects. Future studies should consider monitoring approaches that minimize such influences. In terms of data handling, this study used the LOCF method to address missing data. Although LOCF is computationally simple and commonly used in short-term studies, it assumes that missing data are equivalent to the last observed value, which may lead to an underestimation of variability in the results. If participants drop out due to deteriorating health, LOCF may fail to capture this change, potentially overestimating the intervention effect. Despite these limitations, LOCF was considered an appropriate choice for this study. Future research could consider exploring other advanced methods for handling missing data, such as multiple imputation, to improve the accuracy of data analysis.

Conclusions

This study evaluated the impact of a WeChat-based Diabetes Little Helper video intervention on medication adherence and medication-related outcomes in older adults living with T2DM. The results indicated that the intervention had a significant positive effect on improving patients' medication adherence, knowledge, attitude, behavior, and belief. However, it did not significantly enhance the patients' level of social support. Building on these findings, future research could expand to include more regions and health care institutions and conduct multicenter RCTs to verify the applicability and effectiveness of the Diabetes Little Helper video intervention across different geographic and cultural contexts. This would help to further improve the external validity and reliability of the study's findings. Moreover, future studies could explore the integration of more advanced health tools, such as smart wristbands and smart pillboxes, to monitor the health status and medication adherence of older adults living with T2DM in real time, thereby providing additional support for personalized interventions and data analysis.

Acknowledgments

The authors gratefully acknowledge the support of Universiti Putra Malaysia and 2 hospitals in Henan province, China, for their contributions to this research. The authors also thank Universiti Putra Malaysia for the publication fees support. We also extend our heartfelt thanks to all the experts who offered their valuable guidance, professional insights, and feedback, which were instrumental in enhancing the quality and impact of this intervention study. All authors approved this manuscript for publication.

Funding

This study did not receive research funding from any public, commercial, or nonprofit funding agencies. The publication fees are supported by Universiti Putra Malaysia.

Data Availability

The datasets generated or analyzed during this study are available in the Mendeley data repository.

Authors' Contributions

MW, PYL, KL, SNS, HZT, and QHJ contributed to the conceptualization and design of the study. Data collection was carried out by MW, QHJ, LLH, ZYZ, and YL. MW and PYL conducted the data analysis. Interpretation of the results was performed by MW, PYL, SNS, and HZT. MW was responsible for drafting and revising the manuscript. Supervision was provided by PYL, KL, SNS, and HZT. Project administration was managed by PYL, SNS, and HZT. Ethical approval was obtained by MW and QHJ.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1108 KB - jmir_v28i1e78731_app1.pdf]

XSL•FO
RenderX

References

1. Sun, Wen H, Liu X, Ma Y, Jang J, Yu C. Time trends in type 2 diabetes mellitus incidence across the BRICS from 1990 to 2019: an age-period-cohort analysis. BMC Public Health 2022;22(1):65 [FREE Full text] [doi: 10.1186/s12889-021-12485-y] [Medline: 35012503]
2. Hossain MJ, Al-Mamun M, Islam MR. Diabetes mellitus, the fastest growing global public health concern: early detection should be focused. Health Sci Rep 2024;7(3):e2004 [FREE Full text] [doi: 10.1002/hsr2.2004] [Medline: 38524769]
3. Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, et al. Erratum to "IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045" [Diabetes Res. Clin. Pract. 183 (2022) 109119]. Diabetes Res Clin Pract 2023;204:110945. [doi: 10.1016/j.diabres.2023.110945] [Medline: 37863776]
4. Sinclair A, Saeedi P, Kaundal A, Karuranga S, Malanda B, Williams R. Diabetes and global ageing among 65-99-year-old adults: findings from the International Diabetes Federation Diabetes Atlas, 9 edition. Diabetes Res Clin Pract 2020;162:108078. [doi: 10.1016/j.diabres.2020.108078] [Medline: 32068097]
5. Liu J, Liu M, Chai Z, Li C, Wang Y, Shen M, et al. Projected rapid growth in diabetes disease burden and economic burden in China: a spatio-temporal study from 2020 to 2030. Lancet Reg Health West Pac 2023;33:100700 [FREE Full text] [doi: 10.1016/j.lanwpc.2023.100700] [Medline: 36817869]
6. Babakhanian M, Razavi A, Rahimi Pordanjani S, Hassanabadi S, Mohammadi G, Fattah A. High incidence of type 1 diabetes, type 2 diabetes and gestational diabetes in Central Iran: a six years results from Semnan health cohort. Ann Med Surg (Lond) 2022;82:104749 [FREE Full text] [doi: 10.1016/j.amsu.2022.104749] [Medline: 36268322]
7. Baechle C, Stahl-Pehe A, Prinz N, Meissner T, Kamrath C, Holl R, in cooperation with the German Paediatric Surveillance Unit (ESPED), the DPV initiativethe German Center for Diabetes Research (DZD). Prevalence trends of type 1 and type 2 diabetes in children and adolescents in North Rhine-Westphalia, the most populous federal state in Germany, 2002-2020. Diabetes Res Clin Pract 2022;190:109995. [doi: 10.1016/j.diabres.2022.109995] [Medline: 35853531]
8. Li D, Luo M, Liu Y, Dong J, Geng W, Li X, et al. Increased rates of health management and health education on hypertension and diabetes in Inner Mongolia, China: 10-year population level trends (2009-2018). Int J Environ Res Public Health 2022;19(20) [FREE Full text] [doi: 10.3390/ijerph192013048] [Medline: 36293621]
9. Tomic D, Harding JL, Jenkins AJ, Shaw JE, Magliano DJ. The epidemiology of type 1 diabetes mellitus in older adults. Nat Rev Endocrinol 2025;21(2):92-104. [doi: 10.1038/s41574-024-01046-z] [Medline: 39448829]
10. Han J, Wang H, Liang HH, Guo JX. Correlation of the retinopathy degree with the change of ocular surface and corneal nerve in patients with type 2 diabetes mellitus. Int J Ophthalmol 2021;14(5):750-758 [FREE Full text] [doi: 10.18240/ijo.2021.05.17] [Medline: 34012892]
11. Kirthi V, Perumbalath A, Brown E, Nevitt S, Petropoulos IN, Burgess J, et al. Prevalence of peripheral neuropathy in pre-diabetes: a systematic review. BMJ Open Diabetes Res Care 2021;9(1) [FREE Full text] [doi: 10.1136/bmjdrc-2020-002040] [Medline: 34006607]
12. Wang L, Peng W, Zhao Z, Zhang M, Shi Z, Song Z, et al. Prevalence and treatment of diabetes in China, 2013-2018. JAMA 2021;326(24):2498-2506 [FREE Full text] [doi: 10.1001/jama.2021.22208] [Medline: 34962526]
13. Emerging Risk Factors Collaboration. Life expectancy associated with different ages at diagnosis of type 2 diabetes in high-income countries: 23 million person-years of observation. Lancet Diabetes Endocrinol 2023;11(10):731-742 [FREE Full text] [doi: 10.1016/S2213-8587(23)00223-1] [Medline: 37708900]
14. Jyotsna F, Ahmed A, Kumar K, Kaur P, Chaudhary M, Kumar S, et al. Exploring the complex connection between diabetes and cardiovascular disease: analyzing approaches to mitigate cardiovascular risk in patients with diabetes. Cureus 2023;15(8):e43882 [FREE Full text] [doi: 10.7759/cureus.43882] [Medline: 37746454]
15. Kumar M, Dev S, Khalid M, Siddenthi S, Noman M, John C, et al. The bidirectional link between diabetes and kidney disease: mechanisms and management. Cureus 2023;15(9):e45615 [FREE Full text] [doi: 10.7759/cureus.45615] [Medline: 37868469]
16. Kirya M, Bwayo D, Otim M, Mutoo P, Masaba J, Okibure A, et al. Prevalence of biomarkers and associated factors for chronic kidney disease in adult diabetic out-patients in a tertiary hospital in Eastern Uganda - a cross-sectional study. Res Sq 2024 [FREE Full text] [doi: 10.21203/rs.3.rs-3992049/v1] [Medline: 38496471]
17. Taimur H, Ahmad I, Khan H, Shirayama Y, Okamoto M, Aung MN, et al. A scoping review of type 2 diabetes mellitus in Pakistan investigating the status of glycemic control, awareness, treatment adherence, complications and cost. Front Endocrinol (Lausanne) 2024;15:1441591 [FREE Full text] [doi: 10.3389/fendo.2024.1441591] [Medline: 39649227]
18. Brown MT, Bussell J, Dutta S, Davis K, Strong S, Mathew S. Medication adherence: truth and consequences. Am J Med Sci 2016;351(4):387-399. [doi: 10.1016/j.amjms.2016.01.010] [Medline: 27079345]
19. Weinberg Sibony R, Segev O, Dor S, Raz I. Drug therapies for diabetes. Int J Mol Sci 2023 Dec 05;24(24). [doi: 10.3390/ijms242417147] [Medline: 38138975]
20. Pallangyo P, Millinga J, Bhalia S, Mkojera Z, Misidai N, Swai HJ, et al. Medication adherence and survival among hospitalized heart failure patients in a tertiary hospital in Tanzania: a prospective cohort study. BMC Res Notes 2020;13(1):89 [FREE Full text] [doi: 10.1186/s13104-020-04959-w] [Medline: 32085803]

XSL•FO
RenderX

21. Zhang, Nie X, Yang X, Mei Q, Xiang X, Cheng L. A meta-analysis of effectiveness of mobile health interventions on health-related outcomes in patients with heart failure. J Cardiovasc Med (Hagerstown) 2024;25(8):587-600. [doi: 10.2459/JCM.0000000000001631] [Medline: 38949146]
22. Rasu RS, Xavier C, Rianon N. Dynamic changes in medication burden leading to fall and hospital readmissions in older adults: toward a strategy for improving risk and managing costs. J Manag Care Spec Pharm 2025;31(1):96-100. [doi: 10.18553/jmcp.2025.31.1.96] [Medline: 39745841]
23. He HM, Cai H, Rong QF, Heng JQ. Investigation of factors influencing medication adherence among elderly patients with chronic diseases and strategies for dealing with them. Chin Remedies Clin 2020;20:1622-1624 [FREE Full text] [doi: 10.11655/zgywylc2020.10.013]
24. Zhou HB, Zhang YZ, Yang H, Xia Q, Su H. Investigation on medication compliance of elderly patients with type 2 diabetes mellitus. Diabetes New World 2022 Feb 16;25:13-16 [FREE Full text] [doi: 10.16658/j.cnki.1672-4062.2022.04.013]
25. Khan II, Pulock OS, Pinky SD, Barua B, Dola TA, Chowdhury P, et al. Treatment noncompliance level among patients with type 2 diabetes mellitus: a hospital based cross-sectional study in Bangladesh. PLoS One 2022;17(9):e0271107 [FREE Full text] [doi: 10.1371/journal.pone.0271107] [Medline: 36099268]
26. Zhou JY. Analysis of medication compliance and influencing factors among elderly patients with chronic diseases in Fangshan district, Beijing. Mod Dis Control Prev 2023;34(12):930-957 [FREE Full text] [doi: 10.13515/j.cnki.hnjpm.1006-8414.2023.12.012]
27. Zhang JY, Zhang LQ, Tao MX, Chen XL. Analysis of medication adherence status and influencing factors in elderly type 2 diabetes patients. Pract Prev Med 2024;31(11):1382-1385 [FREE Full text] [doi: 10.3969/j.issn.1006-3110.2024.11.025]
28. Yao YC, Zhong YJ, Yang L. Compliance health education for elderly patients with multidrug use management intervention based on multidisciplinary cooperation. J Prev Med 2023 Jan 23;39(12):1464-1469 [FREE Full text]
29. Li B, Zang S, Ma LP. Health education management in the follow-up of elderly patients with diabetes mellitus. J Qilu Nurs 2023 Feb 05;29(3):111-113 [FREE Full text] [doi: 10.3969/j.issn.1006-7256.2023.03.033]
30. Poonprapai P, Lerkiatbundit S, Saengcharoen W. Family support-based intervention using a mobile application provided by pharmacists for older adults with diabetes to improve glycaemic control: a randomised controlled trial. Int J Clin Pharm 2022;44(3):680-688. [doi: 10.1007/s11096-022-01389-5] [Medline: 35247147]
31. Mohan A, Majd Z, Johnson ML, Essien EJ, Barner J, Serna O, et al. A motivational interviewing intervention to improve adherence to ACEIs/ARBs among nonadherent older adults with comorbid hypertension and diabetes. Drugs Aging 2023;40(4):377-390 [FREE Full text] [doi: 10.1007/s40266-023-01008-6] [Medline: 36847995]
32. Wang ZM, Sun BY, Hu YY. Influence of remote clinical pharmacy service platform on drug use in older patients with type 2 diabetes mellitus. Int J Geriatr 2023 May 20;44(3):280-283 [FREE Full text] [doi: 10.3969/j.issn.1674-7593.2023.03.005]
33. Zhang XL, Liu HL, Wang S. Influence of family pharmacists′ service on medication adherence and efficacy of elderly patients with type 2 diabetes mellitus in community. Chin Pharm 2024;33(22):48-52 [FREE Full text] [doi: 10.3969/j.issn.1006-4931.2024.22.010]
34. Lee S, Kim B. Factors affecting nurses' health promotion behavior during the COVID-19 pandemic based on the information-motivation-behavioral skills model. Medicina (Kaunas) 2022;58(6) [FREE Full text] [doi: 10.3390/medicina58060720] [Medline: 35743983]
35. Park N, Jang S. Effects of mHealth practice patterns on improving metabolic syndrome using the information-motivation-behavioral skills model. Nutrients 2024;16(13) [FREE Full text] [doi: 10.3390/nu16132099] [Medline: 38999847]
36. Nelson LA, Wallston KA, Kripalani S, LeStourgeon LM, Williamson SE, Mayberry LS. Assessing barriers to diabetes medication adherence using the information-motivation-behavioral skills model. Diabetes Res Clin Pract 2018;142:374-384 [FREE Full text] [doi: 10.1016/j.diabres.2018.05.046] [Medline: 29879495]
37. Schoenthaler A, Leon M, Butler M, Steinhaeuser K, Wardzinski W. Development and evaluation of a tailored mobile health intervention to improve medication adherence in black patients with uncontrolled hypertension and type 2 diabetes: pilot randomized feasibility trial. JMIR Mhealth Uhealth 2020;8(9):e17135 [FREE Full text] [doi: 10.2196/17135] [Medline: 32965230]
38. Xu Q, Wu P, Du J, Zhuang W, He X, Ma Y, et al. Online short videos promoting public breast cancer literacy: a pretest-posttest control group trial on efficiency, attitude, and influencing factors. Front Public Health 2023;11:1198780 [FREE Full text] [doi: 10.3389/fpubh.2023.1198780] [Medline: 37397762]
39. Liang H, Pang H. The cognitive, behavioral and interpersonal impacts of virtual practice with short health videos on Chinese ageing women: a discursive approach. Int J Environ Res Public Health 2022;19(12) [FREE Full text] [doi: 10.3390/ijerph19127173] [Medline: 35742422]
40. Gould CE, Carlson C, Wetherell JL, Goldstein MK, Anker L, Beaudreau SA. Brief video-delivered intervention to reduce anxiety and improve functioning in older veterans: pilot randomized controlled trial. JMIR Aging 2024;7:e56959 [FREE Full text] [doi: 10.2196/56959] [Medline: 39652863]
41. Al-Kharouf MS, Abdeljalil MH, Obeidat NM, Oweidat KA, Awwad O. Video-based teach-to-goal intervention on inhaler technique on adults with asthma and COPD: a randomized controlled trial. PLoS One 2023;18(6):e0286870 [FREE Full text] [doi: 10.1371/journal.pone.0286870] [Medline: 37294823]

XSL•FO
RenderX

42. Wang M, Lim PY, Lee K, Sowtali SN, Jiang Q, Thew HZ. Effectiveness of diabetes little helper video intervention on medication adherence among elderly patients with type 2 diabetes mellitus in Henan, China: study protocol for a randomised controlled trial. BMJ Open 2025;15(4):e093191 [FREE Full text] [doi: 10.1136/bmjopen-2024-093191] [Medline: 40258641]
43. Chinese Elderly Type 2 Diabetes PreventionTreatment of Clinical Guidelines Writing Group. [Clinical guidelines for prevention and treatment of type 2 diabetes mellitus in the elderly in China (2022 edition)]. Zhonghua Nei Ke Za Zhi 2022 Jan 01;61(1):12-50. [doi: 10.3760/cma.j.cn112138-20211027-00751] [Medline: 34979769]
44. Lemeshow S, Hosmer DW. Adequacy of Sample Size in Health Studies. New Jersey: Wiley & Sons; 1992:120.
45. Cao Z. Analysis of the impact of clinical pharmacist guidance on rational medication and compliance in elderly patients with chronic diseases. Chin J Mod Drug Appl 2024;18(6):160-163 [FREE Full text] [doi: 10.14164/j.cnki.cn11-5581/r.2024.06.042]
46. Wu X, Qian J, Gong L. Targeted nursing interventions based on SEAMS assessment: effects on blood pressure control and medication adherence in elderly patients with hypertension. Heilongjiang Med Pharm 2024 Apr 8;47(2):84-89 [FREE Full text] [doi: 10.3969/j.issn.1008-0104.2024.02.026]
47. Yan C, Ding Y. Micro-lecture design and development based on Wondershare Animiz. Shoes Technol Des 2025 Jul 28:5-29 [FREE Full text] [doi: 10.31958/je.v3i1.10120]
48. Sowtali SN. Development, implementation and evaluation of family-based educational program on nutrient intake and stroke knowledge for first ever stroke patients. In: Universiti Sains Malaysia. Malaysia: Universiti Sains Malaysia; 2016.
49. Dong XF, Liu YJ, Wang A, Lv PH. Psychometric properties of the Chinese version of the self-efficacy for appropriate medication use scale in patients with stroke. Patient Prefer Adherence 2016;10:321-327 [FREE Full text] [doi: 10.2147/PPA.S101844] [Medline: 27042023]
50. Dong XF, Liu YJ, Wang AX. Reliability and validity of the Chinese version self-efficacy for appropriate medication use scale. J Nurs Sci 2015;30:47-49. [doi: 10.3870/hlxzz.2015.11.047]
51. Xiang GP. A study on the effect of health education on knowledge, attitude, behavior of safe medication for community-dwelling elderly with chronic diseases. In: Central South University. Changsha: Central South University; Nov 2012:1-55.
52. Tang J, Gao LL, Hu XL, Wu JP, Yu KH, Liu SH. The reliability and validity of the Chinese version of beliefs about medicines questionnaire-specific among patients with type 2 diabetes mellitus. J Nurs Sci 2019;34:25-28 [FREE Full text]
53. Huang SF, Chen LX, Ye YM, Gong GM, Shi GF. Preparation of empty nester social support scale and analysis of its reliability and validity. Nurs Res 2017 Mar 09;31(1):537-539 [FREE Full text] [doi: 10.3969/j.issn.1009-6493.2017.05.007]
54. Yang C, Lee DTF, Chair SY. Validation and adaption of an information-motivation-behavioral skills model of medication adherence in older patients with multimorbidity. Patient Educ Couns 2023;113:107756. [doi: 10.1016/j.pec.2023.107756] [Medline: 37156122]
55. Oh E, Kim C, Schlenk EA. A predictive model for medication adherence in older adults with heart failure. Eur J Cardiovasc Nurs 2024 Oct 05;23(6):635-643. [doi: 10.1093/eurjcn/zvae021] [Medline: 38408016]
56. Mirzaei-Alavijeh M, Yarmoradi S, Khashij S, Jalilian F. Application of the information-motivation-behavioral skills model to improve medication adherence among older adults with type 2 diabetes: findings from a quasi-experimental study in Kermanshah, Iran. BMC Geriatr 2025 May 13;25(1):328 [FREE Full text] [doi: 10.1186/s12877-025-06000-w] [Medline: 40361006]
57. Yang C, Lee DTF, Wang X, Chair SY. Effects of a nurse-led medication self-management intervention on medication adherence and health outcomes in older people with multimorbidity: a randomised controlled trial. Int J Nurs Stud 2022;134:104314. [doi: 10.1016/j.ijnurstu.2022.104314] [Medline: 35849886]
58. Xiong YY, Wan J. The influence of WeChat platform technology for the rational drug use and drug compliance of elderly hypertensive patients in the community. Shanghai Med Pharm J 2023 Aug 25;44(16):36-39 [FREE Full text] [doi: 10.3969/j.issn.1006-1533.2023.16.011]
59. Praveen CK, Srinivasan K. Psychological impact and influence of animation on viewer's visual attention and cognition: a systematic literature review, open challenges, and future research directions. Comput Math Methods Med 2022;2022:8802542 [FREE Full text] [doi: 10.1155/2022/8802542] [Medline: 36092789]
60. Chen J, Wei C, Deng L. Study on the influence of multi-faceted health education intervention inhospital-community-family model on the knowledg-attitude-practice of safe and rational drug use in elderly chronic co-patients. Chin J Hosp Pharm 2020 May 17;40(4):452-455 [FREE Full text] [doi: 10.13286/j.1001-5213.2020.04.19]
61. Zhou J. The impact of pharmaceutical care management on the rationality of medication use in elderly patients with multiple coexisting diseases. J North Pharm 2023 Jul 1;20(7):113-115 [FREE Full text] [doi: 10.3969/j.issn.1672-8351.2023.07.038]
62. Wang HH, Wang LY, Hou WY, Zhao YS. Effects of Motivational Interview on Medication Compliance?Self-management and Blood Glucose Control in Elderly Patieuts with Type 2 Diabetes. Henan Medical Research 2022 Mar 31;31(6):1023-1027 [FREE Full text] [doi: 10.3969/j.issn.1004-437X.2022.06.015]
63. Wan X, Cheng X, Wang Y, Li Y. Study on the influence of personalised health education on health behaviour, knowledge, attitude and practice of medication in elderly patients with chronic disease. Chin J Health Educ 2022 Apr 25;37(12):1135-1138 [FREE Full text] [doi: 10.16168/j.cnki.issn.1002-9982.2021.12.016]

XSL•FO
RenderX

64. Yang YX, Fang XX, Zhang HM, Wang H, Zhang CF, Wang ZP, et al. Effect of life retrospective intervention on belief about medication and perceived happiness in elderly patients with chronic kidney disease. Chin J Mod Nurs 2019 Jul 25;25(6):685-688 [FREE Full text] [doi: 10.3760/cma.j.issn.1674-2907.2019.06.006]
65. Li LS, Tang Q, Wang C, Wu WH, Wu HY, Gao ZQ, et al. The influence of rational drug use management mode of the elderly in community on patients' knowledge, attitude and practice of drug use. Chin Gen Pract 2024 Sep 5;27(25):3129-3135 [FREE Full text] [doi: 10.12114/j.issn.1007-9572.2023.0798]
66. Czaja S, Boot W, Charness N, Rogers W, Sharit J. Improving social support for older adults through technology: findings from the PRISM randomized controlled trial. Gerontologist 2018;58(3):467-477 [FREE Full text] [doi: 10.1093/geront/gnw249] [Medline: 28201730]
67. Hu L, Zhang XF, Song ZZ, Wang KJ, Song J, Jia YC, et al. The impact of family collaborative health education on the cancer-related fatigue, psychological resilience, and self-efficacy of elderly lung cancer patients. Chin J Gerontol 2023 Jun 25:3047-3050 [FREE Full text] [doi: 10.3969/j.issn.1005-9202.2023.12.059]
68. Gong YX, Li Y, Chen XL, Shang Y, Feng QQ. Effects of WeChat carrier information-knowledge-attitude-practice collaborative nursing intervention on health literacy and social support in elderly patients with chronic heart failure. Chin Evidence-Based Nurs 2025 Jan 22;11(2):277-280 [FREE Full text] [doi: 10.12102/j.issn.2095-8668.2025.02.016]

Abbreviations

CG: control group
CONSORT: Consolidated Standards of Reporting Trials
DM: diabetes mellitus
I-CVI: Item-level Content Validity Index
IG: intervention group
IMB: information–motivation–behavioral skills
LOCF: last observation carried forward
RCT: randomized controlled trial
S-CVI: Scale-level Content Validity Index
T2DM: type 2 diabetes mellitus

Edited by N Cahill; submitted 08.Jun.2025; peer-reviewed by ZR Fan, H Almomani, AN Ali; comments to author 20.Oct.2025; revised version received 28.Nov.2025; accepted 11.Dec.2025; published 17.Feb.2026.

Please cite as:
Wang M, Lee K, Thew HZ, Sowtali SN, Jiang Q, Zhao Z, Liu Y, Lim PY
Effectiveness of the Mobile-Based Diabetes Little Helper Video Intervention on Medication Adherence Among Older Adults Living With Type 2 Diabetes Mellitus, Henan, China: Randomized Controlled Trial
J Med Internet Res 2026;28:e78731
URL: https://www.jmir.org/2026/1/e78731
doi:10.2196/78731
PMID:

©Meng Wang, Khuan Lee, Hui Zhu Thew, Siti Noorkhairina Sowtali, Qiuhuan Jiang, Zhengyan Zhao, Yu Liu, Poh Ying Lim. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Added Technology for Weight Management in Cardiac Rehabilitation: Biopsychosocial Factors and Rehabilitees' Talk in a Mixed Method Study

Heli Lahtio[1,2*], MSc; Teemu Paajanen[3*], MSc; Hilkka Korpi[1,4*], PhD; Ari Heinonen[1*], PhD; Tuulikki Sjögren[1*], PhD

[1]Faculty of Sport and Health Sciences, University of Jyväskylä, Jyväskylä, Finland

[2]Physical Activity and Functional Capacity Research Group, LAB University of Applied Sciences, Lahti, Finland

[3]Finnish Institute for Health and Welfare, Helsinki, Finland

[4]Welfare and Culture, Oulu University of Applied Sciences, Oulu, Finland

[*]all authors contributed equally

Corresponding Author:
Heli Lahtio, MSc
Faculty of Sport and Health Sciences
University of Jyväskylä
P.O. Box 35
Seminaarinkatu 15
Jyväskylä, 40014
Finland
Phone: 358 14260 1211
Email: heli.lahtio@gmail.com

Abstract

Background: Hardly any previous studies have examined technology-based interventions among cardiac rehabilitees. The added value of technology and biopsychosocial factors in weight management should be studied more in cardiac rehabilitation.

Objective: This mixed method study aimed to examine what biopsychosocial factors predict WC reduction and what factors arise from the talk of cardiac rehabilitees during 12-month cardiac rehabilitation.

Methods: A total of 59 rehabilitees (mean age 60, SD 6 years; 48/59, 81.3% male) were randomly assigned in pairs into the experimental group (n=29) or the reference group (n=30). Biopsychosocial outcomes were measured at baseline and after 6 and 12 months with a 6-minute walk test (6MWT), the World Health Organization Quality of Life Brief Form questionnaire, body mass, BMI, WC, age, and sex. Rehabilitees' experiences were investigated using focus group interviews. Both groups received conventional cardiac rehabilitation. In addition, the experimental group used technology. Multiple linear regression was used to predict WC change in 0-6 months and 0-12 months. The qualitative data were analyzed with thematic analysis. Integrative analysis was used to create a research-based model.

Results: Group allocation predicted WC reduction (P=.007), with the experimental group showing a greater reduction (b=2.5 cm) than the reference group in the 0-6 month analysis. A significant interaction between the reference group and 6MWT was observed (P=.04), meaning that improvement in 6MWT was associated with WC reduction. Baseline WC was marginally associated with a greater reduction in WC (P=.05). None of the independent variables predicted WC reduction in 0- to 12-month analyses. Three themes arose from the interviews with the experimental group: meaningful factors related to weight management, a goal-oriented approach to weight loss, and monitoring body composition. Two themes emerged from the reference group: motivation for change and unstable weight management. According to the integrative analysis, belonging to the experimental group was important. The research-based model highlighted that the added value of remote technology on WC reduction was largest at the beginning of rehabilitation (0-6 months). The rehabilitees' talk in the experimental group was goal-oriented, and they talked about the achieved changes. In the reference group, the talk remained motivational, and their body composition changes were unstable.

Conclusions: The integrative model emphasizes the differences between the experimental and reference groups in weight management. The use of added remote technology seemed to explain the WC reduction. The rehabilitees reflected on behavioral changes in weight management. In the reference group, improved 6MWT distance was associated with decreased WC, but they

XSL•FO
RenderX

talked about motivation for weight management and unstable changes in body composition. In the future, results should be verified by larger research groups and individual interviews, which can be combined with personal quantitative data.

Trial Registration: ISRCTN Registry ISRCTN61225589; https://www.isrctn.com/ISRCTN61225589

(J Med Internet Res 2026;28:e78347) doi:10.2196/78347

KEYWORDS

weight management; waist circumference; cardiac rehabilitation; remote technology; overweight; obesity

Introduction

Background

Obesity is defined as a condition characterized by excess adiposity, with or without abnormal distribution or function of adipose tissue [1]. Obesity is not well understood. The causes are multifactorial, including genetic, metabolic, psychological, and environmental factors [1]. Obesity is a major risk factor of cardiovascular diseases (CVDs), with, for example, tobacco use, unhealthy diets, physical inactivity, and harmful use of alcohol [2]. Abdominal obesity with other risk factors, such as hypertension and diabetes, is associated with an increased risk of myocardial infarction [3]. Rehabilitees with obesity also experience events related to CVDs earlier, have longer years lived with CVD disability, and have shorter average life span [4].

CVDs are a leading cause of premature mortality worldwide. The incidence of CVDs has doubled between 1990 and 2019. Although CVDs can cause death, many rehabilitees continue to live with these conditions for many years [5]. The years lived with disability represent the number of years people live with a disease while experiencing a reduction in functional capacity. In 2019, the total years lived with disabilities attributed to CVDs amounted to 37 million years, reflecting a 106% increase from 1990 to 2019 [6].

Overweight or obesity is typically measured with BMI [1]. However, BMI does not differentiate fat and lean mass, reflect metabolic health, or consider body fat distribution [1]. Therefore, waist circumference (WC), or body fat percentage, might be more accurate for the evaluation of the perception of adiposity and obesity-related risks [1]. In every BMI category, the larger the WC, the higher the morbidity and mortality risk [1].

WC is a measurement to assess abdominal obesity [7,8]. It is a significant predictor of health [9] and cardiometabolic diseases [10]. High WC is a robust indicator of abdominal body fat, and thus is associated with CVDs [4], cardiovascular mortality [7], and recurrent atherosclerotic CVD, especially among men [11]. Increased WC is also associated with hypertension [9,12], diabetes mellitus, hypercholesterolemia, joint and low back pains, hyperuricemia, and obstructive sleep apnea syndrome [9]. Among people with CVDs, decreasing WC is an important target to prevent recurrence [11], morbidity and mortality [7], as well as decrease other well-known risk factors [7].

Weight loss is a complex and multidimensional issue, and therefore, obesity management needs comprehensive and multidimensional interventions [13]. Weight loss interventions should be individualized, multidisciplinary, and focused on realistic goals [14]. In the study by Rogerson et al [15], the rehabilitees experienced weight loss as an enduring and ongoing task. Weight loss involves physical, cognitive, behavioral, social, and environmental dimensions, which can, at the same time, either assist or inhibit weight loss. It has been stated that weight loss interventions concentrating only on the diet and increasing physical activity (PA) usually fail to achieve significant weight loss [16]. An individually designed diet, PA, and behavioral lifestyle interventions should be implemented to improve outcomes and adherence to weight loss interventions [17]. Psychological factors, such as motivation and self-efficacy, as well as behavioral factors, such as eating behavior, can affect the success of weight loss [18].

Technology-based weight loss interventions among rehabilitees with obesity have been studied widely [19-21], but the number of studies concerning technology-based weight loss interventions among cardiac rehabilitees is limited. However, single studies have been implemented. For example, den Uijl et al [22] found that the rehabilitees in cardiac rehabilitation involving activity trackers and behavioral change techniques lost significantly more weight than rehabilitees in standard cardiac rehabilitation. Technology-based interventions have been studied in preventing CVDs [23-25], in improving risk factors of CVDs [26], and in self-management of CVDs [27]. Positive results have been obtained in reducing CVD outcomes [25] and risk factors [24,26], as well as in the prevention and management of CVDs [23]. In health care, the role of artificial intelligence (AI) is increasing, for example, in medical image analysis, clinical decision support, and analyzing data from technological devices such as smartwatches [28]. In cardiac rehabilitation, AI may support the individuality, engagement, motivation, and self-direction of a cardiac rehabilitee [29,30]. In addition, AI-based interventions may be used in predicting CVDs, which enhance preventive strategies [31].

A systematic review, meta-analysis, and meta-regression study [32] found that web- and mobile-based interventions reduced BMI, WC, and body fat percentage. However, the interventions of the included studies were heterogeneous, and it was difficult to identify the real effects of technology-based interventions on body composition. Only 2 studies of 30 included cardiac rehabilitees, and 2 studies involved rehabilitees with cardiovascular risk factors [32]. A cluster randomized trial (CRT) study among cardiac rehabilitees [33] found that remote technology (Fitbit Charge HR activity monitor and Movendos mCoach) brought added value to WC reduction compared with conventional identical multi-professional rehabilitation without remote technology, but the reasons for this difference are unknown. It is important to obtain diverse and comprehensive knowledge about weight loss interventions among cardiac

XSL•FO
RenderX

rehabilitees and the factors influencing weight management in order to improve cardiac rehabilitation and the functional ability of cardiac rehabilitees in the future.

Objective

This mixed method study aimed to examine what biopsychosocial factors predict WC reduction and what factors arise from the speech of cardiac rehabilitees during 12-month cardiac rehabilitation. Based on the quantitative and qualitative results, an integrative research-based model of biopsychosocial factors and themes for weight management is created.

Methods

Study Design

This study (registration number ISRCTN61225589) was a mixed method study with a CRT and a qualitative study. The study design followed the convergent design, where the quantitative and qualitative data were collected at the same time, independently analyzed, and finally merged to combine the results [34].

The duration of the cardiac rehabilitation, implemented in a Finnish rehabilitation center, was 12 months in total. The measurements and interviews were carried out during in-rehabilitation periods at baseline, at 6 months, and at 12 months.

Recruitment of Rehabilitees

A total of 59 rehabilitees were recruited from the rehabilitees attending a coronary heart disease rehabilitation course between September 2015 and May 2017. Participation in a course, which was funded by the Social Insurance Institution of Finland [35], required a physician's referral. In Finland, rehabilitation courses for various diseases are available to all individuals diagnosed with the condition upon a physician's referral. The rehabilitation courses are nationwide, not limited to a single well-being services county. Generally, rehabilitees can choose the rehabilitation center whose courses they want to participate in. In this study, all applicants to a specific rehabilitation center who met the inclusion criteria were enrolled as participants. The eligibility criteria for cardiac rehabilitation in Finland are being older than 18 years, a doctor's statement recommending rehabilitation, and a diagnosed cardiac disease based on an appropriate diagnostic assessment. Further requirements include the following: the rehabilitee has limitations in work, study, or functional capacity caused by the disease, which impairs the ability to manage in everyday environments. The rehabilitee needs to have sufficient independent functional capacity to move in the rehabilitation environment. In addition, the rehabilitee's initial treatment period has been completed, and the condition is stabilized [34].

The eligibility criteria for this study were being older than 18 years, having a diagnosed coronary heart disease, and independent basic-level management of IT and remote technology applications. To ensure the capability of using technology, the exclusion criteria included serious musculoskeletal disorders, cognitive diseases, and memory diseases that affect essential functional abilities. All rehabilitees fulfilled the inclusion criteria.

The study was conducted in a Finnish Rehabilitation Center, where rehabilitees came from all over Finland. The researchers divided an incoming group into 6 groups. The 2 types of groups were a conventional cardiac rehabilitation group with added remote technology (experimental) and a conventional cardiac rehabilitation group (reference) without added remote technology. The technologies used in this study were Movendos mCoach internet software (Movendos Ltd) and Fitbit Charge HR (Fitbit) accelerometers. The remote technology intervention clusters and reference clusters started in autumn (September to November), winter (December to February), and spring (March to May). The study design has been described in more detail in previously published articles [33,36,37].

Rehabilitees

The mean age of the rehabilitees (n=59) was 60 (SD 6, range 41-66) years. A total of 77% (n=23) of the experimental group, and 86% (n=25) of the reference group were male. The rehabilitees had undergone a coronary angioplasty or coronary artery bypass 3-12 months before the study. At baseline, there were no statistically significant differences in the demographic and clinical characteristics between the experimental and reference groups.

Randomization

Randomization was accomplished within the rehabilitation groups with sealed envelopes. A researcher from the gerontological center of the University of Jyväskylä, who was outside the research group, accomplished the randomization 3 times for 2 consecutive groups, under the supervision of 2 researchers (TS, Heikki Kivistö). A total of 59 rehabilitees were randomly assigned in pairs into (1) the remote technology intervention group (experimental group; n=29) or (2) the reference group (n=30). At the beginning of the rehabilitation, the researchers (TS and Heikki Kivistö) informed the rehabilitees about their rehabilitation group. The rehabilitation was implemented in the rehabilitees' natural rehabilitation environment; it was not possible to blind the rehabilitees and caregivers due to the nature of the rehabilitation intervention and for ethical reasons. The outcome assessors were not blinded to the intervention, but the same educated persons who were not part of the research group performed the anthropometric measurements (a nurse specialized in heart diseases) and the 6-minute walk test (6MWT; a physiotherapist, specialized in heart diseases). The assessors came from outside the research group. The statistician (TP) was blinded.

Ethical Considerations

This study was approved by the Ethics Committee of the Central Finland Health Care District (Dnro: 12 U/2015) on October 15, 2015. After randomization, the rehabilitees signed a written consent form about their participation in the study at the rehabilitation center. Participating in the study was voluntary, and rehabilitees were allowed to withdraw from the rehabilitation program at any time. All collected data were anonymized to ensure participant privacy and confidentiality, and all identifiable information was removed from the data. The

XSL•FO
RenderX

rehabilitees did not receive any compensation for their involvement in the study. In Finland, rehabilitation is free for the rehabilitees, including free accommodation, program activities, and meals. In addition, rehabilitees who are employed receive a rehabilitation allowance for the loss of income due to the absence from work.

Intervention

The rehabilitation courses aimed to promote the rehabilitees' biopsychosocial functional ability and work ability [35]. Evidence-based rehabilitation methods and the Finnish Current Care Guidelines [38] were applied. Rehabilitation consisted of three 5-day in-rehabilitation periods in the rehabilitation center. Between these in-rehabilitation periods, the rehabilitees had two 6-month self-rehabilitation periods, when they followed home exercise programs in their own living environment.

Both the experimental and reference groups received conventional cardiac rehabilitation for 12 months, which included, for example, multidisciplinary rehabilitation, medical examination, physiotherapy, aqua therapy, gym training, and group discussions with a nutritionist, social worker, physiotherapist, psychologist, and physician. The rehabilitees received multi-professional information and pamphlets concerning, for example, CVDs and the management of daily activities, such as diets, PA, social security benefits, self-care, and self-rehabilitation while living with CVDs. From the experimental group rehabilitees, the researcher (Heikki Kivistö) collected individual and group-based PA information, stored in the database of Fitbit Charge HR (Fitbit) accelerometers, and sent the data to the physiotherapists. Based on that information, the rehabilitees received feedback on PA during the in-rehabilitation periods. The accelerometers were designed to be worn daily. After the intervention, all rehabilitees received personal information on PA and other biopsychosocial outcomes in paper reports. The only difference being the use of technology, the studied periods were equal in the experimental and reference groups, and the researchers (TS, Heikki Kivistö) from the University of Jyväskylä were in equal contact by phone with all rehabilitees during the different phases of the study.

The only difference between the reference and experimental groups was the use of technology. The reference groups did not have any technical devices, but they used printed materials and self-monitoring with paper and pen. The experimental groups used technological devices in monitoring their progress, communicating with the instructors, goal setting, self-monitoring, and giving feedback. The interventions have been described in more detail in our previous article [33].

Outcome Measurements

This study used both the primary and secondary outcomes [33,36] of the study project. The primary outcome was a PA assessment [36], and the secondary outcomes were based on the biopsychosocial model: 6MWT, the World Health Organization Quality of Life Brief Form (WHOQOL-BREF) questionnaire, weight, BMI, WC, age, and gender. Additional secondary outcomes included the rehabilitees' experiences collected with interviews. The outcome measurements were collected at baseline and after 6 and 12 months in the rehabilitation center.

6MWT is a reliable and valid field test [39] in which a rehabilitee walks at a self-selected pace as long as possible on a flat surface for a 6-minute period [40]. The heart rate, distance walked, and the symptoms were recorded. Anthropometric measurements included body mass, BMI, and WC. Body mass was measured by using a calibrated floor scale. The rehabilitees wore light clothes, such as T-shirts and shorts or tights, and they were without shoes. BMI was determined by dividing body mass by the height square in meters [41], and WC was measured on bare skin at the midpoint between the lowest rib and the iliac crest [7,42]. The WHOQOL-BREF questionnaire was used to measure quality of life. The questionnaire was performed on a paper format, and the researchers were responsible for completing it. The WHOQOL-BREF contains 26 questions. One question is related to overall health, and another question to overall quality of life. The other 24 questions are divided into 4 domains: physical (7 items), psychological (6 items), social (3 items), and environmental (8 items). Higher scores reflect a higher quality of life [43].

Interviews

Semistructured focus group interviews were performed in the rehabilitation center at every follow-up point. The interview themes were formed by a group of researchers (TS, Arja Piirainen), and the interviews were conducted by one researcher (Heikki Kivistö). The interviews at the 6-month follow-up point were themed around experiences of rehabilitation and technology in the rehabilitee's ordinary life and changes related to lifestyle. Additional themes concerned the personalization of rehabilitation and cooperation during rehabilitation. At the 12-month follow-up point, in addition to the previous themes, the interviews inquired about remote guidance and feedback, and meaningful experiences during rehabilitation. The full interview guides are presented in Multimedia Appendix 1. The interviews were implemented on the principle of an open interview, considering the themes. The interviews were recorded, and the recorded data were transcribed. Altogether, this study examined 312 pages of transcribed data, written in font size 12 and line spacing 1.0, including 4 pages of discussion about weight management in the experimental and reference groups. Saturation was not fully achieved because this study included rehabilitees from the rehabilitation course, where the interviews were part of the rehabilitation program. The schedule for all appointments was planned by the Social Insurance Institute of Finland, and the researchers had no possibility to extend the interviews. However, the rehabilitees in all target groups discussed the same themes, and no new themes emerged. In addition, several rehabilitees mentioned at the end of the focus groups that they had nothing further to add.

At the 6-month follow-up point, altogether, 47 rehabilitees participated (38 men and 9 women), and at the 12-month follow-up point, 30 rehabilitees (21 men and 9 women) participated in the interviews. Table 1 presents the number of rehabilitees in the focus groups.

XSL•FO
RenderX

Table 1. Group interviews with cardiac rehabilitees at 6 and 12 months follow-up points during their rehabilitation program.

	6 months		12 months	
	Men, n	Women, n	Men, n	Women, n
Group 1	7	1	2	1
Group 2	5	3	5	3
Group 3	7	2	7	2
Group 4	8	0	2	0
Group 5	7	1	3	2
Group 6	4	2	2	1
Total	38	9	21	9

Quantitative, Qualitative, and Integrative Analyses

Statistical Analysis

Multiple linear regression was the primary method used for quantitative analysis. Before performing that, the Shapiro-Wilk test was performed to ensure the normality of the data. Multicollinearity was tested with a variance-inflating factor. After these tests, 2 separate dependent variables were analyzed using multiple linear regression: WC change from 0 to 6 months and from 0 to 12 months.

Multiple linear regression was performed with R Studio software (version 2024.12.0+467). First, all independent variables were tested with a dependent variable (WC change). The independent variable with the highest *P* value was iteratively removed in a stepwise backward elimination process. Finally, all statistically significant independent variables were included in the final model. If $P<.05$, predictors were considered statistically significant. Separate regression models were constructed for the 0-6-months and 0-12-months. Positive values indicate reductions, as the change variables (0-6 months and 0-12 months) were computed by subtracting the follow-up measurement from the baseline, meaning that baseline values were higher than at follow-up.

The interaction between the experimental group and 6MWT and the interaction between the reference group and 6MWT were tested in both models. The interaction was included in the model because, in the analysis of individual factors (0-6 months), it seemed significant and interesting. In addition, previous studies have found a negative correlation between 6MWT and body weight or BMI [44,45]. However, previous studies have not examined the interaction between factors related to weight loss and biopsychosocial factors, although the interaction seems statistically and clinically relevant from the point of view of the complex topic.

Qualitative Analysis

Thematic analysis of the qualitative data aimed to discover what the rehabilitees talk about their weight management. According to Braun and Clarke [46], one option for thematic analysis is to provide a detailed and nuanced account of one theme, or a group of themes within the data. Typically, this is related to a specific research question or area of interest. This thematic analysis focused on weight management themes, was inductive, and strongly linked to data [46].

Thematic analysis comprises six phases: (1) familiarizing with the data, (2) generating initial codes, (3) searching for themes, (4) reviewing themes, (5) defining and naming themes, and (6) producing the report [46,47]. In this study, researcher triangulation (HL, HK, TS) was used in every phase of the thematic analysis to ensure the quality of the qualitative analysis. In the first phase, the researchers became familiar with the data by reading it carefully and repeatedly. In the second phase, the data was coded. In the third phase, the codes were combined, compared, and analyzed. The analysis was inductive, and the themes were formed from the data (Table 2). In the fourth phase, the themes were processed by reviewing the coded data and by deciding whether the codes formed a coherent pattern or whether they would need reworking. After that, the researchers evaluated the themes in relation to the entire data set and decided whether the themes fit into the thematic map reflecting the meaning of the data. In this phase, the researchers reread the data to examine the themes and recoded the data arising from the newly created themes. In the fifth phase, the researchers created a definition and description of each theme, and in the final phase, they wrote the final analysis and description of the findings. At the initial phase, AI was used for language proofreading.

XSL•FO
RenderX

Table 2. An example of generating themes in qualitative analysis.

Quotes	Coding	Subthemes	Themes	Group and pages
"I've even gotten my wife exercise quite well."	Has exercised with wife...	Weight management together	The meaningful factors related to weight management	G3, p. 15
"She's lost more weight than me."	...and wife has lost more weight than husband	Weight management together	The meaningful factors related to weight management	G3, p. 15
"It would be a personal (diet). So it would support everyone's exercise, and generally help with losing weight as well."	Personal diet in supporting exercising and weight loss	Personal diet in supporting exercise, weight loss and rehabilitation	The meaningful factors related to weight management	G3, p. 17
"But just that it would be personal and concrete – that would support this thing much more strongly."	Personal diet in supporting the rehabilitation	Personal diet in supporting exercise, weight loss and rehabilitation	The meaningful factors related to weight management	G3, p. 17
"Five days, what you get here. It's kind of... Should be more like rehabilitation, or kind of exercise."	There should be more physical activity in the rehabilitation	Increasing the amount of physical activity in rehabilitation to prevent weight gain caused by prolonged sitting	The meaningful factors related to weight management	G5, p. 19
"Well, you could send a message there, like why people are gaining weight. It's because we sit too much here. Yeah. We should just exercise more, and then..."	Sitting in the rehabilitation increases weight	Increasing the amount of physical activity in rehabilitation to prevent weight gain caused by prolonged sitting	The meaningful factors related to weight management	G5, p. 19

Integrative Analysis

In the integrative analysis, the main results from the quantitative and qualitative analyses were integrated. The integrative analysis was made following the steps of Skamagki et al [34]. Joint display was used in the first phase to organize the integration of quantitative and qualitative data. In the second phase, the linkages between quantitative and qualitative findings were defined. In the third phase, establishing relationships, the researchers looked for inconsistencies, alignments, or conflicting findings in the relationships between quantitative and qualitative data [34]. The fourth phase included interpreting and reporting the results.

Results

Overview

The results are presented in three phases: (1) quantitative, (2) qualitative, and (3) integrative study results. The results consider the outcomes from the experimental and the reference group, as well as the differences between them identified by both analysis methods. The results were used to develop a research-based model related to weight management and cardiac rehabilitation.

Biopsychosocial Factors Predicting WC Reduction

In multiple linear regression, group allocation to the experimental and the reference groups significantly predicted WC reduction (*P*=.007), with the experimental group showing a greater reduction (b=2.5 cm) compared with the reference group. A significant interaction between the reference group and 6MWT was observed (*P*=.04): in the reference group, greater improvement in 6MWT was associated with a greater reduction in WC. No such association was observed in the experimental group (Table 2).

In addition, baseline WC was marginally associated with a greater reduction of WC (*P*=.05) in the 0- to 6-month analysis. According to the model, each 1 cm higher baseline WC was associated with a 0.07 cm greater reduction in WC.

None of the independent variables significantly predicted WC reduction in the 0-12 month analyses. For comparison, Table 3 also presents the 0- to 12-month results using the same model that was found to be optimal in the 0- to 6-month analysis (Tables 3 and 4).

Table 3. Results of multiple linear regression on waist circumference (WC) at 0-6 and 0-12 months.

Predictor	Estimate (0-6 months)	*t* test (*df*)	*P* value	Estimate (0-12 months)	*t* test (*df*)	*P* value
Experimental group (vs reference group)	2.52	2.83 (42)	.007	2.88	1.35 (21)	.18
6MWT[a]	–0.02	–1.91 (42)	.06	–0.06	–1.55 (21)	.14
WC at baseline (cm)	0.07	1.98 (42)	.05	0.06	1.38 (21)	.18
Experimental * 6MWT[b]	0.04	2.16 (42)	.04	0.05	1.29 (21)	.21

[a]6MWT: 6-minute walk test.

[b]Experimental * 6MWT is an interaction term between the intervention and 6MWT.

XSL•FO
RenderX

Table 4. Regression model fit and predictive accuracy statistics.

Metric	0-6 months	0-12 months
R^2	0.23	0.19
Adjusted R^2	0.16	0.03
F test (*df*)	3.16 (4, 42)	1.22 (4, 21)
P value	<.02	.33
Residual standard error	2.64	2.91

Themes About Weight Management From the Talk of the Experimental Group

Three themes arose from the talk of the rehabilitees in the experimental group: the meaningful factors related to weight management, the goal-oriented approach to weight loss, and the monitoring of body composition. The group numbers and intervention follow-up points are presented in brackets after the quotations.

The talk about the meaningful factors related to weight management included the following subthemes: weight management together, personal diet instructions, and increasing PA in rehabilitation. The rehabilitees hoped for a personal diet which would support their weight loss, rehabilitation, and PA, as stated in the following example from a group interview:

> *It would be personalized (diet). So, it would still support everyone's physical activity and. And in general, support weight loss.* [Group 3; 6 months]

In addition, the rehabilitees talked about the amount of sitting during rehabilitation, resulting in a weight increase. Therefore, the rehabilitees hoped for more PA during the rehabilitation, as in the following example:

> *Well, you could send a message there asking why the fat keeps increasing. Because we sit here too much when we should... yeah... we should just move more, and then...* [Group 5; 12 months]

The goal-oriented approach to weight loss theme included the following subthemes: weight loss goals during rehabilitation, benefits of weight loss in lowering blood pressure, and the positive changes in muscle mass and body fat percentage. The rehabilitees mentioned that many rehabilitees aim for weight loss during rehabilitation, as in the following example:

> *Especially the goals—probably many have set a goal like 'I want to lose this much weight.* [Group 3; 6 months]

In addition, the rehabilitees remarked that as a result of weight loss, blood pressure decreases. This is illustrated by the following quote:

> *And of course, when the weight drops, it lowers blood pressure and everything else, and...* [Group 3; 6 months]

The monitoring body composition theme included 5 different perspectives. First, the rehabilitees mentioned that the weight remained the same because muscle weighs more than fat. Second, they discussed the results of their earlier body composition measurements. Third, they expressed a wish that measuring body composition would be included in the rehabilitation program. Fourth, they talked about their doubts about the reliability of the body composition measurement, and fifth, the positive changes in their body composition. This is illustrated by the following quote:

> *Why don't they measure at the beginning and the end of the Social Insurance Institute of Finland course? That would tell the truth, then.* [Group 2; 12 months]

Themes About Weight Management From the Talk of the Reference Group

Two themes arose from the talk of the rehabilitees in the reference group: motivation for change and unstable weight management. The motivation for change theme comprised 2 subthemes: the motivating effect of the disease on weight management and lifestyle changes. The following quote illustrates how the rehabilitees talked about the effect of a disease on motivation:

> *Yeah, the motivation came from the heart attack, so it sort of started from there. It was a bit like my own desire to lose some weight and pay attention to what I'm eating and so on.* [Group 4; 6 months]

The rehabilitees mentioned motivating issues that affect rehabilitation, such as a healthy body, family, and the ability to work. This is illustrated by the following example:

> *Well, the motivation comes from wanting to get this, uh, body in shape so that I could be around as long as possible with my family and spend time with the kids, and, you know, do a bit of work on the side too.* [Group 4; 6 months]

The second theme, unstable weight management, contained contradictory discussions about weight change. Some rehabilitees achieved positive changes in their weight, while others did not report any changes in their weight, which disappointed them. An example is presented in the following quote:

> *I've lost a little weight and, and my endurance was better, and…* [Group 1; 12 months]

Some rehabilitees gained more weight during the rehabilitation, and they talked about the hope of losing the weight they gained. This is illustrated by the following examples:

> *Eating habits have changed, but the weight hasn't gone down. That's disappointing, for sure.* [Group 1; 12 months]

XSL•FO
RenderX

I've of course had a bit of a setback, I've gained a couple of kilos. But maybe it'll come off during the fall. [Group 1; 12 months]

Integrative Results of the Quantitative and Qualitative Data

The integrative results of the quantitative and qualitative results are presented in Figure 1. The left side of the figure describes the results of the experimental group, and the right side describes the results of the reference group. The center of the figure describes the quantitative results at 0-6 and 0-12 month follow-up points. The sides of the figure describe the themes of the qualitative data.

The integrative model emphasizes that the experimental and reference groups differed in quantitative and qualitative results in terms of factors related to weight management. Whether the rehabilitee belonged to the experimental or the reference group was important in this study. The group was associated with WC reduction at 0-6 months (the center of Figure 1). The rehabilitees in the experimental group talked more about their weight management (left side of Figure 1) compared with the rehabilitees in the reference group (right side of Figure 1), and they seemed to have advanced in the process of change. The rehabilitees talked about elements of goal-oriented weight management, and they reflected on the changes they had achieved. In the reference group, improved 6MWT distance was associated with WC reduction at 0-6 months. In this group, the talk was more of a motivational speech, and the rehabilitees' body composition changes were unstable. Furthermore, body composition changes arose from the talk of both groups. In the experimental group, the changes in weight management were positive, but in the reference group, the changes were unstable.

Figure 1. Integrative results with the experimental group on the left and the reference group on the right. 6MWT: 6-minute walk test; WC: waist circumference.

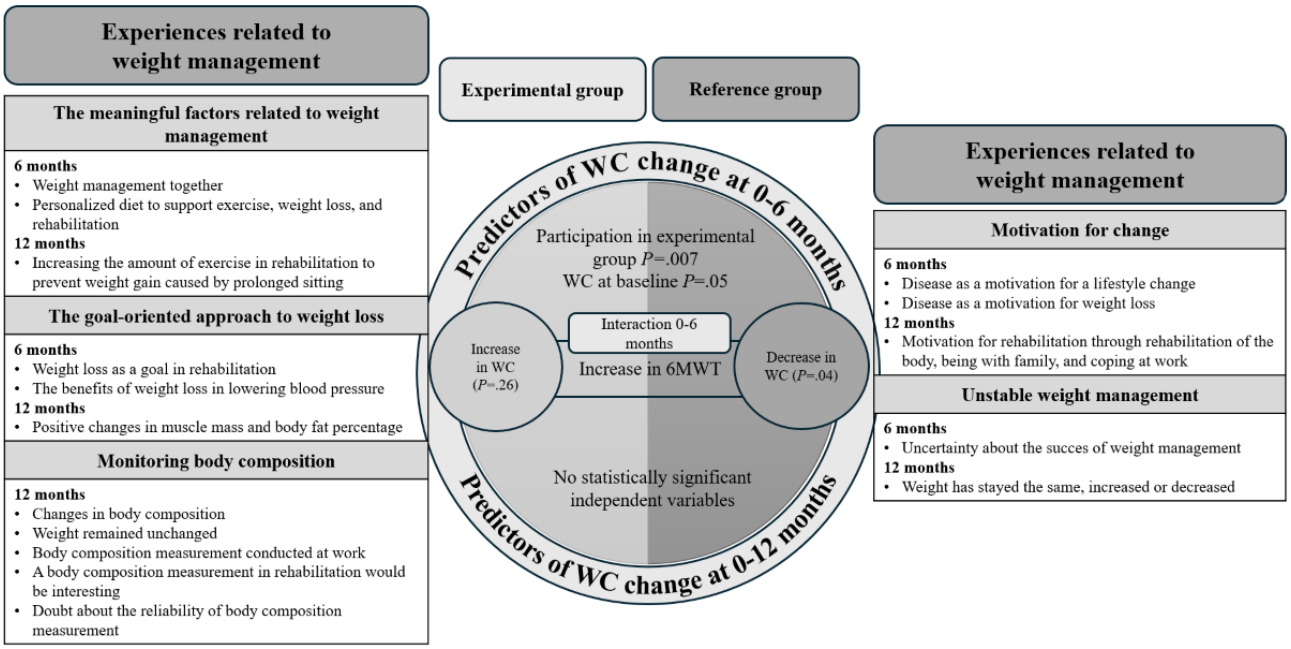

Discussion

Principal Findings

This mixed method study aimed to examine, during a 12-month cardiac rehabilitation, which biopsychosocial factors are related to WC reduction and what themes of weight management are meaningful for cardiac rehabilitees. A main finding was that allocation to the experimental group predicted greater WC reduction compared with the reference group. The integrative model indicated that the experimental and reference groups differed in the quantitative and qualitative results relating to weight management. Adding remote technology to conventional rehabilitation seemed to explain the WC reduction. The talk of rehabilitees was characterized by behavioral change and reflection on weight management. In the reference group, which did not use remote technology, an improved 6MWT distance in meters was associated with decreased WC. The rehabilitees' talk dealt with motivation for weight management and unstable changes in body composition.

Our findings indicate that technology-based interventions are associated with WC reduction. However, other biopsychosocial variables were not associated with WC reduction. A similarity between the groups was found when a larger WC at baseline might predict significant WC reduction at 0-6 months (P=.05). This is an important finding as decreasing WC is essential in weight management because abdominal obesity is associated with CVDs [3]. This finding is in line with previous studies of the importance of baseline value to weight management changes during intervention [48-51]. Iwayama et al [48] found that a higher baseline WC was associated with a larger reduction of WC. Other studies have examined the association between early weight loss and long-term weight loss. Al-Abdullah et al [49] discovered that higher baseline weight and early weight loss were predictors of successful weight loss during an intervention. Höchsmann et al [50] and Farage et al [51] observed that initial weight loss predicted long-term weight loss. However, only one of these previous studies considered WC [48], and none of the studies considered cardiac rehabilitees.

XSL•FO
RenderX

This study found a statistically significant interaction between increased walking distance in 6MWT and decreased WC in the reference group, but not in the experimental group. The change in 6MWT was not statistically significant on its own, meaning that within the groups, the change in the walking distance was not associated with the WC change. One possible explanation for this finding is that the rehabilitees in the experimental group improved their nutrition or other factors that could not be captured with the quantitative measurements. Increased WC has been associated with decreased walking distance in 6MWT [52,53] among rehabilitees with rheumatoid arthrosis [52] and type 2 diabetes [53]. Among cardiac rehabilitees, Lenasi et al [54] found that age, WC, and New York Heart Association classification were predictors of 6MWT in rehabilitees with coronary artery disease. Similarly, Pepera et al [55] found that BMI >25 kg/m^2 and age were independent predictors of poor 6MWT performance in rehabilitees with chronic heart failure. However, Sharma et al [56] studied healthy individuals and discovered a relation between 6MWT with BMI and waist-to-hip ratio. They found that when BMI increased, the 6MWT walking distance also increased. However, when the waist-to-hip ratio increased, the 6MWT walking distance decreased. Results regarding the interaction between 6MWT and other variables are contradictory, and more studies are needed to determine the relationship between 6MWT and weight management-related outcomes.

In this study, the significant predictors of WC change were statistically significant at the 0-6 month follow-up point, but none of the predictors were statistically significant at the 0-12 month follow-up point. Previous studies have found that the effects of interventions are initially more substantial but tend to diminish over the long term. In addition, typically engagement wanes over time as initial enthusiasm diminishes. Whiteley et al [57] found that during 6 months, engagement on technology-based weight loss intervention decreased 33%. In addition, Liljeroos et al [58] and Beleigoli et al [59] found that technology-based interventions were more effective in the short term, but the changes did not last in the long term.

The rehabilitees in the experimental group were more advanced in their weight management reflection compared with the rehabilitees in the reference group. The experimental group discussed the factors related to weight management, a goal-oriented approach to weight loss, and monitoring body composition. They talked about their body composition changes objectively, and their discussion was goal-oriented. The talk in the reference group was rather motivational, including motivation for change and unstable weight management. In line with the transtheoretical model, the rehabilitees in the reference group are in the preparation phase, in which the rehabilitees intend to take an action toward a change [60]. The rehabilitees in the experimental group are in the action phase. They have made changes, and their behavior change has been translated into action [61].

Previous studies have not compared the views of the rehabilitees in experimental and reference groups where the only difference was the use of technology. However, previous studies have examined the rehabilitees' experiences of technology in weight management. Our results have similarities with previous studies. Lambert et al [62] found that an e-coach was beneficial in motivating the rehabilitees for a behavior change at the initial phase. However, the engagement diminished over time. Følling et al [63] found that activity monitors can strengthen or undermine the attempt at weight loss. Some rehabilitees mentioned that activity monitors motivated them to maintain PA and diet, providing proof of their effort. However, some rehabilitees felt that activity monitors highlighted their failure to change their lifestyle.

The rehabilitees in the experimental group discussed personalized diets in supporting weight loss, exercise, and rehabilitation. Previous studies and guidelines have stated that individually tailored weight loss programs can be important in weight loss [21,64,65]. In addition, the rehabilitees in this study mentioned the importance of social support in weight loss. Previous studies have found that social support from family and friends [57,59] as well as peer support is important in the weight loss process [62,66,67]. In addition, both groups talked about the effect of being overweight on diseases. The rehabilitees in the experimental group discussed the benefit of weight loss in lowering blood pressure, and the rehabilitees in the reference group discussed the motivation for weight loss arising from the cardiac disease. This finding is consistent with the study by Spreckley et al [68], who found that intrinsic motivators include the desire to improve health and the effort to improve weight-related conditions.

The experimental group talked about a goal-oriented approach to weight loss and weight loss as a goal in their rehabilitation. The rehabilitees in the reference group did not discuss their weight loss goals. The goal of weight loss might be imposed from outside, for example, by health care professionals, and people with obesity might assume that their goal should be weight loss. According to the study by Liu et al [69], external motivation is the key to initiating dietary changes. However, internal motivation is important for sustaining the behavioral change. This is in line with the self-determination theory, which suggests that intrinsic motivation, which is the most autonomous form of motivation, comes from a rehabilitee's own interests, enjoyment, and satisfaction. Extrinsic motivation comes from outside (for example, from health care providers, social media, and society). When behavior is driven by external forces, long-term changes are typically unlike [70], whereas intrinsic goals result in improved self-regulation and long-term outcomes [71]. It might be important that the initial goal of weight loss comes, for example, from health care professionals, but for long-term changes, internal motivation is essential.

This study is the first mixed method study exploring technology-based weight loss intervention in cardiac rehabilitation. The study focuses on WC and incorporates several aspects that have not been studied before. For example, Dandeneau et al [72] performed a mixed method weight loss study on cardiac patients, but concentrated on adherence to the weight loss program as well as the strengths and limitations of the program. Liljeroos et al [58] conducted a mixed method study on the effects of a mHealth tool for rehabilitees with heart failure. However, they concentrated on the effectiveness and experiences of the mHealth tool with scales and questionnaires.

XSL•FO
RenderX

Leung et al [73] examined, with mixed method study, the psychological factors of dietary and PA adherence in the early weight maintenance phase. However, they concentrated on participants with overweight and obesity, not cardiac rehabilitees.

In health care, obesity-related measurements are universal, and they categorize the rehabilitees to certain classes. This study also used commonly used measures for assessing weight loss. However, these measurements can be very embarrassing for the rehabilitees [74,75]. In behavior change interventions, these measurements are not necessarily optimal because of the stress and anxiety they may cause to the rehabilitees. According to Phelan et al [76], a welcoming and less threatening environment can be created by reducing focus on body weight and concentrating on obesity-related diseases and conditions, as well as encouraging behaviors that improve health and well-being. Anttila et al [37] formed 4 biopsychosocial profiles in technology-based cardiac intervention. The "feeling outsider" group had a high-risk behavior related to inactive behavior and being overweight. The "being uninterested" group had low PA levels and a poor self-assessed physiological quality of life. Both groups had large WC compared with the recommended value, and 6MWT was below the recommended value. The "reflecting benefit" group had high self-efficacy for exercise and an interest in health technology. The "enthusiastic users" group had high self-determination in exercise behavior but lacked self-efficacy in PA. The WC of the rehabilitees in these groups was closer to the recommended value, and their 6MWT was very close to the recommended value [37]. Individual measurements that do not necessarily concentrate on weight should be developed. This view has not been studied, so future studies should evaluate different measurements for overweight rehabilitees.

Because CVDs are a public health challenge and obesity is a risk factor for CVDs [2], weight management should be included in cardiac rehabilitation. However, it has not been a primary focus in there [77]. Technology may increase the engagement in cardiac rehabilitation [78] and lifestyle changes [79] as well as weight management [80]. Therefore, it should be considered to include technological components in cardiac rehabilitation and weight management. Health care systems are undergoing transformation by increasing the usability of technological devices. With technological devices, it is possible to achieve clinical and environmental benefits such as improving disease management, adherence, and access to health care. From an environmental view, digital health technologies decrease carbon emissions, waste generation, and energy consumption [81].

Our studies [32,33] produce gradually deepening information on the effectiveness of the added value of remote technology. Our previous systematic review with meta-analysis and meta-regression highlighted that previous RCT studies have compared highly heterogeneous experimental and control groups [32]. They have not compared experimental and control groups where the only difference was the use of technology. According to the findings of our CRT study [33], rehabilitation interventions using technology decreased WC more noticeably compared with the reference group without technology. The results from this study might indicate that technology brings added value to conventional rehabilitation and may support weight management, but it requires more studies in the future. To our best knowledge, only a few studies have compared experimental and control groups where the only difference is the technology use [25,82]. However, other studies have compared technology-based interventions to very heterogeneous control groups, which do not consider the added value of technology.

Future studies should examine the added value of technology by comparing interventions that are otherwise similar, but one group uses technology, and the other group does not. In addition, future studies should investigate the predictors of WC reduction and weight management, as well as weight management among cardiac rehabilitees.

In the future, AI is predicted to have a major role in smart health care. Smart health care offers several benefits by enhancing diagnostic accuracy, automating clinical workflows, and enabling personalized treatment strategies [28]. AI apps may support the individuality of cardiac rehabilitees' rehabilitation, as well as their engagement, motivation, and self-direction [29,30]. In addition, AI-related solutions may enhance interaction between health care professionals and rehabilitees, support the monitoring of rehabilitation progress [29,30,83], and provide information on, for example, arrhythmias, heart rate variability, and PA [84,85]. However, research related to AI and rehabilitation is still largely limited to descriptive literature reviews [83-87] or studies conducted with non-experimental research designs [29,30,79]. At present, there is a substantial need for more precise evidence on effectiveness—obtained through randomized controlled trial designs—related to technology, including AI, and cardiac rehabilitation [88].

Future studies should investigate AI-based weight loss interventions among cardiac rehabilitees. With AI, it could be possible to, for example, personalize the interventions, provide real-time feedback, and improve adherence [28], which are key components in weight loss interventions. In cardiac rehabilitation, AI may be used in predicting CVDs [31], support early CVD detection, and enhance preventive strategies [31]. Furthermore, explainable AI may improve trust and accountability in health care, which are essential in technology-based interventions [89].

Strengths and Limitations

This study has several strengths. First, it analyzed biopsychosocial factors broadly with multiple linear regressions and deepened the understanding of weight management with interviews. In addition, quantitative and qualitative results were combined in an integrative analysis. Second, the comprehensive cardiac rehabilitation intervention was implemented by a multiprofessional team. Third, the analyses and the randomization process were carefully implemented. The quantitative analyses were performed in collaboration with a statistician (TP), a principal investigator (TS), and a researcher (HL). The qualitative analyses used researcher triangulation throughout the process (HL, HK, and TS).

XSL•FO
RenderX

The strength of this study is that the analyses were discussed collaboratively by the entire research team, and the decisions were reached collectively. In the first linear regression analyses, BMI was included. However, BMI overwhelmed the effects of all other dependent variables. Therefore, BMI was excluded from linear regression. WC was considered more important than BMI, as previous studies have shown that in every BMI category, the larger the WC, the greater the risk of morbidity or mortality [1]. Therefore, WC was considered a more important factor than BMI in this study. One more strength of the study was its CRT design, which was integrated with a longitudinal interview component. To our knowledge, no similar mixed method studies have been conducted before.

A weakness of this study was the relatively small sample size. The protocol for cardiac rehabilitation, which is standardized by the Social Insurance Institute of Finland, and which is also responsible for rehabilitation costs, determined the sample size. The data collection took place over a 1-year period and included all cardiac rehabilitation groups participating in rehabilitation during that time. Due to the study schedule, it was not possible to collect a larger sample. Additionally, the group size was limited to a maximum of 10 rehabilitees. Therefore, the number of rehabilitees was limited. Because of the small sample size, care must be taken in generalizing the results beyond the target population. The rehabilitees in this study were residents across Finland, which increases the generalizability of this study in Finland. Globally, the results can be generalized in high-income countries. However, in low-income countries, for example, access to medical treatment is not necessarily possible. In addition, the number of cardiac rehabilitation studies is limited in low-income countries [90].

The characteristics of the rehabilitees in this study are similar to those in previous studies. For example, the mean age of the rehabilitees in this study was 60 years, and most of the rehabilitees were male. In previous studies, the mean age of the participants varied from 55 to 63 years [25,82,91-93]. In all these studies, most of the rehabilitees were male [25,82,91-93]. The higher proportion of men effects the generalizability of the results, as these results cannot be directly generalized to women. Fat distribution typically differs between women and men. In addition, biopsychosocial factors behind weight management may differ. For example, the prevalence of obesity is typically higher among women. However, men typically have more visceral fat deposits than women [94].

This study applied a qualitative thematic analysis to group interviews, which enabled the inclusion of a wider range of rehabilitees compared with individual interviews. However, focus group interviews may have affected the level of the rehabilitees' talk by decreasing the discussion about weight management. A total of 312 pages of transcribed data were included in the qualitative study, but only 4 pages reflected weight management. Personal interviews would have been more reliable and made it possible to examine the experiences of the rehabilitees at an individual level. Despite the limited scope of discussion, several themes could be distinguished from the data.

The measurements used in this study are widely used but have some limitations. WC can vary across populations and between sexes. Another limitation is that WC does not reflect the accumulation of subcutaneous and visceral fat [1]. 6MWT and WHOQOL-BREF have limitations as well. Even though 6MWT is reliable and valid, small changes in the methodology of the test may affect the results [39]. The learning effect might be associated with improved walking distance. However, 6MWT has been validated for CVDs, and it can provide information about physiological reserve and prognosis [95]. WHOQOL-BREF is a reliable and valid questionnaire, but some studies have reported a ceiling effect [96-98]. However, the reliability of these measurements improved when the same nurse performed the body composition measurements at every measurement point, and one physiotherapist performed the measurements of 6MWT at every follow-up point.

Conclusions

Both quantitative and qualitative results showed that technology-based intervention supported weight management. The integrative model emphasizes that the experimental and reference groups differed in quantitative and qualitative results in terms of factors related to weight management. The added use of remote technology to conventional rehabilitation seemed to explain the observed positive changes in WC. In addition, rehabilitees' talk reflected behavioral changes in weight management, and the rehabilitees of the experimental group were further in their process of change when compared with the reference group. Technological devices may improve the engagement of cardiac rehabilitees in lifestyle changes and weight management and thus improve the well-being of the rehabilitees. In the future, the results should be verified for use by larger research groups, and interviews should be conducted as individual interviews that can be combined with personal quantitative data to achieve more personalized results. In addition, future studies should examine AI-based weight loss interventions for cardiac rehabilitees.

Acknowledgments

The authors would like to thank the Rehabilitation Center of Peurunka, all rehabilitees, and rehabilitation professionals. The authors also thank the project researcher, Heikki Kivistö, for enabling this study and Arja Piirainen for helping with forming the interview themes. Finally, the authors would like to thank Anneli Rinnevalli for language revision. This study was supported by the Social Insurance Institute of Finland. Copilot was used to check and correct English grammar. In addition, Copilot was used to translate previous studies in Finnish and to ensure the right understanding of the studies.

XSL•FO
RenderX

Data Availability

All data generated or analyzed during this study are included in this published article.

Funding

This study was funded by the Social Insurance Institute of Finland and the JYPE research grant.

Authors' Contributions

TS and AH contributed to the conceptualization of the study. TS was responsible for data curation, funding acquisition, project administration, and provision of resources. HL, TP, HK, AH, and TS conducted the formal analysis, investigation, methodology development, and validation. Supervision was provided by TS, HK, and AH. Visualization was carried out by HL, TP, and TS. HL prepared the original draft of the manuscript, and HL, TP, HK, AH, and TS contributed to the review and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Focus group themes at 6- and 12-months follow-up points.
[DOCX File , 42 KB - jmir_v28i1e78347_app1.docx]

References

1. Rubino F, Cummings DE, Eckel RH, Cohen RV, Wilding JPH, Brown WA, et al. Definition and diagnostic criteria of clinical obesity. Lancet Diabetes Endocrinol 2025 Mar;13(3):221-262. [doi: 10.1016/S2213-8587(24)00316-4] [Medline: 39824205]
2. Cardiovascular diseases (CVDs). World Health Organization. 2021. URL: https://www.who.int/europe/news-room/fact-sheets/item/cardiovascular-diseases [accessed 2024-10-10]
3. Teo KK, Rafiq T. Cardiovascular risk factors and prevention: a perspective from developing countries. Can J Cardiol 2021 May;37(5):733-743 [FREE Full text] [doi: 10.1016/j.cjca.2021.02.009] [Medline: 33610690]
4. Powell-Wiley TM, Poirier P, Burke LE, Després JP, Gordon-Larsen P, Lavie CJ, American Heart Association Council on LifestyleCardiometabolic Health; Council on CardiovascularStroke Nursing; Council on Clinical Cardiology; Council on EpidemiologyPrevention;Stroke Council. Obesity and cardiovascular disease: a scientific statement from the American heart association. Circulation 2021 May 25;143(21):e984-e1010 [FREE Full text] [doi: 10.1161/CIR.0000000000000973] [Medline: 33882682]
5. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, GBD-NHLBI-JACC Global Burden of Cardiovascular Diseases Writing Group. Global burden of cardiovascular diseases and risk factors, 1990-2019: update from the GBD 2019 study. J Am Coll Cardiol 2020 Dec 22;76(25):2982-3021 [FREE Full text] [doi: 10.1016/j.jacc.2020.11.010] [Medline: 33309175]
6. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the global burden of disease study 2019: a systematic analysis for the global burden of disease study 2019. Lancet 2021 Dec 19;396(10267):2006-2017 [FREE Full text] [doi: 10.1016/S0140-6736(20)32340-0] [Medline: 33275908]
7. Ross R, Neeland IJ, Yamashita S, Shai I, Seidell J, Magni P, et al. Waist circumference as a vital sign in clinical practice: a consensus statement from the IAS and ICCR working group on visceral obesity. Nat Rev Endocrinol 2020 Mar;16(3):177-189 [FREE Full text] [doi: 10.1038/s41574-019-0310-7] [Medline: 32020062]
8. Fang H, Berg E, Cheng X, Shen W. How to best assess abdominal obesity. Curr Opin Clin Nutr Metab Care 2018 Sep;21(5):360-365 [FREE Full text] [doi: 10.1097/MCO.0000000000000485] [Medline: 29916924]
9. Darsini D, Hamidah H, Notobroto HB, Cahyono EA. Health risks associated with high waist circumference: a systematic review. J Public Health Res 2020 Jul 03;9(2):1811 [FREE Full text] [doi: 10.4081/jphr.2020.1811] [Medline: 32728557]
10. Durrer Schutz D, Busetto L, Dicker D, Farpour-Lambert N, Pryke R, Toplak H, et al. European practical and patient-centred guidelines for adult obesity management in primary care. Obes Facts 2019;12(1):40-66 [FREE Full text] [doi: 10.1159/000496183] [Medline: 30673677]
11. Mohammadi H, Ohm J, Discacciati A, Sundstrom J, Hambraeus K, Jernberg T, et al. Abdominal obesity and the risk of recurrent atherosclerotic cardiovascular disease after myocardial infarction. Eur J Prev Cardiol 2020 Dec;27(18):1944-1952. [doi: 10.1177/2047487319898019] [Medline: 31958380]
12. Sun JY, Ma YX, Liu HL, Qu Q, Cheng C, Kong XQ, et al. High waist circumference is a risk factor of new-onset hypertension: evidence from the China health and retirement longitudinal study. J Clin Hypertens (Greenwich) 2022 Mar;24(3):320-328 [FREE Full text] [doi: 10.1111/jch.14446] [Medline: 35188335]

XSL•FO
RenderX

13. Khattab R. Weight loss programs: why do they fail? a multidimensional approach for obesity management. Curr Nutr Rep 2024 Sep;13(3):478-499. [doi: 10.1007/s13668-024-00551-x] [Medline: 38861120]
14. Yumuk V, Tsigos C, Fried M, Schindler K, Busetto L, Micic D, Obesity Management Task Force of the European Association for the Study of Obesity. European guidelines for obesity management in adults. Obes Facts 2015;8(6):402-424 [FREE Full text] [doi: 10.1159/000442721] [Medline: 26641646]
15. Rogerson D, Soltani H, Copeland R. The weight-loss experience: a qualitative exploration. BMC Public Health 2016 May 04;16:371 [FREE Full text] [doi: 10.1186/s12889-016-3045-6] [Medline: 27142984]
16. Welsh A, Hammad M, Piña IL, Kulinski J. Obesity and cardiovascular health. Eur J Prev Cardiol 2024 Jun 03;31(8):1026-1035 [FREE Full text] [doi: 10.1093/eurjpc/zwae025] [Medline: 38243826]
17. Cornier MA. A review of current guidelines for the treatment of obesity. Am J Manag Care 2022 Dec;28(15 Suppl):S288-S296 [FREE Full text] [doi: 10.37765/ajmc.2022.89292] [Medline: 36525676]
18. Pigsborg K, Kalea AZ, De Dominicis S, Magkos F. Behavioral and psychological factors affecting weight loss success. Curr Obes Rep 2023 Sep;12(3):223-230. [doi: 10.1007/s13679-023-00511-6] [Medline: 37335395]
19. Islam MM, Poly TN, Walther BA, Jack Li YC. Use of mobile phone app interventions to promote weight loss: meta-analysis. JMIR Mhealth Uhealth 2020 Jul 22;8(7):e17039 [FREE Full text] [doi: 10.2196/17039] [Medline: 32706724]
20. Rumbo-Rodríguez L, Sánchez-SanSegundo M, Ruiz-Robledillo N, Albaladejo-Blázquez N, Ferrer-Cascales R, Zaragoza-Martí A. Use of technology-based interventions in the treatment of patients with overweight and obesity: a systematic review. Nutrients 2020 Nov 26;12(12):3634 [FREE Full text] [doi: 10.3390/nu12123634] [Medline: 33255982]
21. Protano C, De Giorgi A, Valeriani F, Mazzeo E, Zanni S, Cofone L, et al. Can digital technologies be useful for weight loss in individuals with overweight or obesity? a systematic review. Healthcare (Basel) 2024 Mar 16;12(6):670 [FREE Full text] [doi: 10.3390/healthcare12060670] [Medline: 38540634]
22. den Uijl I, van den Berg-Emons RJG, Sunamura M, Lenzen MJ, Stam HJ, Boersma E, et al. Effects of a dedicated cardiac rehabilitation program for patients with obesity on body weight, physical activity, sedentary behavior, and physical fitness: the OPTICARE XL randomized controlled trial. Phys Ther 2023 Sep 01;103(9):pzad055 [FREE Full text] [doi: 10.1093/ptj/pzad055] [Medline: 37265452]
23. Kozik M, Isakadze N, Martin SS. Mobile health in preventive cardiology: current status and future perspective. Curr Opin Cardiol 2021 Sep 01;36(5):580-588 [FREE Full text] [doi: 10.1097/HCO.0000000000000891] [Medline: 34224437]
24. Burke LE, Ma J, Azar KMJ, Bennett GG, Peterson ED, Zheng Y, American Heart Association Publications Committee of the Council on EpidemiologyPrevention, Behavior Change Committee of the Council on Cardiometabolic Health, Council on CardiovascularStroke Nursing, Council on Functional GenomicsTranslational Biology, Council on Quality of CareOutcomes Research,Stroke Council. Current science on consumer use of mobile health for cardiovascular disease prevention: a scientific statement from the American heart association. Circulation 2015 Sep 22;132(12):1157-1213 [FREE Full text] [doi: 10.1161/CIR.0000000000000232] [Medline: 26271892]
25. Widmer RJ, Collins NM, Collins CS, West CP, Lerman LO, Lerman A. Digital health interventions for the prevention of cardiovascular disease: a systematic review and meta-analysis. Mayo Clin Proc 2015 Apr;90(4):469-480 [FREE Full text] [doi: 10.1016/j.mayocp.2014.12.026] [Medline: 25841251]
26. Tegegne TK, Shifti DM, Rawstorn JC, Jansons P, Zhang Y, Daryabeygikhotbehsara R, et al. Digital lifestyle interventions for cardiovascular risk reduction: a systematic review and network meta-analysis. Health Policy Technol 2024 Aug;13(3):100879. [doi: 10.1016/j.hlpt.2024.100879]
27. Hanlon P, Daines L, Campbell C, McKinstry B, Weller D, Pinnock H. Telehealth interventions to support self-management of long-term conditions: a systematic metareview of diabetes, heart failure, asthma, chronic obstructive pulmonary disease, and cancer. J Med Internet Res 2017 May 17;19(5):e172 [FREE Full text] [doi: 10.2196/jmir.6688] [Medline: 28526671]
28. Abbas SR, Seol H, Abbas Z, Lee SW. Exploring the role of artificial intelligence in smart healthcare: a capability and function-oriented review. Healthcare (Basel) 2025 Jul 08;13(14):1642 [FREE Full text] [doi: 10.3390/healthcare13141642] [Medline: 40724669]
29. Aharon KB, Gershfeld-Litvin A, Amir O, Nabutovsky I, Klempfner R. Improving cardiac rehabilitation patient adherence via personalized interventions. PLoS One 2022;17(8):e0273815 [FREE Full text] [doi: 10.1371/journal.pone.0273815] [Medline: 36037232]
30. Marketou ME, Anastasiou I, Fourlis A, Alevizaki A, Kochiadakis G. Efficacy of an internet of things-based system for cardiac rehabilitation monitoring: insights from the IntellIoT pivotal trial in heart failure patients. Eur Heart J Digit Health 2025 Mar;6(2):293-297. [doi: 10.1093/ehjdh/ztae093] [Medline: 40110225]
31. Fitriyani NL, Syafrudin M, Chamidah N, Rifada M, Susilo H, Aydin D, et al. A novel approach utilizing bagging, histogram gradient boosting, and advanced feature selection for predicting the onset of cardiovascular diseases. Mathematics 2025 Jul 04;13(13):2194 [FREE Full text] [doi: 10.3390/math13132194]
32. Lahtio H, Rintala A, Immonen J, Sjögren T. The effectiveness of physical activity-promoting web- and mobile-based distance weight loss interventions on body composition in rehabilitation settings: systematic review, meta-analysis, and meta-regression analysis. J Med Internet Res 2022 Mar 24;24(3):e25906 [FREE Full text] [doi: 10.2196/25906] [Medline: 35323126]

XSL•FO
RenderX

33. Lahtio H, Heinonen A, Paajanen T, Sjögren T. The added value of remote technology in cardiac rehabilitation on physical function, anthropometrics, and quality of life: cluster randomized controlled trial. J Med Internet Res 2023 Apr 12;25:e42455 [FREE Full text] [doi: 10.2196/42455] [Medline: 37043264]
34. Skamagki G, King A, Carpenter C, Wåhlin C. The concept of integration in mixed methods research: a step-by-step guide using an example study in physiotherapy. Physiother Theory Pract 2024 Feb;40(2):197-204 [FREE Full text] [doi: 10.1080/09593985.2022.2120375] [Medline: 36069530]
35. Sydänkuntoutuskurssi aikuiselle [cardiac rehabilitation course for adults]. Kela. 2021. URL: https://www.kela.fi/sydansairaudet-kuntoutuskurssit [accessed 2025-04-13]
36. Hakala S, Kivistö H, Paajanen T, Kankainen A, Anttila MR, Heinonen A, et al. Effectiveness of distance technology in promoting physical activity in cardiovascular disease rehabilitation: cluster randomized controlled trial, a pilot study. JMIR Rehabil Assist Technol 2021;8(2):e20299 [FREE Full text] [doi: 10.2196/20299] [Medline: 34142970]
37. Anttila MR, Söderlund A, Paajanen T, Kivistö H, Kokko K, Sjögren T. Biopsychosocial profiles of patients with cardiac disease in remote rehabilitation processes: mixed methods grounded theory approach. JMIR Rehabil Assist Technol 2021;8(4):e16864 [FREE Full text] [doi: 10.2196/16864] [Medline: 34730548]
38. Coronary heart disease. Käypä hoito. URL: https://www.kaypahoito.fi/hoi50130 [accessed 2022-03-23]
39. Holland AE, Spruit MA, Troosters T, Puhan MA, Pepin V, Saey D, et al. An official European respiratory society/American thoracic society technical standard: field walking tests in chronic respiratory disease. Eur Respir J 2014 Dec;44(6):1428-1446 [FREE Full text] [doi: 10.1183/09031936.00150314] [Medline: 25359355]
40. Giannitsi S, Bougiakli M, Bechlioulis A, Kotsia A, Michalis LK, Naka KK. 6-minute walking test: a useful tool in the management of heart failure patients. Ther Adv Cardiovasc Dis 2019;13:1753944719870084 [FREE Full text] [doi: 10.1177/1753944719870084] [Medline: 31441375]
41. Obesity and overweight. World Health Organization. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [accessed 2025-11-30]
42. Body mass index and waist circumference. Finnish Medical Society Duodecim. URL: https://www.kaypahoito.fi/nix00163 [accessed 2025-11-30]
43. WHOQOL: measuring quality of life. World Health Organization. URL: https://www.who.int/tools/whoqol [accessed 2025-11-30]
44. Correia de Faria Santarém G, de Cleva R, Santo MA, Bernhard AB, Gadducci AV, Greve JMD, et al. Correlation between body composition and walking capacity in severe obesity. PLoS One 2015;10(6):e0130268 [FREE Full text] [doi: 10.1371/journal.pone.0130268] [Medline: 26098769]
45. Beriault K, Carpentier AC, Gagnon C, Ménard J, Baillargeon JP, Ardilouze JL, et al. Reproducibility of the 6-minute walk test in obese adults. Int J Sports Med 2009 Oct;30(10):725-727. [doi: 10.1055/s-0029-1231043] [Medline: 19585400]
46. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2008 Jul 21;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
47. Kiger ME, Varpio L. Thematic analysis of qualitative data: AMEE guide no. 131. Med Teach 2020 Aug;42(8):846-854. [doi: 10.1080/0142159X.2020.1755030] [Medline: 32356468]
48. Iwayama Y, Shimba Y, Viswanathan CS, Yano Y. Effect and factors associated with weight and waist circumference reductions in information and communication technology-based specific health guidance. Environ Occup Health Pract 2023;5(1):0001-OA. [doi: 10.1539/eohp.2023-0001-OA] [Medline: 40059931]
49. Al-Abdullah L, Ahern A, Welsh P, Logue J. A predictive model for medium-term weight loss response in people with type 2 diabetes engaging in behavioural weight management interventions. Diabetes Obes Metab 2024 Sep;26(9):3653-3662. [doi: 10.1111/dom.15706] [Medline: 38874091]
50. Höchsmann C, Martin CK, Apolzan JW, Dorling JL, Newton RL, Denstel KD, PROPEL Research Group. Initial weight loss and early intervention adherence predict long-term weight loss during the promoting successful weight loss in primary care in Louisiana lifestyle intervention. Obesity (Silver Spring) 2023 Sep;31(9):2272-2282. [doi: 10.1002/oby.23854] [Medline: 37551762]
51. Farage G, Simmons C, Kocak M, Klesges RC, Talcott GW, Richey P, et al. Assessing the contribution of self-monitoring through a commercial weight loss app: mediation and predictive modeling study. JMIR Mhealth Uhealth 2021 Jul 14;9(7):e18741 [FREE Full text] [doi: 10.2196/18741] [Medline: 34259635]
52. Ferreira MB, Saraiva FA, Fonseca T, Costa R, Marinho A, Oliveira JC, et al. Clinical associations and prognostic implications of 6-minute walk test in rheumatoid arthritis. Sci Rep 2022 Nov 04;12(1):18672 [FREE Full text] [doi: 10.1038/s41598-022-21547-z] [Medline: 36333405]
53. Wang E, Abrahamson K, Liu PJ, Ahmed A. Can mobile technology improve weight loss in overweight adults? a systematic review. West J Nurs Res 2020 Sep;42(9):747-759. [doi: 10.1177/0193945919888224] [Medline: 31762402]
54. Lenasi H, Novak A, Jug B, Dervišević E, Karpljuk D, Videmšek M, et al. An alternative prediction equation for evaluation of six-minute walk distance in stable coronary artery disease patients. Front Physiol 2022;13:844847 [FREE Full text] [doi: 10.3389/fphys.2022.844847] [Medline: 35432004]
55. Pepera G, Ingle L, Sandercock GRH. Predictors of the 6-minute walk test in patients with chronic heart failure. Br J Card Nurs 2015 Sep 02;10(9):454-549. [doi: 10.12968/bjca.2015.10.9.454]

XSL•FO
RenderX

56. Sharma P, Sharma S, Chaudhary V. Relations of six minute walk distance with body mass index and wait hip ratio. Int J Sci Res 2018;7(1). [doi: 10.36106/ijsr]
57. Whiteley JA, Tjaden AH, Bailey CP, Faro J, DiPietro L, Hayman LL, et al. Engagement with digital weight loss intervention components and weight outcomes. J Technol Behav Sci 2024 Dec;9(4):714-723. [doi: 10.1007/s41347-024-00388-9] [Medline: 41122737]
58. Liljeroos M, Arkkukangas M, Strömberg A. The long-term effect of an m-health tool on self-care in patients with heart failure: a pre-post interventional study with a mixed-method analysis. Eur J Cardiovasc Nurs 2024;23(5):470-477 [FREE Full text] [doi: 10.1093/eurjcn/zvad107] [Medline: 38165027]
59. Beleigoli AM, Andrade AQ, Cançado AG, Paulo MN, Diniz MDFH, Ribeiro AL. Web-based digital health interventions for weight loss and lifestyle habit changes in overweight and obese adults: systematic review and meta-analysis. J Med Internet Res 2019;21(1):e298 [FREE Full text] [doi: 10.2196/jmir.9609] [Medline: 30622090]
60. Norcross JC, Krebs PM, Prochaska JO. Stages of change. J Clin Psychol 2011;67(2):143-154. [doi: 10.1002/jclp.20758] [Medline: 21157930]
61. Prochaska JO, Velicer WF. The transtheoretical model of health behavior change. Am J Health Promot 1997;12(1):38-48. [doi: 10.4278/0890-1171-12.1.38] [Medline: 10170434]
62. Lambert JD, Dean SG, Terry RH, Charles N, Greaves C, Campbell JL, et al. Mechanisms of impact of web-based support and self-monitoring to augment and maintain physical activity levels: a qualitative study exploring participants' interactions with the e-coachER, a web-based support programme for people attending exercise referral schemes. BMJ Open 2024;14(10):e080472 [FREE Full text] [doi: 10.1136/bmjopen-2023-080472] [Medline: 39477258]
63. Følling IS, Oldervoll LM, Hilmarsen C, Ersfjord EMI. Exploring use of activity monitors for patients with obesity during weight-loss treatment - a qualitative study. BMC Sports Sci Med Rehabil 2021;13(1):25 [FREE Full text] [doi: 10.1186/s13102-021-00253-9] [Medline: 33731190]
64. Tay A, Hoeksema H, Murphy R. Uncovering barriers and facilitators of weight loss and weight loss maintenance: insights from qualitative research. Nutrients 2023;15(5):1297 [FREE Full text] [doi: 10.3390/nu15051297] [Medline: 36904294]
65. Raaijmakers LCH, Pouwels S, Berghuis KA, Nienhuijs SW. Technology-based interventions in the treatment of overweight and obesity: a systematic review. Appetite 2015;95:138-151. [doi: 10.1016/j.appet.2015.07.008] [Medline: 26165415]
66. Jøranli KT, Vefring LT, Dalen M, Garnweidner-Holme L, Molin M. Experiences of social support by participants with morbid obesity who participate in a rehabilitation program for health-behavior change: a qualitative study. BMC Nutr 2023;9(1):149 [FREE Full text] [doi: 10.1186/s40795-023-00810-0] [Medline: 38098104]
67. Morgan PJ, Scott HA, Young MD, Plotnikoff RC, Collins CE, Callister R. Associations between program outcomes and adherence to social cognitive theory tasks: process evaluation of the SHED-IT community weight loss trial for men. Int J Behav Nutr Phys Act 2014;11:89 [FREE Full text] [doi: 10.1186/s12966-014-0089-9] [Medline: 25011421]
68. Spreckley M, Seidell J, Halberstadt J. Perspectives into the experience of successful, substantial long-term weight-loss maintenance: a systematic review. Int J Qual Stud Health Well-being 2021 Dec;16(1):1862481 [FREE Full text] [doi: 10.1080/17482631.2020.1862481] [Medline: 33455563]
69. Liu K, Choi TST, Zhao L, Teong XT, Hutchison AT, Heilbronn LK. A qualitative exploration of behaviour change and maintenance experience in people with overweight or obesity in a dietary intervention. Nutr Diet 2024 Jun;81(3):296-305. [doi: 10.1111/1747-0080.12855] [Medline: 38057992]
70. Gillison FB, Rouse P, Standage M, Sebire SJ, Ryan RM. A meta-analysis of techniques to promote motivation for health behaviour change from a self-determination theory perspective. Health Psychol Rev 2019 Mar;13(1):110-130. [doi: 10.1080/17437199.2018.1534071] [Medline: 30295176]
71. Santos I, Silva MN, Teixeira PJ. A self-determination theory perspective on weight loss maintenance. EHP 2016;18(5) [FREE Full text]
72. Dandeneau DA, Leon SJ, Cameranesi M, Alexiuk MR, Protudjer JLP, Bohm CJ, et al. Evaluation of an online cardiometabolic and weight loss program: a mixed methods study. Appl Physiol Nutr Metab 2025 Jan 01;50:1-11. [doi: 10.1139/apnm-2024-0190] [Medline: 39854690]
73. Leung AWY, Chan RSM, Sea MMM, Woo J. Psychological factors of long-term dietary and physical activity adherence among Chinese adults with overweight and obesity in a community-based lifestyle modification program: a mixed-method study. Nutrients 2020;12(5):1379. [doi: 10.3390/nu12051379] [Medline: 32408509]
74. Puhl RM, Heuer CA. Obesity stigma: important considerations for public health. Am J Public Health 2010;100(6):1019-1028. [doi: 10.2105/AJPH.2009.159491] [Medline: 20075322]
75. Amy NK, Aalborg A, Lyons P, Keranen L. Barriers to routine gynecological cancer screening for White and African-American obese women. Int J Obes (Lond) 2006 Jan;30(1):147-155. [doi: 10.1038/sj.ijo.0803105] [Medline: 16231037]
76. Phelan SM, Burgess DJ, Yeazel MW, Hellerstedt WL, Griffin JM, van Ryn M. Impact of weight bias and stigma on quality of care and outcomes for patients with obesity. Obes Rev 2015 Apr;16(4):319-326 [FREE Full text] [doi: 10.1111/obr.12266] [Medline: 25752756]
77. Hussein N. Weight management in cardiac rehabilitation: evaluation, assessment, and interventions. J Clin Epigen 2022;7(9) [FREE Full text]

XSL•FO
RenderX

78. Imran TF, Wang N, Zombeck S, Balady GJ. Mobile technology improves adherence to cardiac rehabilitation: a propensity score-matched study. J Am Heart Assoc 2021 Aug 03;10(15):e020482 [FREE Full text] [doi: 10.1161/JAHA.120.020482] [Medline: 34278801]
79. Singh P, Kourav PS, Mohapatra S, Kumar V, Panda SK. Human heart health prediction using GAIT parameters and machine learning model. Biomed Signal Process Control 2024 Feb;88:105696 [FREE Full text] [doi: 10.1016/j.bspc.2023.105696]
80. Pujia C, Ferro Y, Mazza E, Maurotti S, Montalcini T, Pujia A. The role of mobile apps in obesity management: systematic review and meta-analysis. J Med Internet Res 2025 May 06;27:e66887 [FREE Full text] [doi: 10.2196/66887] [Medline: 40327853]
81. Faizan M, Han C, Lee SW. Policy-driven digital health interventions for health promotion and disease prevention: a systematic review of clinical and environmental outcomes. Healthcare (Basel) 2025 Sep 16;13(18):2319 [FREE Full text] [doi: 10.3390/healthcare13182319] [Medline: 41008449]
82. Pfaeffli Dale L, Whittaker R, Jiang Y, Stewart R, Rolleston A, Maddison R. Text message and internet support for coronary heart disease self-management: results from the Text4Heart randomized controlled trial. J Med Internet Res 2015 Oct 21;17(10):e237 [FREE Full text] [doi: 10.2196/jmir.4944] [Medline: 26490012]
83. Raja MA, Loughran R, Caffery FM. A review of applications of artificial intelligence in cardiorespiratory rehabilitation. Inform Med Unlocked 2023;41:101327 [FREE Full text] [doi: 10.1016/j.imu.2023.101327]
84. Marvasti TB, Gao Y, Murray KR, Hershman S, McIntosh C, Moayedi Y. Unlocking tomorrow's health care: expanding the clinical scope of wearables by applying artificial intelligence. Can J Cardiol 2024 Oct;40(10):1934-1945 [FREE Full text] [doi: 10.1016/j.cjca.2024.07.009] [Medline: 39025363]
85. Ahmad OM, Ibrahim R, Odunsi DI, Mohammed M, Mathew BG, Touny M, et al. Role of mobile health and wearable devices in cardiovascular disease prevention: a comprehensive review. Cureus 2025 May;17(5):e83673. [doi: 10.7759/cureus.83673] [Medline: 40486456]
86. Su J, Zhang Y, Ke QQ, Su JK, Yang QH. Mobilizing artificial intelligence to cardiac telerehabilitation. Rev Cardiovasc Med 2022;23(2):45 [FREE Full text] [doi: 10.31083/j.rcm2302045] [Medline: 35229536]
87. Narkhede M, Pardeshi A, Bhagat R, Dharme G. Review on emerging therapeutic strategies for managing cardiovascular disease. Curr Cardiol Rev 2024;20(4):e160424228949. [doi: 10.2174/011573403X299265240405080030] [Medline: 38629366]
88. Laine S, Heilala V, Säteri S, Sjögren T. Tekoälystä hyötyvät niin sydänkuntoutujat kuin fysioterapeutit [AI brings advantages to both cardiac rehab patients and physiotherapists]. Fysioterapia [Physiotherapy] 2025;72(6):36-41.
89. Abbas Q, Jeong W, Lee SW. Explainable AI in clinical decision support systems: a meta-analysis of methods, applications, and usability challenges. Healthcare (Basel) 2025 Aug 29;13(17):2154 [FREE Full text] [doi: 10.3390/healthcare13172154] [Medline: 40941506]
90. Taylor RS, Fredericks S, Jones I, Neubeck L, Sanders J, De Stoutz N, et al. Global perspectives on heart disease rehabilitation and secondary prevention: a scientific statement from the association of cardiovascular nursing and allied professions, European association of preventive cardiology, and international council of cardiovascular prevention and rehabilitation. Eur Heart J 2023;44(28):2515-2525 [FREE Full text] [doi: 10.1093/eurheartj/ehad225] [Medline: 37477626]
91. Varnfield M, Karunanithi M, Lee CK, Honeyman E, Arnold D, Ding H, et al. Smartphone-based home care model improved use of cardiac rehabilitation in postmyocardial infarction patients: results from a randomised controlled trial. Heart 2014 Nov;100(22):1770-1779 [FREE Full text] [doi: 10.1136/heartjnl-2014-305783] [Medline: 24973083]
92. Dorje T, Zhao G, Tso K, Wang J, Chen Y, Tsokey L, et al. Smartphone and social media-based cardiac rehabilitation and secondary prevention in China (SMART-CR/SP): a parallel-group, single-blind, randomised controlled trial. Lancet Digit Health 2019 Nov;1(7):e363-e374 [FREE Full text] [doi: 10.1016/S2589-7500(19)30151-7] [Medline: 33323210]
93. Huang K, Liu W, He D, Huang B, Xiao D, Peng Y, et al. Telehealth interventions versus center-based cardiac rehabilitation of coronary artery disease: a systematic review and meta-analysis. Eur J Prev Cardiol 2015 Aug;22(8):959-971. [doi: 10.1177/2047487314561168] [Medline: 25488550]
94. Kim H, Kim SE, Sung MK. Sex and gender differences in obesity: biological, sociocultural, and clinical perspectives. World J Mens Health 2025 Oct;43(4):758-772 [FREE Full text] [doi: 10.5534/wjmh.250126] [Medline: 40676890]
95. Casano HAM, Ahmed I, Anjum F. Six-minute walk test. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2023.
96. Kalfoss MH, Reidunsdatter RJ, Klöckner CA, Nilsen M. Validation of the WHOQOL-Bref: psychometric properties and normative data for the Norwegian general population. Health Qual Life Outcomes 2021 Jan 07;19(1):13 [FREE Full text] [doi: 10.1186/s12955-020-01656-x] [Medline: 33413455]
97. Archuleta AJ, Prost SG, Golder S. Psychometric properties of the WHOQOL-BREF in a population of older adults in prison: measuring quality of life in a carceral setting. Int J Prison Health 2023 Jul 25;ahead-of-print(ahead-of-print):668-681. [doi: 10.1108/IJPH-07-2022-0042] [Medline: 37480217]
98. Rosén H, Ahlström G, Lexén A. Psychometric properties of the WHOQOL-BREF among next of kin to older persons in nursing homes. Health Qual Life Outcomes 2020 Apr 19;18(1):103 [FREE Full text] [doi: 10.1186/s12955-020-01345-9] [Medline: 32306965]

XSL•FO
RenderX

Abbreviations

6MWT: 6-minute walk test
AI: artificial intelligence
CRT: cluster randomized trial
CVD: cardiovascular disease
PA: physical activity
WC: waist circumference
WHOQOL-BREF: World Health Organization Quality of Life Brief Form

Edited by A Stone; submitted 01.Jun.2025; peer-reviewed by SW Lee; comments to author 24.Nov.2025; accepted 07.Jan.2026; published 02.Mar.2026.

Please cite as:
Lahtio H, Paajanen T, Korpi H, Heinonen A, Sjögren T
Added Technology for Weight Management in Cardiac Rehabilitation: Biopsychosocial Factors and Rehabilitees' Talk in a Mixed Method Study
J Med Internet Res 2026;28:e78347
URL: https://www.jmir.org/2026/1/e78347
doi:10.2196/78347
PMID:

©Heli Lahtio, Teemu Paajanen, Hilkka Korpi, Ari Heinonen, Tuulikki Sjögren. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of the Components of a Digital Multiple Health Behavior Intervention Among University Students (Buddy): Factorial Randomized Trial

Katarina Åsberg[1], PhD; Oskar Lundgren[1,2], PhD; Hanna Henriksson[1,3], PhD; Pontus Henriksson[1], PhD; Ann Catrine Eldh[1,4], PhD; Preben Bendtsen[1,5], PhD; Marie Löf[6], PhD; Marcus Bendtsen[1], PhD

[1]Department of Health, Medicine and Caring Sciences, Linköping University, Linköping, Östergötland, Sweden

[2]Crown Princess Victoria Children's Hospital, Linköping University Hospital, Linköping, Sweden

[3]Unit for Strategic Healthcare, Region Östergötland, Linköping, Sweden

[4]Department of Public Health and Caring Sciences, Uppsala University, Uppsala, Sweden

[5]Medical Specialist Clinic, Motala Hospital, Motala, Sweden

[6]Department of Medicine Huddinge, Karolinska Institutet, Stockholm, Sweden

Corresponding Author:
Marcus Bendtsen, PhD
Department of Health, Medicine and Caring Sciences
Linköping University
Linköping, Östergötland, 581 83
Sweden
Phone: 46 13281000
Email: marcus.bendtsen@liu.se

Abstract

Background: Digital interventions have shown promise in supporting healthy behaviors among university students; however, few interventions support simultaneous change across multiple health behaviors. Moreover, behavioral interventions are typically evaluated as a whole, making it challenging to disentangle the contribution of individual components to the overall effects.

Objective: This study estimated the effects of the components of a digital behavior intervention on alcohol, diet, physical activity, and smoking outcomes among university students in Sweden.

Methods: A double-blind randomized factorial trial with 6 two-level factors was conducted. University students in Sweden were proactively recruited through student health care centers and social media. Participants were eligible if they were aged 18 years or older and had at least one health behavior classified as unhealthy. The effects of 6 components were estimated: screening and feedback; goal-setting and planning; motivation; skills and know-how; mindfulness; and self-authored SMS text messages. Primary outcomes were weekly alcohol consumption and frequency of heavy episodic drinking, average daily fruit and vegetable consumption, weekly sugary drink consumption, weekly moderate-to-vigorous physical activity (MVPA), and 4-week point prevalence of smoking.

Results: A total of 1704 students were randomized. The effectiveness of individual and pairwise components was estimated using available data from 1118 (65.61%) participants at 2 months and 874 (51.29%) at 4 months, with sensitivity analyses conducted using imputed missing data. Most consistently, the evidence indicated that screening and feedback affected fruit and vegetable consumption (2-month mean difference 0.11, compatibility interval [CoI] –0.02 to 0.24; probability of effect [POE] 94.7% and 4-month mean difference 0.12, CoI –0.03 to 0.26; POE 94.4%), as did skills and know-how (2-month mean difference 0.19, CoI 0.06-0.33; POE 99.8% and 4-month mean difference 0.14, CoI 0.01-0.28; POE 96.9%). The combination of these 2 components was even more effective (2-month mean difference 0.30, CoI 0.11-0.48; POE 99.9% and 4-month mean difference 0.26, CoI 0.05-0.46; POE 99.4%). The motivation and mindfulness components, both individually and in combination, increased MVPA at 2 months (combined mean difference 78.0, CoI 28.3-128.2; POE 99.9%); however, this effect was not observed at 4 months. Combining screening and feedback with skills and know-how increased MVPA at 4 months (mean difference 60.1, CoI 3.6-116.5; POE 98.2%). Heavy episodic drinking was reduced at 2 months by screening and feedback (incidence rate ratio 0.87, CoI 0.74-1.02; POE 95.2%), and the effect was greater when combined with goal-setting and mindfulness. There was some evidence that the motivation component was harmful with respect to heavy episodic drinking and that self-authored SMS text messages were harmful with respect to sugary drink consumption.

XSL•FO
RenderX

Conclusions: We dismantled a complex digital multiple behavior intervention and examined it using a factorial design to provide novel insights into the effectiveness of the intervention's different components. Both marginal and synergistic effects were observed across multiple behaviors, providing evidence regarding which components are most promising in complex interventions. These findings should be considered in light of the risk of bias introduced by attrition to follow-up, which was high in this effectiveness trial with low barriers to participation.

Trial Registration: International Standard Randomised Controlled Trial Number (ISRCTN) ISRCTN23310640; https://www.isrctn.com/ISRCTN23310640

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2021-051044

(J Med Internet Res 2026;28:e88884) doi:10.2196/88884

KEYWORDS

behavior change; digital interventions; multiple lifestyle; public health; mHealth intervention

Introduction

Unhealthy behaviors, including alcohol consumption, unhealthy diets, physical inactivity, and smoking, are leading causes of noncommunicable diseases, including cardiovascular diseases, cancers, and diabetes [1-5]. The young adult years are a critical period for the establishment of healthy behaviors [6-8], which may reduce future morbidity [9-14]. Engaging in healthy behaviors also provides several immediate positive effects on both physical and mental health [15-20]. These include fewer mental health problems, such as anxiety and depression [17-20]; improved cognitive function and academic achievement [21,22]; and various aspects of well-being [15,16]. Additionally, reduced alcohol consumption lowers the risk of alcohol-related harms or negative consequences, such as social problems, injuries, or risky sexual behaviors [23,24]. However, a study of university students in 24 countries found a high prevalence of unhealthy behaviors [25]. In addition, unhealthy behaviors tend to co-occur among students; for instance, low fruit and vegetable consumption and low levels of physical activity often occur together. This is concerning because the risks of negative consequences associated with multiple unhealthy behaviors compound [26].

In the digital era, new avenues to support individuals in improving their health behaviors are emerging. One form of support is individual-level digital behavior change interventions, which deliver support materials via personal web portals, SMS text messages, smartphone apps, etc. The effectiveness of digital interventions across various health behaviors has been demonstrated in a wide range of populations and settings [27-33]. As young adults frequently use technology to access health information [34], digital interventions have the potential to efficiently reach this target group, and numerous single–health behavior digital interventions have been evaluated among university and college students [35-38]. In Sweden, digital interventions among university students have shown promising results, particularly with respect to alcohol [39-42], smoking cessation [43], and mental health promotion [44]. However, only a few interventions that support changes in multiple health behaviors simultaneously have been studied [45-47], and even fewer are tailored specifically for students [48]. Therefore, while single-behavior digital interventions show promise, our understanding of the effects of digital interventions targeting multiple health behaviors remains limited, despite the prevalence of co-occurring unhealthy behaviors [49,50].

Most digital behavior change interventions are grounded in behavior change theory and utilize behavior change techniques (BCTs) [51]. Systematic reviews [52,53] support the use of theory and BCTs in developing effective digital interventions. However, findings regarding the effectiveness of BCTs have been heterogeneous [54], indicating that the ways in which BCTs are implemented may vary considerably across interventions. In addition, behavioral interventions are typically evaluated using designs that estimate the effects of interventions as a whole [55], rather than at the component level, making it challenging to disentangle the contribution of specific components and BCTs to the overall effects.

In response to the lack of evidence on the effectiveness of digital interventions targeting multiple health behaviors, and the limited evidence on the effectiveness of different components of digital interventions, we conducted a factorial trial among university and college students in Sweden. The trial aimed to estimate the effectiveness of 6 intervention components of a digital intervention in promoting behavior change related to alcohol consumption, diet, physical activity, and smoking.

Methods

Study Design

We conducted a factorial randomized trial of a digital multiple health behavior intervention among university and college students in Sweden. The factorial trial was designed with 6 two-level factors representing the presence or absence of 6 intervention components. The treatment effect being estimated was thus the effect of the presence versus absence of each specific component, adjusted for the presence and absence of other components, as well as pairwise interactions between components. The trial was part of the MoBILE research program, which includes studies of digital health behavior interventions across the lifespan [56]. A protocol for the trial was published prospectively [57], and the trial was preregistered in the International Standard Randomised Controlled Trial Number (ISRCTN) registry on January 28, 2021 (ISRCTN23310640). This report follows CONSORT (Consolidated Standards of Reporting Trials) guidelines (see Multimedia Appendix 1) [58], including extensions of CONSORT for factorial trials [59].

XSL•FO
RenderX

Students were included in the trial if they met at least one of six inclusion criteria based on unhealthy behaviors as defined by national guidelines. We made an adjustment to the inclusion criteria for sugary drinks after trial registration. Initially, we decided that participants who consumed 3 or more units (33 cl) of sugary drinks per week would be included in follow-up analyses of sugary drink consumption. However, this was changed to 2-3 or more units per week because the questionnaire used did not differentiate between those consuming 2 or more units per week and those consuming 3 or more units per week. No other changes were made to the trial protocol.

Study Setting, Recruitment, and Participants

Students at 18 geographically dispersed universities and colleges across Sweden (out of 50) were invited to participate in the trial. These institutions encompassed a variety of urban and rural settings and offered diverse educational profiles. Prospective participants were recruited using paper advertising (posters and leaflets), digital advertising (email, websites, and social media), and through student health care staff. Students registered their interest by sending an SMS text message to a dedicated phone number, after which they received an immediate reply with a hyperlink to a web page containing informed consent materials (see Multimedia Appendix 2). Consenting students were then asked to complete a baseline questionnaire (see Multimedia Appendix 3), the responses to which were used to assess eligibility. Included students were required to be 18 years or older and to fulfill at least one of the criteria presented in Textbox 1, which were based on Swedish national guidelines for healthy behaviors at the time the trial commenced.

Trial information and intervention materials were in Swedish and delivered to participants' mobile phones; thus, individuals who did not understand Swedish well enough to enroll or who did not have access to a mobile phone were not able to take part in the study. Notably, it has been estimated that 96% of 16-35-year-olds in Sweden own a smartphone, and this percentage is likely to be even higher among university and college students.

Textbox 1. Criteria for inclusion.

1. Weekly alcohol consumption: Consumed 10/15 (women/men) or more standard drinks of alcohol in the past week. A standard drink was defined as 12 g of pure alcohol.

2. Heavy episodic drinking: Consumed 4/5 (women/men) or more standard drinks of alcohol on a single occasion at least once in the past month.

3. Fruit and vegetables: Consumed less than 500 g of fruit and vegetables on average per day in the past week.

4. Sugary drinks: Consumed 2 or more units of sugary drinks in the past week. One sugary drink unit was defined as approximately 33 cl.

5. Moderate-to-vigorous physical activity: Spent less than 150 minutes on moderate-to-vigorous physical activity in the past week.

6. Smoking: Smoked at least one cigarette in the past week.

Interventions

The digital multiple health behavior intervention studied in this trial was a mobile phone–based intervention targeting alcohol, diet, physical activity, and smoking. The intervention, called Buddy, consisted of 6 components aimed at promoting behavior change: screening and feedback (C1); goal-setting and planning (C2); motivation (C3); skills and know-how (C4); mindfulness (C5); and self-authored SMS text messages (C6). The components were informed by social-cognitive models of health [60] and have been included in interventions that have proven successful in facilitating behavior change [61,62]. Table 1 provides a brief description of each of the 6 components, utilizing the BCTTv1 93-item taxonomy [63] to specify BCTs. Please see Multimedia Appendix 4 for full details of the intervention.

XSL•FO
RenderX

Table 1. Description of intervention components, conceptual framework, practical application, and factorial conditions.

Component	Conceptual framework	Practical application	Factorial conditions
Component 1: Screening and feedback	*Self-monitoring* has been shown to be a potentially effective strategy for reducing excessive alcohol consumption [39,64-66] and promoting healthy eating and physical activity [67,68]. This concerned the following BCT[a] components: Discrepancy between current behavior and goal(s) (BCT 1.6), Feedback on behavior (BCT 2.2), Self-monitoring of behavior (BCT 2.3), and Social comparison (BCT 6.2).	Every Sunday afternoon, participants received an SMS text message containing a hyperlink to a questionnaire regarding their current health behaviors. After completing the questionnaire, participants received feedback on their current behaviors in comparison with national guidelines. Thereafter, users were given access to the remaining components (depending on allocation).	When this component was absent, participants were not asked to complete the screening questionnaire but were instead shown the national guidelines without receiving personal feedback.
Component 2: Goal-setting and planning	*Self-regulatory* skills and capacity were addressed via goal-setting and planning. Planning-related activities, such as goal-setting, action planning, and practicing behavior, have been shown to be effective in lifestyle behavior interventions [67,69-74]. This concerned the following BCT components: Goal-setting (BCT 1.1), Problem solving (BCT 1.2), Action planning (BCT 1.4), Prompts/cues (BCT 7.1), and Behavior practice/rehearsal (BCT 8.1).	This component allowed participants to set 1 or more goals for their future behavior. It included action planning for how they would progress toward their goals, preparation for motivational struggles (coping planning), and strategies for rewarding themselves upon success. Participants could also create their own challenges or accept ready-made challenges for the coming week, such as walking for 15 minutes each day or not drinking any alcohol during the coming week. Reminders about their goals and challenges were sent to participants via SMS text messages throughout the week (up to 4 SMS text messages).	When absent, this component was not visible, and goal-setting reminders were not available.
Component 3: Motivation	Digital behavior change interventions have been shown to enhance *self-efficacy*; however, there is a lack of consensus across reviews regarding which content works best to facilitate increases in self-efficacy [75]. This concerned the following BCT components: Information about health consequences (BCT 5.1), Credible source (BCT 9.1), Pros and cons (BCT 9.2), and Comparative imagining of future outcomes (BCT 9.3).	This component contained information and tools to increase participants' awareness of their own motivation, encourage commitment, and boost self-efficacy. It included information on negative health consequences, costs induced by certain behaviors, and reflective tasks delivered via SMS text messages, videos, and exercises. If participants chose, they could also activate motivational SMS text messages derived from previously developed and evaluated interventions [40,43,76-80], which were sent to them throughout the week (a maximum of 8-10 SMS text messages per week).	When absent, this component was not visible, and motivational SMS text messages were not available.
Component 4: Skills and know-how	Intervention content based on shaping knowledge, aimed at increasing an individual's understanding and awareness to facilitate behavior change, has been shown to be effective [70,71,81,82]. This concerned the following BCT components: Instructions on how to perform a behavior (BCT 4.1), Behavior substitution (BCT 8.2), Habit formation (BCT 8.3), and Graded tasks (BCT 8.7).	This component aimed to increase participants' skills and know-how by providing concrete tips on how to initiate and maintain lasting changes in everyday life through repetition and substitution. Participants were given strategies, for example, how to say no to alcoholic beverages or how to introduce vegetables into their meals. If participants chose to do so, they could also activate SMS text messages with tips and know-how, which were sent to them throughout the week (a maximum of 8-10 SMS text messages per week).	When absent, this component was not visible, and SMS text messages with tips and know-how were not available.

XSL•FO
RenderX

Component	Conceptual framework	Practical application	Factorial conditions
Component 5: Mindfulness	The mindfulness exercises were based on previous research and were considered evidence-based methods to improve the mental well-being of clinical populations, whereas their effects on behavior change in nonclinical settings are less well studied [83-87].	This component aimed to increase users' awareness of their own lived experience and strengthen their capacity for a nonreactive, compassionate, and less stressful way of being in the world. Mindfulness exercises, including guided meditations, were offered to participants.	When absent, this component was not visible, and guided meditations were not available.
Component 6: Self-authored SMS text messages	SMS text message self-authorship is generally understudied in the literature, but was included in an effective digital alcohol intervention [88].	This component allowed participants to self-author up to 3 SMS text messages and schedule them to be sent to themselves throughout the week at times of their choosing. For example, a participant could write an SMS text message reminding themselves to eat 2 fruits each day, avoid drinking on Wednesdays, or go for a walk with a friend.	When absent, this component was not visible, and the ability to self-author SMS text messages was not available.

[a]BCT: behavior change technique.

As shown in Figure 1, the components were presented to participants in a menu, allowing access in any order and with any frequency. Participants could therefore navigate between components in ways typical of smartphone apps. Regardless of the inclusion criterion met, all participants received intervention materials regarding alcohol, diet, physical activity, and smoking behaviors. There was no difference in the materials provided based on risk profile, except for the feedback received after screening (C1), where reported behavior was visually matched against national guidelines. The components presented materials that were both behavior-specific and more general to behavior change. Screening and feedback (C1) was conducted for each behavior, with feedback provided separately (eg, showing how current heavy drinking patterns match recommendations). Screening and feedback (C1) was offered to participants each week, allowing them to receive feedback on their behavior multiple times throughout the intervention period. Goal-setting and planning (C2) guided participants to set goals without prompting any specific behavior and also allowed them to create personal challenges; in addition, preprepared challenges were available and categorized by behavior. Motivation (C3) included materials that were not behavior specific and aimed to increase motivation and commitment to behavior change. The C3 component also allowed participants to sign up to receive behavior-specific SMS text messages throughout the week to boost motivation and commitment (participants could sign up for multiple behaviors). Skills and know-how (C4) contained content and optional SMS text messages designed to increase participants' skills and knowledge about how to change, and was provided separately for each behavior, with loose references between behaviors (eg, recommending going for walks when feeling the urge to smoke). Mindfulness (C5) did not mention specific behaviors but instead aimed to strengthen the mental resources needed to change behavior. Self-authored SMS text messages (C6) provided a means for participants to send reminders to themselves throughout the week, with no content provided for any specific behavior.

XSL•FO
RenderX

Figure 1. Screenshot of the Buddy intervention displaying the components in order from top to bottom and left to right: (1) Screening and feedback, (2) Goal-setting and planning, (3) Motivation, (4) Skills and know-how, (5) Mindfulness, and (6) Self-authored SMS text messages.

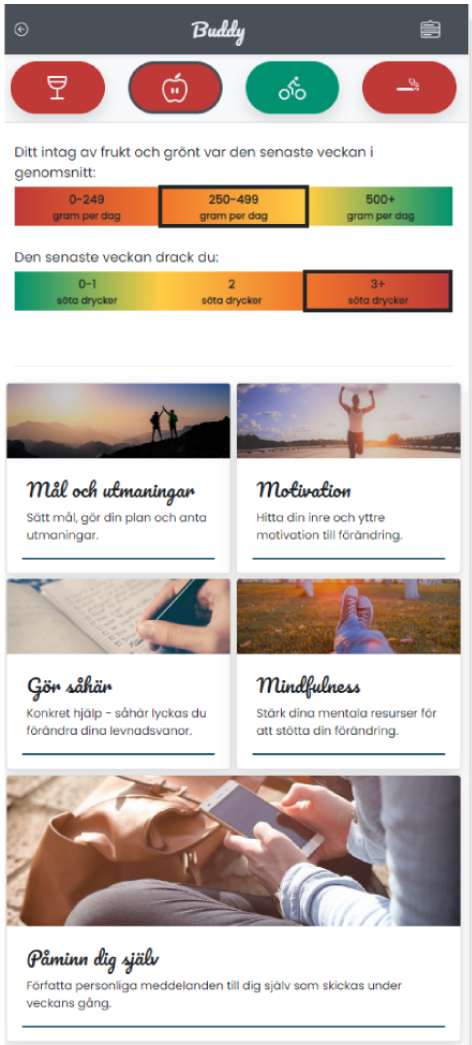

Participants were randomly assigned to 1 of 64 conditions, each representing a unique combination of Buddy's 6 components. The components assigned to individuals, based on their factorial allocation, were available for use at their discretion over a 4-month period, with weekly SMS text message reminders sent on Sunday afternoons to encourage access to the intervention.

Outcomes and Measures

Outcomes are listed in Textbox 2 and described below. All questionnaires used in the trial can be found in Multimedia Appendix 3.

Textbox 2. Primary and secondary outcomes.

1. Primary outcomes

- Alcohol: weekly alcohol consumption; monthly frequency of heavy episodic drinking.
- Diet: average daily consumption of fruit and vegetables; weekly consumption of sugary drinks.
- Physical activity: weekly moderate-to-vigorous physical activity.
- Smoking: 4-week point prevalence of smoking abstinence.

2. Secondary outcomes

- Weekly consumption of candy and snacks.
- BMI.
- Weekly number of cigarettes smoked.
- Perceived stress.

XSL•FO
RenderX

Weekly alcohol consumption was assessed by asking participants to report the number of standard drinks of alcohol they consumed in the past week (short-term recall method [89]). The frequency of heavy episodic drinking was assessed by asking participants how many times they consumed more than 4/5 (women/men) standard drinks of alcohol on 1 occasion in the past month. These 2 outcomes are both part of the core outcome set for brief alcohol interventions [90,91].

Diet and physical activity were assessed using a questionnaire based on a previously published questionnaire by the National Board of Health and Welfare in Sweden [92], which was further modified to include portion sizes. Fruit and vegetable consumption was assessed using 2 questions on the number of portions (100 g) of fruit and vegetables, respectively, that participants consumed on average per day during the past week. Sugary drink consumption was assessed using a question on the number of units (33 cl) of sugary drinks participants consumed in the past week. Moderate-to-vigorous physical activity (MVPA) was assessed by summing responses to 2 questions on the number of minutes spent on MVPA, respectively, during the past week.

BMI was calculated using participants' self-reported weight (without further instructions). Baseline height was used, as it was deemed unlikely to have changed significantly.

Four-week point prevalence of smoking abstinence (no cigarettes in the past 4 weeks) was assessed as a binary question, a measure suggested by the Society for Research on Nicotine and Tobacco [93]. Participants who reported having smoked any cigarettes in the past 4 weeks were also asked to report the number of cigarettes smoked in the past week.

Perceived stress was assessed using the Short-Form Perceived Stress Scale [94].

Follow-Up

There were 2 follow-up intervals for primary and secondary outcomes: 2 and 4 months after randomization. The 2-month follow-up facilitated the assessment of the immediate effects of the components, while the 4-month follow-up facilitated estimation of effects after prolonged access. Additional longer follow-up intervals were not included to reduce participant burden and because attrition was expected to be much higher at longer intervals from baseline. An additional follow-up was conducted at 1 month after randomization to assess proposed mediators of the intervention; these results will be reported separately. All follow-ups were initiated by sending SMS text messages to participants with hyperlinks to questionnaires (see Multimedia Appendix 3). Additional attempts were made to collect data from nonresponders, including 2 SMS text message reminders sent 2 days apart, followed by phone calls with a maximum of 5 attempts.

Randomization and Blinding

We used block randomization with random block sizes of 64 and 128 to allocate participants equally among conditions, where each condition represented a unique combination of components. The process was automated, with the sequence computer-generated and the backend server allocating eligible participants after completion of the baseline. Neither research personnel nor participants were able to influence the allocation. Research personnel were blind to allocation throughout the trial. Participants were not aware of components other than those to which they were allocated and were therefore considered blind to allocation.

Statistical Analysis

We conducted analyses with participants in the conditions to which they were randomly allocated (intention-to-treat). We estimated models using both available data and data with missing values imputed. We used multiple imputation with chained equations, generating 100 datasets using 30 iterations with predictive mean matching. To account for the uncertainty inherent in missing data, imputed analyses were performed for all models as a sensitivity analysis, regardless of the proportion of missingness. All statistical analyses were prespecified in the trial protocol [57].

Longitudinal data were modeled using multilevel models with random intercepts for participants and time-by-component interactions. In addition to estimating the marginal effects of each component, we estimated the effects of pairwise interactions among components to examine potential synergistic effects. This was done by including pairwise interaction terms in the regression models for all outcomes. The treatment effect being estimated was thus the effect of the presence versus absence of each specific component, adjusted for the presence and absence of other components, as well as pairwise interactions between components. The models were adjusted for age, gender, and baseline measures of importance, confidence, and knowledge of how to change.

Bayesian inference was used to estimate the parameters of the models, using standard normal priors for coefficients to incorporate an a priori conservative view of the magnitudes of effects [95]. The standard normal priors induced shrinkage in the parameter estimates to protect against spurious findings. We report the marginal posterior probability of effect (POE), with the median of the posterior distribution representing a point estimate of the magnitude of the effect, along with 95% compatibility intervals (CoIs) defined by the 2.5th and 97.5th percentiles.

Primary and Secondary Outcomes

We analyzed primary outcomes at 2 and 4 months among participants who fulfilled the respective inclusion criteria. For example, the MVPA outcome was analyzed among those who reported less than 150 minutes of MVPA in the past week at baseline. Secondary outcomes at 2 and 4 months were analyzed among all participants, except for the number of cigarettes smoked in the past week, which was analyzed among baseline smokers who continued to smoke.

Weekly alcohol consumption, frequency of heavy episodic drinking per month, number of sugary drinks per week, weekly intake of candy and snacks, and cigarettes smoked in the past week were all analyzed using negative binomial regression. Average intake of fruit and vegetables per day, MVPA minutes per week, BMI, and stress were all analyzed using linear

XSL•FO
RenderX

regression. Point prevalence of smoking abstinence was analyzed using logistic regression.

Ancillary Analyses

We investigated attrition using 2 approaches. First, we estimated the odds of responding to follow-up, conditional on baseline characteristics and the presence or absence of the 6 components (factors), using logistic regression. One model included no interaction terms, while a second model included interaction terms between factors and baseline characteristics. Given the large number of covariates in these models, we used Cauchy priors to promote a parsimonious model (centered at 0 with a standard normal hyperprior for the scale). The use of Cauchy priors encodes an a priori conservative view and pulls estimated associations toward the null unless the data strongly suggest otherwise. Second, if data are missing systematically, nonresponders may be more similar to responders who require more attempts to collect data than to those who require fewer attempts. We therefore modeled primary outcome data conditional on the number of attempts required to collect follow-up data. Associations between the number of attempts and outcome measures may indicate systematic attrition.

Post Hoc Analyses

To investigate whether the number of components that participants had access to affected outcomes, we conducted post hoc analyses in which outcomes were modeled conditional on the number of components. The same model specifications were used as in the primary analyses, with dummy variables for components replaced by a variable representing the component count.

Sample Size

We used a Bayesian sequential design to monitor when to stop recruitment [96]. In such designs, targets are defined and evaluated throughout the course of the trial to determine whether the trial should stop or continue. As 4-month follow-up data became available, we modeled each of the primary outcomes according to the analysis plan, and coefficients for dummy variables representing the presence or absence of each component were compared against targets for effect, harm, and futility (see protocol for details of the criteria [57]).

We stopped recruitment after 30 months. While our prespecified limit was 24 months of recruitment irrespective of targets, we extended recruitment by 6 months because some colleges and universities began their recruitment procedures later than others. Overall, each participating college and university was expected to recruit for at least 24 months.

Ethical Considerations

The trial received ethical approval from the Swedish Ethical Review Authority on December 15, 2020 (Dnr: 2020-05496 and Dnr: 2022-03561-02). All participants were provided with study information and gave informed consent before responding to the baseline questionnaire and subsequently being randomized. Participants received no compensation for participating in the trial. Participants' phone numbers were encrypted and stored in the study database to link baseline and follow-up data. At project completion, the encrypted phone numbers will be deleted, and the data will be anonymized.

Results

Overview

A CONSORT participant flow diagram is presented in Figure 2 (sign up and randomization) and Figure 3 (follow-up). Between April 13, 2021, and October 18, 2023, 2468 individuals expressed interest in the study, of whom 1804 consented. The majority were recruited through the trial's information materials, either digital or leaflets provided by university student health care services (1999/2468, 81%), while the remainder were recruited through online advertisements on social media platforms (469/2468, 19%). A total of 1718 individuals completed the baseline questionnaire, of whom 14 were excluded for not fulfilling any of the behavioral inclusion criteria. The remaining 1704 participants were randomized. Baseline characteristics of the randomized participants were balanced across all factors (see Table 2).

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) participant flow diagram—sign up and randomization.

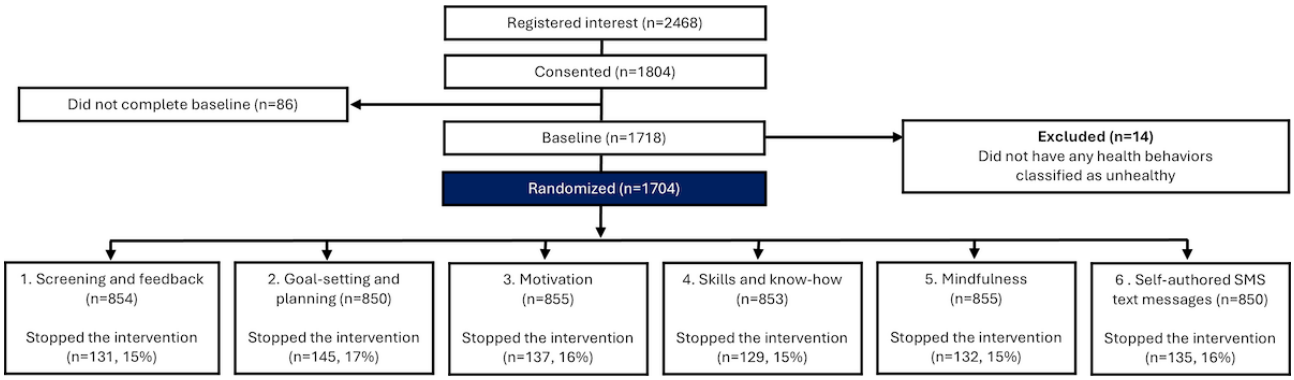

XSL•FO
RenderX

Figure 3. CONSORT (Consolidated Standards of Reporting Trials) participant flow diagram—follow-up.

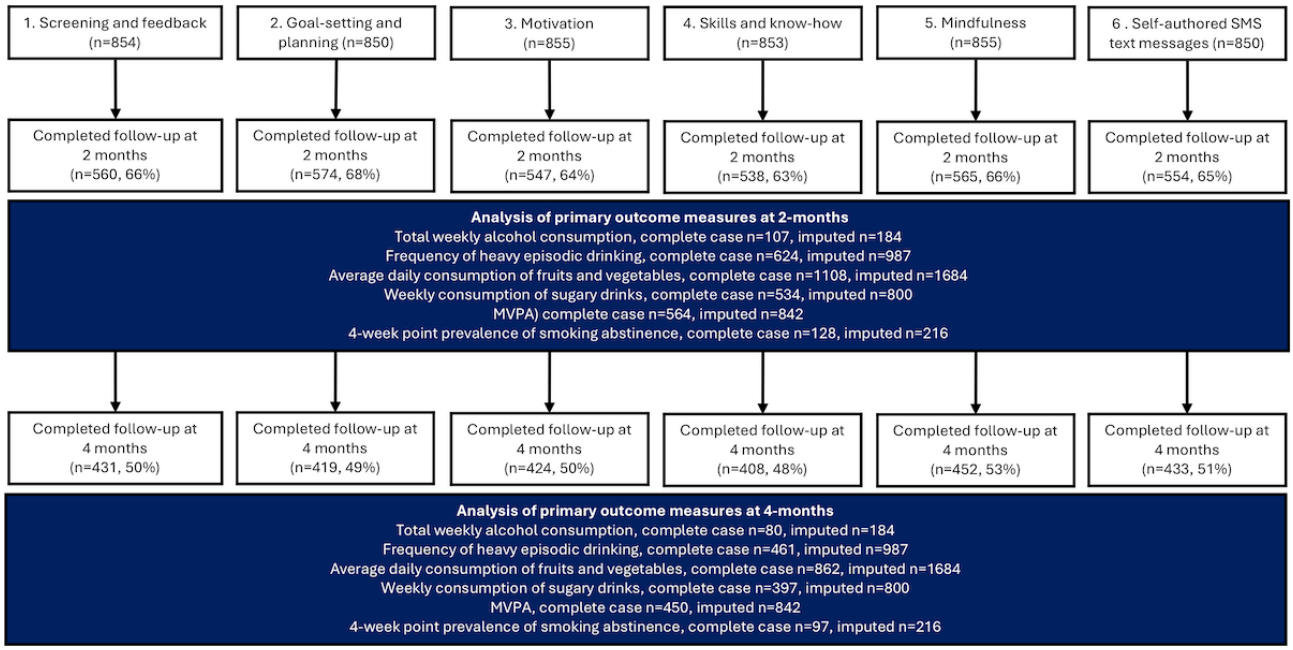

XSL•FO
RenderX

Table 2. Baseline characteristics of randomized participants.

Characteristics	Total (n=1704)	Screening and feedback (n=854)	Goal-setting and planning (n=850)	Motivation (n=855)	Skills and know-how (n=853)	Mindfulness (n=855)	Self-authored SMS text messages (n=850)
Sex, n (%)							
Women	1400 (82.16)	709 (83.02)	704 (82.82)	699 (81.75)	718 (84.17)	697 (81.52)	692 (81.41)
Men	304 (17.84)	145 (16.98)	146 (17.18)	156 (18.25)	135 (15.83)	158 (18.48)	158 (18.59)
Age (years), mean (SD)	28 (6.9)	28.3 (6.9)	28.5 (7.1)	28.2 (7.0)	27.8 (6.8)	28.5 (7.0)	28.4 (7.0)
BMI (kg/m^2), mean (SD)	25.7 (5.5)	25.9 (5.5)	25.6 (5.3)	25.7 (5.2)	25.8 (5.6)	25.7 (5.3)	25.8 (5.3)
Underweight (<18.5), n (%)	60 (3.52)	28 (3.28)	33 (3.88)	30 (3.51)	33 (3.87)	27 (3.16)	19 (2.24)
Normal weight (18.5-24.99), n (%)	844 (49.53)	417 (48.83)	413 (48.59)	415 (48.54)	429 (50.29)	426 (49.82)	432 (50.82)
Overweight (25-29.99), n (%)	458 (26.88)	234 (27.40)	236 (27.76)	239 (27.95)	214 (25.09)	241 (28.19)	227 (26.71)
Obese (≥30.0), n (%)	342 (20.07)	175 (20.49)	168 (19.76)	171 (20.00)	177 (20.75)	161 (18.83)	172 (20.24)
Alcohol, mean (SD)							
Total weekly alcohol consumption	3.7 (5.2)	3.7 (5.2)	3.8 (5.3)	3.7 (5.1)	3.6 (5.1)	3.8 (5.1)	3.7 (5.1)
Frequency of heavy episodic drinking	1.8 (2.7)	1.8 (2.7)	1.8 (2.8)	1.8 (2.7)	1.8 (2.7)	1.7 (2.4)	1.8 (2.5)
Dietary behavior, mean (SD)							
Average daily fruit and vegetable consumption, portions per week	1.3 (1.1)	1.3 (1.1)	1.3 (1.1)	1.3 (1.1)	1.3 (1.1)	1.3 (1.1)	1.3 (1.1)
Sugary drinks, cans per week	2.7 (4.1)	2.7 (4.1)	2.7 (4.2)	2.7 (4.1)	2.6 (3.9)	2.7 (4.0)	2.5 (3.9)
Candy and snacks, portions per week	6.5 (6.4)	6.6 (6.4)	6.4 (6.3)	6.6 (6.5)	6.5 (6.4)	6.5 (6.5)	6.3 (6.4)
Physical activity, mean (SD)							
MVPA (minutes)	207 (214)	209 (221)	214 (225)	213 (214)	211 (213)	207 (210)	196 (207)
Smoking							
Number of smokers, n (%)	218 (12.79)	106 (12.41)	119 (14.00)	106 (12.40)	105 (12.31)	104 (12.16)	101 (11.88)
Number of cigarettes last week, mean (SD)	3.5 (16.6)	3.0 (15.1)	4.2 (18.8)	3.6 (17.7)	2.6 (13.4)	3.3 (16.5)	2.8 (14.9)
Stress, mean (SD)							
Self-perceived stress[a]	7.8 (2.9)	7.8 (2.9)	7.8 (2.9)	7.7 (3.0)	7.8 (2.9)	7.7 (3.0)	7.7 (2.9)
Psychosocial measures, median (quartiles)							
Importance[b]	8 (7-10)	8 (7-10)	8 (7-10)	8 (7-10)	8 (7-10)	8 (7-10)	8 (7-10)
Confidence[b]	6 (5-8)	6 (5-8)	6 (5-8)	6 (5-8)	6 (5-8)	6 (5-8)	6 (5-8)
Know-how[b]	6 (5-8)	6 (5-8)	6 (5-8)	6 (5-8)	6 (5-8)	6 (5-8)	6 (5-8)

[a]The 4-item Perceived Stress Scale with total scores ranging from 0 to 16.

[b]Single item with 1-10 response options.

Among participants, 16% stopped the intervention before the 16-week program had ended. Consistent with the consent provided, we did not investigate the reasons why these individuals chose to stop receiving SMS text messages. The CONSORT flowchart in Figure 2 provides the numbers who stopped per component. All participants, including those who discontinued the intervention, were followed up and included in the main analysis. At the 2-month follow-up interval, primary outcome data were available for 1118 out of 1704 (65.61%)

XSL•FO
RenderX

participants, and for 874 out of 1704 (51.29%) participants at the 4-month follow-up interval.

Attrition

Attrition analyses revealed no marked associations between baseline characteristics and nonresponse at the 2-month follow-up interval. At the 4-month follow-up interval, there was evidence that older participants were more likely to respond to follow-up and that those smoking more cigarettes per week at baseline were less likely to respond. Notably, the association between nonresponse and smoking was driven by a few data points. Neither of these associations was found to differ between component conditions. Please see Multimedia Appendix 5 for full details.

There was some evidence that late responders, that is, those requiring more attempts to collect follow-up data, had healthier behaviors regarding fruit and vegetable intake and alcohol consumption. This association was attenuated among those with access to the screening and feedback (C1) component, who consumed more alcohol and sugary drinks at the 4-month follow-up and less fruit and vegetables at the 2-month follow-up. While there is considerable uncertainty surrounding these analyses, under the assumption that late responders are more similar to nonresponders than early responders, this may indicate systematic attrition. Please see Multimedia Appendix 6 for details.

Primary and Secondary Outcomes

Estimates and Main Effects

Table 3 contains the main effects of individual components on primary outcomes, and Multimedia Appendix 7 presents the main effects of individual components and 2-way interactions between components for the available data. The corresponding analyses with missing data imputed are presented in Multimedia Appendix 8. Here, we highlight findings for which the evidence of effects was strongest. We report estimates from analyses of the available data unless otherwise noted.

XSL•FO
RenderX

Table 3. Estimates of marginal effects of components on primary outcomes at 2- and 4-month follow-up.

Component	C1		C2		C3		C4		C5		C6	
	Estimate[a]	Probability[b], %	Estimate	Probability, %	Estimate	Probability, %	Estimate	Probability, %	Estimate	Probability, %	Estimate	Probability, %
Weekly alcohol consumption												
2 months	1.06 (0.77 to 1.50)	63.7	0.95 (0.68 to 1.32)	63.1	0.79 (0.58 to 1.08)	93.6	0.91 (0.66 to 1.25)	71.9	0.98 (0.72 to 1.34)	53.9	0.90 (0.65 to 1.23)	75.5
4 months	1.01 (0.71 to 1.45)	53.1	1.06 (0.74 to 1.53)	63.4	0.87 (0.62 to 1.22)	79.6	0.96 (0.67 to 1.36)	59.9	1.19 (0.84 to 1.69)	83.4	0.65 (0.45 to 0.93)	99.1
Heavy episodic drinking												
2 months	0.87 (0.74 to 1.02)	95.2	0.89 (0.75 to 1.04)	93.1	0.96 (0.82 to 1.13)	68.7	1.00 (0.86 to 1.18)	51.7	0.89 (0.76 to 1.04)	92.5	1.14 (0.98 to 1.34)	95.0
4 months	0.97 (0.80 to 1.16)	65.2	1.09 (0.91 to 1.32)	82.8	1.18 (0.99 to 1.42)	96.5	1.03 (0.86 to 1.25)	64.2	1.11 (0.92 to 1.33)	86.1	0.93 (0.77 to 1.12)	78.6
Fruit and vegetable consumption												
2 months	0.11 (–0.02 to 0.24)	94.7	–0.09 (–0.22 to 0.04)	90.2	0.09 (–0.04 to 0.23)	92.0	0.19 (0.06 to 0.33)	99.8	0.03 (–0.11 to 0.16)	64.7	–0.01 (–0.14 to 0.12)	54.9
4 months	0.12 (–0.03 to 0.26)	94.4	–0.01(–0.15 to 0.14)	53.6	0.12 (–0.02 to 0.27)	95.4	0.14 (–0.01 to 0.28)	96.9	–0.05 (–0.19 to 0.09)	74.9	–0.09 (–0.23 to 0.05)	88.2
Sugary drinks												
2 months	0.91 (0.77 to 1.06)	88.4	1.03 (0.87 to 1.21)	61.2	0.98 (0.84 to 1.15)	59.2	1.04 (0.88 to 1.22)	66.5	0.94 (0.80 to 1.10)	78.8	1.24 (1.05 to 1.46)	99.4
4 months	1.03 (0.86 to 1.24)	62.2	1.05 (0.87 to 1.26)	68.8	1.09 (0.90 to 1.31)	81.5	1.06 (0.88 to 1.27)	74.2	1.07 (0.89 to 1.29)	76.6	1.16 (0.96 to 1.39)	93.8
Moderate-to-vigorous physical activity												
2 months	9.34 (–25.6 to 45.5)	69.7	–5.73 (–41.9 to 29.5)	63.0	35.8 (–0.41 to 71.9)	97.4	17.1 (–19.2 to 54.2)	82.2	41.2 (5.04 to 77.6)	98.7	5.68 (–30.5 to 41.1)	62.2
4 months	28.1 (–11.3 to 67.6)	92.0	17.6 (–20.6 to 56.5)	81.4	6.63 (–33.1 to 45.2)	63.0	32.8 (–6.1 to 71.8)	95.2	3.02 (–36.3 to 42.0)	56.0	–19.2 (–57.6 to 20.1)	83.6
Smoking cessation												
2 months	1.12 (0.35 to 3.59)	57.4	0.56 (0.17 to 1.76)	84.3	1.43 (0.43 to 4.63)	72.3	0.68 (0.21 to 2.23)	74.5	0.46 (0.14 to 1.49)	90.4	2.15 (0.67 to 6.74)	90.4
4 months	1.45 (0.35 to 6.10)	69.2	0.48 (0.11 to 2.00)	84.3	2.30 (0.52 to 9.56)	87.3	0.77 (0.17 to 3.39)	64.0	0.55 (0.13 to 2.28)	79.6	0.94 (0.22 to 4.02)	53.0

[a]Estimates are presented as the median of the posterior distribution with 95% compatibility intervals. Incidence rate ratios are presented for weekly alcohol consumption, heavy episodic drinking, sugary drinks; odds ratios are presented for smoking cessation; and mean differences are presented for fruit and vegetable consumption and moderate-to-vigorous physical activity.

[b]Probability is the proportion of the posterior distribution above or below the null in the direction of the median.

XSL•FO
RenderX

Primary Outcomes

Total Weekly Alcohol Consumption

Self-authored SMS text messages (C6) reduced total weekly alcohol consumption at 4 months (incidence rate ratio [IRR] 0.65, CoI 0.45-0.93; POE 99.1%). Apart from weak evidence for an effect of motivation (C3) at 2 months, there was no marked evidence of effects for the other components individually. There was evidence that combining motivation (C3) and self-authored SMS text messages (C6) had a larger effect on alcohol consumption at 4 months (IRR 0.53, CoI 0.31-0.88; POE 99.4%). The evidence for the combined effect was also marked in the imputed analyses. Notably, relatively few participants drank above national guidelines with respect to weekly alcohol consumption, and therefore, the numbers included in these analyses were smaller than for other behaviors (see Figure 3 and Multimedia Appendices 7 and 8).

Frequency of Heavy Episodic Drinking

There was considerable variability in the evidence for the effectiveness of components on heavy episodic drinking, with evidence suggesting both benefit and harm. Screening and feedback (C1) reduced the frequency of heavy episodic drinking at 2 months (IRR 0.87, CoI 0.74-1.02; POE 95.2%), which was also evident in the imputed analyses. However, the effect was not sustained at the 4-month follow-up. The strongest evidence for benefit was observed for the combination of screening and feedback (C1) with goal-setting and planning (C2) (IRR 0.78, CoI 0.62-0.98; POE 98.4%), and screening and feedback (C1) with mindfulness (C5) (IRR 0.78, CoI 0.62-0.98; POE 98.4%) at 2 months. Similar evidence was found for the combination of goal-setting and planning (C2) with mindfulness (C5). Evidence of harm was most prominent for motivation (C3), both alone and in combination with goal-setting and planning (C2) or mindfulness (C5) (IRR 1.28, CoI 0.99-1.64; POE 97.1%) at the 4-month follow-up interval. These harmful effects were not evident at the 2-month interval. Notably, the evidence of harm was less prominent in the imputed analyses, whereas the evidence for benefit remained.

Daily Portions of Fruit and Vegetables

The strongest evidence for the beneficial effects of single components on fruit and vegetable consumption was found for screening and feedback (C1) and skills and know-how (C4) at both follow-up intervals. The largest effect was observed for the combination of these 2 components at 2 months (mean difference 0.30, CoI 0.11-0.48; POE 99.9%) and at 4 months (mean difference 0.26, CoI 0.05-0.46; POE 99.4%). While the evidence for motivation (C3) alone was not convincingly strong, in combination with screening and feedback (C1) or skills and know-how (C4), there was evidence of an effect at both the 2- and 4-month intervals on par with the effect observed for the combination of screening and feedback (C1) and skills and know-how (C4). These findings all persisted in the imputed analyses.

Weekly Consumption of Sugary Drinks

There was no strong evidence of the beneficial effects of the intervention components on sugary drink consumption. Rather, there was strong evidence of harm from self-authored SMS text messages (C6) at the 2- and 4-month intervals. The evidence suggested an increase of 15%-20% in weekly consumption of sugary drinks from this component alone, and the effect was larger in combination with goal-setting and planning (C2), motivation (C3), and skills and know-how (C4). These negative effects were partially sustained in the imputed analyses.

Weekly Moderate-to-Vigorous Physical Activity

There was evidence of beneficial effects on MVPA from several combinations of components at the 2-month interval; however, few of these effects persisted at the 4-month interval. The evidence for an effect of single components was strongest for motivation (C3), which increased weekly minutes of MVPA at 2 months (mean difference 35.8, CoI −0.41 to 71.9; POE 97.4%), and mindfulness (C5), which also increased minutes of MVPA at 2 months (mean difference 41.2, CoI 5.04-77.6; POE 98.7%). When motivation (C3) and mindfulness (C5) were combined, the effect on minutes of MVPA was even greater at 2 months (mean difference 78.0, CoI 28.3-128.2; POE 99.9%). Combining motivation (C3) with skills and know-how (C4) also had a positive effect on MVPA at 2 months. The imputed analyses produced similar findings.

While there was only weak evidence that screening and feedback (C1) alone affected MVPA, there was evidence that combining screening and feedback (C1) with skills and know-how (C4) increased minutes of MVPA at 4 months (mean difference 60.1, CoI 3.6-116.5; POE 98.2%), although the evidence was not as strong in the imputed analyses. In addition, there was evidence that screening and feedback (C1) combined with mindfulness (C5) increased minutes of MVPA at the 2-month interval (mean difference 50.6, CoI −0.55 to 101.9; POE 97.4%), with similar findings in the imputed analyses.

Smoking

There was weak evidence of both benefit and harm from the intervention components, and their combinations, on smoking cessation. The relatively strongest evidence for benefit was found for self-authored SMS text messages (C6) at 2 months (odds ratio 2.15, CoI 0.67-6.74; POE 90.4%), and for the combination of screening and feedback (C1) and motivation (C3) at 4 months (odds ratio 4.28, CoI 0.49-39.9; POE 90.7%). The strongest evidence for harm was found for the combination of goal-setting and planning (C2) and mindfulness (C5) at 2 months (odds ratio 0.23, CoI 0.04-1.27; POE 95.4%). The evidence for both benefit and harm persisted in the imputed analyses, although effect estimates were attenuated toward the null.

Secondary Outcomes

Weekly Portions of Candy and Snacks

At the individual component level, there was some evidence that motivation (C3) decreased candy and snack consumption at both 2 and 4 months. When motivation (C3) was combined with screening and feedback (C1), the strongest evidence for benefit was observed at 4 months (IRR 0.85, CoI 0.71-1.01; POE 96.8%). Increases in candy and snack consumption were observed for mindfulness (C5) alone at 4 months and self-authored SMS text messages (C6) alone at 2 months. While the evidence for these effects was not strong, it persisted in the

XSL•FO
RenderX

imputed analyses. In addition, combining mindfulness (C5) and self-authored SMS text messages (C6) increased candy and snack consumption at 4 months (IRR 1.21, CoI 1.01-1.44; POE 97.7%). These findings also persisted in the imputed analyses.

BMI

At the 2-month interval, there was evidence that mindfulness (C5) and self-authored SMS text messages (C6) increased BMI, both individually and when combined (combined mean difference 0.34, CoI 0.09-0.59; POE 99.6%). However, evidence for these effects was not observed at the 4-month interval. There was also evidence that combining screening and feedback (C1) with mindfulness (C5) resulted in higher BMI at both the 2- and 4-month intervals. These findings persisted in the imputed analyses.

Cigarettes Smoked Per Week

Regarding the number of cigarettes smoked per week among those who continued to smoke, there was evidence that the combination of screening and feedback (C1) and goal-setting and planning (C2) had a beneficial effect at 2 months, and that combining motivation (C3) and self-authored SMS text messages (C6) had a harmful effect at 2 months.

Perceived Stress

There was evidence of a beneficial effect on stress from the skills and know-how (C4) component alone at 4 months (mean difference −0.35, CoI −0.71 to 0.01; POE 97.1%). The effect size increased when skills and know-how (C4) was combined with screening and feedback (C1) or motivation (C3). These findings were not markedly different in the imputed analyses.

Post-Hoc Analyses

In Multimedia Appendix 9, we present estimates of the effects of the number of components participants had access to on primary and secondary outcomes. The relatively strongest evidence was found for fruit and vegetable consumption at 2 months (0.05 portion increase per module, CoI 0.001-0.11; POE 97.6%) and 4 months (0.04 portions per module, CoI −0.02 to 0.1; POE 90.7%), as well as for weekly MVPA at the 2-month follow-up (17.5 minutes per module, CoI 2.9-32.2; POE 99.0%) and 4 months (11.8 minutes per module, CoI −3.8 to 27.1; POE 93.1%).

Intervention Access

Of the 854 participants who had access to the screening and feedback component (C1), 807 (94.5%) screened themselves at least once during the intervention period. Among those who screened themselves at least once, 356 (44.1%) did so 4 or more times, while 129 (16.0%) screened themselves 3 times, 161 (20.0%) 2 times, and 161 (20.0%) 1 time.

Of the 850 participants who had access to the goal-setting and planning component (C2), 650 (76.5%) accessed the component at least once. Among all participants, 427 (50.2%) set at least one goal, and 403 (47.4%) selected or self-authored at least one challenge using the intervention's technical solution.

Of the 850 participants who had access to the motivation component (C3), 554 (65.2%) accessed the content at least once. Among all participants, 169 (19.9%) activated SMS text messages for physical activity, 166 (19.5%) for diet, 47 (5.5%) for alcohol, and 21 (2.5%) for smoking. The motivation component provided interactive content on motives, financial costs, and health consequences, with 294 recorded engagements using the intervention's technical solution.

Of the 855 participants who had access to the skills and know-how component (C4), 536 (62.7%) accessed the materials at least once. Among all participants, 166 (19.4%) activated SMS text messages for physical activity, 172 (20.1%) for diet, 51 (6.0%) for alcohol, and 28 (3.3%) for smoking. Most of the content in this module consisted of text and visual materials designed to build knowledge and skills for behavior change. A small number of interactive exercises were also included, with 283 recorded engagements using the intervention's technical solution.

Of the 855 participants who had access to the mindfulness component (C5), 435 (50.9%) accessed the component at least once. Meditations were accessed by 87 (10.2%) participants, and 156 (18.2%) completed at least one task. Among those who completed tasks, 76 (48.7%) completed 3 or more tasks, 33 (21.2%) completed 2 tasks, and 47 (30.1%) completed 1 task. Some tasks required participants to write longer texts, which could be completed either using the intervention's technical solution or offline; however, only activities completed using the technical solution could be measured.

Of the 850 participants who had access to the self-composed SMS text messages component (C6), 406 (47.8%) accessed the component at least once. Among those who accessed the component, a total of 570 SMS text messages were composed by 176 participants.

We had no means of accurately measuring the time spent on each component and therefore, cannot approximate the time spent reading the text and visual information provided in the components. Additionally, we had no means of measuring the extent to which participants comprehended and adopted the information provided. For instance, the skills and know-how component (C4) included mostly tips and information on concrete actions for participants to take, and we had no means of measuring how much time participants spent reading and engaging with the support provided.

Discussion

Principal Findings

This factorial randomized trial estimated the effectiveness of 6 behavior change components within a digital multiple health behavior intervention. We found evidence that both individual components and 2-way interactions between components led not only to notable improvements in health behaviors, but also to notable harm. The components screening/feedback (C1), motivation (C3), and skills/know-how (C4) each demonstrated individual and interacting improvements across various behavioral outcomes, including reductions in alcohol consumption, smoking, and stress, as well as increases in fruit and vegetable consumption and physical activity. Most effects were observed at 4 months; however, changes in heavy episodic

XSL•FO
RenderX

drinking and physical activity were noted only at the 2-month follow-up interval.

Previous factorial trials of health behavior change interventions have shown mixed results regarding the effectiveness of individual components. In general, combining components tends to lead to more substantial improvements in health behaviors compared with individual components alone. For instance, Crane et al [97] found that although individual modules of the Drink Less app did not affect weekly alcohol consumption or Alcohol Use Disorders Identification Test scores, 2-way interactions between components did. Specifically, the combination of normative feedback and cognitive bias retraining reduced weekly alcohol consumption, whereas self-monitoring, feedback, and action planning improved Alcohol Use Disorders Identification Test scores. Similarly, in our study, screening/feedback (C1) reduced heavy episodic drinking at 2 months, with enhanced effects when combined with goal-setting/planning (C2) or mindfulness (C5). Notably, the study by Crane et al [97] was not restricted to university students and focused solely on alcohol consumption.

Schroé et al [98] found that the MyPlan2.0 digital intervention increased MVPA and reduced sedentary behavior through self-monitoring. Additionally, coping planning further boosted MVPA, with the combination of action planning, coping planning, and self-monitoring proving most effective at 5 weeks. Similarly, Fanning et al [99] evaluated a 12-week physical activity app incorporating self-monitoring, goal-setting, and feedback, resulting in an 11-minute increase in daily MVPA (from 34 to 46 minutes). Participants who received points-based feedback engaged in an additional 6 minutes of MVPA per day. Our findings indicated that combining components such as screening/feedback (C1), motivation (C3), skills/know-how (C4), and mindfulness (C5) enhanced short-term physical activity. However, this increase was not sustained over time, suggesting that although the components were effective at 2 months, they did not support the longer-term maintenance of increased physical activity at 4 months. The dissimilarities in findings may be due to both Schroé et al [98] and Fanning et al [99] focusing exclusively on physical activity and not restricting their samples to university students.

The evidence showed no effect for the goal-setting/planning component (C2) when delivered alone. Harms were observed for the interaction between goal-setting/planning (C2) and mindfulness (C5) on smoking cessation at 2 months. Additionally, self-authored SMS text messages (C6) increased sugary drink consumption, with the effect amplified when combined with goal-setting/planning (C2), motivation (C3), or skills/know-how (C4). These findings regarding sugary drinks may be related to the observed decrease in alcohol consumption, as individuals might substitute sugary drinks for alcohol, and behavior substitution was part of the advice provided to participants. Additionally, students may consume energy drinks during studying and physical activity, further contributing to increased sugary beverage intake. Previous research [100] highlighted the complexity of young adults' perceptions of beverage healthfulness, which are influenced by factors including sugar content, packaging, marketing, and context of consumption. For example, a drink may be perceived as healthier if consumed during physical activity or as part of a balanced meal.

Although beneficial, health behavior interventions can increase stress and health anxiety due to heightened health consciousness, dissonance with desired behaviors, or pressure to meet goals, all of which may negatively affect mental well-being [101]. However, our study found that perceived psychological stress decreased at both the 2- and 4-month intervals when participants were provided with the skills/know-how component (C4). This effect became more pronounced over time, highlighting the importance of equipping individuals with practical skills and knowledge to manage their health in daily life.

Taking a broader perspective beyond the scope of factorial trials, our findings differ from previous digital multiple health interventions for young adults and university students, which have reported small effect sizes when evaluating interventions as a whole. For instance, Champion et al [102] reported that a digital multiple health intervention showed no evidence of effect in modifying lifestyle behaviors among adolescents in Australia; however, this study included a much younger demographic (ages 11-13) than our study. The online intervention U@Uni by Cameron et al [103], which targeted new university students, found no evidence of effect on the 4 primary outcomes: portions of fruit and vegetables per day, physical activity in the previous week, alcohol consumption in the previous week, and smoking status at the 6-month follow-up. However, U@Uni also targeted a younger demographic of incoming university students who may not have had the same motivation to change as students who had already established themselves. By contrast, similar to our study, Duan et al [104] demonstrated positive effects on physical activity and fruit and vegetable consumption after 8 weeks of a sequentially delivered web-based intervention among Chinese college students.

Generalizability and Limitations

This trial employed a pragmatic recruitment strategy, reaching students at universities across Sweden through channels typically used by Student Health Care Centers to provide health information and services. This open invitation, with broad inclusion criteria and low barriers to participation, targeted students with at least one health behavior categorized as a significant risk factor for their physical or mental health. Although the validity of trials is not contingent on the sample being representative of the target population, in our case, we anticipate that those who would use the intervention in a real-world rollout would be similar to those who participated in this trial. Therefore, we argue that this trial estimates the effectiveness, rather than the efficacy, of the intervention components. Relatedly, we do not have data to determine which recruitment source led to each individual's participation; thus, a limitation of the study design is that we cannot assess whether effectiveness was moderated by mode of recruitment.

In line with our aim to estimate effectiveness, we designed the study to gain knowledge about how access to components affects outcomes in a large population, rather than to examine the efficacy of different components in controlled environments. Effectiveness, as a concept, seeks to understand what happens under real-world conditions; for this reason, we neither expected

XSL•FO
RenderX

nor observed that all participants would use all of the intervention materials available to them. Nevertheless, in this study, which proactively recruited university students, 768 of 854 (89.9%) with access to the screening component used it at least once, and approximately one-half of participants with access to the other components used them at least once. Based on the measures available to us, the 2 components with the lowest usage were mindfulness (C5) and self-authored SMS text messages (C6). This is not surprising, as both require a stronger cognitive commitment, and mindfulness was not advertised to participants before enrollment; thus, there may have been limited capacity or interest to engage with these components. However, even if a smaller proportion of individuals engage with specific components, it may still be beneficial to include them in interventions, as they may have positive effects from a population perspective. Overall, the findings from this study should be understood as generalizing to the proactive dissemination of digital interventions rather than as controlled treatment of specific individuals.

The invitation reached individuals who were not necessarily actively seeking help but who could benefit from it and were willing to participate when approached. Although recruitment did not target students who had explicitly requested support for change, baseline measures indicated a gap between participants' ratings of importance and their perceived capacity to change, with importance rated, on average, 8 out of 10 and confidence and know-how rated 6 out of 10. This suggests that our proactive recruitment strategy can reach those who may benefit from, and be interested in, support. The relatively low baseline ratings of know-how and confidence also suggest potential to deliver effective behavior change support, as previous research shows that enhancing individuals' confidence in their ability to change and improving their knowledge of how to change can lead to positive behavior modifications, such as reduced alcohol consumption and smoking cessation [105,106]. It should be noted that these 3 measures did not constitute a primary objective of this trial, and to reduce participant burden, we used single-item measures, which cannot capture a more nuanced understanding of confidence, importance, and know-how. A related limitation is that, because participants were not necessarily actively seeking help, baseline screening leading to inclusion may have prompted reflection on behavior for some participants, which in itself may have induced change. Notably, this occurred before randomization; thus, the consequence may be that estimates are biased toward the null.

Attrition rates were high in this study; however, they were consistent with what can be expected in remote trials with low barriers to participation [107]. Attrition analyses did not indicate marked differences across factorial conditions, thereby strengthening the internal validity of the findings. However, there was some evidence that late responders and nonresponders exhibited healthier behaviors compared with early responders. This underscores the importance of the imputed analyses in supporting the main findings.

The factorial design ensured blinding by providing every participant with some form of intervention. Although social desirability is always a concern in studies in which the objectives are clear to participants—and possibly more so when outcomes are self-reported—blinding reduces the risk of compensatory rivalry and other research participation effects [108,109]. Therefore, the risk of bias stemming from the use of self-reported measures in this study is somewhat tempered. Regarding blinding, there was a risk that participants could have deliberately or inadvertently learned about others' participation and shared intervention content, thereby becoming partially aware of other conditions. Although we cannot determine the extent to which this occurred, we judge the risk of bias to be low. If such contamination did occur, our estimates of effectiveness are likely biased toward the null.

Finally, the factorial design used in this trial allowed us to study the individual components of Buddy; however, it limited our ability to estimate the total effect of the intervention as a whole. Therefore, future studies are needed to investigate the overall effects of the Buddy intervention. Indeed, post hoc analyses suggested that access to a greater number of components increased effectiveness with respect to fruit and vegetable consumption and physical activity.

Conclusions

Using a randomized factorial design, we dismantled a complex digital multiple behavior intervention and provided novel insights into the effectiveness of its different components. Both marginal and synergistic effects were estimated across multiple behaviors. Our results highlight which components of digital interventions contribute to effectiveness and demonstrate that component effectiveness is heterogeneous across the targeted behaviors. The findings provide an evidence base for designing more effective digital interventions. These findings should be considered in light of the risk of bias introduced by attrition to follow-up, which was high in this effectiveness trial with low barriers to participation. The trial also relied on self-reported measures, although the risk of bias associated with this design choice may have been partially mitigated by participant blinding.

Funding

This trial was funded by the Swedish Research Council for Health, Working Life and Welfare (grant 2018-01410; principal investigator: ML). The trial was also supported by funding from the Swedish Cancer Society (Cancerfonden; grants 20 0883 Pj and 23 2888 Pj; principal investigator: MB). The funders had no role in the planning, design, conduct, analysis, or reporting of the study.

XSL•FO
RenderX

Data Availability

The datasets used or analyzed in this study are available from the corresponding author (MB) upon reasonable request. Ethical approval is required for data sharing.

Authors' Contributions

Conceptualization: PB, MB, ML
Data curation: KÅ, MB
Formal analysis: KÅ, MB
Investigation: KÅ, MB
Methodology: MB, ML, PH
Project administration: KÅ, MB
Resources: KÅ, OL, HH, PH, ACE, PB, ML, MB
Software: MB
Writing – original draft: KÅ, MB
Writing – review & editing: KÅ, OL, HH, PH, ACE, PB, ML, MB

Conflicts of Interest

MB and PB own a private company (Alexit AB) that develops and distributes digital solutions to the public and private sectors. Alexit AB played no role in the development of the intervention, study design, data analysis, data interpretation, or writing of this report. All other authors declare no conflicts of interest.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 984 KB - jmir_v28i1e88884_app1.pdf]

Multimedia Appendix 2
Informed consent form.
[PDF File (Adobe PDF File), 48 KB - jmir_v28i1e88884_app2.pdf]

Multimedia Appendix 3
Questionnaires.
[PDF File (Adobe PDF File), 75 KB - jmir_v28i1e88884_app3.pdf]

Multimedia Appendix 4
Intervention description and allocation.
[PDF File (Adobe PDF File), 1567 KB - jmir_v28i1e88884_app4.pdf]

Multimedia Appendix 5
Attrition nonresponse.
[PDF File (Adobe PDF File), 841 KB - jmir_v28i1e88884_app5.pdf]

Multimedia Appendix 6
Attrition attempts.
[PDF File (Adobe PDF File), 95 KB - jmir_v28i1e88884_app6.pdf]

Multimedia Appendix 7
Complete-case estimates of effect.
[PDF File (Adobe PDF File), 265 KB - jmir_v28i1e88884_app7.pdf]

Multimedia Appendix 8
Effect estimates based on imputed data.
[PDF File (Adobe PDF File), 267 KB - jmir_v28i1e88884_app8.pdf]

Multimedia Appendix 9
Estimates of the effects of component count.

XSL•FO
RenderX

[PDF File (Adobe PDF File), 87 KB - jmir_v28i1e88884_app9.pdf]

References

1. GBD 2020 Alcohol Collaborators. Population-level risks of alcohol consumption by amount, geography, age, sex, and year: a systematic analysis for the Global Burden of Disease Study 2020. Lancet 2022 Jul 16;400(10347):185-235 [FREE Full text] [doi: 10.1016/S0140-6736(22)00847-9] [Medline: 35843246]
2. GBD 2019 Tobacco Collaborators. Spatial, temporal, and demographic patterns in prevalence of smoking tobacco use and attributable disease burden in 204 countries and territories, 1990-2019: a systematic analysis from the Global Burden of Disease Study 2019. Lancet 2021 Jun 19;397(10292):2337-2360 [FREE Full text] [doi: 10.1016/S0140-6736(21)01169-7] [Medline: 34051883]
3. GBD 2021 Diabetes Collaborators. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet 2023 Jul 15;402(10397):203-234 [FREE Full text] [doi: 10.1016/S0140-6736(23)01301-6] [Medline: 37356446]
4. GBD 2021 Risk Factors Collaborators. Global burden and strength of evidence for 88 risk factors in 204 countries and 811 subnational locations, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet 2024 May 18;403(10440):2162-2203 [FREE Full text] [doi: 10.1016/S0140-6736(24)00933-4] [Medline: 38762324]
5. World Health Organisation (WHO). Fact sheet Noncommunicable diseases. WHO. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases [accessed 2025-01-01]
6. Ashdown-Franks G, Vancampfort D, Firth J, Smith L, Sabiston CM, Stubbs B, et al. Association of leisure-time sedentary behavior with fast food and carbonated soft drink consumption among 133,555 adolescents aged 12-15 years in 44 low- and middle-income countries. Int J Behav Nutr Phys Act 2019 Apr 23;16(1):35 [FREE Full text] [doi: 10.1186/s12966-019-0796-3] [Medline: 31014359]
7. Patrick ME, Evans-Polce RJ, Arterberry BJ, Terry-McElrath Y. Initiation of and escalation to high-intensity drinking in young adults. JAMA Pediatr 2023 Mar 01;177(3):286-293 [FREE Full text] [doi: 10.1001/jamapediatrics.2022.5642] [Medline: 36716022]
8. Poole R, Carver H, Anagnostou D, Edwards A, Moore G, Smith P, et al. Tobacco use, smoking identities and pathways into and out of smoking among young adults: a meta-ethnography. Subst Abuse Treat Prev Policy 2022 Mar 28;17(1):24 [FREE Full text] [doi: 10.1186/s13011-022-00451-9] [Medline: 35346260]
9. Donat-Vargas C, Guerrero-Zotano, Casas A, Baena-Cañada JM, Lope V, Antolín S, et al. Trajectories of alcohol consumption during life and the risk of developing breast cancer. Br J Cancer 2021 Oct 06;125(8):1168-1176 [FREE Full text] [doi: 10.1038/s41416-021-01492-w] [Medline: 34483338]
10. Rodriguez-Ayllon M, Cadenas-Sánchez C, Estévez-López F, Muñoz NE, Mora-Gonzalez J, Migueles JH, et al. Role of physical activity and sedentary behavior in the mental health of preschoolers, children and adolescents: a systematic review and meta-analysis. Sports Med 2019 Sep 16;49(9):1383-1410. [doi: 10.1007/s40279-019-01099-5] [Medline: 30993594]
11. Guasch-Ferré M, Willett WC. The Mediterranean diet and health: a comprehensive overview. J Intern Med 2021 Sep 23;290(3):549-566 [FREE Full text] [doi: 10.1111/joim.13333] [Medline: 34423871]
12. Li Y, Schoufour J, Wang DD, Dhana K, Pan A, Liu X, et al. Healthy lifestyle and life expectancy free of cancer, cardiovascular disease, and type 2 diabetes: prospective cohort study. BMJ 2020 Jan 08;368:l6669 [FREE Full text] [doi: 10.1136/bmj.l6669] [Medline: 31915124]
13. Nyberg ST, Singh-Manoux A, Pentti J, Madsen IEH, Sabia S, Alfredsson L, et al. Association of healthy lifestyle with years lived without major chronic diseases. JAMA Intern Med 2020 May 01;180(5):760-768 [FREE Full text] [doi: 10.1001/jamainternmed.2020.0618] [Medline: 32250383]
14. Steffen LM, Van Horn L, Daviglus ML, Zhou X, Reis JP, Loria CM, et al. A modified Mediterranean diet score is associated with a lower risk of incident metabolic syndrome over 25 years among young adults: the CARDIA (Coronary Artery Risk Development in Young Adults) study. Br J Nutr 2014 Sep 19;112(10):1654-1661. [doi: 10.1017/s0007114514002633]
15. Alosaimi N, Sherar LB, Griffiths P, Pearson N. Clustering of diet, physical activity and sedentary behaviour and related physical and mental health outcomes: a systematic review. BMC Public Health 2023 Aug 18;23(1):1572 [FREE Full text] [doi: 10.1186/s12889-023-16372-6] [Medline: 37596591]
16. Kemel PN, Porter JE, Coombs N. Improving youth physical, mental and social health through physical activity: a systematic literature review. Health Promot J Austr 2022 Jul 15;33(3):590-601. [doi: 10.1002/hpja.553] [Medline: 34735738]
17. Solomou S, Logue J, Reilly S, Perez-Algorta G. A systematic review of the association of diet quality with the mental health of university students: implications in health education practice. Health Educ Res 2023 Jan 20;38(1):28-68 [FREE Full text] [doi: 10.1093/her/cyac035] [Medline: 36441584]
18. Bailey AP, Hetrick SE, Rosenbaum S, Purcell R, Parker AG. Treating depression with physical activity in adolescents and young adults: a systematic review and meta-analysis of randomised controlled trials. Psychol Med 2017 Oct 10;48(7):1068-1083. [doi: 10.1017/s0033291717002653]
19. Yonek JC, Meacham MC, Shumway M, Tolou-Shams M, Satre DD. Smoking reduction is associated with lower alcohol consumption and depressive symptoms among young adults over one year. Drug Alcohol Depend 2021 Oct 01;227:108922 [FREE Full text] [doi: 10.1016/j.drugalcdep.2021.108922] [Medline: 34364192]

XSL•FO
RenderX

20. Taylor G, Lindson N, Farley A, Leinberger-Jabari A, Sawyer K, Te Water Naudé R, et al. Smoking cessation for improving mental health. Cochrane Database Syst Rev 2021 Mar 09;3(3):CD013522 [FREE Full text] [doi: 10.1002/14651858.CD013522.pub2] [Medline: 33687070]
21. Haverkamp BF, Wiersma R, Vertessen K, van Ewijk H, Oosterlaan J, Hartman E. Effects of physical activity interventions on cognitive outcomes and academic performance in adolescents and young adults: a meta-analysis. J Sports Sci 2020 Dec 12;38(23):2637-2660 [FREE Full text] [doi: 10.1080/02640414.2020.1794763] [Medline: 32783695]
22. Seilo N, Paldanius S, Autio R, Kunttu K, Kaila M. Associations between e-health questionnaire responses, health checks and graduation: Finnish register-based study of 2011-2012 university entrants. BMJ Open 2020 Dec 16;10(12):e041551 [FREE Full text] [doi: 10.1136/bmjopen-2020-041551] [Medline: 33328260]
23. Cousins K, Connor JL, Kypri K. Effects of the Campus Watch intervention on alcohol consumption and related harm in a university population. Drug Alcohol Depend 2014 Oct 01;143:120-126. [doi: 10.1016/j.drugalcdep.2014.07.015] [Medline: 25108584]
24. Cho H, Yang Y. Relationship between alcohol consumption and risky sexual behaviors among adolescents and young adults: a meta-analysis. Int J Public Health 2023;68:1605669 [FREE Full text] [doi: 10.3389/ijph.2023.1605669] [Medline: 37153699]
25. Pengpid S, Peltzer K. Prevalence and correlates of multiple behavioural risk factors of non-communicable diseases among university students from 24 countries. J Public Health (Oxf) 2021 Dec 10;43(4):857-866. [doi: 10.1093/pubmed/fdaa138] [Medline: 34918087]
26. Zhang Y, Pan X, Chen J, Cao A, Xia L, Zhang Y, et al. Combined lifestyle factors, all-cause mortality and cardiovascular disease: a systematic review and meta-analysis of prospective cohort studies. J Epidemiol Community Health 2021 Jan 05;75(1):92-99. [doi: 10.1136/jech-2020-214050] [Medline: 32892156]
27. Oosterveen E, Tzelepis F, Ashton L, Hutchesson MJ. A systematic review of eHealth behavioral interventions targeting smoking, nutrition, alcohol, physical activity and/or obesity for young adults. Prev Med 2017 Jun;99:197-206. [doi: 10.1016/j.ypmed.2017.01.009] [Medline: 28130046]
28. Schoeppe S, Alley S, Van Lippevelde W, Bray NA, Williams SL, Duncan MJ, et al. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review. Int J Behav Nutr Phys Act 2016 Dec 07;13(1):127 [FREE Full text] [doi: 10.1186/s12966-016-0454-y] [Medline: 27927218]
29. Mönninghoff A, Kramer JN, Hess AJ, Ismailova K, Teepe GW, Tudor Car L, et al. Long-term effectiveness of mHealth physical activity interventions: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2021 Apr 30;23(4):e26699. [doi: 10.2196/26699]
30. Kaner E, Beyer F, Muirhead C, Campbell F, Pienaar E, Bertholet N, et al. Effectiveness of brief alcohol interventions in primary care populations. Cochrane Database Syst Rev 2018 Feb 24;2(2):CD004148 [FREE Full text] [doi: 10.1002/14651858.CD004148.pub4] [Medline: 29476653]
31. O Rourke L, Humphris G, Baldacchino A. Electronic communication based interventions for hazardous young drinkers: a systematic review. Neurosci Biobehav Rev 2016 Sep;68:880-890. [doi: 10.1016/j.neubiorev.2016.07.021] [Medline: 27453570]
32. Taylor G, Dalili M, Semwal M, Civljak M, Sheikh A, Car J. Internet-based interventions for smoking cessation. Cochrane Database Syst Rev 2017 Sep 04;9(9):CD007078 [FREE Full text] [doi: 10.1002/14651858.CD007078.pub5] [Medline: 28869775]
33. McIntosh J, Jay S, Hadden N, Whittaker P. Do E-health interventions improve physical activity in young people: a systematic review. Public Health 2017 Jul;148:140-148. [doi: 10.1016/j.puhe.2017.04.001]
34. Montagni I, Cariou T, Feuillet T, Langlois E, Tzourio C. Exploring digital health use and opinions of university students: field survey study. JMIR Mhealth Uhealth 2018 Mar 15;6(3):e65 [FREE Full text] [doi: 10.2196/mhealth.9131] [Medline: 29549071]
35. Berman AH, Gajecki M, Sinadinovic K, Andersson C. Mobile interventions targeting risky drinking among university students: a review. Curr Addict Rep 2016 Apr 5;3(2):166-174 [FREE Full text] [doi: 10.1007/s40429-016-0099-6] [Medline: 27226948]
36. Prosser T, Gee KA, Jones F. A meta-analysis of effectiveness of E-interventions to reduce alcohol consumption in college and university students. J Am Coll Health 2018 Apr 19;66(4):292-301. [doi: 10.1080/07448481.2018.1440579] [Medline: 29452058]
37. Dute DJ, Bemelmans WJE, Breda J. Using mobile apps to promote a healthy lifestyle among adolescents and students: a review of the theoretical basis and lessons learned. JMIR Mhealth Uhealth 2016 May 05;4(2):e39 [FREE Full text] [doi: 10.2196/mhealth.3559] [Medline: 27150850]
38. Brown J. A review of the evidence on technology-based interventions for the treatment of tobacco dependence in college health. Worldviews Evid Based Nurs 2013 Aug 19;10(3):150-162. [doi: 10.1111/wvn.12000] [Medline: 23421669]
39. McCambridge J, Bendtsen M, Karlsson N, White IR, Nilsen P, Bendtsen P. Alcohol assessment and feedback by email for university students: main findings from a randomised controlled trial. Br J Psychiatry 2013 Nov 02;203(5):334-340 [FREE Full text] [doi: 10.1192/bjp.bp.113.128660] [Medline: 24072758]

XSL•FO
RenderX

40. Thomas K, Müssener U, Linderoth C, Karlsson N, Bendtsen P, Bendtsen M. Effectiveness of a text messaging-based intervention targeting alcohol consumption among university students: randomized controlled trial. JMIR Mhealth Uhealth 2018 Jun 25;6(6):e146 [FREE Full text] [doi: 10.2196/mhealth.9642] [Medline: 29941417]
41. Bendtsen M. An electronic screening and brief intervention for hazardous and harmful drinking among Swedish university students: reanalysis of findings from a randomized controlled trial using a Bayesian framework. J Med Internet Res 2019 Dec 17;21(12):e14420 [FREE Full text] [doi: 10.2196/14420] [Medline: 31845903]
42. Berman AH, Andersson C, Gajecki M, Rosendahl I, Sinadinovic K, Blankers M. Smartphone apps targeting hazardous drinking patterns among university students show differential subgroup effects over 20 weeks: results from a randomized, controlled trial. J Clin Med 2019 Oct 28;8(11):1807 [FREE Full text] [doi: 10.3390/jcm8111807] [Medline: 31661868]
43. Müssener U, Bendtsen M, Karlsson N, White IR, McCambridge J, Bendtsen P. Effectiveness of short message service text-based smoking cessation intervention among university students: a randomized clinical trial. JAMA Intern Med 2016 Mar 01;176(3):321-328 [FREE Full text] [doi: 10.1001/jamainternmed.2015.8260] [Medline: 26903176]
44. Bendtsen M, Müssener U, Linderoth C, Thomas K. A mobile health intervention for mental health promotion among university students: randomized controlled trial. JMIR Mhealth Uhealth 2020 Mar 20;8(3):e17208 [FREE Full text] [doi: 10.2196/17208] [Medline: 32196462]
45. Chow CK, Redfern J, Hillis GS, Thakkar J, Santo K, Hackett ML, et al. Effect of lifestyle-focused text messaging on risk factor modification in patients with coronary heart disease: a randomized clinical trial. JAMA 2015 Sep 22;314(12):1255-1263. [doi: 10.1001/jama.2015.10945] [Medline: 26393848]
46. Palmer M, Sutherland J, Barnard S, Wynne A, Rezel E, Doel A, et al. The effectiveness of smoking cessation, physical activity/diet and alcohol reduction interventions delivered by mobile phones for the prevention of non-communicable diseases: a systematic review of randomised controlled trials. PLoS One 2018 Jan 5;13(1):e0189801 [FREE Full text] [doi: 10.1371/journal.pone.0189801] [Medline: 29304148]
47. Milne-Ives M, Lam C, De Cock C, Van Velthoven MH, Meinert E. Mobile apps for health behavior change in physical activity, diet, drug and alcohol use, and mental health: systematic review. JMIR Mhealth Uhealth 2020 Mar 18;8(3):e17046 [FREE Full text] [doi: 10.2196/17046] [Medline: 32186518]
48. Atorkey P, Paul C, Wiggers J, Bonevski B, Nolan E, Oldmeadow C, et al. Clustering of multiple health-risk factors among vocational education students: a latent class analysis. Transl Behav Med 2021 Oct 23;11(10):1931-1940. [doi: 10.1093/tbm/ibab068] [Medline: 34155507]
49. Noble N, Paul C, Turon H, Oldmeadow C. Which modifiable health risk behaviours are related? A systematic review of the clustering of Smoking, Nutrition, Alcohol and Physical activity ('SNAP') health risk factors. Prev Med 2015 Dec;81:16-41. [doi: 10.1016/j.ypmed.2015.07.003] [Medline: 26190368]
50. Jonsson KR, Corell M, Löfstedt P, Adjei NK. The clustering of multiple health and lifestyle behaviors among Swedish adolescents: a person-oriented analysis. Front Public Health 2023;11:1178353 [FREE Full text] [doi: 10.3389/fpubh.2023.1178353] [Medline: 37538263]
51. Mair J, Salamanca-Sanabria A, Augsburger M, Frese B, Abend S, Jakob R, et al. Effective behavior change techniques in digital health interventions for the prevention or management of noncommunicable diseases: an umbrella review. Ann Behav Med 2023 Sep 13;57(10):817-835 [FREE Full text] [doi: 10.1093/abm/kaad041] [Medline: 37625030]
52. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010 Feb 17;12(1):e4 [FREE Full text] [doi: 10.2196/jmir.1376] [Medline: 20164043]
53. Dugas M, Gao G, Agarwal R. Unpacking mHealth interventions: a systematic review of behavior change techniques used in randomized controlled trials assessing mHealth effectiveness. Digit Health 2020;6:2055207620905411 [FREE Full text] [doi: 10.1177/2055207620905411] [Medline: 32128233]
54. Aguiar M, Trujillo M, Chaves D, Álvarez R, Epelde G. mHealth apps using behavior change techniques to self-report data: systematic review. JMIR Mhealth Uhealth 2022 Sep 09;10(9):e33247 [FREE Full text] [doi: 10.2196/33247] [Medline: 36083606]
55. Pham Q, Wiljer D, Cafazzo JA. Beyond the randomized controlled trial: a review of alternatives in mHealth clinical trial methods. JMIR Mhealth Uhealth 2016 Sep 09;4(3):e107 [FREE Full text] [doi: 10.2196/mhealth.5720] [Medline: 27613084]
56. Bendtsen M, Bendtsen P, Henriksson H, Henriksson P, Müssener U, Thomas K, et al. The Mobile Health Multiple Lifestyle Behavior Interventions Across the Lifespan (MoBILE) research program: protocol for development, evaluation, and implementation. JMIR Res Protoc 2020 Apr 20;9(4):e14894 [FREE Full text] [doi: 10.2196/14894] [Medline: 32310147]
57. Åsberg K, Lundgren O, Henriksson H, Henriksson P, Bendtsen P, Löf M, et al. Multiple lifestyle behaviour mHealth intervention targeting Swedish college and university students: protocol for the Buddy randomised factorial trial. BMJ Open 2021 Dec 30;11(12):e051044. [doi: 10.1136/bmjopen-2021-051044]
58. Schulz K, Altman D, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010 Mar 23;340:c332 [FREE Full text] [doi: 10.1136/bmj.c332] [Medline: 20332509]
59. Kahan BC, Hall SS, Beller EM, Birchenall M, Chan A, Elbourne D, et al. Reporting of factorial randomized trials: extension of the CONSORT 2010 statement. JAMA 2023 Dec 05;330(21):2106-2114. [doi: 10.1001/jama.2023.19793] [Medline: 38051324]

XSL•FO
RenderX

60. Conner M, Norman P. Predicting and Changing Health Behaviour: Research and Practice with Social Cognition Models. Berkshire, UK: McGraw-Hill Education (UK); 2015.
61. Bendtsen M, Åsberg K, McCambridge J. Effectiveness of a digital intervention versus alcohol information for online help-seekers in Sweden: a randomised controlled trial. BMC Med 2022 May 17;20(1):176 [FREE Full text] [doi: 10.1186/s12916-022-02374-5] [Medline: 35578276]
62. Blomqvist J, Gunnarsson KU, Bendtsen P, Bendtsen M. Effects of a text messaging smoking cessation intervention amongst online help-seekers and primary health care visitors: findings from a randomised controlled trial. BMC Med 2023 Oct 04;21(1):382 [FREE Full text] [doi: 10.1186/s12916-023-03073-5] [Medline: 37794399]
63. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The Behavior Change Technique Taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Mar 20;46(1):81-95. [doi: 10.1007/s12160-013-9486-6]
64. Michie S, Whittington C, Hamoudi Z, Zarnani F, Tober G, West R. Identification of behaviour change techniques to reduce excessive alcohol consumption. Addiction 2012 Apr 17;107(8):1431-1440. [doi: 10.1111/j.1360-0443.2012.03845.x]
65. Bendtsen P, McCambridge J, Bendtsen M, Karlsson N, Nilsen P. Effectiveness of a proactive mail-based alcohol internet intervention for university students: dismantling the assessment and feedback components in a randomized controlled trial. J Med Internet Res 2012 Oct 31;14(5):e142 [FREE Full text] [doi: 10.2196/jmir.2062] [Medline: 23113955]
66. Bendtsen P, Bendtsen M, Karlsson N, White IR, McCambridge J. Online alcohol assessment and feedback for hazardous and harmful drinkers: findings from the AMADEUS-2 randomized controlled trial of routine practice in Swedish universities. J Med Internet Res 2015 Jul 09;17(7):e170 [FREE Full text] [doi: 10.2196/jmir.4020] [Medline: 26159179]
67. Michie S, Abraham C, Whittington C, McAteer J, Gupta S. Effective techniques in healthy eating and physical activity interventions: a meta-regression. Health Psychol 2009 Nov;28(6):690-701. [doi: 10.1037/a0016136] [Medline: 19916637]
68. Murray JM, Brennan SF, French DP, Patterson CC, Kee F, Hunter RF. Effectiveness of physical activity interventions in achieving behaviour change maintenance in young and middle aged adults: a systematic review and meta-analysis. Soc Sci Med 2017 Nov;192:125-133. [doi: 10.1016/j.socscimed.2017.09.021] [Medline: 28965003]
69. Knittle K, Nurmi J, Crutzen R, Hankonen N, Beattie M, Dombrowski SU. How can interventions increase motivation for physical activity? A systematic review and meta-analysis. Health Psychol Rev 2018 Sep 15;12(3):211-230. [doi: 10.1080/17437199.2018.1435299] [Medline: 29385950]
70. Howlett N, Trivedi D, Troop N, Chater A. Are physical activity interventions for healthy inactive adults effective in promoting behavior change and maintenance, and which behavior change techniques are effective? A systematic review and meta-analysis. Transl Behav Med 2019 Jan 01;9(1):147-157 [FREE Full text] [doi: 10.1093/tbm/iby010] [Medline: 29506209]
71. Ashton LM, Sharkey T, Whatnall MC, Williams RL, Bezzina A, Aguiar EJ, et al. Effectiveness of interventions and behaviour change techniques for improving dietary intake in young adults: a systematic review and meta-analysis of RCTs. Nutrients 2019 Apr 11;11(4):825 [FREE Full text] [doi: 10.3390/nu11040825] [Medline: 30979065]
72. Garnett C, Crane D, Brown J, Kaner E, Beyer F, Muirhead C, et al. Behavior change techniques used in digital behavior change interventions to reduce excessive alcohol consumption: a meta-regression. Ann Behav Med 2018 May 18;52(6):530-543 [FREE Full text] [doi: 10.1093/abm/kax029] [Medline: 29788261]
73. McCrabb S, Baker A, Attia J, Skelton E, Twyman L, Palazzi K, et al. Internet-based programs incorporating behavior change techniques are associated with increased smoking cessation in the general population: a systematic review and meta-analysis. Ann Behav Med 2019 Feb 01;53(2):180-195. [doi: 10.1093/abm/kay026] [Medline: 29750240]
74. Ek A, Alexandrou C, Söderström E, Bergman P, Delisle Nyström C, Direito A, et al. Effectiveness of a 3-month mobile phone-based behavior change program on active transportation and physical activity in adults: randomized controlled trial. JMIR Mhealth Uhealth 2020 Jun 08;8(6):e18531 [FREE Full text] [doi: 10.2196/18531] [Medline: 32510462]
75. Newby K, Teah G, Cooke R, Li X, Brown K, Salisbury-Finch B, et al. Do automated digital health behaviour change interventions have a positive effect on self-efficacy? A systematic review and meta-analysis. Health Psychol Rev 2021 Mar 20;15(1):140-158. [doi: 10.1080/17437199.2019.1705873] [Medline: 31847702]
76. Müssener U, Bendtsen M, Karlsson N, White IR, McCambridge J, Bendtsen P. SMS-based smoking cessation intervention among university students: study protocol for a randomised controlled trial (NEXit trial). Trials 2015 Apr 08;16:140 [FREE Full text] [doi: 10.1186/s13063-015-0640-2] [Medline: 25872503]
77. Thomas K, Bendtsen M, Linderoth C, Karlsson N, Bendtsen P, Müssener U. Short message service (SMS)-based intervention targeting alcohol consumption among university students: study protocol of a randomized controlled trial. Trials 2017 Apr 04;18(1):156 [FREE Full text] [doi: 10.1186/s13063-017-1898-3] [Medline: 28372563]
78. Müssener U, Bendtsen M, McCambridge J, Bendtsen P. User satisfaction with the structure and content of the NEXit intervention, a text messaging-based smoking cessation programme. BMC Public Health 2016 Nov 22;16(1):1179 [FREE Full text] [doi: 10.1186/s12889-016-3848-5] [Medline: 27876031]
79. Müssener U, Thomas K, Linderoth C, Leijon M, Bendtsen M. A text message-based intervention targeting alcohol consumption among university students: user satisfaction and acceptability study. JMIR Hum Factors 2018 Jul 10;5(3):e23 [FREE Full text] [doi: 10.2196/humanfactors.9641] [Medline: 29991469]

XSL•FO
RenderX

80. Thomas K, Linderoth C, Bendtsen M, Bendtsen P, Müssener U. Text message-based intervention targeting alcohol consumption among university students: findings from a formative development study. JMIR Mhealth Uhealth 2016 Oct 20;4(4):e119 [FREE Full text] [doi: 10.2196/mhealth.5863] [Medline: 27765732]
81. Kaner E, Beyer F, Garnett C, Crane D, Brown J, Muirhead C, et al. Personalised digital interventions for reducing hazardous and harmful alcohol consumption in community-dwelling populations. Cochrane Database Syst Rev 2017 Sep 25;9(9):CD011479 [FREE Full text] [doi: 10.1002/14651858.CD011479.pub2] [Medline: 28944453]
82. Black N, Johnston M, Michie S, Hartmann-Boyce J, West R, Viechtbauer W, et al. Behaviour change techniques associated with smoking cessation in intervention and comparator groups of randomized controlled trials: a systematic review and meta-regression. Addiction 2020 Nov 16;115(11):2008-2020. [doi: 10.1111/add.15056] [Medline: 32196796]
83. Crane RS, Brewer J, Feldman C, Kabat-Zinn J, Santorelli S, Williams JMG, et al. What defines mindfulness-based programs? The warp and the weft. Psychol Med 2016 Dec 29;47(6):990-999. [doi: 10.1017/s0033291716003317]
84. Creswell JD. Mindfulness interventions. Annu Rev Psychol 2017 Jan 03;68(1):491-516. [doi: 10.1146/annurev-psych-042716-051139] [Medline: 27687118]
85. Wong SYS, Chan JYC, Zhang D, Lee EKP, Tsoi KKF. The safety of mindfulness-based interventions: a systematic review of randomized controlled trials. Mindfulness 2018 Feb 2;9(5):1344-1357. [doi: 10.1007/s12671-018-0897-0]
86. Galante J, Friedrich C, Dawson AF, Modrego-Alarcón M, Gebbing P, Delgado-Suárez I, et al. Mindfulness-based programmes for mental health promotion in adults in nonclinical settings: a systematic review and meta-analysis of randomised controlled trials. PLoS Med 2021 Jan 11;18(1):e1003481 [FREE Full text] [doi: 10.1371/journal.pmed.1003481] [Medline: 33428616]
87. Neff KD, Germer CK. A pilot study and randomized controlled trial of the mindful self-compassion program. J Clin Psychol 2013 Jan 15;69(1):28-44. [doi: 10.1002/jclp.21923] [Medline: 23070875]
88. Crawford J, Collier E, Bendtsen M. Individualized treatment effects of a digital alcohol intervention and their associations with participant characteristics and engagement. Alcohol Alcohol 2024 Jul 21;59(5):agae049 [FREE Full text] [doi: 10.1093/alcalc/agae049] [Medline: 39034147]
89. Rehm J. Measuring quantity, frequency, and volume of drinking. Alcohol Clin Exp Res 1998 Apr;22(2 Suppl):4S-14S. [doi: 10.1097/00000374-199802001-00002] [Medline: 9603301]
90. Shorter GW, Bray JW, Giles EL, O'Donnell AJ, Berman AH, Holloway A, et al. The variability of outcomes used in efficacy and effectiveness trials of alcohol brief interventions: a systematic review. J Stud Alcohol Drugs 2019 May;80(3):286-298. [doi: 10.15288/jsad.2019.80.286]
91. Shorter GW, Bray JW, Heather N, Berman AH, Giles EL, Clarke M, et al. The "Outcome Reporting in Brief Intervention Trials: Alcohol" (ORBITAL) core outcome set: international consensus on outcomes to measure in efficacy and effectiveness trials of alcohol brief interventions. J Stud Alcohol Drugs 2021 Sep;82(5):638-646 [FREE Full text] [doi: 10.15288/jsad.2021.82.638] [Medline: 34546911]
92. Swedish National Board of Health and Welfare. Swedish National Guidelines for Prevention and Treatment of Unhealthy Lifestyle Behaviours [Nationella Riktlinjer: Prevention och Behandling vid Ohälsosamma Levnadsvanor]. Stockholm, Sweden: Swedish National Board of Health and Welfare; 2018. URL: https://www.socialstyrelsen.se/globalassets/sharepoint-dokument/artikelkatalog/nationella-riktlinjer/2018-6-24.pdf [accessed 2026-02-24]
93. SRNT Subcommittee on Biochemical Verification. Biochemical verification of tobacco use and cessation. Nicotine Tob Res 2002 May 1;4(2):149-159. [doi: 10.1080/14622200210123581] [Medline: 12028847]
94. Vallejo MA, Vallejo-Slocker L, Fernández-Abascal EG, Mañanes G. Determining factors for stress perception assessed with the Perceived Stress Scale (PSS-4) in Spanish and other European samples. Front Psychol 2018;9:37 [FREE Full text] [doi: 10.3389/fpsyg.2018.00037] [Medline: 29434563]
95. Bendtsen M. A Gentle Introduction to the Comparison Between Null Hypothesis Testing and Bayesian Analysis: Reanalysis of Two Randomized Controlled Trials. J Med Internet Res 2018 Oct 24;20(10):e10873. [doi: 10.2196/10873]
96. Bendtsen M. Avoiding under- and overrecruitment in behavioral intervention trials using Bayesian sequential designs: tutorial. J Med Internet Res 2022 Dec 16;24(12):e40730 [FREE Full text] [doi: 10.2196/40730] [Medline: 36525297]
97. Crane D, Garnett C, Michie S, West R, Brown J. A smartphone app to reduce excessive alcohol consumption: identifying the effectiveness of intervention components in a factorial randomised control trial. Sci Rep 2018 Mar 12;8(1):4384 [FREE Full text] [doi: 10.1038/s41598-018-22420-8] [Medline: 29531280]
98. Schroé H, Van Dyck D, De Paepe A, Poppe L, Loh WW, Verloigne M, et al. Which behaviour change techniques are effective to promote physical activity and reduce sedentary behaviour in adults: a factorial randomized trial of an e- and m-health intervention. Int J Behav Nutr Phys Act 2020 Oct 07;17(1):127 [FREE Full text] [doi: 10.1186/s12966-020-01001-x] [Medline: 33028335]
99. Fanning J, Roberts S, Hillman CH, Mullen SP, Ritterband L, McAuley E. A smartphone "app"-delivered randomized factorial trial targeting physical activity in adults. J Behav Med 2017 Oct;40(5):712-729. [doi: 10.1007/s10865-017-9838-y] [Medline: 28255750]
100. Brownbill AL, Braunack-Mayer AJ, Miller CL. What makes a beverage healthy? A qualitative study of young adults' conceptualisation of sugar-containing beverage healthfulness. Appetite 2020 Jul 01;150:104675. [doi: 10.1016/j.appet.2020.104675] [Medline: 32184093]

XSL•FO
RenderX

101. Cho H, Salmon C. Unintended effects of health communication campaigns. Journal of Communication 2006;57(2):317. [doi: 10.1111/j.1460-2466.2007.00344.x]
102. Champion KE, Newton NC, Gardner LA, Chapman C, Thornton L, Slade T, et al. Health4Life eHealth intervention to modify multiple lifestyle risk behaviours among adolescent students in Australia: a cluster-randomised controlled trial. The Lancet Digital Health 2023 May;5(5):e276-e287. [doi: 10.1016/s2589-7500(23)00028-6]
103. Epton T, Norman P, Dadzie AS, Harris PR, Webb TL, Sheeran P, et al. A theory-based online health behaviour intervention for new university students (U@Uni): results from a randomised controlled trial. BMC Public Health 2014 Jun 05;14:563 [FREE Full text] [doi: 10.1186/1471-2458-14-563] [Medline: 24903620]
104. Duan Y, Liang W, Wang Y, Lippke S, Lin Z, Shang B, et al. The effectiveness of sequentially delivered web-based interventions on promoting physical activity and fruit-vegetable consumption among Chinese college students: mixed methods study. J Med Internet Res 2022 Jan 26;24(1):e30566 [FREE Full text] [doi: 10.2196/30566] [Medline: 35080497]
105. Bendtsen M, Åsberg K, McCambridge J. Mediators of effects of a digital alcohol intervention for online help-seekers: findings from an effectiveness trial. Drug Alcohol Depend 2023 Oct 01;251:110957 [FREE Full text] [doi: 10.1016/j.drugalcdep.2023.110957] [Medline: 37688979]
106. Crawford J, Blomqvist J, Gunnarsson KU, Bendtsen P, Bendtsen M. Mediated effects of a randomised control trial for a text messaging smoking cessation intervention for online help-seekers and primary care visitors. BMC Public Health 2024 Jul 09;24(1):1824 [FREE Full text] [doi: 10.1186/s12889-024-19273-4] [Medline: 38977972]
107. Gold N, Yau A, Rigby B, Dyke C, Remfry EA, Chadborn T. Effectiveness of digital interventions for reducing behavioral risks of cardiovascular disease in nonclinical adult populations: systematic review of reviews. J Med Internet Res 2021 May 14;23(5):e19688 [FREE Full text] [doi: 10.2196/19688] [Medline: 33988126]
108. MacNeill V, Foley M, Quirk A, McCambridge J. Shedding light on research participation effects in behaviour change trials: a qualitative study examining research participant experiences. BMC Public Health 2016 Jan 29;16:91 [FREE Full text] [doi: 10.1186/s12889-016-2741-6] [Medline: 26825583]
109. Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. Lancet 2002 Feb 23;359(9307):696-700. [doi: 10.1016/S0140-6736(02)07816-9] [Medline: 11879884]

Abbreviations

BCT: behavior change technique
CoI: compatibility interval
CONSORT: Consolidated Standards of Reporting Trials
IRR: incidence rate ratio
ISRCTN: International Standard Randomised Controlled Trial Number
MVPA: moderate-to-vigorous physical activity
POE: probability of effect

Edited by A Stone; submitted 03.Dec.2025; peer-reviewed by S Ondersma; comments to author 05.Jan.2026; revised version received 16.Feb.2026; accepted 17.Feb.2026; published 09.Mar.2026.

Please cite as:
Åsberg K, Lundgren O, Henriksson H, Henriksson P, Eldh AC, Bendtsen P, Löf M, Bendtsen M
Effectiveness of the Components of a Digital Multiple Health Behavior Intervention Among University Students (Buddy): Factorial Randomized Trial
J Med Internet Res 2026;28:e88884
URL: https://www.jmir.org/2026/1/e88884
doi:10.2196/88884
PMID:

©Katarina Åsberg, Oskar Lundgren, Hanna Henriksson, Pontus Henriksson, Ann Catrine Eldh, Preben Bendtsen, Marie Löf, Marcus Bendtsen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Telepathology and Mobile Health System for Province-Wide Pathology Consultation in Henan, China: Retrospective Evaluation Study

Jinming Shi[1*], MEng; Ming Ye[1*], MEng; Dongxu Sun[1*], MEng; Xianying He[1], MSc; Yaoen Lu[1], MEng; Linlin Wang[1], MSc; Haotian Chen[1], MEng; Wenchao Wang[1], MEng; Jie Zhao[1,2], PhD; Fangfang Cui[1,3], MSc

[1]National Engineering Laboratory for Internet Medical Systems and Applications, First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan, China

[2]Shanghai Artificial Intelligence Laboratory, Shanghai, China

[3]Institute of Intelligent Medicine, Henan Academy of Innovations in Medical Sciences, Zhengzhou, Henan, China

[*]these authors contributed equally

Corresponding Author:
Fangfang Cui, MSc
National Engineering Laboratory for Internet Medical Systems and Applications
First Affiliated Hospital of Zhengzhou University
No. 3, Jianshe East Road
Zhengzhou, Henan, 450000
China
Phone: 86 15093159748
Email: fcccuiff@zzu.edu.cn

Abstract

Background: Telepathology has emerged as a transformative digital health solution to address the global shortage of pathologists and the unequal distribution of diagnostic services, particularly in underserved and rural areas. In Henan Province, China, high diagnostic demand, rapid population growth, and limited pathology expertise exacerbate regional health care inequities, leading to delayed diagnoses and restricted access to specialist care.

Objective: This study aimed to design, implement, and evaluate a province-wide telepathology system integrating web and mobile platforms to enhance diagnostic quality, efficiency, and equitable access across health care tiers.

Methods: We conducted a retrospective, multicenter observational study using deidentified data from 120 health care institutions between 2016 and 2024. The system used a 3-tier architecture with virtual private network–secured transmission and a Browser-Server framework, supporting standardized whole-slide image acquisition, remote review, and reporting via web interfaces and a WeChat (Tencent) mini-program. System performance was assessed by consultation volume, turnaround time, concurrency, and diagnostic concordance in a subset of 1027 cases with paired tertiary-hospital expert diagnoses. Economic impact was estimated using previously published per-case savings, reflecting patient travel and ancillary cost reductions. Additional assessments included workflow integration, mobile platform use, and system stability under peak load.

Results: Over 8 years, the network processed 72,916 consultations encompassing 355,104 whole-slide images, supporting 220-300 concurrent users with stable performance. Median turnaround time was 10.06 (IQR 1.63-29.10) hours, with 96.41% (70,298/72,916) of cases completed within 72 hours. County-level hospitals contributed 77.63% (56,603/72,916) of consultations, demonstrating substantial engagement from lower-tier institutions. In the diagnostic subset, originating-site preliminary classifications achieved 0.90 sensitivity and 0.75 specificity relative to expert reference diagnoses, with 17.2% discordance corrected through remote expert review. Estimated annual direct cost savings ranged from US $0.14 to $0.63 million. Mobile-enabled access facilitated remote review and reporting without compromising data security, supporting integration into routine clinical workflows across diverse hospital settings.

Conclusions: The Henan Province telepathology system demonstrates that a centrally coordinated, scalable digital health platform can improve diagnostic efficiency, quality, and equity in resource-constrained settings. High county-level hospital use highlights its potential to reduce geographic and structural diagnostic inequities. Future work should explore formal cost-effectiveness evaluation, artificial intelligence–assisted diagnostic support, and cross-regional interoperability to enable broader adoption and sustainable integration into health care systems.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e75172) doi:10.2196/75172

KEYWORDS

telepathology; diagnostic equity; digital pathology; remote diagnostics; rural health; health systems innovation

Introduction

The digitization of pathology workflows has revolutionized diagnostic practices. High-throughput slide scanning and standardized data management have enabled unprecedented improvements in service accessibility and interinstitutional collaboration [1]. Digital pathology (DP) offers multiple benefits, including enhanced diagnostic accuracy, faster turnaround times (TAT), and increased efficiency in pathology workflows [2-5]. However, translating these technological advancements into equitable real-world health care delivery remains constrained by 2 systemic barriers: the geographic concentration of pathology expertise and the lack of scalable infrastructure to support cross-regional coordination [6]. In Henan Province, China, a region with a population exceeding 100 million and reporting approximately 279,740 new cancer cases annually [7], these challenges are particularly pronounced. The province faces a critical shortage of pathologists, who constitute only 0.4% of the medical workforce [8,9]. This shortage leads to significant delays in diagnosis, especially in rural and underserved areas, where patients often face prolonged wait times for pathology reports. These diagnostic delays directly contribute to adverse patient outcomes, including delayed treatment initiation and reduced survival rates, while exacerbating health care inefficiencies [7,10]. These diagnostic delays are symptomatic of deeper structural inequities in health care access between urban and rural populations.

Telepathology has emerged as a transformative solution to decentralize diagnostic services, offering a viable strategy to overcome these systemic barriers [11-14]. By leveraging digital slide scanning technologies [15], secure data transmission protocols [16], and remote diagnostic interfaces, telepathology enables cross-regional and cross-institutional collaboration among pathologists [17,18]. Numerous studies have demonstrated that telepathology systems can effectively reduce TAT, enhance diagnostic accuracy, and optimize resource distribution, making them valuable tools for both routine diagnostics and specialized consultations [19-24]. For instance, a telepathology diagnostic network established in Tanzania showed TAT performance comparable to conventional intraoperative consultations using optical microscopes [23]. Similarly, the Clinical Telepathology Service at the University of California, Los Angeles, demonstrated improved diagnostic efficiency and cost savings through its cloud-based infrastructure [25].

However, the scalability of telepathology systems remains a critical barrier to widespread adoption. A recent systematic review revealed that the majority of telepathology initiatives are limited to single institutions or pilot programs, with relatively few achieving province- or state-wide coverage [26]. This limitation arises from multiple technical and organizational challenges, including inconsistent standards for image acquisition, storage, and sharing, which hinder interoperability across regions and institutions [26]. Additionally, the large volumes of data generated by whole slide image (WSI) systems contribute to computational bottlenecks, complicating data processing and analysis [15,16]. The absence of tiered governance models, which are essential for coordinating multilevel health care systems, further constrains the effective management and deployment of telepathology solutions [27,28]. Moreover, without equitable deployment strategies, such innovations risk widening existing gaps in diagnostic access.

To address these challenges, this study presents the design, deployment, and evaluation of a province-wide telepathology system across 120 health care institutions in Henan Province, China. Enabled by dedicated funding from the Henan Provincial Health Commission, the system facilitated the deployment of DP scanners in county- and city-level hospitals. Led by the First Affiliated Hospital of Zhengzhou University, the largest medical institution in China, the system operates under a centralized framework for case management, expert consultation, and data integration, supported by a mobile-enabled interface to expand diagnostic access to remote and rural regions.

This study aims to (1) design and implement a scalable province-wide telepathology system to enhance equitable access to diagnostic services across health care tiers; (2) evaluate its performance in terms of diagnostic efficiency, accuracy, and economic impact (direct patient cost savings); (3) demonstrate the feasibility of large-scale telepathology expansion in resource-constrained settings; and (4) offer a scalable model that can be replicated in other health care systems facing similar challenges. These findings contribute to the growing body of evidence supporting digital health solutions in bridging health care gaps and optimizing pathology services worldwide.

Methods

Overview

This Methods section describes the study design, data sources, evaluation methods, and data protection measures used to assess a province-wide telepathology system.

Study Design

This study adopted a multicenter retrospective observational design to evaluate the real-world implementation and performance of a province-wide telepathology system across 120 health care institutions in Henan, China. Routinely collected, deidentified system data generated between January 2016 and June 2024 were analyzed.

The evaluation comprised three components aligned with the study aims: (1) technical performance, assessed using Apache Benchmark stress testing to measure response time and concurrent user capacity; (2) operational use and efficiency, evaluated descriptively using system-generated metrics including consultation volume, contributing hospital level, and diagnostic TAT; and (3) diagnostic accuracy, evaluated in a retrospective

XSL•FO
RenderX

subset of 1027 cases by comparing originating-site preliminary diagnoses with tertiary-hospital expert reference diagnoses recorded in the platform.

Data Collection and Sources

All analyses used routinely collected, deidentified data automatically recorded by the telepathology platform between January 2016 and June 2024. System logs and administrative exports were used to summarize consultation volume, contributing hospital level, annual trends, and TAT, defined as the interval from WSI submission to report issuance.

For frozen-section consultations, the request-to-diagnosis time was summarized based on real-time notification and reporting timestamps recorded in the platform. Patient and diagnostic characteristics (age, sex, benign vs malignant classification, and specimen anatomical site) were obtained from system-generated pathology reports. Platform metadata and storage reports were used to describe supported WSI formats, per-slide file size range, and cumulative storage volume.

Diagnostic Accuracy Assessment

Diagnostic accuracy was evaluated in a retrospective subset of 1027 telepathology consultations that had both (1) an originating-site preliminary diagnosis recorded at submission and (2) a corresponding tertiary-hospital expert diagnosis recorded in the platform for the same case. Cases with missing or noncomparable diagnostic categories were excluded. Diagnoses were dichotomized as malignant (positive) versus non-malignant (negative) according to the recorded report impression. The tertiary-hospital expert diagnosis served as the reference standard. Sensitivity, specificity, and overall agreement were calculated using a confusion matrix.

Economic Impact Estimation

Per-case cost savings were not recalculated from primary economic data in this study. Instead, we adopted the previously published per-case savings estimate for the same provincial telepathology program (378.5 RMB per case; US $1=6.46 RMB) reported by He et al [10], which was based on local consultation fees and assumed travel and food expenses avoided when patients no longer needed to travel to tertiary hospitals for expert consultation. To derive an annual estimate, we multiplied the per-case savings (378.5 RMB; US $1=6.46 RMB) by the number of completed telepathology consultations in each calendar year (Table 1). All RMB values were converted to US dollars using the exchange rate reported in He et al [10].

Table 1. Annual volume of completed telepathology consultations by hospital level in a province-wide observational study conducted in Henan Province, China.

Hospital level	2016	2017	2018	2019	2020	2021	2022	2023	2024
City-level	240	657	1060	923	2100	3503	2970	2439	2421
County-level	2095	3673	6202	8317	7490	6536	6198	8267	7825

System Architecture Design

To address the shortage of pathologists and the uneven distribution of diagnostic resources in Henan Province, China, we designed an integrated province-wide telepathology system optimized for scalability, diagnostic accuracy, and operational efficiency. The system follows a 3-tiered hospital architecture, consisting of tertiary-level hubs (provincial hospitals), secondary-level nodes (city-level hospitals), and primary-level health care institutions (county-level hospitals). This hierarchical structure enables standardized workflows for digital slide acquisition, data transmission, remote diagnosis, and reporting across 120 health care institutions, ensuring equitable access to high-quality pathology services.

The system adopts a Browser-Server architecture, chosen for its flexibility, ease of deployment, and compatibility with diverse IT infrastructures in both urban and rural health care settings. To support real-time access in resource-constrained environments, we incorporated both web-based platforms and mobile access via a WeChat mini-program. This mobile interface was particularly critical for ensuring remote diagnostic access in rural areas where conventional IT infrastructure is limited. The mini-program was selected due to its widespread adoption in China and seamless integration with existing health care communication workflows, eliminating the need for standalone app installations. The system consists of 4 core modules: image acquisition module, transmission link module, data storage module, and teleconsultation platform module. These modules are tightly integrated to support high-throughput data processing, secure transmission, and efficient telepathology consultations (Figure 1).

XSL•FO
RenderX

Figure 1. Architecture of the province-wide telepathology system evaluated in a retrospective multicenter observational study across 120 health care institutions in Henan Province, China.

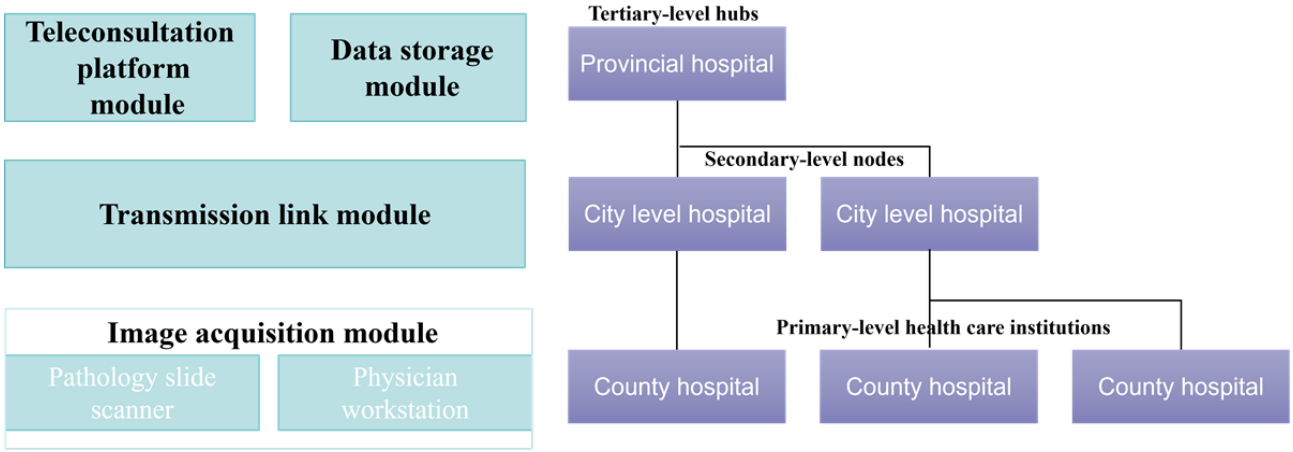

Image Acquisition Module

This module digitizes traditional pathology slides into high-resolution WSIs and prepares them for secure transmission. Each participating hospital is equipped with (1) pathology slide scanner Model KF-PRO-005 (Konfoong Biotech International Co., Ltd.), capable of scanning slides within 40 seconds at 0.47 µm per pixel (20 magnification) and within 100 seconds at 0.5 µm per pixel (40 magnification); and (2) physician workstation powered by Intel Core i5 processors with 8 GB RAM, preloaded with K-Scanner 1.6.0.14 and K-Viewer 1.5.3.1 (Konfoong Bioinformation Tech Co, Ltd., Ningbo, China) for scanner control, image browsing, and management.

The scanned WSIs are processed using a multiresolution pyramid structure, optimizing image loading speed and reducing data size without compromising diagnostic quality.

In addition, a standardized scanning workflow and a 2-step image quality control (QC) procedure were applied across participating sites to ensure that WSIs met diagnostic requirements. Before scanning, slides were checked for correct labeling and basic preparation adequacy, and scanners were routinely calibrated according to the manufacturer's recommendations. After scanning, technicians performed an on-screen QC review to confirm tissue coverage completeness, focus adequacy across the tissue area, and absence of major scanning artifacts (eg, excessive blur, stripe or banding, or large occlusions), as well as reasonable stain and color appearance. Slides failing QC were rescanned, and if persistent issues were attributed to slide preparation, the originating laboratory was notified for recutting or restaining. A second, preanalytical QC check was conducted by triage personnel during case allocation; cases with insufficient image quality were returned for resubmission prior to expert review.

Transmission Link Module

The transmission link module ensures that digitized pathology images can be transmitted from the applying hospital to higher-level hospitals. Considering data security and transmission rates, the system uses the 5G fixed-mobile convergence medical private network, previously established by the Henan Telemedicine Center of China and covering Henan Province, as the transmission link. Digitized pathology slides are scanned and uploaded to the central server for storage and diagnosis through this private network. The architecture of this private network is illustrated in Figure 2.

The telemedicine private network can be classified into wired transmission networks, 5G telemedical transmission networks, and Internet of Things transmission networks. The wired network is primarily used in fixed office scenarios and easily accessible medical institutions, such as large hospitals, for teleconsultations, telepathology, remote education, and applications in medical institutions at the county level and above. The 5G telemedical transmission network mainly supports dynamic scenarios, off-hospital scenes, and the access of medical institutions in remote areas, including remote ward rounds, remote emergency services, remote electrocardiography, remote ultrasound, and access for village and town health clinics. The construction and application of this private network provide effective network support for the transmission of remote pathological slides.

XSL•FO
RenderX

Figure 2. Architecture of the medical private network supporting province-wide telepathology consultations in Henan Province, China.

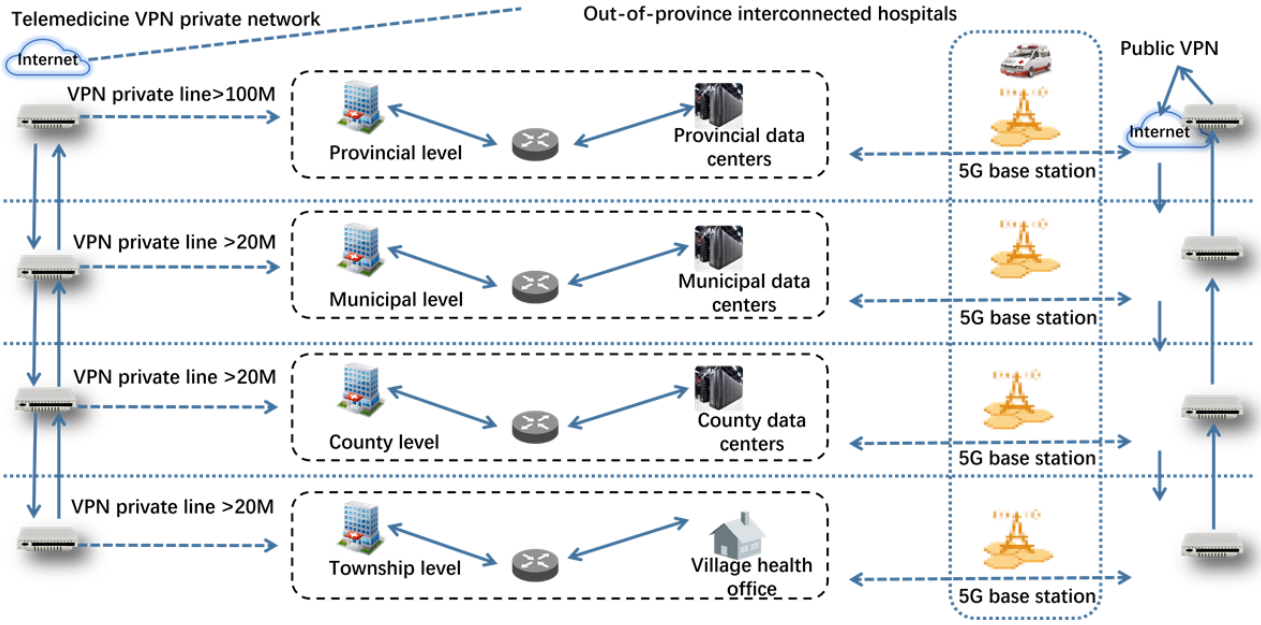

Data Storage Module

The data storage module is designed to ensure the secure, scalable, and efficient storage of WSIs, diagnostic reports, and patient records. To enhance storage efficiency and fault tolerance, we use erasure coding technology, which works by dividing data into multiple fragments and generating parity blocks distributed across different storage nodes. In the event of data corruption or hardware failure, erasure coding enables data reconstruction from the remaining fragments, ensuring high availability and data integrity with minimal storage overhead [29].

For data backup, the system adopts a master-slave data synchronization strategy, implementing asynchronous replication to ensure the consistency and integrity of data. Additionally, binary log (Binlog) technology is used to record all database changes, allowing for rapid data recovery in the event of a system failure. The backup mechanism uses a mixed-based replication model, combining the advantages of statement-level and row-based replication, ensuring efficient and stable data backups.

Teleconsultation Platform Module

Overview

The teleconsultation platform is the core of the system, supporting remote pathology consultations, diagnostic reporting, and interinstitutional case discussions. The software architecture (Figure 3) is organized into 4 layers.

XSL•FO
RenderX

Figure 3. Software architecture of the teleconsultation platform used for province-wide telepathology services in Henan Province, China. AI: artificial intelligence; MQ: message queue; VPN: virtual private network.

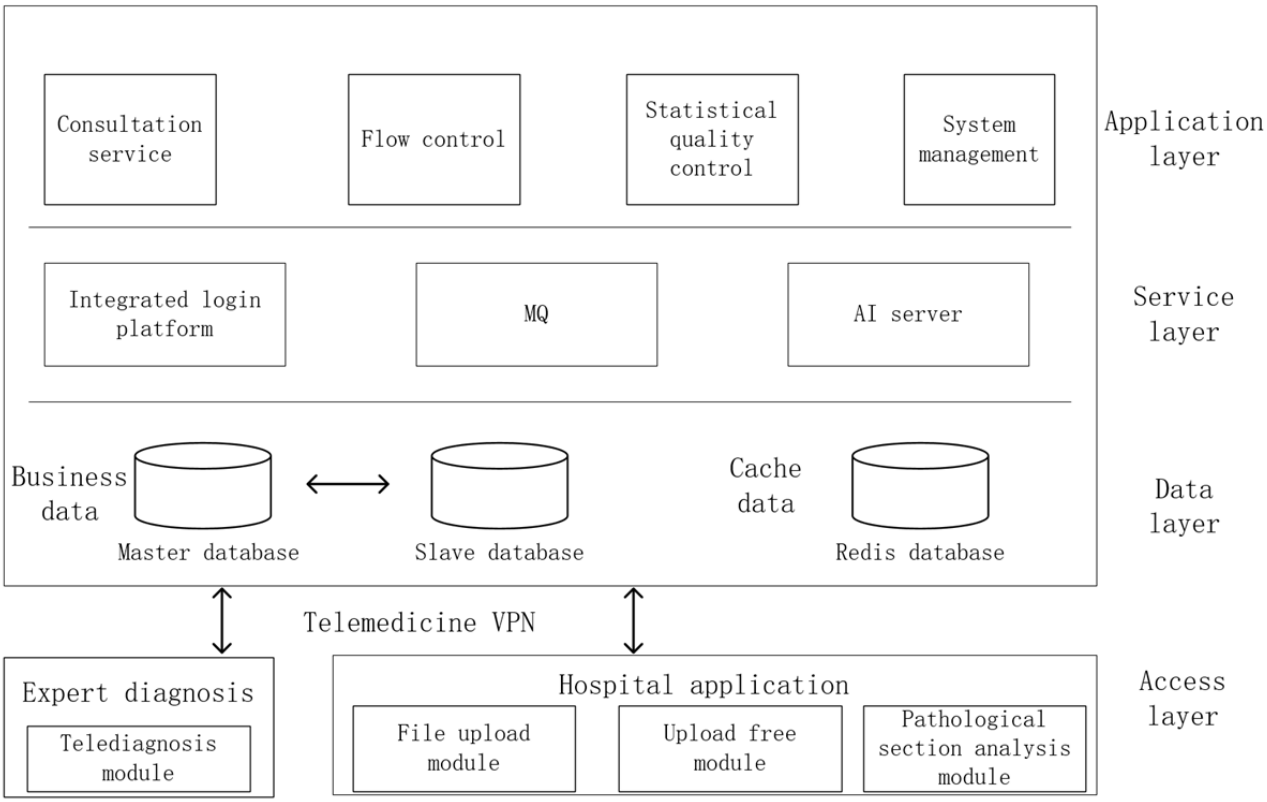

Access Layer

The access layer connects requesting hospitals with expert terminals, enabling seamless interaction and workflow integration. Within this layer, the upload free module eliminates the need for manual file uploads by integrating directly with local DP scanners, ensuring automatic and efficient transmission of WSIs. Additionally, the Pathological Section Analysis Module performs preliminary analysis, including image quality assessment, segmentation, and lesion detection, aiding triage personnel in prioritizing cases. To optimize performance, WSI images undergo segmentation, compression, and pyramid layer processing, allowing for rapid loading and smooth scaling of high-definition images. Experts can access, analyze, and report cases either through a web browser or via the WeChat mini-program, enhancing flexibility and accessibility.

Data Layer

The data layer is responsible for managing both business-related data and cache data to ensure system stability and efficiency. Business data is stored within a MySQL relational database, maintaining structured and secure data management, while cache data, such as user authentication tokens and system configuration information, is stored using Redis, an in-memory database designed to facilitate high-speed data access and enhance system performance.

Service Layer

The service layer provides the system's core functionalities, encompassing consultation process management, diagnostic reporting, and QC statistics. To ensure seamless communication between different modules, message queue technology is implemented, allowing for asynchronous data transmission. This enables efficient handling of high-volume concurrent requests, ensuring that teleconsultation processes remain reliable and scalable without system bottlenecks.

Application Layer

The application layer provides various application functionalities and permissions tailored for triage personnel, system administrators, primary hospital doctors, and others, catering to their specific roles and access levels.

Data protection and privacy. To support province-wide deployment and mobile access, the platform implemented the following safeguards. The WeChat mini-program served only as a mobile access interface. WSIs and clinical records remained in the telemedicine center's private cloud and were accessed through the secured telepathology platform via a dedicated application programming interface gateway (HTTPS with Transport Layer Security encryption, authentication, role-based access control, and an allowlist of permitted application programming interfaces). Access events were logged for audit and monitored with rate limiting when needed. The WeChat-based support chat group was limited to technical troubleshooting and operational coordination; no patient-identifiable information or diagnostic images were shared, and all clinical case review and reporting were performed within the secured telepathology platform.

XSL•FO
RenderX

System Functionality Overview

Overview

During the functional design phase of the system, a multidisciplinary design discussion group was established to ensure that the system effectively met the needs of different stakeholders. This group consisted of 6 members, including 1 pathologist from the First Affiliated Hospital of Zhengzhou University, representing the requirements of tertiary hospitals, 2 pathologists from primary health care institutions who articulated user needs from a clinical perspective, 1 project manager overseeing system development, and 2 software developers responsible for technical implementation. To validate the system's diagnostic reliability, a clinical study was conducted in which pathologists compared WSI-based diagnoses with conventional glass slide diagnoses across various case types. Expert feedback was collected on image interpretability, usability, and diagnostic confidence using both the browser-based platform and the WeChat mini-program. The results demonstrated a high concordance between digital and conventional diagnoses, confirming the system's suitability for clinical application.

Based on functional requirements, users were categorized into 4 distinct roles within the telepathology framework, ensuring an optimized workflow that facilitates seamless collaboration between requesting personnel, triage personnel, diagnostic experts, and system administrators (Figure 4).

Figure 4. Use case diagram illustrating roles and interactions within the province-wide telepathology system in Henan Province, China.

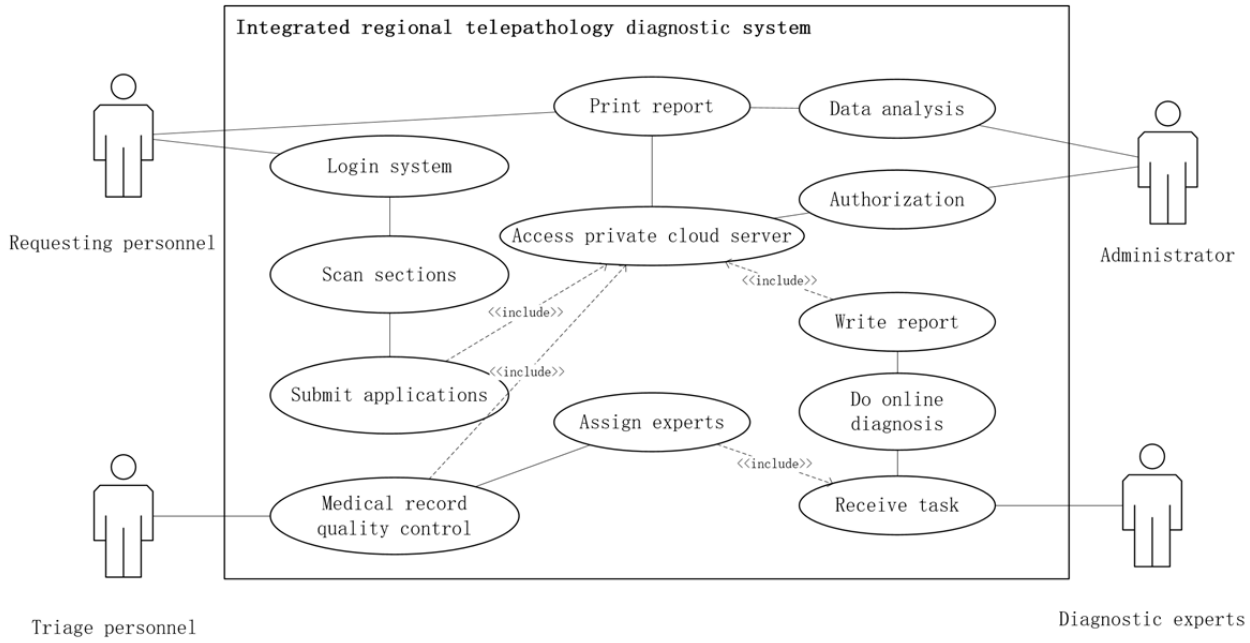

Requesting Personnel

Primarily from county-level (primary) and city-level (secondary) hospitals, initiate telepathology processes by digitizing and scanning pathology slides, completing patient medical records, and submitting telepathology requests to higher-level institutions. This decentralized design enabled even low-resource county hospitals to initiate telepathology consultations independently, contributing to more equitable diagnostic access.

Triage Personnel

As pathologists within city-level (secondary) and provincial-level (tertiary) hospitals, they are responsible for preanalytical QC and case allocation. By leveraging their in-depth knowledge of their hospital's pathology experts and their areas of specialization, they ensure that each case is assigned to the most suitable specialist within the institution.

Diagnostic Experts

Consisting of pathologists within city-level and provincial-level hospitals, analyze the assigned digital slides and generate diagnostic reports, either providing a direct diagnosis or escalating particularly complex cases to a higher-level institution when necessary.

Administrators

They oversee the integrated telepathology system, managing user roles and permissions, ensuring data security, and maintaining system functionality to guarantee a seamless telepathology workflow.

Operational Workflow and User Interaction

Overview

The integrated telepathology system is designed to streamline the end-to-end diagnostic workflow, encompassing 3 primary functionalities: application initiation, triage, and diagnostic review. These functionalities are supported by an intuitive user interface that facilitates efficient interaction between primary health care providers, triage coordinators, and diagnostic experts. This section provides a detailed description of the system's operational flow, user roles, and key technical features.

XSL•FO
RenderX

Application Initiation Workflow

The initiation of a consultation application is a core function performed by requesting personnel at primary health care institutions. This process involves several key steps: logging into the system, scanning pathology slides, and completing the consultation application form. The output of this function is the transmission of consultation application data via the virtual private network (VPN) to the telemedical private cloud server at the Henan Telemedicine Center of China. This information is then made accessible to triage personnel and diagnostic experts at tertiary medical institutions for review and further processing. The use case diagram for this function is illustrated in Figure 5.

Figure 5. Workflow for telepathology consultation request submission in a province-wide telepathology network in Henan Province, China.

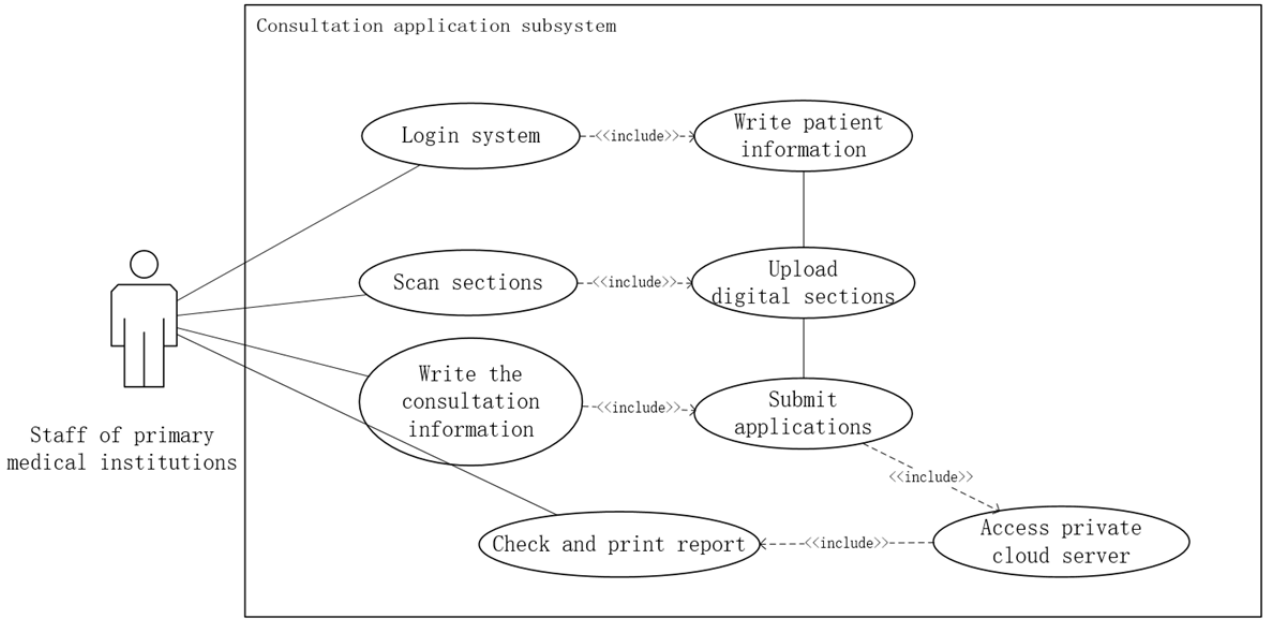

Triage Workflow and QC

The triage function is performed by pathologists serving as triage personnel within tertiary medical institutions, playing a pivotal role in ensuring the integrity of the diagnostic process. This function involves conducting preanalytical QC on received telepathology applications, which includes the following:

1. Assessing the adequacy of scanned WSIs for diagnostic interpretation, including sufficient tissue coverage, effective resolution, and focus across the tissue area, and absence of major artifacts (eg, excessive blur, stripe or banding, or large occlusions) that may compromise interpretation; staining or color issues were assessed in terms of image readability.
2. Verifying the completeness of clinical information, ensuring that relevant patient history, laboratory findings, and any necessary supporting documentation are provided to facilitate accurate diagnosis.

If the WSI quality is insufficient for diagnostic interpretation or the clinical information is incomplete, triage personnel may request additional data or resubmission (eg, rescanning, recutting, or restaining when preparation-related issues are suspected) before assigning the case to the most appropriate diagnostic expert based on the specific pathology subspecialty required. The workflow diagram illustrating this process is depicted in Figure 6.

When resubmission was required, the case was returned within the platform with standardized reasons, and the requesting site was notified through in-platform status updates and SMS alerts to registered staff; unresolved returns were escalated to the site administrator for follow-up.

XSL•FO
RenderX

Figure 6. Workflow for case triage and preanalytical quality control in the province-wide telepathology system in Henan Province, China. QC: quality control.

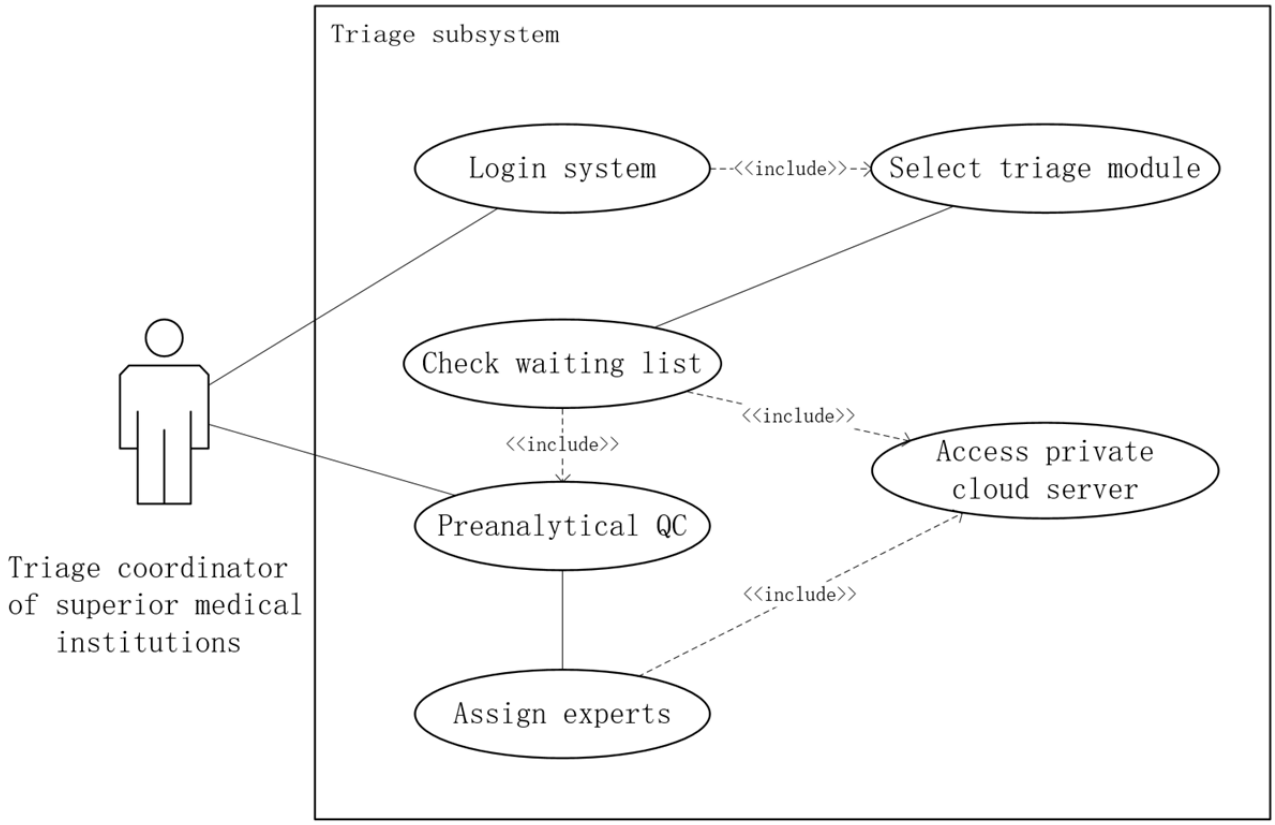

Diagnostic Function Design

The diagnostic function is executed by experts at tertiary hospitals, specifically pathologists, who access the system via a computer interface or the WeChat mini-program. Upon logging in, they review the assigned telepathology cases, analyzing DP slides and accompanying clinical data to provide a comprehensive pathological diagnosis. The diagnostic report includes a pathological description, grading, staging, results of immunohistochemistry, and additional relevant findings (Figure 7).

When an expert at a tertiary hospital determines that additional immunohistochemistry staining is required, they can submit an immunohistochemistry request through the system, specifying the recommended immunohistochemistry markers for further testing. This request is automatically recorded in the system and made accessible to the originating hospital. The originating hospital performs the immunohistochemistry staining locally, digitizes the stained slides, and uploads them back to the original case record within the platform. The tertiary hospital expert then reviews the newly uploaded immunohistochemistry-stained slides, incorporating the findings into a supplementary diagnostic report to ensure a comprehensive pathology assessment. The diagnostic report is completed using a semistructured interface, where predefined sections (eg, “Microscopic Findings,” “Diagnosis,” and “Comments”) guide report writing, while allowing free-text input to maintain clinical flexibility.

XSL•FO
RenderX

Figure 7. Workflow for remote pathology diagnosis and reporting within the province-wide telepathology system in Henan Province, China.

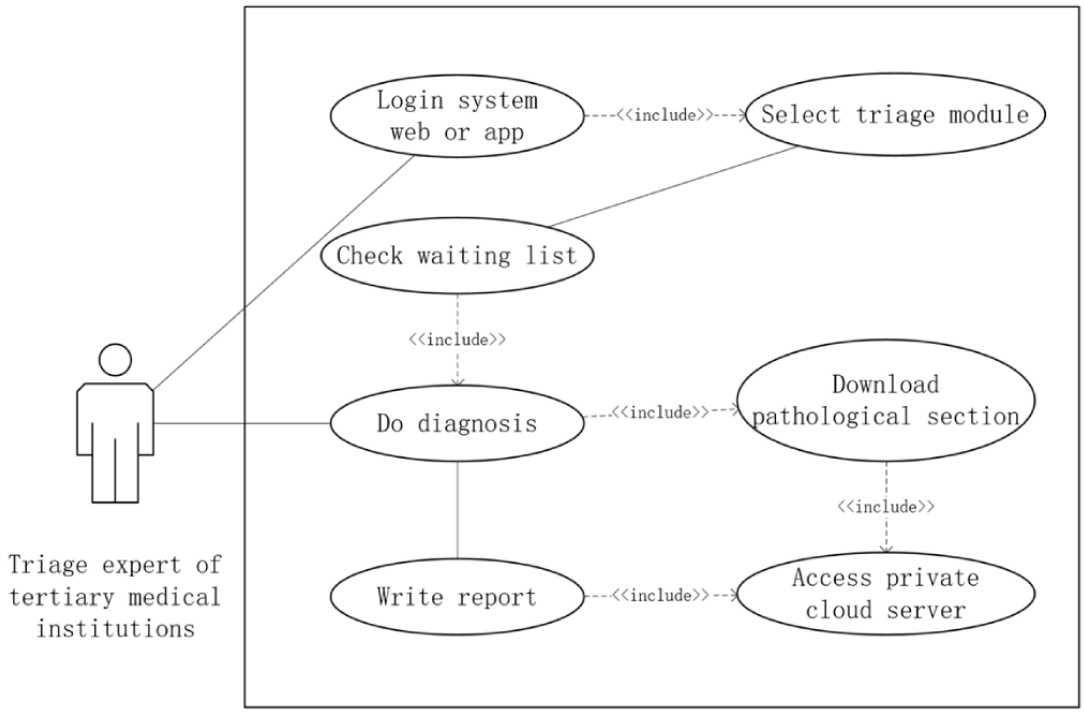

Ethical Considerations

This retrospective multicenter study was approved by the Ethics Committee of the First Affiliated Hospital of Zhengzhou University (approval number 2025-KY-1532-002). The requirement for informed consent was waived because the study involved secondary analysis of deidentified routine telepathology data without direct participant contact.

All data were anonymized prior to analysis, and no personally identifiable information was accessible to the research team. Data were stored on secure institutional servers with restricted access in accordance with applicable data protection regulations. No compensation was provided, as this study involved retrospective analysis of existing data. All procedures adhered to institutional policies and the principles of the Declaration of Helsinki.

Results

System Deployment and Operational Integration

The telepathology system was successfully deployed within the Henan Telemedicine Center's private cloud infrastructure, leveraging a secure VPN to facilitate encrypted data transmission across 120 integrated health care institutions at the provincial, municipal, and county levels. This architecture ensures robust data security, transmission stability, and interoperability, which are critical for maintaining diagnostic accuracy during remote consultations.

The deployment of DP scanners followed a stepwise approach to ensure systematic integration, validation, and training across hospitals. In 2015, the Henan Provincial Health Commission initiated a province-wide infrastructure assessment to evaluate hospital readiness for telepathology implementation. The procurement and phased installation of 120 scanners began in 2016, prioritizing tertiary hospitals first, followed by city and county-level hospitals, based on workload demand and digital infrastructure readiness. Pathologists and technicians were trained in scanner operation, QC, and troubleshooting. By 2017, scanners were fully integrated into routine workflows, with ongoing monitoring and performance optimization to maintain consistency in image acquisition.

To accommodate various clinical scenarios, the system supports both web-based interfaces and mobile access via a WeChat mini-program. The mobile integration significantly enhances accessibility, particularly in rural and resource-limited regions, where mobile technology adoption is high. The user interface was optimized for both platforms to streamline diagnostic workflows, reduce user friction, and improve the TAT for pathology consultations (Figure 8).

Since deployment, the system has facilitated 72,916 telepathology consultations, averaging approximately 22 cases per day. A total of 355,104 WSIs have been processed. The cumulative data volume has reached 220.1 TB, with each slide averaging 0.62 GB. However, slide storage size varies depending on scanning resolution, typically ranging from 500 MB to 2 GB, with 20× scans occupying 500 MB to 1 GB and 40× scans ranging from 1 GB to 2 GB. The system supports multiple whole-slide image formats, including Tagged Image File Format (TIFF) and NDPI (NanoZoomer format), as well as other vendor-specific formats that use pyramidal storage for efficient multiresolution viewing. Unlike Digital Imaging and Communications in Medicine (DICOM), which is primarily used for radiology, these formats support multilayered, ultra-high-resolution whole-slide imaging, ensuring optimal fidelity for pathological assessment.

The system is hosted on Henan Telemedicine Center's private cloud, using a distributed storage system for high availability,

XSL•FO
RenderX

redundancy, and secure access. Diagnosed slides are archived rather than deleted, ensuring long-term availability for clinical review, QC, and research. To optimize storage, frequently accessed images remain in high-performance storage, while older slides are moved to cost-efficient archival storage, maintaining scalability as telepathology adoption grows.

Figure 8. User interfaces of the web-based platform and WeChat mini-program used for telepathology consultations in Henan Province, China.

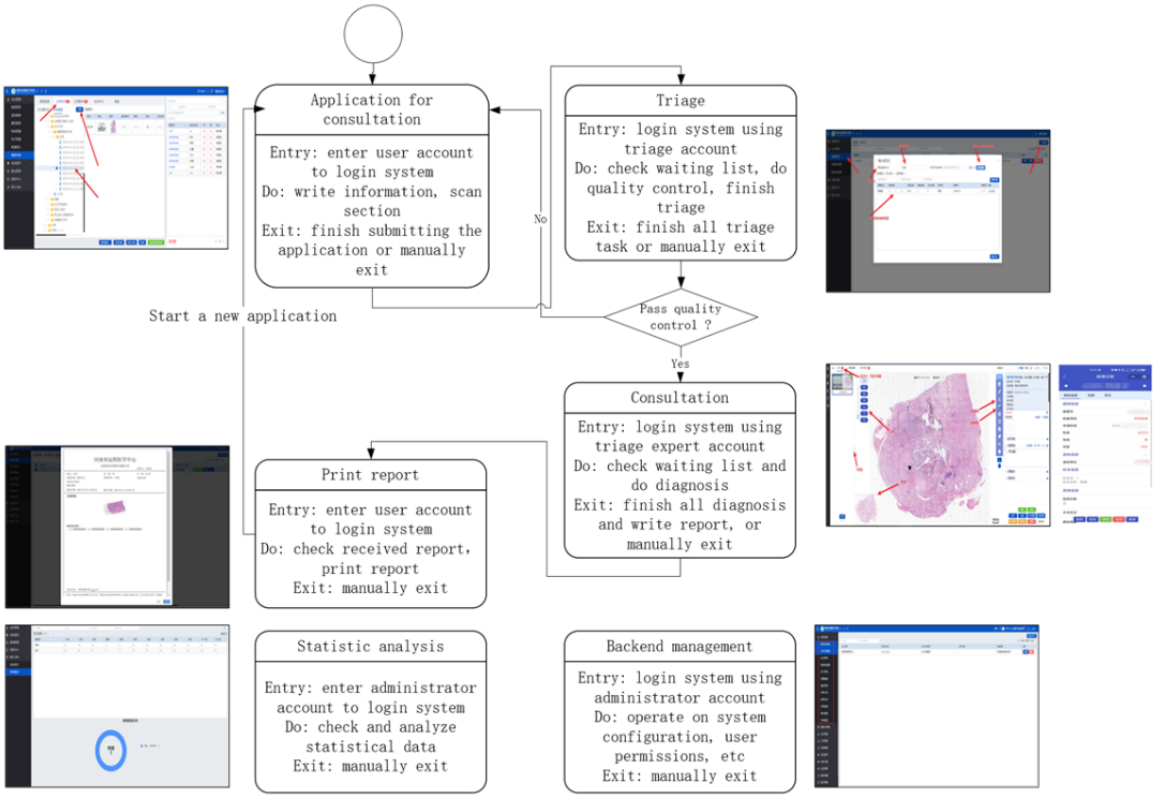

Diagnostic Process Implementation

Since deployment, the telepathology system has streamlined diagnostic workflows across 120 health care institutions, enabling seamless collaboration between county, city, and provincial-level hospitals. The system has facilitated the efficient submission, triage, and assignment of pathology cases, ensuring that cases requiring expert consultation are promptly routed.

The WeChat mini-program has been widely adopted to facilitate remote access and case review, allowing pathologists to view WSIs and provide feedback even outside traditional workstations. However, medical-grade monitors remain the preferred choice for primary diagnosis, while the mobile platform serves as a supplementary tool for preliminary assessments and report verification.

Health care professionals receive ongoing training and technical support to ensure effective system use. Annual training conferences provide updated guidance and case discussions. A comprehensive operations manual is available for reference, covering image acquisition, QC, and reporting workflows.

To optimize medical resource distribution, the system uses a tiered health care architecture, incorporating provincial, municipal, and county-level medical institutions. County-level institutions may escalate complex cases directly to provincial-level institutions or route them through municipal-level institutions. Municipal hospitals act both as requestors and providers, while provincial hospitals primarily manage complex case diagnostics.

System Performance Evaluation

To evaluate the system's technical performance, a stress test was conducted using the Apache Benchmark Tool. The detailed performance metrics are presented in Table 2. Simulated environments with 100 concurrent users performing 100 requests each were used to assess system responsiveness, scalability, and load-bearing capacity under both typical and peak usage conditions.

XSL•FO
RenderX

Table 2. Technical performance metrics of the province-wide telepathology platform under concurrent user load, evaluated using stress testing.

Requests	Response times (ms)					Throughput (transactions/s)	Network (KB/sec)	
Label	Min	Max	90th percentile	95th percentile	99th percentile		Received	sent
Consultation request interface	2568	7619	7013.10	7144.85	7497.85	25.78	114.72	48.81
Application list view page	190	3237	2859.50	2988.70	3206.30	58.88	646.77	30.24
Pending diagnosis list page	200	3017	2838.30	2891.75	3001.79	59.00	648.10	30.59
Diagnosis page	423	4359	3793.00	4031.70	4353.06	43.87	15.72	50.08

The data derived from the system performance evaluation reveals significant capabilities in handling concurrent operations across various interfaces:

1. Consultation request interface: the system can respond to 25.78 requests per second, which supports approximately 125 users operating online simultaneously.
2. Application list view page: the system is capable of handling 58.88 requests per second, allowing for about 300 users to be active online at the same time.
3. Pending diagnosis list page: the system's response rate is 59 requests per second, indicating capacity for approximately 300 users to operate online simultaneously.
4. Diagnosis page: the system manages 43.87 requests per second, suggesting it can accommodate an estimated 220 users online simultaneously.

The results demonstrate that the telepathology system can maintain high performance even under substantial user loads. The integration of a dedicated VPN, coupled with the optimized Browser-Server architecture, ensures that diagnostic workflows are not compromised during peak operational periods. These capabilities are critical for large-scale deployments in health care systems with fluctuating demands.

Real-World Impact and Clinical Outcomes

System Use and Expansion

The scanners deployed in hospitals were used for both routine in-house pathology workflows and telepathology consultations, but only selected cases were uploaded for remote expert review. The telepathology system was not intended to replace all local diagnostic workflows; rather, it served as a platform for second-opinion consultations and specialized expert input when necessary. Cases uploaded for telepathology consultations primarily included diagnostically complex cases requiring additional expert interpretation, subspecialist consultation requests for specific pathology types, and frozen section cases where rapid intraoperative decision-making was required. This selective approach ensured efficient use of DP resources while allowing hospitals to retain routine diagnostics locally. By integrating telepathology into the existing diagnostic framework, hospitals could leverage the expertise of higher-level institutions without overloading the digital infrastructure with routine cases that did not require external review.

Since its implementation, the telepathology system has been integrated into 120 health care institutions across Henan Province. From 2016 to 2024, a total of 72,916 diagnostic cases were completed, with municipal hospitals contributing 16,313 (22.37%) cases and county-level hospitals accounting for 56,603 (77.63%) cases. Provincial hospitals primarily participated as expert consultation centers and triage hubs; although technically able to submit cases, they rarely acted as requestors in routine practice. This hospital-level distribution was used as an indirect indicator of access to expert pathology support for lower-tier institutions. In China's tiered health care system, county-level hospitals typically have more limited access to subspecialty pathology expertise than higher-level centers. Therefore, the predominance of county-level submissions indicates that the platform was primarily used by lower-tier institutions and facilitated routine access to tertiary-hospital expertise for remote sites.

Analysis of annual service volumes revealed a significant upward trend. Municipal hospitals experienced a growth from 240 cases in 2016 to 3503 cases in 2021, representing a 1360% increase. Similarly, county-level hospitals saw an increase from 2095 cases in 2016 to 8317 cases in 2019, reflecting a 297% growth. Despite fluctuations in service volumes in later years, the overall trend indicates sustained adoption and use across health care tiers (Table 1).

These findings suggest that the telepathology system has effectively enhanced diagnostic capacity, particularly in county-level institutions where pathology services were previously limited.

Diagnostic Efficiency and TAT

The system's efficiency was assessed based on TAT, defined as the interval from whole-slide image submission to diagnostic report issuance. The median TAT was 10.06 hours (IQR 1.63-29.10), with 54.28% (39,579/72,916) diagnosed within 12 hours and 96.41% (70,298/72,916) completed within 72 hours.

For frozen section consultations, where rapid diagnostic turnaround is critical, the system facilitated real-time notifications to pathologists upon case submission. From 2016 to 2024, a total of 508 frozen section cases were processed through the telepathology system, with an average time from request to diagnosis of approximately 16 minutes and 6 seconds. This highlights the system's ability to support time-sensitive

XSL•FO
RenderX

intraoperative pathology decisions and enhance surgical workflow efficiency.

Patient Demographics and Diagnostic Distribution

A total of 72,916 patient cases were analyzed. Patient age ranged from 2 to 91 years (mean 51.8, SD 16.9, median 52.6, IQR 45-67 years). Age-group distribution was as follows: 18 years, n=2187 (3%); 19-40 years, n=14,583 (20%); 41-60 years, n=38,652 (53.01%); and >60 years, n=17,494 (23.99%). Overall, 43.58% (n=31,777) of cases were male, and 56.42% (n=41,139) were female. Regarding diagnostic outcomes, 48.80% (35,583 cases) were benign, while 51.20% (37,333 cases) were malignant. Compared with a baseline dataset from 2020 [10], there was an observed increase in the proportion of male patients by 2.49% (from 41.09%) and a decrease in benign lesion diagnoses by 3.38% (from 52.18%).

The most frequently reported anatomical sites were uterine, gastrointestinal, and esophageal tissues. Other common specimen categories included breast, lung, mediastinum, and male genital organs. Additional sites included thyroid, head and neck, and respiratory tract. Although biopsy specimens were included, cases were categorized by anatomical site rather than procedural type (biopsy or resection), and detailed specimen-type classification was not available in the current platform; future updates may incorporate specimen-type fields to improve data granularity.

Economic Impact

Before the implementation of telepathology services, patients in county hospitals seeking expert pathology consultation typically had to travel to tertiary hospitals with their glass slides and medical records, incurring transportation costs, lost wages, and long wait times. Alternatively, some hospitals mailed slides for consultation, but this method was slow and inefficient, often leading to delayed diagnoses.

With telepathology, biopsy slides are now digitized locally and reviewed remotely by experts at tertiary institutions, eliminating the need for physical transport. Using the previously published per-case savings estimate of 378.5 RMB (approximately US $50) [10], and applying it to the annual consultation volumes observed in our network (Table 1), the estimated direct cost savings were approximately US $0.14 to $0.63 million per year across 2016-2024.

In the retrospective subset of 1027 consultations with paired diagnoses, the tertiary-hospital expert diagnosis recorded in the platform was treated as the reference standard (malignant vs nonmalignant). Originating-site preliminary diagnoses showed a sensitivity of 0.90 (481/535) and a specificity of 0.75 (369/492), with an overall agreement of 0.83 (850/1027). Overall, 177 (17.2%) cases were discordant, including 54 false negatives and 123 false positives (Table 3).

Table 3. Confusion matrix comparing originating-site preliminary diagnoses with tertiary-hospital expert reference diagnoses in telepathology consultations.

Originating-site preliminary diagnosis	Expert reference malignant (n=535), n	Expert reference nonmalignant (n=492), n
Originating-site malignant	481[a]	123[b]
Originating-site nonmalignant	54[c]	369[d]

[a]TP: true positive.
[b]FP: false positive.
[c]FN: false negative.
[d]TN: true negative.

Discussion

Principal Results

This study presents the design, deployment, and retrospective evaluation of a province-wide telepathology system implemented across multiple health care institutions in Henan, China. In line with the study aims, the system demonstrated (1) robust technical performance under concurrent load, (2) sustained real-world use with timely reporting, and (3) improved diagnostic accuracy in a retrospective subset compared with originating-site preliminary diagnoses. Overall, these findings suggest that a centrally coordinated telepathology network, supported by secure transmission and optional mobile access, can improve diagnostic access, operational efficiency, and quality assurance at scale.

Interpretation and Implications

A key finding of this study was the high level of engagement from county-level hospitals. This pattern of use suggests that the telepathology network effectively expanded access to specialist pathology expertise for lower-tier and predominantly rural institutions, where in-house subspecialty support is often limited within China's tiered health care system. By lowering geographic and institutional barriers, the system enabled routine consultation with tertiary-hospital experts and may help reduce long-standing disparities in diagnostic capacity.

From an operational perspective, the system supported timely reporting for routine consultations and rapid turnaround for time-sensitive frozen-section cases, indicating that telepathology can be integrated into both routine and intraoperative workflows without compromising clinical responsiveness. The availability of mobile access via a WeChat mini-program further enhanced flexibility for case review and report verification, particularly for clinicians working outside traditional workstations, while maintaining compatibility with established diagnostic practices.

Expert review conducted through the telepathology network corrected a meaningful proportion of preliminary diagnoses originating from lower-tier institutions. This finding underscores the role of telepathology not only in expanding access but also

XSL•FO
RenderX

in strengthening diagnostic quality assurance by enabling second opinions and subspecialty input for complex or ambiguous cases.

In addition to clinical benefits, the system demonstrated potential economic value by reducing the need for patients to travel to tertiary hospitals for expert consultation. These cost-related benefits may be particularly important for rural populations, for whom travel expenses and delays can pose substantial barriers to timely diagnosis and care.

Comparison With Prior Work

Prior studies have shown that telepathology can improve access to subspecialty expertise, shorten diagnostic turnaround, and support pathology services in resource-limited or geographically dispersed settings [19,23,24]. However, most reported implementations are confined to single institutions, pilot programs, or short-term projects. As a result, their impact on system-level diagnostic inequities remains limited [26].

In this context, this study differs in both scale and duration. We report a province-wide telepathology network involving 120 health care institutions across multiple levels of care, sustained over nearly a decade. Previous regional or academic center–based networks, including the California Telepathology Service, have demonstrated feasibility and operational benefits but were implemented within more restricted organizational or geographic boundaries [25,30]. By integrating telepathology into a tiered provincial health care system, our network primarily served county and city-level hospitals while using tertiary centers as expert hubs, a service and governance model that has been less frequently described.

Earlier work has emphasized that successful telepathology implementation relies on standardized image acquisition, secure data transmission, reliable storage, and structured quality assurance workflows [20,31,32]. The system evaluated in this study follows these principles through province-wide scanner deployment, centralized triage, and private-network transmission. At the same time, it addresses practical challenges related to large-scale WSIs storage and concurrent access, which have been identified as key barriers to broader adoption [26,28].

Remote access has become increasingly important for pathology workflows, particularly during periods of workforce shortage or limited on-site availability [33,34]. Most prior studies have focused on web-based or remote desktop solutions for diagnostic review. In our network, the WeChat mini-program was introduced as a complementary tool, supporting case monitoring, preliminary review, and report verification.

Regarding diagnostic quality, previous validation and consultation studies have reported that telepathology-supported second opinions can improve diagnostic confidence and help resolve challenging cases [19,25,34,35]. Our findings are consistent with this literature and indicate that expert review through the telepathology network corrected a substantial proportion of preliminary classifications. This supports the role of telepathology as a quality assurance and capacity building tool, rather than a replacement for local diagnostic workflows.

Economic analyses of telepathology have mainly focused on reductions in patient travel, slide transport, and associated delays [10]. The estimated cost savings in this study, derived by applying a previously published per-case savings estimate to observed consultation volumes, are consistent with prior reports and suggest that large-scale telepathology deployment may yield meaningful aggregate benefits for patients and health care systems. Nonetheless, as highlighted in prior studies, comprehensive cost-effectiveness evaluations that incorporate long-term outcomes and system-level resource use remain scarce and should be addressed in future research [4,26].

Limitations

This study has several limitations. First, use and workflow metrics were derived from routinely collected platform data and reflect cases submitted for telepathology consultation rather than all pathology cases processed locally, which may limit representativeness. Second, diagnostic accuracy was assessed retrospectively in a subset of cases using expert diagnoses as the reference standard, and therefore does not capture interobserver variability or replace prospective multireader validation. Third, the economic impact was estimated rather than assessed through a formal cost-effectiveness analysis that incorporates long-term outcomes and system-level resource use. Specifically, we adopted a published per-case savings estimate for the same program [10] and extrapolated annual savings using observed consultation volumes, which may not capture variation in travel distance, local costs, or indirect costs. Finally, the study was conducted in a single province, and generalizability may vary in regions with different infrastructure and workforce distribution.

Clinical Prospects

This province-wide implementation provides a practical blueprint for scaling telepathology in resource-limited settings by combining standardized workflows, secure transmission, and multiterminal access. Future work should include prospective validation across broader case mixes, formal economic evaluation, and continued investment in training and quality assurance to sustain diagnostic reliability across health care tiers. Incorporating AI-assisted decision support and interoperability standards may further enhance efficiency and consistency as adoption expands [36].

Conclusions

A province-wide telepathology network can improve access to expert pathology services, support timely reporting, and enhance diagnostic performance in real-world practice. The high use of county-level hospitals suggests potential to reduce diagnostic inequity by extending specialist capacity to lower-tier institutions. This model may inform regional pathology capacity building and scalable telepathology deployment in other settings.

XSL•FO
RenderX

Acknowledgments

The authors declare the use of generative artificial intelligence (GenAI) during manuscript preparation. According to the GAIDeT taxonomy (2025), the following tasks were supported by a GenAI tool under full human supervision: English language editing and phrasing optimization, translation, and reformatting. The tool used was ChatGPT (OpenAI; accessed on 2025-03-29). The authors reviewed and verified all AI-assisted edits and are fully responsible for the final manuscript. No GenAI was used for data analysis, statistical modeling, or result generation.

Funding

This study was supported by the Henan Provincial Natural Science Foundation for Young Scholars (No. 252300423337), the Henan Provincial Key Medical Science and Technology Project (Jointly Constructed by the Ministry of Science and Technology and Henan Province; No. SBGJ202402097), and the National Key R&D Program of China (Grant No. 2022ZD0160704). The funders had no role in the study design, data collection, data analysis, interpretation of results, or manuscript preparation.

Data Availability

The study data are not publicly available because of legal, ethical, and privacy restrictions related to provincial health information regulations. The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

JS led the conceptualization of the study, with MY and DS providing supporting input. JS, MY, and DS were responsible for the study methodology and investigation. JS and DS conducted data curation and formal analysis. XH, YL, and LW contributed to system implementation and resources. HC, WW, and JZ contributed to data interpretation, manuscript review, and editing. JS drafted the original manuscript. FC supervised the study, managed project administration, and led funding acquisition.

Conflicts of Interest

None declared.

References

1. Bu H, Li Y. Pathology. Beijing: People's Medical Publishing House; Jul 01, 2018.
2. Hanna MG, Ardon O. Digital pathology systems enabling quality patient care. Genes Chromosomes Cancer 2023;62(11):685-697. [doi: 10.1002/gcc.23192] [Medline: 37458325]
3. Eloy C, Vale J, Curado M, Polónia A, Campelos S, Caramelo A, et al. Digital pathology workflow implementation at IPATIMUP. Diagnostics (Basel) 2021;11(11):2111 [FREE Full text] [doi: 10.3390/diagnostics11112111] [Medline: 34829458]
4. Ho J, Ahlers SM, Stratman C, Aridor O, Pantanowitz L, Fine JL, et al. Can digital pathology result in cost savings? A financial projection for digital pathology implementation at a large integrated health care organization. J Pathol Inform 2014;5(1):33 [FREE Full text] [doi: 10.4103/2153-3539.139714] [Medline: 25250191]
5. Schüffler PJ, Geneslaw L, Yarlagadda DVK, Hanna MG, Samboy J, Stamelos E, et al. Integrated digital pathology at scale: a solution for clinical diagnostics and cancer research at a large academic medical center. J Am Med Inform Assoc 2021;28(9):1874-1884 [FREE Full text] [doi: 10.1093/jamia/ocab085] [Medline: 34260720]
6. Montezuma D, Monteiro A, Fraga J, Ribeiro L, Gonçalves S, Tavares A, et al. Digital pathology implementation in private practice: specific challenges and opportunities. Diagnostics (Basel) 2022;12(2):529 [FREE Full text] [doi: 10.3390/diagnostics12020529] [Medline: 35204617]
7. Zhang J, Zhang S. Cancer Report in Henan. Henan, Zhengzhou: Henan Science and Technology Press; Apr 19, 2023.
8. Andre K. The importance of health information on the internet: how it saved my life and how it can save yours. J Med Internet Res 2019;21(10):e16690 [FREE Full text] [doi: 10.2196/16690] [Medline: 31661084]
9. National Health Commission. China Health Statistics Yearbook. Beijing: China Union Medical University Press; May 17, 2023.
10. He X, Wang L, Wang L, Gao J, Cui F, Ma Q, et al. The effectiveness of a cloud-based telepathology system in China: large-sample observational study. J Med Internet Res 2021;23(7):e23799 [FREE Full text] [doi: 10.2196/23799] [Medline: 34326037]
11. Rodriguez JA, Charles JP, Bates DW, Lyles C, Southworth B, Samal L. Digital healthcare equity in primary care: implementing an integrated digital health navigator. J Am Med Inform Assoc 2023;30(5):965-970 [FREE Full text] [doi: 10.1093/jamia/ocad015] [Medline: 36795062]

XSL•FO
RenderX

12. Verma N, Lehmann H, Alam AA, Yazdi Y, Acharya S. Development of a digital assistant to support teleconsultations between remote physicians and frontline health workers in India: user-centered design approach. JMIR Hum Factors 2023;10:e25361 [FREE Full text] [doi: 10.2196/25361] [Medline: 36729578]
13. Weinstein RS, Holcomb MJ, Krupinski EA. Invention and early history of telepathology (1985-2000). J Pathol Inform 2019;10(1):1 [FREE Full text] [doi: 10.4103/jpi.jpi_71_18] [Medline: 30783545]
14. Orah N, Rotimi O. Telepathology in low resource African settings. Front Public Health 2019;7:264 [FREE Full text] [doi: 10.3389/fpubh.2019.00264] [Medline: 31572705]
15. Brunyé TT, Booth K, Hendel D, Kerr KF, Shucard H, Weaver DL, et al. Machine learning classification of diagnostic accuracy in pathologists interpreting breast biopsies. J Am Med Inform Assoc 2024;31(3):552-562 [FREE Full text] [doi: 10.1093/jamia/ocad232] [Medline: 38031453]
16. Kothari S, Phan JH, Stokes TH, Wang MD. Pathology imaging informatics for quantitative analysis of whole-slide images. J Am Med Inform Assoc 2013;20(6):1099-1108 [FREE Full text] [doi: 10.1136/amiajnl-2012-001540] [Medline: 23959844]
17. Kaplan KJ, Burgess JR, Sandberg GD, Myers CP, Bigott TR, Greenspan RB. Use of robotic telepathology for frozen-section diagnosis: a retrospective trial of a telepathology system for intraoperative consultation. Mod Pathol 2002;15(11):1197-1204. [doi: 10.1097/01.mp.0000033928.11585.42] [Medline: 12429799]
18. Ahuja S, Zaheer S. Advancements in pathology: digital transformation, precision medicine, and beyond. J Pathol Inform 2025;16:100408 [FREE Full text] [doi: 10.1016/j.jpi.2024.100408] [Medline: 40094037]
19. Bashshur RL, Krupinski EA, Weinstein RS, Dunn MR, Bashshur N. The empirical foundations of telepathology: evidence of feasibility and intermediate effects. Telemed J E Health 2017;23(3):155-191 [FREE Full text] [doi: 10.1089/tmj.2016.0278] [Medline: 28170313]
20. Lennerz JK, Pantanowitz L, Amin MB, Eltoum I, Hameed MR, Kalof AN, et al. Ensuring remote diagnostics for pathologists: an open letter to the US Congress. Nat Med 2022;28(12):2453-2455. [doi: 10.1038/s41591-022-02040-6] [Medline: 36266514]
21. Weinstein RS, Descour MR, Liang C, Bhattacharyya AK, Graham AR, Davis JR, et al. Telepathology overview: from concept to implementation. Hum Pathol 2001;32(12):1283-1299. [doi: 10.1053/hupa.2001.29643] [Medline: 11774159]
22. Weinstein RS, Graham AR, Richter LC, Barker GP, Krupinski EA, Lopez AM, et al. Overview of telepathology, virtual microscopy, and whole slide imaging: prospects for the future. Hum Pathol 2009;40(8):1057-1069. [doi: 10.1016/j.humpath.2009.04.006] [Medline: 19552937]
23. Voelker HU, Strehl A, Stuefe A, Stauch G, Mueller-Hermelink H. Re-evaluation of challenging telepathological diagnoses in support of a hospital in Tanzania. J Telemed Telecare 2019;27(3):183-190. [doi: 10.1177/1357633x19866564] [Medline: 31370738]
24. Dietz RL, Hartman DJ, Pantanowitz L. Systematic review of the use of telepathology during intraoperative consultation. Am J Clin Pathol 2020;153(2):198-209 [FREE Full text] [doi: 10.1093/ajcp/aqz155] [Medline: 31618416]
25. Chong T, Palma-Diaz MF, Fisher C, Gui D, Ostrzega NL, Sempa G, et al. The California telepathology service: UCLA's experience in deploying a regional digital pathology subspecialty consultation network. J Pathol Inform 2019;10(1):31 [FREE Full text] [doi: 10.4103/jpi.jpi_22_19] [Medline: 31620310]
26. Jahn SW, Plass M, Moinfar F. Digital pathology: advantages, iimitations and emerging perspectives. J Clin Med 2020;9(11):3697 [FREE Full text] [doi: 10.3390/jcm9113697] [Medline: 33217963]
27. Mori I, Ozaki T, Muragaki Y, Ibata T, Ueda H, Shinagawa T, et al. Construction of web-based remote diagnosis system using virtual slide for routine pathology slides of the rural hospital in Japan. Diagn Pathol 2013;8(S1):1-4. [doi: 10.1186/1746-1596-8-s1-s4]
28. Subramanian H, Subramanian S. Improving diagnosis through digital pathology: proof-of-concept implementation using smart contracts and decentralized file storage. J Med Internet Res 2022;24(3):e34207 [FREE Full text] [doi: 10.2196/34207] [Medline: 35343905]
29. Li X, Cheng K, Shen Z, Lee PPC. Fast proactive repair in erasure-coded storage: analysis, design, and implementation. IEEE Trans Parallel Distrib Syst 2022;33(12):3400-3414. [doi: 10.1109/tpds.2022.3152817]
30. Chong T, Palma-Diaz MF, Fisher C, Gui D, Ostrzega NL, Sempa G, et al. The california telepathology service: UCLA's experience in deploying a regional digital pathology subspecialty consultation network. J Pathol Inform 2019;10(1):31 [FREE Full text] [doi: 10.4103/jpi.jpi_22_19] [Medline: 31620310]
31. Hanna MG, Ardon O, Reuter VE, Sirintrapun SJ, England C, Klimstra DS, et al. Integrating digital pathology into clinical practice. Mod Pathol 2022;35(2):152-164 [FREE Full text] [doi: 10.1038/s41379-021-00929-0] [Medline: 34599281]
32. Evans AJ, Vajpeyi R, Henry M, Chetty R. Establishment of a remote diagnostic histopathology service using whole slide imaging (digital pathology). J Clin Pathol 2021;74(7):421-424. [doi: 10.1136/jclinpath-2020-206762] [Medline: 32611763]
33. Rojansky R, Jhun I, Dussaq AM, Chirieleison SM, Nirschl JJ, Born D, et al. Rapid deployment of whole slide imaging for primary diagnosis in surgical pathology at stanford medicine: responding to challenges of the COVID-19 pandemic. Arch Pathol Lab Med 2023;147(3):359-367 [FREE Full text] [doi: 10.5858/arpa.2021-0438-OA] [Medline: 35802938]
34. Hanna MG, Reuter VE, Ardon O, Kim D, Sirintrapun SJ, Schüffler PJ, et al. Validation of a digital pathology system including remote review during the COVID-19 pandemic. Mod Pathol 2020;33(11):2115-2127 [FREE Full text] [doi: 10.1038/s41379-020-0601-5] [Medline: 32572154]

XSL•FO
RenderX

35. Costello SSP, Johnston DJ, Dervan PA, O'Shea DG. Development and evaluation of the virtual pathology slide: a new tool in telepathology. J Med Internet Res 2003;5(2):e11 [FREE Full text] [doi: 10.2196/jmir.5.2.e11] [Medline: 12857667]
36. Cui M, Zhang DY. Artificial intelligence and computational pathology. Lab Invest 2021;101(4):412-422 [FREE Full text] [doi: 10.1038/s41374-020-00514-0] [Medline: 33454724]

Abbreviations

DICOM: Digital Imaging and Communications in Medicine
DP: digital pathology
QC: quality control
TAT: turnaround time
TIFF: tagged image file format
VPN: virtual private network
WSI: whole-slide image

Edited by A Stone; submitted 29.Mar.2025; peer-reviewed by J Li, W Zhang; comments to author 09.Dec.2025; revised version received 05.Feb.2026; accepted 10.Feb.2026; published 09.Mar.2026.

Please cite as:
Shi J, Ye M, Sun D, He X, Lu Y, Wang L, Chen H, Wang W, Zhao J, Cui F
Telepathology and Mobile Health System for Province-Wide Pathology Consultation in Henan, China: Retrospective Evaluation Study
J Med Internet Res 2026;28:e75172
URL: https://www.jmir.org/2026/1/e75172
doi:10.2196/75172
PMID:

©Jinming Shi, Ming Ye, Dongxu Sun, Xianying He, Yaoen Lu, Linlin Wang, Haotian Chen, Wenchao Wang, Jie Zhao, Fangfang Cui. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Creation of EmpowerMe Website to Promote Self-Efficacy in Survivors of Stroke: Co-Design Study

Elizabeth A Lynch[1], BPTh, PhD; Adrian O'Malley[2]; Zoe Adey-Wakeling[3,4], BMed, PhD; Billie Bonevski[4], BA, PhD; Dominique A Cadilhac[5,6], BSN, PhD; Saran Chamberlain[1,7]; Leonid Churilov[8], BSc, PhD; Robyn Clark[1], BSN, PhD; Richard Cullen[9], BMMC; Erin Godecke[10,11], BSpP, PhD; Gillian Harvey[1], BSN, PhD; Fiona Jones[12], BPTh, PhD; Natasha A Lannin[13,14], BOccThy, PhD; Stacy Larcombe[1], BPTh; Jarrad Law[15,16], BEng; Annette McGrath[17]; Lisa Murphy[9], BMed, PhD; Katie Nesbitt[1], BSN, PhD; Karly Zacharia[18], BNAD, PhD; Niranjan Bidargaddi[15*], BEng, PhD; Coralie English[18,19,20*], BPTh, PhD

[1]Caring Futures Institute, College of Nursing and Health Sciences, Flinders University, College of Nursing and Health Sciences, Flinders University, Adelaide, Australia
[2]Lived Experience of Stroke, Sydney, Australia
[3]Flinders Medical Centre, Adelaide, Australia
[4]Flinders Health and Medical Research Institute, College of Medicine and Public Health, Flinders University, Adelaide, Australia
[5]Stroke and Ageing Research, Department of Medicine, School of Clinical Sciences, Monash University, Melbourne, Australia
[6]Public Health and Health Services Research, Stroke and Critical Care, The Florey, The University of Melbourne, Melbourne, Australia
[7]Lived Experience of Stroke, Adelaide, Australia
[8]Melbourne Medical School, The University of Melbourne, Melbourne, Australia
[9]National Stroke Foundation, Melbourne, Australia
[10]Stroke Recovery and Rehabilitation, Perron Institute for Neurological and Translational Science, Perth, Australia
[11]Allied Health Research, Sir Charles Gairdner Osborne Park Health Care Group, Perth, Australia
[12]Population Health Research Institute, City St George's University of London, London, United Kingdom
[13]Department of Neurosciences, School of Translational Medicine, Monash University, Melbourne, Australia
[14]Alfred Health, Melbourne, Australia
[15]Flinders Digital Health Research Centre, College of Medicine and Public Health, Flinders University, Adelaide, Australia
[16]CareMappr, Adelaide, Australia
[17]Lived Experience of Stroke, Port Elliot, Australia
[18]School of Health Sciences, University of Newcastle Australia, Newcastle, Australia
[19]Heart and Stroke Program, Hunter Medical Research Institute, Newcastle, Australia
[20]Centre for Research Excellence to Accelerate Stroke Trial Innovation and Translation, The University of Sydney, Sydney, Australia
[*]these authors contributed equally

Corresponding Author:
Elizabeth A Lynch, BPTh, PhD
Caring Futures Institute, College of Nursing and Health Sciences, Flinders University
College of Nursing and Health Sciences
Flinders University
GPO Box 2100
Adelaide, 5001
Australia
Phone: 61 0882017647
Email: elizabeth.lynch@flinders.edu.au

Abstract

Background: Digital health innovations are frequently used to support people in managing chronic health conditions. Stroke is common, and people who have survived a stroke and live in the community must learn to manage their health independently. Digital tools can help, but only if designed to match survivors' specific needs. In response to a need expressed by people living

XSL•FO
RenderX

with chronic health conditions, the Australian government created a funding stream to support the development of a digital resource to help individuals gain confidence in managing their health.

Objective: This study aimed to co-design a digital resource to promote self-efficacy to manage life after stroke in community-dwelling survivors of stroke.

Methods: Co-design methodology, which emphasized meaningful engagement with intended end-users throughout the design of the website, was used. The project steering group comprised health professional researchers, digital designers, people with lived experience of stroke (survivors and carers), and representatives from the Stroke Foundation (Australia). A systematic review was conducted to inform the core components of the resource. A lived experience workgroup was convened to advise on features of the digital resource and aspects of its evaluation. Iterative review stages and frequent consultation between the steering group and the lived experience workgroup occurred. Online and in-person usability testing was conducted with survivors and carers.

Results: The lived experience workgroup (workgroup) initially comprised 14 survivors and 1 carer. In total, 11 survivors and 1 carer remained engaged throughout the co-design period. One workgroup member was invited to join the steering group and co-facilitated all co-design workgroup meetings. Defining features of the digital resource were identified by the steering group and the workgroup, including that the resource would be a website that augmented (not reproduced) existing resources, and it needed to be accessible to survivors of stroke with communication changes. Website specifications were determined by the workgroup and included that information needed to be tailored to the individual user, and specific accessibility features were recommended. The workgroup prioritized what content to include on the website and recommended the creation of video stories by Australian survivors and carers. A resource-tailoring tool was created so information could be individualized to the interests of the website user. Overall, 9 web pages containing high-priority content were created, comprising text, video stories, and a downloadable PDF that summarized key information for that page. More than 150 short video stories were created by 26 survivors of stroke and 10 carers for the website. Usability testing indicated that the website was more usable than 83% of all websites.

Conclusions: Authentic co-design with inclusion of people with lived experience of stroke at all stages of development enabled the successful build of a digital resource (website) to improve self-efficacy. An evaluation of the website is underway.

Trial Registration: Australia New Zealand Clinical Trials Registry ACTRN12624001018505; https://anzctr.org.au/Trial/Registration/TrialReview.aspx?id=388255

(J Med Internet Res 2026;28:e76756) doi:10.2196/76756

KEYWORDS

stroke; website; self-efficacy; co-design; digital health

Introduction

Chronic health conditions are the leading causes globally of death and disability [1,2]. Effective ways to support people to manage their chronic health conditions have long been recognized as important to maximize health, quality of life, and reduce health care use [3,4]. As populations have become more familiar with using the internet to find health advice [5,6], and advances have been made in available digital health technologies [7,8], there has been a concurrent growth in digital health care delivery, which expanded exponentially with the COVID-19 pandemic [9,10]. Accordingly, digital health innovations are now commonly used to support people to manage their chronic health conditions [11,12].

In 2023, stroke was the second most common cause of death and disability worldwide [1,2]. Despite reductions in the incidence of stroke in high-income countries, the prevalence of people living with stroke is increasing because of aging populations and better survival rates [13]. Affecting more than 100 million people worldwide [14], stroke can affect an individual's physical ability, energy levels, mood, as well as their ability to communicate, concentrate, and understand [15]. Changes to cognition, communication, and stroke-related fatigue (all affecting at least half of all survivors) [16-18] can present difficulties for survivors in accessing digital health technologies [19,20]. Further, survivors of stroke with communication difficulties and people with cognitive changes are frequently excluded from participating in research [21,22], and little guidance exists in the published literature on how to work with survivors of stroke when designing a digital health innovation.

People living with chronic health conditions were consulted about research priorities to address in Australia and prioritized the development of digital resources to promote health self-efficacy [23]. Self-efficacy is an individual's confidence in their abilities [24], and when teamed with health, it relates to an individual's confidence in their ability to manage their health effectively. A research team was convened and sought to develop a digital resource to build health self-efficacy in survivors of stroke.

For any digital innovation, it is recommended that designers work with intended users during the development phase to optimize acceptability and uptake [25]. A co-design approach was chosen so survivors of stroke and other key stakeholders could be meaningfully involved and engaged throughout the entire research process [26], from conceptualizing and building the digital resource, through to its subsequent evaluation. This study describes how the team co-designed an accessible digital resource to promote self-efficacy in survivors of stroke.

XSL•FO
RenderX

Methods

Overview

The principles of design science in information systems research were used to guide the development of an innovative, purposeful digital resource [27]. The design process was planned to systematically incorporate the needs and preferences of survivors of stroke and carers through each planning and design stage of the digital resource [26]. In total, 4 steps were followed: determining what was required from the resource, determining specifications of the resource, building the resource to the specifications, and testing the resource to determine if it met requirements. These are presented in Figure 1.

Figure 1. Steps followed to design the EmpowerMe website and content.

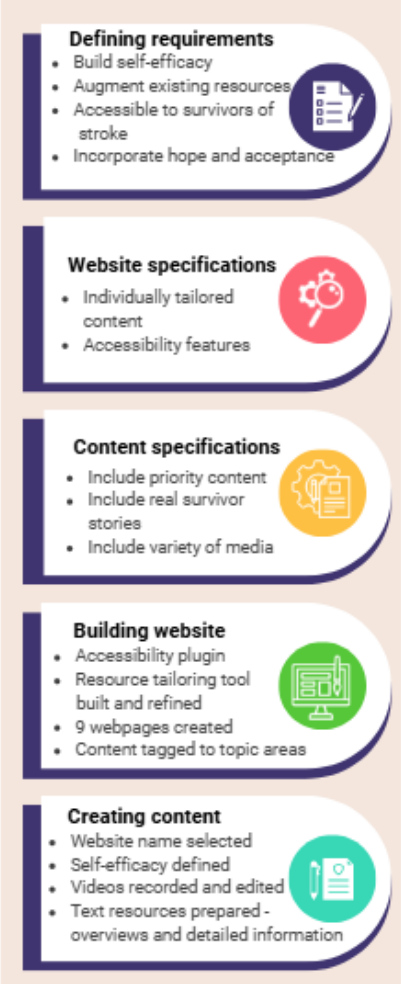

Formation of the Steering Group

The team that developed the research proposal and coordinated the study (hereafter referred to as the steering group) comprised 20 members (16 women), invited by Author EAL for their complementary expertise relevant to the project. The steering group included 14 academic researchers from health, bioinformatics, and biostatistics disciplines, 2 survivors of stroke, 1 carer, health professionals, and representatives from Stroke Foundation (Australia), the national stroke advocacy body. The steering group determined the features of the digital resource that were essential, and the adaptable features that should be determined by people with lived experience of stroke (Figure 2). The steering group met online monthly and had oversight of the resource design, timelines, and budget management.

XSL•FO
RenderX

Figure 2. Core and adaptable features of the digital resource.

Formation of Lived Experience Workgroup

A lived experience workgroup (hereafter referred to as the Workgroup) was convened, comprising survivors of stroke and carers, to advise on the adaptable features of the digital resource. Workgroup members were recruited by advertising expressions of interest via social media (Facebook [Meta] and X [formerly Twitter]). Interested individuals contacted a project team member (KZ), who then met online to explain and answer any questions about the project and discuss skills each individual could contribute to the project, in addition to their lived experience. Written information was provided about the project, and all Workgroup members provided informed consent to contribute their ideas and have online meetings video-recorded. Overall, 14 survivors and 1 carer (10 women) were recruited to the workgroup, representing community-dwelling people of different ages (ranging from 43 to 76 years), stroke-related changes, and time since stroke (6 months to 17 years at project commencement). Workgroup members were from 6 Australian states or territories. Some had extensive experience in research and co-design, whereas 3 had no prior research experience. Over the course of the 12-month co-design phase, 3 workgroup members withdrew due to overseas travel, illness, and being too busy.

Lived Experience Workgroup Meetings

Workgroup meetings were run online via Microsoft Teams and recorded. All Workgroup members were offered support to familiarize themselves with Microsoft Teams at the start of the project. Meetings were co-facilitated by an academic researcher (author EAL) and a survivor of stroke (authors SC or AO'M). Discussion points with or without preliminary reading were sent to attendees before each meeting. Each meeting was attended by a maximum of 5 workgroup members to allow time for contributions from each attendee, so 3 meetings (attended by different workgroup members) were held to discuss each topic.

At the start of each meeting, EAL presented relevant background to the intended discussions and the anticipated outcomes from each meeting. Author SC then facilitated discussions according to the predetermined discussion guide, inviting each person to contribute their ideas. Workgroup members were invited to "think big" and to challenge the status quo when discussing the digital resource design. At the conclusion of each meeting, EAL verbally summarized the key points raised in discussion, and Workgroup members provided agreement or clarification. Discussion points from 1 meeting were presented at subsequent meetings attended by different Workgroup members, to further explore these ideas and determine consensus or conflict. Meeting recordings were reviewed by author EAL, and key points of discussion, consensus, or conflict were documented and distributed to all attendees within a week of the final meeting. Attendee names were not included in meeting summaries. Attendees were invited to provide further details if they considered that information was missing. Facilitators SC and EAL shared key points from workgroup meetings at the bi-monthly steering group meetings, where final decisions about the digital resource design were made (in all instances, Workgroup recommendations were approved). Workgroup members were remunerated for their time at rates recommended by the New South Wales Consumer Alliance.

Overall, 6 rounds of workgroup co-design meetings were conducted between June and December 2023, interspersed with 3 online surveys between meetings, which followed up on workgroup discussions regarding content to include and how to measure the efficacy of the developed digital resource.

Online Surveys

Survey 1 comprised 19 questions, and Workgroup members were asked to rate the importance of including content about 18 different health-related needs (identified from the Australian Stroke Needs Assessment Survey [28]) using a 5-point Likert Scale (definitely include; probably include; consider including;

XSL•FO
RenderX

probably DO NOT include; definitely DO NOT include). An additional question allowed free-text responses seeking any other information to include. In Survey 2, workgroup members were asked to rate the importance (using a slider from 0 not at all important to 100 extremely important) of measuring 28 items related to self-efficacy that were identified in Workgroup discussions and from the Australian Stroke Needs Assessment Survey [28]. Respondents were also asked to nominate their top 3 items from the included list, and a final question allowed free text for any additional comments to consider about measuring the efficacy of the digital resource. In Survey 3, 2 shortlists of 10 items from Surveys 1 and 2 were presented. Workgroup members were asked to rank items in order of importance regarding the priority content to include and the priority aspects of self-efficacy to measure.

Survey Data Analysis

Data from Surveys 1 and 2 were analyzed descriptively. The frequency of "definitely include" responses for items in Survey 1 was calculated, and the means and frequency in the "top 3" were calculated for each item in Survey 2. Data from ranking exercises in Survey 3 were analyzed using a graph theory-based voting system [29], wherein group-based ordinal rankings were generated from the individual ordinal rankings.

Determining Technological Features

A further 3 rounds of co-design meetings with the Workgroup were co-facilitated by a digital designer and a survivor of stroke (author SC) to determine the technological features of the digital resource. The workgroup was reconvened between March and April 2024 to review the digital prototype. The topics discussed at the meetings are presented in Table 1. The project timeline is illustrated in Figure 3.

Table 1. Focus of workgroup meetings.

Meeting number	Main agenda items
1	• Introduction to the project • Discuss ways of working • Definition of self-efficacy • Personal experience of building self-efficacy after a stroke
2	• Terms of reference • Definition of self-efficacy • Do different groups of people have different self-efficacy needs? • Resources to support self-efficacy • Selection of a digital platform
3	• Measuring self-efficacy • Content to include on website
4	• Content to include on website • • Existing resources that should be included • Preferred format • Previous (good and bad) learning experiences using digital technology
5	• Recap of progress to date • Format of resources • Features to include on website
6	• Experiences using websites for information about health or well-being • Features of websites that make them harder or easier to use, more or less enjoyable • Opinions on • • Interactive features on websites • Sharing personal information on websites • Receiving "push" emails or text messages • Any changes to the way information is processed since the stroke
7	• Identifying an individual's information-seeking behaviors and digital preferences
8	• Reviewing the prototype of the quiz to tailor content
9	• Reviewing and refining the prototype of the quiz
10	• Reviewing website prototype • Feedback on layout and color palette • Name of website

XSL•FO
RenderX

Figure 3. Timeline of website co-design and usability testing.

Usability Testing

The tailoring of usability testing procedures and methods to meet the needs of people with language and cognitive changes has been presented elsewhere (Manuscript submitted to *Journal of Medical Internet Research* concurrently). A total of 3 groups of participants were recruited for usability testing: survivors of stroke with no self-reported language changes (n=4), survivors with moderately severe aphasia (a language disorder that affects communication; n=4), and carers of survivors of stroke (n=3). None had been involved in the design or development stage. Participants were offered an Aus $50 (US $35.45) gift voucher on completion of the testing session to thank them for participating.

Usability Testing Procedures

Testing was conducted in a one-to-one moderated format by author KN, with support from author EG for participants with aphasia. In-person testing was organized for all participants with aphasia, and in-person or online participation was offered to other participants. Usability scripts were prepared to guide participants to test navigation around the website and evaluate comprehension of the content (Multimedia Appendix 1).

Data Collection

Demographic data, time taken to complete tasks (for each task and total cumulative time), percentage of tasks completed, knowledge acquisition (questions about definition and importance of self-efficacy), System Usability Scale (SUS), and user experience (survey developed by the research team incorporating specifics of the EmpowerMe website, based on the Website Evaluation Questionnaire).

Usability Data Analysis

Participant demographic data and usability task-time data were analyzed descriptively using IBM SPSS Statistics 28.0.0.0. Knowledge acquisition was analyzed using the chunk and check technique [30,31] and reported narratively. Mean (SD) for overall, individual, and group SUS scores were analyzed using the SUS calculator package [32], and using this package, the raw mean score was normalized and converted to a percentile rank. The percentile rank indicated the relative usability of the website compared to other digital products. Group mean scores for survivors of stroke with and without aphasia were compared using a between-subject 2-sample *t* test (2-tailed), with statistical significance set at $P \leq .05$. Data from the user experience survey were analyzed descriptively and tabulated against task-time and task-completion data to examine usability from different perspectives, providing an opportunity to make targeted website changes when issues were identified.

Ethical Considerations

The project had ethical approval (5886) from Flinders University's Human Research Ethics Committee. People in the lived experience workgroup provided written informed consent for their contributions to be used to shape the digital resource design and for meetings to be video-recorded. Due to the group nature of workgroup meetings, the ideas generated were not confidential and were shared with other workgroup members. Anonymized summaries were shared with the project steering group. Workgroup members were offered Aus $100 (US $70) per meeting, inclusive of preparation time, paid via invoice or electronic gift cards.

Usability testing participants provided written informed consent to participate. The research speech pathologist (EG) prepared an aphasia-friendly participant information sheet and explained the information to each participant with aphasia, using aphasia-friendly language and allowing time for processing of information. She checked the participant's understanding of the information as part of the consent process. Data from usability testing participants were collected anonymously to protect participants' confidentiality, and only the research team had access to the data. Each participant was offered an Aus $50 (US $35) gift card as remuneration for completing the usability testing.

Results

Defining Requirements of the Digital Resource

The lived experience workgroup recommended that a website would be the most acceptable digital resource to support survivors of stroke develop their self-efficacy. Essential requirements determined by the Steering Group before project commencement were that the website would be free to access and needed to augment, not reproduce, information and other

XSL•FO
RenderX

supports already available (such as available through Stroke Foundation and its website EnableMe [33]). The website needed to be accessible to non-experts in stroke, with low or medium levels of health literacy, and accessible to people with stroke-related changes, including changes to cognition, communication, or reading ability. The Workgroup highlighted the importance of including representation of people of different ages, genders, and cultural backgrounds.

Another essential requirement of the website was that it needed to support self-efficacy in Australian survivors of stroke and people in their support network (family members, friends, and carers), so it should incorporate features such as learning from or observing other survivors and carers, and provide information. Elements that were identified by workgroup members as being important to building self-efficacy and a sense of well-being after stroke were having a purpose in life, having self-compassion, having goals to work toward, and being realistic while also having hope. Acceptance that things have changed and having resilience to bounce back if experiencing a setback were also recognized as intertwined with a person's confidence in their abilities. These elements guided the development of resources hosted on the website.

Setting Specifications

Specifications of Website

Workgroup members expressed that the website needed to incorporate a feature so that information could be tailored to each user, who could then "choose their own adventure" to find information or other resources relevant to them. Interactivity in the form of a chatbot was universally rejected by workgroup members. However, a feature where users were asked a series of questions about areas of interest or information needs that would then guide the user to a curated list of relevant resources was deemed to be of benefit. The Workgroup also recommended that the tailoring feature be optional, so users who preferred not to provide information about themselves, or people who prefer to "browse," could look through all the available resources.

In terms of accessibility features for survivors of stroke, the use of consistent layout and spacing was suggested to improve the accessibility of information for people with visual field loss. Use of sans serif font, having aphasia-friendly text options, and availability of captions for videos and adjustable playing speed were recommended to meet the needs of people with communication impairments. The Workgroup recommended having a text-to-voice option so users could choose to listen to rather than read the written information. Use of large boxes for clickable links (rather than underlined text) was recommended to improve functionality for people with motor control difficulties. Avoiding any spontaneous animations, background music, and moderating the tone of voice and volume on videos was important for people who experience sensory overload.

Other recommended features were not specific to the needs of survivors of stroke, including having options to adjust the website colors and having clear navigation pathways.

Specifications of Content to Include

Priority content to include on the website was determined by the workgroup following discussions and surveys. The list of priority content topics is presented in Table 2. The most important topics to address were "What is self-efficacy and why is it important?" and "How to be involved in health care decisions." Due to substantial overlap in content, 2 topics ("Managing concentration problems" and "Managing cognitive changes") were grouped, leaving 9 priority topics to address.

Table 2. Priority content to include and the number of videos created.

Topic number	Topics, prioritized by the lived experience workgroup	Number of videos	
		Survivors	Carers
1	Defining self-efficacy and its importance	50	9
2	Being involved in decisions	7	4
3	Seeking the information you need	10	6
4	Advocating for yourself	12	2
5	Managing fatigue	11	2
6	Emotional problems	10	3
7	Managing cognitive changes[a]	6	3
8	Managing mobility changes	11	2
9	Managing speaking difficulties	10	9
10	Managing concentration problems[a]	__[b]	__[b]

[a]Due to substantial overlap in content, "managing cognitive changes" and "managing concentration problems" were combined into one topic.
[b]Not available.

Results from a systematic review conducted by the research team about self-management interventions to improve self-efficacy in survivors of stroke [34] did not highlight any further components that needed to be incorporated into the website.

XSL•FO
RenderX

Specifications of Content Format

To meet different users' learning styles, the website needed to host content in a variety of media, including video, text (including downloadable information sheets), and audio resources. For written information, having both "bite-size" information chunks and "deep-dive" detailed information for topics was preferred, so a user could browse headlines of topics and then go into detail on topics of interest to them.

To enhance self-efficacy via vicarious learning from peers, the steering group and workgroup planned to include content delivered via testimonials from survivors of stroke and carers. There was a strong preference from workgroup members for these to be real, unscripted video-stories of Australian survivors and carers who shared their struggles and how these were managed, as well as their experiences of recovery. Specifically, Workgroup members recommended using home recordings (eg, via smartphone camera or Zoom videoconferencing) rather than professional videographers to prepare the video stories, as they felt this improved the relatability of the person sharing their story and made the stories and messages more credible. The preferred length of videos was less than 3 minutes.

For the delivery of information prepared by health professionals, Workgroup members preferred to have real people in any videos rather than using graphics or animations, and the use of Australian-specific information when available.

Building the Website and Developing Content

Incorporating Accessibility Features

An accessibility plugin (UserWay) was used to allow users to customize the site to their needs. This included options like increasing contrast, enlarging the text, increasing the text spacing, highlighting links so clickable elements are more prominent, and other options.

Building the Resource Tailoring Tool

The Workgroup members met with a digital designer and shared their poststroke information experiences, device usage patterns, content preferences, and strategies for validating health information. Following this session, the designer developed distinct user personas (Table 3) and used this information to develop a quiz-based tailoring prototype to accommodate all identified personas. The subsequent 2 sessions were used to refine the prototype through iterative feedback, emphasizing content relevance and quiz structure. Workgroup members provided critical input on optimal questionnaire length, preferred response formats, and content delivery mechanisms.

Following development and feedback, the resource tailoring tool was trialed and finalized. It comprised 2-3 (depending on responses) multiple-choice questions that subsequently became more precise to identify topics of interest to each user. Additional questions were added to the start of the tool to collect demographic data of users for research purposes (Textbox 1).

Table 3. Personas of survivors of stroke accessing digital information emerging from the design sessions.

Characteristics	Tech savvy (n=4)	Traditional learner (n=3)	Supported learner (n=3)
Digital skills	High confidence with multiple devices	Prefers structured content	Confident but needs accessibility features
Learning style	Self-directed research and critical evaluation	Relies on medical professionals	Institution-guided learning
Information sources	Multiple sources, fact-checking, and research background	Printed materials from traditional health channels	Verified institutional sources
Preferred content format	Short videos, blogs, and interactive content	Academic content and structured text	Short videos (<15 min) and articles
Device usage	Multiple devices (phone, tablet, and laptop)	Preference for larger screens	Phone and laptop primarily
Support needs	Independent learners	Guided learning approach	Accessibility adaptations

XSL•FO
RenderX

Textbox 1. Questions to guide content curation. Topic was a high priority and had a web page developed as part of this project.

What topics would you like to explore today?

How to build my confidence after a stroke

- Finding the information I need
- Learning from survivors or carers
- Practice
- Support from others
- Skill development

How to speak up for what I want or need

- Self-advocacy
- Involvement in decisions about my care

How to manage the effects of the stroke

- Managing physical changes
  - Using my arm
  - Bladder and bowel problems
  - Falls prevention
  - Mobility changes
  - Swallowing
- Managing changes that are hard to see
  - Vision
  - Cognitive (thinking and memory) changes
  - Speaking difficulties
  - Pain and sensation changes
  - Emotional changes
  - Fatigue

The resource tailoring tool was a plugin to the website built using Flutter Web, which sat on top of the website to add the quiz-based tailoring information. The structure of the Tailoring Tool and the content provided were stored in a Webiny Content Management System. Upon using the Tailoring Tool, relevant resources were pushed to website users – see Organizing Resources within the Website section.

Preferred Language and Definitions

Terminology was considered and workshopped in both workgroup and steering group meetings. Most workgroup members (except for individuals with occupational therapy or education backgrounds) were unfamiliar with the word “self-efficacy” at the start of the project, and some found it hard to pronounce. Many steering group members and workgroup members felt that the word “self-efficacy” was not part of general day-to-day language and were concerned that use of it might be off-putting. However, some workgroup members felt that it was important to educate survivors and carers about self-efficacy, and familiarization with the word “self-efficacy” was an important educational component that could empower survivors and carers. The Steering Group then decided to use lay language (avoiding the word “self-efficacy”) in headings and titles and include the word “self-efficacy” in second-tier text designed to inform and educate users.

After several iterations, the preferred definition of self-efficacy, determined by the Workgroup, was “Belief in my abilities to achieve the things that matter to me.”

Naming the Website

Initially, the Steering group anticipated naming the website ASSET-Stroke according to the project name (derived from co-designing a Digital Solution for Self-Efficacy After Stroke). However, both the steering group and workgroup agreed that this name did not encapsulate the philosophy and purpose of the website and was unlikely to attract the attention of survivors of stroke unfamiliar with the project. Accordingly, discussions were facilitated, and voting was conducted with the workgroup, and the domain name EmpowerMe was selected. This name aligns with others curated by Stroke Foundation, such as EnableMe [33] for survivors of stroke and carers and InformMe [35] for health professionals working with survivors of stroke.

XSL•FO
RenderX

Creating Website Content

Project team members created content and accompanying downloadable PDFs to address each priority topic area. All text was edited until it was deemed to be understandable by people who had completed grade 7 at school by Microsoft Copilot. Text resources for "Managing speaking difficulties" were edited by a speech pathologist with expertise in aphasia (author EG) and then checked and further edited with a survivor of stroke with moderate aphasia.

Expressions of interest were advertised for survivors of stroke and carers to create video stories to be hosted on the website about building self-efficacy after stroke, with a focus on the prioritized topic areas. Twenty-six survivors of stroke and 10 carers recorded video stories, usually over Zoom with a project team member, lasting up to 30 minutes. The project team edited the recordings into numerous short (30-s to 3-min) videos, creating more than 150 videos (Table 2). Each edited video was sent to the individual for approval or further editing before uploading to the Stroke Foundation YouTube channel, and being allocated to a topic area (one of the 9 priority topics when appropriate). A thumbnail with a brief title was provided for each video. Video contributors were reimbursed for their time.

Creating Web Pages

The main website was developed on WordPress, which allowed content to be easily updated by nontechnical team members, if needed, and was compatible with Stroke Foundation's systems to allow Stroke Foundation to manage the website on project completion. The 9 priority topics were organized into 9 web pages. The website home page presented information from the highest priority topic, "Defining self-efficacy and its importance." The resource tailoring tool was presented at the top of the page, presented as a prominent click box to "Find resources relevant to you."

The subsequent 8 topic web pages were created using a standard template, the use of icons, and formatting to ensure consistent layout. Each page had a lay description of the topic, a video about the topic, text about how to build self-efficacy in the topic, written content for carers about the importance of self-efficacy for carers or the carer's role in supporting self-efficacy of survivors, and thumbnails and titles of relevant videos created by survivors and carers relevant to the topic.

Organizing Resources Within the Website

Resources created as part of the project, and existing resources relevant to the priority content areas, were tagged by project team members regarding topic, format, and weighting in the Webiny Content Management System. Resources developed as part of the project were weighted higher than preexisting resources, and Australian resources were weighted higher than internationally produced resources. High-priority topics (identified during the co-design phase) had whole web pages and embedded resources developed, and the web page received top-weighting for the topic.

On using the resource tailoring tool, a curated list of relevant resources was created for each individual based on their responses regarding their areas of interest and associated resources' tags and weightings.

Usability Testing of the Website

Navigation of Website: Task Completion (Time Taken and Percentage Completed)

Survivors of stroke without aphasia (communication difficulties) and carers completed their tasks more quickly and completely than survivors with aphasia. Activities where survivors of stroke with aphasia struggled were reviewed by the team, and adaptations were made to enhance accessibility for this group of survivors.

Comprehension of Content

Overall, 6 participants (4 survivors with aphasia and 2 survivors without aphasia) participated in the evaluation of comprehensibility of website content. In total, 4 participants (2 [50%] survivors with aphasia and 2 [100%] survivors without aphasia) were able to understand and explain 100% of the content included in the knowledge assessment. One survivor with aphasia completed the "chunk and check" for 1 task only (20% of content included in knowledge assessment), and the remaining survivor with aphasia was unable to complete any knowledge tasks. These 2 survivors had more severe aphasia, so it was deemed that the website content was suitable for survivors with mild-moderate aphasia, and more work was required to meet the needs of survivors with severe post-stroke communication difficulties.

SUS

The SUS data were normally distributed, so overall mean scores, group mean scores, and between-subject comparison were calculated (comparing survivors with aphasia to survivors without aphasia, comparing survivors without aphasia to carers). SUS scores of survivors of stroke with aphasia were significantly lower than scores of survivors without aphasia (mean 61.3, SD 6.6 vs mean 86.8, SD 2.4; 95% CI –36.8 to –14.4). There was no statistically significant difference in SUS scores between carers and survivors without aphasia (mean 88.3, SD 11.5 vs mean 86.9, SD 2.4; 95% CI –23.3 to 26.3).

The normalized raw mean SUS score for the website was 78, and following the calculation of the percentile ranking, the usability of the EmpowerMe website was deemed better than 83% of all websites.

User Experience

Suggestions to improve the website from participants, which were also noted by the usability session moderator, included improving the visibility of resources after use of the resource tailoring tool; addressing layouts that were difficult for people with poststroke visual inattention; and improving problematic layouts (text heavy) for people with aphasia. These were subsequently addressed by the design team.

Website Efficacy

The potential of the digital resource "EmpowerMe" to influence self-efficacy, health-related quality of life, participation, and anxiety and depression in different cohorts of survivors

XSL•FO
RenderX

(stratified by age and time since stroke) is currently being evaluated as part of a phase II trial.

Discussion

Principal Results

We have presented how a digital resource in the form of a website can be built specifically to meet the needs of a client group with heterogeneous challenges, including cognitive, language, and motor changes. Rather than retrofit an existing website to the needs of people with poststroke visual, motor, cognitive, or sensory changes, we designed the website proactively to meet the diverse needs of the intended user group. People with lived experience of stroke were included at all stages, from preparation and submission of the grant, being members of the steering group, to facilitating and contributing in lived experience workgroup meetings and contributing to this article.

The Steering Group actively sought to balance the power dynamic between the academic researchers and people with lived experience. Following project commencement, the team recognized the need for more lived experience leadership to facilitate the lived experience workgroup meetings. One workgroup member (SC, survivor of stroke) with experience in both research and in facilitating online peer groups accepted an invitation to join the steering group, cofacilitate workgroup meetings, and was employed as a research assistant for the life of the project to assist in data analysis.

Within a research framework, we were successfully able to embed systems to support people with a lived experience of stroke to contribute meaningfully alongside a team of multidisciplinary experts to co-design a digital resource to address self-efficacy after stroke. Recognizing the different expertise of each party allowed the team to determine who decided each feature of the digital resource, to simultaneously address the needs and preferences of survivors of stroke and carers, and maintain the rigor and aims of the research process.

Survivors of stroke and carers, who comprised the lived experience workgroup, determined the digital platform (website with a feature to direct the user to individually relevant information), the priority content to include, and the priority format of resources (home-made video stories) to be developed by the project team. In contrast, the project Steering Group (which included 2 survivors of stroke and 1 carer) made fewer decisions but worked consistently to incorporate feedback and suggestions from the lived experience workgroup, positioning these within best-practice frameworks of adult learning, self-efficacy, self-management, and digital design.

Comparison With Prior Work

There is a growing body of literature documenting how to facilitate co-design and interactive workshops online, particularly since the COVID-19 pandemic [36-39]. Now that social distancing measures have lifted, online methods of co-designing have remained popular, with obvious benefits of reducing travel time, costs, and increasing accessibility for people who live in regional and remote areas [39,40]. While some researchers in previous work have described using interactive features of online platforms to enhance activity within online co-design meetings [37,41], interactions with the lived experience working group in the current project were mostly dialogue-based, although they were augmented with online surveys between meetings, similar to what others have also described [41].

The project reported in this article is far from the first website to have involved intended users in the design. Supporters of collaborative design methods argue that involving target users when designing a digital resource will improve the resource's impact through better user engagement and satisfaction [40,42-44]. However, the level of involvement of the intended users can vary greatly between studies that report using co-design methods. While co-design has been described as “meaningful end-user engagement in research design and includes instances of engagement that occur across all stages of the research process” [26], it is also acknowledged that the engagement activities “range in intensity from relatively passive to highly active and involved [26]. In a systematic review which tabled the most common co-design activities in studies reporting co-design of digital health solutions, 2 of the most common activities reported were interviews and surveys [45]. Implicit in these activities are the research team setting the agenda, seeking a one-way transference of “data” from participants, and controlling how the data are analyzed, which can be disempowering (rather than engaging) for participants [46]. Other commonly reported co-design activities that may promote more engagement were workshops [45], which tend to be more collaborative and allow for a two-way exchange of information, having been described as occasions in which a group of people can acquire new knowledge or perform creative problem-solving [47].

Intended users have been variably involved in determining whether a website would be of benefit, ranging from not being involved in the decision, (for instance when the rationale of developing a website is based on disease incidence and rates of technology use [43]), to contributing data that is used by researchers to support the project idea (eg, rationale is underpinned by interviews of participants living with the condition under investigation [48]) to sharing responsibility for the decision to develop the website (eg, when people with lived experience are members of the research team and help shape the research proposal [49,50]).

Similarly, the level of input and decision-making responsibility of representatives of the intended user group about features of the co-designed websites varies. Intended users could contribute data to be used by the design team to make decisions (such as people being recruited as participants in interviews, focus groups or co-design sessions) [42,48], they could advise on features (eg, via lived experience advisory groups) [42,48,50] or be decision-making partners (eg, via being members of the research or design team) [49,50]. Just as in our project, several studies reported involving people from the intended user group in different roles within the project [42,48,50], for instance, with some people recruited as participants, others consulting in an advisory capacity, and others partnering on the research team. We have advanced the field beyond existing data tools by explicitly working with the intended end users to embed features

XSL•FO
RenderX

to tailor content and accessibility to survivors of stroke. Unique to this project was the need to develop a website that was accessible to people with language, visual, motor, cognitive, or sensory changes from their stroke. By proactively incorporating accessibility features into the design (rather than retrofitting these to a standard website design), we anticipate the website will address the variety of challenges faced by survivors of stroke. Further, by co-creating content, the voices and preferences of survivors of stroke have been prioritized to support peer learning and peer support.

Compared to prior work, the current project used more active and involved methods of working with people with lived experience of stroke throughout the design of the website than has been reported in most other publications reporting on co-design of websites to support health or well-being. We identified 1 published work presenting the co-design of a website to support physical activity and diet to reduce secondary stroke [50]. This was also conducted in Australia, with 4 academic researchers and 1 team member from the Stroke Foundation also involved in the current project. These 2 Australian projects involved contributions from people with lived experience of stroke and included survivors of stroke in the authorship team; the current project convened a Lived Experience workgroup, and Pogrebnoy et al [50] consulted people with lived experience via a Consumer Advisory Group (1 survivor of stroke advised on both projects). Differences include that the website developed by Pogrebnoy was an adaptation of a newly created telehealth service, whereas in this study, the program and content were co-designed and integrated as part of the website build. Together, the projects demonstrate that collaborative ways of working are feasible, even when working with people who experience communication, cognitive, or physical challenges after their stroke. These ways of working are transferable to other clinical fields, and the methods reported in this study could be used by other research and design teams.

Limitations

By nature of engagement, wherein meetings were conducted online via Microsoft Teams, the lived experience workgroup members who contributed over the duration of the project tended to have high levels of digital literacy. This was a proof-of-concept project, and there was no specific strategy to recruit people from linguistically or culturally diverse backgrounds. Several team members (on the Steering group and Lived Experience Workgroup) were born in countries other than Australia, but most team members spoke English as a first language. Should the website indicate promise at improving outcomes for survivors of stroke, future work will be undertaken to engage with culturally and linguistically diverse groups of survivors and carers, and with people with lower digital literacy to determine their interest in accessing and using a website to promote self-efficacy. Partnership with the community leaders would be sought so that similar methods as have been described in this study could be used to understand and address the needs and preferences of groups of people that have not yet been represented on the EmpowerMe website.

Conclusions

We have demonstrated how a multidisciplinary team of academic researchers, people with lived experience of stroke, and Stroke Foundation representatives can work collaboratively from the inception of a project to co-design a digital resource known as the “EmpowerMe” website, which hosts content to meet the needs of survivors with stroke to support self-efficacy. Our team was convened to answer a question raised by people with lived experience of chronic health conditions, the team comprised people with varied skills and experiences (including people with lived experience of stroke and Stroke Foundation representatives), we convened a lived experience advisory group which we consulted over the course of the project to ensure the resource met the needs of the intended user group. Finally, we had clear roles and responsibilities for the steering group, the lived experience workgroup, and the usability testing participants. In this way, this project can serve as a reference for other co-designed digital health resources and offer practical lessons for similar initiatives.

Acknowledgments

The authors acknowledge the input of members of the Lived Experience Workgroup: Toni Arfaras, Kim Beesley, Brian A. Beh, Saran Chamberlain, Alison Douglas, Emma Gee, Clive Kempson, Letisha Living, Peter Meers, Susan Mueller, Veraina Ryan, and Mabel Ventura.

Funding

This project was funded by a grant from the Australian Government (National Health and Medical Research Council 2021 Medical Research Future Fund Consumer-Led Application 2022828). The sponsor determined the project would involve the development of a digital resource to promote health self-efficacy, but has had no input into the research design or conduct.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

Conceptualization (EAL, AO'M, BB, DC, LC, EG, NL, LM, NB, CE, ZA-W, Robyn Clark, Richard Cullen, GH, FJ, and AM), data curation (EAL, SC, AO'M, KN, SL, KN, and KZ), formal analysis (LC, EAL, and KN), funding acquisition (EAL, AO'M, BB, DC, LC, EG, NL, LM, NB, CE, ZA-W, Robyn Clark, Richard Cullen, GH, FJ, and AM), investigation (EAL, SC, AO'M, KN, EG, and NB), methodology (EAL, NL, LC, CE, Richard Cullen, NB, KN, and EG), project administration (KN and EAL), resources (SL, KN, EAL, KZ, and SC), software (NB and JL), supervision (EAL, CE, and Richard Cullen) visualization (AO'M, SC, Richard Cullen, NB, JL, EAL, and KN) writing – original draft (EAL, NB, and KN), writing – review and editing (AO'M, BB, DC, LC, EG, NL, LM, CA, ZA-W, RC, RC, GH, FJ, and AM). All authors have contributed to the shaping of the website through Steering group and project group discussions.

Conflicts of Interest

NB is cofounder and shareholder of CareMappr Pty Ltd. The other authors declare no conflicts of interest.

Multimedia Appendix 1
Usability testing script and list of customizable features.
[DOCX File , 22 KB - jmir_v28i1e76756_app1.docx]

References

1. GBD 2023 Causes of Death Collaborators. Global burden of 292 causes of death in 204 countries and territories and 660 subnational locations, 1990-2023: a systematic analysis for the global burden of disease study 2023. Lancet 2025;406(10513):1811-1872 [FREE Full text] [doi: 10.1016/S0140-6736(25)01917-8] [Medline: 41092928]
2. GBD 2023 DiseaseInjuryRisk Factor Collaborators. Burden of 375 diseases and injuries, risk-attributable burden of 88 risk factors, and healthy life expectancy in 204 countries and territories, including 660 subnational locations, 1990-2023: a systematic analysis for the global burden of disease study 2023. Lancet 2025;406(10513):1873-1922 [FREE Full text] [doi: 10.1016/S0140-6736(25)01637-X] [Medline: 41092926]
3. Barlow J, Wright C, Sheasby J, Turner A, Hainsworth J. Self-management approaches for people with chronic conditions: a review. Patient Educ Couns 2002;48(2):177-187. [doi: 10.1016/s0738-3991(02)00032-0] [Medline: 12401421]
4. Ahn S, Basu R, Smith ML, Jiang L, Lorig K, Whitelaw N, et al. The impact of chronic disease self-management programs: healthcare savings through a community-based intervention. BMC Public Health 2013;13:1141 [FREE Full text] [doi: 10.1186/1471-2458-13-1141] [Medline: 24314032]
5. Rice RE. Influences, usage, and outcomes of Internet health information searching: multivariate results from the Pew surveys. Int J Med Inform 2006;75(1):8-28 [FREE Full text] [doi: 10.1016/j.ijmedinf.2005.07.032] [Medline: 16125453]
6. Cheng C, Dunn M. Health literacy and the internet: a study on the readability of Australian online health information. Aust N Z J Public Health 2015;39(4):309-314 [FREE Full text] [doi: 10.1111/1753-6405.12341] [Medline: 25716142]
7. Yeung AWK, Torkamani A, Butte AJ, Glicksberg BS, Schuller B, Rodriguez B, et al. The promise of digital healthcare technologies. Front Public Health 2023;11:1196596 [FREE Full text] [doi: 10.3389/fpubh.2023.1196596] [Medline: 37822534]
8. Madanian S, Nakarada-Kordic I, Reay S, Chetty T. Patients' perspectives on digital health tools. PEC Innov 2023;2:100171 [FREE Full text] [doi: 10.1016/j.pecinn.2023.100171] [Medline: 37384154]
9. Senbekov M, Saliev T, Bukeyeva Z, Almabayeva A, Zhanaliyeva M, Aitenova N, et al. The recent progress and applications of digital technologies in healthcare: a review. Int J Telemed Appl 2020;2020:8830200 [FREE Full text] [doi: 10.1155/2020/8830200] [Medline: 33343657]
10. Golinelli D, Boetto E, Carullo G, Nuzzolese AG, Landini MP, Fantini MP. Adoption of digital technologies in health care during the COVID-19 pandemic: systematic review of early scientific literature. J Med Internet Res 2020;22(11):e22280 [FREE Full text] [doi: 10.2196/22280] [Medline: 33079693]
11. Pong C, Tseng RMWW, Tham YC, Lum E. Current implementation of digital health in chronic disease management: scoping review. J Med Internet Res 2024;26:e53576 [FREE Full text] [doi: 10.2196/53576] [Medline: 39666972]
12. Bashi N, Fatehi F, Mosadeghi-Nik M, Askari MS, Karunanithi M. Digital health interventions for chronic diseases: a scoping review of evaluation frameworks. BMJ Health Care Inform 2020;27(1):e100066 [FREE Full text] [doi: 10.1136/bmjhci-2019-100066] [Medline: 32156751]
13. Li L, Scott CA, Rothwell PM, Oxford Vascular Study. Trends in stroke incidence in high-income countries in the 21st century: population-based study and systematic review. Stroke 2020;51(5):1372-1380 [FREE Full text] [doi: 10.1161/STROKEAHA.119.028484] [Medline: 32208842]
14. Global stroke fact sheet 2022. World Stroke Organisation. 2022. URL: https://www.world-stroke.org/assets/downloads/WSO_Global_Stroke_Fact_Sheet.pdf [accessed 2026-02-10]
15. Langhorne P, Bernhardt J, Kwakkel G. Stroke rehabilitation. Lancet 2011;377(9778):1693-1702. [doi: 10.1016/S0140-6736(11)60325-5] [Medline: 21571152]

XSL•FO
RenderX

16. El Husseini N, Katzan IL, Rost NS, Blake ML, Byun E, Pendlebury ST, et al. Cognitive impairment after ischemic and hemorrhagic stroke: a scientific statement from the American Heart Association/American Stroke Association. Stroke 2023;54(6):e272-e291 [FREE Full text] [doi: 10.1161/STR.0000000000000430] [Medline: 37125534]
17. De Cock E, Batens K, Hemelsoet D, Boon P, Oostra K, De Herdt V. Dysphagia, dysarthria and aphasia following a first acute ischaemic stroke: incidence and associated factors. Eur J Neurol 2020;27(10):2014-2021. [doi: 10.1111/ene.14385] [Medline: 32515514]
18. Alghamdi I, Ariti C, Williams A, Wood E, Hewitt J. Prevalence of fatigue after stroke: a systematic review and meta-analysis. Eur Stroke J 2021;6(4):319-332 [FREE Full text] [doi: 10.1177/23969873211047681] [Medline: 35342803]
19. Lee J, Worthy P, Deslandes R, Burton B, Copland DA, Jamieson P, et al. Bridging the digital divide for people with aphasia: a study protocol for codesigning web accessibility tools and guidelines. BMJ Open 2025;15(8):e099273 [FREE Full text] [doi: 10.1136/bmjopen-2025-099273] [Medline: 40784777]
20. Niyomyart A, Ruksakulpiwat S, Benjasirisan C, Phianhasin L, Nigussie K, Thorngthip S, et al. Current status of barriers to mHealth access among patients with stroke and steps toward the digital health era: systematic review. JMIR Mhealth Uhealth 2024;12:e54511 [FREE Full text] [doi: 10.2196/54511] [Medline: 39173152]
21. Brady MC, Fredrick A, Williams B. People with aphasia: capacity to consent, research participation and intervention inequalities. Int J Stroke 2013;8(3):193-196. [doi: 10.1111/j.1747-4949.2012.00900.x] [Medline: 23130972]
22. Taylor JS, DeMers SM, Vig EK, Borson S. The disappearing subject: exclusion of people with cognitive impairment and dementia from geriatrics research. J Am Geriatr Soc 2012;60(3):413-419. [doi: 10.1111/j.1532-5415.2011.03847.x] [Medline: 22288835]
23. Archived grant opportunity view - GO5353. Grant Connect. 2025. URL: https://www.grants.gov.au/Go/Show?GoUuid=00ba3a66-86cd-4b6d-b880-3f00cb9e5152 [accessed 2026-02-10]
24. Bandura A. Social cognitive theory: an agentic perspective. Annu Rev Psychol 2001;52:1-26. [doi: 10.1146/annurev.psych.52.1.1] [Medline: 11148297]
25. Seixas AA, Olaye IM, Wall SP, Dunn P. Optimizing healthcare through digital health and wellness solutions to meet the needs of patients with chronic disease during the COVID-19 era. Front Public Health 2021;9:667654 [FREE Full text] [doi: 10.3389/fpubh.2021.667654] [Medline: 34322469]
26. Slattery P, Saeri AK, Bragge P. Research co-design in health: a rapid overview of reviews. Health Res Policy Syst 2020;18(1):17 [FREE Full text] [doi: 10.1186/s12961-020-0528-9] [Medline: 32046728]
27. Hevner A, March S, Park J, Ram S. Design science in information systems research. MIS Quarterly 2004;28:75-105 [FREE Full text]
28. Andrew NE, Kilkenny M, Naylor R, Purvis T, Lalor E, Moloczij N, et al. Understanding long-term unmet needs in Australian survivors of stroke. Int J Stroke 2014;9Suppl A100:106-112. [doi: 10.1111/ijs.12325] [Medline: 25042019]
29. Utley M, Gallivan S, Mills M, Mason M, Hargraves C. A consensus process for identifying a prioritised list of study questions. Health Care Manag Sci 2007;10(1):105-110. [doi: 10.1007/s10729-006-9003-6] [Medline: 17323658]
30. Maqbool B, Herold S. Potential effectiveness and efficiency issues in usability evaluation within digital health: a systematic literature review. Journal of Systems and Software 2024;208:111881 [FREE Full text] [doi: 10.1016/j.jss.2023.111881]
31. Talevski J, Wong Shee A, Rasmussen B, Kemp G, Beauchamp A. Teach-back: a systematic review of implementation and impacts. PLoS One 2020;15(4):e0231350 [FREE Full text] [doi: 10.1371/journal.pone.0231350] [Medline: 32287296]
32. Sauro J. A Practical Guide to Measuring Usability. Denver, USA: Measuring Usability LLC; 2011.
33. Find everything you need to live well after stroke. Stroke Foundation, Enable me. 2025. URL: https://enableme.org.au/ [accessed 2026-02-10]
34. Lynch EA, Nesbitt K, Gulyani A, Chan RJ, Bidargaddi N, Cadilhac DA, et al. Do self-management interventions improve self-efficacy and health-related quality of life after stroke? A systematic review. Int J Stroke 2025;20(7):786-800 [FREE Full text] [doi: 10.1177/17474930251340286] [Medline: 40275483]
35. Delivering quality stroke care. Stroke Foundation, Inform Me. 2025. URL: https://informme.org.au/ [accessed 2026-02-10]
36. Kennedy A, Cosgrave C, Macdonald J, Gunn K, Dietrich T, Brumby S. Translating co-design from face-to-face to online: an Australian primary producer project conducted during COVID-19. Int J Environ Res Public Health 2021;18(8):4147 [FREE Full text] [doi: 10.3390/ijerph18084147] [Medline: 33919920]
37. Fails JA, Ratakonda DK, Koren N, Elsayed-Ali S, Bonsignore E, Yip J. Pushing boundaries of co-design by going online: lessons learned and reflections from three perspectives. Int J Child Comput Interact 2022;33:100476 [FREE Full text] [doi: 10.1016/j.ijcci.2022.100476] [Medline: 36034969]
38. Istanboulian L, Rose L, Yunusova Y, Dale C. Adapting co-design methodology to a virtual environment: co-designing a communication intervention for adult patients in critical care. Res Involv Engagem 2023;9(1):103 [FREE Full text] [doi: 10.1186/s40900-023-00514-6] [Medline: 37957776]
39. Sanders EBN, Shen Y. Direct vs. remote data collection methods in co-design: navigating the evolving landscape. CoDesign 2025:1-15 [FREE Full text] [doi: 10.1080/15710882.2025.2472172]
40. Ramage ER, Burke M, Galloway M, Graham ID, Janssen H, Marsden DL, et al. Fit for purpose. Co-production of complex behavioural interventions. A practical guide and exemplar of co-producing a telehealth-delivered exercise intervention for

XSL•FO
RenderX

people with stroke. Health Res Policy Syst 2022;20(1):2 [FREE Full text] [doi: 10.1186/s12961-021-00790-2] [Medline: 34980156]
41. Mallakin M, Dery C, Vaillancourt S, Gupta S, Sellen K. Web-based co-design in health care: considerations for renewed participation. Interact J Med Res 2023;12:e36765 [FREE Full text] [doi: 10.2196/36765] [Medline: 36595738]
42. Morse B, Soares A, Ytell K, DeSanto K, Allen M, Holliman BD, et al. Co-design of the transgender health information resource: web-based participatory design. J Particip Med 2023;15:e38078 [FREE Full text] [doi: 10.2196/38078] [Medline: 36626222]
43. Wright M, Thomas KC, Carpenter D, Lee C, Coyne I, Garcia N, et al. Co-designing a website with and for youth, so they can better manage their health. PEC Innov 2023;2:100164 [FREE Full text] [doi: 10.1016/j.pecinn.2023.100164] [Medline: 37384159]
44. Brown M, Lord E, John A. Adaptation of ACTivate your wellbeing, a digital health and well-being program for young persons: co-design approach. JMIR Form Res 2023;7:e39913 [FREE Full text] [doi: 10.2196/39913] [Medline: 37052994]
45. Sanz MF, Acha BV, García MF. Co-design for people-centred care digital solutions: a literature review. Int J Integr Care 2021;21(2):16 [FREE Full text] [doi: 10.5334/ijic.5573] [Medline: 33981193]
46. Kvale S. Dominance through interviews and dialogues. Qualitative Inquiry 2006;12(3):480-500 [FREE Full text] [doi: 10.1177/1077800406286235]
47. Ørngreen R, Levinsen K. Workshops as a research methodology. Electron J e-Learn 2017;15(1):70-81 [FREE Full text]
48. Santin O, McShane T, Hudson P, Prue G. Using a six-step co-design model to develop and test a peer-led web-based resource (PLWR) to support informal carers of cancer patients. Psychooncology 2019;28(3):518-524. [doi: 10.1002/pon.4969] [Medline: 30597666]
49. Donald M, Beanlands H, Straus SE, Smekal M, Gil S, Elliott MJ, et al. A web-based self-management support prototype for adults with chronic kidney disease (My Kidneys My Health): Co-design and usability testing. JMIR Form Res 2021;5(2):e22220 [FREE Full text] [doi: 10.2196/22220] [Medline: 33560245]
50. Ashton L, Beh BA, Burke M, Cullen R, Czerenkowski J, Davey J, et al. Adapting a telehealth physical activity and diet intervention to a co-designed website for self-management after stroke: tutorial. J Med Internet Res 2024;26:e58419 [FREE Full text] [doi: 10.2196/58419] [Medline: 39437389]

Abbreviations

SUS: System Usability Scale

Edited by A Stone; submitted 30.Apr.2025; peer-reviewed by J Hao, Y Qian, T DÜzcÜ; comments to author 01.Dec.2025; revised version received 03.Feb.2026; accepted 04.Feb.2026; published 12.Mar.2026.

Please cite as:
Lynch EA, O'Malley A, Adey-Wakeling Z, Bonevski B, Cadilhac DA, Chamberlain S, Churilov L, Clark R, Cullen R, Godecke E, Harvey G, Jones F, Lannin NA, Larcombe S, Law J, McGrath A, Murphy L, Nesbitt K, Zacharia K, Bidargaddi N, English C
Creation of EmpowerMe Website to Promote Self-Efficacy in Survivors of Stroke: Co-Design Study
J Med Internet Res 2026;28:e76756
URL: https://www.jmir.org/2026/1/e76756
doi:10.2196/76756
PMID:

©Elizabeth A Lynch, Adrian O'Malley, Zoe Adey-Wakeling, Billie Bonevski, Dominique A Cadilhac, Saran Chamberlain, Leonid Churilov, Robyn Clark, Richard Cullen, Erin Godecke, Gillian Harvey, Fiona Jones, Natasha A Lannin, Stacy Larcombe, Jarrad Law, Annette McGrath, Lisa Murphy, Katie Nesbitt, Karly Zacharia, Niranjan Bidargaddi, Coralie English. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Insights and Recommendations From Moderators and Community Members for Keeping Online Peer Support Safe: Thematic Analysis

Hannah Grace Jones[1,2,3], BSc; Grace Lavelle[1], PhD; Elly Aylwin-Foster[1], BA; Ciara Regan[4], BSc; Alan Simpson[4,5], PhD; Ewan Carr[6], PhD; Matthew Hotopf[1*], PhD; Vanessa Lawrence[4*], PhD

[1]Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[2]Lifelong Health, Persistent Pain Research Group, Hopwood Centre for Neurobiology, South Australian Health and Medical Research Institute, Adelaide, South Australia, Australia

[3]IIMPACT (Innovation, IMPlementation and Clinical Translation) in Health, School of Allied Health and Human Performance, Adelaide University, Adelaide, South Australia, Australia

[4]Health Service and Population Research, Institute of Psychiatry, Psychology and Neuroscience, King’s College London, London, United Kingdom

[5]Florence Nightingale Faculty of Nursing, Midwifery & Palliative Care, King’s College London, United Kingdom, London, United Kingdom

[6]Department of Biostatistics & Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King’s College London, United Kingdom, London, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Hannah Grace Jones, BSc
Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neuroscience, King's College London
16 De Crespigny Park
London, SE5 8AF
United Kingdom
Phone: 44 02078480002
Email: hannah.g.jones@kcl.ac.uk

Abstract

Background: Online peer support can help people living with long-term physical health conditions to manage their mental well-being. Although the potential negative events that can occur and risks associated with web-based peer communities are well recognized, our understanding of how best to moderate these spaces is relatively limited, particularly with regard to new communities. Previous work has focused on the experiences of either moderators or community members.

Objective: This study aims to explore the perspectives of both members and moderators of a new online peer support community to evaluate the moderation procedures and inform recommendations for best practice.

Methods: Community members (n=39) who participated in a research trial of a new online peer community, CommonGround, were interviewed. The moderation team (n=5) was invited to a focus group. Community member interviews explored their opinions of moderation policies and the behavior of the moderation team. The moderator focus group explored their experiences of moderating the community, including perceived benefits, common challenges, and areas for improvement. All interviews and the focus group were conducted online, audio-recorded, and transcribed verbatim. An inductive thematic analysis was conducted to sort the data into overarching themes through an iterative process.

Results: Effective moderation was considered critical in creating a safe space that members wanted to engage with and for mitigating any risks, particularly around the spread of medical misinformation. Both moderators and community members felt that the moderation policies and practices were appropriate and applicable to the community. Moderators found navigating the moderation threshold, where they balanced safety against free speech, challenging when determining whether to intervene or not. Being part of a team with mixed clinical expertise helped moderators build confidence in navigating this threshold and also presented other benefits of easy access to support and improving the consistency of their moderation practices. It was suggested that in order for a community to flourish, community members would self-moderate. However, moderators and members felt that the strong community culture and high levels of member engagement that are needed to support self-moderation had not yet evolved. Proposed improvements to moderation included new features to support the efficiency of identifying new content for review and reviewing the rule of anonymity.

XSL•FO
RenderX

Conclusions: Moderation is critical in making online peer communities feel safe and engaging. Moderation practices should be co-produced with the target audience to ensure that they are aligned with the community's unique moderation wants and needs, including clear escalation pathways, transparent communication patterns, and plans to review and update policies or procedures as the community evolves. There should be technological features that promote self-moderation, as the community may shift towards self-moderation as it matures. It is also critical to ensure that moderators feel supported so that they are best placed to support the broader community.

Trial Registration: ClinicalTrials.gov NCT06222346; https://clinicaltrials.gov/study/NCT06222346

International Registered Report Identifier (IRRID): RR2-10.2196/71513

(J Med Internet Res 2026;28:e81943) doi:10.2196/81943

KEYWORDS

chronic conditions; depression; long-term physical health conditions; mental wellbeing; moderation; online peer support; online safety

Introduction

Connecting with peers or other people "like me" who have similar life experiences can be a powerful way of receiving social support and managing one's mental well-being [1,2]. By engaging with peers, people can exchange emotional support and their experiential knowledge [3]. In the last 10 years, there has been a rise in the use of online peer support, with the web-based setting facilitating convenient and flexible engagement that overcomes any physical barriers to accessing in-person support groups [4]. In particular, asynchronous peer community forums have become popular, as they provide the opportunity to exchange support with a potentially unlimited number of peers around the clock [5,6]. Those living with mental health conditions who have engaged with web-based peer support have experienced improved mental health symptoms, quality of life, and medical self-management skills [2,7,8]. Web-based peer support, therefore, offers a highly scalable, low-cost intervention that has the potential to promote mental well-being and reduce the burden of negative mental health on the individual, society, and health care system [4,9].

One population who may particularly benefit from online peer support is people living with long-term physical health conditions, as this population often reports experiencing psychological distress, loneliness, and disconnection from people in their social circles [10-12]. Around 15 million people in the United Kingdom live with a long-term physical health condition [13]. Of these, approximately 3 million are also living with symptoms of low mood that negatively impact their physical health [14], hinder effective self-management, and impair quality of life [15,16]. These symptoms of low mood also put them at risk of developing major depressive disorder, highlighting the critical need to intervene early to mitigate the impacts of mental-physical comorbidities on individuals, families, society, and health services [17-19]. Despite initial evidence documenting the psychosocial benefits of engaging with online peer support for people with long-term conditions [1,20-24], the provision of peer support specifically tailored to the needs of this population is relatively limited [25,26]. Our own qualitative research with people with diverse long-term physical health conditions found that traditional condition-specific communities can limit their ability to connect on other commonalities beyond their diagnosis and may not be appropriate for those with rare or multiple conditions [27,28]. To address this unmet need, we worked alongside people living with long-term health conditions, clinicians, and software developers in an iterative coproduction process to create CommonGround: an online peer support and psychoeducation platform that aims to intervene early to prevent the progression of symptoms of low mood to major depressive disorder [29]. By placing the voices of people with long-term health conditions in the foreground, we have tailored the intervention design and features specifically to their needs, including the all-diagnoses-welcome ethos that reflects the view that many experiences and challenges transcend diagnoses.

While online peer support communities have the potential to enhance the well-being of their members, there are challenges and risks associated with them. People with long-term health conditions have shared concerns about the potential negative outcomes from engaging digitally with their peers and the need for web-based spaces to feel safe [27,30,31]. Web-based forums carry inherent risks of medical misinformation and the sharing or encouragement of risky and unsafe behaviors [9,31]. Hearing about peers' experiences can potentially exacerbate low mood and promote symptom-comparisons that generate anxiety about one's condition [27,32-34]. Behind the screen of anonymity, people can argue, have hostile interactions, and experience cyberbullying or trolling [35,36]. It is well acknowledged that appropriate community governance is essential in reducing the risk of these negative outcomes arising, particularly among communities of people with long-term health conditions who are likely to discuss sensitive, personal, and medical-related topics.

One means of mitigating the risks of online peer support is through moderation by professional moderators. Clear, consistent, and comprehensive moderation policies play an essential role in transparently curating trust and safety in the community. By clearly outlining boundaries and acceptable behavior, community members can make judgments on whether the community is a space they wish to attend and have a clear understanding of how to behave within the community. Professional moderators can play an essential role in implementing these policies, including identifying and removing inappropriate content, preventing conflicts between members, enforcing the community guidelines, and supporting those in potential crisis [37-39]. Often, these moderators consider

XSL•FO
RenderX

themselves as a safety net and draw on their expertise to identify and eliminate hazardous scenarios that could cause harm, such as the sharing of unsafe or fraudulent medical advice and promotions of self-harm or suicide [9,37]. In addition to their safeguarding role, moderators can help establish a supportive, thriving community by creating a safe space for personal disclosures and fostering positive relations between members [40]. Indeed, moderated communities have been found to have higher engagement, perceived social support, and acceptability to members than unmoderated communities [41]. The impact of moderators within a community has also been noted in real-world settings where shifts in community moderation happen. For instance, following the departure of moderators from an established web-based health community, community members felt reluctant to engage with the community and recognized that the professional moderators had played a unique role in fairly and consistently implementing the rules and regulations [31]. Appropriate moderation is therefore often considered essential in laying the foundations of safety and facilitating the success of web-based communities. In the creation of CommonGround, the voices of lived experience experts were central to determining the Community Principles (“terms of use” community members agree to) and the associated governing policies to ensure that they remained aligned with their expectations of moderation and priorities for safety.

Much work has focused on how specific moderation issues are handled in well-established communities [9,30,31], yet there is limited guidance or recommendations on how to integrate moderation practices into a newly founded peer support community [37]. In addition, even though their experiences are inextricably intertwined, many studies consider the experiences of either the moderator [9,30] or the community member in isolation [31], and do not synthesize these perspectives to create a comprehensive understanding of what best practice might look like. Whilst moderators and community members participate with communities for different reasons, they coexist in the same space and their experiences are consistently shaped by the presence and behavior of one another. Moderators help establish the community culture and the boundaries for participation, while the behavior and responses of community members, in turn, influence what moderators do (eg, the content they have to moderate) and how they adapt their moderation responsively to suit the needs of the community. In this way, the experiences and behaviors of moderators and communities are continually shaping one another over time and cannot occur independently of one another. We therefore explored moderation experiences within the newly formed CommonGround community from the perspective of both community members and moderators. Our objective was to evaluate the feasibility and acceptability of the moderation procedures and inform recommendations for best practice in moderating newly forming web-based peer support communities.

Methods

Design

This work exists within the wider context of the CommonGround project, which aimed to coproduce an online peer support and psychoeducation platform and assess the feasibility and acceptability of the intervention in a randomized controlled trial (fRCT [feasibility randomized controlled trial] [42]; ClinicalTrials.gov: NCT06222346). The CommonGround project is underpinned by a pragmatic mixed methods research paradigm with community participation, which prioritizes what works, for whom, and in what context. The fRCT was delivered remotely and randomized 83 participants with long-term physical health conditions and subthreshold depression to the CommonGround intervention (ie, community members) and 42 participants to the control condition of National Health Service Mental Health webpages for 3 months (findings published in full elsewhere; in press). A group of 187 people with long-term physical health conditions who were interested in managing their mental health better, but did not have subthreshold depression, were also onboarded onto the CommonGround platform to increase the size of the community (community engagement cohort [29]). A nested qualitative process evaluation examined the experiences of community members with long-term physical health conditions and moderators. The qualitative process evaluation adopts an interpretivist-constructivist approach that assumes realities are socially constructed through interactions with others in the cultural context in which they are situated. This paper is reported within the SRQR (Standards for Reporting Qualitative Research) guidelines as per EQUATOR (Enhancing the Quality and Transparency of Health Research) recommendations (Multimedia Appendix 1 [43]).

Ethical Considerations

This study involves data from the nested qualitative study of part of a larger feasibility randomized control trial that was conducted remotely at King’s College London (KCL). The trial and nested qualitative study received ethical approval from the South Central - Oxford C Research Ethics Committee (IRAS ID 328175). The protocol was preregistered (ClinicalTrials.gov: NCT06222346) [29] and outlined the larger feasibility randomized controlled trial and the nested qualitative study (ie, community member and control group interviews, moderator focus group). This project is sponsored and organized by King’s College London. The Research Governance Office can be contacted via email at rgo@kcl.ac.uk.

The data from the nested qualitative study were collected in alignment with the original protocol aims of understanding participant perspectives of the intervention, control condition, and trial procedures generally (published elsewhere [29]) and for understanding perspectives and experiences of moderation specifically (this study).

Informed consent was obtained from all community members at two stages: (1) at T_1 when completing the 2 eligibility screenings, and if found to be eligible for the trial, (2) at T_2 when they consented to participate in the trial. All community members had provided consent to take part in the feasibility randomized controlled trial at enrolment. At enrolment, they were informed that they may be invited to interview after the 3-month intervention period, and that additional consent would be sought at that time if selected. Consent was recapped for

XSL•FO
RenderX

those who were invited to interview prior to their agreement to interview.

Community members and moderators were invited to be interviewed or take part in the focus group, respectively, after the intervention period. Five days after the invite, they were contacted again to discuss any questions about what taking part would involve before completing a web-based consent form (via Qualtrics) and scheduling the interview. Participants were allowed more time to consider their participation if needed and asked to contact the research team should they decide they wished to participate.

At the beginning of the interviews and focus group, the information about what the session would involve was repeated verbally, with participants reminded of their right to withdraw and to decline to answer any question, at any time. Community members were reimbursed £20 (US $26.93) for their participation, and moderators were reimbursed in alignment with their hourly wage.

The CommonGround Platform

CommonGround was codesigned in collaboration with people with lived experience of long-term conditions or physical-mental health comorbidities, clinicians, academics, and software developers [28]. The anonymous web-based platform was open to all diagnoses and could be accessed via desktop or mobile devices. The key features included a discussion forum with different posting options, reactions, and the ability to comment (peer support component), alongside a resources page of evidence-based self-help information (psychoeducation component). The self-help information consisted of 91 resources that were hand-selected by the research team from reliable, reputable sources and covered various topics ranging from keeping active, managing mental well-being, and managing finances. The resources were selected for their relevance to people living with any long-term physical health condition (ie, were not condition-specific). The CommonGround platform features, functions, and layout remained static throughout the intervention period with one minor exception. During the first week, the software developers made a minor edit to how the latest activity in the community forum appeared. Initially, at launch, the newest post was displayed at the top of the community forum feed. This was changed so that the post with the most recent activity displayed at the top of the feed (ie, existing post with the most recent comment or new post). Further details of the features of CommonGround and screenshots of the platform are available in Multimedia Appendix 2.

Moderation Policies and Practices

The CommonGround platform was governed by the Community Principles, Moderation Policy, Safeguarding Policy, and Privacy Policy, which were all published on the platform (and are available on the CommonGround projects Open Science Framework webpage [44]). These policies were coproduced in collaboration with our research advisory group (10 members living with a range of long-term health conditions, including 6 females and 4 males (self-reported sex), aged between 25 and 73 years), software developers, and the Samaritans Online Harms Team [45]. Key principles of the moderation approach included that the platform was anonymous, was not moderated 24/7, and did not provide one-to-one or crisis support. The moderation team also had access to a moderator handbook, developed in consultation with a leading disability charity [46], which included guidance on how to moderate different scenarios. The CommonGround platform had an admin backend that allowed moderators to edit and delete posts, delete comments, review flagged posts, and escalate posts for further action. Community members were able to view all policies and "flag" posts for review by moderators. Prior to the launch of the community, moderators were trained by the research team on how to use CommonGround and how to implement the governing policies. The moderator training was developed based on consultations with the community manager at SCOPE [47]. Each moderation session was expected to take approximately 60 minutes per day, and to occur at some time between the hours of 8 AM and 8 PM. Moderators were able to choose whether to moderate across a single or multiple sessions, and exactly when to moderate, dependent on the needs of the community (eg, when there was a moderation concern to monitor) and their personal schedules. Moderators were required to complete a handover after each moderation session of outstanding tasks, posts, or comments to watch, and any other comments. Moderators were instructed to review the previous moderation handover at the beginning of their session and to complete any outstanding tasks as a priority. Then moderators were instructed to review new content and complete any moderation tasks as needed (ie, in response to what had been posted). These moderation tasks may have involved any (or all) of the following: edit posts, delete posts, delete comments, review posts with attachments, review flagged posts, comment on posts to guide discussions toward a productive versus harmful narrative, and contact users to issue warnings. A Microsoft Teams chat was also created to support shared decision-making and provide updates to the moderation team. All moderators were of equal status and consulted with the research team (HGJ, GL, and MH) as required for additional support or to resolve any queries they had on the best course of action. For instance, if a moderator was unsure about whether to moderate or not (eg, whether to delete a comment, an action that could not be reversed), they were encouraged to discuss this with the team to come to a collective decision. If required, any decisions to issue warnings to community members (part of the "three strike" policy) or to remove community members from CommonGround would be made by joint decision of the moderators and research team. At any time, moderators could request an informal debrief with the research team, which included a liaison psychiatrist. A feedback session for moderators to share experiences was conducted halfway through the 3-month intervention period.

There was a separate engagement team whose exclusive focus was on cultivating engagement on CommonGround by sharing discussion topics to stimulate conversation and weekly circular emails with community updates and reminders to log in. The engagement team had no moderating or safeguarding responsibilities, but could use the "flag this post" button if they came across something they felt required attention (as a regular community member could also do). The engagement team consisted of our co-researcher with lived experience and 2 other research team members. The profiles of the Engagement Team

XSL•FO
RenderX

on the platform were clearly identifiable as being distinct from both regular community members and the Moderation Team via their avatar and account username. As the engagement team involved different team members to the moderation team, had unique roles and responsibilities, and were not involved in any moderation activity, the engagement team will not be discussed further in this manuscript.

Participants

CommonGround Community Members

CommonGround community members were recruited in the United Kingdom through multiple channels, including outpatient clinics within King's College Hospital and Guy's and St Thomas's Hospital, public advertisements via charities, and word-of-mouth. The inclusion/exclusion criteria for the feasibility trial are listed in Table 1 and were all self-reported by participants. As mentioned above, to increase the size of the community at the time that CommonGround was launched, people who did not have subthreshold depression (ie, scored 0-4 or 10-14 on 8-item Patient Health Questionnaire) but met all remaining inclusion criteria were also offered the opportunity to join CommonGround as part of a community cohort study (study protocol contains full details [29]). The only difference between participants of the feasibility trial and the community cohort was their 8-item Patient Health Questionnaire scores: feasibility trial participants scored 5-9, whereas community cohort scored either 0-4 or 10-14. Approximately 40 community members who had access to CommonGround (including ≈30 fRCT participants and ≈10 community engagement cohort) were invited to participate in an interview via email with a follow-up phone call. Community members were approached purposefully based on their long-term condition(s), age, gender, ethnicity, and their engagement level with the community. The engagement levels were defined as follows: (1) never joined (never registered their account or logged in), (2) never logged in (created an account but never logged in), (3) readers (logged in but never commented, reacted, or posted), (4) acknowledgers (logged in and commented or reacted, but did not post), and (5) content creators (logged in and created posts).

Table 1. Inclusion and exclusion criteria for the feasibility randomized controlled trial (fRCT) and community cohort.

	Inclusion criteria	Exclusion criteria
fRCT	• Currently living with one or more long-term conditions. A long-term condition was defined as a diagnosed physical medical condition with a duration of ≥6 months that demonstrates recurrence or deterioration and is typically associated with a poor prognosis [48]. There were no restrictions on the type of long-term condition. • Probable subthreshold depression, defined as scoring 5-9 on the PHQ-8[a]. • Age ≥18 years. • Have access to the internet (via phone, smart device, desktop, or laptop at any location, including at home, at the workplace, or free Wi-Fi in public spaces). • Have sufficient English proficiency to engage with the platform. • Ability to give informed consent (ie, no known circumstances/conditions that may affect the capacity to consent).	• Has ever received a clinical diagnosis of severe mental illness, including bipolar disorder, psychosis, posttraumatic stress disorder (PTSD), and schizophrenia. • Has ever received a clinical diagnosis of dementia.
Community cohort	• Currently living with one or more long-term conditions. A long-term condition was defined as a diagnosed physical medical condition with a duration of ≥6 months that demonstrates recurrence or deterioration and is typically associated with a poor prognosis [48]. There were no restrictions on the type of long-term condition. • Scoring 0-4 or 10-14 on the PHQ-8a. • Age ≥18 years. • Have access to the internet (via phone, smart device, desktop, or laptop at any location, including at home, at the workplace, or free Wi-Fi in public spaces). • Have sufficient English proficiency to engage with the platform. • Ability to give informed consent (ie, no known circumstances/conditions that may affect the capacity to consent).	• Has ever received a clinical diagnosis of severe mental illness, including bipolar disorder, psychosis, PTSD, and schizophrenia. • Has ever received a clinical diagnosis of dementia.

[a]PHQ-8: 8-item Patient Health Questionnaire.

Moderators

Moderators were recruited through KCL PhD cohorts and through professional connections of the research team, based on their clinical expertise in mental health treatment and experience of attending or moderating peer support groups. At the end of the trial period, all moderators (n=5) were invited via email to participate in a focus group.

Data Collection

Interviews

Demographic data were collected at baseline as part of the fRCT via anonymous Qualtrics questionnaires. Interviews were conducted after the 3-month intervention period (September 28, 2024, to December 8, 2024) by a member of the research team (CR, HGJ, GL) to explore individual experiences of taking part in the trial and of the CommonGround intervention. Topic

XSL•FO
RenderX

guides were coproduced with the research advisory group (Multimedia Appendix 3). The topic guides included the following open-ended questions around moderation: "What were your experiences, if any, of the Moderation Team on CommonGround?" and "What did you think about CommonGround being an anonymous platform?" All interviews lasted approximately 30-90 minutes and were conducted online via Microsoft Teams and audio-recorded, before being transcribed verbatim and anonymized for analysis.

Focus Group

Demographic data were collected via Qualtrics. A focus group was conducted (by a member of the research team, CR, who had no prior contact with the moderators during the trial) to stimulate debate and discussion around their experiences of moderating the platform, moderation policies and training, the positive and negative experiences they had observed community members have, and how moderation could be improved in the future. Moderators were given the opportunity to elaborate on any topic they felt was relevant that may not have been covered by the topic guides and were prompted to provide specific examples. The focus group was conducted via Microsoft Teams, audio-recorded, and transcribed verbatim and anonymized for analysis. The focus group lasted for approximately 90 minutes.

Data Analysis

Data from the community members (interviews) and moderators (focus group) were analyzed together using NVivo software (version 14, released 2023; Lumivero). An inductive thematic analysis was performed to sort the data into overarching themes that were developed inductively and from an interpretivist-constructivist position [49,50]. The first author (HGJ) familiarized herself with the entire dataset by reading and rereading the transcripts. Once immersed in the data, an initial list of codes was generated, which were constantly compared with one another to identify patterns of shared meaning that formed provisional themes. An experienced qualitative researcher (VL) reviewed the initial coding framework to help refine the codes and provide an additional perspective on interpreting the data. VL and HGJ also met regularly to discuss the analysis and explore different interpretations. The transcripts were then reviewed, and relevant extracts of the data were coded into NVivo nodes (the refined codes). Through an iterative process, codes were refined, added, and condensed as needed before being organized into themes. Following further discussions, a consensus was reached on the analysis, and the themes were named and defined. Throughout, our lived experience co-researcher (EA-F) ensured the coding framework was reflective of lived experience terminologies and definitions.

Results

Participant Characteristics

Community Members

A total of 39 participants who had been provided access to the CommonGround intervention were interviewed (29 fRCT participants, 10 Community Engagement participants), including 21 women, including transgender women, and 18 men, including transgender men, living with a wide variety of long-term health conditions (Table 2). Data saturation was reached after the 39 interviews, and no further interviews were conducted.

XSL•FO
RenderX

Table 2. Demographics of the community members interviewed.

Category	Overall (n=39)
Age (years), mean (SD)	53.2 (13.7)
Gender (n=39), n (%)	
Women, including transgender women	21 (53.8)
Men, including transgender men	18 (46.2)
Ethnicity, n (%)	
Asian or Asian British	—[a]
Black, Black British, Caribbean or African	—
Other ethnic group	—
Prefer not to say	—
White	31 (79.5)
Mixed or multiple ethnic groups	—
Engagement level with CommonGround (n=39), n (%)	
Never joined: never registered their account or logged in	6 (15.4)
Never logged in: created an account but never logged in	5 (12.8)
Readers: logged in but never comments, reacted, or posted	5 (12.8)
Acknowledgers: logged in and commented or reacted, but did not post	7 (17.9)
Content creators: logged in and created posts	16 (41.1)
Long-term physical health conditions: total, n	
Total number of unique LTCs[b] represented	31
Long-term physical health conditions: categories, n	
Autoimmune and inflammatory	11
Cancer	—
Cardiovascular	9
Gastrointestinal and digestive	18
Metabolic and endocrine	6
Musculoskeletal and chronic pain	9
Neurological (eg, Parkinsons)	13
Other	8
Reproductive and gynecological	—
Respiratory	—

[a]Represents <5 participants, removed to protect anonymity.

[b]LTC: long-term condition.

Moderators

The moderation team included 4 women and 1 man (self-reported gender), aged between 26 and 45 years. Three were clinical research fellows. One was a mental health nurse, one a pharmacist, one a clinical psychologist, and 2 were adult psychiatrists. Four had clinical experience in working with people with mental and physical health conditions, and one had experience working with people with mental health conditions only. Their years of clinical experience ranged from 6 to 19 years (mean 10, SD 5.1 years). All the moderators were currently completing PhDs.

Themes and Subthemes

Overview

Data from both groups of participants converged around five key themes: (1) moderation as a pillar of community safety, (2) the moderation blueprint, (3) the moderation threshold, (4) self-moderation is the (achievable) dream, and (5) being part of the moderation team. Table S1 in Multimedia Appendix 4 provides an overview of the themes and illustrative quotations.

Theme 1: Moderation as a Pillar of Community Safety

A widely shared opinion among moderators and community members was that moderation is essential: “moderation, it’s

XSL•FO
RenderX

just important. It's got to be there" (community member 6, acknowledger). In particular, moderation was thought to play a critical role in making the community feel safe; a foundational pillar needed for community members to share, interact, and, ultimately, provide peer support. Regardless of how aware members were of the specific moderation policies or day-to-day activities, the simple presence of the moderation team engendered a sense of safety that was considered unique to CommonGround.

> *It was being overseen or looked after, where, sometimes, on [well-established social media platform], there are some very good support groups, and there are some that are just absolute wild west, and there's people arguing and fighting on them, and they're not healthy places to be. So, there was a sense of safety.* [Community member 28, content creator]

The fact that the community was for people living with long-term physical health conditions augmented the need for moderation and safety. Community members wished to protect the space from people who did not have long-term conditions, as they felt that they were likely to share sensitive and personal information and experiences related to their health.

> *[Moderation is] really important, because we're all quite vulnerable. Those of us with chronic illnesses especially, we're all quite vulnerable, and of course, it comes under the whole Equalities Act 2011 and all that, with hate against disabled people being a hate crime, and all that kind of stuff.* [Community member 9, content creator]

The specific needs of this community also influenced moderator and community members' opinions on the key areas of moderation. For instance, they shared the view that potentially dangerous topics included the spread of medical misinformation and expressions of emotional crisis or self-harm that required follow-up.

> *There would be some things where you'd want to be on it really quickly, like if someone's posting about self-harm methods or how they want to end their life. Something like that, you don't want to wait.* [Moderator 3]

While it was felt that the community was a place to discuss medical and emotional experiences, there was shared concern that these topics require careful review by moderators to ensure that they are discussed in a safe way.

> *I think for a platform offering peer support, there needs to be some moderation to just make sure that, A, you don't go off down wormholes, that it's not taken over by individuals, the information that's posted is accurate, and that any conflict is dealt with quickly and effectively.* [Community member 16, acknowledger]

Theme 2: The Moderation Blueprint

Overview

This theme relates to the policies, community principles, moderator guidance, and training that underpinned the governance of CommonGround. Two subthemes emerged: (1) the importance of this blueprint and (2) the moderator's experience of translating the blueprint into practice.

Subtheme 2a: The Moderation Blueprint Is Essential

In general, community members were satisfied with the moderation and safeguarding policies that were on the platform, recognizing that they were critical for transparently outlining what is and is not acceptable and what to expect in different circumstances. They also provided reassurance that there was a safety net for worst-case scenarios. Typically, community members briefly reviewed the policies at sign-up, often indicating that their review was a relatively light touch and purely to ensure that there were no 'red flags' in the policy. Others made assumptions about the type of content based on their previous experiences or because they considered the owners of the platform (KCL) to be trustworthy.

> *So, the moderation rules are there, and they should be abided by, and people know if they don't, they're parked, and that's how it should be.* [Community member 11, content creator]

The moderators also felt that the comprehensive policies laid out clear, consistent boundaries for how the platform would be moderated. The moderator handbook, which provided detailed guidance on how to implement the policies, was valued for its specific examples of potential events, as this reduced some of the burden in working out how to respond, including the specific language and tone to use. The moderator handovers were also valuable for keeping up to date on recent activity and highlighting points to review in their moderation shift.

> *The guidance that we had was quite helpful in that it had quite specific examples for a lot of those scenarios that might come up. And then an example of how you might respond. I found that was quite helpful. And then I guess, as we started doing more of it, you could also see what other people had posted to similar things from the handover document. And then that would, I guess, we'd all respond in a similar way a lot of the time.* [Moderator 5]

One aspect of the moderation blueprint that split opinions among community members related to the anonymity of CommonGround, with mixed views on whether this constituted a necessary safety feature or stifled members' ability to establish genuine connections. Other community members felt indifferent, or that the best option would be allow members to select their level of anonymity and give individuals the power to disclose their personal information as they see fit. This call for greater flexibility in anonymity was also recognized by the moderators, who suggested that the moderation blueprint could, and should, be reviewed as the community evolves.

> *I think maybe going forward is probably about giving people a choice if they want to stay anonymous or not* [Community member 44, reader]

XSL•FO
RenderX

Subtheme 2b: Translating the Moderation Blueprint Into Practice

In general, the moderators agreed that their moderation shifts were quick, manageable, and quite light touch with minimal need to intervene, edit, or delete posts:

> *It was quite a quick process. Had a checkbox of, check this, check this, check this. And most things had nothing, you know, the flag post, delete posts. Most of them, you're just checking that there was nothing there.* [Moderator 1]

The moderators found the guidance in the moderation handbook useful and applicable to the content that came up, and were confident in applying it (Multimedia Appendix 5 provides an illustration of how comments posted by moderators mapped onto advice in the moderation handbook). It was clear that for moderators to successfully do their job, they needed to feel that it was easy to translate the moderation blueprint into the real-world setting of the web-based community. In general, the moderators felt that this translation was easy to do. The main challenge that moderators faced related to the ease and efficiency with which they identified what content was new since the last moderation session. As noted in the methods, during the trial itself, the software developers updated the forum feed layout to display the most recent activity (comment or post) at the top of the feed. The moderation team shared that this technical update helped to reduce the burden of identifying the new content during the trial and that it could be further developed to improve efficiency in a larger, future community. This initial frustration that moderators expressed at the platform-specific issue provided insight into the fact that if their basic need of easy navigation around the community content is unsatisfied, any form of moderation becomes difficult.

> *Initially, it was more of a system issue. So, we would have to look at comments and threads within the last 24 hours. But initially, the messages were coming up when they were posted, so we wouldn't see the new things immediately. So, having to monitor and try to look at what was new, was really difficult.* [Moderator 5]

Theme 3: The Moderation Threshold

A key experience that moderators spoke about related to navigating the moderation threshold: when to intervene and when to "wait and see" how the situation played out. As time went on, the moderators felt more confident in where the moderation threshold sat, as they learnt how the community worked and reflected on previous moderator activity.

> *I think when we started to notice that certain moderation styles would actually stifle the engagement from the community, it became quite clear that if you come back to what is the purpose of the platform, that then helps you to make a decision about whether you should do something or not. I guess, like a risk, reward thought process.* [Moderator 1]

Moderators and community members shared the view that navigating this threshold was particularly difficult around medical information.

> *[…] By giving space to talk about it, but also disclaiming that that's not a good source, and here's a better source, and having a more balanced opinion, contemplating what they're saying, but also say… It is a fine line, but by acknowledging what they're saying, whilst also posting the counterargument in, it allows people to see.* [Community member 6, acknowledger]

While the moderators considered their clinical experience useful overall, at times, their experience of dealing with clinical populations and fear of "worst case" scenarios could feel unhelpful. Some spoke of the need to recalibrate their threshold for stepping in to suit the current audience. A couple suggested that including people with lived experience of chronic illness in the moderation team could be beneficial, as they may be better placed to judge when an expression of emotion constitutes emotional distress that requires signposting to one-to-one support.

> *I wondered sometimes if our professional backgrounds made us more anxious about some of the comments that people made on the platform. As [moderator] said, sometimes you'd get such a small amount of information indicating that somebody is struggling with something. Then because, I guess, my professional background as a clinical psychologist, you see the real peak of distress in terms of the clients that you might work with, particularly when you're working in a physical health context. So then, I suppose, my mind might go to, this person's really distressed. And then you start thinking of all the people that you've worked with clinically, who've been distressed and whether you might need to intervene, and then in terms of managing risk. Whereas actually, it might be that the likelihood is that person's distressed, but it's a really a much lower level of distress than what I might see professionally, that's therefore, less in need of some intervention from the moderators.* [Moderator 1]

Moderators and community members shared the concern that intervening too early could stifle the community and prevent people from sharing freely. Moderators observed that when they commented, discussion among peers often stopped, leading to a hesitancy to intervene "too early" or unnecessarily. Community members suggested that over-moderation could create a strange dynamic in the community and stagnate conversation. It was also felt to have potential to propel some community members toward "rebellion," creating further issues.

> *I think stifling conversations just makes people more likely to try… Well, not even to try and cause trouble, but that's the end result, but it makes them more likely to have extreme views and to go further down the rabbit hole.* [Community member 6, acknowledger]

Despite the moderators' anxiety about finding the right threshold, community members often felt that the moderators

XSL•FO
RenderX

had struck the right balance. Some suggested that a more authoritarian dynamic may be needed, or necessary, at these earlier stages of community formation to ensure that ground rules are understood and the community grows in a safe manner.

> *I also felt that there was that kind of safeguarding element with the moderators just being in the background, but they weren't, it wasn't like a very controlling element to it. And they almost showed up as equal participants, almost, but just with that, oh, just a reminder of… It didn't feel like an enforcement group or anything.* [Community member 27, content creator]

Theme 4: Self-Moderation Is the (Achievable) Dream

This theme is related to the shared view that self-moderation was the ideal and the hope that this would evolve in the community over time. Self-moderation would include community members jumping in to moderate potentially problematic conversations and preventing any arguments from emerging. Community members and moderators observed that some members of the community not only regulated what they shared on CommonGround but also intervened in situations where people seemed to be entering potential conflict, share incorrect information, or be in need of emotional support. In interviews, community members commented on their responsibility to think before they post or jump into discussions.

> *We found after a while, that other users were self-moderating anyway, so we really didn't have to intervene that much, because by the end of the thread, it would have resolved.* [Moderator 2]

One feature that was built into the platform to support self-moderation was the “flag this post” function that prompted members to determine what they thought was inappropriate and flag it to the attention of the moderators for priority review. This feature was relatively unused by the community, but moderators recognized it as important infrastructure to allow the community to move toward self-moderation if there were to be a larger volume of member-generated content:

> *[...] One of the functions that was on the platform that wasn't used very much, was users at the platform actually flagging posts that they thought were inappropriate or distressing. That's one way that if it [CommonGround] got bigger and someone wasn't looking at every single post, things would still get identified and then would help to reinforce that ownership of the platform. Because then it's like, what do they [users] find unacceptable? What do they think is going against the principles? […] I think that would be good, maybe to see more of that.* [Moderator 5]

The opinion that self-moderation would be a superior model of moderation was in part driven by the moderators’ view that hearing from peers is much more powerful than hearing from a moderator.

> *I felt like the self-moderation was much more powerful in terms of influencing and changing opinion or getting people to think about things in a different way, because it came from a peer.* [Moderator 3]

However, handing the baton of responsibility over to the community members required careful navigation as moderators balanced their responsibility to keep the community safe and the desire to promote members to have ownership over the community:

> *[…] The learning from trial and error. Okay, like, this is something that we were concerned about. We were thinking that possibly it could go down that avenue of concern. And actually, what we saw was someone else commented. That ended up fine, we didn't need to do anything. So you learn from the experience, which I think helped us to feel more confident in leaving things be, a bit more, as well.* [Moderator 5]

Both community members and moderators recognized that a high level of engagement and strong bonds between members were necessary for the community to self-moderate. However, it was recognized that the CommonGround community did not seem to reach this, with more time needed for it to become established and flourish.

> *I think it's quite a short amount of time to form a community, especially online. And I think maybe there was the start of some of that. At times, certainly. Like what [moderator] was referring to earlier, when there were quite, very supportive responses to people who were clearly distressed and sharing of similar experiences. There were moments where it felt like there was good community support.* [Moderator 5]

One moderator noted that the inclusion of lived experience experts as part of the moderation team might help to encourage this shift toward self-moderation. When speaking about the future of moderation on the platform, community members noted that the moderation team would likely still be needed in the background. Some community members explicitly identified a continued responsibility of the professional moderators in ensuring platform safety, highlighting that certain scenarios demand their intervention:

> *Obviously, you’re relying on a moderator to block trolls, or whatever you call them.* [Community member 39, never logged in]

Theme 5: Being Part of the Moderation Team

This theme relates to the positive consequences moderators experienced as a result of being part of a team. The moderators described a sense of shared endeavor in moderating the CommonGround community and had a strong understanding of what their role as a team was, and how their role as a moderator (to keep people safe) was distinct from the engagement team (to promote peer support discussions). Being part of a team helped the moderators to balance the position alongside their other commitments and allow downtime away from the community.

> *I think we had the right number of people as well. It wasn't that you did it all the time. You did it a couple of times a week at most, which I think was about right. It'd be hard to do it every day, I think.* [Moderator 4]

XSL•FO
RenderX

The benefit of being part of a moderation team went beyond the practical aspects, allowing for emotional support and encouragement, and opportunities to debrief after tricky situations. One moderator shared that "[…] It felt very open door, in terms of people [the research team and other moderators] being approachable. If we had any questions, it felt very safe and supported" (Moderator 3). In some way, the moderation team's use of Microsoft Teams chat and handover documents became a form of peer support, with moderators able to seek advice and share how they felt about the role or the community at any time. One moderator recalled feeling well supported by fellow moderators and the research team after they received some negative feedback from a community member whose post they moderated:

> *[…] Then they wrote another one [post] saying, […] and was a bit shirty. But it was nice because I had a chat with the KCL research team as well, and they said, yes, you did the right thing, which was quite reassuring.* [Moderator 4]

Being part of a team also helped to build the moderators' confidence in the role, as they could learn from one another as they implemented moderation policies in a new community. Shared decision-making and reflection helped the team move toward a more efficient method of moderation.

> *[…] It was quite helpful to have other people dipping in, in between your moderation sessions, because then you get a different perspective on what's already come up. And if there are things that are on that boundary of, is this something we need to do something about or not. Then I found it quite helpful to have someone else's eyes on that. Just my own, I guess, reassurance that we're doing the right thing.* [Moderator 1]

The range of expertise within the moderator team was also beneficial, particularly when it came to certain topics that moderators were most concerned about, such as diagnosis-related information or medication. Here, it was understood that it was important for the community to be able to speak about it in a safe way.

> *I do think it was really helpful having some medically trained moderators amongst the group who could be a bit more aware of the comments about medications and treatments and things. Because that was something that I didn't really feel able to moderate that well, because I don't know very much about the medications and the medical treatments for these different health problems.* [Moderator 1]

Future Improvements for Moderation

As moderators reflected on their experiences of moderating the CommonGround platform, they highlighted areas where moderation could be improved in the future. Areas for potential improvements and evidencing quotes are presented in Table 3 below.

XSL•FO
RenderX

Table 3. Potential areas where the CommonGround moderation could be improved, with corresponding evidencing quotations.

Potential improvement to moderation	Quotes
General moderation improvements	
Include lived experience experts, or clinicians with relevant lived experience, on the moderation team	• I think it could have been quite good to have lived experience moderators as well, who would maybe... And maybe that would make a difference to the power dynamic a little bit, in terms of when there were moderator comments, it might be perceived differently? [Moderator 3] • I think having people who aren't healthcare professionals, who are more coming from the perspective of that lived experience might help to get a bit more balance there, possibly, in terms of, I guess, the psychological distress part of it. [Moderator 1] • If anything, I think getting more community ownership over the platform would be a positive, and maybe having lived experience moderators. But still, the community principles are great, and I think those things are really important. But I think to encourage more community ownership of the platform would help make it a more active platform. [Moderator 2]
Explore whether AI[a] or algorithms could aid moderation	• There are lots of massive platforms online that manage to moderate huge amounts of posts. And I guess the way they do that is using algorithms. And they must, because there's no way someone's looking at everything. So, if you scale up, that might be one way to do it. [Moderator 5]
Encourage the community towards self-moderation	• ...One of the functions that was on the platform that wasn't used very much, was users at the platform actually flagging posts that they thought were inappropriate or distressing. That's one way that if it [CommonGround] got bigger and someone wasn't looking at every single post, things would still get identified and then would help to reinforce that ownership of the platform. Because then it's like, what do they [users] find unacceptable? What do they think is going against the principles? […] I think that would be good, maybe to see more of that. [Moderator 4]
Platform-specific improvements	
Improve the feature(s) for identifying new content on the platform since the last moderation session	• I think the efficiency of the moderation could have been way better if you're flagged to anything new that's come up within 24 hours. […] I think it could be improved by having some way that the platform can flag all of these things are new in the last 24 hours. And then you just go through them, almost like a notification system. [Moderator 1] • […] You can almost just have all of them and then you tick them [new comments/posts] off, and you know that they've definitely been moderated. That would be the most efficient way, probably. [Moderator 5] • I think with the addition of the notification system, making it super, super clear, what's new, would help it to be moderated. It would help the moderator, if there were more users. I don't feel like it's necessary for the moderator to see everything, because so much of it is just really positive and helpful and supportive, and day to day chats about books and movies and things like that that don't really need a moderator. [Moderator 1]
Review the community principle of anonymity	• It felt to me that there were a few people who wanted it to be less anonymous. And what "moderator" said earlier, about somebody who wasn't very happy about their personal identifying information being taken out of their posts. I wonder if creating that sense of community for some people is easier, if you are able to share a little bit more information about who you are and what part of the country you live in, and things like that. [Moderator 1]

[a]AI: artificial intelligence.

Discussion

Principal Findings

Online peer support promises a low-cost, easily accessible, and scalable way to support mental well-being and self-management among people living with long-term physical health conditions [1,24]. The risks and potential adverse events that can occur when engaging with peers via web-based communities are well recognized [9,27,31-34]; yet, research into how best to moderate communities to keep them safe is limited, particularly in the context of new web-based health communities [37]. In this work, we explored the perspectives and experiences of both community members and moderators together to evaluate the acceptability and feasibility of our co-designed moderation procedures and identify areas for future improvements. Overall, community members and moderators both had positive views of our moderation procedures. Moderators found it feasible to implement the polices but identified opportunities to improve moderation as the community culture evolves. Based on these findings, we provide recommendations to guide the development and implementation of moderation practices for new communities (Table 4).

XSL•FO
RenderX

Table 4. Recommendations on how to moderate web-based peer support communities.

Recommendation	Description
1. Coproduce clear moderation policies that reflect best practice	• Develop moderation in collaboration with people who have lived experience to ensure alignment with the values of the target end user. Clearly define acceptable and unacceptable behaviors and identify key topics of moderation concern that should be prioritized. Be transparent in why, when, and how any moderation will happen. Lived experience experts should define what is acceptable and appropriate around topics that are important to them, such as crisis support and anonymity. The tone and language that will be used should be agreed upon enhance the acceptability of moderation practices.
2. Establish clear escalation and referral pathways	• Clearly define what support is available for community members in crisis. Ensure moderators are trained in handling high-risk situations in accordance with the escalation and referral pathways. Consider the pathways to be taken when handling complex or ambiguous cases.
3. Determine how the moderators "speak" and be consistent	• Tone and language can dramatically affect how moderation is received in the community, with inappropriate tone or language stifling conversation or creating a "big brother" culture. Consistency is key for moderation to perceived as fair and reduces the risk of members developing a specific attachment to individual moderators. Sharing template phrases and responses to common situations can help reduce moderation burden. In health-related settings, tone and language is particularly important in relation to moderating medical topics and emotional distress.
4. Communicate moderation and moderation decisions transparently	• Share moderation policies transparently in easily accessible formats. Provide clear and respectful feedback to members who are moderated, referring to the relevant sections of the policies. This helps members understand boundaries, reduce hostility toward moderators and minimize the likelihood of future moderation issues.
5. Create features that encourage self-moderation among the community	• Having a community that keeps itself safe and moderates itself (eg, encouraging peers to seek support from their general practice doctor if medical advice is being discussed) with minimal intervention from the moderation team is the ideal. When creating the interface of the online peer community, co-create features that encourage self-moderation, such as "flag this post".
6. Be responsive and evolve moderation processes as the community evolves	• Moderation should evolve with the community. Regularly review what is working well, what is burdensome, and how moderation is impacting members and moderators. Early on, there is a greater need to establish clear boundaries. Over time, community members will grow ownership over the community, and moderator behavior should shift to nurture and accommodate self-moderation. • Adjust the features of the interface and moderation practices as needed. Ensure to seek the perspectives of both community members and moderators to ensure a comprehensive understanding of what is working well and what needs adapting. For example, the moderation team told us that they found it challenging to track new content since their last session. In response, we improved the visibility of new content with new timestamps and post ordering in the community forum. • Guidance on gathering feedback: Moderators are encouraged to reflect on how different moderation approaches and practices are being perceived by the community by reflecting on the community engagement/comments that follow any intervention by moderators. By considering what was done and how the community reacted to the moderation behavior, moderators can adapt their approach to improve how it is delivered and received. Additionally, by reflecting on commonalities among posts that users "flag" to the attention for the moderations, key areas that need further development or greater moderation attention can be detected. With the emergence of AI[a] and algorithms that can monitor community activity, in the future, there could be scope to explore how AI could assist moderation and help identify where moderation is working well or needs adjusting. Whilst further work is needed to explore the usefulness and appropriateness of AI-assisted moderation, this emerging area could offer exciting opportunities for streamlining and facilitating community moderation.
7. Ensure moderators feel supported	• Have ways for the moderators to connect outside of the peer community, for instance, through a Microsoft Teams chat. Creating a space for moderators to discuss situations and come to a collective decision, decompress from emotionally-taxing situations, and share learnings or updates can be beneficial. Provide regular training specifically related to the community's policies, particularly when these change over time.

[a]AI: artificial intelligence.

To our knowledge, this is the first study to integrate the experiences of community members and moderators to understand moderation in a newly launched peer support community. It was universally agreed that moderation is a critical pillar of community safety, reflecting the perspectives of members [31] and moderators [9] of other established communities that moderation is essential for creating a safe space and mitigating risks. As in other web-based health communities, we also found that the community members and moderators had particular concerns about the need to moderate medical misinformation [9,31,37]. This highlights the importance of tailoring moderation practice to the unique aims and sensitivities of the specific community, and the need to identify these through collaboration with the target audience

XSL•FO
RenderX

(Recommendation 1, Table 3). While moderators reported a few instances where they were required to enforce the moderation policies, they felt that the moderation blueprint with the clear escalation procedures was appropriate and applicable to the community. The minimal need for moderator intervention may reflect the relatively low activity levels on CommonGround, where fewer user-generated posts naturally reduced the need for moderation. Equally, it may be because members understood and were happy to abide by the moderation policies [30]. Community members found "no red flags" in the moderation policies, suggesting that these were aligned with their expectations and moderation needs.

The apparent success of moderation also suggests that co-designing moderation practices is critical in ensuring their acceptability among the community, and that having clear boundaries and explicit guidance for moderators can lay the foundations for successful moderation from "Day One" of the community (Recommendations 1-3, Table 4). Community members valued being able to easily access the moderation policies when needed, underlining the need to share moderation practices transparently, which can enhance legitimacy, perceived consistency, and accountability of moderation ([51]; Recommendation 4, Table 4). Improving the communication of reactive moderation decisions may help to increase understanding, promote adherence to norms, and improve future behavior [51,52]. Together, these findings underscore the importance of having clear, comprehensive moderation practices tailored to the community, including the specific content of policies and the transparency, tone, and language of communications. Where the capacity to codesign moderation policies and practices is limited, examples from existing communities can be drawn upon, particularly as key moderation practices and experiences are likely common to any online peer community, such as the use of peer moderators [30].

By integrating the perspectives of community members and moderators, we gained unique insight into areas of moderation that may require review and adaptation as the newly formed CommonGround community evolves. The anonymous nature of web-based communities is often seen as a double-edged sword: anonymity can facilitate stigma-free and honest posting for some [53], whilst others feel that anonymity hinders the ability to form the meaningful and enduring social connections that are the cornerstone of peer support [54,55]. We found similar mixed opinions among our community members. The desire for reduced anonymity may grow as community members become more willing to share personal information if it makes accessing social support from peers easier, even if this could introduce other risks [56,57]. The reputable KCL ownership and branding positively influenced perceptions of safety. However, as some members appreciated the anonymity or acknowledged its value for others, offering members the ability to choose their preferred level of anonymity may be appropriate. Future work should explore with lived experience experts how to implement features that allow personal control over anonymity to ensure that anonymity can serve as a facilitator to engagement for those who desire it, while also removing this potential barrier to social connection for others [58,59]. These divergent views on a core aspect of CommonGround's moderation practice highlight the importance of reflecting on moderation practices by listening to the community and adapting as necessary (Recommendation 6, Table 4).

The moderation practice and style needed for CommonGround as a new community were recognized to be different from those of larger, well-established communities. In well-established communities, community members can be seen to be actively self-moderating by detecting medical misinformation, questioning information shared, reminding peers not to give medical advice, and reporting trolls to moderators [31,37,60]. While CommonGround moderators and community members agreed that peer-to-peer moderation could be more powerful and feel less punitive, they recognized that CommonGround could not self-moderate yet. The limited volume of user-generated content and the weak sense of community and ownership among members is likely to have hindered the emergence of self-moderation. The shift toward self-moderation must occur organically, as the premature withdrawal of moderators could harm community cohesion and reduce perceptions of safety if members are not comfortable acquiring moderation responsibilities [31]. Indeed, as prior work often cites "super users" (ie, those who engage with the community often, and are often observed to be generating larger volumes of content than the average user) in contributing to how this culture evolves, who often only emerge as a community evolves and may also facilitate the shift toward self-moderation [61]. As voiced in our research and prior work [31], there may always be a need for professional moderators to deal with high-risk or serious issues, even if the community takes ownership over other moderation activities. Although used infrequently on CommonGround, the moderators valued features like "flag this post" as critical tools for encouraging a community towards self-moderation. Indeed, moderators of other web-based health communities have urged members to use such self-moderation tools more to help shape moderation norms by defining what content they deem unacceptable [30]. Therefore, while the community culture needed for self-moderation cannot be artificially created, the features and functions that support self-moderation can be implemented early (Recommendation 5, Table 4). Future work should evaluate member and moderator perspectives as a community evolves through different stages of self-moderation to determine what additional practices may help support a successful transition.

Few studies have explored the experiences of moderators, but some key challenges consistently emerge [37]. First, a common challenge lies in efficiently identifying new content to reduce the risk of unmoderated posts escalating into worst-case scenarios [9,37]. Moderators were also mindful of how this challenge could intensify as the community grows, underscoring the need to develop and test novel tools that enhance moderation efficiency as CommonGround is scaled. Another challenge that our moderators faced was navigating the moderation threshold where they must evaluate whether to take action [9,30,51]. The views of the moderators echo those of moderators in other communities, who want to prioritize safety but are concerned that premature or overbearing moderation can stifle conversation and potentially damage community culture [62]. Moderators observed that once they commented, the community ceased to

XSL•FO
RenderX

comment further. This pattern has been observed in other communities [62], with moderators interpreting this to reflect disengagement from the community. However, members typically felt that the moderators had struck the right balance, suggesting that the quiet comment section may instead reflect the generally low engagement we had on CommonGround, where most members were "readers" preferring to observe posts rather than actively create content. This disconnect between how moderators believe their actions are perceived and how the community actually experiences them highlights the importance of considering both perspectives, particularly if informing changes to moderation practices (Recommendation 6, Table 4).

The benefits that the CommonGround moderators experienced as a result of operating as a team have been documented among other moderation teams. For instance, the CommonGround moderators shared how it became easier to navigate the moderation threshold over time and as a result of seeking support from their team, as they were able to exchange experiences to align their moderation practice, seek emotional support, and access the expertise of other members for second opinions on complex situations [63]. This adds to the literature highlighting how the moderation team becomes a form of peer support in itself that can help improve moderation efficiency, consistency, and reduce the moderation burden [9,64]. This highlights how it is important to consider how best to support moderators whilst they are supporting the community and give them a dedicated space to connect and support one another (Recommendation 7, Table 4).

Limitations

There are some limitations that should be considered for the interpretation and transferability of our findings and recommendations. First, this study explored the perspectives of community members and moderators who participated in the new community that was live for only 3 months. Many community members shared that they did not engage with CommonGround as they would have liked to due to personal circumstances or having limited time available during the 3-month trial. While lower levels of engagement and forum activity could be expected during early stages of community development, the experiences shared in this work are contextualized within a small community with relatively low engagement and sense of community in a research setting. As such, our findings and recommendations may not apply to communities launched with larger populations or with high volumes of user-generated content from the outset. Second, our findings should be interpreted relative to the nature of our moderators: professional, paid moderators with clinical expertise who were trained specifically about the moderation practices for CommonGround. Prior research suggests that professional moderators are often perceived as more trustworthy and credible than informal moderators, particularly around their moderation of medical topics [31]. Indeed, our moderators were aware of how their clinical expertise influenced their approach to moderation, and therefore, our findings may not generalize to moderation teams of volunteers, peers, or those without clinical or moderation backgrounds. Third, there is a potential for selection bias in the community members' interviews, as those who consented to participate may have more positive views of web-based peer forums or the moderation and governance of CommonGround. However, we did purposefully sample community members based on their engagement levels, probing those who had never logged in or only logged in a few times for their underlying reasons, motivations for low (or no) engagement, to unpack whether moderation or governance played a role in their disengagement. Fourth, although community members were anonymous to one another, they were aware that the moderators could link their pseudonyms to their personal details if required for moderation purposes. This awareness that they could be identified in exceptional circumstances may have influenced what members posted, meaning that the CommonGround environment may not directly reflect non-research anonymous communities.

All moderators and community members needed internet access to participate in the web-based community and interviews or focus groups. While 96% of households in the United Kingdom have internet access [65], for those without or who have limited access, their ability to engage with the community and interviews may have been limited. Additionally, some participants may have experienced barriers in accessing CommonGround privately, which may have had a downstream impact on the content of the community forum, and therefore the moderation needed [66]. Future work should consider how digital exclusion and privacy influence the nature of web-based communities, in particular their moderation.

Conclusion

Our work adds to our understanding of the critical role moderation has in shaping safety in online peer support communities designed to support mental well-being, and the importance of creating appropriate and trustworthy governance that is specifically tailored to the needs of the community members. By exploring perspectives on moderation in a new community and by integrating the experiences of both community members and moderators, we have gained unique insights into how to enhance our codesigned moderation practices going forward. We have provided a transparent account of the creation, implementation, and reflection on our moderation policies and governance, and recommendations on how to approach the moderation of new communities in the future.

Acknowledgments

We would like to acknowledge our software developers at TOAD and BitJam, who were instrumental in helping to co-design the CommonGround platform and delivering the intervention to participants. We would also like to acknowledge the Online Harms Team at the Samaritans, who kindly reviewed our moderation and safeguarding policies and procedures. And finally, thank you to all the members of our public patient involvement groups and research advisory group who contributed at various

XSL•FO
RenderX

stages throughout the development and delivery of this trial. CommonGround is a co-produced intervention, and this research would not have been possible without their dedication, support, and insightful contributions.

Funding

This project has been funded by King's Health Partners and Guy's and St Thomas Charity from the Multiple Long Term Conditions Challenge Fund (EIC180702 [MH]). The funders had no role in study design, data collection and analysis, decision to publish, or manuscript preparation. The research was part-funded by the National Institute for Health and Care Research (NIHR) Maudsley Biomedical Research Centre (BRC). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Authors' Contributions

All authors listed meet the International Committee of Medical Journal Editors criteria for authorship. MH, AS, VL, and GL conceived the study. All authors co-designed the study. HGJ, CR, and GL collected qualitative data. HGJ completed the qualitative analysis under the supervision of VL and AS, who are experienced qualitative researchers in the areas of mental health and chronic illness. HGJ prepared the manuscript. All authors have read and approved the final manuscript. All the research team members directly involved in data collection and analysis adopt an interpretivist-constructivist approach for this work. Throughout the analysis, collaboration through discussions ensured that the analysis remained grounded in the lived experience and voices of the participants and was not subject to any biases the team may have from their experience in research exploring the interplay between physical and mental health. E-AF has lived experience of multiple long-term conditions, and ensured that this work remained rooted in the lived experience of those with chronic conditions.

Conflicts of Interest

None declared.

Multimedia Appendix 1
SRQR checklist.
[PDF File (Adobe PDF File), 86 KB - jmir_v28i1e81943_app1.pdf]

Multimedia Appendix 2
Screenshots of the final prototype of CommonGround.
[DOCX File , 2293 KB - jmir_v28i1e81943_app2.docx]

Multimedia Appendix 3
Interview and focus group topic guides.
[DOCX File , 36 KB - jmir_v28i1e81943_app3.docx]

Multimedia Appendix 4
Theme descriptions and further evidencing quotes.
[DOCX File , 32 KB - jmir_v28i1e81943_app4.docx]

Multimedia Appendix 5
Comparison of moderator comments on CommonGround to moderation guidance.
[DOCX File , 28 KB - jmir_v28i1e81943_app5.docx]

References

1. Hossain SN, Jaglal SB, Shepherd J, Perrier L, Tomasone JR, Sweet SN, et al. Web-based peer support interventions for adults living with chronic conditions: scoping review. JMIR Rehabil Assist Technol 2021;8(2):e14321 [FREE Full text] [doi: 10.2196/14321] [Medline: 34032572]
2. Fortuna KL, Naslund JA, LaCroix JM, Bianco CL, Brooks JM, Zisman-Ilani Y, et al. Digital peer support mental health interventions for people with a lived experience of a serious mental illness: systematic review. JMIR Ment Health 2020;7(4):e16460 [FREE Full text] [doi: 10.2196/16460] [Medline: 32243256]
3. Wood S, Finnis A, Khan H. At the Heart of Health: Realising the Value of People and Communities. London: Health Foundation and Nesta; 2016:15.
4. Wright KB. Communication in health-related online social support groups/communities: a review of research on predictors of participation, applications of social support theory, and health outcomes. Rev Commun Res 2016;4:65-87 [FREE Full text] [doi: 10.12840/issn.2255-4165.2016.04.01.010] [Medline: 2255-4165]

XSL•FO
RenderX

5. Reidy C, Doble E, Robson A, Kar P. Peer support in chronic health conditions. BMJ 2024 Jul 01;386:e070443. [doi: 10.1136/bmj-2022-070443] [Medline: 38950946]
6. Malik S, Coulson NS. The therapeutic potential of the internet: exploring self-help processes in an internet forum for young people with inflammatory bowel disease. Gastroenterol Nurs 2011;34(6):439-448. [doi: 10.1097/SGA.0b013e318237a9ba] [Medline: 22129797]
7. Druss BG, Singh M, von Esenwein SA, Glick GE, Tapscott S, Tucker SJ, et al. Peer-led self-management of general medical conditions for patients with serious mental illnesses: a randomized trial. Psychiatr Serv 2018;69(5):529-535 [FREE Full text] [doi: 10.1176/appi.ps.201700352] [Medline: 29385952]
8. Cooper RE, Saunders KRK, Greenburgh A, Shah P, Appleton R, Machin K, et al. The effectiveness, implementation, and experiences of peer support approaches for mental health: a systematic umbrella review. BMC Med 2024;22(1):72 [FREE Full text] [doi: 10.1186/s12916-024-03260-y] [Medline: 38418998]
9. Deng D, Rogers T, Naslund JA. The role of moderators in facilitating and encouraging peer-to-peer support in an online mental health community: a qualitative exploratory study. J Technol Behav Sci 2023;8(2):128-139 [FREE Full text] [doi: 10.1007/s41347-023-00302-9] [Medline: 36810998]
10. Tran VT, Barnes C, Montori VM, Falissard B, Ravaud P. Taxonomy of the burden of treatment: a multi-country web-based qualitative study of patients with chronic conditions. BMC Med 2015;13:115 [FREE Full text] [doi: 10.1186/s12916-015-0356-x] [Medline: 25971838]
11. Demain S, Gonçalves AC, Areia C, Oliveira R, Marcos AJ, Marques A, et al. Living with, managing and minimising treatment burden in long term conditions: a systematic review of qualitative research. PLoS One 2015;10(5):e0125457 [FREE Full text] [doi: 10.1371/journal.pone.0125457] [Medline: 26024379]
12. Riegel B, Jaarsma T, Lee CS, Strömberg A. Integrating symptoms into the middle-range theory of self-care of chronic illness. ANS Adv Nurs Sci 2019;42(3):206-215 [FREE Full text] [doi: 10.1097/ANS.0000000000000237] [Medline: 30475237]
13. Long term conditions compendium of information. GOV.UK.: Department of Health; 2012. URL: https://assets.publishing.service.gov.uk/media/5a7c638340f0b62aff6c154e/dh_134486.pdf [accessed 2025-08-08]
14. Gold SM, Köhler-Forsberg O, Moss-Morris R, Mehnert A, Miranda JJ, Bullinger M, et al. Comorbid depression in medical diseases. Nat Rev Dis Primers 2020;6(1):69. [doi: 10.1038/s41572-020-0200-2] [Medline: 32820163]
15. Schmitz N, Gariépy G, Smith KJ, Clyde M, Malla A, Boyer R, et al. Recurrent subthreshold depression in type 2 diabetes: an important risk factor for poor health outcomes. Diabetes Care 2014;37(4):970-978. [doi: 10.2337/dc13-1832] [Medline: 24198303]
16. Yohannes AM, Baldwin RC, Connolly MJ. Prevalence of sub-threshold depression in elderly patients with chronic obstructive pulmonary disease. Int J Geriatr Psychiatry 2003;18(5):412-416. [doi: 10.1002/gps.851] [Medline: 12766917]
17. Hance M, Carney RM, Freedland KE, Skala J. Depression in patients with coronary heart disease. a 12-month follow-up. Gen Hosp Psychiatry 1996;18(1):61-65. [doi: 10.1016/0163-8343(95)00100-x] [Medline: 8666215]
18. Bot M, Pouwer F, Ormel J, Slaets JPJ, de Jonge P. Predictors of incident major depression in diabetic outpatients with subthreshold depression. Diabet Med 2010;27(11):1295-1301. [doi: 10.1111/j.1464-5491.2010.03119.x] [Medline: 20950389]
19. Moussavi S, Chatterji S, Verdes E, Tandon A, Patel V, Ustun B. Depression, chronic diseases, and decrements in health: results from the World Health Surveys. Lancet 2007;370(9590):851-858. [doi: 10.1016/S0140-6736(07)61415-9] [Medline: 17826170]
20. Iliffe LL, Thompson AR. Investigating the beneficial experiences of online peer support for those affected by alopecia: an interpretative phenomenological analysis using online interviews. Br J Dermatol 2019;181(5):992-998 [FREE Full text] [doi: 10.1111/bjd.17998] [Medline: 30972732]
21. Berkanish P, Pan S, Viola A, Rademaker Q, Devine KA. Technology-based peer support interventions for adolescents with chronic illness: a systematic review. J Clin Psychol Med Settings 2022;29(4):911-942 [FREE Full text] [doi: 10.1007/s10880-022-09853-0] [Medline: 35147830]
22. Gerritzen EV, Lee AR, McDermott O, Coulson N, Orrell M. Online peer support for people with parkinson disease: narrative synthesis systematic review. JMIR Aging 2022;5(3):e35425 [FREE Full text] [doi: 10.2196/35425] [Medline: 35896025]
23. Maclachlan LR, Mills K, Lawford BJ, Egerton T, Setchell J, Hall LM, et al. Design, delivery, maintenance, and outcomes of peer-to-peer online support groups for people with chronic musculoskeletal disorders: systematic review. J Med Internet Res 2020 Apr 24;22(4):e15822 [FREE Full text] [doi: 10.2196/15822] [Medline: 32329746]
24. Kingod N, Cleal B, Wahlberg A, Husted GR. Online peer-to-peer communities in the daily lives of people with chronic illness: a qualitative systematic review. Qual Health Res 2017 Jan;27(1):89-99. [doi: 10.1177/1049732316680203] [Medline: 27956659]
25. The living better project: addressing mental health and wellbeing in people living with long-term condition. Mental Health Foundation. 2011. URL: https://www.mentalhealth.org.uk/publications/living-better-report [accessed 2026-02-12]
26. NHS. The Five Year Forward View for Mental Health. England: Mental Health Taskforce NHS England; 2016.
27. Rowlands H, Aylwin-Foster E, Hotopf M, Rayner L, Simpson A, Lavelle G, et al. A qualitative exploration of the psychosocial needs of people living with long-term conditions and their perspectives on online peer support. Health Expect 2023 Oct;26(5):2075-2088 [FREE Full text] [doi: 10.1111/hex.13814] [Medline: 37458403]

XSL•FO
RenderX

28. Jones HL, Lavelle G, Aylwin-Foster E, Carr E, Simpson A, Hotopf M, et al. Co-producing an online platform for people with long-term physical health conditions: a development and usability study. JMIR Preprints . [doi: 10.2196/preprints.79666]
29. Lavelle G, Jones HG, Carr E, Aylwin-Foster E, Lawrence V, Simpson A, et al. Online peer support for long-term conditions: protocol for a feasibility randomized controlled trial. JMIR Res Protoc 2025 Jul 23;14:e71513 [FREE Full text] [doi: 10.2196/71513] [Medline: 40700748]
30. Wood HE, Karampatakis GD, Coulson NS, Sastry N, Li X, Taylor SJC, et al. Moderators' experiences of the safety and effectiveness of patient engagement in an asthma online health community: exploratory qualitative interview study. J Med Internet Res 2025;27:e58167 [FREE Full text] [doi: 10.2196/58167] [Medline: 40279577]
31. Huh J, Marmor R, Jiang X. Lessons learned for online health community moderator roles: a mixed-methods study of moderators resigning from webMD communities. J Med Internet Res 2016;18(9):e247 [FREE Full text] [doi: 10.2196/jmir.6331] [Medline: 27608721]
32. Easton K, Diggle J, Ruethi-Davis M, Holmes M, Byron-Parker D, Nuttall J, et al. Qualitative exploration of the potential for adverse events when using an online peer support network for mental health: cross-sectional survey. JMIR Ment Health 2017;4(4):e49 [FREE Full text] [doi: 10.2196/mental.8168] [Medline: 29084710]
33. Ziebland S, Wyke S. Health and illness in a connected world: how might sharing experiences on the internet affect people's health? Milbank Q 2012;90(2):219-249 [FREE Full text] [doi: 10.1111/j.1468-0009.2012.00662.x] [Medline: 22709387]
34. Lobban F, Coole M, Donaldson E, Glossop Z, Haines J, Johnston R, et al. Improving peer online forums (iPOF): protocol for a realist evaluation of peer online mental health forums to inform practice and policy. BMJ Open 2023;13(7):e075142 [FREE Full text] [doi: 10.1136/bmjopen-2023-075142] [Medline: 37518092]
35. Young C. Community management that works: how to build and sustain a thriving online health community. J Med Internet Res 2013;15(6):e119 [FREE Full text] [doi: 10.2196/jmir.2501] [Medline: 23759312]
36. Kraut RE, Resnick P. Building Successful Online Communities: Evidence-Based Social Design. Cambridge: Mit Press; 2012.
37. Skousen T, Safadi H, Young C, Karahanna E, Safadi S, Chebib F. Successful moderation in online patient communities: inductive case study. J Med Internet Res 2020;22(3):e15983 [FREE Full text] [doi: 10.2196/15983] [Medline: 32181743]
38. Matzat U, Rooks G. Styles of moderation in online health and support communities: an experimental comparison of their acceptance and effectiveness. Comput Hum Behav 2014;36:65-75. [doi: 10.1016/j.chb.2014.03.043]
39. Grimmelmann J. The virtues of moderation. Yale J Law Technol 2015;17:42. [doi: 10.31228/osf.io/qwxf5]
40. Seering J, Wang T, Yoon J, Kaufman G. Moderator engagement and community development in the age of algorithms. New Media Soc 2019;21(7):1417-1443. [doi: 10.1177/1461444818821316]
41. Biagianti B, Quraishi SH, Schlosser DA. Potential benefits of incorporating peer-to-peer interactions into digital interventions for psychotic disorders: a systematic review. Psychiatr Serv 2018;69(4):377-388 [FREE Full text] [doi: 10.1176/appi.ps.201700283] [Medline: 29241435]
42. Lavelle G, Jones HG, Carr E, Aylwin-Foster E, Lawrence V, Simpson A, et al. Online peer support for long-term conditions: protocol for a feasibility randomized controlled trial. JMIR Res Protoc 2025 Jul 23;14:e71513 [FREE Full text] [doi: 10.2196/71513] [Medline: 40700748]
43. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
44. Jones HG, Lavelle G, Aylwin-Foster E, Carr E, Simpson A, Hotopf M, et al. CommonGround: online peer support in long-term conditions: a feasibility randomised controlled trial. Open Science Framework (OSF). 2025. URL: https://osf.io/zhncp/ [accessed 2025-04-08]
45. Samaritans T. Online harms advisory service. Samaritans. 2024. URL: https://www.samaritans.org/about-samaritans/research-policy/internet-suicide/samaritans-online-harms-advisory-service/ [accessed 2026-02-12]
46. Scope. 2025. URL: https://www.scope.org.uk/ [accessed 2025-11-07]
47. Scope Online Forum. 2025. URL: https://forum.scope.org.uk/ [accessed 2025-07-22]
48. O'Halloran J, Miller GC, Britt H. Defining chronic conditions for primary care with ICPC-2. Fam Pract 2004;21(4):381-386. [doi: 10.1093/fampra/cmh407] [Medline: 15249526]
49. Braun V, Clarke V. One size fits all? what counts as quality practice in (reflexive) thematic analysis? Qual Res Psychol 2021;18(3):328-352. [doi: 10.1080/14780887.2020.1769238]
50. Successful qualitative research: A practical guide for beginners. In: Successful Qualitative Research: A Practical Guide for Beginners. California: Sage; Mar 01, 2016:48-50.
51. Jiang JA, Nie P, Brubaker JR, Fiesler C. A trade-off-centered framework of content moderation. ACM Trans Comput-Hum Interact 2023 Mar 07;30(1):1-34. [doi: 10.1145/3534929]
52. Jhaver SA, Bruckman A, Gilbert E. Does transparency in moderation really matter? Proc ACM Hum-Comput Interact 2019;3(CSCW):1-27. [doi: 10.1145/3359252]
53. Saha K, Kiranmai Ernala S. Understanding moderation in online mental health communities. In: Social Computing and Social Media. Participation, User Experience, Consumer Experience, and Applications of Social Computing. Cham, Switzerland: Springer International Publishing; 2020.

XSL•FO
RenderX

54. Strand M, Eng LS, Gammon D. Combining online and offline peer support groups in community mental health care settings: a qualitative study of service users' experiences. Int J Ment Health Syst 2020;14(1):39 [FREE Full text] [doi: 10.1186/s13033-020-00370-x] [Medline: 32514303]
55. Smith-Merry J, Goggin G, Campbell A, McKenzie K, Ridout B, Baylosis C. Social connection and online engagement: insights from interviews with users of a mental health online forum. JMIR Ment Health 2019;6(3):e11084 [FREE Full text] [doi: 10.2196/11084] [Medline: 30912760]
56. Tseng HT, Ibrahim F, Hajli N, Nisar TM, Shabbir H. Effect of privacy concerns and engagement on social support behaviour in online health community platforms. Technol Forecast Soc Change 2022;178:121592. [doi: 10.1016/j.techfore.2022.121592]
57. Zhu M, Wu C, Huang S, Zheng K, Young SD, Yan X, et al. Privacy paradox in mHealth applications: an integrated elaboration likelihood model incorporating privacy calculus and privacy fatigue. Telemat Inform 2021;61:101601. [doi: 10.1016/j.tele.2021.101601]
58. Adjekum A, Blasimme A, Vayena E. Elements of trust in digital health systems: scoping review. J Med Internet Res 2018;20(12):e11254 [FREE Full text] [doi: 10.2196/11254] [Medline: 30545807]
59. Alhammad N, Alajlani M, Abd-Alrazaq A, Epiphaniou G, Arvanitis T. Patients' perspectives on the data confidentiality, privacy, and security of mHealth apps: systematic review. J Med Internet Res 2024;26:e50715 [FREE Full text] [doi: 10.2196/50715] [Medline: 38820572]
60. Maloney-Krichmar D, Preece J. A multilevel analysis of sociability, usability, and community dynamics in an online health community. ACM Trans Comput-Hum Interact 2005;12(2):201-232. [doi: 10.1145/1067860.1067864]
61. De Simoni A, Shah AT, Olivia F, Jasmine P, Aziz S, Pietro P, et al. Superusers' engagement in asthma online communities: asynchronous web-based interview study. J Med Internet Res 2020;22(6):e18185 [FREE Full text] [doi: 10.2196/18185] [Medline: 32573463]
62. Huh J. Clinical questions in online health communities: the case of "See your doctor" threads. 2015 Presented at: Proceedings of the Conference on Computer-Supported Cooperative Work. Conference on Computer-Supported Cooperative Work; 2025 Oct 18; Bergen, Norway p. 1488-1499. [doi: 10.1145/2675133.2675259]
63. Seering J. Reconsidering self-moderation: the role of research in supporting community-based models for online content moderation. Proc ACM Hum-Comput Interact 2020;4(CSCW2):1-28. [doi: 10.1145/3415178]
64. Gibson AD. What teams do: exploring volunteer content moderation team labor on Facebook. Soc Media Soc 2023 Jul 19;9(3):20563051231186109. [doi: 10.1177/20563051231186109]
65. Statistics OFN. Internet access – households and individuals, Great Britain Statistical bulletins. Office for National Statistics. 2020. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/internetaccesshouseholdsandindividuals/previousreleases [accessed 2026-02-12]
66. Schlief M, Saunders KRK, Appleton R, Barnett P, Vera San Juan N, Foye U, et al. Synthesis of the evidence on what works for whom in telemental health: rapid realist review. Interact J Med Res 2022;11(2):e38239 [FREE Full text] [doi: 10.2196/38239] [Medline: 35767691]

Abbreviations

EQUATOR: Enhancing the Quality and Transparency of Health Research
fRCT: feasibility randomized controlled trial
KCL: King's College London
SRQR: Standards for Reporting Qualitative Research

Edited by A Stone; submitted 08.Aug.2025; peer-reviewed by H Wood, E Kalles, A Gupta, S Mittal; comments to author 02.Oct.2025; accepted 30.Jan.2026; published 12.Mar.2026.

Please cite as:
Jones HG, Lavelle G, Aylwin-Foster E, Regan C, Simpson A, Carr E, Hotopf M, Lawrence V
Insights and Recommendations From Moderators and Community Members for Keeping Online Peer Support Safe: Thematic Analysis
J Med Internet Res 2026;28:e81943
URL: https://www.jmir.org/2026/1/e81943
doi:10.2196/81943
PMID:

©Hannah Grace Jones, Grace Lavelle, Elly Aylwin-Foster, Ciara Regan, Alan Simpson, Ewan Carr, Matthew Hotopf, Vanessa Lawrence. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effect of a Digital-Driven Physician-Pharmacist Collaborative Model for Diabetes in Primary Health Care: Cluster Randomized Trial

Jie Xiao[1,2], MM; Qing Wang[1,2], MM; Shenglan Tan[1,2], PhD; Lei Chen[1,2], PhD; Daxiong Xiang[1,2], PhD; Haiyan Yuan[1,2], PhD; Xia Li[3], MD; Shuting Huang[1,2], MM; Bingjie Tang[1,2], MM; Yan Guo[4], MS; Haiying Huang[5], MS; Danhui Zhao[6], BS; Yue Li[7], BS; Li Wang[8], BS; Qun Li[9], MS; Juan Liu[10], MS; Ping Xu[1,2], PhD

[1]Department of Pharmacy, Second Xiangya Hospital, Central South University, Changsha, China
[2]Institute of Clinical Pharmacy, Second Xiangya Hospital, Central South University, Changsha, China
[3]Department of Endocrine, Second Xiangya Hospital, Central South University, Changsha, China
[4]Department of Pharmacy, Taoyuan People's Hospital, Changde, China
[5]Department of Pharmacy, The People's Hospital of Liuyang, Changsha, China
[6]Department of Pharmacy, The First People's Hospital of Pingjiang, Yueyang, China
[7]Department of Pharmacy, People's Hospital of Ningxiang, Changsha, China
[8]Department of Pharmacy, Yueyang Central Hospital, Yueyang, China
[9]Department of Pharmacy, The Second People's Hospital of Huaihua, Huaihua, China
[10]Department of Pharmacy, Yiyang Central Hospital, Yiyang, China

Corresponding Author:
Ping Xu, PhD
Department of Pharmacy
Second Xiangya Hospital
Central South University
No.139 Renmin Road, Furong District
Changsha, 410011
China
Phone: 86 731 85292128
Email: xuping1109@csu.edu.cn

Abstract

Background: Evidence-based physician-pharmacist collaborative clinics have demonstrated significant short-term benefits for patients with type 2 diabetes (T2D), but their long-term effectiveness remains unclear, especially in primary health care settings.

Objective: This study aimed to explore the long-term effectiveness and cost-effectiveness of a novel, digital-driven, multifaceted physician-pharmacist collaborative model for managing patients with T2D in underresourced settings.

Methods: We conducted a 12-month cluster randomized controlled trial from May 2021 to December 2022 across 6 primary health care settings in China. Guided by the theory of planned behavior, the intervention involved routine therapy from physicians along with pharmaceutical interventions from pharmacists. These were delivered through a combination of face-to-face visits and mobile health care. The intervention group received 4 face-to-face visits and biweekly remote education sessions over the 12 months. We conducted intention-to-treat analyses to estimate differences in clinical and behavior indicators between the intervention and control groups. Primary outcomes included glycosylated hemoglobin and 10-year atherosclerotic cardiovascular risk. Data were analyzed using adjusted generalized estimation equations.

Results: This study included 574 patients (291 in the intervention group and 283 in the control group). Over 12 months, patients in the intervention group had significant reductions in hemoglobin A_{1c} (–2.57 vs –1.96, respectively; P<.001; 95% CI –1.027 to –0.238) and 10-year atherosclerotic cardiovascular risk (–1.35 vs 0.01, respectively; P<.001; 95% CI –1.690 to –0.630) compared with the control group. Substantial improvements were also observed in several secondary outcomes, including fasting blood glucose, 2-hour postprandial blood glucose, waist circumference, waist-to-hip ratio, blood pressure, triglyceride, and total cholesterol. Total diabetes-related costs decreased, and patient satisfaction improved significantly in the intervention group. There were no significant differences in BMI, high-density lipoprotein, or low-density lipoprotein.

XSL•FO
RenderX

Conclusions: These findings suggest that the physician-pharmacist collaborative model could improve the long-term quality and efficiency of T2D management and reduce medical costs in underresourced areas globally. Patients with T2D, especially those with central obesity or high cardiovascular risk, may benefit more from collaborative clinics.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2000031839; https://www.chictr.org.cn/showproj.html?proj=51910

(J Med Internet Res 2026;28:e77470) doi:10.2196/77470

KEYWORDS

diabetes mellitus; physician-pharmacist collaboration; central obesity; glycemic control; atherosclerotic cardiovascular risk; cost-effectiveness analysis

Introduction

The global epidemic of diabetes, particularly type 2 diabetes (T2D), has caused severe public health challenges worldwide [1]. The International Diabetes Federation estimates that the prevalence of diabetes worldwide was 10.5% in 2021 and would increase to 12.2% by 2045 due to rapid lifestyle changes and aging of the population [2]. China has the largest population of people living with diabetes worldwide, with an estimated 1.39 million deaths attributable to diabetes or its complications in 2021, and this number continues to rise rapidly [2]. As in other countries, most of the population with diabetes in China has T2D, and half of them present with complications, leading to a huge economic burden [3]. This burden is exacerbated in resource-limited settings, where health care systems are often underequipped to manage the long-term complexities of the disease [1].

Effective, long-term management of T2D is therefore crucial to mitigating these complications and associated costs [4]. Diabetes management is multidimensional, including a low-energy diet, physical activity, medication, diabetes education, and continuous glucose monitoring [4-6]. In vast rural areas of China, primary health care centers are the preferred institutions for more than 900 million people, where diabetes management is still dominated by physicians [7]. However, diabetes management in rural areas in China is currently insufficient due to limited resources and staff shortage in primary health care centers [8]. Moreover, patients in rural areas generally have lower educational background and socioeconomic status, which necessitates more detailed treatment and management support, which puts further strain on the capacity of the current physician-led care model [9].

To address these issues, innovative team-based care models have shown promise [10-13]. One well-established model is the physician-pharmacist collaboration, which improves the quality of outpatient services and reduces physicians' workload by involving pharmacists in medication counseling and patient education [14,15]. However, due to the limited reach and accessibility of in-person pharmaceutical services, it could be difficult to implement established pharmaceutical intervention programs in underresourced areas [16]. Mobile health (mHealth) interventions offer a promising way to enhance and sustain collaborative care in these contexts [17]. Evidence shows that mHealth tools, including medication reminders, mobile apps for tracking, and remote monitoring, can improve medication adherence, self-management behaviors, and clinical outcomes in diabetes care, thereby alleviating the workload of health care providers [18,19]. An mHealth-based physician-pharmacist collaboration model could be adopted in underresourced settings.

Previous studies have demonstrated that the intervention programs based on the theory of planned behavior (TPB) contribute to change behavior patterns, but few studies have tailored the method to comprehensive intervention for diabetes management in underresourced settings [20,21]. Therefore, we have developed a digital-driven, multifaceted intervention program based on physician-pharmacist collaborative models, combining face-to-face visits and mHealth support. We conducted a cluster randomized controlled trial to evaluate the long-term effectiveness of the intervention on hemoglobin A_{1c} (HbA_{1c}), 10-year atherosclerotic cardiovascular disease (ASCVD) risk, body weight, blood pressure, lipids, and cost-effectiveness of the physician-pharmacist collaborative model for T2D management in primary health care centers in China.

Methods

Study Design

The cluster randomized controlled trial was conducted from May 2021 to December 2022 in 6 health care centers in Hunan province, China. A trial protocol was reported in a previous study [22]. Hunan province is located in the center of China, with middle-level economy and general medical condition. To ensure representativeness, all 88 Hunan counties were divided into 3 subgroups by per capita income level, namely, high, medium, and low incomes. We then sent invitations to 27 primary health care centers (the largest by outpatient volume in their respective counties) across these strata. Of the 18 centers that agreed to participate, we randomly selected 2 centers from each income stratum. Within each stratum, the 2 selected centers were then randomly allocated to the intervention and control groups (1 per group), resulting in 3 centers in the intervention group and 3 in the control group. All eligible patients attending these centers received care according to their center's allocation. This study was reported following the CONSORT (Consolidated Standards of Reporting Trials) extension for cluster trials (Multimedia Appendix 1). We held progress meetings monthly and used fidelity checklists to monitor implementation (Multimedia Appendix 2).

Randomization and Masking

Randomization was performed at the level of the health care center (cluster). Using a standard randomization protocol with a computer-based random sequence generation, an independent

XSL•FO
RenderX

statistician performed the randomization [23]. This involved randomly selecting the 6 study centers from the 18 consenting centers, followed by randomly allocating the selected centers to the intervention and control groups with balanced allocation. Because the intervention required knowledge of patient allocation, blinding of participants was not feasible; however, data processors were blinded to group assignment.

Participants

The study sample included patients with T2D who were aged ≥18 years, had HbA_{1c} levels >7.5% in the past 2 months, and provided contact information. Patients were excluded if they were pregnant or had end-stage renal failure, dementia, psychiatric disorders, cancer, congestive heart failure, or pancreatitis. All eligible patients were required to provide written informed consent before participation in the study.

Intervention and Procedures

Pharmacist Training and Qualification

Pharmacists participating in the study had worked as clinical pharmacists for >2 years, held standardized training qualifications, and completed a 3-month comprehensive training course at the Second Xiangya Hospital. The training courses mainly cover 2 aspects: diabetes management and cooperative skills. Topics included knowledge of diabetes, hypoglycemic agents, treatment target, interpretation of laboratory tests, medication adherence, diet, exercise, and lifestyle. Detailed training content is available in Table S1 in Multimedia Appendix 3 [24]. Intermediate and final tests are used to assess learning, and those who fail the tests were excluded from the study [22].

Intervention Group

Centers in the intervention group provided routine physician-led therapy combined with pharmacist-delivered pharmaceutical care through outpatient visits at baseline and at the third, sixth, ninth, and 12th months, along with biweekly remote education via WeChat (the most popular social media platform in China [25,26]) throughout the 12-month study period. The pharmaceutical intervention program was based on the TPB framework and included 3 areas: attitude toward the behavior, subjective norm, and perceived behavior control (Textbox 1). Well-trained pharmacists provided multifaceted pharmaceutical care at each follow-up visit, which lasted at least 20 minutes; handed out health education manuals at the end of visits; and conducted biweekly remote health education through WeChat to regulate the patient attitude toward treatment. Multifaceted pharmaceutical care included diabetes education, medication guidance, lifestyle intervention, treatment of adverse drug reactions, and complication identification, while remote health education involved pharmaceutical consultation, structured educational messages, and regular medication reminders (Table S2 in Multimedia Appendix 3). Subjective norm was achieved through peer support and family support to strengthen patient behavior intention. Meanwhile, pharmacists set health goals for patients and strengthened their medication knowledge to improve the perception of treatment effects.

Textbox 1. The intervention program based on the theory of planned behavior.

Attitude toward the behavior

- Face-to-face pharmaceutical care at each visit, including diabetes education, medication guidance, lifestyle intervention, adverse drug reaction management, and complication identification
- Health education manuals were distributed, covering diabetes knowledge, hazards of complications, introduction of commonly used drugs, healthy lifestyle, and other related knowledge
- Health education was delivered through the WeChat group, including answering questions from patients and their families; providing guidance on daily management; offering disease education; and sharing pharmaceutical knowledge, dietary advice, and exercise plans

Subjective norm

- Peer support, which allows patients to interact with each other within their peer group in WeChat
- Family support, including diabetes care, rational drug use, healthy lifestyle, and other related knowledge taught by pharmacists

Perceived behavior control

- Formulation of health goals, including short-term and long-term health goals
- Medication consultation services, including dose adjustment, medication duration, drug interactions, and adverse drug reactions

Control Group

Patients in the control centers received usual treatment and management by physicians, with no pharmacist available to provide additional patient education. After the study, all patients received a diabetes education manual.

Outcomes

The primary outcomes were HbA_{1c} and 10-year ASCVD risk over 12 months, while glycemic measurements, weight-related measurements, blood pressure, lipid indicators, patient satisfaction, and costs were secondary outcomes. HbA_{1c} provides a better reflection of the patient's average blood glucose level over a 3-month period. The 10-year ASCVD risks were calculated by the Prediction for atherosclerotic cardiovascular disease risk in China model [27]. The biomarkers of glycemic measurements include fasting blood glucose (FBG) and 2-hour postprandial blood glucose (PBG2h). Weight-related measurements included BMI, waist circumference (WC), and waist-to-hip ratio (WHR). The indicators of blood pressure

XSL•FO
RenderX

included systolic blood pressure and diastolic blood pressure. The biomarkers of blood lipid status include triglyceride, total cholesterol (TC), high-density lipoprotein cholesterol, and low-density lipoprotein cholesterol (LDL-C). Patient satisfaction was assessed by the Diabetes Treatment Satisfaction Questionnaire, with a Cronbach alpha index of 0.717. The total cost of a patient's treatment included the costs of drugs, registration, examinations, hospitalization, lost wages, and transportation related to diabetes care.

Sample Size

On the basis of previous studies, we calculated that at 90% power and 5% significance level, the minimum detectable difference in HbA_{1c} in the intervention group was 0.76% with an SD of 1.0%, while minimum detectable difference in 10-year ASCVD risk was 3.98 [28,29]. The intraclass correlation coefficients at the center level were 0.03 for HbA_{1c} and 0.05 for 10-year ASCVD risk [30,31]. Considering the potential effects of clustering and loss to follow-up over 12 months, the study required a sample size of at least 288 patients in this study (144 in each group).

Statistical Analysis

All analyses were conducted using a modified intention-to-treat approach, which excluded 26 patients who withdrew after randomization but before trial implementation. Measurement data were presented as means with SDs or median with IQRs, while enumeration data were presented as counts with percentages. Descriptive statistics were used to summarize the baseline demographic characteristics, while independent sample 2-tailed *t* tests and Mann-Whitney *U* tests were used to compare the differences between groups at baseline. Differences in measurements and costs over 12 months between the groups were detected by generalized estimating equation (GEE) analysis. GEE provides stable estimates of time, group, and interaction effects for repeated measurement data, while GEE model gave valid estimation under the assumption of missing at random. GEE models were adjusted for center, gender, age, comorbidity, smoking history, alcohol history, family history, education, income, occupation, and residence. The standard rate of control was referred to the comprehensive control target of T2D in China (Table S3 in Multimedia Appendix 3). Sensitivity analysis was performed using the per-protocol population, which included patients who adhered to the protocol and were not lost to follow-up. Missing data were minimal (<10% for primary outcomes) and were handled using GEE, which provides valid estimates under the missing-at-random assumption by using all available data without imputation. Additionally, 2-sided *P* values <.05 were considered significant for all analyses. SPSS Statistics (version 25.0; IBM Corp) and SAS (version 9.4; SAS Institute) were used.

Ethical Considerations

This study was approved by the Clinical Research Ethics Committee of the Second Xiangya Hospital of Central South University (2019-213). All participants were informed of the study's purpose, procedures, potential risks, and benefits, and they provided written informed consent prior to enrollment. Strict measures were taken to protect participant privacy and confidentiality; all data were anonymized before analysis. As compensation for their participation, all enrolled patients received free laboratory examinations related to the study.

Results

Baseline Characteristics

The study was conducted in 6 centers (Table S4 in Multimedia Appendix 3) and included 574 patients, of whom 50.7% (n=291) participated in physician-pharmacist collaborative clinics (Table 1). The mean age of the enrolled patients was 55.0 (48.0 to 64.0) years and 52.4% (n=301) were male. Overall, 34% (n=195) of the patients had a history of smoking and 27.2% (n=156) a history of alcohol use. Baseline characteristics of patients were generally comparable between the 2 groups. Patient enrollment is shown in Figure 1.

XSL•FO
RenderX

Table 1. Patients' demographic characteristics (N=574).

Characteristics	Overall	Intervention (n=291)	Control (n=283)
Age (years), median (IQR)	55.0 (48.0-64.0)	54.0 (49.0-60.0)	55.0 (46.0-67.0)
Sex, n (%)			
Male	301 (52.4)	139 (47.8)	162 (57.2)
Female	273 (47.6)	152 (52.2)	121 (42.8)
Comorbidity, n (%)			
Hypertension			
Yes	90 (15.7)	50 (17.2)	40 (14.1)
Hyperlipidemia			
Yes	26 (4.5)	10 (3.4)	16 (5.7)
Coronary heart disease			
Yes	20 (3.5)	11 (3.8)	9 (3.2)
History of smoking, n (%)			
Yes	195 (34)	84 (28.9)	111 (39.2)
History of alcohol intake, n (%)			
Yes	156 (27.2)	83 (28.5)	73 (25.8)
Family history, n (%)			
Diabetes			
Yes	164 (28.6)	89 (30.6)	75 (26.5)
Hypertension			
Yes	94 (16.4)	54 (18.6)	40 (14.1)
Hyperlipidemia			
Yes	45 (7.8)	23 (7.9)	22 (7.7)
Education, n (%)			
Junior high school and below	318 (55.4)	168 (57.7)	150 (53)
Senior high school	152 (26.5)	78 (26.8)	74 (26.1)
College and above	104 (18.1)	45 (15.5)	59 (20.9)
Income per month (CNY)[a], n (%)			
<1000	96 (16.7)	54 (18.5)	42 (14.8)
1000-3000	191 (33.3)	98 (33.7)	93 (32.9)
>3000	287 (50)	139 (47.8)	148 (52.3)
Occupation, n (%)			
Full-time job	395 (68.8)	215 (73.9)	180 (65.9)
Part-time job or no job	179 (31.2)	76 (26.1)	103 (34.1)
Residence, n (%)			
Rural area	321 (55.9)	151 (51.9)	170 (62.3)
Town or urban area	253 (44.1)	140 (48.1)	113 (37.7)
HbA_{1c}[b], mean (SD)	9.83 (2.02)	9.65 (1.87)	10.02 (2.15)
10-year ASCVD[c] risk, mean (SD)	7.56 (5.87)	7.05 (4.92)	8.06 (6.65)
FBG[d] (mmol/L), mean (SD)	10.65 (4.16)	10.60 (3.60)	10.69 (4.66)
PBG2h[e] (mmol/L), mean (SD)	15.19 (5.27)	15.34 (4.64)	15.03 (5.84)
BMI (kg/m^2), mean (SD)	24.87 (3.33)	24.73 (3.38)	25.02 (3.28)

XSL•FO
RenderX

Characteristics	Overall	Intervention (n=291)	Control (n=283)
WC[f] (cm), mean (SD)			
Male	91.92 (9.34)	91.17 (8.89)	92.60 (9.69)
Female	88.73 (10.04)	88.62 (10.23)	88.88 (9.86)
WHR[g], mean (SD)			
Male	0.927 (0.062)	0.924 (0.061)	0.930 (0.062)
Female	0.927 (0.078)	0.927 (0.076)	0.927 (0.078)
SBP[h] (mm Hg), mean (SD)	133.38 (18.33)	132.53 (16.05)	134.26 (20.41)
DBP[i] (mm Hg), mean (SD)	81.20 (9.90)	80.47 (9.03)	81.96 (10.71)
Triglyceride (mmol/L), mean (SD)	2.80 (3.15)	3.09 (3.87)	2.53 (2.20)
TC[j] (mmol/L), mean (SD)	4.98 (1.56)	5.12 (1.74)	4.84 (1.36)
HDL-C[k] (mmol/L), mean (SD)	1.30 (0.57)	1.30 (0.53)	1.30 (0.60)
LDL-C[l] (mmol/L), mean (SD)	2.94 (1.11)	2.96 (1.25)	2.92 (0.94)

[a]CNY 1=US $0.145.
[b]HbA_{1c}: hemoglobin A_{1c}.
[c]ASCVD: atherosclerotic cardiovascular disease.
[d]FBG: fasting blood glucose.
[e]PBG2h: 2-hour postprandial blood glucose.
[f]WC: waist circumference.
[g]WHR: waist-to-hip ratio.
[h]SBP: systolic blood pressure.
[i]DBP: diastolic blood pressure.
[j]TC: total cholesterol.
[k]HDL-C: high-density lipoprotein cholesterol.
[l]LDL-C: low-density lipoprotein cholesterol.

XSL•FO
RenderX

Figure 1. Flowchart of the trial.

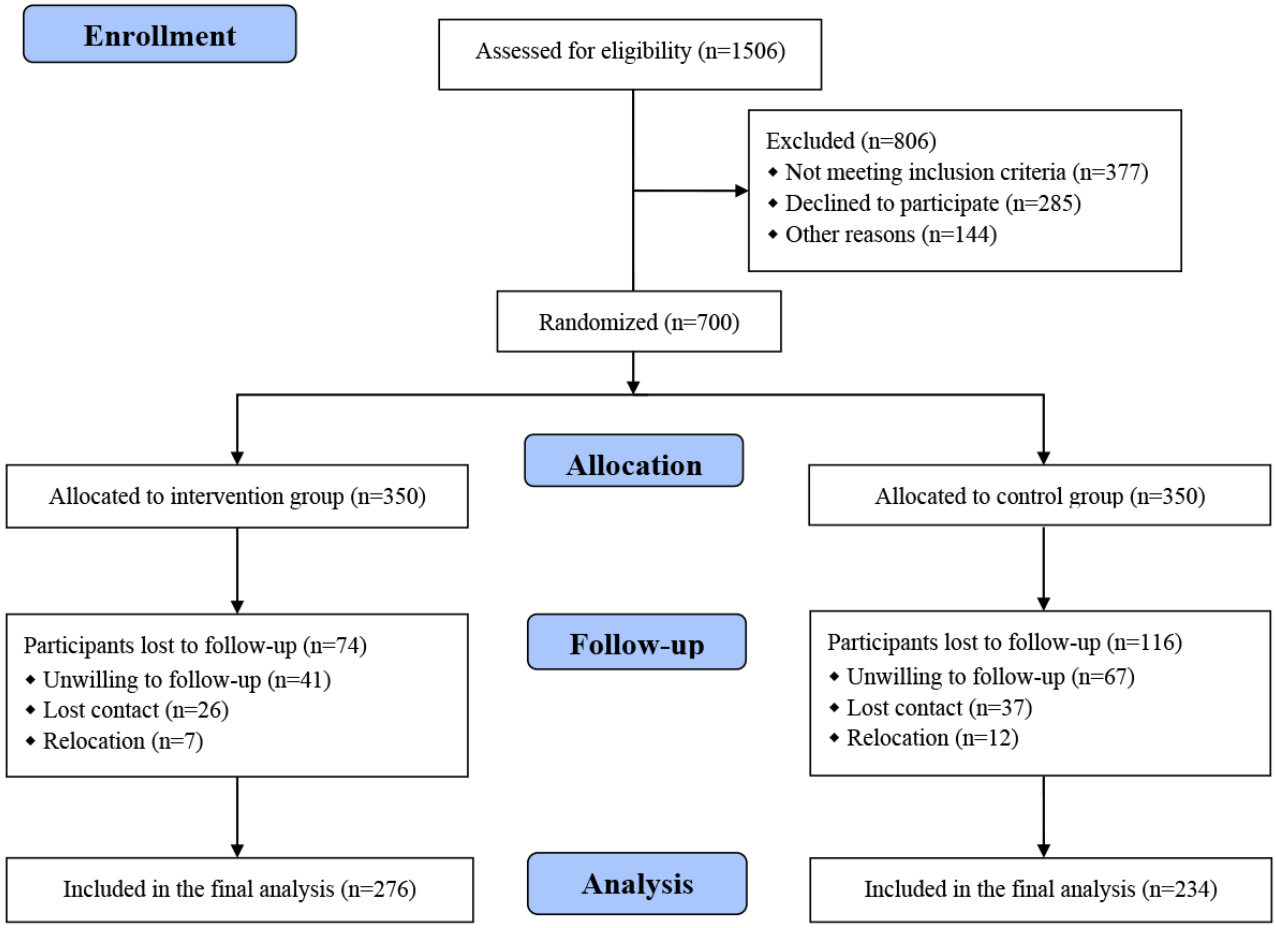

Primary Outcomes

At 12 months, the reduction in HbA_{1c} was significantly greater in the intervention group (−2.57%) than in the control group (−1.96%; β=−0.632; *P*<.001; 95% CI −1.027 to −0.238). Detailed information is shown in Table 2. Correspondingly, the HbA_{1c} control rates increased markedly from baseline to 12 months in both groups, but the improvement was significantly greater in the intervention group (from 0% to 53.61%) than in the control group (from 1.77% to 38.52%; *P*=.001). Refer to Table S5 in Multimedia Appendix 3 for full details. The 10-year ASCVD risk decreased significantly in the intervention group compared with the control group over the study period (β=−1.206; *P*<.001; 95% CI −1.730 to −0.68).

XSL•FO
RenderX

Table 2. The 12-month effectiveness of the physician-pharmacist collaborative clinics on clinical outcomes.

Variables	Intervention group, mean (SD)	Control group, mean (SD)	Unadjusted effect		Adjusted effect[a]	
			P value[b]	β (95% CI)	*P* value	β (95% CI)
HbA_{1c}[c]			<.001	−0.632 (−1.026 to −0.237)	<.001	−0.632 (−1.027 to −0.238)
Baseline	9.65 (1.87)	10.02 (2.15)				
3 months	7.93 (1.53)	8.43 (2.19)				
6 months	7.76 (1.63)	8.15 (1.86)				
9 months	7.55 (1.48)	8.18 (1.87)				
12 months	7.08 (1.13)	8.06 (2.02)				
10-year ASCVD[d] risk			<.001	−1.160 (−1.690 to −0.630)	<.001	−1.206 (−1.730 to −0.682)
Baseline	7.05 (4.92)	8.06 (6.65)				
12 months	5.70 (4.05)	8.07 (6.37)				
FBG[e]			.29	−0.592 (−1.363 to 0.178)	.27	−0.602 (−1.372 to 0.169)
Baseline	10.60 (3.60)	10.69 (4.66)				
3 months	8.38 (2.54)	9.03 (3.32)				
6 months	8.40 (2.77)	8.92 (2.67)				
9 months	8.02 (2.43)	8.31 (2.28)				
12 months	7.58 (2.09)	8.22 (2.67)				
PBG2h[f]			.03	−1.234 (−2.338 to −0.129)	.03	−1.210 (−2.315 to −0.106)
Baseline	15.34 (4.64)	15.03 (5.84)				
3 months	12.38 (3.89)	12.03 (4.54)				
6 months	11.60 (3.67)	11.45 (4.63)				
9 months	11.36 (3.42)	11.06 (4.07)				
12 months	10.44 (3.66)	11.17 (4.10)				
BMI			.21	−0.238 (−0.528 to 0.051)	.23	−0.236 (−0.527 to 0.054)
Baseline	24.73 (3.38)	25.02 (3.28)				
3 months	24.49 (3.22)	24.68 (3.38)				
6 months	24.37 (3.13)	24.89 (3.16)				
9 months	24.23 (3.12)	24.73 (3.29)				
12 months	24.28 (3.14)	24.77 (3.36)				
WC[g]			<.001	−3.570 (−4.200 to −2.940)	<.001	−3.557 (−4.185 to −2.929)
Baseline	89.83 (9.68)	90.98 (9.89)				
3 months	89.06 (9.67)	91.46 (9.83)				
6 months	87.09 (9.21)	91.82 (9.46)				
9 months	87.02 (9.42)	92.50 (9.13)				
12 months	86.84 (9.24)	91.92 (9.85)				
WHR[h]			<.001	−0.020 (−0.025 to −0.015)	<.001	−0.020 (−0.024 to −0.013)
Baseline	0.926 (0.069)	0.929 (0.070)				
3 months	0.920 (0.073)	0.930 (0.080)				

XSL•FO
RenderX

Variables	Intervention group, mean (SD)	Control group, mean (SD)	Unadjusted effect		Adjusted effect[a]	
			P value[b]	β (95% CI)	*P* value	β (95% CI)
6 months	0.911 (0.068)	0.932 (0.073)				
9 months	0.909 (0.070)	0.938 (0.073)				
12 months	0.910 (0.068)	0.933 (0.067)				
SBP[i]			.007	−3.614 (−6.239 to −0.989)	.008	−3.556 (−6.187 to −0.925)
Baseline	132.53 (16.05)	134.26 (20.41)				
3 months	130.21 (13.26)	132.67 (16.37)				
6 months	128.53 (12.14)	132.92 (14.10)				
9 months	127.28 (11.24)	134.30 (17.08)				
12 months	127.30 (11.70)	132.54 (15.20)				
DBP[j]			.01	−2.683 (−4.342 to −1.024)	.01	−2.639 (−4.295 to −0.983)
Baseline	80.47 (9.03)	81.96 (10.71)				
3 months	78.72 (9.04)	81.77 (10.64)				
6 months	77.51 (8.16)	81.53 (9.45)				
9 months	77.12 (8.99)	82.09 (9.94)				
12 months	77.06 (7.94)	81.26 (8.52)				
Triglyceride			.01	−0.815 (−1.293 to −0.337)	.01	−0.814 (−1.293 to −0.335)
Baseline	3.09 (3.87)	2.53 (2.20)				
3 months	2.52 (2.54)	2.40 (2.37)				
6 months	2.37 (2.24)	2.44 (2.70)				
9 months	2.19 (2.30)	2.20 (1.53)				
12 months	1.93 (1.40)	2.24 (1.53)				
TC[k]			.001	−0.544 (−0.804 to 0.284)	.001	−0.540 (−0.798 to −0.282)
Baseline	5.12 (1.74)	4.84 (1.36)				
3 months	4.71 (1.18)	4.60 (1.09)				
6 months	4.43 (0.96)	4.47 (1.14)				
9 months	4.16 (0.77)	4.28 (1.12)				
12 months	4.12 (0.86)	4.37 (0.96)				
HDL-C[l]			.24	0.109 (0.014 to 0.205)	.28	0.109 (0.013 to 0.204)
Baseline	1.30 (0.53)	1.30 (0.60)				
3 months	1.33 (0.66)	1.26 (0.37)				
6 months	1.32 (0.45)	1.24 (0.49)				
9 months	1.33 (0.49)	1.25 (0.49)				
12 months	1.36 (0.39)	1.25 (0.56)				
LDL-C[m]			.40	−0.043 (−0.228 to 0.143)	.42	−0.048 (−0.232 to 0.137)
Baseline	2.96 (1.25)	2.92 (0.94)				
3 months	2.74 (0.96)	2.85 (0.93)				
6 months	2.78 (0.95)	2.78 (0.81)				

XSL•FO
RenderX

Variables	Intervention group, mean (SD)	Control group, mean (SD)	Unadjusted effect		Adjusted effect[a]	
			P value[b]	β (95% CI)	*P* value	β (95% CI)
9 months	2.79 (0.93)	2.77 (0.77)				
12 months	2.78 (0.83)	2.78 (0.80)				
Satisfaction			<.001	−6.271 (−7.191 to −5.352)	<.001	−4.314 (−5.176 to −3.451)
12 months	32.43 (4.38)	26.16 (5.59)				

[a]Adjusted for center, gender, age, comorbidity, history of smoking and alcohol, family history, education, income, occupation, and residence.

[b]Estimated group-by-time interaction effects from generalized estimation equation; time effect and group effect are displayed in Table S6 in Multimedia Appendix 3.

[c]HbA_{1c}: hemoglobin A_{1c}.

[d]ASCVD: atherosclerotic cardiovascular disease.

[e]FBG: fasting blood glucose.

[f]PBG2h: 2-hour postprandial blood glucose.

[g]WC: waist circumference.

[h]WHR: waist-to-hip ratio.

[i]SBP: systolic blood pressure.

[j]DBP: diastolic blood pressure.

[k]TC: total cholesterol.

[l]HDL-C: high-density lipoprotein cholesterol.

[m]LDL-C: low-density lipoprotein cholesterol.

Secondary Outcomes

FBG levels decreased significantly more in the intervention group (mean 7.58, SD 2.09 mmol/L) than in the control group (mean 8.22, SD 2.67 mmol/L; *P*<.001), although the group-by-time interaction effect was not significant (β=−0.602; *P*=.27; 95% CI −1.372 to 0.169). Several other secondary outcomes showed significant group-by-time interaction effects favoring the intervention group, including PBG2h (β=−1.210; *P*=.03; 95% CI −2.315 to −0.106), systolic blood pressure (β=−3.556; *P*=.008; 95% CI −6.187 to −0.925), diastolic blood pressure (β=−2.639; *P*=.01; 95% CI −4.295 to −0.983), triglyceride (β=−0.814; *P*=.01; 95% CI −1.293 to −0.335), and TC (β=−0.540; *P*=.001; 95% CI −0.798 to −0.282). However, there were no significant differences in high-density lipoprotein cholesterol (β=0.109; *P*=.28; 95% CI 0.013-0.204) and LDL-C (β=−0.048; *P*=.42; 95% CI −0.232 to 0.137) between the 2 groups in the long-term interventions. Detailed data are presented in Table S6 in Multimedia Appendix 3.

Significant improvements in indicators of central obesity were also observed. WC (β=−3.557; *P*<.001; 95% CI −4.185 to −2.929) and WHR (β=−0.020; *P*<.001; 95% CI −0.024 to −0.013) decreased significantly in the intervention group, showing a significant effect on weight loss in obese people. However, the change in BMI (β=−0.236; *P*=.23; 95% CI −0.527 to 0.054) was not significant between groups. Patients in the intervention group had a mean satisfaction score of 32.43 (SD 4.38), significantly higher than those in the control group with a mean score of 26.16 (SD 5.59; β=−4.314; *P*<.001; 95% CI −5.176 to −3.451).

Economic Evaluation

Over the 12 months, the cost of diabetes care was significantly lower in the intervention group (mean 5505.86, SD 3939.86 CNY) than in the control group (mean 7902.61, SD 11,218.51 CNY; *P*=.01; Table 3). The costs of registration, examination, lost wages, and transportation of patients were significantly higher in the intervention group than in the control group, while the hospitalization costs were significantly lower, indicating a shift from inpatient care toward more routine outpatient management.

XSL•FO
RenderX

Table 3. The 12-month total costs (in CNY[a]) of patients in the 2 groups (N=574).

Variables	Overall, mean (SD)	Intervention (n=291), mean (SD)	Control (n=283), mean (SD)	*P* value
Drug cost	3986.64 (2268.41)	3835.58 (2006.62)	4141.97 (2503.41)	.12
Registration cost	47.40 (31.40)	56.01 (29.01)	38.54 (31.34)	<.001
Examination cost	520.67 (741.48)	550.88 (323.56)	490.14 (999.86)	<.001
Hospitalization cost	1935.12 (7983.22)	849.72 (3299.63)	3051.21 (10,762.68)	.01
Lost wages	160.87 (276.69)	167.56 (301.06)	154.00 (249.52)	.02
Transportation cost	41.85 (63.14)	56.54 (69.86)	26.75 (51.28)	<.001
Total cost	6687.53 (8440.00)	5505.86 (3939.86)	7902.61 (11,218.51)	.01

[a]CNY 1=US $0.145.

Sensitivity Analysis

The per-protocol population included 510 patients who adhered to the protocol and were not lost to follow-up. There were no significant differences in demographic characteristics between the modified intention-to-treat and per-protocol populations (Table S7 in Multimedia Appendix 3). Notably, significant improvements in primary and secondary outcomes were detected in per-protocol population, including HbA_{1c}, 10-year ASCVD risk, glycemic measurements, weight-related outcomes, blood pressure, and blood lipid parameters (P<.05; Tables S8-S10 in Multimedia Appendix 3).

Discussion

Principal Findings

This study reported a digital-driven, multifaceted physician-pharmacist collaborative model for diabetes treatment and management, adapted for underresourced settings. The study also evaluated the long-term effectiveness and cost-effectiveness of implementing physician-pharmacist collaborative models on participants with diabetes in primary health care. The intervention was found to significantly improve long-term clinical efficacy and associated costs for patients over 12 months in primary health care settings, with the positive effects becoming more significant over time. In all cases, the magnitude of the improvements implied benefits for patients and public health.

The intervention program, based on the TPB framework, aims to enhance self-management in patients with diabetes by facilitating the establishment of behavioral intentions and motivation, was well as stimulating positive therapeutic behaviors [21,32]. Previous studies have found that TPB serves to predict individuals' motivation to engage in health behaviors, including exercise, health checks, disease screening, and health management behaviors [33]. However, most of the available studies are derived from high-income countries and regions, leaving very limited research on TPB-based interventions in rural populations with low education and economic status [21,33,34]. Our study suggests that the TPB-based intervention program contributes to the improvement of health behaviors among patients with diabetes in resource-limited areas, providing a strong basis for the widespread use of TPB in primary health care settings.

Our findings indicated a long-term advantage of physician-pharmacist collaborative models in improving HbA_{1c} and glycemic outcomes. The HbA_{1c} control rate in the intervention group increased to 51.45%, remarkably higher than the average control rate in rural China of 44.1% and close to the urban rate of 54.1%, suggesting a narrowing of the rural-urban disparities in diabetes treatment and management [35]. Moreover, there was an improvement in FBG at the early intervention, whereas the improvement in PBG2h was not observed until 12 months. This is supported by previous studies, showing that postprandial glycemic fluctuations tend to occur in patients with well-controlled diabetes, whereas patients with worsening diabetes have higher FBG [36-38]. Therefore, physician-pharmacist collaborative clinics have shown excellent management of glycemic outcomes in the early stages of diabetes, contributing to the delay of disease progression.

The results were consistent with other studies in high-income countries and regions, which have reported comprehensive improvements in blood glucose, blood pressure, and lipids in patients with diabetes [39,40]. Pharmacist-delivered services improved the quality of health care and reduced the burden of care, generating an annual return on investment of US $1.2 million to US $2.9 million [39]. A study of 2480 patients demonstrated that the pharmacist-physician collaborative care model resulted in significant improvements in HbA_{1c}, blood pressure, and lipid levels [40]. However, it is controversial for improving cardiovascular outcomes, while long-term cardiovascular risk was unclear [15,41]. Few retrospective multicenter cohort studies showed that the collaborative care model effectively decreased the HbA_{1c} levels but did not improve cardiovascular risk in primary care settings [42,43]. The potential explanation is the absence of an explicit theoretical framework to guide interventions, resulting in a lack of long-term motivation for patients to maintain healthy behaviors.

Health economic evaluation illustrated that physician-pharmacist collaborative models significantly reduced the total cost of diabetes care for patients while improving patient outcomes. This provided solid evidence to support the implementation of physician-pharmacist collaborative models in primary health care. Combined with care-seeking behavior, we found that the

XSL•FO
RenderX

number of outpatient visits increased by 78.7% and emergency visits decreased by half in physician-pharmacist collaborative models, indicating that the direct costs due to hospitalization for diabetes were greatly reduced [44]. In conclusion, it is economically feasible to implement physician-pharmacist collaborative clinics for diabetes management in primary health care settings.

In contrast to the previous study, significant reductions in WC and WHR, rather than BMI, were identified over 12 months in physician-pharmacist collaborative care models [45]. This inconsistency could be attributed to the peculiar physical conditions of the Chinese population, which leads to excessive fat deposition in the abdominal region instead of the peripheral region [46]. In China, the obese population primarily exhibits central obesity despite having a normal BMI, and waist-related indicators more accurately reflect obesity trends in this population [47,48]. In our study, the mean BMI at baseline was close to the normal range (BMI<24), and excessive WC was more common (WC≥90 cm for men and WC≥85 cm for women). This study suggests that physician-pharmacist collaborative models achieved significant improvements in WC and WHR, indicating that patients with diabetes and central obesity may benefit more from collaborative clinics.

In addition, our findings supported that physician-pharmacist collaborative models improve blood pressure, TC, and triglyceride levels and significantly reduce the 10-year ASCVD risk. However, we failed to find a significant enhancement in LDL-C, which is inconsistent with previous studies [49]. This discrepancy may be attributed to the interaction of our intervention focus and population characteristics. First, the primary focus of our multifaceted intervention was comprehensive glycemic control; lipid management was addressed through education but was not part of a protocol-driven intervention. Second, consistent with this context, the enrolled population in these primary care settings had a relatively favorable lipid profile, with mean LDL-C already near the recommended target for low-risk patients (LDL-C [without ASCVD]<2.6 mmol/L) [24,50,51]. Studies suggested that pharmacist interventions have more modest effects on LDL-C in such populations [52]. Overall, the collaborative models led to integrated improvements and facilitated effective, comprehensive health management.

Strengths, Generalizability, and Implementation Considerations

This study has several strengths. First, as one of the few implementation studies on physician-pharmacist collaborative clinics with a long-term follow-up, we developed and evaluated a collaborative model for underresourced settings. A key innovation was the integration of the mHealth component, which enabled continuous support and education, overcoming geographical barriers and facilitating sustained patient engagement beyond traditional, infrequent clinic visits. While this model proved effective in primary care settings in China, its broader application requires consideration of local contexts. Key factors may include the existing health care system, cultural acceptance of team-based care, and digital tools. Second, the physician-pharmacist collaborative clinics potentially bridge the rural-urban gap in medical resources disparities for diabetes management. Moreover, the physician-pharmacist collaborative care model could serve as an evidence-based management model for patients with chronic diseases and a collaborative model for team-based care in underresourced settings. In future implementations, a pilot phase should be given priority so that the core components, structured collaboration and digitally augmented support, can be adapted to local resources and patient needs.

Limitations

This study had several limitations. First, although the rate was comparable between the intervention and control groups, and baseline characteristics were similar between intention-to-treat and per-protocol populations, attrition bias cannot be entirely ruled out. Second, unobserved confounding bias could not be excluded from this study. Potential sources include differences in medication adherence or undisclosed lifestyle changes, as well as unmeasured variations in local health care resources or practice patterns. Third, this study was affected by the COVID-19 pandemic, which reduced patients' willingness to seek care and introduced unpredictable bias. Fourth, the sample size was calculated on the strength of the primary outcome; therefore, we may not have sufficient power to detect other meaningful, significant outcomes. Blinding was not feasible in this study, and a Hawthorne effect may have occurred. Finally, while the final 6 study centers were randomly selected and allocated, the initial identification of candidate centers relied on a pragmatic criterion, which may introduce selection bias.

In conclusion, this trial provides strong evidence to support the implementation of the physician-pharmacist collaborative care model in long-term diabetes treatment and management in primary health care settings in underresourced areas. Patients with T2D, especially those with central obesity or high 10-year ASCVD risk, may benefit more from physician-pharmacist collaborative clinics and should be the focus of further widespread implementation. This theory-based intervention program is a promising multidisciplinary model for improving the quality and efficiency of diabetes treatment and management, as well as other chronic diseases in underresourced settings globally.

Acknowledgments

The authors would like to thank all patients, physicians, and pharmacists for participating in the trial.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Funding

This study was funded by the Chronic Disease Management Research Project of the National Health Commission Capacity Building and Continuing Education Center (YXFSC2022JJSJ010), China Medical Board (grant 18-292), and Hunan Provincial Health Commission (grant 2019158). The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Authors' Contributions

PX, DX, HY, and XL conceptualized and designed this study; YG, HH, DZ, YL, LW, QL, and JL collected the data; JX, QW, SH, and BT performed the statistical analysis; and PX, JX, ST, and LC contributed to the interpretation of the results and drafted the manuscript. All authors have reviewed, revised, and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT checklist.
[PDF File (Adobe PDF File), 209 KB - jmir_v28i1e77470_app1.pdf]

Multimedia Appendix 2
Fidelity checklist.
[PDF File (Adobe PDF File), 195 KB - jmir_v28i1e77470_app2.pdf]

Multimedia Appendix 3
Methods and results.
[PDF File (Adobe PDF File), 469 KB - jmir_v28i1e77470_app3.pdf]

References

1. GBD 2016 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet 2017 Sep 16;390(10100):1211-1259 [FREE Full text] [doi: 10.1016/S0140-6736(17)32154-2] [Medline: 28919117]
2. Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, et al. IDF Diabetes Atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes Res Clin Pract 2022 Jan;183:109119 [FREE Full text] [doi: 10.1016/j.diabres.2021.109119] [Medline: 34879977]
3. Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. Nat Rev Endocrinol 2018 Feb;14(2):88-98. [doi: 10.1038/nrendo.2017.151] [Medline: 29219149]
4. Carlson AL, Mullen DM, Bergenstal RM. Clinical use of continuous glucose monitoring in adults with type 2 diabetes. Diabetes Technol Ther 2017 May;19(S2):S4-11 [FREE Full text] [doi: 10.1089/dia.2017.0024] [Medline: 28541137]
5. Kanaley JA, Colberg SR, Corcoran MH, Malin SK, Rodriguez NR, Crespo CJ, et al. Exercise/physical activity in individuals with type 2 diabetes: a consensus statement from the American College of Sports Medicine. Med Sci Sports Exerc 2022 Feb 01;54(2):353-368. [doi: 10.1249/MSS.0000000000002800] [Medline: 35029593]
6. Nauck MA, Wefers J, Meier JJ. Treatment of type 2 diabetes: challenges, hopes, and anticipated successes. Lancet Diabetes Endocrinol 2021 Aug;9(8):525-544. [doi: 10.1016/S2213-8587(21)00113-3] [Medline: 34181914]
7. Lin CW, Abdul SS, Clinciu DL, Scholl J, Jin X, Lu H, et al. Empowering village doctors and enhancing rural healthcare using cloud computing in a rural area of mainland China. Comput Methods Programs Biomed 2014 Feb;113(2):585-592. [doi: 10.1016/j.cmpb.2013.10.005] [Medline: 24296075]
8. Zhao X, Zhang Y, Yang Y, Pan J. Diabetes-related avoidable hospitalisations and its relationship with primary healthcare resourcing in China: a cross-sectional study from Sichuan Province. Health Soc Care Community 2022 Jul;30(4):e1143-e1156. [doi: 10.1111/hsc.13522] [Medline: 34309097]
9. Wang Y, Li Y, Qin S, Kong Y, Yu X, Guo K, et al. The disequilibrium in the distribution of the primary health workforce among eight economic regions and between rural and urban areas in China. Int J Equity Health 2020 Feb 26;19(1):28 [FREE Full text] [doi: 10.1186/s12939-020-1139-3] [Medline: 32102655]
10. Firkus D, McCoy RG, Matulis J3, Kessler M, Mara K, Herges J. Evaluation of pharmacist consults within a collaborative enhanced primary care team model to improve diabetes care. PLoS One 2023;18(1):e0280654 [FREE Full text] [doi: 10.1371/journal.pone.0280654] [Medline: 36662741]
11. Lee JY, Lee SW. Telemedicine cost-effectiveness for diabetes management: a systematic review. Diabetes Technol Ther 2018 Jul;20(7):492-500. [doi: 10.1089/dia.2018.0098] [Medline: 29812965]

XSL•FO
RenderX

12. Munshi MN, Sy S, Lekarcyk J, Sullivan E. A successful diabetes management model of care in long-term care facilities. J Am Med Dir Assoc 2021 Jun;22(6):1322-6.e2. [doi: 10.1016/j.jamda.2020.06.046] [Medline: 32753320]
13. Anderegg MD, Gums TH, Uribe L, MacLaughlin EJ, Hoehns J, Bazaldua OV, et al. Pharmacist intervention for blood pressure control in patients with diabetes and/or chronic kidney disease. Pharmacotherapy 2018 Mar;38(3):309-318 [FREE Full text] [doi: 10.1002/phar.2083] [Medline: 29331037]
14. Pardo A, Ntabaza V, Rivolta M, Goulard A, Sténuit S, Demeester R, et al. Impact of collaborative physician-pharmacist stewardship strategies on prophylactic antibiotic practices: a quasi-experimental study. Antimicrob Resist Infect Control 2022 Jul 26;11(1):100 [FREE Full text] [doi: 10.1186/s13756-022-01138-3] [Medline: 35883189]
15. Norton MC, Haftman ME, Buzzard LN. Impact of physician-pharmacist collaboration on diabetes outcomes and health care use. J Am Board Fam Med 2020;33(5):745-753 [FREE Full text] [doi: 10.3122/jabfm.2020.05.200044] [Medline: 32989069]
16. Wu S, Tannous E, Haldane V, Ellen ME, Wei X. Barriers and facilitators of implementing interventions to improve appropriate antibiotic use in low- and middle-income countries: a systematic review based on the Consolidated Framework for Implementation Research. Implement Sci 2022 May 12;17(1):30 [FREE Full text] [doi: 10.1186/s13012-022-01209-4] [Medline: 35550169]
17. Miles LM, Hawkes RE, French DP. How the behavior change content of a nationally implemented digital diabetes prevention program is understood and used by participants: qualitative study of fidelity of receipt and enactment. J Med Internet Res 2023 Jan 11;25:e41214 [FREE Full text] [doi: 10.2196/41214] [Medline: 36630165]
18. Jiang H, Mi Z, Xu W. Online medical consultation service-oriented recommendations: systematic review. J Med Internet Res 2024 Jul 30;26:e46073 [FREE Full text] [doi: 10.2196/46073] [Medline: 38777810]
19. Seng JJ, Gwee MF, Yong MH, Kwan YH, Thumboo J, Low LL. Role of caregivers in remote management of patients with type 2 diabetes mellitus: systematic review of literature. J Med Internet Res 2023 Sep 11;25:e46988 [FREE Full text] [doi: 10.2196/46988] [Medline: 37695663]
20. Karloh M, Matias TS, de Oliveira JM, de Lima FF, Araújo Pinheiro DH, Barbosa GB, et al. Breaking barriers to rehabilitation: the role of behavior change theories in overcoming the challenge of exercise-related behavior change. Braz J Phys Ther 2023;27(6):100574 [FREE Full text] [doi: 10.1016/j.bjpt.2023.100574] [Medline: 38056192]
21. Paul B, Kirubakaran R, Isaac R, Dozier M, Grant L, Weller D, RESPIRE Collaboration. A systematic review of the theory of planned behaviour interventions for chronic diseases in low health-literacy settings. J Glob Health 2023 Sep 08;13:04079 [FREE Full text] [doi: 10.7189/jogh.13.04079] [Medline: 37681679]
22. Tan SL, Xiao J, Yuan HY, Chen L, Wang Q, Xiang DX, et al. Establishing and evaluating physician-pharmacist collaborative clinics to manage patients with type 2 diabetes in primary hospitals in Hunan province: study protocol of a multi-site randomized controlled trial in the era of COVID-19 pandemic. BMC Health Serv Res 2022 Mar 04;22(1):299 [FREE Full text] [doi: 10.1186/s12913-022-07653-8] [Medline: 35246117]
23. Bellary S, Kyrou I, Brown JE, Bailey CJ. Type 2 diabetes mellitus in older adults: clinical considerations and management. Nat Rev Endocrinol 2021 Sep;17(9):534-548. [doi: 10.1038/s41574-021-00512-2] [Medline: 34172940]
24. Chinese Diabetes Society. Guidelines for the prevention and treatment of type 2 diabetes in China (2020 edition). Chin J Endocrinol Metab 2021;37(4):311-398 [FREE Full text] [doi: 10.3760/cma.j.cn311282-20210304-00142]
25. Bao Y, Wang C, Xu H, Lai Y, Yan Y, Ma Y, et al. Effects of an mHealth intervention for pulmonary tuberculosis self-management based on the integrated theory of health behavior change: randomized controlled trial. JMIR Public Health Surveill 2022 Jul 14;8(7):e34277 [FREE Full text] [doi: 10.2196/34277] [Medline: 35834302]
26. Luo T, Li MS, Williams D, Fritz J, Beiter K, Phillippi S, et al. A WeChat-based smoking cessation intervention for Chinese smokers: a feasibility study. Transl Behav Med 2022 Nov 16;12(10):1018-1027. [doi: 10.1093/tbm/ibac067] [Medline: 36130313]
27. Yang X, Li J, Hu D, Chen J, Li Y, Huang J, et al. Predicting the 10-year risks of atherosclerotic cardiovascular disease in Chinese population: the China-PAR project (Prediction for ASCVD Risk in China). Circulation 2016 Nov 08;134(19):1430-1440. [doi: 10.1161/CIRCULATIONAHA.116.022367] [Medline: 27682885]
28. Olomu A, Tikaria R, Kelly-Blake K, Hart-Davidson W, Wang L, Alroshood Z, et al. Type 2 diabetes patient activation and mHealth interventions decreased cardiovascular disease risk. Am J Manag Care 2022 Nov 01;28(11):e392-e398 [FREE Full text] [doi: 10.37765/ajmc.2022.89263] [Medline: 36374656]
29. Collins C, Limone BL, Scholle JM, Coleman CI. Effect of pharmacist intervention on glycemic control in diabetes. Diabetes Res Clin Pract 2011 May;92(2):145-152. [doi: 10.1016/j.diabres.2010.09.023] [Medline: 20961643]
30. Gold R, Larson AE, Sperl-Hillen JM, Boston D, Sheppler CR, Heintzman J, et al. Effect of clinical decision support at community health centers on the risk of cardiovascular disease: a cluster randomized clinical trial. JAMA Netw Open 2022 Feb 01;5(2):e2146519 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.46519] [Medline: 35119463]
31. Lumu W, Kibirige D, Wesonga R, Bahendeka S. Effect of a nurse-led lifestyle choice and coaching intervention on systolic blood pressure among type 2 diabetic patients with a high atherosclerotic cardiovascular risk: study protocol for a cluster-randomized trial. Trials 2021 Feb 11;22(1):133 [FREE Full text] [doi: 10.1186/s13063-021-05085-z] [Medline: 33573687]

XSL•FO
RenderX

32. Armitage CJ, Conner M. Efficacy of the theory of planned behaviour: a meta-analytic review. Br J Soc Psychol 2001 Dec;40(Pt 4):471-499. [doi: 10.1348/014466601164939] [Medline: 11795063]
33. Khani Jeihooni A, Layeghiasl M, Yari A, Rakhshani T. The effect of educational intervention based on the theory of planned behavior on improving physical and nutrition status of obese and overweight women. BMC Womens Health 2022 Jan 15;22(1):13 [FREE Full text] [doi: 10.1186/s12905-022-01593-5] [Medline: 35033045]
34. Khani Jeihooni A, Jormand H, Saadat N, Hatami M, Abdul Manaf R, Afzali Harsini P. The application of the theory of planned behavior to nutritional behaviors related to cardiovascular disease among the women. BMC Cardiovasc Disord 2021 Dec 07;21(1):589 [FREE Full text] [doi: 10.1186/s12872-021-02399-3] [Medline: 34876014]
35. Wang L, Peng W, Zhao Z, Zhang M, Shi Z, Song Z, et al. Prevalence and treatment of diabetes in China, 2013-2018. JAMA 2021 Dec 28;326(24):2498-2506. [doi: 10.1001/jama.2021.22208] [Medline: 34962526]
36. Moon J, Kim JY, Yoo S, Koh G. Fasting and postprandial hyperglycemia: their predictors and contributions to overall hyperglycemia in Korean patients with type 2 diabetes. Endocrinol Metab (Seoul) 2020 Jun;35(2):290-297 [FREE Full text] [doi: 10.3803/EnM.2020.35.2.290] [Medline: 32615713]
37. Riddle M, Umpierrez G, DiGenio A, Zhou R, Rosenstock J. Contributions of basal and postprandial hyperglycemia over a wide range of A1C levels before and after treatment intensification in type 2 diabetes. Diabetes Care 2011 Dec;34(12):2508-2514 [FREE Full text] [doi: 10.2337/dc11-0632] [Medline: 22028279]
38. Monnier L, Lapinski H, Colette C. Contributions of fasting and postprandial plasma glucose increments to the overall diurnal hyperglycemia of type 2 diabetic patients: variations with increasing levels of HbA(1c). Diabetes Care 2003 Mar;26(3):881-885. [doi: 10.2337/diacare.26.3.881] [Medline: 12610053]
39. Lehn JM, Gerkin RD, Kisiel SC, O'Neill L, Pinderhughes ST. Pharmacists providing palliative care services: demonstrating a positive return on investment. J Palliat Med 2019 Jun;22(6):644-648. [doi: 10.1089/jpm.2018.0082] [Medline: 30570421]
40. Matzke GR, Moczygemba LR, Williams KJ, Czar MJ, Lee WT. Impact of a pharmacist-physician collaborative care model on patient outcomes and health services utilization. Am J Health Syst Pharm 2018 Jul 15;75(14):1039-1047. [doi: 10.2146/ajhp170789] [Medline: 29789318]
41. Aguiar PM, da Silva CH, Chiann C, Dórea EL, Lyra DP, Storpirtis S. Pharmacist-physician collaborative care model for patients with uncontrolled type 2 diabetes in Brazil: results from a randomized controlled trial. J Eval Clin Pract 2018 Feb;24(1):22-30. [doi: 10.1111/jep.12606] [Medline: 27452158]
42. Hinds A, Lopez D, Rascati K, Jokerst J, Srinivasa M. Adherence to the 2013 blood cholesterol guidelines in patients with diabetes at a PCMH: comparison of physician only and combination physician/pharmacist visits. Diabetes Educ 2016 Apr;42(2):228-233. [doi: 10.1177/0145721716631431] [Medline: 26902526]
43. Peterson J, Hinds A, Garza A, Barner J, Hill L, Nguyen M, et al. Impact of physician-pharmacist covisits at a primary care clinic in patients with uncontrolled diabetes. J Pharm Pract 2020 Jun;33(3):321-325. [doi: 10.1177/0897190018807374] [Medline: 30428760]
44. Xiao J, Wang Q, Tan S, Chen L, Tang B, Huang S, et al. Analysis of patient medication compliance and quality of life of physician-pharmacist collaborative clinics for T2DM management in primary healthcare in China: a mixed-methods study. Front Pharmacol 2023 Mar 24;14:1098207 [FREE Full text] [doi: 10.3389/fphar.2023.1098207] [Medline: 37033638]
45. Ledford JL, Hess R, Johnson FP. Impact of clinical pharmacist collaboration in patients beginning insulin pump therapy: a retrospective and cross-sectional analysis. J Drug Assess 2013;2(1):81-86 [FREE Full text] [doi: 10.3109/21556660.2013.815624] [Medline: 27536441]
46. Du T, Sun X, Yin P, Huo R, Ni C, Yu X. Increasing trends in central obesity among Chinese adults with normal body mass index, 1993-2009. BMC Public Health 2013 Apr 10;13:327 [FREE Full text] [doi: 10.1186/1471-2458-13-327] [Medline: 23575244]
47. Wu J, Zou L, Liu Y, Yu H, Yin H, Zhong L, et al. Association between obesity indicators and cardiometabolic disease in Chinese adults. PLoS One 2023 Jan 20;18(1):e0273235 [FREE Full text] [doi: 10.1371/journal.pone.0273235] [Medline: 36662790]
48. Jia Z, Zhou Y, Liu X, Wang Y, Zhao X, Wang Y, et al. Comparison of different anthropometric measures as predictors of diabetes incidence in a Chinese population. Diabetes Res Clin Pract 2011 May;92(2):265-271. [doi: 10.1016/j.diabres.2011.01.021] [Medline: 21334088]
49. Elnaem MH, Rosley NF, Alhifany AA, Elrggal ME, Cheema E. Impact of pharmacist-led interventions on medication adherence and clinical outcomes in patients with hypertension and hyperlipidemia: a scoping review of published literature. J Multidiscip Healthc 2020;13:635-645 [FREE Full text] [doi: 10.2147/JMDH.S257273] [Medline: 32764955]
50. Ruixing Y, Qiming F, Dezhai Y, Shuquan L, Weixiong L, Shangling P, et al. Comparison of demography, diet, lifestyle, and serum lipid levels between the Guangxi Bai Ku Yao and Han populations. J Lipid Res 2007 Dec;48(12):2673-2681 [FREE Full text] [doi: 10.1194/jlr.M700335-JLR200] [Medline: 17890682]
51. Deng B, Luo T, Huang Y, Shen T, Ma J. Prevalence and determinants of hyperlipidemia in moderate altitude areas of the Yunnan-Kweichow plateau in Southwestern China. High Alt Med Biol 2012 Mar;13(1):13-21 [FREE Full text] [doi: 10.1089/ham.2011.1037] [Medline: 22429228]

XSL•FO
RenderX

52. Dixon DL, Baker WL, Buckley LF, Salgado TM, Van Tassell BW, Carter BL. Effect of a physician/pharmacist collaborative care model on time in target range for systolic blood pressure: post hoc analysis of the CAPTION trial. Hypertension 2021 Sep;78(4):966-972 [FREE Full text] [doi: 10.1161/HYPERTENSIONAHA.121.17873] [Medline: 34397278]

Abbreviations

ASCVD: atherosclerotic cardiovascular disease
CONSORT: Consolidated Standards of Reporting Trials
FBG: fasting blood glucose
GEE: generalized estimating equation
HbA_{1c}: glycosylated hemoglobin
LDL-C: low-density lipoprotein cholesterol
mHealth: mobile health
PBG2h: 2-hour postprandial blood glucose
T2D: type 2 diabetes
TC: total cholesterol
TPB: theory of planned behavior
WC: waist circumference
WHR: waist-to-hip ratio

Edited by A Stone; submitted 13.May.2025; peer-reviewed by G Ni, Y Shin; comments to author 02.Jan.2026; accepted 18.Feb.2026; published 13.Mar.2026.

Please cite as:
Xiao J, Wang Q, Tan S, Chen L, Xiang D, Yuan H, Li X, Huang S, Tang B, Guo Y, Huang H, Zhao D, Li Y, Wang L, Li Q, Liu J, Xu P
Effect of a Digital-Driven Physician-Pharmacist Collaborative Model for Diabetes in Primary Health Care: Cluster Randomized Trial
J Med Internet Res 2026;28:e77470
URL: https://www.jmir.org/2026/1/e77470
doi:10.2196/77470
PMID:

©Jie Xiao, Qing Wang, Shenglan Tan, Lei Chen, Daxiong Xiang, Haiyan Yuan, Xia Li, Shuting Huang, Bingjie Tang, Yan Guo, Haiying Huang, Danhui Zhao, Yue Li, Li Wang, Qun Li, Juan Liu, Ping Xu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Coproducing an Online Platform for People With Long-Term Physical Health Conditions: Development and Usability Study

Hannah Grace Jones[1,2,3*], BSc; Grace Lavelle[1,4*], PhD; Elly Aylwin-Foster[1], BA; Ewan Carr[5], PhD; Alan Simpson[6,7], PhD; Matthew Hotopf[1], PhD; Vanessa Lawrence[7], PhD

[1]Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[2]Lifelong Health, Persistent Pain Research Group, Hopwood Centre for Neurobiology, South Australian Health and Medical Research Institute, Adelaide, South Australia, Australia

[3]IIMPACT (Innovation, Implementation and Clinical Translation) in Health, School of Allied Health and Human Performance, Adelaide University, Adelaide, South Australia, Australia

[4]CP-Life Research Centre, School of Physiotherapy, Royal College of Surgeons in Ireland, Dublin, Leinster, Ireland

[5]Department of Biostatistics & Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[6]Florence Nightingale Faculty of Nursing, Midwifery & Palliative Care, King's College London, London, United Kingdom

[7]Health Service and Population Research, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Hannah Grace Jones, BSc
Department of Psychological Medicine
Institute of Psychiatry, Psychology and Neuroscience
King's College London
16 De Crespigny Park
London, SE5 8AB
United Kingdom
Phone: 44 02078480002
Email: hannah.g.jones@kcl.ac.uk

Abstract

Background: There is relatively limited psychological support dedicated to people living with long-term physical health conditions and subthreshold depressive disorder. Online peer support may be an appropriate intervention to help bolster patients' mental well-being to prevent progression of their symptoms to major depressive disorder. For interventions to be successfully integrated into the self-management routines of people with long-term physical health conditions, they should be co-designed to ensure that they align with the wants and needs of the target audience.

Objective: This study aims to coproduce an online peer support intervention with people with lived experience, software experts, clinicians, and academics through an iterative process of co-design and subsequent co-validation through usability testing.

Methods: We followed a 4-stage coproduction process: co-assess, co-design, co-validate, and co-deliver. Our research advisory group was actively involved in all stages, consisting of 1 coinvestigator and 6 people with lived experience of long-term physical and/or mental health comorbidities. The co-assess and co-design stages involved our participatory design panel, which included 10 members living with various long-term conditions. The participatory design panel participated in online focus groups to assess their unmet psychosocial needs and then co-designed the intervention prototype through online workshops with software developers. The co-validation stage involved an additional group of participants (n=12) with long-term physical health conditions. During co-validation, the prototype underwent usability testing, including think-aloud exercises and semistructured interviews. Content analysis identified the priorities for the iterative development that formed the basis of further research advisory group co-design workshops. The next stage, co-delivery, involved coproducing the protocol of a feasibility and acceptability randomized controlled trial.

Results: Participants highlighted that a platform must feel safe and trustworthy for the space to support the mental well-being of those living with long-term health conditions. The participatory design panel co-designed a platform prototype to meet this need. During the co-validation stage, the think-aloud exercises identified common issues related to navigation challenges and feature glitches. Content analysis of the semistructured interviews confirmed that the community forum, resources, and other

XSL•FO
RenderX

platform pages were appropriate and acceptable, but revealed usability concerns. Participants stressed the need for intuitive navigation and suggested new features that would enhance user experience. Facilitators and barriers to engagement were also noted, including the importance of fostering trust in the platform's ethos and branding to create a safe space. Through iterative development and subsequent usability testing, the final prototype was approved.

Conclusions: We have provided a worked example of a comprehensive, coproduction process where we worked alongside people with lived experience to successfully design an online peer support platform with embedded psychoeducation. The platform, called CommonGround, is ready to be evaluated in a feasibility randomized controlled trial.

(J Med Internet Res 2026;28:e79666) doi:10.2196/79666

KEYWORDS

long-term conditions; peer support; coproduction, physical health; mental health; psychosocial support; online peer support; usability testing

Introduction

An estimated 15 million people are living with a long-term physical health condition in the United Kingdom [1]. Up to 20% are living with comorbid major depressive disorder [2,3], and a further 20% are experiencing subthreshold depressive disorder [4]. Subthreshold depressive disorder is characterized by experiencing between 2 and 5 depressive symptoms for at least 2 weeks, one of which must be either low mood or loss of interest/pleasure [5,6]. Although living with subthreshold depressive disorder is a key risk factor for developing major depressive disorder [7,8], the provision of psychological support to help prevent the worsening of these existing depressive symptoms among people with long-term physical health conditions is relatively limited. Targeted approaches to support the mental well-being of this at-risk population are therefore urgently needed.

A critical component in the effective management of long-term physical health conditions is self-management, whereby patients take an active role in the day-to-day management of their symptoms, treatments, and necessary lifestyle modifications [9,10]. As such, in the development of any tool to support the well-being of people with long-term physical health conditions, it is critical that intervention design and implementation recognize this need to self-manage effectively and should not add undue burden [11,12]. Indeed, empowering patients to address their own needs by providing tools that support their self-management has been associated with improved health outcomes [13,14] and health behaviors [15].

One potential solution to support patients' self-management of their psychosocial well-being could be peer support, where people with similar life experiences or characteristics (ie, living with a long-term physical health condition) are able to share emotional, informational, and appraisal support with one another to improve their physical and psychosocial health [16-18]. Peer support can be delivered in a variety of formats, from one-to-one to group-based support, and in-person or online. Recently, the provision of online peer support has increased, as it is typically low-cost, easily scalable, and offers a flexible, convenient solution that overcomes the physical barriers people may face in accessing in-person groups [19]. The positive effects of engaging with online peer support resemble those of offline peer support [20], including reciprocal emotional and social support, patient empowerment, improved quality of life, and mental well-being [17,21-24].

Online peer support interventions therefore have the potential to reduce the burden of long-term conditions and the associated psychological distress [16] but also reduce the burden of self-management by improving the tools one has available to self-manage [25]. However, to date, the development of online peer support interventions has largely been driven by academic researchers and industry with limited input from the intended end user [26]. As a result, interventions may not meet end users' needs and perspectives, and be inappropriately designed, with missing features or poor interfaces [27]. Such poor design can ultimately limit users' long-term motivation to engage with an intervention, for instance, if they encounter difficulties in trying to log-in [27]. The lack of involvement of end users in design is particularly problematic when the end users are people living with long-term physical health conditions, as this population has complex needs and requires interventions that can be easily incorporated into their existing self-management routines [28]. Indeed, in our prior work, people living with a variety of long-term physical health conditions emphasized that this would be essential to ensure that the intervention feels safe, credible, and appropriate for their needs [29]. Coproduction is therefore vital, whereby an online peer support platform is created through an equal partnership between lived experience experts, researchers, and other partners with the underlying ethos of developing *with* and not *for* the users [30].

Coproduction can enhance the relevance, acceptability, and usability of interventions, thereby facilitating patients' long-term engagement. The diversity of perspectives fosters innovative and creative interventions, ensuring that features and functions are uniquely designed and rooted in lived experience [27,31]. A critical aspect of coproduction is ensuring the perspectives of lived experience experts have been accurately translated into components of the intervention [27]. One means of ensuring this is by evaluating the usability and acceptability of the co-designed prototype and allowing time for further iterative development as required [32]. Think-aloud exercises are valuable in usability testing for identifying specific usability issues in common user tasks, for instance, glitches encountered when creating an account or creating a forum post [33]. The analysis of think-aloud exercises can also reveal unmet user needs and highlight areas for improvement, guiding future prototype modification [34,35]. To assess acceptability,

XSL•FO
RenderX

qualitative interviews are useful as users' perceptions of acceptability are deeply contextualized (eg, in relation to their long-term physical health condition) and shaped by social and cultural factors [36]. Together, the findings from usability testing can inform the iterative development of the intervention and its ability to satisfy user needs. Ultimately, this process leads to a final approved prototype, which can then undergo real-world evaluation of safety and efficacy in randomized controlled trials.

Although the benefits of coproduction are widely recognized, there is a significant lack of research evidence documenting the coproduction of real-world interventions [37], particularly those dedicated to supporting psychological well-being in the context of physical health [38]. In this paper, we report how we coproduced "CommonGround," our online peer support and psychoeducation platform dedicated to people living with long-term physical health conditions. This study aimed to coproduce an online peer support platform with embedded psychoeducation for people living with long-term conditions and subthreshold depression to support their psychological well-being. To fulfill this aim, the following objectives were set: (1) to identify the psychosocial needs of people living with long-term physical health conditions, (2) to co-design an online peer support platform prototype that aligns with their wants and needs, and (3) to co-validate the prototype through usability testing and iterative development work to produce a final platform prototype that can be evaluated in a future feasibility randomized controlled trial.

Methods

Design

The CommonGround project followed a 4-stage coproduction framework adapted from Elwyn and colleagues [39] comprising (1) co-assess, (2) co-design, (3) co-validate, and (4) co-deliver (Figure 1).

Figure 1. The coproduction cycle of CommonGround, adapted from Elwyn et al [39]. RCT: randomized controlled trial.

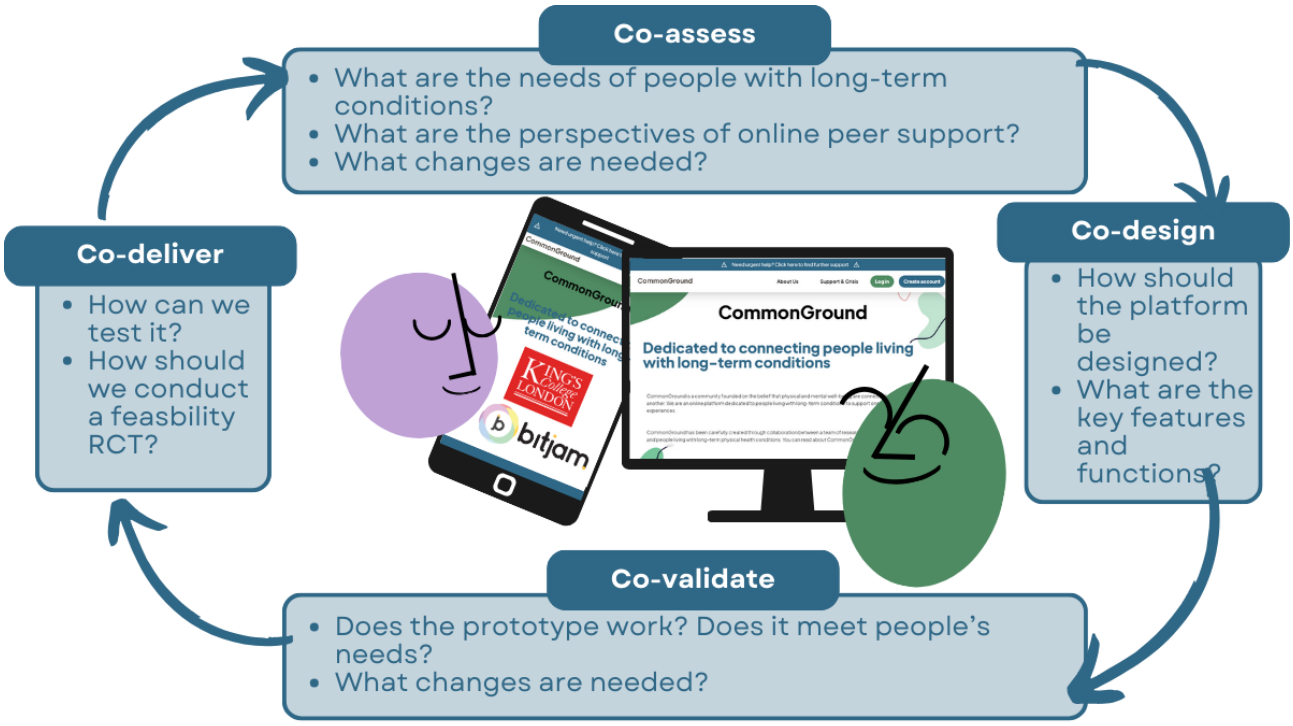

Ethical Considerations

Full ethical approval for the online focus groups and usability testing was sought and granted by King's College London Research Ethics Office, PNM Research Ethics Subcommittee (HR-19/20-14938). Electronic informed consent was obtained before taking part in the relevant research activities (see Multimedia Appendix 1). The procedures to maintain participant confidentiality and anonymity were outlined in the consent forms. All data were pseudoanonymized using unique study IDs, password-protected, and stored on secure servers. Audio-recordings from focus groups and interviews were shared securely with a transcription service, who have signed nondisclosure agreements and deleted audio-recordings immediately following transcription. Participants received a £20 (US $26.9) voucher for participation in the focus groups. The same reimbursement was provided for participation in the usability testing.

Patient and Public Involvement

At each stage of the coproduction cycle, different stakeholders and groups of people with lived experience were involved (Figure 2). One of the authors and a coinvestigator on the project has lived experience of multiple long-term health conditions (cystic fibrosis, type 1 diabetes, and adrenal insufficiency) and, throughout the project, has been a core member of the research team and remunerated for their time. Their role included chairing our independent research advisory group (RAG) of people with lived experience who guided the whole research process, including the coproduction of study documents and key decision-making processes. Our RAG involved 6 members (gender identity: 4 women and 2 men; age range: 31-71 years;

XSL•FO
RenderX

ethnicities: Black British, Caribbean, White British, and White European) living with various long-term physical health conditions. Three members were living with multiple long-term conditions, and 4 had comorbid mental health diagnoses. RAG members were identified and invited to join through our coinvestigator with lived experience and through patient representative groups based within Guy's and St Thomas Hospital and King's College Hospital, and King's College London's Integrated Care Consultation Partners Group, and via snowballing through recruited members. All recruitment materials emphasized the focus on preventing depression in the context of chronic physical illness, and any individual that felt that they had relevant lived experience regarding psychological distress and well-being in the context of their physical health was able to take part (ie, with or without a current or past clinical diagnosis of depression; no formal assessment of depressive symptoms was administered). This inclusive approach not only sought to be representative of a range of lived experiences but was also sensitive to the fact that many people may not recognize that they are experiencing subthreshold depressive symptoms and that these symptoms can also fluctuate over time. Terms of reference were agreed in accordance with UK standards for public involvement [40]. Our RAG members were reimbursed at an hourly rate based on patient and public involvement (PPI) reimbursement recommendations for their time preparing for meetings, attending meetings, and reviewing study documents [41,42]. Members were invited to share feedback outside of any meetings according to their personal preference (eg, via email).

Figure 2. The different stakeholders and groups of people with lived experience involved in the CommonGround coproduction cycle. LTC: long-term condition; RAG: research advisory group.

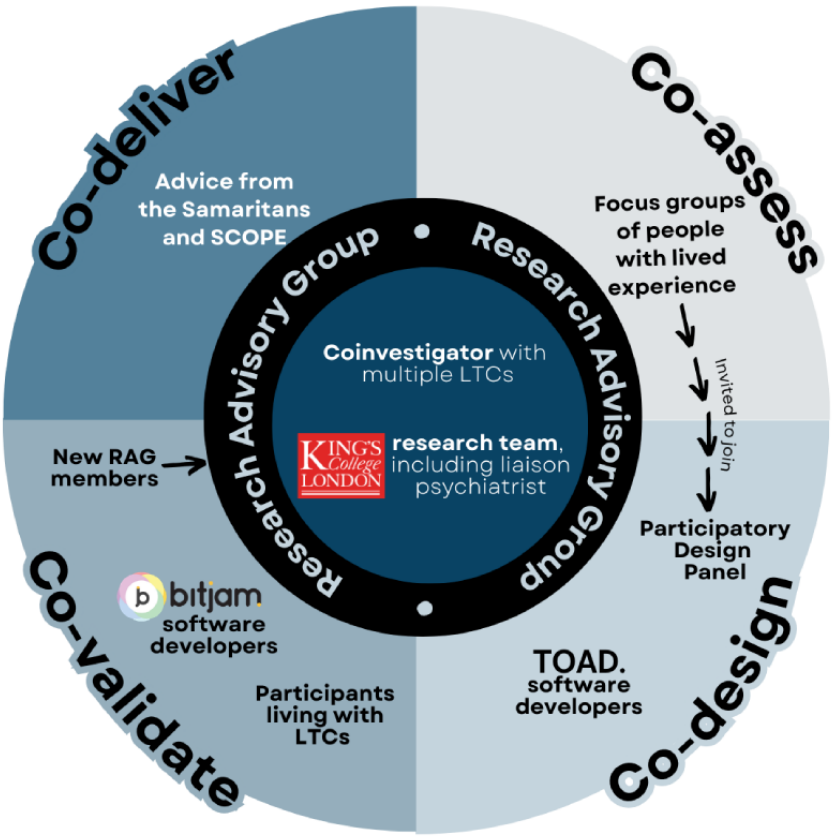

The Coproduction Cycle

The coproduction cycle had 4 stages: co-assess, co-design, co-validate, and co-deliver (Figure 1) [39], which are outlined below.

Co-Assess

The co-assess stage addressed the question: "What are the unmet psychosocial needs of people living with long-term physical health conditions, and how might an online peer support platform address these needs?"

Three online focus groups were conducted, following a topic guide codeveloped with our RAG. Reflexive thematic analysis was conducted using a critical-realist approach [43,44]. The full methodology and findings for this phase have been published elsewhere [29] and laid the foundations for co-designing the online peer support platform. In brief, online focus groups followed a semistructured interview schedule and were transcribed verbatim for analysis. Early stages of the focus group explored their psychosocial needs and how these were linked to their psychological well-being before exploring how these identified needs could be addressed. Participants were asked if, why, and how an online community that connects people living with a variety of long-term physical health conditions would be something that they would use to support the unmet psychosocial needs they identified in the earlier stages. Transcripts were analyzed using reflexive thematic analysis,

XSL•FO
RenderX

adopting a critical-realist approach. Following the analysis, the 10 participants who took part in these interviews were invited to be lived experience experts as part of the participatory design panel alongside other stakeholders, for the co-design stage (described below).

Co-Design

The co-design stage involved the participatory design panel, which was made up of an external software development agency (TOAD), researchers and clinicians from King's College London (KCL), our coinvestigator with lived experience, and lived experience experts from the focus groups (n=10) in the co-assess stage [29]. The lived experience experts were originally recruited from King's College London's Integrated Care Consultation Partners Group, the Guy's and St Thomas' PPI group, and the King's College Hospital PPI group. Many were living with multiple conditions, and a range of musculoskeletal, neurological, cardiovascular, respiratory, gastrointestinal, gynecological, autoimmune, and pain-related disorders were represented by the lived experience experts of our participatory design panel. The majority (8/10, 80%) of the lived experience experts used technology daily; however, only 30% (3/10) had used internet support groups before. No formal assessment of depressive symptoms was administered to participants, but recruitment materials emphasized the focus on preventing depression in the context of chronic illness and invited people who felt that they had relevant lived experience of low mood and psychological distress (current or past, with or without a formal clinical diagnosis) to participate. This inclusive, open approach was considered critical in this coproduction process, particularly with regard to identifying a breadth of different perspectives and experiences.

Six online workshops were conducted over 6 weeks, each lasting approximately 2 hours. The KCL team and software developers guided the workshops following specific agendas to maximize the discussions' applicability for creating the prototype (see Table 1). The panel was invited to share any additional suggestions after the sessions via email. The software developers incorporated feedback from each workshop to create the online peer support platform prototype.

Table 1. Agenda of the 6 development workshops with the participatory design panel to co-design the platform prototype.

Workshop	Agenda
1: How will the platform help people?	This workshop adopted a "users first" approach. This involved exploring questions such as "What relationships in your life are important to you, and how do you navigate them?" and "Have you looked for information on living with a long-term health condition?"
2: What are users' needs and wants for the platform?	This workshop focused on assessing user needs in a features-driven approach. This included brainstorming the features and functions that users would like the platform to have. The associated features or factors that may enhance the benefit of and/or reduce barriers of using the platform were also explored.
3: Platform features and solutions	Establishing the core features and how these should be prioritized in the development of the prototype. This involved asking people to rank their preference for different types of social interaction on a platform, from posting content anonymously to commenting in response to other content, and whom they prefer to socially interact with (eg, people with other long-term conditions, whoever is online at that time, and people in close geographical proximity).
4: Psychoeducation and moderation	Identify the type of educational content that the platform should host. Also identify key moderation principles and rules for communication on the platform by asking "What is the platform?" versus "What the platform isn't?"
5: What is the creative concept of the platform?	Three different creative concepts (eg, color, schemes, and layouts) were presented. Aimed to discuss what will work best and why, to allow further development of the platform features and creative brand.
6: Finalizing the creative bread and prototype	Finalizing the creative brand and prototype, including refining the brand position and name, alongside the visual look and feel.

Co-Validate

Overview of Co-Validation Stage

In the co-validation stage, we sought to review and finalize the platform prototype by answering the following questions: (1) "Have we correctly translated what people wanted into the platform itself?" (2) "Does everything work as we expect it to?" (3) "What, if any, improvements are needed?". This was an iterative process that involved the following steps (see Figure 3): (1) usability testing: think-aloud task and semistructured interviews, (2) iterative platform development, and (3) further prototype testing, final development work, and final prototype approval.

XSL•FO
RenderX

Figure 3. The steps involved in the co-validation stage of the CommonGround project. KCL: King's College London; LTC: long-term condition; RAG: research advisory group; UT: usability testing.

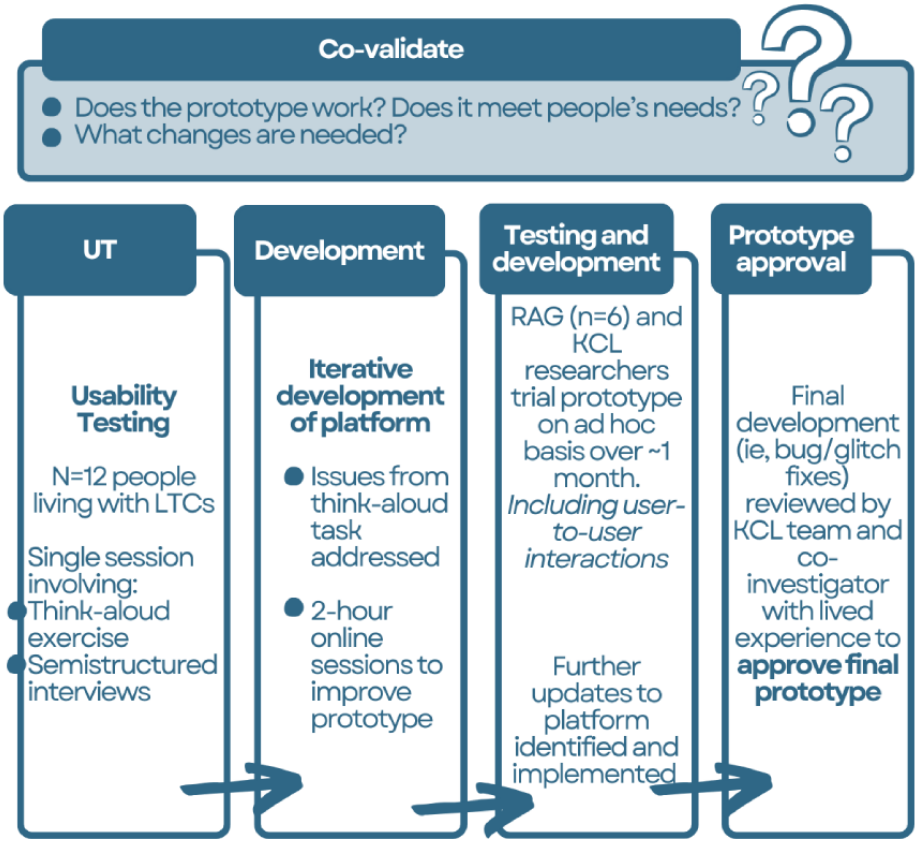

Usability Testing: Think-Aloud Task and Semistructured Interviews

Overview

The usability testing involved participants with different long-term health conditions who had no prior involvement in the project. Participants were recruited through national charities and were reimbursed with a £20 (US $26.6) voucher; ethical approval received from King's College Research Ethics Committee: HR-19/20-14938). The usability testing sessions were conducted via Zoom during the COVID-19 pandemic, as the planned in-person sessions were not possible. Each session included the participant, a research assistant, and a member of the software team. Participants screenshared their view of the platform throughout the session, which involved (1) a think-aloud session (~30 minutes [34]) and (2) a semistructured interview (30-60 minutes).

Think-Aloud Session

Participants were asked to verbalize their thoughts as they explored the platform prototype and worked through a series of tasks that a typical user may complete (see Multimedia Appendix 2 for think-aloud activity tasks). For example, they had to create an account, voicing everything that they were thinking as they read through the pages and completed each step. If participants encountered issues in completing a task, the research assistant and software developer encouraged them to troubleshoot themselves, rather than providing explicit guidance on how to overcome the issues.

Semistructured Interviews

Participants were then interviewed to explore views on the design and usability of the platform prototype, and potential facilitators and barriers to accessing or using peer support platforms in general. The topic guide (see Multimedia Appendix 3) included questions like "How did you find navigating the platform and finding what you were looking for?" and "Was there anything you looked for on the platform that you could not find?" Interviews were conducted until data saturation was reached: when no new themes or insights were emerging during data collection [45].

All usability testing sessions were audio-recorded and transcribed verbatim. The anonymized transcripts were analyzed in NVivo (version 14, released 2023; Lumivero) using content analysis [46] to organize participants' views and experiences into 5 predetermined themes and draw realistic conclusions from the data that could be implemented through further co-design workshops to finalize the prototype. Content analysis involved an iterative process of systematically organizing and condensing the data into a coding framework involving the overarching themes of (1) features and functions, (2) usability, (3) branding, (4) barriers and facilitators, and (5) ethos and values. These overarching themes were selected through discussions with our RAG for their ability to organize the data in accordance with the overall objective of coproducing an online peer support platform. Within the themes, codes were developed inductively in response to emerging patterns on a manifest level (ie, taking the meaning of the text at what the text says). Three coders (HGJ, GL, and VL) independently coded 3 transcripts before meeting to ensure consistency in

XSL•FO
RenderX

interpretations and make final updates to the coding framework. Once the final coding framework was confirmed, the coders analyzed the remaining data. Bugs and glitches were identified (and coded under a separate code) by extracting statements where participants expressed difficulties or problems when completing each think-aloud task.

Iterative Platform Development

First, the software developers from TOAD resolved the glitches and bugs identified in the think-aloud task analysis. The software code was then transferred to the software developers at Bitjam for the remainder of the coproduction. Through a series of 7 two-hour online discussion sessions, the KCL team, our coinvestigator with lived experience, and the RAG agreed on the developments required to improve the existing prototype, and designed the new features and functions that could solve issues or fulfill unmet needs identified in the first stages of usability testing. Based on the RAG discussions, the KCL team prepared user stories and visual mock-ups for all proposed developments to ensure the vision of the RAG was implemented correctly by Bitjam. In these discussions, safety and moderation policies, alongside the related functional features on the platform, were also finalized.

Further Prototype Testing, Final Development Work, and Prototype Approval

Following the iterative development of the platform, a review of the refined prototype was completed by the RAG (who had not accessed the "live" platform during development sessions) and KCL researchers (who had no other involvement in this project). This testing occurred over a 1-month period, where testers were encouraged to interact with the prototype as they would with existing social media platforms or forums, logging in on multiple occasions, exploring all the available features, and interacting with other testers. This usability testing aimed to review the new features and functions and identify any bugs or glitches that occur when multiple users are simultaneously accessing the platform and interacting with each other. All testers shared written feedback about bugs and glitches from which the required edits were extracted. The final edits were then implemented by Bitjam. The final prototype then underwent extensive testing by our research assistant and coinvestigator with lived experience before final platform approval was granted by the KCL team.

Co-Deliver

Following the successful codevelopment of the peer support platform, the next stages involved coproducing a protocol for a feasibility and acceptability randomized controlled trial. The co-deliver stage involved online group discussions with our RAG and coinvestigator with lived experience to guide study design, the review of study documents, questionnaire testing, and cocreating our adverse events protocol. The Samaritans and SCOPE charity organizations were also consulted with regard to moderation and safeguarding policies, and in the development of moderator training. The protocol of the feasibility and acceptability randomized controlled trial is available elsewhere [47].

Results

Co-Assess

The full findings of this work that laid the foundations for co-designing our online platform have been published elsewhere [29]. In brief, findings from 3 online focus groups identified 3 key themes regarding how an online peer support platform could address the unmet psychosocial needs of people living with long-term physical health conditions: (1) relationship between self and the outside world, (2) past experiences of peer support, and (3) philosophy and vision of peer support. Although participants spoke of their mind and body being entwined, they felt this was often neglected in their health care, with their mental and physical health treated as separate entities. The need for their mental health to be supported in the context of their physical health was clear to them. Participants felt that peer support offered both an opportunity to share knowledge and a sense of mutual validation, particularly as people with similar experiences are best placed to understand their circumstances. However, it was clear that online peer support platforms needed to be safe and credible spaces and that to achieve this, platforms must be co-designed with the people who will ultimately use them. Participants also explored the potential benefits and concerns of using online communities, which helped to shape the ethos and philosophy of our platform. Online communities were valued for being easily accessible and easy to participate in; however, it was considered key that a "code of conduct" should be shared and enforced for the space to feel safe and credible. Although many participants were accustomed to the transitions of face-to-face activities to online settings, there were concerns related to how digital interactions may be impaired if equipment or technology does not operate properly. Together, these findings provided the basis for the following co-design stage.

Co-Design

A prototype of the peer support and psychoeducation platform was successfully developed in the co-design phase with the following features and functions:

- A landing page that introduced the platform
- The community principles that govern the community, presented as "terms of use" that people must agree to before registration
- A registration page, where individuals could create their account and set up their profile
- The community feed which hosted a forum page where users could create posts, add comments, and react to posts
- A resource page where users could view, download, react to, and comment on resources relevant to people living with long-term physical health conditions
- The ability for users to save community feed posts and resources to their private "My Garden" page

The web-based prototype was accessible on desktop and mobile devices. During the co-design phase, the platform name "CommonGround" and the "all diagnoses welcome" ethos were developed. This ethos and related manifesto created by our coinvestigator with lived experience (see Multimedia Appendix 4) reflects the view that people living with any long-term

XSL•FO
RenderX

physical health condition have shared lived experiences that they can relate to, from adapting their work or social lives to taking medication and daily self-management. The slogan "share, support, learn, cultivate" was also developed to reflect the main purposes of CommonGround. The term "cultivate" was later replaced with "grow" as our RAG felt this was a more familiar term and better reflected the CommonGround ethos of supporting someone to grow on their self-management journey.

Co-Validate

Usability Testing: Think-Aloud Task and Semistructured Interviews

Overview of Findings

The first stage of usability testing included 12 people living with at least one long-term physical health condition (see Table 2) who had not been involved in the co-assess or co-design stages. The range of diagnoses included asthma, chronic kidney disease, diabetes, epilepsy, fibromyalgia, hypertension, multiple sclerosis, myalgic encephalomyelitis, and rheumatoid arthritis. Participants were aged 25-71 years (median 53.5, IQR 29 years), had access to their own internet-connected device, had internet at home, and used the internet daily. Five (42%) participants had used internet support groups before.

Table 2. Characteristics of participants involved in usability testing through think-aloud exercise and interviews (N=12).

Characteristics	Values, n (%)
Gender identity	
Woman, including transgender woman	11 (92)
Man, including transgender man	1 (8)
Number of participants living with multiple long-term conditions	
1 long-term condition	5 (42)
2 long-term conditions	4 (33)
3+ long-term conditions	4 (33)
Number of participants living with comorbid mental health conditions	3 (25)
Do you have access to the internet at home?	
Yes	12 (100)
No	0 (0)
How frequently do you use the internet?	
Daily	12 (100)
Weekly	0 (0)
Monthly	0 (0)
Devices used to access the internet? (more than one device could be selected)	
Laptop	7 (58)
Desktop	1 (8)
Phone	7 (58)
Tablet device	3 (25)
Have you used internet support groups before?	
Yes	5 (42)
No	7 (58)

From the think-aloud task, explicit verbal indications of bugs and glitches were successfully identified and consolidated (see Table 3 for full results). In total, 9 usability issues were identified. The most common usability issues were related to navigation, layout, understanding instructions, and the multiple steps required to complete a task. The most usability issues were encountered for the "create an account" task. Across ages, gender, and long-term physical health conditions, experiences of issues were similar, with the majority of issues experienced by most participants.

XSL•FO
RenderX

Table 3. Think-aloud tasks and usability testing results for each task.

Usability issues	Description of issues	Example Quotes
Task 1: browse around for 1-2 minutes, providing your feedback [logged out state]		
Navigation	Users had difficulty moving around the home page. Some experienced a sticky scroll. The arrow at the bottom of the page that intended to move the user down did not function properly.	• "So, basically, it doesn't scroll. It doesn't scroll so easy. I think when, when I put in my cursor on the blank space on the side, then it tends to scroll a bit. So, it's a bit erratic, the way it scrolls." [Participant 1] • "Once you click to go back up it's, kind of, doing that thing that if you try and go to go down with the arrow it, kind of, keeps doing that jump thing." [Participant 8]
Layout	Users could not read all the text as some was obscured or not displayed properly.	• "See, I can't read, I can't read that because the pages kind of separates, and you're kind of missing a little bit of something there." [Participant 1] • "The first thing I've noticed, I can't read the top line there." [Participant 6]
Task 2: create an account		
Navigation	Users had difficulty in identifying the button to create an account, often selecting the wrong button.	• "Right, so where, where do I go to join an account, to make an account?" [Participant 2] • "Where do I create an account? Or do I have to login? Right. So, do...? Am I using? Do I put me email address in and stuff?" [Participant 3]
Understanding instructions	Users were confused about what to input for various fields of the registration form and desired further instructions.	• "I think what's really interesting there is it wasn't really clear for you where to start. It's the question's so broad. So, we can look at, you know, maybe having a couple of kind of tips underneath or something." [Participant 1]
Glitchy function	Users struggled with the calendar date selector when inputting date of birth.	• "That was a bit tricky writing down a date of birth... but I did about...I think it was the third time I typed it in. Like I said, here I think a drop-down menu would be easier." [Participant 7] • "Mm, don't even know how to put the date in." [Participant 11]
Workflow	Users often missed a step in the registration form, having to review closely to resolve the issue and proceed.	• "What's that? I've got a hand, so that's saying that's a live link, but it's not allowing me to (click)...Oh, okay. So I've gotta put something else then?" [Participant 7]
Task 3: browse around for 1-2 minutes, providing your feedback [logged in – community feed]		
Visuals	Users were unsure what some of the graphics and labels of buttons meant, and it created confusion about the potential functions of different buttons.	• "I'm not sure what the Add to Garden means. I might be a bit dumb [laughing] if I don't know, but I'm not sure, um, about that, if I'm honest." [Participant 12] • "The, you know, the things you have, the images that you have above the post a thought, share knowledge, ask a question, I was just wondering what that is, what those images depict." [Participant 1]
Task 4: navigating between pages for tasks: (1) find a psychoeducational resource and view it and (2) view posts from other community members		
Navigation	Users often did not notice the navigation bar as it did not stand out on the page. This limited their ability to switch between pages, including to find a resource or to go view their "My Garden."	• "Okay. Okay, so it took me a while to see that thing at the top where I've got the options for, Community Resources In "My Garden". Um, I didn't see that until I'd scrolled all the way down to the bottom." [Participant 11] • "I think it's because I completely didn't realise that those were up there. If I maybe did. Maybe if they were bigger. I'm, I'm trying to...I don't know what other things you could do to, um, make it more noticeable." [Participant 12]
Task 5: create a post		

XSL•FO
RenderX

Usability issues	Description of issues	Example Quotes
Workflow	Some users had trouble in completing the different fields for creating a post.	• "Mmm. I don't know. Is there a way to, sorry, I don't know how to work this. Is there a way to like, type to here, or..." [Participant 4]
Task 6: view posts from other community members		
Navigation	Some users did not know how to comment on a post.	• "I'd be looking for a, um, respond button to be able to get back to [username], you know?" [Participant 10]

The content analysis of the usability testing transcripts grouped the data into the 5 themes predetermined from the agendas of the co-design workshops. The first 4 themes—features and functions, usability, branding, and barriers and facilitators—are presented in full in Table S1 in Multimedia Appendix 5. Table S1 in Multimedia Appendix 5 also describes the platform modifications and subsequent development work that were considered and/or implemented to address participants' feedback. The fifth theme—ethos and values—is depicted in Table S2 in Multimedia Appendix 5. Within each theme, codes and subcodes were also added as appropriate during analysis (see Tables S1 and S2 in Multimedia Appendix 5).

Theme 1: Features and Functions

This theme included the codes of additional features, content and materials, community page, "My Garden," and joining form. In addition to the bugs and glitches participants experienced when creating an account, upon reflection, some expressed apprehension about signing up and agreeing to the community principles (terms of use). For instance,

> *It's almost like these are all our rules and regulations, and you have to agree to all of these rules and regulations before we will even tell you anything. That was, that put me on my guard, really.* [Participant 10]

Participants expressed the need to fully understand "what am I agreeing to," suggesting that background information about CommonGround and who owns it is needed to help people decide whether to join. However, participants shared that the personal information required matched their expectations of what creating an account typically involves. Throughout the feedback, it was clear that making the account creation process as easy and as informative as possible was important for facilitating registration on CommonGround.

> *To be honest, it's, it's really straight forward and it kind of makes it easier because it's, it goes along with what the majority of the registered, sort of, websites go for.* [Participant 8]

Participants generally had positive views of the community, resources, and "My Garden" pages. Specifically, people appreciated the unique features of the community forum, stating that the different types of posts you could create were "something I've not seen ever before" [Participant 2], and that the "reactions are great... the 'I hear you' and this (thinking of you) is helpful, it's quite nice to see that" [Participant 8]. Some participants were unsure what the "My Garden" page was for, but were keen to use it once they understood that they could save posts of interest to their own private "garden": "that really appeals to me, because I can, I can, yeah, grow my own library, oh, that's heaven [laughs]" [Participant 12]. Participants suggested that additional text should be added to the "My Garden" page, and the "save post" button on forum posts should be changed to "add to My Garden," to clearly demonstrate how the page can be used.

Participants also shared the view that new additional features could improve the platform. For instance, participants expressed a clear desire for a search bar to fulfill their need of easily finding specific content of interest and overcoming issues in navigating the platform: "I would seriously just give me a search bar and let me write in what I want to look for" [Participant 10]. Participants also suggested adding pages that share "about us" information and the full moderation policy.

Theme 2: Usability

This theme included the codes of platform navigation, visual design, visual layout, and tone and language. Overall, participants indicated that the platform "is quite user-friendly" [Participant 3], and that they felt comfortable using the platform, even if they did not feel technologically adept themselves:

> *I wouldn't say that I'm fantastic in technology. So, it was...It's, it's...I think it's quite technically friendly, and quite easy to access, to be honest.* [Participant 9]

Participants stressed the importance of the platform being intuitive to navigate with a typical website layout. For instance, participants expected there to be a menu bar as they have "got so used to seeing a menu thing at the top that when I go on to any website, that's the first thing I look for, is that menu" [Participant 10]. Although the prototype had a menu bar, participants "didn't initially notice that they (the menu bar and page tabs) were there at all" [Participant 12], until asked to complete a task involving switching to a different page that forced them to search around the page for it.

Participants appreciated the simple, clear, uniform layout of the community feed, with posts that are "a nice size to be able to digest" [Participant 11]. While many participants liked the colors and graphics of the platform, some of the graphics used to indicate the functionality of buttons were not sufficiently clear: "the images that you have above the post a thought, share knowledge, ask a question, I was just wondering what that is, what those images depict" [Participant 1]. Participants also wanted some clarity on the language used within the site copy, particularly for terms relating to peer support and mental health that may be unfamiliar:

> *Peer to peer...I think at times, it could be slightly wordy...It's a term that I've come across, yes*

XSL•FO
RenderX

> *definitely. I'm not sure it's term that everybody will know, though.* [Participant 1]

Participants also shared that extra explanations would be helpful on the resources page, when introducing the resources overall ("Um, you say end, endorsed by experts...But who, who exactly?" [Participant 2]) and on specific topics such as "keeping active" ("I'm a little bit confused on when you say active, is that more like exercising keeping active sort of thing or...?" [Participant 8]).

Theme 3: Branding

This theme related to how participants felt an online platform should be presented to the world, including the codes of the platform name, platform branding, and advertising. Participants felt that the name "CommonGround" clearly communicated a safe space for people who have the shared experience of living with long-term health conditions to connect. However, participants were unsure whether the name was able to communicate the fact that the platform has a focus on supporting mental well-being:

> *I do think that's a really good name, but, um, I don't know whether it says enough...It depends how people are introduced to CommonGround...If it's something you find, you know, if you Google, if you Googled psychological services, and CommonGround came up, then I don't think it would really matter.* [Participant 8]

These views were also expressed in relation to the logo:

> *I know CommonGround is probably the main logo. Um, reading CommonGround for the first time you wouldn't exactly think it's, like, health-related or anything.* [Participant 6]

Having a brand, including the name and logo, that communicates the purpose of CommonGround was suggested to facilitate engagement with the platform.

The importance of being able to trust the online platforms was emphasized. Participants wanted to know who owns the platform and if their motivation is to genuinely improve the well-being of users or to generate profit. In the case of CommonGround, the KCL logo on the landing page helped develop trust in our motivations and governance: "you're under King's College so that gives you that credibility anyway" [Participant 11]. Participants suggested that advertising the platform through avenues that people with long-term conditions trust, including their health care teams or charities, would also foster trustworthiness in CommonGround:

> *I think it carry, it would carry more weight. You know, if a, if a professional has recommended something.* [Participant 2]

Theme 4: Barriers and Facilitators

During their reflections of exploring CommonGround, participants identified some barriers and facilitators that might affect someone's interest in, and engagement with, a peer support platform. For instance, someone's desire to use the platform might be limited by their perceptions of peer support, and whether peer support is suitable for them or how to engage with it appropriately: "(being) on this platform practically every day. That's not healthy" [Participant 7]. For those who were motivated to use CommonGround, participants believed that a user's technological competency would not be a barrier. Participants who self-identified as "not very tech-savvy" felt "like it's very open to a lot of people" [Participant 8] and "it's quite technically friendly, and quite easy to access" [Participant 9]. However, there were concerns related to the safety of the online community, which may be vulnerable to the sharing of misinformation or malicious links, and infiltration by rogue users:

> *It's just you know how there's like a lot of spam, a lot of scams, especially when it comes to links sort of thing. And you never know what you're clicking on. So, somebody who's got good intention and coming on the website, you don't know if that's going to be everybody's intention.* [Participant 8]

Theme 5: Ethos and Values

The ethos and values theme had codes including platform purpose, the meaning of coproduction, and the power of peer support (see Table S2 in Multimedia Appendix 5). The fact that CommonGround was rooted in coproduction was considered essential, with participants emphasizing the importance of combining lived experience alongside academic experts who bring scientific rigour:

> *Co-design is essential if you're doing anything for a, a particular audience. You, you, you must include representatives of that audience if you want to get a, a, a fully rounded, um, end product.* [Participant 7]

That CommonGround would fulfill a current gap in support for people with long-term conditions was also noted, with participants sharing that CommonGround is "much needed" [Participant 1], particularly for its dual focus on the mind and body: "I mean, as well as your physical health conditions, I think it would help tremendously with my mental health" [Participant 6].The power of peer support was recognized through participants' previous experiences, where the ability to connect anonymously with people "like me" is powerful, as sometimes there was hesitancy to speak with people in their existing social circles:

> *A lot of people I, I, I would imagine would, would talk to somebody on the bus or whatever, before they'd talk to family, or friends even...The good thing about that is nobody knows who I am.* [Participant 5]

Although sharing experiences can help alleviate feelings of "why me," some concerns were shared of becoming overfocused on the effects of their long-term conditions during these exchanges, when they might rather "just get up and get on with my life" [Participant 7]. Participants recognized that a peer support platform can be beneficial at times, but also something to take some space from at other times dependent on their personal circumstances and well-being needs.

Iterative Platform Development

Table S1 in Multimedia Appendix 5 outlines the platform development opportunities identified through content analysis

XSL•FO
RenderX

of the think-aloud exercise and interviews that guided the iterative development process. The software developers were successful in completing the development work agreed by the RAG and KCL team. For instance, in response to the need for improved navigation, the RAG designed a search function. This included adding a search button within the navigation menu and allowing users to search by free text or content tags (similar to hashtags), with the option to filter results by recency or relevance.

Throughout the iterative process, the RAG also created additional features to enhance user experience. One key addition was the "How I Am Feeling Now" feature. The RAG recognized the importance of being able to communicate one's emotional context in forum posts to help their peers understand the type of support they are looking for. The RAG therefore suggested that when users are creating their post, there should be an option to select an emoji that represents their current emotional state or mood which then appears alongside the post's text to give this emotional context.

Further Prototype Testing, Final Development Work, and Final Prototype Approval

The testing by the RAG and KCL volunteers identified minor bugs, slow loading times, and that notifications were not being sent or received properly. Other feedback included suggestions to reorder or reformat text, for instance, by making key information more prominent or highlighted to catch the attention of end users who may skim through the content. The final required edits were successfully implemented, with the final prototype for the feasibility randomized controlled trial completed in May 2024. Screenshots of the final platform prototype are provided in Multimedia Appendix 6.

Discussion

Principal Findings

There is a current lack of psychological support to help people with long-term physical health conditions manage their mental well-being that is easily accessible and tailored to their specific psychosocial needs, particularly for those at risk of developing major depression [48,49]. In response to this gap in psychological support, we sought to coproduce an online peer support platform following a 4-stage, iterative coproduction cycle. While coproduction has become the gold standard for intervention development [28], to date, end users have rarely been involved in all aspects of intervention design, development, and evaluation [27]. Our work serves as an exemplar of how to coproduce an online peer support and psychoeducation platform by creating partnerships between people with long-term physical health conditions, academics, clinicians, and software experts. In the co-assess stage, people living with long-term physical health conditions expressed that an online intervention that provides a safe and trustworthy space for connecting with others like me, while acknowledging the mind-body interplay, could support their psychosocial well-being [29]. Following the co-design phase, which resulted in the successful creation of an initial prototype, usability testing during the co-validation stage confirmed that CommonGround included features and functions that were appropriate and acceptable. Critically, areas for improvement were also identified, such as making platform navigation more intuitive, resolving platform glitches, and building new features to enhance user experiences. Ultimately, through the iterative coproduction cycle that consistently places the voices of lived experience experts at the foreground, we created an intervention that has been specifically tailored to our target audiences' unique characteristics and needs, from the specific features and functions to the "all diagnoses welcome" ethos and platform governance. The CommonGround platform is now viable and ready to be evaluated in a future feasibility randomized controlled trial.

As noted in the development of other mobile health applications (eg, [50]), we found that iterative usability testing was critical for resolving lingering bugs and glitches to minimize the impact of these problems on user engagement [51]. Notably, many of the usability issues our participants experienced are commonly encountered during usability testing, such as navigational barriers and errors in site copy [52]. Our participants' emphasis on the importance of CommonGround being easy to understand and intuitive to navigate offers further evidence of the critical role that language and usability play in facilitating engagement with peer support platforms [16]. This finding also aligns with the technology acceptance model, which states that whether a user will accept or reject an online platform is partially dependent on how easy it is to use [53]. Together, these insights highlight the importance of incorporating structured, iterative usability testing throughout the coproduction of novel online interventions to ensure that they are technically functional, accessible, and aligned with the anticipated end user's needs.

An individual's digital literacy is often cited as a potential barrier to engaging with online interventions [19], yet interestingly, our participants did not identify technological competency as a potential barrier to accessing CommonGround. As participants described the platform as relatively intuitive to navigate, this could suggest that CommonGround is easy to use and does not require a significant level of digital literacy. However, it is important to note that although some participants identified as "not very tech-savvy," they were all daily internet users and may not represent the views of those with limited internet access or digital experience. This highlights the importance of actively recruiting people who may not typically engage with online interventions to comprehensively explore potential barriers to using online peer support forums [16].

Developing an online intervention, however, goes beyond creating the features and functions that may improve end users' experiences. For instance, the importance of CommonGround being a safe space, and underpinned by consistent, reputable, and trustworthy branding, was emphasized throughout the co-design and co-validation stages. This echoes prior work highlighting how the inclusion of reputable branding, logos, and declarations of ownership can positively influence the trustworthiness and credibility of web-based health information [53,54]. Another facet of safety that our participants highlighted was technological security. Concerns about security and data protection are commonly cited as barriers to engaging with online interventions [55], with older users, those from high-income backgrounds, and those who had a previous

XSL•FO
RenderX

experience of a data breach expressing the most apprehension [56]. This highlights the importance of collaborating with software experts to build a robust, protected online interface with security features to minimize potential threats, assure data confidentiality, and maintain privacy. However, given that concerns about security vary between users, it is pertinent to develop security features alongside lived experience experts to strike a balance between allaying fears (a facilitator to engagement) and avoiding excessive measures (a barrier to engagement) [57,58].

The "power of peer support" emanated as a central theme throughout our coproduction process, reinforcing previous findings among people with long-term conditions that online peer support provides a valued space to strengthen social ties [21] and can improve an individual's emotional self-management and self-efficacy [16]. Throughout this work, our PPI and research participants echoed these findings, emphasizing the unique value of peers in providing valuable information and support, independent of their existing social ties [16,59,60] and reducing loneliness by providing connections with "someone like me" [21,49]. However, while the power of peer support was well recognized among our participants, some expressed awareness of potential negative events that can occur in online peer support and warned that not everyone may have such favorable opinions or understand exactly what peer support is. It has been recognized that the term "peer support" has been inconsistently defined in research [16] and is often not colloquially used in patient communities. There is often a lack of clarity or misperceptions of what peer support involves that might serve as a barrier to engagement [61]. Indeed, evidence suggests that some people are deterred from online peer support as they believe that they will not fit in, peer support will be unable to support their well-being, their posts will remain unanswered, or that they will be exposed to incorrect or misleading information [16,59,62]. Given that perceived usefulness of an app can predict engagement, considering the end users' opinion and understanding of peer support may be an important barrier or facilitator to engagement [63]. It will therefore be important to provide clear information to users on what peer support is and how potential adverse effects will be mitigated on CommonGround. Future work should also evaluate participants' views of peer support before and after engaging with online peer support to develop an understanding of the relationship between perceptions of peer support and engagement with interventions.

Limitations

The design of the CommonGround platform and the findings from the formal usability testing (think-aloud exercises, interviews) should be considered within the context of our participant demographics. The majority identified as women, were daily internet users, and had access to the internet at home. Our sample did however have varying levels of prior experience engaging with online forums (30% [3/10] of participatory design panel lived experience experts and 42% [5/12] of usability testing participants had no prior experience of using online forums), representing a range of different baseline expectations of what online forums should be like. Nonetheless, as frequent internet users, they may perceive and prioritize issues differently than those who use the internet less often, particularly as they have greater experience to compare the prototype to. Although 96% of UK households have access to the internet [64], internet usage is lower among those in older adults (eg, 99.5% of those aged 25-34 years versus 85.5% aged 65-74 years using the internet in the last 3 months [65]). As living with (multiple) long-term conditions is more common among older people, who therefore represent a significant proportion of those that could benefit from CommonGround, future work should prioritize exploring the perspectives of this group and investigating the potential role of digital exclusion. By recruiting a larger, more diverse sample as part of a future feasibility randomized controlled trial, we will capture a broader range of experiences, usability concerns, and insights into barriers, facilitators, and intervention improvements. We will purposively interview participants based on various factors, including age and engagement with CommonGround, with topic guides including focus on exploring barriers, such as digital exclusion, in greater depth. Findings from this future work will ensure that CommonGround can be developed further to align with any specific needs identified, including those of older adults and those who may experience digital exclusion.

Our participatory design panel and subsequent RAG also represent a group of patients who are actively engaged in managing their physical-mental well-being, have an interest in peer support, and are frequently online. As such, their perspectives and associated intervention development are potentially rooted in positive experiences and opinions of peer support. Further exploration of the views of individuals who are new to peer support and/or less active in their self-management would provide broader insights into their psychosocial needs and expectations of how a peer support platform should function.

It should also be noted that no formal assessment of depressive symptoms was conducted among participants or RAG members; however, all involved self-identified as having lived experience relevant to preventing depression among people with long-term physical health conditions. While this broad, inclusive approach enabled a range of different perspectives to be incorporated throughout the coproduction process, we cannot conclude that our findings are specifically applicable to those living with subthreshold depression. Future work could explore capturing formal measures of mental health and well-being but should be mindful of excluding the voices of lived experience experts solely based on score cutoffs. Our future feasibility and acceptability trial will offer opportunity to explore the perceived acceptability, feasibility, and appropriateness of the CommonGround platform exclusively among a sample of those living with subthreshold depression.

Due to the COVID-19 pandemic, usability testing and subsequent RAG meetings for iterative development were conducted online via videoconferencing software. For participants unfamiliar with Microsoft Teams, this unfamiliarity may have contributed to nervousness or influenced how they interacted with the platform prototype during the think-aloud exercise. Resource restrictions also meant that usability sessions were conducted by members of the research team involved in prototype development. In future usability testing, involving an

XSL•FO
RenderX

independent facilitator could enhance objectivity and encourage more candid feedback. However, to mitigate any bias as best as possible during the think-aloud exercises, the research team avoided prompting or guiding participants, allowing them to navigate and troubleshoot issues independently.

Finally, our findings may not be applicable beyond the cultural context of the United Kingdom. The willingness of people to discuss their mental and physical health with other people, particularly those who are not included in their existing social circles, likely varies cross-culturally due to different norms around expressing emotions and the role of family/community, and stigma.

Conclusions

For interventions to be successfully integrated into the self-management routines of people with long-term physical health conditions, it is critical that they are developed in collaboration with individuals who have lived experience. By dedicating significant time and resources throughout the coproduction cycle, we have successfully created an online peer support platform with embedded psychoeducation that is designed to support the mental well-being of people living with long-term physical health conditions. Through an iterative process of usability testing and prototype development, we refined the platform to ensure that it aligns with the needs and preferences of the target audience.

The CommonGround platform is now ready to be further evaluated with continued input from our RAG during the delivery of a trial to explore the feasibility and acceptability of CommonGround. The findings from the next stages will be critical in informing the preparation of a future large-scale efficacy trial. If the intervention is found to be feasible, acceptable, and clinically effective at preventing the progression of subthreshold depressive symptoms to comorbid major depressive disorder, CommonGround may offer a low-cost, scalable solution to bridge the current gap in psychological support for people with long-term physical health conditions who are at risk of developing comorbid major depressive disorder.

Acknowledgments

We would like to acknowledge our software developers at TOAD and Bitjam who were instrumental in helping to co-design the CommonGround platform. We also extend our thanks to Hannah Rowlands for their support and time dedicated to the co-assess stage. We would also like to acknowledge the time and contributions of all of the participants who made this work possible. And finally, thank you to all the members of our Public Patient Involvement groups and Research Advisory Group who contributed at various stages throughout the development of this protocol. CommonGround is a coproduced intervention, and this research would not be possible without their dedication, support, and insightful contributions.

Data Availability

Data are available upon request to the corresponding author.

Funding

This project is being funded by King's Health Partners and Guy's and St Thomas Charity from the Multiple Long Term Conditions Challenge Fund [EIC180702 (to MH)]. The funders had no role in study design, data collection and analysis, decision to publish, or manuscript preparation. The research is part-funded by the National Institute for Health and Care Research (NIHR) Maudsley Biomedical Research Centre (BRC). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Authors' Contributions

All authors listed meet the International Committee of Medical Journal Editors criteria for authorship. MH, AS, VL, and GL conceived the study. All authors co-designed the study. HGJ led data analysis alongside supervision from VL and GL. HGJ and GL wrote the manuscript. All authors reviewed and approved the final manuscript.

Conflicts of Interest

EA-F declares voluntary and unpaid advisory to Someone Who Knows, an early-stage technical initiative seeking to create an online platform that matches people with diagnosed long-term health conditions with peers for informal support and virtual conversations. Someone Who Knows is registered as a business in the Netherlands. EA-F has never received payment, shares, or any other form of exchange for this advice. All other authors have no conflicts of interest to declare.

Multimedia Appendix 1
Consent forms.
[DOCX File , 67 KB - jmir_v28i1e79666_app1.docx]

Multimedia Appendix 2
Think-aloud activity tasks.

XSL•FO
RenderX

[DOCX File , 21 KB - jmir_v28i1e79666_app2.docx]

Multimedia Appendix 3
Topic guide for semistructured interviews.
[DOCX File , 25 KB - jmir_v28i1e79666_app3.docx]

Multimedia Appendix 4
The CommonGround platform manifesto written by our coinvestigator with lived experience.
[DOCX File , 21 KB - jmir_v28i1e79666_app4.docx]

Multimedia Appendix 5
The results tables of the content analysis of transcripts of the think-aloud task and semistructured interviews.
[DOCX File , 91 KB - jmir_v28i1e79666_app5.docx]

Multimedia Appendix 6
Screenshots of the final prototype of CommonGround.
[DOCX File , 1104 KB - jmir_v28i1e79666_app6.docx]

References

1. Long Term Conditions Compendium of Information. 3rd ed. UK Department of Health. 2012 May 30. URL: https://assets.publishing.service.gov.uk/media/5a7c638340f0b62aff6c154e/dh_134486.pdf [accessed 2026-03-14]
2. Egede LE. Major depression in individuals with chronic medical disorders: prevalence, correlates and association with health resource utilization, lost productivity and functional disability. Gen Hosp Psychiatry 2007;29(5):409-416. [doi: 10.1016/j.genhosppsych.2007.06.002] [Medline: 17888807]
3. Depression in adults with a chronic physical health problem: recognition and management. National Institute for Health and Care Excellence. 2009 Oct 29. URL: https://www.nice.org.uk/guidance/cg91 [accessed 2026-03-06]
4. Rayner L, Matcham F, Hutton J, Stringer C, Dobson J, Steer S, et al. Embedding integrated mental health assessment and management in general hospital settings: feasibility, acceptability and the prevalence of common mental disorder. Gen Hosp Psychiatry 2014;36(3):318-324 [FREE Full text] [doi: 10.1016/j.genhosppsych.2013.12.004] [Medline: 24630892]
5. Littlewood E, Chew-Graham C, Coleman E, Gascoyne S, Sloan C, Ali S. A psychological intervention by community pharmacies to prevent depression in adults with subthreshold depression and long-term conditions: the CHEMIST pilot RCT. Public Health Res 2022;10(5). [doi: 10.3310/ekze0617] [Medline: 35353469]
6. Rodríguez MR, Nuevo R, Chatterji S, Ayuso-Mateos JL. Definitions and factors associated with subthreshold depressive conditions: a systematic review. BMC Psychiatry 2012 Oct 30;12(1):1-7. [doi: 10.1186/1471-244x-12-181]
7. Hance M, Carney RM, Freedland KE, Skala J. Depression in patients with coronary heart disease: a 12-month follow-up. Gen Hosp Psychiatry 1996;18(1):61-65. [doi: 10.1016/0163-8343(95)00100-x] [Medline: 8666215]
8. Bot M, Pouwer F, Ormel J, Slaets JPJ, de Jonge P. Predictors of incident major depression in diabetic outpatients with subthreshold depression. Diabet Med 2010;27(11):1295-1301. [doi: 10.1111/j.1464-5491.2010.03119.x] [Medline: 20950389]
9. Bodenheimer T, Wagner EH, Grumbach K. Improving primary care for patients with chronic illness. JAMA 2002;288(14):1775-1779. [doi: 10.1001/jama.288.14.1775] [Medline: 12365965]
10. Miller WR, Lasiter S, Bartlett Ellis RB, Buelow JM. Chronic disease self-management: a hybrid concept analysis. Nurs Outlook 2015;63(2):154-161 [FREE Full text] [doi: 10.1016/j.outlook.2014.07.005] [Medline: 25241136]
11. Five year forward view. NHS England. 2014 Oct. URL: https://www.england.nhs.uk/wp-content/uploads/2014/10/5yfv-web.pdf [accessed 2026-03-06]
12. Taylor S, Pinnock H, Epiphanou E, Pearce G, Parke H, Schwappach A. A rapid synthesis of the evidence on interventions supporting self-management for people with long-term conditions: PRISMS (Practical Systematic Review of Self-Management Support) for long-term conditions. Health Serv Deliv Res 2014;2(53). [doi: 10.3310/hsdr02530] [Medline: 25642548]
13. Bodenheimer T, Wagner EH, Grumbach K. Improving primary care for patients with chronic illness: the chronic care model, Part 2. JAMA 2002;288(15):1909-1914. [doi: 10.1001/jama.288.15.1909] [Medline: 12377092]
14. Franek J. Self-management support interventions for persons with chronic disease: an evidence-based analysis. Ont Health Technol Assess Ser 2013;13(9):1-60 [FREE Full text] [Medline: 24194800]
15. Kim S, Park M, Song R. Effects of self-management programs on behavioral modification among individuals with chronic disease: a systematic review and meta-analysis of randomized trials. PLOS One 2021;16(7):e0254995 [FREE Full text] [doi: 10.1371/journal.pone.0254995] [Medline: 34297741]
16. Hossain SN, Jaglal SB, Shepherd J, Perrier L, Tomasone JR, Sweet SN, et al. Web-based peer support interventions for adults living with chronic conditions: scoping review. JMIR Rehabil Assist Technol 2021;8(2):e14321 [FREE Full text] [doi: 10.2196/14321] [Medline: 34032572]

XSL•FO
RenderX

17. Lauckner HM, Hutchinson SL. Peer support for people with chronic conditions in rural areas: a scoping review. Rural Remote Health 2016;16(1):3601 [FREE Full text] [Medline: 26943760]
18. Cooper RE, Saunders KRK, Greenburgh A, Shah P, Appleton R, Machin K, et al. The effectiveness, implementation, and experiences of peer support approaches for mental health: a systematic umbrella review. BMC Med 2024 Feb 29;22(1):72 [FREE Full text] [doi: 10.1186/s12916-024-03260-y] [Medline: 38418998]
19. Wright KB. Communication in health-related online social support groups/communities: a review of research on predictors of participation, applications of social support theory, and health outcomes. Rev Commun Res 2016;4:65-87. [doi: 10.12840/issn.2255-4165.2016.04.01.010] [Medline: 2255-4165]
20. Rafieifar M, Hanbidge AS, Lorenzini SB, Macgowan MJ. Comparative efficacy of online vs. face-to-face group interventions: a systematic review. Res Soc Work Pract 2024 Mar 18;35(5):524-542. [doi: 10.1177/10497315241236966]
21. Kingod N, Cleal B, Wahlberg A, Husted GR. Online peer-to-peer communities in the daily lives of people with chronic illness. Qual Health Res 2016 Dec 12;27(1):89-99. [doi: 10.1177/1049732316680203] [Medline: 27956659]
22. Johansson V, Islind AS, Lindroth T, Angenete E, Gellerstedt M. Online communities as a driver for patient empowerment: systematic review. J Med Internet Res 2021 Feb 09;23(2):e19910. [doi: 10.2196/19910] [Medline: 33560233]
23. Cooper K, Kirkpatrick P, Wilcock S. The effectiveness of peer support interventions for community-dwelling adults with chronic non-cancer pain: a systematic review. JBI Database System Rev Implement Rep 2014;12(5):319-348. [doi: 10.11124/jbisrir-2014-1590]
24. McCaughan E, Parahoo K, Hueter I, Northouse L, Bradbury I. Online support groups for women with breast cancer. Cochrane Database Syst Rev 2017;3(3):CD011652 [FREE Full text] [doi: 10.1002/14651858.CD011652.pub2] [Medline: 28278559]
25. Jennings H, Slade M, Bates P, Munday E, Toney R. Best practice framework for Patient and Public Involvement (PPI) in collaborative data analysis of qualitative mental health research: methodology development and refinement. BMC Psychiatry 2018 Jun 28;18(1):213 [FREE Full text] [doi: 10.1186/s12888-018-1794-8] [Medline: 29954373]
26. Hollis C, Sampson S, Simons L, Davies EB, Churchill R, Betton V, et al. Identifying research priorities for digital technology in mental health care: results of the James Lind Alliance Priority Setting Partnership. Lancet Psychiatry 2018 Oct;5(10):845-854 [FREE Full text] [doi: 10.1016/S2215-0366(18)30296-7] [Medline: 30170964]
27. Brotherdale R, Berry K, Branitsky A, Bucci S. Co-producing digital mental health interventions: a systematic review. Digit Health 2024;10:20552076241239172 [FREE Full text] [doi: 10.1177/20552076241239172] [Medline: 38665886]
28. Thompson DM, Booth L, Moore D, Mathers J. Peer support for people with chronic conditions: a systematic review of reviews. BMC Health Serv Res 2022 Mar 31;22(1):427 [FREE Full text] [doi: 10.1186/s12913-022-07816-7] [Medline: 35361215]
29. Rowlands H, Aylwin-Foster E, Hotopf M, Rayner L, Simpson A, Lavelle G, et al. A qualitative exploration of the psychosocial needs of people living with long-term conditions and their perspectives on online peer support. Health Expect 2023 Oct;26(5):2075-2088 [FREE Full text] [doi: 10.1111/hex.13814] [Medline: 37458403]
30. Sanders EBN, Stappers PJ. Probes, toolkits and prototypes: three approaches to making in codesigning. CoDesign 2014 Mar 06;10(1):5-14. [doi: 10.1080/15710882.2014.888183]
31. Berry N, Machin M, Ainsworth J, Berry K, Edge D, Haddock G, et al. Developing a theory-informed smartphone app for early psychosis: learning points from a multidisciplinary collaboration. Front Psychiatry 2020;11:602861 [FREE Full text] [doi: 10.3389/fpsyt.2020.602861] [Medline: 33362612]
32. Callan JA, Dunbar Jacob J, Siegle GJ, Dey A, Thase ME, DeVito Dabbs A, et al. CBT MobileWork©: user-centered development and testing of a mobile mental health application for depression. Cogn Ther Res 2020 Sep 25;45(2):287-302. [doi: 10.1007/s10608-020-10159-4]
33. Jaspers MWM, Steen T, van den Bos C, Geenen M. The think aloud method: a guide to user interface design. Int J Med Inform 2004 Nov;73(11-12):781-795. [doi: 10.1016/j.ijmedinf.2004.08.003] [Medline: 15491929]
34. Eccles DW, Arsal G. The think aloud method: what is it and how do I use it? Qual Res Sport Exerc Health 2017 May 26;9(4):514-531. [doi: 10.1080/2159676x.2017.1331501]
35. Crane D, Garnett C, Brown J, West R, Michie S. Factors influencing usability of a smartphone app to reduce excessive alcohol consumption: think aloud and interview studies. Front Public Health 2017;5:39 [FREE Full text] [doi: 10.3389/fpubh.2017.00039] [Medline: 28421175]
36. Perski O, Short CE. Acceptability of digital health interventions: embracing the complexity. Transl Behav Med 2021 Jul 29;11(7):1473-1480 [FREE Full text] [doi: 10.1093/tbm/ibab048] [Medline: 33963864]
37. Filipe A, Renedo A, Marston C. The co-production of what? Knowledge, values, and social relations in health care. PLOS Biol 2017;15(5):e2001403 [FREE Full text] [doi: 10.1371/journal.pbio.2001403] [Medline: 28467412]
38. Lambert N, Carr S. 'Outside the original remit': co-production in UK mental health research, lessons from the field. Int J Ment Health Nurs 2018;27(4):1273-1281 [FREE Full text] [doi: 10.1111/inm.12499] [Medline: 29920906]
39. Elwyn G, Nelson E, Hager A, Price A. Coproduction: when users define quality. BMJ Qual Saf 2020;29(9):711-716 [FREE Full text] [doi: 10.1136/bmjqs-2019-009830] [Medline: 31488570]
40. UK standards for public involvement. NIHR. 2019 Nov. URL: https://www.cso.scot.nhs.uk/wp-content/uploads/UK-Standards-for-Public-Involvement_Nov19.pdf [accessed 2026-03-06]

XSL•FO
RenderX

41. Payment guidance for researchers and professionals. NIHR. 2024 Dec 5. URL: https://www.nihr.ac.uk/about-us/who-we-are/policies-and-guidelines/payment-guidance-researchers-and-professionals [accessed 2026-03-06]
42. Payment for public involvement in health and care research: a guide for organisations on employment status and tax. NHS Health Research Authority. 2023 Apr 12. URL: https://www.hra.nhs.uk/planning-and-improving-research/best-practice/public-involvement/resources/payment-public-involvement-health-and-care-research-guide-organisations-employment-status-and-tax/ [accessed 2026-03-06]
43. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
44. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019;11(4):589-597. [doi: 10.1080/2159676x.2019.1628806]
45. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907 [FREE Full text] [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
46. Bengtsson M. How to plan and perform a qualitative study using content analysis. NursingPlus Open 2016;2:8-14. [doi: 10.1016/j.npls.2016.01.001]
47. Lavelle G, Jones HG, Carr E, Aylwin-Foster E, Lawrence V, Simpson A, et al. Online peer support for long-term conditions: protocol for a feasibility randomized controlled trial. JMIR Res Protoc 2025 Jul 23;14:e71513 [FREE Full text] [doi: 10.2196/71513] [Medline: 40700748]
48. Akyirem S, Forbes A, Wad JL, Due-Christensen M. Psychosocial interventions for adults with newly diagnosed chronic disease: a systematic review. J Health Psychol 2022 Jun;27(7):1753-1782 [FREE Full text] [doi: 10.1177/1359105321995916] [Medline: 33586486]
49. Marsh M, Shah SR, Munce SEP, Perrier L, Lee TSJ, Colella TJF, et al. Characteristics of existing online patient navigation interventions: scoping review. JMIR Med Inform 2024;12:e50307 [FREE Full text] [doi: 10.2196/50307] [Medline: 39159443]
50. Ben-Zeev D, Kaiser SM, Brenner CJ, Begale M, Duffecy J, Mohr DC. Development and usability testing of FOCUS: a smartphone system for self-management of schizophrenia. Psychiatr Rehabil J 2013;36(4):289-296 [FREE Full text] [doi: 10.1037/prj0000019] [Medline: 24015913]
51. Jaspers MWM. A comparison of usability methods for testing interactive health technologies: methodological aspects and empirical evidence. Int J Med Inform 2009;78(5):340-353. [doi: 10.1016/j.ijmedinf.2008.10.002] [Medline: 19046928]
52. Broekhuis M, van Velsen L, Peute L, Halim M, Hermens H. Conceptualizing usability for the eHealth context: content analysis of usability problems of eHealth applications. JMIR Form Res 2021 Jul 27;5(7):e18198 [FREE Full text] [doi: 10.2196/18198] [Medline: 34313594]
53. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. Manag Inf Syst Q 1989;13(3):319-340. [doi: 10.2307/249008]
54. Sbaffi L, Rowley J. Trust and credibility in web-based health information: a review and agenda for future research. J Med Internet Res 2017;19(6):e218 [FREE Full text] [doi: 10.2196/jmir.7579] [Medline: 28630033]
55. Adjekum A, Blasimme A, Vayena E. Elements of trust in digital health systems: scoping review. J Med Internet Res 2018;20(12):e11254 [FREE Full text] [doi: 10.2196/11254] [Medline: 30545807]
56. Alhammad N, Alajlani M, Abd-Alrazaq A, Epiphaniou G, Arvanitis T. Patients' perspectives on the data confidentiality, privacy, and security of mHealth apps: systematic review. J Med Internet Res 2024;26:e50715 [FREE Full text] [doi: 10.2196/50715] [Medline: 38820572]
57. Tseng HT, Ibrahim F, Hajli N, Nisar TM, Shabbir H. Effect of privacy concerns and engagement on social support behaviour in online health community platforms. Technol Forecast Soc Chang 2022;178:121592. [doi: 10.1016/j.techfore.2022.121592]
58. Zhu M, Wu C, Huang S, Zheng K, Young SD, Yan X, et al. Privacy paradox in mHealth applications: an integrated elaboration likelihood model incorporating privacy calculus and privacy fatigue. Telemat Inform 2021;61:101601. [doi: 10.1016/j.tele.2021.101601]
59. Kemp E, Koczwara B, Butow P, Turner J, Girgis A, Schofield P, et al. Online information and support needs of women with advanced breast cancer: a qualitative analysis. Support Care Cancer 2018;26(10):3489-3496. [doi: 10.1007/s00520-018-4206-1] [Medline: 29693203]
60. Brady E, Segar J, Sanders C. Accessing support and empowerment online: the experiences of individuals with diabetes. Health Expect 2017 Oct;20(5):1088-1095 [FREE Full text] [doi: 10.1111/hex.12552] [Medline: 28718928]
61. Taylor F, Gutteridge R, Willis C. Peer support for CKD patients and carers: overcoming barriers and facilitating access. Health Expect 2016;19(3):617-630 [FREE Full text] [doi: 10.1111/hex.12348] [Medline: 25649115]
62. Griffiths KM, Reynolds J, Vassallo S. An online, moderated peer-to-peer support bulletin board for depression: user-perceived advantages and disadvantages. JMIR Ment Health 2015;2(2):e14 [FREE Full text] [doi: 10.2196/mental.4266] [Medline: 26543919]
63. Lazard AJ, King AJ. Objective design to subjective evaluations: connecting visual complexity to aesthetic and usability assessments of eHealth. Int J Hum Comput Interact 2019;36(1):95-104. [doi: 10.1080/10447318.2019.1606976]

XSL•FO
RenderX

64. Internet access – households and individuals, Great Britain statistical bulletins. Office for National Statistics. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/internetaccesshouseholdsandindividuals/previousreleases [accessed 2026-03-06]
65. Internet users – Table 2b: Recent and lapsed internet users and internet non-users, by age group (years) and sex, UK, 2013 to 2020. Office for National Statistics. 2021 Apr 6. URL: https://www.ons.gov.uk/businessindustryandtrade/itandinternetindustry/datasets/internetusers [accessed 2026-03-06]

Abbreviations

KCL: King's College London
PPI: patient and public involvement
RAG: research advisory group

Edited by A Stone; submitted 25.Jun.2025; peer-reviewed by A Barooah, E Littlewood; comments to author 20.Oct.2025; revised version received 02.Feb.2026; accepted 18.Feb.2026; published 24.Mar.2026.

Please cite as:
Jones HG, Lavelle G, Aylwin-Foster E, Carr E, Simpson A, Hotopf M, Lawrence V
Coproducing an Online Platform for People With Long-Term Physical Health Conditions: Development and Usability Study
J Med Internet Res 2026;28:e79666
URL: https://www.jmir.org/2026/1/e79666
doi:10.2196/79666
PMID:

©Hannah Grace Jones, Grace Lavelle, Elly Aylwin-Foster, Ewan Carr, Alan Simpson, Matthew Hotopf, Vanessa Lawrence. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of the Components of a Digital Multiple Health Behavior Change Intervention Among Individuals Seeking Help Online (Coach): Factorial Randomized Trial

Joel Crawford[1], PhD; Katarina Åsberg[1], PhD; Jenny Blomqvist[1], MSc; Oskar Lundgren[1,2], MD, PhD; Hanna Henriksson[1,3], PhD; Pontus Henriksson[1], PhD; Preben Bendtsen[1,4], MD, PhD; Marie Löf[5], PhD; Marcus Bendtsen[1], PhD

[1]Department of Health, Medicine and Caring Sciences, Linköping University, Linköping, Östergötland, Sweden

[2]Crown Princess Victoria Children's Hospital, Linköping University Hospital, Linköping, Östergötland, Sweden

[3]Unit for Strategic Healthcare, Region Östergötland, Linköping, Östergötland, Sweden

[4]Motala Hospital, Motala, Östergötland, Sweden

[5]Department of Medicine Huddinge, Karolinska Institutet, Stockholm, Stockholm, Sweden

Corresponding Author:
Marcus Bendtsen, PhD
Department of Health, Medicine and Caring Sciences
Linköping University
Campus US
Linköping, Östergötland, 581 83
Sweden
Phone: 46 13281000
Email: marcus.bendtsen@liu.se

Abstract

Background: Extant digital multiple health behavior change interventions have shown promise in various populations; however, evidence for a broader approach among the general population is lacking. Moreover, existing interventions often contain several components but are typically assessed as a whole, meaning it remains unclear to what extent individual components contribute to intervention effects and how they may interact to influence health outcomes.

Objective: This study estimates the effects of 6 components of a digital health behavior change intervention on alcohol, diet, physical activity, and smoking outcomes among individuals searching for help online.

Methods: A double-blind randomized factorial trial design with 6 two-level factors was used. Adults from the general public in Sweden who were seeking help to change their behaviors were recruited through web searches and social media. Participants were eligible if they were 18 years or older and had at least one health behavior classified as unhealthy. Effects of 6 components were estimated: screening/feedback, goal-setting/planning, motivation, skills/know-how, mindfulness, and self-authored SMS text messages. Primary outcomes were weekly alcohol consumption and frequency of heavy episodic drinking, average daily fruit and vegetable consumption, weekly moderate-to-vigorous physical activity, and 4-week point-prevalence smoking.

Results: A total of 5419 individuals were randomized. Overall, the screening/feedback component was the most effective for changing health behaviors, along with goal-setting/planning and motivation to change. In particular, there was evidence that screening/feedback increased average daily portions of fruit and vegetables at 2 months (mean difference 0.17, compatibility interval [CoI] 0.09-0.25, probability of effect [POE] >99.9%) and at 4 months (mean difference 0.13, CoI 0.04-0.21, POE 99.9%) and reduced the frequency of heavy episodic drinking at 4 months (incidence rate ratio 0.91, CoI 0.81-1.03, POE 94.2%). Components also interacted to further improve health outcomes, most notably the combination of screening/feedback with motivation to change, which further increased fruit and vegetable consumption (2 months: mean difference 0.20, CoI 0.09-0.30, POE >99.9%; 4 months: mean difference 0.17, CoI 0.05-0.29, POE 99.8%).

Conclusions: The results from this study contribute to the development of more effective interventions by providing novel insights into the effects of individual and pairwise components of complex digital health behavior change interventions.

Trial Registration: ISRCTN Registry ISRCTN16420548; http://www.isrctn.com/ISRCTN16420548

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2022-061024

XSL•FO
RenderX

(J Med Internet Res 2026;28:e88881) doi:10.2196/88881

KEYWORDS

behavior change; digital interventions; multiple lifestyle; public health

Introduction

Alcohol consumption, diet, physical activity, and smoking are leading factors determining general health and susceptibility to disease [1] and are prominent causes of noncommunicable diseases (eg, cardiovascular disease, respiratory disease, cancer, and diabetes) [1]. The increasing prevalence of chronic disease is one of the major challenges to global health, and unlike intrinsic etiologies of noncommunicable diseases, health behaviors are modifiable [2]. In light of this, the World Health Organization has set a charter for member states to lower the incidence of noncommunicable diseases by implementing measures to reduce engagement in health risk behaviors while increasing engagement in health promotion behaviors [2]. To achieve this, there is a need to develop and implement effective measures to help individuals improve their health behaviors.

In Sweden, individuals regularly engage in unhealthy behaviors; national surveys have shown that 40% of Swedish individuals drink at levels that indicate risky alcohol use, 66% fail to eat an adequate amount of vegetables, 32% report insufficient physical activity, 5.8% smoke daily, and 51% have obesity or are overweight [3,4]. Despite macro-level interventions aimed at curbing the incidence of unhealthy behaviors [5-7], the prevalence of these behaviors has remained consistent over the last 10 years, with the exception of tobacco smoking [4]. Overall, unhealthy lifestyle behaviors present a significant health burden for premature mortality and morbidity in Sweden, with 15% of all deaths and 2974 disability-adjusted life years (DALYs) attributed to tobacco use, 15% of all deaths and 2204 DALYs attributed to poor dietary behaviors, 5% of all deaths and 1292 DALYs attributed to alcohol use, and 2% of all deaths and 266 DALYs attributed to insufficient physical activity in 2019 (DALYs provided per 100,000 population) [8,9]. Thus, new means of intervening in the general public are warranted to avoid another decade of stagnating or increasing prevalence of unhealthy behaviors and subsequent maladies.

Current intervention efforts have relied on public health recommendations, national ad campaigns, and policy measures such as restrictions on availability and advertising [5-7]. While macro-level approaches do have some merit (ie, the ability to reach a significant portion of the populace), they are unable to address the barriers and intricacies of health behavior change; for example, they provide little support on how to change and do not consider that health behaviors often cluster and interact [10,11]. Evidence indicates that health behaviors consistently co-occur: for example, smoking and alcohol use, smoking and unhealthy diet, and physical inactivity and unhealthy diet [12]. A recent study conducted in Sweden found that 36% of adults had a co-occurrence of at least 2 unhealthy behaviors, with 10% having a co-occurrence of 3 behaviors and 2% having a co-occurrence of 4 behaviors [13]. The clustering of unhealthy behaviors significantly increases the risk of all-cause mortality and morbidity compared with a single behavior alone [14-16]. Current intervention efforts typically focus on a single behavior, meaning they are unable to target underlying clusters that considerably increase health risk [12]; hence, there is a need for interventions that can address the intricacies and barriers this presents to health behavior change.

One method to overcome issues with current prevention efforts is to offer digital support tools for those seeking help online. This approach has the potential to be widely disseminated while also addressing the intricacies of health behavior change. For instance, content can be tailored to personal dimensions of behavior and can also address barriers to health (eg, reducing stigma associated with seeking help) [17]. In addition, a digital approach may be particularly apt for Sweden due to the near-ubiquitous ownership of smartphones and the high prevalence of seeking help online—66% of individuals aged 16-84 years sought health advice online in 2023 [18]. While effectiveness trials for digital behavior change interventions targeting the Swedish general population have shown effectiveness for smoking cessation [19] and reducing alcohol consumption [20], there is currently a lack of digital interventions targeting multiple health behaviors [21]. Existing multiple health behavior change interventions have shown promise for various subpopulations, including university students, patients with a specific malady, and different professions [22-25]; however, evidence for a broader approach for the general population is lacking.

In addition, many current interventions use a sociocognitive framework to target modifiable determinants of health, such as intentions, attitudes, beliefs, self-efficacy, risk perceptions, motivation, planning, and expectancies [26]. Sociocognitive models of health and behavior change (eg, Theory of Planned Behavior, Health Action Process Approach, Capability-Opportunity-Motivation-Behavior [COM-B] model [27-29]) highlight that health behavior is seldom influenced by a single factor, but rather by a configuration of constructs that interact to influence health outcomes; they also indicate that targeting one factor without addressing deficits in others may not result in meaningful change [27,28]. Current interventions that target improvements in these determinants often use various components to instill change; however, they typically assess the effectiveness of the components as a whole [26]. This means that we are unsure to what degree separate components contribute to the effects of the interventions, how they interact to influence health outcomes, or whether certain components are effective only in conjunction with other commensurate components; for example, action planning may only be effective when self-efficacy or motivation is high [27,28]. Hence, gaining knowledge regarding the separate and interactive effects of intervention components would enable a more nuanced understanding of the effects of health behavior change interventions.

We conducted this study in response to the sparsity of studies investigating multiple health behavior change interventions for

XSL•FO
RenderX

the general public and the need to better understand the effects of individual components of digital behavior change interventions. We used a factorial trial design to estimate the effects of 6 separate intervention components of a novel digital intervention targeting multiple health behaviors. The study was carried out among individuals seeking help online. Specifically, the study aimed to achieve the following:

- Estimate the effects of 6 individual intervention components on reducing alcohol use, improving dietary behaviors, increasing physical activity, and increasing smoking cessation.
- Estimate the pairwise interaction effects of the 6 intervention components on reducing alcohol use, improving dietary behaviors, increasing physical activity, and increasing smoking cessation.

Methods

Study Design

We conducted a factorial randomized trial (6 factors representing 6 components, C1-C6, each with 2 levels representing the absence or presence of the component) [30] to estimate the effects of a digital multiple health behavior change intervention among individuals seeking help online. The trial was an extension of the MoBILE (Mobile Health Multiple Lifestyle Behavior Interventions Across the Lifespan) research program [31] and was prospectively registered (ISRCTN16420548). A full protocol and analysis plan were published prospectively [32]. This report follows CONSORT (Consolidated Standards of Reporting Trials; Multimedia Appendix 1) guidelines [33], including extensions for factorial trials [34].

Study Setting, Recruitment, and Participants

The target population was Swedish adults seeking help online to change their health behaviors. Participants were recruited through advertisements on Google (Alphabet Inc), Bing (Microsoft Corporation), Facebook (Meta Platforms, Inc), and websites related to lifestyle and behavior change. Individuals interested in participating sent an SMS text message to a dedicated phone number and received a reply with a link to a page containing informed consent materials (see Multimedia Appendix 2). All individuals who provided consent were then asked to complete a baseline survey, which was also used to assess eligibility (see Multimedia Appendix 3). To be included, individuals had to be 18 years of age or older and fulfill at least one of the conditions presented in Textbox 1. Note that, while there is no safe limit for alcohol consumption [35], we used Swedish national guidelines for alcohol to identify individuals who could benefit greatly from reducing their consumption.

As the study was conducted via participants' smartphones, individuals without access to one were unable to take part. This also applied to those who did not have sufficient understanding of Swedish to comprehend the study materials.

Textbox 1. Criteria for inclusion.

- **Weekly alcohol consumption:** Consumed 10 or more (for women) or 15 or more (for men) standard drinks of alcohol in the past week. In Sweden, a standard drink is defined as 12 g of pure alcohol.
- **Heavy episodic drinking:** Consumed 4 or more (for women) or 5 or more (for men) standard drinks of alcohol on a single occasion at least once in the past month.
- **Fruit and vegetables:** Consumed less than 500 g of fruit and vegetables on average per day in the past week.
- **Moderate-to-vigorous physical activity:** Spent less than 150 minutes on moderate-to-vigorous physical activity in the past week.
- **Smoking:** Smoked at least one cigarette in the past week.

Interventions

The Coach intervention studied in this trial was a smartphone-based intervention comprising 6 components: (C1) screening/feedback, (C2) goal-setting/planning, (C3) motivation to change, (C4) skills/know-how, (C5) mindfulness, and (C6) self-authored SMS text messages. The components were informed by sociocognitive models of health [36] and have been included in interventions that have proven successful in facilitating behavior change [19,20]. While built on what is already known about digital health behavior change interventions, the novelty of the Coach intervention lies in its approach of including multiple behaviors within a single intervention. The intervention was designed so that each component could be presented to participants in a menu list, making it possible to add or remove components based on factorial conditions.

A core feature of the intervention was an SMS text message sent to participants each Sunday afternoon, containing a link and reminder to access the intervention materials. The intervention components, depending on factorial allocation, and weekly reminders were available to participants over 4 months. Table 1 provides an overview of each component, along with its relevant behavior change techniques [37]. Please see Multimedia Appendix 4 for full details of the intervention. As Coach was designed to be a multiple health behavior change intervention, all participants, regardless of their baseline screening, received intervention materials for alcohol, diet, physical activity, and smoking. There was no fixed pathway through the materials; rather, they were presented as a toolbox. Therefore, participants could engage with materials concerning different behaviors according to their own preferences and were not required to engage with materials for behaviors they did not find relevant at the time.

XSL•FO
RenderX

Table 1. Description of intervention components, conceptual framework, practical application, and factorial conditions.

Component	Conceptual framework	Practical application	Factorial conditions
Component 1: screening and feedback	*Self-monitoring* has been shown to be a potentially effective strategy for reducing excessive alcohol [37-40] consumption and promoting healthy eating and physical activity [41,42]. These include BCT[a] 1.6 (Discrepancy between current behavior and goals), BCT 2.2 (Feedback on behavior), BCT 2.3 (Self-monitoring of behavior), and BCT 6.2 (Social comparison).	Every Sunday afternoon, participants received an SMS text message with a hyperlink to a questionnaire regarding their current health behaviors. Once the questionnaire was completed, participants received feedback on their current behaviors in comparison to national guidelines. Thereafter, users were given access to the rest of the components (depending on allocation).	When this component was absent, participants were not asked to respond to the screening questionnaire but were instead shown national guidelines without any personal feedback.
Component 2: goal-setting and planning	*Self-regulatory* skills and capacity via goal-setting and planning. Planning-related activities such as goal-setting, action planning, and practicing behavior have been shown to be effective in lifestyle behavior interventions [41,43-48]. These include BCT 1.1 (Goal-setting), BCT 1.2 (Problem solving), BCT 1.4 (Action planning), BCT 7.1 (Prompts/cues), and BCT 8.1 (Behavior practice/rehearsal).	This component allowed participants to set 1 or more goals for their future behavior. It included action planning for how they were to progress toward their goals, preparation for motivational struggles (coping planning), and strategies for rewarding themselves upon success. Participants could also create their own or accept ready-made challenges for the coming week, such as walking for 15 minutes each day or not drinking any alcohol the coming week. Reminders were sent via SMS text messages to participants about their goals and challenges throughout the week (up to 4 messages).	When absent, this component was not visible, and goal-setting reminders were not available.
Component 3: motivation	Digital health behavior change interventions have been shown to enhance s*elf-efficacy*; however, there is a lack of consensus across reviews as to which content works best to facilitate an increase in self-efficacy [49]. These include BCT 5.1 (Information about health consequences), BCT 9.1 (Credible source), BCT 9.2 (Pros and cons), and BCT 9.3 (Comparative imagining of future outcomes).	This component contained information and tools to increase participants' awareness of their own motivation, encourage commitment, and boost self-efficacy. This included information on negative health consequences, costs induced by certain behaviors, and reflective tasks via texts, videos, and exercises. If participants chose, they could also activate motivational SMS text messages (derived from previously developed and evaluated interventions [50-56]) sent to them throughout the week (a maximum of 8-10 messages per week).	When absent, this component was not visible, and SMS text messages were not available.
Component 4: skills and know-how	Intervention content based on shaping knowledge, aimed at increasing an individual's understanding and awareness to facilitate behavior change, has been shown to be effective [44,45,57,58]. These include BCT 4.1 (Instructions on how to perform a behavior), BCT 8.2 (Behavior substitution), BCT 8.3 (Habit formation), and BCT 8.7 (Graded tasks).	This component aimed to increase participants' skills and know-how by providing concrete tips on how to initiate and maintain lasting changes in everyday life by repetition and substitution. Participants were given strategies (eg, how to say no to alcoholic beverages or how to introduce vegetables into their meals). If participants chose to, they could also activate SMS text messages with tips and know-how, which were sent to them throughout the week (a maximum of 8-10 messages per week).	When absent, this component was not visible, and SMS text messages were not available.
Component 5: mindfulness	The mindfulness exercises were based on previous research and considered to be evidence-based methods to improve the mental well-being of clinical populations, while effects on behavior change in nonclinical settings are less well-studied [59-63].	This component aimed to increase users' awareness of their own lived experience and strengthen their capacity for a nonreactive, compassionate, and less stressful way of being in the world. Mindfulness exercises were offered to participants, including guided meditations.	When absent, this component was not visible, and guided meditations were not available.

XSL•FO
RenderX

Component	Conceptual framework	Practical application	Factorial conditions
Component 6: self-authored SMS text messages	SMS text message self-authorship is generally understudied in the literature but was included in an effective digital alcohol intervention [64].	This component allowed participants to self-author up to 3 SMS text messages and schedule them to be sent to themselves throughout the week at times of their choosing. For example, a participant could write an SMS text message reminding themselves to eat 2 fruits each day, avoid drinking on Wednesdays, or go for a walk with a friend.	When absent, this component was not visible, and the ability to self-author SMS text messages was not available.

[a]BCT: behavior change technique.

Outcomes and Measures

Outcomes are listed in Textbox 2 and subsequently explained. All questionnaires used in the trial can be found in Multimedia Appendix 3.

Weekly alcohol consumption was assessed using a short-term recall method that asked participants to record the number of standard drinks consumed in the past week [65]. The frequency of heavy episodic drinking was assessed by asking participants to report how many times they consumed 4 or more (for women) or 5 or more (for men) standard drinks per session over the past month. These 2 outcomes are both part of the core outcome set for brief alcohol interventions [66,67].

Diet and physical activity were assessed using items based on a questionnaire from the Swedish National Board of Health and Welfare [21]. The items were adapted to include portion sizes (100 g of fruit or vegetables, respectively) and asked how many portions were consumed per day in the preceding week. Sugary drink consumption was assessed by asking participants to report the number of units (33 cl) of sugary drinks consumed in the past week. Moderate-to-vigorous physical activity (MVPA) was assessed with 2 items asking for the number of minutes spent in both moderate and vigorous activity in the past week; responses were summed to provide a total MVPA value. BMI was measured by asking participants to report their height (cm) and weight (kg).

Four-week point-prevalence smoking abstinence (no cigarettes in the past 4 weeks) was assessed using a binary item. Those reporting nonabstinence were asked to record the number of cigarettes smoked in the past week. These items were based on suggestions from the Society for Research on Nicotine [68].

Perceived stress was assessed using the short-form Perceived Stress Scale [69], and quality of life (QoL) was measured using Patient-Reported Outcomes Measurement Information System (PROMIS) Global 10 [70].

Textbox 2. Primary and secondary outcomes.

Primary outcomes

- **Alcohol:** Weekly alcohol consumption; monthly frequency of heavy episodic drinking
- **Diet:** Average daily consumption of fruit and vegetables
- **Physical activity:** Weekly moderate-to-vigorous physical activity
- **Smoking:** Four-week point prevalence of smoking abstinence

Secondary outcomes

- Weekly consumption of sweets and snacks
- Weekly consumption of sugary drinks
- BMI
- Weekly number of cigarettes smoked
- Perceived stress
- Quality of life

Follow-Up

Primary and secondary outcomes were assessed at 2 and 4 months after randomization. The 2-month follow-up facilitated the assessment of the immediate effects of the components, while the 4-month follow-up facilitated the estimation of effects after prolonged access. Additional longer follow-up intervals were not included to reduce participant burden, and because attrition was expected to be much higher the longer the time from baseline. An additional follow-up was conducted at 1 month after randomization to assess the proposed mediators of the intervention. Mediator analyses will be reported separately. Follow-ups were initiated by SMS text messages containing hyperlinks to questionnaires (see Multimedia Appendix 3). A total of 2 reminder SMS text messages were sent, 2 days apart, to nonresponders. If no response was received, participants were contacted by phone and asked to respond to the primary outcome measures only. A total of 5 call attempts were made.

XSL•FO
RenderX

Randomization and Blinding

We used block randomization with random block sizes of 64 and 128 to allocate participants equally among factors. The process was automated, with the sequence computer-generated and the backend server allocating eligible participants after completion of baseline assessments. Neither research personnel nor participants were able to influence allocation. Research personnel were blinded to allocation throughout the trial. Participants were also blinded, as they had access to different components of the trial but were unaware of the other available conditions. All measurements, including baseline and follow-up, were conducted through digital questionnaires completed by participants independently on their smartphones, thereby preserving blinding during measurement. Attempts to collect data from nonresponders via smartphones introduced a risk of disclosure of allocation for some participants; however, research personnel were experienced and instructed to ask only about the outcome measures. In addition, given that there were 64 conditions in the trial, it would have been difficult for participants to inadvertently reveal their allocation.

Statistical Analysis

Intention-to-Treat Analysis and Bayesian Multilevel Modeling

We conducted analyses keeping participants in the conditions to which they were randomly allocated (intention to treat). We estimated models using both available data and data with missing values imputed (multiple imputation with chained equations [71], with 100 datasets and 30 iterations using predictive mean matching). All statistical analyses were prespecified in the study protocol [32].

We contrasted the presence versus absence of components using multilevel regression models with covariates for time-by-component interactions and participant-level adaptive intercepts. We also estimated models with pairwise interactions between components. All models were adjusted for the presence or absence of all components, as well as baseline measures of age, sex, importance, confidence, and know-how. Primary outcomes, perceived stress, and weekly number of cigarettes smoked were additionally adjusted for their respective baseline measures. BMI, sugary drink consumption, and sweets and snacks were adjusted for baseline MVPA minutes per week and average daily intake of fruit and vegetables. QoL was adjusted for perceived stress at baseline.

Bayesian inference was used to estimate the parameters of the models [72]. We used standard normal priors for all covariates and random intercepts to reflect a conservative view of effect magnitudes; thus, the priors induced shrinkage to protect against spurious findings. For each coefficient of interest, we report the marginal posterior probability of effect (POE), using the median as a point estimate of the magnitude of the effect. We also report 95% compatibility intervals (CoIs), defined by the 2.5% and 97.5% percentiles of the posterior distributions.

Primary and Secondary Outcomes

We analyzed primary outcomes among participants who fulfilled each respective inclusion criterion; for example, the MVPA outcome was analyzed among those with less than 150 minutes of MVPA in the past week at baseline. Secondary outcomes were analyzed among all participants, except for the number of cigarettes smoked in the past week, which was analyzed among baseline smokers who continued to smoke.

All count variables, including total weekly alcohol consumption, frequency of heavy episodic drinking, consumption of sugary drinks, sweets and snacks, and cigarettes smoked per week, were analyzed using negative binomial regression models. Average daily intake of fruit and vegetables, minutes of MVPA per week, BMI, QoL, and perceived stress were analyzed using linear regression. Point-prevalence smoking abstinence was analyzed using logistic regression.

Ancillary Analyses

We studied attrition by estimating the odds of responding to follow-up, conditional on baseline characteristics and the presence or absence of the 6 components (factors), using logistic regression. We estimated 1 model without interaction terms and a second model including interaction terms between factors and baseline characteristics. We used Cauchy priors to promote a parsimonious model (centered at 0 with a standard normal hyperprior for scale). We also modeled primary outcome data conditional on the number of attempts to collect follow-up. Under the assumption that late responders are more similar to nonresponders than early responders are, any association between the number of attempts and outcomes may indicate systematic attrition.

Post Hoc Analyses

To investigate whether the number of components to which participants had access affected outcomes, we post hoc modeled outcomes conditional on the number of components. The same model specifications were used as in the primary analyses, with dummy variables for components replaced by a variable for component count.

Sample Size

We used a Bayesian sequential design to monitor when to stop recruitment [73]. As 4-month follow-up data became available, we modeled each of the primary outcomes according to the analysis plan, and effect estimates for each component were assessed against targets for effect, harm, and futility (see protocol for details of the criteria [32]). We anticipated that recruitment would not last more than 12 months; however, we allowed recruitment to continue for a total of 24 months and stopped recruitment at that point, despite targets not being met for all outcomes.

Ethical Considerations

Ethical approval was received on August 11, 2021, from the Swedish Ethical Review Authority (approval Dnr 2021-02855). All participants were provided with study information and gave informed consent before responding to the baseline questionnaire and subsequently being randomized. Participants received no compensation for participating in the trial. Participants' phone numbers were encrypted and stored in the study database to link baseline and follow-up data. At project completion, the

XSL•FO
RenderX

encrypted phone numbers will be deleted and the data will be anonymized.

Results

Overview

A CONSORT participant flow diagram is presented in Figure 1 (sign-up and randomization) and Figure 2 (follow-up). Table 2 provides an overview of participants' baseline characteristics. Between October 19, 2021, and October 19, 2023, a total of 6405 individuals showed interest in the study, of whom 5574 consented. A total of 5445 completed the baseline questionnaire, revealing that 26 participants were not eligible for inclusion. The remaining 5419 individuals were randomized. Among the 5419 participants, 650 (11.99%) decided to stop the reminder SMS text messages before the 16-week program had ended. We did not investigate the reasons why individuals chose to stop the reminders; however, all randomly allocated participants were sent follow-up questionnaires. Primary outcome data were available for 3371 (62.21%) participants at the 2-month follow-up interval and for 2786 (51.41%) at the 4-month interval.

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) participant flow diagram—sign-up and randomization.

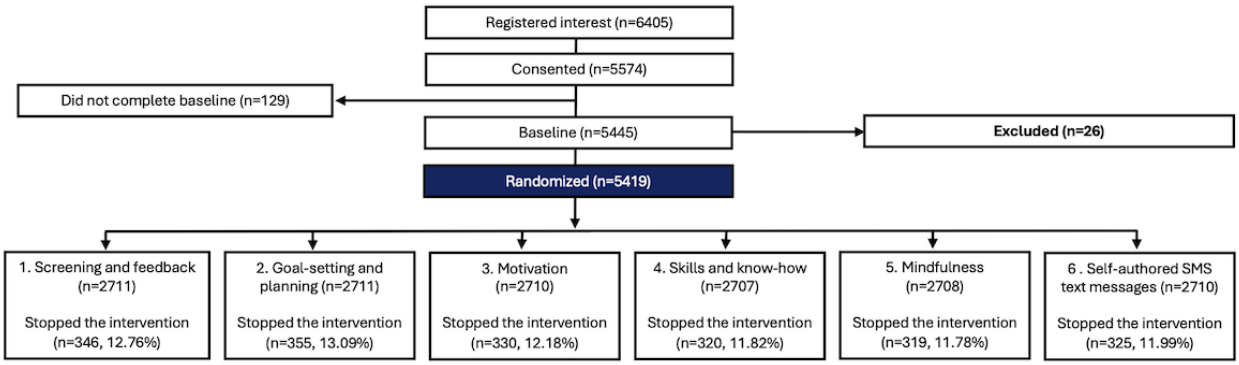

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) participant flow diagram—follow-up.

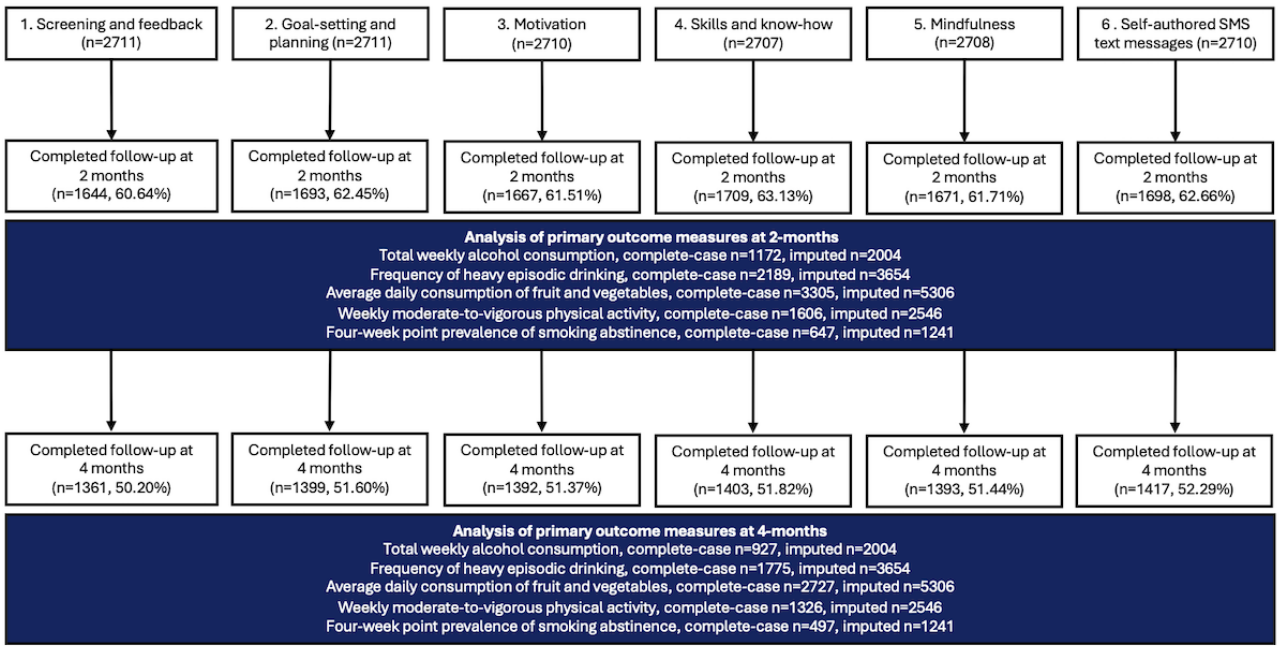

XSL•FO
RenderX

Table 2. Baseline characteristics of randomly allocated participants.

Characteristics	Total (N=5419)	Screening/feedback (n=2711)	Goal-setting/planning (n=2711)	Motivation (n=2710)	Skills/know-how (n=2707)	Mindfulness (n=2708)	Self-authored SMS text message (n=2710)
Sex, n (%)							
Women	4301 (79.37)	2140 (78.94)	2141 (78.97)	2151 (79.37)	2172 (80.24)	2164 (79.91)	2136 (78.82)
Men	1118 (20.63)	571 (21.06)	570 (21.03)	559 (20.63)	535 (19.76)	544 (20.09)	574 (21.18)
Age (years), mean (SD)	50 (11.7)	49.8 (11.8)	50 (11.7)	50.6 (11.6)	50 (11.4)	50.4 (11.6)	51 (11.5)
Alcohol, mean (SD)							
Total weekly alcohol consumption (standard drinks)	8.89 (8.94)	8.88 (8.99)	8.91 (9.00)	8.54 (8.58)	9.04 (8.67)	8.77 (8.97)	9.48 (8.87)
Frequency of heavy episodic drinking	4.80 (7.10)	4.77 (7.07)	4.91 (7.27)	4.54 (6.57)	4.76 (6.57)	4.46 (6.19)	4.22 (5.54)
Smoking							
Number of smokers, n (%)	1240 (22.88)	636 (23.46)	626 (23.09)	632 (23.32)	633 (23.38)	615 (22.71)	636 (23.47)
Number of cigarettes last week, mean (SD)	15.27 (38.0)	15.28 (37.5)	15.04 (38.2)	16.01 (39.9)	15.26 (38.2)	11.89 (34.2)	11.72 (32.3)
Physical activity							
Moderate-to-vigorous physical activity (minutes), mean (SD)	220.4 (226.9)	226.5 (231.4)	208.7 (221.5)	222.0 (232.1)	225.3 (220.3)	210.2 (201.2)	215.3 (234.2)
Dietary behavior							
Average daily fruit and vegetable consumption (portions), mean (SD)	1.49 (1.17)	1.53 (1.18)	1.39 (1.12)	1.47 (1.18)	1.47 (1.17)	1.61 (1.17)	1.34 (1.04)
Stress							
Self-perceived stress[a], mean (SD)	7.62 (2.80)	7.60 (2.77)	7.59 (2.89)	7.67 (2.92)	7.82 (2.69)	7.51 (2.65)	7.55 (2.44)
Psychosocial measures, median (IQR)							
Importance[b]	10 (8-10)	10 (8-10)	10 (8-10)	10 (8-10)	10 (8-10)	10 (8-10)	10 (9-10)
Confidence[b]	7 (5-8)	7 (5-8)	7 (5-8)	7 (5-8)	7 (5-8)	6 (5-8)	6 (5-8)
Know-how[b]	6 (5-8)	6 (5-8)	6 (5-8)	6 (5-8)	6 (5-8)	6 (5-8)	6 (4-8)

[a]The 4-item Perceived Stress Scale with total scores ranging from 0 to 16.

[b]Single item with 1-10 response options.

Attrition

Results from the attrition analyses can be found in Multimedia Appendices 5 and 6. In summary, we found evidence that older participants were more likely to respond to follow-up than younger participants at both follow-up intervals. In addition, those with more frequent episodes of heavy drinking and those who smoked more at baseline were less likely to respond to follow-up. Notably, these associations were not moderated by differential access to components.

We found evidence suggesting that late responders differed from early responders with respect to fruit and vegetable consumption, MVPA, and smoking. The evidence also suggested that these associations were moderated by differential access to components. Under the assumption that late responders are more similar to nonresponders than early responders are, this provides evidence of systematic attrition.

XSL•FO
RenderX

Primary and Secondary Outcomes

Overview

Multimedia Appendix 7 contains tables reporting the main effects and 2-way interactions for the intervention components with available data, whereas Multimedia Appendix 8 contains the corresponding effect estimates with missing data imputed. Here, we highlight findings for which the evidence of effects was strongest, quoting estimates from analyses of available data.

Primary Outcomes

Total Weekly Alcohol Consumption

There was no consistent evidence of effects from any of the components on weekly alcohol consumption, either individually or in combination. The relatively strongest evidence was found for an effect of screening/feedback (C1) at 4 months (incidence rate ratio [IRR] 0.94, CoI 0.83-1.07, POE 82.4%). In addition, there was some evidence of an effect of goal-setting/planning (C2) at 2 months (IRR 0.95, CoI 0.85-1.06, POE 81.8%), but this effect was not observed at 4 months. The evidence suggested that combining screening/feedback (C1) with skills/know-how (C4) resulted in even lower alcohol consumption at 4 months (IRR 0.92, CoI 0.76-1.11, POE 81.4%). These findings persisted when imputing missing data.

Frequency of Heavy Episodic Drinking

Among individual components, the relatively strongest evidence for reduced frequency of heavy episodic drinking was found for screening/feedback (C1) at 4 months (IRR 0.91, CoI 0.81-1.03, POE 94.2%). However, there was considerable evidence of harm from multiple combinations of components at the 4-month follow-up interval. For instance, combining goal-setting/planning (C2) with motivation (C3) (IRR 1.14, CoI 0.97-1.31, POE 93.4%), mindfulness (C5) (IRR 1.15, CoI 0.97-1.37, POE 94.6%), or self-authored SMS text messages (C6) (IRR 1.17, CoI 0.98-1.38, POE 96.3%) was associated with increased frequency of heavy episodic drinking. These findings persisted in analyses with missing data imputed.

Daily Portions of Fruit and Vegetables

Individually, there was strong evidence that screening/feedback (C1) increased daily portions of fruit and vegetables at 2 months (mean difference 0.17, CoI 0.09-0.25, POE >99.9%) and at 4 months (mean difference 0.13, CoI 0.04-0.21, POE 99.9%). There was also evidence that combining screening/feedback (C1) with other components increased effects, most prominently when combined with motivation (C3) at 2 months (mean difference 0.20, CoI 0.09-0.30, POE >99.9%) and at 4 months (mean difference 0.17, CoI 0.05-0.29, POE 99.8%). There was evidence that goal-setting/planning (C2) decreased fruit and vegetable consumption at 4 months (mean difference −0.08, CoI −0.16 to 0.00, POE 97.1%), and even more so when combined with skills/know-how (C4) (mean difference −0.14, CoI −0.26 to −0.02, POE 99.1%). These findings persisted in the imputed data.

Weekly MVPA

There was no consistent evidence of effects from any of the components, nor from combinations of components, on MVPA. The relatively strongest evidence in the available-data analyses suggested that mindfulness increased minutes of MVPA at 4 months (mean difference 10.3, CoI −11.3 to 32.1, POE 82.3%), and even more so when combined with screening/feedback (C1), goal-setting/planning (C2), and motivation (C3). However, in the imputed analyses, there was no evidence of these effects.

Smoking

There were several individual components for which there was evidence of effects on the point prevalence of smoking cessation in the available-data analyses. Notably, motivation (C3) increased the odds of smoking cessation at 4 months (odds ratio [OR] 2.4, CoI 0.98-5.99, POE 97.2%), whereas mindfulness (C5) decreased the odds of smoking cessation at 2 months (OR 0.35, CoI 0.15-0.78, POE 99.5%). These findings persisted in the imputed data, whereas evidence for other individual component effects did not. Evidence for effects of combined components was most consistent across available-data and imputed analyses for the combination of motivation (C3) and self-authored SMS text messages (C6) at 4 months (OR 2.42, CoI 0.70-8.72, POE 91.9%), and the combination of mindfulness (C5) and self-authored SMS text messages (C6) at 2 months (OR 0.23, CoI 0.07-0.71, POE 99.4%). The harmful effects of mindfulness that were prominent at 2 months were less evident at 4 months. There was no consistent evidence suggesting that any of the components, individually or in combination, affected the number of cigarettes smoked per week among those who continued smoking.

Secondary Outcomes

Weekly Sugary Drink Consumption

There was evidence that screening/feedback (C1) initially reduced weekly sugary drink consumption both individually (IRR 0.87, CoI 0.72-1.10, POE 92.2%) and in combination with skills/know-how (C4) and mindfulness (C5). However, there was no strong evidence that these effects persisted at 4 months, and findings were attenuated in the imputed analyses. By contrast, evidence based on available-data and imputed analyses suggested that skills/know-how (C4) reduced sugary drink consumption at 4 months (IRR 0.85, CoI 0.70-1.05, POE 93.8%), and increasingly so when combined with goal-setting/planning (C2) and motivation (C3). There was also evidence that combining motivation (C3) and self-authored SMS text messages (C6) increased sugary drink consumption at 2 months (IRR 1.22, CoI 0.92-1.59, POE 91.6%); however, there was no evidence of this harmful effect at 4 months.

Weekly Portions of Candy and Snacks

There was evidence that screening/feedback (C1) reduced weekly portions of candy and snack consumption at 2 months (IRR 0.93, CoI 0.85-1.02, POE 95.1%) and 4 months (IRR 0.92, CoI 0.83-1.01, POE 96.0%). In addition, there was evidence that motivation (C3) also reduced consumption at 4 months (IRR 0.85, CoI 0.77-0.94, POE 99.9%). Combining screening/feedback (C1) and motivation (C3) led to even greater reductions at 4 months (IRR 0.77, CoI 0.67-0.89, POE >99.9%).

XSL•FO
RenderX

There was also evidence of increased effects when screening/feedback (C1) was combined with goal-setting/planning (C2), as well as with skills/know-how (C4). These findings persisted in the imputed-data analyses.

BMI

The relatively strongest evidence of effects on BMI was found for mindfulness (C5), which reduced BMI at both 2 months (mean difference −0.24, CoI −0.70 to 0.18, POE 86.4%) and 4 months (mean difference −0.23, CoI −0.69 to 0.20, POE 85.0%). There was also evidence that when mindfulness (C5) was combined with skills/know-how (C4), BMI was further reduced at both 2 months (mean difference −0.41, CoI −1.00 to 0.19, POE 90.7%) and 4 months (mean difference −0.47, CoI −1.07 to 0.14, POE 92.9%). These findings persisted in the imputed-data analyses.

Perceived Stress

The most consistent evidence of effects on perceived stress scores was found for the mindfulness (C5) component, both at 2 months (mean difference −0.17, CoI −0.37 to 0.03, POE 95.4%) and at 4 months (mean difference −0.22, CoI −0.44 to 0.0, POE 97.4%). Evidence of additional benefit at the 4-month follow-up interval was found when mindfulness was combined with screening/feedback (C1), as well as with motivation (C3). These findings persisted in the imputed analyses. There was also evidence that skills/know-how (C4) reduced perceived stress scores at 2 months (mean difference −0.25, CoI −0.45 to −0.05, POE 99.2%); however, the evidence suggested that the benefit did not persist at 4 months.

Quality of Life

There was no strong evidence of effects of the components, either individually or in combination, on QoL. Relatively, the strongest evidence was found for motivation (C3), which increased QoL scores both individually (mean difference 0.26, CoI −0.31 to 0.84, POE 81.7%) and in combination with skills/know-how (C4) (mean difference 0.54, CoI −0.28 to 1.35, POE 90.5%). These findings persisted but were attenuated in the imputed-data analyses.

Post Hoc Analyses

In Multimedia Appendix 9, we present estimates of the effects of the number of components to which participants had access on primary and secondary outcomes. The relatively strongest evidence for effects was found for increasing component count with regard to fruit and vegetable consumption and perceived stress at the 2-month follow-up. There was also evidence of reduced candy and snack consumption with increased component count at the 4-month follow-up interval.

Discussion

Primary Findings

This study used a factorial design to estimate the effectiveness of intervention components on multiple health behaviors, along with proxies for health and well-being. The results highlight novel findings regarding both beneficial and harmful effects of specific components, as well as how these components interact to affect health outcomes.

Screening/feedback (C1) was most consistent in reducing alcohol consumption, in terms of both weekly consumption and heavy episodic drinking. Further reductions in weekly consumption were observed when combined with skills/know-how (C4). These findings are consistent with extant digital alcohol interventions focused on screening and feedback, which have been shown to reduce alcohol consumption [20,39,74]. In addition, evidence from a study examining the mediated effects of an alcohol reduction app, “Drink Less,” found that tracking (ie, self-monitoring behavior) mediated the effects of the intervention [75]. However, in this study, notable evidence of harm, in terms of increased frequency of heavy episodic drinking, was found for combinations of components, particularly those including self-authored SMS text messages. We can only speculate why this was the case, but conflicting priorities in behavior change may have partially contributed to these harms. Notably, the intervention content was based on a digital alcohol intervention previously shown to be effective among people seeking help online [20]; however, when this content was combined with materials targeting other behaviors, it may not have been sufficiently tailored to individuals who typically engage in heavy episodic drinking. This highlights the need for caution when offering multiple behavior interventions to individuals whose primary focus should be on reducing heavy episodic drinking rather than improving overall lifestyle.

For smoking cessation, motivation (C3) was the most effective component, along with its interaction with other components, most notably with self-authored SMS text messages. Motivation reflects the importance of quitting smoking, suggesting that motivation to change is influenced by the perceived value of the outcome; the importance of the decision to quit is an integral part of cessation [76]. Evidence indicates that perceived benefits of cessation are associated with increased motivation to quit [77], and that positive abstinence outcome expectancies are associated with greater intentions to quit [78]. An analysis of individualized effects from a recent smoking cessation trial in Sweden showed that those who considered it highly important to quit reported better outcomes than those who considered it less important [79]. By contrast, we found that mindfulness (C5) was harmful in the short term, reducing the odds of smoking cessation at 2 months. Although speculative, many smokers use cigarettes as a stress reliever, and initiating a mindfulness program may initially heighten awareness of previously suppressed feelings, which can feel overwhelming. Previous research has shown that mindfulness interventions may be associated with increased stress during an initial phase before new coping strategies are established [80,81]. Furthermore, this study used self-directed practice, which lacks the emotional support of a group setting and guidance from an experienced teacher. Thus, individuals who smoked may have been more likely to continue smoking in the short term as a coping strategy. Notably, these harmful effects were not observed at the 4-month follow-up.

The screening/feedback component was effective in improving dietary outcomes, including increasing fruit and vegetable consumption, decreasing candy and snack consumption, and, to a lesser extent, reducing sugary drink consumption. This finding is consistent with a review by König et al [82], which

XSL•FO
RenderX

reported that feedback and self-monitoring of dietary behavior may be key to intervention effectiveness. Nonetheless, the literature has yet to establish a dose-response relationship between feedback effects and engagement. Some studies suggest that daily tracking and feedback are effective [83,84], whereas others indicate that this may be too intrusive and therefore less effective [82]; by contrast, the current trial demonstrated that weekly screening and feedback can improve outcomes. Future research is needed to establish consensus on the appropriate level of screening and feedback, potentially by examining how self-perception interacts with external feedback to influence behavior [85]. Screening/feedback also interacted with various components to further improve dietary outcomes, most prominently with motivation (C3). Goal-setting/planning (C2) was also effective in improving candy and snacks consumption, supporting findings that demonstrate action planning and goal-setting as effective approaches [45,86,87]. The impact on dietary outcomes may be explained by how components interact to influence factors necessary for change, as outlined in various sociocognitive models of health [29,88,89]. For example, motivation may increase intentions, goal-setting and planning may help translate intentions into action, and feedback on performance may further enhance efficacy. This suggests that changing dietary behaviors may be best achieved through interventions that target motivation, facilitate the translation of intentions into action, and provide performance feedback. An extant intervention incorporating these components has been shown to improve fruit and vegetable consumption [90], while the current results suggest they may also help reduce consumption of calorie-dense foods and sugary drinks.

There was no consistent evidence that the intervention components, either individually or in pairwise combinations, affected physical activity. This is not consistent with existing literature suggesting that goal-setting and planning are effective intervention approaches for increasing engagement in physical activity [91,92]. There was some evidence indicating that mindfulness was effective in increasing physical activity, which aligns with prior evidence highlighting the effectiveness of this approach [93]. Nonetheless, in the imputed analyses, there was limited evidence of effects on physical activity. There was, however, relatively stronger evidence that mindfulness reduced BMI, supporting findings across trials that highlight the efficacy of mindfulness-based interventions for weight loss [94]. Evidence for skills/know-how (C4) also suggested effectiveness in reducing BMI, consistent with studies demonstrating the efficacy of skills training for weight loss [95,96]. Furthermore, an interaction between mindfulness and skills/know-how resulted in the largest reduction in BMI. This interaction may have promoted more conscious decision-making regarding diet and physical activity. However, it is also possible that the observed effects on BMI were secondary to the components' effects on dietary behaviors, leading to reduced caloric intake.

Finally, mindfulness (C5) was found to be effective in reducing stress, both individually and in combination with other components. Mindfulness-based stress reduction is an established approach for reducing psychological stress and can be effectively delivered either in person or online [97,98]. While there was no strong evidence of effects on QoL, motivation (C3), in combination with skills/know-how (C4), showed some evidence of improving QoL. Extant interventions for improving QoL generally target individuals with chronic health conditions and are typically administered as part of integrated care interventions [99]. This study suggests that QoL may potentially be improved in the general population using digital interventions, possibly as a result of increased perceived self-efficacy arising from improvements in the 4 main health behavior domains.

Generalizability and Limitations

We used a pragmatic approach in this trial and kept barriers to participation low. This strengthens the interpretation of the estimated effects as measures of effectiveness rather than efficacy. However, this comes with a trade-off with respect to attrition. Participants may have enrolled out of curiosity, and dropping out was easy; thus, attrition was, as expected, high. We found notable associations between nonresponse and age, as well as baseline smoking and more frequent episodes of heavy drinking. However, none of these associations were moderated by component access. Furthermore, although the evidence for these associations was relatively strong, the ORs were small. Thus, while the most conservative estimates of effects may be more relevant for older individuals, less frequent heavy drinkers, and smokers, overall, the evidence for systematic attrition with respect to responders versus nonresponders was not sufficiently strong to raise major concern. By contrast, there was evidence that late responders had higher fruit and vegetable consumption at follow-up compared with early responders, which, under the assumption that late responders and nonresponders are similar, suggests a risk of attrition bias. However, this association was not observed among those with access to the screening/feedback component (C1), which showed the most promising effects on fruit and vegetable consumption. Thus, the risk of bias for this finding in particular may be considered lower. Additionally, late responders had more MVPA minutes and were less likely to have quit smoking than early responders. Overall, this suggests that our estimates for these outcomes may be biased due to attrition; however, we do not have sufficient data to determine the direction or magnitude of this bias. Although this limitation should be considered when interpreting the findings, the primary aim of this study should also be emphasized: to examine the relative contributions of different intervention components to changes in health behaviors. The overall conclusions regarding which components are relatively more promising from an effectiveness perspective are less likely to be affected by attrition in this study than the precise effect estimates themselves.

The study used self-report measures to record outcomes. To minimize recall bias, we focused on behavior in the past week (except for heavy episodic drinking and smoking, which considered the past month). This avoided asking participants to recall behavior over several months or a year, which would not only be difficult to recall accurately but also less relevant given the time frame of the intervention. Although the validity of online measures has been established [100], they remain prone to social desirability or impression management biases [101]. However, in our study, blinding reduced the likelihood that these biases differed by allocation. For smoking and

XSL•FO
RenderX

physical activity outcomes, intervention trials have used objective measures such as pedometers or salivary nicotine tests. Evidence suggests that self-reports may be less accurate than direct measures [102,103]; however, in this study, administering direct measures—particularly for smoking—was not feasible. Additionally, the study did not specify a particular type of physical activity; rather, participants reported minutes of activity. The Society for Research on Nicotine and Tobacco also recommends not using salivary tests in population studies to avoid introducing selection bias due to unequal opportunities to complete and return tests [68]. Conducting scientific studies requires decisions about which biases can be controlled and which must be accepted as limitations. In the case of self-report versus device-based measurement, there are trade-offs that are not uniform across studies. Rather than focusing solely on the accuracy of measured behavior—which is subject to bias regardless of method—future work should consider simulation studies to examine the impact of varying levels and types of bias. Estimating how measurement decisions influence effect estimates may help inform and justify study design choices.

An important consideration when assessing the generalizability of the findings from this factorial trial is the level of service utilization in a real-world implementation. If different groups in society are more or less likely to use digital interventions, the health benefits of implementation may disproportionately affect certain groups. Furthermore, the intervention materials may have differential effects across population subgroups, and not all groups may have been adequately represented in our study sample. Our estimates of effects are generalizable to the wider population insofar as it can be assumed that effects are not moderated by factors associated with nonparticipation. However, it is not possible to assess such moderation in this study, and dedicated studies are needed to examine implementation and moderation effects in greater detail.

Finally, this study targeted individuals who had sought help online; thus, participants were expected to have some motivation to change at study entry. This was supported by the high importance reported at baseline. Providing support to individuals actively seeking help is important; however, engaging those who could benefit from behavior change but are not actively seeking support remains a challenge. Future research should focus on proactive approaches to community outreach for multiple behavior support, potentially drawing on brief alcohol intervention screening strategies used in higher education settings [39,40]. This may include offering digital interventions to all individuals visiting primary health care clinics or other central community locations, such as pharmacies, grocery stores, and sports arenas. Relatedly, maintaining motivation throughout the behavior change process is critical for promoting long-term, sustained change. In this study, we provided motivational and tip-based reminders throughout the week; however, the optimal approach to sustaining motivation in digital interventions remains unclear. Strategies such as gamification [104] or habit formation [105] may be effective for some individuals. Future research should explore heterogeneity in sustained motivation and behavior change to better understand what drives long-term success.

Conclusion

This study used a factorial design to estimate the effects of a digital multiple health behavior change intervention at the component level. Studying effects at the component level provided novel insights into the building blocks of complex digital health behavior change interventions. In addition, these effects were examined across multiple behaviors. Overall, the screening/feedback component was the most effective for changing health behaviors, along with goal-setting/planning and motivation to change. Several components interacted to further improve health outcomes, most notably screening/feedback with motivation to change, screening/feedback with skills/know-how, and motivation to change with skills/know-how. By contrast, some components also had detrimental effects on certain behaviors, suggesting that not all components are appropriate for all behaviors. The results from this study contribute to the development of more effective interventions by identifying which components of complex interventions are most effective for specific behaviors.

Acknowledgments

Generative artificial intelligence was not used in this study.

Funding

This trial was funded by the Swedish Cancer Society (Cancerfonden; 20 0883 Pj, 23 2888 Pj; principal investigator: MB) and is an extension of the MoBILE (Mobile Health Multiple Lifestyle Behavior Interventions Across the Lifespan) research program, which is funded by the Swedish Research Council for Health, Working Life and Welfare (grant 2018-01410; principal investigator: ML). The funders had no role in the study planning, design, execution, analysis, or presentation of the findings.

Data Availability

The datasets used or analyzed during this study are available from the corresponding author (MB) on reasonable request. Ethical approval is required for data sharing.

Authors' Contributions

Conceptualization: PB, MB, ML

XSL•FO
RenderX

Data curation: JC, MB
Formal analysis: JC, MB
Intervention content and resources: KÅ, JB, OL, HH, PH, PB, ML, MB
Investigation: JB, KÅ, MB
Methodology: MB, ML, PH
Project administration: JB, KÅ, MB
Software: MB
Writing – original draft: JC, MB
Writing – review & editing: All authors

Conflicts of Interest

MB and PB own a private company (Alexit AB) that develops and distributes digital solutions to the public and private sectors. Alexit AB played no role in the development of the intervention, study design, data analysis, data interpretation, or writing of this report. All other authors declare no conflicts of interest.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 992 KB - jmir_v28i1e88881_app1.pdf]

Multimedia Appendix 2
Informed consent materials.
[PDF File (Adobe PDF File), 61 KB - jmir_v28i1e88881_app2.pdf]

Multimedia Appendix 3
Questionnaires.
[PDF File (Adobe PDF File), 75 KB - jmir_v28i1e88881_app3.pdf]

Multimedia Appendix 4
Intervention description and factorial conditions.
[PDF File (Adobe PDF File), 1873 KB - jmir_v28i1e88881_app4.pdf]

Multimedia Appendix 5
Attrition nonresponse analyses.
[PDF File (Adobe PDF File), 150 KB - jmir_v28i1e88881_app5.pdf]

Multimedia Appendix 6
Attrition attempts analyses.
[PDF File (Adobe PDF File), 103 KB - jmir_v28i1e88881_app6.pdf]

Multimedia Appendix 7
Estimates of effects (available data).
[PDF File (Adobe PDF File), 274 KB - jmir_v28i1e88881_app7.pdf]

Multimedia Appendix 8
Estimates of effects (imputed data).
[PDF File (Adobe PDF File), 309 KB - jmir_v28i1e88881_app8.pdf]

Multimedia Appendix 9
Estimates of effects of number of components.
[PDF File (Adobe PDF File), 94 KB - jmir_v28i1e88881_app9.pdf]

References

1. World Health Organization (WHO). Noncommunicable diseases - fact sheet. WHO. 2025. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases [accessed 2026-03-01]
2. World Health Organization (WHO). Noncommunicable Diseases - Progress Monitor 2022. Geneva, Switzerland: World Health Organization; 2022.

XSL•FO
RenderX

3. Ramstedt M, Guttormsson U. Riskbruk av alkohol i Sverige enligt nya riktlinjer. Kortfakta 04. Centralförbundet för alkohol- och narkotikaupplysning. Stockholm, Sweden: Centralförbundet för alkohol- och narkotikaupplysning; 2024 Feb 22. URL: https://www.can.se/app/uploads/2024/02/can-kortfakta-04-riskbruk-av-alkohol-i-sverige-enligt-nya-riktlinjer.pdf [accessed 2026-03-24]
4. Folkhälsan i sverige - arsrapport 2023. Folkhälsomyndigheten. Stockholm, Sweden: Folkhälsomyndigheten; 2023. URL: https://www.folkhalsomyndigheten.se/publikationer-och-material/publikationsarkiv/f/folkhalsan-i-sverige-arsrapport-2023/ [accessed 2026-03-24]
5. Alkohol ökar risken för sjukdomar och skador Internet. Folkhälsomyndigheten. 2024. URL: https://tinyurl.com/95eehuet [accessed 2026-03-01]
6. Alkohollag. Sveriges Riksdag. URL: https://www.riksdagen.se/sv/dokument-och-lagar/dokument/svensk-forfattningssamling/alkohollag-20101622_sfs-2010-1622/ [accessed 2026-03-01]
7. En hållbar och hälsosam livsmedelskonsumtion. Folkhälsomyndigheten och Livsmedelsverket. Stockholm, Sweden: Folkhälsomyndigheten och Livsmedelsverket; 2024. URL: https://www.folkhalsomyndigheten.se/publikationer-och-material/publikationsarkiv/e/en-hallbar-och-halsosam-livsmedelskonsumtion-aterredovisning-av-regeringsuppdrag/ [accessed 2026-03-24]
8. Organisation for Economic Co-operation and Development (OECD). State of Health in the EU - Sweden Country Health Profile 2023. Brussels, Belgium: European Commission; 2023.
9. Disability-adjusted life years per 100 000 population attributable to dietary risks, low physical activity, tobacco and alcohol use in the EU in 2019. European Commission. URL: https://knowledge4policy.ec.europa.eu/health-promotion-knowledge-gateway/eu-burden-non-communicable-diseases-key-risk-factors-3_en [accessed 2026-03-01]
10. Schuit AJ, van Loon AJM, Tijhuis M, Ocké M. Clustering of lifestyle risk factors in a general adult population. Prev Med 2002 Sep;35(3):219-224. [doi: 10.1006/pmed.2002.1064] [Medline: 12202063]
11. Rabel M, Laxy M, Thorand B, Peters A, Schwettmann L, Mess F. Clustering of health-related behavior patterns and demographics. results from the population-based KORA S4/F4 cohort study. Front Public Health 2018;6:387 [FREE Full text] [doi: 10.3389/fpubh.2018.00387] [Medline: 30723712]
12. Meader N, King K, Moe-Byrne T, Wright K, Graham H, Petticrew M, et al. A systematic review on the clustering and co-occurrence of multiple risk behaviours. BMC Public Health 2016 Jul 29;16:657 [FREE Full text] [doi: 10.1186/s12889-016-3373-6] [Medline: 27473458]
13. Klompstra L, Löf M, Björkelund C, Hellenius M, Kallings LV, Orho-Melander M, et al. Co-occurrence of unhealthy lifestyle behaviours in middle-aged adults: findings from the Swedish CArdioPulmonary bioImage Study (SCAPIS). Sci Rep 2024 Oct 01;14(1):22853 [FREE Full text] [doi: 10.1038/s41598-024-71092-0] [Medline: 39353999]
14. Khaw K, Wareham N, Bingham S, Welch A, Luben R, Day N. Combined impact of health behaviours and mortality in men and women: the EPIC-Norfolk prospective population study. PLoS Med 2008 Jan 08;5(1):e12 [FREE Full text] [doi: 10.1371/journal.pmed.0050012] [Medline: 18184033]
15. Kvaavik E, Batty GD, Ursin G, Huxley R, Gale CR. Influence of individual and combined health behaviors on total and cause-specific mortality in men and women: the United Kingdom health and lifestyle survey. Arch Intern Med 2010 Apr 26;170(8):711-718. [doi: 10.1001/archinternmed.2010.76] [Medline: 20421558]
16. Loef M, Walach H. The combined effects of healthy lifestyle behaviors on all cause mortality: a systematic review and meta-analysis. Prev Med 2012 Sep;55(3):163-170. [doi: 10.1016/j.ypmed.2012.06.017] [Medline: 22735042]
17. Riper H, Hoogendoorn A, Cuijpers P, Karyotaki E, Boumparis N, Mira A, et al. Effectiveness and treatment moderators of internet interventions for adult problem drinking: an individual patient data meta-analysis of 19 randomised controlled trials. PLoS Med 2018 Dec;15(12):e1002714 [FREE Full text] [doi: 10.1371/journal.pmed.1002714] [Medline: 30562347]
18. ICT usage in households and by individuals. Statistiska Centralbyrån. 2023. URL: https://www.scb.se/en/finding-statistics/statistics-by-subject-area/living-conditions/living-conditions/ict-usage-in-households-and-by-individuals/ [accessed 2026-03-01]
19. Blomqvist J, Gunnarsson KU, Bendtsen P, Bendtsen M. Effects of a text messaging smoking cessation intervention amongst online help-seekers and primary health care visitors: findings from a randomised controlled trial. BMC Med 2023 Oct 04;21(1):382 [FREE Full text] [doi: 10.1186/s12916-023-03073-5] [Medline: 37794399]
20. Bendtsen M, Åsberg K, McCambridge J. Effectiveness of a digital intervention versus alcohol information for online help-seekers in Sweden: a randomised controlled trial. BMC Med 2022 May 17;20(1):176 [FREE Full text] [doi: 10.1186/s12916-022-02374-5] [Medline: 35578276]
21. Nationella riktlinjer för prevention och behandling vid ohälsosamma levnadsvanor Internet. Socialstyrelsen. 2018. URL: https://www.socialstyrelsen.se/globalassets/sharepoint-dokument/artikelkatalog/nationella-riktlinjer/2018-6-24.pdf [accessed 2026-03-01]
22. Deitz D, Cook RF, Hersch RK, Leaf S. Heart healthy online: an innovative approach to risk reduction in the workplace. J Occup Environ Med 2014 May;56(5):547-553 [FREE Full text] [doi: 10.1097/JOM.0000000000000148] [Medline: 24806568]

XSL•FO
RenderX

23. Oenema A, Brug J, Dijkstra A, de Weerdt I, de Vries H. Efficacy and use of an internet-delivered computer-tailored lifestyle intervention, targeting saturated fat intake, physical activity and smoking cessation: a randomized controlled trial. Ann Behav Med 2008 Apr;35(2):125-135. [doi: 10.1007/s12160-008-9023-1] [Medline: 18363076]
24. Duncan MJ, Vandelanotte C, Trost SG, Rebar AL, Rogers N, Burton NW, et al. Balanced: a randomised trial examining the efficacy of two self-monitoring methods for an app-based multi-behaviour intervention to improve physical activity, sitting and sleep in adults. BMC Public Health 2016 Jul 30;16:670 [FREE Full text] [doi: 10.1186/s12889-016-3256-x] [Medline: 27473327]
25. Schulz DN, Kremers SPJ, Vandelanotte C, van Adrichem MJG, Schneider F, Candel MJJM, et al. Effects of a web-based tailored multiple-lifestyle intervention for adults: a two-year randomized controlled trial comparing sequential and simultaneous delivery modes. J Med Internet Res 2014 Jan 27;16(1):e26 [FREE Full text] [doi: 10.2196/jmir.3094] [Medline: 24472854]
26. Pham Q, Wiljer D, Cafazzo JA. Beyond the randomized controlled trial: a review of alternatives in mHealth clinical trial methods. JMIR Mhealth Uhealth 2016 Sep 09;4(3):e107 [FREE Full text] [doi: 10.2196/mhealth.5720] [Medline: 27613084]
27. Schwarzer R. Health action process approach (HAPA) as a theoretical framework to understand behavior change. AP 2016 Dec 05;30(121):119. [doi: 10.15517/ap.v30i121.23458]
28. West R, Michie S. A brief introduction to the COM-B model of behaviour and the PRIME theory of motivation. Qeios 2020 Apr 07:1-1. [doi: 10.32388/ww04e6]
29. Ajzen I. The theory of planned behavior. Organizational Behavior and Human Decision Processes 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
30. Montgomery AA, Peters TJ, Little P. Design, analysis and presentation of factorial randomised controlled trials. BMC Med Res Methodol 2003 Nov 24;3:26 [FREE Full text] [doi: 10.1186/1471-2288-3-26] [Medline: 14633287]
31. Bendtsen M, Bendtsen P, Henriksson H, Henriksson P, Müssener U, Thomas K, et al. The mobile health multiple lifestyle behavior interventions across the lifespan (MoBILE) research program: protocol for development, evaluation, and implementation. JMIR Res Protoc 2020 Apr 20;9(4):e14894 [FREE Full text] [doi: 10.2196/14894] [Medline: 32310147]
32. Åsberg K, Blomqvist J, Lundgren O, Henriksson H, Henriksson P, Bendtsen P, et al. Digital multiple health behaviour change intervention targeting online help seekers: protocol for the COACH randomised factorial trial. BMJ Open 2022 Jul 26;12(7):e061024 [FREE Full text] [doi: 10.1136/bmjopen-2022-061024] [Medline: 35882466]
33. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMC Med 2010 Mar 24;8:18 [FREE Full text] [doi: 10.1186/1741-7015-8-18] [Medline: 20334633]
34. Kahan BC, Hall SS, Beller EM, Birchenall M, Chan A, Elbourne D, et al. Reporting of factorial randomized trials: extension of the CONSORT 2010 statement. JAMA 2023 Dec 05;330(21):2106-2114. [doi: 10.1001/jama.2023.19793] [Medline: 38051324]
35. GBD 2016 Alcohol Collaborators. Alcohol use and burden for 195 countries and territories, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet 2018 Sep 22;392(10152):1015-1035 [FREE Full text] [doi: 10.1016/S0140-6736(18)31310-2] [Medline: 30146330]
36. Conner M, Norman P. Predicting and Changing Health Behaviour: Research and Practice with Social Cognition Models. London, England: McGraw-Hill Education (UK); 2015.
37. Michie S, Whittington C, Hamoudi Z, Zarnani F, Tober G, West R. Identification of behaviour change techniques to reduce excessive alcohol consumption. Addiction 2012 Aug;107(8):1431-1440. [doi: 10.1111/j.1360-0443.2012.03845.x] [Medline: 22340523]
38. Bendtsen P, McCambridge J, Bendtsen M, Karlsson N, Nilsen P. Effectiveness of a proactive mail-based alcohol internet intervention for university students: dismantling the assessment and feedback components in a randomized controlled trial. J Med Internet Res 2012 Oct 31;14(5):e142 [FREE Full text] [doi: 10.2196/jmir.2062] [Medline: 23113955]
39. McCambridge J, Bendtsen M, Karlsson N, White IR, Nilsen P, Bendtsen P. Alcohol assessment and feedback by email for university students: main findings from a randomised controlled trial. Br J Psychiatry 2013 Nov;203(5):334-340 [FREE Full text] [doi: 10.1192/bjp.bp.113.128660] [Medline: 24072758]
40. Bendtsen P, Bendtsen M, Karlsson N, White IR, McCambridge J. Online alcohol assessment and feedback for hazardous and harmful drinkers: findings from the AMADEUS-2 randomized controlled trial of routine practice in Swedish universities. J Med Internet Res 2015 Jul 09;17(7):e170 [FREE Full text] [doi: 10.2196/jmir.4020] [Medline: 26159179]
41. Michie S, Abraham C, Whittington C, McAteer J, Gupta S. Effective techniques in healthy eating and physical activity interventions: a meta-regression. Health Psychol 2009 Nov;28(6):690-701. [doi: 10.1037/a0016136] [Medline: 19916637]
42. Murray JM, Brennan SF, French DP, Patterson CC, Kee F, Hunter RF. Effectiveness of physical activity interventions in achieving behaviour change maintenance in young and middle aged adults: a systematic review and meta-analysis. Soc Sci Med 2017 Nov;192:125-133. [doi: 10.1016/j.socscimed.2017.09.021] [Medline: 28965003]
43. Knittle K, Nurmi J, Crutzen R, Hankonen N, Beattie M, Dombrowski SU. How can interventions increase motivation for physical activity? A systematic review and meta-analysis. Health Psychol Rev 2018 Sep;12(3):211-230. [doi: 10.1080/17437199.2018.1435299] [Medline: 29385950]
44. Howlett N, Trivedi D, Troop NA, Chater AM. Are physical activity interventions for healthy inactive adults effective in promoting behavior change and maintenance, and which behavior change techniques are effective? A systematic review

XSL•FO
RenderX

and meta-analysis. Transl Behav Med 2019 Jan 01;9(1):147-157 [FREE Full text] [doi: 10.1093/tbm/iby010] [Medline: 29506209]
45. Ashton LM, Sharkey T, Whatnall MC, Williams RL, Bezzina A, Aguiar EJ, et al. Effectiveness of interventions and behaviour change techniques for improving dietary intake in young adults: a systematic review and meta-analysis of RCTs. Nutrients 2019 Apr 11;11(4) [FREE Full text] [doi: 10.3390/nu11040825] [Medline: 30979065]
46. Garnett CV, Crane D, Brown J, Kaner EFS, Beyer FR, Muirhead CR, et al. Behavior change techniques used in digital behavior change interventions to reduce excessive alcohol consumption: a meta-regression. Ann Behav Med 2018 May 18;52(6):530-543 [FREE Full text] [doi: 10.1093/abm/kax029] [Medline: 29788261]
47. McCrabb S, Baker AL, Attia J, Skelton E, Twyman L, Palazzi K, et al. Internet-based programs incorporating behavior change techniques are associated with increased smoking cessation in the general population: a systematic review and meta-analysis. Ann Behav Med 2019 Feb 01;53(2):180-195. [doi: 10.1093/abm/kay026] [Medline: 29750240]
48. Ek A, Alexandrou C, Söderström E, Bergman P, Delisle Nyström C, Direito A, et al. Effectiveness of a 3-month mobile phone-based behavior change program on active transportation and physical activity in adults: randomized controlled trial. JMIR Mhealth Uhealth 2020 Jun 08;8(6):e18531 [FREE Full text] [doi: 10.2196/18531] [Medline: 32510462]
49. Newby K, Teah G, Cooke R, Li X, Brown K, Salisbury-Finch B, et al. Do automated digital health behaviour change interventions have a positive effect on self-efficacy? A systematic review and meta-analysis. Health Psychol Rev 2021 Mar;15(1):140-158. [doi: 10.1080/17437199.2019.1705873] [Medline: 31847702]
50. Müssener U, Bendtsen M, Karlsson N, White IR, McCambridge J, Bendtsen P. Effectiveness of short message service text-based smoking cessation intervention among university students: a randomized clinical trial. JAMA Intern Med 2016 Mar;176(3):321-328 [FREE Full text] [doi: 10.1001/jamainternmed.2015.8260] [Medline: 26903176]
51. Müssener U, Bendtsen M, Karlsson N, White IR, McCambridge J, Bendtsen P. SMS-based smoking cessation intervention among university students: study protocol for a randomised controlled trial (NEXit trial). Trials 2015 Apr 08;16:140 [FREE Full text] [doi: 10.1186/s13063-015-0640-2] [Medline: 25872503]
52. Thomas K, Bendtsen M, Linderoth C, Karlsson N, Bendtsen P, Müssener U. Short message service (SMS)-based intervention targeting alcohol consumption among university students: study protocol of a randomized controlled trial. Trials 2017 Apr 04;18(1):156 [FREE Full text] [doi: 10.1186/s13063-017-1898-3] [Medline: 28372563]
53. Thomas K, Müssener U, Linderoth C, Karlsson N, Bendtsen P, Bendtsen M. Effectiveness of a text messaging-based intervention targeting alcohol consumption among university students: randomized controlled trial. JMIR Mhealth Uhealth 2018 Jun 25;6(6):e146 [FREE Full text] [doi: 10.2196/mhealth.9642] [Medline: 29941417]
54. Müssener U, Bendtsen M, McCambridge J, Bendtsen P. User satisfaction with the structure and content of the NEXit intervention, a text messaging-based smoking cessation programme. BMC Public Health 2016 Nov 22;16(1):1179 [FREE Full text] [doi: 10.1186/s12889-016-3848-5] [Medline: 27876031]
55. Müssener U, Thomas K, Linderoth C, Leijon M, Bendtsen M. A text message-based intervention targeting alcohol consumption among university students: user satisfaction and acceptability study. JMIR Hum Factors 2018 Jul 10;5(3):e23 [FREE Full text] [doi: 10.2196/humanfactors.9641] [Medline: 29991469]
56. Thomas K, Linderoth C, Bendtsen M, Bendtsen P, Müssener U. Text message-based intervention targeting alcohol consumption among university students: findings from a formative development study. JMIR Mhealth Uhealth 2016 Oct 20;4(4):e119 [FREE Full text] [doi: 10.2196/mhealth.5863] [Medline: 27765732]
57. Kaner EF, Beyer FR, Garnett C, Crane D, Brown J, Muirhead C, et al. Personalised digital interventions for reducing hazardous and harmful alcohol consumption in community-dwelling populations. Cochrane Database Syst Rev 2017 Sep 25;9(9):CD011479 [FREE Full text] [doi: 10.1002/14651858.CD011479.pub2] [Medline: 28944453]
58. Black N, Johnston M, Michie S, Hartmann‐Boyce J, West R, Viechtbauer W, et al. Behaviour change techniques associated with smoking cessation in intervention and comparator groups of randomized controlled trials: a systematic review and meta‐regression. Addiction 2020 Apr 16;115(11):2008-2020. [doi: 10.1111/add.15056]
59. Crane RS, Brewer J, Feldman C, Kabat-Zinn J, Santorelli S, Williams JMG, et al. What defines mindfulness-based programs? The warp and the weft. Psychol Med 2017 Apr;47(6):990-999. [doi: 10.1017/S0033291716003317] [Medline: 28031068]
60. Creswell JD. Mindfulness interventions. Annu Rev Psychol 2017 Jan 03;68:491-516. [doi: 10.1146/annurev-psych-042716-051139] [Medline: 27687118]
61. Wong SYS, Chan JYC, Zhang D, Lee EKP, Tsoi KKF. The safety of mindfulness-based interventions: a systematic review of randomized controlled trials. Mindfulness 2018 Feb 2;9(5):1344-1357. [doi: 10.1007/s12671-018-0897-0]
62. Galante J, Friedrich C, Dawson AF, Modrego-Alarcón M, Gebbing P, Delgado-Suárez I, et al. Mindfulness-based programmes for mental health promotion in adults in nonclinical settings: a systematic review and meta-analysis of randomised controlled trials. PLoS Med 2021 Jan;18(1):e1003481 [FREE Full text] [doi: 10.1371/journal.pmed.1003481] [Medline: 33428616]
63. Neff KD, Germer CK. A pilot study and randomized controlled trial of the mindful self-compassion program. J Clin Psychol 2013 Jan;69(1):28-44. [doi: 10.1002/jclp.21923] [Medline: 23070875]
64. Crawford J, Collier ES, Bendtsen M. Individualized treatment effects of a digital alcohol intervention and their associations with participant characteristics and engagement. Alcohol Alcohol 2024 Jul 21;59(5):agae049 [FREE Full text] [doi: 10.1093/alcalc/agae049] [Medline: 39034147]

XSL•FO
RenderX

65. Rehm J. Measuring quantity, frequency, and volume of drinking. Alcohol Clin Exp Res 1998 Apr;22(2 Suppl):4S-14S. [doi: 10.1097/00000374-199802001-00002] [Medline: 9603301]
66. Shorter GW, Bray JW, Giles EL, O'Donnell AJ, Berman AH, Holloway A, et al. The variability of outcomes used in efficacy and effectiveness trials of alcohol brief interventions: a systematic review. J Stud Alcohol Drugs 2019 May;80(3):286-298. [Medline: 31250793]
67. Shorter GW, Bray JW, Heather N, Berman AH, Giles EL, Clarke M, et al. The "Outcome Reporting in Brief Intervention Trials: Alcohol" (ORBITAL) core outcome set: international consensus on outcomes to measure in efficacy and effectiveness trials of alcohol brief interventions. J Stud Alcohol Drugs 2021 Sep;82(5):638-646 [FREE Full text] [doi: 10.15288/jsad.2021.82.638] [Medline: 34546911]
68. Benowitz NL, Bernert JT, Foulds J, Hecht SS, Jacob P, Jarvis MJ, et al. Biochemical verification of tobacco use and abstinence: 2019 update. Nicotine Tob Res 2020 Jun 12;22(7):1086-1097 [FREE Full text] [doi: 10.1093/ntr/ntz132] [Medline: 31570931]
69. Vallejo MA, Vallejo-Slocker L, Fernández-Abascal EG, Mañanes G. Determining factors for stress perception assessed with the Perceived Stress Scale (PSS-4) in Spanish and other European samples. Front Psychol 2018;9:37 [FREE Full text] [doi: 10.3389/fpsyg.2018.00037] [Medline: 29434563]
70. Hays RD, Bjorner JB, Revicki DA, Spritzer KL, Cella D. Development of physical and mental health summary scores from the Patient-Reported Outcomes Measurement Information System (PROMIS) global items. Qual Life Res 2009 Sep;18(7):873-880 [FREE Full text] [doi: 10.1007/s11136-009-9496-9] [Medline: 19543809]
71. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med 2011 Feb 20;30(4):377-399. [doi: 10.1002/sim.4067] [Medline: 21225900]
72. Bendtsen M. A gentle introduction to the comparison between null hypothesis testing and bayesian analysis: reanalysis of two randomized controlled trials. J Med Internet Res 2018 Oct 24;20(10):e10873 [FREE Full text] [doi: 10.2196/10873] [Medline: 30148453]
73. Bendtsen M. Avoiding under- and overrecruitment in behavioral intervention trials using Bayesian sequential designs: tutorial. J Med Internet Res 2022 Dec 16;24(12):e40730 [FREE Full text] [doi: 10.2196/40730] [Medline: 36525297]
74. Suffoletto B, Pacella M, Huber J, Chung T. Effectiveness of text message interventions with different behavior change techniques on alcohol consumption among young adults: a five-arm randomized controlled trial. Addiction 2023 Feb;118(2):265-275. [doi: 10.1111/add.16074] [Medline: 36305694]
75. Garnett C, Dinu L, Oldham M, Perski O, Loebenberg G, Beard E, et al. Do engagement and behavioural mechanisms underpin the effectiveness of the Drink Less app? NPJ Digit Med 2024 Jun 29;7(1):174 [FREE Full text] [doi: 10.1038/s41746-024-01169-7] [Medline: 38951560]
76. DiClemente CC, Prochaska JO. Self-change and therapy change of smoking behavior: a comparison of processes of change in cessation and maintenance. Addict Behav 1982;7(2):133-142. [doi: 10.1016/0306-4603(82)90038-7] [Medline: 7102444]
77. McKee SA, O'Malley SS, Salovey P, Krishnan-Sarin S, Mazure CM. Perceived risks and benefits of smoking cessation: gender-specific predictors of motivation and treatment outcome. Addict Behav 2005 Mar;30(3):423-435. [doi: 10.1016/j.addbeh.2004.05.027] [Medline: 15718060]
78. Kaufmann A, Malloy EJ, Haaga DAF. Examining outcome expectancies for smoking vs. abstinence among adult daily smokers. Addict Behav 2020 Mar;102:106140 [FREE Full text] [doi: 10.1016/j.addbeh.2019.106140] [Medline: 31706142]
79. Crawford J, Blomqvist J, Ulfsdotter Gunnarsson K, Bendtsen P, Bendtsen M. Individualized treatment effects of a digital smoking cessation intervention among individuals looking online for help: secondary analysis of a randomized controlled trial. JMIR Mhealth Uhealth 2026 Feb 11;14:e63578 [FREE Full text] [doi: 10.2196/63578] [Medline: 41671427]
80. Baer RA, Carmody J, Hunsinger M. Weekly change in mindfulness and perceived stress in a mindfulness-based stress reduction program. J Clin Psychol 2012 Jul;68(7):755-765. [doi: 10.1002/jclp.21865] [Medline: 22623334]
81. Lundgren O, Garvin P, Kristenson M, Jonasson L, Thylén I. A journey through chaos and calmness: experiences of mindfulness training in patients with depressive symptoms after a recent coronary event - a qualitative diary content analysis. BMC Psychol 2018 Sep 13;6(1):46 [FREE Full text] [doi: 10.1186/s40359-018-0252-1] [Medline: 30213276]
82. König LM, Attig C, Franke T, Renner B. Barriers to and facilitators for using nutrition apps: systematic review and conceptual framework. JMIR Mhealth Uhealth 2021 Apr 01;9(6):e20037 [FREE Full text] [doi: 10.2196/20037] [Medline: 34254938]
83. Ambeba EJ, Ye L, Sereika SM, Styn MA, Acharya SD, Sevick MA, et al. The use of mHealth to deliver tailored messages reduces reported energy and fat intake. J Cardiovasc Nurs 2015;30(1):35-43 [FREE Full text] [doi: 10.1097/JCN.0000000000000120] [Medline: 24434827]
84. Turk MW, Elci OU, Wang J, Sereika SM, Ewing LJ, Acharya SD, et al. Self-monitoring as a mediator of weight loss in the SMART randomized clinical trial. Int J Behav Med 2013 Dec;20(4):556-561 [FREE Full text] [doi: 10.1007/s12529-012-9259-9] [Medline: 22936524]
85. Eva KW, Armson H, Holmboe E, Lockyer J, Loney E, Mann K, et al. Factors influencing responsiveness to feedback: on the interplay between fear, confidence, and reasoning processes. Adv Health Sci Educ Theory Pract 2012 Mar;17(1):15-26 [FREE Full text] [doi: 10.1007/s10459-011-9290-7] [Medline: 21468778]
86. Lara J, Evans EH, O'Brien N, Moynihan PJ, Meyer TD, Adamson AJ, et al. Association of behaviour change techniques with effectiveness of dietary interventions among adults of retirement age: a systematic review and meta-analysis of

XSL•FO
RenderX

randomised controlled trials. BMC Med 2014 Oct 07;12:177 [FREE Full text] [doi: 10.1186/s12916-014-0177-3] [Medline: 25288375]
87. Livingstone KM, Rawstorn JC, Partridge SR, Godrich SL, McNaughton SA, Hendrie GA, et al. Digital behaviour change interventions to increase vegetable intake in adults: a systematic review. Int J Behav Nutr Phys Act 2023 Mar 27;20(1):36 [FREE Full text] [doi: 10.1186/s12966-023-01439-9] [Medline: 36973716]
88. Bandura A. Social foundations of Thoughtaction: A Social Cognitive Theory. Englewood Cliffs, NJ: Prentice-Hall, Inc; 1986.
89. Prochaska JO, DiClemente CC. Stages and processes of self-change of smoking: toward an integrative model of change. Journal of Consulting and Clinical Psychology 1983;51(3):390-395. [doi: 10.1037/0022-006X.51.3.390]
90. de Vries H, Kremers SPJ, Smeets T, Brug J, Eijmael K. The effectiveness of tailored feedback and action plans in an intervention addressing multiple health behaviors. Am J Health Promot 2008;22(6):417-425. [doi: 10.4278/ajhp.22.6.417] [Medline: 18677882]
91. Garstang KR, Jackman PC, Healy LC, Cooper SB, Magistro D. What effect do goal setting interventions have on physical activity and psychological outcomes in insufficiently active adults? A systematic review and meta-analysis. J Phys Act Health 2024 Jun 01;21(6):541-553. [doi: 10.1123/jpah.2023-0340] [Medline: 38560998]
92. Rhodes R, Grant S, De BG. Planning and implementation intention interventions. In: Hagger MS, Cameron LD, Hamilton K, Hankonen N, Lintunen T, editors. The Handbook of Behavior Change. Cambridge, UK: Cambridge University Press; 2020:572.
93. Sala M, Rochefort C, Lui PP, Baldwin AS. Trait mindfulness and health behaviours: a meta-analysis. Health Psychol Rev 2020 Sep 11;14(3):345-393. [doi: 10.1080/17437199.2019.1650290] [Medline: 31362588]
94. Carrière K, Khoury B, Günak MM, Knäuper B. Mindfulness-based interventions for weight loss: a systematic review and meta-analysis. Obes Rev 2018 Feb;19(2):164-177. [doi: 10.1111/obr.12623] [Medline: 29076610]
95. Turk MW, Yang K, Hravnak M, Sereika SM, Ewing LJ, Burke LE. Randomized clinical trials of weight loss maintenance: a review. J Cardiovasc Nurs 2009;24(1):58-80 [FREE Full text] [doi: 10.1097/01.JCN.0000317471.58048.32] [Medline: 19114803]
96. Greaves CJ, Poltawski L, van Beurden SB, Price L, Taylor RS, Merrifield R, et al. Addressing the psychology of weight loss and maintenance: a feasibility study of the Skills for Weight Loss and Maintenance Weight Management Programme. Br J Health Psychol 2023 Nov;28(4):1261-1282 [FREE Full text] [doi: 10.1111/bjhp.12683] [Medline: 37487646]
97. Kriakous SA, Elliott KA, Lamers C, Owen R. The effectiveness of mindfulness-based stress reduction on the psychological functioning of healthcare professionals: a systematic review. Mindfulness (N Y) 2021;12(1):1-28 [FREE Full text] [doi: 10.1007/s12671-020-01500-9] [Medline: 32989406]
98. Bossi F, Zaninotto F, D'Arcangelo S, Lattanzi N, Malizia AP, Ricciardi E. Mindfulness-based online intervention increases well-being and decreases stress after Covid-19 lockdown. Sci Rep 2022 Apr 20;12(1):6483 [FREE Full text] [doi: 10.1038/s41598-022-10361-2] [Medline: 35444198]
99. Flanagan S, Damery S, Combes G. The effectiveness of integrated care interventions in improving patient quality of life (QoL) for patients with chronic conditions. An overview of the systematic review evidence. Health Qual Life Outcomes 2017 Sep 29;15(1):188 [FREE Full text] [doi: 10.1186/s12955-017-0765-y] [Medline: 28962570]
100. Kypri K, Wilson A, Attia J, Sheeran P, Miller P, McCambridge J. Social desirability bias in the reporting of alcohol consumption: a randomized trial. J Stud Alcohol Drugs 2016 May;77(3):526-531. [doi: 10.15288/jsad.2016.77.526] [Medline: 27172587]
101. Miller TM, Abdel-Maksoud MF, Crane LA, Marcus AC, Byers TE. Effects of social approval bias on self-reported fruit and vegetable consumption: a randomized controlled trial. Nutr J 2008 Jun 27;7:18 [FREE Full text] [doi: 10.1186/1475-2891-7-18] [Medline: 18588696]
102. Prince SA, Adamo KB, Hamel ME, Hardt J, Connor Gorber S, Tremblay M. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. Int J Behav Nutr Phys Act 2008 Nov 06;5:56 [FREE Full text] [doi: 10.1186/1479-5868-5-56] [Medline: 18990237]
103. Williams J, Rakovac I, Loyola E, Sturua L, Maglakelidze N, Gamkrelidze A, et al. A comparison of self-reported to cotinine-detected smoking status among adults in Georgia. Eur J Public Health 2020 Oct 01;30(5):1007-1012 [FREE Full text] [doi: 10.1093/eurpub/ckaa093] [Medline: 32588045]
104. El-Hilly AA, Iqbal SS, Ahmed M, Sherwani Y, Muntasir M, Siddiqui S, et al. Game on? Smoking cessation through the gamification of mHealth: a longitudinal qualitative study. JMIR Serious Games 2016 Oct 24;4(2):e18 [FREE Full text] [doi: 10.2196/games.5678] [Medline: 27777216]
105. Zhu Y, Long Y, Wang H, Lee KP, Zhang L, Wang SJ. Digital behavior change intervention designs for habit formation: systematic review. J Med Internet Res 2024 May 24;26:e54375 [FREE Full text] [doi: 10.2196/54375] [Medline: 38787601]

Abbreviations

CoI: compatibility interval
COM-B: Capability-Opportunity-Motivation-Behavior

XSL•FO
RenderX

CONSORT: Consolidated Standards of Reporting Trials
DALY: disability-adjusted life year
IRR: incidence rate ratio
MoBILE: Mobile Health Multiple Lifestyle Behavior Interventions Across the Lifespan
MVPA: moderate-to-vigorous physical activity
OR: odds ratio
POE: probability of effect
PROMIS: Patient-Reported Outcomes Measurement Information System
QoL: quality of life

Edited by A Stone; submitted 03.Dec.2025; peer-reviewed by A Eisingerich, L Cheng; comments to author 05.Jan.2026; revised version received 15.Mar.2026; accepted 16.Mar.2026; published 02.Apr.2026.

Please cite as:
Crawford J, Åsberg K, Blomqvist J, Lundgren O, Henriksson H, Henriksson P, Bendtsen P, Löf M, Bendtsen M
Effectiveness of the Components of a Digital Multiple Health Behavior Change Intervention Among Individuals Seeking Help Online (Coach): Factorial Randomized Trial
J Med Internet Res 2026;28:e88881
URL: https://www.jmir.org/2026/1/e88881
doi:10.2196/88881
PMID:

©Joel Crawford, Katarina Åsberg, Jenny Blomqvist, Oskar Lundgren, Hanna Henriksson, Pontus Henriksson, Preben Bendtsen, Marie Löf, Marcus Bendtsen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring Patient Empowerment and Health System Outcomes Associated With MyHealthNB, a Provincial Personal Health Record System: Exploratory Mixed Methods Study

Paula Voorheis[1,2,3], MSc, PhD; Stephan U Dombrowski[4], PhD; Frances Bruno[2,3], RN, MSc; Carolyn Steele Gray[2,3], PhD

[1]School of Pharmacy, University of Wisconsin–Madison, Madison, WI, United States

[2]Science of Care Institute, Sinai Health System, Toronto, ON, Canada

[3]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada

[4]Department of Kinesiology, University of New Brunswick, Fredericton, NB, Canada

Corresponding Author:
Paula Voorheis, MSc, PhD
School of Pharmacy
University of Wisconsin–Madison
Rennebohm Hall
777 Highland Ave
Madison, WI, 53705
United States
Phone: 1 6087161042
Email: paula.voorheis@wisc.edu

Abstract

Background: Personal health record systems (PHRs) have been introduced to support patient empowerment by giving individuals direct access to their personal health information and other key health system resources. MyHealthNB is a province-wide PHR in New Brunswick, Canada, that allows residents to view laboratory results, medication lists, immunization records, imaging reports, and a range of digital health resources. As PHRs continue to expand, it is essential to understand how PHRs like MyHealthNB impact outcomes related to patient empowerment.

Objective: This study uses MyHealthNB as a case example to examine empowerment-related impacts of PHRs on citizens. Building on a conceptual framework linking patient enablement, empowerment, involvement, and engagement, the study is guided by two questions: (1) What perceived impacts of PHR use emerge across enablement, empowerment, involvement, engagement, and cost-related outcomes? (2) Which impacts of PHR use are most prevalent, how are they interrelated, and what characteristics predict variation in these impacts?

Methods: An exploratory sequential mixed methods study design was used. Phase 1 involved qualitative interviews with citizens to explore perceived impacts of using MyHealthNB, which were analyzed using rapid qualitative analysis. Findings informed a Phase 2 cross-sectional survey that measured MyHealthNB users' self-reported impacts across enablement, empowerment, involvement, engagement, and cost-related outcomes. Survey data were analyzed using descriptive statistics, *t* tests, mediation analysis, and multivariable linear regressions to examine impacts, impact pathways, and impact predictors.

Results: Data from 32 interviewees and 885 survey respondents were analyzed. The qualitative analysis showed that MyHealthNB supported a progression from improved access to health information (enablement), to increased confidence (empowerment), to more active participation in health management and health care decisions (involvement and engagement). The survey analysis confirmed significant positive impacts across all 21 outcomes measured that spanned enablement, empowerment, involvement, engagement, and cost-related outcomes ($P<.05$). Mediation analyses showed that higher perceived enablement through MyHealthNB was associated with greater patient involvement and engagement, with empowerment emerging as a central linking factor. Regression models identified key predictors of MyHealthNB impacts, which included satisfaction with MyHealthNB, having a family doctor, provider support of MyHealthNB, digital literacy, and MyHealthNB use frequency.

Conclusions: Exploratory, self-reported citizen data suggest that PHRs may improve outcomes related to patient empowerment, behavior change, and health system benefits. The advantages of PHR use were most prominent when individuals had access to

XSL•FO
RenderX

primary care, received support from health care providers, and had confidence using digital technologies. To fully realize the promise of PHRs, implementers should invest in digital literacy support and strengthen primary care access and integration.

(J Med Internet Res 2026;28:e87521) doi:10.2196/87521

KEYWORDS

personal health records; patient portals; patient empowerment; behavior change; mixed methods

Introduction

The Rise of Personal Health Records

Over the past several decades, health care systems have increasingly embraced patient- and person-centered care, engaging citizens as active partners in their own health care [1-3]. Central to this shift is the recognition that informed, empowered patients are better equipped to manage their health and participate in their care [4-6]. Electronic personal health record systems (PHRs), also commonly referred to as patient portals, have been introduced to support this shift, providing patients with digital access to their personal health information [7-10]. PHRs generally offer individuals secure, web- or app-based access to information such as laboratory results, medication lists, imaging reports, and vaccination records, and additional features such as appointment scheduling, secure messaging, wait times dashboards, and health reminders [11,12].

Toward a Deeper Understanding of PHR Impacts

Growing investment in PHRs has highlighted the need to understand their impact on patient-centered outcomes. A 2021 Cochrane Review examined 10 studies with 78 to 4500 participants on the effects of providing adult patients with access to electronic health records on a range of patient, patient - provider, and health system outcomes [10]. The quality of evidence was assessed as low, and the authors suggested that future studies focus more on assessing patient empowerment and behavioral outcomes, rather than on health - related outcomes alone [10]. This review solidified the need for more comprehensive evaluations that explore why and how PHRs influence outcomes related to patient empowerment.

PHRs and Empowerment-Related Outcomes

To understand how PHRs affect empowerment-related outcomes, it is essential to first define patient empowerment and distinguish it from related concepts such as patient enablement, engagement, and involvement [13-16]. A 2022 systematic review proposed a continuum that distinguishes and connects these interrelated constructs [16]. In this continuum, "patient enablement" refers to the foundational competence through which individuals acquire the knowledge, skills, and confidence needed to understand and engage with their health and health care. Enablement fosters "patient empowerment," which is both a process that enhances a person's ability, motivation, and autonomy to make informed health decisions and a state in which individuals feel confident and capable of actively participating in their care. Empowered patients may then demonstrate 2 types of behavioral change: "patient involvement," where individuals autonomously manage their own health (eg, through self-care or self-management), independent of direct interaction with health care providers; and "patient engagement," where patients actively participate in health care behaviors in collaboration with a health care provider using shared decision-making. Figure 1 illustrates these conceptual distinctions.

Figure 1. A conceptual framework of patient empowerment and related concepts (adapted from Hickmann et al, 2022 [16]). PHR: personal health record.

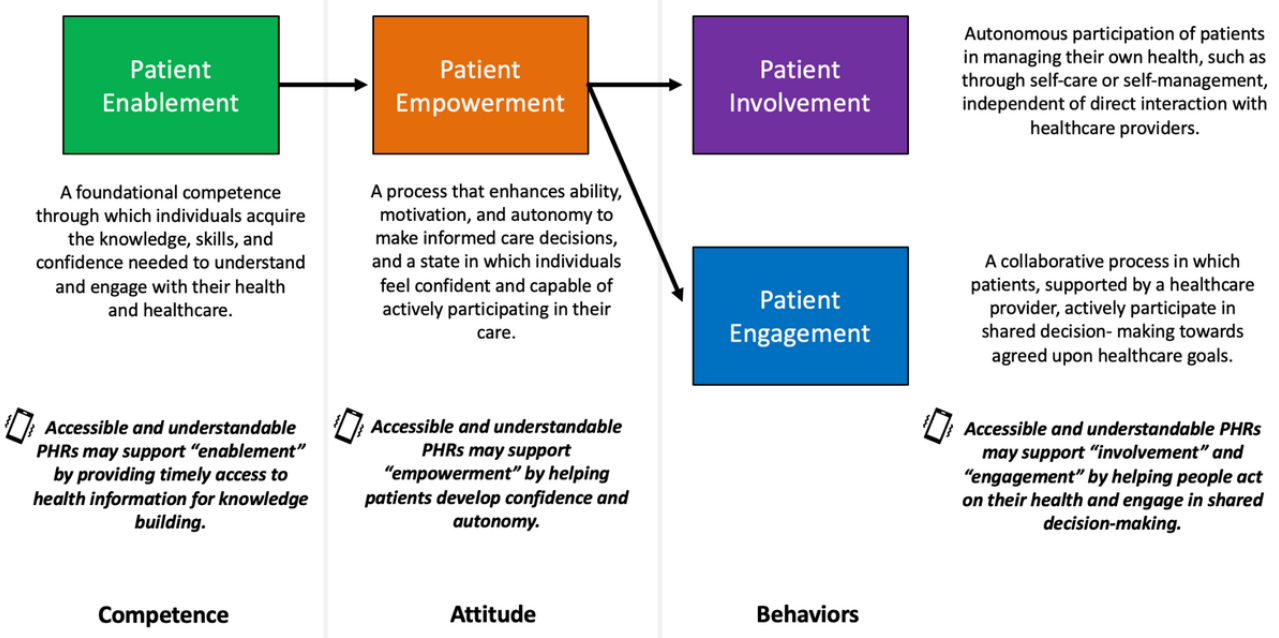

XSL•FO
RenderX

PHRs may support each stage of this progression by providing timely and accessible health information that supports individuals in building foundational knowledge (patient enablement), developing confidence and autonomy (patient empowerment), and engaging in independent health actions and collaborative decision-making (patient involvement and engagement) [7,17]. The conceptual distinctions in Figure 1 provide the foundation for future research to identify relevant measures of each construct and to better understand how PHR use may influence different aspects of these concepts.

Personal and Health System Cost Outcomes

In addition to empowerment-related outcomes, the impact of PHRs on personal and health system usage is a key area of interest in both the literature and public discourse [18-20]. From the citizen perspective, PHRs may help decrease time spent finding and organizing health information, the number of calls or visits to a health care provider to access information, and even emergency room visits [18,19]. Understanding citizens' perceptions of personal and system-related time and usage savings is important, as these experiences may also have broader implications for system costs and efficiency.

PHRs in Canada

Canada has made substantial progress in implementing PHRs, with many provinces and territories developing systems to provide citizens with timely, interoperable access to their health information. Examples of province-wide PHRs include MyHealthNB in New Brunswick, YourHealthNS in Nova Scotia, Health Booklet in Quebec, MySaskHealthRecord in Saskatchewan, MyHealth in Alberta, and Health Gateway in British Columbia. A 2024 report by Canada Health Infoway found that 89% of Canadians want access to their health records via web [21]. Public demand for PHR access creates an opportunity to explore how these systems are perceived and experienced by users, with a focus on patient-centered outcomes. In particular, understanding whether PHRs help citizens become more enabled, empowered, and active in their health and health care can support better design and implementation.

New Brunswick's PHR System: MyHealthNB

MyHealthNB is New Brunswick's province-wide PHR system, available through a mobile app or online web portal at the time of this study [22]. MyHealthNB was first made available to the public in early 2020 in response to the COVID-19 pandemic, providing citizens with access to their COVID-19 test results. Subsequent changes to the information available through MyHealthNB between its launch and the start of this study are outlined in Multimedia Appendix 1. During this study, MyHealthNB's home page connected residents to a range of health-related resources, including access to their personal health information through "MyHealthRecords." Additional resources included health system data (eg, wait times information), guidance on how to access care (eg, how-to links for finding services), information on social supports (eg, programs and services), self-scheduling tools (eg, x-ray booking and eVisitNB), and general health system resources (eg, links to regional health authority websites and mental health supports). Within "MyHealthRecords," users could view their personal laboratory results, medications, immunizations, and imaging reports. At the time of this study, MyHealthNB did not include patient–provider messaging or links to educational information related to test results, tests, or medications. Screenshots of MyHealthNB and its key features are shown in Figure 2. More detailed information about the features and content included in MyHealthNB since its initial launch until the study period can be found in Multimedia Appendix 1.

Figure 2. Images from MyHealthNB at the time of the study.

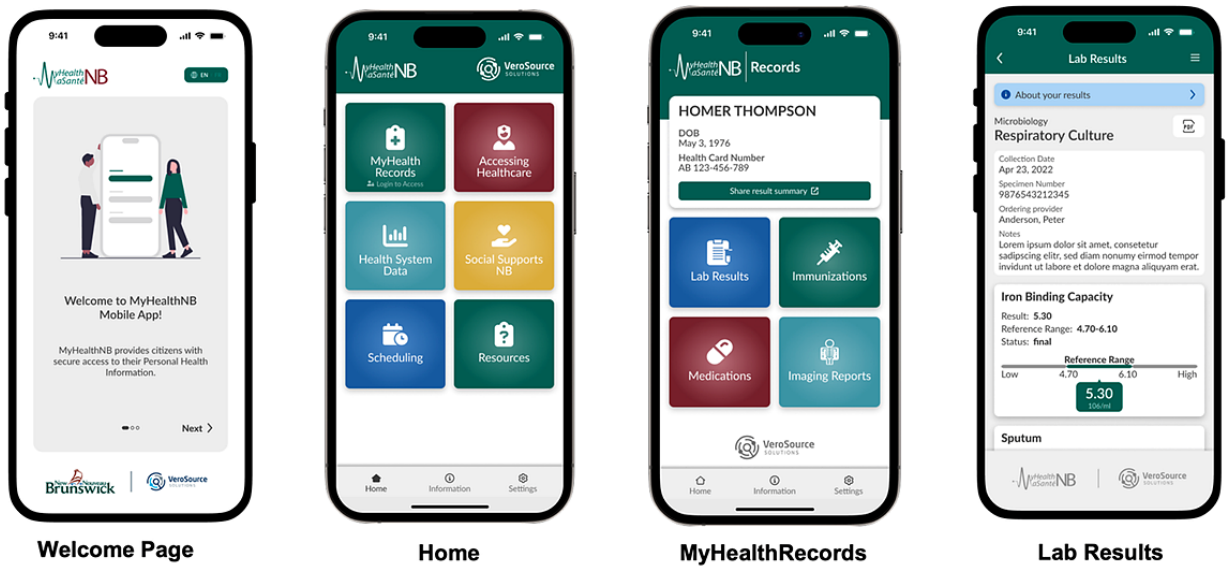

Research Questions

This study uses MyHealthNB as a case example to examine empowerment-related and behavior change impacts (eg, involvement and engagement behaviors) of PHR systems on citizens. The study is guided by the following questions:

XSL•FO
RenderX

1. What are the perceived impacts of MyHealthNB across the domains of enablement, empowerment, involvement, engagement, and cost-related outcomes (eg, personal and health system costs) according to New Brunswick citizens?
2. Which perceived impacts of MyHealthNB are most prevalent among MyHealthNB users? How are these impacts interrelated, and what characteristics (ie, demographic factors, digital literacy, health conditions, health care attachment, and patterns of MyHealthNB use) are associated with variation in these impacts?

By answering these questions, the study aims to generate actionable insights into how PHRs support patient-centered outcomes related to patient-empowerment, and to inform the future design and evaluation of PHRs.

Methods

Study Design and Approach

An exploratory sequential mixed methods design was used, beginning with qualitative data collection to identify perceived impacts, followed by a quantitative phase to broadly measure the identified impacts and their predictors [23].

This study was part of a broader developmental evaluation aimed at assessing and improving the design, implementation, and impact of MyHealthNB in New Brunswick. Developmental evaluation is an approach used during the early stages of an innovation to understand what is effective and what requires adjustment, particularly for complex and evolving programs like MyHealthNB [24,25]. The purpose of developmental evaluation is to learn and adapt the intervention as it is being delivered, ensuring evaluation methods are purposefully selected to allow for rapid transformation in innovation development [24,25]. Unlike traditional evaluation models, a developmental evaluator serves as a collaborative partner rather than an external assessor, embedding evaluation within the innovation process. In this study, the lead evaluator (PV), in support of the academic team (SUD and CSG), worked closely with VeroSource Solutions Inc. (the developer), the New Brunswick Department of Health (the implementer), and Canada Health Infoway (a national supporter of implementation) to build the evaluation strategies. One of the central objectives of the overall developmental evaluation was to understand and measure the impacts MyHealthNB was having on citizens.

Reflexivity

This study was informed by a relativist ontological stance, acknowledging that reality is subjective and shaped by individuals' lived experiences [26]. Epistemologically, the research was guided by a blend of constructivism and pragmatism [27]. Constructivism supported the co-construction of meaning between researchers and participants during the qualitative phase, while pragmatism underpinned the study's emphasis on producing actionable knowledge to inform ongoing PHR implementation. Consistent with this orientation, the study used a naturalistic and practice-oriented approach to inquiry, using descriptive and rapid analyses to generate context-sensitive insights that could support real-time learning and improvement for MyHealthNB [28].

As consistent with established developmental evaluation methods, researchers worked closely with those involved in implementing MyHealthNB. While this proximity supported a strong understanding of context while providing opportunities to provide evaluation feedback to implementers (a core aim of developmental evaluation [24,25]), steps were taken to reduce the risk of bias throughout data collection, analysis, and interpretation. All interviews were conducted by members of the research team only, without the presence of the health system or implementation partners. Qualitative coding and quantitative analyses were also completed solely by the research team, and partners were not involved in analytic decisions or in determining which findings were reported. Throughout the study, the research team engaged in ongoing reflection to discuss partnership dynamics and reflect on the positive and negative findings.

Phase 1: Qualitative Interviews

Participants and Recruitment

New Brunswick citizens were recruited in both English and French to participate in semistructured interviews exploring their experiences and perceptions of MyHealthNB and online access to health information. Recruitment was conducted through established patient and caregiver groups affiliated with the province's regional health authorities, which distributed a recruitment email to their members. The email described MyHealthNB, the study purpose, and what participation would involve. This approach was supplemented by snowball sampling to support demographic and experiential diversity in the sample, including variation in age, race, gender, geographic location, and MyHealthNB usage frequency. Participants were eligible if they were 19 years of age or older and able to complete the interview in English or French. Prior experience with MyHealthNB was not required. Participants without prior experience were instructed that they would be guided through MyHealthNB by the interviewer and would be asked to reflect on their initial perceptions of MyHealthNB and its potential role in their health care journey. Prior use was not required because the broader developmental evaluation aimed to gather citizen feedback on the design, implementation, and impact of MyHealthNB. Including both users and nonusers (eg, those who had attempted but were unable to access the platform, those who had chosen not to use it, or those who were unaware of it) enabled the study to capture both realized and potential impacts of MyHealthNB on citizens' health care experiences. The final sample size was guided by the principle of conceptual depth, with data collection concluding once sufficient richness and relevance to the research objectives had been achieved [29]. Conceptual depth was assessed through iterative review of interview data, focusing on the recurrence of key concepts across participants and the ability of the data to explain how and why MyHealthNB influenced citizens' experiences. Ongoing analytic discussions within the research team (PV and CSG) were used to assess whether additional interviews were yielding new insights and when data collection should stop.

Data Collection

Data were collected using semistructured interview guides and a preinterview survey in either French or English, depending

XSL•FO
RenderX

on participant preference [30]. Interviews were conducted via Zoom or phone and lasted approximately 45 to 60 minutes. The interview guides were piloted and refined based on input from coauthors and partners from VeroSource Solutions Inc. and the New Brunswick Department of Health. Interview guides differed based on prior use of MyHealthNB. For participants with prior MyHealthNB experience, interviews explored past experiences navigating health information within the New Brunswick health care system, how participants became aware of and adopted MyHealthNB, their current engagement with the platform, perceived impacts on their care, likes and dislikes, and recommendations for future development. During these interviews, the interviewer could share their screen to display a mock version of MyHealthNB to facilitate discussion of specific features. For participants without prior MyHealthNB experience, interviews focused on past experiences navigating health information within the New Brunswick health care system, awareness of MyHealthNB, and initial interests or concerns about the platform. The interviewer then shared their screen to present a mock version of MyHealthNB, and participants were asked to discuss their perceptions of the platform and the potential impacts on their care if they were to adopt different features. Participants were subsequently asked about likes, dislikes, and recommendations for the future of MyHealthNB. All interviews were audio-recorded and transcribed with an accuracy check using Zoom's automated recording and transcription feature. Translation was performed with the support of a fluent bilingual French–English speaking team member (FB), and using Microsoft Office Translator in Microsoft Word. The preinterview survey gathered information on participants' demographics and MyHealthNB usage, the results of which are presented later.

Data Analysis

Descriptive statistics were generated to characterize the interview sample. The interview data were analyzed using rapid qualitative analysis in Microsoft Word [31-33]. Rapid qualitative analysis is an action-oriented approach designed to efficiently generate insights that inform decision-making in dynamic or time-sensitive contexts [31-33]. Following established steps for rapid qualitative analysis, the lead researcher (PV) reviewed the interview transcripts and audio recordings to create summary documents that captured key points from each interview. These summaries were structured according to predefined domains aligned with the objectives of the broader developmental evaluation, which focused on identifying design-related facilitators and barriers, implementation-related facilitators and barriers, and the perceived impacts of MyHealthNB. This study focused on perceived impacts. When relevant, illustrative quotes were included alongside summary points. Summary data from each interview were compiled into a larger matrix to allow for cross-interviewee comparison of perceived impacts. Within the matrix, points related to enablement, empowerment, involvement, and engagement were labeled according to previously outlined definitions [16]. The matrix was used to organize the data into a concise, accessible format that could support descriptive interpretation and application. The matrix template was reviewed by CSG.

Phase 2: Integration of Qualitative and Quantitative Steps

The preliminary findings from the qualitative interviews in Phase 1 were used to inform both the sampling strategy and instrument design for the quantitative survey.

Regarding the sampling strategy, interviewees highlighted several population groups that may experience unique impacts from MyHealthNB, including individuals enrolled in school, older adults with caregivers, newcomers to New Brunswick, those living with chronic conditions (eg, diabetes and cancer), individuals involved in community or recreational groups, persons with disabilities, and members of racial minorities or Indigenous communities. In response, PV conducted targeted outreach to publicly available contacts associated with organizations representing these groups to enhance representation in the quantitative sample.

Regarding the instrument design, qualitative insights were used to ensure that survey questions reflected the diverse ways in which MyHealthNB may have impacted participants and identified contextual factors that shaped those experiences. Survey items were developed to capture self-perceived impacts across patient enablement, empowerment, involvement, engagement, and cost-related outcomes. Interview findings also informed the inclusion of additional predictors such as whether respondents were new to New Brunswick, their education level, access to a family doctor, presence of a chronic condition, extent of MyHealthNB use (eg, use of homepage), and whether their providers encouraged use of MyHealthNB. A full description of survey measures is included below and in Multimedia Appendix 1. The alignment between qualitative insights and survey design ensured that the quantitative phase was grounded in the realities of patient experience.

Although validated instruments such as the Patient Activation Measure, Health Empowerment Scale [34] were considered during the design of the quantitative phase, these tools were ultimately not adopted because they were fully responsive to the specific impacts MyHealthNB was intended to have. Given the developmental evaluation focus, survey items were designed to capture PHR-specific perceived impacts across domains of enablement, empowerment, involvement, engagement, and system costs. The final survey was developed iteratively based on qualitative findings, established conceptual distinctions in the academic literature [10,16,35], and measures used in similar evaluations of PHRs in other Canadian jurisdictions [19,20].

Phase 3: Quantitative Cross-Sectional Survey

Participants and Recruitment

Individuals were eligible if they resided in New Brunswick and were 19 years of age or older. The recruitment strategy aimed to obtain a broadly representative sample of the provincial population. Outreach efforts were guided by patient and caregiver advisory groups and informed by insights from the qualitative interviews. Targeted email invitations were sent to organizations representing priority groups, including students, older adults and their caregivers, newcomers to New Brunswick, individuals with chronic health conditions, community and recreational group participants, persons with disabilities, and

XSL•FO
RenderX

members of racialized or Indigenous communities. These organizations distributed the survey invitation through their existing mailing lists and communication channels. Additionally, a survey invitation was posted on the main landing page of the MyHealthNB website. Participation was voluntary and self-selected, which may have resulted in greater participation from individuals who were more digitally engaged or had more notable experiences with MyHealthNB. All recruitment materials and the survey were available in English and French.

Data Collection

An online survey administered through Qualtrics was used to collect data. As previously outlined, survey measures were developed specifically for this developmental evaluation and, therefore, were not subjected to formal psychometric testing. Construct validity was not assessed using methods such as factor analysis or hypothesis testing. Instead, measures were designed to maximize content relevance and interpretability through qualitative grounding, theoretical alignment, and codevelopment with system partners. Many constructs were assessed using single-item measures, which limited the ability to assess internal consistency and may increase susceptibility to measurement error or response bias. Quantitative findings should be interpreted as exploratory indicators of perceived impact rather than precise estimates of underlying constructs. The implications of these measurement limitations for interpretation are discussed further in the Limitations section.

To provide feedback on the impact of MyHealthNB, respondents were required to self-identify as users of MyHealthNB. MyHealthNB users were defined as those who reported using the MyHealthNB to access their personal health information, to help someone else access their personal health information (eg, laboratory results, immunizations, medication lists, and imaging reports), or to access general information about the health system (eg, wait time dashboards) or other health system links (eg, x-ray scheduling or eVisitNB). Respondents who indicated they had not used MyHealthNB were classified as nonusers and were directed to a different survey module that did not include questions about the impacts of MyHealthNB. The measures used to assess self-reported impacts of MyHealthNB and key descriptive variables are detailed below.

Dependent Variables: Impact Measures

Self-reported impact measures captured in the survey aligned with the domains of patient enablement, patient empowerment, patient involvement, patient engagement, and personal and health system costs. These measures were informed by conceptual distinctions and findings provided in the academic literature [10,16,35], surveys conducted in other Canadian jurisdictions [19,20], and descriptions and examples provided by interviewees from the qualitative phase. All measures were codeveloped, piloted, and finalized with partners at the New Brunswick Department of Health and VeroSource Solutions, ensuring both conceptual alignment and practical relevance to PHR implementation. Table 1 outlines the measures and scales used.

XSL•FO
RenderX

Table 1. Survey measures for self-reported impacts from MyHealthNB.

Domain	Question prompt and scale	Measures
Patient enablement	Prompt of "How has having access to your personal health records through MyHealthNB impacted the following?" with a 5-Point Likert Scale including, "Significantly Improved (5)", "Somewhat Improved (4)", "No Change (3)", "Somewhat Worsened (2)", and "Significantly Worsened (1)"	1. Ease in accessing your health information (eg, ability to find a specific health record) 2. Awareness of your health status (eg, awareness of your results, vaccination status, or medication history) 3. Understanding of your health (eg, understanding why you are experiencing a certain issue or why you need a specific treatment) 4. Control over your health information (eg, the ability to identify errors in your record or share your health information with whom you want to)
Patient empowerment	Prompt of "Rate your level of agreement with the following statement" with a 5-Point Likert Scale including, "Strongly agree (5)", "Somewhat agree (4)", "Neither agree nor disagree (3)", "Somewhat disagree (2)", and "Strongly disagree (1)"	1. "MyHealthNB has empowered me to take a more active role in my health and health care by giving me greater access, control, and understanding of my health information." 2. "MyHealthNB has raised concerns, frustration, and stress for me because of the ongoing challenges I face in accessing, managing, and understanding my health information."
Patient involvement	Prompt of "How has having access to your personal health records through MyHealthNB impacted the following?" with a 5-Point Likert Scale including, "Significantly Improved (5)", "Somewhat Improved (4)", "No Change (3)", "Somewhat Worsened (2)", and "Significantly Worsened (1)"	1. Engagement in your health care (eg, doing preventative screenings, scheduling follow up appointments, refilling prescriptions, managing an ongoing issue) 2. Engagement in your health (eg, making healthier decisions, setting goals related to your health)
Patient engagement	Prompt of "How has having access to your personal health records through MyHealthNB impacted the following?" with a 5-Point Likert Scale including, "Significantly Improved (5)", "Somewhat Improved (4)", "No Change (3)", "Somewhat Worsened (2)", and "Significantly Worsened (1)"	1. Preparedness for visits with a health care provider (eg, having a plan for a concern, result, or treatment you want to discuss) 2. Communication with a health care provider (eg, being able to discuss a concern, result, or treatment with a provider) 3. Ability to be a more active participant in health care decision-making (eg, collaborating with a provider to choose the best treatment or next step) 4. Preparedness for a health emergency or unexpected health event (eg, ability to quickly access what you need to share during an emergency) 5. Early detection and action (eg, more timely seeking and addressing an abnormality) 6. Ability to avoid taking a test or treatment you don't need or want (eg, avoiding a duplicate test or treatment) 7. Efficiency of your health care appointments (eg, having more time to talk about what matters to you or getting treatment faster because you know your medical history) 8. Coordination of your health care (eg, ensuring all of your health care providers are on the same page) 9. Overall quality of your health care appointments (eg, satisfaction with your overall experience with visiting a provider)
Personal and health system cost	Prompt of "How has having access to your personal health records through MyHealthNB impacted the following?" with a 5-Point Likert Scale including, "Significantly Decreased (5)", "Somewhat Decreased (4)", "No Change (3)", "Somewhat Increased (2)", and "Significantly Increased (1)"	1. Time spent finding and organizing your health information (eg, time spent retrieving your health results from different providers or clinics) 2. Number of calls you have to make to a health care provider to get health information (eg, information about test results, vaccine records, medication information, and treatment history) 3. Number of times you travel to a health care provider to get health information (eg, travelling to receive a result) 4. Number of times you visit the emergency room

Independent Variables: Other Measures

Other self-reported measures used in the study assessed respondent demographics (ie, age, race, gender, location in New Brunswick, residence length in New Brunswick, language, and education level), digital abilities (ie, digital literacy and digital health navigation ease), health status (ie, chronic condition status), health care attachment (ie, family doctor attachment status), and MyHealthNB usage (ie, MyHealthNB access format, satisfaction, usage frequency, home page usage, and provider support satisfaction). These are fully defined and outlined in Multimedia Appendix 1.

Data Analysis

Data were analyzed using RStudio (RStudio, Inc) and Microsoft Excel. Survey responses were analyzed without weighting, as the primary objective was to examine perceived impacts among users rather than to generate population-representative estimates. Implications of self-selection and underrepresentation of certain groups are addressed in the Limitations section. Analysis

XSL•FO
RenderX

proceeded in 4 sequential phases. In the first phase, descriptive statistics were generated to characterize the survey sample. Survey sample demographics were compared with provincial benchmarks from the 2021 Canadian Census and 2024 New Brunswick Health Council to assess representativeness. In the second phase, descriptive and inferential analyses were conducted to assess perceived impacts of MyHealthNB across 21 self-reported outcomes related to enablement, empowerment, involvement, engagement, and costs. For each outcome, means, SDs, 95% CIs, and one-sample *t* tests against the neutral midpoint of 3 were calculated. Likert-scale distributions were also summarized using frequencies and proportions across all 5 response options to contextualize patterns of perceived improvement, neutrality, or decline. In the third phase, regression-based mediation analyses were conducted to explore whether associations between patient enablement and downstream outcomes were consistent with hypothesized pathways involving empowerment and stress. Composite scores were calculated for enablement, involvement, engagement, and cost-related outcomes by averaging theoretically aligned survey items, and internal consistency was assessed. Empowerment and stress were measured using single-item measures. For each proposed mediator (empowerment and stress), regression models examined associations among enablement, the mediator, and each outcome (engagement, involvement, and cost-related outcomes). Outcomes were first regressed on enablement alone and then on both enablement and the proposed mediator to assess whether the mediator was associated with the outcome and whether both variables remained significant in combined models. Given the cross-sectional design, mediation analyses were intended to explore theory-consistent patterns of association rather than to test formal causal mediation or estimate indirect effects. Patterns in which enablement predicted both the mediator and the outcome, and both remained significant in combined models, were interpreted as consistent with partial mediation. In the fourth phase, multivariable linear regression analyses were conducted to identify predictors of variation in the 21 different impact scores. Independent variables included the 16 other measures outlined in Multimedia Appendix 1. Each impact dependent variable was modeled separately, and standardized beta coefficients, *P* values, and adjusted R^2 values were calculated. Multicollinearity was assessed by calculating variance inflation factors (VIFs) for the regression model examining MyHealthNB's impact on ease of accessing health information, which included the full set of predictors used across analyses. Each predictor was regressed on all remaining predictors, and VIFs were calculated using $VIF = 1/(1-R^2)$. The regression analyses were not stratified by MyHealthNB use type (eg, using MyHealthNB to access one's own personal health information, using MyHealthNB to support others in accessing their health information, or using MyHealthNB to seek general health system information) because these use types reflected overlapping, task-based behaviors rather than stable user groups. Many respondents reported multiple, dynamic forms of use. Additionally, although the conceptual framework suggests directional relationships among enablement, empowerment, involvement, and engagement, the cross-sectional design does not support assumptions of temporal ordering or causal inference. We therefore did not conduct mediation analyses and instead focused on describing associations between predictors and perceived impacts. Full descriptive and inferential statistical outputs for all phases of the analysis are provided in Multimedia Appendix 1.

Ethical Considerations

The study was reviewed and approved by the Research Ethics Boards at the University of New Brunswick, Sinai Health System, and University of Toronto (research ethics board 2023-169). Interview participants provided written consent using email or verbal consent at the start of the Zoom interview. Interview participants were remunerated with a gift card worth US $30 and agreed that at no point in any publications or oral proceedings would they be identified. Survey participants provided consent online using Qualtrics. At the end of the survey, participants were asked if they wanted to enter their name into a raffle to receive 1 of 10 gift cards. If they said "yes," they were taken to another page where they were asked to enter their name and email address. Names and email addresses were in no way linked to their previous survey responses.

Results

Participants

In the qualitative phase, a total of 32 participants were interviewed between April and August 2024 to discuss the perceived impacts of MyHealthNB. Among interviewees, 27 out of 32 (84%) reported using MyHealthNB to access their own personal health information, 17 out of 32 (53%) reported using MyHealthNB to help someone else access health information, and 3 out of 32 (9%) reported no prior use of MyHealthNB. In the quantitative phase, 1160 New Brunswick residents completed the survey between November 27, 2024, and January 11, 2025. Of the 969 respondents who answered the question about prior use, 885 (91%) reported having used MyHealthNB and are included in the analyses below. Survey respondents could indicate multiple reasons for use: 771 out of 885 (87%) reported using MyHealthNB to access their own personal health information, 274 out of 885 (31%) reported using MyHealthNB to help someone else access health information, and 233 out of 885 (26%) reported using MyHealthNB to find general health system information (eg, wait times dashboards) or access related services (eg, x-ray scheduling or eVisitNB). The number of respondents included in each impact analysis is reported alongside the corresponding statistical outputs in Multimedia Appendix 1.

Table 2 presents the participant characteristics for both the interview and survey samples, alongside comparable population data from New Brunswick. The interview sample included a diverse mix of ages, races, genders, geographic locations, and languages. Males were underrepresented in the interview sample, though several female participants spoke on behalf of their male partners. The interview sample showed a similar distribution to the survey sample. The survey sample was broadly representative of the New Brunswick population in terms of age, race, language, location, family doctor attachment status, and chronic health conditions, but underrepresented males and individuals with lower education levels.

XSL•FO
RenderX

Table 2. Interview and survey participant characteristics.

Characteristics	Interview sample (n=32), n (%)	Survey sample (n=885), n (%)	Population data, %
Demographics			
Age (years)			
19-34	9 (28)	197 (22)	21
35-44	4 (13)	158 (18)	15
45-54	5 (16)	156 (18)	16
55-64	4 (13)	161 (18)	20
65-74	9 (28)	158 (18)	17
75-84	3 (9)	45 (5)	8
85+	0 (0)	2 (<1)	3
No response	0 (0)	8 (1)	NR[a]
Race			
Racial majority (White)	28 (88)	770 (87)	94
Racial minority	2 (6)	89 (10)	6
Prefer not to say	1 (3)	14 (2)	NR
No response	1 (3)	12 (1)	NR
Gender			
Male	8 (25)	217 (25)	49
Female	23 (72)	651 (74)	51
Trans or nonbinary	0 (0)	3 (<1)	NR
Prefer not to say	1 (3)	2 (<1)	NR
No response or other	0 (0)	12 (1)	NR
Location			
Major health zone (zone 1-3)	29 (91)	690 (77)	76
Rural health zone (zone 4-7)	3 (9)	137 (15)	24
No response or other	0 (0)	58 (7)	NR
Language			
English only	20 (63)	595 (67)	58
French and English	12 (38)	238 (27)	34
French only	0 (0)	36 (4)	8
No response or other	0 (0)	16 (2)	NR
Residence length			
0-3 years	NR	37 (4)	NR
3-10 years	NR	115 (13)	NR
10-20 years	NR	98 (11)	NR
20+ years	NR	619 (70)	NR
No response or other	NR	16 (2)	NR
Education level			
No advanced degree/certificate	NR	198 (22)	49
Advanced degree/certificate	NR	669 (76)	51
No response or other	NR	18 (2)	NR
Digital literacy			
Web confidence			

XSL•FO
RenderX

Characteristics	Interview sample (n=32), n (%)	Survey sample (n=885), n (%)	Population data, %
Very confident	27 (84)	587 (66)	NR
Somewhat confident	4 (13)	178 (20)	NR
Not very confident	1 (3)	15 (2)	NR
Not confident at all	0 (0)	5 (1)	NR
No response or other	0 (0)	6 (1)	NR
Digital health navigation			
Very easy	7 (22)	200 (23)	NR
Somewhat easy	17 (53)	226 (22)	NR
Somewhat challenging	7 (22)	169 (19)	NR
Very challenging	1 (3)	47 (5)	NR
Have not attempted	0 (0)	115 (13)	NR
No response or other	0 (0)	138 (16)	NR
Health condition			
Ongoing health condition			
Yes	NR	663 (75)	68
No	NR	197 (22)	32
No response or other	NR	15 (2)	NR
Health care attachment			
Family doctor attachment status			
Yes	NR	642 (73)	76
No	NR	202 (23)	24
No response or other	NR	41 (5)	__[b]
MyHealthNB usage			
Access format			
Web only	12 (38)	458 (52)	NR
App only	6 (19)	120 (14)	NR
Web and App	11 (34)	283 (32)	NR
No response	3 (9)	24 (3)	NR
Satisfaction score (Max=4), mean score (n)	3.41 (29)	3.42 (831)	NR
Usage frequency			
Low	17 (53)	487 (55)	NR
Medium	12 (38)	266 (30)	NR
High	0 (0)	80 (9)	NR
No response	3 (9)	52 (6)	NR
Home page use			
Yes	NR	399 (45)	NR
No	NR	413 (47)	NR
No response	NR	73 (8)	NR
Provider support			
Very satisfied	NR	182 (21)	NR
Somewhat satisfied	NR	148 (17)	NR
Neutral	NR	363 (41)	NR

XSL•FO
RenderX

Characteristics	Interview sample (n=32), n (%)	Survey sample (n=885), n (%)	Population data, %
Somewhat dissatisfied	NR	68 (8)	NR
Very dissatisfied	NR	63 (7)	NR
No response	NR	61 (7)	NR

[a]NR: not recorded.

[b]Not available.

Qualitative Results

Overview

Interviews with New Brunswick citizens revealed that MyHealthNB was seen as more than a tool for viewing results and personal health history; it served as a catalyst for behavior change and relational improvements in health care. Interview participants traced a clear progression from gaining access to health information (ie, enablement), to feeling more confident and in control (ie, empowerment), to actively participating in care (ie, engagement and involvement), and noted the emotional and cognitive complexity that accompanied this shift. The sections below outline the foundational ways MyHealthNB impacted participants.

From Enablement to Action: How Patients Used MyHealthNB to Manage Their Health

Interview participants described enablement as MyHealthNB's foundational value, providing timely, understandable access to personal health information:

> *I had all the information in the palm of my hand, whereas before, I was relying on my doctor and hoping that you would talk to a doctor because doctors as so busy. It's so nice to be able to get the information a lot quicker…* [Interviewee #9]

> *I used to get pretty archaic looking lab print outs emailed to me. I had to work hard to figure it out and understand it. So, the difference now is that it's much more intuitive to understand what you're looking at and make sense of it.* [Interviewee #11]

This enablement often initiated a chain reaction toward empowerment and behavior change, particularly in how patients prepared for appointments:

> *For me [MyHealthNB] is awesome, because when I go to the hospital [for cancer treatment] I am able to check my blood results and compare my results from last time. If my blood results are fine and they say to go ahead, I can compare and say I also should be good to go ahead this time. I like to understand the lab results… so it is kind of educational as well at the same time… because you can go further and do some research and talk to your doctor…* [Interviewee #30]

Empowerment Through MyHealthNB: Balancing Feelings of Confidence and Concern

Interview participants consistently described MyHealthNB as a powerful facilitator of patient empowerment, particularly through its ability to enhance feelings of control and peace of mind. For many, access to timely health information allowed them to approach medical appointments with greater confidence and clarity, reducing feelings of helplessness:

> *One big thing is to have that information sooner, which gives you peace of mind. Sometimes when being tested for something your mind starts thinking crazy stuff. So peace of mind is one big thing.* [Interviewee #9]

> *[With MyHeathNB] I could be emotionally and informationally prepared for the appointments to ask specific questions. For people with cancer the wait for results is crucial. Some people can go through depression. For me, my morale is very solid, but I can understand how waiting is hell.* [Interviewee #30]

> *I feel like I am in control, which I should be. [MyHealthNB] gives me peace of mind. I am not worried about my health as much and, as best as I can, I am in control… I feel like I can understand everything, and I know if it's not my doctor who can deal with it, my dietician can deal with it. It gives me the information I need.* [Interviewee #1]

At the same time, many participants emphasized that empowerment through information from MyHealthNB was not without emotional and cognitive complexity. While MyHealthNB provided reassurance and autonomy, it also introduced new challenges that could generate confusion, worry, or stress:

> *Yeah, for me, it's like, if it's not in the green, then that worries me. I'm like, 'Oh, am I deficient in XYZ?' I just don't want to look at [MyHealthNB] anymore. I started to go into that kind of spiral of looking it up online and trying to figure out what everything means.* [Interviewee #29]

> *My poor mom, the condition we have got affects every body system, so she's got so many things going on. She doesn't know which doctor she's supposed to ask which question to, and I don't want to suggest it to her these days because she very anxious normally. So to see all these results when things get out of whack makes her very anxious.* [Interviewee #15]

Although viewing results in MyHealthNB sometimes led to worry and uncertainty, many participants did not see these feelings as contradictory to empowerment. Rather, they viewed them as an inherent part of the empowerment process, recognizing that emotional engagement, both positive and negative, is often intertwined with taking a more active role in one's health:

> *I recently saw something that was usually in the normal range was out of range, and I googled it to*

XSL•FO
RenderX

> *find out it had to do with my liver. This was something new. My doctor has a nurse who speaks to you on the phone first, and I told her about this. Immediately the doctor came on and talked to me about it. I don't know if they ever would have looked at it if I had not seen it. Being able to go on MyHealthNB meant that I could advocate for myself.* [Interviewee #12]

> *It may create some anxiety in people if they're not used to [seeing their records], and if it's a new problem. But people with chronic problems, when they see their doc, and they've been getting tests for years, they know what's good, they know what's bad.* [Interviewee #13]

Patient Involvement: MyHealthNB Allows Patients to Take Control of Their Own Health

Participants described how MyHealthNB supported involvement in their own health in a context where primary care follow-up and communication of test results are often limited. By providing direct access to laboratory results, MyHealthNB allowed patients to review their information, track changes over time, and take action between appointments:

> *I went in [MyHealthNB] and checked my blood test and I saw my blood sugar was a little bit high… So, I said okay, give yourself 6 months and get the sugar out of your diet and try to make a few lifestyle changes to see if you can bring it around…You don't want it to get too far out of line and I certainly don't want to have to take any more medication.* [Interviewee #31]

> *I am a Type 1 Diabetic, and I am very involved in my own care. I see [MyHealthNB] being a 'game changer'. And I hate that terminology, but this is really a game changer for patients to take control of their own health and to learn about their own bodies and what they need to do to improve their results.* [Interviewee #1]

Patient Engagement: MyHealthNB Can Drive Impact Through Enhancing Relational Care

Interview participants emphasized that MyHealthNB helped them engage more actively and collaboratively with their health care providers. Access to health information allowed patients to prepare questions, guide discussions, and engage more collaboratively during clinical encounters when providers were receptive to patient involvement:

> *What I tend to do is highlight [my medical reports] and comment on the things I want answers to. Then I talk to my clinical nurse and ask, "okay, this is that?". I feel very fortunate. If I didn't have the results to download [through MyHealthNB], I wouldn't be able to ask those questions.* [Interviewee #14]

> *I hope physicians understand that [MyHealthNB] will make it easier on them because now we are partners. I've got the information, they've got the information... If my blood tests are bad because of my diet in the last year… do I know enough how to fix my own diet? If not, can they send me to a dietician?* [Interviewee #1]

While MyHealthNB enabled many participants to take greater initiative in clinical conversations, their ability to fully engage often depended on how providers responded. Supportive providers reinforced the shift toward partnership and shared decision-making, whereas dismissive or resistant attitudes risked undermining it:

> *I met [my doctor] in the hospital when I was visiting a friend… I told him about MyHealthNB and its great people can advocate for themselves and he said, 'It's not really good... people going to Dr. Google… you know you'd misinterpret your results". He wasn't very impressed by us having our records in MyHealthNB.* [Interviewee #16]

> *I thought that I should talk to my doctor about [MyHealthNB] and she was not very receptive of it and said 'oh you don't need that'. She didn't tell me about it and it was not something of interest to her… We don't have that close of a relationship and I know they are very taxed and their time is precious.* [Interviewee #28]

System Usage Impacts: Patients Perceive Reduced Burden, Not Increased Demand

While some interviewees acknowledged that they had heard of providers voicing concerns that MyHealthNB could increase the health care system burden through more calls, visits, or anxious follow-ups, many offered a different perspective. Access to information through MyHealthNB often reduces their need to contact providers or visit clinics, streamlining routine tasks and minimizing unnecessary system use:

> *MyHealthNB should save the pharmacy and healthcare providers a lot of work. I recently looked up my COVID vaccination data because I am a caregiver for my mom, a lady with Alzheimer's, and myself. We all go together to get our COVID shots. I would look up on MyHealthNB when our last one was, and it showed me I have to wait until the 22nd of April. It saves me calling the pharmacy, going to the pharmacy, asking the pharmacist about it. So it saves a huge thing on their shoulders.* [Interviewee #1]

Overall Impact of MyHealthNB

Collectively, the qualitative findings suggest that MyHealthNB has the potential to empower citizens to take more proactive and informed roles in their health and health care. While complexity exists around interpreting new information and health system support constraints, most participants felt that the benefits outweighed the risks. When used in conjunction with supportive health care providers, MyHealthNB was perceived as a lever for improving patient preparedness, engagement, and health system efficiency. Multimedia Appendix 1 provides a detailed list of interviewee quotes, organized by specific impacts across the progression from enablement to empowerment, involvement, engagement, and cost-related outcomes.

XSL•FO
RenderX

Quantitative Results

Overview

Building on the findings of the qualitative results, the quantitative results showed similar trends regarding the impact of MyHealthNB across enablement, empowerment, involvement, engagement, and cost-related outcomes.

Impacts of MyHealthNB on Enablement, Empowerment, Involvement, Engagement, and Costs

As illustrated in Figure 3, MyHealthNB was associated with statistically significant positive perceived impacts across all 5 domains of enablement, empowerment, involvement, engagement, and system costs, spanning all 21 measured outcomes (P<.05). The only outcome with a mean score below neutral was agreement that MyHealthNB was stress-producing, indicating that participants generally disagreed with this statement. A full description of the impact results, including descriptive statistics, is included in Multimedia Appendix 1.

Figure 3. Self-reported impacts of MyHealthNB on 21 outcomes.

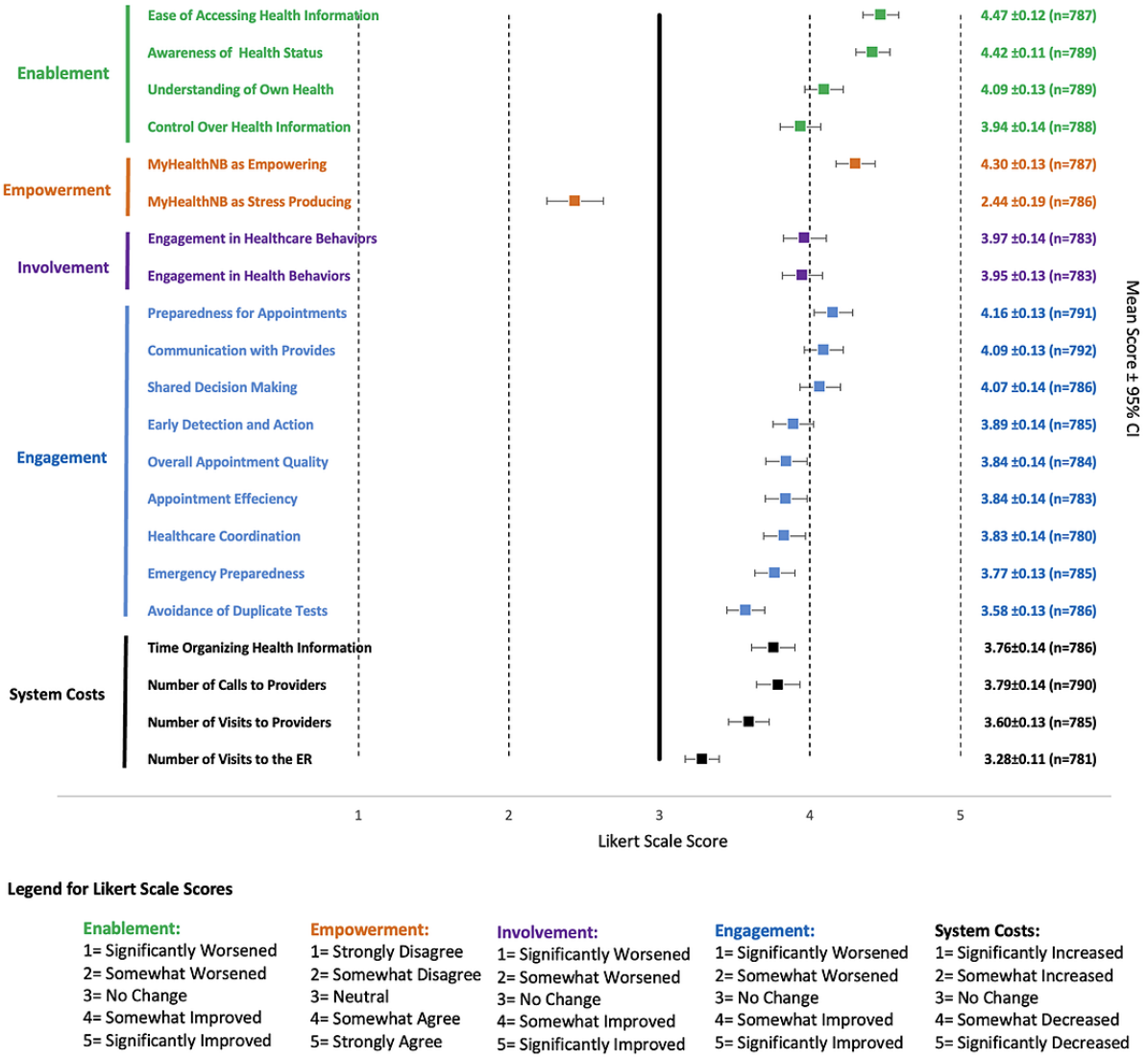

Connection Between MyHealthNB Impacts: Mediation Analysis of Impact Pathways

As shown in Figure 4, the results of the mediation analysis indicate patterns of association consistent with a hypothesized pathway where patients who feel more enabled by the PHR are also more likely to report feeling empowered, which in turn is associated with increased engagement and involvement in care.

Specifically, patients who felt more enabled by MyHealthNB were significantly more likely to feel empowered by it (β=0.94, P<.001), with enablement explaining over half of the variance in empowerment (R^2=0.53). Empowerment, in turn, was associated with engagement and involvement in a manner consistent with partial mediation of the relationship between enablement and these behavioral outcomes. For engagement, enablement alone predicted outcomes (β=0.80, P<.001, R^2=0.58), and both enablement (β=0.62, P<.001) and empowerment (β=0.19, P<.001) remained significant in the combined model, improving model fit (R^2=0.61), consistent

XSL•FO
RenderX

with partial mediation. For involvement, enablement alone was also predictive (β=0.92, P<.001, R^2=0.54), and the addition of empowerment (β=0.11, P=.001) similarly improved model fit (R^2=0.55), again consistent with partial mediation. In contrast, while enablement by MyHealthNB also predicted lower reported stress from MyHealthNB (β=–0.72, P<.001, R^2=0.16), stress did not significantly predict patient engagement (β=–0.02, P=.21), patient involvement (β=0.00, P=.96), or cost-related outcomes (β=–0.02, P=.22), suggesting stress was not associated with outcomes in a manner consistent with a mediating role. For cost-related outcomes, enablement was independently predictive (β=0.57, P<.001, R^2=0.29), but neither empowerment nor stress contributed explanatory value beyond enablement. Multimedia Appendix 1 includes the full mediation analysis results. All composite variables used in the mediation analyses demonstrated acceptable internal consistency. Enablement (4 items; Cronbach α=0.82), involvement (2 items; α=0.84; interitem r=0.73), engagement (9 items; α=0.96), and cost-related outcomes (4 items; α=0.81) supported aggregation of items into averaged composite scores. Empowerment and stress were each measured using single items; therefore, internal consistency reliability could not be assessed for these constructs.

Figure 4. Mediation analysis of MyHealthNB impacts.

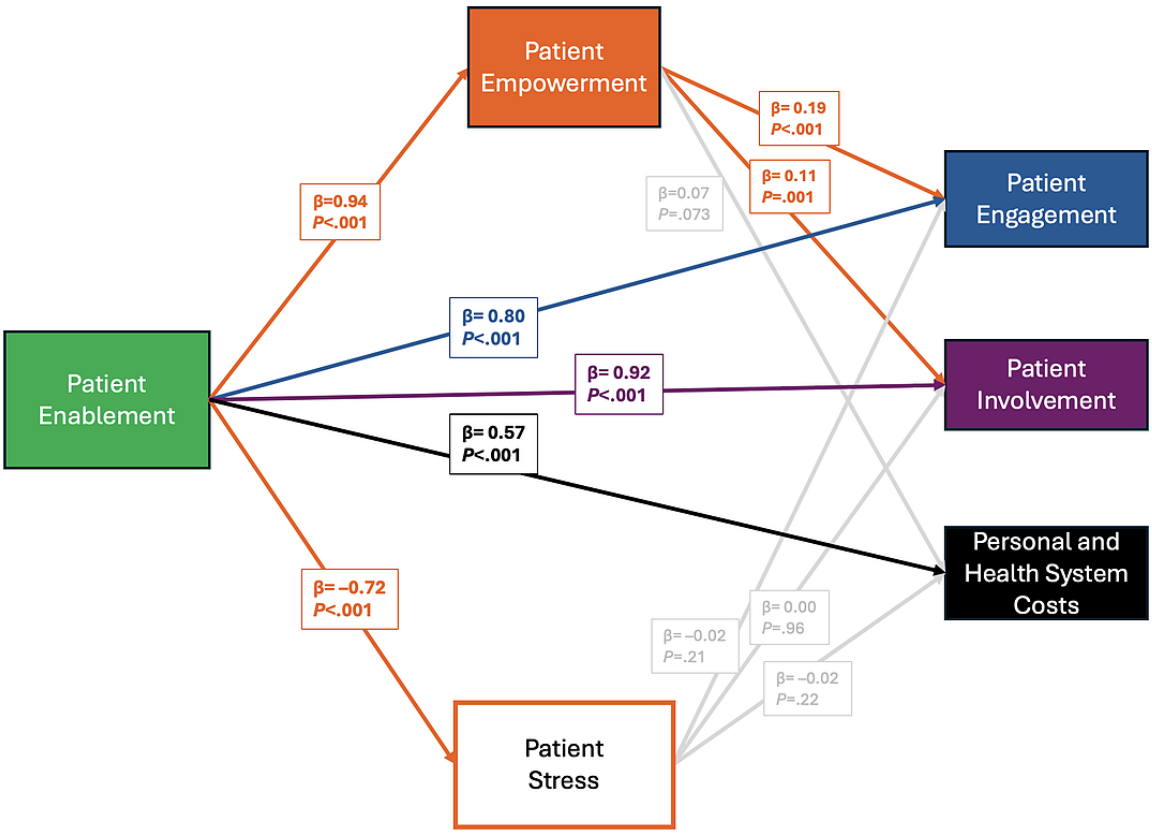

Enabling Environments of MyHealthNB Impacts: Multiple Regression Analysis of Predictors

Figure 5 presents the results of multivariable linear regression models used to identify factors that predict variations in self-reported impacts. A higher resolution image of Figure 5 can be found in the Multimedia Appendix 1. Satisfaction with MyHealthNB and having a family doctor were significantly associated with all 21 impacts (P<.05). In addition, provider support for MyHealthNB, digital literacy, and frequency of MyHealthNB use were significantly associated with 20, 15, and 15 impacts respectively (P<.05). Other variables, including use of the MyHealthNB homepage and education level were also significantly associated with specific impacts, although their associations were less consistent. Full details of these predictive factors, including effect sizes, significance levels, and multiple regression results, are provided in Multimedia Appendix 1**.** VIF values for the model examining predictors of MyHealthNB's impact on ease of accessing health information ranged from 1.07 to 1.45, indicating no evidence of problematic multicollinearity (Multimedia Appendix 1 and Limitations section provide more detail).

XSL•FO
RenderX

Figure 5. Map of MyHealthNB impacts and predictors.

		Dependent Variable	Predictor Variables															
		Impact of MyHealthNB + = Impact	Racial Identity +=Minority	Gender Identity +=Female	Digital Health Service Navigation Ease +=Easier	Residence Length +=Longer	Age +=Older	Managing an Ongoing Health Condition +=Condition	Language Spoken +=French Speaking	Location +=Rural Location	MyHealthNB Access Format +=App usage	Education Level +=More Education	MyHealthNB Home Page Use += Home Page Use	MyHealth NB Use Frequency +=Higher Use	Digital Literacy += More Literate	Provider Support of MyHealthNB +=More Satisfied	Having a Family Doctor +=Family Doctor	Satisfaction with MyHealthNB += More Satisfied
Outcome Variables: Enablement	Improved Ease of Health Information Access	+							-			+			+	+	+	+
	Improved Awareness of Health Status	+											+	+	+	+	+	+
	Improved Understanding of Own Health	+									+		+		+	+	+	+
	Improved Control Over Health Information	+												+	+	+	+	+
Outcome Variables: Empowerment	Agreement with MyHealthNB being Empowering	+										+	+	+	+	+	+	+
	Agreement with MyHealthNB Producing Stress	-				-		-				-			-		-	-
Outcome Variables: Involvement	Improved Engagement in Healthcare Behaviours	+									+		+		+	+	+	+
	Improved Engagement in Health Behaviours	+											+		+	+	+	+
Outcome Variables: Engagement	Improved Preparedness for Appointments	+								-				+	+	+	+	+
	Improved Communication with Providers	+								-			+	+	+	+	+	+
	Improved Shared Decision Making	+										+	+	+	+	+	+	+
	Improved Preparedness for Emergency	+												+	+	+	+	+
	Improved Early Detection and Action	+										+		+		+	+	+
	Improved Avoidance of Duplicate Tests	+					-	-	+					+		+	+	+
	Improved Appointment Efficiency	+											+	+	+	+	+	+
	Improved Healthcare Coordination	+			+								+	+		+	+	+
	Improved Overall Appointment Quality	+									+		+	+	+	+	+	+
Outcome Variables: Costs	Decreased Personal Time Spent	+		-			-							+		+	+	+
	Decreased Calls to Providers	+													+	+	+	+
	Decreased Visits to Providers	+												+		+	+	+
	Decreased ER Visits	+				-				+		-	+	+		+	+	+

Legend

- $P < .05$ Significant (Positive Relationship)
- $P < .05$ Significant (Negative Relationship)
- $P > .05$ Not Significant

+ Holding all other variables constant, **an increase in the predictor variable** is associated with an increase in the outcome variable

− Holding all other variables constant, **a decrease in the predictor variable** is associated with an increase in the outcome variable

Discussion

Summary of Findings

This mixed methods evaluation suggests that MyHealthNB, a province-wide PHR in New Brunswick, Canada, is associated with self-reported improvements in patient enablement, empowerment, involvement, engagement, and personal and health system costs. Qualitative findings revealed that MyHealthNB plays a foundational role in enabling patients through timely, accessible health information. This access fostered empowerment, with users describing greater peace of mind and control in managing their health. While some participants reported moments of increased stress when encountering unfamiliar or unexpected results, most described this stress as occurring alongside greater awareness and involvement in their care. Participants also reported perceived reductions in personal and health system costs, including fewer phone calls, appointments, and administrative burdens, suggesting that MyHealthNB may support more efficient use of both patient and system resources. Overall, participants saw MyHealthNB as a valuable way to promote deeper participation in their health and health care.

Quantitative findings aligned with qualitative insights, demonstrating MyHealthNB statistically significant improvements across 21 measures related to enablement, empowerment, involvement, engagement, and cost-related outcomes. Mediation analyses identified patterns of association consistent with a hypothesized pathway in which higher perceived enablement from MyHealthNB was associated with greater feelings of empowerment, alongside higher levels of patient involvement and engagement. Although greater enablement was also associated with lower reported stress related to MyHealthNB use, stress was not associated with involvement, engagement, or cost-related outcomes. Together, these findings suggest that empowerment may play a more central role than stress in shaping patient involvement and engagement, highlighting empowerment as a potentially important target for future intervention design and evaluation.

Multiple regression analysis identified key predictors of stronger impacts across all outcome domains, including satisfaction with MyHealthNB, having a family doctor, provider support for MyHealthNB, digital literacy, and MyHealthNB use frequency. These findings highlight the importance of both individual skills and broader system-level enablers to maximize the benefits of PHRs. Together, the qualitative and quantitative results indicate that interconnected, province-wide PHRs like MyHealthNB can strengthen empowerment-related outcomes.

Key Implications

PHRs Can Influence Outcomes Related to Patient Empowerment and Behavior Change

This study speaks directly to calls for stronger empirical evidence linking PHR use to changes in patient empowerment and behavioral outcomes [10]. The quantitative and qualitative results suggest that province-wide PHRs like MyHealthNB are more than passive information repositories; they can be active drivers of patient empowerment and health behavior change, especially when use is supported by health care providers. Our

XSL•FO
RenderX

findings echo insights from a large systematic review that assessed the broad impact of PHRs and found improvements in enablement (eg, increased knowledge), empowerment (eg, reassurance and reduced anxiety), involvement (eg, increased adherence to medications), and engagement (eg, improved consultations, patient-provider relationships) outcomes [36]. They also align with a conceptual synthesis and integrative review that examined literature supporting various PHR value propositions [35]. This synthesis identified evidence for PHRs improving value propositions around patient education, enablement, understanding, control, self-management, self-efficacy, preparation for emergencies, and provider communication [35]. This same synthesis also found evidence of reductions in personal and system-level costs, through savings in patient and provider time, reduced duplications of tests and treatments, and reduced burden on health care system resources [35]. Overall, our findings contribute to a growing body of evidence demonstrating PHRs' impact on a range of patient and health system outcomes.

Complex Emotional Reactions Are Inherent to PHR Use

Our qualitative findings revealed that PHRs can simultaneously evoke empowerment and stress when using MyHealthNB. Importantly, these responses were not uniform across users, which led us to include follow-up survey questions examining both effects. Quantitative results showed that most respondents agreed MyHealthNB was empowering and disagreed that it led to stress; however, nearly one-quarter of respondents (23%) reported that using MyHealthNB was stress-inducing, indicating a meaningful minority of users experienced emotional burden.

The quantitative findings further suggested that higher perceived stress related to MyHealthNB use was associated with shorter length of residence in New Brunswick, not managing a chronic condition, lower education, lower digital health literacy, lack of a family doctor, and lower satisfaction with MyHealthNB as a platform. These patterns suggest that stress may arise when users have fewer experiential, informational, or clinical supports to help interpret and act on health information. For example, users with lower digital or health literacy and those newer to the health system may face greater uncertainty when reviewing results, while users without a regular primary care provider may lack opportunities to contextualize findings.

Our findings extend the literature on PHR-related stress [36,37], which has previously suggested that stress from access to health data might cause increased health care usage [38-40]. In our study, stress was not conceptualized as inherently beneficial or harmful, but rather as a natural response that may co-occur with increased awareness and engagement. Psychological research suggests that stress may represent a temporary and motivating form of heightened alertness when individuals feel more in control [41]. Moreover, uncertainty-reduction theory posits that access to reliable information can reduce overall anxiety by decreasing ambiguity, even if initial exposure to results is emotionally challenging [42]. These perspectives may help explain why many participants reported feeling empowered despite moments of stress. However, this interpretation remains speculative and warrants further investigation. Some participants described sustained anxiety or chose to disengage from MyHealthNB after encountering concerning results, underscoring the need for future research on long-term psychological impacts of PHR use.

Overall, the findings from this study underscore the importance of responsible PHR design and implementation that is sensitive to user context. Design strategies such as plain-language explanations, embedded educational resources, and clear guidance on next steps may be particularly valuable for users with fewer supports, helping ensure that greater access to information promotes empowerment rather than distress. Emerging applications of artificial intelligence (AI) may offer additional opportunities to support these goals [43-45]. For example, AI-generated plain-language summaries, contextual explanations of test results, and tailored guidance on urgency and next steps could help users better interpret their health information, particularly those with lower digital literacy or limited access to primary care [43-45]. Future research should evaluate whether such AI-enabled supports can enhance empowerment without compromising accuracy, transparency, and trust.

The Impact of PHRs Is Shaped by User Experience, Primary Care Access, and Provider Support

Our qualitative and quantitative results show that MyHealthNB's impact was significantly shaped by key contextual factors: (1) satisfaction with MyHealthNB, (2) having a family doctor, (3) satisfaction with provider support for MyHealthNB, (4) digital literacy, and (5) MyHealthNB use frequency. In contrast, most demographic variables, such as gender, age, race, language, location of residence, length of residence, and managing an ongoing health condition, had small, inconsistent effects when controlling for other variables. Education level emerged as the most consistent demographic predictor. Higher education was associated with greater impact on self-reported ease of accessing health information, agreement with MyHealthNB being empowering, engagement in shared decision-making, and early detection and action. Conversely, lower education was associated with higher reports of stress from using MyHealthNB. This pattern aligns with prior research showing that individuals with lower education levels are more likely to experience heightened health-related anxiety [46,47].

Among usage-related variables, MyHealthNB home page use was positively associated with most outcomes. This suggests that deeper use of MyHealthNB (eg, accessing more features and more interactions with personal and health systems information) contributed to greater perceived impact across enablement, empowerment, involvement, engagement, and cost-related domains. Satisfaction with MyHealthNB, digital literacy, and MyHealthNB use frequency were among the most consistent and expected predictors of impact, reinforcing prior research showing that ease of use, digital competence, and system familiarity are key factors influencing behavior change through digital health tools [48].

Having a family doctor and feeling supported by a provider were strongly associated with users reporting that MyHealthNB helped them feel more enabled, empowered, involved, engaged, and efficient in managing their care. This is especially relevant in New Brunswick, Canada, where access to primary care

XSL•FO
RenderX

remains a challenge [49], and where 23% of survey respondents reported not having a family doctor. Many of the benefits attributed to MyHealthNB, including improved understanding of health information, better preparation for appointments, and more effective communication with providers, may depend on an ongoing supportive relationship with a provider who can help interpret and act on information. Our results reinforce the importance of embedding digital health tools within supportive care environments [50-52]. Even well-designed PHRs may be contingent on the strength of the health care ecosystem in which they are used. Without adequate access to primary care and active provider engagement, PHRs risk becoming isolated tools rather than integrated supports that improve patient care and system efficiency. PHR implementers should prioritize strategies that strengthen provider engagement and meaningfully integrate PHRs into primary care workflows.

Strengths and Limitations

This study is the first academic mixed methods evaluation of a province-wide PHR in Canada, exploring impacts on patient empowerment-related outcomes. A major strength of this study is its pragmatic, developmental mixed methods design, which enabled iterative refinement of research priorities, alignment with real-world implementations, and integration of qualitative insights into the quantitative survey. The large and relatively representative samples in both phases supported robust analyses, including mediation and regression modeling. Overall, the study demonstrated how a PHR like MyHealthNB can support patient-reported improvements across empowerment-related domains, while also identifying contextual factors that influence its impact.

This study has several limitations. Recruitment was conducted exclusively online (via email and web), which may have limited participation to individuals with internet access and higher levels of digital engagement. Measures related to internet or device access barriers were not included. Participation was voluntary and self-selected, and males and individuals with lower education levels were underrepresented, potentially biasing the sample towards an overestimation of perceived benefits. All data were self-reported, introducing potential for recall bias and social desirability bias. Although we aimed to use neutral language when asking participants to reflect on changes related to MyHealthNB, these responses, particularly those regarding reduced provider calls and visits, should be interpreted with caution. Future research could incorporate objective usage data to validate self-reported impacts. The cross-sectional survey design was appropriate for the study's pragmatic goals, but it limits the ability to draw causal conclusions.

Additionally, while survey items were informed by qualitative findings, prior Canadian surveys, codevelopment with stakeholders, and relevant literature, they were not drawn from validated scales. Many constructs were assessed using single-item measures, which precluded assessment of internal consistency and increased susceptibility to measurement error and response bias, therefore limiting reliability and construct validity. As a result, observed associations should be interpreted as patterns of perceived impact rather than precise estimates of underlying constructs. The findings in this study should be considered exploratory and hypothesis-generating, pending newly developed measures.

Several limitations specific to the mediation analyses should also be noted. Although mediation models were used to examine theoretical impact pathways, the cross-sectional nature of the survey data precludes confirmation of temporal ordering among enablement, empowerment, stress, and downstream behavioral outcomes. As a result, these analyses can only identify associations consistent with hypothesized pathways rather than demonstrate causal mediating processes or directional mechanisms. Formal indirect effects were not estimated, as doing so would risk overstating causal inference in the absence of longitudinal or experimental data. Accordingly, mediation findings should be interpreted as theory-consistent and exploratory, intended to inform future hypothesis testing rather than confirm causal pathways.

Multicollinearity was only formally assessed for the regression model examining MyHealthNB's impact on ease of accessing health information. Although the same predictors were used in all models and similar multicollinearity patterns may be expected, VIFs were not recalculated separately for each model and may vary across analyses due to differences in missing data.

Findings from this study should be interpreted within the broader health system context of New Brunswick, which is characterized by primary care shortages, variable digital literacy, and a PHR that primarily provides access to information rather than fully integrated provider-facing functions (eg, instant messaging). In larger provinces or jurisdictions with more mature digital infrastructure and primary care access, the impacts observed here may be amplified or moderated by existing supports and workflows. Evaluating other Canadian PHRs may suggest that provider messaging, linked educational resources, and clearer clinical follow-up pathways may play an important role in shaping perceived impact. Overall, PHR impacts are not solely a function of the technology itself, but of the surrounding health system context, implementation strategy, and availability of supports.

Finally, this paper reports on one component of a broader evaluation examining the design, implementation, and impacts of MyHealthNB in New Brunswick. Future publications will present additional findings, including barriers and facilitators to design and implementation, and perspectives from individuals who do not use MyHealthNB.

Conclusions

This study used an exploratory sequential mixed methods approach to examine how MyHealthNB, a province-wide PHR in New Brunswick, Canada, impacted patient empowerment-related outcomes. Our findings show that MyHealthNB users see MyHealthNB as more than a way to access information, recognizing it as something that actively supports enablement, empowerment, involvement, engagement, and cost-related improvements. The most impactful experiences with MyHealthNB occurred when patients had access to a family doctor and received support from providers, highlighting the importance of embedding digital tools within trusted care relationships. Although some users experienced stress when

XSL•FO
RenderX

interacting with MyHealthNB, this did not diminish its overall benefits. Overall, citizens report broad benefits from using a PHR, but realizing its full impact depends on user confidence with the digital experience and strong integration with primary care.

Acknowledgments

We would like to thank all of our partners and participants who supported this study. Generative artificial intelligence tools were used to support grammatical editing and improve writing clarity in this publication. The authors take full responsibility for the content, interpretations, and conclusions presented in this study.

Funding

This project was funded under the MITACS Elevate Postdoctoral Fellowship program. It was also supported through funding from the Canada Research Chairs Program.

Data Availability

Data generated or analyzed during this study are not publicly available, as participants did not consent to share the data outside of the research team.

Authors' Contributions

PV, CSG, and SUD contributed to the study design and data analysis. Recruitment was carried out by PV, FB, and SUD, while data collection was conducted by PV and FB. PV, CSG, SUD, and FB contributed to writing and reviewing the manuscript.

Conflicts of Interest

PV holds an MITACS Elevate Postdoctoral Fellowship, which facilitated a partnership between Sinai Health (the academic partner) and VeroSource Solutions (the industry partner). PV holds no equity and receives no direct personal compensation from the industry partner. The industry funder had no role in the study data analysis, interpretation, or the decision to publish. More information about the MITACS Elevate Postdoctoral program can be obtained directly from the MITACS organization.

Multimedia Appendix 1
Additional study details.
[DOCX File , 612 KB - jmir_v28i1e87521_app1.docx]

References

1. Ekman I, Ebrahimi Z. Person-centred care: looking back, looking forward. Eur J Cardiovasc Nurs 2021;20(2):93-95. [doi: 10.1093/eurjcn/zvaa025] [Medline: 33693738]
2. Santana MJ, Manalili K, Jolley RJ, Zelinsky S, Quan H, Lu M. How to practice person-centred care: a conceptual framework. Health Expect 2018;21(2):429-440 [FREE Full text] [doi: 10.1111/hex.12640] [Medline: 29151269]
3. Coulter A, Oldham J. Person-centred care: what is it and how do we get there? Future Hosp J 2016;3(2):114-116 [FREE Full text] [doi: 10.7861/futurehosp.3-2-114] [Medline: 31098200]
4. Munthe C, Sandman L, Cutas D. Person centred care and shared decision making: implications for ethics, public health and research. Health Care Anal 2012;20(3):231-249 [FREE Full text] [doi: 10.1007/s10728-011-0183-y] [Medline: 21786152]
5. Tonelli M, Sullivan MD. Person-centred shared decision making. J Eval Clin Pract 2019;25(6):1057-1062 [FREE Full text] [doi: 10.1111/jep.13260] [Medline: 31407417]
6. Pulvirenti M, McMillan J, Lawn S. Empowerment, patient centred care and self-management. Health Expect 2014;17(3):303-310 [FREE Full text] [doi: 10.1111/j.1369-7625.2011.00757.x] [Medline: 22212306]
7. Irizarry T, Dabbs AD, Curran CR. Patient portals and patient engagement: a state of the science review. J Med Internet Res 2015;17(6):e148 [FREE Full text] [doi: 10.2196/jmir.4255] [Medline: 26104044]
8. Kruse C, Bolton K, Freriks G. The effect of patient portals on quality outcomes and its implications to meaningful use: a systematic review. J Med Internet Res 2015;17(2):e44 [FREE Full text] [doi: 10.2196/jmir.3171] [Medline: 25669240]
9. Han H, Gleason KT, Sun C, Miller HN, Kang SJ, Chow S, et al. Using patient portals to improve patient outcomes: systematic review. JMIR Hum Factors 2019;6(4):e15038 [FREE Full text] [doi: 10.2196/15038] [Medline: 31855187]
10. Ammenwerth E, Neyer S, Hörbst A, Mueller G, Siebert U, Schnell-Inderst P. Adult patient access to electronic health records. Cochrane Database Syst Rev 2021;2(2):CD012707 [FREE Full text] [doi: 10.1002/14651858.CD012707.pub2] [Medline: 33634854]
11. Roehrs A, Costa CD, Righi RDR, de Oliveira KSF. Personal health records: a systematic literature review. J Med Internet Res 2017;19(1):e13 [FREE Full text] [doi: 10.2196/jmir.5876] [Medline: 28062391]

XSL•FO
RenderX

12. Kharrazi H, Chisholm R, VanNasdale D, Thompson B. Mobile personal health records: an evaluation of features and functionality. Int J Med Inform 2012;81(9):579-593 [FREE Full text] [doi: 10.1016/j.ijmedinf.2012.04.007] [Medline: 22809779]
13. Cerezo PG, Juvé-Udina M, Delgado-Hito P. Concepts and measures of patient empowerment: a comprehensive review. Rev Esc Enferm USP 2016;50(4):667-674. [doi: 10.1590/s0080-623420160000500018]
14. Bravo P, Edwards A, Barr PJ, Scholl I, Elwyn G, McAllister M. Conceptualising patient empowerment: a mixed methods study. BMC Health Serv Res 2015;15:252 [FREE Full text] [doi: 10.1186/s12913-015-0907-z] [Medline: 26126998]
15. Castro EM, Van Regenmortel T, Vanhaecht K, Sermeus W, Van Hecke A. Patient empowerment, patient participation and patient-centeredness in hospital care: a concept analysis based on a literature review. Patient Educ Couns 2016;99(12):1923-1939. [doi: 10.1016/j.pec.2016.07.026] [Medline: 27450481]
16. Hickmann E, Richter P, Schlieter H. All together now - patient engagement, patient empowerment, and associated terms in personal healthcare. BMC Health Serv Res 2022;22(1):1116 [FREE Full text] [doi: 10.1186/s12913-022-08501-5] [Medline: 36056354]
17. Risling T, Martinez J, Young J, Thorp-Froslie N. Defining empowerment and supporting engagement using patient views from the citizen health information portal: qualitative study. JMIR Med Inform 2018;6(3):e43 [FREE Full text] [doi: 10.2196/medinform.8828] [Medline: 30201603]
18. Brands MR, Gouw SC, Beestrum M, Cronin RM, Fijnvandraat K, Badawy SM. Patient-centered digital health records and their effects on health outcomes: systematic review. J Med Internet Res 2022;24(12):e43086 [FREE Full text] [doi: 10.2196/43086] [Medline: 36548034]
19. MySaskHealthRecord: Benefits Evaluation. Canada Health Infoway.: Canada Health Infoway; 2020. URL: https://www.infoway-inforoute.ca/en/component/edocman/3868-mysaskhealthrecord-benefits-evaluation/view-document?Itemid=0 [accessed 2025-11-10]
20. Vanderhout S, Taneja S, Kalia K, Wodchis WP, Tang T. Patient experiences and perspectives when MyChart is introduced in a large community hospital: mixed methods study. J Med Internet Res 2025;27:e66353 [FREE Full text] [doi: 10.2196/66353] [Medline: 39847431]
21. Access to Electronic PHI. Canada Health Infoway. 2024. URL: https://insights.infoway-inforoute.ca/2024-access-to-electronic-phi/ [accessed 2025-06-30]
22. MyHealthNB. URL: https://myhealth.gnb.ca/s/?language=en_US [accessed 2025-06-30]
23. Schoonenboom J, Johnson RB. How to construct a mixed methods research design. Kolner Z Soz Sozpsychol 2017;69(2):107-131 [FREE Full text] [doi: 10.1007/s11577-017-0454-1] [Medline: 28989188]
24. Patton M. Developmental Evaluation. Thousand Oaks, CA: Eval Pract Sage Publications Sage CA; 1994:311-319.
25. Patton MQ. State of the art and practice of developmental evaluation. In: Developmental Evaluation Exemplars. USA: Guilford Publications; 2016:1-24.
26. Bradshaw C, Atkinson S, Doody O. Employing a qualitative description approach in health care research. Glob Qual Nurs Res 2017;4:2333393617742282 [FREE Full text] [doi: 10.1177/2333393617742282] [Medline: 29204457]
27. Goldkuhl G. Pragmatism vs interpretivism in qualitative information systems research. Eur J Inf Syst 2017;21(2):135-146. [doi: 10.1057/ejis.2011.54]
28. Sandelowski M. Whatever happened to qualitative description? Res Nurs Health 2000;23(4):334-340. [doi: 10.1002/1098-240x(200008)23:4<334::aid-nur9>3.0.co;2-g] [Medline: 10940958]
29. Nelson J. Using conceptual depth criteria: addressing the challenge of reaching saturation in qualitative research. Qual Res 2016;17(5):554-570. [doi: 10.1177/1468794116679873]
30. Brinkmann S. Unstructured and semi-structured interviewing. In: The Oxford Handbook of Qualitative Research. Oxford: Oxford University Press; 2014:277-299.
31. Hamilton A, Finley EP. Qualitative methods in implementation research: an introduction. Psychiatry Res 2019;280:112516 [FREE Full text] [doi: 10.1016/j.psychres.2019.112516] [Medline: 31437661]
32. Kowalski CP, Nevedal AL, Finley EP, Young JP, Lewinski AA, Midboe AM, et al. Planning for and assessing rigor in rapid qualitative analysis (PARRQA): a consensus-based framework for designing, conducting, and reporting. Implement Sci 2024;19(1):71 [FREE Full text] [doi: 10.1186/s13012-024-01397-1] [Medline: 39394597]
33. Nevedal AL, Reardon CM, Opra Widerquist MA, Jackson GL, Cutrona SL, White BS, et al. Rapid versus traditional qualitative analysis using the consolidated framework for implementation research (CFIR). Implement Sci 2021;16(1):67 [FREE Full text] [doi: 10.1186/s13012-021-01111-5] [Medline: 34215286]
34. Pekonen A, Eloranta S, Stolt M, Virolainen P, Leino-Kilpi H. Measuring patient empowerment - a systematic review. Patient Educ Couns 2020;103(4):777-787. [doi: 10.1016/j.pec.2019.10.019] [Medline: 31767243]
35. Ruhi U, Chugh R. Utility, value, and benefits of contemporary personal health records: integrative review and conceptual synthesis. J Med Internet Res 2021;23(4):e26877 [FREE Full text] [doi: 10.2196/26877] [Medline: 33866308]
36. Tapuria A, Porat T, Kalra D, Dsouza G, Xiaohui S, Curcin V. Impact of patient access to their electronic health record: systematic review. Inform Health Soc Care 2021;46(2):192-204 [FREE Full text] [doi: 10.1080/17538157.2021.1879810] [Medline: 33840342]

XSL•FO
RenderX

37. Ammenwerth E, Schnell-Inderst P, Hoerbst A. The impact of electronic patient portals on patient care: a systematic review of controlled trials. J Med Internet Res 2012;14(6):e162 [FREE Full text] [doi: 10.2196/jmir.2238] [Medline: 23183044]
38. Taneja S, Kalia K, Tang T, Wodchis WP, Vanderhout S. Examining health care provider experiences with patient portal implementation: mixed methods study. J Med Internet Res 2025;27:e65967 [FREE Full text] [doi: 10.2196/65967] [Medline: 39888658]
39. Van Rijt AM, Hulter P, Weggelaar-Jansen AM, Ahaus K, Pluut B. Mental health care professionals' appraisal of patients' use of web-based access to their electronic health record: qualitative study. J Med Internet Res 2021;23(8):e28045 [FREE Full text] [doi: 10.2196/28045] [Medline: 34448705]
40. Van Kuppenveld SI, Van Os-Medendorp H, Tiemessen NA, Van Delden JJ. Real-time access to electronic health record via a patient portal: is it harmful? a retrospective observational study. J Med Internet Res 2020;22(2):e13622 [FREE Full text] [doi: 10.2196/13622] [Medline: 32044753]
41. Le Fevre M, Matheny J, Kolt G. Eustress, distress, and interpretation in occupational stress. J Manag Psychol 2003;18(7-8):726-744. [doi: 10.1108/02683940310502412]
42. Leslie AB, Dawn OB. Uncertainty reduction theory. In: Engaging Theories in Interpersonal Communication. Thousand Oaks, California: Sage Publications Inc; 2008.
43. Kirchner GJ, Kim RY, Weddle JB, Bible JE. Can artificial intelligence improve the readability of patient education materials? Clin Orthop Relat Res 2023;481(11):2260-2267. [doi: 10.1097/CORR.0000000000002668] [Medline: 37116006]
44. Zaretsky J, Kim JM, Baskharoun S, Zhao Y, Austrian J, Aphinyanaphongs Y, et al. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. JAMA Netw Open 2024;7(3):e240357 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.0357] [Medline: 38466307]
45. Biro JM, Handley JL, Malcolm McCurry J, Visconti A, Weinfeld J, Gregory Trafton J, et al. Opportunities and risks of artificial intelligence in patient portal messaging in primary care. NPJ Digit Med 2025;8(1):222 [FREE Full text] [doi: 10.1038/s41746-025-01586-2] [Medline: 40275104]
46. Barbek RME, Makowski AC, Von dem Knesebeck O. Social inequalities in health anxiety: a systematic review and meta-analysis. J Psychosom Res 2022;153:110706. [doi: 10.1016/j.jpsychores.2021.110706] [Medline: 34954602]
47. Norbye AD, Abelsen B, Førde OH, Ringberg U. Distribution of health anxiety in a general adult population and associations with demographic and social network characteristics. Psychol Med 2022;52(12):2255-2262 [FREE Full text] [doi: 10.1017/S0033291720004122] [Medline: 33183380]
48. Arias López MDP, Ong BA, Borrat Frigola X, Fernández AL, Hicklent RS, Obeles AJT, et al. Digital literacy as a new determinant of health: a scoping review. PLOS Digit Health 2023;2(10):e0000279 [FREE Full text] [doi: 10.1371/journal.pdig.0000279] [Medline: 37824584]
49. Johnson C, Bourgoin D, Dupuis JB, Félix JM, LeBlanc V, McLennan D, et al. Exploration of primary care models and timely access to care in New Brunswick (Canada). BMC Prim Care 2024;25(1):366. [doi: 10.1186/s12875-024-02618-8] [Medline: 39402440]
50. Greenhalgh T, Hinder S, Stramer K, Bratan T, Russell J. Adoption, non-adoption, and abandonment of a personal electronic health record: case study of HealthSpace. BMJ 2010;341(nov16 1):c5814-c5814 [FREE Full text] [doi: 10.1136/bmj.c5814] [Medline: 21081595]
51. Mold F, de Lusignan S, Sheikh A, Majeed A, Wyatt JC, Quinn T, et al. Patients' online access to their electronic health records and linked online services: a systematic review in primary care. Br J Gen Pract 2015;65(632):e141-e151 [FREE Full text] [doi: 10.3399/bjgp15X683941] [Medline: 25733435]
52. Singh H, Tahsin F, Nie JX, McKinstry B, Thavorn K, Upshur R, et al. Exploring the perspectives of primary care providers on use of the electronic patient reported outcomes tool to support goal-oriented care: a qualitative study. BMC Med Inform Decis Mak 2021;21(1):366 [FREE Full text] [doi: 10.1186/s12911-021-01734-0] [Medline: 34965860]

Abbreviations

AI: artificial intelligence
PHR: personal health record
VIF: variance inflation factor

XSL•FO
RenderX

Edited by A Stone; submitted 10.Nov.2025; peer-reviewed by S Vanderhout, T Jimoh, O Akinpeloye, K Chen; comments to author 06.Jan.2026; revised version received 19.Feb.2026; accepted 20.Feb.2026; published 28.Apr.2026.

Please cite as:
Voorheis P, Dombrowski SU, Bruno F, Steele Gray C
Exploring Patient Empowerment and Health System Outcomes Associated With MyHealthNB, a Provincial Personal Health Record System: Exploratory Mixed Methods Study
J Med Internet Res 2026;28:e87521
URL: https://www.jmir.org/2026/1/e87521
doi:10.2196/87521

©Paula Voorheis, Stephan U Dombrowski, Frances Bruno, Carolyn Steele Gray. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Clinical Effects of a Digital Health Intervention for Adults With Type 2 Diabetes in the United States: Retrospective Cohort Study

David Kerr[1*], MD; Nita Thingalaya[2*], MD; Praveen Potukuchi[2], PhD; Keni CS Lee[3], PhD; Edward Jonathan Han-Burgess[4], MS, MSc, MBA; Darren Frey[5,6], PhD; Alison Edwards[7], MStat; Xinyan Yu[7], MD; Laura Wilson[2], PharmD; Felix Lee[2], MPharm, MSc, MBA

[1]Center for Health Systems Research, Sutter Health, Santa Barbara, CA, United States
[2]Sanofi, Morristown, NJ, United States
[3]Sanofi, Reading, Berkshire, United Kingdom
[4]Sanofi, Cambridge, MA, United States
[5]Ensemble Insight, Paris, France
[6]Sciences Po, Paris, France
[7]Symphony Health, ICON plc, Blue Bell, PA, United States
[*]these authors contributed equally

Corresponding Author:
Nita Thingalaya, MD
Sanofi
100 Morris St
Morristown, NJ, 07690
United States
Phone: 1 445 245 1602
Email: nita.thingalaya@gmail.com

Abstract

Background: Digital health interventions show promise in supporting adults with type 2 diabetes mellitus (T2DM). Dario is a digital diabetes solution (DDS) that allows individuals with T2DM to automatically log blood glucose (BG) measurements and other variables to facilitate positive behavior changes.

Objective: This retrospective cohort study compared real-world clinical outcomes of adults with T2DM using the DDS with those of adults with T2DM receiving usual care alone.

Methods: Health care and claims data collected from the DDS, the Symphony Health Integrated Dataverse, and the Quest Laboratory were integrated into a single dataset using tokenization. DDS users were compared with a matched cohort of nonusers over the period from January 2017 to October 2021. Inclusion criteria were adults ≥18 years of age with T2DM, treated with antidiabetic medications, with ≥7% glycated hemoglobin (HbA_{1c}) prior to a defined index date, and with valid HbA_{1c} measurements during follow-up. DDS users were matched 1:3 with nonusers using exact-matching criteria and propensity score matching. Difference-in-difference values were determined using a multivariable linear regression model. The primary endpoint was change in HbA_{1c} from baseline to 6 months.

Results: The two cohorts included 568 DDS users and 1699 nonusers with matched baseline characteristics. At 6 months, the generalized linear model least squares mean estimate of the difference in change in HbA_{1c} between the two cohorts was −0.23% (95% CI −0.38% to −0.07%, P=.004). Similarly, HbA_{1c} decreased significantly more in DDS users versus nonusers in sensitivity analyses at 12 months. In four subgroups of individuals with baseline HbA_{1c}>7.5%, >8.0%, >9.0%, and >11.0%, HbA_{1c} decreased significantly more in DDS users than in nonusers at 6 and 12 months. At 6 months, 47.0% of DDS users and 36.8% of nonusers had a ≥1% decrease in HbA_{1c} from baseline (P<.001). Rates of severe hypoglycemia did not increase in either cohort between baseline and 6 months. At 6 months, more DDS users with baseline HbA_{1c}≥8% had achieved HbA_{1c}<8% (P=.002). At 6-month follow-up, medication adherence rates were evaluated as an exploratory endpoint and found to be greater for DDS users than for nonusers (P=.03). Engagement with the DDS was also evaluated as an exploratory endpoint and was associated with a significant decrease in HbA_{1c} from baseline to follow-up (r=−0.28, P<.001). Engagement activities associated with reduced HbA_{1c} were BG measurement and tagging (timing of BG measurement and meal type).

XSL•FO
RenderX

Conclusions: Adults with T2DM and elevated HbA_{1c} levels using the DDS had better outcomes at 6 months than matched nonusers across various baseline HbA_{1c} levels, with no increase in the rates of severe hypoglycemia. These effects were sustained at 12 months. The results confirm that digital interventions that facilitate behavior change can potentially deliver improvements in health outcomes.

(J Med Internet Res 2026;28:e66911) doi:10.2196/66911

KEYWORDS

digital health; type 2 diabetes mellitus; T2DM; glycated hemoglobin; HbA1c; severe hypoglycemia; self-management; medication adherence; glycemic control; diabetic comprehensive care; diabetes; cohort; diabetes mellitus; blood glucose; BG; glucose; adult; adults; antidiabetic medication; antidiabetic medications; glycemia

Introduction

In type 2 diabetes mellitus (T2DM), poor glycemic control is associated with an increased incidence of microvascular and macrovascular complications and reduced quality of life. However, many people with T2DM find it difficult to achieve their individualized glycemic targets [1]. Furthermore, therapeutic inertia is common, leading to sustained hyperglycemia [1,2]. Therapeutic inertia is a multifaceted issue that is influenced by many factors, including lack of knowledge, poor medication adherence, and psychological factors [1,2]. Measures to improve glycated hemoglobin (HbA_{1c}) in people with T2DM are therefore needed, and self-management education and structured self-monitoring of blood glucose (BG) have been shown to be effective strategies [3].

Digital health interventions are becoming increasingly available and show promise in supporting the self-management of diabetes [4-6]. For any digital health technology to be accepted by health care providers and payers in the United States, meaningful and reliable outcomes data are crucial. Few rigorous comparisons to quantify the benefits of digital health interventions have been published, and the studies performed so far have often been of poor quality [7]. Further, digital therapeutics that have been authorized by the US Food and Drug Administration and require a clinician's prescription also have important limitations in the rigor of the underlying evidence [8].

Dario, a digital diabetes solution (DDS), is a nonprescription digital health intervention with a smartphone app that aims to facilitate positive behavior changes leading to improved diabetes outcomes (Multimedia Appendix 1) [9]. The DDS allows the user to monitor BG levels and to log meals, carbohydrate consumption, insulin intake, physical activity, and other parameters through the app interface, as well as generate reports that can be shared with health care providers. The aim of this retrospective study was to compare real-world clinical outcomes of adults with T2DM using the DDS with those not using the DDS.

Methods

Study Design and Data Sources

An abstract describing this study has been published earlier [10], and additional methodological details are included in Multimedia Appendix 2. Briefly, this was a retrospective cohort study in which adults with T2DM who had used the Dario DDS from January 1, 2017, to October 31, 2021, were matched with individuals with T2DM who had not used the Dario DDS in the same period. DDS nonusers were randomly assigned to a single quarter in which they had a medical claim for any diagnosis. An index date was defined as the first registration for DDS users or the first claim date in the quarter for nonusers. The follow-up period for the primary and secondary endpoints was 6 months. Some endpoints were also evaluated at 12 months (Figure 1).

The study used routinely collected real-world data from the DDS, the Symphony Health Integrated Dataverse (IDV), and the Quest Laboratory. The IDV is a database that represents approximately 286 million active patients and includes claims data linked to electronic medical records (EMR) [11]. The study used US pharmacy/medical claims data and EMR data from the IDV. HbA_{1c} results were retrieved from Quest Laboratory data [12] as the primary source, supplemented by EMR, where appropriate. Data from DDS users were deidentified and tokenized with Synoma ID so that the records from DDS, IDV, and Quest data could be triple-linked.

XSL•FO
RenderX

Figure 1. Overview of the retrospective cohort study design. DDS: digital diabetes solution; HbA_{1c}: glycated hemoglobin; T2DM: type 2 diabetes mellitus.

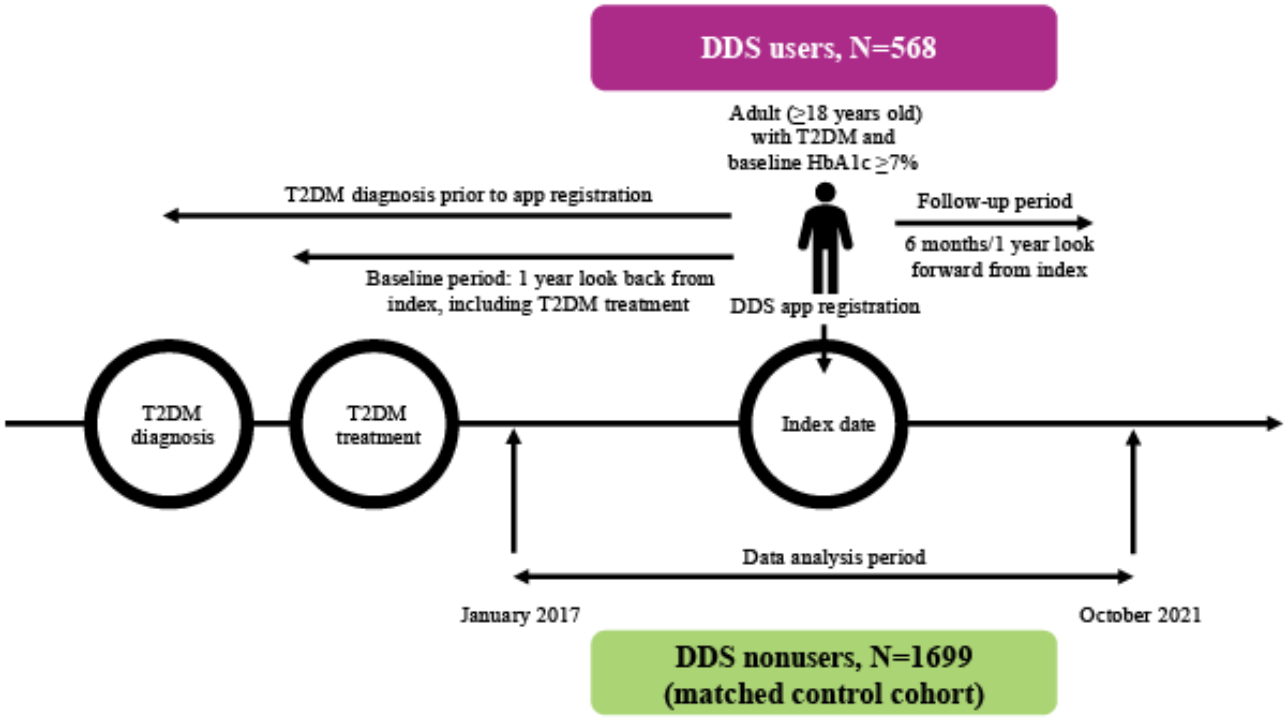

Study Population

Inclusion criteria were as follows: adults ≥18 years of age with T2DM; treated with any antidiabetic medications before the index date; with ≥1 valid HbA_{1c} measurements during the baseline period; with the most recent preindex HbA_{1c}≥7%; with ≥1 valid HbA_{1c} measurements from index date + 31 days to index date + 420 days; and with ≥90 days between baseline and postindex periods. Patients were excluded if they were <18 years old, were using continuous glucose monitoring, or were diagnosed with type 1 diabetes or gestational diabetes.

The DDS cohort consisted of all DDS users who satisfied all the inclusion criteria. The DDS nonuser cohort was selected from the universe of IDV patients meeting all the inclusion criteria and none of the exclusion criteria. Sequential 1:3 exact and propensity score matching (PSM) methods were applied to obtain the study cohort. Sex, payer type, quarter in which the index date fell, and medication type (oral antidiabetic drug [OAD], insulin, any other injectable antidiabetic drug, or any combination) were used for exact matching. Covariates used for PSM additionally included baseline HbA_{1c}, age, race/ethnicity, region, Charlson comorbidity score, comorbidities (hypertension, hyperlipidemia, anemia, depression), and comedications (concomitant medications that could impact BG levels [eg, steroids]).

Endpoints

The primary endpoint was change in HbA_{1c} from baseline to 6 months. Over the same period, secondary endpoints were change in the severe hypoglycemia event rate from baseline, percentage of patients achieving the HbA_{1c} target of <8% [13], and change in the severe hypoglycemia event rate in patients with HbA_{1c} <8%. Severe hypoglycemia events were defined using the *International Classification of Diseases* (ICD)-9 or ICD-10 diagnosis codes from medical claims.

Exploratory endpoints included change in medication adherence and association with HbA_{1c} change and the impact of the engagement activity frequency in DDS users on achieved HbA_{1c} levels, both at 6 months. For DDS users, DDS engagement frequency was defined by the number of active days within the 180-day postindex period. An active day was defined as one in which an individual performed one of the following 10 activities: measuring BG, measuring blood pressure (BP), measuring weight, tagging (timing of BG measurement and meal type), food logging (carbohydrate counting, meal photos, etc), recording the insulin dose, recording physical activity, sharing the logbook, reading an educational article, and interacting with a coach. The change in diabetic comprehensive care measures for DDS users and nonusers at 12 months was also measured. To align with HEDIS (Healthcare Effectiveness Data and Information Set) quality standards, these care measures were defined as the percentage of patients with the following details: HbA_{1c}>9%, HbA_{1c}<8%, nephropathy monitoring, retinopathy exam, and BP control<140/90 mm Hg [14].

Further exploratory endpoints were the difference between DDS users and nonusers in changes in the following parameters at 6 months, extracted from EMR data: fasting BG; total cholesterol, triglycerides, and low-density lipoprotein cholesterol (LDL-c); estimated glomerular filtration rate (eGFR); mean systolic blood pressure (SBP) and diastolic blood pressure (DBP); body weight/BMI; and cardiovascular disease (CVD) risk, as measured using the Framingham risk score [15]. Change in the CVD risk score was also measured at 12 months.

Primary and secondary endpoints were analyzed at 12-month follow-up as a sensitivity analysis. Subgroup analyses were

XSL•FO
RenderX

conducted for the primary and secondary endpoints in various groups according to HbA_{1c} values, age, the presence/absence of severe hypoglycemia, and antidiabetic treatment.

Statistical Methods

It was calculated that 513 individuals were needed in the DDS user cohort and 1539 matched DDS nonusers to achieve 80% power with α=.05 (see Multimedia Appendix 2 for full details).

Continuous variables were statistically compared between the DDS user and nonuser cohorts using *t* tests if the assumption of normality held or the Wilcoxon rank-sum test if a nonparametric test was needed based on a review of histograms. For categorical variables, study cohorts were compared using chi-square tests or Fisher's exact test, whichever was appropriate, given cell sizes and the number of categories. Multivariable linear regression with a fixed-effect model (generalized linear model [GLM]) was generated to determine the association between DDS users and follow-up HbA_{1c} compared with DDS nonusers. For endpoints reported as changes from baseline to follow-up, a difference-in-difference (D-i-D) method was used and least squares means were estimated with 95% CIs.

Medication adherence was measured by drug class using the proportion of days covered (PDC) over the 6 months preceding and 6 months following the index date. Baseline and follow-up measures of the PDC were compared as a continuous variable between the two cohorts. For evaluation of the impact of the engagement activity frequency in DDS users on achieved HbA_{1c} levels, collinearity was evaluated by determining variance inflation factors (VIFs).

Ethical Considerations

This observational study was performed in accordance with ethical principles consistent with the Declaration of Helsinki, the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use Guideline for Good Clinical Practice, Good Pharmacoepidemiology Practices, and the applicable legislation on noninterventional studies and/or observational studies. The investigators performed the observational study in accordance with the regulations and guidelines governing medical practice and ethics in the United States and in accordance with currently acceptable techniques and know-how. Human subjects were not affected by the study as it was a retrospective, anonymized data analysis. The study was reviewed by the Salus Institutional Review Board and granted a waiver of informed consent (IRB #23538).

By providing sensitive personal information (including health information and protected health information), users of the Dario DDS explicitly consented to the collection, use, and sharing of their sensitive personal information in accordance with the privacy policy [16]. In line with the privacy policy consented to by the users upon registering for an account, once personal information is aggregated, deidentified, and/or anonymized, it is no longer considered personal. Therefore, continued use of the Dario DDS is contingent upon this consent, and users do not have a specific option to opt out of their deidentified data being used for research purposes.

Results

Demographic and Baseline Characteristics

The matched cohorts included 568 Dario DDS users and 1699 nonusers. Propensity scores for DDS users and nonusers, before and after PSM, are shown in Figure 2. Demographic and baseline characteristics were similar between the two groups (Table 1).

Figure 2. Mirrored histogram of propensity scores. DDS: digital diabetes solution; LPS: longest prefix suffix; PSM: propensity score matching.

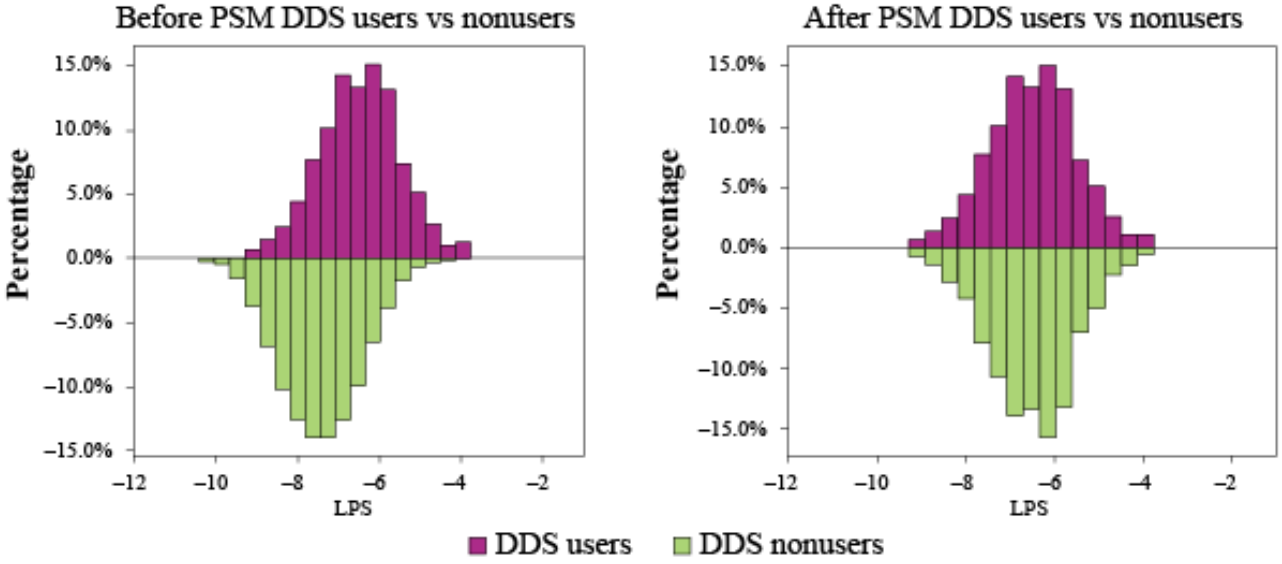

XSL•FO
RenderX

Table 1. Demographic and baseline characteristics of cohorts after matching.

Characteristics	DDS[a] users (N=568)	DDS nonusers (N=1699)
Sex,[b] n (%)		
Female	262 (46.1)	784 (46.1)
Male	306 (53.9)	915 (53.9)
Age[b] (years), mean (SD)	57.3 (10.5)	57.6 (11.6)
Age group (years), n (%)		
18-25	0	12 (0.7)
26-35	14 (2.5)	44 (2.6)
36-45	61 (10.7)	191 (11.2)
46-55	164 (28.9)	445 (26.2)
56-65	202 (35.6)	624 (36.7)
66-75	103 (18.1)	278 (16.4)
≥76	24 (4.2)	105 (6.2)
Race/ethnicity, n (%)		
African American	55 (9.7)	173 (10.2)
Asian	8 (1.4)	22 (1.3)
White	316 (55.6)	933 (54.9)
Hispanic	66 (11.6)	202 (11.9)
Other	10 (1.7)	31 (1.8)
Unknown	113 (19.9)	338 (19.9)
Geographic region, n (%)		
Mid-Atlantic	95 (16.7)	271 (16.0)
Midwest	79 (13.9)	230 (13.5)
Northeast	27 (4.8)	87 (5.1)
Northwest	10 (1.8)	29 (1.7)
Other territories	3 (0.5)	10 (0.6)
Southeast	185 (32.6)	555 (32.7)
Southwest	100 (17.6)	331 (19.5)
West	69 (12.2)	186 (11.0)
Payer type, n (%)		
Cash	3 (0.5)	9 (0.5)
Commercial	398 (70.1)	1190 (70.0)
Medicaid/managed Medicaid	28 (4.9)	84 (4.9)
Medicare	101 (17.8)	303 (17.8)
Other[c]	38 (6.7)	113 (6.7)
Index year, n (%)		
2017	115 (20.3)	345 (20.3)
2018	121 (21.3)	363 (21.4)
2019	103 (18.1)	308 (18.1)
2020	125 (22.0)	374 (22.0)
2021[d]	104 (18.3)	309 (18.2)
HbA_{1c}[a,e] (%)		

XSL•FO
RenderX

Characteristics	DDS[a] users (N=568)	DDS nonusers (N=1699)
Mean (SD)	9.14 (1.78)	9.13 (1.85)
Median (minimum-maximum)	8.6 (7-15.1)	8.6 (7-16.5)
HbA_{1c}≥8%, n (%)	387 (68.1)	1089 (64.1)
HbA_{1c}>9%, n (%)	237 (41.7)	713 (42.0)
Antidiabetic medication, n (%)		
Insulin only	34 (6.0)	101 (5.9)
OAD[f] only	289 (50.9)	864 (50.9)
Any combination/other injectable	245 (43.1)	734 (43.2)
Charlson comorbidity index score		
Mean (SD)	1.5 (1.5)	1.6 (1.6)
Median (minimum-maximum)	1 (0-9)	1 (0-13)

[a]DDS: digital diabetes solution.

[b]Standardized mean differences for sex, age, and baseline HbA_{1c} are reported in Multimedia Appendix 3.

[c]Assistance programs and/or coupons.

[d]2021 was a partial year.

[e]HbA_{1c}: glycated hemoglobin.

[f]OAD: oral antidiabetic drug.

HbA_{1c} Change from Baseline to 6 and 12 Months

Over 6 months, HbA_{1c} for the DDS group fell significantly (−1.02%, 95% CI −1.15 to −0.89], *P*<.001) (Table 2 and Figure 3). In comparison, in the nonuser cohort, HbA_{1c} fell by 0.79% (95% CI −0.87 to −0.71, *P*<.001). The GLM least squares mean estimate of the difference in change in HbA_{1c} between the two cohorts was −0.23% (95% CI −0.38% to −0.07%, *P*=.004).

The change in mean HbA_{1c} was statistically significantly greater among DDS users than among nonusers at 12 months (*P*=.03). See Multimedia Appendix 4 for details.

XSL•FO
RenderX

Table 2. Mean change in HbA_{1c}[a] from baseline to 6 months in all patients and in subgroups by baseline HbA_{1c}.

Endpoints	Baseline		Follow-up		D-i-D[c]	*P* value
	DDS[b] users (N=568)	DDS nonusers (N=1699)	DDS users (N=568)	DDS nonusers (N=1699)		
Primary endpoint: all						
Patients, n (%)	568 (100.0)	1699 (100.0)	568 (100.0)	1699 (100.0)	__[d]	—
HbA_{1c}, mean (SD)	9.14 (1.78)	9.13 (1.85)	8.12 (1.69)	8.34 (1.78)	—	—
Change from baseline (GLM[e]), mean (95% CI)	—	—	−1.02 (−1.15 to −0.89)	−0.79 (−0.87 to −0.71)	−0.23 (−0.38 to −0.07)	.004[f]
P value	—	—	<.001	<.001	—	—
Subgroup analysis: HbA_{1c}>7.5% at baseline						
Patients, n (%)	459 (80.8)	1323 (77.9)	459 (80.8)	1323 (77.9)	—	—
HbA_{1c}, mean (SD)	9.58 (1.70)	9.67 (1.76)	8.26 (1.72)	8.60 (1.85)	—	—
Change from baseline (GLM), mean (95% CI)	—	—	−1.37 (−1.52 to −1.21)	−1.06 (−1.15 to −0.96)	−0.31 (−0.49 to −0.13)	<.001[f]
P value	—	—	<.001	<.001	—	—
Subgroup analysis: HbA_{1c}>8.0% at baseline						
Patients, n (%)	374 (65.8)	1044 (61.4)	374 (65.8)	1044 (61.4)	—	—
HbA_{1c}, mean (SD)	9.99 (1.63)	10.18 (1.64)	8.40 (1.76)	8.83 (1.92)		
Change from baseline (GLM), mean (95% CI)	—	—	−1.68 (−1.87 to −1.50)	−1.31 (−1.42 to −1.20)	0.37 (−0.59 to −0.16)	<.001[f]
P value	—	—	<.001	<.001	—	—
Subgroup analysis: HbA_{1c}>9.0% at baseline						
Patients, n (%)	237 (41.7)	713 (42.0)	237 (41.7)	713 (42.0)	—	—
HbA_{1c}, mean (SD)	10.85 (1.45)	10.96 (1.41)	8.66 (1.91)	9.16 (2.06)	—	—
Change from baseline (GLM), mean (95% CI)	—	—	−2.25 (−2.50 to −1.99)	−1.78 (−1.92 to −1.63)	−0.47 (−0.77 to −0.18)	.002[f]
P value	—	—	<.001	<.001	—	—
Subgroup analysis: HbA_{1c}>11.0% at baseline						
Patients, n (%)	92 (16.2)	291 (17.1)	92 (16.2)	291 (17.1)	—	—
HbA_{1c}, mean (SD)	12.43 (0.93)	12.38 (0.99)	8.88 (2.25)	9.57 (2.3)	—	—
Change from baseline (GLM), mean (95% CI)	—	—	−3.51 (−3.98 to −3.04)	−2.82 (−3.08 to −2.55)	−0.70 (−1.23 to −0.16)	.01[f]
P value	—	—	<.001	<.001	—	—

[a]HbA_{1c}: glycated hemoglobin.

[b]DDS: digital diabetes solution.

[c]D-i-D: difference-in-difference.

[d]Not applicable.

[e]GLM: generalized linear model.

[f]Significant *P* values.

XSL•FO
RenderX

Figure 3. Mean change in HbA_{1c} from baseline to 6-month follow-up in all patients and in subgroups by baseline HbA_{1c}. Bars represent the mean change from baseline. Error bars represent 95% CIs. DDS: digital diabetes solution; D-i-D: difference-in-difference; HbA_{1c}: glycated hemoglobin.

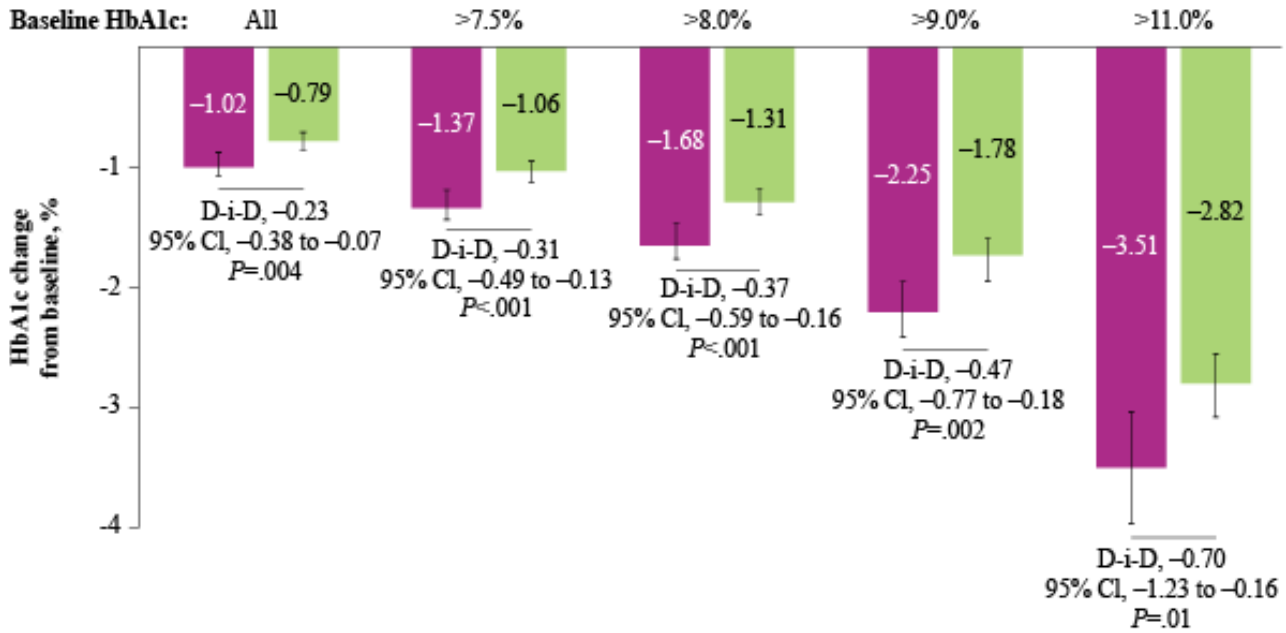

Subgroup Analyses

Analyses were conducted for four subgroups of patients based on their baseline HbA_{1c} (>7.5%, >8.0%, >9.0%, and >11.0%). In each subgroup, HbA_{1c} decreased significantly more in DDS users than in nonusers at 6 months (Table 2 and Figure 3). Similarly, in the sensitivity analyses, HbA_{1c} decreased significantly more in DDS users than in nonusers in each subgroup at 12 months (Multimedia Appendix 4). In both groups, absolute reductions in HbA_{1c} were greater in patients with higher baseline HbA_{1c} levels.

The percentage of patients who experienced an absolute HbA_{1c} decrease of ≥1% from baseline was significantly greater among DDS users compared with nonusers at 6 months (Multimedia Appendix 5). The results did not differ significantly between the groups at 12 months.

The absolute HbA_{1c} change from baseline was compared for DDS users and nonusers at 6 and 12 months in the following subgroups: patients aged ≥65 years, patients on insulin with or without other antidiabetic drugs, and patients on antidiabetic drugs other than insulin. Between-group differences in these subgroups were not statistically significant (data not shown), except for patients on antidiabetic drugs other than insulin. In this subgroup, DDS users were significantly more successful than nonusers at 6 months, with a D-i-D of −0.31% (95% CI −0.5% to −0.1%, *P*=.001). See Multimedia Appendix 5 for details.

To examine the effect of any potential between-group behavioral differences, use of test strips was considered as a proxy for the level of self-motivation in diabetes care. A subgroup analysis was performed on DDS users and nonusers who were using test strips. The decrease in HbA_{1c} did not differ significantly between the two groups at 6 or 12 months (Multimedia Appendix 5).

Secondary Endpoints

The rate of severe hypoglycemia in DDS users and nonusers did not increase between baseline and 6 months. Using the GLM, the estimated mean rate of events per 6 months was 0.0024 among DDS users and 0.0027 among nonusers (*P*=.93). See Table 3 for details.

At 6 months, 174/387 (45.0%) DDS users and 393/1089 (36.1%) nonusers who had a baseline HbA_{1c} of ≥8% had achieved an HbA_{1c} target of <8% (*P*=.002).

The difference in the change in the severe hypoglycemia event rate from baseline to 6-month follow-up between DDS users and nonusers was nonsignificant for the entire cohort and in a subgroup of patients with HbA_{1c}<8% at baseline (Table 3). Similar results were observed at 12 months (Table 3).

At 12 months, the percentage of patients who achieved HbA_{1c}<8% did not differ between DDS users and nonusers with severe hypoglycemia or without severe hypoglycemia (Multimedia Appendix 5).

XSL•FO
RenderX

Table 3. Change in the severe hypoglycemia event rate from baseline to 6 months in all patients and in patients with HbA_{1c}[a]<8%.

Endpoints	Baseline		Follow-up		Incidence rate ratio	*P* value
	DDS[b] users (N=568)	DDS nonusers (N=1699)	DDS users (N=568)	DDS nonusers (N=1699)		
All patients						
Patients, n (%)	568 (100.0)	1699 (100.0)	568 (100.0)	1699 (100.0)	__[c]	—
Patients with severe hypoglycemia, n (%)	7 (1.2)	12 (0.7)	4 (0.7)	6 (0.4)	—	—
Severe hypoglycemia events, n	9	18	11	9	—	—
Severe hypoglycemia events, mean (SD)	0.02 (0.15)	0.01 (0.15)	0.02 (0.26)	0.01 (0.11)	—	—
Follow-up severe hypoglycemia (GLM[d,e]), mean events per 6 months (95% CI)	—	—	0.0024 (0.0003-0.0176)	0.0027 (0.0010-0.0071)	0.91 (0.09-8.76)	.93
HbA_{1c}<8.0%						
Patients with baseline HbA_{1c}≥8% reaching HbA_{1c}<8%, n (%)	174 (30.6)	393 (69.2)	174 (30.6)	393 (69.2)	—	—
Patients with severe hypoglycemia, n (%)	2 (1.1)	5 (1.3)	1 (0.6)	3 (0.8)	—	—
Severe hypoglycemia events, n	3	7	3	6	—	—
Severe hypoglycemia events, mean (SD)	0.02 (0.17)	0.02 (0.18)	0.02 (0.23)	0.02 (0.21)	—	—
Follow-up severe hypoglycemia (GLM[e]), mean events per 6 months (95% CI)	—	—	0.0045 (0.0003-0.0672)	0.0044 (0.0009-0.0212)	1.01 (0.03-25.47)	.99

[a]HbA_{1c}: glycated hemoglobin.

[b]DDS: digital diabetes solution.

[c]Not applicable.

[d]GLM: generalized linear model.

[e]GLM-specified count of severe hypoglycemia events as following a negative binomial distribution.

Exploratory Endpoints

Change in Medication Adherence in Association With HbA_{1c} Change

Adherence rates during the 6-month follow-up period are shown in Table 4 for individuals with medication adherence data available. The D-i-D between the groups for the percentage of patients with ≥80% PDC was 2.16% (95% CI 0.2% to 4.1%, *P*=.03) in favor of the DDS user group.

During the 6-month follow-up period, among individuals with medication adherence data available, the PDC was similar in both groups (*P*=.31), as shown in Table 4. In the follow-up period, the overall model (controlling for baseline medication class and follow-up PDC) showed that patients with ≥80% adherence experienced on average 0.36% more HbA_{1c} reduction than patients with <60% adherence (*P*=.01). DDS users had 0.28% more HbA_{1c} reduction than nonusers (*P*=.006). See Multimedia Appendix 6 for details. When controlling for the follow-up PDC, the reduction in HbA_{1c} did not differ between DDS users and nonusers by medication class, except among patients on combination treatment (Multimedia Appendix 6).

XSL•FO
RenderX

Table 4. Adherence rates and change in adherence among DDS[a] users and nonusers.

Endpoints	Baseline		Follow-up		D-i-D[b]	*P* value
	DDS users (N=568)	DDS nonusers (N=1699)	DDS users (N=568)	DDS nonusers (N=1699)		
Patients with adherence data available,[c] n (%)	498 (87.7)	1494 (87.9)	498 (87.7)	1494 (87.9)	—[d]	—
Patients with ≥80% PDC[c,e] in follow-up period, n (%)	—	—	391 (78.5)	1140 (76.3)	—	.31
PDC (%), mean (SD)	76.5 (30.1)	78.3 (28.3)	88.1 (19.4)	86.3 (20.7)	—	—
Change from baseline (GLM[f] %), mean (95% CI)	—	—	10.6 (8.9-12.3)	8.4 (7.4-9.4)	2.1 (0.2-4.1)	.03[g]

[a]DDS: digital diabetes solution.

[b]D-i-D: difference-in-difference.

[c]Patients with antidiabetic treatment (oral antidiabetic drug [OAD], insulin, other injectable, or any combination) during baseline and follow-up periods.

[d]Not applicable.

[e]PDC: proportion of days covered.

[f]GLM: generalized linear model.

[g]Significant *P* values.

Change in HbA_{1c} by Engagement Activity in DDS Users

The mean overall DDS engagement activity was 75.4 (SD 55.9) days during the 180-day follow-up period. The median overall engagement activity was 64.5 (IQR 25-124) active days.

Descriptive statistics for overall and individual components of DDS engagement are provided in Table 5. Each day of engagement with any one of the 10 DDS engagement activities was associated with a 0.01% change in HbA_{1c} ($P<.001$). The correlation between overall DDS engagement activity and change in HbA_{1c} is shown in Figure 4. Greater engagement with the DDS was associated with a significant decrease in HbA_{1c} from baseline to follow-up ($r=-0.28$, $P<.001$).

Individual engagement activities with significant associations with reduced HbA_{1c} were BG measurement and tagging (timing of BG measurement and meal type). Recording the insulin dose had a significant association with increased HbA_{1c} (Multimedia Appendix 7).

When collinearity was evaluated, VIF values for all included engagement activities were between 1 and 2 (data not shown). As VIF≥10 would suggest collinearity among variables included in the model, collinearity was excluded.

XSL•FO
RenderX

Table 5. Descriptive statistics of DDS[a] engagement and associations with change in HbA_{1c}[b] during 6-month follow-up.

Component of engagement	Patients with ≥1 active day (N=568), n (%)	Active days, median (IQR)	Active days, mean (SD)	β-coefficient (*P* value)
Any	566 (99.6)	64.5 (25-124)	75.4 (55.9)	–0.0105 (<.001[c])
Measuring BG[d]	557 (98.1)	63.5 (24-122)	74.5 (55.8)	–0.0086 (<.001[c])
Tagging (timing of BG measurement and meal type)	450 (79.2)	9 (1-49)	33.9 (48.5)	–0.0046 (.007[c])
Sharing the logbook	425 (74.8)	1 (0-1)	1.3 (2.5)	0.0015 (0.96)
Measuring weight	206 (36.3)	0 (0-1)	3.3 (14.9)	0.0021 (0.64)
Inputting the insulin dose	131 (23.1)	0	9.1 (29.6)	0.0050 (.04[c])
Food logging (carbohydrate counting, meal photos, etc)	126 (22.2)	0	2.6 (13.4)	–0.0064 (.20)
Reading an article	110 (19.4)	0	1.5 (6.5)	0.0007 (.95)
Measuring BP[e]	50 (8.8)	0	3.6 (20.0)	–0.0020 (.59)
Recording physical activity	45 (7.9)	0	0.5 (4.2)	–0.0085 (.61)
Interacting with a coach	5 (0.9)	0	0.1 (0.7)	–0.0618 (.52)

[a]DDS: digital diabetes solution.

[b]HbA_{1c}: glycated hemoglobin.

[c]Significant *P* values.

[d]BG: blood glucose.

[e]BP: blood pressure.

Figure 4. Correlation between overall DDS engagement activity and change in HbA_{1c}, showing that greater engagement with the DDS was associated with a significant decrease in HbA_{1c} from baseline to follow-up (r=−0.28, *P*<.001). DDS: digital diabetes solution; HbA_{1c}: glycated hemoglobin.

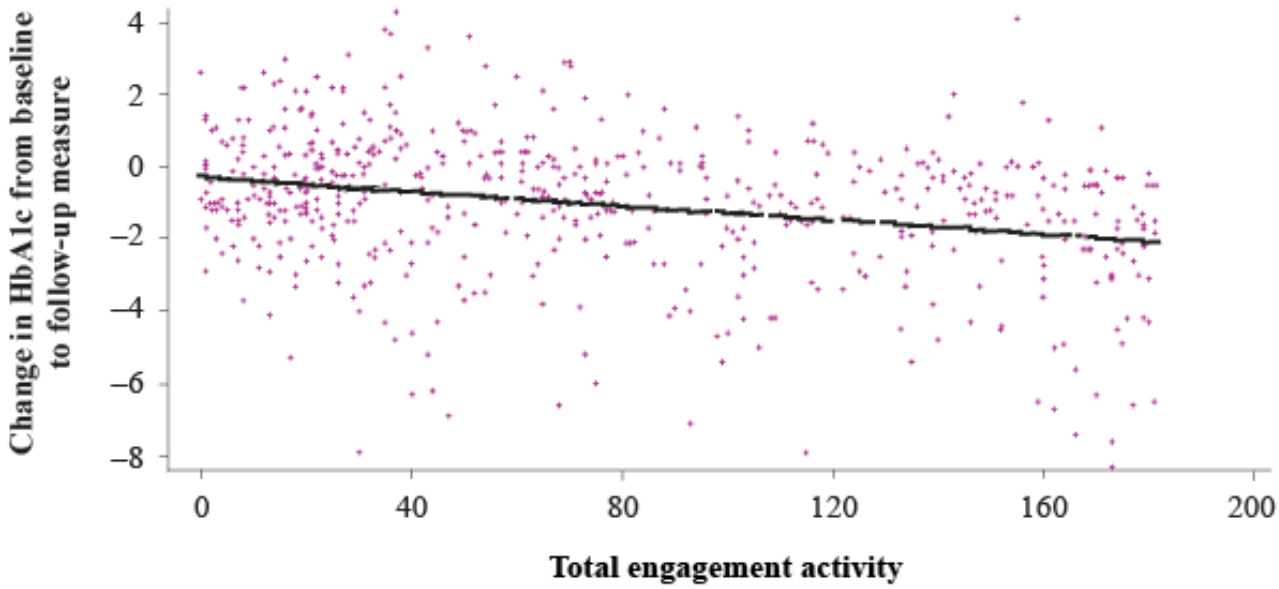

Change in Diabetic Comprehensive Care Measures at 12-month Follow-Up

Of the 387/568 (68.1%) DDS users and 1089/1699 (64.1%) nonusers who had HbA_{1c}≥8% at baseline, 179 (46.3%) and 419 (38.5%), respectively, had HbA_{1c}<8% at 12 months (*P*=.007). See Multimedia Appendix 8 for details.

Similarly, of the 237/568 (41.7%) DDS users and 713/1699 (42.0%) nonusers who had HbA_{1c}>9% at baseline, 86 (36.3%) and 347 (48.7%), respectively, still had HbA_{1c}>9% at 12 months (*P*<.001). See Multimedia Appendix 8 for details.

No statistically significant between-group differences were noted at 12 months for the percentage of patients with BP control, appropriate nephropathy monitoring, or retinopathy exam (Multimedia Appendix 8).

XSL•FO
RenderX

Change in Laboratory Values From Baseline

Changes in the following values were measured in each group at baseline and 6-month follow-up: average fasting BG; total cholesterol, triglycerides, and LDL-c; eGFR; mean SBP and DBP; and body weight/BMI (Multimedia Appendix 9). Change in the CVD risk score was measured at 12 months (Multimedia Appendix 10). No statistically significant differences were noted between groups in changes in any of these measures at follow-up.

Discussion

Principal Findings

This retrospective cohort study examined the effect of 6 months of use of the Dario DDS, a digital health intervention, by adults with T2DM. For DDS users, the decrease in HbA_{1c} between baseline and 6-month follow-up was significantly greater than in nonusers. Furthermore, this improved glycemic control did not increase users' risk of experiencing severe hypoglycemic events.

From the study reported here, several points can be noted about the favorable effect of DDS use on glycemic control. The additional lowering of HbA_{1c} in patients with T2DM who used the DDS is relevant to real-world clinical practice, as good BG control is a core goal of T2DM disease management [13]. The effect was shown to be associated with a sustained clinical impact at 12 months, and improved glycemic control was observed at population as well as individual levels. Improvements were seen in exploratory HEDIS-defined measures of care quality: the number of patients with good glycemic control increased, while those with poor glycemic control decreased. Although the mean HbA_{1c} decrease for the population overall was small, greater reductions from subanalyses were seen in individuals with higher baseline HbA_{1c} levels.

The absolute HbA_{1c} change from baseline did not differ significantly between DDS users and nonusers at 6 and 12 months in subgroups including patients aged ≥65 years or insulin users with or without other antidiabetic drugs. One possible explanation is the inclusion of a more motivated nonuser cohort than would otherwise be expected in a general population: previous studies have suggested that in people with T2DM, the frequency of HbA_{1c} testing is low compared with guideline recommendations [17,18].

The improved glycemic control associated with DDS use occurred in addition to usual care, as an incremental benefit to patients, with no requirement to change drug therapy. The DDS has the potential to provide an additional tool for patients to improve their HbA_{1c} while undergoing standard care. Although it was not evaluated as part of this study, use of the DDS makes it possible for health care providers to monitor patients' health and care episodes between appointments through reports that can be generated from the app [9], potentially informing the patient-physician consultation.

In this study, the DDS appears to have encouraged continuous patient engagement, with associated clinical improvement. Each day of patient engagement with any of the 10 meaningful activities in the DDS was associated with a greater drop in HbA_{1c}. The study also indicated that use of the DDS increased adherence with medication, defined as the PDC. Baseline rates of adherence were already close to 80%, which is considered good adherence, but increased significantly more in DDS users than in nonusers. More analysis of subgroups is needed to further characterize the source of this improvement in medication adherence.

In diabetes care, any improvement in glycemic control with a new measure must be weighed against the intervention's safety profile, specifically with respect to the occurrence of hypoglycemia. The study showed that there was no increase in the rates of severe hypoglycemia, defined as events requiring medical intervention and as recorded in claims data, among DDS users relative to baseline or to nonusers. The definition used did not provide a full picture of all hypoglycemia occurrences but captured the most severe events, which are the greatest concern to patients and have the greatest impact on resource use and costs to the health care system. Although mild hypoglycemia can be a substantial burden on patients with T2DM, it is often underreported in claims data, as patients may not seek out medical care or the hypoglycemic event may not be inputted to the claims database [19,20]. Therefore, including mild hypoglycemic events within the analysis may have reduced the accuracy of the results, but the impact of those events on patient outcomes should not considered inconsequential.

Comparison With Prior Work

Several previous studies had shown that use of the DDS we studied is associated with improved BG control. For example, Hershcovitz et al [21] reported a single-arm retrospective study on use of the DDS in individuals with T2DM and with estimated HbA_{1c} (based on BG readings) >8% (n=148) or >9% (n=220). In that study, reduction≥1.5% in estimated HbA_{1c} versus baseline were reported and were sustained over 2 years, associated with a 20% reduction in glycemic variability. Similarly, a retrospective real-world single-arm analysis followed DDS users with average BG >180 mg/dL over 12 months and found that the average BG level decreased significantly over a year versus baseline, with no differences between White adults and those from racial and ethnic minority groups [22]. The results also showed a positive association between participants' digital engagement and reduction in their glycemia. A retrospective real-world analysis of 998 DDS users found that monthly average BG levels improved significantly in the first 6 months of use, and this improvement was maintained during the following 6 months [23]. Moreover, the initial decrease in glycemia was steeper in individuals who frequently used other DDS features, in addition to BG measurement, compared with those who did not.

Randomized controlled trials of other DDS technologies have been conducted. For example, the efficacy and safety of a digital app delivering cognitive behavioral therapy was assessed in 326 adults with T2DM randomly assigned to use of the app and 343 control patients. Relative to the control arm, the app users had significantly greater reduction in HbA_{1c} from baseline at 90 days [24]. In another randomized controlled trial of an

XSL•FO
RenderX

integrated digital health care platform with artificial intelligence–based dietary management, conducted over 48 weeks in South Korea with the participation of 294 adults with T2DM who were each assigned to one of three arms (routine diabetes care only, DDS alone, or DDS with feedback from medical staff and intermittently applied personal continuous glucose monitoring), use of that DDS was associated with better glycemic control and more weight loss [25]. However, such studies remain rare likely, in large part, because of the substantial resources and organization required to conduct them. Like previous studies of the Dario DDS, most assessments of other DDS technologies have used single-arm studies, such as that of a digitally enhanced diabetes self-management, education, and support program reported by Wilson-Anumudu et al [26], or may rely on the level of retention to stratify patients [27].

Strengths and Limitations

This study had many strengths. Unlike the earlier retrospective studies of the DDS [21-23], this study is the first to include a comparator arm of DDS nonusers. The lack of a comparator arm has been a shortcoming in some studies on digital interventions, such as those analyzed by Kumar et al [8]. Furthermore, the study used well-matched cohorts more commonly associated with randomized controlled trials. This was achieved by linking data sources (including point-of-care data and patient-generated health data to provide a broad perspective) using anonymized and tokenized data, as well as the use of both exact matching and PSM to identify the DDS nonuser cohort, as described in detail elsewhere [10]. The matching methodology was designed to minimize any potential biases, with criteria (eg, comorbidities) chosen considering T2DM characteristics and outcomes. Hypertension and hyperlipidemia are risk factors for CVD and are additional proxies for disease severity/level of CVD risk, anemia affects blood HbA_{1c} levels, and depression is more prevalent among people with T2DM than those without and may be associated with HbA_{1c} levels. The comedications included in the PSM criteria reflected their use in the treatment of patients with these comorbidities [28,29].

As the study period overlapped with the COVID-19 pandemic, it was necessary to ensure that any impact on access to care did not affect the two cohorts differently. Potential biases were overcome by matching the two cohorts by quarter of the study period in which patients were enrolled. The study had a clinical primary endpoint, recognized as a factor that denotes rigor [8], and assessment of glycemic control relied on laboratory-reported HbA_{1c} rather than self-reported fasting BG. A 60-day window was allowed for HbA_{1c} tests to be conducted, which mirrors real-world practice.

The wide patient parameters used mean that the results may be generalizable, for example, to both commercially insured and Medicare-insured patients, as both were represented. Any adult with uncontrolled T2DM in the general US population who may be likely to use the DDS, which requires use of a smartphone, could potentially expect improved glycemic control. However, the study was biased toward more motivated patients in both cohorts, due to the need for baseline and follow-up HbA_{1c} measurements. Consequently, the effects associated with DDS use could potentially be even greater when compared with a less motivated population than the nonuser cohort of the study.

The study also has some limitations. First, as it was an observational study, residual confounding may be present. Baseline behavioral traits that contributed to glycemic control may have differed between DDS users and nonusers. An attempt to control for this by analyzing the use of test strips was limited because it was not possible to account for test strips that were self-paid or sent directly to users. Availability of baseline and follow-up HbA_{1c} measurements was an inclusion criterion, meaning that selection bias, for surviving patients who regularly have their data captured in interactions with health care services, is possible. There is the potential for comorbidities outside those included in the PSM criteria to be confounding, but there is a limit to the criteria that can be feasibly included in PSM, and there was a sound rationale to support the inclusion of the chosen comorbidities that do not apply to other comorbidities. Second, as in any study that relies on database entries, there may have been gaps or errors in the data. Although the coverage of the IDV is extensive (286 million patients in the United States), it is not universal, and as an open claims database, it may not have captured all health care events associated with enrolled patients. Third, individuals must be able to acquire the DDS, through employer benefits, health plan coverage, or the ability to self-pay, which could mean there are socioeconomic differences between DDS users and nonusers. Payer type was included in the PSM criteria, which should partially mitigate the socioeconomic difference between the two groups. Fourth, it was impossible to exclude patients using other digital coaching programs for diabetes, or other mobile technologies, from the DDS nonuser cohort where usage is not captured in clinical records or claims data. Capturing this information would require substantial changes to the IDV and is likely not feasible owing to the intended generalizability of the database. Finally, some subgroups served primarily to highlight directionality for future research and were not powered to show statistical significance. Nonsignificance of results in these subgroups, and between the two cohorts in laboratory values and some comprehensive diabetes health care measures, may have been due to the small sample size of subgroups. Expanding the size of the cohort population could increase the subgroup sample size, but in some cases, the size of population is inherent to the subgroup characteristics.

Conclusion

Health outcomes are determined not only by medical care but also by behavior, social determinants of health, and genetic factors. Digital health interventions show promise in supporting desirable behavior in the self-management of diabetes [4-6] but have often lacked adequate evidence for their effectiveness [7]. This retrospective cohort study provided a rigorous assessment of the impact of a digital health solution on holistic patient management. It compared clinical outcomes of adults with T2DM using a digital health intervention, the Dario DDS, in addition to usual care, with those receiving usual care alone. The use of exact-matching and PSM minimized differences between the DDS user and nonuser cohorts.

XSL•FO
RenderX

At 6-month follow-up, improvements in glycemic control were statistically significantly greater among DDS users compared with nonusers, with no increase in the rate of severe hypoglycemia. The differences in glycemic control were maintained at 12 months, even though the study had not been powered to show this. Adherence with medication increased significantly more among DDS users than among nonusers at 6 months. At 12 months, the number of patients with good glycemic control increased and those with poor glycemic control decreased.

Future research and subgroup analyses of this patient population could identify patients who would most benefit from intervention with a DDS. As increased engagement correlated with change in HbA_{1c}, efforts to identify ways to improve patient engagement with a DDS could further improve outcomes. Expanding the analysis to include patients outside the United States would increase the worldwide generalizability of the study, especially in relation to the differences in health care systems, lifestyle, and disease management.

The study provides evidence on the positive impact of a digital health intervention on the management of T2DM, with findings that are likely generalizable to a wide range of adults with T2DM. The results indicate that in a large cohort of patients with T2DM, digital health interventions that facilitate behavior change can deliver additional improvements to health outcomes, over and above the benefits conferred by drugs and disease education, that is not individually targeted.

Acknowledgments

The authors thank Shruti Pathak, Mathew Barnett, Sanjana Ravikumar, and Yingjia Shen of Symphony Health (ICON plc) for support in the management and analyses of data; Adee Kennedy of Sanofi for overall management of the study; and Charlotte Singh of Sanofi for assisting in coordination of the development of the manuscript and for performing a courtesy review.

Funding

This study was funded by Sanofi. Medical writing and editorial support was provided by Grace Townshend and Colin Glen, on behalf of Lucid Group, funded by Sanofi. The authors were responsible for all content and editorial decisions and received no honoraria related to the development of this publication.

Data Availability

Qualified researchers may request access to data. Further details on Sanofi's data-sharing criteria, eligible studies, and process for requesting access can be found online [30].

Authors' Contributions

NT, PP, AE, XY, FL, and DK made substantial contributions to the conceptualization and methodology of the study. PP, AE, and XY were responsible for data curation and formal analysis of the data. DF and all other authors made substantial contributions to the interpretation of the analysis results. All authors were involved in drafting the manuscript or reviewed it critically for important intellectual content, approved the final version for publication, and take responsibility for the integrity of the work as a whole.

Conflicts of Interest

DK has received consultancy fees from Sanofi, Abbott Rapid Diagnostics, Better Therapeutics, and Proteomics. NT, PP, KCSL, EJH-B, LW, and FL are employees of Sanofi and may hold shares and/or stock options in the company.

Multimedia Appendix 1
Dario DDS: example displays. DDS: digital diabetes solution.
[DOCX File , 292 KB - jmir_v28i1e66911_app1.docx]

Multimedia Appendix 2
Methods (supplementary information).
[DOCX File , 57 KB - jmir_v28i1e66911_app2.docx]

Multimedia Appendix 3
Standardized mean differences in baseline characteristics of unmatched and matched individuals.
[DOCX File , 57 KB - jmir_v28i1e66911_app3.docx]

Multimedia Appendix 4

XSL•FO
RenderX

Sensitivity analysis: mean change in HbA_{1c} from baseline to 12-month follow-up in all patients and in subgroups by baseline HbA_{1c}. HbA_{1c}: glycated hemoglobin.
[DOCX File , 59 KB - jmir_v28i1e66911_app4.docx]

Multimedia Appendix 5
Percentage of patients achieving specific targets and mean change in HbA_{1c} from baseline to 6- and 12-month follow-up in selected subgroups. HbA_{1c}: glycated hemoglobin.
[DOCX File , 324 KB - jmir_v28i1e66911_app5.docx]

Multimedia Appendix 6
Estimated HbA_{1c} reduction according to medication class. HbA_{1c}: glycated hemoglobin.
[DOCX File , 54 KB - jmir_v28i1e66911_app6.docx]

Multimedia Appendix 7
Change in HbA_{1c} from baseline to the 6-month follow-up using individual engagement measures. HbA_{1c}: glycated hemoglobin.
[DOCX File , 55 KB - jmir_v28i1e66911_app7.docx]

Multimedia Appendix 8
Change in diabetic comprehensive care measures at 12-month follow-up.
[DOCX File , 55 KB - jmir_v28i1e66911_app8.docx]

Multimedia Appendix 9
Change in laboratory values from baseline.
[DOCX File , 66 KB - jmir_v28i1e66911_app9.docx]

Multimedia Appendix 10
Change in CVD risk score. CVD: cardiovascular disease.
[DOCX File , 55 KB - jmir_v28i1e66911_app10.docx]

References

1. Karam SL, Dendy J, Polu S, Blonde L. Overview of therapeutic inertia in diabetes: prevalence, causes, and consequences. Diabetes Spectr 2020 Mar;33(1):8-15 [FREE Full text] [doi: 10.2337/ds19-0029] [Medline: 32116448]
2. Okemah J, Peng J, Quiñones M. Addressing clinical inertia in type 2 diabetes mellitus: a review. Adv Ther 2018 Oct 29;35(11):1735-1745 [FREE Full text] [doi: 10.1007/s12325-018-0819-5] [Medline: 30374807]
3. American Diabetes Association Professional Practice Committee. Erratum. 5. Facilitating positive health behaviors and well-being to improve health outcomes: standards of care in diabetes-2025. Diabetes Care 2025; 48(Suppl. 1):S86-S127. Diabetes Care 2025;48(4):665 [FREE Full text] [doi: 10.2337/dc25-er04a] [Medline: 39847507]
4. Nkhoma DE, Soko CJ, Bowrin P, Manga YB, Greenfield D, Househ M, et al. Digital interventions self-management education for type 1 and 2 diabetes: a systematic review and meta-analysis. Comput Methods Programs Biomed 2021 Oct;210:106370 [FREE Full text] [doi: 10.1016/j.cmpb.2021.106370] [Medline: 34492544]
5. Shan R, Sarkar S, Martin SS. Digital health technology and mobile devices for the management of diabetes mellitus: state of the art. Diabetologia 2019 Jul;62(6):877-887 [FREE Full text] [doi: 10.1007/s00125-019-4864-7] [Medline: 30963188]
6. Kerr D, King F, Klonoff DC. Digital health interventions for diabetes: everything to gain and nothing to lose. Diabetes Spectr 2019 Aug;32(3):226-230 [FREE Full text] [doi: 10.2337/ds18-0085] [Medline: 31462878]
7. Silberman J, Wicks P, Patel S, Sarlati S, Park S, Korolev IO, Evidence DEFINED Workgroup, et al. Rigorous and rapid evidence assessment in digital health with the evidence DEFINED framework. NPJ Digit Med 2023 May 31;6(1):101 [FREE Full text] [doi: 10.1038/s41746-023-00836-5] [Medline: 37258851]
8. Kumar A, Ross JS, Patel NA, Rathi V, Redberg RF, Dhruva SS. Studies of prescription digital therapeutics often lack rigor and inclusivity. Health Aff (Millwood) 2023 Dec;42(11):1559-1567 [FREE Full text] [doi: 10.1377/hlthaff.2023.00384] [Medline: 37931187]
9. Dario® blood glucose monitoring system. Digital Therapeutics Alliance. 2023. URL: https://dtxalliance.org/products/dario-blood-glucose-monitoring-system/ [accessed 2023-10-01]
10. Thingalaya N, Kerr D, Potukuchi P, Wilson L, Lee K, Han-Burgess E, et al. 962-P: impact of digital diabetes solution on glycemic control in adults with type 2 diabetes mellitus in the United States—a retrospective cohort study. Diabetes 2023 Jun 20;72(Suppl 1):962-P [FREE Full text] [doi: 10.2337/db23-962-p]

XSL•FO
RenderX

11. Analyse and integrate data. ICON. 2023. URL: https://www.iconplc.com/solutions/symphony-health/health-technology-and-business-intelligence-solutions [accessed 2023-11-01]
12. What we do. Quest Diagnostics. 2023. URL: https://www.questdiagnostics.com/our-company/what-we-do [accessed 2023-11-01]
13. ElSayed NA, Aleppo G, Aroda VR, Bannuru RR, Brown FM, Bruemmer D, on behalf of the American Diabetes Association. 6. Glycemic targets: standards of care in diabetes-2023. Diabetes Care 2023 Jan 01;46(Suppl 1):S97-S110 [FREE Full text] [doi: 10.2337/dc23-S006] [Medline: 36507646]
14. 2023 diabetes recognition program requirements package. Healthcare Effectiveness Data and Information Set (HEDIS). 2023. URL: https://store.ncqa.org/index.php/catalog/product/view/id/5188 [accessed 2023-11-01]
15. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation 2008 Mar 12;117(6):743-753 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.107.699579] [Medline: 18212285]
16. Privacy policy. Dario Health. 2025. URL: https://www.dariohealth.com/privacy-policy/ [accessed 2026-03-31]
17. Duff CJ, Solis-Trapala I, Driskell OJ, Holland D, Wright H, Waldron JL, et al. The frequency of testing for glycated haemoglobin, HbA1c, is linked to the probability of achieving target levels in patients with suboptimally controlled diabetes mellitus. Clin Chem Lab Med 2018 Dec 19;57(2):296-304 [FREE Full text] [doi: 10.1515/cclm-2018-0503] [Medline: 30281512]
18. Imai C, Li L, Hardie R, Georgiou A. Adherence to guideline-recommended HbA1c testing frequency and better outcomes in patients with type 2 diabetes: a 5-year retrospective cohort study in Australian general practice. BMJ Qual Saf 2021 Oct;30(9):706-714 [FREE Full text] [doi: 10.1136/bmjqs-2020-012026] [Medline: 33542066]
19. Liu J, Wang R, Ganz ML, Paprocki Y, Schneider D, Weatherall J. The burden of severe hypoglycemia in type 2 diabetes. Curr Med Res Opin 2018 Jan;34(1):179-186 [FREE Full text] [doi: 10.1080/03007995.2017.1391080] [Medline: 29017368]
20. Uzoigwe C, Hamersky CM, Arbit DI, Weng W, Radin MS. Assessing prevalence of hypoglycemia in a medical transcription database. Diabetes Metab Syndr Obes 2020;13:2209-2216 [FREE Full text] [doi: 10.2147/DMSO.S235298] [Medline: 32612376]
21. Hershcovitz Y, Dar S, Feniger E. 895-P: estimated A1c reduction in high-risk patients over two years of using a digital diabetes management platform. Diabetes 2020 Jun 1;69(Suppl 1):859-P [FREE Full text] [doi: 10.2337/db20-859-p]
22. Gershoni T, Ritholz MD, Horwitz D, Manejwala O, Donaldson-Pitter T, Fundoiano-Hershcovitz Y. Glycemic management by a digital therapeutic platform across racial/ethnic groups: a retrospective cohort study. Appl Sci 2022 Dec 29;13(1):431 [FREE Full text] [doi: 10.3390/app13010431] [Medline: 33599618]
23. Fundoiano-Hershcovitz Y, Hirsch A, Dar S, Feniger E, Goldstein P. Role of digital engagement in diabetes care beyond measurement: retrospective cohort study. JMIR Diabetes 2021 Mar 18;6(1):e24030 [FREE Full text] [doi: 10.2196/24030] [Medline: 33599618]
24. Hsia J, Guthrie NL, Lupinacci P, Gubbi A, Denham D, Berman MA, et al. Randomized, controlled trial of a digital behavioral therapeutic application to improve glycemic control in adults with type 2 diabetes. Diabetes Care 2022 Dec 01;45(12):2976-2981 [FREE Full text] [doi: 10.2337/dc22-1099] [Medline: 36181554]
25. Lee Y, Kim G, Jun JE, Park H, Lee WJ, Hwang Y, et al. An integrated digital health care platform for diabetes management with AI-based dietary management: 48-week results from a randomized controlled trial. Diabetes Care 2023 May 01;46(5):959-966 [FREE Full text] [doi: 10.2337/dc22-1929] [Medline: 36821833]
26. Wilson-Anumudu F, Quan R, Castro Sweet C, Cerrada C, Juusola J, Turken M, et al. Early insights from a digitally enhanced diabetes self-management education and support program: single-arm nonrandomized trial. JMIR Diabetes 2021 Mar 22;6(1):e25295 [FREE Full text] [doi: 10.2196/25295] [Medline: 33616533]
27. Tu Y, Chang Y, Chiou H, Lai K. The effects of continuous usage of a diabetes management app on glycemic control in real-world clinical practice: retrospective analysis. J Med Internet Res 2021 Jul 15;23(7):e23227 [FREE Full text] [doi: 10.2196/23227] [Medline: 34264192]
28. Beran M, Muzambi R, Geraets A, Albertorio-Diaz JR, Adriaanse MC, Iversen MM, European Depression in Diabetes (EDID) Research Consortium. The bidirectional longitudinal association between depressive symptoms and HbA: a systematic review and meta-analysis. Diabet Med 2022 Mar;39(2):e14671 [FREE Full text] [doi: 10.1111/dme.14671] [Medline: 34407250]
29. Ali S, Stone MA, Peters JL, Davies MJ, Khunti K. The prevalence of co-morbid depression in adults with type 2 diabetes: a systematic review and meta-analysis. Diabet Med 2006 Dec;23(11):1165-1173 [FREE Full text] [doi: 10.1111/j.1464-5491.2006.01943.x] [Medline: 17054590]
30. A global clinical research data sharing platform. Vivli. URL: https://www.vivli.org [accessed 2026-03-31]

Abbreviations

BG: blood glucose
BP: blood pressure
CVD: cardiovascular disease

XSL•FO
RenderX

DBP: diastolic blood pressure
DDS: digital diabetes solution
D-i-D: difference-in-difference
eGFR: estimated glomerular filtration rate
EMR: electronic medical records
GLM: generalized linear model
HbA_{1c}: glycated hemoglobin
HEDIS: Healthcare Effectiveness Data and Information Set
IDV: Integrated Dataverse
LDL-c: low-density lipoprotein cholesterol
OAD: oral antidiabetic drug
PDC: proportion of days covered
PSM: propensity score matching
SBP: systolic blood pressure
T2DM: type 2 diabetes mellitus
VIF: variance inflation factor

Edited by T Leung, A Coristine; submitted 30.Sep.2024; peer-reviewed by I Duncan, S Ranjbaran, A Gandra; comments to author 23.Nov.2024; revised version received 24.Mar.2026; accepted 24.Mar.2026; published 09.Jun.2026.

Please cite as:
Kerr D, Thingalaya N, Potukuchi P, Lee KCS, Han-Burgess EJ, Frey D, Edwards A, Yu X, Wilson L, Lee F
Clinical Effects of a Digital Health Intervention for Adults With Type 2 Diabetes in the United States: Retrospective Cohort Study
J Med Internet Res 2026;28:e66911
URL: https://www.jmir.org/2026/1/e66911
doi:10.2196/66911
PMID:

©David Kerr, Nita Thingalaya, Praveen Potukuchi, Keni CS Lee, Edward Jonathan Han-Burgess, Darren Frey, Alison Edwards, Xinyan Yu, Laura Wilson, Felix Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating Co-Designed vs Researcher-Driven Personalized Feedback Formats in a Brief Digital Alcohol Use Intervention: Mixed Methods Study

Antoinette Poulton[1*], PhD; Joanna Bowen[1*], BSc (Hon); Gezelle Dali[1], PhD; Cindy Chew[1], PhD; Katrina Prior[2], PhD; Emma K Devine[2], PhD; Louise Birrell[2], PhD; Lexine A Stapinski[2], PhD; Robert Hester[1], PhD

[1]Melbourne School of Psychological Sciences, The University of Melbourne, Parkville, Victoria, Australia

[2]The Matilda Centre for Research in Mental Health and Substance Use, The University of Sydney, Sydney, New South Wales, Australia

[*]these authors contributed equally

Corresponding Author:
Antoinette Poulton, PhD
Melbourne School of Psychological Sciences
The University of Melbourne
Grattan Street
Parkville, Victoria, 3010
Australia
Phone: 61 3 8344 6377
Fax: 61 3 9347 6618
Email: antoinette.poulton@unimelb.edu.au

Abstract

Background: Increasingly, brief digital interventions are being implemented to combat alcohol misuse. While the utility of incorporating personalized feedback into these interventions has been explored, less work has concentrated on how consumer-driven research might be harnessed to better tailor personalized feedback. Critically, the relative efficacy of co-designed interventions vs those designed by research teams without stakeholder input has not typically been explored.

Objective: This study aimed to use an exploratory sequential mixed methods approach to elicit user-designed personalized feedback to incorporate into an existing brief digital intervention developed to combat hazardous drinking; investigate differences in capacity, opportunity, and motivation (as per the Capacity, Opportunity, Motivation–Behavior model) to alter alcohol use behavior; examine whether participants prefer the co-designed product over those designed by research teams; and evaluate change in actual consumption over time as a function of co-designed (image-based) vs researcher-designed (image- or text-based) personalized feedback.

Methods: Part 1 used an online co-design process involving a convenience sample of young adult focus group participants (N=40; mean age 21.87, SD 7.51 years). Part 2 surveyed a convenience sample of young adults who drank alcohol regularly (N=222; mean age 21.14, SD 4.69 years). Participants were assigned to 1 of 3 personalized feedback conditions (researcher designed, text-based; researcher-designed, image-based; and co-designed, image-based) using block allocation with stratification based on alcohol intake and impulsivity level. Feedback was delivered via email, and impact on capacity, opportunity, and motivation to alter alcohol use behavior, plus preferred format of feedback, was examined (10.76, SD 14.05 days) via mixed and 1-way ANOVAs and post hoc tests. Change in alcohol consumption (89.70, SD 21.59 days) was investigated via a series of repeated-measures ANOVA and post hoc tests.

Results: In Part 1, qualitative thematic analysis of focus group data identified 6 overarching themes: visual representation, clarity, accessibility, brevity, peer comparison, and harm. Following this analysis, personalized feedback regarding alcohol intake and brain health was available in 3 formats: researcher designed, text-based; researcher-designed, image-based; and co-designed, image-based. In Part 2, there were no differences in capacity, opportunity, and motivation to alter alcohol use behavior as a function of feedback format (η_p^2=0.01). Participants preferred co-designed, image-based feedback over researcher-designed, image-based (d=0.55) or researcher designed, text-based (d=0.83) feedback. Regardless of feedback format, all participants significantly reduced total alcohol intake (η_p^2=0.11) and alcohol-related harm (η_p^2=0.05). Participants in image-based conditions

XSL•FO
RenderX

(researcher-designed or co-designed) reduced frequency of drinking to a greater extent (*d*=0.56) than those in the researcher-designed, text-based feedback condition (*d*=0.39).

Conclusions: Though the principles of co-design are undeniably valuable, the efficacy of co-designed interventions, over and above those designed by research teams without user input, requires further research.

Trial Registration: OSF Registries osf.io/rnbv8; https://osf.io/y4bhs

(J Med Internet Res 2026;28:e87393) doi:10.2196/87393

KEYWORDS

alcohol; Capacity, Opportunity, Motivation–Behavior; co-design; COM-B; community-based participatory research; digital health; mixed methods design; personalized intervention

Introduction

Alcohol misuse is a major global health concern. Worldwide, 4.7% of deaths are attributable to alcohol consumption, and 115.9 million disability-adjusted life years are lost [1]. Although tangible costs—such as alcohol-related health care and court expenses—are vast, the intangible costs of alcohol misuse are even greater, encompassing widespread impacts on individuals, their families, friends, and victims of alcohol-related assaults and other crimes [2]. As such, hazardous drinking—defined by the World Health Organization (WHO) as consumption that renders individuals at risk of injury and potentially harmful consequences [1]—is often the target of research. A substantial literature has sought to capitalize on the power of brief digital interventions designed to reduce hazardous drinking [3-5]. Although the efficacy of incorporating personalized feedback into these interventions has been explored [4,6], an underinvestigated area is how consumer-driven research may be harnessed to better shape personalized feedback and the impact of such tailoring.

Consumer-driven co-design research has become a priority worldwide, including in Australia [7-9]. Co-design encompasses practices ranging from facilitated focus groups to collaborative prototyping [10] and aims to engage a diverse pool of people in the design of tools for their own use [11]. By treating people as experts on themselves and their experiences [12], co-design offers an ethical and pragmatic approach to improving digital health services, especially for youth [13-17]. Evidence demonstrates that co-designed products are well-received [18-20] and yield improvements in health literacy [21], awareness of health risks [22], and symptoms of depression [23]. Co-designed interventions targeting alcohol use in adolescents have also been shown to reduce consumption and harm [24]. Nonetheless, none of these studies provide evidence that co-designed products or interventions deliver superior outcomes compared with products designed without stakeholder input.

While some researchers have highlighted the general lack of product and intervention testing in co-design research [11,12,25], there is a particularly salient gap related to investigations that directly compare co-designed products or interventions with those developed without stakeholder input [26]. For instance, a systematic review of co-designed interventions for anxiety and depression found no study compared the effectiveness of co-designed interventions against interventions not developed through co-design [16]. Although identifying a suitable comparison product may be challenging, it is feasible. For example, Thabrew et al [27] found their co-designed app significantly reduced anxiety levels in young users over an 11-week period, with sustained improvement at 3-month follow-up. If researchers had added a comparison group that used a researcher-designed anxiety reduction app, of which there are many [28,29], this would have contributed to the research gap around the relative effectiveness of co-designed interventions vs those designed without stakeholder input.

One way to test the relative efficacy of co-designed products against those developed by researchers would be to use the Capacity, Opportunity, Motivation–Behavior (COM-B) model [30,31]. This framework assesses the impact of a behavior change intervention as a function of an individual's knowledge and skills (capacity), their resources to enact change (opportunity), and the reflective and automatic processes related to planning, decision-making, impulse control, and emotion regulation that underpin motivation [30]. This model has been used to identify barriers and facilitators associated with implementing screening and brief interventions for alcohol use in primary care [32,33], to examine the impact of drinking guidelines on alcohol intake [34], and to explore factors influencing drinking behavior in young Australian adults [35]. By conceptualizing behavior change as a dynamic system, the COM-B model supports a comprehensive assessment of the likelihood of behavior change following a brief digital intervention.

This exploratory sequential mixed methods study aims to use co-design to generate user-designed feedback to incorporate into an existing brief digital intervention developed to combat hazardous drinking [6]. We hypothesize that participants who receive co-designed feedback will score more highly on the COM-B survey relative to those who receive researcher-designed feedback. We also hypothesize that participants will rate the co-designed feedback as more appealing than researcher-designed feedback. In keeping with our previous findings, we additionally hypothesize that individuals will reduce their alcohol intake over time in response to personalized feedback [6]. We are, however, agnostic regarding whether participants who receive co-designed feedback will decrease their drinking to a significantly greater extent than those who receive researcher-designed feedback. Thus, we expect not only to improve our existing intervention but to simultaneously address the research gap regarding the

XSL•FO
RenderX

relative efficacy of co-designed health interventions vs those designed without stakeholder input.

Methods

Overview

This study builds on an existing intervention, which has already been shown to reduce alcohol intake by 32% to 35% in hazardous drinkers [6]. The original intervention involved emailing participants personalized researcher-designed, text-based feedback about their alcohol consumption and brain health (especially impulsivity). In Part 1 of this study, we sought the input of stakeholders to improve the appeal, comprehensibility, and effectiveness of the feedback. In Part 2, we investigated the efficacy of the co-designed product, using the COM-B model to examine capacity, opportunity, and motivation to alter alcohol intake behavior. We also examined the appeal of this product and considered the relative impact on subsequent alcohol intake behavior.

Part 1: Co-Design

Preparation for Co-Design Focus Groups

To maximize youth involvement in the co-design process, the Youth Advisory Board at the Matilda Centre for Research in Mental Health and Substance Use (University of Sydney, Australia) was consulted for advice on running co-design focus groups that are accessible and inclusive. The Youth Advisory Board is comprised of a culturally, linguistically, and gender-diverse group of 10 members aged 16 to 25 years with representatives from metropolitan, rural, regional, and remote areas of Australia [36]. Members bring expertise and an interest in mental health and substance use via their own lived experience (personal or as carers). The Board advised on the design of materials and questions for use in focus group sessions. Members suggested that focus groups be presented with an overview of the overarching research, the original researcher-designed, text-based personalized feedback used in the pilot study (Figure 1), plus other examples of individualized feedback (Figure S1 in Multimedia Appendix 1), all as a springboard for further discussion. Thus, the researcher-designed, image-based feedback was conceived to show how personalized feedback information might be presented more graphically. This feedback was later used in Part 2 of the study to represent the type of image-based feedback that would have been developed by researchers without user input.

Figure 1. Researcher-designed, text-based personalized feedback (derived from the pilot study) used in Part 1 as a prompt for focus group discussions and in Part 2 to represent text-based feedback developed by researchers without user input.

Based on your responses to the alcohol intake protocols in the Cognition and Alcohol Use Behaviour study:

- You consumed 22 drinks in a 14-day period; this means you drank above Australian national recommended weekly intake guidelines (10 standard drinks per week).
- You drank on 4 days; that is, 29% of days.
- You drank 4 or more drinks on one day 2 times; this means you drank above Australian national recommended single occasion (binge) drinking levels 2 times.
- Your score on the World Health Organisation Alcohol Use Disorders Identification Test was 12; this means you are categorised as a hazardous drinker.

Based on your responses to cognitive tasks in this study:

- Your level of impulsivity is on the 70th percentile. This means your level of impulsivity is greater than 70 percent of persons similar to you in age. This places you at higher risk of developing alcohol-related problems.

Why might this be important?
Alcohol consumption is considered a risk factor for the development of Alcohol Use Disorder. Individuals with this disease are characterised by loss of self-control, which manifests as a strong desire to consume alcohol, difficulty regulating intake, and continuing to drink without regard for associated negative consequences. Current theories related to Alcohol Use Disorder implicate heightened impulsivity as a key factor underpinning this loss of self-control. Binge drinking has been identified as an independent risk factor linked to the later development of Alcohol Use Disorder. Binge drinking in adolescence has been found to significantly increase an individual's chance of being diagnosed as dependent in young adulthood, and binge-drinking trajectories throughout young adult years have been found to predict alcohol abuse and dependence ten years later. Given impulsivity is implicated in Alcohol Use Disorder, research concentrating on vulnerability for this disorder in binge drinkers has also focused on this construct. Results to date suggest these individuals might also be characterised by heightened impulsivity. As such, we believe your level of impulsivity may be of interest to you, particularly if the data collected identified you engaged in binge drinking during the experimental period.

Focus Group Participants

A convenience sample of 40 young Australian adult participants was recruited via social media posts, advertisements on the University of Melbourne student portal, sporting clubs, and researcher networks. Beyond a minimum age of 17 years, there were no other exclusion criteria.

XSL•FO
RenderX

Focus Groups Procedure

To promote youth attendance, focus groups were held online and scheduled outside of school and university hours. Groups were organized to include attendees of similar ages to ensure that younger participants were comfortable speaking up without concerns about older participants' perceptions or approval. When presenting to the focus groups, researchers used accessible language and paused regularly to encourage participants to ask questions and offer input. Following an overview of the broader project, the presenting researcher used semistructured questions to prompt discussion but also allowed participants to direct the dialogue where possible. Drawing on the procedure outlined by Debenham et al [22], the presenter stated there were no wrong answers, or stupid questions, nor would criticism of the feedback examples be taken personally. The goal was to create a safe, nonjudgmental, and collaborative atmosphere. So that participants were true partners in producing the co-designed feedback, when opinions differed, the researcher sought consensus about the ideal way to present each component of the feedback [12]. Figure 1 displays the original researcher-designed, text-based personalized feedback used in the pilot study [6]. While we sought general input about the appearance and presentation of the feedback, we also asked for specific detail about how to break down the original text into (1) information about alcohol intake; (2) information about brain health, specifically impulsivity; and (3) general information about the links between impulsivity and alcohol misuse. Figures 2-4 shows some of the springboard examples used in focus groups. Multimedia Appendix 1 contains all examples.

Figure 2. Researcher-designed, image-based feedback about alcohol intake (developed following consultation with the Youth Advisory Board at the Matilda Centre for Research in Mental Health and Substance Use) used in Part 1 as a prompt for focus group discussions and in Part 2 to determine the efficacy of the feedback.

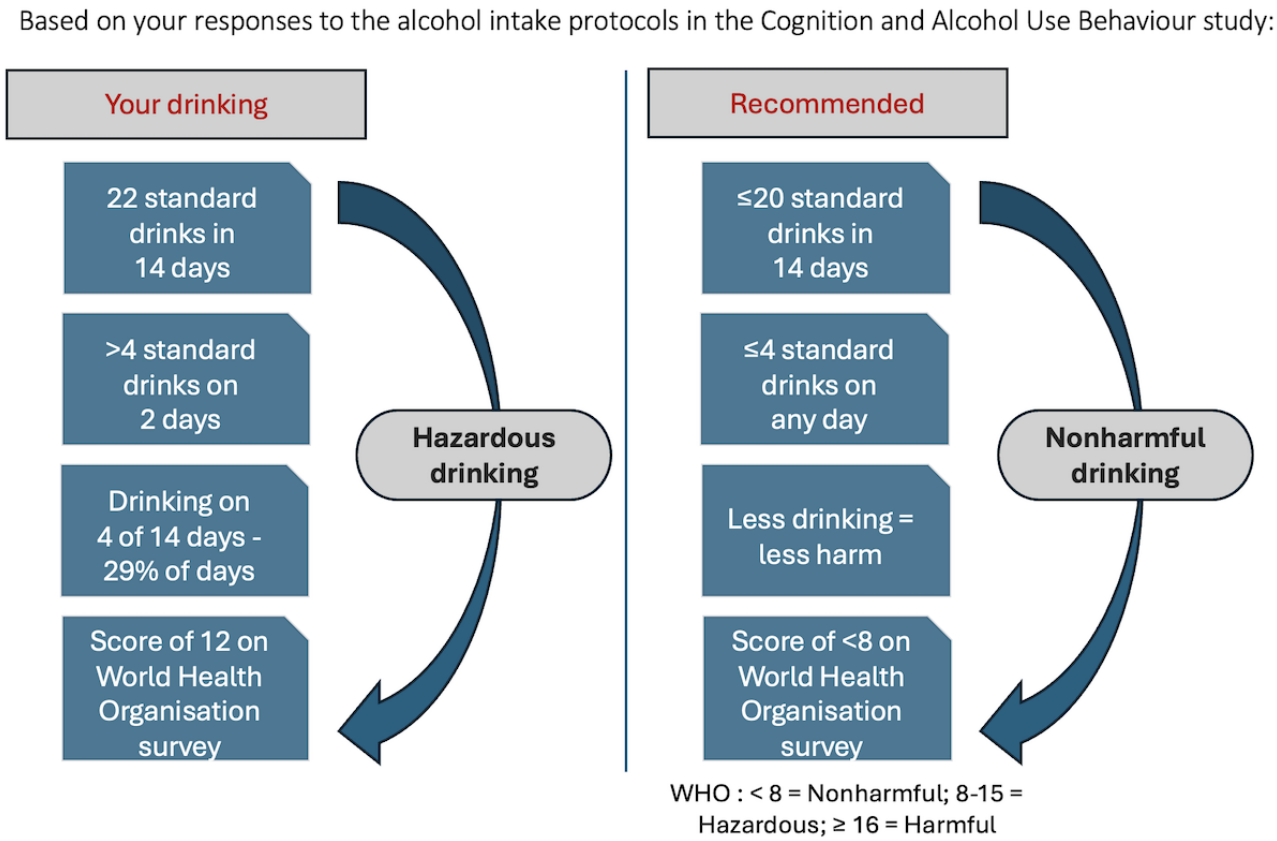

XSL•FO
RenderX

Figure 3. Researcher-designed, image-based feedback about brain health and impulsivity (developed following consultation with the Youth Advisory Board at the Matilda Centre for Research in Mental Health and Substance Use) used in Part 1 as a prompt for focus group discussions and in Part 2 to determine the efficacy of the feedback.

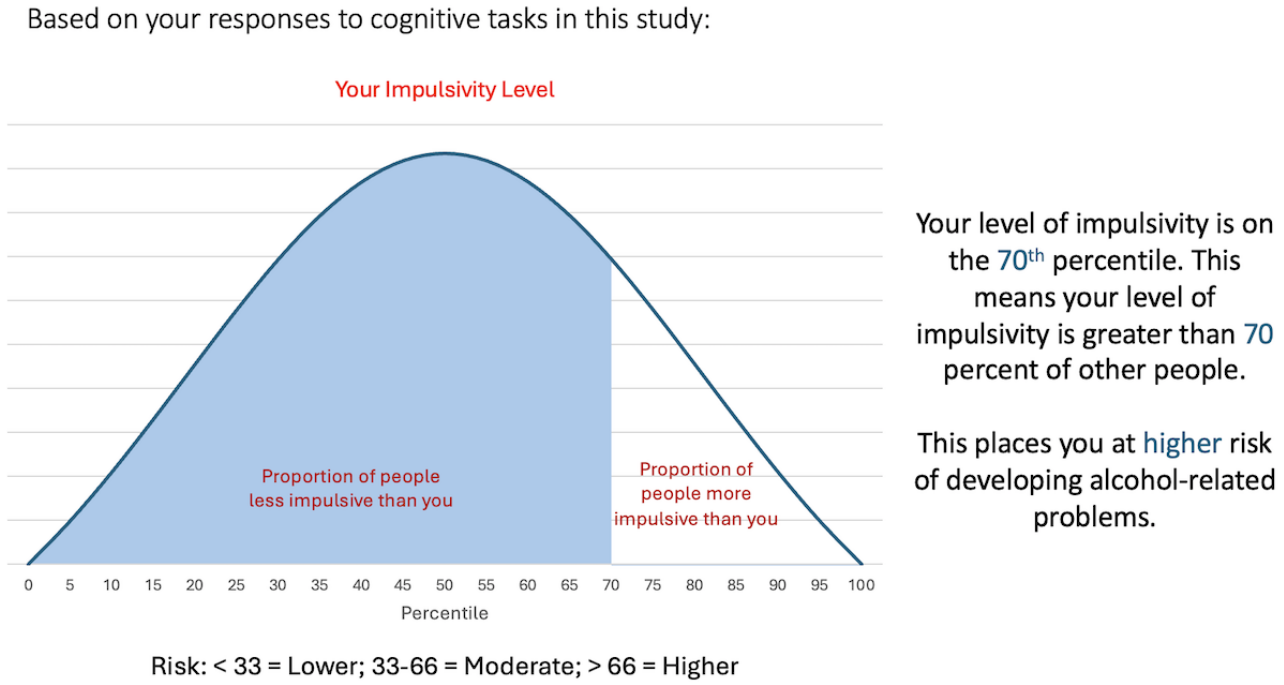

Figure 4. Researcher-designed, image-based feedback with general information (developed following consultation with the Youth Advisory Board at the Matilda Centre for Research in Mental Health and Substance Use) used in Part 1 as a prompt for focus group discussions and in Part 2 to determine the efficacy of the feedback.

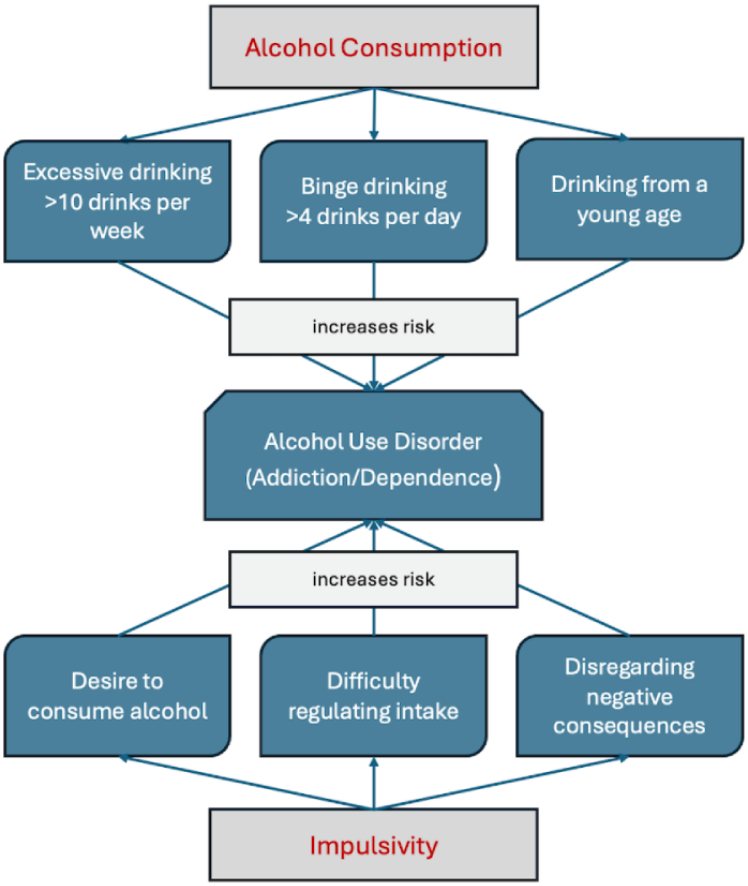

XSL•FO
RenderX

Qualitative Analyses

Focus groups were recorded via Zoom (version 6.4.0; Zoom Communications, Inc) and transcribed (by JB). The transcripts were analyzed using thematic analysis, a flexible and foundational approach to qualitative investigations [37]. Thematic analysis can be either inductive or deductive; this study took a strongly inductive, data-driven approach, focusing on what the data said rather than on a priori theories or conclusions [38]. Thematic analysis has proven useful in other participatory design research [21,25,39].

Following the key steps of thematic analysis, 2 researchers (AP and JB) independently familiarized themselves with the transcripts, then each generated an initial list of codes, which became a preliminary list of themes. The researchers then collaborated to negotiate a core list of themes, which a third researcher (CC) reviewed against the original transcripts. A final meeting was held with all 3 researchers, resulting in a consensus list of themes and feedback. The COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist can be found in Table S1 in Multimedia Appendix 1.

Part 2: Efficacy

Participants

A convenience sample of 222 young Australian adults who drank alcohol regularly were recruited via social media posts, advertisements on the University of Melbourne student portal, and researcher networks. Eligibility criteria specified that they must be aged 14 years or older and that they drink alcohol at least weekly.

Procedure

At T1 (baseline), participants completed an online survey comprising demographic questions, measures of alcohol consumption, anxiety and depression, and cognitive tasks assessing impulsivity. At the completion of T1, participants were allocated to 1 of 3 feedback conditions (original researcher-designed, text-based; researcher-designed, image-based; or co-designed, image-based), with stratification based on impulsivity and alcohol intake levels [6]. Impulsivity was derived from performance on cognitive tasks using normative data [40] and categorized as high, medium or low impulsivity via a tertile split. Alcohol intake was categorized as nonharmful, hazardous, or harmful using Alcohol Use Disorders Identification Test (AUDIT) scores [41,42].

To ensure timely feedback while maintaining stratification, we used dynamic allocation: as T1 data became available, participants were assigned a code (by AP) representing their impulsivity and alcohol category (9 possible combinations). They were then allocated (by JB) to whichever feedback condition had the lowest count in that category at the time, ensuring balanced representation across conditions. The allocation code was concealed from the other authors. Participants were not informed there were different feedback conditions in this study, nor were they explicitly informed which condition they had been allocated to.

Participants were emailed personalized feedback related to alcohol intake, impulsivity, and general information about the links between brain health (especially impulsivity) and alcohol misuse. For those assigned to the original researcher-designed, text-based feedback condition, this feedback was entirely text-based (Figure 1). Those assigned to the researcher-designed, image-based feedback condition were sent feedback identified by focus groups as the best of the examples developed by researchers without user input (Figures 2-4). Participants in the co-designed feedback condition were sent image-based feedback developed following the recommendations of focus groups from Part 1 of this study (Figure 5).

At T2, participants were invited to complete a survey that included a question about how honestly they had reported alcohol use at T1, as well as questions related to capacity, opportunity, and motivation to alter their drinking behavior (COM-B). They were then shown all 3 examples of the feedback in counterbalanced order (using Qualtrics counterbalancing code) and asked questions related to preferred format. They were not provided with information about how each type of feedback had been generated, and the feedback contained no personalized information related to alcohol intake or impulsivity.

At T3, participants were invited to again complete surveys of alcohol consumption and cognitive tasks measuring impulsivity.

XSL•FO
RenderX

Figure 5. Co-designed, image-based feedback (developed following consultation with focus groups) and used in Part 2 to determine the efficacy of the feedback.

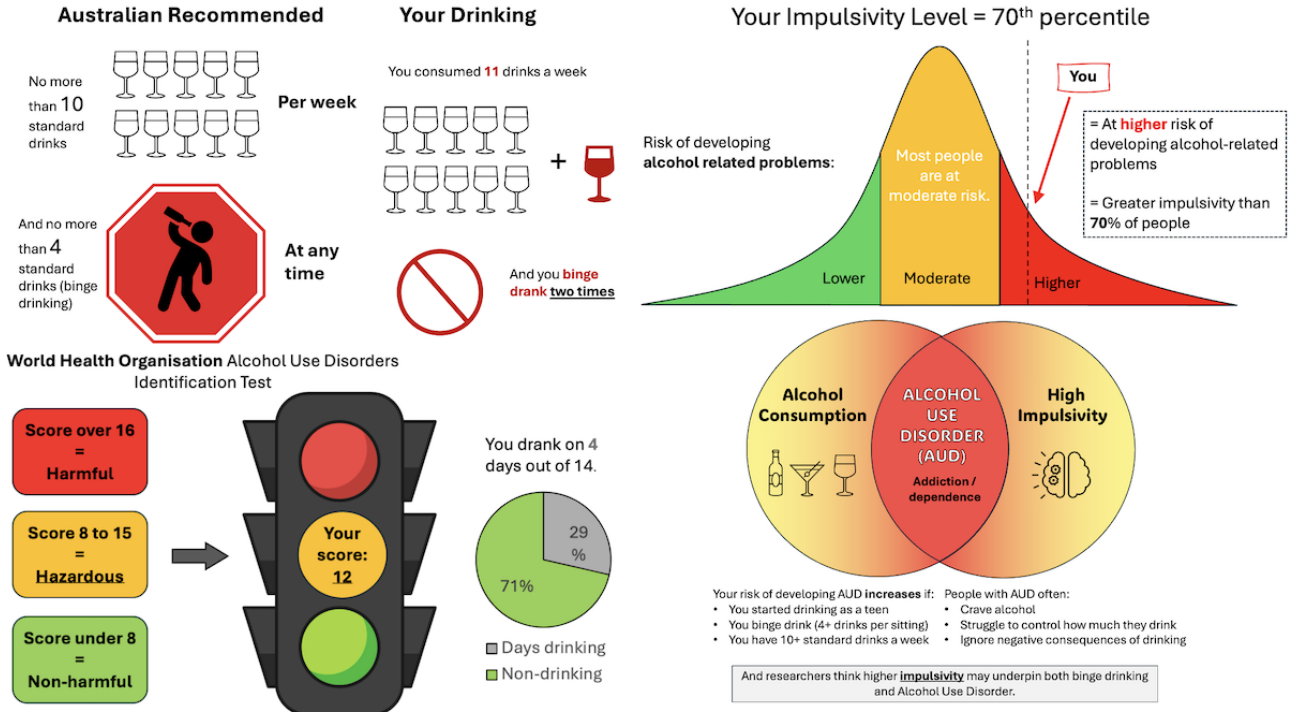

Measures

Monetary Choice Questionnaire

The monetary choice questionnaire (MCQ) is a measure of choice impulsivity, or the propensity to choose immediately rewarding options regardless of delayed consequences. Participants make hypothetical choices between 27 small immediately available monetary rewards and larger rewards obtainable only after some delay [43]. Responses can be expressed as a hyperbolic function. The extent to which an individual discounts a delayed reward is measured by the slope of the function, or overall *k*; larger *k* values indicate greater discounting of the delayed reward, or heightened choice impulsivity [44]. Consistency scores are calculated to capture whether responses are consistent with previous selections; participants with consistency scores below 75% are excluded [44].

Stop Signal Task

The stop signal task (SST) is a widely used measure of inhibitory control that captures the capacity to inhibit a response or reaction, especially once that response has been initiated. Thus, it is a useful measure of how well individuals can control their impulsivity [45,46]. The task, which has been fully described in previous work, consists of a practice block of 32 trials and 3 blocks of 64 experimental trials [46]. The main variable of interest is stop signal reaction time (SSRT), with higher values indicative of greater impulsivity. Participants are excluded if stop accuracy is less than 25% or greater than 75%, go accuracy is less than 60%, go errors are greater than 10%, SSRT is less than 50 ms, or reaction time (RT) on unsuccessful stops (choice errors and premature responses) is greater than RT on go trials [47,48].

AUDIT Measure

The 10-item AUDIT assesses alcohol intake, alcohol-related problems, and alcohol dependence over the preceding 6 months [42]. Scores suggest nonharmful (<8), hazardous (8-15), or harmful (≥16) alcohol use [41].

Timeline Followback

Timeline Followback is a retrospective diary method of collecting alcohol consumption information [49,50]. Participants record the number of standard drinks consumed over the preceding 14 days.

Generalized Anxiety Disorder Scale

The 7-item Generalized Anxiety Disorder-7 (GAD-7) assesses symptoms of anxiety over the previous 2 weeks; scores ≥10 suggest moderate anxiety [51,52].

Patient Health Questionnaire

The 9-item Patient Health Questionnaire-9 (PHQ-9) assesses depressive symptoms over the previous 2 weeks; scores ≥10 indicate moderate depression [53,54].

Advanced Progressive Matrices

A 6-item abridged online untimed version of the Advanced Progressive Matrices (APM) was used to assess intellectual capacity. The APM-6 forms half of the APM-12U, an abridged online untimed 12-item version of the APM [55,56].

COM-B: Alcohol Intake Questionnaire (COM-B:AIQ)

Previous research has successfully used the COM-B model to explore alcohol-related behavior change [34]. Working from West and Michie's guide [30], the 19-item COM-B: Alcohol Intake Questionnaire (COM-B:AIQ) was developed to measure capacity, opportunity, and motivation to alter alcohol intake behavior following the brief intervention under investigation.

XSL•FO
RenderX

Nine items assess capability (knowledge, skills, and self-efficacy), 5 items measure opportunity (social and environmental), while a further 5 items assess motivation (reflective and automatic). A higher score on the COM-B:AIQ indicates a greater likelihood of behavior change. The full measure is available in Table S2 in Multimedia Appendix 1.

Preferred Feedback

Participants rated whether they liked the "look and feel" of the feedback, how easy it was to understand, and how well they thought they would remember the information. Responses were provided via a Likert scale (1=strongly disagree to 6=strongly agree). Higher scores indicate greater liking.

Quantitative Analyses

At T1 (baseline), a percentile score was assigned to each participant for performance on each of the MCQ and SST [6]. The mean of these scores formed the average impulsivity index for each participant. Regarding the MCQ, hyperbolic *k* values were log transformed to correct for nonnormal distribution. Untransformed values (3 decimal places) are displayed in tables to assist with interpretation. With the SST, SSRT was calculated according to the integration method [48]. This involves arranging RTs on go trials (correct, incorrect, and omissions) in ascending order, with the maximum correct go RT value being substituted in the case of any go omissions. Then, the nth RT is identified by multiplying the proportion of unsuccessful stops by the number of go trials. SSRTintegration is calculated by subtracting mean stop signal delay from this nth RT. Due to SST exclusions at T1 (n=19) and T3 (n=9), T3 analyses involving change in impulsivity involved 78 participants. There were no MCQ exclusions.

We conducted 1-way ANOVA and chi-square analyses to determine whether there were any demographic, alcohol intake or cognitive task performance differences as a function of feedback group allocation at T1. We also used 1-way ANOVA to consider whether there were any differences in capacity, opportunity, and motivation to change behavior as a function of feedback group allocation at T2. A mixed measures ANOVA was used at T2 to examine rating or liking of the feedback types and to check whether preferences were influenced by T1 feedback group allocation. Feedback type (original researcher-designed, text-based; researcher-designed, image-based; and co-designed, image-based) formed the within-subjects factor, while T1 feedback group allocation (original researcher-designed, text-based; researcher-designed, image-based; and co-designed, image-based) formed the between-subjects factor. A series of repeated measures ANOVA were used to investigate change in alcohol use behavior (AUDIT score, total standard drinks, frequency of drinking, and occasions when ≥4 drinks were consumed in 1 episode) and impulsivity from T1 to T3 as a function of feedback group allocation. Time (T1 and T3) formed the within-subjects factor, while feedback group allocation (original researcher-designed, text-based; researcher-designed, image-based; and co-designed, image-based) formed the between-subjects factor. Pairwise comparisons were Bonferroni-corrected.

In the case of ANOVAs, effect sizes were calculated using partial eta squared (η_p^2): 0.01=small, 0.06=moderate, and 0.14=large effect [57]. Cohen *d* was used as a measure of effect for post hoc tests: 0.20=small, 0.50=medium, and 0.80=large [58]. For chi-square, Cramer V was used as a measure of effect: 0.07=small, 0.21=medium, and 0.35=large effect [58]. Based on our pilot study [6], we were aware attrition from T1 to T2 would be approximately 30% to 35% but were uncertain about attrition at T3. We therefore conducted an a priori power analysis for T2 and a post hoc analysis for T3. For analyses at T2, the a priori analysis conducted using G*Power (effect size f=0.25; α=.05; power=0.80) showed a total sample of 159 was required. For analyses at T3, the post hoc analysis (effect size f=0.25; α=.05; n=91; number of measurements=2; number of groups=3) showed that power was 0.68. The CONSORT (Consolidated Standard of Reporting Trials) checklist is available in Table S3 in Multimedia Appendix 1.

Ethical Considerations

Study protocols were reviewed and approved by the University of Melbourne Office of Research Ethics and Integrity Human Ethics Committee (approval number 26716). In both Parts 1 and 2 of this study, participants received dedicated ethics-approved plain language statements. If they decided to continue with the study, they then signed dedicated ethics-approved consent forms. Participants from Part 1 were reimbursed AU $75 (US $54) in GiftPay vouchers. Participants from Part 2 received course credit for contributing data at T1 and T2 and were reimbursed AU $15 (US $11) in GiftPay vouchers for T3. To ensure confidentiality and privacy, demographic and survey data were collected via the Qualtrics platform. Qualtrics servers are protected by high-end firewall systems. Exported datasets, audio files, and focus groups transcripts were stored separately on the University of Melbourne OneDrive cloud storage service and were accessible only to AP and JB. Participation in the study was entirely voluntary. The study was not preregistered because we sought to investigate the effectiveness of different types of feedback rather than to test the existing overarching intervention. Nonetheless, we have retrospectively registered the peer-reviewed grant proposal and outcome letter that provides a prospective description of this study [59].

Results

Overview

Figure 6 provides an overview of Parts 1 and 2 of this study, while Figure 7 shows the flow of participants through Part 2, including the numbers allocated to each condition across time points, those lost to follow-up, and those available for analyses.

XSL•FO
RenderX

Figure 6. Flowchart for the mixed methods study evaluating co-designed vs researcher-driven personalized feedback formats in a brief digital alcohol use intervention. YAB: Youth Advisory Board.

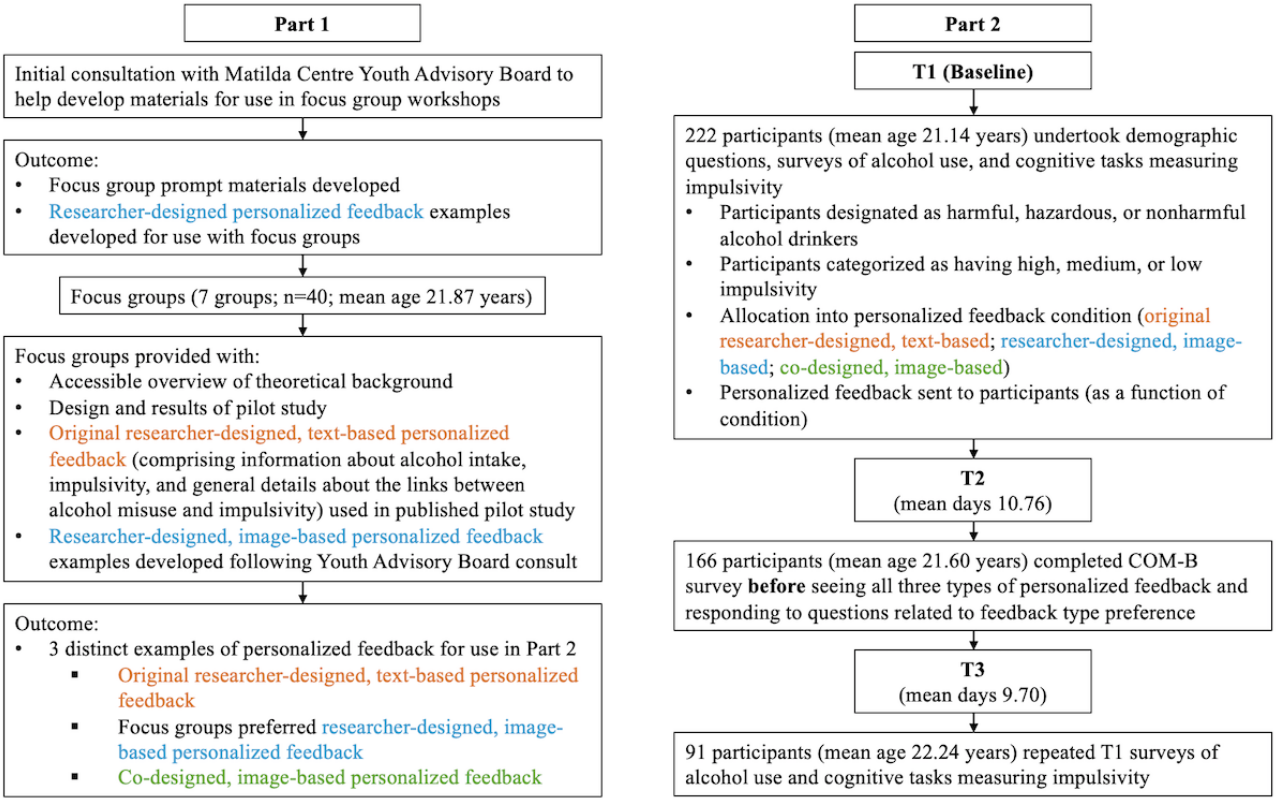

Figure 7. CONSORT (Consolidated Standards of Reporting Trials) flow diagram of participant enrollment in Part 2, allocation, follow-up, and analysis.

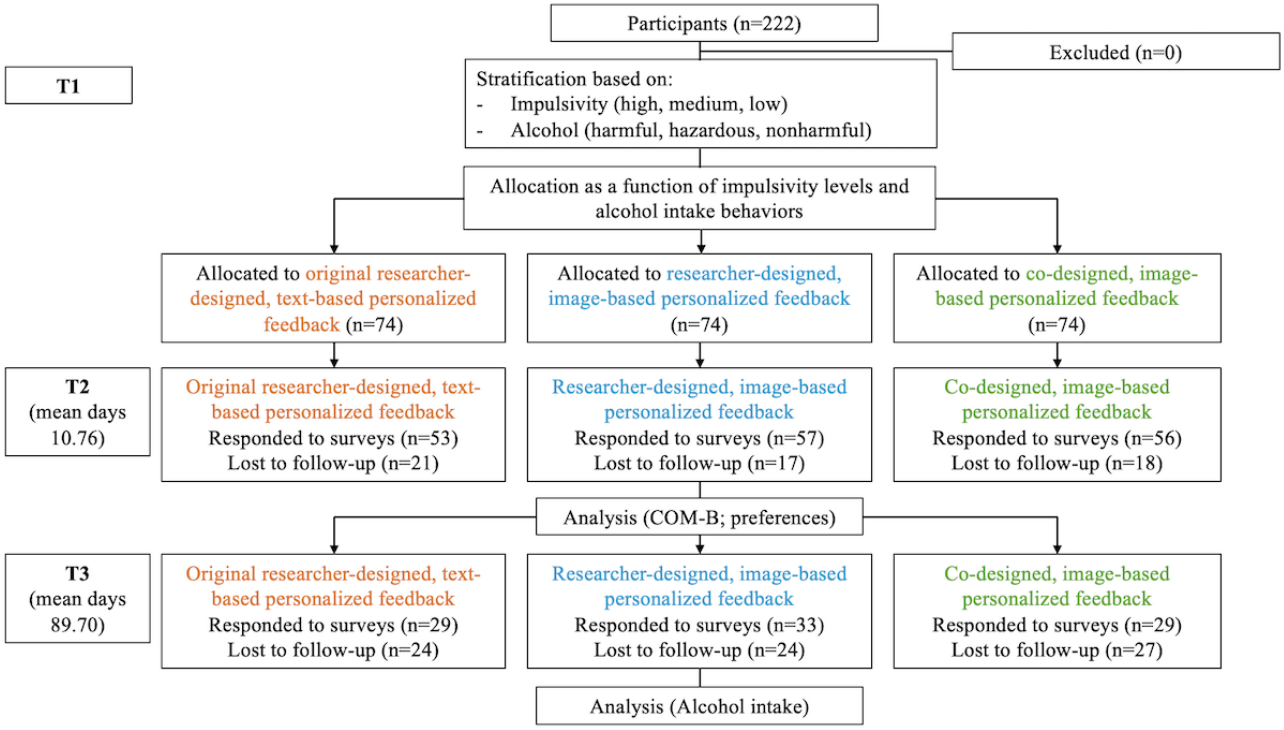

Part 1: Co-Design

A total of 40 participants (mean age 21.87, SD 7.51 years; range 17-47 years; 20/40, 50% female) were involved in 7 focus groups (90 minutes each). A total of 24 participants were aged <20 years, 10 were aged 20-25 years, and 6 were aged >25 years. Average years of education was 13.78 (SD 2.12) years, and the majority (34/40, 85%) hailed from capital cities or large metropolitan areas. In terms of age at first drink, 15% (6/40) had never consumed alcohol, 37.5% (15/40) had consumed their first full drink at age ≤15 years, 27.5% (11/40) at age 16 or 17 years, and 20% (8/40) at age ≥18 years.

XSL•FO
RenderX

Qualitative analysis identified 6 overarching themes, 4 of which centered on presenting information graphically. Theme 1: visual representation. This theme was best summarized as "people love infographics" and "graphics grab attention." Theme 2: clarity. Participants stated, "visuals must be clear, simple and effective" and "a confusing graphic is as bad as a chunk of text." Theme 3: accessibility. Participants suggested, "People are unwilling to work hard to understand something" and "communication must be easy to comprehend." Theme 4: brevity. This was best encapsulated by "wordiness is off-putting." The final 2 themes related to content. Theme 5: peer comparison. Participants wanted to know where they stood in relation to their peers (not the general population). A participant indicated:

> *I think also comparative data is quite important. I want to know how I'm comparing with the females that are my age to see if I am below average, average, because obviously, male, female, we can drink different amounts and feel different ways.*

Theme 6: harm. Participants wanted access to more information about the effects of alcohol, including both long- and short-term effects. They asked for "more info about consequences of drinking small amounts regularly compared to binge drinking" and "more info on alcohol-related diseases." The researchers identified a preconception they had brought—that youth would only be interested in the short-term effects of alcohol misuse, such as hangovers and poor academic performance. However, many young participants expressed the desire for better education about the long-term physical and mental effects of alcohol use, along with more information about the short-term effects of regular drinking below the binge-drinking threshold (≤4 drinks consumed in 1 episode, as per Australian guidelines) [60].

Researchers additionally asked for advice about how to improve the presentation of the feedback examples provided. Reactions to the original researcher-designed, text-based feedback (Figure 1) used in the pilot study [6] were uniformly negative; it was repeatedly referred to as "a chunk of text" and "dense." Of the 3 researcher-designed, image-based feedback examples related to alcohol intake, the graphic in Figure 2 was strongly preferred across all groups (see Figure S1 in Multimedia Appendix 1 for all examples). Participants liked this graphic because they thought it was easy to navigate, well-organized, and appeared logical. Nonetheless, these researcher-designed, image-based examples were also labeled "overwhelming" by 3 participants and were generally considered "confusing." Many participants indicated no preferences across examples, although they agreed they were better than the original researcher-designed text.

Regarding the researcher-designed, image-based feedback examples related to brain health (impulsivity), there was widespread agreement that the visual was an improvement, although most felt the original researcher-designed, text-based feedback was also easy to understand. There was consensus that the use of percentages would be better than percentiles, as participants felt that most people are more likely to understand percentages. The majority preferred the example in Figure 3 (see Figure S1 in Multimedia Appendix 1 for all examples) because it felt the most "accessible." The researcher-designed, image-based feedback example concerning general information (Figure 4) was widely viewed as "overwhelming." However, most participants still preferred it over the original researcher-designed text: "I think it's much better than reading that slab of information." While participants agreed this information was important, they were ambivalent regarding how best to present it.

Table S4 in Multimedia Appendix 1 provides full details of specific suggestions for elements to include in co-designed feedback. Co-designed, image-based feedback was developed from these suggestions, based on majority decisions by the focus groups. As such, the co-designed feedback about alcohol intake incorporated the following focus group recommendations: use graphics, such as drinking icons (especially beer glasses), as a visual representation of drinks; use red and bold text to emphasize drinking above recommended levels; use traffic-light coloring to indicate nonharmful, hazardous, and harmful drinking; use a pie chart to show frequency of drinking; use bullet-point formatting for text; and use a side-by-side layout to juxtapose personalized drinking against national recommendations. The co-designed feedback about brain health, specifically impulsivity, included the following: ensure that colors are clear and contrasting; have a "you are here" indicator with an arrow and red font; split the graph into low, medium, and high categories; color code the risk levels; and use bullet-point formatting for text. General information about the links between impulsivity and alcohol use incorporated the following: use a Venn diagram, simplify language, use colors and bold text, use bullet-point formatting, present information horizontally, and use the active voice. Figure 5 shows the co-designed, image-based feedback.

Part 2: Efficacy

Descriptive Analyses

A total of 222 participants (mean age 21.14, SD 4.69 years; range 18-68 years; 176/222, 79% female) completed T1 (baseline). Of these, 166 participants (mean age 21.60, SD 5.27 years; range 18-68 years; 129/166, 78% female) completed T2 (166/222, 74.77% of the T1 sample) 11 days later, and of these, 91 (mean age 22.24, SD 6.07 years; range 18-68 years; 75/91, 82% female) returned to undertake T3 (91/166, 54.82% of the T2 sample) 3 months later. Figure 7 provides the participation flowchart.

Tables 1 and 2 show demographic statistics for participants who completed T1 (baseline) and T2 of Part 2. Demographics as a function of feedback group allocation (original researcher-designed, text-based; researcher-designed, image-based; and co-designed, image-based) are also detailed. There were no significant differences on these variables as a function of feedback group. Table 3 shows alcohol intake behavior and impulsivity at T1. There were also no significant differences in these variables as a function of feedback group.

XSL•FO
RenderX

Table 1. Part 2: continuous demographic statistics at T1 for all participants who completed T1 and T2, and as a function of original researcher-designed, text-based; researcher-designed, image-based; and co-designed, image-based feedback group allocation.

Variable	Total (n=166), mean (SD)	Original researcher-designed, text-based feedback (n=53), mean (SD)	Researcher-designed, image-based feedback (n=57), mean (SD)	Co-designed, image-based feedback (n=56), mean (SD)	*F* test (*df*)	*P* value	η_p^2
Age (years)	21.60 (5.27)	20.91 (3.76)	22.68 (7.29)	21.16 (3.75)	1.88 (2, 163)	.16	0.02
Education (years)	14.50 (1.85)	14.36 (1.81)	14.85 (1.96)	14.27 (1.73)	1.64 (2, 163)	.20	0.02
APM-6[a]	3.61 (1.65)	3.57 (1.65)	3.83 (1.66)	3.45 (1.65)	0.77 (2, 163)	.46	0.01
GAD-7[b]	6.86 (5.65)	6.83 (5.74)	6.88 (5.09)	6.86 (6.18)	<0.01 (2, 163)	>.99	<0.01
PHQ-9[c]	8.13 (6.13)	7.17 (6.07)	8.93 (6.25)	8.23 (6.06)	1.14 (2, 163)	.32	0.01

[a]APM-6: Advanced Progressive Matrices-6 (abridged 6-item version of Raven's APM).

[b]GAD-7: Generalized Anxiety Disorder-7.

[c]PHQ-9: Patient Health Questionnaire-9.

Table 2. Part 2: categorical demographic statistics at T1 for all participants who completed T1 and T2, and as a function of original researcher-designed, text-based; researcher-designed, image-based; and co-designed, image-based feedback group allocation.

Variable	Total (n=166), n (%)	Original researcher-designed, text-based feedback (n=53), n (%)	Researcher-designed, image-based feedback (n=57), n (%)	Co-designed, image-based feedback (n=56); n (%)	Chi-square (*df*)	*P* value	Cramer V
Sex ratio (male:female:other)	36 (21.7):129 (77.7):1 (0.6)	12 (22.6):41 (77.4):0 (0)	14 (24.6):43 (75.4):0 (0)	10 (17.8):45 (80.4):1 (1.8)	2.7 (4)	.61	0.09
Country of birth					7.3 (8)	.50	0.15
Australasia	76 (45.8)	29 (54.7)	22 (38.6)	25 (44.6)			
Asia	76 (45.8)	19 (35.9)	32 (56.1)	25 (44.6)			
Europe	9 (5.4)	4 (7.6)	2 (3.5)	3 (5.4)			
Americas	4 (2.4)	1 (1.9)	1 (1.8)	2 (3.6)			
Other	1 (0.6)	0 (-)	0 (-)	1 (1.8)			
First language					5.0 (2)	.08	0.17
English	109 (65.7)	41 (77.4)	33 (57.9)	35 (62.5)			
Other	57 (34.3)	12 (22.6)	24 (42.1)	21 (37.5)			

XSL•FO
RenderX

Table 3. Part 2: alcohol intake behavior and impulsivity at T1 for all participants who completed T1 and T2, and as a function of original researcher-designed, text-based; researcher-designed, image-based; and co-designed, image-based feedback group allocation.

Variable	Total (n=166), mean (SD)	Original researcher-designed, text-based feedback (n=53), mean (SD)	Researcher-designed, image-based feedback (n=57), mean (SD)	Co-designed, image-based feedback (n=56), mean (SD)	*F* test (*df*)	*P* value	η_p^2
AUDIT[a] score	9.32 (5.12)	9.98 (5.03)	8.93 (5.25)	9.09 (5.11)	0.66 (2, 163)	.52	0.01
Timeline Followback							
Total standard drinks	17.61 (18.60)	17.42 (13.22)	16.79 (17.65)	18.64 (23.54)	0.14 (2, 163)	.87	<0.01
Drinking days (%)	35.09 (21.19)	32.47 (16.87)	39.18 (24.10)	33.41 (21.44)	1.65 (2, 163)	.20	0.02
Days with ≥4 drinks	1.78 (2.30)	1.81 (1.76)	1.61 (2.32)	1.93 (2.73)	0.27 (2, 163)	.77	<0.01
MCQ[b]							
Overall *k*	0.027 (0.047)	0.022 (0.030)	0.023 (0.042)	0.036 (0.061)	1.86 (2, 163)	.16	0.02
SST[c]							
SSRTintegration[d]	238.70 (51.39)	237.45 (48.53)	246.64 (41.44)	231.65 (62.20)	1.09 (2, 163)	.34	0.02
Impulsivity index	48.63 (21.82)	46.17 (22.48)	49.28 (21.25)	50.31 (22.02)	0.46 (2, 163)	.63	0.01

[a]AUDIT: Alcohol Use Disorder Identification Test.

[b]MCQ: Monetary Choice Questionnaire.

[c]SST: stop signal task.

[d]SSRTintegration: stop signal reaction time calculated using the integration method.

COM-B:AIQ as a Function of Feedback Group

At T2, 77.1% (128/166) and 22.9% (38/166) of the sample, respectively, indicated that they had been either "completely honest" or "mostly honest" when responding to T1 questions about alcohol consumption. No participants selected the "mostly not honest" or "not at all honest" options.

Table 4 shows scores for each of capacity, opportunity, and motivation for behavior change, as well as total COM-B:AIQ score, both for the total sample and as a function of feedback group allocation. There were no differences between groups on total COM-B:AIQ score, $F_{2,163}$=0.77; *P*=.47; η_p^2=0.01, or on subsections related to capacity, $F_{2,163}$=0.73; *P*=.49; η_p^2=0.01, opportunity, $F_{2,163}$=0.68; *P*=.51; η_p^2=0.01, or motivation, $F_{2,163}$=0.61; *P*=.55; η_p^2=0.01.

Table 4. Part 2: Capacity, Opportunity, Motivation–Behavior: Alcohol Intake Questionnaire (COM-B:AIQ) scores at T2 for all participants who completed T1 and T2, and as a function of original researcher-designed, text-based; researcher-designed, image-based; and co-designed, image-based feedback group allocation.

COM-B:AIQ component	Total (n=166), mean (SD)	Original researcher-designed, text-based feedback (n=53), mean (SD)	Researcher-designed, image-based feedback (n=57), mean (SD)	Co-designed, image-based feedback (n=56), mean (SD)
Capacity	27.31 (3.48)	27.76 (3.40)	26.97 (2.93)	27.23 (14.04)
Opportunity	22.37 (3.73)	22.62 (3.42)	22.60 (3.48)	21.89 (4.24)
Motivation	20.63 (3.79)	20.68 (3.44)	21.00 (3.65)	20.21 (4.26)
COM-B[a] total	70.31 (7.47)	71.06 (6.25)	70.56 (6.78)	69.34 (9.07)

[a]COM-B: Capacity, Opportunity, Motivation–Behavior.

To further interrogate responses on the COM-B:AIQ, we considered how participants answered questions specifically addressing Australian national drinking guidelines [60]. In total, 23.5% (39/166) of the sample correctly answered the question probing how many total standard drinks could safely be consumed per week (10 drinks). There was no difference in correct responses as a function of feedback group allocation, χ^2_2=1.2; *P*=.55; V=0.09. Only 15.7% (26/166) correctly

XSL•FO
RenderX

answered the question regarding safe daily consumption (≤4 drinks). Again, there was no difference in correct responses as a function of feedback group allocation, χ^2_2=2.3; *P*=.31; V=0.12. As focus group participants had specifically indicated percentiles might be difficult for people to interpret, we also examined responses to the COM-B:AIQ question ascertaining knowledge around percentiles. In all, 59% (98/166) of the sample answered the question correctly. There was no difference in correct responses as a function of feedback group allocation, χ^2_2=2.3; *P*=.32; V=0.12.

Preferred Feedback Format as a Function of Feedback Group

Participants undertook T2 testing 10.76 (SD 14.05) days after completing T1 (and following delivery of feedback). Table 5 details rating, or liking, of each feedback type, both for the whole sample and as a function of feedback group allocation. A mixed measures ANOVA showed a significant main effect of feedback type, $F_{2,326}$=38.11; *P*<.001; η_p^2=0.19. Pairwise comparisons revealed participants preferred co-designed, image-based over the researcher-designed, image-based (*P*<.001; *d*=0.55) and original researcher-designed, text-based (*P*<.001; *d*=0.83) feedback. There was no main effect of feedback group allocation, $F_{2,163}$=2.18; *P*=.12; η_p^2=0.03, and no interaction between rating and feedback group, $F_{4,326}$=1.84; *P*=.12; η_p^2=0.02.

Table 5. Part 2: preferred feedback format at T2 for all participants who completed T1 and T2, and as a function of original researcher-designed, text-based; researcher-designed, image-based, and co-designed, image-based feedback group allocation.

Feedback format	Total (n=166), mean (SD)	Original researcher-designed, text-based feedback (n=53), mean (SD)	Researcher-designed, image-based feedback (n=57), mean (SD)	Co-designed, image-based feedback (n=56), mean (SD)
Original text-based	20.19 (5.80)	20.13 (5.38)	21.04 (5.62)	19.39 (6.32)
Researcher-designed image-based	21.69 (5.62)	20.81 (5.87)	23.09 (4.97)	21.11 (5.83)
Co-designed	24.65 (4.66)	23.62 (5.82)	24.77 (4.26)	25.48 (3.58)

Change in Alcohol Consumption as a Function of Feedback Group

Participants undertook T3 testing 89.70 (SD 21.59) days after completing T2. In terms of return rate, 54.72% (29/53) of those allocated to the original researcher-designed, text-based feedback group undertook T3 testing, 57.89% (33/57) of those allocated to the researcher-designed, image-based group, and 51.79% (29/56) of those allocated to the co-designed, image-based feedback group.

With regard to the change in AUDIT scores from T1 (mean 9.32, SD 5.12) to T3 (mean 8.53, SD 5.44), there was a significant main effect of time, $F_{1,88}$=4.80; *P*=.03; η_p^2=0.05, representing an 8.5% reduction in harm. There was no significant main effect of feedback group allocation, $F_{2,88}$=0.08; *P*=.92; η_p^2<0.01, and no interaction between time and feedback group, $F_{2,88}$=0.24; *P*=.79; η_p^2=0.01 (Figure 8A).

XSL•FO
RenderX

Figure 8. Change in (A) Alcohol Use Disorders Identification Test (AUDIT) score, (B) total standard drink intake, (C) frequency of drinking, (D) occasions on which 4 or more standard drinks were consumed, and (E) impulsivity following personalized feedback regarding alcohol use and impulsivity from T1 to T3 (89.70 days).

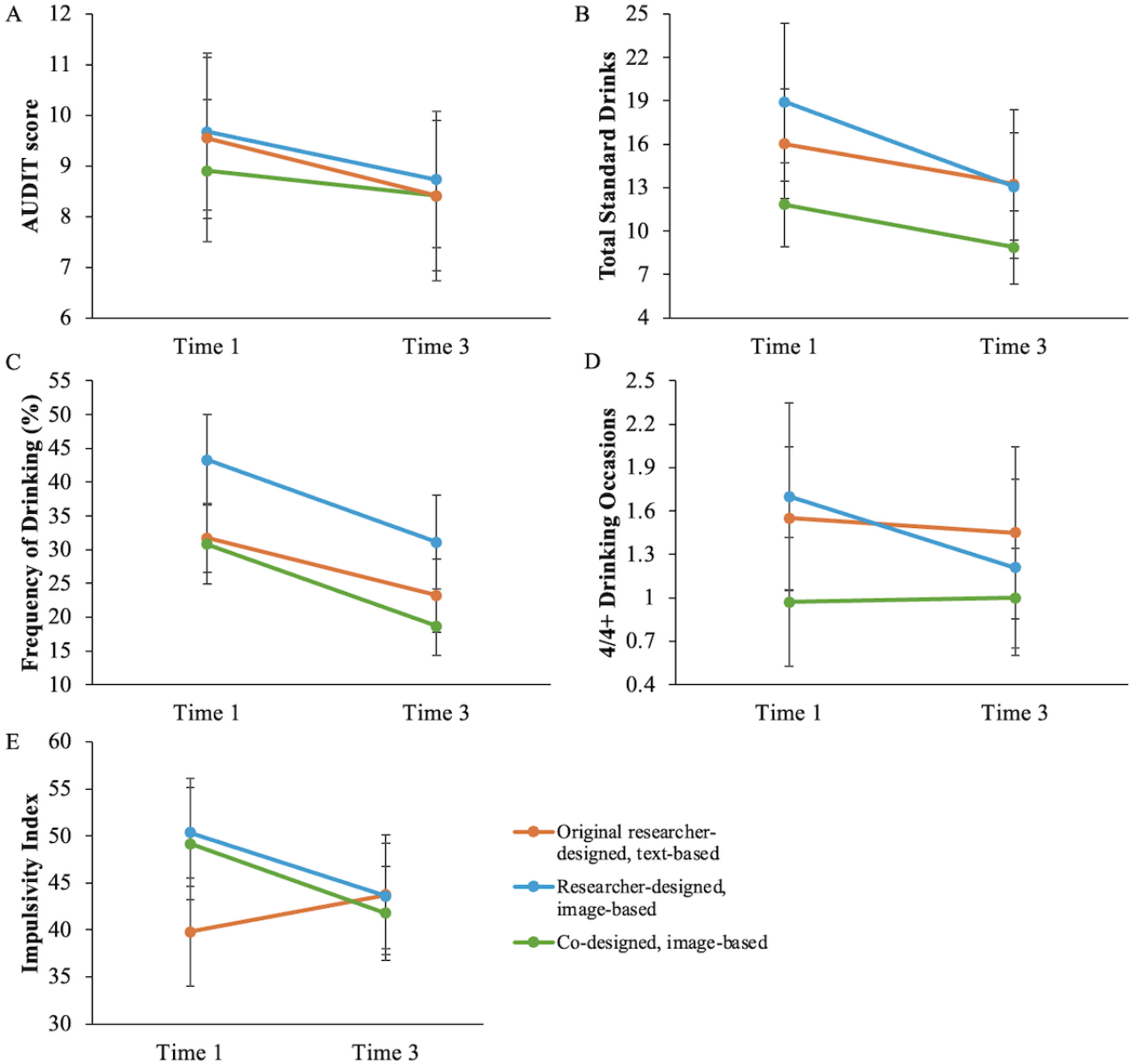

Regarding alcohol intake variables derived from Timeline Followback, there was a significant main effect of time, $F_{1,88}$=10.98; P<.001; η_p^2=0.11, on total standard drinks consumed from T1 (mean 17.61, SD 18.60) to T3 (mean 11.78, SD 14.43), representing a 33% reduction in consumption. There was no significant main effect of feedback group allocation, $F_{2,88}$=1.31; P=.28; η_p^2=0.03, and no interaction between time and feedback group, $F_{2,88}$=0.75; P=.47; η_p^2=0.02 (Figure 8B).

There was a significant main effect of time, $F_{1,88}$=35.43; P<.001; η_p^2=0.29, on frequency of drinking from T1 (mean 35.09, SD 21.19) to T3 (mean 24.63, SD 21.68), representing a 30% reduction in drinking occasions. There was a significant main effect of feedback group allocation, $F_{2,88}$=3.44; P=.04; η_p^2=0.07. Pairwise comparisons showed participants in the researcher-designed, image-based (P=.002; d=0.56) and co-designed, image-based (P=.005; d=0.56) feedback groups significantly reduced frequency of drinking, whereas those in the original researcher-designed, text-based feedback group (P=.15; d=0.39) did not. There was no interaction between time and feedback group, $F_{2,88}$=0.42; P=.66; η_p^2=0.01 (Figure 8C).

There was no significant main effect of time, $F_{1,88}$=1.31; P=.26; η_p^2=0.02, on occasions drinking ≥4 drinks from T1 (mean 1.78, SD 2.30) to T3 (mean 1.22, SD 1.97), although there was still a 31.5% reduction in episodes. There was no significant main effect of feedback group allocation, $F_{2,88}$=0.74; P=.48; η_p^2=0.02, and no interaction between time and feedback group, $F_{2,88}$=0.96; P=.39; η_p^2=0.02 (Figures 8D).

Change in Impulsivity as a Function of Feedback Group

With regard to the change in average impulsivity index from T1 (mean 46.46, SD 21.72) to T3 (mean 43.12, SD 21.07), there was no significant main effect of time, $F_{1,75}$=2.05; P=.16;

XSL•FO
RenderX

η_p^2=0.03, no significant main effect of feedback group allocation, $F_{2,75}$=0.44; P=.65; η_p^2=0.01, and no interaction between time and feedback group, $F_{2,75}$=2.84; P=.06; η_p^2=0.07 (Figure 8E).

Discussion

Summary, Implications, and Comparison With the Literature

Part 1: Co-Design

In Part 1 of this study, we used a co-design process to develop user-designed personalized feedback about alcohol intake and brain health (especially impulsivity) to incorporate into an existing brief digital intervention that has previously been shown to reduce hazardous drinking [6]. Drawing on research and guidelines that advocate for consumer involvement at multiple stages of the research process [7-9], we consulted with young people when preparing prompts and materials and then conducted a series of focus groups with adolescents and young adults using those same resources.

Following consultation, thematic analysis established that young consumers wanted feedback about alcohol intake and brain health to be visual, clear, accessible, and brief. They disliked wordiness or unduly complicated visual representations. They also wanted more information about how their drinking compared with that of their peers and education regarding the long-term physical and mental health impacts of alcohol use. The text-based researcher-designed feedback used in the original study (original researcher-designed, text-based feedback) as well as the image-orientated researcher-designed feedback developed for use with focus groups (researcher-designed, image-based feedback) already included normative information on how drinking compared with Australian national guidelines. While focus group participants indicated that they wanted this information tailored by age group, evidence suggests that this can lead to increased alcohol consumption among young people who drink below what is considered average for their cohort [61]. We were therefore reluctant to include any peer-specific information that might encourage increased drinking but retained the normative information pertaining to national guidelines.

Recommendations to include information on the long-term physical and mental health impacts of alcohol use were unexpected and revealed genuine interest among our participants in improving education on this front. Attempts to include this level of detail in the co-designed, image-based feedback proved difficult, as the result conflicted with preferences for concise infographics. As such, we determined that this level of detail could be incorporated via links to a (planned) website. At the conclusion of Part 1 of this study, we thus had three versions of brief personalized feedback: (1) original researcher-designed, text-based; (2) researcher-designed, image-based; and (3) co-designed, image-based, suitable for investigations examining preference and efficacy.

Part 2: Efficacy

Contrary to our hypothesis, participants who received co-designed, image-based feedback did not score more highly on the COM-B:AIQ, relative to those who received either the original researcher-designed, text-based or researcher-designed, image-based feedback. Groups did not differ in their knowledge of safe drinking levels, their understanding of the impulsivity percentile information, their social and environmental opportunities, or their motivation to change their drinking behavior. Worryingly, irrespective of feedback format, a substantial proportion of the sample remained unaware of guidelines for recommended weekly and maximum daily intake. Large cross-sectional surveys suggest only 32% to 40% of Australians can accurately respond to questions about drinking limits, with younger people demonstrating less understanding than older ones [62]. Several studies have shown even less awareness of daily drinking limits [62,63]. Although all feedback types clearly articulated weekly and daily intake guidelines, the predominance of younger people in our sample may partly account for the relatively low scores on these capacity-based questions. In contrast, and counter to feedback from focus group participants, understanding of percentiles was much higher, although again unrelated to feedback format. As participants were recruited largely from a university student cohort, knowledge related to percentiles might be expected to be higher than in the general population.

As hypothesized, participants preferred the co-designed, image-based feedback over both the original researcher-designed, text-based and the researcher-designed, image-based feedback options. This preference did not vary according to the type of feedback they initially received (as part of investigations into how the different types of feedback impacted consumption) or the order in which they viewed the 3 versions; across all groups, participants rated the co-designed, image-based feedback as having the most appealing look and feel. It is worth noting that no participants involved in Part 1 of this study were recruited into Part 2. Additionally, participants in Part 2 were not informed about how each type of feedback had been generated. Effect sizes indicate that the co-designed feedback was substantially preferred over the researcher-designed, image-based feedback and strongly preferred to the original researcher-designed, text-based feedback. This finding aligns with input from our focus groups, who favored the researcher-designed, image-based feedback over the original researcher-designed, text-based feedback, and is consistent with previous findings that co-designed products are, overall, well-received and well-liked [19,20,64].

Although the co-designed feedback was preferred, there were few significant differences in subsequent alcohol intake behavior as a function of feedback condition. As hypothesized, and replicating the findings of the intervention pilot study [6], there were significant overall reductions in alcohol intake. Across conditions, total standard drinks decreased and there was also a significant reduction in alcohol-related harm. This contributes further evidence for the efficacy of personalized feedback in targeting hazardous drinking, although requires further testing in fully randomized trials. While occasions when 4 or more drinks were consumed in 1 episode decreased across groups,

XSL•FO
RenderX

this change was not significant, again aligning with pilot study findings. Also in keeping with the pilot, and as anticipated, there was no change in impulsivity across time.

One group-based difference emerged: participants in researcher-designed, image-based and co-designed, image-based groups reduced their frequency of drinking, whereas those in the researcher-designed, text-based group did not. In the pilot, which used only the original, text-based personalized feedback, there was likewise no significant reduction in drinking frequency. The present findings thus suggest that visual feedback may be more effective than text-based feedback in reducing how often people drink. This is unsurprising given the reactions of focus group participants to the original researcher-designed, text-based feedback and their strong preferences for the visual examples provided. However, this finding does not directly support the greater efficacy of co-designed, image-based feedback compared with that designed without consumer input, as the groups who received researcher-designed, image-based and co-designed, image-based feedback reduced their drinking frequency by the same amount.

Integration and Future Direction

We used participatory design principles during the preparation phase, throughout the focus groups, and in feedback design. When consulting the Youth Advisory Board at the Matilda Centre, we sought advice on co-design best practice, particularly on facilitating engaging, inclusive, and collaborative focus groups. We aimed to avoid common pitfalls identified in the literature, such as mistaking simple consultation for co-design and limiting participants' decision-making power or creativity [13]. It is important to note that, although the Youth Advisory Board recommended providing visual feedback examples to the focus groups, they did not contribute to the actual design of the researcher-designed, image-based feedback. This ensured it genuinely reflected our unassisted attempts. Finally, we drew on the recommendations of our focus group participants to develop fully co-designed feedback about alcohol intake, brain health, and how these impact vulnerability for alcohol misuse and alcohol use disorder. Consistent with previous findings, our participants expressed a strong preference for feedback developed via co-design [18-20]. Researchers have noted, however, a paucity of research examining the actual effectiveness of co-designed products and interventions [11,12,25,26]. To address this critical gap, we explored the relative impact of co-designed, image-based feedback across several outcomes, including its influence on participants' capacity, opportunity, and motivation towards changing their drinking behavior, as well as their subsequent alcohol intake. Notably, co-designed, image-based feedback was no more efficacious than researcher-designed (image- or text-based) feedback in reducing alcohol consumption.

Bevan Jones et al [26] have suggested that future research should explore how academics experience co-design, as most existing work focuses on participant perspectives. Research should consider whether the time, effort, and resources required for co-design are well spent, and how participatory methods can be improved to ensure the process benefits both developers and end users. This study adds a qualifying voice to the widespread, yet largely unevidenced, enthusiasm for co-design. Nonetheless, the improved likeability of our intervention is a valuable outcome of the co-design process: unless people like the look and feel of a product, they may not use it. Although it is less vital in the context of once-off personalized feedback, preference is especially important for products, such as health apps, that people interact with on a repeated and regular basis. Co-design is arguably essential (on both ethical and pragmatic grounds) when designing tools for people with specific conditions and needs or in specialist contexts. Further research is required to guide the incorporation of co-design within specific and varied scenarios, to ensure it adds sufficient value to the end product and to justify its costs in terms of researcher resources.

Limitations

Some limitations of this study must be noted. Although we endeavored to use participatory principles in Part 1, it is important to recognize that focus groups were supplied with some existing materials. They saw the original text-based personalized feedback as well as examples designed by researchers on the advice of the Youth Advisory Board. This may have influenced their responses and the recommendations they made around the co-designed, image-based feedback. While this challenge is inherent to all participatory processes, future research may seek to address how researchers might tailor studies to maximize participatory input and minimize researcher influence.

In Part 2, returning T2 participants completed the COM-B:AIQ and were then shown all 3 formats of feedback (in counterbalanced order) and asked about preferences. This might have introduced contamination bias and may explain findings at T3 showing participants in all 3 feedback conditions reduced quantity of drinking and harmful alcohol use. Importantly, although participants did see all 3 feedback formats at T2, it was not personalized. No details pertaining to alcohol intake or impulsivity levels were included. Moreover, some condition-based differences were evident at T3. Participants in image-based feedback conditions (researcher-designed and co-designed) reported significant reductions in frequency of drinking, whereas those in the text-based (researcher-designed) condition showed no significant change in drinking frequency. It seems unlikely that having seen all 3 nonpersonalized formats of the feedback had a greater impact on those who received personalized image-based feedback—and only on frequency of drinking—over those who received text-based feedback. Nonetheless, future work will involve a randomized design to minimize the potential for contamination bias.

We did not have a no-feedback control group. As such, we must interpret findings around reductions in alcohol intake behavior and harm cautiously. Nonetheless, there was a control condition in the pilot study, and we saw similar reductions in total alcohol intake, occasions drinking 4 or more drinks in 1 episode, and harm [6]. We also did not ask participants to complete the COM-B:AIQ at T1 (prior to providing feedback) and then again at T2 (following feedback delivery). Thus, we could not explore whether the format of feedback moderated change in capacity, opportunity, or motivation to alter alcohol intake behavior over time. In addition, as with any creative production, it was not

XSL•FO
RenderX

possible to consider all the conceivable impacts of different design choices and how these might have influenced the efficacy of co-designed feedback. The same research team developed the co-designed, image-based and researcher-designed, image-based feedback (using only the graphic tools of PowerPoint); thus, the only difference between the 2 types of feedback was input from the focus groups. This was an intentional decision, to control for other factors that could affect the quality of the designs, such as different tools, or designers with varied skill levels. So, although this may have weakened the impact of the co-designed feedback, it demonstrated the influence of stakeholder input on the design process while holding constant other factors.

While our Part 1 sample comprised equal numbers of male and female participants, a higher proportion of female participants undertook Part 2. Recruitment for Parts 1 and 2 involved advertising on social media and via the University of Melbourne student portal. For Part 2, however, students enrolled in first year psychology subjects could secure course credit for participation. Consequently, our Part 2 sample reflects the demographics of this largely female, high school educated cohort; thus, we must be cautious about the generalizability of findings. Female undergraduate students may not be representative of a group motivated to reduce their alcohol consumption. Community or other samples concerned about their alcohol intake may have been more receptive toward the information presented in the feedback. The sample nonetheless did at least drink as much (or more) than a sample drawn from the general population. At T1, 63% (140/222) of our sample reported binge drinking (≥4 drinks per episode) in the past 2 weeks (compared with 24% of Australians who reported binge drinking at least once a month, and 21% of people aged 18-24 years who reported binge drinking at least weekly). Also, at T1, 29% (64/222) of our sample reported drinking more than 10 standard drinks per week on average (compared with 25% of all Australians, and 28% of those aged 18-24 years who reported consuming more than 10 drinks per week on average) [65]. Further, all participants reported being completely honest or mostly honest about their alcohol intake; so, although an undergraduate university sample may not fully represent the wider community, the research is not limited by insufficient alcohol intake levels for reductions to be observed.

Finally, although 75% (166/222) of T1 participants returned for T2 (comparable to the original pilot where 78% returned for T2), only 55% (91/166) of those returned for T3, and the majority were female (75/91, 82%). This level of attrition was partly due to the nature of collecting data from a largely tertiary cohort. T1 and T2 testing took place within the same semester (and completing the protocols was linked to course credit). By contrast, T3 testing occurred in a subsequent semester (and reimbursement was via eGift vouchers). As such, a proportion of students who completed T1 and T2 could have been on leave or would no longer have been enrolled. Such students would not have received the email invite for T3. Others may have been less motivated to continue with the study given participation was no longer linked to course credit. We must therefore be cautious about the generalizability of findings given the level of attrition at T3.

Conclusions

Using a mixed methods design, we collaborated with consumers to develop co-designed personalized feedback about alcohol intake and brain health to incorporate into an existing brief digital intervention that has previously been shown to successfully reduce hazardous drinking. While participants preferred the co-designed, image-based feedback, it did not appear to influence their capacity, opportunity, or motivation to alter their alcohol consumption behavior. Nonetheless, regardless of the type of feedback participants received, they all significantly reduced their total alcohol intake. There was a significant reduction in harm related to alcohol use and a nonsignificant reduction in binge drinking episodes. Participants in image-based feedback conditions also significantly reduced frequency of drinking. These findings suggest that personalized feedback can be effective in reducing alcohol use and harm even without extensive co-design enhancements. Given co-designed products and interventions require substantial investment of time, resources, and financial commitment, their added value in terms of efficacy should not be assumed. Future research should identify when and for whom co-design yields meaningful gains in effectiveness to ensure the considerable costs associated with this approach are justified and that limited resources are allocated optimally.

Acknowledgments

The authors would like to acknowledge and thank the Matilda Centre Youth Advisory Board along with all the young people and experts for their meaningful engagement and valuable contributions. We acknowledge Aboriginal and Torres Strait Islander peoples of the unceded land on which we work, learn and live. We pay respect to Elders past, present and future, and acknowledge the importance of Indigenous knowledge in the academy.

Data Availability

The datasets generated or analyzed in this study are available in the Open Science Framework [66].

Funding

This study was funded by a National Health and Medical Research Council (NHMRC) Medical Research Future Fund (MRFF) Early to Mid-Career Researcher Grant (#2031836) and a University of Melbourne Medicine, Dentistry and Health Sciences (MDHS) Innovation Seed Grant (#000080).

XSL•FO
RenderX

Authors' Contributions

Conceptualization was undertaken by AP and RH. Data curation was performed by JB. Formal analysis was conducted by AP, JB, and CC. Funding acquisition was carried out by AP, GD, KP, EKD, LB, LAS, and RH. Investigation was conducted by AP and JB. Methodology was developed by AP and JB. Project administration was handled by AP and JB. Supervision was provided by LAS and RH. Validation was undertaken by AP, GD, and RH. Visualization was performed by AP and JB. Writing of the original draft was completed by AP and JB. Writing (review and editing) was carried out by GD, CC, KP, EKD, LB, LAS, and RH.

Conflicts of Interest

None declared.

Editorial Notice

This trial was not registered on an International Committee of Medical Journal Editors (ICJME)–approved clinical trial registry. However, the study protocol was registered retrospectively on the Open Science Framework web platform [59]. The authors attributed this to the exploratory nature of the study, which was designed to guide the development of an intervention for a larger trial, involved a nonclinical population, and had a primary aim of comparing user preferences for co-designed vs researcher-designed feedback on alcohol-related behaviors. The editor granted an exception to the ICMJE requirement for prospective registration of randomized trials on the grounds that risk of bias appears low and the study is considered formative. Readers are nonetheless advised to carefully evaluate any explicit or implicit claims related to primary outcomes or effectiveness, as retrospective registration does not preclude post hoc changes to outcome measures.

Multimedia Appendix 1
Additional data.
[DOCX File , 1322 KB - jmir_v28i1e87393_app1.docx]

References

1. Global status report on alcohol and health and treatment of substance use disorders. World Health Organisation. 2024. URL: https://www.who.int/publications/i/item/9789240096745 [accessed 2025-11-07]
2. Whetton S, Tait RJ, Gilmore W, Dey T, Agramunt S, Abdul Halim S. Examining the social and economic costs of alcohol use in Australia: 2017/18. In: National Drug Research Institute report (Curtin University publication). Perth: National Drug Research Institute, Curtin University; 2021.
3. Sundström C, Schell C, Wardell JD, Godinho A, Cunningham JA. Can brief email guidance enhance the effects of an internet intervention for people with problematic alcohol use? A randomized controlled trial. Subst Use Misuse 2020;55(12):2011-2019 [FREE Full text] [doi: 10.1080/10826084.2020.1788087] [Medline: 32662699]
4. Lee CM, Cadigan JM, Calhoun BH, Cronce JM, Fairlie AM, Rhew IC. A just-in-time adaptive personalized mobile app intervention produces limited short-term reductions in heavy episodic drinking frequency in college students: a randomized control trial during the COVID-19 pandemic. Alcohol Clin Exp Res (Hoboken) 2025;49(6):1321-1336. [doi: 10.1111/acer.70056] [Medline: 40285370]
5. Tanner-Smith EE, Lipsey MW. Brief alcohol interventions for adolescents and young adults: a systematic review and meta-analysis. J Subst Abuse Treat 2015;51:1-18 [FREE Full text] [doi: 10.1016/j.jsat.2014.09.001] [Medline: 25300577]
6. Poulton A, Dali G, McGinness C, Clarke S, Turner T, Liu J. Impact of personalized alcohol intake and cognitive feedback on alcohol use behavior in hazardous drinkers: a quasi-randomized trial. Alcohol Clin Exp Res (Hoboken) 2023;47(7):1377-1389. [doi: 10.1111/acer.15098] [Medline: 37428421]
7. Australian medical research and innovation priorities 2024-2026. Australian Medical Research Advisory Board. 2024. URL: https://www.health.gov.au/resources/publications/australian-medical-research-and-innovation-priorities-2024-2026 [accessed 2025-11-07]
8. Principles for consumer involvement in research funded by the medical research future fund. Medical Research Future Fund Consumer Reference Panel. 2023. URL: https://www.health.gov.au/sites/default/files/2023-03/principles-for-consumer-involvement-in-research-funded-by-the-medical-research-future-fund.pdf [accessed 2026-05-23]
9. Statement on consumer and community involvement in health and medical research. National Health and Medical Research Council. 2016. URL: https://www.nhmrc.gov.au/about-us/publications/statement-consumer-and-community-involvement-health-and-medical-research [accessed 2025-11-07]
10. Javanparast S, Robinson S, Kitson A, Arciuli J. Embedding research codesign knowledge and practice: learnings from researchers in a new research institute in Australia. Res Involv Engagem 2022;8(1):71 [FREE Full text] [doi: 10.1186/s40900-022-00392-4] [Medline: 36476374]
11. Blomkamp E. The promise of co‐design for public policy. Aust J Public Adm 2018;77(4):729-743. [doi: 10.1111/1467-8500.12310]

XSL•FO
RenderX

12. Thabrew H, Fleming T, Hetrick S, Merry S. Co-design of eHealth interventions with children and young people. Front Psychiatry 2018;9:481 [FREE Full text] [doi: 10.3389/fpsyt.2018.00481] [Medline: 30405450]
13. Blanchard M, Fava N. Co-designing new solutions to support young people's mental health and wellbeing. In: White MA, Slemp GR, Murray AS, editors. Future Directions in Well-Being: Education, Organizations and Policy. Cham: Springer International Publishing; 2017:197-202.
14. Malloy J, Partridge SR, Kemper JA, Braakhuis A, Roy R. Co-design of digital health interventions with young people: a scoping review. Digit Health 2023;9:20552076231219117 [FREE Full text] [doi: 10.1177/20552076231219117] [Medline: 38107986]
15. Chinsen A, Berg A, Nielsen S, Trewella K, Cronin TJ, Pace CC, et al. Co-design methodologies to develop mental health interventions with young people: a systematic review. Lancet Child Adolesc Health 2025;9(6):413-425. [doi: 10.1016/S2352-4642(25)00063-X] [Medline: 40311649]
16. Greene Barker T, O'Higgins A, Fonagy P, Gardner F. A systematic review and meta-analysis of the effectiveness of co-designed, in-person, mental health interventions for reducing anxiety and depression symptoms. J Affect Disord 2024;350:955-973 [FREE Full text] [doi: 10.1016/j.jad.2023.12.080] [Medline: 38199405]
17. Hawke LD, Sheikhan NY, Bastidas-Bilbao H, Rodak T. Experience-based co-design of mental health services and interventions: a scoping review. SSM - Mental Health 2024;5:100309. [doi: 10.1016/j.ssmmh.2024.100309]
18. Carrotte E, Hopgood F, Blanchard M, Groot C, Phillips L. A new podcast for reducing stigma against people living with complex mental health issues: co-design study. JMIR Form Res 2023;7:e44412 [FREE Full text] [doi: 10.2196/44412] [Medline: 37145860]
19. Garner K, Thabrew H, Lim D, Hofman P, Jefferies C, Serlachius A. Exploring the usability and acceptability of a well-being app for adolescents living with type 1 diabetes: qualitative study. JMIR Pediatr Parent 2023;6:e52364 [FREE Full text] [doi: 10.2196/52364] [Medline: 38153804]
20. McMullen CA, Williams MV, Smyth SS, Clouser JM, Li J. Co-designing and piloting educational materials with patients and healthcare providers for syncope in the emergency department. PEC Innov 2023;2:100131 [FREE Full text] [doi: 10.1016/j.pecinn.2023.100131] [Medline: 37214525]
21. Ito-Jaeger S, Perez Vallejos E, Curran T, Crawford P. What's up with everyone? A qualitative study on young people's perceptions of cocreated online animations to promote mental health literacy. Health Expect 2022;25(4):1633-1642 [FREE Full text] [doi: 10.1111/hex.13507] [Medline: 35507731]
22. Debenham J, Birrell L, Newton NC, Devine EK, Champion KE, Stapinski LA. Co-design of digital public health substance use resources: a collaboration between young people and experts. Health Promot J Austr 2025;36(1):e874. [doi: 10.1002/hpja.874] [Medline: 38797528]
23. Thabrew H, Kumar H, Steadman E. Acceptability and feasibility of "Village," a digital communication app for young people experiencing low mood, thoughts of self-harm, and suicidal ideation to obtain support from family and friends: mixed methods pilot open trial. JMIR Form Res 2023;7:e41273 [FREE Full text] [doi: 10.2196/41273] [Medline: 36912882]
24. Newton NC, Conrod PJ, Slade T, Carragher N, Champion KE, Barrett EL. The long-term effectiveness of a selective, personality-targeted prevention program in reducing alcohol use and related harms: a cluster randomized controlled trial. J Child Psychol Psychiatry 2016;57(9):1056-1065. [doi: 10.1111/jcpp.12558] [Medline: 27090500]
25. Gan DZQ, McGillivray L, Larsen ME, Bloomfield T, Torok M. Promoting engagement with smartphone apps for suicidal ideation in young people: development of an adjunctive strategy using a lived experience participatory design approach. JMIR Form Res 2023;7:e45234 [FREE Full text] [doi: 10.2196/45234] [Medline: 37279058]
26. Bevan Jones R, Stallard P, Agha SS, Rice S, Werner-Seidler A, Stasiak K. Practitioner review: co-design of digital mental health technologies with children and young people. J Child Psychol Psychiatry 2020;61(8):928-940 [FREE Full text] [doi: 10.1111/jcpp.13258] [Medline: 32572961]
27. Thabrew H, Stasiak K, Kumar H, Naseem T, Frampton C, Merry S. A cognitive behavioral therapy-, biofeedback-, and game-based ehealth intervention to treat anxiety in children and young people with long-term physical conditions (starship rescue): co-design and open trial. JMIR Serious Games 2021;9(3):e26084 [FREE Full text] [doi: 10.2196/26084] [Medline: 34559053]
28. Firth J, Torous J, Nicholas J, Carney R, Rosenbaum S, Sarris J. Can smartphone mental health interventions reduce symptoms of anxiety? A meta-analysis of randomized controlled trials. J Affect Disord 2017;218:15-22 [FREE Full text] [doi: 10.1016/j.jad.2017.04.046] [Medline: 28456072]
29. Leech T, Dorstyn D, Taylor A, Li W. Mental health apps for adolescents and young adults: a systematic review of randomised controlled trials. Child Youth Serv Rev 2021;127:106073. [doi: 10.1016/j.childyouth.2021.106073]
30. West R, Michie S. A brief introduction to the COM-B model of behaviour and the PRIME theory of motivation. Qeios 2020:1-6. [doi: 10.32388/ww04e6.2]
31. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
32. Gargaritano KL, Murphy C, Auyeung AB, Doyle F. Systematic review of clinician-reported barriers to provision of brief advice for alcohol intake in hospital inpatient and emergency settings. Alcohol Clin Exp Res 2020;44(12):2386-2400. [doi: 10.1111/acer.14491] [Medline: 33119905]

XSL•FO
RenderX

33. Rosário F, Santos MI, Angus K, Pas L, Ribeiro C, Fitzgerald N. Factors influencing the implementation of screening and brief interventions for alcohol use in primary care practices: a systematic review using the COM-B system and theoretical domains framework. Implement Sci 2021;16(1):6 [FREE Full text] [doi: 10.1186/s13012-020-01073-0] [Medline: 33413487]
34. Stevely AK, Buykx P, Brown J, Beard E, Michie S, Meier PS, et al. Exposure to revised drinking guidelines and 'COM-B' determinants of behaviour change: descriptive analysis of a monthly cross-sectional survey in England. BMC Public Health 2018;18(1):251 [FREE Full text] [doi: 10.1186/s12889-018-5129-y] [Medline: 29444647]
35. Palmer B, Irwin C, Desbrow B. A review of factors influencing drinking behaviours in young Australian adults using a behavioural framework approach. Drug Alcohol Rev 2024;43(1):170-187. [doi: 10.1111/dar.13760] [Medline: 37830638]
36. Prior K, Ross K, Conroy C, Barrett E, Bock SG, Boyle J. Youth participation in mental health and substance use research: implementation, perspectives, and learnings of the Matilda Centre Youth Advisory Board. Ment Health Prev 2022;28:200251. [doi: 10.1016/j.mhp.2022.200251]
37. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2008;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
38. Thompson J. A guide to abductive thematic analysis. Qual Rep 2022;27(5):1410-1421. [doi: 10.46743/2160-3715/2022.5340]
39. Syan A, Lam JYY, Huang CGL, Smith MSM, Darnay K, Hawke LD, et al. The wellness quest: a health literacy and self-advocacy tool developed by youth for youth mental health. Health Expect 2021;24(2):659-669 [FREE Full text] [doi: 10.1111/hex.13214] [Medline: 33636052]
40. Poulton A, Eastwood O, Bruns LR, Sinnott RO, Hester R. Addressing methodological issues in a study of impulsivity and vulnerability for transition to alcohol use disorder. Alcohol Clin Exp Res 2022;46(2):262-276. [doi: 10.1111/acer.14755] [Medline: 34859438]
41. Babor TF, Higgins-Biddle JC, Saunders JB, Monteiro MG. The Alcohol Use Disorders Identification test: guidelines for use in primary care. World Health Organization. URL: https://www.who.int/publications/i/item/WHO-MSD-MSB-01.6a [accessed 2026-05-23]
42. Saunders JB, Aasland OG, Babor TF, de la Fuente JR, Grant M. Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption--II. Addiction 1993;88(6):791-804. [doi: 10.1111/j.1360-0443.1993.tb02093.x] [Medline: 8329970]
43. Kirby KN, Petry NM, Bickel WK. Heroin addicts have higher discount rates for delayed rewards than non-drug-using controls. J Exp Psychol Gen 1999;128(1):78-87. [doi: 10.1037//0096-3445.128.1.78] [Medline: 10100392]
44. Kaplan BA, Amlung M, Reed DD, Jarmolowicz DP, McKerchar TL, Lemley SM. Automating scoring of delay discounting for the 21- and 27-item monetary choice questionnaires. Behav Anal 2016;39(2):293-304 [FREE Full text] [doi: 10.1007/s40614-016-0070-9] [Medline: 31976983]
45. Poulton A, Hester R. Transition to substance use disorders: impulsivity for reward and learning from reward. Soc Cogn Affect Neurosci 2020;15(10):1182-1191 [FREE Full text] [doi: 10.1093/scan/nsz077] [Medline: 31848627]
46. Poulton A, Chen LPE, Dali G, Fox M, Hester R. Web-based independent versus laboratory-based stop-signal task performance: within-subjects counterbalanced comparison study. J Med Internet Res 2022;24(5):e32922 [FREE Full text] [doi: 10.2196/32922] [Medline: 35635745]
47. Congdon E, Mumford JA, Cohen JR, Galvan A, Canli T, Poldrack RA. Measurement and reliability of response inhibition. Front Psychol 2012;3:37 [FREE Full text] [doi: 10.3389/fpsyg.2012.00037] [Medline: 22363308]
48. Verbruggen F, Aron A, Band G, Beste C, Bissett P, Brockett A. A consensus guide to capturing the ability to inhibit actions and impulsive behaviours in the stop-signal task. Elife 2019;8:e46323. [doi: 10.7554/elife.46323]
49. Sobell LC, Sobell MB. Timeline follow-back: a technique for assessing self-reported alcohol consumption. In: Measuring Alcohol Consumption: Psychosocial and Biochemical Methods. Totowa, NJ: Humana Press; 1992:41-72.
50. Rueger SY, Trela CJ, Palmeri M, King AC. Self-administered web-based timeline followback procedure for drinking and smoking behaviors in young adults. J Stud Alcohol Drugs 2012;73(5):829-833 [FREE Full text] [doi: 10.15288/jsad.2012.73.829] [Medline: 22846247]
51. Löwe B, Decker O, Müller S, Brähler E, Schellberg D, Herzog W, et al. Validation and standardization of the Generalized Anxiety Disorder screener (GAD-7) in the general population. Med Care 2008;46(3):266-274. [doi: 10.1097/MLR.0b013e318160d093] [Medline: 18388841]
52. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
53. Kocalevent RD, Hinz A, Brähler E. Standardization of the depression screener Patient Health Questionnaire (PHQ-9) in the general population. Gen Hosp Psychiatry 2013;35(5):551-555. [doi: 10.1016/j.genhosppsych.2013.04.006] [Medline: 23664569]
54. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
55. Poulton A, Rutherford K, Boothe S, Brygel M, Crole A, Dali G, et al. Evaluating untimed and timed abridged versions of Raven's advanced progressive matrices. J Clin Exp Neuropsychol 2022;44(1):73-84. [doi: 10.1080/13803395.2022.2080185] [Medline: 35658791]

XSL•FO
RenderX

56. Raven J, Raven JC, Court JH. Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 4: Advanced Progressive Matrices, Sets I and II. San Antonio: Pearson; 1998.
57. Kirk RE. Practical significance: a concept whose time has come. Educ Psychol Meas 1996;56(5):746-759 [FREE Full text] [doi: 10.1177/0013164496056005002]
58. Cohen J. Statistical power analysis for the behavioral sciences. 2nd Edition. Hillsdale, NY: Lawrence Erlbaum Associates; 1988.
59. Poulton A, Bowen J, Dali G, Chew C, Prior K, Devine E. Reducing alcohol intake and harm through individualized feedback. OSF. 2023. URL: https://doi.org/10.17605/OSF.IO/FSXKA [accessed 2026-05-07]
60. Australian guidelines to reduce health risks from drinking alcohol. National Health and Medical Research Council. 2020. URL: http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/ds10-alcohol.pdf [accessed 2026-05-23]
61. Miller DT, Prentice DA. Changing norms to change behavior. Annu Rev Psychol 2016;67:339-361. [doi: 10.1146/annurev-psych-010814-015013] [Medline: 26253542]
62. Islam MM, Hoffmann MW, Rahman MB. Knowledge of low-risk drinking and its relationship with a reduction in alcohol consumption: results from six waves of an Australian national survey. Addict Behav 2019;95:172-177. [doi: 10.1016/j.addbeh.2019.03.016] [Medline: 30928662]
63. Livingston M. Perceptions of low-risk drinking levels among Australians during a period of change in the official drinking guidelines. Drug Alcohol Rev 2012;31(2):224-230. [doi: 10.1111/j.1465-3362.2011.00414.x] [Medline: 22220817]
64. Carrotte E, Groot C, Blanchard M, Hopgood F, Phillips L. The wavelengths study: A randomized controlled trial assessing the effectiveness of a codesigned podcast on mental illness stigma. Stigma and Health 2025;10(4):599-609. [doi: 10.1037/sah0000537]
65. National drug strategy household survey 2022-2023. Australian Institute of Health and Welfare. 2024. URL: https://www.aihw.gov.au/reports/illicit-use-of-drugs/national-drug-strategy-household-survey/ [accessed 2025-11-07]
66. Poulton A. Evaluating co-designed versus researcher-driven personalised feedback. OSF. URL: https://osf.io/rnbv8/ [accessed 2025-09-03]

Abbreviations

APM: Advanced Progressive Matrices
AUDIT: Alcohol Use Disorders Identification Test
COM-B: Capacity, Opportunity, Motivation–Behavior
COM-B: AIQ: COM-B: Alcohol Intake Questionnaire
CONSORT: Consolidated Standards of Reporting Trials
COREQ: Consolidated Criteria for Reporting Qualitative Research
GAD-7: Generalized Anxiety Disorder-7
ICJME: International Committee of Medical Journal Editors
MCQ: monetary choice questionnaire
PHQ-9: Patient Health Questionnaire-9
RT: reaction time
SSRT: stop signal reaction time
SST: stop signal task
WHO: World Health Organization

Edited by S Law; submitted 08.Nov.2025; peer-reviewed by D Guertler, LM Rodriguez; comments to author 27.Mar.2026; revised version received 18.May.2026; accepted 19.May.2026; published 12.Jun.2026.

Please cite as:
Poulton A, Bowen J, Dali G, Chew C, Prior K, Devine EK, Birrell L, Stapinski LA, Hester R
Evaluating Co-Designed vs Researcher-Driven Personalized Feedback Formats in a Brief Digital Alcohol Use Intervention: Mixed Methods Study
J Med Internet Res 2026;28:e87393
URL: https://www.jmir.org/2026/1/e87393
doi:10.2196/87393
PMID:

©Antoinette Poulton, Joanna Bowen, Gezelle Dali, Cindy Chew, Katrina Prior, Emma K Devine, Louise Birrell, Lexine A Stapinski, Robert Hester. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License

XSL•FO
RenderX

(https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Web-Based Virtual Environment Behavioral Intervention as Cardiovascular Disease and Metabolic Disease Prevention Education in Persons With HIV: Evaluation of the LEARN Randomized Controlled Trial

S Raquel Ramos[1,2,3*], MBA, MSN, PhD; Lindsie Boerger[4*], BA; Arjee Javellana Restar[2,3*], MPH, PhD; Manali Phadke[5*], MA; Sangchoon Jeon[1*], PhD; Constance Johnson[6,7*], MS, PhD; Gail Melkus[8*], ANP, EdD; Trace Kershaw[2,3*], PhD; Harmony Reynolds[9,10*], MD; Allison Vorderstrasse[11*], RN, DNSc

[1]School of Nursing, Yale University, Orange, CT, United States
[2]Department of Social and Behavioral Sciences, School of Public Health, Yale University, New Haven, CT, United States
[3]Center for Interdisciplinary Research of AIDS, New Haven, CT, United States
[4]Yale Center for Clinical Investigation, New Haven, CT, United States
[5]Yale Center for Analytical Sciences, Yale School of Public Health, New Haven, CT, United States
[6]Cizik School of Nursing, UT Health Houston, Houston, TX, United States
[7]McWilliams School of Biomedical Informatics, UT Health Houston, Houston, TX, United States
[8]Rory Myers College of Nursing, New York University, New York, NY, United States
[9]Department of Medicine, Grossman School of Medicine, New York University, New York, NY, United States
[10]Cardiovascular Clinical Research Center, NYU Langone Health, New York, NY, United States
[11]Elaine Marieb College of Nursing, University of Massachusetts Amherst, Amherst, MA, United States
[*]all authors contributed equally

Corresponding Author:
S Raquel Ramos, MBA, MSN, PhD
School of Nursing
Yale University
400 West Campus Drive
Orange, CT, 06477
United States
Phone: 1 2037372339
Email: raquel.ramos@yale.edu

Abstract

Background: Persons living with HIV have an increased risk of cardiovascular and metabolic diseases compared with HIV-seronegative individuals.

Objective: This study aims to assess the feasibility and acceptability of a virtual environment (VE) for cardiovascular disease and metabolic disease prevention education among men living with HIV through a pilot waitlist-control randomized trial.

Methods: In phase 1, we conducted interviews and VE beta testing with 25 individuals. In phase 2, reported here, we conducted a waitlist-controlled trial using permuted block randomization. Participants were allocated to either the VE intervention (n=40) or the waitlist control arm (n=38). The primary outcomes were feasibility (defined by recruitment and retention metrics) and acceptability (defined by levels of engagement) of the VE. We also examined preliminary effect sizes across several cardiovascular health–related indicators.

Results: Participants had a mean age of 41.8 (SD 10.4) years and had been living with HIV for a mean of 15.9 (SD 9.7) years. Of the 56 total participants, 35 (63%) were Black and 15 (27%) were Hispanic. A total of 14 out of 31 (45%) of participants with baseline data in the intervention arm engaged with the VE, with an average session duration of 110 minutes. The most highly engaged content modules were nutrition (146 engagements), oral health (138 engagements), fitness tips and videos (82 engagements), and relaxation techniques (75 engagements). Retention was 48 out of 78 (61%) at the final assessment, which, to our knowledge, is higher than the typical retention rates reported in remote digital health and behavioral intervention research. Preliminary efficacy

XSL•FO
RenderX

analyses of changes at 3 and 6 months revealed medium effect sizes favoring the intervention for key nutritional outcomes (vegetable consumption: d=0.59-0.66; whole grain intake: d=0.41-0.46) and vigorous physical activity (d=0.38-0.57). Small-to-medium effect sizes were observed for fast food reduction (d=0.37 at 6 months), total physical activity (d=0.37 at 6 months), depressive symptoms (d=0.22 at 3 months), and HIV illness perceptions, including Timeline Acute/Chronic subscale score (d=0.44 at 3 months), Illness Coherence (d=0.33 at 3 months), and Emotional Representations (d=0.23-0.33).

Conclusions: The LEARN (LEveraging A viRtual eNvironment to enhance prevention of HIV-related comorbidities) study demonstrated the feasibility and acceptability of a VE designed for cardiovascular and metabolic disease prevention education among persons living with HIV. High engagement with specific educational content and favorable retention rates highlight the promise of this innovative approach. Preliminary effect sizes suggested positive trends in cardiovascular health indicators and mental well-being, indicating the potential for additional benefits associated with the intervention. These findings support progression to a fully powered randomized controlled trial. Subsequent funding has been awarded based on this work to expand the research in LEARN2. We will integrate lessons learned from this study and focus on HIV-related conditions with shared risk factors to enhance the intervention's impact. By addressing HIV-related conditions with interrelated risk factors, we aim to provide a more comprehensive intervention to improve the overall health and well-being of individuals living with HIV.

Trial Registration: ClinicalTrials.gov NCT05242952; https://clinicaltrials.gov/ct2/show/NCT05242952

International Registered Report Identifier (IRRID): RR2-10.2196/38348

(J Med Internet Res 2026;28:e91145) doi:10.2196/91145

KEYWORDS

virtual environment; gamification; preventative cardiology; chronic illness; HIV; CVD; metabolic disease; AR; VR; health education; lifestyle; RCT; minority groups

Introduction

Cardiovascular disease (CVD) and metabolic disorders pose significant health challenges, particularly among individuals living with HIV. These conditions not only contribute to morbidity and mortality but also complicate HIV management. The incidence of HIV remains highest among Black (26%) and Hispanic (24%) men [1] with incomes at or below the poverty threshold [2]. CVD has emerged as the leading cause of mortality among people with HIV [3-5], primarily because of chronic inflammation associated with the virus, which elevates CVD risk even when viral loads are well controlled [6-9]. Additionally, allostatic load adversely affects the immune, cardiovascular, and metabolic systems over time [10], leading to increased stress responses. These responses elevate cortisol and other inflammatory biomarkers, resulting in physiological disruptions [11] that contribute to cardiometabolic morbidities [12]. Although many CVD risk factors are modifiable, people living with HIV remain at elevated risk because of both shared risk factors with CVD and HIV-specific risk factors.

CVD risk among Black and Hispanic men living with HIV arises from the combined effects of systemic mistreatment, physiological stress responses, and aging with chronic illness. Conceptually referred to as a "cardiovascular conundrum," Thayer and colleagues [13,14] elucidate how stress responses to systemic mistreatment experienced by Black individuals and other underresourced groups [15] lead to elevated systemic vascular resistance, thereby increasing morbidity and mortality risks associated with CVD. Notably, higher vagally mediated heart rate variability, which is typically associated with lower CVD risk, can be negatively affected by consistent exposure to interpersonal conflict, leading to reduced heart rate variability over time [14,16,17]. Compared with the general US population, individuals living with HIV have a CVD risk that is 1.5-2 times higher, with this risk increasing with age [5,15,18]. By 2030, nearly 80% of individuals living with HIV will be 50 years or older and have 1 or more chronic conditions [19], highlighting the urgent need to address cardiovascular risk in this population.

Innovative approaches to lowering CVD risk among men living with HIV, including behavioral prevention, education, and technological strategies, are essential to address pressing health issues in this population. In 2024, a study by the Pew Research Center [20] indicated that the digital divide has significantly narrowed, with up to 90% of adults aged 65 years and older and 91% of individuals living in poverty reporting internet use. A 2022 nationally representative survey conducted by AARP found that 45% of individuals aged 50 years and older engage in video gaming to maintain mental sharpness, reduce stress, entertain themselves, and solve problems [21]. Research using serious games has been applied to social connection and education after stroke [22], diabetes self-management [23], virtual environment (VE)–based exercise among persons with diabetes [24], self-management and lifestyle changes among individuals with bipolar disorder [25], and cardiac rehabilitation [26]. Although core health promotion principles apply across populations, men living with HIV face unique challenges that may benefit from VE-based interventions, particularly because they experience intersecting stigmas related to HIV status, sexual minority identity, and race/ethnicity that may reduce engagement with traditional health care settings [13,14]. VEs offer distinct advantages for CVD prevention education in this population, including anonymity that may reduce stigma-related barriers to engagement, flexibility for self-paced learning that accommodates varying health literacy levels, and immersive gamification elements that can enhance motivation and sustained engagement with health content [22-26]. Taken together, these findings highlight the potential for technology-driven VE-based behavioral interventions to serve as effective tools for health education and engagement, particularly among populations at elevated risk, such as men living with HIV.

XSL•FO
RenderX

The Second Life Impacts Diabetes Education & Self-Management (SLIDES) study [27] and the Learning in a Virtual Environment (LIVE) platform [28] were disease-agnostic VEs previously tested for diabetes self-management. Our team modified the LIVE platform by shifting its focus and content toward CVD and metabolic disease prevention education for sexual minority men living with HIV, entitled "The LEARN Study" ("LEveraging A viRtual eNvironment to enhance prevention of HIV-related comorbidities"). Briefly, the LEARN study incorporated gamification and educational quests to enhance participant engagement while increasing awareness of cardiovascular health, risk factors, and effective prevention strategies. The platform was personalized for cultural salience and content through formative work described elsewhere [28]. The intervention was informed by Social Cognitive Theory, emphasizing self-efficacy and observational learning [29], and Self-Determination Theory, emphasizing autonomous motivation [30]. VE components, including gamification, self-directed exploration, and personalized avatars, were designed to enhance self-efficacy, autonomous motivation, and knowledge acquisition, which were hypothesized to influence behavioral and health outcomes. The aims of the LEARN study were to (1) evaluate the feasibility and acceptability of this pilot intervention among persons living with HIV and (2) examine the preliminary efficacy of the intervention across health indicators.

Methods

Design

Our study evaluation is reported in accordance with the CONSORT (Consolidated Standards of Reporting Trials)-EHEALTH guidelines [31].

To the team's knowledge, the LEARN Study was the first waitlist control randomized trial to test the feasibility and acceptability of a VE for CVD and metabolic disease prevention education among men living with HIV [32,33]. The waitlist design was advantageous in mitigating ethical concerns for the control group because all participants ultimately received the intervention, particularly in pilot and efficacy trials for which no standard-of-care treatment was available [34]. Before the pilot intervention, our team conducted online-only qualitative interviews with 15 individuals and beta testing of the VE with 10 individuals. Findings related to VE design and personalization that informed the LEARN clinical trial are described elsewhere [32,33,35,36]. Data suggested that participants were enthusiastic about the potential of a VE to address health risks associated with HIV and desired greater representation in avatar customization options. In this study, enrolled participants completed the pilot trial over a 6-month period based on their random assignment, with outcomes measured at 3 time points: baseline, 3 months, and 6 months. The 6-month study duration was selected based on evidence that behavioral lifestyle changes typically require 3-6 months to become established and that longer follow-up periods allow assessment of behavior maintenance beyond initial adoption. This time frame aligns with prior VE intervention studies and permits detection of both short-term (3-month) and sustained (6-month) effects [32,33,35,36].

Ethical Considerations

All study procedures were approved by the Yale Human Research Protection Program (institutional review board approval number 2000031403). This approval covered all study phases. The protocol was registered on ClinicalTrials.gov (NCT05242952). Informed consent was obtained verbally by the study coordinator during the screening phone call, and participants provided consent before enrollment and data collection. To protect privacy and confidentiality, participants created pseudonyms for their VE avatars, and no personally identifiable information was displayed within the VE. All data were stored in REDCap (Vanderbilt University), a Health Insurance Portability and Accountability Act (HIPAA)-compliant electronic data capture system with role-based access controls. Participants received compensation of US $30 for each completed assessment (baseline, 3 months, and 6 months), for a maximum total of US $90.

Participants

Eligibility criteria were (1) age 30 years or older, (2) male sex, (3) HIV-seropositive status, (4) ability to participate, and (5) no medical history of serious complications such as heart attack, stroke, or cognitive impairment. Ability to participate was operationalized as (1) having reliable internet access, (2) access to a computer capable of running the VE platform, (3) self-reported ability to read and understand English, and (4) no self-reported visual or motor impairments that would prevent VE navigation. These criteria were assessed during the screening phone call. Additionally, eligibility was not restricted by race/ethnicity; however, recruitment strategies prioritized community partnerships and outreach in settings serving populations with the highest HIV burden.

The rationale for inclusion was grounded in research suggesting that chronic illnesses are increasingly being diagnosed earlier in life, particularly among individuals aged 18-30 years, with the highest prevalence observed among those aged 50 years and older. Data from the Million Hearts study [37] indicate that more than 35% of life-changing cardiac events occur in adults as young as 35 years. Accordingly, our sample age range, beginning at 30 years, was appropriate given the evolving age demographics associated with CVD risk. The focus on men aligns with epidemiologic patterns demonstrating higher HIV incidence among Black and Hispanic men and elevated cardiovascular risk among men aging with HIV [1,2]. The ability to participate was necessary for sustained engagement with a VE that incorporated quests, navigation, and health information. Excluding individuals with prior serious cardiac events or cognitive impairment protected participant safety and avoided potential confounding related to advanced disease or impaired capacity to complete VE activities.

LEARN Intervention

We conducted the LEARN study by customizing the LIVE disease-agnostic platform [28]. The development process for LIVE is described elsewhere [28] and was previously tested for diabetes self-management through the SLIDES study [27].

XSL•FO
RenderX

The LEARN intervention included an orientation packet detailing step-by-step login instructions for the VE (Figure 1), describing the VE districts (eg, lobby, food court, plaza, and reflection garden), and outlining the activities available in each area. Its primary purpose was to familiarize participants with the VE's features and facilitate basic troubleshooting. We also provided a suite of digital support materials. A student research assistant produced a series of brief tutorial videos using Microsoft PowerPoint (each under 6 minutes) covering essential troubleshooting, navigation strategies, and "tips and tricks" for effective engagement within the VE (Figure 2). These orientation videos were distributed via email and posted on the VE's internal message boards.

Informed by LEARN qualitative findings [32,33,35,36], the platform was tailored to participants' preferences and personalized to include a pseudonym for use with their avatar when logged into the VE. Pseudonyms were also used to protect participants' personally identifiable information. Once onboarded, participants engaged in interactive, self-directed exploration within the VE. They customized their avatars (eg, hair color, skin color, clothing, and footwear) to reflect personal preferences and identity considerations, navigated across multiple VE districts, and encountered targeted health content designed to promote behavioral modification. Within each district, participants were presented with information and images relevant to cardiovascular and metabolic health, such as making dietary choices in the food court or exploring physical activity resources in the plaza, to reinforce key prevention messaging in an immersive context. In parallel, a nurse health educator provided 2 live sessions within the VE and created several videos on topics aligned with the American Heart Association Cardiovascular Health framework (Figure 3) [38]. Recorded versions of these sessions were also posted and emailed to participants as asynchronous reinforcement. Although these resources had limited live integration, they served a critical role in participant onboarding, engagement, and knowledge acquisition.

Engagement with the VE was primarily self-directed; however, participants received (1) monthly email newsletters (not containing CVD-related content) to maintain study contact, (2) email notifications when new VE content was posted, and (3) invitations to 2 live health educator sessions within the VE. These communication strategies supported participant retention (discussed further below) while preserving the self-directed nature of the intervention.

Figure 1. LEARN (LEveraging A viRtual eNvironment to enhance prevention of HIV-related comorbidities) orientation packet.

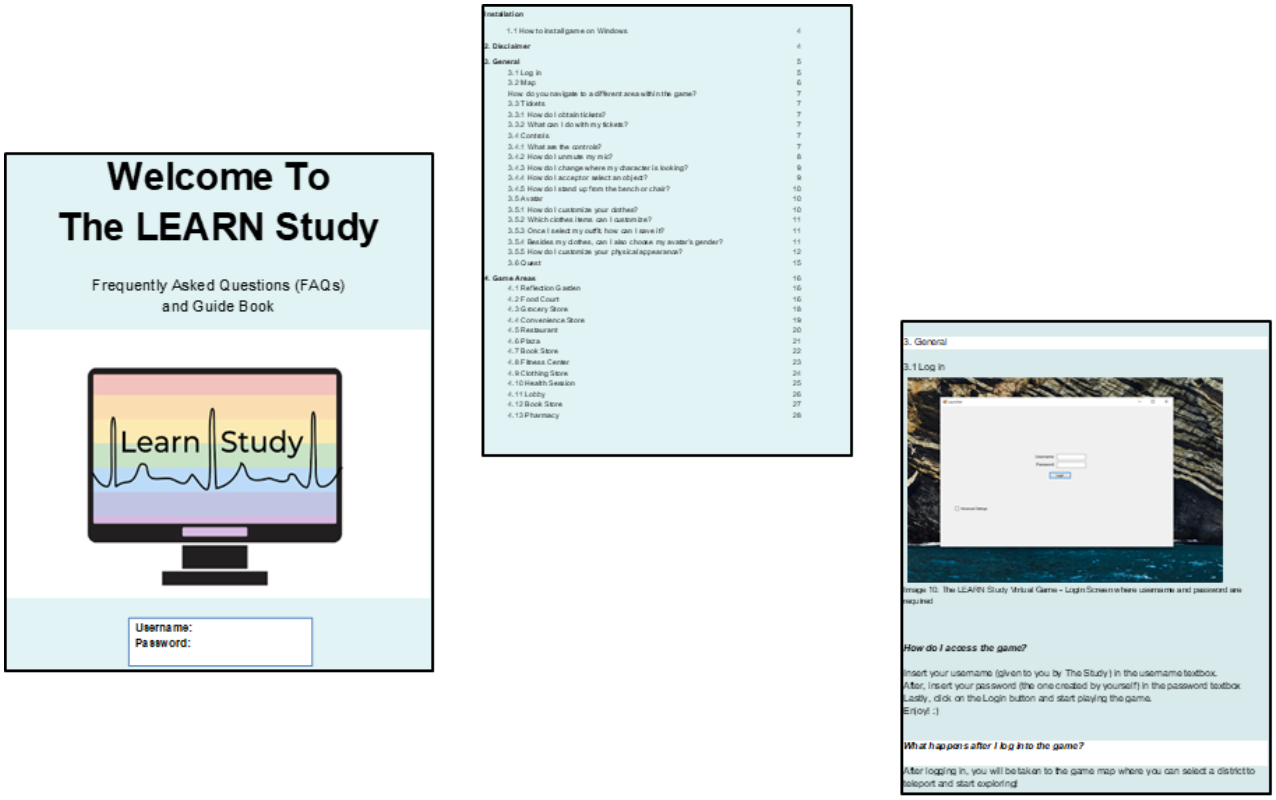

XSL•FO
RenderX

Figure 2. LEARN (LEveraging A viRtual eNvironment to enhance prevention of HIV-related comorbidities) troubleshooting content.

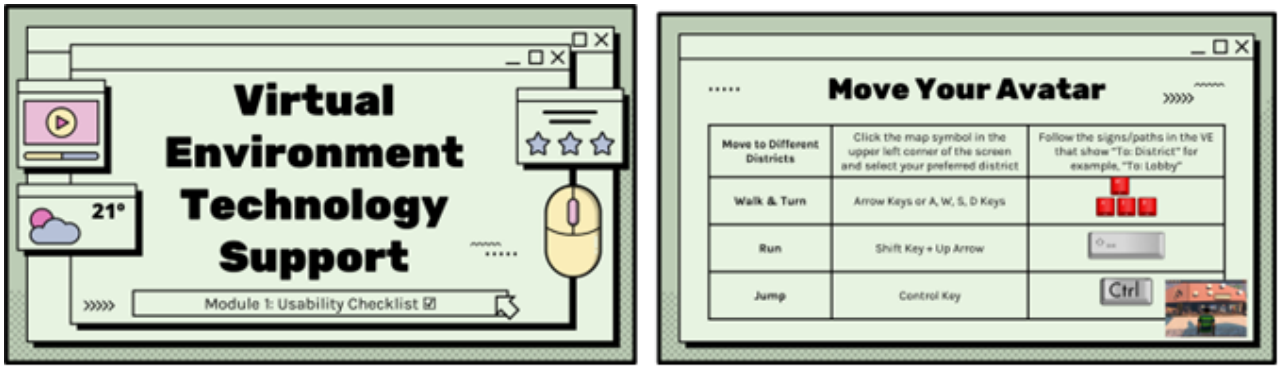

Figure 3. LEARN health video content. LEARN: LEveraging A viRtual eNvironment to enhance prevention of HIV-related comorbidities.

Recruitment, Screening, and Retention Approaches

We collaborated with the Yale Center for Clinical Investigation to support our clinical trial. A research coordinator with expertise in clinical trial screening, enrollment and consent, data collection, tracking, documentation, and regulatory compliance supported the team throughout the study. We used several recruitment strategies. First, we contacted partnering community-based organizations and distributed recruitment flyers, while program managers informed clients about the study through word of mouth. Second, we implemented a MyChart messaging campaign targeting patients who might be eligible for the research and had agreed to be contacted. Third, the research coordinator contacted patients who had previously expressed interest in clinical trials using a recruitment center call list. Fourth, we partnered with an external investigator and used a listserv of 1800 individuals who had consented to be contacted regarding HIV health research studies. Additionally, the protocol was publicly accessible through registration on ClinicalTrials.gov (NCT05242952).

Between July 2023 and December 2024, individuals were screened for participation. Those who met the eligibility criteria provided informed consent to the research coordinator by phone. Once consented, participants were allocated either to immediate access to the VE or to a waitlist control arm using permuted-block randomization with varying block sizes to ensure allocation concealment and balanced assignment between the intervention and control groups throughout the study period [39]. Additionally, the coordinator verified each participant's identity through confirmatory phone calls and email confirmations to prevent individuals from enrolling multiple times.

To enhance retention, the study coordinator maintained regular contact with participants by sending voicemail and email reminders regarding upcoming data collection time points. Monthly newsletters were also distributed via email to support ongoing study communication (Figure 4). These newsletters intentionally excluded content related to CVD or metabolic disease prevention to ensure that retention communications remained distinct from the active intervention and did not influence study outcomes. The primary purpose of the

XSL•FO
RenderX

newsletters was to engage participants indirectly by maintaining an open communication channel and acknowledging seasonal occasions and relevant events meaningful to participants.

Figure 4. LEARN (LEveraging A viRtual eNvironment to enhance prevention of HIV-related comorbidities) Newsletter.

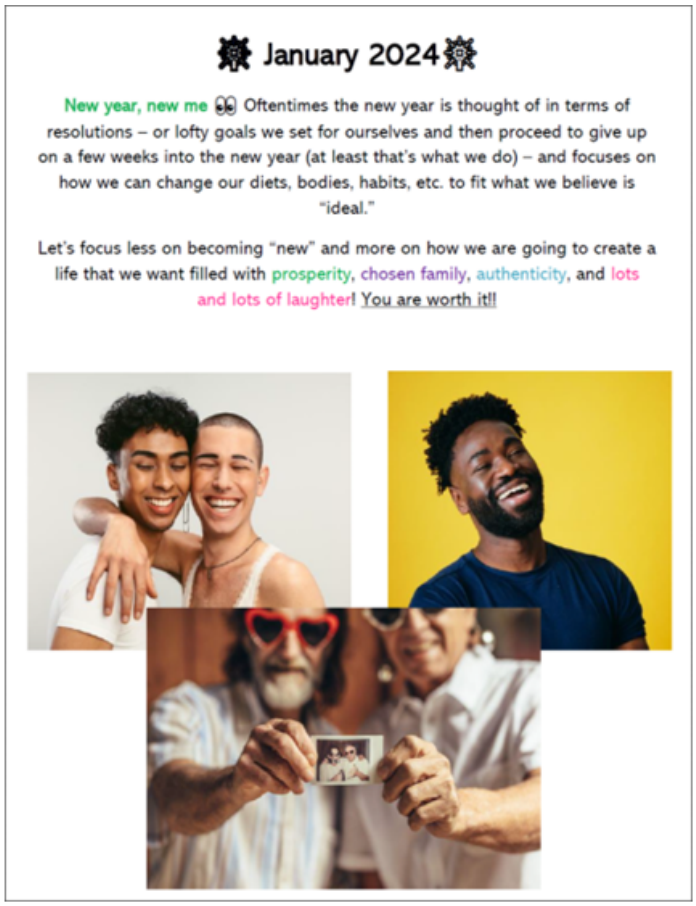

Sample Size Calculation

We evaluated the acceptability and feasibility of the VE while obtaining preliminary effect size estimates of its potential impact on several CVD- and HIV-related health indicators. It is important to clarify that this study was not intended to serve as a definitive test of intervention efficacy. Previous research indicated that feasibility studies with a sample size of at least 10 participants were adequate to achieve these objectives [40-42]. Following recommendations from these studies, sample size was determined based on the precision of effect size estimation rather than statistical power for hypothesis testing, as the study was not designed as a definitive efficacy trial. With 40 participants per group, we anticipated 95% CIs for standardized mean differences with a half-width of approximately 0.45 SD, providing sufficient precision to inform decisions regarding progression to a fully powered trial.

Measures

Self-reported survey data were collected through scheduled phone calls with the study coordinator at baseline, 3-month, and 6-month time points. The coordinator entered participant responses into REDCap, which included demographic, behavioral, and health indicators. Survey data collection required approximately 60-90 minutes. Acceptability data were assessed using engagement metrics derived from the VE platform's metadata. These metadata included participant login records and log file analyses, which are digital records capturing activities within the VE. This information provided insights into login frequency and duration of platform use during the study period and was logged under each participant's avatar pseudonym.

Primary Outcomes

The primary outcomes of this study were the feasibility and acceptability of the VE. Feasibility was operationalized through recruitment and retention metrics. Recruitment metrics were assessed by tracking the proportion of individuals who enrolled among those initially contacted. Retention metrics were measured as the percentage of participants who completed the 3- and 6-month follow-up assessments. Acceptability was evaluated through engagement metrics within the VE, quantified by calculating the total time participants engaged with the platform (in minutes). This included the duration spent visiting various districts (eg, grocery, pharmacy, bookstore, and outlet districts) and interacting with health content modules (eg, nutrition, oral health, fitness tips and videos, and relaxation techniques).

Demographics

Participant demographic characteristics were collected at baseline through self-report questionnaires. Variables assessed included age (continuous, in years), years living with HIV (continuous, in years), race/ethnicity (categorized as Black,

XSL•FO
RenderX

Hispanic/Latine, Asian, Middle Eastern, or other), relationship status (categorized as divorced, married, partnered, or single), education level (categorized as did not complete high school, high school graduate or General Educational Development, some college or associate degree, bachelor's degree, or master's degree), and employment status (categorized as employed, on disability, self-employed, student, unemployed, or seasonal).

Health Indicators

Multiple cardiovascular health indicators were assessed at baseline, 3-month, and 6-month follow-ups using self-report measures aligned with the American Heart Association Life's Essential Eight metrics for cardiovascular health [38]. We assessed self-reported history of hypertension (yes/no) and diabetes (yes/no). Sleep was measured by hours of sleep per night, with adequacy evaluated against the recommended 7-9 hours [43]. BMI was calculated from self-reported height and weight (kg/m^2) and categorized as underweight (<18.5 kg/m^2), normal weight (18.5-24.9 kg/m^2), overweight (25.0-29.9 kg/m^2), or obese (≥30.0 kg/m^2) [44]. Nutritional intake was assessed through average daily servings of vegetables, fruit, and whole grains, along with weekly consumption of sweets, with adherence evaluated against recommendations of 3 servings of vegetables and 2.5 servings of fruit daily [45,46]. Physical activity was assessed using the International Physical Activity Questionnaire, which measures self-reported physical activity in minutes per week, including walking, moderate activity, and vigorous activity [47]. Mental health was evaluated using the 9-item Patient Health Questionnaire-9 to assess depressive symptoms, with participants categorized as having mild, moderate, or severe depression [48]. Tobacco and e-cigarette use were also assessed, including smoking status and quit attempts, using the Behavioral Risk Factor Surveillance System Questionnaire [49].

HIV illness perceptions were assessed using the Revised Illness Perception Questionnaire (IPQR), adapted for HIV, which measures cognitive and emotional representations of illness [50]. The IPQR includes an identity subscale assessing symptoms attributed to HIV and 7 belief subscales measured on a 5-point Likert scale: Timeline Acute/Chronic, Timeline Cyclical, Consequences, Personal Control, Treatment Control, Illness Coherence, and Emotional Representations. Higher scores indicate stronger beliefs within each domain. The IPQR has demonstrated good internal consistency (Cronbach α=.71-.87) across diverse patient populations, including persons living with HIV [50].

We also assessed additional health indicators, including the ability to receive medical care when needed (yes/no) and current health insurance coverage (yes/no). Health literacy was measured using the Standard Assessment of Health Literacy [51], which evaluates an individual's understanding of common medical terms through word recognition and multiple-choice comprehension items. Internalized homophobia was assessed using the Internalized Homophobia Scale [52,53], a validated instrument that measures the extent to which individuals have internalized negative societal attitudes related to identifying as a sexual minority. The scale includes items assessing personal discomfort with one's sexual orientation; desire to change sexual orientation; and negative feelings about being gay, lesbian, or bisexual. Higher scores indicate greater levels of internalized homophobia. Sexual orientation concealment was measured using the Sexual Orientation Concealment Scale [54], which evaluates concealment behaviors across various settings, including the workplace, family, health care settings, and social situations. Items assess both active concealment strategies (eg, avoiding disclosure and presenting as heterosexual) and vigilance related to identity management. Higher scores indicate greater concealment of sexual orientation.

Statistical Analysis

Outcome assessor blinding was not feasible because data collection was conducted by the study coordinator through phone-administered surveys, which required knowledge of participant allocation to coordinate VE access and provide troubleshooting support for intervention participants. To minimize bias, the principal investigator (PI) and study coordinator were excluded from data analysis, which was conducted independently by the biostatistician.

This analysis focused on the 59 participants who completed baseline assessments (31 intervention and 28 control). Means and standard deviations were used to describe continuous variables, whereas frequencies (n) and percentages (%) were used to describe categorical variables, as appropriate. Associations between outcome and demographic variables by group at baseline were assessed using Student *t* tests for continuous outcome variables and chi-square tests of independence for categorical variables. For preliminary efficacy testing, health indicators were collected and analyzed longitudinally across 3 time points (baseline, 3 months, and 6 months) and compared between the intervention and waitlist control groups. A strict intention-to-treat approach was not applied. Rather, participants with at least one data point were included in the longitudinal analyses using linear mixed-effects models, which accommodate missing data under the assumption that data are missing at random. Linear mixed-effects models with restricted maximum likelihood estimation were used to examine differences in each health indicator between the intervention and control groups over time. The model included fixed effects for randomization group, visit time, and the group × time interaction, as well as age, years living with HIV, income category, and education category as covariates. A random intercept was included to account for within-subject correlation across repeated measurements. Type III tests of fixed effects were used to evaluate the significance of each term in the model. Differences in least squares means between the intervention and control groups were estimated at each time point and are presented as effect size estimates with 95% CIs. Statistical significance was set at α=.05. All analyses were conducted using SAS version 9.4 (SAS Institute Inc) [55].

Analyses of engagement and acceptability metrics (VE login sessions, content module access, and session duration) applied only to the intervention arm because control participants did not have access to the VE during the waitlist period. Effect size estimates for health indicators represent between-group differences (intervention vs control) at each time point.

XSL•FO
RenderX

Results

Participant Demographics

Table 1 summarizes participant demographic characteristics. The mean age of participants was 41.8 (SD 10.4) years. The waitlist control group had a mean age of 42.6 (SD 11.2) years, whereas the intervention group had a mean age of 41.1 (SD 9.7) years. Participants had been living with HIV for a mean of 15.9 (SD 9.3) years. The racial/ethnic composition (N=56) was as follows: 35 (63%) participants were Black, 15 (27%) were Hispanic/Latin American, 1 (2%) was Asian, and 1 (2%) was Middle Eastern. Overall, 36 out of 58 (62%) participants were single and 13 out of 58 (22%) were married. Participants reported a range of educational backgrounds, with the largest proportion having completed some college (17/58, 29%), followed by those with bachelor's degrees (14/58, 24.1%) and master's degrees (14/58, 24%). Employment status varied, with 28 out of 58 (48%) employed and 21 out of 58 (36%) either unemployed or working seasonally (Figure 5).

Table 1. LEARN[a] sample demographic characteristics (N=59).[b]

Characteristics	Total, n (%)[c]	Waitlist control (n=28)[c]	Intervention (n=31)[c]	*P* value[d]
Age, mean (SD)	41.8 (10.4)	42.6 (11.2)	41.1 (9.7)	.58
Years living with HIV, mean (SD)	15.9 (9.3)	18.1 (11.1)	13.8 (6.7)	.08
Race/ethnicity, n (%)				.72
Asian	1 (2)	0 (0)	1 (4)	
Black or African American	35 (63)	18 (64)	17 (61)	
Hispanic or Latin American	15 (27)	7 (25)	8 (29)	
Middle Eastern	1 (2)	0 (0)	1 (4)	
Multiple races/mixed races	4 (7)	3 (11)	1 (4)	
White	0 (0)	0 (0)	0 (0)	
Relationship status, n (%)				.42
Divorced	1 (2)	1 (4)	0 (0)	
Married	13 (22)	4 (14)	9 (30)	
Partnered	8 (14)	4 (14)	4 (13)	
Single	36 (62)	19 (68)	17 (57)	
Education , n (%)				.32
Did not complete high school	2 (3)	2 (7)	0 (0)	
High school graduate or General Educational Development	11 (19)	5 (18)	6 (20)	
Some college or associate's degree	17 (29)	9 (32)	8 (27)	
Bachelor's degree	14 (24)	4 (14)	10 (33)	
Master's degree	14 (24)	8 (29)	6 (20)	
Employment status , n (%)				.49
Employed	28 (48)	12 (43)	16 (53)	
On disability	5 (9)	4 (14)	1 (3)	
Self-employed	1 (2)	1 (4)	0 (0)	
Student	3 (5)	1 (4)	2 (7)	
Unemployed or seasonal	21 (36)	10 (36)	11 (37)	

[a]LEARN: LEveraging A viRtual eNvironment to enhance prevention of HIV-related comorbidities.

[b]Summed counts may differ from the total sample because of missing data.

[c]Column percentages.

[d]Chi-square test if categorical or *t* test for continuous variables.

XSL•FO
RenderX

Figure 5. CONSORT (Consolidated Standards of Reporting Trials)-EHEALTH flowchart.

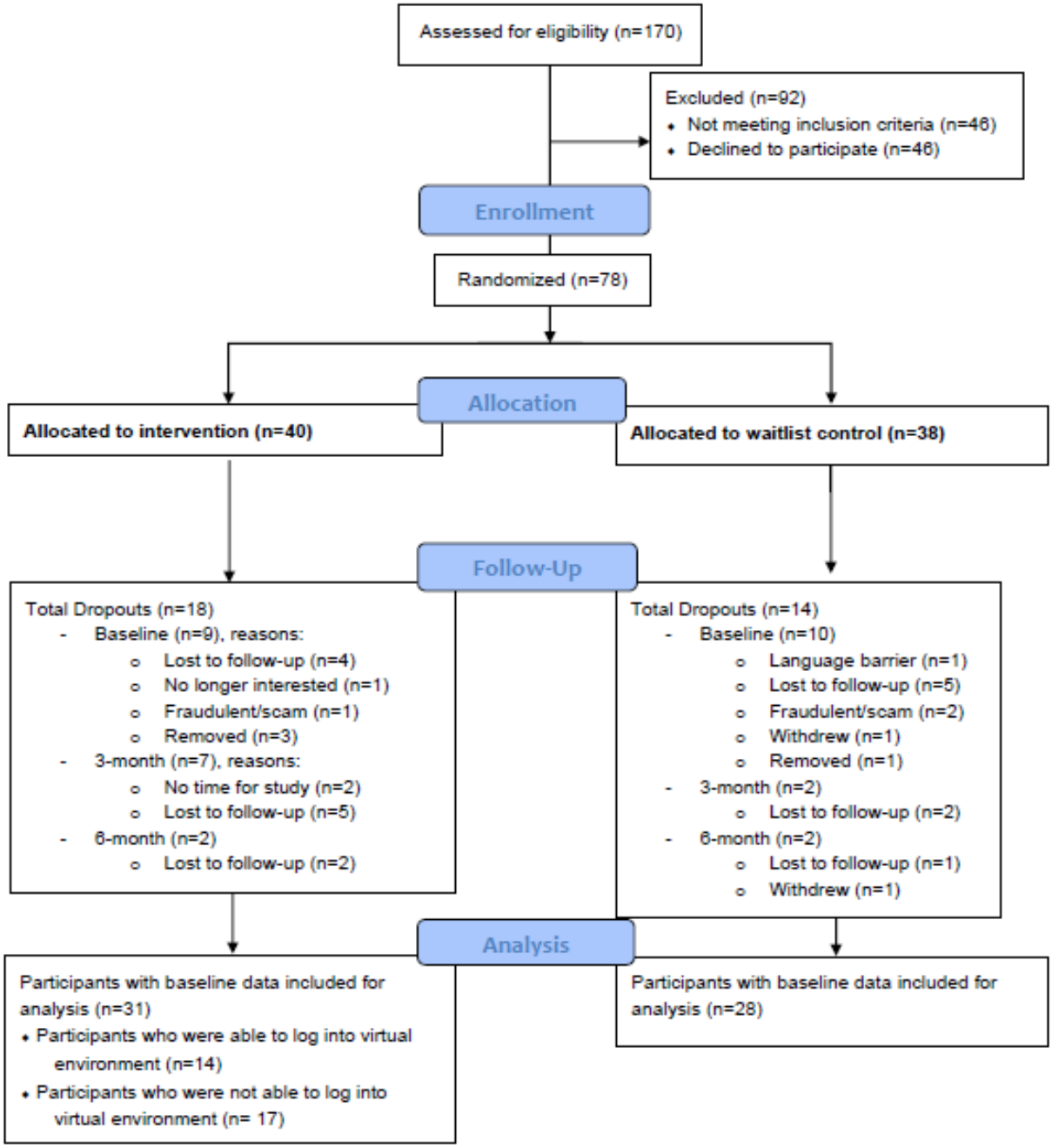

Primary Outcomes

Feasibility

CONSORT-EHEALTH (Multimedia Appendix 1) provides a summary of the feasibility results. A total of 170 individuals were screened and assessed for eligibility, of whom 92 were excluded (n=46 did not meet the inclusion criteria and n=46 declined to participate). In the randomized controlled trial (RCT), 78 of the targeted 80 (98%) participants were enrolled, demonstrating successful recruitment. Of these 78 participants, 40 were assigned to the intervention group and 38 to the waitlist control group.

The final analytic sample size accounted for participant attrition. A total of 59 participants (n=31 in the intervention group and n=28 in the waitlist control group) were included in the final analytic sample following attrition that occurred primarily before baseline data collection, during which 19 (24%) participants dropped out. An additional 8 (10%) participants withdrew before the second assessment (at 3 months), and 3 (4%) participants withdrew before the third assessment (at 6 months). At study completion, 26 participants remained in the waitlist control group and 22 remained in the intervention group. Reasons for dropout across all time points included loss to follow-up (n=19), language barriers (n=1), lack of continued interest in the study (n=1), insufficient time to complete the study (n=2), fraudulent or attempted multiple enrollments (n=3), participant withdrawal

XSL•FO
RenderX

(n=2), and removal from the study (n=4). Following allocation, approximately half of the participants in the immediate intervention arm (n=14, 45%) actively engaged with content in the VE. Within the intervention arm, 14 (45%) participants logged into the VE at least once.

Acceptability

Table 2 summarizes the acceptability results. Participants logged a total of 110 sessions, with an average session duration of 1 hour and 50 minutes. Quests were accessed 14 times, with the most frequently visited quests located in the fitness district (64.0%), followed by the grocery district (14.0%) and the pharmacy, bookstore, and outlet districts (7.0% each). The most frequently accessed content areas were nutrition (146 engagements), oral health (138 engagements), fitness tips and videos (82 engagements), and relaxation techniques (75 engagements) (Multimedia Appendix 2).

Table 2. LEARN[a] acceptability results.

Acceptability	Total
Total sessions logged, n	• 110
Average session duration (minutes)	• 110
Distinct educational quests accessed, n	• 14
Most visited districts	• Fitness: 64.0% • Grocery: 14.0% • Pharmacy: 7.0% • Bookstore: 7.0% • Outlet: 7.0%
Most engaged modules	• Nutrition: k=146 • Oral health: k=138 • Fitness tips and videos: k=82 • Relaxation techniques: k=75 • Health and beauty: k=73 • Skin care: k=71 • Topics on diabetes: k=66 • Health supplies: k=62 • Medication: k=55 • Complications: k=41 • Supplements: k=41 • Fast food: k=39 • Fitness videos: k=37 • Snacks: k=30 • Chain restaurants: k=21 • Prepared foods: k=20 • Beverages: k=18 • Frozen food: k=17 • Groceries: k=17 • Produce: k=16 • Dairy: k=12 • Meat and seafood: k=11 • Health education and technology support: k=8 • Health living: k=7 • Bakery: k=3 • Packaged: k=2

[a]LEARN: LEveraging A viRtual eNvironment to enhance prevention of HIV-related comorbidities.

Health Indicators

Overview

Table 3 provides a summary of health indicators at baseline.

XSL•FO
RenderX

Table 3. LEARN[a] sample health indicator characteristics (n=59).[b]

Characteristics	Total, n (%)[c]	Waitlist control (n=28), n (%)[c]	Intervention (n=31), n (%)[c]	*P* value[d]
Cardiovascular health (history of hypertension or diabetes)				.54
No	42 (71)	21 (75)	21 (68)	
Yes	17 (29)	7 (25)	10 (32)	
Sleep (met recommended hours/night)				.36
No, less than 4 hours	20 (34)	12 (43)	8 (26)	
Somewhat, between 6 and 7 hours	10 (17)	5 (18)	5 (16)	
Yes, 7-9 hours	29 (49)	11 (39)	18 (58)	
BMI (kg/m^2)				.11
Underweight (<18.5)	0 (0)	0 (0.0)	0 (0)	
Normal weight (18.5-24.9)	23 (43)	13 (48)	10 (37)	
Overweight (25.0-29.9)	21 (39)	12 (44)	9 (33)	
Overweight/obese (≥30.0)	10 (19)	2 (7)	8 (30)	
Nutrition intake, mean (SD)				
Vegetable servings	6.4 (4.4)	6.5 (4.7)	6.3 (4.2)	.89
Fruit servings	6.6 (6.8)	7.0 (8.7)	6.3 (4.8)	.70
Whole grain servings	4.2 (5.8)	3.4 (3.1)	5.0 (7.4)	.29
Sweet servings	3.8 (3.4)	3.7 (3.4)	4.0 (3.4)	.71
Fast food meals	1.9 (2.1)	1.6 (2.0)	2.1 (2.1)	.38
Red meat	4.2 (4.1)	5.9 (4.9)	2.7 (2.6)	.004**
Physical activity (International Physical Activity Questionnaire), mean (SD)				
Minutes/week	3725.9 (4151.3)	3601.1 (4346.5)	3869.2 (3992.7)	.81
Walking (minutes)	289.5 (344.7)	331.1 (361.5)	251.9 (330.2)	.38
Moderate activity (minutes)	288.6 (363.9)	302.6 (371.3)	276.0 (362.7)	.78
Vigorous activity (minutes)	199.7 (322.1)	189.9 (285.0)	208.2 (355.8)	.83
Mental health, mean (SD)				.51
9-item Patient Health Questionnaire score	5.7 (6.0)	6.3 (6.2)	5.2 (5.9)	
Smoking status: tobacco or e-cigarette use, n (%)				.67
Current smoker	17 (29)	10 (36)	7 (23)	
Quit 1 to 5 years ago	2 (3)	1 (4)	1 (3)	
Quit ≥5 years ago	9 (15)	5 (18)	4 (13)	
Quit <1 year ago or currently using inhaled nicotine delivery	4 (7)	2 (7)	2 (6)	
Never smoker	27 (46)	10 (36)	17 (55)	
HIV Illness perceptions (Revised Illness Perception Questionnaire), mean (SD)				
Timeline Acute Chronic subscale	7.5 (2.1)	6.9 (1.8)	8.0 (2.3)	.15
Timeline Cyclical subscale	14.6 (3.8)	14.0 (3.9)	15.2 (3.7)	.26
Consequences subscale	15.0 (3.6)	14.8 (3.8)	15.2 (3.4)	.63
Personal Control subscale	10.5 (3.5)	9.5 (3.53)	11.1 (2.7)	.02*
Treatment Control subscale	10.01 (3.0)	11.1 (2.7)	9.1 (2.9)	.007**
Illness Coherence subscale	10.2 (3.8)	10.9 (3.9)	9.5 (3.6)	.18

XSL•FO
RenderX

Characteristics	Total, n (%)[c]	Waitlist control (n=28), n (%)[c]	Intervention (n=31), n (%)[c]	*P* value[d]
Emotional Representations subscale	20.4 (6.4)	18.8 (6.7)	21.9 (5.9)	.07
Number of Symptoms	4.51 (8.8)	5.5 (10.3)	3.7 (7.2)	.43
Other health indicators				
Receive medical care when needed, n (%)				.15
No	10 (17)	3 (11)	7 (23)	
Yes	49 (83)	25 (89)	24 (77)	
Health insurance coverage, n (%)				.80
No	4 (7)	2 (7)	2 (6)	
Yes	54 (93)	25 (93)	29 (94)	
Health literacy, n (%)				.71
Low	8 (14)	3 (11)	5 (16)	
High	51 (86)	25 (89)	26 (84)	
Internalized homophobia, mean (SD)	17.6 (7.1)	18.5 (7.5)	16.7 (6.9)	.33
Sexual orientation concealment, mean (SD)	1.3 (0.5)	1.2 (0.3)	1.3 (0.6)	.65

[a]LEARN: LEveraging A viRtual eNvironment to enhance prevention of HIV-related comorbidities.

[b]Summed counts may differ from the total sample because of missing data.

[c]Column percentages.

[d]Chi-square test if categorical or *t* test for continuous variables.

Cardiovascular Health

Overall, 17 out of 59 (29%) participants reported a history of hypertension or diabetes. A total of 7 (25%) out of 28 participants in the control group reported cardiovascular conditions, compared with 10 out of 31 (32%) in the intervention group. Conversely, 42 (71%) in the total sample reported no history of these conditions (21/28, 75%, in the control group vs 21/31, 68%, in the intervention group).

Sleep

Nearly half of the participants (29/59, 49%) met the recommended 7-9 hours of sleep per night. However, 10 (17%) reported sleeping 6-7 hours per night, and 20 (34%) reported inadequate sleep, with some individuals sleeping fewer than 4 hours per night. These findings suggest a need for improved sleep hygiene among participants who did not meet the recommended guidelines.

BMI

Overall, 23 out of 54 (43%) participants were classified as normal weight, 21 out of 54 (39%) as overweight, and 10 out of 54 (19%) as obese. A higher proportion of participants in the waitlist control group were classified as normal weight (13/27, 48%) or overweight (12/27, 44%) compared with participants in the intervention group (10/27, 37%, and 9/27, 33%, respectively). By contrast, a higher proportion of intervention participants were classified as obese (8/27, 30%) compared with control participants (2/27, 7%).

Nutritional Intake

Vegetable servings averaged 6.4 (SD 4.4) overall, with similar consumption between the control (mean 6.5, SD 4.7) and intervention (mean 6.3, SD 4.2) groups. Fruit servings averaged 6.6 (SD 6.8) overall, with control participants consuming slightly more (mean 7.0, SD 8.7) than intervention participants (mean 6.3, SD 4.8). Whole grain servings averaged 4.2 (SD 5.8) overall, with control participants consuming fewer servings (mean 3.4, SD 3.1) than intervention participants (mean 5.0, SD 7.4). Sweet servings averaged 3.8 (SD 3.4) overall, with similar consumption between the control (mean 3.7, SD 3.4) and intervention (mean 4.0, SD 3.4) groups. Fast-food meals averaged 1.9 (SD 2.1) per week overall, with control participants consuming slightly fewer meals (mean 1.6, SD 2.0) than intervention participants (mean 2.1, SD 2.1). Red meat servings differed significantly between groups, with control participants consuming a mean of 5.9 (SD 4.9) servings compared with 2.7 (SD 2.6) servings among intervention participants.

Physical Activity

Physical activity was assessed using the International Physical Activity Questionnaire. For MET (metabolic equivalent of task)-minutes per week, the mean was 3725.9 (SD 4151.3) for the total sample, indicating moderate-to-high overall physical activity levels. Control participants averaged 3601.1 (SD 4346.5) MET-minutes per week, compared with 3869.2 (SD 3992.7) MET-minutes per week in the intervention group. For minutes spent walking, participants reported a mean of 289.5 (SD 344.7) minutes per week. Control participants averaged 331.1 (SD 361.5) minutes, compared with 251.9 (SD 330.2) minutes in the intervention group. Moderate-intensity physical activity

XSL•FO
RenderX

averaged 288.6 (SD 363.9) minutes per week, with control participants reporting 302.6 (SD 371.3) minutes compared with 276.0 (SD 362.7) minutes in the intervention group. Vigorous-intensity physical activity averaged 199.7 (SD 322.1) minutes per week, with control participants reporting 189.9 (SD 285.0) minutes compared with 208.2 (SD 355.8) minutes in the intervention group.

Mental Health

The mean 9-item Patient Health Questionnaire score was 5.7 (SD 6.0) for the total sample, indicating mild depressive symptoms on average. Scores were similar between groups, with control participants averaging 6.3 (SD 6.2) and intervention participants averaging 5.2 (SD 5.9), with no significant between-group difference.

Tobacco or e-Cigarette

Smoking status was assessed across 3 categories. More than one-quarter of participants (17/59, 29%) were current smokers, 2 (3%) had quit smoking within the past 1-5 years, and 9 (15%) had quit smoking 5 or more years ago. Current smoking rates differed between groups, with 10 out of 28 (36%) control participants reporting current smoking compared with 7 out of 31 (23%) intervention group participants.

HIV Illness Perceptions via IPQR Subscales

Baseline illness perceptions are presented in Table 4. Most IPQR subscale scores were similar between groups. However, 2 subscales demonstrated significant between-group differences: intervention participants reported higher Personal Control scores (mean 11.1, SD 2.7 vs mean 9.5, SD 3.5; P=.02), whereas control participants reported higher Treatment Control beliefs (mean 11.1, SD 2.7 vs mean 9.1, SD 2.9; P=.007). Emotional Representations scores trended higher in the intervention group, although this difference was not statistically significant.

XSL•FO
RenderX

Table 4. LEARN[a] preliminary efficacy results.[b]

Intervention vs control	Delta effect size at 3 months (from baseline)	*t* test (*df*)	*P* value	Delta effect size at 6 months (from baseline)	*t* test (*df*)	*P* value
Cardiovascular health						
BMI	0.03 (–)	0.57	.82	–0.03 (+)	0.61	.83
Nutrition intake						
Vegetable servings	0.59 (+)[c]	1.67	.11	0.66 (+)[c]	1.72	.09
Fruit servings	–0.25 (–)[d]	1.54	.26	0.01 (+)	1.57	.96
Sweets servings	–0.65 (+)[c]	2.18	.31	0.01 (–)	2.25	.98
Whole grains servings	0.41 (+)[c]	1.94	.22	0.46 (+)[c]	2.00	.19
Red meat servings	0.17 (–)	2.13	.74	–0.04 (+)	2.19	.94
Fast food meals	0.17 (–)	1.15	.76	–0.37 (+)[d]	1.18	.52
Physical activity (International Physical Activity Questionnaire)						
Met minutes/week	0.17 (+)	1186.22	.54	0.37 (+)[d]	1216.34	.21
Walking (minutes)	–0.19 (–)	109.06	.55	0.15 (+)	111.71	.65
Moderate activity (minutes)	0.01 (+)	105.80	.98	–0.06 (–)	108.32	.84
Vigorous activity (minutes)	0.38 (+)[d]	102.40	.23	0.57 (+)[c]	104.98	.08
Mental health						
9-item Patient Health Questionnaire total score	0.22 (+)[d]	1.24	.29	–0.05 (–)	1.28	.81
HIV illness perceptions (Revised Illness Perception Questionnaire)						
Timeline Acute Chronic subscale	0.44 (–)[c]	0.66	.17	0.03 (–)	0.69	.94
Timeline Cyclical subscale	0.02 (–)	0.77	.93	0.09 (–)	0.81	.69
Consequences subscale	0.30 (–)[c]	0.93	.25	0.02 (–)	0.97	.95
Personal Control subscale	0.13 (+)	0.83	.60	–0.08 (–)	0.87	.75
Treatment Control subscale	0.13 (+)	0.71	.59	–0.36 (–)[c]	0.74	.15
Illness Coherence subscale	–0.33 (+)[d]	0.88	.16	–0.11 (+)	0.92	.67
Emotional Representations subscale	0.33 (+)[d]	1.12	.07	0.23 (+)[d]	0.18	.22
Number of Symptoms	0.11 (–)	2.62	.70	0.03 (–)	2.74	.93
Other mechanisms						
Internalized Homophobia Total score	–0.03 (+)	1.33	.88	–0.08 (+)	1.36	.68
Sexual Orientation Concealment Scale	0.25 (–)[d]	0.17	.46	0.03 (–)	0.18	.92

[a]LEARN: LEveraging A viRtual eNvironment to enhance prevention of HIV-related comorbidities.

[b]"+" indicates toward the hypothesized direction; "–" indicates away from the hypothesized direction.

[c]Medium effect size (ie, 0.40-0.69).

[d]Small effect size (ie, 0.20-0.39).

Additional Health Indicators

Most participants (49/59, 83%) reported receiving medical care when needed, with a higher proportion of control participants reporting access (25/28, 89%) compared with intervention participants (24/31, 77%). Most participants (54/58, 93%) had health insurance coverage, with 25 out of 27 (93%) control participants and 29 out of 31 (94%) intervention participants reporting coverage. A majority participants demonstrated high health literacy (51/59, 86%), with comparable rates between groups (25/28, 89%, control participants and 26/31, 84%, intervention participants demonstrated high health literacy). The mean Internalized Homophobia Scale total score was 17.6 (SD 7.1) for the total sample. Control participants reported

XSL•FO
RenderX

slightly higher scores (mean 18.5, SD 7.5) than intervention participants (mean 16.7, SD 6.9). Participants reported low levels of sexual orientation concealment, with an overall mean score of 1.3 (SD 0.5). This low score indicates that participants, regardless of group assignment, generally did not conceal their sexual orientation. Scores were nearly identical between the control (mean 1.2, SD 0.3) and intervention (mean 1.3, SD 0.6) groups, suggesting similar levels of openness regarding sexual orientation across groups.

Intervention Effect Size

Overview

Table 4 presents preliminary efficacy results across the health indicators. Below are the effect sizes comparing 3 months with baseline and 6 months with baseline.

Cardiovascular Health

BMI demonstrated minimal change across time points, moving away from the hypothesized direction at 3 months (0.03) and toward the hypothesized direction at 6 months (−0.03).

Nutritional Intake

Vegetable servings showed consistent improvement, with medium effect sizes of 0.59 at 3 months and 0.66 at 6 months, both in the hypothesized direction. Fruit servings showed mixed results, moving away from the hypothesized direction at 3 months with a small effect size (−0.25) but returning toward the hypothesized direction at 6 months (0.01). Sweets consumption showed a notable reduction at 3 months, with a medium effect size (−0.65) in the hypothesized direction; however, this effect was not sustained at 6 months (0.01, away from the hypothesized direction). Similarly, whole grain servings improved with medium effect sizes at both 3 months (0.41) and 6 months (0.46), each in the hypothesized direction. Red meat consumption moved away from the hypothesized direction at 3 months (0.17) but returned toward the hypothesized direction at 6 months (−0.04). Fast-food consumption similarly moved away from the hypothesized direction at 3 months (0.17) but improved by 6 months, with a small effect size (−0.37) in the hypothesized direction.

Physical Activity via International Physical Activity Questionnaire

MET-minutes per week showed progressive improvement, with an effect size of 0.17 at 3 months and a small effect size of 0.37 at 6 months, both in the hypothesized direction. Walking minutes moved away from the hypothesized direction at 3 months (−0.19) but improved toward the hypothesized direction by 6 months (0.15). Moderate-intensity activity minutes showed minimal change at both time points (0.01 at 3 months, in the hypothesized direction, and −0.06 at 6 months, away from the hypothesized direction). Vigorous-intensity activity minutes demonstrated the largest improvements, with a small effect size at 3 months (0.38) and a medium effect size at 6 months (0.57), both in the hypothesized direction.

Mental Health

Total 9-item Patient Health Questionnaire scores for depression showed a small improvement at 3 months, with an effect size of 0.22 in the hypothesized direction, but moved slightly away from the hypothesized direction at 6 months (−0.05).

HIV Illness Perceptions via IPQR

The Timeline Acute/Chronic subscale moved away from the hypothesized direction, with a medium effect size at 3 months (0.44), but moved closer to the hypothesized direction at 6 months (0.03). The Timeline Cyclical subscale moved away from the hypothesized direction at both 3 months (0.02) and 6 months (0.09). The Consequences subscale also moved away from the hypothesized direction, with a small effect size at 3 months (0.30), although this effect was reduced at 6 months (0.02). Personal Control showed improvement at 3 months (0.13, in the hypothesized direction) but moved away from the hypothesized direction at 6 months (−0.08). Treatment Control similarly improved at 3 months (0.13, in the hypothesized direction) but moved away from the hypothesized direction at 6 months, with a small effect size (−0.36). Illness Coherence improved in the hypothesized direction at both 3 months (−0.33, small effect size) and 6 months (−0.11). Emotional Representations improved in the hypothesized direction, with small effect sizes at both 3 months (0.33) and 6 months (0.23). The Number of Symptoms subscale moved away from the hypothesized direction at both 3 months (0.11) and 6 months (0.03).

For internalized homophobia, the intervention group demonstrated improvement at both 3 months (−0.03) and 6 months (−0.08), indicating reductions in internalized negative feelings related to sexual orientation over time. These effect sizes suggest that the intervention may have positively influenced participants' self-perception. By contrast, sexual orientation concealment showed a small effect size at 3 months (0.25), indicating movement away from the hypothesized direction and suggesting increased concealment of sexual orientation. However, this effect size decreased substantially at 6 months (0.03), indicating movement back toward the hypothesized direction and suggesting that concealment levels may have stabilized or improved over time. These trends may have been influenced by the timing of the LEARN study in relation to broader external and societal factors affecting perceptions of sexual orientation.

For the pooled baseline effect size analysis, visit time, age, years since HIV diagnosis, income, and education were included as covariates in the calculation of pooled baseline estimates used to derive effect sizes.

Discussion

Principal Findings

The LEARN Study represents the first known effort to assess the feasibility and acceptability of a VE for cardiovascular and metabolic disease prevention education among men living with HIV. The novelty of this approach lies in its customized platform integration, gamification, and interactive educational quests designed to enhance participant engagement and awareness of heart-healthy behavioral and lifestyle choices. Our findings indicate that the intervention was both feasible to implement

XSL•FO
RenderX

and acceptable to participants, with engagement patterns comparing favorably with those of similar VE interventions.

The findings of the LEARN RCT demonstrated feasibility, as we enrolled 78 participants, achieving 98% of the targeted sample size (N=80). This outcome is noteworthy given that approximately 25% of RCTs fail to meet recruitment goals, even with extended trial periods [56]. The high enrollment rate reflects the effectiveness of the multiple recruitment strategies used, including community partnerships, digital outreach through MyChart messaging campaigns, research listservs, and targeted communications. Retention presented challenges typically associated with technology-driven longitudinal research, with a final retention rate of 61.5% (48 out of 78) at the 6-month assessment. Although modest, our retention outcome exceeded the typical rates observed in remote digital health initiatives, which peak at approximately 56% [57-59]. Attrition primarily occurred before baseline data collection, with 19 out of 78 (24%) participants withdrawing after providing verbal consent, consistent with research indicating that more than half of discontinuations in digital health studies occur within the first week [60]. To enhance retention, various strategies were implemented, including maintaining regular contact through telephone calls and emails and disseminating monthly newsletters celebrating relevant events such as Pride and paying homage to the legacy activists who paved the way (eg, Marsha P. Johnson, Sylvia Rivera, and countless others) toward visibility, representation, and the right to exist. These communications fostered a sense of community connection between the research team and participants. Positive feedback received through “thank-you” emails and phone communications with the study coordinator suggested that maintaining cultural relevance and community acknowledgment, which are key principles of community-based research [61], may be particularly important for engaging underrepresented groups in longitudinal research.

Of note, there were 1 or 2 instances of individuals attempting to enroll multiple times. The study coordinator employed a rigorous, multistep verification process that included phone screenings, email confirmations, checks for geographic inconsistencies, and IP address verification along with location screening. This approach enabled the detection of discrepancies, such as registrations originating from different continents and participants using virtual private networks to mask their actual locations. The challenge of multiple enrollment attempts has become increasingly prevalent in remote research settings, and no evidence-based strategies currently exist to address it [62,63]. However, our comprehensive approach effectively mitigated this risk, resulting in no duplicate enrollments.

Moreover, our findings indicate that the intervention was acceptable. Specifically, engagement metrics demonstrated participants’ willingness to interact with the educational content, with 110 login sessions averaging nearly 2 hours each and 45% (14 out of 31) of intervention participants accessing the VE at least once. These engagement patterns are consistent with findings from similar VE interventions [22,42]. For instance, the SLIDES study, which used a VE for diabetes self-management, reported that participants logged a mean of 2.5 hours per week throughout the study period [42]. The high level of engagement with specific content modules in the LEARN study, particularly nutrition (146 engagements), oral health (138 engagements), fitness tips and videos (82 engagements), and relaxation techniques (75 engagements), suggests that participants were especially interested in informational content that directly affected their daily health choices.

These engagement patterns align with findings from prior VE interventions targeting chronic disease management. For example, the SLIDES study reported that participants logged a mean of 2.5 hours per week over 6 months for diabetes self-management [23,27,42], while Beauchamp et al [22] achieved 93% retention among stroke survivors despite usability challenges, and Touloudi et al [24] demonstrated high acceptance of VE-based exercise in diabetes. However, the LEARN sample differed meaningfully from those in prior studies, as our participants were younger (mean age 41.8 vs 54-59 years), more racially diverse (35/56, 63%, Black; 15/56, 27% Hispanic/Latin American vs 60%-65% White), predominantly single (36/58, 62%, vs 56.3% married), and exclusively male and living with HIV. These demographic differences highlight the importance of tailoring VE interventions to the specific needs and characteristics of target populations, particularly those historically underrepresented in clinical trials and behavioral intervention research. The demographic composition of our sample reflects both the epidemiology of HIV in the United States and our intentional effort to include populations historically underrepresented in clinical trials.

The exploratory effect sizes observed in this study are consistent with, and in some cases exceed, those reported in comparable digital health interventions. For dietary outcomes, medium effect sizes for vegetable consumption (d=0.59 at 3 months to 0.66 at 6 months) and whole grain intake (d=0.41 at 3 months to 0.46 at 6 months) compare favorably with meta-analytic findings from technology-based dietary interventions, which typically report small-to-medium effects [64,65]. The sustained improvement at 6 months suggests that the VE format may support durable behavior change, although the attenuation in sweets reduction (d=–0.65 at 3 months to 0.01 at 6 months) indicates that some dietary behaviors may require ongoing reinforcement. The improvement in vigorous physical activity (d=0.38 at 3 months to d=0.57 at 6 months) is noteworthy and comparable to findings from a meta-analysis of physical activity interventions showing larger effects at 6-9 months, although the extent to which these interventions were delivered virtually or digitally was not specified [66]. The gamified, self-directed nature of the VE may have promoted autonomous motivation, which has been proposed to support sustained physical activity adherence within self-determination theory frameworks [67,68]. We also observed that more participants with higher BMI were retained in the intervention group than in the control group. Although it is plausible that the virtual format reduced weight-related stigma, offered an engaging self-directed experience, or was perceived as more personally relevant by this subgroup, the reasons for this differential retention warrant further study. Moreover, although any improvement resulting from an intervention not directly targeting mental health or

XSL•FO
RenderX

illness perceptions is notable, the small effect sizes for depressive symptoms (d=0.22) and illness perceptions (d=0.23 at 3 months and d=0.33 at 6 months) observed with this low-intensity, self-directed VE intervention warrant further investigation in a fully powered trial. These preliminary shifts in these domains may represent precursors to sustained behavioral change, although larger samples are needed to confirm these effect size patterns.

Limitations

The study has several limitations that warrant consideration. First, the sample was limited to men living with HIV, which may affect the generalizability of the findings to other populations and chronic conditions. Second, the exclusion of women and nonbinary individuals further limits generalizability. This focus was intentional given the epidemiological burden of both HIV and CVD among men, particularly Black and Hispanic men. Future research should examine the feasibility and acceptability of VE-based CVD prevention education among women and gender-diverse populations living with HIV. Third, technical challenges associated with the platform due to its age (eg, freezing, audio problems, and installation difficulties) likely constrained engagement, although the study team worked closely with developers to provide ongoing troubleshooting support. Fourth, given the study's retention rate and the decline in engagement over time, future iterations could address these issues by refining screening and inclusion criteria to identify characteristics associated with participant attrition and by developing a new platform. Future versions should also incorporate additional strategies to sustain participant engagement, such as personalized content, regular updates, and incentives for continued participation. Finally, the exploratory nature of the secondary outcomes means that the study was not powered to detect significant changes among the health indicators examined. Future research should build on these exploratory findings to develop and test more targeted interventions aimed at improving specific health behaviors and outcomes.

Measures of Rigor for Clinical Trials

The methodological rigor of this study helps balance these limitations with several notable strengths. Intervention fidelity was maintained through a comprehensive approach addressing 5 components: (1) intervention design, (2) study staff training, (3) intervention delivery, (4) participant receipt, and (5) outcome assessment. First, all study staff completed ethical research training, participated in role-playing exercises for participant recruitment and enrollment, and followed a detailed scope-of-work document outlining research-related tasks and responsibilities. Second, participant receipt of the intervention was objectively measured using login frequency, engagement duration, and log file data capturing specific activities within the VE. Third, the study coordinator, trained through the Clinical and Translational Science Award program, ensured consistent participant screening, enrollment, consent, and data collection using standardized scripts. Fourth, a statistician designed the permuted block randomization scheme, and although the PI had access to participant allocation, neither the PI nor the study coordinator participated in the data analysis. This approach was intended to minimize bias in outcome ascertainment [69]. Fifth, data collection and storage were conducted using REDCap, a HIPAA-compliant, validated platform recognized for reliability in clinical research. Sixth, the study's educational content was grounded in the American Heart Association Life's Essential 8 framework [38] for cardiovascular health, lending credibility to the prevention messaging delivered through the VE [60]. Finally, the novel approach to cardiovascular and metabolic disease prevention education using gamification and interactive quests distinguishes this intervention from traditional clinical patient education. The high enrollment rate, methodological rigor, and use of validated measures strengthen the validity of the findings. The sample's demographic composition (ie, Black and Hispanic men living with HIV) addresses a critical gap in clinical trial representation. Engagement metrics demonstrating average sessions of nearly 2 hours further highlight participants' favorable reception of the VE content, supporting the potential of this approach to reach underserved populations through innovative health education.

Conclusions

The LEARN Study demonstrated the feasibility and acceptability of a VE designed for cardiovascular and metabolic disease prevention education among persons living with HIV. High engagement with specific educational content and favorable retention rates highlight the promise of this innovative approach. Preliminary effect sizes suggested positive trends in cardiovascular health indicators and mental well-being, indicating the potential for additional benefits associated with the intervention. These findings support progression to a fully powered RCT to evaluate the efficacy of VE-based CVD prevention education. Subsequent funding has been awarded based on this work to expand the research. In the subsequent study, LEARN2, we will integrate lessons learned from this trial and focus on HIV-related conditions with shared risk factors to enhance the intervention's impact. By addressing HIV-related conditions with interrelated risk factors, we aim to provide a more comprehensive intervention to improve the overall health and well-being of individuals living with HIV.

Acknowledgments

We thank Applied Research Associates for their technical support. SRR also thanks the K01 mentorship team for their guidance and support throughout the research process. No generative artificial intelligence tools were used in any aspect of the manuscript revision, including drafting response letters or refining the text. All content was developed and reviewed by the authors, who take full responsibility for the final manuscript.

XSL•FO
RenderX

Data Availability

The datasets generated and analyzed during this study are not publicly available because of the sensitive nature of HIV-related health information and the need to protect participant privacy. Deidentified data may be made available from the corresponding author (SRR) upon reasonable request and subject to appropriate data use agreements.

Funding

This study was supported by the National Heart, Lung, and Blood Institute (grant K01HL145580, principal investigator SRR). The funder had no role in the study design; data collection, analysis, or interpretation; the decision to publish; or preparation of the manuscript.

Conflicts of Interest

SRR is an Associate Editor for JMIR. CJ and AV are the developers of the LIVE (Learning in a Virtual Environment) platform.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 3290 KB - jmir_v28i1e91145_app1.pdf]

Multimedia Appendix 2
LEARN (LEveraging A viRtual eNvironment to enhance prevention of HIV-related comorbidities) acceptability results.
[DOCX File , 15 KB - jmir_v28i1e91145_app2.docx]

References

1. Estimated HIV incidence and prevalence in the United States, 2017-2021: HIV surveillance supplemental report. Centers for Disease Control and Prevention (CDC). 2023. URL: https://stacks.cdc.gov/view/cdc/149080 [accessed 2026-05-25]
2. Centers for Disease Control and Prevention (CDC). HIV prevalence estimates--United States, 2006. Morbidity and Mortality Weekly Report 2008;57(39):1073-1076 [FREE Full text]
3. Tsao C, Aday A, Almarzooq Z, Anderson C, Arora P, Avery C, American Heart Association Council on EpidemiologyPrevention Statistics CommitteeStroke Statistics Subcommittee. Heart disease and stroke statistics-2023 update: a report from the American Heart Association. Circulation 2023 Feb 21;147(8):e93-e621. [doi: 10.1161/CIR.0000000000001123] [Medline: 36695182]
4. Feinstein MJ, Bahiru E, Achenbach C, Longenecker CT, Hsue P, So-Armah K, et al. Patterns of cardiovascular mortality for HIV-infected adults in the United States: 1999 to 2013. Am J Cardiol 2016 Jan 15;117(2):214-220 [FREE Full text] [doi: 10.1016/j.amjcard.2015.10.030] [Medline: 26639041]
5. Losina E, Hyle E, Borre E, Linas B, Sax P, Weinstein M, et al. Projecting 10-year, 20-year, and lifetime risks of cardiovascular disease in persons living with human immunodeficiency virus in the United States. Clin Infect Dis 2017 Oct 15;65(8):1266-1271 [FREE Full text] [doi: 10.1093/cid/cix547] [Medline: 28605504]
6. Titanji B, Gavegnano C, Hsue P, Schinazi R, Marconi VC. Targeting inflammation to reduce atherosclerotic cardiovascular risk in people with HIV infection. JAHA 2020 Feb 04;9(3):e014873. [doi: 10.1161/jaha.119.014873]
7. As HIV patients live longer, heart disease might be their next challenge. American Heart Association. 2019 Oct 22. URL: https://www.heart.org/en/news/2019/06/03/as-hiv-patients-live-longer-heart-disease-might-be-their-next-challenge [accessed 2026-05-14]
8. Feinstein MJ, Hsue PY, Benjamin LA, Bloomfield GS, Currier JS, Freiberg MS, et al. Characteristics, prevention, and management of cardiovascular disease in people living with HIV: a scientific statement from the American Heart Association. Circulation 2019 Jul 09;140(2):e98-e124. [doi: 10.1161/cir.0000000000000695]
9. Gosiker BJ, Lesko CR, Rich AJ, Crane HM, Kitahata MM, Reisner SL, et al. Cardiovascular disease risk among transgender women living with HIV in the United States. PLoS One 2020 Jul 20;15(7):e0236177 [FREE Full text] [doi: 10.1371/journal.pone.0236177] [Medline: 32687532]
10. Seeman TE. Price of adaptation—allostatic load and its health consequences. Arch Intern Med 1997 Oct 27;157(19):2259. [doi: 10.1001/archinte.1997.00440400111013]
11. Duong MT, Bingham BA, Aldana PC, Chung ST, Sumner AE. Variation in the calculation of Allostatic Load Score: 21 examples from NHANES. J Racial Ethn Health Disparities 2017 Jun 28;4(3):455-461. [doi: 10.1007/s40615-016-0246-8] [Medline: 27352114]
12. Fazeli P, Waldrop-Valverde D, Yigit I, Turan B, Edberg J, Kempf M, et al. An exploratory study of correlates of allostatic load in older people living with HIV. J Acquir Immune Defic Syndr 2020 Apr 15;83(5):441-449. [doi: 10.1097/QAI.0000000000002293] [Medline: 31914006]
13. Thayer J, Carnevali L, Sgiofo A, Williams D. Angry in America: psychophysiological responses to unfair treatment. Ann Behav Med 2020 Dec 01;54(12):924-931. [doi: 10.1093/abm/kaaa094] [Medline: 33416840]

XSL•FO
RenderX

14. Hill LK, Thayer JF. The autonomic nervous system and hypertension: ethnic differences and psychosocial factors. Curr Cardiol Rep 2019 Feb 28;21(3):15 [FREE Full text] [doi: 10.1007/s11886-019-1100-5] [Medline: 30820683]
15. Hsue PY, Waters DD. Time to recognize HIV infection as a major cardiovascular risk factor. Circulation 2018 Sep 11;138(11):1113-1115. [doi: 10.1161/circulationaha.118.036211]
16. Hill LK, Hoggard LS, Richmond AS, Gray DL, Williams DP, Thayer JF. Examining the association between perceived discrimination and heart rate variability in African Americans. Cultur Divers Ethnic Minor Psychol 2017 Jan;23(1):5-14 [FREE Full text] [doi: 10.1037/cdp0000076] [Medline: 28045306]
17. Williams DP, Joseph N, Hill LK, Sollers JJ, Vasey MW, Way BM, et al. Stereotype threat, trait perseveration, and vagal activity: evidence for mechanisms underpinning health disparities in Black Americans. Ethn Health 2019 Nov 19;24(8):909-926 [FREE Full text] [doi: 10.1080/13557858.2017.1378803] [Medline: 28922935]
18. Shah AS, Stelzle D, Lee KK, Beck EJ, Alam S, Clifford S, et al. Global burden of atherosclerotic cardiovascular disease in people living with HIV. Circulation 2018 Sep 11;138(11):1100-1112. [doi: 10.1161/circulationaha.117.033369]
19. Smit M, Brinkman K, Geerlings S, Smit C, Thyagarajan K, Sighem AV, et al. Future challenges for clinical care of an ageing population infected with HIV: a modelling study. The Lancet Infectious Diseases 2015 Jul;15(7):810-818. [doi: 10.1016/s1473-3099(15)00056-0]
20. Center P. Internet, broadband fact sheet. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/fact-sheet/internet-broadband/ [accessed 2026-05-21]
21. Kakulla B. Gamers 50-plus are a growing force in the tech market. AARP. 2023. URL: https://www.aarp.org/pri/topics/social-leisure/activities-interests/2023-gamers-50-plus/ [accessed 2026-05-24]
22. Beauchamp JE, Wang M, Leon Novelo LG, Cox C, Meyer T, Fagundes C, et al. Feasibility and user-experience of a virtual environment for social connection and education after stroke: a pilot study. J Stroke Cerebrovasc Dis 2024 Feb;33(2):107515. [doi: 10.1016/j.jstrokecerebrovasdis.2023.107515] [Medline: 38064972]
23. Johnson CM, D'Eramo Melkus G, Reagan L, Pan W, Amarasekara S, Pereira K, et al. Learning in a virtual environment to improve type 2 diabetes outcomes: randomized controlled trial. JMIR Form Res 2023 Apr 20;7:e40359 [FREE Full text] [doi: 10.2196/40359] [Medline: 36962700]
24. Touloudi E, Stavrou V, Galanis E, Bargiota A, Goudas M, Dafoulas G, et al. Virtual reality versus conventional exercise in patients with type 1 diabetes: a feasibility randomized crossover trial. Virtual Worlds 2025 Jul 08;4(3):32. [doi: 10.3390/virtualworlds4030032]
25. Casarez RL, Johnson CM, Soares JC, Meyer TD. Use of a virtual environment to promote self-management and lifestyle changes in persons with bipolar disorder. Arch Psychiatr Nurs 2024 Apr;49:73-82. [doi: 10.1016/j.apnu.2024.02.003] [Medline: 38734458]
26. Brewer LC, Kaihoi B, Zarling KK, Squires RW, Thomas R, Kopecky S. The use of virtual world-based cardiac rehabilitation to encourage healthy lifestyle choices among cardiac patients: intervention development and pilot study protocol. JMIR Res Protoc 2015 Apr 08;4(2):e39 [FREE Full text] [doi: 10.2196/resprot.4285] [Medline: 25857331]
27. Johnson C, Feenan K, Setliff G, Pereira K, Hassell N, Beresford H, et al. Building a virtual environment for diabetes self-management education and support. Int J Virtual Communities Soc Netw 2013;5(3):68-80 [FREE Full text] [doi: 10.4018/ijvcsn.2013070105] [Medline: 25699133]
28. Johnson CM, McIlwain S, Gray O, Willson B, Vorderstrasse A. Creating a sustainable collaborative consumer health application for chronic disease self-management. J Biomed Inform 2017 Jul;71:198-206 [FREE Full text] [doi: 10.1016/j.jbi.2017.06.004] [Medline: 28600025]
29. Bandura A. Social cognitive theory: an agentic perspective. Annu Rev Psychol 2001 Feb;52(1):1-26. [doi: 10.1146/annurev.psych.52.1.1] [Medline: 11148297]
30. Deci E, Ryan R. Self-determination theory. In: Handbook of Theories of Social Psychology. London, UK: SAGE Publications Ltd; 2012:416-436.
31. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
32. Ramos SR, Reynolds H, Johnson C, Melkus G, Kershaw T, Thayer JF, et al. Perceptions of HIV-related comorbidities and usability of a virtual environment for cardiovascular disease prevention education in sexual minority men with HIV: formative phases of a pilot randomized controlled trial. J Med Internet Res 2024 Aug 22;26:e57351 [FREE Full text] [doi: 10.2196/57351] [Medline: 38924481]
33. Ramos SR, Johnson C, Melkus G, Kershaw T, Gwadz M, Reynolds H, et al. Cardiovascular disease prevention education using a virtual environment in sexual-minority men of color with HIV: protocol for a sequential, mixed method, waitlist randomized controlled trial. JMIR Res Protoc 2022 May 17;11(5):e38348 [FREE Full text] [doi: 10.2196/38348] [Medline: 35579928]
34. Sidani S. Health Intervention Research: Understanding Research Design and Methods. Oxford, UK: Sage Publications Ltd; 2015.

XSL•FO
RenderX

35. Ramos SR, Warren R, Shedlin M, Melkus G, Kershaw T, Vorderstrasse A. A framework for using eHealth interventions to overcome medical mistrust among sexual minority men of color living with chronic conditions. Behav Med 2019 Jul 25;45(2):166-176 [FREE Full text] [doi: 10.1080/08964289.2019.1570074] [Medline: 31343963]
36. Ramos SR, O'Hare OM, Hernandez Colon A, Kaplan Jacobs S, Campbell B, Kershaw T, et al. Purely behavioral: a scoping review of nonpharmacological behavioral and lifestyle interventions to prevent cardiovascular disease in persons living with HIV. J Assoc Nurses AIDS Care 2021 Jan 20;32(5):536-547. [doi: 10.1097/jnc.0000000000000230]
37. Wright JS, Wall HK, Ritchey MD. Million hearts 2022: small steps are needed for cardiovascular disease prevention. JAMA 2018 Nov 13;320(18):1857-1858 [FREE Full text] [doi: 10.1001/jama.2018.13326] [Medline: 30193304]
38. Lloyd-Jones DM, Allen NB, Anderson CA, Black T, Brewer LC, Foraker RE, et al. Life's essential 8: updating and enhancing the American Heart Association's construct of cardiovascular health: a Presidential Advisory from the American Heart Association. Circulation 2022 Aug 02;146(5):e18-e43. [doi: 10.1161/cir.0000000000001078]
39. Matts JP, Lachin JM. Properties of permuted-block randomization in clinical trials. Control Clin Trials 1988 Dec;9(4):327-344. [doi: 10.1016/0197-2456(88)90047-5] [Medline: 3203524]
40. Hertzog MA. Considerations in determining sample size for pilot studies. Res Nurs Health 2008 Apr 08;31(2):180-191. [doi: 10.1002/nur.20247] [Medline: 18183564]
41. Cook N, Winkler SL. Acceptance, usability and health applications of virtual worlds by older adults: a feasibility study. JMIR Res Protoc 2016 Jun 02;5(2):e81 [FREE Full text] [doi: 10.2196/resprot.5423] [Medline: 27256457]
42. Johnson C, Feinglos M, Pereira K, Hassell N, Blascovich J, Nicollerat J, et al. Feasibility and preliminary effects of a virtual environment for adults with type 2 diabetes: pilot study. JMIR Res Protoc 2014 Apr 08;3(2):e23 [FREE Full text] [doi: 10.2196/resprot.3045] [Medline: 24713420]
43. Consensus Conference Panel, Watson NF, Badr MS, Belenky G, Bliwise DL, Buxton OM, et al. Joint Consensus Statement of the American Academy of Sleep Medicine and Sleep Research Society on the recommended amount of sleep for a healthy adult: methodology and discussion. J Clin Sleep Med 2015 Aug 15;11(8):931-952 [FREE Full text] [doi: 10.5664/jcsm.4950] [Medline: 26235159]
44. World Health Organization (WHO). Global Health Observatory (GHO) data repository: body mass index. WHO. 2020. URL: https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/body-mass-index [accessed 2026-05-21]
45. What is MyPlate: what foods are in the fruit group? MyPlate. 2025. URL: https://www.myplate.gov/ [accessed 2026-01-26]
46. What is MyPlate: what foods are in the vegetable group? MyPlate. 2025. URL: https://www.myplate.gov/ [accessed 2026-01-26]
47. Center for Disease Control and Prevention (CDC). Adult activity: an overview. CDC. 2025. URL: https://www.cdc.gov/physical-activity-basics/guidelines/adults.html [accessed 2026-01-26]
48. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
49. Center for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System (BRFSS) questionnaire. CDC. 2023. URL: https://www.cdc.gov/brfss/questionnaires/index.htm [accessed 2026-01-22]
50. Moss-Morris R, Weinman J, Petrie K, Horne R, Cameron L, Buick D. The Revised Illness Perception Questionnaire (IPQ-R). Psychology & Health 2002 Jan;17(1):1-16. [doi: 10.1080/08870440290001494]
51. Agency for Healthcare Research and Quality (AHRQ). SAHL-E keys for health literacy measurement tools. AHRQ. 2019. URL: https://www.ahrq.gov/health-literacy/research/tools/sahl-e-keys.html [accessed 2026-05-25]
52. LaPollo AB, Bond L, Lauby JL. Hypermasculinity and sexual risk among Black and White men who have sex with men and women. Am J Mens Health 2014 Sep 01;8(5):362-372 [FREE Full text] [doi: 10.1177/1557988313512861] [Medline: 24297456]
53. Martin J, Dean L. Ego-dystonic homosexuality scale. Unpublished manuscript, Columbia University 1987:1.
54. Jackson SD, Mohr JJ. Conceptualizing the closet: differentiating stigma concealment and nondisclosure processes. Psychology of Sexual Orientation and Gender Diversity 2016 Mar;3(1):80-92. [doi: 10.1037/sgd0000147]
55. SAS Institute Inc. In: SAS. North Carolina: SAS; Jun 11, 2025.
56. Ng C, Bowman S, Ling J, Bagshaw R, Birt A, Yiannakou Y. The future of clinical trials-is it virtual? Br Med Bull 2023 Dec 11;148(1):42-57. [doi: 10.1093/bmb/ldad022] [Medline: 37681298]
57. Griffith Fillipo IR, Pullmann MD, Hull TD, Zech J, Wu J, Litvin B, et al. Participant retention in a fully remote trial of digital psychotherapy: comparison of incentive types. Front Digit Health 2022 Sep 6;4:963741 [FREE Full text] [doi: 10.3389/fdgth.2022.963741] [Medline: 36148211]
58. Frampton GK, Shepherd J, Pickett K, Griffiths G, Wyatt JC. Digital tools for the recruitment and retention of participants in randomised controlled trials: a systematic map. Trials 2020 Jun 05;21(1):478 [FREE Full text] [doi: 10.1186/s13063-020-04358-3] [Medline: 32498690]
59. Bell ML, Kenward MG, Fairclough DL, Horton NJ. Differential dropout and bias in randomised controlled trials: when it matters and when it may not. BMJ 2013 Jan 21;346(jan21 1):e8668-e8668 [FREE Full text] [doi: 10.1136/bmj.e8668] [Medline: 23338004]

XSL•FO
RenderX

60. Pratap A, Neto EC, Snyder P, Stepnowsky C, Elhadad N, Grant D, et al. Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. NPJ Digit Med 2020 Feb 17;3(1):21 [FREE Full text] [doi: 10.1038/s41746-020-0224-8] [Medline: 32128451]
61. Israel B, Schulz A, Parker E, Becker A, Community-Campus Partnerships for Health. Community-based participatory research: policy recommendations for promoting a partnership approach in health research. Educ Health (Abingdon) 2001 Feb 01;14(2):182-197. [doi: 10.1080/13576280110051055] [Medline: 14742017]
62. Burwell N, Mbakop LC, Branch K, Lowenthal ED, Clark RRS, Nightingale KJ. Recruitment strategies to prevent fraudulent virtual enrollment while respecting privacy in vulnerable populations. Qual Health Res 2025 Oct 31:10497323251378095 [FREE Full text] [doi: 10.1177/10497323251378095] [Medline: 41172108]
63. Comachio J, Poulsen A, Bamgboje-Ayodele A, Tan A, Ayre J, Raeside R, et al. Identifying and counteracting fraudulent responses in online recruitment for health research: a scoping review. BMJ Evid Based Med 2025 May 20;30(3):173-182 [FREE Full text] [doi: 10.1136/bmjebm-2024-113170] [Medline: 39715631]
64. Krebs P, Prochaska JO, Rossi JS. A meta-analysis of computer-tailored interventions for health behavior change. Prev Med 2010 Sep;51(3-4):214-221 [FREE Full text] [doi: 10.1016/j.ypmed.2010.06.004] [Medline: 20558196]
65. Barnett A, Wright C, Stone C, Ho NY, Adhyaru P, Kostjasyn S, et al. Effectiveness of dietary interventions delivered by digital health to adults with chronic conditions: systematic review and meta-analysis. J Hum Nutr Diet 2023 Jun 16;36(3):632-656. [doi: 10.1111/jhn.13125] [Medline: 36504462]
66. Murray JM, Brennan SF, French DP, Patterson CC, Kee F, Hunter RF. Effectiveness of physical activity interventions in achieving behaviour change maintenance in young and middle aged adults: a systematic review and meta-analysis. Soc Sci Med 2017 Nov;192:125-133. [doi: 10.1016/j.socscimed.2017.09.021] [Medline: 28965003]
67. Deci E, Ryan R. Intrinsic Motivation and Self-Determination in Human Behavior. London, UK: Springer; May 30, 1985.
68. Teixeira PJ, Carraça EV, Markland D, Silva MN, Ryan RM. Exercise, physical activity, and self-determination theory: a systematic review. Int J Behav Nutr Phys Act 2012 Jun 22;9(1):78 [FREE Full text] [doi: 10.1186/1479-5868-9-78] [Medline: 22726453]
69. Browner W, Newman T, Cummings S, Grady D. Designing Clinical Research. Philadelphia, PA: Lippincott Williams & Wilkins; May 19, 2022.

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
CVD: cardiovascular disease
HIPAA: Health Insurance Portability and Accountability Act
IPQR: Revised Illness Perception Questionnaire
LEARN: LEveraging A viRtual eNvironment to enhance prevention of HIV-related comorbidities
LIVE: Learning in a Virtual Environment
MET: metabolic equivalent of task
PI: principal investigator
RCT: randomized controlled trial
SLIDES: Second Life Impacts Diabetes Education & Self-Management
VE: virtual environment

Edited by A Mavragani; submitted 09.Jan.2026; peer-reviewed by C Xiao, A Ramadas, I Steenstra; comments to author 28.Apr.2026; revised version received 08.May.2026; accepted 19.May.2026; published 15.Jun.2026.

Please cite as:
Ramos SR, Boerger L, Restar AJ, Phadke M, Jeon S, Johnson C, Melkus G, Kershaw T, Reynolds H, Vorderstrasse A
A Web-Based Virtual Environment Behavioral Intervention as Cardiovascular Disease and Metabolic Disease Prevention Education in Persons With HIV: Evaluation of the LEARN Randomized Controlled Trial
J Med Internet Res 2026;28:e91145
URL: https://www.jmir.org/2026/1/e91145
doi:10.2196/91145
PMID:42154285

©S Raquel Ramos, Lindsie Boerger, Arjee Javellana Restar, Manali Phadke, Sangchoon Jeon, Constance Johnson, Gail Melkus, Trace Kershaw, Harmony Reynolds, Allison Vorderstrasse. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any

XSL•FO
RenderX

medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Acceptability and Preliminary Evaluation of a Campus-Integrated Digital Platform (Fruto) for University Students' Mental Health Help-Seeking: Sequential Mixed Methods Study

Myungsung Kim[1], BS; Hyorim Kim[1], BS; Jeong-in Heo[2], MS; Seonmi Lee[2], PhD; Orane Farrah Lahcine[1], BS; Sangil Lee[3], PhD; Hwang Kim[4], PhD; Dooyoung Jung[2,5], PhD, MD

[1]Graduate School of Health Science and Technology, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea
[2]School of Digital Humanities and Computational Social Sciences, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
[3]School of Liberal Arts, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea
[4]Department of Design, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea
[5]Mind Care & Growth Center, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea

Corresponding Author:
Dooyoung Jung, PhD, MD
School of Digital Humanities and Computational Social Sciences
Korea Advanced Institute of Science and Technology
291 Daehak-ro, Yuseong-gu
Daejeon, 34141
Republic of Korea
Phone: 82 01097970807
Email: dooyoung@kaist.ac.kr

Abstract

Background: College students are at heightened risk for mental health problems; yet, professional help-seeking remains low. Although digital mental health tools can improve accessibility and reduce stigma, many focus on isolated functions, such as self-screening, psychoeducation, or symptom management, and are not fully integrated with campus counseling services. Multidomain platforms that combine screening, mental health information, counseling access, and campus service navigation may support help-seeking in university settings; however, their acceptability and implementation value remain underexplored.

Objective: This mixed methods study examined the refinement and preliminary evaluation of Fruto, a campus-integrated, multidomain app developed to support university students' help-seeking attitudes and counseling-related beliefs in a real-world counseling-center setting. Phase 1 used scenario-based prototype sessions to analyze students' interactions with the platform and inform refinement. Phase 2 assessed whether 8 weeks of Fruto use was associated with pre-post changes in attitudinal and counseling-related outcomes.

Methods: We conducted a 2-phase, mixed methods study. Phase 1 involved vignette-based prototype sessions and semistructured interviews with 16 students to explore user experiences with an early version of Fruto. Scenario-based tasks facilitated feedback on the platform, and thematic analysis identified design implications that guided refinement. Phase 2 involved an 8-week single-group pre-post evaluation. A total of 109 students completed the baseline survey, and 70 provided follow-up responses. Surveys assessed help-seeking attitudes, counseling-related beliefs, and perceived app quality. Linear mixed effects models examined pre-post changes using all available data, and exploratory baseline-adjusted regressions examined the association between overall perceived app quality and postuse outcomes.

Results: Scenario-based prototype sessions elicited actionable feedback on how students might use Fruto in realistic help-seeking contexts. Qualitative findings identified 3 refinement priorities that informed subsequent app updates: trusted and identifiable content providers, seamless integration across app features, and relatable self-discovery content to lower psychological barriers to app use. Following these refinements, Phase 2 assessed 8-week pre-post changes using linear mixed effects models. Fruto use was associated with significant increases in positive help-seeking attitudes (B=0.884, SE 0.284, 95% CI 0.327-1.441; *P*=.002) and positive counseling expectations (B=1.585, SE 0.541, 95% CI 0.526-2.645; *P*=.003). No significant changes were observed in negative attitudes, negative counseling beliefs, or socially supportive beliefs. In exploratory baseline-adjusted regressions, overall perceived app quality was associated with positive counseling expectations, but not with positive help-seeking attitudes.

XSL•FO
RenderX

Conclusions: Fruto shows promise as a campus-integrated, multidomain platform associated with more favorable help-seeking attitudes and counseling expectations among university students. These findings suggest that multidomain platforms may strengthen positive, approach-oriented beliefs toward professional support. Future studies with longer follow-up and objective usage or service-use data are needed to examine whether attitudinal changes translate into help-seeking behavior.

Trial Registration: Clinical Research Information Service KCT0010622; https://cris.nih.go.kr/cris/search/detailSearch.do?seq=30274&search_page=L

(J Med Internet Res 2026;28:e78930) doi:10.2196/78930

KEYWORDS

app usability; digital mental health; help-seeking; mixed methods; mobile health; multidomain platform; stigma reduction; university students

Introduction

Mental Health Challenges and Help-Seeking Behavior Among University Students

During the transition from adolescence to adulthood, university students face new roles and environments while managing social, economic, and academic pressures. This life stage, often termed "emerging adulthood" [1], is also when most lifetime mental illnesses first manifest, making it a critical developmental period with substantial individual and societal consequences [2]. Mental health problems among university students are therefore well-documented and pose a significant threat to well-being. In the United States, over one-third of students report depressive symptoms that impair their daily functioning, and approximately 10% seriously consider suicide each year [3]. Yet, despite growing recognition of this burden, the implementation of accessible, acceptable, and student-centered support systems remains limited, particularly regarding how universities can encourage timely professional help-seeking before students' difficulties become severe.

These concerns are especially relevant in South Korea, where university student mental health must be understood within the country's specific context of high psychological distress, academic pressure, and limited use of professional mental health services [4]. South Korea continues to report suicide rates above the Organisation for Economic Co-operation and Development average, and Korean university students report substantial distress related to interpersonal relationships, employment, and career decisions, with prior Korean studies documenting considerable employment- and career-related distress, including job-seeking stress, depressive symptoms, and job-search anxiety [5-7]. According to the Ministry of Health and Welfare, only 22.2% of Koreans seek professional mental health services, compared with 41.3% in the United States and 46.5% in Canada [8]. This treatment gap is concerning because delayed help-seeking may allow student distress to worsen before they reach campus counseling or professional support.

Low use of professional mental health services among students does not necessarily reflect a lack of interest in support. Prior evidence suggests a gap between students' help-seeking intentions and their actual use of professional services, indicating that many students may recognize a need for help yet still face barriers to accessing care. These barriers are multifaceted and include a preference to manage problems independently, lack of time, limited knowledge of available resources, and concerns about privacy or stigma. In Asian cultural contexts, where maintaining social reputation is highly valued, students may also be reluctant to disclose mental health concerns [9-11]. These cultural and practical barriers can reinforce self-stigma, delay professional help-seeking, and reduce students' likelihood of accessing support before difficulties worsen [12,13].

Support Programs and Online Interventions

As awareness of student mental health concerns grows, universities worldwide are facing increasing demand for mental health services [14]. On-campus counseling centers play a key role in identifying and addressing students' emotional challenges, and face-to-face counseling has been shown to reduce symptoms of depression, anxiety, hostility, academic distress, and social anxiety [15]. Universities are increasingly adopting stepped-care models that match students to different levels of support according to need, integrating peer support, group therapy, and psychiatric care alongside traditional face-to-face counseling [16]. This approach can extend support beyond individual counseling by offering earlier, more accessible, and less resource-intensive options at different points in students' help-seeking needs.

Within this stepped-care approach, digital mental health interventions can help expand support by offering students accessible, low-intensity, and less stigmatizing support options [17]. University students are often comfortable using digital platforms and may value the anonymity and flexibility they provide, making online support particularly appealing [12]. Moreover, beyond symptom-focused digital interventions, help-seeking-oriented programs have aimed to improve mental health literacy and increase students' willingness to seek professional support [18]. These approaches often use psychoeducation or self-screening to help students recognize mental health concerns and understand the value of support [19,20]. Still, awareness alone may be insufficient if students do not know how to interpret screening results or what concrete steps to take afterward. Self-screening tools, in particular, may lead to confusion or inaction when they are not paired with clear follow-up guidance.

Yet many existing digital programs focus on single clinical targets, such as depression, anxiety, or general well-being, and are frequently delivered as stand-alone self-help or mobile-based cognitive behavioral therapy [12]. In addition, when digital tools are not aligned with the counseling approaches and service

XSL•FO
RenderX

structures already used on campus, a mismatch may emerge between app-based support and offline care pathways [21]. Although these tools can be useful, they may not fully address students' broader help-seeking needs, such as understanding available services, interpreting self-screening results, reducing hesitation, and taking concrete steps toward campus-based support. Therefore, digital platforms designed for university settings should not only provide information or screening, but also connect students to appropriate next steps, including campus counseling and other in-person support services [21].

Study Purpose

Taken together, these gaps suggest a need for campus-integrated digital platforms that do more than provide isolated information or symptom-focused self-help. In response, this study examines the real-world acceptability, user-informed refinement, and preliminary evaluation of Fruto, a campus-integrated, multidomain platform developed to support university students' mental health help-seeking attitudes and counseling-related beliefs. Fruto incorporates elements commonly used in digital mental health tools, such as psychoeducation and self-screening, but embeds them within a unified help-seeking ecosystem linked to university counseling services.

In the initial stage of the study, we conducted vignette-based semistructured interviews followed by thematic analysis to examine how students perceived and interacted with an early version of Fruto. This phase addressed 2 research questions (RQs): (RQ1) how university students with varying usage motivations engage with and perceive multidomain help-seeking platforms, and (RQ2) what design principles are essential for ensuring that such a platform functions effectively within a real-world campus environment.

Based on the initial findings, we refined Fruto and conducted a pre-post evaluation among campus student users to examine whether platform use was associated with changes in help-seeking attitudes and counseling-related beliefs. This second phase addressed one research question: (RQ3) whether the use of a multidomain help-seeking platform is associated with changes in university students' help-seeking attitudes and counseling-related beliefs. To address this quantitative research question, we established two hypotheses: (H1) students who use Fruto for 8 weeks will show improved attitudes toward seeking professional help, and (H2) students who use Fruto for 8 weeks will demonstrate more positive perceptions of counseling.

In summary, this study first used scenario-based qualitative methods to examine how students engage with a theoretically grounded, multidomain help-seeking platform (RQ1) and to identify design implications for real-world campus implementation (RQ2). These findings informed the refinement of Fruto before a campus-wide pre-post evaluation. The study then examined whether 8 weeks of Fruto use was associated with changes in help-seeking attitudes and counseling-related beliefs (RQ3). This integrated approach aims to clarify how multidomain digital platforms can support help-seeking attitudes and counseling-related beliefs in university settings.

Methods

Multidomain Platform: Fruto

We conceptualize Fruto as a multidomain help-seeking platform, distinct from traditional multicomponent apps that deliver therapeutic tools or multimodule apps that provide sequential intervention content. Each feature within Fruto was designed to support students across different problem and functional domains, thereby facilitating multiple pathways toward mental health help-seeking. The platform includes 5 core features, each developed in-house by the research team and integrated into a cohesive user interface (UI). The home screen includes a navigation bar linking to each core function, along with key announcements and informational resources. The platform design was informed by existing literature on help-seeking behavior and grounded in relevant theoretical frameworks. Figure 1 presents the main UI components. Full UI screenshots and feature descriptions are provided in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. Main user interface components of Fruto.

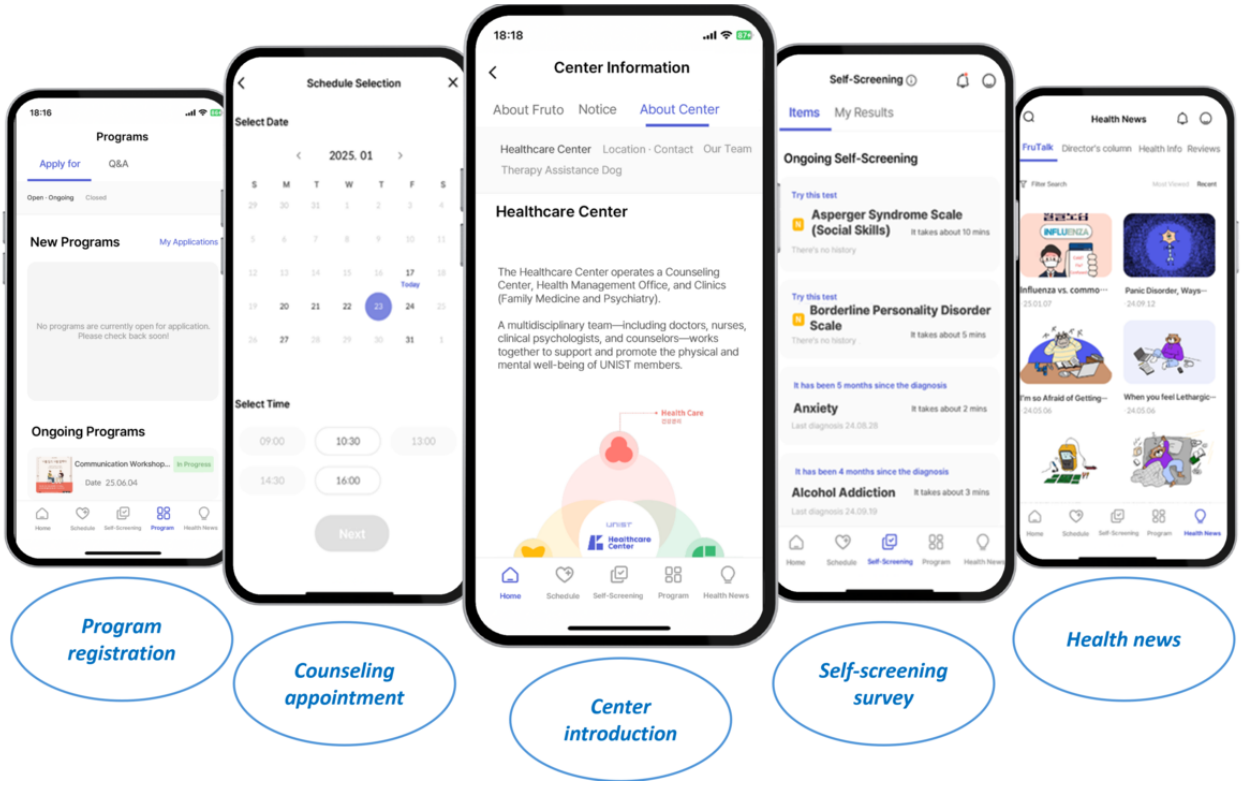

The five core components are described below:

1. Self-screening survey: self-screening is commonly used in mobile mental health interventions to help users reflect on their symptoms and has been recognized as potentially enhancing motivation for self-care or professional support [22]. Previous research has emphasized limitations when self-screening is offered in isolation, as it may leave users uncertain about subsequent steps [23] and does not necessarily lead to direct improvement in depressive symptoms [24]. When results are presented clearly and in a user-friendly manner, without requiring clinical interpretation, they can also assist professionals in making informed treatment decisions [24]. Fruto offers mobile-optimized versions of validated screening tools, including the Patient Health Questionnaire-9 for depression, Generalized Anxiety Disorder-7 for anxiety, Alcohol Use Disorders Identification Test–Korean version for alcohol use, Adult ADHD Self-Report Scale Version 1.1 for attention difficulties, Pittsburgh Sleep Quality Index–Korean version for sleep quality, and the P4 Suicide Risk Screener for suicide risk. Upon completion, users receive immediate, color-coded feedback accompanied by an explanatory summary. To address the limitations of stand-alone screening, Fruto provides follow-up guidance, such as counseling referrals and tailored content recommendations for users with elevated scores or risk indicators. With user consent, these results are also shared with an affiliated counselor, who may incorporate them into consultations or platform-based interventions.
2. Health news: university students often prefer text-based resources, particularly in settings such as dormitories and classrooms where privacy is limited, over audio or visual materials [25]. This preference may be especially relevant in contexts where stigma around mental illness persists [26]. Fruto addresses these specific barriers by offering professionally curated, text-based content to enhance mental health literacy and reduce stigma [8,25]. Developed in collaboration with licensed counselors and psychiatrists, the articles cover a wide range of topics, including psychiatric disorders, well-being, mindfulness, and available campus services. To create a more personalized experience, the content is also tailored to users' self-screening results.
3. Counseling appointment: while online interventions can help reduce psychological barriers to help-seeking [10,21], effective mental health care often requires integration with offline services. However, many students find scheduling appointments in person or over the phone burdensome, especially when experiencing depression-related apathy or social anxiety [27]. Previous studies have shown that digital tools, such as online booking systems, can help mitigate these challenges [28]. Fruto enables students to schedule on-campus counseling sessions directly through the mobile application. The system includes a brief intake form that gathers information about symptoms and concerns, encouraging self-reflection before the session and helping counselors prepare for more efficient, informed consultations.
4. Program registration: although campus mental health services can offer scalable, tailored support and help reduce stigma across the broader student population via wellness programs (eg, group therapy, mindfulness training, and personality assessments) [16,29], participation rates remain low due to poor publicity and limited access to information [30]. Integrating these programs into digital platforms has been proposed as a way to enhance accessibility and sustainability [25,31]. Fruto provides a dedicated interface

XSL•FO
RenderX

for wellness program registration, allowing counselors to share detailed information about each program, its eligibility criteria, and its schedules. Students can easily browse ongoing, upcoming, and past sessions and register directly through the platform. An anonymous comment feature also allows students to express preferences and suggest new topics, helping practitioners better align future programs with student needs.

5. Center introduction: reliable information about on-campus services is essential for building trust and encouraging student engagement. Fruto provides basic information and emergency contact details for university mental health centers [32,33]. Important updates, such as operational hours and event announcements, are regularly posted via a dedicated bulletin board to keep students informed and connected.

To make the platform's design logic more explicit, we mapped each Fruto feature to theory-informed mechanisms relevant to help-seeking, drawing on the theory of reasoned action/theory of planned behavior tradition. Rather than assuming that all features influence help-seeking in the same way, we conceptualized self-screening, psychoeducational content, counseling access, program registration, and center information as targeting different but complementary pathways, including attitudes, perceived norms, and perceived behavioral control. Table 1 maps Fruto's main features, functions, proposed mechanisms, and theoretical constructs.

Table 1. Mapping of Fruto features to theory-informed help-seeking mechanisms and constructs.

Fruto feature	Primary function	Theoretical construct	Proposed mechanism
Self-screening	Symptom recognition and self-monitoring	Attitude toward help-seeking	Makes psychological status more concrete and increases the perceived relevance of professional support
Health news	Mental health literacy and stigma reduction	Subjective norms + attitude	Normalizes help-seeking through expert-curated information and reduces stigma
Counseling appointment	Lowering access barriers	Perceived behavioral control	Simplifies action steps and increases confidence in accessing counseling
Program registration	Participation in preventive programs	Subjective norms	Encourages participation in campus well-being activities and normalizes support use
Center introduction	Service awareness and trust	Attitude + perceived behavioral control	Increases trust in campus services and clarifies how to use them

In a subsequent step, we contextualized Fruto within existing university-oriented digital mental health platforms by conducting a descriptive app comparison using an adaptation of the Lancet Digital Health app evaluation framework [34]. Fruto, TimelyCare [35], YOU at College [36], and IntelliCare for College Students [25] were reviewed based on publicly available information, and each criterion was rated on a 3-point scale through consensus discussion among the research team. The comparison was descriptive rather than evaluative and is provided in Appendix 2.

Technical Implementation and Data Governance

Fruto was deployed within an Ulsan National Institute of Science and Technology (UNIST)-managed cloud environment. Data transmitted between the client and server was protected using Transport Layer Security in accordance with the university cloud network configuration, and individual identifiable information was encrypted at rest using AES-256. User authentication initially relied on email-and-password login and was later expanded to support single sign-on for university-affiliated users, while retaining email-based login for general users. Access to personal data was restricted according to service roles; in the evaluated versions, counselors could access only information voluntarily submitted or shared by students through the platform for service provision, and administrative access to personal information was logged and controlled. Explicit consent procedures were incorporated into both account registration and counseling request workflows.

Study Design

We conducted a sequential mixed methods study to explore the initial acceptability of Fruto's design and features, followed by an evaluation of its potential contribution in a real-world university setting. In Phase 1, we qualitatively investigated how a group of 16 undergraduate and graduate students experienced the prototype. Insights from this phase guided revisions to the platform's design and functionality. In Phase 2, we conducted a large-scale survey to assess changes in help-seeking attitudes and counseling-related beliefs after the platform's implementation. Figure 2 presents a joint display linking Phase 1 themes, corresponding design refinements, and related domains assessed in Phase 2.

XSL•FO
RenderX

Figure 2. Joint display linking Phase 1 themes, corresponding design refinements, and related domains assessed in Phase 2.

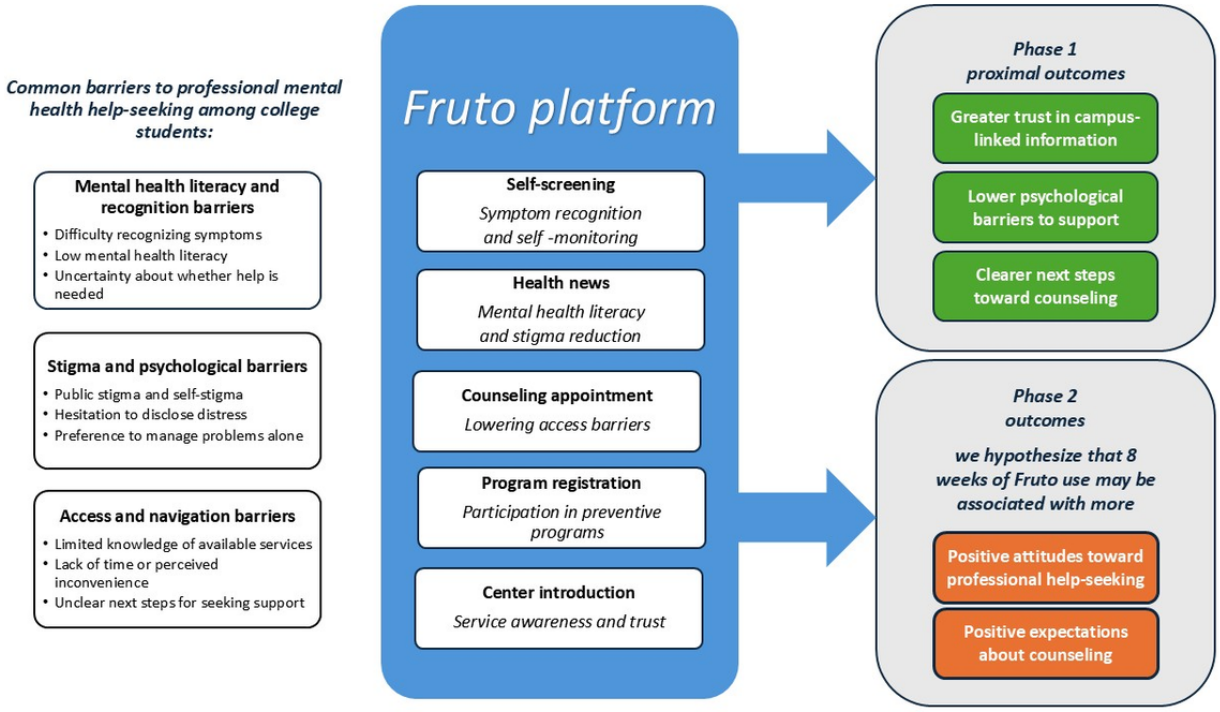

Participant Recruitment

Phase 1: Qualitative Study

Since Fruto was designed for the general university student population, all enrolled undergraduate and graduate students were eligible to participate. Recruitment was conducted through Every Time, a Korean online campus community, where a call for beta testers was posted. Participants provided informed consent and received 20,000 KRW (approximately US $13.6) as compensation for their participation.

Phase 2: Quantitative Study

After incorporating improvements based on Phase 1 findings, Fruto was distributed campus-wide, and recruitment for Phase 2 began. A new survey invitation was posted and promoted through informational posters across campus. Participants were instructed to complete pre- and postuse surveys and to engage freely with the app over an 8-week period. Upon completion, they received compensation of 10,000 KRW (approximately US $6.8).

Ethical Considerations

Phase 1 consisted of vignette-based prototype sessions and interviews focused on user experience. After completion of Phase 1, the protocol and study materials were submitted to the UNIST Institutional Review Board, which determined that the activity met the criteria for exemption from full review (approval number UNISTIRB-26-020-C). Participants provided informed consent for participation and audio recording, and interview data were stored securely under institutional privacy procedures. In addition to research consent procedures, the platform incorporated explicit service-level consent steps for account registration and counseling requests. Personal information was handled according to institutional privacy and security procedures, including encrypted transmission and storage, controlled access, and access logging.

Phase 2, which involved the real-world deployment and evaluation of the Fruto platform, was approved by the Ulsan National Institute of Science and Technology Institutional Review Board (approval number UNISTIRB-24-047-A). Participants received detailed online information about the study purpose, procedures, and potential risks and provided informed consent electronically via Google Forms before participation. Participants were compensated for their time. Phase 1 participants received KRW 20,000 (US $13.6), and Phase 2 participants received KRW 10,000 (US $6.8) upon completion of the study procedures.

Qualitative Data Collection and Analysis

Study Procedure

A vignette-based semistructured interview method was developed to present short scenarios featuring hypothetical individuals in specific situations, prompting participants to reflect on how they might respond in similar contexts [37]. We used this approach to reduce psychological burden and encourage honest feedback on sensitive or complex issues. Prior research suggests that vignette-based techniques allow individuals to express their attitudes, perceptions, and values from a third-person perspective [38].

We designed four vignettes reflecting distinct motivations for using mental health apps, based on prior research identifying key drivers among university students [39]. While many students are motivated by personal stress or needs, external influences such as peer encouragement or previous exposure to mental health issues may also influence app use [40-42]. The four vignette types were (1) supporting a peer in distress, (2) seeking

XSL•FO
RenderX

help for oneself, (3) casual curiosity, and (4) maintaining well-being despite no current problems.

To capture a wider range of intended use contexts while keeping the prototype session manageable, each participant was presented with 2 vignettes in sequence and asked to role-play the main character while exploring the prototype version of Fruto. After a brief app overview, participants freely interacted with its features for approximately 10 minutes, during which their screen activity was recorded. All sessions were conducted face-to-face in a small office or laboratory setting by one of two researchers (MK or Hyorim K). To simulate natural usage, the researcher temporarily left the room during this time.

Following the interaction, a one-on-one semistructured interview lasting approximately 20 minutes was conducted. The interview focused on participants' app usage, identifying helpful or problematic features, and gathering suggestions for improvements. The interview guide was developed in advance and used consistently across participants; no substantive modifications were made during data collection. The full interview guide is provided in Multimedia Appendix 3.

To ensure that the prototype was evaluated across different help-seeking contexts, the scenario design was structured around distinct usage motivations, as described below.

Scenario Design Based on Usage Motivation

Four user types were defined for vignette construction: peer supporter (PS), students who downloaded the app to support friends struggling with mental health issues; help seeker (HS), students whose mental health problems have significantly disrupted their daily lives and prompted them to seek support through the app; curious explorer (CE), students who discovered the app through campus promotions or incentives and tried it out of curiosity; and proactive maintainer (PM), students without current concerns who wished to proactively manage their mental well-being based on personal or vicarious experiences. The original vignette scripts provided to participants are included in Multimedia Appendix 4.

Data Management and Thematic Analysis

All interviews were audio-recorded with participants' informed consent. The 2 interviewers (MK and Hyorim K) manually transcribed the recordings of the interviews they had conducted. Transcripts were securely stored in encrypted cloud storage, accessible only to the research team.

We conducted an inductive thematic analysis to identify meaningful patterns across participant responses [43,44]. This approach was appropriate for exploring how students experienced the prototype across different vignette contexts and for generating design-related insights. MK and Hyorim K repeatedly read all transcripts to familiarize themselves with the dataset and recorded initial analytic notes. Transcripts were then organized in Microsoft Excel into meaning units, with each unit tagged according to vignette type and app feature. The 2 researchers collaboratively assigned inductive codes to these meaning units and iteratively refined the coding framework by merging overlapping codes and clarifying ambiguous codes through repeated analytic discussions.

Related codes were then grouped into candidate themes. Theme development was guided not only by recurrence across interviews but also by the relevance of the patterned responses to the study questions concerning user experience and design implications. Candidate themes were reviewed and refined through repeated discussions between MK and Hyorim K, with additional feedback from coauthors with expertise in digital health, psychiatry, clinical psychology, and design. The first author subsequently refined the final thematic structure, finalized theme labels and descriptions, and selected representative quotations [44,45].

Sample Adequacy, Reflexivity, and Trustworthiness

A total of 16 students participated in Phase 1. Each participant completed 2 vignette scenarios; therefore, the final qualitative dataset comprised 32 vignette-based use episodes, with 8 cases for each vignette type. Rather than applying formal saturation as a fixed stopping rule, we considered sample adequacy in relation to the focused aims of this design-oriented qualitative phase and the coverage achieved across the four intended use contexts. Data collection concluded when all vignette conditions had been adequately represented, and subsequent interviews yielded largely repetitive, design-relevant insights [45-47].

Throughout the analytic process, the researcher's reflexivity was considered. The first author had been involved in the Fruto project since its early planning stages, had prior experience in digital mental health research and qualitative analysis, and had collaborated with the campus counseling center for several years. This position supported strong contextual familiarity with the platform and the institutional setting, but also required reflexive awareness of prior assumptions about the platform and campus mental health services. In contrast, the second author had joined the project more recently, came from a different university background, and brought a relatively less embedded perspective to the data. These differing positions were used during analytic discussions to question assumptions and refine interpretations [45,48].

To enhance trustworthiness, we used repeated engagement with the full dataset, illustrative quotations, multidisciplinary review of developing themes, and documented coding records. Credibility was supported through repeated discussion of interpretations and direct linkage between themes and participant accounts; dependability through use of a consistent interview guide and documented analytic procedures; confirmability through critical review by coauthors from different disciplinary backgrounds; and transferability through contextual description of the institutional setting and platform characteristics [44,48,49].

Design Iteration Process

After completion of the Phase 1 interviews, the research team synthesized the main user experience issues and design implications and organized them into actionable refinement priorities. These priorities were reviewed together with the UI designer and campus counseling center staff to ensure that the proposed changes were both user-centered and feasible within the counseling service workflow. Based on this discussion, the team translated the findings into concrete interface, content,

XSL•FO
RenderX

and service-flow revisions before Phase 2. The UI designer incorporated these agreed-upon revisions into updated interface plans, while counseling center staff provided input on how the revised features should support service delivery and student guidance. The updated plans were then communicated to the development team for implementation, followed by several iterative reviews before campus-wide launch [25,32,50].

Quantitative Data Collection and Analysis

Study Procedure

The second phase used a single-group pre-post design. Following Fruto's official launch in September 2024, students who completed the preuse survey were invited to freely use the app for 8 weeks. At the end of this period, follow-up surveys were emailed to participants. For the primary repeated-measures analyses, we used all available outcome data from the 109 participants who completed the baseline survey and the 70 participants who provided postuse data.

The quantitative phase was not designed to isolate the effects of individual Fruto components or to test feature-specific behavioral mechanisms. Its purpose was instead to examine whether the integrated multidomain platform, used as a whole under real-world campus conditions, was associated with favorable attitudinal and perceptual outcomes during this formative stage of evaluation.

Sample Size Estimation

A priori power analysis was conducted using G*Power (Universität Düsseldorf) to estimate the sample size required for a single-group pre-post design. Prior digital interventions targeting mental health literacy and help-seeking intentions have reported small-to-medium effects. A meta-analysis of digital interventions aimed at improving mental health literacy reported an effect size of approximately 0.45, with similar estimates for stigma reduction of approximately 0.33 [51]. Studies targeting help-seeking intentions have reported effect sizes around 0.48, although these effects were not always statistically significant [52]. Given the preliminary, real-world nature of this study and the use of voluntary campus users rather than a clinical population, we considered several planning scenarios. A conservative scenario using an expected effect size of d_z=0.30 indicated a required sample size of approximately 89 participants with α=.05 and power=.80. Small-to-medium scenarios required smaller samples, including approximately 75 participants for d_z=0.33, 67 participants for d_z=0.35, and 52 participants for d_z=0.40. To account for potential attrition over the 8-week period, we aimed to recruit more than 100 students for the preuse survey. Although the baseline recruitment target was met, the number of participants who completed the postuse survey was below the most conservative complete-case target but within the range required for detecting small-to-medium pre-post effects. Therefore, Phase 2 findings should be interpreted as preliminary.

Measures

In both pre- and postintervention surveys, participants reported their attitudes toward professional help-seeking and perceptions of counseling. App usability and perceived quality were assessed at postintervention.

Attitudes Toward Seeking Professional Psychological Help was used to measure students' trust in and perceived need for counseling services. It includes 2 subscales, positive and negative attitudes, and has been validated for use with Korean populations [53,54].

The Beliefs and Expectations About Counseling Scale was used to assess students' beliefs about counseling. Grounded in multiattribute attitude theory and the theory of reasoned action, this scale assesses students' beliefs about counseling. Validated among both US and Korean university students, it includes subdomains such as positive expectations, negative endogenous beliefs, and socially supportive beliefs [55,56].

The user version of the Mobile App Rating Scale (uMARS) was used after app use to assess app quality from a general user perspective. It includes subscales for engagement, functionality, aesthetics, information, and subjective quality [57]. As no official Korean version was available, the research team used a Korean translation of the uMARS for exploratory purposes in this study. The translated items were reviewed by bilingual members of the research team for conceptual clarity, although formal back-translation and cross-cultural validation were not performed. The Korean translation of the uMARS showed acceptable internal consistency in this sample (Cronbach α=0.77 for engagement, 0.79 for functionality, 0.80 for aesthetics, 0.79 for information, and 0.72 for subjective quality), but it was used for exploratory purposes only and should not be interpreted as a formally validated Korean version.

Data Collection, Pre-Post Analysis, and Sensitivity Analysis for Missing Follow-Up Data

Survey data were collected using Google Forms, with all participants providing informed consent for research use before completing the survey. Internal consistency was evaluated using Cronbach α for each construct. Composite reliability (CR) and average variance extracted (AVE) were also calculated from baseline responses to examine convergent validity. CR values of ≥0.70 and AVE values of ≥0.50 were interpreted as indicating acceptable convergent validity [58]. Potential attrition bias was examined by comparing participants who completed only the baseline survey with those who also provided postintervention data. Categorical baseline characteristics were compared using Fisher exact tests because of sparse cell counts, and continuous baseline summary scores were compared using independent 2-tailed *t* tests. Pre-post changes were examined using linear mixed effects models with a random intercept for participant and a fixed effect of time, baseline vs postintervention, fitted separately for each outcome [59]. This approach allowed us to use all available observations from the 109 baseline respondents and the 70 participants with postintervention data. For each model, we report the fixed effect of time as B, SE, 95% CI, and *P* value.

Because not all baseline participants completed the postintervention survey, we conducted a conservative sensitivity analysis using baseline observation carried forward (BOCF) [60]. For participants without observed follow-up data, the

XSL•FO
RenderX

baseline value for each outcome was carried forward to the postintervention time point. Pre-post comparisons were then repeated in the full 109-participant sample to examine whether the overall pattern of findings was sensitive to this stringent missing-data assumption.

Exploratory Analyses of Perceived App Quality

To explore whether perceived app quality was associated with more favorable postintervention attitudinal outcomes, we conducted baseline-adjusted linear regression analyses using postintervention uMARS data. To provide a focused exploratory model and reduce multiple-testing burden, we used the overall mean uMARS score, calculated as the mean of the 5 uMARS subscale scores: engagement, functionality, aesthetics, information, and subjective quality, rather than entering the individual subscales separately [61].

Exploratory regressions were conducted for the 2 positive-valence outcomes most directly aligned with the platform's intended attitudinal targets: positive Attitudes Toward Seeking Professional Psychological Help (ATSPPH_Pos) and positive expectations about counseling (Positive Expectations About Counseling Scale [BEACS_Pos]). For each model, the corresponding baseline score was entered as a covariate, and the total uMARS score was entered as the predictor of interest. We report unstandardized regression coefficients, SEs, 95% CIs, and standardized coefficients. Model diagnostics included assessment of multicollinearity and residual distributions [62].

Integration of Qualitative and Quantitative Strands

Integration between the qualitative and quantitative strands occurred through both study design and final interpretation [63]. First, integration occurred through building, as Phase 1 findings were used by the design, development, and counseling-content teams to refine the platform's content structure, navigation, and feature emphasis before Phase 2. Second, integration occurred at the interpretation and reporting stage, where qualitative themes were compared with quantitative patterns to examine areas of convergence and complementarity. To make this integration explicit, we provide a mapping table linking major Phase 1 themes, corresponding design revisions, and related Phase 2 evaluation domains.

Results

Comparison With Existing Digital Mental Health Platforms

We reviewed Fruto alongside TimelyCare [35], YOU at College [36], and IntelliCare for College Students [25] using a study-specific adaptation of the Lancet Digital Health app evaluation framework. Based on publicly available information, Fruto appeared to integrate multiple campus-linked functions within a single platform, particularly self-screening, psychoeducation, counseling access, program registration, and counseling center information. Its relative limitations were more apparent in interoperability-related functions, including data export or sharing capability. This comparison is presented to contextualize Fruto's design positioning rather than to make formal claims of superiority over other platforms. The full comparison is provided in Multimedia Appendix 2.

Phase 1: Qualitative Findings

Participant Demographics

A total of 16 students, including both undergraduate and graduate students, participated in Phase 1. Participants ranged in age from 18 to 31 years, with a mean age of 24.25 (SD 4.37) years. The majority were female (n=11, 68.8%), while 5 (31.2%) participants were male. Half of the participants were undergraduate students (n=8, 50%) and half were graduate students (n=8, 50%). Table 2 summarizes participants' demographic characteristics and assigned vignette scenarios.

XSL•FO
RenderX

Table 2. Participant demographics and assigned vignette scenarios.

ID	Sex	Age (years)	Academic status	Vignette
P1	Male	18	Undergraduate	Peer supporter, proactive maintainer
P2	Female	19	Undergraduate	Help seeker, curious explorer
P3	Female	30	Graduate	Help seeker, peer supporter
P4	Female	31	Graduate	Curious explorer, help seeker
P5	Male	22	Undergraduate	Help seeker, curious explorer
P6	Female	23	Graduate	Proactive maintainer, peer supporter
P7	Male	27	Graduate	Peer supporter, curious explorer
P8	Female	28	Graduate	Help seeker, proactive maintainer
P9	Male	22	Undergraduate	Curious explorer, peer supporter
P10	Female	19	Undergraduate	Proactive maintainer, help seeker
P11	Female	23	Undergraduate	Peer supporter, curious explorer
P12	Female	29	Graduate	Proactive maintainer, help seeker
P13	Female	28	Graduate	Curious explorer, proactive maintainer
P14	Female	28	Graduate	Peer supporter, proactive maintainer
P15	Female	20	Undergraduate	Proactive maintainer, help seeker
P16	Male	21	Undergraduate	Curious explorer, peer supporter

Findings

Platform-Based Online Service Experience

Overview

This section explores participants' experiences with the platform's online features, such as mental health news and self-screening tools, as well as their perceptions of the platform's usability and perceived impact.

Increased Trust Through Involvement of Campus Mental Health Professionals

Many participants reported that knowing the platform was developed in collaboration with the campus counseling center enhanced their trust in it. As 1 participant remarked, "If I hear this app was developed by the health center, I naturally trust it" (P2, HS). Participants particularly valued the impression that the mental health content was created or reviewed by professionals, such as school counselors and psychiatrists. One participant explained, "Since this app was developed by our campus health center and the articles are written by experts, I trust the content" (P3, HS).

Some users reported that this sense of trust allowed them to engage with the content with less skepticism. For example, 1 participant noted, "If the information comes from this app's health news section, I'd trust it and read it without needing to filter it" (P4, CE). Others emphasized that citation and professional terminology enhanced the credibility of the content. As 1 participant stated, "The presence of citations in the mental health articles made them feel trustworthy" (P12, PM).

However, not all participants accepted the content uncritically. For instance, P9 expressed a more cautious perspective, stating that they would not assume everything was "100% accurate," but still considered the content potentially helpful when supporting others. "I'm not sure if something like 'how to overcome depression' would really work, but if a friend needs help, I might look into it" (P9, PS).

Emotional Relief and Action Guidance Through Content Linked to Self-Screening Results

Participants responded positively to the way self-screening results were followed by relevant content and counseling recommendations. Several reported feeling emotionally reassured and better equipped to manage their concerns. For example, P12 noted, "After completing the self-assessment, related articles appeared. It gave me the sense that I could resolve my problems" (P12, PM). Similarly, P9 appreciated how the integrated features lowered the barrier to help-seeking: "After doing the anxiety test, I got an article about real counseling cases. It helped me understand how counseling works, and I think this kind of experience might change how hesitant friends perceive it" (P9, HS).

App as a Tool for Supporting Friends

Based on their experience with the app, several participants said they would recommend it to friends struggling with mental health issues. P7 stated, "A lot of my friends at school suffer from stress due to relationships or academics. I think this app would really help them" (P7, PS). P4 emphasized the app's credibility and accessibility due to its institutional backing: "Since it's run by the school, anyone can use it. If a friend needs help, I would say, 'Why not try the self-assessment?'" (P4, CE).

However, some participants expressed concern that recommending a "mental health app" could carry stigma. P13 explained, "I might say it could help, but I'd worry they'd think I'm implying there's something wrong with them" (P13, CE). P10 suggested that a more indirect approach might be more

XSL•FO
RenderX

effective: “There is still a lot of stigma around mental health. So rather than saying, ‘You need help, use this app,’ I would say, ‘The school was promoting this, so I tried it out.’” (P10, HS).

Experience Integrating Online and Offline Services

Overview

This section highlights participants’ experiences with Fruto’s functions for scheduling counseling appointments and registering for offline wellness programs, including meditation sessions and group counseling.

Lowering Barriers to Help-Seeking

Several participants noted that the app’s bright, user-friendly interface helped reduce the psychological burden of seeking counseling. P4 explained, “I do not think my issue is that serious, so visiting the counseling center feels like a big step. But the app auto-fills my name and information, which makes it easier to give it a try.” (P4, CE).

P3 contrasted the app-based process with past experiences: “In the past, writing my information on paper at the center felt burdensome. But typing out my concerns through the app helped me organize my thoughts and reduced the fear of booking a session” (P3, HS). Some participants also suggested including brief information about what to expect during counseling, such as confidentiality policies or whether counseling records affect academic records, to help alleviate anxiety before a first visit.

Selective Use Based on Personal Situations

Participants assigned to the CE or PM scenarios sometimes chose not to use certain features, explaining that they did not feel an immediate need for mental health support. P5 stated, “From the perspective of a CE user, I did not need to book a session, so I did not try that feature” (P5, CE), but added they would consider using it if future issues arose. Similarly, P2 noted, “Even if I am in the CE scenario now, I might become an HS type later. Just knowing these features are available is helpful” (P2, CE), suggesting that preexposure to resources may encourage future help-seeking.

Design Requirements

This section identifies usability issues, technical limitations, and user recommendations for improving the platform.

Challenges With Text-Heavy Content and Complex Structure

Some participants appreciated the conversational tone of the health news articles, finding them easier to read. However, others felt that the amount of text was overwhelming. P9 commented, “People are used to short content such as Shorts or Reels. Long texts can feel burdensome” (P9, CE). One participant noted that the use of clinical terms made some articles difficult to understand and suggested using simpler language. Additionally, some participants found the content categorization confusing. P5 remarked, “There is so much in the health news section that it feels unorganized. You might not even notice the user reviews. It should just focus on actual health information” (P5, HS).

Broadening Content to Include Self-Discovery and Everyday Topics

Participants, particularly those in the CE and PM scenarios, expressed a desire for more accessible and exploratory content. P8 suggested, “Instead of only deep mental health assessments, it’d be nice to have something like an MBTI test, something for self-exploration even without serious issues” (P8, PM). Others expressed interest in topics such as self-esteem, interpersonal relationships, and learning styles. As P8 noted, “A simple learning style test would be fun. Both undergraduate and graduate students would be interested.”

Participants also expressed interest in lighter, more approachable offline programs. P15 shared, “I do not have a mental disorder, so I’d prefer relaxing activities, maybe something like a mood lamp-making session” (P15, PM). HS participants emphasized that programs focused on self-understanding should be offered individually rather than in groups. As P3 noted, “A one-on-one program to build self-esteem would be ideal” (P3, HS).

Minor Functional Errors

Some participants encountered functional issues, such as pages failing to load or back buttons not working as intended. These technical issues were documented and addressed during the refinement process.

Improvements to the Fruto App and Phase 2 Experimental Design

The issues and suggestions identified during the interviews were promptly communicated to the development team and service providers. Most UI-related improvements were subsequently implemented, including the addition of images, paragraph breaks, and bolded key points to improve readability. Additionally, the content-classification system was simplified to improve navigation.

Participants expressed a desire for content that was more relatable and focused on everyday topics such as self-regulation, self-esteem, and attention. They also preferred user-friendly language and practical, real-life examples. In response, new sections, such as “Self-Care,” were added to better address these needs.

The improved version of Fruto was launched campus-wide in September 2024. Qualitative feedback from Phase 1 directly informed app development, streamlining the design process for developers and administrators while enhancing the overall user experience. Within the first 8 weeks after launch, approximately 400 accounts were registered on the platform, providing contextual information on early campus uptake. However, the quantitative analysis was restricted to the 109 students who completed the baseline survey, of whom 70 also provided postintervention data after 8 weeks. Primary repeated-measures analyses used all available observations from these baseline and follow-up assessments.

To clarify how Phase 1 informed subsequent platform refinement and Phase 2 evaluation, we mapped key qualitative themes onto concrete design modifications implemented before campus-wide launch and identified the corresponding domains examined in the quantitative phase. This approach is consistent

XSL•FO
RenderX

with user-centered digital health research, in which user feedback, iterative testing, and implementation-focused refinement are used to improve acceptability, usability, and contextual fit before broader deployment. Figure 3 maps the Phase 1 qualitative themes, corresponding design refinements, and related Phase 2 evaluation domains.

Figure 3. Mapping of Phase 1 qualitative themes, corresponding design refinements, and related Phase 2 evaluation domains. UI: user interface; uMARS: user version of the Mobile App Rating Scale; UX: user experience.

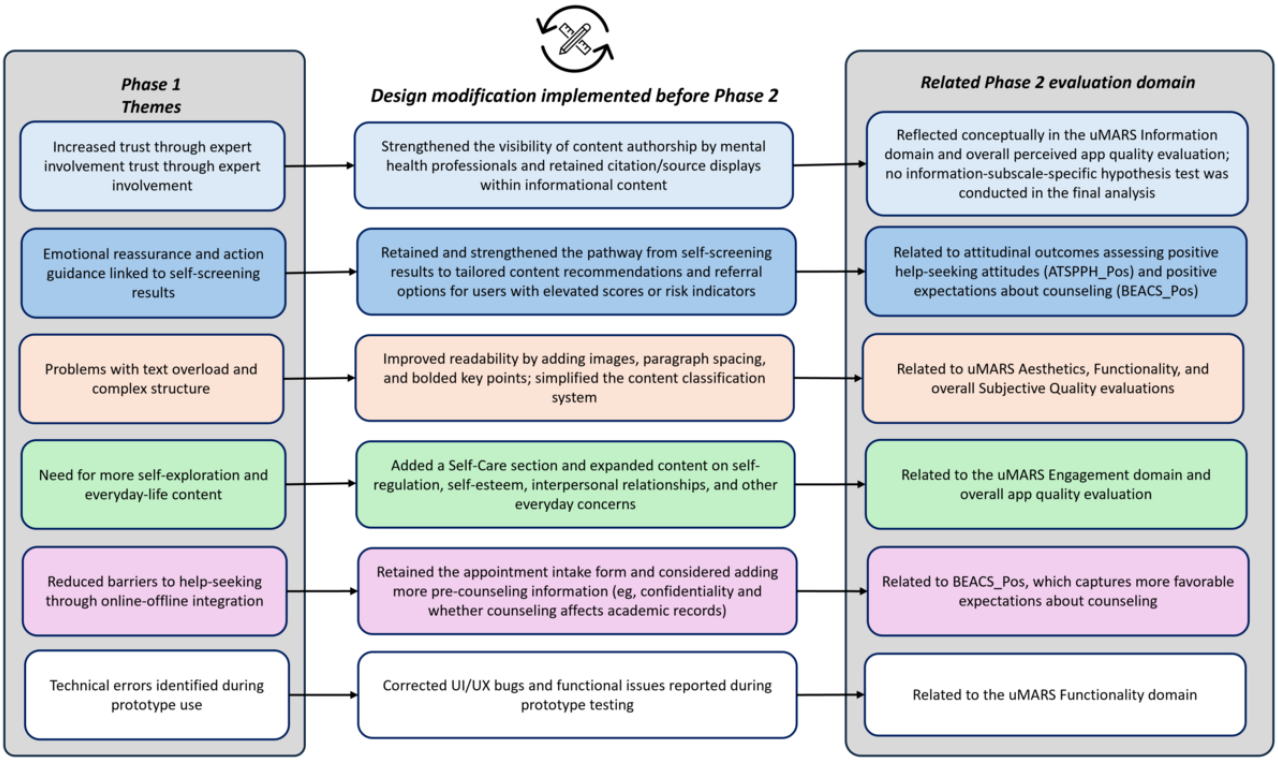

Phase 2: Quantitative Results

Background Characteristics

A total of 109 students completed the preintervention survey, and 70 provided postintervention data after 8 weeks. Because the primary repeated-measures analyses used all available observations, the quantitative analyses included baseline data from all 109 respondents and follow-up data from the 70 participants who completed the postintervention survey. The baseline sample had a mean age of 22.27 (SD 2.96) years, included 62 (56.9%) female participants, and comprised 71 (65.1%) undergraduate students. Among the 70 participants with follow-up data, the mean age was 22.34 (SD 3.13) years, 39 (55.7%) were female, and 45 (64.3%) were undergraduates. To assess potential attrition bias, we compared participants who completed only the preintervention survey (n=39) with those who also provided postintervention data (n=70). No statistically significant baseline differences were observed in sex (Fisher exact test, P=.84), academic status (Fisher exact test, P=>.99), total help-seeking attitudes (P=.50), or total counseling-related beliefs (P=.47), suggesting no clear evidence of systematic baseline differences on the measured variables. Baseline characteristics are summarized in Table 3.

Table 3. Baseline characteristics at week 0.

Variables	Baseline sample (n=109)	Participants with follow-up data (n=70)
Sex, n (%)		
Male	47 (43.1)	31 (44.3)
Female	62 (56.9)	39 (55.7)
Age (years)		
Mean (SD)	22.27 (2.96)	22.34 (3.13)
Median (range)	22 (18-31)	22 (18-30)
Education level, n (%)		
Undergraduate	71 (65.1)	45 (64.3)
Graduate	38 (34.9)	25 (35.7)

XSL•FO
RenderX

The psychometric properties of the instruments in our study were assessed by calculating Cronbach α, AVE, and CR using baseline data from the 109 participants who completed the preintervention survey. The subscales generally demonstrated acceptable reliability, supporting their use in these analyses.

BEACS_Pos showed high internal consistency (Cronbach α=0.88; CR=0.87) and adequate convergent validity (AVE=0.53). Negative Endogenous Beliefs About Counseling Scale (BEACS_Neg) and Socially Supportive Beliefs About Counseling Scale (BEACS_Social) also showed good internal consistency, with Cronbach α values of 0.81 and 0.80, respectively. BEACS_Social additionally demonstrated adequate convergent validity (AVE=0.56; CR=0.83), whereas BEACS_Neg showed adequate CR=0.82 but lower convergent validity (AVE=0.44). ATSPPH_Pos showed acceptable internal consistency (Cronbach α=0.76; CR=0.78), although its AVE was below the conventional threshold (AVE=0.42).

In contrast, negative Attitudes Toward Seeking Professional Psychological Help (ATSPPH_Neg) showed weak psychometric performance, with low internal consistency (Cronbach α=0.44), low convergent validity (AVE=0.17), and low CR (CR=0.46). Accordingly, findings involving ATSPPH_Neg should be interpreted cautiously, and the absence of a significant time effect for this subscale should be regarded as inconclusive rather than as evidence of no change. Detailed reliability and convergent validity metrics are presented in Table 4.

Table 4. Reliability and convergent validity of baseline outcome measures (n=109).

Construct	Cronbach α	AVE[a]	CR[b]
ATSPPH_Pos[c]	0.76	0.42	0.78
ATSPPH_Neg[d]	0.44	0.17	0.46
BEACS_Pos[e]	0.88	0.53	0.87
BEACS_Neg[f]	0.81	0.44	0.82
BEACS_Social[g]	0.80	0.56	0.83

[a]AVE: average variance extracted.

[b]CR: composite reliability.

[c]ATSPPH_Pos: positive Attitudes Toward Seeking Professional Psychological Help.

[d]ATSPPH_Neg: negative Attitudes Toward Seeking Professional Psychological Help.

[e]BEACS_Pos: Positive Expectations About Counseling Scale.

[f]BEACS_Neg: Negative Endogenous Beliefs About Counseling Scale.

[g]BEACS_Social: Socially Supportive Beliefs About Counseling Scale.

Quantitative Changes in Help-Seeking Variables

To examine pre-post changes in key psychological constructs, we fitted linear mixed effects models using all available observations from the 109 baseline respondents and the 70 participants with postintervention data. The fixed effect of time was significant for positive attitudes toward professional help-seeking (ATSPPH_Pos: B=0.884, SE 0.284, 95% CI 0.327-1.441; *P*=.002) and positive expectations about counseling (BEACS_Pos: B=1.585, SE 0.541, 95% CI 0.526-2.645; *P*=.003). Time effects were not significant for negative attitudes toward help-seeking (ATSPPH_Neg: B=–0.157, SE 0.255, 95% CI –0.656 to 0.343; *P*=.54), negative beliefs about counseling (BEACS_Neg: B=–0.751, SE 0.434, 95% CI –1.601 to 0.100; *P*=.08), or socially supportive beliefs about counseling (BEACS_Social: B=0.254, SE 0.338, 95% CI –0.409 to 0.916; *P*=.45). Given the weak psychometric performance of ATSPPH_Neg in this sample, the null finding for negative attitudes should be interpreted as inconclusive rather than as evidence that the intervention did not affect negative attitudes. Overall, the pre-post pattern remained consistent when all available baseline data were incorporated into the mixed effects models. Observed baseline and follow-up values and fixed effects from the linear mixed effects models are presented in Table 5.

As a highly conservative sensitivity analysis, we conducted BOCF analyses, in which baseline values were carried forward to the postintervention time point for the 39 participants without observed follow-up data. The overall pattern remained unchanged: ATSPPH_Pos remained significant (mean change 0.54, 95% CI 0.16-0.92; t_{108}=2.84; *P*=.005; Cohen d_z=0.27) and BEACS_Pos remained significant (mean change 1.04, 95% CI 0.32-1.76; t_{108}=2.86; *P*=.005; Cohen d_z=0.27), whereas ATSPPH_Neg (*P*=.43), BEACS_Neg (*P*=.08), and BEACS_Social (*P*=.25) remained nonsignificant. These results suggest that the positive findings were robust even under a stringent missing-data assumption.

XSL•FO
RenderX

Table 5. Observed outcome values at baseline and follow-up and fixed effects from linear mixed effects models. Baseline and postintervention means are observed descriptive statistics and are shown for transparency. Inferential tests of change are based on the linear mixed effects models using all available observations.

Measure	Baseline (n=109), mean (SD)	Postintervention (n=70), mean (SD)	Time effect, B	SE	95% CI	*P* value
ATSPPH_Pos[a]	10.62 (3.12)	11.57 (2.06)	0.884	0.284	0.327 to 1.441	.002
ATSPPH_Neg[b]	9.02 (2.46)	8.93 (2.09)	–0.157	0.255	–0.656 to 0.343	.54
BEACS_Pos[c]	25.95 (5.27)	27.50 (4.74)	1.585	0.541	0.526 to 2.645	.003
BEACS_Neg[d]	6.20 (5.06)	5.53 (3.68)	–0.751	0.434	–1.601 to 0.100	.08
BEACS_Social[e]	15.18 (3.32)	15.19 (3.68)	0.254	0.338	–0.409 to 0.916	.45

[a]ATSPPH_Pos: positive Attitudes Toward Seeking Professional Psychological Help.

[b]ATSPPH_Neg: negative Attitudes Toward Seeking Professional Psychological Help.

[c]BEACS_Pos: Positive Expectations About Counseling Scale.

[d]BEACS_Neg: Negative Endogenous Beliefs About Counseling Scale.

[e]BEACS_Social: Socially Supportive Beliefs About Counseling Scale.

Follow-Up Analysis: Exploratory Regression Analyses of Postintervention Outcomes

Exploratory app-quality analyses focused on ATSPPH_Pos and BEACS_Pos as the 2 positive-valence outcomes most directly aligned with the platform's intended attitudinal targets and the primary pattern of improvement observed in the mixed effect models. Because the 5 uMARS subscales were substantially intercorrelated, we conducted a focused exploratory analysis using the overall uMARS score rather than individual subscales. Among participants with postintervention uMARS data, higher overall perceived app quality was associated with higher postintervention BEACS_Pos after adjustment for baseline BEACS_Pos (B=2.444, SE 1.066; β=.237, 95% CI 0.317-4.571; *P*=.03). For ATSPPH_Pos, the association with overall uMARS score was positive but did not reach statistical significance after adjustment for baseline ATSPPH_Pos (B=0.761, SE 0.434; β=.170, 95% CI –0.106 to 1.627; *P*=.08). These analyses were exploratory and should be interpreted cautiously.

Discussion

Potential of a Multidomain Platform to Lower Barriers to Help-Seeking

The main quantitative findings from Phase 2 aligned with patterns observed in Phase 1, particularly the qualitative findings presented in the "Lowering Barriers to Help-Seeking" section, where participants described reduced hesitation and lower psychological burden around seeking support. Specifically, students showed small but statistically significant improvements in positive attitudes toward professional help-seeking and positive expectations about counseling after 8 weeks of Fruto use.

In exploratory baseline-adjusted analyses, higher overall perceived app quality was associated with more favorable postintervention positive counseling expectations, while the corresponding association with positive help-seeking attitudes did not reach statistical significance. Given the exploratory nature of these analyses and the absence of objective behavioral indicators, these findings should be interpreted as preliminary self-reported changes in attitudes and counseling-related perceptions rather than evidence of behavioral change. Nevertheless, they are meaningful in light of current trends showing that students often turn to the internet for mental health information before seeking professional help [64]. Self-directed online exploration can increase accessibility, but it may also expose students to unverified, inconsistent, or low-quality content [65]. In contrast, Fruto's professionally curated and locally tailored content may have helped lower barriers by providing credible guidance within a trusted campus service environment.

Structural Design Principles for Multidomain Help-Seeking Platforms

The qualitative findings further suggest that the value of a multidomain platform lies not only in the number of help-seeking functions it offers, but also in how seamlessly these features are interconnected. Participants highlighted that this smooth integration, where one function naturally leads to another, enhanced their sense that support options were connected across different stages of the help-seeking process. For example, after completing a self-assessment, students were directed to tailored articles or given the option to schedule counseling appointments based on their screening results or indicated needs.

Our descriptive comparison with existing university-oriented platforms, such as TimelyCare, YOU at College, and IntelliCare for College Students, also supports this interpretation. Fruto's potential contribution appears to lie less in any single novel component than in the integration of self-screening, psychoeducational content, counseling access, program registration, and service information within one campus-linked platform, whereas these functions are often distributed across separate university services or tools.

A recognized challenge for multidomain platforms is that integrating several functions into one system can increase cognitive burden if navigation, information structure, and transitions between features are not carefully designed [66]. In

XSL•FO
RenderX

the early version of Fruto, participants identified areas where dense content organization and nested categories could make browsing feel overwhelming. Rather than indicating a limitation of the multidomain approach itself, these findings highlighted the importance of user-informed refinement to ensure that the platform remained easy to navigate. Through the Phase 1 prototype sessions, these issues were identified before campus-wide deployment and translated into design improvements, including simplified content classification, clearer feature pathways, and more context-sensitive links between self-screening, information, and counseling access. This process helped preserve the main advantage of a multidomain platform: offering multiple forms of support while maintaining a coherent and low-burden user experience [67].

Contextual Design Principles for Everyday and Low-Stigma Use of Help-Seeking Platforms

Previous studies suggested that university students may be more willing to engage with mental health services when they are framed using positive and familiar terms, such as well-being or self-care, rather than explicitly clinical language [68]. Similar work also suggests that this framing can make digital tools feel more approachable and less stigmatizing [25]. Our qualitative findings extend this evidence in the context of a campus-integrated, multidomain platform. Although Fruto was initially designed to deliver mental health content and programs addressing common student concerns such as depression, anxiety, and sleep disturbances, participants also wanted tools for self-exploration and lighter, everyday themes.

A multidomain platform should therefore function not only as a repository of expert knowledge but also as an accessible and relatable space for everyday use. Students preferred light, engaging content that combined professional guidance with tools for self-reflection and daily life. Content offered at different levels of depth and tone may make the platform feel more approachable than purely clinical resources. Because some participants worried that recommending a “mental health app” could carry stigma, Fruto’s broader well-being and everyday self-care content, together with its connection to university services, may help position the platform as a student well-being resource rather than only a clinical tool.

Strengths and Limitations

This study contributes to the literature on university digital mental health and the relatively underexplored area of campus-integrated, multidomain help-seeking platforms. Fruto combines self-screening, psychoeducation, counseling access, program registration, and service navigation within one system. Unlike stand-alone digital mental health tools, it was evaluated as part of a real university counseling ecosystem, allowing us to examine both user experience and preliminary changes in help-seeking attitudes and counseling-related beliefs.

The platform design was also linked to theory-informed help-seeking mechanisms, allowing each feature to be understood as targeting different but complementary pathways toward support use. The sequential mixed methods design allowed Phase 1 findings to inform concrete platform refinements before Phase 2 evaluation, making the study more closely connected to real implementation needs. In addition, the study provides actionable design guidance for universities seeking to connect digital tools with existing counseling services, including expert visibility, low-burden navigation, everyday self-care content, and clearer pathways from screening to support.

The main quantitative findings were also supported by a conservative BOCF sensitivity analysis, in which the positive outcomes remained significant even when baseline scores were carried forward for participants without follow-up data.

The vignette-based prototype sessions were also a methodological strength. By asking participants to engage with Fruto through scenarios involving seeking help for oneself, supporting a friend, exploring out of curiosity, and maintaining well-being, the interviews elicited a wider range of responses than direct questioning alone may have produced. Participants were not limited to describing only their current personal needs, but could imagine how the platform might be used across different help-seeking situations. This helped identify concrete design requirements for future multidomain platforms, including seamless integration between features, low-burden navigation, visible expert involvement, and content that supports self-understanding and preventive mental health care.

Several limitations should be noted. The quantitative component used a single-group pre-post design without a control group, which limits causal interpretation. Changes in help-seeking attitudes and counseling expectations may have been influenced by unmeasured factors, including natural adaptation to university life during the study period. All outcomes were self-reported, and no objective behavioral indicators, such as app usage intensity, feature-level engagement, counseling appointment completion, or service use, were analyzed. Accordingly, we could not determine how intensively participants used the platform or which features were most commonly used. Therefore, the results should be interpreted as preliminary changes in attitudes and counseling-related perceptions rather than evidence of behavioral change.

Sample size and attrition also require caution. Although the baseline recruitment target was met, the postintervention sample remained below the most conservative complete-case target, which may have reduced power to detect smaller effects. Because participation was voluntary, the sample may also overrepresent students already open to digital mental health tools or university-based support. In addition, ATSPPH_Neg showed weak psychometric performance, and the Korean translation of the uMARS was used for exploratory purposes only. These issues limit the interpretation of negative attitude outcomes and app-quality analyses.

Finally, Fruto was developed and evaluated in a single university context, with specific counseling-center infrastructure, institutional trust, and service workflows. Some features, such as campus-specific content and therapist-representing character mascots, may not generalize directly to other institutions or cultural contexts. Cross-campus deployment and scalability were not formally evaluated. Future studies should include control groups, objective usage and service-use data, validated app-quality measures, and more diverse institutional settings.

XSL•FO
RenderX

Conclusion

Fruto offers a practical example of how universities can organize digital support around students' different help-seeking needs rather than relying on isolated tools. The platform brings together self-screening, credible mental health information, self-care content, counseling access, program registration, and service navigation within one campus-integrated system. In doing so, it may help create a clearer and less burdensome pathway from early awareness to professional support.

The vignette-based interviews showed that students valued professionally developed content, personalized feedback after self-screening, and direct links to counseling and offline programs. Their feedback led to concrete refinements before campus-wide deployment, including a simpler information structure and expanded content on self-discovery and everyday concerns. After 8 weeks of use, students showed significant improvements in positive help-seeking attitudes and positive expectations about counseling, while exploratory analyses suggested that higher perceived app quality may be related to more favorable counseling expectations.

This study may encourage universities to develop connected, low-stigma digital pathways that support students before difficulties become severe and before formal counseling is needed. Future controlled studies should examine longer-term use, objective engagement, service-use outcomes, and whether improvements in attitudes and counseling expectations translate into actual help-seeking behavior.

Acknowledgments

We are grateful to the UNIST Healthcare Center for their active collaboration in this research. The center is a comprehensive health care facility that provides not only psychological counseling and psychiatric care, but also a variety of health promotion programs.

During manuscript preparation, ChatGPT was used only to suggest alternative English phrasing during translation from the authors' original Korean draft; all suggestions were critically reviewed and revised by the authors, and no generative artificial intelligence was used for literature review, data analysis, or the generation of scientific content, tables, or figures. We would like to thank Editage [69] for editing and reviewing this manuscript for the English language.

Funding

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare, Republic of Korea (grant HI22C0646). Additional support was provided by the National Institute of Health research project (2024ER080301) and the AI Science Hub program of the Ministry of Science and ICT (N10260136).

Data Availability

Data supporting the findings of this study are available upon request from the corresponding authors. The data are not publicly available due to privacy concerns.

Authors' Contributions

The author DJ is the corresponding author for this article and is responsible for data and materials, manuscript submission, peer review, publication process, authorship details, and ethics committee approval.
MK and DJ conceived and designed the study. MK, Hyorim K, JH, and OFL performed the analyses. MK and Hyorim K wrote the first draft of the manuscript. MK and Hyorim K collected the data. MK, Hyorim K, JH, OFL, Seonmi L, Sangil L, Hwang K, and DJ edited all manuscript versions. All authors were involved in interpreting the results and have read, commented on, and approved the final version of the manuscript.

Conflicts of Interest

The authors developed the Fruto platform as part of an academic research project; however, they report no commercial or financial interests related to the platform.

Multimedia Appendix 1
Full UI screenshots and descriptions of all features.
[DOCX File , 1689 KB - jmir_v28i1e78930_app1.docx]

Multimedia Appendix 2
Descriptive comparison of Fruto with existing university-oriented digital mental health platforms based on an adapted Lancet Digital Health framework.
[DOCX File , 31 KB - jmir_v28i1e78930_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Full interview guide.
[DOCX File , 18 KB - jmir_v28i1e78930_app3.docx]

Multimedia Appendix 4
The original vignette scripts.
[DOCX File , 22 KB - jmir_v28i1e78930_app4.docx]

References

1. Arnett JJ. Emerging adulthood: a theory of development from the late teens through the twenties. Am Psychol 2000;55(5):469-480. [doi: 10.1037/0003-066x.55.5.469]
2. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005;62(6):593-602. [doi: 10.1001/archpsyc.62.6.593] [Medline: 15939837]
3. American College Health Association. American College Health Association National College Health Assessment spring 2006 reference group data report (abridged). J Am Coll Health 2007;55(4):195-206. [doi: 10.3200/JACH.55.4.195-206] [Medline: 17319325]
4. Roh S, Lee SU, Soh M, Ryu V, Kim H, Jang JW, et al. Int J Ment Health Syst 2016;10:45 [FREE Full text] [doi: 10.1186/s13033-016-0077-3] [Medline: 27257434]
5. Song G, Kim E, Shin J. Effects of appearance satisfaction, self-esteem, and interpersonal relationship satisfaction in university students on employment stress. Asian-Pac J Converg Res Interchange 2021;7(2):23-32. [doi: 10.47116/apjcri.2021.02.03]
6. Lim AY, Lee SH, Jeon Y, Yoo R, Jung HY. Job-seeking stress, mental health problems, and the role of perceived social support in university graduates in Korea. J Korean Med Sci 2018;33(19):e149. [doi: 10.3346/jkms.2018.33.e149] [Medline: 29736162]
7. Kwak Y, Kim Y, Chae H. Job search anxiety and flourishing among university students: the serial mediating effects of social support and strengths use. BMC Psychol 2025;13(1):652 [FREE Full text] [doi: 10.1186/s40359-025-02995-4] [Medline: 40597452]
8. Kim EJ, Yu JH, Kim EY. Pathways linking mental health literacy to professional help-seeking intentions in Korean college students. J Psychiatr Ment Health Nurs 2020;27(4):393-405. [doi: 10.1111/jpm.12593] [Medline: 31954091]
9. Lee S, Lim J, Lee S, Heo Y, Jung D. Group-tailored feedback on online mental health screening for university students: using cluster analysis. BMC Prim Care 2022;23(1):19. [doi: 10.1186/s12875-021-01622-6] [Medline: 35172741]
10. Gibson K, Cartwright C. Young people's experiences of mobile phone text counselling: balancing connection and control. Child Youth Serv Rev 2014;43:96-104. [doi: 10.1016/j.childyouth.2014.05.010]
11. Baldofski S, Kohls E, Efe Z, Eckert M, Saee S, Thomas J, et al. The impact of a messenger-based psychosocial chat counseling service on further help-seeking among children and young adults: longitudinal study. JMIR Ment Health 2023;10:e43780 [FREE Full text] [doi: 10.2196/43780] [Medline: 37195747]
12. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019;21(7):e12869 [FREE Full text] [doi: 10.2196/12869] [Medline: 31333198]
13. Lattie EG, Cohen KA, Hersch E, Williams KD, Kruzan KP, MacIver C, et al. Uptake and effectiveness of a self-guided mobile app platform for college student mental health. Internet Interv 2022;27:100493 [FREE Full text] [doi: 10.1016/j.invent.2021.100493] [Medline: 35036331]
14. Xiao H, Carney DM, Youn SJ, Janis RA, Castonguay LG, Hayes JA, et al. Are we in crisis? National mental health and treatment trends in college counseling centers. Psychol Serv 2017;14(4):407-415. [doi: 10.1037/ser0000130] [Medline: 29120199]
15. Broglia E, Ryan G, Williams C, Fudge M, Knowles L, Turner A, et al. Profiling student mental health and counselling effectiveness: lessons from four UK services using complete data and different outcome measures. Br J Guid Couns 2021;51(2):204-222. [doi: 10.1080/03069885.2020.1860191]
16. Gorman KS. Annual survey: 2020. Association for University and College Counselling Center Directors. 2020. URL: https://www.aucccd.org/assets/documents/Survey/2019-2020%20Annual%20Report%20FINAL%20March-2021.pdf [accessed 2026-06-04]
17. Min Y, Choi YS, Kim BN. Filling the mental service gap on campus: an effectiveness trial testing the utility of app-based mindfulness psychological intervention for college students. Curr Psychol 2023;43(17):15434-15444. [doi: 10.1007/s12144-023-05402-6]
18. Ma KKY, Burn AM, Anderson JK. Review: school-based mental health literacy interventions to promote help-seeking - a systematic review. Child Adolesc Ment Health 2023;28(3):408-424. [doi: 10.1111/camh.12609] [Medline: 36377083]

XSL•FO
RenderX

19. Taylor-Rodgers E, Batterham PJ. Evaluation of an online psychoeducation intervention to promote mental health help seeking attitudes and intentions among young adults: randomised controlled trial. J Affect Disord 2014;168:65-71. [doi: 10.1016/j.jad.2014.06.047] [Medline: 25038293]
20. Howard KA, Griffiths KM, McKetin R, Ma J. Can a brief biologically-based psychoeducational intervention reduce stigma and increase help-seeking intentions for depression in young people? A randomised controlled trial. J Child Adolesc Ment Health 2018;30(1):27-39. [doi: 10.2989/17280583.2018.1467323] [Medline: 29764293]
21. Chan GH. A comparative analysis of online, offline, and integrated counseling among hidden youth in Hong Kong. Child Youth Serv Rev 2020;114:105042. [doi: 10.1016/j.childyouth.2020.105042]
22. BinDhim NF, Alanazi EM, Aljadhey H, Basyouni MH, Kowalski SR, Pont LG, et al. Does a mobile phone depression-screening app motivate mobile phone users with high depressive symptoms to seek a health care professional's help? J Med Internet Res 2016;18(6):e156 [FREE Full text] [doi: 10.2196/jmir.5726] [Medline: 27349441]
23. Kruzan KP, Meyerhoff J, Nguyen T, Mohr DC, Reddy M, Kornfield R. "I Wanted to See How Bad it Was": online self-screening as a critical transition point among young adults with common mental health conditions. Proc SIGCHI Conf Hum Factor Comput Syst 2022;2022:328. [doi: 10.1145/3491102.3501976] [Medline: 35531062]
24. Gilbody S, Sheldon T, House A. Screening and case-finding instruments for depression: a meta-analysis. CMAJ 2008;178(8):997-1003. [doi: 10.1503/cmaj.070281] [Medline: 18390942]
25. Lattie E, Kornfield R, Ringland KE, Zhang R, Winquist N, Reddy M. Designing mental health technologies that support the social ecosystem of college students. Proc SIGCHI Conf Hum Factor Comput Syst 2020;2020:10.1145/3313831.3376362. [doi: 10.1145/3313831.3376362] [Medline: 32656549]
26. Kim SB, Lee YJ. Factors associated with mental health help-seeking among Asian Americans: a systematic review. J Racial Ethn Health Disparities 2022;9(4):1276-1297. [doi: 10.1007/s40615-021-01068-7] [Medline: 34076864]
27. Cohen KA, Graham AK, Lattie EG. Aligning students and counseling centers on student mental health needs and treatment resources. J Am Coll Health 2022;70(3):724-732 [FREE Full text] [doi: 10.1080/07448481.2020.1762611] [Medline: 32432973]
28. Zhao P, Yoo I, Lavoie J, Lavoie BJ, Simoes E. Web-based medical appointment systems: a systematic review. J Med Internet Res 2017;19(4):e134. [doi: 10.2196/jmir.6747] [Medline: 28446422]
29. Saheb R, Mortimer T, Rutherford E, Sperandei S, Reis A. Creating healthy universities: the role of campus-based health promotion events in supporting student well-being. Health Promot J Austr 2021;32(1):13-20. [doi: 10.1002/hpja.305] [Medline: 31724768]
30. Priestley M, Broglia E, Hughes G, Spanner L. Student perspectives on improving mental health support services at university. Couns Psychother Res 2021;22(1). [doi: 10.1002/capr.12391]
31. Eisenberg D, Golberstein E, Gollust SE. Help-seeking and access to mental health care in a university student population. Med Care 2007;45(7):594-601. [doi: 10.1097/MLR.0b013e31803bb4c1] [Medline: 17571007]
32. Alqahtani F, Orji R. Insights from user reviews to improve mental health apps. Health Informatics J 2020;26(3):2042-2066. [doi: 10.1177/1460458219896492] [Medline: 31920160]
33. Aguilar Silvan Y, Hamza S, Fardeheb S, Bird C, Ng LC. Marketing mental health services: a mixed-methods analysis of racially and ethnically diverse college students' engagement with and perspectives on U.S. university mental health clinics' websites. BMC Health Serv Res 2024;24(1):1163. [doi: 10.1186/s12913-024-11652-2] [Medline: 39354461]
34. Henson P, David G, Albright K, Torous J. Deriving a practical framework for the evaluation of health apps. Lancet Digit Health 2019;1(2):e52-e54. [doi: 10.1016/S2589-7500(19)30013-5] [Medline: 33323229]
35. Medical services. TimelyCare. URL: https://timelycare.com/medical/ [accessed 2026-04-29]
36. YOU student impact. YOU at College. URL: https://youatcollege.com/you-student-impact/ [accessed 2026-04-29]
37. Finch J. The vignette technique in survey research. Sociology 1987;21(1):105-114. [doi: 10.1177/0038038587021001008]
38. Port ZMJ, Forsyth DK, Tappin DC. Utilising vignettes as prompts in semi-structured interviewing for employment relations research. NZJER 2024;48(2):1-12. [doi: 10.24135/nzjer.v48ix.160]
39. Borghouts J, Eikey EV, Mark G, De Leon C, Schueller SM, Schneider M, et al. Understanding mental health app use among community college students: web-based survey study. J Med Internet Res 2021;23(9):e27745 [FREE Full text] [doi: 10.2196/27745] [Medline: 34519668]
40. Aziz M, Erbad A, Belhaouari SB, Almourad MB, Altuwairiqi M, Ali R. Who uses mHealth apps? Identifying user archetypes of mHealth apps. Digit Health 2023;9:20552076231152175. [doi: 10.1177/20552076231152175] [Medline: 36714545]
41. Fürtjes S, Gebel E, Kische H, Beesdo-Baum K. Characteristic of mental health app usage: a cross-sectional survey in the general population. BMC Public Health 2024;24(1):3133. [doi: 10.1186/s12889-024-20500-1] [Medline: 39533246]
42. Vera Cruz G, Aboujaoude E, Khan R, Rochat L, Ben Brahim F, Courtois R, et al. Smartphone apps for mental health and wellbeing: a usage survey and machine learning analysis of psychological and behavioral predictors. Digit Health 2023;9:20552076231152164. [doi: 10.1177/20552076231152164] [Medline: 36714544]
43. Clarke V, Braun V. Thematic analysis. J Posit Psychol 2016;12(3):297-298. [doi: 10.1080/17439760.2016.1262613]
44. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Methods 2017;16(1):1609406917733847. [doi: 10.1177/1609406917733847]

XSL•FO
RenderX

45. Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and be(com)ing a researcher. Int J Transgend Health 2023;24(1):1-6. [doi: 10.1080/26895269.2022.2129597] [Medline: 36713144]
46. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
47. Hennink MM, Kaiser BN, Marconi VC. Code saturation versus meaning saturation: how many interviews are enough? Qual Health Res 2017;27(4):591-608 [FREE Full text] [doi: 10.1177/1049732316665344] [Medline: 27670770]
48. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
49. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
50. Scholten H, Granic I. Use of the principles of design thinking to address limitations of digital mental health interventions for youth: viewpoint. J Med Internet Res 2019;21(1):e11528. [doi: 10.2196/11528] [Medline: 31344671]
51. Chen Q, Zhao Z, Bao J, Lin J, Li W, Zang Y. Digital empowerment in mental health: a meta-analysis of internet-based interventions for enhancing mental health literacy. Int J Clin Health Psychol 2024;24(3):100489 [FREE Full text] [doi: 10.1016/j.ijchp.2024.100489] [Medline: 39220119]
52. Nazari A, Garmaroudi G, Foroushani AR, Hosseinnia M. The effect of web-based educational interventions on mental health literacy, stigma and help-seeking intentions/attitudes in young people: systematic review and meta-analysis. BMC Psychiatry 2023;23(1):647 [FREE Full text] [doi: 10.1186/s12888-023-05143-7] [Medline: 37667229]
53. Farina A, Fischer E. Attitudes toward seeking professional psychological help: a shortened form and considerations for research. J Coll Stud Dev 1995;36(4):368-373 [FREE Full text]
54. Nam Sook-Kyung. The construct validity of the Attitudes Toward Seeking Professional Psychological Help Scale-Short Form (ATSPPH-SF). Korean J Couns 2010;11(3):1007-1020 [FREE Full text] [doi: 10.15703/kjc.11.3.201009.1007]
55. Choi Sung-In, Kim Chang-Dae. Validity of the Korean version of the Beliefs and Evaluations about Counseling Scale (K-BEACS). Asian J Educ 2010;11(1):169-195 [FREE Full text] [doi: 10.15753/aje.2010.11.1.008]
56. Choi SI. Measurement of Attitudes Toward Counseling: Scale Development. United States: Ball State University; 2008.
57. Stoyanov SR, Hides L, Kavanagh DJ, Wilson H. Development and validation of the user version of the Mobile Application Rating Scale (uMARS). JMIR Mhealth Uhealth 2016;4(2):e72 [FREE Full text] [doi: 10.2196/mhealth.5849] [Medline: 27287964]
58. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. J Mark Res 1981;18(1):39-50. [doi: 10.1177/002224378101800104]
59. Fitzmaurice GM, Laird NM, Ware JH. Applied Longitudinal Analysis. Hobokon, New Jersey: John Wiley & Sons; 2012.
60. Liu-Seifert H, Zhang S, D'Souza D, Skljarevski V. A closer look at the baseline-observation-carried-forward (BOCF). Patient Prefer Adherence 2010;4:11-16. [doi: 10.2147/ppa.s8135] [Medline: 20165594]
61. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med 2004;66(3):411-421. [doi: 10.1097/00006842-200405000-00021]
62. Vatcheva KP, Lee M, McCormick JB, Rahbar MH. Multicollinearity in regression analyses conducted in epidemiologic studies. Epidemiology (Sunnyvale) 2016;6(2):227 [FREE Full text] [doi: 10.4172/2161-1165.1000227] [Medline: 27274911]
63. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs-principles and practices. Health Serv Res 2013;48(6 Pt 2):2134-2156. [doi: 10.1111/1475-6773.12117] [Medline: 24279835]
64. Chereka AA, Butta FW, Demsash AW, Kassie SY, Shibabaw AA, Dubale AT, et al. Examining internet use for health information seeking and influencing factors among undergraduate health science students in Southwest Ethiopia. Heliyon 2025;11(1):e41545. [doi: 10.1016/j.heliyon.2024.e41545] [Medline: 39845004]
65. Jia X, Pang Y, Liu LS. Online health information seeking behavior: a systematic review. Healthcare (Basel) 2021;9(12):1740. [doi: 10.3390/healthcare9121740] [Medline: 34946466]
66. Galavi Z, Montazeri M, Khajouei R. Which criteria are important in usability evaluation of mHealth applications: an umbrella review. BMC Med Inform Decis Mak 2024;24(1):365 [FREE Full text] [doi: 10.1186/s12911-024-02738-2] [Medline: 39614283]
67. Shneiderman B. Designing the User Interface: Strategies for Effective Human-Computer Interaction. Chennai: Pearson Education India; 1987.
68. Rith-Najarian L, Sun W, Chen A, Chorpita B, Chavira D, Mougalian S, et al. What's in a name? Branding of online mental health programming for university students. J Consult Clin Psychol 2019;87(4):380-391. [doi: 10.1037/ccp0000383] [Medline: 30883165]
69. Editage. URL: https://www.editage.co.kr/ [accessed 2026-06-08]

Abbreviations

ATSPPH_Neg: negative Attitudes Toward Seeking Professional Psychological Help
ATSPPH_Pos: positive Attitudes Toward Seeking Professional Psychological Help
AVE: average variance extracted

XSL•FO
RenderX

BEACS_Neg: Negative Endogenous Beliefs About Counseling Scale
BEACS_Pos: Positive Expectations About Counseling Scale
BEACS_Social: Socially Supportive Beliefs About Counseling Scale
BOCF: baseline observation carried forward
CE: curious explorer
CR: composite reliability
HS: help seeker
PM: proactive maintainer
PS: peer supporter
RQ: research question
UI: user interface
uMARS: user version of the Mobile App Rating Scale
UNIST: Ulsan National Institute of Science and Technology

Edited by A Stone; submitted 12.Jun.2025; peer-reviewed by ME Heidari, R Rahmi; comments to author 06.Mar.2026; revised version received 31.May.2026; accepted 01.Jun.2026; published 22.Jun.2026.

Please cite as:
Kim M, Kim H, Heo JI, Lee S, Lahcine OF, Lee S, Kim H, Jung D
Acceptability and Preliminary Evaluation of a Campus-Integrated Digital Platform (Fruto) for University Students' Mental Health Help-Seeking: Sequential Mixed Methods Study
J Med Internet Res 2026;28:e78930
URL: https://www.jmir.org/2026/1/e78930
doi:10.2196/78930
PMID:

©Myungsung Kim, Hyorim Kim, Jeong-in Heo, Seonmi Lee, Orane Farrah Lahcine, Sangil Lee, Hwang Kim, Dooyoung Jung. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Media for Health Outcomes: Evaluation Study of a Massive Online Open Course

Hannah Melchinger[1], BS, MPH; Sarah Christie[2], MPH, PhD; Abigail Titus[3], MEd; Alyssa Cruz[4], MS, EdD; Chelsey Lepage[5], MA; Tony Foleno[6], MA; Colleen Thompson-Kuhn[6], BA; Adam Terefe Abebe[7], BSc, MBA; G. Nekerwon Gweh[7], MA; Donewell Bangure[7], MPH; Stephen Maina[8], MS; Judy Mwangi[8], MSc; Peninah Thande[8], BCom; Maike Winters[4], MPH, PhD; Amyn A Malik[1], MBBS, MPH, PhD; Aleksandra Kuzmanovic[9], MA; Surangani Abeyesekera[10]; Sara Epperson[3], BS; Whittney Tom[11], MA; Kylie Holmes[12], MS; Angus Thomson[5], PhD; Saad B Omer[1], MBBS, MPH, PhD

[1]Peter O'Donnell Jr. School of Public Health, University of Texas Southwestern Medical Center, Dallas, TX, United States
[2]Yale School of Medicine, New Haven, CT, United States
[3]Poorvu Center for Teaching and Learning, Yale University, New Haven, CT, United States
[4]Yale Institute for Global Health, Yale University, New Haven, CT, United States
[5]Irimi, Thiers, France
[6]Ad Council, New York, NY, United States
[7]Africa Centres for Disease Control and Prevention (Africa CDC), Addis Ababa, Addis Ababa, Ethiopia
[8]Population Services International, Nairobi, Kenya
[9]World Health Organization, Geneva, Switzerland
[10]United Nations Children's Fund (UNICEF), New York, NY, United States
[11]The Permanente Medical Group, Oakland, CA, United States
[12](Formerly: 2018-2023), Meta Platforms, Inc., Menlo Park, CA, United States

Corresponding Author:
Hannah Melchinger, BS, MPH
Peter O'Donnell Jr. School of Public Health
University of Texas Southwestern Medical Center
5323 Harry Hines Blvd.
Dallas, TX, 75390
United States
Phone: 1 2034079398
Email: hannah.melchinger@utsouthwestern.edu

Abstract

Background: Digital media has emerged as a vital tool for the rapid dissemination of public health information to vast audiences. During the COVID-19 pandemic, social media-based health communication kept audiences informed, encouraged health-seeking behaviors, and helped address rampant misinformation. However, the impact of these digital communications was limited by a lack of training in leveraging online advertising tools, behavioral insights, and marketing practices to target, test, and scale social behavioral communications for health. The COVID-19 pandemic highlighted an urgent need for greater capacity among public health organizations to rapidly develop and deploy impactful, evidence-based digital campaigns.

Objective: This study aimed to evaluate changes in learner confidence and perceived ability to implement key skills after taking a massive online open course on digital campaign creation.

Methods: *Digital Media for Health Outcomes* was launched in May 2023 to train public health practitioners in all steps of campaign development, from evidence and insights collection to outcome evaluation. In April 2025, course performance was evaluated using a pre-post survey design. Learners were asked to complete identical surveys before and after completing the course. Question topics included basic demographic information and learners' confidence and experience in applying key campaign development skills. Learners were also asked to rate their confidence in using digital media to drive health outcomes before and after taking the course, and their perceived importance, ease, likelihood, and effectiveness of doing so. Proportions and 95%

XSL•FO
RenderX

confidence intervals were calculated for confidence questions, and proportion changes were compared between the baseline and endline surveys. All analyses were conducted using Stata (Special Edition 18; StataCorp).

Results: As of April 2025, the course had enrolled 14,170 individuals in 9 language versions, representing over 160 unique countries from all global regions. A total of 4392 learners completed the baseline survey, and 634 completed the course and endline survey. Most course completers came from Chile (240/634, 38%) and the United States (69/634, 11%). Learners who completed the course reported significantly higher confidence (+18%, 95% CI 17%-20%) in their perceived ability to use digital media to drive health outcomes and in applying skills associated with effective campaign development, including leveraging behavioral insights (+16%, 95% CI 14%-18%) and evaluating campaign outcomes (+28%, 95% CI 26%-29%). Overall, 92% (583/634) of course completers felt that the course had met most, all, or exceeded their expectations, and 85% (539/634) indicated that they would recommend the course to a colleague.

Conclusions: Performance evaluation of the *Digital Media for Health Outcomes* course suggests the potential of an expert-led, practice-based massive open online course to improve confidence and skill level in leveraging digital media for effective health communication among public health professionals. Further capacity strengthening is needed to encourage skill application.

(J Med Internet Res 2026;28:e85016) doi:10.2196/85016

KEYWORDS

digital health; health campaign; online course; massive online open course; MOOCs; public health campaign; social media

Introduction

Communication is an essential component of public health practice. During the COVID-19 pandemic, the need for effective public health communication became critically apparent. The conflation of emergent and evolving information and public guidance around disease prevention and spread—in addition to the interruption of routine health services, introduction of new vaccines, discovery of information vacuums, and explosion of mis- and disinformation on and offline—eroded public trust in health organizations and messages. Including the United Nations International Children's Emergency Fund (UNICEF), the World Health Organization (WHO), and the United States Centers for Disease Control and Prevention (CDC) on infection prevention, management, and treatment. In the years since the pandemic, social media–based campaigns have become increasingly important in combatting the ongoing "infodemic," characterized by the overload of online information, including mis- and disinformation. [1,2].

The COVID-19 pandemic also highlighted the potential of digital communication—particularly social media—as a cost-effective way to reach vast audiences with relevant, timely, and targeted public health updates. As of 2024, over 94% of global internet users use social media monthly, with platforms like Facebook (Meta; 3 billion users), YouTube (Google; 2.5 billion), Instagram (Meta; 2 billion), WhatsApp (Meta; 2 billion), and TikTok (ByteDance; 1.9 billion) leading global engagement [3-5]. During the pandemic, health professionals and organizations began using or increased their use of these platforms to reach audiences with updates on infection prevalence and mitigation strategies; several studies have found that information shared about COVID-19 via social media influenced audiences' knowledge, attitudes, and self-reported behavior [6-8]. However, few studies have assessed the effect of offline campaigns on online behavior [9].

Effective health communication strategies must be grounded in behavioral evidence and tailored for context. A recent systematic review of insights-based interventions described several effective strategies to shift health attitudes and behaviors, among them message framing, educational campaigns to raise awareness and knowledge, and the use of trusted messengers [10]. These interventions factor in local context, cultural considerations, communication preferences, behavioral evidence, and proven behavior change levers to effectively address concerns and questions among intended audiences [11-14]. Overall, evidence-based behavior change communication relies on a keen understanding of intended audiences, their values, and their concerns to provide trustworthy, impactful health messages.

To effectively communicate and meaningfully influence attitudes and behaviors via digital media, public health communicators need to apply social and behavioral change communication (SBCC) principles to digital communication campaigns but often lack the requisite skill set. In collaboration with the Ad Council, African Union | Africa CDC, Meta, Population Services International, UNICEF, the Yale Institute for Global Health, and the Poorvu Center for Teaching and Learning at Yale University, we designed an online course to lead practitioners through the process of developing and evaluating a digital health SBCC campaign. The *Digital Media for Health Outcomes* (DMHO) course draws on the evidence and real-life experience of diverse partner organizations to deliver accessible, practical, and effective training in digital social and behavior change communication for public health. Since its public launch in May 2023, the DMHO course has enrolled more than 14,170 learners from 162 countries and certified over 630 public health communicators. Here we present lessons learned from course evaluation data and offer recommendations for strengthening capacity in public health communication through training and online education programs.

Methods

Course Development

In 2021, as part of its global response to the COVID-19 pandemic, Meta spearheaded a pilot course with UNICEF, Capulet, and other partners to support public health agencies

XSL•FO
RenderX

and nongovernmental organizations in developing effective health communications for the digital space. In 2022, the DMHO course was proposed to build on this work, and the course syllabus was developed collaboratively during an all-partner meeting in August 2022.

Delivered over 5 modules, the course was designed to support public health agencies and organizations in creating effective digital health behavior change campaigns. From November to December 2022, we piloted the course with a private cohort of 100+ health and communication practitioners from partner organizations over a 6-week period to test, validate, and strengthen content and delivery. The course was launched for public consumption in May 2023. As a part of course maintenance and improvement, we performed regular course evaluations and updates (Figure 1).

Figure 1. Development and evaluation timeline for the Digital Media for Health Outcomes course.

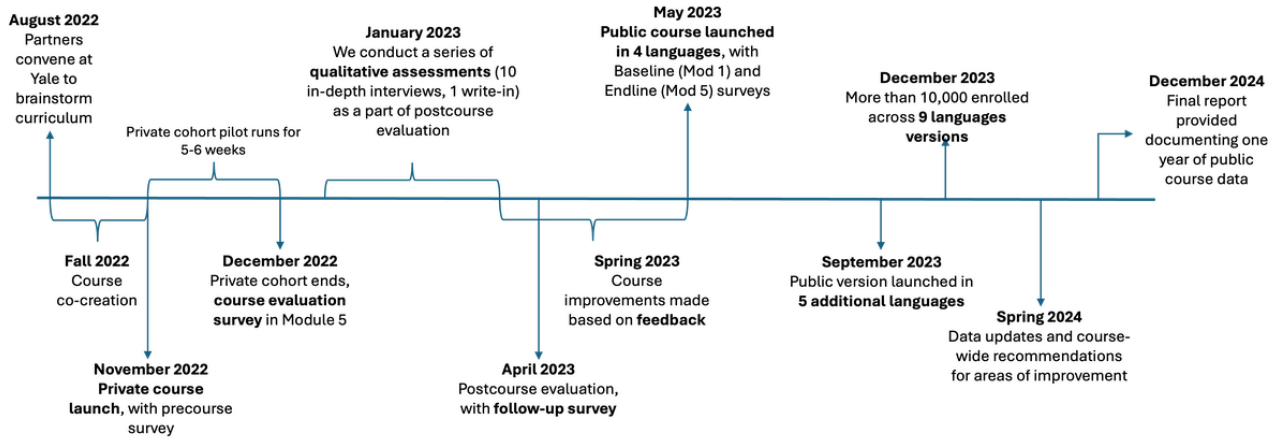

Course Structure

The course consisted of the following 5 modules (Table 1). Each module consisted of instructional videos, in-video quizzes, readings, and expert case studies. Progress was assessed through a simple, relevant homework assignment graded through the Coursera peer review system, by which course learners were able to leave feedback on each other's work. The course was hosted on Coursera, an online learning platform offering free and tiered payment access [15]. The course was available in all 9 United Nations and African Union languages, including Arabic, Bahasa Indonesia, Chinese, English, French, Portuguese, Russian, Spanish, and Swahili. Translations of all course content were done by WeTranslate, a professional translation agency. The course content was available for free; if learners wanted to receive a Yale certificate upon course completion, they could pay for a version of the course costing approximately US $50, payable either by individual learners, learner organization, or a Coursera Plus subscription. However, all those who enrolled could apply for Financial Aid and receive the certificate for free.

Table 1. Course modules.

Module	Title	Description
1	Behavioral Insights as a Foundation	Introduces learners to the concept of behavioral insights and demonstrates how these insights can be identified to inform evidence-based, targeted health communications. Africa CDC[a] presents a case study on Ebola response strategies.
2	Crafting Your Communication Strategy	Learners identify their target audiences and craft a strategic brief, following best practices shared by experts. The Ad Council presents a field perspective on a successful lung cancer screening campaign, Saved by the Scan.
3	Designing for Context: Messaging and Creative that Resonate	Dives into the best practices and key considerations for creating impactful and visually appealing creative assets, with messaging that is meaningful to the target audience. The Ad Council presents a field perspective on creative best practices, while UNICEF[b] highlights vaccine-specific messaging strategies.
4	Tactics for Digital Media and Campaign Implementation	Discusses the practical steps for launching a campaign on social media, focused on best practices for different platforms and metrics of evaluation with a case study from Population Services International.
5	Metrics that Matter: Understanding Impact	Describes several ways to assess health outcomes in response to health communication campaigns to understand the impact of the work. Case studies are presented by the World Health Organization and Yale Institute for Global Health.

[a]CDC: Centers for Disease Control and Prevention.

[b]UNICEF: United Nations International Children's Emergency Fund.

XSL•FO
RenderX

Course Pretest and Launch

The course was first launched for a pilot cohort of public health practitioners from November 2022 to December 2022. Learners, most of whom were UNICEF, WHO, or Ministry of Health employees, were invited based on course partner recommendations to participate in a 6-week, guided version of the course, and asked to provide in-depth feedback on course elements and overall flow. The private cohort included 107 learners representing 42 countries and 55 unique organizations; 83% (88/107) completed the course and submitted postcourse evaluation surveys. Furthermore, 10 completers were randomly selected to participate in individual postcourse interviews, conducted online by the Yale Institute for Global Health and lasting approximately 30-45 minutes. The results of these evaluations were used to update the course for public launch. A public version of the course was officially launched in English, Arabic, French, and Spanish in May 2023. Versions in the additional 5 languages were launched in September 2023.

Survey Development

The course was evaluated using a pre-post survey design hosted on Qualtrics [16]. At the start of the course (module 1), learners were asked to complete a confidential survey (referred to as the "baseline survey") to establish their baseline experience in using digital media to promote health outcomes. At the end of the course, learners were asked to complete a similar survey (referred to as the "endline survey"), which also asks them to reflect on elements of the course, what they liked, and what they did not. Survey questions were validated during the pilot version of the course and are available in Multimedia Appendix 1. Survey translations (from English to other languages) were done using Qualtrics' built-in translation software and verified by knowledgeable team members using DeepL, an online translation program recognized for accuracy. All translations were carefully reviewed for consistency with the English version before survey launch. Within the survey, participants were able to toggle between languages as desired.

Both surveys asked learners a series of demographic questions, including what type of organization they worked for and where their work was based. In both surveys, key performance indicators (KPIs) included learners' confidence and experience in applying six fundamental campaign skills including (1) understanding and identifying behavioral insights, (2) identifying a target audience online, (3) using best practices in designing creatives, (4) planning a digital communication strategy to drive health outcomes, (5) using practical tactics to implement the communication strategy, and (6) evaluating campaign outcomes. Learners were also asked to rank their confidence from low to high (0-10), with scores of 0-5 categorized as low confidence and scores of 6-10 categorized as high confidence. Similarly, learners were asked to report their perceived confidence in and importance of, ease of, likelihood to, and effectiveness of using digital media to drive health outcomes.

Statistical Analysis

Data presented in this study were collected from May 2023 to April 2025. All data were self-reported. Proportions and 95% CIs were calculated for all confidence questions, and proportion change was compared between baseline and endline surveys. Incomplete surveys were removed before starting the analysis.

We also conducted a subanalysis of respondents for whom we could link baseline and endline survey results. Response datasets were merged by respondent age, gender, and current organization, provided in the baseline and endline surveys. Demographics and KPIs were assessed as in the main analysis. Deidentified data from both surveys were analyzed using STATA (StataCorp).

Ethical Considerations

This analysis was classified as a program evaluation (a type of nonhuman participant research) by the Institutional Review Board at the University of Texas Southwestern Medical Center. As such, the study was approved as nonregulated research (STU-2024-0962). Elements of this report were included based on the Western Michigan University Checklist for Program Evaluation Reporting Content. Participants received no compensation for participation. The first page of both surveys advised participants that their participation was entirely voluntary and that their responses would be kept confidential and used to evaluate course performance over time. Respondents provided their consent to participate by proceeding to the next page of the survey.

Results

Baseline Learner Demographics

As of April 2025, a total of 14,170 individuals were enrolled in the DMHO course, and 5971 learners had started the course. A participant flow diagram is presented in Figure 2.

XSL•FO
RenderX

Figure 2. Participant flow diagram.

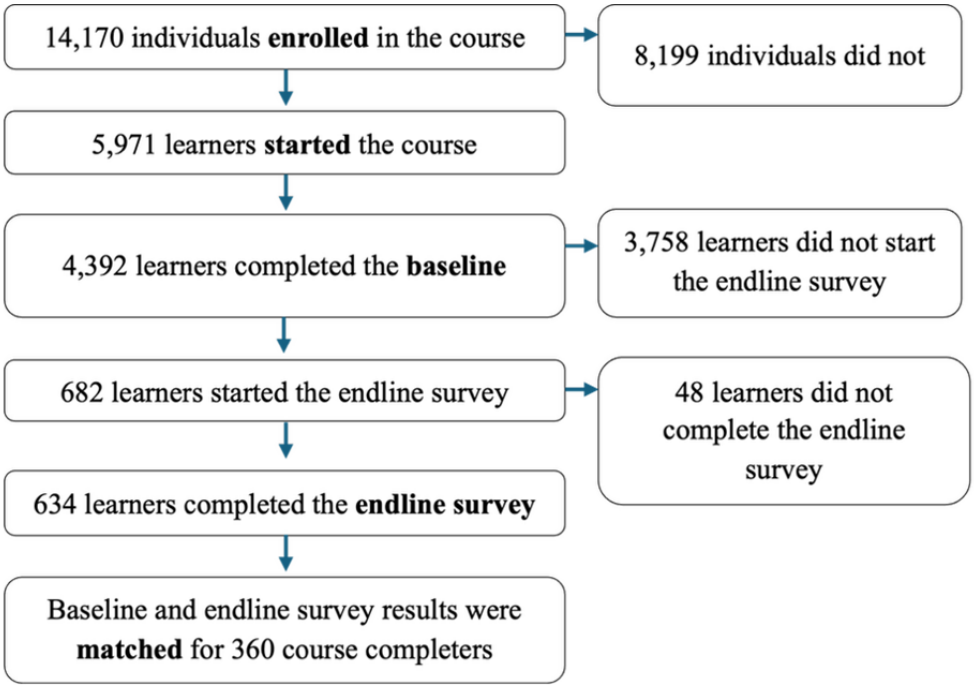

Of these, 4329 (73%) completed the baseline survey (Table 2). The top 5 language versions among learners were English (3394/4329, 78%), Spanish (390/4329, 9%), French (211/4329, 5%), Arabic (116/4329, 3%), and Russian (89/4329, 2%). Most learners identified as women (3010/4329, 69%), were older than 25 years (1974/2464, 80%), and had received at least a bachelor's degree (3519/4329, 81%). Many learners worked at academic institutions (505/4329, 12%), ministries of health and government agencies (447/4329, 10%), and in the private sector (600/4329, 14%). Several learners were university students (427/4329, 10%). Geographically, learners represented 162 countries and all UNICEF regions. The five countries with the greatest shares of learners were the United States (986/4329, 22%), Chile (208/4329, 5%), India (196/4329, 5%), Canada (153/4329, 4%), and Nigeria (140/4329, 3%; Figure 3). Regionally, learners were predominantly from North America (1121/4311, 26%), Latin America and the Caribbean (597/4311, 14%), and Western Europe (582/4311, 14%).

XSL•FO
RenderX

Table 2. Learner demographics for the Digital Media for Health Outcomes course at baseline (April 2025).

Demographic	Baseline survey results, n (%)
Survey language (N=4329)	
Arabic	116 (2.7)
Bahasa Indonesian	16 (0.4)
Chinese	33 (0.8)
English	3394 (78)
French	211 (4.9)
Portuguese	80 (1.9)
Russian	89 (2.1)
Spanish	390 (9)
Swahili	0 (0)
Gender (N=4329)	
Men	1210 (28)
Women	3010 (70)
Other (nonbinary, self-describing, prefer not to say)	109 (2.5)
Age (y; n=2464)	
13-17	40 (1.6)
18-24	451 (18)
25-34	700 (28)
35-44	663 (27)
45-54	363 (15)
55-64	197 (8)
65+	50 (2)
Education level (N=4329)	
No high school	20 (0.5)
High school	391 (9)
Associates degree, Technikon, or Technical	233 (5.4)
University (Bachelor)	1508 (35)
University (Master)	1372 (32)
Graduate degree (MD or PhD)	639 (15)
Other	166 (3.8)
Type of organization (N=4329)	
Academic	505 (12)
Advocacy group	37 (0.85)
International nongovernmental organization (iNGO)	146 (3.4)
Marketing, Media, or Creative	146 (3.4)
Ministry of Health, or Government	447 (10)
Local nongovernmental organization (NGO)	266 (6.1)
Private sector	600 (14)
Self-employed	400 (9)
Student	427 (9.9)
UN Organization (ie, UNICEF[a] and WHO[b])	122 (2.82)
Unemployed	328 (7.6)

XSL•FO
RenderX

Demographic	Baseline survey results, n (%)
Other or not applicable[c]	905 (21)
Countries represented (top 10; N=4329)	
United States	986 (22)
Chile	208 (4.8)
India	196 (4.5)
Canada	153 (3.5)
Nigeria	140 (3.2)
United Kingdom	111 (2.6)
Brazil	104 (2.4)
Kenya	102 (2.4)
Ghana	85 (2)
Egypt	81 (1.9)
UNICEF regions[d] (N=4311)	
East Asia and Pacific	414 (9.6)
Eastern Europe and Central Asia	232 (5.4)
Western Europe	582 (14)
Latin American and Caribbean	597 (14)
Middle East and North Africa	321 (7.5)
North America	1121 (26)
South Asia	321 (7.5)
Eastern and Southern Africa	398 (9.2)
West and Central Africa	325 (0.54)

[a]UNICEF: United Nations International Children's Emergency Fund.

[b]WHO: World Health Organization.

[c]Includes organizations that did not fit predetermined classifications or were not identifiable.

[d]UNICEF regional classifications [17].

Figure 3. Countries represented by Digital Media for Health Outcomes course learners at baseline, April 2025 (n=4392).

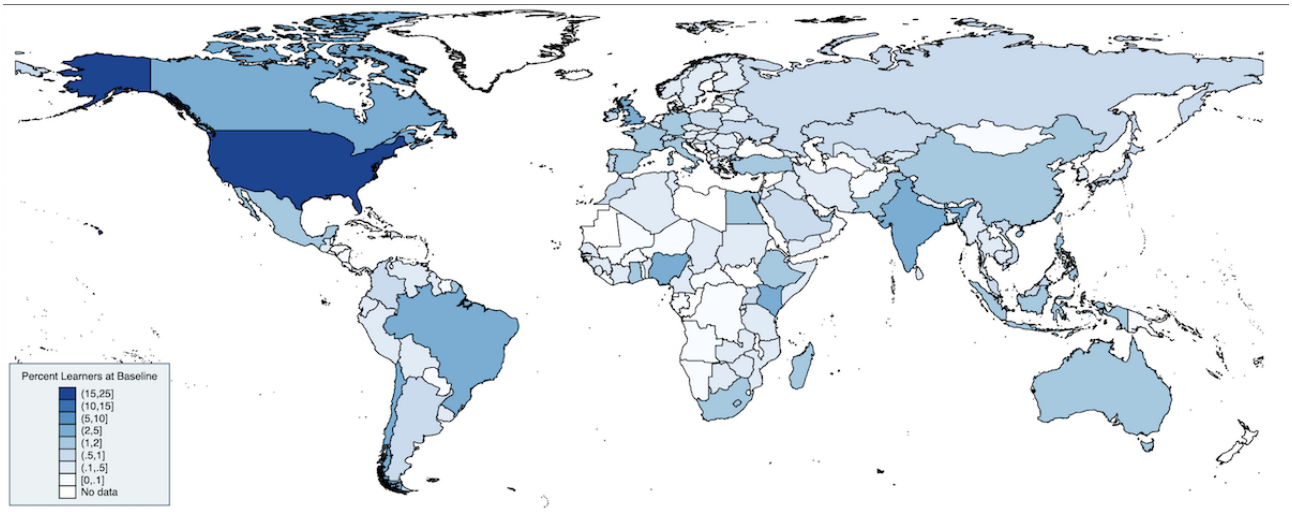

KPIs for Learners at Baseline

When asked to rank which skills they wanted to learn from the course, learners ranked "understanding and identifying behavioral insights" as the most important skill to learn (Figure 4). Learners ranked "identifying target audiences online" second, "using best practices in designing creative [campaign images and graphics]" and "planning a digital communication strategy

XSL•FO
RenderX

to drive health outcomes" third and fourth, "learning practical tactics to implement the communication strategy" fifth, and "evaluating campaign outcomes related to SBCC and health communications" last. Overall, most learners at baseline were confident in using these skills and found using digital media to drive health outcomes to be important, easy, and effective, and reported they were likely to do so (Table 3).

Figure 4. Skills ranked "most important to learn" by Digital Media for Health Outcomes course learners before taking the course (baseline).

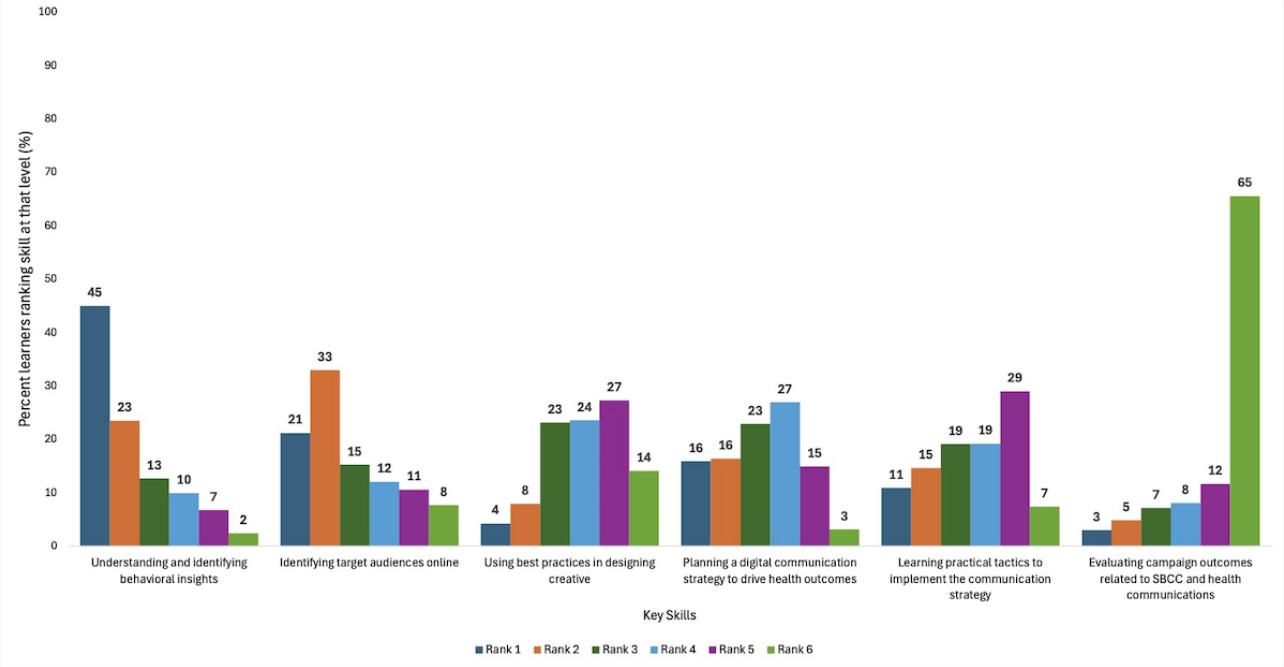

XSL•FO
RenderX

Table 3. Percent point change of learners reporting high confidence in skills before and after taking the Digital Media for Health Outcomes course (April 2025).

Questions	Baseline confidence, n (%)		Endline confidence, n (%)		Percent point change in high confidence (95% CI)	*P* value
	Low (0-5)	High (6-10)	Low (0-5)	High (6-10)		
Please rate how confident you are in each of the following skills on a scale of 0-10						
Understanding and identifying behavioral insights	846 (20)	3483 (80)	23 (3.6)	611 (96)	16% (14%-18%)	<.001
Identifying a target audience online	960 (22)	3369 (78)	20 (3.2)	614 (97)	19% (17%-21%)	<.001
Planning a digital communication strategy to drive health outcomes	1494 (35)	2835 (66)	23 (3.4)	611 (96)	31% (29%-33%)	<.001
Using practical tactics to implement the communication strategy	1428 (33)	2901 (67)	21 (3.3)	613 (97)	30% (28%-32%)	<.001
Evaluating campaign outcomes related to SBCC[a]/health communications	1334 (31)	2995 (69)	21 (3.3)	613 (97)	28% (26%-29%)	<.001
On a scale of 0-10						
How confident are you in your ability to use digital media to drive health outcomes?	867 (20)	3462 (80)	11 (1.7)	623 (98)	18% (17%-20%)	<.001
How important do you think it is for your organization to use digital media to drive health outcomes?	241 (5.6)	4088 (94)	6 (0.95)	628 (99)	5% (4%-6%)	<.001
How easy is it in your workplace to integrate digital media as a tool to drive health outcomes?	535 (12)	3794 (88)	15 (2.4)	619 (97)	9% (7%-11%)	<.001
How likely are you to use digital media to reach your organization's health goals?	321 (7.4)	4008 (93)	8 (1)	626 (99)	6% (5%-7%)	<.001
How effective is digital media in helping you reach your organization's health goals?	410 (9.5)	3919 (91)	8 (1)	626 (99)	8% (7%-9%)	<.001

[a]SBCC: social and behavioral change communication.

When asked how often they used digital media to drive health outcomes, 39% (1668/4329) learners responded "never or rarely," 34% (1467/4329) responded "sometimes," and 28% (1194/4329) responded "often or always." Stratifying by learners' organization type, we saw that students were significantly less likely to use digital media to drive health outcomes. Similarly, when asked how often they evaluated health outcomes in response to digital campaigns, 61% (2551/4157) responded "never or rarely," 28% (1156/4157) responded "sometimes," and 11% (450/4157) responded "often or always" (Table 4). Stratification by the learner organization did not show any significant differences.

Table 4. Frequency of digital media use and evaluation among learners before and after taking the Digital Media for Health Outcomes course (April 2025).

Questions	Baseline, n (%)	Endline, n (%)	Percent point change between baseline and endline (95% CI)	*P* value
How often do you use digital media to drive health outcomes?[a]				
Never or rarely	1668 (39)	206 (33)	–6% (–10% to –2%)	.004
Sometimes	1467 (34)	217 (34)	3% (–4% to 4%)	.89
Often or always	1194 (28)	211 (33)	6% (2% to 10%)	.003
How often do you evaluate health outcomes in response to digital campaigns?[b]				
Never or rarely	2551 (61)	293 (49)	–12% (–16% to –8%)	<.001
Sometimes	1156 (28)	204 (34)	7% (2% to 11%)	.001
Often or always	450 (11)	97 (16)	6% (2% to 9%)	<.001

[a]Baseline N=4329, endline n=634.

[b]Baseline n=4157, endline n=594.

XSL•FO
RenderX

Demographics of Course Completers

Of the 682 (93%) learners who started the endline survey, 634 learners completed it as of April 2025 (Table 5). The English (316/634, 50%) and Spanish (234/634, 37%) versions of the course had the most completers. Most completers identified as women (464/634, 73%), were between 18 and 34 years of age (379/548, 69%), and had earned at least their bachelor's degrees (399/634, 63%). Many course completers were students (226/634, 36%) or worked for ministries of health and government agencies (30/634, 4.7%), academic organizations (59/634, 9%), or the private sector (56/634, 9%). The top three countries represented were Chile (240/634, 38%), the United States (69/634, 11%), and Madagascar (26/634, 4%), and the top 3 regions were Latin America and Caribbean (271/634, 43%), North America (79/634, 13%), and Eastern and Southern Africa (76/634, 12%; Figure 5).

XSL•FO
RenderX

Table 5. Learner demographics for the Digital Media for Health Outcomes course at endline (April 2025).

Demographic	Baseline survey results, n (%)
Survey language (n=634)	
Arabic	6 (0.95)
Bahasa Indonesian	6 (0.95)
Chinese	7 (1.1)
English	316 (50)
French	49 (7.7)
Portuguese	10 (1.6)
Russian	6 (0.95)
Spanish	234 (37)
Swahili	0 (0)
Gender (n=634)	
Men	156 (25)
Women	464 (73)
Other (nonbinary, self-describing, prefer not to say)	14 (2.2)
Age (y; n=548)	
13-17	4 (0.73)
18-24	265 (48)
25-34	114 (21)
35-44	103 (19)
45-54	42 (7.7)
55-64	17 (3.1)
65+	3 (0.55)
Education level (n=634)	
No high school	0 (0)
High school	166 (26)
Associates degree, Technikon, or Technical	31 (4.9)
University (Bachelor)	203 (32)
University (Master)	153 (24)
Graduate degree (MD or PhD)	46 (7.3)
Other	35 (5.5)
Type of organization (n=634)	
Academic	59 (9.3)
Advocacy group	2 (0.32)
International nongovernmental organization (iNGO)	30 (4.7)
Marketing, Media, or Creative	30 (4.7)
Ministry of Health or Government	30 (4.7)
Local nongovernmental organization (NGO)	32 (5.1)
Private sector	56 (8.8)
Self-employed	24 (3.8)
Student	226 (36)
UN Organization (ie, UNICEF[a] and WHO[b])	14 (2.2)
Unemployed	37 (5.8)

XSL•FO
RenderX

Demographic	Baseline survey results, n (%)
Other or not applicable	52 (8.2)
Countries represented (top 10; n=634)	
Chile	240 (38)
United States	69 (11)
Madagascar	26 (4.1)
India	19 (3)
Kenya	16 (2.5)
Nigeria	13 (2.1)
Uganda	12 (1.9)
Indonesia	11 (1.7)
Canada	10 (1.6)
Ethiopia	10 (1.6)
UNICEF regions represented[c] (n=631)	
East Asia and Pacific	47 (7.5)
Eastern Europe and Central Asia	16 (2.5)
Western Europe	57 (9)
Latin American and Caribbean	271 (43)
Middle Easte and North Africa	17 (2.7)
North America	79 (13)
South Asia	32 (5.1)
Eastern and Southern Africa	76 (12)
West and Central Africa	36 (5.7)

[a]UNICEF: United Nations International Children's Emergency Fund.
[b]WHO: World Health Organization.
[c]UNICEF regional classifications [17].

Figure 5. Countries represented by Digital Media for Health Outcomes course completers at endline, April 2025 (n=634).

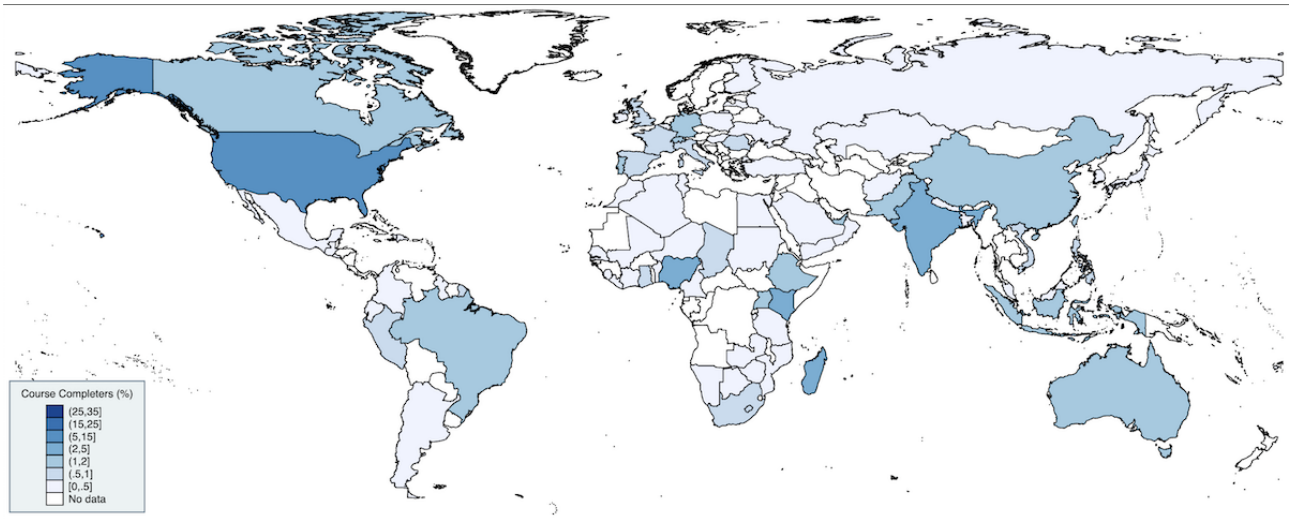

KPIs for Course Completers

There was a significant increase in skill confidence among course completers compared with baseline learners (Table 3). Respondents at endline were significantly more confident in all skills, especially planning a digital communication strategy, with 66% (2835/4329) reporting high confidence at baseline compared to 96% (611/634) at endline (+16%, 95% CI 14%-18%, P<.001), using practical tactics to implement that strategy (+30%, 95% CI 19%-33%, P<.001), and evaluating

XSL•FO
RenderX

health outcomes after campaign implementation (+28%, 95% CI 26%-29%, *P*<.001). Other KPIs also showed significant improvement from baseline to endline. Learners were significantly more confident in their ability to use digital media (+18%, 95% CI 17%-20%, *P*<.001). Similarly, more endline respondents indicated using digital media was important (+5%, 95% CI 4%-6%, *P*<.001), easy (+9%, 95% CI 7%-11%, *P*<.001), effective (+8%, 95% CI 7%-9%, *P*<.001), and that they were likely to use it (+6%, 95% CI 5%-7%, *P*<.001).

When asked how often they used digital media to drive health outcomes, 33% (206/634) responded "never or rarely," 34% (217/634) responded "sometimes," and 33% (211/634) responded "often or always." (Table 4). Stratifying by learners' organization type showed that students were significantly less likely to use digital media as compared with other learners. When asked how often they evaluated health outcomes in response to digital campaigns, 50% (293/594) responded "never or rarely," 34% (204/594) responded "sometimes," and 16% (97/594) responded "often or always." Stratification by organization type did not reveal any differences in the frequency of evaluation. From baseline to endline, there was a significant decrease in the number of respondents who never or rarely used digital media (–6%, *P*=.004), and a significant increase in the number of respondents who often or always used digital media (+6%, *P*=.003; Table 4). Similarly, there was a significant decrease in the number of respondents who never or rarely evaluated health outcomes from baseline to endline (–12%, *P*<.001), and significant increases among respondents who sometimes (+7%, *P*=.001) or often or always evaluated outcomes (+6%, *P*<.001).

Longitudinal Analysis of Course Completers

We compared baseline and endline survey responses for 360 (of 634) completers for whom we could link unique survey submissions (Table 6). Of these, most completed the English (210/360, 58%) and Spanish (120/360, 33%) versions of the course, 79% (285/360) identified as women, 55% (149/271) were between 18 and 24 years old, and most had completed at least their bachelor's degree (252/360, 70%). The top-represented country was Chile (124/360, 34%), followed by the United States (43/360, 12%).

XSL•FO
RenderX

Table 6. Learner demographics matched between baseline and endline surveys for the Digital Media for Health Outcomes course (April 2025).

Demographic	Survey results, n (%)
Survey language (n=360)	
Arabic	1 (0.28)
Bahasa Indonesian	3 (0.83)
Chinese	2 (0.56)
English	210 (58)
French	18 (5)
Portuguese	1 (0.28)
Russian	5 (1.4)
Spanish	120 (33)
Swahili	0 (0)
Gender (n=360)	
Men	68 (19)
Women	285 (79)
Other (nonbinary, self-describing, prefer not to say)	7 (1.9)
Age (y; n=271)	
13-17	2 (0.74)
18-24	149 (55)
25-34	40 (15)
35-44	59 (22)
45-54	16 (5.9)
55-64	4 (1.5)
65+	1 (0.37)
Education level (n=360)	
No high school	0 (0)
High school	87 (24)
Associates degree, Technikon, or Technical	23 (6.4)
University (Bachelor)	119 (33)
University (Master)	92 (26)
Graduate degree (MD or PhD)	16 (4.4)
Other	23 (6.4)
Type of organization (n=360)	
Academic	33 (9.2)
Advocacy group	0 (0)
International nongovernmental organization (iNGO)	20 (5.6)
Marketing, Media, or Creative	3 (0.83)
Ministry of Health or Government	26 (7.2)
Local nongovernmental organization (NGO)	13 (3.6)
Private sector	33 (9.2)
Self-employed	19 (5.3)
Student	117 (33)
UN Organization (ie, UNICEF[a] and WHO[b])	22 (6.1)
Unemployed	22 (6.1)

XSL•FO
RenderX

Demographic	Survey results, n (%)
Other or not applicable[c]	52 (14)
Countries represented (top 10; n=360)	
Chile	124 (34)
United States	43 (12)
Madagascar	15 (4.2)
Kenya	13 (3.6)
India	11 (3.1)
Philippines	10 (2.8)
Canada	9 (2.5)
Nigeria	9 (2.5)
Pakistan	9 (2.5)
Indonesia	8 (2.2)
UNICEF regions[d] (n=359)	
East Asia and Pacific	34 (9.5)
Eastern Europe and Central Asia	10 (2.8)
Western Europe	27 (7.5)
Latin American and Caribbean	137 (38)
Middle East and North Africa	9 (2.5)
North America	52 (15)
South Asia	26 (7.2)
Eastern and Southern Africa	47 (13)
West and Central Africa	17 (4.7)

[a]UNICEF: United Nations International Children's Emergency Fund.

[b]WHO: World Health Organization.

[c]Includes organizations that did not fit predetermined classifications or were not identifiable.

[d]UNICEF regional classifications [17].

As in the main analysis, course completers reported significantly higher confidence in key skills compared to baseline (Table 7). Notably, completers saw increases in confidence of 14% percentage points for planning a digital communications strategy (95% CI 0.09-0.19, $P<.001$) and 16% percentage points for outcome evaluation (95% CI 0.1-0.21, $P<.001$). Similarly, completers' confidence in their ability to use digital media, the likelihood they would do so, and their perception of its effectiveness improved significantly after taking the course (Table 7). Between baseline and endline, no significant change was seen in how often course completers used digital media for health campaigns or evaluated health outcomes following those campaigns.

XSL•FO
RenderX

Table 7. Percent point change of learners reporting high confidence for direct comparison of the Digital Media for Health Outcomes course completers, baseline to endline.

Questions	Baseline confidence, n (%)		Endline confidence, n (%)		Percent point change in high confidence (95% CI)	*P* value
	Low (0-5)	High (6-10)	Low (0-5)	High (6-10)		
Please rate how confident you are in each of the following skills on a scale of 0-10						
Understanding and identifying behavioral insights	58 (16)	302 (84)	30 (8.3)	330 (92)	8% (3% to 13%)	.001
Identifying a target audience online	52 (14)	308 (86)	6 (1.7)	354 (98)	13% (9% to 17%)	<.001
Planning a digital communication strategy to drive health outcomes	78 (22)	282 (78)	27 (7.5)	333 (93)	14% (9% to 19%)	<.001
Using practical tactics to implement the communication strategy	75 (21)	285 (79)	29 (8.1)	331 (92)	13 (8% to 18%)	<.001
Evaluating campaign outcomes related to SBCC[a]/health communications	83 (23)	277 (77)	27 (7.5)	333 (93)	16% (10% to 21%)	<.001
On a scale of 0-10						
How confident are you in your ability to use digital media to drive health outcomes?	43 (12)	317 (88)	21 (5.8)	339 (94)	6% (2% to 10%)	.004
How important do you think it is for your organization to use digital media to drive health outcomes?	11 (3.1)	349 (97)	20 (5.6)	340 (94)	–2.5% (–55% to 0.5%)	.10
How easy is it in your workplace to integrate digital media as a tool to drive health outcomes?	30 (8.3)	830 (92)	25 (6.9)	335 (93)	14% (–2% to 0.5%)	.49
How likely are you to use digital media to reach your organization's health goals?	14 (4.2)	345 (96)	2 (0.56)	358 (99)	3.6% (1.4% to 5.8%)	.001
How effective is digital media in helping you reach your organization's health goals?	22 (6.1)	338 (94)	23 (6.4)	337 (94)	54% (48% to 59%)	<.001

[a]SBCC: social and behavioral change communication.

Course Evaluation

Course completers were asked to evaluate the overall course and specific modules. Generally, it took most learners less than 1 month to complete the course (348/634, 55%) and <4 hours to complete each module (475/634, 75%). Most respondents (570/634, 90%) reported using a laptop or computer to take the course, and 90% (570/634) of learners agreed that, from where they took the course, the content was easy to load and view.

Over 70% (444/634) of respondents reported that they had applied some or all skills from the course to their work. Commonly applied skills included "using behavioral insights to develop social and behavior change communication" (190/634, 30% and 133/634, 21% indicated often or always, respectively), "developing a strategic brief" (197/634, 31% and 139/634, 22%), and "using Canva or a similar platform to create campaign assets" (184/634, 29% and 228/634, 36%).

When asked if they had used Meta to design custom target audiences, used A/B testing to assess their campaign strategy, or evaluated health outcomes following their campaign, the most common response was "not yet, but I plan to" (197/634, 31%, 228/634, 36%, and 208/634, 33% respectively; Figure 6).

XSL•FO
RenderX

Figure 6. Skills applied by course completers since taking the Digital Media for Health Outcomes course, as of April 2025.

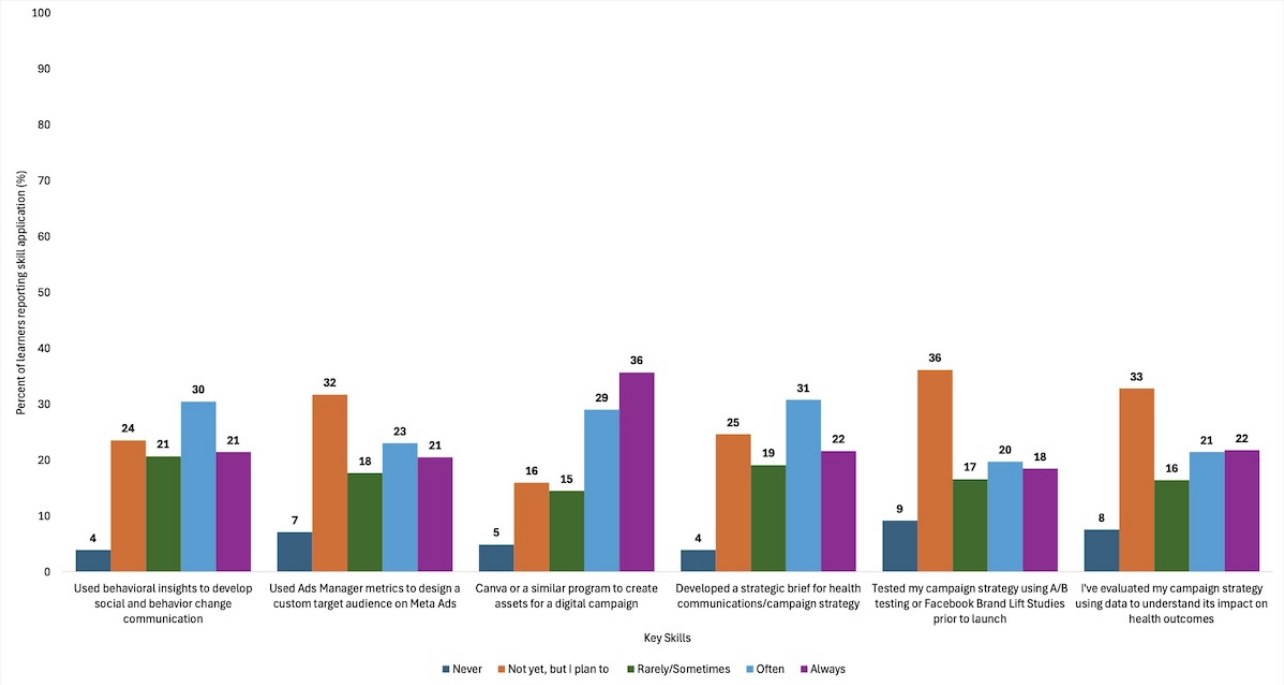

Overall, 92% (583/634) of course completers felt that the course had met most, all, or exceeded their expectations. When asked how satisfied they were with individual course elements, most learners agreed they were satisfied with the instructor videos (577/634, 91%), case studies (583/634, 92%), the Coursera interface (570/634, 90%), course pacing (577/634, 91%), homework assignments (558/634, 88%), quizzes (552/634, 87%), and translation (539/634, 85%). The peer-review process was the least satisfactory, with only 78% (494/634) satisfaction. Anonymous (and optional) comments at the end of the survey cited wishes that all assignments could be reviewed by course instructors or content experts and graded at more regular intervals rather than by the Coursera peer-review system. Individual modules were also well-received by respondents. For each module, the majority of course completers agreed that the material was new to them, easy to understand, and relevant to their work (Table S1 in Multimedia Appendix 2).

Course Recommendation

When asked how likely they were to recommend the course to a colleague or friend, 85% (539/634) indicated that they were likely to do so. When asked whether they would share their certificate of completion (if applicable) on LinkedIn, 64% (406/634) indicated that they would (159/634, 25% indicated "not applicable" either because they chose to complete the course without certification or because they do not have a LinkedIn page).

Discussion

Principal Findings

The DMHO course was designed to be accessible and useful to public health practitioners with limited experience in behavioral evidence-based health communication. Notably, the course has reached thousands of learners from countries in Asia, Africa, and Latin America. A course that is available online, free of charge, and in multiple languages has the potential to make useful training available to audiences worldwide.

The COVID-19 pandemic demonstrated the need for strategic campaigns on digital media, specifically social media, to reach diverse audiences with essential information, especially in the context of widespread online misinformation. Organizations must be equipped both to communicate quickly and effectively during a crisis and to maintain credibility as a trusted voice for public health information. The DMHO course provides an evidence-based framework for the development of campaigns that can be applied to systematically target health behaviors, offer compelling information, and promote demand for services. Encouragingly, the evaluation revealed that the course improved self-reported confidence among learners from baseline to endline, along with self-reported endorsement that using digital media to drive health outcomes was important, easy, and effective. Direct comparison of baseline versus endline survey results for learners who completed both revealed similarly promising results to overall analyses. Learners' perceived ability to apply key skills improved from baseline to endline, as did their confidence in using digital media, how effective they believed it to be, and how likely they were to use it. Moreover, more respondents indicated they were likely to use digital media to achieve their organization's goals at endline than at baseline. These results are especially promising in the context of our diverse learner pool, about half of whom are based in Africa, Latin America, and the Caribbean.

Many organizations, including our partners on the DMHO course, have used digital campaigns to share pertinent information about health topics and encourage recommended behaviors. While few academic studies have documented the direct impact of such campaigns on health outcomes, targeted social and behavior change communication has been used to improve health knowledge, attitudes, and health-seeking behaviors [9,18-20]. To develop the DMHO course, we

XSL•FO
RenderX

combined evidence from multiple experts and experiences to produce a practice-based training that walks learners through the communication process, emphasizing behavioral theory and encouraging empirical evaluation of health outcomes. Designing and implementing campaigns without an evidence-based framework can limit their impact or have unintended consequences [21,22]. Skills-based massive open online courses (MOOCs), like DMHO, have the capacity to improve skills and shift attitudes toward important topics within and beyond public health. Similar evaluations have described successful launches of MOOCs to improve skill-based knowledge and reach a wide and diverse audience at a scale not possible with in-person training [23-25]. These evaluations noted that MOOC success is predicated upon engaging learners with creative and interactive teaching styles, and ensuring material is accessible in terms of format and language, all factors that we considered and tried to address when developing this course.

Our results highlight the need for further capacity strengthening, particularly in terms of applying best practices and evaluating impact. Most practitioners believe these skills are important, but many do not apply them in practice. While course completers reported they were more likely to use digital media often or always than baseline learners, most reported rarely doing so. Notably, many of the learners who completed the course were university students from Chile, and while they may not be using digital media to drive health outcomes now, the course may have exposed them to skills they could apply in the future. Similarly, many course completers indicated they were sometimes, often, or always evaluated health outcomes; however, nearly half reported never or rarely doing so. Evaluation of health outcomes is an important step in the development and implementation of effective digital health campaigns, but it can be challenging, particularly when the intended outcome is a behavior change. More information is needed to understand the say-do gap in the implementation of digital health communications among public health agencies, and how we can encourage public health practitioners to evaluate the impact of their campaigns.

This study has several limitations. All data are self-reported and subject to social desirability and bias, representing those who register and enroll for such a course. The diverse baseline sample mitigates this limitation, as do the results of the completer subanalysis. The subanalysis is limited by its small sample size, which is a function of the anonymity of the surveys. The global and likely heterogeneous nature of our results may obscure regional differences at baseline and endline assessments. While regional analysis was beyond the scope of this initial performance evaluation, we anticipate conducting future analyses at the regional level as course completion increases across regions.

Furthermore, our content is insufficient to keep pace with the dynamic and changing environment of social media platforms and the different ways that various platforms leverage their audience through advertising and other algorithms. Nonetheless, the course provides a foundation for how to approach any SBCC campaign through whatever digital means are prevalent for a target audience. Further investigation is also needed to understand why there is such a high level of enrollment yet a relatively small percentage of completers. Data provided by Coursera course management dashboards shows that most learners who do not complete the course drop off in the first module, either immediately after registering for the course (before the baseline survey) or before the first homework assignment. While this is consistent with other Coursera courses presented by Yale University, a deeper analysis is warranted to promote sustained enrollment in the course through completion.

Finally, this performance evaluation is not sufficient to assess the “practice gap.” Understanding the long-term outcomes of digital campaign development training is essential to improving communication capacity among public health organizations and should be the focus of future studies.

Conclusion

The DMHO course suggests the potential of a practical, theory-based massive online course to improve public health practitioners’ confidence and skills in using digital media to drive health outcomes. However, the need for further capacity strengthening is clear. Digital health communication is an increasingly useful tool for public health practitioners, one which—if wielded strategically—has the continued ability to reach different audiences with the critical information they need and improve health outcomes.

Acknowledgments

The authors would like to acknowledge Chris Purdy (formerly of PSI), Kayla Steinberg (Yale School of Public Health) and Sarah Doty and Julie Szabo (Capulet) for their significant contributions to the success of this course.

No generative artificial intelligence was used for any portion of manuscript development or production.

Funding

Development and launch of the course were supported by gift funds provided by Meta Platforms, Inc (hereafter referred to as “Meta”) to the Yale School of Medicine. Representatives of Meta at the time (WT and KH) were involved in initial course development and launch and were updated periodically on course progress. WT and KH reviewed initial manuscript drafts but did not play a role in the analysis or presentation of results.

Data Availability

Deidentified aggregated data will be available for download from Open Science Framework [26].

XSL•FO
RenderX

Authors' Contributions

All authors were involved in the development and production of the course. HM, SC, A Titus, MW, AAM, and SBO conducted and oversaw course evaluation and initial manuscript drafting. HM, SC, and AAM had access to and verified the data and analyses. All authors reviewed and edited subsequent drafts, and all authors approved the final submission.

Conflicts of Interest

WT and KH are former employees of Meta Platforms, Inc. They were employed by Meta at the time of the study.

Multimedia Appendix 1
Questions for baseline and endline surveys instruments.
[PDF File (Adobe PDF File), 214 KB - jmir_v28i1e85016_app1.pdf]

Multimedia Appendix 2
Table S1. Module-specific evaluation.
[DOCX File , 16 KB - jmir_v28i1e85016_app2.docx]

References

1. Ferreira Caceres MM, Sosa JP, Lawrence JA, Sestacovschi C, Tidd-Johnson A, Rasool MHU, et al. The impact of misinformation on the COVID-19 pandemic. AIMS Public Health 2022;9(2):262-277 [FREE Full text] [doi: 10.3934/publichealth.2022018] [Medline: 35634019]
2. Managing the COVID-19 infodemic: promoting healthy behaviours and mitigating the harm from misinformation and disinformation. World Health Organization. 2020 Sep 23. URL: https://tinyurl.com/325bydez [accessed 2026-06-15]
3. SEO.AI. 2025 Jan 23. URL: https://seo.ai/blog/how-many-users-on-tiktok#:~:text=TikTok%20Key%20Statistics,to%2058%20minutes%20in%202024 [accessed 2026-04-18]
4. Most popular social networks worldwide as of February 2025, by number of monthly active users (in millions). Statista. 2025. URL: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/?srsltid=AfmBOopOPAjroAFMotsGTxxynvMkD4qdVWx69WS3h-ItKTQkvO1OcArK [accessed 2026-04-21]
5. Global social media statistics. DataReportal. URL: https://datareportal.com/social-media-users [accessed 2026-04-18]
6. Tsao S, Chen H, Tisseverasinghe T, Yang Y, Li L, Butt ZA. What social media told us in the time of COVID-19: a scoping review. Lancet Digit Health 2021 Mar;3(3):e175-e194 [FREE Full text] [doi: 10.1016/S2589-7500(20)30315-0] [Medline: 33518503]
7. González-Padilla DA, Tortolero-Blanco L. Social media influence in the COVID-19 pandemic. Int Braz J Urol 2020 Jul;46(suppl.1):120-124 [FREE Full text] [doi: 10.1590/S1677-5538.IBJU.2020.S121] [Medline: 32550706]
8. Ho HY, Chen YL, Yen CF. Different impacts of COVID-19-related information sources on public worry: an online survey through social media. Internet Interv 2020;22:100350 [FREE Full text] [doi: 10.1016/j.invent.2020.100350] [Medline: 33072524]
9. Winters M, Christie S, Lepage C, Malik AA, Bokemper S, Abeyesekera S, et al. Persuasive COVID-19 vaccination campaigns on Facebook and nationwide vaccination coverage in Ukraine, India, and Pakistan. PLOS Glob Public Health 2023;3(9):e0002357 [FREE Full text] [doi: 10.1371/journal.pgph.0002357] [Medline: 37756298]
10. Malik AA, Ahmed N, Shafiq M, Elharake JA, James E, Nyhan K, et al. Behavioral interventions for vaccination uptake: a systematic review and meta-analysis. Health Policy 2023 Nov;137:104894 [FREE Full text] [doi: 10.1016/j.healthpol.2023.104894] [Medline: 37714082]
11. Rosa WE, Levoy K, Doyon K, McDarby M, Ferrell BR, Parker PA, et al. Integrating evidence-based communication principles into routine cancer care. Support Care Cancer 2023 Sep 08;31(10):566 [FREE Full text] [doi: 10.1007/s00520-023-08020-x] [Medline: 37682354]
12. Roundtable on Population Health Improvement, Board on Population Health and Public Health Practice, Institute of Medicine. Effective messaging strategies: a review of the evidence. In: Communicating to Advance the Public's Health: Workshop Summary. Washington, DC: National Academies of Sciences, Engineering, and Medicine; 2015.
13. Riera R, de Oliveira Cruz Latorraca C, Padovez RCM, Pacheco RL, Romão DMM, Barreto JOM, et al. Strategies for communicating scientific evidence on healthcare to managers and the population: a scoping review. Health Res Policy Syst 2023 Jul 10;21(1):71 [FREE Full text] [doi: 10.1186/s12961-023-01017-2] [Medline: 37430348]
14. Winters M, Christie S, Melchinger H, Iddrisu I, Al Hassan H, Ewart E, et al. Debunking COVID-19 vaccine misinformation with an audio drama in Ghana, a randomized control trial. Sci Rep 2025 Mar 15;15(1):8955. [doi: 10.1038/s41598-025-92731-0] [Medline: 40089600]
15. Coursera. URL: https://www.coursera.org [accessed 2026-04-18]
16. Qualtrics. URL: https://www.qualtrics.com [accessed 2026-04-18]
17. Regional classifications. UNICEF. 2023 Jun 07. URL: https://data.unicef.org/regionalclassifications/ [accessed 2026-06-11]

XSL•FO
RenderX

18. Winters M, Christie S, Melchinger H, Arias N, Lirman L, Thomson A, et al. Moral foundations messaging to improve vaccine attitudes: an online randomized experiment from Argentina. PLOS Glob Public Health 2024;4(11):e0003276 [FREE Full text] [doi: 10.1371/journal.pgph.0003276] [Medline: 39546472]
19. Winters M, Sochoń-Latuszek A, Nurzhynska A, Yoruk K, Kukuła K, Bahruddinov M, et al. "Vaccinating your child during an emergency is more important than ever": a randomised controlled trial on message framing among Ukrainian refugees in Poland, 2023. Euro Surveill 2024 Sep;29(39):2400159 [FREE Full text] [doi: 10.2807/1560-7917.ES.2024.29.39.2400159] [Medline: 39328159]
20. Hansen RK, Baiju N, Gabarron E. Social media as an effective provider of quality-assured and accurate information to increase vaccine rates: systematic review. J Med Internet Res 2023 Dec 26;25:e50276 [FREE Full text] [doi: 10.2196/50276] [Medline: 38147375]
21. Swire-Thompson B, Miklaucic N, Wihbey JP, Lazer D, DeGutis J. The backfire effect after correcting misinformation is strongly associated with reliability. J Exp Psychol Gen 2022 Jul;151(7):1655-1665 [FREE Full text] [doi: 10.1037/xge0001131] [Medline: 35130012]
22. Ruggeri K, Vanderslott S, Yamada Y, Argyris YA, Većkalov B, Boggio PS, et al. Behavioural interventions to reduce vaccine hesitancy driven by misinformation on social media. BMJ 2024 Jan 16;384:e076542 [FREE Full text] [doi: 10.1136/bmj-2023-076542] [Medline: 38228339]
23. Ferrari E, Ballegeer AM, Corrochano D, Fuertes M, Herrero Teijo´n P, Delgado Marti´n ML, et al. Improvement of attitudes and skills using a MOOC about the basic science of climate change. Humanit Soc Sci Commun 2024 Jun 08;11(1):739. [doi: 10.1057/s41599-024-03139-6]
24. Herbert M, Smith IM, Guest C. Transforming improvement training at scale with essential digital training skills. BMJ Open Qual 2025 Aug 18;14(3):e003224 [FREE Full text] [doi: 10.1136/bmjoq-2024-003224] [Medline: 40829885]
25. Huang H, Qi D. Is MOOC really effective? Exploring the outcomes of MOOC adoption and its influencing factors in a higher educational institution in China. PLoS One 2025;20(2):e0317701 [FREE Full text] [doi: 10.1371/journal.pone.0317701] [Medline: 39992931]
26. Digital media for health outcomes: evaluation study of a massive online open course. Open Science Framework. URL: https://osf.io/w2adn/ [accessed 2026-06-11]

Abbreviations

CDC: Centers for Disease Control and Prevention
DMHO: Digital Media for Health Outcomes
SBCC: social and behavioral change communication
UNICEF: United Nations' International Children's Emergency Fund
WHO: World Health Organization

Edited by A Stone; submitted 29.Sep.2025; peer-reviewed by L Reid-Chassiakos, R Ruparel, S Çinarli; comments to author 11.Feb.2026; revised version received 11.Apr.2026; accepted 13.Apr.2026; published 25.Jun.2026.

Please cite as:
Melchinger H, Christie S, Titus A, Cruz A, Lepage C, Foleno T, Thompson-Kuhn C, Abebe AT, Gweh GN, Bangure D, Maina S, Mwangi J, Thande P, Winters M, Malik AA, Kuzmanovic A, Abeyesekera S, Epperson S, Tom W, Holmes K, Thomson A, Omer SB
Digital Media for Health Outcomes: Evaluation Study of a Massive Online Open Course
J Med Internet Res 2026;28:e85016
URL: https://www.jmir.org/2026/1/e85016
doi:10.2196/85016
PMID:

©Hannah Melchinger, Sarah Christie, Abigail Titus, Alyssa Cruz, Chelsey Lepage, Tony Foleno, Colleen Thompson-Kuhn, Adam Terefe Abebe, G. Nekerwon Gweh, Donewell Bangure, Stephen Maina, Judy Mwangi, Peninah Thande, Maike Winters, Amyn A Malik, Aleksandra Kuzmanovic, Surangani Abeyesekera, Sara Epperson, Whittney Tom, Kylie Holmes, Angus Thomson, Saad B Omer. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Combined Internet-Based Cognitive Behavioral Therapy and Face-to-Face Physiotherapy in Primary Health Care for Chronic Widespread Pain: Randomized Controlled Trial

Anna Bergenheim[1,2], PhD; Maria EH Larsson[1,2], PhD, Prof Dr; Chan-Mei Ho-Henriksson[1,2,3], PhD; Anna Grimby-Ekman[4], PhD, Prof Dr; Anna Caroline Larsson[2,3], MSc; Marie Persson[2,3,4], PhD; Sandra Weineland[2,3,4,5], PhD; Lena Nordeman[1,2], PhD

[1]Department of Health and Rehabilitation, Sahlgrenska Academy, Institute of Neuroscience and Physiology, University of Gothenburg, Gothenburg, Sweden

[2]Research, Education, Development and Innovation, Primary Health Care, Region Västra Götaland, Sweden

[3]Närhälsan (Public Healthcare Provider), Region Västra Götaland, Sweden

[4]School of Public Health and Community Medicine, Sahlgrenska Academy, Institute of Medicine, University of Gothenburg, Gothenburg, Sweden

[5]Department of Psychology, University of Gothenburg, Gothenburg, Sweden

Corresponding Author:
Anna Bergenheim, PhD
Research, Education, Development and Innovation, Primary Health Care
Vänerparken 15, Vänersborg
Region Västra Götaland, 46235
Sweden
Phone: 46 725046956
Email: anna.bergenheim@gu.se

Abstract

Background: Finding successful treatments for chronic widespread pain (CWP) in primary care is challenging. Interventions addressing both stress and pain may yield synergistic effects. Internet-based cognitive behavioral therapy (iCBT) reduces stress-related pain responses, while physical activity enhances function and resilience. Combined, they may target core CWP mechanisms.

Objective: This study aimed to evaluate the effectiveness of therapist-guided iCBT for stress management combined with physiotherapist-guided physical activity, compared to stand-alone physical activity, on pain and associated symptoms in individuals with CWP.

Methods: Participants with CWP, aged 18-70 years, were recruited in this parallel, multicenter randomized controlled trial via social media in Sweden. The intervention group received a 14-module iCBT program plus a physical activity plan; the control group received the physical activity plan only. Primary outcomes were pain intensity and pain locations. Secondary outcomes included stress, fatigue, depression, and quality of life. Variables were self-assessed using paper questionnaires at baseline and 6 months. Allocation was concealed; participants and researchers were not blinded.

Results: Of 129 participants (64 intervention, 65 control), 82 (64%) completed the 6-month follow-up. In the intervention group, 37% (16/43) completed all 14 iCBT modules. No between-group differences were observed for change in any outcome: pain intensity (mean 1.7, 95% CI –7.5 to 11.0), pain locations (mean –0.8, 95% CI –2.2 to 0.6), Stress and Crisis Inventory (mean 2.0, 95% CI –3.3 to 7.4), Fibromyalgia Impact Questionnaire (mean 2.7, 95% CI –3.7 to 9.2), global fatigue (mean 3.0, 95% CI –7.2 to 13.3), Multidimensional Fatigue Inventory (general fatigue: mean 0.1, 95% CI –1.1 to 1.3; physical: mean 0.4, 95% CI –1.1 to 1.9; mental: mean 0.4, 95% CI –1.1 to 2.0; reduced activity: mean –2.8, 95% CI –0.5 to 2.1; reduced motivation mean –0.5, 95% CI –2.2 to 1.2), Hospital Anxiety and Depression Scale (anxiety: mean –0.8, 95% CI –2.3 to 0.7; depression: mean –0.2, 95% CI –1.7 to 1.3), and Short-Form 36 subscales (physical function: mean 2.9, 95% CI –3.7 to 9.5; role physical: mean –13.1, 95% CI –28.3 to 2.0; role emotional: mean –4.0, 95% CI –22.8 to 14.8; energy/fatigue: mean 3.4, 95% CI –4.0 to 10.8; emotional well-being: mean –0.5, 95% CI –8.2 to 7.3; social functioning: mean –4.6, 95% CI –15.0 to 5.8; pain: mean –3.1, 95% CI –9.8 to 3.6; general health: mean –2.2, 95% CI –9.2 to 4.9). Both groups improved across several outcomes. Main goals were attained by 37% (17/46) in the intervention group vs 19% (8/42) in controls (P=.02), and intermediate goals by 54% (25/46) vs 36% (15/42; P=.01).

XSL•FO
RenderX

Conclusions: This study novelly examines stress-targeting interventions for pain. Clinicians could focus on tailored physical activity plans while considering optional stress management to support behavioral change and goal attainment. Due to high loss to follow-up, results should be interpreted cautiously.

Trial Registration: ClinicalTrials.gov NCT04624139; https://clinicaltrials.gov/study/NCT04624139

(J Med Internet Res 2026;28:e86792) doi:10.2196/86792

KEYWORDS

chronic pain; cognitive behavioral therapy; exercise; internet; physiotherapy

Introduction

Chronic widespread pain (CWP) is pain that persists for more than 3 months, occurring above and below the waist, affecting both the right and left sides of the body, and involving axial pain [1]. The prevalence of CWP is estimated to be between 8% and 11% in the western world, being about twice as common among women as men [2]. Most patients with CWP are managed in primary health care [3], and in addition to negatively impacting the individual's quality of life, CWP imposes a significant economic burden on society [4]. CWP is associated with stiffness, severe fatigue, cognitive difficulties, and psychological distress [5]. The etiology of CWP is not fully understood, but previous research describes various physiological abnormalities, including central sensitization and impaired pain inhibition [6]. Additionally, heightened pain sensitivity in chronic pain populations is linked to dysregulation of the autonomic nervous system and the hypothalamic-pituitary-adrenal axis, although evidence is limited [7]. In women with fibromyalgia, acute stress has been found to be a contributor to momentary pain, while acute pain did not predict momentary stress in the same way [8]. Several demographic, psychosocial, and health-related factors, including female sex, age, family history of pain, depressed mood, widespread pain distribution, poor sleep, obesity, and chronic disease, have been suggested to increase the risk of both progression to and persistence of CWP [9,10]. Additionally, CWP is linked to low physical activity, which has been found to be associated with higher levels of perceived stress [11]. A previous study, examining the long-term symptom changes and predictors of improvement in women with CWP, indicated that a lower degree of stress symptoms at baseline predicted a substantial reduction in pain intensity after 12 years [12]. This finding further strengthens the evidence for an association between stress and pain and highlights the importance of underlying stressors that may trigger, perpetuate, or exacerbate pain.

Fibromyalgia constitutes a subgroup of CWP with more severe symptoms, and although there are criteria for diagnosis, there is not a distinct line for when CWP could be considered as fibromyalgia, since they both lie within the spectrum of nociplastic pain [13], and similar treatment strategies can be applied. European guidelines for treating CWP and fibromyalgia recommend nonpharmacological treatments primarily, such as physical exercise and cognitive behavioral therapy (CBT) [14]. Physical activity contributes to positive outcomes for both stress-related illnesses and chronic pain [15,16]. Various forms of physical exercise show positive effects on symptoms and overall health in individuals with CWP and fibromyalgia [15], and evidence suggests that physical activity could have a positive influence on the autonomic nervous system [17]. Consideration of patients' personal preferences and available resources may help guide the selection of appropriate exercise interventions. Therapist-guided internet-based CBT (iCBT) has demonstrated positive effects for patients with chronic pain [18] and with stress-related illness [19,20]. Combined, CBT and physical exercise have been found to yield small to moderate improvements in pain intensity, functional disability, mental health, and quality of life in patients with chronic pain compared with active control treatments [21]. However, CBT interventions in previous studies have been heterogeneous in format, content, and provider type [21]. Moreover, despite growing evidence linking stress to pain modulation in CWP [7], few interventions have explicitly targeted stress reduction as a primary mechanism. Likewise, although physical activity is recommended, little is known about the added value of combining structured physical activity with stress-focused iCBT in individuals with CWP.

ICBT is an accessible form of CBT that has demonstrated effects comparable to those of face-to-face CBT in populations with primarily mental health disorders [22]. Stress reduction through iCBT could enhance pain modulation by reducing physiological stress responses, while person-centered physical activity can further improve physical function and resilience against stress. The combination of these 2 evidence-based strategies is therefore of interest, as it may address both the psychological and physiological mechanisms that maintain CWP. The novelty of this project is that the iCBT intervention targets stress management, and not primarily pain management.

This study aimed to evaluate the effectiveness of a therapist-guided iCBT program for stress management, combined with a face-to-face physiotherapist-guided physical activity plan, compared to a control group receiving the physical activity plan as a stand-alone treatment. It was hypothesized that the combined intervention would lead to greater improvements in pain, stress-related symptoms, psychological distress, and physical activity in individuals with CWP.

Methods

Trial Design

Overview

This was a multicenter randomized controlled trial (RCT) with parallel groups, conducted at 4 sites in southwestern Sweden and online, stratified by sex and total score on the Stress and Crisis Inventory (SCI)-93; <50 or ≥50, allocation ratio 1:1. The

XSL•FO
RenderX

study is reported in accordance with the Consolidated Standards of Reporting Trials (CONSORT) 2025 guidelines for reporting RCTs [23] (Multimedia Appendix 1), and the Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth (CONSORT-EHEALTH) guidelines [24] (Multimedia Appendix 2).

Changes to Trial Protocol

The actual study start date for participant enrollment was November 4, 2020, which also corresponded to the originally estimated start date. However, this date was not formally updated from "Anticipated" to "Actual" in the trial registry until April 26, 2023, due to an administrative oversight. Recruitment was stopped after the inclusion of 129 participants. Recruitment was substantially affected by the COVID-19 pandemic, which slowed the inclusion process and prolonged the recruitment period. After approximately 2 years of recruitment, it was decided to stop further inclusion to ensure completion of the study within the available funding period and project timeline. This study focuses explicitly on the baseline and 6-month follow-up data. The 12-month follow-up data were not included in the primary analyses because the attrition rate at 12 months was substantial (66/129, 51%), which was considered to limit the validity and interpretability of analyses of intervention effects. The 12-month data are instead reserved for separate, exploratory longitudinal trajectory substudies. Due to the high attrition observed at the 12-month mark, the planned 24-month follow-up was deemed statistically unfeasible and was not conducted; the trial was formally concluded after the completion of the 12-month follow-up assessments. The Short Form-36 (SF-36) version 1 [25] was used instead of Research and Development-36 [26] due to a miscommunication during study implementation. As the questions in the SF-36 version 1 are identical to the Research and Development-36 in content and assess the same domains of health-related quality of life, this change was not considered to affect the study outcomes. The Fibromyalgia Impact Questionnaire (FIQ) [27,28] total score was added after publication of the study protocol to provide a broader measure of overall disease impact and health status, thereby enabling a more comprehensive assessment of the burden of CWP. Additional analyses not mentioned in the study protocol were conducted to investigate differences between groups in participant goal attainment and achievement of the World Health Organization (WHO) recommendations for physical activity [29].

Trial Setting

The physiotherapy intervention was conducted face-to-face at primary care rehabilitation centers in 3 cities in western Sweden. The iCBT was delivered online.

Patient and Public Involvement

There was no patient or public involvement in the design, conduct, or reporting of the trial.

Eligibility Criteria

Women and men aged 18-70 years who met the American College of Rheumatology (ACR) 1990 criteria for CWP—pain persisting for more than 3 months, occurring in areas both above and below the waist, on both the right and left sides of the body, and including axial pain [1]—were included in the study. In Sweden, fibromyalgia is clinically diagnosed according to the ACR 1990 criteria.

Participants were excluded for the following reasons: serious physical or psychological disease or other conditions with restrictions on physical activity, not having a smartphone or computer, ongoing psychotherapy or physiotherapy, and insufficient knowledge of the Swedish language for completing the questionnaires and participating in the iCBT. Participants who had a score of >14 on the subscale for anxiety or depression in the Hospital Anxiety and Depression Scale (HADS) were contacted before randomization by a psychologist by telephone for an assessment about the severity of their psychological distress, and if needed, they could be referred to a relevant service within primary health care.

Recruitment

People with CWP were recruited via social media advertisements on Facebook and Instagram in western Sweden between November 2020 and October 2022. They responded to the advertisement on social media by sending an email to the research project inbox. Research assistant 1 sent written study information via email. Individuals who expressed interest after reading the information were screened for eligibility by research assistant 1 during a telephone call. If the person was eligible, the written consent form, along with a battery of paper-based baseline questionnaires, was sent home by mail to the participant, and the participant returned the forms by mail.

Intervention

Overview

The description of the intervention follows the TIDieR (Template for Intervention Description and Replication) checklist [30] (Multimedia Appendix 3). The combined intervention of therapist-guided iCBT for stress and face-to-face physiotherapist-guided physical activity plan was designed to affect both psychological and physiological aspects of stress and pain. The intervention is grounded in proposed biopsychosocial mechanisms of action, suggesting that CWP is partly maintained by stress-related physiological dysregulation and maladaptive behavioral patterns [31]. Accordingly, the combination of stress-focused iCBT and structured physical activity targets key biological, psychological, and behavioral mechanisms underlying pain. The iCBT was delivered online and guided by a psychologist. The physiotherapy was conducted face-to-face at 4 rehabilitation clinics in primary health care in Region Västra Götaland in western Sweden. The iCBT and the physiotherapy were delivered in parallel with different starting points and were not integrated. The starting point was at baseline for the physical activity plan and at 8 weeks for the iCBT program. The starting points could differ by up to 2 weeks. The time points for the intervention components are illustrated in Figure 1. No modifications to the interventions were made during the study period. Participation in the intervention and access to the iCBT program were free of charge, and the participants received no compensation.

XSL•FO
RenderX

Figure 1. Illustration of the starting points and duration of internet-based cognitive behavioral therapy (iCBT) and face-to-face physical activity plan, delivered in parallel in the combined intervention. A randomized controlled trial in Swedish primary health care, including participants with chronic widespread pain.

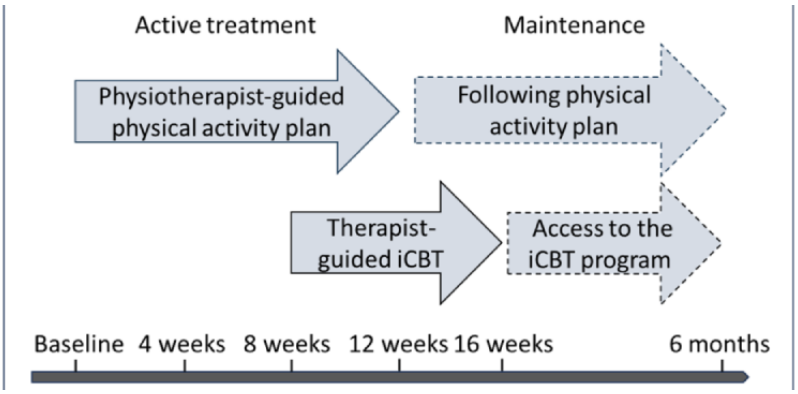

Therapist-Guided iCBT Designed for Stress-Related Illness

The iCBT program used in this study is a stress management program. A previous version of the program has been used and was well accepted among patients with stress-related disorders, resulting in improvements in perceived stress, anxiety, and depression, although it was not superior to a waitlist control group [32]. The program has since been updated, but no version of the stress management program has previously been evaluated in patients whose primary condition is chronic pain. The iCBT started about 8 weeks after baseline. The 8-week digital program was based on traditional CBT techniques, driven by experiential learning and with a focus on increasing psychological flexibility related to feelings. The iCBT program was developed and provided by Livanda AB. The program consisted of 14 modules addressing the following core processes: behavioral activation and exposure, cognitive restructuring, physiological regulation, mindfulness, acceptance, skills training and problem-solving, as well as maintenance and relapse prevention. Together, these aimed to increase psychological flexibility, reduce avoidance, and improve stress management. The contents of the modules are described in Multimedia Appendix 4. The participants were encouraged to complete 2 modules per week. The program guided participants through understanding stress, setting personal goals, practicing relaxation, and developing long-term coping strategies. Each module is built on the previous one, providing practical tools and exercises to reduce stress and support lasting behavioral change. Therapist support was delivered throughout the iCBT program intervention by 2 psychologists. The psychologists were licensed, trained in CBT, and had completed a 1.5-day training in iCBT before starting the study. The iCBT started with a telephone call with the psychologist. During the call, the iCBT treatment was introduced, explained, and motivated for the participant. The psychologists tracked progress, addressed patient queries, and reached out if engagement declined. Brief telephone check-ins were also conducted at mid- and posttreatment. If the participant followed the time plan for the program, it was expected to be completed after 8 weeks. However, they had access to the program for 6 months and could access the platform at any time and proceed at their own pace. The program included written texts, exercise videos, and audio files.

Face-to-Face Physiotherapist-Guided Physical Activity Plan

Participants attended 3 individual 30-minute face-to-face visits with a physiotherapist at baseline, 4 weeks, and 12 weeks, during which a physical activity plan was created, reviewed, and progressively adjusted. The first visit was conducted face-to-face within 2 weeks after baseline. In visit 1 (baseline), the participant and the physiotherapist cocreated an individualized plan of physical activity, according to person-centered principles [33], to be performed at home. The physical activity plan has been used and evaluated in a previous RCT in patients with CWP [34-36]. In this study, the physical activity plan aimed to enable participants to meet the WHO recommendations for weekly physical activity associated with optimal health outcomes: 150-300 minutes of moderate intensity or 75-150 minutes of vigorous intensity physical activity per week, in combination with regular strengthening exercises [29]. The individualized physical activity plan comprised at least 2 components: aerobic exercise and resistance exercise. Aerobic exercise of moderate intensity, aiming toward a total amount of time of 150 minutes a week on moderate intensity or 75 minutes a week on vigorous intensity (or a combination of the two). The resistance exercise program was individualized and performed 2 times a week at a minimum. The exercises started on an individual level with increasing duration and intensity successively over 12 weeks. The physical activity plan also included individual goals that participants aimed to achieve, comprising both main goals and intermediate goals. The main goals referred to overall, long-term outcomes the participant sought through rehabilitation—such as returning to work, improving daily functioning, performing specific tasks important to them, or reducing pain. Intermediate goals were shorter-term, step-by-step objectives that supported progress toward the main goals. These might include improving physical strength, learning pain management techniques, or increasing activity levels. Visit 2 (4 weeks) included follow-up, revision, and progression of the physical activity plan. Visit 3 (12 weeks) included follow-up of attainment of the individual goals and follow-up, revision, and progression of the physical activity plan. The participants were encouraged to continue to follow their plan autonomously.

A total of 9 physiotherapists at the 4 study sites delivered the physiotherapy intervention. All physiotherapists had experience working with chronic pain and stress-related disorders and

XSL•FO
RenderX

cocreating physical activity plans with patients in primary health care. The physiotherapists received both oral and written standardized instructions from the principal investigator, outlining the content of each visit and the procedures for documentation.

Control Group

The control group received the same physiotherapy intervention as the intervention group, but as a stand-alone treatment. It consisted of 3 individual visits: at baseline, 4 weeks, and 12 weeks, during which a physical activity plan was created, revised, and progressively adjusted.

Fidelity and Adherence to the Intervention

Fidelity was ensured by delivering the iCBT content identically to all participants through the digital platform, while the physical activity plan was implemented according to a standardized template. Adherence to the core components of the intervention was assessed based on the number of completed modules and attendance at all 3 physiotherapist visits.

Data Collection

The participants completed a battery of paper-based questionnaires at home at baseline and at 6-month follow-up. Follow-up at 6 months was chosen since the participants had access to the iCBT program for 6 months. The questionnaires were sent by mail, with reminders sent by email. The participants completed the same battery of questionnaires after 6 months. They received and returned the questionnaires by mail.

Primary Outcomes

Pain Intensity

Pain intensity assessed with a visual analog scale (VAS; 0-100 mm) included as a subscale in the FIQ. The participants rated their pain intensity during the previous week [28]. The FIQ is a brief, self-administered instrument designed to evaluate the current health status of women with fibromyalgia syndrome. It consists of 10 items that assess various aspects of health and daily functioning, including physical functioning, work status, depression, anxiety, sleep, pain, stiffness, fatigue, and overall well-being. The FIQ was developed and validated for patients with fibromyalgia and CWP [27].

Number of Pain Locations (0-18)

The participants marked their painful areas on a pain drawing [37] with 18 predefined body areas.

Secondary Outcomes

Symptoms of Stress (0-140)

SCI-93 consists of 35 items assessing clinical manifestations of stress ranging from 0 (not at all) to 4 (very much). The questions include both physical and mental experiences. The total score ranges from 0 to 140, and higher scores indicate a higher level of stress [38]. The following cutoff values have been recommended [39]: (1) 0-25 points equal a normal stress reaction, (2) 26-50 points indicate a mild stress reaction, (3) 51-75 points represent a moderate stress reaction, (4) 75-100 points demonstrate a severe stress reaction, (5) 101-140 points suggest a very severe stress reaction. The SCI-93 has shown satisfactory known-group validity and test-retest reliability in women with fibromyalgia and CWP [38].

Physical Activity (Hours)

Level of physical activity was measured with the Leisure Time Physical Activity Instrument, a questionnaire assessing the amount of physical activity during a typical week. The total score is the number of hours of physical activity [40]. The Leisure Time Physical Activity Instrument has shown satisfactory construct validity and test-retest reliability in women with fibromyalgia [40].

Overall Health Status

The FIQ total score provides a comprehensive overview of the impact of fibromyalgia on a patient's life and is widely used in both clinical practice and research. The total score ranges from 0 to 100, and a higher score indicates more severe symptoms and impairment [27].

Fatigue

The Multidimensional Fatigue Inventory (MFI)-20 consists of 20 statements rated on a 5-point Likert scale, assessing various aspects of fatigue experienced over recent days. It produces 5 subscale scores: general fatigue, physical fatigue, mental fatigue, reduced motivation, and reduced activity. Each subscale score ranges from 4 to 20, with higher scores indicating greater levels of fatigue [41]. The MFI-20 has demonstrated satisfactory convergent construct validity and test-retest reliability in individuals with chronic pain conditions [42]. The FIQ subscale for fatigue (VAS 0-100 mm) was also used as a global measure of fatigue during the last week [28].

Health-Related Quality of Life

Health-related quality of life was assessed with the SF-36 version 1, a generic instrument that consists of 8 subscales ranging from 0 to 100. A higher score indicates better health-related quality of life [25].

Anxiety and Depression

Anxiety and depression were assessed with the HADS, which contains 2 subscales for symptoms of anxiety and depression, ranging from 0 to 21; a higher value means a higher degree of anxiety or depression [43]. A cutoff value of 8 has been suggested to indicate possible anxiety or depression, and a score of ≥15 indicates severe anxiety or depression [44].

Participant's Impression of Change

The participant's impression of change was assessed at the 6-month follow-up with the questionnaire Patient Global Impression of Change (PGIC). The PGIC consists of a single item stating "Since the beginning of the study, my symptoms overall have...." Participants rate this on a 6-point scale ranging from 1 (very much improved) to 6 (very much worsened) [45].

Additional Outcomes

The additional outcomes were collected during the 12-week visit with the physiotherapist and thus not included as outcomes at the 6-month follow-up in the study protocol at ClinicalTrials.gov.

XSL•FO
RenderX

Individual Goals in the Physical Activity Plan

At visit 1 with the physiotherapist, the participant set individual goals, both intermediate goals and main goals that the participant wished to achieve. At the 12-week visit, the participants rated whether they had reached their individual goals completely, partially, or not at all.

WHO Recommendations for Physical Activity

The physiotherapist assessed during the third visit whether the participant met the WHO recommendations for physical activity [29]. Based on standardized questions, the physiotherapist recorded whether the participant met (1) the guideline of performing strength training for the major muscle groups at least twice per week, and (2) the guideline of performing aerobic physical activity for at least 150 minutes per week at moderate intensity, or alternatively 75 minutes per week at vigorous intensity. Each item was rated by the physiotherapist as "Not at all," "Partly," or "Completely" in relation to the guideline.

Descriptive Data

Descriptive data were also collected, such as duration of CWP, marital status, and country of birth. Work and sick leave are presented as the current percentage of full-time work (full-time=40 hours of work per week) divided into 4 groups: 0%, 1%-49%, 50%-79%, and 80%-100%.

Harms

Information about possible adverse or negative effects was not collected systematically. Physical activity often leads to an initial increase in pain intensity for patients with CWP, and adjusting the intensity of activities is a routine part of the follow-up in the physical activity plan.

Sample Size

The power calculation was based on the primary outcome variable, pain intensity, which was assessed using a VAS (0-100 mm) [28]. Many factors impact what could be considered a clinically relevant difference in pain intensity between groups, and a suggestion is 20 mm on VAS (16) and a SD of 40-45, which is applied in the power calculation in this study. Statistical significance is P=.05 and power=80%. To be able to make analyses of subgroups based on the SCI-93 score, the number of patients was calculated to be 136. To be able to use nonparametric statistics, and to adjust for possible confounders as well as consider 20% dropout at the follow-up, the final sample size was calculated to be 200 participants.

Randomization

The randomization was carried out by research assistant 2 and performed with block randomization with blocks of 4, with sealed, opaque, sequentially numbered envelopes prepared by a statistician. The randomization was stratified equally by sex and total score on the questionnaire SCI-93 (total score <50 or ≥50). The stratification by level of stress symptoms was conducted to create groups that were as comparable as possible with respect to stress symptom severity, as measured by the SCI-93. This was done in order to enable subgroup analyses of whether the effect of the intervention differed depending on the level of stress symptoms. Neither the research team nor any of the research assistants had access to the random allocation sequence. Participants and researchers were not blinded in this study. The physiotherapists who carried out the intervention in both groups did not know which group the participants were allocated to, if the participant did not mention it themselves.

Statistical Analyses

The statistical program SPSS (version 28.0.1.1; IBM Corp) was used for data calculations. Descriptive data are presented as mean, median, SD, and range for continuous data and as number and percent for categorical data. The participants were analyzed in trial groups in a randomized trial. Comparisons between groups of change in outcomes were conducted using an independent t test. Within-group changes were calculated using a paired t test. The results for pain intensity are presented both on group level as change in pain intensity over time between the groups and on an individual level, divided into the proportion of participants in each group achieving moderate (≥30%) and major (≥50%) change in pain intensity [46], with differences between groups calculated using the Pearson chi-square test. Effect sizes representing between-group differences in change from baseline to 6-month follow-up were calculated using Cohen d [47]. The primary outcomes, pain intensity and pain locations, were analyzed using linear mixed effects models with repeated measures over time (baseline and 6-month follow-up). Fixed effects included group (intervention vs control), time, and the group-by-time interaction. A random intercept for participants was included to account for within-subject correlation. Potential confounders (sex, age, and pain duration) were first examined by adding each variable separately to model 1. Variables that were significantly associated with pain intensity/pain locations (P<.05) were subsequently included in the adjusted model. The final model (Model 2) included group, time, the group-by-time interaction, and significant confounders. Model assumptions were checked by inspection of residuals. Results are presented as P values, and statistical significance was set at P<.05. Additional analyses of between-group differences at the 12-week visit were conducted for achievement of individual goals using the Mann-Whitney U test, and adherence to WHO recommendations for physical activity and muscle-strengthening exercise was assessed using the Pearson chi-square test. All tests were 2-sided, and statistical significance was set at P<.05. For the subscales in the MFI-20 and SF-36, missing values were handled by replacing them with the mean of the completed items within that subscale, provided that at least half of the items in the subscale had been completed. This procedure allowed a score for the subscale to be calculated. For the SCI-93, the same procedure was applied; up to 10% missing values were allowed to be replaced with the mean of the completed items. For the FIQ total score, the total was calculated as the mean of the subscale scores. If any subscale score was missing, the mean of the remaining subscales was used. A maximum of 3 out of 10 subscales were allowed to be missing to compute the total FIQ score. Missing data were evaluated using the Little missing completely at random (MCAR) test, including baseline and 6-month follow-up outcome measures, treatment group, and selected demographic variables (age, sex, education, and duration of pain). No interim analyses and stopping guidelines were applied in this study.

XSL•FO
RenderX

Ethical Considerations

The study was approved by the Swedish ethical review authority (registration number 2023-04031-0). Written informed consent was obtained from all participants on paper. The original informed consent covered the additional analyses, and no further consent was required. Participant privacy and confidentiality were ensured throughout the study. Data were deidentified prior to analysis, and all records were stored securely with access limited to authorized researchers. The study complied with relevant data protection regulations. The manuscript and supplementary materials do not contain any images in which individual participants can be identified. No financial or nonfinancial compensation was provided to participants for their participation in this study.

Results

Overview

A total of 9 participants who scored >14 were contacted before randomization by a psychologist by telephone for a dialogue about the severity of their psychological distress. For 1 participant, the psychologist and the participant jointly decided that inclusion in the study was not appropriate, due to the participant's mental health. Instead, the participant was referred to a relevant service within primary health care. The other 8 participants were included in the study. The recruitment process was delayed and obstructed by the COVID-19 pandemic, resulting in an interruption before the desired sample size of 200 participants could be reached. Data collection for baseline started in November 2020, and 6 months follow-up was completed in May 2023. A total of 129 participants were included and randomized in the study. A total of 82 (64%) participants completed the 6-month follow-up and were included in the analyses; 43 in the intervention group and 39 in the control group. The participants who were lost to follow-up did not return the questionnaires and did not give reasons for dropping out. The participant flow is described in Figure 2 [23].

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) 2025. Participant flow in the randomized controlled trial in Swedish primary health care, including participants with chronic widespread pain. iCBT: internet-based cognitive behavioral therapy.

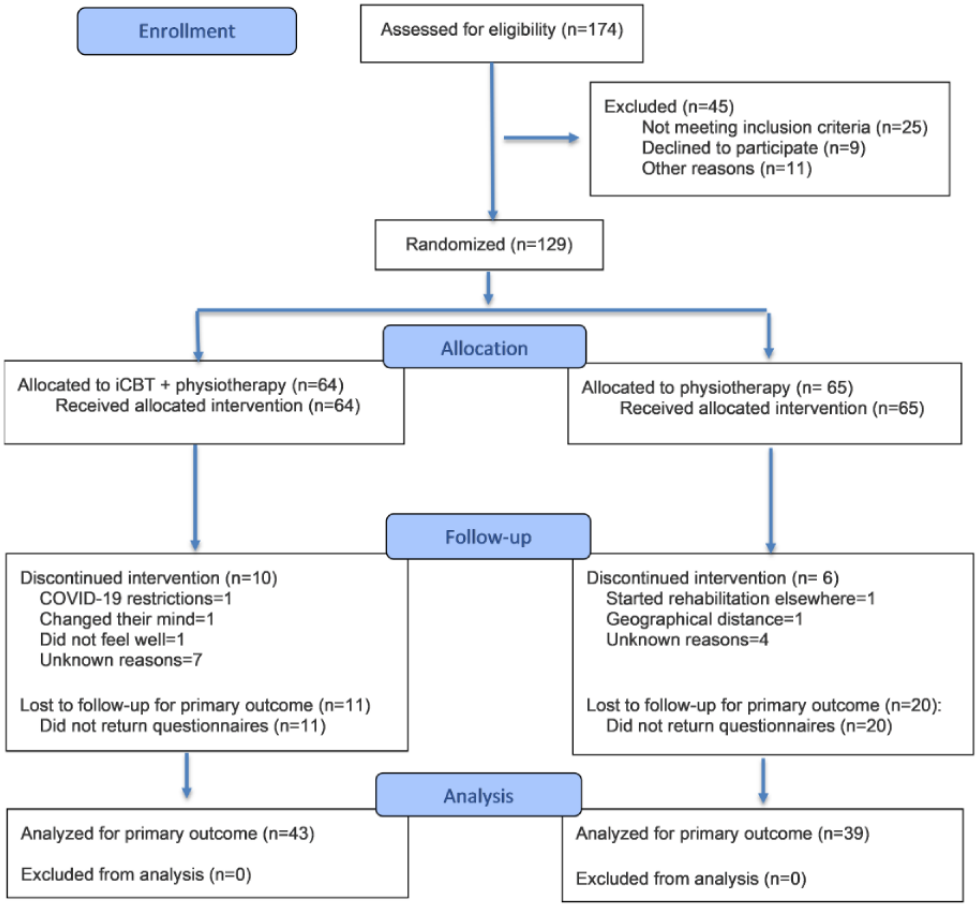

There were no statistically significant baseline differences in age, number of pain locations, pain duration, FIQ pain intensity, FIQ fatigue, SCI-93, level of education, work status, sick leave, or disability pension between the 82 who participated in the 6-month follow-up and the 47 who were lost to the 6-month follow-up.

Little MCAR test was not statistically significant (χ^2_{950}=1007.56; P=.10). Hence, the data could be MCAR, or the power was too low to detect a distinction from this (n=129). A statistically significant test would have indicated not MCAR.

Descriptive data of the participants are presented in Table 1. There were no statistically significant differences in descriptive

XSL•FO
RenderX

data between the intervention group and the control group at baseline.

Table 1. Baseline descriptive data for the participants with chronic widespread pain (n=129), in the randomized controlled trial in primary health care in Sweden, presented separately for the intervention group receiving internet-based cognitive behavioral therapy (iCBT) in combination with a face-to-face physical activity plan, and the control group receiving only the physical activity plan.

	iCBT + physical activity plan (n=64)	Physical activity plan (n=65)
Age (years), mean (SD; range)	51 (10; 30-70)	51 (12; 21-69)
Pain duration (years), mean (SD; range)	14 (9; 1.5-35)[a]	12 (11; 0.3-45)[a]
Number of pain locations, mean (SD; range)	10 (4; 3-18)	10 (4; 4-18)
FIQ[b] pain intensity (mm), mean (SD; range)	62 (18; 10-100)	63 (21; 18-100)
FIQ fatigue (mm), mean (SD; range)	69 (24; 13-100)	66 (25; 3-100)
SCI-93[c] (score), mean (SD; range)	53 (22.2; 15-108)[a]	50 (20.1; 19-111)[a]
Female, n (%)	59 (92)	60 (92)
Education, n (%)		
≤9 years	0 (0)	1 (2)
10-12 years	15 (23)	15 (23)
>12 years	49 (77)	49 (75)
Work, n (%)		
0%	17 (27)	18 (28)
1%-49%	3 (5)	7 (11)
50%-79%	10 (15)	4 (6)
80%-100%	34 (53)	36 (55)
Sick leave, n (%)		
0%	54 (84)	57 (88)
1%-49%	2 (3)	1 (2)
50%-79%	3 (5)	1 (2)
80%-100%	5 (8)	6 (9)
Disability pension, n (%)		
0%	59 (92)	60 (92)
1%-49%	0 (0)	1 (2)
50%-79%	2 (3)	0 (0)
80%-100%	3 (5)	4 (6)

[a]Data were available for 62 or 63 participants due to missing data.

[b]FIQ: Fibromyalgia Impact Questionnaire.

[c]SCI-93: Stress and Crisis Inventory-93.

Adherence to the Intervention

After 6 months, 16 (37%) participants of the 43 in the intervention group had completed all 14 modules in the iCBT program, 19 (42%) participants had completed between 7 and 13 modules, and 8 (19%) participants had completed 0-6 modules. A total of 88 participants (46 in the intervention group and 42 in the control group) completed all 3 physiotherapy visits (baseline, 8 weeks, and 12 weeks).

Primary Outcomes

Pain Intensity

No statistically significant difference between the groups was found for change in pain intensity after 6 months. Within-group improvements in pain intensity were found in both groups (Table 2).

The mixed effects model analysis of pain intensity from baseline to 6-month follow-up is presented in Figure 3A. When potential confounders were examined individually, sex and pain duration were associated with pain intensity and were therefore included in the adjusted final model. There was a significant main effect of time, indicating a reduction in pain intensity over the study

XSL•FO
RenderX

period. No statistically significant main effect of group was observed in any of the models. The group-by-time interaction was not statistically significant, suggesting that changes in pain intensity over time did not differ between the groups. A total of 9 (21%) participants in the intervention group and 13 (33%) participants in the control group (*P*=.79) improved at least 30% in pain intensity, which indicates a moderate improvement. Of these, 6 (14%) participants in the intervention group, and 7 (18%) in the control group (*P*=.89) improved at least 50% in pain intensity, which indicates a major improvement [46].

XSL•FO
RenderX

Table 2. Outcome values at baseline and at 6-month follow-up for the intervention group receiving internet-based cognitive behavioral therapy (iCBT) in combination with a face-to-face physical activity plan, and the control group receiving only the physical activity plan. *P* values for within-group changes and between-group differences, and Cohen d with 95% CIs for change from baseline to 6-month follow-up in a randomized controlled trial of participants with chronic widespread pain in primary health care in Sweden. Paired *t* test was used for within-group analyses and independent *t* test for between-group analyses.

	iCBT + physical activity plan (n=43)			Physical activity plan (n=39)			Between-group difference in change, mean (95% CI)	Between-group, *P* value	Cohen *d* (95% CI)
	Baseline, mean (SD)	6 months follow-up, mean (SD)	Within group, *P* value	Baseline, mean (SD)	6 months follow-up, mean (SD)	Within group, *P* value			
FIQ[a] pain intensity (mm)	62.6 (17.7)	54.9 (23.8)	.04	62.8 (18.0)	53.5 (27.1)	.007	1.7 (–7.5 to 11.0)	.49	0.08 (–0.35 to 0.52)
Pain locations (n)	9.6 (3.6)	7.5 (3.5)	<.001	9.6 (4.1)	8.2 (4.5)	.01	–0.8 (–2.2 to 0.6)	.90	–0.24 (–0.67 to 0.20)
SCI-93[b] (score)	51.4 (22.2)	46.2 (25.0)	.01[c]	47.2 (17.7)	39.4 (19.9)	<.001[d]	2.0 (–3.3 to 7.4)	.24	0.17 (–0.27 to 0.62)
FIQ total (score)	52.6 (17.4)	50.4 (21.9)	.31	50.9 (16.0)	45.9 (21.8)	.04	2.7 (–3.7 to 9.2)	.79	0.19 (–0.25 to 0 to 62)
FIQ Global fatigue (mm)	68.4 (25.0)	69.0 (27.3)	.85	63.2 (25.9)	61.2 (30.0)	.64	3.0 (–7.2 to 13.3)	.82	0.13 (–0.30 to 0.57)
MFI[e] general fatigue (score)	16.9 (2.8)	15.7 (4.3)	.004	16.2 (3.5)	15.0 (4.5)	.02	0.1 (–1.1 to 1.3)	.34	0.03 (–0.40 to 0.46)
MFI physical fatigue (score)	15.6 (3.8)	14.8 (4.7)	.16	16.2 (3.5)	14.9 (4.3)	.02	0.4 (–1.1 to 1.9)	.46	0.12 (–0.31 to 0.56)
MFI mental fatigue (score)	13.7 (3.7)	12.9 (3.8)	.26[c]	12.6 (4.6)	11.4 (4.5)	.04	0.4 (–1.1 to 2.0)	.34	0.13 (–0.32 to 0.57)
MFI reduced activity (score)	13.8 (3.5)	12.6 (4.2)	.03	13.5 (3.9)	13.1 (4.1)	.33	–2.8 (–0.5 to 2.1)	.06	–0.27 (–0.70 to 0.17)
MFI reduced motivation (score)	10.9 (3.8)	10.4 (3.9)	.45	9.6 (3.6)	9.7 (4.0)	.94	–0.5 (–2.2 to 1.2)	.45	–0.12 (–0.56 to 0.31)
HADS[f] depression (score)	7.3 (3.8)	7.0 (4.4)	.64	5.8 (3.4)	6.4 (3.7)	.89	–0.2 (–1.7 to 1.3)	.87	–0.05 (–0.49 to 0.38)
HADS anxiety (score)	8.1 (4.4)	6.3 (4.7)	.004	7.3 (4.6)	5.8 (3.9)	.09	–0.8 (–2.3 to 0.7)	.36	–0.23 (–0.66 to 0.21)
SF-36[g] physical function (score)	66.2 (18.2)	71.2 (17.2)	.04[c]	66.8 (17.7)	67.7 (19.2)	.38[d]	2.9 (–3.7 to 9.5)	.82	0.20 (–0.25 to 0.64)
SF-36 role physical (score)	25.0 (31.7)	30.8 (37.7)	.23	19.2 (27.2)	37.8 (41.3)	<.001	–13.1 (–28.3 to 2.0)	.88	–0.39 (–0.84 to 0.06)
SF-36 role emotional (score)	54.4 (37.3)	59.7 (42.1)	.54[c]	55.6 (42.2)	64.1 (40.0)	.18	–4.0 (–22.8 to 14.8)	.36	–0.10 (–0.54 to 0.35)
SF-36 energy/fatigue (score)	29.3 (20.6)	36.3 (23.8)	.004	35.3 (21.0)	38.8 (27.8)	.24	3.4 (–4.0 to 10.8)	.19	0.20 (–0.23 to 0.64)
SF-36 emotional well-being (score)	63.0 (17.3)	64.4 (23.3)	.63	64.0 (18.6)	65.8 (19.4)	.48	–0.5 (–8.2 to 7.3)	.20	–0.03 (–0.46 to 0 to 41)
SF-36 social functioning (score)	56.4 (25.0)	57.8 (22.8)	.63[c]	57.9 (30.3)	63.6 (23.9)	.08[d]	–4.6 (–15.0 to 5.8)	.53	–0.20 (–0.64 to 0.25)
SF-36 pain (score)	35.7 (14.1)	42.2 (18.3)	.008	30.5 (13.4)	39.0 (21.0)	.002[d]	–3.1 (–9.8 to 3.6)	.45	–0.21 (–0.65 to 0 to 25)
SF-36 general health (score)	44.8 (18.2)	45.1 (22.9)	.67	42.6 (20.8)	45.5 (25.0)	.26	–2.2 (–9.2 to 4.9)	.56	–0.14 (–0.57 to 0.30)
LTPAI[h] total (hours)	6.0 (3.3)	7.0 (4.5)	.11	7.9 (5.9)	8.0 (5.4)	.89	0.9 (–1.3 to 3.1)	.07	0.18 (–0.26 to 0.61)

[a]FIQ: Fibromyalgia Impact Questionnaire.

XSL•FO
RenderX

[b]SCI-93: Stress and Crisis Inventory-93.

[c]Calculated for 39-42 participants due to missing data.

[d]Calculated for 36-38 participants due to missing data.

[e]MFI-20: Multidimensional Fatigue Inventory.

[f]HADS: Hospital Anxiety and Depression Scale.

[g]SF-36: Short-Form 36.

[h]LTPAI: Leisure Time Physical Activity Instrument.

Figure 3. Mixed effect model analysis of (A) pain intensity and (B) number of pain locations. Baseline to 6-month follow-up in a randomized controlled trial of participants with chronic widespread pain in primary health care in Sweden. N/A: not applicable.

(A)

Model 1 (step 1)		Model 1 with confounders (step 2)			Model 2 (step 3, final)
Group	.86	.85	.86	.61	.61
Time	<.001	<.001	<.001	.001	.001
Group x time	.68	.70	.68	.79	.82
Potential confounder		*P values for potential confounders when added one at a time*			
Sex		.017			.017
Age			0.96		N/A
Pain duration				.011	.015

(B)

Model 1 (step 1)		Model 1 with confounders (step 2)			Model 2 (step 3, final)
Group	.72	.73	.72	0.52	.72
Time	<.001	<.001	<.001	<.001	<.001
Group x time	.25	.25	.25	0.18	.25
Potential confounder		*P values for potential confounders when added one at a time*			
Sex		.35			N/A
Age			.95		N/A
Pain duration				.33	N/A

Pain Locations

No statistically significant difference between the groups was found for the change in the number of pain locations after 6 months. Within-group improvements in the number of pain locations were found in both groups (Table 2). The mixed effects model analysis of pain locations from baseline to 6-month follow-up is presented in Figure 3B. When potential confounders were examined individually, none was associated with pain locations. There was a significant main effect of time, indicating a reduction in the number of pain locations over the study period. No statistically significant main effect of group was observed in any of the models. The group-by-time interaction was not statistically significant, suggesting that changes in the number of pain locations over time did not differ between the groups.

Secondary Outcomes

There were no differences found between the intervention and the control groups for change in any of the secondary outcomes after 6 months. Both groups showed statistically significant within-group improvements in SCI-93, MFI general fatigue, HADS anxiety, and SF-36 pain (Table 2). There was no statistically significant difference between the groups in the participants' ratings on the PGIC after 6 months (intervention group: median 3.0, IQR 3.0-4.0; control group: median 3.5, IQR 3.0-4.0; P=.59).

Additional Outcomes

Achievement of individual main goals and intermediate goals in the physical activity plan and achievement of WHO recommendations for physical activity and strengthening exercises are presented in Table 3. The participants in the intervention group were more likely to have reached their main goals and their intermediate goals at the 12-week visit than the participants in the control group. No difference was found between the groups at the 12-week visit in achievement of the WHO recommendations of aerobic physical activity and strengthening exercise (Table 3).

XSL•FO
RenderX

Table 3. Achievement of individual goals and World Health Organization (WHO) recommendations for physical activity and strengthening exercise at the 12-week visit for the participants with chronic widespread pain in the randomized controlled trial in primary health care in Sweden. Data are based on the 88 participants who completed all 3 visits with the physiotherapist.

	iCBT[a] + physiotherapy (n=46), n (%)	Physiotherapy (n=42), n (%)	Comparison between groups, *P* value
Main goals			.02
Completely	17 (37)	8 (19)	
Partially	23 (50)	21 (50)	
Not at all	6 (13)	13 (31)	
Intermediate goals			.01
Completely	25 (54)	15 (36)	
Partially	17 (37)	13 (31)	
Not at all	4 (9)	14 (33)	
Achieving the WHO recommendations			.67
Yes	24 (52)	20 (48)	
No	22 (48)	22 (52)	

[a]iCBT: internet-based cognitive behavioral therapy.

Discussion

Principal Findings

This study hypothesized that the combined intervention of therapist-guided iCBT plus a face-to-face physiotherapist-guided person-centered physical activity plan would result in superior improvements in pain, stress-related symptoms, psychological distress, and physical activity among individuals with CWP than the stand-alone physical activity plan. The findings did not support this hypothesis, as no significant differences in improvement were observed between the groups for either the primary outcomes of pain intensity and pain locations or the secondary outcomes of associated symptoms and amount of physical activity. However, both groups demonstrated modest improvements across several outcomes.

These findings are partly in line with a systematic review investigating the effect of CBT on patients with combined pain and stress-related disorders, in which CBT was found to have no effect on pain, but rather on psychological symptoms [20]. The improvements observed in this study are also supported by a systematic review and meta-analysis of CBT combined with exercise in adults with chronic pain, which reported small to moderate improvements in pain intensity, functional disability, and psychological outcomes such as anxiety and self-efficacy [21]. In this study, the absence of additional effects of iCBT beyond a structured physical activity intervention may suggest that physical activity alone contributes substantially to improvements in both physical and psychological outcomes, potentially limiting the detectable added benefit of iCBT in this context. However, as this study may have been underpowered, a larger sample size might have revealed a between-group difference.

There was a statistically significant difference between the groups in achievement of individual goals at the 12-week visit with the physiotherapist, favoring the combined intervention. Similar findings have been reported in a previous study showing that a CBT-based pain management program may contribute to the attainment of personally meaningful goals among individuals with chronic pain [48]. In the intervention group, the iCBT program started about 8 weeks after the participant's first physiotherapist visit. The iCBT started later because it was considered too demanding to initiate both new physical activity habits and behavioral therapy at the same time. The participants got access to 2 new modules in the iCBT program per week. By the 12-week physiotherapist visit, the participants had access to approximately 8 modules of the total of 12 modules. Thus, the iCBT program could have contributed to improvements or behavioral changes among the participants in the intervention group that enabled them to achieve their individual goals. Approximately half of the participants in each group reached the WHO recommendations for physical activity [29] at the third visit. The physiotherapist-guided person-centered physical activity plan may have contributed to the observed improvements, aligning with previous research on patients with CWP [35]. However, it remains difficult to determine with certainty whether the improvements were attributable to the activity plan, since cohorts with chronic pain have previously shown small symptom improvements over time, particularly in studies with long-term follow-up [12,49].

Limitations and Strengths

The main limitations of this RCT are that it was underpowered due to difficulties with participant recruitment during the COVID-19 pandemic and that the loss to follow-up was high, which affects the validity of the study results. The subgroup analyses based on the SCI-93 score were not performed since statistical power was not reached. One possible reason for the high loss to follow-up is that the questionnaires were completed on paper and had to be returned by mail, which may have required additional effort from the participants. Due to limited statistical power, the study may not have been able to detect potential differences in effects between 2 groups receiving active

XSL•FO
RenderX

treatments. This was also seen in a systematic review that found iCBT to be promising in ameliorating chronic pain and its associated symptoms when compared to passive control treatments, while it may not yield superior effects when compared to active treatments [18].

The follow-up assessment was conducted 6 months after baseline. This time frame was selected because behavioral change and the establishment of regular physical activity routines were expected to require time. However, it would also have been valuable to collect additional follow-up data at 16 weeks in order to assess outcomes immediately after the completion of the intervention. Another limitation is that the questions used to determine whether participants fulfilled the WHO recommendations for physical activity were not validated. In addition, these questions were only collected at the third physiotherapy visit. Consequently, adherence to the WHO recommendations at baseline could not be determined. While the physical activity plan, including the goal setting, has been used and evaluated in terms of content and patient experiences [34,36], the specific rating of goal achievement used in this study has not previously been applied, and its validity has not yet been established, which may affect the level of confidence in the findings.

Although all participants in this RCT fulfilled the inclusion criteria for CWP, they constituted a heterogeneous population. Pain duration varied from 4 months to 45 years, and pain spread ranged from 3 to 18 locations. The variance in symptom severity across participants could have affected the outcomes, as those with more severe symptoms and longer pain duration may have required greater support than what was provided in this intervention. There was also a large range in the study sample regarding the extent of stress-related symptoms as rated on the SCI-93. This was part of the study plan, since the research question concerned whether the participants could benefit from an iCBT program for stress regardless of their stress level. The variability within this CWP population is consistent with previous findings suggesting the existence of subgroups with differing symptom severities in fibromyalgia, a subgroup of CWP [50], thereby supporting the applicability of this study's results. A possible bias could be that the participants were aware that iCBT constituted the intervention of interest rather than the comparator intervention. However, as the intervention did not demonstrate superior effects, this potential expectancy bias may have had a limited impact on the outcomes.

Despite these limitations, this study has several strengths. The randomized controlled design reduces the risk of bias. The study was conducted in a real-world clinical setting in a patient population with complex symptoms, enhancing external validity. In addition, mostly validated outcome measures were used, and appropriate statistical methods were applied to account for repeated measures and potential confounders. The intervention in this study was designed to affect both psychological and physiological aspects of pain using iCBT for stress in combination with a person-centered plan for physical activity. After 6 months, 37% of the participants had completed all 12 modules in the iCBT program, which is in line with previous data from the same region in Sweden showing that 34% of patients participating in iCBT in primary health care completed all modules in an iCBT program [51], suggesting that adherence in this study reflects real-world implementation of iCBT in similar clinical settings. The CBT was chosen to be internet-based in this study since the participants were recruited from a large geographical area in western Sweden, involving many primary health care centers. iCBT could easily be provided to all participants regardless of location. Both the person-centered health plan and the iCBT are provided in Swedish primary health care and require minimal health care resources, offering a potentially cost-effective approach.

Conclusions

This study contributes a novel perspective by examining whether a stress-focused intervention influences pain-related outcomes, addressing an area that has received limited attention in previous research. No differences in pain intensity, number of pain locations, or associated symptoms were observed between the intervention group receiving therapist-guided iCBT focused on stress combined with person-centered physical activity and the control group receiving physical activity alone. However, both groups demonstrated improvements over time across several outcomes. Participants in the intervention group achieved their individually set goals to a greater extent than those in the control group, suggesting a potential added value of the intervention in supporting goal attainment. In a clinical context, these findings support person-centered, collaboratively developed physical activity plans as a central component in the management of CWP, while stress-management interventions may be considered as complementary support to facilitate behavioral change and goal attainment. Although iCBT did not provide additional improvements in pain or other standard outcomes, it may still be beneficial for selected patients, particularly those with higher levels of perceived stress, which warrants further investigation. The relatively high rate of loss to follow-up warrants caution when interpreting the findings, as it may affect both the internal validity and the generalizability of the results.

Acknowledgments

The authors would like to acknowledge all participants in the study, as well as the physiotherapists at the Närhälsan Rehabilitation Centres in Eriksberg, Sannegården, Lidköping, and Uddevalla for conducting the physiotherapy intervention.

The authors declare the use of generative artificial intelligence (GenAI; ChatGPT [OpenAI]) in the research and writing process. According to the GAIDeT (Generative AI Delegation Taxonomy; 2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the outcomes. Declaration submitted by: AB.

XSL•FO
RenderX

Funding

This study was funded by The Operational Healthcare Committee, Region Västra Götaland, Sweden. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

The data used in this study are not readily available, as the sharing of patient data is restricted under Swedish law and the General Data Protection Regulation. Furthermore, access to the data was restricted in the ethical approval to members of the research group only. However, deidentified data may be made available by the corresponding author upon reasonable request and subject to appropriate confidentiality agreements.

Authors' Contributions

AB contributed to the study design, participant recruitment, data collection, data analysis, interpretation of results, and manuscript drafting. CMH-H delivered the physiotherapeutic intervention at one study site and participated in the interpretation of results and manuscript writing. ACL and MP delivered the internet-based cognitive behavioral therapy intervention and contributed to the interpretation of results and manuscript writing. AG-E, MEHL, SW, and LN contributed to the study design, interpretation of results, and manuscript writing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT 2025 checklist.
[PDF File (Adobe PDF File), 258 KB - jmir_v28i1e86792_app1.pdf]

Multimedia Appendix 2
CONSORT E-HEALTH checklist.
[PDF File (Adobe PDF File), 10096 KB - jmir_v28i1e86792_app2.pdf]

Multimedia Appendix 3
TIDieR checklist.
[PDF File (Adobe PDF File), 341 KB - jmir_v28i1e86792_app3.pdf]

Multimedia Appendix 4
The 14 modules in the internet-based cognitive behavioral therapy program.
[PNG File , 387 KB - jmir_v28i1e86792_app4.png]

References

1. Wolfe F, Smythe HA, Yunus MB, Bennett RM, Bombardier C, Goldenberg DL, et al. The American College of Rheumatology 1990 criteria for the classification of fibromyalgia. Report of the multicenter criteria committee. Arthritis Rheum 1990;33(2):160-172. [doi: 10.1002/art.1780330203] [Medline: 2306288]
2. Andrews P, Steultjens M, Riskowski J. Chronic widespread pain prevalence in the general population: a systematic review. Eur J Pain 2018;22(1):5-18. [doi: 10.1002/ejp.1090] [Medline: 28815801]
3. Mansfield KE, Sim J, Croft P, Jordan KP. Identifying patients with chronic widespread pain in primary care. Pain 2017;158(1):110-119 [FREE Full text] [doi: 10.1097/j.pain.0000000000000733] [Medline: 27749607]
4. Morales-Espinoza EM, Kostov B, Salami DC, Perez ZH, Rosalen AP, Molina JO, et al. Complexity, comorbidity, and health care costs associated with chronic widespread pain in primary care. Pain 2016;157(4):818-826. [doi: 10.1097/j.pain.0000000000000440] [Medline: 26645546]
5. Rohrbeck J, Jordan K, Croft P. The frequency and characteristics of chronic widespread pain in general practice: a case-control study. Br J Gen Pract 2007;57(535):109-115 [FREE Full text] [Medline: 17263927]
6. Nijs J, George SZ, Clauw DJ, Fernández-de-Las-Peñas C, Kosek E, Ickmans K, et al. Central sensitisation in chronic pain conditions: latest discoveries and their potential for precision medicine. Lancet Rheumatol 2021;3(5):e383-e392. [doi: 10.1016/S2665-9913(21)00032-1] [Medline: 38279393]
7. Vyverman J, De Baere R, Timmers I, Coppieters I, Van Oosterwijck J, Moerkerke M. The stress-pain connection in chronic primary pain: a systematic review and meta-analysis of physiological stress markers in relation to experimental pain responses. Neurosci Biobehav Rev 2026;184:106604. [doi: 10.1016/j.neubiorev.2026.106604] [Medline: 41679412]

XSL•FO
RenderX

8. Fischer S, Doerr JM, Strahler J, Mewes R, Thieme K, Nater UM. Stress exacerbates pain in the everyday lives of women with fibromyalgia syndrome--the role of cortisol and alpha-amylase. Psychoneuroendocrinology 2016;63:68-77. [doi: 10.1016/j.psyneuen.2015.09.018] [Medline: 26431802]
9. Mundal I, Gråwe RW, Bjørngaard JH, Linaker OM, Fors EA. Prevalence and long-term predictors of persistent chronic widespread pain in the general population in an 11-year prospective study: the HUNT study. BMC Musculoskelet Disord 2014;15:213 [FREE Full text] [doi: 10.1186/1471-2474-15-213] [Medline: 24951013]
10. Larsson B, Björk J, Börsbo B, Gerdle B. A systematic review of risk factors associated with transitioning from regional musculoskeletal pain to chronic widespread pain. Eur J Pain 2012;16(8):1084-1093. [doi: 10.1002/j.1532-2149.2012.00117.x] [Medline: 22362638]
11. Allesøe K, Lau CJ, Buhelt LP, Aadahl M. Physical activity, self-rated fitness and stress among 55,185 men and women in the Danish Capital Region Health Survey 2017. Prev Med Rep 2021;22:101373 [FREE Full text] [doi: 10.1016/j.pmedr.2021.101373] [Medline: 33868905]
12. Bergenheim A, Juhlin S, Nordeman L, Joelsson M, Mannerkorpi K. Stress levels predict substantial improvement in pain intensity after 10 to 12 years in women with fibromyalgia and chronic widespread pain: a cohort study. BMC Rheumatol 2019;3:5 [FREE Full text] [doi: 10.1186/s41927-019-0072-9] [Medline: 31286111]
13. Trouvin AP, Perrot S. New concepts of pain. Best Pract Res Clin Rheumatol 2019;33(3):101415 [FREE Full text] [doi: 10.1016/j.berh.2019.04.007] [Medline: 31703792]
14. Macfarlane GJ, Kronisch C, Atzeni F, Häuser W, Choy EH, Amris K, et al. EULAR recommendations for management of fibromyalgia. Ann Rheum Dis 2017;76(12):e54 [FREE Full text] [doi: 10.1136/annrheumdis-2017-211587] [Medline: 28476880]
15. Wang T, Su H, Zhou Y, Du L, Lv Y, Yu L. Effects of exercise on pain, fatigue, and quality of life in people with fibromyalgia: a systematic review and meta-analysis of randomized controlled trials. Front Med (Lausanne) 2026;13:1782714. [doi: 10.3389/fmed.2026.1782714] [Medline: 41788711]
16. Bendau A, Petzold MB, Kaminski J, Plag J, Ströhle A. Exercise as treatment for "stress-related" mental disorders. Curr Neuropharmacol 2024;22(3):420-436 [FREE Full text] [doi: 10.2174/1570159X22666230927103308] [Medline: 37779399]
17. Meus T, Van Eetvelde J, Meuwissen I, Meeus M, Boullosa D, Timmermans A, et al. Exercise and heart rate variability in chronic musculoskeletal pain: a systematic review. Sports Med Open 2025;11(1):109. [doi: 10.1186/s40798-025-00916-8] [Medline: 41032171]
18. Terpstra JA, van der Vaart R, van Beugen S, van Eersel RA, Gkika I, Erdős D, et al. Guided internet-based cognitive-behavioral therapy for patients with chronic pain: a meta-analytic review. Internet Interv 2022;30:100587 [FREE Full text] [doi: 10.1016/j.invent.2022.100587] [Medline: 36406977]
19. Lindsäter E, Axelsson E, Salomonsson S, Santoft F, Ejeby K, Ljótsson B, et al. Internet-based cognitive behavioral therapy for chronic stress: a randomized controlled trial. Psychother Psychosom 2018;87(5):296-305. [doi: 10.1159/000490742] [Medline: 30041167]
20. Sanabria-Mazo JP, Colomer-Carbonell A, Fernández-Vázquez Ó, Noboa-Rocamora G, Cardona-Ros G, McCracken LM, et al. A systematic review of cognitive behavioral therapy-based interventions for comorbid chronic pain and clinically relevant psychological distress. Front Psychol 2023;14:1200685 [FREE Full text] [doi: 10.3389/fpsyg.2023.1200685] [Medline: 38187407]
21. Zhang YZ, Yang B, Jiang CH, Guo YX, Lai XX, Zhang WQ, et al. Efficacy of cognitive behavioral therapy combined with exercise in patients with chronic pain: a systematic review and meta-analysis. J Pain Res 2026;19:599315. [doi: 10.2147/JPR.S599315] [Medline: 42051752]
22. Hedman-Lagerlöf E, Carlbring P, Svärdman F, Riper H, Cuijpers P, Andersson G. Therapist-supported internet-based cognitive behaviour therapy yields similar effects as face-to-face therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. World Psychiatry 2023;22(2):305-314 [FREE Full text] [doi: 10.1002/wps.21088] [Medline: 37159350]
23. Hopewell S, Chan A, Collins GS, Hróbjartsson A, Moher D, Schulz KF, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ 2025;389:e081123 [FREE Full text] [doi: 10.1136/bmj-2024-081123] [Medline: 40228833]
24. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
25. Sullivan M, Karlsson J, Ware JE. The Swedish SF-36 health survey-I. Evaluation of data quality, scaling assumptions, reliability and construct validity across general populations in Sweden. Soc Sci Med 1995;41(10):1349-1358. [doi: 10.1016/0277-9536(95)00125-q] [Medline: 8560302]
26. Hays RD, Sherbourne CD, Mazel RM. The RAND 36-item health survey 1.0. Health Econ 1993;2(3):217-227. [doi: 10.1002/hec.4730020305] [Medline: 8275167]
27. Burckhardt CS, Clark SR, Bennett RM. The Fibromyalgia Impact Questionnaire: development and validation. J Rheumatol 1991;18(5):728-733. [Medline: 1865419]

XSL•FO
RenderX

28. Hedin PJ, Hamne M, Burckhardt CS, Engström-Laurent A. The Fibromyalgia Impact Questionnaire, a Swedish translation of a new tool for evaluation of the fibromyalgia patient. Scand J Rheumatol 1995;24(2):69-75. [doi: 10.3109/03009749509099287] [Medline: 7747146]
29. Bull FC, Al-Ansari SS, Biddle S, Borodulin K, Buman MP, Cardon G, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. Br J Sports Med 2020;54(24):1451-1462 [FREE Full text] [doi: 10.1136/bjsports-2020-102955] [Medline: 33239350]
30. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: Template for Intervention Description and Replication (TIDieR) checklist and guide. BMJ 2014;348:g1687 [FREE Full text] [doi: 10.1136/bmj.g1687] [Medline: 24609605]
31. Aboushaar N, Serrano N. The mutually reinforcing dynamics between pain and stress: mechanisms, impacts and management strategies. Front Pain Res (Lausanne) 2024;5:1445280 [FREE Full text] [doi: 10.3389/fpain.2024.1445280] [Medline: 39624230]
32. Zetterqvist K, Maanmies J, Ström L, Andersson G. Randomized controlled trial of internet-based stress management. Cogn Behav Ther 2003;32(3):151-160. [doi: 10.1080/16506070302316] [Medline: 16291546]
33. Leplege A, Gzil F, Cammelli M, Lefeve C, Pachoud B, Ville I. Person-centredness: conceptual and historical perspectives. Disabil Rehabil 2007;29(20-21):1555-1565. [doi: 10.1080/09638280701618661] [Medline: 17922326]
34. Larsson A, Barenfeld E, Fors A, Gjertsson I, Bergenheim A, Mannerkorpi K. Person-centred health plans for physical activity in persons with chronic widespread pain (CWP) - a retrospective descriptive review. Disabil Rehabil 2023;45(11):1857-1864 [FREE Full text] [doi: 10.1080/09638288.2022.2077992] [Medline: 35658763]
35. Juhlin S, Bergenheim A, Gjertsson I, Larsson A, Mannerkorpi K. Physical activity with person-centred guidance supported by a digital platform for persons with chronic widespread pain: a randomized controlled trial. J Rehabil Med 2021;53(4):jrm00175 [FREE Full text] [doi: 10.2340/16501977-2796] [Medline: 33576434]
36. Feldthusen C, Bergenheim A, Barenfeld E, Gjertsson I, Larsson A, Mannerkorpi K. Experience of co-creation of a health plan and support for sustainable physical activity among people with chronic widespread pain: a qualitative interview study in Sweden. BMJ Open 2022;12(4):e059432 [FREE Full text] [doi: 10.1136/bmjopen-2021-059432] [Medline: 35396312]
37. Bergman S. Psychosocial aspects of chronic widespread pain and fibromyalgia. Disabil Rehabil 2005;27(12):675-683. [doi: 10.1080/09638280400009030] [Medline: 16012060]
38. Ericsson A, Nyström C, Mannerkorpi K. Psychometric properties of the Stress and Crisis Inventory (SCI-93) in females with fibromyalgia and chronic widespread pain. Nord J Psychiatry 2015;69(1):28-34. [doi: 10.3109/08039488.2014.915580] [Medline: 24848785]
39. Krafft B, Nyström C, Nyström O, Pedersen C. Stressreaktion och välfärdsdata - en populationsstudie [Article in Swedish]. Socialmedicinsk Tidskrift 2016;81(3):222-229. [doi: 10.62607/smt.v81i3.45800]
40. Mannerkorpi K, Hernelid C. Leisure time physical activity instrument and physical activity at home and work instrument. Development, face validity, construct validity and test-retest reliability for subjects with fibromyalgia. Disabil Rehabil 2005;27(12):695-701. [doi: 10.1080/09638280400009063] [Medline: 16012062]
41. Smets EM, Garssen B, Bonke B, De Haes JC. The Multidimensional Fatigue Inventory (MFI) psychometric qualities of an instrument to assess fatigue. J Psychosom Res 1995;39(3):315-325. [doi: 10.1016/0022-3999(94)00125-o]
42. Smets EM, Garssen B, Bonke B, De Haes JC. The Multidimensional Fatigue Inventory (MFI) psychometric qualities of an instrument to assess fatigue. J Psychosom Res 1995;39(3):315-325. [doi: 10.1016/0022-3999(94)00125-o] [Medline: 7636775]
43. Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiatr Scand 1983;67(6):361-370. [doi: 10.1111/j.1600-0447.1983.tb09716.x] [Medline: 6880820]
44. Herrmann C. International experiences with the Hospital Anxiety and Depression Scale--a review of validation data and clinical results. J Psychosom Res 1997;42(1):17-41. [doi: 10.1016/s0022-3999(96)00216-4] [Medline: 9055211]
45. Hurst H, Bolton J. Assessing the clinical significance of change scores recorded on subjective outcome measures. J Manipulative Physiol Ther 2004;27(1):26-35 [FREE Full text] [doi: 10.1016/j.jmpt.2003.11.003] [Medline: 14739871]
46. Dworkin RH, Turk DC, Wyrwich KW, Beaton D, Cleeland CS, Farrar JT, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. J Pain 2008;9(2):105-121 [FREE Full text] [doi: 10.1016/j.jpain.2007.09.005] [Medline: 18055266]
47. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Jacob Cohen: Routledge; 2013.
48. Oliver S, Fisher K, Childs S. What psychological and physical changes predict patients' attainment of personally meaningful goals six months following a CBT based pain management intervention? Disabil Rehabil 2017;39(22):2308-2314. [doi: 10.1080/09638288.2016.1226415] [Medline: 27871195]
49. Nordeman L, Grimby-Ekman A, Ho-Henriksson CM, Enhörning E, Hellgren M, Bergenheim A. Change in self-reported activity in chronic low back pain after 13 years - a prospective longitudinal cohort study in primary healthcare. BMC Musculoskelet Disord 2025;27(1):15 [FREE Full text] [doi: 10.1186/s12891-025-09371-8] [Medline: 41331477]
50. Bartley EJ, Robinson ME, Staud R. Pain and fatigue variability patterns distinguish subgroups of fibromyalgia patients. J Pain 2018;19(4):372-381 [FREE Full text] [doi: 10.1016/j.jpain.2017.11.014] [Medline: 29253551]

XSL•FO
RenderX

51. Gervind E, Salem MB, Svanborg C, Nyström ME, Lilja JL, Kaldo V, et al. The influence of organizational models on the implementation of internet-based cognitive behavior therapy in primary care: a mixed methods study using the RE-AIM framework. Internet Interv 2024;35:100698 [FREE Full text] [doi: 10.1016/j.invent.2023.100698] [Medline: 38174208]

Abbreviations

ACR: American College of Rheumatology
CBT: cognitive behavioral therapy
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
CWP: chronic widespread pain
FIQ: Fibromyalgia Impact Questionnaire
HADS: Hospital Anxiety and Depression Scale
iCBT: internet-based cognitive behavioral therapy
MCAR: missing completely at random
MFI: Multidimensional Fatigue Inventory
PGIC: Patient Global Impression of Change
RCT: randomized controlled trial
SCI: Stress and Crisis Inventory
SF-36: Short Form-36
TIDieR: Template for Intervention Description and Replication
VAS: visual analog scale
WHO: World Health Organization

Edited by S Brini; submitted 11.Nov.2025; peer-reviewed by G Lewis, J Thorlund; comments to author 10.Mar.2026; accepted 09.Jun.2026; published 29.Jun.2026.

Please cite as:
Bergenheim A, Larsson MEH, Ho-Henriksson CM, Grimby-Ekman A, Larsson AC, Persson M, Weineland S, Nordeman L
Combined Internet-Based Cognitive Behavioral Therapy and Face-to-Face Physiotherapy in Primary Health Care for Chronic Widespread Pain: Randomized Controlled Trial
J Med Internet Res 2026;28:e86792
URL: https://www.jmir.org/2026/1/e86792
doi:10.2196/86792
PMID:

©Anna Bergenheim, Maria EH Larsson, Chan-Mei Ho-Henriksson, Anna Grimby-Ekman, Anna Caroline Larsson, Marie Persson, Sandra Weineland, Lena Nordeman. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effectiveness of a Home-Based and Group-Based Tele-Exercise Program for Breast Cancer Survivors: Pilot Randomized Controlled Trial

Yang Gao[1*], PhD; Yan Sun[2*], PhD; Xueying Wang[1], MEd; Oi Kwan Chun[3], MD; Brigitte Kim Yook Fung[4], MD; Wai Wun Sara Fung[3], MD

[1]Department of Sports and Health Sciences, Academy of Wellness and Human Development, Hong Kong Baptist University, AAB 935, Level 9, Academic and Administration Building, Baptist University Road Campus, Hong Kong, China (Hong Kong)

[2]Department of Sports and Health Sciences, Academy of Wellness and Human Development, Faculty of Arts and Social Sciences, Hong Kong Baptist University, Hong Kong, China (Hong Kong)

[3]Department of Surgery, Kwong Wah Hospital, Hong Kong, China (Hong Kong)

[4]Department of Physiotherapy, Kwong Wah Hospital, Hong Kong, China (Hong Kong)

[*]these authors contributed equally

Corresponding Author:
Yang Gao, PhD
Department of Sports and Health Sciences, Academy of Wellness and Human Development, Hong Kong Baptist University, AAB 935, Level 9, Academic and Administration Building, Baptist University Road Campus, Hong Kong, China (Hong Kong)

Abstract

Background: Breast cancer remains the most prevalent cancer among women globally. Adjuvant therapies can cause adverse effects that compromise physical and mental health. Exercise may mitigate these effects; however, many breast cancer survivors remain insufficiently active.

Objective: This pilot study aimed to test the effectiveness of a 7-week theory-informed tele-exercise intervention for breast cancer survivors in Hong Kong.

Methods: We developed a 12-week theory-informed tele-exercise intervention for breast cancer survivors in Hong Kong; this pilot tested a 7-week abbreviated version. In this 2-group randomized controlled pilot trial, 34 individuals were assessed for eligibility, and 27 were randomized; 24 completed baseline and were included in the modified intention-to-treat analyses (12 per group). The intervention comprised a progressively supervised group-based tele-exercise intervention transitioning to unsupervised sessions, combined with psychological counseling. Outcomes were guided by the RE-AIM (reach, effectiveness, adoption, implementation, and maintenance) framework, with emphasis on feasibility, acceptability, and preliminary effects.

Results: Recruitment was 79.4% (27/34) and baseline-to-post retention was 100% (24/24), with satisfactory attendance (87.7%) and compliance (85.3%) in the intervention group. Acceptability was high. Preliminary signals of improvement were observed in cardiorespiratory fitness (primary outcome), lower-extremity strength, balance, affected-side shoulder range of motion, and health-related quality of life.

Conclusions: This pilot supports the feasibility and acceptability of tele-exercise for breast cancer rehabilitation and provides preliminary evidence to justify a larger trial with longer follow-up to assess sustained effects and broader applicability.

Trial Registration: ClinicalTrials.gov NCT06382441; https://clinicaltrials.gov/study/NCT06382441

(J Med Internet Res 2026;28:e79564) doi:10.2196/79564

KEYWORDS

breast cancer survivors; home-based and group-based tele-exercise program; randomized controlled trial; pilot study; telehealth; health-related quality of life

Introduction

Globally, breast cancer is the most prevalent cancer in women; in 2022, an estimated 2.3 million women were diagnosed with breast cancer and 670,000 died from the disease [1]. A review further reported that breast cancer incidence and mortality increased by 20% and 14%, respectively, from 2008 to 2012 [2]. Similarly, in Hong Kong, breast cancer is the most common cancer in females. The latest statistics show that breast cancer accounted for 28.9% of all new cancers in females in 2023 (n=5585) [3]. It is also the third leading cause of cancer

XSL•FO
RenderX

mortality, following lung and colorectal cancers and accounting for 13.1% of cancer mortality in females in 2024 [3].

Adjuvant therapy (ie, treatment after surgery) is the standard treatment. More than 90% of all patients receive an adjuvant therapy [4]. In Hong Kong, most cases after surgery receive a combined treatment of radiation and chemotherapy. Radiation therapy only usually results in short-term side effects (such as fatigue and skin burns), while those receiving chemotherapy may suffer from both short-term (such as fatigue and gastrointestinal symptoms) and long-term consequences (such as weight gain and cognitive dysfunctions) [5-7]. These adverse effects are associated with decreased physical and mental health, decreased health-related quality of life, and increased risk for mortality [2].

Evidence has shown that 70% of newly diagnosed cases are insufficiently active (<150 minutes/week) [8]; >60% of patients reduce physical activity levels postdiagnosis [8], with greater declines observed in those receiving adjuvant radiation therapy with chemotherapy [2]; both insufficient activity and declines in physical activity levels can elevate risks for the side effects of adjuvant treatments [8]. A large body of research has suggested exercise-induced effects in ameliorating cancer and treatment-related symptoms and improving health. Two recent Cochrane reviews have further suggested that exercise can elicit beneficial effects on fatigue, cardiopulmonary fitness, muscular strength, depression, and health-related quality of life with moderate-quality evidence (indicating further studies may change the estimate), while evidence in cognitive function, lymphedema, weight, anxiety, and physical activity behaviors is low- or very low-quality (indicating the estimate is likely or very likely to change with further studies) [2,4,5]. Supervised exercise consisting of both aerobic and resistance training is consistently suggested for breast cancer patients due to its greater effects than unsupervised modes [2,4,5]. However, its sustainability of both exercise habits and beneficial effects after cessation is mixed and critical [2,4,5].

Along with advances in technology, online-based interventions have been emerging in recent years. Particularly, tele-exercise is a novel approach to the delivery of exercise training using information and communication technologies such as the internet, computers, mobile devices, software, and smartphone apps, and allows participants to exercise at home with supervision [9]. The common barriers against exercise in breast cancer survivors include tiredness, transportation, and lack of time [10]. Thus, the home-based tele-exercise approach will help break down these barriers and improve participation and retention of participants. Compared to the traditional face-to-face mode, this approach can also reduce health care burden and costs from both health care providers and patients [9]. Some evidence, though sparse, has suggested that the tele-exercise approach is as effective as the conventional face-to-face exercise to manage disease and promote health [11]. Home-based interventions with minimal exercise equipment were effective and viable alternatives to equipment-intensive regimes [11].

Collectively, exercise can help improve health and reduce treatment-related symptoms among breast cancer survivors. However, strategies are needed to support the sustainability of exercise after supervised programs end, and tele-exercise is a promising approach that warrants evaluation. To address these gaps, we developed a theory-informed tele-exercise program integrating behavior change techniques (BCTs) and a progressive transition from supervised to unsupervised delivery to support habit formation. This paper reports a 7-week pilot study, with the primary objective of assessing acceptability and feasibility and secondary objectives of obtaining preliminary estimates (between-group differences and effect sizes) to inform the design and sample size of the 12-week trial.

Methods

Study Design and Setting

This 2-group pilot randomized controlled trial (RCT) of a tele-exercise program was conducted in Hong Kong in 2024. About 27 participants were randomized; 24 completed the baseline assessment and were included in the modified intention-to-treat (mITT) analyses. The study was reported in accordance with the CONSORT (Consolidated Standards of Reporting Trials) extension for randomized pilot and feasibility trials (Checklist 1), and the evaluation was guided by the RE-AIM (reach, effectiveness, adoption, implementation, and maintenance) framework. Outcome assessors and data analysts were blinded to group allocation. A multidisciplinary task force (health care professionals, exercise science and psychology experts, clinical trial specialists, data analysts, coaches, and a local breast cancer survivor association) supported intervention development and delivery.

Ethical Considerations

Ethical approval was obtained from the Research Ethics Committee of the Hospital Authority of Hong Kong (KC/KE-23‑0122/ER-4). The trial was registered on ClinicalTrials.gov (NCT06382441) and conducted in accordance with the Declaration of Helsinki and ICH-GCP (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use – Good Clinical Practice). All participants provided written informed consent and were informed of their right to withdraw at any time.

Confidentiality was protected through restricted access to study files, secure storage of paper records in locked cabinets, and encryption of electronic data (including SSL/TLS [Secure Sockets Layer/Transport Layer Security] encryption for any web-based data transfer). Personally identifying information and the ID linkage file were stored separately from the analysis dataset, which was deidentified prior to analysis and reporting. Study documents are retained for 5 years after study completion in line with hospital authority/hospital policies and are securely destroyed thereafter. No identifiable participant information is presented in this paper or multimedia appendices.

Participants and Recruitment

Breast cancer survivors in the adjuvant-treatment period (within 6 weeks after completion of chemotherapy, with radiotherapy as per usual care) were recruited from the Breast Centre at Kwong Wah Hospital, Hong Kong, between November 2023 and April 2024. Inclusion criteria included (1) females aged 40‑64 years, (2) within 6 weeks after completion of

XSL•FO
RenderX

chemotherapy, (3) without severe anemia, (4) without cancer metastasis, (5) able to read and communicate in Cantonese or Mandarin, and (6) smartphone users. Those with any medical, physical, or psychological conditions that may limit participation were excluded from the study, such as uncontrolled severe cardiovascular disease, schizophrenia, and severe neurological dysfunction. Eligibility was assessed by a case manager of breast cancer (a nurse consultant), consulting with a surgeon or a physiotherapist if needed.

Previous studies have suggested that 10% of the estimated sample size of the main study (n=130) is recommended for a simple-armed pilot (ie, intervention only). Given that a control group was included, the estimated sample size for the pilot was 24. Several effective strategies for participant recruitment and retention suggested by previous studies were adopted, including emphasizing study benefits to participants (eg, free assessments of the rehabilitation process on physiological and psychological functions) [12].

Randomization and Allocation Concealment

Participants were randomized in a 1:1 ratio to the intervention or control group using permuted block randomization with randomly varying block sizes (4, 6, or 8). The randomization schedule was computer-generated prior to recruitment by an independent staff member not involved in enrollment, intervention delivery, or outcome assessment.

Allocation concealment was implemented using sequentially numbered, opaque, sealed envelopes (SNOSE) prepared in advance and stored by independent research staff. After eligibility confirmation and written consent, the next envelope was opened to reveal the group assignment. Baseline assessments were scheduled immediately after randomization; however, 3 participants withdrew after allocation and before completing the baseline assessment. Outcome assessors and data analysts remained blinded to group allocation at both baseline and postintervention assessments. Intervention coaches were not blinded because they delivered the program, but they were not involved in outcome assessments or data analysis. Coaches did not have access to the randomization sequence/allocation list or outcome data beyond participant contact details and scheduling information required for intervention delivery.

Intervention Group (Group A)

The tele-exercise (Table S1 in Multimedia Appendix 1), consisting of both aerobic and resistance exercises with intensity progressively increasing to a moderate level, was delivered to patients in the intervention group for 7 weeks (150 minutes per week). Delivery mode was transitioned gradually from supervised in week 1 (3 sessions/week, 30‑50 minutes/session) to mixed mode with 1 supervised session and 2 unsupervised sessions in week 7 (Table S2 in Multimedia Appendix 1).

Twelve participants formed two groups (6 participants per group) according to their individual exercise intensity levels determined prior to the intervention for each participant. Each group received a total of 15 online supervised training sessions. Each session consisted of warm-up, aerobic exercise, resistance exercise, stretching, and cool-down, with the exercise period progressively increasing from 30 minutes to 50 minutes and exercise intensity progressively increasing from low to moderate levels. Lighter intensity was adopted when there was a need (eg, for those with severe conditions). Target heart rate during exercise was monitored by a wearable activity monitor watch (Mi Band 5, Xiaomi Corporation) to ensure the real exercise intensity (%HRR, heart rate reserve) fulfilled the individualized requirements. Cameras were requested to be turned on all the time. Sessions were delivered by a certified fitness instructor (coach) and an assistant coach. Both were coinvestigators of the study. The lead coach held a Certificate in Fitness Instruction (Hong Kong Baptist University), and the assistant coach held a master's degree in Sport and Leisure Management (Hong Kong Baptist University). For quality control, YG (the principal investigator) participated in the first 3 sessions in week 1 and attended 1 session per week in weeks 2‑4 without prior notification. Figure 1 illustrates some exercise examples. Starting from week 4, unsupervised sessions were introduced, with one session in weeks 4 and 5 and two sessions in weeks 6 and 7. Participants were asked to complete these sessions at home by watching pre-made videos for them. They were also asked to record their exercise (eg, session duration) in an exercise log and to record heart rate using the Mi Band 5.

XSL•FO
RenderX

Figure 1. Exercises of the tele-exercise for the pilot study.

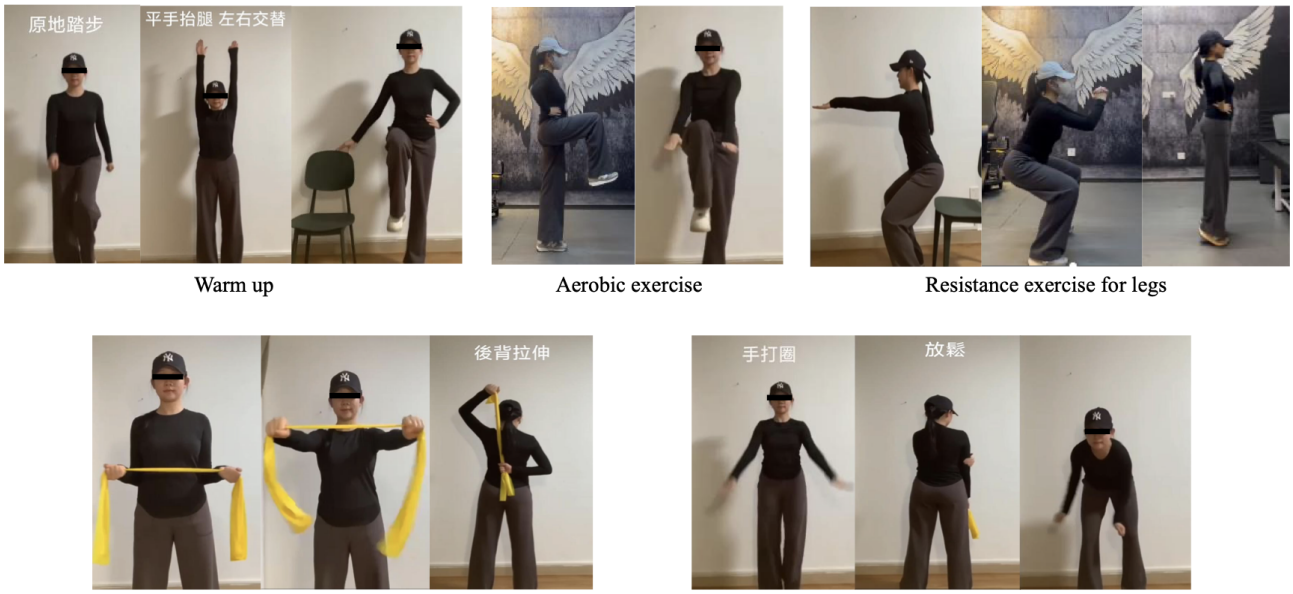

Warm up | Aerobic exercise | Resistance exercise for legs

Resistance exercise for arms | Cool down

In order to promote habit formation of exercise, in each supervised session, we included effective behavior counseling elements. These elements were guided by an integrated theoretical framework targeting motivational, volitional, and habit-formation processes, drawing on dual-process accounts of behavior (reflective vs automatic processes) and habit formation theory to support both behavior change and maintenance. The counseling content was developed based on established psychological theories and BCTs used in our previous studies [12-15] and tailored to unique facilitators and barriers to exercise in breast cancer survivors. Seven psychological counseling sessions (30 minutes per session) were delivered by the same coach and assistant from the tele-exercise sessions at a frequency of once per week for 7 weeks. The lead coach had previous experience as a research assistant in a tele-exercise intervention for patients recovering from COVID-19, co-led by the principal investigator (YG), where she supported the delivery of educational workshops. For the present trial, the coach completed three 1-hour training sessions focused on counseling content and techniques tailored to breast cancer survivors and undertook several practice sessions to further refine counseling delivery.

The psychological strategies aimed at changing motivation, self-regulation skills, and habit formation of exercise. Motivational content targeted reflective motivation (eg, strengthening intentions and outcome expectancies), volitional content targeted self-regulation and action control (eg, planning and self-monitoring), and habit-formation content targeted automaticity by establishing stable contextual cues and repetition. For example, the participants were advised on how to work toward their goals and maintain motivation in the face of barriers (motivational strategy); they were encouraged to set outcome-specific goals and self-monitor their progress toward the goals by monitoring how much time they spent in exercise (self-regulation strategy); and they were asked to recognize and identify situational, contextual, and time-based cues that could prompt them to be active (habit development strategy). Table S3 in Multimedia Appendix 1 details the topics, theoretical strategies, and BCTs involved in each session, and Multimedia Appendix 2 provides the detailed content of session 1 (Chinese only).

On the day of baseline measurements, participants in group A received a face-to-face briefing on installing and using Zoom and on intervention procedures. Additional support was provided before, during, and after the intervention via a hotline and social media (eg, WhatsApp). Participants were asked not to share intervention-related materials (including videos) with other patients to minimize potential contamination.

Safety and potential adverse effects of intervention: It is critical to monitor and record harm or unintended consequences in exercise interventions for breast cancer survivors in order to ensure that the interventions are safe and appropriate. In the intervention arm, a pulse oximeter (Innovo Premium Fingertip Pulse Oximeter) was used to monitor oxygen saturation in the initial exercise sessions to avoid exercise-induced oxyhemoglobin desaturation [16]. Participants with lymphedema wore a compression sleeve during exercise. Possible adverse effects, such as lymphedema, fracture, and other arm/shoulder symptoms, suggested by previous studies were monitored and remedied promptly via diverse means (eg, to inform participants of the possible adverse consequences in advance and provide feasible and practical preventive strategies to them) [16,17]. According to the American College of Sports Medicine's guidelines for exercise, the tele-exercise was stopped immediately in the following situations during exercise: (1) deteriorated upper extremity symptoms; (2) chest pain; (3) central nervous system symptoms (eg, dizziness or near syncope); (4) fatigue, shortness of breath, wheezing, muscular cramps, or claudication; (5) heart rate during exercise ≥90% HRR (the upper limit of vigorous intensity); (6) an absolute decrease in oxygen saturation SpO2≥5% or SpO2≤80% [16].

After the intervention, all participants in the intervention group were interviewed regarding acceptability and feasibility. Secondary health outcomes were measured among all participants before and after the intervention to investigate

XSL•FO
RenderX

preliminary effects of the study and to calculate the sample size of the main study.

Control Group (Group B)

Participants in the control group received 7 educational essays on exercise and health (1 essay per week for 7 weeks). All participants in both groups continued to receive usual care from the study hospital during the study period. Multimedia Appendix 3 provides an example essay (Chinese only).

Outcomes and Measurements

The pilot was assessed following the RE-AIM framework. Both quantitative and qualitative data were collected via semistructured, face-to-face interviews [18]. The interview included a structured 14-item acceptability and feasibility questionnaire, with each item rated on a 5-point Likert scale ("1=strongly disagree" to "5=strongly agree"). For each item, participants were also invited to elaborate through an open-ended follow-up prompt to explain their rating and provide additional comments.

The structured Likert-scale responses were analyzed quantitatively to summarize acceptability and feasibility across domains, including the individualized exercise design, exercise-intensity progression, session frequency/duration, perceived difficulty of the online guided exercises, scheduling and staffing arrangements, transition to independent activity, perceived usefulness and safety, overall satisfaction, willingness to recommend the intervention, and acceptability of the assessment procedures. The open-ended responses were used to complement and contextualize the quantitative findings.

Given the pilot nature of the study and the supportive role of the qualitative data, we conducted a descriptive content summary. One researcher (SY) reviewed the written interview notes, extracted relevant text, and organized comments into broad categories guided by the interview domains (eg, perceived benefits, barriers, facilitators, and recommendations). The qualitative summaries were used to support the interpretation of the quantitative results rather than to generate formal qualitative themes. No qualitative analysis software was used; data were managed in Microsoft Word. Table 1 details the indicators of acceptability and feasibility.

Table . Description of indicators of acceptability and feasibility.

Measure	Description of measurement
Acceptability	Participant-rated acceptability, satisfaction, enjoyment, perceptions of safety, difficulty, and tolerability of treatment (using a 5-point Likert questionnaire), with optional open-ended comments; adverse events were monitored throughout the intervention.
Recruitment rate	The proportion of eligible participants who consented and were randomized. No. of participants randomizedNo. of eligible participants× 100%
Retention rate	The proportion of participants who completed the baseline assessment and remained throughout the entire treatment period. No. of participants at treatment endNo. of participants at baseline× 100%
Attendance rate	The proportion of total sessions offered to participants to the actual number of sessions participants attended. No. of sessions participants attendedNo. of sessions offered× 100%
Compliance rate	The proportion of intervention sessions in which participants achieved the prescribed exercise intensity (based on meeting the target heart rate reserve (HRR) zone, Table S1 in Multimedia Appendix 1); missed sessions were coded as not compliant. No. of sessions meeting target intensityNo. of sessions offered× 100%
Fidelity	Content, frequency, duration, and coverage as originally intended. Assessed by the deliverer's adherence to the originally intended treatment plan.

Health outcomes were assessed pre- and postintervention for all participants, including cardiopulmonary fitness (*primary outcome of the main study*), muscular strength, balance, body composition, affected-side shoulder range of motion (ROM), lymphedema, and health-related quality of life. These outcomes were carefully selected based on literature and clinical importance. Table 2 lists the details of the measurements. A self-administered questionnaire was developed to collect personal information, health-related quality of life, physical activity level, and other related factors.

XSL•FO
RenderX

Table . Description of measurements.

Measure	Description of measurement
Primary outcome of the main study	
Cardiopulmonary fitness	The 6-minute walk test (m) was performed to estimate cardiopulmonary fitness. Participants walked as far as possible in 6 minutes on a floor without a slope and without running. The distance (m) completed was used. The test was taken only once.
Secondary outcomes of the main study	
Muscular strength	Upper extremity muscular strength was estimated with the handgrip test (kg). Participants, while standing, exerted maximum force to squeeze the instrument. Each hand was tested twice, and the best out of the 4 attempts was recorded. Additionally, the higher result for the affected side arm was noted. For participants with both arms affected, the average of the higher results from each arm was used for analysis.
	Lower extremity muscular strength was estimated with the timed stand test (seconds). Participants were asked to hold their arms crossed at the height of the chest and stand up and down from a sitting position to complete 10 times as fast as possible. The faster (shorter time) from 2 attempts was used.
Balance	Estimated using the single leg stance test (seconds). It is used to estimate static postural and balance control and is commonly used in clinical settings to monitor neurological and musculoskeletal conditions [19].
Body composition	Body composition was assessed using a body composition analyzer (InBody 270). Percent body fat and BMI (kg/m²) were recorded.
Shoulder range of motion (ROM) on affected side	An arm goniometer was used to measure shoulder joint ROM in forward flexion (degrees).
Limb circumference and lymphedema	Arm circumference was measured with a nonstretch tape at styloid and every 10 cm intervals from the ulnar styloid up to 40 cm distally [20]. Lymphedema in the affected arm was defined as a difference of >2 cm in arm circumference at any measurement point compared to its value before surgery.
Health-related quality of life (HRQOL)	Functional Assessment of Cancer Therapy – Breast (FACT-B), a 5-point Likert instrument designed to measure 5 domains of HRQOL in breast cancer patients, including physical, social, emotional, and functional well-being, and a breast cancer subscale. It consists of 37 items, processes sound valid and reliable, and is sensitive to patient progression. The traditional Chinese version was adopted.
Other measurements of the main study	
Height (*cm*)	Measured using a portable stadiometer (Seca 284).
Physical activity level	Self-reported using the International Physical Activity Questionnaire (short version). Weekly time spent on moderate to vigorous intensity physical activity (MVPA, min/week) was calculated.
Personal and medical information	Measured using either a self-administered questionnaire or obtained from clinical records.

Data Analysis

Quantitative data were summarized descriptively to evaluate the acceptability and feasibility of the tele-exercise. Qualitative open-ended comments were summarized descriptively to contextualize and supplement the quantitative findings. To preliminarily estimate the health effects of the tele-exercise protocol, changes in health outcomes (eg, cardiopulmonary fitness) were compared between the intervention and control groups using the independent-samples *t* test. Analyses followed an mITT approach including all randomized participants with baseline data. Participants who withdrew after randomization but before baseline assessments were excluded; as 3 participants withdrew postallocation (all before baseline), this may introduce selection bias. In the definitive trial, baseline assessments will be completed prior to allocation disclosure (or allocation will be concealed until after baseline) to minimize postrandomization attrition. Given the pilot nature and limited power, analyses were exploratory and focused on estimation. We report mean differences with 95% CIs and effect sizes (Cohen *d*; interpreted as small 0.2, medium 0.5, large 0.8), and treat *P* values as descriptive with no prespecified threshold for statistical significance [21]. No adjustment for multiple comparisons was applied because outcomes were exploratory. The sample size for the main trial was estimated using the repeated measures ANOVA within-between interaction approach (G*Power,

XSL•FO
RenderX

Version 3.1.9.7), with parameters of α=.05, β=.8, two groups and three measurement waves. Cohen *d* was converted to Cohen *f* using $f=d/\sqrt{3}$ [22].

Results

Feasibility was assessed by the following aspects.

Recruitment and Retention

Of 34 individuals assessed for eligibility, 27 provided consent and were randomized, yielding a recruitment rate of 79.4% (27/34). Of those randomized, 3 participants withdrew after allocation upon learning they had been assigned to the control group and did not complete the baseline assessment; thus, 24 participants completed the baseline assessment and were included in the mITT analyses (12 per group). The remaining 7 eligible individuals declined participation prior to randomization (most commonly due to being busy or lack of interest). All 24 participants who completed baseline also completed the postintervention assessment, resulting in a baseline-to-post retention rate of 100% (24/24). Figure 2 shows the flowchart of the pilot study.

Figure 2. Flowchart of the pilot study.

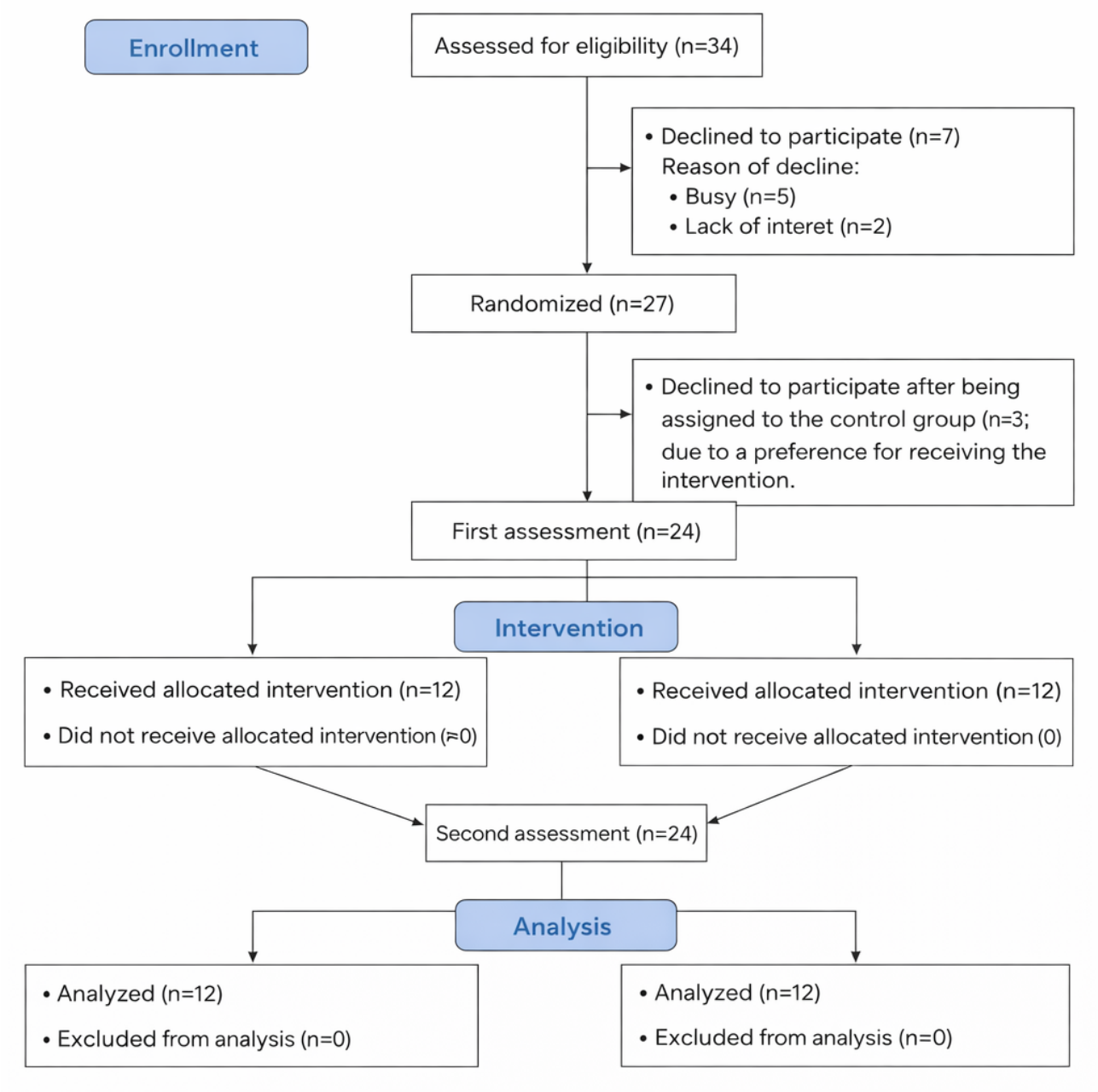

Attendance and compliance were summarized separately for supervised and unsupervised sessions (Table 3). Attendance was 88.3% (159/180) for supervised sessions and 86.1% (62/72) for unsupervised sessions. Compliance was 86.1% (155/180) and 83.3% (60/72), respectively. As shown in Figure 3, attendance and compliance were highest at the start (12/12 participants in sessions 1‑2) and then fluctuated modestly, with no obvious decline following the introduction of unsupervised sessions (diamond markers); attendance generally ranged from 9 to 12 participants and compliance from 8 to 12 participants across sessions.

XSL•FO
RenderX

Table . Attendance and compliance rates of the intervention.

	Overall	Supervised sessions	Unsupervised sessions
Attendance[a]	87.7% (221/252)	88.3% (159/180)	86.1% (62/72)
Compliance[a]	85.3% (215/252)	86.1% (155/180)	83.3% (60/72)

[a]Rates were calculated using session-level denominators based on scheduled sessions for all participants randomized to the intervention group (n=12): supervised sessions (12×15=180) and unsupervised sessions (12×6=72), totaling 252 sessions. For compliance, missed sessions were counted as noncompliant.

Figure 3. Attendance and compliance rates by session.

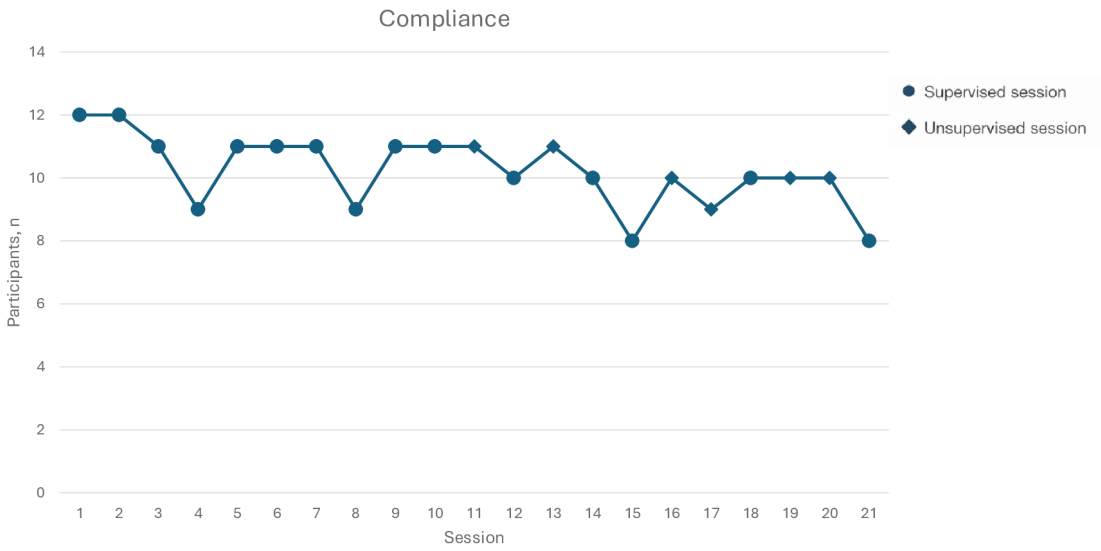

Acceptability of the intervention was assessed through interviews using 14 five-point Likert items among the 12 participants who completed the intervention. All 12 participants completed the acceptability ratings; 5 participants additionally provided optional open-ended written comments. The items covered various aspects, including the individualized approach, exercise intensity, frequency and duration of sessions, difficulty level, time scheduling, staffing arrangement, the transition from supervised to unsupervised delivery modes, usefulness of the psychological elements, safety, perceived effectiveness, satisfaction, willingness to recommend the program to others, and the assessment arrangement. All participants either agreed or strongly agreed with the acceptability of the intervention for half of the items (Table 4). For the remaining items, most responses were positive, with only one neutral response per item. In the optional open-ended comments (n=5), participants reported high satisfaction with the intervention, intention to recommend it to others, and a desire for continuation. One participant described perceived health improvements, including reduced symptoms and improved emotional and mental well-being, and expressed gratitude for the staff's professionalism and support. Regarding the neutral responses, one participant attributed her rating to missing several sessions due to scheduling conflicts with chemotherapy appointments. Another participant indicated a preference for more sessions per week and therefore rated "frequency and duration" as neutral.

XSL•FO
RenderX

Table . Acceptability of the intervention from the interview (n=12).

	Strongly agree, n (%)	Agree, n (%)	Neutral, n (%)	Disagree, n (%)	Strongly disagree, n (%)
Individualized approach	9 (75.0)	3 (25.0)	0 (0.0)	0 (0.0)	0 (0.0)
Intensity of exercise	11 (91.7)	1 (8.3)	0 (0.0)	0 (0.0)	0 (0.0)
Frequency and duration	9 (75.0)	2 (16.7)	1 (8.3)	0 (0.0)	0 (0.0)
Difficulty	10 (83.4)	1 (8.3)	1 (8.3)	0 (0.0)	0 (0.0)
Time arrangement	7 (58.4)	4 (33.3)	1 (8.3)	0 (0.0)	0 (0.0)
Staffing arrangement	9 (75.0)	2 (16.7)	1 (8.3)	0 (0.0)	0 (0.0)
Progressively transitioning from supervised to unsupervised modes	11 (91.7)	1 (8.3)	0 (0.0)	0 (0.0)	0 (0.0)
Usefulness of psychological approach	10 (83.3)	2 (16.7)	0 (0.0)	0 (0.0)	0 (0.0)
Safety	11 (91.7)	1 (8.3)	0 (0.0)	0 (0.0)	0 (0.0)
Effectiveness	10 (83.4)	1 (8.3)	1 (8.3)	0 (0.0)	0 (0.0)
Satisfaction	9 (75.0)	2 (16.7)	1 (8.3)	0 (0.0)	0 (0.0)
Recommendation to others	9 (75.0)	2 (16.7)	1 (8.3)	0 (0.0)	0 (0.0)
Arrangement of assessment				0 (0.0)	0 (0.0)
Questionnaire	10 (83.3)	2 (16.7)	0 (0.0)	0 (0.0)	0 (0.0)
Objective measurements	10 (83.3)	2 (16.7)	0 (0.0)	0 (0.0)	0 (0.0)

Fidelity of the intervention was evaluated in terms of content, frequency, duration, and coverage of the intervention.

- Content: All key components of the sessions were covered with minor deviations due to individual participant needs.
- Frequency: All planned supervised exercise and counseling sessions were delivered as scheduled (dose delivered), while participants attended 87.7% of scheduled sessions (dose received).
- Duration: The recorded durations of the sessions were within the expected range of 90-95% of the time. A few sessions lasted longer than the planned duration due to participant fatigue.
- Coverage: All participants in the intervention group who completed baseline (12/12) completed the postintervention assessment, resulting in a 100% coverage rate. There were no reports of intervention materials being shared with the control group, ensuring the integrity of the study conditions.

The results indicated that the intervention was implemented with high fidelity to the original protocol.

- Safety and adverse events: No adverse events, including no serious adverse events or exercise-related adverse events, were reported during the intervention period.

Health Outcomes

Characteristics of the Participants at Baseline

Table 5 shows the characteristics and the first assessment's results of the 24 participants. The average age of the participants was 57.54 (SD 8.37) years, most were married/cohabitating, and most obtained a secondary education level. The outcome assessment encompassed cardiopulmonary fitness (6-min walk test, *primary outcome*), muscular strength (assessed by handgrip test and timed stand test for upper and lower extremities, respectively), balance (single leg stance test), body composition (percent body fat and BMI), shoulder ROM on the affected side, and health-related quality of life (as indicated by the FACT-B [Functional Assessment of Cancer Therapy – Breast] Trial Outcome Index, FACT-G [Functional Assessment of Cancer Therapy – General] total score, and FACT-B total score). Baseline characteristics were generally comparable between groups, although some numerical differences were observed in handgrip strength.

XSL•FO
RenderX

Table . Baseline characteristics of the participants.

Characteristics	Total (n=24)	Intervention (n=12)	Control (n=12)
Age (years), mean (SD)	57.54 (8.37)	58.58 (8.14)	56.50 (8.83)
Marital status, n (%)			
Married/cohabitating	14 (58.3)	6 (50.0)	8 (66.7)
Divorced/separated/widowed	5 (20.8)	5 (41.7)	0 (0.0)
Others	5 (20.8)	1 (8.3)	4 (33.3)
Education level, n (%)			
Primary	9 (37.5)	6 (50.0)	3 (25.0)
Secondary	12 (50.0)	6 (50.0)	6 (50.0)
Postsecondary	3 (12.6)	0 (0.0)	3 (25.0)
Height (cm), mean (SD)	155.52 (7.83)	154.88 (7.15)	156.17 (8.72)
Weight (kg), mean (SD)	61.56 (13.18)	62.47 (13.35)	60.65 (13.52)
MVPA[a] (min), mean (SD)	46.25 (77.67)	51.67 (97.50)	40.83 (55.18)
6-min walk test (m), mean (SD)	508.54 (76.00)	497.83 (83.33)	519.25 (69.87)
Handgrip test (kg, best), mean (SD)	22.48 (5.18)	25.21 (4.93)	19.75 (3.93)
Handgrip test (kg, affected side)[b], mean (SD)	20.83 (5.64)	24.00 (5.22)	17.67 (4.17)
Timed stand test (seconds), mean (SD)	17.57 (4.47)	17.38 (4.07)	17.79 (5.06)
Single leg stance test (seconds), mean (SD)	39.39 (24.19)	36.97 (22.80)	41.81 (26.28)
Percent body fat (%), mean (SD)	34.47 (6.95)	35.57 (6.23)	33.37 (7.72)
BMI (kg/m^2), mean (SD)	25.33 (4.34)	25.85 (3.84)	24.81 (4.91)
Shoulder ROM[c] (degree, affected side)[b], mean (SD)	165.00 (16.88)	169.58 (10.10)	160.42 (21.16)
Averaged arm circumference (cm, affected side)[b], mean (SD)	23.33 (2.03)	23.95 (1.72)	22.88 (1.66)
FACT-B Trial Outcome Index[d], mean (SD)	53.50 (11.28)	53.17 (9.92)	53.83 (12.95)
FACT-G total score[e], mean (SD)	65.31 (14.57)	63.00 (12.98)	67.61 (16.24)
FACT-B total score[f], mean (SD)	86.81 (18.83)	84.58 (15.75)	89.03 (21.96)

[a]MVPA: moderate to vigorous intensity physical activity.

[b]Calculated and used for these cases.

[c]ROM: range of motion.

[d]Score range: 0 - 96. FACT-B: Functional Assessment of Cancer Therapy – Breast.

[e]Score range: 0 - 108. FACT-G: Functional Assessment of Cancer Therapy – General.

[f]Score range: 0 - 148. FACT-B: Functional Assessment of Cancer Therapy – Breast.

Breast Cancer History and Comorbidities

All participants were diagnosed within 1 year prior to the baseline measurement; laterality was right in 11, left in 10, and bilateral in 3. All invasive tumors were invasive ductal carcinoma, mainly grade 2 or 3. Median invasive tumor size was 18 mm (range: 2 - 65 mm). Local therapy comprised mastectomy in 22 and lumpectomy in 2; axillary surgery was dissection in 14 and sentinel node biopsy in 10; 2 underwent reconstruction. All had received chemotherapy, and 66.7% had received radiotherapy, mostly to the chest wall and regional nodes. Eight were premenopausal, 15 postmenopausal, and 1 had a first-degree family history; all were alive at last follow-up.

Of the 24 participants, 11 (45.8%) had no comorbidities, 5 (20.8%) had one, 3 (12.5%) had two, 3 (8.3% each) had three or four, and 1 (4.2%) had five comorbidities. Hypertension and hyperlipidemia were each present in 6 (25% each); obesity in 5 (20.8%); diabetes in 4 (16.7%); chronic kidney disease and ischemic heart disease in 2 each (8.3% each); and chronic

XSL•FO
RenderX

hepatitis, depression with anxiety, obstructive sleep apnea, anemia, and coronary artery disease in 1 each (4.2% each).

Table 6 presents exploratory between-group comparisons of pre-post change scores. Consistent with the pilot nature of this trial, analyses focused on estimating between-group differences and associated uncertainty rather than formal hypothesis testing for effectiveness. Overall, the direction of effects tended to favor the intervention for the 6-minute walk test (6MWT), timed stand test, single leg stance, affected-side shoulder ROM, and health-related quality-of-life outcomes, indicating outcomes that may be promising to prioritize in a future definitive trial.

Table . Independent-samples *t* test for changes in outcomes (*post minus pre*)[a].

Outcomes	Intervention (n=12)	Control (n=12)	Between-group difference in change (*post minus pre*)				
	Mean (SD)	Mean (SD)	Mean differences	95% CI	*P* value	Cohen *d*	Effect size level[b]
6-min walk test (m, primary outcome)	39.17 (26.19)	24.25 (19.08)	14.92	−4.48 to 34.31	.13	0.651	Medium
Handgrip test (kg, best)	1.08 (1.94)	0.08 (1.82)	1.00	−0.59 to 2.59	.21	0.532	Medium
Handgrip test (kg, affected side)[c]	0.71 (1.78)	0.17 (1.86)	0.54	−1.00 to 2.08	.47	0.298	Small
Timed stand test (seconds)	−2.52 (2.26)	−1.08 (2.40)	−1.44	−3.42 to 0.53	.14	−0.619	Medium
Single leg stance test (seconds)	11.72 (20.01)	3.04 (17.99)	8.69	−7.42 to 24.80	.28	0.457	Small
Percent body fat (%)	−1.19 (1.79)	−0.58 (1.24)	-0.62	−1.92 to 0.69	.23	−0.401	Small
BMI (kg/m^2)	−0.58 (0.70)	−0.28 (0.42)	-0.30	−0.79 to 0.19	.21	−0.522	Medium
Shoulder ROM[d] (degree, affected side)[c]	6.25 (11.31)	−0.83 (10.62)	7.08	−2.21 to 16.37	.13	0.646	Medium
Averaged arm circumference (cm, affected side)[c]	−1.06 (0.80)	−0.66 (0.91)	−0.40	−1.13 to 0.32	.26	−0.470	Small
FACT-B Trial Outcome Index[e]	11.00 (14.18)	1.67 (10.63)	9.33	−1.28 to 19.94	.08	0.745	Medium
FACT-G total score[f]	9.26 (14.45)	−2.28 (13.14)	11.54	−0.15 to 23.24	.05	0.836	Large
FACT-B total score[g]	14.26 (18.68)	−0.78 (17.60)	15.04	−0.33 to 30.41	.06	0.829	Large

[a]Between-group comparisons are exploratory; the study was not powered for effectiveness testing. *P* values are descriptive.

[b]Level: cutoff points of 0.2, 0.5, and 0.8 were used to classify effect sizes into small, medium, and large levels.

[c]Three cases (1 in intervention, 2 in control) had both sides affected by surgery, and then the average was calculated and used for these cases.

[d]ROM: range of motion.

[e]Score range: 0-96. FACT-B: Functional Assessment of Cancer Therapy – Breast.

[f]Score range: 0-108. FACT-G: Functional Assessment of Cancer Therapy – General.

[g]Score range: 0-148. FACT-B: Functional Assessment of Cancer Therapy – Breast.

Absolute effect sizes ranged from 0.298 (handgrip test on the affected side) to 0.836 (FACT-G total score), with half at a medium level, including the primary outcome of the 6-minute walk test (0.651). Based on the effect size of 0.651 for the primary outcome, the required sample size for the main study was calculated to be 102 participants, with 51 participants in each group.

Discussion

Main Findings

This pilot RCT evaluated the feasibility, acceptability, and preliminary effects of a 7-week home-based, group-based tele-exercise intervention (with a progressive transition from

XSL•FO
RenderX

supervised to unsupervised sessions) plus brief psychological counseling among breast cancer survivors undergoing adjuvant therapy. The intervention demonstrated high feasibility, with a recruitment rate of 79.4% and baseline-to-post retention of 100%, and good engagement, with an overall attendance rate of 87.7% and compliance of 85.3% in the intervention group. The intervention was also highly acceptable, with all participants reporting agreement or strong agreement across key acceptability items. Although the study was not powered for definitive effectiveness testing, between-group differences in change generally favored the intervention across cardiopulmonary fitness (6MWT, *primary outcome*) and showed positive trends in lower-extremity strength, balance, shoulder ROM on the affected side, and several health-related quality-of-life outcomes, warranting confirmation in a fully powered trial.

Comparison With Previous Studies

The pilot demonstrated strong feasibility, evidenced by high recruitment and retention rates. These rates appear higher than those reported in many exercise and tele-exercise trials among breast cancer survivors [2,4,5,23], which may reflect the study's pragmatic design and local tailoring (eg, highlighting participant benefits such as free assessments and aligning delivery with local care routines). Engagement was also high: overall attendance was 87.7%, and compliance was 85.3%, comparing favorably with prior exercise interventions [4,5]. The progressive transition from supervised to unsupervised sessions was rated as acceptable (Table 4), and adherence remained broadly comparable between supervised and unsupervised sessions during the intervention period (Table 3 and Figure 3). Figure 3 further suggests that attendance remained stable across the intervention, including after the introduction of unsupervised sessions, which supports the feasibility of reducing supervision without an obvious loss of participation during adjuvant therapy.

In contrast, heart rate–based compliance showed modest session-to-session fluctuations and a slight downward tendency over time. One possible explanation is physiological adaptation: as participants became more familiar with the prescribed workload, the same exercises may have elicited a lower heart-rate response, reducing the likelihood of meeting the original target heart-rate zone despite continued participation [24,25]. This suggests that more dynamic adjustments to the exercise prescription will be necessary in the definitive trial. The brief psychological counseling component may also have helped participants sustain engagement by supporting motivation and self-regulation, although its specific contribution cannot be isolated in this pilot. Tele-exercise delivery at home likely reduced common participation barriers (eg, fatigue, transportation, and time constraints), although longer-term maintenance beyond the intervention period cannot be inferred without follow-up.

Preliminary between-group estimates generally favored the intervention across cardiopulmonary fitness, lower-limb function, balance, shoulder ROM, and health-related quality of life, consistent with meta-analytic evidence in breast cancer survivors [2,4,5,26]. The estimated effect on 6MWT (Cohen d=0.65; 95% CI includes 0) suggests the primary outcome is promising to prioritize in a definitive trial. While functional capacity is clinically relevant, this pilot cannot establish clinical effectiveness or downstream outcomes (eg, morbidity/mortality) [27]. From a clinical perspective, the between-group difference in change in 6MWT in this 7-week pilot was approximately 15 m, which is below commonly cited minimal clinically important differences (MCIDs) of 20 - 30 m in oncology populations [28]; this might be due to the relatively short duration of the intervention. Therefore, these findings should be interpreted as exploratory; the definitive 12-week trial is needed to determine whether the intervention can achieve clinically meaningful improvements in functional walking capacity.

Changes in physical function were also encouraging. Improvements in the timed stand test and single-leg stance are plausible given the repeated lower-body functional tasks and progressive loading embedded in the sessions. For the timed stand test, lower values indicate faster completion and better performance. These measures are clinically relevant because lower-limb weakness and balance impairment can limit mobility and confidence in daily activities. Shoulder ROM on the affected side also improved, which is notable given the frequency of posttreatment shoulder stiffness; embedding brief mobility work within each session may be a pragmatic way to address this without substantially increasing intervention burden.

Upper-extremity strength (handgrip strength) showed only small between-group differences, which may reflect limited progressive upper-body loading in the 7-week protocol and the need for more targeted upper-body resistance progression training in the main trial. BMI and percent body fat decreased modestly; however, body composition was not a primary target, and large changes were not expected over this short, nondiet intervention.

Quality-of-life measures (FACT-B Trial Outcome Index, FACT-G, and FACT-B total score) also tended to improve in the intervention group. Importantly, the between-group differences in change were approximately 9.33 (FACT-B Trial Outcome Index), 11.54 (FACT-G), and 15.04 (FACT-B total), which exceed commonly cited MCIDs for these scales (FACT-B Trial Outcome Index: 5 - 6 points; FACT-G: 5 - 6 points; FACT-B total: 7 - 8 points) [29]. These preliminary clinically meaningful improvements may reflect both gains in physical function from the combined aerobic-resistance training and the brief behavioral counseling elements (eg, motivation, self-regulation, and habit-formation strategies), which may have helped participants manage barriers and sustain activity. Given the pilot design, findings are exploratory and require confirmation in the definitive trial.

Finally, we did not observe lymphedema exacerbation over the intervention period, and perceived safety ratings were high, which is reassuring and consistent with prior evidence that appropriately prescribed exercise is safe for individuals with breast cancer-related lymphedema [27].

Strengths and Limitations

This pilot study has several strengths, including a randomized design, blinded outcome assessment, the use of both objective functional outcomes and patient-reported health-related quality

XSL•FO
RenderX

of life measures, and a pragmatic tele-exercise intervention that is feasible to deliver during adjuvant therapy.

Several limitations should be considered when interpreting the findings. First, as a pilot study, the trial was designed to evaluate feasibility and acceptability and was not powered to draw definitive conclusions about effectiveness. Accordingly, the small sample (n=24) from a single center in Hong Kong limited the precision of estimates and the generalizability of the findings. Second, the pilot intervention period was limited to 7 weeks, as time and resource constraints did not allow completion of the planned 12-week intervention for the main study. This reduced intervention dose may have limited changes in outcomes and may have led to an underestimation of the potential effectiveness and effect sizes achievable with the intended intervention duration. Attendance and compliance observed over 7 weeks may not be representative of adherence over the planned 12-week intervention, as participation often declines with longer intervention duration. Third, follow-up assessments were not conducted; therefore, the longer-term maintenance of behavior change and durability of outcome improvements could not be evaluated. Fourth, eligibility criteria (language and smartphone use) may limit generalizability; future studies should broaden inclusion and provide support for participants with lower digital access and literacy. Fifth, randomization occurred before baseline assessment. This design may increase postallocation withdrawal due to treatment preference and can introduce selection bias into the analyzed sample. In the future study, baseline assessments will be completed prior to randomization (or, alternatively, randomization will be retained but allocation will be disclosed only after baseline assessment) to minimize postrandomization attrition.

Implications for the Main Trial and Future Directions

The results of this pilot study provide a foundation for the design and implementation of a larger-scale main study. Based on the observed primary outcome effect size, the estimated sample size for the main study is 102 participants (51 per group). Future research should (1) include postintervention follow-up to evaluate whether benefits persist after supervision is reduced, (2) strengthen fidelity and adherence monitoring and reporting, and (3) consider enhancing upper-body progressive resistance training if upper-limb strength is a priority outcome. In addition, the integration of psychological theories and BCTs in the intervention design may have contributed to habit formation and sustained exercise participation. These elements should be refined and tailored to address the needs and barriers faced by breast cancer survivors across different contexts and settings.

Conclusions

This pilot RCT demonstrates that a home-based, group-based tele-exercise intervention with brief psychological counseling is feasible and acceptable for breast cancer survivors undergoing adjuvant therapy, with high engagement and no apparent safety concerns. Exploratory between-group estimates for cardiopulmonary fitness, functional performance, affected-side shoulder ROM, and health-related quality of life (effect sizes and 95% CIs) provide parameters to inform the design of a fully powered RCT with longer follow-up to evaluate effectiveness and sustainability.

Acknowledgments

We sincerely thank all the participants who took part in this study for their time and commitment, which made this research possible. We also thank Ms Luyi Mai and Ms Zhennan Xiong for their assistance with data collection. Generative AI was used for English grammar checking only; the authors take full responsibility for the final content.

Funding

This study was supported by the Research Support Fund of the Department of Sports and Health Sciences, Academy of Wellness and Human Development, Hong Kong Baptist University. The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

Deidentified individual participant data and supporting materials are available from the corresponding author (YG) upon reasonable request, subject to approval by the study sponsor or ethics requirements and execution of a data sharing agreement to protect participant privacy.

Authors' Contributions

Conceptualization: YG
Data curation: XW
Formal analysis: YS
Funding acquisition: YG
Investigation: YS, BKYF, OKC, WWSF
Methodology: YG, YS, XW, BKYF, OKC, WWSF
Supervision: YG
Writing – original draft: YS
Writing – review & editing: YG, YS, XW, BKYF, OKC, WWSF
All authors read and approved the final manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed tele-exercise protocol, scheduling, and psychological counseling sessions.
[DOCX File, 35 KB - jmir_v28i1e79564_app1.docx]

Multimedia Appendix 2
Content of session 1 with psychological strategies (as an example) in Chinese.
[PDF File, 261 KB - jmir_v28i1e79564_app2.pdf]

Multimedia Appendix 3
An example essay (in Chinese).
[PDF File, 258 KB - jmir_v28i1e79564_app3.pdf]

Checklist 1
CONSORT checklist.
[PDF File, 152 KB - jmir_v28i1e79564_app4.pdf]

References

1. Breast cancer. World Health Organization. 2026 Apr 16. URL: https://www.who.int/news-room/fact-sheets/detail/breast-cancer [accessed 2026-06-01]
2. Lahart IM, Metsios GS, Nevill AM, Carmichael AR. Physical activity for women with breast cancer after adjuvant therapy. Cochrane Database Syst Rev 2018 Jan 29;1(1):CD011292. [doi: 10.1002/14651858.CD011292.pub2] [Medline: 29376559]
3. Breast cancer. Centre for Health Protection. 2026 Jan 23. URL: https://www.chp.gov.hk/en/healthtopics/content/25/53.html [accessed 2026-06-01]
4. Lipsett A, Barrett S, Haruna F, Mustian K, O'Donovan A. The impact of exercise during adjuvant radiotherapy for breast cancer on fatigue and quality of life: a systematic review and meta-analysis. Breast 2017 Apr;32:144-155. [doi: 10.1016/j.breast.2017.02.002] [Medline: 28189100]
5. Markes M, Brockow T, Resch KL. Exercise for women receiving adjuvant therapy for breast cancer. Cochrane Database Syst Rev 2006 Oct 18;4(4):CD005001. [doi: 10.1002/14651858.CD005001.pub2] [Medline: 17054230]
6. Martins AD, Brito JP, Oliveira R, et al. Relationship between heart rate variability and functional fitness in breast cancer survivors: a cross-sectional study. Healthcare (Basel) 2021 Sep 13;9(9):1205. [doi: 10.3390/healthcare9091205] [Medline: 34574979]
7. Arab C, Dias DPM, Barbosa RTDA, et al. Heart rate variability measure in breast cancer patients and survivors: a systematic review. Psychoneuroendocrinology 2016 Jun;68:57-68. [doi: 10.1016/j.psyneuen.2016.02.018] [Medline: 26943345]
8. Hayes S SK. Is unsupervised exercise following breast cancer safe for all women? Int J Phys Med Rehabil 2014;02(03):1-8. [doi: 10.4172/2329-9096.1000197]
9. McNamara RJ, Dale M, McKeough ZJ. Innovative strategies to improve the reach and engagement in pulmonary rehabilitation. J Thorac Dis 2019 Oct;11(Suppl 17):S2192-S2199. [doi: 10.21037/jtd.2019.10.29] [Medline: 31737346]
10. Pudkasam S, Polman R, Pitcher M, et al. Physical activity and breast cancer survivors: importance of adherence, motivational interviewing and psychological health. Maturitas 2018 Oct;116:66-72. [doi: 10.1016/j.maturitas.2018.07.010] [Medline: 30244781]
11. Jiang Y, Liu F, Guo J, et al. Evaluating an intervention program using WeChat for patients with chronic obstructive pulmonary disease: randomized controlled trial. J Med Internet Res 2020 Apr 21;22(4):e17089. [doi: 10.2196/17089] [Medline: 32314971]
12. Gao Y, Zhong LLD, Quach B, et al. COVID-19 rehabilitation with herbal medicine and cardiorespiratory exercise: protocol for a clinical study. JMIR Res Protoc 2021 May 26;10(5):e25556. [doi: 10.2196/25556] [Medline: 33970864]
13. Sun Y, Wang A, Yu S, et al. A blended intervention to promote physical activity, health and work productivity among office employees using intervention mapping: a study protocol for a cluster-randomized controlled trial. BMC Public Health 2020 Jun 25;20(1):994. [doi: 10.1186/s12889-020-09128-z] [Medline: 32586293]
14. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
15. Sun Y, Gao Y, Ou AYT, et al. Effectiveness of a blended intervention to promote physical activity among office employees: randomized controlled trial. J Med Internet Res 2026;28:e80249. [doi: 10.2196/80249] [Medline: 42172646]

XSL•FO
RenderX

16. Riebe D, Ehrman JK, Liguori G, Magal M. ACSM's Guidelines for Exercise Testing and Prescription, 10th edition: Wolters Kluwer; 2018.
17. Campbell KL, Winters-Stone KM, Wiskemann J, et al. Exercise guidelines for cancer survivors: consensus statement from international multidisciplinary roundtable. Med Sci Sports Exerc 2019 Nov;51(11):2375-2390. [doi: 10.1249/MSS.0000000000002116] [Medline: 31626055]
18. Pfledderer CD, von Klinggraeff L, Burkart S, Wolfenden L, Ioannidis JPA, Beets MW. Feasibility indicators in obesity-related behavioral intervention preliminary studies: a historical scoping review. Pilot Feasibility Stud 2023 Mar 22;9(1):46. [doi: 10.1186/s40814-023-01270-w] [Medline: 36949541]
19. Sibley KM, Straus SE, Inness EL, Salbach NM, Jaglal SB. Clinical balance assessment: perceptions of commonly-used standardized measures and current practices among physiotherapists in Ontario, Canada. Implement Sci 2013 Mar 20;8(1):33. [doi: 10.1186/1748-5908-8-33] [Medline: 23510277]
20. Yuan RZ, Li KP, Wei XL, et al. Effects of free range-of-motion upper limb exercise based on mirror therapy on shoulder function in patients after breast cancer surgery: study protocol for a randomized controlled trial. Trials 2021 Nov 17;22(1):815. [doi: 10.1186/s13063-021-05789-2] [Medline: 34789309]
21. Eldridge SM, Chan CL, Campbell MJ, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016 Oct 24;355:i5239. [doi: 10.1136/bmj.i5239] [Medline: 27777223]
22. Cohen J. Statistical Power Analysis for the Behavioral Sciences: Routledge; 2013. [doi: 10.4324/9780203771587]
23. Galiano-Castillo N, Cantarero-Villanueva I, Fernández-Lao C, et al. Telehealth system: a randomized controlled trial evaluating the impact of an internet-based exercise intervention on quality of life, pain, muscle strength, and fatigue in breast cancer survivors. Cancer 2016 Oct 15;122(20):3166-3174. [doi: 10.1002/cncr.30172] [Medline: 27332968]
24. American College of Sports Medicine. ACSM's Exercise Testing and Prescription: Lippincott Williams & Wilkins; 2017.
25. Mytinger M, Nelson RK, Zuhl M. Exercise prescription guidelines for cardiovascular disease patients in the absence of a baseline stress test. J Cardiovasc Dev Dis 2020 Apr 27;7(2):15. [doi: 10.3390/jcdd7020015] [Medline: 32349219]
26. Schmitz KH, Campbell AM, Stuiver MM, et al. Exercise is medicine in oncology: engaging clinicians to help patients move through cancer. CA Cancer J Clin 2019 Nov;69(6):468-484. [doi: 10.3322/caac.21579] [Medline: 31617590]
27. Jones LW, Habel LA, Weltzien E, et al. Exercise and risk of cardiovascular events in women with nonmetastatic breast cancer. J Clin Oncol 2016 Aug 10;34(23):2743-2749. [doi: 10.1200/JCO.2015.65.6603] [Medline: 27217451]
28. Bohannon RW, Crouch R. Minimal clinically important difference for change in 6-minute walk test distance of adults with pathology: a systematic review. J Eval Clin Pract 2017 Apr;23(2):377-381. [doi: 10.1111/jep.12629] [Medline: 27592691]
29. Eton DT, Cella D, Yost KJ, et al. A combination of distribution- and anchor-based approaches determined minimally important differences (MIDs) for four endpoints in a breast cancer scale. J Clin Epidemiol 2004 Sep;57(9):898-910. [doi: 10.1016/j.jclinepi.2004.01.012] [Medline: 15504633]

Abbreviations

6MWT: 6-minute walk test
BCT: behavior change technique
CONSORT: Consolidated Standards of Reporting Trials
FACT-B: Functional Assessment of Cancer Therapy – Breast
FACT-G: Functional Assessment of Cancer Therapy – General
HRR: heart rate reserve
ICH-GCP: International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use – Good Clinical Practice
MCID: minimal clinically important difference
mITT: modified intention-to-treat
RCT: randomized controlled trial
RE-AIM: reach, effectiveness, adoption, implementation, and maintenance
ROM: range of motion
SNOSE: sequentially numbered, opaque, sealed envelopes
SpO2: oxygen saturation
SSL/TLS: Secure Sockets Layer/Transport Layer Security

XSL•FO
RenderX

Edited by M Balcarras; submitted 25.Jun.2025; peer-reviewed by A Mohammed, I Adefolaju; revised version received 25.Apr.2026; accepted 27.Apr.2026; published 26.Jun.2026.

Please cite as:
Gao Y, Sun Y, Wang X, Chun OK, Fung BKY, Fung WWS
Effectiveness of a Home-Based and Group-Based Tele-Exercise Program for Breast Cancer Survivors: Pilot Randomized Controlled Trial
J Med Internet Res 2026;28:e79564
URL: https://www.jmir.org/2026/1/e79564
doi:10.2196/79564

© Yang Gao, Yan Sun, Xueying Wang, Oi Kwan Chun, Brigitte Kim Yook Fung, Wai Wun Sara Fung. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Literacy and Patient Satisfaction in Telemedicine Follow-Up With In-Person App Instruction Versus Outpatient Department Follow-Up After Upper-Extremity Surgery: A Randomized Controlled Trial

Werapat Ngowroongrueng[1*], BEng, MD; Sitthiphong Suwannaphisit[1*], MD; Chanon Malaikritsanachalee[1*], MD; Sutee Thaveepunsan[1*], MD; Sirisak Chaitantipongse[1*], MD; Theephop Teeragananan[1*], MD; Kanokporn Imsakul[2*], BSN, MBA; Supaira Puntumabumrung[2*], BEd, MPA; Krit Prasittichok[2*], MSc, PhD

[1]Department of Orthopaedics, Faculty of Medicine, Vajira Hospital, Navamindradhiraj University, 681 Samsen Road, Dusit, Bangkok, Thailand

[2]Faculty of Medicine Vajira Hospital, Navamindradhiraj University, Bangkok, Thailand

[*]all authors contributed equally

Corresponding Author:
Sitthiphong Suwannaphisit, MD
Department of Orthopaedics, Faculty of Medicine, Vajira Hospital, Navamindradhiraj University, 681 Samsen Road, Dusit, Bangkok, Thailand

Abstract

Background: Telemedicine is increasingly used for postoperative follow-up in orthopedic surgery, but concerns remain regarding whether patients' digital literacy influences satisfaction with virtual care.

Objective: This study aimed to compare patient satisfaction between telemedicine and in-person outpatient department (OPD) follow-up after upper-extremity surgery and to examine the association between digital literacy and patient satisfaction.

Methods: This single-center, open-label randomized controlled trial enrolled adults undergoing nontraumatic hand or upper-extremity surgery. After trial registration, 70 participants were randomized in a 1:1 ratio to telemedicine (n=35, 50%) or in-person OPD follow-up (n=35, 50%) using block randomization with a fixed block size of 4. All participants received standardized postoperative education at the 2-week visit. Participants in the telemedicine group additionally received brief in-person instruction on the hospital-based Vajira@Home telemedicine app before completing a scheduled video-based follow-up at 6 weeks. Patient satisfaction was assessed at 2 and 6 weeks using a validated 6-item Likert scale questionnaire covering communication, convenience, understanding of care, and overall satisfaction. Digital literacy was measured at baseline using a standardized questionnaire. Between-group comparisons were performed using appropriate parametric or nonparametric tests.

Results: Participants in the telemedicine group were younger than those in the outpatient department (OPD) group (mean 53.3, SD 11.5 years vs mean 59.7, SD 14.2 years) and had higher digital literacy scores (mean 12.0, SD 3.5 vs mean 10.2, SD 2.5). Overall satisfaction scores were high in both groups at 2 and 6 weeks postoperatively (median 5, IQR 5-5 at both time points). No statistically significant differences were observed between groups across any satisfaction domains, including communication (P=.27), convenience (P=.41), and overall satisfaction (P=.35). Satisfaction scores remained stable over time within each group. Digital literacy was not associated with patient satisfaction.

Conclusions: Telemedicine follow-up supported by brief in-person app instruction resulted in patient satisfaction that was not significantly different from traditional in-person outpatient follow-up after upper-extremity surgery. Although age and digital literacy differed modestly between groups, these factors did not influence satisfaction. These findings support telemedicine as an acceptable option for postoperative follow-up in appropriately selected patients.

Trial Registration: Thai Clinical Trials Registry TCTR20250528008; https://trialsearch.who.int/Trial2.aspx?TrialID=TCTR20250528008

(J Med Internet Res 2026;28:e86918) doi:10.2196/86918

KEYWORDS

telemedicine; digital literacy; hand surgery; postoperative care; patient satisfaction; urban areas; Thailand

XSL•FO
RenderX

Introduction

The rapid growth of telemedicine has transformed health care delivery by enabling remote consultations, expanding accessibility, and reducing travel-related burdens for patients [1-3]. In orthopedic surgery, virtual follow-up represents a practical and increasingly favored alternative to conventional outpatient visits, particularly for postoperative assessment and rehabilitation [4-7]. Previous studies [8,9] have shown that telemedicine can maintain patient safety and satisfaction while reducing health care costs and alleviating clinic overcrowding. However, the effectiveness and acceptability of telemedicine depend not only on technological infrastructure but also on patients' ability to engage confidently and effectively with digital platforms [10,11].

Digital literacy, defined as the ability to use digital technologies and critically interpret online information, has emerged as a key factor influencing the success of telemedicine [12]. Although smartphone ownership and internet connectivity have expanded globally, disparities in digital competence persist, especially among older adults and individuals with lower educational attainment [13]. In orthopedic practice, such disparities may hinder equitable access to postoperative care and slow the broader adoption of digital health solutions. Prior research [11] has shown that patients with strong health literacy but limited digital literacy often face challenges when navigating telemedicine systems. Nevertheless, the relationship between digital literacy and patient satisfaction in surgical populations, particularly within middle-income countries, remains insufficiently explored.

Thailand, an emerging digital economy in Southeast Asia, provides an appropriate context in which to investigate this relationship. As health care systems increasingly integrate telemedicine into routine care, understanding whether patient satisfaction differs between virtual and in-person follow-up, and how digital literacy shapes this experience, is essential for guiding clinical practice and health policy [14]. Accordingly, this randomized controlled trial aimed to compare patient satisfaction between telemedicine and in-person postoperative follow-up after upper-extremity surgery in an urban Thai population and to examine the association between digital literacy and patient satisfaction. We hypothesized that telemedicine would result in no significant difference in patient satisfaction compared with traditional in-person follow-up and that higher levels of digital literacy would be associated with greater patient satisfaction.

Methods

Ethical Considerations

This study was approved by the institutional review board of Vajira Hospital, Faculty of Medicine Vajira Hospital, Navamindradhiraj University (COA 012/2568) and conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all participants prior to enrollment.

All data were deidentified before analysis and stored in secure, password-protected hospital systems accessible only to authorized researchers. Telemedicine visits were conducted using a secure, hospital-based platform.

Participants received a modest compensation of THB 100 (US $2.7) upon completion of the 6-week postoperative follow-up visit.

Trial Registration

The study protocol was retrospectively registered with the Thai Clinical Trials Registry (TCTR20250528008) on May 28, 2025, after participant enrollment had begun but before completion of recruitment. Participant screening and enrollment occurred between February and August 2025, with the first participant enrolled on February 2, 2025. No changes were made to the study objectives, outcomes, or planned analyses after trial initiation or registration.

The study protocol, including the study objectives, eligibility criteria, outcome measures, and planned statistical analyses, had been finalized prior to the enrollment of the first participant and remained unchanged throughout the study. Although trial registration occurred after enrollment had begun due to administrative delays, no modifications were made to the study design, outcomes, or analyses after trial initiation or registration.

Study Design

This study was a single-center, open-label, parallel-group randomized controlled trial with a 1:1 allocation ratio comparing telemedicine follow-up with conventional in-person outpatient department (OPD) follow-up after nontraumatic hand and upper-extremity surgery. The trial was conducted at the Department of Orthopedics, Faculty of Medicine Vajira Hospital, Navamindradhiraj University, Bangkok, Thailand. Randomization was performed by an independent coordinator using a computerized block randomization sequence with a fixed block size of 4 to ensure balanced group allocation.

Participants

Adults aged 18 to 80 years who underwent elective, nontraumatic hand or upper-extremity surgery were eligible for inclusion. Diagnoses were categorized using *International Classification of Diseases, Tenth Revision* (*ICD-10*) codes related to upper-extremity conditions and procedures involving the shoulder, arm, elbow, forearm, wrist, and hand. Eligible participants resided in Bangkok or neighboring urban provinces (Nonthaburi, Pathum Thani, Nakhon Pathom, Samut Prakan, Samut Sakhon, and Chachoengsao) and had access to a smartphone, tablet, or computer with a stable internet connection to support telemedicine follow-up.

Exclusion criteria included (1) refusal to participate, (2) conditions requiring radiographic evaluation or direct physical examination at follow-up, as determined by the treating surgeon, and (3) cognitive, visual, or physical limitations that would prevent completion of telemedicine or questionnaire-based assessments. All diagnoses and eligibility determinations were confirmed by a board-certified orthopedic hand surgeon.

Participants were informed that they could withdraw from the study at any time, and those in the telemedicine group could be

XSL•FO
RenderX

transitioned to in-person care if clinically indicated by their attending physician.

Intervention and Follow-Up Protocol

Participants were randomized to receive either telemedicine or in-person outpatient (OPD) follow-up at 2 and 6 weeks postoperatively. All participants received standardized postoperative education during their 2-week clinic visit, including wound care instructions, rehabilitation guidance, and information on potential complications.

Participants allocated to the telemedicine group received an additional structured, in-person training session provided by the Department of Telemedicine at the Faculty of Medicine Vajira Hospital. This training covered installation and login procedures for Vajira@Home, the hospital's dedicated telemedicine app, as well as demonstrations on how to navigate the interface, position the camera for optimal wound visualization, and complete a secure virtual follow-up visit (Multimedia Appendix 1). This brief training ensured that participants were capable of performing their telemedicine consultation independently from home using their own devices and internet connection.

At the 6-week follow-up, participants in the telemedicine group completed their visit through a secure hospital-based videoconferencing platform integrated within the Vajira@Home app. Participants in the OPD group returned for a conventional in-person clinic visit. Both groups followed identical clinical assessment protocols, including history taking, evaluation of wound healing, visual or in-person physical examination (with range-of-motion assessment when appropriate), and standardized documentation. Any findings requiring urgent or additional care were managed according to routine postoperative practice. Telemedicine participants could be converted to in-person evaluation at the discretion of the attending hand surgeon.

Outcome Measures

The primary outcome was overall patient satisfaction at 6 weeks postoperatively, assessed using a validated 6-item Likert scale questionnaire (1=strongly disagree to 5=strongly agree). Satisfaction was also assessed at 2 weeks to evaluate changes over time. The questionnaire covered 4 domains: communication quality, convenience, understanding of postoperative care, and overall satisfaction. Domain-specific and total scores were analyzed, with higher scores indicating greater satisfaction.

The secondary outcome was digital literacy, measured at baseline using a standardized instrument adapted from the Digital Literacy Scale [15]. Participants rated three items assessing (1) fluency in using digital devices, (2) ability to use technology for telemedicine purposes, and (3) comfort communicating with physicians through digital tools. Each item was scored on a 5-point Likert scale, yielding a total score ranging from 3 to 15, with higher scores indicating greater digital literacy. Total scores were also converted to percentage values and categorized as poor (<60%), moderate (60% - 80%), or good (>80%) according to Bloom taxonomy thresholds.

The questionnaire was translated from English into Thai by certified professional translators and reviewed by English-language faculty members to ensure linguistic accuracy and conceptual equivalence. Internal consistency reliability of the adapted instrument was assessed using Cronbach alpha, demonstrating good reliability.

Sample Size Calculation

Given the exploratory nature of this study and the absence of prior effect size estimates in comparable populations, a formal a priori sample size calculation based on a predefined effect size or noninferiority margin was not performed. Instead, a pragmatic sample size approach was adopted, informed by previously published telemedicine studies in orthopedic postoperative care that consistently reported high levels of patient satisfaction with minimal between-group differences.

Based on a prior randomized study by Cha et al [16] evaluating telemedicine follow-up after shoulder arthroscopy, overall patient satisfaction rates exceeding 90% were observed in both telemedicine and in-person groups. Assuming similarly high satisfaction proportions and small absolute differences between groups, a total sample size of approximately 64 participants (32 per group) was considered sufficient to allow exploratory between-group comparisons at a 2-sided significance level of .05.

To account for a potential dropout rate of approximately 5%, the target sample size was increased to 70 participants, with 35 participants allocated to each group.

Post hoc consideration suggests that, given the observed ceiling effects in satisfaction scores, the study may have been underpowered to detect small but clinically meaningful differences between follow-up modalities. Assuming a small effect size (Cohen *d* of approximately 0.3), a substantially larger sample size would have been required to achieve 80% statistical power. Therefore, the findings should be interpreted as exploratory, and larger studies are warranted to confirm these results.

Data Collection and Risk Mitigation

Sociodemographic data collected at enrollment included age, sex, education, occupation, income, residence, and health care coverage. Technical readiness (device ownership, internet access, and privacy for consultations) was also recorded. To minimize risks related to privacy breaches, all data were stored in encrypted, password-protected hospital systems accessible only to authorized researchers. Participants received training in the use of the telemedicine app to prevent errors or delays during virtual sessions.

Blinding

This study was a nonblinded randomized controlled trial, as blinding of participants and clinicians was not feasible due to the nature of the interventions. Participants were aware of their assigned follow-up modality (telemedicine or in-person OPD), and the treating orthopedic surgeons could not be blinded because they directly conducted both types of follow-up assessments.

To minimize potential bias, both groups received identical postoperative education at the 2-week visit, and follow-up

XSL•FO
RenderX

assessments followed a standardized clinical protocol, including structured history taking, wound evaluation, range-of-motion assessment when appropriate, and uniform documentation. Patient satisfaction and digital literacy outcomes were collected using validated, self-administered questionnaires, reducing the risk of measurement bias.

Statistical Analysis

Data were analyzed using SPSS Statistics (version 30; IBM Corp) [17]. Continuous variables are presented as mean (SD) or median (IQR), depending on data distribution. Normality was assessed using the Shapiro-Wilk test. Between-group comparisons of continuous variables were performed using the 2-tailed independent-samples *t* test or the Mann-Whitney *U* test, as appropriate. Categorical variables were analyzed using the chi-square test or Fisher exact test. Because questionnaire responses were measured using Likert scales, nonparametric tests were applied for these variables. Between-group comparisons were conducted using the Mann-Whitney *U* test, and within-group comparisons between week 2 and week 6 satisfaction scores were assessed using the Wilcoxon signed-rank test. A 2-sided P value <.05 was considered statistically significant. All analyses followed the intention-to-treat principle.

Results

Overview

Participants were recruited between February and August 2025, during which all scheduled 2-week and 6-week follow-up assessments were completed. No external or secular events occurred during the study period that could have influenced recruitment, access to telemedicine services, or follow-up procedures. Of the 75 patients screened for eligibility, 93.3% (n=70) were enrolled and randomized equally to the telemedicine group (n=35, 50%) and the OPD group (n=35, 50%). Participant flow through the trial, including enrollment, allocation, follow-up, and analysis, is presented in Figure 1. No participants were lost during follow-up. All randomized participants completed their assigned follow-up modality at both time points, resulting in consistent denominators (n=70) for all primary analyses. Because all participants adhered to and used their assigned follow-up modality as intended (telemedicine: 35/35, 100% virtual visits; OPD: 35/35, 100% in-person visits), analyses were conducted according to the intention-to-treat principle, and no additional per-protocol or user-only analyses were required.

Baseline characteristics are presented in Table 1. Participants in the telemedicine group were younger on average than those in the OPD group (mean 53.3, SD 11.5 vs mean 59.7, SD 14.2 y) and had slightly higher digital literacy scores. Other demographic and clinical variables—including sex, diagnosis, geographic origin, income, employment, and health care coverage—were broadly similar between groups. Educational attainment differed modestly, with a greater proportion of participants in the OPD group holding undergraduate or postgraduate degrees. Consultation duration appeared longer in the telemedicine group (median 15, IQR 11.25-18.75 min) than in the OPD group (median 10, IQR 7.5-12.5 min).

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram. OPD: outpatient department.

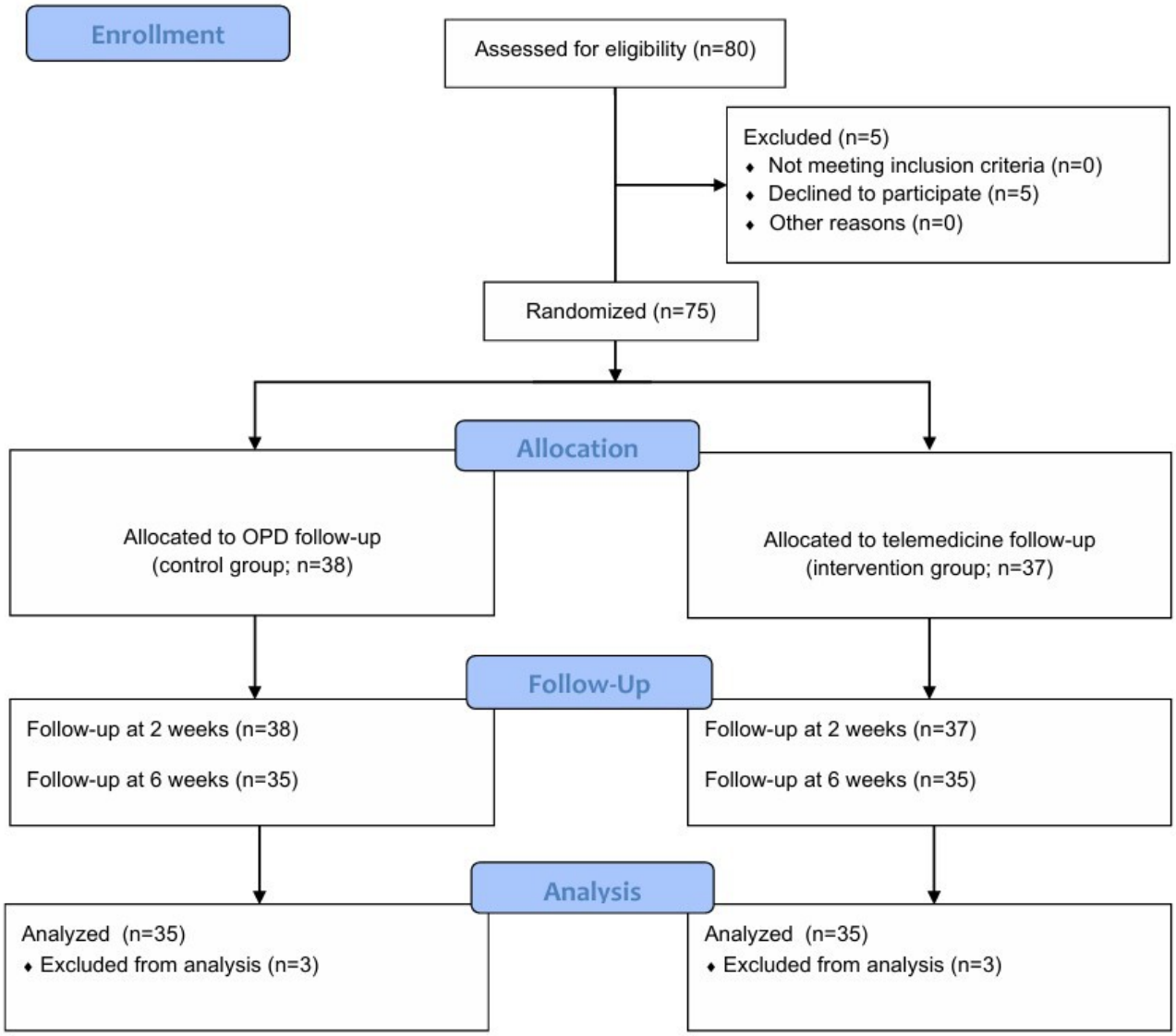

XSL•FO
RenderX

Table . Baseline characteristics and technical access of study participants[a].

Characteristic	Telemedicine group (n=35)	OPD[b] group (n=35)
Age (y), mean (SD)	53.3 (11.5)	59.7 (14.2)
Male, n (%)	12 (34.3)	15 (42.9)
Female, n (%)	23 (65.7)	20 (57.1)
Carpal tunnel syndrome, n (%)	11 (31.4)	8 (22.9)
Trigger finger, n (%)	13 (37.1)	17 (48.6)
De Quervain tenosynovitis, n (%)	4 (11.4)	1 (2.9)
Other diagnoses, n (%)	7 (20)	9 (25.7)
Bangkok, n (%)	27 (77.1)	22 (63.9)
Metropolitan area outside Bangkok, n (%)	8 (22.9)	13 (37.1)
High school education or lower, n (%)	17 (48.6)	11 (31.4)
Undergraduate education or higher, n (%)	18 (51.4)	24 (68.6)
Universal Coverage Scheme, n (%)	8 (22.9)	9 (25.7)
Social Security Scheme, n (%)	13 (37.1)	4 (11.4)
Civil Service Medical Benefits Scheme, n (%)	10 (28.6)	16 (45.7)
Group or private insurance, n (%)	2 (5.7)	2 (5.7)
None, n (%)	2 (5.7)	4 (11.4)
Employed participants, n (%)	28 (80)	23 (65.7)
Unemployed participants, n (%)	7 (20)	12 (34.3)
Consultation duration (min), median (IQR)	15 (7.5)	10 (5.0)
Own a computer, n (%)	17 (48.6)	21 (60)
Own a smartphone, n (%)	35 (100)	35 (100)
Own a tablet, n (%)	13 (37.1)	15 (42.9)
Wi-Fi access at home, n (%)	32 (91.4)	32 (91.4)
Video chat device availability, n (%)	34 (97.1)	31 (88.6)
Able to log in successfully, n (%)	32 (91.4)	31 (88.6)
Private space for visit, n (%)	34 (97.1)	33 (94.3)

[a]Statistical tests were performed for between-group comparisons. Continuous variables were analyzed using the independent-samples *t* test or Mann-Whitney *U* test, as appropriate, and categorical variables were analyzed using the chi-square test or Fisher exact test. Corresponding *P* values are not presented in this table.

[b]OPD: outpatient department.

Digital Literacy and Technological Readiness

All participants owned smartphones and had stable internet access. Ownership of computers and tablets was similar between groups. As shown in Table 1, most participants reported having home Wi-Fi access and a private space suitable for virtual consultations (telemedicine group: 34/35, 97.1%; OPD group: 33/35, 94.3%; *P*=.10). Successful login to the hospital's telemedicine platform was achieved by 91.4% (32/35) of participants in the telemedicine group and 88.6% (31/35) in the OPD group (*P*=.56).

Although the categorical distribution of digital literacy levels did not differ significantly between groups (*P*=.10), the telemedicine group had a significantly higher mean digital literacy score than the OPD group (mean 12.0, SD 3.5 vs mean 10.2, SD 2.5; *P*=.02; Table 2).

Age-stratified analyses of individual digital literacy items showed no statistically significant differences between groups across all domains, including device fluency, telemedicine ability, and comfort communicating with physicians (all *P*>.05; Table 3).

XSL•FO
RenderX

Table . Digital literacy levels[a].

Variable	Telemedicine group, n (%)[b]	Outpatient department group, n (%)[c]	*P* value[d]
Poor (score 3-8)	6 (17.1)	6 (17.1)	—[e]
Moderate (score 9-11)	8 (22.9)	16 (45.7)	—
Good (score 12-15)	21 (60.0)	13 (37.1)	—

[a]Digital literacy was assessed using a 3-item Likert scale (1=strongly disagree to 5=strongly agree; total score range 3 - 15), with higher scores indicating greater digital literacy. Total scores were analyzed as continuous variables using independent-samples *t* tests. Individual Likert scale items were compared using the Mann-Whitney *U* test. Categorical variables were analyzed using the chi-square test or Fisher exact test, as appropriate.

[b]Total score: mean 12.0 (SD 3.5).

[c]Total score: mean 10.2 (SD 2.5).

[d]Total score: *P*=.02.

[e]Not applicable because at least one group contained no participants in this category, preventing statistical comparison.

Table . Age-stratified digital literacy scores[a].

Item	Telemedicine group (<60 y), mean (SD)	OPD[b] group (<60 y), mean (SD)	*P* value	Telemedicine group (≥60 y), mean (SD)	OPD group (≥60 y), mean (SD)	*P* value
Device fluency	4.00 (1.12)	3.94 (0.90)	.83	3.30 (1.70)	3.17 (1.16)	.91
Telemedicine ability	4.00 (1.15)	3.65 (0.86)	.85	3.50 (1.72)	3.06 (1.14)	.88
Comfort communicating	4.36 (1.19)	4.00 (0.79)	.76	4.20 (1.32)	3.47 (1.28)	.86

[a]Individual Likert scale items were compared using the Mann-Whitney *U* test. Categorical variables were analyzed using the chi-square test or Fisher exact test, as appropriate.

[b]OPD: outpatient department.

Patient Satisfaction

Table 4 Patient satisfaction remained uniformly high at both postoperative follow-up assessments. At 2 weeks, the median overall satisfaction score was 5 (IQR 5-5) in both the telemedicine and OPD groups, with 94.3% (33/35) of participants in each group assigning the maximum score of 5 (P=.99). At 6 weeks, the median overall satisfaction score also remained 5 (IQR 5-5) in both groups. All participants in the telemedicine group (35/35, 100%) reported being fully informed about their postoperative care plans, compared with 94.3% (33/35) in the OPD group (P=.49). Ratings for surgeon communication, professionalism, and overall interaction quality were similarly high in both groups, with no statistically significant between-group differences observed (Table 4).

At 6 weeks, satisfaction levels remained unchanged. All telemedicine participants rated their overall care, visit experience, and surgeon interaction at the maximum score of 5. In the OPD group (n=35), 1 (2.9%) participant rated overall care as 4, while the remaining participants (n=34, 97.1%) gave a score of 5. Between-group comparisons revealed no statistically significant differences across any satisfaction domain (*P* values ranged from .15 to >.99).

XSL•FO
RenderX

Table . Patient satisfaction scores at 2 and 6 weeks postoperatively, stratified by follow-up modality and age group[a].

Satisfaction item, time, and score	Telemedicine group (<60 y; n=25), n (%)	OPD[b] group (<60 y; n=17), n (%)	*P* value (<60 y)	Telemedicine group (≥60 y; n=10), n (%)	OPD group (≥60 y; n=18), n (%)	*P* value (≥60 y)
Overall care satisfaction						
2 wk						
Score ≤4	1 (4)	1 (5.9)	>.99	1 (10)	1 (5.6)	>.99
Score 5	24 (96)	16 (94.1)	—[c]	9 (90)	17 (94.4)	—
6 wk						
Score ≤4	0 (0)	1 (5.9)	.41	0 (0)	0 (0)	>.99
Score 5	25 (100)	16 (94.1)	—	10 (100)	18 (100)	—
Satisfaction with follow-up visit						
2 wk						
Score ≤4	1 (4)	1 (5.9)	>.99	1 (10)	1 (5.6)	>.99
Score 5	24 (96)	16 (94.1)	—	9 (90)	17 (94.4)	—
6 wk						
Score ≤4	0 (0)	0 (0)	>.99	0 (0)	0 (0)	>.99
Score 5	25 (100)	17 (100)	—	10 (100)	18 (100)	—
Information about postoperative care						
2 wk						
Score ≤4	0 (0)	2 (11.8)	.16	0 (0)	0 (0)	>.99
Score 5	25 (100)	15 (88.2)	—	10 (100)	18 (100)	—
6 wk						
Score ≤4	0 (0)	1 (5.9)	.41	0 (0)	0 (0)	>.99
Score 5	25 (100)	16 (94.1)	—	10 (100)	18 (100)	—
Satisfaction with surgeon						
2 wk						
Score ≤4	2 (8)	2 (11.8)	>.99	0 (0)	0 (0)	>.99
Score 5	23 (92)	15 (88.2)	—	10 (100)	18 (100)	—
6 wk						
Score ≤4	1 (4)	1 (5.9)	>.99	0 (0)	0 (0)	>.99
Score 5	24 (96)	16 (94.1)	—	10 (100)	18 (100)	—
Recommendation to others						
2 wk						
Score ≤4	1 (4)	0 (0)	>.99	1 (10)	1 (5.6)	>.99
Score 5	24 (96)	17 (100)	—	9 (90)	17 (94.4)	—
6 wk						
Score ≤4	2 (8)	0 (0)	.52	0 (0)	0 (0)	>.99
Score 5	23 (92)	17 (100)	—	10 (100)	18 (100)	—
Overall surgeon rating						
2 wk						
Score ≤4	0 (0)	0 (0)	>.99	1 (10)	0 (0)	.36
Score 5	25 (100)	17 (100)	—	9 (90)	18 (100)	—
6 wk						

XSL•FO
RenderX

Satisfaction item, time, and score	Telemedicine group (<60 y; n=25), n (%)	OPD[b] group (<60 y; n=17), n (%)	*P* value (<60 y)	Telemedicine group (≥60 y; n=10), n (%)	OPD group (≥60 y; n=18), n (%)	*P* value (≥60 y)
Score ≤4	0 (0)	0 (0)	>.99	0 (0)	0 (0)	>.99
Score 5	25 (100)	17 (100)	—	10 (100)	18 (100)	—

[a]Satisfaction was assessed using a validated 6-item Likert scale (1=strongly disagree to 5=strongly agree). No scores below 4 were reported. Between-group comparisons were performed using the chi-square test or Fisher exact test, as appropriate.

[b]OPD: outpatient department.

[c]Not applicable because all participants reported the same score, resulting in no variability for statistical comparison.

Longitudinal Satisfaction Trends

Within-group analyses comparing week 2 and week 6 follow-up assessments showed consistent satisfaction over time in both cohorts (Table 4). In the telemedicine group, mean changes across individual satisfaction items ranged from 0.00 to 0.10, none of which were statistically significant (all $P>.05$). The OPD group exhibited similarly small variations (mean difference ≤0.06; $P>.30$). These findings indicate that satisfaction remained stable throughout the postoperative period, with no evidence of decline in either group.

Summary of Findings

Overall, telemedicine follow-up produced patient satisfaction levels that were not significantly different from those of traditional in-person outpatient visits. Although participants in the telemedicine group were younger and demonstrated slightly higher digital literacy, these differences did not adversely affect satisfaction outcomes. Both models of care were highly acceptable to patients, underscoring telemedicine as a well-accepted and feasible option for postoperative follow-up in appropriately selected patients undergoing upper-extremity surgery.

Discussion

Principal Findings

This randomized controlled trial evaluated patient satisfaction with telemedicine versus conventional in-person OPD follow-up after upper-extremity surgery and examined whether digital literacy influenced satisfaction. Patient satisfaction was uniformly high in both groups at 2 and 6 weeks postoperatively, with no significant differences across satisfaction domains. Although participants in the telemedicine group were younger and had slightly higher digital literacy scores, no significant association was observed between digital literacy and satisfaction. These findings suggest that telemedicine follow-up, when supported by brief in-person app instruction, provides a patient experience that is not significantly different from traditional postoperative care. However, this study was not designed to evaluate clinical safety, complication rates, unplanned visits, or efficiency-related outcomes, such as time savings or cost reduction; therefore, conclusions regarding safety or efficiency should not be inferred from these results.

Interpretation and Comparison with Existing Literature

The absence of statistically significant differences in satisfaction between telemedicine and in-person follow-up is consistent with prior orthopedic studies demonstrating similar levels of patient-reported satisfaction across follow-up modalities [5,17,18]. Previous research in shoulder, knee, hip, and hand surgery has shown that telemedicine can maintain high satisfaction while improving convenience and reducing travel burden [4,19-21]. Our findings extend this evidence to postoperative upper-extremity care in an urban, middle-income country context.

While most telemedicine studies originate from high-income settings [7,18], this study was conducted in Thailand, where variability in digital literacy and access to technology may present implementation challenges [11,13]. Despite this, satisfaction remained high in both groups, suggesting that telemedicine can be effectively integrated into postoperative care when supported by appropriate institutional infrastructure and patient education. A key contribution of this study is the evaluation of digital literacy as a potential determinant of satisfaction. Although digital literacy scores differed slightly between groups, no significant association with satisfaction was observed. This finding suggests that factors such as effective communication, structured postoperative education, and continuity of care may be more influential than baseline digital competence [10,11]. The brief in-person orientation to the telemedicine app may have mitigated usability barriers and supported equitable engagement with virtual follow-up.

Limitations

This study has several limitations. Its single-center design in an urban setting may limit generalizability to rural or resource-limited populations. In addition, the study cohort consisted predominantly of urban residents with universal smartphone ownership and stable internet access, which may not reflect the circumstances of rural populations, older adults, or individuals with lower digital literacy or limited access to digital infrastructure. Implementation of telemedicine in such settings may require additional support strategies, including simplified user interfaces, extended in-person training, caregiver involvement, or hybrid care models combining virtual and in-person follow-up.

The modest sample size limited subgroup analyses according to age, educational level, or degree of digital literacy. In addition, patient satisfaction was assessed using Likert scale questionnaires, which demonstrated a pronounced ceiling effect,

XSL•FO
RenderX

with most participants reporting maximum satisfaction scores across domains. Although this reflects a highly positive overall patient experience, such score clustering reduces the sensitivity of the instrument to detect subtle yet clinically meaningful differences between follow-up modalities. As a result, the absence of statistically significant differences should be interpreted as indicating similarly high overall satisfaction rather than complete equivalence across all experiential dimensions. Future studies may benefit from more nuanced or modality-specific satisfaction measures, as well as alternative metrics such as the Net Promoter Score, to better distinguish levels of patient experience among highly satisfied populations.

In addition, trial registration was completed after participant enrollment had begun. Although the study protocol, including objectives, eligibility criteria, outcome measures, and planned analyses, was finalized prior to enrollment and remained unchanged throughout the study, the timing of registration should be considered a limitation and may raise concerns regarding transparency. Finally, this study focused primarily on satisfaction outcomes and did not evaluate clinical outcomes, complication rates, unplanned visits, or efficiency-related metrics, such as time or cost savings, which should be addressed in future research.

Implications for Future Research

Future multicenter studies involving diverse populations are needed to enhance generalizability. Qualitative methods may provide deeper insights into patient experiences with telemedicine, particularly regarding communication, trust, and usability [10]. Incorporation of clinical outcomes, economic evaluations, and longer-term follow-up would further clarify the role of telemedicine in postoperative orthopedic care.

Conclusions

Telemedicine represents a well-accepted and feasible alternative to traditional in-person follow-up after upper-extremity surgery. The consistently high satisfaction observed across both follow-up modalities suggests that, when supported by adequate digital infrastructure and patient education, telemedicine can be effectively integrated into postoperative care pathways. The observed benefits suggest potential systemic efficiency—such as reduced outpatient clinic burden—and improved patient convenience, including decreased travel and waiting time, as supported by prior research; however, clinical safety outcomes, economic efficiency, and cost-effectiveness were not directly evaluated in this study.

Acknowledgments

The authors sincerely thank all patients who participated in this study. The authors gratefully acknowledge support from the Navamindradhiraj University Research Fund. Editorial assistance for English language and style was supported by generative artificial intelligence (AI) tools, including ChatGPT (OpenAI), which were used solely for language editing, formatting, and clarification of reviewer comments during manuscript preparation. All AI-assisted content was carefully reviewed and verified by the authors, who are solely responsible for the content, scientific accuracy, and conclusions of the final manuscript.

Funding

This study was supported by Navamindradhiraj University Research Fund. The funder had no role in the study design, data collection, analysis, interpretation, or manuscript preparation.

Data Availability

The datasets analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: WN, SS
Data curation: WN, CM, KI, SP
Formal analysis: WN, CM, KP
Funding acquisition: SS
Investigation: WN, SS, CM, ST, SC, TT
Methodology: WN, SS, KP
Project administration: KI, SP
Resources: ST, SC, TT
Supervision: SS
Validation: SS, KP
Writing—original draft: WN, CM
Writing—review and editing: SS, KI, SP, ST, SC, TT
All authors reviewed, revised, and approved the final manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Editorial Notice

This randomized study was retrospectively registered. Authors noted that the registration was conducted retrospectively due to an administrative oversight regarding the timing of trial registration. The editor granted an exception from ICMJE rules mandating prospective registration of randomized trials because the risk of bias appears low and the study information registered with authors' ethics committee matches the trial reporting. However, readers are advised to carefully assess the validity of any potential explicit or implicit claims related to primary outcomes or effectiveness, as retrospective registration does not prevent authors from changing their outcome measures retrospectively.

Multimedia Appendix 1

Screenshots of the Vajira@Home telemedicine platform used for patient login, video consultation, and postoperative follow-up. [PNG File, 15 KB - jmir_v28i1e86918_app1.png]

Checklist 1

CONSORT EHEALTH checklist (V1.6.1).

[PDF File, 9727 KB - jmir_v28i1e86918_app2.pdf]

References

1. Farooq U. Leveraging telemedicine for enhanced healthcare accessibility: a narrative review of patient satisfaction, appointment adherence, and cost-effectiveness. J Health Rehabil Res 2025;5(1):1-6. [doi: 10.61919/jhrr.v5i1.1757]
2. Tomoh BO, Mustapha AY, Mbata AO, Kelvin-Agwu M, Forkuo AY, Kolawole TO. Assessing the impact of telehealth interventions on rural healthcare accessibility: a quantitative study. Int J Appl Res Soc Sci 2025;7(3):203-226. [doi: 10.51594/ijarss.v7i3.1861]
3. Hahn Z, Hotchkiss J, Atwood C, et al. Travel burden as a measure of healthcare access and the impact of telehealth within the Veterans Health Administration. J Gen Intern Med 2023 Jul;38(Suppl 3):805-813. [doi: 10.1007/s11606-023-08125-3] [Medline: 37340257]
4. Kane LT, Thakar O, Jamgochian G, et al. The role of telehealth as a platform for postoperative visits following rotator cuff repair: a prospective, randomized controlled trial. J Shoulder Elbow Surg 2020 Apr;29(4):775-783. [doi: 10.1016/j.jse.2019.12.004] [Medline: 32197766]
5. Muschol J, Heinrich M, Heiss C, et al. Assessing telemedicine efficiency in follow-up care with video consultations for patients in orthopedic and trauma surgery in Germany: randomized controlled trial. J Med Internet Res 2022 Jul 27;24(7):e36996. [doi: 10.2196/36996] [Medline: 35896015]
6. Kumar S, Kumar A, Kumar M, Kumar A, Arora R, Sehrawat R. Feasibility of telemedicine in maintaining follow-up of orthopaedic patients and their satisfaction: a preliminary study. J Clin Orthop Trauma 2020 Oct;11(Suppl 5):S704-S710. [doi: 10.1016/j.jcot.2020.07.026] [Medline: 32837105]
7. Fahey E, Elsheikh MF, Davey MS, Rowan F, Cassidy JT, Cleary MS. Telemedicine in orthopedic surgery: a systematic review of current evidence. Telemed J E Health 2022 May;28(5):613-635. [doi: 10.1089/tmj.2021.0221] [Medline: 34375150]
8. Harst L, Lantzsch H, Scheibe M. Theories predicting end-user acceptance of telemedicine use: systematic review. J Med Internet Res 2019 May 21;21(5):e13117. [doi: 10.2196/13117] [Medline: 31115340]
9. Hossain MA, Amin R, Masud AA, Hossain MI, Hossen MA, Hossain MK. What drives people's behavioral intention toward telemedicine? An emerging economy perspective. Sage Open 2023 Jul;13(3):1-20. [doi: 10.1177/21582440231181394]
10. Nguyen ML, Garcia F, Juarez J, et al. Satisfaction can co-exist with hesitation: qualitative analysis of acceptability of telemedicine among multi-lingual patients in a safety-net healthcare system during the COVID-19 pandemic. BMC Health Serv Res 2022 Feb 14;22:195. [doi: 10.1186/s12913-022-07547-9]
11. Campanozzi LL, Gibelli F, Bailo P, Nittari G, Sirignano A, Ricci G. The role of digital literacy in achieving health equity in the third millennium society: a literature review. Front Public Health 2023;11:1109323. [doi: 10.3389/fpubh.2023.1109323] [Medline: 36891330]
12. Fitzpatrick PJ. Improving health literacy using the power of digital communications to achieve better health outcomes for patients and practitioners. Front Digit Health 2023;5:1264780. [doi: 10.3389/fdgth.2023.1264780] [Medline: 38046643]
13. Soundararajan A, Lim JX, Ngiam NH, et al. Smartphone ownership, digital literacy, and the mediating role of social connectedness and loneliness in improving the wellbeing of community-dwelling older adults of low socio-economic status in Singapore. PLoS One 2023;18(8):e0290557. [doi: 10.1371/journal.pone.0290557] [Medline: 37647263]
14. Chairuengjitjaras P, Nimmanterdwong Z, Petchlorlian A, Praditpornsilpa K, Tangkijvanich P. Exploring the feasibility and acceptability of a telehealth platform for older adults with noncommunicable diseases and chronic viral hepatitis. Telemed J E Health 2025 Mar;31(3):344-353. [doi: 10.1089/tmj.2024.0289] [Medline: 39379066]
15. Amin H, Abid Malik M, Akkaya B. Development and validation of Digital Literacy Scale (DLS) and its implication for higher education. Int J Distance Educ E-Learn 2021;7(1):24-43. [doi: 10.36261/ijdeel.v7i1.2224]
16. Cha EDK, Suraci C, Petrosky D, et al. Patient satisfaction in shoulder arthroscopy: telemedicine vs clinic follow-up visits. Clin Shoulder Elb 2022;25(2):106-111. [doi: 10.5397/cise.2021.00619]

XSL•FO
RenderX

17. Melian C, Kieser D, Frampton C, C Wyatt M. Teleconsultation in orthopaedic surgery: a systematic review and meta-analysis of patient and physician experiences. J Telemed Telecare 2022 Aug;28(7):471-480. [doi: 10.1177/1357633X20950995] [Medline: 32873138]
18. Buvik A, Bugge E, Knutsen G, Småbrekke A, Wilsgaard T. Patient reported outcomes with remote orthopaedic consultations by telemedicine: a randomised controlled trial. J Telemed Telecare 2019 Sep;25(8):451-459. [doi: 10.1177/1357633X18783921] [Medline: 29973130]
19. Grandizio LC, Mettler AW, Caselli ME, Pavis EJ. Telemedicine after upper extremity surgery: a prospective study of program implementation. J Hand Surg Am 2020 Sep;45(9):795-801. [doi: 10.1016/j.jhsa.2020.06.002] [Medline: 32693989]
20. Fabrés Martín C, Ventura Parellada C, Herrero Antón de Vez H, Ordoñez Urgiles CE, Alonso-Rodríguez Piedra J, Mora Guix JM. Telemedicine approach for patient follow-up after total knee and reverse total shoulder arthroplasty: a pilot study. Int J Comput Assist Radiol Surg 2023 Mar;18(3):595-602. [doi: 10.1007/s11548-022-02784-z] [Medline: 36422767]
21. Sharareh B, Schwarzkopf R. Effectiveness of telemedical applications in postoperative follow-up after total joint arthroplasty. J Arthroplasty 2014 May;29(5):918-922. [doi: 10.1016/j.arth.2013.09.019] [Medline: 24342278]

Abbreviations

***ICD-10*:** *International Classification of Diseases, Tenth Revision*
OPD: outpatient department

Edited by A Stone; submitted 01.Nov.2025; peer-reviewed by AJ Umeano, C Aladeokin; revised version received 19.Apr.2026; accepted 28.Apr.2026; published 25.Jun.2026.

Please cite as:
Ngowroongrueng W, Suwannaphisit S, Malaikritsanachalee C, Thaveepunsan S, Chaitantipongse S, Teeragananan T, Imsakul K, Puntumabumrung S, Prasittichok K
Digital Literacy and Patient Satisfaction in Telemedicine Follow-Up With In-Person App Instruction Versus Outpatient Department Follow-Up After Upper-Extremity Surgery: A Randomized Controlled Trial
J Med Internet Res 2026;28:e86918
URL: https://www.jmir.org/2026/1/e86918
doi:10.2196/86918

© Werapat Ngowroongrueng, Sitthiphong Suwannaphisit, Chanon Malaikritsanachalee, Sutee Thaveepunsan, Sirisak Chaitantipongse, Theephop Teeragananan, Kanokporn Imsakul, Supaira Puntumabumrung, Krit Prasittichok. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Telehealth Scale and Artificial Intelligence Adoption Tiers Across Clinical and Operational Domains in US Hospitals: Cross-Sectional Study

Lingbo Liu[1,2], PhD

[1]Thrust of Urban Governance and Design, Society Hub, The Hong Kong University of Science and Technology (Guangzhou), No 1 Du Xue Rd, Nansha District, Guangzhou, Guangdong, China

[2]Center for Geographic Analysis, Harvard University, Cambridge, MA, United States

Corresponding Author:
Lingbo Liu, PhD
Thrust of Urban Governance and Design, Society Hub, The Hong Kong University of Science and Technology (Guangzhou), No 1 Du Xue Rd, Nansha District, Guangzhou, Guangdong, China

Abstract

Background: Telehealth expansion and artificial intelligence (AI) adoption are often described as parallel dimensions of health system digital transformation. However, whether telehealth scale is associated with hospital AI adoption and whether this relationship varies across hospital settings remain unclear.

Objective: This study examined the association of telehealth scale with clinical and operational AI adoption tiers in US hospitals and assessed whether these patterns differed by telehealth reporting behavior and geography.

Methods: This cross-sectional study included 6173 US acute care hospitals using linked 2024 American Hospital Association Annual Survey and Information Technology Supplement data and 2023 Healthcare Cost Report Information System data. Telehealth scale was parameterized using log-transformed telehealth volume, a telehealth nonreporting indicator, and a reported-zero telehealth indicator. Clinical and operational AI adoption tiers were derived from hospital-reported AI capability items and classified into 3 tiers. Both outcomes were modeled using multioutput gradient-boosted tree classifiers, and model behavior was interpreted using Shapley additive explanations, partial dependence plots, and stratified analyses by the Core-Based Statistical Area category.

Results: Telehealth volume was the strongest predictor of both clinical and operational AI adoption tiers and had a larger contribution to the clinical AI model. Telehealth nonreporting was common, occurring in 57% (3521/6173) of hospitals, and was concentrated among hospitals in the lowest clinical AI adoption tier, accounting for 91.4% (3145/3441) of hospitals with no reported clinical AI adoption. Higher telehealth volume was associated with a steep increase in predicted clinical AI adoption tiers at lower telehealth volumes, followed by a plateau at higher volumes. At similar telehealth volumes, rural hospitals showed weaker telehealth-attributed contributions to predicted clinical AI adoption tiers than metropolitan hospitals. Supplementary analyses suggested that telehealth reporting status and telehealth intensity reflected related but distinct structural processes.

Conclusions: Telehealth scale was strongly associated with hospital AI adoption tiers, especially clinical AI adoption tiers. These findings suggest that telehealth capacity may serve as a practical hospital-level marker of broader digital readiness for AI adoption, but the cross-sectional design does not establish whether telehealth expansion precedes or causes AI adoption. Hospitals with telehealth nonreporting and rural hospitals may face additional structural barriers that limit the translation of digital capacity into AI maturity. Policies to reduce inequities in hospital AI adoption may therefore need to pair telehealth expansion with implementation support, interoperability capacity, and organizational resources.

(J Med Internet Res 2026;28:e96762) doi:10.2196/96762

KEYWORDS

artificial intelligence; telemedicine; telehealth; digital health; hospitals; health information technology; machine learning; explainable artificial intelligence; extreme gradient boosting; XGBoost; Shapley additive explanations; United States

Introduction

Telehealth expansion and artificial intelligence (AI) adoption are 2 prominent dimensions of the ongoing digital transformation in health care [1]. Telehealth use increased rapidly during and after the COVID-19 pandemic and has remained an important channel for care delivery, while AI has been introduced across a growing range of clinical and operational functions in hospitals [2-4]. Although these developments are often discussed separately, both depend on a

XSL•FO
RenderX

shared foundation of digital infrastructure, data availability, workflow integration, and organizational capacity [5-7]. Hospitals that are able to deliver care at scale through digital platforms may also be better positioned to implement AI tools that rely on timely, structured, and interoperable information [8,9]. This shared infrastructure perspective is important because hospital digital transformation is shaped not only by organizational and implementation capacity but also by the availability of individual technologies. Adopting and sustaining digital tools requires data infrastructure, workflow integration, governance, implementation planning, and organizational readiness [10]. National studies suggest that advanced digital capabilities remain unevenly distributed across US hospitals, with variation by hospital size, metropolitan status, and local context [11,12]. Telehealth scale may therefore provide an observable hospital-level marker of broader digital readiness.

From a health informatics perspective, AI implementation requires more than access to algorithms. It depends on the ability of health systems to capture digital patient interactions, integrate data across platforms, support technology-enabled workflows, and maintain governance structures that enable safe and sustained use [13-15]. Telehealth may therefore provide a useful systems-level indicator of broader digital readiness. Hospitals with higher telehealth activity may have already invested in patient-facing digital interfaces, staff training, data exchange processes, and operational coordination that also support AI deployment, particularly for clinical applications [16]. By contrast, hospitals with limited telehealth infrastructure may face barriers that also constrain the adoption of AI-enabled tools [17-20].

This question has important equity implications. Prior research suggests that digital capacity is unevenly distributed across hospitals and communities, with rural, safety-net, and socioeconomically disadvantaged settings often facing greater limitations in health information technology infrastructure, broadband access, staffing, and implementation support [21-23]. These constraints may affect not only telehealth access but also the ability to adopt and operationalize AI in routine care. If telehealth scale is associated with hospital AI maturity, then uneven telehealth capacity may signal a broader digital divide in AI readiness [24-27]. Understanding this relationship is therefore relevant not only for innovation strategy but also for equitable health system transformation [28-31].

Despite growing interest in both telehealth and AI, the relationship between telehealth scale and hospital AI maturity has not been well characterized at the national level. In addition, telehealth nonreporting may reflect a structurally different state from explicitly reported-zero telehealth use, but these 2 conditions are often not distinguished analytically. Using data from 6173 US hospitals, this study examined whether telehealth scale was associated with clinical and operational AI adoption tiers, whether telehealth nonreporting and reported-zero telehealth volume showed different patterns, and whether these relationships varied across geographic settings. The study hypothesized that telehealth scale would be positively associated with both domains of AI adoption tiers, with stronger associations for clinical AI, and that rural hospitals would show weaker telehealth-associated AI readiness at comparable telehealth volumes.

Methods

Study Design and Data Sources

This retrospective cross-sectional study followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guideline. The study setting comprised US acute care hospitals represented in the linked 2024 American Hospital Association (AHA) Annual Survey, AHA Information Technology Supplement, and 2023 Healthcare Cost Report Information System (HCRIS) datasets. The hospital, rather than the patient or encounter, served as the unit of analysis. Since this was a secondary analysis of hospital-level administrative and survey data, no participant recruitment was conducted. The 2024 AHA Annual Survey and Information Technology Supplement were linked with hospital utilization and financial data from the 2023 HCRIS to construct a national analytic dataset of US acute care hospitals. Hospitals were matched using facility identifiers and retained if they had valid linkage keys and met acute care eligibility criteria, resulting in a final analytic sample of 6173 hospitals. Financial variables were expressed in 2023 US dollars. Detailed variable definitions and coding rules are provided in Table S1 in Multimedia Appendix 1.

Exposures, Outcomes, and Covariates

The primary exposure was annual telehealth video-visit volume from the AHA survey. To distinguish reporting behavior from observed telehealth use, telehealth was parameterized using 3 variables: log-transformed telehealth volume, a telehealth nonreporting indicator, and a reported-zero indicator among hospitals that submitted a number of telehealth video visits. This parameterization allowed hospitals with no reported telehealth value to be distinguished from hospitals that explicitly reported zero telehealth volume. Telehealth nonreporting was treated as a distinct analytic state rather than ordinary random missingness because it may reflect limited telehealth capability, limited administrative reporting infrastructure, survey response capacity, reimbursement context, or broader digital disadvantage. Reported-zero telehealth volume was interpreted as active nonuse or absence of telehealth service delivery during the reporting period. The primary outcomes were clinical AI adoption tiers and operational AI adoption tiers, constructed from AHA survey items. The clinical AI domain included AI use for diagnosis, patient care activities, clinical decision support, robotic or assisted surgery, precision medicine or genetic analysis, population health management, medication administration, and other clinical applications. The operational AI domain included AI use for revenue cycle management, supply chain management, staffing and scheduling, patient flow and discharge planning, operational efficiency, and applications of information technology or facilities. For each domain, cleaned item responses were summed to generate a raw score, and hospitals were classified into 3 empirical AI adoption tiers: tier 0 for nonresponse across all domain items or a summed score of 0, tier 1 for active adopters at or below the median score among adopters, and tier 2 for active adopters above that median. Complete nonresponse and zero adoption were

XSL•FO
RenderX

combined in tier 0 because both states indicate no reported AI functionality in the available AHA survey items, and the data do not provide sufficient information to reliably distinguish the absence of adoption from the absence of item-level reporting. Therefore, tier 0 was interpreted conservatively as no reported AI adoption rather than as a validated maturity stage or definitive evidence of no AI use. Because no externally validated national hospital AI maturity scale was available in these data, these categories were interpreted as empirical adoption tiers rather than formally validated maturity stages.

Candidate covariates were selected to capture hospital scale, throughput, capital intensity, payer mix, safety-net burden, organizational structure, and geography. The linked datasets did not include direct measures of AI-specific or telehealth-specific investment, such as dedicated digital health budgets, vendor contracts, implementation teams, or specialized technical staffing. Therefore, financial and organizational variables, including total operating expenses, payroll expenses, capital intensity, teaching status, and system affiliation, were used as indirect proxies for organizational resource capacity. These variables included total patient visits, admissions, staffed beds, surgical volume, total operating expenses, payroll, Medicaid and Medicare inpatient days, teaching status, system affiliation, and the Core-Based Statistical Area category. Capital intensity was defined as total operating expenses per staffed bed, and surgical intensity as inpatient surgeries divided by admissions plus 1. Numeric variables were coerced after standardizing missing-value tokens; remaining missing numeric predictors were set to 0 for tree-based modeling. Missing or nonpositive bed counts were replaced with the median of positive values to avoid implausible denominators, and categorical variables were represented using indicator variables.

Statistical Analysis

Hospital characteristics were summarized overall and by clinical AI adoption tier using medians and IQRs for skewed continuous variables and counts with percentages for categorical variables; differences across tiers were assessed using Kruskal-Wallis and χ^2 tests. To visualize regional patterns, hospital-level measures were aggregated to the county level, and local spatial autocorrelation for county-level telehealth intensity and clinical AI was assessed using local indicators of spatial association based on county adjacency. The main predictive analysis used multioutput gradient-boosted tree classification to model clinical and operational AI adoption tiers simultaneously from a shared feature set, implemented as parallel multiclass extreme gradient boosting (XGBoost) models with an 80%/20% train-test split stratified on the joint clinical-operational tier label when feasible and otherwise on the clinical AI tier. The XGBoost classifiers used 300 estimators, a learning rate of 0.03, a maximum tree depth of 5, subsampling of 0.8, column subsampling of 0.8, multiclass soft-probability objectives, and log-loss evaluation. Model performance was evaluated on a held-out 20% test set.

To interpret model behavior, SHAP (Shapley additive explanations) values were estimated for the tier 2 outcome, and global feature importance was summarized using mean absolute SHAP values. SHAP values were used to quantify feature contributions to model predictions and should not be interpreted as causal effects, independent risk ratios, or direct policy levers. Signed SHAP contributions for telehealth nonreporting and reported-zero telehealth volume were also calculated to distinguish their associations with high AI adoption. Partial dependence plots were used to assess nonlinear associations for prespecified drivers, including telehealth intensity, telehealth nonreporting, expenses per bed, and total patient visits. Geographic heterogeneity in telehealth-attributed clinical AI readiness was evaluated by stratifying SHAP dependence by the Core-Based Statistical Area category and smoothing these patterns with locally weighted regression.

As sensitivity analyses, 2 additional sets of models were conducted. First, to assess whether the main results were driven by structural telehealth nonreporting, the analytic sample was restricted to hospitals that reported telehealth volume, and the clinical and operational AI tier models were reestimated after removing the telehealth nonreporting indicator. In these reporter-only models, telehealth scale was represented by log-transformed reported telehealth volume and an indicator for reported-zero telehealth volume. Second, to evaluate whether the findings depended on the distribution-based AI tier definitions, the 3-level clinical and operational AI adoption tier outcomes were replaced with the corresponding continuous raw AI scores, and XGBoost regression models were estimated using the same predictor set as the primary analysis. Two-sided $P<.001$ was considered statistically significant for descriptive comparisons. Analyses were conducted in Python (Python Software Foundation). Detailed model specifications and full performance metrics are provided in Multimedia Appendix 1.

Ethical Considerations

This study used deidentified, institution-level secondary administrative and survey data, with hospitals as the unit of analysis. It did not involve patient-level records, protected health information, or interaction with human participants. The present analysis was not submitted for institutional review board review because it did not constitute human participants' research under relevant institutional policy [32] and applicable regulations for secondary analyses of deidentified institutional data. No informed consent was required. The author had licensed permission to use the AHA data [33], and HCRIS data were used in accordance with its access conditions. No ethics application number was assigned.

Results

Sample Characteristics by the Clinical AI Maturity Tier

The analytic cohort included 6173 US hospitals. When hospitals were stratified by clinical AI adoption tiers, 3441 (55.7%) were classified as tier 0, 1544 (25%) as tier 1, and 1188 (19.2%) as tier 2 (Table 1). Clinical AI scores increased from a median of 0 (IQR 0 - 0) in tier 0 to 22 (IQR 19 - 25) in tier 2, while operational AI scores increased from 0 (IQR 0 - 0) to 14 (IQR 11 - 17; $P<.001$ for both). Telehealth volume increased in parallel, from a median of 0.00 thousand visits (IQR 0.00 - 0.74) in tier 0 to 2.40 (IQR 0.21 - 20.93) in tier 2 ($P<.001$). Tier 2 hospitals also had greater patient throughput and capital intensity

XSL•FO
RenderX

than tier 0 hospitals, with median total visits increasing from 34.17 thousand (IQR 9.14‑107.00) to 160.37 (IQR 75.13‑331.87) and median expenses per bed increasing from $1.08 million (IQR $0.49-$1.90 million) to $1.82 million (IQR $1.33-$2.54 million; *P*<.001 for both). Geographic and institutional characteristics were similarly stratified: tier 2 hospitals were predominantly urban (927/1188, 78%), while rural hospitals were underrepresented in tier 2 (102/1188, 8.6%) and more common in tier 1 (395/1544, 25.6%). Teaching hospitals were also more concentrated in tier 2 than tier 0 (157/1188, 13.2% vs 72/3441, 2.1%; *P*<.001).

Table . Hospital characteristics by the clinical artificial intelligence (AI) adoption tier[a].

Variable	Overall (N=6173)	Tier 0 (no reported adoption; n=3441)	Tier 1 (low; n=1544)	Tier 2 (high; n=1188)	*P* value
Clinical AI score (0‑40), median (IQR)	0 (0-14)	0 (0-0)	11 (8-14)	22 (19-25)	<.001
Operational AI score (0‑30), median (IQR)	0 (0-10)	0 (0-0)	8 (6-10)	14 (11-17)	<.001
Telehealth volume (thousands of visits), median (IQR)	0.41 (0.00-5.44)	0.00 (0.00-0.74)	0.09 (0.00-1.94)	2.40 (0.21-20.93)	<.001
Total patient visits (thousands), median (IQR)	57.88 (15.47-162.81)	34.17 (9.14-107.00)	60.95 (18.51-157.58)	160.37 (75.13-331.87)	<.001
Expenses per bed ($M/bed), median (IQR)	1.32 (0.64-2.13)	1.08 (0.49-1.90)	1.31 (0.65-2.12)	1.82 (1.33-2.54)	<.001
Total operating expenses ($M), median (IQR)	75.84 (30.76-273.50)	54.10 (27.67-169.66)	65.86 (26.24-232.15)	275.28 (93.93-654.33)	<.001
Surgical intensity ratio, median (IQR)	0.76 (0.12-1.59)	0.68 (0.00-1.82)	0.80 (0.06-1.57)	0.83 (0.57-1.25)	<.001
Telehealth reporting status, n (%)					<.001
Telehealth volume not reported	3521 (57.0)	3145 (91.4)	248 (16.1)	128 (10.8)	
Reported-zero telehealth volume (among reporters), n (%)					<.001
True zero	830 (31.3)	180 (60.8)	515 (39.7)	135 (12.7)	
Location, n (%)					<.001
Urban	4223 (68.4)	2373 (69.0)	923 (59.8)	927 (78.0)	
Micro	852 (13.8)	467 (13.6)	226 (14.6)	159 (13.4)	
Rural	1098 (17.8)	601 (17.5)	395 (25.6)	102 (8.6)	
Teaching status, n (%)					<.001
Teaching	278 (4.5)	72 (2.1)	49 (3.2)	157 (13.2)	
Nonteaching	5895 (95.5)	3369 (97.9)	1495 (96.8)	1031 (86.8)	

[a]Values for categorical variables are presented as n/N (%), where percentages are column percentages unless otherwise specified. Reported-zero telehealth volume is calculated among hospitals that reported a telehealth volume value.

Telehealth reporting status also varied sharply across tiers. Overall, 3521 out of 6173 (57%) hospitals did not report telehealth volume. Nonreporting was concentrated in tier 0 (3145/3441, 91.4%) and was uncommon in tier 2 (128/1188, 10.8%; *P*<.001). Among hospitals that reported telehealth volume, reported-zero telehealth use was most frequent in tier 0 (180/296, 60.8%) and least frequent in tier 2 (135/1060, 12.7%), suggesting that telehealth nonreporting and reported-zero use represent related but distinct states of digital capability.

Spatial Concordance of Telehealth Capacity and Clinical AI

County-level maps showed substantial spatial concordance between log-transformed telehealth volume and clinical AI adoption tiers (Figure 1). Counties with higher telehealth activity tended to colocate with higher clinical AI scores, whereas counties with lower telehealth activity tended to colocate with lower clinical AI scores. Local indicators of spatial association further indicated that these patterns were spatially clustered rather than diffuse. High-high clusters for telehealth volume were concentrated along major metropolitan and coastal corridors, including the Northeast and parts of the West Coast,

XSL•FO
RenderX

and clinical AI showed a similar clustering pattern. Low-low clusters extended across parts of the South, Appalachia, and the Great Plains, consistent with regionally organized deficits in digital capacity.

Figure 1. Spatial concordance and clustering of telehealth utilization and clinical artificial intelligence (AI) maturity across continental US counties. (B) County-level mean logarithmic telehealth volume. (A) County-level mean clinical AI score. The 2 raw-value maps show strong spatial concordance, with higher telehealth activity and higher clinical AI colocated in major metropolitan corridors. (C) LISA classification for clinical AI, showing a similar hot spot and cold spot structure. (D) Local indicators of spatial association (LISA) for log telehealth volume, expressed as local Moran I *z*-scores and classified into high-high (HH), low-low (LL), high-low (HL), low-high (LH), and not significant clusters.

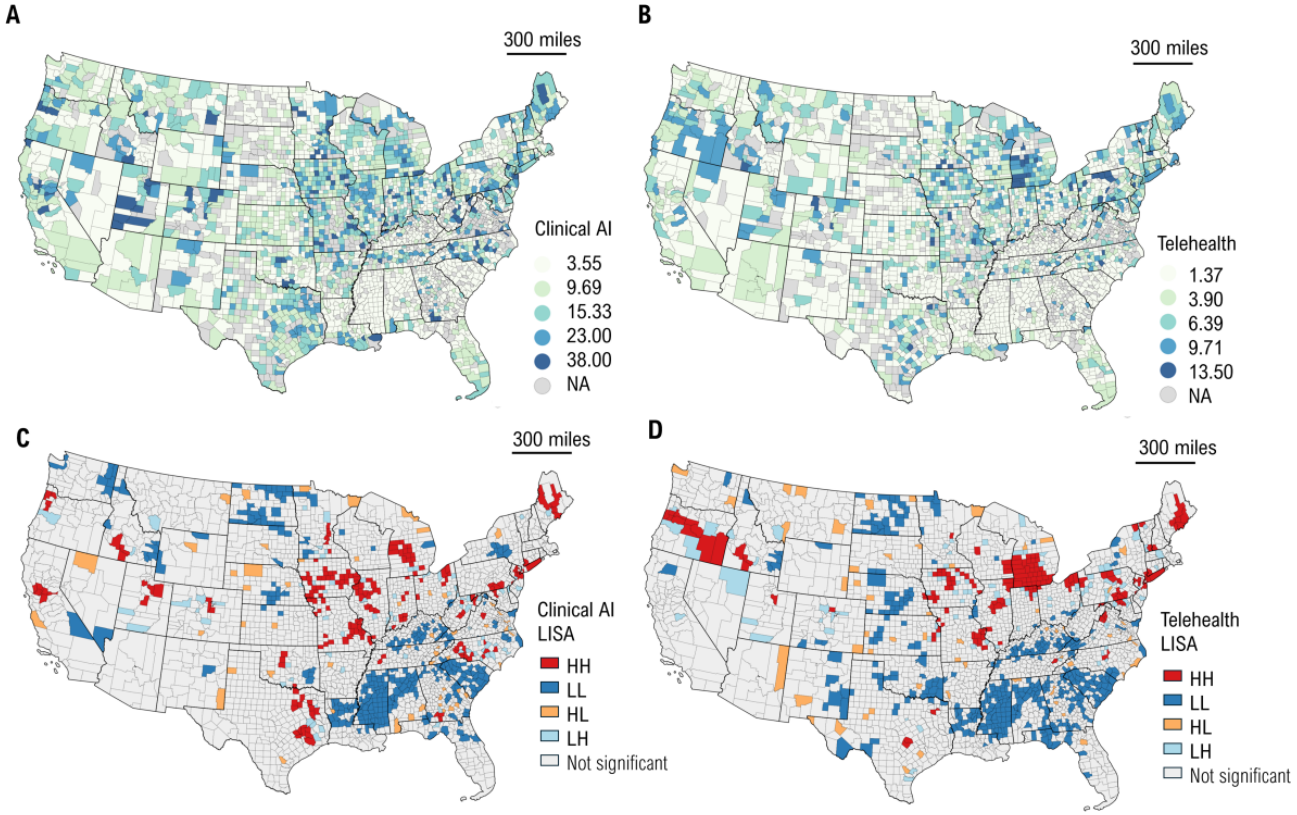

Telehealth Scale and AI Adoption Tiers in Predictive Models

In the multioutput XGBoost models predicting tier 2 adoption, telehealth volume was the strongest predictor of both clinical and operational AI adoption tiers (Figure 2A). The mean absolute SHAP value for telehealth volume was 0.511 in the clinical AI model and 0.327 in the operational AI model, indicating a stronger association with clinical AI than operational AI. Telehealth nonreporting also ranked highly in both models (mean absolute SHAP=0.373 for clinical AI and 0.245 for operational AI). Capital intensity was more predictive for clinical AI than operational AI (mean absolute SHAP value=0.279 vs 0.120), whereas throughput-related variables contributed relatively more to operational AI. Partial dependence plots showed a nonlinear association between telehealth volume and predicted clinical AI adoption tiers (Figure 2B). The predicted probability of tier 2 clinical AI adoption increased steeply during early telehealth scale-up and then plateaued at higher telehealth volumes. By contrast, telehealth nonreporting was associated with a downward shift in predicted adoption probability across the feature range. Capital intensity showed a saturating association, while total visits had a weaker and less specific association with clinical AI readiness. On the held-out test set, the clinical AI tier model achieved an overall accuracy of 0.777 and a Cohen κ of 0.618, while the operational AI tier model achieved an overall accuracy of 0.776 and a Cohen κ of 0.607. Class-specific performance metrics are reported in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 2. Telehealth scale as a model-based marker of hospital artificial intelligence (AI) adoption tiers. (A) Global feature importance from Shapley additive explanations (SHAP), summarized as mean absolute SHAP values for the probability of high adoption, defined as tier 2, comparing clinical AI and operational AI models. Telehealth volume, log-transformed, was the top predictor in both domains and had a larger model contribution in the clinical AI model. (B) Partial dependence plots showing model-predicted probability of tier 2 adoption across telehealth volume, telehealth nonreporting, defined as missing VIDVZ, capital intensity, defined as expenses per bed, and total patient visits; rug marks indicate the observed feature distribution. (C) Geography-stratified SHAP dependence of telehealth volume for the clinical AI model, with LOESS-smoothed trends and point estimates, showing lower telehealth-attributed contributions in rural hospitals than in metropolitan hospitals at comparable volumes. (D) Mean signed SHAP contributions for telehealth nonreporting and reported-zero telehealth volume, defined as VIDVZ=0 among reporting hospitals, shown for clinical and operational AI models; negative values indicate reduced probability of tier 2 adoption. ADMTOT: total admissions; EXPTOT: total operating expenses; Intensity: surgical intensity ratio; LOESS: locally estimated scatterplot smoothing; MAPP5=1: hospital affiliated with a larger health system; MCDIPD: Medicaid inpatient days; MCRIPD: Medicare inpatient days; PATTOT: total patient visits; SUROPTOT: total surgical operations; VIDVZ: number of telehealth video visits.

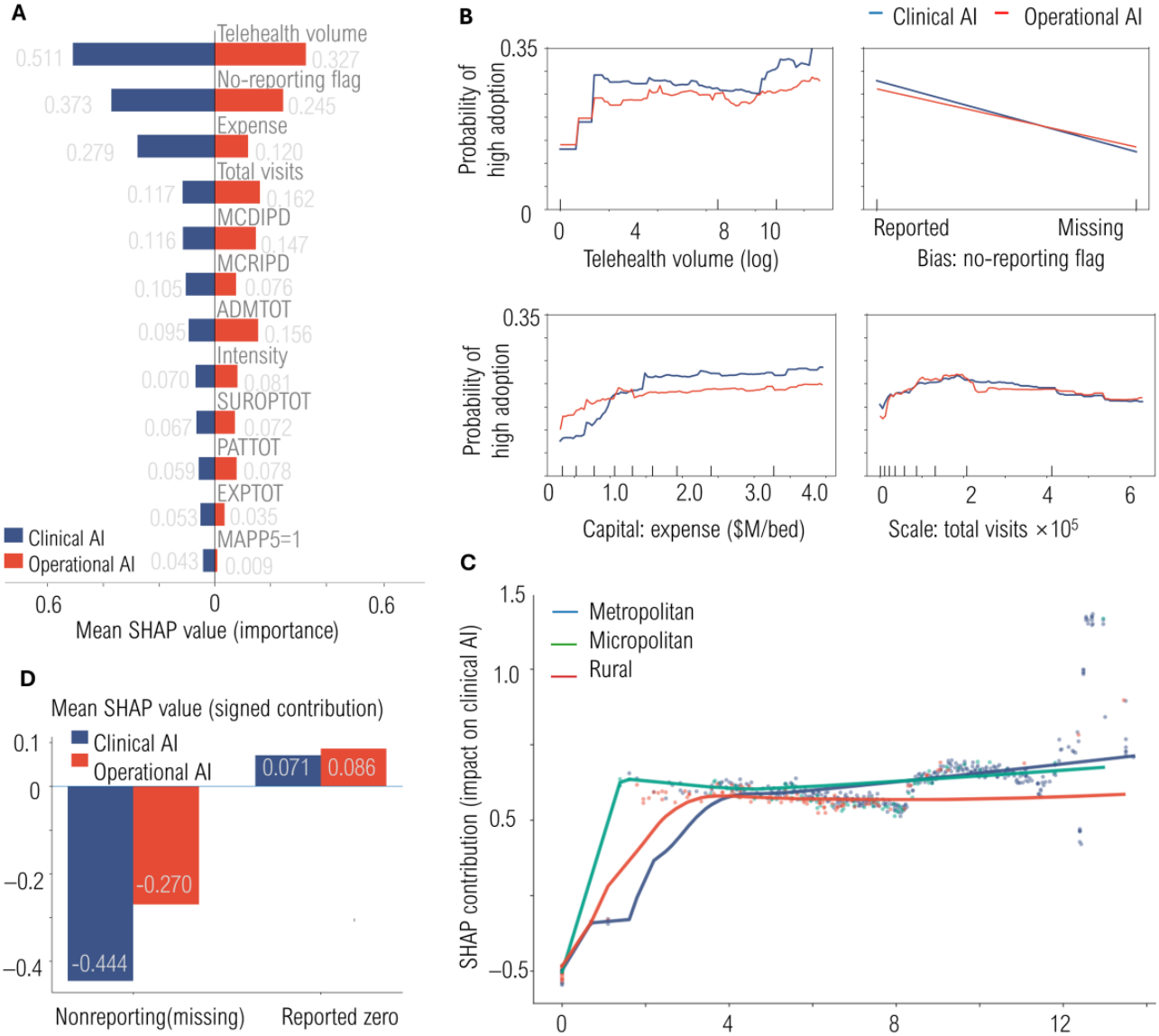

Geographic Heterogeneity, Reporting Mechanisms, and Robustness Analyses

Geographic stratification of SHAP dependence suggested that, at comparable telehealth volumes, rural hospitals consistently had lower telehealth-associated contributions to predicted clinical AI adoption tiers than metropolitan hospitals (Figure 2C). In addition, signed SHAP values suggested that telehealth nonreporting was associated with a negative contribution to the probability of tier 2 adoption, particularly for clinical AI, whereas reported-zero telehealth volume showed a smaller and directionally distinct pattern (Figure 2D).

Supplementary analyses further suggested that telehealth reporting status and telehealth intensity reflected distinct structural processes. A separate gradient-boosted model predicted telehealth nonreporting with good discrimination (test area under the receiver operating characteristic curve [AUC]=0.814), and among reporting hospitals, a telehealth intensity model explained substantial variation in log-transformed telehealth volume (test R^2=0.41; Tables S3 and S4 in Multimedia Appendix 1). In competing-driver analyses, telehealth remained the largest contributor to high adoption in both AI domains, and in ablation analyses, removing telehealth reduced test-set discrimination, with a larger decline for clinical

XSL•FO
RenderX

AI than for operational AI (Figure 3; Tables S5 and S6 in Multimedia Appendix 1).

Figure 3. Decomposing telehealth reporting and intensity and testing robustness. (A) Global feature importance, summarized as mean absolute Shapley additive explanations (SHAP) values, for a reporting model predicting telehealth nonreporting, defined as missing annual telehealth video-visit volume (VIDVZ), using hospital scale, utilization, and financial covariates. (B) Global feature importance, summarized as mean absolute SHAP values, for an intensity model predicting telehealth volume among reporting hospitals, with the outcome defined as log (1+VIDVZ). (C) Partial dependence plots (PDPs) for the reporting model showing the modeled probability of telehealth nonreporting as a function of total visits (VTOT) and capital intensity, defined as expenses per bed. (D) PDPs for the intensity model, restricted to reporting hospitals, showing predicted telehealth volume on the log scale as a function of VTOT and expenses per bed; rug marks indicate the distribution of observations. (E) Competing-drivers analysis of high artificial intelligence (AI) adoption: mean absolute SHAP values, computed on the held-out test set, comparing telehealth measures with alternative explanatory factors, including financial pressure, defined as operating margin; organizational scale, defined as bed size; mission, defined as teaching status; complexity, defined as surgery intensity; payer mix, defined as Medicaid ratio; and operational pressure, defined as occupancy proxy, shown separately for clinical and operational AI high adoption. (F) Ablation test quantifying the percentage drop in the test set area under the receiver operating characteristic curve (AUC) after removing the telehealth feature from the competing-drivers models for clinical and operational AI; larger drops indicate greater dependence of model discrimination on the telehealth-related predictive signal. ADMTOT: total admissions; CBSA: Core-Based Statistical Area; CBSA=METRO: hospital located in a metropolitan CBSA; CBSA=RURAL: hospital located outside metropolitan or micropolitan CBSA; EXPTOT: total operating expenses; MAPP3=1: participating site recognized for one or more programs accredited by Accreditation Council for Graduate Medical Education; EXPTOT: total operating expenses; Intensity: surgical intensity ratio; MAPP5=1: hospital affiliated with a larger health system; MCDIPD: Medicaid inpatient days; MCRIPD: Medicare inpatient days; PAYTOT: total payroll expenses; SUROPTOT: total surgical operations; VTOT: total patient visits.

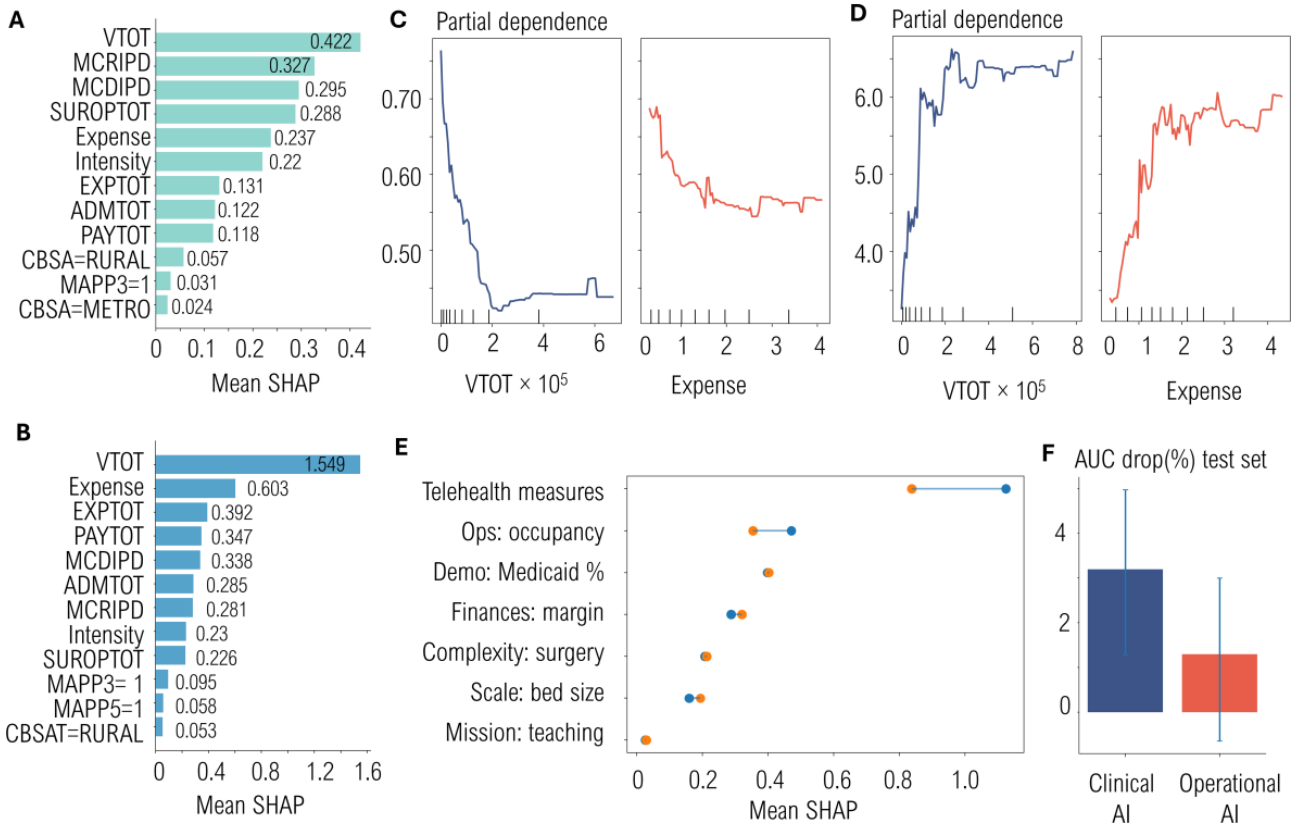

Sensitivity Analyses

Sensitivity analyses further supported the robustness of the main findings while highlighting the structural importance of telehealth nonreporting. In reporter-only models restricted to hospitals with reported telehealth volume (n=2652), log-transformed telehealth volume remained the strongest predictor of high clinical and operational AI adoption. The reporter-only clinical model achieved an accuracy of 0.653, a Cohen κ of 0.374, and a tier 2 AUC of 0.821. The reporter-only operational model achieved an accuracy of 0.633, a Cohen κ of 0.329, and a tier 2 AUC of 0.775. In both models, log-transformed telehealth volume ranked first in SHAP importance for tier 2 adoption, with mean absolute SHAP values of 0.274 for clinical AI and 0.198 for operational AI.

Continuous-score sensitivity analyses showed similar patterns when the clinical and operational AI raw scores were modeled directly rather than categorized into 3 tiers. The clinical AI score model achieved an R^2 of 0.590, a root mean square error of 6.085, and a mean absolute error of 3.851, while the operational AI score model achieved an R^2 of 0.545, a root mean square error of 4.363, and a mean absolute error of 2.763. Telehealth nonreporting was the strongest predictor of both clinical and operational AI raw scores, and log-transformed telehealth volume was the second strongest predictor in both models. These findings suggest that the main results were not driven solely by telehealth nonreporting or by the median-based construction of AI adoption tiers.

XSL•FO
RenderX

Discussion

Principal Findings

In this cross-sectional study of 6173 US hospitals, telehealth volume was strongly associated with hospital AI adoption tiers and was more strongly associated with clinical AI than with operational AI. Across descriptive, spatial, and model-based analyses, 3 findings were notable: telehealth volume showed the largest model contribution to high AI adoption in both domains, the association with clinical AI adoption tiers appeared nonlinear, and telehealth nonreporting and rural location were both associated with lower predicted clinical AI adoption tiers. Together, these findings suggest that telehealth capacity may serve as a model-based marker of broader digital infrastructure and organizational readiness relevant to clinical AI adoption, but the cross-sectional design does not establish whether telehealth expansion precedes or causes AI adoption.

These results also help clarify the relationship between telehealth and AI, which are often discussed in separate policy and implementation literature. Telehealth is commonly framed as an access strategy [5-7], whereas AI is more often framed as an innovation, quality, or efficiency strategy [22]. In these data, however, telehealth volume was more strongly associated with clinical AI maturity than with operational AI maturity. The stronger association with clinical AI than operational AI may reflect the closer alignment between telehealth and patient-facing clinical workflows. Telehealth delivery requires digital patient intake, scheduling, documentation, billing, clinician engagement, and electronic health record integration, which may also support clinical AI applications that depend on structured patient data and clinical decision-support workflows [9,10]. By contrast, operational AI applications may rely more heavily on back-office administrative, staffing, supply-chain, and enterprise resource planning systems. This distinction suggests that clinical and operational AI adoption may be shaped by overlapping but partly different organizational capabilities. This pattern suggests that telehealth may reflect a broader form of digital capability, including routine data capture, platform integration, workflow standardization, and clinician engagement with technology-enabled care, all of which are relevant to equitable digital and AI-enabled care [24]. The weaker association with operational AI is also informative, suggesting that hospital AI adoption may not follow a single pathway and that clinical and operational AI may depend on partly different implementation conditions [9,28].

An additional finding was that telehealth nonreporting was common and concentrated among hospitals in the lowest clinical AI adoption tier. More than half of the hospitals did not report telehealth volume, and additional analyses suggested that nonreporting was itself predictable from hospital structural characteristics and carried a strongly negative contribution to the probability of high AI adoption. This distinction matters because hospitals with missing telehealth data may be systematically different from hospitals with reported-zero telehealth volume. Treating these states as equivalent, or excluding hospitals with missing telehealth values altogether, may therefore obscure an important dimension of the hospital digital divide [18,19]. In this context, telehealth nonreporting should be understood less as a neutral data artifact than as a potential signal of broader digital disadvantage [21].

The geographic findings further highlight the equity implications of these results. County-level clustering showed strong spatial concordance between telehealth intensity and clinical AI maturity, with persistent low-low clusters across parts of the South, Appalachia, and the Great Plains. In addition, at comparable telehealth volumes, rural hospitals showed lower telehealth-associated contributions to predicted clinical AI maturity than metropolitan hospitals. This pattern suggests that telehealth expansion alone may not translate into equivalent AI readiness across settings. Prior work has similarly shown that telemedicine expansion may have limited effectiveness in places with poor internet access, low digital inclusion, and existing health care access constraints, and that hospitals in more socioeconomically deprived service areas are less likely to adopt telehealth and related health information technologies [31]. However, this rural-metro difference should be interpreted as an observed pattern rather than the direct evidence of specific mechanisms. Rural hospitals are heterogeneous, and future work should examine whether this relationship varies by regional context, community deprivation, broadband availability, workforce capacity, and hospital financial resources [8,30].

Policy and Implementation Implications

These findings have several policy implications. First, policies that support telehealth infrastructure and policies that support AI adoption may be more effective if they are designed together rather than as separate initiatives, because broader digital inclusion and health information capacity shape equitable access to digital care [18]. Second, the nonlinear pattern observed here suggests that the association between telehealth volume and clinical AI maturity may be greater at lower levels of telehealth use than at higher levels. Third, telehealth expansion alone may be insufficient to narrow disparities in AI maturity. Hospitals serving rural and structurally disadvantaged communities may require additional implementation support, including interoperability capacity, workflow redesign, technical staffing, governance, and vendor integration [21]. Overall, efforts to reduce digital inequities in AI adoption may require strengthening multiple components of hospital digital capacity rather than focusing on a single technology alone.

Limitations

This study has several limitations. First, the cross-sectional design does not establish temporal ordering or causality between telehealth scale and AI adoption. Second, telehealth nonreporting was common and likely structural. Although nonreporting was modeled separately and reporter-only and continuous-score sensitivity analyses were conducted, these strategies cannot fully resolve nonrandom missingness or endogeneity. Third, the AI adoption tiers were empirical measures derived from AHA survey items and distribution-based thresholds rather than externally validated AI maturity stages. Although continuous-score sensitivity analyses using the raw clinical and operational AI scores produced similar telehealth-related patterns, future work should validate hospital AI maturity measures using external implementation benchmarks,

XSL•FO
RenderX

governance indicators, workflow integration measures, and longitudinal adoption trajectories. Fourth, the available data did not include the direct measures of AI-specific or telehealth-specific investment, such as dedicated digital health budgets, vendor contracts, implementation teams, specialized technical staffing, or governance capacity. Therefore, financial and organizational variables, including operating expenses, payroll expenses, capital intensity, teaching status, and system affiliation, were used as indirect proxies for organizational resource capacity. Fifth, the analysis did not directly model state-level telehealth regulations, reimbursement policies, broadband policy environments, or regional market conditions, which may influence both telehealth reporting and AI adoption.

Conclusions

In this national cross-sectional study of US hospitals, telehealth scale was strongly associated with hospital AI adoption tiers, with a particularly strong relationship for clinical AI. Telehealth nonreporting was common and concentrated among hospitals with the lowest clinical AI adoption tiers, and rural hospitals showed weaker telehealth-associated clinical AI patterns than metropolitan hospitals at similar telehealth volumes. These findings suggest that telehealth capacity may serve as a practical hospital-level marker of broader digital readiness for AI adoption, while telehealth nonreporting may identify hospitals at risk of being left behind in digital transformation. From a digital health and health informatics perspective, the results indicate that equitable AI adoption may depend not only on access to AI tools themselves but also on the organizational and infrastructural conditions that support digital care delivery. Efforts to reduce disparities in hospital AI adoption tiers may therefore require combining telehealth expansion with targeted support for interoperability, technical staffing, workflow redesign, governance, and implementation capacity, especially in rural and structurally disadvantaged settings.

Acknowledgments

The author thanks the reviewers for their constructive comments and support. The author used generative artificial intelligence only for language editing and organizational support during revision.

Funding

The author received no financial support for this work.

Data Availability

Hospital-level data from the American Hospital Association Annual Survey and Information Technology Supplement were obtained under license and are not publicly redistributable by the author; researchers seeking access should apply directly to the American Hospital Association in accordance with its data use conditions.

Authors' Contributions

LL conceptualized the study, developed the methodology, curated and analyzed the data, interpreted the findings, drafted and revised the manuscript, and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The analysis of telehealth scale and clinical and operational artificial intelligence adoption tiers in US hospitals.
[DOCX File, 351 KB - jmir_v28i1e96762_app1.docx]

References

1. Liu L, Wang F. Telehealth accessibility. In: International Encyclopedia of Geography: John Wiley & Sons; 2026:1-7. [doi: 10.1002/9781118786352]
2. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
3. Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bull World Health Organ 2020 Apr 1;98(4):251-256. [doi: 10.2471/BLT.19.237487] [Medline: 32284648]
4. The Lancet Digital Health. Rapid generative AI rollout in health care. Lancet Digit Health 2025 Aug;7(8):100909. [doi: 10.1016/j.landig.2025.100909] [Medline: 40803944]
5. Patel SY, Mehrotra A, Huskamp HA, Uscher-Pines L, Ganguli I, Barnett ML. Variation in telemedicine use and outpatient care during the COVID-19 pandemic in the United States. Health Aff (Millwood) 2021 Feb;40(2):349-358. [doi: 10.1377/hlthaff.2020.01786] [Medline: 33523745]
6. Dorsey ER, Topol EJ. Telemedicine 2020 and the next decade. The Lancet 2020 Mar;395(10227):859. [doi: 10.1016/S0140-6736(20)30424-4]

XSL•FO
RenderX

7. Shachar C, Engel J, Elwyn G. Implications for telehealth in a postpandemic future: regulatory and privacy issues. JAMA 2020 Jun 16;323(23):2375-2376. [doi: 10.1001/jama.2020.7943] [Medline: 32421170]
8. Yan AS, Apathy NC, Chen J. Adoption of health information technologies by area socioeconomic deprivation among US hospitals. JAMA Health Forum 2025 Sep 5;6(9):e253035. [doi: 10.1001/jamahealthforum.2025.3035] [Medline: 40911325]
9. Sahni NR, Carrus B. Artificial intelligence in US health care delivery. N Engl J Med 2023 Jul 27;389(4):348-358. [doi: 10.1056/NEJMra2204673] [Medline: 37494486]
10. Kamel Rahimi A, Pienaar O, Ghadimi M, et al. Implementing AI in hospitals to achieve a learning health system: systematic review of current enablers and barriers. J Med Internet Res 2024 Aug 2;26:e49655. [doi: 10.2196/49655] [Medline: 39094106]
11. Hwang YM, Ng MY, Pillai M, Sahai MP, Hernandez-Boussard T. The landscape of AI implementation in US hospitals. Nat Health 2026;1(1):99-112. [doi: 10.1038/s44360-025-00016-7]
12. Goldfarb A, He X(, Teodoridis F. Patterns of artificial intelligence adoption by hospitals. AEA Pap Proc 2025 May 1;115:40-45. [doi: 10.1257/pandp.20251003]
13. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
14. Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. Health data poverty: an assailable barrier to equitable digital health care. Lancet Digit Health 2021 Apr;3(4):e260-e265. [doi: 10.1016/S2589-7500(20)30317-4] [Medline: 33678589]
15. Bitar H, Alismail S. The role of eHealth, telehealth, and telemedicine for chronic disease patients during COVID-19 pandemic: a rapid systematic review. Digit Health 2021;7:20552076211009396. [doi: 10.1177/20552076211009396] [Medline: 33959378]
16. Liu L, Onega T, Moen EL, et al. Telehealth infrastructure for cancer care in the United States. Sci Data 2026. [doi: 10.1038/s41597-026-07063-z]
17. Ramesh T, McBain RK, Cantor JH, Stein BD, Yu H. Mental health outcomes among patients living in US counties lacking broadband access and psychiatrists. JAMA Netw Open 2023 Sep 5;6(9):e2333781. [doi: 10.1001/jamanetworkopen.2023.33781] [Medline: 37707819]
18. Sieck CJ, Sheon A, Ancker JS, Castek J, Callahan B, Siefer A. Digital inclusion as a social determinant of health. NPJ Digit Med 2021 Mar 17;4(1):52. [doi: 10.1038/s41746-021-00413-8] [Medline: 33731887]
19. Liu L, Onega T, Moen EL, et al. Digital divides in telehealth accessibility for cancer care in the United States. NPJ Digit Med 2025 Aug 19;8(1):534. [doi: 10.1038/s41746-025-01931-5] [Medline: 40830581]
20. Liu L, Alford-Teaster J, Onega T, Wang F. Refining 2SVCA method for measuring telehealth accessibility of primary care physicians in Baton Rouge, Louisiana. Cities 2023 Jul;138:104364. [doi: 10.1016/j.cities.2023.104364] [Medline: 37274944]
21. Wang ML, Gago CM, Rodriguez K. Digital redlining—the invisible structural determinant of health. JAMA 2024 Apr 16;331(15):1267-1268. [doi: 10.1001/jama.2024.1628] [Medline: 38497952]
22. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
23. Schulman KA, Richman BD. Toward an effective innovation agenda. N Engl J Med 2019 Mar 7;380(10):900-901. [doi: 10.1056/NEJMp1812460] [Medline: 30673357]
24. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019 Apr 4;380(14):1347-1358. [doi: 10.1056/NEJMra1814259] [Medline: 30943338]
25. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med 2019 Jan;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
26. Pisano ED. AI shows promise for breast cancer screening. Nature 2020 Jan;577(7788):35-36. [doi: 10.1038/d41586-019-03822-8] [Medline: 31894156]
27. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016 Dec 13;316(22):2402-2410. [doi: 10.1001/jama.2016.17216] [Medline: 27898976]
28. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med 2021 Apr;27(4):582-584. [doi: 10.1038/s41591-021-01312-x] [Medline: 33820998]
29. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018 Apr 3;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
30. Graham AK, Weissman RS, Mohr DC. Resolving key barriers to advancing mental health equity in rural communities using digital mental health interventions. JAMA Health Forum 2021 Jun 4;2(6):e211149. [doi: 10.1001/jamahealthforum.2021.1149] [Medline: 36218752]
31. Loccoh EC, Nguyen A, Kim G, Warraich HJ. Geospatial analysis of access to health care and internet services in the US. JAMA Netw Open 2022 Nov 1;5(11):e2243792. [doi: 10.1001/jamanetworkopen.2022.43792] [Medline: 36416826]
32. US Department of Health and Human Services. 45 CFR §46.102: definitions for purposes of this policy. : Electronic Code of Federal Regulations; 2026 URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46/subpart-A/section-46.102 [accessed 2026-06-02]

XSL•FO
RenderX

33. American Hospital Association. Wharton Research Data Services (WRDS), Wharton School, University of Pennsylvania. URL: https://wrds-www.wharton.upenn.edu/pages/about/data-vendors/american-hospital-association/ [accessed 2026-06-02]

Abbreviations

AHA: American Hospital Association
AUC: area under the receiver operating characteristic curve
HCRIS: Healthcare Cost Report Information System
SHAP: Shapley additive explanations
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
XGBoost: extreme gradient boosting

Edited by I Steenstra; submitted 31.Mar.2026; peer-reviewed by E Ford, F Wang; revised version received 21.May.2026; accepted 22.May.2026; published 12.Jun.2026.

Please cite as:
Liu L
Telehealth Scale and Artificial Intelligence Adoption Tiers Across Clinical and Operational Domains in US Hospitals: Cross-Sectional Study
J Med Internet Res 2026;28:e96762
URL: https://www.jmir.org/2026/1/e96762
doi:10.2196/96762

© Lingbo Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Clinician-Patient Asynchronous Text Messaging Communication in Hospital-at-Home Care: Qualitative Study

Jeremy Soon Leong Seow[1], RN, BSc (Hons); Stephanie Ko[2,3], MBBS, MPH; Shi Yun Low[3], MPH; Shefaly Shorey[1], RN, RM, PhD

[1]Alice Lee Centre for Nursing Studies, Yong Loo Lin School of Medicine, National University of Singapore, Level 5, Centre for Translational Medicine, Block MD6, 14 Medical Drive, Singapore, Singapore

[2]Department of Medicine, Division of Advanced Internal Medicine, National University Hospital, Singapore, Singapore

[3]NUHS@Home, National University Health System, Singapore, Singapore

Corresponding Author:
Shefaly Shorey, RN, RM, PhD
Alice Lee Centre for Nursing Studies, Yong Loo Lin School of Medicine, National University of Singapore, Level 5, Centre for Translational Medicine, Block MD6, 14 Medical Drive, Singapore, Singapore

Abstract

Background: Hospital-at-home (HaH) care models are increasingly being adopted as a strategy to treat older adults with acute care needs and reduce strain on health care systems. Technological innovations, particularly digital communication platforms, have become essential in enabling care delivery beyond traditional hospital settings. Among these, asynchronous messaging tools have the potential to facilitate safe, timely, and coordinated interactions between health care providers, patients, and caregivers. Despite growing interest, little is known about how the content and relational dynamics of such exchanges influence care experiences in real-world HaH contexts.

Objective: This study aimed to examine the content of SMS text messaging exchanged between health care providers and patients or caregivers within a HaH program in Singapore during the COVID-19 pandemic.

Methods: A descriptive qualitative design was used to analyze retrospective WhatsApp messages exchanged between health care providers and patients or caregivers from August 2022 to October 2022. An inductive qualitative content analysis approach was used to systematically identify and categorize emerging patterns in the data.

Results: The analysis of 1218 WhatsApp messages from 354 HaH admissions identified three main categories: (1) clinical checks and advice; (2) administrative and transport arrangements; and (3) quality of interpersonal dynamics, supported by 13 subcategories, reflecting both task-oriented and relational dimensions of communication. The findings highlight how asynchronous messaging enables continuous care coordination while fostering trust and engagement between health care providers, patients, and caregivers.

Conclusions: This study underscores the multifaceted role of digital communication in HaH care, demonstrating its influence beyond clinical coordination to operational efficiency and the quality of interpersonal relationships. The findings provide insights into how digitally mediated interactions supported care delivery and patient engagement within a Singapore HaH program during the COVID-19 pandemic. While situated within a pandemic-specific context, these findings offer transferable considerations for the design of communication strategies and digital tools to enhance the responsiveness, coordination, and patient-centeredness of HaH programs.

(J Med Internet Res 2026;28:e83530) doi:10.2196/83530

KEYWORDS

hospital-at-home; digital communication; asynchronous communication; SMS text messaging; qualitative content analysis; patient-provider communication

Introduction

Globally, health care systems are facing increasing pressure due to constrained hospital bed capacity, rapidly aging populations, and a growing shortage of health care professionals [1]. These systemic challenges have driven interest in alternative care models that enhance capacity and improve care delivery beyond the conventional hospital environment, such as the hospital-at-home (HaH) care model, which delivers hospital-level care to patients within the comfort and familiarity of their home environment [2]. HaH programs are categorized into 2 main models: admission avoidance, which provides acute hospital-level care at home to prevent hospital admission, and early-supported discharge, which facilitates earlier discharge from the hospital with continued care delivered at home [3].

XSL•FO
RenderX

Systematic reviews have shown that HaH programs generally achieve clinical outcomes comparable to conventional inpatient care, while potentially reducing hospital length of stay and health care costs and improving patients' and caregivers' satisfaction [4,5].

During the COVID-19 pandemic, HaH programs were rapidly implemented in many health care systems to manage clinically stable patients in the community, while alleviating hospital bed pressures. Within HaH programs, technological advancements have further expanded the potential of these care models, with digital communication tools playing a pivotal role in enabling clinical care coordination and remote monitoring [6]. Health care services can be delivered more efficiently through eHealth modalities, such as mobile apps, digital communication platforms, and remote monitoring devices, further improving patient outcomes [7]. In HaH care models, digital communication platforms are commonly used to facilitate clinical communication, such as providing clinical updates and coordinating administrative concerns [8,9]. Effective communication among health care providers, patients, and caregivers is central to the success of HaH programs [10,11]. Unlike traditional inpatient care with routine face-to-face interactions, HaH relies on remote monitoring, virtual consultations, and periodic home visits to support timely information exchange [12]. These communication mechanisms are essential in maintaining patient safety, ensuring continuity of care, and facilitating real-time clinical decision-making in the home setting, effectively bridging hospital-level care within the home environment and strengthening the efficiency and viability of HaH programs [13,14].

While digital SMS text messaging has become a core component of communication in HaH programs, the actual content and dynamics of these interactions remain underexplored. Existing research has primarily focused on the operational benefits of digital tools, such as improving triage efficiency, facilitating clinical monitoring, and reducing hospital admissions, without fully examining how these platforms shape care relationships [15,16]. More recently, advances in artificial intelligence (AI)-driven conversational agents have further expanded the potential of digital health communication, with emerging studies examining the use of automated SMS text messaging systems and chatbots to support patient engagement and understand communication patterns in health care settings. However, comparatively little is known about the nature and dynamics of human-generated messaging interactions between health care providers and patients within HaH programs [17,18]. To address this gap, this study explored real-world SMS text messaging–based communication data generated during routine care between health care providers and patients or caregivers via WhatsApp, an asynchronous messaging platform that allows participants to send and receive messages without requiring real-time interaction to uncover patterns, themes, and tones. Beyond identifying informational content, it can also provide insights into how trust, empathy, and clinical authority are negotiated digitally, further illuminating insights into the relational labor embedded in HaH service delivery [19]. Therefore, understanding these digital dialogues is crucial for refining HaH models that are not only clinically effective but also relationally responsive and ethically grounded. As such, this study aimed to explore the content of SMS text messages exchanged between health care providers and patients or caregivers within the HaH program. Specifically, it addresses the research question: *What do the content and dynamics of digital text communication between health care providers, patients, and caregivers reveal about relational experiences in HaH programs?* The goal is for the findings of this study to inform the refinement of current and future HaH models, emphasizing the importance of digitally attuned communication practices and providing valuable insights to enhance health care provider training and strengthen patient engagement.

Methods

Study Design

A qualitative descriptive study design was used to explore the content of SMS text messages exchanged between health care providers and patients or caregivers within the HaH program. This design is appropriate as it contextually facilitates the grounded analysis of real-world SMS text message communications, offering insights into the content, tone, and relational dynamics that shape communication practices in real-world HaH settings [20]. The SRQR (Standards for Reporting Qualitative Research) checklist guided the reporting of this study [21] (Checklist 1).

Sampling and Eligibility

A convenience sampling approach was adopted, drawing on all available deidentified SMS text message communications exchanged between health care providers, patients, and caregivers enrolled in the HaH program during the COVID-19 pandemic. Eligibility criteria included the following: (1) enrollment in the HaH program during the study period, (2) the availability of WhatsApp message records between the health care team and the patient or caregiver, and (3) SMS text messages written in English. All SMS text messages meeting these criteria and exchanged during the HaH admission period were included in the analysis. In routine care, WhatsApp communication occurred using the patient's or caregiver's own device.

Setting

This study used retrospective data extracted between August and October 2022 from WhatsApp, a digital messaging platform used by health care providers and patients enrolled in the National University Health System (NUHS) program NUHS@Home. Established in 2020, NUHS@Home delivers hospital-level care at home for patients who would otherwise require inpatient hospitalization, primarily managing acute internal medicine conditions. During the study period, all patients included in this study were admitted as inpatient-level HaH cases and diagnosed with COVID-19, reflecting the patient population admitted at that time rather than a program-specific focus on COVID-19 care. In addition to SMS text messaging, patients received routine clinical monitoring through teleconsultations, remote monitoring, and home visits when clinically indicated. Patients who experienced clinical deterioration could be escalated for further evaluation or hospital

XSL•FO
RenderX

admission according to established protocols. In this study, all included patients were managed entirely within the HaH setting and did not require escalation to inpatient hospitalization. As such, the dataset captures SMS text message communications during routine home-based care, and interactions related to acute deterioration may therefore be underrepresented.

WhatsApp was used by the clinical care team as part of routine clinical care, with patients agreeing to its use for service communication because of its high local adoption, support for group communication, and convenience. It was integrated into routine care as an asynchronous communication platform to facilitate both clinical and administrative communications between health care providers and patients or their caregivers. These uses included sharing clinical updates, monitoring symptoms, addressing patient queries, and coordinating logistical matters such as appointments and medication delivery.

The study team comprised an internal medicine clinician (SK), a registered nurse (JSLS), a doctoral-trained researcher (SS), and a public health researcher (SYL) with formal training and experience in qualitative methodology. The diversity of professional backgrounds offers a multidisciplinary lens, further enriching the interpretation and contextualization of communication dynamics within the HaH setting.

Ethical Considerations

The study received ethical approval from the National Health Group Domain-Specific Review Board (reference 2021/00896), which serves as the ethics board of the participating institution. Specific research consent was waived by the ethics board because this study involved a retrospective analysis of deidentified routine care data. To ensure anonymity and confidentiality, data from the participants were deidentified and assigned coded pseudonyms.

Data Collection

Qualitative data were retrospectively extracted from WhatsApp comprising all SMS text messages exchanged between the NUHS@Home team and patients or caregivers between August 2022 and October 2022. Demographic characteristics of patients (eg, age, sex, and ethnicity) were obtained from the electronic medical records to describe the study population. Sample messages included conversations related to episodes of care or administrative coordination. As this was an exploratory retrospective qualitative study, no a priori sample size was specified. All eligible messages exchanged during the study period were included in the analysis. During iterative coding, category development was considered sufficiently comprehensive when no substantially new patterns or insights emerged from further review of the dataset [22].

Rigor

Rigor was established using the guidelines of credibility, transferability, dependability, and confirmability developed by Lincoln and Guba [23]. Credibility was enhanced through investigator triangulation during data analysis, which involved interdisciplinary team members interpreting the data. This ensured that multiple perspectives were incorporated and findings were grounded in the evidence. Transferability was maintained by providing detailed contextual descriptions of the setting and participant characteristics, allowing readers to assess relevance to other care environments. Dependability was ensured through the comprehensive documentation of an audit trail, comprising raw SMS text message data, coding, and analytic processes, to provide transparency and consistency. Confirmability was enhanced through reflexive journaling and independent coding by 2 researchers, followed by discussion and consensus-building on coding differences. As emphasized by Teh and Lek [24], reflexivity is a vital component of qualitative research that helps to mitigate unconscious researcher bias. The reflexive journal enabled the research team to critically examine their assumptions and decision-making, thereby enhancing the overall trustworthiness of the study [25].

Data Analysis

Content analysis was used to explore the SMS text messages exchanged between health care providers and patients or caregivers within the HaH program. This method was selected for its established applicability in analyzing communication and media-based data, making it well-suited to the study's focus on clinical SMS text messaging [26]. Content analysis allows researchers to interpret contextual meaning through a systematic coding process, thereby uncovering patterns and themes within communication [27]. Two members of the research team (SS and JSLS), both trained in qualitative research methods, independently reviewed and coded the data. Discrepancies were addressed through collaborative discussions among the research team, ensuring that the findings were accurately represented. Microsoft Word was used to organize and manage the dataset during the coding process. The 3-step inductive approach proposed by Elo and Kyngäs [28] guided the analytical process. In the preparation phase, researchers immersed themselves in the dataset by reading the messages several times to gain familiarity and determine appropriate units of analysis. Message timestamps were used at this stage to establish the chronological sequence of exchanges and to reconstruct conversational threads between health care providers and patients or caregivers. This enabled the identification of message initiation and response patterns within each interaction. In the organizing phase, the data were segmented into meaning units and coded to capture key content. These codes were then grouped into categories and subcategories through a process of abstraction. Finally, in the reporting phase, these categories were refined to reflect the recurring patterns present in the data.

Results

Overview

From August 2022 to October 2022, a total of 354 HaH admissions met the study inclusion criteria and were included in the analysis. All included patients were managed within the HaH settings without requiring inpatient hospitalization, and all patients used asynchronous communication via WhatsApp. The mean age of the participants was 57 (SD 17.8) years. Most study participants were female (199/354, 56.2%), and the group was composed predominantly of individuals with Chinese ethnic backgrounds (246/354, 69.5%; Table 1). The primary diagnosis was COVID-19 (N=354), and the mean length of stay in HaH

XSL•FO
RenderX

was 8 (SD 4) days. A total of 1218 SMS text messages were examined and coded from the conversations between health care providers and patients or caregivers. The median number of WhatsApp messages exchanged between the care team and patients or caregivers per patient interaction was 8 (IQR 13.5). Among the coded messages, 61.1% (364/596) were provider-initiated and 38.9% (232/596) were patient- or caregiver-initiated, based on the initiation of conversational threads. The median response time for health care providers was 11 (IQR 21.5) minutes during office hours (8 AM-6 PM) and 3 (IQR 6.25) minutes after office hours (6 PM).

Table . Sociodemographic, clinical, and SMS text messaging[a] characteristics of patients (N=354).

Collected data	Values
Age (y), mean (SD)	57 (17.8)
Male, n (%)	155 (43.8)
Ethnicity, n (%)	
Chinese	246 (69.5)
Indian	29 (8.2)
Malay	42 (11.9)
Others	37 (10.5)
Primary diagnosis, n (%)	
COVID-19	354 (100)
Length of stay in hospital-at-home (d), mean (SD)	8 (4)
Message thread per patient interaction, median (IQR)	8 (13.5)
Message directionality (n=596), n (%)	
Provider-initiated	364 (61.1)
Patient or caregiver-initiated	232 (38.9)
Response time for health care providers during office hours (8 AM-6 PM; min), median (IQR)	11 (21.5)
Response time for health care providers after office hours (after 6 PM; min), median (IQR)	3 (6.25)

[a]Total number of messages=1218.

Three categories were identified from the analysis of the SMS text messages, reflecting the nature of interactions between the health care providers and patients or caregivers within the HaH program: (1) clinical checks and advice, (2) administrative and transport arrangements, and (3) quality of interpersonal dynamics. These categories were further supported by 13 subcategories (Textbox 1).

XSL•FO
RenderX

Textbox 1. Categories and subcategories identified from the content analysis.

Category 1: clinical checks and advice

- Clinical consultation
- Vitals monitoring by health care providers
- Vitals monitoring by patients and caregivers
- Medication-related inquiries

Category 2: administrative and transport arrangements

- Logistical support and transport coordination
- Administrative transition
- Establishing lines of communication

Category 3: quality of interpersonal dynamics

- Emotional support
- Appreciation
- Trust and feedback
- Courteous communication
- Communication timeliness
- Tone and professional etiquette

Category 1: Clinical Checks and Advice

Overview

A total of 339 coded SMS text messages were categorized under clinical checks and advice, which reflected a range of provider-patient communications focused on monitoring and maintaining the clinical stability of patients in HaH. These were further supported by 4 subcategories: clinical consultation (91 SMS text messages), vitals monitoring by health care providers (157 SMS text messages), vitals monitoring by patients and caregivers (33 SMS text messages), and medication-related inquiries (58 SMS text messages). These subcategories were further elaborated through detailed codes as outlined in the coding framework in Table 2 (see Multimedia Appendix 1 for coding scheme).

Table . Coding framework for clinical checks and advice category (SMS text messages for coding: n=339).

Subcategories and code	SMS text messages for subcategory coding, n (%)	SMS text messages for code coding, n
Clinical consultation	91 (26.8)	
Family member seeks clinical advice from health care provider		55
A health care provider gives clinical advice		8
Updates on tests		28
Vitals monitoring by health care providers	157 (46.3)	
Health care provider following up on patient's abnormal vitals		54
Health care provider instructs on vital monitoring		103
Vitals monitoring by patients and caregivers	33 (9.7)	
Patient or caregiver informs of patient's vitals		33
Medication-related inquiries	58 (17.1)	
Patient or caregiver clarifies about medication		19
Health care provider instructs on administering medication		31
Health care provider clarifies the patient's or caregiver's medication request		8

XSL•FO
RenderX

Clinical Consultation

Three codes were used to categorize 91 SMS text messages related to clinical consultations, which reflected patients or caregivers seeking medical input, health care providers offering clinical advice, and both parties exchanging updates on tests. These SMS text messages often reflected patients or caregivers who sought clarification from health care providers regarding their symptoms and ongoing health concerns, particularly in situations where they were uncertain about the appropriate course of action:

> *Hi my mum is complaining about severe pain at her neck. Do you advise us to [go to] an [accident and emergency department] A&E or [is there] any doctor to review her?* [Caregiver A]
>
> *This [sharing a photo] is the diaper with blood stains, seems to be from the penile area [and the] catheter also shows some blood clots, should we be concerned?* [Caregiver B]

Furthermore, health care providers offered clinical guidance on the symptoms and overall well-being of patients and also interpreted and communicated diagnostic test results to patients and their caregivers:

> *[Team] wanted to let you know that the blood results were good [and] we are planning to drop by tomorrow to remove the PICC [Peripherally-Inserted Central Catheter]* [Doctor A]
>
> *Hi Madam [patient], the results from today's labs show [that] your magnesium is a little low today, so tomorrow morning we will top up with some IV [Intravenous] magnesium.* [Doctor A]

Vitals Monitoring by Health Care Providers

Two codes were used to categorize 157 SMS text messages related to vital signs monitoring by health care providers, which included active tracking and review of patients' parameters such as blood pressure, temperature, oxygen saturation, and glucose levels. These SMS text messages commonly included reminders to submit readings, follow-up assessments based on the submitted data, and clinical interpretations shared with patients:

> *Can you help to recheck the blood pressure in 15 mins again? It is very high now.* [Doctor A]
>
> *Mr [patient], I would like to update you that from tomorrow onwards, you only need to submit your vital signs once in the morning will do.* [Doctor C]

Vitals Monitoring by Patients and Caregivers

Thirty-three SMS text messages were categorized into one code related to vital signs monitoring conducted by patients and caregivers. These messages captured instances where patients or their caregivers independently initiated the reporting of vital parameters, such as blood pressure, temperature, and oxygen saturation, via SMS text messages:

> *Glucose test after breakfast: 7.3 mmol/l.* [Patient B]

Medication-Related Inquiries

Three codes were used to categorize 58 SMS text messages related to medication-related inquiries. These messages reflected questions from patients or caregivers regarding various aspects of medication management, such as prescribed regimens, timing of administration, and potential side effects:

> *May I know can I continue my Lenalidomide this evening?* [Patient C]
>
> *Can [patient] take paracetamol for fever/headaches or not advisable due to chemo.* [Caregiver C]

Similarly, health care providers also provided medication guidance, such as clarification of prescribed drugs and adjustments to dosages based on symptoms reported by patients:

> *Good morning Ms [patient]. I am the nurse whom spoke to [you] earlier [and] regarding the flumucil, there will not be an issue for you to take.* [Nurse A]
>
> *[You] can put ketoprofen plaster, [it] needs to be changed every 12 hourly…* [Doctor A]

Category 2: Administrative and Transport Arrangements

Overview

A total of 213 coded SMS text messages were categorized under administrative and transport arrangements, primarily reflecting logistical coordination, administrative processes, and efforts to establish effective communication between the health care team, patients, and caregivers. These were further supported by 3 subcategories: logistical support and transport coordination (42 SMS text messages), administrative transition (95 SMS text messages), and establishing lines of communication (76 SMS text messages). These subcategories were further elaborated through detailed codes as presented in the coding framework in Table 3.

XSL•FO
RenderX

Table . Coding framework for administrative and transport arrangements category (SMS text messages for coding: n=213).

Subcategories and code	SMS text messages for subcategory coding, n (%)	SMS text messages for code coding, n
Logistical support and transport coordination	42 (19.7)	
Arrangement for delivery/pick-up of logistics		36
Transport arrangement for patients		6
Administrative transition	95 (44.6)	
Discharge arrangements		35
Arrangements made for appointments and follow-up		42
Request for documentation		18
Establishing lines of communication	76 (35.7)	
Health care provider checks the contact availability of the patient or caregiver		19
Establishing a point of contact for the patient		26
Scheduling for a home visit		31

Logistical Support and Transport Coordination

Two codes were used to categorize 42 SMS text messages related to logistical support and transport coordination within the HaH model. The messages captured included arrangements for the delivery of medical devices and the replenishment of consumables, which were crucial to ensuring efficient care delivery:

> *Mdm [patient], what time can we send over the nurse box to your house [as] we need someone to be home to receive it?* [Doctor A]

Additionally, transport coordination emerged as a vital component, with messages detailing the planning and organization of ambulance services for diagnostic procedures or transitions between care settings:

> *Hi, the team is planning for Mdm [patient] to return to NUH for dressing change for the tunnelled catheter. We will be booking [the] transport, kindly standby for the timing.* [Doctor A]

Administrative Transition

Ninety-five SMS text messages were categorized into 3 codes related to administrative processes within the HaH model. These SMS text messages often involved a range of operational and clerical activities, such as the coordination of patient admission and discharge procedures, the completion of documentation requirements, and financial matters:

> *Could you provide me your email so we can send you the discharge documents?* [Doctor A]

> *Can we have [a] document [to] show [that] my father was warded at home for this period?* [Caregiver D]

Establishing Lines of Communication

Three codes were used to categorize 76 SMS text messages capturing efforts to initiate or clarify communication among patients, caregivers, and health care providers. These SMS text messages often included providing contact details, verifying preferred communication channels (eg, text and call), and scheduling home visits:

> *Hi Mdm [patient], let me know when you [are] ready [for us to call].* [Doctor A]

> *I am the granddaughter of Mr [patient]. You may reach out if you need anything.* [Caregiver E]

> *Team, just to check will the nurse come today?* [Patient D]

Category 3: Quality of Interpersonal Dynamics

Overview

A total of 666 SMS text messages were categorized under interpersonal communication quality, emphasizing the central role of relational dynamics and the multifaceted nature of interpersonal interactions within the HaH care model. These were further supported by 6 subcategories: emotional support (14 SMS text messages), appreciation (11 SMS text messages), trust and feedback (10 SMS text messages), courteous communication (48 SMS text messages), communication timeliness (10 SMS text messages), and tone and professional etiquette (573 SMS text messages). These subcategories were further elaborated through codes as presented in the coding framework in Table 4.

XSL•FO
RenderX

Table . Coding framework for the quality of interpersonal dynamics category (SMS text messages for coding: n=666).

Subcategories and code	SMS text messages for subcategory coding, n (%)	SMS text messages for code coding, n
Emotional support	14 (2.1)	
The health care provider provides emotional support to patients or caregivers		14
Appreciation	11 (1.7)	
Patients or caregivers showing appreciation		11
Trust and feedback	10 (1.5)	
Patient feels comfortable sharing concerns with the health care provider		6
Negative feedback		4
Courteous communication	48 (7.2)	
Patient or caregiver positively acknowledges instructions		48
Communication timeliness	10 (1.5)	
Delayed reply from patients or caregivers		2
Delayed reply from health care providers		4
Health care provider responds to patient's missed call		4
Tone and professional etiquette	573 (86)	
Informal tone by patient or caregiver		163
Formal tone by patient or caregiver		94
Informal tone by health care provider		189
Formal tone by health care provider		127

Emotional Support

Fourteen SMS text messages were categorized into 1 code, reflecting expressions of empathy, reassurance, and encouragement conveyed by health care providers through SMS text messages:

> *Do not worry about the NUCOT [National University Centre for Organ Transplantation] blood test, we will do the necessary labs at home during this admission.* [Doctor A]

Appreciation

Eleven SMS text messages were categorized into 1 code, reflecting gratitude from patients or caregivers for acknowledging the health care team's dedication, responsiveness, and compassion:

> *Hi doctor and medical team, thank you for very much for taking care of me.* [Patient E]

Trust and Feedback

Two codes categorized 10 SMS text messages related to expressions of trust and the exchange of feedback between patients, caregivers, and health care providers. These text messages reflected the growing confidence of patients in the clinical decisions made on their behalf and their willingness to voice concerns, ask questions, and engage in open dialogue:

> *My blood pressure [is] very high, 179/113 [mmHg]. I am worried.* [Patient F]

In addition to positive affirmations, a smaller number of messages conveyed dissatisfaction with aspects of care delivery. These instances of negative feedback highlighted areas where expectations were unmet and served as critical input for service improvement and patient-centered responsiveness in home-based care:

> *It is not [a] problem for me to continue [to] submit the vitals. However, I need a clear direction from your side. The direction does not [seem to be] clear [and] now it's just submission of vital, if [give] wrong medication, how?* [Patient G]

Courteous Communication

Forty-eight SMS text messages were categorized into 1 code reflecting the consistent use of courteous language, such as expressions of gratitude and polite message closures exchanged among patients, caregivers, and health care providers. Patients and caregivers frequently acknowledged and affirmed the guidance provided by health care professionals, demonstrating attentiveness and appreciation:

> *Thanks a lot and noted on this.* [Patient H]
>
> *I am thankful to have the [healthcare] team [and] to benefit from the hospital-at-home programme, [it was] very helpful to me and my family.* [Patient I]

XSL•FO
RenderX

Communication Timeliness

Ten SMS text messages were categorized into 3 codes that addressed the timeliness of responses exchanged between patients, caregivers, and health care providers within digital communication. These codes captured a subset of messages highlighting delayed responses related to administrative matters after standard office hours (6 PM), involving both patients and health care providers:

> *Hi Mr [patient], awaiting your email address.* [Doctor A]
>
> *Sorry to revert back so late, we will discharge you tomorrow as planned.* [Doctor A]

In some instances, calls initiated by patients were not immediately answered by health care providers. Nevertheless, health care providers consistently made efforts to follow up and respond as soon as possible:

> *Hi Mr [patient], we got a missed call from you. Is there anything we can help you?* [Doctor A]

Tone and Professional Etiquette

A total of 573 SMS text messages were categorized into 4 codes related to the tone of communication exchanged between patients, caregivers, and health care providers in digital messaging. The tone varied according to context, with both formal and informal language observed across interactions. A formal tone was predominantly used in messages concerning clinical guidance, updates, or health-related discussions, demonstrating professionalism and clarity in clinical communication. In contrast, an informal tone was more common in exchanges regarding administrative matters such as scheduling or logistics, suggesting a more conversational and approachable style suited to routine coordination.

> *Good morning, would like to remind [you] to submit today's morning vitals please. Thank you.* [Doctor A]
>
> *Hi [doctor], my dad agrees to the infusion. Please let me know the arrangements once you have the details, many thanks.* [Caregiver F]
>
> *Hi! Btw can you please help submit the consent form again but this time in your dad's name? haha so sorry thank you!* [Doctor C]

Discussion

Principal Findings

This study examined the content of SMS text messages exchanged between health care providers and patients or caregivers within a HaH program in Singapore, offering insights into clinical communication and the relational dynamics enabling remote care delivery. Although the findings were derived from a cohort composed entirely of patients with COVID-19 during August 2022 to October 2022 and comprised low-acuity cases with no escalation to inpatient care, they offer potentially transferable insights into asynchronous communication processes within HaH settings. However, their applicability to other diagnoses should be interpreted cautiously. To our knowledge, this is the first content analysis of asynchronous communications within a HaH service in Singapore.

Communication related to the clinical aspects of care emerged as a key category in text messages exchanged between patients, caregivers, and health care providers within the HaH program. These interactions were centered on clinical consultations, where patients and caregivers sought medical advice related to symptom management, medication clarification, and general health concerns. In parallel, health care providers used these messages to deliver clinical updates, offer guidance, and adjust care plans based on patient conditions. This closely aligns with studies conducted in the United States and China, suggesting that SMS text messaging–based communication between clinicians and patients frequently involved symptom discussions, disease management support, and medication concerns [29-31]. Additionally, frequent SMS text message exchanges related to the monitoring of clinical parameters, such as reminders to report vital sign readings initiated by both health care providers and patients or caregivers, were observed. These were often followed by health care providers' interpretations of the submitted data, which subsequently guided further clinical decision-making. This communication pattern aligns with findings from a systematic review, which emphasized that remote vital signs monitoring is a foundational element of HaH programs, facilitating the early detection of clinical deterioration and timely intervention when necessary [32]. Collectively, these observations underscore the value of asynchronous communication as a critical enabler of clinical continuity in home-based care. Through timely and responsive digital communications, it bridges the gap between patients and health care providers, allowing for ongoing monitoring, real-time decision support, and patient engagement that aligns with the core principles of the HaH model. As such, future research may benefit from examining how the frequency and quality of digital clinical communication can influence patient outcomes, such as symptom resolution and overall satisfaction with care.

Administrative and transport arrangements were identified as another key category in this study, highlighting the importance of nonclinical operational components in enabling effective care delivery within the HaH model. Numerous SMS text messages reflected the coordination of logistical tasks and administrative and transport-related processes. These interactions included scheduling home visits, organizing the delivery of medications and medical consumables, and facilitating transitions between care settings, such as hospital discharge to home or readmission when clinically indicated. Moreover, administrative communications reflected efforts to clarify documentation requirements, streamline admission and discharge workflows, and address patients' or caregivers' queries related to billing. These findings are consistent with studies conducted in the United States and the Netherlands, suggesting that digital messaging platforms are frequently used to manage a wide range of clerical and logistical tasks, such as administrative coordination, logistical planning, and transport arrangements [33,34]. When effectively streamlined, these processes can contribute to a reduction in delays in care delivery and improved patient experiences within remote care settings [35]. Interestingly, qualitative studies conducted in Singapore have

XSL•FO
RenderX

highlighted that logistical and operational hurdles are inevitable in the implementation of HaH programs, potentially disrupting care continuity and resulting in delays, increased health care provider workload, and reduced patient satisfaction [36,37]. Therefore, addressing these issues is essential to optimize the effectiveness and scalability of HaH services in Singapore. Furthermore, the establishment of clear and reliable communication pathways between health care teams and patients or caregivers was observed. These interactions included identifying preferred modes of communication, whether via SMS text messaging or telephone calls, and designating a consistent point of contact for the care team. Similarly, this is consistent with a study conducted in the United States, highlighting the importance of aligning communication channels with patient preferences to enhance engagement and satisfaction [38]. Thus, clarifying communication preferences and responsibilities not only fosters continuity of care but also supports timely responsiveness and reduces fragmentation in care delivery. All in all, these findings underscore the integral role of digital communication in supporting the administrative and logistical infrastructure of HaH models. Therefore, stakeholders may consider engaging in co-design processes with end users of HaH programs to develop a tailored digital application aimed at streamlining operational workflows, further alleviating the administrative burden on health care teams while enhancing the overall patient experience and satisfaction.

Another key category identified in this study was the quality of interpersonal dynamics reflected in the digital text message exchanges between the health care providers, patients, and caregivers. These interactions consisted of expressions of empathy and emotional reassurance from health care providers and words of appreciation and gratitude from patients and caregivers in response to the care they received. Notably, although these interactions reflected elements of emotional support, explicit discussions of mental health conditions were not observed. This may reflect the clinically oriented nature of routine HaH communication, the likelihood that sensitive psychosocial concerns were addressed through alternative channels (eg, through synchronous audio or video calls), or a reluctance to disclose mental health issues within an SMS text messaging–based clinical exchange [39,40]. Furthermore, respectful and courteous language was observed across both parties, reinforcing the role of digital communication in fostering trust and mutual respect. These findings are consistent with multisite studies conducted in Australia, Ireland, the United Kingdom, Spain, and Germany, which similarly observed empathetic communication, emotional support, and courteous exchanges within digital platforms [40,41]. Moreover, the presence of digital empathy within asynchronous SMS text messaging–based interactions has been shown to foster rapport, encourage participant engagement, and contribute to positive patient experiences in HaH models [42]. Such relational interactions can potentially help to humanize remote care delivery and are instrumental in cultivating therapeutic alliances, which may positively influence patient engagement, satisfaction, and perceived quality of care. Furthermore, with the contemporary workforce gravitating toward asynchronous communication, professional environments have widely integrated digital tools to facilitate collaboration and optimize team productivity [43]. Future research could explore how these relational dynamics shape the experience of patients, caregivers, and health care providers, which could inform the development of culturally sensitive communication frameworks and targeted training programs for health care providers that will support high-quality relational care in digitally mediated HaH settings.

Timeliness of communication and professional tone were identified as another key category in the SMS text message exchanges between health care providers, patients, and caregivers in this study. Several interactions revealed delayed responses to nonclinical inquiries, particularly those sent after standard office hours. This finding aligns with studies conducted in China and the United States, which also reported frequent response delays after office hours [29,44]. Notably, timely communication has been shown to significantly influence patient satisfaction, likely due to its role in reinforcing feelings of safety, emotional reassurance, and trust, which are essential qualities valued in HaH models of care [44,45]. In addition to timeliness, the tone of communication provided further insights into the relational dynamics between health care providers, patients, and caregivers. Our findings highlighted that a formal tone was often used in messages on clinical care or medical guidance, reflecting a focus on clarity and professionalism. In contrast, informal tones were more frequently observed in administrative or logistical messages, indicating a conversational style. These patterns are congruent with studies conducted in the United States, which highlighted that clinically significant messages were communicated in a more formal tone, while operational matters were addressed in a neutral or informal tone [46,47]. Together, these findings suggest that tone modulation according to message purpose may play an important role in balancing clinical professionalism with relational rapport in HaH care. Future research could explore how tone preferences vary across cultural and demographic groups and examine how tone influences perceived empathy, patient engagement, and satisfaction within digital health care communication.

Implications of Future Practice and Research

Our findings suggest several recommendations to strengthen digital communication practices within HaH care. First, 2-way asynchronous communication among patients, caregivers, and providers appears to be an important enabler of HaH care, helping patients and caregivers feel connected and supported, while allowing the care team to convey information efficiently. Second, the use of an existing platform that patients and caregivers are highly familiar with is critical to optimize their responsiveness and experience. Third, guidelines for asynchronous communications for HaH teams will be helpful to optimize clinical effectiveness and patient engagement. This may include types of content suitable for asynchronous communication (eg, instructive or informative rather than discussive), language and tone (eg, professional yet friendly), target response times, and after-hour response expectations. Dedicated training may be necessary for health care providers on the effective use of asynchronous communication tools as their use expands. Understanding the nature of these communications also has implications on the development of future AI-assisted communication tools that can be trained to answer such queries.

XSL•FO
RenderX

Future research directions can explore differences in communication among older adults and how SMS text messaging–based relational elements, such as empathy and mutual respect, shape the experiences of patients and caregivers. Intervention-focused studies could build on these findings using theory-informed approaches. Further work could also examine how tone preferences vary across cultural or demographic groups. Our findings also suggest a need for digital platforms that are better tailored to the operational realities of HaH care to streamline administrative tasks and reduce the burden on health care teams. Future research could also explore how generative AI technologies, such as large language models and conversational agents, could augment SMS text messaging–based communication within HaH programs by supporting routine queries and summarizing message threads. Nevertheless, given the central role of empathy and relational dynamics identified in this study, careful consideration will be critical to ensure that such innovations support the human dimensions of care. These offer valuable opportunities to inform more patient-centered, culturally attuned, and operationally effective communication strategies that support high-quality HaH care.

Limitations

This study has several limitations. First, the dataset primarily comprised English-language communications, which may not fully capture the experiences of non-English-speaking participants. This limits the inclusivity and cultural representativeness of the findings, particularly in a multilingual context like Singapore. Additionally, the sample did not adequately reflect the broader demographic trends of an aging population, who are key users of the HaH services. Hence, future studies may benefit from purposive sampling strategies that include a more diverse linguistic and age profile, such as older adults, to provide a more comprehensive understanding of digital communication in HaH care. Second, the retrospective nature of the SMS text message data limits insights into the broader context of communication exchanges, as asynchronous messages may not have fully captured the full scope of interactions, particularly those that occurred following or preceding a verbal or face-to-face encounter. As such, future studies may consider triangulating message data with complementary methods, such as semistructured interviews or ethnographic observations, to better contextualize communication dynamics. Third, due to the time period of sampling, all patients were admitted for COVID-19. However, although remote monitoring, home visits, and virtual consultations are also used in other HaH diagnostic groups, communication patterns may differ by condition severity, symptom profile, and care trajectory. Therefore, transferability to a broader HaH population should be interpreted with caution. Finally, although qualitative content analysis was used to derive insights, the potential for researcher bias remains inherent due to the interpretative process. To mitigate this, systematic coding procedures and regular team discussions were used within the research team.

Conclusion

This study offered valuable insights into the nuances of digital communication between health care providers, patients, and caregivers within a HaH program in Singapore, in the context of the COVID-19 pandemic. The findings underscore the critical interplay between operational efficiency, the quality of interpersonal relationships, and clinical effectiveness in shaping the overall success of digitally mediated care within HaH settings. These interconnected elements not only facilitate seamless care coordination but also foster patient trust, engagement, and overall satisfaction. These insights underscore the need for communication frameworks and training programs that are clinically robust, operationally efficient, and attuned to the relational dynamics essential for high-quality home-based care. As HaH models continue to expand, future research should integrate these communication principles into health care provider training, digital platform development, and implementation strategies to promote more person-centered, culturally responsive, and sustainable care delivery.

Acknowledgments

The authors would like to extend their gratitude to all participants enrolled in the hospital-at-home program. They would also like to thank Bernedette, Daphne, and Joanne for their preliminary work on this study. The authors declare that generative artificial intelligence tools were not used at any stage in the preparation of this manuscript.

Funding

This work was supported by the National Medical Research Council Population Health Research Grant (PHRGOC23Jan-0002).

Data Availability

The datasets generated or analyzed during this study are not publicly available due to privacy and ethical restrictions; however, they are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: SK, SS
Data analysis: JSLS, SYL, SS
Data curation: JSLS, SYL, SS
Investigation: JSLS, SYL, SS

XSL•FO
RenderX

Methodology: SK, SS
Project administration: JSLS, SS
Supervision: JSLS, SS
Validation: JSLS, SK, SYL, SS
Visualization: JSLS, SYL, SS
Writing – original draft: JSLS, SK, SYL, SS
Writing – review and editing: JSLS, SK, SYL, SS

Conflicts of Interest

None declared.

Multimedia Appendix 1
Coding scheme of text messages
[DOCX File, 21 KB - jmir_v28i1e83530_app1.docx]

Checklist 1
SRQR checklist.
[DOCX File, 18 KB - jmir_v28i1e83530_app2.docx]

References

1. Jones CH, Dolsten M. Healthcare on the brink: navigating the challenges of an aging society in the United States. NPJ Aging 2024 Apr 6;10(1):22. [doi: 10.1038/s41514-024-00148-2] [Medline: 38582901]
2. Kanagala SG, Gupta V, Kumawat S, Anamika F, McGillen B, Jain R. Hospital at home: emergence of a high-value model of care delivery. Egypt J Intern Med 2023;35(1):21. [doi: 10.1186/s43162-023-00206-3] [Medline: 36969500]
3. Wallis JA, Shepperd S, Makela P, et al. Factors influencing the implementation of early discharge hospital at home and admission avoidance hospital at home: a qualitative evidence synthesis. Cochrane Database Syst Rev 2024 Mar 5;3(3):CD014765. [doi: 10.1002/14651858.CD014765.pub2] [Medline: 38438114]
4. Edgar K, Iliffe S, Doll HA, et al. Admission avoidance hospital at home. Cochrane Database Syst Rev 2024 Mar 5;3(3):CD007491. [doi: 10.1002/14651858.CD007491.pub3] [Medline: 38438116]
5. Leong MQ, Lim CW, Lai YF. Comparison of hospital-at-home models: a systematic review of reviews. BMJ Open 2021 Jan 29;11(1):e043285. [doi: 10.1136/bmjopen-2020-043285] [Medline: 33514582]
6. Yeung AWK, Torkamani A, Butte AJ, et al. The promise of digital healthcare technologies. Front Public Health 2023;11:1196596. [doi: 10.3389/fpubh.2023.1196596] [Medline: 37822534]
7. Ezeamii VC, Okobi OE, Wambai-Sani H, et al. Revolutionizing healthcare: how telemedicine is improving patient outcomes and expanding access to care. Cureus 2024 Jul;16(7):e63881. [doi: 10.7759/cureus.63881] [Medline: 39099901]
8. Cerdan de Las Heras J, Andersen SL, Matthies S, et al. Hospitalisation at home of patients with COVID-19: a qualitative study of user experiences. Int J Environ Res Public Health 2023 Jan 10;20(2):1287. [doi: 10.3390/ijerph20021287] [Medline: 36674043]
9. Liu X, Sutton PR, McKenna R, et al. Evaluation of secure messaging applications for a health care system: a case study. Appl Clin Inform 2019 Jan;10(1):140-150. [doi: 10.1055/s-0039-1678607] [Medline: 30812040]
10. Chua CMS, Ko SQ, Lai YF, Lim YW, Shorey S. Perceptions of stakeholders toward “hospital at home” program in Singapore: a descriptive qualitative study. J Patient Saf 2022 Apr 1;18(3):e606-e612. [doi: 10.1097/PTS.0000000000000890] [Medline: 34406987]
11. Chua CMS, Ko SQ, Lai YF, Lim YW, Shorey S. Perceptions of hospital-at-home among stakeholders: a meta-synthesis. J Gen Intern Med 2022 Feb;37(3):637-650. [doi: 10.1007/s11606-021-07065-0] [Medline: 34363185]
12. Pandit JA, Pawelek JB, Leff B, Topol EJ. The hospital at home in the USA: current status and future prospects. NPJ Digit Med 2024 Feb 27;7(1):48. [doi: 10.1038/s41746-024-01040-9] [Medline: 38413704]
13. Paulson MR, Torres-Guzman RA, Avila FR, et al. Telemedicine allows for effective communication between a medically complex cancer patient and his virtual hospital medical team. Clin Case Rep 2022 Oct;10(10):e6456. [doi: 10.1002/ccr3.6456] [Medline: 36245470]
14. Romagnoli KM, Handler SM, Hochheiser H. Home care: more than just a visiting nurse. BMJ Qual Saf 2013 Dec;22(12):972-974. [doi: 10.1136/bmjqs-2013-002339] [Medline: 23940375]
15. Ding J, Freeman M, Hasija S. Can predictive technology help improve acute care operations? Investigating the impact of virtual triage adoption. SSRN. Preprint posted online on Oct 2, 2023. [doi: 10.2139/ssrn.3806478]
16. Kraus S, Schiavone F, Pluzhnikova A, Invernizzi AC. Digital transformation in healthcare: analyzing the current state-of-research. J Bus Res 2021 Feb;123:557-567. [doi: 10.1016/j.jbusres.2020.10.030]

XSL•FO
RenderX

17. Howcroft A, Bennett-Weston A, Khan A, Griffiths J, Gay S, Howick J. AI chatbots versus human healthcare professionals: a systematic review and meta-analysis of empathy in patient care. Br Med Bull 2025 Sep 22;156(1):ldaf017. [doi: 10.1093/bmb/ldaf017] [Medline: 41115171]
18. Profka K, Wang A, Schriver E, et al. Patient interaction phenotypes with an automated SMS text message-based program and use of acute health care resources after hospital discharge: observational study. J Med Internet Res 2025 Jul 18;27:e72875. [doi: 10.2196/72875] [Medline: 40680254]
19. Terry C, Cain J. The emerging issue of digital empathy. Am J Pharm Educ 2016 May 25;80(4):58. [doi: 10.5688/ajpe80458] [Medline: 27293225]
20. Doyle L, McCabe C, Keogh B, Brady A, McCann M. An overview of the qualitative descriptive design within nursing research. J Res Nurs 2020 Aug;25(5):443-455. [doi: 10.1177/1744987119880234] [Medline: 34394658]
21. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
22. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: a systematic review of empirical tests. Soc Sci Med 2022 Jan;292:114523. [doi: 10.1016/j.socscimed.2021.114523] [Medline: 34785096]
23. Lincoln YS, Guba EG. Naturalistic Inquiry: Sage Publications; 1985. [doi: 10.1177/144078338702300329]
24. Teh YY, Lek E. Culture and reflexivity: systemic journeys with a British Chinese family. J Fam Ther 2018 Nov;40(4):520-536. [doi: 10.1111/1467-6427.12205]
25. Buetow S. Apophenia, unconscious bias and reflexivity in nursing qualitative research. Int J Nurs Stud 2019 Jan;89:8-13. [doi: 10.1016/j.ijnurstu.2018.09.013] [Medline: 30316055]
26. Boettger RK, Palmer LA. Quantitative content analysis: its use in technical communication. IEEE Trans Profess Commun 2010;53(4):346-357. [doi: 10.1109/TPC.2010.2077450]
27. Altheide DL, Schneider CJ. Qualitative Media Analysis, 2nd edition: Sage Publications; 2012. [doi: 10.4135/9781452270043]
28. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
29. Huang M, Fan J, Prigge J, Shah ND, Costello BA, Yao L. Characterizing patient-clinician communication in secure medical messages: retrospective study. J Med Internet Res 2022 Jan 11;24(1):e17273. [doi: 10.2196/17273] [Medline: 35014964]
30. Yoo W, Kim SY, Hong Y, Chih MY, Shah DV, Gustafson DH. Patient-clinician mobile communication: analyzing text messaging between adolescents with asthma and nurse case managers. Telemed J E Health 2015 Jan;21(1):62-69. [doi: 10.1089/tmj.2013.0359] [Medline: 25401324]
31. Zhai P, Hayat K, Ji W, et al. Efficacy of text messaging and personal consultation by pharmacy students among adults with hypertension: randomized controlled trial. J Med Internet Res 2020 May 20;22(5):e16019. [doi: 10.2196/16019] [Medline: 32432556]
32. Cheng SA, Tan SI, Goh SLE, Ko SQ. The value of remote vital signs monitoring in detecting clinical deterioration in patients in hospital at home programs or postacute medical patients in the community: systematic review. J Med Internet Res 2025 May 26;27:e64753. [doi: 10.2196/64753] [Medline: 40418800]
33. Hefner JL, MacEwan SR, Biltz A, Sieck CJ. Patient portal messaging for care coordination: a qualitative study of perspectives of experienced users with chronic conditions. BMC Fam Pract 2019 May 3;20(1):57. [doi: 10.1186/s12875-019-0948-1] [Medline: 31053063]
34. van Grootel JWM, Collet RJ, van Dongen JM, et al. Continuity and coordination of care during hospital-to-home transitions: healthcare professionals' perspectives. J Clin Nurs 2026 May;35(5):2429-2439. [doi: 10.1111/jocn.17758] [Medline: 40143559]
35. Norman GJ, Orton K, Wade A, Morris AM, Slaboda JC. Operation and challenges of home-based medical practices in the US: findings from six aggregated case studies. BMC Health Serv Res 2018 Jan 27;18(1):45. [doi: 10.1186/s12913-018-2855-x] [Medline: 29374478]
36. Ko S, Cheong D, Low SY, Shorey S. Hybrid hospital-at-home program in Singapore: ethnographic study. J Med Internet Res 2025 Jun 2;27:e66107. [doi: 10.2196/66107] [Medline: 40456139]
37. Towle RM, Fan PEM, Arba'in J, et al. Enhancing the hospital at home experience. Proc Singap Healthc 2023 Dec;32:20101058231209200. [doi: 10.1177/20101058231209200]
38. Clarke MA, Fruhling AL, Sitorius M, Windle TA, Bernard TL, Windle JR. Impact of age on patients' communication and technology preferences in the era of meaningful use: mixed methods study. J Med Internet Res 2020 Jun 1;22(6):e13470. [doi: 10.2196/13470] [Medline: 32478658]
39. Lee JL, Matthias MS, Huffman M, Frankel RM, Weiner M. Insecure messaging: how clinicians approach potentially problematic messages from patients. JAMIA Open 2020 Dec;3(4):576-582. [doi: 10.1093/jamiaopen/ooaa051] [Medline: 33758796]
40. Luetke Lanfer H, Reifegerste D, Weber W, et al. Digital clinical empathy in a live chat: multiple findings from a formative qualitative study and usability tests. BMC Health Serv Res 2024 Mar 8;24(1):314. [doi: 10.1186/s12913-024-10785-8] [Medline: 38459522]

XSL•FO
RenderX

41. Rey Velasco E, Demjén Z, Skinner TC, Impact Diabetes B2B Collaboration Group. Digital empathy in behaviour change interventions: a survey study on health coach responses to patient cues. Digit Health 2024;10:20552076231225889. [doi: 10.1177/20552076231225889] [Medline: 38528968]
42. Subramanya V, Spychalski J, Coats S, Gottstein E, Chancellor J, Kulshreshtha A. Empathetic communication in telemedicine: a pilot study. PRiMER 2024;8:36. [doi: 10.22454/PRiMER.2024.644242] [Medline: 38946757]
43. Lane JN, Leonardi PM, Contractor NS, DeChurch LA. Teams in the digital workplace: technology's role for communication, collaboration, and performance. Small Group Res 2024 Feb;55(1):139-183. [doi: 10.1177/10464964231200015]
44. Wu DC, Zhao X, Wu J. Online physician-patient interaction and patient satisfaction: empirical study of the internet hospital service. J Med Internet Res 2023 Aug 24;25:e39089. [doi: 10.2196/39089] [Medline: 37616031]
45. Zafar SN, Hazlewood G, Dhiman K, et al. "How are you?" Perspectives from patients and health care providers of text messaging to support rheumatoid arthritis care: a thematic analysis. ACR Open Rheumatol 2024 May;6(5):276-286. [doi: 10.1002/acr2.11652] [Medline: 38376004]
46. Alpert JM, Campbell-Salome G, Gao C, et al. Secure messaging and COVID-19: a content analysis of patient-clinician communication during the pandemic. Telemed J E Health 2022 Jul;28(7):1028-1034. [doi: 10.1089/tmj.2021.0316] [Medline: 34767741]
47. Hogan TP, Luger TM, Volkman JE, et al. Patient centeredness in electronic communication: evaluation of patient-to-health care team secure messaging. J Med Internet Res 2018 Mar 8;20(3):e82. [doi: 10.2196/jmir.8801] [Medline: 29519774]

Abbreviations

AI: artificial intelligence
HaH: hospital-at-home
NUHS: National University Health System
SRQR: Standards for Reporting Qualitative Research

Edited by A Stone; submitted 04.Sep.2025; peer-reviewed by B Petrovic, C Baxter; revised version received 21.Apr.2026; accepted 21.Apr.2026; published 08.Jun.2026.

Please cite as:
Seow JSL, Ko S, Low SY, Shorey S
Clinician-Patient Asynchronous Text Messaging Communication in Hospital-at-Home Care: Qualitative Study
J Med Internet Res 2026;28:e83530
URL: https://www.jmir.org/2026/1/e83530
doi:10.2196/83530

© Jeremy Soon Leong Seow, Stephanie Ko, Shi Yun Low, Shefaly Shorey. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Telehealth Usability, Engagement Patterns, and Technical Infrastructure in Managing Noncommunicable Diseases Among Health Care Professionals in Brazil, Ghana, Honduras, and the United Kingdom: Multinational Cross-Sectional Study

Ken Brackstone[1*], MSc, PhD; Roberta Lins Gonçalves[2,3*], PT, MSc, PhD; Adriana Silvina Pagano[4], MSc, PhD; Zilma Silveira Nogueira Reis[4], MSc, MD, PhD; Seth Kwaku Afagbedzi[5], MSc, MD, PhD; Lysien Ivania Zambrano[6], MD, MSc, PhD; Tainá Costa Pereira Lopes[2], PT, MSc; Sarah Almeida Cordeiro[2], MSc; Julia Macedo Nunes[4], BLitt, MSc; Wagner Meira Jr[4], PhD; James Batchelor[1], MD; Antonio Luiz Pinho Ribeiro[4], MSc, MD, PhD

[1]University of Southampton, University Road, Southampton, England, United Kingdom

[2]Universidade Federal do Amazonas, Manaus, Amazonas, Brazil

[3]Universidade Federal de Juiz de Fora, Juiz de Fora, Minas Gerais, Brazil

[4]Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

[5]University of Ghana, Accra, Greater Accra, Ghana

[6]National Autonomous University of Honduras, Tegucigalpa, Francisco Morazán Department, Honduras

[*]these authors contributed equally

Corresponding Author:
Ken Brackstone, MSc, PhD
University of Southampton, University Road, Southampton, England, United Kingdom

Abstract

Background: Noncommunicable diseases (NCDs) account for over 70% of global deaths, with hypertension and diabetes serving as major contributors. The COVID-19 pandemic disrupted traditional health care services for NCDs and highlighted telehealth as a crucial alternative. Telehealth—encompassing synchronous and asynchronous electronic communication to deliver clinical services remotely—can overcome geographical barriers and enhance patient engagement. However, telehealth usability among health care professionals (HCPs) remains under-studied across low-, middle-, and high-income countries.

Objective: This study aimed to examine which telehealth engagement patterns, technical infrastructure factors, and user profiles were most strongly associated with usability among HCPs and to descriptively compare these across 4 diverse countries: Brazil (high- to middle-income country), Ghana (low- to middle-income country), Honduras (low- to middle-income country), and the United Kingdom (high-income country).

Methods: A multinational cross-sectional survey was conducted with 290 HCPs across 4 countries. Participants completed the System Usability Scale and provided data on telehealth engagement (eg, frequency, duration, and number of systems used), technical infrastructure (connection stability and support satisfaction), and their user profile (demographics, job role, and training received). Descriptive statistics summarized these patterns and usability scores. Multiple linear regression with bootstrap-based sensitivity analyses identified factors associated with telehealth usability. Given the nonprobability design, no formal inferential comparisons were made between countries. Instead, observed patterns were reported descriptively.

Results: Higher telehealth usability scores were associated with greater connection stability (b=5.06, 95% CI 3.06-7.05), higher satisfaction with online support information (b=5.02, 95% CI 3.27-6.75), more frequent use (b=3.05, 95% CI 1.36-4.73), longer duration of use (b=1.59, 95% CI 0.49-2.68), and being a physician by profession (b=3.82, 95% CI 0.23-7.40). Average usability scores were highest among users in Ghana (mean 79.75, SD 14.19) and the United Kingdom (mean 79.00, SD 14.71), followed by Brazil (mean 72.01, SD 14.62) and Honduras (mean 63.09, SD 15.57). According to System Usability Scale guidelines, scores corresponded to “good” usability for users in Ghana, the United Kingdom, and Brazil and were below the “good” threshold for users in Honduras. While most users in Ghana (97/111, 87.4%), Honduras (31/38, 81.6%), and Brazil (57/80, 70.4%) reported using only 1 telehealth system, two-thirds of UK users (40/60, 66.7%) reported using 2 or more systems. User profiles also varied; prepandemic use was highest in Ghana (84/111, 75.7%) and lowest in Honduras (7/38, 18.4%). Other engagement patterns across countries were reported.

Conclusions: Telehealth usability is driven by technical infrastructure reliability, a robust online support infrastructure, and an “experience effect” from frequent and long-term engagement. Descriptive differences in engagement patterns and infrastructure

XSL•FO
RenderX

highlight the need for tailored strategies to address setting-specific challenges. These are essential to optimize telehealth integration and improve health care outcomes for patients with NCDs worldwide.

(J Med Internet Res 2026;28:e64070) doi:10.2196/64070

KEYWORDS

telehealth systems; usability; noncommunicable diseases; low- to middle-income countries; LMICs; high- to middle-income countries; HMICs; high-income countries; HICs

Introduction

Noncommunicable diseases (NCDs) are responsible for over 70% of deaths worldwide, with hypertension and diabetes being major contributors to NCD-related mortality, including cardiovascular disease and stroke [1]. NCDs are often chronic and progressive and require complex and continuous management of physiological, environmental, and behavioral factors [1-3]. Treatment costs for both patients and health services are high and place substantial economic pressure on health care systems [1-3]. Low- to middle-income countries (LMICs) experience higher NCD mortality rates than high-income countries (HICs) [1,4]. This is driven in part by a rapid epidemiological transition characterized by increased consumption of processed foods, reduced physical activity, and rapid urbanization alongside persistent exposure to tobacco and alcohol products [1]. In response, the United Nations 2030 Agenda for Sustainable Development committed to developing new approaches to reduce premature mortality from NCDs by one-third [5]. As a result, innovative and cost-effective solutions to improve care for NCDs are becoming increasingly prevalent. These include telehealth, mobile health apps, remote monitoring systems, and electronic health platforms that facilitate communication between patients and health care professionals (HCPs) [6,7].

While telehealth has existed for decades, the COVID-19 pandemic necessitated a rapid shift in its clinical application. This occurred after a large number of patients with NCDs were reportedly lost to follow-up due to the increased risk of illness and death associated with face-to-face consultations [8,9]. As a result, telehealth quickly transitioned from a supplementary service to an essential platform for HCPs to maintain contact with patients [10]. “Telehealth” is a broad term for health care that is delivered remotely by a variety of clinicians and HCPs. Patients and clinicians do not need to be located in the same clinic or hospital. Instead, HCPs can deliver a range of clinical services—health assessment, consultation, diagnosis, treatment, supervision, and information—to patients remotely regardless of geographic barriers [9,11,12]. Thus, telehealth can help patients become active and more efficient participants in their own health outcomes, particularly in underserved areas, and is used in the comfort, convenience, and safety of the patients’ own homes [13]. Studies have shown similar or better results with telehealth than with traditional care for some chronic conditions [14].

The range of telehealth modalities for HCPs includes synchronous use (eg, face-to-face interactions in real time, including video consultations and web-based videoconferencing) and asynchronous use (eg, email, SMS text messages, teledermatology via image transmission, remote monitoring, and electronic transmission of prescriptions to pharmacists), which can take place in clinical and nonclinical settings [4,11]. These varying engagement patterns represent different technical requirements and clinical workflows. However, to reap the optimal benefits of telehealth, it must be considered usable by both patients and HCPs [15-19]. Studies have shown that usability is an essential attribute of technology systems and is related to acceptability by end users [20-23]. Telehealth is sometimes viewed as disruptive and complex by HCPs, requiring specialized strategies to manage its use—particularly among specific user profiles, such as HCPs with limited telehealth experience or varying levels of prior training [6,16,17,19]. Furthermore, technical infrastructure can impair the usability and effectiveness of telehealth, such as system failures or connectivity issues [17]. Preliminary reports during the early stages of the pandemic indicated that HCPs’ reluctance to use telehealth technologies was one of the main barriers to implementation in clinical practice [17].

Prior to the onset of the COVID-19 pandemic, telehealth was mostly used in clinical settings across HICs and high- to middle-income countries (HMICs) [24]. Conversely, numerous LMICs have encountered and continue to encounter significant infrastructure-related challenges in adopting the technology [25]. Many individuals living in LMICs have restricted admittance to medical services, travel significant distances to receive medical care, and postpone medical care until they experience a health emergency [26,27]. However, since the pandemic, a number of primary care practices across HMICs and LMICs have been forced to implement telehealth in diverse geographical areas [6]. The rapid expansion of telehealth across LMICs has strengthened the potential of telehealth in delivering effective health care, particularly among patients with NCDs.

User acceptance remains an essential ingredient for maintaining telehealth use. Existing research has primarily evaluated telehealth usability from the patient perspective [15-19]. However, it is equally important to assess telehealth usability from the perspectives of HCPs—a topic much less studied [11,16,28]. A systematic review recently concluded that more studies should assess telehealth usability in LMICs to paint a broader picture of challenges faced using telehealth worldwide [28]. Thus, a comparison of the characteristics of telehealth use and usability perceptions across HMICs and LMICs can offer enormous contributions to global health by facilitating telehealth consolidation and assisting in expanding telehealth to countries where health service access is limited.

The overall aim of this study was to examine factors associated with telehealth usability among HCPs managing NCDs across diverse health care settings. Adopting an exploratory approach,

XSL•FO
RenderX

we first sought to identify which telehealth engagement factors, technical infrastructure elements, and user profiles were most strongly associated with perceived system usability. We anticipated that technical reliability—specifically connection stability and satisfaction with support infrastructure—would be primary correlates of usability perceptions. Second, this study sought to describe how these engagement patterns and technical contexts varied across 4 countries representing distinct health care environments: Brazil (HMIC), Ghana (LMIC), Honduras (LMIC), and the United Kingdom (HIC). By identifying these factors, we aimed to provide evidence-based insights to optimize the technical and clinical integration of telehealth for NCD management worldwide.

Methods

Study Design

An international cross-sectional online survey was conducted in Brazil, Ghana, Honduras, and the United Kingdom. The survey was reported in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for cross-sectional studies (Checklist 1 [29]).

Participants and Recruitment

A nonprobability convenience sampling approach was used to recruit HCPs from a range of health care settings, including hospitals, health centers, and clinics specializing in NCD management. We restricted the sample to HCPs working in primary care settings as the long-term management and monitoring of common NCDs such as hypertension and diabetes mostly takes place in primary care [30]. Furthermore, telehealth was widely implemented in these settings during the COVID-19 pandemic to support routine follow-up consultations [31]. Recruitment followed a systematic identification process via professional networks and institutional collaborations. Potential participants were contacted via email, telephone, or personal invitation by the project leaders in each country. Individuals were provided with a standardized information sheet regarding the study's purpose before being sent the digital survey link via email. In Brazil, participants were contacted primarily in the state of Minas Gerais, a large inland state in southeastern Brazil. In Honduras, participants were contacted within Tegucigalpa, the capital city. In Ghana, participants were contacted in Accra, the capital city, located on the southern coast of the country. Finally, in the United Kingdom, participants were contacted in 6 counties from the northeast to the south of England. Data collection was conducted over a 9-month period between August 2022 and May 2023.

The target sample size was determined to provide a baseline for descriptive precision. We initially calculated a minimum requirement of 264 participants (n=66 per country) to estimate mean usability scores with 95% confidence and a 5% margin of error based on previously reported System Usability Scale (SUS) values (mean 71.4, SD 11.6) [31] and an estimated population of 500 HCPs. While this study used a nonprobability design not intended for formal inferential comparisons between countries, this recruitment target was established to ensure sufficient total volume for descriptive reporting by setting and to provide adequate statistical power for the multiple linear regression.

Measures

Engagement Patterns

Participants first reported the number of telehealth systems they used in their current clinical practice (1, 2, or 3 or more) and identified the name of the system used most frequently. All subsequent questions referred to this primary system. Participants then stated their frequency of use (1="every 2-3 months"; 4="every day") and duration of use (1="0-3 months"; 7="5+ years"). After receiving definitions of synchronous and asynchronous telehealth use, participants then rated how often they used each method (1="never"; 5="very often"). Finally, participants selected the telehealth features they used most regularly from a list of 10 (eg, exchange of or sending information to patients) and indicated the number of NCDs they treated via telehealth from a list of 20 (eg, hypertension and diabetes).

User Profiles

Participants indicated their professional background and digital experience. This included whether they received telehealth training (yes or no), whether they had used telehealth systems prior to the COVID-19 pandemic (yes or no), and whether they were required to be on clinical premises to use their telehealth system (yes or no). Finally, participants rated their perceived level of competence with digital tools (1="very poor"; 5="excellent").

Telehealth Usability

Participants completed the SUS, a 10-item scale providing usability scores from 0 to 100 [32-34]. Participants indicated their agreement with 10 items (eg, "This telehealth system is easy to use"; 0="strongly disagree"; 4="strongly agree"). Scores were calculated by summing the individual item scores and multiplying by 2.5. Following established guidelines, scores of 60 to 69 were considered "moderate," scores of 70 to 79 were considered "good," and scores above 80 were considered "excellent."

Technical Infrastructure

Participants rated 2 key infrastructural factors regarding their primary system: their satisfaction with available online help and support information (1="strongly disagree"; 5="strongly agree") and the stability of their internet connection during a typical telehealth session (1="very poor"; 5="excellent").

Demographic Variables

Participants indicated their gender, age (categorized as 24-39, 40-49, 50-59, or ≥60 years), and occupation (physician or nonphysician). In Ghana and Honduras, participants also reported their practice setting (rural or urban).

Statistical Analysis

We first used descriptive statistics to summarize participant demographics. Next, we conducted multiple linear regression to identify factors associated with usability within the overall sample. Given the exploratory nature of this study, variables

XSL•FO
RenderX

with a *P* value below .20 in univariate models were entered into the multiple linear regression model following previous recommendations [35]. Prior to analysis, we evaluated the assumptions of linear regression. Some predictors (eg, connection stability) were ordinal in nature and exhibited negative skew. As such, they were treated as continuous following established precedents for Likert-type scales and to maintain model parsimony [36]. To account for potential biases related to this assumption, we used bootstrap-based sensitivity analyses with 5000 resamples to ensure the robustness of our estimates against violations of normality. We evaluated multicollinearity using tolerance and the variance inflation factor; all tolerance values were within acceptable ranges (tolerance>0.6; variance inflation factor<2)

Given the nonprobability sampling design, we did not perform formal inferential comparisons between countries. Instead, we summarized telehealth usability, user profiles, telehealth engagement patterns, and technical infrastructure by country to observe descriptive patterns across different health care settings. Analyses were conducted using SPSS Statistics (version 30; IBM Corp).

Ethical Considerations

Ethics approval was obtained by the respective institutional review boards in all participating countries: Brazil (Research Ethics Committee of the Federal University of Minas Gerais, Certificate of Submission of Ethical Review 56604122.5.0000.5149; 5.380.538), Honduras (Biomedical Research Ethics Committee of the National Autonomous University of Honduras 075-2022), Ghana (Ghana Health Service Ethics Review Committee 04/19/22), and the United Kingdom (Ethics and Research Governance Online ID: 72962 – University of Southampton). The survey protocol was previously published [37]. Study information was provided on the first page of the digital questionnaire, and all participants provided informed consent electronically by selecting a mandatory agreement box prior to commencing the survey. Participant data were fully anonymized at the point of collection to ensure privacy and confidentiality, with no personally identifiable information stored. Participation was entirely voluntary, and no financial or material compensation was provided.

Results

Demographics and User Profiles

Table 1 shows demographics and user profiles by country. A sample of 290 HCPs completed the survey (mean age 41.00, SD 7.45; range 24-65 years). Telehealth users from Ghana (n=111), Brazil (n=81), the United Kingdom (n=60), and Honduras (n=38) completed the survey. Most participants were men (150/290, 51.7%), aged 40 to 49 years (141/290, 48.6%), physicians by profession (201/290, 69.3%), and based at public hospitals (190/290, 65.5%) apart from users in Ghana, who were based mostly at private hospitals (75/111, 67.6%). In terms of practice setting, most participants in Ghana and Honduras reported working in urban areas (104/111, 93.7%, and 37/38, 97.4%, respectively).

XSL•FO
RenderX

Table . Demographics, user profiles, and telehealth engagement by country (categorical variables; N=290).

Characteristic	Total, n (%)	Brazil (n=81), n (%)	Ghana (n=111), n (%)	Honduras (n=38), n (%)	UK (n=60), n (%)
Gender					
Men	150 (51.7)	19 (23.5)	86 (77.5)	19 (50)	26 (43.3)
Women	140 (48.3)	62 (76.5)	25 (22.5)	19 (50)	34 (56.7)
Age (y)					
24-39	115 (39.7)	31 (38.3)	37 (33.3)	29 (76.3)	18 (30)
40-49	141 (48.6)	42 (51.9)	69 (62.2)	3 (7.9)	27 (45)
50-59	28 (9.7)	7 (8.6)	5 (5.5)	4 (10.5)	12 (20)
≥60	6 (2.1)	1 (1.2)	0 (0)	2 (5.3)	3 (5)
Job role					
Physician	201 (69.3)	14 (17.3)	105 (94.6)	38 (100)	44 (73.3)
Other	89 (30.7)	67 (82.7)	6 (5.4)	0 (0)	16 (26.7)
Sector[a]					
Public	190 (65.5)	73 (92.4)	36 (32.4)	30 (78.9)	51 (85)
Private	98 (33.8)	6 (7.6)	75 (67.6)	8 (21.1)	9 (15)
Used telehealth before the pandemic					
No	128 (44.1)	41 (50.6)	27 (24.3)	31 (81.6)	29 (48.3)
Yes	162 (55.9)	40 (49.4)	84 (75.7)	7 (18.4)	31 (51.7)
Training received					
No	88 (30.3)	19 (23.5)	12 (10.8)	21 (55.3)	36 (60)
Yes	202 (69.7)	62 (76.5)	99 (89.2)	17 (44.7)	24 (40)
Required to be on clinical premises[a]					
No	178 (61.4)	65 (81.3)	37 (33.3)	32 (84.2)	44 (74.6)
Yes	110 (37.9)	15 (18.7)	74 (66.7)	6 (15.8)	15 (25.4)
Telehealth systems used[a]					
1	205 (70.7)	57 (71.3)	97 (87.4)	31 (81.6)	20 (33.3)
2 or more	84 (29.0)	23 (28.7)	14 (12.6)	7 (18.4)	40 (66.7)

[a]Two of the 290 participants did not fully complete all measures (n=1 from Brazil and n=1 from the United Kingdom). Percentages reflect and include these missing participants; therefore, percentages for the variables indicated do not add up to 100%.

A higher proportion of participants in Ghana (84/111, 75.7%) reported using telehealth systems before the COVID-19 pandemic compared with participants in the United Kingdom (31/60, 51.7%), Brazil (40/81, 49.4%), and Honduras (7/38, 18.4%). Furthermore, a higher proportion of participants in Ghana (99/111, 89.2%) and Brazil (62/81, 76.5%) reported having received training for their selected telehealth systems compared to participants in Honduras (17/38, 44.7%) and the United Kingdom (24/60, 40%).

Telehealth Usability, Engagement Patterns, and Technical Infrastructure

Table 2 shows means and SDs of telehealth usability, engagement patterns, and technical infrastructure by country. Descriptively, participants in Ghana (mean 79.75, SD 14.19; range 30-100) and the United Kingdom (mean 79.00, SD 14.71; range 32.50-100) reported higher average usability scores than those in Brazil (mean 72.01, SD 14.62; range 25-100) and Honduras (mean 63.09, SD 15.57; range 7.5-100; Table 2). According to SUS guidelines, scores from users in Ghana, the United Kingdom, and Brazil fell within the “good” range, whereas scores from users in Honduras fell below this threshold.

XSL•FO
RenderX

Table . Telehealth usability, engagement patterns, and technical infrastructure by country (continuous or ordinal variables; N=290).

	Total, mean (SD)	Brazil (n=81), mean (SD)	Ghana (n=111), mean (SD)	Honduras (n=38), mean (SD)	UK[a] (n=60), mean (SD)
System usability (0-100)	75.25 (15.62)	72.01 (14.62)	79.75 (14.19)	63.09 (15.57)	79.00 (14.71)
Frequency of telehealth use (1-4)	3.15 (1.00)	3.00 (0.82)	3.50 (0.81)	1.97 (1.05)	3.43 (0.89)
Duration of telehealth use (1-7)	4.37 (1.62)	4.88 (1.84)	4.21 (1.36)	2.97 (1.44)	4.87 (1.26)
Frequency of synchronous use (1-5)	2.07 (1.25)	2.53 (1.32)	1.38 (0.89)	2.47 (1.22)	2.50 (1.16)
Frequency of asynchronous use (1-5)	3.62 (1.15)	3.03 (1.15)	4.00 (0.89)	2.79 (1.02)	4.23 (1.00)
Connection stability (1-5)	3.82 (0.79)	3.88 (0.80)	3.78 (0.71)	3.37 (0.94)	4.12 (0.69)
Satisfaction with online support information (1-5)	3.84 (0.81)	3.84 (0.89)	4.23 (0.65)	3.00 (1.12)	3.63 (0.99)
Number of telehealth systems used	3.09 (1.67)	2.68 (1.88)	3.33 (1.12)	2.63 (1.65)	3.48 (2.08)
Number of NCDs[b] treated	4.28 (2.97)	2.88 (2.41)	4.57 (2.04)	4.11 (3.03)	5.75 (4.10)
Self-rated digital skills (1-5)	3.88 (0.81)	3.79 (0.83)	4.22 (0.59)	3.79 (0.84)	3.43 (0.85)

[a]UK: United Kingdom.

[b]NCD: noncommunicable disease.

Regarding engagement patterns, most participants in Ghana (97/111, 87.4%), Honduras (31/38, 81.6%), and Brazil (57/81, 70.4%) reported using only 1 telehealth system in their clinical roles, whereas 66.7% (40/60) of the participants in the United Kingdom reported using 2 or more. Descriptively, the frequency of telehealth use was most pronounced among participants in Ghana and the United Kingdom. While Ghanaian and UK participants reported a higher frequency of asynchronous telehealth use, Ghanaian participants reported the lowest nominal frequency of synchronous telehealth use. On average, UK participants reported managing the largest number of NCDs via telehealth (approximately 6 conditions) compared with participants in Brazil, Ghana, and Honduras.

In terms of technical infrastructure, participants in the United Kingdom reported the highest levels of connection stability, whereas Ghanaian participants reported the greatest satisfaction with available help and online support information.

Factors Associated With Telehealth Usability

Table 3 presents the results of the multiple linear regression. Diagnostic residual plots and formal testing indicated a slight negative skew, primarily driven by high ratings for connection stability and support satisfaction. To address these departures from normality and ensure the reliability of our inferences, we performed bootstrap-based sensitivity analyses with 5000 resamples. These analyses yielded results nearly identical to those of the original ordinary least squares regression, which confirmed model robustness. The final regression model was statistically significant ($F_{12,276}$=17.45; P<.001) and accounted for 40.7% of the variance in telehealth usability scores (adjusted R^2=0.407). Higher telehealth usability scores were associated with greater connection stability (b=5.06, 95% CI 3.06-7.05; SE 1.01; β=0.26; P<.001), higher satisfaction with online support information (b=5.02, 95% CI 3.27-6.75; SE 0.88; β=0.30; P<.001), participants who were physicians by profession (b=3.82, 95% CI 0.23-7.40; SE 1.82; β=0.11; P=.04), more frequent telehealth use (b=3.05, 95% CI 1.36-4.73; SE 0.85; β=0.19; P<.001), and longer duration of telehealth use (b=1.59, 95% CI 0.49-2.68; SE 0.55; β=0.17; P=.004).

XSL•FO
RenderX

Table . Multiple linear regression model identifying factors associated with telehealth usability (N=290)[a].

	Univariate model			Multiple linear regression model		
	b	R^2	*P* value	*b* (SE; 95% CI)	β	*P* value
Demographics						
Women (reference=men)	4.97	0.03	.007	2.17 (1.65; −1.08 to 5.42)	0.07	.19
Age	0.25	0.02	.04	−0.02 (0.10; −0.22 to 0.18)	−0.01	.87
Physician (reference=other)	2.71	0.01	.17	3.82 (1.82; 0.23 to 7.40)	0.11	.04
User features						
Training received (reference=no)	8.15	0.06	<.001	2.56 (1.69; −0.77 to 5.89)	0.08	.13
Telehealth use before the pandemic (reference=no)	8.24	0.07	<.001	−1.63 (1.80; −5.17 to 1.91)	−0.05	.37
Skills and competencies with digital tools	3.86	0.04	<.001	0.73 (0.99; −1.22 to 2.67)	0.04	.46
Telehealth use						
Frequency of use	6.34	0.16	<.001	3.05 (0.85; 1.36 to 4.73)	0.19	<.001
Duration of use	3.18	0.11	<.001	1.59 (0.55; 0.49 to 2.68)	0.17	.004
Number of NCDs[b] treated	0.93	0.03	.002	0.22 (0.28; −0.32 to 0.77)	0.04	.42
Number of telehealth features used	0.81	0.01	.14	−0.25 (0.46; −1.15 to 0.66)	−0.03	.59
Stability of connection	8.77	0.20	<.001	5.06 (1.01; 3.06 to 7.05)	0.26	<.001
Satisfaction with online support information	8.18	0.24	<.001	5.02 (0.88; 3.27 to 6.75)	0.30	<.001

[a]The 3 columns for *b*, R^2, and *P* value show predictors that entered the maximum model before variable selection (backward method). All variables had a *P* value of less than .20 in univariate analysis. The 2 columns for β and *P* value show a multiple linear regression containing all β and *P* values ($F_{12,276}$=17.45; *P*<.001; adjusted R2=0.407).

[b]NCD: noncommunicable disease.

Discussion

Principal Findings

This exploratory study examined factors associated with telehealth usability among HCPs managing NCDs across 4 distinct health care settings. We aimed to identify how engagement patterns, technical infrastructure elements, and user profiles influenced perceived usability in these global contexts. Connection stability emerged as a primary correlate, suggesting that technical reliability is a foundational element of telehealth adoption. Unstable connections can disrupt communication and clinical workflows, which may undermine the rapport building necessary for NCD management [17]. These findings align with those of prior research identifying technological reliability as a critical determinant of usability [17,38]. Satisfaction with online support information also emerged as a strong correlate. While system training can be valuable, clear digital guidance may help HCPs better navigate telehealth platforms and manage ongoing support requirements. This provides supporting evidence on the value of a robust technological infrastructure in integrating telehealth systems into the care of patients with NCDs [28].

Furthermore, the association between usability and both the frequency and duration of telehealth use suggests an “experience effect,” consistent with established technology acceptance models [39]. Thus, usability may be a longitudinal skill developed through routine integration and increasing experience as users incorporate telehealth into clinical workflows. This experience effect could also explain why physicians reported higher usability as their clinical roles may require more frequent and varied engagement with platform functionalities for clinical decision-making. Overall, differences in clinical responsibilities

XSL•FO
RenderX

and "on-the-job" experience may be central to how various user profiles evaluate telehealth systems.

Telehealth usability, engagement patterns, and technical infrastructure also varied descriptively by country. HCPs in Ghana, the United Kingdom, and Brazil reported "good" usability according to SUS guidelines, whereas those in Honduras reported "moderate" scores [32,34]. Descriptive findings suggest that this may be linked to lower frequency and duration of telehealth use, poorer connection stability, and lower satisfaction with telehealth support in certain settings. These variations may reflect differing levels of digital maturity and resource allocation across HIC, HMIC, and LMIC contexts. Such disparities highlight the relevance of considering local infrastructure when evaluating how HCPs interact with and perceive digital health tools.

Country-Specific Engagement Patterns and System Integration

Beyond the correlates of usability, this study revealed distinct country-specific engagement patterns that illustrate how telehealth is integrated across different resource settings. Notable differences in telehealth engagement across the 4 countries provide context for understanding adoption in diverse health care settings. Most users in Ghana (97/111, 87.4%), Honduras (31/38, 81.6%), and Brazil (57/81, 71.3%) reported using only 1 telehealth system in their clinical roles compared to UK users (20/60, 33.3%), with 66.7% (40/60) of the latter reporting using 2 or more systems. These engagement patterns may reflect differences in telehealth infrastructure and availability of diverse telehealth solutions among HICs, HMICs, and LMICs. For example, they may reflect a more mature digital landscape in the United Kingdom, where clinicians use a "toolkit" of task-specific telehealth systems. Conversely, HCPs in lower-resourced settings may rely on single, more versatile systems for patient care. However, the high prevalence of prepandemic telehealth use (84/111, 75.7%) and training (99/111, 89.2%)—both key elements of the user profile in Ghana—suggests greater familiarity and integration. This indicates that LMICs such as Ghana leveraged telehealth to address health care delivery challenges before the pandemic.

Telehealth engagement was more frequent among Ghanaian and UK users, possibly reflecting more integrated or better-resourced systems. For example, most Ghanaian participants were based in urban, private hospitals, which generally possess superior technological infrastructure compared to public facilities [40]. Furthermore, the higher frequency of asynchronous engagement among Ghanaian and UK users suggests that these systems serve functions beyond real-time consultations [41]. Interestingly, Ghanaian users descriptively reported the lowest frequency of synchronous telehealth use, potentially reflecting infrastructural challenges such as unreliable internet connection that hinder real-time interactions [8]. While Ghanaian participants nominally reported lower connection stability than those in the United Kingdom, Honduras reported the lowest observed connection stability overall despite maintaining synchronous use rates similar to those reported by Brazil and UK users. This suggests that synchronous functionality may not be as deeply embedded in the technical infrastructure of certain LMICs. These findings should be interpreted alongside contextual factors as differences in infrastructure or engagement patterns may influence how HCPs evaluate connection stability. However, as participants were predominantly from urban settings with generally good connectivity, the results cannot determine whether perceived reliability reflects infrastructure constraints, specific engagement patterns, or other contextual factors.

Finally, the results demonstrated differences in telehealth use for NCD management across countries. Notably, UK users managed the highest number of NCDs (mean 5.75, SD 4.10), which was relatively higher than users in Brazil, Ghana, and Honduras. This discrepancy may reflect the United Kingdom's advanced telehealth infrastructure and broader policy support, enabling more effective clinical integration [39]. Conversely, challenges in Brazil and Honduras—limited technological infrastructure and fewer financial resources—may limit the range of NCDs managed through telehealth platforms [42].

Drivers of Telehealth Usability and Clinical Integration

These findings have several implications for the practice and policy of telehealth systems, extending the literature by examining usability across multiple income-level contexts and highlighting descriptive differences in infrastructure that influence user experience in HICs, HMICs, and LMICs. First, these findings align with literature emphasizing reliable internet connectivity for the success of telehealth initiatives [40]. In LMICs such as Ghana and Honduras, efforts to improve internet infrastructure are essential. Investments in robust connectivity may mitigate implementation challenges and enhance usability perceptions [8]. Furthermore, the findings underscore the need for targeted policies to bolster capabilities in lower-resourced settings. For example, improving telehealth systems to include better synchronous functionality may be highly relevant, particularly in African countries where synchronous functionality may not be as embedded. Addressing these disparities could help promote more equitable access to health care services and optimize the management of NCDs worldwide.

Next, these findings highlight the value of comprehensive online user support, including technical assistance and user manuals. Health care services should focus on developing and maintaining strong online support systems to assist HCPs in navigating telehealth platforms effectively, which could include technical instructions and advice on clinical workflow integration. Furthermore, as the frequency and duration of telehealth use were significantly associated with usability, encouraging consistent use through policy and practice may improve user experience. Health care services could consider integrating telehealth into routine clinical practice and providing incentives for regular use to foster familiarity and efficiency.

Tailored strategies addressing the specific challenges of each country appear foundational to successful implementation. For example, in LMICs with lower internet stability, asynchronous telehealth may be more viable than synchronous methods, as observed in Ghana. Therefore, targeted efforts to enhance telehealth usability are pivotal, including improving technological infrastructure and ensuring reliable support services. Addressing these barriers could support the adoption

XSL•FO
RenderX

and effectiveness of telehealth systems across diverse health care landscapes, potentially contributing to improved health outcomes for patients with NCDs [40,43-45]. International collaborations and investments from HICs may play a valuable supporting role in these efforts.

Limitations and Future Research

To our knowledge, this was the first study to explore factors associated with telehealth usability across HIC, HMIC, and LMIC contexts. However, several limitations exist. First, the sample sizes for each country were small and obtained using nonprobability sampling, which limited generalizability. Data were collected primarily from physicians in urban settings. Consequently, we could not account for broader systemic factors such as regional health care penetration (HCP-to-population ratios) or specific connection technologies (eg, fiber vs wireless). Future research should explore telehealth usability in larger, more diverse samples—specifically targeting rural populations and private sector settings where infrastructure constraints may be more pronounced.

Second, this study relied on cross-sectional self-reported data, which may be subject to response bias. Longitudinal and interventional studies following new users over time would provide deeper insights into adoption trends and predictors of usability, which are likely to vary according to country and other telehealth characteristics. Additionally, qualitative methods such as interviews would provide richer insights into HCPs' perceived challenges. This would inform tailored strategies considering contextual differences—such as health care infrastructure and cultural attitudes toward telehealth—to facilitate successful integration for HCPs managing NCDs.

Finally, we used a binary response to establish baseline presence of telehealth training. This does not capture the depth, quality, or specific nature of such training, which can range from basic technical instruction to complex clinical workflow integration. The use of granular measures to distinguish between basic technical instruction and more comprehensive clinical workflow training is required. We also investigated usability through the SUS, which does not independently measure facets such as effectiveness, efficiency, security, or satisfaction [46], nor did we assess usability for specific telehealth functionalities, such as synchronous and asynchronous use (eg, videoconferencing, emailing, and remote monitoring). Assessing these specific tools is crucial in determining which telehealth functionalities work best across different economic settings as requirements likely differ among HICs, HMICs, and LMICs.

Conclusions

Since the onset of the COVID-19 pandemic, telehealth has flourished and emerged as an indispensable resource for improving access to care worldwide, particularly among HCPs treating patients with NCDs. This study demonstrates that telehealth usability is driven by technical infrastructure (connection stability and support information), user profiles (profession), and an "experience effect" derived from specific engagement patterns (frequency and duration of use). Furthermore, this study uncovered notable differences in these patterns across the 4 participating countries. These findings have potential implications for the practice and policy of telehealth systems in HIC, HMIC, and LMIC contexts, highlighting key areas that warrant attention to enhance system usability and ensure effective health care delivery for patients with NCDs worldwide.

Acknowledgments

We used the generative AI tool (Gemini) for occasional language editing, which was further reviewed and revised.

Funding

This research was funded by the Worldwide Universities Network (WUN), the Brazilian National Council for Scientific and Technological Development (CNPq), the Research Support Foundation of the State of Minas Gerais (FAPEMIG), the Brazilian Innovation Center on Artificial Intelligence for Health (CIIA-S), and the Brazilian Institute for Health Assessment and Translation for Chronic and Neglected Diseases of High Relevance (IATS-CARE).

Data Availability

The datasets generated or analyzed during this study are available in the OSF repository [47].

Conflicts of Interest

None declared.

Checklist 1
STROBE checklist.
[DOCX File, 38 KB - jmir_v28i1e64070_app1.docx]

References

1. Noncommunicable diseases. World Health Organization. 2025. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases [accessed 2026-03-17]

XSL•FO
RenderX

2. GBD 2015 Risk Factors Collaborators. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet 2016 Oct 8;388(10053):1659-1724. [doi: 10.1016/S0140-6736(16)31679-8] [Medline: 27733284]
3. Nikoloski Z, Alqunaibet AM, Alfawaz RA, et al. Covid-19 and non-communicable diseases: evidence from a systematic literature review. BMC Public Health 2021 Jun 5;21(1):1068. [doi: 10.1186/s12889-021-11116-w] [Medline: 34090396]
4. Invisible numbers: the true extent of noncommunicable diseases and what to do about them. World Health Organization. 2022. URL: https://www.who.int/publications/i/item/9789240057661 [accessed 2026-03-23]
5. Transforming our world: the 2030 agenda for sustainable development. United Nations. URL: https://sdgs.un.org/2030agenda [accessed 2025-03-23]
6. Kichloo A, Albosta M, Dettloff K, et al. Telemedicine, the current COVID-19 pandemic and the future: a narrative review and perspectives moving forward in the USA. Fam Med Community Health 2020 Aug;8(3):e000530. [doi: 10.1136/fmch-2020-000530] [Medline: 32816942]
7. Omboni S, McManus RJ, Bosworth HB, et al. Evidence and recommendations on the use of telemedicine for the management of arterial hypertension: an international expert position paper. Hypertension 2020 Nov;76(5):1368-1383. [doi: 10.1161/HYPERTENSIONAHA.120.15873] [Medline: 32921195]
8. Smith AC, Thomas E, Snoswell CL, et al. Telehealth for global emergencies: implications for coronavirus disease 2019 (COVID-19). J Telemed Telecare 2020 Jun;26(5):309-313. [doi: 10.1177/1357633X20916567] [Medline: 32196391]
9. Bitar H, Alismail S. The role of eHealth, telehealth, and telemedicine for chronic disease patients during COVID-19 pandemic: a rapid systematic review. Digit Health 2021;7:20552076211009396. [doi: 10.1177/20552076211009396] [Medline: 33959378]
10. Predmore ZS, Roth E, Breslau J, Fischer SH, Uscher-Pines L. Assessment of patient preferences for telehealth in post-COVID-19 pandemic health care. JAMA Netw Open 2021 Dec 1;4(12):e2136405. [doi: 10.1001/jamanetworkopen.2021.36405] [Medline: 34851400]
11. Monaghesh E, Hajizadeh A. The role of telehealth during COVID-19 outbreak: a systematic review based on current evidence. BMC Public Health 2020 Aug 1;20(1):1193. [doi: 10.1186/s12889-020-09301-4] [Medline: 32738884]
12. Wosik J, Fudim M, Cameron B, et al. Telehealth transformation: COVID-19 and the rise of virtual care. J Am Med Inform Assoc 2020 Jun 1;27(6):957-962. [doi: 10.1093/jamia/ocaa067] [Medline: 32311034]
13. Dorsey ER, Topol EJ. State of telehealth. N Engl J Med 2016 Jul 14;375(2):154-161. [doi: 10.1056/NEJMra1601705] [Medline: 27410924]
14. Flodgren G, Rachas A, Farmer AJ, Inzitari M, Shepperd S. Interactive telemedicine: effects on professional practice and health care outcomes. Cochrane Database Syst Rev 2015 Sep 7;2015(9):CD002098. [doi: 10.1002/14651858.CD002098.pub2] [Medline: 26343551]
15. Kruse CS, Krowski N, Rodriguez B, Tran L, Vela J, Brooks M. Telehealth and patient satisfaction: a systematic review and narrative analysis. BMJ Open 2017 Aug 3;7(8):e016242. [doi: 10.1136/bmjopen-2017-016242] [Medline: 28775188]
16. Nadav J, Kaihlanen AM, Kujala S, et al. How to implement digital services in a way that they integrate into routine work: qualitative interview study among health and social care professionals. J Med Internet Res 2021 Dec 1;23(12):e31668. [doi: 10.2196/31668] [Medline: 34855610]
17. Nepal S, Li J, Jang-Jaccard J, Alem L. A framework for telehealth program evaluation. Telemed J E Health 2014 Apr;20(4):393-404. [doi: 10.1089/tmj.2013.0093] [Medline: 24621384]
18. Sauro J. SUPR-Q: a comprehensive measure of the quality of the website user experience. J User Exp 2015;10(2):68-86 [FREE Full text]
19. Zhang T, Mosier J, Subbian V. Identifying barriers to and opportunities for telehealth implementation amidst the COVID-19 pandemic by using a human factors approach: a leap into the future of health care delivery? JMIR Hum Factors 2021 Apr 9;8(2):e24860. [doi: 10.2196/24860] [Medline: 33779566]
20. Alshamari M. Usability factors assessment in health information system. Intell Inf Manag 2016;08(6):170-180. [doi: 10.4236/iim.2016.86012]
21. Ames HM, Glenton C, Lewin S, Tamrat T, Akama E, Leon N. Clients' perceptions and experiences of targeted digital communication accessible via mobile devices for reproductive, maternal, newborn, child, and adolescent health: a qualitative evidence synthesis. Cochrane Database Syst Rev 2019 Oct 14;10(10):CD013447. [doi: 10.1002/14651858.CD013447] [Medline: 31608981]
22. Odendaal WA, Anstey Watkins J, Leon N, et al. Health workers' perceptions and experiences of using mHealth technologies to deliver primary healthcare services: a qualitative evidence synthesis. Cochrane Database Syst Rev 2020 Mar 26;3(3):CD011942. [doi: 10.1002/14651858.CD011942.pub2] [Medline: 32216074]
23. Klaassen B, van Beijnum BJ, Hermens HJ. Usability in telemedicine systems-a literature survey. Int J Med Inform 2016 Sep;93:57-69. [doi: 10.1016/j.ijmedinf.2016.06.004] [Medline: 27435948]
24. Ohannessian R, Duong TA, Odone A. Global telemedicine implementation and integration within health systems to fight the COVID-19 pandemic: a call to action. JMIR Public Health Surveill 2020 Apr 2;6(2):e18810. [doi: 10.2196/18810] [Medline: 32238336]

XSL•FO
RenderX

25. Mahmoud K, Jaramillo C, Barteit S. Telemedicine in low- and middle-income countries during the COVID-19 pandemic: a scoping review. Front Public Health 2022;10:914423. [doi: 10.3389/fpubh.2022.914423] [Medline: 35812479]
26. Peters DH, Garg A, Bloom G, Walker DG, Brieger WR, Rahman MH. Poverty and access to health care in developing countries. Ann N Y Acad Sci 2008;1136:161-171. [doi: 10.1196/annals.1425.011] [Medline: 17954679]
27. Bukhman G, Mocumbi AO, Atun R, et al. The Lancet NCDI Poverty Commission: bridging a gap in universal health coverage for the poorest billion. Lancet 2020 Oct 3;396(10256):991-1044. [doi: 10.1016/S0140-6736(20)31907-3] [Medline: 32941823]
28. Gonçalves RL, Pagano AS, Reis ZS, et al. Usability of telehealth systems for noncommunicable diseases in primary care from the COVID-19 pandemic onward: systematic review. J Med Internet Res 2023 Mar 16;25:e44209. [doi: 10.2196/44209] [Medline: 36787223]
29. von Elm E, Altman DG, Egger M, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007 Oct 20;335(7624):806-808. [doi: 10.1136/bmj.39335.541782.AD] [Medline: 17947786]
30. WHO package of essential noncommunicable (PEN) disease interventions for primary health care. World Health Organization. 2020. URL: https://www.who.int/publications/i/item/who-package-of-essential-noncommunicable-(pen)-disease-interventions-for-primary-health-care [accessed 2026-03-23]
31. Second round of the national pulse survey on continuity of essential health services during the COVID-19 pandemic: January-March 2021. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/WHO-2019-nCoV-EHS-continuity-survey-2021.1 [accessed 2026-03-23]
32. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud 2009;4(3):114-123. [doi: 10.5555/2835587.2835589]
33. Brooke J. SUS: a 'quick and dirty' usability scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B, editors. Usability Evaluation in Industry: CRC Press; 1996:189-194.
34. Martins AI, Rosa AF, Queirós A, Silva A, Rocha NP. European Portuguese validation of the System Usability Scale (SUS). Procedia Comput Sci 2015;67:293-300. [doi: 10.1016/j.procs.2015.09.273]
35. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. Applied Regression Analysis and Multivariable Methods, 3rd edition: Duxbury Press; 1997.
36. Norman G. Likert scales, levels of measurement and the “laws” of statistics. Adv Health Sci Educ Theory Pract 2010 Dec;15(5):625-632. [doi: 10.1007/s10459-010-9222-y] [Medline: 20146096]
37. Gonçalves RL, Pagano AS, Reis Z, et al. Health professional’s experience of telehealth usability in the care of patients with non-communicable diseases (NCDS), from the COVID-19 pandemic onwards: a cross-sectional study protocol. Princ Pract Clin Res 2023;9(1). [doi: 10.21801/ppcrj.2023.91.10]
38. Jordan DN, Jessen CM, Ferucci ED. Views of patients and providers on the use of telemedicine for chronic disease specialty care in the Alaska native population. Telemed J E Health 2021 Jan;27(1):82-89. [doi: 10.1089/tmj.2019.0284] [Medline: 32286156]
39. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003;27(3):425-478. [doi: 10.2307/30036540]
40. Edirippulige S, Armfield NR. Education and training to support the use of clinical telehealth: a review of the literature. J Telemed Telecare 2017 Feb;23(2):273-282. [doi: 10.1177/1357633X16632968] [Medline: 26892005]
41. Dullet NW, Geraghty EM, Kaufman T, et al. Impact of a university-based outpatient telemedicine program on time savings, travel costs, and environmental pollutants. Value Health 2017 Apr;20(4):542-546. [doi: 10.1016/j.jval.2017.01.014] [Medline: 28407995]
42. Silva AB, da Silva RM, Ribeiro GR, et al. Three decades of telemedicine in Brazil: mapping the regulatory framework from 1990 to 2018. PLoS One 2020;15(11):e0242869. [doi: 10.1371/journal.pone.0242869] [Medline: 33237947]
43. Adongo AA, Azumah FD, Nachinaab JO. A comparative study of quality of health care services of public and private hospitals in Ghana. J Public Health (Berl) 2022 Jul;30:1809-1815. [doi: 10.1007/s10389-021-01479-0]
44. Gajarawala SN, Pelkowski JN. Telehealth benefits and barriers. J Nurse Pract 2021 Feb;17(2):218-221. [doi: 10.1016/j.nurpra.2020.09.013] [Medline: 33106751]
45. Padrini-Andrade L, Balda RC, Areco KC, et al. Evaluation of usability of a neonatal health information system according to the user's perception. Rev Paul Pediatr 2019;37(1):90-96. [doi: 10.1590/1984-0462/;2019;37;1;00019] [Medline: 30569950]
46. Vera F, Noël R, Taramasco C. Standards, processes and instruments for assessing usability of health mobile apps: a systematic literature review. Stud Health Technol Inform 2019 Aug 21;264:1797-1798. [doi: 10.3233/SHTI190653] [Medline: 31438349]
47. Brackstone K. Telehealth usability and usage characteristics in managing non-communicable diseases: a multinational cross-sectional study of healthcare professionals in Brazil, Ghana, Honduras, and the United Kingdom. OSF. 2025. URL: https://osf.io/4hry2/overview [accessed 2026-05-20]

Abbreviations

HCP: health care professional

XSL•FO
RenderX

HIC: high-income country
HMIC: high- to middle-income country
LMIC: low- to middle-income country
NCD: noncommunicable disease
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
SUS: System Usability Scale

Edited by A Mavragani; submitted 18.Dec.2025; peer-reviewed by JA Braimah, R Marshall; revised version received 24.Apr.2026; accepted 30.Apr.2026; published 12.Jun.2026.

Please cite as:
Brackstone K, Gonçalves RL, Pagano AS, Reis ZSN, Afagbedzi SK, Zambrano LI, Lopes TCP, Cordeiro SA, Nunes JM, Meira Jr W, Batchelor J, Ribeiro ALP
Telehealth Usability, Engagement Patterns, and Technical Infrastructure in Managing Noncommunicable Diseases Among Health Care Professionals in Brazil, Ghana, Honduras, and the United Kingdom: Multinational Cross-Sectional Study
J Med Internet Res 2026;28:e64070
URL: https://www.jmir.org/2026/1/e64070
doi:10.2196/64070

© Ken Brackstone, Roberta Lins Gonçalves, Adriana Silvina Pagano, Zilma Silveira Nogueira Reis, Seth Kwaku Afagbedzi, Lysien Ivania Zambrano, Tainá Costa Pereira Lopes, Sarah Almeida Cordeiro, Julia Macedo Nunes, Wagner Meira Jr, James Batchelor, Antonio Luiz Pinho Ribeiro. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Telerehabilitation in Community Stroke Services: Mixed Methods Evaluation of Current Practice and Lessons for Sustained Use

Elizabeth Chandler[1], MSc; Charlotte Dorer[2], MPhil; Valerie M Pomeroy[1,3], PhD; Nicola J Hancock[1], PhD

[1]School of Health Sciences, Faculty of Medicine and Health Sciences, University of East Anglia, The Queen's Building, Norwich Research Park, Norwich, England, United Kingdom

[2]NHS (National Health Service) England–East of England, Cambridge, United Kingdom

[3]Department of Clinical Neurosciences, University of Cambridge; NIHR (National Institute for Health and Care Research) HealthTech Research Centre in Brain Injury, Cambridge, United Kingdom

Corresponding Author:
Nicola J Hancock, PhD
School of Health Sciences, Faculty of Medicine and Health Sciences, University of East Anglia, The Queen's Building, Norwich Research Park, Norwich, England, United Kingdom

Abstract

Background: The delivery of specialist stroke rehabilitation is undergoing a significant transformation, with telerehabilitation increasingly integrated into clinical practice and supported by guidelines and policy. There is a need for the pragmatic evaluation of telerehabilitation in service, which includes insights from clinical teams and people with stroke. This evaluation sought to address that need in the context of community stroke services in the East of England.

Objective: Our evaluation addressed two over-arching aims: (1) to investigate and map contemporary models and experiences of telerehabilitation delivery in community stroke services, examining how it is currently used and perceived by both health care providers and service-users; and (2) to identify practical lessons and enabling factors that support the sustained integration of telerehabilitation into routine community stroke services.

Methods: This study is a two-phase exploratory sequential mixed methods evaluation. Phase one involved discussion groups with stakeholders already using telerehabilitation to explore experiences, attitudes, influences, and behaviors associated with its use. Findings from phase one directly informed phase two by the development of a conceptual framework and in shaping the content of an online survey for clinicians and people with lived experience of stroke. Quantitative and qualitative data were subsequently integrated through triangulation during analysis, interpretation, and reporting stages. Data from the discussion groups were analyzed using a recognized framework for reflexive thematic analysis within a contextualist approach. Descriptive statistics were used to summarize quantitative survey responses.

Results: A total of 20 people attended the discussion groups (n=4 groups total). Further, 96 people responded to the survey. Three themes underpinning successful use of telerehabilitation in this context were derived from triangulation across our data sources: (1) consideration of risks and benefits, with fewer than half of staff viewing telerehabilitation as equivalent in safety (22/49, 45%) or effectiveness (18/49, 37%), but most reporting that they could build rapport remotely (34/49, 69%); (2) the importance of individualized care approaches, where most clinicians reported confidence in identifying which service-users would benefit from telerehabilitation (42/49, 85%), although only 20% (10/49) offered it routinely; and (3) the need for staff support, with up to 34% of staff reporting no training in how to assess suitability for telerehabilitation. Key insights included the potential for telerehabilitation to increase efficiency and address service pressures, the importance of addressing digital exclusion, the value of individualized approaches, and the need for timely and tailored staff training.

Conclusions: Our pragmatic, in-service evaluation demonstrates that telerehabilitation works best not as a replacement for in-person care, but as part of a responsive, blended approach grounded in individual need. These findings highlight that, with appropriate clinician training and flexibility in delivery, telerehabilitation can meet the needs of individuals through personalized approaches while supporting service responsiveness in pressurized clinical environments.

(J Med Internet Res 2026;28:e87741) doi:10.2196/87741

KEYWORDS

stroke; telerehabilitation; stroke rehabilitation; community services; service innovation

XSL•FO
RenderX

Introduction

Telerehabilitation is defined as the delivery of rehabilitation consultations, assessments, and therapies remotely using information and communication technologies [1]. It can involve real-time remote consultation, as well as the sharing of information such as therapy programs and monitoring performance at other times. The possibilities for telerehabilitation to support provision of stroke rehabilitation were explored in a 2020 systematic review and meta-analysis of 22 studies published between 2004 and 2019: comparisons of a small number of studies found that telerehabilitation might be as effective as in-person care at least for activities of daily living (n=2 studies, moderate-quality evidence) and upper limb function (n=3 studies, low-quality evidence) [2]. Telerehabilitation has been proposed as a possible solution to some identified challenges of community rehabilitation, such as remote geographical locations [3].

The use of telerehabilitation was rapidly accelerated by the COVID-19 pandemic, with clinical stroke teams adapting to digital solutions for rehabilitation and care delivery, to continue the provision of some rehabilitation when face-to-face contact was minimal. Following this rapid adoption, telerehabilitation is now an increasing part of specialist stroke rehabilitation delivery and is supported as such in guidelines and policy [4,5]. Two of the UK government's 10-year health plan "big shifts"—policy drivers initiated after the COVID-19 pandemic—speak directly to increased use of delivery beyond hospital settings: "analogue to digital" and "hospital to community care." Fundamental to those shifts is digital transformation to enhance patient care through the use of technology [6]. As a result, the landscape of stroke rehabilitation is changing.

Despite such policy-driven support for telerehabilitation, it is recommended that further pragmatic evaluation of its use, including understanding patient experiences and those of clinical teams, is carried out to support its implementation as a mode of delivery for stroke rehabilitation [7]. Further evaluation, monitoring, and engagement with service users is considered critical to digital service improvement [5]. Following a period of enforced change, there is a need now to strengthen the evidence and understanding in parallel to the service delivery change [7,8]. While safety and feasibility of specific teledelivered stroke programs have been explored, for example, for intense upper limb rehabilitation [9] and for speech and language therapy [10], understanding of how to operationalize telerehabilitation in community stroke rehabilitation and sustain its use is limited [11]. A deeper exploration of the foundational components of successful use of telerehabilitation is required.

To meet the challenge of evaluation alongside existing service delivery, the service evaluation reported here was commissioned by NHS (National Health Service) East of England as part of the UK-wide Stroke Quality Improvement for Rehabilitation initiatives to support the delivery of the UK Integrated Community Stroke Service [5]. The East of England provided a particularly rich setting for the work, due to community health provision across a mix of rural, suburban, and city-based populations.

The overarching aims of the mixed methods evaluation presented here were to (1) investigate and map contemporary models and experiences of telerehabilitation delivery in community stroke services, examining how it is currently used and perceived by both health care providers and service-users; and (2) identify practical lessons and enabling factors that support the sustained integration of telerehabilitation into routine community stroke services.

Specific evaluation questions were the following: (1) In what ways is telerehabilitation currently delivered in community stroke services, and how do experiences vary between providers and service-users? (2) What are the key components of successful use of telerehabilitation in community stroke rehabilitation practice?

Methods

Study Design

An exploratory sequential mixed methods evaluation was conducted between April 2023 and March 2024. A mixed methods evaluation was used due to the known complexity of pragmatic health services evaluation, enabling deep insights unavailable to the evaluation team from either quantitative or qualitative data alone [12]. The current study is reported with attention to the GRAMMS (Good Reporting of a Mixed Methods Study) reporting guidelines for mixed methods health service research [13] (Multimedia Appendix 1).

To address study aims, the evaluation comprised two phases: regarding phase one, discussion groups with clinicians working in community stroke services, the voluntary sector, and members of an established stroke-specific patient and public involvement (PPI) group; and regarding phase two, conceptual framework development and an online survey for clinicians and people with lived experience of stroke.

Using the sequential exploratory design, findings from phase one directly informed and were connected to phase two by shaping the conceptual framework and survey content. Quantitative and qualitative data were subsequently integrated through triangulation during analysis, interpretation, and reporting stages.

Inclusion and Exclusion Criteria

Eligibility criteria were necessarily broad to reflect the range of perspectives across community stroke services. Clinicians working in stroke services in the East of England and PPI members with experience of stroke rehabilitation in the region were eligible to take part. Participation in phase one did not preclude participation in phase two. All were invited to take part in both phases of the evaluation, with the option to take part in phase two regardless of participation in phase one.

For phase one, purposeful sampling was used to recruit clinicians with experience of delivering telerehabilitation, to inform the development of the conceptual framework and survey.

XSL•FO
RenderX

Responses from participants working only in acute services were not included in the analysis.

Recruitment

Recruitment for phase one was conducted via clinical and stroke-specific PPI meetings. At these meetings, the evaluation team presented this study and its aims and invited attendees to participate. Individuals interested in taking part were asked to contact the evaluation team by email to request a participant information sheet with full details of the evaluation and consent procedure (phase one). All participants had the opportunity to ask questions, and written informed consent was obtained.

To maximize accessibility for PPI members, this study was presented at two meetings held at different times of the day.

Fifteen community stroke teams were invited to the clinical meeting; services represented were located within two eastern integrated stroke delivery networks in England.

Phase two differed from phase one in that participation involved the completion of an anonymous survey online rather than enrollment in discussion groups. The survey was accessible via an online link or a QR code. Participation was voluntary, and an introductory page outlined the purpose of this study, data handling practices, and participants' right to withdraw at any time before submission. Consent was implied through survey completion and submission. Participation was advertised at clinical and PPI meetings, as in phase one, but at a later date. To extend the reach across the region, the survey was also promoted at the East of England Stroke Forum and shared via two stroke charities.

Researcher Context and Reflexivity

Data from the discussion groups were analyzed using reflexive thematic analysis within a contextualist approach, enabling flexibility of combining multiple sources of data such as people's experiences and practices [14,15]. This approach highlights the researcher's active role in knowledge production. Codes are understood to represent the researcher's interpretations of patterns of meaning across the dataset. Understanding evaluators' backgrounds in relation to the phenomenon at hand is an important aspect of conducting reflexive qualitative evaluation. Here, the lead author (EC) is a chartered physiotherapist and has worked in NHS community and stroke rehabilitation services. EC conducted all the discussion groups. The principal evaluator (NJH) is a chartered physiotherapist, researcher, and innovation lead working in Higher Education, with extensive experience in NHS stroke rehabilitation service delivery, evaluation, and research. These professional perspectives informed the interpretation of the data, and ongoing reflexive discussion was used to consider how our assumptions and experiences influenced the theme development. NJH was present for 2 of 4 discussion groups. Participants were aware of the evaluators' background when taking part in the discussion groups.

Data Collection

Phase One

Discussion groups were used to explore the use and experiences of telerehabilitation and to identify key attitudes, influences, and behaviors associated with its use. A topic guide was developed a priori (Multimedia Appendix 2: Discussion groups topic guide).

At the start of each discussion group, participants were informed that the conversation would be recorded, reminded of their right to withdraw, and assured that their identities would be pseudo-anonymized.

Four groups were held, facilitated by EC. A choice of online or in-person discussion was offered, resulting in three via videoconferencing (Microsoft Teams [Microsoft Corp]) and one in-person. PPI members and clinical staff attended together to make use of group dynamics, encourage rich discussion, and explore complex issues from both viewpoints. The facilitator used inclusive techniques to support all voices being heard (eg, inviting responses from all participants and managing dominant speakers). Groups were recorded and then transcribed verbatim by an independent third party. Transcriptions were reviewed by EC for accuracy.

Discussion groups were conducted using a pragmatic, time-limited approach consistent with an in-service evaluation. Although formal data saturation was not predefined, few new points of discussion arose in the final group, suggesting adequate coverage of the key perspectives.

Phase Two

Findings from phase one were combined with insights from published literature and clinical guidelines to inform the development of a conceptual framework (Figure 1) and guide the design and content of an online survey. This framework ensured that all relevant topics and variables of interest were addressed, while also facilitating data triangulation by integrating multiple sources and perspectives. The survey aimed to evaluate the current use of telerehabilitation in community stroke services and to explore the beliefs and attitudes influencing its sustained use.

XSL•FO
RenderX

Figure 1. Conceptual diagram illustrating key themes underpinning successful use of telerehabilitation in community stroke services, developed from qualitative interviews with survivors of stroke and rehabilitation clinicians in a mixed methods study conducted in community neurorehabilitation services in the east of England, UK, 2023-2024. MDT comms: multidisciplinary team communication.

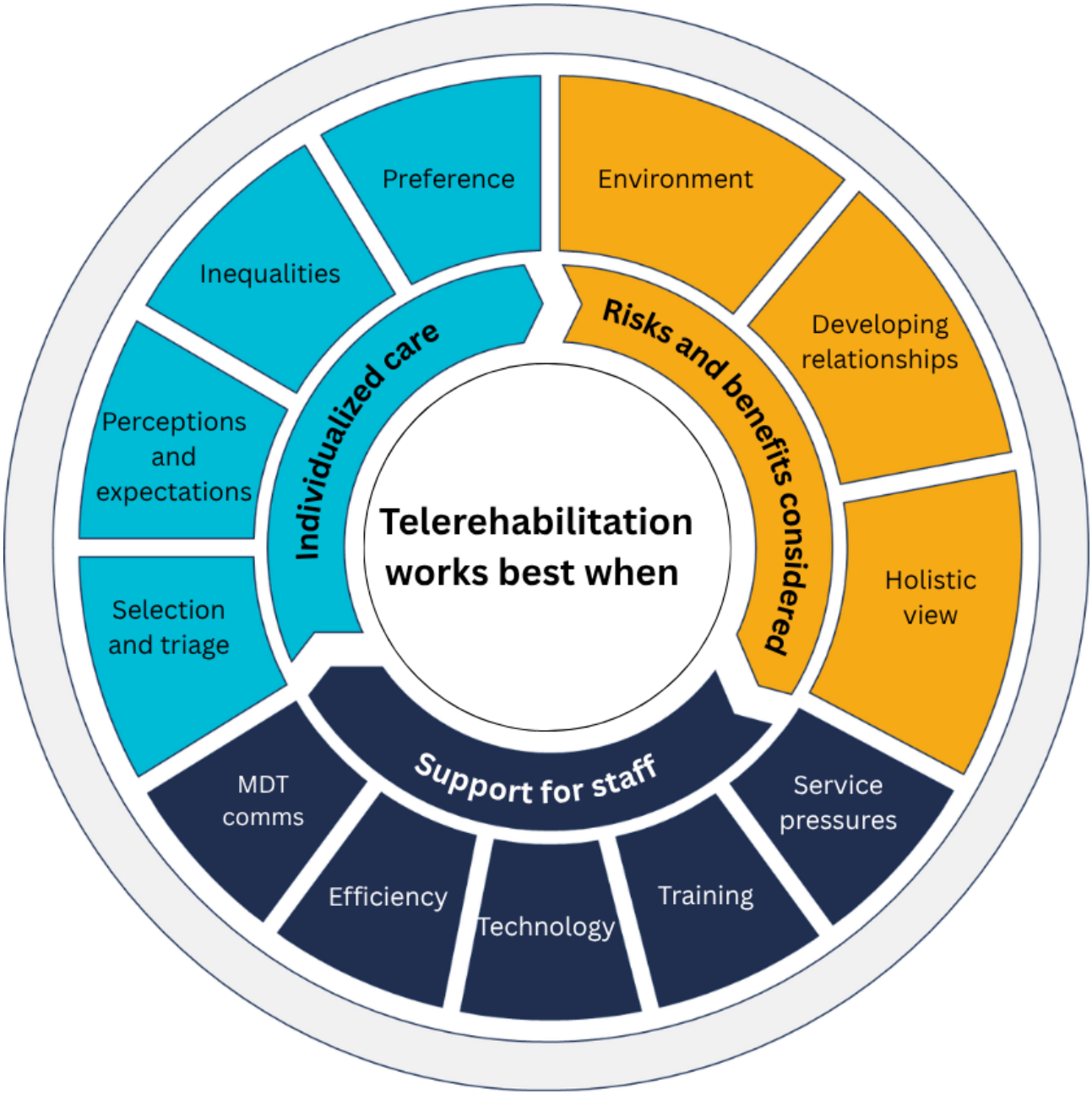

The survey was created using Microsoft Forms (Microsoft Corp) and piloted to test ease of use. It was designed to be mobile device-friendly and take no more than 20 minutes to complete. Filtering logic was used to customize questions based on each respondent's role (clinicians providing stroke services, service users, and unpaid caregivers) and level of experience with telerehabilitation. Responses were anonymous, and no identifying information or IP addresses were collected.

The survey adopted a recommended range of data-gathering styles [16], including open- and closed-ended questions, ranking questions, and 4-point Likert scales with the options: strongly disagree, disagree, agree, and strongly agree.

It was accessible for eight weeks.

Data Analysis

The reflexive thematic analysis was conducted by working iteratively through a six-phase process of data familiarization, generation of initial codes, generating themes, and reviewing, refining, and naming the themes [15]. According to Braun and Clarke [15], coding was conducted inductively, with codes generated from the data rather than a predefined framework.

Qualitative data were analyzed using an inductive coding approach, with codes generated directly from the data. Following coding, relevant published literature and clinical guidelines (including those from related areas of physiotherapy practice) were reviewed to check for potential gaps. This process did not lead to the identification of additional codes.

XSL•FO
RenderX

EC was responsible for all coding and generating themes. EC and NJH then worked collaboratively to reexamine and redefine themes, achieving richer interpretations of the meaning. Differences in interpretation were explored through discussion to deepen understanding and refine themes, consistent with a reflexive approach.

Quote identifiers in the results section include the discussion group number and the participant's role (eg, DG1, OT for an occupational therapist in discussion group 1). Roles include: OT (occupational therapist), physiotherapist (physiotherapist), SLT (speech and language therapist), RA (rehabilitation assistant), SU (service-user with lived experience of stroke), and UC (unpaid caregiver).

Survey data were exported from Microsoft Forms into Microsoft Excel (Microsoft Corp) for analysis. Responses were screened for eligibility, and ineligible submissions (where rehabilitation or service experience did not take place in the East of England) were excluded. Missing data were reviewed; if a respondent left a nonessential question unanswered, their other responses were retained. No imputation methods were used to estimate missing values.

Where contradictory responses were identified, these were reviewed for internal consistency. In one case, a participant reported discontinuing telerehabilitation entirely but also provided detailed examples of its use. Their response was amended to reflect internally consistent use, in line with the examples given.

Descriptive statistics, including frequencies and percentages, were used to summarize demographic characteristics and closed-ended responses. Likert-scale items were summarized using the proportion of respondents selecting each response option.

For questions that allowed respondents to select multiple answers, percentages of respondents who selected each option as part of their responses were calculated.

The ranking data were analyzed by calculating weighted scores for each option based on frequency and rank position. For each response, items ranked in first position were assigned a score of 10, second position a score of 9, and so on down to a score of 1 for the lowest rank. Scores were then summed across responses to produce an overall score for each factor, with higher scores indicating greater perceived importance.

Data Integration

Integration took place at both the design (building) and analysis (merging) stages. Findings from the phase one discussion groups were used to develop a conceptual framework, alongside relevant published literature, and this informed the design of the phase two questionnaire.

After data collection, findings from both phases were considered together. This allowed comparison of perspectives across the two datasets and helped identify areas of agreement and difference.

Ethical Considerations

Ethical approval for this evaluation was granted by the University of East Anglia Faculty of Medicine and Health Sciences Research Ethics Service (ETH2223-1521).

All participants in the phase one discussion groups received a participant information sheet, had the opportunity to ask questions, and provided full informed written consent.

For phase two, the online survey was anonymous. An introductory section outlined the purpose of the evaluation, ethical approval, and data handling. Completion and submission of the survey were taken as implied consent.

Discussion group participants who traveled to the venue were entitled to reimbursement of their travel costs, but participants were not offered financial incentives to take part.

All data were securely stored on encrypted servers in compliance with regulatory requirements.

Results

Participants

Phase One: Discussion Groups

A total of 20 people attended across all discussion groups. They were occupational therapists (n=5), physiotherapists (n=3), speech and language therapists (n=4), rehabilitation assistants (n=4), members of the patient and public volunteer assurance group (n=2; both with lived experience of stroke but not telerehabilitation), and charity representatives from the Stroke Association and Different Strokes (n=2, one a survivor of stroke, also without telerehabilitation experience).

Phase Two: Online Survey

A total of 96 people responded to the survey. Four responses were excluded as the participants had not received stroke rehabilitation in the East of England, leaving 92 valid responses from multidisciplinary team (MDT) members (n=81), individuals who had experienced a stroke (n=10), and 1 unpaid caregiver.

Among the MDT respondents, 14 (17%) worked in acute stroke services, 38 (47%) in early supported discharge teams, and 29 (36%) in community stroke or community neurological services, including stroke rehabilitation. To meet the evaluation aims, only responses from clinical staff working in early supported discharge and community services are reported here (n=67).

Of the survivors of stroke, only one had received telerehabilitation services, while the unpaid caregiver had also experienced telerehabilitation.

To establish MDT participants' experience with telerehabilitation as a baseline for understanding their perspectives and practices, a foundational question, "Have you ever used telerehabilitation?" was asked. A total of 49 (73.1%) of participants had used telerehabilitation (Tables 1 and 2).

Unless otherwise stated, the findings relating to clinical staff reported are based on respondents who reported prior use of telerehabilitation (n=49), as survey filtering directed subsequent

XSL•FO
RenderX

questions to this group. Responses to training-related questions, however, include all clinical participants regardless of telerehabilitation experience (n=67). Service user views are presented alongside clinician findings where available, but are based on small numbers (survivors of stroke n=10; caregiver n=1). Percentages are presented with underlying counts (n/N) to aid interpretation.

Table . Telerehabilitation use by profession among clinicians involved in community stroke rehabilitation services in the East of England, 2023-2024 (total clinician analytic sample, N=67).

Profession	Used telerehabilitation (n=49)	Never used telerehabilitation (n=18)	Total respondents (N=67)
Nonregistered staff	12	4	16
Occupational therapists	11	5	16
Speech and language therapists	11	0	11
Nurses	3	1	4
Psychology	6	1	7
Physiotherapists	4	5	9
Team leads	2	0	2
Dietitian	0	1	1
Therapist (unspecified)	0	1	1
Total	49	18	67

Table . Survey respondent groups (clinicians, survivors of stroke, and caregivers; total respondents N=92).

Respondent group	Values, n
MDT[a] staff from community stroke services	29
MDT staff from early supported discharge	38
MDT staff from acute stroke services	14
Survivors of stroke	10
Unpaid caregivers	1
Total valid respondents	92

[a]MDT: multidisciplinary team.

Evaluation Findings

Overview

A conceptual model was developed based on thematic analysis of phase one discussion group findings (Figure 1) and subsequently informed the design of the phase two survey. This analysis led to three derived themes. To enhance coherence, evaluation findings are here presented within these themes, integrating evidence from phase one (qualitative) and phase two (quantitative) data. Within each theme and subtheme, discussion group findings are presented first, followed by survey results, before being integrated through narrative and considered together.

It is important to note that some overlap exists between themes, reflecting the interconnected and nuanced nature of the factors influencing the use of telerehabilitation.

Theme One: Successful Use of Telerehabilitation—Risks and Benefits

Overview

This theme explores the importance of understanding both the potential and the limitations of telerehabilitation in community stroke services. Clinician reflections suggest that when delivered in isolation, telerehabilitation could increase risk, particularly in relation to safety, safeguarding, and clinical decision-making. However, when integrated within a broader service model, such as blended delivery following face-to-face contact, it is seen as a valuable adjunct to care.

As one participant described:

> *On its own, well, I think it can expose risk, but if it's used again as a blended approach after you've met face to face and you've seen the environment and you've established what support and what relationships there are around them, then you can move forward kind of knowing that you're not missing anything.* [DG2, SLT 1]

This balance between risk and benefit was echoed across the focus groups and shaped three interconnected subthemes: the environment, relationships, and working holistically.

The Environment

Concerns about missing key details in the home environment were consistently raised. Staff spoke of the limitations of video

XSL•FO
RenderX

consultations in capturing the full context of a person's living situation details as compared to face-to-face observations.

An occupational therapist reflected:

> *The little things, like just the space around the toilet, you know, and there's just things you don't always pick up on that you can risk assess in the home environment sometimes.* [DG3, OT1]

Survey responses (phase two) reflected divided views. Among staff who had used telerehabilitation, 53% (26/49) agreed that initial assessments should always be delivered face-to-face, while a similar proportion disagreed (23/49, 47%). Taken together with discussion group accounts, this suggests ongoing uncertainty about when telerehabilitation is appropriate and when in-person assessment is essential for safety and risk management.

Relationships

Participants reflected on how telerehabilitation alters the process of building therapeutic relationships. Many described a more limited sense of the individual and their social environment when interacting remotely.

As one clinician explained:

> *The image, the impression that you get through the video or through the phone is very one-dimensional. So you only get the presentation of that person within that square and you don't get those environmental clues. The relationships between the family or carers. So really those things that we pick up on the minute that we're walking up to somebody's house, we don't get. The difference between face to face and sort of virtual options is that you don't pick up on those other clues, those other things around the person.* [DG2, SLT1]

Survey responses (phase two) reflected this complexity. A strong majority (34/49, 69%) agreed they could build rapport using telerehabilitation, although around 1 in 3 (15/49, 31%) still expressed doubt, suggesting that while feasible, clinical staff still experienced challenges in rapport-building remotely.

Working Holistically

For many staff, the ability to work holistically, seeing beyond the immediate clinical task, was perceived to be limited by remote delivery. In-person visits often revealed issues not directly related to the stroke-specific goal, enabling a more comprehensive and person-centered response.

One participant explained:

> *I work with certain colleagues who … would really prefer not to do virtual because they don't feel that they have the capacity to do that holistic approach in terms of reaching out beyond the actual stroke specific goal and beyond their professional boundaries in terms of you know what their quality of life is looking like and what kind of what's the bigger picture? Basically, both in the immediate home environment, but also kind of their day-to-day?* [DG1, PT1]

Survey findings (phase two) aligned with these concerns. Although 73% (36/49) of staff felt telerehabilitation helped them support the long-term needs of service-users, fewer felt it was as effective (18/49, 37%) or as safe (22/49, 45%) as face-to-face rehabilitation. This aligns with service-user perspectives (phase two), fewer than half of the survivors of stroke and unpaid caregivers surveyed (4/11, 36%) agreed that telerehabilitation was of equal quality to face-to-face therapy, while the majority (7/11, 63%) disagreed or strongly disagreed. Although the survey did not explore their reasoning in depth, these responses suggest that some individuals may have experienced limitations or perceived telerehabilitation as less personal or effective. Coupled with findings from clinicians that recognized the benefits of telerehabilitation, but also its limitations when striving for comprehensive care, these connected findings emphasize the importance of aligning telerehabilitation with individual preferences and ensuring it is not positioned as a universal substitute for in-person care. This is explored further in theme two.

In summary, theme one highlights the complex interplay of risks and benefits when using telerehabilitation in community stroke services. While many staff valued its flexibility and potential to reduce service burden, they also emphasized the need for careful implementation, particularly in relation to understanding the home environment, building relationships, and delivering holistic care. Juxtaposing the perspectives of survivors of stroke and caregivers further underscored that perceived quality and therapeutic connection may be diminished when services are delivered remotely, even when practical needs are met.

Survey items related to this theme are summarized in Multimedia Appendix 3: Survey items theme 1.

Theme 2: Successful Use of Telerehabilitation—Individualized Care

Overview

This theme explores how telerehabilitation was shaped by individual needs, preferences, and circumstances. Staff and service users alike described a complex mix of practical, emotional, and contextual factors that influenced whether it was perceived as acceptable or beneficial. The legacy of the COVID-19 pandemic remained a notable influence on service delivery and personal choice. One clinician reflected:

> *They're still scared of Covid, so they still want their therapy, but they are like, "I don't want people in my home." Still scared and if they've got like someone in their home that's unwell as well I think video works well. At least having a nice balance between the two so you are going in once a week, face-to-face but then having like the video calls I think sometimes just eases that anxiety.* [DG4, RA2]

Participants emphasized that there was no single "type" of service user for whom telerehabilitation would always be appropriate or inappropriate. Instead, they described the importance of tailoring decisions individually. Interrogation of this theme identified four interrelated subthemes: selection and

XSL•FO
RenderX

triage, perceptions and expectations, inequalities, and preferences.

Selection and Triage

Across all staff discussion groups, participants described how they triaged service users to determine suitability for telerehabilitation. Clinical factors such as cognitive impairment, visual or hearing loss, and severe communication difficulties were frequently cited as "red flags" for which telerehabilitation was likely to be inappropriate.

> *It's mainly, receptive or cognitive. Or they got visual or hearing [loss]. Those kind of red flags.* [DG3, SLT 1]

Practical issues also influenced decision-making. For example, where service users lived at a considerable distance from the team, video appointments were sometimes used to supplement face-to-face visits. This approach was seen to increase capacity, particularly for disciplines such as speech and language therapy.

> *If someone lives an hour's drive, you've got to work that in your day, where on the phone you haven't got to.* [DG4 RA5]

Telerehabilitation was also used for brief, focused reviews, particularly where the service user's needs were straightforward.

> *Some people prefer to just have a quick review… They can just sort of sign in to the virtual waiting room to discuss their programme.* [DG2, PT1]

In the survey findings (phase two), over one-third (18/49, 37%) of staff said they offered telerehabilitation selectively, based on whether they believed the person might like it. Only 20% (10/49) said they routinely offered it to all service users. When asked to rank suitability criteria, staff prioritized access to IT equipment, service user preference, and ability to engage in an online conversation above factors such as age, geography, or presence of a caregiver.

Perceptions and Expectations

Clinicians described a shift in how telerehabilitation was perceived, both by staff and service users, since the early days of the pandemic. While some still expressed concern that service users might see telerehabilitation as inferior, others noted positive changes in how expectations were managed and how individuals had adapted.

> *I think a lot of our patients do have certain expectations of what they think therapy is gonna be… I still think they prefer face to face… Maybe they do feel they're not getting a good deal.* [DG3, SLT1]

However, others gave examples of how telerehabilitation could support self-management and build confidence:

> *They are writing down problems before the phone call because they get a little letter before… I think their self-management skills have got better. I think it has helped with their expectations and they're actually quite pleased that they are self-managing a bit better...* [DG3, OT1]

Survey responses (phase two) indicated that 85% (42/49) of staff felt confident identifying who would benefit most from telerehabilitation, only 41% (20/49) believed that service users enjoyed it, and just 14% (7/49) thought it was considered equal in quality to face-to-face therapy. Among survivors of stroke and caregivers, 81% (9/11) said they would be comfortable using telerehabilitation, 100% (11/11) said they wanted to be involved in the decision to use it, and a slightly lower proportion (8/11, 73%) felt confident using technology, suggesting that willingness to engage with telerehabilitation may extend beyond self-confidence in the use of digital technology. Juxtaposition of the views of the clinicians and service users illustrates the importance of dialogue and shared understanding.

Inequalities

Concerns about inequality, particularly digital exclusion, were raised in every discussion group. Some participants reflected on assumptions about age and digital literacy, noting that these did not always hold true in practice:

> *We have a cohort of our patients that are older. So there's that perception that they wouldn't be able to manage with Telerehab. And I think actually we were quite surprised how well that worked.* [DG3, PT1]

> *You assume a 35- or 40-year-old is going to be all right with the tech… but some people don't [have Wi-Fi] and then you've got your 98-year-old and they've got their iMac.* [DG4, RA5]

Others were more cautious, noting that older adults may prefer face-to-face contact to reduce isolation or maintain connection:

> *It's the contact, the physical contact. Well, especially for the older people… and that they're not alone. As such, when it comes to the rehab support.* [DG3, OT1]

Access to equipment was another recurring theme. Staff described how many service users lacked suitable devices or only had limited phone access.

> *So many didn't have any devices… a mobile phone, but it wasn't a smartphone, and they'd only use it very occasionally.* [DG1, SLT1]

These inequalities were also evident in the survey (phase two). Among the 10 factors ranked for telerehabilitation suitability, access to IT equipment was rated highest in importance, followed by service user preference, whereas age, geographical location, and presence of a caregiver were ranked lowest (Figure 2). Factors relating to safety, purpose, and communication occupied the midrange.

XSL•FO
RenderX

Figure 2. Diverging stacking bar-chart of clinical staffs' ranking of 10 factors influencing telerehabilitation suitability decisions for survivors of stroke in community rehabilitation services, derived from quantitative survey data in the mixed methods survey of UK rehabilitation professionals, 2023-2024. Findings are from respondents reporting prior use of telerehabilitation (n=49). Factors are ordered by overall ranking score calculated from the frequency of assignment to each rank position across the 10-item ranking task, with higher ranks contributing greater weight.

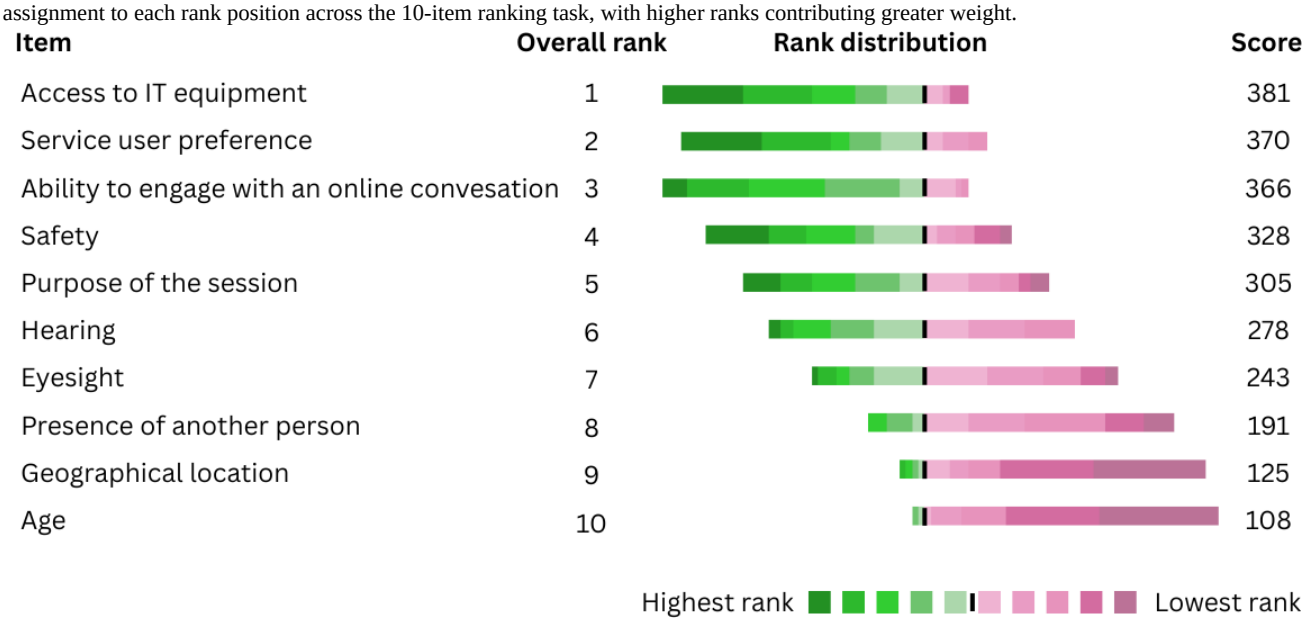

Preferences

Personal preference emerged as a key determinant of successful telerehabilitation. Staff spoke of making assumptions about who would or would not want remote appointments, often based on age or perceived technical ability. However, many recognized the need to move beyond assumptions and to ask service users directly.

> *We probably sort of use a bit of bias and maybe in the patients that we think you know would prefer to do video consultations and maybe ask them.* [DG3, SLT2]

Several clinicians described how shared decision-making could help match delivery method to the individual's needs and context, often using a blended approach.

> *It's about the individual and about choice… being able to have those shared decision-making conversations with someone so that you both feel that you have the choice.* [DG2, PT1]

Despite the recognition of preference, in the survey (phase two), only 1 in 5 clinicians (10/49) said they routinely offered telerehabilitation to all service users, with most choosing to offer it selectively. This contrasts with the voice of service users (phase two), who clearly valued being involved: 100% (11/11) of those surveyed said they either wanted or had wanted to be fully involved in the decision to use telerehabilitation.

This theme highlights the importance of flexibility and personalization in the use of telerehabilitation. While clinical suitability remains critical, particularly for those with cognitive, sensory, or communication challenges, decisions are also shaped by geography, workload, and perceived confidence with technology. These findings suggest that participants did not consider telerehabilitation to be universally applicable or dismissible, and that a thoughtful approach to shared decision-making is required to meet people's needs.

Survey items related to this theme are summarized in Multimedia Appendix 4: Survey items theme 2.

Theme 3: Successful Use of Telerehabilitation—Support for Staff

Overview

This theme explores how staff can be better supported to use telerehabilitation effectively within community stroke services. While telerehabilitation was recognized as a flexible and efficient tool, its success depended on service structures that enabled collaboration, appropriate technology, and sufficient training. For some, telerehabilitation offered a welcome opportunity to work more efficiently and responsively. However, others reflected on the practical and systemic challenges they faced. As one participant put it:

> *So it is an incredibly efficient way of working, but I do appreciate and we do have to do it out of necessity because we're so pressed. But actually, I think there's some value in working that way.* [DG2, SLT 1]

Detailed analysis of this theme identified five subthemes: MDT communication, efficiency, technology, training, and service pressures.

MDT Communication

Participants described how telerehabilitation was used not only to engage with service users but also to collaborate with colleagues. Being able to involve other professionals in real time enhanced coordination and decision-making:

> *The number of times that people write in their notes, liaise with another colleague, whereas you can do that in the here and now… We can get that piece of equipment ordered. It's not then waiting for another visit the following week. And another lot of travel… it's done.* [DG2, SLT1]

XSL•FO
RenderX

This real-time input also reduced the need for repeated assessments:

> *The therapist who has more speciality within that area can ask relevant questions more easily and immediately… So they're not repeating assessments and repeating information.* [DG2, PT1]

Efficiency

Improved communication contributed to greater efficiency, particularly when one staff member could be present with the service user, and another could join remotely:

> *I found it quite helpful or particularly helpful if a rehab assistant was already seeing somebody and I may be needed to review a situation or review where they were with their exercise programmes that I could join, I could be in the office, but I could join that session remotely.* [DG2, PT1]

Telerehabilitation was also seen to reduce wasted time in care home settings:

> *If they're not ready at the time of video, you can just say okay, well, call me back in 10 minutes… Whereas if you're at the care home, you're kind of stuck there.* [DG3, SLT 1]

Survey results (phase two) supported these findings. Over three-quarters (37/49, 76%) of staff agreed that telerehabilitation allowed them to be more responsive in assessment. Further, 73% (36/49) of staff agreed that it enabled increased therapy intensity. Service users also valued this flexibility: 91% (10/11) of staff said they would accept telerehabilitation if it meant being seen more often or more quickly.

Technology

Technology was described as both an enabler and a barrier. While platforms such as WhatsApp (Meta) were familiar and easy to use, some hospital-based systems were harder to navigate:

> *We've had to get really good at troubleshooting when patients… navigated the wrong way.* [DG3, PT1]

Device type also mattered. Some participants noted that smartphones worked best for certain tasks, while laptops were preferred for others. Familiarity, accessibility, and ease of use were all seen as essential.

Training

Participants identified a gap in meaningful training both for themselves and for service users. While basic instructions (such as finding a quiet room) were available, more tailored and in-depth training was lacking:

> *The obvious bits… were well established, but everything else… not so much.* [DG2, SLT1]

Some staff felt that providing training to service users would improve uptake and quality. However, one person with lived experience shared a different perspective:

> *From a patient's point of view, if I've been told at the time that I had to find out quite space, I had to wear certain clothes, I had to do all the things that you've just said that …would have completely put me off full stop. I wouldn't have done it.* [DG2, SU1]

In contrast to analyses restricted to clinical staff with prior telerehabilitation experience, training-related survey items included all clinical staff in the analytical sample (n=67). Survey results (phase two) revealed varied satisfaction with current training. While 47% (35/49) of staff felt confident in their training on using relevant software, fewer were satisfied with training on adapting service-user goals (22/49, 33%), assessing suitability (22/49, 32%), or communicating effectively via telerehabilitation (24/49, 36%). Notably, staff who had never used telerehabilitation were significantly more likely to report receiving no training at all Figure 3.

Figure 3. Horizontal stacked bar chart of clinical staff's responses to four telerehabilitation training statements (software, adapting objectives, assessing suitability, and effective communication), derived from survey data from all staff delivering community stroke rehabilitation in the East of England who responded to this survey, 2023-2024 (n=67). Response categories were "strongly agree," "agree," "disagree," "strongly disagree," and "I have not received training."

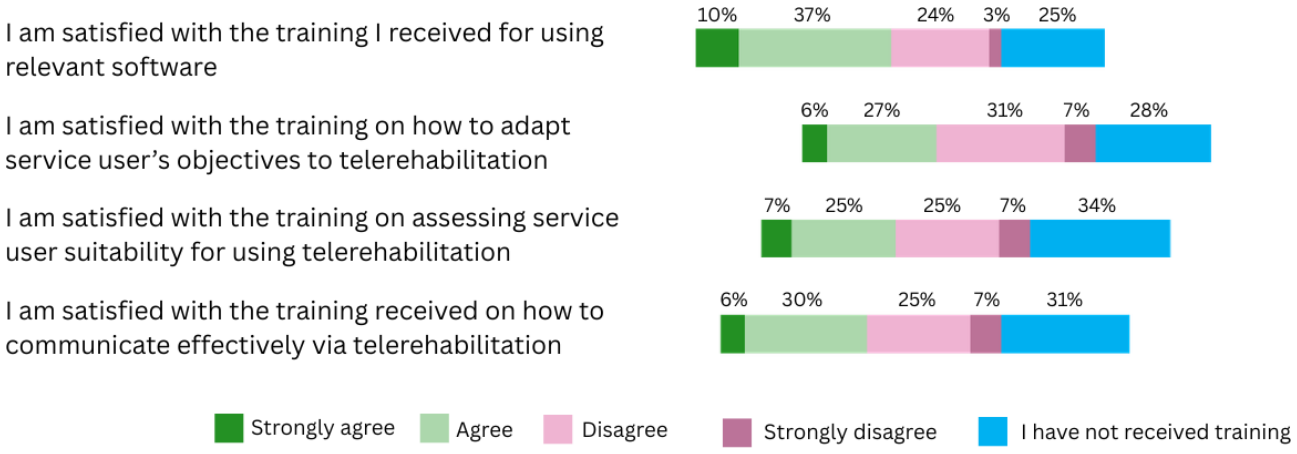

XSL•FO
RenderX

Service Pressures

Participants reflected on how service pressures influenced the use of telerehabilitation. Some acknowledged that it was introduced or used more heavily due to workforce limitations, but still recognized its value:

> *They probably get more telerehab than they would face to face. They also might get seen quicker.* [DG3, SLT1]

In this context, telerehabilitation was not simply a compromise but a means of extending reach and offering timely support. Survey findings (phase two) supported this: 75% (37/49) of staff agreed that telerehabilitation helped them to be more responsive in assessment, and 80% (39/49) agreed it was effective for 6-month reviews.

This theme illustrates that successful telerehabilitation is closely tied to how well staff are supported to use it. Staff welcomed the increased efficiency and collaborative opportunities that telerehabilitation enabled, but identified unmet needs around training, technology, and service delivery models. With appropriate tools, shared platforms, and a needs-led approach to staff development, telerehabilitation can enhance both teamwork and service responsiveness, even in pressured environments.

Survey items supporting this theme are summarized in Multimedia Appendix 5: Survey items theme 3.

Discussion

Interpretation of Principal Findings

This large-scale service evaluation successfully mapped current models and experiences of telerehabilitation delivery in community stroke services by engaging relevant stakeholders, including frontline clinical staff and people with lived experience of stroke, using a sequential mixed methods approach (evaluation aim 1). It also identified practical lessons and enabling factors to support the sustained integration of telerehabilitation into routine stroke rehabilitation pathways, in line with current policy-drivers (evaluation aim 2). As a result of in-depth interrogation of data from multiple sources using recognized methods [15], we developed a conceptual model to underpin and guide the successful sustained use of telerehabilitation in community stroke services (Figure 1). The model coalesced around three themes: risks and benefits; individualized care, and support for staff.

In considering the possible risks and benefits of using telerehabilitation in community stroke services, a complex interaction between the need to reduce burden on services while still building meaningful relationships and understanding personal environments was exposed. While it was clear that immediate practical needs might be met, for clinicians, developing a sense of the person in their social environment and being effective and safe was generally viewed as more challenging using telerehabilitation than face-to-face approaches. Similarly, people with lived experience of stroke generally disagreed that telerehabilitation was comparable to face-to-face delivery. This latter point contrasts with findings of a recent systematic review of randomized-controlled trials investigating factors influencing telerehabilitation delivery in stroke [11], reporting overwhelmingly positive feedback for access to and interaction with therapists via telerehabilitation. More promisingly, in the current evaluation, on deeper interrogation within this theme, many clinicians felt able to adapt and develop rapport, considering themselves able to successfully build relationships with service users using remote communication. Additionally, some evidence of forward-planning and improved self-management by people with stroke was apparent as they prepared for telerehabilitation sessions, in line with previous work exploring possibilities for self-management through the use of technology [17]. Clinicians also recognized clear efficiency gains when telerehabilitation was used appropriately. Reduced travel time, flexible scheduling, and the ability to conduct joint sessions remotely enabled more responsive working and, in some cases, increased therapy intensity without additional service burden. For people with lived experience, albeit a small sample size here, this flexibility was viewed positively when it translated into being seen more often or more quickly. These findings suggest that although telerehabilitation can make some aspects of relationship-building more difficult, it may also offer a pragmatic way of increasing rehabilitation dose within constrained workforce capacity. Further, carefully designed research is needed to understand the nuanced characteristics that underpin this paradigm shift in delivery, in concert with the views of people with lived experience.

It is unsurprising that the need for an approach to using telerehabilitation shaped by individual needs, preferences, and circumstances was a core finding of the evaluation; that a “one-size fits all” approach was not suitable. The importance of personalized care and shared decision-making throughout patient pathways is well-established in modern health services, with improved outcomes and experiences linked to people’s involvement in shaping their care [18,19]. Here, we found that the small group of people with lived experience of stroke who wanted to be fully involved with decisions about use, and that clinicians considered multiple individual factors when deciding whether to use telerehabilitation. These not only included clinical factors such as consideration of cognitive, sensory, or severe communication challenges—often seen as “red flags” for telerehabilitation to be avoided—but decisions were also shaped by factors such as people’s personal preferences and their perceived confidence with technology. A deeper exploration of that latter point revealed similarities with previous work on health technologies in stroke and neurorehabilitation—that the challenge is not to dismiss but to support people in taking advantage of the potential benefits of technological approaches without making assumptions about age or digital literacy [20,21]. In fact, here we demonstrated that age was least important in influencing decision-making by clinical teams, but access to appropriate equipment was considered critical to digital participation and reducing inequalities of provision. This theme demonstrated that thoughtful shared decision-making, supported by access to appropriate equipment and clear communication, can help ensure that remote delivery meets the needs of diverse individuals and avoids reinforcing existing inequalities.

XSL•FO
RenderX

A prevailing body of knowledge indicates that, for both technological advancement and other wide-scale service innovation in stroke and neurorehabilitation, supportive organizational cultures are central to sustained change [21,22]; the evaluation findings presented here reflect that understanding. Specifically, clinicians identified the value of a collaborative, multidisciplinary service approach and a need for meaningful, tailored training for themselves and people with lived experience. Telerehabilitation training needs were exposed on how to work with people to adapt their goals, how to assess suitability for the approach, and how to communicate effectively online—in essence, this approach is about much more than the simple mechanics of opening a laptop or smartphone and making a call, if outcomes for both individuals and the service are to be optimal. As recently captured in an extensive scoping review [23], successful use of telerehabilitation requires a range of competencies from both people with lived experience and clinical teams. Telerehabilitation is a new model of care, implemented at pace in response to urgent need, and as such, workforce transformation through tailored development and a needs-led approach to training is required to embrace its full potential.

Strengths and Limitations

The work met an identified need for reporting of the practicalities and challenges of use of telerehabilitation in NHS services [7] using a recognized approach to combining multiple data sources. This was a time-limited pragmatic service evaluation in a large geographical area of the United Kingdom (East of England) with community health provision across a mix of rural, suburban, and city-based populations. We acknowledge the limitation inherent in evaluating provision only within the UK NHS, but we suggest that key findings on enablers of sustained use of telerehabilitation in community stroke services have potential for application in other systems. The evaluation was supported by the existing integrated stroke delivery network, PPI advisory group, and we acknowledge that our group of participants with lived experience of stroke and their caregivers was small and therefore, while providing considerable insights, cannot represent the views of the wider service-user and caregiver populations. Despite this, we adopted an inclusive model of shared learning from service users and providers at every opportunity throughout. Where discussion groups included people with stroke, the facilitators endeavored to ensure their voices were heard. Similarly, we recognize the risk of bias inherent in surveys on a particular rehabilitation approach—that those with strong views and experiences in either direction might be most likely to respond. We made an attempt to negate this through a variety of recruitment strategies in collaboration with CD as the Stroke Quality Improvement for Rehabilitation lead. Finally, it is possible in pragmatic, rapid mixed methods evaluation that meaningful connections are lost due to limited integration or dominance of one dataset; this evaluation used recognized frameworks for study design, integration, interpretation, and reflexive approaches to minimize this risk.

Conclusions

Digital services have clear potential to address pressing unmet rehabilitation needs as demand ever-escalates, though understanding of implementation and sustained use in a variety of service contexts has, up to now, been lacking [23]. This evaluation has made initial steps to enhance that understanding for telerehabilitation in community stroke services and should provide a platform for further primary research.

Together, the findings presented here suggest that telerehabilitation works best not as a replacement for in-person care, but as part of a responsive, blended approach grounded in individual needs and service-user choice. The evaluation demonstrates that telerehabilitation has the potential to meet people's needs, while supporting service responsiveness and therapy intensity in pressurized clinical environments. A nuanced approach to implementation that includes appropriate training and carefully integrates the individual needs of service users and staff alongside a supportive organizational culture is required to support successful, sustained use.

Acknowledgments

The evaluation team is immensely grateful to NHS (National Health Service) East of England for recognizing the need for a rigorous service evaluation of telerehabilitation and funding the collaboration with the AbiliTec team at the University of East Anglia to provide it. We are very grateful to all clinical staff and people with lived experience who supported the work herein. VMP is supported by the NIHR (National Institute for Health and Care Research) HealthTech Research Centre in Brain Injury. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. Generative artificial intelligence (GenAI) was used for language editing (grammar and sentence structure only). No artificial intelligence tools were used for data analysis, interpretation, or manuscript content generation. The GenAI tool used was Chat GPT 4o (OpenAI). Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the outcomes.

Funding

This work was funded by NHS East of England as part of the Stroke Quality for Rehabilitation, national improvement projects.

Data Availability

Qualitative data (discussion group transcripts) are not publicly available due to the risk of participant identification and the confidential nature of the discussions. Deidentified extracts are presented within this paper. Quantitative survey data underlying

XSL•FO
RenderX

the reported findings are provided in aggregated form in Multimedia Appendices 2-4. The raw data supporting the conclusions of this paper can be made available from the corresponding author upon reasonable request.

Authors' Contributions

EC was the day-to-day project lead and was involved in conceptualization, data curation, analysis, investigation, methodology, project administration, validation, visualization, and co-led writing the original draft of the manuscript and reviewing and editing. CD was the NHS lead involved in funding acquisition, conceptualization, investigation, project administration, writing the original draft of the manuscript, and reviewing and editing. VMP was involved in funding acquisition, conceptualization, project administration, writing the original draft of the manuscript, and reviewing and editing. NJH was the evaluation lead for this study, and as such led on or was closely involved in conceptualization, data curation and analysis, investigation, methodology, managing resources, liaising with the funder, supervision of EC, and validation and co-led writing the original draft of the manuscript and reviewing and editing.

Conflicts of Interest

CD is employed by NHS England, who provided original funding for the Stroke Quality for Rehabilitation projects as part of the Stroke National Improvement programme, and her contributions can be found in the Authors' Contributions statement. All other authors have no conflicts to declare.

Multimedia Appendix 1
Evaluation alignment with reporting guidelines for mixed methods health service research: GRAMMS (Good Reporting of a Mixed Methods Study).
[DOCX File, 21 KB - jmir_v28i1e87741_app1.docx]

Multimedia Appendix 2
Discussion groups topic guide.
[DOCX File, 21 KB - jmir_v28i1e87741_app2.docx]

Multimedia Appendix 3
Survey items theme one.
[DOCX File, 16 KB - jmir_v28i1e87741_app3.docx]

Multimedia Appendix 4
Survey items theme two.
[DOCX File, 16 KB - jmir_v28i1e87741_app4.docx]

Multimedia Appendix 5
Survey items theme three.
[DOCX File, 18 KB - jmir_v28i1e87741_app5.docx]

References

1. Laver K, Walker M, Ward N. Telerehabilitation for stroke is here to stay. But at what cost? Neurorehabilit Neural Repair 2022 May 7;36(6):331-334. [doi: 10.1177/15459683221100492] [Medline: 35527716]
2. Laver KE, Adey-Wakeling Z, Crotty M, Lannin NA, George S, Sherrington C. Telerehabilitation services for stroke. Cochrane Database Syst Rev 2020 Jan 31;2020(1):b3. [doi: 10.1002/14651858.cd010255.pub3] [Medline: 32002991]
3. Appleby E, Gill ST, Hayes LK, Walker TL, Walsh M, Kumar S. Effectiveness of telerehabilitation in the management of adults with stroke: a systematic review. PLoS ONE 2019 Nov 12;14(11):e0225150. [doi: 10.1371/journal.pone.0225150] [Medline: 31714924]
4. National Clinical Guideline for Stroke for the UK and Ireland. : Intercollegiate Stroke Working Party; 2023 May 4. URL: https://www.strokeguideline.org/ [accessed 2026-05-26]
5. National service model for an integrated community stroke service. NHS England. 2022 Mar 16. URL: https://www.england.nhs.uk/publication/national-service-model-for-an-integrated-community-stroke-service/ [accessed 2026-05-26]
6. 10 Year Health Plan for England: fit for the future. GOV.UK. 2025 Jul 30. URL: https://www.gov.uk/government/publications/10-year-health-plan-for-england-fit-for-the-future [accessed 2026-05-26]
7. English C, Ceravolo MG, Dorsch S, et al. Telehealth for rehabilitation and recovery after stroke: state of the evidence and future directions. Int J Stroke 2022 Jan 5;17(5):487-493. [doi: 10.1177/17474930211062480] [Medline: 34983266]
8. Pearce L, Costa N, Sherrington C, Hassett L. Implementation of digital health interventions in rehabilitation: a scoping review. Clin Rehabil 2023 May 2;37(11):1533-1551. [doi: 10.1177/02692155231172299] [Medline: 37132030]

XSL•FO
RenderX

9. Edwards D, Kumar S, Brinkman L, et al. Telerehabilitation initiated early in post-stroke recovery: a feasibility study. Neurorehabilit Neural Repair 2023 Feb 1;37(2-3):131-141. [doi: 10.1177/15459683231159660] [Medline: 36876946]
10. Mallet KH, Shamloul RM, Lecompte-Collin J, Winkel J, Donnelly B, Dowlatshahi D. Telerehab at home: mobile tablet technology for patients with poststroke communication deficits—a pilot feasibility randomized control trial. J Speech Lang Hear Res 2023 Feb 13;66(2):648-655. [doi: 10.1044/2022_jslhr-21-00616] [Medline: 36634230]
11. Stephenson A, Howes S, Murphy PJ, et al. Factors influencing the delivery of telerehabilitation for stroke: a systematic review. PLoS ONE 2022 May 11;17(5):e0265828. [doi: 10.1371/journal.pone.0265828] [Medline: 35544471]
12. Cresswell J. Research Design: Qualitative, Quantitative and Mixed Methods Approaches: Sage Publications; 2009:203-204.
13. O'cathain A, Murphy E, Nicholl J. The quality of mixed methods studies in health services research. J Health Serv Res Policy 2008 Apr;13(2):92-98. [doi: 10.1258/jhsrp.2007.007074] [Medline: 18416914]
14. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
15. Braun V, Clarke V. One size fits all? What counts as quality practice in (reflexive) thematic analysis? Qual Res Psychol 2020 Aug 12;18(3):328-352. [doi: 10.1080/14780887.2020.1769238]
16. Semyonov-Tal K, Lewin-Epstein N. The importance of combining open-ended and closed-ended questions when conducting patient satisfaction surveys in hospitals. Health Policy OPEN 2021 Dec;2:100033. [doi: 10.1016/j.hpopen.2021.100033] [Medline: 37383512]
17. Demain S, Burridge J, Ellis-Hill C, et al. Assistive technologies after stroke: self-management or fending for yourself? A focus group study. BMC Health Serv Res 2013 Aug 22;13(1):334. [doi: 10.1186/1472-6963-13-334] [Medline: 23968362]
18. NHS universal personalised care: implementing the comprehensive model. : NHS; 2019 Jan URL: https://www.england.nhs.uk/wp-content/uploads/2019/01/universal-personalised-care.pdf [accessed 2026-05-26]
19. Ahmad N, Ellins J, Krelle H, Lawrie M. Person-centred care: from ideas to action: bringing together the evidence on shared decision making and self-management support. : The Health Foundation; 2014 Oct URL: https://www.health.org.uk/reports-and-analysis/reports/person-centred-care-from-ideas-to-action [accessed 2026-05-26]
20. Drake A, Sassoon I, Balatsoukas P, et al. The relationship of socio-demographic factors and patient attitudes to connected health technologies: a survey of stroke survivors. Health Informatics J 2022;28(2):14604582221102373. [doi: 10.1177/14604582221102373] [Medline: 35726817]
21. Pradhan S, Levin MF, Hancock NJ, Murphy MA. Uptake of technology for neurorehabilitation in clinical practice: a scoping review. Phys Ther 2024 Feb 1;104(2):pzad140. [doi: 10.1093/ptj/pzad140] [Medline: 37856528]
22. Hancock NJ, Houghton J, Jones F. Integrating an approach to personalised self-management support in stroke and neurorehabilitation service contexts: People1st – a quality improvement initiative. Disabil Rehabil 2022 Oct 27;45(19):3034-3045. [doi: 10.1080/09638288.2022.2127930] [Medline: 36301996]
23. Stark AL, Krayter S, Dockweiler C. Competencies required by patients and health professionals regarding telerehabilitation: a scoping review. Digit HEALTH 2023;9:20552076231218841. [doi: 10.1177/20552076231218841] [Medline: 38107985]

Abbreviations

GRAMMS : Good Reporting of a Mixed Methods Study
MDT: multidisciplinary team
NHS: National Health Service
PPI: patient and public involvement

Edited by A Stone; submitted 20.Nov.2025; peer-reviewed by AJ Umeano, A Ikemefuna, F Torreno, P Adedigba; revised version received 05.May.2026; accepted 05.May.2026; published 11.Jun.2026.

Please cite as:
Chandler E, Dorer C, Pomeroy VM, Hancock NJ
Telerehabilitation in Community Stroke Services: Mixed Methods Evaluation of Current Practice and Lessons for Sustained Use
J Med Internet Res 2026;28:e87741
URL: https://www.jmir.org/2026/1/e87741
doi:10.2196/87741

© Elizabeth Chandler, Charlotte Dorer, Valerie M Pomeroy, Nicola J Hancock. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN

XSL•FO
RenderX

1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development and Preliminary Validation of ORCA-PD, an Online Rapid Cognitive Assessment for Parkinson Disease: Mixed Methods Study

Avigail Lithwick Algon[1], MBA; Sarah Brisman[1]; Chi-Ying R Lin[2], MD; William Saban[1], PhD

[1]Department of Occupational Therapy, Gray Faculty of Medical & Health Sciences, Sagol School of Neuroscience, Center for Accessible Neuropsychology, Tel Aviv University, Chaim Levanon St 55, Tel Aviv, Israel

[2]Department of Neurology, Baylor College of Medicine, Houston, TX, United States

Corresponding Author:
William Saban, PhD
Department of Occupational Therapy, Gray Faculty of Medical & Health Sciences, Sagol School of Neuroscience, Center for Accessible Neuropsychology, Tel Aviv University, Chaim Levanon St 55, Tel Aviv, Israel

Abstract

Background: Traditional in-person neuropsychological tests for Parkinson disease (PD) lack accessibility, scalability, and PD specificity. Mobility impairments hinder access to in-person assessments, and long waiting times for expert evaluation limit scalability. Common tools for cognitive screening, such as the Montreal Cognitive Assessment, are generic and not specific to PD.

Objective: The goal of this study was to address these challenges by leveraging the internet. This study aimed to develop a sensitive tool to detect cognitive impairments in early- to mid-stage PD in an accessible and scalable manner.

Methods: We developed the Online Rapid Cognitive Assessment for Parkinson's Disease (ORCA-PD), a brief (~15 min), fully self-administered tool, tailored to detect PD-specific impairments remotely. In a cross-sectional study, 112 participants from diverse geographical locations completed the ORCA-PD and a usability questionnaire. Task selection for ORCA-PD was guided by meta-analyses and comprehensive reviews, which demonstrated medium-to-high sensitivity to PD compared to healthy controls.

Results: Participants from more than 30 geographical locations showed a 93% (135/145) completion rate, with median usability ratings of 4 to 5 (IQRs 0-1) out of 5, indicating strong usability across a diverse sample. The ORCA-PD score significantly correlated with the Montreal Cognitive Assessment score (ρ=0.45, 95% CI 0.21-0.64; P<.001; n=56), supporting convergent validity. The ORCA-PD scores also followed expected significant trends: neurotypical controls scored higher than participants with PD (β=4.18; P=.048) and participants with PD without mild cognitive impairment scored higher than participants with PD with mild cognitive impairment (β=7.22; P=.046), confirming 2 discriminative abilities.

Conclusions: ORCA-PD demonstrated usability and preliminary validity. ORCA-PD offers an accessible, scalable, and PD-sensitive cognitive screening test that could complement traditional in-person, supervised tools.

(J Med Internet Res 2026;28:e90057) doi:10.2196/90057

KEYWORDS

online; neuropsychological testing; Parkinson disease; accessibility; cognition

Introduction

As global life expectancy increases, the disease burden of age-associated neurodegenerative diseases, such as Parkinson disease (PD), is rising [1]. PD is the most common age-related motor neurodegenerative disorder [2], characterized primarily by motor symptoms [3]. However, PD also leads to nonmotor symptoms, such as a decline in cognition [4]. This decline often remains covert and goes unnoticed by clinicians and people with PD [4]. Despite the importance of cognitive evaluations for people with PD, current in-person assessment tools present several challenges.

Traditional cognitive screening tools for PD have long faced challenges related to accessibility, scalability, and PD specificity [5]. Accessibility is a persistent challenge, particularly for people with PD who face barriers to in-person cognitive assessments [5-7]. It is often challenging for people with PD to attend in-person cognitive assessments, especially those with severe motor and mobility impairments [8]. Motor impairments, including bradykinesia and postural instability, can make travel particularly challenging. In addition, PD-related fatigue may reduce their capacity to participate in testing [9,10]. Cognitive deficits may further complicate participation, as people with PD may struggle with task switching [11], working memory

XSL•FO
RenderX

[12], visuospatial abilities [13], and impulsivity [14]. Thus, people with PD often find it challenging to navigate and commute to laboratories and clinics.

Beyond accessibility concerns, scalability is limited by the fact that most cognitive assessments must be administered by trained professionals, such as neuropsychologists or neurologists, who are often in high demand [15]. This issue is especially pronounced for people with PD living in remote areas with limited specialist access [5]. Although telehealth approaches can extend clinician-led cognitive assessments to remote settings, they remain constrained by specialist availability, scheduling burdens, and clinician time, limiting their scalability for large-scale screening [16]. Furthermore, recruitment challenges often result in small sample sizes (frequently fewer than 30 participants), drawn from single institutions and limited geographic regions [5], raising concerns about sampling bias and the generalizability of findings [17].

Regarding specificity, commonly used cognitive screening tools such as the Montreal Cognitive Assessment (MoCA) [18] and the Mini-Mental State Examination(S) [19] are generic and not tailored to the cognitive profile of PD. Although tasks such as clock drawing and cube copying are intended to measure visuospatial and executive functions—both relevant to cognitive decline in PD—their reliance on motor execution may confound interpretation, since motor impairments can limit performance independently of cognitive status.

We propose using a web-based service to overcome these barriers. Recent work has explored the feasibility of web-based cognitive assessments to enhance accessibility and scalability. Behavioral researchers in different domains have been relying on the internet to reach larger [20] and more diverse populations [21]. Although self-administered cognitive tasks have been evaluated in neurotypical controls (NT) [22-25] and individuals with mild cognitive impairment (MCI) [26], relatively few studies have examined their usability and validity in PD [27-31]. Given the cognitive profile of PD [32] even before PD-dementia, it is important to use sensitive tools capable of detecting cognitive differences between PD and NT, as well as between participants with PD without MCI (PD-non-MCI) and participants with PD with MCI (PD-MCI). Thus, aiming to further address these accessibility, scalability, and PD-specificity challenges, we developed the Online Rapid Cognitive Assessment for Parkinson's Disease (ORCA-PD), a fully self-administered, web-based cognitive screening tool. ORCA-PD was tailored to capture early- to mid-stage PD-specific cognitive impairments without increasing motor demands for in-person visits.

Here, we aimed to evaluate the feasibility and preliminary validity of ORCA-PD as a remote, self-administered cognitive screening tool tailored for PD. First, we evaluated whether ORCA-PD can be feasibly completed outside the clinic in diverse geographical locations and whether users find it usable. Second, to situate ORCA-PD within existing assessment practices, we examined its convergent validity with the MoCA. Finally, we tested its ability to distinguish neurotypical individuals from people with PD, as well as PD-non-MCI from PD-MCI cohorts.

Methods

Overview

First, we developed the ORCA-PD by selecting measures based on prior meta-analyses and reviews reporting medium-to-large effect sizes for differences between PD and NT. Second, to evaluate feasibility, we administered the ORCA-PD to people with PD and NT with or without MCI. Third, we tested whether performance declined as task difficulty increased, assessing the online tasks' internal structure. Fourth, we administered a usability questionnaire.

Participants: Power Analysis

To determine the required sample sizes, we performed a power analysis (α=.05; power=0.99) using effect sizes obtained from 5 in-person studies comparing PD and NT on the MoCA [33-37]. These analyses indicated a minimum required sample size of 14 participants per group. Accordingly, our larger sample size (>50 participants, as given below) provided sufficient power to detect group differences and accounted for the increased variability associated with remote administration (eg, differences in test conditions).

In this cross-sectional study, 112 participants (56 PD and 56 NT) completed the ORCA-PD. The groups were matched for age ($t_{107.06}$=−1.71; P=.09) and gender (χ^2_1=0.00; P>.99). However, the PD group had significantly more years of education than the NT group ($t_{99.30}$=−2.92; P=.004); thus, education was a covariate in the analyses. Table 1 shows the demographic and clinical characteristics of both groups. Within the sample, a subgroup of 56 participants (PD=41; NT=15) also completed the MoCA. Participants completed the MoCA during a videoconference session, while the ORCA-PD was sent to the participants to be completed on their own time. The MoCA required scheduling a videoconference session with the experimenter. Thus, not all participants who completed the ORCA-PD also completed the MoCA (56/112). This higher completion rate of the ORCA-PD compared to the MoCA is another indication of the higher utility of the ORCA-PD as a self-administered task. There were no significant differences in MoCA scores between the groups ($t_{26.87}$=1.35; P=.19).

Importantly, we selected participants with PD who had relatively high cognitive scores, allowing us to test whether our new online cognitive assessment could distinguish between cognitively similar PD and NT groups. This is critical for developing a test that is sensitive to subtle cognitive group differences between participants with PD with early cognitive dysfunction and NT participants.

Participants with PD were primarily recruited through the Center for Accessible Neuropsychology (CAN) database, which includes individuals who had previously participated in CAN assessments or responded to online advertisements (eg, via PD associations such as the Israel Parkinson Association). NT participants were recruited online through the CAN database and through relatives or caregivers of participants with PD.

Inclusion criteria required participants to be aged between 40 and 90 years and fluent in Hebrew or English, with PD

XSL•FO
RenderX

diagnosed after the age of 40 years, with varying degrees of PD stage (Hoehn and Yahr [H&Y] scale ≤4), and with or without MCI (19<MoCA<26). Individuals with other neurological (non-PD) or psychiatric disorders, learning disabilities, or significant visual or auditory impairments were excluded.

Table . Demographic and clinical characteristics for study participants (N=112).

Participant group	Participants, n	Age (y), mean (SD)	Education (y), mean (SD)	Women, n/N (%)	MoCA[a], mean (SD)	MDS-UPDRS III[b], mean (SD)	Disease duration (y), mean (SD)
Parkinson disease	56	67.2 (8.0)	17.5 (4.7)	28/56 (50.0)	25.2 (3.1)	21.9 (14.2)	7.4 (11.4)
Neurotypical controls	56	64.4 (9.4)	15.3 (3.3)	29/56 (51.8)	26.4 (2.9)	—[c]	—

[a]MoCA: Montreal Cognitive Assessment.

[b]MDS-UPDRS III: The International Parkinson and Movement Disorder Society—Unified Parkinson's Disease Rating Scale Part III.

[c]Not applicable.

A total of 145 participants were enrolled in the study, of whom 135 (93%: English=91.5%, Hebrew=96.1%) met the inclusion criteria. Twelve participants were excluded based on eligibility criteria (2 with attention-deficit/hyperactivity disorder, 2 with dyslexia, and 8 who did not meet age requirements), and 1 participant was excluded due to incomplete data, leaving 122 eligible participants. In this sample, the youngest participant with PD was 46 years old. Accordingly, the control group was restricted to the same minimum age to ensure comparability between groups. From this pool, 10 control participants were randomly selected to obtain groups with comparable age distributions, resulting in a final sample of 112 participants (English=76 and Hebrew=36).

Neurological and Neuropsychological Assessment

For a subgroup of participants, we administered neuropsychological testing. We followed the International Protocol for Online Neuropsychological Testing [5-7,38-40]. This protocol includes a script and procedure for participant recruitment through online platforms and for conducting online evaluations [6]. The experimenter contacted interested participants via email to schedule a live interview. Prior to each session, video and audio issues were addressed to minimize disruptions. If technical difficulties arose during the session, they were typically resolved within a few minutes. In rare cases (<5%, 5/112) where the technical difficulties could not be resolved, the session was rescheduled.

After completing the informed consent form, participants completed a demographic and medical questionnaire. Through the questionnaire, the experimenter collected medical history information, including age at diagnosis, medication, primary symptoms, genetic subtype, diet, other neurological or psychiatric conditions, and other relevant information pertaining to their health and lifestyle. Information regarding medications (eg, carbidopa or levodopa, cholinesterase inhibitors, and *N*-methyl-D-aspartate receptor antagonists) was collected through the medical history questionnaire. Because medication type, dosage, and timing varied across participants, and detailed pharmacological modeling was beyond the scope of this validation study, medication effects were not included in the analyses.

The experimenter then administered the MoCA (version 8.1) in the participant's preferred language (Hebrew or English). The MoCA was administered via videoconference following the official guidelines on the MoCA website. Since the MoCA assessment was administered online, several adjustments were made based on previous papers [40-42]. Participants were instructed in advance and at the start of the session to have a pen and paper available. Visual items (eg, visuospatial and naming tasks) were presented via screen sharing. For the Trail Making Test, responses were provided verbally. For the drawing tests, participants were asked to draw at home and show the drawing on the screen to the experimenter. First, participants were shown a cube on the screen with the label "Copy cube" and were instructed to replicate it. Next, participants were presented with the label "Draw clock" and were asked to draw a clock with a specific time given by the experimenter. For Naming, stimuli were presented individually. For orientation, participants were asked to close their eyes and report the date and location. The MoCA took approximately 10 minutes to complete [18,40].

To assess the motor severity of participants with PD, the experimenter next administered the International Parkinson and Movement Disorder Society—Unified Parkinson's Disease Rating Scale Part III (MDS-UPDRS III) as well as the H&Y scale [43]. Then, a link to the self-administered ORCA-PD test and the usability questionnaire was sent to participants' emails to be completed within 2 weeks of the videoconferencing assessment. Participants completed ORCA-PD in the same language as the MoCA to ensure linguistic consistency.

ORCA-PD: Remote, Unsupervised, and Self-Administered Cognitive Tool

The ORCA-PD comprises 8 brief cognitive measures—5 performance-based tasks and 3 questionnaires—tailored to be sensitive to PD (Figure 1). These measures were selected based on converging evidence from meta-analyses and comprehensive reviews identifying cognitive measures that are sensitive to PD, yielding medium-to-large effect sizes when differentiating PD from NT participants [11-13,44,45]. Additionally, all tasks were adapted for online administration and optimized for older participants. For example, visual stimuli were designed to be large and simple, enhancing accessibility for older adults. To reduce the influence of PD-related motor impairments, motor

XSL•FO
RenderX

demands were minimized in each task. For example, instead of using a task that required precise mouse control, which could potentially affect cognitive measurement in PD as a motor impairment-related confounder, the tasks were mostly implemented using simple forced 2-choice response paradigms. In addition, accuracy was emphasized as the primary dependent variable.

Figure 1. The measures of ORCA-PD (Online Rapid Cognitive Assessment for Parkinson's Disease): (A) digit span forward, (B) digit span backward, (C) digit span order, (D) line orientation match, (E) alternate task switching, (F) impulsivity, (G) subjective cognitive complaints—self-assessment, and (H) subjective cognitive complaints—other assessment.

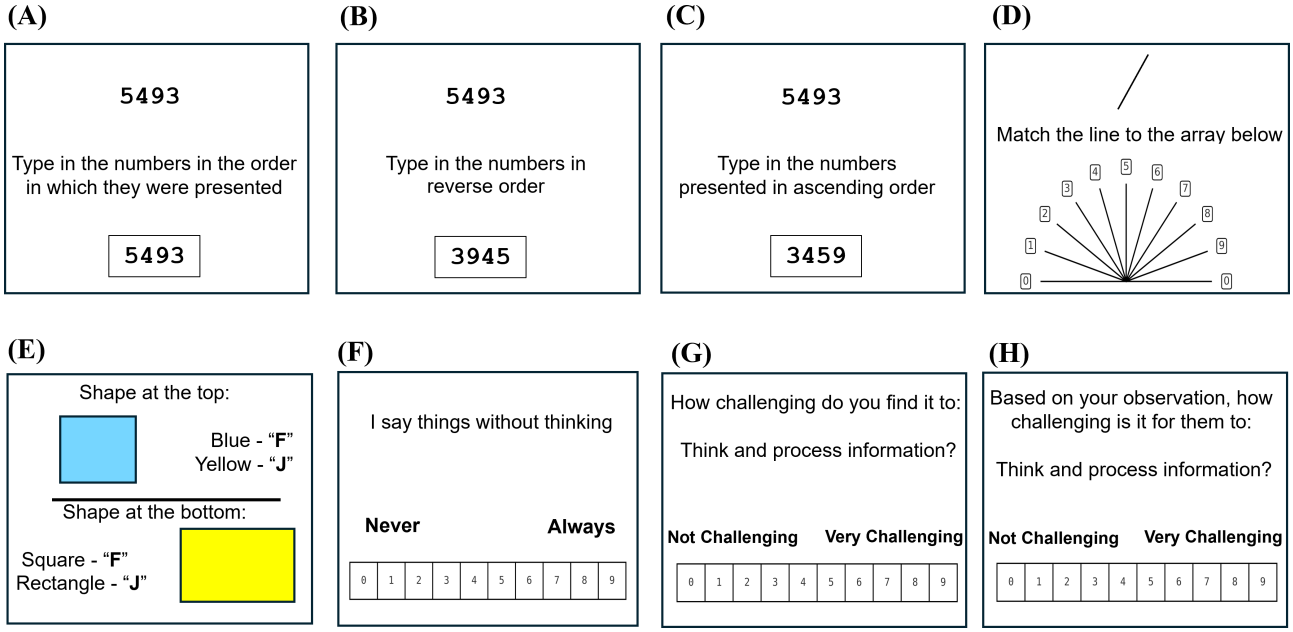

The first 3 tasks in the ORCA-PD were 3 versions of the digit span (DS) task. These consisted of digit span forward (DSF) for establishing baseline abilities, digit span backward (DSB) [11] (effect size=0.46), and digit span order (DSO) [12] (effect size=0.75). Each DS task consisted of 2 difficulty levels, with level 2 (hard) comprising longer digit sequences than level 1 (easy). Sequence lengths for level 1 were 4 (DSF and DSB) or 5 (DSO) and for level 2 were 6 (DSF and DSB) and 7 (DSO). These were selected based on previous work by Grogan et al [46]. Following recommendations from the Movement Disorder Society (MDS) Task Force [32], we adapted the Benton Judgment of Line Orientation task [47] for the fourth task, herein referred to as the line orientation match (LOM) [13] (effect size=0.52). Participants matched pairs of lines based on their orientation (9 options). For the LOM task, difficulty was defined by orientation discriminability, where trials with straight angles (0°, 180°, 90°) were classified as easy, while all other angles (18°, 36°, 54°, 72°, 108°, 126°, 144°, and 162°) were classified as hard. Each orientation was presented twice, resulting in 4 easy trials and 16 hard trials. The fifth cognitive task was an alternate task switching (ATS) task involving shape and color categorization, based on target location, consistent with MDS recommendations and prior meta-analyses [11,32,48]. To reduce motor demands [5], the participant responded by pressing a single key per trial (effect sizes for similar paradigms, such as the Trail Making Test B, were 0.67 to 0.89). Trials were categorized as repeat (same as trial n−1; easy) or switch (different from trial n−1; hard), enabling assessment of task sensitivity. The task comprised 33 trials: 16 repeat trials, 16 switch trials, and 1 initial trial.

The sixth measure was a short questionnaire assessing impulsivity [49]. Based on a recent meta-analysis [44], we used items from the Barratt Impulsiveness Scale (BIS-11). Specifically, we selected items only from the BIS-11 subscales that were found to be sensitive to PD [44] (ie, attention and self-control). From these 2 nonmotor subscales, we selected 7 items that were found to be the most sensitive to PD [50], with responses scored on a scale of 0 to 9. These items represent a targeted subset of the BIS-11 and do not constitute a stand-alone validated impulsivity scale. The final 2 measures assessed subjective cognitive complaints (SCC) [51] using both self-report and informant-report questionnaires. The inclusion of SCC was motivated by prior evidence indicating that SCC are present in approximately 36% of individuals with PD [45]. The questionnaires assessed 6 domains: global cognition and 5 specific cognitive domains (attention, memory, visuospatial function, executive function, and language), consistent with domains recommended by the MDS Task Force [52]. Informant ratings were obtained from individuals familiar with the participant's everyday functioning (eg, spouse, partner, caregiver, or close friend). Responses were scored on a scale of 0 to 9 and normalized to a range of 0 to 1.

Each of the eight measures was scored from 0 to 1, with higher values reflecting better performance. To standardize the questionnaire scores from 0 to 1, each score was adjusted so that higher scores would indicate less impairment by reversing the score where necessary. Accordingly, a greater impulsivity score indicated lower impulsivity, and a higher SCC score indicated lower concern. The total score for each questionnaire was averaged and then divided by 9 to yield a scale ranging from 0 to 1. Finally, the total ORCA-PD score was calculated

XSL•FO
RenderX

as the simple mean of the 8 measures for each participant, such that each task and questionnaire contributed equally to the final score (ORCA-PD naive score).

Upon completing the ORCA-PD, participants were asked to provide feedback using the usability questionnaire comprising 4 questions, each rated on a 5-point Likert scale.

Although partial data could be recovered in cases of interruption, the predefined criterion for analysis was the completion of the full assessment. In cases where a session was interrupted, participants were typically provided with a new access link and asked to restart the assessment. When multiple attempts occurred, only the fully completed session was included in the analysis. In practice, interruptions were uncommon (<12% of sessions), and cases involving mid-test disconnections with incomplete data were rare (<2%).

Statistical Analysis

Data preprocessing and visualization were performed in R (version 4.3.2; RStudio Team, 2023) [53]. The correlation heatmap was generated in Python. To evaluate group differences (eg, NT vs PD), independent samples *t* tests were used for continuous variables (eg, ORCA-PD score), and chi-square tests were used for categorical variables (eg, gender). These comparisons were evaluated using 2-tailed tests. For hypothesis-driven analyses of ORCA-PD outcomes, group comparisons were evaluated using 1-tailed tests when the direction of the effect was specified a priori. Where appropriate, group comparisons were evaluated using linear regression, adjusting for years of education and age as covariates. Correlations between measures were assessed using Pearson correlation coefficient r or Spearman rank correlation ρ, depending on variable distribution and scaling. Convergent validity was assessed via correlations with MoCA scores, where it was hypothesized that ORCA-PD scores would be positively associated with MoCA scores. Construct validity was assessed via group comparisons (NT vs PD; MCI vs non-MCI). It was hypothesized that the MCI group would demonstrate lower ORCA-PD scores than the group without MCI, and that the PD group would demonstrate higher ORCA-PD scores than the NT group.

The internal structure of task-based measures was evaluated by examining whether participants' performance declined with increasing task difficulty. Internal consistency of questionnaire-based measures was assessed using Cronbach α and McDonald ω. In addition, a correlation matrix of all ORCA-PD components was computed to further characterize the internal structure of the ORCA-PD measures. Discriminant validity was also assessed via the association between the ORCA-PD and the MDS-UPDRS III scores. As a secondary exploratory analysis, potential language-related effects were examined by comparing performance between English- and Hebrew-speaking cohorts within each group.

Ethical Considerations

This protocol was approved by the Tel Aviv University Institutional Review Board Committee (#0005713‑6). All participants provided informed consent.

Results

On average, participants completed the ORCA-PD in 15 (SD 4.5; range 12‑25) minutes. This reflects task completion time only and does not include presession setup (eg, audio or video checks) or any additional interaction time. Even with this distinction, this reflects minimal participant burden and supports the ORCA-PD's usability for remote assessment. Participant feedback further supported usability: median ratings indicated that instructions were very clear (5, IQR=1), the display size was comfortable (5, IQR=0), the overall experience was enjoyable (4, IQR=1), and the completion time was good (5, IQR=1) (Figure 2).

XSL•FO
RenderX

Figure 2. Participant (N=112) responses to the usability questionnaire after completing the ORCA-PD (Online Rapid Cognitive Assessment for Parkinson's Disease). Four questions addressed the clarity of questions, font size, completion time, and overall enjoyment of participation. The median is marked with a diamond symbol.

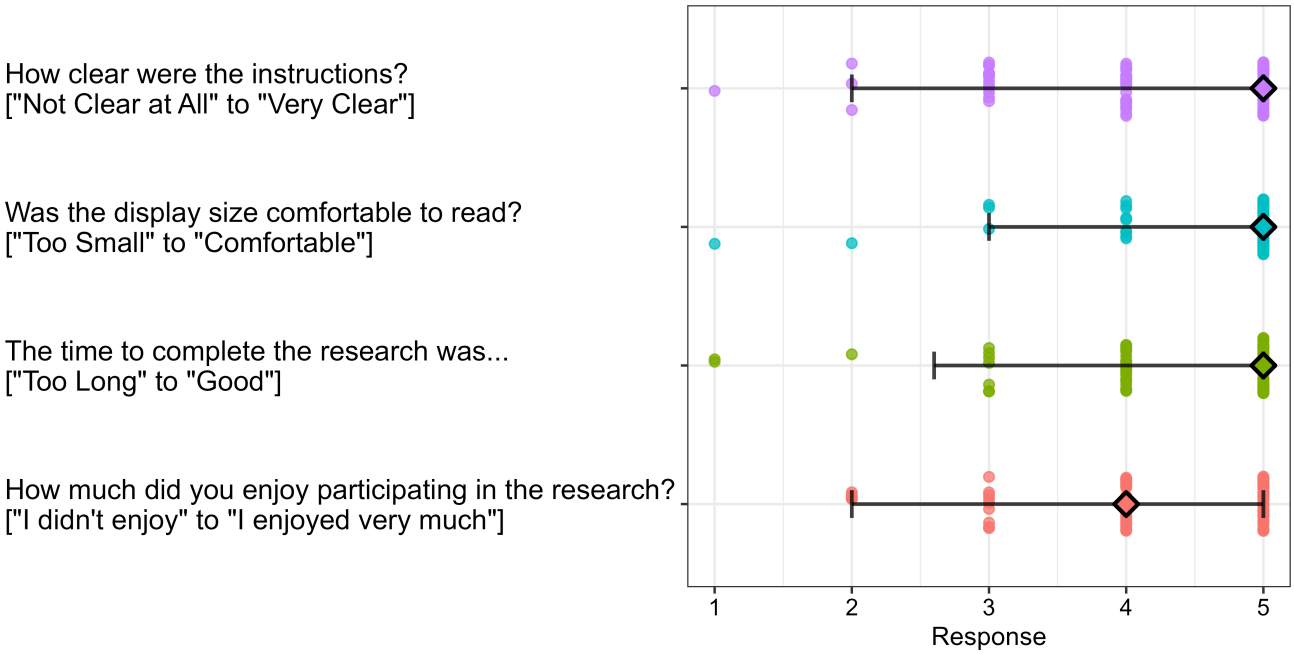

To assess the ORCA-PD's sensitivity to expected patterns of task difficulty, we evaluated whether participant performance declined with increasing task difficulty. Such a pattern would support the expected performance pattern (ie, sanity check) of the online tasks in measuring the intended cognitive constructs (Figure 3). As expected, in 3 versions of the digit span task, participants performed significantly better on the easy level (level 1) than on the hard level (level 2) trials (forward: t_{111}=8.99, P<.001; Cohen d=0.85; backward: t_{110}=10.20, P<.001; Cohen d=0.97; order: t_{111}=11.1, P<.001; Cohen d=1.05), supporting the expected performance pattern of these online tasks. In the LOM task, participants performed significantly better on the easy trials (t_{111}=12.18, P<.001; Cohen d=1.15). In addition, easy trials showed high accuracy (98%), whereas the hard trials showed lower mean accuracy (76.6%). In the ATS, participants performed significantly better on repeat (easy) trials than on switch (hard) trials (t_{111}=6.23, P<.001; Cohen d=0.59).

Figure 3. Participant (N=112) accuracy as a function of difficulty level for all ORCA-PD (Online Rapid Cognitive Assessment for Parkinson's Disease) tasks. Accuracy decreased as task difficulty increased in each of the 5 tasks, supporting the expected performance pattern of the ORCA-PD online measures. (A) Digit span forward, (B) digit span backward, (C) digit span order, (D) line orientation match, and (E) alternate task switching. Each data point represents a single participant. Bars represent group means. The error bars represent SEs.

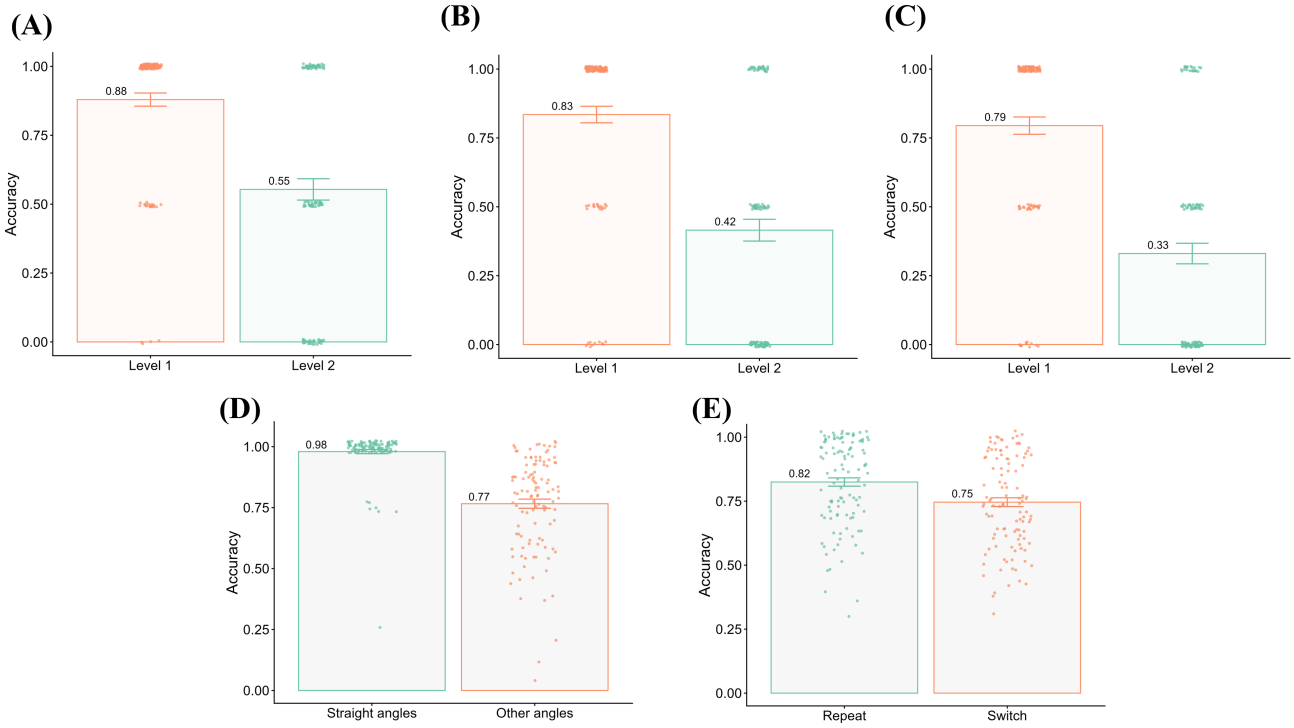

XSL•FO
RenderX

Regarding questionnaires, for impulsivity, internal consistency was medium (Cronbach α=0.68; McDonald ω=0.80). For the self-reported SCC questionnaire, internal consistency was good (Cronbach α=0.83; McDonald ω=0.91). For the informant-reported SCC questionnaire, internal consistency was strong (Cronbach α=0.90; McDonald ω=0.92).

To examine relationships among the different ORCA-PD measures, we computed a correlation matrix of all ORCA-PD components across participants (Figure 4). The heatmap below displays interfeature correlations (range –0.04 to 0.70), which is consistent with the multidimensional structure of the ORCA-PD.

Figure 4. Correlation heatmap of the ORCA-PD (Online Rapid Cognitive Assessment for Parkinson's Disease) task components (N=112). Pearson correlations are displayed using a blue color scale, where deeper shades represent stronger relationships. * indicates *P*<.05. ATS: alternate task switching; DSB: digit span backward; DSF: digit span forward; DSO: digit span order; LOM: line orientation match; SCC: subjective cognitive complaints.

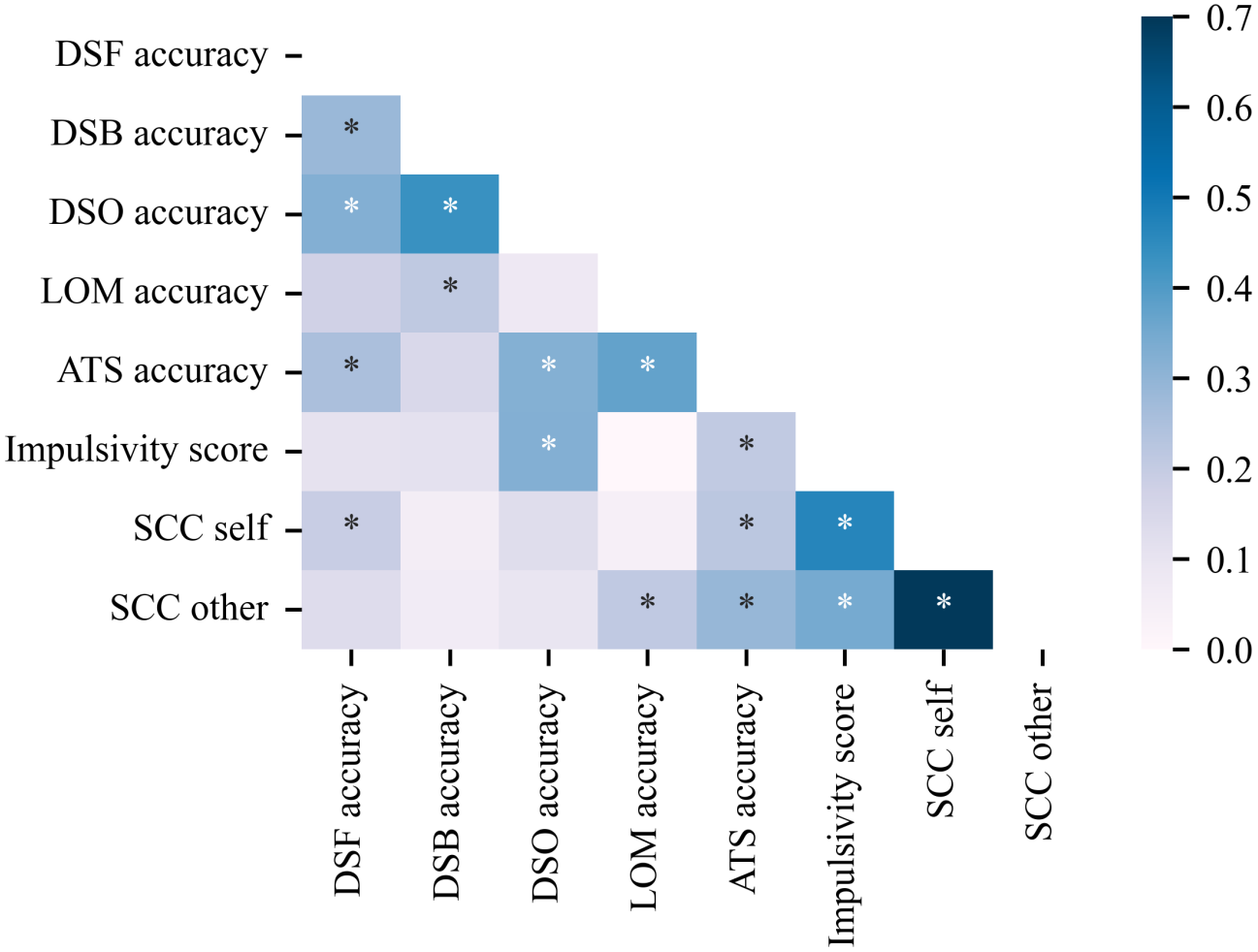

Additionally, to evaluate the discriminant validity of ORCA-PD and its ability to dissociate cognitive from motor functions, correlations were examined with a motor measure, the MDS-UPDRS III. As expected, no significant correlations were found between any ORCA-PD measures and the MDS-UPDRS III scores, supporting the notion that ORCA-PD primarily assesses cognitive abilities (DSF: ρ=–0.25, *P*=.11; DSB: ρ=–0.19, *P*=.25; DSO: ρ=–0.24, *P*=.14; ATS: ρ=–0.07, *P*=.67; and LOM: ρ=–0.09, *P*=.57).

Importantly, to assess convergent validity, we examined the correlation between the ORCA-PD naive score and MoCA scores (Figure 5, n=56). As expected, across both groups, a significant positive correlation was observed (ρ=0.45, 95% CI 0.21-0.64; *P*<.001), indicating a moderate association. When examined separately, significant associations were observed in both the PD (ρ=0.41, 95% CI 0.12-0.64; *P*=.004) and NT (ρ=0.62, 95% CI 0.15-0.86; *P*=.007) groups, supporting the preliminary validity of ORCA-PD.

XSL•FO
RenderX

Figure 5. Scatterplot showing the relationship between ORCA-PD naive scores and MoCA scores (n=56). Dashed lines indicate group-specific trends, while the solid black line indicates the overall trend across both groups. MoCA: Montreal Cognitive Assessment; ORCA-PD: Online Rapid Cognitive Assessment for Parkinson's Disease; PD: Parkinson disease.

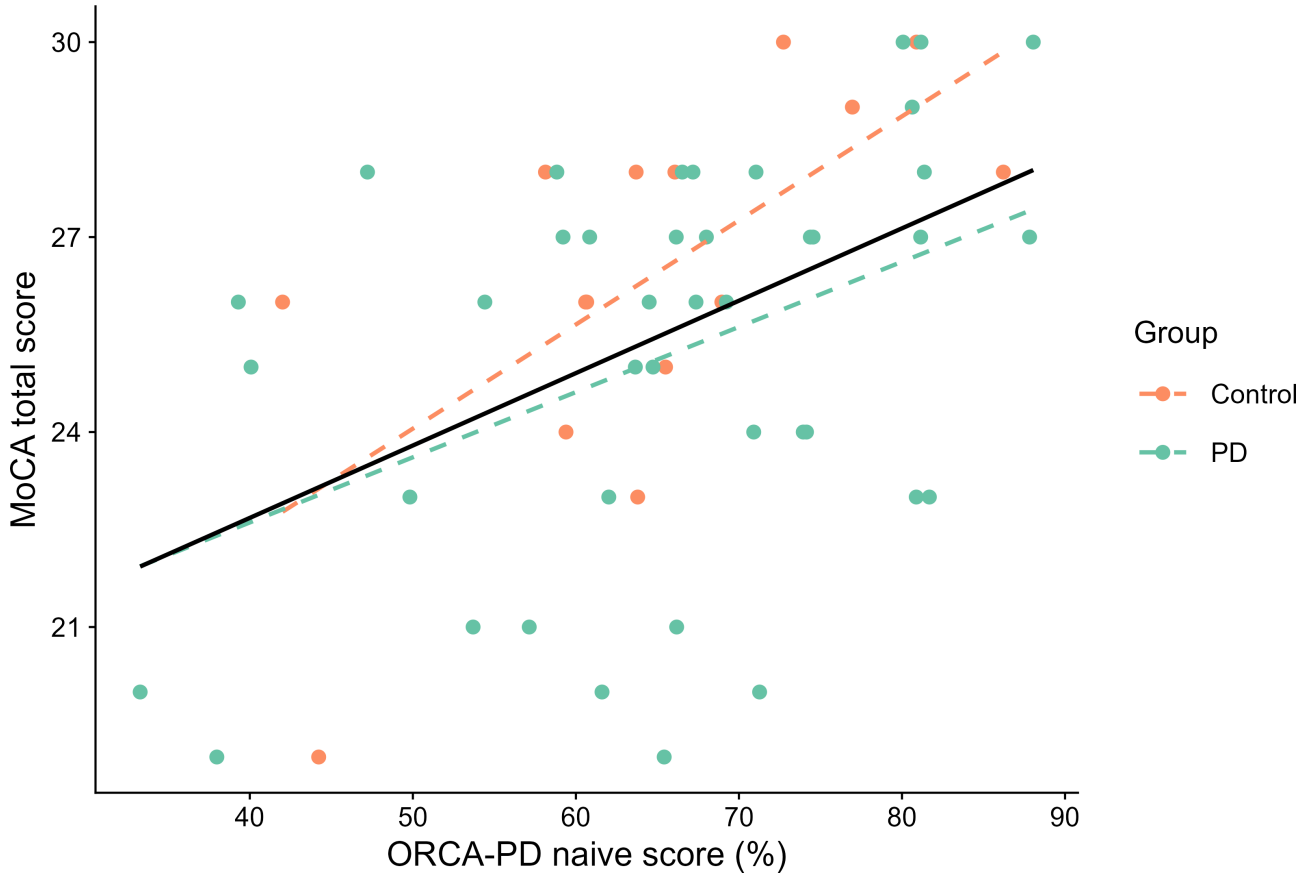

To examine the ability of ORCA-PD to distinguish between people with or without MCI, we classified participants according to established criteria for MCI (ie, MoCA between 19 and 25) [54]. In a linear regression analysis adjusting for education and age, neither education (β=0.37; *P*=.34) nor age (β=–0.27; *P*=.18) was a significant predictor of ORCA-PD naive scores. As expected, the non-MCI group demonstrated higher ORCA-PD naive scores compared to the MCI group (β=6.07; *P*=.046). When restricting the analysis to PD only, neither education (β=0.37; *P*=.40) nor age (β=–0.23; *P*=.40) was a significant predictor of ORCA-PD naive scores. As expected, the non-MCI group again demonstrated higher ORCA-PD naive scores compared to the MCI group (β=7.22; *P*=.046). This analysis was not conducted on NT alone, as there were fewer than 5 participants in this group with MCI. These results support the discriminative ability of the ORCA-PD, demonstrating that it can capture cognitive impairments.

Importantly, the ORCA-PD's discriminative ability was also assessed by comparing ORCA-PD scores between PD and NT (Figure 6). In the regression model, neither education (β=0.27; *P*=.36) nor age (β=–0.19; *P*=.16) was a significant predictor of ORCA-PD naive scores. As expected, the group effect was significant (β=–4.18; *P*=.048), with PD scoring lower than NT.

XSL•FO
RenderX

Figure 6. Comparison between Parkinson disease (PD) and controls (N=112) for (A) ORCA-PD (Online Rapid Cognitive Assessment for Parkinson's Disease) naive score, (B) digit span forward accuracy, (C) digit span backward accuracy, (D) digit span order accuracy, (E) line orientation match accuracy, (F) alternate task switching accuracy, (G) impulsivity score, (H) subjective cognitive complaints—self-assessment, and (I) subjective cognitive complaints—other assessment. * indicates $P<.05$.

XSL•FO
RenderX

To explore the specific cognitive domains contributing to the group differences, the NT and PD groups on all individual measures were compared. NT participants performed significantly better than PD on DSO accuracy (t_{107}=2.67, P=.004). In addition, NT participants scored significantly higher in impulsivity than participants with PD (t_{110}=2.15, P=.02). No significant group differences were found on DSF (t_{110}=0.16, P=.44), DSB (t_{109}=0.87, P=.19), ATS (t_{110}=0.27, P=.39), LOM (t_{110}=–0.06, P=.52), or SCC measures (self: t_{110}=1.28, P=.10 and other: t_{110}=0.44, P=.33).

As an exploratory analysis, which needs to be further validated in future studies, we compared scores between English- and Hebrew-speaking cohorts. At the level of overall ORCA-PD naive scores, differences between English- and Hebrew-speaking participants were not statistically significant in the NT group ($t_{27.49}$=2.02, P=.05). However, in the PD group, the English-speaking cohort had significantly higher scores ($t_{31.67}$=2.19, P=.04). At the task level, language-related differences were observed for digit span forward in the NT group, where performance was significantly higher in the English-speaking cohort ($t_{20.16}$=2.90, P=.009). In the ATS task in the PD group, performance was higher in the English-speaking cohort ($t_{36.49}$=2.19, P=.04). For all other measures, no significant language differences were observed in the NT group (DSB: P=.68; DSO: P=.19; LOM: P=.29; ATS: P=.17; impulsivity: P=.29; SCC-self: P=.79; SCC-other: P=.96) or in the PD group (DSF: P=.05; DSB: P=.92; DSO: P=.16; LOM: P=.72; impulsivity: P=.19; SCC-self: P=.06; SCC-other: P=.08).

Discussion

The current in-person cognitive screening tools face challenges related to accessibility, scalability, and PD specificity. We aimed to develop and tailor a new sensitive tool capable of identifying cognitive impairments in early- to mid-stage PD in a manner that was both accessible and scalable. Accordingly, ORCA-PD was developed as a fully self-administered, web-based cognitive screening tool tailored to capture PD-specific cognitive impairments.

First, as reflected by high completion rates (93%) and participant ratings (4-5 out of 5), ORCA-PD showed high usability, further supporting its feasibility across diverse populations (more than 30 geographical locations). Second, ORCA-PD naive scores correlated with MoCA scores, supporting preliminary validity. While the NT estimate should be interpreted with caution due to the small sample size, the moderate correlation with MoCA suggests that ORCA-PD captures overlapping but not identical constructs, offering information beyond a generic cognitive screener. As with all screening tools designed to enable accessible and fast identification of cognitive impairment, ORCA-PD was not meant to replace comprehensive assessment tools but to prompt early diagnosis of PD with MCI. The moderate correlation with the MoCA supports ORCA-PD's role as a complementary tool, particularly when subtle, domain-specific changes are expected earlier in the PD course [52]. Third, ORCA-PD differentiated between groups, with performance following expected trends: NT>PD and PD-non-MCI>PD-MCI, confirming its discriminative capacity.

When examining previous literature, recent studies have begun to explore the potential of remote cognitive testing in PD. One study [28] demonstrated the feasibility and test-retest reliability of an unsupervised online cognitive battery in PD. However, the study did not assess validity, nor did it evaluate MCI using an additional screening measure. Another study [31] examined smartphone-based cognitive testing in daily contexts, but the sample size was small (n=27) and required a supervised laboratory session. A third study [27] demonstrated the feasibility of large-scale remote data collection in participants with PD using self-report questionnaires, but it lacked performance-based cognitive measures. A fourth pilot study [29] consisted of virtual neuropsychological testing using traditional batteries with a small sample size (n=35) and reliance on supervised administration. A fifth study [30] evaluated a lengthy 19-task online cognitive battery in individuals with PD and individuals with REM (rapid eye movement) sleep behavior disorder, identifying a post hoc 20-minute subset sensitive to group differences. However, the tasks relied heavily on reaction time, including motor-dependent tasks (ie, mouse use), which can confound cognitive measurement in PD. Additionally, the tasks were not chosen based on previous empirical independent findings and were not prospectively tested. In contrast, ORCA-PD was prospectively evaluated. The task selection was guided by prior meta-analyses and reviews that provided independent evidence for their sensitivity to PD. Additionally, the ORCA-PD tasks were specifically designed to reduce reliance on motor skills, thereby reducing the risk of motor-related confounds. This was also demonstrated by the lack of correlation between the ORCA-PD measures and the MDS-UPDRS III scores.

We developed and tailored ORCA-PD as a cognitive assessment for early- to mid-stage PD. In our sample, although inclusion criteria allowed participants with H&Y scores of 4 or lower, 98% had an H&Y score of 3 or lower (none had a score of 5; mean MDS-UPDRS III=21.9, mean disease duration=7.4 y). These findings are consistent with mild-to-moderate motor severity [55].

Of note, ORCA-PD was designed for remote, unsupervised administration, which likely contributed to the high rates of completion and usability observed, supporting its potential clinical applicability. Our sample did not include participants with dementia, as ORCA-PD was designed for early- to mid-stage PD, and self-administered testing may be less feasible in advanced stages. Taken together, validating ORCA-PD in this disease stage is advantageous because it targets the period when mild cognitive changes emerge.

Our study has several limitations. First, some measures (eg, SCC) showed limited sensitivity to between-group differences. Future studies should refine these tasks or replace them with more PD-sensitive alternatives, such as language-based measures. Second, although we assessed validity, we did not evaluate test-retest reliability. Third, the scoring of some tasks (eg, digit span) is relatively coarse, where a small number of trials per difficulty level may limit granularity. Yet, the

XSL•FO
RenderX

ORCA-PD design is typical of screening instruments, such as the MoCA and MMSE. The ORCA-PD, in fact, includes more trials than these widely used measures (4 vs 2 in digit span), which may provide greater sensitivity for detecting impairments. Our primary goal at this stage was to implement a simple approach that allows direct comparison across predefined difficulty levels, serving as an assessment of the expected performance pattern (ie, sanity check) of the online task design. In addition, we prioritized accuracy-based scoring over reaction time to minimize the influence of motor slowing, which may disproportionately affect reaction time measures in PD. While reaction time may provide additional sensitivity, it may also introduce motor confounds. Fourth, MoCA data were available for only a subset of participants, resulting in a smaller sample. Although this subsample exceeded the minimum sample size required by the power analysis, the small sample size may limit the generalizability of these findings. Fifth, the MoCA was used both as a measure for convergent validity and to operationalize MCI classification, introducing a degree of circularity that limits the independence of the validation and may inflate observed associations. Sixth, medication status at the time of testing was not controlled. A majority of participants (<70%) were treated with dopaminergic medication, and the median time since the last medication intake was 120 minutes. Although prior work [56,57] suggests that dopaminergic medication has limited effects on cognitive screening measures such as the MoCA, with more modest and domain-specific effects, variability in medication state may have introduced additional heterogeneity in performance. This study was not designed to examine ON versus OFF medication effects, and future studies should systematically evaluate the influence of medication state and timing of testing on ORCA-PD performance.

Seventh, ORCA-PD scores were not compared against a comprehensive neuropsychological battery, and the use of a single-instrument cutoff does not meet MDS level II criteria. Future research should validate ORCA-PD against independent, gold-standard neuropsychological assessments and include parallel forms to estimate test-retest reliability, while recognizing that ORCA-PD is designed primarily as a screening tool. Eighth, online-based participation may favor individuals who are more technologically adept or less cognitively impaired; therefore, future studies should directly compare online and in-person administrations to assess potential selection and usability biases.

Ninth, some NT participants were recruited through relatives or caregivers of individuals with PD, which may introduce selection bias. Individuals in close proximity to people with a PD diagnosis may differ from the general population in terms of health care engagement and motivation to participate in research. This may limit the external validity of the findings, as the control group may not be fully representative of the general population.

In addition, race, quality of education, area deprivation index, and related factors were not collected. This represents a substantive limitation, particularly given the goal of developing an accessible tool for diverse populations, and it restricts our ability to evaluate generalizability across demographic and socioeconomic contexts. This is an important goal for future studies.

Language-related effects were examined as a secondary exploratory analysis. While some differences were observed (eg, DSF), these effects were not consistent across tasks or groups. However, our study was not designed or powered to evaluate cross-language differences or establish language-specific norms. Additionally, we did not aim to validate ORCA-PD in different languages. The relatively small sample, along with important potential demographic and cultural differences between language groups, further limits interpretability. Future studies should be designed to establish language-specific norms and examine between-language differences. Finally, the cross-sectional design precluded longitudinal inferences, and future longitudinal studies are needed to examine ORCA-PD's sensitivity to cognitive change over time. Addressing these limitations will help maximize the tool's clinical and research utility.

Although we developed ORCA-PD for people with PD, ORCA's framework and principles can be expanded. Additional studies could focus on different populations, with an emphasis on varying levels of education and both rural and urban regions. Given the relatively high education level of our current sample, future validation in less-educated and lower-resource populations will be critical to ensure accessibility and generalizability. Socioeconomic information was not collected in this study, which limits our ability to assess the representativeness of our sample across social classes; future work should incorporate these measures to evaluate equity of access and usability for all. Furthermore, this platform need not be limited to PD. By selecting tasks matched to PD cognitive signatures and adjusting task difficulty and response modalities, our ORCA approach can be adapted to other conditions (eg, cerebellar ataxia). Such adaptations will require independent validation and condition-specific calibrations.

In summary, ORCA-PD is a brief, self-administered cognitive screening test for people with PD that demonstrates feasibility, utility, and preliminary validity that can be used in research and clinical settings upon further validation. By enabling unsupervised home testing, it has the potential to complement clinic-based screening and support between-visit cognitive monitoring. ORCA-PD could provide an accessible, scalable, and specific tool for cognitive screening in PD.

Acknowledgments

The authors thank all participants and research assistants who made this study possible. The research assistants at the Center for Accessible Neuropsychology provided dedicated support in data collection. The authors are grateful to the individuals living with Parkinson disease (PD) and their families for their generous participation and ongoing support. Their willingness to contribute,

XSL•FO
RenderX

along with their thoughtful and constructive feedback, has been instrumental in refining our research approach and developing tools dedicated to improving the well-being of the PD community.

Funding

This work was supported by the Chief Scientist's Office of the Israeli Ministry of Health for WS and ALA (#00370001000). The funders played no role in the study design, data collection, analysis, interpretation of data, or the writing of this manuscript.

Data Availability

The data and code are available upon reasonable request to the corresponding author. Access will be granted in compliance with ethical guidelines, given the patients' confidential information.

Authors' Contributions

WS contributed to the conception. SB contributed to data collection. ALA analyzed the results. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Yang W, Hamilton JL, Kopil C, et al. Current and projected future economic burden of Parkinson's disease in the U.S. NPJ Parkinsons Dis 2020;6(1):15. [doi: 10.1038/s41531-020-0117-1] [Medline: 32665974]
2. Kouli A, Torsney KM, Kuan WL. Parkinson's Disease: Etiology, Neuropathology, and Pathogenesis. In: Stoker TB, Greenland JC, editors. Parkinson's Disease: Pathogenesis and Clinical Aspects: Exon Publications; 2018:3-22. [doi: 10.15586/codonpublications.parkinsonsdisease.2018.ch1]
3. Jankovic J. Parkinson's disease: clinical features and diagnosis. J Neurol Neurosurg Psychiatry 2008 Apr;79(4):368-376. [doi: 10.1136/jnnp.2007.131045] [Medline: 18344392]
4. Aarsland D, Creese B, Politis M, et al. Cognitive decline in Parkinson disease. Nat Rev Neurol 2017 Apr;13(4):217-231. [doi: 10.1038/nrneurol.2017.27] [Medline: 28257128]
5. Binoy S, Lithwick Algon A, Ben Adiva Y, Montaser-Kouhsari L, Saban W. Online cognitive testing in Parkinson's disease: advantages and challenges. Front Neurol 2024;15:1363513. [doi: 10.3389/fneur.2024.1363513] [Medline: 38651103]
6. Saban W, Ivry RB. PONT: a protocol for online neuropsychological testing. J Cogn Neurosci 2021 Oct 1;33(11):2413-2425. [doi: 10.1162/jocn_a_01767] [Medline: 34347867]
7. Binoy S, Woody R, Ivry RB, Saban W. Feasibility and efficacy of online neuropsychological assessment. Sensors (Basel) 2023 May 29;23(11):5160. [doi: 10.3390/s23115160] [Medline: 37299887]
8. Moustafa AA, Chakravarthy S, Phillips JR, et al. Motor symptoms in Parkinson's disease: a unified framework. Neurosci Biobehav Rev 2016 Sep;68:727-740. [doi: 10.1016/j.neubiorev.2016.07.010] [Medline: 27422450]
9. Friedman JH, Brown RG, Comella C, et al. Fatigue in Parkinson's disease: a review. Mov Disord 2007 Feb 15;22(3):297-308. [doi: 10.1002/mds.21240] [Medline: 17133511]
10. Kostić VS, Tomić A, Ječmenica-Lukić M. The pathophysiology of fatigue in Parkinson's disease and its pragmatic management. Mov Disord Clin Pract 2016;3(4):323-330. [doi: 10.1002/mdc3.12343] [Medline: 30363584]
11. Kudlicka A, Clare L, Hindle JV. Executive functions in Parkinson's disease: systematic review and meta-analysis. Mov Disord 2011 Nov;26(13):2305-2315. [doi: 10.1002/mds.23868] [Medline: 21971697]
12. Ramos AA, Machado L. A comprehensive meta-analysis on short-term and working memory dysfunction in Parkinson's disease. Neuropsychol Rev 2021 Jun;31(2):288-311. [doi: 10.1007/s11065-021-09480-w] [Medline: 33523408]
13. Gullett JM, Price CC, Nguyen P, Okun MS, Bauer RM, Bowers D. Reliability of three Benton Judgment of Line Orientation short forms in idiopathic Parkinson's disease. Clin Neuropsychol 2013;27(7):1167-1178. [doi: 10.1080/13854046.2013.827744] [Medline: 23957375]
14. Marinus J, Zhu K, Marras C, Aarsland D, van Hilten JJ. Risk factors for non-motor symptoms in Parkinson's disease. Lancet Neurol 2018 Jun;17(6):559-568. [doi: 10.1016/S1474-4422(18)30127-3] [Medline: 29699914]
15. Hewitt KC, Rodgin S, Loring DW, Pritchard AE, Jacobson LA. Transitioning to telehealth neuropsychology service: considerations across adult and pediatric care settings. Clin Neuropsychol 2020;34(7-8):1335-1351. [doi: 10.1080/13854046.2020.1811891] [Medline: 32842849]
16. Schneider RB, Biglan KM. The promise of telemedicine for chronic neurological disorders: the example of Parkinson's disease. Lancet Neurol 2017 Jul;16(7):541-551. [doi: 10.1016/S1474-4422(17)30167-9] [Medline: 28566190]
17. Foltynie T, Brayne C, Barker RA. The heterogeneity of idiopathic Parkinson's disease. J Neurol 2002 Feb;249(2):138-145. [doi: 10.1007/pl00007856] [Medline: 11985378]

XSL•FO
RenderX

18. Nasreddine ZS, Phillips NA, Bédirian V, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005 Apr;53(4):695-699. [doi: 10.1111/j.1532-5415.2005.53221.x] [Medline: 15817019]
19. Folstein MF, Folstein SE, McHugh PR. “Mini-mental state”. A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975 Nov;12(3):189-198. [doi: 10.1016/0022-3956(75)90026-6] [Medline: 1202204]
20. Adjerid I, Kelley K. Big data in psychology: a framework for research advancement. Am Psychol 2018 Oct;73(7):899-917. [doi: 10.1037/amp0000190] [Medline: 29469579]
21. Casler K, Bickel L, Hackett E. Separate but equal? A comparison of participants and data gathered via Amazon’s MTurk, social media, and face-to-face behavioral testing. Comput Human Behav 2013 Nov;29(6):2156-2160. [doi: 10.1016/j.chb.2013.05.009]
22. Dorociak KE, Mattek N, Lee J, et al. The Survey for Memory, Attention, and Reaction Time (SMART): development and validation of a brief web-based measure of cognition for older adults. Gerontology 2021;67(6):740-752. [doi: 10.1159/000514871] [Medline: 33827088]
23. Kochan NA, Heffernan M, Valenzuela M, et al. Reliability, validity, and user-experience of remote unsupervised computerized neuropsychological assessments in community-living 55- to 75-year-olds. J Alzheimers Dis 2022;90(4):1629-1645. [doi: 10.3233/JAD-220665] [Medline: 36314208]
24. Myers JR, Glenn JM, Madero EN, et al. Asynchronous remote assessment for cognitive impairment: reliability verification of the Neurotrack Cognitive Battery. JMIR Form Res 2022 Feb 18;6(2):e34237. [doi: 10.2196/34237] [Medline: 35179511]
25. Stricker NH, Stricker JL, Karstens AJ, et al. A novel computer adaptive word list memory test optimized for remote assessment: psychometric properties and associations with neurodegenerative biomarkers in older women without dementia. Alzheimers Dement (Amst) 2022;14(1):e12299. [doi: 10.1002/dad2.12299] [Medline: 35280963]
26. Ruano L, Sousa A, Severo M, et al. Development of a self-administered web-based test for longitudinal cognitive assessment. Sci Rep 2016 Jan 8;6(1):19114. [doi: 10.1038/srep19114] [Medline: 26743329]
27. Smolensky L, Amondikar N, Crawford K, et al. Fox Insight collects online, longitudinal patient-reported outcomes and genetic data on Parkinson’s disease. Sci Data 2020 Feb 24;7(1):67. [doi: 10.1038/s41597-020-0401-2] [Medline: 32094335]
28. Balit N, Sun S, Zhang Y, Sharp M. Online unsupervised performance-based cognitive testing: a feasible and reliable approach to scalable cognitive phenotyping of Parkinson’s patients. Parkinsonism Relat Disord 2024 Dec;129:107183. [doi: 10.1016/j.parkreldis.2024.107183] [Medline: 39454300]
29. Gallagher J, Mamikonyan E, Xie SX, Tran B, Shaw S, Weintraub D. Validating virtual administration of neuropsychological testing in Parkinson disease: a pilot study. Sci Rep 2023 Sep 27;13(1):16243. [doi: 10.1038/s41598-023-42934-0] [Medline: 37758767]
30. Bălăeţ M, Alhajraf F, Zerenner T, et al. Online cognitive monitoring technology for people with Parkinson’s disease and REM sleep behavioural disorder. NPJ Digit Med 2024 May 7;7(1):118. [doi: 10.1038/s41746-024-01124-6] [Medline: 38714742]
31. Weizenbaum EL, Fulford D, Torous J, Pinsky E, Kolachalama VB, Cronin-Golomb A. Smartphone-based neuropsychological assessment in Parkinson’s disease: feasibility, validity, and contextually driven variability in cognition. J Int Neuropsychol Soc 2022 Apr;28(4):401-413. [doi: 10.1017/S1355617721000503] [Medline: 33998438]
32. Goldman JG, Holden S, Ouyang B, Bernard B, Goetz CG, Stebbins GT. Diagnosing PD-MCI by MDS Task Force criteria: how many and which neuropsychological tests? Mov Disord 2015 Mar;30(3):402-406. [doi: 10.1002/mds.26084] [Medline: 25449653]
33. Hoops S, Nazem S, Siderowf AD, et al. Validity of the MoCA and MMSE in the detection of MCI and dementia in Parkinson disease. Neurology 2009 Nov 24;73(21):1738-1745. [doi: 10.1212/WNL.0b013e3181c34b47] [Medline: 19933974]
34. Dalrymple-Alford JC, MacAskill MR, Nakas CT, et al. The MoCA: well-suited screen for cognitive impairment in Parkinson disease. Neurology 2010 Nov 9;75(19):1717-1725. [doi: 10.1212/WNL.0b013e3181fc29c9] [Medline: 21060094]
35. Hu MTM, Szewczyk-Królikowski K, Tomlinson P, et al. Predictors of cognitive impairment in an early stage Parkinson’s disease cohort. Mov Disord 2014 Mar;29(3):351-359. [doi: 10.1002/mds.25748] [Medline: 24395708]
36. Kandiah N, Zhang A, Cenina AR, Au WL, Nadkarni N, Tan LC. Montreal Cognitive Assessment for the screening and prediction of cognitive decline in early Parkinson’s disease. Parkinsonism Relat Disord 2014 Nov;20(11):1145-1148. [doi: 10.1016/j.parkreldis.2014.08.002] [Medline: 25176439]
37. Biundo R, Weis L, Antonini A. Cognitive decline in Parkinson’s disease: the complex picture. NPJ Parkinsons Dis 2016;2(1):16018. [doi: 10.1038/npjparkd.2016.18] [Medline: 28725699]
38. Loring DW, Lah JJ, Goldstein FC. Telehealth equivalence of the Montreal cognitive assessment (MoCA): results from the Emory healthy brain study (EHBS). J Am Geriatr Soc 2023 Jun;71(6):1931-1936. [doi: 10.1111/jgs.18271] [Medline: 36762513]
39. Picciotto YD, Lithwick Algon A, Amit I, Vakil E, Saban W. Large-scale evidence for the validity of remote MoCA administration among people with cerebellar ataxia. Clin Neuropsychol 2025 Aug;39(6):1634-1650. [doi: 10.1080/13854046.2024.2397835] [Medline: 39235357]

XSL•FO
RenderX

40. Binoy S, Montaser-Kouhsari L, Ponger P, Saban W. Remote assessment of cognition in Parkinson's disease and cerebellar ataxia: the MoCA test in English and Hebrew. Front Hum Neurosci 2023;17:1325215. [doi: 10.3389/fnhum.2023.1325215] [Medline: 38259338]
41. Gilad T, Algon AL, Vakil E, et al. Breaking geographic and language barriers in neuropsychology: online administration of MoCA in diverse populations. NPJ Digit Med 2025 Jun 20;8(1):379. [doi: 10.1038/s41746-025-01787-9] [Medline: 40537516]
42. Algon AL, Saban W. Characterizing motor-cognitive associations in Parkinson's disease using digital assessments. Front Aging Neurosci 2026;18:1787521. [doi: 10.3389/fnagi.2026.1787521] [Medline: 41809489]
43. Goetz CG, Tilley BC, Shaftman SR, et al. Movement Disorder Society–sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. Mov Disord 2008 Nov 15;23(15):2129-2170. [doi: 10.1002/mds.22340] [Medline: 19025984]
44. Giovannelli F, Gavazzi G, Noferini C, Palumbo P, Viggiano MP, Cincotta M. Impulsivity traits in Parkinson's disease: a systematic review and meta-analysis. Mov Disord Clin Pract 2023 Jul;10(10):1448-1458. [doi: 10.1002/mdc3.13839] [Medline: 37868926]
45. Siciliano M, Tessitore A, Morgante F, Goldman JG, Ricciardi L. Subjective cognitive complaints in Parkinson's disease: a systematic review and meta-analysis. Mov Disord 2024 Jan;39(1):17-28. [doi: 10.1002/mds.29649] [Medline: 38173220]
46. Grogan JP, Knight LE, Smith L, et al. Effects of Parkinson's disease and dopamine on digit span measures of working memory. Psychopharmacology (Berl) 2018 Dec;235(12):3443-3450. [doi: 10.1007/s00213-018-5058-6] [Medline: 30315362]
47. Benton AL, Varney NR, Hamsher KD. Judgment of line orientation [database record]. 2012 Jun 12. URL: https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft11036-000 [accessed 2026-06-02]
48. Hoogland J, van Wanrooij LL, Boel JA, et al. Detecting mild cognitive deficits in Parkinson's disease: comparison of neuropsychological tests. Mov Disord 2018 Nov;33(11):1750-1759. [doi: 10.1002/mds.110] [Medline: 30216541]
49. Patton JH, Stanford MS, Barratt ES. Factor structure of the Barratt impulsiveness scale. J Clin Psychol 1995 Nov;51(6):768-774. [doi: 10.1002/1097-4679(199511)51:6<768::aid-jclp2270510607>3.0.co;2-1] [Medline: 8778124]
50. Aumann MA, Stark AJ, Hughes SB, et al. Self-reported rates of impulsivity in Parkinson's disease. Ann Clin Transl Neurol 2020 Apr;7(4):437-448. [doi: 10.1002/acn3.51016] [Medline: 32227451]
51. Hong JY, Lee PH. Subjective cognitive complaints in cognitively normal patients with Parkinson's disease: a systematic review. J Mov Disord 2023 Jan;16(1):1-12. [doi: 10.14802/jmd.22059] [Medline: 36353806]
52. Litvan I, Goldman JG, Tröster AI, et al. Diagnostic criteria for mild cognitive impairment in Parkinson's disease: Movement Disorder Society Task Force guidelines. Mov Disord 2012 Mar;27(3):349-356. [doi: 10.1002/mds.24893] [Medline: 22275317]
53. RStudio. Posit. 2023. URL: https://posit.co/products/open-source/rstudio [accessed 2026-05-28]
54. Chou KL, Lenhart A, Koeppe RA, Bohnen NI. Abnormal MoCA and normal range MMSE scores in Parkinson disease without dementia: cognitive and neurochemical correlates. Parkinsonism Relat Disord 2014 Oct;20(10):1076-1080. [doi: 10.1016/j.parkreldis.2014.07.008] [Medline: 25085750]
55. Martínez-Martín P, Rodríguez-Blázquez C, et al. Parkinson's disease severity levels and MDS-Unified Parkinson's Disease Rating Scale. Parkinsonism Relat Disord 2015 Jan;21(1):50-54. [doi: 10.1016/j.parkreldis.2014.10.026] [Medline: 25466406]
56. Kübler D, Schroll H, Buchert R, Kühn AA. Cognitive performance correlates with the degree of dopaminergic degeneration in the associative part of the striatum in non-demented Parkinson's patients. J Neural Transm (Vienna) 2017 Sep;124(9):1073-1081. [doi: 10.1007/s00702-017-1747-2] [Medline: 28643101]
57. Xie T, Liao C, Bundy J, et al. Effect of dopaminergic medications on Montreal Cognitive Assessment in Parkinson's disease patients. Parkinsonism Relat Disord 2025 Sep;138:107951. [doi: 10.1016/j.parkreldis.2025.107951] [Medline: 40651129]

Abbreviations

ATS: alternate task switching
BIS-11: Barratt Impulsiveness Scale
CAN: Center for Accessible Neuropsychology
DSB: digit span backward
DSF: digit span forward
DSO: digit span order
H&Y: Hoehn and Yahr scale
LOM: line orientation match
MCI: mild cognitive impairment
MDS: Movement Disorder Society
MDS-UPDRS III: Movement Disorder Society—Unified Parkinson's Disease Rating Scale, Part III
MMSE: Mini-Mental State Examination
MoCA: Montreal Cognitive Assessment
NT: neurotypical controls

XSL•FO
RenderX

ORCA-PD: Online Rapid Cognitive Assessment for Parkinson's Disease
PD: Parkinson disease
SCC: subjective cognitive complaints

Edited by M Balcarras; submitted 23.Dec.2025; peer-reviewed by I Haq, KM Smith; revised version received 14.Apr.2026; accepted 16.Apr.2026; published 08.Jun.2026.

Please cite as:
Lithwick Algon A, Brisman S, Lin CYR, Saban W
Development and Preliminary Validation of ORCA-PD, an Online Rapid Cognitive Assessment for Parkinson Disease: Mixed Methods Study
J Med Internet Res 2026;28:e90057
URL: https://www.jmir.org/2026/1/e90057
doi:10.2196/90057

© Avigail Lithwick Algon, Sarah Brisman, Chi-Ying R Lin, William Saban. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Safety of Telemedicine Versus In-Person Care for Patients With Tracheal Devices: Propensity Score–Matched Cohort Study

Pedro Prosperi Desenzi Ciaralo, MD; Paulo Francisco Guerreiro Cardoso, MD, MSc, PhD; Helio Minamoto, MD, PhD; Benoit Jacques Bibas, MD, PhD; Carlos Roberto Ribeiro Carvalho, MD, PhD; Paulo Manuel Pego-Fernandes, MD, PhD

Instituto do Coracao do Hospital das Clinicas da Faculdade de Medicina da Universidade de Sao Paulo, Rua Dr. Eneas de Carvalho Aguiar 44, São Paulo, Brazil

Corresponding Author:
Pedro Prosperi Desenzi Ciaralo, MD
Instituto do Coracao do Hospital das Clinicas da Faculdade de Medicina da Universidade de Sao Paulo, Rua Dr. Eneas de Carvalho Aguiar 44, São Paulo, Brazil

Abstract

Background: Patients with tracheal diseases often require long-term follow-up after tracheal device placement, with a risk of adverse events that may lead to emergency care and unplanned interventions. Telemedicine has been proposed as an alternative to in-person follow-up to improve access and continuity of care.

Objective: The primary objective of this study was to compare the need for emergency department (ED) visits between telemedicine and in-person groups. Secondary objectives included comparing hospital readmissions, 30-day hospital readmissions, and unplanned interventions between groups.

Methods: This retrospective, single-institution study included adult patients with tracheal devices who underwent telemedicine and in-person outpatient clinic visits between 2020 and 2024. To balance the groups, we used 1:1 propensity score matching. We collected demographic and clinical data and evaluated the need for ED visits, hospital readmissions, 30-day hospital readmissions, and unplanned interventions. Kaplan-Meier estimation of time to first ED visit was performed to assess outcomes after outpatient visits.

Results: A total of 483 patients (n=277, 57% telemedicine and n=206, 43% in-person) underwent 2487 visits (1258 telemedicine and 1229 in-person). After propensity score matching, 336 patients remained (168 in each group). There were no significant differences in the need for ED visits, hospital readmissions, or unplanned interventions. The telemedicine group had significantly fewer 30-day hospital readmissions (odds ratio 0.38, 95% CI 0.16-0.87; *P*=.02). Kaplan-Meier analysis indicated no statistically significant difference in ED-free visits.

Conclusions: Telemedicine follow-up was associated with outcomes comparable to those of in-person follow-up in this cohort of adult patients with tracheal devices, with no evidence of an increased need for ED visits. In the matched analysis, telemedicine was associated with lower odds of 30-day hospital readmission.

(J Med Internet Res 2026;28:e95479) doi:10.2196/95479

KEYWORDS

telemedicine; digital health; electronic health records; tracheal diseases; airway stents; tracheostomy

Introduction

Patients with tracheal devices frequently experience complications [1-3]. Clinical manifestations such as dyspnea, stridor, dysphonia, and infection are common [4], often leading to recurrent emergency department (ED) visits, hospital readmissions, and unplanned interventions [5].

These patients require continuous surveillance to ensure long-term safety and device functionality [6-9]. Proper stoma care, inspection of device integrity, and counseling on daily nebulizer use are essential components of follow-up [10,11]. Due to the chronic nature of tracheal diseases and the high risk of device-related complications, these patients demand prolonged and structured outpatient follow-up to maintain airway patency and stability.

In response to the limited number of specialized centers dedicated to the management of tracheal disease in Brazil [12] and anticipating an increase in the number of patients with tracheal disease following the COVID-19 pandemic [13-16] led us to develop a telemedicine program (TeleTrachea) [17,18] to maintain the support for this patient population. The implementation of TeleTrachea required establishing and standardizing new health care protocols and providing structured education for patients and physicians.

XSL•FO
RenderX

Despite the increasing use of telemedicine across medical specialties [19,20], no studies have addressed the safety of telemedicine follow-up in patients with tracheal devices. This is an important knowledge gap in this vulnerable patient population that requires continuous surveillance to prevent potentially life-threatening complications.

The lack of evidence regarding the safety and effectiveness of telemedicine in this context motivated this study. The primary outcome was the need for ED visits, and the secondary outcomes included hospital readmissions, 30-day hospital readmissions, and unplanned interventions.

Methods

Study Protocol

This is a retrospective cohort study of adult patients wearing tracheal devices who were followed at the Trachea and Airway Outpatient Clinic of the Division of Thoracic Surgery, Hospital das Clinicas, University of Sao Paulo between August 2020 and August 2024. All patients were identified through routine in-person follow-up at the tracheal disease team's outpatient clinic. At the end of each face-to-face consultation, telemedicine follow-up was systematically offered by the attending physician. Patients who declined telemedicine were scheduled for continued in-person follow-up. For those who accepted telemedicine, the patient's telephone number and institutional registration number were forwarded to the support team of the telemedicine department at the Heart Institute of Hospital das Clinicas, University of Sao Paulo (InCor-HCFMUSP).

The support team contacted the patient by telephone to arrange the telemedicine appointment, including the consultation date and time. Before the visit, the team provided instructions on using the platform and conducted audio and video tests to reduce the likelihood of technical issues during the teleconsultation. After the telemedicine visit, the patient was again referred to the support team to coordinate follow-up care, including scheduling return visits, imaging studies, prescriptions, medical reports, and endoscopic procedures, as indicated.

Patients were categorized into the telemedicine or in-person group, and follow-up modality remained unchanged throughout the study period. Patients in the telemedicine group were followed exclusively through telemedicine, whereas those in the in-person group were followed exclusively through face-to-face visits. In the telemedicine group, in-person hospital attendance occurred only when necessary for elective surgery or endoscopic procedures, or complementary diagnostic testing, excluding visits related to study outcomes.

Telemedicine follow-up was systematically offered to all patients evaluated during the in-person visit. Ongoing in-person appointments were maintained for patients who reported limitations to telemedicine, such as a lack of internet access, the absence of a caregiver to assist during the virtual visit, or a clear preference for in-person follow-up. A telemedicine consultation was considered complete when there was uninterrupted synchronous audio and video interaction between the physician and patient.

Ethical Considerations

The study was approved by the research ethics committee for human subjects of InCor-HCFMUSP (protocol number 553422113). Given the retrospective design of the study and the use of data from electronic medical records, the ethics committee waived the requirement for informed consent.

Data were collected from the institution's electronic health records. Access to identifiable information was restricted to the principal investigator (PPDC), who extracted the data and provided the study team with a fully anonymized database. Deidentification was performed using each patient's institutional enrollment number in accordance with institutional ethical standards, ensuring the confidentiality, privacy, and security of patient information.

Patients who underwent telemedicine follow-up had previously signed an informed statement acknowledging and agreeing to this new model of care.

Eligibility Criteria

Inclusion criteria were presence of a tracheal device; aged ≥18 years; regular follow-up at the Trachea and Airway Outpatient Clinic of the Division of Thoracic Surgery, Hospital das Clinicas, University of Sao Paulo; signed informed statement acknowledging and agreeing to online consultation; and to participate in a telemedicine session with simultaneous audio and video.

Study Setting and Telemedicine Platform

Telemedicine consultations were conducted at the InCor-HCFMUSP. The facility has 14 soundproof teleconference rooms, each equipped with desktop computers with dual monitors and headsets with integrated microphones (Figure 1). The support team, located in an adjacent room, provided technical assistance in the event of connection disruptions or audiovisual difficulties (Figure 2) [17,18].

XSL•FO
RenderX

Figure 1. Telemedicine consultation room with desktop computers with cameras, headsets, and integrated microphones. Teleconsultations were conducted by residents and fellows under the supervision of a senior staff thoracic surgeon.

Figure 2. The telemedicine support team workspace is adjacent to the medical team's teleconsultation room.

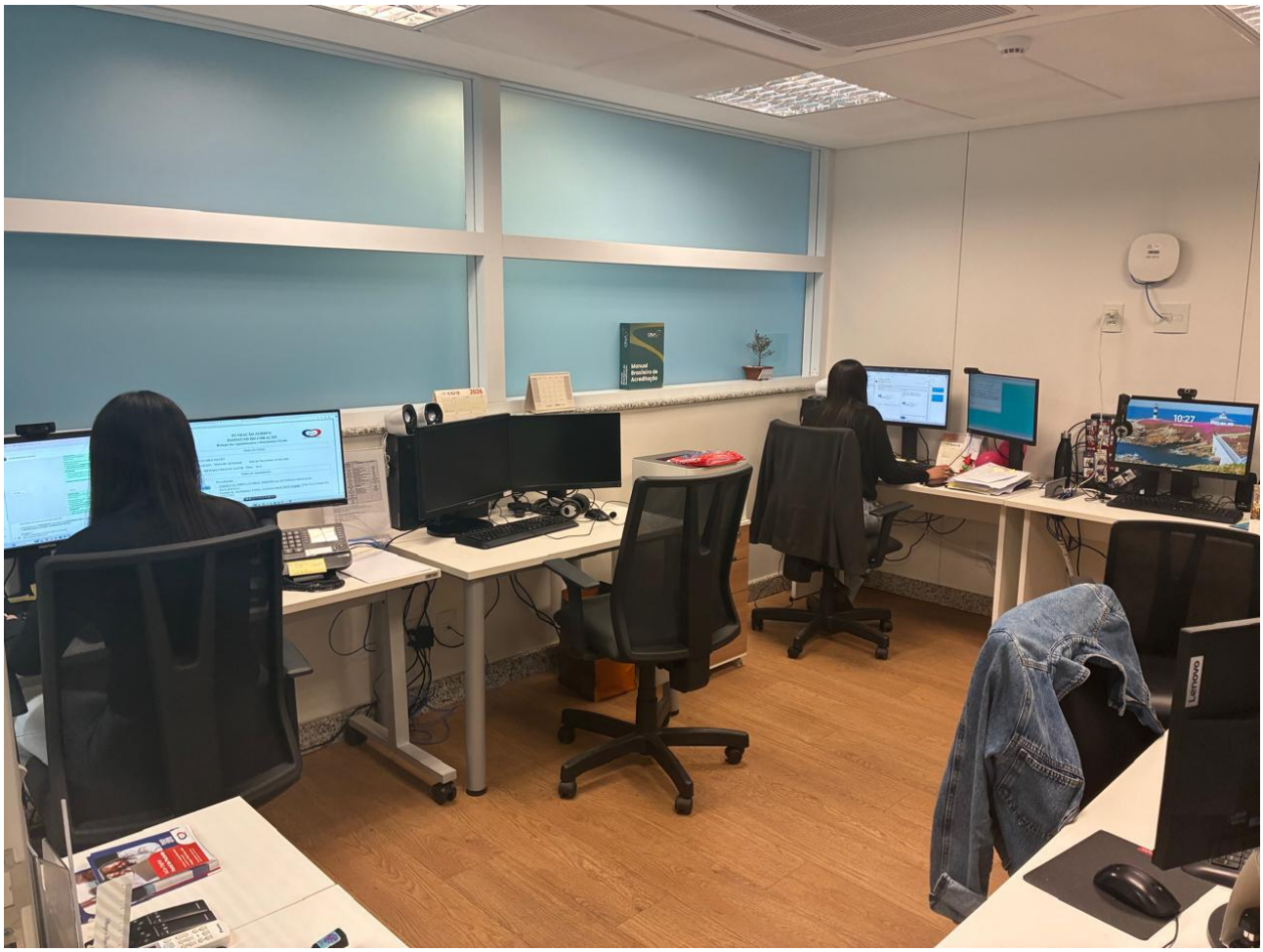

The online consultations were conducted by thoracic surgery residents and fellows under the on-site supervision of a senior

XSL•FO
RenderX

staff thoracic surgeon (PFGC and PPDC). The supervising surgeon reviewed and discussed each case and was available to interact directly with the patient and caregiver when necessary.

The teleconsultation process used the hospital's proprietary institutional teleconferencing platform (iConf, InCor-HCFMUSP; Figure 3). This system was designed to guarantee the secure storage and confidentiality of all service-related data within the institution with no access or manipulation by third parties [21,22]. Digital consultations were documented in the hospital's proprietary electronic medical record system (Si3, InCor-HCFMUSP). This integrated platform consolidates data from outpatient visits, hospital admissions, laboratory, and imaging results. It provides comprehensive access to recent and historical information supporting continuity and coordination of patient care.

Figure 3. The telemedicine institutional platform interface is used for teleconsultations.

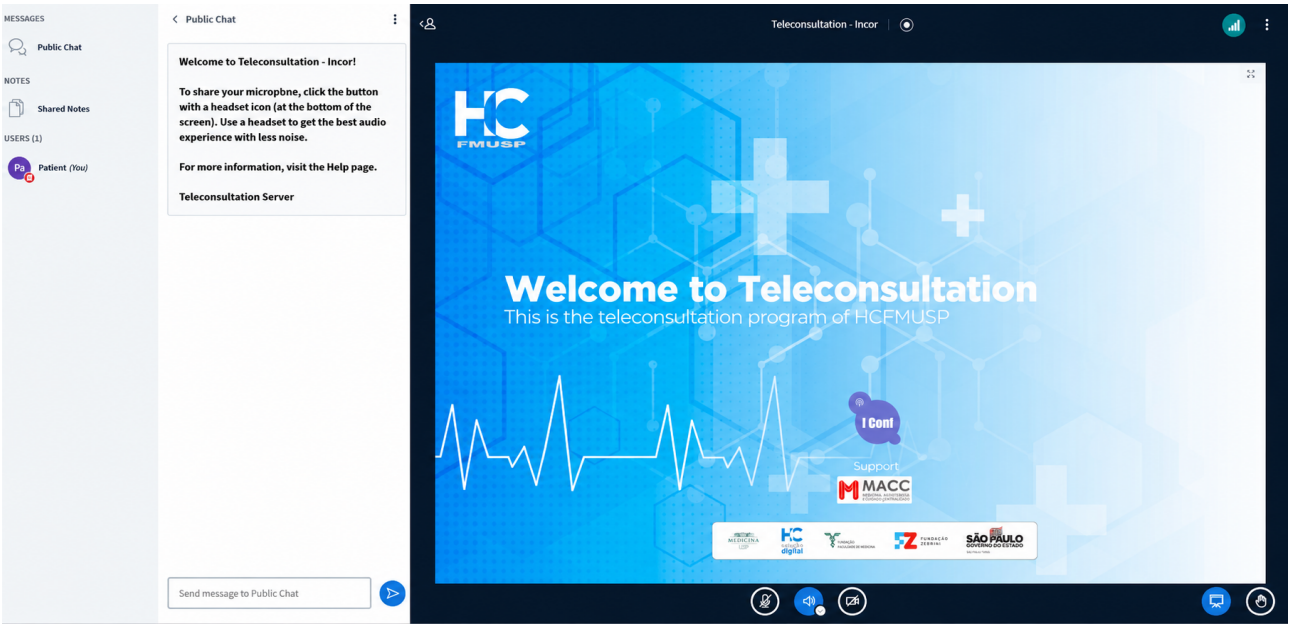

Definitions of Study Outcomes

Outcomes were identified through a review by the principal investigator to confirm eligibility for each outcome category. The electronic medical records, according to the predefined study definitions, ED visits, hospital readmissions, 30-day hospital readmissions, and unplanned interventions were classified according to their association with tracheal disease or device complications.

The primary outcome was whether patients visited the ED during the follow-up period. This was defined as any ED visit motivated by an airway-related symptom (dyspnea, stridor, or infection such as aspiration pneumonia or tracheitis), or for tracheal device–related issues (device fracture, accidental decannulation, loss of tracheostomy cuff pressure, or the need to keep the lateral limb of the T-tube open).

The secondary outcomes were (1) readmissions (defined as unplanned admissions for any clinical deterioration related to the tracheal device occurring during the study period; complications unrelated to the tracheal device were not considered); (2) 30-day readmissions (defined as unplanned admissions for any any clinical deterioration related to the tracheal device occurring within 30 days of the last outpatient visit; complications unrelated to the tracheal device were not considered); and (3) unplanned interventions (defined as any nonscheduled airway procedure during the study period, such as the need to replace a damaged tracheal device, device migration, obstruction, granulation tissue, or bleeding requiring bronchoscopy).

Statistical Analysis

Patient characteristics were summarized as absolute and relative frequencies (%) for categorical variables, and as mean with SD, and median with minimum and maximum values, or IQR for continuous variables. Categorical variables were examined by Pearson chi-square test or Fisher exact test, and continuous variables by the Mann-Whitney *U* test. For paired comparisons, McNemar test was used for categorical variables and the Wilcoxon signed-rank test for continuous variables.

Propensity score matching (PSM) balanced baseline characteristics using a binary logistic regression model with age, race, type of tracheal device, and diagnosis as covariates. The matched cohort was generated via 1:1 nearest-neighbor matching using a caliper of 0.05, and covariate balance before and after matching was assessed using standardized mean differences (Multimedia Appendix 1). Odds ratios with 95% CIs were estimated using logistic regression to compare the occurrence of ED visits, hospital readmissions, 30-day hospital readmissions, and unplanned interventions between groups after PSM.

For the PSM analyses, outcomes were treated as binary variables, defined by the occurrence of at least 1 event per patient during follow-up. The number of readmissions and unplanned interventions per patient was additionally summarized descriptively. For time-to-event analysis, Kaplan-Meier curves were used to estimate ED visit-free; only the first ED visit was considered and defined as the interval from the outpatient visit

XSL•FO
RenderX

to the ED visit. Curves were compared between groups using the log-rank test after PSM.

All statistical tests were 2-sided, with a significance level set at 5%. Analyses were performed using SPSS for Windows (version 25; IBM Corp).

Results

The number of patients is depicted in Figure 4. During the study, 483 patients were included (n=277, 57% telemedicine vs n=206, 43% in-person), with a total of 2487 visits (1258 telemedicine vs 1229 in-person). All patients had a tracheal device (246 tracheostomies, 192 silicone T-tubes, and 45 endoprostheses). The commonest diagnoses were postintubation tracheal stenosis (n=332, 69%), followed by neurological diseases (n=100, 21%), idiopathic stenosis (n=13, 3%), and tracheobronchial tuberculosis (n=12, 2.5%). The mean patient age was 46 (SD 16 years; range 18 - 92 years), and the Charlson comorbidity index score mean was 2 (SD 2).

Figure 4. Flowchart of patient selection and propensity score matching. PSM: propensity score matching.

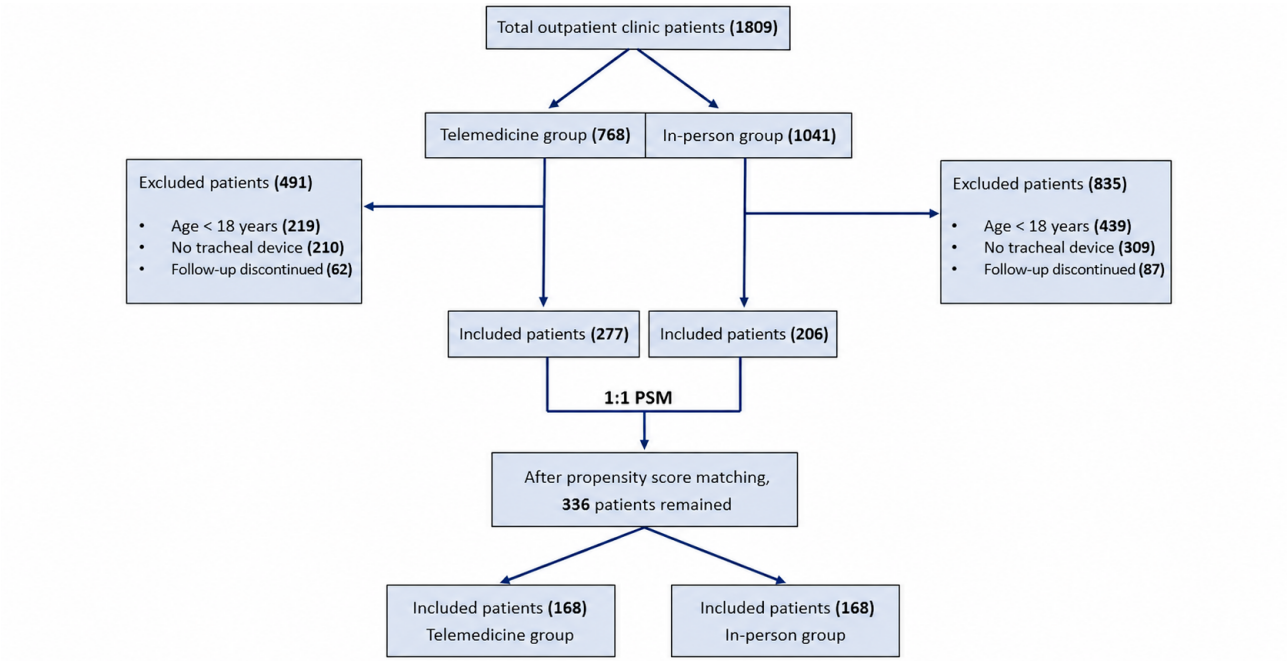

Baseline characteristics are summarized in Table 1. Before matching, the telemedicine group had a higher proportion of patients with silicone T-tubes, whereas tracheostomy was more frequent in the in-person group. Significant differences between the groups were also observed for race and diagnosis, while the remaining baseline variables were generally comparable.

XSL•FO
RenderX

Table . Baseline demographics and clinical characteristics of adult patients with tracheal devices before propensity score matching.

	Group		*P* value
	In person (n=206)	Telemedicine (n=277)	
Sex, n (%)			.92
Female	95 (46.1)	129 (46.6)	
Male	111 (53.9)	148 (53.4)	
Age (years), mean (SD)	48.1 (17.1)	45.5 (16.5)	.09
Race, n (%)			<.001
White	158 (76.7)	249 (89.9)	
Brown	25 (12.1)	18 (6.5)	
Black	23 (11.2)	10 (3.6)	
Education, n (%)			.15
College graduate	21 (11.6)	46 (17.6)	
High school graduate	63 (34.8)	88 (33.7)	
Less than high school	89 (49.2)	120 (46.0)	
No schooling	8 (4.4)	7 (2.7)	
Tracheal device, n (%)			<.001
Tracheostomy	132 (64.1)	114 (41.2)	
Silicone T-tube	58 (28.2)	134 (48.4)	
Endoprosthesis	16 (7.8)	29 (10.5)	
Diagnosis, n (%)			.04
Orotracheal intubation	129 (62.6)	203 (73.3)	
Neurologic diseases	57 (27.7)	43 (15.5)	
Idiopathic diseases	7 (3.4)	6 (2.2)	
Tuberculosis	4 (1.9)	8 (2.9)	
Neoplasm	4 (1.9)	6 (2.2)	
Rheumatologic diseases	3 (1.5)	5 (1.8)	
Others	2 (1.0)	6 (2.2)	
Charlson index, mean (SD)	2.0 (1.9)	2.0 (2.1)	.86
Charlson index (severity), n (%)			.81
Low risk	136 (66.0)	190 (68.6)	
Intermediate risk	44 (21.4)	53 (19.1)	
High risk	26 (12.6)	34 (12.3)	

A total of 140 patients (29%) required ED visits. Patients with tracheal endoprostheses had a higher rate of ED visits than patients with silicone T-tubes or tracheostomies. Patients with tracheobronchial tuberculosis sought ED more frequently when compared to other etiologies. There was no significant difference between the telemedicine and in-person groups (n=85, 30.7% vs n=55, 26.7%; *P*=.34; Table 2).

XSL•FO
RenderX

Table . Baseline demographics and clinical characteristics according to emergency department visits among adult patients with tracheal devices before propensity score matching.

	Emergency department visits		*P* value
	No (n=343)	Yes (n=140)	
Sex, n (%)			.10
Female	151 (67.4)	73 (32.6)	
Male	192 (74.1)	67 (25.9)	
Age (years), mean (SD)	48.1 (17.1)	45.5 (16.5)	.42
Race, n (%)			.11
White	286 (70.3)	121 (29.7)	
Brown	36 (83.7)	7 (16.3)	
Black	21 (63.6)	12 (36.4)	
Education, n (%)			.30
College graduate	43 (64.2)	24 (35.8)	
High school graduate	99 (65.6)	52 (34.4)	
Less than high school	151(72.2)	58(27.8)	
No schooling	13 (86.7)	2 (13.3)	
Tracheal device, n (%)			<.001
Tracheostomy	198 (80.5)	48 (19.5)	
Silicone T-tube	119 (62.0)	73 (38.0)	
Endoprosthesis	26 (57.8)	19 (42.2)	
Diagnosis, n (%)			.01
Orotracheal intubation	222 (66.9)	110 (33.1)	
Neurologic diseases	85 (85.0)	15 (15.0)	
Idiopathic diseases	10 (76.9)	3 (23.1)	
Tuberculosis	7 (58.3)	5 (41.7)	
Neoplasm	7 (70.0)	3 (30.0)	
Rheumatologic diseases	5 (62.5)	3 (37.5)	
Others	7 (87.5)	1 (12.5)	
Charlson index, mean (SD)	2.0 (1.9)	2.0 (2.1)	.95
Charlson index (severity), n (%)			.36
Low risk	234 (71.8)	92 (28.2)	
Intermediate risk	71 (73.2)	26 (26.8)	
High risk	38 (63.3)	22 (36.7)	
Group, n (%)			.34
In-person	151 (73.3)	55 (26.7)	
Telemedicine	192 (69.3)	85 (30.7)	

Regarding unplanned readmissions and interventions, there were no statistically significant differences between the groups. Conversely, the 30-day hospital readmissions were significantly lower in the telemedicine group (n=18, 21.2% vs n=27, 49.1%, *P*=.001; mean 0.2, SD 0.4 vs mean 0.7, SD 0.8, *P*≤.001; Table 3).

XSL•FO
RenderX

Table . Hospital readmission, 30-day hospital readmission, and unplanned interventions outcomes before propensity score matching in adult patients with tracheal devices.

	Group		*P* value
	In person (n=55)	Telemedicine (n=85)	
Readmission, n (%)			.25
No	23 (41.8)	44 (51.8)	
Yes	32 (58.2)	41 (48.2)	
Readmissions, mean (SD)	0.8 (0.8)	0.5 (0.6)	.10
30-day hospital readmission, n (%)			.001
No	28 (50.9)	67 (78.8)	
Yes	27 (49.1)	18 (21.2)	
30-day hospital readmissions, mean (SD)	0.7 (0.8)	0.2 (0.4)	<.001
Unplanned intervention, n (%)			.83
No	21 (38.2)	31 (36.5)	
Yes	34 (61.8)	54 (63.5)	
Unplanned interventions, mean (SD)	0.9 (0.9)	0.8 (0.9)	.66

After PSM, there were 168 patients per group (Table 4). The logistic regression analysis showed no significant differences between the groups in ED visits, hospital readmissions, or unplanned interventions. However, patients in the telemedicine group had significantly fewer 30-day hospital readmissions (odds ratio 0.38, 95% CI 0.16 - 0.87; *P*=.02; Table 5). The Kaplan-Meier analysis showed no statistically significant difference between the groups (*P*=.64), with no meaningful separation between the curves over time (Figure 5).

Table . Baseline demographics and clinical characteristics of adult patients with tracheal devices after propensity score matching.

	Group		*P* value
	In person (n=168)	Telemedicine (n=168)	
Age (years), mean (SD)	48.2 (17.2)	47.5 (17.0)	.67
Race, n (%)			.27
White	152 (90.5)	141 (83.9)	
Brown	9 (5.4)	17 (10.1)	
Black	7 (4.2)	10 (6.0)	
Tracheal device, n (%)			.58
Tracheostomy	103 (61.3)	90 (53.6)	
Silicone T-tube	49 (29.2)	60 (35.7)	
Endoprosthesis	16 (9.5)	18 (10.7)	
Diagnosis, n (%)			.62
Orotracheal intubation	105 (62.5)	113 (67.3)	
Neurologic diseases	49 (29.2)	35 (20.8)	
Tuberculosis	4 (2.4)	5 (3.0)	
Idiopathic diseases	5 (3.0)	3 (1.8)	
Neoplasm	2 (1.2)	5 (3.0)	
Rheumatologic diseases	2 (1.2)	3 (1.8)	
Others	1 (0.6)	4 (2.4)	

XSL•FO
RenderX

Table . Logistic regression analysis of clinical outcomes after propensity score matching in adult patients with tracheal devices followed by telemedicine or in-person care.

Variables	OR[a] (95% CI)	*P* value
Emergency department visit	1.19 (0.74 - 1.90)	.47
Readmission	0.70 (0.31 - 1.54)	.37
30-day hospital readmission	0.38 (0.16 - 0.87)	.02
Unplanned intervention	1.73 (0.74 - 4.04)	.20

[a]OR: odds ratio.

Figure 5. Kaplan-Meier curve for time to first emergency department (ED) visit after propensity score matching.

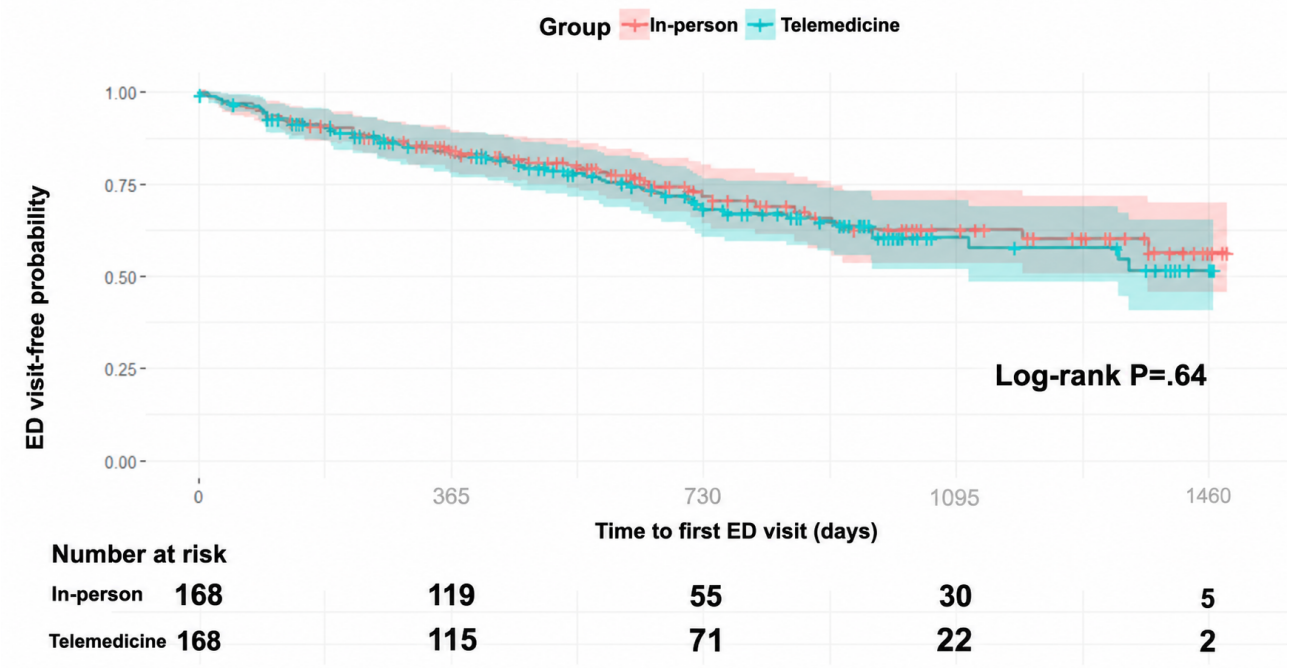

Discussion

Principal Findings

This single-center, propensity score–matched cohort of adult patients with tracheal devices demonstrated that telemedicine follow-up yielded safety outcomes comparable to those of in-person care. Following PSM, there were no significant differences between groups for ED visits, hospital readmissions, or unplanned interventions. Thirty-day hospital readmissions were significantly lower in the telemedicine group. The present findings suggest that a structured telemedicine program may support the longitudinal management of patients with tracheal devices without increasing ED visits. In the matched analysis, telemedicine follow-up was also associated with lower odds of 30-day hospital readmission.

Building on these findings, we must situate our results within the well-documented burden of tracheal device–related complications on patients and health systems [23]. Complications associated with tracheal devices result in substantial emergency care costs. In a recent analysis, tracheostomy-related complications accounted for 6156 ED visits and 2316 readmissions in a single year, with median inpatient charges of US $66,416 [24]. Our study showed no significant differences between the telemedicine and in-person groups in ED visits, but lower 30-day readmissions in the telemedicine group. These findings suggest that adequate remote surveillance may help prevent clinical deterioration and reduce downstream costs.

Studies on follow-up via telemedicine demonstrate that the quality of care is not inferior, regardless of the ED visits and readmissions [25-28]. Telemedicine in thoracic surgery cohorts has been associated with favorable outcomes [29-31]. The study by Xiao et al [31] reported that after esophagectomy, telemedicine follow-up resulted in fewer ED visits than in-person follow-up (mean 0.45, SD 0.76 vs 0.79, SD 1.12 within 6 months; *P*=.03), supporting the potential of telemedicine.

During the telemedicine visit, patients wearing tracheal devices require a visual inspection, as well as an audio capability that may prevent real-time communication. This context can undermine the quality of telemedicine by increasing the risk of misdiagnosis, requiring an ED visit. In this regard, our program offered synchronous audio-video consultations to enable patient and caregiver engagement, standardized checklists for device assessment, and provided technical support for transmission issues. So far, we have not observed worse outcomes, as shown by the absence of significant differences in ED visits, unplanned readmissions, and interventions.

XSL•FO
RenderX

Hospital readmissions affect patients' survival and quality of life, strain the health care resources, and serve as a benchmark of hospital performance [32-35]. Across health care payers, the all-cause 30-day readmission rate is 14% [36]. Although some readmissions are unavoidable, many are unplanned, and approximately 30% are preventable [37]. Therefore, reducing readmissions has become a key objective [38] to lower costs and improve the quality of care [39]. Recent studies have demonstrated a reduction in the risk of 30-day readmissions for patients who underwent telemedicine follow-up [40-42].

Unlike prior studies focused primarily on the immediate postdischarge period, our analysis focused on the outpatient follow-up setting. We hypothesized that telemedicine for outpatient follow-up in patients with tracheal devices may exert distinct and durable effects on safety and device use.

An important point for readmission in our patients is that visits exert a significant impact immediately after the consultation, when preventable failures such as early diagnosis of infection, medication or equipment mismanagement, and caregiver uncertainty are most likely to occur. However, beyond 30 days, the readmissions are mainly driven by the tracheal disease trajectory and factors related to device deterioration. Our study showed a reduction in 30-day readmissions with telemedicine follow-up, and no significant differences were found after this period. However, this finding should be interpreted with caution. The retrospective design does not allow causal inference. Unmeasured confounders such as differences in patient engagement, caregiver availability during follow-up, or clinician triage behavior may have partially contributed to the lower 30-day readmission rate observed in the telemedicine group, although this finding remains clinically relevant.

Other features of in-person follow-up may also influence the outpatient experience, including travel to the medical center, long waiting times [43,44], and the dynamics of a busy academic environment. In contrast, telemedicine consultations in our program were delivered through synchronous audio-video encounters within a structured workflow that included previsit technical support and postvisit review by the telemedicine team, as previously reported [17,18]. These factors may have contributed to continuity of care and reinforcement of recommendations during the follow-up period. In addition, prior studies have suggested that receiving care in the home environment may improve patient comfort and reduce stress during telemedicine visits [45,46]. Although these factors were not directly measured in our study, they may also represent plausible contributors to the lower 30-day readmission rate observed in the telemedicine group.

This study has several limitations. It is performed at a single high-volume center for tracheal diseases with a dedicated airway surgery team, which can limit external generalizability. Despite the planned data collection, the study analyses are retrospective, and even after PSM, the groups can still exhibit selection bias.

Although telemedicine was systematically offered, follow-up modality was not assigned at random. Patients who remained in the in-person group often did so because of limited digital access, lack of caregiver support, or personal preference, factors that may reflect broader differences in social support, health engagement, and adherence to follow-up. These variables may have influenced outcomes even after PSM.

During the COVID-19 pandemic, our academic hospital actively promoted telemedicine by investing in a high-complexity infrastructure and comprehensive staff training to ensure successful implementation and its continuity after the pandemic. Our findings reflect a mature, well-resourced perennial program, which may also limit generalizability to settings without comparable resources or experience. Nevertheless, certain elements of the model are likely essential regardless of setting, including a structured consultation environment with adequate computers and reliable audio-video equipment, technical team support for both patients and clinicians, and continued training in digital care.

Beyond the direct comparison of outcomes, our findings suggest a broader role for structured telemedicine in the longitudinal management of patients with tracheal devices. As these patients often require ongoing specialized follow-up and are frequently concentrated in referral centers, telemedicine may help reduce geographic and logistical barriers to care while maintaining continuity of monitoring. Although these results should be interpreted within the limits of a retrospective single-center study, they support the integration of structured telemedicine pathways into the care of selected patients with tracheal devices and provide a rationale for future prospective evaluation in other settings.

Conclusions

In this cohort of adult patients with tracheal devices, telemedicine outpatient follow-up was associated with clinical outcomes comparable to those of in-person follow-up. Telemedicine follow-up showed similar ED visit rates and was associated with lower odds of 30-day hospital readmission in the matched analysis.

Acknowledgments

No generative artificial intelligence tools were used in the writing, editing, analysis, or preparation of this manuscript. All content was developed exclusively by the authors.

Funding

The authors declared no financial support was received for this work.

XSL•FO
RenderX

Data Availability

Individual-level data from this study are not publicly available because they contain potentially identifiable and sensitive patient information. Deidentified aggregate-level data supporting the findings of this study may be made available by the corresponding author upon reasonable request, subject to submission and approval of a research proposal.

Anonymized data supporting the findings of this study may be shared by the investigative team under the following conditions: (1) collaboration with the study team in any resulting publications, (2) demonstration that the external investigative team is appropriately qualified and has documented training in human subjects protection, and (3) execution of and compliance with applicable interinstitutional data use agreements.

Authors' Contributions

Conceptualization: PPDC, PFGC, HM, CRRC
Data curation: PPDC
Formal analysis: PPDC, PFGC
Investigation: PPDC, PFGC, HM, BJB
Methodology: PPDC, PFGC
Project administration: PPDC and PFGC
Resources: CRRC, PMP-F
Supervision: CRRC, PMP-F
Validation: CRRC, PMP-F
Writing – original draft: PPDC
Writing – review and editing: PFGC, HM, BJB

Conflicts of Interest

None declared.

Multimedia Appendix 1
Absolute standardized mean differences before and after propensity score matching.
[DOCX File, 28 KB - jmir_v28i1e95479_app1.docx]

References

1. Salati M, Bibas BJ. Databases in tracheal diseases. Transl Cancer Res 2020 Mar;9(3):2149-2153. [doi: 10.21037/tcr.2020.02.32] [Medline: 35117571]
2. Terra RM, Bibas BJ, Minamoto H, et al. Decannulation in tracheal stenosis deemed inoperable is possible after long-term airway stenting. Ann Thorac Surg 2013 Feb;95(2):440-444. [doi: 10.1016/j.athoracsur.2012.09.037] [Medline: 23201102]
3. Rodrigues Cremonese M, Bibas BJ, Minamoto H, Pêgo-Fernandes PM, Guerreiro Cardoso PF. Hybrid desobstruction of blind-end subglottic stenosis with long-term stenting. Ann Thorac Surg 2021 Dec;112(6):e393-e395. [doi: 10.1016/j.athoracsur.2021.03.027] [Medline: 33753061]
4. Bianchi ET, Guerreiro Cardoso PF, Minamoto H, et al. Impact of fundoplication for gastroesophageal reflux in the outcome of benign tracheal stenosis. J Thorac Cardiovasc Surg 2019 Dec;158(6):1698-1706. [doi: 10.1016/j.jtcvs.2019.07.111] [Medline: 31587887]
5. Chhajed PN, Eberhardt R, Dienemann H, et al. Therapeutic bronchoscopy interventions before surgical resection of lung cancer. Ann Thorac Surg 2006 May;81(5):1839-1843. [doi: 10.1016/j.athoracsur.2005.11.054] [Medline: 16631682]
6. McGrath BA, Wallace S, Bonvento B, et al. Guidance for: tracheostomy care. : Intensive Care Society; 2023 URL: https://www.ficm.ac.uk/sites/ficm/files/documents/2021-11/2020-08%20Tracheostomy_care_guidance_Final.pdf [accessed 2026-05-13]
7. Brenner MJ, Pandian V, Milliren CE, et al. Global Tracheostomy Collaborative: data-driven improvements in patient safety through multidisciplinary teamwork, standardisation, education, and patient partnership. Br J Anaesth 2020 Jul;125(1):e104-e118. [doi: 10.1016/j.bja.2020.04.054] [Medline: 32456776]
8. Raimonde AJ, Gaston S, Wang CF. Tracheostomy. In: StatPearls: StatPearls Publishing; 2025.
9. Bonvento B, Wallace S, Lynch J, Coe B, McGrath BA. Role of the multidisciplinary team in the care of the tracheostomy patient. J Multidiscip Healthc 2017;10:391-398. [doi: 10.2147/JMDH.S118419] [Medline: 29066907]
10. Lewarski JS. Long-term care of the patient with a tracheostomy. Respir Care 2005 Apr;50(4):534-537. [Medline: 15807917]
11. Oberwaldner B, Eber E. Tracheostomy care in the home. Paediatr Respir Rev 2006 Sep;7(3):185-190. [doi: 10.1016/j.prrv.2006.06.003] [Medline: 16938640]
12. Bibas BJ, Minamoto H, Cardoso PF, Cremonese MR, Pêgo-Fernandes PM, Terra RM. Brazilian airway surgery survey indicates low overall numbers and need for improved teaching skills. Interdiscip Cardiovasc Thorac Surg 2024 Feb 2;38(2):ivad177. [doi: 10.1093/icvts/ivad177] [Medline: 37950507]

XSL•FO
RenderX

13. Piazza C, Filauro M, Dikkers FG, et al. Long-term intubation and high rate of tracheostomy in COVID-19 patients might determine an unprecedented increase of airway stenoses: a call to action from the European Laryngological Society. Eur Arch Otorhinolaryngol 2021 Jan;278(1):1-7. [doi: 10.1007/s00405-020-06112-6] [Medline: 32506145]
14. Potere N, Valeriani E, Candeloro M, et al. Acute complications and mortality in hospitalized patients with coronavirus disease 2019: a systematic review and meta-analysis. Crit Care 2020 Jul 2;24(1):389. [doi: 10.1186/s13054-020-03022-1] [Medline: 32616077]
15. Singer AJ, Morley EJ, Meyers K, et al. Cohort of four thousand four hundred four persons under investigation for COVID-19 in a New York hospital and predictors of ICU care and ventilation. Ann Emerg Med 2020 Oct;76(4):394-404. [doi: 10.1016/j.annemergmed.2020.05.011] [Medline: 32563601]
16. Harrell Shreckengost CS, Foianini JE, Moron Encinas KM, et al. Outcomes of early versus late tracheostomy in patients with COVID-19: a multinational cohort study. Crit Care Explor 2022 Nov;4(11):e0796. [doi: 10.1097/CCE.0000000000000796] [Medline: 36440062]
17. Prosperi Desenzi Ciaralo P, Guerreiro Cardoso PF, Minamoto H, Bibas BJ, Ribeiro de Carvalho CR, Pego-Fernandes PM. Implementation and results of a dedicated telemedicine program (TeleTrachea) for patients with tracheal diseases. Telemed J E Health 2024 May;30(5):1317-1324. [doi: 10.1089/tmj.2023.0524] [Medline: 38109228]
18. Desenzi Ciaralo PP, Guerreiro Cardoso PF, Minamoto H, Bibas BJ, Ribeiro de Carvalho CR, Pego-Fernandes PM. Estimated carbon emissions and support cost savings to telemedicine for patients with tracheal diseases. Int J Med Inform 2025 Mar;195:105757. [doi: 10.1016/j.ijmedinf.2024.105757] [Medline: 39671853]
19. Ye J, He L, Beestrum M. Implications for implementation and adoption of telehealth in developing countries: a systematic review of China's practices and experiences. NPJ Digit Med 2023 Sep 18;6(1):174. [doi: 10.1038/s41746-023-00908-6] [Medline: 37723237]
20. Scott Kruse C, Karem P, Shifflett K, Vegi L, Ravi K, Brooks M. Evaluating barriers to adopting telemedicine worldwide: a systematic review. J Telemed Telecare 2018 Jan;24(1):4-12. [doi: 10.1177/1357633X16674087] [Medline: 29320966]
21. Scudeller PG, Pereira AJ, Cerri GG, et al. Telemedicine in Brazil: teleconsultations at the largest university hospital in the country. Telemed Rep 2023;4(1):193-203. [doi: 10.1089/tmr.2023.0012] [Medline: 37529769]
22. Macedo BR, Garcia MV, Garcia ML, et al. Implementation of tele-ICU during the COVID-19 pandemic. J Bras Pneumol 2021;47(2):e20200545. [doi: 10.36416/1806-3756/e20200545] [Medline: 33950091]
23. Spataro E, Durakovic N, Kallogjeri D, Nussenbaum B. Complications and 30-day hospital readmission rates of patients undergoing tracheostomy: a prospective analysis. Laryngoscope 2017 Dec;127(12):2746-2753. [doi: 10.1002/lary.26668] [Medline: 28543108]
24. Bibas BJ, Cardoso PFG, Hoetzenecker K. The burden of tracheal stenosis and tracheal diseases health-care costs in the 21st century. Transl Cancer Res 2020 Mar;9(3):2095-2096. [doi: 10.21037/tcr.2020.02.59] [Medline: 35117562]
25. Daliya P, Carvell J, Rozentals J, Lobo DN, Parsons SL. Digital follow-up after elective laparoscopic cholecystectomy: a feasibility study. World J Surg 2022 Nov;46(11):2648-2658. [doi: 10.1007/s00268-022-06684-w] [Medline: 35953737]
26. Uppal A, Kothari AN, Scally CP, et al. Adoption of telemedicine for postoperative follow-up after inpatient cancer-related surgery. JCO Oncol Pract 2022 Jul;18(7):e1091-e1099. [doi: 10.1200/OP.21.00819] [Medline: 35263166]
27. Cremades M, Ferret G, Parés D, et al. Telemedicine to follow patients in a general surgery department. A randomized controlled trial. Am J Surg 2020 Jun;219(6):882-887. [doi: 10.1016/j.amjsurg.2020.03.023] [Medline: 32252983]
28. Grygorian A, Montano D, Shojaa M, Ferencak M, Schmitz N. Digital health interventions and patient safety in abdominal surgery: a systematic review and meta-analysis. JAMA Netw Open 2024 Apr 1;7(4):e248555. [doi: 10.1001/jamanetworkopen.2024.8555] [Medline: 38669018]
29. Cerfolio RJ, Ferrari-Light D, Shah S. Telemedicine in thoracic surgery. J Vis Surg 2019;5:54. [doi: 10.21037/jovs.2019.04.04]
30. Harrington CA, Hsu M, Tan KS, et al. Translating telemedicine to thoracic surgical oncological care: performance analysis and patient perceptions during the COVID-19 pandemic. Ann Surg 2023 Jul 1;278(1):e179-e183. [doi: 10.1097/SLA.0000000000005525] [Medline: 35786673]
31. Xiao K, Bolger JC, Allison F, Darling GE, Yeung JC. Impact of telehealth postoperative care on early outcomes following esophagectomy. J Thorac Cardiovasc Surg 2025 Feb;169(2):677-682. [doi: 10.1016/j.jtcvs.2024.02.005] [Medline: 38340955]
32. Pattar BS, Ackroyd A, Sevinc E, et al. Electronic health record interventions to reduce risk of hospital readmissions: a systematic review and meta-analysis. JAMA Netw Open 2025 Jul 1;8(7):e2521785. [doi: 10.1001/jamanetworkopen.2025.21785] [Medline: 40674049]
33. Zogg CK, Pawlik TM, Haider AH. Three common methodological issues in studies of surgical readmission rates: the trouble with readmissions. JAMA Surg 2018 Dec 1;153(12):1074-1076. [doi: 10.1001/jamasurg.2018.2488] [Medline: 30193351]
34. Zogg CK, Olufajo OA, Jiang W, et al. The need to consider longer-term outcomes of care: racial/ethnic disparities among adult and older adult emergency general surgery patients at 30, 90, and 180 days. Ann Surg 2017 Jul;266(1):66-75. [doi: 10.1097/SLA.0000000000001932] [Medline: 28140382]
35. Krumholz HM, Wang K, Normand SL. Hospital-readmission risk - isolating hospital effects. N Engl J Med 2017 Dec 21;377(25):2505. [doi: 10.1056/NEJMc1713509] [Medline: 29262282]
36. Jiang J, Hensche M. Characteristics of 30-day all-cause hospital readmissions, 2016-2020. Agency for Healthcare Research and Quality. 2023 URL: https://hcup-us.ahrq.gov/reports/statbriefs/sb304-readmissions-2016-2020.pdf [accessed 2026-05-19]

XSL•FO
RenderX

37. van Walraven C, Bennett C, Jennings A, Austin PC, Forster AJ. Proportion of hospital readmissions deemed avoidable: a systematic review. CMAJ 2011 Apr 19;183(7):E391-E402. [doi: 10.1503/cmaj.101860] [Medline: 21444623]
38. Joynt KE, Jha AK. A path forward on Medicare readmissions. N Engl J Med 2013 Mar 28;368(13):1175-1177. [doi: 10.1056/NEJMp1300122] [Medline: 23465069]
39. Kripalani S, Theobald CN, Anctil B, Vasilevskis EE. Reducing hospital readmission rates: current strategies and future directions. Annu Rev Med 2014;65:471-485. [doi: 10.1146/annurev-med-022613-090415] [Medline: 24160939]
40. Horman SF, Kviatkovsky M, Castillo E, et al. Virtual transition of care clinics and associated readmission rates: 3-year retrospective cohort study. JMIR Med Inform 2025 Sep 23;13:e73495. [doi: 10.2196/73495] [Medline: 40987450]
41. Kim EJ, Coppa K, Abrahams S, et al. Utilization of transitional care management services and 30-day readmission. PLoS One 2025;20(1):e0316892. [doi: 10.1371/journal.pone.0316892] [Medline: 39752456]
42. Shen E, Koyama SY, Huynh DN, et al. Association of a dedicated post-hospital discharge follow-up visit and 30-day readmission risk in a Medicare advantage population. JAMA Intern Med 2017 Jan 1;177(1):132-135. [doi: 10.1001/jamainternmed.2016.7061] [Medline: 27893040]
43. Goodwin JS, Hunt WC, Samet JM. Determinants of cancer therapy in elderly patients. Cancer 1993 Jul 15;72(2):594-601. [doi: 10.1002/1097-0142(19930715)72:2<594::aid-cncr2820720243>3.0.co;2-#] [Medline: 8319193]
44. Ambroggi M, Biasini C, Del Giovane C, Fornari F, Cavanna L. Distance as a barrier to cancer diagnosis and treatment: review of the literature. Oncologist 2015 Dec;20(12):1378-1385. [doi: 10.1634/theoncologist.2015-0110] [Medline: 26512045]
45. Gajarawala SN, Pelkowski JN. Telehealth benefits and barriers. J Nurse Pract 2021 Feb;17(2):218-221. [doi: 10.1016/j.nurpra.2020.09.013] [Medline: 33106751]
46. Mukamal RC, Gontijo Augusto V, Moraes Dias L, Dias Sarti T, Rego G. Telemedicine in the care of older adults with dementia: caregivers' perceptions and experiences. Geriatrics (Basel) 2025 Dec 17;10(6):169. [doi: 10.3390/geriatrics10060169] [Medline: 41440739]

Abbreviations

ED: emergency department
InCor-HCFMUSP : Heart Institute of Hospital das Clinicas, University of Sao Paulo
PSM: propensity score matching

Edited by A Mavragani; submitted 16.Mar.2026; peer-reviewed by L Ilodigwe, S Venkatesan; revised version received 24.Apr.2026; accepted 30.Apr.2026; published 26.May.2026.

Please cite as:
Ciaralo PPD, Cardoso PFG, Minamoto H, Bibas BJ, Carvalho CRR, Pego-Fernandes PM
Safety of Telemedicine Versus In-Person Care for Patients With Tracheal Devices: Propensity Score–Matched Cohort Study
J Med Internet Res 2026;28:e95479
URL: https://www.jmir.org/2026/1/e95479
doi:10.2196/95479

© Pedro Ciaralo, Paulo Francisco Guerreiro Cardoso, Helio Minamoto, Benoit Jacques Bibas, Carlos Roberto Ribeiro Carvalho, Paulo Manuel Pego-Fernandes. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Monitoring of Pre-Exposure Prophylaxis Users Through Electronic Patient-Reported Outcome Measures and Electronic Patient-Reported Experience Measures: Multicenter Prospective Study on Feasibility, Safety, and Predictive Modeling of Digital Engagement

Gabriel Mercadal-Orfila[1]; Joaquín Ignacio Serrano[2]; Tilman Mijares[1]; Lluís Vidal[1]; Adrian Curran[1]; Salvador Herrera-Pérez[3]

[1]Hospital General Mateu Orfila, Mahon, Balearic Islands, Spain

[2]Hospital Son Llatzer, Mahon, Balearic Islands, Spain

[3]Valencian International University, C. del Pintor Sorolla, 21, Ciutat Vella, Valencia, Spain

Corresponding Author:
Salvador Herrera-Pérez
Valencian International University, C. del Pintor Sorolla, 21, Ciutat Vella, Valencia, Spain

Abstract

Background: Digital health tools integrating electronic patient-reported outcome and experience measures (ePROMs/ePREMs) enable longitudinal monitoring of health-related quality of life (HRQoL), psychological well-being, and treatment satisfaction in pre-exposure prophylaxis (PrEP) users. However, determinants of sustained engagement with digital follow-up platforms remain insufficiently characterized.

Objective: To describe the feasibility of the Naveta-Phemium digital platform for longitudinal monitoring of safety, HRQoL, and treatment satisfaction among PrEP users, and to develop and internally validate a machine learning framework to characterize and predict engagement with digital follow-up.

Methods: A prospective observational study was conducted using the Naveta digital follow-up platform. HIV-negative adults at high risk of HIV infection received tenofovir disoproxil fumarate (245 mg) plus emtricitabine (200 mg). Clinical safety, HRQoL, and satisfaction were assessed using laboratory parameters and validated ePROMs/ePREMs (Hospital Anxiety and Depression Scale, Patient-Reported Outcome Measurement Information System Profile-29, Treatment Satisfaction Questionnaire for Medication, and Person-Centered Coordinated Care Experience Questionnaire). Engagement was defined at the questionnaire level and analyzed using the ALGOPROMIA-Classification framework with repeated stratified cross-validation. Model explainability was assessed using permutation-based Shapley Additive Explanations.

Results: A total of 81 participants contributed repeated questionnaire-level observations (mean PrEP duration 689 d). PrEP was well tolerated, with no moderate or severe adverse events; mild transient symptoms were mainly gastrointestinal (31/45, 68.9%) and neurological (26/45, 57.8%). Renal function remained stable (creatinine: 0.86 [SD 0.13] mg/dL; estimated glomerular filtration rate: P=.498). Psychological well-being and HRQoL remained stable (Hospital Anxiety and Depression Scale<7; Patient-Reported Outcome Measurement Information System Profile-29 near population norms). Treatment satisfaction was consistently high (Treatment Satisfaction Questionnaire for Medication≈85 - 87), and satisfaction with the NAVETA telemedicine model remained stable (8/10). Engagement showed clear sociodemographic and behavioral gradients. Ensemble-based machine learning models achieved good discrimination in predicting engagement (area under the curve≈0.82) across ≈12,300 questionnaire-level observations. Random forest was retained for robustness and consistency. Shapley Additive Explanations analysis highlighted lifestyle-related variables as the most influential predictors, with heterogeneous individual-level effects.

Conclusions: Naveta enabled feasible telemedicine-based PrEP follow-up with preserved HRQoL, high satisfaction, and stable safety. Combining longitudinal ePROMs/ePREMs with explainable machine learning allowed detailed characterization of digital engagement, supporting digitally supported PrEP care optimization and informing future comparative and cost-effectiveness studies.

(J Med Internet Res 2026;28:e87592) doi:10.2196/87592

XSL•FO
RenderX

KEYWORDS

pre-exposure prophylaxis; digital follow-up; health-related quality of life; patient engagement; electronic patient-reported experience measure; electronic patient-reported outcome measure

Introduction

HIV infection continues to be a stigmatized condition in many social contexts. In Spain, pre-exposure prophylaxis (PrEP), based on outpatient hospital pharmacy medicines, has proven to be an effective and safe strategy for HIV prevention among populations at high risk of infection [1]. Following the recommendation of the European Medicines Agency, PrEP with emtricitabine/tenofovir disoproxil was authorized in Spain in August 2016 for HIV-negative individuals aged 16 years or older at high risk of acquisition [2]. In routine clinical practice, this regimen has been consistently reported as safe and well-tolerated [3-5]. Nevertheless, PrEP does not replace other preventive measures for HIV or other sexually transmitted infections [6,7] and must be prescribed within specialized units as part of a comprehensive HIV prevention strategy.

Beyond clinical effectiveness and safety, contemporary health care increasingly emphasizes a more integrated perspective that incorporates health-related quality of life (HRQoL) across physical, emotional, and social dimensions. For individuals using PrEP, this approach is particularly relevant, as persistent stigma surrounding HIV prevention may influence psychological well-being, engagement with health care services, and participation in digital follow-up programs [8]. Previous studies have shown that higher physical, psychological, and environmental HRQoL scores are associated with a greater likelihood of PrEP use [9]. Accordingly, in addition to clinical outcomes, it is important to assess treatment-related satisfaction and HRQoL, as these factors may influence adherence to preventive therapy [10]. Although a positive impact of PrEP on HRQoL has been described [1], further studies using validated assessment frameworks are needed to generate reliable evidence and to support strategies aimed at sustaining adherence to PrEP over time [11-14].

Digital health technologies, particularly telemedicine [15,16], combined with electronic patient-reported outcome measures (ePROMs) and electronic patient-reported experience measures (ePREMs) [17], represent a promising approach to address these needs within person-centered health care models [18]. Digital interventions have demonstrated substantial potential in HIV prevention and care [19], including among individuals with mental health conditions or substance use disorders [20,21], where ePROMs may facilitate early identification of vulnerability [22]. However, adherence to digital prevention and follow-up programs remains highly variable [23-26]. Engagement in such programs is shaped not only by clinical factors but also by psychosocial, demographic, and behavioral determinants, including age, gender, educational level, and health-related behaviors, whose interrelationships are often complex and nonlinear [27-29].

Sustained engagement is critical for the validity of longitudinal digital monitoring and for translating real-world patient-reported data into meaningful clinical insights [30-32]. High engagement enhances the reliability of ePROMs and ePREMs and supports value-based health care strategies by enabling more informed, patient-centered decision-making [33,34]. However, classical statistical approaches may be insufficient to capture the multidimensional and nonlinear patterns underlying digital engagement, particularly when multiple patient-reported measures are considered simultaneously. In this context, machine learning methods provide a robust alternative, as they are well-suited to modeling complex interactions, handling high-dimensional data, and improving discrimination beyond traditional regression techniques.

Within this framework, the Naveta value-based telemedicine initiative has recently been developed as a digital platform for collecting ePROMs and ePREMs [35]. The Naveta-Phemium platform has demonstrated efficiency in chronic disease populations and may be equally applicable to individuals receiving PrEP, who represent preventive users rather than patients. In parallel, the ALGOPROMIA project is an ongoing initiative focused on predicting patient engagement—operationalized as response rates to ePROMs and ePREMs—using machine learning models applied to data generated within the Naveta digital HRQoL follow-up program [36].

On the basis of these considerations, the objectives of this study were twofold. The primary objective was to describe the usefulness of the Naveta-Phemium platform for longitudinal monitoring of effectiveness, safety, HRQoL, and treatment satisfaction among HIV-negative adults receiving PrEP over a 24-month follow-up period. The secondary objective, using the ALGOPROMIA framework, was to develop and internally validate a machine learning–based participant classification system to characterize and predict engagement with the digital follow-up tool among PrEP users.

Methods

Study Design and Participants

This was an observational multicenter study conducted in two public hospitals in the Balearic Islands (Spain). All HIV-negative adults at high risk of HIV infection who were starting or continuing PrEP between November 2022 and March 2025 were considered eligible. Participants were required to be registered in the Naveta-Phemium PrEP HIV program, as previously described by Mercadal-Orfila et al [35], and to provide written informed consent. Participants who provided informed consent (refer to the Ethical Considerations subsection for details) received a link to access the electronic questionnaires, which they could complete on their own device (PC, tablet, or mobile phone). Participants for whom data could not be obtained because of incomplete medical records were excluded from the study.

Participants were recruited consecutively during routine clinical care through the Hospital Pharmacy Service within specialized

XSL•FO
RenderX

PrEP units at the participating public hospitals. Eligible individuals were informed about the study by hospital pharmacists as part of standard care procedures and were invited to participate without additional selection criteria. Participation was voluntary, no incentives were offered, and no replacement or targeted sampling strategies were applied.

Ethical Considerations

This study was approved by the Ethics Committee for Research with Medicines of the Balearic Islands (Comitè d'Ètica de la Investigació amb Medicaments de les Illes Balears, CEIm), under protocol code IB 5117/23 EOm. The committee issued a favorable opinion for this prospective observational study with medicinal products, including approval of the recruitment period extension, in accordance with Spanish regulations governing observational studies with medicines (Royal Decree 957/2020). The study was conducted in compliance with the principles of the Declaration of Helsinki and Good Clinical Practice guidelines.

All participants provided informed consent prior to inclusion in the study through a secure electronic process implemented via the NAVETA platform. Participants accessed the study information and consent materials through a web-based link and electronically provided informed consent. The informed consent documents, participant information materials, and electronic consent procedures were reviewed and approved by the Ethics Committee for Research with Medicines of the Balearic Islands (CEIm), in line with the observational nature of the study.

Participant data were processed in compliance with applicable data protection regulations. All data used for analysis were deidentified prior to processing, and access to identifiable information was restricted to authorized clinical personnel only. Data were stored on secure servers with appropriate technical and organizational safeguards to ensure confidentiality and prevent unauthorized access.

No financial compensation or material incentives were provided to participants for their participation in this study.

Recruitment, Data Collection, and Variables

Patients receiving PrEP were initially assessed and clinically managed by the Infectious Diseases service at public hospitals in the Balearic Islands, Spain, in accordance with established clinical protocols [37]. Once patients met the eligibility criteria and were referred to outpatient hospital pharmacy services for medication dispensing and follow-up, recruitment for the telemonitorization program was conducted exclusively within the hospital pharmacy setting.

During routine outpatient pharmacy consultations, hospital pharmacists informed eligible adult patients—either at PrEP initiation or during routine follow-up—about the objectives of the telemonitorization program and the procedures involved. Patients who expressed interest were provided with a secure link granting access to the NAVETA platform.

Through this platform, participants reviewed the study information, accepted the study conditions, and provided informed consent via an electronic consent process. Following confirmation of consent, participants initiated the telemonitorization process and completed the electronic questionnaires in accordance with the NAVETA_VIH protocol.

Sociodemographic and lifestyle characteristics were collected through an ad hoc questionnaire integrated into the Naveta-Phemium platform. This questionnaire captured information on age, gender, marital status, educational level, employment status, patient association membership, consumption habits, health-related behaviors, drug allergies, and PrEP treatment line. Health-related behaviors included BMI, self-reported adherence to a balanced diet, and level of physical activity, as defined in the Naveta platform [35].

Clinical safety data related to emtricitabine/tenofovir disoproxil were extracted from routine hospital medical records. These included adverse events, treatment discontinuations, serum creatinine levels, and estimated glomerular filtration rate (eGFR, mL/min/1.73 m^2), calculated using the CKD-EPI equation, as well as hepatitis infections, HIV seroconversions, and sexually transmitted infections. Clinical data were collected at baseline and at 6, 12, 18, and 24 months of follow-up.

Assessment Instruments and Follow-Up Schedule

Patient-reported outcomes and experiences were assessed longitudinally using validated instruments: the Treatment Satisfaction Questionnaire for Medication (TSQM) [38], the Patient-Reported Outcome Measurement Information System Profile-29 (PROMIS-29) [39], the Hospital Anxiety and Depression Scale (HADS), and the Person-Centered Coordinated Care Experience Questionnaire (P3CEQ) [40].

Data collection followed the predefined assessment schedule of the Naveta follow-up program. HADS and TSQM were administered at baseline and at 6, 12, 18, and 24 months, whereas PROMIS-29 and P3CEQ were completed at baseline, 12, 18, and 24 months.

Statistical Analysis

All statistical analyses, including descriptive statistics, inferential tests, and machine learning procedures, were conducted in Python (version 3.10) [41]. Baseline demographic and clinical characteristics were summarized using descriptive statistics. Continuous variables are reported as mean (SD), and categorical variables as frequencies and percentages.

Associations between digital engagement and sociodemographic or lifestyle variables were explored by comparing engagement scores (mean, SD) across categories (eg, gender, age range, employment status, BMI, lifestyle habits, and educational level). Correlations among ePROMs were assessed using the Spearman correlation coefficient.

Group comparisons were performed using parametric or nonparametric tests according to distributional assumptions. Normality of continuous variables was assessed using the Shapiro–Wilk test, and homoscedasticity was evaluated using the Levene test. When both assumptions were met (P>.05 for Shapiro–Wilk and Levene tests), one-way ANOVA followed by Tukey post hoc test was applied. When these assumptions were not fulfilled, nonparametric alternatives were used: the Kruskal–Wallis test with Dunn post hoc comparisons for

XSL•FO
RenderX

variables with more than two categories, and the Mann–Whitney *U* test for dichotomous variables (eg, gender, substance use). For all Kruskal–Wallis post hoc analyses, multiple comparisons were adjusted using the Dunn test with Bonferroni correction.

For longitudinal visualization of engagement rates, observed proportions were accompanied by 95% CIs calculated using the Wilson score method, which provides improved coverage for binomial proportions, particularly when sample sizes vary across time points.

Longitudinal Analysis and Missing Data Handling

Longitudinal analyses accounted for variation in sample size across follow-up timepoints, reflecting the asynchronous recruitment on 07/01/2026 05:15:00 PM and follow-up inherent to the real-world design. At each time point, the effective sample size corresponded to participants who completed the relevant questionnaires. Missing data were handled using case-wise exclusion, retaining all available observations for each specific analysis. Completion of ePROMs and ePREMs within the Naveta platform is voluntary at each scheduled timepoint, resulting in intermittent and participant-specific missingness patterns. Because this missingness structure does not satisfy the assumptions required for multiple imputation, case-wise exclusion was considered the most conservative and assumption-free approach. To assess whether noncompletion reflected systematic differences between participants, an exploratory comparison of baseline ePROMs scores was performed between individuals with low completion (≤25% of scheduled assessments) and high completion (≥75%). The results of this analysis are reported in the Results section. No outliers were removed or adjusted, and all analyses reflect the natural variability of the clinical population.

For engagement analyses, Shapiro–Wilk tests indicated nonnormal score distributions, and Levene tests revealed heteroscedasticity across comparison groups. Accordingly, nonparametric methods were applied using the Kruskal–Wallis test with Dunn post hoc comparisons. For longitudinal and subgroup-level comparisons, the primary objective was descriptive characterization rather than formal hypothesis testing. Results are therefore reported using effect direction, distributional patterns, and associated *P* values, without emphasis on full test statistics (eg, F, H, or W values), which are of limited interpretability in the presence of unbalanced repeated observations, heteroscedasticity, and intermittent missingness. This reporting strategy is consistent with recommendations for observational digital health studies, where statistical inference is used to support descriptive trends rather than to test predefined causal hypotheses.

For qualitative interpretation of scores, established thresholds were applied where available. For HADS, subscale scores were interpreted using standard cut-offs: 0‑7=normal, 8‑10=borderline, and ≥11=clinical case [42,43]. PROMIS-29 T-scores (mean 50, SD 10) were interpreted using provisional thresholds of approximately 55 (mild), 60 (moderate), 65 (moderately severe), and 70 (severe) for relevant domains [44]. For TSQM and P3CEQ, no definitive cut points are available for Spanish populations; therefore, scores were interpreted relative to sample distributions and existing literature.

Machine Learning Analysis: ALGOPROMIA-Classification

Although clinical outcomes were assessed according to a predefined 24-month follow-up schedule, analyses of digital engagement and model stability were conducted at the questionnaire level and therefore incorporated all available observations generated within the platform, including those occurring beyond the formal clinical follow-up window. Digital engagement was operationalized as a binary outcome, and participants were classified as “good responders” or “poor responders” based on their mean completion rate of scheduled ePROMs and ePREMs across follow-up (threshold ≥0.50). Three machine learning algorithms were evaluated—histogram-based gradient boosting (HGB), random forest (RF), and logistic regression—given their suitability for binary classification and their ability to model both linear and nonlinear relationships [45].

All predictors were transformed using one-hot encoding, and model hyperparameters were selected according to recommendations from prior work [46]. Importantly, the engagement outcome was defined exclusively by response occurrence to scheduled questionnaire dispatches, independently of whether questionnaire content was completed. Thus, missing ePROMs/ePREMs data did not contribute to outcome definition, avoiding circularity between predictor availability and the target variable.

Because of class imbalance, the RandomOverSampler method was applied exclusively within the training folds to prevent information leakage. Model performance was evaluated using repeated stratified 5-fold cross-validation (five repeats), replacing an initial single train–test split. Oversampling was conducted only within training folds, and evaluation was performed on untouched test partitions. Performance metrics are reported as mean and SD of receiver operating characteristic-area under the curve (ROC-AUC), accuracy, precision, recall, and F_1-score across folds. This framework mitigates overfitting in small real-world datasets. The ALGOPROMIA framework has previously been applied in people living with HIV [47], although this does not constitute external validation for the present PrEP cohort.

To characterize score separation between engaged and nonengaged observations, the distribution of predicted probabilities was examined using kernel density estimates. Predicted probabilities from the final model were stratified by observed engagement status, and smoothed density curves were used to visualize overlap and separation between classes across the full range of decision thresholds.

Model Explainability

After internal validation, a final model was retrained using the fully oversampled dataset to enable interpretability analysis. Feature contributions were examined using SHAP (Shapley Additive Explanations), using a permutation-based estimator compatible with ensemble models [48].

SHAP values quantify each predictor’s contribution to the model output in log-odds space. Mean absolute SHAP values were used to assess overall feature importance independent of

XSL•FO
RenderX

direction, whereas mean signed SHAP values indicated the predominant direction of association (positive or negative) across observations. A feature may therefore show a high mean absolute impact with a near-zero mean signed value when its effect varies in direction across individuals. Consistent with current interpretability guidelines, SHAP values were interpreted as indicators of relative importance and directional influence rather than as absolute changes in predicted probability.

Although the cohort comprised 81 individuals, the classifier was trained on the full set of questionnaire-response events. Because engagement was defined at the questionnaire instance level, the effective sample size reflected repeated observations per participant rather than the number of individuals.

Results

Study Cohort Characteristics, Clinical and Sociodemographic Profile

Our study included 81 (93.1%) individuals enrolled in the Naveta-Phemium PrEP VIH initiative who agreed to participate. A total of 6 (6.9%) individuals declined participation. The most frequently reported reason was lack of interest in participating (3/6, 50%), followed by unspecified reasons (2/6, 38.2%) and limited access to a mobile phone, email, or internet (1/6, 11.8%).

The majority were male (77/81, 95.1%), with most participants (60/81, 74.1%) aged between 21 and 45 years. Regarding social characteristics, 74.3% (n=55) were single, 46.9% (n=38) held a bachelor's degree or higher, and 79.0% (n=64) were employed. Among the health-related behaviors, 54.3% (n=44) have a normal body weight, 48.3% (n=29) reported following a regular balanced diet, while 46.7% (n=28) engaged in physical activity regularly, and another 46.7% (n=27) did so occasionally. Regarding the consumption habits, 48.3% (n=29) had never smoked, and the majority reported occasional (47/81, 58.0%) or weekend (20/81, 24.7%) alcohol consumption. Use of other substances was low, with 85.3% (n=64) reporting no consumption (Table 1). Comorbidities were present in 14.3% of PrEP users. The most frequent condition was hypertension (7.6%), followed by other comorbidities (5.9%) and hypercholesterolemia (3.4%).

XSL•FO
RenderX

Table . Sociodemographic, lifestyle, and health-related characteristics of adult patients receiving HIV pre-exposure prophylaxis (PrEP) enrolled in a prospective observational study conducted at public hospitals in the Balearic Islands (Spain), 2023 - 2024 (N=81).

Participant characteristics	Values, n (%)[a]
Sex	
Female	4 (4.9)
Male	77 (95.1)
Age range (years)	
66 - 75	2 (2.5)
46 - 65	19 (23.5)
21 - 45	60 (74.1)
Marital status	
Widowed	1 (1.4)
Divorced	4 (5.4)
Unknown	2 (2.7)
Married/Cohabiting	12 (16.2)
Single	55 (74.3)
Education level	
Primary education	3 (3.7)
Secondary education	18 (22.2)
Vocational training	21 (25.9)
Bachelor's degree	38 (46.9)
PhD	1 (1.2)
Employment status	
Retired	2 (2.5)
Temporary leave	3 (3.7)
Unemployed	6 (7.4)
Student	6 (7.4)
Employed	64 (79.0)
Not belonging to the patients' association	81 (100)
Consumption habits	
Alcohol consumption	
Daily	5 (6.2)
Weekends	20 (24.7)
Occasionally	47 (58)
Never	9 (11.1)
Tobacco use	
Yes	17 (28.3)
Ex-Smoker	14 (23.3)
Never	29 (48.3)
Other substances (abuse drugs)	
Yes	11 (14.7)
No	64 (85.3)
Health-related behaviors	
BMI range	

XSL•FO
RenderX

Participant characteristics	Values, n (%)[a]
Unknown/Not Classified	1 (1.2)
Obesity	10 (12.3)
Overweight	26 (32.1)
Normal weight	44 (54.3)
Balanced diet	
Never	1 (1.7)
Occasional	30 (50.0)
Regular	29 (48.3)
Physical activity	
Never	4 (6.7)
Occasional	28 (46.7)
Regular	28 (46.7)
Drug allergies	
Yes	6 (7.4)
No	75 (92.6)

[a]n (%): Absolute and relative frequencies. The total number of responses varies due to nonmandatory fields and missing data.

All participants received emtricitabine/tenofovir 200/245 mg, with a mean treatment duration of 688.9 days (95% CI 602.3 - 775.5). During follow-up, 16 out of 81 patients (19.8%) discontinued or interrupted PrEP treatment, a finding partly associated with the use of on-demand rather than continuous regimens. Importantly, no discontinuations were attributed to adverse effects. Regarding emtricitabine/tenofovir effectiveness, no evidence of HIV seroconversions during the period of 24 months was found. On the other hand, some types of sexually transmitted infections were reported in 35.9% of subjects, highlighting *Neisseria gonorrhoeae* (30.4%) and *Chlamydia trachomatis* (13%).

PrEP was well tolerated, with no participant reporting moderate or severe adverse events. The most frequent patient-reported effects were mild gastrointestinal complaints (68.9%) and mild nervous system symptoms, such as headache, dizziness, tingling, or fatigue (57.8%). All events were transient and self-limited. In the analysis of renal function biomarkers (Figure 1), serum creatinine levels remained stable throughout the follow-up period: 0.86 (SD 0.13) mg/dL at baseline (95% CI 0.83 - 0.90), 0.86 (SD 0.12) mg/dL at 6 months (95% CI 0.83 - 0.90), 0.87 (SD 0.12) mg/dL at 12 months (95% CI 0.84 - 0.93), and 0.86 (SD 0.13) mg/dL at 24 months (95% CI 0.78 - 0.93). Data met the assumptions of normality (Shapiro–Wilk, W=0.985, P=.055) and homogeneity of variances (Levene test, P=.935). The one-way Welch ANOVA showed no statistically significant differences between timepoints ($F_{3,\ 58.4}$=0.252, P=.860). In contrast, eGFR values exhibited a slight, nonsignificant decline ($F_{3,\ 58.4}$=0.801, P=.498).

XSL•FO
RenderX

Figure 1. Longitudinal assessment of renal function during digital pre-exposure prophylaxis (PrEP) follow-up in a prospective observational cohort. Temporal evolution of renal function parameters over a 24-month prospective observational follow-up of HIV-negative adults receiving PrEP and monitored through the NAVETA digital telemedicine platform in Spain (n=81). Mean (SD) values of estimated glomerular filtration rate (eGFR; left panel) and serum creatinine (right panel) are shown at baseline, 6, 12, and 24 months. Differences across time points were assessed using Welch ANOVA. AUC: area under the curve; HGB: histogram-based gradient boosting; LR: logistic regression; NPV: negative predictive value; PPV: positive predictive value; RF: random forest.

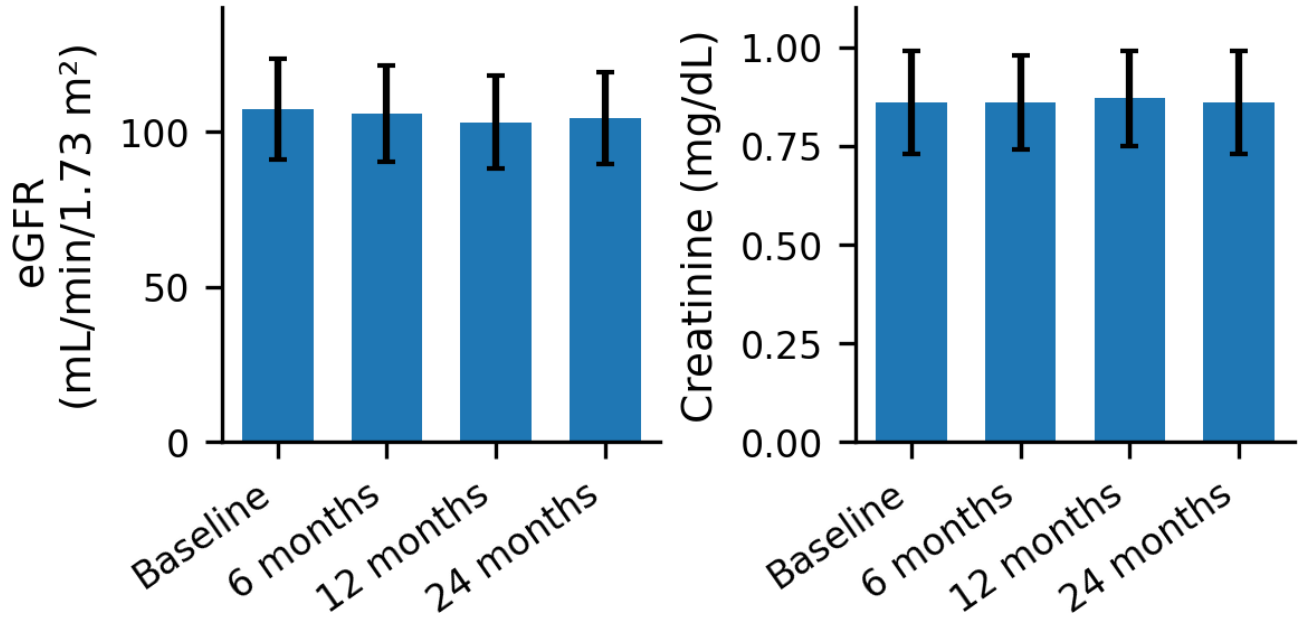

Serum creatinine and eGFR (mL/min/1.73 m²) showed a coherent inverse relationship over time. Strong positive correlations were observed between creatinine measurements at consecutive visits (eg, r=0.63, P<.001 between baseline and 6 mo; r=0.84, P=.003 between baseline and 24 mo), indicating consistent temporal stability of this biomarker. As expected, creatinine and eGFR values were inversely correlated at all timepoints, with the strongest association at baseline (r=–0.78, P<.001). Correlations between creatinine and eGFR at later stages ranged from –0.29 to –0.74. Temporal correlations among eGFR measurements were similarly high (r ≈ 0.73 - 0.92, all P<.001), reflecting robust internal consistency of renal function indices across the 24-month follow-up (Table 2).

Table . Pearson correlation matrix between serum creatinine and estimated glomerular filtration rate (eGFR, mL/min/1.73 m²) across follow-up timepoints in a 24-month prospective observational cohort of HIV-negative adults receiving pre-exposure prophylaxis (PrEP) at public hospitals in the Balearic Islands, Spain (n=81). Table values represent Pearson correlation coefficients between serum creatinine and eGFR measured at baseline (BL), 6, 12, and 24 months.

	Creatinine (BL)	Creatinine (6 months)	Creatinine (12 months)	Creatinine (24 months)	eGFR (BL)	eGFR (6 months)	eGFR (12 months)
Creatinine (6 months)	0.628[a]	—[b]	—	—	—	—	—
Creatinine (12 months)	0.612[a]	0.728[a]	—	—	—	—	—
Creatinine (24 months)	0.841[a]	0.824[a]	0.560[c]	—	—	—	—
eGFR (BL)	−0.779[a]	−0.438[a]	−0.356[c]	−0.531[c]	—	—	—
eGFR (6 months)	−0.392[c]	−0.744[a]	−0.444[c]	−0.571[c]	0.729[a]	—	—
eGFR (12 months)	−0.358[c]	−0.516[a]	−0.705[a]	−0.457[c]	0.710[a]	0.791[a]	—
eGFR (24 months)	−0.671[c]	−0.665[c]	−0.294[c]	−0.775[a]	0.918[a]	0.893[a]	0.752[a]

[a]P<.01.

[b]Not applicable.

[c]P<.05.

XSL•FO
RenderX

Analysis of the ePROMs and ePREMs

Patient-reported outcomes (ePROMs and ePREMs) were evaluated longitudinally to characterize changes in psychological well-being, treatment satisfaction, HRQoL, and perceived person-centered care. A detailed summary of mean (SD) values and sample sizes at each timepoint is provided in Table 3.

Table . Longitudinal trajectories of electronic patient-reported outcome measures (ePROMs) and electronic patient-reported experience measures (ePREMs) over 24 months in a prospective observational cohort of HIV-negative adults receiving pre-exposure prophylaxis (PrEP) and followed through the NAVETA digital monitoring program at public hospitals in the Balearic Islands, Spain (n=81). Numbers (n) indicate the number of participants who completed each questionnaire at each timepoint, reflecting real-world longitudinal follow-up with variable completion rates. *P* values correspond to nonparametric longitudinal comparisons across available timepoints.

ePROMs/ ePREMs and sub-scale	BL[a]		6 months		12 months		18 months		24 months		*P* value
	mean (SD)	n	mean (SD)	n	mean (SD)	n	mean (SD)	n	mean (SD)	n	
HADS[b]											
Anxiety	5.14 (3.14)	76	3.63 (3.49)	35	4.69 (3.97)	32	4.09 (3.78)	32	4.44 (3.71)	32	.088
Depression	3.16 (2.94)	79	2.34 (2.92)	35	3.03 (3.79)	32	3.44 (3.38)	32	2.44 (2.55)	32	.27
TSQM[c]											
Global satisfaction	__[d]	—	76.2 (17.3)	32	76.2 (8.24)	32	75.7 (15.65)	32	86.6 (12.91)	32	.444
PROMIS-29[e] (T-score)											
Depression	49 (7.37)	50	—	—	47.7 (7.49)	34	47.6 (7.67)	25	51.4 (9.3)	19	.4
Anxiety	52.1 (7.34)	50	—	—	50.9 (9)	34	50.4 (8.46)	25	52.7 (9.04)	19	.693
Fatigue	44.8 (7.88)	50	—	—	42.9 (10.05)	34	45 (9.8)	25	44.9 (10)	19	.741
Sleep disturbance	47.6 (7.01)	50	—	—	46.4 (7.26)	34	48.2 (7.11)	25	49.5 (5.74)	19	.397
Ability to participate in social roles and activities	54.9 (7.56)	50	—	—	56.2 (8.14)	34	56.7 (7.8)	25	55.3 (7.45)	19	.734
Physical function	56.7 (1.72)	50	—	—	55.8 (3.86)	34	55.7 (4.59)	25	54.5 (5.22)	19	.163
P3CEQ[f]											
Total score	20.1 (6.55)	74	—	—	21 (6.24)	30	21 (7)	23	22.7 (5.28)	19	.39

[a]BL: baseline.

[b] HADS: Hospital Anxiety and Depression Scale.

[c]TSQM: Treatment Satisfaction Questionnaire for Medication.

[d]Not applicable.

[e]PROMIS-29: Patient-Reported Outcomes Measurement Information System.

[f]P3CEQ: Person-Centered Coordinated Care Experience Questionnaire.

XSL•FO
RenderX

Scores on the HADS indicated low levels of psychological distress throughout follow-up. Both anxiety and depression remained within the normal range (0‑8) and showed no statistically significant changes across time (anxiety: *P*=.09; depression: *P*=.27), consistent with emotional stability over the 24-month period.

Treatment satisfaction, assessed through the TSQM Global Satisfaction domain, remained high and stable across follow-up (*P*=.44), reflecting a persistently favorable perception of effectiveness, tolerability, and convenience. The TSQM was not administered at baseline; the first assessment was conducted at 6 months after treatment initiation, as the instrument is specifically designed to evaluate satisfaction with ongoing treatment.

Across PROMIS-29 domains, mean T-scores remained near the population reference mean (50, SD 10). Emotional health dimensions (depression, anxiety, fatigue, sleep disturbance) consistently fell within expected normative limits (<55), whereas social participation and physical function were slightly above average. None of the PROMIS-29 domains showed statistically significant variation over time (all *P*>.05).

Perceived person-centered coordinated care, measured using the P3CEQ total score, showed a modest but nonsignificant upward trend over follow-up (*P*=.39), indicating stable or slightly improved perceptions of coordination, communication, and shared decision-making. In addition, satisfaction with the NAVETA telemedicine follow-up model, assessed using a platform-specific Likert-type satisfaction scale, remained consistently high throughout follow-up (mean score 8/10), reflecting sustained positive user evaluations of the digital monitoring experience.

To assess potential bias related to missing data, individual completion patterns across scheduled assessment stages (baseline, 6, 12, 18, and 24 mo) were examined. Questionnaire completion decreased progressively over follow-up: 94.8% of participants completed at least 25% of assessments, 76.3% completed ≥50%, 51.5% completed ≥75%, and 30.9% completed all scheduled stages. Individual completion trajectories and summary completion ratios are presented in Figure S1 and Table S2 in Multimedia Appendix 1. To evaluate whether early noncompletion was associated with underlying clinical or psychological differences, baseline ePROMs and ePREMs scores were compared between participants with low completion (≤25% of stages; n=5) and high completion (≥75%; n=50). No statistically significant differences were observed across any domain (HADS, TSQM, PROMIS-29, or P3CEQ; all >.25), and baseline score distributions were closely aligned between groups.

Correlations Among ePROMs/ePREMs During Follow-Up

Correlation analyses among patient-reported outcomes (PROMs and PREMs) revealed consistent interdomain relationships over time. At baseline, moderate associations were observed between overall HRQoL (PROMIS-29) and both psychological well-being (HADS, *r*=0.58) and treatment satisfaction (TSQM, *r*=–0.61), indicating that better emotional status was linked to higher perceived quality of life and satisfaction. Person-centered coordinated care (P3CEQ) showed weak positive correlations with other measures. At 24 months, the pattern strengthened, with high correlations between HADS and PROMIS-29 (*r*≈1.0) and inverse associations with TSQM (*r*≈–0.72), suggesting that improved emotional well-being paralleled higher HRQoL and treatment satisfaction. These findings highlight internal consistency across PROM domains and the mutual influence of emotional, experiential, and satisfaction-related dimensions during follow-up.

Analysis of the Response Profile to Patient-Reported Outcomes Questionnaires

Engagement to the Naveta digital follow-up platform exhibited significant sociodemographic and behavioral gradients (Figure 2). Men demonstrated substantially higher engagement than women (0.62, SD 0.28 vs 0.39, SD 0.14; *P*<.001), and engagement increased progressively with age, being significantly greater in the 46‑65 and 66‑75 years age groups compared with younger participants (both *P*<.001, Dunn test). Participants classified as overweight also showed higher engagement than those with normal weight (*P*<.001), although no differences were found between overweight and obese categories.

XSL•FO
RenderX

Figure 2. Sociodemographic and behavioral determinants of digital engagement during a 24-month telemedicine-based PrEP follow-up in a prospective observational cohort. *P<.05.

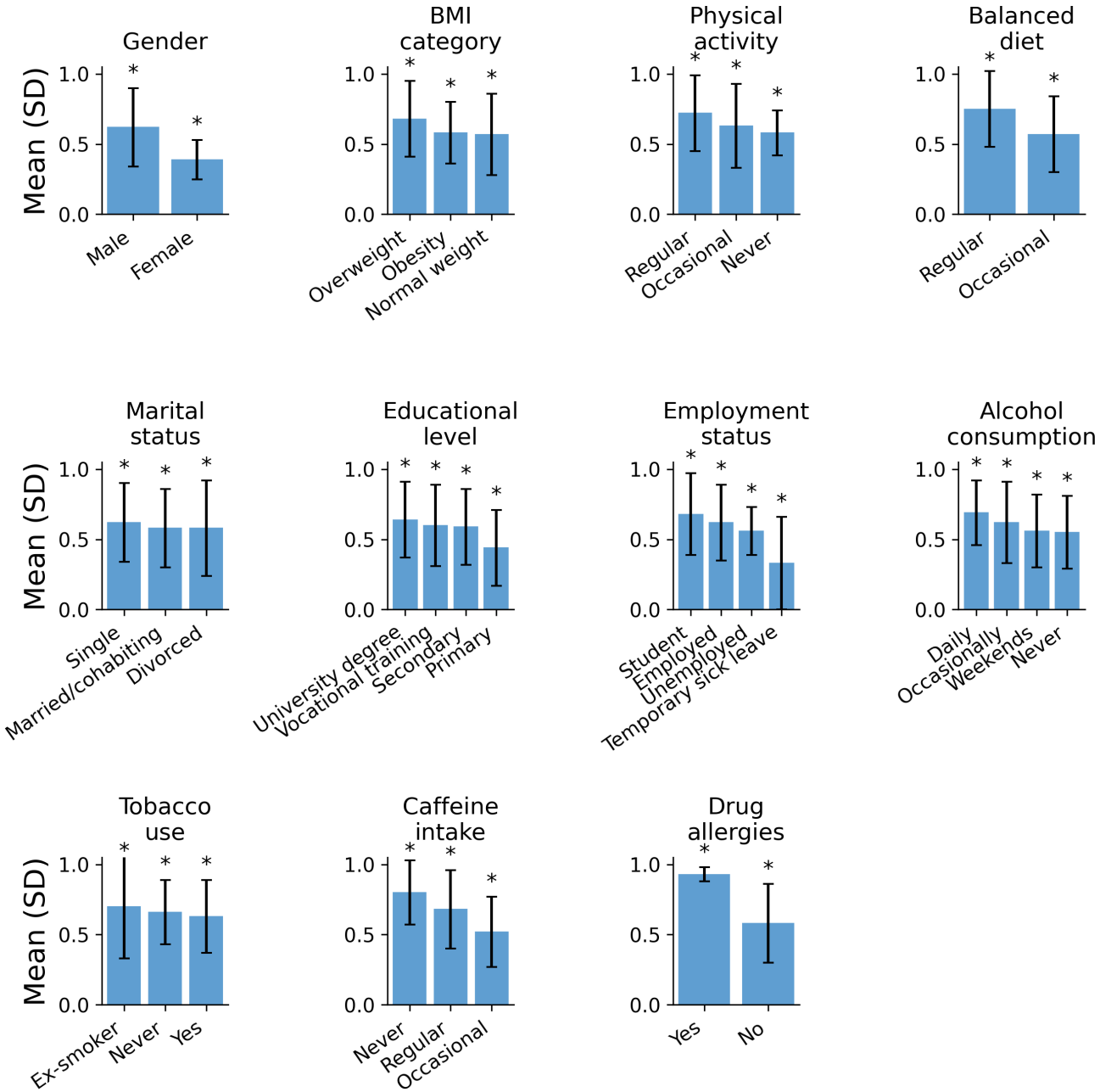

Lifestyle-related variables were among the strongest determinants of engagement. Those reporting regular physical activity (0.72, SD 0.27) exhibited significantly higher engagement compared with participants who were inactive (P<.001) or exercised occasionally (P=.001). Similarly, maintaining a balanced diet was associated with greater continuity than occasional engagement to healthy eating habits (P<.001). Educational level displayed a graded pattern: participants with university degrees showed higher engagement than those with primary or secondary education (P=.005 and P=.03, respectively). Employment status also played a major role, as individuals on temporary sick leave (0.33, SD 0.33) reported markedly lower engagement than employed (P<.001) or student participants (P<.001).

Behavioral factors related to substance use followed similar trends. Daily alcohol consumers presented higher engagement than those drinking only on weekends or never (P=.003). Former smokers were more adherent than current smokers (P=.007), and complete caffeine abstainers showed the highest engagement levels (0.80, SD 0.23), significantly exceeding both regular (P<.001) and occasional consumers (P<.001).

Sociodemographic and behavioral gradients of digital engagement among HIV-negative adults receiving PrEP and followed through the NAVETA digital telemedicine platform at public hospitals in the Balearic Islands, Spain (n=81). Bars represent mean (SD) engagement ratios, defined as the proportion of completed ePROMs/ePREMs over scheduled assessments. Asterisks indicate statistically significant between-group differences based on nonparametric comparisons (Kruskal–Wallis test with Dunn post hoc correction; P<.05).

Although the bivariate associations were statistically robust, classical regression models were insufficient to capture the

XSL•FO
RenderX

underlying nonlinear patterns governing digital engagement [34]. For this reason, we applied our machine learning framework, ALGOPROMIA-Classification, to improve discrimination between engagement profiles within the NAVETA follow-up program (Figure 3, Table 4).

Figure 3. Discrimination performance of machine learning models for predicting digital engagement during pre-exposure prophylaxis (PrEP) follow-up in a prospective observational cohort. Comparative performance of machine learning models for questionnaire-level digital engagement classification in a prospective observational cohort of HIV-negative adults receiving PrEP and followed through the NAVETA digital telemedicine platform at public hospitals in the Balearic Islands, Spain (n=81; 24-mo follow-up). (**A**) Receiver operating characteristic (ROC) curves showing the mean performance of HGB, RF, and LR (saga solver) across repeated stratified 5×5 cross-validation. The y-axis represents sensitivity (true positive rate) and the x-axis represents 1 − specificity (false positive rate). (**B**) Summary of classification metrics (mean, SD) across cross-validation folds, including area under the curve (AUC), F_1-score, PPV, and NPV. AUC: area under the curve; HGB: histogram-based gradient boosting; LR: logistic regression; NPV: negative predictive value; PPV: positive predictive value; RF: random forest.

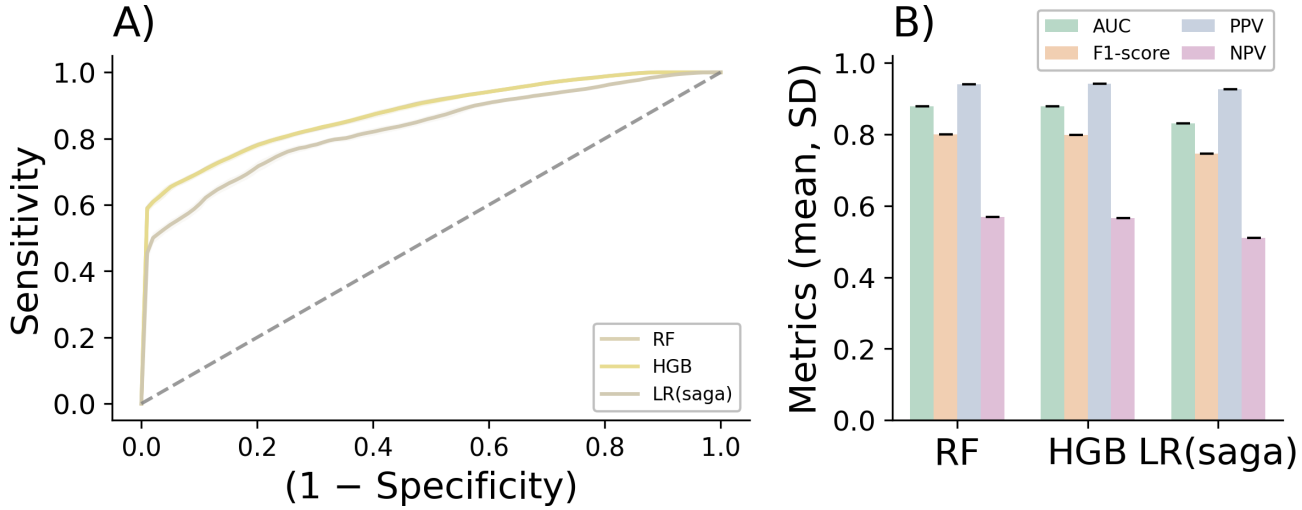

Table . Performance metrics of machine learning models for predicting digital engagement during a 24-month follow-up in a prospective observational cohort of HIV-negative adults receiving pre-exposure prophylaxis (PrEP) and monitored through the NAVETA digital telemedicine platform at public hospitals in the Balearic Islands, Spain (12,299 questionnaire-level events). Performance metrics correspond to questionnaire-level digital engagement classification. Engagement was defined as a binary outcome (engaged vs nonengaged). All metrics are reported as mean (SD) across repeated stratified 5×5 cross-validation. Class imbalance was preserved across training and test folds.

Model	AUC[a], mean (SD)	Accuracy, mean (SD)	Precision, mean (SD)	Recall, mean (SD)	F_1-score, mean (SD)	PPV[b], mean (SD)	NPV[c], mean (SD)
RF[d]	0.817 (0.007)	0.704 (0.007)	0.919 (0.007)	0.637 (0.008)	0.753 (0.006)	0.919 (0.007)	0.500 (0.007)
HGB[e]	0.816 (0.007)	0.704 (0.007)	0.919 (0.007)	0.637 (0.008)	0.752 (0.006)	0.919 (0.007)	0.499 (0.006)
LR[f] (saga)	0.809 (0.007)	0.705 (0.007)	0.918 (0.007)	0.638 (0.008)	0.753 (0.006)	0.918 (0.007)	0.499 (0.007)

[a] AUC: area under the receiver operating characteristic curve.

[b]PPV: positive predictive value.

[c]NPV: negative predictive value.

[d]RF: random forest.

[e]HGB: histogram-based gradient boosting.

[f]LR: logistic regression.

Although the cohort consisted of 81 individuals, the machine learning models were trained on all available questionnaire–response events (12,299 observations), as engagement was defined at the event level rather than the participant level (Table S1 in Multimedia Appendix 1).

Across repeated stratified 5×5 cross-validation, ensemble-based algorithms again showed the strongest overall predictive performance (Figure 1, Table 4). RF and HGB achieved very similar levels of discrimination (mean AUC=0.817, SD 0.007 and 0.816, SD 0.007, respectively), with comparable accuracy (≈0.70) and F_1-scores (≈0.75). Both ensemble models exhibited high positive predictive values (≈0.92), reflecting a strong ability to correctly identify engaged participants, whereas negative predictive values (NPVs) were consistently more modest (NPV≈0.50). Logistic regression (saga) showed slightly lower discrimination (AUC=0.809, SD 0.007), while achieving similar accuracy and F_1 performance. Taken together, these results suggest that ensemble-based approaches offered a small but consistent advantage in modeling the nonlinear behavioral patterns associated with sustained engagement in digital PrEP monitoring.

On the basis of this comparative evaluation, the RF model was retained for downstream analyses. Model performance and behavior were then characterized in greater detail using the full

XSL•FO
RenderX

set of questionnaire-level observations, including temporal stability of predicted versus observed engagement, score separation, subgroup-level discrimination and calibration, and permutation-based SHAP explainability (Figure 4). All analyses were conducted at the questionnaire level; accordingly, n refers to the number of questionnaire-level observations (digital interaction events), with individual participants contributing multiple observations across follow-up stages when applicable.

Figure 4. Temporal stability, subgroup performance, and explainability of the random forest model for predicting digital engagement during a 24-month PrEP follow-up. Evaluation of a random forest classifier applied to questionnaire-level data from a prospective observational cohort of HIV-negative adults receiving PrEP and monitored through the NAVETA digital telemedicine platform at public hospitals in the Balearic Islands, Spain (81 participants contributing 12,299 questionnaire-level observations). (**A**) Temporal stability of predicted engagement probabilities compared with observed engagement rates across follow-up timepoints, with 95% Wilson CIs; gray bars indicate the number of observations per timepoint. (**B**) Distribution of predicted engagement probabilities for engaged and nonengaged observations shown using kernel density estimates; the dashed vertical line indicates the fixed decision threshold (0.5). (**C**) Discriminatory performance across predefined demographic subgroups (sex and age), expressed as AUC with 95% CIs. (**D**) Subgroup-level calibration, expressed as the difference between predicted and observed engagement probabilities (calibration gap). (**E**) Global feature importance derived from permutation-based SHAP analysis, ranking predictors by mean absolute SHAP values. (**F**) Directional contribution of predictors to model output, expressed as mean signed SHAP values. All analyses were conducted at the questionnaire level, with individual participants contributing multiple observations across follow-up. AUC: area under the curve; PrEP: pre-exposure prophylaxis; SHAP: Shapley Additive Explanations.

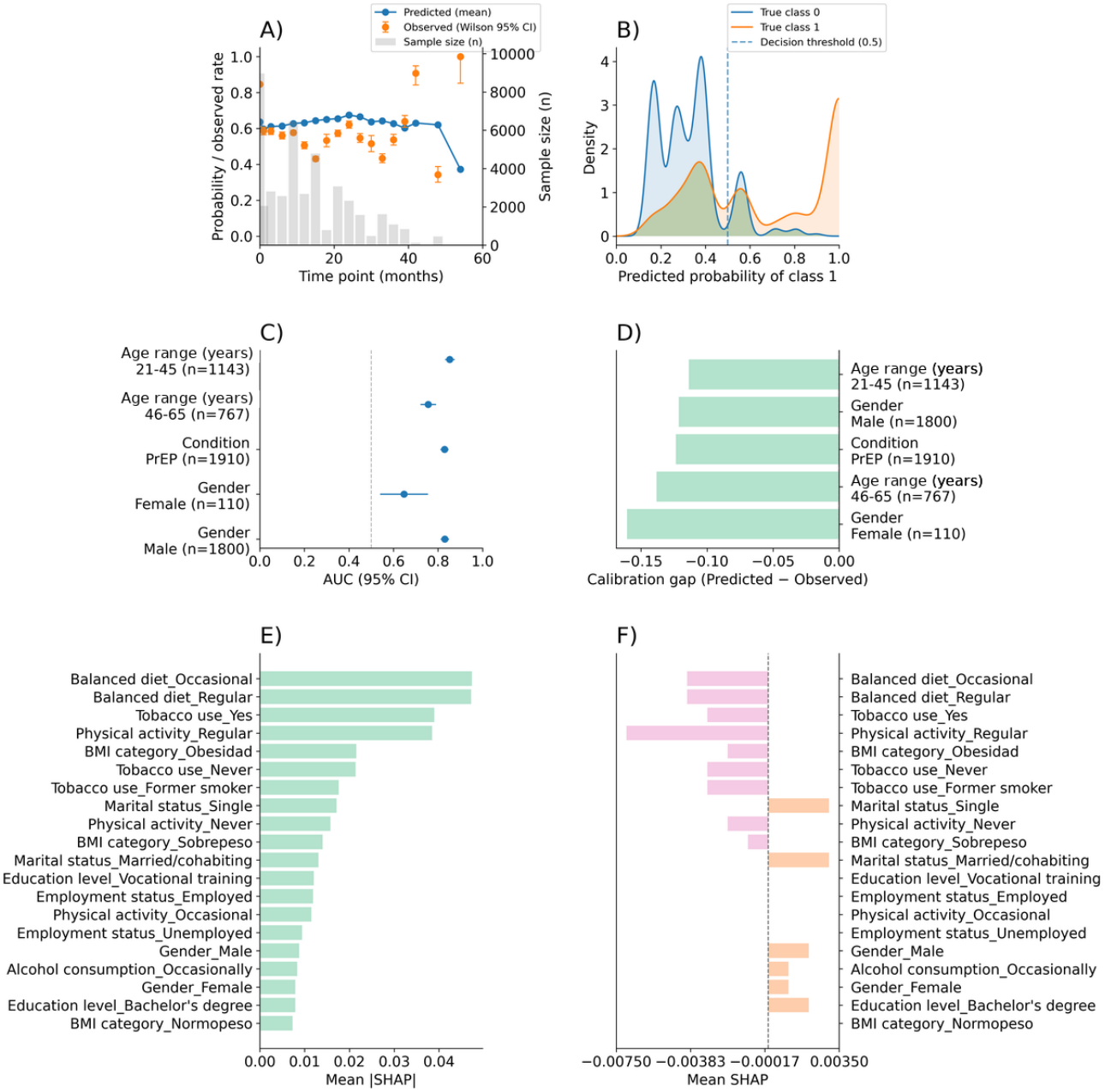

Across monitoring stages (Figure 4A), the mean predicted probability of engagement remained relatively stable over time, ranging from 0.607 to 0.669 across the most frequently observed timepoints. Across all time points, the mean difference between observed and predicted engagement rates was −0.022, indicating a small average deviation. Observed engagement rates varied more widely (0.430 - 0.846). At baseline (0 mo; n=1910), the observed engagement rate was 0.846, and the model-predicted mean probability was 0.636. At 6 months (n=1817), predicted and observed values were closely aligned (0.617 vs 0.616). At

XSL•FO
RenderX

12 months (n=1583), the predicted probability was 0.610, and the observed rate was 0.648. At 15 months (n=1693), the predicted probability was 0.607 while the observed rate was 0.430. At 24 months (n=1235), predicted and observed values again showed close agreement (0.616 vs 0.609). The sample size decreased over the follow-up stages.

Predicted score distributions showed partial separation between engaged and nonengaged observations (Figure 4B). The AUC derived from the predicted probabilities was 0.816. Among engaged observations (class 1; n=1288), the median predicted probability was 0.711 (IQR 0.464 - 1.000); among nonengaged observations (class 0; n=622), the median was 0.321 (IQR 0.174 - 0.607). Using a fixed decision threshold of 0.5, sensitivity was 0.637 and specificity was 0.851 (false positive rate 0.149).

Subgroup analyses indicated comparable discrimination across sex and age strata (Figure 4C). The AUC was 0.825 in men (n=1800; 95% CI 0.800 - 0.850) and 0.795 in women (n=110; 95% CI 0.721 - 0.870). By age group, the AUC was 0.783 in participants aged 21 - 45 years (n=1143; 95% CI 0.755 - 0.811) and 0.835 in those aged 46 - 65 years (n=767; 95% CI 0.807 - 0.863). Calibration gaps (predicted minus observed engagement probability) were −0.037 in men, −0.161 in women, −0.090 in the 21 - 45 year group, and −0.018 in the 46 - 65 year group (Figure 4D).

Permutation-based SHAP analysis identified lifestyle-related variables as the most influential predictors of engagement (Figure 4E). The largest mean absolute SHAP values corresponded to a balanced diet (occasional and regular; mean |SHAP|=0.047 for both), tobacco use (yes: 0.039; never: 0.021; former smoker: 0.018), and regular physical activity (mean |SHAP|=0.039). BMI category (obesity) showed mean |SHAP|=0.021. Mean signed SHAP values indicated variable directionality across predictors (Figure 4F). Regular physical activity had a mean signed SHAP of −0.0075 with a proportion of positive contributions of 0.484. A balanced diet showed mean signed SHAP values of −0.0037 (occasional) and −0.0038 (regular), with positive contribution proportions of 0.343. Marital status (single) showed a mean signed SHAP of +0.00345 with a positive contribution proportion of 0.726, despite a lower overall importance (mean |SHAP|=0.0173).

Discussion

Principal Findings

This study provides descriptive evidence that long-term digital follow-up of PrEP users through integrated ePROMs and ePREMs is operationally feasible and that participants who remained engaged in the program reported stable psychological well-being, experiential outcomes, and HRQoL over 24 months. Across follow-up, no clinically relevant fluctuations were observed in anxiety, depression, treatment satisfaction, or HRQoL domains, suggesting that the cohort, as observed, did not experience deterioration in emotional or quality-of-life outcomes. In parallel, the application of a machine learning–based framework enabled detailed characterization of digital engagement patterns at the questionnaire level, identifying sociodemographic, behavioral, and lifestyle-related factors associated with sustained participation.

Together, these findings address the study objectives by demonstrating the feasibility of digitally supported PrEP monitoring using ePROMs/ePREMs, and by illustrating how predictive analytics can complement descriptive clinical data to better understand engagement dynamics in real-world preventive care.

Clinical Outcomes, HRQoL, and Patient-Reported Experiences

When HRQoL outcomes were examined, participants exhibited stable psychological and HRQoL profiles throughout the monitoring period. The lack of clinically relevant fluctuations across anxiety, depression, satisfaction, and HRQoL domains indicates that the cohort, as observed, did not experience deterioration in emotional well-being over time. These findings are consistent with prior reports suggesting psychological stability among PrEP users [1,2], although comparisons should be interpreted cautiously given differences in study design.

Across PROMIS-29 domains, mean T-scores remained close to the normative population range, indicating preserved HRQoL throughout follow-up. Emotional domains (depression, anxiety, fatigue, and sleep disturbance) showed values within expected ranges, in line with the consistently low and stable HADS scores observed across all visits (normal range 0 - 7). These descriptive patterns are consistent with literature characterizing PrEP users as generally psychosocially stable populations engaged in care [3].

Treatment satisfaction, assessed using the TSQM, remained stable throughout follow-up, in line with earlier observational findings reporting high satisfaction with effectiveness, convenience, and tolerability among PrEP users [4,5]. In addition, perceived person-centered coordinated care, measured with the P3CEQ, showed a modest, nonsignificant upward trend, potentially reflecting increasing familiarity with the digital platform or continuity of interactions with health care professionals, although no inference can be made regarding the effect of digital care on these perceptions.

Correlation analyses among patient-reported outcomes revealed coherent interdomain associations at baseline and at 24 months. HADS and PROMIS-29 emotional domains showed moderate relationships (eg, r=0.58), and treatment satisfaction demonstrated inverse associations with psychological distress (r=–0.61). By 24 months, these associations were slightly stronger, although the observational design precludes interpretation regarding temporal causality or directional effects.

Clinical and Organizational Value of Digital Monitoring

Importantly, the clinical value of the digital monitoring framework in this study does not lie in detecting large group-level changes—which were not observed—but rather in enabling individual-level risk stratification and care reconfiguration that would not have been feasible under standard visit-based follow-up. The NAVETA platform incorporates a configurable clinical dashboard, conceptually analogous to

XSL•FO
RenderX

established systems such as AmbuFlex, including automated alerts for ePROMs scores deviating from predefined normative ranges [35,49]. This functionality allows clinicians to rapidly identify individuals requiring closer attention while maintaining remote oversight of stable patients.

In practice, this digital visibility supported a shift from a rigid, hospital-centric follow-up model toward a responsive, risk-based care pathway. For individuals showing stable ePROMs profiles across physical and psychological domains, clinicians were able to safely reduce routine in-person visits and transition follow-up to a predominantly digital format. This, in turn, enabled decentralized medication dispensing, allowing stable PrEP users to collect treatment at their nearest primary care center rather than attending hospital pharmacy services. Conversely, when moderate or clinically relevant deviations in ePROMs domains were detected at the individual level, the system facilitated expedited telephone or in-person assessments, ensuring timely clinical review despite the absence of statistically significant changes at the group level.

Although this study was not designed to formally quantify the frequency or outcomes of such individual-level interventions, the findings illustrate how clinically meaningful signals can emerge at the patient level even when aggregate trajectories remain stable, highlighting a limitation of relying exclusively on group-level statistical significance to judge clinical utility.

From an implementation perspective, the digital follow-up model was designed to minimize operational burden. Questionnaire administration and scoring were automated, with staff effort concentrated primarily on patient onboarding and periodic review of flagged cases. While no formal cost-effectiveness analysis was conducted, the platform enabled avoidance of unnecessary hospital visits, reduction in patient travel and work absenteeism, and more efficient allocation of clinician time toward individuals with complex needs. These practical efficiencies are supported by consistently high satisfaction ratings regarding the fully telemedicine-based follow-up model, as captured through NAVETA-specific ePROMs and ePREMs.

Taken together, the present findings suggest that the primary added value of intensive digital monitoring in PrEP care lies not in altering average clinical outcomes over time, but in optimizing care delivery, resource allocation, and patient convenience while maintaining clinical safety.

Digital Engagement Patterns and Machine Learning Analysis

While these descriptive analyses characterize overall clinical stability, understanding patterns of digital engagement requires analytical approaches capable of capturing multivariable and nonlinear relationships across repeated observations.

Nonparametric comparisons of engagement across sociodemographic and behavioral variables revealed marked gradients in participation patterns. Men, older participants, individuals reporting healthier lifestyle habits, and those with higher educational attainment or active employment status tended to exhibit greater continuity in digital engagement. These patterns are consistent with previous studies highlighting structural, behavioral, and socioeconomic determinants of participation in digital health interventions [7,8]. However, given the observational design, these associations should be interpreted strictly as correlational and not as evidence of causal effects.

To further characterize the multivariable relationships underlying engagement, we applied the ALGOPROMIA-Classification machine learning framework. Ensemble-based models, including HGB and RF, demonstrated good discriminatory performance, with AUC values around 0.82, indicating a consistent ability to distinguish between engaged and less engaged questionnaire-level observations. Notably, all evaluated algorithms showed very similar performance across discrimination and classification metrics. This convergence likely reflects the structure of the data, characterized by class imbalance and a relatively limited number of dominant predictors [50]. Under these conditions, both linear and nonlinear models can capture the main signal once appropriate regularization, oversampling within training folds, and repeated cross-validation are applied. Consequently, model performance appears to be driven more by the underlying data patterns than by algorithmic complexity per se. The small numerical differences observed between models should therefore be interpreted cautiously, as they primarily reflect averaging across repeated cross-validation folds and rounding precision rather than substantively meaningful differences in predictive capability [51].

On the basis of the comparative model evaluation, the RF algorithm was selected for downstream analyses. This choice was not driven by superior point estimates of discrimination or accuracy, as all evaluated algorithms showed highly comparable performance across metrics. Instead, RF was retained to ensure methodological consistency with previous applications of the ALGOPROMIA framework in related clinical contexts, and because it offers a favorable balance between robustness, interpretability, and stability under class imbalance and repeated cross-validation. Given the convergence in performance across models, this selection reflects a pragmatic and reproducible modeling strategy rather than an optimization-driven decision [52,53].

Model behavior was subsequently examined in greater depth using the full set of questionnaire-level observations, focusing on temporal stability of predicted versus observed engagement, score separation, subgroup-level discrimination and calibration, and permutation-based SHAP explainability (Figure 4). Across monitoring stages, the mean predicted probability of engagement remained relatively stable over time. The average difference between observed and predicted engagement rates across all stages was −0.022, indicating a small overall deviation between model estimates and empirical data. In contrast, observed engagement rates exhibited greater variability (range 0.430 - 0.846). While prediction and observation were closely aligned at several time points (notably at 6 and 24 mo), larger discrepancies were observed at baseline and at 15 months. Despite the overall temporal stability observed, engagement dynamics at specific monitoring stages remain to be fully characterized.

XSL•FO
RenderX

Analysis of predicted score distributions demonstrated partial separation between engaged and nonengaged observations, with an AUC of 0.816 indicating good overall discrimination. However, the operating characteristics of the model revealed an asymmetric performance profile. When applying a fixed decision threshold of 0.5, the classifier prioritized specificity over sensitivity, resulting in high specificity (0.851) and positive predictive value, but more modest sensitivity (0.637) and NPV. From a clinical and implementation perspective, this performance profile is not necessarily optimal for digital follow-up programs, where the primary objective is often the early identification of individuals at risk of disengagement. In this context, false negatives—participants predicted as engaged who subsequently fail to respond—are more consequential than false positives, as they represent missed opportunities for timely intervention.

This asymmetric performance likely reflects both the underlying class imbalance and the limited availability of informative longitudinal behavioral signals early in follow-up. Because engagement is a dynamic and evolving process, richer temporal information—capturing changes in interaction patterns, response latency, and behavioral trajectories over time—may be required to improve sensitivity for disengagement prediction. Accordingly, these findings support the need for iterative model updating as additional longitudinal data become available.

Subgroup analyses showed broadly comparable discriminatory performance across sex and age strata. Although AUC values differed numerically between groups, the CIs overlapped substantially, providing no evidence of systematic performance degradation in any specific subgroup. Calibration gaps were generally small and negative, indicating a tendency toward slight underestimation of engagement probability, particularly among women and younger participants. However, the limited number of women reduces statistical precision in sex-stratified analyses and constrains the generalizability of subgroup-specific estimates, warranting confirmation in more sex-balanced PrEP cohorts [35,54-56].

Explainability and Interpretation of SHAP Results

Permutation-based SHAP analysis provided further insight into the drivers of model predictions. Consistent with prior research in telemedicine and digital health engagement, lifestyle-related variables emerged as the most influential contributors to sustained participation [57,58]. In the present model, indicators of balanced diet, tobacco use, and physical activity exhibited the largest mean absolute SHAP values, highlighting their relative importance for engagement prediction at the questionnaire level. However, mean absolute SHAP values reflect relative importance within the model rather than clinically meaningful changes in engagement probability. SHAP values quantify feature contributions on the model's internal scale and are intended for comparative interpretation across variables, not for direct translation into probability differences or effect sizes.

Importantly, mean signed SHAP values were frequently close to zero, indicating substantial heterogeneity in the direction of feature effects across individuals [59]. For example, regular physical activity showed high global importance but mixed directional influence, contributing positively in some observations and negatively in others. In contrast, certain predictors with lower overall importance—such as marital status (single)—displayed more consistent directional effects.

Taken together, these patterns align with existing evidence linking health-related behaviors to adherence and continuity in digital follow-up programs, while underscoring that such associations should be interpreted strictly as model-based and noncausal. The relatively small magnitude of individual SHAP contributions does not imply that engagement behavior is stochastic or unpredictable. Rather, it reflects the multifactorial and context-dependent nature of digital engagement, particularly when modeled at the questionnaire level [60]. Engagement decisions are shaped by a combination of stable individual characteristics and dynamic situational determinants not explicitly captured in the present model.

Importantly, this work is subject to several limitations inherent to its observational, single-arm design. In the absence of a control group or structured pre–post comparisons, no causal inferences can be drawn regarding the effects of PrEP use or the NAVETA digital monitoring system on psychological, experiential, or HRQoL outcomes. The stable trajectories observed may reflect baseline characteristics of individuals who initiated PrEP or who remained engaged in follow-up, rather than effects attributable to digital monitoring itself. Selection bias and a potential healthy-user effect must therefore be considered, particularly in light of the 19.8% discontinuation rate observed during follow-up.

A key methodological consideration relates to missing data and its potential influence on longitudinal interpretations. Completion rates decreased over the scheduled assessment stages, as expected in a real-world telemedicine context [61]. Although exploratory analyses did not indicate systematic differences in baseline HRQoL or psychological status between participants with high and low questionnaire completion, residual bias cannot be fully excluded. The voluntary nature of ePROMs/ePREMs completion implies that observed longitudinal patterns may over-represent individuals who are more digitally engaged or motivated, limiting generalizability.

Finally, engagement modeling was performed at the questionnaire level and did not incorporate richer temporal or contextual variables that may influence response behavior, such as transient life events or short-term changes in perceived relevance of questionnaires.

Conclusions

Taken together, the present findings indicate that digitally supported PrEP follow-up using integrated ePROMs and ePREMs is feasible in routine clinical practice and can support stable long-term monitoring of psychological well-being, HRQoL, and treatment satisfaction. Beyond descriptive clinical stability, the combination of patient-reported data and explainable machine learning provides actionable insight into engagement patterns, enabling more adaptive, resource-efficient, and person-centered models of preventive care.

While causal effects cannot be inferred, this study demonstrates how digital monitoring infrastructures can be leveraged to optimize care delivery, prioritize clinical attention where needed,

XSL•FO
RenderX

and generate high-resolution engagement data to inform future intervention strategies. These results establish a foundation for subsequent comparative, interventional, and cost-effectiveness studies aimed at evaluating the broader impact of digital monitoring on PrEP care pathways.

Acknowledgments

The authors gratefully acknowledge the study participants, who kindly allowed the use of their anonymized data, as well as the health care professionals at the participating sites who contributed to the study. The authors also thank Lidesec for preliminary editorial assistance provided during the early stages of manuscript preparation. The authors declare the limited use of a generative artificial intelligence (AI)–assisted language editing tool (Grammarly) exclusively for proofreading and stylistic refinement during manuscript preparation, under full human supervision. According to the GAIDeT taxonomy (2025), the task delegated to generative AI was limited to proofreading and editing. No generative AI tools were used for text generation, data generation, statistical analysis, interpretation of results, formulation of conclusions, or selection of references. Full responsibility for the scientific content and the final version of the manuscript rests entirely with the authors. Generative AI tools are not listed as authors and bear no responsibility for the content of this work.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Authors' Contributions

SHP and GMO conceptualized the study. SHP performed data curation, statistical analysis, and developed the ALGOPROMIA predictive model. GMO contributed to the initial draft preparation, provided critical revisions, and was responsible for funding acquisition. JIS, TM, LV, and AC contributed to the review and editing of the final manuscript and provided domain-specific expertise for interpretation and validation. All authors read and approved the final version of the manuscript.

Conflicts of Interest

The authors declare no personal financial competing interests. NAVETA is a licensed digital health platform used within the Spanish National Health System for routine clinical care and research purposes. The authors do not own stocks, stock options, or other financial interests in the company holding the license, nor do they receive personal financial benefit derived from the commercial exploitation of the platform.

Some authors are involved in the academic and clinical use of the NAVETA platform within publicly funded health care and research settings. Their involvement is limited to clinical implementation, evaluation, and research activities and does not include ownership, commercial development, or profit-sharing related to the licensed software.

Multimedia Appendix 1

Participant-level completion trajectories and observation counts across scheduled ePROM and ePREM assessment stages during 24-month digital follow-up.

[DOCX File, 156 KB - jmir_v28i1e87592_app1.docx]

References

1. Takada-de-Souza MK, Picone CDM, Avelino-Silva VI, Freitas AC, Segurado AC. Impact of HIV pre-exposure prophylaxis (PREP) on health-related quality of life (QOL) of patients followed up at a reference center in São Paulo, Brazil. Clinics (Sao Paulo) 2024 Jan;79:100419. [doi: 10.1016/j.clinsp.2024.100419]
2. Commission implementing decision c(2016) 5439 final of 18.8.2016 amending the marketing authorisation granted by decision c(2005)456 for “truvada – emtricitabine / tenofovir disoproxil fumarate”, a medicinal product for human use. : European Commission; 2016 URL: https://ec.europa.eu/health/documents/community-register/2016/20160818135700/dec_135700_en.pdf
3. Dourado I, Soares F, Magno L, et al. Adherence, safety, and feasibility of HIV pre-exposure prophylaxis among adolescent men who have sex with men and transgender women in Brazil (PrEP1519 Study). J Adolesc Health 2023 Dec;73(6S):S33-S42. [doi: 10.1016/j.jadohealth.2023.09.005] [Medline: 37953006]
4. Marcus JL, Hurley LB, Hare CB, et al. Preexposure prophylaxis for HIV prevention in a large integrated health care system: adherence, renal safety, and discontinuation. J Acquir Immune Defic Syndr 2016;73(5):540-546. [doi: 10.1097/QAI.0000000000001129]
5. Tanner MR, Miele P, Carter W, et al. Preexposure prophylaxis for prevention of HIV acquisition among adolescents: clinical considerations, 2020. MMWR Recomm Rep 2020 Apr 24;69(3):1-12. [doi: 10.15585/mmwr.rr6903a1] [Medline: 32324724]

XSL•FO
RenderX

6. Ferrario L, Foglia E, Garagiola E, et al. The impact of PrEP: results from a multicenter health technology assessment into the Italian setting. J Prev Med Hyg 2020 Sep;61(3):E451-E463. [doi: 10.15167/2421-4248/jpmh2020.61.3.1352] [Medline: 33150233]
7. Li J, Dufrene SL, Okulicz JF. Systemic preexposure prophylaxis for HIV: translating clinical data to clinical practice. Ann Pharmacother 2014 Apr;48(4):507-518. [doi: 10.1177/1060028014520880] [Medline: 24473492]
8. Assi L, Chamseddine F, Ibrahim P, et al. A global assessment of eye health and quality of life. JAMA Ophthalmol 2021 May 1;139(5):526. [doi: 10.1001/jamaophthalmol.2021.0146]
9. Gaskins M, Sammons MK, Kutscha F, Nast A, Werner RN. Factors that motivate men who have sex with men in Berlin, Germany, to use or consider using HIV pre-exposure prophylaxis-a multi-methods analysis of data from a multicentre survey. PLoS One 2021;16(11):e0260168. [doi: 10.1371/journal.pone.0260168] [Medline: 34793575]
10. Menditto E, Cahir C, Malo S, et al. Persistence as a robust indicator of medication adherence-related quality and performance. Int J Environ Res Public Health 2021 May 3;18(9):4872. [doi: 10.3390/ijerph18094872] [Medline: 34063641]
11. Sidebottom D, Ekström AM, Strömdahl S. A systematic review of adherence to oral pre-exposure prophylaxis for HIV - how can we improve uptake and adherence? BMC Infect Dis 2018 Nov 16;18(1):581. [doi: 10.1186/s12879-018-3463-4] [Medline: 30445925]
12. Brisson J, Castro-Arteaga M, Apedaile D, Perez-Brumer A. Enhancing daily oral PrEP adherence with digital communications: protocol for a systematic review and meta-analysis. In: Ogunbajo A, editor. PLOS ONE 2024;19(11):e0313322. [doi: 10.1371/journal.pone.0313322] [Medline: 39531430]
13. O Murchu E, Marshall L, Teljeur C, et al. Oral pre-exposure prophylaxis (PrEP) to prevent HIV: a systematic review and meta-analysis of clinical effectiveness, safety, adherence and risk compensation in all populations. BMJ Open 2022 May 11;12(5):e048478. [doi: 10.1136/bmjopen-2020-048478] [Medline: 35545381]
14. Allison BA, Widman L, Stewart JL, Evans R, Perry M. Adherence to pre-exposure prophylaxis in adolescents and young adults: a systematic review and meta-analysis. J Adolesc Health 2022 Jan;70(1):28-41. [doi: 10.1016/j.jadohealth.2021.04.001] [Medline: 34059426]
15. D'Anza B, Pronovost PJ. Digital health: unlocking value in a post-pandemic world. Popul Health Manag 2022 Feb;25(1):11-22. [doi: 10.1089/pop.2021.0031] [Medline: 34042532]
16. Tuckson RV, Edmunds M, Hodgkins ML. Telehealth. N Engl J Med 2017 Oct 19;377(16):1585-1592. [doi: 10.1056/NEJMsr1503323]
17. Bull C, Teede H, Watson D, Callander EJ. Selecting and implementing patient-reported outcome and experience measures to assess health system performance. JAMA Health Forum 2022 Apr 1;3(4):e220326. [doi: 10.1001/jamahealthforum.2022.0326] [Medline: 36218960]
18. Bull C, Callander EJ. Current PROM and PREM use in health system performance measurement: still a way to go. Patient Exp J 2022;9(1):12-18. [doi: 10.35680/2372-0247.1664]
19. Du J, Jin G, Zhang H, et al. Effectiveness of digital health interventions in promoting the pre-exposure prophylaxis (PrEP) care continuum among men who have sex with men (MSM): a systematic review of randomized controlled trials. Curr HIV/AIDS Rep 2025 Dec;22(1). [doi: 10.1007/s11904-025-00733-4]
20. Saberi P, Dawson Rose C, Wootton AR, et al. Use of technology for delivery of mental health and substance use services to youth living with HIV: a mixed-methods perspective. AIDS Care 2020 Aug;32(8):931-939. [doi: 10.1080/09540121.2019.1622637] [Medline: 31132864]
21. Gandhi RT, Bedimo R, Hoy JF, et al. Antiretroviral drugs for treatment and prevention of HIV infection in adults: 2022 recommendations of the international antiviral society-USA panel. JAMA 2023 Jan 3;329(1):63-84. [doi: 10.1001/jama.2022.22246] [Medline: 36454551]
22. Monroe AK, Jabour SM, Peña S, et al. A qualitative study examining the benefits and challenges of incorporating patient-reported outcome substance use and mental health questionnaires into clinical practice to improve outcomes on the HIV care continuum. BMC Health Serv Res 2018 Dec;18(1). [doi: 10.1186/s12913-018-3203-x]
23. Adler L, Vered S, Amran MM, et al. The effect of telemedicine on preventive medicine- a case from Israel. Isr J Health Policy Res 2025 Jun 2;14(1):31. [doi: 10.1186/s13584-025-00696-2] [Medline: 40457339]
24. Chen F, Su W, Becker SH, et al. Clinical and economic impact of a digital, remotely-delivered intensive behavioral counseling program on medicare beneficiaries at risk for diabetes and cardiovascular disease. PLoS ONE 2016;11(10):e0163627. [doi: 10.1371/journal.pone.0163627]
25. Neil J, Mao B, Shao R, et al. A pilot randomized clinical trial of a smartphone-based application to support at-home PSA screening and culturally tailored prostate cancer education for African American men: a study protocol. Contemp Clin Trials 2024 Dec;147:107737. [doi: 10.1016/j.cct.2024.107737]
26. Xin M, Viswanath K, Li AC, et al. The effectiveness of electronic health interventions for promoting HIV-preventive behaviors among men who have sex with men: meta-analysis based on an integrative framework of design and implementation features. J Med Internet Res 2020 May 25;22(5):e15977. [doi: 10.2196/15977] [Medline: 32449685]
27. Agbo CE, Agu MC, Ogieuhi IJ, et al. The role of digital tools in enhancing antiretroviral therapy adherence among adolescents and young adults with HIV: a systematic review of randomized controlled trials. BMC Infect Dis 2025 Mar 13;25(1):353. [doi: 10.1186/s12879-025-10724-w] [Medline: 40082764]

XSL•FO
RenderX

28. Montalescot L, Baussard L, Charbonnier E. Factors associated with digital intervention engagement and adherence in patients with cancer: systematic review. J Med Internet Res 2024 Dec 11;26:e52542. [doi: 10.2196/52542] [Medline: 39661976]
29. Zhang Y, Flannery M, Zhang Z, et al. Digital health psychosocial intervention in adult patients with cancer and their families: systematic review and meta-analysis. JMIR Cancer 2024 Feb 5;10:e46116. [doi: 10.2196/46116] [Medline: 38315546]
30. Ramírez López GA, Gómez Galicia DL, Zagal Jiménez TX, Toledano Jaimes CD, Díaz de León Castañeda C. Salud digital para promover la adherencia al tratamiento antirretroviral en pacientes con VIH/sida: metarevisión. Farm Hosp 2024 Sep;48(5):252-258. [doi: 10.1016/j.farma.2024.02.005]
31. Jongbloed K, Parmar S, van der Kop M, Spittal PM, Lester RT. Recent evidence for emerging digital technologies to support global HIV engagement in care. Curr HIV/AIDS Rep 2015 Dec;12(4):451-461. [doi: 10.1007/s11904-015-0291-7] [Medline: 26454756]
32. Nturibi E, Mecha J, Kaara F, et al. Feasibility and acceptability of a digital health portal to improve HIV care engagement among Kenyan youth: mixed methods study. JMIR Form Res 2025 Jul 7;9:e59661. [doi: 10.2196/59661] [Medline: 40704659]
33. Herrmann S, Power B, Rashidi A, et al. Supporting patient-clinician interaction in chronic HIV care: design and development of a patient-reported outcomes software application. J Med Internet Res 2021 Jul 30;23(7):e27861. [doi: 10.2196/27861] [Medline: 34328442]
34. Halilaj I, Lieverse R, Oberije C, et al. Improving patient engagement in phase 2 clinical trials with a trial-specific patient decision aid: development and usability study. J Med Internet Res 2025 Sep 10;27:e71817. [doi: 10.2196/71817] [Medline: 40931488]
35. Mercadal-Orfila G, Herrera-Pérez S, Piqué N, et al. Implementing systematic patient-reported measures for chronic conditions through the Naveta value-based telemedicine initiative: observational retrospective multicenter study. JMIR Mhealth Uhealth 2024 Jun 28;12:e56196. [doi: 10.2196/56196] [Medline: 38545697]
36. Mercadal-Orfila G, Serrano López de Las Hazas J, Riera-Jaume M, Herrera-Perez S. Developing a prototype machine learning model to predict quality of life measures in people living with HIV. Integr Pharm Res Pract 2025;14:1-16. [doi: 10.2147/IPRP.S492422] [Medline: 39872224]
37. Mascort J, Carrillo R, Alastrue I, et al. Profilaxis pre-exposición de la infección por el VIH y Atención Primaria (AP). Atención Primaria 2020 Mar;52(3):137-139. [doi: 10.1016/j.aprim.2020.02.001]
38. Liberato ACS, São João TM, Jannuzzi FF, Landaas EJ, Wongchareon K, Rodrigues RCM. Treatment satisfaction questionnaire for medication (TSQM version 1.4): ceiling and floor effects, reliability, and known-group validity in Brazilian outpatients with hypertension. Value Health Reg Issues 2020 Dec;23:150-156. [doi: 10.1016/j.vhri.2020.07.578] [Medline: 33158803]
39. Hays RD, Spritzer KL, Schalet BD, Cella D. PROMIS®-29 v2.0 profile physical and mental health summary scores. Qual Life Res 2018 Jul;27(7):1885-1891. [doi: 10.1007/s11136-018-1842-3]
40. Lloyd HM, Fosh B, Whalley B, Byng R, Close J. Validation of the person-centred coordinated care experience questionnaire (P3CEQ). Int J Qual Health Care 2019 Aug 1;31(7):506-512. [doi: 10.1093/intqhc/mzy212] [Medline: 29590365]
41. Chen C, Yang D, Gao S, et al. Development and performance assessment of novel machine learning models to predict pneumonia after liver transplantation. Respir Res 2021 Dec;22(1):94. [doi: 10.1186/s12931-021-01690-3]
42. Bocéréan C, Dupret E. A validation study of the hospital anxiety and depression scale (HADS) in a large sample of French employees. BMC Psychiatry 2014 Dec 16;14(1):354. [doi: 10.1186/s12888-014-0354-0] [Medline: 25511175]
43. Snijkers JTW, van den Oever W, Weerts ZZRM, et al. Examining the optimal cutoff values of HADS, PHQ-9 and GAD-7 as screening instruments for depression and anxiety in irritable bowel syndrome. Neurogastroenterol Motil 2021 Dec;33(12):e14161. [doi: 10.1111/nmo.14161] [Medline: 33938601]
44. PROMIS\textsuperscript® score cut points. HealthMeasures. 2024. URL: https://www.healthmeasures.net/score-and-interpret/interpret-scores/promis/promis-score-cut-points [accessed 2026-01-28]
45. Karri R, Chen YP, Drummond KJ. Using machine learning to predict health-related quality of life outcomes in patients with low grade glioma, meningioma, and acoustic neuroma. PLoS ONE 2022;17(5):e0267931. [doi: 10.1371/journal.pone.0267931] [Medline: 35507629]
46. Kaliappan J, Bagepalli AR, Almal S, Mishra R, Hu YC, Srinivasan K. Impact of cross-validation on machine learning models for early detection of intrauterine fetal demise. Diagnostics (Basel) 2023 May 10;13(10):1692. [doi: 10.3390/diagnostics13101692] [Medline: 37238178]
47. Mercadal-Orfila G, López Sánchez P, Pou Alonso A, et al. TELEPROM psoriasis: enhancing patient-centered care and health-related quality of life (HRQoL) in moderate-to-severe plaque psoriasis. Front Med 2024 Dec 10;11:1465725. [doi: 10.3389/fmed.2024.1465725]
48. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020 Jan;2(1):56-67. [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
49. Schougaard LMV, Larsen LP, Jessen A, et al. AmbuFlex: tele-patient-reported outcomes (telePRO) as the basis for follow-up in chronic and malignant diseases. Qual Life Res 2016 Mar;25(3):525-534. [doi: 10.1007/s11136-015-1207-0] [Medline: 26790427]

XSL•FO
RenderX

50. Luijken K, Groenwold RHH, Van Calster B, Steyerberg EW, van Smeden M. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: a measurement error perspective. Stat Med 2019 Aug 15;38(18):3444-3459. [doi: 10.1002/sim.8183] [Medline: 31148207]
51. Bates S, Hastie T, Tibshirani R. Cross-validation: what does it estimate and how well does it do it? J Am Stat Assoc 2024;119(546):1434-1445. [doi: 10.1080/01621459.2023.2197686] [Medline: 39308484]
52. Belbasis L, Panagiotou OA. Reproducibility of prediction models in health services research. BMC Res Notes 2022 Jun 11;15(1):204. [doi: 10.1186/s13104-022-06082-4] [Medline: 35690767]
53. Peterson RA, McGrath M, Cavanaugh JE. Can a transparent machine learning algorithm predict better than its black box counterparts? a benchmarking study using 110 data sets. Entropy (Basel) 2024 Aug 31;26(9):746. [doi: 10.3390/e26090746] [Medline: 39330080]
54. Wong H, Razvi Y, Hamid MA, Mistry N, Filler G. Age and sex-related comparison of referral-based telemedicine service utilization during the COVID-19 pandemic in Ontario: a retrospective analysis. BMC Health Serv Res 2023 Dec 7;23(1):1374. [doi: 10.1186/s12913-023-10373-2] [Medline: 38062437]
55. Luo J, Tong L, Crotty BH, et al. Telemedicine adoption during the COVID-19 pandemic: gaps and inequalities. Appl Clin Inform 2021 Aug;12(4):836-844. [doi: 10.1055/s-0041-1733848] [Medline: 34496419]
56. Odebunmi OO, Hughes TD, Waters AR, et al. Findings from a national survey of older US adults on patient willingness to use telehealth services: cross-sectional survey. J Med Internet Res 2024 May 23;26:e50205. [doi: 10.2196/50205] [Medline: 38780994]
57. Crane N, Hagerman C, Horgan O, Butryn M. Patterns and predictors of engagement with digital self-monitoring during the maintenance phase of a behavioral weight loss program: quantitative study. JMIR Mhealth Uhealth 2023 Jul 18;11:e45057. [doi: 10.2196/45057] [Medline: 37463017]
58. Coley N, Andre L, Hoevenaar-Blom MP, et al. Factors predicting engagement of older adults with a coach-supported ehealth intervention promoting lifestyle change and associations between engagement and changes in cardiovascular and dementia risk: secondary analysis of an 18-month multinational randomized controlled trial. J Med Internet Res 2022 May 9;24(5):e32006. [doi: 10.2196/32006] [Medline: 35385395]
59. Lamane H, Mouhir L, Moussadek R, Baghdad B, Kisi O, El Bilali A. Interpreting machine learning models based on SHAP values in predicting suspended sediment concentration. Int J Sediment Res 2025 Feb;40(1):91-107. [doi: 10.1016/j.ijsrc.2024.10.002]
60. Idrees AR, Beierle F, Mutter A, et al. Engagement analysis of a persuasive-design-optimized eHealth intervention through machine learning. Sci Rep 2024 Sep 13;14(1):21427. [doi: 10.1038/s41598-024-72162-z] [Medline: 39271759]
61. Mosenia A, Li P, Seefeldt R, Seitzman GD, Sun CQ, Kim TN. Longitudinal use of telehealth during the COVID-19 pandemic and utility of asynchronous testing for subspecialty-level ophthalmic care. JAMA Ophthalmol 2023 Jan 1;141(1):56-61. [doi: 10.1001/jamaophthalmol.2022.4984] [Medline: 36454548]

Abbreviations

AUC: area under the curve
eGFR: estimated glomerular filtration rate
ePREM: electronic patient-reported experience measure
ePROM: electronic patient-reported outcome measure
HADS: Hospital Anxiety and Depression Scale
HGB: histogram-based gradient boosting
HRQoL: health-related quality of life
NPV: negative predictive value
P3CEQ: Person-Centered Coordinated Care Experience Questionnaire
PrEP: pre-exposure prophylaxis
PROMIS-29: Patient-Reported Outcome Measurement Information System Profile-29
RF: random forest
SHAP: Shapley Additive Explanations
TSQM: Treatment Satisfaction Questionnaire for Medication

XSL•FO
RenderX

Edited by A Mavragani; submitted 11.Nov.2025; peer-reviewed by A Famotire, O Akinpeloye; revised version received 28.Jan.2026; accepted 03.Mar.2026; published 13.Apr.2026.

Please cite as:
Mercadal-Orfila G, Serrano JI, Mijares T, Vidal L, Curran A, Herrera-Pérez S
Digital Monitoring of Pre-Exposure Prophylaxis Users Through Electronic Patient-Reported Outcome Measures and Electronic Patient-Reported Experience Measures: Multicenter Prospective Study on Feasibility, Safety, and Predictive Modeling of Digital Engagement
J Med Internet Res 2026;28:e87592
URL: https://www.jmir.org/2026/1/e87592
doi:10.2196/87592

© Gabriel Mercadal-Orfila, Joaquín Ignacio Serrano, Tilman Mijares, Lluís Vidal, Adrian Curran, Salvador Herrera-Pérez. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Accuracy of Visual Inspection Alone to Assess Joint Effusions of the Hand: Cross-Sectional Study

Carrie Ye[1,2], MPH, MD; J Ross Mitchell[1,3], MSc, PhD; Shahad Al-Matar[1], MBBS; Mena Bishay[1], MD; Niall Jones[1], MD; Dalton Sholter[1], MD; Dylan Johnson[1], PhD, MD; Steven Katz[1], MD; Anthony S Russell[1], MD; Mohamed Abdalla[1,3], MSc, PhD

[1]Department of Medicine, Faculty of Medicine & Dentistry, University of Alberta, 8-130 Clinical Sciences Building, 11350 83 Ave NW, Edmonton, AB, Canada

[2]Arthritis Research Canada, Vancouver, BC, Canada

[3]Alberta Machine Intelligence Institute, Edmonton, AB, Canada

Corresponding Author:
Carrie Ye, MPH, MD
Department of Medicine, Faculty of Medicine & Dentistry, University of Alberta, 8-130 Clinical Sciences Building, 11350 83 Ave NW, Edmonton, AB, Canada

Abstract

Background: Physical examination is the cornerstone of diagnosing and monitoring inflammatory arthritis, with the detection of joint effusions being one of the most crucial components of the examination. Rheumatologists largely rely on palpation, supported by other examination techniques, such as evaluating the range of motion and visual inspection, to assess signs of joint swelling. However, remote care is on the rise in rheumatology, and aside from visual inspection, physical examination of joints is limited during telehealth visits. The ability of rheumatologists to accurately detect hand synovitis through photos or videos of the hands, without the benefit of direct tactile examination, is currently unknown.

Objective: This study aimed to assess the accuracy of detecting joint effusions of the hands by remote visual evaluation alone.

Methods: We conducted a prospective cohort study of patients assessed by a rheumatologist in Edmonton, Alberta. Participants were assessed clinically by rheumatologists for the presence of effusions of the metacarpophalangeal (MCP) and proximal interphalangeal (PIP) joints between June 13, 2024, and October 29, 2024. Participants had photos and videos of their hands taken by study staff on the same day. Photos and videos were remotely assessed separately by 4 rheumatologists not involved in the initial clinical assessments for the presence of MCP and PIP joint effusions and compared with the clinical assessments.

Results: The study cohort included 156 patients (mean age 53, SD 14.2 years, n=105, 67.3% were female and n=52, 33.3% had rheumatoid arthritis), for a total of 3120 MCP and PIP joints. Effusions were identified in 12.8% (20/156) of patients and in 2.2% (69/3120) of joints per clinical assessment. The average visual assessment joint-level sensitivity and specificity of photos were 0.14 (95% CI 0.07 - 0.19) and 0.97 (95% CI 0.96 - 0.98), respectively. The average visual assessment joint-level sensitivity and specificity of videos were 0.24 (95% CI 0.13 - 0.33) and 0.98 (95% CI 0.97 - 0.99), respectively. The average person-level visual assessment sensitivity and specificity of photos were 0.44 (95% CI 0.30 - 0.60) and 0.82 (95% CI 0.76 - 0.87), respectively. The average person-level visual assessment sensitivity and specificity of videos were 0.48 (95% CI 0.35 - 0.60) and 0.84 (95% CI 0.79 - 0.89), respectively. Assessor agreement was poor (κ=0.12 - 0.17).

Conclusions: Visual inspection of photos and videos to detect MCP and PIP joint effusions was poor at both the joint and person levels. Patients and rheumatologists should be aware of these limitations when conducting remote telehealth assessments.

(J Med Internet Res 2026;28:e86261) doi:10.2196/86261

KEYWORDS

arthritis; rheumatoid arthritis; effusions; telehealth; rheumatology

Introduction

Rheumatic diseases include a wide range of diagnoses that affect the musculoskeletal system and, according to the World Health Organization, are the leading cause of worldwide disability [1]. Access to specialized rheumatology care remains a significant challenge for many individuals, with female patients and those residing in smaller centers experiencing longer wait times [2,3]. In many cases, patients face long wait times and limited availability of specialists, often resulting in delays in diagnosis and treatment, which can lead to irreparable damage [4]. The Arthritis Society of Canada recently released *The State of Arthritis in Canada Report Card*, which scored all provinces and territories a grade of C or lower (ie, satisfactory,

XSL•FO
RenderX

improvement needed, or significant improvement needed) for access to care, resulting in their top recommendation: "Improve access to arthritis care" [5]. The shortage of trained rheumatologists and the concentration of services in urban centers exacerbate the inequities in access to rheumatology care [5,6]. These barriers can lead to underdiagnosis and inadequate management of conditions such as rheumatoid arthritis, where timely intervention is crucial for preventing joint damage and improving patient outcomes [6-8].

In recent years, the shift toward remote care across medical specialties, including rheumatology, has gained momentum and was accelerated by the COVID-19 pandemic [9,10]. In March 2020, there was a 75.9% decrease in outpatient office visits, with nearly all visits completed by telemedicine, and by September 2021, half of patient encounters remained telemedicine visits [11]. The transition to telehealth appointments has provided a potential solution to address some of the access challenges in rheumatology [12]. An American study found that implementation of electronic consults in medical specialties led to a significant increase in the completion rate of specialist visits and a reduction in wait times [12]. Remote rheumatologist visits offer greater flexibility for patients who are unable to travel to rheumatology clinics. This model of care has the potential to improve health care equity by bridging geographical, economic, and gender gaps, particularly for individuals in underserved regions, those who cannot take time off of work to travel to appointments, or those who face childcare issues [2,13].

Traditionally, physical examination has been the cornerstone of diagnosing joint effusions and other musculoskeletal abnormalities in rheumatology [14]. Rheumatologists largely rely on palpation, supported by other examination techniques, such as evaluating the range of motion and visual inspection, to assess for signs of joint swelling [15]. However, aside from visual inspection, physical examination of joints is limited during telehealth visits. A recent scoping review found several studies that assessed the performance of unassisted physical examinations conducted over telehealth platforms, but no studies examined this from a rheumatology perspective, and it remains unclear how effective visual inspection alone is at detecting joint effusions during telehealth appointments [16]. The ability of rheumatologists to accurately detect hand synovitis, in particular, through photos or videos, without the benefit of direct tactile examination, is a critical question that has not been studied.

The objective of this study was to assess the accuracy with which rheumatologists can remotely detect hand joint effusions through images and videos of patients' hands. By understanding the limitations and potential of remote visual assessment alone, this research seeks to inform best practices for telehealth in rheumatology and ultimately enhance patient care in a more equitable and accessible manner.

Methods

Study Design

We conducted a prospective cohort study at 4 rheumatology practices in Edmonton, Alberta. Participants were assessed by rheumatologists in the clinic for the presence of effusions of the metacarpophalangeal (MCP) and proximal interphalangeal (PIP) joints. After providing individual written consent, participants filled out a short intake survey on baseline characteristics and had photos and videos of their hands taken by a member of the research team on the same day. Photos and videos were assessed separately by rheumatologists not involved in the initial clinic assessments.

Ethical Considerations

This study was approved by the research ethics board at the University of Alberta (Pro00141099). Informed consent was obtained for each participant. Each participant was assigned a unique study ID, and all study data were collected under this unique study ID, separate from the dataset that linked the study ID with the participant's identifying information obtained in the consent form. Participants did not receive any compensation.

Participants

Patients were recruited consecutively from 4 rheumatology clinics between June 13, 2024, and October 29, 2024. The inclusion criteria were being aged 18 years or older at the time of clinic appointment and being able to provide written consent. Patients with and without inflammatory arthritis were included. There were no exclusion criteria. Participants self-reported age, sex, diagnosis, duration of joint symptoms, height, and weight.

Images and Videos

Photos and videos were taken using an 8 MP digital camera on a light blue background. A total of four photos were taken of both hands: (1) palms down, fingers spread; (2) palms down, fists; (3) palms up, fingers spread; and (4) palms up, fists. A single video including both hands was taken of each participant. For the video, each participant was instructed to have their palms down, spread out their fingers, and then make a fist 3 times. They were then instructed to repeat these movements with the palms facing up. Photos and videos included a few centimeters proximal to the wrists, ensuring that all fingertips were included in the images.

Assessment of Joint Effusions

The reference standard labeling of joint effusions, defined as the accumulation of excess fluid in the joint cavity, was completed by the participant's rheumatologist, who assessed each MCP and PIP joint during the physical examination portion of their appointment. All MCP and PIP joints had to be assessed by the rheumatologists, but physical examination techniques were not standardized between rheumatologists. Distal interphalangeal joint findings were not obtained, as these joints are rarely involved in inflammatory arthritis, and we wished to minimize the time burden on physicians. Rheumatologists labeled each joint as positive or negative for MCP or PIP joint effusion and were instructed to label uncertain joints as negative. Each participant was assessed by 1 rheumatologist. The

XSL•FO
RenderX

rheumatologist was also asked to report whether there were visible deformities in the hands.

Asynchronously, 4 separate rheumatologists of varying experience (assessor 1: 1 year; assessor 2: 8 years; assessor 3: 15 years; and assessor 4: 50 years) were asked to assess the photos and videos on a computer screen (unlinked to participants and blinded to additional clinical information) for the presence of any MCP or PIP joint effusions (person-level assessment). If they felt there was MCP or PIP joint effusion present in any individual, they were asked to indicate the exact MCP and/or PIP joints in which they saw effusion (index test). Each evaluator was instructed to label each MCP and PIP joint as 0 (no effusion) or 1 (effusion) and to label uncertain cases as 0. No other specific training was provided to evaluators. Evaluators were allowed to view the images as many times and replay the videos as many times as they wished.

Analysis

The sample size required to achieve a sensitivity of 0.90 with a lower bound of 0.80, power of 0.8, and alpha of .05, with an estimated prevalence of 5% effused joints, was 2140 joints or 107 individuals (20 joints per individual) [17]. For the same parameters, but for specificity, a sample size of 113 was required. To ensure an adequate sample size accounting for missing data, we targeted approximately 150 participants. Summary statistics were used to describe the baseline characteristics of the study cohort and the reference standard physical examination assessment of the MCP and PIP joints. Missing data were handled by available case analysis. Skin tone was assessed from the hand images by a member of the study team (MM) and categorized using a simplified Fitzpatrick skin scale (type 1 or 2=palest, type 3 or 4=medium, and type 5 or 6=darkest) [18]. Visual assessment of the photos and videos (index test) was compared at the joint and person levels with the reference standard assessment and reported as sensitivity (true positives/[true positives+false negatives]), specificity (true negatives/[true negatives+false positives]), positive predictive value (PPV; PPV=true positives/[true positives+false positives]), and negative predictive value (NPV; NPV=true negatives/[true negatives+false negatives]).

Agreement between the 4 evaluators was calculated using Fleiss kappa and pairwise comparisons between evaluators were calculated using Cohen kappa. Adequate visual performance was defined as concordance between physical examination and visual inspection by most of the assessors (at least 3 of 4 assessors). Baseline patient characteristics were compared for association with evaluator performance using chi-square tests. Correlation between assessor's years of experience in rheumatology and performance (sensitivity and specificity) was assessed using the Pearson correlation coefficient. A P value of <.05 was considered statistically significant. All analyses were conducted using Python 3.9 (Python Software Foundation) using the *pandas*, *scikit-learn*, and *statsmodels* packages.

Results

A total of 156 participants were recruited, with an average age of 53.2 (SD 14.2) years (Table 1). Of them, 105 (67.3%) were female. Most (n=90, 57.7%) participants had Fitzpatrick skin types 1 and 2. The most common rheumatologic diagnosis was rheumatoid arthritis (n=52, 33.3%), with an average duration of joint symptoms of 10 years. The average BMI was in the overweight category (29 kg/m^2). Only 6 (3.8%) participants had notable joint deformities.

Table . Participant characteristics (n=156).

Characteristics	Participants
Age (y), mean (SD)	53.2 (14.2)
Sex, n (%)	
Female	105 (67.3)
Male	51 (32.7)
Skin tone, n (%)	
Fitzpatrick 1 and 2	90 (57.7)
Fitzpatrick 3 and 4	52 (33.3)
Fitzpatrick 5 and 6	13 (8.3)
Diagnosis, n (%)	
Rheumatoid arthritis	52 (33.3)
Psoriatic arthritis	21 (13.5)
Osteoarthritis	7 (4.5)
No known joint disease	24 (15.4)
Other	52 (33.3)
Duration of joint symptoms (years), mean (SD)	9.8 (10.1)
BMI (kg/m^2), mean (SD)	29.2 (6.1)
Notable joint deformities, n (%)	6 (3.8)

XSL•FO
RenderX

Of the 156 individuals, 20 (12.8%) had joint effusion in at least one MCP or PIP joint, as assessed by their rheumatologist on physical examination (reference standard; Table 2). On average, there were 3.4 (SD 2.3) effused joints per individual assessed to have joint effusion. In total, 3120 MCP and PIP joints were assessed, with 69 (2.2%) joints found to have effusion per physical examination. Of these, 43 (1.4%) were MCP joints, and 26 (0.8%) were PIP joints.

Table . Rheumatologist physical examination findings (reference standard).

Analysis	Values
Person level (n=156)	
Joint effusion present, n (%)	20 (12.8)
Effused joints or an individual with joint effusion, mean (SD)	3.4 (2.3)
Joint level (n=3120), n (%)	
Joint effusion present	69 (2.2)
MCP[a] joints	43 (1.4)
PIP[b] joints	26 (0.8)

[a]MCP: metacarpophalangeal.

[b]PIP: proximal interphalangeal.

All 156 participants had photos, and 150 (96.1%) participants had videos. Individual assessor's performance on photos and videos is listed in Tables S1 to S4 in Multimedia Appendix 1. Confusion matrices summarizing the overall group performance, using the threshold of at least 3 of 4 assessors labeling the presence of joint effusion being considered a positive index test, are included in Figure S1 in Multimedia Appendix 1. At the joint level, the average sensitivity was 0.14 (95% CI 0.07 - 0.19) for photos and 0.24 (95% CI 0.13 - 0.33) for videos, with a PPV of 0.14 (95% CI 0.06 - 0.19) and 0.33 (95% CI 0.15 - 0.37), respectively (Table 3). The average specificity was 0.97 (95% CI 0.96 - 0.98) for photos and 0.98 (95% CI 0.97 - 0.99) for videos, with an NPV of 0.98 (95% CI 0.98 - 0.98) for both photos and videos. The kappa among all 4 evaluators was 0.12 (95% CI 0.07 - 0.16) for photos and 0.17 (95% CI 0.11 - 0.23) for videos.

Table . Accuracy of visual assessment at the person and joint levels, by image type.

Analysis	Photos, mean (95% CI)	Videos, mean (95% CI)
Joint level[a]		
Sensitivity	0.14 (0.07-0.19)	0.24 (0.13-0.33)
Specificity	0.97 (0.96-0.98)	0.98 (0.97-0.99)
PPV[b]	0.14 (0.06-0.19)	0.33 (0.15-0.37)
NPV[c]	0.98 (0.98-0.98)	0.98 (0.98-0.98)
Agreement[d]	0.12 (0.07-0.16)	0.17 (0.11-0.23)
Person level[e]		
Sensitivity	0.44 (0.30-0.60)	0.48 (0.35-0.60)
Specificity	0.82 (0.76-0.87)	0.84 (0.79-0.89)
PPV	0.31 (0.17-0.35)	0.38 (0.24-0.43)
NPV	0.91 (0.88-0.93)	0.92 (0.89-0.93)
Agreement[d]	0.17 (0.07-0.28)	0.17 (0.04-0.28)

[a]Photos: n=3120; videos: n=3000.

[b]PPV: positive predictive value.

[c]NPV: negative predictive value.

[d]Fleiss kappa between evaluators.

[e]Photos: n=156; videos: n=150.

The average sensitivity of visually detecting the presence of MCP or PIP joint effusion at the person level was 0.44 (95% CI 0.30 - 0.60) from photos and 0.48 (95% CI 0.35 - 0.60) from videos (Table 3), with a PPV of 0.31 (95% CI 0.17 - 0.35) and 0.38 (95% CI 0.24 - 0.43) for photos and videos, respectively. The average specificity was 0.82 (95% CI 0.76 - 0.87) from photos and 0.84 (95% CI 0.79 - 0.89) from videos, with an NPV of 0.91 (95% CI 0.88 - 0.93) and 0.92 (95% CI 0.89 - 0.93) for

XSL•FO
RenderX

photos and videos, respectively. The kappa among all 4 evaluators was 0.17 (95% CI 0.07 - 0.28) for photos and 0.17 (95% CI 0.04 - 0.28) for videos. Pairwise agreements between assessors are listed in Tables S5 and S6 in Multimedia Appendix 1. Illustrative examples of cases with high and low agreement between assessors and concordant and discordant findings, compared with the reference standard, are shown in Figure S1 in Multimedia Appendix 1.

We did not find any patient characteristics that were statistically associated with performance by visual assessment (Table 4). We did not find a significant correlation between years of experience of the assessor and performance (sensitivity: correlation coefficient=0.29; *P*=.70, and specificity: correlation coefficient=0.07; *P*=.93).

Table . Patient characteristics associated with person-level performance[a].

Characteristics	Photos (n=156)		Videos (n=150)	
	Adequate, n/N (%)	*P* value	Adequate, n/N (%)	*P* value
Age (y)		.89		.69
18 - 39	26/32 (81.2)		29/32 (90.6)	
40 - 69	74/99 (74.7)		76/92 (82.6)	
≥70	18/25 (72.0)		19/26 (73.1)	
Sex		.84		.70
Female	78/105 (74.3)		81/100 (81.0)	
Male	40/51 (78.4)		43/50 (86.0)	
Skin tone		.65		.91
Fitzpatrick 1 and 2	63/90 (70.0)		74/87 (85.1)	
Fitzpatrick 3 and 4	43/52 (82.7)		41/51 (80.4)	
Fitzpatrick 5 and 6	12/14 (85.7)		9/12 (75.0)	
BMI (kg/m^2)		.70		.85
<30	71/97 (73.2)		77/94 (81.2)	
≥30	47/59 (79.7)		47/56 (83.9)	
Notable joint deformities present		>.99		.65
Yes	5/6 (83.3)		4/6 (66.7)	
No	113/150 (75.3)		120/144 (83.3)	

[a]Adequate performance was defined as concordance between physical examination and visual inspection by at least 3 of 4 assessors.

Discussion

Principal Findings

The results of this study demonstrated that rheumatologists were not able to accurately detect hand joint effusions through photos and videos, with the sensitivity being particularly low, which means that joint effusions of the MCPs and PIPs were missed by visual inspection alone. Specificity was higher for both photos and videos, suggesting that if joint effusions were found by visual inspection, they were likely present. Sensitivity and PPVs were slightly better at the person level than at the joint level, suggesting that visual inspection may be more appropriate to support triage and follow-up decisions than to assess disease activity. Agreement between assessors at both the person and joint levels was poor, underscoring the limitations of unaided visual inspection. Overall, performance was better with video assessment than with photo assessment, although these differences were not large. We were not able to detect any significant associations between patient characteristics and visual inspection accuracy, nor any correlation between assessors' years of experience in rheumatology and visual inspection performance. These findings have important clinical implications, especially during remote telehealth visits.

The use of telehealth in rheumatology has been studied, but a recent review found that nearly all studies focused on patient and physician satisfaction, improving access to care, and reducing costs [13]. A recent scoping review of the unassisted physical examination conducted over telemedicine found 74 studies, with 8 that compared the telehealth musculoskeletal assessment to in-person assessment [16]. All these studies found that telehealth assessments were equivalent to in-person examination [19-26]. However, all except one of these studies were done in an orthopedic surgery setting, with one that was assessing Parkinson disease. None of these studies examined the performance of telehealth examination in the rheumatology setting. The low sensitivity observed in this study has important implications for telehealth in rheumatology. If joint effusions are missed during remote consultations, this could delay the detection of active inflammation, potentially leading to delays in treatment. Delay in diagnosis and treatment initiation for inflammatory arthritis can result in poor functional outcomes in patients [27]. There is also increasing evidence that

XSL•FO
RenderX

treat-to-target strategies in rheumatoid arthritis improve patient outcomes; inaccurate assessment of swollen joint count and, therefore, disease activity can result in missed or delayed changes in therapy [28,29]. While telehealth appointments are necessary in certain situations and can improve access to specialty care for many patients, these limitations in visual examination for inflammatory arthritis need to be considered. The inability to fully capture all aspects of joint effusions through photos and videos may hinder the opportunity for timely therapeutic interventions, reducing the overall effectiveness of telehealth visits in managing these conditions. Moreover, most disease activity scores also require tender joint count, which can only be assessed through palpation [30].

One of the key strengths of this study is its use of a contemporary reference standard for comparison: the physical examination conducted by rheumatologists on the same day that the photos and videos were taken. This approach ensures that the findings are grounded in current clinical findings, providing a reliable benchmark for assessing the accuracy of photo and video assessments. Additionally, the study incorporated multiple assessors who were blinded to the patients' clinical history and symptoms, minimizing potential bias and ensuring that the results reflect objective visual evaluation of joint effusions. The inclusion of a proportion of inflammatory joints that mirrors real-world clinical practice further strengthens the study's relevance to everyday rheumatologic care.

Limitations

A notable limitation of this study is the predominance of participants with lighter skin tones, limiting our ability to assess performance differences by skin tone. Skin tone may influence the visibility of joint swelling, and visual inspection may be less reliable in individuals with darker skin, potentially leading to differences in diagnostic accuracy. Moreover, the Fitzpatrick skin scale, while the most widely used and validated skin classification system, has not been specifically validated on hands and may perform less reliably when applied to images than in person [18,31]. Similarly, higher BMI and the presence of deformities could obscure the visibility of joint abnormalities, making it more difficult to accurately identify effusions during physical or remote assessments. However, likely due to insufficient sample size to assess these secondary outcomes, we did not detect differences in performance by these patient characteristics. We also did not collect data on the specific types of deformities noted; however, given the limited number of participants with deformities, we would not have been adequately powered to stratify our results based on the type of deformities present.

The prevalence of joint effusions was lower than that used for our sample size calculation, which would reduce the precision of our estimates. Fortunately, missing data were lower than expected, and thus, the recruitment of a larger sample size than calculated resulted in reasonable 95% CIs, suggesting stable estimates. Additionally, all images in this study were taken by someone on the research team, using the same camera, in a clinic setting. Real-life performance, particularly by patients, may vary by camera resolution, lighting, and user. While we used rheumatologists' clinical examination as the reference standard in this study, the sensitivity of detecting clinical swelling in the hand joints by rheumatologists' physical examination (compared with power Doppler ultrasound) is approximately 70% [32]. The addition of corresponding joint imaging would have allowed us to compare visual assessment to a more accurate reference standard and should be considered in future studies. Future studies could also examine the comparative and additive performance of patient-reported outcome measures compared with visual inspection for detecting active inflammatory arthritis [33].

The fact that rheumatologists are not able to accurately detect joint effusions through visual inspection signals promising opportunities for technological advancement. Artificial intelligence has been successfully trained to detect joint effusions of the hands using photos, although it has not been applied in a real-world setting [34]. Improving the reliability of remote assessments through artificial intelligence could make telehealth visits more effective, ultimately enhancing the management of active inflammatory arthritis and improving health care equity by ensuring that all patients, regardless of location and other sociodemographic factors, have access to accurate and timely care. However, the clinical collection and use of such large, representative datasets, particularly those containing sensitive patient images and videos, must be approached with caution and robust safeguards to ensure that patient privacy and data security are paramount.

In conclusion, remote visual inspection of joint effusions via photos and videos lacks the accuracy necessary for reliable detection, a limitation that must be carefully considered when conducting telehealth appointments.

Funding

This research was funded by Arthritis Society Canada (and partners as applicable; grant SOG24-0377). CY is supported by a Canadian Rheumatology Association Foundation (CRAF) Canadian Initiative for Outcomes in Rheumatology Care (CIORA)-Arthritis Society Canada Clinician Investigator Award (Cl-24-0013).

Data Availability

The data that support the findings of this study are available on request from the corresponding author, CY. The data are not publicly available because they contain information that could compromise the privacy of research participants.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: CY
Investigations (data collection): SA-M, MB, NJ, DS, DJ, SK, ASR
Data curation: MA
Formal analysis: MA
Methodology: CY, MA, JRM
Writing—original draft: CY, MA
Writing—review and editing: all authors

Conflicts of Interest

MB has received consultation fees from UCB, Sobi, AbbVie, Janssen, and AmerisourceBergen/Innomar. All other authors have no conflicts of interest to declare.

Multimedia Appendix 1
Table S1-S4: Person- and joint-level performance of visual evaluation of photos and videos by individual assessor. Table S5: Person- and joint-level pairwise agreement for photos and videos. Figure S1 A-D: Confusion matrices for person- and joint-level photo and video overall assessor performance.
[DOCX File, 3578 KB - jmir_v28i1e86261_app1.docx]

References

1. Bilsborrow JB, Peláez-Ballestas I, Pons-Estel B, et al. Global rheumatology research: frontiers, challenges, and opportunities. Arthritis Rheumatol 2022 Jan;74(1):1-4. [doi: 10.1002/art.41980] [Medline: 34535973]
2. Katz SJ, Ye C. Persistent inequality in access to rheumatology care for females after the COVID-19 pandemic. Musculoskeletal Care 2024 Dec;22(4):e70026. [doi: 10.1002/msc.70026] [Medline: 39674905]
3. Katz SJ, Ye C. Biological sex inequality in rheumatology wait times during the COVID-19 pandemic. J Rheumatol 2023 Oct;50(10):1346-1349. [doi: 10.3899/jrheum.221213] [Medline: 36921972]
4. Niemantsverdriet E, Dougados M, Combe B, van der Helm-van Mil AHM. Referring early arthritis patients within 6 weeks versus 12 weeks after symptom onset: an observational cohort study. Lancet Rheumatol 2020 Jun;2(6):e332-e338. [doi: 10.1016/S2665-9913(20)30061-8] [Medline: 38273596]
5. The State of Arthritis in Canada Report Card. : Resilient Healthcare Coalition URL: https://www.resilienthealthcare.ca/publications/the-state-of-arthritis-in-canada-report-card [accessed 2026-04-08]
6. Mbonu I, Tai S, Bartels CM, Putman M. Association between neighborhood deprivation and number of rheumatology providers. Arthritis Care Res (Hoboken) 2023 Jan;75(1):9-13. [doi: 10.1002/acr.25036] [Medline: 36205227]
7. Badley EM, Canizares M, Gunz AC, Davis AM. Visits to rheumatologists for arthritis: the role of access to primary care physicians, geographic availability of rheumatologists, and socioeconomic status. Arthritis Care Res (Hoboken) 2015 Feb;67(2):230-239. [doi: 10.1002/acr.22413] [Medline: 25048206]
8. Wright GC, Okoye GA, Ehrlich AC, et al. Disparities in physician access for rheumatology, dermatology, and gastroenterology: a systematic review. Am J Manag Care 2025 Sep 1;31(9):e270-e277. [doi: 10.37765/ajmc.2025.89792] [Medline: 40966602]
9. George MD, Danila MI, Watrous D, et al. Disruptions in rheumatology care and the rise of telehealth in response to the COVID-19 pandemic in a community practice-based network. Arthritis Care Res (Hoboken) 2021 Aug;73(8):1153-1161. [doi: 10.1002/acr.24626] [Medline: 33973389]
10. Matsumoto RA, Barton JL. Telerheumatology: before, during, and after a global pandemic. Curr Opin Rheumatol 2021 May 1;33(3):262-269. [doi: 10.1097/BOR.0000000000000790] [Medline: 33741808]
11. Kwok TS, Kuriya B, King LK, et al. Changes in service delivery and access to rheumatologists before and during the COVID-19 pandemic in a Canadian universal healthcare setting. J Rheumatol 2023 Jul;50(7):944-948. [doi: 10.3899/jrheum.220658] [Medline: 36182116]
12. Anderson D, Porto A, Angelocci T, Lee I, Macri G. The impact of eConsults on access to specialty care for the uninsured in rural Texas. J Health Care Poor Underserved 2022;33(2):779-789. [doi: 10.1353/hpu.2022.0063] [Medline: 35574876]
13. Avouac J, Marotte H, Balsa A, et al. Teleconsultation in rheumatology: a literature review and opinion paper. Semin Arthritis Rheum 2023 Dec;63:152271. [doi: 10.1016/j.semarthrit.2023.152271] [Medline: 37813005]
14. Villaseñor-Ovies P, Navarro-Zarza JE, Canoso JJ. The rheumatology physical examination: making clinical anatomy relevant. Clin Rheumatol 2020 Mar;39(3):651-657. [doi: 10.1007/s10067-019-04725-9] [Medline: 31446539]
15. Almoallim H, Kalantan D, Alharbi L, Albazli K. Approach to musculoskeletal examination. In: Almoallim H, Cheikh M, editors. Skills in Rheumatology: Springer; 2021:17-65 URL: https://www.ncbi.nlm.nih.gov/books/NBK585755/ [accessed 2026-04-08]
16. Yao P, Adam M, Clark S, et al. A scoping review of the unassisted physical exam conducted over synchronous audio-video telemedicine. Syst Rev 2022 Oct 13;11(1):219. [doi: 10.1186/s13643-022-02085-1] [Medline: 36229830]

XSL•FO
RenderX

17. Bujang MA, Adnan TH. Requirements for minimum sample size for sensitivity and specificity analysis. J Clin Diagn Res 2016 Oct;10(10):YE01-YE06. [doi: 10.7860/JCDR/2016/18129.8744] [Medline: 27891446]
18. Harvey VM, Alexis A, Okeke CA, et al. Integrating skin color assessments into clinical practice and research: a review of current approaches. J Am Acad Dermatol 2024 Dec;91(6):1189-1198. [doi: 10.1016/j.jaad.2024.01.067] [Medline: 38342247]
19. Abel KC, Baldwin K, Chuo J, et al. Can telemedicine be used for adolescent postoperative knee arthroscopy follow-up? J Orthop Physician Assist 2017;5(4):e26. [doi: 10.2106/JBJS.JOPA.17.00014]
20. Cabana F, Boissy P, Tousignant M, Moffet H, Corriveau H, Dumais R. Interrater agreement between telerehabilitation and face-to-face clinical outcome measurements for total knee arthroplasty. Telemed J E Health 2010 Apr;16(3):293-298. [doi: 10.1089/tmj.2009.0106] [Medline: 20406116]
21. Goldstein Y, Schermann H, Dolkart O, et al. Video examination via the smartphone: a reliable tool for shoulder function assessment using the constant score. J Orthop Sci 2019 Sep;24(5):812-816. [doi: 10.1016/j.jos.2018.12.023] [Medline: 30686690]
22. Good DW, Lui DF, Leonard M, Morris S, McElwain JP. Skype: a tool for functional assessment in orthopaedic research. J Telemed Telecare 2012 Mar;18(2):94-98. [doi: 10.1258/jtt.2011.110814] [Medline: 22247538]
23. Lade H, McKenzie S, Steele L, Russell TG. Validity and reliability of the assessment and diagnosis of musculoskeletal elbow disorders using telerehabilitation. J Telemed Telecare 2012 Oct;18(7):413-418. [doi: 10.1258/jtt.2012.120501] [Medline: 23086982]
24. Richardson BR, Truter P, Blumke R, Russell TG. Physiotherapy assessment and diagnosis of musculoskeletal disorders of the knee via telerehabilitation. J Telemed Telecare 2017 Jan;23(1):88-95. [doi: 10.1177/1357633X15627237] [Medline: 26985005]
25. Russell TG, Hoffmann TC, Nelson M, Thompson L, Vincent A. Internet-based physical assessment of people with Parkinson disease is accurate and reliable: a pilot study. J Rehabil Res Dev 2013;50(5):643. [doi: 10.1682/JRRD.2012.08.0148] [Medline: 24013912]
26. Steele L, Lade H, McKenzie S, Russell TG. Assessment and diagnosis of musculoskeletal shoulder disorders over the internet. Int J Telemed Appl 2012;2012:945745. [doi: 10.1155/2012/945745] [Medline: 23193395]
27. Naeem F, Khan SE, Saeed MA, Farman S. Diagnostic and therapeutic delay in rheumatoid arthritis patients: impact on disease outcome. Pak J Med Sci 2021;37(4):1001-1007. [doi: 10.12669/pjms.37.4.3471] [Medline: 34290773]
28. Smolen JS, Breedveld FC, Burmester GR, et al. Treating rheumatoid arthritis to target: 2014 update of the recommendations of an international task force. Ann Rheum Dis 2016 Jan;75(1):3-15. [doi: 10.1136/annrheumdis-2015-207524] [Medline: 25969430]
29. Stoffer MA, Schoels MM, Smolen JS, et al. Evidence for treating rheumatoid arthritis to target: results of a systematic literature search update. Ann Rheum Dis 2016 Jan;75(1):16-22. [doi: 10.1136/annrheumdis-2015-207526] [Medline: 25990290]
30. England BR, Tiong BK, Bergman MJ, et al. 2019 Update of the American College of Rheumatology recommended rheumatoid arthritis disease activity measures. Arthritis Care Res (Hoboken) 2019 Dec;71(12):1540-1555. [doi: 10.1002/acr.24042] [Medline: 31709779]
31. Weir VR, Li Y, Gillis MC, et al. Evaluating skin tone scales for dermatologic dataset labeling: a prospective-comparative study. NPJ Digit Med 2025 Dec 22;8(1):787. [doi: 10.1038/s41746-025-02245-2] [Medline: 41429926]
32. Almoallim H, Attar S, Jannoudi N, et al. Sensitivity of standardised musculoskeletal examination of the hand and wrist joints in detecting arthritis in comparison to ultrasound findings in patients attending rheumatology clinics. Clin Rheumatol 2012 Sep;31(9):1309-1317. [doi: 10.1007/s10067-012-2013-5] [Medline: 22673791]
33. Myasoedova E, De Thurah A, Erpelding ML, et al. Definition and construct validation of clinically relevant cutoffs on the FLARE Assessment in Rheumatoid Arthritis (FLARE-RA) questionnaire. Semin Arthritis Rheum 2020 Apr;50(2):261-265. [doi: 10.1016/j.semarthrit.2019.09.004] [Medline: 31526593]
34. Phatak S, Saptarshi R, Sharma V, et al. Incorporating computer vision on smart phone photographs into screening for inflammatory arthritis: results from an Indian patient cohort. Rheumatology (Oxford) 2025 May 1;64(5):2525-2532. [doi: 10.1093/rheumatology/keae678] [Medline: 39680895]

Abbreviations

MCP: metacarpophalangeal
NPV: negative predictive value
PIP: proximal interphalangeal
PPV: positive predictive value

XSL•FO
RenderX

Edited by A Stone; submitted 21.Oct.2025; peer-reviewed by K Lu, PR Lage-Hansen, L Sapundzhiev; revised version received 24.Mar.2026; accepted 25.Mar.2026; published 16.Apr.2026.

Please cite as:
Ye C, Mitchell JR, Al-Matar S, Bishay M, Jones N, Sholter D, Johnson D, Katz S, Russell AS, Abdalla M
Accuracy of Visual Inspection Alone to Assess Joint Effusions of the Hand: Cross-Sectional Study
J Med Internet Res 2026;28:e86261
URL: https://www.jmir.org/2026/1/e86261
doi:10.2196/86261

© Carrie Ye, J Ross Mitchell, Shahad Al-Matar, Mena Bishay, Niall Jones, Dalton Sholter, Dylan Johnson, Steven Katz, Anthony S Russell, Mohamed Abdalla. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Rapid Deployment of Telemedicine in HIV Care: Mixed Methods Study of Providers' Attitudes and Perceptions

Shannon Galvin[1], MD; Artur Queiroz[2], PhD; Juan Rivera[3], MEd; Mia Calamari[1], MPH; Katherine Wright[3], MPH; Lisa R Hirschhorn[4], MD, MPH; Laura Rusie[3], ScM; Patrick Janulis[1,4], PhD; Rebecca Kumar[5], MD; Frank Palella[1], MD; Mary Clare Masters[1], MD; Jacqueline Bannon[1], PhD, RN; Susheel Reddy[1], MPH; Claudia Hawkins[1], MD

[1]Division of Infetious Disease, Northwestern University, 645 N Michigan Avenue, Chicago, IL, United States

[2]Department of Nursing, Florida State University, Tallahassee, FL, United States

[3]Howard Brown Health Center, Chicago, IL, United States

[4]Department of Medical Social Sciences, Northwestern University, 645 N Michigan Avenue, Chicago, IL, United States

[5]Georgetown University, Washington, DC, United States

Corresponding Author:
Shannon Galvin, MD
Division of Infetious Disease, Northwestern University, 645 N Michigan Avenue, Chicago, IL, United States

Abstract

Background: The COVID-19 pandemic prompted an abrupt transition to telemedicine for HIV care, necessitating the exploration of provider attitudes and experiences to understand its ongoing viability and impact.

Objective: This study aimed to assess providers' attitudes toward and experiences with telemedicine at HIV clinics in the Chicago area during the COVID-19 pandemic at 2 time points and at 2 clinical sites.

Methods: We conducted a convergent mixed methods study to evaluate and explore providers' attitudes toward and experiences with telemedicine at HIV clinics in the Chicago area during the COVID-19 pandemic, applying the RE-AIM (reach, effectiveness, adoption, implementation, and maintenance) framework and the updated Consolidated Framework for Implementation Research. This study assessed HIV providers via surveys at 2 time points, capturing responses on the acceptability, appropriateness, feasibility, and maintenance of telemedicine. Semistructured key informant interviews were conducted on a random selection of 10 participants to explore perceived experiences with telemedicine.

Results: Among emailed providers, 43 of 83 (51.8%) and 27 of 82 (32.9%) responded to survey 1 and survey 2, respectively. The first survey recorded telemedicine usage at 75%, which decreased to 58% by the second survey. Overall, the majority of respondents agreed with statements that assessed telemedicine as appropriate, acceptable, and feasible. There were overall few statistical differences in responses between sites, although more providers at the community site indicated at least some potential value in telemedicine when compared to providers at the university hospital (100% agree or strongly agree vs 82.3%; P=.04) in survey 1. In survey 1, providers with more than 10 years of experience were less likely to report telemedicine helped them see more patients (31.6% agree or strongly agree vs 70.6% for providers with ≤10 y of experience; P=.008). Fewer experienced providers felt they could talk about private issues with their patients during telemedicine (57.9% of more experienced vs 88.3% of less experienced providers; P=.03). Key informant interviews provided qualitative insights into telemedicine integration, revealing mixed sentiments; providers appreciated the flexibility and accessibility that telemedicine offered but preferred in-person visits for their thoroughness, especially for physical examinations and laboratory tests. Despite initial barriers such as technical challenges and patient preferences for in-person visits, telemedicine was deemed feasible for maintaining communication and care quality. Providers highlighted the need for better technological support and ongoing training to optimize telemedicine usage.

Conclusions: Our study underscores telemedicine as a sustainable adjunct to traditional HIV care, emphasizing the importance of addressing technological and training barriers to enhance its efficacy.

(J Med Internet Res 2026;28:e75933) doi:10.2196/75933

KEYWORDS

telemedicine; HIV; provider attitudes; telehealth; mixed methods

XSL•FO
RenderX

Introduction

The COVID-19 pandemic created abrupt and large-scale changes in health care delivery. One notable effect was the rapid increase in the use of telemedicine. Telemedicine is a type of telehealth in which telecommunications technologies support the delivery of medical treatment–related services [1]. A telehealth visit or televisit can be defined as a singular 2-way, synchronous client-provider interaction via telephone or video [2]. Telemedicine use increased from less than 1% of visits prior to the COVID-19 pandemic to as high as 80% in March to April 2020 [3]. While peak use was reported during the pandemic in the spring of 2020, telemedicine use remains high. Recent data show that, overall, 39% to 43% of US adults reported having a televisit in the past year [4,5]. Despite its widespread adoption, the future of telemedicine in the United States is unclear, in part, due to uncertainty in regulations and reimbursement.

HIV care reflected the national trend with a rapid increase in the adoption of telemedicine beginning in 2020. More than 99% of Ryan White providers offered telemedicine after the pandemic, compared to 22% before the pandemic [6]. In the US Department of Veterans Affairs system in 2023, HIV was one of the conditions with the highest proportion of care visits via telemedicine, with 8.8% of visits [7]. While televisits have been shown to be useful in HIV prevention, including improving HIV self-testing [8] and pre-exposure prophylaxis uptake [8-10], the use of telemedicine for HIV care is still evolving. Studies evaluating telemedicine in HIV care before the pandemic showed no difference in virologic outcomes and demonstrated effectiveness in mental health care delivery for persons living with HIV in select settings [11]. Studies conducted after the COVID-19 pandemic have started to document the use of telemedicine in HIV care, showing equivalent or better virologic outcomes when compared to in-person care [12]. The uptake, acceptability, and effectiveness of the rapid, widespread rollout of telemedicine for HIV care during the COVID-19 pandemic and beyond are still not completely understood. Potential benefits include increased patient access, improved retention, greater patient convenience, and a reduction in exposure to COVID-19 in health care settings [13,14]. Risks include a potential decrease in quality of care, concerns about confidentiality of data or personal information, especially related to HIV diagnosis, and inequities in care delivery based on patients' access to technology [13,14].

HIV providers' attitudes and experiences are critical to understanding the risks and benefits of telemedicine in HIV care and to providing important information on barriers and facilitators at the system, clinic, and patient levels. Providers' attitudes are key factors in adopting telemedicine [15]. In a small study of HIV care organizations in South Carolina, providers frequently reported barriers to telemedicine, including insufficient technology support, lack of clear guidelines, lack of ability to assess laboratory data, and both client and provider unfamiliarity with the technology and the procedures [16]. In other studies of single HIV sites, overall provider acceptability of the use of telemedicine in general was high [17,18]. Whether telemedicine will be maintained as a viable strategy requires further information on HIV provider views and experiences.

We conducted a convergent mixed methods study to evaluate and explore providers' attitudes toward and experiences with telemedicine at HIV clinics in the Chicago area during the COVID-19 pandemic, applying the RE-AIM (reach, effectiveness, adoption, implementation, and maintenance) framework and updated Consolidated Framework for Implementation Research frameworks [19,20]. We hypothesized that providers' attitudes toward telemedicine would change over time and be associated with provider and clinic characteristics. Findings from this study could have important practical implications for the continued use of telemedicine in routine HIV clinical practice and future times of system stress.

Methods

Overview

This mixed methods study was nested within a larger study evaluating the processes, implementation outcomes, and effectiveness of an emergently implemented telemedicine program for HIV care during the COVID-19 pandemic, and its effectiveness for people living with HIV managed in an academic (Northwestern University [NU]) and a community outpatient clinic (Howard Brown Health Center [HB]) between March 13, 2020, and December 31, 2021 [21].

Study Location and Telemedicine Implementation During the COVID-19 Pandemic

NU provides care for people living with HIV in an outpatient clinic based at Northwestern Memorial Hospital. The NU patient population is predominantly non-Hispanic Black and White men who have sex with men, most of whom are insured. HB provides care at 10 community-based sites to a racially diverse, gender-diverse, and culturally diverse population of people living with HIV, many (30%) of whom are underinsured or uninsured. People living with HIV managed at both sites are medically complex, with multiple comorbidities that further challenge adherence to care and treatment success.

In March 2020, NU and HB rapidly implemented telemedicine to minimize patient and provider exposure to COVID-19. Each clinic provided patients with the option to convert their scheduled in-person outpatient appointments to visits that could be conducted virtually via telephone or videoconference using Doximity (Doximity Inc.) (NU) or Health Insurance Portability and Accountability Act (HIPAA)–compliant health care version of Zoom (Zoom Communications Inc.) (HB). At both sites, if the provider determined that the patient required an in-person visit (before or during the televisit), they were scheduled with one of the on-site providers. If a patient required laboratory testing but had a televisit, they were instructed to come to the appropriate site for testing.

Study Population and Procedures

All providers (nurse practitioners, physicians, and physician assistants) who provided direct care for persons with HIV at NU and HB during the study period were eligible. Providers were asked to complete 2 anonymous online surveys developed and administered using REDCap (Research Electronic Data Capture) between December 7, 2020, and March 12, 2021 (survey 1) and November 17, 2021, and January 10, 2022

XSL•FO
RenderX

(survey 2). Key informant interviews (KIIs) were then conducted among a subset of providers between August 2021 and November 2021. The mixed methods design was chosen, with a qualitative component (KIIs) giving context to the quantitative survey results.

The RE-AIM framework was selected to guide the quantitative evaluation of implementation outcomes, including reach, adoption, implementation, and maintenance [19]. The Consolidated Framework for Implementation Research (CFIR) was used to guide qualitative data collection and analysis, providing a structured approach to identifying multilevel determinants influencing implementation processes [20]. The use of both frameworks allowed RE-AIM to characterize *what* implementation outcomes were observed, while CFIR supported the interpretation of *why* these outcomes occurred by identifying contextual facilitators and barriers [22].

Survey

A REDCap-generated email was sent to 83 providers (27 NU and 56 HB) that contained a link to the survey instrument. Survey recruitment continued until all providers had 3 opportunities to respond, at which point the recruitment was closed. The survey was adapted from prior validated instruments, including the System Usability Scale [23,24], Normalization Measure Development Questionnaire [25,26], and implementation outcomes modified for local context. The initial survey (survey 1) included questions on acceptability, appropriateness, feasibility, and maintenance of telemedicine during the COVID-19 pandemic, as well as additional questions on frequency and reasons for conducting telemedicine. The same survey (survey 2) was administered a second time in March 2021 to 82 providers (26 NU and 56 HB), with additional questions on how the number of telemedicine visits changed since the prior questionnaire, reasons for the change, if any, and preferences for future telemedicine visits. Survey questions in the domains of acceptability, appropriateness, feasibility, and maintenance were rated by participants using a 5-point Likert scale (ie, strongly agree, agree, neither agree nor disagree, disagree, and strongly disagree; Table S1 in Multimedia Appendix 1).

Key Informant Interviews

KIIs were conducted with 10 participants. Providers were randomly selected from the eligible pool. After 10 participants, saturation of themes was deemed to have been met, consistent with literature assessing sample sizes for qualitative studies in homogeneous populations [27,28]. Study investigators reviewed and edited semistructured interview guides in discussion with other study team members. Study interviews were conducted over Zoom in a private room by the study coordinator, who was trained in qualitative techniques by investigators with this expertise. Study interviews lasted between 30 and 65 minutes and occurred over 2 months. Interview questions were open-ended and explored perceived experiences with telemedicine during the COVID-19 pandemic, features, and processes to integrate telemedicine, barriers and facilitators, and perceived differences in care with telemedicine.

All interviews were conducted in English, digitally recorded, and transcribed. Two authors (CH and SG) first read all transcripts independently and then jointly identified themes to guide the analysis. AQ conducted a lexical analysis with the software IRaMuTeQ-*Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires* (version 0.7), operating as an interface within R (version 4.4.2; R Foundation for Statistical Computing) [29-32].

For the textual analysis, the descending hierarchical classification method was defined, in which the texts are classified according to their respective words, and their set is divided by the frequency of the reduced forms. Thus, classes of text segments were obtained, which were called “preclass.” Subsequently, the organization of key expressions from the interviewed speeches and the identification of central ideas complementing the descending hierarchical classification findings were performed and allowed the delimitation of statements in definite classes [33]. AQ guided the deductive theming of data under each CFIR domain, and disagreements were resolved by discussion between the authors. CFIR’s 5 main domains and multiple constructs used were intended to capture barriers to and facilitators of implementation. The domains we explored included innovation characteristics (eg, relative advantage of telehealth vs in-person appointments, perceived adaptability of telehealth to meet the current local needs of patients during the pandemic), individual characteristics (eg, knowledge and beliefs about the intervention, and personal attributes such as motivation and capability to adapt to telehealth), inner setting (eg, capacity of the organization, the organizational implementation climate, and capacity for change), outer setting (eg, political and societal changes that can influence the implementation), and process (eg, characterized by plans to ease implementation and engagement of community stakeholders). Direct quotations are provided to illustrate the evidence from which the findings were derived, and themes are summarized.

Statistical Analysis

Internal consistency and reliability of items within domains were assessed using raw Cronbach α. Alpha values indicated satisfactory internal consistency for the feasibility (α=0.846) and maintenance (α=0.899) domains. The α values for the appropriateness domain (α=0.672) were equivocal.

Responses were tabulated separately for each survey administration. Individual Likert items were summarized as ordinal categorical measures using frequency counts and percentages across the 5-point Likert scale (strongly agree to strongly disagree). Acknowledging the limited sample size and skewed distributions in many instances, exact Mantel-Haenszel chi-square statistics were used to identify differences in the distribution of responses by site (NU vs HB) and respondent tenure (≤10 y vs >10 y).

Survey administrators were blinded to the identity of respondents across the survey 1 and survey 2 administrations. Changes in perceptions for a given respondent were not tracked as a result, and direct inferential comparisons between survey 1 and survey 2 results were not applied. All analyses were completed using SAS (version 9.4). Data visualizations were

XSL•FO
RenderX

prepared using Tableau Desktop (version 2021.3.1; Tableau Software).

Ethical Considerations

The research was reviewed and approved by the Institutional Review Boards at NU (STU00213441) and HB (DEF 20 - 26). Informed consent was obtained from all participants. Data for qualitative surveys were anonymous, and data for KIIs were deidentified after transcription. No compensation was provided to participants.

Results

Overview

A total of 43 of 83 (51.8%) and 27 of 82 (32.9%) emailed providers responded to survey 1 and survey 2, respectively. Of survey 1 respondents, 30.2% (13/43) had been providing HIV care for over 15 years, 16.3% (7/43) for 11 to 15 years, 41.9% (18/43) for 1 to 10 years, and 11.6% (5/43) for less than 1 year (Table 1). Distribution by tenure for survey 2 respondents was not markedly different, with 33.3% (9/27) providing care for over 15 years, 22.2% (6/27) for 11 to 15 years, 37% (10/27) for 1 to 10 years, and 7.4% (2/27) for less than 1 year.

Table . Respondent characteristics.

Characteristics	Survey 1 (n=43), n (%)	Survey 2 (n=27), n (%)
Tenure (years providing HIV care)		
<1	5 (11.6)	2 (7.4)
1 - 2	8 (18.6)	4 (14.8)
3 - 5	7 (16.3)	5 (18.5)
6 - 10	3 (7)	1 (3.7)
11 - 15	7 (16.3)	6 (22.2)
>15	13 (30.2)	9 (33.3)
Professional category		
Nurse practitioner	10 (23.3)	7 (25.9)
Physician	32 (74.4)	18 (66.7)
Physician assistant	1 (2.3)	2 (7.4)
Any telemedicine in survey periods		
No	7 (16.3)	3 (11.1)
Yes	36 (83.7)	24 (88.9)
Proportion of visits completed using telemedicine[a]		
<25%	19 (52.8)	16 (66.7)
25% - 75%	12 (33.3)	8 (33.3)
>75%	5 (13.9)	0 (0)

[a]Survey 1: n=36; and survey 2: n=24.

The proportion of health care encounters that were telemedicine declined between the 2 survey time points from 75% to 58%. In survey 1, a total of 19 of 36 (52.8%) reported conducting 25% to 75% of their HIV visits via telemedicine, with 12 of 36 (33.3%) conducting less than 25% and 5 of 36 (13.9%) conducting more than 75%, with one of those respondents conducting all visits via telemedicine. The proportion of providers conducting 25% to 75% telemedicine decreased to 33.3% (8/24) in survey 2, and none reported conducting more than 75%. Although telemedicine seemed to comprise a smaller proportion of a provider's encounters in survey 2, more providers in survey 2 reported at least some use of telemedicine, with 7 of 43 (16.3%) reporting not completing any telemedicine visits in survey 1 versus 3 of 27 (11.1%) in survey 2. A minority of telemedicine visits were conducted by video in both time frames, with only 8 of 36 (22.2%) in survey 1 and 4 of 24 (16.7%) in survey 2 reporting more than 75% video use.

Those providers reporting no use of telemedicine cited having visits where procedures and same-day or urgent visits were needed as reasons. One provider reported never conducting telemedicine visits at all during the study, stating, "I do not feel comfortable doing telemedicine."

The main reasons for not conducting a televisit included patients' desire to be seen in person (survey 1: 28/36, 77.8%; and survey 2: 11/24, 45.8%), patients having conditions that required physical examination (survey 1: 29/36, 80.6%; and survey 2: 19/24, 79.2%), or the need for laboratory testing (survey 1: 26/36, 72.2%; and survey 2: 19/24, 79.2%).

Telemedicine Outcomes Among Providers

Provider responses to survey 1 are shown in Figure 1. In survey 1, a total of 20 (55.6%) respondents disagreed with the statement, "I prefer telemedicine over in person visits,"

XSL•FO
RenderX

indicating that they preferred in-person visits, while 8 (22.2%) respondents agreed or strongly agreed that they preferred telemedicine, and another 8 (22.2%) were neutral. In survey 2, these preferences were unchanged, with a slight shift in responses toward more favorable appraisals of telemedicine. However, a plurality continued to indicate a preference for in-person visits (disagree or strongly disagree: n=11, 45.8%). One-third (n=8, 33.3%) of respondents indicated a neutral response, with 20.8% (5/24) preferring telemedicine.

Figure 1. Survey 1.

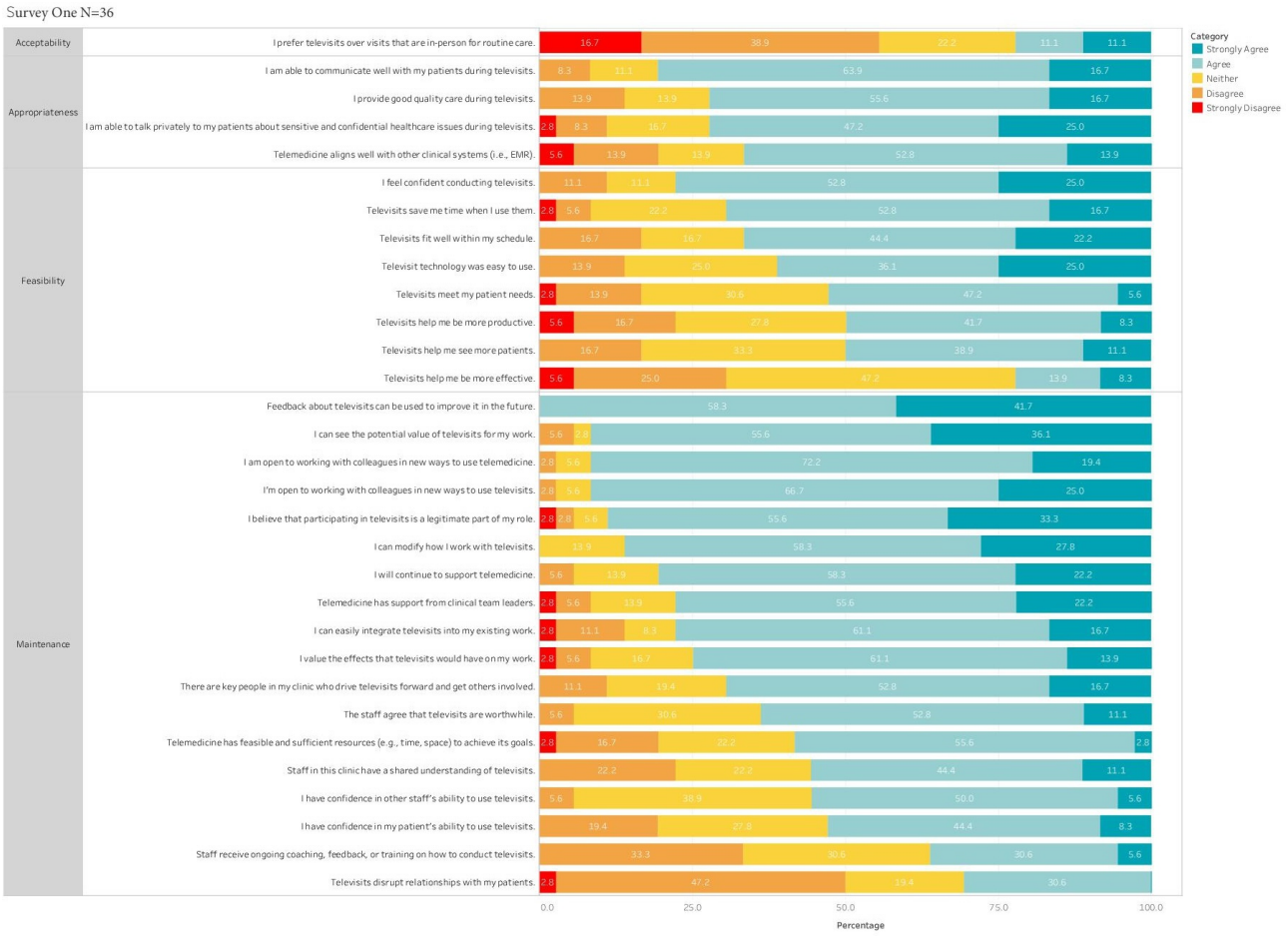

Overall, providers felt telemedicine was appropriate, with high proportions agreeing with statements that televisits provided the ability to communicate (survey 1: 29/36, 80.6%), talk privately about sensitive issues (26/36, 72.2%), and provide good quality care (26/36, 72.2%).

Most providers found telemedicine feasible. In survey 1, providers seemed to agree telemedicine saved them time (survey 1: 25/36, 69.4%), with half (survey 1: 18/36, 50%) agreeing they helped them be more productive. Respondents also reported high levels of confidence conducting televisits, meeting their patient needs using telemedicine, and compatibility with their schedules and ease of use. Fewer participants agreed that telemedicine was more effective (8/36, 22.2% favorable) and efficient (18/36, 50%).

Respondents indicated broad agreement with most of the survey items intended to address maintenance of telemedicine. Most providers strongly agreed or agreed with statements indicating support for telemedicine, that telemedicine was supported by clinical leadership, that they are open to working on new ways of using telemedicine, and that staff supported and were able to use telemedicine. Most providers wanted further feedback on their use of telemedicine. Overall, the reaction was neutral that staff had received sufficient training and coaching on telemedicine in survey 1.

Similar results around appropriateness, feasibility, and maintenance were seen in survey 2 (Figure 2). Questions assessing appropriateness showed overall high or higher proportions agreeing with the ability to communicate (survey 1: 29/36, 80.6% vs survey 2: 22/24, 91.7%), talk privately about sensitive issues (26/36, 72.2% vs 21/24, 87.5%), and provide good quality care (26/36, 72.2% vs 17/24, 70.8%). At survey 2, participants continued to increasingly agree that telemedicine saved them time (survey 1: 25/36, 69.4% and survey 2: 19/24, 79.2%) and helped them be more productive (survey 1: 18/36, 50% and survey 2: 14/24, 58.3%). Assessments of the effectiveness of televisits were more equivocal, although there was some evidence of a shift toward more favorable responses between survey 1 (8/36, 22.2%) and survey 2 (9/24, 37.5%). Notably, a lower proportion of respondents reported that telemedicine helped them see more patients in survey 2 (10/24, 41.7%) even after a long period of using the technology, compared to survey 1 (18/36, 50%). Regarding maintenance, there was overall concordance, except in 2 questions where even more providers disagreed with the statement that providers

XSL•FO
RenderX

had received ongoing training and support (survey 1: 13/36, 36.1% agree and survey 2: 5/24, 20.8% agree). There seemed to be a decrease in reporting disruption in providers' relationship with their patients from survey 1 (11/36, 30.6%) to survey 2 (3/24, 12.5%).

Figure 2. Survey 2.

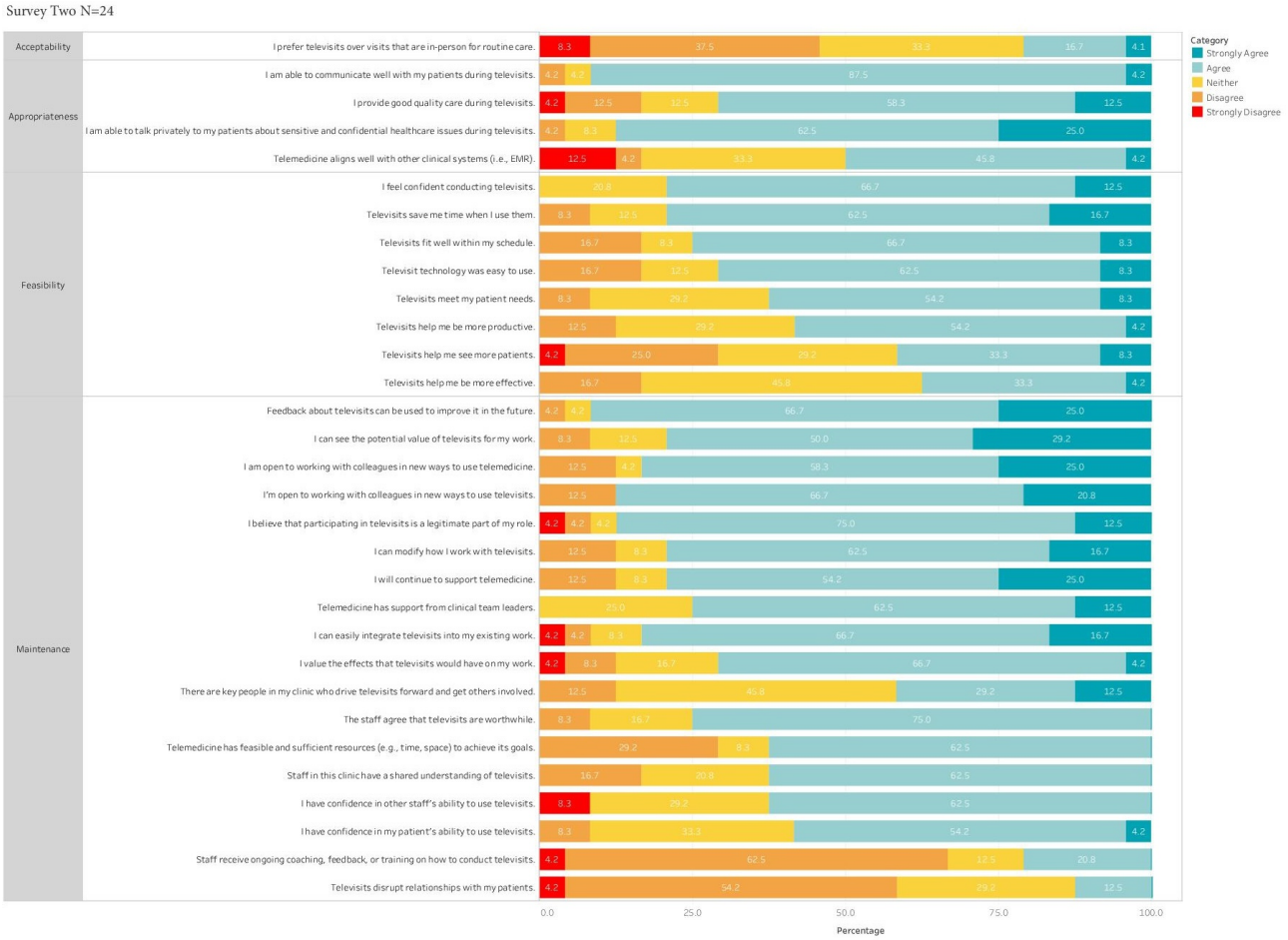

Better videoconferencing, access to laboratories, and software for sharing patient health care data, such as blood pressure, were recommended as ways to make telemedicine more valuable in the future.

Implementation Outcomes With Differences by Site and Provider Type

Overall, there were few differences between sites in views regarding telemedicine (Table 2). More providers at HB indicated at least some potential value in telemedicine when compared to providers at NU (19/19, 100% at HB agree or strongly agree vs 14/17, 82.3% at NU; P=.04) in survey 1. In survey 2, all NU respondents indicated an ability to modify how they work with telemedicine (agree or strongly agree) versus 50% (5/10) of Howard Brown respondents (P=.002). No differences by site were found in both survey 1 and survey 2 among the items in the acceptability, appropriateness, and feasibility domains.

XSL•FO
RenderX

Table . Responses for individual Likert items demonstrating evidence of association by site or tenure.

Domain and item description	Response, n (%)					*P* value[a]	Total[b]
	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree		
Survey 1 by site							
Maintenance							
I can see the potential value of televisits for my work.						.04	
HB[c] (n=22)	9 (47.4)	10 (52.6)	0	0	0		19
NU[d] (n=21)	4 (23.5)	10 (58.8)	1 (5.9)	2 (11.8)	0		17
Total	13 (36.1)	20 (55.6)	1 (2.8)	2 (5.6)	0		36
Survey 1 by tenure							
Feasibility							
Televisits help me see more patients.						.008	
≤10 y (n=23)	3 (17.7)	9 (52.9)	5 (29.4)	0	0		17
>10 y (n=20)	1 (5.3)	5 (26.3)	7 (36.8)	6 (31.6)	0		19
Total	4 (11.1)	14 (38.9)	12 (33.3)	6 (16.7)	0		36
Appropriateness							
I am able to talk privately to my patients about sensitive and confidential healthcare issues during televisits.						.03	
≤10 y (n=23)	7 (41.2)	8 (47.1)	1 (5.9)	1 (5.9)	0		17
>10 y (n=20)	2 (10.5)	9 (47.4)	5 (26.3)	2 (10.5)	1 (5.3)		19
Total	9 (25.0)	17 (47.2)	6 (16.7)	3 (8.3)	1 (2.8)		36
Survey 2 by site							
Maintenance							
I can modify how I work with televisits.						.002	
HB (n=11)	0	5 (50)	2 (20)	3 (30)	0		10
NU (n=16)	4 (28.6)	10 (71.4)	0	0	0		14
Total	4 (16.7)	15 (62.5)	2 (8.3)	3 (12.5)	0		24
Survey 2 by tenure							
Appropriateness						.046	
Telemedicine aligns well with other clinical systems (ie, EMR)[e].							
≤10 y (n=12)	0	2 (22.2)	4 (44.4)	1 (11.1)	2 (22.2)		9
>10 y (n=15)	1 (6.7)	9 (60)	4 (26.7)	0	1 (6.7)		15
Total	1 (4.2)	11 (45.8)	8 (33.3)	1 (4.2)	3 (12.5)		24
Acceptability							
I prefer televisits over visits that are in-person for routine care.						.02	
≤10 y (n=12)	1 (11.1)	3 (33.3)	3 (33.3)	2 (22.2)	0		9
>10 y (n=15)	0	1 (6.7)	5 (33.3)	7 (46.7)	2 (13.3)		15

XSL•FO
RenderX

Domain and item description	Response, n (%)					*P* value[a]	Total[b]
	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree		
Total	1 (4.2)	4 (16.7)	8 (33.3)	9 (37.5)	2 (8.3)		24

[a]*P* values calculated using exact Mantel-Haenszel methodologies. Missing values are not considered when calculating *P* values. Likert responses by site or tenure displayed for items with a *P* value <.05.

[b]Total refers to the number of nonmissing responses. This represents the denominator from which row percentages are calculated.

[c]HB: Howard Brown Health Center.

[d]NU: Northwestern University.

[e]EMR: electronic medical record.

There were some differences between providers based on tenure. In survey 1, providers with more than 10 years of experience were less likely to report that telemedicine helped them see more patients (31.6% agree or strongly agree vs 70.6% for providers with ≤10 years of experience; *P*=.008). Fewer experienced providers felt they could talk about private issues with their patients during telemedicine (57.9% of more experienced providers vs 88.3% of less experienced providers; *P*=.03). However, in survey 2, providers with more than 10 years of experience felt that visits were more integrated into their electronic medical record (66.7% agree or strongly agree vs 22.2% of less experienced providers; *P*=.046) but were less likely to prefer telemedicine to in-person care for routine care (6.7% agree or strongly agree vs 44.4% of less experienced providers; *P*=.02).

Key Informant Interviews

A number of themes were found during the provider interviews that gave additional insights into their experience with telemedicine (Table 3).

Table . Important CFIR domains and constructs identified in key informant interviews.

Domains	Constructs	Determinants identified
Intervention characteristics	• Adaptability • Relative advantage	• Perceptions regarding the benefits and barriers of telemedicine to continue their connection with patients
Inner setting	• Tension for Change • Compatibility • Culture • Recipient centeredness • Structural characteristics • Technology infrastructure	• Need to consider the impact of the COVID-19 pandemic • Staff have concerns over the availability of technology, patient engagement, and workflow disruption • Staff from both sites lacked adequate and adapted spaces for conducting visits with adequate privacy
Outer setting	• Patient needs or resources • Cosmopolitanism	• Patients experience stress, fear of the unknown, and disconnect with health care • Staff feel a lack of support from local and governmental organizations on how to conduct telemedicine care, training on technologies, and support • Linguistic barriers are exacerbated during telemedicine visits
Characteristics of individuals	• Intervention knowledge or beliefs • Other personal attributes • Need	• Knowledge and beliefs regarding technical aspects of telemedicine • Need of the patients to receive care during the pandemic
Process	• Tailoring strategies • Engaging innovation recipients	• Professionals adapted their visits, including additional support to attend to patients' needs

Theme 1: The Rapid Transition Due to COVID-19

The transition to telemedicine due to COVID-19 was rapid and unplanned, and providers saw themselves transitioning to a new reality of care in which their (physical) connection with their patients was severed, without considering disparities in the needs of those patients:

> *Of course, there was a period of time where it really converted to pretty much all telemedicine. After we*

XSL•FO
RenderX

started seeing people in person, I would say I went quickly to half and half.

I have tried it on some of my patients who are really difficult to connect with [for example] people with substantial substance issues.

Some providers adapted to televisits with ease, indicating very little change in their daily activities:

I feel like it's amazing how comfortable I feel doing telemedicine and feeling like I can provide as good care as when in person assuming that they're relatively well and don't have major issues [...]

Theme 2: Barriers to Telemedicine Visits

Overview

Although unplanned, most of the providers rapidly adjusted to this new configuration of care and communication, surprising even themselves with how quickly this adaptation process occurred. However, in situations where external (disruptions in the environment), media (the interface with technology), and internal (lack of patient engagement or focus) changes were occurring rapidly, providers reported barriers, including frustration on perceiving a disconnection with their patients and lacking other tools to overcome this at a distance.

Barrier: Lack of Support From the Institution

Facing these barriers, providers saw themselves without institutional support, having to adjust to this new reality without the proper tools to communicate and provide a quiet and private ambience to meet with their clients:

[...] if they had the same level of attention to providing a space for us, I will not do a televisit with eight people in the room.

Barrier: Privacy

Privacy was a main concern for providers. Providers could not replicate the same discretion of a clinic room while in telemedicine, having to resort to adjusting their speech to adverse situations, at the expense of clear communication. The main reason for this was the desire to protect their confidentiality at a time when many families were suddenly forced to remain confined in their houses, without a proper space to discuss sensitive subjects, like sexual health:

There are times where they're [the patients] like speaking in code. Because they're at work and I'm trying to decipher what they're trying to tell me it just depends on the situation [...], there's other times when they're speaking freely in front of like their grandma. You just never know what you're going to get.

Barrier: Accommodating Emotions During Telemedicine

Providers did not feel comfortable navigating more nuanced subjects during televisits with their patients, such as providing a diagnosis or being the bearer of "bad news" through a screen. Health professionals have developed skills to rely on body language and closer proximity to their patients to know how to handle emotional situations during in-person visits. When deprived of that in televisits, they felt unsure of how to conduct their care, physically and emotionally:

Like interaction of catching people's emotional responses or just like little nuances of being together in person it's not the same in the video so I'm hearing like that human connection that kind of emotional cues are easier to find in in-person visits.

Well, telephone is more difficult because I feel like we end up interrupting each other in a way that we can't see each other. So, I feel like that can impede both the communication but also just the comfort of the relationship.

I certainly don't give people bad news that way if I have anticipated difficult news of course I try to bring them in on occasion I've had to give bad news over the phone I always ask the patient how they want it.

Barrier: Difficulties in Language and Understanding

Providers expressed concerns about their ability to effectively manage and support patients when language and understanding were compromised. Providers found that telemedicine worked better with patients who could clearly understand and follow directions. For those who struggled with language, the absence of face-to-face interaction made it more difficult to ensure understanding and effective communication:

I think it works better with a patient who is really capable and understands directions and that type of stuff.

Honestly I'm not real fond of using it for people who don't speak English well I do use translators but it's hard enough to use a translator in person but in person you've got the visual cues to help you too.

Theme 3: Opportunities and Adjustments

In contrast to the previous theme, demonstrating the diversity of experience collected from the providers, some found that telemedicine gave them the time and opportunity to finally connect to patients who had been more resistant to it before. This may be due to the alleviation of travel burden or more time available for the patients to attend visits.

Flexibility was a key point in the providers' experiences, highlighted here when they mentioned that having the laboratories open during critical periods of the pandemic allowed them to continue their care with minimal disruption:

There's one guy who lives in downstate Illinois. I had one in-person visit with him, but I've been doing a lot of telemedicine because he really requires close follow-up. So, every month, I call him.

Or you end up having less time because you've taken more time for a physical exam or something that's intervened, what you lack in a physical contact you make up for in in-depth of conversation something like that.

Other aspects of the clinics being available during the lockdown revealed the benefits of flexibility to maintain certain functions:

XSL•FO
RenderX

> *Luckily [the clinic] did keep its labs open, so we had the capacity to do labs on a semi-urgent basis.*

Theme 4: Necessary Tools for Telemedicine

Providers stated that technology and communications tools were a central aspect of the telemedicine experience. Providers saw themselves being required to adapt quickly to incorporate those tools into their daily lives in a vital way. The implementation of televisits was conducted without necessarily providing them with specific tools, forcing them to adapt as situations presented themselves. For example, some providers used their personal phone numbers to maintain the best possible communication with their patients despite the risks that come with exposing personal information.

In many cases, providers reported creative solutions (eg, sharing their numbers, inviting a translator to a 3-way call, and sending pictures via text to take advantage of a high resolution) as a way of coping with a lack of institutional support tools dedicated to establishing a stable and safe communication channel with their patients, or proper training:

> *[Talking about adaptations to telemedicine] is it a three-way call, [...] you call the translator first then they put the call into the patient they explain that they're the translator and doctor is on the phone.*

> *I actually have pretty much resorted to giving them my cell phone and asking them to send pictures because you don't get a very good look at it [via zoom] that's how I've augmented it.*

> *[...] it's something to think about or a way for a patient to send cell phone pictures confidentially to the provider without sending it to their cell phone there's got to be a way to do that.*

While some participants seemed to be comfortable or even prefer using their phones, being a familiar technology, the use of personal devices by providers demonstrates a lack of institutional support. Instead of institutions providing technological support, the providers had to use their own devices to perform care:

> *[...] it would be great if somebody actually showed me how to do the zoom call I actually finally grabbed "someone" and she's like this is how you do it then we got it setup right.*

Integration of Quantitative and Qualitative Findings (RE-AIM and CFIR)

To support a convergent mixed methods design, quantitative survey findings organized using the RE-AIM framework were integrated with qualitative interview findings mapped to the CFIR. Quantitative results describe implementation outcomes (eg, acceptability, adoption, and feasibility), while qualitative findings provide contextual explanations for these outcomes by identifying multilevel determinants influencing implementation. Table 4 presents a joint display aligning key RE-AIM outcomes with corresponding CFIR domains and representative qualitative insights.

XSL•FO
RenderX

Table . Integration of RE-AIM[a] implementation outcomes with CFIR[b]-informed qualitative findings.

RE-AIM domain	Key quantitative finding (survey)	Relevant CFIR domains	Qualitative interpretation (KII[c] themes)	Illustrative quote (example)
Reach	Most providers reported using telemedicine during the study period, with telephone visits used more frequently than video visits.	Outer setting: patient needs and resources	Providers described telemedicine (particularly telephone visits) as essential for reaching patients facing barriers such as limited internet access, unstable housing, or privacy concerns. However, limited visual assessment was noted as a trade-off.	"A lot of our patients just don't have stable internet, so phone visits are really the only way to reach them."
Adoption	Providers generally reported adopting telemedicine into their clinical practice, although uptake varied by role and clinical context.	Inner setting; characteristics of individuals	Adoption was influenced by provider comfort with technology, perceived fit with HIV care workflows, and prior experience. Some providers expressed greater enthusiasm than others.	"Some clinicians adapted quickly, others felt it didn't fit how they normally deliver care."
Acceptability	Telemedicine was rated as acceptable by most providers, particularly for follow-up and medication management visits.	Characteristics of individuals	Providers valued telemedicine for its flexibility and ability to maintain continuity of care but emphasized that acceptability decreased for visits requiring physical exams or sensitive discussions.	"It works well for routine follow-ups, but not everything can be done remotely."
Implementation	Providers reported moderate feasibility of delivering care via telemedicine.	Intervention characteristics; process	While telemedicine was feasible, providers identified implementation challenges, including lack of standardized protocols, limited training, and uncertainty around documentation and billing procedures.	"We figured it out as we went along, but there wasn't much guidance early on."
Maintenance	Providers expressed uncertainty about the long-term sustainability of telemedicine beyond the COVID-19 pandemic.	Outer setting; inner setting	Concerns about future reimbursement policies, regulatory changes, and institutional support were prominent and influenced expectations about continued telemedicine use.	"If reimbursement changes again, it will really affect whether we can keep doing this."

[a]RE-AIM: reach, effectiveness, adoption, implementation, and maintenance.

[b]CFIR: Consolidated Framework for Implementation Research.

[c]KII: key informant interview.

While overall quantitative data showed overall high appropriateness, feasibility, and maintenance, qualitative data highlighted concerns.

Concerns with privacy were more notable in KIIs than in the overall response to the survey questions, where most providers answered they were able to talk privately. Likewise, while most respondents agreed that they were able to communicate, qualitative responses showed many concerns around the ease of communication, especially for sensitive subjects, delivery of bad news, and for patients who do not speak English.

Both qualitative and quantitative data showed that feasibility and maintenance concerns correlated most with the inner setting domain around support for using the technology and proper staff and patient training. However, in the survey data, most respondents still agreed with questions that assessed if televisits were feasible and would be continued, while interviews often highlighted barriers they had faced. Possibly, the high degree of adaptability shown by the providers and some of the perceived benefits of the innovation identified in the KIIs ameliorated barriers, allowing for overall positive agreement with statements assessing feasibility.

Areas of agreement in the data showed that overall preference and acceptability for in-person visits remained higher with specific situations such as new patients, ill patients, and sensitive discussions given as examples in qualitative interviews.

The flexibility of telehealth in allowing patients to be seen who live far and providing more opportunity for discussion were also highlighted in provider interviews. As the majority of

XSL•FO
RenderX

respondents in the survey agreed they felt they could adapt and work with leadership, staff, and patients to integrate and improve telehealth, the benefits reported could be strengthened and maintained.

Discussion

Principal Findings

In this mixed methods study conducted at 2 different health care settings in Chicago, most HIV providers reported that telemedicine visits were a feasible, acceptable, and appropriate method of delivering care during the pandemic and supported their continuation. While providers reported a relatively smooth transition to the emergent introduction of telemedicine in our surveys, in-depth interviews with providers identified significant confidentiality and privacy concerns and the potential for suboptimal communication in the virtual setting. These findings highlight the importance of monitoring telemedicine procedures while eliciting ongoing patient and provider input into how to best use this health care modality.

Overall, the uptake of telemedicine was high, although it did vary over the course of the study. Telemedicine was rapidly adopted and comprised a large proportion of visits initially in the COVID-19 pandemic, although it declined from one-half to one-third a year later (survey 2). This is similar to a recent review reporting up to 86.9% use of telemedicine at some HIV clinics after 2020 [18]. The overall reasons for maintaining telemedicine were not related directly to infection prevention but to factors around patient convenience and access, which may be driving use as the pandemic recedes. Furthermore, the importance of long-term engagement makes factors that improve patient access especially relevant to HIV care. Follow-up data on the use of telemedicine in HIV care and its effect on long-term patient engagement and adherence will be important.

During the initial phase of the COVID-19 pandemic, providers experienced a significant transition, moving swiftly from predominantly in-person visits to a mix of telemedicine and in-person consultations. Surprisingly, given this rapid expansion, HIV providers provided mostly positive responses in the surveys about telemedicine, reporting high levels of confidence with this mode of health care delivery. However, exploring this transition in more depth with providers during their interviews revealed important concerns about the impact it had on patient care. At least one provider reported that the transition occurred without adequate consideration for the varying needs and circumstances of patients, while others were concerned about an apparent disconnect with patients. Despite these concerns, there was significant support for continuing to support telemedicine in the future, as indicated in individual responses in the maintenance domain of the provider survey.

Although most providers in the survey indicated that telemedicine was appropriate in that they could communicate and speak privately with patients, interviews indicated that privacy and confidentiality were frequent concerns. Confidentiality and privacy concerns have been shown to be a major barrier to accessing HIV care. In a survey of 80 people living with HIV and 60 HIV physicians in Italy, physicians had more concerns than people living with HIV that telemedicine use would increase isolation, create difficulties in interaction, worsen the disconnect between patients and providers, and result in a lower quality of care for patients. Both groups agreed they were willing to use telemedicine but that it should not replace personal visits completely [34]. Another recent study in Florida, United States, found that some people living with HIV felt they faced more potential confidentiality concerns with in-person care, especially if they received care in a clinic solely focused on HIV, but others had more privacy concerns with telehealth [35]. In our study, privacy was less of a concern in survey 2, possibly as clinics had more consistent procedures in place for delivering telemedicine and patients became more comfortable participating in televisits in their local environments. Other work has found that providers noted the ability to assess their patients' home environments could actually be a benefit [36], but this was not commonly noted in our study.

Adequate communication is a major component of appropriate health care and was a common theme identified by providers. Moreover, trust in health care workers requires communication with respect, partnership, and appropriate time [37]. Trust in their providers has been linked to better health literacy, better retention in care, and antiretroviral therapy adherence for people living with HIV [38]. Overall, providers responded that they could communicate well with their patients in the surveys; however, during interviews, they expressed concerns about the lack of nonverbal cues and physical contact to deliver bad news and discuss emotional issues. Providers also noted that many visits require a physical examination to be effective and expressed concern about a loss of connection, including lack of physical touch, which is important to the provider-patient relationship [36]. Literature suggests patients overall do not prioritize physical connection nor the need for physical diagnosis as much as providers [39]. Our study suggests that telemedicine may not fully replicate the interpersonal dynamics and communication of in-person consultations, particularly in emotionally charged situations and with vulnerable populations.

Some providers did note that telemedicine visits could provide more time for discussion compared to in-person visits, thereby improving access due to more flexibility on both the patient and provider side and the decreased need for travel. A similar observation was made by Walker et al [40] in assessing provider attitudes to telemedicine in California, United States. In this study, providers felt telemedicine reduced patient barriers to care and offered increased flexibility to providers, although there were concerns over the complexity of workflows for telemedicine, technical challenges with video platforms, and an overall high degree of variation among provider views. In a separate qualitative study of HIV providers, there was overall support for patients and staff to continue telemedicine, with perceptions that telemedicine was useful to keep patients in care, reporting specifically that it was easier and more flexible for patients. However, the authors also cited technology challenges, such as a lack of smartphones for both patients and providers in addition to reports of poor connectivity during visits and privacy concerns [41].

In our study, we found most visits did not use video, which perhaps reflected uncertainty or hesitancy using this then-new

XSL•FO
RenderX

technology. Similar observations have been made in other studies [3,40]. A recent systematic review showed video may have better diagnostic accuracy than phone for acute conditions or those requiring physical findings, but in many outpatient settings, clinical outcomes were equivalent for phone or video visits [42]. Interestingly, patient satisfaction was noninferior for telephone compared to video televisits in 1 study [43]. Given the rapid rollout of telemedicine during this study, it is unclear if increasing video use would have increased provider frustration or positively or negatively affected their privacy concerns. Further information on the acceptability and efficacy of telephone versus video modalities is needed, especially for HIV care, along with clarity regarding reimbursement for the differing types of visits.

Despite the challenges, the study identified several opportunities associated with telemedicine. Some providers reported improved connectivity with patients who had previously been resistant to traditional health care settings. Flexibility emerged as a key advantage, allowing for increased access to care and more efficient use of time for both providers and patients. The availability of essential health care services, such as laboratory testing, during lockdowns underscored the resilience of telemedicine in maintaining continuity of care during crises.

Limitations

The strengths of our study include a heterogeneous sample of providers of different backgrounds and experience at 2 different sites (academic and community health center) and 2 different periods. However, our study suffered from an inability to track a given provider's responses across survey administrations and small sample sizes that, taken together, limited our ability to apply statistical inferences. Changes in attitudes over time for participants were not considered, as respondents' identifiers were randomly assigned between survey administrations and unlinked to participants' personal identifiers to further maintain anonymity from the prospective peer respondents, which prevented the possibility of paired analyses. We did generate mean domain scores that were intended to summarize feasibility, maintenance, and appropriateness, with each constituent Likert item scored on a scale of 1 to 5, with higher scores indicating a more favorable assessment. These summaries can be found in Table S2 in Multimedia Appendix 2 (available on request). Summarizing consolidated items by domain revealed largely equivocal results. Although there is some evidence of a difference in the assessment of appropriateness by tenure in the first survey, we had some concerns regarding the raw Cronbach α of 0.672 obtained when testing this domain using these data. Given these findings and a primary interest in item-specific differences by site and tenure, we opted to exclude these analyses. We also considered and tested the possibility of a multivariable analysis of these data using proportional odds models, where the 5-category Likert outcomes were consolidated into 3 categories (ie, agree, neither agree nor disagree, and disagree). Sample sizes made this difficult, however, and the interpretation of the resulting odds ratios was a concern. Ultimately, we determined that reporting a simple analysis emphasizing the descriptive results with a minimum of applied inference was a better approach. In addition, qualitative and quantitative responses from individual participants could not be compared.

The single urban setting from only 2 clinics limited the generalization of our results to other settings. It should be noted that a relatively high rate of providers in both clinics chose not to respond to the survey, which possibly reflected the demands of COVID-19 on providers at that time. Therefore, results may reflect a response bias toward those with either more positive or more negative views. Additional limitations include reliance on self-reported attitudes rather than objective provider metrics, such as visit number and duration, and a potential that responses may suffer from social desirability bias as participants were employees of the organization. Nonetheless, this study demonstrates an initial indication that characterizes the use of telemedicine for HIV care as largely acceptable, feasible, appropriate, and sustainable. These findings recommend the implementation of larger, robust longitudinal investigations rigorously evaluating a potentially promising means of augmenting and facilitating HIV patient care.

Conclusions

In this mixed methods study conducted at 2 different health care settings in Chicago, most HIV providers reported that telemedicine visits overall were a feasible, acceptable, and appropriate method of delivering care during the pandemic and supported their continuation. The findings of this study have several implications for future health care practice and policy. Addressing the identified challenges, such as privacy concerns and the need for institutional support, will be crucial in ensuring the sustainability and effectiveness of telemedicine initiatives. Investments in technological infrastructure, training programs, and regulatory frameworks are essential to maximize the potential of telemedicine while mitigating its limitations. Additionally, future research should focus on longitudinal studies to assess the long-term impacts of telemedicine on patient outcomes and health care delivery models, especially in HIV care. Overall, our results suggest that support for telemedicine as a continuing care modality is warranted.

Acknowledgments

The authors would like to acknowledge the patients and staff at Northwestern University ID Center and Howard Brown Health Centers.

XSL•FO
RenderX

Funding

Funding for the study was provided by P30AI117943 Third Coast Center for AIDS Research Supplement: Evaluation of an Emergency Implemented Telemedicine Intervention, its Acceptability and Impact on HIV care and Treatment Among Persons Living with HIV.

Data Availability

The quantitative survey datasets generated during the study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey questions and domains.
[DOCX File, 16 KB - jmir_v28i1e75933_app1.docx]

Multimedia Appendix 2
Supplementary table.
[DOCX File, 21 KB - jmir_v28i1e75933_app2.docx]

References

1. Telehealth, telemedicine, and telecare: what's what? FCC. URL: https://www.fcc.gov/general/telehealth-telemedicine-and-telecare-whats-what [accessed 2026-01-02]
2. Why use telehealth? HHS. URL: https://telehealth.hhs.gov/patients/why-use-telehealth [accessed 2026-01-23]
3. Karimi M, Lee E, Couture S, et al. National survey trends in telehealth use in 2021: disparities in utilization and audio vs. video services. : Assistant Secretary for Plannning and Evaluation; 2022 URL: https://aspe.hhs.gov/sites/default/files/documents/4e1853c0b4885112b2994680a58af9ed/telehealth-hps-ib.pdf [accessed 2025-05-10]
4. Chang E, Penfold RB, Berkman ND. Patient characteristics and telemedicine use in the US, 2022. JAMA Netw Open 2024 Mar 4;7(3):e243354. [doi: 10.1001/jamanetworkopen.2024.3354] [Medline: 38517438]
5. Spaulding EM, Fang M, Commodore-Mensah Y, Himmelfarb CR, Martin SS, Coresh J. Prevalence and disparities in telehealth use among US adults following the COVID-19 pandemic: national cross-sectional survey. J Med Internet Res 2024 May 10;26:e52124. [doi: 10.2196/52124] [Medline: 38728070]
6. Labisi T, Regan N, Davis P, Fadul N. HIV care meets telehealth: a review of successes, disparities, and unresolved challenges. Curr HIV/AIDS Rep 2022 Oct;19(5):446-453. [doi: 10.1007/s11904-022-00623-z] [Medline: 36063284]
7. Ferguson JM, Greene L, Van Campen J, Zulman DM, Wray CM. Variability in primary care telehealth delivery methods across chronic conditions. JAMA Netw Open 2025 Mar 3;8(3):e251988. [doi: 10.1001/jamanetworkopen.2025.1988] [Medline: 40136302]
8. Siegler AJ, Koh SHE, Schukraft T, et al. Telemedicine preexposure prophylaxis prescribing from a large online US company. JAMA Netw Open 2025 Dec 1;8(12):e2546792. [doi: 10.1001/jamanetworkopen.2025.46792] [Medline: 41324957]
9. Calvetti S, Lei B, Stocks JB, et al. Feasibility and acceptability of a mHealth patient navigation intervention to increase pre-exposure prophylaxis uptake in racially and ethnically diverse sexual and gender minority youth in Los Angeles (PrEPresent): pilot randomized controlled trial. JMIR Form Res 2025 Oct 31;9:e69255. [doi: 10.2196/69255] [Medline: 41171997]
10. Chen S, Yu FY, Fang Y, et al. Chatbot-delivered real-time support to improve HIV self-testing rates. JAMA Netw Open 2025 Nov 3;8(11):e2544821. [doi: 10.1001/jamanetworkopen.2025.44821]
11. Junkins A, Psaros C, Ott C, et al. Feasibility, acceptability, and preliminary impact of telemedicine-administered cognitive behavioral therapy for adherence and depression among African American women living with HIV in the rural South. J Health Psychol 2021 Dec;26(14):2730-2742. [doi: 10.1177/1359105320926526] [Medline: 32515245]
12. Rehman N, Mbuagbaw L, Mertz D, Muraca GM, Jones A, Ontario HIVTNCS. Association between virtual visits and health outcomes of people living with HIV: a cross-sectional study. PLoS One 2025;20(5):e0315880. [doi: 10.1371/journal.pone.0315880] [Medline: 40435167]
13. Ezenwaji CO, Alum EU, Ugwu OPC. Bridging the gap: telemedicine as a solution for HIV care inequities in rural and vulnerable communities. Int J Equity Health 2025 Jul 14;24(1):205. [doi: 10.1186/s12939-025-02584-2] [Medline: 40660269]
14. Kwena ZA, Odhiambo FA, Kulzer JL, et al. Exploring patients' and healthcare providers' knowledge and perceptions of benefits, challenges, and resource requirements in integrating telemedicine services into routine HIV care in western Kenya. AIDS Care 2025 Dec 21:1-10. [doi: 10.1080/09540121.2025.2604607] [Medline: 41423422]

XSL•FO
RenderX

15. Connolly SL, Gifford AL, Miller CJ, Bauer MS, Lehmann LS, Charness ME. Provider perceptions of virtual care during the coronavirus disease 2019 pandemic: a multispecialty survey study. Med Care 2021 Jul 1;59(7):646-652. [doi: 10.1097/MLR.0000000000001562] [Medline: 34009880]
16. Yelverton V, Qiao S, Weissman S, Olatosi B, Li X. Telehealth for HIV care services in South Carolina: utilization, barriers, and promotion strategies during the COVID-19 pandemic. AIDS Behav 2021 Dec;25(12):3909-3921. [doi: 10.1007/s10461-021-03349-y] [Medline: 34173137]
17. Budak JZ, Scott JD, Dhanireddy S, Wood BR. The impact of COVID-19 on HIV care provided via telemedicine-past, present, and future. Curr HIV/AIDS Rep 2021 Apr;18(2):98-104. [doi: 10.1007/s11904-021-00543-4] [Medline: 33616811]
18. Meyer D, Slone SE, Ogungbe O, Duroseau B, Farley JE. Impact of the COVID-19 pandemic on HIV healthcare service engagement, treatment adherence, and viral suppression in the United States: a systematic literature review. AIDS Behav 2023 Jan;27(1):344-357. [doi: 10.1007/s10461-022-03771-w] [Medline: 35916951]
19. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
20. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implementation Sci 2009 Dec;4(1):50. [doi: 10.1186/1748-5908-4-50]
21. Masters MC, Rivera J, Calamari M, et al. Telemedicine and HIV care quality measures during the COVID-19 pandemic. J Acquir Immune Defic Syndr 2023 Sep 1;94(1):46-52. [doi: 10.1097/QAI.0000000000003238] [Medline: 37368925]
22. Li SA, Jeffs L, Barwick M, Stevens B. Organizational contextual features that influence the implementation of evidence-based practices across healthcare settings: a systematic integrative review. Syst Rev 2018 May 5;7(1):72. [doi: 10.1186/s13643-018-0734-5] [Medline: 29729669]
23. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the System Usability Scale. International Journal of Human–Computer Interaction 2008 Jul 29;24(6):574-594. [doi: 10.1080/10447310802205776]
24. Lewis JR. The System Usability Scale: past, present, and future. International Journal of Human–Computer Interaction 2018 Jul 3;34(7):577-590. [doi: 10.1080/10447318.2018.1455307]
25. Finch T, Girling M, May C, Mair F, Murray E, Treweek S, et al. Can we measure implementation progress using normalization process theory? Development and validation of the nomad survey tool. Int J Behav Med 2016 Nov;23:S135-S135. [doi: 10.1007/s12529-016-9586-3]
26. Finch TL, Girling M, May CR, et al. Improving the normalization of complex interventions: part 2 - validation of the NoMAD instrument for assessing implementation work based on normalization process theory (NPT). BMC Med Res Methodol 2018 Nov 15;18(1):135. [doi: 10.1186/s12874-018-0591-x] [Medline: 30442094]
27. Hennink MM, Kaiser BN, Weber MB. What Influences saturation? Estimating sample sizes in focus group research. Qual Health Res 2019 Aug;29(10):1483-1496. [doi: 10.1177/1049732318821692] [Medline: 30628545]
28. Turner-Bowker DM, Lamoureux RE, Stokes J, et al. Informing a priori sample size estimation in qualitative concept elicitation interview studies for clinical outcome assessment instrument development. Value Health 2018 Jul;21(7):839-842. [doi: 10.1016/j.jval.2017.11.014] [Medline: 30005756]
29. Mendes IAC, Ventura CAA, Queiroz AAFLN, de Sousa Á. Global health education programs in the Americas: a scoping review. Ann Glob Health 2020 Apr 21;86(1):42. [doi: 10.5334/aogh.2745] [Medline: 32346523]
30. Queiroz AAFLN, Sousa AFLD. PrEP Forum: an on-line debate on pre-exposure prophylaxis in Brazil. Cad Saude Publica 2017 Nov 21;33(11):e00112516. [doi: 10.1590/0102-311X00112516] [Medline: 29166481]
31. Queiroz AAFLN, Sousa ÁD, Matos MCB, Araújo TME, Reis RK, Moura MEB. Knowledge about HIV/AIDS and implications of establishing partnerships among Hornet® users. Rev Bras Enferm 2018;71(4):1949-1955. [doi: 10.1590/0034-7167-2017-0409] [Medline: 30156682]
32. Souza MARD, Wall ML, Thuler ACDMC, Lowen IMV, Peres AM. The use of IRAMUTEQ software for data analysis in qualitative research. Rev Esc Enferm USP 2018 Oct 4;52:e03353. [doi: 10.1590/S1980-220X2017015003353] [Medline: 30304198]
33. Ramos MG, Lima VMR, Amaral-Rosa MP. IRAMUTEQ software and discursive textual analysis: interpretive possibilities. In: Costa A, Reis L, Moreira A, editors. Computer Supported Qualitative Research: Springer; 2019:58-72. [doi: 10.1007/978-3-030-01406-3_6]
34. Massaroni V, Delle Donne V, Ciccarelli N, et al. Use of telehealth for HIV care in Italy: are doctors and patients on the same page? A cross-sectional study. Int J Med Inform 2021 Dec;156:104616. [doi: 10.1016/j.ijmedinf.2021.104616] [Medline: 34695728]
35. Carnes N, Koenig LJ, Wilkes AL, Gelaude D, Salabarría-Peña Y, Johnston M. Addressing stigma and privacy through telemedicine: qualitative findings on enhancing HIV care engagement among racial and ethnic minority groups. J Racial Ethn Health Disparities 2026 Apr;13(2):1106-1114. [doi: 10.1007/s40615-025-02319-7] [Medline: 40029481]
36. Gomez T, Anaya YB, Shih KJ, Tarn DM. A qualitative study of primary care physicians' experiences with telemedicine during COVID-19. J Am Board Fam Med 2021 Feb;34(Suppl):S61-S70. [doi: 10.3122/jabfm.2021.S1.200517] [Medline: 33622820]

XSL•FO
RenderX

37. Dawson-Rose C, Cuca YP, Webel AR, et al. Building Trust and Relationships Between Patients and Providers: An Essential Complement to Health Literacy in HIV Care. J Assoc Nurses AIDS Care 2016;27(5):574-584. [doi: 10.1016/j.jana.2016.03.001] [Medline: 27080926]
38. Whetten K, Leserman J, Whetten R, et al. Exploring lack of trust in care providers and the government as a barrier to health service use. Am J Public Health 2006 Apr;96(4):716-721. [doi: 10.2105/AJPH.2005.063255] [Medline: 16507725]
39. Santos LAD, Deus LFAD, Unsain RF, Leal AF, Grangeiro A, Couto MT. Agreements and disagreements between professionals and users about the experience of a Telehealth Service for HIV Pre-Exposure Prophylaxis (TelePrEP): qualitative Interview Study. J Med Internet Res 2025 Apr 2;27:e67445. [doi: 10.2196/67445] [Medline: 40173449]
40. Walker D, Moucheraud C, Butler D, et al. Experiences with telemedicine for HIV care in two federally qualified health centers in Los Angeles: a qualitative study. BMC Health Serv Res 2023 Feb 15;23(1):156. [doi: 10.1186/s12913-023-09107-1] [Medline: 36793080]
41. Harsono D, Deng Y, Chung S, et al. Experiences with telemedicine for HIV care during the COVID-19 pandemic: a mixed-methods study. AIDS Behav 2022 Jun;26(6):2099-2111. [doi: 10.1007/s10461-021-03556-7] [Medline: 35064390]
42. Caffery LJ, Catapan SDC, Taylor ML, et al. Telephone versus video consultations: a systematic review of comparative effectiveness studies and guidance for choosing the most appropriate modality. J Telemed Telecare 2025 Aug;31(7):909-918. [doi: 10.1177/1357633X241232464] [Medline: 38419502]
43. Danila MI, Sun D, Jackson LE, et al. Satisfaction with modes of telemedicine delivery during COVID-19: a randomized, single-blind, parallel group, noninferiority trial. Am J Med Sci 2022 Nov;364(5):538-546. [doi: 10.1016/j.amjms.2022.06.021] [Medline: 35793732]

Abbreviations

CFIR: Consolidated Framework for Implementation Research
HB: Howard Brown Health Center
HIPAA: Health Insurance Portability and Accountability Act
KII: key informant interview
NU: Northwestern University Infectious Diseases Clinic
RE-AIM: reach, effectiveness, adoption, implementation, and maintenance
REDCap: Research Electronic Data Capture

Edited by A Stone; submitted 13.Apr.2025; peer-reviewed by A Avila, S Chen, V Yelverton; revised version received 19.Feb.2026; accepted 19.Feb.2026; published 27.Apr.2026.

Please cite as:
Galvin S, Queiroz A, Rivera J, Calamari M, Wright K, Hirschhorn LR, Rusie L, Janulis P, Kumar R, Palella F, Masters MC, Bannon J, Reddy S, Hawkins C
Rapid Deployment of Telemedicine in HIV Care: Mixed Methods Study of Providers' Attitudes and Perceptions
J Med Internet Res 2026;28:e75933
URL: https://www.jmir.org/2026/1/e75933
doi:10.2196/75933

© Shannon Galvin, Artur Queiroz, Juan Rivera, Mia Calamari, Katherine Wright, Lisa R Hirschhorn, Laura Rusie, Patrick Janulis, Rebecca Kumar, Frank Palella, Mary Clare Masters, Jacqueline Bannon, Susheel Reddy, Claudia Hawkins. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Application of Digital Health Technologies in Cardiac Rehabilitation for Patients With Coronary Heart Disease: Scoping Review

Xinyu Zhu[1], MSN; Lei Liu[2], PhD, MD, Prof Dr; Yingjie Wang[3], MSN; Hongyuan Li[4], MSN; Min Zang[2], MSN; Jiayu Wang[2], MSN; Yu Tian[5], MSN; Zihan Li[1], MSN

[1]Liaoning University of Traditional Chinese Medicine, Shenyang, China

[2]School of Nursing, Liaoning University of Traditional Chinese Medicine, 79 Chongshan Road, Beita Street, Shenyang, China

[3]Department of Nursing, Affiliated Hospital of Liaoning University of Traditional Chinese Medicine, Shenyang, China

[4]Department of Nursing, Dalian Friendship Hospital, Dalian, China

[5]Liaoning Cancer Hospital & Institute, Shenyang, China

Corresponding Author:
Lei Liu, PhD, MD, Prof Dr
School of Nursing, Liaoning University of Traditional Chinese Medicine, 79 Chongshan Road, Beita Street, Shenyang, China

Abstract

Background: The high mortality and recurrence rates associated with coronary heart disease (CHD) impose substantial health care costs and economic burdens globally. Identifying effective interventions to improve patient outcomes is paramount. Digital health technologies (DHTs) offer novel solutions to overcome the challenge of low participation rates in traditional cardiac rehabilitation (CR).

Objective: This review aims to systematically map the scope of application, intervention objectives, and evaluation metrics of DHTs in CR for patients with CHD, thereby providing a structured evidence base for future research and practice.

Methods: This scoping review adheres to the Joanna Briggs Institute's methodology and is reported according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines. A systematic search was conducted across 5 major databases, PubMed, Web of Science, Embase, Cochrane Library, and EBSCO, covering the period from inception to February 2026. Inclusion criteria were developed based on the participants, concept, and context framework. Studies focused on the application of various DHTs within CR settings for patients with CHD. Eligible literature comprised randomized controlled trials, quasi-randomized controlled trials, and longitudinal before-and-after studies published in peer-reviewed journals. Two researchers (XZ and ZL) independently conducted literature screening and data extraction. Findings were presented through a comprehensive narrative synthesis and evidence gap maps.

Results: A total of 43 studies were included, predominantly randomized controlled trials (n=40). Findings revealed (1) diverse technological formats, categorized into 3 main types: digital health tools, real-time remote support, and asynchronous communication. Multitechnology combined interventions have become the mainstream model (36/43, 83.7%). (2) Intervention objectives were multifaceted, consolidating into 4 dimensions: motivation and guidance, knowledge and skills, monitoring and security, and social and group dynamics. (3) Evaluation metrics were multidimensional, encompassing clinical physiological indicators, health behaviors, patient-reported outcomes, service use rates, and technological feasibility. DHTs demonstrated positive effects in improving short-term physiological function and health behaviors; however, evidence remains insufficient regarding their impact on long-term clinical outcomes such as reducing adverse events.

Conclusions: The innovation of this scoping review lies in integrating highly heterogeneous evidence to reveal the field's evolution from isolated tools toward systematic, integrated solutions. Research confirms that DHTs effectively overcome temporal and spatial constraints, enhancing rehabilitation accessibility and engagement. They serve as crucial strategic tools for bridging geographical disparities in health care resources and advancing equity in cardiovascular health services. However, the evidence base remains limited, including insufficient long-term efficacy data and inadequate exploration of vulnerable populations such as older people and those with low digital literacy. Future research urgently requires large-scale, long-term follow-up clinical trials, alongside enhanced studies on adaptability for specific populations and considerations of health equity. This will propel digital CR toward greater scientific rigor, universal applicability, and precision.

(J Med Internet Res 2026;28:e85917) doi:10.2196/85917

KEYWORDS

digital health technologies; coronary heart disease; cardiac rehabilitation; scoping review; telemedicine

XSL•FO
RenderX

Introduction

Background

The World Health Organization's (WHO) report on the world's top 10 causes of death indicates that cardiovascular disease (CVD) claims the highest number of lives, with coronary heart disease (CHD) accounting for 13% of global mortality [1]. Beyond the health burden, CVD, particularly CHD, imposes a significant economic strain. Globally, CHD accounts for 42% of total CVD expenditure, with annual per capita expenditure on CHD reaching 4.9% to 137.8% of per capita gross domestic product [2]. This underscores how CHD has become a "heavy burden" weighing upon individuals, families, society, and health care systems. More notably, despite substantial investment, the overall prognosis for patients with CHD has yet to improve effectively.

In response to this challenge, cardiac rehabilitation (CR) has demonstrated significant value as a comprehensive intervention. Research indicates that CR can reduce cardiovascular adverse events by 28%, 1-year readmission rates by 31%, and mortality by 24% while effectively reducing health care expenditure [3-5]. It is recommended as a class Ia evidence-based intervention in clinical guidelines [6,7]. However, participation rates in CR programs remain universally low among patients with CHD globally [8-10]. Rates range from 9.7% to 22.5% in Germany, 20% to 30% in the United States, and a mere 41.5% in the United Kingdom [8-10]. Asian nations present similarly unfavorable figures: Singapore at 12.3% and Japan and South Korea between 14% and 50% [4,11-13]. As the most populous developing nation, China has CR centers constituting merely 0.06% of all medical institutions, with underdeveloped regions accounting for a mere 8.8% [14]. Beyond awareness factors, limitations inherent to traditional CR, such as transport difficulties, time conflicts, and uneven resource distribution, constitute primary barriers to participation [15].

At this intersection of practical need and technological innovation, digital health technologies (DHTs) have transcended the constraints of conventional CR, forging novel pathways for its implementation [16]. In 2019, the WHO formally introduced the concept of DHTs, defining them as the field of developing and using digital technologies to disseminate health knowledge and facilitate related practices [17]. This encompasses applications of technologies such as the Internet of Things and artificial intelligence within health management [17]. Digital devices such as pedometers, accelerometers, and smartphones enable daily activity tracking, exercise intensity assessment, and personalized exercise guidance for patients with CHD [18,19]. Smart pillboxes and "digital pills" facilitate real-time monitoring of medication adherence [20,21]. Thus, DHTs overcome temporal and spatial constraints to deliver more accessible rehabilitation support. They effectively alleviate resource scarcity issues and hold promise for extending benefits to a broader population with CHD [22,23].

In summary, DHTs offer novel solutions to low CR participation rates. However, their highly heterogeneous delivery formats result in fragmented evidence. Compared to traditional review methodologies, scoping reviews can integrate heterogeneous evidence and define research boundaries [24-26]. Therefore, in this study, we use a scoping review approach to systematically collate evidence in this field, providing a holistic perspective for subsequent research and policy formulation. This aims to bridge gaps in cardiovascular health accessibility and advance the scientific, universal, and sustainable development of digital CR.

Objectives and Research Questions

In this study, we aim to systematically review the scope of DHT applications in CR for patients with CHD through a scoping literature review methodology. It seeks to provide evidence-based guidance for the diversified development and effective implementation of future CR.

Our research will clarify (1) the application strategies and scenarios of existing digital technologies in CR for patients with CHD; (2) the key performance indicators determining the effectiveness of current DHT applications, alongside identifying their primary challenges; and (3) how DHTs can be more effectively applied to CR for patients with CHD and future research directions.

Methods

Overview

In this study, we strictly followed the Joanna Briggs Institute's scoping review methodology framework to ensure methodological rigor and transparency in the research process [27,28]. This scoping review adhered strictly to a structured research process, using standardized methodologies to ensure the reliability of findings and their practical applicability. The research involved comprehensive systematic literature searches, data extraction, and evidence synthesis analysis, culminating in a narrative synthesis of studies concerning the application of DHTs in CR for patients with CHD. As a scoping review, our primary objective is to systematically map the current application, intervention formats, and outcome measures of DHTs in CR for patients with CHD. It does not evaluate intervention effectiveness or evidence quality grades; consequently, no rigorous methodological quality assessment of included studies was conducted [29]. Our review was reported according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [26]. The PRISMA-ScR checklist is provided in Checklist 1.

Eligibility Criteria

The inclusion criteria for this scoping review were based on the Joanna Briggs Institute Scope Review Methodology Guide and structured using the participants, concept, context framework [30]. The inclusion and exclusion criteria are presented in Textbox 1.

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria. Eligibility criteria for the screening and inclusion process, including target population type, research setting, intervention measures, and study type.

Inclusion criteria

- All patients diagnosed with coronary heart disease, irrespective of nationality, gender, or ethnic background. Coronary heart disease encompasses, but is not limited to, the following clinical presentations: stable angina pectoris, acute coronary syndromes (including unstable angina pectoris, non–ST-segment elevation myocardial infarction, and ST-segment elevation myocardial infarction), and patients who have undergone percutaneous coronary intervention or coronary artery bypass grafting.
- Focus on the various digital health technologies used in cardiac rehabilitation (CR), including but not limited to mobile health apps, wearable devices, telemedicine or remote monitoring platforms, educational modules delivered via web-based or online platforms, virtual reality, and text messaging.
- The context of interest pertains to the application of digital health technologies within CR, where such technologies serve as complementary, alternative, augmenting, or extended means to traditional CR. Their purpose is to support, optimize, or enhance the delivery of CR services. This context is not restricted to specific countries or health care systems, permitting the inclusion of studies from diverse cultural, geographical, or medical settings.
- The types of literature included are empirical studies, which must be published in peer-reviewed journals. The study designs encompass randomized controlled trials, quasi-randomized controlled trials, and longitudinal before-and-after studies.

Exclusion criteria

- Research participants who have undergone cardiac or cardiopulmonary transplantation or patients with chronic heart failure.
- Studies involving participants younger than 18 years of age.
- Non-English language literature, duplicated publications, gray literature, studies where the full text is unavailable, conference abstracts, review papers, or qualitative research.

Information Sources

We conducted systematic searches of the following 5 electronic databases: PubMed, Web of Science (Clarivate), Embase (Elsevier), Cochrane Library (Wiley), and EBSCO (EBSCOhost). Searches were performed independently within each database interface, without using cross-database simultaneous search functionality.

Search Strategy

The literature search process for this study was reported in accordance with the PRISMA-S (extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews) guidelines [31]. The search strategy was independently developed by the research team based on the databases' subject term lists, without direct adaptation or use of other published scoping review strategies. The complete search strategies for each database are detailed in Multimedia Appendix 1. The researchers first conducted a preliminary search in the PubMed database to expand the keywords and use MeSH to determine standardized subject terms. Following this, a search strategy was developed, and an initial search was conducted in PubMed, with a brief analysis of the results. Two researchers collaboratively developed the final search strategy, which was reviewed by a third researcher. The PubMed search strategy is detailed in Textbox 2. To maintain sensitivity, no restrictions were applied regarding study design, language, or publication type. The search time frame spanned from the inception of each database to February 2026. Furthermore, as this scoping review aimed to map published evidence to delineate the existing evidence base, clinical trial registries were excluded from the search.

XSL•FO
RenderX

Textbox 2. PubMed search strategy.

((((((((((((((“Coronary Disease”[Mesh]) OR (“Myocardial Infarction”[Mesh])) OR (“Coronary Artery Disease”[Mesh])) OR (“Coronary Heart Disease*“[Title/Abstract])) OR (“Heart Attack*“[Title/Abstract])) OR (“Myocardial Infarct*“[Title/Abstract])) OR (“Cardiovascular Stroke*“[Title/Abstract])) OR (“acute coronary syndrome”[Title/Abstract])) OR (“angina pectoris”[Title/Abstract])) OR (“STEMI”[Title/Abstract])) OR (“NSTEMI”[Title/Abstract])) OR (“PCI”[Title/Abstract])) OR (“percutaneous coronary intervention”[Title/Abstract])) OR (“CABG”[Title/Abstract])) OR (“coronary artery bypass grafting”[Title/Abstract]))

AND

((((((((((((((((((((((((((((“Telemedicine”[Mesh]) OR (“Wearable Electronic Devices”[Mesh])) OR (“Digital Health”[Mesh])) OR (“Remote Patient Monitoring”[Mesh])) OR (“Text Messaging”[Mesh])) OR (“Virtual Medicine”[Title/Abstract])) OR (“Tele-Referral*“[Title/Abstract])) OR (“Mobile Health”[Title/Abstract])) OR (“mHealth”[Title/Abstract])) OR (“Telehealth”[Title/Abstract])) OR (“eHealth”[Title/Abstract])) OR (“Tele Intensive Care”[Title/Abstract])) OR (“Tele Care”[Title/Abstract])) OR (“Wearable Device*“[Title/Abstract])) OR (“Wearable Technolog*“[Title/Abstract])) OR (“Wearable Computer”[Title/Abstract])) OR (“Digital Health Technolog*“[Title/Abstract])) OR (“Health Technolog*“[Title/Abstract])) OR (“Short Message Service”[Title/Abstract])) OR (“digital therapeutics”[Title/Abstract])) OR (“smartwatch”[Title/Abstract])) OR (“fitness tracker”[Title/Abstract])) OR (“activity tracker”[Title/Abstract])) OR (“tele-rehabilitation”[Title/Abstract])) OR (“virtual care”[Title/Abstract])) OR (“mobile phone”[Title/Abstract])) OR (“cell phone”[Title/Abstract])) OR (“application”[Title/Abstract])) OR (“internet-based”[Title/Abstract])) OR (“web-based”[Title/Abstract])) OR (“online program”[Title/Abstract])))

AND

((((((((((((((((((“Cardiac Rehabilitation”[Mesh]) OR (“Secondary Prevention”[Mesh])) OR (“Cardiac Rehabilitation*“[Title/Abstract])) OR (“Cardiovascular Rehabilitation*“[Title/Abstract])) OR (“Secondary Prevention*“[Title/Abstract])) OR (“Disease Prevention*“[Title/Abstract])) OR (“Secondary Disease Prevention*“[Title/Abstract])) OR (“Early Therap*“[Title/Abstract])) OR (“Relapse Prevention*“[Title/Abstract])) OR (“exercise training”[Title/Abstract])) OR (“physical activity”[Title/Abstract])) OR (“lifestyle modification”[Title/Abstract])) OR (“behavior change”[Title/Abstract])) OR (“self-management”[Title/Abstract])) OR (“Exercise Therapy”[Title/Abstract])) OR (“Patient Education”[Title/Abstract])) OR (“Risk Factor Management”[Title/Abstract])) OR (“Medication Adherence”[Title/Abstract]))

Selection of Sources

The literature retrieval and screening process were independently conducted by 2 researchers trained in evidence-based medicine and possessing cardiovascular research experience, with the entire procedure subject to third-party oversight. The specific workflow was as follows: (1) Search results from all databases were imported into EndNote (Clarivate) reference management software, which automatically identified and excluded duplicate records; (2) 2 researchers independently conducted an initial screening of the remaining literature based on predefined inclusion and exclusion criteria, reviewing titles and abstracts; (3) screening results were cross-checked; discrepancies were resolved through third-party discussion until consensus was reached; and (4) full-text evaluation of initially selected papers determined final inclusion in the study.

Data Charting Process and Items

Data extraction was conducted using predesigned standardized forms by 2 researchers independently using Microsoft Excel to extract content covering the following core elements: (1) basic information: author, year of publication, and country; (2) study design: study type, sample size, follow-up duration, and control group configuration; (3) population characteristics: disease type; (4) intervention details: type of DHT, intervention duration, frequency of use, combination use, and combination method; (5) intervention objective; (6) outcome measures: primary outcomes and secondary outcomes; and (7) key findings: intervention effectiveness. In the event of disagreement during data extraction, a third researcher shall arbitrate the resolution.

Results

Selection of Sources of Evidence

As illustrated in Figure 1, this study strictly adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) process for literature screening [32]. The initial search yielded 8156 publications. After removing 2320 duplicates using EndNote, 2 researchers (XZ and ZL) independently screened titles and abstracts. Based on predefined inclusion criteria, 5836 publications were excluded. The remaining 407 publications underwent full-text assessment, resulting in the exclusion of 364 publications that did not meet the requirements. Ultimately, 43 studies were included in the analysis.

XSL•FO
RenderX

Figure 1. PRISMA flow diagram showing the identification of sources from databases and the screening and inclusion processes. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

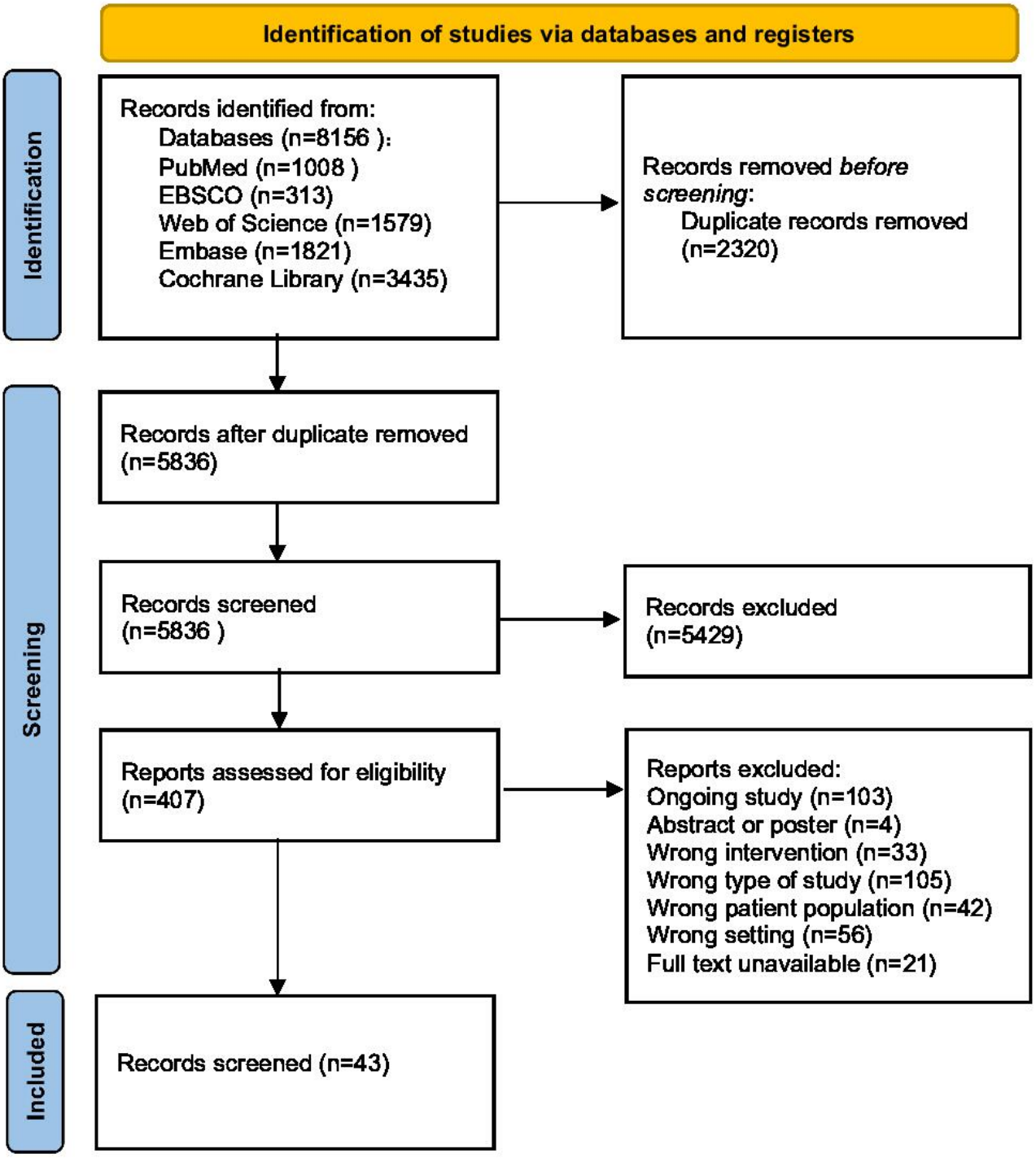

Characteristics of Sources of Evidence

Figure 2 presents an overview of the 43 included studies by year and geographical distribution. The publication span of the included studies ranges from 2014 to 2026, with a marked increase in relevant research since 2021, with 60.5% of studies published from 2021 to 2026. The geographical coverage spans 18 countries, with China accounting for the highest proportion of studies at 34.9% (15/43) [19,33-46].

Of the 43 studies included in total, the vast majority used randomized controlled designs, comprising 33 randomized controlled trials [33,36,38,39,41,43-69] and 7 randomized controlled pilot studies [34,35,37,42,70-72]. Among the remaining studies, 2 were quasi-randomized controlled trials [40,73], and 1 was a controlled before-and-after study [74]. Detailed characteristics of the studies included in this scoping review are presented in Table 1.

XSL•FO
RenderX

Figure 2. Publication trends for the 43 included studies by country and year (2014‐2026).

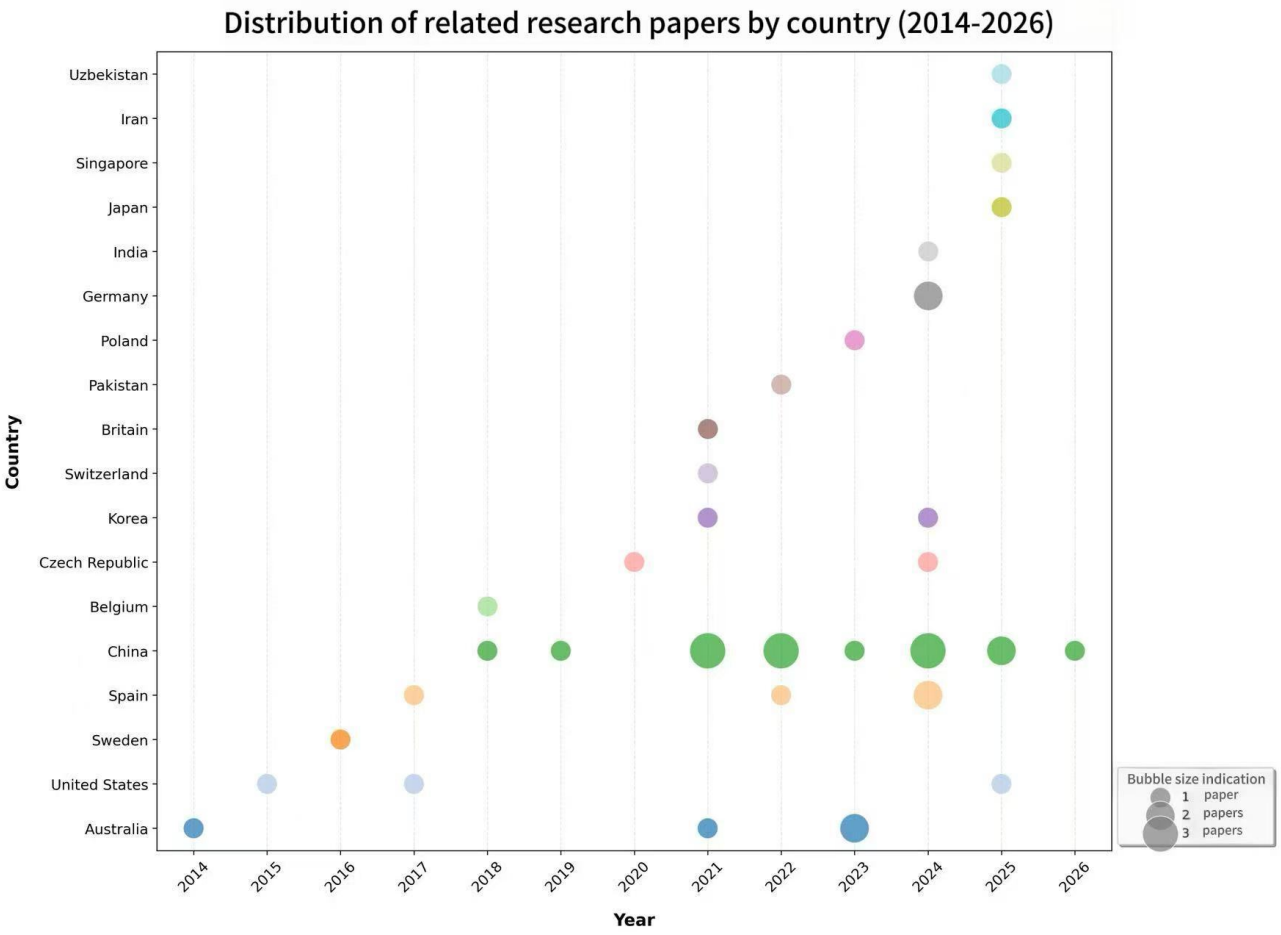

XSL•FO
RenderX

Table . Study characteristics[a].

Author (year)	Country	Design	Population (sample size)	Type and duration of intervention	Outcome measures	Statistically significant	Forms of intervention	Objectives of the intervention
Krzowski et al (2023) [47]	Poland	RCT[b]	Patients after acute myocardial infarction (n=100)	• Intervention group: Standard rehabilitation+afterAMI app • Control group: Standard rehabilitation • Duration: 6 months	• Primary: Rehospitalization or emergency department attendance within 6 months • Secondary: Risk factor control, NT-proBNP[c], disease knowledge	Partially valid	• App	• Health education • Alerts and reminders • Data monitor
Hong et al (2021) [33]	China	RCT	Patients with coronary artery disease (n=60)	• Intervention group: Health IT system grounded in self-efficacy theory+self-monitoring devices+fortnightly telephone interviews • Control group: Standard care (receiving identical intervention after a waiting period) • Duration: 3-month intervention period, 6-month follow-up	• Primary: Systolic blood pressure at 3 months • Secondary: Self-management behaviors, quality of life, diastolic blood pressure	Valid	• Website • Wearable devices • Telephone	• Health education • Data monitor • Feedback • Ensuring safety

XSL•FO
RenderX

Author (year)	Country	Design	Population (sample size)	Type and duration of intervention	Outcome measures	Statistically significant	Forms of intervention	Objectives of the intervention
Chan et al (2022) [34]	China	RCT pilot	Patients with stable coronary artery disease (n=139)	• Intervention group: 15-minute face-to-face ZTEx introduction+ZTEX email+ZTEX app • Control group: Equivalent duration and quantity of information on healthy eating and breathing exercises • Duration: 12 weeks	• Primary: Physical activity (IPAQ[d]), physical fitness • Secondary: Self-efficacy in exercise, well-being, quality of life	Valid	• App • Social media platform	• Health education • Data monitor • Alerts and reminders
Varnfield et al (2014) [48]	Australia	RCT	Patients after myocardial infarction (n=120)	• Intervention group: CAP-CR • Control group: Traditional centralized CR[e] • Duration: 6-week intervention period, followed by a 6-month maintenance phase	• Primary: Rehabilitation acceptance rate, adherence rate, completion rate • Secondary: Lifestyle factors, clinical indicators, quality of life	Valid	• App • Wearable devices • Digital video	• Health education • Data monitor • Personalization • Goal setting
Duan et al (2018) [35]	China	RCT pilot	Patients with coronary artery disease (n=136)			Valid	• Website	• Health education • Data monitor • Goal setting • Peer effect • Feedback

XSL•FO
RenderX

Author (year)	Country	Design	Population (sample size)	Type and duration of intervention	Outcome measures	Statistically significant	Forms of intervention	Objectives of the intervention
				• Intervention group: An 8-week web-based intervention grounded in the health action process approach (HAPA) model • Control group: A waiting-list control group • Duration: 8 weeks	• Primary: Physical activity, fruit and vegetable intake • Secondary: Healthy lifestyle, social cognitive indicators, health outcomes			
Xu et al (2024) [36]	China	RCT	Patients with CHD[f] after PCI[g] (n=147)	• Intervention group: Remote rehabilitation strategy based on the SCeiP[h] model • Control group: Standard exercise rehabilitation • Duration: 3 months	• Primary: Exercise adherence (number of days, duration) • Secondary: CR awareness, exercise program, exercise commitment	Valid	• Social media platform • Wearable devices • Telephone	• Health education • Data monitor • Feedback • Remotely adjust prescription
Avila et al (2018) [49]	Belgium	RCT	Patients with coronary artery disease (n=90)		• Primary: Peak oxygen uptake (VO_2 max) • Secondary: Physical activity, risk factors, quality of life	Valid	• Website • Wearable devices • Email • Telephone	• Health education • Data monitor • Feedback • Personalization • Goal setting

XSL•FO
RenderX

Author (year)	Country	Design	Population (sample size)	Type and duration of intervention	Outcome measures	Statistically significant	Forms of intervention	Objectives of the intervention
				• Intervention group 1: Home-based rehabilitation+remote monitoring • Intervention group 2: In-hospital rehabilitation • Control group: Standard care • Duration: 12 weeks				
Nishio et al (2025) [70]	Japan	RCT pilot	Patients with coronary artery disease (n=50)	• Intervention group: Wearable device + real-time monitoring+weekly text message and monthly videoconference guidance • Control group: Wearable device only • Duration: 12 weeks	• Primary: Changes in VO_2 max and anaerobic threshold oxygen uptake • Secondary: Daily activities, anxiety, quality of life	Partially valid	• Wearable devices • SMS text messaging • Remote counseling sessions	• Health education • Data monitor • Personalization
Li et al (2022) [37]	China	RCT pilot	Patients with coronary artery disease (n=300)	• Intervention group: Traditional follow-up+self-management app • Control group: Traditional hospital follow-up • Duration: 1 year		Valid	• App	• Health education • Data monitor • Personalization • Alerts and reminders

XSL•FO
RenderX

Author (year)	Country	Design	Population (sample size)	Type and duration of intervention	Outcome measures	Statistically significant	Forms of intervention	Objectives of the intervention
					• Primary: Proportion of patients receiving all guideline-recommended medications at 12 months • Secondary: Proportion receiving medication at 6 months, lipid control rate, blood pressure control rate			
Cruz-Cobo et al (2024) [50]	Spain	RCT	Patients with acute coronary syndrome after PCI (n=300)	• Intervention group: Standard care + eMOTIVA app • Control group: Standard care • Duration: 6 months	• Primary: Mediterranean diet adherence, physical activity, sedentary time, functional capacity, smoking cessation, disease knowledge, app satisfaction • Secondary: No	Valid	• App • Remote counseling sessions • Digital video • SMS text messaging	• Health education • Data monitor • Personalization • Alerts and reminders • Goal setting • Feedback • Gamification • Reward mechanism
Bernal-Jiménez et al (2024) [51]	Spain	RCT	Patients with CHD after PCI (n=128)			Valid	• App • Remote counseling sessions • Email • Telephone • SMS text messaging	• Health education • Data monitor • Personalization • Alerts and reminders • Goal setting • Feedback

XSL•FO
RenderX

Author (year)	Country	Design	Population (sample size)	Type and duration of intervention	Outcome measures	Statistically significant	Forms of intervention	Objectives of the intervention
				• Intervention group: Standard care+Interactive mHealth App (EVITE app) • Control group: Standard care • Duration: 9 months	• Primary: Mediterranean diet adherence, food frequency, physical activity, smoking, knowledge, treatment adherence, quality of life, satisfaction • Secondary: No			
Bae et al (2021) [52]	Korea	RCT	Patients after the first PCI (n=879)	• Intervention group: Standard care + website + 4 weekly lifestyle text messages • Control group: Standard care • Duration: 6 months	• Primary: Low-density lipoprotein cholesterol, systolic blood pressure, BMI • Secondary: Lifestyle modifications, adherence to health behaviors	Invalid	• Wearable devices • SMS text messaging	• Health education • Alerts and reminders
Su and Yu (2021) [38]	China	RCT	Patients with coronary artery disease (n=146)	• Intervention group: Standard care+NeCR • Control group: Standard care • Duration: 12 weeks	• Primary: Lifestyle behavior changes • Secondary: Self-efficacy, quality of life, physiological indicators	Valid	• Social media platform • Remote counseling sessions • Digital video	• Health education • Alerts and reminders • Feedback • Goal setting • Social support • Data monitor • Emotional support or counseling
Dodson et al (2025) [53]	United States	RCT				Partially valid		

XSL•FO
RenderX

Author (year)	Country	Design	Population (sample size)	Type and duration of intervention	Outcome measures	Statistically significant	Forms of intervention	Objectives of the intervention
			Patients aged 65 years or older with ischemic heart disease (n=400)	• Intervention group: mHealth-CR software on tablet devices+remote monitoring+weekly telephone guidance • Control group: standard care • Duration: 3 months	• Primary: Change in 6MWD[i] from baseline to 3 months • Secondary: Health status, residual angina, impairment in activities of daily living		• App • Remote counseling sessions • Telephone • Wearable devices	• Health education • Feedback • Goal setting • Data monitor
Dorje et al (2019) [19]	China	RCT	Patients with CHD after PCI (n=312)	• Intervention group: SMART-CR/SP • Control group: Standard care • Duration: 6 months	• Primary: Change in 6MWD at 2 months and 6 months • Secondary: No	Valid	• Social media platform • Wearable devices • Telephone • Remote counseling sessions	• Health education • Feedback • Data monitor • Alerts and reminders
Hisam et al (2022) [54]	Pakistan	RCT	After an acute coronary syndrome (n=160)	• Intervention group: Standard care+MCard • Control group: Standard care • Duration: 24 weeks	• Primary: Health-related quality of life • Secondary: No	Valid	• App • Wearable devices • SMS text messaging	• Health education • Data monitor • Alerts and reminders • Emotional support or counseling
Zheng et al (2024) [39]	China	RCT	Patients with CHD after PCI (n=106)	• Intervention group: Standard care+HCT • Control group: Standard care • Duration: 3 months	• Primary: 6-minute walk test, quality of life (SF-12[j]), disease burden (FBIS[k]), cardiac function (LVEF[l]) • Secondary: No	Valid	• App • Wearable devices • Telephone	• Health education • Data monitor • Goal setting • Remotely adjust prescription • Feedback
Ma et al (2021) [40]	China	Q-RCT[m]				Valid		

XSL•FO
RenderX

Author (year)	Country	Design	Population (sample size)	Type and duration of intervention	Outcome measures	Statistically significant	Forms of intervention	Objectives of the intervention
			Patients with CHD after PCI (n=335)	• Intervention group: HBCR • Control group: Routine care • Duration: 42 months	• Primary: Incidence of major adverse cardiovascular and stroke events • Secondary: Exercise capacity, quality of life, risk factor control		• App • Wearable devices • Digital video	• Health education • Data monitor • Training courses • Ensuring safety • Alerts and reminders
Xu et al (2024) [41]	China	RCT	Patients with coronary artery disease (n=108)	• Intervention group 1 (individual): Gamified intervention • Intervention group 2 (team): Gamification+social interaction • Control group: Daily step target • Duration: 12-week intervention, 12-week follow-up	• Primary: Daily step count variation, proportion of days achieving target • Secondary: Autonomy, sense of connection, intrinsic motivation	Valid	• App	• Goal setting • Feedback • Social support • Reward mechanism • Gamification
Widmer et al (2017) [55]	United States	RCT	Patients with acute coronary syndrome after PCI (n=80)	• Intervention group: Standard CR+DHI • Control group: Standard CR • Duration: 3-month intervention, 6-month outcome assessment	• Primary: Related emergency department visits and readmissions • Secondary: Risk factors and lifestyle factors at 90 days	Partially valid	• Website • Remote counseling sessions	• Health education • Data monitor
Yudi et al (2021) [56]	Australia	RCT				Valid		

XSL•FO
RenderX

Author (year)	Country	Design	Population (sample size)	Type and duration of intervention	Outcome measures	Statistically significant	Forms of intervention	Objectives of the intervention
			Patients with acute coronary syndromes (n=206)	• Intervention group: Standard care+S-CRP • Control group: Standard care • Duration: 8 weeks	• Primary: Change in 6MWD at 8 weeks • Secondary: CR participation, risk factors, psychological indicators		• App • Remote counseling sessions	• Health education • Data monitor • Feedback • Ensuring safety • Personalization
Gallagher et al (2023) [57]	Australia	RCT	Patients with coronary artery disease (n=390)	• Intervention group: MyHeartMate app • Control group: Standard care • Duration: 6 months	• Primary: Self-reported physical activity (METs[n]) • Secondary: Blood lipids, blood pressure, BMI, smoking	Invalid	• App • Digital video • Email	• Gamification • Data monitor • Social support • Reward mechanism • Training courses
Wohlfahrt et al (2024) [71]	Czech Republic	RCT pilot	Patients after myocardial infarction (n=64)	• Intervention group: Smartwatch step tracking+remote monitoring by nurses • Control group: 150 minutes per week of moderate-intensity exercise recommendation • Duration: 3 months, followed by crossover after 3 months	• Primary: VO_2 max • Secondary: Body weight, 6MWD, quality of life	Valid	• Wearable devices • Telephone	• Data monitor • Alerts and reminders • Real-time monitoring • Goal setting
	Singapore	RCT pilot				Valid		

XSL•FO
RenderX

Author (year)	Country	Design	Population (sample size)	Type and duration of intervention	Outcome measures	Statistically significant	Forms of intervention	Objectives of the intervention
Ramachandran et al (2025) [72]			Patients after acute myocardial infarction (n=50)	• Intervention group: Home-based remote rehabilitation • Control group: Centralized CR • Duration: 6 weeks	• Primary: Use of CR • Secondary: Functional capacity, risk factors, self-reported behaviors		• App • Website • Wearable devices • Telephone • Remote counseling sessions	• Health education • Training courses • Data monitor • Alerts and reminders • Goal setting • Remotely adjust prescription
Jo et al (2024) [58]	Korea	RCT	Patients after acute myocardial infarction (n=48)	• Intervention group: Mobile app-based rehabilitation • Control group: Conventional home-based rehabilitation+biweekly telephone supervision • Duration: 6 weeks	• Primary: VO_2 max • Secondary: Resting heart rate, blood pressure, quality of life, psychological indicators	Invalid	• App • Telephone • Wearable devices	• Health education • Data monitor • Goal setting • Ensuring safety • Real-time monitoring • Training courses
Fallah et al (2025) [59]	Iran	RCT	Patients with myocardial infarction (n=144)	• Intervention group: HAPA-based mobile app • Control group: No specific intervention • Duration: 8 weeks	• Primary: Physical activity (IPAQ) • Secondary: HAPA-related psychological constructs	Valid	• App	• Health education • Training courses • Feedback • Social support • Personalization
Li et al (2023) [42]	China	RCT pilot	Patients with acute myocardial infarction after PCI (n=60)			Valid	• Website • Wearable devices	

XSL•FO
RenderX

Author (year)	Country	Design	Population (sample size)	Type and duration of intervention	Outcome measures	Statistically significant	Forms of intervention	Objectives of the intervention
				• Intervention group: 5G IoT platform • Control group: Conventional CR training within the hospital. • Duration: 3 months.	• Primary: Cardiorespiratory fitness (VO_2 max, MET) • Secondary: Physiological indicators, psychological indicators, adherence, satisfaction.			• Health education • Training courses • Feedback • Remotely adjust prescription • Personalization • Emotional support or counseling
Waranski et al (2024) [73]	Germany	Q-RCT	Patients with coronary artery disease (n=169)	• Intervention group: Personalized messages (twice weekly) • Control group: Routine outpatient care • Duration: 6 months	• Primary: Routine physical activity (≥150 minutes per week) and daily activities at 6 months • Secondary: Psychological indicators, self-efficacy, quality of life	Valid	• App • SMS text messaging • Telephone	• Health education • Alerts and reminders • Goal setting • Personalization
Ni et al (2022) [43]	China	RCT	Patients with coronary artery disease (n=230)	• Intervention group: WeChat Message Express • Control group: WeChat only • Duration: 60-day intervention, 30-day follow-up	• Primary: Medication Adherence Score (Voils Extent Scale) • Secondary: Heart rate, systolic blood pressure, diastolic blood pressure	Valid	• SMS text messaging • Social media platform	• Health education • Alerts and reminders
Liu et al (2026) [44]	China	RCT	Patients after PCI (n=180)			Valid		

XSL•FO
RenderX

Author (year)	Country	Design	Population (sample size)	Type and duration of intervention	Outcome measures	Statistically significant	Forms of intervention	Objectives of the intervention
				• Intervention group: Standard rehabilitation+eHealth platform based on the Persuasive systems design (PSD) model • Control group: Standard rehabilitation • Duration: 12-week intervention, 12-week follow-up	• Primary: Physical activity level (IPAQ) • Secondary: Exercise endurance (6MWD), self-perceived fatigue, exercise self-efficacy, quality of life		• App • Social media platform • Telephone	• Health education • Alerts and reminders • Training courses
Bruggmann et al (2021) [60]	Switzerland	RCT	Patients with acute coronary syndromes (n=60)	• Intervention group: Routine care+viewing of interactive video+brief interview with pharmacist • Control group: Routine care • Duration: 6 months	• Primary: Differences in medication adherence at 1, 3, and 6 months (ARMS[o]) • Secondary: Disease knowledge, readmission, emergency department visits, satisfaction	Partially valid	• Website	• Health education • Training courses • Feedback
Zhang et al (2025) [45]	China	RCT	Patients with coronary artery disease (n=62)	• Intervention group: Smartwatch-assisted CR • Control group: Standard CR • Duration: 3 months		Valid	• Wearable devices • App	• Health education • Real-time monitoring • Data monitor • Feedback • Alerts and reminders

XSL•FO
RenderX

Author (year)	Country	Design	Population (sample size)	Type and duration of intervention	Outcome measures	Statistically significant	Forms of intervention	Objectives of the intervention
					• Primary: HBCR adherence (HET-AQ[p]) • Secondary: cardiopulmonary function, anxiety, depression, quality of life			
Patterson et al (2023) [61]	Australia	RCT	Patients with coronary artery disease (n=120)	• Intervention group: Conventional rehabilitation+Vire app • Control group: Conventional rehabilitation • Duration: 12 months	• Primary: Nonelective hospital admissions and emergency department visits • Secondary: Sedentary behavior, BMI, waist circumference, quality of life, cost-effectiveness	Invalid	• Wearable devices • App	• Data monitor • Feedback • Alerts and reminders • Personalization
Batalik et al (2020) [62]	Czech Republic	RCT	Patients with coronary artery disease (n=56)	• Intervention group: Home-based remote rehabilitation • Control group: Routine outpatient rehabilitation • Duration: 12 weeks	• Primary: VO_2 max • Secondary: Quality of life (SF-36[q]), training adherence	Invalid	• Wearable devices • App • Telephone	• Data monitor • Feedback • Emotional support or counseling
Dalli Peydró et al (2022) [63]	Spain	RCT	Patients with acute coronary syndromes (n=67)			Valid	• Wearable devices • App	• Data monitor • Feedback • Personalization • Training courses

XSL•FO
RenderX

Author (year)	Country	Design	Population (sample size)	Type and duration of intervention	Outcome measures	Statistically significant	Forms of intervention	Objectives of the intervention
				• Intervention group: Remote CR • Control group: Center-based CR • Duration: 8 weeks	• Primary: Self-reported physical activity (IPAQ) • Secondary: VO_2 max, blood lipids, body weight, quality of life, time to return to work			
Bravo-Escobar et al (2017) [64]	Spain	RCT	Patients with stable, intermediate-risk coronary artery disease (n=28)	• Intervention group: Weekly hospital sessions+3 home training sessions • Control group: Routine hospital rehabilitation (3 sessions per week) • Duration: 2 months	• Primary: Physical fitness, risk profile, cardiovascular complications, quality of life • Secondary: No	Partially valid	• Wearable devices • App	• Data monitor • Feedback • Goal setting
Widmer et al (2015) [74]	United States	CBA[r]	Patients after myocardial infarction (n=42)	• Intervention group 1: During CR+PHA • Intervention group 2: Following CR+PHA • Control group: Standard CR during the corresponding time period • Duration: 3 months	• Primary: Changes in risk factors and readmissions or emergency department visits after 3 months • Secondary: No	Valid	• Website	• Data monitor • Feedback • Alerts and reminders • Health education
	Sweden	RCT				Valid		

XSL•FO
RenderX

Author (year)	Country	Design	Population (sample size)	Type and duration of intervention	Outcome measures	Statistically significant	Forms of intervention	Objectives of the intervention
Johnston et al (2016) [65]			Patients after myocardial infarction (n=174)	• Intervention group: Full-featured CR app • Control group: Simplified version app • Duration: 6 months	• Primary: Nonadherence score based on app records • Secondary: Risk factors, quality of life, device satisfaction		• App • SMS text messaging • Telephone	• Data monitor • Feedback • Alerts and reminders • Health education
Kumar et al (2024) [66]	India	RCT	Patients after CABG (n=40)	• Intervention group: eMedia-supported exercise rehabilitation • Control group: Standard care • Duration: 12 weeks	• Primary: Functional capacity (6MWT), quality of life (WHOQOL-BREF[s]), physical activity (GPAQ[t]) • Secondary: No	Valid	• Website • App	• Feedback • Health education • Ensuring safety • Data monitor
Bretschneider et al (2024) [67]	Germany	RCT	Patients with coronary artery disease (n=354)	• Intervention group: Standard care plus Mebix • Control group: Standard care • Duration: 12 months	• Primary: Disease-specific quality of life (HeartQoL[u]) and body weight • Secondary: Cardiovascular risk, occupational prognosis	Valid	• App • SMS text messaging • Telephone	• Data monitor • Feedback • Alerts and reminders • Health education • Training courses
Herring et al (2021) [68]	Britain	RCT	Patients with coronary artery disease (n=291)			Invalid	• Group meeting • SMS text messaging	• Health education • Alerts and reminders

XSL•FO
RenderX

Author (year)	Country	Design	Population (sample size)	Type and duration of intervention	Outcome measures	Statistically significant	Forms of intervention	Objectives of the intervention
				• Intervention group: 2 structured educational sessions+follow-up text message support • Control group: Standard care • Duration: 12 months	• Primary: Changes in overall physical activity at 12 months (GENEActiv) • Secondary: Functional, cardiovascular, biochemical, and patient-reported outcomes			
Li et al (2025) [46]	China	RCT	Patients with coronary artery disease (n=294)	• Intervention group: Personalized face-to-face education+i-CARE app+pedometer • Control group: Standard care+pedometer • Duration: 6 months	• Primary: Self-care behavior in CHD (SC-CHDI[v]) • Secondary: Health status, quality of life, physiological indicators	Valid	• App • Wearable devices	• Data monitor • Feedback • Health education • Real-time monitoring • Social support • Personalization
Khikmatova Madina et al [69] (2025)	Uzbekistan	RCT	Patients after myocardial infarction (n=300)	• Intervention group: Standard care+wearable activity monitoring device and accompanying app • Control group: Standard care • Duration: 6 months	• Primary: Rehabilitation adherence • Secondary: Readmission rate, mortality rate, ejection fraction, exercise capacity	Valid	• App • Wearable devices • Remote counseling sessions	• Health education • Data monitor • Real-time monitoring • Alerts and reminders

XSL•FO
RenderX

[a]Information regarding authors, publication year, participants, study design, study outcomes, and methodology for 43 studies. Digital health technology intervention name: zero-time exercise (ZTEx), care assessment platform of cardiac rehabilitation (CAP-CR), nurse-led eHealth cardiac rehabilitation (NeCR), mobile health cardiac rehabilitation (mHealth-CR), smartphone and social media–based cardiac rehabilitation and secondary prevention (SMART-CR/SP), mobile health augmented cardiac rehabilitation (MCard), home-based cardiac telerehabilitation (HCT), home-based cardiac rehabilitation (HBCR), digital health intervention (DHI), smartphone-based cardiac rehabilitation program (S-CRP), fifth generation mobile communication technology Internet of Things platform (5G IoT platform), persuasive systems design (PSD), coronary artery bypass grafting (CABG), internet-based cardiac rehabilitation enhancement (i-CARE).

[b]RCT: randomized controlled trial.

[c]NT-proBNP: N-terminal pro-brain natriuretic peptide.

[d]IPAQ: International Physical Activity Questionnaire.

[e]CR: cardiac rehabilitation.

[f]CHD: coronary heart disease.

[g]PCI: percutaneous coronary intervention.

[h]SCeiP: Self-Monitoring, Coaching, e-Health, Interactive Feedback, and Personalization.

[i]6MWD: 6-minute walk distance.

[j]SF-12: Short Form 12 Health Survey.

[k]FBIS: Framingham Burden of Illness Scale.

[l]LVEF: left ventricular ejection fraction.

[m]Q-RCT: quasi-randomized controlled trial.

[n]MET: metabolic equivalents of task.

[o]ARMS: Adherence to Refills and Medications Scale.

[p]HETAQ: Home-Based Cardiac Rehabilitation Adherence Questionnaire.

[q]SF-36: Short Form 36 Health Survey.

[r]CBA: controlled before-after study.

[s]WHOQOL-BREF: World Health Organization Quality of Life-BREF.

[t]GPAQ: Global Physical Activity Questionnaire.

[u]HeartQoL: Heart Disease-Specific Quality of Life Questionnaire.

[v]SC-CHDI: Self-Care Behaviour in Coronary Heart Disease Inventory.

Results of Individual Sources of Evidence

In this study, we analyzed data from 43 research papers and generated an evidence gap map (Table 2) illustrating the application forms and objectives of DHT interventions.

XSL•FO
RenderX

Table . Evidence gap analysis of digital health technologies in cardiac rehabilitation for patients with coronary heart disease, based on 43 included studies.

Author (year)	Forms of intervention										Objectives of the intervention													
	Wearable devices	Application	Website	Digital video	Social media platform	Telephone	Group meeting	Remote counseling sessions	Email	SMS text messaging	Goal setting	Feedback	Reward mechanism	Gamification	Alarts and reminds	Personalization	Health education	Training courses	Data monitor	Real-time monitoring	Ensuring safety	Remotely adjust prescription	Peer effect	Social support	Emotional support	Counseling
Krzowski et al (2022) [47]		✓													✓		✓		✓							
Hung et al (2021) [33]	✓		✓			✓						✓					✓		✓		✓					
Chan et al (2022) [34]		✓			✓										✓		✓		✓							
Vanfield et al (2014) [48]	✓	✓		✓							✓					✓	✓		✓							
Duan et al (2019) [35]			✓								✓	✓					✓		✓				✓			
Xu et al (2024) [36]	✓				✓	✓						✓	✓	✓			✓		✓			✓		✓		
Avila et al (2019) [49]	✓		✓			✓			✓		✓	✓				✓	✓		✓							
Nilio et al (2025) [70]	✓							✓		✓						✓	✓		✓							
Li et al (2022) [37]		✓													✓	✓	✓		✓							

XSL•FO
RenderX

Author (year)	Forms of intervention								Objectives of the intervention														
Gómez-Cobo et al (2024) [50]		✓		✓			✓		✓	✓	✓	✓	✓	✓	✓	✓		✓					
Barahona-Juárez et al (2024) [51]		✓				✓		✓	✓	✓	✓	✓			✓	✓	✓		✓				
Bae et al (2021) [52]	✓								✓						✓		✓						
Su and Yu (2021) [38]				✓	✓			✓			✓	✓			✓		✓		✓			✓	✓	✓
Dunson et al (2025) [53]	✓	✓				✓		✓			✓	✓					✓		✓				
Dorje et al (2019) [19]	✓				✓	✓		✓				✓			✓		✓		✓				
Hsin et al (2022) [54]	✓	✓							✓						✓		✓		✓				✓	✓
Zheng et al (2024) [39]	✓	✓				✓				✓	✓						✓		✓		✓		
Ma et al (2021) [40]	✓	✓		✓											✓		✓	✓	✓			✓	
Xu et al (2024) [41]		✓								✓	✓												

XSL•FO
RenderX

Author (year)	Forms of intervention								Objectives of the intervention											
Wilmer et al (2017) [55]			✓			✓							✓		✓					
Yudi et al (2021) [56]		✓				✓				✓			✓	✓		✓		✓		
Gallagher et al (2023) [57]		✓		✓			✓				✓	✓			✓	✓			✓	
Wu et al (2024) [71]	✓				✓				✓				✓			✓	✓			
Ramadanandan et al (2025) [72]	✓	✓	✓		✓	✓			✓				✓		✓	✓	✓			✓
Jo et al (2024) [58]	✓	✓			✓				✓						✓	✓	✓	✓	✓	
Fallah et al (2025) [59]		✓								✓			✓	✓	✓				✓	
Li et al (2023) [42]	✓		✓							✓			✓	✓	✓				✓		✓	✓
Wiwanski et al (2024) [73]		✓			✓			✓	✓				✓	✓	✓					
Ni et al (2022) [43]				✓				✓					✓		✓					

XSL•FO
RenderX

Author (year)	Forms of intervention								Objectives of the intervention									
Liu et al (2016) [44]		✓		✓	✓						✓		✓	✓				
Bruggeman et al (2021) [60]			✓							✓			✓	✓				
Zhang et al (2025) [45]	✓	✓								✓		✓			✓	✓		
Patterson et al (2023) [61]	✓	✓								✓		✓	✓			✓		
Banik et al (2020) [62]	✓	✓			✓					✓					✓			✓ ✓
Dalli Peydró et al (2022) [63]	✓	✓								✓			✓	✓	✓			
Bravo-Escobar et al (2017) [64]	✓	✓							✓	✓					✓			
Wilmer et al (2016) [74]			✓							✓		✓	✓		✓			
Johnston et al (2016) [65]		✓			✓			✓		✓		✓	✓		✓			

XSL•FO
RenderX

Author (year)	Forms of intervention								Objectives of the intervention									
Kumar et al (2024) [66]		✓	✓						✓			✓		✓		✓		
Rhinder et al (2024) [67]		✓			✓			✓	✓		✓	✓	✓	✓				
Hering et al (2024) [68]						✓		✓			✓		✓					
Li et al (2025) [46]	✓	✓							✓			✓	✓		✓	✓		✓
Kikmatova et al (2025) [69]	✓	✓					✓				✓		✓		✓	✓		

The Form of DHT in CR for Patients With CHD

The 43 studies included in this research demonstrate that DHTs exhibit significant diversity in their application within CR for patients with CHD. These technologies can be categorized into 3 core groups (Table 3): digital health tools, real-time remote support, and asynchronous communication. Among these, digital health tools represent the most prevalent intervention form, enabling patients with CHD to undertake self-management and monitoring primarily through devices or software. This includes apps (28/43, 65.1%) [34,37,39-41,44-48,50,51,53,54,56-59,61-67,69,72,73], wearable devices (22/43, 51.1%) [19,33,36,39,40,42,45,46,49,52-54,58,61-64,69-72], websites (9/43, 20.9%) [33,35,42,49,55,60,66,72,74], and social media platforms (6/43, 13.9%) [19,34,36,38,43,44]. Wearable devices encompass smartwatches, heart rate monitors, fitness trackers, and pedometers, primarily used for real-time monitoring of physiological indicators and exercise data. Real-time remote support involves direct interpersonal interaction via voice or video, covering telephone (15/43, 34.8%) [19,33,36,38,39,44,49,51,53,57,62,65,67,72,73], remote counseling sessions, and group meetings. Asynchronous communication delivers reminders, education, and support through non–real-time information exchange, chiefly via SMS text messaging (10/43, 23.2%) [43,50-52,65,67,68,70,73] and email. As shown in Table 2, multitechnology combined interventions have become the predominant model. A substantial 83.7% (36/43) of studies used combinations of 2 or more digital technologies, such as “app+wearable device” [39,40,45,46,48,53,54,58,61-64,69,72] and “social media platform+wearable device+telephone” [19,36]. Some studies further integrated digital technologies with traditional rehabilitation methods like face-to-face guidance and offline education, forming blended online-offline rehabilitation models. This landscape not only reflects varying levels of technological support, from standalone tools to interpersonal interactions, but also signals the trend toward systematized and diversified digital CR.

XSL•FO
RenderX

Table . A total of 3 categories and 10 specific forms of digital health technology application in cardiac rehabilitation interventions.

Type	Core functionality	Content
Digital health tools	Provide patients with tools for independent health management through a technology platform, emphasizing self-monitoring and personalized interaction.	• Wearable devices • App • Website • Digital video • Social media platform
Real-time remote support	Provides real-time, person-to-person professional support or peer interaction through synchronous communication technology, with high interactivity.	• Telephone • Group meeting • Remote counseling sessions
Asynchronous communication	Reminders, education, and support are provided through non–real-time information transmission methods, which are flexible and not restricted by time or space.	• Email • SMS text messaging

The Objectives of DHT in CR for Patients With CHD

From the perspective of intervention objectives, the application of DHTs in CHD rehabilitation exhibits distinct functional stratification. Health education (36/43, 83.7%) [19,33-40,58-60,65-70,72-74], data monitor (34/43, 79.1%) [19,33-40,45-51,53-58,61-67,69-74], and reminders and alerts (18/43, 41.9%) [19,34,37,38,40,43-45,47,50-52,54,55,71-73] form the core functional layer, each accounting for over 80% of applications in the included studies. Feedback, goal setting, and personalized interventions constitute the secondary core functional layer, with application rates ranging between 50% and 70%. Additionally, some studies integrated distinctive features such as gamification, reward mechanisms, social support, emotional support or counseling, and remotely adjust prescription to address patients' diverse rehabilitation needs. Building upon this, this study systematically categorized the intervention objectives across 43 publications, identifying 15 specific types grouped into 4 major categories: first, motivation and guidance, encompassing goal setting, feedback, reward mechanisms, gamification, and reminders and alerts, aimed at incentivizing patients with CHD to complete rehabilitation behaviors and enhance adherence; second, foundation of knowledge and skills, centered on health education and training courses to help patients with CHD build the knowledge base and self-management capabilities required for disease management; third, monitoring and security, including data monitoring, real-time monitoring, and ensuring safety for physiological indicator tracking, risk assessment, and safety protection during rehabilitation; and fourth, social and group dynamics, leveraging peer effects and social support mechanisms to use social relationships and group interactions to promote patient adherence to rehabilitation behaviors. The specific composition is detailed in Table 4. As shown in Table 2, health education emerged most frequently, underscoring the central role of knowledge transfer in contemporary digital rehabilitation practice. Notably, the vast majority of studies adopted multipurpose integrated intervention strategies, organically combining educational, motivational, monitoring, and social support functions rather than relying on single technological approaches. It is precisely this composite application model that has transformed DHTs from fragmented tools into systematic rehabilitation support systems, significantly enhancing the holistic nature and continuity of rehabilitation interventions.

Table . A total of 4 key intervention objectives and 15 specific types of digital health technologies in cardiac rehabilitation for individuals with coronary heart disease.

Type	Core objective	Content
Motivation and guidance	Motivate patients and guide them to complete specific behaviors.	• Goal setting • Feedback • Reward mechanism • Gamification • Alerts and reminders • Personalization
Foundation of knowledge and skills	Provide necessary information and cultivate patients' self-management skills.	• Health education • Training courses
Monitoring and security	Track data, assess risks, and provide a safety net.	• Data monitor • Real-time monitoring • Ensuring safety • Remotely adjust prescription
Social and group dynamics	Using social relationships and group dynamics to promote patient adherence and change.	• Peer effect • Social support • Emotional support or counseling

XSL•FO
RenderX

Evaluation Criteria of DHT in CR for Patients With CHD

The outcome measures included in the study encompass 5 major categories: clinical physiological indicators, rehabilitation behavioral indicators, patient-reported outcomes, rehabilitation service use rates, and technical feasibility. Clinical physiological indicators include peak oxygen uptake, 6-minute walk distance, and blood pressure. Rehabilitation behavioral indicators include exercise adherence, medication adherence, physical activity levels, dietary adherence, and sedentary time. Patient-reported outcomes encompassed quality of life, self-efficacy, anxiety and depression levels, disease knowledge, and rehabilitation satisfaction. Rehabilitation service use metrics included rehabilitation acceptance rate, adherence rate, completion rate, readmission rate, and emergency department visit rate. Technical feasibility referred to patient satisfaction with the DHT used.

Clinical Efficacy and Physiological Indicators

Clinical efficacy and physiological indicators constitute the core dimensions for evaluating the effectiveness of DHTs, with over 60% of studies incorporating them as primary outcomes. These encompass 3 specific levels: cardiopulmonary function, physical capacity and strength, and clinical end-point events. Cardiopulmonary function stands as the most critical indicator, with peak oxygen uptake [39,42,44,49,58,66,70,71] and 6-minute walk distance [42,49,58,62,63,70,71] being the most widely applied measures. Combining remote monitoring, wearable devices, and online guidance can effectively improve patients' cardiopulmonary function, as measured by peak oxygen uptake and 6-minute walk distance [49,70]. Physical function and strength serve as supplementary dimensions, encompassing muscle endurance and overall physical capacity. Dalli Peydró et al [63] confirmed that remote rehabilitation improves patients' physical activity capabilities. Regarding clinical end points, over 20 studies evaluated blood pressure, lipid profiles, N-terminal pro-brain natriuretic peptide, and left ventricular ejection fraction. Li et al [37] found that app-based interventions increased lipid control rates; however, no consistent conclusions have emerged regarding long-term outcomes such as readmission rates and major adverse cardiovascular events. For instance, Krzowski et al [47] did not demonstrate a significant advantage of digital interventions in reducing readmission rates, suggesting that further research is needed to substantiate long-term efficacy.

Health Behavior and Lifestyle

Health behaviors and lifestyle constitute core factors in improving the long-term prognosis of patients with CHD, with over half of the studies incorporating them into evaluations. These are categorized into 2 dimensions: exercise behavior and daily lifestyle. Regarding exercise behavior, key indicators include physical activity levels, exercise adherence, number of exercise days, and duration. Xu et al [41] demonstrated in a telerehabilitation study based on the Self-Monitoring, Coaching, e-Health, Interactive Feedback, and Personalization model that the intervention group exhibited significantly superior exercise adherence and duration compared to the control group. Varnfield et al [48] confirmed that smartphone-based home rehabilitation effectively enhances physical activity levels in patients with postmyocardial infarction. Optimization of daily lifestyle habits has also garnered significant attention, encompassing indicators such as medication adherence, dietary compliance, fruit and vegetable intake, sedentary time, and smoking cessation behavior. Interventions incorporating digital tools have shown positive effects on dietary adherence, sedentary time, and smoking cessation [50,51]. DHTs effectively enhance patient motivation for behavioral change through personalized reminders, real-time feedback, and adaptive goal-setting, thereby promoting the sustained maintenance of healthy behaviors.

Patient-Reported Outcomes and Cognitive Function

Patient-reported outcomes and cognitive-related indicators constitute crucial dimensions for evaluating the humanistic value of DHTs. A total of 34 such indicators were incorporated into the studies as assessment criteria, encompassing domains such as quality of life, social cognition and support, disease knowledge, and psychological state [33-36,38-42,44-51,53,54,56,58-68,70,71,73]. Quality of life emerged as the most frequently assessed outcome, with multiple studies using scales such as the Short Form 36 Health Survey, EQ-5D, and Heart Disease-Specific Quality of Life Questionnaire. Dodson et al [53] demonstrated positive trends in mobile health interventions improving health status among older patients, while Hisam et al [54] found that mobile health interventions significantly enhanced quality of life in patients with postacute coronary syndrome. Regarding knowledge and self-management capabilities, health education emerged as the most prevalent intervention objective. Its efficacy was evaluated through indicators such as cardiovascular risk factor knowledge and self-management competence. Chan et al [34] confirmed that 0-time exercise interventions can enhance patients' exercise self-efficacy. In the domain of social cognition and support, Duan et al [35] incorporated social cognitive outcomes into its evaluation. DHTs effectively enhance patient cognition through personalized education and interactive feedback, providing crucial support for the long-term maintenance of rehabilitation outcomes.

Program Participation and Adherence

Participation rates and adherence in CR are core indicators for assessing the real-world feasibility of DHTs. These encompass rehabilitation program participation rates, adherence rates, completion rates, alongside patient satisfaction and perceived acceptability. Varnfield et al [48] found that smartphone-based home rehabilitation significantly improved rehabilitation uptake, adherence, and completion rates among patients with postmyocardial infarction compared to conventional rehabilitation, providing robust evidence for digital technologies enhancing rehabilitation engagement. Ramachandran et al [72] further validated the advantages of home-based remote rehabilitation in improving rehabilitation use rates. Patient satisfaction, acceptability, and perceived ease of use of the technology are also crucial evaluation components. Studies by Bernal-Jiménez et al [51] and Cruz-Cobo et al [50] both incorporated application satisfaction into their evaluation frameworks. Multiple studies indicate that DHTs, through their accessibility, convenience, and interactive features, significantly reduce participation barriers such as geographical constraints

XSL•FO
RenderX

and time conflicts, laying a solid foundation for improving rehabilitation participation rates and adherence.

Technical Feasibility, Safety, and Use

Against the backdrop of rapid advancements in DHTs, evaluating their feasibility, safety, and impact on health care service use is particularly crucial. Assessments of technical feasibility encompass device operational stability, data collection integrity, and user-friendliness. Wohlfahrt et al [71] demonstrated in their study that smart device step tracking exhibits good feasibility and compliance among patients with postmyocardial infarction. Safety assessments involve adverse event monitoring, data privacy protection, and risk alert mechanisms. Ma et al [40] demonstrated in a long-term follow-up study that digital interventions did not increase the risk of major adverse cardiovascular events. Health care use metrics include readmission rates and emergency department visit rates. Widmer et al [55] observed a downward trend in readmissions and emergency visits within the digital intervention group, though this did not reach statistical significance. Several studies have mentioned the need for cost-effectiveness analysis of digital rehabilitation, and preliminary explorations suggest that it may have potential economic advantages, but more empirical evidence is needed, though further evidence accumulation remains necessary.

Discussion

Principal Findings

In this study, we used a scoping review methodology to systematically evaluate the current application of DHTs in CR for patients with CHD. The research revealed its core characteristics, including diverse forms of technological application, multidimensional intervention objectives, and multilevel assessment indicators. It integrated 3 categories of technological application forms, 4 categories of intervention objectives, and 5 types of outcome assessment indicators. Findings indicate that DHTs have evolved from supplementary aids to systematic solutions, effectively overcoming the temporal and spatial constraints of traditional rehabilitation. This advance has significantly improved patient engagement in CR and treatment adherence in patients with CHD.

Diversity and Integration of DHTs

Through this scoping review, we found that DHTs show obvious diversity and integration in the form of technology app, which can be divided into 3 main types: digital health tools, real-time remote support, and asynchronous communication. These encompass 10 specific formats including wearable devices, app, website, digital video, social media platform, telephone, group meeting, remote counseling sessions, email, and SMS text messaging. In this scoping review, we found that apps are the main intervention tool for digital CR, and wearable devices are key for real-time data monitoring; their combined application represents the most prevalent model.

From a technical support perspective, digital health tools are primarily oriented toward patient self-management and health monitoring. This aligns with the findings of van Olmen et al [75], who concluded that digital health tools can effectively empower individuals to engage in self-management and advance the achievement of relevant health goals. Real-time remote support emphasizes direct interpersonal interaction between clinicians and patients, preserving the inherent humanistic care inherent in traditional health care [76]. For instance, studies by Ryan et al [77] integrated empathy and care into remote interactions, revealing no significant difference in perceived humanistic care compared to in-person consultations. Asynchronous communication, leveraging flexible information delivery, provides patients with continuous health reminders and educational support [15].

These 3 complementary levels synergistically construct a multitiered, multidimensional, and comprehensive rehabilitation support system spanning patient self-management to real-time clinician-patient interaction. This provides a viable pathway for developing personalized, multimodal CR models. Notably, some studies further integrate offline face-to-face guidance, forming a blended rehabilitation model combining online and offline approaches [37,38,54,56]. This aligns with findings from Thomas et al [78], confirming that a comprehensive digital technology app significantly enhances the individualized adaptability of CR. Compared to traditional rehabilitation methods, this blended model partially addresses limitations such as relatively monotonous formats and insufficient consideration of individual differences [79].

The Multifaceted Application Objectives of DHTs

In this scoping review, we found that DHTs exhibit multifaceted features in terms of intervention goals, which can be divided into 4 main dimensions: motivation and guidance, foundation of knowledge and skills, monitoring and security, and social and group dynamics. A total of 15 specific objectives have been identified. Health education, data monitoring, and reminders and alerts form the core layer; feedback, goal setting, and personalized interventions constitute the secondary layer; while some studies incorporate distinctive features such as gamification, rewards, and social support.

All interventions integrate multidimensional objectives, with health education appearing most frequently, highlighting the central role of knowledge transfer [80]. Features like gamification and social support effectively address patient issues such as lack of motivation and difficulty sustaining behavior [41,50,57]. This aligns with the WHO Global Digital Health Strategy's advocacy for "patient-centered approaches to achieve sustainable behavioral change" [17]. This multidimensional goal integration enables DHTs to systematically tackle key barriers to participation in traditional CR [78]. Knowledge gaps can be addressed through health education, motivation deficits remedied by gamified incentives, and sustained support ensured via social interaction [81]. It is precisely this composite application model that has evolved DHTs from fragmented tools into systematic rehabilitation support systems. This significantly enhances the comprehensiveness and continuity of rehabilitation interventions, providing a crucial pathway for advancing patients with CHD understanding of CR and enabling precision-targeted interventions [15].

XSL•FO
RenderX

Effect Evaluation

Analysis of multidimensional assessment indicators across 43 included studies demonstrates that DHTs exhibit clear short-term intervention value for CR of patients with CHD. However, evidence for long-term clinical outcomes remains scarce, strongly aligning with the findings of positive short- to medium-term effects and insufficient long-term evidence.

Regarding clinical physiological indicators, DHTs significantly improve patients' cardiopulmonary function and physical fitness levels, with statistically significant improvements in core metrics like peak oxygen uptake and 6-minute walk distance observed in intervention groups. They also positively influence blood pressure and lipid control, validating the effectiveness of real-time monitoring and personalized guidance in short-term physiological optimization [39,42,44,49,58,66,70,71]. At the health behavior level, digital technologies significantly enhance rehabilitation adherence in areas like exercise, medication, and diet through mechanisms such as scheduled reminders and real-time feedback. They also correct unhealthy lifestyle habits like prolonged sitting, aligning with findings where health behaviors serve as core assessment dimensions [41,48,51]. Regarding patient-reported outcomes and rehabilitation service use, digital interventions effectively improve patients' quality of life, disease awareness, and self-efficacy. They also significantly overcome temporal and spatial constraints, increasing rehabilitation participation and completion rates while alleviating barriers to traditional rehabilitation, consistent with outcome-related findings [48,53,54,72].

However, existing studies have not reached a unified conclusion regarding the assessment of long-term clinical outcomes. While some studies observed a downward trend in readmission rates, they failed to demonstrate statistical significance in reducing major adverse cardiovascular events or long-term mortality [40,47,55]. This is closely related to the limited sample sizes and short follow-up periods in most studies, as well as the significant heterogeneity in intervention designs and the lack of systematic long-term rehabilitation management systems. Research on the long-term cost-effectiveness and sustainability of these interventions is also scarce, necessitating further exploration [82].

Advantages and Challenges of DHT

The most significant advantage of DHTs lies in their ability to effectively overcome geographical constraints and economic barriers, substantially enhancing the accessibility of CR and patient participation rates [21]. In this study, we found that by providing easily accessible, user-friendly, and highly interactive digital technologies, it is possible to significantly reduce structural barriers commonly encountered in traditional rehabilitation models, such as transport difficulties, time conflicts, and uneven resource distribution [4]. Research by Varnfield et al [48] confirmed that smartphone-based home rehabilitation significantly outperformed traditional rehabilitation in terms of uptake, adherence, and completion rates among patients with postmyocardial infarction. Ramachandran et al [72] further validated the superiority of home-based remote rehabilitation in enhancing rehabilitation use rates. This finding provides a pathway to high-quality rehabilitation services for remote areas with scarce medical resources and for patients with CHD with limited mobility, positioning DHTs as a crucial strategic tool for bridging geographical disparities in health care resources and advancing equity in cardiovascular health services. However, the widespread adoption of digital technologies also presents a new challenge: the "digital divide" [15]. Older adults, low-income groups, and patients with lower educational attainment may encounter significant difficulties in operating digital devices, using apps, or accessing information [17,83,84].

Therefore, advancing DHT must prioritize equity and inclusivity as core principles [17]. Simplified interfaces and voice-assisted features tailored for older people and low-skilled users should be developed, alongside personalized, face-to-face training in digital technology use [85]. Exploring device subsidies or digital equipment loan schemes for vulnerable groups is essential to overcome digital barriers and encourage active participation in digital CR [83]. Furthermore, key challenges for scaling DHTs include technical feasibility, data security and privacy protection, sustainable cost-effectiveness, and seamless integration of digital interventions into existing clinical workflows [15]. Currently, while some studies have begun examining implementation-level indicators for DHTs, such as the feasibility of smart device monitoring and remote rehabilitation, overall evidence remains limited. Greater practical research and systematic evaluation are urgently needed to advance the standardized application and long-term development of digital CR models [71,86].

Limitations

Although we systematically reviewed the current application of DHTs in CR for patients with CHD through a scoping review methodology, several limitations remain. First, regarding literature sources, we only included peer-reviewed empirical research published in English. While this approach helps ensure the quality of included studies, it may overlook important literature published in other languages and relevant gray literature, thereby affecting the comprehensiveness of the study conclusions. Second, significant methodological heterogeneity among the included studies limited the integration and comparability of results. We varied significantly in intervention design, technology combinations, intervention duration, use frequency, participant characteristics, selected outcome measures, and follow-up periods. Furthermore, most studies featured small sample sizes and short follow-up durations, lacking comprehensive assessments of long-term clinical outcomes, cost-effectiveness, intervention sustainability, and impacts on health equity. This limits a thorough evaluation of the long-term value of DHTs.

Conclusions

In this study, we used a scoping review methodology to systematically examine the current application and practical value of DHTs in CR for patients with CHD. Findings confirm that DHTs effectively improve patients' short-term physiological function and optimize health behaviors. Simultaneously, they overcome limitations in traditional CR regarding spatial-temporal constraints and health care resource allocation, significantly enhancing patient engagement and adherence to rehabilitation programs. In clinical practice, health care

XSL•FO
RenderX

providers can integrate multiple DHTs to develop personalized rehabilitation plans tailored to individual characteristics such as patient age, digital literacy, and disease severity, thereby enhancing the precision and adaptability of CR. Future research should prioritize large-scale, multicenter, long-term follow-up randomized controlled trials to thoroughly investigate the impact of DHTs on long-term clinical outcomes for patients with CHD and explore potential mechanisms of action, such as long-term mortality and major adverse cardiovascular events. This will provide more robust evidence-based support for validating their long-term efficacy and advancing standardized clinical implementation.

Acknowledgments

The authors wish to express their gratitude to the librarians who assisted in formulating the retrieval strategy, as well as to all scholars in relevant fields for their accumulated prior research. The authors did not use any artificial intelligence generation tool in the study.

Funding

This work was supported by grants from 2025 Basic Scientific Research Projects of Colleges and Universities in Humanities and Social Sciences Category of Liaoning Provincial Department of Education (LJ112510162012); 2025 Nursing Discipline Research Projects of the Chinese Medical Association Publishing House (CMAPH-NRC2025011). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Authors' Contributions

Conceptualization: XZ (lead), LL (supporting)
Data curation: XZ (lead), ZL (supporting), LL (supporting), JW (supporting)
Formal analysis: XZ (lead), JW (supporting), ZL (supporting)
Funding acquisition: LL (lead), JW (equal)
Investigation: XZ (lead), ZL (supporting), LL (supporting), JW (supporting)
Methodology: XZ (lead), LL (supporting)
Project administration: MZ (lead), YW (supporting), HL (supporting)
Resources: MZ (lead), YW (equal), HL (equal)
Supervision: LL (lead), HL (supporting), YW (supporting), MZ (supporting)
Writing—original draft: XZ (lead), JW (supporting), YT (supporting)
Writing—review and editing: XZ (lead), JW (supporting), YW (supporting), YT (supporting), LL (supporting), MZ (supporting), HL (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies for each database.
[PDF File, 126 KB - jmir_v28i1e85917_app1.pdf]

Checklist 1
PRISMA-ScR checklist.
[PDF File, 119 KB - jmir_v28i1e85917_app2.pdf]

References

1. The top 10 causes of death. World Health Organization. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death [accessed 2025-08-20]
2. Shakya S, Shrestha A, Robinson S, et al. Global comparison of the economic costs of coronary heart disease: a systematic review and meta-analysis. BMJ Open 2025 Jan 21;15(1):e084917. [doi: 10.1136/bmjopen-2024-084917] [Medline: 39842921]
3. Witt BJ, Jacobsen SJ, Weston SA, et al. Cardiac rehabilitation after myocardial infarction in the community. J Am Coll Cardiol 2004 Sep 1;44(5):988-996. [doi: 10.1016/j.jacc.2004.05.062] [Medline: 15337208]
4. Resurrección DM, Moreno-Peral P, Gómez-Herranz M, et al. Factors associated with non-participation in and dropout from cardiac rehabilitation programmes: a systematic review of prospective cohort studies. Eur J Cardiovasc Nurs 2019 Jan;18(1):38-47. [doi: 10.1177/1474515118783157] [Medline: 29909641]

XSL•FO
RenderX

5. Kotseva K, Wood D, De Backer G, De Bacquer D, EUROASPIRE III Study Group. Use and effects of cardiac rehabilitation in patients with coronary heart disease: results from the EUROASPIRE III survey. Eur J Prev Cardiol 2013 Oct;20(5):817-826. [doi: 10.1177/2047487312449591] [Medline: 22718794]
6. Heidenreich PA, Bozkurt B, Aguilar D, et al. 2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. Circulation 2022 May 3;145(18):e895-e1032. [doi: 10.1161/CIR.0000000000001063] [Medline: 35363499]
7. Authors/Task Force Members, McDonagh TA, Metra M, et al. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: developed by the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). With the special contribution of the Heart Failure Association (HFA) of the ESC. Eur J Heart Fail 2022 Jan;24(1):4-131. [doi: 10.1002/ejhf.2333] [Medline: 35083827]
8. Gabrys L, Schmidt C. Prescription and utilization of sports therapy programs following cardiac rehabilitation 2006-2013. Rehabilitation (Stuttg) 2020 Feb;59(1):42-47. [doi: 10.1055/a-0869-9810] [Medline: 31096291]
9. Ades PA, Keteyian SJ, Wright JS, et al. Increasing cardiac rehabilitation participation from 20% to 70%: a road map from the Million Hearts cardiac rehabilitation collaborative. Mayo Clin Proc 2017 Feb;92(2):234-242. [doi: 10.1016/j.mayocp.2016.10.014] [Medline: 27855953]
10. Sumner J, Grace SL, Doherty P. Predictors of cardiac rehabilitation utilization in England: results from the national audit. J Am Heart Assoc 2016 Oct 21;5(10):e003903. [doi: 10.1161/JAHA.116.003903] [Medline: 27792657]
11. Poh R, Ng HN, Loo G, et al. Cardiac rehabilitation after percutaneous coronary intervention in a multiethnic Asian country: enrollment and barriers. Arch Phys Med Rehabil 2015 Sep;96(9):1733-1738. [doi: 10.1016/j.apmr.2015.05.020] [Medline: 26073758]
12. Im HW, Baek S, Jee S, Ahn JM, Park MW, Kim WS. Barriers to outpatient hospital-based cardiac rehabilitation in Korean patients with acute coronary syndrome. Ann Rehabil Med 2018 Feb;42(1):154-165. [doi: 10.5535/arm.2018.42.1.154] [Medline: 29560336]
13. Kanazawa N, Yamada S, Fushimi K. Trends in the use of cardiac rehabilitation in Japan between 2010 and 2017—an epidemiological survey. Circ Rep 2021 Oct 8;3(10):569-577. [doi: 10.1253/circrep.CR-21-0018] [Medline: 34703934]
14. Shen T, Ren C, Shang Z, et al. Current situation of cardiac rehabilitation and exercise evaluation under the background of the construction of chest pain center. Chin J Intervent Cardiol 2023;31(7):509-513. [doi: 10.3969/j.issn.1004-8812.2023.07.003]
15. Yang Z, Jin D, Huang H, Zheng X, Liu S, Wang A. Nudging health behavior change among home-based cardiac rehabilitation patients: a scoping review. J Multidiscip Healthc 2025;18:1639-1653. [doi: 10.2147/JMDH.S515523] [Medline: 40125302]
16. Guasti L, Dilaveris P, Mamas MA, et al. Digital health in older adults for the prevention and management of cardiovascular diseases and frailty. A clinical consensus statement from the ESC Council for Cardiology Practice/Taskforce on Geriatric Cardiology, the ESC Digital Health Committee and the ESC Working Group on e-Cardiology. ESC Heart Fail 2022 Oct;9(5):2808-2822. [doi: 10.1002/ehf2.14022] [Medline: 35818770]
17. Global strategy on digital health 2020-2027. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240116870 [accessed 2025-08-20]
18. Wang W, Jiang Y. The evolving mHealth-based cardiac rehabilitation. Lancet Digit Health 2019 Nov;1(7):e326-e327. [doi: 10.1016/S2589-7500(19)30155-4] [Medline: 33323205]
19. Dorje T, Zhao G, Tso K, et al. Smartphone and social media-based cardiac rehabilitation and secondary prevention in China (SMART-CR/SP): a parallel-group, single-blind, randomised controlled trial. Lancet Digit Health 2019 Nov;1(7):e363-e374. [doi: 10.1016/S2589-7500(19)30151-7] [Medline: 33323210]
20. Martani A, Geneviève LD, Poppe C, Casonato C, Wangmo T. Digital pills: a scoping review of the empirical literature and analysis of the ethical aspects. BMC Med Ethics 2020 Jan 8;21(1):3. [doi: 10.1186/s12910-019-0443-1] [Medline: 31914995]
21. Piotrowicz E, Piotrowicz R. Cardiac telerehabilitation: current situation and future challenges. Eur J Prev Cardiol 2013 Jun;20(2 Suppl):12-16. [doi: 10.1177/2047487313487483c] [Medline: 23702985]
22. Taylor RS, Dalal HM, McDonagh STJ. The role of cardiac rehabilitation in improving cardiovascular outcomes. Nat Rev Cardiol 2022 Mar;19(3):180-194. [doi: 10.1038/s41569-021-00611-7] [Medline: 34531576]
23. Ometov A, Shubina V, Klus L, et al. A survey on wearable technology: history, state-of-the-art and current challenges. Comput Networks 2021 Jul;193:108074. [doi: 10.1016/j.comnet.2021.108074]
24. Colquhoun HL, Levac D, O'Brien KK, et al. Scoping reviews: time for clarity in definition, methods, and reporting. J Clin Epidemiol 2014 Dec;67(12):1291-1294. [doi: 10.1016/j.jclinepi.2014.03.013] [Medline: 25034198]
25. Tricco AC, Lillie E, Zarin W, et al. A scoping review on the conduct and reporting of scoping reviews. BMC Med Res Methodol 2016 Feb 9;16(1):15. [doi: 10.1186/s12874-016-0116-4] [Medline: 26857112]
26. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
27. Pollock D, Peters MDJ, Khalil H, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. JBI Evidence Synthesis 2023;21(3):520-532. [doi: 10.11124/JBIES-22-00123]
28. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evidence Synthesis 2020;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167]

XSL•FO
RenderX

29. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018 Nov 19;18(1):143. [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
30. JBI Manual for Evidence Synthesis. JBI. URL: https://jbi-global-wiki.refined.site/space/MANUAL/355862707/10.2.4+Inclusion+criteria [accessed 2026-02-27]
31. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
32. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
33. Hong PC, Chen KJ, Chang YC, Cheng SM, Chiang HH. Effectiveness of theory-based health information technology interventions on coronary artery disease self-management behavior: a clinical randomized waitlist-controlled trial. J Nurs Scholarsh 2021 Jul;53(4):418-427. [doi: 10.1111/jnu.12661] [Medline: 33844425]
34. Chan NPT, Lai AYK, Choy HK, et al. Feasibility and potential effectiveness of a smartphone zero-time exercise intervention for promoting physical activity and fitness in patients with coronary heart disease: a pilot randomized controlled trial. Front Public Health 2022;10:865712. [doi: 10.3389/fpubh.2022.865712] [Medline: 35910893]
35. Duan YP, Liang W, Guo L, Wienert J, Si GY, Lippke S. Evaluation of a web-based intervention for multiple health behavior changes in patients with coronary heart disease in home-based rehabilitation: pilot randomized controlled trial. J Med Internet Res 2018 Nov 19;20(11):e12052. [doi: 10.2196/12052] [Medline: 30455167]
36. Xu D, Xu D, Wei L, Bao Z, Liao S, Zhang X. The effectiveness of remote exercise rehabilitation based on the “SCeiP” model in homebound patients with coronary heart disease: randomized controlled trial. J Med Internet Res 2024 Nov 5;26:e56552. [doi: 10.2196/56552] [Medline: 39499548]
37. Li Y, Gong Y, Zheng B, et al. Effects on adherence to a mobile app-based self-management digital therapeutics among patients with coronary heart disease: pilot randomized controlled trial. JMIR Mhealth Uhealth 2022 Feb 15;10(2):e32251. [doi: 10.2196/32251] [Medline: 34906924]
38. Su JJ, Yu DSF. Effects of a nurse-led eHealth cardiac rehabilitation programme on health outcomes of patients with coronary heart disease: a randomised controlled trial. Int J Nurs Stud 2021 Oct;122:104040. [doi: 10.1016/j.ijnurstu.2021.104040] [Medline: 34333211]
39. Zheng Y, Guo J, Tian Y, Qin S, Liu X. Effect of home-based cardiac telerehabilitation in patients after percutaneous coronary intervention: a randomized controlled trial. Comput Inform Nurs 2024 Dec 1;42(12):898-904. [doi: 10.1097/CIN.0000000000001167] [Medline: 39087895]
40. Ma J, Ge C, Shi Y, et al. Chinese home-based cardiac rehabilitation model delivered by smartphone interaction improves clinical outcomes in patients with coronary heart disease. Front Cardiovasc Med 2021;8:731557. [doi: 10.3389/fcvm.2021.731557] [Medline: 34676252]
41. Xu L, Tong Q, Zhang X, et al. Smartphone-based gamification intervention to increase physical activity participation among patients with coronary heart disease: a randomized controlled trial. J Telemed Telecare 2024 Oct;30(9):1425-1436. [doi: 10.1177/1357633X221150943] [Medline: 36794484]
42. Li X, Zhao L, Xu T, et al. Cardiac telerehabilitation under 5G internet of things monitoring: a randomized pilot study. Sci Rep 2023;13(1):18886. [doi: 10.1038/s41598-023-46175-z]
43. Ni Z, Wu B, Yang Q, Yan LL, Liu C, Shaw RJ. An mHealth intervention to improve medication adherence and health outcomes among patients with coronary heart disease: randomized controlled trial. J Med Internet Res 2022 Mar 9;24(3):e27202. [doi: 10.2196/27202] [Medline: 35262490]
44. Liu Y, Huang X, Dai Z, et al. Effects of an eHealth cardiac exercise rehabilitation platform for patients after percutaneous coronary intervention based on the persuasive systems design model: randomized controlled trial. J Med Internet Res 2026 Jan 14;28:e71450. [doi: 10.2196/71450] [Medline: 41534075]
45. Zhang S, Wang Y, Wu J, Ma C, Meng X. Effectiveness of smartwatch device on adherence to home-based cardiac rehabilitation in patients with coronary heart disease: randomized controlled trial. JMIR Mhealth Uhealth 2025 Sep 18;13:e70848. [doi: 10.2196/70848] [Medline: 40966536]
46. Li PWC, Yu DSF, Chan BS, Wong CW, Wong CWY, Chung HSW. i-CARE: a smartphone-based intervention to enhance cardiac rehabilitation in coronary artery disease—a randomized controlled trial. Eur J Prev Cardiol 2025 Aug 11:zwaf509. [doi: 10.1093/eurjpc/zwaf509] [Medline: 40796126]
47. Krzowski B, Boszko M, Peller M, et al. Mobile app and digital system for patients after myocardial infarction (afterAMI): results from a randomized trial. J Clin Med 2023 Apr 15;12(8):2886. [doi: 10.3390/jcm12082886] [Medline: 37109223]
48. Varnfield M, Karunanithi M, Lee CK, et al. Smartphone-based home care model improved use of cardiac rehabilitation in postmyocardial infarction patients: results from a randomised controlled trial. Heart 2014 Nov;100(22):1770-1779. [doi: 10.1136/heartjnl-2014-305783] [Medline: 24973083]
49. Avila A, Claes J, Goetschalckx K, et al. Home-based rehabilitation with telemonitoring guidance for patients with coronary artery disease (short-term results of the TRiCH study): randomized controlled trial. J Med Internet Res 2018 Jun 22;20(6):e225. [doi: 10.2196/jmir.9943] [Medline: 29934286]

XSL•FO
RenderX

50. Cruz-Cobo C, Bernal-Jiménez M, Calle G, et al. Efficacy of a mobile health app (eMOTIVA) regarding compliance with cardiac rehabilitation guidelines in patients with coronary artery disease: randomized controlled clinical trial. JMIR Mhealth Uhealth 2024 Jul 25;12:e55421. [doi: 10.2196/55421] [Medline: 39052330]
51. Bernal-Jiménez M, Calle G, Gutiérrez Barrios A, et al. Effectiveness of an interactive mHealth app (EVITE) in improving lifestyle after a coronary event: randomized controlled trial. JMIR Mhealth Uhealth 2024 Apr 22;12:e48756. [doi: 10.2196/48756] [Medline: 38648103]
52. Bae JW, Woo SI, Lee J, et al. mHealth interventions for lifestyle and risk factor modification in coronary heart disease: randomized controlled trial. JMIR Mhealth Uhealth 2021 Sep 24;9(9):e29928. [doi: 10.2196/29928] [Medline: 34559058]
53. Dodson JA, Adhikari S, Schoenthaler A, et al. Rehabilitation at home using mobile health for older adults hospitalized for ischemic heart disease: the RESILIENT randomized clinical trial. JAMA Netw Open 2025 Jan 2;8(1):e2453499. [doi: 10.1001/jamanetworkopen.2024.53499] [Medline: 39775808]
54. Hisam A, Haq ZU, Aziz S, Doherty P, Pell J. Effectiveness of mobile health augmented cardiac rehabilitation (MCard) on health-related quality of life among post-acute coronary syndrome patients: a randomized controlled trial. Pak J Med Sci 2022;38(3Part-I):716-723. [doi: 10.12669/pjms.38.3.4724] [Medline: 35480536]
55. Widmer RJ, Allison TG, Lennon R, Lopez-Jimenez F, Lerman LO, Lerman A. Digital health intervention during cardiac rehabilitation: a randomized controlled trial. Am Heart J 2017 Jun;188:65-72. [doi: 10.1016/j.ahj.2017.02.016] [Medline: 28577682]
56. Yudi MB, Clark DJ, Tsang D, et al. SMARTphone-based, early cardiac REHABilitation in patients with acute coronary syndromes: a randomized controlled trial. Coron Artery Dis 2021 Aug 1;32(5):432-440. [doi: 10.1097/MCA.0000000000000938] [Medline: 32868661]
57. Gallagher R, Chow CK, Parker H, et al. The effect of a game-based mobile app “MyHeartMate” to promote lifestyle change in coronary disease patients: a randomized controlled trial. Eur Heart J Digit Health 2023 Jan;4(1):33-42. [doi: 10.1093/ehjdh/ztac069] [Medline: 36743873]
58. Jo HS, Kim HM, Go CH, Yu HY, Park HK, Han JY. Effectiveness of home-based cardiac rehabilitation with optimized exercise prescriptions using a mobile healthcare app in patients with acute myocardial infarction: a randomized controlled trial. Life 2024;14(9):1122. [doi: 10.3390/life14091122]
59. Fallah Z, Feizi A, Sadeghi M, Hadavi MM, Shahnazi H. The effect of home-based virtual exercise rehabilitation on myocardial infarction patients using the health action process approach: a randomized controlled trial. BMC Sports Sci Med Rehabil 2025;18(1):35. [doi: 10.1186/s13102-025-01492-w]
60. Bruggmann C, Adjedj J, Sardy S, Muller O, Voirol P, Sadeghipour F. Effects of the interactive web-based video “Mon Coeur, Mon BASIC” on drug adherence of patients with myocardial infarction: randomized controlled trial. J Med Internet Res 2021 Aug 30;23(8):e21938. [doi: 10.2196/21938] [Medline: 34459744]
61. Patterson K, Davey R, Keegan R, et al. Testing the effect of a smartphone app on hospital admissions and sedentary behavior in cardiac rehabilitation participants: ToDo-CR randomized controlled trial. JMIR Mhealth Uhealth 2023 Oct 3;11:e48229. [doi: 10.2196/48229] [Medline: 37788043]
62. Batalik L, Dosbaba F, Hartman M, Batalikova K, Spinar J. Benefits and effectiveness of using a wrist heart rate monitor as a telerehabilitation device in cardiac patients. Medicine (Baltimore) 2020;99(11):e19556. [doi: 10.1097/MD.0000000000019556]
63. Dalli Peydró E, Sanz Sevilla N, Tuzón Segarra MT, Miró Palau V, Sánchez Torrijos J, Cosín Sales J. A randomized controlled clinical trial of cardiac telerehabilitation with a prolonged mobile care monitoring strategy after an acute coronary syndrome. Clin Cardiol 2022 Jan;45(1):31-41. [doi: 10.1002/clc.23757] [Medline: 34952989]
64. Bravo-Escobar R, González-Represas A, Gómez-González AM, et al. Effectiveness and safety of a home-based cardiac rehabilitation programme of mixed surveillance in patients with ischemic heart disease at moderate cardiovascular risk: a randomised, controlled clinical trial. BMC Cardiovasc Disord 2017 Feb 20;17(1):66. [doi: 10.1186/s12872-017-0499-0] [Medline: 28219338]
65. Johnston N, Bodegard J, Jerström S, et al. Effects of interactive patient smartphone support app on drug adherence and lifestyle changes in myocardial infarction patients: a randomized study. Am Heart J 2016 Aug;178:85-94. [doi: 10.1016/j.ahj.2016.05.005] [Medline: 27502855]
66. Kumar R M, T SK, Vinod Kumar B, S S, Natarajan V. Effects of an e-Media-supported, exercise-based Phase II cardiac rehabilitation in coronary artery bypass grafting surgery patients: a randomized controlled trial. Cureus 2024 Aug;16(8):e67557. [doi: 10.7759/cureus.67557] [Medline: 39314565]
67. Bretschneider MP, Mayer-Berger W, Weine J, Roth L, Schwarz PEH, Petermann F. Results of a digital multimodal motivational and educational program as follow-up care for former cardiac rehabilitation patients: randomized controlled trial. JMIR Cardio 2024 Dec 11;8:e57960. [doi: 10.2196/57960] [Medline: 39661085]
68. Herring LY, Dallosso H, Schreder S, et al. Physical Activity after Cardiac EventS (PACES): a group education programme with subsequent text message support designed to increase physical activity in individuals with diagnosed coronary heart disease: a randomised controlled trial. Open Heart 2021 Feb;8(1):e001351. [doi: 10.1136/openhrt-2020-001351] [Medline: 33637567]

XSL•FO
RenderX

69. Khikmatova M, Togayeva B, Ganieva N, et al. Digital health interventions for post-myocardial infarction rehabilitation: a randomized trial on wearable technology adherence and cardiac outcomes. Rev Latinoam Hipertens 2025;20(7):504-510. [doi: 10.5281/zenodo.16122962]
70. Nishio R, Dohi T, Yokoyama M, et al. Wearable devices in remote cardiac rehabilitation with and without weekly online coaching for patients with coronary artery disease: randomized controlled trial. JMIR Mhealth Uhealth 2025 May 12;13:e63797. [doi: 10.2196/63797] [Medline: 40395066]
71. Wohlfahrt P, Jenča D, Melenovský V, et al. Remote, smart device-based cardiac rehabilitation after myocardial infarction: a pilot, randomized cross-over SmartRehab study. Mayo Clin Proc Digit Health 2024 Sep;2(3):352-360. [doi: 10.1016/j.mcpdig.2024.06.001] [Medline: 40206122]
72. Ramachandran HJ, Yeo TJ, Seah CWA, et al. Feasibility and effectiveness of an integrative cardiac rehabilitation employing smartphone technology (I-CREST): a pilot randomized controlled trial. Eur J Cardiovasc Nurs 2025 Sep 5;24(6):959-970. [doi: 10.1093/eurjcn/zvaf053]
73. Waranski M, Garbsch R, Kotewitsch M, Teschler M, Schmitz B, Mooren FC. A behavioral change-based mobile intervention for promoting regular physical activity in medical rehabilitation maintenance of patients with coronary artery disease: controlled trial. J Med Internet Res 2024 Oct 8;26:e56480. [doi: 10.2196/56480] [Medline: 39378432]
74. Widmer RJ, Allison TG, Lerman LO, Lerman A. Digital health intervention as an adjunct to cardiac rehabilitation reduces cardiovascular risk factors and rehospitalizations. J Cardiovasc Trans Res 2015 Jul;8(5):283-292. [doi: 10.1007/s12265-015-9629-1]
75. van Olmen J. The promise of digital self-management: a reflection about the effects of patient-targeted e-Health tools on self-management and wellbeing. Int J Environ Res Public Health 2022 Jan 26;19(3):1360. [doi: 10.3390/ijerph19031360] [Medline: 35162383]
76. Elliott T, Tong I, Sheridan A, Lown BA. Beyond convenience: patients' perceptions of physician interactional skills and compassion via telemedicine. Mayo Clin Proc Innov Qual Outcomes 2020 Jun;4(3):305-314. [doi: 10.1016/j.mayocpiqo.2020.04.009]
77. Ryan BL, Brown JB, Freeman TR, DaSilva M, Stewart M, Terry AL. Safeguarding compassion in virtual family physician care. J Am Board Fam Med 2025 Nov 24;38(4):661-674. [doi: 10.3122/jabfm.2024.240460R1] [Medline: 41052935]
78. Thomas RJ, Beatty AL, Beckie TM, et al. Home-based cardiac rehabilitation: a scientific statement from the American Association of Cardiovascular and Pulmonary Rehabilitation, the American Heart Association, and the American College of Cardiology. J Am Coll Cardiol 2019 Jul 9;74(1):133-153. [doi: 10.1016/j.jacc.2019.03.008] [Medline: 31097258]
79. Liu Y, Wang X, Li J, Huang X, Yu J, Chen M. Effects of personalized structured telemedicine-based exercise cardiac rehabilitation on health outcomes in patients with coronary heart disease: a systematic review and meta-analysis. Eur J Cardiovasc Nurs . [doi: 10.1093/eurjcn/zvaf195]
80. Cowell AJ. The relationship between education and health behavior: some empirical evidence. Health Econ 2006 Feb;15(2):125-146. [doi: 10.1002/hec.1019] [Medline: 16167321]
81. Patel MS, Small DS, Harrison JD, et al. Effect of behaviorally designed gamification with social incentives on lifestyle modification among adults with uncontrolled diabetes: a randomized clinical trial. JAMA Netw Open 2021 May 3;4(5):e2110255. [doi: 10.1001/jamanetworkopen.2021.10255] [Medline: 34028550]
82. Savović J, Jones HE, Altman DG, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med 2012 Sep 18;157(6):429-438. [doi: 10.7326/0003-4819-157-6-201209180-00537] [Medline: 22945832]
83. Choi NG, Dinitto DM. The digital divide among low-income homebound older adults: Internet use patterns, eHealth literacy, and attitudes toward computer/Internet use. J Med Internet Res 2013 May 2;15(5):e93. [doi: 10.2196/jmir.2645] [Medline: 23639979]
84. Falter M, Scherrenberg M, Dendale P. Digital health in cardiac rehabilitation and secondary prevention: a search for the ideal tool. Sensors (Basel) 2020 Dec 22;21(1):12. [doi: 10.3390/s21010012] [Medline: 33374985]
85. Yao R, Zhang W, Evans R, Cao G, Rui T, Shen L. Inequities in health care services caused by the adoption of digital health technologies: scoping review. J Med Internet Res 2022 Mar 21;24(3):e34144. [doi: 10.2196/34144] [Medline: 35311682]
86. Bäck M, Leosdottir M, Ekström M, et al. Feasibility, safety and patient perceptions of exercise-based cardiac telerehabilitation in a multicentre real-world setting after myocardial infarction—the remote exercise SWEDEHEART study. Eur Heart J Digit Health 2025 May;6(3):508-518. [doi: 10.1093/ehjdh/ztaf014] [Medline: 40395424]

Abbreviations

CHD: coronary heart disease
CR: cardiac rehabilitation
CVD: cardiovascular disease
DHT: digital health technology
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews

XSL•FO
RenderX

PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
WHO: World Health Organization

Edited by S Brini; submitted 15.Oct.2025; peer-reviewed by K Kitagaki, Z Yang; revised version received 12.Mar.2026; accepted 12.Mar.2026; published 29.Apr.2026.

Please cite as:
Zhu X, Liu L, Wang Y, Li H, Zang M, Wang J, Tian Y, Li Z
Application of Digital Health Technologies in Cardiac Rehabilitation for Patients With Coronary Heart Disease: Scoping Review
J Med Internet Res 2026;28:e85917
URL: https://www.jmir.org/2026/1/e85917
doi:10.2196/85917

© Xinyu Zhu, Lei Liu, Yingjie Wang, Hongyuan Li, Min Zang, Jiayu Wang, Yu Tian, Zihan Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Characteristics of Patients Accessing Outpatient Oncology Services Virtually and Predictors of Subsequent Unplanned Emergency Department Presentations in 78,323 Adults in Australia: Retrospective Cohort Study

Ramya Walsan[1], PhD; Reema Harrison[1], PhD; Johanna Westbrook[1], PhD; Ashfaq Chauhan[1], PhD; Michelle Moscova[2], PhD; Anita Vandyke[3], MD; Elizabeth Manias[4], PhD; Tracey Webster[5], MSc; Natalie Taylor[6], PhD; Prince Peprah[1], PhD; Rebecca Mitchell[7], PhD

[1]Centre for Health Systems and Safety Research, Australian Institute of Health Innovation, Macquarie University, Level 6, 75 Talavera Road, North Ryde, New South Wales, Australia

[2]Health ANSWERS, Shoalhaven, NSW, Australia

[3]Commerical & Capital Advisory, Change & Innovation, NSW Health Infrastructure, Sydney, NSW, Australia

[4]Monash Nursing and Midwifery, Faculty of Medicine, Nursing and Health Sciences, Monash University, Clayton, Victoria, Australia

[5]Transformation, Quality, & Safety Division, Clinical Leadership Effectiveness & Outcomes, Northern Health, Epping, Victoria, Australia

[6]Faculty of Medicine and Health, School of Population Health, UNSW Sydney, Sydney, New South Wales, Australia

[7]Centre for Healthcare Resilience and Implementation Science, Australian Institute of Health Innovation, Macquarie University, North Ryde, New South Wales, Australia

Corresponding Author:
Ramya Walsan, PhD
Centre for Health Systems and Safety Research, Australian Institute of Health Innovation, Macquarie University, Level 6, 75 Talavera Road, North Ryde, New South Wales, Australia

Abstract

Background: Virtual care has become increasingly integrated into oncology services since the COVID-19 pandemic, yet little is known about which patients use it most and how its use is associated with unplanned emergency care among people living with cancer.

Objective: The study aims to identify sociodemographic and clinical predictors of virtual care use among patients accessing outpatient oncology services and quantify their association with unplanned emergency department (ED) visits.

Methods: A retrospective cohort study was conducted using linked administrative health data for 78,323 adults with cancer who accessed outpatient oncology services in Victoria and Queensland, Australia, between January 2018 and December 2020, with a 1-year follow-up. Virtual care use and unplanned ED visits were categorized as none, low (1-3), or high (≥4), and analyzed using modified Poisson models with robust variance estimation, adjusted for sociodemographic and clinical factors.

Results: Out of 78,323 patients, 37,706 (48.1%) did not use virtual care (only in-person), 24,196 (30.9%) had low use, and 16,421 (20.9%) were high users. Higher virtual care use was associated with rural (vs urban) residence (relative risk [RR] 1.23, 95% CI 1.19 - 1.28), mental health disorders (vs none; RR 1.28, 95% CI 1.24 - 1.33), Charlson comorbidities (vs none; RR 1.12, 95% CI 1.09 - 1.28), and receiving care during the COVID-19 pandemic (vs nonpandemic period; RR 3.03, 95% CI 2.92 - 3.15). In contrast, older age (≥75 y vs 18 - 44 y; RR 0.78, 95% CI 0.74 - 0.83) and being born overseas (vs Australia; RR 0.83, 95% CI 0.80 - 0.86) were associated with lower virtual care use. High virtual care use (vs none) was associated with an increased risk of ≥4 unplanned ED visits (RR 2.64, 95% CI 2.52 - 2.79).

Conclusions: The use of virtual outpatient oncology services varied based on patients' demographic and clinical characteristics. Higher virtual outpatient use was associated with increased unplanned ED presentations. Further research using causal analytic approaches is needed to clarify the relationship between virtual care and unplanned acute care use.

(J Med Internet Res 2026;28:e87694) doi:10.2196/87694

KEYWORDS

cancer; virtual care; telehealth; unplanned health service use; emergency department visits

XSL•FO
RenderX

Introduction

Cancer is a complex, often chronic condition requiring multidisciplinary, coordinated care from diagnosis through survivorship [1]. While cancer care was traditionally delivered in inpatient settings, advancements in treatment safety, increasing cost pressures, and a growing emphasis on patient-centered care have driven a shift toward outpatient and ambulatory models [2,3]. Virtual care, defined as the use of telephone, video, and other technological tools to deliver health care [4], was initially adopted in cancer services to improve access for rural and remote populations [4,5]. The COVID-19 pandemic, however, significantly accelerated its uptake across all health care settings [6]. This rapid integration yielded substantial benefits, including reduced travel burden [7], minimal disruption to work or education [8], and improved access for underserved populations [7].

Virtual care use in oncology has received relatively little attention, with existing studies primarily focusing on quality of life [9] and patient experiences [10]. To support postpandemic integration of virtual care, it is important to understand which patient groups are more or less likely to engage with virtual care, as well as its association with acute health service use, particularly unplanned emergency department (ED) visits. Unplanned ED presentations are a recognized indicator of gaps in outpatient care, often reflecting inadequate care coordination, limited access to timely support, and insufficient patient navigation services [11]. Although virtual care is expected to reduce unplanned service use by improving access and enabling earlier intervention, evidence on the nature of this association remains limited, particularly in Australia and among populations living with cancer [12,13].

A recent Australian study describing patients living with cancer found a higher likelihood of unplanned ED visits among those who used virtual services at least once, compared to those receiving in-person care [14]. However, it is not yet known whether these associations vary by the level of virtual care use, that is, whether higher levels of virtual care use are associated with higher unplanned ED use. This study aims to describe the sociodemographic and clinical characteristics associated with virtual care engagement and to examine the association between levels of virtual care use and unplanned ED visits among patients living with cancer.

Methods

Study Design and Setting

This retrospective cohort study was conducted using linked administrative health data from public nonadmitted patient or outpatient oncology services in Victoria and Queensland, Australia. The study included individuals aged 18 years or older who received nonadmitted patient or outpatient oncology services in Victoria and Queensland, Australia, between January 1, 2018, and December 31, 2020, and who had a confirmed cancer diagnosis in the Victorian or Queensland Cancer registries, respectively. These services deliver specialist cancer care within hospital-based outpatient clinics across the 2 states. Participants were followed for 1 year from their index outpatient encounter or were censored at death.

Ethical Considerations

Ethical approval for the study was obtained from the Victorian Department of Health and the Department of Families, Fairness and Housing Human Research Ethics Committee (HREC/97793/DOH-2023-383794(v3)) and was approved by Queensland Health under Section 282 of the Public Health Act 2005 (Qld) s 282 (PHA 97793). As this study involved a retrospective analysis of routinely collected administrative data, a waiver of informed consent was granted by the approving ethics committees. Data linkage was conducted by the Centre for Victorian Data Linkage and the Queensland Statistical Services Branch, and researchers accessed only deidentified datasets within the Secure Unified Research Environment to ensure participant privacy and confidentiality. No compensation was provided to participants.

Data Sources

Outpatient data from the Victorian Integrated Non-Admitted Health (VINAH) minimum dataset and the Queensland Health Non-Admitted Patient Data Collection (QHNAPDC) were linked to ED presentations, cancer registries, hospital admissions, and mortality datasets. VINAH and QHNAPDC capture all public outpatient services in Victoria and Queensland, including patient demographics, service dates, service types, contact streams, and delivery modes. Cancer-related data were obtained from the Victorian and Queensland Cancer Registries, which are population-based registries receiving notifications on cancer diagnoses, excluding nonmelanoma skin cancers. ED presentation data from the Victorian Emergency Minimum Dataset and Queensland Emergency Data Collection capture patient demographics, presentation details, and visit types. Comorbidity information was sourced from the Victorian Admitted Episodes Dataset and Queensland Hospital Admitted Patient Data Collection, which include admissions to both public and private hospitals classified using International Classification of Diseases, 10th Revision, Australian Modification (ICD-10-AM). Mortality data from the respective registries of births, deaths, and marriages provided the fact and date of death. Data linkage was performed by the Centre for Victorian Data Linkage and Queensland Statistical Services Branch.

Case Inclusion Criteria

The study cohort included adults aged 18 years or older residing in Victoria or Queensland who received outpatient oncology services between January 1, 2018, and December 31, 2020, and had a cancer diagnosis (ICD-10-AM codes: C00–C43, C45–C96, D45, D46, D47.1, D47.3–D47.5) recorded in a cancer registry prior to or within 7 days of their first outpatient oncology encounter. Oncology service use was identified using contact stream ID 110 in the VINAH and Tier 2 clinic codes (oncology, medical oncology consultation, chemotherapy, and radiation oncology—consultation, simulation, and planning) in the QHNAPDC. Virtual care was defined as consultations delivered via telephone or video, based on the mode-of-delivery variable. To reduce selection bias, inclusion criteria were applied uniformly across both jurisdictions.

XSL•FO
RenderX

Clinical Characteristics and Comorbidities

Cancer type was identified using primary site ICD-10-AM codes from cancer registry records, supplemented by inpatient data. The 4 most common cancer types in the dataset, such as breast, digestive (including colorectal), lung, and prostate, were analyzed individually, with all others combined (Table S1 in Multimedia Appendix 1). Comorbidities were determined from admission diagnoses, based on the 17 Charlson comorbidities, with malignant neoplasm excluded to prevent collinearity with cancer diagnosis, and were classified as present or absent. In addition, inpatient records were used to identify mental health disorders (ICD-10-AM: F00–F99) and tobacco use (ICD-10-AM: F17.0–F17.9, P04.2, T65.2, Z58.7, Z71.6, Z72.0, Z81.2, Z86.43). To account for prior outpatient health service use, individuals were categorized based on their preindex outpatient activity during the year preceding the study period (January 1 to December 31, 2017) as "low users," defined as 0 to 3 outpatient encounters, while "high users" had ≥4 encounters [15].

Sociodemographic Characteristics

Patients' residential location in the outpatient data was classified using Statistical Area Level 2 (SA2) for Queensland and using postcode for Victoria, based on the Australian Statistical Geography Standard Remoteness classification. The 5 remoteness categories were collapsed into urban (major cities) and rural (ie, inner regional, outer regional, remote, and very remote areas). Socioeconomic disadvantage was assessed using the Index of Relative Socioeconomic Disadvantage, derived from the patient's Statistical Area Level 2 (SA2)/postcode of residence, and classified into quintiles (Q1=most disadvantaged; Q5=least disadvantaged). Country of birth was categorized as Australian—or overseas—born from inpatient and outpatient records. Patients were classified as receiving treatment during the COVID-19 lockdown period if any portion of their 12-month follow-up period intersected with the defined COVID-19 lockdown period in both states (March 1, 2020-July 31, 2021).

Study Outcomes

Outcomes were levels of virtual outpatient service use and unplanned ED visits during the 12-month follow-up. Based on the number of virtual consultations, patients were categorized into 3 groups: no virtual care use, low (1 - 3 visits), and high (≥4 visits) use. Unplanned ED visits were identified using the visit type code classified as an "emergency presentation" and categorized as no, low (1 - 3 visits), or high (≥4 visits) use [15].

Data Management and Analysis

All analyses were performed using SAS 9.4 (SAS Institute Inc) within the Secure Unified Research Environment. Cohort characteristics and outcomes were summarized using descriptive statistics and Pearson chi-square tests. Modified Poisson regression models [16,17] with robust variance estimation were used to examine the association between patient characteristics and virtual care use and between virtual care use and unplanned ED visits within the same 12-month follow-up period. Among patients with both events, only 6% (4699/78,323) had an ED visit prior to their first virtual care encounter, indicating that, in the majority of cases, virtual care preceded ED presentation.

Separate fully adjusted models were fitted for each outcome and for each level, comparing low and high levels of each outcome with no use as the reference category [16]. Covariates were selected for their known association with service use and availability [18-20]. These included age group, gender, country of birth, Charlson comorbidities, remoteness, socioeconomic status, mental disorders, cancer type and morphology, COVID-19 lockdown period, and prior outpatient use.

Collinearity between covariates was assessed using pairwise correlation coefficients. Tobacco use was excluded due to strong collinearity with mental health disorders (r=0.90). Backward stepwise regression sequentially removed nonsignificant variables using a removal criterion of P>.05, and relevant 2-way interactions were tested. Analyses were conducted using complete cases and the proportion of missing data reported in the tables. A sensitivity analysis excluded unplanned ED visits with a COVID-19 diagnosis (ICD-10-AM code U07 as the principal or among the first 20 additional diagnoses). The results were reported as adjusted relative risks (RRs) with 95% CIs, with P≤.05 denoting significance.

Results

Characteristics of the Study Sample

The cohort included: 39,099 (49.9%) adults from Queensland and 39,224 (50.1%) from Victoria who accessed nonadmitted patient oncology services. Of these, 37,706 (48.1%) did not use virtual care during the 12-month follow-up, 24,196 (30.9%) had low virtual care use, and 16,421 had high (20.9%) virtual care use. High virtual care users had the highest proportion of unplanned ED visits, with 38.9% (n=4030) experiencing high ED use compared to 31.4% (n=3252) of low virtual care users and 29.8% (n=3091) of nonusers (Table 1, Table S2 in Multimedia Appendix 1).

XSL•FO
RenderX

Table . Unplanned ED[a] visits among individuals aged 18 years or older who used virtual outpatient oncology services in Queensland and Victoria, Australia, from 2018 to 2021 (retrospective cohort study; N=78,323).

Characteristics	No unplanned ED use (n=39,512), n (%)	Low unplanned ED use (≤3 use/y) (n=28,438), n (%)	High unplanned ED use (≥4 use/y) (n=10,373), n (%)	*P* value
Virtual outpatient oncology use				
No use	22,680 (57.4)	11,935 (42)	3091 (29.8)	<.001
Low use (≤3 use/y)	11,440 (29)	9504 (33.4)	3252 (31.4)	
High use (≥4 use/y)	5392 (13.7)	6999 (24.6)	4030 (38.9)	
Age categories (y)				
18 - 44	3177 (8)	2336 (8.2)	879 (8.5)	<.001
45 - 64	12,529 (31.7)	8976 (31.6)	3309 (31.9)	
65 - 74	11,416 (28.9)	7795 (27.4)	2872 (27.7)	
≥75	12,390 (31.4)	9331 (32.8)	3313 (31.9)	
Gender[b]				
Male	17,261 (43.7)	14,197 (49.9)	5654 (54.5)	.09
Female	22,249 (56.3)	14,241 (50.1)	4719 (45.5)	
Country of birth[c]				
Australia	26,888 (69.5)	19,544 (69.6)	7276 (71)	<.001
Other	11,796 (30.5)	8530 (30.4)	2968 (29)	
Rurality[d]				
Urban	25,969 (65.8)	17,141 (60.4)	5598 (54.1)	<.001
Rural	13,492 (34.2)	11,252 (39.6)	4760 (45.9)	
Charlson comorbidities (excluding malignant neoplasm)[e]				
No	20,595 (70.9)	13,156 (62.5)	4658 (53.3)	<.001
Yes	8453 (29.1)	7882 (37.5)	4077 (46.7)	
Other comorbidities				
Mental health disorder (yes)[e]	3620 (12.5)	4637 (22)	2892 (33.1)	<.001
Tobacco use (yes)[e]	13,210 (46.1)	11,239 (53.7)	5265 (60.6)	<.001
Socioeconomic status[f]				
Q1 (most disadvantaged)	7463 (18.9)	6718 (23.6)	2825 (27.2)	<.001
Q2	6461 (16.4)	5806 (20.4)	2347 (22.6)	
Q3	7741 (19.6)	5759 (20.3)	2035 (19.6)	
Q4	8054 (20.4)	5345 (18.8)	1704 (16.4)	
Q5 (most advantaged)	9784 (24.8)	4808 (16.9)	1461 (14.1)	
COVID-19 lockdown				
No	23,624 (59.8)	17,091 (60.1)	6319 (60.9)	.11
Yes	15,888 (40.2)	11,347 (39.9)	4054 (39.1)	
Cancer type				
Breast	10,540 (26.7)	5111 (18)	1248 (12)	<.001
Digestive organs including colorectal	5516 (14)	5950 (20.9)	2602 (25.1)	
Lung	2960 (7.5)	3975 (14)	1836 (17.7)	
Prostate	4643 (11.8)	2173 (7.6)	559 (5.4)	

XSL•FO
RenderX

Characteristics	No unplanned ED use (n=39,512), n (%)	Low unplanned ED use (≤3 use/y) (n=28,438), n (%)	High unplanned ED use (≥4 use/y) (n=10,373), n (%)	*P* value
Other cancers/unknown primary site/ill-defined	15,853 (40.1)	11,229 (39.5)	4128 (39.8)	
Cancer morphology				
Well or moderately differentiated	11,607 (29.4)	7664 (27)	2688 (25.9)	<.001
Poorly or undifferentiated	10,492 (26.6)	7273 (25.6)	2365 (22.8)	
Unspecified	17,413 (44.1)	13,501 (47.5)	5320 (51.3)	
Preindex outpatient use				
No use/low use	32,023 (81.1)	22,472 (79)	7798 (75.2)	<.001
High use (≥3/y)	7489 (18.9)	5966 (21)	2575 (24.8)	

[a]ED: emergency department.

[b]Data on gender were missing for 2 individuals and were not included in the analysis.

[c]Data on country of birth were missing for 1321 individuals and were not included in the analysis.

[d]Data on rurality were missing for 111 individuals and were not included in the analysis.

[e]Data on Charlson comorbidities, mental health disorders, and tobacco use were missing for 19,502 individuals and were not included in the analysis.

[f]Data on socioeconomic status were missing for 12 individuals and were not included in the analysis.

Predictors of Virtual Outpatient Oncology Use

High users of virtual care had a slightly higher proportion of individuals aged 18 to 44 years (ie, n=1627, 9.9%) compared to low users (n=1864, 7.7%) and nonusers (n=1864, 7.7%) and a higher proportion residing in regional/rural areas (n=7163, 43.6% vs n=9827, 40.6% and n=12,514, 33.2% for low users and nonusers, respectively). Mental health disorders were slightly more common among high users (n=3090, 22.9%) than among low users (n=3087, 18.4%) and nonusers (n=4972, 17.4%). Among high users, there was a higher proportion of digestive cancers (n=3919, 23.9%) and a lower proportion of prostate cancer (n=840, 5.1%). A higher proportion of high virtual care users (n=9929, 60.5%) received care during the COVID-19 period (Table S2 in Multimedia Appendix 1).

Older individuals had a lower likelihood of high virtual care use compared to adults aged 18 to 44 years (eg, ≥75 y: RR 0.78; 95% CI 0.74 - 0.83). Overseas-born individuals also had a lower likelihood of high virtual care use compared to adults born in Australia (RR 0.83; 95% CI 0.80 - 0.86). Individuals from socioeconomically advantaged areas had a lower likelihood of virtual care use; for example, adults in the least disadvantaged areas had a 23% lower likelihood of high virtual care use compared with patients in the most disadvantaged areas (RR 0.77; 95% CI 0.74 - 0.81).

Mental health disorders were associated with higher virtual care use (RR 1.28; 95% CI 1.24 - 1.33), as were Charlson comorbidities (RR 1.12; 95% CI 1.09 - 1.16), compared to individuals without these conditions. Compared to individuals with breast cancer, adults with digestive cancers (RR 1.24; 95% CI 1.19 - 1.30) or lung cancer (RR 1.14; 95% CI 1.07 - 1.20) had a higher rate of high virtual care use, while males with prostate cancer had a lower likelihood (RR 0.53; 95% CI 0.49 - 0.57).

Virtual care uptake was higher during the COVID-19 period, with individuals more likely to have low use (RR 1.47; 95% CI 1.43 - 1.51) and over 3 times more likely to have high use (RR 3.03; 95% CI 2.92 - 3.15) compared to the non–COVID-19 period. Additionally, individuals with high preindex outpatient use were more likely to have high virtual care use (RR 1.38; 95% CI 1.31 - 1.45) (Table S3 in Multimedia Appendix 1).

Association Between Virtual Care Use and Unplanned ED Visit

The higher proportion of virtual care use was observed among adults with higher unplanned ED use. Among individuals with no ED visits, 57.4% (n=22,680) had not used virtual care, whereas among adults with high ED use, only 29.8% (n=3091) had not used virtual care and 38.9% (n=4030) had high virtual care use.

Among individuals with high unplanned ED use, 46.7% (n=4077) had Charlson comorbidities, 33.1% (n=2892) had mental health disorders, 60.6% (n=5265) reported tobacco use, 24.8% (n=2575) had high preindex outpatient use, and 45.9% (n=4760) resided in rural areas. Socioeconomic disadvantage was more common among high unplanned ED users, with 27.2% (n=2825) in the most disadvantaged quintile, compared with 23.6% (n=6718) among those with low ED use and 18.9% (n=7463) among those with no ED use. Lung (17.7%, n=1836) and digestive cancers (25.1%, n=2602) were more prevalent among high unplanned ED users, whereas breast (12.0%, n=1248) and prostate cancers (5.4%, n=559) were less common (Table 1).

Compared to nonvirtual care users, individuals with low virtual care use had 1.27 times the risk (95% CI 1.1.24 - 1.31) and adults with high virtual care use had 2.64 times the risk (95% CI 2.52 - 2.79) of high unplanned ED use. Older age was associated with lower risk, with individuals aged 75 years or older having a 35% lower risk of high unplanned ED use compared to adults aged 18 to 44 years (RR 0.65, 95% CI

XSL•FO
RenderX

0.60 - 0.70). Females were less likely than males to have high unplanned ED use (RR 0.89, 95% CI 0.85 - 0.93) (Figure 1 and Table S4 in Multimedia Appendix 1).

Compared to referent groups, individuals born outside Australia had a slightly higher risk of high ED use (RR 1.08; 95% CI 1.03 - 1.13), as did rural residents (RR 1.15; 95% CI 1.10 - 1.20), adults with Charlson comorbidities (RR 1.38; 95% CI 1.33 - 1.44), mental health disorders (RR 1.64; 95% CI 1.57 - 1.71), and high preindex outpatient users (RR 1.18; 95% CI 1.12 - 1.24). Individuals from the least disadvantaged socioeconomic quintiles were less likely to experience unplanned ED use (RR 0.66; 95% CI 0.62 - 0.70).

Receiving care during the COVID-19 period was associated with a lower risk of high unplanned ED use (RR 0.88; 95% CI 0.84 - 0.92). Individuals with lung (RR 1.82; 95% CI 1.67 - 1.98) or digestive (RR 1.74; 95% CI 1.61 - 1.88) cancers had a higher risk of high unplanned ED use compared to adults with breast cancer (Figure 1 and Table S4 in Multimedia Appendix 1). No interaction effects were observed. Supplementary analysis excluding COVID-19-related encounters yielded similar results (Table S5 in Multimedia Appendix 1).

Discussion

Figure 1. Virtual outpatient oncology use and unplanned ED visits among individuals aged ≥18 years who used outpatient oncology services in Queensland and Victoria, Australia, from 2018 to 2021 (retrospective cohort study; N=78,323). ED: emergency department.

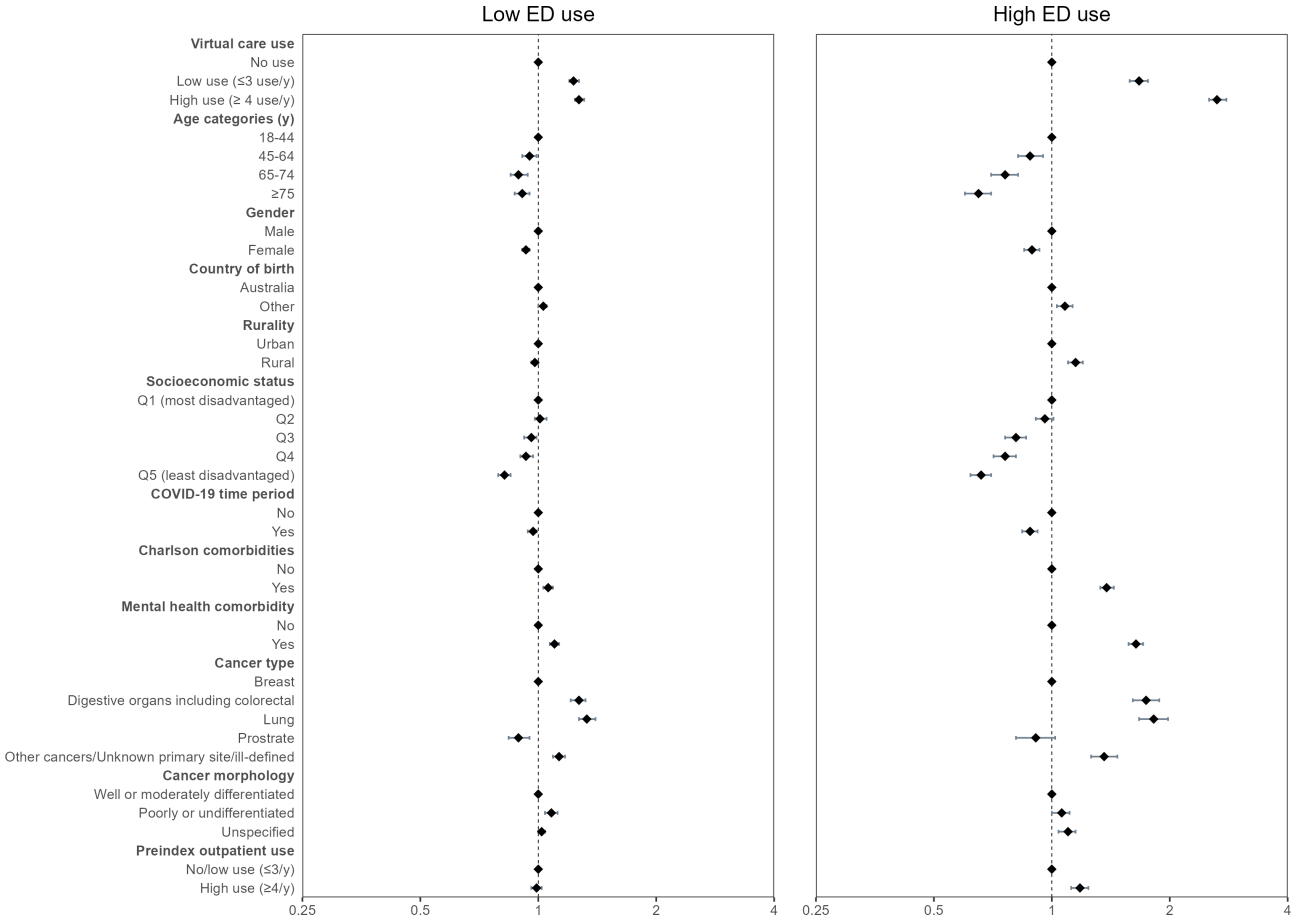

Main Findings

This study investigated factors associated with virtual care use and its association with unplanned ED visits among patients with cancer receiving outpatient care in Victoria and Queensland. Virtual care was more frequently used by younger adults, individuals residing in rural areas, and those with mental health conditions or a greater burden of Charlson comorbidities. Use was also higher during the COVID-19 lockdown period. In contrast, older adults and individuals born overseas were less likely to use virtual outpatient services. Notably, higher levels of virtual care use were associated with increased unplanned ED presentations.

Comparison With Existing Literature

The observed disparities in virtual care use among older adults and individuals born overseas align with previous research [21,22] and could reflect barriers, such as digital literacy, limited access to devices or reliable internet, and lower confidence in using technology for health purposes [23]. Similarly, individuals born overseas may encounter language barriers, cultural preferences for in-person care, and reduced access to digital services [24]. Adults with a higher comorbidity burden, including mental health disorders, were more frequent users of virtual care, potentially due to an increased need for regular monitoring, symptom management, and multidisciplinary input [25]. Additionally, virtual care uptake was higher during the COVID-19 lockdown period, as health services shifted online to maintain continuity and minimize infection risk. Cancer type

XSL•FO
RenderX

was also associated with virtual care use, with higher use observed among individuals with digestive or lung cancers compared to adults with breast cancer. This may reflect greater clinical complexity and a higher need for ongoing monitoring and management in these cancers [26,27].

Higher virtual care use was associated with a greater likelihood of unplanned ED presentations in this study. Similar associations have been reported in prior studies [28,29], whereas others have reported no difference or reduced use [30,31]. It is important to note that these associations may reflect underlying differences in patient characteristics, and a causal relationship with the mode of care cannot be inferred. A higher likelihood of unplanned ED visits may reflect greater clinical complexity among virtual care users, who were more likely to have comorbidities and high prior service use, all markers of greater care needs. In oncology, where rapid deterioration is common, virtual consultations may prompt early symptom detection and precautionary ED referrals for evaluation when in-person assessment is not feasible (eg, directing a patient reporting chest pain or fever straight to the ED). Notably, only 2.6% of ED visits in our sample were coded under ICD-10-AM Z00–Z13 (examination and investigation), suggesting that most ED presentations were not investigation-only encounters; however, these codes are assigned after clinical assessment and do not capture referral rationale. Alternatively, this observed association may indicate a more proactive care-seeking behavior among virtual care users or limitations in virtual care's ability to manage complex needs.

Older adults with cancer had a lower likelihood of unplanned ED visits, consistent with prior research among cancer survivors suggesting decreased ED use with age [18], possibly due to preferences for conservative care or greater engagement with other services or death [32]. Women had lower odds of unplanned ED visits, aligning with higher acute care use reported among male patients with cancer [19]. Patients in the socioeconomically least disadvantaged areas were less likely to present to the ED for an unplanned visit, potentially indicating better access to timely, coordinated care [33,34]. In contrast, individuals born overseas and adults living in rural areas had higher odds of unplanned ED use, which may be associated with systemic barriers, including limited access to primary or specialist care [35], language barriers [36], and longer travel times [37]. These findings align with literature on greater ED reliance in rural and underserved populations [35,38]. The COVID-19 lockdown period was associated with reduced unplanned ED presentations, reflecting international trends during the pandemic [39]. This decrease may be related to public health messaging, infection fears, and system-level measures discouraging hospital attendance [40].

Comorbidities, including mental health disorders, were associated with unplanned ED use, consistent with their known impact on clinical deterioration and crisis-driven health care utilization [25]. Patients with digestive and lung cancers had the highest likelihood of unplanned ED presentations, likely due to acute complications and rapid disease progression [26,27,41]. Increased preindex engagement with outpatient services was also associated with higher ED use, suggesting that prior care intensity may signal greater underlying need or a tendency toward more proactive care-seeking.

While conducted in a health system with relatively higher digital capacity, these findings are relevant to other settings, particularly those with larger digital divides or constrained oncology resources, where inequities in virtual care access may be amplified. Targeted interventions, such as digital literacy support for older adults and multilingual telehealth services for migrant populations, may help promote equitable access [42,43]. Further research is needed to better understand the content of virtual encounters, subsequent acute care pathways, and the extent to which observed associations reflect causal effects versus underlying patient complexity.

Strengths and Limitations

A key strength of this study is the use of large, population-based linked administrative datasets from both Victoria and Queensland, providing valuable insights into virtual oncology service use and unplanned ED presentations among patients living with cancer. However, its cross-sectional design limits interpretation to associations and precludes causal inference. Virtual care use and outcomes were measured within the same follow-up window, and despite virtual care preceding ED presentation for most patients, residual confounding is possible. Information on cancer stages and treatment types was not available, limiting the ability to account for treatment intensity, eligibility for virtual care use, and disease severity. Comorbidities were limited to hospital-recorded diagnoses, likely capturing only the most severe cases and resulting in underenumeration. However, the 1-year lookback period likely reduced but did not eliminate this limitation. Residual confounding may persist, as adults with more severe symptoms could preferentially choose virtual care, partly explaining its association with unplanned ED use. Virtual care use was classified using absolute visit counts, which may mask heterogeneity in clinical patterns across categories, particularly among higher-use patients. The use of a complete case approach may have introduced selection bias if missing data were not random. Preferred language was not included in the analysis due to substantial missing data (47%, n=36,812) and strong collinearity with country of birth. The provision of treatment during the COVID-19 pandemic further complicated interpretation, given service disruptions, rapid telehealth expansion, and varying digital readiness. However, a COVID-19 lockdown time period was included as a covariate in the analysis to distinguish nonpandemic and pandemic phases of care delivery. A stratified prepandemic and postpandemic analysis was not feasible due to low virtual care provision in the pre-COVID period. Finally, the content, quality, the modality of virtual encounters, or administrative data were not able to be validated.

Conclusion

This study highlighted disparities in virtual oncology use, with uptake highest among younger patients, rural residents, and adults with comorbidities, and lowest among older adults and individuals born overseas. To promote equity, policymakers and service planners should work to identify the factors driving these disparities and ensure appropriate use of virtual care.

XSL•FO
RenderX

Although virtual care enhances timely access, its association with higher unplanned ED presentations likely reflects greater clinical complexity and differences in care pathways. Further research, including studies using causal analytic approaches, is needed to better understand how virtual consultations are conducted, how they are integrated into broader care pathways, and how these factors influence acute care use.

Acknowledgments

The authors gratefully acknowledge the Queensland and Victorian Departments of Health for access to nonadmitted patient, emergency department visit, hospitalization, and mortality data, Cancer Alliance Queensland and the Victorian Cancer Registry for access to the state cancer registers, and the Queensland Statistical Services Branch and the Centre for Victorian Data Linkage for undertaking the record linkage. The authors also thank our Smarter Hospitals project partners, Health Infrastructure NSW; Western NSW, South Western Sydney, and Northern Health Districts, Townsville Hospital and Health Service, University of New South Wales Sydney, Western Health, Monash University, the University of Leeds, and Macquarie University for their invaluable support.

Funding

This work is funded by the National Health and Medical Research Council Partnership Scheme (grant number 2015544). The funder was not involved in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The data that support the findings of this study are available from the Victorian and Queensland Department of Health. Restrictions apply to the availability of these data, which were used under license for this study, so are not publicly available.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Cancer type classification, patient characteristics, predictors of virtual outpatient oncology use, and its association with unplanned emergency department presentations.
[DOCX File, 37 KB - jmir_v28i1e87694_app1.docx]

References

1. Phillips JL, Currow DC. Cancer as a chronic disease. Collegian 2010;17(2):47-50. [doi: 10.1016/j.colegn.2010.04.007] [Medline: 20738055]
2. Wu IQ, Lim F, Koh LP. Outpatient care. In: The Comprehensive Cancer Center: Development, Integration, and Implementation: Springer; 2022:21-33. [doi: 10.1007/978-3-030-82052-7_4]
3. Williamson TS. The shift of oncology inpatient care to outpatient care: the challenge of retaining expert oncology nurses. Clin J Oncol Nurs 2008 Apr;12(2):186-189. [doi: 10.1188/08.CJON.186-189] [Medline: 18390454]
4. Harrison R, Manias E. How safe is virtual healthcare? Int J Qual Health Care 2022 Apr 11;34(2):mzac021. [doi: 10.1093/intqhc/mzac021] [Medline: 35404457]
5. Harkey LC, Jung SM, Newton ER, Patterson A. Patient satisfaction with telehealth in rural settings: a systematic review.. Int J Telerehabil 2020 Dec 8;12(2):53-64. [doi: 10.5195/ijt.2020.6303] [Medline: 33520095]
6. Seivert S, Badowski ME. The rise of telemedicine: lessons from a global pandemic. EMJ Innov 2021:64-69. [doi: 10.33590/emjinnov/20-00239]
7. Jue JS, Spector SA, Spector SA. Telemedicine broadening access to care for complex cases. J Surg Res 2017 Dec;220:164-170. [doi: 10.1016/j.jss.2017.06.085] [Medline: 29180178]
8. Rodin D, Lovas M, Berlin A. The reality of virtual care: implications for cancer care beyond the pandemic. Healthc (Amst) 2020 Dec;8(4):100480. [doi: 10.1016/j.hjdsi.2020.100480] [Medline: 33129178]
9. Larson JL, Rosen AB, Wilson FA. The effect of telehealth interventions on quality of life of cancer patients: a systematic review and meta-analysis. Telemed J E Health 2018 Jun;24(6):397-405. [doi: 10.1089/tmj.2017.0112] [Medline: 29112484]
10. Cox A, Lucas G, Marcu A, et al. Cancer survivors' experience with telehealth: a systematic review and thematic synthesis. J Med Internet Res 2017 Jan 9;19(1):e11. [doi: 10.2196/jmir.6575] [Medline: 28069561]
11. Alishahi Tabriz A, Turner K, Hong YR, Gheytasvand S, Powers BD, Elston Lafata J. Trends and characteristics of potentially preventable emergency department visits among patients with cancer in the US. JAMA Netw Open 2023 Jan 3;6(1):e2250423. [doi: 10.1001/jamanetworkopen.2022.50423] [Medline: 36656584]

XSL•FO
RenderX

12. Sina M, Mitchell R, Walsan R, et al. Using virtual models of care for chronic disease management in outpatient services: a systematic review of quality of care outcomes. Telemed J E Health 2025 Sep;31(9):1049-1063. [doi: 10.1089/tmj.2025.0008] [Medline: 40323664]
13. Shaffer KM, Turner KL, Siwik C, et al. Digital health and telehealth in cancer care: a scoping review of reviews. Lancet Digit Health 2023 May;5(5):e316-e327. [doi: 10.1016/S2589-7500(23)00049-3] [Medline: 37100545]
14. Walsan R, Harrison R, Braithwaite J, et al. Predictors of Unplanned Emergency Department Visits and Hospitalization Among Adults Accessing Virtual Outpatient Oncology Services in Queensland, Australia. Telemedicine and e-Health . [doi: 10.1177/15305627261440755]
15. Soril LJJ, Leggett LE, Lorenzetti DL, Noseworthy TW, Clement FM. Characteristics of frequent users of the emergency department in the general adult population: a systematic review of international healthcare systems. Health Policy 2016 May;120(5):452-461. [doi: 10.1016/j.healthpol.2016.02.006] [Medline: 26947060]
16. Camey SA, Torman VBL, Hirakata VN, Cortes RX, Vigo A. Bias of using odds ratio estimates in multinomial logistic regressions to estimate relative risk or prevalence ratio and alternatives. Cad Saude Publica 2014 Jan;30(1):21-29. [doi: 10.1590/0102-311x00077313] [Medline: 24627010]
17. Iglesias-Rios L, Harlow SD, Reed BD. Depression and posttraumatic stress disorder among women with vulvodynia: evidence from the population-based woman to woman health study. J Womens Health (Larchmt) 2015 Jul;24(7):557-562. [doi: 10.1089/jwh.2014.5001] [Medline: 25950702]
18. Abdel-Rahman O. Gender, socioeconomic status and emergency department visits among cancer survivors in the USA: a population-based study. J Comp Eff Res 2021 Aug;10(12):969-977. [doi: 10.2217/cer-2020-0278] [Medline: 34156279]
19. Lash RS, Bell JF, Reed SC, et al. A systematic review of emergency department use among cancer patients. Cancer Nurs 2017;40(2):135-144. [doi: 10.1097/NCC.0000000000000360] [Medline: 26925998]
20. Handley NR, Schuchter LM, Bekelman JE. Best practices for reducing unplanned acute care for patients with cancer. J Oncol Pract 2018 May;14(5):306-313. [doi: 10.1200/JOP.17.00081] [Medline: 29664697]
21. Mistry SK, Shaw M, Raffan F, et al. Inequity in access and delivery of virtual care interventions: a scoping review. Int J Environ Res Public Health 2022 Aug 1;19(15):9411. [doi: 10.3390/ijerph19159411] [Medline: 35954768]
22. Abdel-Rahman O. Patient-related barriers to some virtual healthcare services among cancer patients in the USA: a population-based study. J Comp Eff Res 2021 Feb;10(2):119-126. [doi: 10.2217/cer-2020-0187] [Medline: 33448874]
23. Hall Dykgraaf S, Desborough J, Sturgiss E, Parkinson A, Dut GM, Kidd M. Older people, the digital divide and use of telehealth during the COVID-19 pandemic. Aust J Gen Pract 2022 Sep;51(9):721-724. [doi: 10.31128/AJGP-03-22-6358] [Medline: 36045630]
24. Mahumud RA, Shahjalal M, Dahal PK, et al. Emerging burden of post-cancer therapy complications on unplanned hospitalisation and costs among Australian cancer patients: a retrospective cohort study over 14 years. Sci Rep 2025 Feb 8;15(1):4709. [doi: 10.1038/s41598-025-89247-y] [Medline: 39922897]
25. Sarfati D, Koczwara B, Jackson C. The impact of comorbidity on cancer and its treatment. CA Cancer J Clin 2016 Jul;66(4):337-350. [doi: 10.3322/caac.21342] [Medline: 26891458]
26. Grewal K, Calzavara A, McLeod SL, et al. Emergency department use before cancer diagnosis in Ontario, Canada: a population-based study. CMAJ 2024 Nov 3;196(37):E1252-E1261. [doi: 10.1503/cmaj.240952] [Medline: 39496352]
27. Thompson CA, Sheridan P, Metwally E, et al. Emergency department involvement in the diagnosis of cancer among older adults: a SEER-Medicare study. JNCI Cancer Spectr 2024 Apr 30;8(3):pkae039. [doi: 10.1093/jncics/pkae039] [Medline: 38796687]
28. Hatef E, Lans D, Bandeian S, Lasser EC, Goldsack J, Weiner JP. Outcomes of in-person and telehealth ambulatory encounters during COVID-19 within a large commercially insured cohort. JAMA Netw Open 2022 Apr 1;5(4):e228954. [doi: 10.1001/jamanetworkopen.2022.8954] [Medline: 35471570]
29. Waseem N, Boulanger M, Yanek LR, Feliciano JL. Disparities in telemedicine success and their association with adverse outcomes in patients with thoracic cancer during the COVID-19 pandemic. JAMA Netw Open 2022 Jul 1;5(7):e2220543. [doi: 10.1001/jamanetworkopen.2022.20543] [Medline: 35797048]
30. Shah VV, Villaflores CW, Chuong LH, et al. Association between in-person vs telehealth follow-up and rates of repeated hospital visits among patients seen in the emergency department. JAMA Netw Open 2022 Oct 3;5(10):e2237783. [doi: 10.1001/jamanetworkopen.2022.37783] [Medline: 36282505]
31. Bange EM, Li Y, Kumar P, et al. The association between telemedicine, advance care planning, and unplanned hospitalizations among high-risk patients with cancer. Cancer 2024 Feb 15;130(4):636-644. [doi: 10.1002/cncr.35116] [Medline: 37987207]
32. Tolppanen AM, Lamminmäki A, Kataja V, Tyynelä-Korhonen K. Specialized palliative outpatient clinic care involvement associated with decreased end-of-life hospital costs in cancer patients, a single center study. BMC Palliat Care 2024 Dec 28;23(1):302. [doi: 10.1186/s12904-024-01633-x] [Medline: 39731088]
33. Walsan R, Mayne DJ, Feng X, Pai N, Bonney A. Examining the association between neighbourhood socioeconomic disadvantage and type 2 diabetes comorbidity in serious mental illness. Int J Environ Res Public Health 2019 Oct 15;16(20):3905. [doi: 10.3390/ijerph16203905] [Medline: 31618833]

XSL•FO
RenderX

34. Sheringham J, Asaria M, Barratt H, Raine R, Cookson R. Are some areas more equal than others? Socioeconomic inequality in potentially avoidable emergency hospital admissions within English local authority areas. J Health Serv Res Policy 2017 Apr;22(2):83-90. [doi: 10.1177/1355819616679198] [Medline: 28429977]
35. Stephens AS, Dinh MM, Kinsman L. Patterns of emergency department use in rural and metropolitan New South Wales by socioeconomic status: a population-based study. Emerg Med Australas 2023 Jun;35(3):489-495. [doi: 10.1111/1742-6723.14155] [Medline: 36571146]
36. Chu JN, Wong J, Bardach NS, et al. Association between language discordance and unplanned hospital readmissions or emergency department revisits: a systematic review and meta-analysis. BMJ Qual Saf 2024 Jun 19;33(7):456-469. [doi: 10.1136/bmjqs-2023-016295] [Medline: 38160059]
37. Cerni J, Rhee J, Hosseinzadeh H. End-of-life cancer care resource utilisation in rural versus urban settings: a systematic review. Int J Environ Res Public Health 2020 Jul 9;17(14):4955. [doi: 10.3390/ijerph17144955] [Medline: 32660146]
38. Credé SH, Such E, Mason S. International migrants' use of emergency departments in Europe compared with non-migrants' use: a systematic review. Eur J Public Health 2018 Feb 1;28(1):61-73. [doi: 10.1093/eurpub/ckx057] [Medline: 28510652]
39. Patt D, Gordan L, Diaz M, et al. Impact of COVID-19 on cancer care: how the pandemic is delaying cancer diagnosis and treatment for American seniors. JCO Clin Cancer Inform 2020 Nov;4:1059-1071. [doi: 10.1200/CCI.20.00134] [Medline: 33253013]
40. Lange SJ, Ritchey MD, Goodman AB, et al. Potential indirect effects of the COVID-19 pandemic on use of emergency departments for acute life-threatening conditions - United States, January-May 2020. Am J Transplant 2020 Sep;20(9):2612-2617. [doi: 10.1111/ajt.16239] [Medline: 32862556]
41. Galdas PM, Cheater F, Marshall P. Men and health help-seeking behaviour: literature review. J Adv Nurs 2005 Mar;49(6):616-623. [doi: 10.1111/j.1365-2648.2004.03331.x] [Medline: 15737222]
42. Jung J, You J, Kim D. Effective but sustainable? A case of a digital literacy program for older adults. Educ Inf Technol 2025 Jul;30(10):13309-13330. [doi: 10.1007/s10639-025-13364-4]
43. Gallegos-Rejas VM, De Guzman KR, Kelly JT, Smith AC, Thomas EE. Strategies to improve telehealth access for culturally and linguistically diverse communities: a systematic review. J Public Health (Oxf) 2025 Aug 29;47(3):e374-e390. [doi: 10.1093/pubmed/fdaf030] [Medline: 40106706]

Abbreviations

ED: emergency department
ICD-10-AM: International Classification of Diseases, 10th Revision, Australian Modification
QHNAPDC: Queensland Health Non-Admitted Patient Data Collection
RR: relative risk
VINAH: Victorian Integrated Non-Admitted Health

Edited by A Mavragani; submitted 12.Nov.2025; peer-reviewed by EP Greenup, E Bange, Z Liu; revised version received 02.Feb.2026; accepted 24.Mar.2026; published 30.Apr.2026.

Please cite as:
Walsan R, Harrison R, Westbrook J, Chauhan A, Moscova M, Vandyke A, Manias E, Webster T, Taylor N, Peprah P, Mitchell R Characteristics of Patients Accessing Outpatient Oncology Services Virtually and Predictors of Subsequent Unplanned Emergency Department Presentations in 78,323 Adults in Australia: Retrospective Cohort Study
J Med Internet Res 2026;28:e87694
URL: https://www.jmir.org/2026/1/e87694
doi:10.2196/87694

© Ramya Walsan, Reema Harrison, Johanna Westbrook, Ashfaq Chauhan, Michelle Moscova, Anita Vandyke, Elizabeth Manias, Tracey Webster, Natalie Taylor, Prince Peprah, Rebecca Mitchell. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Understanding How a Digital Platform for Chronic Disease Management Can Enable and Limit Patient Self-Care: Qualitative Study

Lysanne Lessard[1,2], PhD; Mark de Reuver[3], PhD; Kerri-Anne Mullen[4,5], PhD

[1]Telfer School of Management, University of Ottawa, Ottawa, ON, Canada

[2]Institut du Savoir Montfort, Ottawa, ON, Canada

[3]Engineering Systems and Services, Technology Policy and Management, Delft University of Technology, Postbus 5, 2600 AA, Delft, The Netherlands

[4]Ottawa Heart Institute, Ottawa, ON, Canada

[5]School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

Corresponding Author:
Mark de Reuver, PhD
Engineering Systems and Services, Technology Policy and Management, Delft University of Technology, Postbus 5, 2600 AA, Delft, The Netherlands

Abstract

Background: A growing segment of the population requires ongoing care and support for managing their chronic diseases. Digital platforms for self-management are rapidly emerging to meet this need, but patients' experiences with these platforms vary significantly. This may be due to the complexity and flexibility of digital platforms, where the wide array of available features can generate unexpected impacts.

Objective: This study aims to explore how a digital platform can both enable and limit patients with a chronic disease in managing their own health.

Methods: We conducted semistructured qualitative interviews with patients to better understand their experience of using a digital platform for self-managing their chronic diseases. Patients who had been using a digital platform (the Chronic Care Platform) for at least 1 month were invited to participate. Twenty-four patients were recruited and interviewed in person or by phone. The collected data were analyzed using template analysis, which is a type of thematic analysis that allows inductive identification of themes from data and deductive application of theory-informed themes. We leveraged Self-Care Theory to understand how patients' motivation to use the platform and their subsequent use of its features generated perceived value and challenges in achieving this value.

Results: The platform was shown to support patients' development of core self-care abilities (cognitive, psychosocial, and sociocultural abilities) and self-care behaviors (maintenance, monitoring, management), but it did not provide any support to the development of physiological abilities. Moreover, results indicate important limitations in the way in which the digital platform supported all self-care abilities and behaviors—in particular, self-care management. Hence, while the platform was viewed as valuable overall, patients reported several challenges in effectively using the Chronic Care Platform for self-care.

Conclusions: Digital platforms for chronic disease management can enhance patient self-care by providing valuable resources and support for reinforcing desired behaviors. However, gaps in platform features can limit patients' ability to comprehensively care for themselves. This study shows that relating platform features to specific dimensions of self-care can help identify missing features, providing a fine-grained understanding of how a given platform is generating positive impacts and how it may be improved to fully support self-care.

(J Med Internet Res 2026;28:e71875) doi:10.2196/71875

KEYWORDS

digital platform; chronic disease management; self-care; qualitative study; self-management; patient experience

Introduction

There has been a key shift in disease prevalence across the globe over the past several decades. As improvements in the treatment of communicable diseases have led to a global decrease in premature mortality, aging populations are now increasingly affected by one or more noncommunicable diseases, also called chronic diseases [1]. As a result, chronic diseases such as cardiovascular disease, diabetes, and chronic respiratory diseases have become leading causes of death, accounting for over 74% of deaths globally every year [2]. Yet, the burden of chronic

XSL•FO
RenderX

diseases can be significantly diminished by reducing their risk factors, which include modifiable behaviors such as tobacco use, unhealthy diet, and lack of exercise. While varied types of interventions can help individuals manage these behaviors, self-management support interventions are both the most frequently used and the most successful in terms of improvements in patient-level outcomes [3].

Information and communication technologies are increasingly used to deliver self-management interventions for chronic disease management [4], as well as a wide variety of related interventions, including medical consultations [5], education, and early detection [6]. Solutions focused on self-management allow patients to play an active role in the management of their disease and facilitate interactions and collaborative care among patients and health care professionals [7]. Moreover, they can be key to scaling up chronic disease interventions in a cost-effective manner [8,9]. Studies of digital interventions often show clinical and health benefits for chronic conditions. For example, a review of 65 systematic reviews and randomized controlled trials showed that interventions using mobile health technology for chronic diseases such as chronic respiratory disease, diabetes, chronic kidney disease, and cardiovascular diseases generally have positive behavioral and clinical effects [10]. Similarly, a meta-analysis of 26 articles consisting of 22 studies with 2645 participants showed that smartphone-based self-management interventions for diabetes patients have positive impacts on self-efficacy, self-care activities, and health-relevant outcomes such as quality of life [11].

Recently, digital platforms offering a multitude of information and communication technology–based interventions for self-management have emerged [12]. These platforms provide wide-ranging support, such as learning resources, personalized self-monitoring, and communication with peers and care professionals. Digital platforms can flexibly adapt and update these interventions through generic hardware, software, and data components [13]. While studies of patients' experiences with platform-based interventions for chronic disease management show overall positive results, mixed experiences and expectations regarding aspects such as data entry burden, digital literacy, and provider feedback can hinder the adoption and perceived value of these interventions [14,15]. Such ambivalent experiences [14] may result from the wide array of platform features that can generate unexpected impacts, as one feature may reinforce or hinder another. The complexity and flexibility of digital platforms thus necessitate new approaches to understanding their impacts on self-management outcomes, beyond classical approaches such as clinical trials focused on individual interventions.

This study aims to explore how a digital platform can both enable and limit patients suffering from a chronic condition in managing their own health. The context for this study is a Canadian chronic disease management program for people who are at risk of cardiovascular disease or who have suffered from cardiovascular issues. A digital platform, referred to as the Chronic Care Platform in this study, is used to deliver the health program. We use Self-Care Theory [16] to articulate the positive and negative value that patients perceive from the use of the Chronic Care Platform. Self-Care Theory reframes the concept of self-management into several abilities and behaviors that are required for individuals to successfully manage their conditions [17,18]. As a result, we identify important gaps in the way in which the platform enables patients to care for their health. Beyond the Chronic Care Platform, this study illustrates an approach to analyzing the comprehensive interplay of platform features and self-care abilities and behaviors. The results of this study can also be used to support the configuration and design of digital platforms for chronic disease self-management.

Methods

Overview

This study adopts a qualitative inquiry approach, which seeks to build an understanding of the world from the experiences, beliefs, and ideas of people while acknowledging that multiple interpretations of reality may coexist [19]. This approach was chosen because it allows for an understanding of patients' experiences using a digital platform. Semistructured interviewing was chosen as the data collection method as it is useful for gaining an understanding of participants' perspectives related to a domain of interest while remaining open to exploring other relevant topics during the course of the interview [20]. The study focuses on a single digital platform to enable a richer understanding of the relationship between specific platform features and patient experiences. We sought to recruit patients who had used this platform for at least 1 month to ensure sufficient experience for the study.

We used the SRQR (Standards for Reporting Qualitative Research) reporting guideline when editing this manuscript [21], and its associated checklist is included in Checklist 1.

Ethical Considerations

The study was conducted in compliance with the Declaration of Helsinki, including anonymizing research participant data throughout the analysis and reporting process, ensuring that participants did not feel any undue pressure to take part in the study nor to answer interview questions, and holding interviews at times and locations that were considered convenient for each participant. Written informed consent was obtained from all participants before data collection. Participants were not compensated. The study protocol and all associated methods were approved by the University of Ottawa's Office of Research Ethics and Integrity (S-11-19-5244).

Chronic Care Platform

The study focuses on patients' experiences with a digital platform referred to as the Chronic Care Platform for the purpose of this study. The platform is used to deliver a chronic disease management program for individuals who have suffered or are at risk of suffering from cardiovascular disease. The program was created and is managed by a hospital providing specialized secondary and tertiary care and had been underway for several years when the study started and remains available. The program targets 2 populations: patients who have received heart disease–related care at this hospital (post discharge) and individuals at risk of experiencing heart disease (prevention). The goal of this program is to help both populations improve

XSL•FO
RenderX

their risk factors in order to prevent future life-threatening events.

The Chronic Care Platform is designed to facilitate access to the program and allow the hospital to serve a larger number of individuals. The platform is accessible through a web browser and an app offering features desired by patients to manage their conditions [22]: an online library with varied educational material related to medication, prognosis, and care; a personal health record where patients can record, access, and manage their health information; care and wellness plan; and online rehabilitation support with exercise guidelines. Additionally, the Chronic Care Platform allows scheduling meetings with a health coach for patients who are given access to as part of the program, sharing information with individuals such as family members and physicians as desired by the patient, and interacting with other users through a forum. The platform also supports the development of care and wellness plans, and the tracking and monitoring of varied metrics such as blood pressure, steps, weight, exercise, and mood. It provides visual displays to understand trends related to these metrics (eg, weight gain, daily exercise duration). Information can be tracked through manual entry or by syncing with wearable and health-monitoring devices and apps. As such, the Chronic Care Platform reflects typical digital platforms for self-management.

Participant Recruitment

Participant recruitment followed a convenience sampling strategy. Patients enrolled in the chronic care management program, delivered through the Chronic Care Platform, were invited to participate in the study. Invitations to participate in the study were sent to platform users by the health care program coordinator using the platform messaging system. The message invited interested patients to contact the researchers directly. Recruitment material was also posted on the walls of the hospital where the chronic care management program was being administered. Patients who were using or had used the Chronic Care Platform for at least 1 month were invited to participate in the study. No additional inclusion or exclusion criteria were used in order to allow the recruitment of participants with varied experiences. However, since participants were recruited within the chronic care management program, they were adults (aged 18 y or older) at risk of heart disease or with chronic heart disease. All interested and eligible participants were recruited for the study since the experiences of any user of the platform were of interest to the study.

Data Collection

The first author and a research assistant conducted semistructured interviews with the 24 recruited patients. The first author had experience conducting such interviews and provided training to the research assistant prior to starting interviews. They did not have any preexisting relationship or relationship of authority with the participants. Following the same predefined interview protocol for each interview, patients were asked about their motivation for using the platform, how they used the platform, and the resulting value perceived alongside any challenges in realizing this value (Table 1). Each interview concluded with an open-ended question that encouraged patients to share anything they thought to be relevant about the Chronic Care Platform. Nine interviews were conducted in person and the remainder by phone. In-person interviews were held at locations chosen for their convenience by the participants; these locations included private rooms at the hospital delivering the chronic disease management program and public spaces such as coffee shops. Participants chose a space or table that they felt was appropriate for the interview in the latter case.

Table . Categories and questions asked to participants.

Category	Question
Context	• How did you hear about the Chronic Care Platform? • Are you using it as part of a postdischarge care plan, for preventing future issues, or both? • How long have you been using it?
Motivation	• Why did you decide to use the Chronic Care Platform? • Do you feel that it allows you to reach certain goals or solve certain problems? Which ones?
Platform use	• Who do you interact with through the Chronic Care Platform? • How do you use the Chronic Care Platform in order to reach the goals or solve the problems that you stated previously? • Which platform content, services, functionalities, or tools do you use?
Perceived value and challenges	• How is the Chronic Care Platform or its contents beneficial to you in your everyday life? • Have you experienced challenges in using the Chronic Care Platform?
Open	• Would you like to share anything else about the Virtual Care platform or your experience with it?

Interviews lasted between 20 and 60 minutes, with an average of 30 minutes. They were audio-recorded with the consent of participants and transcribed verbatim by a research assistant and verified for accuracy by the first author. Any identifying

XSL•FO
RenderX

information was removed from the transcripts prior to data analysis, and each participant was labeled with a number (eg, P1). As the number of participant interviews increased, substantial redundancies emerged in reported motivations, usage, and perceptions, indicating that the dataset had reached saturation [23]. However, data saturation was not used as a criterion for concluding data collection, and all recruited participants were interviewed.

Data Analysis

Template analysis was used to analyze interview data. It is a type of thematic analysis that allows both the inductive identification of themes from data and the deductive application of theory-informed themes [24,25]. Each participant interview was thus first coded inductively using labels that reflected the collected data, and similar codes across participant interviews were then merged into initial themes that reflected broader meanings. This process allowed us to discern emerging patterns of use through recurring themes regarding patients' motivation for using the platform (eg, "understanding," "accessing care"), how the platform was used (eg, "tracking," "self-educating"), and the kinds of benefits and challenges that patients experienced (eg, "being encouraged," "irrelevant information"). Two researchers were involved in this process, comparing code application, discussing any discrepancy, and clarifying label and theme definitions.

We then turned to existing theories that could provide a relevant explanation for these patterns. We initially reviewed behavior change theories, including the Theory of Planned Behavior [26] and the Health Belief Model [27]. However, while these theories focus on explaining what may drive individuals' intended and observed behaviors (eg, the concept of "perceived benefits" in the Health Belief Model), they did not provide sufficient coverage of the emerging themes, nor did they help to achieve the study's objective. These theories and others that are frequently cited in informatics, such as the Technology Acceptance Model [28], propose concepts that can be used to predict the adoption of healthy behaviors and supporting technology. However, their focus on behavioral intentions does not provide the concepts needed to unpack the behaviors themselves such as the platform uses that were present in our data (eg, "self-educating"). We thus turned to Self-Care Theory [16], which provided a natural fit to our data and emerging themes.

Self-care refers to purposeful actions or activities that are performed deliberately and continuously by an individual to maintain their life, healthy functioning, and well-being [29,30]. These activities can be initiated independently by the individual or in collaboration with health care professionals [30]. The notion of self-care thus aligns with common definitions of self-management [31]. In the context of chronic conditions, self-care is key to empowering people in assuming responsibility for their own health [32]. This understanding of self-care has been formalized in Self-Care Theory, which distinguishes between an individual's capacity to engage in self-care behavior (self-care abilities), and the actions that can be performed to realize this care (self-care behaviors [17,18,33]).

Self-care abilities encompass 4 groups of factors: cognitive, psychosocial, physiological, and sociocultural [17,18,34] (Textbox 1). While some authors also include demographic factors such as age, we do not include them here as they can be viewed as basic conditioning factors whose impacts are better captured through specific self-care ability domains [34]. Self-care behaviors refer to 3 interrelated processes that are initiated by individuals on their own or in collaboration with health care professionals: self-care maintenance, self-care monitoring, and self-care management [17,18] (Textbox 1).

XSL•FO
RenderX

Textbox 1. Factors supporting self-care ability and processes of self-care behavior.

Factors supporting self-care ability are as follows:

1. *Cognitive factors* such as learning skills provide an individual with the cognitive skills needed to fulfill self-care behaviors. They include cognitive function, learning skills, memory, problem-solving and goal-setting skills, awareness and knowledge of health condition, and knowledge of self-care behaviors.
2. *Psychosocial factors*, also referred to as emotional factors, focus on an individual's feelings, values, and motivations that can influence if and how they will perform self-care. These factors include an individual's self-esteem, self-discipline ability, personality traits, perceived self-competence, self-efficacy, motivation, and perception that the behavior to be carried out is efficacious.
3. *Physiological factors*, such as psychomotor skills, also referred to as physical factors, impact the ability of an individual to perform needed self-care behaviors. They relate to an individual's dexterity, psychomotor skills, functional or movement level, as well as their disability or injury status.
4. *Sociocultural factors*, such as an individual's family system, can enable or constrain one's self-care behaviors. These contextual factors include an individual's family system, cultural beliefs and practices, social support, health beliefs and values, available resources, and access to care.

Processes of self-care behavior are as follows:

1. *Self-care maintenance* refers to the selection of appropriate activities to improve well-being or manage health problems. For chronic conditions, these activities often reflect those recommended by health care providers and include lifestyle-related activities such as preparing healthy foods and those related to following a medical regimen such as taking medication as prescribed.
2. *Self-care monitoring* is a process of surveillance of one's body that aims at allowing an individual to recognize that a change has occurred. In the context of chronic conditions, this process is key to enabling the recognition of health problems through changes in signs or symptoms.
3. *Self-care management* is a process of implementing a treatment in response to changes indicating a deterioration in condition, as well as evaluating the effectiveness of that treatment. The treatment may involve an independent modification of self-care maintenance activities or a change requiring consultation with a health care provider. For patients with a chronic condition, the preference is for empowering them and their care partners (eg, spouse) to independently select and engage in self-care activities.

The Self-Care Theory concepts (in italics in Textbox 1) were deductively applied to identified themes and their data by 2 researchers, comparing code application, discussing any discrepancy, and clarifying code definition. Relationships between concepts were also identified when present in the data (eg, "Cognitive abilities—supporting—self-care maintenance"). Each interview was then coded by 1 researcher, with frequent revision of code application by the second author and inputs on the interpretations by the third author. Both the inductive and deductive phases of data analysis were informed by the research team's training and experience in digital technology and platforms developed and used in health care contexts (LL, MdR) as well as in cardiovascular health and prevention (KAM). Their combined viewpoints allowed the team to identify meaningful information that could have been glossed over from any single perspective and to compare varied assumptions and interpretations.

Results

Overview

A total of 33 potential participants contacted the researchers. Two of them were excluded because they had used the platform less than 1 month, 6 participants did not respond to follow-up communications, and 1 withdrew from the study prior to being interviewed. A sample of 24 participants consisting of 9 (37.5%) women and 15 (62.5%) men was thus recruited. Among them, 79% (19/24) participants were part of the postdischarge program and 21% (5/24) participants belonged to the prevention program. Twelve of 24 (50%) participants were receiving or had received advice from a health coach provided by the program for the first 6 months of use. The frequency at which they used the platform varied from occasional to daily, with 42% (10/24) accessing the platform on a weekly or daily basis. All participants had been using the platform for at least 1 month, 12 (50%) of them having been regular users for 1 year or more. Table 2 provides more detailed information on participant characteristics.

XSL•FO
RenderX

Table . Participant characteristics (N=24).

Participant number	Sex	Program	Health coach	Frequency of use	Length of use
P1	Male	Postdischarge	No	Daily	<3 mo
P2	Female	Postdischarge	No	Daily	3 - 6 mo
P3	Female	Prevention	Yes	>Weekly	>1 y
P4	Male	Postdischarge	Initially	Rarely	>1 y
P5	Female	Postdischarge	Yes	>Weekly	3 - 6 mo
P6	Female	Postdischarge	Initially	>Monthly	>1 y
P7	Male	Postdischarge	No	Daily	>1 y
P8	Male	Postdischarge	Initially	Monthly	>1 y
P9	Male	Post-discharge	No	>Year	>1 y
P10	Male	Postdischarge	No	Weekly	6 mo to 1 y
P11	Female	Prevention	Initially	Never	6 mo to 1 y
P12	Male	Postdischarge	No	>Monthly	6 mo to 1 y
P13	Female	Prevention	Initially	Monthly	6 mo to 1 y
P14	Male	Postdischarge	No	Daily	>1 y
P15	Male	Postdischarge	Initially	>Weekly	>1 y
P16	Male	Postdischarge	Yes	>Weekly	<3 mo
P17	Male	Postdischarge	Yes	Rarely	6 mo to 1 y
P18	Male	Postdischarge	No	Never	6 mo to 1 y
P19	Female	Prevention	No	>Monthly	a[a]
P20	Male	Postdischarge	No	>Weekly	>1 y
P21	Male	Postdischarge	Initially	a[a]	>1 y
P22	Male	Postdischarge	Yes	>Weekly	<3 mo
P23	Female	Postdischarge	No	Monthly	>1 y
P24	Female	Prevention	No	Rarely	>1 y

[a]“a” indicates that the participant did not answer.

Platform Support to Self-Care

In our analysis, we investigated how motivations for self-care translated into platform use, which subsequently not only generated perceived value but also highlighted challenges for patients in effectively using the platform to achieve the desired goals. Overall, we found that the platform offered support for all self-care abilities and behaviors except physiological abilities (Table 3). The platform, however, shows limitations for self-care management support and important challenges in its support to all self-care abilities and behaviors.

XSL•FO
RenderX

Table . Summary of the identified patients' patterns of use of the Chronic Care Platform for each self-care ability and behavior.

Self-Care Theory concepts	Motivations	Platform features (and external resources)	Platform use	Perceived value	Challenges
Cognitive abilities	• Understanding how to change behaviors and improve health using a trusted source of information	• Health-related generic information • Health-related questionnaires • Personalized care plans, workbooks, and reports • Space for manual medical data entry and goals • (External: websites containing health-related information)	• Accessing health-related information and prescribed activities (eg, exercise) • Entering personal goals and medical information • Filling out surveys, consulting workbooks, and reports	• Feeling educated, inspired, and able to achieve goals	• Missing personal tailoring and more advanced information • Not understanding platform functioning or where to find desired content • Difficulty in processing information due to disease-related cognitive disability
Psychosocial abilities	• Finding positive validation and reaffirming motivation to change behavior	• Information on mental health • Reward system providing automated positive feedback based on tracked measures	• Educating oneself about anxiety and stress • Consulting reward points and automated congratulations	• Feeling reassured, motivated, and empowered	• Not perceiving tangible benefit from reward system
Physiological abilities	• No data	• No data	• No data	• No data	• No data
Sociocultural abilities	• Connecting with other patients and health care providers • Feeling safe, watched over	• Discussion forum • Scheduling and emailing feature • Data sharing feature • (External: social apps, in-person visits to general practitioner, in-person interactions with family members and peers)	• Reading content written by other patients on the forum • Communicating and sharing data with health coach • Adding chosen individuals to account, allowing them to view data and send reminders	• Interacting and learning from peers • Getting support from a health coach	• Missing true social interaction • Sharing data with external physicians is difficult
Self-care maintenance	• Changing lifestyle and developing healthy behaviors, routines, and solutions	• Exercise recommendations and descriptions • (External: classes and equipment at local gym)	• Initiating lifestyle changes and exercising independently	• Integrating new behaviors into lifestyle	• Lacking proactive recommendations/guidance
Self-care monitoring	• Efficiently measuring own activity, food intake, and health data • Understanding cause and effect among metrics	• Manual data entry • Data importation from compatible tracking devices • Dashboard showing trends over time • (External: tracking devices and apps)	• Monitoring health-related measures and symptoms • Viewing trends on platform dashboard	• Having relevant insights into progress, reinforcing willingness to continue self-care efforts	• Entering data and synchronizing with external devices are challenging

XSL•FO
RenderX

Self-Care Theory concepts	Motivations	Platform features (and external resources)	Platform use	Perceived value	Challenges
Self-care management	• Assessing whether lifestyle changes are effective	• Dashboard showing trends over time	• Viewing trends on platform dashboard	• Visualizing progress over time	• Unable to self-assess if health status requires physician input • No alerts going to health care provider

Patterns of Use for Self-Care Abilities

Cognitive Abilities

Patients are looking for reliable information and insight into how to change behavior. To do so, they combine generic health information, such as articles related to diet, exercise, and lifestyle, and patient-specific information with interactive resources such as exercise workbooks. As a result, they feel educated, inspired, and reminded to change behavior:

> *I think one of the biggest things is that it's a great reminder to me. I was reading something on diabetes, now I'm well versed on diabetes but it never hurts to be reminded.* [P3]

Patients also feel connected, able to set goals, and structure their lifestyle to achieve them. However, some patients were disappointed by the generic and repetitive nature of available information, especially patients looking for more expert information:

> *I would like to see more [...content...] in the library...that pointed directly to me...because there's a lot of associations or references...which doesn't really apply to me.* [P7]

This led some patients to use external, invalidated information rather than the platform:

> *I'm active online, so I get...Dr. I can't remember who,...and they offer advice - regular advice online, and so I find other - I'm bombarded with all sorts of information that seems to come at me.* [P8]

Psychosocial Abilities

Patients seek positive validation and reaffirming motivation to change behavior. Patients educate themselves about how their mental health can impact their motivation and remain accountable and motivated through automated feedback provided by the platform:

> *When you log something they send you messages right away, keep up the good work. Positive reinforcement which is really great.* [P2]

As a result, they feel reassured that they are not alone in this process and feel empowered and motivated to continue. However, some patients felt that actual rewards would have motivated them more:

> *Because I have 6600 points and I don't know what those points are for. I'm not getting a prize or a cup of coffee at the end.* [P2]

Sociocultural Abilities

Three dimensions of sociocultural abilities were highlighted through data analysis: availability of resources, social support, and access to care. The first dimension concerns the convenience of using the platform to avoid using personal resources such as time and money to receive in-person support and care:

> *We only have one car between the two of us and we are going different ways all the time...and then the parking is atrocious there. Just getting a spot and then once you get it's expensive.* [P4]

This motivation is thus not related to a specific platform functionality.

The second dimension, social support, is motivated by the desire to connect with others. The platform supports this dimension through the forum, which enables reading content written by other patients and interacting with them, and the data sharing feature allows patients to add individuals to their environment, typically family members:

> *[...] so data from my [external apps and devices] would find a way onto the platform, which I found extremely useful for me for keeping a record, but also for those who were sharing that information, and I had my physiotherapist, my wife, the health coach.* [P7]

As a result, patients feel reassured and enjoy interacting and learning from peers. However, few participants reported using the forum, and some participants felt that real social interactions were missing, particularly in relation to the forum:

> *It would have been nice to say hey everyone, let's be buddies or whatever, because the social aspect would be nice.* [P10]

The third dimension, access to care, is related to the desire to get advice from health care providers by sharing information with them, as well as to feel safe and cared for:

> *[...] part of my rehabilitation was to change my lifestyle and I felt that the platform gave me a virtual link to somebody...the health coach to make sure that I was on the right track. It just gave me some reassurance that somebody was interested, besides my wife and my family of course.* [P4]

Patients used meeting and scheduling features to interact with the health coach but were sometimes disappointed that the data on the platform were not used by their physicians, either because they are unwilling or unaware:

XSL•FO
RenderX

> *Okay I had a six month interview with [a physician external to the intervention] about 2 - 3 weeks ago. I...showed him the blood pressure history—but...he didn't seemed to be overly interested.* [P5]

Relatedly, participants were uncertain who would benefit from the effort of inserting data.

Physiological Abilities

The collected data did not generate any results regarding participants' physiological abilities. Hence, when participants were asked about their motivation for using the Chronic Care Platform, the way in which they used it, and the value and challenges that they perceived through that use (Table 1), participants did not refer to elements that could be categorized as physiological or physical abilities that allowed them to carry out self-care behaviors, or hindered them from doing so.

Patterns of Use for Self-Care Behaviors

Self-Care Maintenance

Participants want to change their lifestyle and develop healthy behaviors, routines, and solutions. The platform offers exercise recommendations and descriptions to do so. Patients use these features, at times in combination with external resources such as a local gym, to exercise independently and initiate lifestyle changes:

> *I prefer to be on my own, do my own workouts and updating the system. To me that was a natural.* [P4]

As a result, participants experience benefits by integrating new behaviors into their lifestyle. However, some participants expressed the desire for more proactive recommendations and guidance.

Self-Care Monitoring

Patients are interested in efficiently measuring their own activity, food intake, and health data, and in understanding cause and effect among metrics. To do so, they both manually enter data such as weight, food intake, and blood sugar levels and connect external tracking devices such as smart watches to the platform. They also make use of nonconnected tracking devices and diaries. They are thus able to monitor an array of health-related measures and symptoms, and to view resulting trends through the platform dashboard:

> *There really is something about being able to look back over the month or the six months and see the graphs...something about that was just - I found it really encouraging.* [P6]

This provides them with relevant insights on their health, which reinforces their willingness to continue their efforts toward self-care:

> *[...] you can't sort of fool yourself in terms of your diet...It sort of reinforces your whole diet program in a sense that you know if you can stick to a certain calorie intake you are going to lose weight which is positive reinforcement.* [P5]

However, participants perceived the need to manually enter data as burdensome and experienced challenges in importing data from external trackers.

Self-Care Management

The importance of this pattern is the ability to assess whether lifestyle changes are effective. The only related platform support is a dashboard feature that patients use to visualize trends related to tracked metrics; it is also used for self-care monitoring, which enables one to see the progress over time. However, participants expressed negative perceptions related to their inability to judge if trends reflected improvements or deterioration in health status and when clinical interventions may be appropriate:

> *[...] it'd be nice to have a metric to kind of know, oh hey, you're running your life ragged and you're in danger of having a heart attack, because a lot of this is prevention.* [P10]

Several participants were also unsure if these metrics were being monitored by anyone else:

> *Does a medical doctor look at if the blood pressure readings are bad? Say, hey somebody better flag this guy and get him in. I don't know.* [P4]

Feedback Loops Among Platform Features and Self-Care

Data show that self-care abilities and behaviors impact each other through the platform's patterns of use. For example, the informational material supporting the development of cognitive and psychosocial activities (eg, educational material related to stress, recipes) generates learning, which is key to integrating effective lifestyle changes as part of self-care maintenance. Another example is that the platform's tracking features provide incentive (psychosocial abilities) to enter or import and view health-related data (self-care monitoring), which in turn provides positive reinforcement (psychosocial abilities), supports interactions with the coach (sociocultural abilities), and enables patients to integrate behavioral change into their current lifestyle (self-care maintenance). The results of these interactions are thus positive loops among platform features, abilities, and behaviors.

Conversely, challenges in using the platform and missing features were found to hinder the realization of positive loops among self-care components. For example, a participant (P16) stopped tracking their data since the results were stable and they did not see any reason to continue. Putting this decision in the context of the very limited features provided by the platform for self-care management—features that would include, among others, the ability for users to assess the effectiveness of their lifestyle changes—points to negative impacts of gaps within the Chronic Care Platform that could affect patients' health outcomes over time.

Discussion

Contributions

This study leverages Self-Care Theory [16,18] to analyze patients' patterns of use—related patient motivations, platform

XSL•FO
RenderX

features, patient activities, and resulting perceptions of value—of the Chronic Care Platform. This provides a fine-grained understanding of how platform features can support self-care abilities and behaviors. Such explanations go beyond the insights that can be gained from exploring patients' willingness or intention to use a platform using the Theory of Planned Behavior and the Health Beliefs Model [4,35], or from overall perceptual measures such as usability, ease of use, and usefulness of digital health technology [13,35].

Understanding the role of the Chronic Care Platform in facilitating self-care provides context to clinical studies that are evaluating the behavioral and clinical impacts of the program being delivered by the platform. An earlier 12-month study of the program conducted with 478 participants showed significant improvements in participants' cardiovascular disease risk—with an average reduction of 93 (19.5%) in Framingham risk score, improvements in anxiety, perceived stress, physical activity, and more [36]. The program, at that time, offered telephone-based virtual coaching, but was not being delivered using the Chronic Care Platform. A more recent study evaluating the program for postpartum women, delivered using the Chronic Care Platform, showed similar results [37]. A total of 190 women taking part in the 12-month study showed a 102 (53.7%) reduction in Framingham scores alongside a 43 (22.6%) average reduction in Lifetime Risk Scores, significant decrease in body weight, hypertension, and other clinical outcomes, increase in physical activity levels, and healthy eating.

These results contribute to the body of knowledge on the impact of digital health interventions on chronic disease self-management and the digital health technology features that are needed to successfully deliver them. For example, 2 systematic studies, one on platform-based interventions for self-management of chronic diseases [38] and the other on the efficacy of smartphone-based digital health interventions [39], arrived at similar results. That is, both reviews found that the majority of included studies showed significant improvements in behavioral outcomes such as physical activity. However, the results were mixed for clinical outcomes such as blood pressure and body weight. While several included studies showed positive outcomes (sometimes with minimal effects, however), the majority failed to demonstrate statistically significant improvements.

Better understanding of the features and role of the digital health technologies used to deliver such interventions may help to understand variability in clinical outcomes. Digital health technology includes various software and devices used to process and exchange health information on the internet, including websites, wearable devices, smartphones, and more [40]. Their most common functionalities relate to communication (eg, messaging), self-monitoring (eg, tracking of health measures), tailored self-care support (eg, personalized goals and reminders), self-care education (eg, links to reference material), care planning (eg, personal care plan), and community forums [39,40]. The Chronic Care Platform reflects the proposition for digital health platforms to integrate these functionalities into an environment that is easily accessible through various devices. A recurring observation across studies of digital health technology and platforms is that despite their obvious potential for self-managing chronic diseases, they are yet to meet patients' expectations regarding tailored, personalized support with an acceptable combination of automated and human guidance [38,39,41]. The challenge in meeting these expectations is not solely technological, in particular with the growing availability of large datasets and artificial intelligence (AI) capabilities that can be used to truly personalize care plans, prevention recommendations, alerts, dietary recommendations, and more [42,43]. The results of this study, in line with other studies of digital health technology [38,39], rather emphasize that the gap lies in using these features in a manner that creates value and generates patient engagement and retention.

The use of Self-Care Theory also helps to situate commonly identified patient needs. For example, while other studies have also found that patients express the need for personalized and relevant information when using digital solutions to support chronic disease management [4,44], they do not justify why such information is important. Self-Care Theory helps to understand that personalized information is needed to develop the cognitive abilities required by patients to care for themselves. Another example is the request for features facilitating the organization of health community activities to increase social engagement [13], which can be understood as supporting the development of sociocultural abilities in Self-Care Theory. As new technologies such as AI become integrated into self-care digital platforms [44], this understanding is key to differentiating needs that are core to chronic disease management from those that are specific to AI-based systems.

Moreover, the use of Self-Care Theory to make sense of this study's results helps to identify gaps in the Chronic Care platform's features. Figure 1 identifies key features of the platform for each self-care ability and behavior, as well as missing features that could address the challenges experienced by participants. A key gap, the lack of platform support for the development of physiological self-care abilities, can be observed across studies of digital platforms for chronic disease management. This points to a missed opportunity to support chronic disease management more comprehensively, for example, by making platforms interoperable with movement-focused sensors that could assess a patient's initial state and lead to tailored exercise recommendations. Such features would also help to improve the platform's limited support for Self-Care Management, which shows important gaps in enabling normative judgment (eg, assessing if a trend is healthy or unhealthy) and providing actionable recommendations (eg, recommendation to call one's physician). The integration of AI capabilities could address several of these gaps by leveraging sensor-generated and other patient data to better support all self-care behaviors, for example, facilitating meal tracking through image recognition (self-care monitoring) [43]; predicting and alerting the patient about low blood glucose events (self-care management) [42]; and providing tailored nutritional intervention messages to prevent future events (self-care maintenance). However, patients' reticence in using features such as AI chatbots would need to be mitigated [41,45].

XSL•FO
RenderX

Figure 1. Existing and missing platform features for self-care abilities and behaviors. Bullet points identify key features of the chronic care platform for each self-care ability and behavior. Bullet points in italics refer to missing features in the chronic care platform.

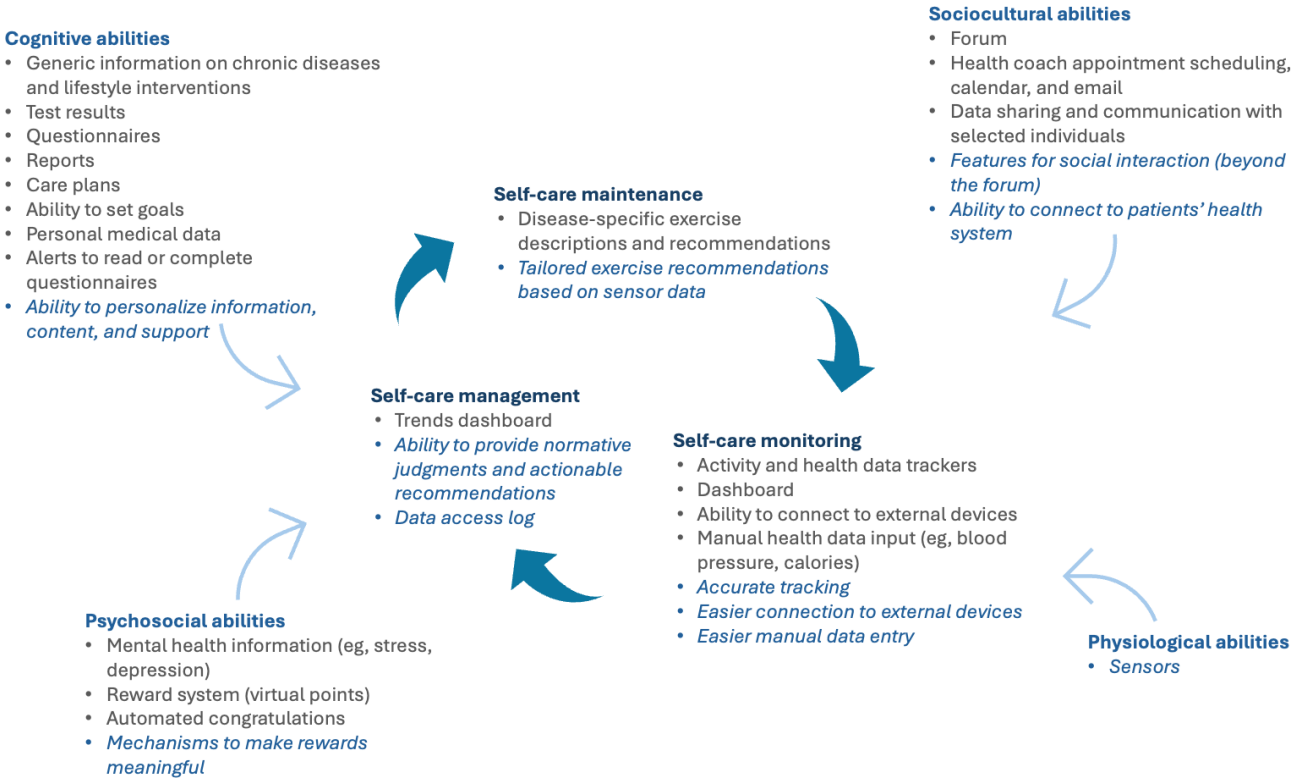

Other gaps pertain to the quality or the way in which functionalities have been designed. For example, the platform does contribute to the development of psychosocial abilities through information and feedback, but could have more impact if the feedback had real-life implications for participants. Similarly, while the ability to share data with selected individuals generated positive value perceptions, the forum only partially supported the development of sociocultural abilities. In both of these examples, the gaps are not functional—information is provided as planned, the forum functions as intended—but rather relate to the inability of available features to meet participants' expectations.

An apparent disconnect between patient expectations and what is offered through the platform was also apparent in relation to features that facilitate interactions with a health coach, which were positively evaluated by many participants. More generally, results showed that several participants sought reassurance and feedback from health care providers through the platform, which is in line with similar studies [41]. However, virtual care for chronic disease management is often seen as a way to increase efficiency, whereby scarce human resources can be leveraged to care for a greater patient population by transferring self-care responsibilities to the patient while maintaining or improving patient outcomes [46]. Providing features that enable normative judgment and personalized, actionable recommendations may partially address patients' need to be reassured in their progress. Beyond such automated support, the benefits of added human support should be evaluated against costs when creating digital health interventions for self-management of chronic conditions [38].

Nevertheless, when making design choices for digital health platforms in this context, it is important to account for the ripple effects of individual features across several self-care abilities and behaviors. For example, the support provided by a health coach improves sociocultural abilities for self-care, but these improved abilities in turn affect almost all other abilities and self-care behaviors. Conversely, the lack of mechanisms to make rewards meaningful (gap in psychosocial abilities support) could diminish a patient's motivation to track their progress (self-care monitoring behavior), hindering their ability to detect and react to changes in health status (self-care management behavior). The same functionalities are now typically integrated across digital health technology and platforms for self-care [39,40], and these functionalities are expected by patients [41]. The challenge of designing new solutions hence does not lie, in the choice or creation of discrete functionalities, but rather in the interlinking of these functionalities to trigger virtuous cycles of self-care abilities and behaviors. This can be achieved in varied ways, for example, by codesigning digital health platforms with target end users to understand if and how features support desired patterns of use and by ensuring software and hardware interoperability (eg, automated data import from self-tracking wearable devices) to facilitate seamless flow among self-care behaviors.

Limitations

Despite its contributions, the study has several limitations. The study is based on a single digital platform within a single health care setting (cardiovascular disease), which could limit the applicability of the results to other platforms. This limit was however mitigated by analyzing data using a relevant theory that provides generalizable explanations for self-care. As such, the patterns of use identified in this study and resulting

XSL•FO
RenderX

recommendations for fully realizing patient value are likely to apply to other digital platforms aiming to support self-care for patients suffering from other chronic diseases. The small number of participants (n=24) also limits the generalizability of results, while being in line with similar qualitative studies [41]. Nevertheless, future research on digital health platforms for chronic disease self-management would benefit from adopting a mixed methods approach with quantitative analysis of representative datasets. Methodological limitations related to analysis and interpretation of the data in qualitative studies were mitigated by involving 3 researchers in the coding process and by iteratively moving between interpretations and data to ensure that findings are grounded in the data as collected. The use of multiple researchers in data analysis corresponds to investigator triangulation, while the consideration of alternative theories as a source of concepts for the deductive analysis phase corresponds to theoretical triangulation [47]. Both types of triangulation used in the study aimed at increasing the trustworthiness of study results.

Conclusions

This study examined patients' patterns of use of a digital platform to understand how the platform both enabled and limited their ability to self-manage their chronic disease. Digital platforms are increasingly being used in the context of chronic disease self-management, without a full understanding of how their features may influence patients' experiences and generate desired health outcomes. We contribute to this issue by analyzing what motivates patients to use a chronic self-care platform, how they subsequently use the platform's features, and how this use generates both positive and negative values. The resulting patterns of use are interpreted as supporting the development of required self-care abilities and the realization of core self-care behaviors [16,18]. Experienced challenges lead to the identification of gaps in platform features that limit patients' ability to fully care for themselves. This study thus provides a fine-grained understanding of how a given platform can generate positive impacts and how it may be improved to fully support self-care.

Acknowledgments

The authors thank Raphaël Ménard-Grenier for his help in collecting data and in the inductive data analysis phase of this study and Manpreet Sandhu for his help in creating the Table of Contents image for this article. The authors also thank Karine Savard for helping us to better understand the chronic care program and platform investigated in this study, including by providing us with materials and performing demonstrations of the platform, and for facilitating participant recruitment. Artificial intelligence tools were not used at any stage of preparing this manuscript.

Funding

This work was supported by a Telfer School of Management Research Grant, which was secured by the first author. Work by the second author was partly supported through the "High Tech for a Sustainable Future" capacity building program of the 4TU Federation in The Netherlands.

Data Availability

The datasets generated or analyzed during the current study are available from the corresponding author on reasonable request.

Authors' Contributions

LL and MdR conceptualized the study and acquired its funding. LL developed the study's methodology, was responsible for the study's investigation, and supervised Raphaël Ménard-Grenier. MdR advised on the methods, including interview guidelines, and KAM advised on participant recruitment and data collection. LL and MdR analyzed collected data and wrote the original manuscript draft. KAM provided insights on the results and advised on the conceptualization and writing of the manuscript. All authors made substantial contributions to the intellectual content, reviewing and editing of the manuscript; they all read and approved the final manuscript.

Conflicts of Interest

None declared.

Checklist 1
Reporting items and their location based on the Standards for Reporting Qualitative Research (SRQR) and its associated checklist.
[DOCX File, 42 KB - jmir_v28i1e71875_app1.docx]

References

1. James SL, Abate D, Abate KH, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet 2018 Nov;392(10159):1789-1858. [doi: 10.1016/S0140-6736(18)32279-7]

XSL•FO
RenderX

2. Noncommunicable diseases. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases [accessed 2024-05-29]
3. Reynolds R, Dennis S, Hasan I, et al. A systematic review of chronic disease management interventions in primary care. BMC Fam Pract 2018 Jan 9;19(1):11. [doi: 10.1186/s12875-017-0692-3] [Medline: 29316889]
4. Shen H, van der Kleij R, van der Boog PJM, et al. Digital tools/eHealth to support CKD self-management: a qualitative study of perceptions, attitudes and needs of patients and health care professionals in China. Int J Med Inform 2022 Sep;165:104811. [doi: 10.1016/j.ijmedinf.2022.104811] [Medline: 35753175]
5. Randine P, Sharma A, Hartvigsen G, Johansen HD, Årsand E. Information and communication technology-based interventions for chronic diseases consultation: scoping review. Int J Med Inform 2022 Jul;163:104784. [doi: 10.1016/j.ijmedinf.2022.104784] [Medline: 35525127]
6. Abqari U, van ’t Noordende AT, Richardus JH, Isfandiari MA, Korfage IJ. Strategies to promote the use of online health applications for early detection and raising awareness of chronic diseases among members of the general public: a systematic literature review. Int J Med Inform 2022 Jun;162:104737. [doi: 10.1016/j.ijmedinf.2022.104737] [Medline: 35358894]
7. Nikayin F, De Reuver M, Itälä T. Collective action for a common service platform for independent living services. Int J Med Inform 2013 Oct;82(10):922-939. [doi: 10.1016/j.ijmedinf.2013.06.013] [Medline: 23891564]
8. Gaziano TA, Galea G, Reddy KS. Scaling up interventions for chronic disease prevention: the evidence. Lancet 2007 Dec;370(9603):1939-1946. [doi: 10.1016/S0140-6736(07)61697-3] [Medline: 18063028]
9. Kyaw TL, Ng N, Theocharaki M, Wennberg P, Sahlen KG. Cost-effectiveness of digital tools for behavior change interventions among people with chronic diseases: systematic review. Interact J Med Res 2023 Feb 16;12:e42396. [doi: 10.2196/42396] [Medline: 36795470]
10. Fan K, Zhao Y. Mobile health technology: a novel tool in chronic disease management. Intell Med 2022 Feb;2(1):41-47. [doi: 10.1016/j.imed.2021.06.003]
11. Aminuddin HB, Jiao N, Jiang Y, Hong J, Wang W. Effectiveness of smartphone-based self-management interventions on self-efficacy, self-care activities, health-related quality of life and clinical outcomes in patients with type 2 diabetes: a systematic review and meta-analysis. Int J Nurs Stud 2021 Apr;116:103286. [doi: 10.1016/j.ijnurstu.2019.02.003] [Medline: 30827741]
12. Breeman LD, Keesman M, Atsma DE, et al. A multi-stakeholder approach to eHealth development: promoting sustained healthy living among cardiovascular patients. Int J Med Inform 2021 Mar;147:104364. [doi: 10.1016/j.ijmedinf.2020.104364] [Medline: 33373949]
13. Park M, Bui LK, Jeong M, et al. ICT-based person-centered community care platform (IPC3P) to enhance shared decision-making for integrated health and social care services. Int J Med Inform 2021 Dec;156:104590. [doi: 10.1016/j.ijmedinf.2021.104590] [Medline: 34619572]
14. Taylor ML, Thomas EE, Vitangcol K, et al. Digital health experiences reported in chronic disease management: an umbrella review of qualitative studies. J Telemed Telecare 2022 Dec;28(10):705-717. [doi: 10.1177/1357633X221119620] [Medline: 36346938]
15. van de Vijver S, Hummel D, van Dijk AH, et al. Evaluation of a digital self-management platform for patients with chronic illness in primary care: qualitative study of stakeholders’ perspectives. JMIR Form Res 2022 Aug 3;6(8):e38424. [doi: 10.2196/38424] [Medline: 35921145]
16. Denyes MJ, Orem DE, Bekel G, et al. Self-care: a foundational science. Nurs Sci Q 2001 Jan;14(1):48-54. [doi: 10.1177/089431840101400113] [Medline: 11873354]
17. Sidani S, Doran DI. Development and validation of a self-care ability measure. Can J Nurs Res 2014 Mar;46(1):11-25. [doi: 10.1177/084456211404600103] [Medline: 29509462]
18. Riegel B, Jaarsma T, Strömberg A. A middle-range theory of self-care of chronic illness. ANS Adv Nurs Sci 2012;35(3):194-204. [doi: 10.1097/ANS.0b013e318261b1ba] [Medline: 22739426]
19. Liamputtong P. Qualitative inquiry. In: Handbook of Research Methods in Health Social Sciences: Springer; 2019:9-25. [doi: 10.1007/978-981-10-5251-4_53]
20. Adeoye‐Olatunde OA, Olenik NL. Research and scholarly methods: semi‐structured interviews. J Am Coll Clin Pharm 2021 Oct;4(10):1358-1367. [doi: 10.1002/jac5.1441]
21. O’Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
22. Bally ELS, Cheng D, van Grieken A, et al. Patients’ perspectives regarding digital health technology to support self-management and improve integrated stroke care: qualitative interview study. J Med Internet Res 2023 Apr 4;25:e42556. [doi: 10.2196/42556] [Medline: 37014677]
23. Saunders B, Sim J, Kingstone T, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907. [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
24. King N, Brooks J, Tabari S. Template analysis in business and management research. In: Ciesielska M, Jemielniak D, editors. Qualitative Methodologies in Organization Studies: Palgrave Macmillan; 2018:179-206. [doi: 10.1007/978-3-319-65442-3_8]

XSL•FO
RenderX

25. Cheng VWS, Piper SE, Ottavio A, Davenport TA, Hickie IB. Recommendations for designing health information technologies for mental health drawn from self-determination theory and co-design with culturally diverse populations: template analysis. J Med Internet Res 2021 Feb 10;23(2):e23502. [doi: 10.2196/23502] [Medline: 33565985]
26. Ajzen I. From intentions to actions: a theory of planned behavior. In: Kuhl J, Beckmann J, editors. Action Control From Cognition to Behavior 1985:11-39. [doi: 10.1007/978-3-642-69746-3_2]
27. Rosenstock IM. The health belief model and preventive health behavior. Health Educ Monogr 1974 Dec;2(4):354-386. [doi: 10.1177/109019817400200405]
28. Rahimi B, Nadri H, Lotfnezhad Afshar H, Timpka T. A systematic review of the technology acceptance model in health informatics. Appl Clin Inform 2018 Jul;9(3):604-634. [doi: 10.1055/s-0038-1668091] [Medline: 30112741]
29. Orem DE, Taylor SG, Renpenning KM. Nursing: Concepts of Practice, 6th edition: Mosby; 2001.
30. Sidani S. Self-care. In: Doran DM, editor. Nursing Sensitive Outcomes The State of the Science: Jones and Bartlett; 2011:79-130.
31. Richardson J, Loyola-Sanchez A, Sinclair S, et al. Self-management interventions for chronic disease: a systematic scoping review. Clin Rehabil 2014 Nov;28(11):1067-1077. [doi: 10.1177/0269215514532478] [Medline: 24784031]
32. Riegel B, Dunbar SB, Fitzsimons D, et al. Self-care research: where are we now? Where are we going? Int J Nurs Stud 2021 Apr;116:103402. [doi: 10.1016/j.ijnurstu.2019.103402] [Medline: 31630807]
33. Hartweg DL. Dorothea Orem: Self-Care Deficit Theory: Sage; 1991.
34. Burks KJ. A nursing practice model for chronic illness. Rehabil Nurs 1999;24(5):197-200. [doi: 10.1002/j.2048-7940.1999.tb02177.x] [Medline: 10754910]
35. Chen Y, Xu Q. The willingness to use mobile health and its influencing factors among elderly patients with chronic heart failure in Shanghai, China. Int J Med Inform 2022 Jun;158:104656. [doi: 10.1016/j.ijmedinf.2021.104656] [Medline: 34933173]
36. Prince SA, Reid RD, Pipe AL, McDonnell LA. An evaluation of CardioPrevent: a technology-enabled, health-behavior change program for the global reduction of cardiovascular risk. Curr Opin Cardiol 2017 Sep;32(5):580-589. [doi: 10.1097/HCO.0000000000000438] [Medline: 28639974]
37. Quansah DY, Martin N, McDonnell L, et al. Association between a technology-enabled postpartum program and cardiovascular disease risk reduction in women with prior hypertensive disorders of pregnancy. O G Open 2025 Jun;2(3):e087. [doi: 10.1097/og9.0000000000000087] [Medline: 41000084]
38. Tighe SA, Ball K, Kensing F, Kayser L, Rawstorn JC, Maddison R. Toward a digital platform for the self-management of noncommunicable disease: systematic review of platform-like interventions. J Med Internet Res 2020 Oct 28;22(10):e16774. [doi: 10.2196/16774] [Medline: 33112239]
39. Suresh S, Bhardwaj S, Rodrigues HC, et al. Smartphone-based digital health interventions: a comprehensive systematic review of efficacy for cardiovascular and cerebrovascular outcomes. J Med Syst 2025 Aug 26;49(1):109. [doi: 10.1007/s10916-025-02236-6] [Medline: 40856892]
40. Wannheden C, Åberg-Wennerholm M, Dahlberg M, et al. Digital health technologies enabling partnerships in chronic care management: scoping review. J Med Internet Res 2022 Aug 1;24(8):e38980. [doi: 10.2196/38980] [Medline: 35916720]
41. Tighe SA, Ball K, Kayser L, Kensing F, Maddison R. Qualitative study of the views of people living with cardiovascular disease, and healthcare professionals, towards the use of a digital platform to support cardiovascular disease self-management. BMJ Open 2022 Nov 1;12(11):e056768. [doi: 10.1136/bmjopen-2021-056768] [Medline: 36319055]
42. Shomali M, Mora P, Aleppo G, et al. The critical elements of digital health in diabetes and cardiometabolic care. Front Endocrinol (Lausanne) 2024;15:1469471. [doi: 10.3389/fendo.2024.1469471] [Medline: 39351525]
43. Lee YB, Kim G, Jun JE, et al. An integrated digital health care platform for diabetes management with AI-based dietary management: 48-week results from a randomized controlled trial. Diabetes Care 2023 May 1;46(5):959-966. [doi: 10.2337/dc22-1929] [Medline: 36821833]
44. Wang B, Asan O, Zhang Y. Shaping the future of chronic disease management: insights into patient needs for AI-based homecare systems. Int J Med Inform 2024 Jan;181:105301. [doi: 10.1016/j.ijmedinf.2023.105301] [Medline: 38029700]
45. Wu PF, Summers C, Panesar A, Kaura A, Zhang L. AI hesitancy and acceptability-perceptions of AI chatbots for chronic health management and long COVID support: survey study. JMIR Hum Factors 2024 Jul 23;11:e51086. [doi: 10.2196/51086] [Medline: 39045815]
46. Carter HE, Wallis S, McGowan K, et al. Economic evaluation of an integrated virtual care programme for people with chronic illness who are frequent users of health services in Australia. BMJ Open 2023 Apr 5;13(4):e066016. [doi: 10.1136/bmjopen-2022-066016] [Medline: 37019493]
47. Dootson S. An in-depth study of triangulation. J Adv Nurs 1995 Jul;22(1):183-187. [doi: 10.1046/j.1365-2648.1995.22010183.x] [Medline: 7560528]

Abbreviations

AI: artificial intelligence
SRQR: Standards for Reporting Qualitative Research

XSL•FO
RenderX

Edited by A Stone; submitted 31.Jan.2025; peer-reviewed by I Milani, O Ibikunle, VSK Kancharla; revised version received 23.Feb.2026; accepted 24.Feb.2026; published 14.May.2026.

Please cite as:
Lessard L, de Reuver M, Mullen KA
Understanding How a Digital Platform for Chronic Disease Management Can Enable and Limit Patient Self-Care: Qualitative Study
J Med Internet Res 2026;28:e71875
URL: https://www.jmir.org/2026/1/e71875
doi:10.2196/71875

© Lysanne Lessard, Mark de Reuver, Kerri-Anne Mullen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Telemedicine Adoption for Managing Chronic and Rare Diseases in Indonesia During and Beyond the COVID-19 Era: Qualitative Study

Christine Gracia Pratama[1], MBA; Rima Nurlianti[1], MSc; Cornelius Herstatt[2,3], Prof Dr oec publ; Moritz Goeldner[1,4], Prof Dr

[1]Technology and Innovation Management, Hamburg University of Technology (TUHH), Am Schwarzenberg-Campus 4, Hamburg, Germany

[2]Leuphana University of Lüneburg, Lüneburg, Germany

[3]University of Auckland, Auckland, New Zealand

[4]Working Group for Data-Driven Innovation, Hamburg University of Technology, Hamburg, Germany

Corresponding Author:
Christine Gracia Pratama, MBA
Technology and Innovation Management, Hamburg University of Technology (TUHH), Am Schwarzenberg-Campus 4, Hamburg, Germany

Abstract

Background: Telemedicine has emerged as a valuable tool for improving health care delivery, especially in low-resource and geographically isolated regions. In Indonesia, the COVID-19 pandemic highlighted the need for digital health solutions to manage chronic diseases and improve access to specialists.

Objective: This study aimed to examine the rapid adoption of telemedicine in Indonesia, focusing on its role in managing chronic and rare diseases and highlighting both its benefits and challenges to long-term viability during and post pandemic.

Methods: A qualitative study was used, following the Consolidated Criteria for Reporting Qualitative Research 32-item checklist and guided by established theories and frameworks. Recruitment followed a systematic multistage procedure suited to the context of Indonesia, an upper-middle-income country and Group of Twenty member characterized by wide regional and socioeconomic disparities. Data were collected through 15 semistructured interviews conducted between October and December 2024. The interviews were divided into 2 phases: first with physicians from various medical specialties and 6 out of 7 regions within Indonesia, and then with patients diagnosed with autoimmune diseases. All interviews were audio-recorded, transcribed verbatim, and analyzed inductively to identify recurrent themes and subthemes.

Results: Before the COVID-19 pandemic, telemedicine adoption in Indonesia was limited, with many physicians recalling its nonexistence before 2015. The pandemic accelerated the adoption of virtual platforms, which became essential for patient consultation, follow-ups, and medication management. Though its use decreased after the pandemic, telemedicine remains valuable, particularly for chronic and rare disease management in remote areas with limited access to specialists. At least three overarching themes were identified: (1) acceleration and normalization of telemedicine (the pandemic catalyzed the integration of virtual consultations into routine care, transforming initial skepticism into broad acceptance); (2) accessibility and cost efficiency (telemedicine reduced financial and travel burdens, especially for patients in remote areas, with several reporting cost savings from hybrid consultation models); and (3) limitations of virtual care (physicians highlighted constraints related to the absence of physical examination).

Conclusions: The findings reveal that telemedicine has become a valuable tool in Indonesia for managing chronic and rare diseases, particularly for follow-up care and medical specialist access across diverse geographical areas. While telemedicine has improved health care accessibility and demonstrated significant cost benefits by reducing transportation and consultation costs, challenges such as limited infrastructure and physician burnout remain. Long-term success will depend on the development of sustainable regulatory frameworks, continued investment in digital infrastructure, and a focus on optimizing cost-effectiveness in health care delivery.

(J Med Internet Res 2026;28:e83462) doi:10.2196/83462

KEYWORDS

digital health; telemedicine; chronic disease; rare disease; COVID-19; Indonesia

XSL•FO
RenderX

Introduction

Background

Digital technologies offer new ways to improve health care coverage and quality, particularly in settings with limited resources and widespread geographic dispersion. Integrating digital technologies has made care more efficient, accessible, and patient-centered. The global adoption of digital health care, such as telemedicine, was accelerated by the COVID-19 pandemic [1-3]. Telemedicine has gained prominence as a crucial component of health care delivery. It was strongly advocated during the pandemic for its potential to address immediate health care needs while reducing the risk associated with in-person visits [4-6]. This shift was not just a temporary solution; it marked the beginning of a permanent transformation in health care delivery models in both the Global North and the Global South. Importantly, telemedicine also holds potential to support progress toward sustainable development goal (SDG) 3 (good health and well-being) and goal 10 (reducing inequalities) by enabling more equitable access to health care services across diverse socioeconomic and geographic contexts.

With a population exceeding 270 million and more than 17,500 islands, Indonesia faces significant disparities in health care access due to its fast-growing population, urbanization, and huge distances between health care providers and patients. This intensified during the pandemic, particularly among individuals with chronic, noncommunicable diseases (NCDs) [7]. Therefore, digital health solutions have the potential to ensure equitable and continuous care in this region. Early evidence suggests that telemedicine could reduce transportation costs by 4% and shorten patient travel times without compromising the level of care [8]. However, amidst the effectiveness of telemedicine, barriers such as digital literacy, educational level, age, rural residence, and internet access continue to affect its adoption [9-11].

Evolution, Definitions, and Current Relevance

The digital transformation of health care systems has the potential to greatly improve quality while reducing costs [12,13]. Telemedicine, which originated in the early 20th century with the use of telephone lines to transmit medical data, has evolved significantly and become a vital component of modern health care delivery. Early applications, such as remotely transmitting electrocardiograms and X-rays, laid the groundwork for telemedicine's role in improving health care access, particularly in emergencies and rural areas [14].

The World Health Organization defines telemedicine as the use of information and communication technology to provide health care remotely, particularly in regions with limited access to medical services [12]. Telemedicine is highly patient-centered because it is convenient and caters to patients' digital preferences. This makes telemedicine an essential component of patient care management in areas with limited traditional options [15]. Telemedicine enables communication between health care providers and patients through technology [16]. It encompasses both synchronous and asynchronous methods of delivering health care services remotely and is provided through various digital platforms, including secure messaging, mobile apps, and video conferencing [16-19]. Telemedicine is a broad health care service that uses telecommunication and electronic technologies to support remote consultations, monitoring, nursing, and rehabilitation between patients and health care providers [4,20].

Telemedicine Adoption in Asia-Pacific and Organization for Economic Cooperation and Development Countries

In the Asia-Pacific region, telemedicine experienced substantial growth during the COVID-19 pandemic, nearly doubling between 2019 and 2021. Australia recorded a ninefold increase in usage during this period [20,21]. However, postpandemic trends show a stabilization or decline in some countries. For instance, between 2021 and 2023, Singapore and Australia each saw a 9 percentage points rise in telemedicine adoption (from 34% to 43% and 45% to 54%, respectively), while Indonesia's adoption increased by 7 percentage points (from 51% to 58%) [22,23]. Conversely, from the same study, India experienced a 4 percentage points decline (from 59% to 55%), reflecting regional disparities in digital infrastructure, regulatory support, and user engagement.

In Organization for Economic Cooperation and Development countries, telemedicine use was limited before the pandemic, with an average of only 0.6 teleconsultations per patient annually in 2019 [24]. Sweden led in adoption, with 47% of Swedish adults reporting in early 2021 that they had at least 1 teleconsultation since the start of the pandemic [25,26]. France also saw a dramatic spike, from 136,882 consultations in 2019 to 4.5 million in April 2020 alone, driven by policy changes such as full reimbursement [27]. Germany's growth was short-lived; after a surge from 3000 consultations in 2019 to 2.7 million in 2020, usage declined by 40% in 2022 [28,29]. Physician adoption mirrored this trend: in Sweden, 50% of physicians conducted 25% - 75% of consultations online weekly, compared to just 6% in France and 14 % in Germany [30]. These patterns underscore the need for adaptive digital health policies tailored to each country's health care context [31].

Digital Health Adoption and COVID-19

The COVID-19 pandemic catalyzed an unprecedented acceleration in digital adoption across health care systems globally, particularly in the area of telemedicine. Mobile apps offering teleconsultation, contact tracing, symptom monitoring, and remote patient management became essential tools for maintaining health care continuity during lockdowns. Regional distinctions shaped the design and use of these platforms: for example, Asian countries emphasized user accessibility and functional simplicity to promote widespread adoption, while North American and European apps primarily focused on disseminating health information with limited interactivity [32]. Additionally, global differences in regulatory environments, such as Germany's stringent data privacy laws compared to India's more rapid, utilitarian deployment of contact-tracing apps, reflect diverse policy priorities during the pandemic [33]. These variances underscore the importance of localized strategies for telehealth implementation and regulation, particularly as countries transition into the postpandemic phase.

XSL•FO
RenderX

Post-COVID-19 telemedicine is increasingly recognized as a foundational element of modern health care, extending far beyond its initial role as an emergency response. The World Health Organization emphasized its potential not only for acute care but also as a sustainable solution for managing long-term health conditions. Chronic diseases, which demand ongoing monitoring and personalized care, have emerged as a central focus for telehealth apps [5,6]. This is particularly relevant for rare diseases, which often require specialist consultations and coordinated care that may not be readily available in all geographic regions. The decentralized model enabled by telemedicine can mitigate geographic and systemic barriers, facilitating access to specialist expertise for rare disease patients and improving outcomes through timely interventions and continuity of care [34,35].

Indonesia exemplifies the strategic integration of telemedicine into national health planning in the postpandemic context. With a population exceeding 270 million distributed across thousands of islands, the government has prioritized digital health to overcome logistical challenges and expand health care accessibility [36]. The Ministry of Health's "Blueprint of Digital Transformation Strategy 2024" developed in collaboration with the United Nations Development Program outlines a comprehensive framework for digital health system development [37,38]. This includes infrastructure to support the management of chronic and rare diseases, particularly in view of the increasing prevalence of NCDs in the country [7,39]. The blueprint reflects a broader regional trend toward embedding telehealth within long-term health care policy, signaling a shift from reactive digital health adoption to proactive, system-wide transformation. Such strategies are vital to ensuring equitable care for patients with chronic and rare diseases in both urban and remote settings.

Use of Telemedicine for Chronic and Rare Disease Management

Chronic and rare disease management requires a continuous, coordinated, and patient-centered approach, focusing on monitoring, medication adherence, education, and preventive care to improve outcomes and quality of life [40,41]. Chronic NCD diseases such as diabetes, cardiovascular conditions, and autoimmune disorders—including rare diseases like systemic lupus erythematosus and Sjögren's disease—demand sustained health care interventions due to their complexity and long-term nature [42,43]. Telemedicine has emerged as a critical enabler of chronic care, particularly during and after the COVID-19 pandemic, by bridging geographic gaps, reducing travel costs, and supporting remote management for patients in underserved areas [32,44,45]. In Indonesia, for example, the Program Pengelolaan Penyakit Kronis under Badan Pengelola Jaminan Sosial Kesehatan (BPJS; National Health Insurance Agency) illustrates the integration of telemedicine in chronic disease management, offering education, routine monitoring, and preventive care for diabetes and hypertension [24,46].

The post-pandemic era has highlighted telemedicine's broader clinical use in managing both chronic and rare diseases. Studies demonstrate its efficacy in improving medication adherence and reducing psychological distress among patients with conditions such as rheumatoid arthritis, while also enhancing satisfaction in postoperative and psychiatric care when delivered via video consultations [47-50]. Although many autoimmune patients still prefer initial face-to-face consultations, hybrid models—where video is used for follow-ups—have been well received, with 84% reporting video consultations as effective when paired with subsequent in-person visits [43]. Beyond synchronous video, asynchronous tools such as SMS have proven particularly effective among younger patients, promoting continuous engagement through reminders and 2-way communication [51]. To ensure quality and patient safety, frameworks like the Asia-Pacific League of Associations for Rheumatology telemedicine guidelines recommend time-limited follow-ups, prioritization through preconsultation screening, and strict adherence to data privacy standards [52]. Collectively, these strategies underscore telemedicine's critical role in enhancing access, continuity, and personalization of care for patients with complex and chronic conditions.

Cost-Benefit Analysis

Telemedicine has emerged as a cost-effective solution for improving health care access and quality, particularly in regions facing shortages of medical professionals. In Association of Southeast Asian Nation countries, including Indonesia, the scarcity of physicians in rural areas exacerbates health care challenges. With only 7.5 medical practitioners per 10,000 population, Indonesia lags behind other nations like Malaysia and Singapore, where the availability of physicians is much higher. Telemedicine closes this gap by offering remote consultations, which reduces the need for patients to travel and minimizes health care costs for patients and providers alike [49]. The increasing adoption of telemedicine has proven to reduce consultation, travel, and time-related expenses, while also enhancing the overall quality of care.

Studies from Europe also highlight the cost-effectiveness of telemedicine. A comprehensive analysis found that telemedicine was cost-effective in 73.3% of cases, with reduced consultation costs and travel expenses being significant contributors to savings [53]. In Sweden, for example, digital consultations cost 41.5% less than in-person visits [54]. Similarly, a remote monitoring program for diabetic patients in France demonstrated cost savings, with the remote group spending €1334 (US $1513, December 2021 exchange rate) less compared to the control group over 1 year [55]. These findings underscore the potential of telemedicine to reduce both health care costs and improve patient outcomes, particularly in systems where teleconsultations are integrated into public health systems with reimbursement options.

Telemedicine also shows considerable promise in reducing nonmedical expenses, such as transportation costs and lost time from work or school. Studies from Bangladesh and pediatric rheumatology research highlight the substantial savings telemedicine provides by cutting travel time by 56% and associated costs by 94% [43,56]. In Indonesia, the cost of teleconsultations varies depending on the platform, with independent companies offering more affordable options compared to hospital-integrated services. However, telemedicine remains largely inaccessible through public health insurance,

XSL•FO
RenderX

as it is only reimbursed via private insurance or self-payment [57]. Despite these limitations, text messaging and other telemedicine tools continue to be cost-effective solutions for chronic disease management, particularly in underserved areas with limited health care access [58,59]. This qualitative study aims to examine how access to telemedicine supports care for managing chronic and rare diseases in Indonesia, drawing on perspectives from both patients and health care professionals in the post-pandemic context.

Methods

Ethical Considerations

Informed consent was obtained from all participants prior to data collection. Data were collected, anonymized, and stored securely in line with international ethical standards. No personal or social information was collected, and all data was fully anonymized to ensure privacy and confidentiality. Participants were not offered any compensation for taking part in the study.

This study was conducted in accordance with the principles of the Declaration of Helsinki. Ethical approval was obtained from the German Association for Experimental Economic Research Institutional Review Board (certificate number syBSVy9b). No identifying images or personal or clinical details of participants are included.

Research Design

This study adopts a qualitative research design following the Consolidated Criteria for Reporting Qualitative Research (COREQ) 32-item checklist as provided as Checklist 1 [60,61]. The topic guides were developed by RN and CGP in accordance with the COREQ 32-item checklist to ensure comprehensive coverage of key methodological domains. Guide development was informed by prior literature on telemedicine adoption and qualitative interviewing and subsequently reviewed by MG to establish content validity. A deductive analytic approach was applied, drawing on established theories and frameworks to inform the research questions and interview guide development.

The study comprises two primary participant groups: physicians and patients. Semistructured interviews were chosen as the main data collection method due to their suitability for eliciting in-depth responses on telemedicine experiences and challenges [62,63]. The interviews were conducted in 2 phases. First, physicians from different regions and medical specialties were interviewed. Next, case studies of patients diagnosed with autoimmune diseases were examined.

Data Collection

Recruitment

Recruitment followed a systematic multistage procedure suited to the context of Indonesia, an upper-middle-income country and Group of Twenty member characterized by wide regional and socioeconomic disparities [64]. For physicians, the inclusion criteria required active medical practice in Indonesia, prior experience using or familiarity with telemedicine in managing chronic or rare diseases, and willingness to participate in an online interview. For patients, inclusion criteria included having a clinically confirmed chronic or rare autoimmune condition and experience with telemedicine services. A social media-based strategy was used, as it has proven effective in engaging participants and achieving broader outreach across socioeconomic categories [65-68]. Physicians were approached through professional platforms (LinkedIn and ResearchGate), email, and Instagram, while patients were recruited through content-sharing platforms (Instagram and TikTok) and WhatsApp community groups. This multiplatform strategy enabled access to both health care professionals and hard-to-reach patient populations [66,68,69]. Previous studies underscored notable gender disparities, revealing a predominance of female patients, which corresponds with the higher incidence of autoimmune diseases observed in women [69].

Furthermore, standardized invitations outlining the study, eligibility criteria, and confidentiality were shared, thus participation occurred according to respondents' availability. Digital consent was obtained prior to interviews. Although social media enabled broad outreach, engagement from marginalized groups remained uneven [65].

For this study, Indonesia was categorized into Western and Eastern regions. Western Indonesia comprises Sumatra, Java, and Kalimantan, while Eastern Indonesia includes Nusa Tenggara, Sulawesi, Maluku, and Papua. Participants were drawn from 6 of Indonesia's 7 major regions, including Sumatra, Java, Kalimantan, Nusa Tenggara, Sulawesi, and Papua. Recruitment from the Maluku Islands proved challenging due to the absence of referral connections with physicians or patients with autoimmune conditions, as well as the region's small population.

National telecommunications statistics indicate substantial regional disparities in digital connectivity. Western Indonesia generally demonstrates higher internet access, with an average of 80% of the population, compared to around 40% in the Eastern regions [70]. The lowest internet penetration was observed in parts of Papua, where access declined to approximately 6% in 2024 [70]. Similar results were found for cellular signal coverage across 7 Indonesian regions [70].

Interviews With Physicians and Patients

A total of 35 physicians were approached, resulting in 15 semistructured interviews conducted between August and November 2024. The participants, representing 6 medical specialties, were chosen for their experience with telemedicine, research involvement, or public health advocacy. For physicians, the inclusion criteria required active medical practice in Indonesia, prior experience using or familiarity with telemedicine in managing chronic or rare diseases, and willingness to participate in an online interview. For patients, inclusion criteria included having a clinically confirmed chronic or rare autoimmune condition and experience with telemedicine services. Each interview, conducted in Bahasa Indonesia via Zoom (Zoom Communications, Inc) or WhatsApp (Meta Platforms, Inc), lasted 25 to 63 minutes (mean 48, SD 13.4), following a 5-section interview guide with 27 questions as shown in Appendix S1 (Multimedia Appendix 1). In total, the interviews yielded 743 minutes of recordings and 163 pages of transcripts. The interviews were primarily conducted by RN,

XSL•FO
RenderX

with several sessions co-led or supported by CGP. Physicians were recruited through social media, professional platforms, and referrals, with 1 session including a passive observer (a community leader).

In the second phase, 15 patients diagnosed with autoimmune diseases were contacted, leading to 9 interviews conducted between November 15 and December 15, 2024. All interviews were conducted by RN. Recruitment took place through patient communities on Instagram (Meta Platforms, Inc) and WhatsApp (Meta Platforms, Inc), with consent from community leaders. Each interview had a typical duration of 26 to 70 minutes (mean 54, SD 16.1), guided by a 6-section, 34-question format, and transcribed verbatim. Appendix S3 (Multimedia Appendix 1) contains the interview guide, and a detailed interview summary is attached in Appendix S4 (Multimedia Appendix 1). The total output was 539 minutes of recordings and 197 transcript pages, with 1 session also involving a nonparticipant community leader. Table 1 summarizes key details of both phases.

Table . Overview of interview characteristics and participant details.

Category	Physicians	Patients
Interview period	October 10-November 10, 2024	November 15-December 15, 2024
Participants reached	35 physicians	15 patients
Number of participants	15 physicians	9 patients
Recruitment method	Direct outreach via ResearchGate, LinkedIn, Instagram, and referral through mutual connections	Community posts and broadcasts on Instagram and WhatsApp
Language	Bahasa Indonesia	Bahasa Indonesia
Interview duration	25-63 min (mean 48, SD 13.4)	26-70 min (mean 54, SD 16.1)
Interview format	Semistructured (5 sections and 27 questions)	Semistructured (6 sections and 34 questions)
Total recorded time	743 min	539 min
Transcript length	163 pages	197 pages

Data Analysis

Thematic analysis is used to systematically identify, analyze, and interpret patterns within qualitative data [71]. Transcripts were initially reviewed for accuracy and completeness. Zoom (Zoom Communications, Inc) interviews were transcribed using built-in features, while WhatsApp (Meta Platform) recordings were processed using QuickTime Player (Apple Inc) and TurboScribe (Leif Erikson Ventures, LLC). All transcripts were manually checked and then translated into English using ChatGPT 4.0 (OpenAI) and DeepL (DeepL SE) to ensure consistency and accessibility for non-Indonesian—speaking readers. All translated transcripts were then reviewed again manually by 2 of the authors, both of whom are native speakers and 1 of whom has medical knowledge. To maintain confidentiality and reduce participant burden, transcripts were not returned to participants for verification. Identifying details were anonymized throughout. The data analysis process is displayed in Figure 1.

Figure 1. Steps involved in the qualitative data analysis process.

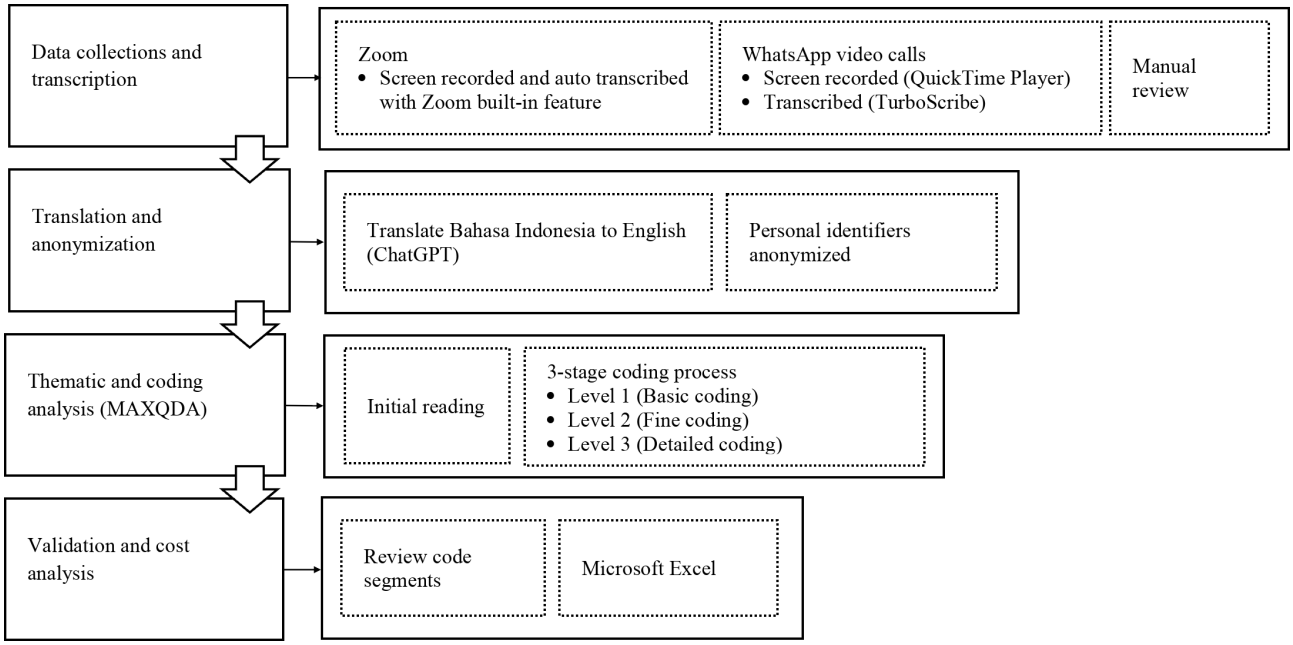

XSL•FO
RenderX

Data were analyzed using MAXQDA (VERBI Software GmbH), following a structured, 3-level coding process [72]. First-level coding identified broad themes that served as the foundation for analysis. Second-level coding, or fine coding, further refined these categories into specific subcategories for deeper analysis of telemedicine experiences. Third-level coding introduced additional granularity, breaking down subcategories into more specific classifications to capture nuanced patterns within the data. Coding was primarily conducted by RN, while CGP and MG independently reviewed portions of the transcripts to validate emerging codes and themes. Theme validation extended beyond partial co-coding through iterative comparison of the final themes with the original transcripts to ensure accurate representation of participants' perspectives. Interpretive discrepancies were discussed among the authors and resolved through consensus. In addition, the researchers also held regular discussions throughout the coding process and participated in debriefing sessions with MG to further strengthen analytical rigor. Data saturation was determined by RN during the coding process and subsequently confirmed through joint discussions with CGP and MG. Saturation was considered achieved once participants from all major Indonesian regions were represented and recurring patterns consistently emerged across interviews. This determination was further supported by verification through field notes and preliminary coding reviews, with no additional first-order concepts or second-order themes identified [73].

In addition to thematic analysis, cost-related data mentioned in patient interviews (eg, transportation, consultation, and medication costs) were extracted and compiled using Microsoft Excel (Microsoft Corporation) to support an exploratory cost-benefit dimension.

Interviews with physicians were thematically coded across 6 domains: general experience, telemedicine usage, perceived benefits, challenges, health care accessibility, and future outlook. Physicians highlighted telemedicine's effectiveness in monitoring stable patients, especially in remote areas, citing improved accessibility and cost-efficiency. However, they also noted challenges such as diagnostic limitations, poor internet connectivity, legal concerns, and disparities in patient digital literacy. Telemedicine adoption varied by region, case complexity, and health care setting, with usage patterns influenced by the COVID-19 pandemic and shifting postpandemic practices. The coding structure is visualized in Appendix S5 (Multimedia Appendix 1).

Patient interviews were classified according to demographics, medical history, telemedicine usage, health care access and costs, challenges, and overall perceptions. Participants reported diverse experiences shaped by illness severity, hospital class, payment methods (eg, BPJS and private insurance), and regional access to specialists. Many appreciated the reduced travel and financial burden telemedicine offered, while others preferred in-person care for complex diagnostics. Barriers included infrastructure limitations, insurance gaps, and varying levels of digital literacy. Overall, patients valued the flexibility of telemedicine but saw it as a complementary tool rather than a full replacement for face-to-face consultations. Cost-related findings were supported by comparative analysis using MAXQDA (VERBI Software GmbH) and Microsoft Excel (Microsoft Corporation) to highlight reductions in nonmedical expenses. Appendix S6 (Multimedia Appendix 1) presents the coding analysis level of the patient.

Results

Geographic Distribution of Participants

Figure 2 illustrates the geographic distribution of study participants across Indonesia. Physicians and patients are represented according to their respective regions. This highlights the coverage across 6 major regions and shows regional representation.

Figure 2. Distribution of physician and patient interviews across Indonesia.

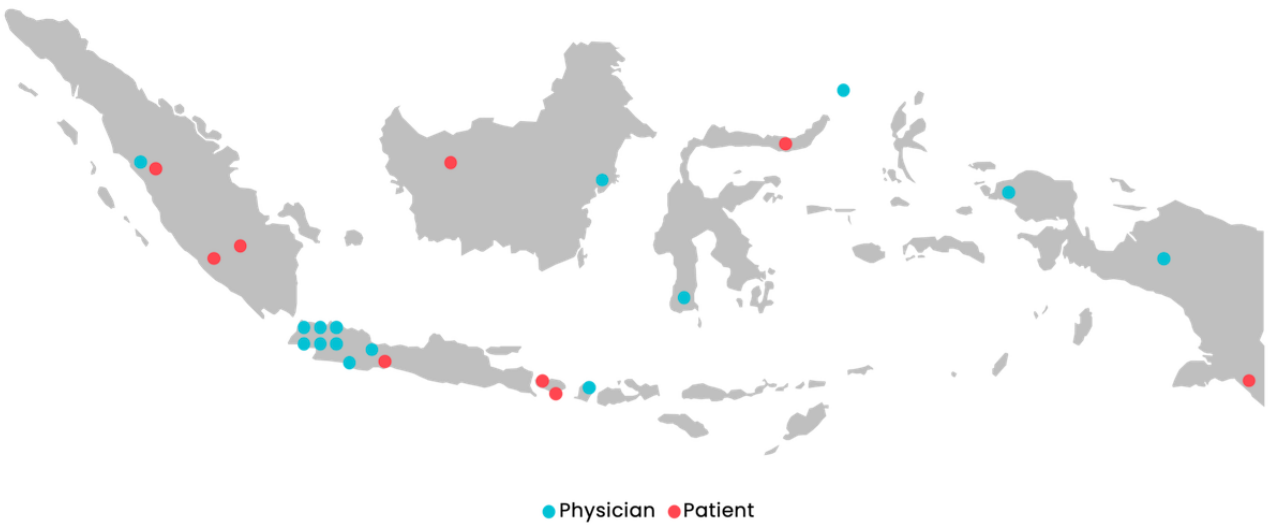

XSL•FO
RenderX

Physicians' Perspectives

Overview

The study involved 15 physicians from 6 regions in Indonesia, as illustrated in Figure 2, representing a range of specialties: general practice (n=5), immunology (n=4), rheumatology (n=3), plastic surgery (n=1), pediatrics (n=1), and ophthalmology (n=1). Participants' clinical experience varied, with 2 participants having over 20 years of practice, 6 having 16 - 20 years, 2 having 6 - 10 years, and 5 having between 1 - 5 years of experience.

Additionally, the physicians practiced across all hospital classifications, as defined in Article 340/MENKES/PER/III/2010, which categorizes hospitals based on service capabilities, infrastructure, and available resources [74-76]. The classification system categorizes hospitals into four types: class A, B, C, and D. Class A hospitals, primarily located in major cities, serve as top referral centers with extensive facilities and a wide range of medical and subspecialty services. Class B hospitals, mostly found in regional areas, provide at least 11 specialist and limited subspecialist services. Class C hospitals focus on essential specialist care and are equipped with a minimum number of medical specialties. Class D hospitals, commonly referred to as Community Health Centers (Pusat Kesehatan Masyarakat), offer basic medical services with limited specialized care.

Evolution of Telemedicine Adoption

Before the COVID-19 pandemic, telemedicine adoption in Indonesia was sporadic and limited. Among 15 physicians surveyed, 12 reported using telemedicine. Additionally, 1 physician recalled that "telemedicine didn't exist yet" around 2015, reflecting the nascent state of digital health at the time.

The pandemic became a turning point. A total of 11 out of 15 physicians began using virtual platforms such as WhatsApp (Meta Platforms, Inc), Zoom (Zoom Communications, Inc), standalone health tech apps like HaloDoc (PT Media Dokter Investama) and Alodokter (PT Alodokter Teknologi Indonesia), as well as hospital-based telemedicine systems [65-68]. During this period, telemedicine was particularly valuable for triaging patients, conducting follow-ups, and managing medications. The need to reduce physical contact due to infection risks made remote consultations not just practical but essential. Many hospitals and clinics swiftly adopted telemedicine services, though often with basic tools or improvised infrastructure.

As COVID-19 cases declined, physicians observed a reduction in telemedicine usage as patients gradually returned to in-person visits. Still, telemedicine continues to serve an important role, especially for follow-up care and for managing chronic or rare conditions. This transition is reflected in 1 physician's experience:

> *At one point, I was handling up to 23 patients a day, but during a single Zoom call, there could be five family members present. Afterward, I would continue treating the rest of my patients. That was mostly in 2019 - 2020, but then the numbers started declining. Post-COVID, it's been fewer, I primarily manage autoimmune patients from remote areas.* [D13]

While full replacement of in-person care remains unlikely, the pandemic has established telemedicine as a lasting complement within hybrid care models. Half of the physicians viewed it as unsuitable for initial diagnoses, but they now use it for monitoring, education, second opinions, and psychological support. This approach is particularly effective for extending care and reducing financial burdens to patients in remote locations. In this context, pediatric cardiologists in Jakarta conduct monthly follow-ups via video consultation with patients in Aceh, located in the westernmost part of Indonesia. Given the long waiting times and the need for continuous monitoring, a process known as the surgical conference is used to assess and prioritize patient conditions. One physician shared:

> *They are usually asked to have monthly follow-ups via the online clinic. They also check the status of the surgical conference queue. So, it just keeps moving along.* [D4]

Challenges

In total, 14 out of 15 physicians identified the lack of physical examination as a key limitation in telemedicine, as noted:

> *However, if I've never met a patient in person, never done a physical examination, or never had direct contact, it's challenging to diagnose complex diseases [...] for diagnoses that require physical examinations, telemedicine can't fully replace in-person visits.* [D7]

Telemedicine has also increased workloads. Two physicians reported burnout due to additional consultation demands. A physician explained:

> *Right now, telemedicine is just a replacement for face-to-face consultations. It does not reduce the workload; it adds to it. [...] So, telemedicine adds extra time slots. To reduce burnout, we need more doctors and better networking.* [D7]

Legal uncertainty adds another layer of concern. Five physicians questioned data ownership when using third-party apps, emphasizing "[a]nother issue is significant patient privacy concerns regarding data security and encryption measures" (D9).

Future Directions

Telemedicine is advancing toward greater accessibility, efficiency, and diagnostic accuracy. A key innovation is the creation of specialized hubs linking medical teams across hospitals. As explained by a physician, this hub-and-spoke model is underway at his or her hospital, beginning with neurosurgery:

> *For example, at our hospital, what we plan for the future is to have this. Currently, our neurosurgery team is already working on it. The main hub for the neurosurgery team is at our hospital. But then, in every hospital branch, there is a hub that becomes part of the neurosurgery network.* [D7]

XSL•FO
RenderX

Another promising development is artificial intelligence (AI)–assisted diagnostics. A physician highlighted a project targeting diabetic retinopathy:

> *We are currently developing AI tools for diabetic retinopathy [...] The AI analyzes the photos to categorize them automatically [...] In remote regions, access to equipment is still a challenge.* [D8]

Additionally, a physician cited virtual reality (VR) integration in Japan to support remote assessments:

> *As currently being implemented at the University of Nagasaki, Japan [...] Patients visit a clinic on a remote island, and physicians remotely assess their condition using VR.* [D15]

Patients' Perspectives

Participant Characteristics

The data is based on interviews with 9 patients diagnosed with various autoimmune conditions, including rheumatoid arthritis, systemic lupus erythematosus, Sjögren's disease, psoriasis, antiphospholipid syndrome, Graves' disease, and scleroderma, with time since diagnosis ranging from 2 to 17 years. Participants, aged between 26 and 59 years, came from diverse professional backgrounds and lived in both urban and rural areas. Seven were covered by Indonesia's national health insurance (BPJS), while 2 financed their health care independently.

Telemedicine Adoption Shaped by Health Care Access

Patients with autoimmune diseases in Indonesia face significant disparities in health care access. Accurate diagnosis typically requires referral to class A hospitals located in major cities, due to the scarcity of subspecialists in rural or outer island regions. The average diagnosis delay exceeded 1 year. One patient was diagnosed in Malaysia due to the lack of domestic expertise during the early 2000s, as local specialists were unavailable at the time.

While some patients now receive initial evaluations at class B hospitals, advanced care still necessitates travel to central facilities. A patient from eastern Indonesia reported a 400 km journey, taking 10 hours by bus through mountainous terrain, costing up to IDR 10 million (€580) per visit.

> *The eight to ten hours by land is really far, especially with the poor infrastructure with mountain roads and winding paths [...], so my mother called instead. There is no consultation fee; only the medicine has a cost, but it is expensive.* [P6]

Due to the physical and financial burden, the patient was later transferred to a nearer class B hospital, albeit with limited specialist support.

Telemedicine has emerged as a critical solution for bridging these access gaps. Among the 9 patients interviewed, those living in urban or semiurban areas within 5 km of a hospital preferred in-person care, citing convenience and minimal travel time.

> *The distance is probably one kilometer. You can walk or ride a motorbike. It usually takes about 5-10 minutes by motorbike.* [P9]

In contrast, 6 patients residing over 50 km from the nearest hospital relied heavily on telemedicine for follow-ups and flare-ups. Poor infrastructure meant that a 50 km trip could take up to 3 hours, pushing patients to communicate directly with their physicians via WhatsApp text and video calls. This is an informal, yet reliable channel in the absence of more robust telehealth systems.

> *I once had a check-up, and everything was fine, with no complaints for a month. Then, just a week after the check-up, I suddenly experienced pain in my hand, swelling, and an unusual toothache. When I consulted a dentist, they found no inflammation, but the pain persisted. [...] I messaged the doctor directly and sent a photo of the dental examination results.* [P2]

Notably, 3 patients with multiple chronic conditions reported using WhatsApp video calls during emergencies to replace in-person visits. They maintained direct contact with their physicians with whom they had been in care for more than 5 years.

> *I was able to video call my personal doctor directly via WhatsApp while I was in isolation due to COVID-19 to discuss my symptoms along with my autoimmune conditions. The consultation lasted 10 - 15 minutes and was free of charge.* [P5]

Telemedicine was also used for prescription renewals and noncritical follow-ups. Two patients receiving care at specialized autoimmune clinics reported being notified via WhatsApp when their condition was stable, skipping the usual in-person appointment. Interestingly, none of the patients used commercial health-tech apps (eg, Halodoc), preferring direct contact with trusted physicians. As 1 patient explained:

> *I already know the doctor, so I just message them directly.* [P2]

In remote regions where specialist access remains sparse and health literacy is rising, patients have begun turning to social media for medical advice. This noted the role of social media in improving access to medical information and direct physician interaction [66,67].

> *I asked him on Instagram or TikTok, "Doctor, is there an alternative to Myfortic because it's hard to find?"* [P1]

Together, these findings highlight how geography, infrastructure, and personal networks shape health care access and telemedicine use. Patients in remote areas, particularly those with long-term physician relationships, increasingly rely on digital tools to overcome systemic barriers to consistent and affordable care.

Cost-Effectiveness

Telemedicine significantly reduced financial burdens for patients requiring frequent monitoring. Of the 9 participants, 6 reported measurable cost savings after switching from monthly in-person visits to a hybrid model with 1 in-person visit and 5 virtual consultations over 6 months.

XSL•FO
RenderX

At the time of the study, the exchange rate was approximately €1=17,400 IDR (US $1=IDR 15,500 [2024 average]) [77,78]. The average cost per in-person visit was IDR 2 million (€115), with transportation and accommodation expenses increasing the total cost substantially for remote patients. In comparison, teleconsultations conducted via WhatsApp were free. Table 2 below summarizes the financial impact, as reported by multiple participants in this study.

These results reflect a cost reduction of 65%, from IDR 42.0 million (€2484) to IDR 16.5 million (€869), corresponding to an absolute savings of IDR 25.5 million (€1615), without compromising care quality for stable patients. However, public insurance (BPJS) does not fully cover all lab tests or medications. Out-of-pocket costs varied from IDR 200,000 (€12) to IDR 7 million (€405) per month, depending on the disease progression. One patient noted paying IDR 7 million monthly during the first year post diagnosis due to uncovered medications.

Hence, telemedicine offers significant financial and logistical relief for patients in eastern Indonesia, reducing the need for costly and exhausting multileg hospital trips. Thus, telemedicine is highly valued, not only for its economic advantages but also for its emotional and physical relief.

Table . Exemplary comparison of cost savings based on patient interviews.

Category	In-person visits	Hybrid model	Cost savings
Visit frequency (6 mo)	6 in-person visits (monthly)	1 in-person visit+5 telemedicine consultations	__[a]
Cost per in-person visit (IDR)[b]	2,000,000	2,000,000	0
Total in-person visit cost (IDR)	12,000,000 (IDR 2,000,000x6)	2,000,000 (1 visit)	10,000,000
Telemedicine consultation cost (IDR)	0	0 (via WhatsApp)[c]	0
Medication costs (IDR)	9,000,000 (IDR 1,500,000 x 6)	9,000,000 (IDR 1,500,000 x 6)	0
Transportation costs (IDR)	12,000,000 (round-trip for all visits)	2,000,000 (1 trip)	10,000,000
Additional costs (accommodation, meals) (IDR)	9,000,000 (overnight stays, meals)	1,500,000	7,500,000
Total cost (IDR)	42,000,000	16,500,000	25,500,000 saved

[a]Not applicable.

[b]All costs are reported in Indonesian Rupiah (IDR). Conversion to US dollars (US $) was performed using an exchange rate of US $1=IDR 15,500 (2024 average).

[c]Free, under specific conditions (eg, a close physician-patient relationship, typically established for at least 5 years since the initial diagnosis).

Discussion

Overview

This study examines the impact of telemedicine adoption on the management of chronic and rare diseases in Indonesia. Telemedicine has demonstrated significant potential in enhancing chronic illness care and has been recognized as a critical component of Indonesia's health care response during the COVID-19 pandemic [79-81]. By exploring the perspectives of both patients and health care professionals, this research offers valuable insights into the effectiveness, challenges, and opportunities of telemedicine for managing chronic and rare diseases in the postpandemic era.

Principal Findings and Alignment With Prior Literature

The results confirm existing theoretical insights into telemedicine adoption, particularly in Indonesia. The findings align with previous studies, which highlight both synchronous (real-time consultations) and asynchronous (store-and-forward) telemedicine models [16,18]. Physicians' reliance on WhatsApp, Zoom, and hospital-based platforms demonstrates the transition from minimal telemedicine adoption prepandemic to a more structured implementation during and post–COVID-19.

However, the results reveal that despite growing adoption, telemedicine is still not evenly distributed. This reflects reports that, although telemedicine adoption increased during the pandemic, its growth has since plateaued or declined in some regions [22,23]. The Indonesian data aligns particularly well with the trend observed in India, where slow adoption is due to remote infrastructure challenges.

From the perspectives of both patients and health care providers, telemedicine was associated with improved access to care, reduced travel time and costs, and was particularly useful for follow-ups and managing chronic diseases. However, shared concerns included clinical limitations, unreliable internet connections, and legal uncertainties. Referring to Figure 3, health care professionals benefit from expanded patient reach and improved remote monitoring but face limitations such as the inability to conduct physical examinations and increased administrative burdens. On the patient side, benefits like improved access to specialists and reduced travel are offset by challenges such as low digital literacy and poor internet access.

However, this analysis may not fully capture variations across rural Indonesia. Some areas, particularly those with digitally literate younger chronic patients and stable internet access, face physician shortages rather than technological barriers. In such cases, patients often turn to social media messaging platforms

XSL•FO
RenderX

to seek timely medical guidance from appropriate specialists, reflecting the evolving role of asynchronous telemedicine in bridging health care gaps [49,51,66]. Telemedicine is most effective and feasible for patients with high digital literacy, stable internet access, and low disease activity [43]. This further supports the trend in Indonesia, where telemedicine serves as a viable alternative for patients who meet these criteria but lack immediate access to physicians. Our study contributes that access to medical specialists is very limited outside the capital, and that telemedicine could significantly improve care across the country by improving access and lowering cost. Regarding chronic and rare disease management, the physicians' insights reflect the Asia-Pacific League of Associations for Rheumatology recommendations that telemedicine is effective for follow-ups but should not replace face-to-face visits for initial diagnoses [52]. This aligns with findings that emphasize the importance of physical examinations in clinical decision-making [49]. Moreover, the results echo research showing that over half of autoimmune patients prefer in-person consultations, particularly for first-time visits [43]. However, once a patient-physician relationship is established, they are more open to telemedicine follow-ups, reflecting evidence that patients prefer continuity with familiar physicians [15].

Figure 3. Summary of the benefits and limitations of telemedicine in Indonesia.

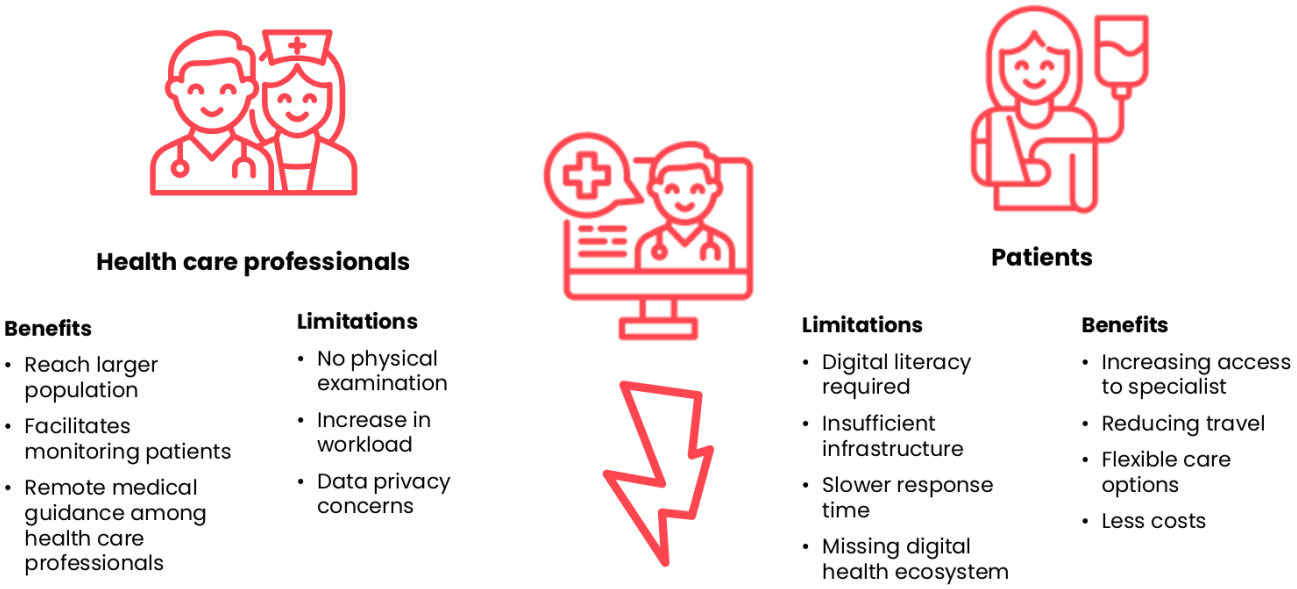

The cost-effectiveness of telemedicine in chronic disease management is also validated. Telemedicine has been shown to reduce travel costs and improve medication adherence, a pattern echoed in physicians' reports of its benefits for remote patients in Kalimantan and Sulawesi [45,47]. These results reinforce previous findings that telemedicine reduces costs for patients and providers, reducing inequalities in health care delivery and thus advancing SDGs 3 and 10. A systematic review of studies conducted in high-income countries further indicates that patients with chronic diseases show a measurable willingness to pay for telemedicine, ranging between 19% and 70% across studies, with higher willingness linked to greater distance from health facilities, while older patients were less willing to pay [82]. In Sweden, digital consultations led to a 41.5% reduction in health care expenses [54], a trend mirrored in Indonesia, where chat-based teleconsultations are considerably more affordable than face-to-face visits [83,84]. However, the study also highlights barriers to affordability. Unlike Sweden and France, where telemedicine consultations are reimbursed by statutory health insurances, Indonesia's system still relies on private insurance or out-of-pocket payments [55,85]. This limits access for low-income populations and supports concerns that telemedicine may exacerbate health care inequalities if not implemented equitably [49,86].

Beyond cost savings, telemedicine improves access to specialists for patients in remote areas who require consultations with rare disease specialists without geographical constraints [87,88]. This is evident in Indonesia's surgical conference model, where pediatric cardiologists conduct video consultations to assess and monitor children awaiting congenital heart disease surgery. Therefore, this study confirms existing challenges from the literature.

Practical Implications: Potential for Future Telemedicine Advancement

While telemedicine presents substantial benefits, physicians continue to recognize its limitations, particularly the necessity of physical examinations for certain medical fields such as rheumatology, neurology, and surgery [89]. Although effective for initial screenings and follow-ups, these specialties often require in-person assessments. Additionally, concerns have been raised about physician burnout, as virtual consultations expand patient reach beyond local boundaries and increase workload [59,90]. While the phenomenon of an increased workload and the resulting risk of burnout is not new to health care professionals, it certainly needs more attention in the context of telemedicine, from scholars and regulators alike. Similarly, concerns have been raised in the Association of Southeast Asian Nation-Japan Center report, where Indonesian health care providers express uncertainty over data security and ownership when using independent telemedicine platforms [91]. While unregulated platforms such as WhatsApp and Zoom are widely used at the moment (also in other health-related contexts such

XSL•FO
RenderX

as nutrition monitoring), it remains to be seen whether regulators will enforce other platforms in the future to increase privacy and security, while maintaining low entry barriers and an easy-to-use system [92]. Furthermore, infrastructural challenges—specifically unreliable electricity and limited internet connectivity in remote areas—continue to hinder telemedicine implementation in rural Indonesia [6,33].

Despite these challenges, the findings point to significant opportunities for advancing telemedicine. Physicians' vision for the development of telemedicine hubs across Indonesia aligns with Sweden's integrated telehealth model, which connects regional hospitals and primary care centers with specialized institutions to optimize care delivery [26]. Technological innovations such as VR can enhance remote diagnostics and treatment by offering immersive patient experiences, particularly beneficial in underserved regions [16,93]. Moreover, AI-assisted tools, including automated retinal screening for diabetic retinopathy, show promise in increasing diagnostic accuracy and early intervention. Integrating such tools into national health programs like Program Pengelolaan Penyakit Kronis could significantly strengthen chronic disease management, especially for diabetes, by facilitating proactive care and reducing complications [30].

Study Limitations and Directions for Future Research

While this study provides valuable insight into telemedicine implementation in a low-resource setting such as Indonesia, it is not without limitations. First, the sample size and scope were limited to a specific group of physicians and patients, which may not fully capture the diversity of Indonesia's health care landscape. Although our qualitative analysis did not reveal notable differences in telemedicine use based on gender or socioeconomic status within the small patient sample, we acknowledge that the composition of the sample was limited and skewed. Most of the participants were female and were predominantly insured by BPJS. Second, the study did not sufficiently account for regional disparities. Although adoption trends were discussed, variations in infrastructure quality, digital literacy, and health care accessibility between urban and rural areas were not comprehensively addressed. These disparities significantly influence the rate of telemedicine uptake and its overall effectiveness. Third, the study adopts a primarily short-term perspective, focusing on current patterns of telemedicine use without evaluating its long-term impact on patient outcomes, health care systems, or cost-efficiency. Longitudinal studies are needed to assess whether telemedicine can sustainably support health care delivery in Indonesia over time. At last, while the study acknowledges certain challenges related to technology and policy, it does not provide an in-depth analysis of regulatory environments, reimbursement structures, or technological readiness. These factors are crucial for understanding the scalability and sustainability of telemedicine services.

Future research should aim to expand sample diversity by including a wider range of health care professionals, patient populations, facility types, and geographic regions. In particular, studies with larger and more demographically diverse participant groups are needed to explore potential variations in telemedicine access, use, and user experience across gender, income, and insurance categories. This broader scope would enable a more representative and comprehensive understanding of telemedicine adoption across Indonesia. Additionally, long-term studies are necessary to evaluate the effects of telemedicine on patient health outcomes, cost-effectiveness, and physician workload. Exploring alternatives to in-person examinations, such as VR-based assessments, as trialed in Japan, may offer innovative solutions to address diagnostic challenges in remote settings. Furthermore, a detailed analysis of policy frameworks, legal regulations, data security protocols, and funding mechanisms is essential to assess their influence on telemedicine integration. Comparative studies with countries facing similar health care challenges could also provide valuable benchmarks and help identify best practices for more effective and equitable telemedicine implementation in Indonesia.

Conclusions

The findings of this study align with existing literature while providing additional insights into the regional disparities and infrastructure challenges related to telemedicine adoption in a major South-East Asian country. A noticeable shift in telemedicine usage occurred before, during, and after the COVID-19 pandemic. Prior to the pandemic, telemedicine adoption was minimal, with limited engagement from both physicians and patients. During the pandemic, there was a significant increase in the use of telemedicine platforms, as remote consultations became essential for managing patient volumes through tools like WhatsApp, Zoom, independent health technology applications, and hospital-based systems. Therefore, targeted investment in digital infrastructure and internet connectivity, particularly in underserved regions, would help reduce geographic inequities.

Postpandemic, the use of telemedicine for general consultations has decreased, though it continues to be important for follow-up care, the management of chronic and rare diseases, and improving access to specialists in remote areas, thereby supporting progress toward SDGs 3 and 10. While telemedicine has contributed to increased health care accessibility and cost-effectiveness, challenges remain, such as inadequate infrastructure, physician burnout, legal uncertainties, and the inability to conduct physical examinations remotely, which can hinder its impact on health equity. Hence, implementing support mechanisms for health care professionals, such as workload management systems and mental health progress, is crucial for long-term adoption.

Emerging models, such as telemedicine hubs, VR-based diagnostics, and AI-assisted screening, offer potential solutions to bridge these gaps and further promote equitable health care delivery in line with SDG 3 and SDG 10. However, the long-term success and broader adoption of telemedicine in Indonesia require sustainable regulatory frameworks, adequate funding models, consideration of the well-being of health care professionals, and continued investment in digital infrastructure. Addressing these challenges will be essential to maximize telemedicine's contribution to more inclusive and resilient health care systems that advance global commitments to SDG 3 and 10.

XSL•FO
RenderX

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) tools in the research and manuscript preparation process. In accordance with the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and language editing and translation of interview data. The GAI tools used were ChatGPT (versions 4o and 5; OpenAI) and DeepL (DeepL SE). The authors take full responsibility for the integrity, accuracy, and final content of the manuscript. GAI tools are not listed as authors and do not bear responsibility for the study's findings or conclusions. Declaration submitted by: CGP, RN, CH, and MG.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

The datasets generated and/or analyzed during the current study are not publicly available due to privacy and confidentiality agreements with participants, but deidentified data are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: CGP, RN, CH, MG
Formal analysis: CGP, RN, MG
Investigation: CGP, RN
Methodology: CGP, RN, MG
Supervision: CH, MG
Validation: CGP, RN, MG
Writing – original draft: CGP, RN, MG
Writing – review & editing: CGP, RN, CH, MG

Conflicts of Interest

None declared.

Multimedia Appendix 1
Qualitative research design.
[DOCX File, 71 KB - jmir_v28i1e83462_app1.docx]

Checklist 1
COREQ checklist.
[DOCX File, 20 KB - jmir_v28i1e83462_app2.docx]

References

1. Amagai S, Vonesh E, Adams J, Luo Y. Closing the gap: addressing telehealth disparities across specialties in the sustained pandemic era. NPJ Digit Med 2024 Aug 21;7(1):217. [doi: 10.1038/s41746-024-01201-w] [Medline: 39164391]
2. Hodgkins M, Barron M, Jevaji S, Lloyd S. Physician requirements for adoption of telehealth following the SARS-CoV-2 pandemic. NPJ Digit Med 2021 Feb 9;4(1):19. [doi: 10.1038/s41746-021-00390-y] [Medline: 33564102]
3. Lytras MD, Housawi AA, Alsaywid BS, editors. Digital Transformation in Healthcare in Post-Covid-19 Times: Next Generation Technology Driven Personalized Medicine and Smart Healthcare: Academic Press; 2023. URL: https://www.sciencedirect.com/book/edited-volume/9780323983532/digital-transformation-in-healthcare-in-post-covid-19-times [accessed 2026-02-18]
4. Haleem A, Javaid M, Singh RP, Suman R. Telemedicine for healthcare: capabilities, features, barriers, and applications. Sens Int 2021;2:100117. [doi: 10.1016/j.sintl.2021.100117] [Medline: 34806053]
5. Implementing telemedicine services during COVID-19: guiding principles and considerations for a stepwise approach. WHO. 2020. URL: https://www.who.int/publications/i/item/WPR-DSE-2020-032 [accessed 2024-08-09]
6. WHO issues consolidated guide to running effective telemedicine services: when delivered effectively, telemedicine can broaden healthcare access. WHO. 2022. URL: https://www.who.int/news/item/10-11-2022-who-issues-new-guide-to-running-effective-telemedicine-services [accessed 2024-08-09]
7. Indonesia. IHME. 2024. URL: https://www.healthdata.org/research-analysis/health-by-location/profiles/indonesia [accessed 2024-08-10]

XSL•FO
RenderX

8. Digital health transformation and telehealth regulation in improving health access in indonesia. UNDP Indonesia YouTube page. 2022 Jan 9. URL: https://www.youtube.com/watch?v=fbNezaOdWLY [accessed 2026-02-28]
9. Bakibinga P, Kamande E, Kisia L, Omuya M, Matanda DJ, Kyobutungi C. Challenges and prospects for implementation of community health volunteers' digital health solutions in Kenya: a qualitative study. BMC Health Serv Res 2020 Sep 21;20(1):888. [doi: 10.1186/s12913-020-05711-7] [Medline: 32957958]
10. Danila MI, Gavigan K, Rivera E, et al. Patient perceptions and preferences regarding telemedicine for autoimmune rheumatic diseases care during the COVID‐19 pandemic. Arthritis Care Res (Hoboken) 2022 Jul;74(7):1049-1057. [doi: 10.1002/acr.24860]
11. Kaihlanen AM, Virtanen L, Buchert U, et al. Towards digital health equity - a qualitative study of the challenges experienced by vulnerable groups in using digital health services in the COVID-19 era. BMC Health Serv Res 2022 Feb 12;22(1):188. [doi: 10.1186/s12913-022-07584-4] [Medline: 35151302]
12. Consolidated telemedicine implementation guide. : World Health Organization; 2022 URL: https://iris.who.int/server/api/core/bitstreams/d119a69e-e56d-4c78-a2b9-52572e14217c/content [accessed 2026-02-19]
13. Goeldner M, Gehder S. Digital health applications (DiGAs) on a fast track: insights from a data-driven analysis of prescribable digital therapeutics in Germany from 2020 to mid-2024. J Med Internet Res 2024 Aug 29;26:e59013. [doi: 10.2196/59013] [Medline: 39208415]
14. Higgins C, Dunn E, Conrath D. Telemedicine: an historical perspective. Telecomm Pol 1984 Dec;8(4):307-313. [doi: 10.1016/0308-5961(84)90044-2]
15. Kruse CS, Molina-Nava A, Kapoor Y, Anerobi C, Maddukuri H. Analyzing the effect of telemedicine on domains of quality through facilitators and barriers to adoption: systematic review. J Med Internet Res 2023 Jan 5;25:e43601. [doi: 10.2196/43601] [Medline: 36602844]
16. Sugawara Y, Hirakawa Y, Iwagami M, et al. Issues in the adoption of online medical care: cross-sectional questionnaire survey. J Med Internet Res 2024 Nov 27;26:e68439. [doi: 10.2196/68439] [Medline: 39602799]
17. Ronquillo Y, Meyers A, Korvek SJ. Digital health. In: StatPearls: StatPearls Publishing; 2025. URL: https://www.ncbi.nlm.nih.gov/books/NBK470260/ [accessed 2026-02-19]
18. Eswaran H, Dawson L, Hernandez M, et al. Telehealth: current definitions and future trends. : The Rural Telehealth Evaluation Center; 2022 URL: https://idhi.uams.edu/rtec/wp-content/uploads/sites/4/2022/05/Telehealth-Definitions-Paper-06MAY2022-1.pdf [accessed 2026-02-19]
19. Haimi M, Gesser-Edelsburg A. Application and implementation of telehealth services designed for the elderly population during the COVID-19 pandemic: a systematic review. Health Informatics J 2022;28(1):14604582221075561. [doi: 10.1177/14604582221075561] [Medline: 35175881]
20. Online doctor consultations. Statista. 2024. URL: https://www.statista.com/outlook/hmo/digital-health/online-doctor-consultations/indonesia [accessed 2026-02-19]
21. Osman S, Churruca K, Ellis LA, Luo D, Braithwaite J. The unintended consequences of telehealth in Australia: critical interpretive synthesis. J Med Internet Res 2024 Aug 27;26:e57848. [doi: 10.2196/57848] [Medline: 39190446]
22. Telehealth adoption rate in the Asia-Pacific region in 2019 and 2021, with a forecast for 2024, by selected country. Statista. 2024. URL: https://www.statista.com/statistics/1294089/apac-telehealth-adoption-rate-by-country/ [accessed 2026-02-19]
23. Kapur V, Boulton A, Angus D, Sukhrani D. Asia-pacific telemedicine is here to stay. Bain & Company. 2024. URL: https://www.bain.com/insights/asia-pacific-telemedicine-is-here-to-stay-snap-chart [accessed 2026-02-19]
24. Keelara R, Sutherland E, Almyranti M. Beyond the pandemic: leading practices for the future of telemedicine. : OECD; 2025 URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/01/beyond-the-pandemic_0b39f409/98863928-en.pdf [accessed 2026-02-19]
25. Sweden: country health profile 2023. : OECD Publishing; 2023 URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/12/sweden-country-health-profile-2023_1f0c8967/ec938a6d-en.pdf [accessed 2026-02-19]
26. The digital transformation of healthcare. Health by Sweden. 2025. URL: https://www.healthbysweden.com/the-digital-transformation-of-healthcare/ [accessed 2025-02-05]
27. Chittim G, Pappas A, Bomba J. The changing fortunes of telemedicine in Europe – past, present, and future beyond COVID-19. Health Advances. 2020. URL: https://healthadvances.com/insights/blog/the-changing-fortunes-of-telemedicine-in-europe [accessed 2025-02-15]
28. Messal H, Müller T, Richter L, Silberzahn T. Germany's eHealth transformation makes good but uneven progress. McKinsey & Company. 2022. URL: https://www.mckinsey.com/industries/life-sciences/our-insights/germanys-ehealth-transformation-makes-good-but-uneven-progress [accessed 2025-02-01]
29. Beratung am Bildschirm: videosprechstunden werden wieder deutlich weniger genutzt [Web page in German]. Welt. 2024. URL: https://www.welt.de/regionales/hamburg/article252891742/Beratung-am-Bildschirm-Videosprechstunden-werden-wieder-deutlich-weniger-genutzt.html [accessed 2025-02-05]
30. The COVID-19 pandemic and the future of telemedicine. : OECD; 2023 URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/01/the-covid-19-pandemic-and-the-future-of-telemedicine_1c878192/ac8b0a27-en.pdf [accessed 2026-02-19]

XSL•FO
RenderX

31. Holl F, Kircher J, Hertelendy AJ, Sukums F, Swoboda W. Tanzania's and Germany's digital health strategies and their consistency with the World Health Organization's global strategy on digital health 2020-2025: comparative policy analysis. J Med Internet Res 2024 Mar 18;26:e52150. [doi: 10.2196/52150] [Medline: 38498021]
32. Omboni S, Padwal RS, Alessa T, et al. The worldwide impact of telemedicine during COVID-19: current evidence and recommendations for the future. Connect Health 2022 Jan 4;1:7-35. [doi: 10.20517/ch.2021.03] [Medline: 35233563]
33. Davalbhakta S, Advani S, Kumar S, et al. A systematic review of smartphone applications available for coronavirus disease 2019 (COVID-19) and the assessment of their quality using the Mobile Application Rating Scale (MARS). J Med Syst 2020 Aug 10;44(9):164. [doi: 10.1007/s10916-020-01633-3] [Medline: 32779002]
34. Dorsey ER, Topol EJ. Telemedicine 2020 and the next decade. Lancet 2020 Mar 14;395(10227):859. [doi: 10.1016/S0140-6736(20)30424-4] [Medline: 32171399]
35. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. Nat Biomed Eng 2021 Jun;5(6):493-497. [doi: 10.1038/s41551-021-00751-8] [Medline: 34131324]
36. Indonesia population. World Population Review. 2024. URL: https://worldpopulationreview.com/countries/indonesia-population [accessed 2024-07-14]
37. Indonesia launches a blueprint on digital health to expand inclusive health care coverage. UNDP Indonesia. 2021. URL: https://www.undp.org/indonesia/press-releases/indonesia-launches-blueprint-digital-health-expand-inclusive-health-care-coverage [accessed 2026-02-19]
38. Digitalizing Indonesia's health sector, a critical step towards SDGs achievement. UNDP Indonesia. 2021. URL: https://www.undp.org/indonesia/news/digitalizing-indonesias-health-sector-critical-step-towards-sdgs-achievement#:~:text=Indonesia%20Ministry%20of%20Health%20emphasized%20needed%20herd%20immunity%20level [accessed 2026-02-19]
39. Wijaya JH, Octavius GS, Hwei LRY. A literature review of telemedicine in Indonesia: past, present, and future prospective. J Administr Kesehat Indones 2022;10(2):261-272. [doi: 10.20473/jaki.v10i2.2022.261-272]
40. Bodenheimer T, Wagner EH, Grumbach K. Improving primary care for patients with chronic illness. JAMA 2002 Oct 9;288(14):1775-1779. [doi: 10.1001/jama.288.14.1775] [Medline: 12365965]
41. Bashshur RL, Shannon GW, Smith BR, et al. The empirical foundations of telemedicine interventions for chronic disease management. Telemed J E Health 2014 Sep;20(9):769-800. [doi: 10.1089/tmj.2014.9981] [Medline: 24968105]
42. Classification of rare diseases. Orphanet. URL: https://www.orpha.net/en/disease/classification [accessed 2025-05-20]
43. Jackson LE, Edgil TA, Hill B, et al. Telemedicine in rheumatology care: a systematic review. Semin Arthritis Rheum 2022 Oct;56:152045. [doi: 10.1016/j.semarthrit.2022.152045] [Medline: 35843158]
44. Riaz MM, Mahmood SBZ, Nasir N. Telemedicine for rheumatology patients in COVID-19 pandemic: perspectives of patients. J Patient Exp 2022;9:23743735221092635. [doi: 10.1177/23743735221092635] [Medline: 35441090]
45. Patel KB, Turner K, Alishahi Tabriz A, et al. Estimated indirect cost savings of using telehealth among nonelderly patients with cancer. JAMA Netw Open 2023 Jan 3;6(1):e2250211. [doi: 10.1001/jamanetworkopen.2022.50211] [Medline: 36626174]
46. Nappoe SA, Djasri H, Kurniawan MF. Chronic Disease Management Programme (PROLANIS) in Indonesia. : World Health Organization; 2023 URL: https://iris.who.int/server/api/core/bitstreams/6546593d-9cd6-4dc5-b47b-4c6031b28c3f/content [accessed 2026-02-19]
47. Ma Y, Zhao C, Zhao Y, et al. Telemedicine application in patients with chronic disease: a systematic review and meta-analysis. BMC Med Inform Decis Mak 2022 Apr 19;22(1):105. [doi: 10.1186/s12911-022-01845-2] [Medline: 35440082]
48. Sharareh B, Schwarzkopf R. Effectiveness of telemedical applications in postoperative follow-up after total joint arthroplasty. J Arthroplasty 2014 May;29(5):918-922.e1. [doi: 10.1016/j.arth.2013.09.019] [Medline: 24342278]
49. Greenhalgh T, Vijayaraghavan S, Wherton J, et al. Virtual online consultations: advantages and limitations (VOCAL) study. BMJ Open 2016 Jan 29;6(1):e009388. [doi: 10.1136/bmjopen-2015-009388] [Medline: 26826147]
50. Choi NG, Hegel MT, Marti N, Marinucci ML, Sirrianni L, Bruce ML. Telehealth problem-solving therapy for depressed low-income homebound older adults. Am J Geriatr Psychiatry 2014 Mar;22(3):263-271. [doi: 10.1097/JGP.0b013e318266b356] [Medline: 23567376]
51. Nelson LA, Spieker A, Greevy R, LeStourgeon LM, Wallston KA, Mayberry LS. User engagement among diverse adults in a 12-month text message–delivered diabetes support intervention: results from a randomized controlled trial. JMIR Mhealth Uhealth 2020 Jul 21;8(7):e17534. [doi: 10.2196/17534] [Medline: 32706738]
52. Ahmed S, Grainger R, Santosa A, et al. APLAR recommendations on the practice of telemedicine in rheumatology. Int J Rheum Dis 2022 Mar;25(3):247-258. [doi: 10.1111/1756-185X.14286] [Medline: 35043576]
53. Market study on telemedicine. : European Commission; 2018 URL: https://health.ec.europa.eu/system/files/2019-08/2018_provision_marketstudy_telemedicine_en_0.pdf [accessed 2026-02-19]
54. Ekman B. Cost analysis of a digital health care model in Sweden. Pharmacoecon Open 2018 Sep;2(3):347-354. [doi: 10.1007/s41669-017-0059-7] [Medline: 29623633]
55. Mounié M, Costa N, Gourdy P, et al. Cost-effectiveness evaluation of a remote monitoring programme including lifestyle education software in type 2 diabetes: results of the Educ@dom study. Diabetes Ther 2022 Apr;13(4):693-708. [doi: 10.1007/s13300-022-01207-1] [Medline: 35133640]

XSL•FO
RenderX

56. Sorwar G, Rahamn MM, Uddin R, Hoque MR. Cost and time effectiveness analysis of a telemedicine service in Bangladesh. Stud Health Technol Inform 2016;231:127-134. [Medline: 27782024]
57. Key Contacts in Indonesia. DLA Piper. 2021. URL: https://www.dlapiperintelligence.com/telehealth/countries/index.html?c=ID&t=contacts-section [accessed 2025-02-05]
58. Hall AK, Cole-Lewis H, Bernhardt JM. Mobile text messaging for health: a systematic review of reviews. Annu Rev Public Health 2015 Mar 18;36:393-415. [doi: 10.1146/annurev-publhealth-031914-122855] [Medline: 25785892]
59. Kruse CS, Krowski N, Rodriguez B, Tran L, Vela J, Brooks M. Telehealth and patient satisfaction: a systematic review and narrative analysis. BMJ Open 2017 Aug 3;7(8):e016242. [doi: 10.1136/bmjopen-2017-016242] [Medline: 28775188]
60. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
61. Donnelly C, Janssen A, Shah K, Harnett P, Vinod S, Shaw TJ. Qualitative study of international key informants' perspectives on the current and future state of healthcare quality measurement and feedback. BMJ Open 2023 Jun 7;13(6):e073697. [doi: 10.1136/bmjopen-2023-073697] [Medline: 37286326]
62. Kallio H, Pietilä AM, Johnson M, Kangasniemi M. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. J Adv Nurs 2016 Dec;72(12):2954-2965. [doi: 10.1111/jan.13031] [Medline: 27221824]
63. Khanassov V, Ilali M, Ruiz AS, Rojas-Rozo L, Sourial R. Telemedicine in primary care of older adults: a qualitative study. BMC Prim Care 2024 Jul 17;25(1):259. [doi: 10.1186/s12875-024-02518-x] [Medline: 39020277]
64. Harapan BN, Harapan T, Theodora L, Anantama NA. From archipelago to pandemic battleground: unveiling Indonesia's COVID-19 crisis. J Epidemiol Glob Health 2023 Dec;13(4):591-603. [doi: 10.1007/s44197-023-00148-7] [Medline: 37707715]
65. Pathak L, Hernandez-Ramos R, Rosales K, et al. Using social media to recruit a diverse sample of participants for a mobile health (mHealth) intervention to increase physical activity: exploratory study. JMIR Mhealth Uhealth 2025 Apr 28;13:e56329. [doi: 10.2196/56329] [Medline: 40294424]
66. Darmawan I, Bakker C, Brockman TA, Patten CA, Eder M. The role of social media in enhancing clinical trial recruitment: scoping review. J Med Internet Res 2020 Oct 26;22(10):e22810. [doi: 10.2196/22810] [Medline: 33104015]
67. Uy L, Safwan N, Ahmad R, et al. Research recruitment strategies in the digital era: opportunities and challenges of social media. Contemp Clin Trials 2025 Sep;156:108033. [doi: 10.1016/j.cct.2025.108033] [Medline: 40738218]
68. Topolovec-Vranic J, Natarajan K. The use of social media in recruitment for medical research studies: a scoping review. J Med Internet Res 2016 Nov 7;18(11):e286. [doi: 10.2196/jmir.5698] [Medline: 27821383]
69. Angum F, Khan T, Kaler J, Siddiqui L, Hussain A. The prevalence of autoimmune disorders in women: a narrative review. Cureus 2020 May 13;12(5):e8094. [doi: 10.7759/cureus.8094] [Medline: 32542149]
70. Statistik Telekomunikasi Indonesia 2024 [Article in Indonesian]. : Badan Pusat Statistik Indonesia; 2025 URL: https://tinyurl.com/mrxedn4j [accessed 2026-02-19]
71. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
72. Rädiker S, Kuckartz U. Focused Analysis of Qualitative Interviews with MAXQDA: Step by Step: Maxqda Press; 2020. [doi: 10.36192/978-3-948768072]
73. Gehder S, Goeldner M. Unpacking performance factors of innovation systems and studying Germany's attempt to foster the role of the patient through a market access pathway for digital health applications (DiGAs): exploratory mixed methods study. J Med Internet Res 2025 Jan 6;27:e66356. [doi: 10.2196/66356] [Medline: 39761562]
74. Peraturan menteri kesehatan nomor 340 tahun 2020 tentang klasifikasi rumah sakit [Article in Indonesian]. : Kementerian Kesehatan Republik Indonesia; 2010 URL: https://pelayanan.jakarta.go.id/download/regulasi/peraturan-menteri-kesehatan-nomor-340-tentang-klasifikasi-rumah-sakit.pdf [accessed 2026-02-19]
75. The republic of Indonesia health system review. WHO. 2017. URL: https://iris.who.int/bitstream/handle/10665/254716/9789290225164-eng [accessed 2025-03-01]
76. Sungkono KR, Sarno R. Determine types of Indonesian hospital by criteria-based proses model, k-means cluster, and hierarchical average linkage. Presented at: 2019 International Seminar on Application for Technology of Information and Communication (iSemantic); Sep 21-22, 2019. [doi: 10.1109/ISEMANTIC.2019.8884299]
77. JISDOR. Bank Indonesia. URL: https://www.bi.go.id/id/statistik/informasi-kurs/jisdor/Default.aspx [accessed 2024-12-31]
78. Euro foreign exchange reference rates. European Central Bank. 2024. URL: https://www.ecb.europa.eu/stats/policy_and_exchange_rates/euro_reference_exchange_rates/html/index.en.html [accessed 2024-12-31]
79. Indonesia's telemedicine industry one of fastest-growing: minister. Antara News. URL: https://en.antaranews.com/news/291951/indonesias-telemedicine-industry-one-of-fastest-growing-minister [accessed 2024-08-11]
80. Corbett JA, Opladen JM, Bisognano JD. Telemedicine can revolutionize the treatment of chronic disease. Int J Cardiol Hypertens 2020 Dec;7:100051. [doi: 10.1016/j.ijchy.2020.100051] [Medline: 33330846]
81. Shaqeel M. President Jokowi gives appreciation for the contribution of various parties in handling the COVID-19 pandemic. Industry.co.id. 2023. URL: https://en.industry.co.id/read/8067/

XSL•FO
RenderX

president-jokowi-gives-appreciation-for-the-contribution-of-various-parties-in-handling-the-covid-19-pandemic#google_vignette [accessed 2024-08-11]
82. Chua V, Koh JH, Koh CHG, Tyagi S. The willingness to pay for telemedicine among patients with chronic diseases: systematic review. J Med Internet Res 2022 Apr 13;24(4):e33372. [doi: 10.2196/33372] [Medline: 35416779]
83. Dokter umum - konsultasi online terpercaya 24 jam [Article in Indonesian]. Halodoc. 2020. URL: https://www.halodoc.com/tanya-dokter/kategori/dokter-umum [accessed 2025-02-23]
84. MySiloam. Siloam Hospitals. 2025. URL: https://www.siloamhospitals.com/en/mysiloam [accessed 2025-02-23]
85. Key Contacts in Sweden. DLA Piper. 2021. URL: https://www.dlapiperintelligence.com/telehealth/countries/index.html?c=SE&t=contacts-section [accessed 2025-02-25]
86. Yao R, Zhang W, Evans R, Cao G, Rui T, Shen L. Inequities in health care services caused by the adoption of digital health technologies: scoping review. J Med Internet Res 2022 Mar 21;24(3):e34144. [doi: 10.2196/34144] [Medline: 35311682]
87. Rahman FF. Indonesia's healthcare landscape: embracing innovation in the new health regime. Curr Med Res Opin 2024 May 6;2024:1-5. [doi: 10.1080/03007995.2024.2349732] [Medline: 38683759]
88. Bin KJ, Santana Alves PG, Costa R, et al. User experience regarding digital primary health care in Santarém, Amazon: evaluation of patient satisfaction and Doctor's feedback. JMIR Form Res 2023 Jan 11;7:e39034. [doi: 10.2196/39034] [Medline: 36630164]
89. Kaissar S, Zara S, Fabrice F, et al. Physicians' satisfaction with the use of teleconsultation in France. J Hosp Manag Health Policy 2023;7:1. [doi: 10.21037/jhmhp-22-76]
90. Mandal S, Wiesenfeld BM, Mann DM, Szerencsy AC, Iturrate E, Nov O. Quantifying the impact of telemedicine and patient medical advice request messages on physicians' work-outside-work. NPJ Digit Med 2024 Feb 14;7(1):35. [doi: 10.1038/s41746-024-01001-2] [Medline: 38355913]
91. Navallo K, Detros K. Assessing digital health adoption in ASEAN. : ASEAN-Japan Centre; 2024 URL: https://www.asean.or.jp/main-site/wp-content/uploads/2024/07/Assessing-Digital-Health-Adoption-in-ASEAN.pdf [accessed 2026-02-19]
92. Aisyah DN, Mayadewi CA, Utami A, et al. Using WhatsApp for nutrition surveillance among children under 5 years in West Java, Indonesia: cross-sectional survey and feasibility study. JMIR Pediatr Parent 2025 May 15;8:e58752. [doi: 10.2196/58752] [Medline: 40372943]
93. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. NPJ Digit Med 2023 Oct 9;6(1):186. [doi: 10.1038/s41746-023-00927-3] [Medline: 37813960]

Abbreviations

AI: artificial intelligence
BPJS: Badan Pengelola Jaminan Sosial Kesehatan (National Health Insurance Agency)
COREQ: Consolidated Criteria for Reporting Qualitative Research
GAI: generative artificial intelligence
NCD: noncommunicable disease
SDG: sustainable development goal
VR: virtual reality

Edited by A Stone; submitted 03.Sep.2025; peer-reviewed by A Rasool, IAP Tarquino; accepted 02.Feb.2026; published 19.Mar.2026.

Please cite as:
Pratama CG, Nurlianti R, Herstatt C, Goeldner M
Telemedicine Adoption for Managing Chronic and Rare Diseases in Indonesia During and Beyond the COVID-19 Era: Qualitative Study
J Med Internet Res 2026;28:e83462
URL: https://www.jmir.org/2026/1/e83462
doi:10.2196/83462

© Christine Gracia Pratama, Rima Nurlianti, Cornelius Herstatt, Moritz Goeldner. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Telehealth Intervention to Reduce Sedentary Behavior in Older Adults With Type 2 Diabetes: Development and Feasibility Study

Xiaoyan Zhang[1,2*], PhD; Dan Yang[3*], PhD; Sihan Chen[1], BSN; Lei Wang[4], BSN; Xin Li[1], BSN; Ziyan Wang[1], BSN; Meiqi Meng[1], MSN; Xuejing Li[1], PhD; Hongzhan Jiang[1], MSN; Yufang Hao[1], PhD; Jiayin Luo[2], MSN

[1]School of Nursing, Beijing University of Chinese Medicine, Liangxiang High Education Park, Fangshan District, Beijing, China

[2]Department of Vascular Surgery, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing, China

[3]Department of Nursing Administration, Fangshan Hospital Beijing University of Chinese Medicine, Beijing, China

[4]Department of Nursing, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Yufang Hao, PhD
School of Nursing, Beijing University of Chinese Medicine, Liangxiang High Education Park, Fangshan District, Beijing, China

Abstract

Background: Sedentary behavior (SB) is a modifiable risk factor for complications in older adults with type 2 diabetes mellitus (T2DM). Despite widespread adoption of digital health platforms, theory-driven telehealth interventions specifically targeting SB reduction remain limited, particularly those incorporating cultural adaptation and behavioral change frameworks.

Objective: This study aims to develop and evaluate the feasibility of a theory-based personalized telehealth intervention to reduce SB in older adults with T2DM in China.

Methods: The intervention was developed over 14 months (January 2022-February 2023) following the intervention mapping and Behavior Change Wheel frameworks. A panel of 19 multidisciplinary experts (90.5% response rate) refined the program through a systematic iterative process. Subsequently, a 7-week quasi-experimental study (pre-post self-controlled design) was conducted to assess feasibility. We recruited 30 community-dwelling older adults with T2DM via WeChat-based convenience sampling. The primary outcome was SB measured by the Measure of Older Adults' Sedentary Time for Type 2 Diabetes Mellitus questionnaire. Secondary outcomes included cardiovascular risk (blood pressure), glycemic control (fasting blood glucose), Diabetes-Specific Quality of Life, social isolation, BMI, and fall incidence. Pre-post changes from baseline to 7 weeks were statistically evaluated to assess the intervention's feasibility and preliminary impact.

Results: The intervention comprises 5 components: an eHealth education manual, a motion graphics library, an SMS text messaging library, a WeChat Q&A group, and a material incentive package. These components address "knowledge," "social support," and "intention" determinants through "education," "enablement," and "incentivisation" functions, respectively. All components used the "service provision" policy and various behavior change techniques. Preliminary feasibility testing (n=31) showed reduced sedentary time by 1.12 hours/day ($P<.001$) and improved social connectivity scores ($P=.001$).

Conclusions: This study demonstrates the feasibility and potential impact of a systematically developed telehealth intervention for reducing SB in older adults with T2DM in China. The integration of intervention mapping with the Behavior Change Wheel provides a replicable framework for developing theory-driven digital health interventions. With significant reductions in sedentary time and improved social connectivity, this culturally adapted approach offers a scalable model for chronic disease self-management in aging populations. The systematic methodology and positive preliminary outcomes support further large-scale evaluation of evidence-based telehealth solutions for behavioral modification in diabetes care.

(J Med Internet Res 2026;28:e80827) doi:10.2196/80827

KEYWORDS

intervention mapping; behavior change intervention; sedentary behavior; older adults; mixed methods; telehealth interventions

Introduction

Type 2 diabetes mellitus (T2DM) represents a global health challenge characterized by insufficient insulin secretion and/or dysfunction [1]. The International Diabetes Federation reported that diabetes prevalence among older adults (aged ≥65 y) reached 19.3% globally in 2020, with projections indicating an increase to 27.6% by 2045 [2]. This trend is particularly

XSL•FO
RenderX

pronounced in upper-middle-income countries (UMICs), where age-standardized mortality rates remain substantially elevated [3]. China exemplifies this challenge with a T2DM prevalence of 30.2% among its older population [4]. Given the profound impact, lifestyle modifications offer a potentially cost-effective and sustainable method for improving T2DM outcomes [5], with sedentary behavior (SB) emerging as a critical modifiable risk factor.

SB is defined as any waking behavior characterized by an energy expenditure of 1.5 metabolic equivalents or less while sitting, reclining, or lying [6]. Distinct from physical inactivity, adverse health effects may persist even when recommended activity levels are met [7]. Globally, 67% of adults aged 60 years and older spend over 8.5 hours daily in SB [8], with 69.8% of Chinese older adults with diabetes classified as sedentary [9]. Extensive research confirms SB's adverse impacts, including increased all-cause mortality [10], cardiovascular disease [11], cancer [12], and metabolic disorders [13-16]. SB represents a significant modifiable risk factor requiring targeted intervention in T2DM management.

While the World Health Organization (WHO) and various countries' guidelines recommend reducing sedentary time [17-19], promoting physical activity alone proves insufficient [20,21]. Current SB interventions face challenges: moderate activity often replaces light activity rather than sedentary time [22], and recommended activity levels may paradoxically increase compensatory SB [23]. Notably, a 2021 Cochrane systematic review revealed no statistically significant effects of current SB interventions for community-dwelling older adults [24]. As the global chronic disease burden increases, telemedicine emerges as a crucial tool for providing accessible care [25]. While small-scale studies suggest potential efficacy [26], patient-oriented behavioral interventions remain fragmented—a challenge prevalent globally in digitally advancing health care systems [27,28]. There is a pressing need for theory-based, culturally adapted approaches that systematically integrate theoretical frameworks into remote SB intervention design.

Traditional implementation strategy development often relies on the "It Seemed Like a Good Idea at the Time" principle, targeting barriers without theoretical foundation, thereby compromising outcome interpretation and failing to elucidate mechanisms of action [29]. The UK Medical Research Council recommends systematic, theoretically grounded approaches [29]. Implementation science has emerged as a discipline exploring how to systematically integrate evidence-based interventions into routine practice [30-32]. To operationalize these principles, the intervention mapping (IM) framework [33,34] provides a structured, theory-driven approach that constructs logical problem models, identifies behavioral determinants, and matches them with behavioral objectives. Successfully applied in SB reduction interventions across diverse populations [35-38], IM ensures systematic development of theoretically grounded strategies.

Our foundational work involved completing step 1 of the IM framework through previously published research [39]. This foundational phase used a 2-phase sequential explanatory mixed methods design based on the Behavior Change Wheel (BCW). Initially, we used an established conceptual framework (Figure 1) to conduct a cross-sectional survey, which validated key SB patterns and determinants [39]. Subsequently, qualitative interviews were conducted to enrich the model with patient perspectives and identify contextual influencing factors (X Zhang et al, PhD, unpublished data, August 2025). Through the integration of these findings, we finalized the SB logic model (Figure 2), which systematically delineates 20 modifiable determinants across capability, motivation, and opportunity dimensions, as well as a 3-stage behavioral trajectory: multipoint triggering, immersive maintenance, and transitional breaking contexts. This logic model (Figure 2) serves as the theoretical and empirical foundation for the systematic development and testing of the "Double-S" intervention in this study.

XSL•FO
RenderX

Figure 1. Preliminary conceptual framework of sedentary behavior for older adults with type 2 diabetes mellitus (T2DM). This framework was developed prior to empirical data collection by integrating the capability, opportunity, motivation, and behavior model, the theoretical domains framework, and existing literature. The model depicts a logical flow from potential determinants mapped onto COM-B dimensions on the left, through various behavioral contexts in the center, to potential health outcomes on the right. This initial framework was subsequently validated and refined through 2-phase empirical studies ([39]; X Zhang et al, PhD, unpublished data, August 2025), leading to the finalized logic model presented in Figure 2.

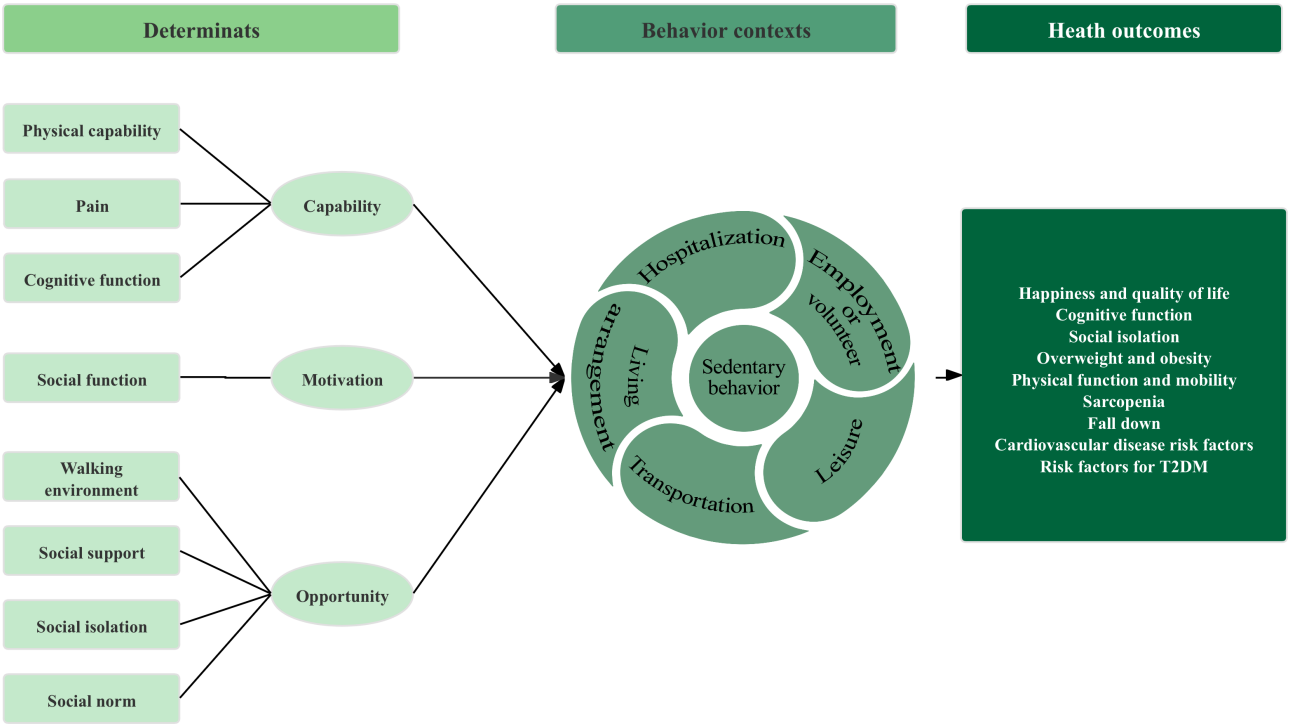

Figure 2. Logical model of sedentary behavior in older adults with type 2 diabetes mellitus (T2DM).

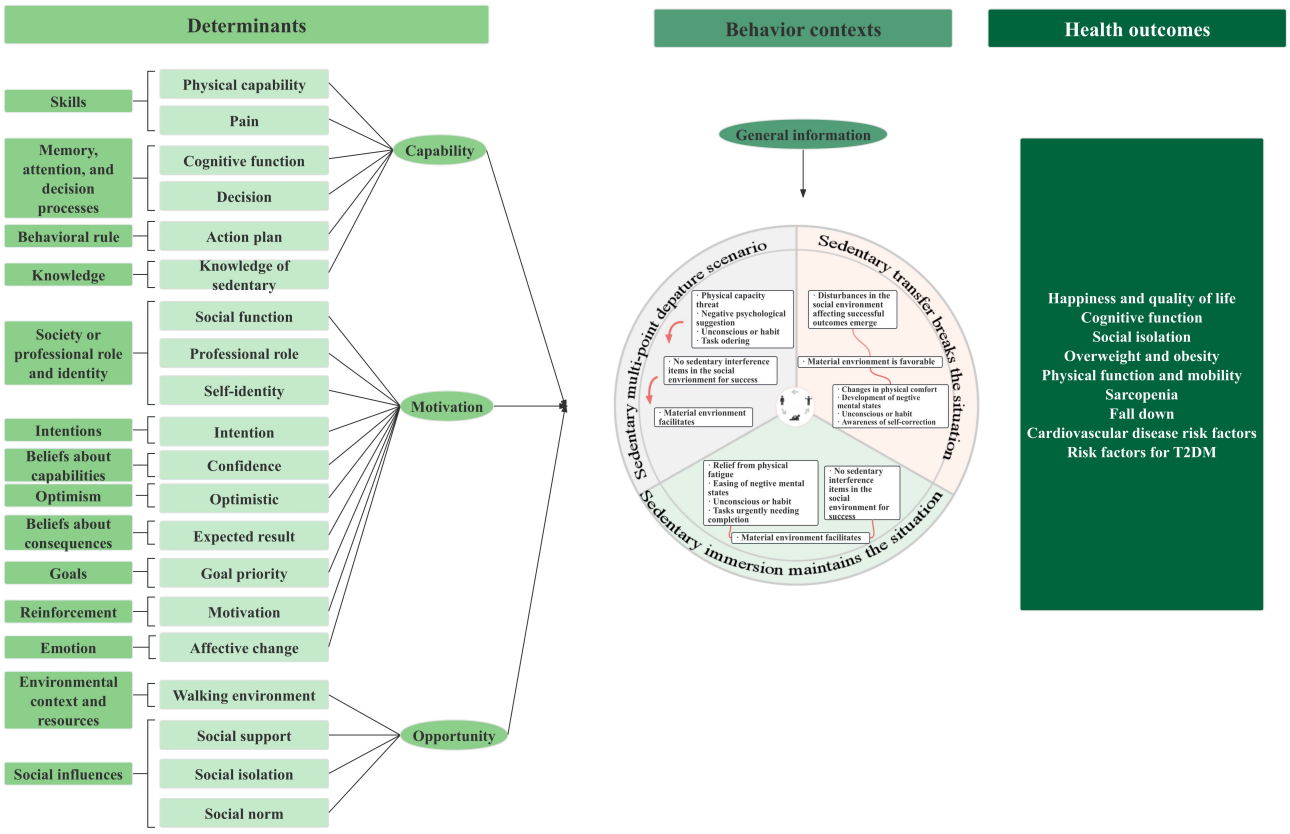

Building on these foundational insights, this study aims to translate identified behavioral determinants into a practical, technology-enabled application. We introduced the theory-driven remote intervention program “Double-S” (stop SB, stand up), designed to specifically address the identified determinants and reduce SB among community-dwelling older

XSL•FO
RenderX

adults with T2DM. This research sought to bridge the gap between theoretical frameworks and practical telehealth delivery by addressing the following specific research questions: (1) How can a multicomponent telehealth intervention be systematically developed for older Chinese adults with T2DM using the IM [33,34,40] and BCW [41] frameworks? (2) Is the resulting "Double-S" program feasible and acceptable for this population, and what are its preliminary impacts on behavioral and health-related outcomes?

Methods

Study Design

This study used the IM framework to develop and evaluate a theory-based behavior change intervention. The process focused on IM steps 2 through 6, conducted between January 2022 and February 2023, spanning 14 months (Figure 3). Building on the logic model established in step 1 (Figure 2), which was previously completed and published [39], this study focuses on IM steps 2 through 6. A detailed summary of the sampling strategies, recruitment channels, and participant involvement (and the independence of participants) across all stages is provided in Multimedia Appendix 1. IM consists of six iterative steps: (1) completed previously [39], (2) formulating program outcomes and objectives, (3) designing the program, (4) producing program components and materials, (5) planning program implementation, and (6) planning for evaluation. Each step used core processes to identify key literature, apply theory, and collect additional data [40]. Experts were purposively recruited through professional networks and institutional recommendations between January 2022 and February 2022. The research team identified potential participants based on their expertise and extended personal invitations via email, accompanied by detailed program information. The panel comprised experts from various fields: implementation science, evidence-based practice, education, psychology, geriatric nursing, clinical nursing, and diabetes specialist nursing. Eligibility criteria required professionals to hold a relevant postgraduate degree (master's or higher) or possess over 10 years of specialized work experience. Of the 21 professionals identified and invited, 19 accepted the invitation (90.5% response rate). The recruitment was conducted through the research team's professional database and academic societies related to geriatric nursing in Beijing. No formal interviews were conducted; selection was based solely on professional credentials and alignment with the program's objectives.

Figure 3. Intervention mapping (IM) process and study timeline for the "Double-S" program. APEASE: Acceptability, Practicability, Effectiveness, Affordability, Safety, Equity; BCW: Behavior Change Wheel; SB: sedentary behavior.

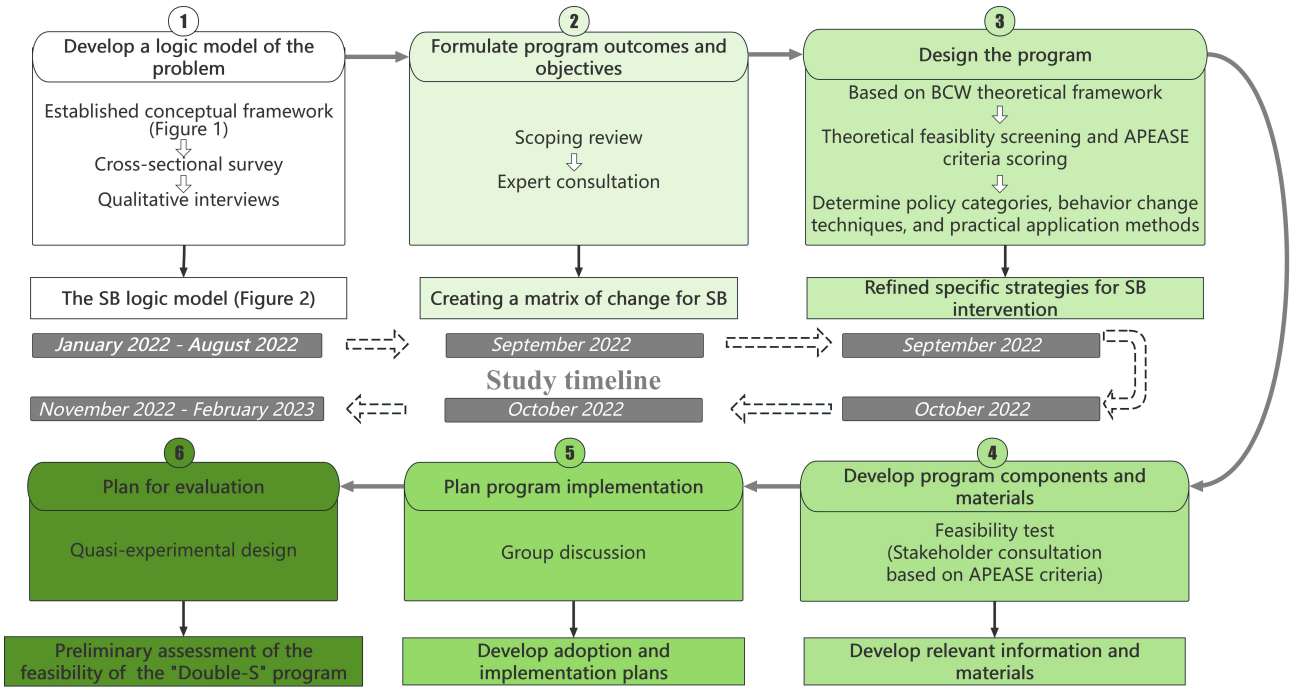

Ethical Considerations

The study protocol was approved by the ethics committee of Beijing University of Chinese Medicine (2022BZYLL0505). Digital or written informed consent was obtained from all participants prior to their enrollment. To protect participant privacy, all data were anonymized and deidentified during analysis. All participants (experts, stakeholders, and older adults) participated voluntarily and did not receive financial compensation for their involvement.

Determining Program Outcomes and Change Objectives

This phase aimed to establish behavioral outcomes for the SB intervention and articulate performance objectives (POs) for each outcome, specifying what participants must learn, do, or change to achieve particular results. The process encompassed three key steps. First, based on preliminary literature research, we defined expected behavioral change outcomes for SB intervention strategies by synthesizing evidence from multiple sources: (1) the Cochrane systematic review [24] examined the effectiveness of interventions to reduce sedentary time in adults

XSL•FO
RenderX

with type 2 diabetes, providing evidence on the feasibility and health benefits of reducing daily sedentary time; (2) our scoping review [42] systematically mapped existing evidence on SB interventions in older adults, complementing the disease-specific evidence from the Cochrane review with age-related considerations; and (3) the clinical practice guideline for exercise therapy of T2DM in China (2024 edition) [43] provided specific recommendations on the maximum duration of single sedentary bouts for patients with T2DM. These reviews complemented each other by providing both disease-specific evidence (population with T2DM) and age-specific considerations (older adults), enabling us to establish appropriate behavioral targets for our specific population of older adults with T2DM. For these expected outcomes, we set POs based on the previously developed 3-stage SB trajectory model (specifically, sub-behaviors of the target behavior). These POs specified what the study participants needed to learn, implement, or change to achieve the specific outcomes. Second, selecting important and modifiable behavioral determinants. Guided by the SB logic model, we developed an expert consultation questionnaire to systematically evaluate each determinant. This evaluation was conducted by the intervention strategy development group, consisting of a multidisciplinary panel of 19 experts with a mean professional or research experience of 7.39 (SD 6.94) years. The panel's expertise spanned various fields, including implementation science, evidence-based practice, education, psychology, geriatric nursing, community nursing, and specialized diabetes nursing. Members of the intervention strategy development group (n=19) independently rated each determinant on 2 dimensions using a 5-point Likert scale (1="very low" to 5="very high"): (1) importance—the extent to which the determinant influences SB in older adults, and (2) modifiability—the extent to which the determinant can be changed through intervention strategies. Determinants scoring 3 or more (indicating moderate to high levels) on both dimensions were selected as priority targets for intervention development. This scoring and threshold system was established to ensure that the resulting intervention would focus on determinants with the highest potential for meaningful behavior change and practical feasibility. Third, creating a matrix of change for SB. The intervention strategy group constructed a change matrix based on the screened determinants and established POs. This step aimed to operationalize objectives, enhancing their practicality and feasibility.

Program Design

Framework

This phase aimed to translate change objectives into practical intervention strategies using a theory-driven approach. We adopted the BCW [41] as our overarching framework, as it systematically integrates the Capability, Opportunity, Motivation, and Behavior (COM-B) model, the theoretical domains framework (TDF), and behavior change techniques (BCTs). The COM-B model explains behavior change mechanisms by analyzing the interplay of 3 factors: capability, opportunity, and motivation. Building on COM-B, the TDF provides a more detailed classification of behavioral determinants (Multimedia Appendix 2). BCTs are specific intervention measures designed to modulate behavioral determinants [44]. Based on this theoretical framework and its components, intervention strategies were developed through the following steps.

Identifying Intervention Functions

Based on the relationship between determinants identified by the TDF and intervention functions recommended by the BCW (Multimedia Appendix 3), we screened theoretically feasible intervention functions. Subsequently, the intervention development group applied the APEASE (Affordability, Practicability, Effectiveness/Cost-effectiveness, Acceptability, Side-Effects/Safety, Equity) criteria [45] to score the prescreened intervention functions. Functions with the highest scores were selected for strategy design. The APEASE scoring system ranged from 1 to 3, with higher scores indicating greater development potential for the intervention function (Multimedia Appendix 4).

Determining Policy Categories, BCTs, and Practical Application Methods

First, we identified theoretically applicable policy categories to support the selected intervention functions, primarily based on the BCW-recommended matching between intervention functions and policy categories (Multimedia Appendix 3). Subsequently, we referred to the results of our previous literature review [42] on SB interventions for older adults and the BCTs recommended in the BCW guide that correspond to specific intervention functions. Group members then determined the BCTs by mapping the selected intervention functions to the specific needs of older adults with T2DM, guided by both the BCW manual and findings from our previous literature review [42]. Finally, considering the identified intervention functions, policy categories, and BCTs, as well as the context and environment of intervention implementation, the intervention development group further refined specific strategies for SB intervention in community-dwelling older adults with T2DM.

Program Production

This phase aimed to refine the structure of intervention strategies and to draft, test for feasibility, and produce relevant information and materials. The intervention strategy group developed an initial draft of SB intervention strategies for community-dwelling older adults with T2DM, based on previously identified functions, policy categories, and BCTs. Throughout the iterative process, POs, change objectives, and practical strategies were continuously optimized.

To test the feasibility of strategies, a stakeholder consultation method was used. A panel of 9 stakeholders was recruited through purposive sampling, including experts in evidence-based medicine and nursing, specialists in education, clinicians in diabetes care, and patient representatives. These 9 stakeholders were independent of the 19 experts involved in step 2 to ensure an objective external review, although they shared similar eligibility criteria. The consultation was conducted through (eg, individual email-based or face-to-face sessions). Each stakeholder was provided with a detailed description of the proposed intervention and used a custom-designed APEASE scoring sheet to evaluate intervention strategies across 6 dimensions: acceptability, practicability, effectiveness or

XSL•FO
RenderX

cost-effectiveness, affordability, side effects or safety, and equity. Each dimension was rated on a 3-point Likert scale. Data were analyzed using descriptive statistical methods.

Program Implementation Plan

This phase focused on developing adoption and implementation plans in consultation with the intervention development group. Key stakeholders from evidence-based medicine or nursing, education, and diabetes care were engaged to discuss the integration of intervention strategies into existing services, as well as the logistics of delivery (how, when, and by whom). Training guidelines and curriculum plans were developed to equip community older people's health and diabetes management professionals with the tools to address the primary barriers contributing to SB in older adults with T2DM. In designing remote interventions, we specifically considered the resource constraints and technological infrastructure challenges common in UMICs. We opted for low-cost, easily implementable remote intervention solutions that do not require expensive equipment or complex technical support.

Evaluation Plan

This phase aimed to preliminarily assess the feasibility of the "Double-S" program for community-dwelling older adults with T2DM. A quasi-experimental design (pre-post self-controlled study) was used (registration: ER20220111).

The study population comprised community-dwelling older adults with T2DM. A convenience sampling method was used to recruit participants via WeChat in November 2022. Recruitment flyers were posted in 3 purposively selected WeChat groups of community-dwelling older adults and shared through the "Moments" functions of community health workers. Interested individuals contacted the research team directly via WeChat or phone. The researchers then screened these potential participants against the inclusion criteria and explained the study's purpose. Eligible candidates who provided informed consent were then purposively selected to ensure diversity in age and health status and formally enrolled for interviews. Inclusion criteria were: aged 60 years and older, diagnosed with T2DM, permanent community residents, ability to use WeChat on smartphones, and voluntary participation. Exclusion criteria included communication disorders, mental disorders, physical disabilities preventing standing, and comorbid severe illnesses. No formal sample size calculation was conducted due to the exploratory nature of this feasibility study; a target sample of 30 patients was set.

The "Double-S" program was delivered over a 7-week period using a WeChat-based telehealth approach. The intervention was facilitated by a trained research assistant and consisted of four primary components as follows:

- Digital education: an eHealth education manual, designed as a cloud-based document, was provided to participants for real-time access.
- Behavioral prompts: standardized reminders were selected from a predeveloped prompt library and sent to the WeChat group to encourage participants to reduce and break sedentary bouts.
- Visual learning: motion graphics (MG) animations were distributed through the WeChat group and a synchronized public video channel to demonstrate sedentary-breaking exercises.
- Interactive support: real-time Q&A sessions were conducted within the WeChat group to address participant concerns and provide peer support.

Participants underwent structured assessments at 2 specific time points: baseline (T0; week 0) and postintervention (T1; immediately following the completion of week 7). To ensure data accuracy and efficiency, all assessments were conducted online via the "Wenjuanxing" professional survey platform (similar to SurveyMonkey). The research team provided remote guidance to participants who required assistance in navigating the electronic questionnaires.

The intervention implementation adhered to the plan formulated in step 5, with assessment focusing on primary and secondary outcome measures. The primary outcome, SB, was measured using the Measure of Older Adults' Sedentary Time for T2DM Patients questionnaire [39]. Secondary outcomes encompassed a comprehensive range of health and psychosocial factors: cardiovascular disease risk (assessed by the most recent morning blood pressure readings), T2DM risk factors (evaluated using the most recent fasting blood glucose levels), quality of life (measured with the Diabetes-Specific Quality of Life scale) [46], social isolation (assessed using the Chinese version of the Lubben Social Network Scale-6 [LSNS-6]) [47,48], overweight and obesity (determined by BMI), and fall incidence (recorded as the number of individuals experiencing falls and their consequences within the past week).

Data were analyzed by SPSS Statistics v25.0 (IBM Corp). Descriptive statistics were used for continuous variables. Paired 2-tailed *t* tests or Wilcoxon signed-rank tests were used to compare preintervention and postintervention differences, depending on data distribution. The McNemar test was used for paired categorical variables. The significance level for all statistical tests was set at .05.

Results

Program Outcomes and Change Objectives

Behavioral Change Outcomes and POs

Based on the previously established logic model ([39]; X Zhang et al, PhD, unpublished data, August 2025), the intervention strategy group developed the following behavioral change outcome and POs. First, older adults with T2DM should reduce their average sedentary time by 1 hour per day (based on evidence synthesis from Cochrane systematic review [24] and our scoping review [42]), with a recommendation that single bouts of SB not exceed 30 minutes (based on the *Guideline for Exercise Therapy of T2DM in China, 2024 Edition* [43]). Second, to achieve the expected behavioral outcome, 4 POs were established based on the 3-stage trajectory of SB (Multimedia Appendix 5) as follows:

- PO1: patients should comprehend the risks and consequences of SB (prebehavior).

XSL•FO
RenderX

- PO2: patients select appropriate strategies to address SB (during behavior).
- PO3: patients implement reminders and interruptions during sedentary periods (during behavior).
- PO4: patients engage in appropriate physical activity to alleviate discomfort from SB (postbehavior).

Selection of Important and Modifiable Determinants of SB

Following the evaluation by the multidisciplinary panel described in the "Program Production" section, the assessment results (Table 1) identified knowledge, social influences, and intention as the priority TDF domains meeting both importance and modifiability criteria. These identified determinants were mapped directly to the TDF framework during the analysis. Further analysis within the social influences domain revealed that it encompassed social support, social isolation, and social norms. Considering feasibility constraints, social support was selected as the representative factor for social influence.

Table . Degree of support, importance, and modifiability scores for the determinants of sedentary behavior.

TDF[a] domains (behavioral determinants)	Literature review [49]	Cross-sectional study [39]	Qualitative research[b]	Importance, mean (SD)	Modifiability, mean (SD)
Knowledge	N[c]	N	Y[d]	4.00 (0.79)	4.37 (0.81)
Memory, attention, and decision processes	Y	Y	Y	4.74 (0.55)	2.68 (0.92)
Behavioral regulation	N	N	Y	4.58 (0.75)	2.84 (0.74)
Skills	N	Y	Y	4.00 (0.79)	2.79 (0.77)
Environmental context and resources	N	Y	Y	3.95 (1.05)	2.53 (1.09)
Social influences	N	Y	Y	3.63 (0.98)	3.47 (0.60)
Society or professional roles and identity	N	Y	Y	3.84 (0.87)	2.74 (0.85)
Intentions	N	N	Y	4.26 (0.91)	3.16 (0.49)
Beliefs about capabilities	N	N	Y	4.16 (0.67)	2.95 (0.76)
Optimism	N	N	Y	3.58 (1.09)	2.79 (0.83)
Beliefs about consequences	N	N	Y	3.47 (1.14)	2.74 (0.64)
Goals	N	N	Y	3.58 (1.04)	2.32 (0.46)
Reinforcement	N	N	Y	3.47 (0.60)	2.89 (0.64)
Emotion	N	N	Y	3.21 (0.77)	2.47 (0.60)

[a]TDF: theoretical domains framework.

[b]X Zhang et al, PhD, unpublished data, August 2025.

[c]N: not yet discovered.

[d]Y: support.

Change Matrix for SB

Based on the established POs and the selected determinants, specific change objectives were formulated to create a comprehensive change matrix, as shown in Table 2.

XSL•FO
RenderX

Table . Matrix of change objectives for sedentary behavior reduction: linking performance objectives (POs) with theoretical determinants.

PO	Determinants		
	Knowledge	Social support	Intentions
PO1: the patient should comprehend the risks and consequences of SB (before behavior).	The patient should comprehend the hazards of SB[a] (current situation, consequences, and its relationship with T2DM[b]).	Researchers or family members provide social support to the patient regarding PO1.	The patient expresses an intention to reduce SB.
PO2: the patient chooses appropriate strategies to cope with SB (during behavior).	The patient should comprehend strategies to counteract SB, such as reminders and behavior reversal.	Researchers or family members provide social support to the patient regarding PO2.	The patient expresses an intention to adopt appropriate coping strategies.
PO3: The patient is reminded and interrupted during SB (during behavior).	Family members should comprehend the hazards of SB (current situation, consequences, and its relationship with T2DM).	Researchers or family members provide social support to the patient regarding PO3.	Family members express an intention to monitor and remind the patient.
PO4：the patient engages in appropriate physical activities to alleviate discomfort from SB (after behavior).	The patient should comprehend how to engage in physical activity.	Researchers or family members provide social support to the patient regarding PO4.	The patient expresses an intention to engage in physical activity.

[a]SB: sedentary behavior.

[b]T2DM: type 2 diabetes mellitus.

Program Design

Selection of Intervention Functions

Intervention functions were evaluated using the APEASE criteria, with results presented in Multimedia Appendix 6. The BCW framework was used to align intervention functions with determinants: (1) knowledge: education (mean total score of 16.89, SD 1.10), (2) intentions: incentivization (mean highest score of 15.21, SD 1.08), and (3) social support: enablement (mean highest score of 15.10, SD 1.33).

Determination of Policy Categories, BCTs, and Practical Applications

The theoretical constructs were systematically translated into practical intervention strategies through integration of the following components: behavioral change outcomes, POs, determinants, change matrix, intervention functions, and policy categories. The resulting practical applications are comprehensively outlined in Multimedia Appendix 7.

Intervention Production

Draft Intervention Strategies

Through 2 rounds of expert panel discussions (n=19), the group determined that simple SMS text messaging reminders and social media–based support groups were particularly effective in reducing SB. These low-cost, highly acceptable intervention methods may be applicable in other resource-limited settings. Following two rounds of expert panel discussions involving 19 intervention group specialists, a comprehensive intervention strategy for SB in the community-dwelling older adults with T2DM was formulated. The strategy comprised 5 key components: eHealth education manual, MG animation library, SMS text messaging library, WeChat Q&A group, and material incentive package. The eHealth education manual, MG animation library, and SMS text messaging library targeted the "knowledge" determinant of SB. The WeChat Q&A group primarily addressed the "social support" determinant, while the material incentive package focused on the "intentions" determinant. Full descriptions of these intervention strategies are provided in Multimedia Appendix 8.

Feasibility Assessment of Intervention Strategies

Eight multidomain stakeholders (mean age of 44.75, SD 15.56 y) participated in the feasibility assessment, including 2 implementation science and evidence-based nursing or medicine stakeholders, 2 education stakeholders, 2 diabetes clinical nursing stakeholders, and 2 patient representatives. APEASE scoring results are presented in Multimedia Appendix 9. The eHealth education manual, MG animation library, SMS text messaging library, and material incentive package all achieved mean scores of 3.00 (SD 0.00), while the WeChat Q&A group scored 2.88 (SD 0.35).

Program Implementation Plan

Specific POs for intervention implementation are outlined in Table 3. These objectives were established by the research team based on the behavioral determinants identified in the logic model (Table 2). To achieve the adoption and implementation change objectives, the intervention design group selected a series of BCTs grounded in theory and integrated them with the 5 intervention components (Multimedia Appendix 10): eHealth education manual, MG animation library, SMS text messaging library, WeChat Q&A group, and material incentive package.

XSL•FO
RenderX

Table . Adoption and implementation outcomes and performance objectives (POs) for the intervention program.

PO	Context
PO1	The patient should comprehend the hazards of SB[a] (before behavior).
PO2	The patient should comprehend strategies to counteract SB, such as reminders and behavior reversal.
PO3	Family members should comprehend the hazards of SB (current situation, consequences, and its relationship with T2DM[b]).
PO4	The patient should comprehend how to engage in physical activity.

[a]SB: sedentary behavior.

[b]T2DM: type 2 diabetes mellitus.

Evaluation Results

A total of 31 community-dwelling older adults with T2DM (mean age 66.35, SD 3.94 years) completed the 7-week intervention. All 31 participants successfully completed both the baseline (T0) and postintervention (T1) assessments. Detailed participant characteristics are provided in Multimedia Appendix 11.

Following the implementation of the intervention, significant reductions in SB were observed among participants. Total sedentary time decreased by 1.12 hours per day compared to preintervention levels ($P<.001$). Notably, screen-based sedentary time showed a significant reduction of 0.33 hours per day ($P=.01$). However, no statistically significant changes were observed in sedentary time related to reading ($P=.71$), socializing ($P=.87$), transportation ($P=.32$), hobbies ($P=.85$), or other activities ($P=.56$).

The intervention also yielded improvements in social network scores as measured by the LSNS-6. Post-intervention LSNS-6 scores showed significant improvement ($P<.001$), with the family dimension demonstrating marked enhancement ($P<.001$). The friend dimension, however, did not show statistically significant changes ($P=.46$).

For participants with T2DM, no statistically significant changes were observed in morning blood pressure, fasting blood glucose, Diabetes-Specific Quality of Life scores, or BMI. No fall incidents were reported either before or during the intervention period.

Discussion

Principal Findings

This study aimed to develop and evaluate a theory-driven remote intervention strategy, “Double-S,” to reduce SB among older adults with T2DM in China. Our findings suggest that the systematic application of the IM framework is a feasible approach for developing targeted behavioral interventions. To our knowledge, this research represents one of the first theory-driven and evidence-based remote intervention strategies specifically targeting SB in older Chinese adults with T2DM in community settings. By integrating theoretical frameworks and addressing the specific needs of a rapidly aging population in a digitally evolving environment, this work contributes to the growing body of knowledge on culturally adapted, evidence-based interventions for chronic disease management in diverse global settings. Specifically, the intervention led to a promising reduction in total sedentary time by 1.12 hours per day and screen-based sedentary time by 0.33 hours per day. Furthermore, it significantly enhanced social network scores within the family dimension. However, these results should be interpreted as preliminary indicators of the program’s potential rather than definitive proof of efficacy, given the feasibility nature and limited sample size of this pilot evaluation.

Comparison With Prior Work and Implications

The core advantage of IM lies in its ability to explicitly link intervention strategies with theory, supporting hypothesis testing and deepening our understanding of causal mechanisms [33,50]. By addressing the gap in theory-grounded, culturally adapted digital health tools, our study offers a novel design contribution to the chronic disease management market in China and other similar settings. In diverse health care settings, IM is crucial for developing systematic approaches to use available resources and translate research findings into routine practice. Building on the BCW theory, this study constructed a problem logic model for SB among older adults with T2DM. The model integrates determinants (capability, motivation, and opportunity), the Chinese community context, the 3-stage SB trajectory, and health outcomes. By addressing this gap, our study offers new perspectives on improving behavior change interventions in diverse health care environments.

While IM has been widely applied in Western health care contexts [51-53], our research explores its applicability in Asian settings. Previous reviews [42] indicated that Western interventions predominantly target individual psychological aspects through face-to-face counseling or wearable devices.

However, SB is a habitual, unconscious behavior deeply influenced by cultural and environmental factors. The perception of SB and social norms varies across cultural contexts, affecting not only individual behavior but also health system design. Consequently, Western strategies need to be re-examined and adapted for Asian health care settings to ensure their effectiveness.

In community-based settings like China, high-cost individualized interventions pose challenges due to population aging and diabetes pressures [54]. Thus, our research shifted focus towards social and policy-level intervention strategies. Initial findings suggest that older adults in China often obtain health information

XSL•FO
RenderX

through public media, aligning with current science popularization policies. The high feasibility and acceptability scores from both multidisciplinary stakeholders and older users highlight significant opportunities for scaling up the "Double-S" program.

The widespread use of smartphones offers new channels for rapid health information dissemination. Remote interventions combined with digital tools provide advantages such as broad coverage and cost-effectiveness. Given the deep integration of digital devices into daily life, leveraging digital means for diabetes management has become increasingly necessary [55].

Beyond testing under controlled conditions, the preliminary evaluation results from both stakeholder and user perspectives provide a strong foundation for future implementation testing. This includes exploring how the "Double-S" program can be integrated into routine community health care systems for large-scale application. However, while the "Double-S" proved feasible under controlled conditions, real-world implementation may be constrained by structural barriers. This finding highlights potential challenges in transitioning from experimental settings to practical applications.

Digital infrastructure in rural areas and technological adaptability among older populations generally lag behind urban areas [56]. This can subject beneficiaries to "double exclusion" regarding technology access and service availability. Policymakers should prioritize bridging the gap between "technology popularization" and "institutional adaptation" rather than merely hardware upgrades.

Regarding effectiveness, the results showed significant improvements in social isolation but limited impact on physiological indicators like BMI or blood glucose. This aligns with findings from Ferguson et al [57] and Chastin et al [24], who indicated that metabolic improvements from SB reduction are typically modest and nonsignificant in the short term.

These findings reveal the dual nature of remote interventions: while cost-effective in enhancing social support networks, they may face challenges in reversing long-term metabolic disorders quickly. Heterogeneity in intervention intensity and duration likely compromises cross-study comparability, suggesting that determining core parameters within specific cultural contexts is crucial.

Limitations

This study has several limitations that should be acknowledged. First, the feasibility study used a convenience sampling method with a small sample size (N=31), which may limit the generalizability of the findings to the broader population with T2DM in China. Importantly, as this was a feasibility study, it lacked the statistical power to definitively assess clinical effectiveness; thus, the observed reductions in sedentary time should be viewed as preliminary. Second, the 7-week intervention period was relatively short, which likely explains the lack of significant changes in physiological and metabolic outcomes. Third, data collection relied on self-reported questionnaires, which are subject to recall bias. Future research should use randomized controlled trials with longer follow-up periods and objective measures of SB to further validate these findings.

Conclusions

The "Double-S" program demonstrates that theory-driven telehealth interventions can be successfully developed and adapted to specific cultural and technological contexts to address the public health challenge of SB. Beyond the immediate reduction in sedentary time, this research highlights the critical role of social media ecosystems and family-based social support in managing chronic diseases among older adults. Rather than providing definitive evidence of efficacy, this study offers a scalable design model and a validated methodological foundation for health care systems in UMICs. Successfully scaling such digital health solutions requires moving beyond hardware deployment toward institutional adaptation and the development of localized, culturally sensitive outcome sets. The positive outcomes from this pilot phase support the progression toward large-scale implementation trials to further validate the program's impact.

Acknowledgments

The authors would like to thank the health care professionals and patients who participated in the stakeholder consultations and pilot study.

Funding

This study was funded by the National High Level Hospital Clinical Research Funding (BJ-2024-198), the Fundamental Research Funds for the Central Universities (3332024098), and the Research Funding of the Chinese Nursing Association (ZHKYQ202416). The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Data Availability

All of the data are available upon reasonable request to the corresponding author.

Data Availability

All of the data are available on reasonable request to the corresponding author.

XSL•FO
RenderX

Authors' Contributions

XZ and DY conceptualized and designed the study, and are co-first authors of this paper. LW, XL, and JL participated in the investigation and data curation. The formal data analysis was conducted by XZ, DY, SC, ZW, MM, and HJ. Throughout the study, YH and JL were responsible for program administration and provided methodological guidance. XZ and DY wrote the original draft of the manuscript. SC and ZW reviewed and edited the manuscript. LW, XL, ZW, MM, XL, HJ, YH, and JL reviewed the manuscript and provided valuable revisions. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed recruitment and participant flow.
[DOCX File, 13 KB - jmir_v28i1e80827_app1.docx]

Multimedia Appendix 2
The matching relationship between the determinants identified by theoretical domains framework and the capability, opportunity, motivation, and behavior components.
[DOCX File, 19 KB - jmir_v28i1e80827_app2.docx]

Multimedia Appendix 3
Mapping relationships between behavioral determinants, intervention functions, and policy categories.
[DOCX File, 24 KB - jmir_v28i1e80827_app3.docx]

Multimedia Appendix 4
APEASE (Affordability, Practicability, Effectiveness/Cost-effectiveness, Acceptability, Side-Effects/Safety, Equity) criteria.
[DOCX File, 13 KB - jmir_v28i1e80827_app4.docx]

Multimedia Appendix 5
Three-stage trajectory of sedentary behavior.
[DOCX File, 439 KB - jmir_v28i1e80827_app5.docx]

Multimedia Appendix 6
Intervention function APEASE (Affordability, Practicability, Effectiveness/Cost-effectiveness, Acceptability, Side-Effects/Safety, Equity) score.
[DOCX File, 17 KB - jmir_v28i1e80827_app6.docx]

Multimedia Appendix 7
The practical application of intervention strategies converts results.
[DOCX File, 29 KB - jmir_v28i1e80827_app7.docx]

Multimedia Appendix 8
“Double-S” intervention strategy for older adult patients with type 2 diabetes.
[DOCX File, 18 KB - jmir_v28i1e80827_app8.docx]

Multimedia Appendix 9
“Double-S” program APEASE (Affordability, Practicability, Effectiveness/Cost-effectiveness, Acceptability, Side-Effects/Safety, Equity) score results.
[DOCX File, 15 KB - jmir_v28i1e80827_app9.docx]

Multimedia Appendix 10
The 5 components of the “Double-S” Plan.
[DOCX File, 1216 KB - jmir_v28i1e80827_app10.docx]

Multimedia Appendix 11

XSL•FO
RenderX

General patient information (N=31).
[DOCX File, 16 KB - jmir_v28i1e80827_app11.docx]

References

1. Alberti K, Zimmet PZ. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. Diabet Med 1998 Jul;15(7):539-553. [doi: 10.1002/(SICI)1096-9136(199807)15:7<539::AID-DIA668>3.0.CO;2-S] [Medline: 9686693]
2. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in diabetes prevalence and treatment from 1990 to 2022: a pooled analysis of 1108 population-representative studies with 141 million participants. Lancet 2024 Nov 23;404(10467):2077-2093. [doi: 10.1016/S0140-6736(24)02317-1] [Medline: 39549716]
3. Liu J, Bai R, Chai Z, Cooper ME, Zimmet PZ, Zhang L. Low- and middle-income countries demonstrate rapid growth of type 2 diabetes: an analysis based on Global Burden of Disease 1990-2019 data. Diabetologia 2022 Aug;65(8):1339-1352. [doi: 10.1007/s00125-022-05713-6] [Medline: 35587275]
4. Zhao W. Interpretation on clinical guidelines for prevention and treatment of type 2 diabetes mellitus in the elderly in China (2022 edition). Med J Peking Union Med Coll Hosp 2022;13(4):574-580. [doi: 10.12290/xhyxzz.2022-0199]
5. Ning X, Chen X. The new role of lifestyle interventions in the clinical prevention and treatment of type 2 diabetes. J Pract Med 2025;41(17):2611-2616. [doi: 10.3969/j.issn.1006-5725.2025.17.001]
6. Tremblay MS, Aubert S, Barnes JD, et al. Sedentary behavior research network (SBRN)–terminology consensus project process and outcome. Int J Behav Nutr Phys Act 2017 Jun 10;14(1):75. [doi: 10.1186/s12966-017-0525-8] [Medline: 28599680]
7. Lee IM, Shiroma EJ, Lobelo F, Puska P, Blair SN, Katzmarzyk PT. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. The Lancet 2012 Jul;380(9838):219-229. [doi: 10.1016/S0140-6736(12)61031-9]
8. Harvey JA, Chastin SFM, Skelton DA. Prevalence of sedentary behavior in older adults: a systematic review. Int J Environ Res Public Health 2013 Dec 2;10(12):6645-6661. [doi: 10.3390/ijerph10126645] [Medline: 24317382]
9. Li H, Zhang J, Gu R, Li H, Li L, Jin Y. Status and influencing factors of sedentary behavior in community-dwelling elderly diabetic patients. Chin Nurs Manage 2023;23(5):698-703. [doi: 10.3969/j.issn.1672-1756.2023.05.012]
10. Xu C, Furuya-Kanamori L, Liu Y, et al. Sedentary behavior, physical activity, and all-cause mortality: dose-response and intensity weighted time-use meta-analysis. J Am Med Dir Assoc 2019 Oct;20(10):1206-1212.e3. [doi: 10.1016/j.jamda.2019.05.001] [Medline: 31272857]
11. Cabanas-Sánchez V, Guallar-Castillón P, Higueras-Fresnillo S, Rodríguez-Artalejo F, Martínez-Gómez D. Changes in sitting time and cardiovascular mortality in older adults. Am J Prev Med 2018 Mar;54(3):419-422. [doi: 10.1016/j.amepre.2017.10.010] [Medline: 29305068]
12. Schmid D, Leitzmann MF. Television viewing and time spent sedentary in relation to cancer risk: a meta-analysis. J Natl Cancer Inst 2014 Jul;106(7):dju098. [doi: 10.1093/jnci/dju098] [Medline: 24935969]
13. Lee SH, Son C, Yeo S, Ha IH. Cross-sectional analysis of self-reported sedentary behaviors and chronic knee pain among South Korean adults over 50 years of age in KNHANES 2013-2015. BMC Public Health 2019 Oct 26;19(1):1375. [doi: 10.1186/s12889-019-7653-9] [Medline: 31655569]
14. Chastin SFM, Mandrichenko O, Helbostadt JL, Skelton DA. Associations between objectively-measured sedentary behaviour and physical activity with bone mineral density in adults and older adults, the NHANES study. Bone 2014 Jul;64:254-262. [doi: 10.1016/j.bone.2014.04.009] [Medline: 24735973]
15. Beunza JJ, Martínez-González MA, Ebrahim S, et al. Sedentary behaviors and the risk of incident hypertension: the SUN Cohort. Am J Hypertens 2007 Nov;20(11):1156-1162. [doi: 10.1016/j.amjhyper.2007.06.007] [Medline: 17954361]
16. Ohlsson C, Gidestrand E, Bellman J, et al. Increased weight loading reduces body weight and body fat in obese subjects—a proof of concept randomized clinical trial. EClinicalMedicine 2020 May;22:100338. [doi: 10.1016/j.eclinm.2020.100338] [Medline: 32510046]
17. Izquierdo M, Merchant RA, Morley JE, et al. International exercise recommendations in older adults (ICFSR): expert consensus guidelines. J Nutr Health Aging 2021;25(7):824-853. [doi: 10.1007/s12603-021-1665-8] [Medline: 34409961]
18. Katzmarzyk PT, Powell KE, Jakicic JM, et al. Sedentary behavior and health: update from the 2018 Physical Activity Guidelines Advisory Committee. Med Sci Sports Exerc 2019 Jun;51(6):1227-1241. [doi: 10.1249/MSS.0000000000001935] [Medline: 31095080]
19. Bull FC, Al-Ansari SS, Biddle S, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. Br J Sports Med 2020 Dec;54(24):1451-1462. [doi: 10.1136/bjsports-2020-102955] [Medline: 33239350]
20. Prince SA, Saunders TJ, Gresty K, Reid RD. A comparison of the effectiveness of physical activity and sedentary behaviour interventions in reducing sedentary time in adults: a systematic review and meta-analysis of controlled trials. Obes Rev 2014 Nov;15(11):905-919. [doi: 10.1111/obr.12215] [Medline: 25112481]
21. Martin A, Fitzsimons C, Jepson R, et al. Interventions with potential to reduce sedentary time in adults: systematic review and meta-analysis. Br J Sports Med 2015 Aug;49(16):1056-1063. [doi: 10.1136/bjsports-2014-094524] [Medline: 25907181]

XSL•FO
RenderX

22. Gomersall S, Maher C, English C, Rowlands A, Olds T. Time regained: when people stop a physical activity program, how does their time use change? A randomised controlled trial. PLoS ONE 2015;10(5):e0126665. [doi: 10.1371/journal.pone.0126665] [Medline: 26023914]
23. Melanson EL. The effect of exercise on non-exercise physical activity and sedentary behavior in adults. Obes Rev 2017 Feb;18(Suppl 1):40-49. [doi: 10.1111/obr.12507] [Medline: 28164451]
24. Chastin S, Gardiner PA, Harvey JA, et al. Interventions for reducing sedentary behaviour in community-dwelling older adults. Cochrane Database Syst Rev 2021 Jun 25;6(6):CD012784. [doi: 10.1002/14651858.CD012784.pub2] [Medline: 34169503]
25. Gentili A, Failla G, Melnyk A, et al. The cost-effectiveness of digital health interventions: a systematic review of the literature. Front Public Health 2022;10:787135. [doi: 10.3389/fpubh.2022.787135] [Medline: 36033812]
26. Howland C, Wakefield B. Assessing telehealth interventions for physical activity and sedentary behavior self-management in adults with type 2 diabetes mellitus: an integrative review. Res Nurs Health 2021 Feb;44(1):92-110. [doi: 10.1002/nur.22077] [Medline: 33091168]
27. China's digital health development report was released, and digital health community leads the new trend of the industry. Global Times. 2020 Jun 24. URL: https://www.huanqiu.com/article/3ymeOncYqfv [accessed 2025-12-30]
28. Cui F, Ma Q, He X, et al. Implementation and application of telemedicine in China: cross-sectional study. JMIR Mhealth Uhealth 2020 Oct 23;8(10):e18426. [doi: 10.2196/18426] [Medline: 33095175]
29. Skivington K, Matthews L, Simpson SA, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. Int J Nurs Stud 2024 Jun;154:104705. [doi: 10.1016/j.ijnurstu.2024.104705] [Medline: 38564982]
30. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implementation Sci 2009 Dec;4(1):50. [doi: 10.1186/1748-5908-4-50]
31. Proctor EK, Landsverk J, Aarons G, Chambers D, Glisson C, Mittman B. Implementation research in mental health services: an emerging science with conceptual, methodological, and training challenges. Adm Policy Ment Health 2009 Jan;36(1):24-34. [doi: 10.1007/s10488-008-0197-4] [Medline: 19104929]
32. Proctor EK, Powell BJ, McMillen JC. Implementation strategies: recommendations for specifying and reporting. Implement Sci 2013 Dec 1;8:139. [doi: 10.1186/1748-5908-8-139] [Medline: 24289295]
33. Bartholomew LK, Parcel GS, Kok G. Intervention mapping: a process for developing theory- and evidence-based health education programs. Health Educ Behav 1998 Oct;25(5):545-563. [doi: 10.1177/109019819802500502] [Medline: 9768376]
34. Fernandez ME, Ruiter RAC, Markham CM, Kok G. Intervention mapping: theory-and evidence-based health promotion program planning: perspective and examples. Front Public Health 2019;7:209. [doi: 10.3389/fpubh.2019.00209] [Medline: 31475126]
35. De Decker E, De Craemer M, De Bourdeaudhuij I, et al. Using the intervention mapping protocol to reduce European preschoolers' sedentary behavior, an application to the ToyBox-Study. Int J Behav Nutr Phys Act 2014 Feb 19;11:19. [doi: 10.1186/1479-5868-11-19] [Medline: 24552138]
36. Huang Y, Benford S, Price D, et al. Using Internet of Things to reduce office workers' sedentary behavior: intervention development applying the behavior change wheel and human-centered design approach. JMIR Mhealth Uhealth 2020 Jul 29;8(7):e17914. [doi: 10.2196/17914] [Medline: 32723716]
37. Ezeugwu VE, Manns PJ. Using intervention mapping to develop and implement a home-based sedentary behavior change intervention after stroke. Transl Behav Med 2020 Feb 3;10(1):87-95. [doi: 10.1093/tbm/iby128] [Medline: 30566661]
38. Direito A, Walsh D, Hinbarji M, et al. Using the intervention mapping and behavioral intervention technology frameworks: development of an mHealth intervention for physical activity and sedentary behavior change. Health Educ Behav 2018 Jun;45(3):331-348. [doi: 10.1177/1090198117742438] [Medline: 29216765]
39. Zhang X, Yang D, Luo J, et al. Determinants of sedentary behavior in community-dwelling older adults with type 2 diabetes based on the behavioral change wheel: a path analysis. BMC Geriatr 2024 Jun 6;24(1):502. [doi: 10.1186/s12877-024-05076-0]
40. Kok G, Peters LWH, Ruiter RAC. Planning theory- and evidence-based behavior change interventions: a conceptual review of the intervention mapping protocol. Psicol Reflex Crit 2017 Oct 18;30(1):19. [doi: 10.1186/s41155-017-0072-x] [Medline: 32026109]
41. Michie S, van Stralen MM, West R. The Behaviour Change Wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
42. Zhang X, Li X, Yang D, Meng M, Hao Y. Intervention strategies for reducing sedentary behavior among older adults:a scoping review. Chin J Nurs 2024;59(8):1012-1020. [doi: 10.3761/j.issn.0254-1769.2024.08.017]
43. National Center of Gerontology, Chinese Diabetes Society, China Sport Science Society. Guideline for exercise therapy of type 2 diabetes mellitus in China. Chin J Sports Med 2024;43(6):419-452. [doi: 10.12114/j.issn.1007-9572.2024.A0019]
44. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]

XSL•FO
RenderX

45. Global strategy on digital health 2020-2025. : World Health Organization; 2021 URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2026-02-27]
46. Ye H, Zhu X, Zhu X, Lin Y, Zhuang J. Evaluation of reliability, validity and reactivity of Diabetes Specific Quality of Life scale in rural-urban fringe diabetic patients. J Nurs 2009;16(24):60-61. [doi: 10.16460/j.issn1008-9969.2009.24.026]
47. Chang Q, Sha F, Chan CH, Yip PSF. Validation of an abbreviated version of the Lubben Social Network Scale ("LSNS-6") and its associations with suicidality among older adults in China. PLoS ONE 2018;13(8):e0201612. [doi: 10.1371/journal.pone.0201612] [Medline: 30071067]
48. Yang K. Status of social network and its influencing factors among elderly patients with type 2 diabetes in Beijing. J Nurs Sci 2020;35(10):94-97. [doi: 10.3870/j.issn.1001-4152.2020.10.094]
49. Zhang X. Sedentary behavior mechanism and intervention strategies in community-dwelling elderly patients with T2DM [PhD Thesis]. : Beijing University of Chinese Medicine; 2018.
50. Lewis CC, Klasnja P, Powell BJ, et al. From classification to causality: advancing understanding of mechanisms of change in implementation science. Front Public Health 2018;6:136. [doi: 10.3389/fpubh.2018.00136] [Medline: 29868544]
51. Fernandez ME, Ten Hoor GA, van Lieshout S, et al. Implementation mapping: using intervention mapping to develop implementation strategies. Front Public Health 2019;7:158. [doi: 10.3389/fpubh.2019.00158] [Medline: 31275915]
52. Hurley DA, Murphy LC, Hayes D, et al. Using intervention mapping to develop a theory-driven, group-based complex intervention to support self-management of osteoarthritis and low back pain (SOLAS). Implementation Sci 2016 Dec;11(1):56. [doi: 10.1186/s13012-016-0418-2]
53. Mirzaei-Alavijeh M, Jouybari TA, Jalilian F, Motlagh ME, Jalilian F. Using intervention mapping approach to finding socio-cognitive determinants of diabetes preventive behaviors. J Prev Med Hyg 2019 Sep;60(3):E237-E242. [doi: 10.15167/2421-4248/jpmh2019.60.3.1159] [Medline: 31650060]
54. Yan Y, Wu T, Zhang M, Li C, Liu Q, Li F. Prevalence, awareness and control of type 2 diabetes mellitus and risk factors in Chinese elderly population. BMC Public Health 2022 Jul 19;22(1):1382. [doi: 10.1186/s12889-022-13759-9] [Medline: 35854279]
55. Sriram RD, Reddy SSK. Artificial intelligence and digital tools: future of diabetes care. Clin Geriatr Med 2020 Aug;36(3):513-525. [doi: 10.1016/j.cger.2020.04.009] [Medline: 32586478]
56. Ming Y, Li Y, Liu Y. Digital technologies as solutions to China's aging population: a systematic review of their opportunities and challenges in rural development. Front Public Health 2024;12:1416968. [doi: 10.3389/fpubh.2024.1416968] [Medline: 39882121]
57. Ferguson T, Olds T, Curtis R, et al. Effectiveness of wearable activity trackers to increase physical activity and improve health: a systematic review of systematic reviews and meta-analyses. Lancet Digit Health 2022 Aug;4(8):e615-e626. [doi: 10.1016/S2589-7500(22)00111-X] [Medline: 35868813]

Abbreviations

APEASE: Affordability, Practicability, Effectiveness/Cost-effectiveness, Acceptability, Side Effects/Safety, Equity
BCT: behavior change technique
BCW: Behavior Change Wheel
COM-B: capability, opportunity, motivation, and behavior
IM: intervention mapping
LSNS-6: Lubben Social Network Scale-6
MG: motion graphics
PO: performance objective
SB: sedentary behavior
T2DM: type 2 diabetes mellitus
TDF: theoretical domains framework
UMIC: upper-middle-income country
WHO: World Health Organization

Edited by A Stone; submitted 17.Jul.2025; peer-reviewed by IB Felix, LF Belo; revised version received 12.Feb.2026; accepted 18.Feb.2026; published 26.Mar.2026.

Please cite as:
Zhang X, Yang D, Chen S, Wang L, Li X, Wang Z, Meng M, Li X, Jiang H, Hao Y, Luo J
Telehealth Intervention to Reduce Sedentary Behavior in Older Adults With Type 2 Diabetes: Development and Feasibility Study
J Med Internet Res 2026;28:e80827
URL: https://www.jmir.org/2026/1/e80827
doi:10.2196/80827

XSL•FO
RenderX

© Xiaoyan Zhang, Dan Yang, Sihan Chen, Lei Wang, Xin Li, Ziyan Wang, Meiqi Meng, Xuejing Li, Hongzhan Jiang, Yufang Hao, Jiayin Luo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Association Between Physician Communication Features and Patient Outcomes in Telemedicine: Retrospective Cross-Sectional Observational Study

Yipei Wang[1,2*], MBA, PhD; Shu Wang[3*], BA; Ke Zhang[1,4*], MD; Zhijie Liu[5]; Qingbian Ma[6], MD; Hong Ji[7], MS; Zheng Hou[1,8], MD; Tracy Xiao Liu[3], PhD; Xuedong Xu[9], MBA; Xinxia Wu[1], MS; Changxiao Jin[2], MS

[1]Department of Medical Affairs, Office of Internet Hospital, Peking University Third Hospital, Beijing, China
[2]Institute of Hospital Management, Peking University Third Hospital, 49 North Huayuan Road, Haidian District, Beijing, China
[3]Department of Economics, School of Economics and Management, Tsinghua University, Beijing, China
[4]Department of Otolaryngology, Peking University Third Hospital, Beijing, China
[5]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
[6]Department of Emergency Medicine, Peking University Third Hospital, Beijing, China
[7]Information Management and Big Data Center, Peking University Third Hospital, Beijing, China
[8]Department of Obstetrics and Gynecology, Peking University Third Hospital, Beijing, China
[9]Peking University People's Hospital, Beijing, China
[*]these authors contributed equally

Corresponding Author:
Changxiao Jin, MS
Institute of Hospital Management, Peking University Third Hospital, 49 North Huayuan Road, Haidian District, Beijing, China

Abstract

Background: Asynchronous telemedicine is a crucial component of multichannel health care, where effective communication drives satisfaction. However, the effectiveness of communication features remains poorly understood. Prior research relied on subjective surveys or small-scale simulations, failing to link features to objective outcomes. Understanding these features is critical for optimizing physician engagement and establishing quality indicators to enhance the patient experience.

Objective: This study aimed to bridge this gap by leveraging a large-scale real-world dataset to quantify the association between physicians' communication features—including response modalities, length, and sequence—and patient repurchase behavior, as well as review scores, within a high-autonomy health care setting.

Methods: This retrospective cross-sectional study analyzed 304,337 paid, patient-initiated virtual visits from a Chinese academic medical center (2021 - 2023), which included 823,135 physician responses. The sample was selected after applying a series of exclusion criteria, such as free consultations, team-based visits, and outlier data. The key exposures were the modality of physician responses, response length, and response sequence. Outcomes included patient loyalty and satisfaction. Loyalty was operationalized as follow-up visits within 6 months, with a 30-day exclusion period applied to same-physician (*fv1*) and same-department (*fv2*) revisits to filter out clinical necessity, but not to hospital-wide revisits (*fv3*). Satisfaction was measured by the review scores. We used probit and ordinary least squares regressions to examine the relationships between communication features and patient outcomes.

Results: Regarding loyalty, audio-only visits were associated with the lowest *fv1*, with an average marginal effect (AME) of −0.030 (95% CI −0.043 to −0.016, $P<.001$), translating to a 30.9% (0.030/0.097) reduction compared to text-only visits. Regarding satisfaction, audio messages were associated with a significantly increased likelihood of patients providing reviews, with an AME of 0.041 (95% CI 0.006 - 0.076, $P=.02$), but they did not affect review scores after adjusting for inverse Mills ratios. Increased numbers of text and audio replies were (marginally) associated with improved *fv1*, with AMEs of 0.009 (95% CI 0.006 - 0.011, $P<.001$) and 0.007 (95% CI −0.000 to 0.016, $P=.06$), respectively. Visits beginning with a sub-5-second audio response and ending with text had significantly higher *fv1* than text-only visits, with an AME of 0.069 (95% CI 0.018 - 0.120, $P=.008$). The same patterns hold for *fv2* and *fv3*. Based on the Bonferroni test, coefficients with a P value smaller than $\alpha=.050/3=.017$ or $\alpha=.50/2=.025$ were regarded as significant when evaluating the association with patient loyalty or satisfaction, respectively.

Conclusions: Physician communication practices were significantly associated with patient loyalty and satisfaction. This study is innovative in leveraging large-scale real-world data to systematically examine physician communication. It differs from existing studies by transcending prior survey-based research limitations. It introduces an effective hybrid approach, balancing human

XSL•FO
RenderX

connection with text clarity in the field. Its implication in the real world is providing data-driven evidence to guide clinicians and policymakers in designing high-quality telemedicine services.

(J Med Internet Res 2026;28:e86977) doi:10.2196/86977

KEYWORDS

telemedicine; asynchronous visits; physician-patient communication; patient loyalty; patient satisfaction; behavioral pattern; quality of virtual care; health-seeking behavior

Introduction

Telemedicine has emerged as a crucial component of the multichannel service delivery [1-3], significantly expanding access to care beyond the COVID-19 pandemic [4-6]. While granting increased convenience [7,8], enhanced accessibility [9,10], and improved clinical outcomes [11,12], the shift to virtual settings strips away physical interactions, making the communication skills of physicians essential, as they positively influence patient satisfaction and engagement [13-16]. In the specific context of asynchronous telemedicine, where visual cues are lacking, physicians' choices of text and/or audio messages become critical. Physicians face a dilemma: text messages offer clarity and ease of review, whereas audio communication enables faster expression and conveys a sense of human presence [17-19]. Striking the right balance between these modalities is essential for establishing a robust physician-patient connection [20,21].

Nonetheless, few studies have comprehensively investigated which features of physician response behaviors most effectively enhance the quality of virtual care. Prior research on this topic predominantly focused on in-person visit settings, examining the influence of computer-based electronic medical record writing and patient demographic factors on physician-patient interactions [22,23]. In the context of telemedicine, studies have explored the effect of audio use on satisfaction [24], yielding inconsistent results [18]. Some studies have quantified the length of physician responses and examined its impact on diagnostic accuracy and patient satisfaction using standardized patients [25,26]. Although these studies have value, they exhibit several shortcomings, including the subjectivity of patient satisfaction as an indicator, the artificiality of the standardized patient approach, and constraints on disease categories and sample sizes. Taken together, there is a scarcity of large-scale real-world evidence linking physician communication features to objective behavioral outcomes such as patient loyalty.

The Chinese health care context offers a unique setting to examine patient loyalty due to its high degree of consumer autonomy. Unlike systems with strict gatekeeping, China lacks a mandatory primary care referral system, granting patients direct access to specialists [27-29]. Furthermore, the government-led universal health insurance eliminates restrictive provider networks, allowing patients to seek care from any public institution without facing out-of-network penalties [30]. This autonomy is particularly pronounced in the asynchronous telemedicine examined in this study. In contrast to synchronous virtual visits, which are typically scheduled by physicians for postdischarge or routine follow-up visits based on clinical necessity, asynchronous consultations are patient-initiated and on-demand. Patients voluntarily seek advice, and physicians respond during flexible intervals. Therefore, unconstrained by either insurance policies or provider-mandated appointment requirements, the asynchronous setting effectively isolates patient behavior from structural and clinical constraints. Consequently, within the context of asynchronous consultations in China, patient repurchase acts as a revealed preference, driven predominantly by the overall care experience rather than by insurance constraints or referral mandates.

Our study leverages this unique setting to address the research gap by systematically investigating the association between physician communication features and patient outcomes. The primary objective of this cross-sectional study is to examine how response modality, length, and sequence relate to objective patient loyalty and subjective patient satisfaction. By using large-scale real-world data from 304,337 asynchronous visits at an academic medical center (AMC) in China, we aim to elucidate the distinct utility of these communication features and identify effective engagement patterns. Ultimately, this work is expected to establish an empirical basis for defining quality indicators and improving the patient experience of telemedicine.

Methods

Inclusion and Exclusion Criteria

This cross-sectional study was designed and reported in accordance with the Journal Article Reporting Standards (JARS) [31]. Our study used all administrative records (368,317 visits) of telemedicine visits from Peking University Third Hospital, a large AMC in China, covering the period from January 2021 to December 2023. These records were directly exported from the hospital information system (HIS) of the AMC. In this study, a "visit" refers to a complete, billable asynchronous consultation episode, distinct from a single message or reminder. To reflect the dynamics of individual physicians responding to patients' requests on a daily basis, the following criteria were applied (Multimedia Appendix 1): (1) free telemedicine visits provided by physicians on a voluntary basis were excluded, as physicians' behavior might change (n=35,804); (2) visits responded to by a medical team, such as multidisciplinary treatments, and those by physicians in diagnostic and therapeutic service departments, were excluded (n=2217); (3) visits initiated by physicians were excluded (n=97); (4) visits where physicians only prescribed medication without replying with a response were excluded (n=7075); (5) to ensure that our results were not skewed by specific physicians, we further excluded visits of "outlier physicians" from the analysis. Specifically, we focused on high-volume physicians, defined as those whose visit counts exceeded the median. We then analyzed the distributions of

XSL•FO
RenderX

response volume (number of messages) and duration (length of text/audio). Upon observing structural discontinuities at the extreme tails of these distributions, we applied a gap-based exclusion criterion: the top-ranked physician was removed if their mean value was more than twice that of the second-ranked physician. This conservative threshold was chosen to eliminate extreme anomalies likely driven by nonstandard usage patterns (eg, systematic templates) that could act as high leverage points and skew the regression estimates, while retaining valid high-performing physicians (n=12,086); and (6) given our research interest, we excluded visits containing image responses sent by physicians (n=6701). Altogether, the final dataset contained 304,337 (82.63% of 368,317 typical visits) telemedicine visits provided by 781 unique physicians, requested by 140,466 unique patients.

Participant Characteristics and Dataset Description

The 781 unique physicians exhibited a balanced gender distribution, with the average age in the mid-forties. Around 85% of individuals were from the fields of internal medicine, surgery, and obstetrics & gynecology. The proportions of different professional titles, such as chief, deputy chief, and attending physicians, were approximately equal (Table 1). Over the 3-year period, the average number of asynchronous visits per patient was 2.069, 1.810, and 1.786 per year. Over 60% of patients attended only once each year (Table 1).

XSL•FO
RenderX

Table . Baseline characteristics at the visit level (N=304,337), physician level (N=781), and patient level (N=140,466) at Peking University Third Hospital, Beijing, China (2021 - 2023).

	2021	2022	2023
Physician level			
Total, n (%)	582 (100)	698 (100)	706 (100)
Field, n (%)			
Obstetrics and gynecology	70 (12)	90 (12.9)	94 (13.3)
Internal medicine	229 (39.3)	280 (40.1)	279 (39.5)
Surgery	185 (31.8)	214 (30.7)	226 (32)
Pediatrics	27 (4.6)	30 (4.3)	28 (4)
Ophthalmology and otorhino-laryngology	71 (12.2)	84 (12)	79 (11.2)
Position at the year, n (%)			
Attending physician	149 (25.6)	224 (32.1)	226 (32)
Deputy chief physician	239 (41.1)	250 (35.8)	252 (35.7)
Chief physician	194 (33.3)	224 (32.1)	228 (32.3)
Gender, n (%)			
Female	303 (52.1)	353 (50.6)	358 (50.7)
Male	279 (47.9)	345 (49.4)	348 (49.3)
Age at the year, mean (SD)	44,735 (8.235)	44,838 (8.381)	45,072 (8.573)
Number of visits per physician, mean (SD)	145,847 (495.547)	145,479 (413.860)	167,011 (503.085)
Patient level			
Total, n (%)	41,033 (100)	56,093 (100)	66,013 (100)
Number of visits, n (%)			
1	24,786 (60.4)	37,733 (67.3)	44,669 (67.7)
2	7268 (17.7)	9011 (16.1)	10,643 (16.1)
3	3416 (8.3)	3815 (6.8)	4443 (6.7)
4	1885 (4.6)	1951 (3.5)	2284 (3.5)
5	1191 (2.9)	1175 (2.1)	1336 (2)
6	757 (1.8)	781 (1.4)	815 (1.2)
≥7	1730 (4.2)	1627 (2.9)	1823 (2.8)
Visit level			
Total, n (%)	84,883 (100)	101,544 (100)	117,910 (100)
Field, n (%)			
Obstetrics and gynecology	47,863 (56.4)	47,980 (47.3)	53,948 (45.8)
Internal medicine	18,622 (21.9)	27,356 (26.9)	28,554 (24.2)
Surgery	13,080 (15.4)	16,733 (16.5)	27,185 (23.1)
Pediatrics	3428 (4)	6192 (6.1)	4417 (3.7)
Ophthalmology and otorhino-laryngology	1890 (2.2)	3283 (3.2)	3806 (3.2)
Position at the year, n (%)			
Attending physician	11,041 (13)	12,685 (12.5)	15,417 (13.1)
Deputy chief physician	26,011 (30.6)	30,704 (30.2)	38,440 (32.6)
Chief physician	47,831 (56.3)	58,155 (57.3)	64,053 (54.3)

XSL•FO
RenderX

	2021	2022	2023
Total text length, mean (SD)	70.564 (95.751)	69.445 (92.141)	69.876 (95.651)
Total audio length, mean (SD)	24.261 (41.606)	24.411 (43.024)	22.078 (40.886)

Around half of the 304,337 visits pertained to obstetrics and gynecology, whereas roughly a quarter were related to internal medicine. The Department of Reproductive Medicine in the field of obstetrics and gynecology had the highest number of asynchronous visits, making up around 40% of the total visits (Figure 1A and Multimedia Appendix 2).

Figure 1. Study overview and distribution of 823,135 physician responses across 304,337 retrospective asynchronous telemedicine visits at Peking University Third Hospital, Beijing, China (2021 - 2023).

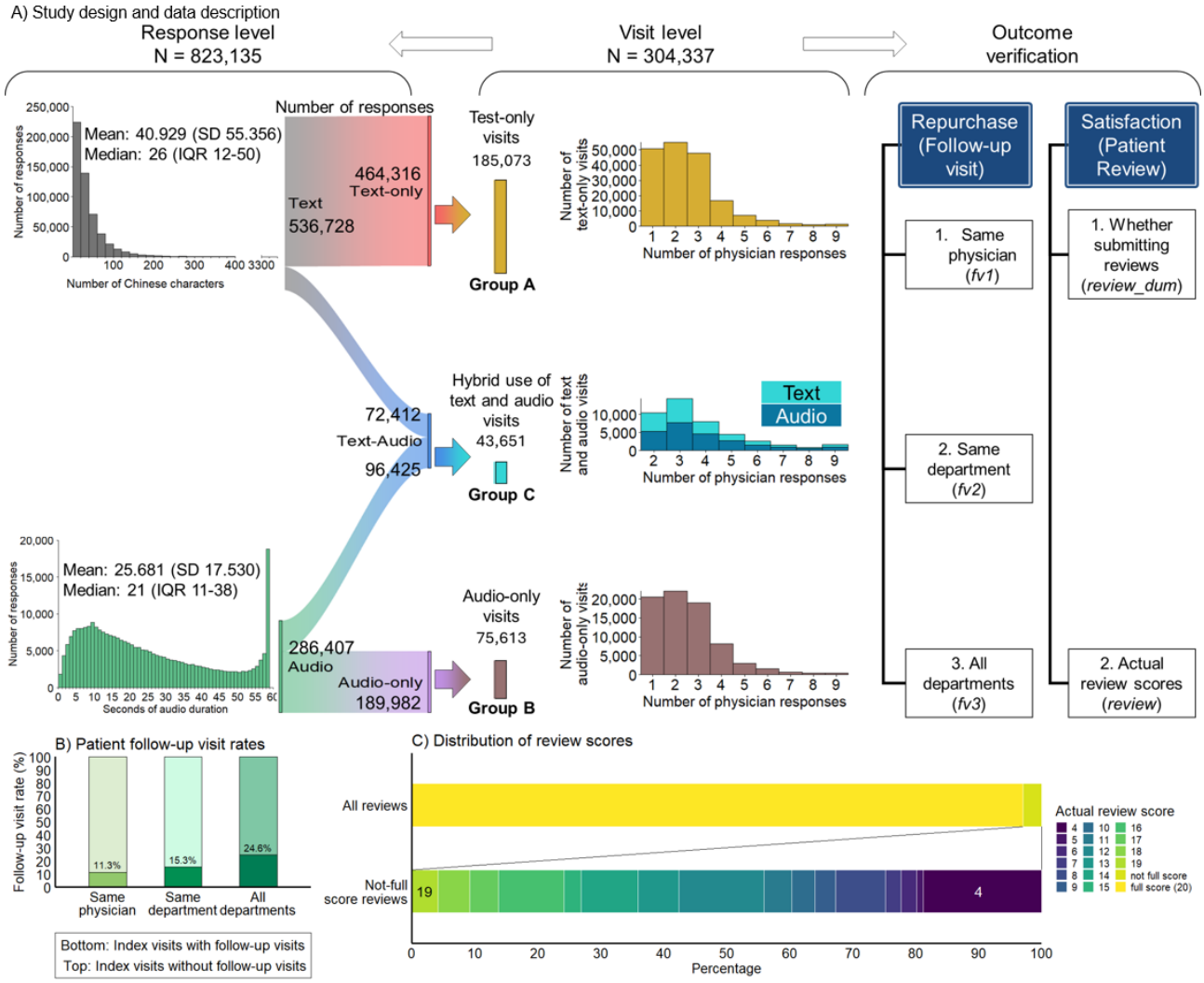

Between 2021 and 2023, the number of visits exhibited a considerable compound annual growth rate of 17.9%. A modest rise in the number of physicians offering asynchronous treatment was observed, with a notable surge in 2022. In 2023, there was a 14.8% (21,533/145,478) rise in the average workload per physician, with significant heterogeneity noticed among individuals.

These 304,337 visits included 823,135 physician responses, translating into 2.705 responses per visit. Of all 823,135 responses, 65.2% (n=536,728) were delivered in text modality and 34.8% (n=286,407) in audio modality. Text responses had no word restriction; however, each audio response was restricted to a maximum duration of 59 seconds, similar to the widely used instant communication application WeChat in China. On average, each text response contained 40.929 Chinese characters, whereas each audio response lasted 25.681 seconds (Figure 1A).

A total of 464,316 text responses formed 185,073 text-only visits, 189,982 audio responses formed 75,613 audio-only visits, and the additional 43,651 visits included both text and audio information. Regardless of the modalities, most visits consisted of no more than 3 physician responses (Figure 1A). Among all index visits, the follow-up visit rates within 6 months were 11.3% (13,937/123,504), 15.3% (15,203/99,126), and 24.8% (19,791/80,557) when considering the scope of the same physician (*fv1*), same department (*fv2*), and all departments (*fv3*) (Figure 1B). Additionally, the percentage of visits reviewed by patients (*review_dum*) was 5% (15,268/304,337), with an

XSL•FO
RenderX

average score (*review*) of 19.299 (SD 2.981) out of 20 (Figure 1C).

Sampling Procedures

This study used a retrospective cross-sectional observational design. All visits that met the predefined eligibility criteria during the study period were included. Therefore, no sampling was performed; instead, the study used a census of all eligible visit records. Accordingly, there were no self-selection concerns in our sample. The dataset was obtained directly from the hospital in Beijing in April 2024 and was analyzed subsequently.

Sample Size, Power, and Precision

As this was a retrospective study using existing data from HIS, a formal a priori power analysis was not conducted. The sample size was determined by the data in the Peking University Third Hospital within the specified period (ie, January 2021 to December 2023). All eligible records within this timeframe were included to maximize statistical power. Thus, the intended sample size was 368,317 visits. Following the inclusion and exclusion criteria, we achieved an actual sample size of 304,337 visits. The difference resulted from the exclusion process (steps 1 - 6 described in Multimedia Appendix 1).

With this substantial sample size, the study is adequately powered to conduct statistical analyses, ensuring high precision in parameter estimates.

Measures and Covariates

Our study focused on 3 features of physician responses, including the modality of physician responses (text-only, audio-only, or hybrid), response length, and response sequence. Two outcomes (patient loyalty and patient satisfaction) were used to examine the consequences of these physician characteristics.

Outcome: Patient Loyalty

Revisit intention is commonly used to measure consumers' recognition of a product or service and has been studied across various fields, such as tourism, catering, and health care [32-34]. However, revisit intention is a subjective measure that may lack accuracy. Thus, we focused on actual revisits, which provide an objective description of behavior. To distinguish behavioral loyalty from clinical dependency and avoid confusion with the negative quality indicator "revisit," we specifically used "follow-up visit" as a proxy for patient loyalty.

This measurement strategy is grounded in 2 key premises. First, revisits within the acute phase (eg, 14, 21, or 30 d) often serve as negative indicators in health care, reflecting unresolved medical issues or unmet expectations requiring additional treatment in a relatively short period of time [35-41]. Second, in this setting where patients have sufficient autonomy to choose providers for ongoing care, a voluntary return after the acute phase reflects a revealed preference rather than a lack of alternatives [42]. In this context, a follow-up visit is defined as the patient reusing the asynchronous service after a specific exclusion period following the index visit. Based on this rationale, a subsample of index visits was constructed to accurately capture this long-term behavioral loyalty.

To do so, we defined 2 dimensions—period and scope—to help formulate the concept of an index visit and a follow-up visit. The period was considered to be 6 months (182 d), which was the typical timeframe that the Clinician & Group Survey used to assess the lasting effects of a visit [43,44]. Meanwhile, the scope varied across different measurements of visit loyalty. Specifically, the scopes of (1) the same physician, (2) the same department, and (3) all departments applied to the 3 measures: follow-up visit 1 (*fv1*), follow-up visit 2 (*fv2*), and follow-up visit 3 (*fv3*), respectively. The 3 subsamples of index visits were thus defined as the visits without any visit with the same physician, in the same department, or in all departments in the 6 months preceding them (Figure 2).

XSL•FO
RenderX

Figure 2. Schematic illustration of index and follow-up visits used to measure patient loyalty across 3 hierarchical scopes: same physician (*fv1*, light blue circle), same department (*fv2*, dark blue circle), and same hospital (*fv3*, gold circle).

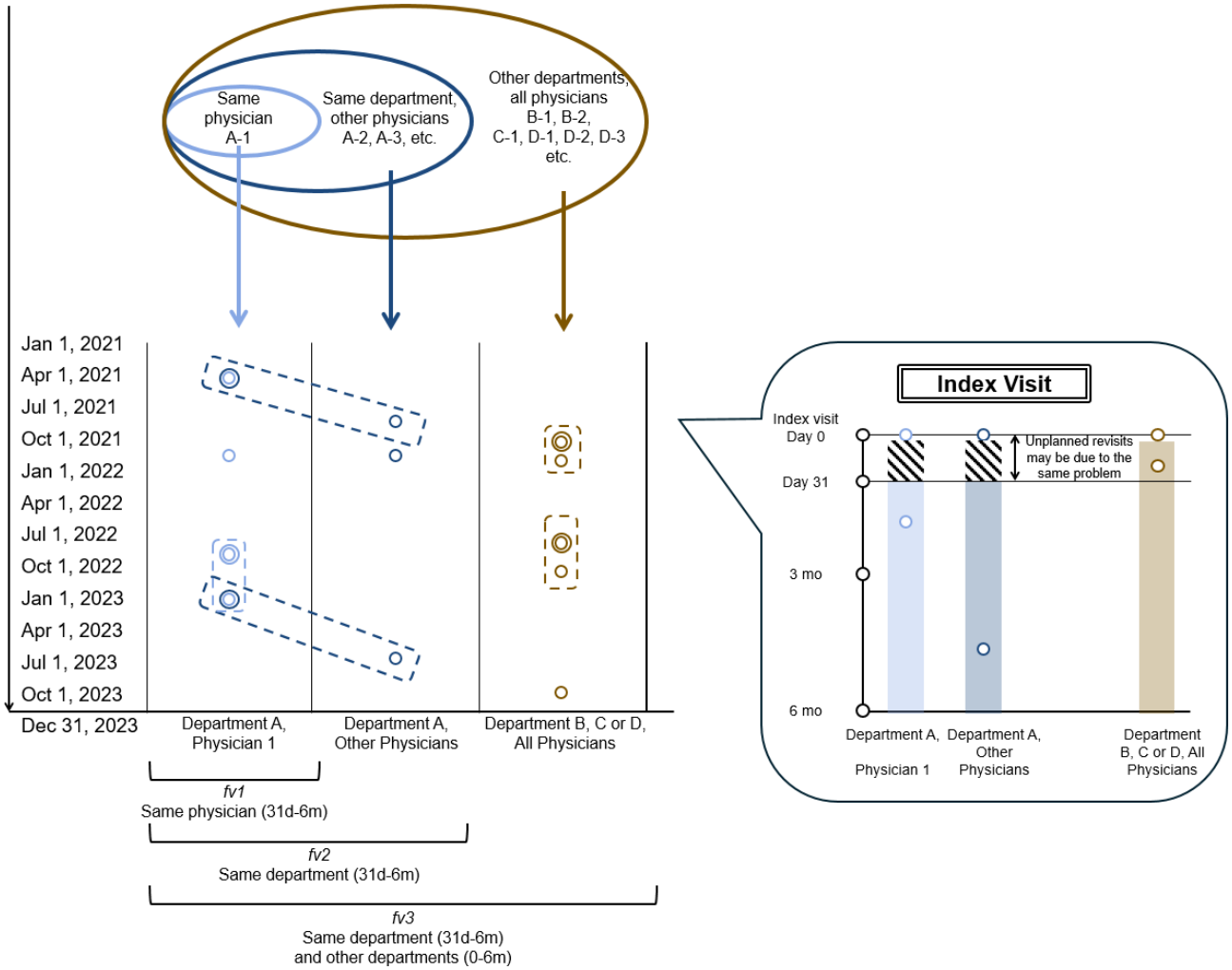

If there was at least one follow-up visit with the same physician, within the same department, or across all departments within 6 months after the corresponding index visit, then *fv1*, *fv2*, or *fv3* equaled 1; otherwise, it was equal to 0. Meanwhile, index visits followed by a visit to the same physician within the subsequent 30 days were excluded for *fv1*, whereas those followed by a visit within the same department were excluded for *fv2* and *fv3*, as immediate follow-up visits were often regarded as a signal of poor quality [35-41]. Note that when the scope was extended to all departments (ie, *fv3*), only follow-up visits within the same department within 30 days were excluded, whereas cross-department visits were retained, as these cross-department follow-up visits were unlikely to reflect poor quality of care and were more likely to indicate new demand. Finally, to ensure we could observe the entire 6-month period before and after the index visits, we excluded 2 types of index visits for all 3 measurements: (1) those in the last 6 months of our sample period and (2) those in the first 6 months of our sample period. In this way, we obtained 123,504, 99,126, and 80,556 index visits for *fv1*, *fv2*, and *fv3*, respectively.

Outcome: Patient Satisfaction

Patient satisfaction is commonly used as an evaluation metric for services in health care settings [45]. Listed as 1 of the 5 domains of the Quintuple Aim, patient satisfaction is strongly related to quality [46-48]. Furthermore, research has demonstrated that it is positively associated with patients' intention to revisit [34,45,49].

In our study, patients could review their experience on a voluntary basis after having asynchronous visits. The review asked the patients to rate their experience on a scale ranging from a minimum of 4 points to a maximum of 20 points. Based on this, 2 measurements were constructed: a dummy variable capturing whether the patient submitted a review (*review_dum*) and the corresponding review score conditional on the submission (*review*).

Control Variables

We defined some other variables to describe the characteristics of physicians and added them as control variables in the regression analyses. Based on the classification of departments in the sample hospital and different disease types, all departments were divided into internal medicine, surgery, obstetrics and gynecology, pediatrics, and ophthalmology and otorhinolaryngology (including ophthalmology, otolaryngology [ENT], and stomatology). Each field was defined as a dummy variable. Furthermore, 3 dummy variables were defined for attending physicians, deputy chief physicians, and chief physicians, respectively. Additionally, we counted the number

XSL•FO
RenderX

of Chinese characters and the seconds of audio duration by physicians per visit as control variables.

Data Collection

Data were directly extracted from the HIS of Peking University Third Hospital. The dataset integrates administrative records (eg, timestamps and payment status) and clinical contents (eg, message modality and response length).

Quality of Measurements

We used all data exactly as recorded in the system. This direct-output process did not involve human data collectors, thus eliminating any potential bias or intercollector variability associated with manual data collection.

Instrumentation

In this study, we developed a set of instruments to comprehensively assess patient loyalty and satisfaction. Loyalty was evaluated through follow-up visits, defined within different periods and scopes. Satisfaction was measured based on the raw review score provided by the patient, including whether a review was submitted and the review score itself.

Masking

Because this was a retrospective study rather than a randomized controlled trial, blinding was not applicable, as data collection occurred after clinical interactions had been completed.

Conditions and Design

This study used a nonexperimental design (ie, no experimental manipulation). As an observational design, it used retrospective data.

Data Diagnostics

There were no missing data in the dataset, ensuring a complete set of records for analysis. All visit-level continuous variables were censored at the top 1% to alleviate concerns related to outliers.

Analytic Strategy

All statistical analyses and visualizations were performed using Stata (MP 18.0), R (R/2024.04.2+764), and Origin (2024bSr0H).

To examine the role of the response modality, we split the whole sample into 3 subsamples: text-only visits (Group A), audio-only visits (Group B), and visits with both modalities (Group C) (Figure 1A). We also created 3 dummy variables to indicate each subsample, respectively. In the regression, we included 2 dummies for Groups B and C, as well as control variables for physician title (a dummy for deputy chief physicians and another for chief physicians). Moreover, the day-of-week fixed effect was added to capture features specific to the day of the week, such as weekdays and weekends. The year-month fixed effect was added to capture month-specific shocks common to all physicians, such as seasonal fluctuations in workload. The physician fixed effect was also included to capture individual behavioral habits. Standard errors were clustered at the physician level. Probit regressions were applied when the dependent variables were *fv1*, *fv2*, *fv3*, and *review_dum* using the Stata command *probit*, and the average marginal effects (AMEs) were reported. Ordinary least squares regressions were applied when the dependent variable was *review* using the Stata command *reghdfe*. Notably, only a subsample of visits contained reviews, which might lead to sample selection concerns. To account for such bias, we followed the standard procedure of the Heckman two-step method [50] by calculating the inverse Mills ratio and adding it into the regression in the second step. To further examine the differences between visits with both text and audio responses and those with only audio responses, we used the Wald test with the Stata command *test* after the regression to examine whether the coefficients of being in Group B and Group C were equivalent.

In a similar vein, we conducted regression analysis using the same models and dependent variables to evaluate the extent to which the length of responses affects patient loyalty and patient satisfaction. Specifically, we included the number of text responses (*text_num*) and the number of audio responses (*audio_num*) as the main independent variables and controlled for the total Chinese characters and total audio duration. Additionally, we added the same set of control variables and fixed effects as in the analysis regarding the response modality. Moreover, as a companion exercise, we further considered the extreme cases of visits with only audio responses, where the total audio duration falls within the range of 55 - 59 seconds (Group B1). We investigated whether splitting long audio responses mattered by including *audio_num* while controlling for the total audio duration.

We next considered the role of the sequence of audio responses. Specifically, we constructed a subsample that included 2 types of visits, both of which included only one audio response. The first started with the audio response (Group C1), whereas the second ended with the audio response (Group C2). We constructed a dummy variable to indicate the second type of visit. By incorporating the dummy variable, as well as controlling for *text_num*, *audio_num*, total Chinese characters, total audio duration, the physician title, and 3 sets of fixed effects, the coefficient of the dummy represented the difference in patient loyalty and satisfaction associated with moving the audio response from the opening to the ending.

The 3 sets of examinations corresponded to our prespecified hypotheses regarding response modalities, response length, and response sequence, each comprising 5 tests. To control the family-wise Type I error rate arising from multiple comparisons, Bonferroni correction was applied. The 3 outcomes related to patient loyalty (*fv1*, *fv2*, and *fv3*) and the 2 outcomes related to patient satisfaction (whether a review was provided and the review score) were each treated as a separate family. Accordingly, P values smaller than α=.050/3=.017 for loyalty outcomes and α=.050/2=.025 for satisfaction outcomes were considered statistically significant under the Bonferroni correction.

Then, as an exploratory analysis, the length and position of audio responses were jointly considered using additionally constructed subsamples that also contained 2 types of visits. The first type required the visit to start with the audio response, while the length was required to be shorter than the threshold t, ranging from 2 to 60 seconds. The second type included all text-only visits (Group A). After including the dummy variable

XSL•FO
RenderX

for Group A as an independent variable, we repeated the above regression analysis using different subsamples with different thresholds *t* combined with Group A, thus obtaining 59 regression results.

Beyond the dimension of audio length, the variation in position was also considered. We constructed a subsample containing visits with one audio response shorter than 5 seconds while not in the first position, as well as text-only visits (Group A). All visits were required to have at least 2 physician responses to facilitate the comparison. Using the subsample and constructing a dummy for the visits with audio responses, we could examine the difference between the above 2 types of visits regarding outcomes.

Ethical Considerations

This study protocol was approved by the Institutional Review Board (IRB) of Peking University Third Hospital (approval number IRB00006761-M2024701). The requirement for informed consent was waived by the IRB due to the retrospective nature of the study and the use of deidentified data. All data were strictly anonymized and deidentified prior to analysis to protect participant privacy and confidentiality. No compensation was provided to participants for this study. No identifiable images or personal details of participants were included in this manuscript.

Results

Response Modalities

As for the follow-up visit rates, the AME for pure audio response visits was −0.030 (95% CI −0.043 to −0.016, *P*<.001), −0.030 (95% CI −0.045 to −0.016, *P*<.001), and −0.037 (95% CI −0.054 to −0.019, *P*<.001) for *fv1*, *fv2*, and *fv3*, respectively, indicating that the follow-up rates for pure audio visits were 3 - 4 percentage points lower than those for pure text visits (Figure 3). While the combination of text and audio surpassed pure audio in eliciting follow-up visits, its coefficients remained negative in comparison to pure text. In summary, text-only visits (Group A) were associated with the highest patient loyalty, followed by combinations of text and audio (Group C), while audio-only visits (Group B) performed the worst, obtaining 16.6% - 30.9% lower follow-up visit rates than those achieved by text-only visits.

Figure 3. Association between response modality and outcomes (N=304,337 visits) at Peking University Third Hospital, Beijing, China (2021 - 2023). Each subfigure reports the results from one regression. The blue dots represent average marginal effects on *fv1*, *fv2*, *fv3*, and *review_dum* in Probit regressions (A-D) and coefficients for *review* in the ordinary least squares regression (E). Light blue vertical lines with caps represent 95% CIs. Wald tests were used to examine whether the coefficients of the 2 dummies were equal.

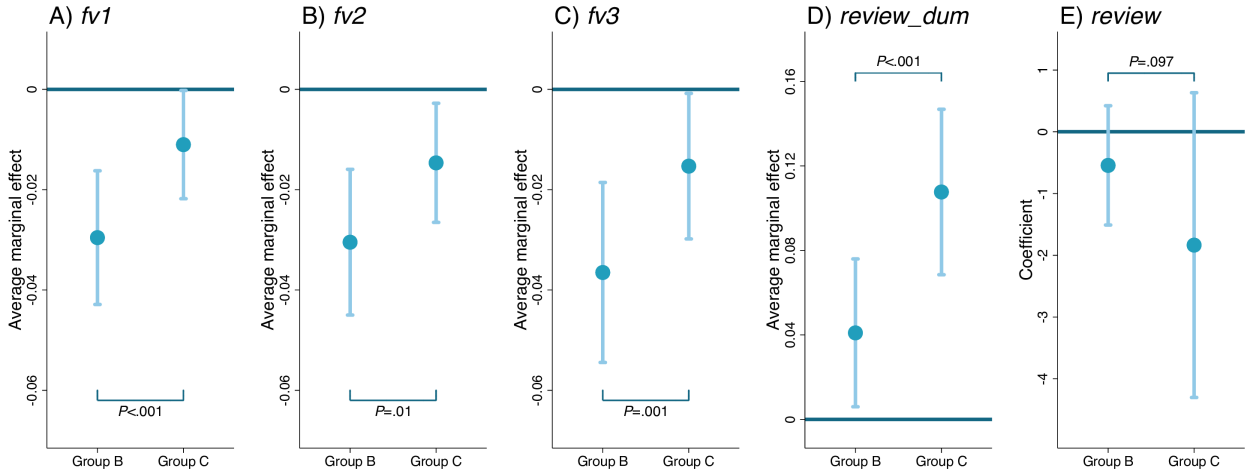

As for the review, whether the use of audio would promote the likelihood of receiving reviews was examined first. The AME of pure audio visits was 0.041 (95% CI 0.006 - 0.076, *P*=.02), indicating that physicians were more likely to receive patient reviews following the provision of medical services in the pure audio modality. The AME for the combination of text and audio was 0.108 (95% CI 0.068 - 0.147, *P*<.001), suggesting that visits in mixed information modalities were most evaluated by patients, followed by those in pure audio, while those in pure text were the least. Subsequently, the linkage between various response modalities and review scores was investigated. Upon adjusting for the inverse Mills ratio, it was determined that there were no statistically significant differences in the review scores among Groups A, B, and C. Note that Group B could achieve a higher level of satisfaction than Group A when the inverse Mills ratio was not controlled (Multimedia Appendix 3).

Response Length

Alongside the modality of response, length is a critical measure in telemedicine. Therefore, the association between response length and outcomes was further investigated. The number of responses was used as a metric to measure visit length when the total information load was fixed.

As for patient loyalty, results demonstrated that the AMEs of *text_num* were 0.009 (95% CI 0.006 - 0.011, *P*<.001), 0.008 (95% CI 0.005 - 0.011, *P*<.001), and 0.008 (95% CI 0.004−0.012 *P*<.001) for *fv1*, *fv2*, and *fv3*, respectively. The AMEs of *audio_num* were 0.007 (95% CI −0.000 to 0.016, *P*=.06), 0.011 (95% CI 0.005−0.018, *P*<.001), and 0.013 (95% CI 0.004 - 0.022, *P*=.006), respectively (whole sample, Figure 4). The findings indicated that irrespective of the modality—text or audio—used for telemedicine services, when the information

XSL•FO
RenderX

load was fixed, an increased number of physician responses correlated with enhanced patient loyalty.

Figure 4. Association between response length and outcomes at Peking University Third Hospital, Beijing, China (2021 - 2023). Each subfigure reports the results of 2 regressions, using the whole sample (N=304,337 visits) and Group B1 (n=3674 visits), respectively. Blue dots represent average marginal effects on *fv1*, *fv2*, *fv3*, and *review_dum* in probit regressions (A-D) and coefficients for *review* in ordinary least squares regressions (E). Light blue vertical lines with caps represent 95% CIs.

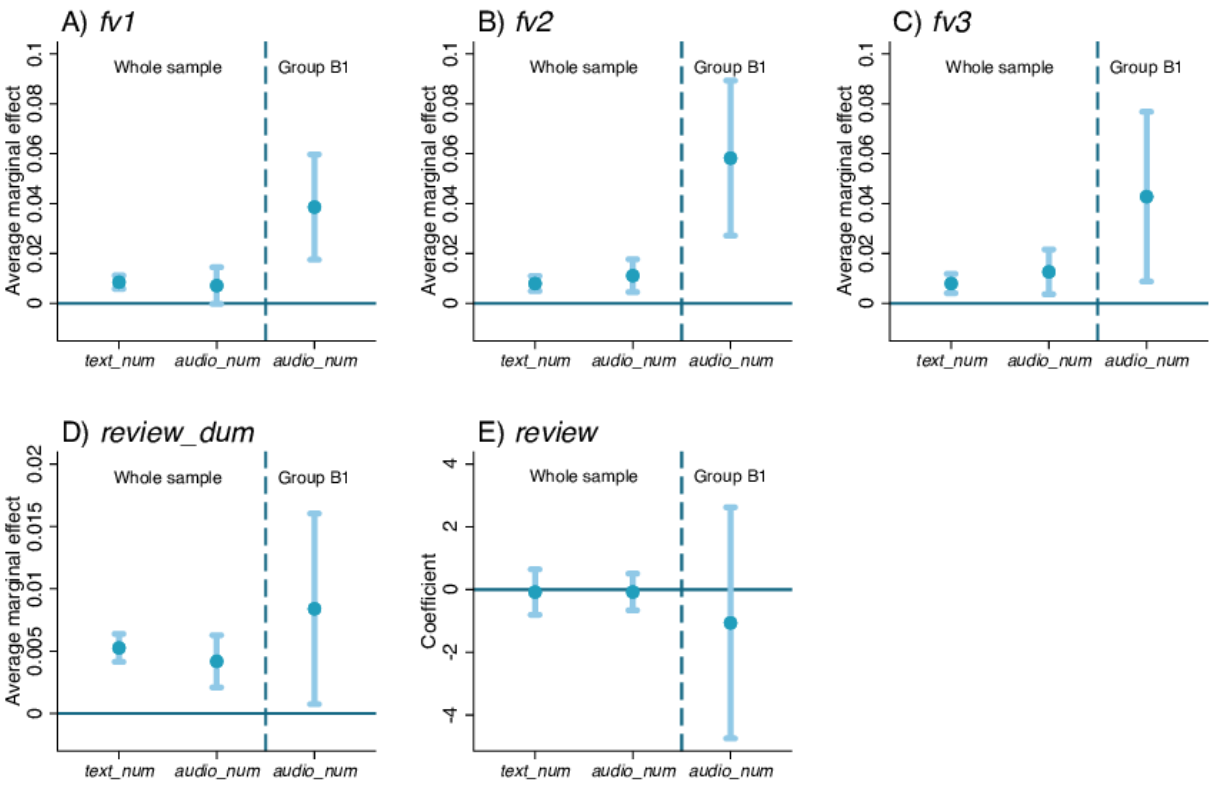

To further validate the above conclusions, we leveraged the design of the telemedicine platform, where a single audio response was limited to a maximum of 59 seconds, to create a more extreme scenario. Specifically, cases where the total audio length of physician responses in a visit fell between 55 and 59 seconds were selected, and we investigated whether there would be differences in the outcome measures as the number of audio responses increased (Group B1, Figure 4).

The results showed that the AMEs of *audio_num* were 0.039 (95% CI 0.017 - 0.060, *P*<.001), 0.058 (95% CI 0.027 - 0.089, *P*<.001), and 0.043 (95% CI 0.009 - 0.077, *P*=.014) for *fv1*, *fv2*, and *fv3*, respectively (Figure 4). This positive association further confirmed that patient loyalty improved when lengthy audio information was divided into several shorter segments.

As for the review, each additional text and audio response was linked to a 0.5% (95% CI 0.4% - 0.6%, *P*<.001) and 0.4% (95% CI 0.2% - 0.6%, *P*<.001) increase in the probabilities of patients providing reviews, respectively. Yet, both additional text and audio responses did not significantly alter the actual review score, indicating that segmenting text or audio into shorter responses was not associated with improved patient satisfaction (Figure 4).

Sequence

In addition to response length, the sequence in which text and audio messages were presented may also be connected with outcomes in visits containing both modalities. We restricted the sample to Group C to examine the association of audio responses positioned at various rounds and the outcomes. Specifically, 2 groups were constructed from the sample. The first group comprised visits in which the first response was in the audio modality and the last response was not (audio-opening visits, Group C1). The second group encompassed visits that ended with an audio response and did not begin with an audio response (audio-ending visits, Group C2).

Firstly, the follow-up visit rates associated with these 2 groups of visits were evaluated. The findings indicated that the AMEs for being in Group C2 were −0.019 (95% CI −0.033 to −0.005, *P*=.007), −0.022 (95% CI −0.042 to −0.001, *P*=.04), and −0.023 (95% CI −0.045 to −0.000, *P*=.046) on *fv1*, *fv2*, and *fv3*, respectively, compared to Group C1 (Table 2). This demonstrated that patients exhibited increased loyalty to the same physician when physicians used audio responses at the beginning of a telemedicine session instead of at the end. Note that the associations for *fv2* and *fv3* were no longer significant after Bonferroni correction.

XSL•FO
RenderX

Table . Association between the position and sequence of the audio responses and outcomes at Peking University Third Hospital, Beijing, China (2021 - 2023).[a]

	fv1	*fv2*	*fv3*	*review_dum*	*review*
Comparing the audio-opening visits (Group C1) and audio-ending visits (Group C2)					
Group C2 (95% CI, *P* value)	−0.019 (−0.033 to −0.005, *P*=.007)	−0.022 (−0.042 to −0.001, *P*=.04)	−0.023 (−0.045 to −0.000, *P*=.046)	0.005 (−0.001 to 0.011, *P*=.098)	0.170 (−1.017 to 1.358, *P*=.78)
Comparing the visits with nonbeginning short audio group and text-only visits (Group A)					
Visits with nonbeginning short audio (95% CI, *P* value)	0.152 (−0.032 to 0.337, *P*=.11)	0.189 (−0.045 to 0.422, *P*=.11)	0.212 (−0.150 to 0.575, *P*=.25)	−0.055 (−0.188 to 0.077, *P*=.41)	−1.960 (−15.598 to 11.677, *P*=.78)
Controls	Y	Y	Y	Y	Y
Physician FE[b]	Y	Y	Y	Y	Y
Day-of-week FE	Y	Y	Y	Y	Y
Year-month FE	Y	Y	Y	Y	Y

[a]The coefficients in the first row were estimated using the subsample Group C1 and Group C2 (n=33,758 visits), and those in the second row were estimated using the subsample Group A and visits with only one nonbeginning short audio response of less than 5 seconds in Group C, both restricted to visits with at least 2 physician responses (n=134,698 visits). Average marginal effects for *fv1*, *fv2*, *fv3*, and *review_dum* from probit regressions are reported in the first 4 columns, and coefficients for *review* from ordinary least squares regressions are reported in the last column.

[b]FE: fixed effect ____.

Secondly, the association between audio position and reviews was examined. The findings indicated no significant association between audio position and the probability of patients submitting reviews or the actual review scores.

Role of Audio

The previous results suggested that, while the audio modality did not facilitate the follow-up visit rate, as reflected by a diminished patient loyalty for Group C compared to Group A, brevity and being in the opening position of audio responses were individually associated with elevated follow-up visit rates. Therefore, the follow-up visit rates of Group C1 and Group A were compared to examine if the decrease in follow-up visit rates due to audio use could be mitigated when the audio was short and positioned at the beginning.

Results showed that when the length of the introductory audio was less than 5 seconds, the follow-up visit rates in this group were significantly higher than those of visits with only text, with the AMEs of 0.069 (95% CI 0.018 - 0.120, *P*=.008), 0.077 (95% CI 0.018 - 0.136, *P*=.011), and 0.099 (95% CI 0.001 - 0.197, *P*=.049) on *fv1*, *fv2*, and *fv3*, respectively (Figure 5). As for reviews, no matter how the length of the introductory audio response changed, there were no significant differences between Group C1 and the text-only visits in terms of the probability of getting reviews or the actual review scores.

XSL•FO
RenderX

Figure 5. Association between the length of the first audio response and outcomes at Peking University Third Hospital, Beijing, China (2021 - 2023). Each subfigure contains results from 59 regressions, each comparing visits with only 1 audio response placed at the first position and shorter than *t* seconds (2≤t≤60) with visits consisting solely of text. The blue curves represent average marginal effect (AMEs) of being the visit with an introductory audio response on *fv1*, *fv2*, *fv3*, and *review_dum* in probit regressions in (A-D), and its coefficients on *review* in ordinary least squares regressions (E). The light blue shades represent 95% CIs. The blue dots and vertical lines with caps represent the AMEs or coefficient and 95% CIs, respectively, when *t*=5 (meaning subsample with pure-text visits and visits with an introductory audio response that are shorter than 5 s).

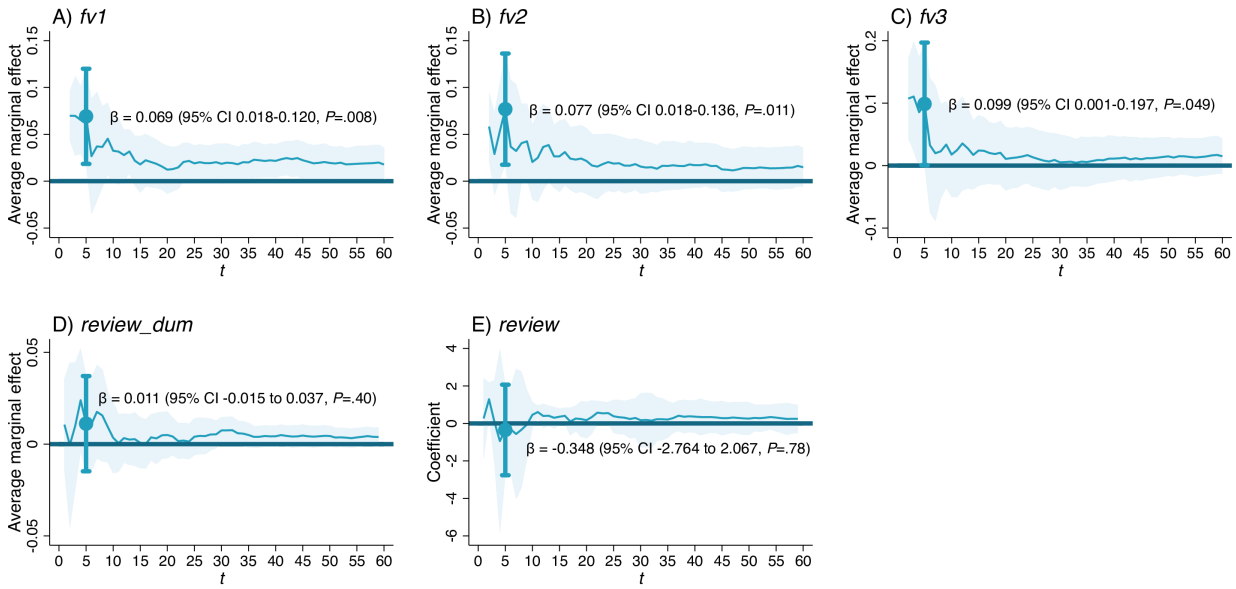

Therefore, visits featuring solely brief introductory audio responses yielded a higher follow-up visit rate than those with only text responses. In addition, this effect diminished when the audio length reached 5 seconds.

Subsequently, short duration (less than 5 s) was maintained as a favorable condition, and whether changing the position of short audio responses would be associated with elevated or diminished follow-up visit rates compared to Group A was assessed. We identified visits in Group C that contained only one short audio response, which was not positioned in the first place, as the nonbeginning audio response group, then compared this group to text-only visits with at least 2 physician responses. No significant effects were observed between the 2 groups of visits (Table 2).

Discussion

Principal Results

This study is the first to systematically examine the effects of physicians' communication modality, length, and sequence on patient behavioral outcomes in asynchronous telemedicine. Our primary findings indicated that the repurchase rate was highest in text-only visits, followed by text-audio hybrid visits, and lowest in audio-only visits. The reduced average length of both text and audio responses was correlated with higher patient loyalty. We also identified a response pattern in which virtual visits begin with a brief introductory audio message followed by text, which was associated with the highest patient loyalty. Additionally, audio responses were connected to a higher likelihood of patients providing reviews without changing the review scores. Our research illustrates specific behavioral patterns of physicians in telemedicine, elucidates the association between various communication features and outcomes, and serves as a reference for delivering high-quality and efficient telemedicine services.

This study improves our current comprehension of physician behaviors in telemedicine by leveraging a large-scale dataset of real-world interactions. A key methodological strength lies in our dual-outcome measurement, which integrates objective longitudinal repurchase behavior with subjective satisfaction scores. While satisfaction may accurately represent the immediate consumption experience, it remains a subjective assessment prone to selection bias, often suffering from a low response rate, which is estimated to be only 5% - 10% in natural settings [51]. In contrast, repurchase behavior serves as an objective, long-term metric free from selection bias [52,53]. By combining these 2 metrics, our study overcomes the limitations of relying solely on self-reported data [34,54,55], offering a more comprehensive evaluation of patient outcomes. In addition, our study thoroughly characterized various dimensions of physician responses, including the modality, length, and sequence. This multidimensional analysis approach establishes a robust empirical foundation for understanding physician-patient communication behaviors.

Central to the validity of this multidimensional framework is the rigorous construction process of the patient loyalty proxy, which was carefully designed to distinguish behavioral preference from clinical necessity or dependency. Unlike in general consumer services, a repeat visit in health care does not inherently signal satisfaction; it may instead reflect unresolved medical needs [56,57]. To address this problem, we imposed specific temporal and scope constraints to isolate true loyalty.

XSL•FO
RenderX

Regarding the temporal dimension, we established a 1‑ to 6-month window to balance the exclusion of clinical necessity against the attribution of care quality. On the lower bound, we excluded revisits occurring within 30 days [38]. In clinical contexts, immediate returns often serve as proxies for unplanned revisits, signaling ineffective initial communication or unresolved symptoms rather than satisfaction [35-41]. On the upper bound, we restricted the observation to 6 months. This cutoff is empirically grounded in the temporal decay of visit influence; for instance, prior research indicates that while initial patient experience significantly predicts retention at 6 months, this association fades and becomes nonsignificant by 12 months [43,44]. Thus, limiting the window ensures that the measured follow-up behavior remains attributable to the impact of the index visit rather than long-term habituation.

Regarding the scope of loyalty, we stratified follow-up visits into 3 levels—the same physician, the same department, and all departments—to capture the multilayered nature of patient trust [58]. While returning to the same physician (*fv1*) serves as the most direct measure of interpersonal loyalty, patients in high-demand academic centers often manifest institutional trust by transferring their confidence to a physician's colleagues [59,60]. By broadening the scope to the same department (*fv2*) and all departments (*fv3*), we capture the spillover effect of effective communication, where a positive interaction with an individual provider enhances the perceived reliability of the entire medical team [59,60].

Crucially, our exclusion criteria incorporated a nuanced interaction between time and scope to account for disease relatedness. We applied the 30-day exclusion period strictly to revisits involving the same physician or department, as these were highly likely to address the same unresolved medical needs. In contrast, visits to different departments within a month were not subject to this exclusion constraint. This distinction is pivotal, as an immediate return to the same specialty suggests a failure to resolve the initial problem, whereas a visit to a different specialty typically indicates a new, unrelated health need [61]. Thus, retaining these cross-specialty visits allows us to capture a broader form of hospital-wide loyalty without confounding it with the negative quality signal of clinical redundancy. Yet, given that returning to the same individual physician offers the most direct and unmediated reflection of the specific interaction's quality, we designated *fv1* as our primary outcome measure to maximize internal validity.

Our focus on repurchase behavior complements the broader literature on patient nonattendance [16]. While nonattendance metrics capture passive adherence to scheduled appointments, the asynchronous workflow in our study operates on a patient-initiated, on-demand basis where prebooking is absent [62]. In this context, engagement is not measured by the absence of a no-show but by the active decision to repurchase. Thus, our loyalty metric represents a deliberate reselection of the provider, serving as a distinct, proactive indicator of the patient-provider connection.

The specific aspects that these 2 outcome indicators highlight align with the different factors that patients encounter during a telemedicine visit. In a virtual visit, a patient's holistic experience is derived from 2 distinct aspects: the comprehension of medical information, which constitutes an evaluation of factual matters, and the perception of physicians' attitudes, which represents an evaluation of interpersonal interactions. We shall examine the importance of the findings from these 2 aspects [63].

Research indicated that comprehension of information was a significant determinant in fostering repurchase behavior [64]. Textual information surpasses auditory in improving understanding of lengthy and complicated material [65]. Meanwhile, medicine is considered a complex field that most individuals have limited knowledge of, resulting in significant information asymmetry [66]. These could explain our findings that text-only visits resulted in more repurchasing behaviors. Because text responses may improve patients' comprehension of physicians' information and provide convenient access for future reference, thereby fostering recognition and promoting repurchases, the outcome metric from an objective and long-term perspective.

Another crucial dimension of outcomes is patient satisfaction, which primarily reflects patients' feelings and attitudes toward physicians. Prior research has indicated that audio communication is more effective than text in bridging gaps among those unable to meet in person [42], and it correlated with increased satisfaction [24]. However, in this study, the use of audio responses resulted in an enhancement of the probability of providing reviews, rather than review scores after adjusting for the inverse Mills ratio. Yet, in the absence of controlling for selection bias, the review scores were higher in the audio-only visits (P=.04), leading us to hypothesize that the increase in satisfaction levels attributed to audio might be a result of selection bias.

Beyond the response modality, a notable discovery is that shorter responses had better outcomes not only for audio responses but also for text responses. When the load of information from the physician was consistent, dividing the content into shorter responses correlated with a higher probability of follow-up visits and an enhanced inclination to provide feedback. This finding aligns with previous empirical research conducted in online health communities, which demonstrated that although physicians' knowledge-sharing behaviors could effectively enhance patient engagement, overly lengthy medical information might inadvertently induce patient anxiety [67], suggesting the importance of response segmentation in physician-patient communication.

Our findings indicated that audio responses were adversely associated with patients' repurchase behavior. The repurchase rate for audio-only or hybrid visits was inferior to that of text alone. Does this imply that audio should not be used at all? Not particularly. We identified that visits featuring a brief introductory audio message (<5 s) followed by text yielded the highest patient loyalty, highlighting it as a particularly effective communication pattern within our sample. This served as evidence that audio itself was not inherently flawed. Given the inability to access substantive content based on communication norms, we hypothesize that these initial, sub–5-second audio clips likely function as personalized greetings or empathetic

XSL•FO
RenderX

openings rather than complex medical explanations. This initial human touch may help establish an emotional connection, while the subsequent text ensures the clarity and reviewability of the medical advice.

Limitations

This study used physician-patient interaction records in the real world and thoroughly examined the roles of text and audio usage, offering a reference for physicians in telemedicine practice; nevertheless, it also had specific limitations.

First, this study focused on observable communication features and did not examine the substantive content of asynchronous care. As a result, the informational richness and clinical appropriateness of medical advice, as well as their roles in shaping communication practices, could not be directly assessed. Future research may extend this work by incorporating content-based or semantic analysis with data on medical specialization and case characteristics to provide a more comprehensive evaluation of how communication quality and clinical complexity jointly affect telemedicine outcomes.

Second, while our study documents associations between physicians' response features and patient behavioral outcomes, it does not allow us to infer causal relationships. Retrospective investigations do indeed possess such limitations. To alleviate the concern, we prioritized controlling for a series of covariates and added abundant fixed effects at the physician level, day-of-week level, and year-month level. Therefore, our results are unlikely to be driven by differences across physicians or time periods. While we cannot fully eliminate endogeneity in a retrospective study, we have taken extensive steps to mitigate such concerns in our analysis.

Third, the external validity of our findings may be influenced by institutional, platform, and cultural factors. The data were drawn from a single large AMC in China, where high patient volume and strong professional norms may shape communication patterns differently from those in primary or community care settings. In addition, platform-specific design features, such as the 59-second audio length limit, may act as technological nudges encouraging concise responses, which is a common constraint across major Chinese platforms. These factors could limit generalizability to other health care systems or international contexts with different technical parameters and cultural expectations around physician-patient interactions. Future cross-institutional, cross-platform, and cross-cultural studies are needed to examine how these elements interact to shape effective communication practices.

Conclusions

This study is innovative in using a large-scale dataset to thoroughly examine the distinct links between specific communication features and patient loyalty outcomes in telemedicine. It differs from existing studies by leveraging a large-scale real-world dataset and objective indicators, diverging from prior research that relied on small-scale simulations or subjective surveys. It provides critical empirical evidence to the field, establishing that while text-only visits facilitate content comprehension and audio encourages feedback, a hybrid pattern with a brief audio introduction followed by text is associated with the highest patient loyalty. Its implication in the real world is offering critical knowledge on how physicians can strategically select engagement tactics to improve service quality and patient retention.

Acknowledgments

Generative artificial intelligence played no role in the writing of this manuscript.

Funding

This work was supported by the Capital Foundation of Medical Development to KZ (2024-2-40915) and the National Natural Science Foundation of China to YW (72204005) and TXL (72342032). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets used in this study are not publicly available due to their proprietary nature but may be made accessible upon reasonable request and with the permission of the corresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Flowchart of virtual visits after applying exclusion and inclusion criteria.
[DOCX File, 188 KB - jmir_v28i1e86977_app1.docx]

Multimedia Appendix 2
The number of visits (N=304,337 visits) and physicians (N=781) across years and departments at Peking University Third Hospital, Beijing, China (2021-2023). Departments with visit volume of less than 1% each year are consolidated into "Others."
[DOCX File, 21 KB - jmir_v28i1e86977_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
The impact of response modality on satisfaction without controlling inverse Mills ratio (N=304,337 visits) at Peking University Third Hospital, Beijing, China (2021-2023).
[DOCX File, 17 KB - jmir_v28i1e86977_app3.docx]

References

1. Chang E, Penfold RB, Berkman ND. Patient characteristics and telemedicine use in the US, 2022. JAMA Netw Open 2024 Mar 4;7(3):e243354. [doi: 10.1001/jamanetworkopen.2024.3354] [Medline: 38517438]
2. Dorn SD. Backslide or forward progress? Virtual care at U.S. healthcare systems beyond the COVID-19 pandemic. NPJ Digit Med 2021 Jan 8;4(1):6. [doi: 10.1038/s41746-020-00379-z] [Medline: 33420420]
3. Dorsey ER. The new platforms of health care. NPJ Digit Med 2021 Jul 15;4(1):112. [doi: 10.1038/s41746-021-00478-5] [Medline: 34267312]
4. Shaver J. The state of telehealth before and after the COVID-19 pandemic. Prim Care 2022 Dec;49(4):517-530. [doi: 10.1016/j.pop.2022.04.002] [Medline: 36357058]
5. Appointments in general practice. : NHS England; 2024 URL: https://digital.nhs.uk/data-and-information/publications/statistical/appointments-in-general-practice/june-2024 [accessed 2026-02-28]
6. Amagai S, Vonesh E, Adams J, Luo Y. Closing the gap: addressing telehealth disparities across specialties in the sustained pandemic era. NPJ Digit Med 2024 Aug 21;7(1):217. [doi: 10.1038/s41746-024-01201-w] [Medline: 39164391]
7. Roth M. Telehealth: how asynchronous communication creates provider efficiencies. Accreditation & Quality Compliance Center. URL: https://www.accreditationqualitycenter.com/articles/telehealth-how-asynchronous-communication-creates-provider-efficiencies [accessed 2026-02-28]
8. Rzewuska Díaz M, Locock L, Keen A, Melvin M, Myhill A, Ramsay C. Implementation of a web-based outpatient asynchronous consultation service: mixed methods study. J Med Internet Res 2024 Jun 4;26:e48092. [doi: 10.2196/48092] [Medline: 38833695]
9. Wang Y, Zhang P, Xing Y, et al. Telemedicine integrated care versus in-person care mode for patients with short stature: comprehensive comparison of a retrospective cohort study. J Med Internet Res 2024 Nov 19;26:e57814. [doi: 10.2196/57814] [Medline: 39560983]
10. Barbosa W, Zhou K, Waddell E, Myers T, Dorsey ER. Improving access to care: telemedicine across medical domains. Annu Rev Public Health 2021 Apr 1;42:463-481. [doi: 10.1146/annurev-publhealth-090519-093711] [Medline: 33798406]
11. Nguyen OT, Alishahi Tabriz A, Huo J, Hanna K, Shea CM, Turner K. Impact of asynchronous electronic communication-based visits on clinical outcomes and health care delivery: systematic review. J Med Internet Res 2021 May 5;23(5):e27531. [doi: 10.2196/27531] [Medline: 33843592]
12. Leighton C, Cooper A, Porter A, Edwards A, Joseph-Williams N. Effectiveness and safety of asynchronous telemedicine consultations in general practice: a systematic review. BJGP Open 2024 Apr;8(1). [doi: 10.3399/BJGPO.2023.0177] [Medline: 37783479]
13. Levinson W, Lesser CS, Epstein RM. Developing physician communication skills for patient-centered care. Health Aff (Millwood) 2010 Jul;29(7):1310-1318. [doi: 10.1377/hlthaff.2009.0450] [Medline: 20606179]
14. Drossman DA, Chang L, Deutsch JK, et al. A review of the evidence and recommendations on communication skills and the patient-provider relationship: a Rome foundation working team report. Gastroenterology 2021 Nov;161(5):1670-1688. [doi: 10.1053/j.gastro.2021.07.037] [Medline: 34331912]
15. Mason AN. The most important telemedicine patient satisfaction dimension: patient-centered care. Telemed J E Health 2022 Aug;28(8):1206-1214. [doi: 10.1089/tmj.2021.0322] [Medline: 34882032]
16. Greenup EP, Best D. Systematic review and meta-analysis of no show or non-attendance rates among telehealth and in-person models of care. BMC Health Serv Res 2025 May 9;25(1):663. [doi: 10.1186/s12913-025-12826-2] [Medline: 40340966]
17. He X, Wu D. Does additional audio really work? A study on users' cognitive behavior with audio-visual dual-channel in panoramic digital museum. Inf Manag 2023 Jun;60(4):103791. [doi: 10.1016/j.im.2023.103791]
18. Xia A, Perera SC, Ahmed MU, Tang J, Wang JJ. Voice or text? The role of physician media choice on patient experience in online medical communities. Decis Sci 2024 Dec;55(6):620-638. [doi: 10.1111/deci.12654]
19. Valenzuela-Gálvez ES, González-Benito Ó, Garrido-Morgado Á. Sound in mode on: how sound stimuli influence customer engagement? Electron Commer Res 2026 Feb;26(1):783-804. [doi: 10.1007/s10660-024-09902-w]
20. Culmer N, Smith TB, Stager C, et al. Asynchronous telemedicine: a systematic literature review. Telemed Rep 2023;4(1):366-386. [doi: 10.1089/tmr.2023.0052] [Medline: 38143795]
21. Yan M, Tan H, Jia L, Akram U. The antecedents of poor doctor-patient relationship in mobile consultation: a perspective from computer-mediated communication. Int J Environ Res Public Health 2020 Apr 9;17(7):2579. [doi: 10.3390/ijerph17072579] [Medline: 32283741]
22. Street RL, Gordon H, Haidet P. Physicians' communication and perceptions of patients: is it how they look, how they talk, or is it just the doctor? Soc Sci Med 2007 Aug;65(3):586-598. [doi: 10.1016/j.socscimed.2007.03.036] [Medline: 17462801]

XSL•FO
RenderX

23. Alkureishi MA, Lee WW, Lyons M, et al. Impact of electronic medical record use on the patient-doctor relationship and communication: a systematic review. J Gen Intern Med 2016 May;31(5):548-560. [doi: 10.1007/s11606-015-3582-1] [Medline: 26786877]
24. Liu Y, Zhang X, Liu L, Lai KH. Does voice matter? Investigating patient satisfaction on mobile health consultation. Inf Process Manag 2023 Jul;60(4):103362. [doi: 10.1016/j.ipm.2023.103362]
25. Zeng Z, Xu DR, Cai Y, Gong W. Assessing quality of direct-to-consumer telemedicine in China: a cross-sectional study using unannounced standardised patients. BMJ Qual Saf 2024 Aug 16;33(9):609-612. [doi: 10.1136/bmjqs-2024-017072] [Medline: 38688711]
26. Gong X, Hou M, Guo R, Feng XL. Investigating the relationship between consultation length and quality of tele-dermatology E-consults in China: a cross-sectional standardized patient study. BMC Health Serv Res 2022 Sep 22;22(1):1187. [doi: 10.1186/s12913-022-08566-2] [Medline: 36138410]
27. Peeters KMM, Reichel LAM, Muris DMJ, Cals JWL. Family physician-to-hospital specialist electronic consultation and access to hospital care: a systematic review. JAMA Netw Open 2024 Jan 2;7(1):e2351623. [doi: 10.1001/jamanetworkopen.2023.51623] [Medline: 38214930]
28. Cheng TC, Yip W. Policies, progress, and prospects for internet telemedicine in China. Health Syst Reform 2024 Dec 17;10(2):2389570. [doi: 10.1080/23288604.2024.2389570] [Medline: 39437237]
29. Geng J, Li R, Wang X, et al. Exploring the lack of continuity of care in older cancer patients under China's "integrated health system" reform. Age Ageing 2024 Oct 1;53(10):afae213. [doi: 10.1093/ageing/afae213] [Medline: 39373574]
30. Yip W, Fu H, Jian W, et al. Universal health coverage in China part 1: progress and gaps. Lancet Public Health 2023 Dec;8(12):e1025-e1034. [doi: 10.1016/S2468-2667(23)00254-2] [Medline: 38000882]
31. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
32. Yan X, Wang J, Chau M. Customer revisit intention to restaurants: evidence from online reviews. Inf Syst Front 2015 Jun;17(3):645-657. [doi: 10.1007/s10796-013-9446-5]
33. Alegre J, Cladera M. Analysing the effect of satisfaction and previous visits on tourist intentions to return. Eur J Mark 2009 May 29;43(5/6):670-685. [doi: 10.1108/03090560910946990]
34. Woo S, Choi M. Medical service quality, patient satisfaction and intent to revisit: case study of public hub hospitals in the Republic of Korea. PLoS ONE 2021;16(6):e0252241. [doi: 10.1371/journal.pone.0252241] [Medline: 34181662]
35. Glock H, Jakobsson U, Borgström Bolmsjö B, Milos Nymberg V, Wolff M, Calling S. eVisits to primary care and subsequent health care contacts: a register-based study. BMC Prim Care 2024 Aug 12;25(1):297. [doi: 10.1186/s12875-024-02541-y] [Medline: 39135150]
36. Zeltzer D, Einav L, Rashba J, Balicer RD. The impact of increased access to telemedicine. J Eur Econ Assoc 2024 Apr 3;22(2):712-750. [doi: 10.1093/jeea/jvad035]
37. Uscher-Pines L, Mehrotra A. Analysis of Teladoc use seems to indicate expanded access to care for patients without prior connection to a provider. Health Aff (Millwood) 2014 Feb;33(2):258-264. [doi: 10.1377/hlthaff.2013.0989] [Medline: 24493769]
38. Duseja R, Bardach NS, Lin GA, et al. Revisit rates and associated costs after an emergency department encounter: a multistate analysis. Ann Intern Med 2015 Jun 2;162(11):750-756. [doi: 10.7326/M14-1616] [Medline: 26030633]
39. Ye H, Yi J. Patient-physician race concordance, physician decisions, and patient outcomes. Rev Econ Stat 2023 Jul 11;105(4):766-779. [doi: 10.1162/rest_a_01236]
40. Chan DC. Teamwork and moral hazard: evidence from the emergency department. J Pol Econ 2016 Jun;124(3):734-770. [doi: 10.1086/685910]
41. Chan DC. The efficiency of slacking off: evidence from the emergency department. Econometrica 2018;86(3):997-1030. [doi: 10.3982/ECTA13565]
42. Sprecher S. Initial interactions online-text, online-audio, online-video, or face-to-face: effects of modality on liking, closeness, and other interpersonal outcomes. Comput Human Behav 2014 Feb;31:190-197. [doi: 10.1016/j.chb.2013.10.029]
43. Examples of physician quality measures for consumers. Agency for Healthcare Research and Quality. URL: https://www.ahrq.gov/talkingquality/measures/setting/physician/examples.html [accessed 2026-02-28]
44. Quigley DD, Reynolds K, Dellva S, Price RA. Examining the business case for patient experience: a systematic review. J Healthc Manag 2021;66(3):200-224. [doi: 10.1097/JHM-D-20-00207] [Medline: 33960966]
45. Saini GK, Matta JK. Service quality and re-visit intention: a comparative study of charitable and corporate hospitals. J Health Manag 2014 Dec;16(4):595-612. [doi: 10.1177/0972063414548562]
46. Juwaheer TD, Kassean H. Exploring quality perceptions of health care operations: a study of public hospitals of Mauritius. J Hosp Mark Public Relations 2006;16(1-2):89-111. [doi: 10.1300/J375v16n01_07] [Medline: 17194688]
47. Nundy S, Cooper LA, Mate KS. The quintuple aim for health care improvement: a new imperative to advance health equity. JAMA 2022 Feb 8;327(6):521-522. [doi: 10.1001/jama.2021.25181] [Medline: 35061006]
48. Kruse CS, Krowski N, Rodriguez B, Tran L, Vela J, Brooks M. Telehealth and patient satisfaction: a systematic review and narrative analysis. BMJ Open 2017 Aug 3;7(8):e016242. [doi: 10.1136/bmjopen-2017-016242] [Medline: 28775188]

XSL•FO
RenderX

49. Olsen SO. Comparative evaluation and the relationship between quality, satisfaction, and repurchase loyalty. J Acad Mark Sci 2002 Jan 1;30(3):240-249. [doi: 10.1177/0092070302303005]
50. Heckman JJ. Sample selection bias as a specification error. Econometrica 1979 Jan;47(1):153-161. [doi: 10.2307/1912352]
51. Online review statistics: the definitive list. Luisa Zhou. URL: https://luisazhou.com/blog/online-review-statistics [accessed 2024-12-30]
52. Mittal V, Kamakura WA. Satisfaction, repurchase intent, and repurchase behavior: Investigating the moderating effect of customer characteristics. J Mark Res 2001 Feb;38(1):131-142. [doi: 10.1509/jmkr.38.1.131.18832]
53. Curtis T, Abratt R, Rhoades D, Dion P. Customer loyalty, repurchase and satisfaction: a meta-analytical review. J Consum Satisfaction, Dissatisfaction Compl Behav 2011;24:1-26 [FREE Full text]
54. Petrie S, Laur C, Rios P, et al. Quality measures of virtual care in ambulatory healthcare environments: a scoping review. BMJ Open 2024 Apr 2;14(4):e078214. [doi: 10.1136/bmjopen-2023-078214] [Medline: 38569716]
55. Wu DC, Zhao X, Wu J. Online physician-patient interaction and patient satisfaction: empirical study of the internet hospital service. J Med Internet Res 2023 Aug 24;25:e39089. [doi: 10.2196/39089] [Medline: 37616031]
56. Namgung M, Lee DH, Bae SJ, et al. A comparison of emergency department revisit rates of pediatric patients between pre-COVID-19 and COVID-19 periods. Children (Basel) 2022 Jul 4;9(7):1003. [doi: 10.3390/children9071003] [Medline: 35883987]
57. Juang WC, Cai ZX, Chen CM, You ZH. Assessing revisit risk in emergency department patients: machine learning approach. JMIR AI 2025 Aug 7;4:e74053. [doi: 10.2196/74053] [Medline: 40773678]
58. Meyer SB, Brown P, Calnan M, et al. Development and validation of the trust in multidimensional healthcare systems scale (TIMHSS). Int J Equity Health 2024 May 8;23(1):94. [doi: 10.1186/s12939-024-02162-y] [Medline: 38720303]
59. Greene J, Samuel-Jakubos H. Building patient trust in hospitals: a combination of hospital-related factors and health care clinician behaviors. Jt Comm J Qual Patient Saf 2021 Dec;47(12):768-774. [doi: 10.1016/j.jcjq.2021.09.003] [Medline: 34654668]
60. Taylor LA, Nong P, Platt J. Fifty years of trust research in health care: a synthetic review. Milbank Q 2023 Mar;101(1):126-178. [doi: 10.1111/1468-0009.12598] [Medline: 36689251]
61. Hutchinson CL, Curtis K, McCloughen A. Characteristics of patients who return unplanned to the ED, and factors that contribute to their decision to return: integrated results from an explanatory sequential mixed methods inquiry. Australas Emerg Care 2024 Mar;27(1):71-77. [doi: 10.1016/j.auec.2023.08.004] [Medline: 37741746]
62. AlOsaimi RB, Aljehani NM, Aljuhani T. Examining the reasons for patient no-show in an outpatient adult rehabilitation department in King Abdulaziz Medical City, Riyadh: a cross-sectional study. Health Sci Rep 2025 May;8(5):e70783. [doi: 10.1002/hsr2.70783] [Medline: 40309620]
63. Yang Z, Velasco Vizcaíno F, Jin M. Enhancing patient wellbeing through telemedicine services: the impact of cognitive and affective service quality ratings and physicians' long-term orientation. Int J Inf Manage 2025 Jun;82:102873. [doi: 10.1016/j.ijinfomgt.2025.102873]
64. Hollebeek LD, Macky K. Digital content marketing's role in fostering consumer engagement, trust, and value: framework, fundamental propositions, and implications. J Interact Mark 2019 Feb;45(1):27-41. [doi: 10.1016/j.intmar.2018.07.003]
65. Leahy W, Sweller J. Cognitive load theory and the effects of transient information on the modality effect. Instr Sci 2016 Feb;44(1):107-123. [doi: 10.1007/s11251-015-9362-9]
66. Arrow KJ. Uncertainty and the welfare economics of medical care. Am Econ Rev 1963;53(5):941-973 [FREE Full text]
67. Meng F, Liu Y, Zhang X, Liu L. General knowledge-sharing and patient engagement in online health communities: an inverted U-shaped relationship. J Knowl Manag 2024 Mar 18;28(3):763-788. [doi: 10.1108/JKM-12-2022-0986]

Abbreviations

AMC: academic medical center
AME: average marginal effect
HIS: hospital information system
IRB: Institutional Review Board
JARS: Journal Article Reporting Standards

XSL•FO
RenderX

Edited by S Brini; submitted 03.Nov.2025; peer-reviewed by EP Greenup, R Zhang; revised version received 03.Feb.2026; accepted 18.Feb.2026; published 26.Mar.2026.

Please cite as:
Wang Y, Wang S, Zhang K, Liu Z, Ma Q, Ji H, Hou Z, Liu TX, Xu X, Wu X, Jin C
Association Between Physician Communication Features and Patient Outcomes in Telemedicine: Retrospective Cross-Sectional Observational Study
J Med Internet Res 2026;28:e86977
URL: https://www.jmir.org/2026/1/e86977
doi:10.2196/86977

© Yipei Wang, Shu Wang, Ke Zhang, Zhijie Liu, Qingbian Ma, Hong Ji, Zheng Hou, Tracy Xiao Liu, Xuedong Xu, Xinxia Wu, Changxiao Jin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Primary Health in Rwanda: Qualitative Study of User Experiences and Implementation Lessons From Babyl's Telemedicine Platform

Sabine Musange Furere[1], MD; Gashaija Absolomon[1], MPH; Nathalie Umutoni[1], MD; Stella Matutina Umuhoza[2,3], MPH; Felix K Rubuga[1,2], PhD, MPH; Yvonne Delphine Nsaba-Uwera[4], MSN (MWH); Patrick Singa Muhoza[5], MPH; Regis Hitimana[2,6], PhD; Jeanine Condo[1,2,7], PhD, MD; James Humuza[2], PhD, MD

[1]Centre for Impact, Innovation and Capacity Building for Health Information Systems and Nutrition (CIIC-HIN), Kigali, Rwanda

[2]School of Public Health, University of Rwanda, KK 19 Ave, Kigali, Rwanda

[3]Partners In Health - Inshuti Mu Buzima, Kigali, Rwanda

[4]School of Nursing and Midwifery, University of Rwanda, Kigali, Rwanda

[5]Irembo, Kigali, Rwanda

[6]Rwanda Social Security Board (RSSB), Kigali, Rwanda

[7]Tulane School of Public Health and Tropical Medicine, New Orleans, LA, United States

Corresponding Author:
James Humuza, PhD, MD
School of Public Health, University of Rwanda, KK 19 Ave, Kigali, Rwanda

Abstract

Background: Digital health innovations address health care accessibility challenges in low- and middle-income countries. Babyl, Rwanda's largest telemedicine platform, reached 450 of 510 health facilities and enrolled 2 million patients before halting in September 2023 for system redesign. Limited research has explored implementation experiences and user perspectives that influenced its sustainability.

Objective: This study aims to explore user experiences and implementation lessons from Babyl's digital health platform, examining drivers that supported or hindered adoption and scale-up. This qualitative study uniquely examines the lived experiences of diverse stakeholders, active users, lapsed users, nonusers, health care providers, and Babyl agents to understand implementation challenges that contributed to the platform's halt.

Methods: A qualitative, cross-sectional study used 20 focus group discussions (FGDs) and 32 key informant interviews (KIIs) across 12 health centers in ten districts with diverse utilization rates, geographic locations, and Babyl agent availability. FGDs captured collective community perspectives while KIIs provided in-depth individual experiences, enabling data triangulation. FGDs included active users, lapsed users, registered nonusers, and eligible nonregistrants. KIIs involved health center heads, health care providers, and Babyl agents. Data were analyzed using thematic analysis following Braun and Clarke's framework. Data saturation was achieved when no new themes emerged from the last 3 FGDs and 5 KIIs. All transcripts were validated through member checking with a subset of participants, and intercoder reliability was established with a Cohen kappa of 0.82 across 2 independent coders.

Results: Five themes emerged: (1) knowledge and perceptions of digital health, (2) enablers and barriers to utilization, (3) experience and satisfaction, (4) benefits, and (5) improvement suggestions. Participants held positive perceptions of digital health for improving access and reducing wait times. Key enablers included qualified providers, convenience, privacy, and Babyl agents. Major barriers included negative perceptions of remote care quality, service delays, limited digital literacy, device access challenges, and inadequate health facility integration. Users reported high satisfaction with consultations but experienced process confusion. Patient and provider perspectives diverged: patients emphasized convenience, while providers expressed concerns about diagnostic limitations without physical examination. Digital literacy and smartphone access were pronounced barriers among rural and older participants. Recommendations included community mobilization, universal agent deployment, expanded coverage, and sustainable financing.

Conclusions: Multiple implementation challenges at individual, community, health system, and policy levels contributed to Babyl's discontinuation. Critical lessons include the importance of genuine health system integration, sustainable financing, stakeholder engagement, and gradual scaling. Findings provide insights for Rwanda's health sector digitalization and other African nations investing in telemedicine platforms.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e84832) doi:10.2196/84832

KEYWORDS

digital health; telemedicine; implementation science; qualitative research; Rwanda; low- and middle-income countries; health system integration

Introduction

Digital health innovations have emerged as transformative solutions to address health care accessibility challenges, particularly in low- and middle-income countries (LMICs) where geographic barriers, provider shortages, and infrastructure limitations impede effective health care delivery [1]. The World Health Assembly of the World Health Organization (WHO) unanimously adopted the use of telehealth in 2005 as a means of improving access to specialized health care services and contributing towards achieving Universal Health Coverage goals [2,3]. The expansion of mobile phone usage in emerging markets has created unprecedented opportunities for telemedicine platforms to revolutionize health care access and convenience, offering remote consultation services that can bridge the gap between patients and health care providers [4,5].

Rwanda's health care landscape exemplifies both the potential and challenges of implementing digital health solutions at scale. The country's well-structured health care system operates through a hierarchical model, with 499 health centers (HCs) serving as the backbone of primary health care delivery, addressing approximately 80% of the disease burden at the peripheral level [6,7]. Each HC provides a package of services, including curative, preventive, promotional, and rehabilitation services, while supervising health posts and community health workers in their catchment areas. This systematic approach to health care delivery created a conducive environment for digital health integration.

Babyl, an innovative digital health care platform, was introduced in Rwanda with the mission to transform health care service delivery through digital consultations facilitated by experienced health care providers. Leveraging Rwanda's high mobile phone penetration rate, Babyl's platform connected citizens with health care professionals through artificial intelligence–powered chatbots that triaged patients' medical concerns, offered treatment recommendations, and facilitated appointments with remote physicians based on symptom severity [7,8]. The platform's integration with Rwanda's existing health care infrastructure represented a significant commitment to leveraging technology for improved health outcomes, underscored by the Government of Rwanda's initial 10-year partnership agreement with the platform [8].

The scale of Babyl's implementation was remarkable, reaching 450 out of 510 primary health facilities across Rwanda, representing over 80% of the primary health care market coverage before it halted in September 2023 as the system is being redesigned. By August 2020, the platform had enrolled 2 million patients, with approximately half having actively used the service to seek care. The service primarily catered to Community-Based Health Insurance (CBHI) members, who comprised over 80% of Babyl's client base, alongside Rwandaise d'Assurance Maladie members and private clients. Rwandaise d'Assurance Maladie is the public medical insurance scheme that primarily covers employees in the public sector and also offers coverage to private sector employees and pensioners. It is coordinated by the Rwanda Social Security Board [9].

Despite the promising scale of implementation, the platform's halt in September 2023 for system redesign highlighted critical gaps in understanding the complex factors that influence the adoption, utilization, and sustainability of large-scale digital health interventions. While previous studies, including Dalberg's initial impact assessment, examined effects on access, equity, patient satisfaction, and cost-effectiveness compared with traditional care [7,10-12]. These evaluations did not provide insights into the deeper implementation challenges and user experiences that ultimately determine the success or failure of such platforms.

Notably, this study distinguishes itself from prior evaluations by examining the lived experiences and perspectives of diverse stakeholder groups, including active users, lapsed users, nonusers, health care providers, and Babyl agents, to understand the multifaceted barriers that contributed to the platform's discontinuation. Unlike previous quantitative assessments focused primarily on utilization metrics and clinical outcomes, this qualitative investigation explores the contextual, social, and systemic factors that shaped user experiences and implementation outcomes. This study contributes to the literature through comprehensive stakeholder inclusion, implementation-focused analysis, and multi-level examination of factors influencing digital health platform adoption and sustainability.

The complexity of implementing digital health solutions extends beyond technical capabilities to encompass individual, community, and health system factors that can either facilitate or hinder adoption. Health care providers must adapt to new diagnostic approaches without physical examination capabilities, potentially altering clinical decision-making processes and provider-patient interactions [13,14]. Patients may face challenges in communicating illness concerns and treatment preferences through digital platforms, while also navigating new technologies that require different levels of digital literacy and access to reliable communication infrastructure.

Understanding these multifaceted implementation experiences becomes crucial as countries across sub-Saharan Africa and other low- and middle-income settings continue to explore and invest in digital health solutions [15,16]. The WHO Digital Health Guidelines emphasize the importance of understanding contextual factors and user experiences in digital health implementation [17]. The COVID-19 pandemic accelerated the global adoption of telemedicine and digital care models, making

XSL•FO
RenderX

it imperative to learn from large-scale implementations like Babyl to inform future digital health initiatives [18].

This study uniquely addresses critical knowledge gaps by providing the first comprehensive, multi-stakeholder qualitative evaluation of Rwanda's Babyl telemedicine platform, explicitly contrasting perspectives across patients (active users, lapsed users, registered nonusers, and eligible nonregistrants), health care providers, HC administrators, and Babyl agents. Unlike previous evaluations focused primarily on utilization metrics or single stakeholder groups, this study systematically examines the complex interplay between individual, community, health system, and technological factors shaping implementation outcomes, while explicitly linking demographic characteristics, including age, education, geographic location, and digital literacy, to adoption patterns and user experiences. These insights extend beyond Rwanda's context to inform digital health implementation strategies across sub-Saharan Africa and other LMICs pursuing universal health coverage through technological innovations.

Methods

Study Design and Setting

This qualitative, cross-sectional study was conducted across Rwanda's health care system to explore key drivers that supported or hindered the adoption and scale-up of Babyl's digital health services in Rwanda. The study used focus group discussions (FGDs) and key informant interviews (KIIs) to capture perspectives from multiple stakeholder groups involved in or affected by the digital health platform implementation. The combination of FGDs and KIIs was strategically chosen to capture both collective community perspectives and in-depth individual experiences. FGDs enabled exploration of shared experiences, social norms, and community-level barriers through group interactions, while KIIs provided detailed accounts from health care providers and program implementers who could speak to operational challenges and system-level implementation issues. This methodological triangulation strengthened the validity and comprehensiveness of findings by allowing cross-verification of themes across different data sources and participant perspectives.

Sampling Strategy and Participant Selection

A purposive, maximum variation sampling strategy was used to ensure representation of diverse experiences and perspectives across different contexts. The study used a multi-stage sampling approach. First, HCs were purposively selected based on three criteria: (1) Babyl service utilization rates (high, medium, and low), (2) geographic distribution across Rwanda's provinces, and (3) presence or absence of Babyl agents at the facility level. This stratification ensured representation of diverse implementation contexts and user experiences.

Twelve HCs were selected across 10 districts, distributed as follows: two from the City of Kigali (urban), four from the Eastern Province (mixed urban-rural), three from the Southern Province (predominantly rural), two from the Western Province (rural), and one from the Northern Province (rural). This geographic distribution captured variations in infrastructure, digital connectivity, and socioeconomic contexts that might influence digital health service adoption.

Within each selected HC catchment area, recruitment occurred through multiple channels. HC staff identified potential participants from patient registers and community health worker networks. Community meetings were held at centrally located venues to explain the study's purpose and answer questions. Interested individuals were screened for eligibility using a structured screening form. Recruitment continued until data saturation was achieved and the target sample size was reached. For FGDs, four distinct user categories were identified:

1. Active users: CBHI members who had registered for and actively used Babyl services within the past 6 months.
2. Lapsed users: CBHI members who had registered but stopped using Babyl services.
3. Registered nonusers: CBHI members who registered for Babyl but never used the services.
4. Eligible nonregistrants: CBHI members aware of Babyl services but never registered.

Participants were eligible if they were 18 years or older, CBHI members in the selected catchment areas, and able to provide informed consent. For KII participants, eligibility criteria included being used at the selected HC or serving as a Babyl agent for at least 6 months, ensuring adequate experience with the platform implementation. For KIIs, three categories of key informants were purposively selected:

1. HC heads: Administrative leaders overseeing HC operations.
2. Health care providers: Nurses, clinical officers, and physicians who interacted with Babyl-referred patients.
3. Babyl agents: Community-based facilitators who assisted users with platform registration and navigation.

The final sample size of 192 participants (160 in FGDs and 32 in KIIs) was determined based on several criteria: (1) Data saturation confirmed when no new themes emerged from the final 3 FGDs and 5 KIIs; (2) Representation of all four user categories across diverse geographic settings; (3) Multi-site coverage, including stakeholders from all 12 HCs; (4) Methodological guidelines suggesting 15‑20 FGDs for thematic saturation in multi-site studies; and (5) Distribution of 4‑5 FGDs per user category. The sample size exceeded the minimum recommendations.

Data Collection and Procedures

Data collection was conducted between October 2022 and January 2023 by a team of 6 trained research assistants with prior qualitative research experience. All data collectors underwent a 3-day training covering qualitative interviewing techniques, research ethics, study objectives, and the semi-structured interview guides. Mock interviews were conducted to standardize data collection approaches across team members.

The research team comprised public health researchers with backgrounds in health systems strengthening, digital health, and implementation science. The principal investigator is a PhD holder with 8 years of experience in health policy research. The

XSL•FO
RenderX

study coordinator is a female MPH graduate with expertise in qualitative methods. Research assistants included 4 females and 2 males, all with bachelor's degrees in public health. No team members had prior professional relationships with Babyl or the participants.

Semi-structured interview guides were developed based on a literature review, consultation with digital health experts, and the research team's expertise in health systems and implementation science. The interview guides were adapted from WHO's framework for digital health implementation research [19] and incorporated constructs from the Consolidated Framework for Implementation Research [20]. The guides were pretested with 5 users and 2 health care providers from a nonstudy HC. Based on feedback, questions were refined for clarity, cultural appropriateness, and to reduce technical jargon.

The guides covered key domains including: (1) awareness and perceptions of digital health services, (2) facilitators and barriers to adoption, (3) user experiences with service delivery processes, (4) satisfaction with care quality, (5) perceived benefits and challenges, and (6) recommendations for improvement. Content validity was established through expert review by 3 implementation science researchers and 2 clinicians with telemedicine experience. Interview guides are provided as Multimedia Appendices 1-4.

Each FGD included 6‑10 participants of the same user category to promote open discussion and ensure homogeneity within groups. FGDs were conducted in Kinyarwanda (local language) at community meeting spaces near the HCs, lasted 60‑90 minutes, and were audio-recorded with participant consent. Two team members facilitated each FGD: one as moderator and one as note-taker to ensure comprehensive documentation of verbal and nonverbal data.

KIIs were conducted individually in private rooms at HCs or Babyl agent offices, lasted 45‑60 minutes, and were also audio-recorded. Interviews were scheduled at times convenient for participants to minimize disruption to their work responsibilities.

Field notes were taken during all interviews to capture contextual information, nonverbal cues, and preliminary observations. A reflexive journal was maintained by the research team lead to document potential biases, decision-making processes, and emerging insights throughout data collection.

In addition to qualitative data, demographic and contextual information were collected from all participants using structured questionnaires administered prior to FGDs and KIIs. For FGD participants, the demographic questionnaire captured age, gender, education level, geographic location (urban/rural), smartphone ownership status, digital literacy self-assessment, health insurance type, and Babyl service utilization history. For KII participants, the questionnaire captured professional role, years of professional experience, duration of Babyl involvement, health facility location, and service coverage characteristics.

Data Analysis

All audio recordings were transcribed verbatim in Kinyarwanda within 48 hours of data collection by professional transcribers trained in qualitative transcription standards. Transcripts were then translated into English by bilingual team members. A back-translation process was used for 20% of transcripts to ensure translation accuracy and conceptual equivalence. Transcripts were anonymized, with all identifying information removed and replaced with participant codes.

Transcript quality was verified through a validation process where the lead researchers reviewed 30% of transcripts against audio recordings. Additionally, summary transcripts were shared with a subset of 15 participants for member checking to ensure accurate representation of their views.

Data were analyzed using iterative, inductive thematic analysis following Braun and Clarke's 6-phase framework [18] : (1) familiarization with data, (2) generating initial codes, (3) searching for themes, (4) reviewing themes, (5) defining and naming themes, and (6) producing the report.

Two researchers independently coded the first 4 transcripts to develop an initial codebook. Codes were compared and discussed to achieve consensus, resulting in a refined codebook with clear code definitions. This codebook was then applied to all transcripts using NVivo (version 12; Lumivero) qualitative analysis software. Intercoder reliability was assessed on 20% of transcripts coded independently by 2 researchers, achieving a Cohen kappa of 0.82, indicating strong agreement.

The coding process began with open coding, where segments of text were assigned descriptive codes. A total of 127 initial codes were identified during open coding. These codes were then grouped into broader categories through axial coding, examining relationships and patterns. Finally, selective coding identified overarching themes that captured the essence of participants' experiences and perspectives.

The analysis was conducted iteratively, with constant comparison between emerging themes and raw data. Deviant cases were actively sought and analyzed to ensure a comprehensive understanding. Regular team meetings were held to discuss emerging themes, resolve discrepancies, and refine the thematic structure.

Thematic saturation was formally assessed by monitoring theme emergence across sequential data collection phases. Data collection occurred in 3 waves. When analysis of the final 3 FGDs and 5 KIIs revealed no new themes, data collection was concluded.

Data credibility was enhanced through (1) methodological triangulation using both FGDs and KIIs, (2) investigator triangulation with multiple researchers involved in coding and analysis, (3) participant validation through member checking, (4) peer debriefing sessions with external qualitative research experts, and (5) maintaining a detailed audit trail documenting all analytical decisions.

Quantitative data were analyzed using descriptive statistics and chi-square tests to examine associations between demographic characteristics and technology access patterns. Statistical significance was set at $P<.05$. These quantitative findings are presented alongside qualitative data to provide a comprehensive understanding of barriers.

XSL•FO
RenderX

Ethical Considerations

Ethical approval for this study was obtained from the University of Rwanda, College of Medicine and Health Sciences Institutional Review Board (approval number 335/CMHS IRB/2022). Additional administrative clearance was granted by the Rwanda Ministry of Health and participating district health offices. The study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki.

Written informed consent was obtained from all participants after providing detailed information about the study purpose, procedures, risks, benefits, and their right to withdraw at any time without consequences. For participants with limited literacy, the consent form was administered orally in the presence of a witness, and verbal consent was audio-recorded.

FGD participants received 5000 Rwandan francs (approximately US $5) as compensation for transportation and time. KII participants (health care providers, administrators, and Babyl agents) were interviewed during working hours with supervisor approval and did not receive monetary compensation. Refreshments were provided during all FGDs and interviews.

Participant privacy and confidentiality were ensured by conducting interviews in private settings, anonymizing transcripts, and storing audio recordings and transcripts on password-protected computers accessible only to the research team. Data will be securely stored for 5 years before permanent deletion. No adverse events were reported during the study.

Results

Overview

The analysis of 20 FGDs and 32 KIIs revealed five major themes that emerged from combining client and health facility personnel responses: (1) knowledge and perceptions of digital health services, (2) enablers and barriers to using digital services, (3) experience and satisfaction using Babyl digital services, (4) benefits of using digital services, and (5) suggestions to improve uptake of digital services. The KIIs also provided opportunities to better understand Babyl's operations related to agent recruitment processes, client facilitation experiences, and engagement with health facilities.

Participant Characteristics

The demographic distribution of FGD and KII participants is summarized in Tables 1 and 2, respectively.

Table . Demographic characteristics of focus group discussion participants (n=160).

Characteristic	Active users (n=40)	Lapsed users (n=40)	Registered nonusers (n=40)	Eligible nonregistrants (n=40)
Sex, n (%)				
Female	25 (62.5)	27 (67.5)	28 (70.0)	21 (52.5)
Male	15 (37.5)	13 (32.5)	12 (30.0)	19 (47.5)
Age group (years), n (%)				
18 - 30	12 (30.0)	8 (20.0)	6 (15.0)	10 (25.0)
31 - 50	26 (65.0)	28 (70.0)	30 (75.0)	28 (70.0)
>50	2 (5.0)	4 (10.0)	4 (10.0)	2 (5.0)
Age (years), mean (SD)	35.2 (8.4)	36.8 (9.1)	38.4 (8.7)	42.1 (10.2)
Education level, n (%)				
Primary or less	10 (25.0)	15 (37.5)	25 (62.5)	28 (70.0)
Secondary	24 (60.0)	20 (50.0)	13 (32.5)	10 (25.0)
Tertiary	6 (15.0)	5 (12.5)	2 (5.0)	2 (5.0)
Location, n (%)				
Urban	11 (27.5)	9 (22.5)	10 (25.0)	6 (15.0)
Rural	29 (72.5)	31 (77.5)	30 (75.0)	34 (85.0)
Phone ownership, n (%)				
Own phone	38 (95.0)	36 (90.0)	33 (82.5)	27 (67.5)
Shared/no phone	2 (5.0)	4 (10.0)	7 (17.5)	13 (32.5)
Insurance type, n (%)				
CBHI[a]	35 (87.5)	32 (80.0)	33 (82.5)	34 (85.0)
Other/None	5 (12.5)	8 (20.0)	7 (17.5)	6 (15.0)

[a]CBHI: Community-Based Health Insurance.

XSL•FO
RenderX

Table . Characteristics of key informant interview participants (n=32).

Characteristic	Health center heads (n=6)	Health care providers (n=18)	Babyl agents (n=8)
Sex, n (%)			
Female	2 (33.3)	10 (55.6)	5 (62.5)
Male	4 (66.7)	8 (44.4)	3 (37.5)
Age group (years), n (%)			
25 - 35	1 (16.7)	6 (33.3)	4 (50.0)
36 - 45	3 (50.0)	9 (50.0)	3 (37.5)
>45	2 (33.3)	3 (16.7)	1 (12.5)
Age (years), mean (SD)	42.5 (7.2)	38.4 (6.8)	34.2 (5.9)
Professional role, n (%)			
Nurses	—[a]	12 (66.7)	—
Lab technicians	—	4 (22.2)	—
Pharmacists	—	2 (11.1)	—
Facility-based	6 (100)	—	8 (100)
Years of experience, n (%)			
<2	0 (0)	2 (11.1)	3 (37.5)
2 - 5	1 (16.7)	8 (44.4)	5 (62.5)
6 - 10	3 (50.0)	6 (33.3)	0 (0)
>10	2 (33.3)	2 (11.1)	0 (0)
Years of experience, mean (SD)	10.2 (4.8)	6.3 (4.2)	2.8 (1.5)
Health center type, n (%)			
Urban	2 (33.3)	7 (38.9)	3 (37.5)
Rural	4 (66.7)	11 (61.1)	5 (62.5)
Babyl experience (years), n (%)			
<1	1 (16.7)	4 (22.2)	1 (12.5)
1 - 2	2 (33.3)	8 (44.4)	3 (37.5)
>2	3 (50.0)	6 (33.3)	4 (50.0)
Babyl experience (years), mean (SD)	2.4 (1.1)	1.8 (0.9)	2.1 (0.8)

[a]Not applicable.

A total of 192 participants were enrolled in this study, comprising 160 individuals in 20 FGDs and 32 key informants in individual interviews (KIIs). Among FGD participants, 98 (61.3%) were female, and 62 (38.7%) were male. The majority (112/160, 70.0%) were aged 31 - 50 years, with 36 (22.5%) participants aged 18 - 30 years, and 12 (7.5%) participants older than 50 years.

Demographic characteristics significantly influenced digital health adoption patterns. Twenty-eight younger participants (18 - 30 y) demonstrated greater digital literacy and smartphone ownership (76.5%) compared with 4 participants older than 50 years (34.8%), facilitating easier platform adoption. Among the 12 participants older than 50 years, 11 (91.7%) participants reported difficulty navigating the digital platform. Female participants more frequently emphasized privacy as a key benefit, particularly for reproductive health consultations. Male participants more commonly highlighted time efficiency and workplace convenience. Rural participants (92/160, 57.5% of FGDs) faced compounded barriers: lower smartphone ownership (34.8% vs 76.5% in urban areas), poorer network connectivity (78.3% reported “poor” or “very poor” connection vs 23.5% urban), and limited Babyl agent presence (available in only 2 of 7 rural HCs vs all 5 urban centers).

Among active users (n=40), most were CBHI members (35/40, 87.5%), aged 25 - 45 years, predominantly female (25/40, 62.5%), rural (29/40, 72.5%), and phone owners (38/40, 95%), with higher digital literacy and positive digital health experiences. Lapsed users (n=40) had similar demographic profiles but reported technology challenges and service dissatisfaction. Registered nonusers (n=40) were mainly female (28/40, 70%), older (mean age of 38 y), with lower digital literacy (25/40, 62.5% had only primary education), and largely

XSL•FO
RenderX

rural (30/40, 75%). Eligible nonregistrants (n=40) were the oldest group (mean age of 42 y) with the lowest digital literacy and phone ownership (27/40, 67.5%), often relying on shared devices.

Table 2 shows that key informants had substantial professional experience and adequate exposure to Babyl services to provide informed perspectives. HC heads had the longest professional experience (mean 10.2 y), providing institutional memory and leadership perspective on digital health integration. Health care providers, predominantly nurses (12/18, 66.7%), had moderate experience (mean 6.3 y) and represented frontline staff managing the interface between digital and traditional services. Babyl agents were younger (mean age 34.2 y), had shorter professional tenure (mean 2.8 y), but substantial Babyl-specific experience (mean 2.1 y), positioning them to understand both technical platform issues and community-level barriers. The predominance of rural HCs (61.1% - 66.7% across groups) ensured representation of contexts where implementation challenges were most pronounced. Variation in Babyl experience duration (1 - 2+ years across groups) captured perspectives from both early implementation and more mature operational phases.

Table 3 presents key findings organized by stakeholder group, revealing both convergent and divergent perspectives across patients/users, health care providers, HC heads, and Babyl agents. While all stakeholders recognized Babyl's potential to improve health care access, critical disconnects emerged: patients emphasized front-end barriers, including registration complexity and agent availability; health care providers focused on back-end challenges including workflow integration and lack of access to consultation records; HC heads highlighted coordination difficulties and unclear protocols, and Babyl agents identified gaps in training, equipment, and support systems. These divergent priorities reflect each group's position within the implementation ecosystem and underscore the importance of multi-stakeholder engagement in digital health design and implementation.

XSL•FO
RenderX

Table . Key findings by stakeholder group.

Theme	Patients/Users (n=160)	Health care providers (n=12)	Health center heads (n=6)	Babyl agents (n=14)
Knowledge and Perceptions	Mixed awareness levels; valued convenience but had an incomplete understanding of coverage and eligibility; trust concerns about remote diagnosis quality	Skepticism about diagnostic accuracy without physical examination; concerns about consultation quality and clinical decision-making limitations	Recognized potential for expanded access but concerned about facility burden, resource allocation, and unclear role in digital health integration	Strong belief in service value and community health impact; frustration when unable to assist users due to system limitations or lack of support
Enablers	Convenience and time savings; privacy for sensitive health issues; cost savings from avoided travel; access to qualified doctors; medication prescription and dispensing	Potential for reduced outpatient department congestion; access to physician consultation for complex cases; opportunity to refer patients requiring specialist input	Expanded health care access for the catchment population; potential to reduce referrals to district hospitals; alignment with national health coverage goals	Direct patient interaction and community engagement; satisfaction from helping users access health care; visible community health impact when services worked well
Barriers	Digital literacy gaps and limited smartphone access; agent absence at health centers; registration process complexity; poor network connectivity (especially rural); service coverage exclusions (chronic diseases, pediatrics)	No access to Babyl consultation records; medication stock-outs disrupting prescribed treatments; workflow disruptions from parallel service delivery; lack of feedback mechanisms to learn from consultations; diagnostic uncertainty without physical examination	Unclear integration protocols and coordination mechanisms; additional workload without corresponding resources; medication supply chain issues; ambiguous accountability when problems arose	Inadequate technical and patient education training; equipment failures and resource constraints; inconsistent presence at facilities; blamed for system issues beyond their control; limited escalation support
Experience and Satisfaction	High satisfaction when services functioned properly; frustration with technical issues, delays, and process complexity; pronounced age-related challenges (older adults required agent assistance 100%); lapsed users cited process friction exceeding perceived benefit	Frustration with incomplete patient information arriving at facilities; inability to provide continuity of care without consultation histories; diagnostic uncertainty affecting clinical confidence; felt Babyl operated as a parallel rather than an integrated system	Difficulty coordinating between Babyl and facility-based services; unclear accountability and escalation pathways when problems arose; concerned about sustainability and long-term resource implications	Satisfaction when helping users successfully; demotivation from facing blame for system failures; limited technical support for troubleshooting; felt undervalued despite frontline role
Recommendations	Universal agent availability at all health centers; simplified registration process; expanded service coverage (chronic diseases, pediatrics); improved rural network infrastructure; clearer information about eligibility and services	Full health management information system integration; access to consultation histories and prescribed treatments; regular coordination meetings with Babyl; clear referral and feedback protocols; assurance of medication availability	Clear operational guidelines and integration protocols; regular stakeholder coordination meetings; dedicated resources for digital health integration; sustainable financing clarity and commitment	Enhanced training (technical troubleshooting + patient education); reliable equipment and backup systems; consistent facility presence and defined roles; clear escalation procedures; recognition and support for frontline work

Theme 1: Knowledge and Perceptions of Digital Health Services

Overview

Three common subthemes emerged: (1) sources of awareness about Babyl services, (2) perceptions of service quality and legitimacy, and (3) understanding of eligibility and coverage. These subthemes reveal how information dissemination and initial perceptions shaped enrollment decisions.

Sources of Awareness

Awareness of Babyl services came primarily through 3 channels. Health facility visits were the most common source, where patients learned about services during routine consultations or registrations. Radio advertisements were mentioned in 7 of 8 urban FGDs but only 3 of 8 rural FGDs, indicating geographic disparities in media reach. One participant stated,

> *I heard about it from the radio. They said we can talk to doctors without going to the health center.* [active user, FGD participant #3]

Word-of-mouth through community networks was particularly important in rural areas, where 6 of 8 FGDs cited neighbors or community health workers as information sources. Registered nonusers and eligible nonregistrants frequently mentioned incomplete or unclear information as a reason for nonengagement. As one participant explained,

> *I heard about it, but I wasn't sure if it was free with my insurance.* [eligible nonregistrant, FGD participant #8]

Perceptions of Quality and Legitimacy

Perceptions of service quality varied substantially across user categories. Active users expressed confidence in the

XSL•FO
RenderX

qualifications of Babyl doctors, emphasizing their ability to diagnose and prescribe appropriately. A typical comment was,

> *They have doctors, real doctors who can prescribe medicine.* [registered nonuser, FGD participant #2]

However, skepticism was common among nonregistrants and registered nonusers, who questioned whether remote consultations could match in-person care. This skepticism was more pronounced in rural areas (6 of 8 rural FGDs) compared with urban areas (3 of 8 urban FGDs) and among older participants (>50 y).

While 89% of CBHI members had heard of Babyl services, only 34% could accurately describe the service scope and eligibility. This awareness-understanding gap created adoption barriers across user groups. Active users learned through trial and error with agent assistance, while registered nonusers held persistent misunderstandings (eg, believing services required payment). Provider skepticism about remote diagnosis reinforced community doubts, creating a negative cycle where patients avoided services that providers did not endorse. This demonstrates that awareness campaigns alone are insufficient; comprehensive education addressing service scope, appropriate use cases, and complementarity with facility-based care is essential.

Theme 2: Enablers and Barriers to Digital Service Utilization

Overview

Four subthemes emerged operating at different levels: (1) individual-level factors, including technology access and digital literacy, (2) community-level factors such as social norms and trust, (3) health system factors, including integration with existing services, and (4) technological infrastructure. Findings are organized by stakeholder perspectives to highlight divergent experiences.

Patient Perspectives on Enablers

Patients, particularly active users, identified several key enablers. Convenience and time savings were universally emphasized, with one participant noting,

> *I don't have to travel to the health center or wait in line.* [active user, FGD participant #5]

Privacy emerged as a significant enabler, especially for female participants (mentioned in 68% of female vs 42% of male participant comments) and for sensitive health concerns. As one woman explained,

> *I can get advice without anyone knowing my problem.* [active user, FGD participant #9]

For working-age adults (31 - 50 y), the ability to consult without missing work was critical.

> *I don't have to miss work to see a doctor.* [active user, FGD participant #10]

Access to qualified doctors and prescribed medications was an additional enabler, particularly valued in rural areas where specialist access is limited.

Patient Perspectives on Barriers

Demographic and digital literacy data were collected using a structured questionnaire administered to all FGD participants (n=160). The questionnaire captured age, gender, education level, smartphone ownership, phone sharing patterns, and self-reported digital literacy. Quantitative data were analyzed using descriptive statistics and chi-square tests to examine associations between demographic characteristics and technology access patterns. Statistical significance was set at $P<.05$. These quantitative findings are presented alongside qualitative data to provide a comprehensive understanding of barriers.

Individual-level barriers were substantial and strongly correlated with demographic characteristics. Technology access challenges included a lack of smartphone ownership (affecting 65.2% of adults older than 50 y old vs 23.5% of 18 - 30 y olds), reliance on shared devices (creating privacy concerns), and device limitations. One lapsed user described,

> *My phone is very old and slow. The app takes too long to open.* [lapsed user, FGD participant #2]

Smartphone ownership ranged from 76.5% (18 - 30 y) to 34.8% (>50 y), $P<.001$. Rural participants older than 45 years had lower ownership (23.8%) than urban counterparts (71.4%, $P<.001$). Educational attainment predicted ownership: 85.7% secondary-educated versus 38.9% primary-educated owned smartphones ($P<.001$). Female participants were 2.3× more likely to share phones (62.5% vs 27.4% male, $P<.001$), creating privacy barriers for reproductive health discussions.

Digital literacy barriers were pronounced, with 91.7% of participants older than 50 years reporting difficulty using smartphone apps compared with 22.2% of 18 - 30-year-olds. An eligible nonregistrant stated,

> *I don't know how to use these apps. I need someone to help me.* [eligible nonregistrant, FGD participant #8]

Digital literacy showed a 4-fold age gradient: 91.7% of those older than 50 years needed continuous assistance versus 22.2% of 18 - 30 years ($P<.001$). Educational patterns were equally pronounced: 89.7% primary-educated versus 23.4% secondary-educated required assistance ($P<.001$). All 8 participants, both older than 50 years and primary-educated, required help for every interaction, averaging 4.2 registration attempts.

Process complexity deterred many potential users, particularly the registration process, which required multiple steps, including phone verification, insurance validation, and profile setup.

> *I tried to register but got confused with all the steps.* [registered nonuser, FGD participant #6]

Trust concerns centered on data privacy, consultation quality without physical examination, and medication dispensing procedures. Service coverage limitations, particularly the exclusion of chronic diseases (diabetes, hypertension) and pediatric care, created significant barriers. Network connectivity issues disproportionately affected rural users, with 78.3%

XSL•FO
RenderX

reporting poor or intermittent connections compared with 23.5% of urban users.

> *The network doesn't work well in my village.* [eligible nonregistrant, FGD participant #3]

Health Care Provider Perspectives

Health care providers identified both enablers and barriers from their operational standpoint. Key enablers included reduced outpatient department (OPD) congestion for simple cases and access to physician expertise for complex cases. As one provider explained,

> *For patients who only need simple consultations, this reduces our workload.* [health care provider, KII]

However, providers raised concerns about diagnostic limitations, noting,

> *Without physical examination, it's difficult to make accurate diagnoses.* [health center head, KII]

Workflow disruptions occurred when Babyl patients arrived at HCs without proper documentation or when consultation records were not accessible to facility staff.

> *We don't know what the doctor prescribed or why patients were referred.* [health care provider, KII]

Medication stock-outs created frustration for both providers and patients.

> *Sometimes the medication is out of stock when patients come to collect it.* [health care provider, KII]

The lack of feedback mechanisms meant providers could not learn from Babyl consultations or understand clinical outcomes.

Babyl Agent Perspectives

Babyl agents identified community relationships and collaborative HC staff as key enablers, but faced significant operational barriers. Equipment inadequacy was a persistent challenge.

> *Sometimes the equipment doesn't work, and I have to use my own phone.* [Babyl agent, KII]

Agents experienced blame for system failures beyond their control, affecting their motivation and community relationships.

> *When the system fails, or verification takes long, patients blame me.* [Babyl agent, KII]

Insufficient training on both technical troubleshooting and patient education limited their effectiveness. The absence of dedicated agents at many HCs (particularly rural facilities) was identified as a critical gap, with patients having no one to assist with registration or resolve issues.

Barriers compound multiplicatively for vulnerable populations. Older rural women with primary education faced simultaneously: limited smartphone ownership, agent absence, poor connectivity, registration complexity, and service coverage gaps. These disadvantages are concentrated among populations with the greatest potential benefit. While patients emphasized front-end barriers (registration, digital literacy, agent availability), providers focused on back-end challenges (workflow disruption, information access), indicating multilevel intervention needs. That 100% of participants older than 50 years required agent assistance, yet agents were absent from 41.7% of HCs, exemplifies how self-service assumptions created insurmountable barriers for vulnerable groups.

Theme 3: Experience and Satisfaction Using Babyl Services

Overview

Three subthemes characterized user experiences: (1) registration and enrollment processes, (2) consultation quality and clinical interactions, and (3) factors influencing continued use or discontinuation. Experiences varied significantly by user category and demographic characteristics.

Patient Registration Experiences

Age was the strongest predictor of registration experience. Younger participants (18‑30 y) predominantly enrolled independently (89.5%), using smartphones and navigating the app without assistance.

> *I was able to register easily on my own using the app.* [active user, FGD participant #2]

In contrast, 100% of participants older than 50 years required agent assistance for registration.

> *The agent helped me step by step to register and call.* [active user, FGD participant #6]

This agent dependency created a critical vulnerability when agents were absent, older adults could not access services.

Registration success was age-stratified: 89.5% of 18‑30 years registered independently versus 0% of older than 50 years (100% required assistance). Educational attainment was equally predictive: 74.3% secondary-educated versus 15.6% primary-educated registered independently. Rural participants required 2.8 versus 1.3 registration attempts (urban, $P<.001$). Independent registrants were 3.7× more likely to initiate consultations (67.2% vs 18.1%), suggesting registration difficulty created psychological barriers to utilization.

Consultation Experiences and Discontinuation

Among lapsed users, the most common pattern was initial satisfaction with clinical outcomes but frustration with process complexity, leading to discontinuation.

> *The medicine worked, but the process was too confusing to use again.* [lapsed user, FGD participant #9]

Circumstantial discontinuation occurred when users' health needs fell outside service coverage, particularly for chronic disease management.

> *I tried it once for a cough, but now I just go to the health center.* [lapsed user, FGD participant #2]

Registered nonusers typically completed enrollment but never initiated consultations due to persistent trust concerns or lack of immediate need.

XSL•FO
RenderX

Provider and Agent Experiences

Health care providers reported operational challenges, including the inability to access consultation records, medication stock-outs disrupting prescribed treatments, and a lack of clinical feedback limiting learning opportunities. Babyl agents experienced satisfaction when assisting successful users but felt blamed for system failures beyond their control, including verification delays and network connectivity problems.

User experiences revealed distinct trajectories: active users progressed through initial challenges (2‐3 attempts) to routine utilization; lapsed users reported satisfaction but discontinued due to process friction; registered nonusers completed enrollment but never initiated consultations. Age predicted experience quality through process navigation capability; older adults required sustained agent assistance. Without agent availability, consultation initiation was impossible regardless of need severity. Provider experiences paralleled patient frustrations: initial enthusiasm diminished when medication stock-outs and inability to access consultation records created operational challenges.

Theme 4: Benefits of Digital Health Services

Overview

Four subthemes emerged regarding benefits: (1) access to specialized care and qualified providers, (2) time and cost savings, (3) health system efficiency improvements, and (4) innovation and learning opportunities. Benefits were emphasized differently by different stakeholder groups.

Patient-Identified Benefits

Patients most frequently emphasized access to doctors and specialists.

> *I talked to a real doctor from Kigali who understood my problem.* [active user, FGD participant #4]

This access benefit was particularly valued in rural areas where specialist consultations require traveling to district or referral hospitals. Time savings were mentioned more frequently by male participants (72%) and working-age adults than by female or older participants. Cost savings, including transport, lost work time, and informal payments, were emphasized more in rural FGDs (7 of 8) compared with urban FGDs (3 of 8).

> *I saved money on transport and time away from my business.* [active user, FGD participant #9]

Privacy and convenience benefits allowed participants to seek care for sensitive health issues and at convenient times.

Benefit valuations varied by demographics. Male participants emphasized time savings more than females (72.6% vs 48.0%), while females emphasized privacy (68.4% vs 41.9%). Age influenced perceptions: younger participants valued convenience, middle-aged participants emphasized time/cost, and older participants valued doctor access. Rural participants saved 2000‐3000 RWF (approximately US $1.90-2.86) versus urban 500‐1000 RWF (approximately US $0.48-0.95) per avoided visit, reflecting greater travel distances.

Provider and Agent-Identified Benefits

Health care providers valued the ability to refer complex cases to physicians.

> *Patients can consult doctors from Kigali for complex cases.* [health center head, KII]

Reduced OPD congestion for routine consultations allowed providers to focus on patients requiring physical examination.

> *We can focus on patients who really need physical examination.* [health care provider, KII]

However, providers noted that the lack of feedback limited their learning from consultations. Babyl agents identified community health impact and the satisfaction of facilitating access as key benefits, noting that successful users often became service advocates.

Stakeholder benefit perceptions diverged critically. Patients emphasized individual benefits (time savings, cost reduction, and privacy); providers emphasized system-level advantages (reduced congestion and specialist access). However, providers' anticipated benefits remained unrealized due to implementation gaps. Babyl operated in parallel to facility services; patients still presented for medication collection. Inability to access consultation records created fragmented care. This disconnect created sustainability challenges: patients experienced sufficient benefit to continue utilization, but providers experienced more burdens than benefits, threatening institutional buy-in essential for scaling.

Theme 5: Suggestions for Service Improvement

Overview

Four subthemes emerged regarding improvement recommendations: (1) community-level awareness and engagement strategies, (2) health facility-level infrastructure and staffing, (3) service delivery modifications, and (4) policy and sustainability considerations. Recommendations varied by stakeholder role and context.

Community-Level Recommendations

Expanded awareness campaigns using multiple channels (radio, community meetings, health facility outreach) were universally recommended.

> *They should announce it at community meetings in villages.* [eligible nonregistrant, FGD participant #1]

Participants suggested deploying community health worker mobilizers in cells and villages, particularly in rural areas. Clear, simplified information about eligibility, covered services, costs, and registration steps was emphasized across all user categories.

Health Facility-Level Recommendations

The most critical recommendation was ensuring dedicated Babyl agents at all HCs, especially rural facilities.

> *There should be an agent at every health center to help people.* [lapsed user, FGD participant #4]

Additional needs included adequate equipment (smartphones, tablets, backup power), improved network infrastructure, comprehensive agent training covering technical troubleshooting

XSL•FO
RenderX

and patient education, and better integration of Babyl consultation records with health facility systems.

Service Delivery Recommendations

Service expansion to include pediatric consultations (for caregivers) was frequently requested. Participants suggested accommodating phone sharing through mechanisms like designated family members or community phone points. Partnerships with health posts for last-mile service delivery in remote areas were proposed. Chronic disease coverage was the most common service expansion request.

> *It should cover chronic diseases like diabetes and hypertension.* [registered nonuser, FGD participant #1]

Rural network infrastructure improvement was critical.

> *The network infrastructure in rural areas needs improvement.* [Babyl agent, KII]

Simplified registration processes, medication stock management improvements, and patient feedback mechanisms were additional operational recommendations.

Policy and Sustainability Considerations

Health care providers and HC heads emphasized sustainable financing mechanisms, clear governance and accountability structures, and formal integration into national health policies and strategies. These systemic recommendations reflected concerns about program continuity and scale-up.

Demographic Patterns in Digital Health Adoption

Demographics functioned as intersecting determinants, creating compounded barriers. Older rural women with primary education faced simultaneously: smartphone nonownership (65.2%), digital illiteracy (91.7% needed assistance), poor connectivity (78.3% rural), agent absence (5 of 7 rural facilities), and service coverage gaps. Successful adopters were advantaged: 74.6% secondary-educated (vs 29.7% nonregistrants, $P<.001$), 87.5% urban, 95.0% smartphone owners, mean age 33.2 years (vs 42.1, $P<.001$). Each additional disadvantage multiplicatively reduced adoption: 87.5% utilization (zero disadvantages) versus 12.3% (3‑4 disadvantages). Digital health primarily reached already-advantaged populations, potentially exacerbating inequities.

Thematic synthesis reveals critical interdependencies among findings that explain implementation outcomes and point toward multilevel intervention requirements. Knowledge gaps identified in theme 1 directly amplified barrier impacts documented in theme 2: participants unaware that Babyl agents could provide registration assistance perceived enrollment as impossible and never attempted it, while informed participants with equivalent digital literacy limitations successfully navigated identical technical challenges with agent support. Service experiences captured in theme 3 created feedback loops affecting subsequent adoption: positive experiences generated word-of-mouth promotion that encouraged community members' service uptake ("My neighbor used it successfully, so I tried it too," FGD #7), while negative experiences, particularly older adults' repeated inability to access services without agents, circulated through social networks actively deterring potential users from attempting enrollment. Perceived benefits in theme 4 remained theoretical rather than experienced for most participants because theme 2 barriers prevented service access; the value proposition was sound, but inaccessibility meant benefits could not be realized. Provider integration challenges documented in theme 2 directly undermined patient-perceived benefits highlighted in theme 4: medication unavailability at HCs negated Babyl consultations' convenience advantages when patients still required facility visits, while lack of clinical follow-up information prevented the care continuity that would optimize health outcomes. Participants' improvement recommendations in theme 5 implicitly recognized these interdependencies: universal agent deployment simultaneously addresses theme 1 knowledge gaps and theme 2 barrier navigation challenges; health management information system integration resolves theme 2 provider workflow challenges while enabling theme 3 positive experiences through care continuity; expanded service coverage addresses theme 2 coverage gaps while enhancing theme 4 benefit realization. The finding that no single intervention type dominated stakeholder recommendations participants proposed community-level awareness strategies, facility-level agent deployment and equipment improvements, service-level coverage expansion and process simplification, and policy-level financing and governance solutions—indicates that stakeholders intuitively understood that multilevel implementation challenges require correspondingly multilevel solutions. This systems perspective, emerging organically from participants' lived experiences rather than being imposed by implementation science frameworks, suggests that user-centered digital health design must attend to complex interactions among individual capabilities, community resources, health system structures, technological infrastructure, and policy environments.

Discussion

Principal Findings

This qualitative study provides insights into the user experiences and implementation lessons from Babyl's large-scale digital health platform in Rwanda, representing one of the most extensive telemedicine implementations in sub-Saharan Africa. Despite Babyl's impressive reach, covering 450 of 510 primary health facilities and enrolling 2 million patients, the eventual closure in September 2023 underscores the critical importance of understanding implementation challenges beyond initial uptake metrics [7]. Importantly, the findings from this study are now informing Rwanda's current redesign of its digital health strategy, providing evidence-based insights for developing more sustainable and effective digital health interventions. The findings reveal a complex interplay of individual, community, health system, and technological factors influencing digital health adoption, utilization, and sustainability, contributing valuable lessons for future initiatives in similar contexts and supporting Rwanda's broader health system strengthening goals [21,22].

This study advances digital health implementation science through three distinct contributions to the literature. First, the

XSL•FO
RenderX

comprehensive stakeholder inclusion approach sampled diverse groups, including active users, lapsed users (who discontinued services), registered nonusers (who never used despite registering), eligible nonregistrants (who were aware but never registered), health care providers, HC heads, and Babyl agents. Unlike previous evaluations that focused primarily on active users or clinical outcomes [7,10-12], this comprehensive sampling reveals the full spectrum of adoption patterns and identifies barriers across the complete user journey from awareness through sustained utilization. This methodological approach demonstrates that evaluating only successful adopters creates selection bias that obscures critical implementation barriers.

Second, the implementation-focused analytical framework examined why and how implementation succeeded or failed across different contexts, rather than solely measuring what outcomes were achieved. By examining facilitators, barriers, and contextual factors through an implementation science lens informed by WHO digital health guidelines [19] and the Consolidated Framework for Implementation Research [20], the study provides actionable insights for future digital health initiatives. This focus on implementation processes complements outcome-focused evaluations by revealing the mechanisms underlying utilization patterns.

Third, the multi-level analysis examined implementation challenges at individual levels (user characteristics, digital literacy), community levels (social norms, agent availability), health system levels (provider readiness, workflow integration), and policy levels (financing, governance structures). This holistic framework demonstrates that digital health implementation challenges cannot be addressed through single-level interventions. The finding that 100% of participants older than 50 years required agent assistance, yet 41.7% of HCs lacked agents, exemplifies how individual-level needs and system-level resource allocation interact to determine access outcomes. This multi-level perspective provides a comprehensive understanding of the complex, interacting factors that influenced Babyl's trajectory and ultimate discontinuation.

The findings reveal profound equity considerations that must inform future digital health implementations. The demographic analysis demonstrates that digital health adoption is not merely a matter of individual choice but is fundamentally shaped by structural inequities in age, education, gender, and geography. Older adults faced substantial barriers, including limited smartphone ownership and digital literacy challenges. Educational attainment strongly predicted utilization patterns, with active users more likely to have secondary or higher education. Geographic disparities were pronounced, with rural participants experiencing lower smartphone ownership and substantially worse network connectivity. Gender differences emerged particularly in privacy concerns, with female participants emphasizing confidential consultations more frequently than male participants. These overlapping disadvantages create compounded barriers for the most vulnerable populations; older, less educated, rural, female participants face multiple simultaneous obstacles to digital health access. Future digital health interventions must proactively address these equity considerations through targeted strategies, including age-appropriate interfaces, low-literacy design, substantial rural infrastructure investments, and gender-sensitive implementation approaches.

Participants' experiences highlight the critical importance of digital literacy support and trusted community channels for adoption. The finding that older participants consistently required agent assistance for registration, while younger participants more often enrolled independently, demonstrates that digital health platforms cannot rely on user self-service alone.

The absence of dedicated Babyl agents at many HCs, identified as the most critical implementation gap across all stakeholder groups, meant that older and less digitally literate potential users had no pathway to access services. Participants' emphasis on word-of-mouth promotion and community health worker involvement reflects the importance of trusted local channels for health information dissemination [8]. Reports of users deleting verification codes due to cyberfraud concerns illustrate how broader digital security anxieties can undermine adoption, suggesting that digital health implementations require accompanying digital literacy and security education components [23].

Health care provider perspectives revealed critical health system integration gaps that limited clinical effectiveness and provider buy-in. Providers reported an inability to access Babyl consultation records, creating continuity of care challenges and workflow disruptions when patients arrived at facilities. The frequent medication stock-outs for prescribed treatments undermined patient confidence and created provider frustration. Lack of feedback mechanisms meant providers could not learn from Babyl consultations or understand clinical outcomes, limiting professional development opportunities. While providers valued reduced OPD congestion for simple cases and access to physician expertise for complex cases, concerns about diagnostic limitations without physical examination and parallel rather than integrated service delivery prevented full provider engagement. These findings demonstrate that technical platform functionality is insufficient; digital health implementations require genuine integration with clinical workflows, information systems, and pharmaceutical supply chains [5,17].

Digital health implementation in LMICs faces distinct challenges, including health care provider shortages, limited infrastructure access, inadequate education opportunities, greater disease burden, and insufficient data management capabilities compared with high-income countries [24,25]. Babyl's experience can be contextualized within broader LMIC digital health implementations. India's e-Sanjeevani platform delivered over 10 million teleconsultations through government-led integration within the national digital health strategy, while Kenya's M-TIBA platform connected 4 million users with private-sector financing models integrated into existing health insurance schemes. Brazil's Portal Telemedicina demonstrates infrastructure integration connecting remote primary care providers with specialists nationwide [24,25]. Unlike these government-integrated approaches, Babyl operated as a parallel system with limited health management information system integration and donor-dependent financing. Participants'

XSL•FO
RenderX

emphasis on national health system integration and expanded insurance coverage reflects awareness of structural requirements for digital health sustainability.

The findings highlight critical service design considerations for future implementations. The exclusion of chronic disease management (diabetes, hypertension) and pediatric consultations created significant coverage gaps, with many lapsed users discontinuing after initial acute illness treatment when ongoing health needs fell outside service scope. Participants' recommendations for telephone sharing mechanisms and health post partnerships for remote areas reflect practical solutions to device access barriers. The registration process complexity, requiring multiple steps including phone verification, insurance validation, and profile setup, deterred potential users, particularly older adults and those with lower digital literacy. Successful registration required either independent technological capability or access to agent support; the absence of agents at many facilities meant the registration barrier was insurmountable for substantial population segments. These implementation details, often overlooked in digital health planning, proved critical to adoption and sustained utilization.

Implementation Lessons and Policy Implications

The findings yield several implementation lessons for future digital health initiatives in similar contexts. First, sustainable financing mechanisms integrated within health system structures are essential. Babyl's closure despite technical success and user satisfaction demonstrates that donor-dependent models face inherent sustainability challenges. Second, comprehensive health system integration from implementation outset is critical; parallel systems create workflow disruptions, continuity of care gaps, and provider resistance that undermine effectiveness regardless of platform technical capabilities. Third, differentiated implementation approaches are required for diverse population segments; one-size-fits-all strategies fail to address the distinct needs of older versus younger, rural versus urban, and high versus low digital literacy users. Fourth, adequate human resource allocation for user support is nonnegotiable. The absence of agents at HCs created insurmountable barriers for populations most in need of support. Finally, explicit governance frameworks clarifying digital health service accountability within existing health management structures are essential to prevent implementation confusion and ensure sustainability [2,3,12].

Strengths and Limitations

This study has several strengths. The purposive sampling strategy across 4 distinct user categories (active users, lapsed users, registered nonusers, and eligible nonregistrants) enabled examination of the full adoption spectrum rather than only successful users, revealing critical barriers typically missed in implementation evaluations. Multi-stakeholder inclusion of patients, health care providers, HC heads, and Babyl agents provided triangulated perspectives on implementation facilitators and challenges from different vantage points within the health system. Geographic diversity across 6 districts with both urban and rural representation ensured findings captured varied contexts. Data collection timing during active implementation enabled exploration of real-time experiences rather than retrospective recall. The use of established qualitative methods with rigorous analysis procedures, including independent coding, regular team discussions, member checking, and audit trails, enhanced analytical validity [18].

Several limitations should be considered when interpreting findings. (1) The study relied on self-reported experiences and perceptions, which may be subject to recall bias and social desirability response patterns, particularly when discussing reasons for non-adoption or discontinuation. (2) Data collection occurred during Babyl's active implementation period before closure was announced; participants' perceptions may have differed had they known about impending service discontinuation. (3) Eligible nonregistrants who were unaware of Babyl services were not included, meaning perspectives from the least-reached population segments are absent. (4) The study did not directly assess clinical outcomes, cost-effectiveness, or health impact, limiting conclusions to user experiences and implementation processes rather than clinical or economic outcomes. (5) Although data saturation was documented through iterative data collection and analysis, the possibility remains that additional perspectives may exist within the broader population. (6) The specific context of Rwanda's health system, insurance coverage structures, and telecommunications infrastructure may limit direct generalizability to other LMICs, though conceptual transferability to similar contexts is likely. (7) Language translation from Kinyarwanda to English may have introduced subtle meaning shifts despite back-translation validation procedures. (8) Selection bias may have occurred despite purposive sampling efforts, as participants willing to discuss digital health services may differ from those who declined participation. (9) The cross-sectional design captured experiences at a single time point; longitudinal follow-up would have enabled examination of evolving perceptions and utilization patterns over time.

Conclusions

This qualitative study explored user experiences and implementation challenges of Babyl's telemedicine platform in Rwanda, which operated from 2016 until its closure in September 2023. Our findings reveal that while participants recognized specific benefits of digital consultation, particularly convenience, reduced travel costs, and privacy, multiple interconnected barriers at individual, community, health system, and policy levels impeded widespread adoption and sustained utilization. Through analysis of experiences from active users, lapsed users, registered nonusers, and eligible nonregistrants, alongside perspectives from health care providers, HC heads, and Babyl agents, the study identified 5 major implementation themes that collectively explain the platform's challenges.

Key insights from participants emphasize that successful digital health implementation extends beyond technological functionality to encompass critical human and structural elements. First, participants across all categories emphasized that digital literacy support and human assistance are nonnegotiable requirements, not optional enhancements. All participants (100%) older than 50 years required agent assistance, yet agent absence at HCs was identified as the single most critical implementation gap. Second, health care providers

XSL•FO
RenderX

consistently reported that the platform operated as a parallel system rather than an integrated component of health service delivery, creating workflow disruptions, continuity of care gaps, and medication stock-out frustrations that undermined both provider satisfaction and patient confidence. Third, participants highlighted significant equity implications: older adults, individuals with lower educational attainment, rural residents, and those lacking smartphones faced compounded barriers that the service design failed to adequately address.

These findings carry practical implications for Rwanda's current digital health redesign process and for similar initiatives in comparable contexts. Participants' recommendations, validated across stakeholder groups, converge on five critical requirements: (1) genuine integration with existing health information systems and clinical workflows rather than parallel operation; (2) sustainable financing models embedded within national health system structures rather than donor-dependent arrangements; (3) universal availability of trained agents providing enrollment and troubleshooting support; (4) differentiated implementation strategies that accommodate varying levels of digital literacy, device access, and infrastructure; and (5) explicit governance frameworks clarifying roles, responsibilities, and accountability within health management structures. While this study focused specifically on Babyl's experience in Rwanda, these implementation lessons may inform digital health initiatives in settings with similar health system structures, resource constraints, and digital infrastructure development stages.

Acknowledgments

We acknowledge the contributions of all study participants who shared their experiences and insights. We thank the Rwanda Ministry of Health, University of Rwanda School of Public Health, and CIIC-HIN for their support in conducting this research. We also acknowledge the Bill and Melinda Gates Foundation for funding the broader evaluation project of which this qualitative study was a component. All authors declared that they had insufficient or no funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article. No artificial intelligence tools or large language models were used in the design, conduct, analysis, interpretation, or writing of this study. All data collection, thematic analysis, manuscript writing, and editing were performed exclusively by the research team members listed as authors.

Funding

This research was supported by the Centre for Impact, Innovation and Capacity Building for Health Information Systems and Nutrition (CIIC-HIN) with funding from the Bill & Melinda Gates Foundation (Grant Number: INV-030342). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The views expressed are those of the authors and do not necessarily reflect the official policy or position of the funding organizations.

Data Availability

The deidentified data that support the findings of this study are included in the manuscript in the form of anonymized quotes and thematic analyses. Full interview transcripts cannot be shared publicly due to privacy concerns and ethical considerations regarding participant confidentiality. However, the interview guide and coding framework are available from the corresponding author upon reasonable request, subject to approval by the ethical review board.

Authors' Contributions

The original draft was written by SMF, JH, JC, and GA. The methodology and formal analysis were developed and performed by SMF, JH, JC, and GA. The manuscript underwent review and editing by MSF, GA, UN, USM, FRK, YDN, PS, RH, JC, and JH. All authors contributed to the conceptualization and design of the study. Data collection was supervised by JH and JC. Critical revisions and intellectual content were provided by all co-authors. All authors reviewed and approved the final manuscript for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Focus Group Discussion Guide for Babyl registered clients who never used digital services and nonregistered eligible CBHI (Community-Based Health Insurance) members who are aware of Babyl services.
[DOCX File, 22 KB - jmir_v28i1e84832_app1.docx]

Multimedia Appendix 2
Key Informant interview guide for health care providers.
[DOCX File, 24 KB - jmir_v28i1e84832_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Key Informant interview guide for Babyl agents.
[DOCX File, 68 KB - jmir_v28i1e84832_app3.docx]

Multimedia Appendix 4
Key Informant interview guide with health center staff.
[DOCX File, 23 KB - jmir_v28i1e84832_app4.docx]

References

1. Ahmed MM, Okesanya OJ, Olaleke NO, et al. Integrating digital health innovations to achieve universal health coverage: promoting health outcomes and quality through global public health equity. Healthcare (Basel) 2025 May 5;13(9):1060. [doi: 10.3390/healthcare13091060] [Medline: 40361838]
2. Fifty-eighth world health assembly. : World Health Organization (WHO); 2005 URL: https://apps.who.int/gb/ebwha/pdf_files/wha58-rec1/english/a58_2005_rec1-en.pdf [accessed 2026-03-30]
3. Al-Shorbaji N. The World Health Assembly resolutions on eHealth: eHealth in support of universal health coverage. Methods Inf Med 2013;52(6):463-466. [doi: 10.1055/s-0038-1627062] [Medline: 24310395]
4. Hollander JE, Carr BG. Virtually perfect? Telemedicine for Covid-19. N Engl J Med 2020 Apr 30;382(18):1679-1681. [doi: 10.1056/NEJMp2003539] [Medline: 32160451]
5. Ibrahim MS, Mohamed Yusoff H, Abu Bakar YI, Thwe Aung MM, Abas MI, Ramli RA. Digital health for quality healthcare: a systematic mapping of review studies. Digit Health 2022;8:20552076221085810. [doi: 10.1177/20552076221085810] [Medline: 35340904]
6. Primary health care systems (primasys). Nigeria Governors' Forum. 2017. URL: https://ngfrepository.org.ng:8443/jspui/bitstream/123456789/3163/1/PRIMARY HEALTH CARE SYSTEMS-CASE STUDY FROM NIGERIA.pdf [accessed 2025-08-15]
7. Telehealth innovation in Rwanda: lessons and insights from the Babyl experience. Percept. 2021. URL: https://percept.co.za/wp-content/uploads/2021/05/Brief-3-babyl.pdf [accessed 2025-08-28]
8. Mitgang EA, Blaya JA, Chopra M. Digital health in response to COVID-19 in low- and middle-income countries: opportunities and challenges. Glob Policy 2021 Jul;12(Suppl 6):107-109. [doi: 10.1111/1758-5899.12880] [Medline: 34230840]
9. Nyakabau MR. Digital-first integrated care: Rwanda's innovative digital health care service. : Transform Health Coalition; 2021 Nov URL: https://transformhealthcoalition.org/wp-content/uploads/2021/11/Rwanda.pdf [accessed 2026-03-18]
10. Totten AM, McDonagh MS, Wagner JH. The Evidence Base for Telehealth: Reassurance in the Face of Rapid Expansion during the COVID-19 Pandemic: Rockville (MD): Agency for Healthcare Research and Quality (US); 2020. [Medline: 32479040]
11. Jiang X, Ming WK, You JHS. The cost-effectiveness of digital health interventions on the management of cardiovascular diseases: systematic review. J Med Internet Res 2019 Jun 17;21(6):e13166. [doi: 10.2196/13166] [Medline: 31210136]
12. Digital-in-health: unlocking the value for everyone digit unlocking value everyone. Washington (DC): World Bank. 2023. URL: https://www.worldbank.org/en/topic/health/publication/digital-in-health-unlocking-the-value-for-everyone [accessed 2026-03-18]
13. Ashwood JS, Mehrotra A, Cowling D, Uscher-Pines L. Direct-To-consumer telehealth may increase access to care but does not decrease spending. Health Aff (Millwood) 2017 Mar;36(3):485-491. [doi: 10.1377/hlthaff.2016.1130]
14. Singh H. Diagnostic errors: moving beyond "no respect" and getting ready for prime time. BMJ Qual Saf 2013 Oct;22(10):789-792. [doi: 10.1136/bmjqs-2013-002387] [Medline: 24048615]
15. Madhivanan A, Venugopal V, Dongre AR. Physical violence against doctors: a content analysis from online Indian newspapers. Indian J Community Med 2020;45(1):108-109. [doi: 10.4103/ijcm.IJCM_215_19] [Medline: 32029997]
16. Cory N, Stevens P. Building a global framework for digital health services in the era of COVID-19. Inf Technol Innov Found 2020 [FREE Full text]
17. Labrique A, Agarwal S, Tamrat T, Mehl G. WHO digital health guidelines: a milestone for global health. NPJ Digit Med 2020;3(1):120. [doi: 10.1038/s41746-020-00330-2] [Medline: 33015373]
18. Eslami Jahromi M, Ayatollahi H. Utilization of telehealth to manage the Covid-19 pandemic in low- and middle-income countries: a scoping review. J Am Med Inform Assoc 2023 Mar 16;30(4):738-751. [doi: 10.1093/jamia/ocac250]
19. World Health Organization. Monitoring and Evaluating Digital Health Interventions A Practical Guide to Conducting Research and Assessment: World Health Organization; 2016:1-144 URL: https://www.who.int/publications/i/item/9789241511766 [accessed 2026-03-27]
20. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implementation Sci 2009 Dec;4(1):50. [doi: 10.1186/1748-5908-4-50]

XSL•FO
RenderX

21. 4X4 reform: executive summary towards quadrupling the current workforce in 4 years (4X4); uplifting quantity and quality of care. Ministry of Health. 2024. URL: https://www.moh.gov.rw/index.php?eID=dumpFile&t=f&f=92859&token=3c26da47182b639efd749f9c6829633dc0fcada4 [accessed 2025-10-25]
22. Health sector strategic plan V. MOH-Rwanda. URL: https://www.moh.gov.rw/index.php?eID=dumpFile&t=f&f=117507&token=f2a527e089201f73bec29f8ce6d6c4dc55ed241e [accessed 2025-11-05]
23. Global Strategy on Digital Health 2020-2025: World Health Organization; 2021. URL: https://iris.who.int/ [accessed 2026-03-18]
24. McKinsey & Company. Unlocking digital healthcare in lower- and middle-income countries. McKinsey. 2021 Sep 22. URL: https://www.mckinsey.com/industries/healthcare/our-insights/unlocking-digital-healthcare-in-lower-and-middle-income-countries [accessed 2026-11-18]
25. Bhushan I, Anandampillai K, Agarwal S. Navigating complexities: agile digital health initiatives in developing countries. Oxf Open Digit Health 2024;2:oqae032. [doi: 10.1093/oodh/oqae032] [Medline: 40230973]

Abbreviations

CBHI: Community-Based Health Insurance
FGD: focus group discussion
HC: health center
KII: key informant interview
LMIC: low- and middle-income country
OPD: outpatient department
RAMA: Rwandaise d'Assurance Maladie
WHO: World Health Organization

Edited by A Stone; submitted 25.Sep.2025; peer-reviewed by KL Mauco, R Ranjan; revised version received 10.Jan.2026; accepted 28.Jan.2026; published 01.Apr.2026.

Please cite as:
Musange Furere S, Absolomon G, Umutoni N, Matutina Umuhoza S, K Rubuga F, Nsaba-Uwera YD, Singa Muhoza P, Hitimana R, Condo J, Humuza J
Digital Primary Health in Rwanda: Qualitative Study of User Experiences and Implementation Lessons From Babyl's Telemedicine Platform
J Med Internet Res 2026;28:e84832
URL: https://www.jmir.org/2026/1/e84832
doi:10.2196/84832

© Sabine Musange Furere, Gashaija Absolomon, Nathalie Umutoni, Stella Matutina Umuhoza, Felix K Rubuga, Yvonne Delphine Nsaba-Uwera, Patrick Singa Muhoza, Regis Hitimana, Jeanine Condo, James Humuza. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 1.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Telehealth Use and Modality Choice Among US Adults: Shorrocks-Shapley Decomposition of a 2022 Cross-Sectional National Survey

Corneliu Bolbocean[1], BA, MA, MPPM, PhD; Corey Hayes[2], MPH, PhD; Cari Bogulski[3], PhD

[1]Nuffield Department of Primary Care Health Sciences, University of Oxford, Radcliffe Primary Care Building, Radcliffe Observatory Quarter, Woodstock Rd, Oxford, United Kingdom

[2]Department of Pharmacy Practice, University of Arkansas for Medical Sciences, Little Rock, AR, United States

[3]Department of Biomedical Informatics in the College of Medicine, UAMS, Little Rock, AR, United States

Corresponding Author:
Corneliu Bolbocean, BA, MA, MPPM, PhD
Nuffield Department of Primary Care Health Sciences, University of Oxford, Radcliffe Primary Care Building, Radcliffe Observatory Quarter, Woodstock Rd, Oxford, United Kingdom

Abstract

Background: Telehealth use surged during the COVID-19 pandemic and has stabilized at levels substantially above prepandemic baselines. However, concerns persist that the digital divide may reproduce or widen disparities in access. Understanding the determinants of telehealth use—and particularly modality choice between video and audio—is essential for designing policies that promote equitable access in the post–public health emergency era.

Objective: This study aims to identify determinants of telehealth use and modality among US adults in 2022 and quantify the relative contributions of digital, geographic, clinical, and socioeconomic domains.

Methods: We conducted a cross-sectional secondary analysis of the sixth cycle of the Health Information National Trends Survey, administered in 2022 by the National Cancer Institute, a nationally representative, 2-stage stratified random probability survey of civilian, noninstitutionalized US adults aged 18 years or older. Sampled households were recruited via mailed invitations, and 1 adult per household was randomly selected using the next birthday method and invited to complete a self-administered questionnaire between February 2022 and November 2022 (N=6252). The primary analytic sample included respondents with nonmissing telehealth modality responses (n=6046, 59.4% female; mean age of 55.1 y). Individual-level data were linked to county-level American Community Survey socioeconomic indicators and broadband availability measures. The primary outcome was telehealth use, categorized as video (n=1641, 27.2%; 95% CI 25.5%-29.1%), audio-only (n=876, 12.1%; 95% CI 10.9%-13.4%), or none (n=3529, 60.7%; 95% CI 58.6%-62.7%). We estimated 4 binary contrasts using survey-weighted linear probability models with jackknife variance estimation, reporting absolute risk differences in percentage points (pp) with 95% CIs. We applied Shorrocks-Shapley decomposition to quantify each predictor domain's contribution to explained variance.

Results: Nationally, 39.3% (n=2517; 95% CI 37.3% - 41.4%) reported any telehealth use in the past 12 months. In survey-weighted linear probability models (α=.05), significant predictors of any telehealth vs none included: male sex (−9.7 pp, 95% CI −14.0 to −5.4), disability status (+22.5 pp, 95% CI 16.1-28.8), and health app use (+18.4 pp, 95% CI 12.0-24.8). For video vs audio-only telehealth, insurance coverage increased video use (+21.2 pp, 95% CI 13.0-29.3), while basic cell phone only (vs smartphone) decreased video use (−20.1 pp, 95% CI −33.5 to −6.8). Shorrocks-Shapley decomposition revealed that digital access and eHealth behaviors explained 40.4% of variance in video vs audio choice and 33.4% of video vs none; geography explained 40.5% of audio vs none; digital factors (25.7%), geography (19.7%), and health status and needs (15.5%) all contributed substantially to any vs none.

Conclusions: Digital access and eHealth behaviors collectively explain more variance in modality choice than traditional sociodemographic factors. Telehealth uptake reflects a combination of digital factors, geography, and clinical need, whereas video modality specifically hinges on digital readiness. Interventions pairing sustained insurance coverage with targeted investments in device access, affordable high-speed connectivity, and digital literacy training are most likely to narrow persistent telehealth gaps.

(J Med Internet Res 2026;28:e81879) doi:10.2196/81879

KEYWORDS

telehealth; digital divide; eHealth literacy; Shorrocks-Shapley decomposition; Health Information National Trends Survey; HINTS

XSL•FO
RenderX

Introduction

The landscape of health care delivery in the United States has been reshaped by the expansion of telehealth—the use of electronic information and telecommunications technologies to support clinical care, education, and administration [1-3]. Services delivered by telephone, video, remote monitoring, or asynchronous platforms can lower geographic and logistical barriers to care [3]. However, as telehealth has scaled, concerns persist that the "digital divide" may reproduce or even widen disparities in access and outcomes [4-9].

Evidence documents persistent differences in telehealth access across sociodemographic groups. Older adults, racial and ethnic minorities, individuals with lower educational attainment, and those residing in rural areas consistently demonstrate lower telehealth use rates [10]. A systematic review found that telehealth adoption during the pandemic was significantly lower among Black and Hispanic patients compared to White patients after accounting for clinical and geographic factors [11]. Recent analyses of post-pandemic telehealth patterns reveal that while overall use has stabilized, disparities by age, income, and digital literacy persist [12]. Critically, modality choice—whether patients use video or audio-only visits—varies substantially by socioeconomic status (SES) and digital access, with audio-only visits more common among older, lower-income, and rural populations [13].

Use of telehealth surged during the initial phase of the COVID-19 public health emergency (PHE) and then declined from its peak, but it has stabilized at levels substantially higher than before the pandemic. For example, among Medicare fee-for-service beneficiaries, telehealth use in 2021 was approximately 38-fold higher than prepandemic levels (from 0.1% in 2019 to 3.8% of visits in 2021) [14]. By late 2022, about 15% of Medicare beneficiaries received telehealth in a given quarter, and in the last quarter of 2023, more than 1 in 10 beneficiaries (ie, 13%) used telehealth at least once, well above the pre-2020 baseline [15,16]. Also, in population-based surveys, 37% of US adults reported telemedicine use in 2021 and 43% of adults with a health care visit reported telemedicine use in 2022 (70% video; 30% audio) [17,18]. Taken together, these data indicate that telehealth has become a substantial component of routine care, with meaningful variation by modality (video vs audio).

The PHE also saw major federal investments in household internet affordability that intersect directly with telehealth access. The Affordable Connectivity Program (ACP) launched at the end of 2021 expanded subsidized broadband, which increased adoption among lower-income households and provided a discount for eligible households and was active throughout 2022 [19,20]. By the end of 2022, roughly 15.4 million households had enrolled (around 28% of eligible households), growing to more than 23 million by the February 2024 enrollment freeze [21,22]. The ACP ended in June 2024 when Congress did not appropriate additional funding, potentially affecting an estimated 23 million households [23]. Evidence after the ACP wind-down suggests that broadband affordability shocks can translate into foregone telehealth among subsidy-eligible households: in a post-ACP consumer survey of over 68,000 lifeline and/or former ACP participants fielded in November 2024, 36.1% of respondents who were previously enrolled in ACP reported being unable to continue with telehealth, rehabilitation, or remote medical monitoring without ACP [24]. These policy changes have altered the digital access landscape in ways that may affect who can use—and how they use—telehealth.

The digital divide is multidimensional, encompassing access to devices and connectivity, the quality and affordability of home internet, and the skills and engagement required to use digital tools effectively [25]. These dimensions map directly onto telehealth modality: video visits typically require higher-bandwidth connectivity, compatible devices, and greater comfort with app-based workflows, whereas audio-only visits may remain feasible when bandwidth, device capability, or digital skills are constrained [26,27]. However, much of the existing literature emphasizes individual predictors without quantifying the relative explanatory importance of digital, geographic, clinical, and socioeconomic domains. A domain-level approach can help distinguish which levers are most policy-relevant for improving overall uptake vs enabling equal access to video-based care.

Despite growing research on telehealth disparities, significant gaps remain in understanding the relative importance of different determinant domains. Most existing studies examine individual predictors in isolation rather than quantifying how much each domain—demographics, SES, health needs, digital access, and geography—contributes to explaining telehealth use patterns. Furthermore, few studies have simultaneously examined predictors of both any telehealth use and modality choice (video vs audio-only) in the post-PHE period, when several temporary telehealth flexibilities required repeated legislative extensions and telehealth patterns stabilized into new norms.

We use nationally representative data from the sixth cycle of the Health Information National Trends Survey (HINTS 6), administered in 2022 by the National Cancer Institute, to describe contemporary patterns of telehealth use (phone or video) and to evaluate how sociodemographic, clinical, area-level and geographic measures, psychosocial, and digital access factors relate to (1) any telehealth use and (2) modality choice. Using the Shorrocks-Shapley decomposition models, we quantify the relative contribution of each explanatory domain to explained variance in telehealth outcomes [28].

Our objectives were to (1) estimate national prevalence of telehealth use by modality (video, audio-only, and none) among US adults in 2022; (2) identify individual and contextual determinants of telehealth use and modality choice using survey-weighted regression models; and (3) quantify the relative contribution of digital access or eHealth behaviors, geography, clinical need, and socioeconomic factors to explained variance using the Shorrocks-Shapley decomposition.

XSL•FO
RenderX

Methods

Research Design

This study used a cross-sectional observational design using secondary analysis of nationally representative survey data linked to area-level contextual measures. The primary aim was to identify individual and contextual determinants of telehealth use and modality choice among US adults in 2022. We report this cross-sectional observational study in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for cross-sectional studies [29].

Participants

Inclusion and Exclusion Criteria

The study population included civilian, noninstitutionalized adults aged 18 years and older residing in the United States. Eligibility was determined by the HINTS sampling frame, which excluded individuals living in institutionalized settings (eg, prisons and nursing homes), those without a valid mailing address, and residents of US territories. For this analysis, we included all respondents with valid responses to the telehealth use question.

Sampling Procedures

We analyzed the HINTS 6, a nationally representative survey of civilian US adults administered by the National Cancer Institute from February 2022 through November 2022 [30,31]. HINTS uses a complex 2-stage stratified random probability sampling design: first, a stratified sample of addresses was selected from a comprehensive address file, then 1 adult per household was randomly selected to complete the survey. The survey provides person-level sampling weights and 50 jackknife replicate weights to allow reliable national inference. The restricted-use version of HINTS (available to researchers through a data use agreement with the National Cancer Institute) contains geocoding variables (county Federal Information Processing Series [FIPS] codes) not included in the public-use file, enabling linkage with area-level datasets.

Sample Size

The HINTS 6 cycle included 6252 respondents. The primary analytic sample for telehealth modality outcomes included respondents with nonmissing telehealth modality responses (n=6046). This sample size provides adequate statistical power to detect small-to-medium effect sizes (Cohen $d \geq 0.10$) with 80% power at $\alpha.05$ for the primary regression analyses, accounting for the design effect of the complex survey design.

Data Linkage

Consistent with prior literature, we appended area-level measures to each respondent via county FIPS codes. County socioeconomic context for 2022 (total population, poverty rate, unemployment rate, percent without a high school diploma, and percent aged ≥65 y) was derived from the American Community Survey [32]. County-level broadband availability was obtained from the National Institute on Minority Health and Health Disparities (NIMHD) health disparities (HD) Pulse portal as the percent of residents with broadband access [33]. We additionally obtained the Centers for Disease Control and Prevention/Agency for Toxic Substances and Disease Registry (CDC/ATSDR) Social Vulnerability Index (SVI) to characterize local vulnerability [34].

Ethical Considerations

This study analyzed the 2022 HINTS 6 restricted-use deidentified dataset. The University of Arkansas for Medical Sciences Institutional Review Board determined this study did not constitute human participants research (protocol 297801). The original HINTS data collection was approved by the Westat Institutional Review Board, and all survey respondents provided implied informed consent by completing and returning the questionnaire. The restricted-use dataset contained no direct identifiers; geographic information was limited to county-level codes. Respondents received no compensation for the secondary analysis; compensation for the original HINTS survey included a US $2 preincentive and an opportunity for a US $25 postsurvey incentive. No images containing identifiable individuals were included in this study.

Measures

Outcome Variable

Telehealth exposure in the 12 months preceding the survey was categorized into 3 mutually exclusive groups—video, audio-only, and no telehealth—yielding 1641 (27.2%, 95% CI 25.5% - 29.1%), 876 (12.1%, 95% CI 10.9% - 13.4%), and 3529 (60.7%, 95% CI 58.6% - 62.7%) respondents, respectively (Table 1). Respondents reporting any video visit were assigned to the video group (even if they also had phone visits); audio-only indicates telephone visits with no video exclusively. We used “telehealth” to denote synchronous video or phone visits. Other digital health activities (apps, portals, and information seeking) were considered model predictors, not modalities of telehealth delivery.

XSL•FO
RenderX

Table . Descriptive characteristics by telehealth modality (video, audio only, and none) among US adults[a].

Characteristics	Video	Audio only	No telehealth
Participants, n (%)	1641 (27.2)	876 (12.1)	3529 (60.7)
Age (y), mean (SD)	52.84 (16.46)	58.57 (17.55)	56.26 (17.58)
Female, n (%)	1071 (65.3)	566 (64.6)	2012 (57)
High school graduation, n (%)	1499 (96.7)	762 (92.1)	3120 (92.5)
Race and ethnicity, n (%)			
Asian	51 (3.1)	29 (3.3)	116 (3.3)
Black	241 (14.7)	135 (15.4)	551 (15.6)
Hispanic	262 (16)	177 (20.2)	542 (15.4)
White	1155 (76.5)	566 (71.7)	2353 (72.6)
Other	169 (10.3)	119 (13.6)	422 (12)
Metropolitan status, n (%)			
Large central metro	684 (41.7)	355 (40.5)	1220 (34.6)
Large fringe metro (suburban)	393 (23.9)	176 (20.1)	769 (21.8)
Medium or small metro	401 (24.4)	250 (28.5)	1016 (28.8)
Micropolitan	96 (5.9)	55 (6.3)	307 (8.7)
Noncore (rural)	66 (4)	40 (4.6)	216 (6.1)
Income group (US $, n (%)			
Low ($20,000)	213 (14.4)	155 (20)	555 (17.5)
Middle (20,000-75,000)	557 (37.6)	351 (45.2)	1457 (45.9)
High (≥75,000)	788 (48)	270 (34.8)	1292 (36.6)
Employed, n (%)	912 (55.6)	349 (41.9)	1807 (51.2)
Cancer history, n (%)	258 (15.7)	134 (15.3)	487 (13.8)
Hypertension, n (%)	731 (44.5)	452 (51.6)	1477 (41.9)
Disabled, n (%)	172 (11)	96 (11.5)	211 (6.3)
PROMIS[b] Isolation T-score, mean (SD)	47.11 (9.97)	46.36 (9.46)	45.40 (9.30)
PROMIS Meaning T-score, mean (SD)	54.16 (10.29)	55.10 (9.75)	55.61 (10.08)
Health Insurance, n (%)	1559 (95.1)	813 (92.8)	3165 (89.7)
Health numeracy, n (%)			
Easy or very easy	1241 (75.6)	624 (71.2)	2495 (70.7)
Hard or very hard	369 (22.5)	231 (26.4)	951 (26.9)
Use of health or wellness apps, n (%)			
Used app	1103 (67.2)	414 (47.3)	1441 (40.8)
Did not use or no apps or no smart device	498 (30.3)	385 (43.9)	1871 (53)
Sought health information online, n (%)			
No	128 (7.8)	95 (10.8)	506 (14.3)
Yes	1367 (83.3)	586 (66.9)	2294 (65)
Not internet user	120 (7.3)	153 (17.5)	586 (16.6)
Satisfaction with home internet for health needs, n (%)			
Satisfied (extremely or very)	1075 (65.5)	439 (50.1)	1789 (50.7)
Somewhat	354 (21.6)	194 (22.1)	799 (22.6)

XSL•FO
RenderX

Characteristics	Video	Audio only	No telehealth
Not satisfied or nonuser	185 (11.3)	207 (23.6)	802 (22.7)
Device ownership, n (%)			
Any smart device (tablet or smartphone or multiple)	1582 (96.4)	750 (85.6)	3091 (87.6)
No smart device (basic cell only or none)	54 (3.3)	120 (13.7)	417 (11.8)
Census division, n (%)			
New England	80 (4.9)	39 (4.5)	108 (3.1)
Middle Atlantic	192 (11.7)	83 (9.5)	399 (11.3)
East North Central	170 (10.4)	98 (11.2)	481 (13.6)
West North Central	51 (3.1)	17 (1.9)	210 (6)
South Atlantic	444 (27.1)	204 (23.3)	878 (24.9)
East South Central	51 (3.1)	39 (4.5)	202 (5.7)
West South Central	232 (14.1)	117 (13.4)	567 (16.1)
Mountain	126 (7.7)	58 (6.6)	255 (7.2)
Pacific	295 (18)	221 (25.2)	429 (12.2)

[a]n is unweighted and percentage is survey-weighted.

[b]PROMIS: Patient-Reported Outcomes Measurement Information System.

Explanatory Variables

A comprehensive set of explanatory variables to explain telehealth use was constructed based on prior literature on the digital divide and health care access [7,30,31,35]. Variable definitions and coding for individual-level demographics, SES, health status, digital access and literacy (device ownership, satisfaction with home internet for health needs, health numeracy, health or wellness app use, and online health information seeking), and geographic indicators (US Census division and urbanicity) are summarized in Appendix A in Multimedia Appendix 1.

Digital Readiness Variables

We modeled digital readiness using separate indicators for device ownership, internet satisfaction, health app use, and online health information seeking rather than combining them into a single composite score via factor analysis. This approach was chosen for 3 reasons. First, these constructs are theoretically distinct: device ownership reflects material access, internet satisfaction captures infrastructure quality, and app use or information seeking represents behavioral engagement and digital skills [36]. Second, separate coefficients provide policy information—interventions targeting device distribution differ from those promoting digital literacy or improving broadband infrastructure [37]. Finally, the Shorrocks-Shapley decomposition explicitly partitions explained variance across covariate domains, allowing us to quantify the relative importance of “digital access” vs “health literacy and eHealth” without collapsing distinct mechanisms into a single latent factor.

Psychosocial Variables

HINTS 6 includes validated Patient-Reported Outcomes Measurement Information System (PROMIS) short forms measuring psychosocial constructs [38,39]. The PROMIS Social Isolation 4-item short form assesses perceptions of being excluded, disconnected, or isolated from others; scores are reported as T-scores (population: mean 50, SD 10), with higher scores indicating greater isolation. The PROMIS Meaning and Purpose 4-item short form assesses sense of meaning, purpose, and engagement in life, with higher T-scores indicating greater sense of purpose.

PROMIS measures were included as exploratory covariates to assess whether psychosocial context contributed meaningfully beyond digital, geographic, clinical, and socioeconomic domains. Social isolation has been associated with reduced health care use, lower technology adoption, and barriers to accessing care [40,41]. Conversely, individuals with a greater sense of meaning and purpose may be more motivated to engage with health services and adopt health-promoting technologies [42]. These constructs may therefore influence both the decision to use telehealth and the choice of modality, independent of traditional sociodemographic and health-related factors.

Analytic Strategy

Primary Analyses

We report descriptive statistics and estimate multivariable associations using survey-weighted linear probability models (LPMs) for four contrasts: (1) video vs audio-only, (2) video vs no telehealth, (3) audio-only vs no telehealth, and (4) any telehealth (video or audio) vs no telehealth. The LPM directly models the probability of the outcome as a linear function of covariates and with survey weights provides easily interpretable absolute risk differences in percentage points [43,44].

XSL•FO
RenderX

Coefficients are interpreted as the percentage-point (pp) change in the probability of the outcome for a 1-unit change in the predictor, holding other variables constant.

Survey Design Adjustments

All adjusted analyses incorporated the complex survey design of HINTS 6. We used the final person-level weight and the 50 jackknife replicate weights to obtain design-consistent point estimates and robust standard errors, following HINTS technical guidance [45]. This approach yields nationally representative estimates with SEs and *P* values adjusted for stratification and clustering. Replicate-weight variance estimation used the jackknife mean squared error (MSE).

Decomposition Analysis

To quantify the explanatory contribution of prespecified covariate domains, we applied a Shorrocks-Shapley decomposition of the model R^2, partitioning the total R^2 into additive contributions from groups of predictors (demographics, SES, health status and needs, health system access, geography, urbanicity, psychosocial factors, digital access, health literacy and eHealth, and area context) [28,46]. Decompositions were implemented for each survey-weighted LPM using the *shapley2* module in Stata (StataCorp LLC) [47]. The Shorrocks-Shapley approach does not imply causal ordering or construct hierarchy; it quantifies predictive importance.

Missing Data Analysis

We evaluated item-level missingness across all study variables prior to model estimation. Missingness was assessed for continuous measures (including PROMIS T-scores and self-rated health) and for sociodemographic covariates (including employment status, disability status, race or ethnicity, and household income). To minimize loss of observations in adjusted models and to preserve the survey sample structure, selected categorical covariates with modest item nonresponse and/or structurally nonapplicable responses (indicators related to digital access and online behaviors) were coded with an explicit “missing/unknown” category. These missingness categories were included as indicator levels in the regression models to avoid listwise deletion for those variables, and coefficients for missingness indicators were not interpreted substantively.

To evaluate the plausibility of common missing-data mechanisms, we first applied the Little Missing Completely at Random (MCAR) test [48]. In this way, we assessed whether observed values differed systematically across missing data. Second, because income exhibited the largest item nonresponse among key sociodemographic measures, we performed a supplementary missingness model in which an indicator for income nonresponse was regressed with minimal missingness (age, census division, cancer history, and insurance status). These diagnostics were used to guide the selection of an appropriate missing-data handling strategy for sensitivity analyses.

As a sensitivity analysis, we conducted multiple imputation by chained equations under a missing at random (MAR) framework conditional on observed data. We generated *m*=20 imputed datasets. Variables imputed included household income, race or ethnicity, employment status, disability status, self-rated health, sex, marital status, education, and PROMIS psychosocial measures. The imputation models incorporated all variables used in the primary analyses, along with auxiliary predictors expected to be associated with missingness and/or the imputed values (including age, census division, urbanicity, cancer history, high blood pressure, and insurance status). After imputation, we re-estimated the primary survey-weighted regression models in each imputed dataset and pooled point estimates and SEs across imputations using standard combining rules. We then compared pooled estimates from the imputed analyses with the complete-case results to evaluate robustness; full imputation model specifications, diagnostics, and pooled regression outputs are provided in Multimedia Appendix 1.

Software: Statistical Threshold

All analyses were conducted in Stata/MP 18 (StataCorp LLC) using survey procedures (svy: prefix) with the provided jackknife replicate weights and MSE option, consistent with HINTS methodological guidelines. Tests were 2-sided with an α of .05.

Results

Sample Characteristics by Telehealth Modality

Table 1 summarizes characteristics by telehealth modality. Among 6252 respondents in HINTS 6, 1641 reported any video telehealth, 876 reported exclusively audio-only telehealth, and 3529 reported no telehealth in the past 12 months. These correspond to survey-weighted prevalence of 27.2% (95% CI 25.5% - 29.1%), 12.1% (95% CI 10.9% - 13.4%), and 60.7% (95% CI 58.6% - 62.7%), respectively. Video users were younger on average (52.8, SD 16.2 y) than those with no telehealth (56.3, SD 17.1 y) and audio-only users (58.6, SD 16.8 y). Women comprised a larger share of video (1071/1641, 65.3%) and audio-only (566/876, 64.6%) users than nonusers (2012/3529, 57%). Compared with nonusers, video users more frequently reported higher income (788/1641, 48% vs 1292/3529, 36.6% at ≥US $75,000), current employment (912/1641, 55.6% vs 1807/3529, 51.2%), and insurance coverage (1559/1641, 95.1% vs 3165/3529, 90%). Digital readiness was notably higher among video users: any smart device ownership (1582/1641, 96.4% vs 3091/3529, 87.6%), health or wellness app use (1103/1641, 67.2% vs 1441/3529, 40.8%), and online health information seeking (1367/1641, 83.3% vs 2294/3529, 65%). Satisfaction with home internet for health needs was also higher in the video group (1075/1641, 65.5% extremely or very satisfied vs 1789/3529, 50.7% among nonusers). Geographic composition differed across modalities, with video visits more common among residents of large central metros (684/1641, 41.7%) and large fringe metro or suburban areas (393/1641, 23.9%) compared with nonusers (1220/3529, 34.6% and 769/3529, 21.8%, respectively). White respondents made up a similar share across video, audio-only, and no telehealth groups, while Hispanic respondents were somewhat more represented among audio-only users than among video users or those who did not use telehealth.

XSL•FO
RenderX

Regression Estimates

Key Predictors

Table 2 reports key predictors from survey-weighted LPMs for all 4 contrasts, presenting marginal effects as pp differences. Below we summarize the key patterns; full coefficient estimates and CIs for all covariates appear in Tables S5-S8 in Multimedia Appendix 1.

Table . Key predictors from survey-weighted linear probability models of telehealth outcomes, Health Information National Trends Survey (HINTS 6)[a].

Variables	Video vs audio[b] (n=2069)	Video vs none[c] (n=4317)	Audio vs none[d] (n=3632)	Any vs none[e] (n=5009)
Demographics and socioeconomics				
Male	−0.2 (−5.5 to 5.2; *P*=.94)	−8.5 (−12.4 to −4.6; *P*<.001)	−5.5 (−8.1 to −2.8; *P*<.001)	−9.7 (−14.0 to −5.4; *P*<.001)
White	2.4 (−3.4 to 8.2; *P*=.42)	6.4 (2.3 to 10.6; *P=0.001*)	3.7 (0.7 to 6.7; *P*=.02)	7.1 (2.2 to 12.0; *P<0.001*)
Employed	9.1 (2.6 to 15.5; *P=0.001*	3.8 (−1.6 to 9.2; *P*=.17)	−3.3 (−6.8 to 0.2; *P*=.07)	0.9 (−3.3 to 5.1; *P*=.67)
Health status and access				
Disabled	9.6 (1.6 to 17.6; *P*=.02)	23.7 (17.8 to 29.6; *P*<.001)	11.9 (5.1 to 18.7; *P*<.001)	22.5 (16.1 to 28.8; *P*<.001)
High blood pressure	−1.7 (−7.8 to 4.3; *P*=.58)	5.8 (1.8 to 9.8; *P=0.001*)	5.8 (2.6 to 8.9; *P*<.001)	8.0 (4.8 to 11.2; *P*<.001)
Health insurance	21.2 (13.0 to 29.3; *P*<.001)	13.3 (9.2 to 17.3; *P*<.001)	1.1 (−4.0 to 6.3; *P*=.67)	11.5 (4.6 to 18.4; *P*=.001)
Digital access				
Basic cell only (reference: landline or none)	−20.1 (−33.5 to −6.8; *P=0.001*)	−7.4 (−16.4 to 1.7; *P*=.11)	2.6 (−3.9 to 9.1; *P*=.44)	−4.0 (−10.6 to 2.6; *P*=.23)
Internet: extremely satisfied	8.6 (1.3 to 15.9; *P*=.02)	7.7 (1.1 to 14.4; *P*=.02)	0.1 (−5.4 to 5.6; *P*=.97)	5.1 (−1.3 to 11.5; *P*=.12)
eHealth behaviors				
Used health apps	5.5 (−3.4 to 14.5; *P*=.23)	17.5 (12.5 to 22.6; *P*<.001)	9.1 (5.5 to 12.7; *P*<.001)	18.4 (12.0 to 24.8; *P*<.001)
Seeks health information online	6.6 (−5.5 to 18.7; *P*=.28)	7.2 (1.0 to 13.4; *P*=.02)	1.7 (−1.0 to 4.4; *P*=.22)	7.2 (2.4 to 12.0; *P=0.001*)

[a]Coefficients represent percentage-point differences in probability of outcome (95% CI; *P* value). Models additionally controlled for: age, marital status, income, education, self-rated health, cancer history, census division, urbanicity, Patient-Reported Outcomes Measurement Information System (PROMIS) isolation and meaning scores, other device categories, internet satisfaction levels, health numeracy, and county-level characteristics.

[b]R^2 for "video vs audio" is 0.08.

[c]R^2 for "video vs none" is 0.13.

[d]R^2 for "audio vs none" is 0.07.

[e]R^2 for "any vs none" is 0.12.

Demographics and Socioeconomic Factors

Gender differences were consistent: men were 8 to 10 pp less likely than women to use telehealth across all uptake contrasts (*P*<.001 for video and any telehealth vs none). White respondents had 4 to 7 pp higher use than non-White respondents, but race or ethnicity was not significantly associated with video vs audio modality choice. Adults aged 35 to 49 years showed elevated telehealth use (approximately 6 - 7 pp) compared with older age groups. Among telehealth users, employment and insurance coverage were strongly associated with choosing video over audio (9.1 pp and 21.2 pp, respectively).

Health Status and Needs

Health-related factors showed the largest effect sizes. Disability was strongly associated with telehealth use: 22 to 24 pp for video and any telehealth vs none, and 12 pp for audio-only vs none. Among telehealth users, disability was also associated with choosing video over audio (9.6 pp). Cancer history (6 - 7 pp) and hypertension (6 - 8 pp) were positively associated with all forms of telehealth uptake, but were not significant predictors of choosing video vs audio. Insurance coverage was consistently positive (11 - 13 pp for uptake; 21 pp for video vs audio choice).

Digital readiness

Digital factors differentiated video from audio use. Basic cell phone only (vs smartphone or tablet) was associated with 20 pp lower video use among telehealth users. Internet satisfaction strongly favored video: "extremely" or "very satisfied" users were 8 to 9 pp more likely to choose video over audio. Health app use showed the largest digital association with uptake: 17 to 18 pp for video and any telehealth vs none. Online health information seeking was positively associated with video (7 pp) but not audio-only uptake.

Geographic Variation

Geography was especially salient for audio-only use. Relative to New England, audio-only use was 9 to 19 pp lower across multiple census divisions (East North Central, West North

XSL•FO
RenderX

Central, South Atlantic, and West South Central). For any telehealth vs none, the West North Central showed the largest deficit (−21 pp). Geographic variation for video use was more modest, with only West North Central showing a significant deficit (−13 pp vs New England).

Psychosocial Factors

PROMIS measures showed small associations. Social isolation was weakly associated with any telehealth use (0.2 pp per T-score point). Among telehealth users, greater meaning in life was associated with lower video use (−0.3 pp per point), though this effect was small in magnitude.

Shorrocks-Shapley Decomposition

Explained Variation and Shapley Decomposition

Model fit, expressed as Shapley-based R^2, was modest and varied by contrast: 0.08 (video vs audio), 0.13 (video vs none), 0.07 (audio vs none), and 0.12 (any vs none; Table 3). Decomposition of R^2 across prespecified domains highlighted distinct drivers for modality choice vs telehealth uptake. For video vs audio, digital access was the leading contributor (25.1% of explained variance), followed by SES (23.5%), health literacy and eHealth (15.3%), and geography (13.8%). Together, digital access+health literacy and eHealth accounted for 40.4% of explained variance, underscoring that device ownership, internet satisfaction, and online engagement are central to video over audio.

Table . Shapley decomposition of survey-weighted linear probability models (LPMs)[a].

Variable group	Video vs audio-only, R^2 (%)	Video vs no telehealth, R^2 (%)	Audio-only vs no telehealth, R^2 (%)	Any telehealth vs no telehealth, R^2 (%)
Demographics	0.0029 (3.65)	0.0164 (12.61)	0.0036 (5.13)	0.0142 (11.87)
Socioeconomic status	0.0188 (23.49)	0.0154 (11.83)	0.0049 (7.03)	0.0095 (7.91)
Health status and needs	0.0057 (7.12)	0.0153 (11.79)	0.0148 (21.08)	0.0186 (15.53)
Health system access	0.0030 (3.80)	0.0065 (5.00)	0.0013 (1.83)	0.0057 (4.76)
Geography	0.0110 (13.78)	0.0158 (12.15)	0.0283 (40.49)	0.0237 (19.71)
Urbanicity	0.0011 (1.34)	0.0042 (3.25)	0.0013 (1.91)	0.0038 (3.19)
Psychosocial factors	0.0016 (1.95)	0.0104 (7.97)	0.0011 (1.53)	0.0088 (7.34)
Digital access	0.0201 (25.07)	0.0186 (14.31)	0.0010 (1.36)	0.0099 (8.23)
Health literacy and eHealth	0.0122 (15.27)	0.0248 (19.06)	0.0062 (8.79)	0.0209 (17.45)
Area context	0.0036 (4.52)	0.0026 (2.03)	0.0076 (10.85)	0.0048 (4.03)
Total	0.08 (100)	0.13 (100)	0.07 (100)	0.12 (100)

[a]Shapley decomposition apportions the total R^2 of each model into the marginal contributions from distinct groups of predictor variables. All models were survey-weighted. Variable groups defined as (1) demographics: gender, marital status, age group, and race or ethnicity (White vs non-White); (2) socioeconomic status: household income, education level, and employment status; (3) health status and needs: disability status, self-rated health, cancer history, and high blood pressure history; (4) health system access: insurance status; (5) geography: US Census division; (6) urbanicity: urban, suburban, and rural; (7) psychosocial factors: Patient-Reported Outcomes Measurement Information System (PROMIS)—"Social Isolation" and "Meaning and Purpose" scores; (8) digital access: device ownership, satisfaction with home internet for health needs, and county-level percent broadband access; (9) health literacy and eHealth: self-reported health numeracy, use of health or wellness apps, and online health information-seeking behavior; and (10) area context: county-level characteristics including total population, poverty rate, unemployment rate, percent without a high school diploma, and percent of population aged ≥65 years.

For video vs no telehealth, the pattern remained digitally oriented but more diffuse: health literacy and eHealth (19.1%) and digital access (14.3%) ranked highly, as did demographics (12.6%), geography (12.2%), SES (11.8%), and health status and needs (11.8%). The digital stack contributed 33.4%, while demographics and SES jointly explained 24.4%, and geography, urbanicity, and area context explained 17.4%.

For audio-only vs no telehealth, the leading domain shifted decisively to place: geography alone explained 40.5%, followed by health status and needs (21.1%) and area context (10.9%). The digital stack was comparatively small (10.2%), while the aggregate place–based share reached 53.2%, indicating that where people live (division and local socioeconomic context) is especially salient for audio uptake. Finally, for any telehealth vs none, contributions were balanced across domains: geography (19.7%) and health literacy and eHealth (17.5%) led, followed by health status and needs (15.5%) and demographics (11.9%). The digital stack accounted for 25.7%, place-based factors for 26.9%, and demographics+SES for 19.8%. Across results, insurance and urbanicity contributed modest shares, whereas psychosocial factors contributed minimally (2%) for video vs audio but were more substantial for video vs none (8%) and

XSL•FO
RenderX

any vs none (7.3%), suggesting a limited role in modality choice but some role in overall uptake.

Missing Data Patterns and Multiple Imputation Sensitivity Analysis

Item-level missingness varied across variables. Key continuous variables had moderate missingness: PROMIS Isolation T-scores (n=334, 5.3%), PROMIS Meaning T-scores (n=314, 5%), and self-rated health (n=234, 3.7%). Sociodemographic variables with missingness included: employment status (n=390, 6.2%), disability status (n=390, 6.2%), race or ethnicity (n=619, 9.9%), and income (n=732, 11.7%). For categorical variables with modest missingness, we used a "missing-as-its-own-category" approach: cancer history (5.9% missing or unknown), digital access indicators (0.6% - 14.8% missing or not applicable), and health app use or online health information seeking (3.7% - 5.8% missing).

The null hypothesis of the MCAR test [48] was rejected (χ^2_{530}=982.1; *P*<.001), indicating that data were not MCAR. As a supplementary analysis, we conducted logistic regression examining whether missingness on income (the variable with the highest missingness) was predicted by fully observed covariates; this analysis also indicated systematic missingness (χ^2_1=1354.6, *P*<.001), with missingness associated with age, census division, cancer history, and insurance status. Together, these results indicate that missingness is related to observed covariates. Accordingly, we conducted multiple imputation by chained equations under a MAR assumption as a recommended sensitivity approach for item nonresponse in observational studies [49,50].

Multiple Imputation Sensitivity Analysis

Given the MAR finding, we conducted a sensitivity analysis using multiple imputation by chained equations with *m*=20 imputations. Imputed variables included income, race, employment, disability, self-rated health, gender, marital status, education, and PROMIS scores. Imputation models included all analysis variables plus auxiliary predictors (age, census division, urbanicity, cancer history, high blood pressure, and insurance status). Results from multiple imputations analyses were consistent with complete-case estimates, supporting the robustness of our primary findings (see Table S1 in Multimedia Appendix 1). Complete multiple imputations regression results with 95% CIs are provided in Tables S2–S4 and accompanying data files in Multimedia Appendix 1.

Discussion

Principal Findings

In a nationally representative survey of US adults in 2022, we found that overall telehealth use (any vs none) showed large regression effect sizes for health needs (eg, disability, +22.5 pp) and insurance coverage (+11.5 pp), whereas video vs audio-only modality choice among telehealth users was most strongly differentiated by digital readiness (device access and satisfaction with home internet) and eHealth behaviors. Shorrocks-Shapley decomposition revealed distinct variance patterns: for any telehealth vs none, geography (19.7%), digital factors (25.7%), and health status (15.5%) all contributed substantially; digital domains explained the largest share of variation in video vs audio choice (40.4%), while geography and area context contributed most to explaining audio-only uptake (53.2% combined for audio vs none). Men were less likely than women to use telehealth, consistent with gender differences in health care utilization [51], and telehealth users with limited device capability were more likely to rely on audio-only visits. Hypertension was positively associated with telehealth uptake but not with video vs audio choice, potentially reflecting that patients with chronic conditions requiring frequent visits benefit from any telehealth modality without a strong preference for video.

This study advances prior work in 3 ways. First, we distinguish modality choice (video vs audio) from uptake (any telehealth vs none) in a unified, survey-weighted framework, showing that digital access+health literacy and eHealth domains together explain over 40% of the variance in video vs audio choice (40.4%) and one-third in video vs none (33.4%). Second, we explicitly separate individual digital behaviors (app use and health information seeking) from device ownership and satisfaction, demonstrating that behavioral engagement is strongly linked to uptake above and beyond device availability. Third, we link person-level data with area context (division, urbanicity, and county characteristics) and apportion explained variance across domains using a Shorrocks-Shapley decomposition adapted to complex surveys—offering decision makers or researchers a transparent view of the relative leverage points for potential interventions.

A key strength of this study relative to the telehealth disparities literature is the granularity of the person-level technology data available in HINTS 6, which allows us to operationalize digital readiness beyond proxies such as binary internet access or area-level broadband coverage alone. Specifically, HINTS captures multiple, policy-relevant dimensions of digital readiness—device access and ownership, perceived adequacy of home internet for meeting health-related needs, and behavioral engagement with digital health. We leveraged this richness in a two-level way to avoid inconsistent aggregation: (1) we modeled these measures as distinct indicators in the regression analyses to preserve interpretability and map directly onto different intervention levers, and (2) we grouped them into prespecified domains in the Shorrocks-Shapley decomposition to provide a coherent summary of the relative explanatory contribution of the broader digital readiness "stack" compared with sociodemographic, clinical, and place-based domains. Future work could apply psychometric methods to develop and validate a composite digital readiness scale. In contrast, our indicator plus–domain approach prioritizes actionable interpretability for policy or practice [52,53].

The decomposition analysis shows 3 patterns. First, the digital access+health literacy and eHealth consistently explain a large fraction of variation, but its weight is contrast-specific: it is largest for video vs audio and video vs none, smaller for any vs none, and smallest for audio vs none. Second, geography+urbanicity+area context is most influential for audio vs none, moderate for any vs none and video vs none, and least for video vs audio. Third, the combined health status and

XSL•FO
RenderX

needs+health system access contribution is largest when the contrast is framed as use vs nonuse (audio vs none and any vs none), and smaller when the choice is video vs audio. Finally, because the effect size for social isolation was small and imprecise across modality contrasts, we interpret this as suggestive at best. It may reflect unmeasured factors (eg, mobility or transport barriers or care preferences) correlated with isolation rather than a direct effect.

Taken together, these patterns imply that (1) modality is largely a function of digital readiness and skills, while (2) adoption (any use vs none) is a joint function of clinical need, geography or place, and digital readiness. Despite modest R^2—typical for behavioral use models—the decomposition points to clear, policy-addressable levers: device access and digital engagement for video-based care; and place-based strategies (alongside coverage and need) to broaden overall telehealth uptake.

A prior nationwide study in older adults shows that living in a neighborhood at the lowest tertile of broadband internet subscription is associated with a significantly lower likelihood of using telehealth [54]. We similarly observe strong associations between digital readiness and video use. Lacking a smart device is strongly associated with choosing audio over video but was not significantly associated with any telehealth vs none in adjusted models. In uptake models, app use and online health information seeking are positive, whereas device category is not significantly associated with any telehealth vs none. However, our adjusted county-level percent broadband access shows small associations with any and audio telehealth. This likely reflects differences between availability vs quality or speed and aggregation by counties, which might dilute neighborhood heterogeneity. Together, the findings underscore that closing gaps requires more than network presence; adoption and affordability are important. Our modality-specific results align: device richness and internet satisfaction push use toward video over audio, while those with fewer devices or lower satisfaction are more likely to rely on audio. This is consistent with a study that emphasized the importance of optimal broadband connectivity for improved mental health access and raised quality considerations as visual cues may provide useful clinical information and improve patient-clinician rapport/access to video-based care [55].

Furthermore, a recent study by Tilhou et al [56] reported that telehealth expansion exposed differences in the use of telehealth services that persisted even among beneficiaries with high-speed internet and that expansion of telehealth service or access to high-speed internet might not close gaps in use. Despite strong associations for digital access and eHealth behaviors, geography and social context still explain sizable variance. Thus, high-quality connectivity is likely important but not sufficient without attention to coverage, device access, and digital skills or engagement. Our domain-level results empirically support that digital inclusion is considered a “super” social determinant of health [37] as the device access and skills or engagement carry a substantial explanatory weight, while area context signals persistent place-based differences. The decomposition attributes important shares to geography and smaller but nontrivial shares to area context even after detailed individual controls, reinforcing that place effects remain operative for digital health adoption, consistent with a study conducted in Chicago which found that segregation and concentrated poverty influence why residents do not have home internet access [57].

Policy and Practice Implications

Our findings point to complementary levers. First, sustaining insurance coverage and targeting populations with elevated needs (eg, disability and chronic conditions) may yield large absolute gains in telehealth use. Second, programs that upgrade device access and digital engagement (eg, app enablement and coaching in online health information seeking) should accompany broadband deployment; these address the “devices” and “skills or training” pillars alongside connectivity. Third, where home connectivity is constrained, offering high-bandwidth alternatives (community clinics and libraries) and maintaining audio-only options preserves access while longer-term infrastructure and affordability gaps are addressed. Finally, because geography and area context explain substantial variance, especially for audio telehealth–place-based investments (including affordability supports) are warranted. These implications are directly supported by our Shorrocks-Shapley decomposition, which quantifies that digital readiness is the dominant explanatory domain for video vs audio modality choice; whereas place-based factors play a larger role for audio-only uptake.

The overarching goal of telehealth is to democratize high-value care by overcoming geographic and logistical barriers. However, our findings indicate that this promise might remain unfulfilled. The significant differences identified in our nationally representative analysis reveal that telehealth, in its current form, may risk exacerbating the very differences it is intended to solve [58]. These results carry direct implications for policymakers, health care systems, and technology developers aiming to ensure the expansion of virtual care is available to all individuals.

In our data, “geography” primarily captures regional policy environments, market or provider supply for virtual care, as well as digital infrastructure and adoption. “Underserved” in this context refers to areas with lower broadband adoption or affordability and speeds, higher area deprivation, rural or micropolitan location, and limited provider supply for video-capable care. For audio only vs no telehealth, geography explained the largest share of the model’s variance—substantially more than in other contrasts—underscoring that a one-size-fits-all strategy is insufficient. The significant reductions in the probability of telehealth use by geography, for example, for residents of the Midwest and South compared to the Northeast point to potential structural issues. This aligns with research highlighting how state-level policies, particularly those governing reimbursement and physician licensure, create a fragmented regulatory landscape that hinders uniform adoption [59]. To address this, federal and state policies should prioritize targeted investments in underserved regions [60,61]. This includes not only expanding community infrastructure like broadband internet, a known barrier [60-64], but also incentivizing provider adoption through sustainable reimbursement models and advancing interstate

XSL•FO
RenderX

licensure compacts to improve the supply of virtual care in these areas.

Overall, since affordability is a binding constraint for adoption, recent federal policy changes—including 2025 adjustments to Broadband Equity, Access, and Deployment program affordability requirements and the termination of Digital Equity Act grant programs—threaten to undermine gains in telehealth use, particularly in high-need regions where subsidized connectivity and digital inclusion supports have been critical [65,66].

Second, our analysis provides robust empirical weight to the emerging concept of digital determinants of health [63,64]. We found that the combined domains of health literacy and digital access are as important as geography in explaining telehealth use. Critically, this difference is not merely about access to a device but about the capability to use it effectively for health. We did not find consistent negative associations for low health numeracy after adjustment. Instead, app use and online information seeking were the behavioral factors most strongly linked to higher telehealth uptake. One interpretation is that these behaviors proxy digital readiness—familiarity with app-based health workflows similar to those required for scheduling, check-in, and joining video visits. Patient studies of video visits report that technical and setup challenges (eg, connecting to the visit, audio or video problems, and confusing initial setup) remain common and systematic reviews identify limited technical literacy and technology barriers as recurring obstacles to telehealth use in older adults. Accordingly, this pattern is hypothesis-generating and suggests that the usability burden of telehealth platforms could disproportionately deter patients with less prior digital health experience [5,8,9,67].

Overall, initiatives should address both digital access and digital literacy [68]. This requires a 2-pronged approach: funding for community-based digital health literacy programs and a push for user-centered design principles in the development of patient-facing technologies to reduce overall burden on users.

Limitations

Cross-sectional data preclude causal inference and telehealth measures were self-reported and lacked encounter frequency or quality. County-level broadband access may mis-measure adoption, affordability, and speed (and operates at a coarser geography); collinearity with urbanicity and digital behaviors may attenuate or reverse associations for the county broadband variable. Model R^2 values are modest, indicating unmeasured determinants (eg, provider supply). Future research should examine finer gradations of educational measure to distinguish groups that may face distinct barriers to telehealth adoption. Nevertheless, the survey-weighted national scope, explicit modality contrasts, and Shorrocks-Shapley decomposition across domains provide a comprehensive, policy-relevant portrait of telehealth determinants in 2022. Furthermore, understanding intersectional perspectives across race, income, education, and broadband availability might be an area of fruitful research.

Conclusions

Video telehealth uptake in the United States reflects intersecting determinants with modality choice driven most strongly by digital readiness and overall uptake shaped by digital factors, place-based characteristics, and health needs. Policies that simultaneously advance connectivity, affordability, device access, and digital skills support are likely to narrow persistent gaps in telehealth use across modalities. Methodologically, this study advances the telehealth disparities literature by applying Shorrocks-Shapley decomposition to survey-weighted national data to quantify the relative explanatory contribution of determinant domains rather than examining predictors in isolation. Our approach provides a domain-level ranking of where disparities are most strongly concentrated, distinguishing digital readiness mechanisms from place-based constraints. Leveraging HINTS' rich person-level technology measures, we show that digital access and eHealth behaviors account for a larger share of explained variation in video-vs-audio modality choice than traditional sociodemographic characteristics. These findings support a dual strategy: maintain reimbursable audio-only care to preserve access while digital constraints persist and pair sustained coverage with targeted investments in devices, affordable high-speed connectivity, and digital skills supports to enable more equitable access to video-based telehealth.

Acknowledgments

No generative artificial intelligence tools were used to write or edit the scientific content of this manuscript. Any grammar or spell checking tools did not alter the authors' interpretation of findings.

Funding

This research has been funded by Health Resources and Services Administration (grant U3GRH40001)

Data Availability

This study used the Health Information National Trends Survey restricted-use dataset containing county identifiers; access requires approval through the National Cancer Institute restricted data request process and a data use agreement, and the authors are not permitted to redistribute the restricted-use file. County-level contextual indicators were obtained from the American Community Survey, NIMHD HD Pulse broadband indicators, and the CDC/ATSDR Social Vulnerability Index. Supplementary data files are provided in Appendix D in Multimedia Appendix 1, including complete LPM regression results with 95% CIs for all four contrasts (LPM Video vs Audio Complete.csv, LPM Video vs NoTelehealth Complete.csv, LPM Audio vs NoTelehealth Complete.csv, LPM AnyTelehealth vs NoTelehealth Complete.csv); multiple imputation results (MI regression results CI.csv, MI vs CompleteCase

XSL•FO
RenderX

comparison.csv, MI regression results.csv); and missingness diagnostics (littles mcar test results.txt). Analysis code will be made available in a public repository upon acceptance; during peer review, code is available from the corresponding author upon reasonable request.

Authors' Contributions

Concept and design: C Bolbocean.
Data acquisition of data: C Bolbocean, C Bogulski.
Drafting of the manuscript: C Bolbocean.
Statistical analysis: C Bolbocean.
Interpretation of data: C Bolbocean, CH, C Bogulski.
Critical revision of the paper for important intellectual content: C Bolbocean, CH, C Bogulski.
Obtaining funding: C Bolbocean, CH, C Bogulski.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials including definition of variables, complete linear probability model results, multiple imputation sensitivity analysis, and supplementary data files.
[PDF File, 265 KB - jmir_v28i1e81879_app1.pdf]

References

1. What is telehealth? Health Resources and Services Administration, US Department of Health and Human Services. 2024. URL: https://www.hrsa.gov/telehealth/what-is-telehealth [accessed 2025-06-10]
2. Kichloo A, Albosta M, Dettloff K, et al. Telemedicine, the current COVID-19 pandemic and the future: a narrative review and perspectives moving forward in the USA. Fam Med Community Health 2020 Aug;8(3):e000530. [doi: 10.1136/fmch-2020-000530] [Medline: 32816942]
3. Bashshur RL, Howell JD, Krupinski EA, Harms KM, Bashshur N, Doarn CR. The empirical foundations of telemedicine interventions in primary care. Telemed J E Health 2016 May;22(5):342-375. [doi: 10.1089/tmj.2016.0045] [Medline: 27128779]
4. Lyles CR, Schillinger D, Sarkar U. Connecting the dots: health information technology expansion and health disparities. PLoS Med 2015 Jul;12(7):e1001852. [doi: 10.1371/journal.pmed.1001852] [Medline: 26172977]
5. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. J Am Med Inform Assoc 2018 Aug 1;25(8):1080-1088. [doi: 10.1093/jamia/ocy052] [Medline: 29788380]
6. Kates J, Ranji U, Beamesderfer A, Salganicoff A, Dawson L. Health and access to care and coverage for lesbian, gay, bisexual, and transgender (LGBT) individuals in the US. : Kaiser Family Foundation; 2018 URL: https://collections.nlm.nih.gov/master/borndig/101740255/Issue-Brief-Health-and-Access-to-Care-and-Coverage-for-LGBT-Individuals-in-the-US.pdf [accessed 2026-02-14]
7. Nouri S, Khoong EC, Lyles CR, Karliner L. Addressing equity in telemedicine for chronic disease management during the COVID-19 pandemic. NEJM Catalyst Innov Care Deliv 2020;1(3). [doi: 10.1056/CAT.20.0123]
8. Ternes S, Lavin L, Vakkalanka JP, et al. The role of increasing synchronous telehealth use during the COVID-19 pandemic on disparities in access to healthcare: a systematic review. J Telemed Telecare 2025 Sep;31(8):1091-1113. [doi: 10.1177/1357633X241245459] [Medline: 38646804]
9. Chang JE, Lai AY, Gupta A, Nguyen AM, Berry CA, Shelley DR. Rapid transition to telehealth and the digital divide: implications for primary care access and equity in a post‐COVID era. Milbank Q 2021 Jun;99(2):340-368. [doi: 10.1111/1468-0009.12509] [Medline: 34075622]
10. Garfan S, Alamoodi AH, Zaidan BB, et al. Telehealth utilization during the COVID-19 pandemic: a systematic review. Comput Biol Med 2021 Nov;138:104878. [doi: 10.1016/j.compbiomed.2021.104878] [Medline: 34592585]
11. Vakkalanka JP, Gadag K, Lavin L, et al. Telehealth use and health equity for mental health and substance use disorder during the COVID-19 pandemic: a systematic review. Telemed J E Health 2024 May;30(5):1205-1220. [doi: 10.1089/tmj.2023.0588] [Medline: 38227387]
12. Eberly LA, Khatana SAM, Nathan AS, et al. Telemedicine outpatient cardiovascular care during the COVID-19 pandemic: bridging or opening the digital divide? Circulation 2020 Aug 4;142(5):510-512. [doi: 10.1161/CIRCULATIONAHA.120.048185] [Medline: 32510987]
13. Forcino RC, Johnson JK, Elwyn G. Telehealth modality preferences for video and voice-only visits among US clinicians and the public: cross-sectional, web-based survey study. J Med Internet Res 2025 Jun 13;27:e72276. [doi: 10.2196/72276] [Medline: 40513030]

XSL•FO
RenderX

14. Samson LW, Couture SJ, Creedon TB, Jacobus-Kantor L, Sheingold S. Updated Medicare FFS telehealth trends by beneficiary characteristics, visit specialty, and state, 2019-2021. : Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services; 2023 Mar URL: https://aspe.hhs.gov/sites/default/files/documents/cb83f6f25c25c3a3529807f23cd2327d/medicare-telehealth-updated-trends-report.pdf [accessed 2026-03-06]
15. Saharkhiz M, Rao T, Parker-Lue S, Borelli S, Johnson K, Cataife G. Telehealth expansion and Medicare beneficiaries' care quality and access. JAMA Netw Open 2024 May 1;7(5):e2411006. [doi: 10.1001/jamanetworkopen.2024.11006] [Medline: 38739388]
16. What to know about medicare coverage of telehealth. Kaiser Family Foundation. 2024. URL: https://www.kff.org/medicare/what-to-know-about-medicare-coverage-of-telehealth/ [accessed 2025-10-14]
17. Lucas JW, Villarroel MA. Telemedicine use among adults: United States, 2021. NCHS Data Brief 2022 Oct(445):1-8. [Medline: 36255940]
18. Chang E, Penfold RB, Berkman ND. Patient characteristics and telemedicine use in the US, 2022. JAMA Netw Open 2024 Mar 4;7(3):e243354. [doi: 10.1001/jamanetworkopen.2024.3354] [Medline: 38517438]
19. FCC launches Affordable Connectivity Program. Federal Communications Commission. 2021 Dec. URL: https://www.fcc.gov/fcc-launches-affordable-connectivity-program [accessed 2026-02-14]
20. Affordable Connectivity Program. Federal Communications Commission. 2024. URL: https://www.fcc.gov/acp [accessed 2026-02-14]
21. Half of ACP-eligible households still unaware of the program. Benton Institute for Broadband & Society. 2023 Mar. URL: https://www.benton.org/blog/half-acp-eligible-households-still-unaware-program [accessed 2026-02-14]
22. ACP enrollment and claims tracker. Universal Service Administrative Company. 2024. URL: https://www.usac.org/about/affordable-connectivity-program/acp-enrollment-and-claims-tracker [accessed 2026-02-14]
23. Affordable connectivity program wind down. Federal Communications Commission. 2024. URL: https://www.fcc.gov/acp-wind-down [accessed 2026-03-06]
24. Annual consumer survey fielded november 1–20, 2024; 68,000+ respondents. National Lifeline Association. 2025. URL: https://eqm2782zctq.exactdn.com/wp-content/uploads/2025/01/NaLA-Annual-Consumer-Survey-1.22.25-FINAL.pdf [accessed 2026-03-06]
25. van Deursen AJ, van Dijk JA. The first-level digital divide shifts from inequalities in physical access to inequalities in material access. New Media Soc 2019 Feb;21(2):354-375. [doi: 10.1177/1461444818797082]
26. Rodriguez JA, Betancourt JR, Sequist TD, Ganguli I. Differences in the use of telephone and video telemedicine visits during the COVID-19 pandemic. Am J Manag Care 2021 Jan;27(1):21-26. [doi: 10.37765/ajmc.2021.88573] [Medline: 33471458]
27. Sachs JW, Graven P, Gold JA, Kassakian SZ. Disparities in telephone and video telehealth engagement during the COVID-19 pandemic. JAMIA Open 2021 Jul;4(3):ooab056. [doi: 10.1093/jamiaopen/ooab056] [Medline: 34632322]
28. Shorrocks AF. Decomposition procedures for distributional analysis: a unified framework based on the Shapley value. J Econ Inequal 2013 Mar;11(1):99-126. [doi: 10.1007/s10888-011-9214-z]
29. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. The Lancet 2007 Oct;370(9596):1453-1457. [doi: 10.1016/S0140-6736(07)61602-X]
30. Blake KD, Moser RP, D'Angelo H, Gaysynsky A, Vanderpool RC. The evolution of NCI's Health Information National Trends Survey: methods, data, and future directions. J Natl Cancer Inst 2025 Jun 1;117(6):1110-1116. [doi: 10.1093/jnci/djae317] [Medline: 39657961]
31. Finney Rutten LJ, Blake KD, Skolnick VG, Davis T, Moser RP, Hesse BW. Data resource profile: the national cancer institute's health information national trends survey (HINTS). Int J Epidemiol 2020 Feb 1;49(1):17-17j. [doi: 10.1093/ije/dyz083] [Medline: 31038687]
32. American Community Survey data. US Census Bureau. URL: https://www.census.gov/programs-surveys/acs/data.html [accessed 2026-02-14]
33. Broadband access for United States by county: all races (includes Hispanic/Latino), both sexes, all ages, 2019-2023 (sorted by percent). National Institute on Minority Health and Health Disparities. URL: https://hdpulse.nimhd.nih.gov/data-portal/physical/table?age=001&age_options=ageall_1&demo=235&demo_options=broadband_1&physicaltopic=006 [accessed 2026-02-14]
34. SVI data & documentation download. Centers for Disease Control and Prevention (CDC) and Agency for Toxic Substances and Disease Registry (ATSDR). URL: https://www.atsdr.cdc.gov/place-health/php/svi/svi-data-documentation-download.html [accessed 2026-03-06]
35. Lam K, Lu AD, Shi Y, Covinsky KE. Assessing telemedicine unreadiness among older adults in the United States during the COVID-19 pandemic. JAMA Intern Med 2020 Oct 1;180(10):1389-1391. [doi: 10.1001/jamainternmed.2020.2671] [Medline: 32744593]
36. van Dijk J. Digital divide research, achievements and shortcomings. Poetics 2006 Aug;34(4-5):221-235. [doi: 10.1016/j.poetic.2006.05.004]

XSL•FO
RenderX

37. Sheon A, Khoong EC. Pillars, policies, and plausible pathways linking digital inclusion and health equity. Health Aff 2024. [doi: 10.1377/hpb20240327.505]
38. Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH roadmap cooperative group during its first two years. Med Care 2007 May;45(5 Suppl 1):S3-S11. [doi: 10.1097/01.mlr.0000258615.42478.55] [Medline: 17443116]
39. Hays RD, Bjorner JB, Revicki DA, Spritzer KL, Cella D. Development of physical and mental health summary scores from the Patient-Reported Outcomes Measurement Information System (PROMIS) global items. Qual Life Res 2009 Sep;18(7):873-880. [doi: 10.1007/s11136-009-9496-9] [Medline: 19543809]
40. Nicholson NR. Social isolation in older adults: an evolutionary concept analysis. J Adv Nurs 2009 Jun;65(6):1342-1352. [doi: 10.1111/j.1365-2648.2008.04959.x] [Medline: 23163719]
41. Hawkley LC, Cacioppo JT. Loneliness matters: a theoretical and empirical review of consequences and mechanisms. Ann Behav Med 2010 Oct;40(2):218-227. [doi: 10.1007/s12160-010-9210-8] [Medline: 20652462]
42. Roepke AM, Jayawickreme E, Riffle OM. Meaning and health: a systematic review. Appl Res Qual Life 2014 Dec;9(4):1055-1079. [doi: 10.1007/s11482-013-9288-9]
43. Lee MJ, Lee G, Choi JY. Linear probability model revisited: why it works and how it should be specified. Sociol Methods Res 2025;54(1):173-186. [doi: 10.1177/0049124123117685]
44. Wooldridge JM. Econometric Analysis of Cross Section and Panel Data, 2nd edition: MIT press; 2002.
45. Moser RP, Naveed S, Cantor D, et al. Health Information National Trends Survey: Integrative analytic methods using population-level cross-sectional data. : National Cancer Institute; 2013 URL: https://hints.cancer.gov/docs/hints_ida_report.pdf [accessed 2026-02-14]
46. Suo Z, Shao L, Lang Y. A study on the factors influencing the utilization of public health services by China's migrant population based on the Shapley value method. BMC Public Health 2023 Nov 24;23(1):2328. [doi: 10.1186/s12889-023-17193-3] [Medline: 38001411]
47. Juarez FWC. Shapley2: Stata module to compute additive decomposition of estimation statistics by regressors or groups of regressors. Statistical Software Components. 2015. URL: https://ideas.repec.org/c/boc/bocode/s457543.html [accessed 2026-02-14]
48. Little RJA. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 1988 Dec;83(404):1198-1202. [doi: 10.1080/01621459.1988.10478722]
49. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009 Jun 29;338:b2393. [doi: 10.1136/bmj.b2393] [Medline: 19564179]
50. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med 2011 Feb 20;30(4):377-399. [doi: 10.1002/sim.4067] [Medline: 21225900]
51. Bertakis KD, Azari R, Helms LJ, Callahan EJ, Robbins JA. Gender differences in the utilization of health care services. J Fam Pract 2000 Feb;49(2):147-152. [Medline: 10718692]
52. Greco S, Ishizaka A, Tasiou M, Torrisi G. On the methodological framework of composite indices: a review of the issues of weighting, aggregation, and robustness. Soc Indic Res 2019 Jan;141(1):61-94. [doi: 10.1007/s11205-017-1832-9]
53. Nardo M, Saisana M, Saltelli A, Tarantola S, Hoffmann A, Giovannini E. Handbook on Constructing Composite Indicators: Methodology and User Guide: OECD Publishing; 2008. [doi: 10.1787/9789264043466-en]
54. Okoye SM, Mulcahy JF, Fabius CD, Burgdorf JG, Wolff JL. Neighborhood broadband and use of telehealth among older adults: cross-sectional study of national survey data linked with census data. J Med Internet Res 2021 Jun 14;23(6):e26242. [doi: 10.2196/26242] [Medline: 34125071]
55. O'Shea AMJ, Howren MB, Mulligan K, Haraldsson B, Shahnazi A, Kaboli PJ. Quantifying the digital divide: associations of broadband internet with tele-mental health access before and during the COVID-19 pandemic. J Gen Intern Med 2023 Jul;38(Suppl 3):832-840. [doi: 10.1007/s11606-023-08120-8] [Medline: 37340258]
56. Tilhou AS, Jain A, DeLeire T. Telehealth expansion, internet speed, and primary care access before and during COVID-19. JAMA Netw Open 2024 Jan 2;7(1):e2347686. [doi: 10.1001/jamanetworkopen.2023.47686] [Medline: 38180762]
57. Mossberger K, Tolbert CJ, Bowen D, Jimenez B. Unraveling different barriers to internet use: urban residents and neighborhood effects. Urban Aff Rev 2012;48(6):771-810. [doi: 10.1177/1078087412453713]
58. SteelFisher GK, McMurtry CL, Caporello H, et al. Video telemedicine experiences in COVID-19 were positive, but physicians and patients prefer in-person care for the future: study examines patient and physician opinion of telemedicine experiences during COVID-19. Health Aff 2023;42(4):575-584. [doi: 10.1377/hlthaff.2022.01027]
59. Turner Lee N, Karsten J, Roberts J. Removing regulatory barriers to telehealth before and after COVID-19. : Brookings Institution; 2020 URL: https://www.brookings.edu/wp-content/uploads/2020/05/Removing-barriers-to-telehealth-before-and-after-COVID-19_PDF.pdf [accessed 2026-02-14]
60. Sheerah HA, AlSalamah S, Alsalamah SA, et al. The rise of virtual health care: transforming the health care landscape in the Kingdom of Saudi Arabia: a review article. Telemed J E Health 2024 Oct;30(10):2545-2554. [doi: 10.1089/tmj.2024.0114] [Medline: 38984415]

XSL•FO
RenderX

61. Epstein JA, Lkhagvajav Z, Young T, Bertram A, Yeh HC, Taylor CO. Will the doctor “see” you now? The development and implementation of a targeted telemedicine system for primary care. ACI open 2023 Jul;7(2):e71-e78. [doi: 10.1055/s-0043-1776038] [Medline: 37900978]
62. Kane CK, Gillis K. The use of telemedicine by physicians: still the exception rather than the rule. Health Aff (Millwood) 2018 Dec;37(12):1923-1930. [doi: 10.1377/hlthaff.2018.05077] [Medline: 30633670]
63. Jahnel T, Dassow HH, Gerhardus A, Schüz B. The digital rainbow: digital determinants of health inequities. Digit Health 2022;8:20552076221129093. [doi: 10.1177/20552076221129093] [Medline: 36204706]
64. Chidambaram S, Jain B, Jain U, et al. An introduction to digital determinants of health. PLOS Digit Health 2024 Jan;3(1):e0000346. [doi: 10.1371/journal.pdig.0000346] [Medline: 38175828]
65. Broadband equity, access, and deployment (BEAD) program: BEAD restructuring policy notice. : National Telecommunications and Information Administration; 2025 URL: https://www.ntia.gov/sites/default/files/2025-06/bead-restructuring-policy-notice.pdf [accessed 2026-02-14]
66. NTIA broadband programs: semiannual status report. : US Department of Commerce, Office of Inspector General; 2025 URL: https://www.oig.doc.gov/wp-content/OIGPublications/OIG-25-031-I_FinalReport-SECURED.pdf [accessed 2026-02-14]
67. Kruse C, Fohn J, Wilson N, Nunez Patlan E, Zipp S, Mileski M. Utilization barriers and medical outcomes commensurate with the use of telehealth among older adults: systematic review. JMIR Med Inform 2020 Aug 12;8(8):e20359. [doi: 10.2196/20359] [Medline: 32784177]
68. Roberts ET, Mehrotra A. Assessment of disparities in digital access among Medicare beneficiaries and implications for telemedicine. JAMA Intern Med 2020 Oct 1;180(10):1386-1389. [doi: 10.1001/jamainternmed.2020.2666] [Medline: 32744601]

Abbreviations

ACP: Affordable Connectivity Program
FIPS: Federal Information Processing Series
HINTS: Health Information National Trends Survey
HINTS: Health Information National Trends Survey
LPM: linear probability model
MAR: missing at random
MCAR: missing completely at random
PHE: public health emergency
pp: percentage point
PROMIS: Patient-Reported Outcomes Measurement Information System
SES: socioeconomic status
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by S Brini; submitted 05.Aug.2025; peer-reviewed by A Sheon, M Okonkwo, O Okunbor; accepted 30.Jan.2026; published 18.Mar.2026.

Please cite as:
Bolbocean C, Hayes C, Bogulski C
Telehealth Use and Modality Choice Among US Adults: Shorrocks-Shapley Decomposition of a 2022 Cross-Sectional National Survey
J Med Internet Res 2026;28:e81879
URL: https://www.jmir.org/2026/1/e81879
doi:10.2196/81879

© Corneliu Bolbocean, Corey Hayes, Cari Bogulski. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Attitudes Toward Video Consultations From the Perspective of Physicians and Psychotherapists in German Outpatient Care After the COVID-19 Pandemic: Survey Study

Lara Kleinschmidt[1], MSc; Juergen Wasem[1], Prof Dr; Nikola Blase[1], Dr med; Beatrice Nauendorf[2], MSc; Juliane Malsch[2], MSc; Matthias Brittner[3], MA; Paul Brandenburg[4], MA; André Aeustergerling[5], MA; Theresa Hüer[1], Dr rer pol

[1]Institute for Health Care Management and Research, University of Duisburg-Essen, Thea-Leymann-Straße 9, Essen, Germany

[2]Kassenärztliche Vereinigung Berlin, Berlin, Germany

[3]Kassenärztliche Vereinigung Westfalen-Lippe, Dortmund, Germany

[4]Kassenärztliche Vereinigung Schleswig-Holstein, Bad Segeberg, Germany

[5]Kassenärztliche Vereinigung Mecklenburg-Vorpommern, Schwerin, Germany

Corresponding Author:
Lara Kleinschmidt, MSc
Institute for Health Care Management and Research, University of Duisburg-Essen, Thea-Leymann-Straße 9, Essen, Germany

Abstract

Background: Although video consultations (VCs) are permitted in German outpatient care and have seen a notable rise during the COVID-19 pandemic, their use still does not seem to have become established in Germany.

Objective: This survey aims to evaluate the attitudes of physicians and psychotherapists with regard to the use of VC after the COVID-19 pandemic, in particular in the context of types of treatment and suitable medical fields.

Methods: A standardized questionnaire was sent out to all 34,095 physicians and psychotherapists in 4 German regions. The analysis consisted of both descriptive and inferential statistics. Subgroup analysis included gender, age groups, community size of practice location, VC experience, type and ownership of practice, and area of medical care. Binary logistic regression was conducted to determine whether physicians' and psychotherapists' individual factors, organizational factors, or area of medical care were associated with at least monthly VC provision or interest in VC provision.

Results: The response rate was 17.9%, including a total of 5930 participants in the analysis. About 40% (2216/5863) of the physicians and psychotherapists surveyed stated that they offer VC at least once a month. In the area of medical care, the odds ratio (OR) of at least monthly VC provision in psychotherapeutic care was about 8.2 (95% CI 7.4 - 1.64; $P<.001$) compared to primary care, whereas in specialist care, the odds for monthly VC provision were approximately 50% lower than in primary care (OR 0.5, 95% CI 0.43-0.59; $P<.001$). Further, female participants have higher odds to provide VC at least once a month (OR 1.163, 95% CI 1.01 - 1.34; $P=.03$). The odds for monthly VC provision in older age groups are approximately 60% higher than in the <40 years old age group (OR 0.41, 95% CI 0.32-0.52; $P<.001$). Around 80% (4347/5442) of the participants expressed interest in VC use. The most common occasions for which treatment by VC was reported to be suitable were discussing test results (1422/1896, 75.0%), taking the patient's medical history (1195/2147, 55.7%), issuing prescriptions for drugs and remedies (793/1204, 65.9%), and the issuing of incapacity certificates for work (677/1042, 65.0%).

Conclusions: There has been an increase in the self-reported uptake of VC among physicians and psychotherapists compared to prepandemic levels, although this remains at a relatively low level in primary and specialist care. A significant proportion of doctors and psychotherapists have expressed an interest in using VC after the pandemic period. However, this self-reported use is not yet reflected in actual usage data, suggesting the need for further investigation into the underlying factors influencing the gap and identifying potential enablers. Further, these self-assessments by service providers on suitable types of treatment and suitable medical fields can inform political decision-making.

(J Med Internet Res 2026;28:e73757) doi:10.2196/73757

KEYWORDS

eHealth; telehealth; telemedicine; remote care; mental health services; provider views; virtual medicine; attitude of health personnel; COVID-19; video consultation; survey; suitable types of treatment; medical fields; German outpatient medical care; physician and psychotherapist perspective; provider perspective

XSL•FO
RenderX

Introduction

The utilization of telemedicine, in particular the implementation of video consultation (VC), has become increasingly popular since the onset of the global COVID-19 pandemic in 2020 [1-3]. In Germany, VC could be billed for certain diagnoses or check-ups in outpatient care since 2017. In 2019, the legislative initiative of the Digital Healthcare Act was implemented with the objective of accelerating the digital transformation of the nation's health care infrastructure. Thereafter, VCs have been eligible for billing in almost all medical specialties, as physicians and psychotherapists have been granted permission to use them at their discretion, without the prior restrictions [4].

However, uptake in its use was only reported during the COVID-19 pandemic as it was a relevant option for preventing infections [4-6]. Toward the end of the pandemic, its use has declined again and does still not appear to have become established into everyday medical practice [7]. A representative survey of 2800 medical practices in Germany by Albrecht et al [8] revealed that in 2021, 20% of practices offered VC, compared to 25% during the beginning of the COVID-19 pandemic in 2020. Among psychotherapy practices, VCs are used more frequently, with about three quarters of practices offering the service in 2020 and 2021 [8]. Claims data analysis by the statutory health insurance fund BARMER reported that in 2021 around 80% of the VCs have been provided by psychiatrists and psychotherapists, followed by general practitioners (GPs) who provided 14% of all VCs. Less prevalent are the medical specialties gynecologists (1.8%) and pediatricians (1.5%) [7]. According to the latest figures reported by the Zentralinstitut kassenärztliche Versorgung, the use of VC in Germany did not demonstrate an upward trend from 2022 to 2024, with approximately 2.7 million VCs recorded in both years [9].

For the period before and during the COVID-19 pandemic, there are study results that have identified a number of factors that influence the provision of VC services. Besides medical specialty, these include the age of the service provider, the community size of practice location, and the size of the service provider's practice. A greater uptake in VC use during the COVID-19 pandemic was observed among younger physicians and psychotherapists [6,8] and among providers in an urban practice [6,10]. In terms of practice type, Kane et al [11] and Albrecht et al [8] reported higher use in larger practices, multispecialty practices, or practices with a nonphysician ownership structure. Conversely, Knörr et al [10] found no associations for use but did find associations for perceived benefit with group practices rather than solo practices. At that time, VCs were most frequently employed for the purposes of discussing examination results and taking a patient's medical history. VCs have previously been reported as being inadequate for the purposes of diagnosis and determining medical indication [8].

Moreover, VCs have been most commonly used as an adjunct to in-person treatment. However, in rural regions with a lower density of physicians, VCs have been increasingly provided without additional personal contact in a billing quarter, that is, in 3 months [6]. This indicates that VC can also be a suitable method in the context of a shortage of medical care. Internationally, service providers have reported positive experiences with chronic disease management, mental health support, and medication management during the pandemic [12]. Despite the fact that experiences with VC during the pandemic have frequently been characterized as positive overall [13], health care providers appear to be rather hesitant about implementing them into routine practice [14].

The sudden and unanticipated need for VC adoption during the course of the pandemic was widely regarded as an opportunity for their long-term integration into routine treatment. At the same time, it accelerated ongoing political initiatives in Northern Europe to digitize health care [15-17]. However, in Germany, measures to promote the use of VC do not seem to have been effectively implemented, which has created uncertainty about the sustainability of VC. This underscores the need for research on how VC use evolved. Although there is plenty of research on providers' experiences with VC during the pandemic, given the potential shifts in attitudes toward such services resulting thereof, it is crucial to examine the subject from a provider's perspective and their attitudes toward telemedicine services in a postpandemic context [14]. Hence, the objective of this study is to examine the attitudes of physicians and psychotherapists with regard to the use of VC after the COVID-19 pandemic in German outpatient care. Specifically, the study examines factors associated with VC provision and interest in VC provision. Further, the study investigates which types of treatments and medical fields are deemed suitable by the medical and psychotherapeutic service providers in nonpandemic conditions. This survey of service providers' perspectives can assist in ascertaining how VC can be used sensibly. For many, their familiarity with VC was built in the context of the pandemic, when it was quickly implemented. There is now a need for evaluation to establish an evidence base for its optimal use [16,18].

Methods

Study Design

This survey was conducted as part of the German study "Preference-based use of video consultations in urban and rural regions," which was funded by the Innovation Fund of the German Federal Joint Committee (funding no. 01VSF20011). A study protocol was published previously [19]. Based on groundwork research within the study—a literature review and focus group discussions—a standardized questionnaire was developed.

The questionnaire was divided into three segments: (1) attitudes and experiences regarding VC provision, (2) preference survey using discrete choice experiments, and (3) sociodemographic information as well as information about the medical profession.

First, the survey asked whether VC was offered and, if so, how frequently. A follow-up question enquired how long VC had been part of their medical treatment. The first section also asked about suitable types of treatment and medical fields for which VC could be (at least partially) suitable. There were also

XSL•FO
RenderX

questions about the exclusive use of VC (eg, no additional face-to-face contact in a quarter for the treatment case), the use of VC in treating children and adolescents, and potential future VC provision. All questions had categorical responses except the questions about suitable types of treatment and suitable medical fields for treatment via VC, which were 5-scale Likert questions (highly unsuitable, unsuitable, to some extent, suitable, and very suitable). All Likert scales had an "I don't know" option, which was coded as missing during the analysis. Likert scales also included open-text response options to allow for further explanation of the participants' experiences.

The attitude of inhibiting and promoting factors for the provision of VC and the results of the second part of the survey (preference elicitation) are published separately. Thus, no further details are given here.

The third section included sociodemographic questions and information about the participants' medical profession. These included membership at one of the four participating regional Associations of Statutory Health Insurance Physicians (ASHIPs), community size of practice location, employment status, area of medical care, and medical specialty or psychotherapy practice.

Prior to conducting the survey, a pretest with think-aloud and probing methods was done [20]. The pretest included 6 physicians with different medical specialties.

Ethical Considerations

At the beginning of the survey, all participants received information about the objective of the study and how their data would be dealt with. Participation was anonymous and voluntary, and participants could opt out at anytime. Entries could be made on paper or online using a QR code using QuestionPro online survey software (QuestionPro, Inc). The incentive to participate in the survey was amplified by the prospect of winning a tablet in a raffle. The local ethics committee of the Medical Faculty of the University of Duisburg-Essen approved the study (reference: 21‑10283-BO). Completion time of the survey was around 20 minutes. An excerpt of the survey can be found in Multimedia Appendix 1.

Recruitment

The 4 participating ASHIPs in the regions Berlin, Westphalia-Lippe, Mecklenburg-Western Pomerania, and Schleswig-Holstein sent the survey via postal delivery to all their members (physicians and psychotherapists) who met the inclusion criterion. Inclusion criterion was being eligible to provide VC. In Germany, outpatient medical care in the field of psychotherapeutic/psychiatric disorders is provided by medical care providers (eg, medical psychotherapists, psychiatrists, neurologists, family physicians) and nonphysician care providers called psychological psychotherapists. As both groups are allowed to provide VC and are reimbursed under the same remuneration system, psychological psychotherapists were also included in the study. The ASHIPs cover rural as well as urban regions.

In order to gain insight into the postpandemic perspective, the survey was conducted between November 2022 and March 2023. No reminder was sent out, as the target of 10% response rate was reached early.

Statistical Analysis

Descriptive data are presented as frequencies (n) and percentages (%). All 49 medical specialty categories were aggregated to 12 final medical specialist groups to simplify the evaluation. For the gender, participants who answered "diverse" were coded as missing in the gender variable for the purposes of subgroup analysis. This decision was taken to safeguard the privacy of those involved because of the relatively limited number of participants with this response. As not all questions were mandatory, missing answers were possible, and no further conclusions can be drawn about the number of "diverse" cases.

Subgroup analyses included gender (male or female), age groups (<40 y, 40‑50 y, 51‑60 y, and >60 y), community size in reference to the classification "Stadt-und Gemeindetyp" as rural or urban by the Federal Institute for Research on Building, Urban Affairs, and Spatial Development (rural community, small town, middle town, or large city) [21], VC experience (yes or no), type of practice (individual practice, joint practice, group practice, or medical care unit), ownership of practice (self-employed and employed), and area of medical care (primary care, specialist care, or psychotherapeutic care). Physicians and psychotherapists were categorized as having VC experience if they stated they offered VC at least once a month. In Germany, the distinction between joint practice and group practice is primarily evident in the manner of billing. In joint practice, a joint billing system is employed, whereas in group practice, the physicians share the facilities but operate independently of one another. Medical care units are outpatient facilities in which physicians are employed exclusively and which can be operated by physicians, hospitals, or other licensed providers. With regard to the variable type of practice, multiple responses were permitted in the questionnaire; however, cases with multiple answers were excluded for the subgroup analysis. Regarding areas of medical care, in primary care, all family physicians, including GPs, general internists, and pediatricians, are displayed. Specialist care, on the other hand, comprised physicians who have indicated that they work in specialist care but not in the areas of family medicine and psychiatry or psychotherapy. With regard to psychotherapeutic care, all physicians or nonphysician service providers (psychological psychotherapists) who have reported working in this field were included. All participants working in psychotherapy, psychiatry, or neurology care were referred to in the study as "psychotherapists," unless otherwise stated.

First, a descriptive analysis was conducted, followed by an association analysis as a second step and binary logistic regression as a third step. Correlation and association coefficients were calculated using the Stuart-Kendall Tau-c for ordinal variables and Pearson χ^2 categorical variables. Effect sizes were indicated by Stuart-Kendall Tau-c for ordinal variables and Cramer *V* for categorical variables [22,23]. In accordance with the criteria established by Cohen (*d*), effect sizes that fall below the threshold of 0 are considered inconsequential and thus excluded from the reporting. Effect sizes greater than 0.1 and up to 0.3 are considered weak, while

XSL•FO
RenderX

those greater than 0.3 are considered moderate, and those greater than 0.5 are considered strong [24].

Binary logistic regression was conducted to determine whether physicians' and psychotherapists' individual factors (gender, age, and region), organizational factors (ownership or type of practice), or area of medical care are associated with at least monthly VC provision (yes or no). Similarly, it was then examined whether these factors are related to the general interest in VC provision (yes or no). The aim is to test whether the predictors influence the probability of VC provision or interest in VC and to determine their odds. Multicollinearity diagnostics indicated no concerns for the predictor variables (all tolerance values >0.20; all variance inflation factor values <5). All independent variables were included in the analysis simultaneously in order to examine their contribution to predicting each dependent variable. The level of statistical significance was set at α=.05, with a *P* value of ≤.05 deemed significant. Statistical analyses were performed using the statistical software IBM SPSS Statistics version 28 (IBM Corp).

Results

Response Rate and Participants' Characteristics

The study questionnaire was sent to 34,095 physicians and psychotherapists, of whom 33,994 participants could be contacted. The remaining 101 questionnaires could not be delivered. The total number of returned questionnaires was 6112, resulting in a response rate of 17.9%. About 60.5% (3700/6112) of participants responded online and 39.5% (2412) responded by paper. Following the exclusion of 182 questionnaires that were either implausible or blank, a total of 5930 respondents were included in the analysis. They represent around 17.4% (5930/34,095) of physicians and psychotherapists in the four regions surveyed.

The study sample comprised a slight overrepresentation of women (3140 to 2378; 56.9% to 32.1%) and a slight underrepresentation of medical service providers in the older age groups (>60 y: 1359/5539, 24.5% to 10,261/34,095, 30.1%) compared to the basic population. Table 1 summarizes the sociodemographic characteristics and information on the medical profession.

XSL•FO
RenderX

Table . Characteristics and medical profession of survey participants and basic population.

	Participants, n (%)	Physicians and psychotherapists in selected regions, n (%)[a]
Total	5930 (100)	34,095 (100)
Gender (n=5518)		
Female	3140 (56.9)	17,452 (51.2)
Male	2378 (43.1)	16,643 (48.8)
Age groups (aggregated) (n=5539)		
<40 years	749 (13.5)	3538 (10.4)
40 - 50 years	1478 (26.7)	8317 (24.4)
51 - 60 years	1952 (35.2)	11,979 (35.1)
>60 years	1359 (24.5)	10,261 (30.1)
Community size of practice location (n=5525)		—[b]
Rural community	372 (6.7)	
Small town	965 (17.5)	
Middle town	1411 (25.5)	
Large city	2777 (50.3)	
Employment status[c] (n=5542)		—
Self-employed	4421 (79.8)	
Employed	1161 (20.9)	
Type of practice[c] (n=5542)		—
Individual practice	2864 (50.1),	
Joint practice[d]	1526 (26.7)	
Group practice[d]	807 (14.1)	
Medical care unit[d]	521 (9.1)	
Area of medical care (n=5531)		—
Primary care	1591 (28.8)	
Specialist care	1747 (31.6)	
Psychotherapeutic care	2193 (39.6)	

[a]Data of all 34,095 physicians and psychotherapists from the Associations of Statutory Health Insurance Physicians (ASHIPs) in the regions Berlin, Westphalia-Lippe, Mecklenburg-Western Pomerania, and Schleswig-Holstein. All specialist groups that are theoretically authorized to provide video consultation in Q4 2023 are included. Source: Data provided by the participating ASHIPs.

[b]Not available.

[c]Multiple answers possible.

[d]German terms: joint practice: Berufsausübungsgemeinschaft; group practice: Praxisgemeinschaft; medical care unit: Medizinisches Versorgungszentrum.

About half (2777/5525) of the participants have their practice location in a large city, 25.5% (1411/5525) run their practice in a medium-sized town, 17.5% (965/5525) in a small town, and 6.7% (372/5525) in the countryside. Most of the participants (4549/5542, 82.1%) are self-employed. Around half of the participants (2873/5542, 51.8%) run an individual practice, 27.6% (1530/5542) a joint practice, and 14.6% (811/5542) a group practice. Only 9.5% (529/5542) of the participants state that they are employed in a medical care unit. The participants also were asked to indicate their area of medical care. About 28.8% (1591/5531) of the participants practice in primary care, while a comparable proportion (1747/5531, 31.6%) works in specialist care. A larger proportion of participants (2193/5531, 39.6%) is engaged in psychotherapeutic/psychiatric care.

Most common medical care areas are primary care (1438/5484, 26.1%), nonphysician care providers in psychological psychotherapy (1293/5484, 23.6%), physicians in the field of psychotherapy, psychiatry, or neurology (464/5484, 8.5%), and psychotherapists for children and adolescents (458/5484, 8.4%). Other, less frequent areas of medical care in which the participants are trained in include gynecology, gynecological endocrinology, and reproductive medicine (315/5484, 5.7%), orthopedics and trauma surgery (182/5484, 3.3%), pediatrics

XSL•FO
RenderX

and adolescent medicine in primary care (153/5484, 2.8%), otorhinolaryngology (130/5484, 2.4%), and venereal diseases and dermatology (113/5484, 2.1%). Table 2 presents a summary of the (aggregated) medical specialties displayed in this survey in comparison to the basic population.

Table . Medical specialty (aggregated).

	Participants, n (%)	Physicians and psychotherapists in selected regions, n (%)[a]
Total	5484 (100)	34,104[b] (100)
Primary care (excluding pediatrics in primary care)	1438 (26.1)	10,736 (31.5)
Psychological psychotherapy	1293 (23.6)	5446 (16.0)
Psychotherapy/psychiatry/neurology (excluding psychological psychotherapists and psychotherapists for children/adolescents)	464 (8.5)	2389 (7.0)
Psychotherapists for children and adolescents	458 (8.4)	1590 (4.7)
Gynecology, gynecological endocrinology, and reproductive medicine	315 (5.7)	2394 (7.0)
Orthopedics and trauma surgery	182 (3.3)	1702 (5.0)
Pediatrics (in primary care)	153 (2.8)	1350 (4.0)
Otorhinolaryngology	130 (2.4)	898 (2.6)
Surgery (excluding orthopedics and trauma surgery)	128 (2.3)	863 (2.6)
Venereal diseases and dermatology	113 (2.0)	793 (2.3)
Pediatrics (specialist care)	44 (0.8)	144 (0.4)
Other	766 (13.2)	5799 (17.0)

[a]Data of all 34,095 physicians and psychotherapists from the Associations of Statutory Health Insurance Physicians (ASHIPs) in the regions Berlin, Westphalia-Lippe, Mecklenburg-Western Pomerania, and Schleswig-Holstein. All specialist groups that are theoretically authorized to provide video consultation in Q4 2023 are included. Source: Data provided by the participating ASHIPs.

[b]Since in a few cases more than one specialist group was assigned to a physician or psychotherapist in an ASHIP, the number in this table (34,104) differs slightly from the total number of respondents (34,095).

Provision of Video Consultations

Provision of VC

Among the physicians and psychotherapists who have already used VC, 78.3% (1722/2199) have offered them since the beginning of the COVID-19 pandemic; 11.2% (246/2199) started using VC later on during the pandemic (from 2021 onwards). However, only 10.2% (231/2199) have been providing VC since before the COVID-19 pandemic.

Of the physicians and psychotherapists, 37.8% (2216/5863) reported using VC at least once a month, with 36.6% using VC several times a week, 33.6% at least once a week, and 29.5% at least once a month. Meanwhile, 62.2% of participants (3647/5863) report that they have rarely or never offered VC. Significant associations with gender, age, community size of practice location, type of practice, ownership of practice, and area of medical care were identified with VC provision (see Multimedia Appendix 2).

The findings of the logistic regression analysis demonstrate that at least monthly VC provision is associated with each included predictor except community size of practice location. For the area of medical care, the odds ratio (OR) of at least monthly VC provision in psychotherapeutic care was about 8.2 compared to primary care (95% CI 7.4 - 1.64; P<.001), whereas in specialist care, the odds for monthly VC provision were approximately 50% lower than in primary care (OR 0.5, 95% CI 0.43-0.59; P<.001). Regarding medical specialty, within psychotherapeutic care, medical professionals who offer VC most frequently are psychological psychotherapists (985/1282, 76.8%), psychotherapists for children and adolescents (333/456, 73%), and psychotherapists, psychiatrists, and neurologists (245/457, 53.6%). At a considerable distance from the frequency observed in psychotherapeutic care, 22.7% (10/44) of pediatric specialists, 20% (285/1422) of physicians in primary care, and 19.6% (22/112) of specialists for venereal diseases and dermatology have stated to have VC experience.

Individual factors such as gender and age were also significant predictors for monthly VC provision. Female participants have higher odds to provide VC at least once a month (OR 1.163, 95% CI 1.01 - 1.34; P=.03). The odds for monthly VC provision in older age groups are approximately 60% higher than in the <40 years old age group (OR 0.41, 95% CI 0.32-0.52; P<.001). With regard to organizational factors, self-employed participants had about 50% higher odds of providing VC than employed participants (OR 0.480, 95% CI 0.39-0.59; P<.001). Regarding the type of practice, participants working in group practice (OR 1.33; 95% CI 1.07 - 1.66; P=.01) and medical care units (OR

XSL•FO
RenderX

1.39, 95% CI 1.04 - 1.85; *P*=.03) had significantly higher odds to provide VC at least once a month compared to those in individual practice. The model was statistically significant, χ^2_{13}=1937.52, *P*<.001, and explained approximately 41% according to Nagelkerke pseudo-R^2. Further, the Hosmer–Lemeshow test was nonsignificant, χ^2_8=12.32, *P*=.14, indicating good model fit.

Interest in VC Provision

The majority of the physicians and psychotherapists (4347/5442, 79.9%) express interest in potentially providing VC. There appears to be a significant association between interest in use by gender, age group, VC experience, and area of medical care (see Multimedia Appendix 3).

The patterns evident in current at least monthly VC provision are also reflected in the results of the logistic regression analysis of general interest in VC provision. Interest in providing VC is significantly associated with gender, age group, practice location, type of practice, and area of medical care. For instance, logistic regression indicated a significant association between age group and the likelihood of interest in VC provision (*P*<.001). Compared to the reference group of physicians and psychotherapists under 40 years, those aged 40 - 50 years had 51% lower odds of the outcome (OR 0.49, 95% CI 0.34-0.69; *P*<.001). The 50 - 60 years age group also exhibited lower odds compared to the reference group (OR 0.3, 95% CI 0.21-0.42; *P*<.001). The strongest effect was observed among individuals aged over 60 years, who had approximately 80% lower likelihood of interest in VC provision (OR 0.2, 95% CI 0.15-0.28; *P*<.001). The overall model for interest in VC provision was statistically significant, χ^2_{13}=574.23, *P*<.001, indicating that the set of predictors reliably distinguished interest in use of VC provision. The model explained approximately 17% according to Nagelkerke pseudo-R^2. The Hosmer–Lemeshow goodness-of-fit test was nonsignificant, χ^2_8=13.77, *P*=.09, suggesting acceptable model fit. Table 3 presents details on the results of the logistic regression models.

XSL•FO
RenderX

Table . Binary logistic regression for at least monthly VC[a] provision and interest in VC provision.

Predictor	At least monthly VC provision[b] (n=5232)		Interest in VC provision [c] (n=5173)	
	P value	OR (95% CI)	*P* value	OR (95% CI)
Gender (reference: male)				
Female	.03	1.163 (1.01-1.34)	.03	1.315 (1.13-1.53)
Age group (reference: <40 y)				
40 - 50 y	<.005	0.711 (0.56-0.90)	<.001	0.486 (0.34-0.69)
51 - 60 y	<.008	0.729 (0.58-0.9)	<.001	0.300 (0.21-0.42)
>60 y	<.001	0.409 (0.32-0.52)	<.001	0.204 (0.15-0.28)
Community size of practice location (reference: rural community)				
Small town	.10	0.783 (0.59-1.05)	.79	0.958 (0.7-1.31)
Middle town	.20	0.833 (0.63-1.10)	.96	0.993 (0.73-1.35)
Large city	.38	1.128 (0.864-1.47)	.21	1.210 (0.9-1.63)
Ownership of practice (reference: self-employed)				
Employed	<.001	0.480 (0.39-0.59)	.30	0.893 (0.72-1.11)
Type of practice (reference: individual practice)				
Joint practice	.05	1.175 (0.999-1.38)	<.001	1.560 (1.31-1.85)
Group practice	.01	1.331 (1.067-1.66)	<.001	1.385 (1.08-1.78)
Medical care unit	.03	1.387 (1.039-1.85)	.01	2.713 (1.96-3.76)
Area of medical care (reference: primary care)				
Specialist care	<.001	0.503 (0.425-0.59)	.01	0.807 (0.68-0.96)
Psychotherapeutic care	<.001	8.917 (7.472-10.64)	<.001	3.850 (3.11-4.77)
Constant	0.33	0.845	<.001	5.606

[a]VC: video consultation.

[b]At least monthly VC provision: pseudo-R^2=0.41.

[c]Interest in VC provision: pseudo-R^2=0.17.

Exclusive Use of VC

Almost half (2702/5651, 47.8%) of the physicians and psychotherapists surveyed can imagine treating patients exclusively via VC without further face-to-face appointments for the treatment case in the same quarter. There are differing attitudes toward the exclusive use of VC, depending on whether the participants have VC experience (χ^2_1=320.997, *P*<.001; Cramer *V*=0.239), their age group (χ^2_3=126.589, *P*<.001; Cramer *V*=0.152), or their medical specialty (χ^2_{11}=146.092, *P*<.001; Cramer *V*=0.164). Physicians and psychotherapists with VC experience (1345/2676, 63%) are more likely to imagine treating patients in a treatment case exclusively via VC than participants who have not or rarely provided VC before. Further, medical specialists with specialist training in venereal diseases and dermatology (80/113, 70.8%), GPs (720/1423, 50.6%), nonphysician care providers in psychological psychotherapy (721/1285, 56.1%), and physicians who specialized in psychotherapy, psychiatry, and neurology (232/459, 50.5%) are more likely to consider treating patients in a treatment case exclusively via VC.

Suitable Medical Fields

Regarding suitable medical fields, the participants were asked if parts of their treatment could be carried out via VC for the enquired medical fields. Results only include participants with

XSL•FO
RenderX

previous VC experience, as it is assumed that a more precise assessment can be made. The participants indicated a potential for using VC, in particular, in the area of mental and behavioral disorders. In the field of mental disorders, nonorganic sleep disorders (very suitable or suitable: 1296/1995, 65.0%), affective disorders (very suitable or suitable: 1123/1991, 56.4%), anxiety disorders (very suitable or suitable: 1110/2014, 55.1%), sexual dysfunctions (very suitable or suitable: 952/1836, 51.3%), or obsessive-compulsive disorders (very suitable or suitable: 952/1974, 48.2%) are considered to be very suitable or suitable most often. An exception is cases of schizophrenia (highly unsuitable or unsuitable: 1047/1764, 59.4%) as well as schizotypal and delusional disorders caused by psychotropic substances (highly unsuitable or unsuitable: 824/1837, 44.9%), which are rather reported as highly unsuitable or suitable for VC.

Among chronic illnesses, the treatment of chronic pain (very suitable or suitable: 936/1517, 61.7%), metabolic disorders (very suitable or suitable: 464/1009, 46.0%), and allergies (very suitable or suitable: 413/1020, 40.5%) as well as dermatological diseases (very suitable or suitable: 368/989, 37.2%) are considered very suitable or suitable for VC treatment.

Acute illnesses are rather considered highly unsuitable or unsuitable, with the exception of acute headaches (to some extent suitable: 481/1238, 38.9%) or dermatoses (to some extent suitable: 337/922, 36.6%), which are considered suitable to some extent. Further details on the attitude of suitable medical fields from the perspective of physicians and psychiatrists with VC experience can be found in Multimedia Appendix 4.

Types of Treatment

To assess potential suitable types of treatment, only findings of participants with VC experience are displayed as a better understanding is assumed. Significant associations have been identified by participants who have experience in VC provision for all queried types of treatment. They are more likely to consider the types of treatment as suitable for VC provision than participants with little or no experience; for details on their associations and effect sizes, see Multimedia Appendix 5.

According to over 60% of these physicians and psychotherapists, discussing test results (1422/1896, 75%), issuing of prescriptions for drugs and remedies (793/1204, 65.9%), and issuing of incapacity certificates (677/1042, 65.0%) and treatment planning (1379/2131, 64.7%) are very suitable or suitable types of treatment via VC (see Table 4). Follow-up appointments (771/1350, 57.1%) and taking a patient's medical history (1195/2147, 55.7%) are considered very suitable or suitable by over half of the participants. Diagnostic procedures appear to be the least suitable form of treatment for VC, with only 24.6% of participants (very suitable or suitable: 523/2126), indicating agreement with these types of treatment.

Table . Attitudes of physicians and psychotherapists with VC[a] experience towards the suitability of types of treatment for VC.

	Highly unsuitable, n (%)	Unsuitable, n (%)	To some extent, n (%)	Suitable, n (%)	Very suitable, n (%)
Discussion of test results (n=1896)	29 (1.5)	101 (5.3)	344 (18.1)	706 (37.2)	716 (37.8)
Medical history (n=2147)	85 (4.0)	343 (16.0)	524 (24.4)	633 (29.5)	562 (26.2)
Issuing prescriptions for drugs and remedies (n=1204)	49 (4.1)	136 (11.3)	226 (18.8)	429 (35.6)	364 (30.2)
Issuing incapacity certificate (n=1042)	44 (4.2)	100 (9.6)	221 (21.2)	366 (35.1)	311 (29.8)
Therapy treatment planning (n=2131)	35 (1.6)	172 (8.1)	545 (25.6)	919 (43.1)	460 (21.6)
Individual psychiatric/psychotherapeutic consultations (n=2017)	27 (1.3)	122 (6.0)	546 (27.1)	743 (36.8)	579 (28.7)
Follow-up checks (eg, wound healing, medication) (n=1350)	34 (2.5)	130 (9.6)	415 (30.7)	485 (35.9)	286 (21.2)
Group sessions (eg, in psychotherapy) (n=977)	178 (18.2)	381 (39.0)	256 (26.2)	105 (10.7)	57 (5.8)
(Further) diagnostic work-up (n=2126)	151 (7.1)	669 (31.5)	783 (36.8)	375 (17.6)	148 (7.0)

[a]VC: video consultation.

Regarding area of medical care of participants with VC experience, significant associations have been identified for taking a patient's medical history (χ^2_8=21.266, P<.001; Cramer V=0.229), further diagnostic work-up (χ^2_8=85.333, P<.001; Cramer V=0.145), issuing an incapacity certificate (χ^2_8=36.051, P<.001; Cramer V=0.135), the discussion of test results

XSL•FO
RenderX

(χ^2_8=164.154, *P*<.001; Cramer *V*=0.213), follow-up checks (χ^2_8=71.743, *P*<.001; Cramer *V*=0.167), and individual psychiatric or psychotherapeutic consultations (χ^2_8=127.499, *P*<.001; Cramer *V*=0.182). Taking a patient's medical history via VC is deemed very suitable or suitable rather by providers in primary care (very suitable or suitable: 259/308, 84.1%) and specialist care (very suitable or suitable: 141/198, 71.2%), less than by providers in psychotherapeutic care (very suitable or suitable: 739/1544, 47.9%). It is similar in the case of discussing test results, 91.2% (281/308) of GPs and 90.2% (175/194) of specialists report them to be a very suitable or suitable option via VC, while 68.6% (895/1305) of psychotherapists consider them very suitable or suitable. Conversely, the attitudes of the suitability of (further) diagnostics are the opposite. Although (further) diagnostics are deemed highly unsuitable or suitable, health care professionals engaged in the delivery of psychotherapeutic care (very suitable or suitable: 441/1526, 28.9%) are more likely to view them as suitable than GPs (very suitable or suitable: 37/303, 12.2%) and specialists (very suitable or suitable: 19/201, 9.5%). For follow-up checks, 64.5% (517/802) of psychotherapists and about half of specialists (98/184, 53.3%) consider these very suitable or suitable for VC, whereas only 38.3% (116/303) of GPs agree on their suitability. In particular, in the context of individual psychotherapeutic consultations, psychotherapists (very suitable or suitable: 1079/1569, 68.8%) consider these to be suitable for VC, and also approximately 50% of GPs (very suitable or suitable: 122/241, 50.6%) and specialist care (very suitable or suitable: 55/111, 49.5%) hold this view. For further associations or correlations and effect sizes, see Multimedia Appendix 6.

Discussion

Principal Findings

This study examined the attitude toward the use of VC in German outpatient care after the COVID-19 pandemic from a provider perspective. For the majority of the surveyed physicians and psychotherapists, the advent of the COVID-19 pandemic served as a catalyst, with approximately 80% commencing VC use during the initial phase of the pandemic since 2020, particularly within the area of psychotherapeutic care. VC provision is reported more often in psychotherapeutic care (73%) and still less often in primary (20%) and specialist care (12%). Despite over half of service providers reporting little or no VC provision, with approximately 40% stating otherwise, this represents a notable increase in providers who have used VC so far compared to the prepandemic period. However, the survey results do not reveal whether they have incorporated VC into their daily routine or whether it is merely a sporadic occurrence necessitated by the circumstances during the pandemic. Current claims data analyses indicate that VC provision is not regular, as use has experienced a decline once again following the conclusion of the COVID-19 pandemic [7].

According to our findings, the odds of VC provision are higher among female physicians and psychotherapists, younger age groups, and those living in urban areas. These results are coherent with previously published research on German claims data done earlier in the study by Hüer et al [6] covering the time period during the COVID-19 pandemic. Therefore, there seem to be no changes in the analyzed factors associated with VC provision during and after the COVID-19 pandemic. In terms of organizational factors, practice types with shared facilities demonstrated higher odds for VC provision than participants in an individual practice setting. During the pandemic, Hüer et al [6] and Knörr et al [10] did not observe a higher frequency of use by group practices. However, Knörr et al [10] noted that the perceived benefit is greater in group practices. Recent systematic reviews emphasized the importance of perceived usefulness and ease of use in the adoption of telemedicine [25,26]. The findings of this study suggest that shared facilities may result in a reduction of organizational barriers to VC use and offer enhanced practical benefits or ease of use.

Conversely, self-employed physicians also demonstrated higher odds for VC provision. The current literature suggests that the barriers are particularly related to the expected usefulness within their professional environment, making the indications inconclusive.

Globally, many countries have integrated telehealth into routine care to some extent, although the degree varies widely [27]. The 2022 International Health Policy Survey of General Practitioners, conducted by the Commonwealth Fund in 10 high-income countries, examined opinions on the user-friendliness and effectiveness of telemedicine in the period following the peak of the pandemic. In the majority of countries, physicians reported a number of advantages, including financial compensation, improved timeliness of care, and enhanced ability for mental health care. German physicians demonstrated the lowest levels of satisfaction across the majority of stated metrics, while medical professionals in the United Kingdom, Australia, Canada, and the United States reported higher levels of satisfaction and ease of implementation of the telehealth platform. In Germany, telehealth expanded more slowly postpandemic, with only about 30% - 40% of German primary care doctors stating telehealth to be easy to implement, citing cost of the platform as a barrier [28]. This indicates that besides readiness of users and providers, national policies and technical infrastructure have an influence on VC adoption [16,28]. Nevertheless, the results suggest that there is potential to establish use among medical service providers, as approximately 80% expressed interest in potentially using VC. With younger providers in particular using and being interested in using VC, it is likely that an age-related dip in VC provision will diminish in the future. In addition, these results can also help identify subgroups that might perceive more barriers in VC adoption. Elder and male physicians and psychotherapists showed less odds of VC adoption. Research suggests that those groups show less readiness and might lack familiarity or trust in new technologies [26,29]. For the time being, it is recommended that efforts are made to facilitate use [26], with a particular focus on engaging older age groups, as even around 70% of physicians and psychotherapists with little or no VC experience reported interest in using VC. This objective could be realized through the implementation of knowledge pertaining to the advantages of VC in the obligatory Continuing Medical Education administered by the German Medical Associations.

XSL•FO
RenderX

Although awareness of VC has been heightened by the COVID-19 pandemic and there is considerable interest in its use, VC still appears to be used irregularly at present [7], suggesting that there may currently be inhibiting factors inhibiting the expansion of VC provision (eg, concerns about a deterioration in the physician–patient relationship [13] or factors relating to the health care system [16]). This issue merits a need for further research to investigate the reasons for the current reluctance to use VC. In order to explore why the uptake of VC has not been sustainable and to develop strategies that can be used to ensure its long-term integration, a look at hindering factors, particularly in the German health care system, or qualitative research on sustainable implementation strategies would also be an appropriate approach.

Globally, many countries have integrated telehealth into routine care to some extent, although the degree varies widely [27]. The 2022 International Health Policy Survey of General Practitioners, conducted by the Commonwealth Fund in 10 high-income countries, examined opinions on the user-friendliness and effectiveness of telemedicine in the period following the peak of the pandemic. In the majority of countries, physicians reported a number of advantages, including financial compensation, improved timeliness of care, and enhanced ability for mental health care. German physicians demonstrated the lowest levels of satisfaction across the majority of stated metrics, while medical professionals in the United Kingdom, Australia, Canada, and the United States reported higher levels of satisfaction and ease of implementation of the telehealth platform. In Germany, telehealth expanded more slowly postpandemic, with only about 30% - 40% of German primary care doctors stating telehealth to be easy to implement, citing cost of the platform as a barrier [28]. This indicates that besides readiness of users and providers, national policies and technical infrastructure have an influence on VC adoption [16,28]. Nevertheless, the results suggest that there is potential to establish use among medical service providers, as approximately 80% expressed interest in potentially using VC. With younger providers in particular using and being interested in using VC, it is likely that an age-related dip in VC provision will diminish in the future. In addition, these results can also help identify subgroups that might perceive more barriers in VC adoption. Elder and male physicians and psychotherapists showed less odds of VC adoption. Research suggests that those groups show less readiness and might lack familiarity or trust in new technologies [26,29]. For the time being, it is recommended that efforts are made to facilitate use [26], with a particular focus on engaging older age groups, as even around 70% of physicians and psychotherapists with little or no VC experience reported interest in using VC. This objective could be realized through the implementation of knowledge pertaining to the advantages of VC in the obligatory Continuing Medical Education administered by the German Medical Associations. In light of the heightened awareness of VC brought on by the COVID-19 pandemic and given the great interest in the use of VC, VCs still seem to be used irregularly [7], which suggests that there might currently be inhibiting factors that impede the expansion of VC provision (eg, fear of deterioration in the physician–patient relationship [13] or health system factors [16]). This issue merits a need for further research to investigate the reasons for the current reluctance to use VC. In order to explore why the uptake of VC has not been sustainable and to develop strategies that can be used to ensure its long-term integration, a look at hindering factors, particularly in the German health care system, or qualitative research on sustainable implementation strategies would also be an appropriate approach.

Regarding VC as an add-on service or as a service without face-to-face patient contact in the same quarter for the treatment case, there is no clear agreement or rejection in favor of its exclusive use. According to the findings of the study, a slightly below-average number of respondents were in favor of exclusive use. However, it should be noted that in Germany, discounts in honorarium are applied when VC is used exclusively during a given billing quarter [4]. Hence, the attitude of suitability of exclusive VC use may be biased due to the German billing regulations. Nevertheless, the findings suggest that exclusive use may be more appropriate for certain medical specialties. Physicians in venereal diseases and dermatology, as well as nonphysician care providers in psychological psychotherapy, demonstrated a greater tendency to consider the exclusive use of VC.

This research suggests that VC may be particularly suited to the care of patients with mental and behavioral health problems [30]. However, there are limitations for certain diagnoses, such as schizophrenia, schizotypal, and delusional disorders caused by psychotropic substances. This assessment has been widely acknowledged in relevant literature and is accompanied by an increase in the use of VC and perceived suitability by providers of psychotherapeutic care [31,32]. Furthermore, the results suggest that VCs appear to be suitable for selected chronic illnesses. Current research also validates this for selected indications, for example, for diabetes [33], obstructive sleep apnea [34], or dermatological diseases such as atopic dermatitis [35].

Regarding the findings on types of treatment, there are indications that physicians and psychotherapists consider VC to be particularly suitable to those types of treatment which are mainly based on conversations (eg, discussing test results or taking the patient's medical history, therapy treatment planning) or which require a higher degree of administrative work (eg, issuing of prescriptions for drugs and remedies and issuing of incapacity certificates). This may particularly apply to well-known patients and those with long-term conditions [12,13]. The results of our study, as well as existing research, indicate that diagnostics may not be the optimal fit for this particular modality of treatment in primary and specialist care. For psychotherapeutic care, further diagnostic workup has been rated more positively, which is supported by another German survey [8]. During the COVID-19 pandemic, follow-up checks were reported to be a well-accepted type of treatment [12,36,37]. According to this study, after the COVID-19 pandemic, this was not fully confirmed, with only 33% finding it suitable and 36% suitable to some extent. The variability of the suitability of follow-up checks seems to be dependent on different treatments. In the area of psychotherapeutic care, follow-up visits are more likely to be considered appropriate than in primary care or specialist care. Moreover, differences are

XSL•FO
RenderX

observed with regard to the presence or absence of VC experience. This suggests that as a result of the adoption of VC by physicians and psychotherapists, attitudes have shifted toward a more positive outlook. It can be reasonably concluded that training programs may assist in reducing the current deficit of knowledge and in fostering greater confidence in the provision of VC.

Limitations

The study achieves a satisfactory response rate of around 18% and thus results in a large sample size of 5930, which is typically challenging to obtain when studying physicians or psychotherapists. This was made possible due to the involvement of ASHIPs in contacting all of the physicians in their respective regions. However, it is important to note that the participants do not fully represent the original population. The sample comprises a slightly higher proportion of female participants and a slightly higher proportion of individuals under the age of 40 than is observed in the basic population. The mode of data collection did not influence the variables related to the main findings. However, differences were observed in demographic characteristics such as the age subgroup between participants recruited online and those surveyed via paper-pencil. The option to participate on paper next to online via a QR code was incorporated with the objective of mitigating the exclusion of offline participants, a demographic which is predominantly composed of older individuals. Moreover, it should be acknowledged that there is a considerably greater number of physicians and medical service providers from the psychotherapeutic care area than represented in the surveyed basic population. This could potentially introduce a selection bias, as psychotherapists are more likely to use VC and may therefore have responded differently from the average due to their vested interest in VC. Psychotherapists may perceive VC as being congruent with their professional practice, particularly with regard to continuity of care and emergency situations [1,2,30,38]. Conversely, some medical practitioners, particularly those specializing in physical disciplines, may regard VC as being incongruent with the necessity for physical examinations [39]. Subsequently, the results may lead to more positive attitudes toward VC, which could restrict the generalizability of the findings to other outpatient providers. Subgroup analysis differentiated according to the areas of primary care, and specialist and psychotherapeutic cares was done in order to be able to make a more precise distinction.

While the rollout of VCs has been comprehensively researched in many countries since the start of the pandemic, there has been limited investigation into this for the German outpatient sector up to now. This study, therefore, adds value by taking a closer and more detailed look at the situation in outpatient care in Germany. Yet, physicians and psychiatrists, as well as psychotherapists working exclusively in the private sector, were not included. However, as approximately 90% of insured persons in Germany have statutory health insurance and only members of ASHIPs are permitted to bill patients with statutory health insurance, the relevance of exclusively private practitioners is negligible. Nevertheless, since only German physicians and psychotherapists were surveyed, the transferability of the results to other countries with different regulatory systems may be limited. Further, studies could explore the use of VC in the inpatient or care settings where there may also be potential [40,41]. Additionally, all specific medical specialties in the field of psychiatric/neurological and psychotherapeutic care are referred to as “psychotherapists” in this study. It is important to note that there are differences in practicing depending on specific medical specialty. Further research could go into depth according to specific medical specialties.

While the logistic regression models provided useful insights, several limitations should be noted. First, as this was a cross-sectional study, the associations identified cannot be interpreted causally. The logistic regression model was employed for the purpose of exploratory analysis, with the objective of ascertaining which variables exert a significant effect when controlling for the other variables. Second, although the first model showed substantial explanatory power (Nagelkerke R^2=0.41) and the second model moderate power (Nagelkerke R^2=0.17), a certain proportion of variance remains unexplained, suggesting that additional factors not included in the analysis may influence the outcome. These might include digital maturity, technical infrastructure, and remuneration terms [42-44].

As this was an exploratory cross-sectional study, a self-constructed questionnaire was used. Although no full psychometric validation was conducted, the instrument was informed by prior qualitative analyses and pretesting to ensure content relevance and comprehensibility. This limits the extent to which the results can be generalized, and the approach is appropriate for an exploratory design and provides valuable initial insights to guide future research. Furthermore, systematic errors cannot be completely excluded. In this study, the use of batteries of items (eg, Likert scales) may have led to a tendency for the questions to be averaged or not to be differentiated. Concerning measures of association or correlation, the calculated effect sizes are according to Cohen value categorized as weak and a few as moderate.

Conclusion

Although the actual use of VC in general practice and specialist care is still relatively low, most physicians and psychotherapists are in favor of their use after the COVID-19 pandemic. A wide range of medical indications is considered to be suitable, particularly in the area of psychotherapeutic care and use for chronic illnesses. The findings indicate that German service providers do not appear to be averse to the use of VC as they show high interest in their provision. Potential barriers to the exclusive use of VC may include factors such as remuneration or the medical specialty of the practitioner. Furthermore, in the case of certain medical specialties, treatment, particularly diagnosis, cannot be provided without a physical examination. It is advisable to incorporate the providers’ perspectives into the ongoing refinement of VC policies and practices, as these insights are important for ensuring the successful implementation of VC as a therapeutic modality. This research on physicians’ and psychotherapists’ attitudes adds to a baseline knowledge of different attitudes by providers and can provide grounds to create a shared vision, which has shown to be a driver

XSL•FO
RenderX

in the adoption process of digital technology [45]. Besides, the attitude and conduct of medical professionals play a fundamental role in ensuring that patients experience a sufficient level of comfort during the course of their treatment [46]. Given that physicians and psychotherapists are more likely to endorse video-based treatment modalities when they have prior experience with them, expanding training programs may help reduce knowledge gaps, alleviate uncertainty, and foster greater confidence in their use in outpatient care. The expansion of the digital infrastructure in the future, which facilitates the dissemination of information, has the potential to diminish potential barriers. Nevertheless, the results do not provide insights into the hurdles that exist in daily use and how exactly their use can be promoted. Further research is necessary to ascertain the full implications.

Funding

This study was entirely funded by the Innovation Fund of the German Federal Joint Committee (number 01VSF20011). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

Primary data of the survey are not publicly available due to data protection reasons.

Authors' Contributions

All authors contributed to the study organization and conception of the surveys. LK was responsible for the survey analysis and was the major contributor in writing this manuscript. The analysis and the manuscript draft were critically revised by TH. All authors contributed to the manuscript in different stages and also read and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Excerpt of the survey.
[DOCX File, 63 KB - jmir_v28i1e73757_app1.docx]

Multimedia Appendix 2
Subgroup analysis for provision of video consultations.
[DOCX File, 17 KB - jmir_v28i1e73757_app2.docx]

Multimedia Appendix 3
Subgroup analysis for interest in video consultation provision.
[DOCX File, 16 KB - jmir_v28i1e73757_app3.docx]

Multimedia Appendix 4
Attitudes on suitable medical fields (only for participants with video consultation experience).
[DOCX File, 22 KB - jmir_v28i1e73757_app4.docx]

Multimedia Appendix 5
Association and effect size of suitable types of treatment for participants with video consultation experience.
[DOCX File, 15 KB - jmir_v28i1e73757_app5.docx]

Multimedia Appendix 6
Association/correlation and effect size of suitable types of treatment (only for participants with video consultation experience).
[DOCX File, 19 KB - jmir_v28i1e73757_app6.docx]

References

1. Zangani C, Ostinelli EG, Smith KA, et al. Impact of the COVID-19 pandemic on the global delivery of mental health services and telemental health: systematic review. JMIR Ment Health 2022 Aug 22;9(8):e38600. [doi: 10.2196/38600] [Medline: 35994310]
2. Mann DM, Chen J, Chunara R, Testa PA, Nov O. COVID-19 transforms health care through telemedicine: evidence from the field. J Am Med Inform Assoc 2020 Jul 1;27(7):1132-1135. [doi: 10.1093/jamia/ocaa072] [Medline: 32324855]

XSL•FO
RenderX

3. Stein DJ, Naslund JA, Bantjes J. COVID-19 and the global acceleration of digital psychiatry. Lancet Psychiatry 2022 Jan;9(1):8-9. [doi: 10.1016/S2215-0366(21)00474-0] [Medline: 34921798]
4. Bericht des Bewertungsausschusses und des ergänzenden Bewertungsausschusses zur telemedizinischen Leistungserbringung im Einheitlichen Bewertungsmaßstab [Report in German]. : Deutscher Bundestag; 2022 URL: https://dserver.bundestag.de/btd/20/049/2004982.pdf [accessed 2025-11-28]
5. Smith AC, Thomas E, Snoswell CL, et al. Telehealth for global emergencies: implications for coronavirus disease 2019 (COVID-19). J Telemed Telecare 2020 Jun;26(5):309-313. [doi: 10.1177/1357633X20916567] [Medline: 32196391]
6. Hüer T, Walendzik A, Kleinschmidt L, et al. Use of video consultation between 2017 and 2020 in outpatient medical care in Germany and characteristics of their user groups: analysis of claims data. JMIR Form Res 2025 Mar 14;9:e60170. [doi: 10.2196/60170] [Medline: 40085136]
7. Mangiapane S. Fünf jahre videosprechstunde in der vertragsärztlichen versorgung: hat die pandemie zum durchbruch geführt [Book in German]. In: Repschläger U, Schulte C, Osterkamp N, editors. Gesundheitswesen Aktuell 2023 Beiträge Und Analysen: Barmer Institut fur Gesundheitssystemforschung; 2023:82-99 URL: https://www.bifg.de/media/dl/gesundheitswesen-aktuell/2023/gwa-2023_mangiapane.pdf [accessed 2025-11-28]
8. Albrecht M, Otten M, Bernhard J. PraxisBarometer digitalisierung 2022: befragung von vertragsärztinnen/-ärzten und vertrags [Report in German]. : iGES; 2022 URL: https://www.iges.com/sites/igesgroup/iges.de/myzms/content/e6/e1621/e10211/e27603/e27639/e27640/e27642/attr_objs27737/IGES_PraxisBaroDigit2021_Kurzbericht_ger.pdf [accessed 2025-11-28]
9. Mangiapane S, Kretschmann J, Czihal T, Stillfired DV. Zi-trendreport zur vertragsärztlichen versorgung: bundesweiter tabellarischer report vom 1 [Report in German]. : Zentralinstitut kassenarztliche Versorgung; 2025 URL: https://www.zi.de/fileadmin/Downloads/Service/Publikationen/Zi-Trendreport_2023-Q2.pdf [accessed 2024-07-22]
10. Knörr V, Dini L, Gunkel S, et al. Use of telemedicine in the outpatient sector during the COVID-19 pandemic: a cross-sectional survey of German physicians. BMC Prim Care 2022 Apr 23;23(1):92. [doi: 10.1186/s12875-022-01699-7] [Medline: 35461212]
11. Kane CK, Gillis K. The use of telemedicine by physicians: still the exception rather than the rule. Health Aff 2018 Dec;37(12):1923-1930. [doi: 10.1377/hlthaff.2018.05077] [Medline: 30633670]
12. Ward K, Vagholkar S, Sakur F, Khatri NN, Lau AYS. Visit types in primary care with telehealth use during the COVID-19 pandemic: systematic review. JMIR Med Inform 2022 Nov 28;10(11):e40469. [doi: 10.2196/40469] [Medline: 36265039]
13. Wanderås MR, Abildsnes E, Thygesen E, Martinez SG. Video consultation in general practice: a scoping review on use, experiences, and clinical decisions. BMC Health Serv Res 2023 Mar 30;23(1):316. [doi: 10.1186/s12913-023-09309-7] [Medline: 36997997]
14. Saeed S, Singhal M, Kaur KN, et al. Acceptability and satisfaction of patients and providers with telemedicine during the COVID-19 pandemic: a systematic review. Cureus 2024 Mar;16(3):e56308. [doi: 10.7759/cureus.56308] [Medline: 38628988]
15. Hoff T, Lee DR. Physician satisfaction with telehealth: a systematic review and agenda for future research. Qual Manag Health Care 2022;31(3):160-169. [doi: 10.1097/QMH.0000000000000359] [Medline: 35132008]
16. Assing Hvidt E, Atherton H, Keuper J, et al. Low adoption of video consultations in post-COVID-19 general practice in Northern Europe: barriers to use and potential action points. J Med Internet Res 2023 May 22;25:e47173. [doi: 10.2196/47173] [Medline: 37213196]
17. Tyre MJ, Orlikowski WJ. Windows of opportunity: temporal patterns of technological adaptation in organizations. Organ Sci 1994 Feb;5(1):98-118. [doi: 10.1287/orsc.5.1.98]
18. Fahy N, Williams GA, Habicht T, et al. Use of Digital Health Tools in Europe: Before, During and After COVID-19: European Observatory on Health Systems and Policies; 2021. [Medline: 35099866]
19. Kleinschmidt L, Walendzik A, Wasem J, et al. Preference-based implementation of video consultations in urban and rural regions in outpatient care in Germany: protocol for a mixed methods study. JMIR Res Protoc 2024 Apr 11;13:e50932. [doi: 10.2196/50932] [Medline: 38602749]
20. Häder M. Empirische Sozialforschung [Book in German]: Springer Fachmedien Wiesbaden; 2019.
21. Milbert A, editor. Raumabgrenzungen und Raumtypen des BBSR [Book in German]: Bonn; 2012:111.
22. Stuart A. The estimation and comparison of strengths of association in contingency tables. Biometrika 1953 Jun;40(1/2):105. [doi: 10.2307/2333101]
23. Cramér H. Mathematical Methods of Statistics (PMS-9): Princeton University Press; 2016:1593.
24. Cohen J. In: Hillsdale NJ, editor. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Erlbaum; 1988:567. [doi: 10.4324/9780203771587]
25. Garavand A, Aslani N, Nadri H, Abedini S, Dehghan S. Acceptance of telemedicine technology among physicians: a systematic review. Inform Med Unlocked 2022;30:100943. [doi: 10.1016/j.imu.2022.100943]
26. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023 Sep 18;6(1):161. [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
27. Ndwabe H, Basu A, Mohammed J. Post pandemic analysis on comprehensive utilization of telehealth and telemedicine. Clinical eHealth 2024 Dec;7:5-14. [doi: 10.1016/j.ceh.2023.12.002]

XSL•FO
RenderX

28. Gunja M, Gumas ED, Williams R. How primary care physicians experience telehealth: an international comparison—findings from the 2022 international health policy survey of primary care physicians. : The Commonwealth Fund; 2023 URL: https://tinyurl.com/3pxu6chz [accessed 2025-11-27]
29. Weik L, Fehring L, Mortsiefer A, Meister S. Understanding inherent influencing factors to digital health adoption in general practices through a mixed-methods analysis. NPJ Digit Med 2024 Feb 27;7(1):47. [doi: 10.1038/s41746-024-01049-0] [Medline: 38413767]
30. Appleton R, Williams J, Vera San Juan N, et al. Implementation, adoption, and perceptions of telemental health during the COVID-19 pandemic: systematic review. J Med Internet Res 2021 Dec 9;23(12):e31746. [doi: 10.2196/31746] [Medline: 34709179]
31. Shaker AA, Simonsen E, Tarp K, et al. Capturing patients' and clinicians' experiences of using video consultations in mental health outpatient services: qualitative thematic analysis. JMIR Form Res 2024 Aug 21;8:e50580. [doi: 10.2196/50580] [Medline: 39167796]
32. Shaker AA, Austin SF, Storebø OJ, et al. Psychiatric treatment conducted via telemedicine versus in-person modality in posttraumatic stress disorder, mood disorders, and anxiety disorders: systematic review and meta-analysis. JMIR Ment Health 2023 Jul 5;10:e44790. [doi: 10.2196/44790] [Medline: 37277113]
33. Sotomayor F, Hernandez R, Malek R, Parimi N, Spanakis EK. The effect of telemedicine in glycemic control in adult patients with diabetes during the COVID-19 era - a systematic review. J Clin Med 2023 Aug 31;12(17):5673. [doi: 10.3390/jcm12175673] [Medline: 37685740]
34. Alsaif SS, Kelly JL, Little S, et al. Virtual consultations for patients with obstructive sleep apnoea: a systematic review and meta-analysis. Eur Respir Rev 2022 Dec 31;31(166):220180. [doi: 10.1183/16000617.0180-2022] [Medline: 36517048]
35. Verma L, Turk T, Dennett L, Dytoc M. Teledermatology in atopic dermatitis: a systematic review. J Cutan Med Surg 2024;28(2):153-157. [doi: 10.1177/12034754231223694] [Medline: 38205736]
36. Richter JG, Chehab G, Reiter J, et al. Evaluation of the use of video consultation in German rheumatology care before and during the COVID-19 pandemic. Front Med (Lausanne) 2022;9:1052055. [doi: 10.3389/fmed.2022.1052055] [Medline: 36507506]
37. Johnsen TM, Norberg BL, Kristiansen E, et al. Suitability of video consultations during the COVID-19 pandemic lockdown: cross-sectional survey among Norwegian general practitioners. J Med Internet Res 2021 Feb 8;23(2):e26433. [doi: 10.2196/26433] [Medline: 33465037]
38. Galvin E, Desselle S, Gavin B, et al. Stakeholder perspectives and experiences of the implementation of remote mental health consultations during the COVID-19 pandemic: a qualitative study. BMC Health Serv Res 2023 Jun 13;23(1):623. [doi: 10.1186/s12913-023-09529-x] [Medline: 37312119]
39. Kulkarni AJ, Thiagarajan AB, Skolarus TA, Krein SL, Ellimoottil C. Attitudes and barriers toward video visits in surgical care: insights from a nationwide survey among surgeons. Surgery 2024 Jul;176(1):115-123. [doi: 10.1016/j.surg.2024.03.033] [Medline: 38734503]
40. Asiri A, AlBishi S, AlMadani W, ElMetwally A, Househ M. The use of telemedicine in surgical care: a systematic review. Acta Inform Med 2018 Oct;26(3):201-206. [doi: 10.5455/aim.2018.26.201-206] [Medline: 30515013]
41. May S, Jonas K, Fehler GV, Zahn T, Heinze M, Muehlensiepen F. Challenges in current nursing home care in rural Germany and how they can be reduced by telehealth - an exploratory qualitative pre-post study. BMC Health Serv Res 2021 Sep 6;21(1):925. [doi: 10.1186/s12913-021-06950-y] [Medline: 34488746]
42. Kerr G, Greenfield G, Li E, et al. Factors associated with the availability of virtual consultations in primary care across 20 countries: cross-sectional study. J Med Internet Res 2025 Mar 19;27:e65147. [doi: 10.2196/65147] [Medline: 40105882]
43. Mold F, Cooke D, Ip A, Roy P, Denton S, Armes J. COVID-19 and beyond: virtual consultations in primary care-reflecting on the evidence base for implementation and ensuring reach: commentary article. BMJ Health Care Inform 2021 Jan;28(1):e100256. [doi: 10.1136/bmjhci-2020-100256] [Medline: 33436372]
44. Donaghy E, Atherton H, Hammersley V, et al. Acceptability, benefits, and challenges of video consulting: a qualitative study in primary care. Br J Gen Pract 2019 Sep;69(686):e586-e594. [doi: 10.3399/bjgp19X704141] [Medline: 31160368]
45. Kateb S, Ruehle RC, Kroon DP, van Burg E, Huber M. Innovating under pressure: adopting digital technologies in social care organizations during the COVID-19 crisis. Technovation 2022 Jul;115:102536. [doi: 10.1016/j.technovation.2022.102536]
46. Orrange S, Patel A, Mack WJ, Cassetta J. Patient satisfaction and trust in telemedicine during the COVID-19 pandemic: retrospective observational study. JMIR Hum Factors 2021 Apr 22;8(2):e28589. [doi: 10.2196/28589] [Medline: 33822736]

Abbreviations

ASHIP: Association of Statutory Health Insurance Physicians
GP: general practitioner
OR: odds ratio
VC: video consultation

XSL•FO
RenderX

Edited by J Sarvestan; submitted 12.Mar.2025; peer-reviewed by C Doarn, E Galvin; revised version received 15.Nov.2025; accepted 17.Nov.2025; published 06.Jan.2026.

Please cite as:
Kleinschmidt L, Wasem J, Blase N, Nauendorf B, Malsch J, Brittner M, Brandenburg P, Aeustergerling A, Hüer T
Attitudes Toward Video Consultations From the Perspective of Physicians and Psychotherapists in German Outpatient Care After the COVID-19 Pandemic: Survey Study
J Med Internet Res 2026;28:e73757
URL: https://www.jmir.org/2026/1/e73757
doi:10.2196/73757

© Lara Kleinschmidt, Juergen Wasem, Nikola Blase, Beatrice Nauendorf, Juliane Malsch, Matthias Brittner, Paul Brandenburg, André Aeustergerling, Theresa Hüer. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Communication Challenges and Mitigation Strategies in Primary Care Virtual Consultations: Qualitative Study

Ahmed Alboksmaty[1], MBBS, MPH; Tetiana Lunova[1], PhD; Ara Darzi[1], OM, KBE, PC; Ana-Luisa Neves[1,2], PhD

[1]Department of Surgery and Cancer, Institute of Global Health Innovation, Imperial College London, St Mary's Campus, Norfolk Place, London, United Kingdom

[2]Department of Primary Care and Public Health, Imperial College London, London, United Kingdom

Corresponding Author:
Ahmed Alboksmaty, MBBS, MPH
Department of Surgery and Cancer, Institute of Global Health Innovation, Imperial College London, St Mary's Campus, Norfolk Place, London, United Kingdom

Abstract

Background: The growing reliance on virtual consultations in primary care has reshaped traditional general practitioner (GP)–patient communication dynamics, presenting new challenges that affect care quality and safety.

Objective: This study explores communication challenges and gaps, particularly relevant to virtual consultations compared with face-to-face interactions, as well as identifying mitigation strategies from both GPs' and patients' perspectives.

Methods: This qualitative study employed 4 online focus group discussions with a purposive sample of UK-based GPs and patients. Data were analyzed using a deductive-inductive thematic approach with NVivo software. The extended Shannon-Weaver communication model and the Capability, Opportunity, Motivation and Behavior model guided the analysis of communication challenges and mitigation strategies, respectively. The Consolidated Criteria for Reporting Qualitative Research were followed to ensure rigorous reporting.

Results: A total of 21 participants (12 patients and 9 GPs) took part in 4 online focus group discussions, 2 for patients and 2 for GPs. Six key themes on communication challenges emerged: 5 aligned with the extended Shannon-Weaver communication model (related to the sender-encoder, message, channel, receiver-decoder-feedback, and context), and a new one was inductively identified (patient autonomy and inclusivity). GPs, as senders, highlighted missing visual cues, affecting message clarity in remote communication channels. Patients, as receivers, reported difficulties explaining symptoms remotely, reduced emotional connection, and perceived empathy, linked to contextual challenges and the need for inclusive communication. Mitigation strategies were mapped to the Capability, Opportunity, Motivation and Behavior model: capability (training/resources), opportunity (triage/tools), and motivation (patient engagement/system adaptability), with participants emphasizing tailored training, standardized approaches, and flexible models to support effective and inclusive virtual communication.

Conclusions: This study highlights communication gaps in virtual consultations and proposes actionable mitigation strategies. Tailored use of virtual modalities, supported by structured training and policy efforts, is essential to ensure effective and safe remote communication.

(J Med Internet Res 2026;28:e79399) doi:10.2196/79399

KEYWORDS

communication; remote care; patient-physician relationship; virtual care; telehealth

Introduction

Primary care has become increasingly complex and demanding, driven by the variety of patients' conditions, the multifaceted nature of health care systems, the introduction of new health technologies, and the changing needs and preferences of patients and communities [1,2]. These factors have proposed a shift toward virtual consultations, emerging a transformation in primary care delivery [3]. This trend gained considerable momentum during the COVID-19 pandemic, as telephone and video consultations became essential for maintaining health care access amid disruptions to traditional face-to-face services [3,4]. While these virtual modalities have enhanced the accessibility and sustainability of services, they also necessitate a much-needed reevaluation of the core components of communication in health care to uphold quality and safety standards [3,4].

Effective communication is fundamental to high-quality primary care, acting as the foundation for building trust, fostering mutual understanding, and delivering patient-centered care [5]. The Shannon-Weaver communication model (SWCM), a foundational framework in communication theory, offers a

XSL•FO
RenderX

valuable perspective for critically analyzing GP-patient interactions across different consultation modalities (virtual and face-to-face) [6]. The extended SWCM comprises 9 components, ie, sender, encoder, message, channel, noise, decoder, receiver, feedback, and context, which collectively outline the classic communication process and can be adapted to various contexts [6].

Traditionally, the primary channel for communication between general practitioners (GPs) and patients has been face-to-face consultations. While this method facilitates direct engagement, it is not without challenges, such as time constraints, language and cultural barriers, and environmental distractions, factors that both GPs and patients have learned to navigate and challenge [7]. However, the shift to virtual consultation modalities disrupts these established dynamics, introducing new communication barriers that require tailored mitigation strategies to ensure effective and equitable care delivery [3].

Virtual consultations, whether conducted via phone or video calls, present an additional range of challenges, including technical difficulties, communication barriers, privacy and confidentiality concerns, limitations in clinical assessment, and adaptation challenges for both GPs and patients [3,8,9]. While issues with mutual understanding during in-person consultations might arise from the use of jargon or complex medical terminology, virtual consultations introduce the added difficulty of missing nonverbal cues and expressions, which can exacerbate communication gaps [10]. These distinctive challenges in remote interactions can impact the quality and safety of primary care, indicating the importance of assessing these issues in practice from the perspectives of both GPs and patients [3].

Extensive research has explored communication challenges, gaps, required skills, and improvement strategies in traditional face-to-face medical consultations within primary care [5,11,12]. However, there is limited evidence on the unique challenges and implications introduced by virtual consultation modalities. This study aims to address this gap by exploring the communication barriers and gaps specific to virtual consultations compared to traditional in-person consultations in primary care. It further seeks to identify potential strategies for mitigating these challenges based on the COM-B (capability, opportunity, motivation, and behavior) model [13], which offers a framework for identifying and presenting strategies for improvement. The assessment incorporates perspectives from both GPs and patients, who are the primary participants in the consultation process.

Methods

Study Design

This study adopted a qualitative approach, employing focus group discussions. The study adhered to the Consolidated Criteria for Reporting Qualitative Research (COREQ) checklist to ensure comprehensive reporting of the research processes and outcomes [14].

Study Participants and Recruitment

Participants included patients and GPs with experience of being involved in primary care consultations, both face-to-face and remote interactions within the United Kingdom. Adult patients aged 18 years or older who could communicate in English and had lived experience of virtual consultations (video or audio) with GPs were targeted to facilitate meaningful engagement and interaction during the discussions. Patients were recruited through Valuing Our Intellectual Capital and Experience, an online platform for community engagement and involvement in research [15].

GPs were recruited through a combination of convenience and purposive sampling via the research team's networks, followed by snowball sampling to ensure diversity among participants [16]. GP participants were required to have professional experience conducting virtual consultations (video or audio) in primary care within the UK National Health Service to ensure shared familiarity with the primary care context, including relevant policies and regulations.

All participants received, via email, a detailed participant information leaflet describing the study background, proposed methodology, and objectives. Participants were encouraged to ask questions for clarification before providing informed consent and participating in the focus groups.

Data Collection

Four online focus groups were conducted via Microsoft Teams between June and August 2024: 2 with GPs and 2 with patients. Each group participated independently to capture distinct perspectives. The focus groups were conducted as part of a larger study aimed at identifying the safety implications of virtual consultations in primary care. A researcher (TL) with a clinical background, PhD qualification, and expertise in patient safety and digital health in primary care moderated all 4 focus groups. A senior researcher (ALN) comoderated the first GP focus group to ensure alignment of the discussions with the study's objectives.

A semistructured topic guide was developed to facilitate interactive discussions during the focus groups, covering key aspects of communication challenges and potential solutions. The guide was piloted within the research team to ensure clarity and a logical flow. The questions in the topic guide aimed at exploring communications issues and strategies are presented in Textbox 1.

XSL•FO
RenderX

Textbox 1. Questions relevant to discussing communication issues in the topic guide.

- Communication problems in virtual consultations (this section aimed to explore the communication issues experienced by participants [GPs and patients] during virtual consultations).
  - What communication issues do you think can arise in virtual consultations?
  - Have you experienced any communication problems when having a virtual consultation?
  - How do you think these issues can be solved?

Data Analysis

Microsoft Teams' transcription feature was used to generate automatic transcripts of the discussions, which were later reviewed against the audio recordings by the focus group moderator to ensure accuracy before finalization and analysis. An interpretivist approach was adopted for this qualitative research [17], considering that participants construct meaning through their experiences and regular involvement in consultation dynamics. The analytical focus was therefore on understanding how participants interpreted and made sense of communication within virtual consultations. Transcripts were not returned to participants for further review; instead, the transcripts were directly analyzed by the research team to identify recurring themes and insights into participants' experiences and perspectives.

A deductive-inductive approach was employed for data analysis to ensure a comprehensive assessment while allowing for new themes to emerge [18]. The deductive analysis was guided by the components of the extended SWCM model [6], which were set as proposed themes for communication challenges and gaps. The extended SWCM originally included 9 components, which were reviewed and refined by the research team into 7 key components (Table 1), representing the major deductive themes for this analysis. For the analysis of suggested mitigation strategies, the COM-B model was adopted as the framework for deductive themes, as illustrated in Figure 1.

Figure 1. Redefined COM-B model's components, adapted to be used for the deductive analysis of the mitigation strategies to improve communication during virtual consultations. COM-B: Capability, Opportunity, Motivation and Behavior; GP: general practitioner.

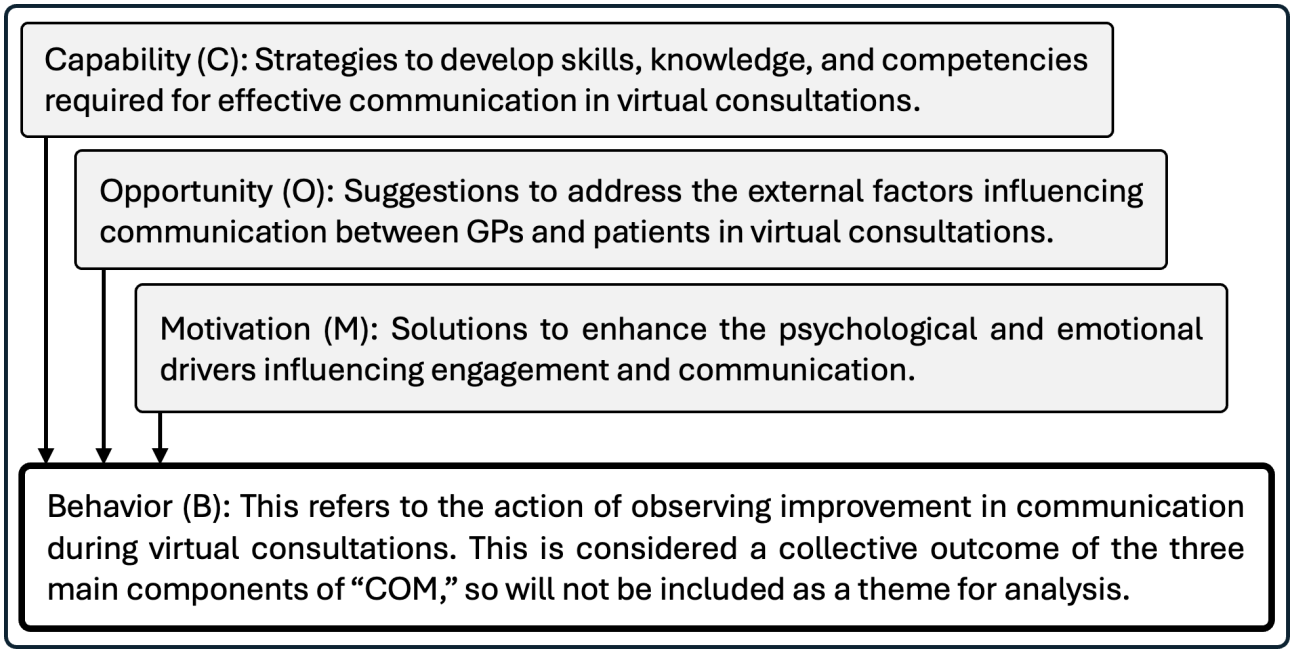

XSL•FO
RenderX

Table . Adapting the SWCM's[a] components for the deductive analysis of communication challenges and gaps.

SWCM components	Adapted definitions
Sender-Encoder	• Merged definition: Combines the roles of the sender (GP[b]) and the encoder, as they are closely intertwined. The sender is the person who initiates the communication, and the encoder refers to the process of transforming thoughts into verbal and nonverbal expressions and talks. • Contextualization to our research: How GPs articulate and communicate their questions, advice, and thoughts with their patients during virtual consultations.
Message	• Definition: The content of the communication, including clinical advice, questions, and emotional cues. • Contextualization to our research: How the clinical discussion is constructed, its clarity, and its comprehensiveness through different modalities.
Channel	• Definition: The medium through which the communication is delivered, virtual (eg, video call, telephone) or face-to-face. • Contextualization to our research: The limitations and merits of each modality in conveying messages effectively, particularly regarding technological or physical limitations.
Noise	• Definition: Any factors that disrupt or distort the communication process. • Contextualization to our research: This would include, for example, poor connectivity, background distractions, emotional and cognitive barriers.
Decoder-Receiver	• Merged definition: Combines the roles of the decoder (interpreting the message) and the receiver (the person/patient for whom the message is intended). • Contextualization to our research: How effectively patients understand the information, particularly the challenges posed by reduced nonverbal cues in phone consultations.
Feedback	• Definition: The responses or reactions from the receiver (patient) that indicate whether the message was understood as intended. • Contextualization to our research: How feedback mechanisms differ between virtual and face-to-face settings.
Context	• Definition: The situational factors influencing the communication process, such as the patient-clinician relationship, and the differences in medical conditions. • Contextualization to our research: How different contexts impact the communication dynamics, including the adaptability of both patients and clinicians to virtual or face-to-face settings.

[a]SWCM: Shannon-Weaver communication model.

[b]GP: general practitioner.

The initial analysis was conducted by a researcher (AA) with expertise in qualitative data analysis and the study topic. Analyses of the patients' and GPs' focus group datasets were first conducted independently, followed by triangulation across both sources to support a comprehensive and coherent interpretation of the findings [19,20]. This process combined a deductive approach, using the predetermined models, with an inductive approach to identify any newly emerging themes. NVivo software was used to facilitate the analysis [21].

Data adequacy was supported by the concept of information power [22], given the focused discussions, the relevance of participants recruited through purposive sampling, and the strong alignment between the data and the study objectives. During analysis, the research team observed thematic stability across the focus groups, indicating that additional data collection was unlikely to generate new insights and suggesting a level of data saturation appropriate for the study aims [23]. Preliminary themes were further refined through iterative analysis and discussions among all members of the research team to reach a consensus on the final themes.

Ethical Considerations

Ethics approval for this study was obtained from the Imperial College Research Ethics Committee (ICREC 6833567) before commencing the recruitment and study processes. All

XSL•FO
RenderX

participants provided informed consent prior to taking part in the focus group discussions. This consent was obtained after participants had reviewed comprehensive participant information leaflets detailing the study's aims and procedures, and after they had the opportunity to ask questions and receive clarification on any aspect of the study. To protect privacy and confidentiality, data were pseudoanonymized; identifiable details were removed after verification of transcription accuracy, and only deidentified qualitative data were used in the analysis. Participants received a £30 (US $40.36) voucher as a token of appreciation for their time and contribution.

Results

Participant Characteristics

A total of 21 participants (12 patients and 9 GPs) were involved in 4 focus group discussions; 2 patient groups, each with 6 participants, and 2 GP groups with 5 and 6 participants, respectively. None of the invited participants refused to take part. Each online focus group lasted for an average duration of 90 minutes. All participating GPs (n=9) were based in England, along with the majority of patients (10/12). The participating GPs had a mean of 5.33 (SD 4.06) years of National Health Service experience, with a range spanning from 2 to 16 years. Table 2 provides an overview of the participants' characteristics.

XSL•FO
RenderX

Table . Self-reported participant characteristics.

Characteristic	Patients	GPs[a]
Total number of participants, n (%)	12 (57.14)	9 (42.86)
Age bands, n (%)		
18- - 29 years (n=1, 4.76%)	1 (4.76)	0 (0)
30- - 39 years (n=10, 47.62%)	3 (14.29)	7 (33.33)
40- - 49 years (n=4, 19.05%)	2 (9.52)	2 (9.52)
50- - 59 years (n=2, 9.52%)	2 (9.52)	0 (0)
≥60 years (n=4, 19.05%)	4 (19.05)	0 (0)
Missing (n=0, 0%)	0 (0)	0 (0)
Sex, n (%)		
Female (n=13, 61.90%)	8 (38.10)	5 (23.81)
Male (n=8, 38.10%)	4 (19.05)	4 (19.05)
Missing (n=0, 0%)	0 (0)	0 (0)
Geographic location, n (%)		
Bournemouth (n=1, 4.76%)	1 (4.76)	0 (0)
Bristol (n=1, 4.76%)	1 (4.76)	0 (0)
Cambridge (n=2, 9.52%)	0 (0)	2 (9.52)
Edinburgh (n=1, 4.76%)	1 (4.76)	0 (0)
Leicester (n=1, 4.76%)	0 (0)	1 (4.76)
Lisburn (n=1, 4.76%)	1 (4.76)	0 (0)
London (n=9, 42.86%)	3 (14.29)	6 (28.57)
Newcastle Upon Tyne (n=4, 19.05%)	4 (19.05)	0 (0)
Walsingham (n=1, 4.76%)	1 (4.76)	0 (0)
Missing (n=0, 0%)	0 (0)	0 (0)
Ethnic group, n (%)		
Arab (n=2, 9.52%)	1 (4.76)	1 (4.76)
Asian or Asian British (n=6, 28.57%)	2 (9.52)	4 (19.05)
Black or Black British (n=2, 9.52%)	2 (9.52)	0 (0)
Mixed (n=1, 4.76%)	1 (4.76)	0 (0)
Other (n=1, 4.76%)	0 (0)	1 (4.76)
White (n=9, 42.87%)	6 (28.57)	3 (14.29)
Missing (n=0, 0%)	0 (0)	0 (0)

[a]GP: general practitioner.

Communication Challenges and Gaps

Overview

Suggestions from GPs and patients were overall aligned, reflecting shared struggles in communication during virtual consultations and a common desire for improvement. Six themes were identified through the data analysis of both patient and GP focus groups, comprising 5 deductive themes derived from merging some of the SWCM components and 1 inductive theme [6]. The 5 deductive themes included communication challenges related to (1) sender-coder, (2) message, (3) channel, (4) receiver-decoder-feedback, and (5) context with modality appropriateness. A newly identified inductive theme was labeled patient autonomy and inclusivity. A description of each theme and its subthemes is provided in Table 3.

XSL•FO
RenderX

Table . Major themes, description, and subthemes identified through the data analysis.

Major theme	Description	Subthemes
Theme 1: Sender–Encoder	How GPs[a] articulate and communicate their questions, advice, and thoughts with their patients during virtual consultations.	• Issues in continuity and coordination of care • Misunderstanding, bias, and missing clinically important points • Lack of empathy and emotional engagement
Theme 2: Message	How the clinical discussion is constructed, its clarity, and its comprehensiveness through different modalities.	• Less clarity and comprehensiveness in virtual consultations • Limited diagnostic visualization during phone consultations
Theme 3: Channel	The limitations and merits of each consultation modality in conveying messages effectively, particularly regarding technological or physical limitations.	• Technical and organizational barriers • Physical barriers in virtual communication • Accessibility and reliability issues for stable connection in virtual consultations
Theme 4: Receiver–Decoder–Feedback	How patients interpret and understand clinical information and discussions and how their responses and interactions indicate understanding and engagement.	• Lack of trust and confidence in virtual consultations • Communication issues that affect understanding, including language barriers • Resistance by patients to engage in virtual consultations • Unclear understanding and missing engagement cues • Missing natural interactions and non-verbal cues in phone consultations
Theme 5: Context and modality appropriateness	The situational factors and suitability of communication methods that influence the effectiveness and engagement of interactions between GPs and patients.	• System and structural limitations • Influence of cultural and social norms • Questionable patient expectations and trust • No standardized system to decide on the consultation type
Theme 6: Patient autonomy and inclusivity	How patient autonomy and inclusivity influence communication between GPs and patients, focusing on the extent to which patients can control their consultation choices and the barriers to full participation in virtual care.	• Barriers to choosing consultation modality • Technical challenges faced by some patient groups in engaging with virtual consultations • Patient preferences for in-person consultations • Inequality in access to technology

[a]GP: general practitioner.

Theme 1: Sender–Encoder (GPs)

Virtual consultations occasionally disrupted clinicians' and patients' ability to interpret emotional and clinical cues, leading to a mutual sense of disconnection that affects relational quality. GPs expressed concerns about missing nonverbal and visual clinical signs during phone consultations, negatively affecting diagnostic decision-making. They also highlighted the extra efforts needed for maintaining engaging conversations with patients remotely, a difficulty further exacerbated by increasing workload pressures. Patients, in turn, reported feeling a lack of empathy and emotional connection during virtual consultations, highlighting a perceived disconnect between themselves and their GPs in virtual settings, particularly in phone consultations.

> *Missing non-verbal communication, the stuff we wouldn't get in a virtual [phone] consultation, you can't see them [patients] walk into the room, can't see how they're sitting, or notice the little things that might make you think differently.* [GP 3.3]

> *Many of the emotions and empathy, at least towards my case, are almost completely gone when it comes to phone and video consultations.* [Patient 2.1]

Theme 2: Message (Content of the Medical Consultation)

The lack of direct interaction in virtual consultations complicated symptom description and clinical explanation, making it harder for both patients and GPs to achieve clarity, especially when addressing complex or sensitive issues. Patients raised significant concerns about the difficulty of clearly describing symptoms and medical complaints over the phone. Likewise, GPs reported challenges in effectively explaining diagnostic reasoning and management plans to patients and ensuring their understanding during virtual consultations, which is of more concern during phone calls. These issues became particularly challenging when discussing sensitive or complex medical conditions.

XSL•FO
RenderX

> *…Because the doctor can't pick up on non-verbal cues that might explain a complaint or tell if the patient wants to say something else.* [Patient 3.2]

Theme 3: Channel

Participants reflected that the risk of technical unreliability in remote communication tools essentially reduces the perceived quality of virtual consultations. Patients expressed increased concerns about the timing of their appointments, often disrupted by delays in calls, the reliability of their internet or phone connections, and the challenge of effectively describing symptoms over the phone, particularly for non-native speakers. Similarly, GPs reported concerns about dropped calls or lost connections during consultations, as well as issues with the quality of sound and images shared online, which could hinder effective communication.

> *The number of times I've called patients—too many—they've had a really bad signal, no alternative number to call, or they're out in a supermarket or somewhere else, making me just can't understand a word they're saying.* [GP 4.3]

Theme 4: Receiver–Decoder–Feedback

Persistent challenges in trust, confidence, and equitable access to care influenced patient engagement with virtual consultations, with the absence of nonverbal interactions over the phone often weakening the mutual understanding essential for safe and reassuring care. Despite the normalization of virtual consultations for GP appointments in the United Kingdom, particularly since the COVID-19 pandemic, patients continued to express a lack of confidence and trust in this method for addressing serious medical conditions. This has been perceived by GPs as a form of resistance, especially among elderly patients, toward accepting virtual consultations as the only method of communication, even when deemed clinically safe.

> *There's a lot of reluctance towards virtual consultations across the country… some people say there's no place for virtual consultations because they've been brought up with the face-to-face discussion model.* [GP 1.4]

Patients further expressed concerns about equity in accessing and benefiting from remote care, particularly among individuals with special needs, such as those with hearing impairments, mental health issues, in need of an interpreter, or learning difficulties. Both GPs and patients reported ongoing challenges with interactive feedback during virtual consultations, largely due to the reliance on verbal cues alone, with potentially facial reactions in video consultations, to convey messages and management plans on most occasions.

Theme 5: Context and Modality Appropriateness

Both GPs and patients emphasized that effective communication depends on aligning consultation modality with clinical needs and contextual expectations. GPs highlighted the need for a robust triage system and effective coordination policies to ensure patients receive the most appropriate and convenient consultation modality tailored to their individual needs. Patients echoed this sentiment, emphasizing that communication during medical appointments extends beyond the consultation itself and is influenced by health care system structure, social norms, and community culture. Bridging these perspectives, both groups stressed that aligning the consultation modality with patient expectations and needs is crucial for building trust and ensuring effective communication.

> *I needed to be present with the doctor. I needed to be there, face-to-face, when talking about sensitive and difficult things. I think sometimes GPs need to acknowledge that a quick call or video call isn't the right way to talk to a patient at certain times.* [Patient 1.1]

Theme 6: Patient Autonomy and Inclusivity

Patients shared a wish to have their voices heard beyond their clinical symptoms, particularly in deciding the type of consultation they receive with their GP. They further explained that some patients struggle to engage, express their emotions, or maintain focus during remote conversations, making in-person appointments more suitable for their personal traits, regardless of their symptoms. Conversely, patients with demanding jobs or caregiving responsibilities may prefer communicating remotely for their convenience. GPs acknowledged this perspective, emphasizing that person-centered and equitable care requires understanding each patient's preferences and capabilities while also assessing their medical needs.

> *I would like a bit more patient choice over which consultations you may have. There may be some conditions where you're quite happy to talk about them over the phone, while with others, you know you'd be better off having an in-person or video consultation.* [Patient 3.2]

Mitigation Strategies to Improve Communication

The proposed mitigation strategies targeted the identified challenges and gaps, which were categorized into 3 deductive themes (capability, opportunity, and motivation) based on the COM-B model [13]. Table 4 outlines these major themes and the associated subthemes based on the data analysis.

XSL•FO
RenderX

Table . Theme of the mitigation strategies based on the capability, opportunity, motivation, and behavior model.

Major theme	Subthemes
Theme 1: Capability	• Training and skill development for virtual consultations • Resource provision and standardization • Enhancing patient understanding
Theme 2: Opportunity	• Triage and risk mitigation • Enhancing communication through supplementary materials • Addressing barriers to effective communication
Theme 3: Motivation	• Patient and doctor preferences in consultation modalities • Supporting patient preparedness and engagement • Monitoring and adapting models of care

Theme 1: Capability

Training and skill development for virtual communication was highlighted by both patients and GPs as essential for improving interactions during virtual consultations. It was noted that strong communication skills in face-to-face settings do not necessarily translate to effective communication over the phone, and vice versa. Therefore, it was suggested that medical students and doctors in training should gain practical experience in conducting virtual consultations and leading virtual clinical appointments. Additionally, there was a shared call for equipping GPs with modern tools and reliable infrastructure to enable seamless and effective communication with patients remotely.

> *I think we need to train our GP trainees using a different model of communication.* [GP 1.4]
>
> *…training [of GPs] on using specific platforms and building rapport over the phone to aid patient-doctor communication, I think, is important, how you do that over the phone to achieve the maximal effect.* [Patient 3.2]

Theme 2: Opportunity

The primary recommendation in this theme was to set a standardized triage system and clear protocols to determine the most suitable consultation modality based on each patient's individual needs, preferences, and characteristics. Offering video consultations as a standard option, in addition to phone appointments, was also suggested as a way to enhance engagement and build stronger professional doctor-patient relationships. Another proposal involved maintaining a registry of patients who face barriers to virtual consultations, such as those living in areas with poor signal, experiencing language barriers, suffering from hearing impairments, or having complex medical conditions.

> *In our practices, there are two completely separate approaches. One practice has a policy that if patients want virtual, they get virtual, and if they want face-to-face, they get face-to-face. In the other practice, the receptionist decides.* [GP 3.4]

Theme 3: Motivation

A GP's comfort and familiarity with different consultation modalities were identified as important factors influencing communication quality, whether virtually or face-to-face. Some participants proposed a system that allocates patients to GPs based on their preferred consultation modality, for instance, assigning virtual consultations to GPs who favor them, while those who prefer direct communication handle face-to-face appointments. However, concerns were raised about the potential impact of this approach on continuity of care.

> *One approach could be to ask patients whether they would prefer a virtual consultation with a new doctor or one they are familiar with. Personally, I would feel more comfortable talking to a doctor I know in person.* [GP 2.4]

Patients also suggested offering incentives to encourage engagement in virtual consultations, such as implementing a fast-track appointment booking system for those opting for virtual appointments. However, some participants noted potential inequality issues with such incentives, as certain patients may be unable to choose virtual consultations due to personal or medical needs. Additionally, a system to promptly update patients about any changes to their appointment times was recommended to help them prepare accordingly and feel more confident in the system.

Discussion

Summary of Results

This study adopted a qualitative approach using focus groups with GPs and patients to identify communication challenges in primary care virtual consultations compared to traditional face-to-face appointments, as well as exploring mitigation strategies. Four online focus groups were conducted, 2 with GPs (n=9) and 2 with patients (n=12).

Six major themes emerged regarding the challenges and gaps in virtual consultations, 5 of which align with the SWCM model: sender-encoder, message, channel, receiver-decoder-feedback, and context with modality appropriateness. An additional theme, patient autonomy and inclusivity, was identified inductively. For GPs as sender-encoders, misinterpretations and missed clinically significant cues were major concerns. The "message" was often compromised by limited clarity in dialog and difficulties in communicating diagnostic information virtually. Technical issues, poor connectivity, and accessibility barriers complicated the remote "channels" of communication. Patients, as receiver-decoders, reported a lack of trust in discussing

XSL•FO
RenderX

complex medical needs remotely and perceived reduced empathy and emotional engagement from GPs in virtual consultations. Social norms and cultural factors, especially among elderly patients, shaped preferences and communication styles, highlighting the importance of "context and modality appropriateness." Furthermore, patient autonomy and inclusivity in choosing virtual consultations influenced their engagement and willingness to communicate effectively during virtual primary care appointments.

Proposed mitigation strategies were mapped to the COM-B model: capability (training and resource provision), opportunity (triage systems and supplementary communication tools), and motivation (patient engagement and system adaptability). Participants highlighted the need for tailored capacity building through virtual-communication-focused training and skills development, supported by resource provision and standardized approaches. Recommendations to enhance the opportunity for better communication in virtual consultations included implementing triage and risk mitigation strategies, using supplementary materials to support remote interactions, and continuously identifying and addressing communication barriers. To strengthen motivation among both GPs and patients, participants emphasized the importance of respecting individual preferences for consultation modality, offering preparatory guidance and materials to support patient engagement, and ensuring flexibility and adaptability in care models.

Findings in Context of Existing Literature

Consistent with our findings, previous studies have identified common challenges in virtual consultations. Patients often struggle to articulate complex symptoms over the phone, while GPs face difficulties ensuring patient understanding of medical information and management plans [24]. These challenges are particularly pronounced in sensitive discussions, cases involving complex health needs, conditions requiring physical examination [25], and with new patients unfamiliar to the GP [26,27]. Some conditions, such as mental health concerns, raise questions about the appropriateness of virtual consultations on some occasions [28]. Our participants described a perceived lack of empathy in virtual appointments, contributing to dissatisfaction in mental health care. However, previous studies reported contrasting evidence, suggesting that offering timely virtual appointments may enhance mental health care by enhancing accessibility to medical advice whenever needed [3,28,29]. This inconsistency in evidence for some medical conditions highlights the importance of contextualizing and adapting care and communication models, including the choice of consultation modality, to align with individual patient needs and clinical circumstances [11]. It is also important to note that patient preferences for communicating remotely versus face-to-face are influenced not only by their health needs but also by personal characteristics [30]. Evidence shows that younger individuals, those with busy schedules, and highly educated patients tend to favor virtual appointments for convenience and accessibility [3,31].

Existing literature has tended to identify similar communication challenges in primary care consultations, but these are often presented in a fragmented manner [26,29,32]. Our study synthesizes these challenges within a structured, theoretically informed framework [6], with particular emphasis on virtual consultations as an emerging norm in primary care [3]. By framing communication issues as a shared responsibility between patients and GPs [33], we offer a more nuanced interpretation that supports practical approaches to mitigation. Through the application of the adapted SWCM framework [6], we further highlight how patient, clinician, technological, and contextual factors intersect, enabling us to propose targeted interventions that may lead to meaningful and sustained improvements in communication practices.

Importantly, communication challenges in virtual consultations extend beyond the consultation itself, involving both pre- and post-encounter interactions [32,34,35]. A qualitative study conducted in Australia highlighted pre-consultation patient-related factors that influence GP-patient communication, particularly health literacy and familiarity with digital platforms [26]. Many patients struggle with completing pre-consultation forms required for describing their symptoms and concerns for booking virtual appointments, especially when selecting appropriate terms to describe their symptoms [32]. From a GP's perspective, familiarity with the patient, as well as access to and time for reviewing medical records, plays a crucial role in shaping communication style and ensuring a high-quality virtual consultation [27,36]. Existing tools also provide structured checklists and guidance to support high-quality virtual consultations, with communication highlighted as a central component [37]. For example, the Telehealth Etiquette Competency Checklist (TECC) emphasizes communication alongside technological, environmental, and confidentiality considerations [37].

Effective communication remains essential even after the consultation itself, particularly in follow-up messaging, referrals, and sharing components of the agreed-upon management plan [34]. The widespread adoption of virtual consultations has raised concerns regarding continuity of care, as it may challenge the foundation of building a strong professional relationship between GPs and their patients, which is an essential factor in enhancing the consultation experience and overall quality of care [38]. Ensuring clarity in follow-up plans after virtual consultations is crucial, particularly in guiding patients on how to seek further care if their concerns persist or if the initial virtual visit does not fully address their health needs [8,35]. Clear and efficient remote communication can help reduce unnecessary follow-up visits, whether in primary care or to hospital and ambulatory care [3].

To mitigate the identified communication barriers and enhance interactions during virtual consultations, GPs often use verbal affirmations, such as "yes" and "I see," to prove active listening [33]. They may also verbally narrate any concurrent tasks, such as reviewing investigation results or previous reports, to maintain transparency with patients for reassurance. Such capacity-building requirements for these supporting skills have been proposed by the participants in our study to improve communication.

Evidence suggests that, when feasible, video consultations may be preferable for initial appointments to establish rapport, for

XSL•FO
RenderX

assessments requiring observation of physical signs [25], and for counseling appointments. Existing literature also offers evidence-based guidelines, emphasizing aspects regarding the knowledge, skills, attitudes, and teaching strategies required for high-quality videoconferencing, which primary care teams can adapt to their local context [39]. However, no clear advantage or superiority of video consultations over face-to-face interactions has been established [40]. The collective evidence from broader literature indicates that both GPs and patients generally prefer communication through face-to-face consultations, followed by video calls when feasible, with phone consultations being the least preferred [41]. However, this hierarchy may vary depending on the context; for instance, when the consultation is for routine medication renewal, or patients live far away from their GP surgeries, phone or video consultations could be more practical [42].

Strengths and Limitations

This study focuses on the core aspect of virtual consultation -communication- as a critical determinant of both satisfaction and patient preference in clinical practice [5]. Focus groups were employed due to their ability to foster an in-depth dialog among participants and to identify shared concerns effectively [43,44]. However, focus groups could potentially allow vocal participants to dominate the discussion and the possibility that some individuals may share less in-depth reflections in a group setting [45]. To mitigate these risks, the moderator had substantial experience in conducting similar research, ensuring balanced participation and providing sufficient time for individuals to share their views [45]. Additionally, in terms of methodology, future research using direct observation or other objective methods could reveal communication nuances that focus group discussions may miss [46]. This would enable a clearer assessment of communication aspects in different consultation modalities.

The analysis was strengthened by the use of 2 theoretical frameworks to structure the data analysis and guide its interpretation: the SWCM framework [6], which addresses the challenges and gaps, and the COM-B model [13], which informs the development of mitigation strategies. Our analysis integrated insights from both GP and patient transcripts, emphasizing that effective communication is inherently reciprocal and requires active engagement from both parties.

While we aimed to recruit a diverse patient population, further research is needed to explore communication challenges among particular groups with potential distinct needs, such as patients requiring interpreters or those with mental health conditions and complex chronic illnesses like cancer. Additionally, all participating GPs were based in England, with an average of 5.4 years of experience. Future studies should include GPs at varying career stages and across different practice settings to identify role-specific communication challenges and potential training needs. Moreover, our study categorized both video and telephone consultations under the broad term of virtual consultations; however, each modality presents unique communication dynamics that warrant further investigation.

Implications for Research, Policy, and Practice

Virtual consultations have become an integral part of the “norm” in primary care, whether in the UK or globally. Therefore, it is essential to optimize remote communication to uphold patient safety and care quality [3]. The communication challenges and mitigation strategies identified in this study, structured using theoretical frameworks, provide actionable themes to be considered by policymakers and practitioners for improvement plans. For example, targeted communication training for GPs, ensuring that GP practices are equipped with stable and reliable technology for virtual consultations, and implementing guidelines to allocate virtual appointments to patients who are both medically and personally suited for them are all key areas for intervention.

Further research should focus specifically on GP-patient communication during virtual consultations, examining its impact on patient satisfaction and clinical outcomes, including safety. Policy efforts should extend beyond the provision of technology and infrastructure to include structured training programs that equip GPs and other clinicians with the skills necessary for effective remote communication. The training efforts should target not only qualified GPs but also undergraduate medical students and other healthcare professionals who are involved in patient care provision, potentially through virtual appointments [12]. In practice, the allocation of virtual consultations should consider both the GP’s level of experience in remote communication and the patient’s health condition and background. Communication in healthcare should be viewed holistically, not merely as spoken language, but as the means through which care is delivered, tailored to the needs of each patient.

Conclusion

This study highlights the pivotal role of communication in virtual consultations, highlighting existing gaps and potential strategies for improvement in primary care. Through focus groups with GPs and patients, we identified key communication challenges and potential mitigation strategies. While virtual consultations offer convenience, their use should be tailored to individual patient needs and clinical contexts. Future research should explore the distinct communication dynamics of different virtual modalities in primary care, informing policymakers and practitioners with avenues for improvement to ensure equitable, high-quality, and patient-centered primary care delivery.

Funding

This study was supported by the National Institute for Health and Care Research (NIHR) North-West London Patient Safety Research Collaboration (NIHR NWL PSRC, Ref. NIHR204292), with infrastructure support from the NIHR Imperial Biomedical Research Centre. ALN is also supported by the NIHR Applied Research Collaboration North-West London. The views expressed

XSL•FO
RenderX

in this publication are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. The funders had no role in study design, data collection and analysis, decision to publish, or writing the manuscript.

Authors' Contributions

AD, ALN, and TL contributed to the conceptualization of the study. TL (she/her), a postdoctorate research associate, supported by ALN, a senior clinical lecturer in digital health, led the conduct and transcription of the interviews. AA and ALN led the conceptualization of the data analysis approach and the selection of the theoretical frameworks guiding the analysis. AA conducted the initial thematic data analysis. TL reviewed and discussed the initial analysis outcomes with AA, while ALN reviewed and confirmed the final agreed analysis outcomes. All authors contributed to and confirmed the interpretation of the results. AA drafted the initial full manuscript. AD, ALN, and TL reviewed and provided feedback on the manuscript. All authors read and approved the final version of the manuscript. The corresponding author confirms that all listed authors meet the authorship criteria and that no individuals meeting the criteria have been omitted.

Conflicts of Interest

None declared.

References

1. Mount JK, Massanari RM, Teachman J. Patient care complexity as perceived by primary care physicians. Fam Syst Health 2015 Jun;33(2):137-145. [doi: 10.1037/fsh0000122] [Medline: 25893538]
2. Neve G, Fyfe M, Hayhoe B, Kumar S. Digital health in primary care: risks and recommendations. Br J Gen Pract 2020 Dec;70(701):609-610. [doi: 10.3399/bjgp20X713837] [Medline: 33243917]
3. Campbell K, Greenfield G, Li E, et al. The impact of virtual consultations on the quality of primary care: systematic review. J Med Internet Res 2023 Aug 30;25:e48920. [doi: 10.2196/48920] [Medline: 37647117]
4. Ezeamii VC, Okobi OE, Wambai-Sani H, et al. Revolutionizing healthcare: how telemedicine is improving patient outcomes and expanding access to care. Cureus 2024 Jul;16(7):e63881. [doi: 10.7759/cureus.63881] [Medline: 39099901]
5. King A, Hoppe RB. “Best practice” for patient-centered communication: a narrative review. J Grad Med Educ 2013 Sep;5(3):385-393. [doi: 10.4300/JGME-D-13-00072.1] [Medline: 24404300]
6. Shannon CE. A mathematical theory of communication. Bell Syst Tech J 1948;27(3):379-423. [doi: 10.1002/j.1538-7305.1948.tb01338.x]
7. Payne R, Dakin F, MacIver E, et al. Challenges to quality in contemporary, hybrid general practice a multi-site longitudinal case study. Br J Gen Pract 2024 Jan;75(750):e1-e11. [doi: 10.3399/BJGP.2024.0184] [Medline: 39117426]
8. Turner A, Morris R, Rakhra D, et al. Unintended consequences of online consultations: a qualitative study in UK primary care. Br J Gen Pract 2022 Feb;72(715):e128-e137. [doi: 10.3399/BJGP.2021.0426] [Medline: 34903520]
9. Williams S, Barnard A, Collis P, et al. Remote consultations in primary care across low-, middle- and high-income countries: implications for policy and care delivery. J Health Serv Res Policy 2023 Jul;28(3):181-189. [doi: 10.1177/13558196221140318] [Medline: 36484225]
10. Murphy M, Scott LJ, Salisbury C, et al. Implementation of remote consulting in UK primary care following the COVID-19 pandemic: a mixed-methods longitudinal study. Br J Gen Pract 2021;71(704):e166-e177. [doi: 10.3399/BJGP.2020.0948] [Medline: 33558332]
11. Moulaei K, Sheikhtaheri A, Fatehi F, Shanbehzadeh M, Bahaadinbeigy K. Patients’ perspectives and preferences toward telemedicine versus in-person visits: a mixed-methods study on 1226 patients. BMC Med Inform Decis Mak 2023 Nov 15;23(1):261. [doi: 10.1186/s12911-023-02348-4] [Medline: 37968639]
12. Thuraisingham C, Abd Razak SS, Nadarajah VD, Mamat NH. Communication skills in primary care settings: aligning student and patient voices. Educ Prim Care 2023 May;34(3):123-130. [doi: 10.1080/14739879.2023.2210097] [Medline: 37194600]
13. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
14. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
15. Newcastle University. Voice. URL: https://voice-global.org [accessed 2026-01-10]
16. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health 2015 Sep;42(5):533-544. [doi: 10.1007/s10488-013-0528-y] [Medline: 24193818]
17. Schwandt TA. Constructivist, interpretivist approaches to human inquiry. In: Denzin NK, Lincoln YS, editors. Handbook of Qualitative Research: Sage; 1994:118-137 URL: https://www.scirp.org/reference/referencespapers?referenceid=435054 [accessed 2026-01-10]

XSL•FO
RenderX

18. Bonner C, Tuckerman J, Kaufman J, et al. Comparing inductive and deductive analysis techniques to understand health service implementation problems: a case study of childhood vaccination barriers. Implement Sci Commun 2021 Sep 15;2(1):100. [doi: 10.1186/s43058-021-00202-0] [Medline: 34526141]
19. Ramani S, Mann K. Introducing medical educators to qualitative study design: twelve tips from inception to completion. Med Teach 2016 May;38(5):456-463. [doi: 10.3109/0142159X.2015.1035244] [Medline: 25897710]
20. Carter N, Bryant-Lukosius D, DiCenso A, Blythe J, Neville AJ. The use of triangulation in qualitative research. Oncol Nurs Forum 2014 Sep;41(5):545-547. [doi: 10.1188/14.ONF.545-547] [Medline: 25158659]
21. NVivo 15 - the most trusted qualitative analysis software (QDA) is even better. Lumivero. 2020. URL: https://lumivero.com/products/nvivo [accessed 2026-01-10]
22. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
23. Saunders B, Sim J, Kingstone T, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907. [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
24. Khanji MY, Gallagher AM, Rehill N, Archbold RA. Remote consultations: review of guiding themes for equitable and effective delivery. Curr Probl Cardiol 2023 Aug;48(8):101736. [doi: 10.1016/j.cpcardiol.2023.101736] [Medline: 37075908]
25. Honarbakhsh S, Sporton S, Monkhouse C, Lowe M, Earley MJ, Hunter RJ. Remote clinics and investigations in arrhythmia services: what have we learnt during Coronavirus disease 2019? Arrhythm Electrophysiol Rev 2021 Jul;10(2):120-124. [doi: 10.15420/aer.2020.37] [Medline: 34401185]
26. Nguyen AD, White SJ, Tse T, et al. Communication during telemedicine consultations in general practice: perspectives from general practitioners and their patients. BMC Prim Care 2024 Sep 4;25(1):324. [doi: 10.1186/s12875-024-02576-1] [Medline: 39232645]
27. Schers H, van den Hoogen H, Bor H, Grol R, van den Bosch W. Familiarity with a GP and patients' evaluations of care. A cross-sectional study. Fam Pract 2005 Feb;22(1):15-19. [doi: 10.1093/fampra/cmh721] [Medline: 15640289]
28. Bulkes NZ, Davis K, Kay B, Riemann BC. Comparing efficacy of telehealth to in-person mental health care in intensive-treatment-seeking adults. J Psychiatr Res 2022 Jan;145:347-352. [doi: 10.1016/j.jpsychires.2021.11.003] [Medline: 34799124]
29. Antonio S, Joseph D, Parsons J, Atherton H. Experiences of remote consultation in UK primary care for patients with mental health conditions: a systematic review. Digit Health 2024;10:20552076241233969. [doi: 10.1177/20552076241233969] [Medline: 38465292]
30. Lunova T, Hurndall KH, Crespo R, et al. Impact of the cost-of-living crisis on patient preferences towards virtual consultations. J Telemed Telecare 2025 Sep;31(8):1175-1185. [doi: 10.1177/1357633X241255411] [Medline: 38767152]
31. Crook RL, Iftikhar H, Moore S, Lowdon P, Modarres P, Message S. A comparison of in-person versus telephone consultations for outpatient hospital care. Future Healthc J 2022 Jul;9(2):154-160. [doi: 10.7861/fhj.2022-0006] [Medline: 35928204]
32. Atherton H, Eccles A, Poltawski L, Dale J, Campbell J, Abel G. Investigating patient use and experience of online appointment booking in primary care: mixed methods study. J Med Internet Res 2024 Jul 8;26:e51931. [doi: 10.2196/51931] [Medline: 38976870]
33. White SJ, Nguyen AD, Roger P, et al. Tailoring communication practices to support effective delivery of telehealth in general practice. BMC Prim Care 2024 Jun 27;25(1):232. [doi: 10.1186/s12875-024-02441-1] [Medline: 38937674]
34. Lyles CR, Gupta R, Tieu L, Fernandez A. After-visit summaries in primary care: mixed methods results from a literature review and stakeholder interviews. Fam Pract 2019 Mar 20;36(2):206-213. [doi: 10.1093/fampra/cmy045] [Medline: 29846584]
35. Reed M, Huang J, Graetz I, Muelly E, Millman A, Lee C. Treatment and follow-up care associated with patient-scheduled primary care telemedicine and in-person visits in a large integrated health system. JAMA Netw Open 2021 Nov 1;4(11):e2132793. [doi: 10.1001/jamanetworkopen.2021.32793] [Medline: 34783828]
36. Jeffers H, Baker M. Continuity of care: still important in modern-day general practice. Br J Gen Pract 2016 Aug;66(649):396-397. [doi: 10.3399/bjgp16X686185] [Medline: 27481958]
37. Pittmann R, Danaher-Garcia N, Adair White BA, Thompson A. Development and validation of the Telehealth Etiquette Competency Checklist: a Delphi study. J Telemed Telecare 2025 Oct;31(9):1308-1316. [doi: 10.1177/1357633X241279494] [Medline: 39311041]
38. de Visser RO, Nwamba C, Brearley E, Shafiei V, Hart L. Remote consultations in primary care: patient experiences and suggestions for improvement. J Health Psychol 2024 Oct;29(12):1321-1335. [doi: 10.1177/13591053241240383] [Medline: 38581309]
39. Koppel PD, De Gagne JC, Webb M, et al. Guidelines for rapport-building in telehealth videoconferencing: interprofessional e-Delphi study. JMIR Med Educ 2025 Aug 7;11:e76260. [doi: 10.2196/76260] [Medline: 40773683]
40. Thiyagarajan A, Grant C, Griffiths F, Atherton H. Exploring patients' and clinicians' experiences of video consultations in primary care: a systematic scoping review. BJGP Open 2020;4(1):bjgpopen20X101020. [doi: 10.3399/bjgpopen20X101020] [Medline: 32184212]
41. Carrillo de Albornoz S, Sia KL, Harris A. The effectiveness of teleconsultations in primary care: systematic review. Fam Pract 2022 Jan 19;39(1):168-182. [doi: 10.1093/fampra/cmab077] [Medline: 34278421]

XSL•FO
RenderX

42. Powell RE, Henstenburg JM, Cooper G, Hollander JE, Rising KL. Patient perceptions of telehealth primary care video visits. Ann Fam Med 2017 May;15(3):225-229. [doi: 10.1370/afm.2095] [Medline: 28483887]
43. Hamilton AB, Finley EP. Reprint of: Qualitative methods in implementation research: an introduction. Psychiatry Res 2020 Jan;283:112629. [doi: 10.1016/j.psychres.2019.112629] [Medline: 31735374]
44. Kitzinger J. Qualitative research. Introducing focus groups. BMJ 1995 Jul 29;311(7000):299-302. [doi: 10.1136/bmj.311.7000.299] [Medline: 7633241]
45. Guest G, Namey E, O'Regan A, Godwin C, Taylor J. Comparing interview and focus group data collected in person and online. : Patient-Centered Outcomes Research Institute (PCORI); 2020. [doi: 10.25302/05.2020.ME.1403117064]
46. Amelung D, Whitaker KL, Lennard D, et al. Influence of doctor-patient conversations on behaviours of patients presenting to primary care with new or persistent symptoms: a video observation study. BMJ Qual Saf 2020 Mar;29(3):198-208. [doi: 10.1136/bmjqs-2019-009485] [Medline: 31326946]

Abbreviations

COM-B: Capability, Opportunity, Motivation and Behavior model
COREQ: Consolidated Criteria for Reporting Qualitative Research
GP: general practitioner
SWCM: Shannon-Weaver communication model

Edited by A Stone; submitted 20.Jun.2025; peer-reviewed by C Gray, R Pittmann; revised version received 29.Nov.2025; accepted 31.Dec.2025; published 20.Jan.2026.

Please cite as:
Alboksmaty A, Lunova T, Darzi A, Neves AL
Communication Challenges and Mitigation Strategies in Primary Care Virtual Consultations: Qualitative Study
J Med Internet Res 2026;28:e79399
URL: https://www.jmir.org/2026/1/e79399
doi:10.2196/79399

© Ahmed Alboksmaty, Tetiana Lunova, Ara Darzi, Ana-Luisa Neves. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Gender Concordance and Patient Outcomes in Indian Telemedicine: Retrospective Cross-Sectional Quantitative Study of 286,000 Consultations

Nafisa Vaz[1], PhD; Vishalkumar Jani[2], PhD

[1]Healthcare Management Department, Goa Institute of Management, Sattari, Sanquelim, Goa, India

[2]Practo Technologies Pvt Ltd., Bangalore, Karnataka, India

Corresponding Author:
Nafisa Vaz, PhD
Healthcare Management Department, Goa Institute of Management, Sattari, Sanquelim, Goa, India

Abstract

Background: Gender concordance (GC) between patients and physicians has been linked to trust and satisfaction in traditional health care. However, its role in telemedicine, especially in culturally complex settings like India, is underexplored. In India's culturally diverse and gender-sensitive context, understanding GC becomes particularly relevant for specialties such as gynecology, dermatology, psychiatry, and urology, where discussions often involve intimate or stigmatized concerns. Despite rapid telemedicine expansion, little empirical evidence exists on whether GC affects patient-reported outcomes in this context.

Objective: This study examined whether GC significantly influences patient satisfaction and self-reported recovery in teleconsultations across India, with a focus on specialty-specific effects in culturally sensitive specialties.

Methods: We conducted a retrospective cross-sectional analysis of 286,196 anonymized teleconsultation records from a national telemedicine platform (January 2023–December 2024) spanning across 20 medical specialties using binary logistic regression. Records missing gender or satisfaction data were excluded from the analysis; recovery analyses included only consultations with completed day-21 follow-up surveys (n=1170, 0.4%). Outcomes included patient satisfaction (ratings 4‑5 on a five-point scale) and self-reported recovery at follow-up. Logistic regression models (Stata 17.0) tested associations between GC and outcomes, controlling for consultation time, duration, and physician experience. Subgroup analyses were conducted for the top 5 specialties. Each record included key data on consultation duration, timing, physician experience, specialty type, patient satisfaction rating, and self-reported recovery status. The study excluded the pediatrics specialty from the analysis to control for the parental bias.

Results: Of the 286,196 consultations, 164,008 (60.4%) were gender-concordant. Overall, 261,213 of 286,196 (91.3%) patients reported good satisfaction. GC had a statistically significant negative association with patient satisfaction (odds ratio [OR] 0.87, 95% CI 0.85‑0.90; $P<.001$). Across gender, the male doctor received higher satisfaction. In gynecology, female patient–female doctor pairs had significantly higher odds of reporting recovery (OR 4.53, 95% CI 0.8‑25.3; $P=.099$). Overall, consultation timing (OR 0.99, 95% CI 0.998‑0.999; $P<.001$) and patient satisfaction (OR 20.13, 95% CI 12.06‑35.38; $P<.001$) were stronger predictors of self-reported recovery than GC.

Conclusions: GC in telemedicine has a context-dependent impact. While it does not independently predict clinical recovery, it meaningfully shapes patient satisfaction. These findings highlight that gender sensitivity training and context-specific communication approaches may enhance telemedicine experiences in culturally sensitive domains. Integrating awareness of gender dynamics into telehealth design and policy could strengthen patient trust and engagement in virtual care. Future research should explore specialty-specific dynamics and improve follow-up response rates to better assess clinical outcomes.

(J Med Internet Res 2026;28:e78311) doi:10.2196/78311

KEYWORDS

gender concordance; telemedicine; India; patient satisfaction; clinical outcomes; culturally sensitive specialties; physician-patient dyads

XSL•FO
RenderX

Introduction

Telemedicine and Gender Dynamics in the Indian Context

The COVID-19 crisis has triggered a revolutionary change in health care delivery and expedited the worldwide penetration of telemedicine as an acceptable mode of consultation [1]. In India, with sharp health care access inequalities, especially across rural and socioeconomically disadvantaged populations, telemedicine presented itself as a vital mechanism for overcoming geographic and resource-based barriers [2,3]. Latest figures indicate a sweeping rise in virtual consults, with millions of patients depending on telehealth facilities for medical advice, diagnosis, and treatment [4].

Despite the operational success and scalability of telemedicine in India, patient experience and perceived quality are varying [5]. While aspects like connectivity, clinical competence, and usability of the platform are extensively researched [2,6], one aspect that is not well examined is the influence of gender in the construction of the virtual doctor-patient interaction. More specifically, the way the gender match or mismatch between patient and provider, also known as gender concordance (GC), has an effect in the Indian telemedicine context.

GC, alignment between the gender identity of the health care provider and the patient, has been linked to improved trust, communication, and outcomes, particularly in intimate specialties like gynecology and psychiatry [7]. But the impact of GC in telemedicine, a medium that lacks presence and is based primarily on audio-only interactions, is not well understood, particularly against the Indian backdrop [8].

India's sociocultural landscape presents an additional level of complexity. Strongly rooted gender norms of modesty and patriarchal formations tend to shape health care–seeking behavior [7-10]. Women will be reluctant to share certain symptoms with a male physician, and male patients will find female health care providers more empathic across certain specialties [11,12]. Such factors pose essential questions regarding the impact of GC on patient satisfaction and perceived effectiveness in telemedicine.

Recent telehealth studies [7,11,13] show that GC may influence trust and comfort differently across cultural contexts, but evidence from non-Western settings remains scarce. Few large-scale analyses have examined patient-reported outcomes in virtual care, especially in India. Previous studies in this area have been carried out in Western societies with varying cultural values and health care infrastructure [14,15]. Consequently, it is not possible to transfer findings from such studies directly to India. Moreover, studies so far have concentrated mostly on clinical results or communication patterns, with a lack of proper incorporation of patient-reported measures like satisfaction or self-perceived recovery. There is an utmost need for large-scale evidence from a contextualized setting that can be used for developing culture-sensitive design and implementation of telemedicine services for India.

This paper attempts to fill this lacuna by investigating whether GC is a predictor of 2-principal patient-reported outcomes, consultation satisfaction, and self-reported recovery, in telemedicine consultations across a broad spectrum of medical specialties in India. Using a database of more than 286,000 anonymous telemedicine consultations conducted by one of India's biggest telemedicine platforms, we explore the interaction between gender dynamics and other factors, including consultation timing, physician experience, and specialty type, in shaping patient perceptions of quality of care.

GC in Health Care: Global Evidence

GC, the match between a patient's and a provider's gender, has typically been linked with increased patient trust, satisfaction, and communication in conventional health care situations, physical face-to-face consultations, especially in developed economies [6,15,16]. The important point to note here is that literature about GC in clinical settings has been using self-reported sex or gender by the patients and doctors as a gender variable and does not really differentiate between gender and sex of a person like some other social science studies would have done. For instance, GC in primary care is related to longer consultation times and more patient-oriented communication [16], and, in oncology, female patients tend to be more comfortable with female providers because of impressions of empathy and emotional support [17].

Yet the results are still unclear. Previous studies [14,18] discovered minimal or no effect of GC on clinical outcomes or satisfaction, which suggests that GC by itself is not a strong predictor of health outcomes. One of the studies also stresses the differences in GC effects depending on the clinical specialty and setting in a related systematic review [15].

With the development of telemedicine, questions arise as to whether the impact of GC is present in the virtual setting. While nonverbal cues are diminished in teleconsultations, there is evidence that rapport and trust can still be established [19]. When voice-only consultation is used, as it is in resource-poor environments such as rural India, GC can facilitate comfort and disclosure by alleviating gender-based communication barriers [18]. However, the inconclusive global evidence underscores the necessity for context-oriented, culture-sensitive studies.

Sociocultural Context and Gender Preferences in India

India is a specific context to study GC because of the entrenched gender norms, especially in specialties with sensitive health matters [20]. Research reveals that female patients are likely to seek female providers, particularly in psychiatry and gynecology, where personal revelation is necessary [2,14]. Hofstede cultural dimensions, specifically masculinity-femininity and power distance, can explain why the health care expectations of the patients in India tend to be influenced by gendered constructs of authority and nurturing [21]. These sociocultural undercurrents can shape whether GC is likely to build trust or create new tensions.

Specialty-Specific Insights

GC's impact varies by specialty, widely preferred in gynecology for perceived empathy, considered in pediatrics, and supportive of disclosure in mental health, especially for women [6,8,21,22]. Those differences highlight the need to examine GC as a

XSL•FO
RenderX

specialty-specific rather than a general concept. For pediatric consultations, GC was determined based on the gender of the primary caregiver who initiated the consultation, minimizing potential bias from child accounts.

Empirical Gaps in Indian Telemedicine

Despite growing health care reach and diminished geographic barriers [1,3], telemedicine's capacity for supporting patient-provider relationships within gender-sensitive frameworks is uncertain. More satisfaction was evidenced in GC-based video interactions by Verma et al [23], but outcomes of recovery did not vary significantly. More critical factors, for example, consultation duration and doctor experience, could play an even more powerful role in shaping outcomes, and so analyses including these factors with GC as control variables are needed.

Empirical Lacunae in Indian Research

With the increasing utilization of telemedicine in India [5], there is still a scarcity of empirical evidence regarding the impact of GC on the delivery of health care virtually, especially from the patient's side. Female patients in rural and traditional areas can avoid or underdeclare their problems during interactions with male practitioners, particularly for reproductive or sexual health [9]. Previous studies [2,3,22,24] underscored the importance of system design with a gender awareness that ensures equal access. No large-scale, quantitative studies so far have thoroughly analyzed the effect of GC on both satisfaction and health outcomes in telemedicine in India.

With the mixed international evidence, this research will bridge the gaps in existing research by investigating the outcomes of GC for patient satisfaction and self-perceived recovery from teleconsultation in various medical specialties in India. The issue under investigation is relevant to both policy and practice of allocation algorithms, provider communication and behavior, and training needs for providers working for telemedicine platforms in India. The research is relevant to all telemedicine platforms—public and private.

Study Aims

This research is directed to assess the role of GC between patient and doctor in customer satisfaction (CSAT) and patient-reported medical efficacy (PRME), recovery rate, in the Indian telemedicine context. The study hopes to provide evidence-based insights into the telemedicine encounter as influenced by sociodemographic congruence between providers and patients. We expect the findings to have direct practical applications for telehealth platform design, provider training, and patient matchmaking algorithms. Guided by expectation confirmation theory (ECT) and the patient-centered care model (PCCM), this study examines whether GC predicts (1) patient satisfaction and (2) self-reported recovery in teleconsultations across India. We hypothesize that GC may have specialty-specific effects, particularly in gynecology.

Methods

Data Source and Study Design

This study draws on a dataset provided by one of the leading private Indian telemedicine platforms, recognized for its extensive reach and service diversity across specialties [25]. The platform was selected due to its nationwide footprint and volume of consultations, making it a strong representative sample for analyzing virtual care patterns across India. As of 2024, the platform reports over 400 million registered users across 22 states, positioning it among the largest telehealth providers in the country. The platform provides close to 2 million teleconsultations annually.

This retrospective cross-sectional study used secondary, anonymized data from a leading Indian telemedicine platform. The dataset covered consultations from January 2023 to December 2024.

Sample Selection and Data Processing

The dataset consists of 286,196 telemedicine consultations from various specialties and systems of medicine, such as Allopathic, Ayurveda, and Homeopathy. The current study used the anonymized data provided by the telemedicine platform. The organization has robust processes to anonymize the data and does not share any personally identifiable information, even within the organization. The data were transferred to authors in a spreadsheet where each transaction was identified with a unique identification number. For all purposes, the data were secondary in nature for the investigators.

The platform, over the period of 2 years, had done close to 3 million teleconsultations. However, for the study purposes, the platform shared only the data for consultations for which the patient had responded to the patient satisfaction survey (CSAT). The health outcome medical efficacy (PRME) survey started in April 2024, and hence it is available in very limited proportion. The inclusion of transactions was based on the fact that the CSAT or PRME survey was responded back by the patient because outcome data were important for the study. To utilize the provided data, authors had to filter the data. The transactions with reported CSAT and PRME were separated. Patients' and doctors' genders were reported, so GC was derived as a variable and utilized to categorize the transactions.

Dataset Screening Process

A structured screening process was applied to 286,196 for missing data and to form the analytical sample.

- Exclusion criteria for satisfaction analysis were as follows:
  - Missing patient gender (n=14,578)
  - Missing doctor gender (n=151)
  - Unclear satisfaction rating response (n=3)
- Exclusion criteria for recovery analysis were as follows:
  - All consultations without a completed day-21 recovery follow-up (n=284,874)
  - Pediatric consultations (n=29,288) due to caregiver-provided responses

XSL•FO
RenderX

Patients accessed the platform through a web or mobile interface. The current structure of the telemedicine platform automatically assigns a doctor to a patient based on the specialty or symptom that the patient enters into the system. An algorithm that is based on the availability of doctors on the platform at that time connects a patient with a doctor. So, there is no choice from either the doctor or patient side involved in matching.

The current study excludes the pediatrics consultations from the assessment of the impact of GC on the CSAT and PRME. The reason here is the response gets confounded by the parent or parents' preferences or biases. The patient himself or herself is not the one responding to either of the surveys.

Variables and Measures

Dependent Variables

Patient Satisfaction [csat_good]

This binary variable indicates whether the patient reported high satisfaction with the consultation. This variable is created from the 1-5 scale-based question asked at the end of each consultation. The ratings of 1, 2, and 3 were considered to be dissatisfied (csat_good=0), and the ratings of 4 and 5 were considered to be satisfied (csat_good=1).

PRME-Recovery Rate

This binary variable indicates whether the patient reported improvement in symptoms after consultation. This recovery rate question is floated to all patients on the 21st day after consultation. Patients who reported recovery are captured as prme_good=1, and 0 otherwise.

Independent Variables

GC [gender_cc]

This binary variable indicates whether the patient's gender matched the physician's gender. The platform collects this demographic information as part of preconsultation user inputs; patients self-report their sex when registering or booking a consultation. Physicians do not collect or verify this information during the consultation process. While the system technically captures biological sex, we acknowledge that much of the existing literature in this domain uses the term "GC" rather than "sex concordance." In line with prevailing academic convention, we continue using the term GC throughout this study, while recognizing this conceptual limitation in our definition of the research question.

Time of Day [time_day]

This captures consultation timing, categorized into morning and night. The platform uses a 9 AM to 9 PM window as usual day timing from the doctor's practice perspective. Hence, consultation happening in this time period is marked as time_day=1.

Consultation Duration [total_dura]

This was the total duration [in seconds] of the teleconsultation.

Physician Experience [dr_experience]

This was measured in years of practice.

Data Analysis

The study started analysis with the finding of proportions of transactions reported with positive CSAT and positive PRME across patient-doctor gender-concordant dyads and nonconcordant dyads. The study employed a *z* test to check if the 2 groups were statistically different in CSAT or PRME.

To assess the impact of GC on patient outcomes, the study employed a logistic regression model. Exponentiated coefficients (odds ratios [OR]) were reported to interpret the impact of GC and other control variables. The analysis also included the regressions at the specialty level. These specialties were selected by consultation volume. These were identified using the metric of total number of completed consultations within the study period. Logistic regression analysis requires an adequate number of outcome events relative to the number of predictor variables to ensure the stability and reliability of parameter estimates. Following conventional guidelines, a minimum of 10 outcome events per predictor variable is recommended to avoid overfitting and ensure model validity [24]. Additionally, the Central Limit Theorem supports a minimum sample size of 30 as a general rule of thumb for approximating normality in sampling distributions [26]. These principles jointly informed the decision to set 30 observations as the minimum threshold for conducting specialty-specific regression analyses in this study. This gave us 9 specialties for CSAT and 7 specialties for PRME. However, the result tables reported only the 5 specialties that had the highest number of responses available for CSAT or PRME. Analyses were conducted using Stata (version 17.0; StataCorp LLC). Binary logistic regression models assessed associations between GC and 2 outcomes, patient satisfaction, and self-reported recovery, while controlling for consultation duration, physician experience, and time of day. Specialty-level regressions were conducted where event counts were ≥30.

Ethical Considerations

This study was approved by the Board of Research Ethics (BORE) of Goa Institute of Management (Approval No. GIM/BHCM042507). The study was conducted in compliance with ethical research standards. Since the data were anonymized and obtained with necessary permissions from the telemedicine platform, no personally identifiable information was accessed. The research adheres to data protection laws and ethical guidelines related to patient confidentiality. This retrospective study was approved by the institutional review board of Goa Institute of Management. Data used were deidentified secondary records collected with informed consent for service use; the platform's terms permit anonymized secondary analyses. All identifiers were removed prior to researcher access; only aggregate data were analyzed. No compensation was involved. No identifiable images or patient details are included.

Results

Overview of the Sample

Of the 286,196 consultations analyzed, 60.4% (n=164,008) were gender concordant and 47.8% (n=129,848) involved female patients. Table 1 shows the specialty-wise shares in total

XSL•FO
RenderX

transactions. Pediatrics is shown in descriptive statistics but excluded from outcome analyses due to caregiver response bias.

XSL•FO
RenderX

Table . Summary of data and variables.

Number of teleconsultations	Count (N=286,196)
Patient age (y), median (IQR)	30.0 (25-37)
Patient sex, n (%)	
Male	141,670 (52)
Female	129,848 (48)
Doctor experience (y), median (IQR)	8.0 (5-12)
Doctor sex, n (%)	
Male	169,363 (59)
Female	116,682 (41)
Specialty, n (%)	
Ayurveda	2021 (0.7)
Cancer	600 (0.2)
Cardiology	5290 (1.8)
Dental	3329 (1.2)
Dermatology	38,824 (14)
Diabetology	3729 (1.3)
Diet and nutrition	6069 (2.1)
Ear, Nose, Throat	11,282 (3.9)
Eye and Vision	5506 (1.9)
Gastroenterology	5654 (2)
General physician	65,794 (23)
General surgery	967 (0.3)
Gynecology	37,876 (13)
Homeopathy	2754 (1)
Nephrology	1643 (0.6)
Neurology	7212 (2.5)
Orthopedics	11,871 (4.1)
Pediatrics	29,288 (10)
Physiotherapy	1427 (0.5)
Psychiatry	5688 (2)
Psychological counseling	3993 (1.4)
Pulmonology	3216 (1.1)
Rheumatology	777 (0.3)
Sexology	10,192 (3.6)
Stomach and digestion	10,176 (3.6)
Urology	9798 (3.4)
Unknown	1220 (0.4)
Time of consultation, n (%)	
Peak hours—day (9 AM to 9 PM)	193,612 (68)
Nonpeak hours—night (9 PM to 9 AM)	92,583 (32)
Consultation duration (s), median (IQR)	382 (237-600)
CSAT[a] score (1-5), n (%)	
1	12,461 (4.4)

XSL•FO
RenderX

Number of teleconsultations	Count (N=286,196)
2	3732 (1.3)
3	8787 (3.1)
4	39,024 (14)
5	222,191 (78)
CSAT (good vs bad), n (%)	
Bad	24,980 (8.7)
Good	261,213 (91)
PRME[b] responses, n (%)	1322 (0.46)
PRME (good vs bad), n (%)	
Not recovered	322 (24)
Recovered	1000 (76)
GC[c], n (%)	
Discordant	107,369 (40)
Concordant	164,008 (60)

[a]CSAT: customer satisfaction.

[b]PRME: patient-reported medical efficacy.

[c]GC: gender concordance.

Table 2 shows that 135,755 out of 148,803 (91.23%) transactions with a gender-concordant patient-doctor dyad reported good CSAT compared with 86,680 out of 93,971 (92.24%) transactions where doctor-patient GC was not present. Even when the sample was categorized for patient gender, it showed similar results. Gender-concordant transactions showed lower CSAT compared with nonconcordant transactions.

Table . Statistical difference in CSAT[a] and PRME[b] across different groups.

	CSAT			PRME		
Group	GC[c]=0, n/N (%)	GC=1, n/N (%)	*z* score (*P* value)	GC=0, n/N (%)	GC=1, n/N (%)	*z* score (*P* value)
All	86,680/93,971 (92.24)	135,755/148,803 (91.23)	8.753 (<.001)	389/502 (77.49)	495/669 (73.99)	1.38 (.17)
Female patients	51,511/55,808 (92.3)	55,057/61,003 (90.25)	12.358 (<.001)	233/304 (76.64)	203/261 (77.78)	–0.32 (.75)
Male patients	35,169/38,163 (92.19)	80,698/87,800 (91.87)	1.469 (.14)	156/198 (78.79)	292/408 (71.57)	1.9 (.06)

[a]CSAT: customer satisfaction.

[b]PRME: patient-reported medical efficacy.

[c]GC: gender concordance.

For PRME, the available transactions were only 1322, and after excluding the pediatrics specialty, the available transactions were 1199. Out of the available data on PRME, 923 of 1199 (77.01%) transactions have good PRME; that is, 77.01% patients reported that they had recovered. Out of these 1199 transactions, 1170 transactions had data on patient and doctor gender. Among these, 669 (57.17%) transactions had patient-doctor GC. Table 2 shows that gender-concordant cohort of transactions reported that PRME is 495 out of 669 (73.99%), compared with 388 out of 501 (77.49%) in the nonconcordant cohort. This difference is not statistically significant, unlike what we found with CSAT. Patient gender-wise categorization also found insignificant difference between concordant and nonconcordant cohorts. However, for the female patient cohort, the gender-concordant group reported higher PRME compared to the nonconcordant group. In contrast, it was the other way around for the male patient cohort.

Table 3 shows that, on average, male patients report higher satisfaction compared with their female counterparts, and male doctors receive better satisfaction scores compared with female doctors. However, irrespective of patient gender, a gender-discordant dyad results in a better CSAT score.

XSL•FO
RenderX

Table . CSAT[a] for different doctor genders as per patient genders.

	Patient gender		
Doctor gender	All patients, n/N (%)	Female patients, n/N (%)	Male patients, n/N (%)
All doctors	222,564/242,914 (91.6)	106,681/116,935 (91.23)	115,883/125,979 (91.98)
Female doctors	12,398/13,713 (90.41)	7678/8592 (89.36)	4720/5121 (92.17)
Male doctors	210,166/229,201 (91.69)	99,003/108,343 (91.38)	111,163/120,858 (91.98)

[a]CSAT: customer satisfaction.

GC and Patient Satisfaction

This section is focused on the logistic regression results assessing the impact of GC and other independent variables on the CSAT. The study employed logistic regression on the overall sample and separately for 9 specialties individually. Table 4 reports the regression results for the overall sample and 5 specialties that had the highest number of transactions with CSAT reported.

Table . Relationship between CSAT[a] and gender concordance in telemedicine.

	csat_good					
Variable	All, OR[b] (95% CI; *P* value)	General physician, OR (95% CI; *P* value)	Gynecology, OR (95% CI; *P* value)	Dermatology, OR (95% CI; *P* value)	Orthopedics, OR (95% CI; *P* value)	ENT, OR (95% CI; *P* value)
Gender concordance	0.874 (0.81-0.94; <.001)	0.9796 (0.92-1.04; .49)	1.079 (0.97-1.20; .16)	0.917 (0.85-0.98; .02)	1.023 (0.89-1.18; .74)	0.841 (0.73-0.97; .02)
Consultation during office hours	1.081 (1.04-1.12; <.001)	1.114 (1.05-1.18; <.001)	1.038 (0.97-1.11; .31)	1.052 (0.98-1.13; .19)	0.867 (0.74-1.02; .08)	1.229 (1.06-1.42; <.001)
Duration of the consultation	1.0004 (1.0003-1.0005; <.001)	1.0003 (1.0002-1.0004; <.001)	1.001 (1.001-1.001; <.001)	1.001 (1.001-1.001; <.001)	1.000 (1.00-1.00; .05)	1.000 (1.00-1.00; .09)
Doctor experience	1.002 (1.00-1.002; .11)	1.003 (1.00-1.01; .06)	1.001 (1.00-1.01; .63)	0.995 (0.99-1; .16)	0.955 (0.93-1; <.001)	0.985 (0.97-1; .008)
Intercept	9.46 (8.55-10.47; <.001)	10.235 (8.96-11.69; <.001)	5.5912 (4.74-6.6; <.001)	8.6154 (7.28-10.2; <.001)	16.6125 (13.43-20.55; <.001)	11.9619 (10.27-13.93; <.001)
N	228,640	65,174	36,185	38,183	11,770	11,213

[a]CSAT: customer satisfaction.

[b]OR: odds ratio.

In the full sample, gender discordance was a statistically significant predictor of patient satisfaction (OR 0.874, 95% CI 0.85 - 0.90; *P*<.001), suggesting that the presence of GC results in a lower satisfaction rate. This aligns with the findings of Table 2, showing that female patient-male doctor and male patient-female doctor dyads showed higher satisfaction rates.

At the specialty level, GC was associated with lower satisfaction in dermatology (OR 0.917, 95% CI 0.85-0.98; *P*=.02) and ENT (OR 0.841, 95% CI 0.73-0.97; *P*=.02). In the specialty of Gynecology, where patient sensitivity about doctor gender is high [27,28], GC showed a positive but statistically nonsignificant association with satisfaction. The effect of GC was not uniform, suggesting that specialty context moderates its influence on patient experience.

Except for orthopedics, all other specialties showed that transaction happening during 9 AM to 9 PM had a positive impact on the CSAT. Except for the general physician group, other specialties showed that the younger (less experienced) the doctor, the better the CSAT rate.

GC and PRME

Table 5 reports the logistic regression results for the impact of GC on the PRME—the recovery rate. It presents 7 regressions: the first 2 for the whole sample for which all required variables were available, and the next 5 were separate specialty levels. The first model included CSAT as one of the explanatory variables, whereas the other 6 models did not include it.

XSL•FO
RenderX

Table . Relationship between patient-reported medical efficacy and gender concordance in telemedicine.

Variable	prme_good						
	All, OR[a] (95% CI; *P* value)	All, OR (95% CI; *P* value)	General physician, OR (95% CI; *P* value)	Gynecology, OR (95% CI; *P* value)	Dermatology, OR (95% CI; *P* value)	Orthopedics, OR (95% CI; *P* value)	ENT, OR (95% CI; *P* value)
Customer satisfaction	20.1334 (8.11-50.02; <.001)	__[b]	—	—	—	—	—
Gender concordance	0.8492 (0.63-1.15; .29)	0.8401 (0.64-1.01; .21)	0.855 (0.51-1.43; .55)	4.53 (0.75-27.26; .099)	1.102 (0.6-2.03; .76)	1.168 (0.33-4.11; .81)	1.733 (0.62-4.83; .29)
Consultation during office hours	1.142 (0.82-1.59; .43)	1.2524 (0.93-1.69; .14)	1.492 (0.84-2.64; .17)	2.099 (0.77-5.75; .15)	1.324 (0.68-2.57; .41)	0.423 (0.09-2.09; .29)	0.716 (0.23-2.28; .57)
Duration of the consultation	0.999 (0.99-1.00; <.001)	0.999 (0.99-1.00; .001)	0.999 (0.99-1.00; .18)	1.001 (1.00-1.002; .18)	1 (0.99-1.002; .95)	0.998 (0.995-1; .17)	0.998 (0.99-1; .19)
Doctor experience	1.0032 (0.98-1.02; .75)	1.007 (0.99-1.03; .44)	0.998 (0.97-1.02; .87)	1.09 (1.00-1.19; .04)	0.965 (0.91-1.02; .20)	0.999 (0.81-1.19; .99)	1.054 (0.97-1.14; .19)
Intercept	0.2854 (0.2-0.42; <.001)	3.5164 (2.4-5.15; <.001)	5.2211 (3.16-8.62; <.001)	0.1725 (0.02-1.58; .12)	3.448 (1.47-8.1; .005)	6.9175 (0.00-9.98; .91)	2.7454 (0.68-11.06; .16)
N	1170	1170	430	114	243	46	79

[a]OR: odds ratio.

[b]not applicable.

Based on the 1170 transactions, GC had no statistically significant effect on overall recovery outcomes (OR 0.84, 95% CI 0.64 - 1.1; *P*=.21). While marginal significance was noted in gynecology (*P*=.099), given the small number of transactions (n=1170) for recovery follow-ups, findings must be interpreted cautiously.

The first model showed that patient satisfaction was the strongest and most consistent predictor of self-reported recovery (OR 20.13, 95% CI 12.06 - 35.38; *P*<.001).

Longer consultations were weakly correlated with satisfaction, not recovery. The experience of a doctor did not show a positive impact on the patient-reported recovery rate, except in the gynecology specialty.

Discussion

Principal Findings

This study examined the impact of GC on patient satisfaction and self-reported recovery in more than 286,000 Indian telemedicine consultations. Contrary to conventional assumptions [7], GC did not uniformly enhance satisfaction or recovery. In fact, nonconcordant dyads reported higher satisfaction, challenging existing evidence and expectations of demographic matching improving health care experiences [6,15,16].

These findings align with 2 theories: ECT and PCCM.

ECT explains how satisfaction arises when experiences exceed expectations [26,27], as seen in male patient-female doctor dyads. In concordant dyads, higher expectations may have gone unmet, reducing perceived trust [29]. Multiple studies indicate that male patients interacting with female physicians frequently experience their expectations being exceeded, particularly in the domains of interpersonal communication and empathy [30], leading to higher trust. Research reveals that male patient-female doctor dyads are often characterized by higher patient-reported satisfaction and quality scores, largely due to the “expectation-surprise” effect: male patients, who may hold lower initial expectations of female physicians’ technical and interpersonal abilities, report greatly enhanced perceptions when those expectations are surpassed in real clinical interactions [13]. Female doctors are consistently highlighted for their superior listening, attentiveness, and clarity, which not only challenge prevailing gender stereotypes but also foster increased trust and comfort among male patients. This dynamic has been observed across different care settings and cultural contexts, underscoring the potential benefits of such cross-gender provider-patient pairings for improving patient satisfaction and health care quality [31]. These superseded expectations may meet with better treatment adherence, leading to significantly different recovery rates in male patient-female doctor dyads compared to male-male dyads.

The female patient, and the overall sample, showed higher satisfaction rates when consulted with a male doctor, which is in corroboration with general perception of men being preferred for technical roles [32] and previous evidence of this female patients rating male doctors better than the female doctors [33].

PCCM emphasizes aligning care with patient preferences, values, and context [23,34]. In telemedicine, where nonverbal cues are limited, empathy, trust, and communication quality become more crucial than demographic similarity. This helps explain why patient satisfaction, rather than GC, was the strongest predictor of self-reported recovery [17]. Figure 1 illustrates how these 2 broad theories influence patient satisfaction and self-reported health outcomes through communication and trust.

XSL•FO
RenderX

Figure 1. Theoretical framework.

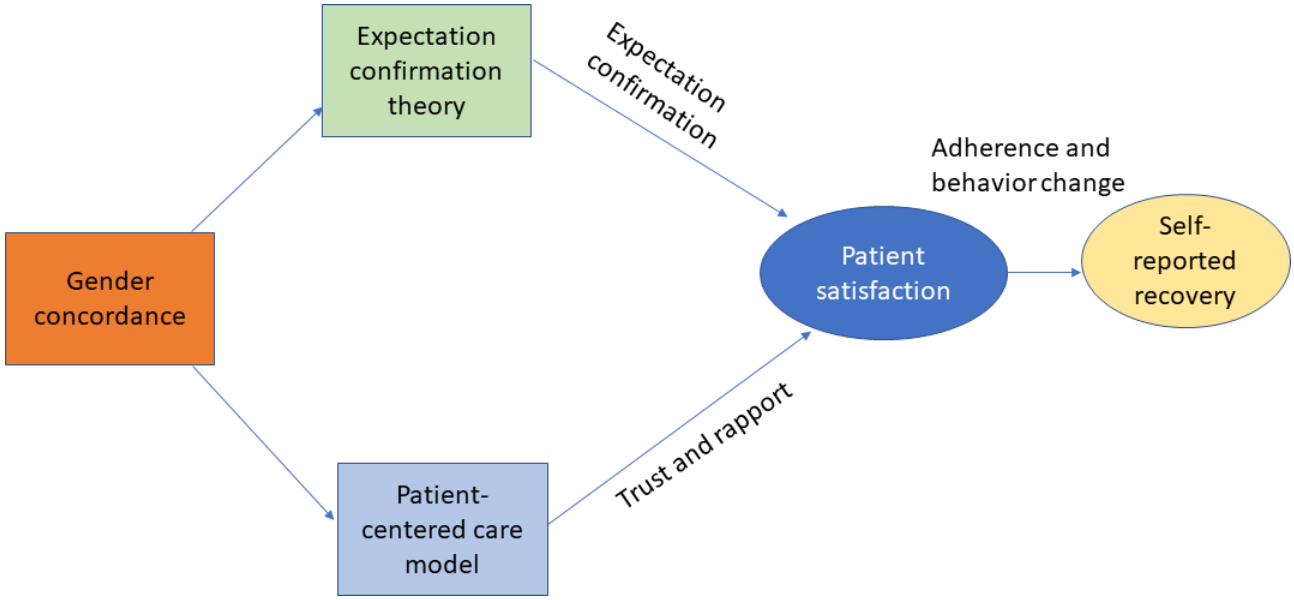

Comparison With Prior Work

These findings suggest telemedicine platforms, especially catering to Indian clients, should not default to GC-based matching. Instead, emphasis should be placed on enhancing communication skills and digital empathy, particularly in sensitive specialties such as gynecology. Male doctors are considered to be good at the technical skills and, hence, the trust and satisfaction [32,33]; and female doctors turn out to be more empathetic and better at communication, resulting in higher satisfaction [16,17]. Unlike evidence from the developed world, in the case of the Indian telemedicine context, GC is not playing any significant role. Hence, the training in active listening and patient engagement can be preferred tools over demographic alignment through matching algorithms [35].

Moreover, patient preferences, specialty, and combinations of them play an important role in final satisfaction and health outcome—recovery rate. The patient preferences should guide provider matching, particularly in specialties involving intimate care, without enforcing rigid defaults.

Strengths and Limitations

This study is among the first to explore GC and patient-reported outcomes using a very large real-world dataset from India, covering multiple specialties. Its strength lies in its scale, cross-specialty comparisons, and integration of both satisfaction and recovery indicators.

However, several limitations must be noted. First, PRME data represented <1% of consultations, limiting statistical power. Second, pediatric cases may involve caregiver responses, introducing bias; hence, the study could not include it in the analysis. Third, recall and nonresponse bias could affect postconsultation surveys. Fourth, this was a single-platform dataset, limiting generalizability. Moreover, generalizability to a very different country would also be limited, as the current evidence suggests that the Indian context behaved differently compared to Western evidence. Fifth, gender identity was operationalized as binary sex due to data limitations. The current findings are presented keeping in mind all these limiting factors. So, there is a need for further research, as described previously.

Future Directions

Further research is warranted to explore intersectional factors and evaluate objective health outcomes beyond self-reports. There is a definite need to also assess similar hypotheses of GC on health outcomes in physical outpatient department settings for understanding Indian evidence. Specialty-level research and insights would help inform not only patients but also the current medical students choosing their specialization and super-specialization. Future research should replicate this analysis across multiple telemedicine platforms, incorporating both quantitative and qualitative approaches to capture patients' subjective experiences. Including gender identity, language preference, and digital literacy could offer richer insights into patient-provider rapport.

Longitudinal or experimental designs could further assess how targeted interventions, such as empathy training or communication workshops for physicians, impact satisfaction and clinical outcomes.

Conclusions

This study examined how matching the gender of patients and doctors influenced satisfaction and patient-reported recovery rate, drawing on data from more than 286,000 telemedicine consultations across various medical specialties in India. The GC did not consistently predict recovery but did influence satisfaction. In fact, at the overall level, GC between patients and doctors resulted in lower satisfaction.

Patient satisfaction emerged as the strongest factor linked to self-reported recovery, underscoring how crucial trust, empathy, and clear communication are in virtual consultations [36]. However, because only a small fraction of consultations, less than 1% of the total data, included recovery data, conclusions drawn about health outcomes should be viewed with caution.

These findings suggest that Indian telemedicine platforms should prioritize strategies that enhance communication skills, cultural

XSL•FO
RenderX

sensitivity, and patient education. Indian patients apparently showed that gender-discordant dyads would be more effective. In specialties such as gynecology, where GC showed more pronounced effects on health outcomes, customizable matching features or patient-preference options may be warranted.

The current study has implications for telemedicine platforms, especially those catering to the Indian population. The evidence suggests that, unlike the Western world, patient satisfaction and health outcomes are not influenced by the GC; hence, there is no support for matching algorithms to consider this as a variable. The doctors need to be aware of varying preferences as per patient gender. There is also a case of further research on specialty-level evidence that may inform future doctors to choose their specializations. This may lead to a more effective patient-centric health care delivery.

In summary, while GC is not a major factor across the board in Indian telemedicine, it does matter in specific contexts. Its impact depends on the medical specialty, cultural expectations, and the quality of communication, highlighting the importance of personalized approaches to patient-provider interactions.

Acknowledgments

No generative AI tools were used in drafting or revising this manuscript.

All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

The anonymized data that support this study are available from the corresponding author upon reasonable request but are not publicly shared due to confidentiality agreements with the telemedicine provider. The datasets of this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: NV, VJ
Data curation: VJ, NV
Formal analysis: VJ, NV
Writing – original draft: NV, VJ
Writing – review & editing: NV, VJ

Conflicts of Interest

None declared.

References

1. Perrin PB, Pierce BS, Elliott TR. COVID-19 and telemedicine: a revolution in healthcare delivery is at hand. Health Sci Rep 2020 Jun;3(2):e166. [doi: 10.1002/hsr2.166] [Medline: 32500101]
2. Kruse CS, Krowski N, Rodriguez B, Tran L, Vela J, Brooks M. Telehealth and patient satisfaction: a systematic review and narrative analysis. BMJ Open 2017 Aug 3;7(8):e016242. [doi: 10.1136/bmjopen-2017-016242] [Medline: 28775188]
3. Sharma P, Rao S, Krishna Kumar P, et al. Barriers and facilitators for the use of telehealth by healthcare providers in India-A systematic review. PLoS Digit Health 2024 Dec;3(12):e0000398. [doi: 10.1371/journal.pdig.0000398] [Medline: 39642108]
4. Mehrotra A, Ray K, Brockmeyer DM, Barnett ML, Bender JA. Rapidly converting to “virtual practices”: outpatient care in the era of COVID-19. NEJM Catal 2020 Apr 1;1(2). [doi: 10.1056/CAT.20.0091]
5. Aneja J, Goyal T, Verma M, Kaur G, Mirza M, Gupta S. Client satisfaction with telemedicine services during COVID-19 pandemic: a cross-sectional survey from a teaching institute of North India. J Family Med Prim Care 2022 Sep;11(9):5187-5193. [doi: 10.4103/jfmpc.jfmpc_2217_21] [Medline: 36505639]
6. Kaur KN, Niazi F, Thakur R, Saeed S, Rana S, Singh H. Patient satisfaction for telemedicine health services in the era of COVID-19 pandemic: a systematic review. Front Public Health 2022;10:1031867. [doi: 10.3389/fpubh.2022.1031867] [Medline: 36589954]
7. Kitole FA, Ali Z, Song J, et al. Exploring the gender preferences for healthcare providers and their influence on patient satisfaction. Healthcare (Basel) 2025 May 5;13(9):1063. [doi: 10.3390/healthcare13091063] [Medline: 40361841]
8. Pinedo-Torres I, Garcia-Villasante EJ, Gutierrez-Ortiz CC, et al. The doctor-patient relationship and barriers in non-verbal communication during teleconsultation in the era of COVID-19: a scoping review. F1000Res 2023 Jun 15;12:676. [doi: 10.12688/f1000research.129970.1]

XSL•FO
RenderX

9. Banerjee A, Sanyal D. Dynamics of doctor-patient relationship: a cross-sectional study on concordance, trust, and patient enablement. J Family Community Med 2012 Jan;19(1):12-19. [doi: 10.4103/2230-8229.94006] [Medline: 22518353]
10. Patel V, Rodrigues M, DeSouza N. Gender, poverty, and postnatal depression: a study of mothers in Goa, India. Am J Psychiatry 2002 Jan;159(1):43-47. [doi: 10.1176/appi.ajp.159.1.43] [Medline: 11772688]
11. Martinez KA, Rothberg MB. Physician gender and its association with patient satisfaction and visit length: an observational study in telemedicine. Cureus 2022 Sep;14(9):e29158. [doi: 10.7759/cureus.29158] [Medline: 36258932]
12. Dash S, Aarthy R, Mohan V. Telemedicine during COVID-19 in India-a new policy and its challenges. J Public Health Policy 2021 Sep;42(3):501-509. [doi: 10.1057/s41271-021-00287-w] [Medline: 34012012]
13. Si Y, Chen G, Zhou Z, Yip W, Chen X. The impact of physician-patient gender match on healthcare quality: an experiment in China. Soc Sci Med 2025 Sep;380:118166. [doi: 10.1016/j.socscimed.2025.118166] [Medline: 40451062]
14. Greenwood BN, Carnahan S, Huang L. Patient-physician gender concordance and increased mortality among female heart attack patients. Proc Natl Acad Sci U S A 2018 Aug 21;115(34):8569-8574. [doi: 10.1073/pnas.1800097115] [Medline: 30082406]
15. Lau ES, Hayes SN, Volgman AS, et al. Does patient-physician gender concordance influence patient perceptions or outcomes? J Am Coll Cardiol 2021 Mar 2;77(8):1135-1138. [doi: 10.1016/j.jacc.2020.12.031] [Medline: 33632488]
16. Kreps GL. Doctors talking with patients/patients talking with doctors: improving communication in medical visits. Health Commun 1995 Jan;7(1):67-71. [doi: 10.1207/s15327027hc0701_5]
17. Street RL, Makoul G, Arora NK, Epstein RM. How does communication heal? Pathways linking clinician-patient communication to health outcomes. Patient Educ Couns 2009 Mar;74(3):295-301. [doi: 10.1016/j.pec.2008.11.015] [Medline: 19150199]
18. Plichta JK, Williamson H, Sergesketter AR, et al. It's not you, It's me: the influence of patient and surgeon gender on patient satisfaction scores. Am J Surg 2020 Nov;220(5):1179-1188. [doi: 10.1016/j.amjsurg.2020.07.036] [Medline: 32847689]
19. Elsamadicy AA, Reddy GB, Nayar G, et al. Impact of gender disparities on short-term and long-term patient reported outcomes and satisfaction measures after elective lumbar spine surgery: a single institutional study of 384 patients. World Neurosurg 2017 Nov;107:952-958. [doi: 10.1016/j.wneu.2017.07.082] [Medline: 28743671]
20. Eggermont D, Kunst AE, Groenewegen PP, Verheij RA. Social concordance and patient reported experiences in countries with different gender equality: a multinational survey. BMC Prim Care 2024 Mar 23;25(1):97. [doi: 10.1186/s12875-024-02339-y] [Medline: 38521895]
21. Hofstede G. Dimensionalizing cultures: the Hofstede model in context. Online Read Psychol Cult 2011 Dec 1;2(1). [doi: 10.9707/2307-0919.1014]
22. Gideon J, Asthana S, Bisht R. Health systems in India: analysing barriers to inclusive health leadership through a gender lens. BMJ 2024 Jul 17;386:e078351. [doi: 10.1136/bmj-2023-078351] [Medline: 39019544]
23. Verma N, Buch B, Taralekar R, Acharya S. Diagnostic concordance of telemedicine as compared with face-to-face care in primary health care clinics in rural India: randomized crossover trial. JMIR Form Res 2023 Jun 23;7:e42775. [doi: 10.2196/42775] [Medline: 37130015]
24. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996 Dec;49(12):1373-1379. [doi: 10.1016/s0895-4356(96)00236-3] [Medline: 8970487]
25. Bagga B. Telemedicine in india: tech-driven democratization of healthcare. APAC News Network. 2024. URL: https://apacnewsnetwork.com/2024/09/telemedicine-in-india-tech-driven-democratization-of-healthcare/ [accessed 2026-01-02]
26. Kwak SG, Kim JH. Central limit theorem: the cornerstone of modern statistics. Korean J Anesthesiol 2017 Apr;70(2):144-156. [doi: 10.4097/kjae.2017.70.2.144] [Medline: 28367284]
27. Nimbi FM, Galizia R, Rossi R, et al. The biopsychosocial model and the sex-positive approach: an integrative perspective for sexology and general health care. Sex Res Soc Policy 2022 Sep;19(3):894-908. [doi: 10.1007/s13178-021-00647-x]
28. Aromaa A, Kero K, Grönlund J, et al. Let's talk about sexuality - a web-based survey of self-reported competence in sexual problems among obstetrician-gynecologists in Finland. Acta Obstet Gynecol Scand 2023 Feb;102(2):190-199. [doi: 10.1111/aogs.14492] [Medline: 36515100]
29. Zhu L, Jiang X, Cao J. Factors affecting continuance intention in non-face-to-face telemedicine services: trust typology and privacy concern perspectives. Healthcare (Basel) 2023 Jan 28;11(3):374. [doi: 10.3390/healthcare11030374] [Medline: 36766949]
30. Madanay F, Bundorf MK, Ubel PA. Physician gender and patient perceptions of interpersonal and technical skills in online reviews. JAMA Netw Open 2025 Feb 3;8(2):e2460018. [doi: 10.1001/jamanetworkopen.2024.60018] [Medline: 39951262]
31. Derose KP, Hays RD, McCaffrey DF, Baker DW. Does physician gender affect satisfaction of men and women visiting the emergency department? J Gen Intern Med 2001 Apr;16(4):218-226. [doi: 10.1046/j.1525-1497.2001.016004218.x] [Medline: 11318922]
32. AlSamhori JF, Rayyan R, Hammouri M, et al. Factors influencing gender preference towards surgeons among Jordanian adults: an investigation of healthcare bias. Sci Rep 2023 Jul 18;13(1):11614. [doi: 10.1038/s41598-023-38734-1] [Medline: 37464087]

XSL•FO
RenderX

33. Wiltshire J, Allison JJ, Brown R, Elder K. African American women perceptions of physician trustworthiness: a factorial survey analysis of physician race, gender and age. AIMS Public Health 2018;5(2):122-134. [doi: 10.3934/publichealth.2018.2.122] [Medline: 30094275]
34. Mead N, Bower P. Patient-centredness: a conceptual framework and review of the empirical literature. Soc Sci Med 2000 Oct;51(7):1087-1110. [doi: 10.1016/s0277-9536(00)00098-8] [Medline: 11005395]
35. Roter D, Hall JA. Doctor-patient communication: why and how communication contributes to the quality of medical care. In: Gellman MD, Turner JR, editors. Encyclopedia of Behavioral Medicine: Springer International Publishing; 2020:693-698. [doi: 10.1007/978-3-030-39903-0_1223]
36. Jameel A, Sahito N, Guo W, Khan S. Assessing patient satisfaction with practitioner communication: patient-centered care, hospital environment and patient trust in the public hospitals. Front Med (Lausanne) 2025;12:1544498. [doi: 10.3389/fmed.2025.1544498] [Medline: 40470044]

Abbreviations

CSAT: customer satisfaction
ECT: expectation confirmation theory
GC: gender concordance
OR: odds ratio
PCCM: patient-centered care model
PRME: patient-reported medical efficacy

Edited by A Schwartz, TDA Cardoso; submitted 30.May.2025; peer-reviewed by A Saxena, D Verran; revised version received 14.Nov.2025; accepted 24.Nov.2025; published 20.Jan.2026.

Please cite as:
Vaz N, Jani V
Gender Concordance and Patient Outcomes in Indian Telemedicine: Retrospective Cross-Sectional Quantitative Study of 286,000 Consultations
J Med Internet Res 2026;28:e78311
URL: https://www.jmir.org/2026/1/e78311
doi:10.2196/78311

© Nafisa Vaz, Vishalkumar Jani. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Virtual Delivery of Early Psychosis Care: Retrospective Cohort Study of Factors Associated With Initial Engagement

Trinity Vey[1,2], MSc; Nicole Kozloff[1,3,4], SM, MD; George Foussias[1,3], MD, PhD; Wei Wang[1,3], PhD; Albert HC Wong[1,3], MD, PhD; Aristotle N Voineskos[1,3], MD, PhD; Nicole Davis-Faroque[1,3], MD; Alexia Polillo[1,3,4], PhD

[1]Centre for Addiction and Mental Health, 1025 Queen Street West, Fourth Floor, Toronto, ON, Canada

[2]Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

[3]Department of Psychiatry, University of Toronto, Toronto, ON, Canada

[4]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Alexia Polillo, PhD
Centre for Addiction and Mental Health, 1025 Queen Street West, Fourth Floor, Toronto, ON, Canada

Abstract

Background: There is evidence that virtual delivery of early psychosis intervention (EPI) is well received by youth and has benefits such as reported improvements in accessibility, convenience, and comfort; however, potential barriers remain, including the digital divide and privacy concerns. Although initial engagement in EPI services is important for long-term recovery, little is known about initial engagement in the context of virtual care and the role of health equity and service use factors.

Objective: This study aimed to identify factors associated with attendance at the initial EPI consultation appointment when most were being delivered virtually.

Methods: This retrospective cohort study used electronic medical record data from patients aged 16 to 29 years who were referred to a large EPI program. The EPI program received 301 unique referrals that met study eligibility criteria from April to December 2020. Self-reported demographic variables were derived from the Centre for Addiction and Mental Health's structured health equity form and included age, gender, racial and ethnic group, country of birth, and sexual orientation. Service use factors derived from clinical documentation included referral source, days to consultation, and attendance at the consultation appointment, which was the primary outcome. Comparisons were made with 2018 to 2019 data from 999 participants from the same site prior to virtual care implementation using chi-square tests for categorical variables, independent *t* tests for continuous variables, and binary logistic regression.

Results: Patients had a mean age of 23.2 (SD 3.3) years, and 214 (71.1%) participants identified as male. Compared with pre–virtual care, there were significantly higher rates of inpatient referral (114/301, 37.9%) and lower rates of referral from outpatient and other providers (122/301, 40.5%) post–virtual care (χ^2_2=18.7, P<.001), with a small effect size and moderately narrow CI (Cramér V=0.120, 95% CI 0.06 to 0.17). Following univariable tests and stepwise backward selection, identifying as Black (odds ratio 0.45, 95% CI 0.21 to 0.97) and being referred from the emergency department or bridging clinic (odds ratio 0.24, 95% CI 0.08 to 0.72) were associated with decreased odds of attendance at the consultation appointment in the final adjusted model. All tests were 2-sided with an α level of .05.

Conclusions: This study is innovative in that it examines the self-reported health equity and service use factors that may contribute to nonattendance when most EPI appointments are delivered virtually, unlike previous studies that focused solely on differences in attendance rates. Although it was during the COVID-19 pandemic and may not be representative of virtual care in real-world practice, this study suggests that virtual care may improve initial engagement in EPI services; however, barriers to care still exist for Black patients and those referred from the emergency department. A hybrid model may improve connection to EPI, though targeted approaches are needed to bridge the digital divide and ensure that structurally marginalized and high-acuity patients have equitable access to care.

(J Med Internet Res 2026;28:e81313) doi:10.2196/81313

KEYWORDS

early psychosis intervention; psychosis; youth; virtual care; health equity; access to care; engagement

XSL•FO
RenderX

Introduction

The COVID-19 pandemic presented an urgent need to transition mental health services from in-person to virtual delivery, leading to research on feasibility and effectiveness. It also raised questions about the digital divide, referring to disparities between communities in access to technology, and the implications for structurally marginalized groups already facing barriers to accessing care [1]. Like many services, early psychosis intervention (EPI) programs rapidly shifted to virtual care during the COVID-19 pandemic, with the hope of maintaining access to and engagement in services. This was particularly crucial, given that approximately half of youth with psychosis do not access treatment and one-third disengage from EPI services early [2-5], despite the emphasis on identification and treatment early in the course of illness.

Research suggests that barriers preventing youth from accessing EPI treatment generally include stigma and a lack of knowledge about psychosis and where to seek treatment [6-8], with youth often viewing pathways to EPI care as complex [8]. Families can represent key sources of support throughout the help-seeking process and can help maintain engagement with services, particularly given the typical age of onset for psychosis [9,10]. Unfortunately, pathways to EPI treatment often begin with referrals from acute services, including emergency departments (EDs) or inpatient units [11,12], which also face barriers to ensuring timely outpatient follow-up. These include patient-level barriers, such as transportation or financial issues, and health systems–level issues, such as coordination between the ED and outpatient services, insufficient funding, and political disinterest [13,14]. Building partnerships between EPI and external services, including schools and shelters, can facilitate early detection, increase referrals from nonacute pathways, and improve access to vulnerable patients [14,15].

Emerging evidence suggests that virtual EPI care is well received by youth with psychosis. In recent cross-sectional studies, youth expressed satisfaction with the virtual delivery of EPI services and found it comparable to in-person treatment [16], highlighting the convenience, ease of use, accessibility, and comfort of virtual care [17]. Virtual care can also be conducive to a client-centered approach, bolstering youth autonomy and decision-making power over care [17]. While many youth have acknowledged the value of virtual care, some have reported feeling more isolated and disconnected from clinicians, finding it more difficult to express themselves during virtual appointments [17]. Other, more practical challenges include technological difficulties and privacy/confidentiality concerns. The extent to which the benefits of virtual care extend to improvements in initial EPI attendance is unclear.

Using service use data from EPI programs in the United States and Australia, 3 studies examined attendance rates before and after the implementation of virtual care, yielding inconsistent results. One study found an increase in missed appointments after virtual care implementation compared with pre–virtual care (13.3% and 7%, respectively) [17], whereas another found a 5% increase in attendance at EPI appointments following virtual care implementation [18]. A more recent study examined the ranges of missed appointments before and after virtual care implementation and found greater variability in the percentage of missed appointments in the post–virtual care (2.7% - 9%) than pre–virtual care implementation (2.8% - 6.4%) [19]. Another study examining EPI services before and after the pandemic noted a significant increase in video appointments offered postpandemic; however, no significant difference in video appointments attended [20]. Given these conflicting findings, more research is needed to evaluate the impact of virtual care implementation on EPI appointment attendance and identify factors that may lead to nonattendance.

We recently completed a study that examined factors associated with attendance at the EPI consultation appointment from 2018 to 2019, when this appointment was offered exclusively in person [5]. We found that older patients and those referred from the ED were less likely to attend the consultation appointment. Additionally, older patients and those who identified as Black or belonging to other racial and ethnic groups were more likely to be referred to EPI services from the ED compared with White patients, indicating that structurally marginalized groups are facing barriers accessing care early in the course of illness (before it becomes an emergency) [5]. While the digital divide and disparities in access to technology may be expected to increase barriers to care [19], it remains unclear if the aforementioned health equity and service use factors continue to affect access to EPI care when services are delivered virtually.

Using retrospective data from a large EPI program, we examined factors associated with attendance at the first EPI consultation appointment following the transition to most of these being scheduled virtually, with a focus on self-reported health equity and service use factors. Initial engagement was defined as attendance at the consultation appointment. We hypothesized that patients who were older, referred from the ED, or identified as Black would be less likely to attend the consultation appointment when it was delivered virtually [5].

Methods

Setting

The Centre for Addiction and Mental Health (CAMH) EPI program is the largest in Canada, serving downtown Toronto, a large urban center. The program provides consultation and 3 years of coordinated specialty care for people up to 29 years of age with affective, nonaffective, and substance-induced psychosis. The program receives referrals from the ED, CAMH inpatient units, CAMH outpatient psychiatrists, or externally through primary care providers or external inpatient and outpatient psychiatrists. The ED also houses a “bridging” clinic, where they may triage patients who are determined to be at lower acuity levels during business hours; it is also used to provide short-term follow-up after an inpatient discharge. External referrals are triaged by nurses to appropriate services. The program aims to offer consultation appointments within 2 weeks [21].

Study Design and Population

This was a retrospective cohort study that followed Strengthening the Reporting of Observational Studies in

XSL•FO
RenderX

Epidemiology (STROBE) reporting guidelines for cohort studies [22]. We replicated methods from our previous paper examining initial engagement with in-person services [5]. Electronic medical record (EMR) data were obtained for all patients aged 16 to 29 years who were referred to the CAMH EPI program from April 2020 to December 2020, the period immediately following virtual care implementation. We classified consultation appointments between 2018 and 2019 as "pre–virtual care," as they were offered exclusively in-person during this period. Appointments between April and December 2020 were classified as "post–virtual care," as the vast majority were conducted virtually, consistent with institutional practices and jurisdictional restrictions at the time. We focused on this specific period because in-person consultations resumed during subsequent phases of the pandemic. Referral data were excluded from the analysis if the referral was canceled for any reason and for patients who were previously enrolled in the CAMH EPI program. Missing data were manually extracted from EMR charts when possible, and data from second and third referrals were excluded from the analysis. Statistical comparisons were made with 2018 to 2019 data prior to virtual care implementation [5].

Variables and Missing Data

The primary outcome was rate of attendance at the first consultation appointment. Patients were coded as having not attended if they declined services, could not be reached to schedule their appointment, or did not attend their scheduled appointment. Demographic variables were self-reported by patients through CAMH's standardized health equity form that is routinely completed around the time of their first appointment and during ED visits and inpatient admissions. These variables included age, gender, racial and ethnic group ("other racial and ethnic groups" included Indigenous, Latin American, Middle Eastern, and other not specified), country of birth, and sexual orientation. Low prevalence categories were combined for data cells of fewer than 5 people to protect the privacy of participants and facilitate statistical comparisons. Recoded variables are outlined in Multimedia Appendix 1. Service use factors were derived from clinical documentation: referral source (inpatient, ED or bridging clinic, and those from outpatient psychiatrists, primary care providers, or other external providers) and days to consult (calculated as the number of days from referral to consultation appointment). We retained "do not know" or "prefer not to answer" responses in the analysis because we considered them valid responses to self-reported health equity questions. Nonresponse data were categorized as missing, and data were missing for three variables (25.9%-27.6%). We performed Little's MCAR test, which indicated that data were MCAR (χ^2_{13}=11.09, P=.60). However, we did not impute missing data because it is routinely completed by patients around the time of their first appointment and during ED visits and inpatient admissions. To minimize potential sources of bias, we applied consistent inclusion and exclusion criteria and used standardized EMR data fields to minimize misclassification.

Statistical Analysis

Descriptive statistics were calculated, including means, SDs, and percentages, to describe the demographic and service use characteristics of the full sample. We used chi-square tests for categorical variables and independent t tests for continuous variables to compare service use pre– versus post–virtual care. Effect sizes were calculated using Cohen d for continuous variables and Phi (2 × 2 tables) and Crame´r V (2 × 3 tables) for categorical variables. We used binary logistic regression to model the odds of attendance at the consultation appointment, controlling for demographic (ie, age, gender, racial and ethnic group, sexual orientation, and country of birth) and service use (ie, referral source and days to consult) factors. Simulation work by Vittinghoff and McCulloch (2007) demonstrated that logistic regression can yield reliable estimates with 5 to 9 events-per-variable (EPV) under commonly encountered modeling conditions (ie, when covariates are not extremely sparse, effect sizes are moderate, and covariates are not highly collinear). Our data satisfy all these characteristics, and the candidate covariates were prespecified based on theoretical and clinical consideration [23]. Univariable tests were performed and variables statistically significant at an a priori level of P<.20 were included in the multivariable model. Backward stepwise selection was used to determine the final adjusted model by removing variables with a P≥.20. This criterion was used as a heuristic to safeguard against type II error during model development, following recommendations suggesting that conservative thresholds (eg, .05 or .10) may prematurely exclude important variables [24]. However, we retained age and gender in the adjusted model, regardless of statistical significance, due to their theoretical and clinical importance. A 2-sided P<.05 was considered statistically significant. All statistical analyses were performed using Stata BE (version 17.0; StataCorp) [25].

Ethical Considerations

This study was approved by the Research Ethics Board at CAMH (#060/2020 - 01). Informed consent was not required by CAMH's Research Ethics Board, as retrospective deidentified EMR data were used for the analysis and posed minimal risk. Data were deidentified and additional efforts were taken to protect the privacy of participants, including combining data cells of fewer than 5 people for low prevalence categories. There was no compensation for participants in this study, as it was an analysis of retrospective EMR data. We have taken efforts to ensure that no identification of individual participants would be possible via the data in the results or supplementary material. There were no images included in the manuscript that could lead to identification of participants.

Results

Participant Characteristics

CAMH's EPI program received 383 patient referrals between April 2020 and December 2020. Referral data that did not meet study eligibility were excluded (70/383, 18.3%), and second and third referrals were filtered out of the analysis (12/313, 3.8%), leaving 301 unique patient referrals. Table 1 displays the demographic and service use characteristics of patients. Patients had a mean age of 23.2 (SD 3.3) years, 71.1% (214/301) identified as male, 29.2% (88/301) identified as White, 40.2% (121/301) identified as heterosexual, and 46.2% (139/301) were born in Canada.

XSL•FO
RenderX

Table . Characteristics of patients referred to the Centre for Addiction and Mental Health early psychosis intervention program in 2018 - 2019 and 2020.

Characteristics	All referrals	
	Pre–virtual care, 2018 - 2019 (n=999[a])	Post–virtual care, 2020 (n=301)
Age (y), mean (SD)	22.5 (3.5)	23.2 (3.3)
Age (y), median (IQR)	22 (20 - 25)	23 (21 - 26)
Gender, n (%)		
Male	654 (65.5)	214 (71.1)
Female	323 (32.3)	81 (26.9)
Trans, nonbinary, two-spirit, other, or prefer not to answer[b]	22 (2.2)	6 (2.0)
Racial and ethnic group, n (%)		
Asian	199 (19.9)	36 (12.0)
Black	176 (17.6)	55 (18.3)
White	384 (38.4)	88 (29.2)
Other racial and ethnic groups[c]	143 (14.3)	28 (9.3)
Do not know or prefer not to answer	24 (2.4)	16 (5.3)
Missing	73 (7.3)	78 (25.9)
Sexual orientation, n (%)		
Heterosexual	667 (66.8)	121 (40.2)
LGBTQ2S+[d,e]	159 (15.9)	41 (13.6)
Do not know or prefer not to answer[e]	54 (5.4)	56 (18.6)
Missing	119 (11.9)	83 (27.6)
Born in Canada, n (%)		
No	294 (29.4)	69 (22.9)
Yes	606 (60.7)	139 (46.2)
Do not know or prefer not to answer	26 (2.6)	15 (5.0)
Missing	73 (7.3)	78 (25.9)
Referral source, n (%)		
Outpatient psychiatrists, PCPs[f], or other external providers	525 (52.6)	122 (40.5)
ED[g]/Bridging	217 (21.7)	65 (21.6)
Inpatient	257 (25.7)	114 (37.9)
Days to consult, mean (SD)[h]	18.3 (13.9)	12.6 (8.9)
Days to consult, median (IQR)	15 (10-22)	11 (7-15)
Monthly referrals, mean (SD)	44.2 (9.9)	34 (4.4)

[a]Pre–virtual care data from Polillo et al [5].

[b]2018 - 2019 data includes two-spirit and prefer not to answer.

[c]Other racial and ethnic groups include Indigenous, Latin American, Middle Eastern, and other not specified.

[d]LGBTQ2S+: lesbian, gay, bisexual, trans, queer (or sometimes questioning), and two-spirited.

[e]2018 - 2019 data includes two-spirit and 2020 data includes prefer not to answer.

[f]PCP: primary care provider.

[g]ED: emergency department.

[h]Includes data only for participants who booked a consultation appointment (n=280).

XSL•FO
RenderX

Service Use

Service use characteristics are displayed in Tables 1 and 2. Approximately one-fifth of patients (65/301, 21.6%) were referred from the ED or bridging clinic, 37.9% (114/301) from inpatient units, and 40.5% (122/301) from other referral sources. There were significantly higher rates of inpatient referral (114/301, 37.9%) and lower rates of referral from outpatient and other providers (122/301, 40.5%) post–virtual care compared with pre–virtual care (χ^2_2=18.7, *P*<.001), with a small effect size and moderately narrow CI (Cramér *V*=0.120, 95% CI 0.06 to 0.17). The mean number of days from referral to consultation appointment was 12.6 (SD 8.9), which was significantly lower (t_{1149}=6.44, 95% CI 3.95 to 7.41; *P*<.001) than wait times pre–virtual care (mean 18.3, SD 13.9), with a moderate effect size and narrow CI (Cohen *d*=0.44, 95% CI 0.31 to 0.58). The mean number of monthly referrals post–virtual care (mean 34, SD 4.4) was significantly lower (t_{1298}=−17.3, 95% CI −10.9 to −8.75; *P*<.001) than pre–virtual care (mean 44.2, SD 9.9), with a large effect size and a narrow CI (Cohen *d*=−1.14, 95% CI −1.27 to −1.00). Overall, 84.1% (253/301) of patients attended their consultation appointment post–virtual care, which was significantly higher than attendance pre–virtual care (770/999, 77.1%); however, it was a small effect with a moderately wide CI (χ^2_1=6.71, ϕ=0.072, 95% CI 0.02 to 0.12; *P*=.01). Post–virtual care, patients referred from the ED or bridging clinic had the highest rate of nonattendance (18/65, 27.7%) at the consultation appointment compared with those referred from inpatient units (19/114, 16.7%) or other providers (11/122, 9%).

Table . Rates of attendance at early psychosis intervention consultation appointments at the Centre for Addiction and Mental Health by referral source in 2018 - 2019 and 2020.

Referral source	Outcome of referral					
	Pre–virtual care, 2018-2019[a]			Post–virtual care, 2020		
	Patients, n	Attended consult, n (%)	Did not attend consult[a], n (%)	Patients, n	Attended consult, n (%)	Did not attend consult[b], n (%)
All referral sources	999	770 (77.1)	229 (22.9)	301	253 (84.1)	48 (15.9)
Inpatient	257	215 (83.7)	42 (16.3)	114	95 (83.3)	19 (16.7)
ED[c]/Bridging	217	145 (66.8)	72 (33.2)	65	47 (72.3)	18 (27.7)
Outpatient psychiatrists, PCPs[d], or other external providers	525	410 (78.1)	115 (21.9)	122	111 (91.0)	11 (9.0)

[a]Pre–virtual care data from Polillo et al [5].

[b]Nonattendance at consult includes those who declined services, could not be reached for booking, or booked and did not attend.

[c]ED: emergency department.

[d]PCP: primary care provider.

Factors Associated With Attendance at Consultation Appointment

Following univariable tests and stepwise backward selection, identifying as Black (odds ratio 0.45, 95% CI 0.21 to 0.97) and being referred from the ED or bridging clinic (odds ratio 0.24, 95% CI 0.08 to 0.72) were associated with decreased odds of attendance at the consultation appointment in the final adjusted model (Table 3).

XSL•FO
RenderX

Table . Logistic regression analysis of factors associated with attendance at early psychosis intervention consultation appointment at the Centre for Addiction and Mental Health during virtual care delivery in 2020 (n=301).

Variable	Attendance at EPI[a] consultation appointment			
	Univariable		Multivariable	
	OR[b] (95% CI)	*P* value	OR (95% CI)	*P* value
Age (y)	0.94 (0.85 - 1.03)	.18	0.91 (0.80 - 1.02)	.11
Gender				
Male	Reference	—[c]	Reference	—
Female, trans, nonbinary, or other	0.98 (0.50 - 1.94)	.96	1.24 (0.51 - 3.00)	.64
Racial and ethnic group				
Asian[d]	0.65 (0.24 - 1.82)	.42	—	—
Black	0.42 (0.18 - 0.99)	.046	0.45 (0.21 - 0.97)	.04
White	Reference	—	Reference	—
Other racial and ethnic groups, do not know, or prefer not to answer[d,e]	1.58 (0.48 - 5.21)	.45	—	—
Sexual orientation				
Heterosexual	Reference	—	Reference	—
LGBTQ2S+[f], do not know, or prefer not to answer	1.39 (0.67 - 2.88)	.37	—	—
Born in Canada				
Yes	Reference	—	Reference	—
No, do not know, or prefer not to answer	0.91 (0.45 - 1.87)	.80	—	—
Referral source				
Outpatient psychiatrists, PCPs[g], or other external providers	Reference	—	Reference	—
ED[h]/Bridging	0.26 (0.11 - 0.59)	.001	0.24 (0.08 - 0.72)	.01
Inpatient	0.50 (0.22 - 1.09)	.08	0.48 (0.17 - 1.41)	.18
Days to consult	1.04 (0.98 - 1.10)	.22	—	—

[a]EPI: early psychosis intervention.

[b]OR: odds ratio

[c]Not applicable

[d]Variables removed from the adjusted model through backward stepwise selection.

[e]Other racial and ethnic groups include Indigenous, Latin American, Middle Eastern, and other not specified.

[f]LGBTQ2S+: lesbian, gay, bisexual, trans, queer (or sometimes questioning), and two-spirited.

[g]PCP: primary care provider

[h]ED: emergency department

Discussion

Principal Findings

Using retrospective cohort data from 301 patients, we found that patients referred to EPI services after the transition to virtual care had higher rates of attendance at the consultation appointment and encountered shorter wait times compared to those referred pre–virtual care. We also found higher rates of inpatient referral and lower rates of referral from outpatient and other providers post–virtual care compared to pre–virtual care. Approximately one-third of patients did not attend their consultation appointment; ED or bridging clinic referrals had the highest rate of nonattendance at the consultation appointment compared with other referral sources. Equity-related and service use factors, notably identifying as Black and being referred from

XSL•FO
RenderX

the ED or bridging clinic, were associated with decreased odds of attending the consultation appointment when it was mostly delivered virtually. These findings suggest that virtual care can improve initial engagement in EPI services; however, patients from structurally marginalized groups and those referred from acute sources still face barriers to care, even when appointments are delivered virtually.

Comparison With Prior Work

We observed a 7% increase in attendance at the EPI consultation appointment following the implementation of virtual care, which is higher than the 5% increase found across all appointments in a prior study [18]. This difference is likely attributable to the focus on the consultation appointment (rather than all appointments). Findings from both studies, when taken together, lend support to virtual care as a tool for improving initial engagement in EPI services, as well as throughout treatment; however, it is important to acknowledge that other factors, such as the COVID-19 lockdown which restricted nonessential outings and may have made youth more available to attend their appointments, may have influenced appointment attendance. We also observed a decrease in wait times from EPI referral to the consultation appointment post–virtual care compared to pre–virtual care, consistent with emerging evidence that virtual care may allow for increased efficiency and streamlined processes [17]. Although the association of shorter wait times with virtual care is an important finding, it is also important to note that monthly referrals to EPI were significantly lower post–virtual care compared with pre–virtual care, which, in turn, could be influencing the shorter wait times. Nevertheless, longer wait times for EPI services have been associated with worse patient outcomes, not only during the initial period of untreated psychosis but also later in treatment [26].

Our results also align with existing research on disparities in virtual care access. Patients referred from the ED or bridging clinic had decreased odds of attending the consultation appointment post–virtual care, consistent with our prior work showing lower attendance rates at in-person initial appointments for those referred from the ED or bridging clinic [5]. Individuals seeking care from acute services may experience greater symptom burden, such as disorganization, paranoia, and a fear of surveillance, which may affect the utilization of virtual care. For example, patients may be too disorganized to navigate connecting online at the prescribed time [27]. Additionally, these patients may be more likely to experience socioeconomic disadvantage, which is linked to increased ED usage, decreased engagement with outpatient mental health care [28], and digital equity barriers. In contrast to findings from our previous study examining in-person attendance [5], we also found that Black patients had a decreased odds of attending the consultation appointment post–virtual care, suggesting that virtual care may amplify health inequities through the digital divide. This is consistent with other research reporting that Black patients were less likely to use telehealth services compared to White patients [29,30], likely due to systemic barriers that can limit access to devices, the internet, or digital literacy skills [31,32]. This is layered upon a long history of racism, oppression, and stigmatization of Black people by the medical system, and ongoing inequities, including a greater risk of police involvement and involuntary hospitalization for Black youth experiencing psychosis [33]. Overcoming systemic barriers to psychosis care for Black patients will require systemic solutions, such as increasing the racial and ethnic diversity of clinicians, ensuring a trauma-informed and anti-racist approach to services, and including lived experience perspectives of Black patients and their families in EPI evaluation [34].

Implications

These findings have several implications for the design and delivery of EPI services in a postpandemic context. The use of a hybrid model in EPI, integrating virtual care with in-person appointments, can be a way to engage patients in treatment in a way that meets their needs while still maintaining the benefits of an in-person therapeutic relationship. Global surveys of clinicians and clinical leaders show strong support for hybrid models, though challenges exist around infrastructure, training, and digital literacy [35]. Mental health care organizations, including EPI programs, should consider building up technological infrastructure and capacity to support clinicians' ability to offer hybrid models of care, although we acknowledge that this may be challenging for more low-resourced settings. Additionally, our results showed a shift in referral patterns post–virtual care, with increased inpatient and decreased outpatient referrals. This may reflect delays in early detection of psychosis due to school closures or reduced access to primary care and other outpatient providers. It is clear that outpatient providers represent irreplaceable sources for early detection and EPI access. Leveraging hybrid care models can improve connection to outpatient providers, though consideration of systemic solutions to reduce inequities in digital access is required.

Limitations

This study has several limitations. Due to the wide CIs in the final adjusted model, the findings should be interpreted with caution. Variables, including race/ethnicity, sexual orientation, and country of birth, had more than 10% of missingness, which was not imputed due to it being self-reported health equity data. There were increases in the percentage of missing data for certain demographic categories from pre–virtual care to post–virtual care, including racial and ethnic group and sexual orientation. These changes highlight potential challenges with the reliability of self-reported data, particularly in the virtual care context. While the overall missingness is reported as more than 10%, the variability in the self-reported health equity data between the two timepoints may affect the accuracy and generalizability of the findings. Future research should explore methods to improve the completeness and reliability of self-reported data within virtual settings. The time period of the study was also a limitation, as it was during the COVID-19 pandemic and may not be representative of virtual care in real-world practice. Thus, findings may be partially related to pandemic effects rather than the transition to virtual care. The use of data from 2020 may also limit the relevance of our findings, given changes in healthcare delivery and policy over time [36]. We were not able to include other relevant patient- and system-level factors, such as socioeconomic status and illness severity, as covariates in the model. These factors may

XSL•FO
RenderX

better explain the associations observed in our final adjusted model. For example, lower referral volumes and treatment delays may have led to patients presenting to care when they were more ill and had a higher symptom burden [37]. As such, not having specific data on illness severity limits our ability to understand its potential impact on EPI attendance and engagement and should be an area for future investigation. It is also important to note that patients referred from the ED may differ from those referred through outpatient and inpatient settings [5], limiting the generalizability of our findings across referral sources. Another limitation is that the study setting was a well-resourced hospital with infrastructure and technological support dedicated to virtual care implementation, the findings may not be generalizable to lower-resourced mental health settings.

Conclusions

This retrospective cohort study of patients referred to EPI services is innovative in that it examines the self-reported health equity and service use factors that may contribute to nonattendance when most EPI appointments are delivered virtually. This is unlike previous studies, which solely focused on differences in attendance rates. Although the time period of the study is a limitation, as it was during the COVID-19 pandemic and may not be representative of virtual care in real-world practice, we found improved attendance rates at the first EPI consultation appointment and shorter wait times, suggesting that virtual care may improve initial engagement in EPI services. Despite this, barriers to care still exist for Black patients and those referred from the ED. The use of a hybrid model can be a way to improve connection to EPI, though targeted approaches are needed to improve the digital divide and ensure that structurally marginalized and high-acuity patients have equitable access to EPI care. Future research should examine virtual engagement in EPI services in a postpandemic context.

Acknowledgments

There was no use of generative artificial intelligence (AI) technology in the generation of text, figures, or other informational content of this manuscript.

Funding

This work was supported by the University of Toronto Department of Psychiatry's Reasons for Hope Fund and the Canadian Institutes of Health Research.

Data Availability

The data that supports the findings of this study are available from the corresponding author, AP, upon reasonable request to protect the privacy and confidentiality of participants.

Authors' Contributions

Conceptualization**:** AP (lead), NK (equal)
Formal analysis**:** AP (lead), TV (equal)
Methodology**:** AP (lead), NK (equal), AHCW (support), GF (support), WW (support), AV (support)
Project administration**:** AP (lead), NK (support)
Supervision**:** AP
Visualization**:** TV (lead), AP (support)
Writing – original draft**:** TV (lead), AP (support)
Writing – review & editing**:** All authors

Conflicts of Interest

None declared.

Multimedia Appendix 1
Recoded variables.
[DOCX File, 15 KB - jmir_v28i1e81313_app1.docx]

Checklist 1
STROBE checklist.
[PDF File, 176 KB - jmir_v28i1e81313_app2.pdf]

References

1. Litchfield I, Shukla D, Greenfield S. Impact of COVID-19 on the digital divide: a rapid review. BMJ Open 2021 Oct 12;11(10):e053440. [doi: 10.1136/bmjopen-2021-053440] [Medline: 34642200]

XSL•FO
RenderX

2. Anderson KK, Norman R, MacDougall AG, et al. Disparities in access to early psychosis intervention services: comparison of service users and nonusers in health administrative data. Can J Psychiatry 2018 Jun;63(6):395-403. [doi: 10.1177/0706743718762101] [Medline: 29562748]
3. Doyle R, Turner N, Fanning F, et al. First-episode psychosis and disengagement from treatment: a systematic review. Psychiatr Serv 2014 May 1;65(5):603-611. [doi: 10.1176/appi.ps.201200570] [Medline: 24535333]
4. Mascayano F, van der Ven E, Martinez-Ales G, et al. Disengagement from early intervention services for psychosis: a systematic review. Psychiatr Serv 2021 Jan 1;72(1):49-60. [doi: 10.1176/appi.ps.201900375] [Medline: 33234052]
5. Polillo A, Foussias G, Wang W, et al. Care pathways and initial engagement in early psychosis intervention services among youths and young adults. JAMA Netw Open 2023 Sep 5;6(9):e2333526. [doi: 10.1001/jamanetworkopen.2023.33526] [Medline: 37703014]
6. dos Santos Martin I, Giacon BCC, Vedana KGG, Zanetti ACG, Fendrich L, Galera SAF. Where to seek help? Barriers to beginning treatment during the first-episode psychosis. Int J Nurs Sci 2018 Jul 10;5(3):249-254. [doi: 10.1016/j.ijnss.2018.06.007] [Medline: 31406833]
7. Tiller J, Maguire T, Newman-Taylor K. Early intervention in psychosis services: a systematic review and narrative synthesis of barriers and facilitators to seeking access. Eur Psychiatry 2023 Nov 6;66(1):e92. [doi: 10.1192/j.eurpsy.2023.2465] [Medline: 37929296]
8. Anderson KK, Fuhrer R, Malla AK. “There are too many steps before you get to where you need to be”: help-seeking by patients with first-episode psychosis. J Ment Health 2013 Aug;22(4):384-395. [doi: 10.3109/09638237.2012.705922] [Medline: 22958140]
9. Del Vecchio V, Luciano M, Sampogna G, et al. The role of relatives in pathways to care of patients with a first episode of psychosis. Int J Soc Psychiatry 2015 Nov;61(7):631-637. [doi: 10.1177/0020764014568129] [Medline: 25614470]
10. Lucksted A, Stevenson J, Nossel I, Drapalski A, Piscitelli S, Dixon LB. Family member engagement with early psychosis specialty care. Early Interv Psychiatry 2018 Oct;12(5):922-927. [doi: 10.1111/eip.12403] [Medline: 27863039]
11. Anderson KK, Fuhrer R, Malla AK. The pathways to mental health care of first-episode psychosis patients: a systematic review. Psychol Med 2010 Oct;40(10):1585-1597. [doi: 10.1017/S0033291710000371] [Medline: 20236571]
12. Anderson KK, Fuhrer R, Wynant W, Abrahamowicz M, Buckeridge DL, Malla A. Patterns of health services use prior to a first diagnosis of psychosis: the importance of primary care. Soc Psychiatry Psychiatr Epidemiol 2013 Sep;48(9):1389-1398. [doi: 10.1007/s00127-013-0665-3] [Medline: 23429939]
13. Walker ER, Fukuda J, McMonigle M, Nguyen J, Druss BG. A qualitative study of barriers and facilitators to transitions from the emergency department to outpatient mental health care. Psychiatr Serv 2021 Nov 1;72(11):1311-1319. [doi: 10.1176/appi.ps.202000299] [Medline: 33887957]
14. O’Connell N, O’Connor K, McGrath D, et al. Early intervention in psychosis services: a systematic review and narrative synthesis of the barriers and facilitators to implementation. Eur Psychiatry 2021 Dec 16;65(1):e2. [doi: 10.1192/j.eurpsy.2021.2260] [Medline: 34913421]
15. Cheng C, Dewa CS, Goering P. Matryoshka Project: lessons learned about early intervention in psychosis programme development. Early Interv Psychiatry 2011 Feb;5(1):64-69. [doi: 10.1111/j.1751-7893.2010.00255.x] [Medline: 21272277]
16. Dong F, Jumper MBE, Becker-Haimes EM, et al. Tele-mental health transitions for pennsylvania coordinated specialty care programs for early psychosis during the COVID-19 pandemic. Psychiatr Q 2023 Jun;94(2):89-102. [doi: 10.1007/s11126-023-10015-0] [Medline: 36820952]
17. Randall LA, Raisin C, Waters F, Williams C, Shymko G, Davis D. Implementing telepsychiatry in an early psychosis service during COVID-19: experiences of young people and clinicians and changes in service utilization. Early Interv Psychiatry 2023 May;17(5):470-477. [doi: 10.1111/eip.13342] [Medline: 35943177]
18. Chaudhry S, Weiss A, Dillon G, O’Shea A, Hansel TC. Psychosis, telehealth, and COVID-19: successes and lessons learned from the first wave of the pandemic. Disaster Med Public Health Prep 2022 Oct;16(5):1785-1788. [doi: 10.1017/dmp.2021.42] [Medline: 33588969]
19. Nelson EB, Franco OH, Patton BA, et al. Telehealth of coordinated specialty care in early psychosis during COVID-19. J Clin Psychiatry 2022 Nov 30;84(1):21m14259. [doi: 10.4088/JCP.21m14259] [Medline: 36449475]
20. Duggan A, Buckley S, Fletcher E, et al. The impact of the COVID-19 pandemic on first-episode psychosis presentations in two early intervention in psychosis services. Ir J Psychol Med 2024 Dec 10:1-8. [doi: 10.1017/ipm.2024.61] [Medline: 39655482]
21. Early psychosis intervention program standards. : Ministry of Health and Long-Term Care; 2011 URL: https://epiontario.org/wp-content/uploads/2024/10/EPI-Program-Standards-2011_.pdf [accessed 2026-01-28]
22. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
23. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol 2007 Mar 15;165(6):710-718. [doi: 10.1093/aje/kwk052] [Medline: 17182981]
24. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression: John Wiley & Sons; 2013. URL: https://onlinelibrary.wiley.com/doi/book/10.1002/9781118548387 [accessed 2026-01-28]

XSL•FO
RenderX

25. Stata. URL: https://www.stata.com/ [accessed 2026-01-28]
26. Reichert A, Jacobs R. The impact of waiting time on patient outcomes: evidence from early intervention in psychosis services in England. Health Econ 2018 Nov;27(11):1772-1787. [doi: 10.1002/hec.3800] [Medline: 30014544]
27. Sugarman DE, Horvitz LE, Greenfield SF, Busch AB. Clinicians' perceptions of rapid scale-up of telehealth services in outpatient mental health treatment. Telemed J E Health 2021 Dec;27(12):1399-1408. [doi: 10.1089/tmj.2020.0481] [Medline: 33600272]
28. Sweeney S, Air T, Zannettino L, Galletly C. Psychosis, socioeconomic disadvantage, and health service use in South Australia: findings from the second Australian National Survey of Psychosis. Front Public Health 2015;3:259. [doi: 10.3389/fpubh.2015.00259] [Medline: 26636059]
29. Adepoju OE, Chae M, Ojinnaka CO, Shetty S, Angelocci T. Utilization gaps during the COVID-19 pandemic: racial and ethnic disparities in telemedicine uptake in federally qualified health center clinics. J Gen Intern Med 2022 Apr;37(5):1191-1197. [doi: 10.1007/s11606-021-07304-4] [Medline: 35112280]
30. Buis LR, Brown LK, Plegue MA, et al. Identifying inequities in video and audio telehealth services for primary care encounters during COVID-19: repeated cross-sectional, observational study. J Med Internet Res 2023 Sep 29;25:e49804. [doi: 10.2196/49804] [Medline: 37773609]
31. Curtis ME, Clingan SE, Guo H, Zhu Y, Mooney LJ, Hser YI. Disparities in digital access among American rural and urban households and implications for telemedicine-based services. J Rural Health 2022 Jun;38(3):512-518. [doi: 10.1111/jrh.12614] [Medline: 34355427]
32. Iasiello JA, Rajan A, Zervos E, Parikh AA, Snyder RA. Racial differences in patient-reported access to telehealth: an important and unmeasured social determinant of health. JCO Oncol Pract 2023 Dec;19(12):1215-1223. [doi: 10.1200/OP.23.00006] [Medline: 37844269]
33. Knight S, Jarvis GE, Ryder AG, Lashley M, Rousseau C. 'It just feels like an invasion': Black first-episode psychosis patients' experiences with coercive intervention and Its influence on help-seeking behaviours. J Black Psychol 2023 Apr;49(2):200-235. [doi: 10.1177/00957984221135377]
34. Davis B, Anglin DM, Oluwoye O, Keshavan M. The unfulfilled promise of equitable first episode care for Black-Americans: a way forward. Schizophr Res 2022 Mar;241:171-173. [doi: 10.1016/j.schres.2022.01.046] [Medline: 35124436]
35. Torous J. Hybrid and virtual care: interest outstrips infrastructure. NEJM Catal Innov Care Deliv 2025 Jan 15;6(2). [doi: 10.1056/CAT.25.0009]
36. Danseco E, Kurzawa J, Sundar P, Brown J, Huang C. Evaluating the sector-wide implementation of virtual child and youth mental health services in response to the COVID-19 pandemic: perspectives from service providers, agency leaders and clients. Implement Res Pract 2021;2:26334895211045690. [doi: 10.1177/26334895211045690] [Medline: 37089999]
37. Penttilä M, Jääskeläinen E, Hirvonen N, Isohanni M, Miettunen J. Duration of untreated psychosis as predictor of long-term outcome in schizophrenia: systematic review and meta-analysis. Br J Psychiatry 2014 Aug;205(2):88-94. [doi: 10.1192/bjp.bp.113.127753] [Medline: 25252316]

Abbreviations

CAMH: Centre for Addiction and Mental Health
ED: emergency department
EMR: electronic medical record
EPI: early psychosis intervention
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by S Brini; submitted 29.Jul.2025; peer-reviewed by D Hilty, L Agrawal; revised version received 01.Dec.2025; accepted 04.Dec.2025; published 19.Feb.2026.

Please cite as:
Vey T, Kozloff N, Foussias G, Wang W, Wong AHC, Voineskos AN, Davis-Faroque N, Polillo A
Virtual Delivery of Early Psychosis Care: Retrospective Cohort Study of Factors Associated With Initial Engagement
J Med Internet Res 2026;28:e81313
URL: https://www.jmir.org/2026/1/e81313
doi:10.2196/81313

© Trinity Vey, Nicole Kozloff, George Foussias, Wei Wang, Albert HC Wong, Aristotle N Voineskos, Nicole Davis-Faroque, Alexia Polillo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Rising to the Challenge of Early Screening in Primary Health Care Through the Web Italian Network for Autism Spectrum Disorder (Win4ASD) in the Pediatric Population: Retrospective Observational Study

Noemi Buo[1*], MSc; Eleonora Rosi[1*], MSc; Silvia Busti Ceccarelli[1], MSc; Mariarosa Ferrario[2], MD; Valerio Maiorca[2]; Erika Morandi[2], MSc; Nicole Viganó[1], MSc; Ivan Limosani[3], MD; Laura Falcone[3], MSc; Massimo Molteni[1], MD; Paola Colombo[1], MSc, PhD

[1]Child Psychopathology Unit-Smart Lab, Scientific Institute, Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Eugenio Medea, Bosisio Parini, Lecco, Lazio, Italy

[2]Azienda Socio Sanitaria Territoriale della Valle Olona - Dipartimento di salute mentale e delle dipendenze, Gallarate, Varese, Italy

[3]Direzione Generale Welfare, Lombardy Region, Milan, Italy

[*]these authors contributed equally

Corresponding Author:
Eleonora Rosi, MSc
Child Psychopathology Unit-Smart Lab
Scientific Institute, Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Eugenio Medea
via don Luigi Monza 20
Bosisio Parini, Lecco, Lazio, 23842
Italy
Phone: 39 031877623
Email: eleonora.rosi@lanostrafamiglia.it

Abstract

Background: Early identification of autism spectrum disorder (ASD) and neurodevelopmental disorders is a key element in the ability to intervene early and improve children's outcomes. The American Academy of Pediatrics recommends routine autism screening for children aged 18 to 24 months. Telemedicine (health care services delivered remotely using digital technologies) is a valuable tool in supporting this objective. As highlighted in the literature, telemedicine increases the availability of treatment, reduces diagnostic waiting times, and supports the monitoring of neurodevelopmental disorders. The Web Italian Network for Autism Spectrum Disorder (Win4ASD) platform is a telemedicine-based screening tool designed to detect ASD early within the pediatric health care system in Italy.

Objective: This study investigated the integration of the Win4ASD platform into the Italian National Health Service, providing a detailed account of its operational features and use patterns among family pediatricians. It further examined the platform's contribution to reinforcing preventive health care strategies and enhancing coordination across child health services. Finally, through a retrospective observational analysis, this study evaluated the clinical effectiveness of Win4ASD in facilitating the early detection of ASD and improving screening outcomes.

Methods: The Win4ASD platform, which is active in the Lombardy region, helps family pediatricians screen toddlers for autism at an early age (18-24 months) using the Checklist for Autism in Toddlers. Furthermore, children identified as at risk are referred for further diagnostic evaluation to the local specialized autism unit—a mental health specialist team comprising child neuropsychiatrists, psychologists, and developmental therapists. This platform, which created an integrated network for early identification of autism, ensures rapid access to diagnostic and therapeutic services. This study analyzed data collected from the platform between January 2022 and September 2024 to assess family pediatricians' commitment and effectiveness in identifying children at risk of autism using descriptive analyses performed in SPSS.

Results: The first key finding was the commitment of family pediatricians from the Regional Public Health Services in Lombardy to use the platform. In total, 72.5% (909/1253) of all family pediatricians working for the Regional Public Health Services actively used the platform. The platform enabled the screening of 58,419 infants from the general population in Lombardy, of whom 596 (1%) were identified as "at risk" and referred for fast-track diagnostic evaluation, which is consistent with global prevalence

XSL•FO
RenderX

rates. In addition, diagnostic outcomes from a subsample are presented, showing the effectiveness of the platform in facilitating the early diagnosis of neurodevelopmental disorders.

Conclusions: Early identification of risk and timely intervention can prevent the progressive development of behavioral atypicalities, have a significant impact on the quality of life of the child and their family, and allow for a reduction in health care costs and a consequent improvement in the organization of the health care system.

(J Med Internet Res 2026;28:e74302) doi:10.2196/74302

KEYWORDS

screening Checklist for Autism in Toddlers; CHAT; online screening; tele–mental health; primary care; early detection; autism spectrum disorder; health care policy; surveillance system

Introduction

Prevention of Neurodevelopmental Disorders and Telemedicine

Telemedicine refers to the full range of health care services delivered through innovative digital technologies whereby the health professional and the patient are not in the same location [1,2]. Recently, telemedicine interventions have gained significant interest and expanded into many areas of medicine, including mental health, both before and after the COVID-19 pandemic [3-5]. The pandemic first increased the demand for effective health support, and the growing need for easy-to-access remote services has substantially boosted the use of telehealth since then [5,6]. According to recent reviews, telehealth can potentially increase treatment availability; reduce diagnostic waiting times; and aid in the monitoring of mental disorders, particularly neurodevelopmental disorders [5,7].

Neurodevelopmental disorders, such as autism spectrum disorder (ASD) and attention-deficit/hyperactivity disorder, typically begin in childhood and may persist as chronic conditions throughout life [8]. Neurodevelopmental conditions account for the highest estimated prevalence in children aged <15 years worldwide, ranging from 4.1% (<5 years) to 7% (10-14 years) [9]. The World Health Organization emphasizes that identifying children at risk of neurodevelopmental disorders is a critical first step to establishing a strong relationship between parents and health care professionals to ensure early intervention [10]. Early intervention aims to prevent or minimize motor, cognitive, and emotional impairments in young children disadvantaged by biological or environmental risk factors [11].

In summary, the high prevalence of neurodevelopmental disorders and the need for early intervention make prevention and screening more crucial than ever. Telemedicine can be an effective tool in identifying these disorders at an early stage.

The Health Care System in Italy

The Italian National Health Service (NHS) is a public, regionally based system organized at the national, regional, and local levels [12] and provides free primary care, inpatient care, and health screenings for all residents. Public health care is financed primarily through national and regional taxes (97%) and patient copayments. Private health insurance plays a limited role in Italy's health coverage system [13].

The pediatric health care system in Italy is part of the NHS and is organized in the traditional 3 levels of intervention: primary care (first access), secondary care (hospital care), and tertiary care (specialized hospital care) [14]. Children are required to register with a family pediatrician until the age of 6 years, with the option to remain under pediatric or general practitioner care until the age of 14 years. At 14 years, patients transition to general practitioners, although extensions with pediatricians may be granted for chronic or disabling conditions. Both family pediatricians and general practitioners serve as primary care providers within the system [15].

Children's development is monitored by family pediatricians at predetermined time points, totaling 10 visits from birth to the age of 13 years, during well-child checkups. Additionally, family pediatricians are responsible for implementing specific health promotion initiatives, providing parents with valuable information to assess their children's growth and health status, diagnosing and treating illnesses, prescribing medications and diagnostic tests when necessary, referring patients to specialized facilities, and collaborating with other health care professionals.

Early Prevention System for ASD

ASD is a neurodevelopmental disorder characterized by social and communication impairments and restricted or repetitive behaviors [8,16]. It has a median global prevalence of 65 in 10,000 and 1 in 31 children according to Centers for Disease Control and Prevention data [17-19]. Specifically, a recent systematic review reported a prevalence of ASD between 1.70% and 1.85% in American children aged 4 and 8 years, respectively, whereas the prevalence in Europe ranges between 0.38% and 1.55% [20]. In Italy, a very recent nationwide population study estimated a prevalence of 13.4 (CI 11.3-16.0) per 1000 children aged 7 to 9 years, with a male-to-female ratio of 4.4:1 [21].

The presentation and onset patterns of ASD are variable. In many cases, the core diagnostic features of autism can be identified during toddlerhood [22]. Interventions are more beneficial when delivered earlier in life [23,24], and the American Academy of Pediatrics recommends routine autism screening at 18 to 24 months [16]. Furthermore, over the past 12 years, the American Academy of Pediatrics has reported an increase in the prevalence rates of ASD in children, an expanded understanding of potential risk factors, improved awareness of co-occurring medical conditions and genetic contributions to the etiology, and substantial growth in research supporting evidence-based interventions [16]. Recent studies have emphasized the importance of early ASD detection through screening [7,23,24]. Standardized, digital, high-fidelity ASD

XSL•FO
RenderX

screening during pediatric well-child visits facilitates the identification of children at higher risk of autism at a younger age, including those with more subtle clinical manifestations [7,25-27].

In Italy, guidelines on the diagnosis and treatment of ASD in children and adolescents have been approved and published [28,29], emphasizing the importance of early diagnosis.

Given the importance of telemedicine in the prevention and active surveillance of mental health, a screening system, the Web Italian Network for Autism Spectrum Disorder (Win4ASD) platform [30,31], has been developed in Lombardy, a highly populated Italian region with 10,012,054 inhabitants [32]. It provides early identification of individuals at risk of neurodevelopmental disorders using the Checklist for Autism in Toddlers (CHAT) [33]. Owing to computerized scoring, children identified as being at risk are immediately referred by the platform to the local NHS child psychiatric services for a fast-track diagnostic evaluation. The platform aims to create an integrated, interdisciplinary network, enabling early identification of ASD risk; facilitating rapid activation of diagnostic pathways; and ensuring care, habilitation, and rehabilitation based on priorities and intensities determined by the child's age group and functioning profile.

On the basis of these premises, this paper aims to illustrate how the Win4ASD platform has become part of the public NHS and its potential as a preventive medicine model for the general population through a retrospective observational study.

The primary objective of this study was to illustrate the integration of the Win4ASD platform within the NHS, detailing its functionalities and analyzing the associated screening pathways. Furthermore, this study aimed to evaluate the platform's contribution to the reorganization of public health care services, the strengthening of preventive strategies, and the enhancement of coordination among health care professionals and facilities. Finally, the clinical effectiveness of the platform was assessed in a subsample of users through a retrospective observational study with the objective of determining how the use of Win4ASD may facilitate the early diagnosis of ASD and improve screening outcomes.

Methods

Overview

The Win4ASD platform enables early screening for autism in toddlers (from 18 months; –2 months to +2 months) by filling out the CHAT [33].

The family pediatrician accesses the platform to register the patient and record project consent and then administers the screening. The CHAT screening test involves the administration of 14 items: 9 questions directed to the parents and 5 direct observations of the child's behavior, which are assessed in person by the pediatrician, who provides the corresponding responses for these observational items and enters them into the platform. Upon completion, the platform provides automated scoring in real time; stores the results; and, based on the screening outcome, recommends appropriate follow-up actions for the family pediatrician to undertake.

If the screening outcome is "generic" or "medium risk," the assessment can be repeated after a month. In the case of a "high risk" outcome at the first screening and a "generic risk," "medium risk," or "high risk" outcome at the repeated screening, the Win4ASD platform allows for the patient's referral to fast-track diagnostic evaluation. This customized platform is designed to connect primary care (family pediatricians) and mental health specialist services in safe ways. Specifically, 8 health care protection agencies in Lombardy provided a list of all NHS pediatricians, who then received invitations to use the platform. Within the platform, each family pediatrician has the option to refer high-risk children to the local specialized autism unit (SAU), a multidisciplinary and specialized team focused on the diagnosis and treatment of ASD, based on the gold standard. Specifically, each diagnostic assessment includes standardized procedures—comprising parent or caregiver interviews, direct child observations, and administration of validated diagnostic instruments (eg, the Autism Diagnostic Observation Schedule, Second Edition [ADOS-2]; Autism Diagnostic Interview–Revised [ADI-R]; Vineland Adaptive Behavior Scales, Second Edition; Leiter International Performance Scale; Griffiths Scales of Child Development, Third Edition; Psychoeducational Profile, Third Edition; Protocol for a Preventive Assessment of the Motor Coordination Disorder; *Primo Vocabolario del Bambino*; and Adaptive Behavior Assessment System–II)—conducted by clinical psychologists and child psychiatrists. The final diagnosis is formulated by the child psychiatrist, who integrates all clinical and psychometric data to develop the child's functional profile and therapeutic plan. While the diagnostic process is standardized across local units, individual teams may incorporate additional assessments according to clinical needs [34-43].

The different services communicate through the platform bulletin board, where pediatricians and mental health specialists can write observations and attach documents or reports. In addition, e-learning training courses and information on how to use the platform for family pediatricians are visible and easily accessible from the platform dashboard. The dashboard aggregates key information for real-time monitoring of screening data, patient data, and the activity trends of family pediatricians. The platform is described in detail in the study by Colombo et al [31]. For this study, we considered family pediatricians' activities and screenings performed within the platform from January 1, 2022, to September 30, 2024. The effectiveness of the screening process was studied further by analyzing data from a subsample for which outcome data were available.

The platform is now being used throughout the Lombardy region following the approval of the Autism Operational Plan (regional deliberations XI/5415 of October 25, 2021, and XII/3686 of December 20, 2024) [34].

Ethical Considerations

This study was conducted in accordance with the ethical principles for research involving human participants and with the guidelines of the Declaration of Helsinki. It was reviewed and approved by the Eugenio Medea Scientific Institute Ethics

XSL•FO
RenderX

and Committee (protocol code 352; September 22, 2016). Informed consent was obtained from the caregivers of all participating children before data collection through pediatric clinical practice. All data used in this research were pseudonymized to ensure participants' privacy and confidentiality. No financial or material compensation was provided for participation in the study. The dataset analyzed does not contain any information that could allow for the identification of individual participants.

Materials

Win4ASD Platform

We used the input from the Win4ASD platform, representing "performance data" together with data resulting from screening. Specifically, we calculated the number of family pediatricians per participating district in the Lombardy region and subsequently categorized them into 3 groups: those who were active and regularly used the platform, those who had never accessed it, and those who had logged in but did not engage with it.

Screening Tool

The screening tool is the Italian version of the CHAT [33,44]—a checklist for detecting children aged 18 months (–2 months to +2 months) who are at risk of autism. It was validated in England with a population of 16,000 children and was designed to be administered by primary health care services. The checklist consists of 2 sections: section A comprises 9 questions for parents asked by the family pediatricians, while section B comprises 5 items completed by the pediatrician based on direct observation of the child's behavior. All items are rated as "yes" or "no." This standardized screening tool is available to family pediatricians through the web application. It takes 5 to 10 minutes to administer and requires the family pediatricians' direct observation and involvement in the entire process, thereby increasing their compliance. The screening result can be "high risk" (failure of all critical items), "medium risk" (failure of a few critical items), "generic risk" (a risk of other neurodevelopmental disorders; failure of a few noncritical items), or "no risk." If the outcome of the first evaluation is "medium risk" or "generic risk," the family pediatrician has the option of repeating the assessment after 1 month (second evaluation). If the outcome is "high risk" either at the first or repeated screening or "medium risk" or "generic risk" at the repeated evaluation, the family pediatrician refers the patient for fast-track diagnostic evaluation through the platform.

Diagnostic Outcome

The diagnostic results are available for a subsample derived from a specific SAU in a regional district (Valle Olona). The diagnostic procedures described above (eg, clinical interviews and observations of the child) were carried out, and in particular, the 2 gold standard tools for autism were used: the ADOS-2 [35] and ADI-R [45]. The ADOS-2 is a semistructured, standardized assessment of autism symptomatology across communication, reciprocal social interaction abilities, play, and restricted and repetitive behaviors and yields diagnostic classifications or concern classification scores. The ADI-R is an investigator-based semistructured diagnostic interview that was designed to provide a developmental history framework for a lifetime differential diagnosis of pervasive developmental disorders and information about current functioning (over the previous 3 months) for individuals with a mental age of ≥2 years from early childhood to adult life. It is usually conducted with the parents or main carer of the child.

Statistical Analysis

Analyses were performed both on the total Win4ASD platform sample and on a specific subsample from the Valle Olona SAU. Descriptive statistics were used to summarize demographic characteristics and screening outcomes. Continuous variables were expressed as means and SDs, and categorical variables were expressed as absolute and relative frequencies. The 95% CIs were calculated for the main outcome proportions to assess the precision of estimates. Additionally, comparisons between male and female participants were conducted to explore potential Sex-related differences in screening outcomes. All statistical analyses were performed using SPSS (version 26; IBM Corp) [46], and a 2-sided significance level of .05 was considered.

Data Exclusion

We excluded the CHATs completed outside the age range of 16 to 22 months from the overall sample. In addition, as mentioned previously, we focused on a time frame that excluded all CHATs completed before January 2022 and after September 30, 2024. Of the total sample, 16 participants were excluded from the repeat assessment because they repeated the CHAT despite being classified as "no risk" at the first assessment.

Results

Win4ASD Platform Data

Pediatricians' Commitment

A total of 1253 family pediatricians were registered on the platform during the indicated time frame, of whom 909 (72.5%) were active and regularly used the platform, 188 (15%) had never logged on, and 163 (13%) had logged on but were not using the platform. The average number of screening CHATs performed by family pediatricians was 70.5 (SD 53.9; range 1-317). Through the platform's interactive dashboard, all information is presented and synthesized simply and intuitively.

The data shown in Figure 1 illustrate the pediatricians' use of the platform across different regional districts and their inclusion in the e-learning program (n=346). The bar chart in Figure 2 provides information about family pediatricians' engagement over time.

XSL•FO
RenderX

Figure 1. Family pediatricians' commitment to the web platform (total number of pediatricians, pediatricians who have taken e-learning courses, pediatricians divided by local ATS department). ATS: Italian name for the 8 local health authorities.

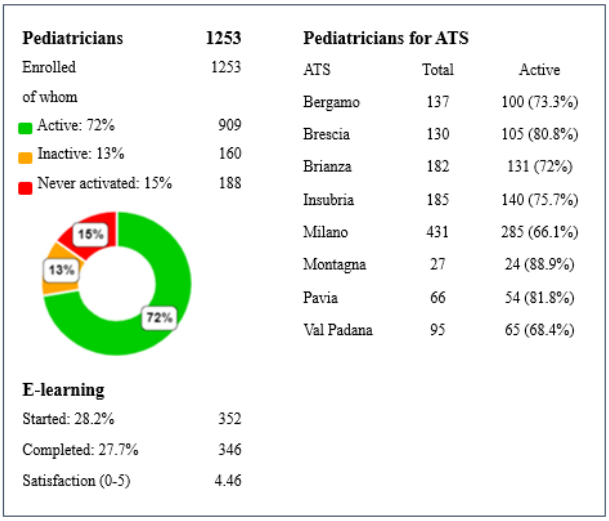

Figure 2. Family pediatricians' commitment over time: the x-axis shows the different months from January 2022 to September 2024, whereas the y-axis represents the number of pediatricians who registered on the platform and actively used it.

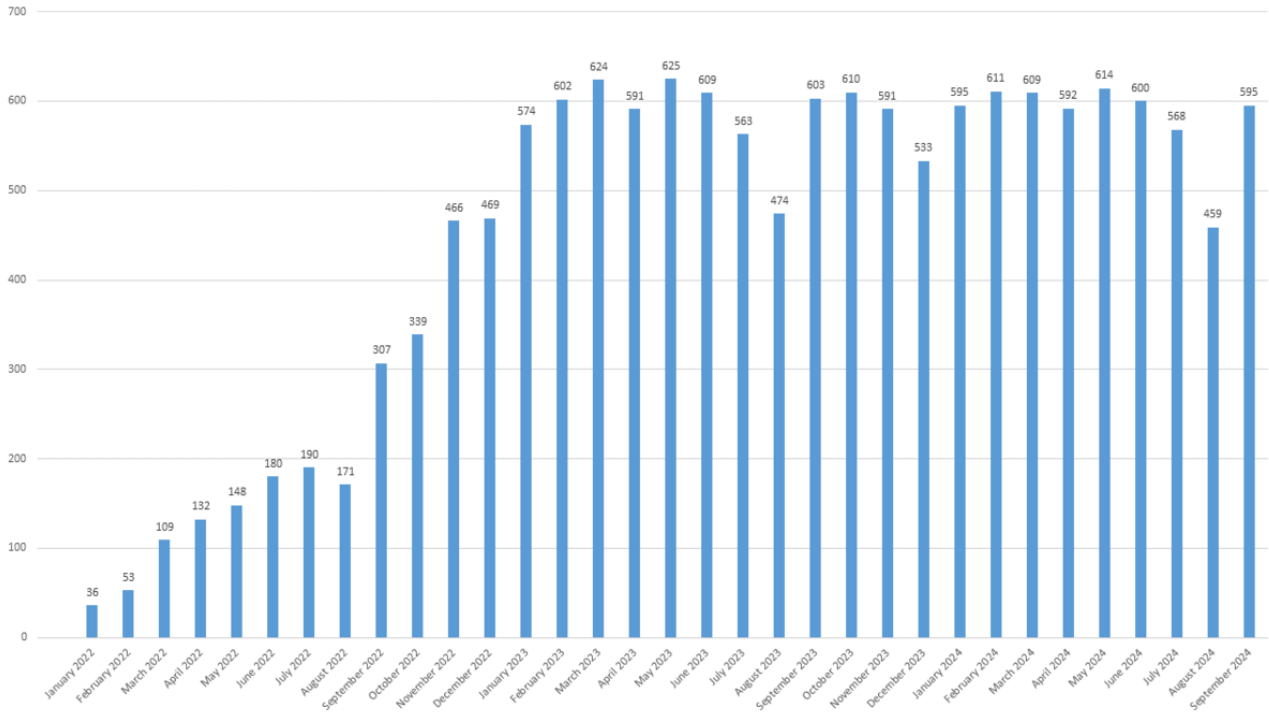

Early Screening System

A total of 58,419 toddlers from the general population in the Lombardy region were screened during the 16- or 18-month well-child checkup visit (n=29,980, 51.3% male and n=28,439, 48.7% female; mean age 17.32, SD 1.35; range 16-22 months).

From January 2022 to September 2024, the number of screenings gradually increased from 532 in the first quarter of 2022 to 5666 in the third quarter of 2024. The largest number of screenings was performed by family pediatricians in the first quarter of 2023, reaching 8775 CHATs (Figure 3).

A flowchart (Figure 4) shows the outcomes for patients with completed first and repeated screening CHATs, as well as referrals for a further diagnostic investigation.

XSL•FO
RenderX

Figure 3. Number of screenings by family pediatricians from January 2022 to September 2024 shown by quarterly trends.

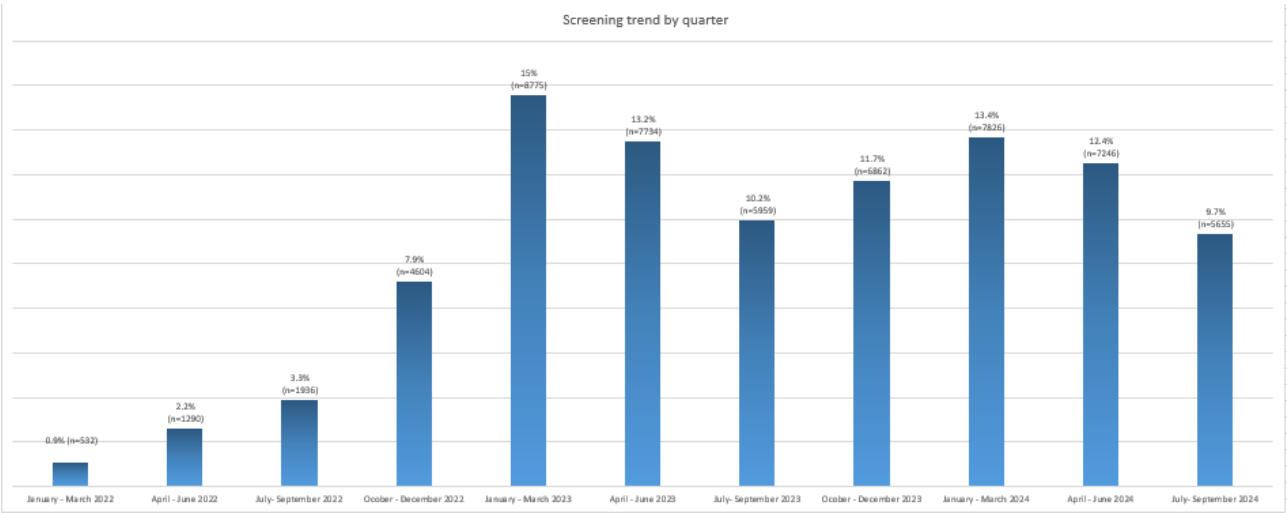

Figure 4. Screening outcomes: flowchart of screenings conducted by family pediatricians from January 2022 to September 2024. CHAT: Checklist for Autism in Toddlers.

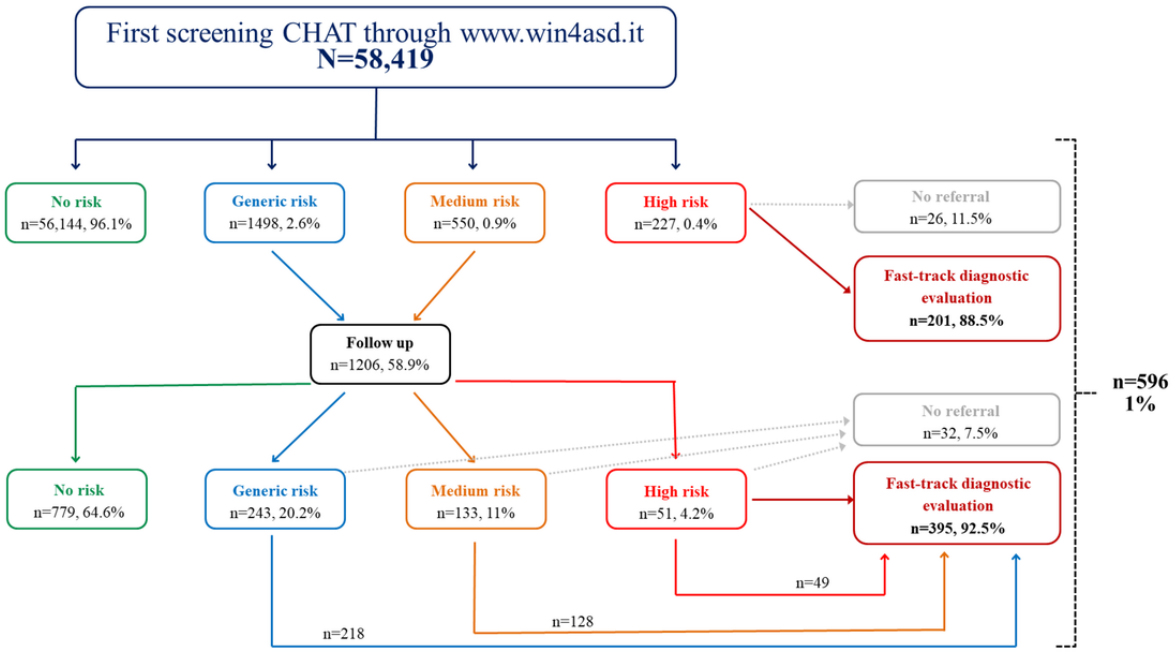

On the basis of the results, 96.1% (56,144/58,419) of the children were at "no risk" of a neurodevelopmental disorder at first screening, 2.6% (1498/58,419) had a "generic risk," 0.9% (550/58,419) were at "medium risk," and 0.4% (227/58,419) were at "high risk." Table 1 details the proportions of each screening outcome with corresponding 95% CIs in the overall screened population. A total of 0.3% (201/58,419) of the children screened in total (201/227, 88.5% of "high risk" children) were referred for fast-track diagnostic evaluation to the SAU. The detailed screening outcomes with a comparison of demographic variables is shown in Table 2.

Table 1. Proportions of each screening outcome with corresponding 95% CIs in the overall screened population (N=58,419).

Screening outcome	Participants, n (%; 95% CI)
No risk	56,144 (96.1; 95.95-96.26)
Generic risk	1498 (2.6; 2.44-2.69)
Medium risk	550 (0.9; 0.86-1.02)
High risk	227 (0.4; 0.34-0.44)

XSL•FO
RenderX

Table 2. Comparison of demographic and screening characteristics between male and female participants (N=58,419).

Variable	Male (n=29,980)	Female (n=28,439)	*P* value
Age at screening (months), mean (SD)	17.33 (1.36)	17.31 (1.33)	.03
Screening outcome, n (%)			<.001
No risk	28,552 (95.2)	27,592 (97)	
Generic risk	917 (3.1)	581 (2)	
Medium risk	581 (1.9)	192 (0.7)	
High risk	153 (0.5)	74 (0.3)	

In total, 11.5% (26/227) of the participants identified as "high risk" by the CHAT but who were not referred to the SAU were considered out of sample for several reasons. Of these 26 participants, 2 (7.7%) were excluded as they had moved away and were no longer receiving care at the appropriate SAU. In total, 23.1% (6/26) of the participants were already receiving care at other facilities. Regarding the remaining 69.2% (18/26), no information was available; lack of data may have been due to a family pediatrician's failure to use the platform correctly or a shared decision with the family not to proceed with referral of the child to the SAU.

Participants found to be at "generic risk" and "medium risk" (2048/58,419, 3.5%) repeated the screening after a month, as outlined in the protocol; among these 2048 participants, 1206 (58.9%) follow-ups were made.

Of the 1206 patients for whom screening was repeated (mean age at repeat screening 18.41 months), 779 (64.6%) were classified as "no risk," and 427 (35.4%) were identified as at risk. Of these 1206 patients, 243 (20.1%) were at generic risk, 133 (11%) were at medium risk, and 51 (4.2%) were at high risk. In total, 92.5% (395/427) of the patients at risk (395/58,419, 0.7% of the total sample) were referred for fast-track diagnostic evaluation to the SAU. In total, 1% (596/58,419) of the children in the total sample (399/596, 66.9% male; 197/596, 33.1% female) were referred for fast-track evaluation to the SAU.

In total, 7.5% (32/427) of the patients identified as "at risk" at the repeated assessment but who were not referred to the SAU were considered out of sample. No information was available. Again, the lack of data may have been due to a pediatrician's failure to use the platform or a shared decision with the family not to proceed with referral.

Assessing the Effectiveness of Screening

The performance of the CHAT screening test was evaluated by comparing the actual test results with the patients' true diagnostic status, as determined through a gold standard assessment. Among the children who underwent screening, diagnostic investigation outcomes were available for a subsample drawn from the Valle Olona SAU. This selection was driven by clinical and organizational considerations and, therefore, represents a convenience sampling [47]. Although this sampling approach may limit generalizability, a comparison between the Valle Olona SAU cohort and the rest of the population was performed using 2-tailed *t* tests and chi-square tests (Table 3). Given the large overall sample size (N=58,419), even minimal absolute differences reached statistical significance (eg, mean age 17.31, SD 1.34 months vs mean age 17.46, SD 1.42 months in the other SAUs and the Valle Olona SAU, respectively; *P*<.001).

Table 3. Comparison of demographic and screening characteristics among the overall screened population, the population at other centers, and the Valle Olona specialized autism unit (SAU; N=58,419).

Variable	Total sample	Valle Olona SAU (n=3599)	Other SAUs (n=54,820)	*P* value
Age at screening (months), mean (SD)	17.32 (1.35)	17.46 (1.42)	17.31 (1.34)	<.001
Sex, n (%)				.62
Male	29,980 (51.3)	1832 (50.9)	28,148 (51.3)	
Female	28,439 (48.7)	1767 (49.1)	26,672 (48.7)	
Screening outcome, n (%)				.02
No risk	56,144 (96.1)	3427 (95.2)	52,717 (96.2)	
Generic risk	1498 (2.6)	107 (3)	1391 (2.5)	
Medium risk	550 (0.9)	47 (1.3)	503 (0.9)	
High risk	227 (0.4)	18 (0.5)	209 (0.4)	

The Valle Olona SAU included 62 family pediatricians, of whom 56 (90%) regularly used the screening platform, 2 (3%) had never logged onto it, and 4 (6%) had logged in but did not use it actively. During the study period, 3599 CHAT screenings were performed by primary care pediatricians and specialists. Figure 5 shows the outcomes for patients who completed the CHAT screening at both assessments, including referrals for further diagnostic investigation. In total, 0.1% (2/3599) of the

XSL•FO
RenderX

patients were not referred because they had moved away and were no longer followed up on by the SAU. Overall, 1.5% (55/3599) of the total screened children (37/55, 67.3% male; 18/55, 32.7% female) were referred for fast-track diagnostic evaluation at the SAU. Of these 55 children, 27 (49.1%) were diagnosed with ASD and 20 (36.4%) were diagnosed with other neurodevelopmental disorders, including 3 (15%) language disorders (2 male and 1 female), 14 (70%) cases of intellectual disability (n=9, 64.3% male; n=5, 35.7% female), 1 (5%) genetic syndrome (male), and 2 (10%) unspecified neurodevelopmental disorders (1 male and 1 female). In total, 1.8% (1/55) of the children were still awaiting a diagnosis, whereas 12.7% (7/55) exhibited clinical signs but did not meet the diagnostic threshold; in these cases, targeted intervention programs were planned at home and at school, with ongoing 6- and 12-month clinical follow-ups.

Finally, a post hoc power analysis conducted using PASS 2024 (NCSS, LLC) [48] confirmed the adequacy of the Valle Olona SAU subsample (n=3599) to estimate the proportion of children identified as "at risk" (*P*=.01). At a 95% confidence level, the estimated interval ranged from 0.0070 to 0.0138, corresponding to a margin of error of –0.0034 to +0.0034, thus supporting the precision and robustness of the subsample-based estimates.

Figure 5. Effectiveness of screening: flowchart of screenings and diagnostic outcomes conducted by family pediatricians from January 2022 to September 2024 in the Valle Olona specialized autism unit subsample. ASD: autism spectrum disorder; CHAT: Checklist for Autism in Toddlers.

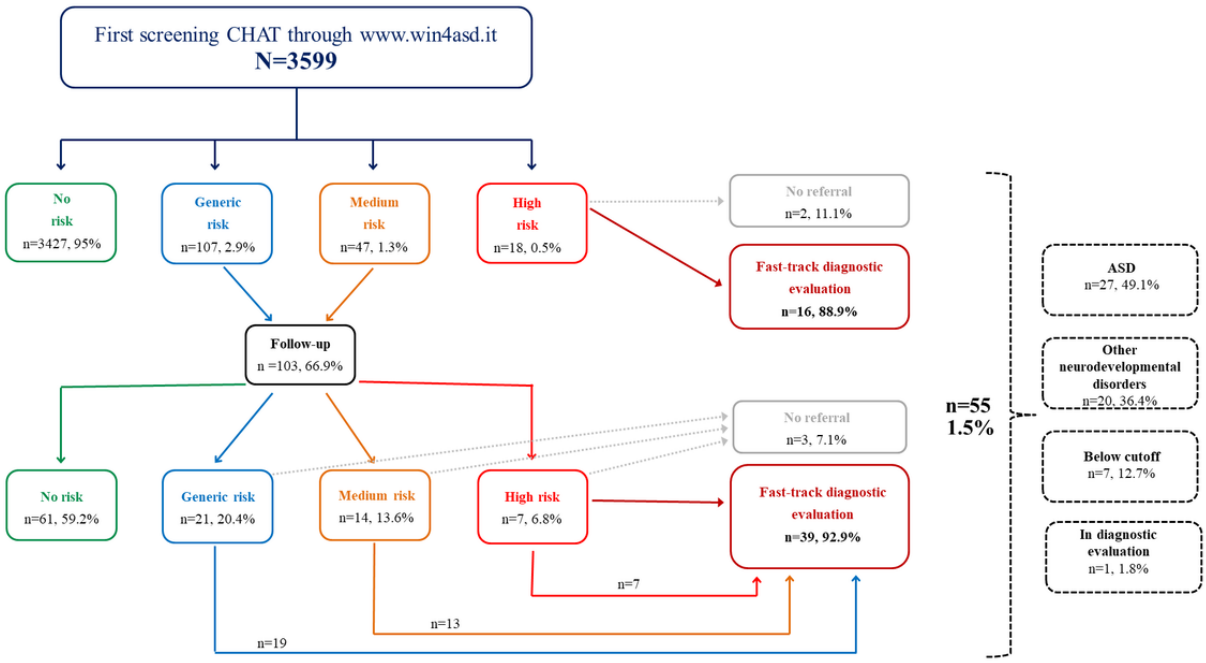

Discussion

Principal Findings

In this work, we presented the Win4ASD platform, an innovative digital tool designed to support early identification and diagnosis of children with ASD within the general population, as a model of preventive medicine. It holds significant potential for building a care network that supports children with ASD from early diagnosis to timely clinical intervention. We described how the platform was integrated into the regional NHS and demonstrated its clinical effectiveness for early ASD screening through a retrospective observational study. The results of this work highlight the political and institutional efforts behind the platform, as well as its clinical impact on the general population.

The first key finding is the commitment of family pediatricians to using the platform owing to the institutional agreement (a supplementary agreement for the Lombardy region) [49]. In total, 72.5% (909/1253) of all family pediatricians of the regional NHS were actively using the platform. This high level of participation makes ASD screening an integral part of the pediatric surveillance system through the systematic use of the CHAT. Furthermore, family pediatricians' engagement increased over time. We anticipate that continued use of the platform, along with ongoing dissemination and training through institutional channels, will further improve this percentage.

Additionally, our platform enabled the screening of 58,419 toddlers from the general population in Lombardy. According to Italian statistical data (Italian National Institute of Statistics), the estimated pediatric population aged 18 to 22 months between 2022 and 2024 was 69,474 [50]. On the basis of these figures, it can be estimated that 84.08% of children in Lombardy may have received well-child checkups and ASD screening. While a detailed analysis of screening adherence was not possible in this study, a large proportion of families in the region chose to use this free screening service. Moreover, compared with other similar studies that have screened for or diagnosed ASD through digital platforms or telemedicine, our study had a much larger sample size [5,25-27] and is innovative in terms of the network between family pediatricians and local care units for child psychiatry, which the platform succeeded in creating.

XSL•FO
RenderX

On the basis of the total sample (N=58,419), 1% (596/58,419) of the children were classified as “at risk” at the first or repeated screening and were referred for fast-track diagnostic evaluation. Although this proportion is numerically similar to global ASD prevalence estimates, it should be interpreted as the rate of screening positivity, not as a measure of confirmed ASD prevalence [18-20]. The results also show an overall higher number of male than female children. However, the statistically significant comparison between male and female children is probably due to the large sample size, as the absolute differences were quantitatively minimal.

Looking at the data on the effectiveness of screening from a specific SAU, 3599 screenings were performed, and 55 (1.5%) of these children were referred for fast-track diagnostic evaluation. This selection was driven by clinical and organizational considerations and represents a convenience sampling. From further comparative analyses with the general sample and from the post hoc analysis, it can be inferred that all differences between the 2 groups (general and Valle Olona SAU) were quantitatively negligible, supporting the overall homogeneity of the populations, the representativeness of the Valle Olona SAU subsample, and the precision and robustness of the subsample-based estimates.

In this SAU, 90% (56/62) of pediatricians actively used the platform, which led to an adherence rate higher than the regional average, making this SAU an effective model for the screening system. As mentioned above, the data for the total screenings and the percentage of children referred for further diagnosis in this SAU are in line with prevalence data [18-20]. Of the 55 patients referred for evaluation, 45 (81.8%) received a diagnosis, demonstrating the effectiveness of the screening system in identifying clinical signs. Specifically, of the 55 children, 27 (49.1%) received an ASD diagnosis and 20 (36.4%) were diagnosed with other neurodevelopmental disorders. This is in line with the findings of the authors of the CHAT, Baron-Cohen et al [51], who point out that the CHAT intercepts the risk of other neurodevelopmental disorders, especially in the medium or generic risk.

In short, the Win4ASD platform is a viable model for improving the NHS by enhancing continuity of care for children with ASD. Looking ahead, we anticipate even more efficient implementation of the platform, spreading this model further within the NHS and increasing screening coverage across different age groups of the child population. As part of upcoming developments, the Win4ASD platform is being enhanced with the integration of an additional screening tool, the Strengths and Difficulties Questionnaire (SDQ) [52]. The SDQ is a brief behavioral screening questionnaire for children aged 2 to 17 years. It comes in various versions to cater to the needs of researchers, clinicians, and educators. The version for children aged 2 to 4 years is now part of the platform. Similar to the CHAT, the platform allows pediatricians to complete all items of the SDQ, automatically score the screening, and facilitate seamless communication between pediatricians and the SAUs. The SDQ will be available on the platform for children aged 23 to 36 months, providing extended coverage for the active surveillance of neurodevelopmental disorders. The service is free of charge and serves as an additional screening layer within the NHS.

Limitations

This study has some limitations that must be considered when interpreting its findings. First, this work is not experimental research, which may introduce potential methodological bias. Specifically, we lacked statistical data on the number of children aged 18 to 22 months residing in Lombardy between 2022 and September 2024. Additionally, we did not have data on the number of children in Lombardy diagnosed with ASD or other neurodevelopmental disorders by public or private agencies to compare with the data from the platform.

In addition, we acknowledge that the use of data from a single SAU, along with the lack of information on children at generic, medium, or high risk who, for clinical reasons or parental choice, did not complete the screening and diagnostic process monitored by the platform, represents a limitation regarding the generalizability of our findings. To address this, we plan to expand the outcome dataset by incorporating cases from other SAUs, which will provide a broader and more representative perspective. Nevertheless, despite the potential of the platform, the emerging network, and the high number of screenings carried out, our results cannot yet be generalized to the wider regional context.

Furthermore, as illustrated in the flowcharts (Figures 4 and 5), some children identified as being at risk of neurodevelopmental disorders through the CHAT were not referred to SAU services for diagnostic assessment. This may be due to a variety of clinical and nonclinical factors, for instance, parental decisions regarding the continuation of diagnostic evaluation or a family’s relocation to another region.

Conclusions

Owing to the political and institutional support that the platform has received over the years, it has been possible to make improvements that have led to an increasingly widespread dissemination of the screening system. Early detection of risk and timely intervention can prevent the progressive development of behavioral atypicalities; have a significant impact on the quality of life of the child and their family; and, finally, allow for a reduction in health care costs and a consequent improvement in the organization of the health care system.

Acknowledgments

The authors acknowledge all partners and collaborators of the project. For this paper, the generative AI tool ChatGPT (OpenAI [53]) was used to assist with language revision.

XSL•FO
RenderX

Funding

This work has been supported by the Italian Ministry of Health (Ricerca Corrente 2025-2026) and the *Fondo per la cura dei soggetti con disturbo dello spettro autistico* of the Italian Ministry of Health through the Regione Lombardia *Attivare Una risposta Territoriale INclusiva e Continuativa per le persone con Autismo* (AUTINCA) Project (deliberation XII/277; May 15, 2023).

Authors' Contributions

PC, NB, IL, LF, and MM designed the study. NB, ER, SBC, and NV conducted literature searches. NB, NV, SBC, MF, EM, and VM collected the data. NB, PC, and ER analyzed the data. NB and ER wrote the first draft of the manuscript. MM, IL, LF, and PC critically reviewed the manuscript. All authors have approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Telemedicina. Ministero della Salute. URL: https://tinyurl.com/523uxb4a [accessed 2025-01-18]
2. Telehealth: defining 21st century care. American Telemedicine Association. URL: https://www.americantelemed.org/resource/why-telemedicine/ [accessed 2025-02-21]
3. Galmiche S, Rahbe E, Fontanet A, Dinh A, Bénézit F, Lescure FX, et al. Implementation of a self-triage web application for suspected COVID-19 and its impact on emergency call centers: observational study. J Med Internet Res 2020 Nov 23;22(11):e22924 [FREE Full text] [doi: 10.2196/22924] [Medline: 33147165]
4. Basch E, Deal AM, Dueck AC, Scher HI, Kris MG, Hudis C, et al. Overall survival results of a trial assessing patient-reported outcomes for symptom monitoring during routine cancer treatment. JAMA 2017 Jul 11;318(2):197-198 [FREE Full text] [doi: 10.1001/jama.2017.7156] [Medline: 28586821]
5. Valentine AZ, Hall SS, Young E, Brown BJ, Groom MJ, Hollis C, et al. Implementation of telehealth services to assess, monitor, and treat neurodevelopmental disorders: systematic review. J Med Internet Res 2021 Jan 20;23(1):e22619 [FREE Full text] [doi: 10.2196/22619] [Medline: 33326409]
6. Holmes EA, O'Connor RC, Perry VH, Tracey I, Wessely S, Arseneault L, et al. Multidisciplinary research priorities for the COVID-19 pandemic: a call for action for mental health science. Lancet Psychiatry 2020 Jun;7(6):547-560 [FREE Full text] [doi: 10.1016/S2215-0366(20)30168-1] [Medline: 32304649]
7. Liu M, Ma Z. A systematic review of telehealth screening, assessment, and diagnosis of autism spectrum disorder. Child Adolesc Psychiatry Ment Health 2022 Oct 08;16(1):79. [doi: 10.1186/s13034-022-00514-6] [Medline: 36209100]
8. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5). Washington, DC: American Psychiatric Association; 2013.
9. Global report on children with developmental disabilities: from the margins to the mainstream. UNICEF. 2023 Sep 15. URL: https://www.unicef.org/documents/global-report-children-developmental-disabilities [accessed 2025-02-21]
10. World report on disability. World Health Organization. 2011 Dec 14. URL: https://www.who.int/publications/i/item/9789241564182 [accessed 2025-02-21]
11. Cioni G, Inguaggiato E, Sgandurra G. Early intervention in neurodevelopmental disorders: underlying neural mechanisms. Dev Med Child Neurol 2016 Mar 29;58 Suppl 4(S4):61-66 [FREE Full text] [doi: 10.1111/dmcn.13050] [Medline: 27027609]
12. Donatini A. The Italian health care system. Emilia Romagna Regional Health Authority, IT: International profiles of health care systems; 2015. URL: https://www.issuelab.org/resources/25100/25100.pdf#page=97 [accessed 2026-02-05]
13. France G, Taroni F, Donatini A. The Italian health-care system. Health Econ 2005 Sep;14(S1):S187-S202 [FREE Full text] [doi: 10.1002/hec.1035]
14. Corsello G, Ferrara P, Chiamenti G, Nigri L, Campanozzi A, Pettoello-Mantovani M. The child health care system in Italy. J Pediatr 2016 Oct;177S:S116-S126. [doi: 10.1016/j.jpeds.2016.04.048] [Medline: 27666260]
15. Crescere in salute: rete pediatrica. Agenzia di Tutela della Salute di Bergamo. URL: https://ats-bg.it/crescere-salute-rete-pediatrica-provincia-bergamo [accessed 2025-02-21]
16. Hyman SL, Levy SE, Myers SM, Council on Children With Disabilities, Section on Developmental and Behavioral Pediatrics. Identification, evaluation, and management of children with autism spectrum disorder. Pediatrics 2020 Jan;145(1):e20193447. [doi: 10.1542/peds.2019-3447] [Medline: 31843864]
17. Zeidan J, Fombonne E, Scorah J, Ibrahim A, Durkin MS, Saxena S, et al. Global prevalence of autism: a systematic review update. Autism Res 2022 May;15(5):778-790 [FREE Full text] [doi: 10.1002/aur.2696] [Medline: 35238171]
18. Maenner MJ, Warren Z, Williams AR, Amoakohene E, Bakian AV, Bilder DA, et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2020. MMWR Surveill Summ 2023 Mar 24;72(2):1-14 [FREE Full text] [doi: 10.15585/mmwr.ss7202a1] [Medline: 36952288]
19. Data and statistics on autism spectrum disorder. Centers for Disease Control and Prevention. 2025 May 27. URL: https://www.cdc.gov/autism/data-research/index.html [accessed 2025-02-21]

XSL•FO
RenderX

20. Bougeard C, Picarel-Blanchot F, Schmid R, Campbell R, Buitelaar J. Prevalence of autism spectrum disorder and co-morbidities in children and adolescents: a systematic literature review. Front Psychiatry 2021;12:744709 [FREE Full text] [doi: 10.3389/fpsyt.2021.744709] [Medline: 34777048]
21. Scattoni ML, Fatta LM, Micai M, Sali ME, Bellomo M, Salvitti T, et al. Autism spectrum disorder prevalence in Italy: a nationwide study promoted by the Ministry of Health. Child Adolesc Psychiatry Ment Health 2023 Oct 28;17(1):125 [FREE Full text] [doi: 10.1186/s13034-023-00673-0] [Medline: 37898807]
22. Lord C, Brugha TS, Charman T, Cusack J, Dumas G, Frazier T, et al. Autism spectrum disorder. Nat Rev Dis Primers 2020 Jan 16;6(1):5 [FREE Full text] [doi: 10.1038/s41572-019-0138-4] [Medline: 31949163]
23. Guthrie W, Wetherby AM, Woods J, Schatschneider C, Holland RD, Morgan L, et al. The earlier the better: an RCT of treatment timing effects for toddlers on the autism spectrum. Autism 2023 Mar 15;27(8):13623613231159153 [FREE Full text] [doi: 10.1177/13623613231159153] [Medline: 36922406]
24. Dimian AF, Symons FJ, Wolff JJ. Delay to early intensive behavioral intervention and educational outcomes for a Medicaid-enrolled cohort of children with autism. J Autism Dev Disord 2021 Apr;51(4):1054-1066. [doi: 10.1007/s10803-020-04586-1] [Medline: 32642958]
25. Vivanti G, Algur Y, Ryan V, McClure LA, Fein D, Stahmer AC, et al. The impact of using standardized autism screening on referral to specialist evaluation for young children on the autism spectrum: a cluster-randomized controlled trial. J Am Acad Child Adolesc Psychiatry 2025 Jun;64(6):686-698 [FREE Full text] [doi: 10.1016/j.jaac.2024.08.502] [Medline: 39454668]
26. Abbas H, Garberson F, Glover E, Wall DP. Machine learning approach for early detection of autism by combining questionnaire and home video screening. J Am Med Inform Assoc 2018 Aug 01;25(8):1000-1007 [FREE Full text] [doi: 10.1093/jamia/ocy039] [Medline: 29741630]
27. Abbas H, Garberson F, Liu-Mayo S, Glover E, Wall DP. Multi-modular AI approach to streamline autism diagnosis in young children. Sci Rep 2020 Mar 19;10(1):5014 [FREE Full text] [doi: 10.1038/s41598-020-61213-w] [Medline: 32193406]
28. Raccomandazioni della Linea Guida per la diagnosi e il trattamento di bambini e adolescenti con disturbo dello spettro autistico. Istituto Superiore di Sanità. 2021 Nov 30. URL: https://www.iss.it/-/raccomandazioni-lg-diagnosi-trattamento-di-bambini-adolescenti-con-asd [accessed 2025-02-21]
29. Morgano GP, Fulceri F, Nardocci F, Barbui C, Ostuzzi G, Papola D, et al. Introduction and methods of the evidence-based guidelines for the diagnosis and management of autism spectrum disorder by the Italian National Institute of Health. Health Qual Life Outcomes 2020 Mar 26;18(1):81 [FREE Full text] [doi: 10.1186/s12955-020-01320-4] [Medline: 32216833]
30. Win4ASD. URL: https://www.win4asd.it [accessed 2025-02-21]
31. Colombo P, Buo N, Busti Ceccarelli S, Molteni M. Integrating a new online platform in primary care for early detection, referral, and intervention in autism spectrum disorder: the first Italian pivotal project. Brain Sci 2022 Feb 12;12(2):256 [FREE Full text] [doi: 10.3390/brainsci12020256] [Medline: 35204020]
32. Dati sulla popolazione e l'occupazione. ISTAT. URL: https://esploradati.istat.it/databrowser/#/it/dw/categories [accessed 2025-02-11]
33. Baron-Cohen S, Allen J, Gillberg C. Can autism be detected at 18 months? The needle, the haystack, and the CHAT. Br J Psychiatry 1992 Dec 02;161(6):839-843. [doi: 10.1192/bjp.161.6.839] [Medline: 1483172]
34. Piano operativo regionale autismo. Regione Lombardia. 2024. URL: https://tinyurl.com/5n6bhue9 [accessed 2025-02-11]
35. Lord C, Rutter M, Dilavore P, Risi S, Gotham K, Bishop S. Autism Diagnostic Observation Schedule, Second Edition. Torrence, CA: Western Psychological Services; 2012.
36. Lanfranchi S, Rea M, Ferri R, Vianello R. Studio di validazione e standardizzazione italiana delle Griffiths III. Catalogo dei Prodotti Della Ricerca. 2019. URL: https://iris.uniroma1.it/handle/11573/1355472 [accessed 2025-02-21]
37. Roid GH, Miller LJ, Pomplun M, Koch C. Leiter International Performance Scale, Third Edition. Wood Dale, IL: Stoelting Company; 2013.
38. Schopler E, Lansing M, Reichler R, Marcus L. Psychoeducational Profile Third Edition (PEP‐3). Austin, TX: Pro-Ed Inc; 2004.
39. Schopler E, Lansing MD, Reichler RJ, Marcus LM. Pep 3 Examiner's Manual With Tables. Austin, TX: Pro-Ed Inc; 2005.
40. Sparrow SS, Cicchetti DV, Balla DA. Vineland Adaptive Behavior Scales, Second Edition (Vineland-II). London, UK: Pearson; 2005.
41. Harrison PL, Oakland T. ABAS-II assessment methods. In: Oakland T, Harrison PL, editors. Adaptive Behavior Assessment System-II. Cambridge, UK: Academic Press; 2008:37-49.
42. Sabbadini L. APCM-2 Abilità Prassiche e della Coordinazione Motoria – 2a Edizione. Göttingen, Germany: Hogrefe; 2015.
43. Caselli MC, Bello A, Rinaldi P, Stefanini S, Pasqualetti P. Il Primo Vocabolario Del Bambino: Gesti, Parole e Frasi. Milan, Italy: Franco Angeli; 2018.
44. CHAT – checklist for autism in toddler. Autism Research Centre. URL: https://docs.autismresearchcentre.com/tests/CHAT_Italian.pdf [accessed 2025-02-18]
45. Lord C, Rutter M, Le Couteur A. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. J Autism Dev Disord 1994 Oct;24(5):659-685. [doi: 10.1007/BF02172145] [Medline: 7814313]

XSL•FO
RenderX

46. IBM SPSS statistics for Windows; version 26.0. IBM Corp. 2019. URL: https://www.ibm.com/products/spss-statistics [accessed 2026-02-17]
47. Stratton S. Population research: convenience sampling strategies. Prehosp Disaster Med 2021 Jul 21;36(4):373-374 [FREE Full text] [doi: 10.1017/s1049023x21000649]
48. PASS 2026. NCSS Statistical Software. URL: https://www.ncss.com/ [accessed 2025-11-02]
49. Accordo integrativo regionale pediatria di famiglia. Regione Lombardia. 2025. URL: https://tinyurl.com/fwf77s44 [accessed 2025-02-18]
50. Italia, regioni, province. ISTAT. URL: https://esploradati.istat.it/databrowser/#/it/dw/categories/IT1,POP,1.0/POP_POPULATION/DCIS_POPRES1/IT1,22_289_DF_DCIS_POPRES1_1,1.0 [accessed 2025-02-03]
51. Baron-Cohen S, Wheelwright S, Cox A, Baird G, Charman T, Swettenham J, et al. Early identification of autism by the CHecklist for Autism in Toddlers (CHAT). J R Soc Med 2000 Oct 01;93(10):521-525 [FREE Full text] [doi: 10.1177/014107680009301007] [Medline: 11064690]
52. Goodman R. The Strengths and Difficulties Questionnaire: a research note. J Child Psychol Psychiatry 1997 Jul;38(5):581-586. [doi: 10.1111/j.1469-7610.1997.tb01545.x] [Medline: 9255702]
53. ChatGPT. OpenAI. URL: https://chatgpt.com/c/698b0793-c320-8389-b4d0-4f2f40f6f63d

Abbreviations

ADI-R: Autism Diagnostic Interview–Revised
ADOS-2: Autism Diagnostic Observation Schedule, Second Edition
ASD: autism spectrum disorderCHAT: Checklist for Autism in ToddlersNHS: National Health Service
SAU: specialized autism unit
SDQ: Strengths and Difficulties Questionnaire
Win4ASD: Web Italian Network for Autism Spectrum Disorder

Edited by A Stone; submitted 21.Mar.2025; peer-reviewed by RS Kucukardali, HC Koh; comments to author 09.Sep.2025; revised version received 21.Jan.2026; accepted 21.Jan.2026; published 24.Feb.2026.

Please cite as:
Buo N, Rosi E, Busti Ceccarelli S, Ferrario M, Maiorca V, Morandi E, Viganó N, Limosani I, Falcone L, Molteni M, Colombo P Rising to the Challenge of Early Screening in Primary Health Care Through the Web Italian Network for Autism Spectrum Disorder (Win4ASD) in the Pediatric Population: Retrospective Observational Study
J Med Internet Res 2026;28:e74302
URL: https://www.jmir.org/2026/1/e74302
doi:10.2196/74302
PMID:

©Noemi Buo, Eleonora Rosi, Silvia Busti Ceccarelli, Mariarosa Ferrario, Valerio Maiorca, Erika Morandi, Nicole Viganó, Ivan Limosani, Laura Falcone, Massimo Molteni, Paola Colombo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Effect of Veteran Race and Socioeconomic Status on Enrollment in Remote Patient Monitoring for Hypertension: Retrospective Observational Cross-Sectional Study

Hannah Robin Friedman[1], PhD; Hillary J Mull[1,2], MPP, PhD; Jenice Ria Guzman-Clark[3], GNP-BC, PhD; Daniel J Sturgeon[4,5], MS; Marva V Foster[1,2], RN, PhD

[1]Center for Health Optimization and Implementation Research, VA Boston Healthcare System, Boston, MA, United States
[2]Chobanian and Avedisian School of Medicine, Boston University, Boston, MA, United States
[3]Southern Arizona VA Health Care System, Tucson, AZ, United States
[4]Stanford Medicine, Palo Alto, CA, United States
[5]Health Economics Resource Center, VA Palo Alto Health Care System, Palo Alto, CA, United States

Corresponding Author:
Hannah Robin Friedman, PhD
Center for Health Optimization and Implementation Research
VA Boston Healthcare System
Lafayette City Center
2 Ave de Lafayette
Boston, MA, 02111
United States
Phone: 1 617 232 9500
Email: hannah.friedman2@va.gov

Abstract

Background: Black veterans and veterans from lower socioeconomic backgrounds are more likely to have uncontrolled hypertension. One potential explanatory factor is reduced access to specific treatments that result in improved chronic disease management. In the Veterans Health Administration (VHA), veterans with hypertension may enroll in a remote patient monitoring (RPM) program, which consists of patient education, daily home blood pressure (BP) monitoring, health coaching, and case management. Barriers for socioeconomically disadvantaged patients may exist for similar programs in other health systems; however, the VHA is an integrated health care system, and these barriers may differ for veteran populations.

Objective: The objective of this study was to assess the relationship between veteran race and neighborhood socioeconomic status and the likelihood of enrolling in the VHA RPM program.

Methods: The study sample included VHA-enrolled veterans with a diagnosis of hypertension (average BP >130/80 mm Hg on ≥2 BP readings) between fiscal years 2020 and 2023. We ran random-effects logistic regression models to assess the relationship between veteran race and Area Deprivation Index and RPM enrollment each year, controlling for potential demographic and clinical confounders. For sensitivity analysis, we limited our sample to veterans with stage 2 hypertension (BP >140/90 mm Hg) and on antihypertensive medication.

Results: Overall use of RPM was low, with only 4.1% (56,553/1,390,995; 95% CI 4%-4.1%) of veterans being enrolled in RPM. Black veterans, who represented 26.6% (n=35,096) of all veterans, were more likely (odds ratio [OR] 1.65, 95% CI 1.59-1.70) to enroll in RPM compared to White veterans. Asian American or Pacific Islander veterans were less likely to enroll (OR 0.83, 95% CI 0.74-0.94). We found no meaningful association between Area Deprivation Index and RPM enrollment (OR 1.00, 95% CI 0.99-1.00). When limiting our sample to those with stage or grade 2 hypertension, we found a similar association (OR 1.61, 95% CI 1.50-1.72) between Black race and RPM enrollment but no significant association with Asian or Pacific Islander race (OR 1.02, 95% CI 0.80-1.29).

Conclusions: Prior research on RPM in veterans has examined duration or outcomes of RPM enrollment but not the probability of initial enrollment. We found higher enrollment rates in the VHA RPM program among Black veterans but slightly lower enrollment among Asian American or Pacific Islander veterans. Higher enrollment among Black veterans and among those with higher comorbidity burden suggests that the VHA RPM program is successfully reaching those who could most benefit, despite

XSL•FO
RenderX

low overall enrollment. Given the low enrollment in RPM, future research should focus on improving uptake among veterans who could additionally benefit from the program. Non-VHA systems, particularly those serving low-income or socioeconomically disadvantaged areas, should explore subsidized or free RPM programs for eligible patients similar to the VHA's no-cost model for veterans.

(J Med Internet Res 2026;28:e78423) doi:10.2196/78423

KEYWORDS

hypertension; veterans' health; remote patient monitoring; telehealth; health equity

Introduction

Remote patient monitoring (RPM) is a form of telehealth that allows clinicians to track patient vital signs and health metrics outside traditional health care settings [1]. This can include using wireless-enabled blood pressure (BP) cuffs, thermometers, pulse oximeters, scales, and glucometers to transmit vital signs to a nurse from the patient's home on a regular basis. Integrating RPM into chronic disease management can potentially improve patient quality of life and mitigate health care costs by enabling the timely detection of health deterioration, thereby reducing unnecessary emergency department visits and hospital admissions [2,3]. The Veterans Health Administration (VHA) is an integrated health care system that provides health care for more than 9 million of the 18 million veterans in the United States [4]. As a comprehensive health care system, the VHA offers RPM services at no cost to veterans and has existed since the early 2000s serving veterans with a variety of chronic conditions [5]. The VHA RPM model is unique due to its nationwide coverage, centralized funding, and focus on a specific patient population. The essence of RPM, as implemented in the VHA, is ongoing assessment, monitoring, patient education, health coaching, and nurse case management of veterans in their place of residence. It also includes the provision of appropriate information to the veterans' primary care team or specialty care team, as well as the health care system, to enable timely care [5].

Veterans targeted for RPM are those who have high health service use (eg, frequent emergency room visits) due to chronic diseases, who have uncontrolled chronic diseases, and who agree to participate in daily submission of health data for monitoring [6,7]. Referrals can be made by any VHA-licensed independent clinician (physician, nurse practitioner, or physician assistant) or other clinicians when following an approved local protocol. Once enrolled, veterans are expected to submit vital sign and other relevant biometric data every day and respond to daily disease-specific questions.

Hypertension, or high BP, is one of the most common chronic medical conditions managed by the VHA RPM program [8]. Hypertension is a major contributor to cardiovascular morbidity and mortality in the United States and remains one of the most important modifiable cardiovascular risk factors [9]. Prevalence data indicate that 47% of adults in the United States have hypertension, which is equivalent to an estimated 122 million adults [10]. More than half of these individuals with hypertension have BP that is inadequately controlled [11]. There are racial and ethnic differences in hypertension prevalence and treatment in the United States, with more Black, Hispanic, and Asian adults having poor control of their BP compared to White adults despite Black adults having a higher prevalence of hypertension [12]. Hypertension is the most common chronic condition among patients in the VHA. Recent research has found that 87% of veterans have high BP and, of those, most (66%) are considered to have uncontrolled hypertension [13]. Effective control of high BP is of key importance for reducing the risk of hypertension-related cardiovascular disease complications.

However, several factors have been found to impact hypertension control, key among them being patient residence in socioeconomically disadvantaged neighborhoods [14-16]. In the VHA, Black veterans have the poorest control rates for hypertension compared to other ethnic backgrounds even after adjusting for access to care and socioeconomic status [17]. Black veterans also have considerably higher rates of cardiovascular-related hospitalizations, an indicator of poor hypertension control [18]. RPM has been posited as a potential intervention to decrease this racial and ethnic disparity in hypertension control and has been shown to improve hypertension control in patients of a large health care organization [3]; however, a small pilot study that compared a sample of primary care patients who mostly comprised minority groups or groups with low socioeconomic status who used RPM to a control group who did not use RPM found no difference in BP control after 90 days, but found that RPM enrollment was associated with decrease in all-cause clinic visits [19].

There is no singular socioeconomic determinant implicated in disadvantaged neighborhoods. Multidimensional factors such as economic (eg, low income and blue-collar jobs), social (eg, low educational level, nonvoter status, and high unemployment), and physical (eg, graffiti and low-rent housing) conditions, among others, can influence a neighborhood [20,21]. These neighborhood-level characteristics have been increasingly recognized as playing a critical role in health [22]. The use of RPM has been suggested as a potential solution to overcome the geographical limitations of health care services [23,24]. However, a review of existing literature on non-VHA populations has shown that some studies indicate the expected directionality (ie, greater neighborhood socioeconomic disadvantage is associated with decreased use of RPM), whereas others report mixed and null findings [25-28]. The pathway between neighborhood deprivation and RPM use for hypertension management in the VHA is poorly understood. Previous VHA studies have examined adherence to RPM and enrollment rates in RPM but have not examined the role of neighborhood deprivation on RPM enrollment [8,29-31]. Non-VHA studies have identified barriers to RPM among Black patients, including lower overall participation in clinical trials

XSL•FO
RenderX

and reduced precision and accuracy of wearable devices on people with darker skin tones [32,33]. The overall evidence points to the fact that living in a socioeconomically disadvantaged neighborhood may impact hypertension and possibly RPM use in non-VHA populations.

What is not known is whether neighborhood deprivation impacts enrollment in RPM among veterans; such knowledge will identify populations for whom appropriate interventions can be implemented. Therefore, the purpose of this study was to investigate whether veteran race or residence in a socioeconomically disadvantaged neighborhood was associated with the probability that a patient with hypertension will enroll in the VHA RPM program.

Methods

We used the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines to ensure completeness and accuracy when reporting our methods and results [34].

Population and Sample

This retrospective cross-sectional study included veterans with a diagnosis of hypertension between October 1, 2019, and September 30, 2023. We used data from the VHA's Corporate Data Warehouse to identify all covariates and outcomes for this study. Our initial sample consisted of all veterans with an *International Classification of Diseases, 10th Revision*, diagnosis code for hypertension on or before October 1, 2022 [35]. We excluded observations in which the veteran was not enrolled in the VHA at the start of the fiscal year (FY), did not have a diagnosis of hypertension before the start of the FY, or died before the end of the FY. We then limited our sample to veterans with uncontrolled hypertension.

Definition of Uncontrolled Hypertension

The American College of Cardiology and American Heart Association have defined hypertension as a BP of at least 130/80 mm Hg over a minimum of 2 BP readings [36,37]. To ensure that our sample consisted of those with hypertension in accordance with clinical practice guidelines, we limited the study sample to veterans with ≥2 days of recorded BP readings in the prior year and with an average BP reading of at least 130/80 mm Hg across all readings. For sensitivity analysis, we further limited the sample to those with an average BP of at least 140/90 mm Hg (hypertension stage or grade 2) despite being prescribed antihypertensive medications [37].

Measures

Outcome

Our outcome was any enrollment in the VHA's hypertension RPM program each year. Veterans were considered to have enrolled in RPM if any encounter had an RPM VHA stop code (683, 684, 685, or 686) included as a primary or secondary stop code in electronic health record data. VHA stop codes are administrative codes used to categorize and track outpatient clinical services and have been used in previous studies [38,39]. We did not differentiate between new enrollment or continued use of RPM from patients enrolled in RPM during the prior year. We also did not consider the duration of enrollment as the focus of this study was veteran enrollment in the RPM program.

Primary Independent Variables

For enrollment-adjusted regression models, our primary independent variables comprised veteran race (White; Black; Asian American, Native Hawaiian, Pacific Islander; American Indian, Alaska Native; and other race, unknown, or multiracial) and the Area Deprivation Index (ADI) of the census block area of the veteran's primary address. When possible, veteran race in the Corporate Data Warehouse is self-reported, and prior research has found high agreement between Corporate Data Warehouse race and self-reported race and ethnicity [40]. The ADI is an established composite measure of an area's overall socioeconomic vulnerability and includes measures of median home value, poverty rate, and education, among others. For analysis, the unit of change was a 10-percentile change in ADI national ranking.

Control Variables

We controlled for patients' overall health, specific comorbidities, age, sex, rurality, VHA priority group, and distance to the nearest VHA facility. Veterans' overall health was assessed using the Elixhauser Comorbidity Index, and we controlled separately for a diagnosis of diabetes mellitus due to complications associated with comorbid diabetes and hypertension, as well as the presence of a separate RPM program for veterans with diabetes. We also controlled for the type of hypertension (uncomplicated or with complications) as defined by the Elixhauser Comorbidity Index [41]. We described veterans' rurality using the VHA classification of urban, rural, and highly rural, which is originally based on Rural-Urban Commuting Area codes. A VHA priority group is a category assigned to veterans to determine the amount, if any, they may have to pay for health care. Veterans are classified into priority groups 1 to 8. We aggregated priority groups 1 to 3 (service-connected disability), priority group 4 (non–service-connected catastrophic disability or receiving VHA aid benefits), priority group 5 (low income or Medicaid eligible), and priority groups 6 to 8 (service history and copayments required for some services). All variables reflected the status of the patient at the beginning of the FY. For covariates for which data were missing (rurality, priority group, ADI, and comorbidities), we used multiple imputation to generate replacement values. Data processing and analysis were conducted using SAS Enterprise Guide (SAS Institute) and Stata (version 18; StataCorp).

Analysis

For descriptive analysis, we assessed the veteran characteristics by RPM enrollment; we used chi-square tests (for categorical variables) and Wilcoxon rank sum tests (for continuous variables) to assess the significance of differences between White veterans, Black veterans, and veterans of other races. We ran random-effects logistic regression models to assess the likelihood of a veteran being enrolled in RPM each year from FYs 2020-2023, with random intercepts for each individual veteran. We separated the cohort into person-years with 1 observation per patient per year for a maximum of 4

XSL•FO
RenderX

observations per patient, although not all patients were eligible for inclusion in all 4 years. Separating patients into person-years allowed for more accurate assessments of patients' potential confounders, such as comorbidities, age, location, and BP control, all of which could change over the course of 4 years. Using person-year as the unit of analysis also allowed us to account for the COVID-19 pandemic, which began in FY 2020 and substantially impacted health care use patterns, including RPM participation. We used multiple tests and metrics to assess model goodness of fit using log-likelihood ratio tests and also tested for potential collinearity using variance inflation factors.

Sensitivity Analysis

For sensitivity analysis, we limited our sample to those with stage or grade 2 hypertension (average BP reading>140/90 mm Hg) and presence of at least one hypertension medication in the 6 months before the start of the FY to focus on those who might be most in need of closer monitoring.

Ethical Considerations

This study was approved by the US Department of Veterans Affairs (VA) Boston Institutional Review Board (1765472-5; approval date: January 18, 2024). As this study involved only retrospective analysis of a large observational dataset derived from electronic health record data where risk was deemed minimal to study participants, we received a waiver for protected health information and a waiver of informed consent from the institutional review board. All identifiable study data were stored and analyzed on a secure server operated by the VHA Informatics and Computing Infrastructure, and no identifiable data were ever moved from that server. No identification of individual participants is included in this paper. As this study consisted of analysis of a secondary dataset, no participants received compensation.

Results

Overview

Among those with complete covariate data and who met the eligibility criteria (Figure 1), we identified 2,411,787 total person-years during FYs 2019-2023 among 1,390,995 unique veterans.

Figure 1. Veterans with hypertension eligible for inclusion in study sample between fiscal years 2019 and 2023.

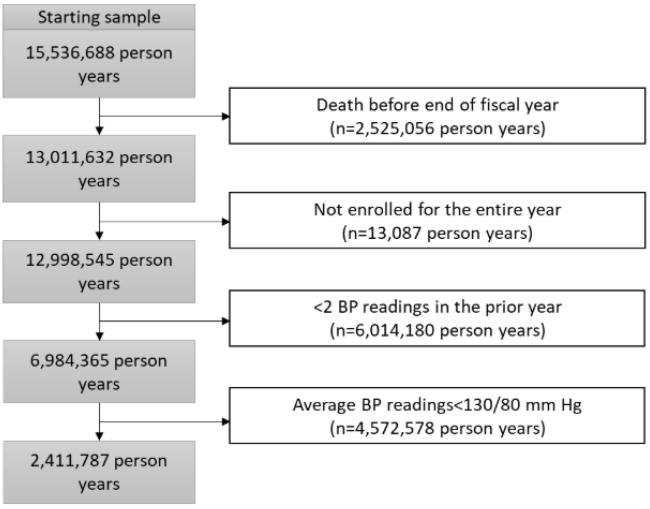

In Table 1, we share the characteristics of all veterans in our sample stratified by whether they were ever enrolled in RPM during our study period. If veterans appeared more than once in our sample, time-varying characteristics (eg, age and comorbidity) reflect their status during the first year they were eligible for inclusion in our sample. For priority group (n=6482 person-years), rurality (n=2320 person-years), ADI (n=36,351 person-years), and comorbidity (Elixhauser Comorbidity Index; diabetes and hypertension with complications; n=58,249 person-years), we used multiple imputation to replace missing values. Overall rates of RPM use were low, with 4.1% (56,553/1,390,995; 95% CI 4.0%-4.1%) of the total sample ever enrolling in RPM. There were statistically significant differences between those who enrolled and those who never enrolled across all veteran characteristics. Those with any RPM enrollment were more likely to be Black compared to individuals with no enrollment (19,457/56,533 veterans, 34.4%, 95% CI 34.0%-34.8% vs 350,391/1,334,462, 26.3%, 95% CI 26.2%-26.3%), and there was a very small but statistically significant difference by ADI percentile (the mean percentile was 0.3 higher for those with no RPM enrollment). Veterans who ever enrolled in RPM were more likely than those who never enrolled to be in VHA priority groups 1 to 3 (service-connected disability; 39,807/56,533, 70.4%, 95% CI 70%-70.8% vs 906,648/1,334,462, 67.9%, 95% CI 67.9%-68%), reside in urban areas (38,555/56,533, 68.2%, 95% CI 67.8%-68.6% vs 864,851/1,334,462 64.8%, 95% CI 64.7%-64.9%), and have 4 or more comorbidities (22,841/56,533, 40.4%, 95% CI 40.4%-40.8% vs 369,619/1,334,462, 27.7%, 95% CI 27.6%-27.8%).

XSL•FO
RenderX

Table 1. Characteristics of veterans with hypertension in the study sample between fiscal years 2020 and 2023 by enrollment in remote patient monitoring (N=1,390,995).

Characteristic	Never enrolled (n=1,334,462)	Ever enrolled (n=56,533)	Total	*P* value
Area Deprivation Index, mean (SD)	58.9 (25.2)	59.2 (25.2)	58.9 (25.2)	.001
Race n (%; 95% CI)				<.001
Hispanic, any race	83,061 (6.2; 6.2-6.3)	3421 (6.1; 5.9-6.3)	86,482 (6.2; 6.2-6.3)	
Non-Hispanic American Indian or Alaska Native	8,697 (0.7; 0.6-0.7)	309 (0.5; 0.5-0.6)	9006 (0.6; 0.6-0.7)	
Non-Hispanic Asian American and Pacific Islander	22,678 (1.7; 1.7-1.7)	747 (1.3; 1.2-1.4)	23,425 (1.7; 1.7-1.7)	
Non-Hispanic Black	350,391 (26.3; 26.2-26.3)	19,457 (34.4; 34.0-34.8)	369,848 (26.6; 26.5-26.7)	
Non-Hispanic White	805,018 (60.3; 60.2-60.4)	30,052 (53.2; 52.7-53.6)	835,070 (60; 60.0-60.1)	
Non-Hispanic other	64,617 (4.8; 4.8-4.9)	2548 (4.5; 4.3-4.7)	67,165 (4.8; 4.8-4.9)	
Fiscal year, n (%; 95% CI)				<.001
2020	636,001 (47.7; 47.6-47.7)	30,390 (53.8; 53.3-54.2)	666,391 (47.9; 47.8-48.0)	
2021	229,647 (17.2; 17.1-17.3)	10,252 (18.1; 17.8-18.5)	239,899 (17.2; 17.2-17.3)	
2022	233,225 (17.5; 17.4-17.5)	8772 (15.5; 15.2-15.8)	241,997 (17.4; 17.3-17.5)	
2023	235,589 (17.7; 17.6-17.7)	7120 (12.6; 12.3-12.9)	242,709 (17.4; 17.4-17.5)	
Sex, n (%; 95% CI)				<.001
Female	99,072 (7.4; 7.4-7.5)	5382 (9.5; 9.3-9.8)	104,454 (7.5; 7.5-7.6)	
Male	1,235,388 (92.6; 92.5-92.6)	51,151 (90.5; 90.2-90.7)	1,286,539 (92.5; 92.4-92.5)	
Age (y), n (%; 95% CI)				<.001
<65	673,661 (50.5; 50.4-50.6)	28,522 (50.5; 50.0-50.9)	702,183 (50.5; 50.4-50.6)	
65-74	499,645 (37.5; 37.4-37.5)	21,557 (38.1; 37.7-38.5)	521,202 (37.0; 37.4-37.6)	
75-84	124,783 (9.0; 9.3-9.4)	5172 (9.1; 8.9-9.4)	129,955 (9.3; 9.3-9.4)	
≥85	36,371 (2.8; 2.7-2.8)	1282 (2.3; 2.1-2.4)	37,653 (2.7; 2.7-2.7)	
Priority group, n (%; 95% CI)				<.001
1-3	906,648 (67.9; 67.9-68.0)	39,807 (70.4; 70.0-70.8)	946,455 (68.0; 68.0-68.1)	
4	15,105 (1.1; 1.1-1.2)	782 (1.4; 1.3-1.5)	15,887 (1.1; 1.1-1.2)	
5	211,435 (15.8; 15.8-15.9)	9138 (16.2; 15.9-16.5)	220,410 (15.8; 15.8-15.9)	
6-8	201,435 (15.1; 15.0-15.2)	6806 (12; 11.8-12.3)	208,241 (15.0; 14.9-15.0)	
Rurality, n (%; 95% CI)				<.001
Urban	864,851 (64.8; 64.7-64.9)	38,555 (68.2; 67.8-68.6)	903,406 (64.9; 64.9-65.0)	
Rural	415,852 (31.2; 31.1-31.2)	15,965 (28.2; 27.9-28.6)	431,817 (31.0; 31.0-31.1)	
Highly rural	53,759 (4.0; 4.0-4.1)	2014 (3.6; 3.4-3.7)	55,773 (4.0; 4.0-4.0)	
Elixhauser comorbidities, n (%; 95% CI)				<.001
1	262,970 (19.7; 19.6-19.8)	7055 (12.5; 12.2-12.8)	270,025 (19.4; 19.3-19.5)	
2	384,950 (28.8; 28.8-28.9)	13,145 (23.3; 22.9-23.6)	398,095 (28.6; 28.5-28.7)	
3	316,923 (23.7; 23.7-23.8)	13,492 (23.9; 23.5-24.2)	330,415 (23.8; 23.7-23.8)	
≥4	369,619 (27.7; 27.6-27.8)	22,841 (40.4; 40.0-40.8)	392,460 (28.2; 28.1-28.3)	
Hypertension type, n (%; 95% CI)				<.001
Uncomplicated	1,224,737 (91.8; 91.7-91.8)	49,383 (87.4; 87.1-87.6)	1,274,120 (91.6; 91.6-91.6)	
Complications[a]	109,725 (8.2; 8.2-8.3)	7150 (12.6; 12.4-12.9)	116,875 (8.4; 8.4-8.4)	

XSL•FO
RenderX

Characteristic	Never enrolled (n=1,334,462)	Ever enrolled (n=56,533)	Total	*P* value
Diabetes type 1 or 2, n (%; 95% CI)				<.001
No	899,914 (67.4; 67.4-67.5)	31,719 (56.1; 55.7-56.5)	931,633 (67.0; 66.9-67.1)	
Yes	434,548 (32.6; 32.5-32.6)	24,814 (43.9; 43.5-44.3)	459,362 (33.0; 32.9-33.1)	

[a]With hypertension-related heart disease or kidney disease.

In Table 2, we share the results of our first regression model. Compared to White veterans, we found that Black veterans (odds ratio [OR] 1.65, 95% CI 1.59-1.70) and veterans of “other” race (OR 1.11, 95% CI 1.03-1.19) had higher odds of RPM enrollment. Asian American and Pacific Islander (AAPI) veterans were slightly less likely (OR 0.83, 95% CI 0.74-0.94) to enroll. We found that, while technically statistically significant, there was no meaningful association between a 10-percentile change in ADI ranking and the odds (OR 1.00, 95% CI 0.99-1.00) of RPM enrollment. We tested different variable specifications of ADI (ADI quartiles and specific cutoff points [ie, 90th percentile]) and found that those in the highest quartile of ADI (most vulnerable) had slightly lower odds of RPM enrollment compared to those in the first quartile (least vulnerable) but that the magnitude (OR 0.93, 95% CI 0.90-0.97) was still quite small (Multimedia Appendix 1).

XSL•FO
RenderX

Table 2. Odds ratios from logistic regression modeling the odds of remote patient monitoring enrollment among veterans with hypertension from fiscal years 2020 to 2023.

Characteristic	Odds ratio (95% CI)
Race	
Hispanic, any race	1.09 (1.02-1.16)
Non-Hispanic American Indian or Alaska Native	0.90 (0.74-1.09)
Non-Hispanic Asian American and Pacific Islander	0.83 (0.74-0.94)
Non-Hispanic Black	1.65 (1.59-1.70)
Non-Hispanic White	__[a]
Non-Hispanic other	1.11 (1.03-1.19)
Area Deprivation Index (per 10-percentile change)	1.00 (0.99-1.00)
Fiscal year	
2020	—
2021	1.22 (1.18-1.26)
2022	1.12 (1.08-1.16)
2023	1.30 (1.26-1.34)
Sex	
Female	—
Male	0.72 (0.69-0.76)
Age (y)	
<65	—
65-74	1.04 (1.01-1.08)
75-84	1.12 (1.06-1.17)
≥85	0.95 (0.87-1.05)
Priority group	
1-3	—
4	1.20 (1.05-1.36)
5	1.01 (0.97-1.05)
6-8	0.83 (0.79-0.87)
Rurality	
Urban	—
Rural	0.91 (0.88-0.94)
Highly rural	0.94 (0.87-1.02)
Elixhauser comorbidities	
1	—
2	1.24 (1.18-1.29)
3	1.63 (1.56-1.71)
≥4	2.47 (2.36-2.58)
Hypertension type	
Uncomplicated	—
With complications[b]	1.57 (1.51-1.64)
Comorbid diabetes	
No	—
Yes	1.47 (1.42-1.51)

XSL•FO
RenderX

[a]Reference group.

[b]With hypertension-related heart disease or kidney disease.

In addition to veteran race, we found that male veterans were less likely (OR 0.72, 95% CI 0.69-0.76) to enroll in RPM and those between the ages of 65 and 84 years were slightly more likely to enroll in RPM than younger veterans. Among clinical covariates, veterans with diabetes (OR 1.47, 95% CI 1.42-1.51), hypertension-associated complications (eg, heart disease or kidney disease; OR 1.57, 95% CI 1.51-1.64), and more chronic health conditions (OR 1.24-2.47) were all significantly associated with higher odds of RPM enrollment.

Sensitivity Analysis

For sensitivity analyses, we further limited our sample to those with grade or stage 2 hypertension (average BP>140/80 mm Hg across ≥2 BP measurements) and who also were prescribed at least one antihypertensive medication. These requirements further limited our sample to 464,971 person-years across 348,265 unique veterans, or 19.3% (464,971 /2,411,782 person-years) of the sample in the original analysis. Among these veterans, compared to the initial sample, a higher proportion (17,762/348,265, 5.1%, 95% CI 5.0%-5.2%) were enrolled in RPM.

In the adjusted analysis for this population (Table 3), Black veterans still had higher odds (OR 1.61, 95% CI 1.50-1.72) of being enrolled in RPM compared to White veterans. There was no longer a statistically significant association between AAPI race and RPM enrollment. However, there was a slightly larger association between “other” race and RPM enrollment (OR 1.19, 95% CI 1.03-1.37). For ADI, the association with RPM enrollment remained very small. While most variables that were significant in the first model remained so in the second model, the magnitude was attenuated; for example, those with 4 or more comorbidities now had only 2.04 (95% CI 1.86-2.24) times the odds of RPM enrollment compared to 2.47 (95% CI 2.36-2.58) times the odds among the larger sample. The magnitude of other covariates also shifted, with some associations now not significant either due to differences in the patient population or due to the smaller sample size.

XSL•FO
RenderX

Table 3. Odds ratios from logistic regression modeling the odds of remote patient monitoring enrollment among veterans with stage 2 hypertension (blood pressure >140/90 mm Hg) and antihypertensive medication use from fiscal years 2020 to 2023.

Characteristic	Odds ratio (95% CI)
Race	
Hispanic, any race	1.27 (1.12-1.45)
Non-Hispanic American Indian or Alaska Native	1.12 (0.77-1.65)
Non-Hispanic Asian American and Pacific Islander	1.02 (0.80-1.29)
Non-Hispanic Black	1.61 (1.50-1.72)
Non-Hispanic White	__[a]
Non-Hispanic other	1.19 (1.03-1.37)
Area Deprivation Index (per 10-percentile change)	0.99 (0.98-1.00)
Fiscal year	
2020	—
2021	1.14 (1.05-1.22)
2022	1.07 (0.99-1.15)
2023	1.26 (1.17-1.36)
Sex	
Female	—
Male	0.68 (0.61-0.75)
Age (y)	
<65	—
65-74	0.98 (0.92-1.05)
75-84	1.07 (0.93-1.02)
≥85	1.03 (0.77-1.37)
Priority group	
1-3	—
4	1.25 (0.93-1.68)
5	0.94 (0.86-1.02)
6-8	0.92 (0.84-1.00)
Rurality	
Urban	—
Rural	0.94 (0.88-1.01)
Highly rural	0.80 (0.67-0.95)
Elixhauser comorbidities	
1	—
2	1.18 (1.08-1.29)
3	1.48 (1.35-1.63)
≥4	2.04 (1.86-2.24)
Hypertension type	
Uncomplicated	—
With complications[b]	1.63 (1.51-1.77)
Comorbid diabetes	
No	—
Yes	1.36 (1.28-1.45)

XSL•FO
RenderX

[a]Reference group.

[b]With hypertension-related heart disease or kidney disease.

Discussion

Principal Results

RPM for hypertension may provide disease management support; however, we found that only 4.1% (56,533/1,390,995) of potentially eligible veterans were ever enrolled during our study period. This finding aligns with the relatively low enrollment in RPM for other chronic diseases within the VHA [38,42]. While having different inclusion criteria, other studies examining RPM in the VHA have found relatively low enrollment [38]. In our retrospective national cross-sectional study of veterans with hypertension from FYs 2020-2023, we found mixed data on racial disparities in RPM enrollment and no association with the socioeconomic vulnerability of the neighborhood where the veterans lived. Black veterans and veterans of other races, including those identifying as multiracial or with unknown race, were more likely to be enrolled in RPM, whereas AAPI veterans were less likely to enroll. We found no meaningful difference in the likelihood of enrollment by neighborhood socioeconomic characteristics.

Comparison With Prior Work

Controlling for potential demographic and clinical confounders, we found that Black veterans were more likely than White veterans to enroll in RPM for hypertension and AAPI veterans were less likely to enroll. Previous research found that overall awareness of hypertension and hypertension control was lower among Asian American patients compared to non-Hispanic White patients, although there was considerable variation by ethnic group [14,43]. Prior research involving the use of RPM among patients without hypertension found that Black patients were more likely to be engaged in the program [44], whereas other research on different populations (nonveterans and those with heart failure) found that female patients and Black patients were less likely to enroll in RPM [32]. In our study, veterans with multiple comorbid conditions and those with diabetes or hypertension-associated complications were more likely to enroll in RPM. These findings indicate that VHA RPM is reaching the appropriate population as those targeted for additional support by VHA RPM are patients who have high health service use (eg, frequent emergency room visits) due to multimorbidity [5,31].

We found no meaningful difference in enrollment in RPM by ADI. This finding held true even when we tested different approaches to modeling, such as ADI quartiles or specified cutoffs. This finding supports previous research among rural and low-income patients that found that socioeconomic status was not a barrier to RPM use [45]. However, it contrasts other research that found that those in more disadvantaged neighborhoods are less likely to have access to and enroll in RPM [44,46-48]. These studies have often cited factors such as differential access to care, low health literacy, limited access to technology, or broadband availability (ie, internet access) as barriers to RPM use [49-51]. Barriers to RPM in the VHA are in part ameliorated because access to RPM is available to all eligible veterans at no cost and the technology used in the VHA does not necessitate broadband availability [5]. This may explain why ADI may not have impacted RPM use in our study population.

Among a smaller subset of patients with documented stage or grade 2 hypertension despite the use of hypertension medications, we still found that Black veterans were more likely than White veterans and AAPI veterans to enroll in RPM, but the magnitude of the association shrank, as it did for most clinical covariates. Perhaps the reason Black veterans were more likely to enroll in RPM was linked to their health and morbidity status as Black veterans are more likely to report circulatory-related chronic conditions than non-Hispanic White veterans [52]. Previous research has found that patients with hypertension and concomitant multimorbidity incur higher health care costs and have more complex health care needs [53-55]. While Black veterans were more likely than White veterans to enroll in the VHA RPM program, future research is needed to encourage uptake of RPM and determine whether RPM is equally effective across all participants. Additionally, given the low (approximately 4%) overall rate of RPM enrollment, RPM is likely one of many tools and approaches required to improve hypertension control in the larger population. Related interventions, such as video BP visits, have successfully reduced BP in veterans, but the intervention did not assess for whom the intervention was most effective [56]. Recent literature on hypertension among veterans shows that Black veterans, followed closely by American Indian and Alaska Native veterans, have the lowest overall rates of hypertension control [17]. Studies of RPM programs that specifically evaluate their effect on racial disparities in hypertension outcomes are promising but limited [57].

Limitations

While our sample included more than 1 million unique veterans over 4 years, our findings may not be generalizable to other populations. Veterans are not representative of the general American population as they are much more likely to be male, White, and older, although younger veteran populations are more diverse. Furthermore, we limited our sample to those with ≥2 BP readings on ≥2 occasions with an average BP of >130/80 mm Hg. This approach ensured that our study cohort had an active case of hypertension according to clinical guidelines but likely nonrandomly eliminated veterans who were healthier (and had fewer medical encounters in which BP was taken) and those who were less reliant on VHA care and less likely to have BP readings documented in VHA data. This approach further limits the generalizability of our study to the veteran population as a whole. We also excluded veterans who died, which likely eliminated veterans with poorer health, which also limits the implications of our study. Our study also only evaluated the VHA's RPM program, whereas other similar programs to monitor and control hypertension may be structured differently. Participants of other non-VHA programs may also have different barriers (such as lack of insurance coverage and changes in health care systems) that are less present in the VHA veteran

XSL•FO
RenderX

population, where many, though not all, veterans receive much of their care in the same system with low or no copayments. Those in lower priority groups (6-8) were the least likely to enroll in RPM, perhaps because they received less care at the VHA overall and were subject to copayments for some services [58,59].

Additionally, while we controlled for a number of clinical and demographic characteristics, there may still be some confounding information that we were unable to capture. Although widely used with established best practices at identifying the most accurate value, Corporate Data Warehouse data on race do not always consist of veteran self-reported race. Primary care provider awareness of RPM as an option for their patients may vary substantially, and there could also be differences between facilities in terms of awareness, uptake, and support for this program for veterans. Future research will examine these other factors. This study also did not evaluate how long veterans stayed in the RPM program, only whether they enrolled. Recent research has examined enrollment duration, and data on reasons for enrollment and disenrollment are stored externally to the VHA electronic health record by a third-party server, which was not accessible for this study [8]. Finally, while there are limitations to the use of ORs for reporting regression results, for rare outcomes such as ours, ORs approximate the risk ratio and eliminate these potential concerns [60].

Conclusions

Controlling for potential confounders, Black veterans were more likely than White veterans to be enrolled in the VHA hypertension RPM program, but we found no difference in RPM enrollment by neighborhood socioeconomic status. These results were consistent when limited to a smaller sample of veterans with hypertension that remained uncontrolled despite the use of antihypertensive medications. Prior research has assessed the duration of enrollment and outcomes after enrolling, but no prior studies have examined the probability of any enrollment in the RPM program for controlling hypertension in a veteran population. In an integrated health care system in which most patients do not pay for health care services, we found that patients from socioeconomically disadvantaged areas and Black veterans were not less likely to be enrolled in RPM. Our study results demonstrate that, within the VHA-enrolled veteran population, RPM is reaching an appropriate patient population (ie, patients with more comorbidities and patients with stage or grade 2 hypertension). However, overall enrollment was low, suggesting that RPM could still benefit additional patients who are not currently enrolled.

For non-VHA settings, reducing or eliminating financial barriers may improve access for similar nonveteran populations. As overall enrollment in RPM for hypertension was low, future research should focus on improving uptake among veterans who could additionally benefit from the program, including White and AAPI veterans. While the VHA is a unique health care system in the United States, being the largest integrated health care system in the country and serving a unique population, outside the VHA, RPM programs may be able to effectively reach minority populations and those living in socioeconomically disadvantaged areas. This study demonstrated that RPM use was not impacted by neighborhood socioeconomic status when the use of those services was either fully or partially paid for and the technology was not reliant on broadband availability. Non-VHA systems, particularly those serving low-income or socioeconomically disadvantaged areas, should explore subsidized or free RPM programs for eligible patients similar to the VHA's no-cost model for veterans.

Acknowledgments

The opinions expressed in this paper are the authors' own and do not necessarily represent the views of the US Department of Veterans Affairs or the US government. No generative artificial intelligence was used at any point in the development of this manuscript.

Funding

HRF was supported by the US Department of Veterans Affairs Office of Academic Affiliations Advanced Fellowship program in Health Systems Research at the Center for Healthcare Organization and Implementation Research, US Department of Veterans Affairs Boston Healthcare System.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to Veterans Health Administration policies but are available from the corresponding author on reasonable request. All requests are subject to compliance with all security requirements of the US Department of Veterans Affairs data sharing agreements.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sensitivity analysis.
[DOCX File , 14 KB - jmir_v28i1e78423_app1.docx]

XSL•FO
RenderX

References

1. Introduction to telehealth and remote patient monitoring. Telehealth.HHS.gov. URL: https://telehealth.hhs.gov/providers/best-practice-guides/telehealth-and-remote-patient-monitoring [accessed 2025-03-31]
2. Farias FA, Dagostini CM, Bicca YD, Falavigna VF, Falavigna A. Remote patient monitoring: a systematic review. Telemed J E Health 2020 May 17;26(5):576-583. [doi: 10.1089/tmj.2019.0066] [Medline: 31314689]
3. Smith W, Colbert BM, Namouz T, Caven D, Ewing JA, Albano AW. Remote patient monitoring is associated with improved outcomes in hypertension: a large, retrospective, cohort analysis. Healthcare (Basel) 2024 Aug 09;12(16):1583 [FREE Full text] [doi: 10.3390/healthcare12161583] [Medline: 39201142]
4. Rasmussen P, Farmer C. The promise and challenges of VA community care: veterans' issues in focus. Rand Health Q 2023 Jun;10(3):9 [FREE Full text] [Medline: 37333666]
5. Remote patient monitoring - home telehealth (RPMHT). U.S. Department of Veterans Affairs. URL: https://www.va.gov/lexington-health-care/programs/remote-patient-monitoring-home-telehealth-rpmht/ [accessed 2025-03-31]
6. At home. U.S. Department of Veterans Affairs. URL: https://telehealth.va.gov/type/home [accessed 2025-06-02]
7. Home telehealth operations manual. VHA Office of Connected Care. 2017 Dec. URL: https://www.vendorportal.ecms.va.gov/FBODocumentServer/DocumentServer.aspx?DocumentId=4869870&FileName=36C25718R0247-030.pdf [accessed 2025-12-21]
8. Gray KE, Rao M, Gunnink E, Eschenroeder L, Geyer JR, Nelson KM, et al. Home telehealth in the Veterans Health Administration: trends and correlates of length of enrollment from 2010 to 2017. J Gen Intern Med 2022 Sep;37(12):3089-3096 [FREE Full text] [doi: 10.1007/s11606-022-07452-1] [Medline: 35230624]
9. Bludorn J, Railey K. Hypertension guidelines and interventions. Prim Care 2024 Mar;51(1):41-52. [doi: 10.1016/j.pop.2023.07.002] [Medline: 38278572]
10. Tsao CW, Aday AW, Almarzooq ZI, Anderson CA, Arora P, Avery CL, et al. Heart disease and stroke statistics-2023 update: a report from the American Heart Association. Circulation 2023 Feb 21;147(8):e93-621. [doi: 10.1161/CIR.0000000000001123] [Medline: 36695182]
11. Chobufo MD, Gayam V, Soluny J, Rahman EU, Enoru S, Foryoung JB, et al. Prevalence and control rates of hypertension in the USA: 2017–2018. Int J Cardiol Hypertens 2020 Jul 31;6:100044. [doi: 10.1016/j.ijchy.2020.100044]
12. Aggarwal R, Chiu N, Wadhera RK, Moran AE, Raber I, Shen C, et al. Racial/ethnic disparities in hypertension prevalence, awareness, treatment, and control in the United States, 2013 to 2018. Hypertension 2021 Dec;78(6):1719-1726 [FREE Full text] [doi: 10.1161/HYPERTENSIONAHA.121.17570] [Medline: 34365809]
13. Yamada M, Wachsmuth J, Sambharia M, Griffin BR, Swee ML, Reisinger HS, et al. The prevalence and treatment of hypertension in Veterans Health Administration, assessing the impact of the updated clinical guidelines. J Hypertens 2023 Jun 01;41(6):995-1002 [FREE Full text] [doi: 10.1097/HJH.0000000000003424] [Medline: 37071434]
14. Abrahamowicz AA, Ebinger J, Whelton SP, Commodore-Mensah Y, Yang E. Racial and ethnic disparities in hypertension: barriers and opportunities to improve blood pressure control. Curr Cardiol Rep 2023 Jan 09;25(1):17-27. [doi: 10.1007/s11886-022-01826-x] [Medline: 36622491]
15. Xu J, Lawrence KG, O'Brien KM, Jackson CL, Sandler DP. Association between neighbourhood deprivation and hypertension in a US-wide Cohort. J Epidemiol Community Health 2022 Mar 17;76(3):268-273. [doi: 10.1136/jech-2021-216445] [Medline: 34789553]
16. Satapathy P, Khatib MN, Gaidhane S, Zahiruddin QS, Gaidhane AM, Rustagi S, et al. Association of neighborhood deprivation and hypertension: a systematic review and meta-analysis. Curr Probl Cardiol 2024 Apr;49(4):102438 [FREE Full text] [doi: 10.1016/j.cpcardiol.2024.102438] [Medline: 38301916]
17. Lind KE, Wong MS, Frochen SE, Yuan AH, Washington DL. Variation in hypertension control by race and ethnicity, and geography in US veterans. J Am Heart Assoc 2025 Jan 21;14(2):e035682 [FREE Full text] [doi: 10.1161/JAHA.123.035682] [Medline: 39791424]
18. Krishnamurthi N, Francis J, Fihn SD, Meyer CS, Whooley MA. Leading causes of cardiovascular hospitalization in 8.45 million US veterans. PLoS One 2018 Mar 22;13(3):e0193996. [doi: 10.1371/journal.pone.0193996] [Medline: 29566396]
19. Makutonin M, Dare J, Heekin M, Salancy A, Hood C, Dominguez LW. Remote patient monitoring for hypertension: feasibility and outcomes of a clinic-based pilot in a minority population. J Prim Care Community Health 2023;14:21501319231204586 [FREE Full text] [doi: 10.1177/21501319231204586] [Medline: 37815085]
20. Galobardes B, Shaw M, Lawlor DA, Lynch JW, Davey Smith G. Indicators of socioeconomic position (part 1). J Epidemiol Community Health 2006 Jan;60(1):7-12 [FREE Full text] [doi: 10.1136/jech.2004.023531] [Medline: 16361448]
21. Krieger N, Williams DR, Moss NE. Measuring social class in US public health research: concepts, methodologies, and guidelines. Annu Rev Public Health 1997 May;18(1):341-378. [doi: 10.1146/annurev.publhealth.18.1.341] [Medline: 9143723]
22. van Ham M, Manley D, Bailey N, Simpson L, Maclennan D. Neighbourhood Effects Research: New Perspectives. Dordrecht, The Netherlands: Springer; 2012.
23. Parati G, Ochoa JE, Postel-Vinay N, Pellegrini D, Torlasco C, Omboni S, et al. Home blood pressure telemonitoring: conventional approach and perspectives from mobile health technology. In: Stergiou G, Parati G, Mancia G, editors. Home Blood Pressure Monitoring. Cham, Switzerland: Springer; 2020:103-119.

XSL•FO
RenderX

24. Park SH, Shin JH, Park J, Choi WS. An updated meta-analysis of remote blood pressure monitoring in urban-dwelling patients with hypertension. Int J Environ Res Public Health 2021 Oct 09;18(20):10583 [FREE Full text] [doi: 10.3390/ijerph182010583] [Medline: 34682329]
25. Najarian M, Goudie A, Bona JP, Rezaeiahari M, Young SG, Bogulski CA, et al. Socioeconomic determinants of remote patient monitoring implementation among rural and urban hospitals. Telemed J E Health 2023 Nov 03;29(11):1624-1633. [doi: 10.1089/tmj.2022.0412] [Medline: 37010391]
26. Rivera V, Aldridge MD, Ornstein K, Moody KA, Chun A. RESEARCHRacial and socioeconomic disparities in access to telehealth. J Am Geriatr Soc 2021 Jan 24;69(1):44-45 [FREE Full text] [doi: 10.1111/jgs.16904] [Medline: 33075143]
27. Darrat I, Tam S, Boulis M, Williams AM. Socioeconomic disparities in patient use of telehealth during the coronavirus disease 2019 surge. JAMA Otolaryngol Head Neck Surg 2021 Mar 01;147(3):287-295 [FREE Full text] [doi: 10.1001/jamaoto.2020.5161] [Medline: 33443539]
28. Rocque GB, Caston NE, Hildreth K, Deng L, Henderson NL, Williams CP, et al. Engagement among diverse patient backgrounds in a remote symptom monitoring program. JCO Oncol Pract 2024 Oct;20(10):1426-1435. [doi: 10.1200/OP.24.00066] [Medline: 38917385]
29. Guzman-Clark J, Wakefield BJ, Farmer MM, Yefimova M, Viernes B, Lee ML, et al. Adherence to the use of home telehealth technologies and emergency room visits in veterans with heart failure. Telemed J E Health 2021 Sep 01;27(9):1003-1010 [FREE Full text] [doi: 10.1089/tmj.2020.0312] [Medline: 33275527]
30. Guzman-Clark JR, van Servellen G, Chang B, Mentes J, Hahn TJ. Predictors and outcomes of early adherence to the use of a home telehealth device by older veterans with heart failure. Telemed J E Health 2013 Mar;19(3):217-223. [doi: 10.1089/tmj.2012.0096] [Medline: 23268695]
31. Guzman-Clark J, Farmer MM, Wakefield BJ, Viernes B, Yefimova M, Lee ML, et al. Why patients stop using their home telehealth technologies over time: predictors of discontinuation in Veterans with heart failure. Nurs Outlook 2021 Mar;69(2):159-166 [FREE Full text] [doi: 10.1016/j.outlook.2020.11.004] [Medline: 33279151]
32. Mastoris I, DeFilippis EM, Martyn T, Morris AA, Van Spall HG, Sauer AJ. Remote patient monitoring for patients with heart failure: sex- and race-based disparities and opportunities. Card Fail Rev 2023 Feb 1;9:e02 [FREE Full text] [doi: 10.15420/cfr.2022.22] [Medline: 36891178]
33. Chen S, Li J. Participation of Black US residents in clinical trials of 24 cardiovascular drugs granted FDA approval, 2006-2020. JAMA Netw Open 2021 Mar 01;4(3):e212640 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.2640] [Medline: 33755163]
34. Ghaferi AA, Schwartz TA, Pawlik TM. STROBE reporting guidelines for observational studies. JAMA Surg 2021 Jun 01;156(6):577-578. [doi: 10.1001/jamasurg.2021.0528] [Medline: 33825815]
35. Beckman KD. How to document and code for hypertensive diseases in ICD-10. Fam Pract Manag 2014;21(2):5-9 [FREE Full text] [Medline: 24693838]
36. Reboussin DM, Allen NB, Griswold ME, Guallar E, Hong Y, Lackland DT, et al. Systematic review for the 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Hypertension 2018 Jun;71(6):e116-e135 [FREE Full text] [doi: 10.1161/HYP.0000000000000067] [Medline: 29133355]
37. Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Dennison Himmelfarb C, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Hypertension 2018 Jun 26;71(6):e13-115 [FREE Full text] [doi: 10.1161/HYP.0000000000000065] [Medline: 29133356]
38. Mohr NM, Vakkalanka JP, Holcombe A, Carter KD, McCoy KD, Clark HM, et al. Effect of chronic disease home telehealth monitoring in the Veterans Health Administration on healthcare utilization and mortality. J Gen Intern Med 2023 Nov 08;38(15):3313-3320. [doi: 10.1007/s11606-023-08220-5] [Medline: 37157039]
39. Ferguson JM, Jacobs J, Yefimova M, Greene L, Heyworth L, Zulman DM. Virtual care expansion in the Veterans Health Administration during the COVID-19 pandemic: clinical services and patient characteristics associated with utilization. J Am Med Inform Assoc 2021 Mar 01;28(3):453-462 [FREE Full text] [doi: 10.1093/jamia/ocaa284] [Medline: 33125032]
40. Peltzman T, Rice K, Jones KT, Washington DL, Shiner B. Optimizing data on race and ethnicity for Veterans Affairs patients. Mil Med 2022 Jul 01;187(7-8):e955-e962. [doi: 10.1093/milmed/usac066] [Medline: 35323934]
41. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care 1998 Jan;36(1):8-27. [doi: 10.1097/00005650-199801000-00004] [Medline: 9431328]
42. Reddy A, Gunnink EJ, Rojas J, Nelson K, Wong ES. Impact of home telehealth expansion on high-cost utilization among Veterans Health Administration patients with diabetes. J Gen Intern Med 2025 May 31;40(6):1340-1349. [doi: 10.1007/s11606-024-09169-9] [Medline: 39482477]
43. Jung MY, Lee S, Thomas SB, Juon HS. Hypertension prevalence, treatment, and related behaviors among Asian Americans: an examination by method of measurement and disaggregated subgroups. J Racial Ethn Health Disparities 2019 Jun 7;6(3):584-593 [FREE Full text] [doi: 10.1007/s40615-018-00557-6] [Medline: 30618006]

XSL•FO
RenderX

44. Kirkland E, Schumann SO, Schreiner A, Heincelman M, Zhang J, Marsden J, et al. Patient demographics and clinic type are associated with patient engagement within a remote monitoring program. Telemed J E Health 2021 Aug 01;27(8):843-850 [FREE Full text] [doi: 10.1089/tmj.2020.0535] [Medline: 34115942]
45. Clark D3, Woods J, Zhang Y, Chandra S, Summers RL, Jones DW. Home blood pressure telemonitoring with remote hypertension management in a rural and low-income population. Hypertension 2021 Dec;78(6):1927-1929 [FREE Full text] [doi: 10.1161/HYPERTENSIONAHA.121.18153] [Medline: 34757773]
46. Fritz BA, Ramsey B, Taylor D, Shoup JP, Schmidt JM, Guinn M, et al. Association of race and neighborhood disadvantage with patient engagement in a home-based COVID-19 remote monitoring program. J Gen Intern Med 2022 Mar;37(4):838-846 [FREE Full text] [doi: 10.1007/s11606-021-07207-4] [Medline: 34993862]
47. Akar JG, Bao H, Jones P, Wang Y, Chaudhry SI, Varosy P, et al. Use of remote monitoring of newly implanted cardioverter-defibrillators: insights from the patient related determinants of ICD remote monitoring (PREDICT RM) study. Circulation 2013 Nov 26;128(22):2372-2383. [doi: 10.1161/CIRCULATIONAHA.113.002481] [Medline: 24043302]
48. Varma N, Piccini JP, Snell J, Fischer A, Dalal N, Mittal S. The relationship between level of adherence to automatic wireless remote monitoring and survival in pacemaker and defibrillator patients. J Am Coll Cardiol 2015 Jun 23;65(24):2601-2610 [FREE Full text] [doi: 10.1016/j.jacc.2015.04.033] [Medline: 25983008]
49. Foster M, Etchin A, Pope C, Hartmann CW, Emidio O, Bosworth HB. The impact of COVID-19 on hypertension and hypertension medication adherence among underrepresented racial and ethnic groups: a scoping review. Curr Hypertens Rep 2023 Nov 25;25(11):385-394. [doi: 10.1007/s11906-023-01262-4] [Medline: 37624472]
50. Patel P, Dhindsa D, Eapen DJ, Khera A, Gulati M, Stone NJ, et al. Optimizing the potential for telehealth in cardiovascular care (in the era of COVID-19): time will tell. Am J Med 2021 Aug;134(8):945-951. [doi: 10.1016/j.amjmed.2021.03.007] [Medline: 33845033]
51. Bress AP, Cohen JB, Anstey DE, Conroy MB, Ferdinand KC, Fontil V, et al. Inequities in hypertension control in the United States exposed and exacerbated by COVID-19 and the role of home blood pressure and virtual health care during and after the COVID-19 pandemic. J Am Heart Assoc 2021 Jun;10(11):e020997 [FREE Full text] [doi: 10.1161/JAHA.121.020997] [Medline: 34006116]
52. Ward RE, Nguyen XM, Li Y, Lord E, Lecky V, Song RJ, et al. Racial and ethnic disparities in U.S. veteran health characteristics. Int J Environ Res Public Health 2021 Mar 02;18(5):2411. [doi: 10.3390/ijerph18052411] [Medline: 33801200]
53. Wong MC, Wang HH, Cheung CS, Tong EL, Sek AC, Cheung NT, et al. Factors associated with multimorbidity and its link with poor blood pressure control among 223,286 hypertensive patients. Int J Cardiol 2014 Nov 15;177(1):202-208. [doi: 10.1016/j.ijcard.2014.09.021] [Medline: 25499379]
54. Nguyen TN, Kalia S, Hanlon P, Jani BD, Nicholl BI, Christie CD, et al. Multimorbidity and blood pressure control in patients attending primary care in Canada. J Prim Care Community Health 2023 Dec 14;14:21501319231215025 [FREE Full text] [doi: 10.1177/21501319231215025] [Medline: 38097504]
55. Alshahawey M, Jafari E, Smith SM, McDonough CW. Characterizing apparent treatment resistant hypertension in the United States: insights from the All of Us Research Program. J Am Med Inform Assoc 2024 Dec 01;31(12):2899-2907. [doi: 10.1093/jamia/ocae227] [Medline: 39181122]
56. Carrico M, Frosch C, Craig K, Carter M, Falk J, Guerrero S, et al. Implementation of video blood pressure visits in the Veterans Health Administration. Telemed J E Health 2024 Apr 01;30(4):1006-1012. [doi: 10.1089/tmj.2023.0269] [Medline: 37935031]
57. Khoong EC, Commodore-Mensah Y, Lyles CR, Fontil V. Use of self-measured blood pressure monitoring to improve hypertension equity. Curr Hypertens Rep 2022 Nov 24;24(11):599-613 [FREE Full text] [doi: 10.1007/s11906-022-01218-0] [Medline: 36001268]
58. VA priority groups. U.S. Department of Veterans Affairs. URL: https://www.va.gov/health-care/eligibility/priority-groups/ [accessed 2025-04-10]
59. Hynes DM, Koelling K, Stroupe K, Arnold N, Mallin K, Sohn MW, et al. Veterans' access to and use of Medicare and Veterans Affairs health care. Med Care 2007 Mar;45(3):214-223. [doi: 10.1097/01.mlr.0000244657.90074.b7] [Medline: 17304078]
60. Cummings P. The relative merits of risk ratios and odds ratios. Arch Pediatr Adolesc Med 2009 May 04;163(5):438-445. [doi: 10.1001/archpediatrics.2009.31] [Medline: 19414690]

Abbreviations

AAPI: Asian American and Pacific Islander
ADI: Area Deprivation Index
BP: blood pressure
FY: fiscal year
OR: odds ratio
RPM: remote patient monitoring
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

XSL•FO
RenderX

VA: US Department of Veterans Affairs
VHA: Veterans Health Administration

Edited by S Brini; submitted 02.Jun.2025; peer-reviewed by C Exeter, L Tieu, J Edwards; comments to author 05.Sep.2025; accepted 19.Nov.2025; published 09.Mar.2026.

Please cite as:
Friedman HR, Mull HJ, Guzman-Clark JR, Sturgeon DJ, Foster MV
The Effect of Veteran Race and Socioeconomic Status on Enrollment in Remote Patient Monitoring for Hypertension: Retrospective Observational Cross-Sectional Study
J Med Internet Res 2026;28:e78423
URL: https://www.jmir.org/2026/1/e78423
doi:10.2196/78423
PMID:

©Hannah Robin Friedman, Hillary J Mull, Jenice Ria Guzman-Clark, Daniel J Sturgeon, Marva V Foster. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impact of Tele-Intensive Care Units on the Clinical Outcomes of Critically Ill Patients With COVID-19: Retrospective Cohort Study

Thais Dias Midega[1], MD; Fabio Barlem Hohmann[1], PhD; Renato Carneiro de Freitas Chaves[1,2], PhD; Guacyra Margarita Batista de Almeida[3], MD; Vivian Jaqueline Lima Leoneza[4], MD; Jennifer Ferreira Figueiredo Cabral[5]; Bianca Veloso Vitalino[6], MD; Emanuelle de Araújo Camboim[7], MBA; Nelma de Jesus Nogueira Machado[8]; Ricardo Fernando Batista de Melo[9], MD; Jorge Patrick Oliveira Feliciano[10]; Breno Mendes Cardoso[11], MD; Leonardo José Rolim Ferraz[1], MD; Thiago Domingos Corrêa[1], PhD; Maura Cristina dos Santos[12]; Renata Albaladejo Morbeck[12]; Adriano José Pereira[1,12], PhD

[1]Department of Critical Care Medicine, Hospital Israelita Albert Einstein, São Paulo, Sao Paulo, Brazil
[2]Massachusetts Institute of Technology, Cambridge, MA, United States
[3]Department of Critical Care Medicine, Hospital de Emergência Dr. Daniel Houly, Arapiraca, Brazil
[4]Department of Critical Care Medicine, Santa Casa de Misericórdia de Sorocaba, Sorocaba, Brazil
[5]Department of Critical Care Medicine, Hospital Regional do Cariri, Juazeiro do Norte, Brazil
[6]Department of Critical Care Medicine, Santa Casa de Misericórdia de São João del-Rei, São João del-Rei, Brazil
[7]Department of Critical Care Medicine, Hospital Metropolitano Sul Dom Helder Câmara, Cabo de Santo Agostinho, Brazil
[8]Department of Critical Care Medicine, Santa Casa de Misericórdia do Pará, Belém, Brazil
[9]Department of Critical Care Medicine, Hospital Santa Lúcia, Poços de Caldas, Brazil
[10]Department of Critical Care Medicine, Dom Orione Hospital, Araguaína, Brazil
[11]Department of Critical Care Medicine, Hospital Municipal Senhora Santana, Brasília de Minas, Brazil
[12]Department of Telemedicine, Hospital Israelita Albert Einstein., São Paulo, Brazil

Corresponding Author:
Thais Dias Midega, MD
Department of Critical Care Medicine
Hospital Israelita Albert Einstein
Rua Inhambu 1125
São Paulo, Sao Paulo
Brazil
Phone: 55 11997076207
Email: thais.dmidega@einstein.br

Abstract

Background: The COVID-19 pandemic imposed an unprecedented demand for intensive care unit (ICU) resources in Brazil, where shortages of trained intensivists prompted the implementation of telemedicine-based critical care support strategies.

Objective: This study aimed to evaluate the association between adherence to the Tele-ICU COVID-19 Brazil Program and clinical outcomes of ICU patients with COVID-19.

Methods: We conducted a retrospective cohort study including all ICUs participating in the Tele-ICU COVID-19 Brazil Program between April and December 2020. Program adherence was assessed at 2 levels: patient coverage, defined as the number of daily multidisciplinary rounds per patient divided by the patient's total ICU length of stay (LOS), and ICU coverage, defined as the number of daily multidisciplinary round days in the ICU divided by the total number of patient-days in that ICU. We compared outcomes between groups categorized by an empirically defined 50% cutoff: low patient coverage (<50%) versus high patient coverage (≥50%) and low ICU coverage (<50%) versus high ICU coverage (≥50%). Multilevel mixed-effects models accounting for ICU-level clustering were used to assess outcomes: logistic regression for ICU mortality (adjusted odds ratios) and linear mixed-effects regression with log-transformed ICU LOS (exponentiated coefficients, exp[β]).

Results: A total of 1680 patients were included. Compared with the low patient coverage group (<50%), patients in the high patient coverage (≥50%) had lower Sequential Organ Failure Assessment scores (median 2, IQR 0-5 vs median 3, IQR 0-6; *P*=.007); shorter ICU LOS (median 6, IQR 3-11 days vs median 11, IQR 6-20 days; *P*<.001); and shorter hospital LOS (median

XSL•FO
RenderX

9, IQR 5-16 days vs median 14, IQR 8-26 days; *P*<.001). In unadjusted analyses, ICU mortality did not differ significantly between the low and high patient coverage groups (50.1% vs 46.3%; *P*=.16). In multilevel analysis, mechanical ventilation and vasopressor use were independently associated with higher ICU mortality. Higher patient coverage was independently associated with lower ICU mortality (adjusted odds ratio 0.52, 95% CI 0.27-0.99; *P*=.048). In the log-transformed mixed-effects model for ICU LOS, a higher Sequential Organ Failure Assessment score (exp[β] 1.037, 95% CI 1.02-1.05; *P*<.001) and use of mechanical ventilation (exp[β] 1.23, 95% CI 1.05-1.43; *P*=.01) were associated with longer ICU LOS, whereas higher patient coverage was independently associated with shorter ICU LOS (exp[β] 0.17, 95% CI 0.13-0.21; *P*<.001). ICU coverage was not independently associated with ICU mortality or ICU LOS.

Conclusions: Greater patient-level coverage by remote intensivist–led multidisciplinary rounds within the Tele-ICU program was independently associated with lower ICU mortality and shorter ICU LOS. These findings support the potential contribution of tele–critical care strategies to expanding specialist support during public health emergencies.

(J Med Internet Res 2026;28:e64996) doi:10.2196/64996

KEYWORDS

coronavirus; COVID-19; SARS-CoV-2; intensive care unit; telemedicine; tele–intensive care unit; Tele-ICU; clinical outcomes; mortality

Introduction

Background

COVID-19 has become a major public health care concern, with almost half a billion cases diagnosed and more than 6 million deaths reported across the globe [1]. During the pandemic, waves of increased number of newly diagnosed patients with SARS-CoV-2 infection were reported worldwide, with varied degrees of disease severity [2]. The rapid spread of SARS-CoV-2 increased the demand for intensive care units (ICUs) and the pressure on health care systems [2].

In Brazil, the high demand for ICU beds coupled with the shortage of intensivists to care for patients who are critically ill—particularly in rural and remote regions far from major urban centers—motivated the search for alternative strategies to deliver efficient care [3]. In this scenario, in a partnership between the Brazilian Ministry of Health (via the Institutional Development Support Program of the Unified Health System) and some of the top-ranked Brazilian hospitals, the Tele-ICU COVID-19 Brazil Program was created to provide remote intensivists to guide daily multidisciplinary rounds (DMRs) in public COVID-19 ICUs from Brazil with patients with COVID-19. The program consisted of DMRs (structured patient-centered discussions) with the remote intensivist and the local ICU multidisciplinary team, aiming to review diagnostic hypotheses, list active problems, and establish joint therapeutic goals.

Different Tele-ICU models exist and vary in structure and intensity. The most common is the centralized “hub-and-spoke” model, where a central team remotely monitors several ICUs [4]. Alternatively, decentralized models allow remote specialists to provide support from multiple locations on demand. These configurations may include continuous monitoring, scheduled virtual rounds, or reactive consultations, depending on institutional needs and available infrastructure [4].

Observational studies indicated that implementing Tele-ICU programs may be an effective strategy for improving clinical outcomes [5-11]. However, reported impacts vary across studies, likely due to differences in Tele-ICU models and applications, including second opinion consultations, real-time evaluations, remotely conducted rounds, and others [12-14]. More recently, a randomized clinical trial evaluating an intensivist-led, telemedicine-based strategy that incorporated DMRs found no significant improvement in ICU length of stay (LOS) or patient mortality [15].

However, evidence regarding the impact of Tele-ICU interventions on clinical outcomes, particularly during large-scale health care crises such as the COVID-19 pandemic, remains limited [7,8]. In this context, this study aimed to evaluate the association of the Tele-ICU COVID-19 Brazil Program and clinical outcomes of patients with COVID-19. We hypothesized that higher adherence to the program’s structured DMRs would be associated with lower ICU mortality and shorter ICU LOS among patients admitted to public ICUs across Brazil during the COVID-19 pandemic.

Objectives

This study aimed to evaluate the association of the adherence to the Tele-ICU COVID-19 Brazil Program and clinical outcomes of patients with COVID-19.

Methods

Study Design

We performed a retrospective study to evaluate the impact of Tele-ICU COVID-19 Brazil Program on clinical outcomes of ICU patients with COVID-19.

Ethical Considerations

The study was approved by Brazilian National Ethics Committee on August 27, 2020, with waiver of informed consent (CAAE: 31459120.0.0000.0071, opinion number: 4.240.138). All data used in the study were deidentified before analysis to protect participant privacy and ensure confidentiality. No compensation or reimbursement was provided to participants.

Tele-ICU Program

The Tele-ICU COVID-19 Brazil Program was led by the Tele-ICU Department of the Hospital Israelita Albert Einstein, a private quaternary care hospital located in São Paulo, Brazil.

XSL•FO
RenderX

Participating ICUs that accepted the invitation (a convenience sample) were located in public hospitals across different geographic regions of the country (1 ICU per hospital) and were designated by the Brazilian Ministry of Health between April and December 2020. The recruitment process occurred at the institutional level and was coordinated directly by the Ministry of Health, which was responsible for identifying, evaluating, and selecting eligible ICUs. ICUs were eligible if they belonged to public or philanthropic hospitals and did not have a board-certified intensivist providing daily on-site care. The Ministry also established criteria for replacing units during the program: ICUs could be removed if they missed daily tele-rounds for 5 consecutive days or for more than 10 nonconsecutive days within a month.

As critical care specialists were not available daily in the participating ICUs during this period, DMRs were conducted entirely by the program's remote intensivists. Discussions were conducted by a board-certified intensivist trained in telemedicine, based at the remote center (tele-intensivist, in São Paulo, Brazil), in collaboration with the local multidisciplinary team (physician, nurse, and physiotherapist, in any part of the country). DMRs took place from Monday to Friday at a predetermined time (mostly during the mornings), from April to December 2020, and approached all patients admitted to the participating ICUs. The main objective of the DMR conducted by the tele-intensivist was to review diagnostic hypotheses, list active problems, and jointly create a treatment plan until the next DMR. Tele-intensivists made recommendations based on updated scientific evidence, suitable to the local context. Clinical protocols in texts and figures were made accessible through a dedicated application developed for the physicians and multidisciplinary team of the participating ICUs.

Study Participants

All the 17 participant ICUs of the program were invited to participate in this study. All patients admitted to these ICUs during the Tele-ICU program, from April to December 2020, were included in this analysis.

We excluded patients who meet the following criteria: (1) aged <18 years, (2) receiving exclusive palliative care at ICU admission, and (3) incomplete data about ICU outcomes.

Data Collection and Study Variables

All study data were collected by the Tele-ICU intensivist during the DMRs performed during the program and complemented by the local staff of the participant ICUs. Data were stored on a server at Hospital Israelita Albert Einstein, São Paulo, Brazil, and retrieved by the study authors after the approval of the ethics committee.

Collected data included demographics, comorbidities, Sequential Organ Failure Assessment (SOFA) score [16], number of DMRs, resource use, and organ support (vasopressors, noninvasive ventilation, and mechanical ventilation) during ICU stay, palliative care, ICU and hospital LOS, and ICU and in-hospital mortality. Variables were collected during the DMRs (and lately complemented by invited professionals of the participant ICU). SOFA score refers to the maximum value of SOFA during ICU stay.

Definitions

To evaluate the impact of the Tele-ICU program, we developed 2 variables: patient coverage and ICU coverage. *Patient coverage* quantifies how often each patient's case was discussed during DMRs based on the hypothesis that more frequent discussions could enhance clinical decision-making and continuity of care. *ICU coverage* reflects the overall participation of each ICU in the program, hypothesizing that higher engagement would promote local learning and greater adherence to evidence-based practices through regular interaction with remote intensivists. These variables were defined as follows:

- Patient coverage=(number of DMRs per patient)/(patient's ICU LOS)
- ICU coverage=(number of DMR days in the ICU)/(total ICU patient-days)

A cutoff of 50% was empirically defined to categorize groups. Patients were classified into *low patient coverage* (<50%) and *high patient coverage* (≥50%) groups based on their individual patient coverage. Similarly, ICUs were categorized into low ICU coverage (<50%) and high ICU coverage (≥50%) groups based on their overall ICU coverage.

Outcomes

Our primary outcome of interest was ICU mortality. The secondary outcomes were ICU LOS and use of resources (eg, mechanical ventilation, noninvasive ventilation, use of vasopressors, and renal replacement therapy).

Statistical Analysis

Categorical variables are presented as absolute and relative frequencies. Continuous variables are presented as median with IQR. Normality was assessed using the Shapiro-Wilk test.

Comparisons were made between groups: low patient coverage group vs high patient coverage group and low ICU coverage group vs high ICU coverage group. Continuous variables were compared using the independent 2-tailed *t* test or the Mann-Whitney *U* test in case of nonnormal distribution. Univariate analyses were performed to identify which predictors were associated with ICU mortality and ICU LOS.

We performed multivariate analyses using multilevel mixed modeling, with hospitals as random effect. All variables tested with a *P* value <.20 in the univariate analyses were included as fixed effects. We performed a multicollinearity analysis before the backward elimination procedure. A mixed model was undertaken to obtain adjusted odds ratio (OR) along with 95% CI and to define which variables were independently associated with ICU mortality.

A linear mixed-effects model with log-transformed ICU LOS as the outcome was used, with hospital as a random effect. Exp(β) along with the 95% CI were obtained to define which variables were independently associated with ICU LOS. To perform the analysis of the predictor ICU LOS, probability distribution analyses of the outcome were conducted. We carried out inferential tests to assess the fit of various probability distributions. The only distribution that demonstrated adherence

XSL•FO
RenderX

was the normal distribution for data transformed by natural logarithm. Thus, the outcome variable ICU LOS was transformed by taking the natural logarithm (ln).

All statistical tests were 2-tailed, and $P<.05$ was considered statistically significant. Analyses were performed using R software (version 4.1.0; R Foundation for Statistical Computing).

Results

From April to December 2020, a total of 1945 patients from 17 different ICUs across different Brazilian geographic regions were included in the database and had their plan of care defined by the remote tele-intensivist on at least 1 day of the Tele-ICU program. We excluded 233 (11.9%) patients who lacked data on ICU outcomes (leading to the exclusion of an entire ICU, as none of the patients from this ICU had available outcome data); 18 (0.9%) patients aged younger than 18 years; and 14 (0.7%) patients receiving palliative care since the ICU admission, as previously defined in our study protocol. Finally, 1680 patients from 16 different ICUs were included in the study.

Baseline characteristics of patients included in the study are presented in Table 1. The median age in this cohort was 66 (IQR 55-76) years, and 56.4% (865/1534) of the patients were men. The median SOFA score was 4 (IQR 1-9), and the median number of DMRs per patient was 3 (IQR 2-6). In this cohort, during ICU stay, 56.8% (769/1354) of the patients required mechanical ventilation, 46.3% (627/1354) of the patients required vasopressors, 16.9% (229/1354) of the patients required renal replacement therapy, and 9.7% (132/1361) of the patients required noninvasive ventilation. ICU mortality was 48.3% (812/1680), and the hospital mortality was 51.1% (847/1658). The median ICU and hospital LOS were, respectively, 8 (IQR 4-15) days and 11 (IQR 6-20) days (Table 1).

XSL•FO
RenderX

Table 1. Baseline characteristics and clinical outcomes of intensive care unit (ICU) patients with COVID-19 admitted to public ICUs participating in the Tele-ICU COVID-19 Brazil Program (retrospective multicenter cohort study, Brazil, April to December 2020; N=1680)a.

Characteristics	Patients
Male, n (%)	865 (56.4)
Age (years), median (IQR)	66 (55-76)
Comorbidities, n (%)	
Systemic hypertension	684 (62.8)
Diabetes mellitus	441 (40.5)
Congestive heart failure	74 (6.8)
Acute cerebral stroke	60 (5.5)
Coronary arterial disease	17 (1.6)
Previous myocardial infarction	21 (1.9)
Asthma	27 (2.5)
COPD[b]	86 (7.9)
Chronic kidney disease	44 (4)
Chronic kidney disease requiring RRT[c]	19 (1.7)
Locoregional cancer	4 (0.4)
Metastatic cancer	5 (0.5)
Hematologic cancer	1 (0.1)
Cognitive impairment and dementia	19 (1.7)
Liver cirrhosis	1 (0.1)
ICU mortality, n (%)	812 (48.3)
Hospital mortality, n (%)	847 (51.1)
ICU LOS[d], median (IQR)	8 (4-15)
Hospital LOS, median (IQR)	11 (6-20)
SOFA[e] score, median (IQR)	4 (1-9)
Number of DMRs[f] per patient, median (IQR)	3 (2-6)
Support on the first day of the Tele-ICU visit, n (%)	
Mechanical ventilation	639 (47.2)
Noninvasive ventilation	132 (9.7)
Vasopressors	425 (31.4)
RRT	128 (9.5)
Support in the first 24 hours of ICU stay, n (%)	
Mechanical ventilation	243 (36.9)
Noninvasive ventilation	71 (10.8)
Vasopressors	164 (24.9)
RRT	26 (3.9)
Support at any time during ICU stay, n (%)	
Mechanical ventilation	769 (56.8)
Noninvasive ventilation	168 (12.4)
Vasopressors	627 (46.3)
RRT	229 (16.9)

[a]Percentages were calculated based on available data for each variable.

XSL•FO
RenderX

[b]COPD: chronic obstructive pulmonary disease.

[c]RRT: renal replacement therapy.

[d]LOS: length of stay.

[e]SOFA score: Sequential Organ Failure Assessment score, ranges from 0 to 24, with higher scores indicating more severe organ dysfunction.

[f]DMR: daily multidisciplinary round.

Compared with the low patient coverage group, patients in the high patient coverage group had shorter ICU LOS (median 6, IQR 3-11 days vs median 11, IQR 6-20 days; *P*<.001) and hospital LOS (median 9, IQR 5-16 days vs median 14, IQR 8-26 days; *P*<.001; Table 2). SOFA score was higher in the low patient coverage group (median 3, IQR 0-6 vs median 2, IQR 0-5; *P*=.007). ICU mortality (50.1% vs 46.3%; *P*=.16) did not differ significantly between the low and high patient coverage groups. During ICU stay, patients in the high patient coverage group used mechanical ventilation less frequently, but used noninvasive ventilation more frequently when compared with the low patient coverage group.

Table 2. Comparison of clinical outcomes and resource use according to patient-level adherence to tele–intensive care unit (ICU) daily multidisciplinary rounds among ICU patients with COVID-19 (Brazil, April to December 2020).

	Low patient coverage group (n=673)	High patient coverage group (n=862)	*P* value
ICU deaths, n (%)	337 (50.1)	399 (46.3)	.16[a]
Hospital deaths, n (%)	351 (52.2)	417 (48.4)	.17[a]
ICU LOS[b], median (IQR)	11 (6-20)	6 (3-11)	<.001[c]
Hospital LOS, median (IQR)	14 (8-26)	9 (5-16)	<.001[c]
SOFA[d] score, median (IQR)	3 (0-6)	2 (0-5)	.007[c]
Support in the first 24 hours of ICU stay, n (%)			
Mechanical ventilation	86 (45.7)	154 (33)	.003[a]
Noninvasive ventilation	13 (6.9)	58 (12.4)	.06[a]
Vasopressors	56 (29.8)	106 (22.7)	.07[a]
RRT[e]	11 (5.9)	14 (3)	.14[a]
Support at any time during ICU stay, n (%)			
Mechanical ventilation	328 (60.3)	357 (51.6)	.003[a]
Noninvasive ventilation	58 (10.7)	108 (15.6)	.01[a]
Vasopressors	261 (48)	301 (43.5)	.13[a]
RRT	105 (19.3)	100 (14.5)	.03[a]

[a]*P* values were calculated using the chi-square test.

[b]LOS: length of stay.

[c]*P* values were calculated using the Mann-Whitney *U* test.

[d]SOFA score: Sequential Organ Failure Assessment score, ranges from 0 to 24, with higher scores indicating more severe organ dysfunction.

[e]RRT: renal replacement therapy.

Compared with the low ICU coverage group, patients in the high ICU coverage group had lower SOFA scores (median 2, IQR 0-4 vs median 3, IQR 1-7; *P*<.001), lower ICU mortality (44.3% vs 52.5%; *P*=.004), lower ICU LOS (median 7, IQR 4-12 days vs median 9, IQR 5-17 days; *P*<.001), and lower hospital LOS (median 8, IQR 5-16 days vs median 14, IQR 7-23 days; *P*<.001). During ICU stay, patients in the high ICU coverage group used mechanical ventilation, noninvasive ventilation, vasopressors, and renal replacement therapy less frequently (Table 3).

XSL•FO
RenderX

Table 3. Comparison of clinical outcomes and resource use according to intensive care unit (ICU)–level adherence to the Tele-ICU COVID-19 Brazil Program (Brazil, April to December 2020).

	Low ICU coverage group (n=844)	High ICU coverage group (n=836)	*P* value
ICU deaths, n (%)	442 (52.5)	370 (44.3)	.004[a]
Hospital deaths, n (%)	451 (53.5)	396 (48.7)	.05[a]
ICU LOS[b], median (IQR)	9 (5-17)	7 (4-12)	<.001[c]
Hospital LOS, median (IQR)	14 (7-23)	8 (5-16)	<.001[c]
SOFA[d] score, median (IQR)	3 (1-7)	2 (0-4)	<.001[c]
Palliative care, n (%)	146 (17.3)	123 (14.7)	.17[a]
Support in the first 24 hours of ICU stay, n (%)			
Mechanical ventilation	142 (44)	101 (30.1)	<.001[a]
Noninvasive ventilation	32 (9.9)	39 (11.6)	.56[a]
Vasopressors	93 (28.8)	71 (21.1)	.03[a]
RRT[e]	24 (7.4)	2 (0.6)	<.001[a]
Support at any time during ICU stay, n (%)			
Mechanical ventilation	471 (63.1)	298 (49)	<.001[a]
Noninvasive ventilation	109 (14.6)	59 (9.7)	.008[a]
Vasopressors	384 (51.5)	243 (40)	<.001[a]
RRT	168 (22.5)	61 (10)	<.001[a]

[a]*P* values were calculated using the chi-square test.

[b]LOS: length of stay.

[c]*P* values were calculated using the Mann-Whitney *U* test.

[d]SOFA score: Sequential Organ Failure Assessment score; ranges from 0 to 24, with higher scores indicating more severe organ dysfunction.

[e]RRT: renal replacement therapy.

Independent predictors of ICU mortality were the use of mechanical ventilation (OR 3.22, 95% CI 2.056-5.043; *P*<.001) and vasopressors (OR 1.87, 95% CI 1.236-2.824; *P*=.003) during ICU stay. The use of noninvasive ventilation (OR 0.54, 95% CI 0.331-0.881; *P*=.01) and patient coverage (OR 0.520, 95% CI 0.272-0.993; *P*=.048) were identified as a protective factors (Table 4). Neither ICU coverage nor total time of ICU participation in the project remained in the final model.

Table 4. Multilevel mixed-effects logistic regression analysis of factors associated with intensive care unit (ICU) mortality among critically ill patients with COVID-19 enrolled in the Tele-ICU COVID-19 Brazil Program (Brazil, April to December 2020).

Fixed effects	OR[a] (95% CI)	*P* value
SOFA[b] score	1.238 (0.860-1.317)	.052
Use of MV[c] during ICU stay	3.222 (2.056-5.043)	<.001
Use of NIV[d] during ICU stay	0.540 (0.331-0.881)	.01
Use of vasopressor during ICU stay	1.870 (1.236-2.824)	.003
Patient coverage	0.520 (0.272-0.993)	.048

[a]OR: odds ratio.

[b]SOFA: Sequential Organ Failure Assessment.

[c]MV: mechanical ventilation.

[d]NIV: noninvasive ventilation.

Independent predictors of higher ICU LOS included in the final adjusted model were SOFA (exp[β] 1.037, 95% CI 1.02-1.05; *P*<.001), use of mechanical ventilation during ICU stay (exp[β] 1.228, 95% CI 1.05-1.43; *P*=.01), and patient coverage as protective factor (exp[β] 0.166, 95% CI 0.13-0.21; *P*<.001; Table 5). As in the first model for ICU mortality, neither ICU

XSL•FO
RenderX

coverage nor total time of ICU participation in the project remained in this final adjusted model.

Table 5. Multilevel linear mixed-effects model evaluating factors associated with intensive care unit (ICU) length of stay among critically ill patients with COVID-19 participating in the Tele-ICU COVID-19 Brazil Program (Brazil, April to December 2020).

Fixed effects	Exp(β)[a] (95% CI)	*P* value
SOFA[b] score	1.037 (1.02-1.05)	<.001
Use of MV[c] during ICU stay	1.228 (1.05-1.43)	.01
Patient coverage	0.166 (0.13-0.21)	<.001

[a]Exp(β): exponentiated regression coefficient from the linear mixed-effects model with log-transformed ICU LOS; values >1 indicate longer ICU LOS, and values <1 indicate shorter ICU LOS.

[b]SOFA score: Sequential Organ Failure Assessment score; ranges from 0 to 24, with higher scores indicating more severe organ dysfunction.

[c]MV: mechanical ventilation.

Discussion

Principal Findings

The main finding of this study was that patients with higher patient coverage by Tele-ICU, defined as a greater proportion of ICU days during which patients were discussed in DMRs led by a remote intensivist during the Tele-ICU COVID-19 Brazil Program, were associated with lower ICU mortality and LOS. We postulate that Tele-ICU may have improved clinical outcomes by implementing multidisciplinary rounds, thereby enhancing adherence to best practices and promoting continuity of care in participating ICUs.

Our results are in line with some previous studies regarding the association between the use of Tele-ICU and lower mortality and LOS [5-10]. For instance, a prospective study found that the implementation of a Tele-ICU intervention, where the Tele-ICU team participated in critical care processes throughout the day, was associated with reduced adjusted odds of mortality and reduced hospital LOS, as well as with changes in best practice adherence and lower rates of preventable complications [17]. Similarly, a before-and-after study observed a notable rise in the percentage of patients receiving a daily sedative interruption with Tele-ICU support, which involved the incorporation of a third shift of Tele-ICU pharmacist assistance [18].

In addition, recently, a cluster randomized controlled trial found that telemedical quality improvement program increased adherence to 7 evidence-based German performance indicators in acute ICU care [19]. The quality indicators included were sedation, analgesia and delirium, ventilation, weaning from ventilation, infection management, enteral nutrition, patient and family communication, and early mobilization [19].

Nevertheless, there is significant heterogeneity among various Tele-ICU models, making these studies potentially noncomparable [8,9]. There are notable discrepancies concerning various characteristics in the previous studies regarding the technology and hospitals used, as well as the autonomy of Tele-ICU intensivists [8,9,18].

Tele-ICU COVID-19 Brazil Program was based on the implementation of multidisciplinary rounds conducted by a board-certified tele-intensivist in ICUs lacking daily critical care specialists. Daily rounds conducted by an intensivist have been previously associated with a 3-fold reduction in hospital mortality in surgical ICUs [20]. Furthermore, a study exploring organizational factors in Brazilian ICUs did not identify a direct impact of regular multidisciplinary rounds on mortality; however, the findings suggest that collaborative multidisciplinary efforts among ICU care providers positively influence patient outcomes [21].

The Tele-ICU program was also linked to an increase in the use of noninvasive ventilation and a decrease in the need of invasive mechanical ventilation. We hypothesize that the adoption of noninvasive ventilation increased as intensivists encouraged its use, and local teams observed that patients responding positively to noninvasive ventilation could safely be managed (especially considering the fear of aerosolization and spread of SARS-CoV-2 in the environment, which was highly prevalent in the first COVID-19 wave).

Interestingly, no independent association was found between overall ICU engagement (ICU coverage) or duration of ICU participation in the program and patient outcomes. Although we initially hypothesized that greater program exposure or broader ICU involvement would lead to better results, our findings suggest that the direct and consistent implementation of daily Tele-ICU rounds may have been the most influential factor, rather than general program duration or institutional engagement.

The program may also have enhanced continuity of care, as the remote intensivist overseeing the rounds remained mostly consistent from Monday to Friday, whereas the local team members often changed daily during the weekdays. Continuity of care is known to be a fundamental aspect of effective medical treatment and is linked to enhanced patient satisfaction, greater use of care services, and reduced health care costs [22]. Maintaining continuity of care in the ICU environment is particularly challenging, especially due to the prolonged duty hours that can impact the well-being of physicians [23]. In this scenario, the Tele-ICU solution emerges as a potential solution, providing a more stable and continuous presence to support ongoing patient care.

Limitations

Finally, our study has some limitations. First, it has a retrospective design, thus reporting associations rather than

XSL•FO
RenderX

cause-and-effect relationships. Second, we used the maximum SOFA score to quantify the severity of illness in patients, which may overestimate the severity of patients with COVID-19 included in our analysis. Third, although our study includes a considerable number of patients from multiple ICUs across different regions, the Tele-ICU program itself was coordinated by a single center, which may limit the external validity of the findings. Fourth, we acknowledge the potential variability in the presence of local intensivists across participating ICUs. While all selected ICUs lacked full-time intensivist coverage, which was a key inclusion criterion, there may have been occasional availability of specialists in some sites. This unmeasured variability may have influenced local care dynamics and should be considered as a potential confounding factor in interpreting the results. Finally, data on tracheostomy procedures were not collected in this study, which prevents evaluation of their potential impact on outcomes such as ICU mortality, duration of mechanical ventilation, and LOS [24]. Therefore, these should be considered when assessing generalizability, and the study should be considered as hypothesis generating.

Conclusions

During the COVID-19 pandemic, the Tele-ICU COVID-19 Brazil Program showed that greater exposure to DMRs led by remote intensivists was associated with lower ICU mortality and shorter ICU stays. These findings suggest that structured tele–critical care support may help extend specialized expertise, promote adherence to best practices, and enhance continuity of care in settings with limited intensivist availability. The results also highlight the potential value of Tele-ICU models in low- and middle-income countries and during periods of health care system strain. Further research is needed to confirm these findings, identify which program components drive the greatest benefit, and evaluate different Tele-ICU configurations and their cost-effectiveness to support future implementation efforts.

Acknowledgments

The authors would like to thank the intensive care unit (ICU) physicians, nursing staff, physical therapists, and all members of the multidisciplinary teams of the participating ICUs and of the Tele-ICU COVID-19 Brazil Program, who managed patients during the COVID-19 pandemic. The authors would also like to thank the Brazilian Ministry of Health for its support of the Tele-ICU COVID-19 Brazil Program. The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the Generative Artificial Intelligence Disclosure and Transparency Taxonomy (GAIDeT) taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: text and image generation. The GenAI tool used was ChatGPT (OpenAI). Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes. ChatGPT was used to refine sentence structure and correct grammar and punctuation errors.

Funding

The Tele-ICU COVID-19 Brazil Program was primarily funded by the Brazilian Ministry of Health through the Institutional Development Program of the Unified Health System (PROADI-SUS), which covered the costs of physician (tele-consultant) services. No specific funding was received for this study, and the funder had no role in the study design, conduct, or publication.

Data Availability

The datasets analyzed during this study are not publicly available due to ethical and regulatory restrictions, including requirements from Brazilian regulatory agencies. However, anonymized data may be made available from the corresponding author upon reasonable request, subject to appropriate regulatory approvals.

Conflicts of Interest

None declared.

References

1. Weekly epidemiological update on COVID-19 - 1 September 2023. World Health Organization. 2023 Sep 1. URL: https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---1-september-2023 [accessed 2026-02-27]
2. Van Damme W, Dahake R, Delamou A, Ingelbeen B, Wouters E, Vanham G, et al. The COVID-19 pandemic: diverse contexts; different epidemics-how and why? BMJ Glob Health 2020 Jul 27;5(7):e003098 [FREE Full text] [doi: 10.1136/bmjgh-2020-003098] [Medline: 32718950]
3. Bastos LS, Ranzani OT, Souza TM, Hamacher S, Bozza FA. COVID-19 hospital admissions: Brazil's first and second waves compared. Lancet Respir Med 2021 Aug;9(8):e82-e83. [doi: 10.1016/s2213-2600(21)00287-3]
4. Merola R, Marra A, Simone SD, Vargas M. Telemedicine in intensive care unit: current practice and future prospect. J Intensive Care Med 2025 Apr;40(4):456-463. [doi: 10.1177/08850666251325782] [Medline: 40123239]
5. Rosenfeld BA, Dorman T, Breslow MJ, Pronovost P, Jenckes M, Zhang N, et al. Intensive care unit telemedicine: alternate paradigm for providing continuous intensivist care. Crit Care Med 2000 Dec;28(12):3925-3931. [doi: 10.1097/00003246-200012000-00034] [Medline: 11153637]

XSL•FO
RenderX

6. Thomas EJ, Lucke JF, Wueste L, Weavind L, Patel B. Association of telemedicine for remote monitoring of intensive care patients with mortality, complications, and length of stay. JAMA 2009 Dec 23;302(24):2671-2678. [doi: 10.1001/jama.2009.1902] [Medline: 20040555]
7. Wilcox ME, Adhikari NK. The effect of telemedicine in critically ill patients: systematic review and meta-analysis. Crit Care 2012 Jul 18;16(4):R127 [FREE Full text] [doi: 10.1186/cc11429] [Medline: 22809335]
8. Chen J, Sun D, Yang W, Liu M, Zhang S, Peng J, et al. Clinical and economic outcomes of telemedicine programs in the intensive care unit: a systematic review and meta-analysis. J Intensive Care Med 2018 Jul 22;33(7):383-393. [doi: 10.1177/0885066617726942] [Medline: 28826282]
9. Young LB, Chan PS, Lu X, Nallamothu BK, Sasson C, Cram PM. Impact of telemedicine intensive care unit coverage on patient outcomes: a systematic review and meta-analysis. Arch Intern Med 2011 Mar 28;171(6):498-506. [doi: 10.1001/archinternmed.2011.61] [Medline: 21444842]
10. Fusaro MV, Becker C, Scurlock C. Evaluating tele-ICU implementation based on observed and predicted ICU mortality: a systematic review and meta-analysis. Crit Care Med 2019 Apr;47(4):501-507. [doi: 10.1097/CCM.0000000000003627] [Medline: 30688718]
11. Watanabe T, Ohsugi K, Suminaga Y, Somei M, Kikuyama K, Mori M, et al. An evaluation of the impact of the implementation of the tele-ICU: a retrospective observational study. J Intensive Care 2023 Mar 07;11(1):9 [FREE Full text] [doi: 10.1186/s40560-023-00657-4] [Medline: 36882878]
12. Vranas KC, Slatore CG, Kerlin MP. Telemedicine coverage of intensive care units: a narrative review. Ann Am Thorac Soc 2018 Nov;15(11):1256-1264 [FREE Full text] [doi: 10.1513/AnnalsATS.201804-225CME] [Medline: 30382786]
13. Lilly CM, Fisher KA, Ries M, Pastores SM, Vender J, Pitts JA, et al. A national ICU telemedicine survey: validation and results. Chest 2012 Jul;142(1):40-47. [doi: 10.1378/chest.12-0310] [Medline: 22518025]
14. Fuhrman SA, Lilly CM. ICU telemedicine solutions. Clin Chest Med 2015 Sep;36(3):401-407. [doi: 10.1016/j.ccm.2015.05.004] [Medline: 26304277]
15. Pereira AJ, Noritomi DT, Dos Santos MC, Corrêa TD, Ferraz LJ, Schettino GP, et al. Effect of tele-ICU on clinical outcomes of critically ill patients: the TELESCOPE randomized clinical trial. JAMA 2024 Dec 03;332(21):1798-1807. [doi: 10.1001/jama.2024.20651] [Medline: 39382244]
16. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Med 1996 Jul;22(7):707-710. [doi: 10.1007/bf01709751]
17. Lilly CM, Cody S, Zhao H, Landry K, Baker SP, McIlwaine J, et al. Hospital mortality, length of stay, and preventable complications among critically ill patients before and after tele-ICU reengineering of critical care processes. JAMA 2011 Jun 01;305(21):2175-2183. [doi: 10.1001/jama.2011.697] [Medline: 21576622]
18. Forni A, Skehan N, Hartman CA, Yogaratnam D, Njoroge M, Schifferdecker C, et al. Evaluation of the impact of a tele-ICU pharmacist on the management of sedation in critically ill mechanically ventilated patients. Ann Pharmacother 2010 Mar;44(3):432-438. [doi: 10.1345/aph.1M576] [Medline: 20164471]
19. Spies CD, Paul N, Adrion C, Berger E, Busse R, Kraufmann B, et al. Effectiveness of an intensive care telehealth programme to improve process quality (ERIC): a multicentre stepped wedge cluster randomised controlled trial. Intensive Care Med 2023 Feb;49(2):191-204 [FREE Full text] [doi: 10.1007/s00134-022-06949-x] [Medline: 36645446]
20. Pronovost PJ, Jenckes MW, Dorman T, Garrett E, Breslow MJ, Rosenfeld BA, et al. Organizational characteristics of intensive care units related to outcomes of abdominal aortic surgery. JAMA 1999 Apr 14;281(14):1310-1317. [doi: 10.1001/jama.281.14.1310] [Medline: 10208147]
21. Soares M, Bozza FA, Angus DC, Japiassú AM, Viana WN, Costa R, et al. Organizational characteristics, outcomes, and resource use in 78 Brazilian intensive care units: the ORCHESTRA study. Intensive Care Med 2015 Dec 28;41(12):2149-2160. [doi: 10.1007/s00134-015-4076-7] [Medline: 26499477]
22. Sharma G, Freeman J, Zhang D, Goodwin JS. Continuity of care and intensive care unit use at the end of life. Arch Intern Med 2009 Jan 12;169(1):81-86 [FREE Full text] [doi: 10.1001/archinternmed.2008.514] [Medline: 19139328]
23. Parshuram CS, Amaral AC, Ferguson ND, Baker GR, Etchells EE, Flintoft V, et al. Patient safety, resident well-being and continuity of care with different resident duty schedules in the intensive care unit: a randomized trial. CMAJ 2015 Mar 17;187(5):321-329 [FREE Full text] [doi: 10.1503/cmaj.140752] [Medline: 25667258]
24. Merola R, Iacovazzo C, Troise S, Marra A, Formichella A, Servillo G, et al. Timing of tracheostomy in ICU patients: a systematic review and meta-analysis of randomized controlled trials. Life (Basel) 2024 Sep 14;14(9):1165 [FREE Full text] [doi: 10.3390/life14091165] [Medline: 39337948]

Abbreviations

DMR: daily multidisciplinary round
ICU: intensive care unit
LOS: length of stay
OR: odds ratio

XSL•FO
RenderX

SOFA: Sequential Organ Failure Assessment

Edited by A Mavragani, T de Azevedo Cardoso; submitted 01.Aug.2024; peer-reviewed by D Schnell, R Merola; comments to author 21.Apr.2025; revised version received 19.Dec.2025; accepted 30.Dec.2025; published 19.Mar.2026.

Please cite as:
Midega TD, Hohmann FB, Chaves RCDF, Almeida GMBD, Leoneza VJL, Cabral JFF, Vitalino BV, Camboim EDA, Machado NDJN, Melo RFBD, Feliciano JPO, Cardoso BM, Ferraz LJR, Corrêa TD, Santos MCD, Albaladejo Morbeck R, Pereira AJ
Impact of Tele-Intensive Care Units on the Clinical Outcomes of Critically Ill Patients With COVID-19: Retrospective Cohort Study
J Med Internet Res 2026;28:e64996
URL: https://www.jmir.org/2026/1/e64996
doi:10.2196/64996
PMID:

©Thais Dias Midega, Fabio Barlem Hohmann, Renato Carneiro de Freitas Chaves, Guacyra Margarita Batista de Almeida, Vivian Jaqueline Lima Leoneza, Jennifer Ferreira Figueiredo Cabral, Bianca Veloso Vitalino, Emanuelle de Araújo Camboim, Nelma de Jesus Nogueira Machado, Ricardo Fernando Batista de Melo, Jorge Patrick Oliveira Feliciano, Breno Mendes Cardoso, Leonardo José Rolim Ferraz, Thiago Domingos Corrêa, Maura Cristina dos Santos, Renata Albaladejo Morbeck, Adriano José Pereira. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Supporting Access to Care Through Peripheral Devices and Patient-Generated Health Data: Qualitative Study

Ashley C Griffin[1,2], PhD; Caroline Gray[1], PhD; Victoria Ngo[3], PhD; Camila Chaudhary[1], MPH; Stephanie L Shimada[4,5,6], PhD; Amanda M Midboe[1,7], PhD; Donna M Zulman[1,8], MD, MS

[1]Center for Innovation to Implementation, VA Palo Alto Health Care System, Menlo Park, CA, United States

[2]Center for Biomedical Informatics Research, School of Medicine, Stanford University, Stanford, CA, United States

[3]New England Geriatric Research Education and Clinical Center, VA Bedford Healthcare System, Bedford, MA, United States

[4]Center for Health Optimization & Implementation Research, VA Bedford Healthcare System, Bedford, MA, United States

[5]Department of Health Law, Policy, and Management, Boston University School of Public Health, Boston, MA, United States

[6]Department of Population and Quantitative Health Sciences, University of Massachusetts Chan Medical School, Worcester, MA, United States

[7]Department of Public Health Sciences, Division of Health Policy and Management, School of Medicine, University of California Davis, Davis, CA, United States

[8]Division of Primary Care and Population Health, Department of Medicine, School of Medicine, Stanford University, Stanford, CA, United States

Corresponding Author:
Ashley C Griffin, PhD
Center for Innovation to Implementation
VA Palo Alto Health Care System
795 Willow Road
Menlo Park, CA, 94025
United States
Phone: 1 650 493 5000
Email: griffina@stanford.edu

Abstract

Background: In 2016, the US Department of Veterans Affairs (VA) implemented a national initiative to distribute video-enabled tablets and peripheral devices, such as blood pressure monitors and weighing scales, to patients facing geographic, clinical, or socioeconomic challenges. Such patients could potentially benefit from health monitoring in conjunction with video-based care, as peripheral devices offer opportunities to enrich care received during a video visit and support tracking of health-related data collected outside of clinical care, or patient-generated health data. However, little is known about experiences with the devices and how they could support improved access to care.

Objective: We explored patients' experiences with VA-issued peripheral devices and their impact on video-based care and health monitoring outside of clinical visits.

Methods: We conducted in-depth semistructured interviews among patients who received VA-issued tablets and peripheral devices between 2023 and 2024. Purposive sampling was used to gather views based on gender, age, race or ethnicity, and rurality. Interviews were transcribed and analyzed using rapid qualitative analysis, guided by the Unified Theory of Acceptance and Use of Technology.

Results: Among 25 patients, most received a blood pressure monitor (21/25, 84%), a weight scale (14/25, 56%), and/or a pulse oximetry device (12/25, 48%). The majority reported using their peripheral devices (23/25, 92%) and tablets (19/25, 76%) to monitor their vital signs and attend video visits. Qualitative analysis yielded ten themes reflecting experiences and impacts of the devices, organized by the Unified Theory of Acceptance and Use of Technology constructs: "effort expectancy" consisted of (1) familiar and easy to use devices and (2) challenges of Bluetooth pairing and measurement; "performance expectancy" consisted of (3) integration with video visits, (4) health monitoring for peace of mind, (5) perceptions of improved vital signs and lifestyle behaviors, (6) removing obstacles to in-person care, and (7) desiring an overall picture of health; "social influence" consisted of (8) fostering care team connections and (9) promoting awareness of tablets and peripheral devices; and "facilitating conditions" consisted of (10) supportive help desk infrastructure. Overall, patients described using peripheral devices during virtual visits by syncing data to the tablet for real-time access by their care team. They also reported manually tracking and sharing patient-generated health data with their care team. Despite some challenges with Bluetooth pairing, patients found the devices easy to use and

XSL•FO
RenderX

contributed to improved health and motivation. Devices also reduced logistical burdens of in-person visits, especially for those with limited mobility, visual impairments, mental health needs, or transportation barriers.

Conclusions: Patients perceive that peripheral devices can enhance video-based care and support health care access and chronic disease management. Patients reported benefits to health, behavior, and communication with care teams. To maximize the impact, program enhancements should prioritize device interoperability, accessible training, and expanded outreach.

(J Med Internet Res 2026;28:e88405) doi:10.2196/88405

KEYWORDS

remote patient monitoring; self-management; health services accessibility; patient generated health data; veterans; mobile phone

Introduction

Background

Multiple forces are aligning to integrate patients' remote health data into care settings, showing strong potential to improve health outcomes. With 91% of adults in the United States owning a smartphone [1], the proliferation of consumer health technologies presents numerous opportunities for patients to track and monitor their health through apps and peripheral devices, such as blood pressure monitors, weighing scales, or pulse oximetry devices. A growing proportion of the US population is estimated to use a wearable device to monitor their health (40%), of which the majority (73%) are willing to share these data with their care team [2]. In addition to patient interest in tracking and sharing their health data, substantial progress has been made in the remote monitoring regulatory and interoperability ecosystems [3-5]. Notably, the Fast Healthcare Interoperability Resources standard [6] provides a framework to exchange data between health information systems, devices, apps, and electronic health records (EHRs). These regulatory and technological advancements have given rise to a plethora of patient-generated health data (PGHD), or data created and recorded by patients outside of the clinical setting [7].

Efforts within health systems to leverage PGHD are beginning to flourish [8-11], yet numerous studies have identified digital health disparities in access and use of peripheral devices [12-14]. Patients with lower socioeconomic status, limited internet access or connectivity, barriers to care, and older adults have been most impacted by the digital divide. These disparities can impact care and outcomes, as the use of peripheral devices and PGHD has demonstrated early value in personalized medicine and improvements in physiological measures and symptoms across a variety of chronic conditions [9,10,15,16]. Given the value of peripheral devices, system-wide efforts are needed to support greater access and more equitable use to allow all patients to receive these benefits.

To combat digital disparities, the US Department of Veterans Affairs (VA) distributes internet-connected tablets and peripheral devices to VA patients with barriers to accessing in-person care [17]. Patients are eligible for a tablet if they lack a device or connectivity, have a barrier to care (ie, geographical, clinical, or social), and require a VA video-based appointment within the next 90 days. In addition to the tablet, depending on the patient's clinical needs, they can receive 1 or more peripheral devices including a blood pressure monitor, pulse oximetry device, weighing scale, stethoscope, or thermometer. Peripheral devices are intended to be connected to the tablet via Bluetooth and used during a video visit to share a patient's vital signs with the care team. The tablets also contain a variety of preinstalled VA apps, allowing patients to engage with the VA patient portal or VA Mobile App Store [18]. Previous work has shown increased registration and use of VA's patient portal after receiving a tablet [19] and demonstrated how tablets foster opportunities to use VA apps for multiple health management needs [20].

Despite growing evidence supporting remote monitoring and PGHD integration, little is known about patient experiences using peripheral devices in conjunction with video-enabled tablets, and how they use these devices outside of clinical visits to collect and track health data. We also have a limited understanding of how these technologies are experienced by patients who face barriers to care. Previous work has examined patients' expectations and experiences with remote monitoring across various chronic conditions, highlighting how patients gain knowledge about their condition, feel reassured about their health, and how the devices support self-management and shared decision-making [21]. Patients have also reported reluctance to use new technologies, lack of trust in technology, concerns about costs, and fear of their data being lost or losing face-to-face communication with their care team [21]. Much of the existing telehealth and remote monitoring literature has focused on disease-specific interventions or on patient populations who already have access to technology [22-24].

National health system initiatives designed to mitigate these inequities, such as the VA tablet and peripheral device distribution program, aim to address structural barriers to care by providing devices, connectivity, and technical support to patients with documented challenges [17]. Previous evaluations of this initiative have demonstrated improvements in access to mental health services, continuity of care, and reductions in emergency department visits among VA patients receiving tablets [25,26]. These evaluations have mostly focused on usage and clinical outcomes rather than the patient's lived experience using peripheral devices and the data generated from them. We extend existing work within the context of patients who experience barriers to accessing health care services, including physical and mental health challenges, socioeconomic stressors, and transportation difficulties. Patients with these challenges potentially stand to gain substantial benefits from peripheral devices.

XSL•FO
RenderX

Objective

Among VA patients who received a tablet and 1 or more peripheral devices, we aimed to assess their experiences with the devices, including setup, use, technical and social support, and integration with their video-enabled tablet. We also examined how patients used PGHD generated from the peripheral devices to enhance their health, including through health monitoring, communication with care teams, and access to care.

Methods

Setting

The VA Office of Rural Health and Office of Connected Care tablet initiative provides tablets (ie, Apple iPads with T-Mobile [Deutsche Telekom] or Verizon [Verizon Communications Inc] data plans) and peripheral devices to VA patients based on an assessment by clinicians and social workers using a Digital Divide Consult template in the VA's EHR [27]. As of March 2026, more than 10,500 VA patients have been shipped both a peripheral device and a tablet. The primary goal of peripheral device distribution is to use the devices during a video visit, in which devices are paired via Bluetooth to the tablet. During the video visit, data are displayed to the care team within the VA Video Connect app and can be entered into the EHR. Devices can also be used without Bluetooth, but data are not shared with VA care teams when used without Bluetooth. Upon receipt of a device, patients receive a phone call from the VA Connected Device Support Team to assist with setting up and pairing the Bluetooth device with the tablet.

Ethical Considerations

This evaluation was conducted as part of the Virtual Access Quality Enhancement Research Initiative and was designated as nonresearch by the Stanford University institutional review board (#70128). All patients provided consent to participate in the interview and received a US $50 gift card upon completion. All recruitment and interview data were stored in a private and secure location to ensure confidentiality.

Sample and Sampling

To identify patients for in-depth semistructured interviews, we used EHR data from the VA Corporate Data Warehouse (CDW) that included sociodemographic and peripheral device characteristics. Patients were eligible if they were shipped a peripheral device between February 2023 and February 2024, and had a physical mailing address, email address, and phone number on file. We used purposive sampling to include perspectives from at least 25% (7/25) of women, older adults (older than 65 y), patients of diverse race or ethnicity, and those who resided in a rural area. Rurality was based on Census tract Rural Urban Commuting Areas codes corresponding to each patient's ZIP code [28].

We also aimed to gather views among patients who received different devices (ie, blood pressure cuff, weighing scale, pulse oximetry, thermometer, and stethoscope), those who received a device setup call from the VA Connected Device Support Team, and those who received a peripheral device and tablet but did not use it. Setup calls were tracked on a VA dashboard and then linked to our sample in the CDW. Use of the peripheral device and tablet was reported during the interview prescreening process and noted in our recruitment sheet. We also characterized the sample based on the most common chronic physical and mental health conditions from the CDW based on the list from Yoon et al [29].

Recruitment

Patients were first contacted through an email explaining the project and asking if they were interested in participating. If they expressed interest, they received a phone call to confirm that they had obtained a peripheral device and to schedule an interview. If there was no response to the initial recruitment email, 2 follow-up phone calls were attempted. Recruitment occurred between March 2024 and August 2024.

Interview Guide Development

To examine patients' experiences with the devices, an interview guide was developed using constructs from the Unified Theory of Acceptance and Use of Technology (UTAUT) [30], which is a well-established framework to assess acceptance and behavioral intention to use health technologies. Interview questions related to UTAUT constructs were developed to elicit input on various topics, including device setup and use, technical and social support, communication with the care team, integration with the tablet, use of data generated from the device, and improvements in health (Table 1).

XSL•FO
RenderX

Table 1. Interview questions mapped to constructs from the Unified Theory of Acceptance and Use of Technology.

Construct	Definition applied to this study	Questions
Effort expectancy	The degree of ease associated with using the peripheral device	• What was your experience like setting up the device? • What was it like learning to use the device? • How often do you use it?
Performance expectancy	The degree to which a patient believes that using the peripheral device will help them to improve their health	• Do you consider the device to be useful? (Why or why not?) • Have you used the device before or during a video visit or in-person visit with your care team? (If so, please describe how you used the device.) • How did receiving a device help you talk with your doctor? • How have you used the data generated from the device? Has it resulted in any lifestyle changes? • In your opinion, can the device improve your [blood pressure, weight, other health monitoring]?
Social influence	The degree to which a patient perceives that important others believe they should use the peripheral device	• Who is supportive of you using the device? • Do you have anyone in your support circle (family or friends) who helps you use the device?
Facilitating conditions	The degree to which a patient believes that organizational and technical infrastructure exists to support use of the peripheral devices	• Have you had prior training on how to set up and use the device? • What did you find most challenging about setting up the device? Using the device? • Are you able to get technical support on the device? • What other information or support do you need to use the device?

Interviews

Interviews were conducted by 2 trained qualitative researchers (ACG and CG) using Microsoft Teams software. Each interview included the interviewer and patient, with the exception of 1 interview that also involved the patient's caregiver and 1 interview with a late patient's family member. Thus, 26 individuals were interviewed to represent the 25 patients. Demographic information is presented for the 25 patients who received the peripheral devices. Interviews lasted for 30-60 minutes and were audio and/or video-recorded upon consent. Interviews were completed between April 2024 and August 2024.

Analysis

Interview data were transcribed verbatim by Keystrokes Transcription Service and imported into ATLAS.ti qualitative data analysis software (version 23; Lumivero). Patient narratives within the transcriptions were analyzed using rapid qualitative analysis, guided by the Planning for Assessing Rigor in Rapid Qualitative Analysis framework [31]. One team member (CG) reviewed the transcripts and interviewer's notes to summarize the content into a matrix based on the initial topics from the interview guide using ATLAS.ti. The matrix was then exported into Microsoft Excel, where a second team member (ACG) reviewed the data summaries and referenced the notes and recordings as needed. The resulting summaries were then discussed between the 2 team members (ACG and CG) and iteratively grouped into themes within Excel using a hybrid inductive-deductive approach to allow for emergent themes. The larger project team also provided input on the themes. Illustrative quotes were identified within each theme. To organize the themes and allow for comparison with related studies, we grouped each theme under the most relevant UTAUT construct.

The study team actively engaged in ongoing reflexive practices throughout data collection and analysis. We are aware that our personal and professional backgrounds shape interactions with patients and the interpretation of findings. Similarly, our role as researchers affiliated with the VA may influence patients' social desirability by potentially overreporting positive experiences or underreporting challenges with the devices because they perceive the interviewers as part of their care system. The study team had no established relationships with patients, and interviews were conducted using open-ended questions and neutral language. Iterative discussions with our multidisciplinary study team helped us remain aware of our positioning and avoid assumptions when interpreting the data.

Results

Patient Characteristics

Among 156 patients contacted, 25 (16%) participated in an interview. The mean age was 62 (SD 11) years, 32% (8/25) were female, 40% (10/25) were Black and/or African American, and 44% (11/25) lived in a rural area. Patients had a range of common chronic physical and mental health conditions (Table 2). Based on the EHR Digital Divide Consult, (1) a mental health diagnosis or participation in therapy (7/25, 28%) or (2) living farther than 30 miles from a VA Medical Center (6/25, 24%) were the most common reasons for receiving a tablet and peripheral device. The majority received a blood pressure monitor (21/25, 84%), weighing scale (14/25, 56%), and/or pulse oximetry device (12/25, 48%) from the VA (Table 2). Nearly all (23/25, 92%) reported using at least one of their peripheral devices. The 2 patients who were unable to use their peripheral devices received only a blood pressure monitor and tablet. They explained that the blood pressure monitor was missing a hardware component (ie, arm cuff and display

XSL•FO
RenderX

monitor) that prevented them from taking their blood pressure. Others reported using some of their devices but not all of them due to difficulty with Bluetooth setup, already having their own device at home, or not being sure why they received certain devices. Approximately three-quarters (19/25, 76%) reported using their tablet to do a video visit with their care team. Of the 18 patients who used both the tablet and peripheral devices, 9 (50%) were able to successfully connect them via Bluetooth so that their care team could access data in real-time during a video visit. Of the 6 patients who had not done a video visit, 3 described being unable to operate the tablet despite receiving a phone call from the VA Connected Device Support Team, 2 patients did not recall receiving a support call and were unable to start the video visit on their own, and 1 preferred using their desktop for video visits. According to the VA Connected Device Support Team dashboard, most received assistance to set up the peripheral devices and/or tablet, either by the VA initiating contact (15/25, 60%) or the patient initiating contact (6/25, 24%).

XSL•FO
RenderX

Table 2. Sociodemographic, clinical, social, and technological characteristics of patients who received a peripheral device (n=25).

Characteristics	Patients, n (%)
Sociodemographic	
Age group (y; mean 62, SD 11 y)	
18-44	3 (12)
45-64	10 (40)
65-74	9 (36)
75-84	3 (12)
Sex	
Female	8 (32)
Male	17 (68)
Race or ethnicity	
Black and/or African American	10 (40)
Hispanic or Latino	2 (8)
More than one race	1 (4)
Native American or Alaska Native	1 (4)
White	10 (40)
Unknown	1 (4)
Rurality	
Rural	11 (44)
Urban	14 (56)
Clinical and social	
Physical health conditions	
Hypertension	17 (68)
Heart failure, heart disease, or atrial fibrillation	9 (36)
Hyperlipidemia	16 (64)
Diabetes	9 (36)
Obesity	8 (32)
Lower back pain, fibromyalgia, or chronic migraine	17 (68)
Rheumatoid arthritis or osteoarthritis	6 (24)
Cataract, glaucoma, or visual impairments	8 (32)
Mental health conditions	
Depression	14 (56)
Anxiety	13 (52)
PTSD[a]	13 (52)
Drug or alcohol use disorder	6 (24)
Digital divide consult reasons (clinicians can select multiple reasons)	
Mental health diagnosis or participating in therapy	7 (28)
Lives father than 30 miles from VA[b] medical center	6 (24)
Health issues make the patient homebound	4 (16)
Difficulty attending visits at VA facility (eg, psychological distress or immunocompromised)	4 (16)
Work, school, or caregiver commitments make in-person visits challenging	4 (16)

XSL•FO
RenderX

Characteristics	Patients, n (%)
Other health issues (ie, hospitalized in last 90 days, hospice diagnosis, requires frequent video visits)	4 (16)
Other reasons (ie, social isolation, costs, or difficulty with public transport)	3 (12)
Technological	
Peripheral devices (patients can receive multiple devices)	
Blood pressure monitor	21 (84)
Weighing scale	14 (56)
Pulse oximetry device	12 (48)
Thermometer	7 (28)
Stethoscope	5 (20)
Ever used a peripheral device	23 (92)
Ever had a video visit on tablet	19 (76)
Tablet and/or peripheral device setup call	
VA-initiated setup complete	15 (60)
Patient-initiated setup complete	6 (24)
Patient requested in-person setup at local VA	2 (8)
Declined assistance or did not answer	2 (8)

[a]PTSD: posttraumatic stress disorder.

[b]VA: US Department of Veterans Affairs.

Experiences With Peripheral Devices

Overall, patients described using the peripheral devices for a variety of purposes, ranging from monitoring blood pressure post surgery, participating in virtual diabetes management classes, and engaging in home-based primary care programs, among other medical reasons. Patients reported that the devices offered many benefits when used in combination with the tablet, which was the intended use of receiving a peripheral device as part of the VA's national telehealth initiative. Patients also described valuable aspects of using the devices in between visits and having the ability to track and reflect on their health data over time. They discussed how the devices were simple to set up and use, and they highlighted opportunities to improve their distribution and usability. The following interviews revealed 10 themes reflecting patients' insights about and impacts of using the peripheral devices (Textbox 1).

XSL•FO
RenderX

Textbox 1. Themes for patient experiences with peripheral devices corresponding to constructs from the Unified Theory of Acceptance and Use of Technology.

Effort expectancy

- Familiar and easy to use devices
- Challenges of Bluetooth pairing and measurement

Performance expectancy

- Integration with video visits
- Health monitoring for peace of mind
- Perceptions of improved vital signs and lifestyle behaviors
- Removing obstacles of in-person care
- Desiring an overall picture of health

Social influence

- Fostering care team connections
- Promoting awareness of tablets and peripheral devices

Facilitating conditions

- Supportive help desk infrastructure

Effort Expectancy

Familiar and Easy to Use Devices

The devices were easy to set up and use, and patients discussed how they had used similar devices before. One patient described,

> *It was really easy. The first couple of times I learned it and just kept track of everything… I did that [set up] on my own, and the instructions were pretty easy.* [P5]

Among the 2 caregivers, both reported using similar devices and felt they were easy to operate. The caregiver of a home-based patient explained,

> *I've been taking care of him for a long time because I'm his caregiver plus his wife… I check his blood pressure, his temperature, and write it down in my little tablet. I'm up-to-date on it. You get the hang of everything after a while.* [P2]

Challenges of Bluetooth Pairing and Measurement

The greatest challenge with using the peripheral devices was connecting them to the iPad (Apple Inc) via Bluetooth. One patient felt,

> *It was just an issue connecting it… And I got frustrated and just used my own scale.* [P10]

Bluetooth pairing issues were mentioned across all types of peripheral devices, and some patients were unsure if their devices were Bluetooth-enabled. Patients who were unable to pair the device felt that having a Bluetooth connection would be useful so that their care team could see their readings.

There were also some concerns about the accuracy of the devices. Sometimes patients felt their blood pressure readings were too high or too low and brought in their devices to the clinic for help. One person reported,

> *I think I had been using [the blood pressure cuff] wrong at first because I kept getting these really low readings. So I called them and told them it didn't work, and when I went in and got to see the nurse in one of my appointments, she showed me how I should be using it. Now that I use it right and more regularly, I feel comfortable about the fact that I'm tracking my blood pressure regularly.* [P8]

It was recommended to have the option for classes to set up and understand how to take accurate measurements,

> *Sometimes you need to understand a little more about your blood pressure than just the cuff and the button that you push because it makes a difference on your feet being on the floor and stuff like that.* [P24]

Performance Expectancy

Integration With Video Visits

Patients described using the peripheral devices during video visits in which the devices were paired via Bluetooth to the tablet. This allowed patients to take their vital signs during the visit and display them automatically on the screen for their care team. These patients were usually part of a VA program, such as home-based primary care or a diabetes management program. One patient explained,

> *The weight scale is Bluetooth connected to the iPad. So when I step on it, it will just tell me to step off, and it'll put my weight on the iPad for the provider that I'm talking to so that they get the information. The same thing with the arm cuff… It automatically shows up on the screen.* [P4]

XSL•FO
RenderX

Patients could also take their vitals right before the visit, with one patient discussing how,

> *I describe my data to my care team and it automatically sends it to them too. It's all connected in one system there with the tablet.* [P16]

The stethoscope and thermometer, while less frequently discussed compared with the other types of devices, were commonly used in real-time during video visits.

Health Monitoring for Peace of Mind

Patients found the devices helpful for monitoring and feeling in control of their health. One patient stated,

> *I feel better and less anxious about my health because I have some control over some of the things. I don't have to wait for an appointment to get my blood pressure checked or to get weighed in and see if I gained or lost weight. So, it's just a feeling of more control and it reduces my anxiety and stress.* [P8]

The devices also provided reassurance for some patients that their medications were working properly and helped them feel less stressed about their blood pressure. Patients mentioned the benefits of monitoring their blood pressure after starting new medications (eg, steroids) and for certain medications that they were instructed not to take when their blood pressure was too low (eg, beta blockers).

The blood pressure monitor, weighing scale, and pulse oximeter were commonly used on a daily basis, such as to check blood pressure before taking medicine, before pacemaker battery replacement, or if something felt off. One patient explained,

> *On the oximeter, sometimes [I use it] two or three times a day. When I come back in from getting the groceries or whatever, and I have a hard time breathing, I take my oxygen level.* [P6]

Patients discussed how being able to take their vitals at home gave them comfort in helping to identify or distinguish between different health issues. For example, one patient described using the pulse oximeter to determine the difference between a panic attack versus shortness of breath for physical reasons. They explained how if they were experiencing mild shortness of breath, but their oxygen levels were high, they could use techniques to calm down and then let their doctor know about the episode. One patient also described using the blood pressure cuff to identify a headache from elevated blood pressure versus a migraine,

> *I get migraines but also get headaches from high blood pressure. So, [the blood pressure monitor] identifies just how high it is, do I need to call someone, and if it's a headache due to high blood pressure… If my blood pressure is high enough for a headache, I need to go in.* [P19]

Perceptions of Improved Vital Signs and Lifestyle Behaviors

Patients reflected on how the devices made a positive impact on their health and lifestyle behaviors. For example, monitoring blood pressure and weight fostered other healthy habits, such as eating less salt, drinking more water, and walking more. One patient reported,

> *I have increased my walking and watched what I'm eating, and last month I went to my doctor – he said I had lost 12 pounds.* [P17]

Others self-reported losing weight or lowering their blood pressure after receiving the devices. Patients also elaborated on how the devices strengthened their motivation to improve their health and helped them stay accountable since a member of their care team was checking on the data. Notably, one person indicated they were able to be taken off their current blood pressure medication, and another described how they did not have to start a medication.

Removing Obstacles of In-Person Care

Patients described how taking their vitals at home through the peripheral devices helped them avoid the challenges of attending in-person care. Patients described having mobility and visual impairments, needing to transport medical equipment (eg, oxygen tanks and wheelchairs) or arrange ambulatory transportation. One patient elaborated on using the peripheral devices in combination with video visits, which avoided some of the in-person challenges with their wheelchair.

> *I'm just very, very thankful that I can go through a video chat rather than have to load my powerchair up every time… it's kind of rough for me to get all that hooked up, and go up there and unhook and then come out and hook it back up again. And then if it's raining, I've got to fight with the cover that goes over the chair so it doesn't ruin.* [P16]

Other patients mentioned how attending in-person visits required them to schedule time off from work or drive long distances, which was especially challenging for some patients with anxiety.

> *I live in a smaller suburban area and most of my appointments are in the city, so that really gets my anxiety up… Sometimes it's more beneficial to be able to be in the safety of your own little habitat, to stay and do your appointment there. Sometimes that's beneficial because you're not distracted and all worked up and anxious by the time you get there. And then you have to spend half the appointment trying to calm down and focus on what you're doing there and nervous about having to make the drive back.* [P25]

Desiring an Overall Picture of Health

Patients were interested in viewing combined visualizations of their health within 1 interface, including their weight trends, blood pressure readings, and other vital signs. One patient described,

> *It'd be nice to be able to see something that I could reflect on over time, even if it was just six months at a time, to see the trends and how it's going because I suspect that everything's going to be worse when I'm depressed... Opening an app and putting that information in would be really awesome because then I could look at the whole entire picture.* [P19]

XSL•FO
RenderX

Patients also reported tracking their peripheral device data in combination with personal device data, such as physical activity, nutrition, or glucose, to reflect on improvements in their health. Another patient stated, "*knowledge is power*" [P24] when referring to their interest and perceived benefit in tracking and visualizing data from devices over time.

Social Influence

Fostering Care Team Connections

Patients explained how using the devices supported communication and connection with their care team. One patient stated,

> *I believe this [blood pressure monitor] device and having the connection with the physical therapist from cardiology improved my desire to get better. It had an immense effect on not just my mental side but the physical side, as well. And at that time, it had more to do with my mental health more than even the physical health.* [P23]

Patients also communicated with their care team as a result of data from the devices, such as asking for advice about fluctuating blood pressure measurements or weight. Patients also discussed how a combination of technologies supported communication with their care team. For example, the VA's My Health*e*Vet patient portal was used to communicate with their care team about data from the devices, such as sending messages or screenshots of their data when they had questions. Others reflected on reviewing their personal health information in the portal and their communication with their care team.

Promoting Awareness of VA-Issued Tablets and Peripheral Devices

Patients heard about receiving the tablets and devices from other veterans, which prompted them to reach out to their care team. For example,

> *I found out about [the devices] through another veteran, and I called right then and got the ball rolling… I try to spread the information as much as I can… The people in my building don't know that they can get a tablet and make it easier to communicate with the hospital.* [P24]

Patients recommended that more veterans need to know about this program and that their care team should discuss the opportunity to provide a tablet and any needed peripheral devices to support the veteran.

Facilitating Conditions

Supportive Help Desk Infrastructure

Patients described receiving a phone call from the VA help desk to assist with setting up the peripheral devices to ensure they could take their vitals at home and start a video visit. Patients found the help desk personnel to be very helpful and effective in training them to use the technologies, with one patient stating how,

> *I feel more confident now that when I go to log on I'm going to get right on [to the video visit].* [P8]

Patients could also contact the help desk at any time, as a contact number was provided on the back of the peripheral devices. Patients described calling the help desk for support with pairing the device via Bluetooth but were usually not able to set up the Bluetooth connection over the phone. Overall, patients did not need any ongoing technical support with the devices but reported knowing how to reach out to the help desk if any technical issues came up.

Discussion

Principal Findings

This qualitative study examined the experiences of patients who received VA-issued peripheral devices and tablets as part of a national telehealth initiative designed to address access barriers to care. Overall, nearly all patients who were interviewed reported using the peripheral devices and discussed the benefits for their health management, improved communication with their care team, and how the devices reduced their burden of attending in-person health care services. Patients frequently used peripheral devices to supplement virtual appointments by syncing data to the tablet for real-time clinician access or by reporting the data to their clinician during the appointment. Despite some usability challenges with Bluetooth data transmission, patients found the devices easy to use and integrated them into their daily routine. These findings support the role of peripheral devices and PGHD in supporting care and reducing access disparities among patients with complex medical and mental health conditions.

Our findings align with previous telehealth and remote monitoring literature, including the enabling role of home-based monitoring, the importance of technical and social support, and ongoing challenges related to interoperability [9,21,22]. Our study, however, extends existing research by demonstrating how and why these dynamics operate within a unique patient population supported by a national digital equity intervention. Unlike previous literature that examines early adopters or patients already engaged in digital health, our cohort consisted of patients identified through an EHR digital divide assessment as having significant barriers to accessing in-person care. These barriers included limited previous access to digital technologies, high prevalence of mental health conditions, mobility limitations, rural residence, and transportation challenges. Although these characteristics have historically been associated with lower engagement in telehealth and patient portals [14,22,32], we found that most patients engaged with the peripheral devices and incorporated them into day-to-day health management.

Guided by UTAUT, we identified several aspects that support the adoption and use of peripheral devices. Patients found the devices were helpful in managing their health (performance expectancy) and described numerous health benefits, including reduced anxiety, increased confidence in their medications, and improved diet and physical activity. These findings resonate with existing work showing that PGHD integration can increase self-management and activation, thereby leading to lifestyle changes and better adherence to treatment plans and medication regimens [33]. Overall, the devices were simple to set up and

XSL•FO
RenderX

operate (effort expectancy), which may be attributed to support from the VA's help desk (facilitating conditions) that was used by 84% (21/25) of patients. However, challenges with Bluetooth pairing and data syncing emerged as key concerns. These technical hurdles of usability and interoperability have been identified in previous studies as limiting factors for the effective adoption of peripheral devices [9,12]. Our findings provide new insight into how PGHD is used in real-world settings where technical integration is incomplete and suggest that PGHD can still retain clinical and relational value despite technical challenges. We also found that care team and social connections shape a patient's engagement with technology, including relationships with the care team, caregivers who assisted in device use, and fellow veterans who promoted awareness of the program (social influence), in accordance with existing work [34-36].

Peripheral Devices and PGHD Play a Role in Reducing Barriers to Care

Patients described how peripheral devices and tablets mitigated logistical barriers to in-person care. Patients with physical disabilities, mental health concerns, or long travel distances emphasized how home monitoring tools reduced the burden of attending in-person medical visits. These experiences reinforce the potential of virtual care to reduce care fragmentation and expand access to care for patients who face challenges in receiving in-person care. Patients expressed that avoiding stressful travel to appointments improved their ability to focus during visits. This is particularly important for patients with conditions that may impact driving or navigating public transportation, such as anxiety, limited mobility, or visual impairments. Our qualitative findings build upon our team's earlier quantitative evaluations of the VA's tablet distribution program, which showed reduced emergency department visits, improved continuity of mental health care, and demonstrated how tablets are reaching veterans from multiple high-need groups [25,26,37]. They also speak to how remote care and PGHD can foster more equitable access to care [13,38].

Patients also discussed using PGHD to spark conversations with their clinicians. In contrast to previous work documenting patient concerns with peripheral devices leading to less face-to-face time with care teams [21], patients in this study described how using their data helped facilitate more frequent and productive communication with care teams. Patients also noted improved understanding of data presented in the My Health*e*Vet patient portal, which highlights the complementary value of access to different digital tools [19,20]. These findings also touch upon considerations for integrating more technical and health literacy education into remote monitoring programs, as patients expressed interest in more in-depth training on device measurement techniques.

Opportunities for System Integration

An emergent theme in our study was the desire for an integrated platform that could aggregate and visualize data from different devices over time to facilitate self-management and health insights. Although the key focus of this telehealth initiative was for patients to use the peripheral device during a video visit, patients desired additional ways to use their data in meaningful ways. For example, patients expressed interest in tracking trends across multiple health domains (eg, blood pressure, glucose, and weight), which could enhance their understanding of personal health trajectories. These suggestions reflect a broader push in the informatics community toward integrated personal health records and highlight the importance of designing tools that offer patients a comprehensive view of their health [6,15].

Recently, the VA has prioritized integration of PGHD into research and clinical efforts, with the development of an enterprise-wide PGHD database that incorporates third-party apps, wearables, medical devices, and patient portal data [11]. Taking an interoperable approach, PGHD are standardized using Fast Healthcare Interoperability Resources to ensure data consistency across the different data sources so that it can be effectively linked with VA data. Robust system integration of heterogenous data streams can also set the foundation for future data-driven tools and algorithms that can improve patient care.

Limitations

These findings should be interpreted within the context of several limitations. First, our sample was purposively selected to include diverse perspectives (eg, age, race, gender, geography, and peripheral device or tablet use) but may not be representative of all VA patients who received peripheral devices. As the majority of patients used their peripheral devices (23/25, 92%) or tablet (19/25, 76%), our findings may overestimate positive experiences, including effort expectancy and ease of use, because nonusers were underrepresented. Study participants, and tablet and peripheral device recipients in general, may also have been more engaged, technologically inclined, or positively disposed toward digital tools than the overall VA patient population. Second, although we recruited patients who received a device within the previous 6 months, perspectives on health improvements may be subject to recall biases, which could impact the accuracy of self-reported improvements. Third, our interview guide focused on the core constructs of UTAUT and did not include topics on privacy, trust, or confidentiality of data generated from the peripheral devices. However, privacy concerns and perceived risks may also influence behavioral intention and use of technology [39,40]. Fourth, we did not conduct member checking to validate patient narratives or themes, but during the interviews, patients were asked to clarify responses as needed to ensure their perspectives were understood and properly documented. The insights generated from this work offer valuable real-world perspectives that can inform the ongoing design, implementation, and expansion of remote monitoring programs and PGHD use across large, national health systems.

Conclusion

As the use of peripheral devices and PGHD continues to grow, input from patients about their experiences can maximize the benefit of these tools for patients and health systems. Our findings suggest that peripheral devices can enhance health monitoring, reduce care barriers, and foster patient engagement, especially when accompanied by other virtual care tools and responsive technical support. These findings provide evidence to support continued investment in health system–wide connected care initiatives, with emphasis on device

XSL•FO
RenderX

interoperability, education on device use, and awareness to ensure equitable access.

Acknowledgments

We thank the veterans who shared their experiences and participated in this project. We also appreciate the analytic contributions of Liberty Green (MS), James Van Campen (MS), and Cindie Slightam (MPH). We also thank Emily Drucker (LCSW) and Nilesh Shah (MD) for their input on the manuscript.

Funding

This work was supported by a Stanford Medicine Spectrum Population Health Sciences pilot grant (principal investigator ACG) and a grant from the US Department of Veterans Affairs Health Systems Research (CDA 23-145, principal investigator ACG). The views expressed are those of the authors and do not represent those of the US Department of Veterans Affairs or those of the US government.

Data Availability

The qualitative data generated during this project are not publicly available to protect the privacy of individuals who participated. Deidentified data are available from the corresponding author upon reasonable request, subject to institutional or regulatory approval.

Authors' Contributions

All authors contributed to the conceptualization, design, and methodology. ACG and CG conducted the qualitative analysis. ACG was primarily responsible for the initial draft of the manuscript, with input from all authors. All authors participated in manuscript writing, revision, and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Mobile fact sheet. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2025-07-31]
2. Health information national trends survey 7. National Cancer Institute. 2024. URL: https://hints.cancer.gov/view-questions/all-hints-questions.aspx [accessed 2025-11-24]
3. 21st century cures act: Interoperability, information blocking, and the ONC health IT certification program. Health and Human Services Department. 2020. URL: https://www.federalregister.gov/documents/2020/05/01/2020-07419/21st-century-cures-act-interoperability-information-blocking-and-the-onc-health-it-certification [accessed 2025-07-31]
4. Interoperability and patient access final rule. Centers for Medicare & Medicaid Services. 2020. URL: https://www.cms.gov/priorities/burden-reduction/overview/interoperability/policies-and-regulations/cms-interoperability-and-patient-access-final-rule-cms-9115-f [accessed 2025-07-31]
5. 2024 State of FHIR survey results. Health Level Seven International and Firely. URL: https://www.hl7.org/documentcenter/public/white-papers/2024%20StateofFHIRSurveyResults_final.pdf(31 [accessed 2025-07-31]
6. Enabling health interoperability through FHIR. Health Level Seven International. 2025. URL: http://fhir.org/ [accessed 2025-07-31]
7. Patient-generated health data. Assistant Secretary for Technology Policy. 2025. URL: https://www.healthit.gov/topic/scientific-initiatives/patient-generated-health-data [accessed 2025-07-31]
8. Austin E, Lee JR, Amtmann D, Bloch R, Lawrence SO, McCall D, et al. Use of patient-generated health data across healthcare settings: implications for health systems. JAMIA Open 2020;3(1):70-76 [FREE Full text] [doi: 10.1093/jamiaopen/ooz065] [Medline: 32607489]
9. Tiase VL, Hull W, McFarland MM, Sward KA, Del Fiol G, Staes C, et al. Patient-generated health data and electronic health record integration: a scoping review. JAMIA Open 2020;3(4):619-627 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa052] [Medline: 33758798]
10. Foster C, Schinasi D, Kan K, Macy M, Wheeler D, Curfman A. Remote monitoring of patient- and family-generated health data in pediatrics. Pediatrics 2022;149(2):e2021054137 [FREE Full text] [doi: 10.1542/peds.2021-054137] [Medline: 35102417]
11. Newton TJ, Shah N, Lewis K, Zocchi MS, Bixler FR, Etingen B, et al. Developing infrastructure to realize the value of patient-generated health data in a large integrated health care system: the Veterans Health Administration experience. J Med Internet Res 2025;27:e70755 [FREE Full text] [doi: 10.2196/70755] [Medline: 40479588]

XSL•FO
RenderX

12. Grossman LV, Masterson Creber RM, Benda NC, Wright D, Vawdrey DK, Ancker JS. Interventions to increase patient portal use in vulnerable populations: a systematic review. J Am Med Inform Assoc 2019;26(8-9):855-870 [FREE Full text] [doi: 10.1093/jamia/ocz023] [Medline: 30958532]
13. Estacio EV, Whittle R, Protheroe J. The digital divide: Examining socio-demographic factors associated with health literacy, access and use of internet to seek health information. J Health Psychol 2019;24(12):1668-1675. [doi: 10.1177/1359105317695429] [Medline: 28810415]
14. Clare CA. Telehealth and the digital divide as a social determinant of health during the COVID-19 pandemic. Netw Model Anal Health Inform Bioinform 2021;10(1):26 [FREE Full text] [doi: 10.1007/s13721-021-00300-y] [Medline: 33842187]
15. Treadwell JR, Reston JT, Rouse B. Automated-Entry Patient-Generated Health Data for Chronic Conditions: The Evidence on Health Outcomes. Rockville (MD): Agency for Healthcare Research and Quality; 2021.
16. Gresham G, Hendifar AE, Spiegel B, Neeman E, Tuli R, Rimel BJ, et al. Wearable activity monitors to assess performance status and predict clinical outcomes in advanced cancer patients. NPJ Digit Med 2018;1:27 [FREE Full text] [doi: 10.1038/s41746-018-0032-6] [Medline: 31304309]
17. Zulman DM, Wong EP, Slightam C, Gregory A, Jacobs JC, Kimerling R, et al. Making connections: nationwide implementation of video telehealth tablets to address access barriers in veterans. JAMIA Open 2019;2(3):323-329 [FREE Full text] [doi: 10.1093/jamiaopen/ooz024] [Medline: 32766533]
18. VA app store. U.S. Department of Veterans Affairs. 2025. URL: https://mobile.va.gov/appstore [accessed 2025-07-31]
19. Griffin AC, Troszak LK, Van Campen J, Midboe AM, Zulman DM. Tablet distribution to veterans: an opportunity to increase patient portal adoption and use. J Am Med Inform Assoc 2022;30(1):73-82 [FREE Full text] [doi: 10.1093/jamia/ocac195] [Medline: 36269168]
20. Griffin AC, Greene L, Slightam C, Wray CM, Wielgosz J, Kuhn E, et al. Use of mobile applications and mental health services among veterans who received internet-connected tablets. JAMIA Open 2024;7(4):ooae119. [doi: 10.1093/jamiaopen/ooae119] [Medline: 39534216]
21. Walker RC, Tong A, Howard K, Palmer SC. Patient expectations and experiences of remote monitoring for chronic diseases: systematic review and thematic synthesis of qualitative studies. Int J Med Inform 2019;124:78-85 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.01.013] [Medline: 30784430]
22. Paul MM, Khera N, Elugunti PR, Ruff KC, Hommos MS, Thomas LF, et al. The state of remote patient monitoring for chronic disease management in the United States. J Med Internet Res 2025;27:e70422 [FREE Full text] [doi: 10.2196/70422] [Medline: 40570326]
23. Alvarez P, Sianis A, Brown J, Ali A, Briasoulis A. Chronic disease management in heart failure: focus on telemedicine and remote monitoring. Rev Cardiovasc Med 2021;22(2):403-413 [FREE Full text] [doi: 10.31083/j.rcm2202046] [Medline: 34258907]
24. Hirko K, Kerver J, Ford S, Szafranski C, Beckett J, Kitchen C, et al. Telehealth in response to the COVID-19 pandemic: implications for rural health disparities. J Am Med Inform Assoc 2020;27(11):1816-1818 [FREE Full text] [doi: 10.1093/jamia/ocaa156] [Medline: 32589735]
25. Jacobs JC, Blonigen DM, Kimerling R, Slightam C, Gregory AJ, Gurmessa T, et al. Increasing mental health care access, continuity, and efficiency for veterans through telehealth with video tablets. Psychiatr Serv 2019;70(11):976-982. [doi: 10.1176/appi.ps.201900104] [Medline: 31378193]
26. Gujral K, Van Campen J, Jacobs J, Kimerling R, Blonigen D, Zulman DM. Mental health service use, suicide behavior, and emergency department visits among rural US veterans who received video-enabled tablets during the COVID-19 pandemic. JAMA Netw Open 2022;5(4):e226250 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.6250] [Medline: 35385088]
27. Bridging the digital divide. U.S. Department of Veterans Affairs. 2025. URL: https://telehealth.va.gov/digital-divide [accessed 2025-07-31]
28. ZIP code RUCA approximation. Rural Health Research Center. URL: https://depts.washington.edu/uwruca/ruca-approx.php#:~:text=Because%20the%20smallest%20geographic%20identifiers,the%20ZIP%20code%20geographic%20unit [accessed 2025-07-31]
29. Yoon J, Chee CP, Su P, Almenoff P, Zulman DM, Wagner TH. Persistence of high health care costs among VA patients. Health Serv Res 2018;53(5):3898-3916 [FREE Full text] [doi: 10.1111/1475-6773.12989] [Medline: 29862504]
30. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: Toward a unified view. MIS Quarterly 2003;27(3):425-478. [doi: 10.5555/2017197.2017202]
31. Kowalski CP, Nevedal AL, Finley EP, Young JP, Lewinski AA, Midboe AM, et al. Planning for and assessing rigor in rapid qualitative analysis (PARRQA): a consensus-based framework for designing, conducting, and reporting. Implement Sci 2024;19(1):71 [FREE Full text] [doi: 10.1186/s13012-024-01397-1] [Medline: 39394597]
32. Nouri S, Khoong EC, Lyles CR. Addressing equity in telemedicine for chronic disease management during the COVID-19 pandemic. NEJM Catalyst Innovations in Care Delivery 2020;1(3). [doi: 10.1056/CAT.20.0123]
33. Conceptualizing a data infrastructure for the capture, use, and sharing of patient-generated health data in care delivery and research through 2024. The Office of the National Coordinator for Health Information Technology. 2018. URL: https://www.healthit.gov/sites/default/files/onc_pghd_final_white_paper.pdf [accessed 2025-07-31]

XSL•FO
RenderX

34. Fortuna KL, Brooks JM, Umucu E, Walker R, Chow PI. Peer support: a human factor to enhance engagement in digital health behavior change interventions. J Technol Behav Sci 2019;4(2):152-161 [FREE Full text] [doi: 10.1007/s41347-019-00105-x] [Medline: 34337145]
35. Peng R, Chang J, Zhu Y, Feng H, Cao Z. Unlocking the power of peer support in digital use and digital health interventions for older adults: a scoping review. J Adv Nurs 2026;82(1):162-173. [doi: 10.1111/jan.16963] [Medline: 40195282]
36. Tahsin F, Armas A, Kirakalaprathapan A, Kadu M, Sritharan J, Steele Gray C. Information and communications technologies enabling integrated primary care for patients with complex care needs: scoping review. J Med Internet Res 2023;25:e44035 [FREE Full text] [doi: 10.2196/44035] [Medline: 37074779]
37. Ferguson JM, Van Campen J, Slightam C, Greene L, Heyworth L, Zulman DM. Evaluation of the veterans health administration's digital divide consult for tablet distribution and telehealth adoption: cohort study. J Med Internet Res 2024;26:e59089 [FREE Full text] [doi: 10.2196/59089] [Medline: 39250183]
38. Campos-Castillo C, Anthony D. Racial and ethnic differences in self-reported telehealth use during the COVID-19 pandemic: a secondary analysis of a US survey of internet users from late March. J Am Med Inform Assoc 2021;28(1):119-125 [FREE Full text] [doi: 10.1093/jamia/ocaa221] [Medline: 32894772]
39. Liu Y, Lu X, Zhao G, Li C, Shi J. Adoption of mobile health services using the unified theory of acceptance and use of technology model: self-efficacy and privacy concerns. Front Psychol 2022;13:944976 [FREE Full text] [doi: 10.3389/fpsyg.2022.944976] [Medline: 36033004]
40. Su J, Wang Y, Liu H, Zhang Z, Wang Z, Li Z. Investigating the factors influencing users' adoption of artificial intelligence health assistants based on an extended UTAUT model. Sci Rep 2025;15(1):18215 [FREE Full text] [doi: 10.1038/s41598-025-01897-0] [Medline: 40414992]

Abbreviations

CDW: Corporate Data Warehouse
EHR: electronic health record
PGHD: patient-generated health data
UTAUT: Unified Theory of Acceptance and Use of Technology
VA: US Department of Veterans Affairs

Edited by A Coristine, A Stone; submitted 24.Nov.2025; peer-reviewed by S A, M Lennon; comments to author 22.Dec.2025; revised version received 09.Mar.2026; accepted 09.Mar.2026; published 06.Apr.2026.

Please cite as:
Griffin AC, Gray C, Ngo V, Chaudhary C, Shimada SL, Midboe AM, Zulman DM
Supporting Access to Care Through Peripheral Devices and Patient-Generated Health Data: Qualitative Study
J Med Internet Res 2026;28:e88405
URL: https://www.jmir.org/2026/1/e88405
doi:10.2196/88405
PMID:

©Ashley C Griffin, Caroline Gray, Victoria Ngo, Camila Chaudhary, Stephanie L Shimada, Amanda M Midboe, Donna M Zulman. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Telehealth Delivery of the Homeostasis–Enrichment–Plasticity Approach for Premature Infants With Developmental Risks: Exploratory Feasibility Study

Aymen Balikci[1], PhD; Gamze Cagla Sirma[2], MSc; Izgi-Miray Demirbag[1], MSc; Teresa A May-Benson[3], ScD; Hatice Gulhan Sozen[4], MD; Ayse Firdevs Aracikul Balikci[5], PhD; Gul Ilbay[6], Prof Dr

[1]Sense On Ltd, Istanbul, Turkey
[2]Department of Occupational Therapy, Faculty of Health Sciences, Fenerbahçe University, Istanbul, Istanbul, Turkey
[3]TMB Education, Norristown, PA, United States
[4]Department of Child Health and Diseases, Faculty of Medicine, Bahcesehir University, Istanbul, Turkey
[5]Department of Special Education, Faculty of Education, Istanbul Medipol University, Istanbul, Turkey
[6]Department of Physiology, Faculty of Medicine, Kocaeli University, Kocaeli, Turkey

Corresponding Author:
Gamze Cagla Sirma, MSc
Department of Occupational Therapy
Faculty of Health Sciences
Fenerbahçe University
Atatuk Mah Metropol Istanbul, Atasehir Blv.
Istanbul, Istanbul, 34758
Turkey
Phone: 90 2169101907
Email: cagladirgen@hotmail.com

Abstract

Background: Preterm delivery is an increasing worldwide health concern linked to increased neurodevelopmental risks. Early intervention is crucial for harnessing neuroplasticity to enhance developmental and functional performance outcomes; however, access to early intervention is frequently hindered by logistical, financial, and labor constraints. The Homeostasis–Enrichment–Plasticity (HEP) Approach is a family-centered early intervention model based on enriched environments, designed to improve infants' sensory-motor, cognitive, and socio-emotional development.

Objective: This study aimed to assess the feasibility, safety, acceptability, and outcomes sensitivity to change of implementing the HEP Approach through telehealth for premature infants at developmental risk.

Methods: A pre-post exploratory feasibility study was performed, including 16 preterm infants (aged 4-12 months corrected age), of whom 14 completed the study. The 12-week intervention included weekly remote sessions focused on environmental enrichment, active exploration, and parental guidance. The feasibility and acceptability were evaluated using a 24-item questionnaire. Developmental outcomes were assessed with the Young Children's Participation and Environment Measure, Ages and Stages Questionnaire (ASQ), Alberta Infant Motor Scale, Infant Motor Profile, and Depression Anxiety Stress Scales.

Results: High adherence (14/14, 100%) and retention (14/16, 87.5%) rates demonstrated robust feasibility. Parents indicated 86%-100% agreement across all feasible criteria, affirming safety, satisfaction, and acceptability. No adverse incidents were reported. Changes were identified in participation (Young Children's Participation and Environment Measure), motor development (Alberta Infant Motor Scale, Infant Motor Profile, and ASQ), communication and social-emotional domains (ASQ), and caregiver well-being (Depression Anxiety Stress Scales) ($P<.05$).

Conclusions: The telehealth implementation of the HEP Approach demonstrated feasibility, safety, and strong acceptance among families, along with quantifiable developmental and psychosocial changes. These initial findings endorse the model's viability as an accessible, family-oriented telehealth framework for infants born preterm. Future randomized controlled and longitudinal studies are necessary to validate intervention efficacy and scalability.

(J Med Internet Res 2026;28:e86883) doi:10.2196/86883

XSL•FO
RenderX

KEYWORDS

early intervention; environmental enrichment; premature infants; telehealth; adherence; retention

Introduction

Approximately 1 in every 10 children worldwide (9.9% of all births) is born preterm (<37 weeks of gestation) [1]. Preterm delivery is associated with substantial long-term health consequences, spanning multiple domains such as respiratory and cardiovascular health as well as neurodevelopmental concerns. Within the neurodevelopmental domain, impairments may range from severe conditions, such as cerebral palsy—particularly among those born at the lowest gestational ages—to less severe but still clinically significant outcomes [1-3]. Recent evidence indicates that even births occurring just a few weeks early increase the risk of adverse developmental outcomes [1,4,5]. This is particularly relevant as approximately 85% of preterm births occur between 32 and 37 weeks of gestation, meaning that a large proportion of preterm infants are at risk, challenging the earlier assumption that only extremely preterm infants face vulnerability [1,4,6]. Moreover, numerous studies have demonstrated that, compared with their full-term-born peers, preterm infants are at heightened risk for difficulties across multiple developmental domains, including cognitive [7-9], motor [7,10-12], sensory [11-13], language [11,14,15], and social-emotional functioning [5,15,16].

Overall, these data emphasize the importance of early developmental support for infants born preterm [17,18]. Given the significant neuroplasticity in early life [19,20], early intervention (EI) can influence multiple domains related to prematurity [17]. Evidence indicates that structured services during infancy improve cognitive, motor, sensory, language, and social-emotional outcomes, while also fostering long-term neurodevelopmental pathways and functional engagement [12,21-24]. Consequently, EI may influence neurodevelopmental trajectories, reducing risks associated with prematurity and enhancing the quality of life for infants and their families [12,17,18,24-26].

Although EI provides substantial benefits for the development of preterm infants, many families face persistent challenges in accessing these services. The most prominent barriers include shortages of qualified professionals, transportation difficulties, and the difficulty of organizing time to attend services—factors that are especially evident for families living in rural or underserved areas. While systemic issues such as delayed referrals or fragmented communication may also play a role, it is these fundamental constraints of workforce availability, accessibility, and time that most directly limit timely participation in EI and reduce its potential impact during critical developmental windows [27-30].

With advances in digital technologies, telehealth has become a promising alternative for addressing barriers to EI in preterm infants [31,32]. Studies during the COVID-19 pandemic showed that telemedicine maintained developmental follow-up and medical care when in-person visits were restricted, mitigating workforce and logistical challenges [33]. Telehealth services are valuable for infants discharged from the neonatal intensive care unit, as they decrease needless hospital visits and limit exposure to transmissible infections [31,32,34]. Furthermore, studies demonstrate that telehealth can be as efficacious as in-person care for many health conditions and interventions [31,35,36] and may offer a cost-effective health care delivery paradigm [31,32,37,38].

Building on the growing evidence for telehealth as a viable and cost-effective model of early intervention, the Homeostasis–Enrichment–Plasticity (HEP) Approach [39] represents a structured framework that can be adapted to remote delivery while preserving its core principles of parent coaching and enriched developmental support. The HEP Approach, developed by Balikci [39], is a comprehensive EI model designed to promote sensory-motor, cognitive, and social-emotional development in infants. The HEP Approach incorporates 10 fundamental principles—homeostasis, safety, multisensory experiences, novelty, sustained engagement, and active exploration—based on environmental enrichment paradigms, through a clinic-based program that includes parent coaching and concurrent home implementation [24,39,40]. The approach provides infants with enhanced sensory and motor experiences in the clinic, while parents obtain systematic coaching that is later integrated into everyday routines and natural settings, with telehealth follow-up reinforcing these procedures at home. Rooted in frameworks including Ecological Perception Theory, Perception–Action Theory, Sensory Integration, the Person–Environment–Occupation (PEO) Model, and Dynamic Systems Theory, the HEP Approach equips caregivers with adaptable, goal-focused strategies that enhance co-regulation, sensory-motor learning, and adaptive participation [24,39,40].

While previous research has demonstrated the feasibility and effectiveness of the HEP Approach in the traditional clinic-based/home-monitored format [24,40], its applicability through telehealth has not yet been systematically examined. This study represents the first attempt to evaluate the feasibility of delivering the HEP Approach via telehealth.

The aim of this study is to assess the feasibility of implementing the HEP Approach in a telehealth format and to determine whether this model provides an accessible, acceptable, and practical option for supporting the developmental outcomes of premature infants and their families.

Methods

Study Design

This is a pre-post feasibility experimental study. The study was conducted in Istanbul, Türkiye, with interventions delivered remotely via telehealth between January 2025 and July 2025.

Ethical Considerations

The study was conducted in accordance with the Declaration of Helsinki and approved by the Clinical Research Ethics Committee of Istanbul Medipol University

XSL•FO
RenderX

(E-10840098-202.3.02-7368). All caregivers provided written informed consent before participation and were informed of their right to withdraw from the study at any time. Participant privacy and confidentiality were strictly maintained, and all data were anonymized prior to analysis. No identifiable personal information was included in the manuscript. No financial or material compensation was provided to the participants.

Study Procedures

The study comprised 11 structured phases that guided the implementation of the HEP Approach via telehealth [40]. Phase 1 covered participant recruitment via physician referral and digital record review, while Phase 2 introduced families to the HEP philosophy and telehealth process. Phase 3 of the study comprised baseline evaluations completed exclusively remotely via Zoom. The evaluation process comprised caregiver interviews, unstructured observations, and structured observations tailored to the requirements of standardized assessment instruments. All evaluations, including the implementation of standardized assessments, were conducted remotely via videoconferencing (Zoom). In Phases 4-5, assessment findings were collaboratively interpreted with families, and hypotheses regarding sensory-motor, emotional, social, and environmental factors were formulated and discussed. Phase 6 focused on interactive goal setting, followed in Phases 7-8 by individualized intervention strategies delivered through weekly telehealth coaching, demonstrations, and integration into daily family routines, supported by video sharing via WhatsApp. Phase 9 emphasized adaptation and generalization of strategies, while Phase 10 ensured continuous monitoring with therapist feedback and reflective questioning to enhance family problem-solving. Finally, Phase 11 included postintervention assessments, feasibility questionnaires, and feedback interviews conducted via videoconferencing. A comprehensive description of each phase is provided in Table S1 in Multimedia Appendix 1.

Fidelity to the intervention was maintained through multiple strategies. Treating therapists completed 120 hours of theoretical and practical training, including supervision, provided by the first author, who developed the model. In addition, fidelity was reinforced through regular mentoring and the first author's systematic review of recorded telehealth sessions. All procedures were implemented in accordance with the HEP Approach Implementation Manual [22], and session recordings were evaluated using the fidelity checklists included in the manual to ensure high implementation reliability.

Participants and Recruitment

Convenience sampling, a nonprobability method, was used [40]. Neonatologists from 3 hospitals in Istanbul referred premature infants who met the designated inclusion criteria. The inclusion criteria were (1) corrected age ranging from 4 to 12 months, (2) birth at ≤36 weeks and 6 days of gestation, (3) absence of systemic diseases or congenital defects, and (4) familial consent to participate consistently in the study process. The exclusion criteria included the following conditions: (1) substantial impairment of vision or hearing; (2) history of febrile seizures; (3) medical conditions that impede active participation in the study, such as dependence on oxygen; and (4) prior participation in other experimental rehabilitation studies. Following assessment of the qualifying criteria, informed consent was secured.

A total of 39 preterm infants were assessed for eligibility. Eighteen did not meet the inclusion criteria, and five families declined to participate, resulting in 16 infants enrolled in the HEP telehealth intervention. All 16 received the intervention. During follow-up, 2 infants discontinued participation due to illness, leaving 14 who completed the program and were included in the final analysis. All pre- and postintervention evaluations were conducted by a physical therapist with over 5 years of pediatric experience and an advanced Master of Science degree. This therapist had received formal training in the administration and scoring of outcome measures as part of their graduate program and was blinded to the intervention. The HEP Approach itself was delivered by 2 physiotherapists, each with more than five years of pediatric clinical experience and advanced Master's-level training. The participant flow is illustrated in Figure 1.

XSL•FO
RenderX

Figure 1. Flowchart of participant and outcome assessment through the trial. HEP: Homeostasis–Enrichment–Plasticity.

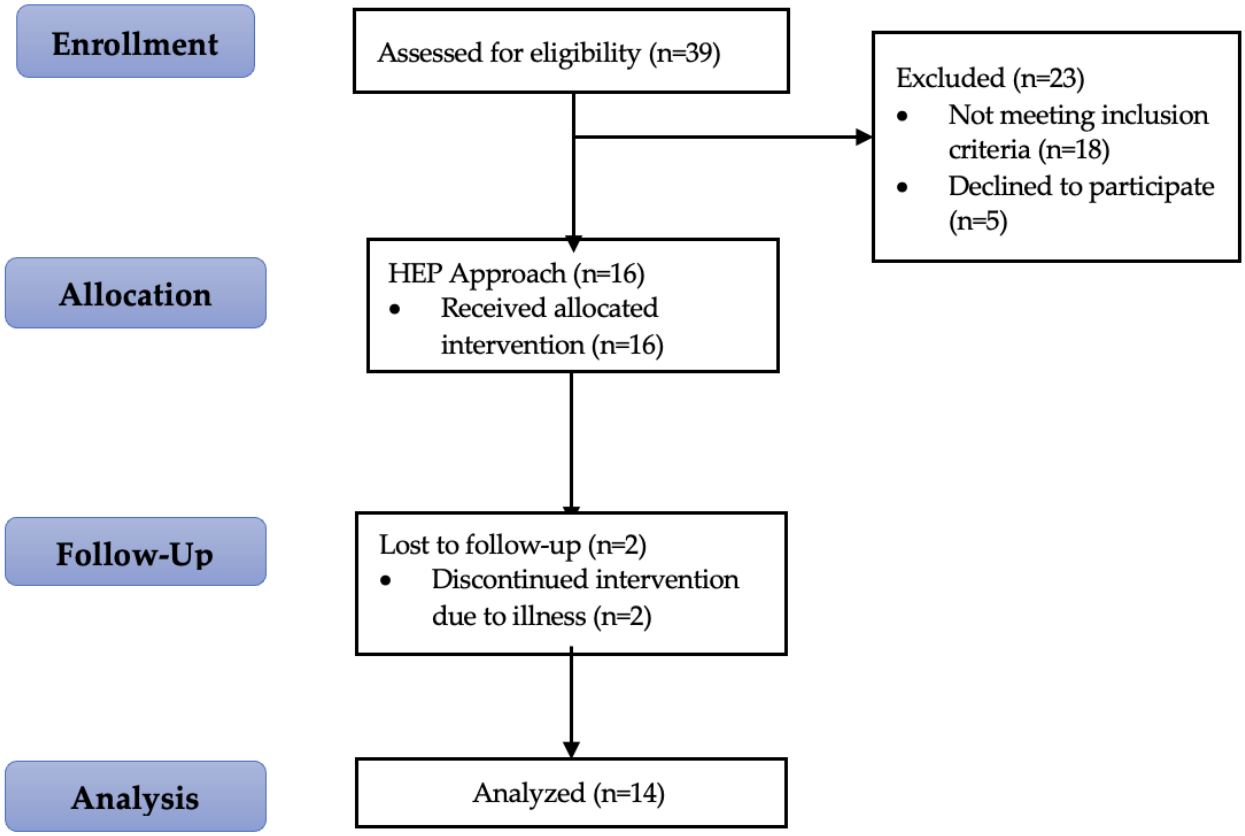

Feasibility Measure

Feasibility and acceptability of the telehealth-based HEP Approach were evaluated using a 24-item Likert-type questionnaire, adapted from the feasibility study by Balikci et al [40]. The survey was structured to assess four domains: feasibility (items 1-7), acceptability (items 8-12), satisfaction (items 13-19), and safety (items 20-24). Each item was rated on a five-point scale ranging from strongly agree to strongly disagree. For analytic purposes, responses of “Strongly Agree” and “Agree” were categorized as agreement, whereas all other responses were classified as disagreement. Alongside the structured items, 3 open-ended questions were incorporated at the conclusion of the form, inviting parents to provide their responses. The qualitative data were subsequently examined and thematically grouped according to the 3 questions. Feedback on the intervention’s feasibility, safety, and perceived effectiveness was also collected from both therapists through postintervention reflections and monitoring logs maintained throughout the study period.

Outcome Measures

Caregivers first completed a demographic questionnaire capturing infant and maternal health, gestational and postnatal history, corrected age at enrollment, gender, birth weight, gestational age, maternal age, family income, and caregiver educational level.

Five standardized outcome measures were used to evaluate developmental progress across multiple domains. The Turkish translated version of the Young Children’s Participation and Environment Measure (YC-PEM) was used to assess participation frequency, involvement, desire for change, and environmental support in both home and community contexts [41-43]. Broader developmental domains, including social-emotional, communication, gross motor, fine motor, problem-solving, and personal-social skills, were measured with the Turkish translated version of the Ages and Stages Questionnaire (ASQ) [44]. Motor development was further assessed using the Alberta Infant Motor Scale (AIMS) [45-49] and the Infant Motor Profile (IMP) [50-52], which captured dimensions such as variation, adaptability, symmetry, fluency, and performance. Finally, parental mental health and well-being were evaluated with the Depression Anxiety Stress Scales (DASS-21) [53]. The selection of outcome measures was guided by the study’s feasibility objectives rather than by an intention to establish intervention efficacy. AIMS and IMP were selected as standardized, performance-based tools reported in remote or telehealth assessment contexts and among the limited motor assessments adaptable to video-based administration. In this study, these instruments were used to evaluate (1) whether they could be feasibly administered in a telehealth format within our sample and (2) whether they were capable of detecting measurable change over a 12-week period, regardless of whether such change reflected intervention-related or developmental factors.

Caregiver-reported measures (YC-PEM, ASQ, and DASS-21) were included to capture domains not directly assessable via remote standardized performance testing and to examine the feasibility of administering structured family-report instruments within a telehealth early intervention framework.

XSL•FO
RenderX

The Young Children's Participation and Environment Measure-Turkish Version

The Turkish translation of the YC-PEM was used in this study. The YC-PEM is a caregiver-completed measure designed to assess children's participation up to 5 years of age across home, daycare/preschool, and community settings. It evaluates four domains: frequency of participation, level of involvement, parental desire for change, and environmental supports and barriers. Previous studies demonstrated good internal consistency (α=0.67-0.96), fair-to-good test-retest reliability (ICC=0.59-0.94) in home and community contexts, and adequate construct validity in differentiating participation across age and disability groups [41,42]. The Turkish translation and cross-cultural adaptation of the YC-PEM were conducted following standardized procedures, and subsequent psychometric evaluation in a large sample of 367 children, both with and without disabilities, demonstrated good reliability (Cronbach α=0.68-0.94; ICC=0.69-0.89) together with evidence of discriminant and concurrent validity, thereby supporting the YC-PEM-T as a reliable and culturally appropriate instrument for assessing children's participation and environmental influences in Türkiye [43]. The YC-PEM was provided to families in electronic (PDF) format. Caregivers independently completed the measure at their convenience and returned the completed forms electronically.

The Ages and Stages Questionnaire-Turkish Version

The Turkish version of the Ages and Stages Questionnaire –Turkish Version (ASQ-TR) was used. The ASQ is a parent-completed developmental screening instrument designed for children aged 4 to 60 months (extended up to 72 months in some adaptations). Each form consists of age-specific items that evaluate five developmental domains: communication, gross motor, fine motor, problem-solving, and personal-social skills. Parents complete the questionnaires in approximately 10-15 minutes, based on their observations and knowledge of their children's performance, and respond to items with options such as "yes," "sometimes," or "not yet." Responses are scored to determine whether the child's development is on schedule, requires monitoring, or needs referral for further evaluation.

The original ASQ demonstrated good psychometric qualities, with reported sensitivity values typically ranging from 75% to 90% and specificity from 77% to 88%, alongside strong reliability indices and evidence of construct validity against standardized developmental assessments. The Turkish version (ASQ-TR), evaluated in a large sample of 833 children aged 3-72 months, demonstrated satisfactory reliability and validity. Concurrent validity analyses showed overall agreement rates of 70.95% when one domain was below the cutoff and 88.24% when 2 domains were below the cutoff, supporting the 2-domain criterion as a more accurate classification approach. Internal consistency coefficients (Cronbach α) ranged from 0.38 to 0.95 across domains and age forms, generally increasing with age. Domain-total correlations (r=0.54-0.96) and interdomain correlations (r=0.65-0.79) were all statistically significant. Test-retest reliability over a 2-week interval indicated high agreement across domains (ranging from 0.82 to 0.97), and interrater reliability between mothers and teachers had agreement levels between 73.68% and 87.18%, depending on the criterion used.

The use of a 2-domain cutoff criterion considerably improved diagnostic accuracy in this version of the measure, increasing specificity to 85% and positive predictive value to 75%, while maintaining high sensitivity (94%). These findings support the ASQ-TR as a reliable and culturally appropriate developmental screening tool for use in Türkiye. Families received the ASQ-TR in PDF format; caregivers completed the measure at their convenience and submitted the forms electronically.

The Alberta Infant Motor Scale

The AIMS is a standardized observational tool developed to identify motor delays and describe infants' gross motor performance. It evaluates spontaneous movements, posture, weight shifts, and antigravity control over a 20-30 minute assessment, often supported by age-appropriate toys to encourage activity. The scale includes 58 items across four positions (21 prone, 9 supine, 12 sitting, 16 standing), scored dichotomously and summarized on age-specific percentile charts ranging from the 5th to the 90th percentile. Based on these scores, infants' motor performance can be classified as delayed, suspect, or normal [45,46]. Due to its low cost, ease of administration, reproducibility, and minimal handling of the infant, AIMS is widely recognized as a practical measure in both clinical and community health contexts, and has been validated as reliable for high-risk populations The AIMS demonstrated strong concurrent validity with the Bayley Scales of Infant Development (r=0.95; P<.01) and high interrater reliability across all tested ages (intraclass correlation coefficients ranging from 0.76 to 0.99), supporting it as a valid and reliable tool for assessing motor development in high-risk infants [47]. More recently, studies have demonstrated that the AIMS can be successfully applied in telehealth settings, with strong agreement between remote and face-to-face assessments for neurodevelopmentally high-risk infants [48,49]. In this study, the AIMS was administered remotely via videoconferencing (Zoom). The evaluating specialist provided real-time verbal guidance to caregivers on how to position their child for each assessment item and scored the child's performance based on live observations. All sessions were video-recorded, enabling the evaluator to revisit the recordings when needed to verify or adjust scores. This administration procedure aligned with protocols reported in prior research [48,49].

The Infant Motor Profile

The IMP is a video-based observational tool designed to evaluate motor behavior in infants aged 3 to 18 months, or until a few months after independent walking is achieved. It consists of 80 items across five domains—movement variation, adaptability, symmetry, fluency, and performance—capturing both quantitative and qualitative aspects of motor development. Grounded in the Neuronal Group Selection Theory, the IMP emphasizes the importance of motor variability and adaptive strategy selection as indicators of central nervous system integrity [50].

The IMP has demonstrated strong psychometric properties, including acceptable interrater reliability in high-risk infants,

XSL•FO
RenderX

with Spearman correlation coefficients ranging from 0.47 to 0.94 across domains; reliability was highest for the performance domain (r=0.90-0.93) and lowest for symmetry (r=0.47-0.74), supporting its use as a reliable tool for assessing early motor development and identifying physiotherapy needs in the Turkish sample [51]. Moreover, the IMP has shown excellent concurrent validity with the AIMS (ρ=0.76; P<.001), a significant association with the General Movement Assessment, and high diagnostic accuracy in predicting neurodevelopmental disorders at 5 months (AUC=0.92), with strong sensitivity (93%) and specificity (81%), confirming its validity and clinical utility for evaluating early motor behavior and reflecting the severity of brain injury in at-risk infants. [52]. Additionally, studies indicate that the IMP is responsive to change, making it a valuable outcome measure in early intervention research [52]. Collectively, these findings support the IMP as a reliable and valid tool for assessing motor development and detecting early deviations in infants at risk of neurodevelopmental disorders. The IMP was administered by the examiner, who provided structured instructions for various positions and object manipulations during the initial session; the examiner recorded the session and took detailed notes while guiding the caregiver through the tasks, and later used the video recording to score and analyze the child's performance.

Depression Anxiety Stress Scales-21-Turkish Version

The DASS-21 is a self-report instrument designed to assess levels of depression, anxiety, and stress [53]. The Turkish cultural adaptation, validity, and reliability study of the DASS-21 was conducted by Yılmaz et al [54], who refined the existing Turkish translation of the scale through a comprehensive linguistic and cultural adaptation process and subsequently confirmed its psychometric soundness. Proven to be a valid and reliable tool for screening depressive and anxiety disorders, the DASS-21 effectively evaluates the severity of these 3 psychological conditions through 21 items. The scale consists of 3 subscales: depression (7 items), anxiety (7 items), and stress (7 items). Each item is rated on a 4-point Likert scale ranging from 0 (never) to 3 (always). Lower scores on the subscales indicate better psychological well-being [54]. The electronic (PDF) version of the DASS-21 in Turkish was given to families. At their convenience, caregivers completed the measure themselves and submitted the completed forms via WhatsApp.

Intervention

The HEP Approach, developed by Balıkcı [39], is a systematic early intervention model grounded in EE paradigms, principles of neuroplasticity, and ecological-dynamic theories of human development. Drawing on experimental EE research and mechanisms of brain plasticity, the model translates these principles into clinical practice through an integrated theoretical framework that combines Ecological Theory, Dynamic Systems Theory, Gibson's perception-action perspective, the Theory of Neuronal Group Selection, sensory integration theory, and the PEO model. Within this framework, development is conceptualized as emerging from the continuous interaction among the child, task demands, environmental affordances, and temporal processes [39,40].

Ecological and PEO perspectives inform the alignment of environmental supports and task structures with the infant's regulatory and developmental capacities, particularly within the zone of proximal development. Dynamic Systems Theory underpins the emphasis on non-linear change and self-organization, operationalized clinically through graded challenge, structured variation, and strategic manipulation of environmental and task constraints to facilitate spontaneous motor and behavioral reorganization. Gibson's perception-action framework guides the design of environments that invite active exploration, ensuring that sensory information is directly coupled with meaningful action. The Theory of Neuronal Group Selection and neuroplasticity principles support the systematic use of novelty, repetition within variation, and sustained enriched exposure to strengthen adaptive neural networks through experience-dependent plasticity.

Central to the HEP model is physiological homeostasis, conceptualized as a foundational prerequisite for exploration, engagement, and learning. Regulation of sleep, arousal, stress modulation, and coregulation processes is addressed prior to and throughout intervention to optimize the infant's readiness for adaptive change. Enriched environment principles are operationalized through 10 structured intervention components: physiological homeostasis, safety, provision of multisensory experiences, spatial adaptation, environmental and object novelty, just-right challenge, enjoyment, social opportunities, continuity, and active engagement and exploration. These components guide assessment, hypothesis generation, and intervention planning by systematically aligning environmental affordances, task demands, and individual sensory-motor and regulatory capacities.

In practice, the HEP Approach integrates structured therapist guidance with caregiver coaching to embed enriched, developmentally attuned experiences into daily routines and natural contexts. More detailed information regarding the theoretical foundations and conceptual framework of the HEP model is available in previously published studies [22,24,39,40].

The approach consists of 11 structured phases, beginning with referral and family introduction, followed by comprehensive assessment, identification of child and family strengths and challenges, hypothesis formulation, collaborative goal setting, selection of outcome measures, intervention planning, implementation, monitoring, and evaluation [40]. In this study, all phases of the HEP Approach were adapted for telehealth delivery and implemented over 12 one-hour weekly sessions across 3 months. Families, guided through Zoom, arranged their home environments according to their own resources, preferences, and the child's developmental needs, transforming everyday spaces into therapeutic contexts. Assessments and observations were conducted via structured caregiver interviews, live video interaction, and parent-recorded video clips, while therapists engaged parents in reflective discussions to interpret child behaviors and environmental affordances. Intervention planning and delivery progressed systematically: early sessions emphasized supporting homeostasis and self-regulation; subsequent sessions focused on reorganizing physical and social environments to foster active exploration; later sessions encouraged the diversification and generalization of newly

XSL•FO
RenderX

acquired skills across settings and materials; and final sessions concentrated on empowering parents to independently adapt environments and sustain enriched opportunities.

Each telehealth session also followed a five-step therapeutic sequence to ensure fidelity and caregiver engagement: (1) a check-in and reflection phase, where parents and therapists discussed the past week and reviewed parent-recorded videos; (2) an observation of natural parent-infant interaction, with therapists guiding camera placement to capture spontaneous exploration; (3) a guided reflective discussion, in which therapists and caregivers collaboratively interpreted the infant's strengths, challenges, and responses to environmental affordances; (4) a real-time application phase, where parents directly practiced strategies with their infant during the session under therapist support; and (5) a closure and collaborative planning phase, setting goals and strategies for home practice until the next meeting (Figure 2).

Figure 2. Therapeutic process cycle in telehealth sessions.

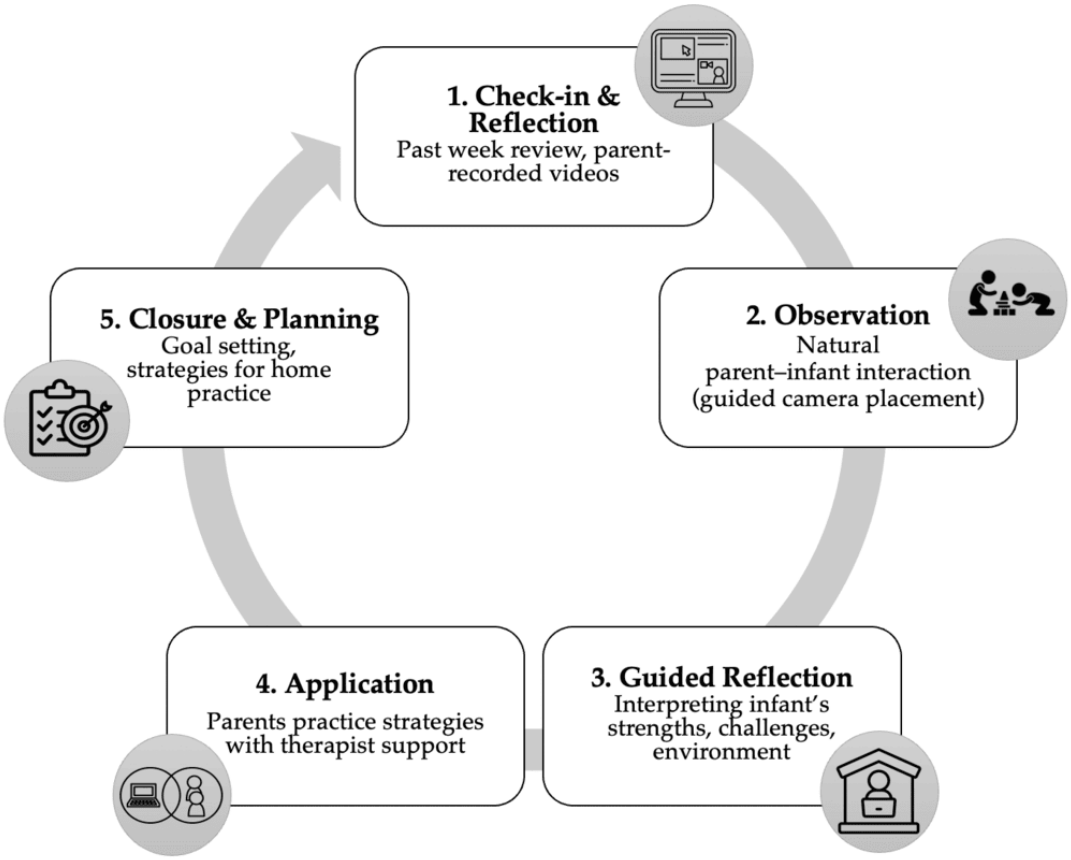

Grounded in 10 core principles—including maintenance of homeostasis, sensory experiences, safety, novelty, spatial adaptation, just-right challenge, enjoyment, social opportunities, continuous engagement, and active exploration—the intervention sought not only to promote self-regulation, sensory processing, motor and functional skills in children, but also to enhance parenting self-efficacy, thereby enabling families to provide ongoing enriched environments across contexts [24,39,40]. Fidelity to the approach was assured through adherence to the HEP manual [22].

Statistical Analysis

Study data were analyzed using IBM SPSS Statistics for Windows (version 22.0; SPSS Inc). Descriptive and demographic variables were summarized with frequency and percentage analyses. Outcome measure scores were examined using means and SDs. Normality of the research variables was assessed through skewness and kurtosis, with values between −1.5 and +1.5 for skewness and between −2.0 and +2.0 for kurtosis considered indicative of a normal distribution [55,56]. All variables met these criteria, confirming normal distribution. Accordingly, parametric statistical methods were applied. Two-sided paired *t* tests were conducted to compare repeated measures within dependent groups, and a *P* value of <0.05 was accepted as statistically significant. Given the exploratory feasibility design of the study, statistical analyses were conducted to examine whether measurable pre-post differences could be detected and to explore the sensitivity of the selected instruments to change within a telehealth context. No formal correction for multiple comparisons was applied, as the analyses were descriptive and hypothesis-generating rather than confirmatory.

Results

Demographic Characteristics of the Participants

The study included 14 infants, of whom 6 (42.9%) were female, and 8 (57.1%) were male. The mean maternal age at the time of birth was 33.42 (SD 3.69, range 25-39) years, while the mean

XSL•FO
RenderX

paternal age was 34.85 (SD 4.46, range 27-44) years. Mothers predominantly held a bachelor's degree (11/14, 78.6%), with a small proportion having completed high school (2/14, 14.3%) or postgraduate education (1/14, 7.1%). Similarly, most fathers had attained a bachelor's degree (12/14, 85.7%), while 7.1% (n=1) had completed high school and 7.1% (n=1) held postgraduate qualifications. The mean gestational age at birth was 32.14 (SD 3.03, range 26-36) weeks, and the mean birth weight was 1951.57 (SD 748.08, range 855-3400) grams. At baseline, the corrected mean infant age was 8.03 (SD 1.53, range 4.5-10.5) months. Detailed demographic characteristics of the participants are presented in Table 1.

Table 1. Participant demographic information (N=14).

Characteristic	Values
Sex, n (%)	
Female	6 (42.9)
Male	8 (57.1)
Mother's educational level, n (%)	
High school	2 (14.3)
Bachelor's degree	11 (78.6)
MSc or PhD	1 (7.1)
Father's educational level, n (%)	
Secondary school	0 (0)
High school	1 (7.1)
Bachelor's degree	12 (85.7)
MSc or PhD	1 (7.1)
Maternal age at infant's birth (years)	
Mean (SD)	33.42 (3.69)
Min-Max	25-39
Father's age at infant's birth (years)	
Mean (SD)	34.85 (4.46)
Min-Max	27-44
Birth gestational age (week)	
Mean (SD)	32.14 (3.03)
Min-Max	26-36
Corrected age at baseline (months)	
Mean (SD)	8.03 (1.53)
Min-Max	4.5-10.5
Birth weight (grams)	
Mean (SD)	1951.57 (748.08)
Min-Max	855-3400

Feasibility, Safety, Acceptability, and Parental Satisfaction in Telehealth

The trial enrolled 16 infants; however, 2 were withdrawn due to medical complications. One infant developed recurrent respiratory distress, and the other experienced feeding difficulties requiring hospitalization. As a result, both were unable to attend four consecutive telehealth sessions and were excluded from the study. Consequently, the intervention was completed with 14 infants who participated in all sessions and constituted the final sample for analysis. Over the 12-week period, all remaining caregivers and their infants attended the weekly sessions regularly and completed all 12 sessions without any disruptions.

The feasibility survey found highly positive results across all domains—feasibility, safety, acceptability, and parental satisfaction—regarding the implementation of the HEP Approach via telehealth. Overall, 86%-100% (n=12-14) of participants selected "Agree" or "Strongly Agree" across all survey items, indicating strong endorsement of the program. In terms of feasibility, 86% (n=12) of families agreed or strongly agreed that the HEP Approach was clear and understandable, and that the recommended activities and environmental arrangements were easy to apply within their daily routines.

XSL•FO
RenderX

Additionally, 93% (n=13) found participation in remote sessions convenient, and the process of regularly sharing home activity videos with the clinician practical and manageable. The remaining responses represented participants who chose neutral options, reflecting uncertainty or hesitation rather than disagreement, as these respondents did not indicate negative perceptions of the telehealth process or program content.

Regarding acceptability, all participants (100%) agreed or strongly agreed that the suggested activities and home arrangements were appropriate and suitable for their infants and families. They emphasized that the therapist's communication and approach toward both parents and infants were respectful and supportive, and that remote sessions were compatible with family dynamics and needs.

Within the satisfaction domain, 100% (N=14) of participants expressed that they were satisfied or highly satisfied with the program. They reported that the intervention effectively met their infants' developmental needs and enhanced their confidence and competence as parents. All families stated that they would recommend the program to other parents.

Regarding safety, 100% (N=14) of families agreed or strongly agreed that all recommended activities were safe for their infants, that they received sufficient guidance to ensure a secure home environment, and that the therapist consistently attended to the infants' and parents' sense of safety. Additionally, 86% (n=12) indicated that no adverse incidents (such as falls or collisions) occurred during the program, while the remaining 14% (n=2) neither agreed nor disagreed, reporting no negative experiences but maintaining a neutral stance.

Qualitative comments from caregivers corroborated these quantitative results. Parents highly valued the program's remote accessibility, the promptness of therapist replies, and the practicality of advice that could be incorporated into everyday routines. Initially, several families encountered difficulties integrating recommended activities into their routines and maintaining their infants' attention during remote sessions; however, these obstacles diminished with time as they adapted to the approach. Participants characterized the program as coherent, efficient, and highly advantageous, noting that it not only facilitated their infants' developmental progress but also enhanced their parental awareness, confidence, and involvement. Additionally, insights obtained from the 3 open-ended survey questions identified the following principal themes.

- The program's most valued features: Parents consistently emphasized that the curriculum was lucid, pragmatic, and easily implementable within the home setting, using resources readily accessible in everyday life. They highlighted that the therapist's instructions were clear and easily comprehensible, which enhanced their confidence in securely applying procedures. A mother remarked, "The activities were straightforward, and I was able to implement them immediately." For instance, I discovered that my infant required visual exposure to his hands before using them, and that maintaining an upright posture was crucial for engaging with his surroundings. Another participant remarked, "I appreciated that I could utilize materials we already possessed at home and did not require any special purchases." For example, I assisted my infant's sitting with a laundry basket, used cardboard boxes as a diminutive table, and crafted vibrant rattles using lentils and plastic bottles—these were inventive, enjoyable, and secure concepts." Numerous families appreciated the inclusive, home-centered format of the meetings, which facilitated the involvement of additional caregivers. One parent stated, "Due to its home-based nature, my husband and grandmother were able to participate, enhancing the experience and providing greater support for everyone involved."
- Suggestions for program improvement: Although parents expressed predominantly favorable opinions, a minority proposed slight enhancements to improve involvement and continuity. They suggested creating a straightforward activity tracking checklist and a visual family guide featuring images and sequential examples of activities. These parents suggested that these tools would help them recall, organize, and confidently recreate activities between sessions. One participant stated, "A straightforward checklist would assist me in adhering to the activities more consistently and monitoring our weekly progress." Another suggested, "A concise booklet featuring images that illustrate the proper positions and materials would enhance ease and safety in practice." These recommendations indicated families' need for organized assistance mechanisms to provide activities that enhance both safety and autonomy at home.
- Supplementary observations and pragmatic advantages: In their supplementary comments, parents underscored the practical and logistical benefits of the home-based arrangement. Avoiding traffic and congested public areas—particularly in major urban centers like Istanbul—was seen as alleviating stress, saving time, and reducing health risks for newborns. Families valued the flexibility to arrange lessons around their child's daily schedule, including nap and feeding times. A mother remarked, "We no longer had to endure prolonged traffic, resulting in a calmer environment for both the baby and us, thereby enhancing the productivity of the sessions." Another remarked, "The most advantageous aspect was the ability to select the most appropriate time for our child." If the baby became fatigued, we could postpone the session to another time, which enhanced its effectiveness.

These thoughts emphasized that the program's remote and home-based format not only improved safety and accessibility but also bolstered family engagement and comfort, strongly correlating with the favorable quantitative results.

Young Children's Participation and Environment Measure

Pre-post comparisons demonstrated statistically significant differences across both home and community participation domains. In the home setting, mean frequency scores increased from 3.67 (SD 0.37) to 4.32 (SD 0.62) ($P<.001$), and involvement scores increased from 2.43 (SD 0.57) to 3.49 (SD 0.47) ($P<.001$). Caregivers reported a decrease in desire for change (58.16, SD 5.63 to 21.43, SD 9.15; $P<.001$) and higher

XSL•FO
RenderX

environmental support scores (78.75, SD 13.39 to 87.36, SD 7.35; *P*=.002).

In the community setting, frequency and involvement scores also showed statistically significant increases (*P*<.001), while desire for change decreased (58.91, SD 10.20 to 39.31, SD 10.30; *P*<.001). Environmental support scores increased from 68.13 (SD 12.75) to 76.19 (SD 13.95) (*P*<.001).

Overall, the YC-PEM captured measurable pre-post changes over the 12-week telehealth delivery period. Further details are presented in Table 2.

XSL•FO
RenderX

Table 2. Outcome changes after 12 weeks of intervention (N=14).

	BI[a], mean (SD)	AI[b], mean (SD)	Estimate of effect (95% CI)	*t* test (*df*)	*P* value
YC-PEM[c] Home					
Frequency	3.67 (0.37)	4.32 (0.62)	–0.65 (–0.96 to –0.35)	–4.655 (13)	<.001[d]
Involvement	2.43 (0.57)	3.49 (0.47)	–1.06 (–1.32 to –0.80)	–8.720 (13)	<.001[d]
Desire for change	58.16 (5.63)	21.43 (9.15)	36.74 (30.25 to 43.22)	12.243 (13)	<.001[d]
Environmental support	78.75 (13.39)	87.36 (7.35)	–8.61 (–13.31 to –3.91)	–3.954 (13)	.002[d]
YC-PEM Community					
Frequency	1.55 (0.51)	2.00 (0.54)	–0.45 (–0.64 to –0.26)	–5.036 (13)	<.001[d]
Involvement	1.46 (0.46)	2.29 (0.78)	–0.83 (–1.07 to –0.58)	–7.351 (13)	<.001[d]
Desire for change	58.91 (10.20)	39.31 (10.30)	19.60 (11.52 to 27.69)	5.236 (13)	<.001[d]
Environmental support	68.13 (12.75)	76.19 (13.95)	–8.06 (–11.41 to –4.71)	–5.195 (13)	<.001[d]
ASQ[e]					
Social emotional	32.50 (16.37)	16.79 (10.85)	15.71 (6.56 to 20.89)	2.393 (13)	<.001[d]
Communication	45.36 (8.65)	50.71 (6.75)	–5.36 (–10.22 to –2.38)	–2.252 (13)	.03[d]
Gross motor	32.50 (16.02)	48.21 (10.30)	–15.71 (–23.55 to –4.33)	–3.626 (13)	.001[d]
Fine motor	44.64 (11.00)	54.64 (4.98)	–10.00 (–16.98 to –3.09)	–3.231 (13)	.009[d]
Problem solving	46.79 (12.19)	53.21 (4.64)	–6.43 (–15.52 to 2.66)	–1.528 (13)	.15
Personal-social	38.93 (12.89)	51.79 (5.75)	–12.86 (–21.16 to –3.35)	–3.841 (13)	.005[d]
AIMS[f]	31.07 (9.93)	47.50 (8.19)	–16.43 (–20.13 to –9.58)	–7.716 (13)	<.001[d]
IMP[g]					
Variation	81.93 (6.81)	96.50 (4.80)	–14.57 (–19.04 to –10.10)	–7.041 (13)	<.001[d]
Adaptability	73.25 (8.95)	91.79 (9.40)	–18.08 (–23.25 to –12.92)	–7.706 (13)	<.001[d]
Symmetry	98.21 (5.21)	99.14 (2.18)	–0.93 (–2.93 to 1.08)	–1.000 (13)	.34
Fluency	88.71 (14.04)	99.14 (3.21)	–10.43 (–18.22 to –2.64)	–2.893 (13)	.01[d]
Performance	70.14 (9.35)	88.93 (6.60)	–18.79 (–22.29 to –11.59)	–11.587 (13)	<.001[d]
Total	82.86 (5.32)	95.00 (4.37)	–12.14 (–14.66 to –10.41)	–10.408 (13)	<.001[d]
DASS-21[h]					
Depression	3.29 (2.40)	1.64 (1.01)	1.64 (0.50 to 2.79)	3.097 (13)	.008[d]
Anxiety	2.36 (0.84)	0.57 (0.65)	1.79 (1.38 to 2.19)	9.555 (13)	<.001[d]
Stress	6.57 (2.68)	4.00 (2.22)	2.57 (1.67 to 3.47)	6.188 (13)	<.001[d]
Total	12.36 (4.80)	6.21 (3.31)	6.14 (4.64 to 7.64)	8.849 (13)	<.001[d]

[a]BI: before intervention.
[b]AI: after intervention.
[c]YC-PEM: Young Children's Participation and Environment Measure.
[d]Statistically significant results ($P<.05$).
[e]ASQ: Ages and Stages Questionnaire.
[f]AIMS. Alberta Infant Motor Scale.
[g]IMP: Infant Motor Profile.
[h]DASS-21: Depression Anxiety Stress Scales.

XSL•FO
RenderX

Ages and Stages Questionnaire

Pre-post comparisons demonstrated statistically significant differences across several ASQ domains. Communication scores increased from 45.36 (SD 8.65) to 50.71 (SD 6.75) (P=.03), gross motor scores from 32.50 (SD 16.02) to 48.21 (SD 10.30) (P=.001), fine motor scores from 44.64 (SD 11.00) to 54.64 (SD 4.98) (P=.009), and personal-social scores from 38.93 (SD 12.89) to 51.79 (SD 5.75) (P=.005). Social-emotional scores decreased from 32.50 (SD 16.37) to 16.79 (SD 10.85) (P<.001).

Problem-solving scores increased over the 12-week period; however, this change did not reach statistical significance (P=.15). Overall, the ASQ detected measurable pre-post differences across multiple developmental domains during the telehealth delivery period. Further details are presented in Table 2.

Alberta Infant Motor Scale

Pre-post analysis revealed a statistically significant difference in AIMS total scores over the study period. The mean total score increased from 31.07 (SD 9.93) at baseline to 47.50 (SD 8.19) at follow-up (P<.001). These results reflect measurable changes in gross motor performance across the 12-week telehealth interval.

Infant Motor Profile

Pre-post comparisons indicated statistically significant differences across multiple IMP domains. Variation scores increased from 81.93 (SD 6.81) to 96.50 (SD 4.80) (P<.001), adaptability from 73.25 (SD 8.95) to 91.79 (SD 9.40) (P<.001), fluency from 88.71 (SD 14.04) to 99.14 (SD 3.21) (P=.01), and performance from 70.14 (SD 9.35) to 88.93 (SD 6.60) (P<.001). Symmetry scores did not differ significantly (P=.34). The overall IMP total score increased significantly from baseline to follow-up, from 82.86 (SD 5.32) to 95.00 (SD 4.37) (P<.001). Further details are presented in Table 2.

Depression Anxiety Stress Scales

Analysis of caregiver-reported outcomes showed statistically significant differences between baseline and follow-up scores. Depression scores declined from 3.29 (SD 2.40) to 1.64 (SD 1.01) (P=.008), anxiety scores from 2.36 (SD 0.84) to 0.57 (SD 0.65) (P<.001), and stress scores from 6.57 (SD 2.68) to 4.00 (SD 2.22) (P<.001). The total DASS-21 score decreased from 12.36 (SD 4.80) to 6.21 (SD 3.31) (P<.001). Detailed results are presented in Table 2.

Discussion

Principal Findings

This study evaluated the feasibility, acceptability, satisfaction, and safety of delivering the HEP Approach through telehealth for premature infants at developmental risk. Fourteen of 16 enrolled infants completed the 12-week program (87.5% retention), and session adherence reached 100%, indicating strong feasibility. Between 86% and 100% of caregivers endorsed all feasibility, acceptability, satisfaction, and safety items. No adverse events were reported. Therapists likewise confirmed that the intervention was practical to deliver remotely and supported appropriate environmental adaptations. Qualitative feedback further emphasized the clarity of instructions, ease of implementation with household materials, logistical advantages (eg, avoiding travel and exposure to illness), and opportunities for broader family involvement. These findings are consistent with previous clinic-based feasibility research and randomized controlled trial outcomes examining the HEP Approach [24,40], supporting the model's applicability across delivery formats.

The present results align with a growing body of telehealth-based early intervention literature demonstrating high parental engagement and satisfaction [31,33,34,57-59]. Telepractice models enabling routine-based home support have been well-received and promote active caregiver involvement [58]. Similar feasibility and satisfaction findings have been reported in the Baby Bridge telehealth model [32], neonatal intensive care unit follow-up telemedicine services [34], and telehealth physiotherapy programs for infants at risk of cerebral palsy [60,61]. Collectively, these studies reinforce the practicality and ecological validity of telehealth in early intervention contexts.

Significant improvements were observed in children's participation, as measured by the YC-PEM. Increases in home and community frequency and involvement, alongside reduced caregiver desire for change and improved environmental support, suggest meaningful enhancement in functional engagement. These findings are consistent with prior studies demonstrating the YC-PEM's sensitivity to intervention-related change across service models [62,63]. Similarly, previous HEP feasibility research using the Goal Attainment Scale reported functional improvements in daily routines [40]. Participation-focused enriched models such as the STEP protocol have also shown gains in YC-PEM outcomes [62,63], and early intervention intensity has been associated with improved home participation and environmental support [41]. Together, this evidence supports the relevance of participation-centered, family-focused intervention frameworks.

Motor development improved substantially. AIMS scores indicated significant gains in gross motor performance, consistent with previous clinic-based HEP studies [24,40]. These findings align with enriched environment-based interventions such as SAFE [21], STEP [62], and telehealth physiotherapy models [60], which have demonstrated positive effects on motor outcomes. The AIMS appears sensitive to motor change across both in-person and telehealth formats.

IMP findings further demonstrated significant improvements in variation, adaptability, fluency, performance, and total scores, reflecting enhancement in both motor quality and milestone acquisition. The absence of change in symmetry likely reflects high baseline symmetry, as similarly reported in CareToy studies [64]. Comparable improvements in IMP domains have been observed in enriched and family-centered models, including CareToy [64], COPCA [65], and EMI-Heart telehealth interventions [66]. Previous HEP studies have also demonstrated positive motor outcomes across multiple assessment tools [22,24,39,40,67]. Together, these results support the IMP's

XSL•FO
RenderX

responsiveness and reinforce the motor benefits of enriched, parent-guided early intervention.

ASQ findings revealed improvements in communication, gross motor, fine motor, personal-social, and social-emotional domains. These gains parallel performance-based motor improvements (AIMS and IMP) and are consistent with prior HEP research [24,40]. Similar multidomain improvements have been reported in enriched early intervention trials, such as the GAME [68,69] and structured physiotherapy programs [70], as well as in Alberta FICare models [71]. These findings suggest that family-centered enriched interventions may extend beyond motor domains, supporting broader sociocommunicative and cognitive development. In telehealth settings, the ASQ provides a practical, sensitive parent-reported measure for monitoring multidimensional progress.

Caregivers demonstrated significant reductions in depression, anxiety, and stress scores. These results mirror prior HEP findings [24] and align with evidence from other early intervention programs for parents of preterm infants [72]. The Mother-Infant Transaction Program [73], home-based postdischarge models [74], and family-centered neonatal care programs [75] have similarly reported improvements in parental psychological well-being. The consistent use of the DASS-21 across these studies supports its utility for detecting meaningful changes in caregiver emotional health.

Despite promising findings, several limitations must be acknowledged. The small sample size and lack of a control group limit causal inference and generalizability. Developmental maturation cannot be entirely excluded as a contributing factor. Long-term effects were not evaluated. Future studies should include larger randomized controlled trials, longitudinal follow-up, hybrid delivery models, cost-effectiveness analyses, and examination of cultural and socioeconomic influences on participation.

Conclusions

This exploratory feasibility research established that the telehealth implementation of the HEP Approach is a feasible, safe, and well-received early intervention paradigm for premature infants at developmental risk. High retention, complete session adherence, and elevated caregiver satisfaction percentages demonstrate that remote implementation of the HEP Approach is both feasible and based on family needs. In addition to feasibility, notable enhancements were evident across various developmental domains, including gross and fine motor skills, communication, social-emotional development, participation outcomes, and caregiver psychological well-being. The findings indicate that the fundamental elements of the HEP model, which prioritize environmental enrichment, parental guidance, and active engagement, can be effectively adapted to telehealth formats without undermining therapeutic integrity.

Despite the promising results, they must be approached with caution due to the limited sample size and the single-group design. Future research must include larger randomized controlled trials and longitudinal follow-up to assess the long-term effects, cost-effectiveness, and generalizability of the telehealth-based HEP Approach across diverse target groups and cultural contexts. The current findings, however, offer initial evidence that the HEP telehealth model constitutes a promising and scalable approach to accessible, family-centered early developmental care for preterm infants.

Acknowledgments

The authors thank all of the families who participated in this study.

Data Availability

The data are not publicly available due to ethical and privacy restrictions, but are available from the corresponding author upon reasonable request.

Funding

This research received no external funding.

Authors' Contributions

AB, AFAB, GCS, and TM-B contributed to the conceptualization of the study. The methodology was developed by HGS, GCS, IMD, AB, and TM-B. Formal analysis was conducted by AFAB, GI, HGS, and TM-B. The investigation was carried out by IMD and GCS. Resources were provided by AFAB, GI, HGS, IMD, TM-B, and AB. Data curation was performed by AFAB, HGS, GI, IMD, GCS, and AB. AB and GCS prepared the original draft of the manuscript. AB, GI, and TM-B contributed to the review and editing of the manuscript. Supervision was provided by AB and TM-B. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of the Homeostasis–Enrichment–Plasticity Approach phases in telehealth delivery.
[DOCX File , 19 KB - jmir_v28i1e86883_app1.docx]

XSL•FO
RenderX

References

1. Bradley E, Blencowe H, Moller A, Okwaraji YB, Sadler F, Gruending A, et al. Born too soon: global epidemiology of preterm birth and drivers for change. Reprod Health 2025;22(Suppl 2):105 [FREE Full text] [doi: 10.1186/s12978-025-02033-x] [Medline: 40551192]
2. Hee Chung E, Chou J, Brown KA. Neurodevelopmental outcomes of preterm infants: a recent literature review. Transl Pediatr 2020;9(Suppl 1):S3-S8 [FREE Full text] [doi: 10.21037/tp.2019.09.10] [Medline: 32206579]
3. Sarda S, Sarri G, Siffel C. Global prevalence of long-term neurodevelopmental impairment following extremely preterm birth: a systematic literature review. J Int Med Res 2021;49(7):3000605211028026 [FREE Full text] [doi: 10.1177/03000605211028026] [Medline: 34284680]
4. You J, Shamsi BH, Hao MC, Cao CH, Yang WY. A study on the neurodevelopment outcomes of late preterm infants. BMC Neurol 2019;19(1):108 [FREE Full text] [doi: 10.1186/s12883-019-1336-0] [Medline: 31146703]
5. Song I. Neurodevelopmental outcomes of preterm infants. Clin Exp Pediatr 2023;66(7):281-287 [FREE Full text] [doi: 10.3345/cep.2022.00822] [Medline: 36596743]
6. Chawanpaiboon S, Vogel JP, Moller AB, Lumbiganon P, Petzold M, Hogan D, et al. Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis. Lancet Glob Health 2019;7(1):e37-e46 [FREE Full text] [doi: 10.1016/S2214-109X(18)30451-0] [Medline: 30389451]
7. Khurana S, Kane AE, Brown SE, Tarver T, Dusing SC. Effect of neonatal therapy on the motor, cognitive, and behavioral development of infants born preterm: a systematic review. Dev Med Child Neurol 2020;62(6):684-692 [FREE Full text] [doi: 10.1111/dmcn.14485] [Medline: 32077096]
8. Zhang Q, Dong X, Jin W, Fan J. Early brain cognitive development in late preterm infants: an event-related potential and resting EEG study. Ital J Pediatr 2024;50(1):26 [FREE Full text] [doi: 10.1186/s13052-023-01567-9] [Medline: 38355639]
9. Nivins S, Padilla N, Kvanta H, Ådén U. Gestational age and cognitive development in childhood. JAMA Netw Open 2025;8(4):e254580 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.4580] [Medline: 40227687]
10. Ko J, Lim HK. Motor development comparison between preterm and full-term infants using alberta infant motor scale. Int J Environ Res Public Health 2023;20(5):3819 [FREE Full text] [doi: 10.3390/ijerph20053819] [Medline: 36900838]
11. Kobaş M, Kızıldere E, Doğan I, Aktan-Erciyes A, Demir-Lira Ö, Akman I, et al. Motor skills, language development, and visual processing in preterm and full-term infants. Curr Psychol 2023;42(15):12463-12475 [FREE Full text] [doi: 10.1007/s12144-021-02658-8]
12. Yildiz R, Yildiz A, Zorlular R, Elbasan B. Relationship between sensory processing skills and motor skills in 12-month-old infants. Brain Behav 2024;14(9):e70052 [FREE Full text] [doi: 10.1002/brb3.70052] [Medline: 39317996]
13. Sırtbaş-Işık G, Yardımcı-Lokmanoğlu BN, Livanelioğlu A, Mutlu A. Sensory processing and sleep characteristics in preterm infants in the early period of life. Sleep Med 2023;106:78-83. [doi: 10.1016/j.sleep.2023.03.025] [Medline: 37054558]
14. Vandormael C, Schoenhals L, Hüppi PS, Filippa M, Borradori Tolsa C. Language in preterm born children: atypical development and effects of early interventions on neuroplasticity. Neural Plast 2019;2019:6873270 [FREE Full text] [doi: 10.1155/2019/6873270] [Medline: 30930944]
15. Palomero-Sierra B, Sánchez-Gómez V, Magán-Maganto M, Bejarano-Martín Á, Ruiz-Ayúcar I, de Vena-Díez VB, et al. Early social communication and language development in moderate-to-late preterm infants: a longitudinal study. Front Psychol 2025;16:1556416 [FREE Full text] [doi: 10.3389/fpsyg.2025.1556416] [Medline: 40248834]
16. Hass J, Panceri C, Procianoy RS, Silveira RDC, Valentini NC. Risk factors for cognitive, motor and language development of preterm children in the first year of life. Rev Paul Pediatr 2022;41:e2021165 [FREE Full text] [doi: 10.1590/1984-0462/2023/41/2021165] [Medline: 36169521]
17. Fan J, Wang J, Zhang X, He R, He S, Yang M, et al. A home-based, post-discharge early intervention program promotes motor development and physical growth in the early preterm infants: a prospective, randomized controlled trial. BMC Pediatr 2021;21(1):162 [FREE Full text] [doi: 10.1186/s12887-021-02627-x] [Medline: 33827496]
18. Orton J, Doyle LW, Tripathi T, Boyd R, Anderson PJ, Spittle A. Early developmental intervention programmes provided post hospital discharge to prevent motor and cognitive impairment in preterm infants. Cochrane Database Syst Rev 2024;2(2):CD005495 [FREE Full text] [doi: 10.1002/14651858.CD005495.pub5] [Medline: 38348930]
19. DeMaster D, Bick J, Johnson U, Montroy JJ, Landry S, Duncan AF. Nurturing the preterm infant brain: leveraging neuroplasticity to improve neurobehavioral outcomes. Pediatr Res 2019;85(2):166-175. [doi: 10.1038/s41390-018-0203-9] [Medline: 30531968]
20. Rocha-Ferreira E, Hristova M. Plasticity in the neonatal brain following hypoxic-ischaemic injury. Neural Plast 2016;2016:4901014 [FREE Full text] [doi: 10.1155/2016/4901014] [Medline: 27047695]
21. Apaydın U, Yıldız R, Yıldız A, Acar Ş, Gücüyener K, Elbasan B. Short-term effects of SAFE early intervention approach in infants born preterm: a randomized controlled single-blinded study. Brain Behav 2023;13(10):e3199 [FREE Full text] [doi: 10.1002/brb3.3199] [Medline: 37534617]
22. Balikci A, Aracikul Balikci AF, Demirbag M, Sirma GC, Beaudry-Bellefeuille I, May-Benson TA. Implementation of the homeostasis-enrichment-plasticity (HEP®) approach for an infant at risk for autism spectrum disorder: a case report. J Child Sci 2025;15(1) [FREE Full text] [doi: 10.53391/2474-5871.1005]

XSL•FO
RenderX

23. Bernabe-Zuñiga JE, Rodriguez-Lucenilla MI, Alias-Castillo AJ, Rueda-Ruzafa L, Roman P, Del Mar Sanchez-Joya M. Early interventions with parental participation and their implications on the neurodevelopment of premature children: a systematic review and meta-analysis. Eur Child Adolesc Psychiatry 2025;34(3):853-865 [FREE Full text] [doi: 10.1007/s00787-024-02528-1] [Medline: 39028424]
24. Sirma G, Zengin Alpozgen A, Balikci A. Effects of the HEP® (Homeostasis-Enrichment-Plasticity) approach in preterm infants with increased developmental risk: a randomized controlled study. Front Pediatr 2025;13:1606490 [FREE Full text] [doi: 10.3389/fped.2025.1606490] [Medline: 41080046]
25. Filippa M, Lordier L, Lejeune F, De Almeida JS, Hüppi PS, Barcos-Munoz F, et al. Effect of an early music intervention on emotional and neurodevelopmental outcomes of preterm infants at 12 and 24 months. Front Psychol 2024;15:1443080 [FREE Full text] [doi: 10.3389/fpsyg.2024.1443080] [Medline: 39498332]
26. Smith L, McParlan A, Niss T, Fucile S. Longitudinal outcomes of a parent-administered sensorimotor intervention in preterm infants at 4 and 18 months. Early Hum Dev 2025;201:106171. [doi: 10.1016/j.earlhumdev.2024.106171] [Medline: 39662236]
27. Little A, Kamholz K, Corwin BK, Barrero-Castillero A, Wang CJ. Understanding barriers to early intervention services for preterm infants: lessons from two states. Acad Pediatr 2015;15(4):430-438 [FREE Full text] [doi: 10.1016/j.acap.2014.12.006] [Medline: 26142069]
28. Altunalan T, Sari Z, Saglam S, Çelik M, Utku Umut G. Investigating enrollment and continuity of parents to early childhood intervention services. Turk J Appl Soc Work 2020;3(2):45-54 [FREE Full text]
29. Decker K, Williams ER, Cook GA, Fry MM. The early intervention referral process for rural infants and toddlers with delays or disabilities: a family perspective. Matern Child Health J 2021;25(5):715-723 [FREE Full text] [doi: 10.1007/s10995-020-03067-2] [Medline: 33201452]
30. Namiiro F, Nolens B, Rujumba J, Kiguli S, Batte A, van den Akker T. "My baby is fine, no need for more clinic visits." Facilitators and barriers for utilisation of follow-up services for children born preterm in low-resource setting: parents' perceptions. Trop Med Int Health 2023;28(3):194-202 [FREE Full text] [doi: 10.1111/tmi.13857] [Medline: 36708235]
31. Sarik DA, Matsuda Y, Terrell EA, Sotolongo E, Hernandez M, Tena F, et al. J Pediatr Nurs 2022;67:139-147. [doi: 10.1016/j.pedn.2022.09.003] [Medline: 36116347]
32. Pineda R, Kellner P, Ghahramani S, Dusing S, Chapman R, Peden CJ. Feasibility of a telehealth model of therapy service delivery after neonatal intensive care unit (NICU) discharge. OTJR (Thorofare N J) 2025:15394492251362721. [doi: 10.1177/15394492251362721] [Medline: 40814234]
33. DeMauro S, Duncan AF, Hurt H. Telemedicine use in neonatal follow-up programs - what can we do and what we can't - lessons learned from COVID-19. Semin Perinatol 2021;45(5):151430 [FREE Full text] [doi: 10.1016/j.semperi.2021.151430] [Medline: 33892961]
34. Das A, Cina L, Mathew A, Aziz H, Aly H. Telemedicine, a tool for follow-up of infants discharged from the NICU? experience from a pilot project. J Perinatol 2020;40(6):875-880. [doi: 10.1038/s41372-020-0593-5] [Medline: 31959907]
35. Peskin A, Barth A, Andrew Rothenberg W, Turzi A, Formoso D, Garcia D, et al. New therapy for a new normal: comparing telehealth and in-person time-limited parent-child interaction therapy. Behav Ther 2024;55(1):106-121. [doi: 10.1016/j.beth.2023.05.002] [Medline: 38216225]
36. Rabner J, Norris LA, Olino TM, Kendall PC. A comparison of telehealth and in-person therapy for youth anxiety disorders. J Clin Child Adolesc Psychol 2025;54(5):555-566. [doi: 10.1080/15374416.2024.2372770] [Medline: 38995683]
37. Lung T, Hayes A, Wen LM, Xu H, Brown V, Baur LA, et al. Socioeconomic differences in the cost-effectiveness of a telephone-based intervention for obesity prevention in early childhood. Int J Obes (Lond) 2026;50(1):152-158. [doi: 10.1038/s41366-025-01904-4] [Medline: 40914767]
38. Salsabilla A, Azzahra AB, Syafitri RIP, Supadmi W, Suwantika AA. Cost-effectiveness of telemedicine in Asia: a scoping review. J Multidiscip Healthc 2021;14:3587-3596 [FREE Full text] [doi: 10.2147/JMDH.S332579] [Medline: 35002248]
39. Balıkcı A. Exploring effects of the HEP (homeostasis-enrichment-plasticity) approach as a comprehensive therapy intervention for an infant with cerebral palsy: a case report. J Child Sci 2022;12(01):e182-e195 [FREE Full text] [doi: 10.1055/s-0042-1757913]
40. Balikci A, May-Benson TA, Sirma GC, Kardas A, Demirbas D, Aracikul Balikci AF, et al. The homeostasis-enrichment-plasticity (HEP) approach for premature infants with developmental risks: a pre-post feasibility study. J Clin Med 2024;13(18):5374 [FREE Full text] [doi: 10.3390/jcm13185374] [Medline: 39336861]
41. Khetani M, Graham JE, Davies PL, Law MC, Simeonsson RJ. Psychometric properties of the young children's participation and environment measure. Arch Phys Med Rehabil 2015;96(2):307-316 [FREE Full text] [doi: 10.1016/j.apmr.2014.09.031] [Medline: 25449189]
42. Williams U, Law M, Hanna S, Gorter J. Using the young children’s participation and environment measure (YC-PEM) to describe young children’s participation and relationship to disability and complexity. J Dev Phys Disabil 2018;31(1):135-148 [FREE Full text] [doi: 10.1007/s10882-018-9637-6]
43. Seyhan-Bıyık K, Delioğlu K, Özal C, Üneş S, Tunçdemir M, Kerem-Günel M. Young children's participation and environment measure (YC-PEM): cross-cultural adaptation, reliability, and validity of Turkish version. Percept Mot Skills 2024;131(6):2125-2149. [doi: 10.1177/00315125241287971] [Medline: 39340534]

XSL•FO
RenderX

44. Kapci E, Kucuker S, Uslu R. How applicable are ages and stages questionnaires for use with Turkish children? Top Early Child Spec Educ 2010;30(3):176-188 [FREE Full text] [doi: 10.1177/0271121410373149]
45. Piper MC, Pinnell LE, Darrah J, Maguire T, Byrne PJ. Construction and validation of the Alberta infant motor scale (AIMS). J Public Health 1992;83(Suppl 2):S46-S50 [FREE Full text] [Medline: 1468050]
46. Piper MC, Darrah J. Motor Assessment of the Developing Infant. Philadelphia, PA: Saunders Co; 1994.
47. Almeida K, Dutra MVP, Mello RRD, Reis ABR, Martins PS. Concurrent validity and reliability of the Alberta infant motor scale in premature infants. J Pediatr (Rio J) 2008;84(5):442-448. [doi: 10.2223/JPED.1836] [Medline: 18923790]
48. Davies S, Lucas BR, Dwyer GM. Reliability of the Alberta infant motor scale (AIMS) when used via telehealth for neurodevelopmentally high-risk infants. Phys Occup Ther Pediatr 2025;45(4):558-571 [FREE Full text] [doi: 10.1080/01942638.2025.2451406] [Medline: 39865795]
49. Rawnsley K, Cheong JLY, Bennett K, Mahady ML, Smith SE, Zannino D, et al. Synchronous telehealth and face-to-face administration of the Alberta infant motor scale. Dev Med Child Neurol 2026;68(1):91-98 [FREE Full text] [doi: 10.1111/dmcn.16391] [Medline: 40576266]
50. Hadders-Algra M, Heineman KR. The Infant Motor Profile. New York, NY, USA: Routledge; 2021.
51. Akbaş AN, Çankaya Ö, Seyhan K, Günel MK. Inter-rater reliability of infant motor profile in 3-24-month-old high-risk infants: Turkey sample. Arc Health Sci Res 2020;6(7) [FREE Full text] [doi: 10.5152/archealthscires.2020.575433]
52. Rizzi R, Menici V, Cioni ML, Cecchi A, Barzacchi V, Beani E, Clinical CareToy-R Consortium. Concurrent and predictive validity of the infant motor profile in infants at risk of neurodevelopmental disorders. BMC Pediatr 2021;21(1):68 [FREE Full text] [doi: 10.1186/s12887-021-02522-5] [Medline: 33549070]
53. Lovibond PF, Lovibond SH. The structure of negative emotional states: comparison of the depression anxiety stress scales (DASS) with the beck depression and anxiety inventories. Behav Res Ther 1995;33(3):335-343. [doi: 10.1016/0005-7967(94)00075-u] [Medline: 7726811]
54. Yılmaz Ö, Boz H, Arslan A. Depresyon anksi˙yete stres ölçeği˙ni˙n (DASS 21) Türkçe kisa formunun geçerli˙li˙k-güveni˙li˙rli˙k çalişmasi [validity and reliability study of the Turkish short form of the Depression Anxiety Stress Scale (DASS-21)]. Res Financ Econ Soc Stud 2017;2:78-91 [FREE Full text]
55. Tabachnick B, Fidell L, Ullman J. Using Multivariate Statistics. Boston, MA, USA: Pearson; 2013.
56. George D, Mallery M. SPSS for Windows Step by Step: A Simple Guide and Reference. Boston, MA, USA: Pearson; 2010.
57. Cheung W, Aleman-Tovar J, Johnston AN, Little LM, Burke MM. A qualitative study exploring parental perceptions of telehealth in early intervention. J Dev Phys Disabil 2023;35(3):353-373 [FREE Full text] [doi: 10.1007/s10882-022-09853-w] [Medline: 35789590]
58. Hinton V, Akemoğlu Y, Tomeny K, McWilliam RA. Supporting families from a distance: implementing routines-based home visits via telepractice. Early Child Educ J 2023:1-8 [FREE Full text] [doi: 10.1007/s10643-023-01464-0] [Medline: 37360610]
59. Jimenez-Arberas E, Casais-Suarez Y, Fernandez-Mendez A, Menendez-Espina S, Rodriguez-Menendez S, Llosa JA, et al. Evidence-based implementation of the family-centered model and the use of tele-intervention in early childhood services: a systematic review. Healthcare (Basel) 2024;12(1):112 [FREE Full text] [doi: 10.3390/healthcare12010112] [Medline: 38201017]
60. Schlichting T, Martins da Silva K, Silva Moreira R, Marques de Moraes MV, Cicuto Ferreira Rocha NA, Boyd RN, et al. Telehealth program for infants at risk of cerebral palsy during the covid-19 pandemic: a pre-post feasibility experimental study. Phys Occup Ther Pediatr 2022;42(5):490-509. [doi: 10.1080/01942638.2022.2057209] [Medline: 35341469]
61. Hielkema T, Hamer EG, Boxum Anke G, La Bastide-Van Gemert S, Dirks T, Reinders-Messelink HA, L2M 0-2 Study Group. LEARN2MOVE 0-2 years, a randomized early intervention trial for infants at very high risk of cerebral palsy: neuromotor, cognitive, and behavioral outcome. Disabil Rehabil 2020;42(26):3752-3761 [FREE Full text] [doi: 10.1080/09638288.2019.1610508]
62. Lima C, Abreu RWFD, Verdério BN, Brugnaro BH, Santos MMD, Dos Santos AN, et al. Early intervention involving specific task-environment-participation (step) protocol for infants at risk: a feasibility study. Phys Occup Ther Pediatr 2023;43(3):303-320. [doi: 10.1080/01942638.2022.2142084] [Medline: 36329671]
63. Lima C, de Abreu RWF, Brugnaro BH, Rocha NACF. Participation of infants at biological risk is facilitated by remote intervention carried out by parents – step protocol: randomized clinical trial. Braz J Phys Ther 2024;28:100674 [FREE Full text] [doi: 10.1016/j.bjpt.2024.100674]
64. Sgandurra G, Lorentzen J, Inguaggiato E, Bartalena L, Beani E, Cecchi F, CareToy Consortium. A randomized clinical trial in preterm infants on the effects of a home-based early intervention with the 'CareToy System'. PLoS One 2017;12(3):e0173521 [FREE Full text] [doi: 10.1371/journal.pone.0173521] [Medline: 28328946]
65. Akhbari Ziegler S, von Rhein M, Meichtry A, Wirz M, Hielkema T, Hadders-Algra M. The coping with and caring for infants with special needs intervention was associated with improved motor development in preterm infants. Acta Paediatr 2021;110(4):1189-1200 [FREE Full text] [doi: 10.1111/apa.15619] [Medline: 33047325]
66. Mitteregger E, Dirks T, Theiler M, Kretschmar O, Latal B. The implementation of EMI-heart, a family-tailored early motor intervention in infants with complex congenital heart disease, in practice: a feasibility RCT. Pilot Feasibility Stud 2024;10(1):105 [FREE Full text] [doi: 10.1186/s40814-024-01532-1] [Medline: 39095881]

XSL•FO
RenderX

67. Balikci A, May-Benson TA, Sirma GC, Ilbay G. HEP (homeostasis-enrichment-plasticity) approach changes sensory-motor development trajectory and improves parental goals: a single subject study of an infant with hemiparetic cerebral palsy and twin anemia polycythemia sequence (TAPS). Children (Basel) 2024;11(7):876 [FREE Full text] [doi: 10.3390/children11070876] [Medline: 39062325]
68. Morgan C, Novak I, Dale RC, Guzzetta A, Badawi N. Single blind randomised controlled trial of GAME (goals - activity - motor enrichment) in infants at high risk of cerebral palsy. Res Dev Disabil 2016;55:256-267. [doi: 10.1016/j.ridd.2016.04.005] [Medline: 27164480]
69. Gündoğmuş E, Bumin G, Yalçın SS. Effect of early intervention on developmental domains and parent-child interaction among children with developmental delay: a randomized controlled study. Am J Occup Ther 2024;78(6):7806205110. [doi: 10.5014/ajot.2024.050706] [Medline: 39418650]
70. Ochandorena-Acha M, Terradas-Monllor M, López Sala L, Cazorla Sánchez ME, Fornaguera Marti M, Muñoz Pérez I, et al. Early physiotherapy intervention program for preterm infants and parents: a randomized, single-blind clinical trial. Children (Basel) 2022;9(6):895 [FREE Full text] [doi: 10.3390/children9060895] [Medline: 35740832]
71. Moe AM, Kurilova J, Afzal AR, Benzies KM. Effects of Alberta family integrated care (FICare) on preterm infant development: two studies at 2 months and between 6 and 24 months corrected age. J Clin Med 2022;11(6):1684 [FREE Full text] [doi: 10.3390/jcm11061684] [Medline: 35330009]
72. Finlayson F, Olsen J, Dusing SC, Guzzetta A, Eeles A, Spittle A. Supporting play, exploration, and early development intervention (SPEEDI) for preterm infants: a feasibility randomised controlled trial in an Australian context. Early Hum Dev 2020;151:105172. [doi: 10.1016/j.earlhumdev.2020.105172] [Medline: 33137579]
73. Landsem I, Handegård BH, Tunby J, Ulvund SE, Rønning JA. Early intervention program reduces stress in parents of preterms during childhood, a randomized controlled trial. Trials 2014;15:387 [FREE Full text] [doi: 10.1186/1745-6215-15-387] [Medline: 25282345]
74. Fan J, He R, He S, Yang M, Tao X, Zhou M, et al. Effectiveness of a home-based, post-discharge early intervention program for very preterm infants in reducing parental stress: a randomized controlled trial. BMC Public Health 2024;24(1):2476 [FREE Full text] [doi: 10.1186/s12889-024-19969-7] [Medline: 39261874]
75. Pourkhani S, Chehrzad MM, Reza Masouleh S, Kazem Nezhad Leyli E. The effect of family-based care on stress, anxiety, and depression of mothers with premature infants. J Holist Nurs Midwifery 2018;28(2):121-128 [FREE Full text] [doi: 10.29252/hnmj.28.2.121]

Abbreviations

AIMS: Alberta Infant Motor Scale
ASQ: Ages and Stages Questionnaire
DASS-21: Depression Anxiety Stress Scales
EI: early intervention
HEP: Homeostasis–Enrichment–Plasticity
IMP: Infant Motor Profile
PEO: Person–Environment–Occupation
YC-PEM: Young Children's Participation and Environment Measure

Edited by M Balcarras; submitted 31.Oct.2025; peer-reviewed by U Apaydin, B Elbasan; comments to author 11.Feb.2026; revised version received 04.Mar.2026; accepted 04.Mar.2026; published 07.Apr.2026.

Please cite as:
Balikci A, Sirma GC, Demirbag IM, May-Benson TA, Sozen HG, Aracikul Balikci AF, Ilbay G
Telehealth Delivery of the Homeostasis–Enrichment–Plasticity Approach for Premature Infants With Developmental Risks: Exploratory Feasibility Study
J Med Internet Res 2026;28:e86883
URL: https://www.jmir.org/2026/1/e86883
doi:10.2196/86883
PMID:

©Aymen Balikci, Gamze Cagla Sirma, Izgi-Miray Demirbag, Teresa A May-Benson, Hatice Gulhan Sozen, Ayse Firdevs Aracikul Balikci, Gul Ilbay. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of Telemedicine vs Face-to-Face Consultation in Fighting COVID-19: Retrospective Cohort Study of Adult Patients With COVID-19 in a Primary Care Setting

Fangfang Jiao[1], PhD; Ka Ming Ho[1], MBChB; Lapkin Chiang[1], MBChB; Siu Hin Ko[1], MBChB; Catherine Xiaorui Chen[1], PhD

Department of Family Medicine and Primary Health Care, Queen Elizabeth Hospital, Hong Kong SAR, China (Hong Kong)

Corresponding Author:
Fangfang Jiao, PhD
Department of Family Medicine and Primary Health Care
Queen Elizabeth Hospital
Room 622, 6 Floor, Nurse Quarter
30 Gascoigne Road, Kowloon
Hong Kong SAR, China
China (Hong Kong)
Phone: 852 35068888
Email: francesjiao0401@gmail.com

Abstract

Background: Telemedicine use expanded rapidly during the COVID-19 pandemic. The Hong Kong Hospital Authority (HA) launched both tele-designated clinics (Tele-DCs) and face-to-face physical designated clinics (PDCs) to manage mild cases. However, the comparative effectiveness of these models remains unclear.

Objective: This study aimed to compare clinical outcomes, specifically hospitalization and severe complications, between patients with mild COVID-19 managed via Tele-DCs versus PDCs in Hong Kong's public primary care setting.

Methods: We conducted a retrospective cohort study involving all patients with COVID-19, aged 18 years or older, who visited a PDC (n=23,031) or a Tele-DC (n=38,628) at the Kowloon Central Cluster in Hong Kong from July 28, 2022, to January 29, 2023. Patients were matched 1:1 using propensity score matching based on age, sex, smoking status, comprehensive social security assistance (CSSA) status, and the Charlson comorbidity score, resulting in 17,199 patients per group. The primary outcome was the hospital admission rate between day 1 and day 28. Secondary outcomes included severe complications, mortality, accident and emergency department (AED) use, the antiviral prescription rate, and DC revisit.

Results: The average age of patients in the Tele-DC and PDC groups was 58.55 (SD 17.53) and 58.53 (SD 17.54) years, respectively (*P*=.93). In both groups, 9.05% (n=1557) of patients were on CSSA, and 11% (n=1892) were smokers. Compared to the PDC group, the Tele-DC group demonstrated similar hospital admission rates (Tele-DC: n=497, 2.89%; PDC: n=471, 2.74%; between-group difference 0.15%, 95% CI –0.20% to 0.50%, *P*=.40), lengths of stay (Tele-DC: mean 6.92, SD 0.47 days; PDC: mean 6.61, SD 0.50 days; between-group difference 0.31 days, 95% CI –1.65 to 1.04, *P*=.66), severe complication rates (Tele-DC: n=46, 0.27%; PDC: n=33, 0.19%; between-group difference 0.08%, 95% CI –0.03% to 0.18%, *P*=.18), and mortality rates (Tele-DC: n=23, 0.13%; PDC: n=18, 0.10%; between-group difference 0.03%, 95% CI –0.04% to 0.10%, *P*=.39). However, the Tele-DC group exhibited a higher AED visit rate (n=641, 3.73%, vs n=542, 3.15%; between-group difference 0.58%, 95% CI 0.19%-0.96%, *P*.003) and DC revisit rate (n=1446, 8.41%, vs n=1287, 7.48%; between-group difference 0.93%, 95% CI 0.09%-1.50%, *P*.002). In addition, the Tele-DC group had a lower antiviral prescription rate (n=9872, 57.4%, vs n=10,797, 62.78%; between-group difference –5.38%, 95% CI –6.41% to –4.32%, *P*<.001).

Conclusions: The tele-DC demonstrated clinical safety comparable to the PDC regarding hospitalization and severe complications for patients with mild COVID-19. By validating a scalable model without complex home monitoring, these findings challenge the strict necessity of physical examinations for safe triage and support a digital-first strategy for future infectious surges. However, the disparities observed in AED visits and antiviral prescription rates suggest that integrated remote monitoring tools and improved medication logistics are needed to fully replicate the efficacy of conventional care.

(J Med Internet Res 2026;28:e74046) doi:10.2196/74046

XSL•FO
RenderX

KEYWORDS

telemedicine; designated clinic; COVID-19; health service use; tele-designated clinic; physical designated clinic

Introduction

COVID-19 was a global pandemic that had a devastating impact on health care worldwide. COVID-19 is highly contagious, with more than 762 million identified cases and 6.8 million deaths reported worldwide as of April 2023 [1]. Although severe and critical complications predominantly occur in older adults or those with certain underlying medical comorbidities [2], the sheer volume of mild-to-moderate cases has historically overwhelmed health care systems, necessitating rapid adaptations in service delivery models to maintain continuity of care, while minimizing transmission risks [3].

In January 2022, Hong Kong experienced the fifth wave of the COVID-19 pandemic. This wave proved to be significantly more severe than its predecessors, with thousands of daily infections placing unprecedented strain on existing health care facilities. To meet the increasing service demand from patients with confirmed COVID-19 in the community, the Hong Kong Hospital Authority (HA) converted several general outpatient clinics (GOPCs) into COVID-19 designated clinics (DCs) starting in January 2022 [4]. Furthermore, to curb the spread of COVID-19 and improve health care accessibility, the HA introduced telehealth services in the form of tele-designated clinics (Tele-DCs) on July 28, 2022. The objective of both physical designated clinics (PDCs) and Tele-DCs was to provide prompt treatment for patients with mild COVID-19 and in a stable clinical condition. Both PDCs and Tele-DCs operated until January 29, 2023, when the Hong Kong government lifted the isolation order for patients with COVID-19 [5].

Telemedicine has several well-established advantages over traditional face-to-face consultation for managing infectious diseases. First, remote consultation via teledevices reduces the risk of physical transmission of communicable illnesses. Second, it facilitates clinical triage, avoiding overwhelming emergency and in-patient services [6]. Third, it improves access to medical care, particularly for patients with mobility issues. Consequently, teleconsultation rapidly gained popularity following the pandemic's onset. It has been effectively used to manage a range of conditions, including diabetes mellitus [7], cancer [8], acute pediatric conditions [9], orthopedic conditions, psychological diseases, cardiovascular diseases, and chronic pain [10].

However, the rapid implementation of telemedicine has outpaced evidence regarding its comparative safety for acute respiratory infections. A critical limitation of teleconsultation is the inability to perform physical examinations, such as lung auscultation, or to obtain immediate vital signs, which are traditional cornerstones of pneumonia assessment [11,12]. Although remote patient monitoring using home oximetry has shown promise in mitigating these risks, such programs are resource intensive and difficult to scale during massive surges [13,14].

The shift to telemedicine in primary care during the COVID-19 pandemic has been extensively documented [15,16], yet studies evaluating its effectiveness in COVID-19–specific care have yielded mixed results. A large cohort study in Spain found that the proactive telemonitoring and teleconsultation for patients with COVID-19 are associated with fewer emergency department (ED) visits, lower hospitalization rates, a shorter length of stay (LOS), and lower mortality when compared with other models of care [17]. Conversely, a cohort study in the United States for post–ED visit follow-up care patients with COVID-19 found that telehealth is associated with increased rates of ED visits and hospitalizations compared with in-person care [18]. Although some studies have indicated high patient satisfaction with telemedicine for various conditions, including COVID-19 [19], robust comparative effectiveness studies, particularly cluster-based cohort studies that account for the safety of consultation-only telemedicine, remain scarce.

This knowledge gap is critical: health care systems need to know whether a scalable, consultation-based telemedicine model (Tele-DC) is as safe as a face-to-face model involving physical examination (PDC) for triaging mild COVID-19 cases. To address this gap, this study aimed to compare the risk of hospitalization and severe complications among patients with mild COVID-19 managed via Tele-DCs versus those managed via PDCs in a public primary care setting in Hong Kong. We hypothesized that Tele-DCs are associated with hospitalization and severe complications rates compared to PDCs. We believe that the findings from this study will provide essential evidence for future service planning, specifically regarding the safety of substituting physical examinations with virtual assessments during infectious surges.

Methods

Study Design

We conducted a retrospective cohort study to compare clinical outcomes between Tele-DCs and PDCs in the management of patients with COVID-19. Patients in both groups were matched using propensity score matching based on age, sex, smoking status, comprehensive social security assistance (CSSA) status, and the Charlson comorbidity score [20]. This study adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [21].

Propensity score matching pairs subjects based on observable characteristics that indicate a similar probability of receiving treatment (similar propensity score), even though they receive different interventions. The propensity score is independent of the outcomes and represents the conditional probability of receiving an intervention, given the observed baseline characteristics. Propensity score matching is suitable for studies with a large sample size and many covariates as it helps balance the distribution of covariates between treatment groups [22]. We used propensity score matching to reduce confounding at baseline.

The Charlson comorbidity index is a weighted index used to predict the risk of death within 1 year of hospitalization for patients with specific comorbid conditions. This index includes 17 conditions, and each condition is assigned a weight ranging

XSL•FO
RenderX

from 1 to 6. These weights are determined based on the estimated 1-year mortality hazard ratio derived from a Cox proportional hazards model. These weights were summed up to produce the Charlson comorbidity score. The conditions incorporated into the Charlson comorbidity score encompassed a wide range of comorbidities identified as risk factors for severe COVID-19. These included cancer, chronic lung disease, stroke, chronic kidney disease (CKD), chronic liver disease, diabetes mellitus, heart conditions, HIV, and more [23].

Study Setting

Since July 28, 2022, the HA has launched DC services in telehealth mode (Tele-DCs) for the management of confirmed cases of COVID-19 with mild symptoms in the community. Individuals aged 18 years and above who had received a positive diagnosis for COVID-19 infection through either a rapid antigen test (RAT) or a COVID-19 polymerase chain reaction (PCR) test could book appointments for the Tele-DC service using the government-developed mobile app HA GO. This app features encrypted real-time video capabilities for teleconsultations, allowing Tele-DC doctors to conduct consultations on an iPad. During the consultation, doctors assessed patients' condition, provided medical treatment, and offered self-care advice to suitable patients via the electronic device. Unstable patients or those at high risk of developing critical conditions were advised to visit the accident and emergency department (AED), as determined by the doctors. For consultations conducted during the morning session, prescribed medications were delivered to patients' registered addresses by the afternoon on the same day. For consultations conducted during the afternoon session, medications were delivered to patients on the following morning. In addition, patients in need also received a pulse oximeter, along with the medications.

The PDC was established in January 2022 and provided face-to-face consultations to patients with mild COVID-19. Prior to consultations, nurses checked every patient's temperature and oxygen saturation using a pulse oximeter. SSurgery tests were available in the PDC, including electrocardiography (ECG) , blood glucose monitoring using Glucometer (Contour Plus Elite, PHC Corporation), and Urine Multistix test. The doctors assessed the patients' condition, provided treatment, and advised unstable patients to attend the AED in the same manner as in the Tele-DC. Following the consultation, medications were dispensed to the patients right away.

Both the Tele-DC and the PDC were free of charge to patients. The attending doctors were the same as the GOPC doctors and had a similar working experience. The assigned consultation time for the Tele-DC and the PDC was similar. As a small proportion of patients had multiple visits at the Tele-DC or the PDC, patient groups were determined based on their first visit to either DC. The Tele-DC and PDC groups were defined as the intervention and control groups, respectively.

Participants

The inclusion criteria were that all patients with COVID-19 had visited one HA Kowloon Central Cluster PDC or Tele-DC from July 28, 2022, to January 29, 2023. During this period, both DC modes were available in the HA for patients to choose from. A COVID-19 diagnosis was identified using International Classification of Primary Care-2 (ICPC-2) A77 code, which includes a diagnosis of COVID-19; Coxsackie virus infection; dengue fever; hand, foot, and mouth disease; and adenovirus infection. When patients booked a PDC appointment over the phone or a Tele-DC appointment through the mobile app, they were reminded that only patients with confirmed COVID-19 were eligible for the booking. Therefore, patients with other illnesses under code A77 were excluded at the time of booking. Furthermore, when nurses initially assessed patients in the PDC or the Tele-DC, they asked and documented the diagnostic method for COVID-19 (either PCR or RAT).

The exclusion criteria were patients younger than 18 years old or patients without a diagnosis code of A77.

Assessments

The primary outcome was the hospital admission rate from day 1 to day 28 after the first visit to the Tele-DC or the PDC. The secondary outcomes included AED visits between day 1 and day 28, the hospital admission rate and AED visit on the same day (day 0) after Tele-DC or PDC visit, the LOS, intensive care unit (ICU) or high-dependency unit (HDU) admission, the revisit rate of the Tele-DC or the PDC after the first consultation at the Tele-DC or the PDC, prescriptions of antivirals, mortality, and development of COVID-19–related severe and critical complications (including pneumonia, sepsis, encephalopathy/encephalitis, myocarditis, systemic inflammatory response syndrome, and shock) within 28 days after the first Tele-DC or PDC visit. To cancel out the effect of antivirals on the severity of COVID-19, we further conducted a subgroup analysis of health service use by prescription of antivirals.

Measurements

Data were collected from the Clinical Management System (CMS) or retrieved from the Clinical Data Analysis and Reporting System (CDARS) of the HA and included the following:

- Baseline characteristics of patients, including sex, age, smoking status, CSSA status, and existing comorbidities. The CSSA is a welfare scheme that provides supplementary payments to Hong Kong residents whose income is not sufficient to meet basic needs. It is used as an indicator of socioeconomic status. The comorbidities at baseline were identified using the *International Classification of Disease, Ninth Revision, Clinical Modification* (ICD-9-CM) codes and ICPC-2 codes (Table S1 in Multimedia Appendix 1).
- Prescription of antivirals (ie, Paxlovid or molnupiravir, two choices available at both the Tele-DC and the PDC).
- Health service use rate on day 0 of visit to the Tele-DC or the PDC, which included AED visits and hospital admission.
- Health service use rates between day 1 and day 28 after the first consultation at the Tele-DC or the PDC, including AED visits, hospital admission, the LOS, and ICU/HDU admission.

XSL•FO
RenderX

- The revisit rate of the Tele-DC or the PDC within 28 days after the first consultation at the Tele-DC or the PDC.
- The development of severe complications within 28 days after COVID-19 diagnosis. These complications, including pneumonia, sepsis, encephalopathy/encephalitis, myocarditis, systemic inflammatory response syndrome, and shock, were identified using ICD-9-CM codes (Table S2 in Multimedia Appendix 1).
- Mortality within 28 days after the first visit to at the Tele-DC or the PDC.

As patients with COVID-19 were managed only in the public health sector in Hong Kong, and the clinical data of the public health sector were all available in the CMS and the CDARS, there were no missing data or loss of data at follow-up in this study. All patients were followed up for 28 days.

Sample Size

We estimated that 3% patients with COVID-19 would require inpatient care during the fifth wave in Hong Kong [24]. We assumed a relative risk of 1.3 between the two groups as clinical significance, with a noninferior limit of 0.9%. We used 80% power (1 – β=80%) and a 95% CI (α .05). Using a sample size calculator [25] for noninferiority studies, 6156 individuals would be needed in each group. To allow for a 20% exclusion rate, a minimum of 7388 individuals would be needed in each group for data analysis. To reduce selection bias, we included all eligible patients in the data analysis.

Data Analysis

For descriptive statistics, continuous variables were expressed as mean (SD) and categorical variables were expressed as percentages. We used the independent *t* test or chi-squared test, as appropriate, to compare demographics, clinical parameters, and health service use between groups. To assess noninferiority for the primary outcome of hospital admission rate, we calculated the absolute risk difference between the Tele-DC and PDC groups, along with its 95% CI. Noninferiority was considered established if the upper bound of the 95% CI for the risk difference was less than the predefined noninferiority margin of 0.9%. The *psmatch2* STATA package was used to conduct propensity score matching. We estimated the average treatment effect on the treated (ATT). We adopted a one-to-one matching without replacement with a caliper of 0.0001, which means the differences in propensity scores for each matched pair were no greater than 0.0001. Unmatched individuals were discarded. No sensitivity analysis was performed in this study. All statistical analyses were performed using STATA/MP (StataCorp), and $P<.05$ was considered statistically significant.

Ethical Considerations

This study was approved by the Research Ethics Committee (Kowloon Central/Kowloon East; reference #KC/KE-23-0092/ER-3). Informed consent was waived by the ethics committee because this retrospective observational study involved only the analysis of deidentified data collected for routine clinical care, with no direct patient contact. All the study data were deidentified. No compensation was provided to participants, as there was no direct contact with patients and no active participation required. No images or supplementary materials in this manuscript contain identifiable information about individual participants. No patient photographs or personal health information are included.

Results

Basic Patient Characteristics

A total of 61,659 patients visited either the Tele-DC or the PDC during the study period. Among them, 23,031 (37.35%) visited the PDC, of which 4300 (18.67%) were younger than 18 years old. After excluding patients without a diagnosis of COVID-19, there were 36,951 (59.93%) patients in the Tele-DC group and 18,371 (79.77%) patients in the PDC group (Figure 1).

Before propensity score matching, patients in the Tele-DC group were significantly younger than those in the PDC group (mean 55.30, SD 17.63, years vs mean 59.31, SD 17.57, years; $P<.001$). The proportion of patients on CSSA was significantly higher in the PDC group (n=2039, 11.1%) than in the Tele-DC group (n=2465, 6.67%; $P<.001$). The PDC group also showed baseline comorbidities, as indicated by the Charlson comorbidity index (mean 0.10, SD 0.42) compared to the Tele-DC group (mean 0.04, SD 0.28; $P<.001$), as shown in Table 1.

XSL•FO
RenderX

Figure 1. Flowchart of participant selection. DC: designated clinic; ICPC-2: International Classification of Primary Care-2; PDC: physical designated clinic; Tele-DC: tele-designated clinic.

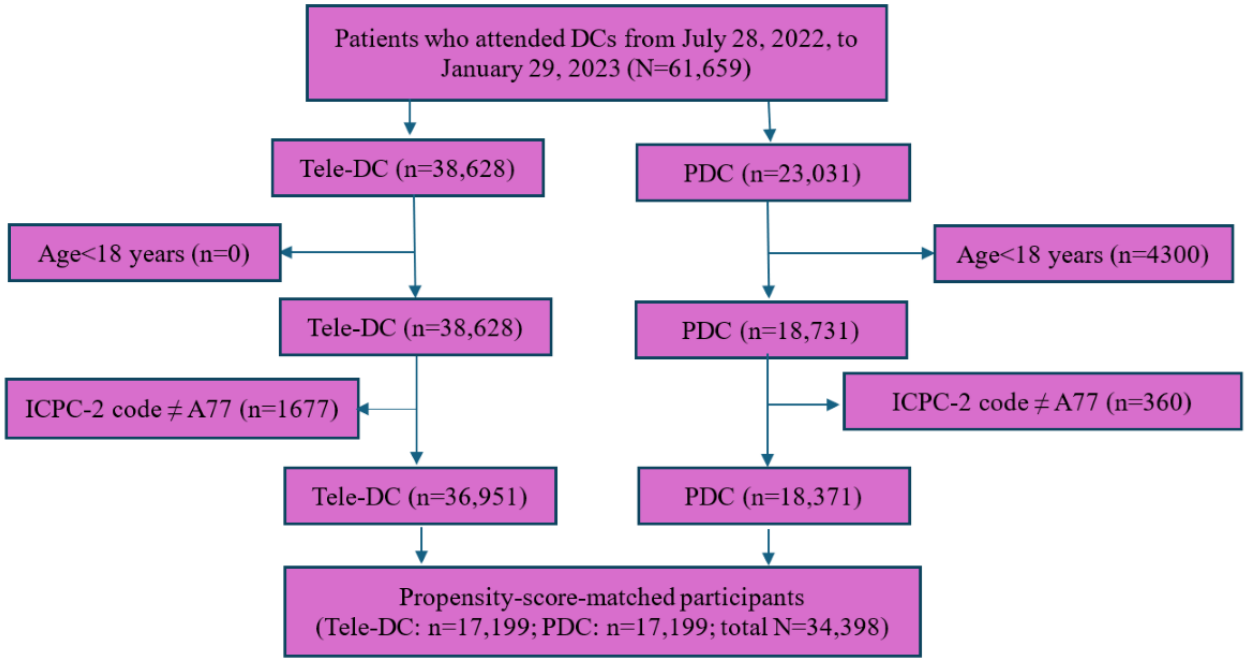

Table 1. Basic characteristics of study participants before propensity score matching.

Characteristics	Tele-DC[a] (n=36,951)	PDC[b] (n=18,371)	Standardized mean difference	*P* value
Age (years), mean (SD)	55.30 (17.63)	59.31 (17.57)	–22.80	<.001[c]
Age group (years), n (%)				
18-29	2727 (7.38)	1034 (5.62)	—[d]	<.001[c]
30-44	8853 (23.96)	3331 (18.13)	—	—
45-59	8477 (22.94)	3794 (20.65)	—	—
60-74	11,806 (31.95)	6617 (36.02)	—	—
≥75	5088 (13.77)	3599 (19.59)	—	—
Female, n (%)	22,766 (61.61)	10,819 (58.89)	5.60	<.001[c]
Current or ex-smokers, n (%)	3769 (10.20)	2379 (12.95)	9.00	<.001[c]
Patients on CSSA[e], n (%)	2465 (6.67)	2039 (11.10)	–15.60	<.001[c]
Charlson comorbidity index, mean (SD)	0.04 (0.28)	0.10 (0.42)	–17.50	<.001[c]

[a]Tele-DC: tele-designated clinic.

[b]PDC: physical designated clinic.

[c]Not applicable.

[d]Significant *P* values.

[e]CSSA: comprehensive social security assistance.

After propensity score matching, each group had 17,199 patients. The average age in both groups was 58.55 (SD 17.53) years. Females accounted for almost the same proportion in both groups: 59.53% (n=10,239) of patients in the Tele-DC group and 59.50% (n=10,233) of patients in the PDC group. Similarly, the proportion of patients on CSSA was 9.05% (n=1557) in both groups. The Charlson comorbidity index was also similar in each group (Tele-DC: mean 0.05, SD 0.28; PDS: mean 0.04, SD 0.27; *P*=.77), as shown in Table 2.

XSL•FO
RenderX

Table 2. Basic characteristics of study participants after propensity score matching.

Characteristics	Tele-DC[a] (n=17,199)	PDC[b] (n=17,199)	Standardized mean difference	*P* value
Age (years), mean (SD)	58.55 (17.53)	58.53 (17.54)	–0.001	.93
Age group (years), n (%)				
18-29	1013 (5.89)	1018 (5.92)	__[c]	.99
30-44	3287 (19.11)	3287 (19.11)	—	—
45-59	3658 (21.27)	3658 (21.27)	—	—
60-74	6094 (35.43)	6094 (35.43)	—	—
≥75	3147 (18.30)	3142 (18.27)	—	—
Female, n (%)	10,239 (59.53)	10,233 (59.50)	0.001	.96
Current or ex-smokers, n (%)	1894 (11.01)	1892 (11.00)	0.002	.99
Patients on CSSA[d], n (%)	1557 (9.05)	1557 (9.05)	0.000	.99
Charlson comorbidity index, mean (SD)	0.05 (0.28)	0.04 (0.27)	0.002	.77

[a]Tele-DC: tele-designated clinic.

[b]PDC: physical designated clinic.

[c]Not applicable.

[d]CSSA: comprehensive social security assistance.

Health Service Use

The primary outcome was the hospital admission rate from day 1 to day 28 after visiting the DCs, which was similar between the Tele-DC (2.89%) and and PDC (2.74%) groups, respectively (*P*=.40). The absolute between-group difference was 0.15% (95% CI –0.20% to 0.50%). Because the upper bound of the 95% CI (0.50%) was well below the predefined noninferiority margin of 0.9%, the Tele-DC was deemed noninferior to the PDC for the primary outcome. The average LOS was comparable between the two groups (Tele-DC: mean 6.92, SD 0.47 days; PDC: mean 6.61, SD 0.50 days; *P*=.66). In both groups, 15 (0.09%) patients required ICU or HDU admission. The proportion of patients admitted to the hospital on day 0 was similar (Tele-DC: n=108, 0.63%; PDC: n=86, 0.50%; *P*=.11). However, more patients in the Tele-DC group (n=641, 3.73%) visited the AED within 1-28 days than in the PDC group (n=542, 3.15%; *P*=.003). The proportion of patients visiting the AED on day 0 was similar (Tele-DC: n=183, 1.06%; PDC: n=175, 1.02%; *P*=.67), as shown in Table 3.

XSL•FO
RenderX

Table 3. Comparison of health service use.

Health care service use	Tele-DC[a] (n=17,199)	PDC[b] (n=17,199)	Between-group difference[c]		*P* value[d]
	Value	Value	Percentage (95% CI)	Mean (95% CI)	
Hospital admission within 1-28 days after DC[e] visit					
Overall admission rate, n (%)	497 (2.89)	471 (2.74)	0.15 (–0.20 to 0.50)	—[f]	.40
LOS[g], mean (SD)	6.92 (0.47)	6.61 (0.50)	—	0.31 (–1.65 to 1.04)	.66
HDU[h]/ICU[i] admission rate, n (%)	15 (0.09)	15 (0.09)	0	—	.99
LOS in HDU/ICU, mean (SD)	3.07 (0.70)	7.33 (3.75)	—	–4.27 (–3.55 to 12.08)	.27
Hospital admission on day 0 after DC visit					
Overall admission rate, n (%)	108 (0.63)	86 (0.50)	0.13 (–0.03 to 0.29)	—	.11
LOS, mean (SD)	13.76 (1.57)	10.55 (1.21)	—	3.21 (–7.30 to 0.87)	.12
HDU/ICU admission rate, n (%)	5 (0.03)	1 (0.01)	—	—	—
LOS in HDU/ICU, mean (SD)	9.00 (6.76)	3 (—)	—	—	—
AED[j] visit, n (%)					
Within 1-28 days after DC visit	641 (3.73)	542 (3.15)	0.58 (0.19 to 0.96)	—	.003[k]
On day 0 after DC visit	183 (1.06)	175 (1.02)	0.05 (–0.17 to 0.26)	—	.67
Proportion of patients with repeated DC visit, n (%)					
Patients with second DC visit	1446 (8.41)	1,287 (7.48)	0.93 (0.09 to 1.50)	—	.002[k]
Tele-DC visit	1220 (7.09)	373 (2.17)	—	—	<.001[k]
PDC visit	226 (1.31)	914 (5.31)	—	—	—
Patients with third DC visit	150 (0.87)	129 (0.75)	0.12 (–0.07 to 0.31)	—	.21[k]
Tele-DC visit	118 (0.69)	55 (0.32)	—	—	<.001[k]
PDC visit	32 (0.19)	74 (0.43)	—	—	—

[a]Tele-DC: tele-designated clinic.
[b]PDC: physical designated clinic.
[c]Between-group difference = Tele-DC group mean or percentage – PDC group mean or percentage.
[d]*P* values calculated with the independent *t* test or the chi-squared test, as appropriate.
[e]DC: designated clinic.
[f]Not applicable.
[g]LOS: length of stay.
[h]HDU: high-dependency unit.
[i]ICU: intensive care unit.
[j]AED: accident and emergency department.
[k]Significant *P* values.

After visiting the PDC, 7.48% (n=1287) of patients made a second visit to the DC; the majority of these (n=914, 71.01%) chose to revisit the PDC. In contrast, in the Tele-DC group, the revisit rates were slightly higher, with 8.41% (n=1446) of patients visiting the DC again; the majority of these (n=1220, 84.37%) opted to visit the Tele-DC for their second visit (Table 3).

Prescription of Antivirals

The PDC group exhibited a higher prescription rate for antivirals (n=10,797, 62.78%) than the Tele-DC group (n=9872, 57.40%; *P*<.001). Paxlovid accounted for ~75% of the prescriptions in both groups. Furthermore, the PDC group also had a higher proportion of patients aged 60 years or older who received antivirals (n=7714, 83.47%) compared to the Tele-DC group (n=7147, 77.37%; *P*<.001), as shown in Table 4.

XSL•FO
RenderX

Table 4. Prescription of antivirals.

Patient group and antivirals	Tele-DC[a], n (%)	PDC[b], n (%)	Between-group difference[c], percentage (95% CI)	*P* value[d]
All patients (n=17,199 in each group)				
Any antivirals	9872 (57.40)	10,797 (62.78)	–5.38 (–6.41 to –4.34)	<.001
Paxlovid	7561 (43.96)	8085 (47.01)	–3.05 (–3.04 to –2.00)	<.001
Monulpiravir	2311 (13.44)	2712 (15.77)	–2.33 (–3.08 to –1.59)	<.001
Patients aged ≥60 years (Tele-DC: n=9237; PDS: n=9242)				
Any antivirals	7147 (77.37)	7714 (83.47)	–6.10 (–7.23 to –4.95)	<.001
Paxlovid	5196 (56.25)	5411 (58.55)	–2.30 (–3.72 to –0.87)	.002
Monulpiravir	1951 (21.12)	2303 (24.92)	–3.8 (–5.01 to –2.59)	<.001

[a]Tele-DC: tele-designated clinic.

[b]PDC: physical designated clinic.

[c]Between-group difference = Tele-DC group percentage – PDC group percentage.

[d]*P* values calculated with the chi-squared test. All *P* values were significant.

Clinical Outcomes of Patients With COVID-19

The proportion of patients who developed severe complications was similar between the Tele-DC and PDC groups. Among patients who visited the AED with or without admission within 1-28 days after visiting the DC, 46 (0.27%) patients in the Tele-DC group and 33 (0.19%) patients in the PDC group developed severe complications (*P*=.18). In addition, 25 (0.15%) patients the Tele-DC group and 17 (0.10%) patients in the PDC group visited the AED on the same day after their DC visit and developed severe complications (*P*=.28). The mortality rate was also similar between the two groups among patients who visited the AED within 1-28 days after visiting the DC: 23 (0.13%) in the Tele-DC group and 18 (0.10%) in the PDC group (*P*=.39), as shown in Table 5.

Table 5. Clinical outcomes.

Patient group and AED[a] visits	Tele-DC[b] (n=17,199), n (%)	PDC[c] (n=17,199), n (%)	Between-group difference[d], percentage (95% CI)	*P* value[e]
Development of severe complications[f]				
AED visit within 1-28 days after DC[g] visit	46 (0.27)	33 (0.19)	0.08 (–0.03 to 0.18)	.18
AED visit on day 0 after DC visit	25 (0.15)	17 (0.10)	0.05 (–0.03 to 0.12)	.28
Death within 28 days after visit				
AED visit within 1-28 days after DC visit	23 (0.13)	18 (0.10)	0.03 (–0.04 to 0.10)	.39
AED visit on day 0 after DC visit	10 (0.06)	6 (0.03)	0.03 (–0.02 to 0.07)	.45

[a]AED: accident and emergency department.

[b]Tele-DC: tele-designated clinic.

[c]PDC: physical designated clinic.

[d]Between-group difference = Tele-DC group percentage – PDC group percentage.

[e]*P* values calculated with the chi-squared test.

[f]Severe complications included pneumonia, sepsis, encephalopathy/encephalitis, myocarditis, systemic inflammatory response syndrome, and shock.

[g]DC: designated clinic.

Subgroup Analysis of Health Service Use by Prescription of Antivirals

To account for the effects of antiviral use on outcomes, we conducted a subgroup analysis stratified by antiviral prescription. Among patients with an antiviral prescription, the Tele-DC and PDC groups had similar admission rates within 1-28 days after DC visits (n=357, 3.62%, vs n=370, 3.43%; *P*=.46) and a similar LOS (mean 6.42, SD 0.50, vs mean 6.29, SD 0.54; *P*=.87). Among patients without an antiviral prescription, the Tele-DC and PDC groups had comparable admission rates within 1-28 days after DC visits as well (n=140, 1.91%, vs n=101, 1.58%; *P*=.14). The hospital admission rate on day 0 after DC visits was also similar between the Tele-DC and PDC groups, regardless of whether antivirals were prescribed. Of note, among patients without an antiviral prescription, the Tele-DC group had more AED visits within 1-28 days (n=252, 3.44%) compared to the PDC group (n=152,

XSL•FO
RenderX

2.37%; *P*<.001), while among patients with an antiviral prescription, the Tele-DC group had a higher proportion of patients with a second DC visit (n=906, 9.18%) compared to the PDC group (n=856, 7.93%; *P*<.001), as shown in Table 6.

Table 6. Comparison of health service use by antiviral prescription subgroups.

Health care service use	Tele-DC[a] (n=17,199)	PDC[b] (n=17,199)	Between-group difference[c]		*P* value[d]
	Value	Value	Percentage (95% CI)	Mean (95% CI)	
Patients with antiviral prescription and hospital admission within 1-28 days after DC[e] visit (Tele-DC: n=9872; PDC: n=10,797)					
Overall admission rate, n (%)	357 (3.62)	370 (3.43)	0.19 (–0.31 to 0.69)	—[f]	.46
LOS[g], mean (SE)	6.42 (0.50)	6.29 (0.54)	—	0.13 (–1.34 to 1.59)	.87
Patients without antiviral prescriptionand hospital admission within 1-28 days after DC visit (Tele-DC: n=7327; PDC: n=6402)					
Overall admission rate, n (%)	140 (1.91)	101 (1.58)	0.33 (–0.10 to 0.77)	—	.14
LOS, mean (SE)	7.86 (0.95)	7.78 (1.26)	—	0.08 (–3.02 to 3.18)	.96
Patients with antiviral prescription and hospital admission on day 0 after DC visit (Tele-DC: n=9872; PDC: n=10,797)					
Overall admission rate, n (%)	20 (0.20)	20 (0.19)	0.01 (–0.10 to 0.13)	—	.78
LOS, mean (SE)	13.40 (4.84)	12.95 (3.70)	—	0.45 (–11.88 to 12.78)	.94
Patients without antiviral prescription and hospital admission on day 0 after DC visit (Tele-DC: n=7327; PDC: n=6402)					
Overall admission rate, n (%)	88 (1.20)	66 (1.03)	0.17 (–0.18 to 0.52)	—	.35
LOS, mean (SE)	13.84 (1.61)	9.82 (1.12)	—	4.02 (–0.11 to 8.16)	.06
Patients with antiviral prescription and AEDh visit (Tele-DC: n=9872; PDC: n=10,797), n (%)					
Within 1-28 days after DC visit	389 (3.94)	390 (3.61)	0.33 (–0.19 to 0.85)	—	.22
On day 0 after DC visit	35 (0.35)	48 (0.44)	–0.09 (–0.26 to 0.08)	—	.31
Patients without antiviral prescription and AED visit (Tele-DC: n=7327; PDC: n=6402), n (%)					
Within 1-28 days after DC visit	252 (3.44)	152 (2.37)	1.07 (0.51 to 1.63)	—	**<.001[i]**
On day 0 after DC visit	148 (2.02)	127 (1.98)	0.04 (–0.43 to 0.51)	—	.88
Patients with antiviral prescription and repeated DC visit (Tele-DC: n=9872; PDC: n=10,797), n (%)					
Second DC visit	906 (9.18)	856 (7.93)	1.25 (0.49 to 2.01)	—	**<.001[i]**
Third DC visit	96 (0.97)	95 (0.88)	0.09 (–0.17 to 0.35)	—	.49
Patients without antiviral prescription and repeated DC visit (Tele-DC: n=7327; PDC: n=6402), n (%)					
Second DC visit	540 (7.37)	431 (6.73)	0.64 (–0.22 to 1.50)	—	.15
Third DC visit	54 (0.74)	34 (0.53)	0.21 (–0.06 to 0.47)	—	.13

[a]Tele-DC: tele-designated clinic.
[b]PDC: physical designated clinic.
[c]Between-group difference = Tele-DC group mean or percentage – PDC group mean or percentage.
[d]*P* values calculated with the chi-squared test.
[e]DC: designated clinic.
[f]Not applicable.
[g]LOS: length of stay.
[h]AED: accident and emergency department.
[i]Significant *P* values.

Discussion

Principal Findings

To the best of our knowledge, this is the first study to compare clinical outcomes between telemedicine and face-to-face consultations in managing patients with mild COVID-19 in a primary care setting using a large cluster-based cohort. Our findings revealed that the Tele-DC and PDC groups had comparable hospital admission rates, LOS, rates of severe complications, and mortality rates. However, the Tele-DC group exhibited a higher rate of visits at the AED and revisits at the DCs.

XSL•FO
RenderX

It is important to acknowledge a notable demographic difference between the Tele-DC and PDC groups prior to matching. The Tele-DC group consisted of younger individuals with fewer comorbidities, a lower prevalence of smoking, and a comparatively higher socioeconomic status. Similar findings have been observed in other studies comparing telemedicine groups with usual-care groups [26,27]. Telemedicine is a form of medical practice that heavily relies on internet technology as an extension of a doctor's work environment [28]. Consequently, younger and more affluent individuals are more inclined to choose telemedicine as their preferred choice to receive medical care. To mitigate the impact of age, comorbidities, and socioeconomic status on clinical outcomes, we conducted matching between these two groups by age, sex, smoking status, CSSA status, and the Charlson comorbidity index. It should be noted that despite its potential, telemedicine can inadvertently widen health inequities by relying on digital access, a reliable internet connection, and digital literacy [29]. The use of telemedicine is limited by reduced access among socioeconomically disadvantaged populations who cannot afford the necessary devices or may not have internet access. Social justice and financial concerns that limit the equitable provision of services could be overcome through public funding and government investment [30].

Regarding our primary outcome, the comparable rates of hospital admission, LOS, and severe outcomes between the Tele-DC and PDC groups provide robust evidence that the lack of physical examination in a telemedicine setting does not compromise patient safety in mild COVID-19 cases. These findings suggest that for patients triaged as mild, clinical history taking and visual assessment via video or audio are sufficient to identify risk factors and manage the illness course. The similarity in HDU/ICU admissions and mortality rates further indicates that the Tele-DC model successfully identifies patients requiring escalation of care, with accuracy comparable to in-person clinicians, ensuring that high-risk patients are not missed despite the remote interface.

Future Implications

The results of this study have significant implications for future pandemic preparedness and primary care resource allocation. First, establishing the clinical safety of Tele-DCs supports their use as a standard first line of defense during infectious disease surges, allowing health care systems to reserve physical clinic space and personal protective equipment for patients with more complex needs. Second, the data support a shift in policy toward a digital-first triage model for respiratory infections, reducing the risk of nosocomial transmission in waiting rooms. Finally, the comparable LOS suggests that patients managed remotely do not present to the hospital too late or in a worse condition than those seen in person, validating the effectiveness of the safety-net advice given during teleconsultations. Our findings align with the literature demonstrating the safety of telemedicine for acute respiratory infections. A study in Spain similarly reported that remote monitoring of patients with COVID-19 results in low hospitalization rates and low mortality rates, without an increase in adverse events compared to standard care [17].

However, as a physical examination could not be performed and some vital signs could not be checked for the Tele-DC group, the patients were understandably more inclined to seek face-to-face consultations at the AED for reassurance when their clinical condition changed. This may explain why the AED visit rate among patients in the Tele-DC group was slightly higher than that of the PDC group. This finding is similar to that of a previously large cohort study in the United States [31]. However, despite the higher AED visit rate, we did not observe a corresponding increase in the hospital admission rate. This suggests that most of the patients in the Tele-DC group visiting the AED did not exhibit severity warranting hospital admission.

The AED visit rate on day 0 after DC visits was similar between the two groups. The hospital admission rate on day 0 after Tele-DC visits was similar to that of the PDC group. This indicates that both groups have a similar ability to identify patients with potentially severe conditions requiring immediate AED or inpatient care. Although the PDC and Tele-DC were, indeed, designed to manage mild COVID-19, the disease's dynamic nature means patients can rapidly deteriorate [32]. Our DC service serves as a critical safety net, identifying and escalating patients whose condition worsens. Therefore, potential severe cases with hospital admission or AED visits on the same day were not excluded but served as vital safety outcomes. These outcomes reflect the utility of the service in distinguishing patients suitable for community management and those requiring immediate higher-level care [33]. Excluding these cases would artificially inflate perceived success, introduce significant selection bias, and mask the real-world performance of our service. Their inclusion provides a more comprehensive and robust evaluation of our Tele-DC's crucial role in patient management and safety.

The DC revisit rate for patients in the Tele-DC group after their initial visit was higher than that for the PDC group. This higher rate of follow-up visits may indicate that the patients' needs were not adequately addressed at the first consultation, requiring additional unplanned care [15]. Beyond the lack of physical examinations, several other factors likely contributed to the increased AED visits and DC revisit rate in the Tele-DC group. First, provider caution and diagnostic uncertainty played a significant role. Faced with a novel and potentially rapidly progressing disease like COVID-19, and without the full spectrum of diagnostic information available in a face-to-face setting, doctors using telemedicine may have adopted a lower threshold for recommending an AED visit or an in-person follow-up. This cautious approach, although intended to ensure patient safety, can lead to overtriage and increased resource use. Second, patient anxiety and the need for reassurance cannot be underestimated. The psychological impact of a pandemic, coupled with the uncertainty surrounding COVID-19 symptoms, may have led patients to seek in-person care for reassurance [34], even for symptoms objectively deemed mild, if their virtual consultation did not fully alleviate their concerns. We did not have data on the exact reasons for AED visits and DC revisits.

These findings carry significant implications, particularly concerning the sufficiency of telemedicine for the triage of high-risk patients with COVID-19. Although our study demonstrated favorable outcomes for mild cases, the observed

XSL•FO
RenderX

association with an increase in AED visits and DC revisits in the Tele-DC group strongly suggests its potential limitations for definitively triaging patients with moderate-to-severe symptoms or those at high risk of deterioration. The higher AED use, even if not accompanied by severe outcomes in our cohort, underscores a systemic challenge in balancing patient safety with efficient resource allocation when relying solely on virtual care for potentially complex presentations [12].

Interestingly, we observed 84.7% (1220/1446) patients continued with the Tele-DC after their initial visit, while only 71.0% (914/1287) of patients stuck to the PDC for their second visit. Several explanations may account for this observation. First, the Tele-DC provided easy access and saved travel time and money for patients. Therefore, if the purpose of the second visit was for symptomatic treatment of prolonged symptoms, patients were more likely to choose the Tele-DC. This was particularly true for individuals with limited mobility due to comorbidities, who were more prone to experiencing prolonged symptoms. Second, previous studies have demonstrated high patient satisfaction rates, shorter waiting times, and adequate communication with physicians in telemedicine settings [35,36]. As a result, ~80% of patients expressed willingness to choose telemedicine for future consultations after their initial visit [36].

However, the Tele-DC group exhibited a lower prescription rate of antivirals in both the overall patient population and older patients aged 60 years or above. According to the guideline for antiviral prescription from the Hong Kong Centre for Health Protection, antivirals are recommended for two groups: (1) patients ≥60 years old or <60 years old but with high-risk conditions, such as smoking, previous smoking history, or certain comorbidities, and (2) patients seen within 5 days since the onset of symptoms [37]. For Tele-DC consultations scheduled during the afternoon session, antivirals were delivered the next day. However, for patients in the PDC group, the antivirals were dispensed immediately after the consultation, regardless of whether it took place in the morning or in the afternoon. Based on the experience of Tele-DC doctors, a small proportion of patients in the Tele-DC group were not prescribed antivirals because their disease course would have exceeded 5 days by the next day of consultation. This may partly explain the lower prescription rate of antivirals observed in the Tele-DC group. Similarly, two previous studies focusing on telemedicine for managing acute upper respiratory tract infections, including sinusitis and pharyngotonsillitis, have found that the telemedicine leads to a lower prescription rate for antibiotics [26,38]. Therefore, Tele-DC doctors may have exercised caution in prescribing antivirals, considering the absence of a physical examination and a lack of vital sign assessment. Furthermore, the Tele-DC was popular and needed time to book. Maybe many patients in the Tele-DC group were already in days 3-4 of their disease course and, therefore, were not so eager to receive antiviral therapy. In addition, patients in the Tele-DC group were usually more well educated and health conscious; they were perhaps not willing to try the new antivirals due to their potential side effects, despite medical advice.

Our subgroup analysis revealed that both Tele-DC and PDC models were generally associated with comparable hospital admission rates, both on day 0 and within 1-28 days postvisits, regardless of whether patients received antiviral treatment. This suggests that Tele-DCs are associated with safety outcomes comparable to traditional face-to-face care for hospitalized patients with COVID-19 [39]. However, among patients without an antiviral prescription, the Tele-DC group exhibited a significantly higher rate of AED visits within 1-28 days. This finding suggests that although the Tele-DC may have been associated with similar hospital admission rates, it correlated with an increase in urgent care seeking in the absence of antiviral therapy, possibly due to a lack of direct physical assessment or differing thresholds for patient reassurance and escalation of care. Moreover, among patients receiving antiviral treatment, the Tele-DC was associated with a higher proportion requiring a second DC visit, potentially indicating a need for more frequent follow-up consultations to manage treatment or address ongoing concerns, even as admission rates and the LOS remained similar to those of the PDC.

Building on the discussion of potential health inequities, it is equally critical to acknowledge the profound ethical considerations inherent in telemedicine implementation. These include ensuring clinical safety in remote diagnostic and treatment scenarios, establishing robust data protection measures to safeguard sensitive patient information, developing clear and comprehensive protocols for informed consent in virtual care settings, maintaining and assessing professional competency for health care providers operating remotely, and upholding principles of distributive justice to ensure equitable access and quality of care for all populations [40]. A previous literature review found that the ethical and legal issues related to the practice of telehealth or telemedicine services still need standard and specific rules of application in order to guarantee equitable access, quality of care, sustainable costs, professional liability, respect of patient privacy, and data protection and confidentiality. In fact, telemedicine services could be used only as complementary or supplementary to the traditional health care services and not as a complete substitute [41].

Strengths and Limitations

The main strength of this study is the inclusion of all eligible patients with COVID-19 at multiple public primary care clinics. This large sample size helped us ensure greater representativeness and reduced selection bias. Furthermore, we used a propensity score to match the two groups on factors that could potentially affect clinical outcomes, thereby minimizing confounding. Overall, these data analysis strategies contribute to the robustness of the findings and enhance the validity of the conclusions drawn.

This study also has several limitations. First, our participants were matched by age, smoking status, and existing comorbidities, which are known risk factors for severe COVID-19 [23]. Although various disease severity scores have been developed to predict outcomes in patients with COVID-19, such as the Quick COVID-19 Severity Index [42], the PRIEST COVID-19 Clinical Severity Score [43], the CHOSEN risk score for COVID-19 [44], and the 4C Mortality Score for COVID-19 [45], we did not use these severity scores to match our participants. This is because these scores were developed for hospitalized patients and were not applicable to our

XSL•FO
RenderX

outpatient management approach. Our study lacks detailed, standardized data on well-established risk factors for severe COVID-19, such as the BMI; the duration, type, and severity of symptoms at initial presentation; and the duration of disease at presentation. Crucially, patients' self-selected duration might have selectively chosen the PDC over the Tele-DC. As such, there may be residual confounding by indication that our matching process could not fully capture. Our study focused on mild patients, and those experiencing desaturation, respiratory distress, or symptoms of severe disease were advised to visit the AED for inpatient care. This unmeasured confounding remains a limitation.

Second, COVID-19 vaccination records were not included as baseline characteristics in this study due to data limitations. However, as of July 2022, ~94% of all Hong Kong residents had already received two doses of COVID-19 vaccines, with over 84% having received three doses [46]. Therefore, we have good reason to believe the vaccination rate between the two groups should be comparable. Third, a comprehensive socioeconomic status, including education levels and monthly incomes, was not included as a baseline variable in this study. These factors could be associated with smartphone usage rates and may have influenced the results. Fourth, there could be potential for misclassification bias arising from the use of ICPC-2 code A77 to identify COVID-19 cases. Because code A77 broadly encompasses other viral infections, such as Coxsackie virus; dengue fever; hand, foot, and mouth disease; and adenovirus, there is a possibility that patients without COVID-19 were inadvertently included in the cohort. Although we implemented strict mitigation measures, including restricting appointment bookings to self-reported confirmed COVID-19 cases and requiring nurses to verify the COVID-19 diagnostic method (PCR or RAT) during the initial assessment, the retrospective nature of the data means we cannot entirely rule out the chance of misclassification. Fifth, our study did not prospectively investigate the specific reasons for AED visits or DC revisits. Consequently, we were unable to definitively ascertain whether these visits were primarily driven by actual clinical deterioration, heightened patient anxiety, provider caution due to diagnostic uncertainty, or other factors (eg, non-COVID-19–related issues). This lack of detailed causal data limits our ability to distinguish between potentially avoidable visits and those that were clinically necessary, thereby obscuring clinically relevant distinctions regarding the precise drivers of utilization patterns.

Sixth, although doctors in both groups were drawn from the same pool of family physicians, the modality of consultation itself could have influenced clinical decision-making and treatment practices. The inherent information asymmetry in the Tele-DC, coupled with the heightened perceived risk and uncertainty during a pandemic, may have led to differences in antiviral prescription patterns, follow-up recommendations, or referral thresholds to emergency services. These unmeasured variations in physician practice, although not directly assessed in our study, could have subtly affected our observed outcomes, particularly the usage patterns. Seventh, we did not include a formal assessment of patient satisfaction with Tele-DC services. Although our findings shed light on clinical outcomes, understanding the patient experience is paramount for a comprehensive evaluation of telemedicine's effectiveness and acceptability. Patient satisfaction directly influences adherence to treatment, engagement with care, and overall perception of health care quality, all of which are critical for the successful and sustainable integration of telemedicine into routine practice. Future research should prioritize the inclusion of validated patient satisfaction measures to provide a more holistic understanding of telemedicine's impact from the end-user perspective, thereby informing strategies to optimize service delivery and improve patient engagement. Finally, our statistical approach used the chi-square test to compare binary outcomes over a fixed 28-day period. Although our centralized database ensured complete follow-up data for this window, this approach does not account for specific time-to-event dynamics or completing risks (eg, mortality) in the way that survival analyses such as Cox proportional hazards regression would. Future studies should incorporate time-to-event analyses to provide a more nuanced understanding of clinical trajectories.

Conclusion

In conclusion, this study demonstrated that Tele-DCs were associated with clinical outcomes comparable to face-to-face consultations regarding hospitalization and severe complications among patients with mild COVID-19 in a primary care setting. Distinct from the existing literature that often focuses on complex remote patient monitoring using home devices, our work is innovative in validating a scalable, consultation-based model that relies primarily on clinical history rather than physical data. This brings to the field critical evidence that a physical examination is not strictly necessary for the safe triage of mild respiratory infections, challenging the traditional reliance on in-person assessment. The real-world implications of these findings support a digital-first strategy for future infectious surges to conserve resources and minimize transmission. However, the observed associations with higher AED visit and lower antiviral prescription rates highlight that telemedicine cannot operate in isolation; future implementation must integrate remote patient-monitoring tools and robust logistical support to fully replicate the efficacy of conventional care.

Acknowledgments

This study was funded by the Hong Kong College of Family Physician Trainees Research Fund 2024 and the Kowloon Central Cluster Research Grant 2024.

XSL•FO
RenderX

Data Availability

The datasets generated and analyzed during the study are not publicly available as all the data are patients' data under the Hong Kong Hospital Authority, but they are available from the corresponding author upon reasonable request.

Authors' Contributions

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. Responsibility for the final manuscript lies entirely with the authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
ICD-9-CM and ICPC-2 codes of comorbidities included in the Charlson comorbidity index and ICD-9-CM codes of severe complications of COVID-19. ICD-9-CM: *International Classification of Disease, Ninth Revision, Clinical Modification*; ICPC-2: International Classification of Primary Care-2.
[DOCX File , 17 KB - jmir_v28i1e74046_app1.docx]

References

1. Du J, Lang H, Ma Y, Chen A, Qin Y, Zhang X, et al. Global trends in COVID-19 incidence and case fatality rates (2019-2023): a retrospective analysis. Front Public Health 2024 Jul 29;12:1355097 [FREE Full text] [doi: 10.3389/fpubh.2024.1355097] [Medline: 39135930]
2. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature 2020 Aug 08;584(7821):430-436 [FREE Full text] [doi: 10.1038/s41586-020-2521-4] [Medline: 32640463]
3. Ledesma JR, Chrysanthopoulou SA, Lurie MN, Nuzzo JB, Papanicolas I. Health system resilience during the COVID-19 pandemic: a comparative analysis of disruptions in care from 32 countries. Health Serv Res 2024 Dec 18;59(6):e14382 [FREE Full text] [doi: 10.1111/1475-6773.14382] [Medline: 39295092]
4. Hospital Authority adjusts designated clinics and general out-patient clinic service arrangements. The Government of the Hong Kong Special Administrative Region. 2022 Sep 8. URL: https://www.info.gov.hk/gia/general/202209/08/P2022090800399.htm [accessed 2026-05-06]
5. LCQ19: designated clinics and tele-consultation service under the Hospital Authority. The Government of the Hong Kong Special Administrative Region. 2023 Feb 15. URL: https://www.info.gov.hk/gia/general/202302/15/P2023021500405.htm [accessed 2026-05-06]
6. Farzandipour M, Nabovati E, Sharif R. The effectiveness of tele-triage during the COVID-19 pandemic: a systematic review and narrative synthesis. J Telemed Telecare 2023 Jan 23;30(9):1367-1375 [FREE Full text] [doi: 10.1177/1357633x221150278]
7. Jiang S, Gao X, Diao H, Zhang Y, Lu G, Liu X, et al. Clinical improvements from telemedicine interventions for managing type 2 diabetes compared with usual care: systematic review, meta-analysis, and meta-regression. JMIR Mhealth Uhealth 2026 Feb 18;14:e70429 [FREE Full text] [doi: 10.2196/70429] [Medline: 41707176]
8. Higashi RT, Etingen B, Lee J, Cole S, Mansour JC, Zhao A, et al. Patient and clinician perspectives on expanding telehealth use for older adults across the cancer control continuum: mixed methods study. JMIR Cancer 2026 Feb 09;12:e73058 [FREE Full text] [doi: 10.2196/73058] [Medline: 41662673]
9. Castillo-Rodenas M, Solanas Bacardit N, Farràs Company C, Miró Catalina Q, Solà Reguant L, Fuster-Casanovas A, et al. Feasibility, diagnostic accuracy, and satisfaction of an acute pediatric video interconsultation model in rural primary care in catalonia: prospective observational study. JMIR Pediatr Parent 2026 Jan 26;9:e82133 [FREE Full text] [doi: 10.2196/82133] [Medline: 41589566]
10. Bitar H, Alismail S. The role of eHealth, telehealth, and telemedicine for chronic disease patients during COVID-19 pandemic: a rapid systematic review. Digit Health 2021 Apr 19;7:20552076211009396 [FREE Full text] [doi: 10.1177/20552076211009396] [Medline: 33959378]
11. Ftouni R, AlJardali B, Hamdanieh M, Ftouni L, Salem N. Challenges of telemedicine during the COVID-19 pandemic: a systematic review. BMC Med Inform Decis Mak 2022 Aug 03;22(1):207 [FREE Full text] [doi: 10.1186/s12911-022-01952-0] [Medline: 35922817]
12. Wamsley C, Kramer A, Kenkel J, Amirlak B. Trends and challenges of telehealth in an academic institution: the unforeseen benefits of the COVID-19 global pandemic. Aesthet Surg J 2021 Jan 01;41(1):109-118 [FREE Full text] [doi: 10.1093/asj/sjaa212] [Medline: 32697289]
13. Cornelis J, Christiaens W, de Meester C, Mistiaen P. Remote patient monitoring at home in patients with COVID-19: narrative review. JMIR Nurs 2024 Nov 19;7:e44580 [FREE Full text] [doi: 10.2196/44580] [Medline: 39287362]

XSL•FO
RenderX

14. Robert P. The effect of a primary care remote patient monitoring program on emergency department usage in COVID-19. Ann Fam Med 2025 Sep;23((Suppl 1)):7689 [FREE Full text] [doi: 10.1370/afm.23.s1.7689]
15. Reed M, Huang J, Somers M, Hsueh L, Graetz I, Millman A, et al. Telemedicine versus in-person primary care: treatment and follow-up visits. Ann Intern Med 2023 Oct;176(10):1349-1357 [FREE Full text] [doi: 10.7326/m23-1335]
16. Venkatesh KP, Raza MM, Kvedar J. Has increased telehealth access during COVID-19 led to over-utilization of primary care? NPJ Digit Med 2022 Dec 13;5(1):178 [FREE Full text] [doi: 10.1038/s41746-022-00740-4] [Medline: 36513808]
17. Casariego-Vales E, Blanco-López R, Rosón-Calvo B, Suárez-Gil R, Santos-Guerra F, Dobao-Feijoo M, et al. Efficacy of telemedicine and telemonitoring in at-home monitoring of patients with COVID-19. J Clin Med 2021 Jun 29;10(13):2893 [FREE Full text] [doi: 10.3390/jcm10132893] [Medline: 34209725]
18. Shah VV, Villaflores CW, Chuong LH, Leuchter RK, Kilaru AS, Vangala S, et al. Association between in-person vs telehealth follow-up and rates of repeated hospital visits among patients seen in the emergency department. JAMA Netw Open 2022 Oct 03;5(10):e2237783 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.37783] [Medline: 36282505]
19. Chen K, Lodaria K, Jackson HB. Patient satisfaction with telehealth versus in-person visits during COVID-19 at a large, public healthcare system. J Eval Clin Pract 2022 Dec 22;28(6):986-990 [FREE Full text] [doi: 10.1111/jep.13770] [Medline: 36148479]
20. Charlson M, Peterson J, Szatrowski TP, MacKenzie R, Gold J. Long-term prognosis after peri-operative cardiac complications. J Clin Epidemiol 1994 Dec;47(12):1389-1400 [FREE Full text] [doi: 10.1016/0895-4356(94)90083-3] [Medline: 7730848]
21. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, STROBE Initiative. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. PLoS Med 2007 Oct 16;4(10):e297 [FREE Full text] [doi: 10.1371/journal.pmed.0040297] [Medline: 17941715]
22. Dehejia RH, Wahba S. Propensity score-matching methods for nonexperimental causal studies. Rev Econ Stat 2002 Feb;84(1):151-161 [FREE Full text] [doi: 10.1162/003465302317331982]
23. Underlying medical conditions associated with high risk for severe COVID-19: information for healthcare providers Jan 2023. Centers for Disease Control and Prevention. 2023 Jan. URL: https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-care/underlyingconditions.html [accessed 2026-05-06]
24. Statistics on 5th wave of COVID-19. Centre for Health Protection of the Department of Health and the Hospital Authority. 2023. URL: https://www.chp.gov.hk/files/pdf/5th_wave_statistics_20220704.pdf [accessed 2026-05-06]
25. Power (sample size) calculators. Sealed Envelope. URL: https://www.sealedenvelope.com/power/binary-noninferior/ [accessed 2026-05-06]
26. Shi Z, Mehrotra A, Gidengil CA, Poon SJ, Uscher-Pines L, Ray KN. Quality of care for acute respiratory infections during direct-to-consumer telemedicine visits for adults. Health Aff (Millwood) 2018 Dec;37(12):2014-2023 [FREE Full text] [doi: 10.1377/hlthaff.2018.05091] [Medline: 30633682]
27. Kim D, Chu H, Min BK, Moon Y, Park S, Kim K, et al. Telemedicine Center of Korean Medicine for treating patients with COVID-19: a retrospective analysis. Integr Med Res 2020 Sep;9(3):100492 [FREE Full text] [doi: 10.1016/j.imr.2020.100492] [Medline: 32802745]
28. Wang H, Yuan X, Wang J, Sun C, Wang G. Telemedicine maybe an effective solution for management of chronic disease during the COVID-19 epidemic. Prim Health Care Res Dev 2021 Sep 29;22:e48 [FREE Full text] [doi: 10.1017/s1463423621000517] [Medline: 34583801]
29. Bokolo Anthony J. Use of telemedicine and virtual care for remote treatment in response to COVID-19 pandemic. J Med Syst 2020 Jun 15;44(7):132 [FREE Full text] [doi: 10.1007/s10916-020-01596-5] [Medline: 32542571]
30. Oncu F, Balcioglu YH. Virtualization of mental health care in the midst of chaos: is telepsychiatry a silver lining? Psychiatry Neurol Sci 2021;34:219-222 [FREE Full text] [doi: 10.14744/DAJPNS.2021.00141]
31. Casey SD, Huang J, Parry DD, Lieu TA, Reed ME. Health care utilization with telemedicine and in-person visits in pediatric primary care. JAMA Health Forum 2024 Nov 01;5(11):e244156 [FREE Full text] [doi: 10.1001/jamahealthforum.2024.4156] [Medline: 39576615]
32. Tsai P, Lai W, Lin Y, Luo Y, Lin Y, Chen H, et al. Clinical manifestation and disease progression in COVID-19 infection. J Chin Med Assoc 2021 Jan 01;84(1):3-8 [FREE Full text] [doi: 10.1097/JCMA.0000000000000463] [Medline: 33230062]
33. Scott AM, Sanders S, Atkins T, van der Merwe M, Sunner C, Clark J, et al. The impact of telehealth care on escalation to emergency care: a systematic review and meta-analysis. J Telemed Telecare 2025 Sep;31(8):1059-1077 [FREE Full text] [doi: 10.1177/1357633X241259525] [Medline: 38839244]
34. Pedrosa AL, Bitencourt L, Fróes ACF, Cazumbá MLB, Campos RGB, de Brito SBCS, et al. Emotional, behavioral, and psychological impact of the COVID-19 pandemic. Front Psychol 2020 Oct 2;11:566212 [FREE Full text] [doi: 10.3389/fpsyg.2020.566212] [Medline: 33117234]
35. Tan LO, Ganapathy S. A single centre study of the level of parents' satisfaction with the COVID-19 telemedicine consultation. Eur J Pediatr 2024 Jan 20;183(1):213-218 [FREE Full text] [doi: 10.1007/s00431-023-05276-7] [Medline: 37861792]
36. Tan N, Yang LW, Tan MZ, Chng J, Tan MH, Tan C. Virtual care to increase military medical centre capacity in the primary health care setting: a prospective self-controlled pilot study of symptoms collection and telemedicine. J Telemed Telecare 2020 Oct 04;28(8):603-612 [FREE Full text] [doi: 10.1177/1357633x20959579]

XSL•FO
RenderX

37. Interim recommendation on clinical management of adult cases with coronavirus disease 2019 (COVID-19). HA Central Committee on Infectious Diseases and Emergency Response (CCIDER). 2023. URL: https://www.scribd.com/document/919287980/Interim-Recommendation-on-Clinical-Management-of-Adult-Cases-With-COVID-19 [accessed 2026-05-06]
38. Accorsi T, Moreira F, Pedrotti C, Amicis KD, Correia RFV, Morbeck RA, et al. Telemedicine diagnosis of acute respiratory tract infection patients is not inferior to face-to-face consultation: a randomized trial. Einstein (Sao Paulo) 2022;20:eAO6800 [FREE Full text] [doi: 10.31744/einstein_journal/2022AO6800] [Medline: 35649057]
39. Huang J, Gopalan A, Muelly E, Hsueh L, Millman A, Graetz I, et al. Primary care video and telephone telemedicine during the COVID-19 pandemic: treatment and follow-up health care utilization. Am J Manag Care 2023 Jan 01;29(1):e13-e17 [FREE Full text] [doi: 10.37765/ajmc.2023.89307] [Medline: 36716159]
40. Gajarawala SN, Pelkowski JN. Telehealth benefits and barriers. J Nurse Pract 2021 Feb;17(2):218-221 [FREE Full text] [doi: 10.1016/j.nurpra.2020.09.013] [Medline: 33106751]
41. Solimini R, Busardò FP, Gibelli F, Sirignano A, Ricci G. Ethical and legal challenges of telemedicine in the era of the COVID-19 pandemic. Medicina (Kaunas) 2021 Nov 30;57(12):1314 [FREE Full text] [doi: 10.3390/medicina57121314] [Medline: 34946259]
42. Haimovich AD, Ravindra NG, Stoytchev S, Young HP, Wilson FP, van Dijk D, et al. Development and validation of the Quick COVID-19 Severity Index: a prognostic tool for early clinical decompensation. Ann Emerg Med 2020 Oct;76(4):442-453 [FREE Full text] [doi: 10.1016/j.annemergmed.2020.07.022] [Medline: 33012378]
43. Goodacre S, Thomas B, Sutton L, Burnsall M, Lee E, Bradburn M, et al. Derivation and validation of a clinical severity score for acutely ill adults with suspected COVID-19: the PRIEST observational cohort study. PLoS One 2021 Jan 22;16(1):e0245840 [FREE Full text] [doi: 10.1371/journal.pone.0245840] [Medline: 33481930]
44. Levine DM, Lipsitz SR, Co Z, Song W, Dykes PC, Samal L. Derivation of a clinical risk score to predict 14-day occurrence of hypoxia, ICU admission, and death among patients with coronavirus disease 2019. J Gen Intern Med 2021 Mar 03;36(3):730-737 [FREE Full text] [doi: 10.1007/s11606-020-06353-5] [Medline: 33274414]
45. Knight SR, Ho A, Pius R, Buchan I, Carson G, Drake TM, ISARIC4C investigators. Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: development and validation of the 4C Mortality Score. BMJ 2020 Sep 09;370:m3339 [FREE Full text] [doi: 10.1136/bmj.m3339] [Medline: 32907855]
46. Statistics on government COVID-19 vaccination programme: statistics as at 30 April 2026. Centre for Health Protection of the Department of Health, Government of the Hong Kong Special Adminitrative Region. 2026 May 5. URL: https://www.chp.gov.hk/en/features/106989.html [accessed 2026-05-07]

Abbreviations

AED: accident and emergency department
CDARS: Clinical Data Analysis and Reporting System
CMS: Clinical Management System
CSSA: comprehensive social security assistance
DC: designated clinic
ED: emergency department
GAI: generative artificial intelligence
GOPC: general outpatient clinic
HA: Hospital Authority
HDU: high-dependency unit
ICD-9-CM: *International Classification of Disease, Ninth Revision, Clinical Modification*
ICPC-2: International Classification of Primary Care-2
ICU: intensive care unit
LOS: length of stay
PCR: polymerase chain reaction
PDC: physical designated clinic
RAT: rapid antigen test
Tele-DC: tele-designated clinic

XSL•FO
RenderX

Edited by S Brini; submitted 17.Mar.2025; peer-reviewed by YH Balcioglu, Z Guo; comments to author 12.Sep.2025; revised version received 13.Apr.2026; accepted 14.Apr.2026; published 14.May.2026.

Please cite as:
Jiao F, Ho KM, Chiang L, Ko SH, Chen CX
Effectiveness of Telemedicine vs Face-to-Face Consultation in Fighting COVID-19: Retrospective Cohort Study of Adult Patients With COVID-19 in a Primary Care Setting
J Med Internet Res 2026;28:e74046
URL: https://www.jmir.org/2026/1/e74046
doi:10.2196/74046
PMID:

©Fangfang Jiao, Ka Ming Ho, Lapkin Chiang, Siu Hin Ko, Catherine Xiaorui Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Virtual Consultations for People With Intellectual Disabilities in General Practice and Community Care: Mixed Methods Qualitative Study

Freda Mold[1], BSc, PhD; Anna Cox[1], BSc, MSc, PhD; Vicki Tsianakas[2], PhD; Harm van Marwijk[3], BMBS, PhD; Paul Shanahan[4], PhD; Treena Parsons[1], DipSW, BSc, MSc, PhD; Jo Armes[1], RGN, BSc, MSc, PhD

[1]School of Health Sciences, Faculty of Health & Medical Sciences, University of Surrey, Guildford, United Kingdom

[2]Florence Nightingale Faculty of Nursing, Midwifery & Palliative Care, King's College London, London, United Kingdom

[3]Department of Primary Care and Public Health, Brighton and Sussex Medical School, Brighton, United Kingdom

[4]Your Healthcare CIC, Surbiton, United Kingdom

Corresponding Author:
Freda Mold, BSc, PhD
School of Health Sciences
Faculty of Health & Medical Sciences
University of Surrey
Kate Granger Building
Priestly Road, Surrey Research Park
Guildford, GU2 7YH
United Kingdom
Phone: 44 01483 684636
Email: Freda.mold@surrey.ac.uk

Abstract

Background: Virtual consultations (VCs) using video or telephone were embraced at speed in general practice (GP) and community care during the COVID-19 pandemic. People with intellectual disabilities, their families, and support workers (SWs), along with health care professionals (HCPs), had to adapt quickly to this change in provision, but little is known about how this new way of working was experienced.

Objective: This study aims to explore the views and experiences of people with intellectual disabilities, their families, SWs, GP, and community care professionals on the quality and safety of VCs.

Methods: This paper reports on users' experiences of VCs, as part of a larger Experience-Based Co-design study. This paper relates to 2 stages of data collection. Observed video consultations in GP and community care (n=3), and semistructured interviews with people with intellectual disabilities, their family members or SWs, GP, and community care professionals (n=34). The 30-month study was conducted from November 2021 ending in April 2024. Data were analyzed using framework analysis.

Results: Integrated results are presented through 5 themes, encompassed under an overarching theme of safety and quality. The five themes highlight critical factors in planning, delivery, and aftercare of VCs in GP and community services for people with intellectual disabilities in the United Kingdom: (1) context, space, and purpose—showing the importance of safe spaces to talk, and having clear consultation objectives and purpose; (2) choice—facilitating choice over time about modality of health care contact; (3) familiarity, online relationships, and trust—the building blocks for quality consultations; (4) prepare and personalize—to ensure that HCPs are aware of reasonable adjustments, and recognition of caregiver involvement; and (5) continued connection—where patients or families are offered continued contact with a named or same HCP enhancing access to regular or ongoing care. All participants were aware of the limitations of VC, which may impact safety, such as gaps in home monitoring due to the absence of appropriate equipment or recording, inability to identify vital risk indicators, and limited field of vision on screen. However, participants were also aware of the distinct benefits they offer in terms of quality provision, such as timeliness of care, building and sustaining comfortable relationships, support for more frequent attendance, and continuous connection to health teams.

Conclusions: VCs offer an opportunity to improve digital inclusion in health care for people with intellectual disabilities. However, the quality and safety of VCs for this population are dependent on continuous review of patients' needs over time and ensuring that their choices and preferences are considered when planning and providing ongoing care.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e81173) doi:10.2196/81173

KEYWORDS

access to care; patient safety; general practice; community care; digital inclusion; information and communication technologies; remote consultations; video conferencing; intellectual disabilities; virtual consultations

Introduction

The post–COVID-19 general practice (GP) and community health care landscape has evolved significantly [1], with digital access now a key priority in health care systems, including the National Health Service (NHS) [2-4]. Recent guidance includes a range of online services, for example, online appointment booking, test results and record access, online prescription ordering, and telephone and video consultations [4].

Online service provision, such as virtual consultations (VCs) in GP and community care, has brought into sharp focus how health care provision might be adapted using new technologies to support widening access to diverse patient groups [5,6]. However, online services have potential limitations for both patients and service providers. Patients face challenges related to access, affordability, and the skills or support required, which may limit their ability to navigate and use these services effectively [7-9]. For health care professionals (HCPs), providing online services can impact their confidence and skills in using different modes of remote care delivery [10]. At an organizational level, there is a need to adhere to local governance or NHS Trust guidance, which may fluctuate over time [11,12].

Although there is good-quality research outlining the use of various digital platforms for people with intellectual disabilities to support social contacts [13], there is little evidence about whether and how they use technology to manage their health. While the impact of video or telephone consultations (VCs) has been widely studied, their effects on specific patient groups, such as people with intellectual disabilities, remain less well understood [14].

In this paper, we use the term VCs, as it is favored by our study participants, community partners, patient and public involvement and engagement (PPIE) contributors, and the Experts by Experience (EbE) group. This term represents both telephone and video modes of consultation to assist any aspect of health care (eg, to support an annual health check), which complement in-person care delivery.

During COVID-19, people with intellectual disabilities, their families, and support workers (SWs) experienced a reduction and change in health care provision [8]. These compounded challenges encountered when accessing health care and exacerbated existing health inequalities experienced by people with intellectual disabilities [15,16]. Since COVID-19, the challenge has been to embed the use of technologies more widely despite having fewer resources and needing to work differently. With greater reliance on technologies to speak to HCPs, there is a need to explore the experience and views of people with intellectual disabilities who might struggle more with virtual service provision. While several UK policies highlight the need to improve uptake of key services for this group, such as annual health checks [2] and digital access [17], greater understanding is required about whether and how VC may widen opportunities for care access and integration across GP and community care providers.

The overall aim of the Experience-Based Co-Design (EBCD) study was (1) to explore the views and experiences of VC from the perspective of people with intellectual disabilities, their families, SWs, GP, and community care professionals and (2) to coproduce practical resources and guidelines to support the access and use of VCs. In this paper, we report findings for objective 1, users' views and experiences of VC, as related to quality and safety. The development of the co-designed resources from objective 2 will be reported in a subsequent paper.

Methods

Early Project Design

Our study developed from early PPIE work, comprising a series of meetings over 24 months and a brief survey conducted in 2020. This accessible survey of people with intellectual disabilities explored how they talk to HCPs and their views of interacting via video or telephone calls. Findings from the meetings and survey helped shape and refine the aims of this study and informed the study design, ensuring that our methods were accessible and the purpose was meaningful to participant groups.

Design

We adopted an EBCD research design [18]. This participatory and collaborative method promotes service-user involvement to drive health care quality improvement [19,20]. Using participatory, narrative, and visual methods, it provides a practical and collaborative way to engage people with intellectual disabilities [21]. This method also aligns with national guidance about how services should be planned and delivered, emphasizing coproduction with people with intellectual disabilities, their families, and SWs [22]. In keeping with inclusive research principles, we established an EbE group to work with us throughout the study to help inform the accessibility of participant information, advise on recruitment strategies, and support data analysis (among other key tasks) to meet study milestones. This group comprised 9 people with intellectual disabilities and/or family members of people with intellectual disabilities.

The overall EBCD project comprised five sequential stages: (1) rapid review of existing guidance; (2) nonparticipant observation of video consultations in GP and community care; (3) qualitative semistructured interviews with people with intellectual disabilities, their family members or SWs, and GP and community care professionals; (4) production of narrative catalyst film, based on interview video clips; and (5) series of

XSL•FO
RenderX

co-design and priority-setting events and meetings with participants. Stages 2 and 3 were undertaken concurrently. In this paper, we focus on data collected during observations (stage 2) and interviews (stage 3). The film production and co-design process and outputs will be reported elsewhere. Figure 1 shows the time frames associated with each stage.

Figure 1. Time frame of each EBCD study method. EBCD: Experience-Based Co-Design.

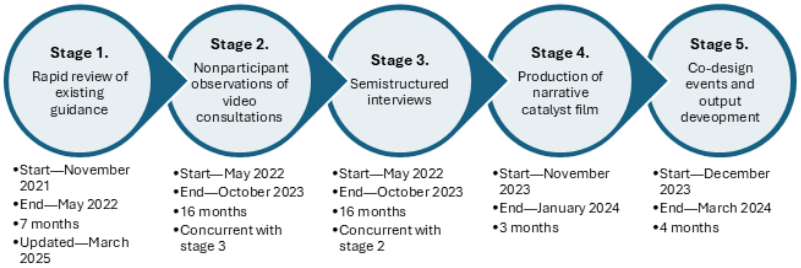

Sample and Recruitment

Participants included adults with intellectual disabilities, their family members or SWs, and HCPs recruited through GPs, community teams, and third-sector organizations across the Southeast of England. A purposeful maximum variation sampling strategy [23] was used to ensure participation from across a large geographical area (Kent, Surrey, Sussex, and South London) and diversity according to specific demographic variables, including gender, ethnicity, and locality. Eligibility criteria for people with intellectual disabilities included being ≥18 years and having a formal diagnosis or self-identifying as having mild to moderate intellectual disabilities. We extended eligibility criteria to also include family members or SWs for individuals with profound or severe intellectual disabilities due to their critical role in managing the person's health. Participants could be supported throughout the study by a family member or SW. Participants needed internet access or a telephone (mobile or landline) connection. The involvement of family members and SWs was essential, particularly for participants with profound disabilities and those living in supported accommodation. Two people with intellectual disabilities consented to participate in the study but withdrew due to being unavailable for the interview.

Multiple recruitment strategies were used. For patients and families, recruitment was conducted via word of mouth, posters and flyers to our community intellectual disability partners, attendance and presentations at Intellectual Disabilities Communities of Practice meetings, and project notifications via their online or printed newsletters. Posters were also displayed in GP surgeries aimed at patients and their families or caregivers. Recruitment was also supported by word-of-mouth project introduction by HCPs in GP sites and via the research team's professional network to promote the study. Audio and easy-read information were used throughout to support diverse recruitment.

HCPs were recruited via newsletters and support from primary care networks and clinical research networks. Presentations at community or local authority caregiver meetings were undertaken, followed by online meetings for professionals working in supported living or daycare facilities. Physical posters were also used in GP settings.

Eligible professionals were involved in the delivery or organization of health care services, including VC and annual health checks, within GP or community care teams. All participants resided or worked within Southeast England or South London, an area covering 4 integrated care boards or NHS trusts. Clinical sites' withdrawal was due to misalignment of the recruitment period with work demands (COVID-19 and seasonal flu or pneumonia vaccine rollouts; 1 GP site), poor timing (1 supported living home), limited capacity to take part in research (1 community team), and reduced research and design (R&D) capacity combined with needing to maintain business continuity (1 community team).

Data Collection

Data were collected from May 2022, with the entire project ending in 2024. All observations and interviews were conducted online using audio and visual recordings by an experienced social care researcher (TP). Adults with intellectual disabilities, their family members, and SWs were based either in their family home or a supported living home. HCPs participated from their workplace. Interview topic guides were informed by existing literature, our previous PPIE work, co-designed with our community partners, and piloted with our EbE group. We offered a flexible approach to project participation, with participants choosing to take part in 1 or more stages of data collection (ie, observation, interview, and co-design). Participants were unknown to the research team prior to the commencement of the study. Data were collected using Microsoft Teams, and all recordings were uploaded into a Trusted Research Environment [24,25].

Ethical Considerations

Project information packs, comprising project information and consent forms, were available in various accessible formats, including easy-read, audio, and printed paper. Seven types of project information packs were developed to best suit our different participant groups (patients, family members, SWs, and HCPs) and to ensure that we addressed diverse participant groups' accessibility needs. We obtained written and oral consent at each stage. Efforts undertaken to maintain participants' privacy included anonymization of transcripts, removal of identifying information, secure storage, and access to data using a Trusted Research Environment secure server. We gained ethics approval from the NHS Health Research Authority (HRA reference: 303071) and the University of Surrey

XSL•FO
RenderX

(reference: 2021/15 FHMS). Our EbE group and co-design participants were compensated for their contribution to the project, with interview and observation participants generously volunteering their time.

Nonparticipant Observation

Two of the research team (FM and TP) members drove the observation recruitment. Once potential participants contacted us, we posted or emailed out the relevant project information pack. Once we received consent, we contacted the patient, family member or SW, or HCP to answer any questions and talk about the logistics of the nonparticipant observation. Questions tended to be about how the observation would work, the time taken, what would happen in the consultation, and what would happen at the end. Participants who were willing to proceed were then invited to a new meeting to test their laptop or phone connectivity and other device functions (camera, sound, etc) being used for the observation. This additional meeting was optional, but all participants chose to meet with us, as they found it reassuring ahead of the observation. The researcher (TP) was present at the start of the VC to introduce themselves and remind them that the VC would be recorded. The recording then started, and the researcher withdrew from the VC, leaving the HCP, patient, and SW to conduct the consultation. An observation schedule was developed and used after observation to ensure consistent recording of field notes (available on request). The schedule was informed by key concepts such as communication style, as outlined by Roter and Larson [26] and Miller and Nelson [27], and constructs from the Consolidated Framework for Implementation Research [28]. The schedule comprised 97 topic areas, which were evidenced as being important in health consultations, such as the consultation environment, technical support, consultation purpose or aim, communication, and patient-health care staff relationship. The researcher also made observations and reflection notes when reviewing consultation data. The goal of the observation was to observe real-world consultations to better understand the experiences of VC participants, gaining insights into the process by which VCs are planned, conducted, and followed up, and the interpersonal and relational impact. By viewing the VC, the researchers would also gain a better understanding of participants' experiences, as reported in subsequent stages of the study.

Three observed consultations took place between HCPs and patients, with a family member or SW if preferred. Although we initially aimed to observe up to 12 consultations, practical constraints, such as extended recruitment and preparation timelines, led us to reassess our approach. Following 3 completed observations, we determined that sufficient data had been gathered to support meaningful interpretation. Consequently, we discontinued further recruitment for observation. In total, 10 participants were involved in these 3 consultations. Observations lasted between 20 and 75 minutes and were all video consultations.

Interviews (n=34) were supported by an interview schedule comprising 17 questions across 7 topic areas, focusing on the context of VCs, likes and dislikes of VCs, access and choice, relational and communication, safety and quality, workload and VC service fit, and finally, future use and advice for people thinking of using VCs. Table S1 in Multimedia Appendix 1 outlines the interview schedule topics and example questions. Interview schedules varied depending on whether the participant had taken part in the VC observation, used VCs before, or by their role (patient, family member, SW, or care professional). Postinterview field notes were completed to record any researcher observations and reflections. Interviews lasted 20-90 minutes. Two patients chose to be supported by an SW at the interview.

Our target interview number across geographical areas and participant groups was 36. Although we had arranged the remaining 2 interviews, these were not conducted, as interviewees became unresponsive after they requested to reschedule their interview date. Toward the end of the interview period, it became evident that no new themes or insights were emerging. Consequently, we determined that we had collected sufficient interview data to support robust and meaningful analysis.

Data Analysis

Analysis was conducted in several stages. First, the data were analyzed separately according to the data collection method. Observational data comprising field notes and observation schedule data were tabulated to become familiar with the data across observations, followed by inductive and deductive coding. These codes were informed by key communication styles [26,27] and other frameworks [28] and brought together with the interview data at the indexing stage.

Framework analysis was used to analyze interview data, supported by NVivo (version 14; QSR International Pty Ltd). Framework analysis consists of 5 analytic stages (familiarization, identifying themes, indexing, charting, and mapping and interpretation) [29,30]. Interview data were transcribed verbatim, and inductive and deductive codes were attributed to segments of data, staying close to the context and participants' meaning. We created codes based on the emerging data, but we were also mindful of key frameworks [26-28]. Two researchers (TP and FM) conducted independent coding, after which they met daily for 1 week to agree on codes, descriptors, and index codes, which shared commonalities [29,30]. Charting, where data were rearranged or clustered, to create order, was undertaken by the same researchers in a series of meetings before finalizing at 2 consensus meetings with the wider research team (AC, VT, JA, PS, and HM). These meetings were used to map and begin interpreting the data, focusing on quality and safety issues across different time points (before, during, and after a VC), and examining whether the data were concordant, discordant, or complementary. Over 219 codes were initially created, which were refined to 25 themes and subthemes.

Data Integration and Theoretical Frameworks

Themes from observation and interview data were tabulated across participant groups (patient, family member or SW, and HCP) and at key time points in the VC (before, during, and after). The data triangulation method was used to identify recurring, complementary, contrasting, and important overarching or cross-cutting themes [23]. Integrated data were

XSL•FO
RenderX

presented as a bubble figure, aiding data visualization when discussed and refining data interpretation by the wider research team meetings (FM, PS, AC, HM, JA, and VT). At this stage, we refined themes constructed from our interpretations of data in 3 online meetings with our EbE. Data integration identified several overarching themes emerging from the analysis. The first of which was quality and safety, as it was the primary concern reported by participants. The mapping and interpretation of quality and safety data were synthesized into 5 subthemes. The study is reported in accordance with the COREQ (Consolidated Criteria for Reporting Qualitative Research) framework) [31].

Results

Participant and Care Setting Characteristics

Three VC observations were undertaken, with 10 participants. A total of 34 interviews were conducted, and of these, 22 were with a range of HCPs in GP and community care settings (Table 1).

Table 1. General practice and community care professions and care setting (n=22).

Characteristics	Care setting		
	Primary care or general practice (n=5), n (%)	Community care teams (n=13), n (%)	Supported or care professional (n=4), n (%)
Gender			
Men	—[a]	—	1 (25)
Women	5 (100)	13 (100)	3 (75)
Professions			
General practitioner	1 (20)	—	—
Practice nurse	1 (20)	—	—
Intellectual disability liaison nurse	1 (20)	—	—
Care coordinator	2 (40)	—	—
Intellectual disability community matron	—	1 (8)	—
Speech and language therapists	—	3 (23)	—
Specialist nurse	—	2 (15)	—
Physiotherapist or lead	—	3 (23)	—
Clinical psychologist	—	1 (8)	—
Clinical psychiatrist	—	1 (8)	—
Community care professionals (coordinators and managers)	—	2 (15)	—
Home managers	—	—	3 (75)
Care professional	—	—	1 (25)
Geographical			
Kent (n=7)	2 (40)	4 (31)	1 (25)
Surrey (n=8)	2 (40)	4 (31)	2 (50)
Sussex (n=5)	1 (20)	3 (23)	1 (25)
Out of area (n=2)	—	2 (15)	—

[a]Not applicable.

Patient and family participants comprise people with mild to moderate intellectual disabilities (n=6), family members, including those of people with profound intellectual disabilities (n=2), and SWs (n=4; total n=12). Patient, family, and SW interview demographics are outlined in Table 2. Patients were equally split between men and women. SWs (n=4) worked in supported living homes (n=2) or day centers (n=2). All SWs were women. Family members were a parent of adult patients and a brother. Patients lived in a range of homes, independently in their own home (n=1), family home (n=2), and supported living homes (n=3).

XSL•FO
RenderX

Table 2. Patient or family and support worker interview characteristics (n=12).

Characteristics	Participant		
	People with intellectual disabilities (n=6)	Family members (n=2)	Support workers (n=4)
Gender, n (%)			
Men	3 (50)	1 (50)	—[a]
Women	3 (50)	1 (50)	4 (100)
Age range (years)	26-68	—	—
Residence or work setting, n (%)			
Home (with family)	2 (33)	—	—
Home (independent)	1 (17)	—	—
Supported living home	3 (50)	—	2 (50)
Day care service	—	—	2 (50)
Relation, n (%)			
Parents of adults with profound intellectual disabilities	—	1 (50)	—
Brother	—	1 (50)	—

[a]Not applicable.

Summary of Overarching Themes

Overall, a mixed picture emerged from our analysis. There was agreement across all participant groups that the best practice care delivery for people with intellectual disabilities was in-person appointments. However, under certain conditions, VCs may complement existing health care services, particularly for patients who experience significant stress or discomfort during in-person appointments. As an HCP said, “I think it [VCs] will complement. I don’t think it can replace” (D027 HCP).

Five themes were identified, which encompass key considerations for conducting high-quality VC safely: (1) context, space, and purpose; (2) choice; (3) familiarity, online relationships, and trust; (4) prepare and personalize; and (5) continued connection. Key themes and the associated subthemes are shown in Table 3. Illustrative quotes are presented in Table S2 in Multimedia Appendix 1.

XSL•FO
RenderX

Table 3. Key themes and subthemes by participant group.

Participant group	Key insights
Theme 1: context, space, and purpose	
Subtheme 1: home and work context	
People with intellectual disabilities	• Privacy to talk • Safe and familiar physical space
Family members and support workers	• Privacy to talk • Agency and choice to leave or stay in the room • VC[a] as a tool to improve access and confidence
HCPs[b] (GP[c] and community)	• Knowing who is in the room and privacy • Missing vital signs • Ability to control who is in the room and agency • Appropriate workspaces
Subtheme 2: urgency, pain, and purpose	
People with intellectual disabilities	• Deciding urgency • Purpose of the appointment
Family members and support workers	• Assessing urgency, pain, and timeliness • Considering travel and transportation
HCPs (GP and community)	• Assessing urgency and pain • Importance of timely access and clear purpose • Diagnostic challenges (eg, overshadowing)
Subtheme 3: physical checks	
People with intellectual disabilities	• Medication review or understanding • Seeking help when worried
Family members and support workers	• Role ambiguity or drift and delegation • Hybrid consultation, home equipment, and perceived risk
HCPs (GP and community)	• Complement in-person care and limitations • Hybrid consultation
Theme 2: choice	
People with intellectual disabilities	• Support from family or support workers in practical tasks • Facilitating comprehension and support • Importance of choice and autonomy
Family members and support workers	• Flexibility around patient and family preferences • Confidence in using technology
HCPs (GP and community)	• Responsiveness to patient preferences and service flexibility • Confidence in technology use
Theme 3: familiarity, online relationships, and trust	
People with intellectual disabilities	• Developing familiarity and unique communication styles • Building trust
Family members and support workers	• Building familiarity • Building trust • Confidence in what is being seen and heard • Relationships with HCPs

XSL•FO
RenderX

Participant group	Key insights
HCPs (GP and community)	• Building familiarity and individual communication • Building trust • Trust in what is being seen and heard • Foundations for ongoing contact • HCPs' relationships with families
Theme 4: prepare and personalize	
People with intellectual disabilities	• HCP preparedness, including awareness of reasonable adjustments • Importance of HCPs' understanding communication styles (verbal and nonverbal)
Family members and support workers	• HCP preparedness, including frustrations over the lack of awareness of reasonable adjustments • Importance of HCPs' understanding varied communication styles (verbal and nonverbal)
HCPs (GP and community)	• Presence of family members during VCs • Need to establish contingency planning • Value of collaborative learning within care teams
Theme 5: continued connection	
People with intellectual disabilities	• Feeling connected • Loss of connection and quality • Avoiding interruptions • Post-VC information in accessible formats
Family members and support workers	• Ensuring continuity • Connecting long-distance families or support workers • Mixed experiences of connection and community • Minimizing interruptions • Helpfulness of post-VC information
HCPs (GP and community)	• Ensuring continuity, including service ethos • Importance of family members or support workers in maintaining connection • Potential for fragmented care, the unintentional consequence of VCs • Consulting space and interruptions • Importance of post-VC information

[a]VC: virtual consultation.

[b]HCP: health care professional.

[c]GP: general practice.

Theme 1: Context, Space, and Purpose

Consideration of the patient's context, such as their home environment and medical needs, and the purpose of the consultation, including whether any physical checks were required, was reported to be essential in determining the quality and safety of conducting a consultation virtually, rather than in-person. Context was seen as critically important at all stages of VC planning, delivery, and follow-up.

Home and Work Context

For HCPs, prior knowledge of the patient's home context (ie, shared, independent, or family homes) and whether the person had the necessary support (eg, family, spouse, sibling, or SW) to use video or telephone devices was the first step for determining the appropriateness of a VC.

HCPs recognized the potential benefits of VCs, but they reported the need to continuously weigh these against possible risks. HCPs raised concerns about their limited awareness or control over who was in the virtual "room."

> *... we've certainly had situations where we haven't been clear, who's been in the room when we've been doing virtual consultations and we've not been told that there's like a whole bunch of other people sitting on the sofa listening. And that made us think like, Ohh, actually we need to specifically ask, like, who's in the room because it just felt very odd to think that there were people that were listening but not contributing to a conversation.* [E007 HCP interview]

HCPs, therefore, encouraged everyone participating in a VC to introduce themselves (if unknown) and to be visible on screen. They were also concerned about missing vital signs or

XSL•FO
RenderX

identifying risks, such as bad smells (urine), or client neglect (not dressed or unclear clothes). HCPs reported drawing on existing knowledge and prior experience of risk factors within the patient's home context to inform consultation planning and decisions about the appropriateness of offering a VC.

People with intellectual disabilities, their families, and SWs emphasized the value of a VC occurring in a physical space, which facilitated feeling safe and a sense of privacy so they could talk openly. These home environments were often described as quiet and uncluttered spaces. However, HCPs questioned whether patients, or even themselves, had access to appropriate meeting spaces to meet virtually. HCPs reported sometimes working in dark, nonsoundproofed rooms, or workspaces that had unexplained issues, such as random, frequent alarms. They also worried about inadvertently screen-sharing other patients' personal details, if using multiple screens, during the appointment, which could impact privacy.

Staff also reported factors related to the sense of agency when conducting VC with patients in their home setting. HCPs spoke about their limited ability to control who was present during consultations and which rooms the VCs were conducted in. Limited control over who was present also limited HCPs' feelings that they could request parents of adults with intellectual disabilities to leave the room when necessary. HCPs also mentioned the "limited return" of VCs, observing that patients could easily walk away during consultations and choose not to return. Walking away was less common during in-person consultations. While HCP generally viewed the presence of family members or SWs during VCs as beneficial, they did note that in certain situations, such as when parents were present, it could hinder open discussions of sensitive topics, including sexual health with young adults.

For people with intellectual disabilities and their families or SWs, being able to leave a consultation was sometimes experienced positively, giving them a greater sense of agency and control. This raises questions about whether it is easier or harder to "leave the room" in an in-person versus a VC appointment and suggests whether contingency planning is needed, in case patients "opting out" of virtual appointments altogether. As such, the rules of engagement were more challenging to navigate online.

HCPs also questioned whether VCs empowered or disengaged patients, particularly when multiple participants attended, such as in multidisciplinary care planning meetings, or when rooms were unsuitable for VCs. For SWs, VC may offer patients greater autonomy, and with this comes the autonomy to disengage from appointments. In contrast, SWs reported feeling more confident in VCs, rather than in in-person appointments, when talking to clinical staff and other agency workers:

> *... it's easier for me to sort of like bounce ideas and stuff rather than being in a room full of people where I'd be a bit, maybe a bit more quiet.* [B015 SW interview]

This was because they were in a familiar environment and were comfortable with working online. This suggests how VC can impact professional dynamics, with some professionals feeling more empowered to share their views virtually.

For people with intellectual disabilities, VCs were sometimes preferred, with family members reporting how VCs were beneficial when patients were too nervous to leave the home for in-person appointments. Family members also reported how VCs may appeal to the heterogeneity of people with intellectual disabilities across age groups, widening their potential appeal.

In summary, several key factors impacted the perceived safety of VCs conducted in the home, and these include having an appropriate quiet, usable virtual space, consideration about who would be in attendance, and all participants' confidence (patients, families, SWs, and HCPs) when talking online or using technology.

Urgency, Pain, and Purpose

The timeliness and urgency of the appointment (urgent or nonurgent routine) were critical in assessing whether an in-person or virtual appointment was needed. Decisions were influenced by factors such as pain levels, type of service needed (ie, speech and language therapy and general practitioner), and travel logistics (eg, cost, distance, and mode of transport). Families or SWs played a crucial role in recognizing when someone was in pain, further shaping these decisions. HCPs recognized the importance of assessing urgency or pain levels to determine whether virtual or in-person care was required. In making these decisions, many HCPs highlighted diagnostic overshadowing, where symptoms can be misattributed to an existing diagnosis, such as an intellectual disability, leading to missed or delayed care [32-34]. Diagnostic overshadowing was a significant challenge, particularly in virtual settings, where identifying pain was more difficult.

> *They [carers] often say this thing of "they've got a high pain threshold" which all it means is that the adult can't express their pain. They're still feeling the pain, they're just not able to express it. Whereas sometimes, if you'll actually see the adult, you can see if they move differently or their guarding themselves, you can see they're probably in pain, but you can't really get that virtually all the time.* [E006 HCP interview]

Missing expressions of pain emerged as a primary concern for HCPs and SWs alike. Identifying the signs of pain depended on HCPs' knowledge of the patient and their unique characteristics, health history, and how they express discomfort.

Confidence in decision-making relied on recognizing individual pain expressions and determining whether escalation was needed. In such circumstances, telephone and video appointments were reported to be more challenging, as greater emphasis was placed on HCPs' ability to assess what is seen or unseen in virtual appointments. None of the HCP participants mentioned using established tools to identify pain [32].

Although VCs were seen as more challenging to conduct for HCPs, they offered advantages in delivering timely care. While physical checks required in-person visits, the urgency of appointment decisions was weighed against appointment

XSL•FO
RenderX

availability, travel feasibility, and pain levels. Further risk considerations arose when patients were known to underreport key symptoms. In this situation, HCPs, families, and SWs highlighted the need to be familiar with the patient to assess urgency and for HCPs to offer prompt appointments when needed. Prior knowledge of a person's typical responses was key in distinguishing visible and subtle signs of pain. This was considered by HCPs as central to identifying pain or discomfort and/or signs of neglect or abuse. Being familiar with the patient helped, especially if SWs or HCPs were concerned about the patient underreporting pain to avoid treatment or in-person care.

Having a clear purpose for the VC was important for people with intellectual disabilities. In this study, we identified several practical uses of VC, including pre- and postcare information gathering, medication checks, imparting nonurgent test results, and care planning meetings before and after annual health checks. VCs were also used to support different types of appointments, including those with speech and language therapy, physiotherapy, and neurology. HCPs and SWs reported VC as being particularly useful for people with a history of nonattendance and preexisting conditions, including those who are housebound, have phobias, are immunocompromised, and/or have challenging behavior. Some of these issues were also relevant for family members, with one noting that being immunocompromised themselves made in-person attendance a less favorable option. People with intellectual disabilities took part in VCs for a range of activities, such as a medication review or follow-up on a new treatment. HCPs and SWs also worked together to collate physical measurement (ie, weight) data when in-person appointments were considered too stressful for patients or when travel or transport barriers made attendance difficult.

In summary, assessing urgency and identifying pain were the main factors influencing HCPs' and SWs' views of VC. These factors were important in understanding the quality of VCs. Assessing urgency requires being familiar with the patient and knowing their signs of distress. For people with intellectual disabilities, having a clear understanding of the VC's purpose was paramount, preferring in-person appointments if the conversation was sensitive, serious, or complex.

Physical Checks

There was some agreement between all participant groups about how and when VC could be used to support physical checks. As noted earlier, acceptance of use was initially based on the topic under discussion. Mental and sexual health topics were generally considered unsuitable for VCs, as these were seen as potentially triggering, unsafe, or inappropriate when talking to patients, especially when families or SWs were present. HCPs reported specific activities that were successfully undertaken online, such as information gathering before annual health checks, medication reviews, or as a stand-alone VC to follow up on nonurgent test results, check medication or equipment delivery or use, or update on next steps in care planning. HCPs also used VC to prepare for in-person appointments, but only when needed.

There were mixed views about VC use to support activities such as medication review. While most HCPs favored VC to support medication reviews, people with intellectual disabilities often found phone consultations challenging, particularly due to difficulties in understanding information clearly.

> *Sometimes I can't understand, the pharmacist [on the telephone], they've got a pharmacist there, which I don't understand .... With the doctor it's OK, because I understood. But with the pharmacist querying the medication, I couldn't. Don't know. I couldn't hear him properly ....* [D025 patient interview]

People with intellectual disabilities spoke about needing prior knowledge about the topic to be discussed in the VC before deciding whether an in-person or a VC was best. For patients, in-person appointments were preferred if a physical check was needed or if they were worried about something.

In relation to physical checks, VCs were reported to sometimes blur boundaries between the roles of SWs, HCPs, and family members. For example, a researcher noted in one observation how a practice nurse asked a female SW if a breast examination had previously been undertaken. When it had not, the nurse then asked if the SW could assist the patient in performing the examination. The SW appeared reluctant, as the patient was known to be uncomfortable with being touched. The researcher's observation notes recorded the SW's facial expressions, showing discomfort and frowning to show a lack of agreement with the request. The researcher noted that the SW's reaction was likely because such an activity fell outside their typical responsibilities. Similarly, the HCP went on to ask the SW if they were able to take vital signs, like a blood pressure (BP) reading. This potential drift of delegation might have significant implications for the quality and safety of health care, as reliance on SW's skills or knowledge and willingness to perform these tasks becomes paramount. This drift also has the potential to blur professional and quality boundaries, especially for SW who might not feel comfortable or informed about taking vital or assisting in health-related tasks.

Several strategies were adopted when physical monitoring was needed, but in-person attendance was not possible due to stress or other impediments. In these situations, a hybrid consultation style was observed with HCP asking SW or family members to undertake and report readings such as body weight and BP.

HCPs and family members considered a hybrid style of care delivery useful, but in very defined situations. Hybrid working relies on SWs or families accurately recording data and ensuring that readings are reported back to HCPs. This reliance was reciprocal, as HCPs also relied on families or SWs to help with essential care when options for in-person appointments were limited (eg, when communicating vital sign information, care, or medication instructions). The challenges of monitoring vital signs (ie, BP, urine, and temperature) also emerged in interviews with families. They were concerned about HCPs' expectations that families had the skills and equipment at home to perform physical readings.

Overall, despite the obvious limitations of using VC, HCPs and families also described some benefits of this hybrid approach to care, but these were dependent on several factors, such as whether the appointment covered low-risk or routine topics. They also spoke about their potential benefits in supporting

XSL•FO
RenderX

more frequent attendance, over that of just annual health checks. This suggests that VCs may play a role in widening access opportunities for patients, enabling patients to consult more frequently and complementing in-person care.

Theme 2: Choice

During interviews, people with intellectual disabilities reported that they sometimes required support from a family member or SW. Partners, parents, or siblings often assumed practical responsibilities, such as taking notes or providing supplementary information throughout the entire VC process: before, during, and after the consultation. They also assisted with communication by rewording, rephrasing, or clarifying information to reduce the risk of misunderstandings and ensure that the individual's verbal or nonverbal communication was accurately understood. This also included support with comprehension and information retention—including explaining instructions, note-taking, and arranging reminders: "Information, my key worker or my parents will explain things to me" (D026 patient interview).

Patients spoke of how having someone with them facilitated comprehension, before, during, and after appointments. Having appropriate support helped allay worries:

> *It means that you can have you can be calm and relaxed and you can have the people you need around you. Like in my case where I've got my other half and my brother, I can have them both here and they can hear exactly what's being said.* [D002 patient interview]

For people with intellectual disabilities, feeling comfortable was also linked to practical considerations, such as knowing the purpose of the appointment, who will be there, such as a trusted advocate, and knowing what, if any, activities will take place. Other practical considerations were knowing how long the appointment would be. The need for support was particularly important when planning and undertaking the first VC, as previous negative experiences influenced attitudes toward future VCs for almost all groups.

Challenges related to ensuring privacy during VCs were evident for HCPs when both patients and their family members were present on the call. In some cases, individuals wished to speak privately with the HCPs but were unable to do so within the constraints of the virtual format. However, similar privacy concerns were also reported during in-person appointments, particularly when family members were accompanied by younger children due to the unavailability of childcare.

Knowing and prioritizing patients' and family members' preferences and choices for the mode of consultation emerged as a quality indicator. Patients and family members valued HCPs who prioritized their preferences, whether for an in-person appointment or a VC, and were acutely aware of when VCs were used in ways that appeared to prioritize the service needs over their own preferences. This tended to occur when services had limited staff availability, time, staff transportation issues, or their home was geographically inaccessible.

Being flexible about offering choice and taking account of preferences, and how these may change over time, was acknowledged by HCPs as an important consideration when building familiarity and rapport. Participants with intellectual disabilities frequently emphasized the importance of having choice in how care was delivered, while also acknowledging that preferences might vary depending on the specific health concern or the need for physical examinations. This flexibility also included consideration of staff preferences, about how they chose to work, and if they themselves had reasonable adjustments or disabilities that impacted on delivering care virtually (ie, having sensory impairments).

Finally, supporting choice, in practical terms, may depend on patients, families or SWs, and HCPs' confidence in using technology and their ability to organize, set up, and open VC appointment links. Many patients owned or had access to tablets, smartphones, and laptops and were confident in using them to communicate online through different platforms:

> *I'm used to the iPad cause I did loads of things on my iPad. I use zoom. I use Microsoft. And to adapt the way we are now.* [D026 patient interview]

For patients in supported living, they frequently had access to laptops owned by the home manager. As previously noted, family members and SWs played a central role in supporting communication and comprehension during VCs. Unlike in-person appointments, though, families or SWs also needed to have some technological competency to effectively support communication in the virtual environment.

Offering patients a choice in the mode of contact (in-person, video, or telephone) was recognized as an indicator of high-quality care. However, this is contingent upon HCPs' ongoing assessment of patients' evolving needs and preferences. Facilitating choice also requires consideration of various contextual factors, including the role of families and SWs in supporting VC, the availability of appropriate digital devices, and the digital literacy of users.

Theme 3: Familiarity, Online Relationships, and Trust

Building familiarity was seen as essential to building good online relationships. Once established, a good rapport contributed to multiple forms of trust, including trust in HCPs' competence in using VC, their ability to deliver meaningful consultations, trust in the technology itself, and confidence that care would remain continuous and that choice in care delivery would be maintained. Building familiarity was seen as knowing the patient, their health status and care plans, family context, support network, and ideally, their interests, hobbies, or jobs. Building familiarity was the foundation for the first contact and successful ongoing engagement:

> *I mean the thing more that helped because I'd already met her face to face. So I knew who I was talking to, so it might have been more nervous if I hadn't known her. It was that I was going to have on the screen.* [D002 patient interview]

HCPs reported that building familiarity was the focus in the first VC, before any subsequent meetings focusing on clinical issues. This first VC was especially important for those reluctant

XSL•FO
RenderX

to attend in-person appointments. Sometimes an in-person appointment was needed first, before moving online, when regular ongoing care was needed, and there were concerns that regular in-person attendance would not be possible. Developing familiarity with HCPs was perceived by all participants as contributing to more positive future interactions, whether through virtual or in-person appointments. Building familiarity also aided feelings of comfort and was a key part of laying the foundations of future care. For both patients and HCPs, this sense of familiarity helped foster trust, reduce anxiety, and improve communication through better rapport, all of which are critical to enabling person-centered care. Linked to familiarity, all participant groups mentioned the pros and cons of communicating online and the need to build familiarity to help understand individuals' communication styles, unique characteristics, and sense of humor. Indeed, HCPs often noted their surprise at the extent to which meaningful relationships could be established online, even in cases where they had not previously met the patient, family, or SW in person:

> *You know, we're building up a relationship with them and and it's surprised me how well that worked for people that haven't met us in person. And then when we would go out and see them in person ... they would know who we were because they'd seen us on the screen ....* [E004 HCP interview]

Familiarity between VC participants also extended to knowing, not just the patient, but also the SW, supported living or day care staff and/or home managers by name, as their roles were essential in initiating VC, building rapport between professional groups, and maintaining ongoing contact.

Trust in the use of a VC was also shaped by what families or SWs and HCPs could observe and hear during video or telephone appointments. A limited or partial view of the patient, such as only seeing them from the waist up, was a concern for HCPs, as it restricted their ability to assess the patients holistically. However, the impact of this limitation varied across professional roles, with some specialist staff adapting their use of VC to suit their clinical needs. For instance, a community physiotherapist repositioning the camera to capture the patient walking around the room.

For HCPs, having advanced communication skills was considered key to establishing a trusted relationship to ensure VCs went well. HCPs reported using multiple skills aimed at getting to know the patient, such as taking time for chit-chat, authentically asking questions about clients' interests, and advanced skills needed to build rapport (tone, pace, eye contact, and body language). The absence of this authentic contact could otherwise reduce HCP to be, "... just a head in a box" (E011 HCP interview).

Building familiarity and relationships with families or SWs was important for HCPs, as VC contributed to efficient care planning over time. The flexibility of using VC also allowed for the inclusion of family members and other care agencies (eg, social care professionals) in appointments, regardless of their physical location. This was particularly beneficial in accommodating family members who worked shifts or had other scheduling constraints, enabling broader participation in patient care.

Overall, there were mixed views about the ease of building online relationships and the many factors needed to establish and support trusted relationships to develop over time.

Theme 4: Prepare and Personalize

Preparing for VCs was considered critical at different time points, before, during, and after a VC, as the tasks varied according to the context and purpose of the appointment. Examples included sending preappointment reminders or post-VC information checks. Only patients, family members, or SWs spoke about the importance of HCPs being prepared for online or telephone appointments. They emphasized the need for HCPs to be familiar with the patients' needs, such as existing health conditions, reasonable adjustments, communication or support needs, and whether a patient would be accompanied by a family member or SW. Nevertheless, patients and families reported that this preparation was not often undertaken and sometimes overlooked entirely: "I think have health professions are so overworked that they don't read the notes" (D002 patient interview).

Family members expressed frustration with HCPs' lack of preparation for VCs, particularly when required to repeat essential information such as reasonable adjustments or communication needs, which were already documented in the electronic medical record or when such records or notes were not reviewed in advance.

HCPs occasionally questioned the presence of family members on the call. Recurring advice from people with intellectual disabilities for HCPs was to get to know the patient, as this would aid their understanding of the patient and improve the quality of the conversations once in the virtual "room." Patients and families emphasized the importance of HCPs understanding their communication styles. Preparation was considered essential for enhancing both verbal and nonverbal communication, as the more prepared HCPs were for the VC, the more comfortable patients and families felt engaging with them.

Inevitably, technical issues arose during VC. HCPs emphasized the importance of early preparation, including establishing contingency plans at the outset of the consultation and, where possible, reinforced through prior correspondence. This plan typically involved informing patients that, in the event of internet connectivity failure, the clinician would immediately initiate a telephone call to continue the appointment. The presence of such an alternative arrangement provided reassurance to patients and reflected the thoroughness of clinicians' preparatory practices for VCs.

Strategies to enhance VC use among HCP teams took several forms. For instance, HCPs routinely documented consultation experiences, briefed team members, and established procedures for consistently communicating summary notes or follow-up actions to patients and their families when appropriate. This collective learning aimed not only to build professional competencies and support the exchange of effective practices but also to mitigate the recurrence of poor practice. Ultimately, these efforts aimed to strengthen best practices and ensure the sustained quality of VC.

XSL•FO
RenderX

Overall, the success of VCs was dependent on key preparatory tasks carried out at different time points, including contingency planning by HCPs, awareness of reasonable adjustments, and the implementation of team-based improvement strategies.

Theme 5: Continued Connection

For individuals with intellectual disabilities, repeated contact with the same HCP was especially valued, as it promoted continuity and reduced the cognitive and emotional demands associated with engaging with staff, particularly those mediated by technology. Patients liked and expected to see the same professional when they visited their GP or community team in person. This was similarly echoed in VC. Feeling connected to HCPs and services emerged in several ways throughout the study.

Continued connection was an important theme, before, during, and after a VC, and was seen as a quality indicator. HCPs reported their intention to plan future appointments for patients with the same staff member. This was seen as important not only to monitor health through regular checks but also to keep the momentum of being seen regularly and build on prior relationships and familiarity. For HCPs, continuity was embedded into the ethos of some services, with some explicitly stating that supporting continuity was one of their core values.

> *I just liaise with them and work around what's best for them, who they want to see if they want continuity, if there's a particular clinician they want to see or if there's a site because we've got four sites they might want to only go to a particular site. So I'm always on the phone speaking to them.* [D028 HCP interview]

VCs could facilitate ongoing engagement in care, particularly for family members and SWs who wished to be involved in decision-making but were unable to attend appointments in person due to geographic distance or other commitments. VCs offered a flexible means of participation, enabling individuals to join consultations partially or fully, thus maintaining involvement in the care process. This continued connection was perceived as a valuable opportunity to share information more efficiently and to respond more promptly to emerging needs in care planning and management. For patients, VCs may serve as an initial point of contact with HCPs, allowing for the development of rapport even before any in-person appointments, again helping to reinforce the sense of connection through familiarity with HCPs.

Although VCs may facilitate connection with health care services, family members, SWs, and HCPs expressed concerns that VCs could also lead to a loss of connection or a perception of impersonal care, describing them as more "clinical":

> *I feel like having the kind of virtual side of it kind of depersonalises it again and the whole part of my work is personalisation and, kind of, making every kind of contact count. I don't know if necessarily sometimes video chats is the best way forward for it really.* [E003 CP interview]

For patients, families or SWs, and HCPs, feeling connected was enhanced if VCs were uninterrupted, such as when HCPs put patients on hold or were disturbed by colleagues asking questions. These interruptions were thought to impact the flow of conversation and level of engagement, making a difference to whether patients felt comfortable or not. For HCPs, being interrupted was strongly linked to whether they had an appropriate (quiet) meeting space or had all the information they needed prior to starting the appointment.

The conclusion of a VC was regarded as a critical phase for ensuring continuity of care. Health care staff emphasized the importance of clearly communicating the next steps, including any required follow-up, aftercare, or planned activities. Explicit articulation of postconsultation actions ensured that all participants were informed about subsequent processes, which helped to manage expectations and facilitated a smooth transition to ongoing care. This included sending post-VC written summary notes, preferably in accessible formats, using plain language, easy-read, and different communication modes (letters and telephone calls).

All user groups expressed the importance of continuity of care, specifically the value of meeting the same HCP during VCs. Feelings of disconnection were alleviated by simple measures such as ensuring that calls were uninterrupted. A sustained sense of connectedness was perceived as integral to the quality provision over time.

Discussion

Principal Findings

In this paper, we report the first UK-based insights into VC and the factors influencing the quality and safety of use for individuals with intellectual disabilities, their families, SWs, and HCPs. Our findings highlight vital areas for development to inform how services are designed and delivered in the future and enable sustainable, safe, individualized, and effective care [35,36], contributing to a deeper understanding of VCs as a form of online service provision for all user groups. This learning can be used to underpin how services are designed and delivered in the future, enabling safe, high-quality, person-centered care. We found a complex perspective on the acceptability, use, and effectiveness of VCs in managing health care for people with intellectual disabilities [37].

Our study reports mixed views about the acceptability and use of VC in relation to quality and safety as defined in UK policy [38,39]. Quality and safety are crucial aspects outlined in the UK Health and Care Act 2022 [38] and the NHS Outcomes Framework 2014 [39], which impact 3 key aspects of care: clinical effectiveness, safety, and patient experience. As implementation of the NHS 10-year plan [2] advances, it is vital to ensure that digital services are of high quality and safe. Although not reporting on clinical effectiveness, our data identify vital and interconnected elements needed to ensure effective service delivery. Namely, the VC space, online relationship building, personalization, and continued connection were critically important areas that were reported as being key to ensuring the quality and safety of care provision.

More broadly, our findings align with concerns about VC use among patients with sensory, neurological, and musculoskeletal impairments [7]. Similar issues have been noted in other groups

XSL•FO
RenderX

at risk of digital exclusion, such as older adults in rural areas [40] and those with limited language proficiency [41-43]. Building on the work of Halas et al [35], who explored the balance between virtual and in-person care, we found that the risks associated with VCs may be heightened for individuals with intellectual disabilities, potentially compromising care quality [44]. While our data reflect some of these concerns, it also aligns with other research, indicating that there are opportunities for positive online experiences [45], especially when supported or implemented effectively. If we are to improve access to care, such as annual health checks, more needs to be done to facilitate access to alternative routes to care [46].

Comparison to Previous Research

Safety was ingrained across many of the key themes we identified, particularly linked to the spaces where VCs take place, both at home and at work. Consistent with previous research, both the patients' home environment [47] and the HCPs' virtual workspace played a significant role in shaping the quality of VCs. The importance of the VC consultation space has been highlighted in previous studies involving service users from marginalized groups, such as asylum seekers, and supported living residents [43].

In line with other studies, we found uncertainty about privacy within home or work settings for VCs, as it was reported to inhibit engagement, especially for HCPs who might not have suitable workspaces [48] or feel less confident working online [10]. Risk was also heightened if HCPs were in any doubt about the patient's home environment [42]. Clinicians also reported mixed views about the use of VCs, especially when linked to practical tasks, such as when a physical examination [49] was required, or there were concerns about their ability to identify pain or care urgency. Consistent with Halas et al [35], we found instances where HCPs needed to balance risks when managing patients virtually. These skills rely on advanced communication abilities whereby HCPs must depend on what is visible, heard, and said on screen [50]. People with intellectual disabilities often prefer video calls to telephone consultations, primarily because they enable visibility of facial expressions and improve communication [51]. We found that although telephone calls retained value in low-risk contexts or for quick information exchange, they could be less effective, particularly if affected by accents or a lack of visual cues. Nevertheless, like in-person care, HCPs need specific skills such as "empathetic verbalization" to enhance active listening [51]. Moreover, it could be argued that communication skills are even more important in VC, as there is a greater risk of missing vital information.

We found that patient experience was heavily linked to the development of trust through building good relationships, preparing for appointments and personalizing care, and critically ensuring continued connection, again to ensure uptake of future appointments. HCPs were vocal about the need to be familiar with the patient, which, in turn, built trust and paved the way for future appointments, whether in person or via VCs. Similar to other studies, we identified how prior relationships can support a reverse shift, from online interactions to in-person appointments, as these interactions had built a foundation of trust [52]. As such, VC may create new opportunities for therapeutic connections, especially when trust has already been established [37].

In our study, VC participants were aware of how the use of technology may have the potential to impact the patient-HCP relationship [35,44,53]. HCPs were observed to use multiple strategies to build familiarity and mitigate or eliminate the power imbalance that could occur in VCs. Building familiarity was based on recognizing VC participants' unique communication styles and their ability to pick up on visual cues that aided understanding during appointments. Like other studies, our study found that body language and facial cues were key in supporting communication [50]. Our study also showed a close link between familiar and established relationships and communication effectiveness [54] due to families' or caregivers' ability to recognize the needs of patients and their role in supporting technology use. VCs offered opportunities for patients who may struggle with verbal communication in telephone or in-person appointments to access care in a way that made them feel more comfortable, such as patients with autism spectrum disorders. Consistent with prior research on remote care access for marginalized groups [55], VC can lead to different care experiences depending on how they are planned, participants' familiarity with one another, and the context of their work or home environments.

VC requires preparation for planning, conducting, and undertaking post-VC tasks, and this can be a lengthy process. Previous studies have also emphasized the importance of HCP preparedness [56] and the need to balance access with workload demands [57]. Preparation takes time and includes a range of tasks over time to ensure choice is given, the VC is prepared for, and there is a continued connection. If any of these issues are absent, then there is a danger of future nonattendance and reduced opportunities for future VC engagement. As such, preparation is critical if care delivery is to be safe and of high quality. Similar to other studies, we found that a lack of HCP preparedness often led to patient and family frustration, especially if reasonable adjustments were not read before appointments or acted upon [58,59].

Finally, we identified that feeling connected and having a sense of continuity of care in VC was important for patients and their families. Continuity, especially seeing the same HCP, was linked to higher-quality care through improved rapport and familiarity. This aligns with existing evidence on sources of resilience [60], confirming the importance of connectedness, even in virtual spaces, and the concept of "accumulated knowledge" that underpins person-centered care [61]. Continuity may be even more important when consulting online, as the bond established can be argued to be more fragile than in-person appointments, especially if patients are unfamiliar with the HCP and trust has not been established.

Implications for Practice, Education, and Research

Our findings point to implications for practice in terms of supplying HCPs with adequately equipped VC workspaces, IT equipment and support, and private or quiet workspaces. There is also a need to develop guidelines to support rules of engagement [62] or "netiquette" guidelines for online interaction

XSL•FO
RenderX

[8,50,53]. Guidance on the rules of engagement, rapport, and relationship building online may also impact HCPs' and SWs' confidence in using VC and have implications for improved interaction. These considerations are all within the context of VC as a complement to service delivery, rather than a replacement for in-person care.

Not all HCPs were comfortable with video or telephone conferencing, but all wanted to improve their digital or communication skills. This highlights the need for education, training, and skill development at all levels of higher education and service types. HCPs suggested shared training opportunities with colleagues from related services to enhance IT skills and build confidence in conducting quality VC. It is essential that training is collaboratively designed by a diverse range of staff and includes digital support for all user groups, particularly for individuals with intellectual disabilities and family members or caregivers who are less confident IT users who may struggle with accessing digital health care [63].

Further research is needed to design user-friendly digital platforms that avoid excessive information and complex language [64,65]. Understanding how to foster successful online relationships over time is also crucial for both HCPs in different health care settings (HCPs) and individuals with intellectual disabilities [52,66]. Studies should also examine how VCs impact pain identification, waiting times, and care quality and satisfaction [67]. While some research addresses diagnostic overshadowing [33,68], for individuals with intellectual disabilities, knowledge about how technology-mediated health care affects timely access and health outcomes remains limited. Future studies should involve people with intellectual disabilities through co-design processes [69] to create inclusive technologies, focusing on areas such as cyber etiquette [70], online safety [71,72], and collaborative learning to establish best practices.

This study provides valuable insights into the use of VC across care settings. However, the transferability of the findings should be considered, as more data were derived from community teams rather than from GP professionals. Nonetheless, key findings, such as the importance of relationship-building, preparation, and continuity of care, are likely to be transferable and relevant across diverse settings and professional groups.

Future research could examine how these findings translate into other settings with differing technologies, information systems, or staffing models to better understand more subtle differences and the applicability of VC care delivery. Moreover, greater research into the possible ethical or medical-legal implications of professional role blurring may also be important, not only for health care employers to protect staff, but also more broadly in terms of providing ethical quality care.

For patients and families, key safety concerns arise when using VC, particularly in shared home settings and when there is a lack of practical support and guidance. Outputs from this study, including guidance, go some way to addressing these needs and may offer valuable insights for local authorities, health services, and policymakers to ultimately support the inclusive and responsible use of VC, complementing in-person provision. See Multimedia Appendix 1 for the web link to resources.

Finally, our project highlights several policy implications across GP, community care, and SW groups. In alignment with other research on digital inclusion [73], our research shows the need to align working practices across care settings to ensure consistent delivery of VC practices, ensuring appropriate workplace fit, and being attentive to supporting access and use of virtual care. Moreover, changes to current policies and best practice guidance are needed to overcome professional, service, and system-level barriers. Clarifying delegated tasks, between HCP and families or SWs, when taking vital signs, may also help define role boundaries between groups.

There is also a need to clarify when and how to use VC effectively. This should be supported by adequate resourcing in improved workspaces, targeted HCP or SW training, and co-designed resources for users. Additionally, VC platform design must improve to enhance digital accessibility, including visual and easy-read navigation tools and information for diverse user groups.

Limitations

On reflection, one limitation was the number of observations undertaken. Initially, we aimed to observe up to 12 videoconferences. However, due to the challenges of arranging these sessions and time constraints within the project, we were only able to observe 3, 2 of which were in GP. A significant factor affecting recruitment was the pandemic itself, as most of our data collection took place during the latter end of this challenging period. Furthermore, although we captured a broad representation of community care professionals across geographical areas, our final sample lacked sufficient diversity of family members, especially those caring for people with profound intellectual disabilities and primary care clinicians, which may have caused sampling bias.

We gained valuable methodological insights from conducting this study, some of which highlighted real-world issues impacting research practices (both positive and negative, which will be reported elsewhere). As such, a further limitation might be that several interviews took place in unconventional but private spaces, such as home hallways or remote buildings not usually used for the service, as these were the only locations where family members, SWs, or HCPs could be interviewed and where internet connections were available or reliable.

These limitations were mitigated by the study's co-design approach, focusing on quality and not speed. In this study, we collaborated with individuals with intellectual disabilities, their families, SWs, and community partners using co-design methods, all of which took time. This limitation was also this study's strength, as this inclusive research approach was designed to ensure the collection of meaningful data that addresses the project's goals and produces outputs that are useful, practical, and relevant in real-world settings.

Conclusions

This study provides a deeper understanding of VCs as a form of online service provision. The findings reveal mixed opinions regarding the acceptability and use of VCs among individuals with intellectual disabilities, their families, SWs, and HCPs. Our results show how VCs can be conducted safely while

XSL•FO
RenderX

ensuring high-quality, person-centered care. However, quality and safety can be compromised if not implemented effectively. We found that while VCs can play a significant role in delivering timely care, participants felt that they should not be considered a substitute for in-person consultations, nor should they be conducted in isolation. Instead, VCs can effectively complement in-person care, supporting existing care arrangements and even wider care access in specific circumstances. They also provide an opportunity for those who may not be able to attend in-person meetings to connect with HCPs in a way that meets their needs.

Acknowledgments

The authors would first like to thank all the participants who took the time to contribute to this study. Their stories and experiences have all been invaluable in showing us their views about the topic and have helped to co-design practical study outputs, which will be available via the project website: Virtual consultations for people with learning disabilities, their families and health care providers: a co-design study to aid implementation in everyday practice, University of Surrey. The research team would also like to thank our lead Experts by Experience (Michele Cobb and Katy Bassent), our network of Experts by Experience, and members of the project advisory group for their guidance when undertaking this project. The authors confirm that the basic version of Grammarly was used to check spelling and grammar during the manuscript preparation. No other generative artificial intelligence technologies were used to produce the text, figures, or informational content.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to controlled access arrangements but are available from the corresponding author on reasonable request. Extracted data are supplied in Multimedia Appendix 1, available to download along with the published manuscript.

Funding

This research was funded by the National Institute for Health and Care Research (NIHR) Applied Research Collaboration Kent, Surrey, Sussex. The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR, or the Department of Health and Social Care. JA receives salary support from the NIHR Applied Research Collaboration Kent, Surrey, Sussex. For the purpose of Open Access, the author has applied a Creative Commons Attribution (CC BY) public copyright license to any author-accepted manuscript version arising from this submission.

Authors' Contributions

Conceptualization: FM (leading)
Formal analysis: FM (leading), AC (supporting), VT (supporting), PS (supporting), TP (supporting), HM (supporting), and JA (supporting)
Funding acquisition: FM (leading), AC (supporting), VT (supporting), PS (supporting), HM (supporting), and JA (supporting)
Investigation: TP (leading), FM (supporting)
Methodology: FM (leading), AC (supporting), VT (supporting), PS (supporting), HM (supporting), and JA (supporting)
Project administration: FM (leading), TP (supporting)
Supervision: FM (leading), JA (supporting)
Visualization: FM (leading), TP (supporting)
Writing—original draft: FM (leading)
Writing—review and editing: FM (leading), AC (supporting), VT (supporting), PS (supporting), HM (supporting), JA (supporting),

Conflicts of Interest

JA receives salary support from the National Institute for Health and Care Research Applied Research Collaboration Kent, Surrey, Sussex.

Multimedia Appendix 1
Quotes and resources weblink.
[DOCX File , 99 KB - jmir_v28i1e81173_app1.docx]

References

1. Song K, Hey M, Payne R. Media depictions of primary care teleconsultation safety: a thematic analysis of UK newspapers. Br J Gen Pract 2024;74(747):e695-e701 [FREE Full text] [doi: 10.3399/BJGP.2023.0543] [Medline: 38621808]
2. 10 year health plan for England: fit for the future. Department of Health and Social Care. 2025 Jul 03. URL: https://www.gov.uk/government/publications/10-year-health-plan-for-england-fit-for-the-future [accessed 2025-07-05]
3. Digitally enabled triage. NHS England. 2023. URL: https://www.england.nhs.uk/long-read/digitally-enabled-triage/ [accessed 2024-11-25]

XSL•FO
RenderX

4. Online health and prescription services. NHS Services. 2023. URL: https://www.nhs.uk/nhs-services/gps/online-health-and-prescription-services/ [accessed 2025-01-31]
5. Payne R, MacIver E, Clarke A. Do new models of primary care risk exacerbating existing inequity? Br J Gen Pract 2024;74(747):436-437 [FREE Full text] [doi: 10.3399/bjgp24X739449] [Medline: 39327083]
6. Patient safety healthcare inequalities reduction framework. NHS England. URL: https://www.england.nhs.uk/long-read/patient-safety-healthcare-inequalities-reduction-framework/ [accessed 2025-06-25]
7. Pettersson L, Johansson S, Demmelmaier I, Gustavsson C. Disability digital divide: survey of accessibility of eHealth services as perceived by people with and without impairment. BMC Public Health 2023;23(1):181 [FREE Full text] [doi: 10.1186/s12889-023-15094-z] [Medline: 36707791]
8. Chadwick D, Ågren KA, Caton S, Chiner E, Danker J, Gómez‐Puerta M, et al. Digital inclusion and participation of people with intellectual disabilities during COVID-19: a rapid review and international bricolage. J Policy Pract Intellect Disabil 2022;19(3):242-256 [FREE Full text] [doi: 10.1111/jppi.12410]
9. Clarke G, Dias A, Wolters A. Access to and delivery of general practice services: a study of patients at practices using digital and online tools. The Health Foundation. 2022. URL: https://www.health.org.uk/publications/access-to-and-delivery-of-general-practice-services [accessed 2024-03-18]
10. Statton S, Jones R, Thomas M, North T, Endacott R, Frost A, et al. Professional learning needs in using video calls identified through workshops. BMC Med Educ 2016;16:140 [FREE Full text] [doi: 10.1186/s12909-016-0657-6] [Medline: 27165431]
11. Digital requirements guidance. NHS England. 2024. URL: https://www.england.nhs.uk/gp/investment/gp-contract/digital-requirements-guidance/ [accessed 2024-11-25]
12. Digital primary care: the good practice guidelines for GP electronic patient records—(GPGv5). NHS England. 2023. URL: https://www.england.nhs.uk/digital-gp-good-practice-guidance/ [accessed 2025-07-05]
13. Flynn S, Hatton C. Health and social care access for adults with learning disabilities across the UK during the COVID-19 pandemic in 2020. Tizard Learn Disabil Rev 2021;26(3):174-179 [FREE Full text] [doi: 10.1108/tldr-06-2021-0014]
14. Selick A, Durbin J, Salonia C, Volpe T, Orr E, Hermans H, et al. The nuts and bolts of health care: evaluating an initiative to build direct support professional capacity to support the health care of individuals with intellectual disabilities. J Appl Res Intellect Disabil 2022;35(2):623-632. [doi: 10.1111/jar.12975] [Medline: 34985176]
15. Health and care of people with LD, experimental statistics 2022-2023. NHS England. URL: https://digital.nhs.uk/data-and-information/publications/statistical/health-and-care-of-people-withlearning-disabilities/experimental-statistics-2022-to-2023# [accessed 2024-03-27]
16. White A, Sheehan R, Ditzel N, Ding J, Roberts C, Magill N, et al. LeDeR Annual Report: Learning from lives and deaths people with a learning disability and autistic people. The Institute of Psychiatry, Psychology and Neuroscience (IoPPN), King's College London. 2023. URL: https://www.kcl.ac.uk/ioppn/assets/pdfs/leder/2023-final-updated.pdf [accessed 2023-09-07]
17. Priorities and operational planning guidance 2024/25. NHS England. 2024. URL: https://www.england.nhs.uk/publication/priorities-and-operational-planning-guidance-2024-25/ [accessed 2025-07-05]
18. Bate S, Robert G. Bringing User Experience to Health Care Improvement: The Concepts, Methods and Practices of Experience-Based Design. Oxford: Radcliffe Publishing; 2007.
19. Tsianakas V, Robert G, Maben J, Richardson A, Dale C, Griffin M, et al. Implementing patient-centred cancer care: using experience-based co-design to improve patient experience in breast and lung cancer services. Support Care Cancer 2012;20(11):2639-2647 [FREE Full text] [doi: 10.1007/s00520-012-1470-3] [Medline: 22544223]
20. Tsianakas V, Robert G, Richardson A, Verity R, Oakley C, Murrells T, et al. Enhancing the experience of carers in the chemotherapy outpatient setting: an exploratory randomised controlled trial to test impact, acceptability and feasibility of a complex intervention co-designed by carers and staff. Support Care Cancer 2015;23(10):3069-3080. [doi: 10.1007/s00520-015-2677-x] [Medline: 25744288]
21. Kaley A, Hatton C, Milligan C. More than words: the use of video in ethnographic research with people with intellectual disabilities. Qual Health Res 2019;29(7):931-943. [doi: 10.1177/1049732318811704] [Medline: 30520345]
22. Learning disabilities and behaviour that challenges: service design and delivery. Nice Guideline Ref NG93. National Institute for Health and Care Excellence (NICE). 2018 Mar 28. URL: https://www.nice.org.uk/guidance/ng93 [accessed 2026-04-23]
23. Patton MQ. Qualitative Research Methods: Integrating Theory and Practice. London: Sage; 2015.
24. Dunford M. Unlocking data secrets safely with Trusted Research Environments. Lifebit. 2025 Jun 16. URL: https://lifebit.ai/blog/trusted-research-environment/ [accessed 2025-12-14]
25. UK Health Data Research Alliance, NHSX. Building Trusted Research Environments—principles and best practices; towards TRE ecosystems. Zenodo. 2021. URL: https://zenodo.org/record/5767586 [accessed 2026-04-23]
26. Roter D, Larson S. The roter interaction analysis system (RIAS): utility and flexibility for analysis of medical interactions. Patient Educ Couns 2002;46(4):243-251. [doi: 10.1016/s0738-3991(02)00012-5] [Medline: 11932123]
27. Miller EA, Nelson E. Modifying the roter interaction analysis system to study provider-patient communication in telemedicine: promises, pitfalls, insights, and recommendations. Telemed J E Health 2005;11(1):44-55. [doi: 10.1089/tmj.2005.11.44] [Medline: 15785220]

XSL•FO
RenderX

28. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009;4:50 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
29. Ritchie J, Spencer L. Qualitative data analysis for applied policy research. In: Bryman A, Burgess B, editors. Analyzing Qualitative Data. London: Routledge; 1994:173-194.
30. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013;13:117 [FREE Full text] [doi: 10.1186/1471-2288-13-117] [Medline: 24047204]
31. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
32. Clinical guide for frontline staff to support the management of patients with a learning disability and autistic people—relevant to all clinical specialties. NHS England. URL: https://tinyurl.com/48p8jnb6 [accessed 2025-05-05]
33. Mason J, Scior K. 'Diagnostic overshadowing' amongst clinicians working with people with intellectual disabilities in the UK. J Appl Res Intellect Disabil 2004;17(2):85-90 [FREE Full text] [doi: 10.1111/j.1360-2322.2004.00184.x]
34. Dell'Armo K, Tassé MJ. Diagnostic overshadowing of psychological disorders in people with intellectual disability: a systematic review. Am J Intellect Dev Disabil 2024;129(2):116-134. [doi: 10.1352/1944-7558-129.2.116] [Medline: 38411245]
35. Halas G, Baldwin A, LaBine L, MacKay K, Singer A, Katz A. A phenomenological inquiry of the shift to virtual care delivery: insights from front-line primary care providers. Healthcare (Basel) 2024;12(8):861 [FREE Full text] [doi: 10.3390/healthcare12080861] [Medline: 38667623]
36. Hammersley V, Donaghy E, Parker R, McNeilly H, Atherton H, Bikker A, et al. Comparing the content and quality of video, telephone, and face-to-face consultations: a non-randomised, quasi-experimental, exploratory study in UK primary care. Br J Gen Pract 2019;69(686):e595-e604 [FREE Full text] [doi: 10.3399/bjgp19X704573] [Medline: 31262846]
37. Lake J, Jachyra P, Volpe T, Lunsky Y, Magnacca C, Marcinkiewicz A, et al. The wellbeing and mental health care experiences of adults with intellectual and developmental disabilities during COVID-19. J Ment Health Res Intellect Disabil 2021;14(3):285-300 [FREE Full text] [doi: 10.1080/19315864.2021.1892890]
38. UK Parliament. Health and Care Act 2022. Legislation.Gov.UK.: London: The Stationary Office; 2022. URL: https://www.legislation.gov.uk/ukpga/2022/31/contents [accessed 2025-07-05]
39. NHS Outcomes Framework 2015 to 2016. Department of Health and Social Care. 2015 Mar 26. URL: https://www.gov.uk/government/publications/nhs-outcomes-framework-2015-to-2016#full-publication-update-history [accessed 2025-05-07]
40. Landgren S, Cajander Å. Non-use of digital health consultations among Swedish elderly living in the countryside. Front Public Health 2021;9:588583 [FREE Full text] [doi: 10.3389/fpubh.2021.588583] [Medline: 34568247]
41. Locked out: digitally excluded people's experiences of remote GP appointments. Healthwatch. 2021. URL: https://www.healthwatch.co.uk/report/2021-06-16/locked-out-digitally-excluded-peoples-experiences-remote-gp-appointments [accessed 2025-07-05]
42. Kaihlanen A, Virtanen L, Buchert U, Safarov N, Valkonen P, Hietapakka L, et al. Towards digital health equity—a qualitative study of the challenges experienced by vulnerable groups in using digital health services in the COVID-19 era. BMC Health Serv Res 2022;22(1):188 [FREE Full text] [doi: 10.1186/s12913-022-07584-4] [Medline: 35151302]
43. Moschogianis SF, Darley S, Coulson T, Peek N, Cheraghi-Sohi S, Brown B. Patient experiences of an online consultation system: a qualitative study in English primary care post-COVID-19. Br J Gen Pract 2024;74(744):e489-e497 [FREE Full text] [doi: 10.3399/BJGP.2023.0076] [Medline: 38164550]
44. Selick A, Bobbette N, Lunsky Y, Hamdani Y, Rayner J, Durbin J. Virtual health care for adult patients with intellectual and developmental disabilities: a scoping review. Disabil Health J 2021;14(4):101132 [FREE Full text] [doi: 10.1016/j.dhjo.2021.101132] [Medline: 34127406]
45. Caton S, Bradshaw J, Gillooly A, Hatton C, Flynn S, Oloidi E, et al. Digital participation of people with profound and multiple learning disabilities during the Covid‐19 pandemic in the UK. Br J Learn Disabil 2022;51(2):163-174 [FREE Full text] [doi: 10.1111/bld.12501]
46. Owens J, Ravindrarajah R, Norman G, Hopkin E, Shi C, Lovell K, et al. Access, inequalities and annual health checks (AHCs) for adults living with severe mental illness in the UK: a mixed-methods systematic review. BMJ Open 2025;15(8):e093426. [doi: 10.1136/bmjopen-2024-093426] [Medline: 40759520]
47. Sachdeva N, Tuikka A, Kimppa K, Suomi R. Digital disability divide in information society: a framework based on a structured literature review. J Inf Commun Ethics Soc 2015;13(3/4):283-298. [doi: 10.1108/jices-10-2014-0050]
48. Turner A, Morris R, Rakhra D, Stevenson F, McDonagh L, Hamilton F, et al. Unintended consequences of online consultations: a qualitative study in UK primary care. Br J Gen Pract 2022;72(715):e128-e137 [FREE Full text] [doi: 10.3399/BJGP.2021.0426] [Medline: 34903520]
49. Lunsky Y, Bobbette N, Selick A, Jiwaet MI. "The doctor will see you now": direct support professionals' perspectives on supporting adults with intellectual and developmental disabilities accessing health care during COVID-19. Disabil Health J 2021;14(3):101066 [FREE Full text] [doi: 10.1016/j. dhjo.2021.101066]

XSL•FO
RenderX

50. Selick A, Durbin J, Hamdani Y, Rayner J, Lunsky Y. "Can you hear me now?": a qualitative exploration of communication quality in virtual primary care encounters for patients with intellectual and developmental disabilities. BMC Prim Care 2023;24(1):105 [FREE Full text] [doi: 10.1186/s12875-023-02055-z] [Medline: 37081380]
51. Vennik J, Hughes S, Lyness E, McDermott C, Smith KA, Steele M, et al. Patient perceptions of empathy in primary care telephone consultations: a mixed methods study. Patient Educ Couns 2023;113:107748 [FREE Full text] [doi: 10.1016/j.pec.2023.107748] [Medline: 37141693]
52. Scheffers F, van Vugt E, Moonen X. Resilience in the face of adversity in adults with an intellectual disability: a literature review. J Appl Res Intellect Disabil 2020;33(5):828-838 [FREE Full text] [doi: 10.1111/jar.12720] [Medline: 32153087]
53. Desroches ML, Stych J, Bannett G, Guttentag R, Ailey SH, Fisher K. Establishing best practices in telehealth care for adults with developmental disabilities in the United States: an e-Delphi study. Telemed J E Health 2024;30(8):e2373-e2382 [FREE Full text] [doi: 10.1089/tmj.2023.0547] [Medline: 38919995]
54. Menschik C, Kunze C, Renner G, Etges T. Mainstream technologies in facilities for people with intellectual disabilities: multiple-methods study using the nonadoption, abandonment, scale-up, spread, and sustainability framework. JMIR Rehabil Assist Technol 2024;11:e59360 [FREE Full text] [doi: 10.2196/59360] [Medline: 39499922]
55. Humphrey A, Cummins S, May C, Stevenson F. GP remote consultations with marginalised patients and the importance of place during care: a qualitative study of the role of place in GP consultations. BJGP Open 2025;9(1):BJGPO.2024.0050 [FREE Full text] [doi: 10.3399/BJGPO.2024.0050] [Medline: 39074881]
56. Boardman L, Bernal J, Hollins S. Communicating with people with intellectual disabilities: a guide for general psychiatrists. Adv Psychiatr Treat 2018;20(1):27-36. [doi: 10.1192/apt.bp.110.008664]
57. Neve G, Fyfe M, Hayhoe B, Kumar S. Digital health in primary care: risks and recommendations. Br J Gen Pract 2020;70(701):609-610 [FREE Full text] [doi: 10.3399/bjgp20X713837] [Medline: 33243917]
58. Accessible information standard. NHS England. URL: https://www.england.nhs.uk/accessible-information-standard/ [accessed 2025-07-03]
59. DAPB4019: Reasonable Adjustment Digital Flag. NHS England. 2023. URL: https://digital.nhs.uk/data-and-information/information-standards/governance/latest-activity/standards-and-collections/dapb4019-reasonable-adjustment-digital-flag/ [accessed 2025-07-03]
60. Scheffers F, Moonen X, van Vugt E. Assessing the quality of support and discovering sources of resilience during COVID-19 measures in people with intellectual disabilities by professional carers. Res Dev Disabil 2021;111:103889 [FREE Full text] [doi: 10.1016/j.ridd.2021.103889] [Medline: 33578230]
61. Hjortdahl P. Continuity of care: general practitioners' knowledge about, and sense of responsibility toward their patients. Fam Pract 1992;9(1):3-8. [doi: 10.1093/fampra/9.1.3] [Medline: 1634024]
62. Mold F, Cooke D, Ip A, Roy P, Denton S, Armes J. COVID-19 and beyond: virtual consultations in primary care-reflecting on the evidence base for implementation and ensuring reach: commentary article. BMJ Health Care Inform 2021;28(1):e100256 [FREE Full text] [doi: 10.1136/bmjhci-2020-100256] [Medline: 33436372]
63. Chadwick D, Wesson C, Fullwood C. Internet access by people with intellectual disabilities: inequalities and opportunities. Future Internet 2013;5(3):376-397 [FREE Full text] [doi: 10.3390/fi5030376]
64. Cooper‐Moss N, Umpleby K, Roberts C, Garner C, Edwards AH, Ditzel N, et al. Barriers to healthcare for people with a learning disability from ethnic minorities: perspectives of self‐advocates and carers. Brit J Learn Disabil 2024;53(2):301-311. [doi: 10.1111/bld.12636]
65. Savolainen K, Kujala S. Testing two online symptom checkers with vulnerable groups: usability study to improve cognitive accessibility of eHealth services. JMIR Hum Factors 2024;11:e45275 [FREE Full text] [doi: 10.2196/45275] [Medline: 38457214]
66. Reasonable Adjustment Flag. NHS England. 2025. URL: https://digital.nhs.uk/services/reasonable-adjustment-flag [accessed 2025-01-31]
67. Pereira Gray DJ, Sidaway-Lee K, White E, Thorne A, Evans PH. Continuity of care with doctors-a matter of life and death? A systematic review of continuity of care and mortality. BMJ Open 2018;8(6):e021161 [FREE Full text] [doi: 10.1136/bmjopen-2017-021161] [Medline: 29959146]
68. Jopp DA, Keys CB. Diagnostic overshadowing reviewed and reconsidered. Am J Ment Retard 2001;106(5):416-433. [doi: 10.1352/0895-8017(2001)106<0416:DORAR>2.0.CO;2] [Medline: 11531461]
69. Dakin FH, Rybczynska-Bunt S, Rosen R, Clarke A, Greenhalgh T. Access and triage in contemporary general practice: a novel theory of digital candidacy. Soc Sci Med 2024;349:116885 [FREE Full text] [doi: 10.1016/j.socscimed.2024.116885] [Medline: 38640742]
70. Caton S, Chapman M. The use of social media and people with intellectual disability: a systematic review and thematic analysis. J Intellect Dev Disabil 2016;41(2):125-139 [FREE Full text] [doi: 10.3109/13668250.2016.1153052]
71. Triantafyllopoulou P, Clark-Hughes C, Langdon PE. Social media and cyber-bullying in autistic adults. J Autism Dev Disord 2022;52(11):4966-4974 [FREE Full text] [doi: 10.1007/s10803-021-05361-6] [Medline: 34799787]
72. Triantafyllopoulou P, Newsome J, Tsang W, McCarthy M, Jones K. Safer online lives: internet use and online experiences of adults with intellectual disabilities-a survey study. J Appl Res Intellect Disabil 2025;38(3):e70061. [doi: 10.1111/jar.70061] [Medline: 40302647]

XSL•FO
RenderX

73. Alami H, Lehoux P, Shaw SE, Papoutsi C, Rybczynska-Bunt S, Fortin J. Virtual care and the inverse care law: implications for policy, practice, research, public and patients. Int J Environ Res Public Health 2022;19(17):10591 [FREE Full text] [doi: 10.3390/ijerph191710591] [Medline: 36078313]

Abbreviations

BP: blood pressure
COREQ: Consolidated Criteria for Reporting Qualitative Research
EBCD: Experience-Based Co-Design
EbE: Experts by Experience
GP: general practice
HCP: health care professional
NHS: National Health Service
PPIE: patient and public involvement and engagement
SW: support worker
VCs: virtual consultations

Edited by A Stone, A Mavragani; submitted 23.Jul.2025; peer-reviewed by J Vennick, S Narayan, J Raju; comments to author 15.Aug.2025; revised version received 01.Mar.2026; accepted 02.Mar.2026; published 20.May.2026.

Please cite as:
Mold F, Cox A, Tsianakas V, van Marwijk H, Shanahan P, Parsons T, Armes J
Virtual Consultations for People With Intellectual Disabilities in General Practice and Community Care: Mixed Methods Qualitative Study
J Med Internet Res 2026;28:e81173
URL: https://www.jmir.org/2026/1/e81173
doi:10.2196/81173
PMID:

©Freda Mold, Anna Cox, Vicki Tsianakas, Harm van Marwijk, Paul Shanahan, Treena Parsons, Jo Armes. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Mode Effects Between Mobile Web and Telephone Surveys on Patient Experience Scores in South Korea: Secondary Analysis of a Randomized Controlled Trial Under Various Missingness Scenarios

Young-Geun Choi[1,2], PhD; Bon Mi Koo[3], PhD; Yeongchae Song[4,5], MS; Young Kyung Do[3,4], MD, PhD

[1]Department of Mathematics Education, Sungkyunkwan University, Jongno-gu, Seoul, Republic of Korea

[2]Institute of Pure and Applied Mathematics, Sungkyunkwan University, Jongno-gu, Seoul, Republic of Korea

[3]Institute of Health Policy and Management, Seoul National University Medical Research Center, Jongno-gu, Seoul, Republic of Korea

[4]Department of Health Policy and Management, Seoul National University College of Medicine, 103 Daehak-ro, Jongno-gu, Seoul, Republic of Korea

[5]Review and Assessment Research Department, Health Insurance Review and Assessment Service, Wonju, Gangwon, Republic of Korea

Corresponding Author:
Young Kyung Do, MD, PhD
Institute of Health Policy and Management, Seoul National University Medical Research Center, Jongno-gu, Seoul, Republic of Korea

Abstract

Background: Patient experience surveys are essential tools for assessing health care quality, yet the potential influence of survey mode on patient experience scores remains understudied. This study investigates the mode effects between mobile web and telephone surveys within South Korea's Patient Experience Assessment.

Objective: This study aimed to examine the presence and extent of the mode effects of mobile web versus telephone surveys on patient experience scores. The primary outcome was defined as the total score across all 21 survey items, rescaled to 0 - 100.

Methods: This is a secondary analysis using experimental data from a parallel-group randomized controlled trial involving 3200 patients (adults aged ≥19 years, hospitalized >1 day, discharged 2 - 56 days before the survey) from 4 general hospitals between October and November 2022, equally allocated to telephone and mobile web survey modes. An independent survey company generated the random allocation sequence using computer-generated random numbers and assigned participants to the survey modes. Due to the nature of the intervention, blinding of participants, interviewers, and outcome assessors was not feasible after assignment. We calculated unadjusted score differences among respondents and estimated adjusted differences accounting for nonresponse using inverse probability weighting (IPW) and multiple imputation (MI) under the missing-at-random assumption. Sensitivity analyses, using the delta-adjustment method based on the missing-not-at-random assumption, assessed robustness to departures from the missing-at-random assumption. Subgroup analyses by sex, age group, and field of care were also conducted.

Results: Of 3200 patients randomized (1600 per mode), 878 completed the survey (520 mobile web and 358 telephone). Analyses included all randomized participants (n=3200), with nonresponse addressed through IPW and MI. No adverse events were reported in this survey-based study. The total patient experience score was significantly lower in the mobile web group (mean 81.5, SD 16.4) than in the telephone group (mean 84.9, SD 14.3; unadjusted difference –3.41 points, 95% CI –5.51 to –1.31; IPW-adjusted –4.11, 95% CI –6.17 to –2.04; MI-adjusted –4.59, 95% CI –7.45 to –1.73). Similar patterns were observed across most subdomains. Subgroup analyses revealed consistent mode effects across different demographic categories. Sensitivity analyses using the delta-adjustment method confirmed the robustness of these findings under various missing data scenarios.

Conclusions: Mobile surveys may yield substantially lower patient experience scores than telephone surveys. Unlike previous studies, our study analyzes randomized experimental data under various missingness scenarios and provides evidence that this mode effect is unlikely to be attributable to analytical methods or heterogeneity in respondent characteristics between the 2 survey administration modes. Accordingly, caution is warranted when comparing patient experience scores obtained from traditional telephone surveys with those from mobile surveys. Methodologically, our sensitivity analysis approach provides a robust framework for assessing and addressing potential nonresponse bias in patient experience assessments.

Trial Registration: Clinical Research Information Service KCT0011374; https://tinyurl.com/3e3u5mjs

(J Med Internet Res 2026;28:e79398) doi:10.2196/79398

XSL•FO
RenderX

KEYWORDS

patient experience survey; telephone survey; mobile web survey; mode effect; missing data; sensitivity analysis

Introduction

Background and Rationale

Incorporating the patient's perspective in assessing health care quality is gaining prominence globally, with various countries and international organizations highlighting its significance [1-4]. Consequently, there is a growing effort to systematically and consistently collect data, focusing not only on patient-centered health outcomes but also on patient experience [5,6]. Patient experience concerns the nontechnical aspect of health care quality, as illustrated by the communication between inpatients and their health care providers throughout their hospital stay [7]. While measuring and publicly reporting patient experience ultimately aims to enhance patient experience across the health system, individual hospitals are also key actors in this process [8]. These hospitals are keenly interested in the publicly reported results of patient experience surveys, as these results may significantly affect their reputation and competitiveness in the health care market. A positive correlation has been noted between patient experience and hospital profitability [9]. The stakes are heightened when the results of patient experience surveys are used for third-party payers to adjust hospital reimbursement, aligning with the notion of value-based care [10,11]. For these reasons, the imperative of ensuring a fair comparison becomes a critical methodological consideration in generating hospital-level patient experience results. Nevertheless, hospitals vary in many ways, most notably in terms of different patient characteristics, which can affect hospitals' patient experience scores [12]. This challenge is further compounded by the effects of survey administration mode on who chooses to participate and who does not, as well as on how respondents answer, thereby introducing the central issue of missing data due to nonresponse [12]. As web-based survey administration gains popularity alongside traditional modes such as mail and telephone surveys in patient experience initiatives, a growing challenge is effectively accounting for the influence of these diverse survey modes on differential missingness and the resulting bias in estimating causal mode effects.

The term "mode effect" refers to the difference between the measured score obtained through one survey mode and the counterfactual score that would have been measured if the survey had been conducted by another mode [13,14]. Estimating mode effects, therefore, necessitates an experimental study and careful statistical adjustments based on appropriate assumptions [15]. In the context of patient experience surveys, several countries have conducted rigorous experiments to understand these effects. In the United States, the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey uses multiple modes: mail, telephone, mail with telephone follow-up, and interactive voice response (IVR). An experimental study in 2006 reported that the telephone and IVR modes generally yielded more positive evaluations compared to mail surveys. In England, the National Health Service Adult Inpatient Survey has been exploring the transition to a mixed-mode methodology. A pilot study conducted in 2019 tested two experimental push-to-web approaches against the traditional paper-based method [16]. The study found that online responses tended to be more negative than paper responses.

Statistically, a critical step in estimating mode effects is adjusting for selective nonresponse. The missing-at-random (MAR) assumption is commonly used, positing that score distributions for respondents and nonrespondents remain consistent after accounting for survey mode and other variables influencing both mode selection and scores. If valid, this assumption allows for unbiased estimation of true mode effects by adjusting for respondent composition. Elliott et al [12] reported adjusted mode effects under MAR, while Ipsos MORI (2020) [16] presented unadjusted estimates. However, growing evidence suggests limitations of the MAR assumption in estimating survey mode effects. Multiple studies [17-23] indicate that late responders and nonresponders often report less favorable health care experiences than early responders. Consequently, assessing the robustness of MAR-based results is crucial. Experts recommend sensitivity analyses by re-estimating mode effects under missing-not-at-random (MNAR) scenarios, where distributions of observed and unobserved outcome scores may differ even after adjusting for response probability factors [24-26]. MNAR analyses, while free from the MAR assumption, require additional assumptions about nonrespondents' score distributions. Two common MNAR strategies are the pattern mixture model [27] and the sample selection model [28]. The "delta-adjustment" approach based on the pattern mixture model is often preferred for its fewer restrictive assumptions and easily interpretable results [26,29,30]. Combining MAR-based primary analysis with MNAR-based sensitivity analysis can yield more robust mode effect estimates by examining their stability across different missing data assumptions.

Objectives

This study examines mode effects between telephone and mobile web surveys in South Korea's Patient Experience Assessment (PXA). Our previous study [31] focused on the influence of survey mode on response rates in the PXA. This study, a companion to our previous work, examines the impact of survey mode on patient experience scores, a distinct research question using different outcome measures and analytical approaches. Specifically, we aim to: (1) quantify differences in patient experience scores between telephone and mobile web surveys, (2) examine adjusted differences accounting for potential nonresponse bias under the MAR assumption, and (3) conduct sensitivity analyses using the delta adjustment to assess the robustness of MAR-based findings under the MNAR assumption. By addressing these objectives, we aim to provide valuable insights into the policy considerations for the PXA in South Korea and to contribute to the broader international discourse on survey methodology in patient experience surveys. As this study involves survey administration without clinical interventions, no harm to participants was anticipated.

XSL•FO
RenderX

Methods

Trial Design, Trial Setting, and Eligibility Criteria

This study is a secondary analysis of a randomized controlled trial conducted by our research team at 4 general hospitals in South Korea (2 in Seoul, 1 in Ulsan, and 1 in Gyeongju). Eligible patient lists were obtained from participating hospitals from early October 2022, and the survey was conducted from October 24 (first participant enrolled) to November 18, 2022. The trial was registered retrospectively (December 23, 2025). This study adheres to the CONSORT 2025 (Consolidated Standards of Reporting Trials) guidelines for reporting randomized trials (Checklist 1) [32]. The sample selection in the experiment followed the current PXA framework in South Korea, with minor modifications. Specifically, inclusion criteria were: (1) adults aged 19 years and older, (2) inpatients for >1 day, and (3) discharged 2 - 56 days before survey participation. Exclusion criteria included patients from day clinics, palliative care, pediatrics, neuropsychiatry, and those without completed personal information consent forms. Hospital staff identified eligible patients. The study recruited a total of 4800 patients from 4 participating general hospitals. To ensure representativeness of the subpopulations, stratified sampling was performed based on sex (male and female), age group (19 - 39, 40 - 59, 60 - 69, and 70 years and older), field of care (medical and surgical and others), and hospital of origin (A, B, C, and D). The sampled patients were randomly assigned to 1 of 3 survey modes (telephone, mobile web, and mixed) in a 3-arm parallel-group exploratory design with a 1:1:1 allocation ratio, with individual patients as the unit of randomization. As the goal of our study is to accurately estimate the mode effect between telephone and mobile web surveys, we used the data from 3200 subjects from the telephone and mobile web survey modes (1600 for each). A total of 878 participants (358 for the telephone mode and 520 for the mobile web mode) provided responses to all questions.

Randomization and Blinding

Simple randomization was used, with no blocking or stratification applied. An independent survey company generated the random allocation sequence using computer-generated random numbers (via the NEWID() function in Microsoft SQL Server) and assigned participants to the survey modes. No other parties, including the research team, hospital staff who enrolled participants, interviewers, and participants themselves, had access to the allocation sequence prior to assignment. Due to the nature of survey mode interventions, blinding was not feasible after assignment.

Intervention and Comparator

Survey procedures varied by mode. In the telephone survey, trained interviewers, who received standardized training on the PXA survey protocol prior to the study, made up to 5 daily call attempts, following a standardized protocol. In the mobile web survey, participants received up to 5 survey link messages via an official messenger app. The survey process involved clicking through domain-specific screens and submitting responses. The mobile survey interface was optimized based on usability testing. All participants were informed about the survey beforehand. Each mode followed a similar sequence of identity verification, information provision, questionnaire administration, and conclusion.

Patient and Public Involvement

There was no patient or public involvement in the design, conduct, or reporting of this trial.

Changes to Trial Protocol

No changes were made to the trial protocol after commencement. Subgroup analyses and sensitivity analyses under MNAR scenarios were not prespecified in the original protocol but were conducted post hoc (see Statistical methods).

Outcomes

The prespecified primary outcome was the total patient experience score, calculated as the sum of responses across all 21 questions of the PXA questionnaire. The questionnaire comprises six domains: nurse (Q1–Q4, 4 items), doctor (Q5–Q8, 4 items), medication and treatment (Q9–Q13, 5 items), hospital environment (Q14–Q15, 2 items), patient's rights (Q16–Q19, 4 items), and overall ratings (Q20–Q21, 2 items). Most items use a 4-point Likert scale (1=never or not at all to 4=always or very much), while Q13 uses a binary (yes or no) format, and the 2 overall rating items (Q20–Q21) use a 0 - 10 scale. All item scores were rescaled to a maximum of 100 points per item before summation, adhering to the current policy convention in the Korean PXA. The prespecified secondary outcomes comprised the sum scores for each of the 6 domains, also rescaled to 100 points. The timepoint of outcome assessment was 2 to 56 days after hospital discharge, corresponding to the survey administration period. Outcomes were self-reported by participants in the mobile web group and collected by the trained interviewers in the telephone group.

Harms

As this study involved only survey administration (telephone calls or mobile web survey links) without any clinical interventions, no physical or clinical harm was anticipated. Therefore, systematic harm monitoring was not conducted.

Sample Size

Sample size calculation was based on detecting a 1-point difference in patient experience scores (SD=2, 80% power, α=.05) using G*Power (version 3.1; HHU), requiring 64 respondents per group. Assuming a 10% response rate, this necessitated recruiting 640 participants per mode. To ensure sufficient power for planned subgroup analyses, the study recruited 1600 per mode. No interim analyses were planned or conducted for this study.

Statistical Methods

As a basic characteristic, response rates by survey mode and study participants' characteristics (sex, age group, field of care, and hospital of origin) were reported as counts and percentages. The chi-square test was used to identify statistically significant differences in response rates between modes.

For the analysis of the overall score and six subscores, an unadjusted average score by mode was calculated from

XSL•FO
RenderX

respondents who completed the survey. The unadjusted score difference between the two modes was also reported. As the main analysis, an adjusted difference was estimated under the MAR assumption, ie, the outcome distribution of the nonrespondent group was assumed to be the same as that of the respondent group for each stratum of mode, sex, age group, field of care, and hospital of origin. Both inverse probability weighting (IPW) and multiple imputation (MI) methods were used for the adjustment. The IPW-adjusted difference calculates the weighted mean difference between the respondents in the two modes, where the weights adjust for nonresponse so that the respondent sample better represents the full randomized population. The propensity score model for the weights used logistic regression to predict response probability from the aforementioned stratification variables. Variance for IPW estimates was calculated using robust sandwich estimators to account for weight estimation uncertainty. The MI-adjusted difference calculates the unweighted difference between the two modes from the full dataset where the missing scores of the nonrespondents are imputed by regression. Specifically, we used Bayesian linear regression with the stratification variables as predictors, generating 50 imputed datasets. Variance was obtained by applying Rubin's rules, which properly combine within- and between-imputation variability. It is worth noting that both methods are complementary, as they rely on different statistical assumptions; IPW relies on the assumption that the probability of response for each stratum was accurately estimated, while MI assumes that the mean outcome (score) for each stratum was accurately estimated. For clarity, the target estimand is formally defined in Supplementary Material A.1 (Multimedia Appendix 1), representing the average causal effect of survey mode on patient experience scores among our eligible patient population. It is known that both the IPW and MI-adjusted difference unbiasedly estimate the target estimand under the MAR assumption.

For subgroup analyses, the same outcome analyses were reconducted across subpopulations categorized by sex, age group, and field of care, as these are known to be associated with both survey response patterns and patient experience scores [12,31].

To check the robustness of the results against the MAR assumption, we conducted sensitivity analyses by the delta-adjustment method under the pattern mixture model that assumes the MNAR scenario. The delta-adjustment method is popular for the MAR-sensitivity analyses due to its simplicity for interpretation and implementation [26,29,30]. Specifically, the adjusted score difference between the 2 mode groups was recalculated, assuming that the mean outcomes of the nonrespondent telephone survey group and the nonrespondent mobile web survey group are shifted by δ_1 and δ_2, respectively, compared to their respondent counterparts. The values of δ_1 and δ_2 must be specified by the analyst, and vary from −5 to 5 points in our study. A detailed description and the model formula are deferred to Supplement Material A (Multimedia Appendix 1). For example, if δ_1 is −3, the nonrespondent telephone survey group is assumed to have scores that are 3 points lower than the respondent telephone survey group. The case where $\delta_1=\delta_2=0$ coincides with the MAR assumption. The delta-adjustment method is available via the MI method, which naturally incorporates the shift in mean outcomes.

The primary analysis examining mode effects on patient experience scores was prespecified in the original study protocol. Subgroup analyses and sensitivity analyses using the delta-adjustment method were conducted post hoc to further assess the robustness and generalizability of the primary findings.

All statistical analyses were performed using R (version 4.4.1; R Foundation for Statistical Computing). A *P* value of <.05 was considered statistically significant.

Ethical Considerations

An application was submitted to the Seoul National University Hospital's Institutional Review Board (IRB No. H-2207-134-1342), as well as the IRBs of 3 other general hospitals (IRB Nos. 2022‐1284, Asan Medical Center; UUH2022-07-043, Ulsan University Hospital; and 110757‐202207-HR-03-04, Dongguk University Gyeongju Hospital). All these boards approved the study, which involved patients discharged from these hospitals. It is important to note that the study only included patients who, at the time of admission, had agreed to the use of their personal information by signing the respective hospital's consent form. This was a part of the exclusion criteria to guarantee voluntary participation. All participant data were deidentified before analysis, with only study identification numbers used to link survey responses. No compensation was provided to participants for survey completion. This manuscript does not include any images or information that could identify individual participants.

Clinical Research Information Service KCT0011374 was registered retrospectively on December 23, 2025 (first participant enrolled: October 24, 2022). The trial protocol, the full statistical analysis plan, and preliminary results are accessible through this registry.

Results

Participant Flow, Recruitment, and Baseline Data

Of the 3200 participants randomized to telephone and mobile web modes, 878 completed the survey (Figure 1 for detailed participant flow). Both interventions were delivered as planned, with no deviations from the protocol. No concomitant interventions that could have affected the outcomes were identified. Table 1 presents response rates by survey mode and study participants' characteristics. The overall response rate was 27.4% (878/3200), with a significantly higher rate in the mobile web group (520/1600, 32.5%) compared to the telephone group (22.4%, 358/1600; *P*<.001). Notably, response rates varied across sex, age group, field of care, and hospital of origin, with significant differences between the 2 modes in most categories. This heterogeneity in respondent composition reaffirms the need for appropriate statistical adjustment methods under the MAR or MNAR scenarios to obtain unbiased estimates of mode effects on survey outcomes.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram of the study participants. All 4800 participants were randomly allocated to one of three survey modes (n=1600 each). To estimate the mode effects accurately, the current analysis included only the telephone (n=358) and mobile web (n=520) groups (total n=878). The mixed-mode group was excluded in this comparative analysis.

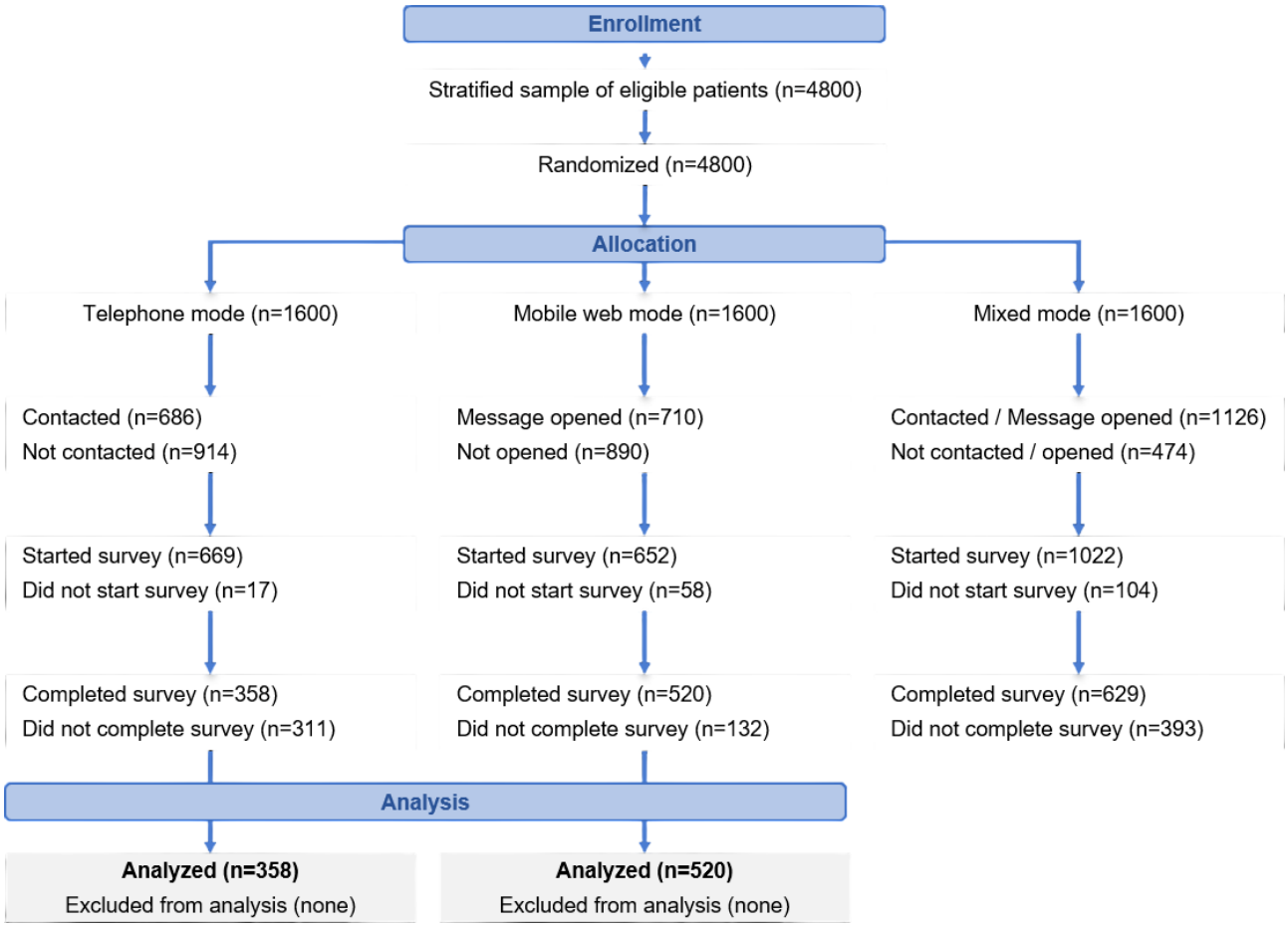

XSL•FO
RenderX

Table . Response rates of the study sample by survey mode and participant characteristics[a].

Characteristics	All (n=3200)	Telephone (n=1600)	Mobile web (n=1600)	*P* value
	Respondents/randomized (%)	Respondents/randomized (%)	Respondents/randomized (%)	
Total	878/3200 (27.4)	358/1600 (22.4)	520/1600 (32.5)	<.001
Sex				
Male	445/1583 (28.1)	195/799 (24.4)	250/784 (31.9)	<.001
Female	433/1617 (26.8)	163/801 (20.3)	270/816 (33.1)	
Age group (years)				
19 - 39	115/414 (27.8)	31/189 (16.4)	84/225 (37.3)	<.001
40 - 59	359/1030 (34.9)	128/525 (24.4)	231/505 (45.7)	
60 - 69	253/878 (28.8)	116/439 (26.4)	137/439 (31.2)	
70+	151/878 (17.2)	83/447 (18.6)	68/431 (15.8)	
Field of care				
Medical	310/1268 (24.4)	123/612 (20.1)	187/656 (28.5)	<.001
Surgical and other	568/1932 (29.4)	235/988 (23.8)	333/944 (35.3)	
Hospital of origin				
A	314/1000 (31.4)	136/500 (27.2)	178/500 (35.6)	<.001
B	288/1002 (28.7)	127/502 (25.3)	161/500 (32.2)	
C	158/666 (23.7)	50/334 (15)	108/332 (32.5)	
D	118/532 (22.2)	45/264 (17)	73/268 (27.2)	

[a]Data from a randomized controlled trial conducted in October–November 2022 at four general hospitals in South Korea. A total of 3200 adult inpatients (discharged 2–56 d prior) were randomly allocated to telephone (n=1600) or mobile web (n=1600) survey modes. The chi-square test was used to compare response rates between modes. Response rate (%) is defined as the number of complete responses divided by the number of patients randomized to each mode.

Numbers Analyzed, Outcomes, and Estimation

Table 2 presents the unadjusted and adjusted estimates for mode effects on the total score and domain scores. The unadjusted estimates were based on the 878 respondents (358 telephone and 520 mobile web), while adjusted estimates using IPW and MI incorporated all 3200 full randomized participants to account for nonresponse. The primary outcome, the total score, was significantly lower in the mobile web group compared to the telephone group, with an unadjusted difference of −3.41 points (95% CI −5.51 to −1.31; *P*=.001). This difference remained significant after adjusting for missing data using the IPW method (−4.11 points; 95% CI −6.17 to −2.04; *P*<.001) and the MI method (−4.59 points; 95% CI −7.45 to −1.73; *P*=.007). Among the secondary outcomes, the hospital environment domain score showed the largest difference between the 2 modes, with an unadjusted difference of −6.63 points (95% CI −9.30 to −3.96; *P*<.001) and adjusted differences of –8.36 points (IPW; 95% CI –10.95 to −5.77; *P*<.001) and −7.81 points (MI; 95% CI −11.42 to −4.20; *P*=.001). The scores in the nurse, doctor, and medication and treatment domains also showed significantly lower scores in the mobile web group compared to the telephone group, both in the unadjusted and adjusted analyses. The patient's rights domain score showed a marginally significant difference in the IPW-adjusted analysis (−2.60 points; 95% CI −5.16 to −0.04; *P*=.05) but not in the unadjusted (−2.05 points; 95% CI −4.61 to 0.52; *P*=.12) or MI-adjusted analyses (−2.92 points; 95% CI −6.93 to 1.09; *P*=.15). This inconsistency between IPW and MI adjustment may be due to the difference in underlying assumptions of the adjustment methods: IPW assumes correct specification of the response propensity model, while MI relies on appropriate modeling of the outcome distribution. Finally, the overall rating score did not show a statistically significant difference between the 2 modes in either the unadjusted or adjusted analyses.

We then assessed whether the mode effect on the total score differed across levels of sex, age group, and field of care (Tables S1–S7 in Multimedia Appendix 1). For the total score and the 6 domain scores, the *P* values for interaction were generally greater than .05 across the subgroups, indicating that the mode effect was consistent within each subgroup and did not significantly differ between subgroups.

XSL•FO
RenderX

Table . Unadjusted and adjusted estimates for mode effects on the total score and domain scores.[a]

Outcomes	Telephone mean (SD)	Mobile mean (SD)	Unadjusted[b] Diff.[c] (95% CI)	*P* value	IPW–Adjusted[b] Diff.[c] (95% CI)	*P* value	MI–Adjusted[b] Diff.[c] (95% CI)	*P* value
Primary outcome								
Total	84.92 (14.34)	81.51 (16.35)	−3.41 (−5.51 to −1.31)	.001	−4.11 (−6.17 to −2.04)	<.001	−4.59 (−7.45 to −1.73)	.007
Secondary outcomes								
Nurse	89.40 (15.20)	86.02 (17.06)	−3.38 (−5.58 to −1.18)	.003	−4.44 (−6.60 to −2.29)	<.001	−4.79 (−7.80 to −1.79)	.005
Doctor	82.51 (18.79)	78.43 (19.93)	−4.08 (−6.71 to −1.46)	.002	−4.97 (−7.56 to −2.39)	<.001	−5.21 (−8.77 to −1.65)	.01
Medication and treatment	85.08 (16.69)	81.98 (18.78)	−3.10 (−5.52 to −0.68)	.01	−3.93 (−6.33 to −1.52)	.001	−4.48 (−7.77 to −1.19)	.01
Hospital environment	86.13 (18.28)	79.50 (20.85)	−6.63 (−9.30 to −3.96)	<.001	−8.36 (−10.95 to −5.77)	<.001	−7.81 (−11.42 to −4.20)	.001
Patient's rights	80.81 (19.83)	78.76 (18.47)	−2.05 (−4.61 to 0.52)	.12	−2.60 (−5.16 to −0.04)	.05	−2.92 (−6.93 to 1.09)	.15
Overall rating	85.22 (16.94)	84.12 (19.10)	–1.10 (−3.56 to 1.36)	.38	0.45 (–1.99 to 2.89)	.72	–0.77 (–3.92 to 2.37)	.63

[a]Results from a randomized controlled trial of 3200 adult inpatients from four general hospitals in South Korea (October-November 2022), with 878 complete responses (358 telephone and 520 mobile web).

[b]Unadjusted estimates are based on complete responses. IPW (inverse probability weighting) and MI (multiple imputation) adjusted estimates account for potential nonresponse bias under the missing-at-random (MAR) assumption.

[c]"Diff." means difference in mean scores (mobile web – telephone). Negative differences indicate lower scores in the mobile web survey compared to the telephone survey. All scores are scaled to a maximum of 100 points.

Harms

No adverse events or unintended effects were identified in either the telephone or mobile web survey group during the study period. This was anticipated, as the intervention involved only survey administration without any clinical procedures. No complaints or adverse reactions were reported during the survey process.

Ancillary Analyses

A sensitivity analysis under MNAR scenarios was conducted to assess the robustness of the MAR assumption by varying the values of δ_1 and δ_2, which represent the assumed differences in mean scores between nonrespondents and respondents for the telephone and mobile web survey groups, respectively. Figure 2 depicts the estimated difference of scores (=mobile web – telephone) for given values of δ_1 and δ_2. For instance, δ_1=0 and δ_2=−2.5 represents a scenario where telephone nonrespondents would have had similar patient experience scores to telephone respondents, while mobile web nonrespondents would have had scores 2.5 points lower than mobile web respondents; under this scenario, the mode effect remains substantial at −6.84 points (95% CI −9.13 to −4.55) for the total score. The figure reveals that as δ_1 increases and δ_2 decreases, the mode effect between telephone and mobile web surveys becomes more pronounced in the negative direction. In other words, if nonrespondents in the telephone survey are assumed to give higher scores than respondents, while nonrespondents in the mobile web survey are assumed to give lower scores than respondents, the negative mode effect is amplified. Conversely, when δ_1 decreases and/or δ_2 increases, the mode effect is attenuated, suggesting that if nonrespondents in the telephone survey are assumed to give lower scores than respondents, while nonrespondents in the mobile web survey are assumed to give higher scores than respondents, the difference between the 2 modes is reduced. Overall, most of the plots for the total score and domain scores show a negative mode effect, consistent with the IPW and MI results presented in Table 2. The telephone survey would only yield better scores than the mobile web survey under the

XSL•FO
RenderX

artificial assumption that δ_1 is −5 and δ_2 is positive. Even if δ_1 is set to 0, δ_2 would have to be less than 2.5 for the telephone survey to show higher scores than the mobile web survey across most domains. These findings suggest that, even under the MNAR scenarios, the possibility of the mobile web survey yielding higher scores than the telephone survey is highly unlikely, except under some unusual combinations of δ_1 and δ_2 values.

Figure 2. Estimated mode effects by the delta-adjustment method under a missing-not-at-random (MNAR) scenario, varying values of δ_1 and δ_2. MI: multiple imputation.

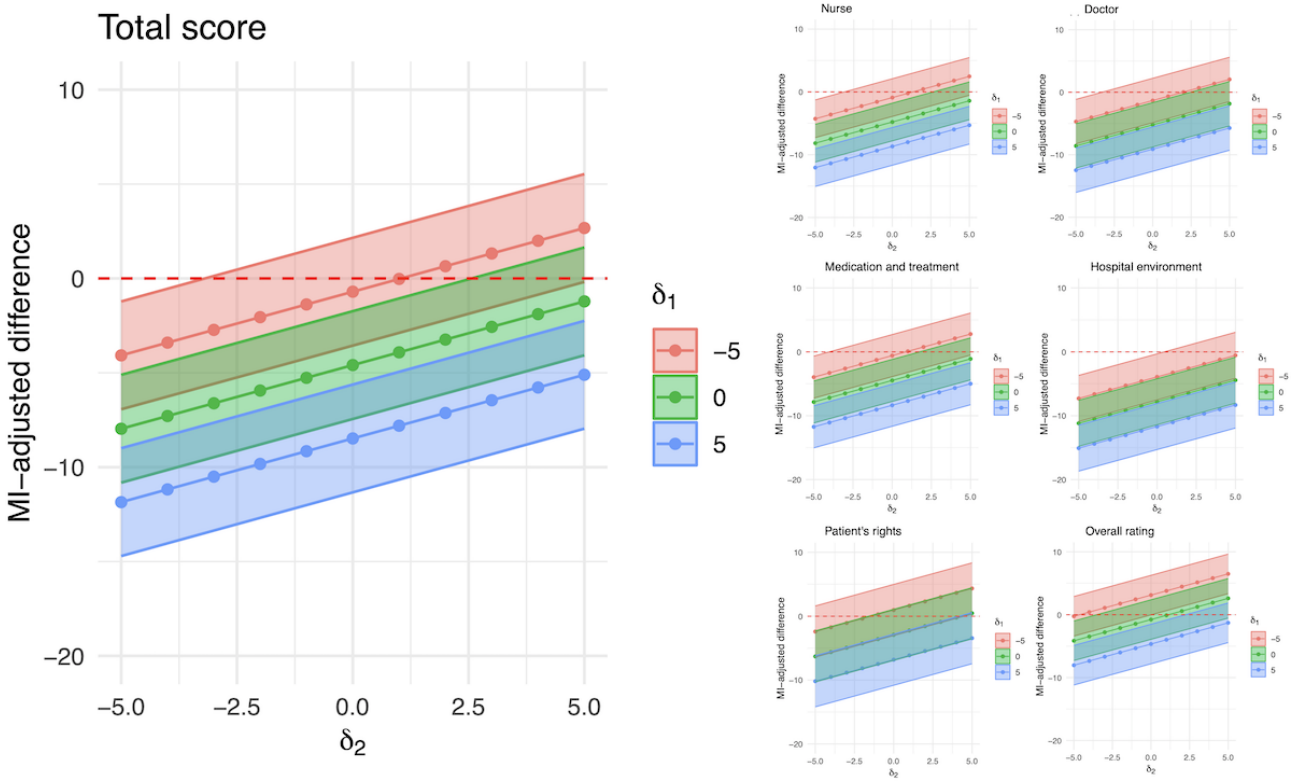

Discussion

Interpretation

In accordance with our study objectives, we examined mode effects between mobile web and telephone surveys on patient experience scores under various missing data assumptions. The main analyses consistently showed that, under the widely accepted MAR assumption, the mobile web survey yielded significantly lower scores compared to the telephone survey for both the total score and almost all domain scores. The sensitivity analysis under the MNAR assumption, our third objective, further supported this finding, revealing that in most plausible scenarios, the mobile web survey scores remained lower than the telephone survey scores, even when considering potential departures from the MAR assumption. Only under extreme and unlikely scenarios where the nonrespondent telephone survey group was assumed to have substantially lower scores than the respondent telephone survey group (eg, δ_1=−5) while simultaneously assuming the nonrespondent mobile web survey group had higher scores than the respondent mobile web survey group (eg, δ_2>0), did the pattern reverse. These findings highlight the robustness of the observed mode effect, suggesting that the lower scores in the mobile web survey are not merely an artifact of the MAR assumption. Moreover, the randomized controlled design ensures that these mode effects are not attributable to confounding by respondent characteristics between the 2 survey modes.

Findings in Relation to Other Evidence

Our findings are consistent with the existing literature indicating that interviewer-led surveys, whether conducted face-to-face or via telephone, often yield more positive responses to scale questions compared to self-administered surveys, such as those completed by mail or online [12,33-38]. This pattern extends well beyond health care surveys. Leeuw [37] meta-analysis established that interviewer-led modes consistently elicit more socially desirable responses than self-administered modes. Dillman et al [34] experimentally demonstrated that aural modes (telephone and IVR) produced significantly more extreme positive responses to satisfaction questions than visual modes (mail and web). A large-scale experiment by the Pew Research Center in 2015 [38], randomizing respondents across 60 questions, reported an average telephone–web difference of approximately 5 percentage points, with the largest gaps on topics susceptible to social desirability. Together, these findings indicate that the tendency for interviewer-led surveys to yield more positive responses is robust across diverse topics, populations, and cultural contexts.

In the context of patient experience surveys, our findings align with Elliott et al HCAHPS study [12], showing consistent mode effects across different cultural contexts and health care systems. Both studies found that telephone surveys yielded more positive

XSL•FO
RenderX

evaluations than self-administered modes (mail or mixed in HCAHPS, mobile web in ours). Interestingly, both studies found a significant mode effect for the Nurse domain and a nonsignificant effect for the Overall ratings domain. These similarities across different health care systems, languages, and survey designs further support the robustness of mode effects and highlight the importance of considering these effects in patient experience assessments, regardless of the specific context.

The observed mode effect can be attributed to various factors [33-35]. First, the social setting of the survey itself can influence responses: interview modes involve a social interaction between respondent and interviewer, which is absent in self-administered modes, potentially affecting response patterns [33]. Second, social desirability bias may lead respondents to provide more socially acceptable answers in the presence of an interviewer, resulting in an overreporting of desirable behaviors and higher scores in telephone surveys [33,34]. Third, acquiescence bias, or the tendency to agree with statements or questions when in doubt, may be more prominent in interview situations [33,34]. Fourth, the recency effect—where respondents tend to select the last option presented in an auditory mode—can lead to higher selection of positive top-box categories in telephone surveys, as these are typically read last [34,35]. These factors may collectively contribute to the higher scores observed in the telephone survey mode compared to the self-administered mobile web mode across most domains.

Research Implications

The sensitivity analysis in this study helps to assess the robustness of results against potential deviations from the MAR assumption. Multiple imputation with delta adjustment offers a method for imputing missing data under various MNAR scenarios [26,29,30]. By systematically varying the delta parameter, researchers can explore how their conclusions might change under different missing data mechanisms. This approach is particularly valuable when nonresponse is prevalent [39-41], as in the context of patient experience surveys, where mode-specific nonresponse patterns may introduce bias that standard MAR-based methods cannot detect.

While our analysis assumed a constant shift parameter (ie, a fixed degree of departure from MAR) for all individuals within each survey mode group, one could explore the possibility of varying delta values across different individuals in the same group. For instance, in an HIV longitudinal study [40], for participants without actual outcomes (serological HIV test results), the researchers assumed different delta values based on the participants' self-reported HIV status. This approach allowed for a more nuanced sensitivity analysis, accounting for potential differences in the missing data mechanism among subgroups within the same study group.

Policy Implications

Our results have 2 important policy implications for South Korea's PXA. First, by demonstrating that survey modes have substantial effects on patient experience scores, our findings imply that mixing different modes in South Korea's PXA would introduce additional methodological challenges for making fair comparisons across hospitals. Efforts to improve response rates in the PXA have included introducing a mobile web survey either in place of or alongside the existing telephone survey [31]. However, mixing 2 different modes would have required adjustments for mode effects. In 2023, the PXA survey switched to a single mode of mobile web, thereby eliminating the need to account for mode effects between survey methods [42]. Second, caution is still needed when comparing PXA scores between 2023 and previous years. Based on our findings, PXA scores in 2023 are expected to show a decline, all else being equal, due to the mode effect of the mobile web survey compared to the telephone survey used in prior years. Policymakers, the general public, and the media should anticipate lower scores in 2023 due to this mode effect and should not interpret this potential decline as an indication of worsening patient experience over time.

Limitations

This study has several limitations. First, using data from patients at 4 specific hospitals in South Korea may limit the generalizability of the findings to other health care and cultural settings. Second, because the study compares mobile web with telephone survey modes, the results are less generalizable to populations with relatively low mobile device usage. Third, the use of sum scores for analysis, while consistent with current policy practices, may obscure nuances in individual item responses that could be affected differently by survey mode. Fourth, our study cannot identify or disentangle the specific mechanisms underlying the observed mode effects, including social interaction effects, social desirability bias, acquiescence bias, and the recency effect.

Conclusions

Mobile surveys may yield substantially lower patient experience scores than telephone surveys. Unlike previous studies, our study analyzes randomized experimental data under various missingness scenarios and provides evidence that this mode effect is unlikely to be attributable to analytical methods or heterogeneity in respondent characteristics between the 2 survey administration modes. Accordingly, caution is warranted when comparing patient experience scores obtained from traditional telephone surveys with those from mobile surveys. Methodologically, our sensitivity analysis approach provides a robust framework for assessing and addressing potential nonresponse bias in patient experience assessments.

Acknowledgments

Data for this study were made available with support from the Health Insurance Review and Assessment Service (HIRA). The views expressed in this article are those of the authors and not necessarily the views of the HIRA.

Disclosure of generative artificial intelligence (GenAI) use

XSL•FO
RenderX

The authors declare the use of GenAI in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: code generation, visualization, reproducibility testing, proofreading and editing, and reformatting. The GenAI tool used was Claude Opus 4.6. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

YGC's work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No. RS-2026-25484084). The funder had no role in the design, conduct, data analysis, interpretation, or reporting of this study.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to data-use restrictions of the Health Insurance Review and Assessment Services (HIRA). Interested readers may contact HIRA for data access. The statistical code (R) used for the analyses is available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Statistical analysis details and additional results.
[DOCX File, 52 KB - jmir_v28i1e79398_app1.docx]

Checklist 1
CONSORT 2025 checklist.
[DOCX File, 37 KB - jmir_v28i1e79398_app2.docx]

Checklist 2
CONSORT e-HEALTH (V1.6.1) checklist.
[PDF File, 884 KB - jmir_v28i1e79398_app3.pdf]

Checklist 3
Items to include when reporting a randomized trial in a journal or conference abstract.
[PDF File, 31 KB - jmir_v28i1e79398_app4.pdf]

References

1. Coulter A, Cleary PD. Patients' experiences with hospital care in five countries. Health Aff (Millwood) 2001;20(3):244-252. [doi: 10.1377/hlthaff.20.3.244] [Medline: 11585173]
2. Valentine N, Darby C, Bonsel GJ. Which aspects of non-clinical quality of care are most important? Results from WHO's general population surveys of "health systems responsiveness" in 41 countries. Soc Sci Med 2008 May;66(9):1939-1950. [doi: 10.1016/j.socscimed.2007.12.002] [Medline: 18313822]
3. Cleary PD. Evolving concepts of patient-centered care and the assessment of patient care experiences: optimism and opposition. J Health Polit Policy Law 2016 Aug;41(4):675-696. [doi: 10.1215/03616878-3620881] [Medline: 27127265]
4. Fujisawa R, Klazinga N. Measuring patient experiences (PREMS): progress made by the OECD and its member countries between 2006 and 2016. : Organisation for Economic Co-operation and Development; 2017. [doi: 10.1787/893a07d2-en]
5. de Bienassis K, Kristensen S, Hewlett E, Roe D, Mainz J, Klazinga N. Measuring patient voice matters: setting the scene for patient-reported indicators. Int J Qual Health Care 2021 Feb 20;34(Suppl 1):ii3-ii6. [doi: 10.1093/intqhc/mzab002] [Medline: 33575802]
6. Knapp A, Harst L, Hager S, Schmitt J, Scheibe M. Use of patient-reported outcome measures and patient-reported experience measures within evaluation studies of telemedicine applications: systematic review. J Med Internet Res 2021 Nov 17;23(11):e30042. [doi: 10.2196/30042] [Medline: 34523604]
7. Horton DJ, Yarbrough PM, Wanner N, Murphy RD, Kukhareva PV, Kawamoto K. Improving physician communication with patients as measured by HCAHPS using a standardized communication model. Am J Med Qual 2017;32(6):617-624. [doi: 10.1177/1062860616689592] [Medline: 28693347]
8. Contandriopoulos D, Champagne F, Denis JL. The multiple causal pathways between performance measures' use and effects. Med Care Res Rev 2014 Feb;71(1):3-20. [doi: 10.1177/1077558713496320] [Medline: 23877955]
9. Richter JP, Muhlestein DB. Patient experience and hospital profitability: is there a link? Health Care Manage Rev 2017;42(3):247-257. [doi: 10.1097/HMR.0000000000000105] [Medline: 27050925]

XSL•FO
RenderX

10. Petrullo KA, Lamar S, Nwankwo-Otti O, Alexander-Mills K, Viola D. The Patient Satisfaction Survey: what does it mean to your bottom line? JHA 2013;2(2):1. [doi: 10.5430/jha.v2n2p1]
11. Hong YR, Nguyen O, Yadav S, et al. Early performance of hospital value-based purchasing program in Medicare: a systematic review. Med Care 2020 Aug;58(8):734-743. [doi: 10.1097/MLR.0000000000001354] [Medline: 32692140]
12. Elliott MN, Zaslavsky AM, Goldstein E, et al. Effects of survey mode, patient mix, and nonresponse on CAHPS hospital survey scores. Health Serv Res 2009 Apr;44(2 Pt 1):501-518. [doi: 10.1111/j.1475-6773.2008.00914.x] [Medline: 19317857]
13. Vannieuwenhuyze J, Loosveldt G, Molenberghs G. A method for evaluating mode effects in mixed-mode surveys. Public Opin Q 2010 Jan 1;74(5):1027-1045. [doi: 10.1093/poq/nfq059]
14. Clarke PS, Bao Y. Estimating mode effects from a sequential mixed-mode experiment using structural moment models. Ann Appl Stat 2022;16(3):1563-1585. [doi: 10.1214/21-AOAS1557]
15. Burroughs TE, Waterman BM, Cira JC, Desikan R, Claiborne Dunagan W. Patient satisfaction measurement strategies: a comparison of phone and mail methods. Jt Comm J Qual Improv 2001 Jul;27(7):349-361. [doi: 10.1016/s1070-3241(01)27030-8] [Medline: 11433626]
16. NHS adult inpatient survey: findings from the mixed-mode methodology pilot. : Ipsos MORI Social Research Institute; 2020 URL: https://nhssurveys.org/wp-content/surveys/06-development-work/06-engagement-work/2020/Adult%20Inpatient%202019%20Mixed-mode%20pilot%20results.pdf [accessed 2026-02-20]
17. Rubin HR. Can patients evaluate the quality of hospital care? Med Care Rev 1990;47(3):267-326. [doi: 10.1177/107755879004700302] [Medline: 10108049]
18. Barkley WM, Furse DH. Changing priorities for improvement: the impact of low response rates in patient satisfaction. Jt Comm J Qual Improv 1996 Jun;22(6):427-433. [doi: 10.1016/s1070-3241(16)30245-0] [Medline: 8806045]
19. Etter JF, Perneger TV, Rougemont A. Does sponsorship matter in patient satisfaction surveys? a randomized trial. Med Care 1996 Apr;34(4):327-335. [doi: 10.1097/00005650-199604000-00004] [Medline: 8606557]
20. Lasek RJ, Barkley W, Harper DL, Rosenthal GE. An evaluation of the impact of nonresponse bias on patient satisfaction surveys. Med Care 1997 Jun;35(6):646-652. [doi: 10.1097/00005650-199706000-00009] [Medline: 9191708]
21. Mazor KM, Clauser BE, Field T, Yood RA, Gurwitz JH. A demonstration of the impact of response bias on the results of patient satisfaction surveys. Health Serv Res 2002 Oct;37(5):1403-1417. [doi: 10.1111/1475-6773.11194] [Medline: 12479503]
22. Zaslavsky AM, Zaborski LB, Cleary PD. Factors affecting response rates to the consumer assessment of health plans study survey. Med Care 2002 Jun;40(6):485-499. [doi: 10.1097/00005650-200206000-00006] [Medline: 12021675]
23. Elliott MN, Edwards C, Angeles J, Hambarsoomians K, Hays RD. Patterns of unit and item nonresponse in the CAHPS hospital survey. Health Serv Res 2005 Dec;40(6 Pt 2):2096-2119. [doi: 10.1111/j.1475-6773.2005.00476.x] [Medline: 16316440]
24. The Prevention and Treatment of Missing Data in Clinical Trials: National Research Council; 2010.
25. Permutt T. Sensitivity analysis for missing data in regulatory submissions. Stat Med 2016 Jul 30;35(17):2876-2879. [doi: 10.1002/sim.6753] [Medline: 26567763]
26. Little RJ, D'Agostino R, Cohen ML, et al. The prevention and treatment of missing data in clinical trials. N Engl J Med 2012 Oct 4;367(14):1355-1360. [doi: 10.1056/NEJMsr1203730] [Medline: 23034025]
27. Glynn RJ, Laird NM, Rubin DB. Selection modeling versus mixture modeling with nonignorable nonresponse. In: Wainer H, editor. Drawing Inferences from Self-Selected Samples: Springer; 1986:115-142. [doi: 10.1007/978-1-4612-4976-4_10]
28. Heckman JJ. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Ann Econ Soc Meas 1976;5(4):475-492 [FREE Full text]
29. Liublinska V, Rubin DB. Sensitivity analysis for a partially missing binary outcome in a two-arm randomized clinical trial. Stat Med 2014 Oct 30;33(24):4170-4185. [doi: 10.1002/sim.6197] [Medline: 24845086]
30. Rubin DB. Multiple Imputation for Nonresponse in Surveys: Wiley; 2009.
31. Koo BM, Song Y, Choi YG, et al. Comparing response rates between mobile web and telephone surveys for patient experience: a randomised experimental study in South Korea. BMJ Open 2026 Mar 24;16(3):e103862. [doi: 10.1136/bmjopen-2025-103862] [Medline: 41876163]
32. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ 2025 Apr 14;389:e081123. [doi: 10.1136/bmj-2024-081123] [Medline: 40228833]
33. Bowling A. Mode of questionnaire administration can have serious effects on data quality. J Public Health (Bangkok) 2005 Sep 1;27(3):281-291. [doi: 10.1093/pubmed/fdi031]
34. Dillman DA, Phelps G, Tortora R, et al. Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the internet. Soc Sci Res 2009 Mar;38(1):1-18. [doi: 10.1016/j.ssresearch.2008.03.007]
35. Elliott MN, Beckett MK, Hambarsoomian K, et al. Updated adjustment of the HCAHPS survey for new modes of survey administration and patient mix. Med Care 2025;63(5):358-365. [doi: 10.1097/MLR.0000000000002127]
36. Christensen AI, Ekholm O, Glümer C, Juel K. Effect of survey mode on response patterns: comparison of face-to-face and self-administered modes in health surveys. Eur J Public Health 2014 Apr;24(2):327-332. [doi: 10.1093/eurpub/ckt067] [Medline: 23766339]

XSL•FO
RenderX

37. Leeuw E. Data Quality in Mail, Telephone and Face to Face Surveys, th edition: TT-Publikaties; 1992.
38. Keeter S, McGeeney K, Igielnik R, Mercer A, Mathiowetz NA. From telephone to the web: the challenge of mode of interview effects in public opinion polls. Pew Research Center. 2015. URL: https://www.pewresearch.org/methods/2015/05/13/from-telephone-to-the-web-the-challenge-of-mode-of-interview-effects-in-public-opinion-polls/ [accessed 2026-02-20]
39. Gachau S, Quartagno M, Njagi EN, Owuor N, English M, Ayieko P. Handling missing data in modelling quality of clinician-prescribed routine care: sensitivity analysis of departure from missing at random assumption. Stat Methods Med Res 2020 Oct;29(10):3076-3092. [doi: 10.1177/0962280220918279] [Medline: 32390503]
40. Leacy FP, Floyd S, Yates TA, White IR. Analyses of sensitivity to the missing-at-random assumption using multiple imputation with delta adjustment: application to a tuberculosis/HIV prevalence survey with incomplete HIV-status data. Am J Epidemiol 2017 Feb 15;185(4):304-315. [doi: 10.1093/aje/kww107] [Medline: 28073767]
41. García-Hernandez A, Pérez T, Del Carmen Pardo M, Rizopoulos D. An illness-death multistate model to implement delta adjustment and reference-based imputation with time-to-event endpoints. Pharm Stat 2024;23(2):219-241. [doi: 10.1002/pst.2348] [Medline: 37940608]
42. Release of the 4th (2023) patient experience assessment results - introduction of user-friendly mobile web survey, all six domains averaging above 80 points [Webpage in Korean]. Health Insurance Review and Assessment Service. 2024. URL: https://www.hira.or.kr/bbsDummy.do?pgmid=HIRAA020041000100&brdScnBltNo=4&brdBltNo=11264&pageIndex=1 [accessed 2026-02-20]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
HCAHPS: Hospital Consumer Assessment of Healthcare Providers and Systems
IPW: inverse probability weighting
IVR: interactive voice response
MAR: missing-at-random
MI: multiple imputation
MNAR: missing-not-at-random
PXA: Patient Experience Assessment

Edited by S Brini; submitted 23.Jun.2025; peer-reviewed by S Park, Y Lee; revised version received 14.Mar.2026; accepted 17.Mar.2026; published 14.May.2026.

Please cite as:
Choi YG, Koo BM, Song Y, Do YK
Mode Effects Between Mobile Web and Telephone Surveys on Patient Experience Scores in South Korea: Secondary Analysis of a Randomized Controlled Trial Under Various Missingness Scenarios
J Med Internet Res 2026;28:e79398
URL: https://www.jmir.org/2026/1/e79398
doi:10.2196/79398

© Young-Geun Choi, Bon Mi Koo, Yeongchae Song, Young Kyung Do. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Fusing Specialized Surveys of Rare Populations to Larger Surveys for Generalized Inference: Cross-Sectional Survey Study

Karilynn M Rockhill[1], PhD; Elizabeth A Bemis[1], MPH; Nicole Schow[1], MS; Heather A Olsen[1], MPH; Kyle Beekman[1], MPH; Evelyn J Fox[1], MPH; Andrew A Monte[1,2], MD, PhD; Joshua C Black[1], PhD

[1]Rocky Mountain Poison & Drug Safety, Denver Health and Hospital Authority, 777 Bannock Street, M/C 0180, Denver, CO, United States

[2]Emergency Medicine Department, University of Colorado School of Medicine, Aurora, CO, United States

Corresponding Author:
Joshua C Black, PhD
Rocky Mountain Poison & Drug Safety, Denver Health and Hospital Authority, 777 Bannock Street, M/C 0180, Denver, CO, United States

Abstract

Background: Mainstays of pharmacoepidemiology are large, representative, behavioral surveys, which focus on many drugs with few detailed behaviors. Smaller, targeted studies measure drug-specific patterns but without explicit generalizability assumptions; the evidence generated is narrow.

Objective: In this cross-sectional survey study, we outline an estimation framework based on data fusion and combine two surveys: a representative, anchor survey and an enriched survey about psychedelic drugs in the United States. Application of calibration weighting transports estimates from the enriched survey to the anchor survey.

Methods: The psychedelic-focused enriched survey was sampled twice from a commercial online panel of adults from April 19 to June 25, 2024, (n=2306; 40.4% female, 33.1 y median age) and January 24 to March 21, 2025 (n=2023; 39.6% female, 35.2 y median age). The anchor survey was sampled twice from a different online panel from March 13 to May 6 2024 (n=28,679 total; 2430 using psychedelics) and 14 August to 9 October 2024 (n=29,040 total; 2309 using psychedelics). Internal consistency (transport bias, the absolute difference between the weighted estimates from the anchor survey and weighted fused survey) and external validity (root-mean-square error, RMSE, of self-reported demographic, health, and substance use estimates to probability-based benchmarks) metrics were calculated. The methodology was applied to estimate reasons for using specific psychedelic drugs.

Results: Without adjustments, the enriched surveys had lower percentages of male and White respondents, lower self-perceived health, and higher cigarette use. A total of 2048 weighting schemes were tested, with good internal consistency. Average transport biases with the final weighting scheme were: demographics, 0.09 percentage points; health characteristics, 0.35 percentage points; and substance use, 0.22 percentage points. Estimates after fusion were externally consistent with benchmarks. RMSE increased by 3.3% for demographic estimates (1.82 unweighted to 1.88 weighted); larger decreases were observed for health RMSE (7.30 to 3.38, 53.7% decrease) and for substance use RMSE (6.56 to 6.03, 8.1% decrease). Alcohol use substantially increased the RMSE, likely due to question differences (without alcohol, the RMSE decreased from 6.03 to 1.55). Using the fused dataset, recreational use of psilocybin (92.9%, 95% CI 91.1, 94.7), LSD (93.2%, CI 90.1, 96.4, and MDMA (93.3%, 91.0, 95.6) was more common than medical use (30.9%, CI 27.6, 34.2; 26.4%, CI 21.1, 31.7; and 21.1%, CI 17.5, 24.7, respectively).

Conclusions: Building upon past data fusion research, this study fused two surveys for the purpose of surveillance. This methodology, termed the “fused survey design,” is a rigorous but accessible approach for surveilling rare behaviors like drug use, and we demonstrated constructs absent from anchor surveys may be measured with generalizable inference. This expands the surveillance epidemiology toolbox, giving researchers an actionable process to field enriched surveys with specialized questions that would be impractical to add to larger surveys due to space constraints and respondent fatigue.

(J Med Internet Res 2026;28:e86059) doi:10.2196/86059

KEYWORDS

online survey; data fusion; psychedelics; statistical adjustment; psilocybin; lysergic acid diethylamide; LSD; methylenedioxymethamphetamine; MDMA; representativeness

XSL•FO
RenderX

Introduction

Problem

Mainstays of epidemiology and surveillance of public health trends are large, representative surveys of human behavior. In pharmacoepidemiology, these surveys focus on drug use behaviors and health outcomes, of which there are several surveys across many countries [1-6]. These surveys study many drugs and thereby include only a few questions per drug; in essence, prioritizing breadth of knowledge over depth of knowledge about specific drugs. Large surveys provide essential information about substance use prevalence and correlates with demographics and predictors of health across multiple drug classes [7-11]. Smaller, targeted studies frequently measure drug-specific patterns, generating insight into how specific drugs are used in context, perceptions of use and harm reduction, and subgroup analyses that are challenging to field at scale [12-16]. Results from these smaller surveys are meaningful within the context of their sampling frame, but without explicit assumptions detailing the generalizability, the evidence generated may be narrow.

Review of Relevant Scholarship

The lack of generalizable estimates from smaller surveys presents a specific problem for surveilling psychedelic drug use. While use patterns are changing globally across many classes [17], psychedelic drugs, particularly psilocybin have rapidly grown in popularity [18,19]. From 2015 to 2022, the Global Drug Survey demonstrated increased prevalence of use of psilocybin, lysergic acid diethylamide (LSD), 3,4-methylenedioxymethamphetamine (MDMA), ketamine (a dissociative drug with psychedelic properties), and to a lesser extent 4-bromo-2,5-dimethoxyphenethylamine across Europe, the United States, Australia, and Brazil [1]. Policy makers will likely need to adapt regulations, such as has already occurred across the United States [20], presenting an immediate surveillance need to study the motivations, experiences, and attendant health benefits and risks of targeted subgroups in a generalizable framework [21,22]. However, despite increasing use and popularity, prevalence remains rare compared to other substances like alcohol, tobacco, or cannabis, and there is a vast spectrum of behaviors that may influence health and safety. In response, psychedelic drugs have become a subject for targeted survey studies using anonymous online surveys [21,23,24], which may be advantageous for encouraging accurate reporting of stigmatized behaviors. However, online recruitment suffers this problem of lack of generalizability to target populations of interest [25]. Therefore, solving the generalizability problem would substantially improve surveillance capacity for psychedelic drugs, presenting better data for policy-making [22].

Data fusion and transport weighting methodology can overcome this problem [26-28]. Selection bias between data sources must be corrected to successfully fuse data if descriptive inference is to be generalizable, and, if causal inference is the objective, selection diagrams can decode the necessary assumptions [26]. Prior research demonstrates the utility of fusing multiple surveys to enhance data collection. An early example fused two surveys to improve accuracy of prevalence estimates of diabetes and cardiovascular disease [29]. Analogous data fusion procedures have been suggested for nonprobability surveys to fill gaps in low response-rate probability-based surveys [30], used to compare different generations of veterans' health and social wellbeing [31], and used to combine multiple survey data for depression diagnoses [32]. In these cases, the topic under study was prevalent in the applicable populations. In the present study, we expand upon this foundation, demonstrating valid inference for combining data from different surveys to study rare subpopulations.

Study Objectives

Our objectives were to: (1) describe the application of data fusion to population survey inference, (2) apply a quantitative framework for testing external validity of estimates produced from fusing surveys, and (3) demonstrate a practical application by estimating reasons for psychedelic drug use in the United States by fusing a small, targeted survey to a large, representative survey.

Methods

Design Overview and Assumptions

We created a two-step data fusion inspired approach to combine surveys for enhanced inference [26,27,33]. Figure 1 outlines the conceptual framework describing two survey samples, collected separately, with different recruitment strategies, with the intent of ultimately generating inference to the general population (Figure 1C, black outline). A large general survey without enriched recruitment (Figure 1B, blue outline) is collected with a well-established sampling frame; we term this the "anchor survey." The smaller, targeted survey is collected, with additional selection biases due to targeted sampling; we term this the "enriched survey" (Figure 1A, green outline).

XSL•FO
RenderX

Figure 1. Conceptual diagram of fusing an enriched sample with an anchor sample to create generalizable estimates. A is an enriched sample collected with selection bias due to the targeted recruitment regarding the specialized content. B is an anchor sample collected with selection bias but without specific intent upon recruiting a profile of behavior. C is the target population where inference is intended.

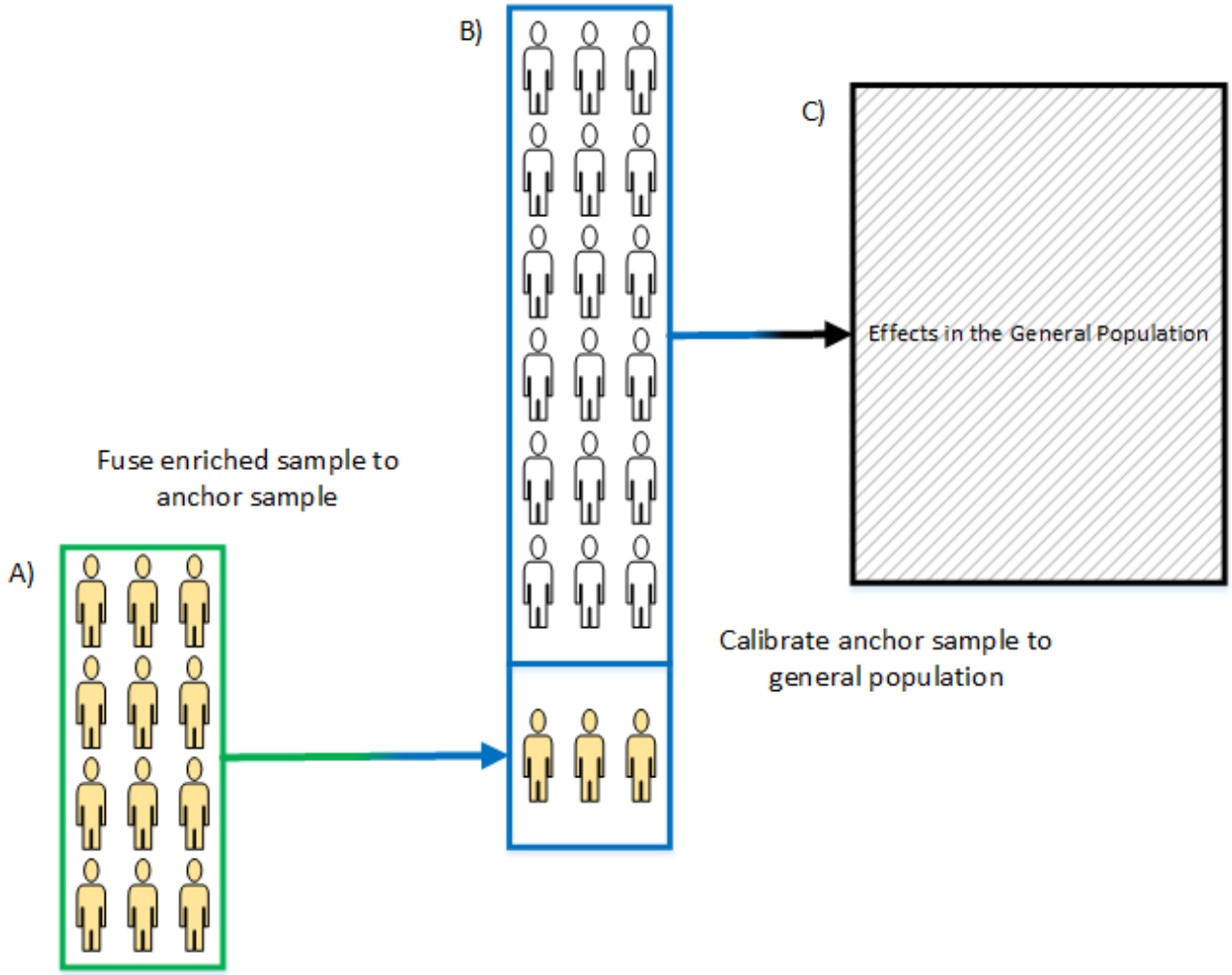

For clarity, we define terms describing related phenomena which have slightly different meanings. In this context, “data fusion” refers to the process of combining two surveys into a single dataset for analysis. “Calibration” refers to the statistical process of adjusting weights so that variable distributions match between the surveys or benchmark population. “Transport weights” refer to individual-level values that, when used to generate statistical estimates, calibrate the enriched sample to the anchor sample; “transport weighting scheme” refers to the set of variables used to create the transport weights.

This framework partitions correction of selection bias into two steps requiring three primary assumptions. First, intersurvey calibration makes the enriched sample match the population corrected values of the analogous subsample in the anchor survey (shown in yellow in Figure 1B). Accurate calibration (A to B in Figure 1) requires an assumption that all factors that induce a selection bias between surveys are accounted for and measurable between the two surveys. If a factor produces differential probability of recruitment between the two surveys and is associated with the outcome, then this would induce a bias to estimates. In our practical application, we verified this assumption by examining internal consistency between the fused survey and anchor survey across a range of metrics likely associated with psychedelic use. Second, the fused survey is calibrated for population inference, which assumes generalizing weights are identifiable (B to C in Figure 1). We verify this assumption in our practical application by examining external validity across a range of metrics found in benchmark surveys. Both steps use generalized raking to calibrate [34]. Generalized raking is well suited for this because many variables associated with selection bias need adjustment and raking does not require complete cross-stratified distributions. This approach also assumes differential measurement bias between samples is negligible, and that any difference that could be observed was due to differences in selection.

Data Sources and Sampling Procedures

The enriched survey, the National Survey Investigating Hallucinogenic Trends (NSIHT), collected detailed information about last 12-month use of psychedelic drugs. This enriched survey was fused to an existing, well-established anchor survey of the general adult population, the Survey of Non-Medical Use of Prescription Drugs (NMURx) [3]. The study plan with design details was pre-registered at the Open Science Framework, but is briefly summarized here and deviations are described in the supplement [35].

XSL•FO
RenderX

The psychedelic-focused enriched survey was sampled from a commercial online panel administered by the commercial research company YouGov (Palo Alto, California, US) in two waves, from April 19 to June 25 , 2024 and from January 24 to March 21, 2025. Respondents were required to answer all questions; therefore, there was no missing data. The first wave was used to develop and select the least biased transport weights, and the second wave replicated the results and tested external validity. For both waves, participants consented to being surveyed, were 18 - 110 years old, and self-reported using at least one psychedelic drug in the last 12 months. Quotas based on age by biological sex and region of residence balanced initial distribution of these groups. The recruitment pool was a group of individuals who take surveys for modest compensation on a variety of topics (eg, health, sports, politics, consumer products) and are a neutral pool of participants with regards to drug use. As an online survey, respondents demonstrating careless responding (ie, failing to give necessary attention to answering questions) [36] were removed to reduce measurement error from misclassification.

The anchor survey, NMURx, was recruited from Kantar's Profiles division (London, United Kingdom) proprietary online panel network, which has an established set of survey weights previously validated for drug use estimates [3]. As with the enriched survey, respondents were 18 - 110 years old, and were required to answer all questions. This survey included an identifiable subsample who used psychedelic drugs in the last 12 months. Much like other large drug use surveys, NMURx prioritized breadth, not depth of psychedelic use. The survey was fielded twice, from March 13 to May 6 and from August 14 to October 9, 2024. The first wave was used to develop and select least biased transport weights and the second used to replicate the results and generate final estimates.

Two-Step Calibration

In the fusion step (1A to 1B in Figure 1), intersurvey recruitment differences, and therefore differential selection forces, were calibrated using generalized raking. We took an empirical approach to satisfying the assumption that selection differences were accounted for in the transport weighting scheme, because we did not *a priori* know which variables may best represent selection differences between surveys. Using a previously published framework for studying psychedelic use in the real world as a guide [22], several possible variables were included on the enriched survey that may represent selection differences specifically related to psychedelic use. We included variables describing demographics, recent substance use, health, spirituality, and trust. Through multiple variable sets, we iterated testing how well each reduced overall bias relative to federal benchmark estimates. Variables in the enriched survey were calibrated to match the weighted values of the psychedelic use subsample in the anchor survey, resulting in transport weights that correct for composition differences between the two surveys (shown in yellow in Figure 1B). To test different weighting schemes, variables were split into two groups: required and iterated (Table 1). Required variables included a base set of demographic and health variables that we have demonstrated are important to correct for when using online panel data [3]. Iterated variables are those we hypothesized may differ between surveys and drive selection differences. Each combination of the iterated variables with the full set of required variables was tested (total of 2048 schemes) with required 0.1 percentage point convergence. The enriched survey with transport weights then replaced the entire anchor sub-sample, where the weighted marginal distributions of variables matched between enriched and anchor surveys. This is the fused survey.

Table . Known and hypothesized variables to correct for selection bias.

Required variables (Known)	Iterated variables (Hypothesized)
Age[a]	Survey language
Gender[a]	Income
Census division[a]	Marital status
Limited in daily activities[a]	Self-assessment of general health[a]
Current cigarette use[a]	Trust in people[a]
Ethnicity[a]	Spirituality
Self-reported race[a]	Number of psychedelic drugs used in last year
	Number of illicit drugs used in last year
	Number of psychoactive prescription drug classes used on a regular basis last year
	Number of lifetime diagnoses of mental health illnesses
	Drug Abuse Screening Test (DAST-10) severity score

[a]Included in final transport weighting scheme to create the fused sample

In the generalizing step (1B to 1C in Figure 1), differences in selection between the fused survey and the national population were calibrated. Here, we used our previously validated method of calibrating to national benchmarks [3]. Briefly, three demographic and two health benchmarks were raked against estimates from federal benchmark surveys—the American

XSL•FO
RenderX

Community Survey and National Health Interview Survey. After the two calibration steps, final estimates from the fused survey were generalizable to the target population, adults in the US.

Statistical Analysis

Validity of the fusion was evaluated in two ways. Internal consistency was assessed by comparing an independent set of evaluation variables, i (Table S1 in Multimedia Appendix 1):

$$(1) Y_i = |(\hat{X}_{iA} - \hat{X}_{iF})$$

where, transport bias Y_i represents the absolute difference between the weighted population estimate from the anchor survey ($\hat{X}_{iA}$) to the weighted population estimate from the fused survey ($\hat{X}_{iF}$). Y_i would equal zero if no transport bias was introduced by the fusion. A parsimonious transport weighting scheme was chosen that minimized the mean Y_i across evaluation variables across three categories (demographic, health, and substance use). Internal consistency was measured for full population estimates and for the sub-population of adults reporting last 12-month psychedelic drug use. The selected weighting scheme was then applied to the independent second wave to test replicability. Weight stability was assessed by comparing convergence behavior, weight distributions, and the effective sample size (ESS) ratio. The ESS ratio is interpreted as how much of the sample was effectively retained after fusion, relative to a hypothetical simple random sample with equal variance.

To assess external validity, the population estimates from the independent second wave were compared to benchmark estimates from US federal surveys (Table S2 in Multimedia Appendix 1). One national survey was selected as the "gold standard," but multiple estimates are shown to demonstrate variability across probability surveys where available. The root-mean-square error (RMSE) of the benchmark estimates, k, was calculated as:

$$(2) RMSE(weighted) = \sum_{k=1}^{K} (\hat{X}_{kGS} - \hat{X}_{kF})^2 K$$

$$(3) RMSE(unweighted) = \sum_{k=1}^{K} (\hat{X}_{kGS} - \hat{X}_{kE})^2 K$$

where the estimates from the gold standard ($\hat{X}_{kGS}$) and the estimates from the weighted fused sample ($\hat{X}_{kF}$) and unweighted, not transported, enriched survey ($\hat{X}_{kE}$) were stratified by category.

Finally, the percentage and 95% CIs for reason of use among adults who used three psychedelic substances (ie, psilocybin, LSD, and MDMA) in the last 12 months were estimated by Taylor series expansion to demonstrate new information produced through the fusion method. All analyses were conducted using SAS (version 9.4; SAS Institute Inc.). This study is reported following the Journal Article Reporting Standards for Quantitative Research (JARS-Quant) guidelines [37].

Ethical Considerations

The Colorado Multiple Institutional Review Board first approved the anchor sample study on July 5, 2016 (#16‐0922) and the enriched sample study on February 02, 2024 (#23‐2426), allowing ongoing secondary analyses of these data without additional consent. All participants consented to being surveyed in the online questionnaires, the survey was confidentially administered, and study datasets were deidentified by the panel companies prior to transferring them to the researchers. No identification of individual participants in any parts of the manuscript or supplementary material is possible. In both panels, participants were compensated with points, which could be redeemed for gift cards of approximately US $5.

Results

Selecting a Transport Weighting Scheme

The first wave of the enriched survey invited 18,907 panelists (Figure S1 in Multimedia Appendix 1), of which 16,423 participated (16,423/18,907; 86.9% participation rate) and 15,430 completed it (15,430/16,423; 94.0% completion rate). The final enriched sample included 2306 respondents. The enriched survey was 40.4% female (932/2306) and had a median age of 33.1 (interquartile range [IQR]: 25.8‐39.7). The first wave of the anchor survey had a total sample of 28,679 respondents, where 2430 respondents reported using at least one psychedelic drug in the last 12 months and were weighted to represent 12.7 million adults (95% CI 11.9, 13.4). Table S3 in Multimedia Appendix 1 compares the enriched survey to the anchor survey weighted subpopulation using psychedelic drugs. Without adjustments, the enriched survey had lower percentages of male and White respondents, lower self-perceived health, and higher current cigarette use.

Fusion introduced minimal transport bias relative to the anchor survey, demonstrating good internal consistency (Figure 2). Across all weighting schemes, the absolute transport bias in demographic, health, and substance use metrics was less than 0.5 percentage points after fusion. The selected transport weighting scheme, shown by a diamond, was relatively parsimonious, including two variables (trust in people and self-assessment of general health) in addition to the required demographics. The average transport biases with the selected weighting scheme were: demographics, 0.09 percentage points; health characteristics, 0.35 percentage points; and substance use, 0.22 percentage points.

XSL•FO
RenderX

Figure 2. Average transport bias in evaluation estimates, internal consistency in fused dataset. Internal consistency was tested using data from the first half of 2024. Green dots represent average transport bias across all iterations of the transport weighting schemes. Purple diamond is the selected model, which was a parsimonious scheme with low bias across all three categories.

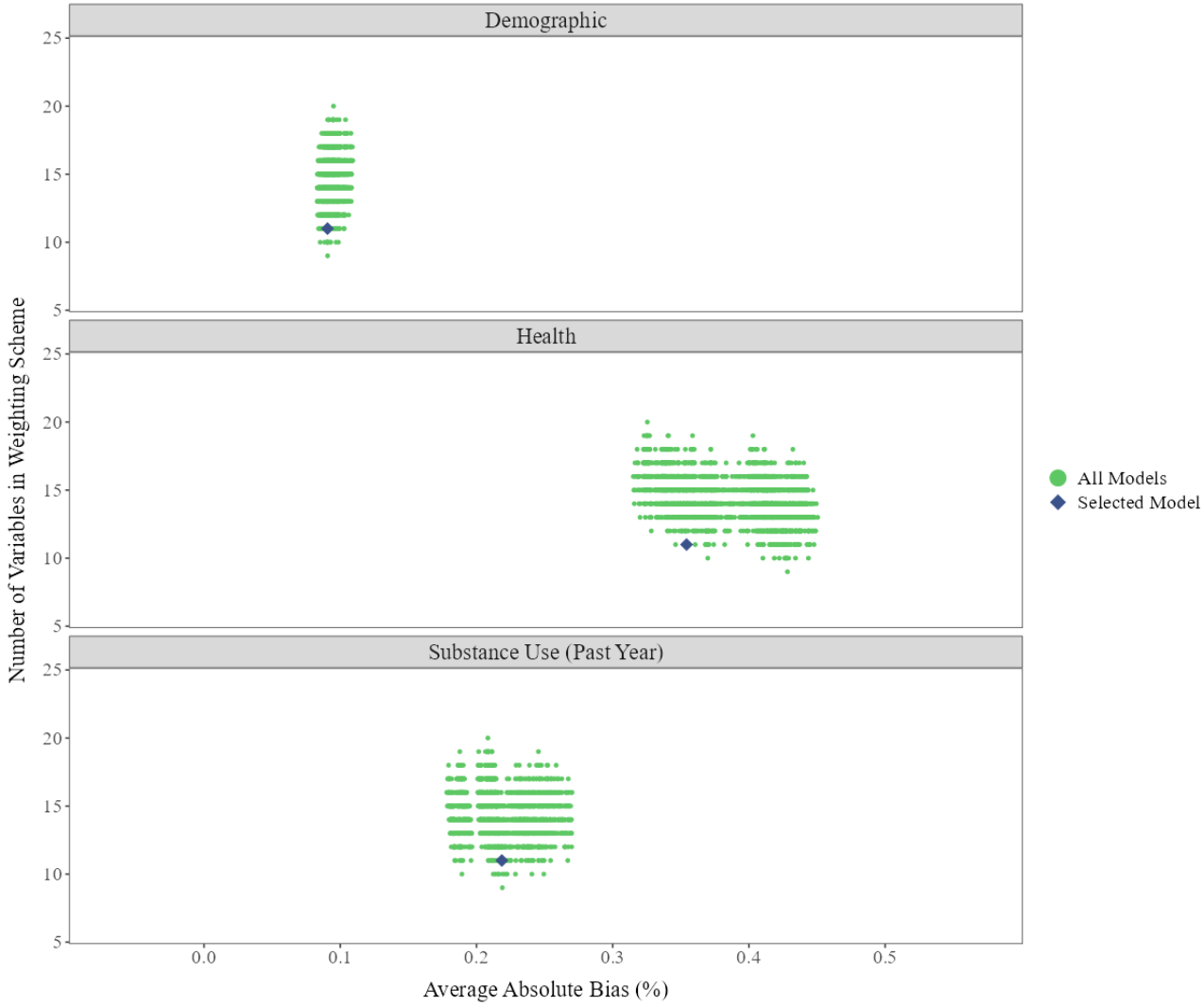

Replicability and Weight Stability

The selected transport weighting scheme was then applied to the second wave of enriched survey where 36,592 panelists were invited (Figure S2 in Multimedia Appendix 1), of which 27,868 participated (27,868/36,592; 76.2% participation rate) and 25,842 completed it (25,842/27,868; 92.7% completion rate). The final enriched survey included 2023 respondents. This survey was 39.6% female (801/2,023) and had a median age of 35.2 (IQR 26.8 - 44.4). The second wave of the anchor survey had a total sample of 29,040 respondents, where 2309 respondents reported using at least one psychedelic drug in the last 12-months and were weighted to represent 12.7 million adults (95% CI: 11.9, 13.5).

Fusion did not substantially perturb the demographics of the subpopulation using psychedelic drugs in either the first or second samples (Figures S3,and S4 in Multimedia Appendix 1). Geography, age, biologic sex, employment, and most health and substance use metrics were not perturbed by fusion. Health and substance use metrics had larger shifts in proportions after fusion compared to demographics, but shifts were not persistent when retested. Two metrics were consistently different across sample pairs. The fused sample had higher proportion of adults with high school or equivalent education and lower mental health treatment.

The weighting scheme converged in a similar number of iterations across waves (44 in Wave 1 vs 48 in Wave 2), indicating comparable optimization difficulty. Extreme weights were rare in both waves and did not increase in the second launch (three observations exceeded median+5×IQR in Wave 1 vs. two observations in Wave 2), suggesting little, if any, instability in the upper tail of the weight distribution. Transport weight distributions between waves had similar central tendencies and overall ranges (median 4888 vs 5,304; maximum 37,671 vs 34,793). Although the 99th percentile increased modestly in Wave 2 (14,384 vs 18,794), this change was not accompanied by notable increases in the maximum weight or the number of extreme weights and did not result in meaningful differences in ESS. ESS ratios were nearly identical across waves (0.78 in Wave 1 vs 0.75 in Wave 2), indicating stable transport weighting performance and no meaningful change in variance inflation in the second wave (Table 2).

XSL•FO
RenderX

Table . Weight stability assessment.

Weight characteristics	Wave 1, Early 2024	Wave 2, Late 2024
Mean (SD)	5498.31 (2914.28)	6275.43 (3616.25)
Minimum	422.06	451.88
Maximum	37,671.13	34,793.16
Median	4887.76	5304.47
Upper tail: 99%	14,383.51	18,794.18
Effective sample size ratio[a]	0.78	0.75

[a]Interpreted as how much of sample size effectively retained after weighting

External Validity

For external validity, fusion did not substantially change the demographic RMSE (1.82 unweighted to 1.88 weighted, a 3.3% increase); larger decreases were observed for health RMSE (7.30 to 3.38, 53.7% decrease) and for substance use RMSE (6.56 to 6.03, 8.1% decrease). Final population estimates were similar to benchmark estimates (Table 3, Figure 3). Geography, age, sex, and spirituality [38] estimates in the fused survey were nearly identical to benchmarks. Small differences (<5% generally) persisted in race estimates, marital status, education, and employment status after weighting, though race questions were asked differently between surveys. The data fusion process corrected health estimates by increasing self-perceived health estimates and decreasing distress, hospital stays, and mental health estimates. After fusion, most substance use estimates were nearly equal to benchmark values. Notably, alcohol use estimates were substantially lower than the benchmark and substantially increased the RMSE; this is possibly due to differences in how the questions were asked between surveys (without alcohol, the RMSE decreased from 6.03 to 1.55).

XSL•FO
RenderX

Figure 3. Impact of transport weighting on pre-defined benchmarks of demographics, health, and substance use estimates, external validity in fused survey. External validity was tested using data from the second half of 2024 and early 2025. US national gold standard benchmarks (triangles) and other survey (squares) estimates were substantially similar after transport weights (closed circle), and transport weighted values were corrected towards benchmark estimates in most health and substance use factors compared to unweighted estimates (open circle). Alcohol use questions were asked differently between the gold standard and this study. GAD-2, General Anxiety Disorder, 2 Item; PHQ-9, Patient Health Questionnaire, 9 Item.

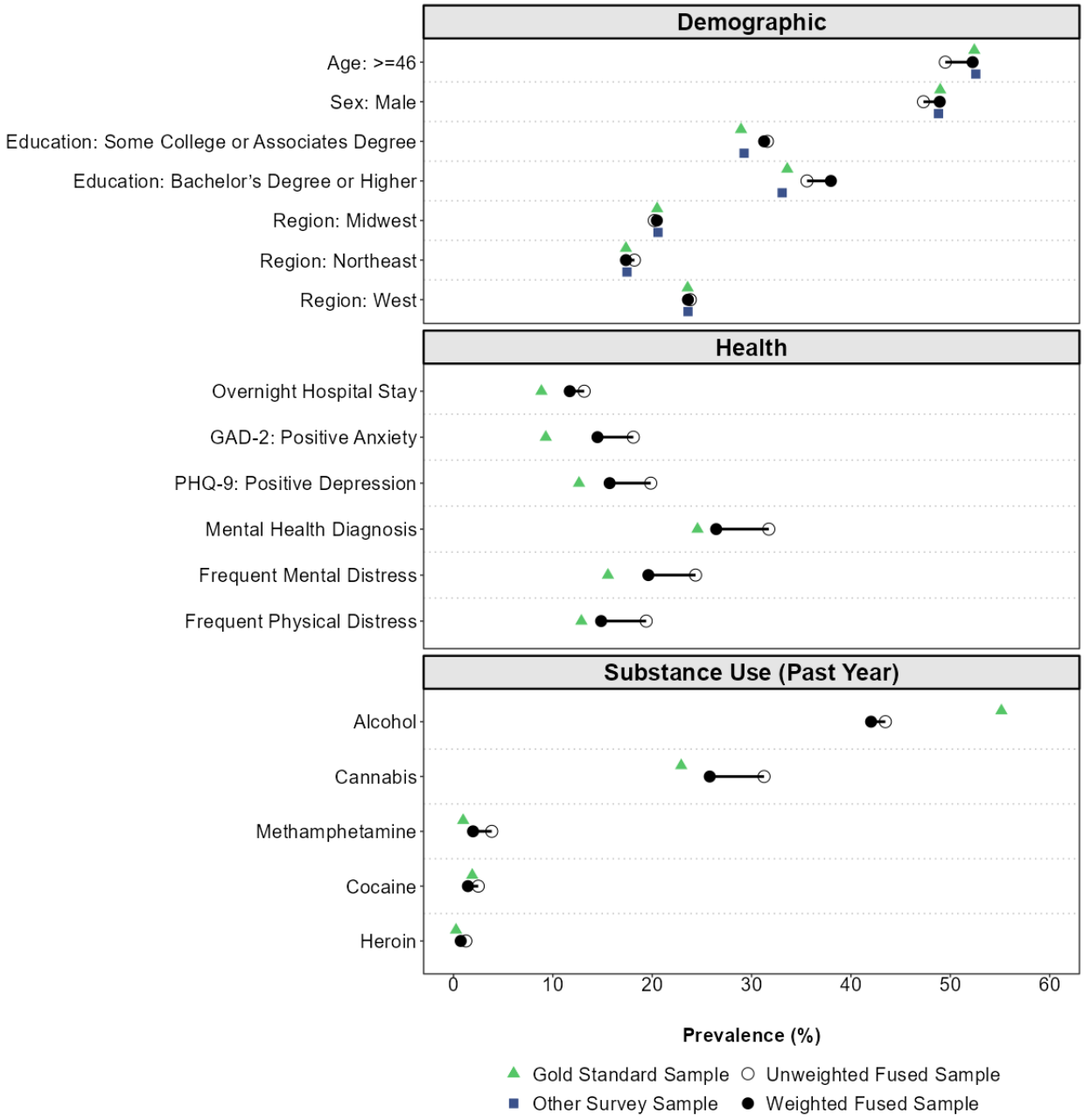

XSL•FO
RenderX

Table . Comparison of estimates in fused dataset, after transport weighting, to external survey benchmarks, second half of 2024 and early 2025.

Characteristic	Wave 2, Late 2024 Unweighted fused % (95% CI)	Wave 2, Late 2024 Weighted fused % (95% CI)	Additional benchmarks % (95% CI)	Survey
Number of adults				
Sample size, n	28,754	28,754		
Estimated number of adults, in millions	—	266 (264, 268)		
Age (years)				
≥46	49.5 (48.9, 50.1)	52.2 (51.5, 53.0)	52.4 (52.3, 52.5)	ACS[a]
Median (IQR)	44.8 (34.4, 58.2)	46.5 (32.4, 62.5)	46.6 (31.8, 62.5)	ACS
Sex				
Female	52.7 (52.1, 53.3)	51.1 (50.3, 51.8)	51.0 (50.9, 51.1)	ACS
Colorado or Oregon resident				
Yes	3.3 (3.1, 3.5)	3.1 (2.9, 3.4)	3.1 (3.1, 3.1)	ACS
Ethnicity and race				
Hispanic or Latino	16.7 (16.2, 17.1)	17.2 (16.6, 17.7)	17.6 (17.5, 17.6)	ACS
Black or African American alone	12.5 (12.1, 12.8)	12.4 (11.9, 12.9)	11.8 (11.8, 11.9)	ACS
White alone	73.4 (72.8, 73.9)	72.4 (71.7, 73.1)	62.9 (62.8, 63.0)	ACS
Other race alone	11.2 (10.8, 11.5)	12.3 (11.8, 12.8)	14.3 (14.2, 14.3)	ACS
Multiple races	3.0 (2.8, 3.2)	2.9 (2.6, 3.1)	11.0 (11.0, 11.1)	ACS
Marital status				
Never married	29.0 (28.5, 29.5)	30.7 (30.0, 31.4)	31.1 (31.0, 31.1)	ACS
Now married	47.7 (47.1, 48.3)	47.6 (46.9, 48.3)	50.5 (50.4, 50.6)	ACS
Widowed, divorced, separated	23.3 (22.8, 23.8)	21.7 (21.1, 22.2)	18.4 (18.3, 18.5)	ACS
Limited in the kind or amount of work				
Yes	28.7 (28.2, 29.2)	19.4 (18.9, 19.9)	19.5 (18.9, 20.1)	NHIS[b]
Current employment				
Yes	58.1 (57.6, 58.7)	55.5 (54.8, 56.3)	60.9 (60.1, 61.6)	NHIS
General health				
Good, Very Good, or Excellent	81.5 (81.0, 81.9)	84.8 (84.3, 85.3)	84.9 (84.3, 85.4)	NHIS
Poor or Fair	18.5 (18.1, 19.0)	15.2 (14.7, 15.7)	15.1 (14.6, 15.7)	NHIS
Trust in People: 0-You can't be too careful to 10-Most people can be trusted				
≥5 People Can Be Trusted	55.8 (55.3, 56.4)	56.3 (55.6, 57.0)		_[c]
Spirituality				
Not Spiritual or Don't Know	20.3 (19.8, 20.8)	21.3 (20.7, 21.9)	21.0	Pew American Trends Panel 2023
Current cigarette use				
Yes	37.2 (36.7, 37.8)	10.8 (10.6, 11.1)	10.8 (10.3, 11.3)	NHIS
Number of psychedelic substances used past year				
1	4.1 (3.9, 4.3)	2.7 (2.5, 2.9)		_[c]

XSL•FO
RenderX

Characteristic	Wave 2, Late 2024 Un-weighted fused % (95% CI)	Wave 2, Late 2024 Weighted fused % (95% CI)	Additional benchmarks % (95% CI)	Survey
2	1.3 (1.2, 1.5)	0.9 (0.8, 1.0)		_[c]
3 - 4	0.9 (0.8, 1.0)	0.6 (0.5, 0.7)		_[c]
≥5	0.2 (0.2, 0.3)	0.2 (0.1, 0.2)		_[c]
Number of illicit substances used past year				
1 - 3	7.6 (7.3, 7.9)	4.3 (4.1, 4.6)		_[c]
≥4	1.5 (1.3, 1.6)	0.9 (0.8, 1.0)		_[c]
Number of psychoactive prescription drugs used on a regular basis past year				
≥3	4.2 (3.9, 4.4)	2.8 (2.6, 3.0)		_[c]
Number of lifetime mental health diagnoses				
≥3	10.7 (10.4, 11.1)	7.7 (7.4, 8.1)		_[c]
Problematic substance use				
≥3 DAST-10[d] Severity Score	10.2 (9.9, 10.6)	6.5 (6.2, 6.8)		_[c]

[a]ACS: American Community Survey.

[b]NHIS: National Health Interview Survey.

[c]No available benchmark

[d]DAST: Drug Abuse Screening Test, 10 Item

Estimates presented in this table are not inclusive of all gold standard and other survey estimates calculated.

Fusion reduced the proportions of people using prescription, psychedelic, and illicit drugs and reduced the proportion with multiple mental health diagnoses. Using the transport weighting scheme, reasons for use differed by psychedelic drug. Recreational use of psilocybin (92.9%, CI: 91.1, 94.7), LSD (93.2%, CI 90.1, 96.4], and MDMA (93.3%, CI 91.0, 95.6) was more common than medical use (30.9%, CI 7.6, 34.2; 26.4%, CI 21.1, 31.7; and 21.1%, CI 17.5, 24.7, respectively). (Figure 4).

XSL•FO
RenderX

Figure 4. Reason for psychedelic substance use estimates using fused dataset, second half of 2024 and early 2025**.** Among adults in the United States who used each of the psychedelic substances below, the proportion who used for each reason differed with recreational use being most common while treating a medical symptom being least common. LSD,: lysergic acid diethylamide; MDMA: 3,4-methylenedioxymethamphetamine.

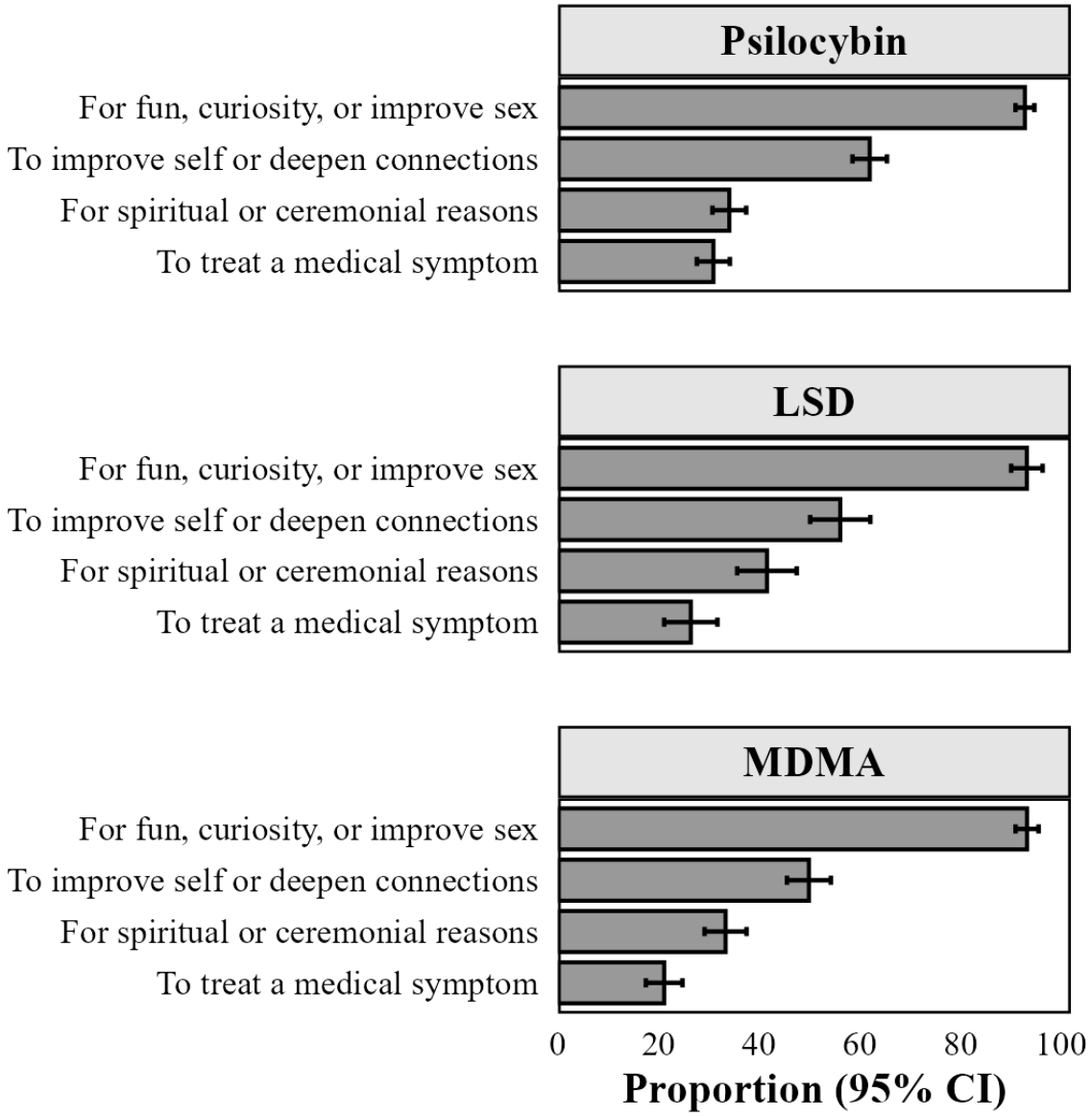

Discussion

Principal Findings

This study outlined a novel "fused survey design," applied a quantitative framework to test external validity, and demonstrated a practical application by estimating reasons for using specific psychedelic drugs among the general US adult population. This method of fusing surveys generated internally consistent estimates of demographics, drug use, and health, which were replicated across two independently fielded survey waves. The methodology also reduced differences in prevalence and demographic estimates relative to national benchmarks, thereby demonstrating the method can create estimates externally valid to a well-defined target population. Application of the validated methodology showed recreational reasons for use were more common than medical reasons across three psychedelic substances (ie, psilocybin, LSD, MDMA).

Specific Implications for Psychedelic Drug Research and Policy

In the United States, states have decriminalized and legalized medical access models, which has led to more people using [19]. Surveilling why people use psychedelic drugs would give policymakers evidence to enact regulations that supports safe, equitable, and effective access. These results suggest regulations that enable access should also account for substantial recreational use alongside medical use. The data fusion framework presented here produced generalizable estimates about a specific kind of behavior involving psychedelic drugs, which otherwise would not have been available in either large population surveys alone or in small, non-generalizable enriched

XSL•FO
RenderX

surveys. The growing prevalence of psychedelic use presents an emergent health phenomenon where detailed depth of measurement is required to understand the health implications of increased use [22], and the fused survey design can address future questions in this research area.

General Implications for Applying Data Fusion to Survey Samples

The framework presented in this work solves problems that arise when combining drug use using surveys for surveillance [30], and complements other survey-based applications of data fusion and transport weighting [39], which has historically focused on using surveys as targets for fusion with trial data, such as demonstrating the impact of a drug use interventions in more diverse populations [40,41]. The transport weighting scheme corrected for differences in demographic, health, and attitudinal perspectives of the enriched survey. These differences are correlated to medical outcomes of interest (eg, psychological symptoms, healthcare interactions), and if left uncorrected, would have introduced bias in the final estimates [42]. Nonprobability online panels are inherently not representative [25,30], and calibration of health and substance use was required to generate valid domain-specific estimates [25]. Data fusion would permit analysis between new populations not found in the enriched survey; for example in this study, characteristics between those who use psychedelics and those who use other medications without psychedelic use may be explored within the fused survey, but not within either survey alone. For psychedelic research, comparing to those who do not use is essential for developing effective policy and health recommendations, and drug-specific surveys which do not recruit those who do not use psychedelic drugs are unable to analyze that contrast [21].

This framework may be applied to a wide variety of research questions and may be particularly useful for rare subpopulations where direct sampling enrichment in a large survey is not feasible. Researchers should ensure that fusion samples are matched to similar subgroups of interest, that transport variables are measured as similarly as possible between enriched and anchor surveys, and that metrics are available to confirm the validity of the transport. Assumptions may not be directly testable, so demonstrating internal consistency and external validity, as has been done here, would strengthen results using this method. The methodology can also be used to fuse nonprobability surveys with probability-based surveys. Given the low response rate in many government surveys, in the US and elsewhere, nonprobability sampling may fill important gaps in surveillance in many disciplines, not just pharmacoepidemiology [30].

Limitations

The assumption that there was no measurement bias was not directly tested. Our research group controlled both the anchor and enriched surveys, so all transport variables had the same question wording, similar digital interfaces, and skip logic. This strengthened the validity of the measurement assumption for this pilot. If measurements were not equivalent, then calibration approach may not fully account for selection bias, because measurement and selection bias would be conflated. This may result in over- or under-correction, depending on the nature of measurement mismatch. Questions used to test validity of fusion should also match. In our results, there were differences in how past year alcohol use was measured. In our surveys, it was a single question asking whether a person had 12 drinks in the past 12 months. In the NHIS benchmark, two questions asked about lifetime use then about past 12-month use [43]. The NHIS question also provided a list of types of drinks, which may have improved recall. Data fusion and validity testing are influenced by measurement differences [26], and researchers who use this method should minimize such differences.

While the methodology overcomes many inferential limitations, the survey remains a rare population survey, where subgroups within the rare populations (eg, Indigenous peoples using psychedelics) may still be infrequently sampled. Weighting cannot eliminate stochastic uncertainty from low sample size [44]. However, fusion methods may allow multiple survey sets to be combined, thereby increasing subgroup sample size. Recent advances weighting individual variables may enable additional transport options for subgroups [45]. Additionally, there were small residual differences between the fusion estimates and benchmarks, potentially due to interactions between variables not accounted for in calibration. While other unmeasured selection differences may be present, this study examined a wide variety of variables that may bias selection, and we selected a final scheme that minimized differences against a national benchmark in a data-driven and parsimonious manner. Future work may incorporate multi-level calibration, which may further reduce bias by accounting for such higher-order interactions [46]. Measurement differences in how questions were asked or fielding date differences may contribute. Finally, this method requires the subpopulations between samples be aligned. In this pilot work, the studies had slightly mismatched drug groups, where approximately 15% of the enriched sample endorsed a drug not examined in the anchor sample. These respondents were kept in the fused dataset for this pilot, which may result in unadjusted transport bias in estimates involving data about these drugs.

Conclusion

Building upon past data fusion research, this study fused two surveys for the purpose of surveillance. This methodology, termed the “fused survey design,” is a rigorous but accessible approach for surveilling rare behaviors like drug use, and we demonstrated constructs absent from anchor surveys may be measured with generalizable inference. This expands the surveillance epidemiology toolbox, giving researchers an actionable process to field enriched surveys with specialized questions that would be impractical to add to larger surveys due to space constraints and respondent fatigue.

XSL•FO
RenderX

Funding

This methodology work was funded through contract #SAM31857 to investigator Andrew Monte from Substance Abuse and Mental Health Services Administration (SAMHSA) and the National Institute On Drug Abuse (NIDA) of the National Insitutes of Health (NIH) award R01DA064612 to Drs. Monte and Black. The funders had no involvement in the study design, analysis, interpretation, or the writing of the manuscript. The views expressed here are those of the authors and do not represent the position of SAMHSA, NIDA, NIH or the US Department of Health and Human Services.

Data Availability

Due to the sensitive and potentially illicit nature of the subject matter (drug use), data are not publicly available. Data requests may be made to the corresponding author.

Artificial intelligence was not used to develop, conceive, analyze, write, or review this manuscript.

Conflicts of Interest

This work was conducted by the staff at Rocky Mountain Poison and Drug Safety (RMPDS). RMPDS is a division of nonprofit Denver Health and Hospital Authority, a political subdivision of the State of Colorado, USA. Outside of this work, RMPDS provides independent surveillance, research, and reporting services to pharmaceutical manufacturers, government, and non-government agencies. No manufacturer or agency participated in the conception, analysis, drafting, or review of this manuscript.

Multimedia Appendix 1
Additional results.
[DOCX File, 402 KB - jmir_v28i1e86059_app1.docx]

References

1. Winstock A, Munksgaard R, Davies E, Ferris J, ZhuParris A, Barratt M. Global Drug Survey. 2022 May 19. URL: http://www.globaldrugsurvey.com/wp-content/uploads/2022/02/2022WowReport_Final.pdf
2. Key substance use and mental health indicators in the United States: Results from the 2023 National Survey on Drug Use and Health Rockville. In: Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration 2024. URL: https://www.samhsa.gov/data/sites/default/files/reports/rpt47095/National%20Report/National%20Report/2023-nsduh-annual-national.pdf [accessed 2024-05-19]
3. Black JC, Rockhill K, Forber A, et al. An online survey for pharmacoepidemiological investigation (Survey of Non-Medical Use of Prescription Drugs Program): validation study. J Med Internet Res 2019 Oct 25;21(10):e15830. [doi: 10.2196/15830] [Medline: 31654568]
4. Drug misuse in england and wales: year ending march 2024. ONS Centre for Crime and Justice. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/drugmisuseinenglandandwales/yearendingmarch2024 [accessed 2024-05-19]
5. European web survey on drugs. Lisbon, Portugal: European Union Drugs Agency. URL: https://www.euda.europa.eu/ewsd2024_en [accessed 2024-05-19]
6. Canadian Substance Use Survey (CSUS): summary of results for 2023.: Health Canada URL: https://www.canada.ca/en/health-canada/services/canadian-alcohol-drugs-survey/2023-summary.html [accessed 2024-05-19]
7. Chen JJ, Berg CJ, Yang YT. Hallucinogen use in the United States, 2021-2023: diverging trends and subgroup patterns. Drug Alcohol Depend Rep 2026 Mar;18:100410. [doi: 10.1016/j.dadr.2026.100410] [Medline: 41626531]
8. Cherro M, Itani H, Ghossoub E, Maalouf F. Predictors of suicide attempts among adolescents with suicidal ideations and a plan: results from the National Survey on Drug use and Health (NSDUH). PLoS One 2025;20(8):e0331261. [doi: 10.1371/journal.pone.0331261] [Medline: 40875615]
9. Han B, Jones CM, Volkow ND, et al. Prevalence of cannabis consumption methods among people with medically recommended and nonmedical cannabis use in the United States. Addiction 2025 May;120(5):962-974. [doi: 10.1111/add.16741] [Medline: 39838511]
10. Han B, Jones CM, Volkow ND, et al. Prescription stimulant use, misuse, and use disorder among US adults aged 18 to 64 years. JAMA Psychiatry 2025 Jun 1;82(6):572-581. [doi: 10.1001/jamapsychiatry.2025.0054] [Medline: 40105821]
11. Stroud LR, Gautam P, Jao NC, Sharma E. Menthol cigarette use in US pregnancies: prevalence, racial and ethnic disparities, and associated characteristics from the 2010-2019 National Survey on Drug Use and Health (NSDUH). Nicotine Tob Res 2025 Oct 22;27(11):2120-2128. [doi: 10.1093/ntr/ntaf008] [Medline: 39869405]
12. Lenton S, Potter G, Fortin D, et al. Growing practices and the use of potentially harmful chemical additives from a web survey of mainly small-scale cannabis growers in 18 countries. Int J Drug Policy 2025 Oct;144(Pt 3):104662. [doi: 10.1016/j.drugpo.2024.104662] [Medline: 39609120]

XSL•FO
RenderX

13. Marotta PL, Biaid M, Heimer R, et al. Rural providers' attitudes toward integrating harm reduction strategies and PrEP prescribing into rural primary care settings in the US. Southeast and Midwest. Addict Sci Clin Pract 2025 Sep 12;20(1):73. [doi: 10.1186/s13722-025-00584-9] [Medline: 40940646]
14. Music J, Sterling B, Charlebois S, Goedhart C. Comparison of perceptions in Canada and USA regarding cannabis and edibles. J Cannabis Res 2024 Jan 3;6(1):1. [doi: 10.1186/s42238-023-00213-9] [Medline: 38167591]
15. Örüm MH, Kapıcı Y, Sönmez D, et al. Practices and attitudes of adult psychiatrists regarding methamphetamine-associated psychotic disorder: an internet based survey conducted in Turkey. BMC Health Serv Res 2025 Jan 4;25(1):20. [doi: 10.1186/s12913-024-12134-1] [Medline: 39754117]
16. Wilkins C, Romeo J, Rychert M, Graydon-Guy T. Exploring the substitution of cannabis for alcohol and other drugs among a large convenience sample of people who use cannabis. Harm Reduct J 2024 Nov 5;21(1):192. [doi: 10.1186/s12954-024-01111-w] [Medline: 39501355]
17. Zhu DT, Kwon YIC, Lai A, Park AMG, Barnes AJ, Chapman DA. Global burden of disease due to opioid, amphetamine, cocaine, and cannabis use disorders, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021. PLoS One 2025;20(8):e0328276. [doi: 10.1371/journal.pone.0328276] [Medline: 40839682]
18. Bhave A. Magic of the mushrooms: effects of psilocybin decriminalization. J Psychoactive Drugs 2025;57(4):421-430. [doi: 10.1080/02791072.2024.2376755] [Medline: 38984875]
19. Black JC, Bau GE, Cook RR, Bemis EA, Monte AA. Psilocybin trends in states that decriminalized use. JAMA 2026 Mar 12. [doi: 10.1001/jama.2026.1952] [Medline: 41817506]
20. Siegel JS, Daily JE, Perry DA, Nicol GE. Psychedelic drug legislative reform and legalization in the US. JAMA Psychiatry 2023 Jan 1;80(1):77-83. [doi: 10.1001/jamapsychiatry.2022.4101] [Medline: 36477830]
21. Lake S, Lucas P. The Global Psychedelic Survey: consumer characteristics, patterns of use, and access in primarily anglophone regions around the world. Int J Drug Policy 2024 Aug;130:104507. [doi: 10.1016/j.drugpo.2024.104507] [Medline: 38936219]
22. Black JC, Monte AA, Dasgupta N, et al. Optimizing real-world benefit and risk of new psychedelic medications: the need for innovative postmarket surveillance. Nat Mental Health 2024;2(5):469-477. [doi: 10.1038/s44220-024-00233-1]
23. Kopra EI, Ferris JA, Winstock AR, Kuypers KP, Young AH, Rucker JJ. Investigation of self-treatment with lysergic acid diethylamide and psilocybin mushrooms: findings from the Global Drug Survey 2020. J Psychopharmacol 2023 Jul;37(7):733-748. [doi: 10.1177/02698811231158245] [Medline: 36876583]
24. Use of psychedelic substances in the United States. Results from the National Survey Investigating Hallucinogenic Trends (NSIHT).: Rocky Mountain Poison & Drug Safety, Denver Health and Hospital Authority; 2024. URL: https://www.rmpds.org/nsiht [accessed 2024-05-19]
25. Mercer AW, Kreuter F, Keeter S, Stuart EA. Theory and practice in nonprobability surveys. Public Opin Q 2017;81(S1):250-271. [doi: 10.1093/poq/nfw060]
26. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. Proc Natl Acad Sci U S A 2016 Jul 5;113(27):7345-7352. [doi: 10.1073/pnas.1510507113] [Medline: 27382148]
27. Josey KP, Yang F, Ghosh D, Raghavan S. A calibration approach to transportability and data-fusion with observational data. Stat Med 2022 Oct 15;41(23):4511-4531. [doi: 10.1002/sim.9523] [Medline: 35848098]
28. Yang S, Ding P. Combining multiple observational data sources to estimate causal effects. J Am Stat Assoc 2020;115(531):1540-1554. [doi: 10.1080/01621459.2019.1609973] [Medline: 33088006]
29. Aluja-Banet T, Daunis-I-Estadella J, Brunsó N, Mompart-Penina A. Improving prevalence estimation through data fusion: methods and validation. BMC Med Inform Decis Mak 2015 Jun 24;15(49):49. [doi: 10.1186/s12911-015-0169-z] [Medline: 26104747]
30. Si Y, Wagner JR, Kessler RC. Probability, probability-based, and nonprobability surveys in psychiatric epidemiological research. JAMA Psychiatry 2026 Mar 1;83(3):298-305. [doi: 10.1001/jamapsychiatry.2025.3652] [Medline: 41335451]
31. Vogt D, Borowski S, Maguen S, et al. Strengths and vulnerabilities: comparing post-9/11 U.S. veterans' and non-veterans' perceptions of health and broader well-being. SSM Popul Health 2022 Sep;19:101201. [doi: 10.1016/j.ssmph.2022.101201] [Medline: 36046065]
32. Lee S, Lee S, Lee J, Jo YT, Park E, Cha J. Fusion of multiple self-diagnostic questionnaires into optimal diagnostic cut-offs and factor analysis for depression characterization of the Korean university student group. BMC Psychiatry 2024 Nov 19;24(1):820. [doi: 10.1186/s12888-024-06295-w] [Medline: 39563273]
33. Pearl J, Bareinboim E. External Validity: From Do-Calculus to Transportability Across Populations Probabilistic and Causal Inference: The Works of Judea: Association for Computing Machinery; 2022:451-482. [doi: 10.1145/3501714.3501741]
34. Deville JC, Särndal CE, Sautory O. Generalized Raking Procedures in Survey Sampling. J Am Stat Assoc 1993 Sep;88(423):1013-1020. [doi: 10.1080/01621459.1993.10476369]
35. Rockhill K, Bemis EA, Schow N, Beekman K, Olsen HA, Fox EJ, et al. Establishment of External Validity of the National Survey Investigating Hallucinogenic Trends (NSIHT).: Center for Open Science; 2025 Jan 23. URL: https://osf.io/856dx
36. Ward MK, Meade AW. Dealing with Careless Responding in Survey Data: Prevention, Identification, and Recommended Best Practices. Annu Rev Psychol 2023 Jan 18;74(577-96):577-596. [doi: 10.1146/annurev-psych-040422-045007] [Medline: 35973734]

XSL•FO
RenderX

37. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
38. Spirituality Among Americans.: Pew Research Center; 2023 Jun 6. URL: https://www.pewresearch.org/religion/2023/12/07/spirituality-among-americans [accessed 2024-05-19]
39. Vuong Q, Metcalfe RK, Ling A, Ackerman B, Inoue K, Park JJ. Systematic review of applied transportability and generalizability analyses: A landscape analysis. Ann Epidemiol 2025 Apr;104(61-70):61-70. [doi: 10.1016/j.annepidem.2025.03.001] [Medline: 40064249]
40. Ackerman B, Lesko CR, Siddique J, Susukida R, Stuart EA. Generalizing randomized trial findings to a target population using complex survey population data. Stat Med 2021 Feb 28;40(5):1101-1120. [doi: 10.1002/sim.8822] [Medline: 33241607]
41. Inoue K, Hsu W. Transportability analysis-a tool for extending trial results to a representative target population. JAMA Netw Open 2024 Jan 2;7(1):e2346302. [doi: 10.1001/jamanetworkopen.2023.46302] [Medline: 38289608]
42. Black JC, Rockhill KM, Schow N, Monte AA. Proactive bias mitigation when using online survey panels for self-reported use of illicitly manufactured fentanyl in the general adult population. JAMA Health Forum 2025 Nov 7;6(11):e254011. [doi: 10.1001/jamahealthforum.2025.4011] [Medline: 41201814]
43. 2023 NHIS Questionnaires, Datasets, and Documentation.: National Center for Health Statistics; 2023. URL: https://www.cdc.gov/nchs/nhis/documentation/2023-nhis.html [accessed 2024-05-19]
44. Elvira V, Martino L, Robert CP. Rethinking the effective sample size. Int Statistical Rev 2022 Dec;90(3):525-550. [doi: 10.1111/insr.12500]
45. Tripet A, Tillé Y. Calibration and optimal transport approaches for harmonizing survey weights. Stat Methods Appl 2025 May;34(2):195-210. [doi: 10.1007/s10260-024-00776-8]
46. Ben-Michael E, Feller A, Hartman E. Multilevel calibration weighting for survey data. Polit Anal 2024 Jan;32(1):65-83. [doi: 10.1017/pan.2023.9]

Abbreviations

ACS: American Community Survey
DAST-10: Drug Abuse Screening Test, 10 Item
ESS: Effective sample size
GAD-2: General Anxiety Disorder, 2 Item
JARS-Quant: Journal Article Reporting Standards for Quantitative Research
LSD: lysergic acid diethylamide
MDMA: 3,4-methylenedioxymethamphetamine
NHIS: National Health Interview Survey
NMURx: Survey of Non-Medical Use of Prescription Drugs
NSIHT: National Survey Investigating Hallucinogenic Trends
PHQ-9: Patient Health Questionnaire, 9 Item
RMSE: root-mean-square error

Edited by S Brini; submitted 17.Oct.2025; peer-reviewed by A Patel, S Mirji; revised version received 06.Apr.2026; accepted 07.Apr.2026; published 27.Apr.2026.

Please cite as:
Rockhill KM, Bemis EA, Schow N, Olsen HA, Beekman K, Fox EJ, Monte AA, Black JC
Fusing Specialized Surveys of Rare Populations to Larger Surveys for Generalized Inference: Cross-Sectional Survey Study
J Med Internet Res 2026;28:e86059
URL: https://www.jmir.org/2026/1/e86059
doi:10.2196/86059

© Karilynn M Rockhill, Elizabeth A Bemis, Nicole Schow, Heather A Olsen, Kyle Beekman, Evelyn J Fox, Andrew A Monte, Joshua C Black. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

After-Hours Use of the Electronic Health Record Among Medical and Surgical Specialists After Implementation of a System-Wide Integrated Clinical Information System in Alberta, Canada: Longitudinal Descriptive Study

Robert P Pauly[1], MD, MSc; Tania Stafinski[2], PhD; Helia Koosha[3], MSc; Melita Avdagovska[2], PhD; Maliheh Hadizadeh[2], PhD; David Bigam[4], MD, MSc; Timothy Graham[5], MD, MSc; Craig Kuziemsky[6,7], PhD; Narmin Kassam[1], MD, MHPE; Devidas Menon[2], MHSA, PhD

[1]Department of Medicine, University of Alberta, 11-107 Clinical Sciences Building, 8440 112 Street NW, Edmonton, AB, Canada
[2]School of Public Health, University of Alberta, Edmonton, AB, Canada
[3]Faculty of Engineering, University of Alberta, Edmonton, AB, Canada
[4]Department of Surgery, University of Alberta, Edmonton, AB, Canada
[5]Department of Emergency Medicine, University of Alberta, Edmonton, AB, Canada
[6]Office of Research Services, MacEwan University, Edmonton, AB, Canada
[7]School of Business, MacEwan University, Edmonton, AB, Canada

Corresponding Author:
Robert P Pauly, MD, MSc
Department of Medicine, University of Alberta, 11-107 Clinical Sciences Building, 8440 112 Street NW, Edmonton, AB, Canada

Abstract

Background: Studies suggest that the introduction of electronic health records (EHRs) has decreased the efficiency of clinical practice and increased clinician workload for US-based physicians. Most studies involve clinicians in primary care settings. Less is known about other health care settings, subspecialist clinicians, or whether markers of efficiency and workload change over time.

Objective: This study aimed to describe 2 common metrics of after-hours use of the EHR (pajama time and time outside scheduled hours [TOSH]) among diverse specialists and track these parameters longitudinally in a Canadian setting.

Methods: In this longitudinal descriptive study, medical and surgical specialists were observed starting from the introduction of a system-wide EHR in 2019 to 2022 at a large quaternary teaching hospital in Edmonton, Alberta. Pajama time and TOSH were extracted from the EHR on an Epic system platform and monitored over time. Clinicians were stratified according to clinical group (medical and surgical) and workload (clinical full-time equivalent).

Results: A total of 71 medical and surgical specialists participated in this study, spending approximately 24 to 40 minutes per day on pajama time and 32 to 55 minutes per day on TOSH depending on clinician grouping. Both metrics increased over the observation period, as reflected in the longitudinal plots and the higher values observed at the end vs the beginning of follow-up.

Conclusions: After-hours EHR use in this Canadian cohort of medical and surgical specialists was similar to what is reported in the US literature, although the drivers may be different. Perhaps surprisingly, these markers increased over time despite presumed improved familiarity with the EHR. The extent to which this affects clinician well-being and work-life integration cannot be determined from these results, although there may be cause for concern.

(J Med Internet Res 2026;28:e76872) doi:10.2196/76872

KEYWORDS

electronic health record; EHR; after hours; pajama time; time outside scheduled hours; TOSH; efficiency; workload; specialists; clinicians; longitudinal; Epic; Canada; Edmonton; quaternary; hospital; medical; surgical; informatics; burnout; well-being; workflow

Introduction

The purported benefits of implementing electronic health records (EHRs) in the past decade were to decrease costs, more effectively manage health information, increase guideline adherence, and reduce medical errors [1,2]. However, delivering on these promises has come at the expense of other aspects of care, such as increased clinician workload and decreased

XSL•FO
RenderX

efficiency [1,3,4]. The impact of EHRs on clinician well-being can be viewed through the lens of the job demands–resources (JD-R) framework, where well-being is the integration of job demands and available resources [5]. EHRs are an important resource and simultaneously represent a cognitive burden with inbox management, order entry, and documentation; a temporal burden with EHR work encroaching on traditional nonwork time; and a boundary-blurring strain on work-life balance with few constraints on system accessibility and system engagement expectation. Studies suggest that clinicians spend considerable time on the EHR, particularly after typical work hours, such as on evenings and weekends [6,7]. This after-hours engagement with the system is also called work outside work [8] and has even led to the coining of a new term, *pajama time*, to describe the time at home in the evening when clinicians re-engage with the EHR to complete tasks that would otherwise have been done at the office during regular working hours. After-hours work such as this may impact clinician well-being, an area that is receiving more attention as the physician workforce is facing increasing pressures from within health systems and the public, with significant rates of burnout reported in Canada, the United States, and elsewhere [9-12]. In fact, there is an increasing body of evidence that links the excessive burden of EHR use directly with burnout and other measures of clinician dissatisfaction [13-16]. The specific aspects of the EHR that drive this dissatisfaction are less clear, although after-hours engagement with the electronic system is likely a key contributor [17-20]. An important limitation of the existing literature investigating EHR use is that most studies have been conducted in the United States, where documentation is influenced in large part by insurance claim requirements, billing, quality reporting, and litigation risks. These drivers of documentation (and, hence, use of the EHR) vary from country to country and could easily result in systematic differences in clinician documentation behavior in different health care systems. This raises the question of whether the widely reported results of EHR use in general and use outside of regular working hours in particular are generalizable to other health systems.

In 2019, Alberta Health Services began system-wide implementation of a comprehensive EHR in Alberta, Canada. This provided a unique opportunity to track clinician EHR use longitudinally in a Canadian setting. This study aimed to describe 2 common metrics of after-hours use, pajama time and work time outside scheduled hours (TOSH), and track trends over time among medical and surgical specialists in a large quaternary care setting after the implementation of the EHR. Ours is a descriptive study summarizing differences and temporal patterns in aggregated monthly measures rather than aiming to test prespecified causal hypotheses.

Methods

Study Design, Setting, and Participants

This was a longitudinal descriptive study conducted at the University of Alberta Hospital in Edmonton, Alberta. We collected physician EHR use data between November 1, 2019 (system rollout), and July 30, 2022. A more detailed study protocol was previously published [21]. A purposive sampling method was used to recruit clinicians from the Department of Medicine and the Department of Surgery inclusive of all medical and surgical subspecialists. Specifically, the study aims and methodology were advertised and promoted during departmental rounds, and participants were tracked using REDCap (Research Electronic Data Capture; Vanderbilt University). The study population consisted of clinicians who had an ambulatory care practice and at least 7 months of experience with the EHR. Clinicians were asked to identify their department or division, EHR username (to link individuals to their EHR data), gender, and self-reported clinical full-time equivalent (FTE) percentage. The latter was used as a marker of overall clinical workload given that many clinicians also have substantial research, teaching, and administrative portfolios. No demographic or other personal information was obtained other than EHR use data (see below). Participants were grouped into medical specialists with a clinical FTE of 0.5 or less, medical specialists with a clinical FTE of more than 0.5, and surgical specialists with a clinical FTE of more than 0.5 (no surgeon reported a clinical FTE≤0.5).

Ethical Considerations

The study protocol was approved by the University of Alberta Hospital Research Ethics Board (case number Pro00119194), and individual participants provided written consent allowing their anonymized EHR use data to be transferred from Alberta Health Services (the legal custodian of the EHR) to the study team for analysis. Participants were not compensated for taking part.

Data Source and Metrics

Connect Care is the provincial EHR customized on an Epic platform (Epic Systems). Data were obtained from Signal, an analytics tool developed by Epic Systems using EHR user action log data and integrated into the EHR platform. Data related to outpatient visits were extracted in aggregate 1–calendar month reporting periods.

The two key metrics used to describe clinician engagement with the EHR outside typical ambulatory clinical working hours were (1) pajama time, defined as the average number of minutes that a clinician spent in charting activities on weekdays outside the hours of 7 AM to 5:30 PM or outside scheduled hours on weekends or nonscheduled holidays; and (2) TOSH, defined as the average number of minutes that a provider spent on the system outside of scheduled hours, where scheduled hours represented scheduled clinic appointments with a 30-minute time buffer before and after those scheduled clinic appointments. The calculation of these metrics is outlined in Table S1 in Multimedia Appendix 1. TOSH was captured for the duration of the study, whereas pajama time was captured from October 2020 to study termination in July 2022 as this metric was not available within the system for the first year after rollout (see Figure S1 in Multimedia Appendix 1 for a graphic representation of the observation periods).

System-imposed restrictions (ie, nonmodifiable Signal defaults) limited availability to only those data for which participants registered seeing at least 1 ambulatory care patient during the reporting period. There was an additional system-imposed

XSL•FO
RenderX

restriction for one specific metric contributing to the total time that a participant spent engaged with the EHR: time in the in-basket per appointment. A participant needed to have 5 or more appointments per week within the reporting period to trigger the collection of this variable.

Statistical Analysis

Data were summarized using SAS (version 9.4; SAS Institute) and Tableau (version 2021.4.3; Tableau Software LLC) as month-level group-weighted means for each metric to describe typical monthly group levels (ie, aggregated summaries rather than clinician-level longitudinal effects). For example, the monthly weighted mean of pajama time was calculated by summing the pajama time minutes of all participating clinicians in a given month and dividing it by the sum of the number of days that all clinicians spent on the system during that month. Because consecutive months are serially related in a longitudinal time series, we did not treat month-level summaries as independent observations for primary inference. Accordingly, the results are presented descriptively. Longitudinal figures display a 4-month simple moving average for visualization only. As a sensitivity analysis, we compared the distribution of month-level group summaries between groups using the Mann-Whitney *U* test; these results are reported as supportive only and should be interpreted cautiously given the independence assumption. Missing data were identified for pajama time for April 2021, July 2021, and September 2021 and for TOSH for April 2021; these months were excluded using listwise deletion.

Results

A total of 71 participants from 19 medical and surgical specialties were enrolled in this study, with 27 (38.0%) being medical specialists with clinical FTE of 0.5 or less, 26 (36.6%) being medical specialists with clinical FTE of more than 0.5, and 18 (25.4%) being surgical specialists with clinical FTE of more than 0.5 (no surgeon reported a clinical FTE≤0.5). In total, 39.4% (28/71) of all specialists were women (details on specialist participants can be found in Table S2 in Multimedia Appendix 1).

The average number of appointments per day, pajama time, and TOSH is summarized in Table 1. Surgeons had a higher number of appointments per scheduled ambulatory day than medical specialists, and after-hours EHR time ranged from approximately 24 to nearly 40 minutes per scheduled day for pajama time and 32 to 55 minutes per scheduled day for TOSH depending on clinician group. Among medical specialists, clinicians with clinical FTE of 0.5 or less had lower after-hours EHR time than clinicians with clinical FTE of more than 0.5 for both pajama time and TOSH. Over the observation period, appointments per day, pajama time, and TOSH showed an overall upward pattern in all clinician groups (Table 2 and Figure 1). To provide an interpretable descriptive summary of change over time, Table 3 compares the average of the first 3 months vs the last 3 months of available data for pajama time and TOSH. For example, surgeons averaged 30.1 minutes of pajama time per scheduled day in the first 3 months and 47.3 minutes per scheduled day in the last 3 months of observation.

Table . Weighted monthly means of the reported metrics for medical and surgical groups.

Metric	All specialties (95% CI[a])	Medical specialties (95% CI)			All surgical specialties[b] (95% CI)
		All medical specialties	Medical specialties with FTE[c] ≤0.5	Medical specialties with FTE >0.5	
Number of appointments per day	8.94 (8.20-9.67)	6.42 (6.31-6.52)	6.47 (6.33-6.61)	6.36 (6.20-6.52)	13.98 (13.50-14.47)
Pajama time (minutes)	32.10 (29.82-34.39)	28.21 (26.17-30.25)	24.05 (21.77-26.34)	32.37 (30.20-34.53)	39.89 (36.31-43.47)
Time outside scheduled hours (minutes)	42.27 (37.88-42.67)	43.44 (40.27-46.61)	32.33 (30.26-34.41)	54.54 (52.28-56.80)	33.94 (31.69-36.19)

[a]95% CIs were calculated from the distribution of the month-level weighted means across the observation period for each group. These CIs reflect between-month variability of aggregated summaries.

[b]All surgeons reported a clinical full-time equivalent of more than 0.5.

[c]FTE: full-time equivalent.

XSL•FO
RenderX

Table . Trend analysis of the reported metrics over time for the medical and surgical groups.

Metric	Medical specialties with FTE[a] ≤0.5		Medical specialties with FTE >0.5		Surgical specialties	
	Slope	Trend[b]	Slope	Trend	Slope	Trend
Appointments per day	0.0008	Increasing	0.0014	Increasing	0.0015	Increasing
Pajama time[c]	0.0118	Increasing	0.0161	Increasing	0.0275	Increasing
Time outside scheduled hours[d]	0.0140	Increasing	0.0137	Increasing	0.0098	Increasing

[a]FTE: full-time equivalent.

[b]Direction is based on the sign of the slope from a linear fit to the month-level series displayed for visualization (Figure 1). These trend summaries are descriptive and are not intended as confirmatory inference.

[c]Trend spanning October 2020 to July 2022.

[d]Trend spanning November 2019 to July 2022.

XSL•FO
RenderX

Figure 1. Longitudinal trends in outcome measures for (A) appointments per day, (B) pajama time, and (C) time outside scheduled hours. The green line corresponds to the medical group with clinical full-time equivalent (FTE) of 0.5 or less, the red line corresponds to the medical group with clinical FTE of more than 0.5, and the purple line corresponds to the surgical group. Graphs are shown as simple moving averages (SMAs).

XSL•FO
RenderX

Table . Changes in pajama time and time outside scheduled hours (TOSH) over the observation period.

Specialist group	Pajama time (minutes per scheduled day), mean		TOSH (minutes per scheduled day), mean	
	First 3 months[a]	Last 3 months[b]	First 3 months[c]	Last 3 months[b]
Medical specialties with FTE[d] ≤0.5	21.8	22.6	26.5	34.8
Medical specialties with FTE >0.5	26.9	34.8	43.1	56.4
Surgical specialties	30.1	47.3	31.2	43.6

[a]October 2020 to December 2020 inclusive.

[b]May 2022 to July 2022 inclusive.

[c]November 2019 to January 2020 inclusive.

[d]FTE: full-time equivalent.

Discussion

To our knowledge, this is the only study to track physician after-hours EHR use in ambulatory care across multiple medical and surgical specialties from the time of EHR implementation for almost 3 years in a Canadian setting. Adoption of EHRs in Canada has accelerated in the past decade, but there is little to no quality publicly available data to benchmark clinician use; this study fills that void. Clinician workload, as measured using the number of appointments per clinic day, increased over time, accompanied by concurrent trends in increased after-hours engagement with the EHR regardless of clinician group or job description. The average amount of after-hours time that clinicians engaged with the EHR equated to approximately 11 to 19 hours per month for pajama time and approximately 15 to 25 hours per month for TOSH depending on specialty group.

This represents significant infringement on time that might reasonably be considered outside routine clinical service. Sinsky et al [6] reported after-hours EHR use at 53 minutes per day measured prospectively through clinician self-reporting for a diverse group of primary care and internal medicine specialists, although approximately a quarter of the evenings and nights under observation were skewed by clinicians also being on call (when one might anticipate accessing the EHR for patient care as part of expected duties). This compares to 1.4 hours per day per 1.0 clinical FTE reported by Arndt et al [7], 58 minutes per day per 1.0 clinical FTE reported by Attipoe et al [22], or 48 minutes per workday reported in the latter's subsequent study [23]. These US-based studies are limited by their cross-sectional design. Our study found that both pajama time and TOSH increased over nearly 2 and 3 years since EHR implementation, respectively, which may be interpreted as a worrisome sign. Specifically, familiarity with the EHR presumably increased over time but coincided with increased after-hours engagement with the system rather than a decrease, as might have been reasonably expected.

However, the observation that after-hours clinician work in this study was within the general range observed in US studies indicates that Canadian clinicians are similar to their US colleagues, at least as far as their after-hours EHR use is concerned. This suggests that country differences in EHR engagement do not have as much of a net effect on clinician time as might be predicted (ie, specific documentation requirements in the United States appear to be balanced by other EHR activities in Canada, although this is speculative). Indeed, a recent multinational comparison of EHR use with the Epic Systems global ambulatory care customer base suggests that country-specific contextual differences play a significant role in EHR use [24].

There is no agreed-upon gold standard metric for after-hours EHR use [8]. Pajama time is tied to clock-based time and may be a reasonable marker of after-hours system engagement for clinicians with predictable schedules involving daytime ambulatory practice. This metric is not well suited for clinicians with varied inpatient and outpatient practices or shift work. TOSH, on the other hand, is tied to clinicians' ambulatory care appointment schedules. At first glance, this may appear to solve the disadvantages of pajama time, but it is not well suited for clinicians with fragmented ambulatory clinic schedules or those whose clinical duties predominantly involve inpatient care [25]. Thus, comparisons of after-hours work between studies should be interpreted with caution. First, there is variation in how after-hours work is defined and assessed. Some studies cite clinician self-reported EHR time [6]. Others use externally unvalidated proprietary vendor predefined metrics (as is the case in this study) that variably use clock time–based metrics vs schedule-based metrics [24,26]; a combination of mouse clicks, keystrokes, and mouse miles (pixels) traveled to assess EHR activity [27]; or event logging data [22,23], which is believed to capture activity with the highest degree of precision but is expensive to perform and requires significant analytic power and expertise. Each of these approaches has its limitations [8]. Second, for studies capturing vendor-defined use metrics, there is no consistency in defining these metrics across platforms, and comparisons of EHR use from studies of health systems using different vendors remain unvalidated [25]. This is particularly relevant in this study: while pajama time and TOSH are well-recognized concepts, in our study, they are Epic-generated defaults, and the precise methods to calculate these may differ from those used by other vendors. Third, much of the literature on after-hours EHR use focuses on primary adult and pediatric care, with much less literature centered on specialist care. There is some evidence that surgeons, for example, have different use patterns from those of either primary care providers or medical specialists [26], a finding that this

XSL•FO
RenderX

study corroborates. Fourth, there is no standardized method to account for physicians' clinical workload and the bias that this may introduce into after-hours EHR use metrics. Some studies have normalized reporting to clinical FTE [7,19], whereas in most cases, normalization remains unreported or ambiguous [6,22-24,26,27]. Others have adjusted for relative value units, a metric used in the US health care system as a marker of volume and type of clinical service delivered [22]. Even comparing this study to the existing literature is imprecise and fraught due to different reporting methodologies. Finally, for those studies using event or access log metadata, there is no agreed-upon classification of EHR events that cumulatively defines activities that clinicians perform within the system.

Notwithstanding these sources of inherent bias encompassing the entire body of literature, this study has a number of important strengths. First, this is one of only a few reports from a non-US health system; therefore, it provides an opportunity to benchmark after-hours EHR use in a contemporary Canadian setting. Because this study includes a wide range of medical and surgical specialists, it also adds to only a small number of other studies investigating EHR use outside of primary care. Finally, this was a longitudinal study tracking changes in after-hours EHR use from EHR implementation for nearly 3 years, so we were able to report on changes over time. To our knowledge, all other studies save 1 [28] are cross-sectional. Interestingly, this sole other longitudinal study involved primary care physicians in the United States and also demonstrated increased after-hours EHR workload between 2019 and 2023.

Nonetheless, this study has several limitations. First, it is descriptive in nature and, thus, is agnostic to the impact of clinicians' after-hours work on well-being, burnout, or work-life integration. Linking EHR use metrics to measures of professional satisfaction is a future research priority. Second, this study is likely to underestimate true after-hours EHR work because of underreporting of time due to EHR system-imposed restrictions on data capture. For example, a participant needed to have 5 or more appointments per week to trigger the collection of in-basket metrics (which, presumably, is expected to make up a meaningful proportion of the after-hours time that clinicians spend on the EHR during pajama time or TOSH)—thus raising concerns that pajama time and TOSH may be significantly underestimated for those clinicians with low ambulatory clinical volume, such as surgeons with irregular schedules or internists with high inpatient care service but little outpatient demands. Third, this study did not include any information on clinician schedules or patient complexity so that even comparisons between specialists are limited by lack of normalization to clinical activity. Fourth, the rollout of the EHR in Alberta (and, hence, this study) coincided with the COVID-19 pandemic, when routine ambulatory practice temporarily and intermittently transitioned to virtual care, where use of an EHR might well have shifted practice patterns, including after-hours EHR use. Fifth, data for pajama time were not available in the EHR for the first 12 months of observation, so this metric could not be reported from the time of system rollout in the same way as such baseline data were only available for TOSH. This may limit the interpretation of trends over time (Table 2), and it is not known whether this dampens or exaggerates the effect that we report. Sixth, because participants were recruited using purposive sampling, there is inherent self-selection bias in terms of clinicians opting to participate, potentially affecting generalizability to physicians at large. Generalizability may also be limited as study participants were derived from a single quaternary hospital setting among clinicians with an ambulatory practice, which may not represent after-hours use of the EHR among clinicians without an ambulatory practice or in community hospital settings. Finally, while this study is larger than most in this research space, a larger participant pool would have enabled more robust analysis of relevant strata of clinician ages, genders, and areas of specialization.

Our study confirms (notwithstanding the limitations mentioned above) that the metrics of after-hours use are similar in range to those reported in other studies. When considering the JD-R framework, it may be helpful to consider as well the proposed matrix in Figure 2 as an approach to mitigating the intrusiveness of the EHR and optimizing system efficiency. The matrix itemizes the core activities for which clinicians use the EHR (ie, documentation, note management, chart review, navigation, and order entry) and considers the role played by industry, regulators, hospital systems, clinical groups, and individuals. System optimization from clinicians' perspectives might include a range of interventions, such as improving the IT itself (eg, the structure of the user interface and system navigation, frequency and intrusiveness of alerts, and curation of information to limit information overload or chaos), streamlining regulatory requirements for using scribes or integrating artificial intelligence tools, adopting rules of engagement between system users (eg, when and how to communicate between members of interdisciplinary teams and roles and responsibilities pertaining to laboratory result reviews among team members), having hospital systems reappropriate clerical tasks to clerical support personnel to free up clinician time, and emphasizing personalization of the system for individual users (eg, templating, voice recognition software, and shortcuts). Interventions at all levels may be expected to rebalance after-hours time spent on the EHR. As per the JD-R model, increasing demands are not by themselves necessarily harmful as long as they are sufficiently supported by resources to accommodate them.

Ultimately, ours is an important longitudinal study on after-hours EHR use in the Canadian context with potential implications for clinician well-being and system usability.

XSL•FO
RenderX

Figure 2. A proposed intervention matrix for considering improvements in the usability and intrusiveness of electronic health records.

	Industry	Regulators	Hospital systems	Clinical groups	Individuals
Documentation					
Inbox management					
Chart review					
Navigation					
Order entry					

Funding

This study was supported by a Catalyst Grant from the Canadian Institutes of Health Research (grant CGQAE 180993).

Data Availability

Data are stored at the Health Technology and Policy Unit in the School of Public Health at the University of Alberta. Data are not publicly available as the investigators are not the custodians of data from the provincial EHR and additional permission would need to be sought from the health authority prior to releasing or transferring such data; all data queries may be directed to the Health Technology and Policy Unit by contacting the corresponding author.

Authors' Contributions

RPP contributed to study conceptualization, methodology, investigation, writing the original draft, supervision, and funding acquisition. TS contributed to conceptualization, methodology, investigation, writing the original draft, project administration, and funding acquisition. HK contributed to the formal analysis, data curation, and writing the original draft. MA contributed to study conceptualization, methodology, data curation, and funding acquisition. MH contributed to investigation and manuscript review and editing. DB contributed to investigation, manuscript review and editing, and funding acquisition. TG contributed to study conceptualization, methodology, manuscript review and editing, and funding acquisition. CK contributed to study conceptualization, methodology, manuscript review and editing, and funding acquisition. NK contributed to study investigation and funding acquisition. DM contributed to study conceptualization, methodology, manuscript review and editing, supervision, project administration, and funding acquisition.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Metrics included in the data analysis and their definitions, breakdown of participants by subspecialty, and timeline of observation periods.
[DOCX File, 219 KB - jmir_v28i1e76872_app1.docx]

References

1. Tsai CH, Eghdam A, Davoody N, Wright G, Flowerday S, Koch S. Effects of electronic health record implementation and barriers to adoption and use: a scoping review and qualitative analysis of the content. Life (Basel) 2020 Dec 4;10(12):12. [doi: 10.3390/life10120327] [Medline: 33291615]
2. Uslu A, Stausberg J. Value of the electronic medical record for hospital care: update from the literature. J Med Internet Res 2021 Dec 23;23(12):e26323. [doi: 10.2196/26323] [Medline: 34941544]
3. Baumann LA, Baker J, Elshaug AG. The impact of electronic health record systems on clinical documentation times: a systematic review. Health Policy 2018 Aug;122(8):827-836. [doi: 10.1016/j.healthpol.2018.05.014] [Medline: 29895467]
4. Boonstra A, Jonker TL, van Offenbeek MAG, Vos JFJ. Persisting workarounds in electronic health record system use: types, risks and benefits. BMC Med Inform Decis Mak 2021 Jun 8;21(1):183. [doi: 10.1186/s12911-021-01548-0] [Medline: 34103041]
5. Demerouti E, Bakker AB, Nachreiner F, Schaufeli WB. The job demands-resources model of burnout. J Appl Psychol 2001 Jun;86(3):499-512. [Medline: 11419809]

XSL•FO
RenderX

6. Sinsky C, Colligan L, Li L, et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. Ann Intern Med 2016 Dec 6;165(11):753-760. [doi: 10.7326/M16-0961] [Medline: 27595430]
7. Arndt BG, Beasley JW, Watkinson MD, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Ann Fam Med 2017 Sep;15(5):419-426. [doi: 10.1370/afm.2121] [Medline: 28893811]
8. Sinsky CA, Rule A, Cohen G, et al. Metrics for assessing physician activity using electronic health record log data. J Am Med Inform Assoc 2020 Apr 1;27(4):639-643. [doi: 10.1093/jamia/ocz223] [Medline: 32027360]
9. Shanafelt TD, West CP, Sinsky C, et al. Changes in burnout and satisfaction with work-life integration in physicians and the general US working population between 2011 and 2020. Mayo Clin Proc 2022 Mar;97(3):491-506. [doi: 10.1016/j.mayocp.2021.11.021] [Medline: 35246286]
10. CMA 2021 National Physician Health Survey. : Canadian Medical Association; 2022 URL: https://digitallibrary.cma.ca/link/digitallibrary17 [accessed 2026-03-16]
11. McKinley N, McCain RS, Convie L, et al. Resilience, burnout and coping mechanisms in UK doctors: a cross-sectional study. BMJ Open 2020 Jan 27;10(1):e031765. [doi: 10.1136/bmjopen-2019-031765] [Medline: 31988223]
12. Rotenstein LS, Torre M, Ramos MA, et al. Prevalence of burnout among physicians: a systematic review. JAMA 2018 Sep 18;320(11):1131-1150. [doi: 10.1001/jama.2018.12777] [Medline: 30326495]
13. Babbott S, Manwell LB, Brown R, et al. Electronic medical records and physician stress in primary care: results from the MEMO Study. J Am Med Inform Assoc 2014 Feb;21(e1):e100-e106. [doi: 10.1136/amiajnl-2013-001875] [Medline: 24005796]
14. Melnick ER, Dyrbye LN, Sinsky CA, et al. The association between perceived electronic health record usability and professional burnout among US physicians. Mayo Clin Proc 2020 Mar;95(3):476-487. [doi: 10.1016/j.mayocp.2019.09.024] [Medline: 31735343]
15. Tajirian T, Stergiopoulos V, Strudwick G, et al. The influence of electronic health record use on physician burnout: cross-sectional survey. J Med Internet Res 2020 Jul 15;22(7):e19274. [doi: 10.2196/19274] [Medline: 32673234]
16. Li C, Parpia C, Sriharan A, Keefe DT. Electronic medical record-related burnout in healthcare providers: a scoping review of outcomes and interventions. BMJ Open 2022 Aug 19;12(8):e060865. [doi: 10.1136/bmjopen-2022-060865] [Medline: 35985785]
17. Olson K, Sinsky C, Rinne ST, et al. Cross-sectional survey of workplace stressors associated with physician burnout measured by the Mini-Z and the Maslach Burnout Inventory. Stress Health 2019 Apr;35(2):157-175. [doi: 10.1002/smi.2849] [Medline: 30467949]
18. Peccoralo LA, Kaplan CA, Pietrzak RH, Charney DS, Ripp JA. The impact of time spent on the electronic health record after work and of clerical work on burnout among clinical faculty. J Am Med Inform Assoc 2021 Apr 23;28(5):938-947. [doi: 10.1093/jamia/ocaa349] [Medline: 33550392]
19. Adler-Milstein J, Zhao W, Willard-Grace R, Knox M, Grumbach K. Electronic health records and burnout: time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. J Am Med Inform Assoc 2020 Apr 1;27(4):531-538. [doi: 10.1093/jamia/ocz220] [Medline: 32016375]
20. Akbar F, Mark G, Warton EM, et al. Physicians' electronic inbox work patterns and factors associated with high inbox work duration. J Am Med Inform Assoc 2021 Apr 23;28(5):923-930. [doi: 10.1093/jamia/ocaa229] [Medline: 33063087]
21. Avdagovska M, Kuziemsky C, Koosha H, et al. Exploring the impact of in basket metrics on the adoption of a new electronic health record system among specialists in a tertiary hospital in Alberta: descriptive study. J Med Internet Res 2024 Apr 29;26:e53122. [doi: 10.2196/53122] [Medline: 38684079]
22. Attipoe S, Huang Y, Schweikhart S, Rust S, Hoffman J, Lin S. Factors associated with electronic health record usage among primary care physicians after hours: retrospective cohort study. JMIR Hum Factors 2019 Sep 30;6(3):e13779. [doi: 10.2196/13779] [Medline: 31573912]
23. Attipoe S, Hoffman J, Rust S, et al. Characterization of electronic health record use outside scheduled clinic hours among primary care pediatricians: retrospective descriptive task analysis of electronic health record access log data. JMIR Med Inform 2022 May 12;10(5):e34787. [doi: 10.2196/34787] [Medline: 35551055]
24. Holmgren AJ, Downing NL, Bates DW, et al. Assessment of electronic health record use between US and non-US health systems. JAMA Intern Med 2021 Feb 1;181(2):251-259. [doi: 10.1001/jamainternmed.2020.7071] [Medline: 33315048]
25. Arndt BG, Micek MA, Rule A, Shafer CM, Baltus JJ, Sinsky CA. Refining vendor-defined measures to accurately quantify EHR workload outside time scheduled with patients. Ann Fam Med 2023;21(3):264-268. [doi: 10.1370/afm.2960] [Medline: 37217321]
26. Rotenstein LS, Holmgren AJ, Downing NL, Bates DW. Differences in total and after-hours electronic health record time across ambulatory specialties. JAMA Intern Med 2021 Jun 1;181(6):863-865. [doi: 10.1001/jamainternmed.2021.0256] [Medline: 33749732]
27. Anderson J, Leubner J, Brown SR. EHR overtime: an analysis of time spent after hours by family physicians. Fam Med 2020 Feb;52(2):135-137. [doi: 10.22454/FamMed.2020.942762] [Medline: 32050270]

XSL•FO
RenderX

28. Arndt BG, Micek MA, Rule A, Shafer CM, Baltus JJ, Sinsky CA. More tethered to the EHR: EHR workload trends among academic primary care physicians, 2019-2023. Ann Fam Med 2024;22(1):12-18. [doi: 10.1370/afm.3047] [Medline: 38253499]

Abbreviations

EHR: electronic health record
FTE: full-time equivalent
JD-R: job demands–resources
REDCap: Research Electronic Data Capture
TOSH: time outside scheduled hours

Edited by J Sarvestan; submitted 09.May.2025; peer-reviewed by N Shaw, R Koppel, S Sivarajkumar; revised version received 23.Feb.2026; accepted 24.Feb.2026; published 15.Apr.2026.

Please cite as:
Pauly RP, Stafinski T, Koosha H, Avdagovska M, Hadizadeh M, Bigam D, Graham T, Kuziemsky C, Kassam N, Menon D
After-Hours Use of the Electronic Health Record Among Medical and Surgical Specialists After Implementation of a System-Wide Integrated Clinical Information System in Alberta, Canada: Longitudinal Descriptive Study
J Med Internet Res 2026;28:e76872
URL: https://www.jmir.org/2026/1/e76872
doi:10.2196/76872

© Robert P Pauly, Tania Stafinski, Helia Koosha, Melita Avdagovska, Maliheh Hadizadeh, David Bigam, Timothy Graham, Craig Kuziemsky, Narmin Kassam, Devidas Menon. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Trust Building in Internet-Based Home Care Among Loyal Patients: Qualitative Study

Yan Yang, MSN; Fei Lu, MSN; Xinxin Wang, MSN; Leiwen Tang, MSN; Luchen Pan, MSN; Yu Zhang, BSc; Hongling Sun, BSc; Jia Feng, MSN; Chenling Zhu, MSN; Meijuan Lan, MSN

The Second Affiliated Hospital of Zhejiang University School of Medicine, 88 Jiefang Road, Shangcheng District, Hangzhou, China

Corresponding Author:
Meijuan Lan, MSN
The Second Affiliated Hospital of Zhejiang University School of Medicine, 88 Jiefang Road, Shangcheng District, Hangzhou, China

Abstract

Background: The growing demand for home-based care, driven by rapid population aging, has accelerated the development of internet-based home care. Despite its emerging status, a subset of loyal patients has consistently used internet-based home care with high frequency, showing strong commitment and a willingness to recommend it to others. Understanding their experiences with trust-building among loyal patients is essential to optimize and scale this service model. However, few studies have specifically explored the experience of trust-building among this patient group, leaving a significant gap in the literature.

Objective: This study aims to explore the experiences of trust-building among loyal patients in internet-based home care, guided by the cognitive-affective-conative model.

Methods: A descriptive qualitative design was used. A purposive sampling method was used to select 15 loyal patients in internet-based home care (mean age 65, SD 18.6, range 30-92 y; mean number of care visits 168, range 101-359) in Zhejiang Province, China. Semistructured interviews, informed by the cognitive-affective-conative framework, were conducted between June 2025 and August 2025. Data were analyzed using directed content analysis, facilitated by NVivo 12.0 software (Lumivero).

Results: Three core themes emerged. The first theme, cognitive dimension—building the foundation of trust through rational appraisal—included channels of trusted information, recognition of professional competence, convenience of digital services, and concerns about safety and privacy. The second theme, affective dimension—deepening trust through emotional and cultural connection—involved from professional interaction to “quasi-family” bonds, personalized care to emotional comfort, and filial piety culture as an emotional and trust catalyst. The third theme, conative dimension—translating trust into loyal behaviors within rational limits—underscored trust-driven advocacy and word-of-mouth, willingness to pay as a monetization of trust, engaged participation in service improvement, and trust-based decisions to continue or terminate services.

Conclusions: Patient trust in internet-based home care is shaped by the interplay of cognitive appraisal, emotional connection, and behavioral intention. This study shows that loyalty first originates from a rational understanding of service quality, which generates trust. During the service process, emotional connections and cultural factors, such as filial piety, further enhance this trust, gradually leading to loyalty behaviors based on rational choice, such as frequent use and word-of-mouth promotion. This study demonstrates that loyalty is not only a rational choice based on service quality but also a commitment anchored in long-term emotional relationships, further reinforced by cultural factors like filial piety that embed the services within family values. The cognitive-affective-conative model effectively captures this multifaceted experience. Consequently, moving beyond mere technical use, future development must strategically integrate emotional support and cultural sensitivity to nurture deep, sustainable trust.

(J Med Internet Res 2026;28:e88860) doi:10.2196/88860

KEYWORDS

home care; loyalty; trust; experience; qualitative study

Introduction

The rising worldwide need for long-term care, driven by rapid population aging and a growing burden of chronic diseases, has accelerated the development of internet-based home care [1]. This emerging model uses digital technology to deliver skilled nursing care directly to patients’ homes, thus improving health care accessibility and efficiency [2,3]. By acting as a bridge over distance and mobility gaps, this model shows significant potential for underserved populations, including older adults, individuals with chronic illnesses, and rural residents [4,5]. Following national pilot programs such as the “Internet Plus Nursing” initiative in 2019, China has actively promoted internet-based home care to reduce health care disparities and support aging in place [4,6].

XSL•FO
RenderX

Despite this potential, a widespread lack of public trust remains a critical barrier to large-scale adoption [7,8]. Trust, a multifaceted psychosocial construct, is pivotal in health care as it underpins patient engagement, adherence, and loyalty [9,10]. It encompasses confidence in the reliability, competence, and integrity of a service or provider. Loyalty represents the culmination of sustained trust and comprises both favorable attitudes and consistent behaviors [11]. Loyal patients are thus defined by significant loyal behaviors—characterized by long-term, high-frequency service use alongside a strong propensity for recommendation [11-13]. Extensive quantitative research across traditional health care and commercial domains has consistently shown that trust enhances outcomes such as patient satisfaction, treatment adherence, and repeat patronage [14-16]. Furthermore, qualitative research findings clearly emphasize the need for health care practitioners and public health officials to prioritize trust-building in community care work, because trust fosters patient engagement, enhances treatment cooperation, and strengthens the patient-provider relationship [17].

However, within the novel and rapidly evolving context of internet-based home care, understanding how trust is built becomes even more critical. This model introduces unique dynamics involving digital interfaces, in-home service delivery, and the mediation of care through a platform, which collectively reshape traditional patient-provider interactions and risk perceptions [7,9]. Investigating how trust is built in this specific setting is therefore essential not merely to fill a gap in literature but to generate actionable insights that can optimize service design, enhance patient adoption, and ensure the sustainable integration of digital solutions into long-term care ecosystems. A deep comprehension of the trust-building experience, particularly among those who have demonstrably embraced the service (loyal patients), is crucial for informing strategies that foster wider acceptance, improve care quality, and ultimately realize the potential of internet-based home care to address pressing societal health challenges.

To systematically investigate how trust is built among loyal patients in this novel service model, this study uses the cognitive-affective-conative (CAC) model [18], which, unlike models focusing primarily on cognitive determinants such as the technology acceptance model [19] and the unified theory of acceptance and use of technology [20], offers a more holistic lens by integrating cognitive appraisal with emotional attachment and behavioral loyalty to capture the dynamic evolution of trust. The CAC model is a well-established theoretical framework that explains how attitudes are formed and transformed across 3 sequential dimensions: cognitive (beliefs and knowledge), affective (emotional responses and attachment), and conative (behavioral intentions and actions). It has been widely applied in health technology adoption research, such as studies on middle-aged adults' usage of health apps [21] and users' continued engagement with health education videos [22]. As an attitude, trust is defined as a positive evaluation of the reliability and integrity of a service or provider [9,10]. Trust is conceptualized as a dynamic, multistage psychosocial construct that evolves across these 3 interrelated dimensions—from cognitive appraisal to emotional attachment to behavioral commitment [15]. This operationalization aligns with the CAC model's sequential logic: cognition shapes initial trust, affective connections deepen trust, and conative behaviors actualize and sustain trust.

Guided by the CAC model framework, this study used a descriptive qualitative design to examine the trust-building experiences of loyal patients receiving internet-based home care in Zhejiang Province, China.

Methods

Research Design

Descriptive qualitative research is particularly apt for exploring phenomena that have not been extensively researched [23], as it permits a nuanced and detailed depiction of participants' lived experiences and sentiments without imposing pre-existing theoretical constraints. Given the paucity of studies specifically examining the health care experiences of loyal patients in internet-based home care, this methodological approach was deemed most appropriate. Furthermore, the reporting of this study meticulously adhered to the SRQR (Standards for Reporting Qualitative Research) guidelines (Checklist 1) [24].

Setting and Sampling

This study used purposive sampling, augmented by a maximum variation sampling strategy to ensure participant diversity across age, gender, educational attainment, and types of services received. Participants were recruited between June 2025 and August 2025 from the "Zheli Nursing" application platform in Zhejiang Province, China.

Inclusion criteria were as follows: (1) being aged 18 years or older, (2) having clinical care needs matching the scope of internet-based home care services (primarily including malignant tumors, chronic illnesses such as diabetes and hypertension, and postoperative care needs such as wound management and stoma care after major surgeries), (3) sustained use of internet-based home care services for a prolonged period (defined as ≥6 months [25,26]), (4) high-frequency service use (assessed as receiving at least 100 service visits between January 2023 and May 2025; this threshold was determined via analysis of user frequency data from the "Zheli Nursing" platform, representing a level near the median usage among platform users), (5) a favorable attitude toward recommending the same health care provider (assessed via phone screening), and (6) cognitive lucidity and the ability to articulate their perspectives accurately.

Exclusion criteria were as follows: the presence of severe cardiac, neurological, or other conditions that would impede participation in interviews.

Sample size determination adhered to the principle of data saturation, wherein recruitment ceased when no new information or themes emerged from successive interviews [27]. A total of 19 eligible patients were initially enrolled. Among them, 3 participants completed preliminary (pilot) interviews, which were conducted solely to test and refine the interview guide; data from these pilot interviews were not included in the final analysis. Subsequently, 1 enrolled participant was hospitalized

XSL•FO
RenderX

due to a change in their medical condition and declined to proceed with the formal interview, resulting in withdrawal from the study. Consequently, a final sample of 15 participants completed the in-depth interviews and were included in the data analysis.

Data Collection

Data collection was performed using semistructured interviews. The preliminary interview guide, developed based on the CAC model and previous literature [15,18,28], was pilot-tested with 3 eligible patients to identify potential issues in wording, question sequencing, or comprehensibility and subsequently revised accordingly. Two methodologists with extensive qualitative research experience independently reviewed and validated the refined interview guide. The final interview guide encompassed the following key areas, mapped to the CAC dimensions. First, in the cognitive dimension, the key areas were How did you first learn about internet-based home care, and what led you to try it? How do you assess the professionalism and skills of the nurses who provide care? What advantages or conveniences do you perceive in using this digital service compared to hospital visits? Have you had any concerns about safety, privacy, or reliability when using the service? If so, how were they addressed? Second, in the affective dimension, the key areas were How would you describe your relationship with the nurses who visit you? Has it changed over time? In what ways, if any, has the service provided emotional support or comfort to you or your family? Does using this service affect the life of you or your family feel about caregiving and family responsibilities? Finally, in the conative dimension, the key areas were What has motivated you to continue using the service over a long period? Have you recommended internet-based home care to others? If so, why and to whom? Under what circumstances would you consider discontinuing the service? What suggestions do you have to improve the service in the future?

Before each interview, researchers introduced themselves, clarified the study's purpose and procedures, and obtained written informed consent, informing participants that all interactions would be audio-recorded. Researchers worked to build trust with patients. Interviews were conducted in a quiet, private environment selected by the participant, with no other people present. Ten interviews were conducted in participants' homes, while 5 were conducted via video call. Throughout the interviews, researchers meticulously observed and documented nonverbal cues such as facial expressions and body language, carefully noting emotional shifts and attitudinal changes. Participants were actively encouraged to articulate their perspectives freely. If the responses were ambiguous or required further elaboration, researchers used techniques such as strategic rephrasing of key information or asking probing questions to ensure data completeness and accuracy. Interview durations ranged from 35 to 57 minutes.

Data Analysis

During the analysis, interview recordings and other non-numeric data were coded, identified, and summarized using Lumivero NVivo 12 qualitative analysis software [29]. This study used a directed content analysis approach [30], guided by the CAC model as its theoretical framework. This method is particularly well-suited for validating or extending existing theoretical frameworks [31]. The analysis was conducted independently by 2 researchers (XW and FL), both of whom have training in qualitative methods. The analytical approach comprised the following steps. First, the analysis used a hybrid deductive-inductive approach. The CAC model provided an initial theoretical framework for deductive guidance, while the researchers remained open to inductive insights emerging directly from the data. Two researchers first independently immersed themselves in the interview transcripts to achieve familiarity. An initial codebook was developed based on the CAC model and relevant literature. Second, to ensure the robustness of the coding scheme, a pilot phase was conducted. Three transcripts were randomly selected and independently coded by the 2 researchers using the initial codebook. Their results were compared, and discrepancies were discussed. This pilot test led to revisions, including refining definitions, adding new codes for uncovered concepts, and clarifying categorization criteria. This iterative process resulted in a finalized codebook. Third, using the finalized codebook, the two researchers independently coded all 15 transcripts. The text was examined line by line to identify meaningful units, which were labeled with corresponding codes. New codes were created inductively when necessary. Related codes were then grouped based on their conceptual relationships and properties to form broader categories and subcategories. Fourth, the categorized data were further abstracted and organized to identify overarching themes that captured the core patterns relevant to the research questions. Data collection and analysis were conducted concurrently until thematic saturation was achieved, meaning that no new codes or themes emerged from the final interviews. Finally, throughout the analysis, negative case analysis was actively used. The researchers consciously sought out and interrogated data that contradicted or nuanced the emerging themes, ensuring that the findings accounted for disconfirming evidence and reflected the full complexity of participants' experiences.

This rigorous, iterative process resulted in a structured coding tree (Multimedia Appendix 1). The coding process and its relationship to the subthemes and overarching theme are exemplified in the coding tree (Textbox 1).

XSL•FO
RenderX

Textbox 1. Example of a coding tree.

Raw data

P3: *Before being discharged, I worried about how to manage my wound at home. The doctor told me I could book a nurse for home dressing changes, and I immediately felt relieved.*

My wound is both deep and extensive, and compounded by my diabetes, its management is considerably challenging.

The visiting nurse demonstrates exceptional professionalism; she specializes in treating such cases at the hospital.

One time, Nurse Mu rushed back from out of town to help me with my wound. She was 5 hours late due to traffic, but I was willing to wait for her just like I would for my daughter.

Meaning-bearing content

- Information from trusted health care providers reduces anxiety and builds initial trust.
- Recognition of nurses' specialized skills and professional competence builds trust.
- Strong personal commitment to specific nurses reflects deep trust bonds.

Code

- Trusted information from clinicians
- Recognition of professional competence
- Patient patience and understanding

Subtheme

- Channels of trusted information
- Recognition of professional competence
- From professional interaction to "quasi-family" bonds

Overarching theme

- Cognitive dimension—building the foundation of trust through rational appraisal.

Quality Control and Research Rigor

The rigor of this qualitative study was comprehensively ensured across 4 key dimensions: credibility, dependability, confirmability, and transferability [32,33].

All interviews were conducted by the first author in collaboration with a clinical nurse. Both were female, held master's degrees in nursing, and were trained in qualitative methodologies. The first author, with extensive experience in nursing management, actively promoted home care initiatives. During data analysis, 2 researchers (XW and FL) independently coded all transcripts and conducted multiple comparative analyses and discussions until consensus was reached. If consensus could not be reached, a third senior researcher was consulted to arbitrate and make the final decision. The refined themes and subthemes were presented back to the participants for member checking (ie, validating the researchers' interpretation with participants) to confirm their resonance with the lived experiences. Participants did not raise any objections to the results. To strengthen dependability, all interview recordings were transcribed verbatim independently by the 2 researchers within 24 hours postinterview, followed by cross-verification of the transcribed texts against the original audio. In instances of ambiguous phrasing or incomplete information during transcription, researchers promptly contacted participants for clarification or conducted supplementary interviews to ensure data accuracy and completeness. Regarding confirmability, the research team maintained continuous reflective journals throughout the data collection and analysis phases. These journals focused on critically examining aspects such as adequate respect for participant perspectives, avoidance of personal assumptions, leading language and potential biases, appropriateness of analytical methods, vigilance against the omission of crucial information, and the potential influence of these factors on data interpretation. The refined themes and subthemes were presented back to the participants for member checking to validate their resonance with the lived experiences. Participants did not raise any objections to the results. Finally, the transferability of the study findings was ensured through a comprehensive and detailed description of the research context, participant sociodemographic characteristics, and methodological procedures.

Researcher Reflexivity and Positionality

The primary researchers hold master's degrees in nursing and are actively engaged in promoting home care initiatives. This professional background provided deep insider knowledge and facilitated rapport with participants, likely encouraging richer discussions about care experiences. However, this same positionality carried a potential risk of pro-innovation bias—a subconscious tendency to view the service positively or to overlook its shortcomings. To mitigate this, we implemented several safeguards. First, the interview guide was designed with open, neutral questions that explicitly asked about challenges and negative experiences. Second, data analysis was conducted

XSL•FO
RenderX

independently by 2 researchers who were not involved in data collection, bringing a degree of outsider perspective to counteract potential collection biases. Third, negative case analysis was deliberately applied throughout the analytic process to identify and interrogate data that contradicted or refined the emerging themes of positive trust-building. This reflexive practice aimed to balance the benefits of insider understanding with disciplined critical distance throughout the research process.

Ethical Considerations

Ethical approval for this study was obtained from the medical ethics committee of the Second Affiliated Hospital of Zhejiang University School of Medicine (2025‑1196). Researchers rigorously adhered to the principles of voluntary participation and participant confidentiality. Formal interviews were initiated only after participants had provided written informed consent. Participants did not receive any form of compensation for their involvement in this study. To safeguard personal identity information, all participants were assigned unique identifiers (eg, P1, P2, and so on), which were used in interview records and throughout the analysis. Any incidental mentions of third parties within the interview transcripts were redacted before data analysis. Access to raw data or unredacted materials was strictly restricted to the research team.

Results

General Characteristics of Participants

A total of 15 loyal patients in internet-based home care participated in this study (Table 1). Patients ranged in age from 30 to 92 years, with a mean age of 65 (SD 18.6) years. The home visits they received ranged from 101 to 359, with an average of 168. In terms of education, 46.7% (n=7) of patients had a secondary school education. The vast majority of care (n=14, 93.3%) for patients was provided by family members.

XSL•FO
RenderX

Table . General characteristics of interviewees (N=15).

Participant	Sex	Age (y)	Education level	Primary caregiver	Primary diagnosis	Key nursing procedures	Number of visits	Same nurse[a]
P1	Male	60	Higher education	Care worker	Paraplegia and pressure injury	Pressure injury care	359	Yes
P2	Male	92	Primary education	Offspring	Pressure injury, diabetes, and hypertension	Pressure injury care	146	Yes
P3	Male	72	Secondary education	Offspring	Heatstroke, hypertension, and diabetes	Chronic difficult wound	104	Yes
P4	Female	73	Uneducated	Offspring	Postcolorectal cancer surgery	Wound care and ostomy care	101	Yes
P5	Female	41	Secondary education	Spouse	Malignant tumor	Chronic wound and lymphedema care	103	Yes
P6	Male	76	Primary education	Offspring	Postliver cancer surgery	Pain management, drain bag change, and venipuncture	105	Partially
P7	Female	64	Primary education	Offspring	Postmalignant tumor surgery, chronic heart failure, and nephropathy	Subcutaneous injection, PICC[b] care, and drain bag change.	292	No
P8	Female	86	Secondary education	Offspring	Stroke, hypertension, and diabetes	Pressure injury care, bladder irrigation, and blood glucose monitoring.	154	Yes
P9	Female	33	Higher education	Parents	Postfracture surgery	Wound care and pain management	151	Yes
P10	Female	72	Uneducated	Offspring	Malignant tumor	PICC care and pain management	170	Yes
P11	Female	87	Secondary education	Offspring	Cerebrovascular accident, hemiplegia, and hypertension.	Nasogastric feeding care, catheter care, and subcutaneous injection.	114	No
P12	Female	67	Secondary education	offspring	Stroke, hypertension, and diabetes	Pressure injury care, catheter care, and nasogastric feeding care	163	Yes
P13	Female	50	Secondary education	Spouse	Malignant tumor	PICC care and pain management	110	No

XSL•FO
RenderX

Participant	Sex	Age (y)	Education level	Primary caregiver	Primary diagnosis	Key nursing procedures	Number of visits	Same nurse[a]
P14	Female	69	Secondary education	offspring	Malignant tumor	Venipuncture, infusion port care, and pressure injury care	158	Yes
P15	Female	33	Higher education	Spouse	Pregnancy and threatened abortion	Subcutaneous or intramuscular injection, venipuncture, and neonatal care	292	No

[a]Same nurse: whether the patient designates the same nurse for every visit.

[b]PICC: peripherally inserted central catheter.

Primary diagnoses of patients predominantly included malignant tumors, chronic illnesses, and post–major surgical conditions. The 36 distinct nursing services provided were grouped into specialized and general procedures based on their operational complexity. The performance of nursing procedures is strictly governed by staff qualifications, which are defined by operational complexity. For specialized procedures such as complex wound management and implantable port care, provincial or higher specialized certifications or senior titles are mandated. For general procedures, including subcutaneous injections and urinary catheter changes, the criteria are an intermediate title coupled with a minimum of 5 years of clinical experience.

The interviewed loyal patients shared their views on the experiences and drivers that build trust. From the interviews, 3 main categories emerged (Figure 1). It is worth noting, however, that when talking about direct care, all informants in this study referred to nurses collectively. The term “staff,” without distinguishing between nurses with different educational levels, was used, and the patients carefully avoided mentioning any names. Only when talking about social relationships with the nurses did a few patients use the first name of a specific nurse with whom they had a closer relationship.

Figure 1. The hierarchical diagram of the 3 main themes and their associated subthemes.

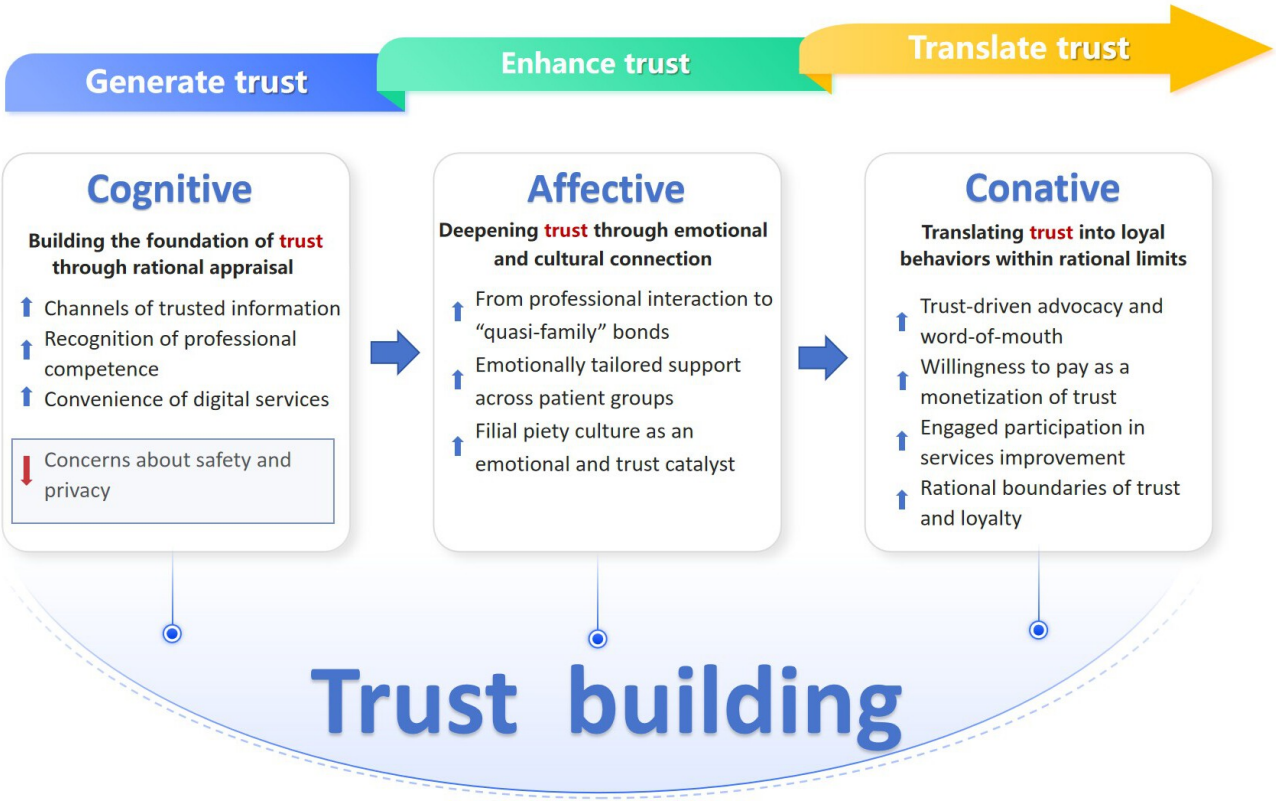

XSL•FO
RenderX

Cognitive Dimension: Building the Foundation of Trust Through Rational Appraisal

Within this category, 4 subcategories were generated: channels of trusted information, recognition of professional competence, convenience of digital services, and concerns about safety and privacy, reflecting patients' cognitive assessment of internet-based home care at the rational level, which form the basis for trust and loyalty.

Channels of Trusted Information

The process of building trust in Internet-based home care began with patients becoming aware of the services, with the information source playing a critical role in shaping their initial level of confidence. Participants consistently reported that information from interpersonal and professional channels—such as recommendations from treating clinicians during hospital discharge or from friends and family with direct experience—was paramount in alleviating anxieties and building foundational trust. These trusted sources acted as a powerful validation of the services' reliability and safety. In contrast, impersonal channels, like media advertisements, were effective for raising general awareness but often required reinforcement from personal networks to translate awareness into trust:

> *Before being discharged, I worried about how to manage my wound at home. The doctor told me I could book a nurse for home dressing changes, and I immediately felt relieved.* [P3]

> *A friend recommended I use my phone to book a nurse for injections at home, saying his father used it and it was great. Actually, I had seen advertisements on WeChat videos before, but I didn't take it seriously until a friend recommended it; then I tried it.* [P5]

Recognition of Professional Competence

Interviewed patients consistently expressed confidence in the nurses' ability to manage their health issues at home. The professional competence of the nursing staff was emphasized as a fundamental driver in building initial trust and encouraging continued use of services. This reliance was particularly evident among patients with complex or high-risk care needs, who showed a strong preference for nurses with specialized technical skills:

> *My wound is both deep and extensive, and, compounded by my diabetes, its management is considerably challenging. The visiting nurse demonstrates exceptional professionalism; she specializes in treating such cases at the hospital.* [P3]

> *The infusion port is a new device, and general nurses usually don't know how to handle it. I always ask nurse Huang to come to my home.* [P14]

Convenience of Digital Services

Patients reported that the services significantly reduced the time and effort associated with hospital visits, highlighting the convenience of receiving care at home. The platform's ability to offer a wide range of choices was also frequently mentioned. Even older adult patients, who faced digital barriers, experienced this practical benefit with caregiver assistance, which reinforced their trust and willingness to continue using the services:

> *I have bone metastasis from breast cancer, and I need to go to the hospital 2 times a week. It used to require an ambulance and two family members to accompany me. Now that nurses come to my home, it's not only convenient for my family, but I no longer need an ambulance.* [P5]

> *Booking on my phone is very convenient. I can choose the hospital, nurse, and time based on my situation, and I can even book several services together.* [P7]

> *I'm old and don't know how to use a smartphone, but my daughter handles the booking. She says it's very easy to just click a few buttons on the phone.* [P4]

Concerns About Safety and Privacy

Some patients reported concerns regarding the reliability, safety, and privacy protection of services during their initial use. While trust in government-endorsed platforms and transparent information could alleviate some safety doubts, perceived risks of privacy breaches during actual service delivery remained a potential hurdle to the full establishment of trust:

> *My wife is concerned about the safety of strangers coming into our home. But I'm not worried because I booked through the Provincial Health Commission's internet hospital platform and chose a nurse from a well-known hospital.* [P6]

> *Once, when the nurse was here, my wife and daughter-in-law argued. I felt very awkward and worried the nurse would talk about it outside, spreading our "family shame."* [P6]

Affective Dimension: Deepening Trust Through Emotional and Cultural Connection

In this category, 3 subcategories emerged: from professional interaction to "quasi-family" bonds, personalized care to emotional comfort, and filial piety culture as an emotional and trust catalyst, revealing the emotional dependencies and cultural drivers formed by loyal patients during long-term service interactions, which are pivotal in transforming initial trust into deep-seated trust and loyalty.

From Professional Interaction to "Quasi-Family" Bonds

Patients described how, over multiple visits and as their health needs were consistently met, their relationships with nurses gradually deepened beyond a purely professional dynamic. Many recounted developing strong personal trust with their nurses, often using family-like terms to describe these bonds. In their narratives, patients highlighted small, personal gestures and emotional support that made them feel cared for beyond the clinical procedures. This sense of personal connection contributed significantly to their overall trust in the services. The strength of these relational bonds was also evident in patients' willingness to accommodate unusual circumstances. Some expressed a high degree of patience and understanding when service delays occurred, reflecting the depth of their personal commitment to specific nurses:

XSL•FO
RenderX

> *Nurse Zhang is like a member of my family. She encourages me, and even brings local snacks during holidays. I've been bedridden for 37 years, and I'm most afraid of pressure injuries, but with her these past 2 years, I feel at ease.* [P1]

> *One time, nurse Mu rushed back from out of town to help me with my wound. She was 5 hours late due to traffic, but I was willing to wait for her just like I would for my daughter.* [P3]

> *Nurse Li is like my daughter. She is professional in wound care, chats with me.* [P2]

Personalized Care to Emotional Comfort

The pathways through which emotional connection deepened trust varied significantly among patients, closely aligned with their specific health conditions and corresponding care expectations. Participants with complex, high-risk needs—such as difficult wound management—associated emotional security directly with the nurse's specialized technical competence. The assurance that a highly skilled professional was managing their care provided them with substantial reassurance. In contrast, for patients requiring routine procedures, such as catheter changes, emotional comfort arose primarily from the service's overall reliability and convenience. Consistent and accessible care alleviated their anxiety, rendering the specific identity of the nurse less important. Younger patients, including new mothers, derived emotional comfort from the efficiency of the service and its seamless integration into their daily lives, which contributed to a practical sense of ease:

> *My complex wound can't be handled by ordinary nurses. I have to ask nurse X to help me—her skills are that good.* [P3]

> *I need regular catheter changes and enteral feeding tube care. For these, as long as a nurse can come to my home to handle these issues, that's fine—convenience is key, and it doesn't matter who comes. However, when it comes to wound care, I want a specialist nurse to do it.* [P8]

> *Having a nurse come to my home for injections has been fantastic. It means I can avoid hospital trips entirely and have real peace of mind during my pregnancy. After my baby was born, I booked neonatal home care, so convenient.* [P15]

Filial Piety Culture as an Emotional and Trust Catalyst

Some patients frequently described how internet-based home care helped alleviate feelings of being a burden on their families, with their narratives often highlighting a sense of relief as the services not only addressed their clinical needs but also provided a modern solution to the challenge of long-term care—enabling their offspring to fulfill filial duties despite practical constraints, in keeping with traditional beliefs. The motivation of their offspring to book these services was often interpreted by patients as a direct expression of care and filial devotion, strengthening their trust in both the services and their family relationships. For families with members abroad, the services were portrayed as a vital digital bridge. Overseas offspring could actively participate in their parents' care remotely, transforming physical absence into engaged, thoughtful involvement and easing anxieties on both sides:

> *Every hospital visit in the past meant my offspring had to take turns using their precious leave to accompany me. It was hard on their jobs. The guilt would weigh on my conscience. Now, with the nurses coming, no one needs to take leave. Everyone is so much happier.* [P11]

> *My offspring don't have the time to be constantly present, nor the specialized skills needed, like changing wound dressings. Now, with home-based nursing care, this issue is beautifully resolved.* [P8]

> *My daughter insisted on choosing the services for me. She was worried about hospital infections.* [P4]

> *My offspring are abroad, I'm lying at home. They are very filial, although they can't be here to look after me. I don't talk much about my condition for fear of worrying them, but that just makes them worry more. Now, they book nurses for me online. Every time a nurse visits, they join a video call and talk with her about my health. They feel much more at ease abroad.* [P11]

Conative Dimension: Translating Trust Into Loyal Behaviors Within Rational Limits

The third theme captured how trust is behaviorally enacted, focusing on the conative dimension of patient loyalty. Subcategories included trust-driven advocacy and word-of-mouth, willingness to pay as a monetization of trust, engaged participation in service improvement, and trust-based decisions to continue or terminate services.

Trust-Driven Advocacy and Word-of-Mouth

Most interviewed patients expressed a strong willingness to engage in trust-driven word-of-mouth dissemination. Their recommendations were often directed toward other patients and families in similar health situations, highlighting a pattern of advocacy rooted in shared experiences and the recognized value of services:

> *If there are long-term bedridden patients who need care, I will recommend it to them.* [P9]

> *Internet-based home care is especially beneficial for patients discharged after major surgery with weakened immunity. It saves them from going to the hospital and reduces cross-infection. I will recommend it to these people.* [P4]

Willingness to Pay as a Monetization of Trust

A clear willingness to pay a premium for the services was expressed by many patients, even when acknowledging the higher costs compared to in-hospital care and limited insurance coverage. In explaining their stance, participants frequently framed their decision not primarily in terms of financial cost, but through the lens of the overall value received—weighing professional competence, relational continuity, and emotional support against the monetary outlay:

XSL•FO
RenderX

> *Each visit costs 2 to 3 hundred yuan, and after over 100 visits, the financial pressure is not small, but I still insist on choosing it.* [P10]
>
> *Although it's more expensive than the hospital, compared to the cost of an ambulance and my family's lost work time, it's still more cost-effective.* [P5]

Engaged Participation in Services Improvement

Participants provided substantive suggestions for service improvement, reflecting a proactive engagement with the internet-based home care platform. Their input frequently addressed perceived barriers to wider adoption, such as cost, and extended to visionary proposals for enhancing care delivery. These included aspirations for multidisciplinary team visits and the integration of advanced technologies to support aging in place:

> *I hope it can be covered by basic medical insurance, reimbursing part of the cost, or by commercial insurance for full reimbursement.*
>
> *[P8]*
>
> *I hope ultrasound can come to the home; doctors, rehabilitation therapists, and nurses can all visit together.* [P1]
>
> *In the future, smart monitoring equipment can be rented, with data linked. When we are recovering or aging at home, it can intelligently remind us to eat, take medicine, and exercise. And give me a robot to talk to me and to push my wheelchair when I want to go out.* [P10]

Trust-Based Decisions to Continue or Terminate Services

Participants described clear boundaries to their use of internet-based home care, delineating the conditions under which their trust and loyalty would rationally apply. Their continued reliance was contingent upon the services' capacity to safely and effectively meet their clinical needs at home. Once those needs evolved beyond the scope of home-based care or were resolved entirely, patients acknowledged they would discontinue the services, indicating a pragmatic and context-dependent dimension to their trust:

> *Unless my condition changes and I must be hospitalized, otherwise I will continue to choose it.* [P4]
>
> *Once fully recovered and mobile, I won't need nurses to come home.*
>
> *[P9]*

Discussion

To our knowledge, this study is the first to investigate how trust is built among loyal patients in internet-based home care, guided by the CAC model. Our findings reveal that trust develops through a dynamic, multistage process: it is initially formed through cognitive appraisal of the service's use and professionalism, deepened via affective and cultural connections, and ultimately enacted through conative behaviors that reinforce and sustain loyalty.

Principal Findings

The foundation of trust is established within the cognitive dimension, where patients conduct a rational evaluation of the services' core value and reliability. Consistent with the CAC model's premise that cognition forms the basis of attitude [18], our findings confirm that trust originates from logical appraisals. First, access to trusted information channels—particularly interpersonal recommendations from treating clinicians or family members—serves as a critical initial trigger. This aligns with prior research informed by perceived risk theory, which emphasizes that information from credible sources significantly reduces uncertainty and facilitates initial trust formation [34]. Our results further substantiate that, unlike impersonal channels such as advertisements, interpersonal endorsements provide essential social validation that enhances trustworthiness, corroborating earlier findings on the primacy of relational cues in health services adoption. Second, the recognition of professional competence—especially in managing complex care needs such as difficult wounds—constitutes a cornerstone of cognitive trust. This observation resonates with French and Raven's [35] bases of social power, particularly the concept of expert power, wherein trust is grounded in perceived knowledge and skill. Our data reinforce prior studies indicating that patients' reliance on nurses with specialized certifications reflects a non-negotiable expectation of professional credibility in health care contexts—a consistent theme across trust-related literature in both digital and conventional care settings. Third, the convenience afforded by digital services—including reduced travel burden and flexible scheduling—was shown to enhance perceived usefulness and ease of use, core constructs of the technology acceptance model [19]. This finding is consistent with earlier work on technology-mediated care, which highlights practical use as a key facilitator of trust. However, our study also reveals that this nascent trust remains fragile and can be undermined by persistent concerns about safety and privacy. This echoes prior research underscoring the need for systemic safeguards, such as government endorsement and transparent service protocols [36], to solidify cognitive trust—a recurring recommendation in studies on digital health adoption.

Moving beyond rationality, trust is profoundly deepened in the affective dimension through relational and cultural elements. The evolution of nurse–patient relationships from professional interactions to "quasi-family" bonds is pivotal. This transition, characterized by emotional companionship and personalized care, fulfills patients' needs for belonging and emotional security, resonating with the principles of emotional social support theory [37]. Such emotional bonds are particularly crucial for elderly or chronically ill patients, as they mitigate feelings of isolation and transform the service into a source of psychosocial sustenance—thereby deepening trust beyond mere transactional satisfaction. These findings align with prior qualitative studies in long-term and home-based care, which similarly identify emotional support and personalized interaction as key mechanisms for sustaining patient engagement and trust [15,16]. Furthermore, this study highlights the underexplored role of filial piety culture as an emotional and trust catalyst. In a societal context where adult offspring often struggle to provide hands-on care due to geographic or professional constraints,

XSL•FO
RenderX

internet-based home care offers a culturally congruent solution. It enables offspring to discharge their filial duties by procuring professional care, thus embedding the service within a valued cultural narrative [38]. This cultural resonance generates a powerful, affect-based trust that is particularly potent in the Chinese context. Our results corroborate and extend recent discrete choice experiments indicating that Chinese patients and families highly value home-based care models that align with filial norms [37]. This demonstrates how trust is built not only through service quality but also through alignment with deeply held cultural values—a dimension less emphasized in Western-centric trust models. Importantly, the underlying mechanisms illuminated by this culturally specific finding may hold broader relevance. The core idea—that trust is deepened when a service resonates with users' fundamental relational values and emotional needs—is likely generalizable. In Western contexts, different cultural or social values (eg, autonomy, independence, and community support) might serve as analogous trust catalysts [39]. For instance, a home care service that effectively supports an older adult's desire to maintain independence or that strengthens their connection to a local community could generate similarly powerful affective trust [40]. Therefore, while the specific cultural artifact (filial piety) is context-bound, the broader theoretical implication is that trust-building strategies must identify and align with the core relational and emotional value systems of the target population.

The conative dimension represents the behavioral manifestation of trust. Loyal patients do not merely hold trust internally; they enact it through tangible behaviors. Trust-driven advocacy and word-of-mouth are direct translations of trust into promotion. Patients become voluntary ambassadors, primarily recommending the services to peers in similar situations—a behavior rooted in empathy and shared experience. This aligns with Dick and Basu's [11] conceptualization of loyalty, wherein a positive attitude leads to supportive behavioral intentions, and extends prior findings that recommendation intent is a core component of attitudinal loyalty in health care settings [14]. Similarly, the willingness to pay a premium, despite limited insurance coverage, demonstrates the monetization of trust. Patients conduct a cost-benefit analysis, as suggested by expectation-confirmation theory [41], and their willingness to pay reflects a holistic valuation of the trust they have developed—encompassing both cognitive and affective benefits. This finding resonates with economic loyalty models that frame price tolerance as a function of perceived value and relational investment [10], yet it further specifies that in home care, this valuation integrates clinical competence with emotional and cultural use. Moreover, patients' engaged participation in service improvement signals a high level of trust and co-ownership. This proactive engagement indicates that trust has fostered a partnership dynamic, where patients feel invested in the service's success. Such behaviors reflect principles of co-production often discussed in service-dominant logic and patient-centered care literature [14,16], but our study situates them within the digital home care context as outcomes of accumulated trust. Crucially, this study also identifies the rational boundaries of trust and loyalty. Trust is not blind. Patients' loyalty remains contingent on the service's ability to meet their clinical needs safely and effectively. This rational boundary underscores a mature and context-dependent form of trust, ensuring that the trust-building process remains grounded in realistic expectations of the service's scope. This nuanced view challenges the sometimes static portrayals of loyalty in earlier literature [42], highlighting its conditional and dynamic nature in ongoing service relationships.

Theoretical and Practical Implications

In synthesis, the process of how trust is built in internet-based home care is a reinforcing cycle: cognitive assessments establish its foundation, emotional and cultural connections deepen its roots, and conative behaviors both manifest and reinforce it, creating a positive feedback loop that sustains trust and loyalty. This dynamic aligns with and extends the CAC model [18], while also resonating with social exchange theory in relational contexts, where reciprocal reinforcement between trust and behavior has been widely observed [11]. This study's identification of filial piety as a distinct emotional and trust catalyst extends the application of the CAC model in digital long-term care in two significant ways. First, it demonstrates the model's contextual adaptability, showing how a core cultural value permeates all 3 trust-building dimensions. Filial norms influence cognitive appraisals (eg, interpreting offspring's service booking as an act of care), fuel affective bonds by transforming professional care into "quasi-family" relationships, and motivate conative loyalty as acts of family duty [38]. Thus, filial piety acts as an active, culturally embedded force that strengthens the entire CAC sequence. Second, and more importantly, it refines the model's affective dimension by specifying a powerful, culturally shaped source of emotional attachment—thereby revealing a generalizable mechanism. While the CAC model acknowledges that emotions deepen trust [18], it often treats their origins as individual, idiosyncratic, or vague. Our findings demonstrate that affective responses can be systematically rooted in shared normative values that function as collective motivators [38]. This theoretical insight moves beyond the specific case of filial piety; in any cultural context, trust is deepened when a service resonates with users' fundamental relational values and emotional needs [39]. This mechanism is culturally universal, though its manifestations are context-specific. In China, filial piety serves as the dominant cultural frame; in Western contexts, analogous trust catalysts may include autonomy and independence (eg, services that enable older adults to age in place without relying on family), community connectedness (eg, services that integrate home care with local social networks), or reciprocity and dignity (eg, models that frame care as partnership rather than dependency) [40,41]. Consequently, the CAC model evolves from a general psychological sequence into a culturally informed framework that can better predict which emotional drivers will be effective in specific settings.

For practice, this multifaceted understanding demands integrated strategies that are both universally applicable and culturally adaptable. To build cognitive trust, platforms must ensure information transparency, showcase nurse credentials, optimize usability, and implement robust safety and privacy protocols. These measures directly address the cognitive drivers identified in this study and are consistent with prior recommendations for mitigating perceived risk in digital health services [34,36]. To

XSL•FO
RenderX

foster affective trust, service models should encourage relational continuity (eg, through the same nurse assignment) and train nurses in empathetic communication and emotional support—strategies that are effective across cultures. However, the cultural framing of marketing and service narratives must be locally tailored. In China, highlighting the service's role in enabling filial piety has proven powerful [38]; in other contexts, promotional messages might emphasize preserving independence, strengthening community ties, or upholding personal dignity. The common principle is not the specific value used, but the strategic alignment with what patients and families culturally hold dear. To facilitate conative trust, providers should create structured channels for patient feedback and co-design, reflecting the partnership dynamic observed in our data. Meanwhile, policymakers should explore insurance reforms to reduce financial barriers, thereby validating patients' willingness to invest in a trusted service—a finding supported by both our results and economic models of loyalty [10,41]. These integrated strategies not only respond to the multi-stage trust-building process identified in this study but also align with emerging calls in the literature for more culturally and emotionally informed approaches to digital home care implementation [2,6,38].

Limitations

There are various limitations to consider when interpreting the results. First, the operational definition of "loyalty" is based on a specific behavioral threshold and participants' self-reported word-of-mouth recommendation intent. Future research would benefit from incorporating multidimensional, objective behavioral indicators, such as frequency, duration, and cross-temporal continuity of use, for more robust validation, or could further explore the robustness of findings under different operational thresholds to define and measure patient loyalty more comprehensively. Second, during data collection, requiring participants to retrospectively recount their experiences with internet-based home care may have introduced recall bias. The scope of this study's "loyal customer" interview sample was restricted to Zhejiang Province, China. Given the substantial regional variations in economic conditions, cultural contexts, and health care resources across China, these geographic limitations may impact the generalizability of our findings. Third, by focusing exclusively on loyal patients, this study offers deep insight into the processes of trust consolidation but does not capture experiences of trust erosion, dissatisfaction, or service discontinuation. Consequently, our understanding of trust dynamics remains partial. Future research should explicitly address this gap by incorporating comparative designs, for example, interviewing nonloyal users, former users who discontinued the service, or those who experienced trust violations, to provide a more holistic view of trust formation, maintenance, and attrition in internet-based home care. Finally, while qualitative research offers rich, in-depth insights, the generalizability of its conclusions warrants further substantiation. We recommend that future studies develop standardized scales and integrate quantitative methods to rigorously analyze the influence pathways and weights of the cognitive, affective, and conative dimensions on loyal behavior. Subsequent research should also broaden the sample scope and conduct multicenter studies nationwide to comprehensively and profoundly investigate variations in internet-based home care experiences regarding their trust across diverse populations.

Conclusions

This study, grounded in the CAC model, is the first to systematically reveal the dynamic process of trust-building among loyal patients in internet-based home care. The findings indicate that trust formation originates from a rational cognitive appraisal of information channels, professional competence, service convenience, and safety and privacy concerns. Through sustained interactions, nurse-patient relationships evolve into "quasi-family" bonds, and under the influence of filial piety culture, emotional connections deepen, embedding trust within both affective and cultural dimensions. Ultimately, trust translates into loyal behaviors, including continued use, word-of-mouth advocacy, willingness to pay, and engaged participation in service improvement, all bounded by rational considerations of clinical need and safety.

The originality and contribution of this study are threefold. Theoretically, it extends the CAC model to the context of digital home care and identifies "filial piety culture" as a key emotional and trust catalyst, enhancing the model's explanatory power and cultural adaptability. Empirically, by focusing on the under-researched group of loyal patients, it maps the complete trajectory of trust evolution from cognitive appraisal through affective deepening to behavioral enactment, offering a novel framework for understanding sustained trust in digitally mediated care. Practically, the findings provide clear guidance for service optimization: strengthening nurses' dual professional and emotional competencies, implementing "same-nurse" systems, improving platform usability, enhancing privacy protection, promoting insurance coverage, and aligning service design with cultural values such as filial piety to support families in fulfilling care responsibilities.

In conclusion, through its theoretical, empirical, and practical contributions, this study offers a theoretically grounded and actionable roadmap for building a trust-driven, patient-centered digital home care ecosystem in aging societies.

Acknowledgments

The authors sincerely thank all the interviewees and research team members who participated in this study and extend heartfelt gratitude to Yao Meiqi, Zong Hongying, and Zhu Jieying for their support and suggestions. The authors used DeepSeek for language polishing during the writing of the paper. This tool did not participate in the generation or interpretation of academic content, and all artificial intelligence–assisted editing content was reviewed and approved by the authors.

XSL•FO
RenderX

Funding

This study was funded by the Zhejiang Provincial Department of Education of China through a general research project (project Y202351911). The funders provided only financial support and did not participate in the research design, data collection, data analysis, result interpretation, or paper writing.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: ML, YZ
Data curation: LP
Formal analysis: XW, FL
Funding acquisition: YY
Investigation: YY, HS, LP
Methodology: XW
Resources: YY
Software: YY, ML
Supervision: YY, ML
Validation: YY, ML
Visualization: YY, CZ
Writing – original draft: YY
Writing – review & editing: JF, LT

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete coding tree.
[DOCX File, 19 KB - jmir_v28i1e88860_app1.docx]

Checklist 1
SRQR checklist.
[DOCX File, 20 KB - jmir_v28i1e88860_app2.docx]

References

1. Yang J, Ma B, Chen S, et al. Nurses' preferences for working in Uber-style "Internet plus" nursing services: a discrete choice experiment. Int J Nurs Stud 2025 Jan;161:104920. [doi: 10.1016/j.ijnurstu.2024.104920] [Medline: 39378739]
2. Jiang M, Zhao J, Zhao R, et al. Resident preferences for telemedicine services in China in the digital health era: mixed methods study. J Med Internet Res 2025 Sep 3;27:e67390. [doi: 10.2196/67390] [Medline: 40903005]
3. Wu M, Zhang Y, Song L, Linghu Y. Analysis of the current status and frontiers of domestic research on "Internet + Nursing Services" [Article in Chinese]. Chin J Clin Res 2021;34(2):279-282. [doi: 10.13429/j.cnki.cjcr.2021.02.035]
4. Notice from the General Office of the National Health Commission on launching pilot programs for "Internet Plus Nursing Services" [Article in Chinese]. National Health Commission of the People's Republic of China. 2019 Feb 12. URL: https://www.nhc.gov.cn/yzygj/c100068/201902/5b2c18d54d0c487db84562fe2c780114.shtml [accessed 2005-10-25]
5. Ageing and health. World Health Organization. 2025 Oct 1. URL: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health [accessed 2005-10-25]
6. Notice from the General Office of the National Health Commission on further promoting the pilot program of "Internet + Nursing Services". National Health Commission of the People's Republic of China. 2020. URL: https://www.nhc.gov.cn/wjw/c100175/202012/5888310873ae422c98f1365a033f1324.shtml [accessed 2005-10-25]
7. Xu X, Li N, Gao F, et al. Residents' willingness and awareness to participate in internet+nursing service in China: a meta-analysis. Medicine (Abingdon) 2025;104(2):e41219. [doi: 10.1097/MD.0000000000041219]
8. Fang S, Yu J, Hu F, Wang Z, Jiang H. Potential profile analysis of community residents' willingness to use "Internet + Nursing Services" and its influencing factors [Article in Chinese]. Int J Nurs 2025;44(11):2091-2096. [doi: 10.3760/cma.j.cn221370-20240919-00458]
9. Zhao Y, Zhao Q. Current situation and influencing factors of "Internet + Nursing Service" use in Shanghai from the perspective of patients [Article in Chinese]. Chin J Prac Nurs 2023;39(3):195-201. [doi: 10.3760/cma.j.cn211501-20220821-02645]

XSL•FO
RenderX

10. Belli A, O'Rourke AM, Carrillat FA, Pupovac L, Melnyk V, Napolova E. 40 years of loyalty programs: how effective are they? Generalizations from a meta-analysis. J Acad Mark Sci 2022 Jan;50(1):147-173. [doi: 10.1007/s11747-021-00804-z]
11. Dick AS, Basu K. Customer loyalty: toward an integrated conceptual framework. J Acad Mark Sci 1994 Mar 1;22(2):99-113. [doi: 10.1177/0092070394222001]
12. Taylor AH, Taylor RS, Ingram WM, et al. Adding web-based behavioural support to exercise referral schemes for inactive adults with chronic health conditions: the e-coachER RCT. Health Technol Assess 2020 Nov;24(63):1-106. [doi: 10.3310/hta24630] [Medline: 33243368]
13. Dávila Espuela N, Reina Paz MD, Sevilla Sevilla C. Social networks and customer loyalty: review of loyalty keys and main social networks publications' characteristics. Front Psychol 2023;14:1286445. [doi: 10.3389/fpsyg.2023.1286445] [Medline: 37928572]
14. Olesen K, Bathula H. A meta-analysis of the determinants of patient satisfaction and loyalty. Health Mark Q 2022;39(2):191-210. [doi: 10.1080/07359683.2022.2050000] [Medline: 35306972]
15. Hu HL, Kuo CL, Wang SY, Chuang YH. Perceptions of empathy among nursing assistants in long-term care facilities: a qualitative study. J Nurs Manag 2024;2024(1):3072064. [doi: 10.1155/2024/3072064] [Medline: 40224795]
16. Li X, Dorstyn D, Mpofu E, et al. Nursing assistants and resident satisfaction in long-term care: a systematic review. Geriatr Nurs 2021;42(6):1323-1331. [doi: 10.1016/j.gerinurse.2021.08.006] [Medline: 34560527]
17. Slatton BC, Cambrice FG, Omowale SS. COVID-19 vaccine hesitancy among Black women in the US. JAMA Netw Open 2025 Jan 2;8(1):e2453511. [doi: 10.1001/jamanetworkopen.2024.53511] [Medline: 39786776]
18. Rosenberg MJ, Hovland CI, McGuire WJ, Abelson RP, Brehm JW. Attitude Organization and Change: An Analysis of Consistency among Attitude Components: Yale University Press; 1960:1-14 URL: https://psycnet.apa.org/record/1962-00924-000 [accessed 2026-03-20]
19. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
20. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]
21. Huang YM, Lou SJ, Huang TC, Jeng YL. Middle-aged adults' attitudes toward health app usage: a comparison with the cognitive-affective-conative model. Univ Access Inf Soc 2019 Nov;18(4):927-938. [doi: 10.1007/s10209-018-0621-9]
22. Xuan W, Phongsatha T, Hao L, Tian K. Impact of health science popularization videos on user perceived value and continuous usage intention: based on the C-A-C and ECM model framework. Front Public Health 2024;12:1382687. [doi: 10.3389/fpubh.2024.1382687] [Medline: 39011330]
23. Doyle L, McCabe C, Keogh B, Brady A, McCann M. An overview of the qualitative descriptive design within nursing research. J Res Nurs 2020 Aug;25(5):443-455. [doi: 10.1177/1744987119880234] [Medline: 34394658]
24. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
25. Peyrusqué E, Buckinx F, Kergoat MJ, Aubertin-Leheudre M. Exercise guidelines to counteract physical deconditioning in long-term care facilities: what to do and how to do it? J Am Med Dir Assoc 2023 May;24(5):583-598. [doi: 10.1016/j.jamda.2023.01.015] [Medline: 36822232]
26. Fong MC, Russell D, Brickner C, Gao O, Vito S, McDonald M. Medicaid long-term care workforce training intervention and value-based payment metrics. Health Serv Res 2022 Apr;57(2):340-350. [doi: 10.1111/1475-6773.13930] [Medline: 34921725]
27. Saunders B, Sim J, Kingstone T, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907. [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
28. Kallio H, Pietilä AM, Johnson M, Kangasniemi M. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. J Adv Nurs 2016 Dec;72(12):2954-2965. [doi: 10.1111/jan.13031] [Medline: 27221824]
29. Dhakal K. NVivo. J Med Libr Assoc 2022 Apr 1;110(2):270-272. [doi: 10.5195/jmla.2022.1271] [Medline: 35440911]
30. Graneheim UH, Lundman B. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse Educ Today 2004 Feb;24(2):105-112. [doi: 10.1016/j.nedt.2003.10.001] [Medline: 14769454]
31. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
32. Ahmed SK. The pillars of trustworthiness in qualitative research. J Med Surg Public Health 2024 Apr;2:100051. [doi: 10.1016/j.glmedi.2024.100051] [Medline: 16204405]
33. Shenton AK. Strategies for ensuring trustworthiness in qualitative research projects. Educ Inf 2004;22(2):63-75. [doi: 10.3233/EFI-2004-22201]
34. Lăzăroiu G, Neguriţă O, Grecu I, Grecu G, Mitran PC. Consumers' decision-making process on social commerce platforms: online trust, perceived risk, and purchase intentions. Front Psychol 2020;11:890. [doi: 10.3389/fpsyg.2020.00890] [Medline: 32499740]

XSL•FO
RenderX

35. French JRP, Raven B. The bases of social power. In: Cartwright D, editor. Studies in Social Power: Michigan University Press; 1959:150-167 URL: https://atl.web.baylor.edu/sites/g/files/ecbvkj1781/files/2024-01/nathan_alleman_byron_newberry_and_sarah_madsen-_faculty_authority.pdf [accessed 2026-03-20]
36. Bouyer F, Thiongane O, Hobeika A, et al. Epidemic intelligence in Europe: a user needs perspective to foster innovation in digital health surveillance. BMC Public Health 2024 Apr 6;24(1):973. [doi: 10.1186/s12889-024-18466-1] [Medline: 38582850]
37. Roosen I, Salway S, Osei-Kwasi HA. Transnational social networks, health, and care: a systematic narrative literature review. Int J Equity Health 2021 Jun 12;20(1):138. [doi: 10.1186/s12939-021-01467-6] [Medline: 34118934]
38. Liu Y, Du S, Liu C, Xue T, Tang Y. Preference of primary care patients for home-based healthcare and support services: a discrete choice experiment in China. Front Public Health 2024;12:1324776. [doi: 10.3389/fpubh.2024.1324776]
39. Callaghan T, Moll H, Rakoczy H, et al. Early social cognition in three cultural contexts. Monogr Soc Res Child Dev 2011 Aug;76(2):vii-viii. [doi: 10.1111/j.1540-5834.2011.00603.x] [Medline: 21767264]
40. Abdi S, Spann A, Borilovic J, de Witte L, Hawley M. Understanding the care and support needs of older people: a scoping review and categorisation using the WHO international classification of functioning, disability and health framework (ICF). BMC Geriatr 2019 Jul 22;19(1):195. [doi: 10.1186/s12877-019-1189-9] [Medline: 31331279]
41. Pan G, Mao Y, Song Z, Nie H. Research on the influencing factors of adult learners' intent to use online education platforms based on expectation confirmation theory. Sci Rep 2024 Jun 4;14(1):12762. [doi: 10.1038/s41598-024-63747-9] [Medline: 38834737]
42. Marzo RR, Bhattacharya S, Ujang NB, et al. The impact of service quality provided by health-care centers and physicians on patient satisfaction. J Educ Health Promot 2021;10(1):160. [doi: 10.4103/jehp.jehp_979_20] [Medline: 34222535]

Abbreviations

CAC: cognitive-affective-conative
SRQR: Standards for Reporting Qualitative Research

Edited by A Stone; submitted 02.Dec.2025; peer-reviewed by D Chen, N Cui, S Wang, X Wu; revised version received 27.Feb.2026; accepted 27.Feb.2026; published 15.Apr.2026.

Please cite as:
Yang Y, Lu F, Wang X, Tang L, Pan L, Zhang Y, Sun H, Feng J, Zhu C, Lan M
Trust Building in Internet-Based Home Care Among Loyal Patients: Qualitative Study
J Med Internet Res 2026;28:e88860
URL: https://www.jmir.org/2026/1/e88860
doi:10.2196/88860

© Yan Yang, Fei Lu, Xinxin Wang, Leiwen Tang, Luchen Pan, Yu Zhang, Hongling Sun, Jia Feng, Chenling Zhu, Meijuan Lan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Child Vaccination Status and Behavioral and Social Drivers of Vaccination Among Their Caregivers in the Philippines: Cross-Sectional Survey Study Comparison of Household, Mobile, and Online Modes

Kimberly E Bonner[1], MPA, PhD; Mikka Hipol[2], MPA; Dominique Sy[2], MSc; Rivandra Royono[2], MPA/ID; Douglas Johnson[2], MPA; Isabel del Rosario[2], MSc; Alice Redfern[2], BMBCh; Darahlyn Biel-Romualdo[2], BSc; Man Kai Wong[1], MPH; Eugene Lam[1], MSc, MSPH, MD; Shibani Kulkarni[1], MPH, PhD; Kirsten Ward[1], MPH(Hons), DrPH; Romel Lacson[3], MPH, PhD; Meng-Yu Chen[4], MSc, MD; James Matthew Miraflor[4], BS; Devon Ray Pacial[5], MPH; Rowena Bunoan[5], MMPA; Hugo Catan[5], BA; Talya Shragai[1], PhD

[1]Global Immunization Division, Centers for Disease Control and Prevention, 1600 Clifton Rd NE, Atlanta, GA, United States

[2]IDInsight, Makati, Philippines

[3]Centers for Disease Control and Prevention, Global Health Center, Atlanta, GA, United States

[4]The Task Force for Global Health, Atlanta, GA, United States

[5]Republic of the Philippines Department of Health, Manila, Philippines

Corresponding Author:
Kimberly E Bonner, MPA, PhD
Global Immunization Division, Centers for Disease Control and Prevention, 1600 Clifton Rd NE, Atlanta, GA, United States

Abstract

Background: The World Health Organization recommends that countries routinely collect data on the behavioral and social drivers (BeSD) of vaccination to inform public health interventions that increase vaccine uptake. There is a need to identify data collection methods that can rapidly and inexpensively collect representative data, particularly in low- and middle-income countries.

Objective: This study aimed to understand BeSD drivers of vaccination in the Philippines and assess the trade-offs between survey methods. We compared responses to household, mobile, and online surveys in terms of demographics, vaccination status, responses to BeSD questions, and cost.

Methods: We conducted concurrent household, mobile (SMS text messaging and interactive voice response), and online surveys among caregivers of children 2 years of age and below in Regions V and XII of the Philippines, with sampling differing by survey method. We assessed, for each survey method, (1) respondent demographics (sex, age, region, and socioeconomic status) and (2) the weighted proportion of responses from caregivers of children who received at least one dose of diphtheria-pertussis-tetanus (DPT)–containing vaccine. We estimated the weighted proportion of each BeSD survey response option and calculated the financial cost (monetary outlays) per survey response from an implementer's perspective by summing the costs incurred in each survey method and dividing by the number of responses received.

Results: We surveyed a total of 1201 household respondents, 2153 mobile respondents, and 398 online respondents from January to March 2025. We found that online and mobile survey respondents were more likely to be male and have completed high school than household survey respondents. The weighted proportion of respondents indicating that their child had received at least one dose of DPT vaccine was 91.8% (n=1090; 95% CI 90% - 93.3%) for the household survey, 90.3% (n=1853) for the mobile survey, and 85% (n=346) for the online survey. With regard to vaccine demand, more than 85% of respondents in each survey method indicated that vaccines are very important, very safe, supported by family, and that they knew where to bring a child for vaccination. More than 30% of mobile and online survey respondents indicated that it was not easy to pay for vaccination. The financial cost to conduct the survey per survey response was US $2.61 for the online survey, US $6.93 for the mobile survey, and US $29.38 for the household survey.

Conclusions: In the Philippines, household, mobile, and online survey methods reached caregivers of children who were unvaccinated against DPT, and these proportions were similar across survey methods. BeSD responses indicated high vaccine demand and challenges in caregivers' cost to access vaccination. Determining the most appropriate survey method depends on trade-offs between representativeness and costs. However, areas with strong connectivity and high mobile device ownership can consider mobile and online methods as a lower-cost alternative to rapidly collect BeSD data.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e81059) doi:10.2196/81059

KEYWORDS

mHealth; low- and middle-income countries; health surveys; reproducibility of results; data collection methods; vaccination; behavioral and social drivers; mobile surveys; mobile health

Introduction

In low- and middle-income countries (LMICs), there is a pressing need for methods of collecting health data beyond household surveys that can inform public health practice [1], particularly for measuring drivers of vaccine demand [2,3].

While household surveys are considered a gold standard for population-representative data [4], there are drawbacks to this method, including declining participation in urban areas, household survey costs, exclusion of nonstationary households [5], and infrequent periodicity [6]. Mobile phone surveys and online surveys offer promising alternatives to household surveys, particularly in areas with high mobile phone ownership rates and network penetration [7]. One key concern of both mobile and online surveys is that the population responding to these surveys might not be representative of the overall population of interest [8,9], and the responses to survey questions might differ by survey modality [10]. While some comparisons between remote data collection via self-administered SMS text messaging surveys have been made in LMICs, these assessments focused on nutrition indicators [10]. While mobile methods for social mobilization and vaccination reminders have been extensively employed in vaccination activities [11], few studies have examined bidirectional messaging or data collection.

The World Health Organization recommends that countries routinely monitor the behavioral and social drivers (BeSD) of vaccination at the community level using a series of validated questions derived from the BeSD framework [12]. An understanding of the drivers of vaccination in a population enables public health programs to develop tailored interventions that address specific population needs and concerns regarding vaccination. Of particular importance is understanding the needs of unvaccinated, also known as zero-dose, populations, defined as receiving zero doses of diphtheria-pertussis-tetanus (DPT)–containing vaccine. There is a need to identify other survey methods that can rapidly collect BeSD data at scale [13]. In the context of BeSD surveys, there is a need to understand how the proportion of respondents with unvaccinated and under-vaccinated children differs by survey method, including its associated costs.

The Philippines has high mobile phone penetration, widespread internet coverage, a high population of children who have not received recommended vaccines (zero-dose) [14], and documented barriers to vaccine demand [15]. Mobile phone ownership was 122 per 100 people [16], with an estimated 92% of households possessing at least 1 mobile phone in 2022 [17]. Cellular network coverage was 99%, and third generation of cellular network technology (3G) network access, which enables internet connectivity, was 96% in 2024 [18]. In addition, the Philippines has a longstanding community health worker system of Barangay Health Workers (BHWs), based in the smallest administrative unit (barangay). BHWs support primary care service delivery by maintaining contact with a designated list of families in their catchment area [19]. Self-administered mobile surveys distributed by BHWs offer an opportunity to leverage the potential speed and scalability of mobile data methods while garnering an improved response rate over random digital dial through the use of trusted BHW distribution networks. In addition, the Department of Health has established an online survey platform that can be readily accessed through a web search, Facebook, or Viber.

The overall objective of this study is to compare respondent characteristics and BeSD survey responses using 3 different data collection methods: household surveys, self-administered mobile surveys distributed via SMS text messaging by BHWs, and online surveys. Specific aims of this study are to (1) describe respondent demographics and vaccination status by data collection method, (2) compare how the distribution of responses to BeSD questions varies by survey method, and (3) describe the process of implementing the 3 survey methods, including financial costs from an implementer's perspective and a data quality assessment of response accuracy in the mobile survey.

Methods

Recruitment

Participants for each survey method were recruited from 2 regions in the Philippines: Region V (Bicol) and Region XII (SOCCSKSARGEN). These regions were selected due to mobile penetration above 90% [17], a high population of children who received zero doses of DPT or with an incomplete vaccination series (30.1% unvaccinated in Region XII [17]; 64.7% under-vaccinated in Region V [20]), and security conditions that permit in-person data collection. Participants for each survey method were eligible if they were 18 years of age or older, were caregivers of a child 2 years old or younger, and were residents of Region V or XII during the survey period.

Sampling differed by survey method (Table 1). In the household survey, a 2-stage cluster sampling approach was used, drawing from the barangay, the smallest administrative units in each region. In first-stage cluster sampling, 91 barangays in Region V and 63 barangays in Region XII were selected using probability proportional to size, drawn from 2020 census data. The Philippines health system engages BHWs to maintain a comprehensive list of households in their catchment area to ensure that these households receive relevant health information and services. The second stage drew a simple random sample of 8 households with a child under 2 years of age from the BHW listing from each barangay, for a total target of 1200 households. The estimated sample size of 1200 for the household survey was calculated to ensure a margin of error on the outcome of interest of 5 percentage points or fewer, assuming 150 clusters, 8 households per cluster, an intraclass correlation of 0.3 on DPT

XSL•FO
RenderX

vaccination status, and a control group mean of 0.5 (SD 0.5), based on the Demographic Health Survey estimate of 58% of children in these regions reported as fully immunized. Trained data collectors from the study team approached each of the randomly selected households in person and conducted an informed consent process with those who were eligible. If no eligible person was available at the household, data collectors visited each household a second time, and, if no eligible person was available again, they then randomly selected another household in the barangay as a replacement. Following informed consent, data collectors verbally administered a face-to-face 65-question survey in English, Tagalog, or an applicable regional language (including Cebuano/Bisaya, Hiligaynon, or Bikol), according to the preference of the participant. When available and with caregiver consent, data collectors verified the reported vaccination history by directly examining the child's vaccine card. Deidentified data were recorded on a password-protected device, and survey responses were uploaded to a secure web server in accordance with the 2012 Data Privacy Act [21]. Participants with unvaccinated or incompletely vaccinated children were counseled to seek vaccination services from the nearest vaccination site.

Table . Summary of sampling, survey administration, survey questions, and weights by survey method: Philippines Regions V and VII, January-March 2025.

Survey method	Sampling	Administration	Number of questions	Analysis
Household	Two-stage cluster sampling: Stage 1: barangays selected using probability proportional to size. Stage 2: simple random sample of 8 households with a child under 2 y of age from the Barangay Health Worker (BHW) household listing.	In-person, interviewer-administered on a password-protected device.	65	Base weights for the probability of selecting each barangay and household based on probability proportional to size. Poststratification weights were used to match the marginal totals for strata defined by the combination of educational attainment of the household head, main water source, and region.
Mobile	Two-stage cluster sampling: Stage 1: barangays selected using probability proportional to size, drawn from 2020 census data. Stage 2: simple random sample of 2 BHWs per barangay who distributed survey invitation links to households listed in their catchment areas.	SMS text messaging invitation link sent by BHWs to eligible households in their catchment areas. Self-administered via SMS text messaging or interactive voice response, as selected by the participant.	15	Poststratification weights were used to match the marginal totals for strata defined by the combination of educational attainment of the household head, main water source, and region.
Online	Nonprobabilistic respondent-driven internet recruitment sampling; opt-in via Knowledge Informs Responsible Action chatbot.	Self-administered online survey	15	Poststratification weights were used to match the marginal totals for strata defined by the combination of educational attainment of the household head, main water source, and region.

In the mobile phone survey, barangays, the smallest administrative units in each region, were selected using probability proportional to size sampling in the same municipalities, but not the same barangays, as the in-person survey sites. A simple random sample of 2 BHWs per barangay was selected to distribute the mobile survey via SMS text messaging invitation link, using the BHW lists, which contained contact information, including mobile phone numbers, for each of the households with children 2 years of age and below. These BHWs were briefed on the purpose of the survey and asked to share an opt-in survey keyword via SMS text messaging with all eligible individuals using their catchment area household listing. To facilitate this outreach, each BHW received a one-time communication allowance of Php 300 (~US $5.40). As a quality check, BHWs were asked to report the total number of eligible individuals in their catchment area and the number of eligible individuals who had been sent the survey invitation message. The survey was hosted on the EngageSPARK platform, a self-service platform that sends interactive automated phone calls (interactive voice response [IVR] surveys) and 2-way SMS text messaging campaigns and collates survey responses. Recipients of the SMS text messaging survey invitation could opt into the survey by entering the survey shortcode included in the survey invitation to access the EngageSPARK platform. Following an initial screening questionnaire to confirm age, location, and whether there were children 2 years of age or younger in the household, participants were asked to proceed within the survey if they consented to participate. Next, participants could select to proceed in English or Tagalog and via SMS text messaging or by IVR.

In the online survey, we employed nonprobabilistic respondent-driven internet recruitment sampling, where regionally geo-targeted Facebook advertisements encouraged

XSL•FO
RenderX

eligible respondents to opt into the survey. Respondents could opt into participating in the online survey through the pre-existing Knowledge Informs Responsible Action chatbot, known locally as *Katuwang na Impormasyon para sa Responsableng Aksyon*, hosted by the Philippines Department of Health (DOH) [22] and maintained by AI4GOV. The Knowledge Informs Responsible Action chatbot service offers information, collects survey data, and can be accessed through the DOH webpage, DOH Facebook page, and via Viber.

Measures

Across the 3 surveys, we used consistent measures to ascertain eligibility, respondent demographics, child vaccination status, and responses to the BeSD questions.

The interviewer-administered household survey included 16 questions on caregiver demographics, 28 on child demographics and vaccination status, and 21 BeSD questions (Table S1 in Multimedia Appendix 1). Each survey method used consistent questions, and the household survey included additional questions at the behest of local authorities.

The self-administered mobile survey consisted of 15 questions, including 4 demographic questions, 7 core BeSD of vaccination questions, and 4 questions on demographics and vaccination status of the youngest child under 2 years of age in the household (Table S1 in Multimedia Appendix 1). We opted to assess a subset of the household survey questions, in alignment with the number and estimated completion time of other mobile surveys [7].

The self-administered online survey began with a welcome, eligibility confirmation, and informed consent process. Participants completed the 15-question survey in English, Tagalog, or Cebuano/Bisaya (Table S1 in Multimedia Appendix 1). The survey consisted of 4 demographic questions, 7 core BeSD questions, and 4 questions on demographics and vaccination status of the youngest child under 2 years of age in the household.

All responses were according to self-report, although DPT vaccination status was also recorded by card verification when available. Region was classified as either Region V, Region XII, or missing. The age of the respondent at the last birthday was recorded in years. The sex of the respondent was classified as male, female, or missing.

To assess proxies for socioeconomic status, we assessed educational attainment and main water source. Education was classified as did not graduate from elementary, elementary graduate, high school graduate, college graduate, postgraduate studies, or missing, and then collapsed to a binary (completed vs did not complete high school) for the analysis. Water source was classified as piped water, dug well, tube well or borehole, water from a spring, water refilling station, others, or missing.

We classified children under 2 years of age who received zero doses of DPT-containing vaccine as zero-dose, consistent with the literature [23]. Vaccination status was classified as received at least 1 dose of DPT or did not receive DPT. This variable was assessed by vaccine card verification when available in the household survey, and by self-report in all other cases (mobile and internet surveys, plus household survey respondents without available vaccine cards). Self-reported DPT vaccination status was used in the analysis, and a sensitivity analysis compared differences in DPT vaccination status between respondents who provided vaccination card verification and those who provided only self-report.

The BeSD questions were drawn from 7 of the WHO BeSD validated core questions [12] (Table S1 in Multimedia Appendix 1). We assessed confidence in vaccine importance with the question, “How important do you think vaccines are for your child’s health?” with response options of Very important, Moderately important, A little important, and Not at all important. Similarly, confidence in vaccine safety was assessed with the question, “How safe do you think vaccines are for your child?” with response options of Very safe, Moderately safe, A little safe, and Not at all safe. We assessed social norms with the question, “If it was time for your child to get vaccinated, would the mother need permission from any household member to take your child to the clinic?” with response options of Yes or No. In the domain of practical issues, 3 questions assessed access issues. Knowledge of where to access the vaccine was assessed with the question “Do you know where to go to get your child vaccinated?” with response options of Yes and No. Difficulty of access was assessed with the question, “How easy is it to get vaccination services for your child?” with ordinal response options of Not at all easy, A little easy, Moderately easy, and Very easy. Vaccine costs, including opportunity costs, were assessed with the question, “How easy is it to pay for vaccination?” with ordinal response options of Not at all easy, A little easy, Moderately easy, and Very easy. Vaccination intent was assessed with the question, “The Philippines has a schedule of vaccines for children. Do you want your child to get none of these vaccines, some of these vaccines or all of these vaccines?” with response options All, Some, or None.

Statistical Analysis

Statistical analysis was conducted in Stata 18 [24].

Household survey data were analyzed using base weights for the probability of selecting each barangay and household based on the probability proportional to size, drawn from the population projections in the 2020 census. No base weights were incorporated in the mobile or online data, given the sampling design of these methods assumes uniformity across the areas. For mobile and online data, CIs are not reported, as they would not accurately reflect the true uncertainty due to the nonprobabilistic sampling approach; instead, we focus on point estimates while acknowledging the limitations in generalizability [25].

For the household, mobile, and online data, poststratification weights were used to match the marginal totals for strata defined by the combination of educational attainment of the household head, main water source, and region.

We reported respondent demographics, using the raw counts and weighted proportions for each demographic variable.

To assess the proportion of survey respondents whose children had received at least 1 dose of DPT by survey method, we calculated the poststratification weighted proportion of DPT

XSL•FO
RenderX

vaccination status for each survey method, as well as the 95% CI for the household survey.

We assessed the distribution for 6 of the core BeSD questions listed above by estimating the weighted proportion of responses for each survey method. We did not assess the seventh BeSD question on vaccination intent, as intent was not directly linked with DPT vaccination status. We further explored the distribution of BeSD responses by DPT vaccination status by estimating the weighted proportion of BeSD responses by vaccination status for each survey method. We report the 95% CIs for the household survey data. We assumed missingness completely at random and used a complete records analysis.

Process Description and Survey Costing

We describe the process of implementing each of the 3 survey methods, assessed programmatic costs, and conducted a data quality assessment for the mobile survey responses. We assessed the financial cost (monetary outlays) from the implementer's perspective using a retrospective top-down costing approach. Measures were total cost per survey method and the cost per survey response. Financial expenses incurred by the implementer for each of the 3 survey methods were obtained from the implementer's budget. Costs included in the analysis were all financial costs for personnel time (paid labor), per diem, transportation, equipment, BHW worker incentives, survey participant incentives, venue rental, supplies, and advertising and platform costs for each data collection method. We included costs from these program activities: planning and data collection (survey administration), cleaning, validation, and storage. Costs were assessed in 2025 Philippine Pesos (Php). We used the conversion rate on Wise on April 4, 2025 (US $1=Php 57.01) to calculate the exchange rate from Peso to US dollars, with no inflationary adjustments. These overall costs were divided by the number of survey responses completed to yield the financial cost per completed survey for each survey method. Values are presented in nominal US dollars.

The data quality assessment focused on accuracy of responses at 2 time points. First, a preprocessing assessment identified ineligible or duplicate responses automatically immediately following survey responses. Responses were deemed ineligible if the response included a respondent age below 18 years, a respondent region of residence outside of Philippines Regions V (Bicol) and XII (SOCCSKSARGEN), or a respondent-reported child age above 2 years. Second, we randomly selected 10% of mobile survey respondents for a callback to confirm eligibility. Among the mobile survey respondents who received a callback, we assessed the number flagged for quality concerns, tabulated the weighted proportion for each of the demographics and BeSD responses, and then calculated the Pearson chi-squared design-based *F* statistic and *P* values (Table S2 in Multimedia Appendix 1). These statistical analyses were possible within a single survey method, although it was not possible to conduct these analyses between the 3 survey methods, given the variations in sampling and weighting approaches. It was not possible to conduct a quality assessment for the online survey, given that we could not assess duplicate entries from the same source and contact information was not collected for follow-up verification.

Ethical Considerations

This study was assessed and approved by the Philippines Department of Health Single Joint Ethics Review Board (SJREB-2024 - 55) on November 18, 2024. This activity was reviewed by the Centers for Disease Control and Prevention (CDC)'s Human Subjects Office, deemed not research, and was conducted consistent with applicable federal law and CDC policy (see, eg, §45 CFR part 46, 21 CFR part 56; 42 USC §241(d); 5 USC §552a; 44 USC §3501 et seq). For the household survey, all participants received Php 100 (~US $1.80) to thank them for their time. For the mobile survey, participants received Php 50 (~US $0.90) upon survey completion. Participants did not receive any remuneration for their participation in the online survey.

Results

Respondent Characteristics

Household, mobile, and online survey methods were implemented during January to March 2025. Among the 1605 households visited for the household survey, 1208 respondents were available, of whom 1206 consented, and 1201 respondents completed the household survey (Figure S1 in Multimedia Appendix 2). Among the estimated 4451 eligible households in the participating CHW's catchment areas, 2987 consented to participate, of whom 2500 completed the mobile survey. Following quality checks of the mobile survey data, 272 responses were excluded, including 252 perfect duplicates within the same municipality, 18 ineligible upon backchecking, and 2 with incomplete information. Among the 2135 mobile responses retained in the analytic dataset, more than 99% (2135/2153) had been completed via SMS text messaging, compared to IVR (18/2153). Among the online respondents, 26,464 initiated the survey, 996 completed it, but only 396 of them reported that they resided in the 2 eligible regions. The final analytic dataset included 1201 household survey responses, 2153 mobile responses, and 398 online responses, for a total of 3752 responses (Table 2). Excluded respondent demographics differed significantly from the analytic dataset (Table S2 in Multimedia Appendix 1).

XSL•FO
RenderX

Table . Demographics of respondents by survey method, by number and weighted proportion: Philippines Regions V (Bicol) and XII (SOCCSKSARGEN), January–March 2025.

Variables	Household		Mobile		Online	
	Value, n	Weighted % (95% CI)[a]	Value, n	Weighted %[b]	Value, n	Weighted %[b]
Region						
Region V (Bicol)	697	53.2	749	53.2	296	58.4
Region XII (SOCCSKSARGEN)	504	46.8	1404	46.8	102	41.6
Sex						
Male	55	4.5 (3.3 - 6.1)	433	17.7	20	15.1
Female	1146	95.5 (93.9 - 96.7)	1712	82.3	378	84.9
Missing	__[c]	—	8	—	—	—
Age (y)						
18 - 25	326	29.3 (26.4 - 32.3)	1078	45.4	35	13.7
26 - 40	724	58.3 (54.9 - 61.6)	940	48.7	342	78.6
41 - 60	130	10.6 (9 - 12.5)	110	5.6	18	3
>60	21	1.8 (1.2 - 2.9)	7	0.3	3	4.7
Missing	—	—	18	—	—	—
Education						
Did not complete high school	321	26.8	436	37.1	17	24.5
High school graduate	877	73.2	1687	62.9	381	75.5
Missing	3	—	30	—	—	—
Main drinking water source						
Piped water	455	39	753	39	56	39.8
Tube well or borehole	133	20.1	326	20.1	11	18.2
Dug well	49	7.6	167	7.6	5	3.9
Water from a spring	83	8.9	184	8.9	26	8.7
Water refilling station	458	21.8	672	21.8	276	26.2
Others	23	2.6	51	2.6	24	3.2

[a]Household survey data were analyzed using base weights for the probability of selecting each barangay and household based on the probability proportional to size, drawn from the population projections in the 2020 census, along with poststratification weights to match the marginal totals of each stratum of age and sex to the regional distribution of educational attainment and main water source for each region.

[b]Poststratification weights were used to match the marginal totals of each stratum of age and sex to the regional distribution of educational attainment and main water source for each region. We do not report CIs for the mobile and online surveys, given that CI estimates would substantially underestimate the true variability due to their sampling approaches.

[c]Entries with no inputs (zero responses) and entries where weighted proportions cannot be calculated due to zero responses are indicated as "—".

The unweighted count and weighted proportion of respondents by demographic characteristic varied substantially by survey method. The household survey had the smallest proportion of male respondents at 4.5% (n=55; 95% CI 3.3% - 6.1%), followed by the online survey at 15.1% (n=20), and the mobile survey at 17.7% (n= 433). Age also varied by survey method, with 78.6% (n=342) of online respondents reporting an age between 26 and 40 years, compared with 58.3% (n=724; 95% CI 54.9% - 61.6%) in the household survey and 48.7% (n=940) of respondents in the mobile survey. Online survey respondents were the most likely to be high school graduates at 75.5% (n=381), followed by household respondents at 73.2% (n=877), and mobile respondents at 62.9% (n=1687).

XSL•FO
RenderX

The probability of reaching a household with a child vaccinated against DPT was consistent between household and mobile surveys at 91.8% (n=1090; 95% CI 90% - 93.3%) and 90.3%, respectively, but substantially lower at 85% among online respondents (Figure 1). Among the 93.1% of household survey respondents who gave permission for vaccination card validation, 92.6% (n=1017; 95% CI 91% - 94%) of children received at least 1 dose of DPT vaccine. We found caregiver-reported DPT vaccination status (n=73, 85.2%; 95% CI 76.7% - 91%) was somewhat lower than that of those with card-verified data.

At least 85% of respondents in each survey method reported the highest level of vaccine demand, indicating that vaccines are very important, very safe, supported by family, and that they knew where to access vaccination (Figure 2). In response to the BeSD question on vaccine importance, 94.5% (n=1139; 95% CI 92.4% - 96%) of household survey respondents, 89.3% (n=2258) of mobile respondents, and 92.2% (n=381) of online respondents indicated that vaccines are very important. In response to the BeSD question on vaccine safety, 88.5% (n=1056; 95% CI 86% - 90.6%) of household survey respondents, 89.7% (n=2240) of mobile respondents, and 92.8% (n=386) of online respondents indicated that vaccines are very safe. In response to the BeSD question on family support for vaccination, 97.3% (n=1166; 95% CI 96.1% - 98.2%) of household survey respondents, 97.8% (n=2449) of mobile respondents, and 94.6% (n=390) of online respondents indicated that their families supported vaccination. In response to the BeSD question on knowing where to access vaccination, 99.8% (n=1198; 95% CI 99.2% - 99.9%) of household survey respondents, 96.3% (n=2362) of mobile respondents, and 95.1% (n=393) of online respondents indicated that they knew where to access vaccination.

XSL•FO
RenderX

Figure 1. Diphtheria-pertussis-tetanus vaccination (at least 1 dose) by survey method: Philippines Regions V (Bicol) and XII (SOCCSKSARGEN), January–March 2025. DPT: diphtheria-pertussis-tetanus.

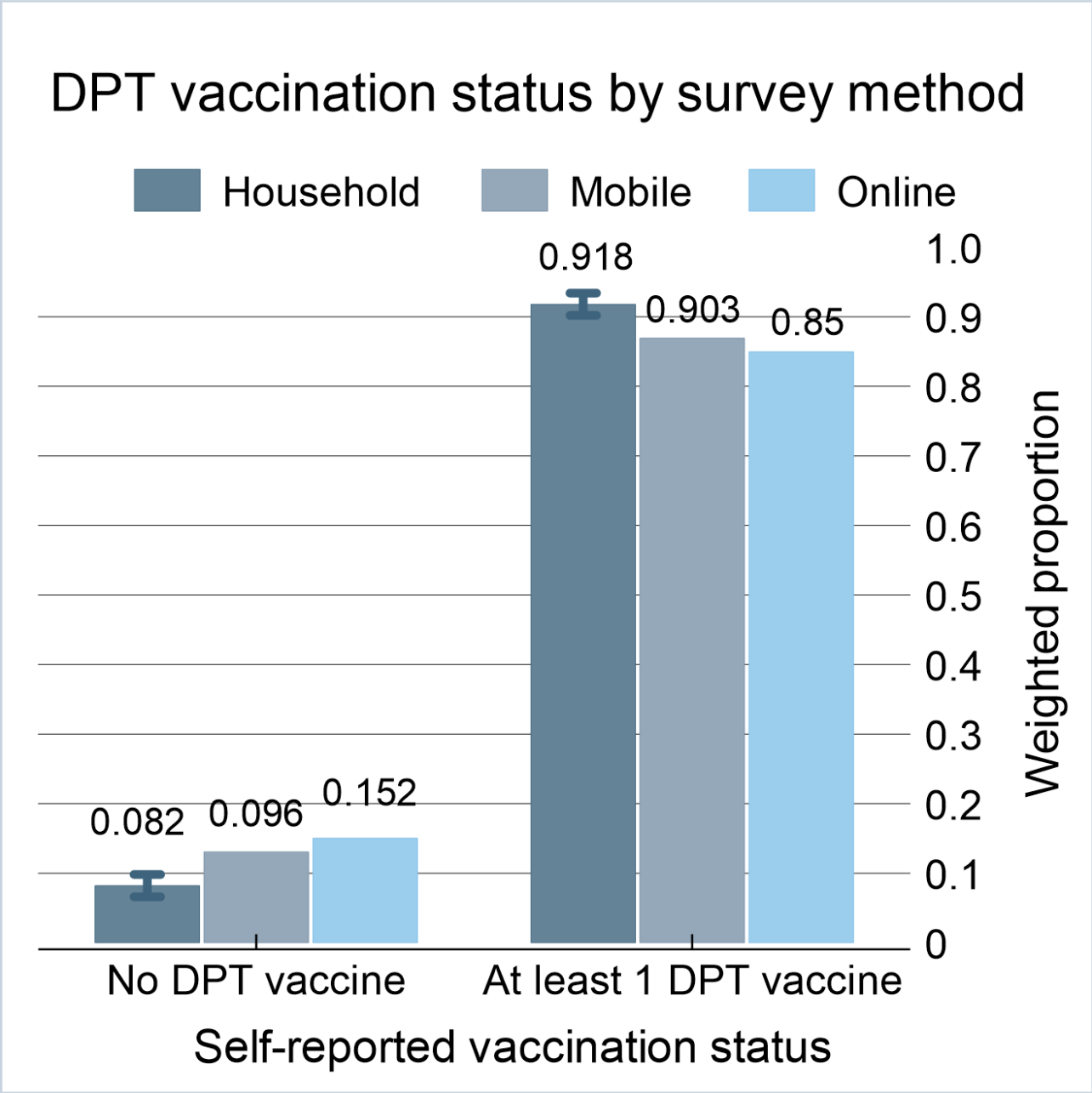

Reported ease of accessing vaccination varied by survey method, with the highest proportion of mobile survey respondents indicating that it was not at all easy to access vaccination (Figure 2). Among household survey respondents, 75.6% (n=899; 95% CI 72.1%‑78.9%) indicated it was very easy to access vaccination, compared with 53.9% (n=1205) of mobile respondents and 72% (n=294) of online respondents. Similarly, reported ease of paying for vaccination varied by survey method, with more than 30% of mobile and online survey respondents indicating it was not at all easy to pay for vaccination, compared with 9.4% (n=112; 95% CI 7.9%‑11.2%) of household respondents. Conversely, 62.1% (n=754; 95% CI 58.3%‑65.7%) of household survey respondents indicated that it was very easy to pay for vaccination, compared with 32.1% (n=744) of mobile respondents and 41.4% (n=169) of online respondents.

The responses to BeSD questions on vaccine importance, vaccine safety, and family support for vaccination did not vary substantially by survey method when the distribution of responses was categorized by DPT vaccination status (Figure S2 in Multimedia Appendix 3).

XSL•FO
RenderX

Figure 2. Behavioral and social drivers of vaccination by survey method: Philippines Regions V (Bicol) and XII (SOCCSKSARGEN), January–March 2025.

XSL•FO
RenderX

Process Description

Among the 2153 responses retained in the analytic dataset for the mobile survey, 348 mobile survey respondents received a quality callback, of whom 117 (33.6%)o JMIR reviewer callbacks indicated that the initial survey response was ineligible or not concordant with the original responses to demographic survey questions. The demographic distribution of the respondents flagged for quality concerns differed significantly from the demographic distributions of those not flagged for quality concerns (Table S2 in Multimedia Appendix 1). However, the BeSD responses for vaccine importance, safety, family support for vaccination, and ease of paying for vaccination did not vary significantly between responses flagged for quality concerns and those that were not flagged for quality concerns (Table S2 in Multimedia Appendix 1).

The financial cost per survey from the implementer's perspective, in nominal US $, was US $2.10 per survey response for the online survey, US $6.93 for the mobile survey, and US $29.38 for the household survey (Table S3 in Multimedia Appendix 1). Personnel time paid labor and per diem were the main cost drivers for the household survey, respondent incentives were the main cost driver for the mobile survey, and advertising and platform costs were the main cost drivers for the online survey.

Discussion

Principal Findings

Our study comparing 3 different methods of collecting BeSD of vaccination data in 2 regions of the Philippines found that respondent characteristics varied by method, with mobile survey respondent demographics and child vaccination status closer to household survey responses than online survey responses. Of particular interest, the proportion of caregivers reporting a child unvaccinated against DPT was 9.7% for the household survey, 8.2% for the mobile survey, and 15% for the online survey, indicating that all 3 survey methods were able to access unvaccinated populations, with online respondents reporting the highest levels of unvaccinated children.

Overall, respondents for each survey method indicated high vaccine demand, with more than 85% of respondents in each survey method indicating that vaccines are very important, very safe, supported by family, and that the site to access vaccination was known. Within these overall trends, the mobile survey responses were more closely reflective of the household responses than the online survey responses. However, responses to questions on ease of accessing and paying for vaccination illustrated more heterogeneity in survey responses and between survey methods, with household survey respondents more likely to indicate that it was very easy to access and pay for a vaccine than mobile and online survey respondents. These differential responses to the BeSD questions on challenges could be due to underlying differences in population characteristics, or they could be attributed to differential effects of social desirability bias between the in-person household survey and the self-administered mobile and online surveys.

The cost per survey response increased threefold from online to mobile and fourfold from mobile to household survey. A quality check reassessed 10% of mobile survey responses through callbacks and identified 37 instances of inconsistent responses; a quality assessment did not identify substantive differences in magnitude or significance for 4 of the core BeSD questions between responses flagged in sensitivity checks and those that were not flagged.

Each survey method offered benefits and drawbacks. The household survey generated responses that could be considered representative of their respective regions, though the costs per response were considerably higher than the other 2 methods. The mobile survey could be conducted at a fraction of the cost of the household survey, and it yielded similar responses to the BeSD questions. However, the demographic distribution of the respondents differed substantially from that of the household survey, and the quality assessment identified nearly one-third of the responses as ineligible, based on region or caregiver status. The online survey was the least expensive to implement and also yielded similar responses to the BeSD questions as the other 2 methods. However, online survey respondent demographics were substantially different from those of household survey respondents, and there was no means to ascertain respondent eligibility by regional residence or caregiver status.

The WHO has recommended that all countries incorporate questions from BeSD validated surveys [26] in routine data collection and recommends that those with low or inequitable coverage of childhood immunization conduct national or subnational surveys every 2 to 3 years [27]. These indicators have been deployed in contexts ranging from human papillomavirus vaccination in Zimbabwe [28] to COVID-19 vaccination in the Philippines [29]. However, the routine implementation of these validated survey questions poses challenges in resource-constrained settings, particularly if rapid data collection is needed following an event that causes a rapid drop in vaccine confidence.

Our work adds to the growing literature that compares mobile and online data collection methods to household surveys in LMICs [30]. A Cochrane review identified app-based data collection methods as noninferior to other methods of data collection [31], but there were no studies included from LMICs. In a systematic review comparing survey methods in LMICs, only 3 of the included studies compared remotely delivered surveys to household surveys, of which 2 focused on infant feeding programs [31,32] and a third examined household poverty levels following economic shocks. To our knowledge, this study is the first comparison of survey methods regarding BeSD of vaccination at the community level in an LMIC.

Compared to mobile and online methods, household surveys are typically the most representative of population demographics, with mobile and online data drawing a relatively lower proportion of older respondents, rural respondents, and those with less than a high school education and less relative wealth [10,33,34]. Our study identified similar trends in representativeness by survey method [35], although the key variable of interest, DPT vaccination status, differed between

XSL•FO
RenderX

data collection methods, indicating that mobile and online surveys are less reliable than household surveys in ascertaining vaccination status.

The literature suggests that the reliability of responses by survey method depends on the type of question asked [36], though mobile surveys have been demonstrated to be noninferior to in-person data collection in survey responsiveness for questions deemed less sensitive [37]. An analysis comparing sequential calling of IVR and computer-assisted telephone interviewing in Bangladesh and Tanzania found high consistency in response to questions on alcohol consumption type and smoking, moderate consistency for history of diabetes or hypertension, and low consistency for routine activities, including physical activity and diet [38]. In a study comparing farmers in India who were surveyed both in person and by phone, significant differences in production were detected due to differential answers by survey method, rather than differential attrition [33]. Our study identifies inconsistencies in demographics by survey method but identifies similarities in perceptions of vaccine importance, safety, family support, and knowledge of where to access vaccination across survey methods.

We found that the financial cost per survey method differed substantially, with the online survey the least expensive and the household survey the most. These findings are consistent with other studies that evaluate costs per survey method between SMS text messaging, IVR, and computer-assisted telephone interviewing [39] and those that examine costs of including mobile phones in existing survey modalities [40]. Our direct comparison by survey method allows decision-makers to assess the preferred survey method in the context of representativeness, responses to core BeSD questions, and financial costs from the implementer's perspective.

To our knowledge, our study represents the first direct comparison of different survey methods for BeSD data. The implications of our study illustrate that while the survey respondents substantially differ by data collection method, responses to the BeSD questions were markedly similar across all 3 methods. These findings point to the potential of mobile and online methods to complement household surveys, particularly in contexts where mobile ownership or internet connectivity is high and there is reported homogeneity in vaccine drivers across subpopulations.

Limitations

This study has at least 5 limitations. First, the survey methods did not use the same sampling frame; thus, it was not possible to assess how the underlying population demographics varied between the 3 sampling frames. The differences in sampling frames limit our ability to assess the extent to which differences between surveys were attributable to the underlying population differences in sampling frames vs the differences in respondents by survey type. We aimed to address this limitation by sampling adjacent barangays for the household and mobile survey methods, and we adjusted for key demographic and socioeconomic factors. In addition, we do not draw conclusions on comparative representativeness by survey type due to this limitation. Second, our data do not permit an assessment of response rates for each data method, as the denominator for the mobile surveys and online data was not available. We report only the number of responses and compare the representativeness of respondents between the survey methods. Third, with the exception of vaccination card verification in the household survey method, the responses to the survey questions could not be independently verified and could have been subject to social desirability bias or recall limitations. To help mitigate this, in-person enumerators assured caregivers that responses were anonymous and, when possible, referred to the child's vaccination card for verification. While the BeSD questions cannot be independently verified, we sought to mitigate any unmeasured social desirability bias through the remote administration methods and by training household survey data collectors in appropriate data collection techniques that mitigate the risk of these biases. In addition, the mobile survey method included a data quality assessment of response accuracy to quantify the proportion of respondents whose responses to demographic questions indicated that they were not eligible by age, region of residence, or being caregivers of children 2 years of age or younger. Fourth, by distributing the mobile survey by BHWs, the mobile survey sampling frame was not structured to reach households that might be missed by household survey methods and could therefore be subject to similar biases in missing transitory or hard-to-reach populations. Furthermore, the mobile survey distribution via BHWs also limits the generalizability of these findings to contexts where trusted community health workers maintain listings of vaccine-eligible households in their catchment areas. While we acknowledge this limitation, we also note that the mobile methods yielded a similar proportion of children unvaccinated against DPT. Finally, we note that the financial costing of survey administration from the implementer's perspective does not include the economic costs, such as the in-kind labor cost of workers whose salaries are covered by the Department of Health, or the systems cost of integrating BeSD data collection into Department of Health data platforms. Furthermore, financial costs presented here are those reported by the implementer on its budgets and may differ from the actual expenses incurred. Similarly to the survey response rates, costs of each survey method are not directly comparable because their representativeness and length differ.

Conclusions

Although respondent demographics and child vaccination status differed between household, mobile, and online data collection methods, respondents consistently indicated high vaccine demand as well as challenges in accessing vaccination and in the cost of accessing vaccination. Selecting the most suitable survey method requires a consideration of trade-offs between methods, including cost, access to technology, and target demographics of caregivers or children. In areas with high mobile phone ownership and strong mobile and internet connectivity, routine mobile and online vaccine demand data collection methods can supplement periodic household surveys.

XSL•FO
RenderX

Acknowledgments

The authors wish to thank Dr Janis Asuncion Bunoan-Macazo, Dr Tony Mounts, Dr Joseph Breese, Dr Nelly Mejia Gonzalez, the data collection team, Barangay Health Workers, and participants involved in this study. The coauthors attest that we did not use generative AI tools for any part of the study, including study design, study question development, data analysis, figure generation, or manuscript drafting.

Funding

This study was funded by the US Centers for Disease Control and Prevention.

Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the US Centers for Disease Control and Prevention.

Data Availability

Data can be made available upon request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey questions and response options for each survey method, mobile survey quality assessment, and total financial cost and financial cost per survey response.
[DOCX File, 55 KB - jmir_v28i1e81059_app1.docx]

Multimedia Appendix 2
Flow diagram of household, mobile, and online survey participants: Philippines Regions V (Bicol) and XII (SOCCSKSARGEN), January–March 2025.
[PNG File, 84 KB - jmir_v28i1e81059_app2.png]

Multimedia Appendix 3
Behavioral and social drivers of vaccination by DPT vaccination status and survey method: Philippines Regions V (Bicol) and XII (SOCCSKSARGEN), January–March 2025.
[PNG File, 283 KB - jmir_v28i1e81059_app3.png]

References

1. Hogan DR, Stevens GA, Hosseinpoor AR, Boerma T. Monitoring universal health coverage within the sustainable development goals: development and baseline data for an index of essential health services. Lancet Glob Health 2018 Feb;6(2):e152-e168. [doi: 10.1016/S2214-109X(17)30472-2] [Medline: 29248365]
2. Brewer NT, Chapman GB, Rothman AJ, Leask J, Kempe A. Increasing vaccination: putting psychological science into action. Psychol Sci Public Interest 2017 Dec;18(3):149-207. [doi: 10.1177/1529100618760521] [Medline: 29611455]
3. Rego RT, Zhukov Y, Reneau KA, et al. Promoting data harmonization to evaluate vaccine hesitancy in LMICs: approach and applications. BMC Med Res Methodol 2023 Nov 24;23(1):278. [doi: 10.1186/s12874-023-02088-z] [Medline: 38001442]
4. Cutts FT, Claquin P, Danovaro-Holliday MC, Rhoda DA. Monitoring vaccination coverage: defining the role of surveys. Vaccine 2016 Jul 29;34(35):4103-4109. [doi: 10.1016/j.vaccine.2016.06.053] [Medline: 27349841]
5. Thomson DR, Bhattarai R, Khanal S, et al. Addressing unintentional exclusion of vulnerable and mobile households in traditional surveys in Kathmandu, Dhaka, and Hanoi: a mixed-methods feasibility study. J Urban Health 2021 Feb;98(1):111-129. [doi: 10.1007/s11524-020-00485-z] [Medline: 33108601]
6. Meyer BD, Mok WKC, Sullivan JX. Household surveys in crisis. J Econ Perspect 2015 Nov 1;29(4):199-226. [doi: 10.1257/jep.29.4.199]
7. Gibson DG, Pereira A, Farrenkopf BA, Labrique AB, Pariyo GW, Hyder AA. Mobile phone surveys for collecting population-level estimates in low- and middle-income countries: a literature review. J Med Internet Res 2017 May 5;19(5):e139. [doi: 10.2196/jmir.7428] [Medline: 28476725]
8. Collins E, Warren S, Lamke C, Contreras I, Henderson S, Rosenbaum M. Representativeness of remote survey methods in LMICs: a cross-national analysis of pandemic-era studies. SSRN Journal 2023 May 31. [doi: 10.2139/ssrn.4582588]
9. Lambrecht I, van Asselt J, Headey D, et al. Can phone surveys be representative in low- and middle-income countries? An application to Myanmar. PLoS One 2023;18(12):e0296292. [doi: 10.1371/journal.pone.0296292] [Medline: 38134041]

XSL•FO
RenderX

10. Greenleaf AR, Gibson DG, Khattar C, Labrique AB, Pariyo GW. Building the evidence base for remote data collection in low- and middle-income countries: comparing reliability and accuracy across survey modalities. J Med Internet Res 2017 May 5;19(5):e140. [doi: 10.2196/jmir.7331] [Medline: 28476728]
11. Gibson DG, Tamrat T, Mehl G. The state of digital interventions for demand generation in low- and middle-income countries: considerations, emerging approaches, and research gaps. Glob Health Sci Pract 2018 Oct 10;6(Suppl 1):S49-S60. [doi: 10.9745/GHSP-D-18-00165] [Medline: 30305339]
12. Behavioural and social drivers of vaccination: tools and practical guidance for achieving high uptake. : World Health Organization; 2022 URL: https://iris.who.int/server/api/core/bitstreams/c9f4a77a-001e-477d-a9d1-b05832820b64/content [accessed 2025-03-14]
13. Jäckle A, Roberts C, Lynn P. Assessing the effect of data collection mode on measurement. Int Statistical Rev 2010 Apr;78(1):3-20. [doi: 10.1111/j.1751-5823.2010.00102.x]
14. Burki T. Zero-dose immunization programme reaches milestone. Lancet Infect Dis 2024 Apr;24(4):e228-e229. [doi: 10.1016/S1473-3099(24)00167-1] [Medline: 38521077]
15. Reducing vaccine hesitancy in the Philippines: findings from a survey experiment. : World Bank; 2021 URL: https://thedocs.worldbank.org/en/doc/9b206c064482a4fbb880ee23d6081d52-0070062021/original/Vaccine-Hesitancy-World-Bank-Policy-Note-September-2021.pdf [accessed 2026-03-14]
16. Kemp S. Digital 2025: the Philippines. DataReportal. URL: https://datareportal.com/reports/digital-2025-philippines [accessed 2025-03-14]
17. 2022 Philippine National Demographic and Health Survey (NDHS): final report. : Philippine Statistics Authority (PSA) and ICF; 2023 URL: https://www.dhsprogram.com/pubs/pdf/FR381/FR381.pdf [accessed 2025-03-14]
18. Digital development dashboard Philippines. : International Telecommunication Union (ITU); 2024 URL: https://www.itu.int/en/ITU-D/Statistics/Documents/DDD/ddd_PHL.pdf [accessed 2025-03-14]
19. Phillips DR. Primary health care in the Philippines: banking on the Barangays? Soc Sci Med 1986;23(10):1105-1117. [doi: 10.1016/0277-9536(86)90269-8] [Medline: 3823975]
20. Bernadas KA. PBBM urges public to support ‘catch-up’ immunization among Bicol kids. Philippine Information Agency. 2024. URL: https://pia.gov.ph/news/pbbm-urges-public-to-support-catch-up-immunization-among-bicol-kids/ [accessed 2025-05-14]
21. Republic act 10173 data privacy act of 2012, Republic of the Philippines Congress of the Philippines Metro Manila Fifteenth Congress Second Regular Session (2013). National Privacy Commission. URL: https://privacy.gov.ph/data-privacy-act/ [accessed 2025-05-14]
22. Distor C, Moon MJ. Artificial intelligence and big data in COVID-19 response: lessons from the KIRA chatbot and data management platform of the Department of Health, Philippines. : Yonsei University; 2022 URL: https://www.researchgate.net/publication/362620886_Artificial_Intelligence_and_Big_Data_in_COVID-19_Response_Lessons_from_the_KIRA_Chatbot_and_Data_Management_Platform_of_the_Department_of_Health_-_Philippines [accessed 2026-03-14] [doi: 10.13140/RG.2.2.36497.51047/1]
23. Wonodi C, Farrenkopf BA. Defining the zero dose child: a comparative analysis of two approaches and their impact on assessing the zero dose burden and vulnerability profiles across 82 low- and middle-income countries. Vaccines (Basel) 2023 Sep 28;11(10):1543. [doi: 10.3390/vaccines11101543] [Medline: 37896946]
24. Stata statistical software: release 16. StataCorp LLC. 2019. URL: https://www.stata.com/stata16/ [accessed 2022-08-16]
25. Baker R, Brick JM, Bates NA, et al. Summary report of the AAPOR Task Force on Non-probability Sampling. J Surv Stat Methodol 2013 Nov 1;1(2):90-143. [doi: 10.1093/jssam/smt008]
26. Wiley KE, Levy D, Shapiro GK, et al. A user-centered approach to developing a new tool measuring the behavioural and social drivers of vaccination. Vaccine (Auckl) 2021 Oct 8;39(42):6283-6290. [doi: 10.1016/j.vaccine.2021.09.007] [Medline: 34538695]
27. World Health Organization. Understanding the behavioural and social drivers of vaccine uptake WHO position paper – May 2022. Wkly Epidemiol Rec 2022;97:209-224 [FREE Full text]
28. Chuma DM, Machacha R, Adjagba AO, January J. Behavioural and social drivers (BeSD) of HPV vaccination in Zimbabwe: a rapid scoping review of literature. Asian Pac J Cancer Prev 2025 Mar 1;26(3):775-783. [doi: 10.31557/APJCP.2025.26.3.775] [Medline: 40156393]
29. Silvestre CJ, Sornillo BJT, Endoma V, et al. Newness, unfamiliarity, and cultural beliefs; social and behavioural barriers to COVID-19 vaccination among the Dumagat Remontado, an Indigenous population in the Philippines. Health Place 2025 May;93:103444. [doi: 10.1016/j.healthplace.2025.103444] [Medline: 40187120]
30. Du X, Wang W, Helena van Velthoven M, et al. mHealth Series: text messaging data collection of infant and young child feeding practice in rural China - a feasibility study. J Glob Health 2013 Dec;3(2):020403. [doi: 10.7189/jogh.03.020403] [Medline: 24363921]
31. Marcano Belisario JS, Jamsek J, Huckvale K, O’Donoghue J, Morrison CP, Car J. Comparison of self-administered survey questionnaire responses collected using mobile apps versus other methods. Cochrane Database Syst Rev 2015 Jul 27;2015(7):MR000042. [doi: 10.1002/14651858.MR000042.pub2] [Medline: 26212714]

XSL•FO
RenderX

32. Li Y, Wang W, van Velthoven MH, et al. Text messaging data collection for monitoring an infant feeding intervention program in rural China: feasibility study. J Med Internet Res 2013 Dec 4;15(12):e269. [doi: 10.2196/jmir.2906] [Medline: 24305514]
33. Anderson E, Lybbert TJ, Shenoy A, Singh R, Stein D. Does survey mode matter? Comparing in-person and phone agricultural surveys in India. J Dev Econ 2024 Jan;166:103199. [doi: 10.1016/j.jdeveco.2023.103199] [Medline: 38164439]
34. Mahfoud Z, Ghandour L, Ghandour B, Mokdad AH, Sibai AM. Cell phone and face-to-face interview responses in population-based surveys. Field methods 2015 Feb;27(1):39-54. [doi: 10.1177/1525822X14540084]
35. Rane MS, Kochhar S, Poehlein E, et al. Determinants and trends of COVID-19 vaccine hesitancy and vaccine uptake in a national cohort of US adults: a longitudinal study. Am J Epidemiol 2022 Mar 24;191(4):570-583. [doi: 10.1093/aje/kwab293] [Medline: 34999751]
36. Ballivian A, Azevedo JP, Durbin W. Using mobile phones for high-frequency data collection. In: Toninelli D, Pinter R, de Pedraz P, editors. Mobile Research Methods: Opportunities and Challenges of Mobile Research Methodologies: Ubiquity Press; 2015:21-39. [doi: 10.5334/bar.c]
37. Chasukwa M, Choko AT, Muthema F, et al. Collecting mortality data via mobile phone surveys: a non-inferiority randomized trial in Malawi. PLoS Glob Public Health 2022;2(8):e0000852. [doi: 10.1371/journal.pgph.0000852] [Medline: 36962430]
38. Pariyo GW, Greenleaf AR, Gibson DG, et al. Does mobile phone survey method matter? Reliability of computer-assisted telephone interviews and interactive voice response non-communicable diseases risk factor surveys in low and middle income countries. PLoS One 2019;14(4):e0214450. [doi: 10.1371/journal.pone.0214450] [Medline: 30969975]
39. Lau CQ, Cronberg A, Marks L, Amaya A. In search of the optimal mode for mobile phone surveys in developing countries. a comparison of IVR, SMS, and CATI in Nigeria. Surv Res Methods 2019;13(3):305-318. [doi: 10.18148/srm/2019.v13i3.7375]
40. Vehovar V, Berzelak N, Lozar Manfreda K. Mobile phones in an environment of competing survey modes: applying metric for evaluation of costs and errors. Soc Sci Comput Rev 2010 Aug;28(3):303-318. [doi: 10.1177/0894439309353035]

Abbreviations

BeSD: behavioral and social drivers
BHW: Barangay Health Worker
CDC: Centers for Disease Control and Prevention
DOH: Department of Health
DPT: diphtheria-pertussis-tetanus
IVR: interactive voice response
LMIC: low- and middle-income country

Edited by A Mavragani; submitted 21.Jul.2025; peer-reviewed by A Greenleaf, S Agha; accepted 31.Dec.2025; published 10.Apr.2026.

Please cite as:
Bonner KE, Hipol M, Sy D, Royono R, Johnson D, del Rosario I, Redfern A, Biel-Romualdo D, Wong MK, Lam E, Kulkarni S, Ward K, Lacson R, Chen MY, Miraflor JM, Pacial DR, Bunoan R, Catan H, Shragai T
Child Vaccination Status and Behavioral and Social Drivers of Vaccination Among Their Caregivers in the Philippines: Cross-Sectional Survey Study Comparison of Household, Mobile, and Online Modes
J Med Internet Res 2026;28:e81059
URL: https://www.jmir.org/2026/1/e81059
doi:10.2196/81059

© Kimberly E Bonner, Mikka Hipol, Dominique Sy, Rivandra Royono, Douglas Johnson, Isabel del Rosario, Alice Redfern, Darahlyn Biel-Romualdo, Man Kai Wong, Eugene Lam, Shibani Kulkarni, Kirsten Ward, Romel Lacson, Meng-Yu Chen, James Matthew Miraflor, Devon Ray Pacial, Rowena Bunoan, Hugo Catan, Talya Shragai. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Illness Experiences in Young and Middle-Aged Patients With Dilated Cardiomyopathy on Chinese Social Media: Qualitative Study

Weijing Chen*, BS; Hanyu Xiao*, BS; Chenchen Feng, BS; Jialin Li, BS; Jinhua Jin, MS; Ruimin Mu, BS

Nursing Department, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, 3 Qingchun East Road, Hangzhou, China

*these authors contributed equally

Corresponding Author:
Jinhua Jin, MS
Nursing Department, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, 3 Qingchun East Road, Hangzhou, China

Abstract

Background: Dilated cardiomyopathy (DCM), characterized by ventricular dilation and systolic dysfunction, has a 10-year survival rate of 25%. The number of young and middle-aged patients with DCM is increasing, with progressive physical limitations and psychosocial distress substantially impairing quality of life. However, illness experiences in this population remain underexplored, particularly in non-Western contexts. Social media platforms such as Zhihu and Weibo offer a novel avenue for exploring patient narratives and peer support.

Objective: A qualitative descriptive study was conducted to explore the multidimensional experiences of young and middle-aged Chinese patients with DCM through the social narratives shared on social media.

Methods: A qualitative reflexive thematic analysis was conducted on 872 questions and responses from Zhihu, a Chinese Q&A platform. Data were extracted using a Python (Python Software Foundation) application programming interface, manually filtered, and analyzed with NVivo 14.0 (Lumivero). Thematic codes were developed iteratively, with researcher reflexivity and team discussions ensuring analytical rigor.

Results: Five themes emerged: (1) bodily control loss (persistent sleep disturbances, severe respiratory compromise, and physical activity limitations), (2) enmeshed in emotional turmoil (anxiety over survival challenges, guilt over burdening family, and heterogeneous responses to death), (3) social and family role disruption (impacts on academic and career paths and family responsibility interruptions), (4) support needs in health care (urgent demand for information on diseases and treatments and desire for continuous care and guidance), and (5) life reconstruction—between disease and life (lifestyle modifications, exploring diversified treatment paths, peer support and experience sharing, and reconstructing the meaning of life).

Conclusions: This study provides an interpretive account of the intertwined physical, psychological, and social experiences of young and middle-aged patients with DCM in China, emphasizing the influence of life-course stage and sociocultural context. Patients' perceptions of family responsibilities, societal role expectations, and death shape their understanding of illness and their processes of meaning reconstruction. These findings underscore the importance of incorporating life-stage and culturally sensitive perspectives in future research and highlight the potential value of tailored psychological support and digital health platforms in improving access to illness-related information and ensuring continuity of care.

(J Med Internet Res 2026;28:e76918) doi:10.2196/76918

KEYWORDS

dilated cardiomyopathy; young adults; qualitative research; social media; nursing

Introduction

Dilated cardiomyopathy (DCM) is a diverse myocardial illness marked by ventricular dilation and reduced cardiac contractility [1,2]. Its pathophysiology is complex, involving genetic, viral, immunological, and metabolic factors. The World Health Organization (WHO) classifies DCM as a significant cardiovascular disorder that is a leading cause of heart failure, malignant arrhythmias, and sudden cardiac death [3,4]. Epidemiological studies show that patients with DCM have a 5-year survival rate of about 50%, which drops further after 10 years [5,6], with an estimated mortality rate of 5.9 per 100,000 people [7].

In addition to poor clinical results, patients with DCM commonly experience recurrent illness exacerbations or symptom deterioration, which are accompanied by significant psychosocial distress, such as sadness, anxiety, and low self-esteem [8]. However, psychosocial assistance within current

XSL•FO
RenderX

care pathways is noticeably insufficient. According to the UK Cardiomyopathy Charity's Cardiomyopathy Care Report (2023), more than half of adult respondents reported difficulties with emotional regulation in the previous 6 months as a result of cardiomyopathy, while only 9% received any form of psychological support during their care [9].

Notably, as heart failure develops at a younger age, so does the affected population, with the prevalence among young and middle-aged individuals increasing by around 50% [10,11]. Young and middle-aged patients with DCM have significant clinical and psychological features. Disease progression and outcomes may differ from those seen in older individuals [12], and this demographic faces more complex work, familial, and societal demands when dealing with chronic illness. Adults with DCM report significantly lower quality of life in the physical, emotional, and social domains, according to a cross-sectional study by Steptoe et al [13], highlighting the need for specialized psychological interventions and encouraging counseling [14]. However, the majority of current research has concentrated on improving clinical treatment approaches and clarifying pathophysiological mechanisms [15,16]. The development of patient-centered and precision care models is hampered by the lack of research on the illness experiences and psychosocial needs of young and middle-aged patients with DCM, despite the fact that psychosocial adaptation has been studied in younger and middle-aged populations with conditions like diabetes and breast cancer [17,18]. A thorough understanding of patient experiences may be hindered by the limitations of traditional data-gathering techniques like interviews, which frequently face issues such as recollection bias and contextual constraints [19].

The rapid development of digital health technologies presents fresh opportunities to close this disparity. In today's society, the internet has emerged as a crucial medium for health communication [20]. These platforms enable patients to interact with others who are dealing with the same medical issues, offering guidance and emotional support that would not be accessible in conventional health care settings. Additionally, they provide a way for patients to express their individual health journeys outside the limitations of formal clinical settings. Nearly 69.2% of China's 1.108 billion internet users as of December 2024 were young and middle-aged people [21]. This group is more likely to use social media sites like Weibo and Zhihu to seek peer support and discuss health-related experiences. About 15% of the content on Zhihu, one of the biggest Q&A sites in China [22], relates to discussions among its 14.2 million subscribers [23,24] and over 100 million monthly active users. Researchers can record a variety of real-time narratives of lived experiences that might not surface in conventional clinical interviews thanks to this realistic, peer-to-peer contact, which offers insightful information about the multifaceted experiences of young and middle-aged persons with DCM.

The concept of "illness experience" is critical for understanding the psychosocial and emotional effects of DCM on patients. The theoretical underpinning for this notion comes from the Engel biopsychosocial model of illness [25], which holds that illness is more than just a biological condition; it is a complex interaction of biological, psychological, and social elements that affect patients' total well-being. This paradigm highlights the necessity of taking into account not only medical symptoms but also how people adapt emotionally and socially to their condition. In this study, illness experience is defined as the multidimensional and dynamic process by which patients perceive, understand, and respond to the physical, emotional, and social demands of living with a chronic illness. This encompasses not only obvious physiological symptoms but also emotional suffering, shifts in social roles, and coping techniques used in reaction to the disease. By using this approach, we hope to gain a better understanding of how patients with DCM deal with the physical, emotional, and social aspects of their illness.

As a result, this study examines conversational posts written by young and middle-aged patients with DCM on the Zhihu platform to learn about their illness experiences. The study uses qualitative research methodologies to explore the physiological, psychological, and social experiences of young and middle-aged patients with DCM through their online narratives.

These findings may serve as a theoretical foundation for the development of targeted therapies and support programs aimed at improving self-management and quality of life in young and middle-aged patients with DCM.

Methods

Research Design

This descriptive qualitative study used media-based naturalistic narratives to capture the illness experiences of patients with DCM, preserving the richness and depth of their perspectives. This approach offers valuable insights into the development of targeted interventions for patients with DCM. The study adhered to the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines (Checklist 1) [26].

Data Collection

On March 30, 2025, we used Python (Python Software Foundation) to collect data on DCM among young and middle-aged adults on Zhihu via the platform's application programming interface. The search included phrase combinations associated with DCM and the target age group, such as "young people + dilated cardiomyopathy," "middle-aged dilated cardiomyopathy," "youth dilated cardiomyopathy," and "young and middle-aged adults dilated cardiomyopathy." To balance data richness with analytical practicality and account for application programming interface limits, we collected only the top 50 threads per search phrase, as determined by Zhihu's default popularity algorithm. To reduce data loss, the number of collected replies for each thread was manually validated against the platform's response count. This procedure yielded 200 question threads with 8727 user responses.

To create a high-quality collection of patient experience narratives, a manual screening process was performed on the initially collected 200 questions and 8727 associated replies using predetermined inclusion and exclusion criteria. Two researchers screened independently, and conflicts were settled by a third reviewer through consensus or arbitration. The inclusion criteria were as follows: (1) confirmation of patient identity, defined as explicit self-reporting of a DCM diagnosis

XSL•FO
RenderX

(eg, "I was diagnosed with dilated cardiomyopathy") [27]—responses describing relatives with DCM were included only if they evolved into detailed first-person accounts of the respondent's own illness experiences; (2) age eligibility, with explicit indication that the respondent was aged 18 to 59 years, with age determined from self-reported information in the posts; (3) content relevance, in which responses were first-person narratives describing personal experiences with DCM, including symptoms, psychological states, and disease-related life impacts; and (4) temporal relevance, defined as responses posted between April 2020 and March 2025, corresponding to the 5 years preceding data collection.

Exclusion criteria included (1) insufficient narrative depth, defined as responses containing fewer than 100 words or content too vague to support thematic analysis; (2) nonexperiential or informational content, such as popular science explanations, reposted medical information, comments from health care professionals, caregivers, or other third parties, as well as responses limited to motivational or general well-wishing content; and (3) commercial or promotional material, including product recommendations, institutional advertisements, embedded promotional links, or text exhibiting explicit marketing intent, as identified by assessing writing purpose, content balance, and the respondent's posting history. Throughout the analytic process, the research team engaged in iterative conversations and reflexive deliberation to refine codes, challenge interpretations, and improve analytic rigor.

Zhihu enables multilingual content, but the platform is largely in Chinese. All search queries were in Chinese, and the screening process was based on Chinese-language proficiency; therefore, all included data consisted of Chinese-language texts. After deleting irrelevant and duplicate questions and responses, the final dataset consisted of 872 postings with valid patient narratives that met all inclusion criteria. This multistage screening method ensured a true representation of the intended patient group while reducing any confounding factors and bias.

Analysis Strategy

Data were examined using reflexive thematic analysis, a qualitative method for identifying shared patterns of meaning throughout a dataset. This method was chosen for its conceptual flexibility and emphasis on researchers' active participation and reflexive posture throughout the coding and theme generation phases [28]. This approach views researchers as active contributors to knowledge development rather than passive spectators. The analysis followed Braun and Clarke's [29] 6-phase framework, which included data familiarization, initial coding, topic generation, theme review, theme definition and naming, and report authoring. To retain semantic nuances, cultural connotations, and contextual coherence, all analyses were carried out in Chinese, the primary language spoken by both the researchers and the study participants. Following the completion of the analysis, illustrative quotations and synthesized findings were translated into English by the first author for reporting purposes. NVivo (version 14.0; Lumivero) was used to help with data management, coding, and theme creation. Throughout the analytic process, the research team engaged in iterative conversations and reflexive deliberation to refine codes, challenge interpretations, and improve analytic rigor.

Analysis Procedure

Researcher WC led and directed all analytic phases, with other members of the research team providing input on crucial analytic conclusions via a series of crystallization talks. This collaborative process facilitated multidisciplinary reflection while also increasing the depth and rigor of the study. In the first phase (data familiarization), WC manually cleaned the information, deleting all private usernames and references to specific geographic regions to protect user privacy. Throughout this procedure, the full dataset was reviewed multiple times to achieve immersion, and analytic reflections were recorded in a reflexive log (Multimedia Appendix 1). The second phase (initial coding) involved 2 researchers independently coding 20% of a randomly selected portion of the data. The coding results were then addressed at the first crystallization conference, where the research team presented preliminary interpretations and perspectives on potential meanings in the data. Building on these conversations, WC used descriptive coding on the remaining data. This method generated 176 initial codes. A second crystallization meeting was held to review these preliminary codes. Following the meeting, WC merged overlapping codes, deleted redundant or conceptually shallow codes, and modified the code set to increase analytic focus. This continual refinement resulted in the final set of 52 codes (Multimedia Appendix 1). In the third phase (theme generation), WC organized the refined codes into larger conceptual patterns and created 5 tentative themes. During the fourth and fifth phases (theme review and definition), the research team held a third crystallization meeting to assess the themes' coherence with respect to the overall dataset and to refine theme boundaries. The themes were examined, improved, and given final names that best matched the data's main narratives (Multimedia Appendix 1).

Reflexivity

Throughout the study, the research team engaged in ongoing reflexive practice to investigate how personal beliefs, professional backgrounds, and underlying assumptions influenced the research process and the interpretation of findings. The team included personnel with a variety of clinical and academic positions, including a clinician specialized in chronic disease management (YP), a head nurse in cardiology (JJ), a cardiology education nurse (CF), and master's-level nursing graduate students. JJ has considerable clinical experience caring for patients with cardiovascular problems, whereas CF specializes in patient communication and cardiology-focused health education, allowing him to have a thorough understanding of patients' views and needs. The principal researcher's overall study program focuses on understanding the lived experiences of people with cardiovascular diseases, with a special emphasis on the bidirectional linkages between psychological and physiological processes. In addition, WC and HX received professional training in reflexive thematic analysis and qualitative data management using NVivo software. These diverse professional responsibilities and experiential backgrounds were identified

XSL•FO
RenderX

as possible influencers on data interpretation. At the same time, they were considered a source of analytic strength, providing complementary views and boosting both the depth and credibility of the study findings [30].

Rigor

This study followed best-practice guidelines proposed by Korstjens and Moser [31] to ensure trustworthiness in qualitative research, with particular attention to the principles of transferability, credibility, and reflexivity. To enhance credibility, a collaborative analytic approach involving multiple researchers was adopted. During the coding process, 2 researchers independently coded approximately 20% of the dataset and subsequently discussed discrepancies until consensus was reached, thereby reducing potential bias arising from a single researcher's interpretation. In addition, regular team meetings were conducted throughout the analytic process to continuously review and refine the coding framework and the development of themes. To improve transparency and traceability of the analytic process, examples of corresponding codes and their thematic categorization are provided in Multimedia Appendix 1. Furthermore, reflexive journals were maintained throughout the study to document researchers' positions and consider how these perspectives might influence data interpretation, thereby further strengthening reflexivity and the methodological rigor of the study.

Ethical Considerations

Research using social media data mining is in a gray area, necessitating not only institutional ethical control but also careful consideration of participant privacy, data use, and the vulnerability of individuals sharing personal health experiences online. Although Zhihu users agree to the platform's privacy policy [32] and understand that their content may be made public, those who upload illness narratives may not foresee their use in academic research. As a result, the gathering and analysis of publicly available data in this study were found to be ethically acceptable, but obtaining individual informed consent was deemed impractical. All data were obtained solely from Zhihu, with no access to nonpublic or restricted content, and all procedures adhered to the platform's terms of service and data use policy. This study was reviewed by the ethics committee of Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, and was determined to be exempt from formal ethical approval, with a waiver of informed consent granted due to the use of publicly available, nonidentifiable data. Several protections were put in place to reduce the risk of harm. To decrease the possibility of indirect identification, usernames were replaced with unique identification codes, and textual quotes were paraphrased or deidentified [33,34]. Raw data were only accessible to the research team and were kept on an encrypted server. These parameters struck a balance between the scientific usefulness of evaluating publicly published illness narratives and respect for participants' privacy, dignity, and vulnerability.

Results

Data Description

A total of 872 questions and responses relevant to young and middle-aged patients with DCM were analyzed. Five themes emerged from the analysis: (1) loss of bodily control, (2) enmeshed in emotional turmoil, (3) social and family role disruption, (4) support needs in health care, and (5) life reconstruction—between disease and life. Textbox 1 lists the 5 core themes and their subthemes. Table 1 lists the number of questions and responses for each topic.

XSL•FO
RenderX

Textbox 1. The 5 core themes and subthemes.

Theme 1: bodily control loss

- Persistent sleep disturbances
- Severe respiratory compromise
- Physical activity limitations

Theme 2: enmeshed in emotional turmoil

- Anxiety over survival challenges
- Guilt over burdening family
- Heterogeneous responses to death

Theme 3: social and family role disruption

- Impacts on academic and career paths
- Family responsibility interruptions

Theme 4: support needs in health care

- Urgent demand for information on diseases and treatments
- Desire for continuous care and guidance

Theme 5: Life reconstruction—between disease and life

- Lifestyle modifications
- Exploring diversified treatment paths
- Peer support and experience sharing
- Reconstructing the meaning of life

Table . Details of the data, including the number of questions and responses for each theme.

Themes	Questions and responses (N=872), n (%)
Bodily control loss	96 (11)
Enmeshed in emotional turmoil	164 (18.8)
Social and family role disruption	110 (12.6)
Support needs in health care	279 (31.9)
Life reconstruction—between disease and life	223 (25.7)

Thematic Analysis

We identified 5 core themes and their subthemes from the data, which highlight the multidimensional challenges and coping strategies experienced by young and middle-aged patients with DCM. These themes encompass physical impairments, psychological distress, social role disruptions, health care support needs, and corresponding adaptive strategies, including life-rebuilding coping behaviors (Textbox 1).

Theme 1: Bodily Control Loss

Patients often describe a rapid decline in physical function due to the disease, marked by irreversible damage to physiological functions and significant limitations in daily activities.

Persistent Sleep Disturbances

Most respondents reported being unable to lie flat due to heart failure symptoms, forcing them to adopt a seated position for sleep, which significantly impaired sleep quality. Some patients rely on sedative medications to achieve brief periods of sleep, while others compare their current physical state to their predisease condition, leading to a strong sense of loss of control:

> *I just can't fall asleep—my heart feels like it's about to burst. It's not like it's skipping beats; it feels more like I'm being smothered. That bubbling, squeezing heartbeat sound—it's terrifying.* [P45]
>
> *I can't get proper rest, but if I take Stilnox, it helps me squeeze in maybe 3 hours of sleep.* [P112]
>
> *What kills me is remembering how it used to be. I used to sleep straight through the night—now I wake up gasping, scrambling to sit upright just to breathe.* [P83]

Severe Respiratory Compromise

Patients with DCM may have a low cardiac ejection fraction and reduced blood pumping ability, potentially leading to

XSL•FO
RenderX

pulmonary congestion and edema, which can cause dyspnea. Severe respiratory deprivation may also trigger a sensation of suffocation, akin to drowning:

> *When my heart acts up, it feels like death knocking. My chest tightens, pounds, and aches.* [P27]
>
> *Lying flat turns me into a wheezing mess—I gasp like I'm underwater, choking on foamy phlegm.* [P156]

Physical Activity Limitations

Respondents generally reported a gradual decline in their ability to perform daily activities due to DCM symptoms. Basic tasks—such as dressing, climbing stairs, and walking—required considerable effort. These functional impairments led some patients to reduce their activity levels, with some even considering staying home as their only option:

> *It takes me 3 separate attempts just to take off a shirt now—every little thing feels like a battle against my own body.* [P09]
>
> *It feels like I've been fast-forwarded through time. Climbing 2 flights of stairs leaves me breathless, like I just ran a marathon. A 1-kilometer run that I once completed easily now feels impossible, and even stairs have become an insurmountable obstacle.* [P191]
>
> *I sweat a lot while sleeping, and I feel very weak. I've been lying in bed at home, afraid to go outside even for a walk.* [P63]

Theme 2: Enmeshed in Emotional Turmoil

The study revealed the multifaceted psychological and emotional experiences that patients face as they adapt to the disease, as outlined in the following subsections. A total of 3 subthemes were identified.

Anxiety Over Survival Challenges

Some patients experience anxiety and helplessness when confronted with the pressures of survival, with financial strain and career limitations being major sources of distress. On one hand, the high cost of treatment contrasts with the family's insufficient financial resources to support the expensive treatment. On the other hand, patients face employment difficulties due to physical decline, which prevents them from securing a stable income:

> *We just can't keep up with the treatment costs—it's draining money we don't have. How long do I even have left? A year? Two? Honestly, I'm terrified I could be gone tomorrow.* [P134]
>
> *I finally started feeling a bit stronger and tried to imagine rebuilding my life. But how? You can't rely on your parents forever. Then it hit me—society can be ruthless. Suddenly, I lost all hope.* [P72]

Guilt Over Burdening Family

Patients with DCM often depend on their family members for care and companionship. However, some middle-aged patients, despite being in their prime years, rely on older adult parents for support, leading to feelings of guilt and self-blame. Some patients even choose to end intimate relationships to avoid burdening their partners, while others express guilt about the cost of treatment. Additionally, concerns regarding genetic risks associated with childbearing were also mentioned:

> *What I feel most sorry about is my parents—I can no longer support them in their old age.* [P38]
>
> *After I got sick, I didn't want to burden my boyfriend, so I ended the relationship that had lasted from graduate school through the start of my career.* [P155]
>
> *Most of my family's income goes toward my treatment. I feel deeply sorry for them.* [P102]
>
> *I know that heart disease may be passed on to the next generation. If this disease is hereditary, I won't have children. Even if I did, I wouldn't be able to take care of them.* [P89]

Heterogeneous Responses to Death

Confronted with a limited life expectancy, patients demonstrate individualized coping strategies and emotional responses. Most respondents reported difficulty accepting the prospect of living only 5 more years, describing feelings of fear and despair while living with what they referred to as the "5-year curse":

> *I never thought I'd get such a severe illness. It's really hard to accept… What should I do now?* [P167]
>
> *I once had a Holter monitor test, and my heart rate dropped to 34 beats per minute while I was sleeping. It felt like I might not wake up. I'm only 20 years old, not yet married, and I still have hope for life. I don't want to die yet.* [P13]

Some patients calmly accept the reality of facing death while maintaining an optimistic attitude:

> *I don't know if I can live for another 10 years. I plan to follow the doctor's advice, take my medication on time, and live happily each day.* [P94]
>
> *I've heard stem cell technology is going to be applied clinically soon. I'll keep fighting to live. Maybe one day medical technology will make a breakthrough!* [P182]

Some patients engage in avoidance behaviors, attempting to challenge the medical diagnosis or refusing medical examinations to escape the prospect of death:

> *After my last scan showed an ejection fraction of 23%—a tiny bump from last time—I quit my meds. Why bother with checkups? Ignorance feels safer than bad news.* [P41]
>
> *Is there a high misdiagnosis rate for DCM? Is there another possibility?*
>
> *[P109]*

Theme 3: Social and Family Role Disruption

Following the onset of disease, younger patients frequently experience disruptions to their educational and early career trajectories, whereas middle-aged patients are more likely to encounter strain or disruption in family roles due to the

XSL•FO
RenderX

substantial responsibilities associated with being primary household providers.

Impacts on Academic and Career Paths

The academic progress of some young patients was significantly hindered by their illness. Some patients who were pursuing a master's degree found that their health condition inevitably became a major factor in their future prospects:

> *Because of this disease, I had to take a year off from school, and since then, my grades have dropped dramatically.* [P55]
>
> *I'm a graduate student, in the prime of my life, but being diagnosed with DCM is a huge catastrophe.* [P112]

Additionally, the impact of their health condition often forces difficult decisions, such as changing jobs or facing unemployment:

> *Regardless of the risk of the operation, it is difficult to successfully match a suitable heart donor. So I chose conservative treatment and stayed at home without looking for a job.* [P29]
>
> *I used to love cosmetics, and before the disease, I was a cosmetic formula designer for a big company, but after the illness, I was let go.* [P173]

Family Responsibility Interruptions

Many middle-aged patients face the dual responsibilities of supporting both their older adult parents and raising young children. However, due to their health conditions, they are often unable to fulfill the obligations associated with these roles, leading to feelings of powerlessness:

> *My children are still young, and my parents need support. How can I just give up?* [P87]
>
> *The more often I go to the hospital, the more afraid I become. I'm young, but I have both elderly parents and young children. I feel overwhelmed, but I have no choice but to face it.* [P142]

Theme 4: Support Needs in Health Care

Illness experiences shared by young and middle-aged patients on the Zhihu platform highlight challenges in navigating health care and obtaining disease-related information. Some patients reported delayed medical consultations attributable to limited disease-related knowledge, whereas others expressed heightened concerns regarding disease information and described difficulties in accessing continuous care and professional guidance.

Urgent Demand for Information on Diseases and Treatments

During the early stages of symptom onset, young and middle-aged patients with DCM often underestimate the seriousness of their condition due to limited disease-related knowledge. After receiving a diagnosis, many actively seek information on Zhihu regarding disease etiology, symptoms, dietary recommendations, and treatment options, with the aim of gaining a clearer understanding of disease progression and management strategies:

> *At 34, my initial symptoms felt like a cold, and I had a cough for 3 months. I thought it was just a flare-up of my chronic pharyngitis. But when I started coughing up pink sputum and couldn't lie down to sleep, I realized something was wrong and went to the hospital.* [P121]
>
> *I'm 28 years old and have been suffering from this disease for 5 years. I hope for a miracle—does anyone know about the latest treatment progress?* [P154]
>
> *I am 49 years old and was diagnosed with DCM last week. What should I do now? Should special dietary restrictions be noted?* [P88]

Desire for Continuous Care and Guidance

Some patients face a lack of continuous care during the process of seeking medical treatment:

> *The local hospital doesn't have doctors who can treat DCM, so we had to rely on larger hospitals.* [P192]
>
> *I am 25 years old and have DCM with stage IV heart failure. At my worst, my ejection fraction was 9%, and I had fluid accumulation in multiple parts of my body. I visited many hospitals, but they all gave me a poor prognosis and advised transferring to another hospital.* [P33]

Theme 5: Life Reconstruction—Between Disease and Life

Amid the shadow of medical prognoses, young and middle-aged patients with DCM actively reconstruct their lives through multidimensional adaptive strategies. These patients adopt various approaches—ranging from strict self-discipline in diet and lifestyle to exploring integrated treatments that combine Western and traditional Chinese medicine—seeking to balance disease management with daily living. Additionally, through mutual support with fellow patients, they share experiences, foster hope, and engage in psychological struggles that reshape their understanding of life's meaning.

Lifestyle Modifications

Confronted with the threat of DCM, patients adjust their daily routines to manage the physical changes caused by the disease. Many quit smoking and drinking, maintain a regular schedule, follow a lighter diet, take medications on time, and engage in moderate exercise:

> *I've quit smoking and cut out alcohol completely—no more late nights either. In bed by 10, up at 7 sharp, meds on schedule. Even added a nightly 30-minute walk after dinner.* [P117]
>
> *Since my diagnosis, I've turned into a health-conscious person—early bedtimes, bland meals, everything. It feels like I'm living like a retiree at 40.* [P62]

Exploring Diversified Treatment Paths

In recent years, integrated treatment approaches that combine Western and traditional Chinese medicine have offered new hope to patients. Young and middle-aged patients with DCM

XSL•FO
RenderX

are open to exploring various therapeutic options, hoping to find an effective cure through diverse treatment pathways:

> *I'm currently taking Western medicine with supplementary traditional Chinese medicine treatment.* [P179]
>
> *A local traditional Chinese medicine doctor recommended making tea from red dates and lotus seeds, although I don't know if it's effective, I thought it couldn't hurt to try.* [P25]
>
> *After I decided to adopt an integrated traditional Chinese and Western medicine treatment, both mitral and tricuspid regurgitation improved significantly.* [P2]

Peer Support and Experience Sharing

Throughout their journey with DCM, patients actively engage with others on platforms like Zhihu, offering life advice, treatment tips, and psychological support. They share real-life experiences to inspire hope in others. Peer support and the exchange of experiences have become vital sources of strength in combating the illness. Emotional support from fellow patients, particularly those of a similar age, is especially significant.

> *If you have heart problems, it's important to eat light meals and stick to a regular routine. Also, make sure to stay warm. I feel discomfort in my heart when the seasons change or when I am in cold wind for too long.* [P141]
>
> *This disease may seem scary, but we need to manage it wisely. Take it seriously by following your doctor's treatment plan while staying positive. Remember, this disease does not always require a heart transplant.* [P166]
>
> *I was diagnosed with this disease when I was 18. Don't give up on treatment, especially if you're around my age. We still have a chance for recovery. Let's live our lives to the fullest!* [P104]

Reconstructing the Meaning of Life

Under the looming shadow of death, many patients reassess the meaning of life. Some attempt to accept their mortality by engaging in meaningful activities during their remaining time. Others, after confronting the brink of life and death, find new meaning in their lives by focusing on family and love, which leads to posttraumatic growth. Notably, some patients shift the focus of their lives from self-actualization to familial responsibility, forming a new ethical contract of "living for others":

> *I don't want to die like this; I want to fight. I want to do meaningful things in the time I have left, to love those who care about me and make the most of every minute.* [P93]
>
> *After experiencing a near-death experience, I want to cherish my life even more and treat my parents with greater respect.* [P132]
>
> *Living well isn't just about me anymore. I now feel like my life belongs to my children, my younger sister, and my parents. I cannot give up easily because they are always there for me.* [P07]

Discussion

Principal Findings

This study presents the first systematic analysis of illness experiences among young and middle-aged patients with DCM. Using reflexive thematic analysis, 5 overarching themes with associated subthemes were identified through the analysis of patient narratives. Themes 1 to 3 capture the complex aspects of patients' illness experiences, stressing issues in the physiological, psychological, and social realms. The physical restrictions mentioned in theme 1 frequently increase the emotional distress presented in theme 2, and changes in social roles, as discussed in theme 3, exacerbate both the physical and psychological burdens. Theme 4 describes patients' challenges in obtaining disease-related information and accessing ongoing treatment. Finally, theme 5 focuses on the active processes by which patients translate their lived experiences and disease-related challenges (as themes 1 - 4) into actual activities and the creation of personal meaning.

The subthemes of theme 1 (bodily control loss) are interconnected. Severe respiratory compromise causes recurrent sleep disturbances because breathing difficulties prevent patients from sleeping flat, thereby reducing sleep quality. Sleep deprivation, in turn, causes weariness and exacerbates respiratory problems, further restricting physical activity. The interplay of these subthemes emphasizes the compounding impact of disease symptoms on patients' daily lives, as well as the physical hardships they bear. Among these multidimensional experiences, abrupt bodily changes are frequently the first and most noticeable effects felt by young and middle-aged individuals. Compared to older patients, these individuals experience a "sense of bodily betrayal"—a rapid deterioration in heart function that causes deep loss of body control and disrupts life goals. This behavior has rarely been described in previous DCM investigations, likely because younger patients compare themselves to healthier peers, exacerbating perceived loss and causing "asynchronous" life-course stress that impedes role adaptability. Charmaz [35] also found physical alienation among adults with chronic disease, accompanied by submission to illness. Comparable experiences occur in adolescent and young adult patients with cancer, who feel alienated due to treatment-related physical changes shaped by social contexts [36]. Body image disturbances in DCM reflect both functional decline and broader psychosocial consequences. Health care providers should promote body acceptance in clinical or community settings by conducting body image–focused psychological screenings and offering targeted interventions. Cognitive behavioral therapy has been shown to improve self-acceptance and alleviate distress in people with chronic illnesses [37].

When the body is no longer reliable, its impact often extends to the patient's existing social roles and responsibilities, leading to deeper psychological distress. The twin demands of high treatment expenses and employment obstacles frequently intensify anxiety, as evidenced by Abidan's case study of young

XSL•FO
RenderX

and middle-aged patients with DCM [38]. Patients frequently experience anxiety as a result of their failure to fulfill tasks as economic providers or caregivers, which causes feelings of guilt for "burdening" their family members. This ethical quandary has also been observed in investigations of patients receiving maintenance hemodialysis and survivors of stroke in China [39,40]. Patients' feelings of shame are ingrained in the familial and societal institutions they inhabit. While the family plays a significant role in chronic disease care across cultures [41], its specific expressions and moral consequences differ. In the Chinese setting, filial piety is fundamentally normative, requiring adult children to provide financial, emotional, and end-of-life care for their parents, and failing to do so may be viewed as "unfilial" [42,43]. Rigid expectations about the "breadwinner" role exacerbate shame in young and middle-aged adults. In this study, some patients even ended personal connections to alleviate the perceived burden on their families, demonstrating "ethically driven guilt" aimed at collective family well-being [44]. These findings also emphasize how the cultural environment influences the emotional experiences of patients with chronic disease [45]. Accordingly, nursing interventions should respect cultural differences, take into account family relational networks, and give tailored, culturally sensitive assistance. Family gatherings, intergenerational communication, and family life planning might help patients renegotiate role expectations and reduce guilt and survival anxiety.

Under prolonged stress, patients inevitably confront the issue of death. Patients who are stressed for an extended period of time are inevitably confronted with the possibility of death. The majority express acute terror, while a minority remain optimistic, hoping for medical breakthroughs. Threatening medical statistics, such as "5-year survival rates," are sometimes viewed as a "countdown to life," which heightens anxiety. In traditional cultural contexts, some patients adopt a fatalistic attitude, which may impact treatment adherence and psychological adjustment [46]. A few people use negative coping mechanisms, like denial or avoidance. These disparate views, driven by fear and avoidance, reflect both a lack of scientific understanding and an insufficient focus in clinical practice on the death-related concerns of young and middle-aged patients with DCM [47]. Without adequate emotional support and opportunities for discussion, negative cognitions can be reinforced by recurrent symptom experiences, resulting in a "poor prognosis—fear of death" loop. Therefore, incorporating assessments of patients' views on death into routine care is critical for understanding patients' perspectives and correcting cognitive biases. Developing multimedia-based death education for young and middle-aged patients with DCM may improve their psychological well-being [48].

It is vital to highlight that patients' illness experiences vary greatly depending on their life stage. While a group examination of young and middle-aged patients with DCM can indicate common characteristics, social roles and psychosocial obstacles vary by developmental stage. Middle-aged patients frequently have family, economic, and caregiving duties, with disease-related disruptions to job and financial stability causing shame and social role crises. Young patients, on the other hand, face abrupt pauses in their education, careers, or personal relationships, as well as physical limitations that impair their self-esteem and future expectations. Tutelman et al [49] observed that adolescents and young adults are at a key developmental stage, and life-threatening illnesses cause distinct psychological demands that necessitate tailored interventions. Young and middle-aged patients with DCM, unlike adolescent and young adult patients with cancer, do not receive adequate psychological assistance for relationship and social role reconstruction, putting them at risk of social isolation and developmental suffering [50,51]. Evidence linking social integration in adults to all-cause and cardiovascular mortality [52] emphasizes the significance of taking life stage into account when treating psychosocial needs. Differentiated support can help people retain social participation and role continuity, which improves their quality of life. This study classified patients generally into young and middle-aged groups; further age stratification may reveal significant differences.

Furthermore, throughout medical therapy, young and middle-aged patients with DCM frequently report difficulties in obtaining adequate disease information and continuity of care, indicating a lack of disease awareness and difficulties in coping with the unpredictability of a sudden chronic condition. The diagnosis disrupts established life plans, increasing susceptibility to symptom variations and treatment outcomes. Uncertainty about disease progression and prognosis contributes to increased anxiety and a sense of loss of control. Our findings show that patients actively seek disease-related information to better understand their condition and restore a sense of control, which is especially important after discharge due to long-term self-monitoring and limited access to professional help. Digital health solutions have developed as instruments for addressing health care system issues and increasing universal health coverage [53]. Future interventions should assure ongoing delivery of intelligible, evidence-based information via social media content, supplemented by remote consultations or regular follow-ups, to assist patients in developing a stable understanding of their disease while reducing fear and feelings of powerlessness.

Despite numerous hurdles, DCM does not disrupt life's order but rather encourages the construction of a new one [54]. Patients rebuild their lives by changing their behaviors, researching treatment choices, and seeking peer support, thereby redefining their sense of purpose and exhibiting resilience. Young and middle-aged patients frequently reduce anxiety by sharing their experiences on the Zhihu platform, likely due to the limited options for ongoing communication and emotional support in official health care settings. Peer support therapies have been proven to improve heart function markers in middle-aged patients with heart failure [55], implying that online peer groups may offer both emotional support and practical guidance. Familial duties also influence life reconstruction, with many patients changing their focus away from personal development and toward helping and safeguarding family members. This reflects the family-centered value orientation in Chinese culture [56], contrasting with the individualistic goals common in Western societies [57]. These findings highlight the need for culturally sensitive psychological and nursing

XSL•FO
RenderX

interventions that help patients balance disease management with life responsibilities.

Strengths and Limitations

This is the first qualitative study to examine the illness experiences of young and middle-aged patients with DCM, filling a knowledge gap and providing insights into their multidimensional experiences spanning physiological, psychological, and role-adjustment dimensions. Our findings expand the biomedical focus of DCM research, offering health care providers a deeper understanding of patient needs and enhancing the effectiveness of nursing care. Unlike previous studies, this research highlights patients' developmental life stages and cultural context, addressing the lack of psychosocial research in non-Western populations.

Nonetheless, the study contains drawbacks. Data were gathered from a single social media network, which may have biased the results toward technologically literate and emotionally involved consumers. In addition, patients were classified generically as "young" or "middle-aged," with no tighter age stratification, thus limiting the examination of within-group variations. Furthermore, in this study, fewer individuals reported the duration of their disease. Instead, their accounts focused primarily on their experiences during the illness. As a result, the absence of accurate information about the length of the illness may hinder a thorough understanding of how disease progression affects patients' lives.

Recommendations for Further Research

Future research could expand in several directions. First, similar qualitative studies across diverse health care settings and cultural contexts are needed to deepen the understanding of the illness experiences and challenges faced by young and middle-aged patients. Second, more detailed age stratification or developmental stage frameworks should be adopted to collect more comprehensive data on factors such as the duration of the disease, in order to explore the similarities and differences in psychological and social challenges, role conflicts, and the construction of life meaning between young and middle-aged patients. Third, longitudinal qualitative designs could capture dynamic changes in patients' cognition, emotions, and coping strategies throughout disease progression. Finally, given the prominence of death in patients' narratives in this study, future research could focus on young and middle-aged patients' understanding of death, providing an empirical basis for refining theoretical frameworks.

Implications and Conclusion

This study provides a qualitative analysis of the multidimensional illness experiences of young and middle-aged patients with DCM in China. Findings indicate that physical decline, psychological distress, and social role crises are closely intertwined, shaped by patients' life stages and cultural context. Family responsibilities, societal role expectations, and perceptions of death collectively influence patients' understanding of their illness and the reconstruction of life meaning. Theoretically, the study highlights the importance of examining the experiences of young and middle-aged patients with DCM within a biopsychosocial-cultural framework, offering a foundation for future research, supportive interventions, and cross-cultural chronic disease management. Practically, clinical interventions should account for patients' developmental stages and cultural contexts, provide tailored psychological support, and leverage digital health platforms to enhance comprehension of disease information and ensure continuity of care.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT (Generative AI Delegation Taxonomy; 2025), the following tasks were delegated to GAI tools under full human supervision: proofreading, editing, and translation. The GAI tool used was ChatGPT (GPT-5.2). Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Critical revision of the manuscript: JL, RM, CF
Data collection: WC, HX, JJ
Drafting: WC, RM, CF
Drafting and editing of the manuscript: WC, HX, JJ
Study design: WC, HX, JJ, JL, RM, CF
Qualitative analysis: WC, HX, JJ

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Analysis procedure record.
[PDF File, 152 KB - jmir_v28i1e76918_app1.pdf]

Checklist 1
COREQ checklist.
[PDF File, 106 KB - jmir_v28i1e76918_app2.pdf]

References

1. Chinese Society of Cardiology. Chinese guidelines for the diagnosis and treatment of dilated cardiomyopathy [Article in Chinese]. J Clin Cardiol 2018;34(5):421-434. [doi: 10.13201/j.issn.1001-1439.2018.05.001]
2. Weintraub RG, Semsarian C, Macdonald P. Dilated cardiomyopathy. Lancet 2017 Jul 22;390(10092):400-414. [doi: 10.1016/S0140-6736(16)31713-5] [Medline: 28190577]
3. Richardson P, McKenna W, Bristow M. Report of the 1995 World Health Organization/International Society and Federation of Cardiology Task Force on the definition and classification of cardiomyopathies. Circulation 1996 Mar;93(5):841-842. [doi: 10.1161/01.CIR.93.5.841] [Medline: 8598070]
4. Qiu H, Sun Y, Wang X, et al. Lamin A/C deficiency-mediated ROS elevation contributes to pathogenic phenotypes of dilated cardiomyopathy in iPSC model. Nat Commun 2024 Aug 14;15(1):7000. [doi: 10.1038/s41467-024-51318-5] [Medline: 39143095]
5. Tayal U, Ware JS, Lakdawala NK, Heymans S, Prasad SK. Understanding the genetics of adult-onset dilated cardiomyopathy: what a clinician needs to know. Eur Heart J 2021 Jun 21;42(24):2384-2396. [doi: 10.1093/eurheartj/ehab286] [Medline: 34153989]
6. Verdonschot JAJ, Hazebroek MR, Krapels IPC, et al. Implications of genetic testing in dilated cardiomyopathy. Circ Genom Precis Med 2020 Oct;13(5):476-487. [doi: 10.1161/CIRCGEN.120.003031] [Medline: 32880476]
7. Lozano R, Naghavi M, Foreman K, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 2012 Dec 15;380(9859):2095-2128. [doi: 10.1016/S0140-6736(12)61728-0] [Medline: 23245604]
8. Yu J, Li C. Evaluation of the effect of psychological nursing intervention on patients with dilated cardiomyopathy [Article in Chinese]. Chin Rural Health Ser Admin 2012;32(2):151-152. [doi: 10.19955/j.cnki.1005-5916.2012.02.020]
9. The State of Cardiomyopathy Care: National Cardiomyopathy Report 2023. : Cardiomyopathy UK; 2023 URL: https://www.cardiomyopathy.org/sites/default/files/State%20of%20cardiomyopathy%20care%202022%20survey%20report.pdf [accessed 2026-04-17]
10. Andersson C, Vasan RS. Epidemiology of cardiovascular disease in young individuals. Nat Rev Cardiol 2018 Apr;15(4):230-240. [doi: 10.1038/nrcardio.2017.154] [Medline: 29022571]
11. Groenewegen A, Rutten FH, Mosterd A, Hoes AW. Epidemiology of heart failure. Eur J Heart Fail 2020 Aug;22(8):1342-1356. [doi: 10.1002/ejhf.1858] [Medline: 32483830]
12. Xie R, Yang P, Liu J, et al. Lanthanide-functionalized metal-organic frameworks based ratiometric fluorescent sensor array for identification and determination of antibiotics. Talanta 2021 Aug 15;231:122366. [doi: 10.1016/j.talanta.2021.122366] [Medline: 33965031]
13. Steptoe A, Mohabir A, Mahon NG, McKenna WJ. Health related quality of life and psychological wellbeing in patients with dilated cardiomyopathy. Heart 2000 Jun;83(6):645-650. [doi: 10.1136/heart.83.6.645] [Medline: 10814621]
14. Thomas SA, Friedmann E, Khatta M, Cook LK, Lann AL. Depression in patients with heart failure: physiologic effects, incidence, and relation to mortality. AACN Clin Issues 2003 Feb;14(1):3-12. [doi: 10.1097/00044067-200302000-00002] [Medline: 12574698]
15. Zhang R, Ma S, Chen C. Study on the influencing factors of psychosocial adaptation level in patients with type 2 diabetes [Article in Chinese]. Med Soc 2020;33(7):86-89. [doi: 10.13723/j.yxysh.2020.07.020]
16. Ruiqin Z, Cuiyong S, Huimin G, Min G, Lanqing G. Correlation between family resilience, social support and psychosocial adjustment for young breast cancer patients [Article in Chinese]. J Nurs Rehabil 2021;20(4):10-15 [FREE Full text]
17. Xuan H, Fan C, Bai X, et al. Protein disulfide isomerase involvement in dilated cardiomyopathy caused by filamin C deficiency in male mice. J Cell Mol Med 2025 Mar;29(6):e70493. [doi: 10.1111/jcmm.70493] [Medline: 40099936]
18. Altman NL, Gill EA, Kahwash R, et al. Heart rate reduction is associated with reverse left ventricular remodeling and mechanism-specific molecular phenotypes in dilated cardiomyopathy. Circ Heart Fail 2025 Apr;18(4):e012484. [doi: 10.1161/CIRCHEARTFAILURE.124.012484] [Medline: 40052260]
19. Mayfield A. A commanders strategy for social media. : Indian Strategic Knowledge Online; 2008 URL: https://indianstrategicknowledgeonline.com/web/mayfield_strat_for_soc_media.pdf [accessed 2026-04-17]
20. The 50th CNNIC Statistical Report on China's Internet Development. : China Internet Network Information Center; 2022 URL: https://www.21cloudbox.com/statistical-report-on-china-internet.html [accessed 2026-04-30]

XSL•FO
RenderX

21. Liyang H. The 55th Statistical Report on China's Internet Development in Beijing [Report in Chinese]. : China Internet Network Information Center; 2025 URL: https://cnnic.cn/n4/2025/0117/c208-11228.html [accessed 2026-04-30]
22. Li P, Xu L, Tang T, Wu X, Huang C. Users' willingness to share health information in a social question-and-answer community: cross-sectional survey in China. JMIR Med Inform 2021 Mar 30;9(3):e26265. [doi: 10.2196/26265] [Medline: 33783364]
23. Zhihu Inc. reports unaudited fourth quarter and fiscal year 2023 financial results. : Zhihu Inc; 2024 Mar 26 URL: https://tools.eurolandir.com/tools/PressReleases/GetPressRelease/?ID=7363825&lang=en-GB&companycode=services [accessed 2026-05-01]
24. Zhihu Inc. reports unaudited fourth quarter and fiscal year 2024 financial results. : Zhihu Inc; 2025 Mar 26 URL: https://ir.zhihu.com/media/dygpwalz/4q-2024.pdf [accessed 2026-04-17]
25. Engel GL. The need for a new medical model: a challenge for biomedicine. Science 1977 Apr 8;196(4286):129-136. [doi: 10.1126/science.847460] [Medline: 847460]
26. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
27. Chun L. Re-partitioning population age group and its implications [Article in Chinese]. Popul Res 2017;41(5):16-25 [FREE Full text]
28. Braun V, Clarke V. Thematic Analysis: A Practical Guide: SAGE Publications; 2021. URL: https://in.sagepub.com/en-in/ind/thematic-analysis/book248481 [accessed 2026-04-17]
29. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Aug 8;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
30. Noble H, Smith J. Issues of validity and reliability in qualitative research. Evid Based Nurs 2015 Apr;18(2):34-35. [doi: 10.1136/eb-2015-102054] [Medline: 25653237]
31. Korstjens I, Moser A. Series: practical guidance to qualitative research. Part 4: trustworthiness and publishing. Eur J Gen Pract 2018 Dec;24(1):120-124. [doi: 10.1080/13814788.2017.1375092] [Medline: 29202616]
32. Zhihu agreement [Web Page in Chinese]. Zhihu. URL: https://www.zhihu.com/term/zhihu-terms [accessed 2026-04-27]
33. McDermott E, Roen K, Piela A. Hard-to-reach youth online: methodological advances in self-harm research. Sex Res Soc Policy 2013 Jun;10(2):125-134. [doi: 10.1007/s13178-012-0108-z]
34. Reagle J. Disguising Reddit sources and the efficacy of ethical research. Ethics Inf Technol 2022;24(3):41. [doi: 10.1007/s10676-022-09663-w] [Medline: 36105629]
35. Charmaz K. The body, identity, and self: adapting to impairment. Sociol Q 1995 Sep;36(4):657-680. [doi: 10.1111/j.1533-8525.1995.tb00459.x]
36. Moore JB, Canzona MR, Puccinelli-Ortega N, et al. A qualitative assessment of body image in adolescents and young adults (AYAs) with cancer. Psychooncology 2021 Apr;30(4):614-622. [doi: 10.1002/pon.5610] [Medline: 33275802]
37. Pearl RL, Wadden TA, Bach C, et al. Effects of a cognitive-behavioral intervention targeting weight stigma: a randomized controlled trial. J Consult Clin Psychol 2020 May;88(5):470-480. [doi: 10.1037/ccp0000480] [Medline: 31971410]
38. Abulimiti A, Nishitani-Yokoyama M, Kato T, et al. The importance of psychosocial support in youth dilated cardiomyopathy patient who underwent cardiac rehabilitation. J Cardiol Cases 2023 Jul;28(1):44-48. [doi: 10.1016/j.jccase.2023.03.004] [Medline: 37360831]
39. Xie C, Li L, Li Y. "Alive Day is the Day": a qualitative study of experiences of learned helplessness in maintenance haemodialysis patients. Risk Manag Healthc Policy 2023;16(231-245):231-245. [doi: 10.2147/RMHP.S401205] [Medline: 36815199]
40. Chen J, Li Y, Liu J, et al. Symptom clusters of early-stage poststroke depression: a mixed-methods study. Nurs Open 2021 Sep;8(5):2488-2497. [doi: 10.1002/nop2.772] [Medline: 33471441]
41. Gilliss CL, Pan W, Davis LL. Family involvement in adult chronic disease care: reviewing the systematic reviews. J Fam Nurs 2019 Feb;25(1):3-27. [doi: 10.1177/1074840718822365] [Medline: 30755110]
42. Rongping Y, Rong G, Anni H, Xiaohua W, Yuan X. Influence of filial piety values on care-giving burden among caregivers of patients with chronic heart failure [Article in Chinese]. Chin J Modern Nurs 2019;25(22):2847-2850. [doi: 10.3760/cma.j.issn.1674-2907.2019.22.017]
43. Qi X. Remaking Families in Contemporary China: Oxford University Press; 2021.
44. Shuang W. An analysis of the differences between Chinese and American family cultures from the perspective of collectivism and individualism [Article in Chinese]. Abil Wisdom 2017;2:190-191 [FREE Full text]
45. Li L, Sun J, Wu H. Stigma in patients with chronic diseases: a review [Article in Chinese]. Nurs J Chin PLA 2020;37(3):75-78 [FREE Full text]
46. Li M, Liu W, Mi S, et al. Clinical and echocardiographic risk factors of adverse outcomes in young patients with dilated cardiomyopathy. Cardiovasc Ther 2025;2025:2122089. [doi: 10.1155/cdr/2122089] [Medline: 40225593]
47. Xiaohui J, Xinxin D, Ming X, Lingjuan M, Yajuan P, Lei Y. Status quo and influencing factors of death anxiety in middle-aged and young adults with chronic heart failure [Article in Chinese]. Chin J Modern Nurs 2024;30(18):2423-2427. [doi: 10.3760/cma.j.cn115682-20240119-00366]

XSL•FO
RenderX

48. Tieman J, Miller-Lewis L, Rawlings D, Parker D, Sanderson C. The contribution of a MOOC to community discussions around death and dying. BMC Palliat Care 2018 Feb 20;17(1):31. [doi: 10.1186/s12904-018-0287-3] [Medline: 29458347]
49. Tutelman PR, Moran C, Beattie SM, et al. Acceptability, feasibility and preliminary effects of an online group psychotherapy intervention for adolescents and young adults with cancer. Psychooncology 2024 Mar;33(3):e6335. [doi: 10.1002/pon.6335] [Medline: 38526517]
50. Canzona MR, Victorson DE, Murphy K, et al. A conceptual model of fertility concerns among adolescents and young adults with cancer. Psychooncology 2021 Aug;30(8):1383-1392. [doi: 10.1002/pon.5695] [Medline: 33843104]
51. Schilstra CE, Fardell JE, Burns MA, et al. Determinants of social functioning among adolescents and young adults with cancer: a systematic review. Psychooncology 2021 Oct;30(10):1626-1642. [doi: 10.1002/pon.5740] [Medline: 34081823]
52. Tan J, Wang Y. Social integration, social support, and all-cause, cardiovascular disease and cause-specific mortality: a prospective cohort study. Int J Environ Res Public Health 2019 Apr 27;16(9):1498. [doi: 10.3390/ijerph16091498] [Medline: 31035622]
53. Gong E, Wang H, Zhu W, et al. Bridging the digital divide to promote prevention and control of non-communicable diseases for all in China and beyond. BMJ 2024 Oct 18;387:e076768. [doi: 10.1136/bmj-2023-076768] [Medline: 39424328]
54. Ide-Okochi A, Yamazaki Y, Tadaka E, Fujimura K, Kusunaga T. Illness experience of adults with cervical spinal cord injury in Japan: a qualitative investigation. BMC Public Health 2013 Jan 24;13(69):69. [doi: 10.1186/1471-2458-13-69] [Medline: 23347726]
55. Chen T, Niu Y, Chen P, Zhou M, Chen Z. Effect of peer support on psychological status and quality of life of middle-aged patients with heart failure [Article in Chinese]. J Nurs Adm 2024;24(1):26-31 [FREE Full text]
56. Eshleman JR. Introduction to the Family [Book in Chinese]: China Social Sciences Press; 1991.
57. Wang L. Autonomy in medical decision-making under the background of Confucian culture [Article in Chinese]. Med Soc 2022;35(1). [doi: 10.13723/j.yxysh.2022.01.006]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
DCM: dilated cardiomyopathy
WHO: World Health Organization

Edited by A Stone; submitted 04.May.2025; peer-reviewed by B Jiménez-Alonso, Y Luo; revised version received 01.Apr.2026; accepted 07.Apr.2026; published 05.May.2026.

Please cite as:
Chen W, Xiao H, Feng C, Li J, Jin J, Mu R
Illness Experiences in Young and Middle-Aged Patients With Dilated Cardiomyopathy on Chinese Social Media: Qualitative Study
J Med Internet Res 2026;28:e76918
URL: https://www.jmir.org/2026/1/e76918
doi:10.2196/76918

© Weijing Chen, Hanyu Xiao, Chenchen Feng, Jialin Li, Jinhua Jin, Ruimin Mu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Sociodemographic Paradoxes and Enrollment Differences in In-Person Versus Online Recruitment to a Mobile Health Smoking Cessation Intervention for Food-Insecure Adults: Secondary Analysis of a Randomized Controlled Trial

Charles E Hoogland[1], PhD; Steven K Sutton[2,3,4], PhD; Sarah R Jones[1], MPH; Bence Nagy[1], MPH; Samuel J Brockway[1], MS; David Himmelgreen[5], PhD; Thomas Mantz[6], BA; Michael S Businelle[7,8], PhD; Ya-Chen Tina Shih[9], PhD; Jennifer I Vidrine[1,3,4*], PhD; Damon J Vidrine[1,3*], DrPH

[1]Department of Health Outcomes & Behavior, Moffitt Cancer Center, 12902 USF Magnolia Drive, Tampa, FL, United States

[2]Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, FL, United States

[3]Department of Oncologic Sciences, Morsani College of Medicine, University of South Florida, Tampa, FL, United States

[4]Department of Psychology, University of South Florida, Tampa, FL, United States

[5]Department of Anthropology, Center for the Advancement of Food Security and Healthy Communities, University of South Florida, Tampa, FL, United States

[6]Feeding Tampa Bay, Tampa, FL, United States

[7]TSET Health Promotion Research Center, Stephenson Cancer Center, University of Oklahoma Health Sciences Center, Oklahoma City, OK, United States

[8]Department of Family and Preventive Medicine, University of Oklahoma Health Sciences Center, Oklahoma City, OK, United States

[9]Jonsson Comprehensive Cancer Center, Program in Cancer Health Economics Research, Department of Radiation Oncology, School of Medicine, UCLA Jonsson Comprehensive Cancer Center, Los Angeles, CA, United States

[*]these authors contributed equally

Corresponding Author:
Damon J Vidrine, DrPH
Department of Health Outcomes & Behavior, Moffitt Cancer Center, 12902 USF Magnolia Drive, Tampa, FL, United States

Abstract

Background: Little is known about (1) sociodemographic, psychosocial, or smoking-related differences among individuals recruited to smoking cessation randomized controlled trials (RCTs) using in-person versus online recruitment methods or (2) the relative speed of recruitment using these 2 approaches. This secondary analysis is the first to examine these comparisons in a smoking cessation RCT for people experiencing food insecurity, a vulnerable special population for whom quitting is especially urgent.

Objective: To compare (1) baseline sociodemographic, smoking-related, and psychosocial characteristics; and (2) screening, eligibility, and enrollment rates of in-person versus online recruits to a smoking cessation RCT for people experiencing food insecurity.

Methods: Participants completed a brief eligibility questionnaire and a baseline assessment via tablet (in person) or personal electronic device (after clicking an online advertisement). Eligibility required past-30-day food aid use, smoking ≥5 cigarettes per day, and willingness to attempt quitting within 7 days post enrollment. Responses were compared using chi-squared and Fisher exact tests (categorical variables) and 2-tailed *t* tests (continuous variables).

Results: Enrollees recruited online endorsed greater food insecurity (mean 4.5, SD 1.9 vs mean 3.0, SD 2.3; $P<.001$) and were more likely to be educated beyond high school or equivalent (69% vs 49%; $P<.001$), have household income of US $20,000 or more (46% vs 36%; $P=.03$), and be non-Hispanic White (77% vs 50%; $P<.001$). Online recruits indicated lower motivation to quit smoking (Contemplation Ladder; mean 7.2, SD 2.4 vs mean 8.0, SD 2.8; $P<.001$) and smoking cessation self-efficacy (mean 20.5, SD 8.0 vs mean 23.2, SD 8.6; $P<.001$). Online recruits also reported lower subjective social status (mean 4.6, SD 2.0 vs mean 5.9, SD 2.2; $P<.001$), greater financial strain (mean 17.9, SD 6.3 vs mean 16.2, SD 6.6; $P=.004$), more depressive symptoms (mean 8.6, SD 6.3 vs mean 7.4, SD 6.1; $P=.04$), greater loneliness (mean 6.0, SD 2.1 vs mean 5.2, SD 2.0; $P<.001$), less resilience (mean 19.5, SD 5.1 vs mean 20.5, SD 4.3; $P=.02$), less alcohol misuse (27% vs 37%; $P=.02$), and more past-30-day cannabis use (25% vs 15%; $P=.01$). Enrollment rates were higher online (64.8 per month; n=324) than in-person (7.7 per month; n=178). Although screened eligible at similar rates whether recruited online or in person (79% vs 75%; $P=.10$), eligible online individuals were more likely to enroll (71% vs 49%; $P<.001$).

XSL•FO
RenderX

Conclusions: This study is the first to compare baseline participant characteristics by recruitment method (in person vs online) in a cessation RCT for people experiencing food insecurity and to evaluate the relative pace of recruitment via those methods. Online and in-person recruits were demographically and psychosocially distinct, and online recruitment was associated with faster accrual than in-person recruitment. These findings inform recruitment strategies for cessation interventions, especially those targeting food-insecure individuals.

Trial Registration: ClinicalTrials.gov NCT05004662; https://clinicaltrials.gov/study/NCT05004662

International Registered Report Identifier (IRRID): RR2-10.1186/s12889-022-12840-7

(J Med Internet Res 2026;28:e80530) doi:10.2196/80530

KEYWORDS

recruitment strategies; smoking cessation; food insecurity; food assistance; randomized controlled trial; social status

Introduction

Background and Rationale

Cigarette smoking remains the leading cause of preventable disease and death in the United States [1]. Although the prevalence of smoking has been declining for decades, smoking rates have declined more slowly in socioeconomically disadvantaged groups [1,2]. One such group facing unique challenges in quitting smoking is people experiencing food insecurity or a lack of consistent access to sufficient food to lead a healthy, active life [3]. A person's level of food security is largely determined by purchasing power, which is mostly derived from wages, government transfers (eg, Social Security payments), and household composition [4], and money spent on cigarettes becomes unavailable to be spent on essential goods, including food. Although food insecurity is associated with both financial strain and psychological distress, it is also an independent predictor of smoking status and intensity and likely impedes smoking cessation success via multiple mechanisms, such as the reinforcement of smoking via nicotine's appetite suppressant effects and negative affect reduction [5]. Thus, there is a critical need to directly target individuals experiencing food insecurity with evidence-based smoking cessation interventions. In this population, successful long-term smoking cessation might be especially beneficial, as it may lead not only to abstinence-based health improvements but also to potential increases in food security due to both eliminating spending on cigarettes and reducing spending on smoking-related health care costs.

Recruiting special populations (eg, patients with cancer) to smoking cessation intervention trials can be very challenging, with eligibility percentages among those contacted sometimes in the low single digits [6]. Given these and other obstacles to obtaining desired sample sizes, researchers targeting various special populations have often used innovative recruitment strategies. For example, a recent feasibility trial for a text-based smoking cessation intervention for sexual and gender minorities yielded 79 participants following the posting of approximately 700 flyers in a variety of public establishments in the local catchment area, 1500 emails sent to potentially eligible participants on ResearchMatch (who were overwhelmingly from outside the catchment area), and the posting of advertisements on Facebook, Instagram, and Craigslist. The screening yield rate for each modality was below 6%, although the enrollment rate among those screened was approximately 50% [7].

Latino Americans are another special population that can be difficult to successfully reach and enroll in smoking cessation randomized controlled trial (RCTs), necessitating multiple types of recruitment approaches. For example, a secondary analysis of data from *Decídetexto*, a mobile health smoking cessation RCT for Latino Americans conducted across 4 states, compared the effectiveness of outbound, direct recruitment strategies, in which study staff contacted potential participants, to inbound, mass recruitment strategies (eg, flyers and television, newspaper, or Facebook advertisements), in which interested individuals had to contact study personnel to be screened [8]. Although most enrollees were recruited via "high effort" direct strategies (eg, in-person community outreach or personal phone calls to patients in registries), individuals recruited through either mass recruitment or "low effort" direct strategies (eg, emailing or texting patients in registries) were more likely to screen eligible and to enroll in the trial. In terms of personal characteristics, mass-recruited enrollees differed considerably from enrollees recruited via direct strategies, including having lower socioeconomic status, being less likely to have health insurance, and being less acculturated (eg, being more likely to primarily speak Spanish).

A contrasting pattern emerged from a smoking cessation RCT conducted by our team for another special population, cervical cancer and precancer survivors [9,10]. Enrollees recruited online (vs enrollees recruited from clinics at a National Cancer Institute–designated cancer center) had higher levels of education and health literacy but did not differ in age, race, household income, smoking behaviors, or motivation to quit smoking [9]. Thus, online recruitment strategies (eg, social media advertisements) cannot be presumed to consistently yield more- or less-disadvantaged members of special populations than in-person strategies.

To date, few studies have specifically addressed the process of recruiting people who smoke and are experiencing food insecurity to smoking cessation RCTs. One notable exception, however, was a pilot study that supported the feasibility of in-person recruitment of such individuals at food pantries to a smoking cessation treatment study [11]. The study entailed 22 research staff visits distributed across 4 food pantries in the Greater Cleveland region. During each site visit, the research staff set up an outreach table and remained there for between 1.5 and 3.5 hours, during which approximately 28% of the clients (173 of an estimated 628 total pantry visitors) approached their table. Of them, 31% (54/173) expressed interest in

XSL•FO
RenderX

information on quitting smoking, and of those interested in information on quitting, 78% (42/54) were able to be contacted and screened for eligibility via phone. Ultimately, 18% (31/173) of eligible individuals gave consent to have their contact information sent to the Ohio Tobacco Quitline (ie, 5% [31/628] of individuals served by the food pantries during recruitment) [11]. Thus, the recruitment yield was low.

With the limited exceptions of the studies described earlier that recruited Latino Americans and cervical cancer or precancer survivors [8,9], few studies have examined (1) potential sociodemographic, psychosocial, or smoking-related differences among individuals recruited to smoking cessation trials using in-person versus online recruitment methods or (2) differences in recruitment yield and enrollment outcomes associated with these 2 methods [6,9]. Indeed, to our knowledge, the current secondary analysis represents the first such comparison using data from a smoking cessation RCT for people experiencing food insecurity, a vulnerable special population for whom quitting is especially urgent.

Objectives

The primary objective of this secondary analysis was to compare baseline sociodemographic, smoking-related, and psychosocial characteristics of participants recruited via in-person versus online methods to a National Cancer Institute–funded smoking cessation trial evaluating the efficacy of a mobile health intervention for individuals experiencing food insecurity, as such differences may have clinically meaningful implications for trial design and outcome interpretation. A secondary objective was to describe screening, eligibility, and enrollment rates among individuals recruited in person versus online.

Methods

Trial Design

The study underlying this secondary analysis is a smoking cessation RCT evaluating the efficacy of a fully automated smartphone-delivered intervention (automated treatment [AT]) versus connections to state quitlines (standard treatment [ST]).

Trial Setting

Individuals reporting current smoking and recent receipt of food assistance (eg, food bank or food pantry or electronic benefits transfer [EBT] or food stamps) were recruited in person (March 2022 to February 2024) and online (February 2024 to June 2024). All assessments were conducted online via REDCap (Research Electronic Data Capture), and participants were followed for 12 months [12]. Additional procedural information has been published previously [13].

Intervention and Comparator

The AT intervention was fully automated after enrollment, whereas each ST participant's contact information was provided to the appropriate state quitline, and the quitline proactively contacted the participant. All participants received a 10-week supply of combination nicotine replacement therapy (ie, nicotine patches and lozenges).

Outcomes

This secondary analysis focuses on participant recruitment via online and in-person modalities; however, the primary and secondary outcomes of the parent RCT are described here for context. The primary outcome of the RCT was self-reported smoking abstinence (missing=smoking) at the 12-month follow-up. The secondary outcomes included abstinence verified biochemically via salivary cotinine samples and self-reported continuous, 30-day, and 24-hour abstinence, all at the 12-month follow-up. For each outcome in the parent RCT, the proportion of participants abstinent in AT versus ST was compared. These outcomes are not reported in this secondary analysis.

Sample Size

The target sample size was 500 (250 per group), as that was estimated to provide 80% power to detect an increase in the self-reported smoking abstinence rate (missing=smoking) at 12 months of 7% in AT (vs ST) [13].

Randomization

After consenting to participate and completing the baseline assessment using REDCap, participants were randomly assigned to AT or ST following stratification on sex and nicotine dependence. Allocation was implemented via REDCap's randomization module, and enrolling staff did not have access to the allocation sequence (see [13]).

Blinding

Participants were not blind to treatment condition.

Patient and Public Involvement

In-person recruitment was facilitated via a partnership with Feeding Tampa Bay, which aided the researchers in the conduct of the trial. There was no other public or patient involvement in the design, conduct, and reporting of the trial.

Changes to Trial Protocol

To increase recruitment, the study team collaborated with an advertising partner to develop an online advertising campaign (eg, online advertisements and the study landing page), which used targeted social media (such as Facebook and Instagram) and search engine advertising and ran for approximately 5 months, from February to June 2024.

Harms

Although this was a low-risk smoking cessation RCT, participant safety was systematically monitored in accordance with the study's data and safety monitoring plan by the study multiple principal investigators (JIV and DJV) throughout the trial, with adverse events being defined as any unexpected physical, psychological, or social harms occurring during the trial period.

Data Collection and Management

The AT participants used a unique link sent to their email address to download the treatment app. Data collected on the app were stored via a backend database meeting all Moffitt Cancer Center data security policies and procedures. Participants in the ST condition were connected to state quitlines that were free of charge to callers. All participants were assigned unique identification numbers for use in data analyses and for data

XSL•FO
RenderX

transfer, in the event of a request. Only study identification numbers were used in analysis files.

Participant Recruitment Strategies

In-Person Recruitment

Trained research coordinators, who generally worked in teams of 2 to 4, made 246 recruitment visits to food distribution events held at 86 sites throughout West Central Florida for 23 months (March 2022 to February 2024). The research coordinators wore matching, study-branded clothing and approached clients waiting in their automobiles (at events where clients received their food in their trunk) or sat at an event table and talked with food bank clients who approached the table (at the events where food bank clients were handed boxes of food or served meals in a restaurant-type setting).

Online Recruitment

The online recruitment targeting parameters were as follows: nationwide, smoking ≥5 cigarettes per day, having received food assistance in the past 3 months, being ≥18 years old, speaking English or Spanish, and not currently pregnant. Creation of the advertising partner's proprietary advertisement targeting algorithm was based on machine learning and data mining using aggregated data sets from social media, online searches, health websites, and online patient communities. Advertisements did not change during the course of the nationwide campaign. Although optimization services were available in the event that recruitment targets were not being met, they were not used because participant accrual was relatively rapid throughout the online campaign. In the third month of the campaign (April 2024), the campaign budget was lowered by half to decrease the rate at which potential participants were recruited, an action that was taken to preserve staff bandwidth. Recruitment ended in June 2024, shortly after reaching the study's target sample size (ie, 500 enrolled participants).

Materials and Measures

Eligibility Criteria

Potential participants completed an electronic screener to determine eligibility. Tablets were used to administer the questionnaire to individuals approached in person. Primary eligibility requirements included receiving food assistance in the past 30 days (food bank, EBT, or food aid from a family member or friend), smoking ≥5 cigarettes per day, and willingness to attempt quitting within 7 days of enrollment [13].

Baseline Assessment

Participants indicated their degree of food insecurity by completing the 6-item short form version of the United States Department of Agriculture's US Household Food Security Survey (eg, "In the last 12 months, did you ever eat less than you felt you should because there wasn't enough money for food?") [14]. Scores ranged from 0 to 6, with higher scores indicating greater food insecurity [14]. Other participant characteristics measured at baseline included sociodemographic, smoking-related, and psychosocial variables [13].

Statistical Methods

Recruitment groups were compared on categorical variables using chi-squared and Fisher exact tests and on continuous variables using independent samples *t* tests. Effect sizes were gauged using Cohen *d* for continuous variables and either Cramer *V* or odds ratios for categorical variables. CIs for Cohen *d* were created using the noncentral *t* distribution method, creating exact interval estimates for the standardized mean differences [15], while 95% CIs for Cramer *V*, which is bounded between 0 and 1 and whose magnitude can be interpreted similarly to a Pearson correlation coefficient [16,17], were constructed using a nonparametric percentile bootstrapping procedure with 10,000 resamples [18-20]. The 95% CIs for the odds ratios were calculated using the standard, natural logarithm transformation–based formula [21].

To explore whether differences in self-reported food insecurity between the in-person and online groups might have been a consequence of basic demographic differences between the groups, a hierarchical linear regression analysis was performed. Little's MCAR (missing completely at random) test was used to ascertain whether the data were MCAR [22]. Although there were no missing data on in-person versus online group membership, nor on food insecurity, cases with missing data on any of the predictors present in one or both models tested were deleted listwise, which resulted in a nominal 5% reduction of the analytic sample size from 502 to 475. To identify model covariates, univariate demographic predictors of either food insecurity ($P<.10$) or recruitment modality ($P<.05$) were identified using simple logistic (categorical variables) or linear (continuous variables) regression analyses [9]. (Candidate covariates included age, ethnicity or race, sex, sexual orientation, relationship status, educational attainment, health insurance coverage, household income, and employment status.) Next, variables meeting those criteria were entered as a set to estimate a covariates-only model (model 1). Before model inclusion, categorical variables were dichotomized (eg, income was approximately median split to form less than US $20,000 per year and US $20,000 or more per year categories) to prevent estimation problems (eg, coefficient instability and inflated SEs) [23]. The next step entailed adding recruitment modality to the model as a predictor (model 2). Differences in model fit were examined using an *F* test; a significant difference would indicate an improvement in model fit, meaning that there was an association between food insecurity and recruitment modality even after accounting for the effects of relevant demographic covariates. The 95% CIs for the proportion of variance in self-reported food insecurity explained by models 1 and 2 (ie, model R^2) were based on a noncentral *F* distribution and created using R package MBESS' function "ci.R2" with parameter Random.Predictors set to "FALSE," as the model predictors were fixed, rather than random [15,24]. To generate the increment in R^2 between models 1 and 2 (ie, ΔR^2), nonparametric bootstrapping with 10,000 resamples was used. An inspection of the histogram of the bootstrapped ΔR^2 estimates revealed slight positive skew in the distribution, and so both percentile and bias corrected and accelerated (BCa) bootstrap CIs were calculated and reported [20] (for details, see Multimedia Appendix 1).

XSL•FO
RenderX

To compare differences in enrollment patterns between participants recruited in person versus online, the flow of potential participants between prescreening and full enrollment in the trial was presented, keeping the in-person and online recruitment denominators (and thus percentages) as comparable as possible within the constraints of the available data. Prescreening numbers were defined as the number of people assessed for eligibility after being approached at food pantry sites and the number of people who were assessed for eligibility after clicking online advertisements. Regardless of whether they were recruited in person or online, potential participants who were not excluded at the prescreening stage completed the study's eligibility questionnaire.

Potential participants who prescreened eligible (eg, aged ≥18 y, self-reported smoking, and receiving food assistance) were considered "unreachable" after not being able to be contacted following approximately 2 weeks of staff outreach attempts using a structured outreach schedule. For the first week following prescreening, 3 REDCap notifications were sent via email and text (using Twilio integration) to remind potential participants of the screening survey, and study staff attempted contact via phone, text, or email 3 to 5 times. The second week consisted of 3 contact attempts approximately every other day. A similar procedure was followed at other steps of the enrollment process. For example, a participant who consented to participate but did not complete the baseline questionnaire was considered unreachable after 2 weeks of contact attempts.

Next, those who screened eligible and were not excluded (either because they were unreachable or because they declined to participate) completed the informed consent process. The onboarding completion rate was defined as the number of people who enrolled in the trial divided by the number of people who completed the consent process. Enrollment rates for those who both consented to participate and subsequently completed the preenrollment baseline questionnaire were calculated by dividing the number of people who completed the baseline questionnaire by the number of people who enrolled in the trial.

To provide a fuller sense of the intensity of recruitment efforts by recruitment modality, time (in days) from screening completion to enrollment was reported, along with the number of contact attempts made prior to enrollment for online recruits. Owing to the positive skew in the distributions of these values, medians and IQRs were reported in addition to means and SDs.

Ethical Considerations

Ethical Approval and Informed Consent

The RCT was approved by Advarra, which serves as Moffitt Cancer Center's institutional review board (IRB 00000971). The IRB and the informed consent document allowed this secondary analysis without additional consent. Informed consent was obtained from all participants. A waiver of written documentation of informed consent was obtained, allowing potential enrollees recruited in person to verbally consent, while potential participants recruited online read the consent document themselves via a consent screen on REDCap and acknowledged consent without needing to sign electronically.

Privacy and Confidentiality

All data were securely stored and managed at Moffitt Cancer Center using REDCap, with access limited to authorized study personnel. No identification of individual participants or users is possible in any images of this manuscript or supplementary material.

Compensation

No compensation was provided before enrollment in the trial (eg, individuals were not compensated for completing the screening questionnaire). Potential participants recruited in person and online were made aware of the study's compensation structure during the consent process and received identical compensation for all study-related tasks. Over the course of the 12-month study period, participants could potentially earn up to a total of US $510 [13].

Other Information

Reporting Guidelines

The trial was reported in accordance with the CONSORT (Consolidated Standards of Reporting Trials) 2025 statement (Checklist 1) [25] and the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist (Checklist 2) [26].

Trial Oversight

The study's multiple principal investigators were responsible for compliance with all federal and institutional IRB policies and the study's data and safety monitoring plan.

Results

Baseline Data

Online (vs in person) recruits reported greater food insecurity (mean 4.5, SD 1.9 vs mean 3.0, SD 2.3; d=0.74, 95% CI 0.55-0.93; t_{500}=7.93; P<.001). As illustrated in Figure 1, 64% (n=208) recruited online (vs in person: n=66, 37%) were classified as having very high food insecurity (odds ratio [OR] 3.04, 95% CI 2.05-4.53; χ^2_1=34.1; P<.001). Online recruits were also younger (45.7 vs 52.9 y; d=−0.70, 95% CI −0.89 to −0.51; t_{500}=−7.48; P<.001) and less likely to be Hispanic or Latino (n=22, 7% vs n=39, 22%; OR 0.25, 95% CI 0.14-0.47; χ^2_1=24.9; P<.001). Participants recruited online were more likely to be female (n=250, 77% vs n=117, 66%; OR 1.76, 95% CI 1.15-2.69; χ^2_1=7.6; P=.006), non-Hispanic White (n=249, 77% vs n=88, 50%; OR 3.35, 95% CI 2.26-5.06; χ^2_1=38.3; P<.001), and LGBT+ (n=57, 18% vs n=11, 6%; OR 3.20, 95% CI 1.60-6.96; χ^2_1=12.5; P<.001). Moreover, online recruits were more likely to be educated beyond high school or a general educational development certificate (n=225, 69% vs n=87, 49%; OR 2.37, 95% CI 1.60-3.53; χ^2_1=20.7; P<.001) and to have annual household income of US $20,000 or more (n=150, 46% vs n=64, 36%; OR 1.52, 95% CI 1.03-2.26; χ^2_1=4.8; P=.03).

XSL•FO
RenderX

Figure 1. Secondary analysis of data from a mobile health smoking cessation randomized controlled trial for people with food insecurity: food insecurity by recruitment modality. Note: Error bars are 95% Wilson CIs.

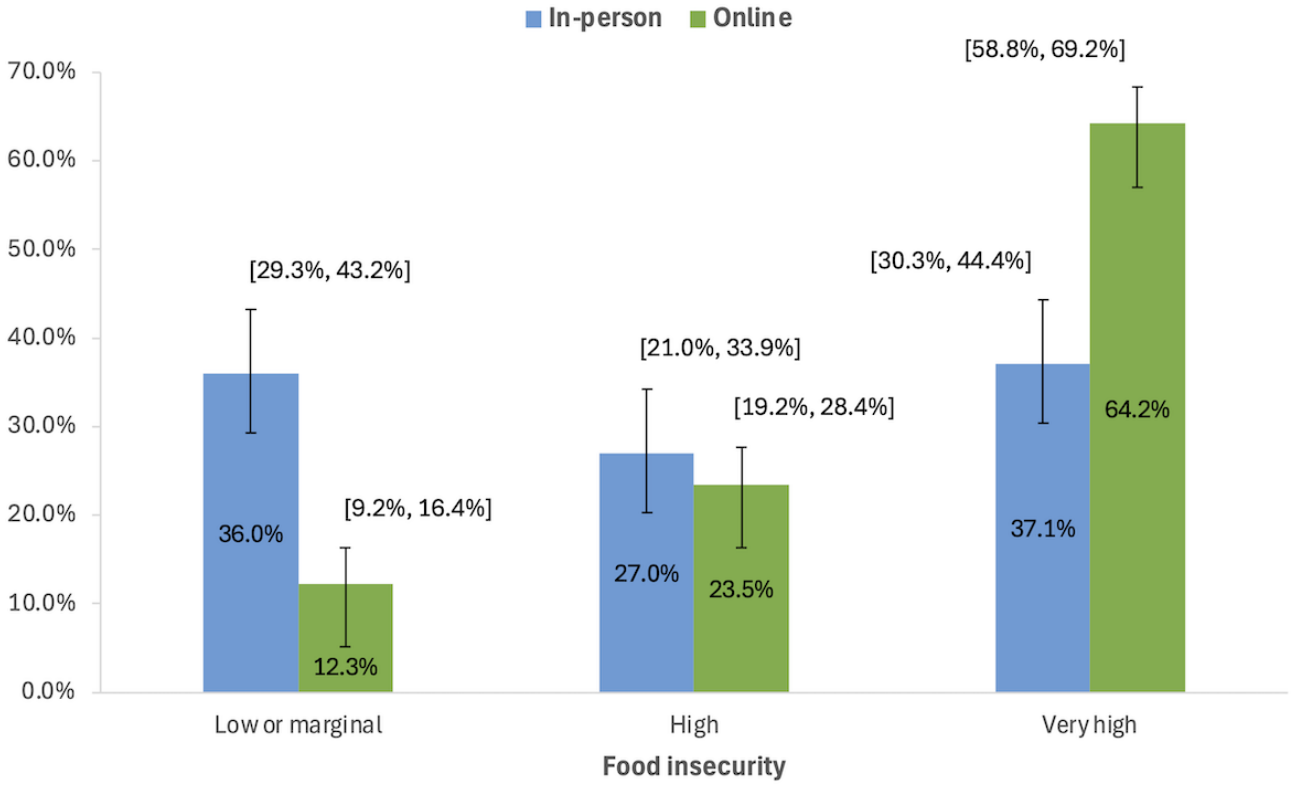

Online recruits reported lower motivation to quit smoking (Contemplation Ladder; mean 7.2, SD 2.4 vs mean 8.0, SD 2.8; d=−0.315, 95% CI −0.499 to −0.131; t_{500}=3.38; P<.001), smoking cessation self-efficacy (mean 20.5, SD 8.0 vs mean 23.2, SD 8.6; d=−0.318, 95% CI −0.502 to −0.134; t_{499}=3.40; P<.001), and subjective social status (mean 4.6, SD 2.0 vs mean 5.9, SD 2.2; d=−0.63, 95% CI −0.82 to −0.44; t_{498}=−6.75; P<.001). They also reported more financial strain (mean 17.9, SD 6.3 vs mean 16.2, SD 6.6; d=0.27, 95% CI 0.09-0.46; t_{486}=2.89; P=.004), depressive symptoms (mean 8.6, SD 6.3 vs mean 7.4, SD 6.1; d=0.19, 95% CI 0.01-0.38; t_{498}=2.07; P=.04), and loneliness (mean 6.0, SD 2.1 vs mean 5.2, SD 2.0; d=0.40, 95% CI 0.22-0.59; t_{499}=4.33; P<.001), as well as lower resilience (mean 19.5, SD 5.1 vs mean 20.5, SD 4.3; d=−0.22, 95% CI −0.40 to −0.03; t_{493}=−2.30; P=.02).

Results were mixed with respect to substance use. Online recruits reported lower levels of past-year alcohol misuse (ie, binge drinking or heavy drinking, per National Institute on Alcohol Abuse and Alcoholism guidelines [27]; n=88, 27% vs n=66, 37%; OR 0.63, 95% CI 0.42-0.95; χ^2_1=5.3; P=.02). However, online participants were more likely to report having used one or more recreational drugs on a monthly or more frequent basis in the past year (n=74, 23% versus n=17, 10%; OR 2.78, 95% CI 1.56-5.22; χ^2_1=13.5; P<.001) and to report recreational cannabis use in particular within the past 30 days (n=80, 25% vs n=27, 15%; OR 1.82, 95% CI 1.10-3.07; χ^2_1=6.1; P=.01).

Further results (eg, on tobacco product use) are available in Table 1. Notably, using the Benjamini-Hochberg procedure [28,29] to control the false discovery rate (or FDR) by limiting the proportion of the significant analyses in Table 1 due to false positives to 0.05, only 3 (11%) of the 28 (out of 39 total) significance tests became nonsignificant. Specifically, differences between recruitment groups on typical cigarette type (regular vs menthol), age of smoking initiation, and depressive symptoms had P values <.05 but were nonsignificant according to the FDR-adjusted results.

XSL•FO
RenderX

Table . Secondary analysis of data from a mobile health smoking cessation randomized controlled trial for people with food insecurity: participant characteristics at baseline by recruitment modality.

Participant characteristics	Overall	Recruitment modality		*P* value[a]	Cohen *d* (95% CI) or Cramer *V* (95% CI)[b]
		In person	Online		
Food insecurity score (0 - 6) (n=502), mean (SD)	3.9 (2.2)	3.0 (2.3)	4.5 (1.9)	<.001	−0.740 (−0.928 to −0.551)
Food insecurity group (n=502), n (%)				<.001	0.306 (0.222 to 0.394)
Low or marginal food insecurity (0 - 1)	104 (21)	64 (36)	40 (12)		
High food insecurity (2-4)	124 (25)	48 (27)	76 (23)		
Very high food insecurity (5-6)	274 (55)	66 (37)	208 (64)		
Age (n=502), mean (SD)	48.2 (10.9)	52.9 (11.7)	45.7 (9.5)	<.001	0.698 (0.510 to 0.886)
Ethnicity or race (n=501), n (%)				<.001	0.336 (0.257 to 0.426)
Non-Hispanic White	337 (67)	88 (50)	249 (77)		
Non-Hispanic Black or African American	71 (14)	43 (24)	28 (9)		
Non-Hispanic mixed race or other	32 (6)	7 (4)	25 (8)		
Hispanic or Latino	61 (12)	39 (22)	22 (7)		
Sex (n=502), n (%)				.006	0.123 (0.034 to 0.213)
Male	135 (27)	61 (34)	74 (23)		
Female	367 (73)	117 (66)	250 (77)		
Gender identity (n=501), n (%)				.025	0.119 (0.046 to 0.212)
Male	137 (27)	61 (34)	76 (23)		
Female	360 (72)	115 (65)	245 (76)		
Transgender	4 (<1)	1 (<1)	3 (<1)		
Sexual orientation (n=500), n (%)				<.001	0.171 (0.099 to 0.248)
Straight or heterosexual	424 (85)	160 (91)	264 (81)		
Gay or lesbian, bisexual, or prefer to self-describe	68 (14)	11 (6)	57 (18)		
Prefer not to say	8 (2)	5 (3)	3 (<1)		
Relationship status (n=500), n (%)				.85	0.025 (0.011 to 0.130)
Single	203 (41)	71 (40)	132 (41)		
Married, living with significant other, or partnered	183 (37)	68 (38)	115 (36)		
Divorced, widowed, or separated	114 (23)	39 (22)	75 (23)		
Educational attainment (n=502), n (%)				<.001	0.207 (0.132 to 0.299)

XSL•FO
RenderX

Participant characteristics	Overall	Recruitment modality		*P* value[a]	Cohen *d* (95% CI) or Cramer *V* (95% CI)[b]
		In person	Online		
Less than high school	48 (10)	25 (14)	23 (7)		
High school diploma or GED[c]	142 (28)	66 (37)	76 (23)		
Some college or vocational training or associate degree	256 (51)	73 (41)	183 (56)		
Four-year degree or more	56 (11)	14 (8)	42 (13)		
Health insurance coverage (n=479), n (%)				.003	0.135 (0.042 to 0.227)
Yes	400 (84)	128 (77)	272 (87)		
No	79 (16)	39 (23)	40 (13)		
Household income (US$; n=501), n (%)				.001	0.189 (0.121 to 0.283)
<10,000	139 (28)	58 (33)	81 (25)		
10,000-19,999	130 (26)	43 (24)	87 (27)		
20,000-29,999	89 (18)	34 (19)	55 (17)		
≥30,000	125 (25)	30 (17)	95 (29)		
Refuse to answer	18 (4)	12 (7)	6 (2)		
Employment status (n=502), n (%)				.55	0.065 (0.030 to 0.170)
Employed for wages or self-employed	153 (30)	53 (30)	100 (31)		
Out of work	96 (19)	29 (16)	67 (21)		
Homemaker, student, retired, or refuse to answer	119 (24)	47 (26)	72 (22)		
Unable to work	134 (27)	49 (28)	85 (26)		
Lives in household with another person who smokes (n=500), n (%)				.83	0.010 (0.001 to 0.102)
Yes	185 (37)	64 (36)	121 (37)		
No	315 (63)	112 (64)	203 (63)		
Lifetime quit attempts (n=494), n (%)				.41	0.077 (0.034 to 0.184)
0	49 (10)	18 (11)	31 (10)		
1 - 4	268 (54)	84 (49)	184 (57)		
5 - 8	76 (15)	29 (17)	47 (15)		
≥9	101 (20)	40 (23)	61 (19)		
Abstinence goal (n=499), n (%)				.20	0.109 (0.063 to 0.207)
Total abstinence, never use again	335 (67)	130 (73)	205 (64)		
Total, but could slip and maintain abstinence	81 (16)	24 (13)	57 (18)		

XSL•FO
RenderX

Participant characteristics	Overall	Recruitment modality		*P* value[a]	Cohen *d* (95% CI) or Cramer *V* (95% CI)[b]
		In person	Online		
Occasional use when urges strongly felt	43 (9)	13 (7)	30 (9)		
Temporary abstinence	12 (2)	5 (3)	7 (2)		
No goal of abstinence at this time	28 (6)	6 (3)	22 (7)		
Cigarette type (n=501), n (%)				.04	0.090 (0.010 to 0.180)
Nonmenthol	271 (54)	85 (48)	186 (57)		
Menthol	230 (46)	92 (52)	138 (43)		
Other tobacco products used in past month (n=502), n (%)					
None	404 (80)	127 (71)	277 (85)	<.001	0.171 (0.080 to 0.262)
Cigars	43 (9)	27 (15)	16 (5)	<.001	0.175 (0.082 to 0.266)
Little cigars, cigarillos, bidis, or Black and Milds	57 (11)	21 (12)	36 (11)	.82	0.010 (0.001 to 0.104)
Pipe with tobacco	6 (1)	1 (<1)	5 (2)	.43	0.043 (0.003 to 0.098)
Dip, chew, or snus	5 (1)	3 (2)	2 (<1)	.35	0.051 (0.002 to 0.137)
Ever used e-cigarettes or vaped (n=501), n (%)				<.001	0.215 (0.128 to 0.305)
Yes	332 (66)	93 (53)	239 (74)		
No	169 (34)	84 (47)	85 (26)		
Last time used e-cigarettes or vaped (n=331), n (%)				.008	0.146 (0.044 to 0.248)
Within the past 30 d	134 (40)	27 (29)	107 (45)		
More than 30 d ago	197 (60)	66 (71)	131 (55)		
Cigarettes per day (n=502), n (%)				.11	0.111 (0.050 to 0.212)
0 - 10	133 (26)	51 (29)	82 (25)		
11 - 20	240 (48)	77 (43)	163 (50)		
21 - 30	97 (19)	33 (19)	64 (20)		
≥31	32 (6)	17 (10)	15 (5)		
Time to first cigarette after waking (n=502), n (%)				<.001	0.222 (0.143 to 0.319)
After 60 min	29 (6)	17 (10)	12 (4)		
31-60 min	43 (9)	26 (15)	17 (5)		
6-30 min	195 (39)	70 (39)	125 (39)		
Within 5 min	235 (47)	65 (37)	170 (52)		
Heaviness of Smoking Index score (0 - 6) (n=502), mean (SD)	3.3 (1.4)	3.1 (1.5)	3.4 (1.3)	.01	−0.233 (−0.416 to −0.049)

XSL•FO
RenderX

Participant characteristics	Overall	Recruitment modality		*P* value[a]	Cohen *d* (95% CI) or Cramer *V* (95% CI)[b]
		In person	Online		
Age initiated smoking (n=492), mean (SD)	17.6 (6.2)	18.4 (7.2)	17.2 (5.6)	.04	0.199 (0.014 to 0.385)
Years smoking (n=502), mean (SD)	27.6 (11.6)	30.3 (12.9)	26.2 (10.6)	<.001	0.355 (0.171 to 0.539)
Contemplation Ladder (0 - 10) (n=502), mean (SD)	7.5 (2.6)	8.0 (2.8)	7.2 (2.4)	<.001	0.315 (0.131 to 0.499)
Smoking cessation self-efficacy (9-45) (n=501), mean (SD)	21.5 (8.3)	23.2 (8.6)	20.5 (8.0)	<.001	0.318 (0.134 to 0.502)
Subjective social status (1-10) (n=500), mean (SD)	5.1 (2.2)	5.9 (2.2)	4.6 (2.0)	<.001	0.632 (0.444 to 0.820)
Financial strain (0 - 26) (n=488), mean (SD)	17.3 (6.4)	16.2 (6.6)	17.9 (6.3)	.004	−0.275 (−0.462 to −0.088)
Depressive symptoms (0 - 24) (n=500), mean (SD)	8.2 (6.2)	7.4 (6.1)	8.6 (6.3)	.04	−0.194 (−0.378 to −0.010)
Perceived stress (4-20) (n=497), mean (SD)	10.7 (3.1)	10.5 (2.8)	10.9 (3.2)	.19	−0.124 (−0.308 to 0.060)
Resilience (5-30) (n=495), mean (SD)	19.8 (4.9)	20.5 (4.3)	19.5 (5.1)	.02	0.216 (0.031 to 0.401)
Sense of control (8-32) (n=493), mean (SD)	23.9 (3.0)	23.6 (2.6)	24.1 (3.2)	.15	−0.135 (−0.319 to 0.050)
Loneliness (3-9) (n=501), mean (SD)	5.7 (2.1)	5.2 (2.0)	6.0 (2.1)	<.001	−0.405 (−0.589 to −0.220)
Alcohol misuse, past year (n=499), n (%)				.02	0.103 (0.017 to 0.194)
Yes	154 (31)	66 (37)	88 (27)		
No	345 (69)	111 (63)	234 (73)		
Recreational drug use frequency, past year (n=501), n (%)				.004	0.175 (0.118 to 0.257)
Daily or almost daily	56 (11)	9 (5)	47 (15)		
Weekly	19 (4)	3 (2)	16 (5)		
Monthly	16 (3)	5 (3)	11 (3)		
Once or twice	73 (15)	31 (18)	42 (13)		
Never	337 (67)	129 (73)	208 (64)		
Recreational cannabis use, past 30 days (n=501), n (%)				.01	0.110 (0.026 to 0.190)
Yes	107 (21)	27 (15)	80 (25)		
No	394 (79)	150 (85)	244 (75)		

[a]Independent samples *t* test; Pearson χ^2 test; Fisher exact test.

[b]Effect size measures: Cohen *d* [95% CI] for continuous outcomes [15]; Cramer *V* [16,17] [95% CI] for categorical outcomes [18-20].

[c]GED: general educational development certificate (high school equivalent).

XSL•FO
RenderX

Recruitment

Participant Flow

As detailed in Figure 2, the large majority of potential recruits via both in-person and online modalities were excluded via prescreening before completing the eligibility screener. Of 12,691 approaches at food pantries, 12,211 (96%) did not complete the eligibility screener, predominantly because they did not currently smoke cigarettes (11,344/12,211, 93%). Potential online recruits were excluded at a lower rate of 78% (1999/2574 total potential online recruits); most were either not receiving food assistance (1094/1999, 55%) or were unreachable (502/1999, 25%).

Figure 2. Secondary analysis of data from a mobile health smoking cessation randomized controlled trial for people with food insecurity: CONSORT-style flow diagram by recruitment modality. CONSORT: Consolidated Standards of Reporting Trials.

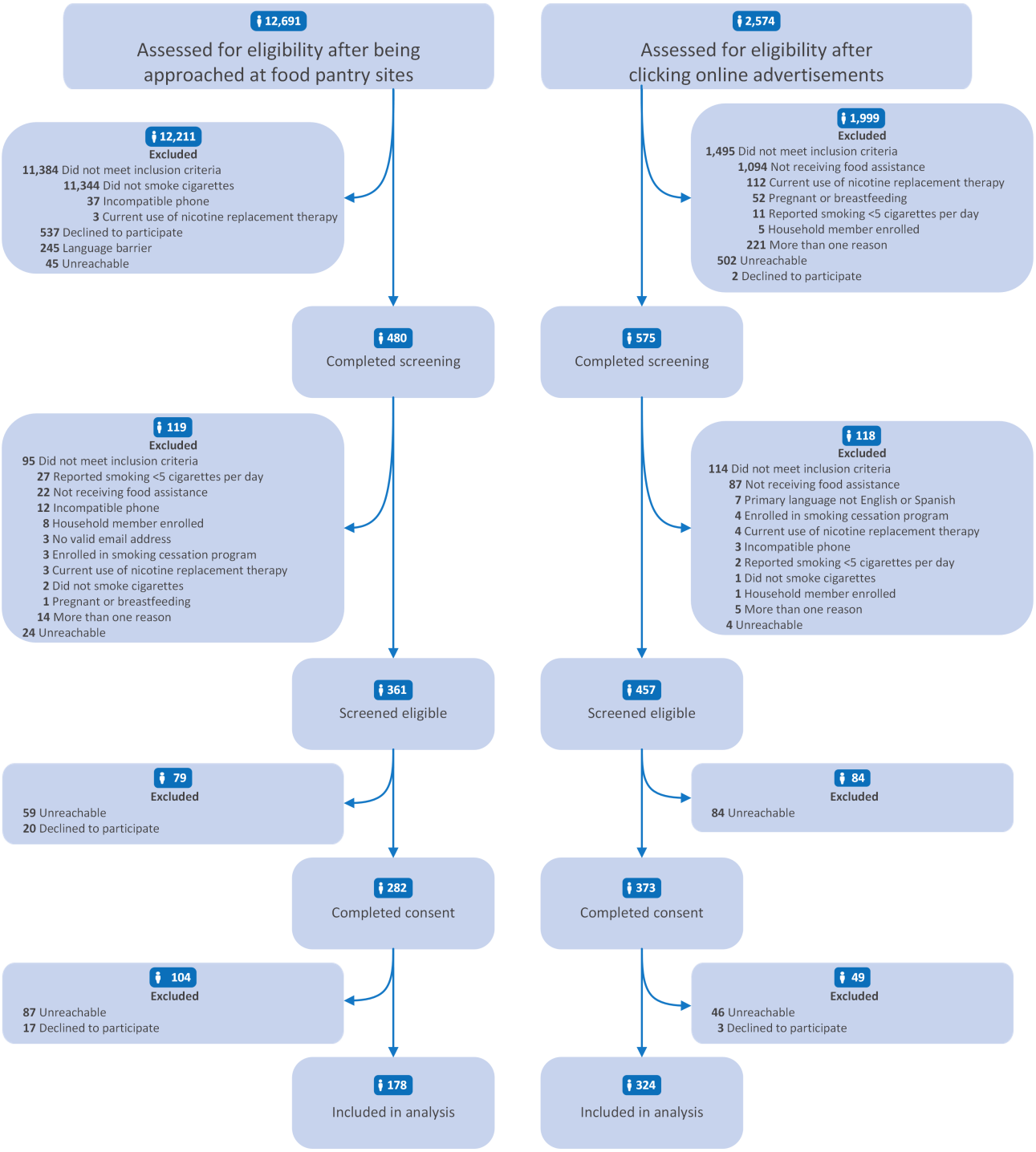

Intervention and Comparator Delivery

Site visits yielded 178 participants over 23 months, an in-person enrollment rate of 7.7 per month. Online recruitment garnered 324 participants over 5 months—an enrollment rate of 64.8 per month—resulting in a final sample size of 502. A total of 1055 individuals (online: 575, 55% and in person: 480, 45%) completed the eligibility screener. Potential online and in-person recruits did not differ significantly on study eligibility (457/575,

XSL•FO
RenderX

79%) versus (361/480, 75%; OR 1.27, 95% CI 0.95-1.72; χ^2_1=2.7; *P*=.10). However, eligible individuals recruited online were more likely to enroll (324/457, 71%) than eligible individuals recruited in person (178/361, 49%; OR 2.50, 95% CI 1.86-3.38; χ^2_1=39.7; *P*<.001). Among those who completed the eligibility screener, enrollment was higher for online recruits (324/575, 56%) than for in-person recruits (178/480, 37%; OR 2.19, 95% CI 1.70-2.83; χ^2_1=38.9; *P*<.001). Thus, recruitment modality was primarily associated with enrollment among eligible participants, rather than eligibility itself.

Online advertisement impression data were not available. A total of 2574 individuals clicked an advertisement for the study. Among those who clicked an advertisement, 12.6% (324/2574) enrolled in the trial. In contrast, 1.4% (178/12,691) of all individuals approached in person ultimately enrolled; however, enrollment was 13.2% (178/1347) among those not immediately ruled out at prescreening due to self-reported nonsmoking (12,691 approached minus 11,344 reporting nonsmoking).

The onboarding completion rate (ie, full enrollment) among those who completed consent was higher for online recruits (324/373, 86.9%) than for in-person recruits (178/282, 63.1%). Enrollment rates among those who both completed consent and the baseline questionnaire were 97.0% (324/334) for online recruits and 89.4% (178/199) for in-person recruits. Screening to enrollment times were shorter and less variable for online recruits (median 1, IQR 0-2 d; and mean 2.2, SD 5.1 d) than for in-person recruits (median 2, IQR 1-8 d; mean 10.0, SD 39.6 d). Similarly, fewer contact attempts were made prior to enrollment for online recruits (median 0, IQR 0-1; and mean 0.5, SD 0.9) than for in-person recruits (median 1, IQR 1-3; and mean 2.5, SD 3.1).

Harms

No adverse events were reported, nor were any stopping rules triggered during the trial.

Ancillary Analyses

Little's MCAR test indicated that the data used in the exploratory hierarchical linear regression analysis were not MCAR (χ^2_{32}=52.9; *P*=.01). To evaluate the sensitivity of the results to the handling of missing data, multiple imputation by chained equations was conducted, generating 20 imputed datasets using predictive mean matching [30]. The pooled regression results using the multiply-imputed datasets did not differ substantively on predictor significance or effect magnitude from those obtained using listwise deletion, and conclusions were unchanged. Thus, for parsimony and ease of interpretation, results presented below are based on the complete-case analyses.

Exploratory hierarchical linear regression analysis results (Table 2) indicated that, as a set, the demographic covariates explained a significant proportion of the variance in self-reported food insecurity (R^2=0.098, 95% CI 0.040-0.136; $F_{7,467}$=7.21; *P*<.001). When recruitment modality was added as a predictor, the resulting model (ie, model 2) remained significant (R^2=0.150, 95% CI 0.081-0.194; $F_{8,466}$=10.29; *P*<.001); the improvement in model fit was significant ($F_{1,466}$=28.86; *P*<.001). Critically, recruitment modality accounted for approximately an additional 50% of the variability in self-reported food insecurity above and beyond the demographic factors alone (ΔR^2=0.053; percentile bootstrapped 95% CI 0.019-0.100; BCa bootstrapped 95% CI 0.020-0.101). Briefly, in model 2, age (*b*=−0.03, 95% CI −0.05 to −0.01; *P*=.003) and having an annual household income ≥$20,000 (*b*=−0.52, 95% CI −0.91 to −0.14; *P*=.008) were significantly associated with lesser food insecurity. In contrast, female sex (*b*=0.49, 95% CI 0.06-0.92; *P*=.03) and being recruited online (vs in person; *b*=1.22, 95% CI 0.78-1.67; *P*<.001) were significantly associated with greater food insecurity.

XSL•FO
RenderX

Table . Secondary analysis of data from a mobile health smoking cessation randomized controlled trial for people with food insecurity: predicting food insecurity (0 - 6) with associated demographic variables (model 1) versus demographic variables plus recruitment modality (model 2).

Predictor	Model 1[a] (n=475[b])		Model 2[c] (n=475[b])[d]	
	b (95% CI)	*P* value	b (95% CI)	*P* value
Age (y)	−0.05 (−0.06 to −0.03)	<.001	−0.03 (−0.05 to −0.01)	.003
Non-Hispanic White	0.61 (0.20 to 1.03)	.004	0.24 (−0.19 to 0.67)	.27
Female	0.56 (0.12 to 1.00)	.01	0.49 (0.06 to 0.92)	.03
LGBT+[e]	0.26 (−0.31 to 0.83)	.37	0.14 (−0.42 to 0.69)	.63
Educated beyond HS[f] or GED[g]	0.47 (0.07 to 0.88)	.02	0.26 (−0.15 to 0.66)	.21
No health insurance coverage	0.23 (−0.30 to 0.75)	.40	0.38 (−0.13 to 0.89)	.15
Income of US $20,000 or more per year	−0.54 (−0.94 to −0.15)	.007	−0.52 (−0.91 to −0.14)	.008
Recruited online (vs in person)	__[h]	—	1.22 (0.78 to 1.67)	<.001

[a]R^2: b=0.098, 95% CI 0.040-0.136.

[b]Cases with any missing predictor data (n=27) were deleted listwise, yielding the analytic sample of 475 participants.

[c]R^2: b=0.150, 95% CI 0.081-0.194.

[d]Change in model fit: $F_{1, 466}$=28.86; P<.001; ΔR^2=0.053, 95% CI 0.019-0.100.

[e]LGBT+: lesbian, gay, bisexual, transgender, and others.

[f]HS:high school.

[g]GED: General educational development certificate (high school equivalent).

[h]Not applicable.

Discussion

Interpretation

This secondary analysis examined differences in baseline sociodemographic, psychosocial, and smoking-related characteristics of participants recruited in person and online to a smoking cessation RCT for people with food insecurity and compared differences in recruitment pace and enrollment throughput across in-person versus online recruitment methods. Results indicated that participants recruited in person at food distribution events in West Central Florida differed in important and sometimes paradoxical ways from those recruited online via social media. Given our use of these 2 different modalities at different points in the study, a secondary objective was to examine the relative speed of in-person versus online recruitment. Following relatively slow in-person recruitment, switching to an online, national recruitment strategy was associated with a substantially higher enrollment rate; however, other concurrent differences between the strategies and their consequences, such as how in-person and online recruits became aware of the study (ie, face-to-face interactions with study personnel vs online via advertisements on social media or internet search results) and geographical reach (regional vs national) are important to consider.

To our knowledge, this study provided the most comprehensive comparison of the personal characteristics of participants recruited via different modalities to a smoking cessation RCT for a special population of any kind. The comparison revealed noteworthy sociodemographic similarities between groups, including relationship status, living in a household with another person who smoked, and employment status. However, there were also important group differences. In-person recruits had lower household income and educational attainment. They were also significantly older, more ethnically diverse, and more likely to be male. Despite having lower socioeconomic status, participants recruited at in-person food distribution events reported lower levels of loneliness, greater resilience, fewer depressive symptoms, higher subjective social status, and lower levels of perceived financial strain (despite having lower incomes). These protective characteristics [31,32] may have reflected greater engagement in their communities [33] and receiving services from a trusted community organization (ie, Feeding Tampa Bay) [34]. They could also represent self-selection into receiving on-site food assistance services by people who already felt a relatively strong connection to their communities [35]. That said, it remains possible that regular in-person engagement with community-based resources may have engendered a sense of belonging and support [36]. Furthermore, the additional resources to address food insecurity provided by food banks (above and beyond all other possible sources of food aid, such as EBT or food stamps) may have resulted in the in-person recruits having more food available on average than the individuals recruited online, potentially helping to explain the considerable gap in food insecurity between the recruitment groups [37].

It is also notable that relative to online recruits, in-person recruits scored more favorably on a variety of characteristics associated

XSL•FO
RenderX

with smoking cessation success. These included considerably higher levels of food security [5], lesser financial strain [38], higher subjective social status [39], better psychosocial functioning [40,41], and greater readiness to quit smoking [42]. An important caveat is that although in-person recruits were less likely to report using recreational drugs, they were more likely to endorse alcohol misuse and more likely to consume noncigarette tobacco products (eg, cigars). In keeping with vaping's greater frequency among younger cohorts [2], however, in-person recruits reported both having tried vaping and current vaping less frequently than online recruits. Finally, the exploratory hierarchical linear regression analysis suggested that the observed difference in food security between the groups was not solely a byproduct of contrasting basic demographic characteristics. The results were consistent with the possibility that differences in behavioral and psychological characteristics between the groups might help explain the sizable difference in food insecurity observed between the groups. Although the analyses could not establish causal pathways, the in-person and online recruits nonetheless differed considerably in terms of risk and protective factors found to predict successful smoking cessation in previous studies.

On average, individuals recruited online may have been more technologically adept, comfortable with digital platforms, and willing to engage with smartphone-based care than individuals recruited in person, who had not proactively self-selected into screening for participation via a personal electronic device but rather had been approached at a food bank and invited to complete the screener via tablet [43,44]. Comfort and skill with technology might not only be relevant to recruitment outcomes but also might have downstream consequences for intervention uptake, follow-up completion, and cessation outcomes [45,46].

Taken together, findings suggest that a 2-pronged recruitment approach involving both in-person and online methods may be needed to effectively target the wide spectrum of individuals seeking food assistance. It is possible that differences might not have been observed if online recruitment had been conducted only in the same geographic area that in-person recruitment took place [47]. Thus, online recruitment may not always serve as a complete substitute for "boots on the ground" approaches, depending on researchers' goals and priorities. Furthermore, results from the pilot trial on smoking cessation treatment recruitment at food pantries in Greater Cleveland suggest that in-person recruitment at food distribution events resulted in a slightly higher proportion of people seeking food assistance who enrolled. Nonetheless, the enrollee yield rate in that study was still very modest [11]. Given the divergence in the personal profiles of the in-person and online recruits, it could be beneficial to use stratified or adaptive recruitment methods to improve representation of groups of interest recruited via a given method. For example, recruitment modality quotas could be prespecified or online advertisement spending reallocated toward underrepresented subgroups in real time.

Intriguingly, although participants recruited in person and online were equally likely to screen eligible, the probability of enrolling in the study was much higher for individuals recruited online, suggesting a need for future research designed to carefully evaluate differences in recruitment yield by recruitment strategy. For example, the relatively private, impersonal online approach might encourage some otherwise reluctant individuals to enroll [48], while others might more effectively be persuaded to enroll via in-person interactions with study personnel [49]. Learning which approach is likely to be more effective for different groups of people who smoke could improve both recruitment speed and the reach of future smoking cessation trials and programs.

Limitations

First, it should be noted that in-person approaches and online advertisement clicks represent different stages of study exposure and intent to participate (staff-initiated initial contact vs self-selected engagement). Thus, comparisons of the percentages of individuals recruited via in-person versus online modalities who enrolled should be contextualized in terms of those differences, although the conversion percentages become more directly comparable later in the enrollment process (eg, enrollment rates among those who provided consent to participate and completed the baseline questionnaire). Moreover, possible geography-based differences between in-person and online recruits included overall rurality, state and local smoking ordinances, and the prevalence of smoking in participants' immediate environments [47]. Furthermore, online and in-person recruitment occurred during different periods, which might have influenced both receptiveness and observed differences. Future researchers could avoid related concerns by recruiting via both modalities simultaneously and, if feasible and desired, ensuring that online and in-person participants live in the same region (eg, a cancer center's catchment region).

Contextual priming effects (eg, answering questions in the food bank environment) [50], social desirability bias (eg, wanting to please study staff) [51], and differential modes of survey administration (eg, using personal phones vs tablets provided by study staff to fill out questionnaires) were also possible [43,44]. Nevertheless, mode of survey administration was not systematically recorded, and some participants filled out the baseline questionnaire while at the food bank (albeit in a private area); thus, contextual priming, social desirability, and survey administration mode–based effects cannot be conclusively ruled out as alternative explanations for observed psychological differences between recruitment groups. Future researchers could record device type and administration context (eg, on-site with staff administration, on-site without staff administration, or offsite) to directly test for such effects.

Finally, given the large sample size involved and the number of comparisons conducted, some significant differences might have been of limited practical importance [52] or due to random chance [53]. On the other hand, key differences between the groups (eg, age) had effect size CIs indicating that the magnitude of the group difference was medium to large, in addition to having very low *P* values, and the FDR-adjusted analyses for Table 1 suggested that nearly all significant results were not chance occurrences resulting from having conducted multiple comparisons.

XSL•FO
RenderX

Conclusions

Prior to this secondary analysis, differences in the baseline characteristics of participants enrolled in person versus online, and differences in recruitment pace across methods, had not been examined in a smoking cessation RCT for people with food insecurity. Nonetheless, it is important to stress that this study is a descriptive comparison within a specific trial context, rather than strong evidence for the comparative advantage of one recruitment modality over another. Regardless, using both recruitment strategies when conducting smoking cessation intervention trials for people with food insecurity may broaden reach, although, as seen in this RCT, the pace of enrollment might differ considerably between the 2 approaches. Additionally, the use of both online and in-person strategies might yield a sample that more closely reflects the underlying population of individuals who smoke and are food insecure. This study expands the RCT recruitment strategy evidence base by providing concrete results to inform researchers designing and conducting smoking cessation trials for special populations, particularly people with food insecurity. Furthermore, examining whether treatment effects and follow-up assessment completion status are influenced by recruitment modality will be important to understanding the potential substantive implications of recruiting from this population via online versus in-person methods.

Acknowledgments

We are grateful to Feeding Tampa Bay for their partnership on this project.

Funding

This research was supported by a National Cancer Institute grant (R01CA231952) awarded to principal investigators DJV and JIV. This research was also supported by the Biostatistics and Bioinformatics Shared Resource at Moffitt Cancer Center (P30CA76292). The funders had no role in the design, conduct, analysis, interpretation, or reporting of this secondary analysis.

Data Availability

Deidentified participant data, the data dictionary, and analysis code will be made available upon request to the first or senior authors, contingent on institutional requirements and execution of a data use agreement.

Authors' Contributions

CEH contributed to conceptualization, data curation, formal analysis, visualization, and writing of the original draft, as well as review and editing of the manuscript. SKS contributed to data curation, formal analysis, writing of the original draft, and manuscript review and editing. SRJ contributed to project administration, data curation, and manuscript review and editing. BN, SJB, DH, TM, MSB, and YCTS contributed to manuscript review and editing. JIV contributed to conceptualization, methodology, formal analysis, funding acquisition, and manuscript review and editing and served as co-senior author with DJV. DJV contributed to conceptualization, methodology, formal analysis, funding acquisition, and manuscript review and editing and served as co-senior author with JIV.

Conflicts of Interest

None declared.

Multimedia Appendix 1
R Code for obtaining 95% CI for ΔR^2.
[DOCX File, 16 KB - jmir_v28i1e80530_app1.docx]

Checklist 1
CONSORT 2025 checklist.
[PDF File, 231 KB - jmir_v28i1e80530_app2.pdf]

Checklist 2
CONSORT E-HEALTH checklist.
[PDF File, 457 KB - jmir_v28i1e80530_app3.pdf]

References

1. Health UDo, Services H. Smoking Cessation: A Report of the Surgeon General: US Department of Health and Human Services; 2020. [Medline: 32255575]
2. Meza R, Cao P, Jeon J, Warner KE, Levy DT. Trends in US adult smoking prevalence, 2011 to 2022. JAMA Health Forum 2023 Dec 1;4(12):e234213. [doi: 10.1001/jamahealthforum.2023.4213] [Medline: 38038988]

XSL•FO
RenderX

3. Farrelly MC, Shafer PR. Comparing trends between food insecurity and cigarette smoking among adults in the United States, 1998 to 2011. Am J Health Promot 2017 Sep;31(5):413-416. [doi: 10.1177/0890117116660773] [Medline: 27493199]
4. Berkowitz SA, Seligman HK, Palakshappa D. Understanding food insecurity risk in the United States: a longitudinal analysis. SSM Popul Health 2024 Mar;25:101569. [doi: 10.1016/j.ssmph.2023.101569] [Medline: 38156292]
5. Kim-Mozeleski JE, Pandey R. The intersection of food insecurity and tobacco use: a scoping review. Health Promot Pract 2020 Jan;21(1_suppl):124S-138S. [doi: 10.1177/1524839919874054] [Medline: 31908208]
6. Martinez E, Tatum KL, Weber DM, et al. Issues related to implementing a smoking cessation clinical trial for cancer patients. Cancer Causes Control 2009 Feb;20(1):97-104. [doi: 10.1007/s10552-008-9222-x] [Medline: 18758971]
7. Tundealao S, Oubda MC, Klaff R, Reininger B, McNeill L, Tamí-Maury I. Pragmatic recruitment strategies for a text-based smoking cessation intervention for sexual and gender minority groups. Discov Public Health 2025;22(1):743. [doi: 10.1186/s12982-025-01117-0] [Medline: 41306822]
8. Arana-Chicas E, Cartujano-Barrera F, Rieth KK, et al. Effectiveness of recruitment strategies of Latino smokers: secondary analysis of a mobile health smoking cessation randomized clinical trial. J Med Internet Res 2022 Jun 27;24(6):e34863. [doi: 10.2196/34863] [Medline: 35759320]
9. Fennell BS, Jones SR, Sutton SK, et al. In-clinic versus online recruitment of women with a history of cervical intraepithelial neoplasia or cervical cancer to a smoking cessation trial: a post hoc comparison of participant characteristics, study retention, and cessation outcomes. Nicotine Tob Res 2024 Aug 22;26(9):1264-1270. [doi: 10.1093/ntr/ntae049] [Medline: 38452212]
10. Vidrine JI, Sutton SK, Wetter DW, et al. Efficacy of a smoking cessation intervention for survivors of cervical intraepithelial neoplasia or cervical cancer: a randomized controlled trial. J Clin Oncol 2023 May 20;41(15):2779-2788. [doi: 10.1200/JCO.22.01228] [Medline: 36921237]
11. Kim-Mozeleski JE, Smell A, Castele MC, Ogden E, Trapl ES. Assessing the feasibility of conducting smoking cessation outreach in food pantries: a pilot intervention study. Nicotine Tob Res 2024 Jan 1;26(1):46-53. [doi: 10.1093/ntr/ntad137] [Medline: 37531409]
12. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
13. Vidrine JI, Shih YCT, Businelle MS, et al. Comparison of an automated smartphone-based smoking cessation intervention versus standard quitline-delivered treatment among underserved smokers: protocol for a randomized controlled trial. BMC Public Health 2022 Mar 22;22(1):563. [doi: 10.1186/s12889-022-12840-7] [Medline: 35317789]
14. Service USDoAER. US Household Food Security Survey Module: Six-Item Short Form: US Department of Agriculture, Economic Research Service Washington; 2012.
15. Kelley K. Confidence intervals for standardized effect sizes: theory, application, and implementation. J Stat Soft ;20(8):N/A. [doi: 10.18637/jss.v020.i08]
16. Tomczak M, Tomczak E. The need to report effect size estimates revisited. an overview of some recommended measures of effect size. TRENDS Sport Sci 2014;1(21):19-25 [FREE Full text]
17. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Lawrence Erlbaum Associates, Publishers; 1988.
18. Mangiafico SS. Summary and Analysis of Extension Program Evaluation in R, Version 1241, Revised 2026: Rutgers Cooperative Extension. URL: https://rcompanion.org/handbook/ [accessed 2026-01-09]
19. Banjanovic ES, Osborne JW. Confidence intervals for effect sizes: applying bootstrap resampling. Practical Assessment, Research, and Evaluation 2016;21(1):1-18. [doi: 10.7275/dz3r-8n08]
20. Rousselet GA, Pernet CR, Wilcox RR. The percentile bootstrap: a primer with step-by-step instructions in R. Advances in Methods and Practices in Psychological Science 2021 Jan;4(1):2515245920911881. [doi: 10.1177/2515245920911881]
21. Szumilas M. Explaining odds ratios. J Can Acad Child Adolesc Psychiatry 2010 Aug;19(3):227-229. [Medline: 20842279]
22. Little RJA. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 1988 Dec;83(404):1198-1202. [doi: 10.1080/01621459.1988.10478722]
23. Cohen J, Cohen P, West SG, Aiken LS. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences: Lawrence Erlbaum Associates; 2003.
24. Kelley K. Methods for the behavioral, educational, and social sciences: an R package. Behav Res Methods 2007 Nov;39(4):979-984. [doi: 10.3758/bf03192993] [Medline: 18183915]
25. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. Lancet 2025 Apr 14:S0140-6736(25)00672-5. [doi: 10.1016/S0140-6736(25)00672-5] [Medline: 40245901]
26. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
27. Understanding alcohol drinking patterns. NIoAAA. URL: https://www.niaaa.nih.gov/alcohols-effects-health/alcohol-drinking-patterns [accessed 2025-01-31]
28. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B 1995 Jan 1;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]

XSL•FO
RenderX

29. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Statist 2001;29(4):1165-1188. [doi: 10.1214/aos/1013699998]
30. Buuren S, Groothuis-Oudshoorn K, Robitzsch A, Vink G, Doove L, Jolani S. mice: multivariate imputation by chained equations (computer software). Comprehensive R Archive Network. 2015. URL: https://cran.r-project.org/web/packages/mice/index.html [accessed 2026-06-06]
31. van Wijk EC, Landais LL, Harting J. Understanding the multitude of barriers that prevent smokers in lower socioeconomic groups from accessing smoking cessation support: a literature review. Prev Med 2019 Jun;123:143-151. [doi: 10.1016/j.ypmed.2019.03.029] [Medline: 30902700]
32. Hiscock R, Bauld L, Amos A, Fidler JA, Munafò M. Socioeconomic status and smoking: a review. Ann N Y Acad Sci 2012 Feb;1248(1):107-123. [doi: 10.1111/j.1749-6632.2011.06202.x] [Medline: 22092035]
33. Denney JT, Sharp G, Kimbro RT. Community social environments and cigarette smoking. SSM Popul Health 2022 Sep;19:101167. [doi: 10.1016/j.ssmph.2022.101167] [Medline: 35879966]
34. Rizvi A, Wasfi R, Enns A, Kristjansson E. The impact of novel and traditional food bank approaches on food insecurity: a longitudinal study in Ottawa, Canada. BMC Public Health 2021 Apr 22;21(1):771. [doi: 10.1186/s12889-021-10841-6] [Medline: 33882881]
35. Reitzel LR, Kendzor DE, Castro Y, et al. The relation between social cohesion and smoking cessation among Black smokers, and the potential role of psychosocial mediators. Ann Behav Med 2013 Apr;45(2):249-257. [doi: 10.1007/s12160-012-9438-6] [Medline: 23135831]
36. Haim-Litevsky D, Komemi R, Lipskaya-Velikovsky L. Sense of belonging, meaningful daily life participation, and well-being: integrated investigation. Int J Environ Res Public Health 2023 Feb 25;20(5):4121. [doi: 10.3390/ijerph20054121] [Medline: 36901132]
37. An R, Wang J, Liu J, Shen J, Loehmer E, McCaffrey J. A systematic review of food pantry-based interventions in the USA. Public Health Nutr 2019 Jun;22(9):1704-1716. [doi: 10.1017/S1368980019000144] [Medline: 30834852]
38. Cook S, Curtis J, Buszkiewicz JH, Brouwer AF, Fleischer NL. Financial strain and smoking cessation and relapse among U.S. adults who smoke: a longitudinal cohort study. Am J Prev Med 2025 Jan;68(1):164-171. [doi: 10.1016/j.amepre.2024.09.012] [Medline: 39293702]
39. Reitzel LR, Mazas CA, Cofta-Woerpel L, et al. Subjective social status affects smoking abstinence during acute withdrawal through affective mediators. Addiction 2010 May;105(5):928-936. [doi: 10.1111/j.1360-0443.2009.02875.x] [Medline: 20219054]
40. Weinberger AH, Kashan RS, Shpigel DM, et al. Depression and cigarette smoking behavior: a critical review of population-based studies. Am J Drug Alcohol Abuse 2017 Jul;43(4):416-431. [doi: 10.3109/00952990.2016.1171327] [Medline: 27286288]
41. Wootton RE, Greenstone HSR, Abdellaoui A, et al. Bidirectional effects between loneliness, smoking and alcohol use: evidence from a Mendelian randomization study. Addiction 2021 Feb;116(2):400-406. [doi: 10.1111/add.15142] [Medline: 32542815]
42. Piñeiro B, López-Durán A, Del Río EF, Martínez Ú, Brandon TH, Becoña E. Motivation to quit as a predictor of smoking cessation and abstinence maintenance among treated Spanish smokers. Addict Behav 2016 Feb;53:40-45. [doi: 10.1016/j.addbeh.2015.09.017] [Medline: 26441045]
43. Tourangeau R, Maitland A, Rivero G, Sun H, Williams D, Yan T. Web surveys by smartphone and tablets: effects on survey responses. Public Opin Q 2017;81(4):896-929. [doi: 10.1093/poq/nfx035]
44. Wenz A, Keusch F. Increasing the acceptance of smartphone-based data collection. Public Opin Q 2023;87(2):357-388. [doi: 10.1093/poq/nfad019] [Medline: 37457396]
45. Di Palo MP, Di Spirito F, Garofano M, et al. Effectiveness and adherence of standalone digital tobacco cessation modalities: a systematic review of systematic reviews. Healthcare (Basel) 2025 Aug 26;13(17):2125. [doi: 10.3390/healthcare13172125] [Medline: 40941478]
46. Cobos-Campos R, Cordero-Guevara JA, Apiñaniz A, et al. The impact of digital health on smoking cessation. Interact J Med Res 2023 Mar 15;12:e41182. [doi: 10.2196/41182] [Medline: 36920468]
47. Holford TR, McKay L, Jeon J, et al. Smoking histories by state in the U.S. Am J Prev Med 2023 Apr;64(4 Suppl 1):S42-S52. [doi: 10.1016/j.amepre.2022.08.018] [Medline: 36653233]
48. Brøgger-Mikkelsen M, Ali Z, Zibert JR, Andersen AD, Thomsen SF. Online patient recruitment in clinical trials: systematic review and meta-analysis. J Med Internet Res 2020 Nov 4;22(11):e22179. [doi: 10.2196/22179] [Medline: 33146627]
49. Hung M, Mohajeri A, Almpani K, et al. Successes and challenges in clinical trial recruitment: the experience of a new study team. Med Sci (Basel) 2024 Aug 14;12(3):39. [doi: 10.3390/medsci12030039] [Medline: 39189202]
50. Albarracín D, Dai W. The impact of the environment on behavior. In: Advances in Experimental Social Psychology Elsevier: Elsevier; 2024:151-201.
51. Heerwegh D. Mode differences between face-to-face and web surveys: an experimental investigation of data quality and social desirability effects. Int J Public Opin Res 2009 Mar 1;21(1):111-121. [doi: 10.1093/ijpor/edn054]
52. Wasserstein RL, Lazar NA. The ASA statement on P -values: context, process, and purpose. Am Stat 2016 Apr 2;70(2):129-133. [doi: 10.1080/00031305.2016.1154108]

XSL•FO
RenderX

53. Cao J, Zhang S. Multiple comparison procedures. JAMA 2014 Aug 6;312(5):543-544. [doi: 10.1001/jama.2014.9440] [Medline: 25096694]

Abbreviations

AT: automated treatment
BCa: bias corrected and accelerated
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
EBT: electronic benefits transfer
FDR: false discovery rate
MCAR: missing completely at random
OR: odds ratio
RCT: randomized controlled trial
REDCap: Research Electronic Data Capture
ST: standard treatment

Edited by S Brini; submitted 11.Jul.2025; peer-reviewed by JM Unger, W Guo; revised version received 22.Apr.2026; accepted 27.Apr.2026; published 11.Jun.2026.

Please cite as:
Hoogland CE, Sutton SK, Jones SR, Nagy B, Brockway SJ, Himmelgreen D, Mantz T, Businelle MS, Shih YCT, Vidrine JI, Vidrine DJ
Sociodemographic Paradoxes and Enrollment Differences in In-Person Versus Online Recruitment to a Mobile Health Smoking Cessation Intervention for Food-Insecure Adults: Secondary Analysis of a Randomized Controlled Trial
J Med Internet Res 2026;28:e80530
URL: https://www.jmir.org/2026/1/e80530
doi:10.2196/80530

© Charles E Hoogland, Steven K Sutton, Sarah R Jones, Bence Nagy, Samuel J Brockway, David Himmelgreen, Thomas Mantz, Michael S Businelle, Ya-Chen Tina Shih, Jennifer I Vidrine, Damon J Vidrine. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Willingness to Share Internet Use Data for Research on Early Disease Detection: Cross-Sectional Survey

Christina Derksen[1], BSc, MSc, PhD; Mel Ramasawmy[1], BSc, MSc, PhD; Sanjula Arora[1], BSc, MSc; Wynsee Lau[2]; Suzanne E Scott[1], BSc, MSc, PhD

[1]Wolfson Institute of Population Health, Queen Mary University of London, Charterhouse Square, London, United Kingdom

[2]Faculty of Medicine, Imperial College London, London, United Kingdom

Corresponding Author:
Christina Derksen, BSc, MSc, PhD
Wolfson Institute of Population Health, Queen Mary University of London, Charterhouse Square, London, United Kingdom

Abstract

Background: Preliminary research has suggested that internet use data could offer digital signals of early disease and has the potential to facilitate early detection and improve patient outcomes. However, there are significant challenges in linking individual-level internet use data with health outcomes. One key aspect is that the public might not be willing to share data for research or that selective data sharing might create bias in datasets and increase inequalities.

Objective: Our study aimed to investigate the willingness of the public to share their internet use data for medical research and to identify key criteria that affect willingness to share.

Methods: We conducted a web-based, cross-sectional online survey with 2390 UK adults with and without a history of cancer, heart disease, and depression using quota sampling. Participants were randomly assigned to explore willingness to share different types of internet use data for 1 of 3 health conditions (cancer, heart disease, and depression) and for provision of a pictorial example of internet use data. Logistic regression analysis (α=.05) for each condition was used to determine key factors of willingness to share, including sociodemographics and attitudes toward sharing. Open-ended comments regarding facilitators of sharing and concerns were analyzed thematically.

Results: Willingness to share internet use data was high across conditions (74% - 77%, 95% CI 70.5% - 80.3%), especially for health app data (73% - 76%, 95% CI 69.8% - 79.1%). The pictorial example of browsing history did not affect willingness to share. For all conditions, factors consistently associated with willingness to share were perceived benefits (odds ratios [ORs] 5.692 - 8.850; all P<.001) and concerns (ORs 0.343 - 0.432; all P<.001). Key concerns were data privacy, potential for misuse, and lack of relevance. Suggestions to increase willingness to share included contributing to society and research, data security assurances, clarification of research purposes, and monetary incentives. Familiarity with internet use data was related to lower willingness to share for heart disease detection (OR 0.740, 95% CI 0.561 - 0.976). Asian ethnicity was associated with lower willingness to share internet use for cancer detection (OR 0.234, 95% CI 0.076 - 0.723). Younger age (OR 0.975, 95% CI 0.951 - 0.999) and male gender (OR 2.615, 95% CI 1.511 - 4.526) were associated with higher willingness to share data for depression detection.

Conclusions: This first large-scale assessment of public willingness to share internet use data for early disease detection adds novel insights by comparing conditions and examining sociodemographic factors alongside perceived benefits and risks. It highlights that understanding of internet data is limited yet willingness to share for research is high. Clear communication of benefits, strong privacy protections, and incentives may increase participation and reduce bias. The findings inform consent design, targeted outreach to underrepresented groups, and data governance for safe use of personal digital data. Future research should focus on improving public communication, particularly among less willing groups at risk of inequality.

(J Med Internet Res 2026;28:e85637) doi:10.2196/85637

KEYWORDS

early detection; internet use data; willingness to share online data; public perceptions; data privacy

Introduction

Across many health conditions, early diagnosis is crucial for successful treatment and long-term outcomes, potentially impacting survival and health care costs [1-3]. Early diagnosis depends on detection of signs and symptoms, presentation to health care professionals, and timely diagnostic tests. However, patients often present to health care weeks or months after the onset of symptoms, and a key reason for this delay is the time taken to interpret symptoms as potentially serious [4,5]. Thus,

XSL•FO
RenderX

the understanding of timing of symptom perception, symptom interpretation, and help-seeking behavior is vital in efforts to encourage early diagnosis.

"Digital footprints" capture huge amounts information, including online activity, in-store purchases, wearable device data, and social media engagement [6]. These digital data may provide insights into symptoms, changes in behavior, and the decision to seek help [7,8], and using these data may tackle current research limitations in this area. To date, researchers studying time to presentation have relied on retrospective self-report from patients recently diagnosed with the condition under investigation [9]. These data are subject to recall bias and could be affected by hindsight from the recent diagnosis [10,11].

In the United Kingdom, there are almost 50 million health-related searches per year using Google [12]. Internet use data include internet search history, calendar data, bookmarks, contacts, emails, photos, reminders, profiles, and documents stored online. Given the richness of these data, there is growing interest to explore the use of internet search data and other digital footprints as digital signals of serious disease. Preliminary research has begun to examine whether data on internet search history could be used to predict diagnosis of cancers and mental health conditions, with promising results. Changes in the frequency, content, and tone of online searches were evident up to 60 days before suicide attempts or general practitioner referral for gynecological cancers, up to 21 weeks before pancreatic cancer diagnosis, and up to 1 year before psychiatric hospitalization [13-16].

Despite the promise of the digital signals, there are multiple questions that need to be explored regarding the acceptability of this kind of research for patients and the public, especially if research involves sharing internet use data on an individual level and linking with future health outcomes such as a cancer diagnosis. Previous studies have mixed results regarding people's willingness to share different types of digital data over "traditional" types of health data such as biomedical samples or questionnaire data [17,18]. It is unclear if internet users are aware that these data exist, where they are stored, and how they could be used in research or if effective ways of explaining these issues to potential participants exist. This is key to informed consent.

Willingness to share might depend on several factors, including the type of data being shared, the way internet data are explained, the purpose of the research, privacy concerns, and sociodemographic factors [19]. If some groups are less willing to provide internet use data, findings from future research might not be representative and could thus perpetuate health inequalities. For example, gender, age, ethnicity, or socioeconomic status such as employment might influence willingness to share internet use data and other types of data such as location data or measures from health and fitness apps [19-21]. Further, the type of condition being studied and the relevance to an individual (eg, through personal history or experiences of friends and family being affected by a mental or physical health condition) can influence willingness to participate in medical research [22], yet it is unknown if this applies to willingness to share internet use data.

This study therefore investigated public perspectives on the use of internet use data for medical research to identify key criteria that could enhance acceptability and avoid inequalities in engagement. Our research questions were exploratory:

1. How willing are the public to share different types of internet use data for early diagnosis research?
2. Does willingness to share internet use data vary for different mental and physical health conditions?
3. Does the way internet use data is explained impact willingness to share internet use data?
4. Are sociodemographic characteristics, personal history of different mental and physical health conditions, familiarity with internet use data, and attitudes toward data sharing associated with willingness to share internet use data?

Methods

Study Design

We used a cross-sectional design to conduct an online survey on willingness to share different types of internet use data. We randomly assigned participants using an automated algorithm based on 2 factors: (1) the medical condition about which participants are asked whether they would be willing to share their data (cancer, heart disease, and depression) and (2) the way internet use data was explained (written text plus a screenshot as a pictorial example of internet use data vs written text only). Cancer, depression, and heart disease were selected, as these are common conditions for which early diagnosis is key and there is emerging evidence for use of digital data in prediction models [13-16]. We reported the study in line with Journal Article Reporting Standards Quantitative Research Design (JARS-Quant) standards [23].

Ethical Considerations

Ethical approval for research with human subjects was granted by the Queen Mary Ethics of Research Committee (reference number QME25.0924). Participants received the participant information sheet for download at the beginning of the survey and provided informed consent for participation and data sharing before they could access the survey. All data were collected anonymously. In line with Dynata's company policies, participants received tokens through the loyalty programs set up through Dynata partnerships automatically upon survey completion, which could be redeemed for vouchers or gifts. Incentives were valued at approximately £0.75 (US $1.00). No identification of individual participants in any images of the manuscript or appendices is possible.

Participants and Procedure

Participants were adult members of the public (18 years or older), both with and without a personal history of cancer, heart disease, or depression. The survey was hosted on the online platform Qualtrics. Participants were recruited through Dynata from participants who had previously signed up to an online panel from July 1, 2025, to July 23, 2025. As the panel was English-speaking and conducted initial checks with new participants, we assumed all participants would be proficient in the English language. Before full recruitment, the survey was pilot tested with patient and public representatives (see the

XSL•FO
RenderX

following sections). To ensure data quality, it was then released to 10% of the final population to prevent technical problems and unclear questions before full roll-out.

The survey was administered using quota sampling based on gender (approximately 50% male and 50% female), age (approximately 50% of the population over the age of 50 years), income (approximately 50% with disposable household income levels <£35,000), and ethnicity (maximum 85% of the population from a White ethnic group). A minimum of 300 participants with a history of cancer, heart disease, or depression was sought. We chose to sample for age to increase the proportion of older participants, who have a higher incidence of cancer and cardiovascular disease. Online survey panels often have a disproportionately higher share of younger respondents [24], and we anticipated that younger participants would be more likely to take part in a survey on the topic of internet use and data.

Measures

The survey assessed willingness to share data for early diagnosis research, attitudes toward data sharing, familiarity about internet use data, and sociodemographics. All items and the pictorial example can be found in Multimedia Appendix 1.

Willingness to Share Internet Use Data for Early Diagnosis Research and Attitudes Toward Data Sharing

The main outcome measure was willingness to share data, assessed as a single item adapted from Hirst et al [19] with 4 answer options (definitely yes to definitely no). Where appropriate, this was recoded to a binary variable (yes, no) for analyses. The same question was used to assess willingness to share specific types of internet use data (eg, internet search history or online purchases), with an option to choose "Not applicable (I do not use this)."

Open-ended questions were used to ask participants what would increase their willingness to share internet use data, as well as their concerns about data sharing.

Attitudes toward data sharing were assessed as perceived benefits and concerns or worries using an adaptation of scales by Sanderson et al [24] and Seltzer et al [18]. We used matrix questions to include perceived benefits and concerns and calculated mean scores (Cronbach α=0.96 for perceived benefits and 0.94 for concerns).

Familiarity With Internet Use

We assessed participants' self-reported familiarity with internet use data by adapting the scale by Hirst et al [19] on awareness about General Data Protection Regulation (GDPR). We asked whether participants had heard about internet use data before. Objective knowledge was assessed by asking what participants thought was part of internet use data from a list of examples (eg, internet search history as example of internet use data or content of emails as distractor). Correct answers were counted to calculate a knowledge score (range 1 - 10).

Sociodemographics and Current Internet Use

For sociodemographic characteristics, we assessed age, gender, educational level, employment, marital status, ethnicity, and preferred language. Items were adapted from the DISTINCT project [25]. We assessed disposable household income (<£35,000 or >£35,000) and postcodes (to determine levels of deprivation via English indices of deprivation 2019 [26], Postcode Deprivation Finder 2019 [27], and Multiple Deprivation Measures 2017 Lookup Tool [28]). Finally, we measured self-reported history of cancer, cardiovascular disease, and depression and whether their relatives had a history of these conditions. Answer options were adapted from the Health Survey for England, 2021 [29].

To assess current internet and online service use, we used 2 questions from the UK Digital Landscape survey [30] about frequency and purpose of internet use.

Data Preparation

There were no missing data in the study variables such as willingness to share internet use data, self-reported familiarity with internet use data, and perceived benefits or concerns, as items were set to mandatory. In the sociodemographic data, <1% were missing, except for disposable household income (2.3%) and deprivation quintile (7.8%). Online surveys are open to fraudulent responses [31,32]; as a precaution, we introduced a number of steps to ensure credibility of the data. An attention check was added halfway through the survey, asking participants to select "strongly agree" with answer options appearing in a random order. Participants were excluded if they failed the attention check. We also excluded participants who were flagged by Qualtrics as both likely bot *and* fraudulent respondent (Recaptcha Score<0.5 and Fraud Score>30) or as duplicate (flagged by Qualtrics or on inspection of the data), as well as participants who answered the survey in less than 2 minutes, which was considered to be insufficient time to read through the participant information sheet and questionnaire items. Finally, we excluded participants with the common placeholder "lorem ipsum" text or questionable open answers (eg, "Banana") on both open-ended questions.

Data Analysis

We descriptively report differences in the willingness to share internet use data for early diagnosis research using percentages of participants indicating they would be willing to share. We conducted χ^2 tests for differences in proportions willing to share internet use data between medical conditions, the way internet use data were explained, and history of the condition. Logistic regression analysis was used to determine how sociodemographic characteristics, perceived benefits and concerns, knowledge, and familiarity with internet use data affect willingness to share internet use data, coded as a binary outcome. Separate regression models were run for willingness to share internet use data for each medical condition (cancer, heart disease, depression).

Responses to open-ended questions were analyzed using a thematic content analysis approach and reported alongside how willing participants were to share internet use data. Categories were developed by two authors (CD and WL) and discussed with the research team including patient and public involvement (PPI) representatives. Answers not in English (one Romanian and one Polish) were translated using Google Translate, as

XSL•FO
RenderX

translated content fit the questions, showing that participants understood English but answered in a different preferred language.

The full syntax and the final datasets can be found on Open Science Framework (OSF) [33].

Power Analysis

Based on previous research [19], we made the conservative assumption that willingness to share internet use data would likely vary between 45% and 55% of participants being willing to share. For an a priori sample size calculation, we therefore assumed a difference of 0.1 in probabilities of being willing to share data or not for a binary factor (0.55 in one group, 0.45 in the other; odds ratio [OR] 1.49). To reach a power of 80% given an α-error level of .05, we determined a sample size of 786 participants per regression. As logistic regression analysis was repeated for each condition, we aimed for a sample size of 2358 participants (3 × 786). However, the sample sizes for the logistic regression analyses were lower than the targeted sample size, so analyses were repeated without the variable with the most missing data (level of deprivation; see the following sections) for sensitivity analysis.

Patient and Public Involvement

We presented initial ideas for survey items to 3 PPI representatives who were older than 55 years and from a minority background. They were recruited from the Centre for Cancer Screening, Prevention and Early Diagnosis PPI pool at Queen Mary University of London, which is a database of more than 150 people with an interest in contributing to research. They endorsed the research but asked to add questions about emotions (fear and embarrassment) to the items on attitudes toward data sharing. The group asked that the participant information sheet emphasized that data sharing was not required for participation in this study and to change the wording of some items to facilitate understanding. Following these changes, we piloted the survey with 2 PPI representatives in think-aloud sessions to ensure face validity [34]. The PPI group was also involved in interpreting the results and discussing implications for future research and practice.

Results

Participant Characteristics

Of the 2864 participants who gave informed consent and completed the survey, 474 were excluded: 365 failed the attention check, 32 were flagged as both likely bot and fraudulent respondent, 64 were duplicate respondents, 6 completed the survey in less than 2 minutes, and 7 gave questionable open answers. Hence, 2390 participants with a mean age of 50.2 (SD 16.86, range 18 - 90) years were included in the analyses. Participant characteristics are provided in Table 1.

XSL•FO
RenderX

Table . Participant characteristics from a cross-sectional survey in 2025 conducted with UK participants recruited through an online survey panel (N=2390).

Characteristic	Results, n (%)
Gender	
Man	1250 (52.3)
Woman	1135 (47.5)
Nonbinary	4 (0.2)
Prefer not to say	1 (<0.1)
Employment status	
Employed (including self-employed)	1433 (60)
Not employed	352 (14.7)
Retired	592 (24.8)
Prefer not to say	13 (0.5)
Highest educational qualifications	
No formal qualifications	73 (3.1)
GCSEs[a] or equivalent	408 (17.1)
AS[b], A level or equivalent	296 (12.4)
NVQ[c] up to level 3 or equivalent	256 (10.7)
Qualification at degree level or above	1250 (52.3)
Apprenticeship	51 (2.1)
Other or equivalent unknown	52 (2.2)
Prefer not to say	4 (0.2)
Disposable household income (£)	
<35,000	1206 (50.5)
>35,000	1127 (47.2)
Prefer not to say	57 (2.3)
Deprivation quintile based on postcode	
1 (most deprived)	464 (19.4)
2	522 (21.8)
3	425 (17.8)
4	385 (16.1)
5 (least deprived)	408 (17.1)
Missing	186 (7.8)
Marital status	
Single	662 (27.7)
Married or in legal relationship	1368 (57.2)
Widowed, separated or divorced	348 (14.6)
Prefer not to say	12 (0.5)
Ethnicity	
White	1978 (82.8)
Asian or Asian British	127 (5.3)
Black, Black British, Caribbean or African	216 (9)
Mixed or multiple ethnic groups	51 (2.1)
Other	8 (0.3)

XSL•FO
RenderX

Characteristic	Results, n (%)
Prefer not to say	10 (0.4)
Preferred language	
English	2376 (99.4)
Other	12 (0.5)
Prefer not to say	2 (0.1)
Personal history of medical conditions	
Cancer	398 (16.7)
Heart disease or vascular problems	422 (17.7)
Depression	745 (31.2)
None of the above	1047 (43.8)
Not sure/not formally diagnosed	126 (5.3)
Prefer not to say	16 (0.7)
Relatives' history of medical conditions	
Cancer	656 (27.4)
Heart disease or vascular problems	633 (26.5)
Depression	535 (22.4)
None of the above	1005 (42.1)
Not sure/not formally diagnosed	159 (6.7)
Prefer not to say	24 (1)
Knowledge[d]	
Internet search history	2143 (89.7)
Internet browsing history	2141 (89.6)
Health app data	1717 (63.5)
Online banking details (incorrect)	1021 (42.7)
App store data	1490 (62.3)
Content of emails (incorrect)	1050 (43.9)
Purchases made online	1650 (69)
Location history	1831 (76.6)
Internet searches in "incognito" mode (incorrect)	813 (34)
Video streaming history	1720 (72)
Self-reported familiarity with internet use data	
Not familiar	792 (33.1)
Heard about internet use data	850 (35.6)
Some knowledge	512 (21.4)
Extensive knowledge	236 (9.9)

[a]GCSE: General Certificate of Secondary Education.

[b]AS: Advanced Subsidiary.

[c]NVQ: National Vocational Qualification.

[d]Number and percentage of participants indicating this is part of internet use data.

Willingness to Share Internet Use Data

There were no differences in proportions of respondents willing to share internet use data between the 3 medical conditions (for early detection of cancer=77%; heart disease=74%; depression=74%). When asked about different types of data, participants were most willing to share their health app data (73% - 76%, 95% CI 69.8% - 79.1%), whereas participants were less willing to share other types of data, especially regarding purchases made online (52% - 57%, 95% CI

XSL•FO
RenderX

48.7% - 60.3%) and their location history (58% - 59%, 95% CI 54.2% - 62.2%). Details are provided in Table 2. There was variability in what participants understood to be a part of internet use data, with more than 40% of participants incorrectly believing that the content of their emails (1050/2390, 43.9%) and online banking history (1021/2390, 42.7%) would be included (see Table 1).

Table . Willingness to share internet use data for research on different medical conditions from a cross-sectional survey in 2025 with UK participants recruited through an online survey panel.

Type	Willingness to share internet use data for research on early detection of…[a]									Comparison, χ^2 (df)	*P* value
	Cancer			Heart disease			Depression				
	Total users, n	Users willing to share, n (%)	95% CI	Total users, n	Users willing to share, n (%)	95% CI	Total users, n	Users willing to share, n (%)	95% CI		
Internet use data in general	795	615 (77.4)	74.4 - 80.3	798	587 (73.6)	70.5 - 76.6	797	592 (74.3)	71.2 - 77.3	4.64 (2,2390)	.18
Internet search history	779	532 (68.3)	65.0 - 72.6	777	509 (65.5)	62.2 - 68.9	787	541 (68.7)	65.5 - 72.0	2.18 (2,2343)	.34
Internet browsing history	784	519 (66.2)	62.9 - 69.5	783	493 (63.0)	59.6 - 66.4	792	535 (67.6)	64.3 - 70.8	3.87 (2,2359)	.14
Health app data	682	498 (73.0)	69.7 - 76.4	670	508 (75.8)	72.6 - 79.1	688	503 (73.1)	69.8 - 76.4	1.78 (2,2040)	.41
App store data	730	446 (61.1)	57.6 - 64.6	732	455 (62.2)	58.6 - 65.7	748	490 (65.5)	62.1 - 68.9	3.37 (2,2210)	.19
Purchases made online	770	402 (52.2)	48.7 - 55.7	777	414 (53.3)	49.8 - 56.8	788	448 (56.9)	53.4 - 60.3	3.72 (2,2335)	.16
Location history	771	453 (58.8)	55.3 - 62.2	768	443 (57.7)	54.2 - 61.2	774	447 (57.8)	54.3 - 61.2	0.23 (2,2313)	.89
Video streaming history	750	463 (61.7)	58.3 - 65.2	751	471 (62.7)	59.3 - 66.2	756	499 (66.0)	62.6 - 69.4	3.26 (2,2257)	.20

[a]Total numbers and percentages for willingness to share include valid participants who are using these services and answered, “definitely yes” and “probably yes.”

Participants were descriptively less willing to share their internet use data if a pictorial example of internet use history was provided alongside written text compared with written text only, but this did not reach statistical significance for research on any of the 3 medical conditions (see Table 3).

Having a personal or family history of the condition being researched was not associated with willingness to share internet use history for that condition, with the exception of those with a personal history of depression. Participants who had a history of depression were more likely to be willing to share their internet use data for the early detection of depression (79.7%, 95% CI 74.5% - 84.8%) than participants without a history of depression (72.2%, 95% CI 67.8% - 76.5%; *P*=.03; Table 3).

XSL•FO
RenderX

Table . Willingness to share internet use data in relation to how the internet search history was described and personal and family history of the medical condition being researched from a cross-sectional survey in 2025 with UK participants recruited through an online survey panel.

Characteristic	Willingness to share internet use data for research on early detection of…[a]											
	Cancer				Heart disease				Depression			
	N (%)	95% CI	Comparison, χ^2 (df)	*P* value	N (%)	95% CI	Comparison, χ^2 (df)	*P* value	N (%)	95% CI	Comparison, χ^2 (df)	*P* value
Description of internet use			3.56 (1,795)	.06			2.05 (1,798)	.15			2.02 (1,797)	.16
Written description	312 (80.2)	76.2 - 84.2			298 (75.8)	71.6 - 80.1			294 (76.6)	72.3 - 80.8		
Pictorial example and written description	303 (74.6)	76.2 - 84.2			289 (71.4)	66.9 - 75.8			298 (72.2)	67.8 - 76.5		
Personal history of medical condition being researched			3.60 (1,790)	.058			0.36 (1,790)	.55			4.82 (1,794)	.03
Yes	116 (83.5)	77.2 - 89.7			103 (75.7)	68.4 - 83.0			188 (79.7)	74.5 - 84.8		
No	495 (76.0)	72.8 - 79.3			479 (73.2)	69.8 - 76.6			403 (72.2)	68.5 - 76.0		
Family history of medical condition being researched			3.17 (1,792)	.08			0.95 (1,789)	.33			3.17 (1,792)	.08
Yes	171 (81.8)	76.6 - 87.1			159 (76.4)	70.6 - 82.3			133 (76.9)	70.5 - 83.2		
No	442 (75.8)	72.3 - 79.3			424 (73.0)	69.4 - 76.6			449 (73.4)	69.9 - 76.9		

[a]Total numbers and percentages for willingness to share include participants answering, “definitely yes” and “probably yes.”

Key Factors Associated With Willingness to Share Internet Use Data

Multivariable logistic regression analysis on willingness to share internet use data for early detection of cancer indicated that willingness to share was associated with perceived benefits and perceived concerns of data sharing. Higher endorsement of benefits of sharing data was associated with a greater willingness to share (OR 8.588, 95% CI 5.687 - 12.970). A higher endorsement of concerns was associated with a lower willingness to share (OR 0.432, 95% CI 0.290 - 0.643). Ethnicity was significantly associated with willingness to share, with individuals of Asian or Asian British ethnicity demonstrating a significantly lower willingness to share than White participants (OR 0.234, 95% CI 0.076 - 0.723). No other demographic factors, including age, gender, employment, income, marital status, education level, personal or family history of cancer, or level of deprivation, were associated with a willingness to share in the multivariable regression analysis, although some (education, employment, ethnicity and income) had indicated associations in univariate analyses (see Multimedia Appendix 1). Factors related to digital literacy (knowledge about internet use data, the number of online services used, or self-reported familiarity with internet use data) were not significantly related to willingness to share internet use data for cancer research in the multivariable regression analysis.

Willingness to share internet use data for the early detection of heart disease and depression followed similar patterns, with perceived benefits being significantly associated with a higher willingness to share internet use data (heart disease: OR 5.692, 95% CI: 4.050 - 7.998; depression: OR 8.850, 95% CI 5.998 - 13.059) and perceived concerns being significantly associated with a lower willingness to share internet use data (heart disease: OR 0.424, 95% CI 0.292 - 0.617; depression: OR 0.343, 95% CI 0.228 - 0.516). For research on detection of heart disease, self-reported familiarity with internet use data was associated with a lower willingness to share internet use data (OR 0.740, 95% CI 0.561 - 0.976). For research on the early detection of depression, older participants were less willing to share their data (OR 0.975, 95% CI 0.951 - 0.999), and men were more likely to be willing to share than women (OR 2.615, 95% CI 1.511 - 4.526).

All results are displayed in Table 4.

Sample sizes for the logistic regression analyses were lower than intended due to missing sociodemographic data, especially for levels of deprivation. We therefore conducted sensitivity analyses by repeating the logistic regression analyses without levels of deprivation. Results were similar to the initial logistic regression analysis (see Multimedia Appendix 1), except for depression, where age was no longer associated with willingness to share their data (OR 0.981, 95% CI 0.959 - 1.003) and those with a higher knowledge score about internet use data were less willing to share internet use data (OR 0.842, 95% CI 0.723 - 0.981). For early detection of cancer, Asian and Asian

XSL•FO
RenderX

British were not significantly less likely to share their internet use data (OR 0.351, 95% CI 0.122 - 1.011).

XSL•FO
RenderX

Table . Logistic regression on willingness to share internet use data for research on the early detection of cancer from a cross-sectional survey in 2025 with UK participants recruited through an online survey panel.

Variable	Willingness to share internet use data for research on early detection of…											
	Cancer (n=670)[a]				Heart disease (n=646)[b]				Depression (n=673)[c]			
	Willing, n (%)	OR[d] (95% CI)	SE	*P* value	Willing, n (%)	OR (95% CI)	SE	*P* value	Willing, n (%)	OR (95% CI)	SE	*P* value
Description of internet use												
Written description	269 (80.5)	1	—[e]	—	246 (78.8)	1	—	—	244 (76.0)	1	—	—
Pictorial example and written description	257 (76.5)	0.617 (0.357-1.066)	0.279	.08	243 (72.8)	0.735 (0.450-1.198)	0.250	.22	263 (74.7)	1.200 (0.723-1.989)	0.258	.48
Gender												
Women	240 (80.5)	1	—	—	228 (74.8)	1	—	—	224 (70.2)	1	—	—
Men	286 (76.9)	0.980 (0.560-1.716)	0.286	.94	261 (76.5)	1.017 (0.607-1.704)	0.263	.95	283 (79.9)	2.615 (1.511-4.526)	0.280	<.001
Education												
Degree level or above	296 (80.9)	1	—	—	263 (79.7)	1	—	—	294 (80.1)	1	—	—
Basic education	98 (72.1)	0.590 (0.300-1.159)	0.344	.13	101 (71.6)	0.688 (0.343-1.380)	0.36	.29	90 (67.7)	0.614 (0.291-1.296)	0.381	.20
Advanced education	132 (78.6)	0.754 (0.390-1.455)	0.336	.40	125 (71.4)	0.627 (0.347-1.130)	0.301	.12	123 (71.1)	0.689 (0.362-1.313)	0.329	.26
Employment												
Employed	324 (83.7)	1	—	—	297 (76.3)	1	—	—	329 (78.5)	1	—	—
Unemployed	69 (67.6)	0.730 (0.353-1.509)	0.371	.40	60 (74.1)	1.172 (0.518-2.656)	0.417	.70	66 (75.0)	1.303 (0.575-2.950)	0.417	.53
Retired	133 (73.5)	1.194 (0.521-2.732)	0.422	.68	132 (75.0)	1.251 (0.583-2.687)	0.390	.57	112 (67.5)	1.770 (0.826-3.792)	0.389	.14
Income (£)												
<35,000	251 (74.5)	1	—	—	254 (73.6)	1	—	—	239 (72.2)	1	—	—
>35,000	275 (82.6)	1.327 (0.703-2.503)	0.324	.38	235 (78.1)	0.826 (0.459-1.487)	0.300	.52	268 (78.4)	0.775 (0.406-1.479)	0.330	.44
Marital status												

XSL•FO
RenderX

Variable	Willingness to share internet use data for research on early detection of…											
	Cancer (n=670)[a]				Heart disease (n=646)[b]				Depression (n=673)[c]			
	Willing, n (%)	OR[d] (95% CI)	SE	*P* value	Willing, n (%)	OR (95% CI)	SE	*P* value	Willing, n (%)	OR (95% CI)	SE	*P* value
Married or in legal partnership	310 (80.9)	1	—	—	299 (79.7)	1	—	—	314 (79.1)	1	—	—
Single	131 (74.4)	1.250 (0.625-2.499)	0.353	.53	123 (70.7)	0.669 (0.364-1.229)	0.310	.20	128 (69.6)	1.020 (0.537-1.936)	0.327	.95
Divorced/widowed/separated	85 (76.6)	1.057 (0.511-2.186)	0.371	.88	67 (69.1)	0.754 (0.362-1.571)	0.374	.45	65 (70.7)	1.294 (0.575-2.915)	0.414	.53
Ethnicity												
White	445 (78.1)	1	—	—	416 (75.0)	1	—	—	408 (73.8)	1	—	—
Asian or Asian British	21 (65.6)	0.234 (0.076-0.723)	0.576	.01	22 (81.5)	0.843 (0.231-3.084)	0.661	.80	32 (80.0)	1.277 (0.394-4.133)	0.599	.68
Black or Black British	48 (92.3)	0.806 (0.189-3.431)	0.739	.77	43 (82.7)	0.478 (0.177-1.291)	0.507	.15	58 (89.2)	0.682 (0.225-2.065)	0.565	.50
Mixed or multiple ethnic groups	12 (75.0)	0.251 (0.062-1.019)	0.714	.053	8 (66.7)	0.474 (0.098-2.283)	0.802	.35	9 (60.0)	0.273 (0.063-1.177)	0.745	.08
Level of deprivation												
IMD[f] Q1 (most deprived)	107 (80.5)	1	—	—	107 (78.7)	1	—	—	120 (78.4)	1	—	—
IMD Q2	130 (77.4)	1.664 (0.766-3.613)	0.396	.20	113 (75.8)	1.116 (0.541-2.302)	0.370	.77	114 (79.7)	2.043 (0.910-4.583)	0.412	.08
IMD Q3	93 (76.9)	0.965 (0.417-2.234)	0.428	.93	99 (72.3)	1.153 (0.531-2.503)	0.395	.72	93 (73.8)	1.450 (0.638-3.295)	0.419	.38
IMD Q4	94 (77.0)	1.162 (0.495-2.729)	0.436	.73	75 (75.0)	1.868 (0.809-4.315)	0.427	.14	83 (69.7)	1.896 (0.811-4.432)	0.433	.14
IMD Q5 (least deprived)	102 (81.0)	1.943 (0.793-4.763)	0.457	.15	95 (76.6)	1.155 (0.512-2.603)	0.415	.73	97 (73.5)	1.142 (0.496-2.628)	0.425	.76
Personal history of condition of interest												

XSL•FO
RenderX

Variable	Willingness to share internet use data for research on early detection of…											
	Cancer (n=670)[a]				Heart disease (n=646)[b]				Depression (n=673)[c]			
	Willing, n (%)	OR[d] (95% CI)	SE	*P* value	Willing, n (%)	OR (95% CI)	SE	*P* value	Willing, n (%)	OR (95% CI)	SE	*P* value
No	422 (77.3)	1	—	—	395 (75.1)	1	—	—	349 (73.0)	1	—	—
Yes	104 (83.9)	1.526 (0.720-3.236)	0.383	.27	94 (78.3)	0.892 (0.444-1.794)	0.357	.75	158 (81.0)	1.345 (0.720-2.513)	0.319	.35
Family history of condition of interest												
No	370 (76.8)	1	—	—	352 (73.9)	1	—	—	390 (74.4)	1	—	—
Yes	156 (83.0)	1.189 (0.651-2.171)	0.307	.57	137 (80.6)	1.037 (0.564-1.907)	0.311	.91	117 (78.5)	1.458 (0.758-2.804)	0.334	.26
Age	—	0.983 (0.958-1.009)	0.013	.21	—	0.984 (0.962-1.008)	0.012	.18	—	0.975 (0.951-0.999)	0.012	.04
Perceived benefits about sharing internet use data	—	8.588 (5.687-12.970)	0.210	<.001	—	5.692 (4.050-7.998)	0.174	<.001	—	8.850 (5.998-13.059)	0.198	<.001
Perceived concerns about sharing internet use data	—	0.432 (0.290-0.643)	0.203	<.001	—	0.424 (0.292-0.617)	0.191	<.001	—	0.343 (0.228-0.516)	0.208	<.001
Knowledge about internet use data (score)	—	1.061 (0.906-1.242)	0.080	.46	—	1.011 (0.866-1.180)	0.079	.891	—	0.858 (0.727-1.012)	0.084	.07
Number of online services used	—	0.923 (0.826-1.031)	0.057	.16	—	1.040 (0.942-1.148)	0.051	.440	—	1.033 (0.929-1.149)	0.054	.55

XSL•FO
RenderX

Variable	Willingness to share internet use data for research on early detection of…											
	Cancer (n=670)[a]				Heart disease (n=646)[b]				Depression (n=673)[c]			
	Willing, n (%)	OR[d] (95% CI)	SE	*P* value	Willing, n (%)	OR (95% CI)	SE	*P* value	Willing, n (%)	OR (95% CI)	SE	*P* value
Familiarity with internet use data	—	1.108 (0.814-1.509)	0.158	.52	—	0.740 (0.561-0.976)	0.141	.033	—	0.854 (0.625-1.167)	0.160	.32

[a]Nagelkerke R=0.573.

[b]Nagelkerke R=0.516

[c]Nagelkerke R=0.588.

[d]OR: odds ratio.

[e]Not applicable.

[f]IMD: index of multiple deprivation.

Statements on Willingness to Share and Concerns

The open-ended question on what would make them more willing to share their internet use data was answered by 2307 (2307/2390, 96.5%) participants. Percentages of statements relative to numbers of those definitely willing, probably willing, probably unwilling, and definitely unwilling to share are presented in Figure 1. The main suggestions given by those currently not willing to share were strong assurance of data security (66/596, 11.1%) and incentives (37/596, 6.2%); however, the majority of these participants stated “nothing” would make them more willing (definitely not: 189/236, 80.1%; probably not: 196/360, 54.4%). Those who were “probably” or “definitely” willing to share their internet use data offered a wider range of suggestions, including a sense of contribution to research and society (“helping others”: 480/1794, 26.8%); data security assurances (211/1794, 11.8%); clarification of research purposes (157/1794, 8.6%); incentives (110/1794, 6.1%); strict governance, personal control, and building trust (106/1794, 5.9%); and personal relevance (147/1794, 8.2%).

XSL•FO
RenderX

Figure 1. Suggestions to increase willingness to share internet use data from a cross-sectional survey in 2025 with UK participants recruited through an online survey panel. Statements could be coded as multiple categories. Participants saying nothing would make them more willing are not included in this figure.

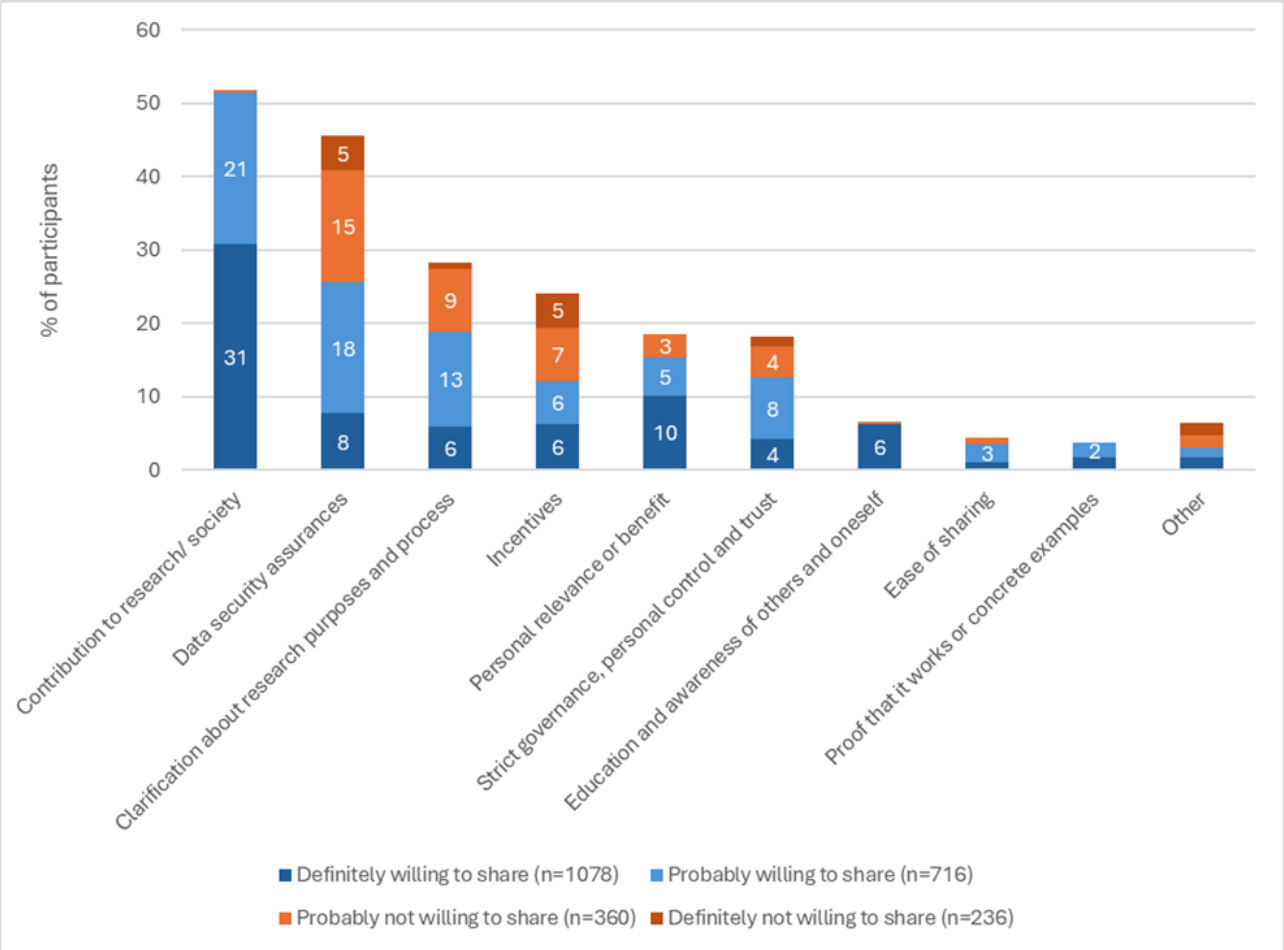

The question about concerns surrounding sharing internet use data was answered by 2318 (2318/2390, 97%) participants (Figure 2). Concerns of participants who were “probably not” or “definitely not” willing to share their internet use data mostly focused on data security and the invasion of privacy (160/596, 26.8%) and concerns about data misuse, including sharing with private (marketing) companies (39/596, 6.5%). Participants also questioned the relevance of their internet use data for early detection of disease (76/596, 12.8%). Participants who were “probably” or “definitely” willing to share their internet use data noted similar concerns, although many indicated no concerns (probably willing: 368/716, 51.4%; definitely willing: 794/1078, 73.7%). Examples are provided in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 2. Concerns about sharing internet use data from a cross-sectional survey in 2025 with UK participants recruited through an online survey panel. Statements could be coded as multiple categories. Participants saying they had no concerns are not included in this figure.

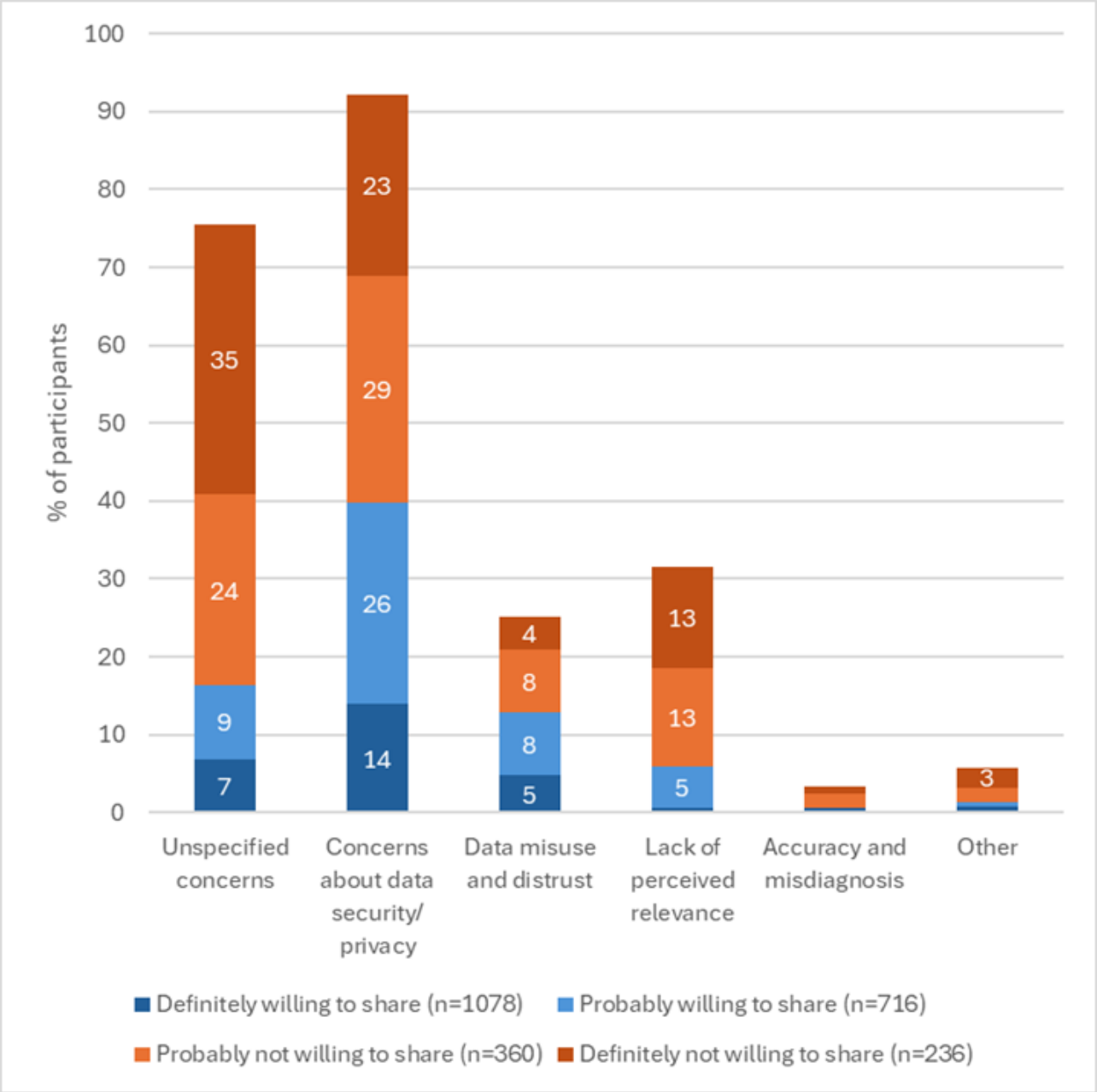

Discussion

Principal Findings

This study examined willingness to share internet use data for research into early detection of cancer, heart disease, and depression. Participants showed a similarly high willingness to share internet use data for all conditions. Willingness differed by data type, being lowest for online purchases and location history and highest for health app data. Viewing a screenshot of browsing history did not significantly affect willingness for any condition.

Perceived benefits and concerns were key factors associated with willingness across all conditions. Answering open questions about willingness to share, participants reported that contributing to research, personal relevance, data security, incentives, clear research purposes, concrete examples, strong governance, and trust in researchers would improve their willingness to share. In contrast, they reported that privacy concerns, fear of misuse, and low relevance could reduce willingness to share. Those who were more familiar with internet use data were less likely to share data for heart disease detection, but this was not associated with willingness to share data for research on cancer or depression. There were few sociodemographic factors associated with willingness to share, but some did exist: Asian and Asian British participants were less willing to share their internet use data for cancer detection (not replicated in the sensitivity analyses), while younger (not replicated in the sensitivity analyses) and male participants were more willing to share data for depression detection.

XSL•FO
RenderX

Comparison With Prior Work

We found a surprisingly high willingness to share internet use data for health research compared with recent studies. For example, Grande et al [20] classified approximately 13% of their participants as "agreeable to sharing," whereas about 45% of our participants indicated they were "definitely" willing to share. Similarly, Hirst et al [19] found that 61% of participants were willing to share their data with research organizations, and only 43% were willing to share with private organizations, compared with 75% in this study. Although we did not differentiate the kind of organization data would be shared with, we consistently found that a higher proportion of participants were willing to share across different types of data than in these studies. Our proportions are based on users of the respective types of data. However, willingness to share was similar when participants were asked more generally about "internet use data."

As we used a similar recruitment strategy as previous studies [19], it seems that the *purpose* of the data sharing described in our study may be responsible for the higher willingness to share internet use data. In our survey, we asked participants specifically about sharing their internet use data for the early detection of health conditions. Previous research has found public support for screening and early detection [35,36], especially if personally relevant [37]. This is line with open-ended comments regarding increased willingness to share if there were personal benefits (eg, receiving a quicker diagnosis). Beyond personal benefits, participants in our study who were willing to share their internet use data described a general sense of contribution to research and society, which our PPI group summarized as an act of "charity" in terms of "donating data." The Smart Data Research UK report [12] reports how the public grew more enthusiastic about using smart data in research for the "public good" when learning and discussing potential applications. However, information about the value of data sharing, as well as ensuring personal control over sharing potentially sensitive information is crucial to ensure acceptability of sharing data [38], which is reflected in our findings.

Another suggestion for improving willingness to share internet use data in our study was monetary incentives, which was not captured in our quantitative scales but reported in the free text. Even participants who were currently unwilling to share mentioned incentives as potentially increasing their willingness to share. This indicates a "trade-off" between privacy and payment, although the exact value of individual privacy is difficult to determine [39].

Our findings regarding higher willingness to share health app data than to share other types of internet use data are in line with previous research [20]. For example, Ackermann et al [40] found that a better "intuitive match" between the type of data requested and what the data are used for increases willingness to share internet use data. In this survey study, this would be the case for health app data but not necessarily for online purchases or location data, which participants were less willing to share. However, researchers developing risk prediction models based on or including internet use data might rely on changes in online behavior beyond concrete health app data (eg, if an individual repeatedly searches for nonspecific symptoms or home remedies). As individuals are likely to look up symptoms online or change the tone and semantics of their searches long before visiting their general practitioner, capturing these early changes might benefit early detection [13-16].

Types of data that are not directly related to medical contexts might also be seen as more personal and more likely to be misused, such as by marketing companies [19]. Concerns about privacy and misuse of data were associated with a lower willingness to share across all conditions in this study. They were commonly endorsed and reported in both the quantitative and qualitative part of the survey, even by participants who were "definitely" willing to share. The Smart Data Research UK report [12] also shows that the public is skeptical about private sector motivation and concerned about potential inequalities and data security.

Implications for Future Research and Practice

There is a clear need for transparent and open communication around the purpose of using internet data for early detection research, as well as data processing after it is shared, including which data, how, and why they are collected and stored [12,38]. Researchers and data scientists need to work toward clarifying which data are needed for what purpose, how to limit data sharing to what is necessary and filter out any other information, and how to communicate purpose and process of data sharing, especially concerning minority groups [41,42]. We found no effect of showing participants a pictorial example of internet use data on willingness to share. Our PPI group noted that the example only included a generic internet history without particularly sensitive data (eg, regarding financial details or a child's local school). Future research could examine the effect of different types of examples, preferably using participants' own internet use data.

Only 10% of participants in our sample noted that they knew a lot about internet use data, and 34% to 44% mistakenly believed that internet searches in "incognito" mode, online banking details, and content of emails are part of internet use data. This indicates that internet users might not be well-informed about the personal data they might be sharing online [43,44]. There is a need for more public information and education around what internet use data entail and what would need to be accessed for medical research to facilitate informed consent [45]. Our findings suggest that higher familiarity and, potentially, knowledge might negatively impact willingness to share internet use data for medical research. However, this should be interpreted with caution, as univariate analyses indicated the opposite trend. Actual familiarity and knowledge about internet use data and the consequences for willingness to share data should be investigated in more detail. Clear communication including which data are shared with whom and ensuring personal control are central to enable informed participation in internet research [45].

Our findings show how concerns around sharing internet use data might be more pronounced in some ethnic minority groups and corroborate research showing that cancer is associated with fear and stigma in minority communities [46]. Future research should focus on mitigation strategies to avoid inequalities in

XSL•FO
RenderX

this new stream of internet-based medical research, especially for vulnerable populations. Stigma around mental health conditions might drive the findings that younger age and male gender were associated with a higher willingness to share internet use data for the detection of depression [47]. As traditional norms of masculinity can reduce symptom expression and inhibit help-seeking for mental health conditions [48,49], sharing internet use data might be a more anonymous way to disclose potential symptoms.

Although many participants reported having concerns about sharing internet use data, a high proportion did not specify their concerns. This might have been enforced by the wording of the open-ended question around concerns but could also reflect a general unease with data sharing due to a lack of knowledge and transparency [50]. A qualitative study identified similar concerns around data security and misuse [51] as well as actions taken by individuals to protect their privacy, such as use of incognito mode for sensitive searches, that could impact future research. Future research, especially qualitative research, should examine how these concerns can be mitigated in more detail. Additionally, those who reported being "definitely not" willing to share data often said "nothing" would make them more willing. Future research should therefore investigate why some people would not be willing to share internet use data under any circumstances.

Strengths and Limitations

This study is a large survey with population-based quotas that builds on a previous qualitative study [51], previously used measures [18,19,24], and extensive patient and public input. However, the survey assessed self-reported willingness to share rather than the actual behavior, as well as self-reported diagnosis, which might have led to an overestimation of those with a confirmed clinical diagnosis, especially in those reporting to have depression [52]. It remains to be seen whether these results will be confirmed in future research and practice when shared internet use data are required.

We recruited through an online panel and therefore included participants who are comfortable answering online research surveys for incentives and attracted those with higher education (approximately 52% with at least a degree level of education, compared with the UK average of 33% - 34%, [53]). This might have contributed to the high willingness to share internet use data and frequent mentioning of incentives to increase willingness in the open-ended questions. Nevertheless, internet use data can only be shared by people who are using the internet in the first place, and internet use and knowledge about internet use data varied in the sample, suggesting the sample represents a range of internet users. The power for some of the logistic regression analyses might have been too low to detect differences. Although we ensured a representative percentage of the sample (approximately 17%) was from ethnic minority groups, the numbers within minority ethnic groups limit accurate comparisons.

Other potential methodological limitations due to the short survey design include the use of area-based index of multiple deprivation quintiles for individual-level data, potentially leading to ecological fallacy as individuals living in affluent areas may still experience deprivation and vice versa [54]. Nevertheless, it is still the most comprehensible method that could be applied in a short questionnaire, and sensitivity analyses were used to replicate logistic regression analyses without this deprivation variable [55]. We combined those "definitely" and "probably" (not) willing to share for analyses, although the groups might differ. Dichotomizing questions about willingness to share internet use data might have reduced the nuance of the data.

Finally, we had to exclude a substantial proportion (16.6%) of responses due to low data quality, including those flagged as bots and fraudulent responses and those failing attention checks. This does not negatively impact the quality of our final sample, as we were able to exclude and replace these participants during recruitment. However, this shows that researchers will need to account for "impostor" responses when planning for sample sizes for online surveys, implement strategies to detect low quality data (eg, attention checks or bot detection), and carefully check responses, leading to additional workload. If low-quality responses are not addressed, we may not be able to trust similar research results [56].

Conclusion

This study is innovative as the first large-scale survey, to the best of our knowledge, to examine public willingness to share internet use data for early detection research across multiple common health conditions while assessing perceived benefits, risks, and sociodemographic differences. This survey indicated willingness to share internet use data is high across research on early detection of different common health conditions. Willingness to share was associated with perceived benefits for oneself and society in general and limited by concerns around data privacy and potential for misuse. Participants were more willing to share data that were plausibly related to the purpose of early detection of health conditions (eg, health app data).

The survey results advance the understanding of participation in research using internet use data by identifying both barriers and enablers of data sharing. Although sociodemographic differences were limited, variations by age, gender, and ethnicity suggest potential inequities. The findings inform consent design, communication strategies, and data governance, highlighting the need for clear purpose limitation and robust data protection to foster trust and reduce inequalities. They emphasize the need to clarify which and how data are used for research into early detection and investigate how communication about internet use data could improve willingness to share while avoiding perpetuating inequalities. Data sharing needs to be limited to specific research purposes, and resolute data protection strategies need to be put in place.

XSL•FO
RenderX

Acknowledgments

We would like to thank our patient and public involvement (PPI) group for their valuable inputs during this project and for reviewing and testing the survey. We would also like to thank our participants for their time and interest in this study.

The Data Liberation Front (based at Google) was a collaborator on the study and advised on accurate description about internet use data. The research team members (CD, MR, SES) were the sole decision-makers on research design and outputs.

Funding

Google provided unrestricted funding for this research through a Privacy Research Faculty Award to SES. The funder had no role in the study design; collection, analysis, or interpretation of data; writing of the paper; and/or decision to submit for publication. SES is supported by Barts Charity (G-001520; MRC&U0036).

Data Availability

The full syntax and the final datasets can be found on OSF [56].

Authors' Contributions

SES acquired funding. SES, CD, MR and SA contributed to conceptualization. SES and CD were responsible for project administration. CD undertook the data collection and analysis, with input from SES, MR and WL. CD drafted the original manuscript. All authors contributed to writing – review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
WISER survey, additional regression analyses, and definitions of categories for suggestions to increase willingness to share and for concerns.
[DOCX File, 139 KB - jmir_v28i1e85637_app1.docx]

References

1. Oude Wolcherink MJ, Behr CM, Pouwels X, Doggen CJM, Koffijberg H. Health economic research assessing the value of early eetection of cardiovascular disease: a systematic review. Pharmacoeconomics 2023 Oct;41(10):1183-1203. [doi: 10.1007/s40273-023-01287-2] [Medline: 37328633]
2. Vogel B, Acevedo M, Appelman Y, et al. The Lancet Women and Cardiovascular Disease Commission: reducing the global burden by 2030. Lancet 2021 Jun 19;397(10292):2385-2438. [doi: 10.1016/S0140-6736(21)00684-X] [Medline: 34010613]
3. Crosby D, Bhatia S, Brindle KM, et al. Early detection of cancer. Science 2022 Mar 18;375(6586):eaay9040. [doi: 10.1126/science.aay9040] [Medline: 35298272]
4. Whitaker KL, Scott SE, Wardle J. Applying symptom appraisal models to understand sociodemographic differences in responses to possible cancer symptoms: a research agenda. Br J Cancer 2015 Mar 31;112 Suppl 1(Suppl 1):S27-S34. [doi: 10.1038/bjc.2015.39] [Medline: 25734385]
5. Pedersen SD, Nielsen BD, Assmann ML, Hauge EM, de Thurah A. Early diagnosis of rheumatoid arthritis: associations between patients' perceptions of initial symptoms and the timing of seeking help from the general practitioner. Scand J Rheumatol 2025 Jul;54(4):242-251. [doi: 10.1080/03009742.2025.2464457] [Medline: 40013732]
6. Asch JM, Asch DA, Klinger EV, Marks J, Sadek N, Merchant RM. Google search histories of patients presenting to an emergency department: an observational study. BMJ Open 2019 Feb 20;9(2):e024791. [doi: 10.1136/bmjopen-2018-024791] [Medline: 30787088]
7. Padrez KA, Ungar L, Schwartz HA, et al. Linking social media and medical record data: a study of adults presenting to an academic, urban emergency department. BMJ Qual Saf 2016 Jun;25(6):414-423. [doi: 10.1136/bmjqs-2015-004489] [Medline: 26464519]
8. D'Alfonso S, Coghlan S, Schmidt S, Mangelsdorf S. Ethical dimensions of digital phenotyping within the context of mental healthcare. J technol behav sci 2024 Jun 25;10(1):132-147. [doi: 10.1007/s41347-024-00423-9]
9. Koo MM, Hamilton W, Walter FM, Rubin GP, Lyratzopoulos G. Symptom signatures and diagnostic timeliness in cancer patients: a review of current evidence. Neoplasia 2018 Feb;20(2):165-174. [doi: 10.1016/j.neo.2017.11.005] [Medline: 29253839]
10. Ben-Zeev D, Young MA. Accuracy of hospitalized depressed patients' and healthy controls' retrospective symptom reports: an experience sampling study. J Nerv Ment Dis 2010 Apr;198(4):280-285. [doi: 10.1097/NMD.0b013e3181d6141f] [Medline: 20386257]

XSL•FO
RenderX

11. Lackner JM, Jaccard J, Keefer L, et al. The accuracy of patient-reported measures for GI symptoms: a comparison of real time and retrospective reports. Neurogastroenterol Motil 2014 Dec;26(12):1802-1811. [doi: 10.1111/nmo.12466] [Medline: 25424582]
12. Smart Data Research UK. Smart Data Research UK public dialogue report 2025. 2025 URL: https://www.sdruk.ukri.org/our-work/publications/public-dialogue-report/sdr-uk-public-dialogue-report-2025/ [accessed 2026-03-21]
13. Areán PA, Pratap A, Hsin H, et al. Perceived utility and characterization of personal Google search histories to detect data patterns proximal to a suicide attempt in individuals who previously attempted suicide: pilot cohort study. J Med Internet Res 2021 May 6;23(5):e27918. [doi: 10.2196/27918] [Medline: 33955838]
14. Barcroft JF, Yom-Tov E, Lampos V, et al. Using online search activity for earlier detection of gynaecological malignancy. BMC Public Health 2024 Mar 11;24(1):608. [doi: 10.1186/s12889-024-17673-0] [Medline: 38462622]
15. Paparrizos J, White RW, Horvitz E. Screening for pancreatic adenocarcinoma using signals from web search logs: feasibility study and results. J Oncol Pract 2016 Aug;12(8):737-744. [doi: 10.1200/JOP.2015.010504] [Medline: 27271506]
16. Birnbaum ML, Wen H, Van Meter A, et al. Identifying emerging mental illness utilizing search engine activity: a feasibility study. PLOS ONE 2020;15(10):e0240820. [doi: 10.1371/journal.pone.0240820] [Medline: 33064759]
17. Brall C, Berlin C, Zwahlen M, Ormond KE, Egger M, Vayena E. Public willingness to participate in personalized health research and biobanking: a large-scale Swiss survey. PLoS ONE 2021;16(4):e0249141. [doi: 10.1371/journal.pone.0249141] [Medline: 33793624]
18. Seltzer E, Goldshear J, Guntuku SC, et al. Patients' willingness to share digital health and non-health data for research: a cross-sectional study. BMC Med Inform Decis Mak 2019 Aug 8;19(1):157. [doi: 10.1186/s12911-019-0886-9] [Medline: 31395102]
19. Hirst Y, Stoffel ST, Brewer HR, Timotijevic L, Raats MM, Flanagan JM. Understanding public attitudes and willingness to share commercial data for health research: survey study in the United Kingdom. JMIR Public Health Surveill 2023 Mar 23;9:e40814. [doi: 10.2196/40814] [Medline: 36951929]
20. Grande D, Mitra N, Iyengar R, et al. Consumer willingness to share personal digital information for health-related uses. JAMA Netw Open 2022 Jan 4;5(1):e2144787. [doi: 10.1001/jamanetworkopen.2021.44787] [Medline: 35072717]
21. Shandhi MMH, Singh K, Janson N, et al. Assessment of ownership of smart devices and the acceptability of digital health data sharing. NPJ Digit Med 2024 Feb 22;7(1):44. [doi: 10.1038/s41746-024-01030-x] [Medline: 38388660]
22. Trauth JM, Musa D, Siminoff L, Jewell IK, Ricci E. Public attitudes regarding willingness to participate in medical research studies. J Health Soc Policy 2000;12(2):23-43. [doi: 10.1300/J045v12n02_02] [Medline: 11184441]
23. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
24. Sanderson SC, Brothers KB, Mercaldo ND, et al. Public attitudes toward consent and data sharing in biobank research: a large multi-site experimental survey in the US. Am J Hum Genet 2017 Mar 2;100(3):414-427. [doi: 10.1016/j.ajhg.2017.01.021] [Medline: 28190457]
25. Lidington E, Stiles M, Maudsley J, et al. Establishing a set of acceptable demographic questions for use in health research through public consultation. Res Involv Engagem 2026 Jan 28;12(1):24. [doi: 10.1186/s40900-026-00836-1] [Medline: 41606646]
26. English indices of deprivation 2019. Ministry of Housing, Communities & Local Government. 2019. URL: https://imd-by-postcode.opendatacommunities.org/imd/2019 [accessed 2026-03-21]
27. Arias J. Postcode Deprivation Finder. URL: https://alasdairrae.github.io/postcodez/ [accessed 2026-03-21]
28. NI multiple deprivation measures. Northern Ireland Statistics and Research Agency. URL: https://deprivation.nisra.gov.uk/ [accessed 2026-03-21]
29. Health Survey for England, 2021 part 2. NHS England. 2023. URL: https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england/2021-part-2/adult-health-general-health [accessed 2026-03-21]
30. GOV UK. Research and analysis: Digital Landscape Research. 2012. URL: https://www.gov.uk/government/publications/digital-landscape-research/digital-landscape-research [accessed 2026-03-21]
31. Johnson MS, Adams VM, Byrne J. Addressing fraudulent responses in online surveys: insights from a web‐based participatory mapping study. People and Nature 2024 Feb;6(1):147-164 [FREE Full text] [doi: 10.1002/pan3.10557]
32. Lawlor J, Thomas C, Guhin AT, et al. Suspicious and fraudulent online survey participation: introducing the REAL framework. Methodological Innovations 2021 Sep;14(3):20597991211050468. [doi: 10.1177/20597991211050467]
33. Derksen C, Ramasawmy M, Scott S. WISER survey. Open Science Framework (OSF). URL: https://osf.io/bzn4q/overview [accessed 2026-03-21]
34. Willis GB, Artino AR. What do our respondents think we're asking? Using cognitive interviewing to improve medical education surveys. J Grad Med Educ 2013 Sep;5(3):353-356. [doi: 10.4300/JGME-D-13-00154.1] [Medline: 24404294]
35. Schmeising-Barnes N, Waller J, Marlow LAV. Attitudes to multi-cancer early detection (MCED) blood tests for population-based screening: a qualitative study in Great Britain. Soc Sci Med 2024 Apr;347:116762. [doi: 10.1016/j.socscimed.2024.116762] [Medline: 38520825]

XSL•FO
RenderX

36. Waller J, Osborne K, Wardle J. Enthusiasm for cancer screening in Great Britain: a general population survey. Br J Cancer 2015 Feb 3;112(3):562-566. [doi: 10.1038/bjc.2014.643] [Medline: 25535731]
37. Cheong AT, Khoo EM, Tong SF, Liew SM. To check or not to check? A qualitative study on how the public decides on health checks for cardiovascular disease prevention. PLoS ONE 2016;11(7):e0159438. [doi: 10.1371/journal.pone.0159438] [Medline: 27415432]
38. Shiells K, Di Cara N, Skatova A, et al. Participant acceptability of digital footprint data collection strategies: an exemplar approach to participant engagement and involvement in the ALSPAC birth cohort study. Int J Popul Data Sci 2020;5(3):1728. [doi: 10.23889/ijpds.v7i1.1728] [Medline: 35519823]
39. Skatova A, McDonald R, Ma S, Maple C. Unpacking privacy: valuation of personal data protection. PLoS ONE 2023;18(5):e0284581. [doi: 10.1371/journal.pone.0284581] [Medline: 37134067]
40. Ackermann KA, Burkhalter L, Mildenberger T, Frey M, Bearth A. Willingness to share data: contextual determinants of consumers' decisions to share private data with companies. J of Consumer Behaviour 2022 Mar;21(2):375-386 [FREE Full text] [doi: 10.1002/cb.2012]
41. Baines R, Stevens S, Austin D, et al. Patient and public willingness to share personal health data for third-party or secondary uses: systematic review. J Med Internet Res 2024 Mar 5;26:e50421. [doi: 10.2196/50421] [Medline: 38441944]
42. Yusuf ZK, Dixon WG, Sharp C, Cook L, Holm S, Sanders C. Building and sustaining public trust in health data sharing for musculoskeletal research: semistructured interview and focus group study. J Med Internet Res 2024 Oct 15;26:e53024. [doi: 10.2196/53024] [Medline: 39405526]
43. Bartol J, Prevodnik K, Vehovar V, Petrovčič A. The roles of perceived privacy control, Internet privacy concerns and Internet skills in the direct and indirect Internet uses of older adults: conceptual integration and empirical testing of a theoretical model. New Media & Society 2024 Aug;26(8):4490-4510. [doi: 10.1177/14614448221122734]
44. Raibulet C, Wang K. Awareness of privacy and data collection: exploring privacy policy effectiveness in Google Maps. Front Comput Sci 2025 Sep 24;7:1568179. [doi: 10.3389/fcomp.2025.1568179]
45. Bietz MJ, Bloss CS, Calvert S, et al. Opportunities and challenges in the use of personal health data for health research. J Am Med Inform Assoc 2016 Apr;23(e1):e42-e48. [doi: 10.1093/jamia/ocv118] [Medline: 26335984]
46. Licqurish S, Phillipson L, Chiang P, Walker J, Walter F, Emery J. Cancer beliefs in ethnic minority populations: a review and meta-synthesis of qualitative studies. Eur J Cancer Care 2017 Jan;26(1):e12556 [FREE Full text] [doi: 10.1111/ecc.12556]
47. Bradbury A. Mental health stigma: the impact of age and gender on attitudes. Community Ment Health J 2020 Jul;56(5):933-938. [doi: 10.1007/s10597-020-00559-x] [Medline: 32002711]
48. Seidler ZE, Dawes AJ, Rice SM, Oliffe JL, Dhillon HM. The role of masculinity in men's help-seeking for depression: a systematic review. Clin Psychol Rev 2016 Nov;49:106-118. [doi: 10.1016/j.cpr.2016.09.002] [Medline: 27664823]
49. Wirback T, Forsell Y, Larsson JO, Engström K, Edhborg M. Experiences of depression and help-seeking described by young Swedish men. Psychol Men Masc 2018 Jul;19(3):407-417. [doi: 10.1037/men0000110]
50. Baruh L, Secinti E, Cemalcilar Z. Online privacy concerns and privacy management: a meta-analytical review. J Commun 2017 Feb;67(1):26-53 [FREE Full text] [doi: 10.1111/jcom.12276]
51. Gradwell NC, Ramasawmy M, Arora S, et al. Acceptability of sharing internet browsing history for cancer research: think-aloud and interview study. JMIR Cancer 2026 Feb 2;12:e82009. [doi: 10.2196/82009] [Medline: 41628357]
52. Thombs BD, Kwakkenbos L, Levis AW, Benedetti A. Addressing overestimation of the prevalence of depression based on self-report screening questionnaires. CMAJ 2018 Jan 15;190(2):E44-E49. [doi: 10.1503/cmaj.170691] [Medline: 29335262]
53. Education, England and Wales: census 2021. Office for National Statistics. 2021. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/educationandchildcare/bulletins/educationenglandandwales/census2021 [accessed 2026-03-21]
54. Bryere J, Pornet C, Copin N, et al. Assessment of the ecological bias of seven aggregate social deprivation indices. BMC Public Health 2017 Jan 17;17(1):86. [doi: 10.1186/s12889-016-4007-8] [Medline: 28095815]
55. Capuano AW, Dawson JD, Ramirez MR, Wilson RS, Barnes LL, Fields RW. Modeling Likert scale outcomes with trend-proportional odds with and without cluster data. Methodology (Gott) 2016 Apr;12(2):33-43. [doi: 10.1027/1614-2241/a000106] [Medline: 37583928]
56. Comachio J, Poulsen A, Bamgboje-Ayodele A, et al. Identifying and counteracting fraudulent responses in online recruitment for health research: a scoping review. BMJ EBM 2025 Jun;30(3):173-182. [doi: 10.1136/bmjebm-2024-113170]

Abbreviations

GDPR: General Data Protection Regulation
JARS-Quant: Journal Article Reporting Standards Quantitative Research Design
OR: odds ratio
OSF: Open Science Framework
PPI: patient and public involvement

XSL•FO
RenderX

Edited by S Brini; submitted 10.Oct.2025; peer-reviewed by N Calanzani, N Howe; revised version received 07.Mar.2026; accepted 07.Mar.2026; published 25.Mar.2026.

Please cite as:
Derksen C, Ramasawmy M, Arora S, Lau W, Scott SE
Willingness to Share Internet Use Data for Research on Early Disease Detection: Cross-Sectional Survey
J Med Internet Res 2026;28:e85637
URL: https://www.jmir.org/2026/1/e85637
doi:10.2196/85637

© Christina Derksen, Mel Ramasawmy, Sanjula Arora, Wynsee Lau, Suzanne Scott. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

From Searching to Coping, How Chinese Patients With Breast Cancer Navigate Web-Based Health Information: Semistructured Interview Study

Jiahui Liu[1], MSc; Xingfeng Li[1], MSc; Xuying Li[2], PhD; Jiejun Chen[1], MSc; Dan Zhang[2], MSc; Cuiling Qiu[2], MSc

[1]Department of Breast Medicine, Hunan Cancer Hospital, 582 Xianjiahu Road, Yuelu District, Changsha, Hunan, China

[2]Department of Nursing, Hunan Cancer Hospital, 582 Xianjiahu Road, Yuelu District, Changsha, Hunan, China

Corresponding Author:
Cuiling Qiu, MSc
Department of Nursing, Hunan Cancer Hospital, 582 Xianjiahu Road, Yuelu District, Changsha, Hunan, China

Abstract

Background: With the development of digital health platforms, patients with breast cancer are increasingly relying on web-based resources to search for disease-related information. Proper usage of web-based health information by patients with breast cancer is crucial for understanding disease information and participating in treatment decisions. However, in the face of the large amount and complexity of information, it is still unclear how patients can make psychological adjustments and behavioral responses. Problems such as variable information quality and conflicting information are also affecting the cognitive and treatment decision-making process of patients with breast cancer.

Objective: This study aims to explore the real experiences of Chinese patients with breast cancer in their search for web-based health information from a phenomenological perspective, providing insights for optimizing future web-based health information support for patients.

Methods: This qualitative study used semistructured, in-depth face-to-face interviews to collect data. Through purposive and convenience sampling, 18 female patients with breast cancer were recruited from a tertiary cancer hospital in China. The data saturation principle was observed to determine the endpoint of data collection. The collected data were analyzed using thematic analysis.

Results: From 18 original interview documents, three themes and 11 subthemes were categorized as follows: (1) driving force of information search (emotion-based information search, problem-solving–oriented information search), (2) cognitive judgments amidst the information fog (interweaving of multichannel information, judgment of information authenticity, information applicability assessment, cognitive confusion in the context of information conflict, and construction of information meaning), and (3) adaptation under the pressure of web-based information (transform information into action, emotional regulatory coping, build a support network, and acceptance and adjustment of expectations).

Conclusions: This study reveals that the experiences of patients with breast cancer within web-based health information environments resemble an information navigation journey. Patients continuously search, evaluate, and adjust within the sea of information to maintain cognitive clarity and emotional equilibrium. The findings offer valuable insights for clinical health care providers, health information platform developers, and policymakers. They can help optimize digital health services and design personalized information support that better meets patients' needs.

(J Med Internet Res 2026;28:e80363) doi:10.2196/80363

KEYWORDS

breast cancer; information seeking; decision-making; coping strategy; qualitative research

Introduction

According to the latest global cancer data in 2020, breast cancer has surpassed lung cancer to become the world's leading cancer, being the main cause of cancer incidence, disability, and mortality among women worldwide, imposing a huge burden on global health services [1]. With the advancement of medical technology and the promotion of early screening, more and more patients with breast cancer can receive early diagnosis and treatment, and their survival prognosis has significantly improved compared with before [2,3]. However, due to the complex pathological characteristics of breast cancer, diverse treatment options, and long-term rehabilitation needs, patients' demand for health information is extremely urgent. A survey study found that among the patients with breast cancer interviewed, as many as 85% expressed a desire to obtain more information about their disease [4]. Sheehy et al [5] conducted a prospective longitudinal investigation and found that the

XSL•FO
RenderX

information needs of patients with breast cancer remained high throughout the entire follow-up period of 1 year, 3 years, and 5 years after diagnosis. The research conducted by Ludwigson et al [6] revealed that patients with breast cancer were dissatisfied with the health information they received regarding disease prognosis, symptom management, and surgical reconstruction. It can be seen that patients with breast cancer have long-term demands for health information, but these demands have not been met.

Therefore, with the rapid development of the internet and the increasing inclusiveness, accessibility, and comprehensibility of digital media, an increasing number of patients are beginning to actively seek health information through web-based channels to assist in their treatment decisions [7]. It has been reported that the internet has become the preferred source of information for patients with cancer, surpassing health care professionals [8]. According to relevant research reports, breast cancer is one of the most frequently searched cancer topics on the internet [9]. Online health information search (OHIS) plays a significant role in the health management of patients with chronic diseases, enabling them to understand their conditions and facilitating their participation in the decision-making process regarding treatment [10,11].

However, the current digital health information environment also presents numerous challenges. Given the large amount of inaccurate web-based information, people can easily become misinformed [12]. A study has shown that misleading web-based health information is associated with poor patient outcomes [13]. Moreover, there is an abundance of web-based health information, but the quality of this information is difficult to guarantee [14]. In the face of the challenges brought by web-based information, patients with breast cancer often need to make psychological trade-offs, forming a complex coping mechanism. The “Stress and Coping Theory” proposed by Lazarus and Folkman [15] provides a theoretical framework for understanding this phenomenon. According to this theory, when individuals encounter potentially threatening events, they undergo primary appraisal and secondary appraisal and, based on the situation, choose appropriate coping strategies [16]. This theory proposes two primary coping strategies: (1) problem-focused coping, aimed at resolving the problem or situation; and (2) emotion-focused/avoidance coping, aimed at regulating stress-related emotions [17]. Problem-focused coping involves acquiring relevant information, while emotion-focused coping entails avoiding thoughts about the threat or reevaluating the threat without altering the stressful reality. These 2 coping strategies may either promote or mutually hinder each other in stressful situations [18]. This theory has been widely applied to analyze the psychological adaptation mechanisms of patients in disease management and health behavior decision-making [19,20].

Although some studies have preliminarily explored the OHIS behaviors of patients with breast cancer, limited research has explored their subjective experiences, cognitive processes, and coping strategies—particularly within the Chinese cultural context [8,21]. Most of the existing literature focuses on the objective characteristics of information behavior. The perspective of phenomenological research emphasizes the genuine understanding of individual experiences [22]. Therefore, this study aims to explore the experiences of patients with breast cancer in web-based health information contexts from a phenomenological perspective. To facilitate a deeper understanding of the psychological mechanisms reflected in participants’ experiences, the Stress and Coping Theory was drawn upon as an interpretive reference. By filling this underexplored research gap, this study has deepened our understanding of how breast cancer patients search for, evaluate, and apply health information in a complex information environment. The research results provide practical guidance for designing personalized and operational web-based health interventions and information support strategies in clinical practice and may also offer insights for the development of digital health services for other chronic diseases or vulnerable groups.

Methods

Study Design

Guided by a phenomenological perspective and using qualitative research methods, this study aims to explore in depth the experiences and feelings of Chinese patients with breast cancer as they search for web-based health information during their disease journey [22]. Data were collected through semistructured face-to-face in-depth interviews. This study drew upon Lazarus and Folkman’s [15] stress-coping theory in designing the interview guide. Following thematic analysis, researchers used this theoretical framework to interpret patients’ cognitive evaluations and coping strategies within their web-based information-seeking behaviors, thereby enhancing the theoretical significance of the findings. Originally proposed in 1984, this classic psychological theory explains how individuals assess and respond to stressful events through a dynamic regulatory process involving person–environment interactions. The study reporting followed the COREQ (Consolidated Criteria for Reporting Qualitative Research Checklist; Checklist 1) [23].

Setting and Participants

From September 2024 to May 2025, participants were recruited through a combination of convenience and purposive sampling in 3 breast cancer inpatient wards at Hunan Cancer Hospital in China. Hunan Province is a typical province in central China, comprising 14 prefecture-level cities. Its diversity in language and culture, urban-rural structure, health care resources, economic development, and residents’ digital literacy provides a suitable setting for examining the contextual variations in information-seeking behaviors [24].

Two female researchers (LJH and ZD) were Advanced Practice Nurses in breast oncology with training in breast disease management and qualitative interviewing. The researcher has experience practicing in participants’ workplaces and is responsible for participant screening and interview scheduling. First, based on convenience sampling principles, the researcher reviewed daily inpatient records and made preliminary assessments of eligibility based on medical documentation. Subsequently, clinicians and charge nurses assisted in notifying eligible patients. Researchers then conducted face-to-face interactions with patients, explaining the study objectives,

XSL•FO
RenderX

procedures, timeline, and privacy safeguards to invite participation. To enrich the data dimensions, the research team purposefully included patients of varying ages, educational levels, disease stages, duration since diagnosis, and urban/rural residence after initial recruitment, continuing until data saturation was achieved.

Participant Recruitment and Enrollment

Participants were female patients diagnosed with breast cancer who had experience in searching for information on the internet. Inclusion criteria included the following: (1) aged 18 years and older; (2) diagnosed with breast cancer; (3) able to express themselves clearly, willing to participate in the study, and sign an informed consent form; and (4) previously actively searched for health information on the internet or mobile platforms during the disease. Exclusion criteria included the following: (1) comorbid serious mental illness affecting the interview communication; (2) those who refused to record or were unable to complete the interview.

Data Collection

Data were collected through semistructured, in-depth interviews. Based on the study aims, a review of relevant literature, and the theoretical framework of stress and coping, the interview guide was developed collaboratively by the research team members (LJH and QCL) with guidance from a qualitative research expert (LXY). To enhance clarity, exploratory questions were incorporated into the primary interview questions to better capture the depth and complexity of participants' experiences, as shown in Table 1. Two participants who met the inclusion criteria were invited for a pilot interview. Based on the feedback from the pilot participants, the wording and sequence of the interview questions were adjusted, and the exploratory questions were also optimized.

Table . Interview guide.

Main question	Exploratory questions
a. Please tell us about your motivation and initial experience of online health information searching.	• Under what circumstances did you begin using the internet to search for health information related to breast cancer? • What kind of information were you initially looking for? • What is the primary motivation for searching for information online?
b. Please tell us about the process of your information search.	• Which platforms or websites do you typically use to access online health information? • What challenges have you encountered during the information search process? • How do you screen and evaluate online health information? • Do you have any criteria or methods you rely on?
c. Please tell us about the impact of online health information on your emotions.	• How do you feel when searching for and reading online health information? • How has online health impacted you? • Could you describe your emotional experience in more detail?
d. Please tell us about the impact of online health information on your treatment decision-making.	• l How does online health information influence your treatment decisions? • How do you think searching for health information online affects your communication with your doctor? • Can you discuss how you question or offer suggestions to your doctor based on online health information?
e. Please tell us about your overall evaluation of and expectations for online health information.	• What role do you think online health information played in your journey to overcome breast cancer? • If given the opportunity, how would you improve your information-seeking process? • What kind of assistance or support would you like to receive?

The interview began after obtaining the participants' verbal and written informed consent. The entire process was audio-recorded after obtaining their consent, and at the same time, the researchers took field notes to capture nonverbal information and key immediate thoughts. During the interview, the researcher flexibly adjusted the order and depth of questions according to the participants' expressions and responses and encouraged patients to describe their information-seeking experience in-depth and truthfully. The average duration of the interviews is approximately 45 minutes (range 20 - 60 min). All interviews were conducted in quiet hospital spaces designated for patient–clinician communication to ensure privacy and comfort. To minimize interviewer bias and ensure the authenticity of participants' expressions, the researcher adopted a neutral and nonevaluative stance during the interviews, avoided leading or suggestive questions, and respected each participant's subjective experiences. When no new topics are generated, data collection is halted [25]. When no new significant themes emerge after analyzing 2 consecutive rounds of interviews, data saturation is achieved, and sampling ceases.

XSL•FO
RenderX

Data Analysis

The interview recordings were transcribed verbatim into Chinese text by the researcher (LJH) within 24 hours of completion and verified for accuracy by the researcher (ZD). Transcribed texts are managed using Excel, with each record containing participant ID, raw text, open coding, thematic categorization, and researcher notes. Excel files are uniformly stored in a controlled environment and encrypted to ensure participant information security and privacy.

Using thematic analysis as the methodological basis for theme summarization [26]. Two researchers (LJH and ZD) independently conducted the initial coding and then discussed and reached a consensus. In case of disagreement, a third-party expert (LXF) was invited to participate in the discussion to ensure the reliability and authenticity of the analysis. After coding completion, the researcher (LJH) categorized and merged relevant codes to form preliminary themes. These themes underwent discussion and review by all researchers to ensure alignment with the original data. Subsequently, themes were defined and named, with descriptive explanations drafted for each. Finally, themes were linked to corresponding original data citations. The researchers provided the preliminary analysis results to some participants for feedback and solicited their opinions to ensure the authenticity and representativeness of the analysis. To ensure the accuracy of language conversion, all citations and theme descriptions intended for publication were translated from the original Chinese into English by professional bilingual personnel with medical backgrounds. Subsequently, all members of the research team conducted back-translation and concept [27].

Reflexivity

Throughout the research process, the team maintained a reflective stance to enhance transparency, credibility, and rigor. Given that the interviewers had professional backgrounds in the breast department, face-to-face semistructured interviews fostered trust, enabling participants to be more willing to share their personal experiences. However, this also risked leading researchers to overinterpret participants' perspectives during interviews, potentially overlooking subtle variations in their experiences. Therefore, researchers maintained reflective journals before and after each interview, documenting their preconceptions, emotional responses, interview interactions, and any subjective impressions that might influence interpretation. These records were repeatedly reviewed during subsequent analysis. During data analysis, researchers repeatedly returned to original recordings to ensure initial coding stemmed from participants' authentic expressions. Researchers took particular care to separate clinical experience from the analytical process, avoiding the substitution of professional judgments for meanings presented in participants' narratives. Throughout theme development, the research team continuously challenged, compared, and refined interpretations through ongoing peer discussions and reflective dialogues. This multiperspective examination of coding consistency and thematic robustness enhanced the credibility and confirmation of research findings. Additionally, the team deliberately revisited earlier hypotheses and analytical pathways during the later stages of analysis [28].

Ethical Considerations

The study was approved by the Hunan Cancer Hospital Medical Ethics Review Committee (number SBQLL-2024 - 180). All participants signed written informed consent forms prior to the commencement of face-to-face interviews, ensuring their participation was voluntary, information was anonymized, and data were used solely for this study. Furthermore, no financial compensation was provided to any participants. All raw interview data and participant information underwent anonymization and were stored on an encrypted cloud storage drive to ensure confidentiality.

Results

Demographic Characteristics

A total of 18 female patients with breast cancer participated, originating from Wuhan, Hubei Province (n=1), Jiamusi City, Heilongjiang Province (n=1), and 9 prefecture-level cities in Hunan Province (Changsha, Yiyang, Changde, Hengyang, Loudi, Zhuzhou, Shaoyang, Yongzhou, and Yueyang) (n=16). All the participants were of Han ethnicity. All patients with breast cancer are currently undergoing treatment in the hospital. Except for one participant who had bilateral breast cancer, all the other participants had unilateral breast cancer. The time of cancer diagnosis for the participants ranged from 1 month to 20 years. The most commonly used digital search platform by the participants was Baidu. The general information of the participants is shown in Table 2.

XSL•FO
RenderX

Table . Characteristics of the participants.

Variables	Participants (N=18)
Age (years), mean (SD)	42.29 (10.60)
Education background, n (%)	
Junior high school and below	5（27.8）
Senior high school	3（16.7）
Bachelor degree	8（44.4）
Graduate degree	2（11.1）
Employment, n (%)	
Employed	6（33.3）
Unemployed	10（55.6）
Retired	2（11.1）
Marital status, n (%)	
Unmarried	1（5.56）
Married	15（83.3）
Divorced	2（11.1）
Medical insurance payment methods, n (%)	
Employee Medical Insurance	12（66.7）
Resident Medical Insurance	3（16.7）
New Rural Cooperative Medical Scheme	3（16.7）
Year of breast cancer diagnosis, n (%)	
＜1	9（50.0）
1-4	4（22.2）
5-9	2（11.1）
＞10	3（16.7）
Clinical stages, n (%)	
Stage Ⅰ-II	9（50.0）
Stage Ⅲ-IV	9（50.0）
Relapse and metastasis, n (%)	
Yes	9（50.0）
No	9（50.0）
Digital seeking platform, n (%)	
Baidu[a]	15（83.3）
XiaoHongShu[b]	5（27.8）
TikTok[c]	7（38.9）
Deepseek[d]	7（38.9）
WeChat[e]	4（22.2）
TengXunYuanBao[d]	1（5.56）

[a]Baidu: Comprehensive search engine.
[b]Xiaohongshu: Social sharing platform primarily featuring user-generated content.
[c]TikTok: Short-video-centric information dissemination platform.
[d]DeepSeek and TengXunYuanBao: Artificial intelligence-based intelligent service platform.
[e]WeChat: Comprehensive social platform.

XSL•FO
RenderX

This study identified three core themes and 11 subthemes of OHIS in patients with breast cancer through thematic analysis: (1) the driving forces of information search were emotion-driven and problem-oriented; (2) cognitive judgments in the information fog included the interweaving of multiple channels of information, assessment of information authenticity and applicability, cognitive confusion under information conflicts, and the construction of information meaning; (3) adaptive strategies under the pressure of network information encompassed information conversion into action, emotion regulation, support network construction, and expectation acceptance and adjustment. Overall, the OHIS behaviors of patients with breast cancer present a dynamic process driven by motivation, cognitive processing, and adaptive strategies. The overall coding framework, including the hierarchical relationship between themes and subthemes, is shown in Figure 1.

Figure 1. Coding framework of the web-based health information navigation process among patients with breast cancer.

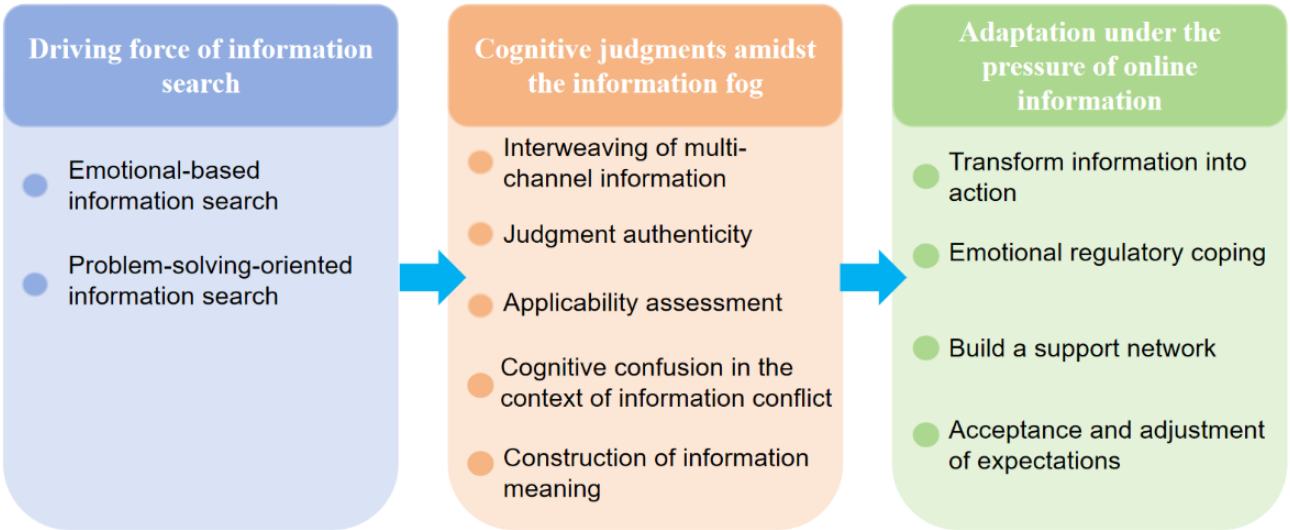

Driving Force of Information Search

Overview

This theme reflected the multiple drivers behind information-seeking behaviors of patients with breast cancer, stemming from the combined influence of emotional needs and practical concerns. On one hand, information retrieval served to alleviate uncertainty and provide emotional reassurance; on the other, patients also used web-based resources to understand their disease, support treatment decisions, and address daily management needs.

Emotion-Based Information Search

The information-seeking behavior of participants was an emotionally driven experience process. It began with the anxiety, fear, and uncertainty triggered by the diagnosis, and the participants attempted to obtain psychological comfort and a sense of control through searching.

> *I just wanna see if it's really breast cancer. It's just for some peace of mind. If it's not, I'll definitely feel better, right? I'm also really scared about myself…* [P6]

> *That kind of bad news really makes me anxious… Especially when I see stuff about having to take medicine long - term and go through all these treatments. It makes me super anxious, like the treatments just never end!* [P5]

> *I'm worried if there's still hope for a cure… 'Cause I've got old folks to take care of and kids to raise. Wanna check if there are any such cases…* [P1]

However, web-based health information not only serves a knowledge function; its content itself could directly influence patients' emotional states, becoming a significant trigger that amplifies positive or negative emotions. When search results were interpreted as "positive," patients often experienced a brief sense of hope and emotional uplift; conversely, negative or uncertain information could significantly exacerbate anxiety and emotional distress.

> *If the search results are good, I'll feel a bit happier. If they're bad, I'll definitely worry more.* [P6]

> *When you see something good, you'll feel great. But when you see something bad, it'll affect you. Your mood can easily be influenced by the internet. Especially when you come across something negative, you might feel depressed for days…* [P4]

In the collectivist culture of Chinese families, illness was regarded as a shared family event. Family members would proactively intervene in the patient's information-seeking behavior and act as an "emotional firewall" by discouraging the patient from searching for information. As a result, patients were protected and constrained by family emotions from the very beginning.

> *My husband doesn't want me to search randomly online. He said, "Don't look at random stuff on the Internet. Just follow what the doctor says.* [P6]

Problem-Solving–Oriented Information Search

The information-seeking behavior of participants revolved around a series of specific questions, reflecting their use of web-based resources as a vital supplement to inadequate clinical

XSL•FO
RenderX

information. Through information retrieval, patients sought to enhance their sense of control over the disease, understand diagnostic outcomes and disease staging, explore treatment options, and support decision-making, while also addressing symptoms, side effects, and issues related to recovery and lifestyle adjustments. The participants described their experience of understanding disease staging through web-based information.

> *I searched for it online. The doctor told me not to look it up, but I did anyway. T stands for the diameter, like the size of the tumor. So if it's T2, that means it's in category 2, right? And N1 means there's lymph node metastasis. I couldn't get an answer from the doctor, so later, when I checked my discharge record, I saw that I'm T1N1M0, which means there's no distant metastasis. After I found that out online, I was like, okay, I get it.* [P17]

In the management of symptoms and side effects, information retrieval was also used to guide practical interventions.

> *Mainly, I wanna search for the side effects of the injection. 'Cause I felt really awful after getting the red potion injected four times before. Every time, my stomach was killing me, and I'd throw up. So I wanna look it up online to see which time the side effects would be the worst and stuff. Then I'll see if the side effects will be less severe after I switch to a different medicine.* [P6]

> *Like, if my white blood cell count is low and I'm wondering what kind of medicine to take, or if I've got a bad taste in my mouth and my tongue is really white and thick and I wanna know how to relieve it, I'll just search for it online.* [P3]

Several participants described that the brevity of clinical consultations often left their information needs insufficiently addressed, leading them to rely on web-based resources for additional medical explanations.

> *You can see how busy the doctor is here. He can't talk to you for more than 10 minutes. He has so many patients, so there are many things you have to learn by yourself from other sources.* [P7]

Cognitive Judgments Amidst the Information Fog

Overview

When patients with breast cancer sought web-based health information across various platforms, they found themselves immersed in a landscape of multiple sources and diverse perspectives. This interweaving of information from different channels offered patients more clues for understanding their illness and treatment options, while also increasing the complexity of their cognitive judgment. Patients had to repeatedly weigh the authenticity of the information and its suitability for their specific disease stage and physical condition. When web-based information conflicted with existing knowledge or medical advice, they often experienced cognitive dissonance and indecision. Throughout this process, patients attempted to compare, filter, and reflect upon fragmented or conflicting information, assigning it personalized meaning in alignment with their own experiences.

Interweaving of Multichannel Information

Participants obtained health information through various web-based channels. Some initiated inquiries using general search engines, such as Baidu, and sought the truth by comparing multiple search results.

> *Yeah, I used Baidu, no other search engines. I think it's pretty reliable 'cause most of the stuff I saw was posted by doctors. Like which doctor from which hospital… For the same question, check out a few more results. If they're more or less the same, then it should be pretty accurate. Just search the same question a couple more times.* [P6]

Some participants searched for real-life experiences shared by other patients on social media platforms, while remaining highly vigilant against advertisements and potential scams.

> *Sometimes I also used Douyin. There are some people who are fellow patients, and they post videos directly on it. For example, like the Pyrrotinib I'm taking, they'll show its side - effects on Douyin, and it's more vivid there… What's on Douyin is a real reflection. Like some people do live - streaming to talk about their illness progress. They're quite honest and won't tell lies. Sometimes when I compare, it seems like I have the same symptoms too. Of course, some people might be just trying to sell stuff, so I stop watching after that…* [P7]

Some participants also viewed artificial intelligence (AI) tools as efficient and objective sources for interpreting medical terms and integrating health knowledge.

> *Besides the imaging data, I also had a carcinoembryonic antigen test. These are the two main factors, right? … I read on DeepSeek that it mainly looks at CA and CA153, 'cause these two things are characteristic of a high response to breast cancer. … My daughter checked both DeepSeek and ChatGPT for me, and the results were pretty much the same. … I think this stuff is great. It's all about the data, and now AI can be trusted. Isn't there an AI hospital opened in Shanghai now?* [P13]

Most participants tended to trust recommendations from doctors or hospital platforms, with their choices often determined by assessments of these channels' authority, convenience, and personal usage habits.

> *Can you still modify this interface? This was the essence of deep thinking. I don't often use Deepseek. I clicked it a few times, then stopped. Sometimes it got stuck, but once it got stuck, I couldn't get out of the interface.* [P12]

> *I think what Deepseek provides was a more objective perspective. They won't say something is definitely bad or good. There are no vested interests involved.* [P11]

XSL•FO
RenderX

Judgment of Information Authenticity

In assessing the authenticity of web-based health information, participants often cast themselves as information detectives, proactively using various cognitive strategies to discern truth from falsehood. Confronted with complex sources and uneven-quality web-based content, patients attempted to mitigate potential risks through personal experience and repeated verification. Some participants noted that on platforms dominated by user-generated content, the authenticity of information relied heavily on individual judgment. Since these platforms featured both content based on genuine experiences and information driven by advertising or commercial promotion, participants had to continually discern between them while reading.

> *On Xiaohongshu, you have to make your own judgment. Because there might be many that are relatively genuine, and there might also be some that are advertising-oriented, aiming to tell you what medicine to take or what to do. So they tend to write in that direction.* [P11]

To enhance the reliability of their judgments, some participants cross-verified information by conducting cross-platform or multiple searches. This reflected their active efforts to strengthen confidence in the authenticity of information through consistency validation.

> *For the same question, I would look at several results. If they are roughly the same, then they should not be too different. Search a few more times.* [P6]

Additionally, some participants emphasized the authority of information sources and the scientific rigor of content, often choosing to remain cautious when a foundation of trust was lacking.

> *What was said online can be quite confusing. You need to distinguish between what is true and what is false…Like those kinds of things on TikTok that are uncertain, where I don't have any real experience, or it's not with friends or anything like that. I don't dare to try them; I don't dare to take any risks or give up the treatment here and try something else (on the internet), because the stakes are too high.* [P2]

Several participants further developed verification strategies for authoritative personal accounts, reflecting patients' critical stance toward superficial authority.

> *First of all, I definitely need to check if this person is real… I'll look at his videos, the words he says to each patient. Where could he find so many shills? It's impossible. So many different patients.* [P14]

However, a portion of participants acknowledged that, when confronted with a complex information environment, their own judgment abilities were limited, which often left them feeling uncertain and powerless in verifying information.

> *I'm not sure whether the information on the internet is good or bad. Some of it is correct, while some is incorrect.* [P8]

Information Applicability Assessment

Upon obtaining information, participants typically did not adopt it immediately. Instead, they analyzed its applicability in light of their personal condition, treatment stage, and individual differences. One participant noted that when using generative AI tools, they did not directly follow the suggestions provided but rather treated them as supplementary reference material.

> *DeepSeek really thought things through and gave more detailed advice. For example, when I asked about eating peaches, it warned me to be cautious. Then, based on that advice and my own situation—like whether I had mouth ulcers or diarrhea—I decided whether to eat them or not. So, if I had diarrhea that day, I knew not to eat peaches because they might make it worse. I appreciated how it reminded me of these things so I could choose whether it was okay to eat or better to avoid.* [P18]

At the same time, participants recognized that web-based information was often based on group data and might not be fully applicable to individual circumstances. This awareness prompted them to maintain a certain distance from such information and to avoid applying it rigidly.

> *Regarding the internet, it is also based on big data, right? It's not an individual, either. For instance, how this person behaves might also be the case for you. It's just a proportion or percentage of the big data…* [P9]

Some participants prioritized seeking information that aligned closely with their personal needs, using this as the primary criterion for judging the information's value.

> *I won't browse random online information anymore. Whenever I come across any content suggesting a poor prognosis for triple-negative cases, I won't even bother to read it. I only look at those that are of the same type as mine and say the prognosis is relatively good.* [P14]

Several participants mentioned that they adjusted the content of their information search based on their needs at different times in treatment.

> *If after my radiotherapy is completed, how should I recover, how to exercise, learn to do sports, I will continue to search online, based on the actual situation. I need to change the search content according to my own disease development. For example, in the later stage, if I have just finished chemotherapy, I will start radiotherapy. What should I pay attention to?* [P3]

Moreover, the assessment of information relevance demonstrated clear phase-specific characteristics. Several participants indicated that the content of their information-seeking was continually adjusted throughout the treatment process.

> *After getting sick, I started searching for information about the disease online… I rested for half a year and didn't search again after returning to work. This time,*

XSL•FO
RenderX

the illness recurred, and I was more diligent in checking the information. [P1]

Cognitive Confusion in the Context of Information Conflict

When web-based health information conflicted with or contradicted treatment plans prescribed by their primary physicians, participants often experienced cognitive dissonance, which led to decision-making difficulties and a crisis of trust. In such situations, they had to repeatedly weigh conflicting advice from different authoritative sources, considerably heightening their psychological burden. Some participants described feeling confused and hesitant when they encountered discrepancies between web-based information and their doctors' advice.

> *The internet also says so. The doctors here say that radiotherapy is needed. Of course, there are doubts. Don't you think so? I'm just a patient, and I'm very confused. As long as someone says you don't have to have radiotherapy or chemotherapy, you'll be torn. It's all like this, right?* [P14]

> *It's still a very difficult decision-making process… The information on the internet is somewhat contrary to that of the doctor. It can affect the treatment plan your doctor gives you… Then your confidence in your attending doctor may decrease.* [P4]

However, when web-based information conflicted with what their doctors said, patients often kept their doubts to themselves, for fear of challenging medical authority. This reflected patients' reluctance to question professional authority within a high power-distance cultural context.

> *I didn't tell the doctor I got the info online 'cause I'm afraid the doc's gonna diss me. Like, the doc might say, "If you trust the internet so much, go ask those people online. Don't come to me.* [P4]

The direct consequence of information contradiction was a widespread deadlock in choice and a psychological burden. This contradiction left patients feeling caught in the middle.

> *Most doctors seem to say pretty much the same thing. But then it's like, traditional Chinese medicine has its own way of looking at it, and Western medicine has another.* [P2]

> *It's really tough to figure out what's true and what's not on your own—there's just so much conflicting info out there.* [P14]

Faced with decision-making difficulties, some participants would repeatedly verify and dwell on these contradictory pieces of information, reflecting their underlying desire for cognitive consistency.

> *...I've looked online, and some of them talk about how it's advanced and how it's impossible to cure this disease anyway, that you live with cancer, but one of the points I'm struggling with right now is that some of them talk about how your body will definitely break down if you do chemo, and that's 100% true. But if you don't do chemo, and this disease just continues, which option is better?* [P8]

Construction of Information Meaning

By integrating medical data with personal experiences, patients were able to derive new meaning from information. Dissatisfied with her doctor's brief explanation, one participant independently conducted a detailed assessment, comparing earlier imaging reports and using AI tools to perform calculations.

> *I myself have tested for carcinoembryonic antigen… Besides the imaging data, I also have the carcinoembryonic antigen. These are the two main factors, right?" … My daughter checked DeepSeek and ChatGPT for me, and the results were similar… We also calculated the volume… For example, for magnetic resonance imaging (MRI), I used MRI as a comparison… Then we calculated its volume… This indicates that the size has decreased… Overall, it has shrunk by 68.95%. I'm very happy that it's reduced by two-thirds… This proves that my own thinking and reaction were correct.* [P13]

After repeated comparisons and deliberation, participants developed a stable perception of the roles played by different information sources. One participant clearly distinguished the value of AI tools from that of doctors.

> *Deepseek is just a reference sometimes. It might explain things in a more detailed and nitpicky way. Doctors, on the other hand, might just go through things quickly. If you don't understand, you can ask. Docs talk so fast that you might not get what they mean at all. You can ask Deepseek, and it'll give you a more detailed explanation. But like those indicators I asked you about last time, well, you know, professionals should do professional jobs. It can't replace doctors and nurses.* [P18]

Some participants also constructed their own visions of the future based on the individual stories of their fellow patients.

> *You know, when you're scrolling through stuff these days, there are a bunch of really well - known breast cancer doctors with their own personal accounts… If you come across the good cases they share, it'll have a positive impact on you. But sometimes, if you see the not - so - good stuff, it can mess with your mood. There's this one doctor's video. He said one of his patients was in a similar situation to me, and it seemed like the prognosis was okay. That patient also lived a pretty long time.* [P4]

Through long-term engagement with information, patients gradually developed their own knowledge frameworks, transforming fragmented information into a coherent understanding of their illness.

> *I've been looking up new drugs and treatments—'cause, let's be real, everyone just wants to survive. When I first got diagnosed, I pretty much did all the research I could. I got a decent idea of the*

XSL•FO
RenderX

> *main treatment options—first-line, second-line meds, all that. So I kind of accepted, okay, I'll go through those during this year of treatment. But then I started thinking—what happens after that? Will there be any new drugs by then? Any new treatment approaches I could try next?* [P10]

Participants defined the process of information searching as an empowerment process, through which they enhanced their information literacy by means of continuous exploration and summarization.

> *Searching for reading materials is also a form of learning and it helps us, y'know. I think it's pretty good to have this kind of resource. At least it can guide me. Doctors are so busy taking care of so many patients. There's no way they can cater to every single patient's needs. Sometimes, you gotta rely on yourself.* [P7]

> *No matter how smart AI is, it's not like it can read your mind. But I do think it's something you have to learn how to use. The better you understand it, the more likely it is to give you the answers you actually want. I'm not super familiar with it yet—like, I haven't really studied it in a structured way or anything. But now there are more and more tools out there, like DeepSeek and other platforms, to help people learn.* [P11]

Through ongoing exploration, comparison, and summarization of information, these participants gradually built a firm understanding of their disease, treatment options, and personal responses. They subsequently translated this knowledge into actionable guidance for both decision-making and daily life.

Adaptation Under the Pressure of Web-Based Information

Overview

Facing the vast and complex web-based health information, patients with breast cancer have developed various adaptation strategies to alleviate the cognitive load, emotional stress, and decision-making difficulties brought about by the information. These strategies mainly fall into 4 categories: translating information into actions, emotional regulation and coping, establishing support networks, and acceptance and expectation adjustment.

Transform Information Into Action

This subtheme illustrated how participants translated acquired knowledge into concrete daily actions after obtaining, filtering, and verifying web-based health information. This process strengthened their self-management capacities and improved communication with health care providers. Some participants mentioned modifying their diets based on recommendations from tools like DeepSeek, tailoring the information to their individual health conditions.

> *Like, Deepseek says you gotta be careful when eating peaches. If you've got oral ulcers or diarrhea, don't eat 'em. So I'll judge based on how I'm feeling today. If I'm having diarrhea, I know not to eat 'em. But if my appetite's normal, no oral ulcers, and I don't feel nauseous, then I can eat 'em. So I can choose whether to eat 'em or not according to its suggestions.* [P18]

In information-driven self-care practices, some participants proactively integrated traditional Chinese health preservation wisdom, regarded as a crucial action resource, into their daily health management.

> *I watched a video on Video Account about the Three-Bean Drink. It seems pretty good. I mean, it's basically a food therapy, you know?* [P18]

In medical communication, participants drew on web-based information to engage with their physicians in seeking personalized guidance. This practice demonstrated how patients translated external information into concrete actions and interaction strategies, thereby obtaining a sense of psychological reassurance and decision support through such exchanges.

> *I saw someone on Xiaohongshu saying that taking Sanguisorba Tablets can increase white blood cell count. So I went to ask my doctor if I could take those pills. I was like, "Doc, I saw this on Xiaohongshu, and I even took a screenshot." Then I showed it to my doctor.* [P12]

Participants further integrated web-based rehabilitation information into their daily training routines, achieving a phased transition from cognition to action. Through practice, they continuously tested and adjusted the applicability of the information while cultivating a sense of self-efficacy. This process reinforced the connection between information empowerment and tangible rehabilitation outcomes.

> *Sometimes people on the Internet tell me to do exercises. My hands have recovered better than most people because I do exercises well. They tell me what exercises to do online, and I do them for 40 minutes. I jog for 20 minutes and do exercises for another 20 minutes. Actually, I stick to it every day. On the third and fourth days, I start with 10 minutes, then 15 minutes, and then 20 minutes. I gradually increase the time.* [P3]

Emotional Regulatory Coping

Faced with vast and often complex web-based health information, participants used multiple strategies to regulate their emotions, alleviate anxiety, and restore psychological equilibrium. These strategies encompassed both the active pursuit of positive information and the conscious limitation of exposure to content in order to mitigate the negative emotions elicited by web-based sources. For instance, some participants sought psychological comfort by deliberately reviewing positive content and reading about the recovery experiences of fellow patients.

> *I think the cure rate for breast cancer was relatively low before. I saw online that the survival period is now over five years, and the cure rate has reached several tens of percent. I think the current cure rate is quite high. I think this boosts my confidence… Because I have triple-negative breast cancer. But I*

XSL•FO
RenderX

> *saw that many of my friends with triple-negative cancer, as long as they survive for these five years, basically, they won't have a recurrence. This gives me a lot of confidence.* [P3]

Additionally, some participants maintained emotional stability through self-motivation and psychological self-regulation. They not only regulated their emotions by acquiring information but also reinforced positive feelings and beliefs through intentional behavioral practices.

> *I will browse the internet… I will copy some good sentences to encourage myself. Now, even a doctor shouting slogans has become quite famous on the video platform. On it, I have also learned to shout slogans. This mental victory method can change my inner emotional state.* [P13]

A small number of participants also halted their information retrieval, reduced their search frequency, and avoided specific platforms or negative topics to mitigate the emotional interference caused by web-based information.

> *I've searched before, and I don't do it anymore now. Now… even if I search, it won't be of any use.* [P5]
>
> *But the more I searched, the more anxious I became, and the more I didn't want to look.* [P1]

For some of the participants, the cell phone was never far away, and they could only reduce searching or reading web-based information by diverting their attention. However, there were also participants who had to passively accept web-based information.

> *I don't even want to search anymore. I just watch some funny videos every day.* [P 2]
>
> *Anyway, it's just for some gaming… Sometimes, it would come up on TikTok by itself, but I still wanted to check if there were any similar things related to me, like what others use for treatment. Just to see what others are posting on TikTok… Without searching, it just shows up on TikTok.* [P1]

Build a Support Network

Participants sought assistance from others within complex information environments to alleviate their decision-making pressure and emotional burdens.

> *I joined those patient groups. Then I'd listen to what they said and also do some searches on the group. There's a lot of info there. You know, I just wanna learn more about it.* [P17]

Patients often referred to the experiences of fellow patients and sought decision-making advice from doctors or nursing staff, or had family members screen and verify information on their behalf.

> *Seeing this information, one would surely be confused about how to choose this path and which one to take… But my daughter will always help me make the right decision.* [P8]
>
> *I talked to the doctor about online info. Like, I asked if I could have radiotherapy for bone metastases. He said I didn't need it and should just listen to doctors. Doctors have seen more cases. What I've seen is just one-sided.* [P7]

Acceptance and Adjustment of Expectations

Through information acquisition and practical experience, some participants gradually adjusted their expectations regarding treatment outcomes, shifting their focus from defeating the disease toward managing specific symptoms and maintaining daily functioning.

> *I just checked online, and some info said it's in the advanced stage. Anyway, it's impossible to cure your illness. You gotta live with this cancer…* [P8]

Upon recognizing the gap between their own knowledge and the professional system, participants learned to delegate decision-making authority to professional authorities after thoroughly exploring information.

> *Anyway, when you're in a tough spot trying to make a decision, I always think doctors have more experience than you, right? You've been looking up info for just a few months, but doctors have seen tons of patients. You gotta trust them as much as you can.* [P4]

Discussion

Summary of Results

The findings of this study authentically reflect the web-based information-seeking experiences of patients with breast cancer, which can be summarized under 3 core themes: drivers of information seeking, cognitive judgments amid information fog, and adaptive mechanisms under web-based information pressure. These themes are not isolated categories but rather depict a dynamic psychosocial process: how patients navigate a challenging digital information environment from their initial needs to ultimately develop personalized coping strategies.

Findings

This study found that information-seeking behavior among patients with breast cancer is driven by both emotional and problem-solving needs. The anxiety and uncertainty brought by a diagnosis drive emotional information-seeking centered on seeking reassurance, while simultaneously prompting problem-solving searches to understand the disease and participate in decision-making. Consistent with the classification of web-based help-seeking behaviors of patients with cancer in previous studies, the study by Feng et al [29] mainly focused on web-based health communities and categorized patients' help-seeking behaviors into informational and emotional support needs. This study further expands this perspective by extending the scope of analysis from help-seeking behaviors in web-based community interactions to a broader context of OHIS. Notably, these 2 drivers do not switch off and on simply; they often coexist and reinforce each other. This finding resonates with research by Liu et al [30], based on a dual-factor model, which confirmed that pediatric patients' OHISs are simultaneously influenced by complex factors—both facilitators (such as perceived stress) and inhibitors (such as learned helplessness),

XSL•FO
RenderX

revealing the inherent tension within behavioral drivers. In the Chinese context, this driving force is further moderated by the family system. Phenomena observed in this study, such as spousal discouragement and proxy searching by children, align with Feng's [31] findings that Chinese family communication patterns influence information-seeking behaviors undertaken for family members or oneself. This highlights how, within collectivist cultures, disease management is often viewed as a shared family unit event.

This study reveals patients' cognitive judgment processes within uncertain information environments. Patients develop strategies such as cross-verifying multiple channels, tracing information sources, and assessing applicability to construct personal cognitive safety nets amid information fog. The core paradox of this process lies in the dual role of information: while it may enhance patients' sense of control over their illness, excessive web-based information, poor quality content, and conflicting viewpoints can also trigger decision-making difficulties, diminished trust in information, and persistent anxiety [32,33]. This study particularly highlights the critical role of doctor-patient communication within the information fog. When patients conceal their web-based doubts for fear of challenging medical authority, conflicting information becomes internalized as a psychological burden. This finding strongly corroborates the study by Mahmood et al [34], which demonstrated that patient-centered communication significantly reduces frustration during information-seeking among cancer survivors. It cannot be overlooked that while OHIS has to some extent challenged the traditional paternalistic medical model, its role in clinical practice is increasingly being reinterpreted. Far from threatening physician authority, it may actually enhance the depth and quality of doctor-patient communication [35]. Furthermore, patients' heightened focus on dietary adjustments and nutritional therapies when evaluating information is deeply rooted in local health beliefs and traditional practices. This provides empirical support for moving beyond universal health literacy models and constructing culturally sensitive analytical frameworks.

In contrast to the results of the review by Chen et al [8], our study found that Chinese patients with breast cancer did not use the websites of national or nonprofit cancer organizations to search for health information. Chen et al [8] showed that patients with breast cancer, including those in China, often relied on these websites for information. This disparity may indicate that the patients in this sample have relatively low digital health literacy and have limited access to information through professional web-based channels. Digital health literacy specifically refers to a series of technical abilities and cognitive skills that individuals need to discover, understand, evaluate, and apply web-based health information [36]. The patients with breast cancer in this study's reliance on social media and general search engines may directly indicate their deficiencies in the 2 key aspects of digital literacy, namely identifying the authority of information and conducting effective web-based navigation. In this context, digital health care approaches, such as hospital official apps, web-based patient communities, or intelligent question-answering systems, can provide patients with alternative channels for information acquisition, thereby reducing the obstacles caused by insufficient digital literacy [37,38]. The demand of patients with cancer for digital health functions is closely related to their digital health literacy levels, and there are also differences in the preferences for digital health interventions among patients with different literacy levels. This indicates that digital literacy is an important factor influencing the use of digital nursing tools [39]. Through these platforms, patients can not only access timely and reliable health information but also pose questions, verify information, and receive personalized advice in an interactive environment, which in practice contributes to enhancing their ability to understand and apply information. Meanwhile, research indicates that there are significant disparities in digital health literacy among breast cancer patients. Patients with low literacy levels face greater difficulties in participating in eHealth interventions, highlighting the need to consider patients' literacy levels when designing digital nursing resources [40].

One of our findings is that the web-based information search behavior of patients with breast cancer changes with the disease stage, similar to the results of Madge, but their research focused on patients who had undergone breast cancer surgery [41]. Our participants had a wider range of disease stages, including both patients who had just undergone breast cancer surgery and those with recurrent breast cancer. The phased changes in information search behavior were more significant, suggesting that web-based health information intervention should be time-sensitive.

What is worth noting is that many participants with breast cancer, when searching for web-based health information, not only used traditional search engines and social media but also showed a high preference and trust in AI platforms. As shown in related studies, while traditional web-based resources remain dominant, LLM-based chatbots are becoming a resource for health information for some users, especially those who are younger and have a higher level of trust in AI [42]. However, this AI preference also comes with potential risks. The study found that the web-based health advice provided by AI models contains information misguidance and inaccurate information sources [43]. It can be seen that AI preference may trigger new cognitive trust traps. Therefore, we suggest that when promoting AI-assisted information services, it must be incorporated into the collaborative doctor-patient communication process, and through educational courses, web-based lectures, etc, improve patients' digital health literacy, thereby enhancing the accuracy of information acquisition and decision-making quality for patients [44].

Theoretical Extensions in Web-Based Health Information Contexts

This study's findings align with key processes in the stress-coping theory, thereby providing empirical support for understanding this theory in digital health information contexts. Concurrently, this study theorizes the complex web-based health information environment itself as a significant digital stressor, thereby expanding the application context of stress and coping theories. Findings indicate that for patients, this environment serves not only as a repository of resources for seeking support but may also constitute a persistent source of psychological stress due to information overload, contradictions, and

XSL•FO
RenderX

uncertainties. In essence, an overwhelming or inconsistent volume of information can trigger anxiety and stress responses, potentially leading to avoidance behaviors [45]. Research findings also indicate that information avoidance and reduced retrieval frequency are not merely passive reactions, but rather contextualized emotion regulation strategies. This discovery complements traditional coping classifications' understanding of avoidance strategies. Unlike the original theory's focus on immediate coping, this study found that patients construct meaning around their illness through repeated comparison and integration of information from multiple sources, thereby adjusting their expectations for the future. This demonstrates the long-term and developmental nature of the coping process. It is worth noting that since this theory was not originally designed for digital health information scenarios, some new phenomena related to web-based information conflict and digital uncertainty lack direct conceptual counterparts within the theory. The findings of this study provide new perspectives on the applicability and extension of this theory within digital health information environments.

Limitations

This study has the following limitations. Although most participants came from different cities and prefectures within Hunan Province of China, there are differences in language, economic development, and urban-rural structure within the province. However, the sample size of the qualitative research is limited, and it may still not cover the diversity of digital literacy and socioeconomic backgrounds. Participants were patients from a specialized cancer hospital, and these findings may not fully reflect the experiences or perspectives of health care providers in other medical systems or regions with differing technological infrastructure. Although semistructured in-depth interviews were adopted to obtain data, the respondents' recollections might have a selective bias, and some emotional or behavioral experiences might not be accurately expressed or recalled. At the same time, qualitative research itself has a strong interpretative and subjective nature, and different analysts might form different understandings of the same material. This study was conducted during a treatment stage of patients with breast cancer and did not longitudinally track the web-based information search process and behaviors of patients with breast cancer. Future research could further focus on the differences in information coping patterns among people from different cultures and with different health literacy levels, and also examine how coping strategies evolve dynamically over the course of the treatment from a longitudinal perspective. This study analyzed the research results based on the theory of stress and coping. It provided theoretical support for understanding web-based information-seeking behavior. However, these theories may have limitations in terms of applicability in the specific cultural context of China. Future research could further validate these theories by integrating cultural adaptation theories.

Implications of Research Findings for Policy and Future Research

This study reveals the complex experiences of patients with breast cancer in navigating web-based health information in the digital age. These findings not only deepen our understanding of patients' health information behaviors but also provide significant insights for improving health policies, optimizing clinical practices, and guiding future research directions. The key points are as follows:

1. Promote the construction of a health information support system: Health policy-makers and medical institutions should take the lead in building national or regional official disease information and support platforms. These platforms should integrate evidence-based medical guidelines, treatment pathway descriptions, side-effect management, and rehabilitation knowledge and present them in easily understandable forms such as videos and interactive question-and-answer sessions to reduce information risks.
2. Incorporate patient information behavior management into routine nursing and doctor-patient communication norms: It is recommended that modules for assessing and coaching patients' health information literacy be added to nursing practice guidelines and doctor-patient communication training. Health care providers should proactively inquire about patients' information sources and concerns and provide tailored information prescriptions. They should also guide patients in translating web-based information into safe, actionable self-management behaviors.
3. Conduct digital health literacy education: In the future, health information literacy education courses should be developed for patients with cancer. Special attention should be paid to digitally vulnerable groups such as older people and those with low educational levels, and barrier-free digital health literacy intervention measures should be designed to promote fairness in health information access.
4. Deepen local research and interdisciplinary cooperation: Future research should explore in depth how culture-specific factors shape information behavior patterns and carry out interdisciplinary cooperation among nursing, psychology, information science, and computer science. In the future, intelligent health information question and answer and decision-making assistance tools based on AI and in line with the Chinese cultural context can be developed.

Conclusions

The web-based information-seeking experience of Chinese patients with breast cancer is a dynamic process involving motivation, cognition, and adaptation. Their search behavior stems from dual needs: emotional drive and problem-solving. Within a complex and conflicting information environment, they assess authenticity and applicability while constructing meaning. Through strategies such as information transformation, emotional regulation, support network building, and expectation adjustment, they achieve individual adaptation. Theoretically, this finding enriches the research framework of digital health literacy and information behavior and provides empirical evidence for understanding patients' coping mechanisms in a complex information environment. In practice, this study suggests that clinical and nursing staff should pay attention to patients' emotional needs, cognitive judgment abilities, and adaptation strategies when designing digital health interventions and information support strategies. They should provide multichannel, easy-to-understand, and actionable information resources and combine social support and emotional regulation

XSL•FO
RenderX

guidance to help patients obtain, understand, and apply health information more effectively.

Acknowledgments

The authors extend their gratitude to the participants for sharing their experiences to make this study possible. As a nonnative English-speaking research team, we have made limited use of generative artificial intelligence tools for English language polishing and enhancement to ensure our research findings can be disseminated globally. Beyond this, no artificial intelligence was used in the creation of the remaining content.

Funding

This research was supported by the High-Level Talent Support Program of Hunan Cancer Hospital and the 2025 Hunan Normal University-Hunan Cancer Hospital Joint Research Project (grant: HNUCRI2025010).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

LJH and QCL contributed to conceptualization. LJH and ZD contributed to the data collection. LJH, LXF, and ZD conducted a thematic analysis. QCL contributed to project administration. LJH managed resources. QCL, CJJ, and LXY performed supervision. LJH and ZD contributed to writing the original draft. All authors contributed to reviewing and editing. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Checklist 1
COREQ checklist.
[DOCX File, 23 KB - jmir_v28i1e80363_app1.docx]

References

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021 May;71(3):209-249. [doi: 10.3322/caac.21660] [Medline: 33538338]
2. Caetano Dos Santos FL, Michalek IM, Wojciechowska U, Didkowska J. Changes in the survival of patients with breast cancer: Poland, 2000-2019. Breast Cancer Res Treat 2023 Feb;197(3):623-631. [doi: 10.1007/s10549-022-06828-5] [Medline: 36509986]
3. Quang DT, Luong Thi T, Nguyen Di K, Vu Thi Quynh C, Nguyen Thi Hoa H, Phan Ngoc Q. Illuminating the breast cancer survival rates among Southeast Asian women: a systematic review and meta-analysis spanning four decades. Curr Probl Cancer 2024 Feb;48:101062. [doi: 10.1016/j.currproblcancer.2024.101062] [Medline: 38309146]
4. Kimiafar K, Sarbaz M, Shahid Sales S, Esmaeili M, Javame Ghazvini Z. Breast cancer patients' information needs and information-seeking behavior in a developing country. Breast 2016 Aug;28:156-160. [doi: 10.1016/j.breast.2016.05.011] [Medline: 27318171]
5. Sheehy EM, Lehane E, Quinn E, Livingstone V, Redmond HP, Corrigan MA. Information needs of patients with breast cancer at years one, three, and five after diagnosis. Clin Breast Cancer 2018 Dec;18(6):e1269-e1275. [doi: 10.1016/j.clbc.2018.06.007] [Medline: 30153977]
6. Ludwigson A, Huynh V, Vemuru S, et al. Characterizing informational needs and information seeking behavior of patients with breast cancer. Am J Surg 2024 Jan;227:100-105. [doi: 10.1016/j.amjsurg.2023.09.047] [Medline: 37863795]
7. Deng J, Savjani RR, Lee P. Online search behavior for cancer immunotherapy resources and readability analysis: an opportunity to aid in medical decision-making. Cureus 2019 Oct 7;11(10):e5857. [doi: 10.7759/cureus.5857] [Medline: 31763080]
8. Chen J, Duan Y, Xia H, Xiao R, Cai T, Yuan C. Online health information seeking behavior among breast cancer patients and survivors: a scoping review. BMC Womens Health 2025 Jan 3;25(1):1. [doi: 10.1186/s12905-024-03509-x] [Medline: 39754199]
9. Foroughi F, Lam AKY, Lim MSC, Saremi N, Ahmadvand A. "Googling" for cancer: an infodemiological assessment of online search interests in Australia, Canada, New Zealand, the United Kingdom, and the United States. JMIR Cancer 2016 May 4;2(1):e5. [doi: 10.2196/cancer.5212] [Medline: 28410185]

XSL•FO
RenderX

10. Ferraris G, Monzani D, Coppini V, et al. Barriers to and facilitators of online health information-seeking behaviours among cancer patients: a systematic review. Digit Health 2023;9:20552076231210663. [doi: 10.1177/20552076231210663] [Medline: 38107979]
11. Zhao YC, Zhao M, Song S. Online health information seeking among patients with chronic conditions: integrating the health belief model and social support theory. J Med Internet Res 2022 Nov 2;24(11):e42447. [doi: 10.2196/42447] [Medline: 36322124]
12. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. Annu Rev Public Health 2020 Apr 2;41:433-451. [doi: 10.1146/annurev-publhealth-040119-094127] [Medline: 31874069]
13. Southwell BG, Niederdeppe J, Cappella JN, et al. Misinformation as a misunderstood challenge to public health. Am J Prev Med 2019 Aug;57(2):282-285. [doi: 10.1016/j.amepre.2019.03.009] [Medline: 31248741]
14. Lim HM, Wong SS, Yip KC, et al. Online health information-seeking behaviour of patients attending a primary care clinic in Malaysia: a cross-sectional study. Fam Pract 2022 Jan 19;39(1):38-45. [doi: 10.1093/fampra/cmab099] [Medline: 34423368]
15. Lazarus RS, Folkman S. Stress, Appraisal, and Coping: Springer; 1984.
16. Folkman S, Moskowitz JT. Coping: pitfalls and promise. Annu Rev Psychol 2004;55(745-74):745-774. [doi: 10.1146/annurev.psych.55.090902.141456] [Medline: 14744233]
17. Lambert WW, Lazarus RS. Psychological stress and the coping process. Am J Psychol 1970 Dec;83(4):634. [doi: 10.2307/1420698]
18. Mehrabizadeh M, Zaremohzzabieh Z, Zarean M, Ahrari S, Ahmadi AR. Narratives of resilience: understanding Iranian breast cancer survivors through health belief model and stress-coping theory for enhanced interventions. BMC Womens Health 2024 Oct 8;24(1):552. [doi: 10.1186/s12905-024-03383-7] [Medline: 39375669]
19. Ko E, Lee Y. The effects of coping strategies between uncertainty and quality of life of Korean women with gynecological cancer: evaluation of uncertainty in illness theory and stress and coping theory. ANS Adv Nurs Sci 2024;47(3):E84-E95. [doi: 10.1097/ANS.0000000000000507] [Medline: 37498078]
20. Oh YS. Communications with health professionals and psychological distress in family caregivers to cancer patients: a model based on stress-coping theory. Appl Nurs Res 2017 Feb;33(5-9):5-9. [doi: 10.1016/j.apnr.2016.09.008] [Medline: 28096023]
21. Melhem SJ, Nabhani-Gebara S, Kayyali R. Evaluating online health information utilisation and its psychosocial implications among breast cancer survivors: qualitative explorations. Health Promot Perspect 2024 Mar;14(1):61-69. [doi: 10.34172/hpp.42682] [Medline: 38623349]
22. Neubauer BE, Witkop CT, Varpio L. How phenomenology can help us learn from the experiences of others. Perspect Med Educ 2019 Apr;8(2):90-97. [doi: 10.1007/s40037-019-0509-2] [Medline: 30953335]
23. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
24. Guo Y, Xi S, Zhang M. Provincial division of economic zones based on the improved urban gravity model: a case study of Hunan Province, China. PLoS ONE 2021;16(12):e0261205. [doi: 10.1371/journal.pone.0261205]
25. Kerr C, Nixon A, Wild D. Assessing and demonstrating data saturation in qualitative inquiry supporting patient-reported outcomes research. Expert Rev Pharmacoecon Outcomes Res 2010 Jun;10(3):269-281. [doi: 10.1586/erp.10.30] [Medline: 20545592]
26. Braun V, Clarke V, Hayfield N, Terry G. Thematic analysis. In: Liamputtong P, editor. Handbook of Research Methods in Health Social Sciences: Springer Singapore; 2019:843-860. [doi: 10.1007/978-981-10-5251-4_103]
27. Chen HY, Boore JR. Translation and back-translation in qualitative nursing research: methodological review. J Clin Nurs 2010 Jan;19(1-2):234-239. [doi: 10.1111/j.1365-2702.2009.02896.x] [Medline: 19886874]
28. Olmos-Vega FM, Stalmeijer RE, Varpio L, Kahlke R. A practical guide to reflexivity in qualitative research: AMEE guide no. 149. Med Teach 2022 Apr 7:1-11. [doi: 10.1080/0142159X.2022.2057287] [Medline: 35389310]
29. Feng X, Hu Y, Pfaff H, Liu S, Wang H, Wang Y. Emotional or informational? understanding cancer patients' multifaceted help-seeking behaviors in online health communities. Soc Sci Med 2025 Apr;371:117902. [doi: 10.1016/j.socscimed.2025.117902] [Medline: 40056535]
30. Liu F, Kong X, Xia T, Guo H. Exploring the impact of discharged patients' characteristics on online health information-seeking behaviors: insights from patients' dilemmas. Front Psychol 2025;16:1500627. [doi: 10.3389/fpsyg.2025.1500627] [Medline: 40066181]
31. Feng Y. For myself or for others? the influence of family communication patterns on family health history communication and online health information seeking. J Health Commun 2025;30(1-3):72-81. [doi: 10.1080/10810730.2025.2450617] [Medline: 39804582]
32. Wilhelm E, Vivilaki V, Calleja-Agius J, et al. Effects of the modern digital information environment on maternal health care professionals, the role of midwives, and the people in their care: scoping review. J Med Internet Res 2025 Feb 25;27:e70108. [doi: 10.2196/70108] [Medline: 39998875]

XSL•FO
RenderX

33. Zheng H, Chen X, Jiang S, Sun L. How does health information seeking from different online sources trigger cyberchondria? the roles of online information overload and information trust. Inf Process Manag 2023 Jul;60(4):103364. [doi: 10.1016/j.ipm.2023.103364]
34. Mahmood A, Kedia S, Ogunsanmi DO, Kabir U, Entwistle C. Patient-centered communication and cancer information-seeking experiences among cancer survivors: a population-based study in the United States. Patient Educ Couns 2025 Jun;135:108710. [doi: 10.1016/j.pec.2025.108710] [Medline: 40010060]
35. Starcevic V. The impact of online health information seeking on patients, clinicians, and patient-clinician relationship. Psychother Psychosom 2024;93(2):80-84. [doi: 10.1159/000538149] [Medline: 38531342]
36. van der Vaart R, Drossaert C. Development of the digital health literacy instrument: measuring a broad spectrum of health 1.0 and health 2.0 skills. J Med Internet Res 2017 Jan 24;19(1):e27. [doi: 10.2196/jmir.6709] [Medline: 28119275]
37. Lazarou I, Krooupa AM, Nikolopoulos S, et al. Cancer patients' perspectives and requirements of digital health technologies: a scoping literature review. Cancers (Basel) 2024 Jun 21;16(13):2293. [doi: 10.3390/cancers16132293] [Medline: 39001356]
38. Kirsch EP, Kunte SA, Wu KA, et al. Digital health platforms for breast cancer care: a scoping review. J Clin Med 2024 Mar 27;13(7):1937. [doi: 10.3390/jcm13071937] [Medline: 38610702]
39. Yoo SH, Sung JH, Lee K, et al. The needs for digital health and eHeath literacy of cancer patients, caregivers, and healthcare providers: a multicenter, descriptive correlational study. Eur J Oncol Nurs 2024 Jun;70:102581. [doi: 10.1016/j.ejon.2024.102581] [Medline: 38749385]
40. Moon Z, Zuchowski M, Moss-Morris R, Hunter MS, Norton S, Hughes LD. Disparities in access to mobile devices and e-health literacy among breast cancer survivors. Support Care Cancer 2022 Jan;30(1):117-126. [doi: 10.1007/s00520-021-06407-2] [Medline: 34236506]
41. Madge OL, Marinca AM, Daha C, Simion L. Health information seeking behaviour and decision making by patients undergoing breast cancer surgery: a qualitative study. Health Info Libr J 2025 Mar;42(1):42-53. [doi: 10.1111/hir.12480] [Medline: 36861854]
42. Yun HS, Bickmore T. Online health information-seeking in the era of large language models: cross-sectional web-based survey study. J Med Internet Res 2025 Mar 31;27:e68560. [doi: 10.2196/68560] [Medline: 40163112]
43. Yau JYS, Saadat S, Hsu E, et al. Accuracy of prospective assessments of 4 large language model chatbot responses to patient questions about emergency care: experimental comparative study. J Med Internet Res 2024 Nov 4;26:e60291. [doi: 10.2196/60291] [Medline: 39496149]
44. Álvarez-Pérez Y, Duarte-Díaz A, Toledo-Chávarri A, et al. Digital health literacy and person-centred care: co-creation of a massive open online course for women with breast cancer. Int J Environ Res Public Health 2023 Feb 22;20(5):3922. [doi: 10.3390/ijerph20053922] [Medline: 36900935]
45. Soroya SH, Farooq A, Mahmood K, Isoaho J, Zara SE. From information seeking to information avoidance: understanding the health information behavior during a global health crisis. Inf Process Manag 2021 Mar;58(2):102440. [doi: 10.1016/j.ipm.2020.102440] [Medline: 33281273]

Abbreviations

AI: artificial intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research Checklist
OHIS: online health information search

Edited by A Stone; submitted 09.Jul.2025; peer-reviewed by M Rosenlund, R Grasso; accepted 30.Jan.2026; published 26.Mar.2026.

Please cite as:
Liu J, Li X, Li X, Chen J, Zhang D, Qiu C
From Searching to Coping, How Chinese Patients With Breast Cancer Navigate Web-Based Health Information: Semistructured Interview Study
J Med Internet Res 2026;28:e80363
URL: https://www.jmir.org/2026/1/e80363
doi:10.2196/80363

©Jiahui Liu, Xingfeng Li, Xuying Li, Jiejun Chen, Dan Zhang, Cuiling Qiu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Sociodemographic Drivers of Recruitment and Attrition in Digital Neurological Research: Longitudinal Cohort Study

Peyman Nejat[1,2], MD; Ashley D Bachman[3], BS; Vicki M Stubbs[3], BS; Joseph R Duffy[3], PhD; John L Stricker[4], MS, PhD; Vitaly Herasevich[1,2], MD, PhD; David T Jones[3], MD; Rene L Utianski[3], PhD; Hugo Botha[2,3], MBChB

[1]Department of Anesthesiology and Perioperative Medicine, Mayo Clinic, Rochester, MN, United States

[2]Office of Digital Innovation, Center for Clinical And Translational Science, Mayo Clinic, Rochester, MN, United States

[3]Department of Neurology, Mayo Clinic, 200 First St SW, Rochester, MN, United States

[4]Department of Information Technologies, Mayo Clinic, Rochester, MN, United States

Corresponding Author:
Hugo Botha, MBChB
Office of Digital Innovation, Center for Clinical And Translational Science, Mayo Clinic, Rochester, MN, United States

Abstract

Background: Digital recruitment methods offer opportunities to address challenges in clinical research participation, particularly in neurology. However, the impact of digital approaches across socioeconomic and demographic groups remains inadequately understood.

Objective: This study investigates the influence of sociodemographic factors on recruitment and attrition in a remote neurological research cohort, mapping participation pathways and identifying disparities to inform inclusive digital strategies.

Methods: We conducted a nonexperimental, observational longitudinal cohort study at Mayo Clinic using patient-portal invitations between March and July 2024 as part of a remote speech capture study. Eligibility criteria included age 18 years and older, US residence, and English proficiency. Of 5846 invited patients, progression was tracked across checkpoints (invitation, eligibility screening, electronic consent, and task completion) using Epic (Epic Systems Corporation) to obtain demographic information, Qualtrics (Qualtrics, LLC) for screening, PTrax (a Mayo Clinic–developed Participant Tracking System) for consent tracking, and the recording platform. Socioeconomic context was assessed using the Housing-based Socioeconomic Status (HOUSES) index, where higher values indicate higher socioeconomic status, and the Area Deprivation Index (ADI), where higher values reflect greater neighborhood disadvantage. Data diagnostics included Anderson-Darling tests for non-normality and Little missing completely at random (MCAR) test to characterize missingness. Associations between participation outcomes and age, sex, urbanicity, and socioeconomic indices were examined using nonparametric tests. Exact *P* values and 95% CIs are reported. Analyses were conducted using BlueSky Statistics (BlueSky Statistics, LLC) and the Python *SciPy* package.

Results: Overall, 415 out of 5846 participants (7.1%) completed all study requirements. Completers were older (median age 66.4, IQR 56.0-72.5; 95% CI 65.1 - 67.6 years) than noncompleters (median age 62.8, IQR 47.5-72.7; 95% CI 62.2 - 63.2 years; $P<.001$). Participants from more socioeconomically disadvantaged neighborhoods were less likely to respond (invitation nonresponder median ADI 45.0, IQR 29.0-63.0 vs interested median ADI 42.0, IQR 27.0-59.0; $P<.001$), and completers had slightly lower ADI ranks than noncompleters (median 41.0, IQR 27.0-56.0 vs median 44.5, IQR 28.0-62.0; $P=.04$). Urban participants enrolled faster (median 32.0, IQR 9.0-58.0; 95% CI 31.0 - 37.0 days) than rural (median 41.0, IQR 22.0-65.0; 95% CI 37.0 - 49.0 days; $P=.01$). Female participants responded slower (median 38.5, IQR 14.8-66.3; 95% CI 35.0 - 41.0 days) than males (median 32.0, IQR 8.0-57.5; 95% CI 29.0 - 38.0 days; $P=.01$). No significant differences were observed for the HOUSES index, and device type was unrelated to completion or timelines. Missingness for key variables was completely at random (MCAR $\chi^2_3=3.45$; $P=.24$).

Conclusions: Digital recruitment does not overcome traditional barriers to participation and may introduce new disparities related to age, urbanicity, and neighborhood disadvantage. These findings inform inclusive digital research strategies, including multichannel outreach, age-specific engagement, and rural technical support. This study applies longitudinal pathway analysis to digital neurology recruitment, offering actionable insights for improving inclusivity in remote research.

(J Med Internet Res 2026;28:e83432) doi:10.2196/83432

KEYWORDS

digital recruitment; neurological research; participation disparities; sociodemographic factors; digital divide

XSL•FO
RenderX

Introduction

Participant recruitment is a critical component of medical research studies, and the rate of participation varies significantly depending on the type of study, the setting, and the population involved. As an example, recent data suggest that enrollment rate in cancer trials is around 6.3% to 7.1% [1,2]. Enrollment rates may also differ among different populations; specifically, they can be lower among minority, pediatric, and geriatric populations [3,4]. Poor recruitment often results in underpowered studies, which lack the necessary sample size to detect meaningful differences between distinct groups. This can lead to statistically nonsignificant results even when there are clinically relevant effects [5]. Despite these efforts and the growing adoption of digital recruitment tools, recent analyses confirm that these challenges persist with participation patterns continuing to vary by age, sex, and socioeconomic status (SES) [6]. This underscores a critical point that digital methods may not fully mitigate traditional barriers and could introduce new inequities that influence who participates in research.

Digital tools, including social media platforms, mobile apps, electronic health records, patient portals, and electronic consent, have expanded the reach of research recruitment efforts by enabling targeted outreach to specific demographics and geographic areas [7-10]. These approaches can engage populations previously hard to reach through traditional methods, but evaluations report inconsistent effects on diversity and efficiency across settings [11]. A recent systematic review cataloged the spectrum of digital technologies deployed for recruitment and highlighted the still-limited evidence that any one approach reliably improves inclusion of underrepresented groups [6].

A central challenge for digital recruitment is the digital divide, defined as persistent inequities in broadband access, device availability, and digital literacy that map closely onto socioeconomic, geographic, and age-related lines [12]. These inequities can depress response rates and create systematic attrition in specific subgroups even when outreach is delivered digitally. Recent studies and reviews caution that digitalization can reproduce or widen participation gaps if considerations are not built into recruitment strategies [13]. Within health care systems, portal-based recruitment shows promise but exhibits differential response patterns across patient characteristics, underscoring the need to track impacts on disparity as programs scale. Moreover, scoping reviews in specific domains suggest that mixed, multichannel recruitment strategies (digital and offline) may improve inclusivity compared with relying on a single modality [11]. These disparities are often more pronounced among disadvantaged groups, including racial and ethnic minorities and older adults, who may have limited access to the internet and lower digital literacy levels [14-16]. For instance, studies have shown that older adults and African American patients are less likely to use digital health portals compared to their younger and White counterparts, highlighting a significant gap in technology use [16]. Here, the term “White” reflects the racial classification used in the original study, which reported race and ethnicity separately and did not provide ethnicity-specific breakdowns. Similarly, individuals from lower SES neighborhoods often have reduced access to the internet and lower health literacy, which can hinder their ability to engage with digital health technologies effectively [15]. While digital tools have the potential to improve the representativeness of trial participants, there is limited evidence supporting their effectiveness in recruiting underrepresented groups [8,17]. This underscores the need for targeted interventions and strategies to bridge the digital divide and ensure equitable access to digital health resources [18,19].

Digital recruitment has become especially popular with the increased interest in artificial intelligence (AI) in health care, which requires large and representative datasets to train. For example, the Bridge2AI-Voice program has the explicit goal to create “an ethically sourced flagship dataset to enable future research in artificial intelligence” [20]. Such datasets are typically assembled through digital recruitment workflows, underscoring the need to understand how these methods shape participant diversity. Our group has similarly endeavored to obtain a large bank of speech recordings focused on neurological disorders, with the goal of subsequently using these data to train AI models, through a primarily digital recruitment approach. Given concerns about the digital divide and its potential impact on representativeness, the broader speech capture study offered the ideal setting to formally investigate these issues in depth. This study aims to characterize recruitment and attrition patterns in a remote neurology cohort, quantify associations with sociodemographic factors (age, sex, neighborhood deprivation, housing-based SES, and urbanicity), and identify drop-off points to inform strategies for digital recruitment for all participants.

Methods

Study Design and Setting

This analysis was nonexperimental and observational, using a longitudinal cohort design to evaluate sociodemographic factors influencing recruitment and attrition in a remote speech capture study. This study was conducted in accordance with the American Psychological Association (APA) Journal Article Reporting Standards (JARS; American Psychological Association, 2018; refer to Multimedia Appendix 1 for the completed JARS-Quant checklist) [21]. The overarching speech capture study aims to remotely collect speech samples from patients with neurologic diseases to develop an easy-to-use and cost-effective screening tool for predicting disease progression. The parent speech capture study extends beyond the March-July 2024 analysis window; for this report, we used data collected within that period. The research was conducted at Mayo Clinic.

Inclusion and Exclusion Criteria

Individuals were eligible if they met the following criteria: (1) adults aged 18 years or older, (2) residing in the United States, and (3) able to communicate in English via spoken language.

Data Collection and Participant Characteristics

Patient identification was conducted using the Mayo Clinic Electronic Health Record (Epic, developed by Epic Systems Corporation), and invitations to complete an eligibility survey were sent via the patient portal using Qualtrics (developed by Qualtrics, LLC) [22]. The eligibility survey also assessed

XSL•FO
RenderX

participants' understanding of the study and asked whether they had a legally authorized representative responsible for financial or health care decisions. Once interest and eligibility were confirmed, participants received a PDF of the consent form to sign electronically via AdobeSign (Adobe Inc) through Mayo Clinic–developed Participant Tracking System (PTrax). PTrax is an institutional research software program designed to streamline informed consent processes, manage participant status, track enrollments and accruals, and provide reporting and analytics. Recruitment followed a convenience sampling approach, targeting patients with upcoming neurology appointments accessible via the institutional patient portal. After consent was obtained, participants were sent a secure message in the patient portal with a link to the speech recording platform and instructions.

We exported longitudinal record data for all eligible patients invited to the speech capture study between March and July 2024. This resulted in a total sample of 5846 participants, reflecting all patients meeting the inclusion criteria during the recruitment window. Race was extracted from the electronic health record, where "White" reflects a race category and is distinct from ethnicity. Ethnicity (eg, Hispanic or Latinx origin) was not included in the dataset for this secondary analysis; therefore, individuals categorized as "White" may include participants of diverse ethnic backgrounds. The invited cohort had a median (IQR) age of 63 (48-72) years, with 56.2% (3283/5846) female and 93.7% (5478/5846) identifying as White. Urbanicity distribution was 56.5% (3303/5846) urban, 23.3% (1363/5846) rural, and 20.2% (1180/5846) urban cluster. This secondary analysis included all eligible patients invited during the recruitment window of the overarching speech capture study. No a priori power calculation was performed because the analysis was observational and descriptive. Precision was conveyed using 95% CIs for medians and IQRs. Data were drawn from Epic (demographics), Qualtrics (survey responses), PTrax (consent tracking), and the speech recording platform (task completion).

Measures and Covariates

SES was assessed using the Housing-based Socioeconomic Status (HOUSES) index and the Area Deprivation Index (ADI) national rank. The HOUSES index is a practical and adaptable tool for assessing SES using housing data. It effectively correlates with traditional SES measures and predicts various health outcomes [23]. Higher HOUSES index values indicate higher SES, while lower values indicate lower SES. The ADI national rank measures neighborhood socioeconomic disadvantage based on factors such as income, education, employment, and housing characteristics [24,25]. Higher scores indicate greater disadvantage. ADI is widely used in public health research to assess the impact of socioeconomic context on health outcomes.

Primary outcomes included study completion and time to enrollment. Exposures and predictors included age, sex, urbanicity (urban, rural, or urban cluster), ADI national rank, and the HOUSES index. Potential confounders considered were device type and urbanicity. Urbanicity was treated as an exposure in primary analyses and as a potential confounder in models examining associations between socioeconomic indices and participation outcomes. The collected data included participants' age, sex assigned at birth (hereafter referred to as sex), date of invitation, recruitment process checkpoints and dates, and devices used for speech recording. Residence information was used to measure the HOUSES and ADI index and to classify participants by urbanicity (urban, rural, or urban cluster).

Quality of Measurements

This analysis relied exclusively on data exported from established operational systems (Epic for demographics, Qualtrics for eligibility responses, PTrax for participant status tracking, and the recording platform for task completion). No new training, instrumentation, or measurement procedures were implemented specifically for this analysis. All timestamps and statuses were generated by the source systems as part of routine workflows. Data were deidentified before analysis.

Analytic Strategy

The longitudinal time series data for each patient undergoing the participation process were standardized to align with specific checkpoints, as demonstrated in Figure 1. At each step, it was possible for patients not to respond to the research coordination team, which was defined as the "No Response" stage. After providing consent, participants were also free to withdraw it at any time. While most individuals followed one of the typical pathways depicted in Figure 1, there were 83 cases (1.4% of the total 5846 cases) in which participants deviated from these paths, necessitating intervention by research coordinators. These atypical cases were primarily attributed to personal circumstances or the involvement of a legally authorized representative.

XSL•FO
RenderX

Figure 1. Simplified view of the participation enrollment pathways in a longitudinal digital neurology research study. PTrax: Participant Tracking System.

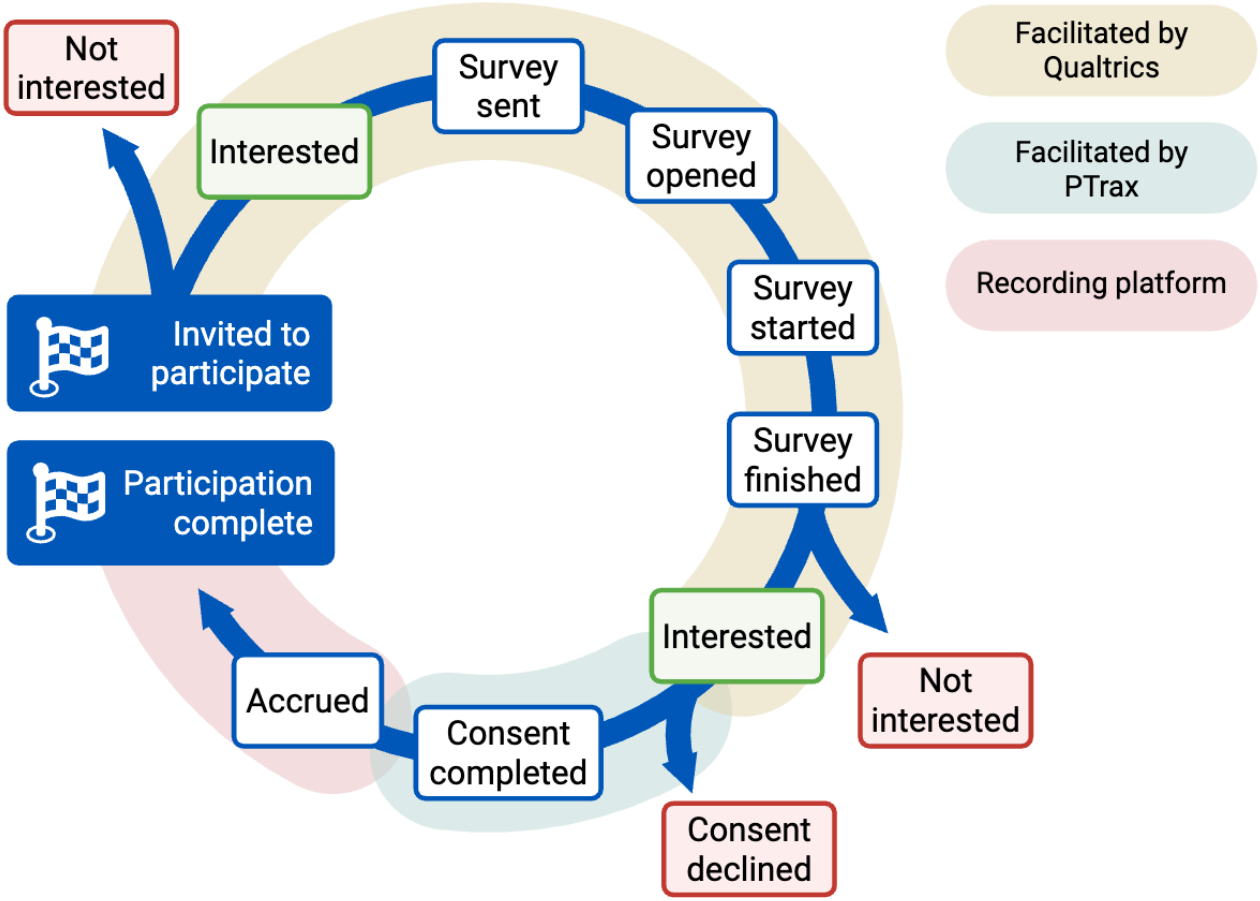

Kruskal-Wallis and Wilcoxon rank-sum tests were used to compare the median age, socioeconomic indices, and time taken to reach different steps of the study. These nonparametric tests were selected due to non-normal distribution of key variables, as confirmed by the Anderson-Darling test ($P<.001$). They are appropriate for comparing medians across groups and handling skewed data. Analysis was conducted at 2 levels: at each checkpoint and through an end-to-end investigation of participants who completed the study. For each path, a participant could take at each checkpoint, age, ADI national rank, and HOUSES index were compared to identify statistically significant differences, using a 2-sided alpha level of .05. At the end-to-end level, additional comparisons were made across sex, urbanicity (urban, rural, and urban cluster), ADI national rank, and HOUSES index to evaluate differences in time taken to complete participation, from initial invitation to accrual, and whether participants completed the study, regardless of the path taken. Analyses were conducted using BlueSky Statistics version 10.3.4 (developed by BlueSky Statistics, LLC) and Python *SciPy* package version 1.16. Exact *P* values are reported, with $P<.001$ used where appropriate. No effect sizes were calculated for analyses.

Data Diagnostics

Since both Kruskal-Wallis and Wilcoxon rank-sum tests require at least 5 data points in each comparison category, pathways with fewer than 5 participants without missing values were excluded from the analysis. This threshold was applied to ensure statistical validity and avoid unreliable comparisons in small subgroups, as no randomized assignments or masking strategies were implemented. To further assess the nature of missing data across key sociodemographic variables, including age, ADI national rank, and HOUSES index, we conducted Little's missing completely at random (MCAR) test [26]. This test evaluates whether data are MCAR, which informs the appropriate handling strategy. Based on the results, we adopted a pairwise deletion approach for statistical analyses, allowing each test to include all available cases for the specific variable of interest. This method was chosen to preserve sample size and maintain statistical power, particularly in subgroups with limited data. This method was chosen to preserve sample size and maintain statistical power, particularly in subgroups with limited data. No outlier removal or variable transformations were applied.

Ethical Considerations

The overarching speech capture study was reviewed and approved by the Mayo Clinic Institutional Review Board (#22-002430). Informed consent was obtained electronically for the primary speech capture study via Adobe Sign. For this secondary analysis of recruitment and attrition patterns, which used deidentified data from the primary study, the institutional review board determined that additional approval was not required. All data were deidentified prior to analysis to ensure participant confidentiality. No compensation was provided for

XSL•FO
RenderX

participation. No identifiable images of participants are included in the manuscript or any supplementary materials.

Results

Overview of the Invited Cohort

A total of 5846 patients were invited to participate in the study between March and July 2024. Of these participants, 3283 (56.2%) were female, 2560 (43.8%) were male, and 3 (0.1%) were unknown. Most participants (5478/5846, 93.7%) identified as White. The age distribution ranged from 18 to 96 years, with a median (IQR) age of 63 (48-72) years (N=5846; 95% CI 62‐63). The narrow CI suggests high precision in estimating the median age of the invited cohort. Regarding urbanicity, 56.5% (3303/5846) of invited participants resided in urban areas, 23.3% (1363/5846) in rural areas, and 20.2% (1180/5846) in urban clusters (Table 1). Following accrual completion, participants used various devices to access the recording platform. Apple-based mobile devices (iPhone and iPad) were most frequently used (141/415, 34.0%), followed by Windows-based computers (134/415, 32.3%), Apple-based computers (82/415, 19.8%), Android-based mobile devices (52/415, 12.5%), and other devices (6/415, 1.4%).

Table . Demographic characteristics of 5846 patients invited to participate in a remote speech capture study for neurological research at Mayo Clinic between March and July 2024.

Characteristic	Participants, n (%)
Sex	
Female	3283 (56.2)
Male	2560 (43.8)
Unknown	3 (0.1)
Race	
White	5478 (93.7)
Black or African American	97 (1.7)
Choose not to disclose	65 (1.1)
Other	206 (3.5)
Age (years)	
18‐33	569 (9.7)
34‐49	999 (17.1)
50‐64	1607 (27.5)
65‐80	2203 (37.7)
≥81	468 (8.0)
Population	
Urban area	3303 (56.5)
Rural area	1363 (23.3)
Urban cluster	1180 (20.2)
Device used for participation	
Apple-based mobile device	141 (34.0)
Windows-based computer	134 (32.3)
Apple-based computer	82 (19.8)
Android-based mobile device	52 (12.5)
Other devices	6 (1.4)

The ADI national rank of participants spanned the entire range from 1 to 100, with a median (IQR) of 44 (28-61; n=5403; 95% CI 43‐45), indicating representation across diverse socioeconomic backgrounds. Similarly, the HOUSES index percentile ranged from 1 to 100, with a median (IQR) of 70 (43-88; n=5439; 95% CI 69‐71), demonstrating considerable variability in housing conditions among participants. The tight CI reflects reliable estimation despite the wide IQR, which indicates heterogeneity in housing-based SES. Neither age, ADI national rank, nor HOUSES index exhibited normal distribution according to the Anderson-Darling test (P<.001 for all variables), confirming the appropriateness of nonparametric statistical methods used in subsequent analyses. Figure 2 illustrates the distribution of the HOUSES index percentile and ADI national rank, with panel A showing ADI national ranks and panel B showing HOUSES index percentiles.

XSL•FO
RenderX

Figure 2. Distribution of socioeconomic measures among participants in a remote speech capture study for neurological research. ADI: Area Deprivation Index. HOUSES: Housing-based Socioeconomic Status.

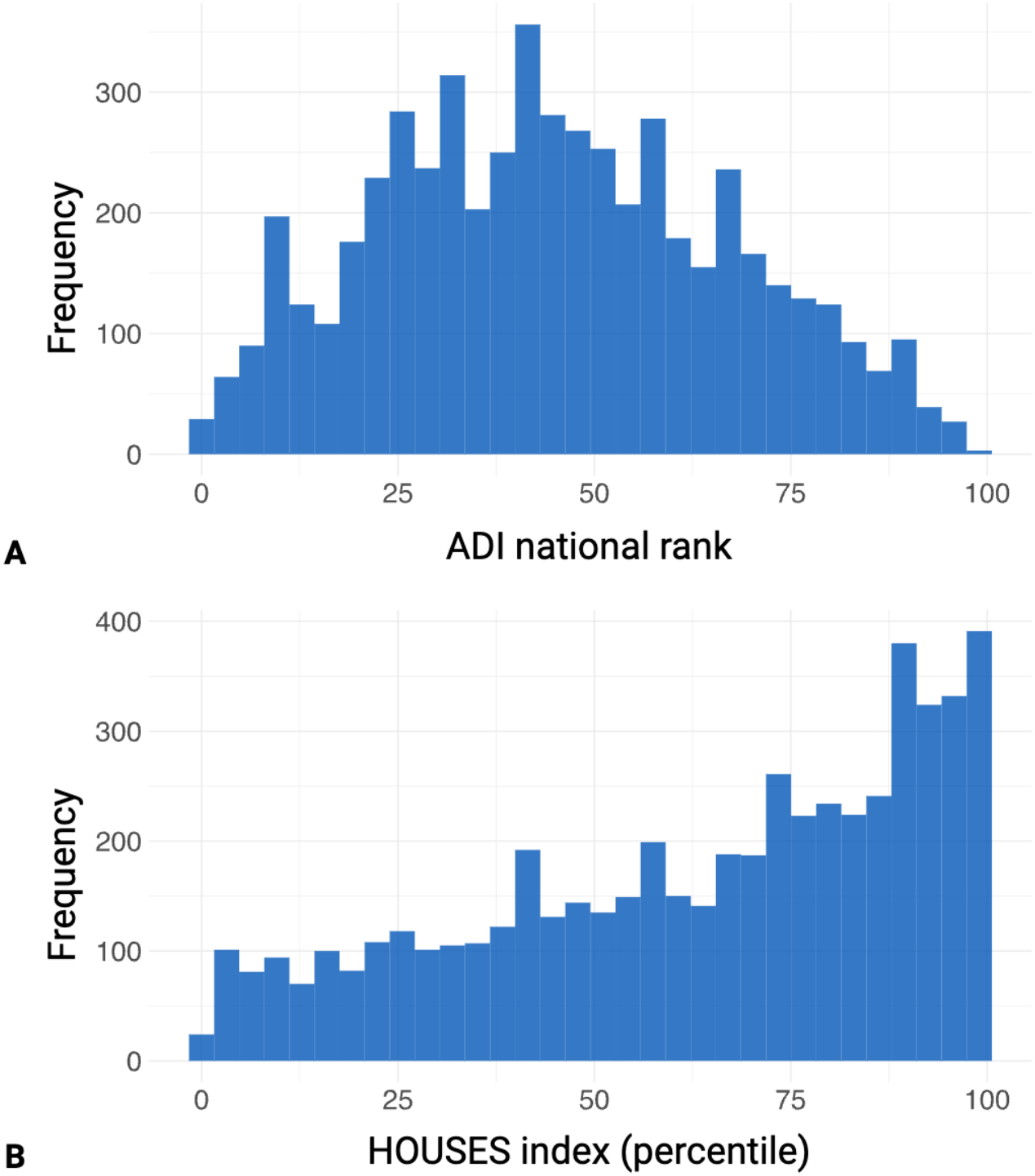

To assess the nature of missing data, Little MCAR test was conducted on age, ADI national rank, and HOUSES index and yielded χ^2_3=3.45; *P*=.24. This nonsignificant result indicates that missingness was not systematically related to other variables, supporting the assumption that data were MCAR. Accordingly, pairwise deletion was applied in subsequent analyses, a strategy chosen to preserve sample size and maintain statistical power while ensuring valid comparisons between study completers and noncompleters.

Comprehensive Participation Analysis

Significant differences were observed between participants who completed the study and those who did not across several demographic and socioeconomic factors. Age was higher among completers (median 66.4, IQR 56.0-72.5; 95% CI 65.1 - 67.6 years) compared to noncompleters (median 62.8, IQR 47.5-72.7; 95% CI 62.2 - 63.2 years; *P*<.001), suggesting that older individuals were more likely to participate (Table 2). The narrow CIs for both groups indicate high precision in estimating age differences, reinforcing the robustness of this finding.

XSL•FO
RenderX

Participants who completed the study also resided in slightly less socioeconomically disadvantaged areas, as indicated by lower ADI national ranks (median 41.0, IQR 27.0-56.0; 95% CI 39.0 - 43.0 vs median 44.5, IQR 28.0-62.0; 95% CI 43.0 - 45.0; *P*=.04). Although the difference is statistically significant, the overlapping CIs suggest that the magnitude of this effect is modest and should be interpreted cautiously. No significant differences were found in HOUSES percentiles, indicating that individual HOUSES may exert less influence on participation compared to neighborhood-level disadvantage.

Table . Comparison of socioeconomic and demographic characteristics between participants who completed the remote speech capture and those who did not. The Wilcoxon rank-sum test was applied to assess differences between groups.

Variable and group	Number of participants	Median (IQR)	95% CI	*P* value
Age (years)				<.001
Participation not complete	5431	62.8 (47.5 - 72.7)	62.2 - 63.2	
Participation complete	415	66.4 (56.0 - 72.5)	65.1 - 67.6	
ADI[a] national rank				.04
Participation not complete	5018	44.5 (28.0 - 62.0)	43.0 - 45.0	
Participation complete	385	41.0 (27.0 - 56.0)	39.0 - 43.0	
HOUSES[b] percentile				.76
Participation not complete	5052	70.0 (42.0 - 88.0)	68.0 - 71.0	
Participation complete	387	71.0 (44.0 - 88.0)	66.0 - 75.0	

[a]ADI: Area Deprivation Index.

[b]HOUSES: Housing-based Socioeconomic Status.

Sex differences in enrollment time were significant, with female participants taking longer to enroll than males (median 38.5, IQR 14.8-66.3; 95% CI 35.0 - 41.0 vs median 32.0, IQR 8.0-57.5; 95% CI 29.0 - 38.0 days; *P*=.01; Table 3). The relatively narrow CIs for both groups indicate precise estimates of enrollment time differences, reinforcing the statistical significance of this finding. Additionally, urbanicity influenced the time to complete enrollment. Participants from urban areas enrolled more quickly than those from rural or urban cluster regions (median 32.0, IQR 9.0-58.0; 95% CI 31.0 - 37.0 vs median 41.0, IQR 22.0-65.0; 95% CI 37.0 - 49.0 days and median 40.0, IQR 13.0-71.0; 95% CI 33.0 - 49.0, respectively; *P*=.01). The wider CIs for rural and urban cluster groups suggest greater variability in enrollment time compared to urban participants, possibly reflecting differences in access or engagement. However, no significant differences were found in completion time across sex or urbanicity. No significant associations were detected between HOUSES indices and study completion. Additionally, no significant relationship was found between the device type used for task completion and the completion or participation time. These null findings suggest that SES and device type did not influence completion dynamics in this cohort.

XSL•FO
RenderX

Table . Enrollment and completion times stratified by urbanicity and sex among participants in a remote speech capture study for neurological research. Statistical significance was assessed using the Wilcoxon rank-sum test or the Kruskal-Wallis test, depending on the category count.

Variable	Number of participants	Median (IQR)	95% CI	*P* value
Enrollment time (days)				.01
Rural	150	41.0 (22.0 - 65.0)	37.0 - 49.0	
Urban area	532	32.0 (9.0 - 58.0)	31.0 - 37.0	
Urban cluster	141	40.0 (13.0 - 71.0)	33.0 - 49.0	
Completion time (days)				.70
Rural	74	20.0 (12.3 - 36.0)	15.5 - 22.0	
Urban area	269	20.0 (10.0 - 32.0)	17.0 - 22.0	
Urban cluster	72	19.5 (7.0 - 38.8)	12.5 - 27.0	
Enrollment time (days)				.01
Male	343	32.0 (8.0 - 57.5)	29.0 - 38.0	
Female	480	38.5 (14.8 - 66.3)	35.0 - 41.0	
Completion time (days)				.95
Male	181	20.0 (10.0 - 32.0)	16.0 - 22.0	
Female	234	20.0 (9.0 - 35.0)	15.5 - 22.0	

Step-by-Step Participation Analysis

The majority of invited participants either did not read or did not respond to the initial invitation via Epic (n=2736) or expressed no interest (n=1752). Among the 1358 participants who initially expressed interest, 415 (30.6%) ultimately completed the study in its entirety. Throughout various stages of the recruitment process, a total of 3346 participants failed to respond to follow-up communications from the research coordination team.

Analysis of participant age across different pathways revealed that individuals who did not respond to the invitation or eligibility check were significantly younger than those who proceeded toward study completion. This age disparity contributed to an increase in the median age of participants completing the study (66.4, IQR 56.0-72.5; 95% CI 65.1 - 67.6) years compared to the overall invited cohort (62.8, IQR 47.5-72.7; 95% CI 62.2 - 63.2) years. The narrow CIs for both estimates indicate high precision, reinforcing confidence in the observed age-related attrition pattern. A similar pattern was observed among the 95 participants who withdrew consent after initially providing it but before being accrued for the recording session, with these participants having a median age of 55.8 years compared to 66.3 years for those who continued with the recording. This suggests that younger participants were disproportionately represented among early dropouts.

While no significant differences in the HOUSES index were observed across different participation pathways, participants who did not respond to the initial invitation had significantly higher ADI national ranks (median 45.0, IQR 29.0-63.0; 95% CI 44.0 - 46.0) compared to those who expressed interest (median 42.0, IQR 27.0-59.0; 95% CI 39.0 - 43.0), indicating residence in more socioeconomically disadvantaged neighborhoods. The narrow CIs for these ADI estimates suggest precise measurement of this disparity, underscoring the influence of neighborhood-level socioeconomic disadvantage on initial engagement. No other significant differences in ADI national ranks were observed across subsequent recruitment steps. Figure 3 details the participant flow and checkpoint-specific distributions.

XSL•FO
RenderX

Figure 3. Step-by-step analysis of participant recruitment and attrition in a longitudinal digital neurology research study conducted at Mayo Clinic between March and July 2024. The figure illustrates participant flow across recruitment checkpoints, with descriptive statistics for Area Deprivation Index (ADI), age, and Housing-based Socioeconomic Status (HOUSES) index reported for each pathway as median (IQR; 95% CI), arranged from top to bottom. Significant differences, based on a significance level of .05, are highlighted in green. Pie charts indicate the relative frequencies of participants at each stage. Paths with fewer than 5 participants are excluded from the analysis. LAR: Legally Authorized Representative.

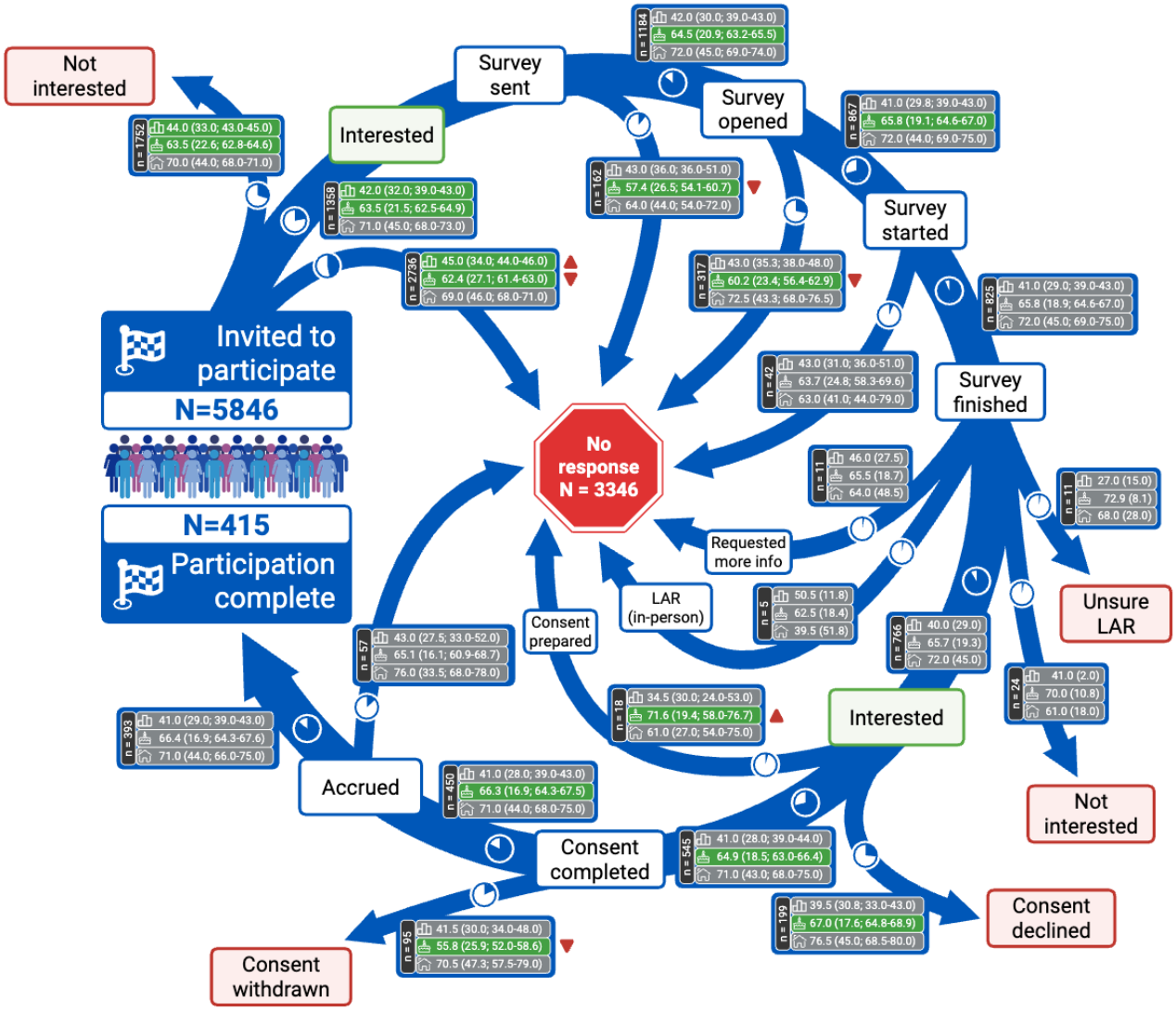

Discussion

Summary of Principal Findings

This study provides a comprehensive analysis of participant recruitment pathways in a digital speech research study, revealing important associations between sociodemographic factors and participation outcomes. Understanding these pathways is critical for improving equity and efficiency in digital clinical research [17]. First, recruitment remains a fundamental challenge in clinical research, with participation rates often below 10 percent in specialized studies, limiting generalizability and slowing innovation [1,2]. Second, as digital recruitment methods become increasingly prevalent, it is essential to assess whether these approaches truly reduce barriers or inadvertently perpetuate existing disparities, an area where evidence remains limited [8]. By analyzing longitudinal recruitment data from 5846 invited patients, our findings demonstrate that systematic pathway analysis can uncover patterns of participation and attrition that may not be apparent when examining only final enrollment outcomes. These results support our primary hypothesis that sociodemographic factors, including age, urbanicity, and neighborhood disadvantage, are associated with recruitment and attrition in digital research workflows. This finding challenges the assumption that digital methods inherently improve inclusivity and highlights the need for targeted strategies, such as age-specific engagement and rural digital support, to promote retention in remote research.

Age-Related Participation Patterns

The significant age differences observed at various dropout points suggest that digital recruitment methods may be less effective for younger populations. Our analysis showed that the median age increased from 63 (IQR 48-72; 95% CI 62 - 63) years in the invited cohort to 66.4 (IQR 56.0-72.5; 95% CI 65.1 - 67.6) years among completers, indicating a robust and precise trend toward older participant retention. This finding challenges the conventional wisdom that digital methods inherently appeal to younger participants and suggests that age-specific engagement strategies may be necessary throughout

XSL•FO
RenderX

the recruitment process. While younger individuals may be more comfortable with technology, our results and others suggest that younger participants are more likely to disengage or drop out of digital studies over time [27,28]. Possible explanations include competing priorities, lower perceived relevance of neurological research, and reduced tolerance for multistep enrollment processes [29]. This emphasizes the importance of age-sensitive retention strategies that extend beyond initial recruitment, including tailored messaging and incentives to support long-term engagement.

Disparities by Urbanicity in Enrollment Timing

The observation that participants from urban areas completed enrollment significantly faster than those from rural areas or urban clusters highlights potential disparities by urbanicity in research accessibility. Urban participants enrolled in a median of 32.0 (IQR 9.0-58.0; 95% CI 31.0 - 37.0) days compared to 41.0 (IQR 22.0-65.0; 95% CI 37.0 - 49.0) days for rural and 40.0 (IQR 13.0-71.0; 95% CI 33.0 - 49.0) days for urban clusters, underscoring a consistent and precise difference. This finding aligns with broader concerns about the urban-rural digital divide and suggests that digital recruitment, while theoretically boundaryless, may still be influenced by infrastructure, digital literacy, or health care engagement [30,31]. Rural participants may face slower internet speeds, limited device availability, and less familiarity with patient portals, which could delay enrollment [32]. This contrasts with some findings in telehealth adoption, which suggest that technology can overcome location-based care barriers [33]. To mitigate these disparities, future strategies should include technical support for rural participants, alternative enrollment options (eg, phone-based consent), and targeted outreach through local health care networks.

Socioeconomic Disadvantage and Participation

Perhaps most notably, our analysis revealed that participants from neighborhoods with higher socioeconomic disadvantage (higher ADI national ranks) were significantly less likely to respond to initial invitations. This may reflect limited broadband access, lower digital literacy, and competing priorities in disadvantaged neighborhoods [34]. This finding is consistent with prior research demonstrating that individuals from more disadvantaged areas experience greater barriers to engaging with digital health, including higher no-show rates and lower uptake of telehealth services [35,36]. Our study adds nuance by comparing neighborhood-level (ADI) and housing-based (HOUSES) SES measures, an approach rarely examined in digital recruitment literature. This finding suggests that digital recruitment methods may perpetuate existing socioeconomic disparities in research participation if not specifically designed to address these barriers. To mitigate these barriers, recruitment strategies should include targeted outreach in areas with high ADI, provide technical support, and incorporate offline options. The absence of significant differences in the HOUSES index across participation pathways, despite differences in the ADI national rank, suggests that neighborhood context likely influences digital engagement more than individual housing characteristics because infrastructure and community resources shape access and literacy.

Sex Differences in Enrollment Dynamics

Females took longer to complete the enrollment process than males, a difference that warrants further investigation. This may reflect differences in time availability, competing responsibilities, or engagement with digital health platforms that could impact recruitment strategies. Prior studies suggest that social and structural factors, such as unequal distribution of caregiving and household responsibilities, affect time availability for research participation, particularly in remote studies [37]. Recruitment workflows could incorporate flexible scheduling and simplified enrollment steps to reduce time burden. While underexplored in the current literature, these insights underscore the need for sex- and gender-aware design in digital recruitment, which may improve inclusivity and reduce attrition.

Recruitment Funnel Attrition

Of 5846 individuals invited, a large proportion did not respond or declined participation, and only 415 of all the invited individuals (7.1%) completed the study. This significant attrition, consistent with patterns observed in other digital recruitment efforts [38,39], may reflect perceived complexity of enrollment, lack of immediate incentives, and competing priorities among participants. These findings highlight the need for iterative, multitouch recruitment strategies that re-engage potential participants and address barriers to active enrollment. Such strategies could include reminder messages, simplified consent processes, and personalized follow-ups to maintain engagement. Our findings reinforce prior evidence that digital recruitment alone is insufficient and highlight the importance of hybrid approaches combining digital and traditional outreach strategies [40].

Limitations

Several limitations should be considered when interpreting these findings. First, this was an observational analysis conducted within a single academic health system, which may limit generalizability to other settings with different patient demographics or digital infrastructure. Second, the study relied on electronic health record–based digital recruitment and patient portal messaging, which presumes access to broadband internet and digital literacy; these factors were not directly measured and may have influenced participation patterns. Findings may generalize to similar academic health systems using portal-based recruitment but may differ in settings with lower portal adoption or different sociodemographic profiles. Third, the cohort was predominantly White (93.7%), limiting the ability to examine racial and ethnic disparities in digital recruitment. Future studies should prioritize inclusion of more diverse populations to assess whether similar sociodemographic patterns persist across racial and ethnic groups. Fourth, while we examined neighborhood-level (ADI) and housing-based (HOUSES) socioeconomic measures, other dimensions of socioeconomic status, such as income, education, and employment, were not available and could provide additional insights. Fifth, attrition analysis was based on recruitment checkpoints rather than qualitative data on participant motivations or barriers, which constrains the interpretation of underlying causes for dropout. Finally, the study period was relatively short (March to July

XSL•FO
RenderX

2024), and findings may not reflect seasonal or long-term trends in digital recruitment dynamics.

To address these challenges, we recommend several strategies to enhance inclusivity and effectiveness in digital recruitment. First, adopting a multichannel approach is essential. Digital methods should be complemented with traditional outreach, particularly for populations with limited online access, such as those in rural areas or socioeconomically disadvantaged communities. Second, age-specific engagement should be prioritized by developing tailored messaging and user experiences for different age groups, with special attention to reducing dropout rates among younger participants. Similarly, geographic barriers must be addressed by providing technical support and offering alternative participation options for both rural and urban clusters. To ensure continuous improvement, it is critical to track recruitment analytics. Monitoring data will help identify dropout points and demographic trends, enabling real-time adjustments to recruitment strategies. Additionally, efforts should be made to minimize participation burden by streamlining enrollment processes to reduce time demands, which is particularly beneficial for individuals with limited availability, such as those with caregiving responsibilities. Socioeconomic factors also require consideration. Recruitment materials should be inclusive, and incentives or support should be offered to offset participation costs, thereby improving socioeconomic accessibility. Furthermore, while device type did not affect completion time in our findings, platforms should still be optimized for device diversity, ensuring mobile-friendly and cross-device accessibility. Ultimately, intentional design and continuous evaluation are key to ensuring that digital methods promote inclusivity rather than hinder it. Future work should explore the mechanisms behind demographic and socioeconomic disparities, test interventions to address these patterns, and evaluate whether similar dynamics occur across other clinical domains.

Conclusion

This study demonstrates that digital recruitment methods in neurological research are subject to demographic, urbanicity, and socioeconomic influences that affect the representativeness of study populations. By mapping these factors, this study provides actionable insights for designing recruitment strategies that improve participation in remote neurological research and similar digital health initiatives. Our findings suggest that although digital recruitment expands reach, it does not eliminate traditional barriers and introduces new challenges. The digital divide appears to manifest in nuanced ways throughout the recruitment process, potentially influencing who participates in neurological research and, consequently, who benefits from its findings. The relatively low overall completion rate (7.1% of invited participants) underscores the persistent challenge of recruitment in specialized medical research, even with digital methods. Collectively, these findings reinforce the importance of refining digital recruitment strategies to bridge the persistent digital divide and promote research participation.

Acknowledgments

Some figures were created with BioRender. The authors declare the use of generative artificial intelligence (AI) in the research and writing process. According to the GAIDeT taxonomy (2025), tasks delegated to generative AI tools under full human supervision included literature search and systematization, code generation and optimization, proofreading and editing, text summarization, and reformatting. The generative AI tool used was GPT-5 within Microsoft Copilot. Responsibility for the final manuscript lies entirely with the authors; generative AI tools are not listed as authors and do not bear responsibility for the final outcomes. This declaration is submitted by PN.

Funding

This project was supported by Grant Number UL1TR002377 from the National Center for Advancing Translational Sciences (NCATS) and Grant Number R01AG083832 from the National Institute on Aging (NIA). The funding agencies had no role in the study design, data collection, analysis, interpretation of results, or writing of the manuscript. The contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to the sensitive nature of patient information and restrictions imposed by the Mayo Clinic Institutional Review Board. Deidentified data may be made available by the corresponding author upon reasonable request and contingent on appropriate institutional approvals.

Authors' Contributions

Conceptualization: HB (lead), PN (supporting)
Project administration: ADB, VMS
Data curation: ADB, VMS
Methodology: HB
Formal analysis: PN
Writing - original draft: PN
Writing - review & editing: JRD, JLS, VH, DTJ, RLU, HB

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
American Psychological Association JARS-Quant guidelines, completed checklist.
[PDF File, 193 KB - jmir_v28i1e83432_app1.pdf]

References

1. Unger JM, Fleury M. Nationally representative estimates of the participation of cancer patients in clinical research studies according to the Commission on Cancer. J Clin Oncol 2021 Oct 1;39(28):74. [doi: 10.1200/JCO.2020.39.28_suppl.74]
2. Unger JM, Shulman LN, Facktor MA, Nelson H, Fleury ME. National estimates of the participation of patients with cancer in clinical research studies based on Commission on Cancer accreditation data. J Clin Oncol 2024 Jun 20;42(18):2139-2148. [doi: 10.1200/JCO.23.01030] [Medline: 38564681]
3. Glickman SW, Anstrom KJ, Lin L, et al. Challenges in enrollment of minority, pediatric, and geriatric patients in emergency and acute care clinical research. Ann Emerg Med 2008 Jun;51(6):775-780. [doi: 10.1016/j.annemergmed.2007.11.002] [Medline: 18191297]
4. Khazak A, Parsa K, Elias E, Liaw BCH. Geriatric enrollment patterns in clinical trials at a National Cancer Institute. J Clin Oncol 2024 Jun 1;42(16_suppl):e13762-e13762. [doi: 10.1200/JCO.2024.42.16_suppl.e13762]
5. Ashery RS, McAuliffe WE. Implementation issues and techniques in randomized trials of outpatient psychosocial treatments for drug abusers: recruitment of subjects. Am J Drug Alcohol Abuse 1992;18(3):305-329. [doi: 10.3109/00952999209026069] [Medline: 1329493]
6. Bikou AG, Deligianni E, Dermiki-Gkana F, et al. Improving participant recruitment in clinical trials: comparative analysis of innovative digital platforms. J Med Internet Res 2024 Dec 18;26:e60504. [doi: 10.2196/60504] [Medline: 39693619]
7. Jackson J. Expanding opportunities to maximise research recruitment and data collection using digital tools. Int J Soc Res Methodol 2023 Nov 2;26(6):721-731. [doi: 10.1080/13645579.2022.2091258]
8. Kasahara A, Mitchell J, Yang J, Cuomo RE, McMann TJ, Mackey TK. Digital technologies used in clinical trial recruitment and enrollment including application to trial diversity and inclusion: a systematic review. Digit Health 2024;10:20552076241242390. [doi: 10.1177/20552076241242390] [Medline: 38559578]
9. Whitaker C, Stevelink S, Fear N. The use of Facebook in recruiting participants for health research purposes: a systematic review. J Med Internet Res 2017 Aug 28;19(8):e290. [doi: 10.2196/jmir.7071] [Medline: 28851679]
10. Schwinne M, Woods E, Chan B, et al. Transforming research recruitment: leveraging EHR systems and patient portals. J Clin Transl Sci 2025;9(1):e10. [doi: 10.1017/cts.2024.692] [Medline: 39911926]
11. Iflaifel M, Hall CL, Green HR, et al. Strategies to improve recruitment in mental health clinical trials: a scoping review (RE-MIND study). Trials 2024 Dec 18;25(1):832. [doi: 10.1186/s13063-024-08665-x] [Medline: 39695709]
12. Biga R, Nottebaum S, Kozlakidis Z, Psomiadis S. Digitalization of healthcare in lmics: digital health and the digital divide based on technological availability and development. In: Kozlakidis Z, Muradyan A, Sargsyan K, editors. Digitalization of Medicine in Low- and Middle-Income Countries: Paradigm Changes in Healthcare and Biomedical Research: Springer International Publishing; 2024:185-193. [doi: 10.1007/978-3-031-62332-5_18]
13. König LM, Krukowski RA, Kuntsche E, et al. Reducing intervention- and research-induced inequalities to tackle the digital divide in health promotion. Int J Equity Health 2023 Dec 4;22(1):249. [doi: 10.1186/s12939-023-02055-6] [Medline: 38049789]
14. Litchfield I, Shukla D, Greenfield S. Impact of COVID-19 on the digital divide: a rapid review. BMJ Open 2021 Oct 12;11(10):e053440. [doi: 10.1136/bmjopen-2021-053440] [Medline: 34642200]
15. Livingood WC, Bautista MAB, Smotherman C, et al. Comparative study of different SES neighborhood clinics for health literacy and internet access. Digit Health 2022;8:20552076221123715. [doi: 10.1177/20552076221123715] [Medline: 36081750]
16. Walker DM, Hefner JL, Fareed N, Huerta TR, McAlearney AS. Exploring the digital divide: age and race disparities in use of an inpatient portal. Telemed J E Health 2020 May;26(5):603-613. [doi: 10.1089/tmj.2019.0065] [Medline: 31313977]
17. Tomiwa T, Wong E, Miller HN, et al. Leveraging digital tools to enhance diversity and inclusion in clinical trial recruitment. Front Public Health 2024;12:1483367. [doi: 10.3389/fpubh.2024.1483367] [Medline: 39529717]
18. Bradway M, Årsand E. Diverse recruitment strategies are needed to reduce digital divide: results from a workshop addressing digital divide and effects of pandemic restrictions. Stud Health Technol Inform 2022 May 25;294:811-812. [doi: 10.3233/SHTI220592] [Medline: 35612212]
19. Friis-Healy EA, Nagy GA, Kollins SH. It is time to REACT: opportunities for digital mental health apps to reduce mental health disparities in racially and ethnically minoritized groups. JMIR Ment Health 2020;8(1):e25456. [doi: 10.2196/25456]
20. Bensoussan Y, Sigaras A, Rameau A, Elemento O, Powell M, Dorr D, et al. Bridge2AI-Voice: An Ethically-Sourced, Diverse Voice Dataset Linked to Health Information (Version 200): PhysioNet; 2025. [doi: 10.13026/3xt6-rf05]

XSL•FO
RenderX

21. APA Style Journal Article Reporting Standards (JARS). American Psychological Association. 2018. URL: https://apastyle.apa.org/jars [accessed 2026-01-29]
22. Qualtrics AI + XM platform_. Qualtrics. 2024. URL: https://www.qualtrics.com/en-au/ [accessed 2026-01-29]
23. Juhn YJ, Beebe TJ, Finnie DM, et al. Development and initial testing of a new socioeconomic status measure based on housing data. J Urban Health 2011 Oct;88(5):933-944. [doi: 10.1007/s11524-011-9572-7] [Medline: 21499815]
24. About the Neighbothood Atlas® and Area Deprivation Index. Center for Health Disparities Research University of Wisconsin School of Medicine and Public Health. 2024. URL: https://www.neighborhoodatlas.medicine.wisc.edu [accessed 2026-01-29]
25. Kind AJH, Buckingham WR. Making neighborhood-disadvantage metrics accessible - the Neighborhood Atlas. N Engl J Med 2018 Jun 28;378(26):2456-2458. [doi: 10.1056/NEJMp1802313] [Medline: 29949490]
26. Little RJA. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 1988 Dec;83(404):1198-1202. [doi: 10.1080/01621459.1988.10478722]
27. Cho PJ, Olaye IM, Shandhi MMH, Daza EJ, Foschini L, Dunn JP. Identification of key factors related to digital health observational study adherence and retention by data-driven approaches: an exploratory secondary analysis of two prospective longitudinal studies. Lancet Digit Health 2025 Jan;7(1):e23-e34. [doi: 10.1016/S2589-7500(24)00219-X] [Medline: 39722250]
28. Pratap A, Neto EC, Snyder P, et al. Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. NPJ Digit Med 2020;3(1):21. [doi: 10.1038/s41746-020-0224-8] [Medline: 32128451]
29. Gutiérrez Á, Cain R, Aranda MP. The digital divide exacerbates disparities in Latinx recruitment for Alzheimer's disease and related dementias online education during COVID-19. Gerontol Geriatr Med 2022;8:23337214221081372. [doi: 10.1177/23337214221081372] [Medline: 35252476]
30. Jongebloed H, Anderson K, Winter N, et al. The digital divide in rural and regional communities: a survey on the use of digital health technology and implications for supporting technology use. BMC Res Notes 2024 Mar 28;17(1):90. [doi: 10.1186/s13104-024-06687-x] [Medline: 38549176]
31. Kelpin S, Brockman TA, Decker PA, et al. Increasing digital equity to promote online smoking cessation program engagement among rural adults: a randomized controlled pilot trial. Commun Med (Lond) 2024 Oct 8;4(1):194. [doi: 10.1038/s43856-024-00624-6] [Medline: 39375517]
32. Quinton JK, Ong MK, Vangala S, et al. The association of broadband internet access and telemedicine utilization in rural western Tennessee: an observational study. BMC Health Serv Res 2021 Aug 3;21(1):765. [doi: 10.1186/s12913-021-06746-0] [Medline: 34344377]
33. Dorsey ER, Glidden AM, Holloway MR, Birbeck GL, Schwamm LH. Teleneurology and mobile technologies: the future of neurological care. Nat Rev Neurol 2018 May;14(5):285-297. [doi: 10.1038/nrneurol.2018.31] [Medline: 29623949]
34. Price-Haywood EG, Arnold C, Harden-Barrios J, Davis T. Stop the divide: facilitators and barriers to uptake of digital health interventions among socially disadvantaged populations. Ochsner J 2023;23(1):34-42. [doi: 10.31486/toj.22.0101] [Medline: 36936477]
35. Brown SH, Griffith ML, Kripalani S, Horst SN. Association of health literacy and area deprivation with initiation and completion of telehealth visits in adult medicine clinics across a large health care system. JAMA Netw Open 2022 Jul 1;5(7):e2223571. [doi: 10.1001/jamanetworkopen.2022.23571] [Medline: 35877125]
36. Ettman CK, Ringlein GV, Dohlman P, et al. Trends in mental health care and telehealth use across area deprivation: an analysis of electronic health records from 2016 to 2024. PNAS Nexus 2025 Feb;4(2):pgaf016. [doi: 10.1093/pnasnexus/pgaf016] [Medline: 39959710]
37. Jolliff A, Holden RJ, Valdez R, et al. Investigating the best practices for engagement in remote participatory design: mixed methods analysis of 4 remote studies with family caregivers. J Med Internet Res 2024 Dec 3;26:e60353. [doi: 10.2196/60353] [Medline: 39626228]
38. Brøgger-Mikkelsen M, Ali Z, Zibert JR, Andersen AD, Thomsen SF. Online patient recruitment in clinical trials: systematic review and meta-analysis. J Med Internet Res 2020 Nov 4;22(11):e22179. [doi: 10.2196/22179] [Medline: 33146627]
39. Pagliaro JA, Wash LK, Ly K, et al. Enrollment and retention outcomes from the Veterans Health Administration for a remote digital health study: multisite observational study. JMIR Form Res 2025 Aug 15;9:e68676. [doi: 10.2196/68676] [Medline: 40334212]
40. Frampton GK, Shepherd J, Pickett K, Griffiths G, Wyatt JC. Digital tools for the recruitment and retention of participants in randomised controlled trials: a systematic map. Trials 2020 Jun 5;21(1):478. [doi: 10.1186/s13063-020-04358-3] [Medline: 32498690]

Abbreviations

ADI: Area Deprivation Index
AI: artificial intelligence
APA: American Psychological Association
HOUSES: Housing-based Socioeconomic Status
JARS: Journal Article Reporting Standards

XSL•FO
RenderX

MCAR: missing completely at random
PTrax: Participant Tracking System
SES: socioeconomic status

Edited by S Brini; submitted 02.Sep.2025; peer-reviewed by B Agyare-Aggrey, C Ahmadu; revised version received 07.Jan.2026; accepted 09.Jan.2026; published 25.Feb.2026.

Please cite as:
Nejat P, Bachman AD, Stubbs VM, Duffy JR, Stricker JL, Herasevich V, Jones DT, Utianski RL, Botha H
Sociodemographic Drivers of Recruitment and Attrition in Digital Neurological Research: Longitudinal Cohort Study
J Med Internet Res 2026;28:e83432
URL: https://www.jmir.org/2026/1/e83432
doi:10.2196/83432

© Peyman Nejat, Ashley D Bachman, Vicki M Stubbs, Joseph R Duffy, John L Stricker, Vitaly Herasevich, David T Jones, Rene L Utianski, Hugo Botha. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Using Epidemiological Test Diagnostics to Select Fraud Detection Methods: Secondary Analysis of Quantitative Cross-Sectional Survey Data

Rachel Willard-Grace[1], MPH; Tali Klima[2], PhD; Mansi Dedhia[1], MPH, MS; Emily Lo[1], BA, BS; Annie Nisnevich[1], BS; Allison Gray[3], MA; Holly Henry[3], MHS, PhD

[1]Department of Family and Community Medicine, University of California, San Francisco, Box 1315 2540 23rd St., Floor 5, 5511, San Francisco, CA, United States

[2]Practical Research Solutions, Sunnyvale, CA, United States

[3]Lucile Packard Foundation for Children's Health, Palo Alto, CA, United States

Corresponding Author:
Rachel Willard-Grace, MPH
Department of Family and Community Medicine, University of California, San Francisco, Box 1315 2540 23rd St., Floor 5, 5511, San Francisco, CA, United States

Abstract

Background: Survey research has the potential to elevate the experiences and opinions of marginalized populations. The rising number of bot attacks, a method of participant fraud that creates multiple records in survey data using automated software, threatens to drown out those voices and produce inaccurate findings. Rapid identification and mitigation of bot attacks are vital; however, there is limited guidance for researchers on scalable approaches to address this problem.

Objective: This study aimed to assess how well recommended methods detect fraud using an epidemiological diagnostic test framework to inform web-based survey researchers on how best to identify and shut down bot attacks.

Methods: We analyzed data from a cross-sectional web-based statewide survey on access to pediatric subspecialty care in California that used Qualtrics survey software. Caregivers of children with chronic conditions were recruited through family resource centers (FRCs), nonprofit agencies serving families with developmental delays and chronic medical conditions. The survey was sent out to 17 FRCs, whose staff distributed anonymous links to their clients through listservs and flyers. Respondents who completed the survey received a US $30 gift card. Prior to launch, we designed a protocol to identify and respond to bot attacks and reviewed responses for markers of fraudulent activity. If markers were identified or there was a spike in responses, a senior member of our research team reviewed patterns among all submitted surveys for each FRC to look for signs of bot attacks. We calculated epidemiologic measures of diagnostic test accuracy, such as sensitivity, specificity, positive predictive value, and negative predictive value, which describe a test's ability to distinguish "disease" (in this case, fraudulent records) from normal cases, to better understand the utility of recommended strategies to identify bot attacks.

Results: We received 646 valid survey records and 905 fraudulent records resulting from bot attacks. The primary indicator of a bot attack was a sudden spike in responses to the survey. Differences in demographics and outcomes, including wait times for pediatric subspecialty care and use of health care services, between the valid and fraudulent data indicated that failure to remove fraudulent records would have substantially altered the survey results. Most recommended methods in the literature for identifying fraudulent responses had low sensitivity to detect bot attacks, and only 2 were better than chance alone at correctly identifying bot attacks. Combinations of fraud markers and blocks of repeated responses were particularly useful to identify bot attacks.

Conclusions: Fraudulent data entry using bots is increasing in survey research. Sharing flexible protocols to identify and mitigate them in a way that is responsive to their ever-changing nature is vital to ensuring that researchers elevate the voices of real people within survey research to inform policy and programmatic discussions.

(J Med Internet Res 2026;28:e85161) doi:10.2196/85161

KEYWORDS

health care surveys; bots; surveys; questionnaires; survey methodology; web-based data collection; specificity; sensitivity

XSL•FO
RenderX

Introduction

Survey research has the potential to elevate the experiences and opinions of individuals who may be difficult to reach. While ideal survey methods include the use of individualized, single-use survey links shared with a defined and verified set of respondents, in reality, there are important populations who are not neatly vetted and included in registries and, in some cases, whose safety depends on being invisible and unregistered [1,2]. Since creating single-use survey links matched to known populations is challenging in these situations, many surveys use generic survey links, which carry a greater risk of use by fraudulent responders for profit or manipulative purposes.

The rising number of fraudulent responses threatens to drown out those voices. Our nonexhaustive literature review of 13 studies exploring fraudulent responses found a range of 13% to 97% of survey responses were identified as fraudulent, with a pooled proportion of 61%. Eight of 13 studies reported that more than half of responses were fraudulent (Multimedia Appendix 1) [1-10]. The onslaught of fraudulent responses has a number of consequences; key among these is the distortion of study findings, leading to inaccurate results if fraudulent records are not excluded [4]. For the purpose of this paper, we define a "fraudulent record" as any record in which false information is presumed to be purposely provided. Bot attacks comprise a subset of these fraudulent records and are distinguished by the use of software to generate a large number of responses [11].

Rapid identification and mitigation of bot attacks are vital; however, there is limited guidance for survey researchers in the face of rapidly evolving and ever-more-sophisticated fraudulent attacks. We did not find any prior studies that used standard epidemiological methods to assess the effectiveness of various bot detection approaches. Standard recommendations include identifying duplicate responses or off-topic responses, inconsistencies between location stamp and location reported, similar or identical response or pattern of responses is provided across surveys, identifying inconsistencies between different survey items (eg, age and year of birth), embedding "honeypot" items that are visible to bots but not humans, or short survey completion time [10]. Most of these recommendations are based on assumptions of manual fraud, where individuals submit multiple surveys by hand (formerly known as "ballot box stuffing"), or relatively unintelligent technology that uses simple strategies and does not learn how to game the survey (eg, overcome screening items).

Based on our experience and anecdotal evidence from others, researchers have already observed failures of many of these methods, which are less applicable to modern bot capabilities. Moreover, among survey researchers seeking to reach populations with limited digital literacy, some suggested methods pose the additional danger of excluding eligible but less digitally savvy participants [6]. For example, we abandoned use of CAPTCHA screening (identification of images, such as motorcycles, in a matrix) on our survey platform after technologically savvy staff were consistently unable to pass the screening themselves.

Little is known about the ability of current fraud detection methods to successfully identify fraudulent records (sensitivity, in clinical epidemiological terms) or their ability to correctly identify valid records (ie, not to label a valid record as a fraudulent one, hereafter referred to as "specificity"). Using data from a California-wide survey and a diagnostic approach that treats bot attacks as a disease to be detected by various possible tests, we describe the ability of available tools to correctly identify "diseased cases" (positive predictive value) and the risk that they will miss such cases (negative predictive value). Insights from this analysis may provide guidance on how best to identify and shut down large-scale bot attacks on surveys.

Methods

We conducted a secondary analysis of data from a survey of caregivers of children with chronic conditions in California. Results of the study have been previously reported [12-15].

Population and Setting

We recruited caregivers of children with chronic conditions who were clients of 17 family resource centers (FRCs) across California. FRCs are nonprofit community agencies serving families of children with developmental delays and chronic medical conditions through education and advocacy. For example, they frequently provide classes and peer support to help caregivers of children with chronic conditions to secure social and educational services, and to connect with other families of children with similar conditions. Due to restricted budgets, they often have fewer organizational resources, including paid staff, and rely on volunteer caregivers to advise and provide services to families. They frequently communicate with clients through email listservs. Eligible participants were English- or Spanish-speaking adult caregivers residing in California whose child had a visit with a new pediatric subspecialist in the previous 12 months.

Distribution of the Survey

The study team initially proposed sharing a list of unique, single-use survey links with FRCs to be distributed to their clients, but FRCs were unable to use these unique links with their listserv software and were restricted from sharing their listservs with outside organizations, like the study team, which could have sent unique links directly to FRC clients. As a result, we generated a unique generic link for each FRC. FRCs distributed an invitation to the survey from the study team and the survey link through their email listservs of clients and printed flyers with a QR code link to the survey to post in their offices for access by families coming in for services. Data was collected between August and October 2023.

Fraud Detection Protocol

Based on the REAL (Reflect, Expect, Analyze, Label) framework for online survey fraud [16], we considered in advance how to prevent or identify fraudulent responses at each stage of the process, including pre–data collection (eg, recruitment through community-based partners), data collection (eg, integration of fraud detection questions in the survey design), and post–data collection (eg, a systematic review

XSL•FO
RenderX

process). We also borrowed ideas from Anomaly Detection Theory, grounded in engineering sciences, which posits that detection of anomalies requires identifying data that do not follow established patterns [17].

We anticipated that the distribution of the survey through an anonymous link to large groups of people could make the survey vulnerable to fraudulent responses, and in particular, to bot attacks in which an individual writes code to automatically submit multiple survey responses that appear to originate from eligible respondents. Based on sparse literature and consultation with colleagues, the team utilized 10 different fraud detection tools when the survey was fielded, including proprietary tools that can be turned on within the Qualtrics survey platform, such as invisible ReCAPTCHA technology, and other flags, like the Relevant ID Duplicate, Fraud scores, and Ballot Box Stuffing flags. In addition, we embedded questions that required internally consistent logic (eg, birth year and age), monitored open-ended questions for nonsensical or repeating responses, reviewed longitude and latitude to ensure respondent location in California, considered records based on completion times (including shorter-than-expected response time and duplicate submission times), and verified IP addresses of respondents through an external server of suspicious addresses (see Multimedia Appendix 2). The survey also included a series of 13 screening questions designed to remove ineligible participants before reaching the body of the survey, which we believed would discourage fraudulent responses.

The number of responses was monitored multiple times each day, and spikes in responses were immediately investigated. We completed a logbook entry for responses that triggered concern. Most indicators of fraudulent activity have caveats; for example, IP addresses triggering warnings from the fraudulent database could result from a previous infection of a legitimate device by malware that does not affect the survey (see Multimedia Appendix 2 for additional examples). As a result, no single flag was treated as a sole determinant of fraud. Classification of all records with flags was conducted through consensus between a senior researcher and a research associate. Fraudulent responses were often confirmed through patterns of responses, such as blocks of questions with the same responses (see Figure 1 for example). In the minority of cases in which the coding team was uncertain about a determination, the research associate followed up with the participant, requesting additional information that they could check against the survey to confirm that the response was valid.

Figure 1. Examples of fraudulent response patterns (repetition of blocks across multiple records) in a cross-sectional survey of caregivers of children with new pediatric subspecialty referrals (August-October 2023).

scr_age	how_long_ago	wait_time	travel_time	travel_diff	travel_diff	travel_diff	travel_diff	t
11-18 years old	Within last 6 months	4 weeks (1 month)						
0-5 years old	Within last 6 months	8 weeks (2 months)	2 hours - 2 hours, 59 minutes			covered		
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			

XSL•FO
RenderX

Data Analysis

We cleaned datasets from each FRC and reviewed records for final disposition (eg, valid, ineligible based on various criteria, fraudulent). We examined differences in demographic characteristics and health care utilization (outcomes in the primary study) between valid and fraudulent records in order to ascertain how our results might differ if fraudulent data were not identified and removed.

To assess the validity of different fraud detection methods, we calculated the sensitivity, specificity, and positive and negative predictive value for each method (see Figure 2). Sensitivity is the ability of a test or method to correctly identify individuals with a disease, in this case, fraudulent records, while specificity is the ability of a test to correctly identify individuals without a disease, here, valid records [18]. Epidemiologists do not define universal thresholds of acceptable sensitivity and specificity, taking into consideration the danger of having false-positive or negative findings based on the severity and prevalence of the condition, for example. However, one rule of thumb is that the sum of the two values should be between 150% and 200%, meaning that it is at least more likely than chance alone to result in a correct identification [19]. We accordingly calculated the sum of sensitivity and specificity.

Figure 2. Epidemiological measures of diagnostic test accuracy.

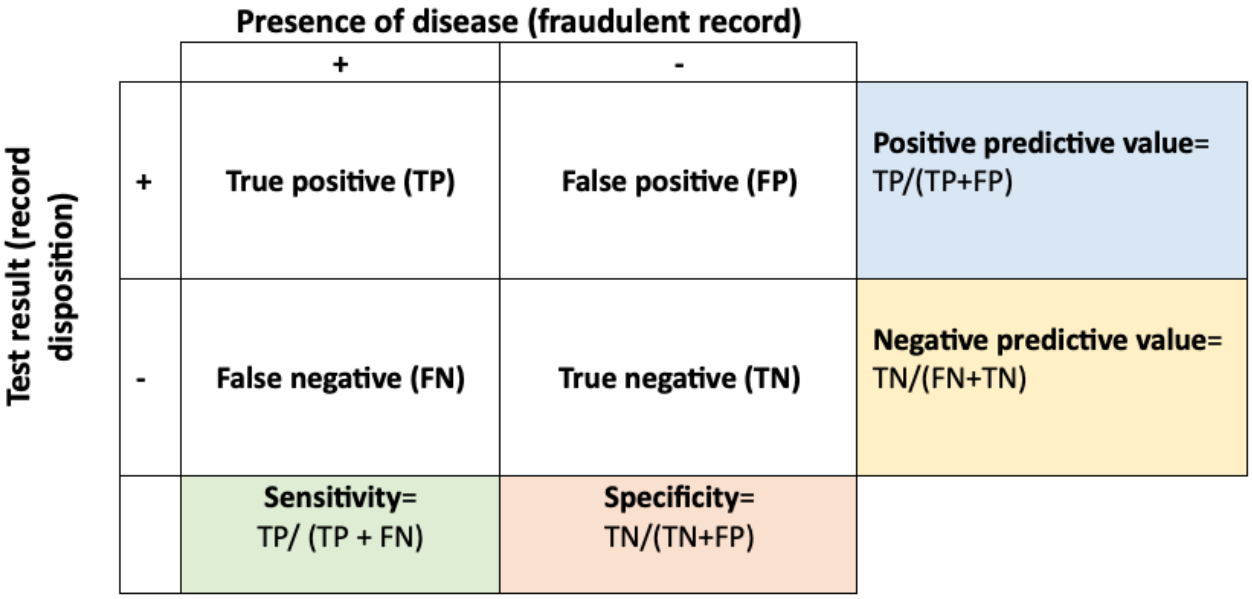

Positive predictive value is the proportion of individuals testing positive (given a fraudulent study disposition) who actually have the disease (are actually fraudulent). Negative predictive value is the proportion of individuals testing negative (given a valid disposition) who truly do not have the disease (are, in fact, valid records). Both positive and negative predictive values depend not only on the sensitivity and specificity of the tests, but also on the prevalence of the condition in the population (in this case, the prevalence of bot attacks in the pool of survey records).

Ethical Considerations

The protocol for the original study was approved by the University of California, San Francisco (UCSF) Institutional Review Board (23‑39316). The board approved implied consent for the study, whereby participants clicking on the survey link first viewed an information page approved by the board and then opted into participation by clicking on a link to take part in the survey. Data was deidentified prior to analysis. Respondents received a US $30 electronic gift card for completing the survey, once the validity of the survey was confirmed by the study team (see Multimedia Appendix 3).

Results

Of 2220 records, 646 (29%) were valid and complete (hereafter "valid"), 905 (41%) were classified as fraudulent by the protocol, 263 (12%) were valid but ineligible based on the eligibility screeners, 308 (14%) were incomplete surveys, and 98 (4%) were duplicate records. Duplicate records were distinguished from fraudulent records as those that made no effort to disguise their identity or device as identical; these appeared to result from respondents being unsure if their previous response had been received. Subsequent analyses were limited to valid and fraudulent records (n=1551).

There were a number of statistically significant differences in the reported demographics of valid versus fraudulent records (Table 1). When valid responses were compared with the overall responses (for valid + fraudulent records), valid responses were more likely to report having younger children (317/646, 49.1% aged 0‑5 y vs 536/1472, 36.4% for all responses); being Asian American and Pacific Islander (106/639, 16.4% vs 125/1368, 8.1%) or Latino/a (352/639, 55.1% vs 382/1368, 27.9%); living in a big city (218/635, 34.3% vs 697/1343, 51.9%); and having commercial health insurance (226/640, 35.3% vs 344/1353, 25.4%).

XSL•FO
RenderX

Table . Demographic characteristics of all, valid, and fraudulent respondents to a cross-sectional survey for caregivers of children with new pediatric subspecialty referrals in California (August-October 2023).

Characteristic	All (valid + fraudulent records), n/N (%)	Valid records, n/N (%)	Fraudulent records, n/N (%)	Difference between valid and fraudulent records	
				Pearson chi-square (*df*)	*P* value
Age (y)				104.12 (1)	<.001
0-5	536/1472 (36.4)	317/646 (49.1)	219/825 (26.5)		
6-10	527/1472 (35.8)	148/646 (22.9)	379/825 (45.9)		
11-20	408/1472 (27.7)	181/646 (28.0)	227/825 (27.7)		
Child's gender[a]				1.41 (1)	.24
Boy	829/1364 (60.8)	398/640 (62.2)	431/724 (59.5)		
Girl	530/1364 (38.9)	237/640 (37)	293/724 (40.5)		
Another gender	5/1364 (0.4)	5/640 (0.8)	0/724 (0.0)		
Race/ethnicity[b]					
Black/African American	190/1368 (13.9)	38/639 (5.9)	152/728 (20.9)	63.4 (1)	<.001
American Indian/Alaskan Native	34/1368 (2.5)	31/639 (4.8)	3/728 (0.3)	27.7 (1)	<.001
Asian/Pacific Islander	125/1368 (8.1)	106/639 (16.4)	19/728 (2.1)	104.2 (1)	<.001
Hispanic/Latino	382/1368 (27.9)	352/639 (55.1)	30/728 (4.1)	439 (1)	<.001
White	799/1368 (58.5)	258/639 (40.4)	541/728 (74.3)	161.4 (1)	<.001
Other	20/1368 (1.5)	18/639 (2.8)	2/728 (0.3)	—[c]	—
Urbanicity				149 (2)	<.001
Big city	697/1343 (51.9)	218/635 (34.3)	479/708 (67.7)		
Small city or suburb	557/1343 (41.5)	358/635 (56.4)	199/708 (28.1)		
Rural area	89/1343 (6.6)	59/635 (9.3)	30/708 (4.2)		
Insurance[b]					
Medicaid	986/1353 (72.9)	488/640 (76.3)	494/713 (69.9)	7 (1)	.008
Commercial	344/1353 (25.4)	226/640 (35.3)	118/713 (16.6)	62.6 (1)	<.001

[a]Reflects caregiver-reported gender. Because of small cell sizes, "another gender" was excluded from testing of differences across groups.

[b]May report more than 1 response, so responses may sum to more than 100%. For multiple response comparisons, used Pearson X^2 analysis to compare responses for each option across valid versus fraudulent responses.

[c]Not applicable.

With respect to health care utilization, which was the key outcome in the primary study, there were significant differences between valid versus fraudulent responses. Valid responses generally reported lower utilization than overall responses, such as having at least one hospitalization in the last year (172/639, 26.9% for valid vs 578/1352, 42.8% of overall responses; Table 2), at least one emergency department visit (324/638, 50.8% vs 799/1351, 59.1%), and at least 1 hour per week in medical visits (469/636, 73.7% vs 1139/1347, 84.6%). On the other hand, valid respondents were more likely to report 3-month waits or more for pediatric subspecialty care compared to the overall sample (156/638, 24.5% vs 271/1382, 19.6%).

XSL•FO
RenderX

Table . Outcomes for health care utilization of all, valid, and fraudulent respondents to a cross-sectional survey for caregivers of children with new pediatric subspecialty referrals in California (August-October 2023).

Outcome	All (valid + fraudulent) records, n/N (%)	Valid records, n/N (%)	Fraudulent records, n/N (%)	Difference between valid and fraudulent records	
				Pearson chi-square (*df*)	*P* value
Hospital visits in last year				124.1 (1)	<.001
0 times	774/1352 (57.3)	467/639 (73.1)	307/713 (43.1)		
1 or more times	578/1352 (42.8)	172/639 (26.9)	406/713 (56.9)		
ED[a] visits in last year				34.9 (1)	<.001
0 times	552/1351 (40.9)	314/638 (49.2)	238/713 (33.4)		
1 or more times	799/1351 (59.1)	324/638 (50.8)	475/713 (66.6)		
Wait time for pediatric subspecialist				17.6 (1)	<.001
2 mo or less	1111/1382 (80.4)	482/638 (75.6)	629/744 (84.5)		
3 mo or more	271/1382 (19.6)	156/638 (24.5)	115/744 (15.5)		
Hours per week in medical visits				108 (1)	<.001
<1 h/wk	208/1347 (15.4)	167/636 (26.3)	41/711 (5.8)		
1 or more hours/week	1139/1347 (84.6)	469/636 (73.7)	670/711 (94.2)		

[a]ED: emergency department.

We examined the validity of different fraud detection methods using epidemiological measures of diagnostic tests (Table 3). Most methods had low sensitivity (poor ability to detect fraudulent records), including comparisons of paired items (eg, age and y of birth), short completion time, duplicate IP addresses, and review of open-ended responses for duplicate or nonsensical answers. Even ReCAPTCHA technology, which relies on information that bots can see but humans cannot, only detected about half of fraudulent responses. The most sensitive indicators appeared to be a fraudulent IP address detection service, where a low threshold setting (ie, one that identified not only IP addresses with confirmed fraudulent activity but also those with suspicious activity) yielded a sensitivity of 88%, and the proprietary Fraud ID feature on the survey software, with a sensitivity of 67%.

XSL•FO
RenderX

Table . Sensitivity, specificity, positive and negative predictive value of fraud detection methods at identifying fraudulent responses in a cross-sectional survey for caregivers of children with new pediatric subspecialty referrals (August-October 2023) (n=1551).

Method of detecting fraudulent responses	True positive, n	False positive, n	False negative, n	True negative, n	Sensitivity (%)	Specificity (%)	Sensitivity + specificity (%)	Positive predictive value (%)	Negative predictive value (%)
Nonsensical open-ended responses	0	0	905	646	0	100	100	__[a]	41.7
Repeated open-ended responses	0	0	905	646	0	100	100	—	41.7
Qualtrics Ballot Box stuffing	0	0	905	646	0	100	100	—	41.7
Duplicate IP addresses	82	45	823	601	9.1	93	102.1	64.6	42.2
Short completion time (<4 min)	135	1	770	645	14.9	99.8	114.7	99.3	45.6
Discrepancy between paired items (age and year of birth)	183	15	642	631	22.2	97.7	119.9	92.4	49.6
Fraudulent IP address database (lower threshold)[b]	758	337	106	309	87.7	47.8	135.5	69.2	74.5
Longitude/latitude	389	9	637	516	37.9	98.3	136.2	97.7	44.8
Fraudulent IP address database (higher threshold)[b]	422	8	442	638	48.8	98.8	147.6	98.1	59.1
Qualtrics ReCAPTCHA functionality	466	0	439	646	51.5	100	151.5	100	59.5
Qualtrics Fraud ID	609	27	296	619	67.3	95.8	163.1	95.8	67.7
Combination: Qualtrics ReCAPTCHA or Fraud ID	675	37	230	609	74.6	94.3	168.9	94.8	72.6

[a]Not available.

[b]The service used to look up fraudulent IP addresses had two thresholds: a lower threshold indicating that the IP address had been used in suspicious activity, and a higher threshold, indicating that it had been used in fraudulent activity.

Sensitivity is typically balanced against specificity, the ability of a diagnostic test to accurately predict which people do not have a disease (in this case, to identify valid records). Almost all individual indicators yielded relatively high specificity, suggesting that they correctly identified the valid records in the sample, even as most of these methods could not identify the fraudulent ones well (Table 3).

When we examined the sum of sensitivity and specificity and used the proportional threshold of 150% or greater as an indicator of the ability of a test to identify cases, only fraud detection tools embedded in the survey software (ReCAPTCHA and Fraud ID within the Qualtrics system) met the criteria of being more useful than chance alone in identifying fraudulent records. Notably, each of these methods used alone would have missed between 33% and 48% of fraudulent records, based on

XSL•FO
RenderX

their modest sensitivity. Using the two measures together (ie, flagging if either was positive) would have resulted in a sensitivity of 75% and a specificity of 94%, with a combined total of 168.9%.

When we calculated predictive values for each method, we found that several had positive predictive values in excess of 90%, meaning that if they were activated (a positive test), it was a fairly good indicator that a record was indeed fraudulent, but negative predictive power less than 60%, indicating that the failure of a record to activate one of these indicators was not assurance that the record was valid. For example, activation of the survey software's ReCAPTCHA indicators was strongly suggestive of a fraudulent record (positive predictive values of 100%), but its negative predictive value (59.5%) suggests that it would miss around 40% of fraudulent records.

Discussion

Similar to past studies [1,3,5-10], in this large statewide survey, we observed a high rate of fraudulent records compared to valid ones (1.4 fraudulent records for every valid record) that would have significantly altered the characterization of the sample and study findings if undetected. Strategies recommended for survey fraud detection performed poorly as judged by epidemiological standards for detecting the "disease" of fraud. These included a number of tried-and-true methods, such as the use of paired items (age and year of birth) at opposite ends of the survey [2,5,8-10], which detected only 1 in 5 fraudulent records, or looking for duplicate IP addresses, which detected only 1 in 10 fraudulent records. Only 2 methods, both proprietary survey software technologies, performed better than chance alone in identifying fraudulent surveys, but even the combination of the 2 methods failed to detect about a quarter of fraudulent survey responses.

Fraudulent responses, if undetected, would have significantly altered the characterization of the sample as well as the study findings. For example, the responses from bot attacks indicated greater health care utilization and shorter wait times for pediatric subspecialty care compared to valid responses. Thus, failure to remove bot responses would have resulted in the incorrect conclusion that Californian families have greater access to pediatric subspecialty care than they actually do.

None of the 10 recommended fraud detection methods used in this study was effective as a stand-alone method capable of identifying fraudulent records, even those that other studies have reported to be highly effective. Moreover, our study represents a small number of bot attacks, and comparison to past literature suggests that bot attacks can have highly individualized signatures. For example, MacKinnon et al [20] were able to identify that over 31% of responses were fraudulent based on nonsensical open-ended questions, duplicate IP addresses, geocoding, and short completion time. In our study, these combined methods would have detected only 43% of fraudulent cases. Rather than use our results to suggest discarding or prioritizing particular measures, we echo the call of Dewitt et al [6] for the use of dynamic protocols, which use as many methods as possible to flag suspicious cases and human reviewers to manually review flagged records. The method is analogous to installing motion sensors, alarms at entry points, and safe meddling safeguards at a bank, but activating human guards to investigate if any of these detection mechanisms are activated.

Combating fraudulent survey responses poses a fundamental tension between data quality, inclusion, and resources. Manual review of responses and follow-up with suspect cases was highly labor-intensive in this study. Yet fraud detection protocols that require prior authentication of participants (eg, through requiring participants to upload documents [3]) or seek to expeditiously classify cases based on a priori algorithms risk exclusion of legitimate respondents. Populations with lower income, racial and ethnic minorities, and older adults are already underrepresented in internet research [21-23]. Adding additional barriers may pose not only an inconvenience but also a threat to the inclusion of historically marginalized voices [3]. In the short term, combating bot attacks will require greater allocation of resources. Developing strategies to improve bot detection, reviewing results in real time, and following up with participants when the authenticity of responses is in question are time-consuming and raise the cost of research. Ultimately, the field must invest in developing more cost-effective strategies for identifying and removing fraudulent records.

Advances in artificial intelligence (AI) will likely increase the frequency and sophistication of bot attacks, including mimicking human response biases, in an effort to evade detection. On the other hand, they may also contribute to the development of important bot detection tools that can co-evolve with bot attacks. Machine learning models may be well equipped over time to identify newly emerging patterns of fraudulent activity and to adapt as fraudulent tactics evolve. However, the emerging tool of AI does not make current research on patterns of fraudulent survey activity irrelevant, because development of technology for detecting bot attacks will initially rely on existing datasets, such as the one used for this study, using human classification of responses as valid or fraudulent. Anomaly Detection Theory advances a number of approaches to identify problem cases, such as a search for outliers [17]; however, when there are more fraudulent than valid responses to a survey, as was the case with our survey and has been with others, fraudulent cases may cease to be outliers. The need for accurate data for learning models may, therefore, lend even greater urgency to accurate identification within the datasets that are fed to the AI models.

Limitations of this study include a lack of a true gold standard for identifying fraudulent responses, which meant that we relied on manual review and consensus determination of case dispositions as reflective of the correct disposition of fraudulent or valid. On the other hand, we employed a high-effort strategy to distinguish valid from fraudulent responses, reviewing each response manually and following up with respondents when the validity of the response was in doubt, and we observed patterns unlikely to have occurred through true responses. Because fraud detection technologies such as Qualtrics ReCAPTCHA are proprietary, we cannot offer insight into the mechanisms of their ability to detect some fraudulent attacks or their failure to detect others. We collected data in August–October 2023, and bot-facilitating technology and methods may have evolved since that time. This study was conducted through web-based surveys

XSL•FO
RenderX

using anonymous links shared through community-based organizations serving families of children with chronic conditions. Further study is needed to generalize findings to understand patterns of response as bot attacks become more sophisticated.

The incidence of fraudulent responses to web-based surveys using bots has been rapidly increasing in survey research and threatens the integrity of important research. Our study identified two methods that performed better than chance alone in identifying fraudulent cases; the combination of these two tools, both proprietary tools of the survey software, identified 75% of fraudulent cases. Literature suggests that bot attacks may have diverse phenotypes, and our findings support calls for "dynamic protocols" that capture as many potential markers of fraud as possible for review and follow-up. To avoid the need to expend intensive human resources in future research, it is imperative that researchers begin to employ more sophisticated technology, investing in AI-enabled bot detection approaches that can evolve along with fraudulent attacks. As such, interdisciplinary collaborations with computer scientists and tech entrepreneurs who currently solve security problems could be fruitful to elevate the voices of real people within policy and programmatic discussions.

Acknowledgments

We acknowledge the caregiver participants who generously shared their time and expertise to take part in this study. We extend special thanks to the 17 family resource centers across California that took part in the study, as well as Dr Yvette Baptiste, Robert Rochin, Jacqui Knudsen, and Ana Seda from the Family Resource Center Network of California for helping to coordinate data collection. Alicia DiGiammarino provided invaluable support with copyediting.

Funding

Funding for data collection for the original study was provided by the Lucile Packard Foundation for Children's Health (grant 2023-09171). This secondary analysis was conducted without additional funding. Foundation staff contributed to the design of the original study but were not involved in design of this secondary analysis. Foundation staff AG and HH contributed to the manuscript revision and are listed as coauthors.

Data Availability

Requests for deidentified study data will be reviewed by the authors and made available where appropriate.

Authors' Contributions

This analysis was conceptualized and carried out by RW-G. The fraud detection protocol described herein was designed by TK and RW-G. AG, HH, TK, and RW-G contributed to instrument design and interpretation of analyses. RW-G and MD guided data collection and cleaning, with assistance from EL and AN. RW-G drafted the manuscript. All authors reviewed and provided feedback on the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Literature review, including methods tested, and proportion of fraudulent responses.
[DOCX File, 35 KB - jmir_v28i1e85161_app1.docx]

Multimedia Appendix 2
Approaches to detecting fraudulent responses.
[DOCX File, 22 KB - jmir_v28i1e85161_app2.docx]

Multimedia Appendix 3
Sample language in institutional review board/consent form.
[DOCX File, 19 KB - jmir_v28i1e85161_app3.docx]

References

1. Zuniga C, Ragosta S, Thompson TA. Recruiting foreign-born individuals who have sought an abortion in the United States: lessons from a feasibility study. Front Glob Womens Health 2023;4:1114820. [doi: 10.3389/fgwh.2023.1114820] [Medline: 37143605]

XSL•FO
RenderX

2. Myers KJ, Jaffe T, Kanda DA, et al. Reaching the “hard-to-reach” sexual and gender diverse communities for population-based research in cancer prevention and control: methods for online survey data collection and management. Front Oncol 2022;12:841951. [doi: 10.3389/fonc.2022.841951] [Medline: 35756657]
3. Campbell CK, Ndukwe S, Dubé K, Sauceda JA, Saberi P. Overcoming challenges of online research: measures to ensure enrollment of eligible participants. JAIDS, J Acquired Immune Defic Syndr 2022;91(2):232-236. [doi: 10.1097/QAI.0000000000003035]
4. Pratt-Chapman M, Moses J, Arem H. Strategies for the identification and prevention of survey fraud: data analysis of a web-based survey. JMIR Cancer 2021 Jul 16;7(3):e30730. [doi: 10.2196/30730] [Medline: 34269685]
5. Choi SH, Mitchell J, Lipkus I. Lessons learned from an online study with dual-smoker couples. Am J Health Behav 2017 Jan;41(1):61-66. [doi: 10.5993/ajhb.41.1.6] [Medline: 27935791]
6. Dewitt J, Capistrant B, Kohli N, et al. Addressing participant validity in a small internet health survey (the restore study): protocol and recommendations for survey response validation. JMIR Res Protoc 2018 Apr 24;7(4):e96. [doi: 10.2196/resprot.7655] [Medline: 29691203]
7. Griffin M, Martino RJ, LoSchiavo C, et al. Ensuring survey research data integrity in the era of internet bots. Qual Quant 2022;56(4):2841-2852. [doi: 10.1007/s11135-021-01252-1] [Medline: 34629553]
8. Mournet AM, Kleiman EM. Internet-based mental health survey research: navigating internet bots on Reddit. Cyberpsychol Behav Soc Netw 2023 Feb;26(2):73-79. [doi: 10.1089/cyber.2022.0173] [Medline: 36724303]
9. Pekarsky C, Skiffington J, Leijser LM, Slater D, Metcalfe A. Social media recruitment strategies to recruit pregnant women into a longitudinal observational cohort study: usability study. J Med Internet Res 2022 Dec 12;24(12):e40298. [doi: 10.2196/40298] [Medline: 36508244]
10. Pozzar R, Hammer MJ, Underhill-Blazey M, et al. Threats of bots and other bad actors to data quality following research participant recruitment through social media: cross-sectional questionnaire. J Med Internet Res 2020 Oct 7;22(10):e23021. [doi: 10.2196/23021] [Medline: 33026360]
11. Storozuk A, Ashley M, Delage V, Maloney EA. Got bots? Practical recommendations to protect online survey data from bot attacks. Quant Meth Psych 2020;16(5):472-481. [doi: 10.20982/tqmp.16.5.p472]
12. Klima T, Dedhia M, Willard-Grace R. A survey of caregivers of children and youth with special health care needs: family perspectives on peer support. : Lucile Packard Foundation for Children’s Health; 2024 URL: https://lpfch.org/wp-content/uploads/2024/06/20240520-ES-ATC-P2P-JT7.pdf [accessed 2026-02-17]
13. Access to care in California: the CYSHCN family experience. : Lucile Packard Foundation for Children’s Health; 2024 URL: https://childrens-coalition.org/wp-content/uploads/2024/02/Access-to-Care-in-California-The-CYSHCN-Family-Experience.pdf [accessed 2026-02-17]
14. Travel time and family preferences for in-person care. : Lucile Packard Foundation for Children’s Health; 2024 URL: https://lpfch.org/wp-content/uploads/2024/06/FINAL-Travel-Time_1pgr_JT5.pdf [accessed 2026-02-17]
15. Delays to access in pediatric subspecialty care in California. University of California San Francisco, Center for Excellence in Primary Care. URL: https://cepc.ucsf.edu/delays-access-pediatric-subspecialty-care-california [accessed 2026-02-19]
16. Lawlor J, Thomas C, Guhin AT, et al. Suspicious and fraudulent online survey participation: introducing the REAL framework. Methodol Innov 2021 Sep;14(3):20597991211050467. [doi: 10.1177/20597991211050467]
17. Dogo EM, Nwulu NI, Twala B, Aigbavboa C. A survey of machine learning methods applied to anomaly detection on drinking-water quality data. Urban Water J 2019 Mar 16;16(3):235-248. [doi: 10.1080/1573062X.2019.1637002]
18. Trevethan R. Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. Front Public Health 2017;5:307. [doi: 10.3389/fpubh.2017.00307] [Medline: 29209603]
19. Power M, Fell G, Wright M. Principles for high-quality, high-value testing. Evid Based Med 2013 Feb;18(1):5-10. [doi: 10.1136/eb-2012-100645] [Medline: 22740357]
20. MacKinnon KR, Khan N, Newman KM, et al. Introducing novel methods to identify fraudulent responses (sampling with Sisyphus): web-based LGBTQ2S+ mixed-methods study. J Med Internet Res 2025 Mar 17;27:e63252. [doi: 10.2196/63252] [Medline: 40096683]
21. Jang M, Vorderstrasse A. Socioeconomic status and racial or ethnic differences in participation: web-based survey. JMIR Res Protoc 2019 Apr 10;8(4):e11865. [doi: 10.2196/11865] [Medline: 30969173]
22. Kelfve S, Kivi M, Johansson B, Lindwall M. Going web or staying paper? The use of web-surveys among older people. BMC Med Res Methodol 2020 Oct 8;20(1):252. [doi: 10.1186/s12874-020-01138-0] [Medline: 33032531]
23. Corey KL, McCurry MK, Sethares KA, Bourbonniere M, Hirschman KB, Meghani SH. Utilizing internet-based recruitment and data collection to access different age groups of former family caregivers. Appl Nurs Res 2018 Dec;44:82-87. [doi: 10.1016/j.apnr.2018.10.005] [Medline: 30389065]

Abbreviations

AI: artificial intelligence
FRC: family resource center
REAL: Reflect, Expect, Analyze, Label

XSL•FO
RenderX

Edited by A Stone; submitted 02.Oct.2025; peer-reviewed by M Ekwueme, Y Deng; revised version received 11.Jan.2026; accepted 30.Jan.2026; published 05.Mar.2026.

Please cite as:
Willard-Grace R, Klima T, Dedhia M, Lo E, Nisnevich A, Gray A, Henry H
Using Epidemiological Test Diagnostics to Select Fraud Detection Methods: Secondary Analysis of Quantitative Cross-Sectional Survey Data
J Med Internet Res 2026;28:e85161
URL: https://www.jmir.org/2026/1/e85161
doi:10.2196/85161

© Rachel Willard-Grace, Tali Klima, Mansi Dedhia, Emily Lo, Annie Nisnevich, Allison Gray, Holly Henry. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Feasibility of Smartwatch Micro–Ecological Momentary Assessment for Tracking Eating Patterns of Malaysian Children and Adolescents in the South-East Asian Community Observatory Child Health Update 2020: Cross-Sectional Study

Richard Lane[1], PhD; Louise A C Millard[2], PhD; Ruth Salway[3], PhD; Chris J Stone[4], BSc (Hons), RSci; Andy L Skinner[5], PhD; Sophia M Brady[6], PhD; Jeevitha Mariapun[7], PhD; Sutha Rajakumar[7], PhD; Amutha Ramadas[7], PhD; Hussein Rizal[7], MSc; Laura Johnson[3], PhD; Tin Tin Su[7], DrMed; Miranda Elaine Glynis Armstrong[8,9], PhD

[1]Jean Golding Institute, University of Bristol, Bristol, United Kingdom

[2]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom

[3]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom

[4]Integrative Cancer Epidemiology Programme, School of Psychological Science, University of Bristol, Bristol, United Kingdom

[5]Bristol Medical School, Integrative Cancer Epidemiology Programme, University of Bristol, Bristol, United Kingdom

[6]Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, United Kingdom

[7]South East Asia Community Observatory (SEACO), Jeffrey Cheah School of Medicine and Health Sciences, Monash University Malaysia, Subang Jaya, Malaysia

[8]Department for Health, University of Bath, Bath, England, United Kingdom

[9]NIHR Bristol Biomedical Research Centre, University Hospitals Bristol and Weston NHS Foundation Trust, Bristol, United Kingdom

Corresponding Author:
Richard Lane, PhD
Jean Golding Institute
University of Bristol
Beacon House
Queen's Road
Bristol, BS8 1QU
United Kingdom
Phone: 44 117 928 9000
Email: mh19137@bristol.ac.uk

Abstract

Background: Mobile phone ecological momentary assessment (EMA) methods are a well-established measure of eating and drinking behaviors, but compliance can be poor. Micro-EMA (µEMA), which collects information with a single tap response to brief questions on smartwatches, offers a novel application that may improve response rates. To our knowledge, there is no data evaluating µEMA to measure eating habits in children or in low-to-middle-income countries.

Objective: In this study, we investigated the feasibility of micro-EMA to measure eating patterns in Malaysian children and adolescents.

Methods: We invited 100 children and adolescents aged 7-18 years in Segamat, Malaysia, to participate in 2021-2022. Smartwatches were distributed to 83 children and adolescents who agreed to participate. Participants were asked to wear the smartwatch for 8 days and respond to 12 prompts per day, hourly, from 9AM to 8PM, asking for information on their meals, snacks, and drinks consumed. A questionnaire captured their experiences using the smartwatch and µEMA interface. Response rate (proportion of prompts responded to) assessed participants' adherence. We explored associations between response rate with time of day, across days, age, and sex using multilevel binomial logistic regression modeling.

Results: Eighty-two participants provided usable smartwatch data. The median number (IQR) of meals, drinks, and snacks per day was 2 (2-4), 3 (1-5), and 1 (0-2), respectively, on the first day of the study. The median response rate across the study was 68% (IQR 50-83). The response rate decreased across study days from 74% (68-78) on Day 1 to 40% (30-50) on Day 7 (odds ratio [OR] per study day 0.73, 95% CI 0.64-0.83). Response rate was lowest at the start of the day and highest between the hours

XSL•FO
RenderX

of 12 PM and 2 PM. Female participants responded to more prompts than male participants (OR 1.72, 95% CI 1.03-2.86). There was no evidence of differential response by age (OR 0.73, 95% CI 0.41-1.28). Most participants (65%) rated their experience using the smartwatch positively, with 33% saying they were happy to participate in future studies using the smartwatch. For children that did not wear the smartwatch for the full study duration (n=22), discomfort was the most common complaint (41%).

Conclusions: In this study of the feasibility of µEMA on smartwatches to measure eating in Malaysian children, we found the method was acceptable. However, response rates declined across study days, resulting in substantial missingness. Future studies (eg, through focus groups) should explore approaches to improving response to event prompts, trial alternative devices to increase children's comfort, and evaluate revised protocols for reporting of intake events.

(J Med Internet Res 2026;28:e73435) doi:10.2196/73435

KEYWORDS

eating behavior; ecological momentary assessment; EMA; Malaysia; LMIC; children; adolescents; micro-interaction EMA; µEMA; smartwatch

Introduction

Noncommunicable diseases (NCDs) are the most common cause of death worldwide [1]. In Malaysia [2], NCDs particularly impact lower-income households [3]. Therefore, health surveillance in this population is required to better understand policy interventions that may improve health outcomes in Malaysia. Dietary risk factors accounted for 10% of all deaths globally in 2021 [4]; therefore, measuring eating is a crucial component of health surveillance. Traditional methods for measuring eating and dietary intake include food diaries, 24-hour recalls, diet histories, or food frequency questionnaires. While methods relying on memory of past behavior are subject to error like recall bias, prospective methods like diaries are affected by reactivity, where real or reported behavior is altered owing to the process of documenting food intake in real time. Underreporting is common in all existing methods, with an estimated 263 kcal per day typically missing from self-reported intakes compared with objective measures [5]. Underreporting varies with food type and eating occasion, with snacks and snack foods more likely to be left out of a self-reported record [6-9]. Online tool, such as Intake24 (Newcastle University), that guide users through a 24-hour recall process aim to reduce researcher burden in coding data collected. Photographic methods, where participants are asked to take pictures of their meals rather than write down each food and drink along with its portion size, aim to offer a more objective approach to add portion size estimation and reduced participant burden for capturing real-time food intake [10-12]. However, moving 24-hour recall online has not yet altered estimated underreporting [13,14], and issues with remembering to take photos before consuming foods as well as automating the estimation of foods and nutrients [15] in photos mean that outstanding challenges in dietary assessment methods remain [12]. Therefore, the feasibility of using alternative methods needs to be considered.

Ecological momentary assessment (EMA) is the repeated sampling of current behaviors in real-time in a natural environment [16]. EMA has evolved to be primarily delivered using mobile phones (mEMA), which have improved response rates compared with original pen and paper methods [17-19]. There is a large volume of literature on EMA using smartphones (n=796 studies) [20]. While diet is the second most commonly studied topic, it still only accounts for 4% (35/796) of these studies. Studies of diet using EMA in young people are primarily in the United States and Europe, with just 2 studies in Asia, in China, and Taiwan [18,21].

Liao [22] highlighted that response rates and compliance with EMA protocols were rarely reported. Since then, reporting of compliance has improved, but the response latency remains unknown from many studies [21]. Response rates to mEMA of diet are a median of 74% [23], which is similar to mEMA of all topics (mean 75% (IQR 64%-84%) [20]. A review of mEMA for diet in young people (16-30 years) showed response rates mostly exceeded 80% [21], whereas at younger ages poorer responses <80% are more often observed [18]. Lower response rates have also been associated with weekends versus weekdays [21], when participants receive more prompts during the day [17,20], and in males versus females [21].

Smartwatches are an emerging technology for collecting data alongside sensor data using micro-EMA (µEMA) protocols. This captures information using single-tap responses to brief questions, which is suitable for the small screens on these devices [19,23]. In adults, µEMA has been found to yield higher compliance rates despite more frequent sampling than mEMA and is perceived by users as less distracting [24]. Further, the use of µEMA significantly improves response rate (mean 72% vs 82%) but remains rare, with only 12 studies on any topic in any age group [20]. Despite these advantages, some limitations of µEMA have been reported in the literature, including limited battery life and technical problems such as problems with charging [25].

In children and adolescents, the use of pen and paper EMA to measure diet has typically been implemented outside of school hours [18]. Internet-connected devices such as mobile phones are often used for mEMA data collection [26]. These may be less suitable for child and adolescent populations, where 40% of education systems now ban the use of smartphones in school [27]. Further, devices such as smartwatches that can function without an internet connection may be better suited to rural, semirural, and low-resource settings where communication infrastructure may be less well-developed [28]. Therefore, smartwatches offer the potential to implement EMA across the whole day, with the potential for additional advantages such as improved compliance and response rates [20,24]. To our knowledge, only two diet studies involving adults in the United

XSL•FO
RenderX

Kingdom and the United States have reported on EMA with smartwatches, and none have involved children [18,23,29].

Therefore, this study investigates the feasibility of using smartwatch-based µEMA to record eating patterns in Malaysian children and adolescents. The collected µEMA data are used to examine the completeness of the collected data and factors associated with response rates, alongside survey responses assessing participants' experience during the study. Establishing the feasibility of this novel dietary measurement tool is an important first step to inform utility and any required refinement prior to deployment for dietary measurement more widely.

Methods

SEACO-CH20 Study

The South-East Asian Community Observatory (SEACO) health and demographic surveillance cohort is a dynamic cohort of 13,335 households in Segamat, a semirural region in the state of Johor Darul Takzim, Malaysia. The cohort was established in 2012, with surveys, blood tests, and physical measurement data collected from participants. In 2013 and 2018, health surveys were conducted on ~25,000 adults and children, 25,168 in 2013 and 24,710 in 2018.

All households (18,602) in 5 subdistricts, which SEACO operates, were invited to participate in the 2017 census. Altogether 11,617 households (40,184 residents) accepted our invitation. In 2018, participants who were involved in the 2017 census and were older than 5 years were invited to participate in the 2018 health round data survey, to which a total of 24,710 participants agreed. Potential participants from 3 subdistricts were preselected and approached via telephone using the existing health database (HR 2028). Participants' parents were approached via telephone for recruitment before the home visit.

Children and adolescents aged 7-18 years who were part of the SEACO cohort were invited to participate in the SEACO Child Health 2020 update (SEACO-CH20) study; a systematic review of EMA studies in youth recommended 7 as a lower age limit for EMA [26]. The eligibility of households was limited by location due to the safety measures implemented during the COVID-19 pandemic to reduce the risk to participants, households, and fieldworkers. Therefore, the 1993 children and adolescents invited to participate were from only 3 of the 5 SEACO subdistricts (Jabi, Sungai Segamat, and Gemereh) in the Segamat district.

Data collection visits to individual households were performed in person from November 1, 2021, to July 31, 2022. The data were collected as part of a larger study and included surveys, physical measurements, such as height, weight, blood pressure, waist and hip circumference, and blood sample collection. Participants were given wrist-worn Axivity AX6 6-axis accelerometers to monitor their physical activity, which were worn 24 hours per day over 7 days [30]. A random subset of the participants were also given TicWatch C2 (Mobvo) Android smartwatches to record eating and drinking with µEMA as part of this feasibility study, using a smartwatch µEMA system developed within the research team. The smartwatch system used was an adaptation of a µEMA system first used in a study involving high-temporal density longitudinal measurement of alcohol use [31] by a subset of the research team who developed this system. Participants wore these devices over the same 7-day period as the accelerometers. As they are both wrist-worn devices, this may have affected the acceptability of the smartwatch. Participants were briefed on the use of these devices by the data collectors, including how to charge them and how to replace them after charging. The original data collection plan can be seen in Figure S1 in Multimedia Appendix 1. A completed Checklist for Reporting Ecological Momentary Assessment Studies (CREMAS) [22] can be found in Multimedia Appendix 2.

Study Participants

Participants for the SEACO-CH20 study were selected from the larger SEACO cohort. Parents of participants provided consent for their children to participate in SEACO-CH20. A random subsample of 100 participants were each invited to wear a smartwatch.

Data Collection

SEACO-CH20 fieldworkers performed 2 home visits to collect the data. The smartwatch was distributed on the initial visit, and the participants were briefed on how to use and charge it. They were instructed to wear the device for the next 8 days, on "the wrist that [they] use to write." The smartwatches were collected during the second home visit, and the participants were asked to complete a questionnaire on their experience with the devices. Questionnaires assessed the participants' attitudes toward several aspects of the smartwatch study, including ease of use, their attitude toward charging, and their overall experience. Since children as young as 10 were asked to complete the questionnaires, a reduced set of acceptance questions was used to reduce burden. This was based on similar pilot work using novel methods in the ALSPAC G2 study [32] and included the following questions:

- Overall, how would you rate your experience of using the smartwatch during the study, on a scale from 1 (didn't like it at all) to 5 (really liked it)?
- If you were asked to use the smartwatch again in another study, would you participate?
- How many days in total did you wear the smartwatch for?
- If you wore the smartwatch for less than 8 days, what were the main reasons for not wearing it longer?

Parents of participants aged 7-9 years completed the survey on behalf of their children, while participants aged 10 years and older filled out the survey themselves. The full text of the survey can be found in Figure S2 in Multimedia Appendix 1.

Smartwatch µEMA Questions

During the study, the smartwatch prompted participants once every hour to enter any food or drink that they had consumed in the last hour. These prompts were scheduled to appear once every hour from 9 AM to 8 PM, so participants were expected to interact with the smartwatch 12 times throughout the day. We chose this hourly prompt frequency to maximize the chance that eating and drinking events were less likely to be missed and to capture more fine-grained temporal patterns in eating

XSL•FO
RenderX

behavior. As this was a feasibility study, this choice was justified, given that µEMA has been shown to improve compliance despite more prompts than mEMA in adults [26]. The smartwatch interface included the following 5 questions that the participants completed for each item consumed:

1. Have you had any food or drink in the last hour? Options: yes, no
2. What did you have? Options: meal, snack, drink
3. What size was it? Options: small, medium, large
4. What did you use to eat? Options: hands, fork/spoon, chopsticks
5. Where were you? Options: home, school, elsewhere

A possible future use of this methodology, once fully refined, could include automated eating detection [33]. Therefore, we included a question on the type of cutlery used ("What did you use to eat?"), as lack of information on utensil type has been highlighted as a limitation of some datasets used for algorithm development related to automated eating detection [34].

After entering this information for one item consumed, they were asked, "Any more food or drink to record?" and could then start again to add another entry. Therefore, each consumption entry either indicates that the participant did not eat or drink in the last hour or contains the answers to the above questions for a particular meal, drink, or snack, linked to an hour period within a day. If participants ignored the prompts, they would receive a reminder prompt after 1 minute; if they continued to ignore the prompt for a further 1 minute, the prompt would disappear and "no response" would be recorded by the smartwatch.

Participants could choose "back" on each question screen to return to a previous question and update their response. However, after submitting their answers for a particular item (ie, completing the "where were you" question for that item), they would not be able to return to that entry.

An additional prompt (the "catch-up") was scheduled every morning at 8 AM asking if they had consumed any food or drink on the previous day that had not been recorded on the smartwatch. If they indicated "yes," they were asked the same questions as above. Catch-up entries did not have an associated eating time but were labeled as catch-up–type events, indicating that they applied to the previous day.

The smartwatch study was co-created and piloted with the Malaysian research team and the English was translated into Malay for use on the smartwatch. All the original data collection was in Malay. The smartwatch protocol, including the prompts and possible responses, can be seen in Tables S1-S3 in Multimedia Appendix 1.

Smartwatch Data Cleaning

Smartwatches were distributed by fieldworkers partway through the day, and µEMA responses on this distribution day were removed from analyses. The study period is taken to be the subsequent 7 days after this distribution day.

The version of the EMA software we used did not save the hour period to which each entry belonged. Therefore, we needed to infer this from the submission timestamp, the date and time a particular entry was submitted. As entries for the same hour period are submitted one after the other, we used a time window to group nearby entries into a single "eating event," which is intended to capture the participants' responses to one prompt. A 30-minute window was chosen to group nearby prompts, as we expect this to collect entries from the same eating event without grouping prompts from adjacent hours. Previous work from diet diaries suggests that 30 minutes is a reasonable cut-point to distinguish independent eating occasions [35]. Occasionally, there may be participants with more than 12 eating events per day, for example, if they took more than 30 minutes to finish responding to a prompt. This occurred on 26 occasions, less than 5% of the total 574 (82 participants multiplied by 7 days) study days.

The µEMA data used was restricted to the 7 days after the distribution day. During the data review, we identified an issue with the collected data where there were sometimes multiple identical entries for a given intake event due to an issue with the µEMA software. Therefore, duplicate entries were identified as any pair of entries with identical contents (same meal type, portion size, utensil, and location), for the same hour period, and entered within 5 minutes of each other. The first such entry was kept in each case. Around 588 duplicate responses were removed of 10,539. Data cleaning was performed in Python (version 3.10.0; Python Software Foundation).

Response Rate

The response rate was calculated as the proportion of prompts responded to (with either at least one item consumed or an entry stating that they did not eat or drink anything in the previous hour). The response rate tells us the extent that participants engaged with the smartwatch app throughout the day but not the extent that the data entered are complete, that is, whether all intake events were recorded. We therefore summarized the number of each type of meal entry (meal, drink, snack, or no food or drink) submitted per day across our sample. For these summaries we included only participants who took part in the study outside the Ramadan fasting period (April 3, 2022-May 1, 2022; n=67), since fasting participants are likely to enter fewer eating events during the day. The mean of participants' response rates per day was recorded, and the median and quartiles of these were reported.

Attrition from the study was examined by identifying the last day each participant responded to any smartwatch prompt; participants who had ceased responding to the prompts are referred to as "inactive."

Statistical Analyses

We summarized response rates to each individual prompt of the smartwatch µEMA and the participants' experiences based on the survey questions. We used mean for continuous variables, n (%) for categorical variables, and median and IQR for ordinal or nonnormally distributed continuous variables.

We used a mixed-effect logistic regression model for the response (yes or no) to an individual prompt on a specific study day of data collection for each participant. A fixed effect term was included for study day (from the first to the seventh day) as a continuous linear trend. The time of day was also included

XSL•FO
RenderX

as a fixed effect to capture nonlinearity in response throughout the day, grouping the prompts by the nearest hour as follows to decrease the number of parameters in the model:

- Morning (9-11 AM)
- Lunchtime (12-2 PM)
- Afternoon (3-5 PM)
- Evening (6-8 PM)

Random intercept and random slope terms were included for study day within each participant. Estimates are provided as odds ratios (OR) and 95% CIs, interpreted as the multiplicative change in the odds of a participant responding to an individual prompt. The degree of difference between participants was summarized in the intraclass correlation coefficient.

To evaluate if changes in participation across wear days differed in boys versus girls, we repeated this base model, adding a fixed term for sex and an interaction term between sex and study day. Similarly, we explored differences by age group (Malaysian primary school age: 7-12 years versus secondary school age: 13-18 years) by adding a fixed term for age group and an interaction term between age group and study day to the base model.

Analyses were performed in Python version 3.10.0 and R (version 4.2.2; R Core Team) [36]. All of our analysis code is publicly available [37]. Git tag 3.0 (The Git Project) corresponds to the version of the analyses presented here.

Ethical Considerations

Written informed consent was obtained from parents or guardians on behalf of the participants. Children and adolescents were also asked to provide their written assent to participate in the study. Ethical approval was obtained from the Monash University Human Research Ethics Committee on March 17, 2020 (Project ID: 23271) and the University of Bristol REC Case no. 2020–4208 (ID nr: 1304255) prior to any data collection. The study was conducted in accordance with the Declaration of Helsinki for experiments involving humans.

Participants were given a token worth up to RM25 to compensate for their time participating in the study. This was divided into RM5 for completion of each of the following components: (1) questionnaires, (2) health check, (3) blood sample, (4) activity monitor, and (5) smartwatch. They were also given a free health screen and a direct referral to the government primary health care clinic if they were identified as high risk.

Study data have been deidentified and can be freely requested from SEACO, Monash University Malaysia Institutional Data Access at “mum.seaco@monash.edu” for researchers meeting the criteria for access to confidential data. Please refer to the web resource hosted on Monash University’s website [38] for more information.

Results

Participants

A flowchart showing the study participants can be seen in Figure 1. Parents of 728 participants consented to their children’s participation in SEACO-CH20, of which 626 provided demographic (age, sex, and ethnicity) and accelerometer data for the larger study. Of these, 100 participants were randomly invited to wear a smartwatch for this smaller feasibility study. Of the 100 participants invited to participate in the smartwatch substudy, 83 participants agreed. The reasons for nonparticipation included concern the device was not comfortable (n=3) and allergies (n=2). The remaining participants rejected the smartwatch without comment. One further participant accepted the smartwatch study but removed the EMA app from the watch during the study period, rendering their data unrecoverable, resulting in 82 participants who provided smartwatch data.

The sex, ethnicity, and age breakdown for all participants who took part in the smartwatch study can be seen in Table 1.

XSL•FO
RenderX

Figure 1. Study flowchart. Eligible participants were selected from the SEACO Health Round Survey 2018 (SEACO HR-2018) cohort. Reasons for rejecting the smartwatch study included concern about discomfort and allergies. SEACO-CH20: South-East Asian Community Observatory Child Health 2020; SEACO HR-2018: South-East Asian Community Observatory Health Round Survey 2018;.

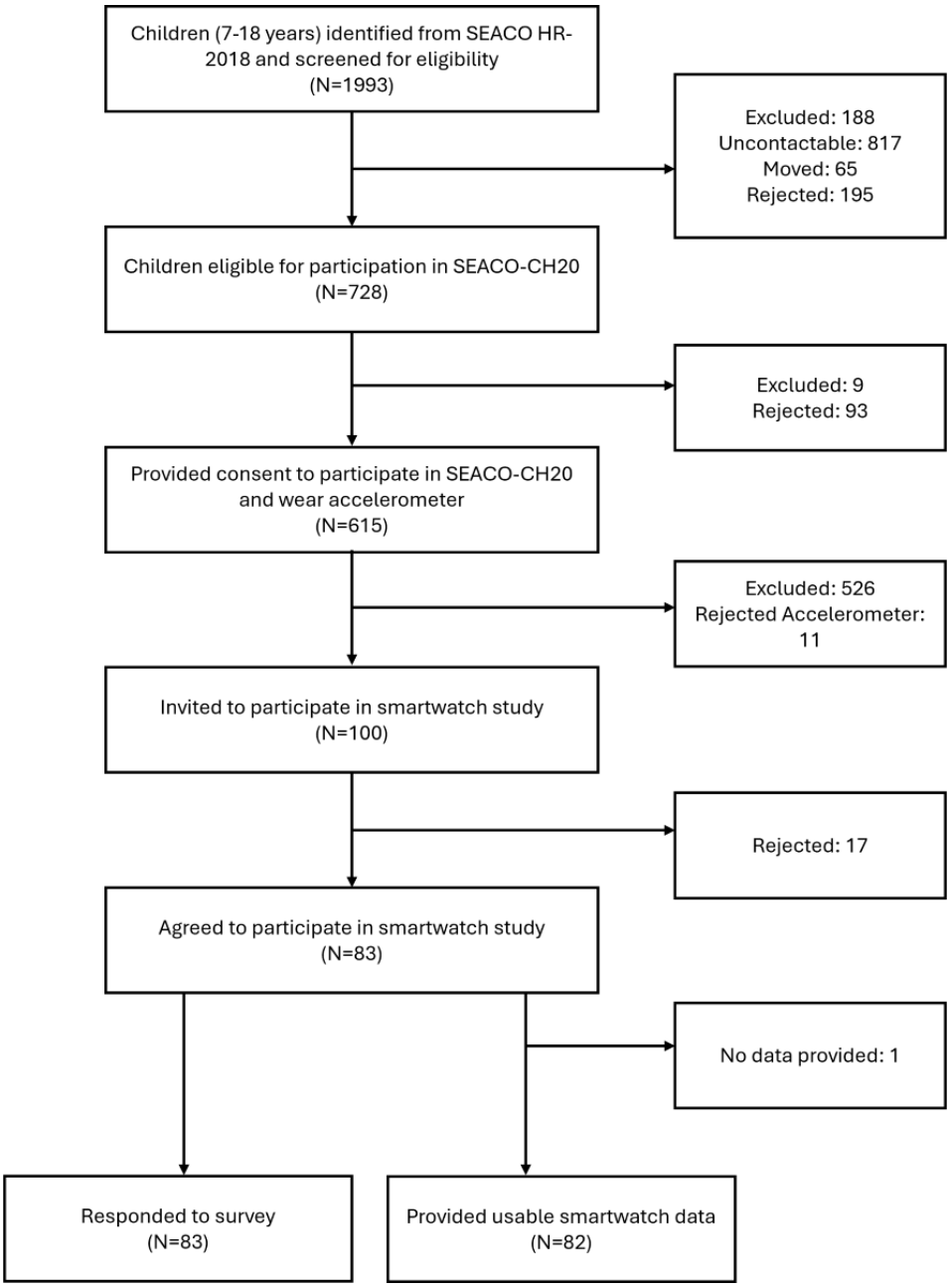

XSL•FO
RenderX

Table 1. Summary of participant demographics (N=83).

Participant characteristic	Smartwatch participants, n (%)
Sex	
Female	53 (64)
Male	30 (36)
Ethnicity	
Malay	73 (88)
Non-Malay	10 (12)
Age (years)	
7-12	24 (29)
13-15	28 (34)
16-18	31 (37)

Smartwatch Responses

The median prompt response rate was 69% (IQR 52%-82%). The number of participants who became inactive on each day can be seen in Figure 2. The majority (55/82, 67%) of participants were active until day 7, that is, they responded to at least 1 prompt on day 7.

Figure 2. The number of participants who became inactive on each day (N=82), that is, having responded to no µEMA prompts after this day. All participants were active for at least one day.

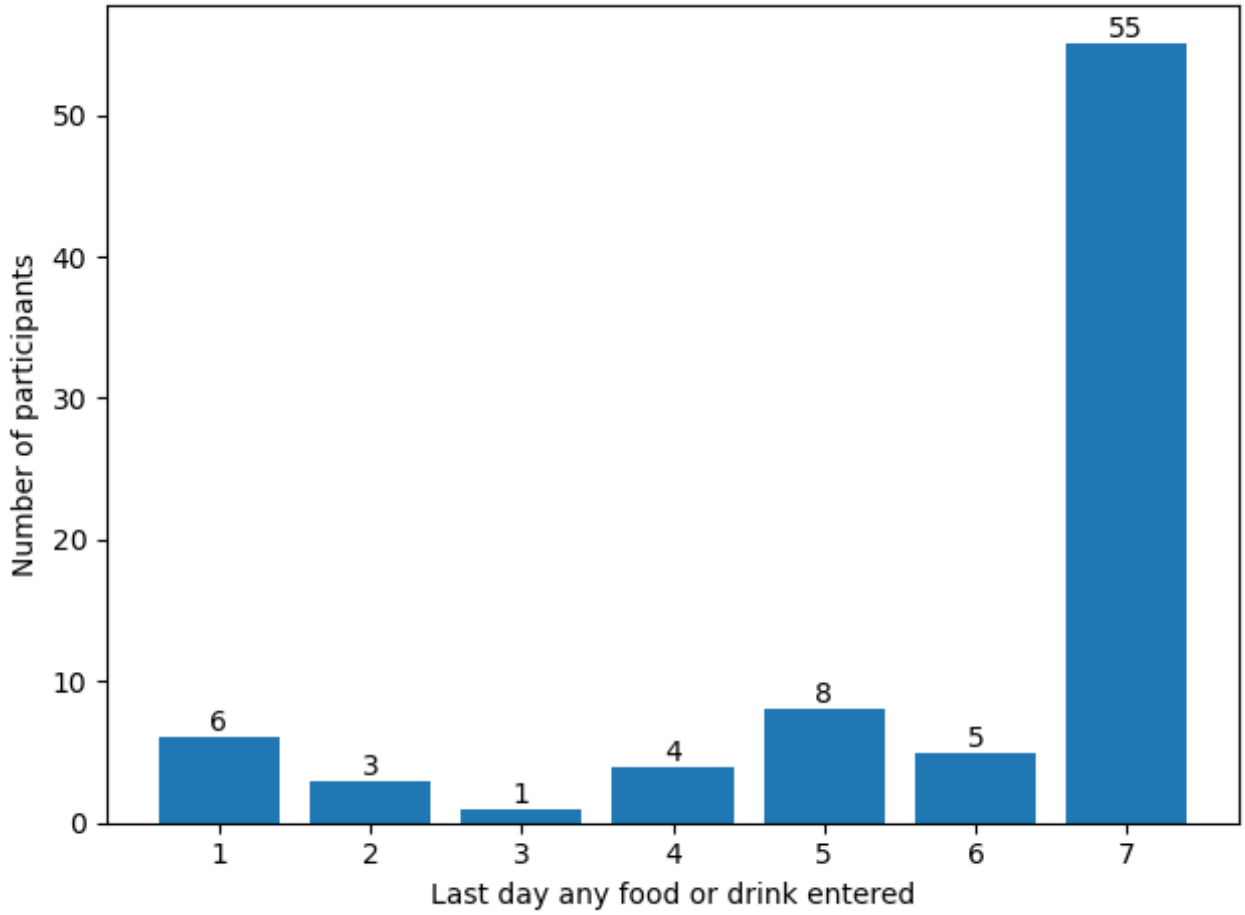

The median and IQR in the number of entries of each type made by each participant per day are summarized in Table 2. Fifteen participants took part (at least partially) during Ramadan and so were excluded from these summaries. Only a minority of intake events were submitted as catch-up entries (N=125 catch-up entries versus 4705 noncatch-up).

XSL•FO
RenderX

Table 2. The median and IQR of the number of noncatch-up entries per day per participant, for participants whose study period did not intersect with Ramadan (N=67).

Study day	Meal, median (IQR)	Drink, median (IQR)	Snack, median (IQR)
1	2 (2-4)	3 (1-5)	1 (0-2)
2	2 (1-3)	2 (1-5)	1 (0-2)
3	2 (1-3)	2 (0-3)	0 (0-1)
4	2 (1-3)	1 (0-3)	1 (0-1)
5	1 (0-2)	1 (0-3)	0 (0-1)
6	1 (0-2)	1 (0-2)	0 (0-1)
7	1 (0-2)	0 (0-2)	0 (0-1)

Response Rate Across and Within Study Days

The response rate for individual prompts had a median (IQR) of 67% (50-83). The response rate on each day ranged from 83% (66-92) on day 1 to 58% (33-75) on day 7.

Figure 3 shows the response rate with study day and time. The response rate decreased across study days (OR for each additional day of the study: 0.73 (95% CI 0.64-0.83). The response rate was lowest at the beginning of the day; the OR and 95% CIs are summarized in Table 3. The intraclass correlation coefficient was 0.207, which indicates that approximately 21% of the total variance in prompt response behavior was attributable to between-participant differences.

Figure 3. Participants' response rates against study day (left) and time of day (right; N=82).

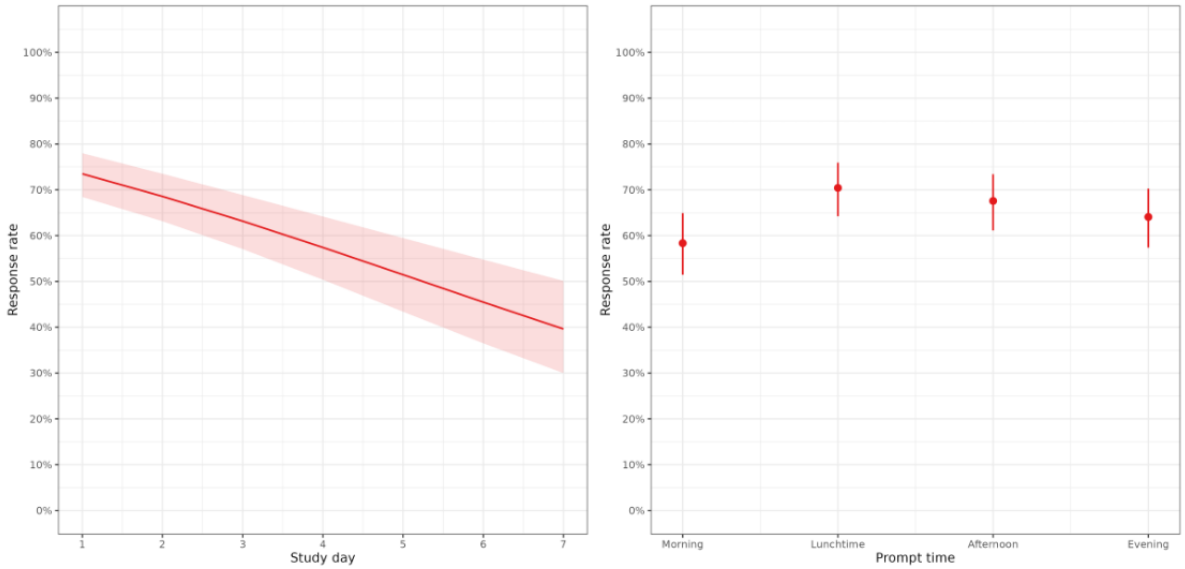

Table 3. The odds ratios for responses at different times of day, taking breakfast time (9-11AM) as the reference level. The response rate was lowest at the beginning of the day.

Time	Odds ratio (95% CI)
Breakfast (9-11 AM)	1[a] (0-0)
Lunchtime (12-2 PM)	1.69 (1.43-2.12)
Afternoon (3-5 PM)	1.49 (1.25-1.76)
Evening (6-8 PM)	1.27 (1.07-1.51)

[a]Reference level.

The results of analyses estimating differences due to sex and age are shown in Figure 4. Girls responded more often to the µEMA prompts compared with boys (OR 1.71, 95% CI 1.03-2.84). However, the daily patterns were similar for both sexes (interaction term OR 1.07, 95% CI 0.93-1.23). Response rate did not differ between age groups (OR 0.73, 95% CI 0.42-1.27), and daily response patterns were similar for the 2 age groups (study day-by-age interaction OR 1.11, 95% CI 0.95-1.29).

XSL•FO
RenderX

Figure 4. Association of micro-interaction ecological momentary assessment (μEMA) prompt response with study day stratified by sex (left) and age (right).

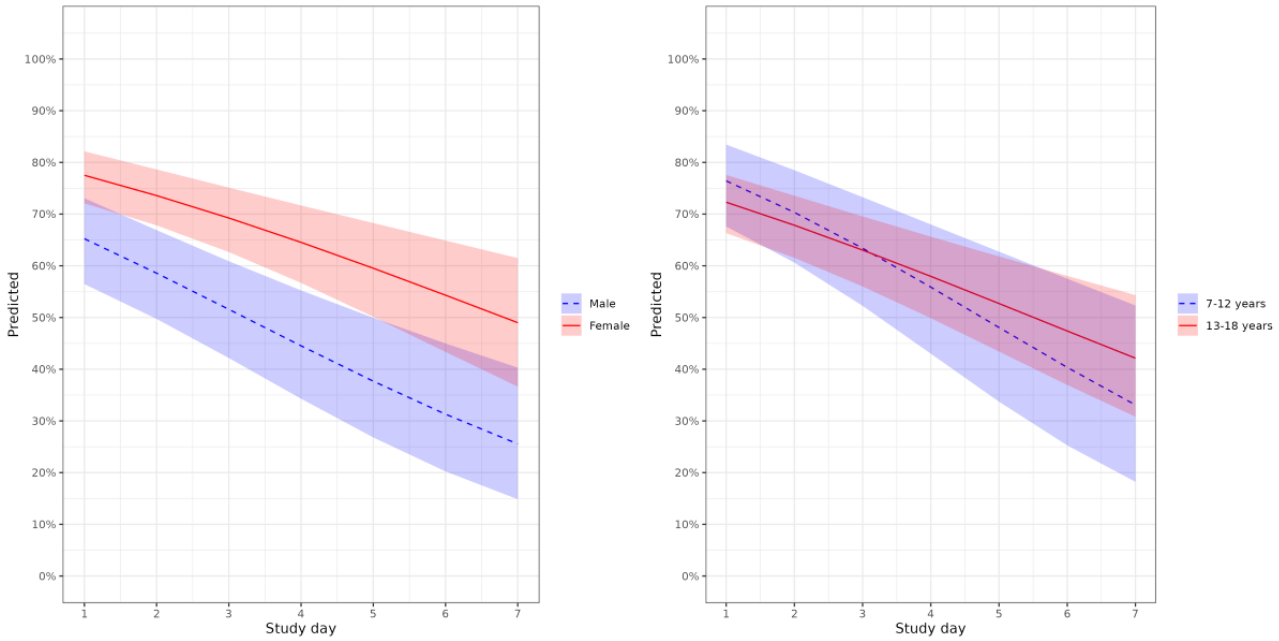

Evaluating Participants' Survey Responses on Acceptability

A summary of responses to questions about smartwatch acceptability and wear time is provided in Table 4.

A total of 54 out of 83 (65%) participants rated their experience using the smartwatch positively (a rating of 4 or 5 out of 5), 20 out of 83 (24%) gave a neutral rating (3 out of 5), and 8 out of 83 (10%) rated it negatively (1 or 2 out of 5). In addition, 27 out of 83 (33%) participants said that they would be happy to participate in future studies using the smartwatch, while 39 out of 83 (47%) said maybe, and 12 out of 83 (14%) said no. The majority of participants who responded (61/83, 73.5%) reported wearing the smartwatch for the entire duration of the study (8 days).

Table 4. The survey questions. Participants were directed to wear the smartwatch on the day that the watch was distributed and for the 7 subsequent days, making 8 days total. Missing data and participants who refused to respond are not included (N=83).

Question and response	Participants, n (%)
Overall, how would you rate your experience of using the smartwatch during the study, on a scale from 1 (didn't like it at all) to 5 (really liked it)?	
Negative (1 or 2)	8 (10)
Neutral (3)	20 (24)
Positive (4 or 5)	54 (65)
If you were asked to use the smartwatch again in another study, would you participate?	
No	12 (14)
Maybe	39 (47)
Yes	27 (33)
How many days in total did you wear the smartwatch for?	
5 or fewer	7 (9)
6	7 (8)
7	8 (10)
8	61 (73)

For those who reported that they did not wear the smartwatch for the entire duration of the study (22 participants), the most common reason was that they did not find it comfortable to wear (9/22, 41%). Other reasons included forgetting, that they did not see the benefit when they could not see the data, they were forbidden to wear it by school, and it ran out of battery (all 3 or fewer responses).

Summaries of the participants' responses to the remaining survey questions can be found in Tables S4-S8 and Figure S3 in Multimedia Appendix 1. Further summary statistics, including

XSL•FO
RenderX

catch-up events and participants who took part during Ramadan, can be found in Tables S9-S11 in Multimedia Appendix 1.

Discussion

Principal Findings

In this feasibility study of a smartwatch-based µEMA method to collect data on eating habits over 7 days in Malaysian children, we found that most participants (55/83, 67%) remained responsive to prompts up to the last day of the study. Participants were least likely to respond to prompts between 9 and 11 AM and most likely between 12 and 2 PM. The intraclass correlation coefficient was 20.7%, suggesting that while some variation in response pattern is attributable to between-participant differences, the majority of the variation (79.3%) was due to within-participant differences. The response rate dropped off day-on-day and was higher for female than male participants; no association was found between participant age group and response rate.

Our average response rate of 69% was lower than the average of 78% found in a meta-analysis of EMA in children and adolescents, including studies that prompted between 2 and 9 times daily [17]. That study found that prompting participants more often had a large negative effect on completion rate, which is further supported by Kraft et al [20], which found a negative correlation of –0.12 between increased number of prompts and response rate (P=.009). Participants in our study were prompted 12 times a day, plus an additional catch-up prompt in the morning. We justified our original prompt frequency choice, as µEMA has been shown in adults to improve compliance despite a higher number of prompts than mEMA. However, in our study using a child and adolescent population, it is likely a higher prompt frequency may have had a negative effect on response rate, especially in the case of repeated "No food/drink" entries. It has been reported [17] in nonclinical studies that "a higher average compliance rate was observed in studies that prompted participants 2-3 times daily (91.7%) compared with those that prompted participants more frequently (4-5 times: 77.4%; 6+ times: 75.0%)." This suggests that compliance may be improved by prompting participants less frequently, for example, by having 3 prompts daily at 11 AM, 3 PM, and 7 PM, although longer time intervals increase the reliance on memory, potentially affecting the completeness of recorded consumption events. It is also possible that our lower response rate might have been related to the number of questions asked in our µEMA protocol. A study that compared the deployment of 6 back-to-back multiple-choice questions delivered via a smartwatch versus a mobile phone found no difference in compliance between these 2 modalities. However, compliance was improved when single questions requiring a one-touch response were asked via a smartwatch, despite an increase in prompt frequency [39]. While on average, the participants (65%) rated the study protocol positively (either 4 or 5 out of 5), the response rate fell day-on-day. Participants were less likely to respond to prompts at the beginning or end of the day, compared with the middle of the day. Focus group or interview discussions were not feasible in this study due to COVID-19 restrictions but should be explored in future studies to determine the reasons for missing event prompts and nonresponses, which may include forgetting or being involved in a competing activity when the prompt is sent [40].

Female participants had a higher response rate than male participants, consistent with previous findings [21,41]. There was little evidence of difference in the relationship between response and study day for male versus female participants. An analysis of the SEACO-CH20 accelerometer dataset [30] found that a similar proportion of males and females had usable accelerometer data, suggesting that this was specific to the smartwatches rather than a difference with wrist-worn devices generally. Little evidence of a difference was found between response rates in the 7-12 years old and 13-18 years old age groups.

Although the subjective indicators suggested that most participants enjoyed wearing the smartwatch, only a minority of participants (27/83, 33%) indicated that they would be willing to participate in a similar study again; 39 out of 83 (47%) responded "Maybe." Potential changes to the protocol that may improve compliance could include only wearing the smartwatch instead of both the smartwatch and accelerometer.

Strengths and Limitations

This is the first study exploring the feasibility of using smartwatch-based EMA in a population of children and adolescents from a low-to-middle-income country. This study was part of the SEACO study, using the SEACO-CH20 dataset, and lays the foundation for an improved understanding of the potential for wearable devices for measuring relationships between eating and cardiometabolic health. Data on 24-hour eating behaviors are important for informing future policy that may reduce cardiometabolic risk among children and adolescents and prevent progression to cardiometabolic disease in adulthood. While a food frequency questionnaire was completed as part of the larger SEACO-CH20 study and reported elsewhere [42], this current study has assessed an alternative for recording behavior in real time.

However, this study did have some limitations. The number of questions in prompts may have affected compliance and should be discussed with participants to optimize the protocol for future studies in this population. It has been suggested [17] that compliance can be improved by incentivizing participants with a monetary reward or raffle entries. Since this study concerns young people, one potential incentive method could be to gamify the EMA process using a level-up or promotion system in the app [43]. Previous studies have explored adding an end-of-day catch-up prompt, which has been found to improve the reporting of dinner [40]. Replacing the morning catch-up prompt with one in the evening may improve response rate, especially given that we found that participants were more likely to respond to prompts in the evening than in the morning. Future studies may additionally consider using the catch-up entries to impute missed entries on the previous day, which could give more complete data. Another suggestion could be to incorporate a short period of training to improve response rates, where responses are monitored in real time by researchers and participants are prompted directly by researchers if missing responses are

XSL•FO
RenderX

common. Such an approach has been used to improve the accuracy of real-time food photography methods [12].

The questionnaires used for acceptability and acceptance were not standard because we were motivated to use a reduced set of questions (that have been used in a previous publication [31]) to reduce the burden on the young participants. Future studies should consider if the emerging, more standard approaches to exploring acceptability and acceptance for wearable devices (eg, those based on Technology Acceptance Models) have been developed to the point at which varying levels of participant burden can be accommodated.

Unbalanced statistics limited our ability to assess differences across age, sex, and ethnic group. The larger SEACO-CH20 accelerometer study [34] from which participants for this study were selected had more balanced statistics (49% female, 67% Malay, and 44% <13 years old), which suggests that the cohort used for smartwatch data may not represent the overall SEACO-CH20 cohort. In particular, we only had 4 participants aged 7-9 years, so further studies are required to better understand the feasibility of dietary µEMA in the younger participants. Participants in this study began wearing the devices on different days of the week, and it is possible that the day of the week could affect participation; for example, whether it is a weekend or weekday. A lower response rate on weekends has been previously documented by Battaglia [21]. We did not account for study start day due to the small sample size, and because schooling was disrupted throughout the study period due to the COVID-19 pandemic [44].

An issue with entry duplication meant that some entries may have been removed that were actual events, not due to the software issue. This bug with the smartwatch software has since been fixed. Additionally, only the response time of the participant was recorded, and not the time that the prompt was sent. This means we had to infer which hour window the entries corresponded to; this could be programmed in the software.

Discomfort was the most common reason for nonwear cited by the participants, which may be unique to our study protocol that required participants to wear 2 wrist-worn devices on the same arm. Furthermore, the smartwatch used in our study was not specifically designed to fit smaller children. Efforts to make smartwatches less intrusive, for example, by making them smaller, may further improve response rate and study uptake.

Ramadan, a culturally important event in Malaysian society, which includes fasting in some population groups, took place during the course of the data collection period. This is likely to have affected the eating behaviors of participants who took part during this time. To ensure that the number of EMA entries was not influenced by Ramadan, we excluded participants from our analyses who wore the smartwatch during the Ramadan period, thereby further limiting our sample size.

Conclusions

This study extends previous eating behavior studies by exploring the use of µEMA in a population of children and adolescents in Malaysia and is the first such study to do so. Willingness to take part in the µEMA study was high, but poor response rates suggest that the number of questions asked per prompt or the high number of prompts per day may be too burdensome. While our smartwatch-based EMA app was largely based on the µEMA methods originally developed by Intille et al [24], a key aspect of true µEMA implementation is the presentation of only one question at a time. In our approach, we chained questions to capture details on food and drink type, size, and consumption context, making it more accurately described as a modified µEMA. Further work is needed to explore different µEMA variations, including using fewer questions and/or fewer prompts, and identify devices that may be more comfortable for child and adolescent participants. The growing use of smartwatches amongst children, particularly in Southeast Asia may offer more opportunities for further study [45].

Acknowledgments

For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) license to any author accepted manuscript version arising from this submission.

The authors would like to express their appreciation to the SEACO Field Teams and survey participants. The SEACO [46] funded the research detailed in this paper. The authors' opinions, however, are their own, and there is no real or implied sponsorship from SEACO.

Funding

The Medical Research Council (MR/T018984/1) and the Ministry of Higher Education/UK-MY Joint Partnership on Non-Communicable Diseases (2019/MR/T018984/), both provided funding in support of this research. The SEACO health and demographic surveillance system is supported by Monash University. The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript. Sophia M Brady is funded by the National Institute for Health and Care Research (NIHR) Applied Research Collaboration (ARC) North East and North Cumbria (NENC; NIHR200173). The NIHR Bristol Biomedical Research Centre funds Miranda EG Armstrong (NIHR203315). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

XSL•FO
RenderX

Data Availability

Data cannot be shared publicly for confidentiality and ethical reasons. Deidentified data are available and can be freely requested from the South East Asia Community Observatory, Monash University Malaysia Institutional Data Access at "mum.seaco@monash.edu" for researchers who meet the criteria for access to confidential data. For more information, please refer to [38].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary information, including the original data collection plan; survey instruments; smartwatch EMA protocols and prompts; and additional smartwatch and survey tables and figures.
[DOC File , 321 KB - jmir_v28i1e73435_app1.doc]

Multimedia Appendix 2
Checklist for Reporting Ecological Momentary Assessment Studies (CREMAS).
[PDF File (Adobe PDF File), 105 KB - jmir_v28i1e73435_app2.pdf]

References

1. Invisible numbers: the true extent of noncommunicable diseases and what to do about them. World Health Organization. 2022. URL: https://www.who.int/publications/i/item/9789240057661 [accessed 2025-03-20]
2. Low W, Lee Y, Samy AL. Non-communicable diseases in the Asia-Pacific region: prevalence, risk factors and community-based prevention. Int J Occup Med Environ Health 2015;28(1):20-26. [doi: 10.2478/s13382-014-0326-0] [Medline: 26159943]
3. Isamail N, Said R, Ismail NW, Haron SA. Non-communicable diseases impact low-income households in Malaysia. Malays J Med Sci 2024;31(1):124-139 [FREE Full text] [doi: 10.21315/mjms2024.31.1.11] [Medline: 38456101]
4. Diet. Institute for Health Metrics and Evaluation. URL: https://www.healthdata.org/research-analysis/health-risks-issues/diet-research-library [accessed 2025-03-25]
5. Mehranfar S, Jalilpiran Y, Jafari A, Jayedi A, Shab-Bidar S, Speakman JR, et al. Validity of dietary assessment methods compared with doubly labeled water in children: a systematic review and meta-analysis. Obes Rev 2024;25(9):e13768. [doi: 10.1111/obr.13768] [Medline: 38783784]
6. Poppitt SD, Swann D, Black AE, Prentice AM. Assessment of selective under-reporting of food intake by both obese and non-obese women in a metabolic facility. Int J Obes Relat Metab Disord 1998;22(4):303-311. [doi: 10.1038/sj.ijo.0800584] [Medline: 9578234]
7. Garden L, Clark H, Whybrow S, Stubbs RJ. Is misreporting of dietary intake by weighed food records or 24-hour recalls food specific? Eur J Clin Nutr 2018;72(7):1026-1034 [FREE Full text] [doi: 10.1038/s41430-018-0199-6] [Medline: 29789710]
8. Chan V, Davies A, Wellard-Cole L, Lu S, Ng H, Tsoi L, et al. Using wearable cameras to assess foods and beverages omitted in 24 hour dietary recalls and a text entry food record app. Nutrients 2021;13(6):1806 [FREE Full text] [doi: 10.3390/nu13061806] [Medline: 34073378]
9. Gemming L, Ni Mhurchu C. Dietary under-reporting: what foods and which meals are typically under-reported? Eur J Clin Nutr 2016;70(5):640-641. [doi: 10.1038/ejcn.2015.204] [Medline: 26669571]
10. Simpson E, Bradley J, Poliakov I, Jackson D, Olivier P, Adamson AJ, et al. Iterative development of an online dietary recall tool: INTAKE24. Nutrients 2017;9(2):118 [FREE Full text] [doi: 10.3390/nu9020118] [Medline: 28208763]
11. Gemming L, Utter J, Ni Mhurchu C. Image-assisted dietary assessment: a systematic review of the evidence. J Acad Nutr Diet 2015;115(1):64-77. [doi: 10.1016/j.jand.2014.09.015] [Medline: 25441955]
12. Höchsmann C, Martin CK. Review of the validity and feasibility of image-assisted methods for dietary assessment. Int J Obes (Lond) 2020;44(12):2358-2371 [FREE Full text] [doi: 10.1038/s41366-020-00693-2] [Medline: 33033394]
13. Park Y, Dodd KW, Kipnis V, Thompson FE, Potischman N, Schoeller DA, et al. Comparison of self-reported dietary intakes from the automated self-administered 24-h recall, 4-d food records, and food-frequency questionnaires against recovery biomarkers. Am J Clin Nutr 2018;107(1):80-93 [FREE Full text] [doi: 10.1093/ajcn/nqx002] [Medline: 29381789]
14. Foster E, Lee C, Imamura F, Hollidge SE, Westgate KL, Venables MC, et al. Validity and reliability of an online self-report 24-h dietary recall method (Intake24): a doubly labelled water study and repeated-measures analysis. J Nutr Sci 2019;8:e29 [FREE Full text] [doi: 10.1017/jns.2019.20] [Medline: 31501691]
15. Skinner A, Toumpakari Z, Stone C, Johnson L. Future directions for integrative objective assessment of eating using wearable sensing technology. Front Nutr 2020;7:80 [FREE Full text] [doi: 10.3389/fnut.2020.00080] [Medline: 32714939]

XSL•FO
RenderX

16. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
17. Wen CKF, Schneider S, Stone AA, Spruijt-Metz D. Compliance with mobile ecological momentary assessment protocols in children and adolescents: a systematic review and meta-analysis. J Med Internet Res 2017;19(4):e132 [FREE Full text] [doi: 10.2196/jmir.6641] [Medline: 28446418]
18. Mason TB, Do B, Wang S, Dunton GF. Ecological momentary assessment of eating and dietary intake behaviors in children and adolescents: a systematic review of the literature. Appetite 2020;144:104465 [FREE Full text] [doi: 10.1016/j.appet.2019.104465] [Medline: 31541670]
19. Berkman ET, Giuliani NR, Pruitt AK. Comparison of text messaging and paper-and-pencil for ecological momentary assessment of food craving and intake. Appetite 2014;81:131-137 [FREE Full text] [doi: 10.1016/j.appet.2014.06.010] [Medline: 24930596]
20. Kraft R, Reichert M, Pryss R. Mobile crowdsensing in ecological momentary assessment mHealth studies: a systematic review and analysis. Sensors (Basel) 2024;24(2):472 [FREE Full text] [doi: 10.3390/s24020472] [Medline: 38257567]
21. Battaglia B, Lee L, Jia SS, Partridge SR, Allman-Farinelli M. The use of mobile-based Ecological Momentary Assessment (mEMA) methodology to assess dietary intake, food consumption behaviours and context in young people: a systematic review. Healthcare (Basel) 2022;10(7):1329 [FREE Full text] [doi: 10.3390/healthcare10071329] [Medline: 35885855]
22. Liao Y, Skelton K, Dunton G, Bruening M. A systematic review of methods and procedures used in ecological momentary assessments of diet and physical activity research in youth: an adapted STROBE Checklist for Reporting EMA Studies (CREMAS). J Med Internet Res 2016;18(6):e151 [FREE Full text] [doi: 10.2196/jmir.4954] [Medline: 27328833]
23. Schembre SM, Liao Y, O'Connor SG, Hingle MD, Shen S, Hamoy KG, et al. Mobile ecological momentary diet assessment methods for behavioral research: systematic review. JMIR Mhealth Uhealth 2018;6(11):e11170 [FREE Full text] [doi: 10.2196/11170] [Medline: 30459148]
24. Intille S, Haynes C, Maniar D, Ponnada A, Manjourides J. µEMA: microinteraction-based Ecological Momentary Assessment (EMA) using a smartwatch. Proc ACM Int Conf Ubiquitous Comput 2016;2016:1124-1128 [FREE Full text] [doi: 10.1145/2971648.2971717] [Medline: 30238088]
25. Beukenhorst AL, Howells K, Cook L, McBeth J, O'Neill TW, Parkes MJ, et al. Engagement and participant experiences with consumer smartwatches for health research: longitudinal, observational feasibility study. JMIR Mhealth Uhealth 2020;8(1):e14368 [FREE Full text] [doi: 10.2196/14368] [Medline: 32012078]
26. Heron KE, Everhart RS, McHale SM, Smyth JM. Using mobile-technology-based Ecological Momentary Assessment (EMA) methods with youth: a systematic review and recommendations. J Pediatr Psychol 2017;42(10):1087-1107. [doi: 10.1093/jpepsy/jsx078] [Medline: 28475765]
27. To ban or not to ban? UNESCO. URL: https://www.unesco.org/en/articles/smartphones-school-only-when-they-clearly-support-learning [accessed 2025-11-04]
28. Dawood S. Digital divide and poverty eradication in the rural region of Northern Peninsular Malaysia. Indones J Geogr 2019;51(2):172.
29. Maugeri A, Barchitta M. A systematic review of ecological momentary assessment of diet: implications and perspectives for nutritional epidemiology. Nutrients 2019;11(11):2696 [FREE Full text] [doi: 10.3390/nu11112696] [Medline: 31703374]
30. Brady SM, Salway R, Mariapun J, Millard L, Ramadas A, Rizal H, et al. Accelerometer-measured 24-hour movement behaviours over 7 days in Malaysian children and adolescents: a cross-sectional study. PLoS One 2024;19(2):e0297102 [FREE Full text] [doi: 10.1371/journal.pone.0297102] [Medline: 38377079]
31. Stone C, Adams S, Wootton RE, Skinner A. Smartwatch-based ecological momentary assessment for high-temporal-density, longitudinal measurement of alcohol use (AlcoWatch): feasibility evaluation. JMIR Form Res 2025;9:e63184 [FREE Full text] [doi: 10.2196/63184] [Medline: 40131326]
32. Lawlor DA, Lewcock M, Rena-Jones L, Rollings C, Yip V, Smith D, ALSPAC Executive. The second generation of the Avon Longitudinal Study of Parents and Children (ALSPAC-G2): a cohort profile. Wellcome Open Res 2019;4:36 [FREE Full text] [doi: 10.12688/wellcomeopenres.15087.2] [Medline: 31984238]
33. Mekruksavanich S, Jantawong P, Jitpattanakul A. Smartwatch-based eating detection and cutlery classification using a deep residual network with squeeze-and-excitation module. Washington DC: IEEE; 2022 Presented at: 45th International Conference on Telecommunications and Signal Processing (TSP); 2022 July 13-15; Prague, Czech Republic. [doi: 10.1109/tsp55681.2022.9851333]
34. Stankoski S, Jordan M, Gjoreski H, Luštrek M. Smartwatch-based eating detection: data selection for machine learning from imbalanced data with imperfect labels. Sensors (Basel) 2021;21(5):1902 [FREE Full text] [doi: 10.3390/s21051902] [Medline: 33803121]
35. Ibacache F, Northstone K, Zou M, Johnson L. Investigating eating architecture and the impact of the precision of recorded eating time: a cross-sectional study. Am J Clin Nutr 2025;121(3):685-694 [FREE Full text] [doi: 10.1016/j.ajcnut.2025.01.012] [Medline: 39805560]
36. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2024.

XSL•FO
RenderX

37. JGIBristol / seaco_smartwatch_paper. Github. 2025. URL: https://github.com/JGIBristol/seaco_smartwatch_paper [accessed 2026-01-06]
38. How to Collaborate with SEACO. MONASH University. 2022. URL: https://www.monash.edu.my/seaco/research-and-training/how-to-collaborate-with-seaco [accessed 2025-11-18]
39. Ponnada A, Haynes C, Maniar D, Manjourides J, Intille S. Microinteraction ecological momentary assessment response rates: effect of microinteractions or the smartwatch? Proc ACM Interact Mob Wearable Ubiquitous Technol 2017;1(3):92 [FREE Full text] [doi: 10.1145/3130957] [Medline: 30198012]
40. Ziesemer K, König LM, Boushey CJ, Villinger K, Wahl DR, Butscher S, et al. Occurrence of and reasons for "Missing Events" in Mobile dietary assessments: results from three event-based ecological momentary assessment studies. JMIR Mhealth Uhealth 2020;8(10):e15430 [FREE Full text] [doi: 10.2196/15430] [Medline: 33052123]
41. Ashurst J, van Woerden I, Dunton G, Todd M, Ohri-Vachaspati P, Swan P, et al. The association among emotions and food choices in first-year college students using mobile-ecological momentary assessments. BMC Public Health 2018;18(1):573 [FREE Full text] [doi: 10.1186/s12889-018-5447-0] [Medline: 29716572]
42. Ramadas A, Rizal H, Rajakumar S, Mariapun J, Yasin MS, Armstrong MEG, et al. Dietary intake, obesity, and metabolic risk factors among children and adolescents in the SEACO-CH20 cross-sectional study. Sci Rep 2024;14(1):11265 [FREE Full text] [doi: 10.1038/s41598-024-61090-7] [Medline: 38760446]
43. van Berkel N, Goncalves J, Hosio S, Kostakos V. Gamification of mobile experience sampling improves data quality and quantity. Proc ACM Interact Mob Wearable Ubiquitous Technol 2017;1(3):1-21. [doi: 10.1145/3130972]
44. School sessions to cease rotation system starting today: Radzi. The Vibes.com. 2022. URL: https://www.thevibes.com/articles/education/58838/school-sessions-to-cease-rotation-system-starting-today-radzi [accessed 2023-02-24]
45. Global Smartwatch Shipments in 2024: Market Declines for First Time, China Leads for First Time. Counterpoint Technology Market Research. 2025. URL: https://www.counterpointresearch.com/insight/global-smartwatch-market-in-2024 [accessed 2025-04-01]
46. South East Asia Community Observatory. Monash University Malaysia. URL: https://www.monash.edu.my/seaco [accessed 2026-01-18]

Abbreviations

CREMAS: Checklist for Reporting Ecological Momentary Assessment Studies
EMA: ecological momentary assessment
mEMA: mobile ecological momentary assessment
NCD: noncommunicable disease
OR: odds ratio
SEACO: South-East Asian Community Observatory
SEACO-CH20: South-East Asian Community Observatory Child Health 2020
µEMA: micro-interaction ecological momentary assessment

Edited by A Mavragani; submitted 01.May.2025; peer-reviewed by C Wang, P Delir Haghighi; comments to author 17.Aug.2025; revised version received 18.Nov.2025; accepted 30.Dec.2025; published 06.Feb.2026.

Please cite as:
Lane R, Millard LAC, Salway R, Stone CJ, Skinner AL, Brady SM, Mariapun J, Rajakumar S, Ramadas A, Rizal H, Johnson L, Su TT, Armstrong MEG
The Feasibility of Smartwatch Micro–Ecological Momentary Assessment for Tracking Eating Patterns of Malaysian Children and Adolescents in the South-East Asian Community Observatory Child Health Update 2020: Cross-Sectional Study
J Med Internet Res 2026;28:e73435
URL: https://www.jmir.org/2026/1/e73435
doi:10.2196/73435
PMID:

©Richard Lane, Louise A C Millard, Ruth Salway, Chris J Stone, Andy L Skinner, Sophia M Brady, Jeevitha Mariapun, Sutha Rajakumar, Amutha Ramadas, Hussein Rizal, Laura Johnson, Tin Tin Su, Miranda Elaine Glynis Armstrong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Integrating a Large Language Model to Streamline Nursing Handover Documentation Across Multiple Hospitals in Taiwan: Development and Implementation Study

Ray-Jade Chen[1,2,3*], MD; Mai-Szu Wu[4,5,6*], MD; Lung-Wen Tsai[7,8,9], PhD; Shy-Shin Chang[10,11], MD, PhD; Shu-Tai Shen Hsiao[12,13], PhD; Yu-Sheng Lo[14,15], PhD

[1]Department of Surgery, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
[2]College of Medicine, Professional Master Program in Artificial Intelligence in Medicine, Taipei Medical University, Taipei, Taiwan
[3]Department of Surgery, Division of Pediatric Surgery, Taipei Medical University Hospital, Taipei, Taiwan
[4]Department of Internal Medicine, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
[5]Division of Nephrology, Department of Internal Medicine, Shuang Ho Hospital, New Taipei City, Taiwan
[6]Taipei Medical University Research Center of Urology and Kidney (TMU-RCUK), Taipei Medical University, Taipei, Taiwan
[7]Graduate Institute of Data Science, College of Management, Taipei Medical University, New Taipei City, Taiwan
[8]Department of Information Technology Office, Taipei Medical University Hospital, Taipei, Taiwan
[9]Department of Medicine Research, Taipei Medical University Hospital, Taipei, Taiwan
[10]Department of Family Medicine, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
[11]Department of Family Medicine, Taipei Medical University Hospital, Taipei, Taiwan
[12]Office of Superintendent, Taipei Medical University Hospital, Taipei, Taiwan
[13]School of Nursing, Taipei Medical University, Taipei, Taiwan
[14]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, New Taipei City, Taiwan
[15]Center of Digital Innovation and Development, Taipei Medical University, New Taipei City, Taiwan
[*]these authors contributed equally

Corresponding Author:
Yu-Sheng Lo, PhD
Graduate Institute of Biomedical Informatics
College of Medical Science and Technology
Taipei Medical University
No. 301, Yuantong Road, Zhonghe District
New Taipei City, 235
Taiwan
Phone: 886 6620 2589
Email: Loyusen@tmu.edu.tw

Abstract

Background: The global nursing shortage, exacerbated by heavy workloads and high turnover rates associated with the COVID-19 pandemic, continues to undermine care quality and nurse well-being. Although digital health technologies have enhanced coordination, improved communication, and reduced clinical errors in nursing practice, they have also increased nurses' documentation burden. Advances in large language models (LLMs) and other generative artificial intelligence (GenAI) tools facilitate the generation of accurate reports from electronic medical records (EMRs), thereby streamlining documentation workflows, saving time, and reducing nurses' workloads. Accordingly, integrating LLMs into electronic nursing documentation systems warrants further exploration.

Objective: This study examines the integration of an LLM into an in-house nursing information system (NIS) implemented across 3 hospitals in Taiwan to reduce the time and effort required for nursing handover documentation and to preliminarily assess the operational and economic implications of GenAI-assisted workflows.

Methods: A multidisciplinary team of nursing specialists and information technology experts at Taipei Medical University (TMU) restructured the organization's existing nursing handover documentation process to facilitate interaction with the LLM. The team also developed prompt-based interfaces to automatically generate section-specific content for the nursing handover

XSL•FO
RenderX

document. The LLM-integrated NIS was subsequently deployed across 3 hospitals in Taiwan: Taipei Medical University Hospital (TMUH), Wan Fang Hospital (WFH), and Shuang Ho Hospital (SHH). We then extracted and analyzed NIS log data to compare documentation times before and after LLM implementation, thereby quantifying time savings.

Results: Integration of the LLM into nursing handover documentation was associated with shorter per-patient documentation time in routine clinical use across TMUH, WFH, and SHH. Based on preintegration NIS logs (September 2024), the average handover document completion time per patient ranged from 3.45 (SD 3.82) to 4.32 (SD 4.48) minutes across hospitals and shifts, providing a preliminary baseline for subsequent comparisons. In postintegration NIS logs (October-December 2024), the overall handover document completion time per patient (mean) was substantially lower, ranging from 1.17 (SD 1.86) to 2.54 (SD 2.82) minutes across hospitals and shifts. Using monthly patient volume to estimate time savings, 113-273, 160-314, and 198-391 hours were saved per month at TMUH, WFH, and SHH, respectively, corresponding to aggregate savings of 474-981 hours per month across hospitals during the study period.

Conclusions: We integrated an LLM into an NIS to generate nursing handover documents without altering existing workflows. Across 3 hospitals within TMU's health system, GenAI assistance was associated with shorter documentation time and a positive net labor value from October to December 2024. Prompts were constrained, and nurse verification was required to mitigate hallucinations. Future work will enhance logging to capture reliability and editing metrics, compare LLM-generated drafts with nurse-finalized notes to inform prompt refinement, and assess generalizability to other documentation workflows.

(J Med Internet Res 2026;28:e81604) doi:10.2196/81604

KEYWORDS

nursing shortage; large language model; generative artificial intelligence; artificial intelligence; nursing information system; electronic medical records

Introduction

The global nursing shortage has been exacerbated by several factors and remains a critical challenge for health care systems, threatening health care quality and access [1,2]. The COVID-19 pandemic has increased nursing workloads, resulting in increased patient acuity, greater infection control demands, and longer working hours—all of which have contributed to nurses experiencing physical exhaustion, moral distress, and mental health challenges [3,4]. Moreover, the pandemic worsened the nursing shortage, with 100,000 nurses leaving the US health care workforce during the pandemic and an additional 900,000 planning to exit by 2027 [5]. This shortage has compromised care quality and increased burnout and turnover risks. According to the World Health Organization (WHO), although the global nursing workforce grew from 27.9 million in 2018 to 29.8 million in 2023, a shortage of 5.8 million nurses persists, which is projected to decrease to 4.1 million by 2030 [6].

Although Taiwan experienced early success in controlling the COVID-19 pandemic, the pandemic exposed structural problems within the country's nursing sector. Nursing shortages, particularly in intensive care and emergency departments, persisted during the late-pandemic and postpandemic periods. High turnover rates and uneven workloads contributed to increased work-related stress [7]. Furthermore, interviews with public health nurses revealed that frequent changes in temporary assignments and information systems during the pandemic exacerbated fatigue and created additional challenges for frontline nurses [8].

The ongoing digital transformation in health care has rendered digital technologies essential to nursing practice. Studies have indicated that the adoption of digital tools by health care professionals improves care quality, enhances patient participation in their own care [9,10], fosters better coordination among health care professionals [11], reduces medication errors and adverse events [12,13], and improves documentation accuracy [14,15]. However, this shift has led to more time being spent on indirect tasks, particularly activities related to maintaining electronic medical records (EMRs), including data collection, entry, and confirmation [16]. Documentation accounts for a considerable portion of nurses' work time, and this burden is exacerbated by high patient turnover [17] and the use of multiple recording systems [18,19]. A systematic review revealed that nurses spend between 13% and 31% of their work time on documentation [20]; thus, considering nurses' heavy workloads, reducing documentation time has become a critical issue in nursing.

Large language models (LLMs), a category of artificial intelligence (AI) tools, perform deep learning with millions or billions of parameters on extensive data to generate human-like text [21]. LLM platforms, such as ChatGPT (OpenAI), Gemini (Google LLC), and Llama (Meta Platforms, Inc.) [22,23], are applied in various domains, including text generation, question answering, translation, and creative content creation. Moreover, large multimodal models have been developed to process several types of data, such as text, images, video, and audio, thereby broadening the application potential of generative AI (GenAI) tools [24,25]. In health care, LLMs and other GenAI tools enable coherent text generation for tasks such as drug discovery, personalized medicine, radiology diagnosis, automated clinical documentation, medical research, and health information consultation [26-34]. These tools are driving advancements in diverse health care applications; however, ensuring the accuracy, reliability, and fairness of the generated results remains critical [35].

AI applications in nursing are rapidly expanding, and AI tools have the potential to become essential parts of nursing practice [36,37]. By integrating clinical data with language generation tools, GenAI systems can improve documentation quality,

XSL•FO
RenderX

support clinical decision-making, and reduce administrative workload. Additionally, these systems can considerably shorten documentation time and enhance handover efficiency by rapidly producing consistent records, thereby allowing nurses to devote more time to patient care and ultimately improving the quality and safety of health care delivery [38,39].

To explore the potential of incorporating LLMs into electronic nursing documentation systems, this study integrated an LLM into an in-house nursing information system (NIS) to reduce the time and effort required for nursing handover documentation. We evaluated the LLM-integrated NIS by comparing the time spent by nurses on nursing handover documentation before and after integration. We additionally conducted a preliminary cost-effectiveness assessment and quantified the revision rate of GenAI-generated drafts before final submission.

Methods

Settings

This study was conducted at Taipei Medical University (TMU), which operates 3 hospitals: Taipei Medical University Hospital (TMUH), Wan Fang Hospital (WFH), and Shuang Ho Hospital (SHH). The combined capacity of these hospitals is 2675 acute beds (TMUH: 735; WFH: 761; SHH: 1179), and they handle a total of over 400,000 outpatient visits per month. Owing to nursing shortages, operational capacity decreased to 2195 acute beds as of May 2025 (TMUH: 608; WFH: 576; SHH: 1011). Nursing workflows are supported by an in-house NIS managed by the TMU Center of Digital Innovation and Development.

Google Gemini

We selected a HIPAA (Health Insurance Portability and Accountability Act)-compliant enterprise instance of Gemini 1.5 Pro (Google LLC) for integration into the in-house NIS. Gemini 1.5 Pro is a mid-sized multimodal LLM optimized for a broad range of reasoning tasks and can generate text output in response to diverse inputs, including text, image, video, and audio. Moreover, it offers capabilities for output streaming, conversational interactions, and system-level instruction handling. The model can be used to conduct various language-related tasks, such as text generation, text editing, code generation, problem-solving, recommendation generation, information extraction, and data generation [40].

Conventional Workflow in Nursing Handover Documentation

In hospitals across Taiwan, nursing schedules are typically structured around 3 daily shifts, and nurses are required to complete handover documentation twice per day. The nursing handover document follows the WHO-endorsed ISBAR (Identify, Situation, Background, Assessment, Recommendation) framework structured into 5 key components [41].

Figure 1 illustrates the conventional workflow for nursing handover documentation with the in-house NIS. Initially, the NIS displays a list of patients requiring handover documentation. Subsequently, nurses collect and analyze individualized patient information using various EMR systems. They then sequentially write a nursing handover document for each patient. Finally, the documented handover information is manually summarized to formulate the corresponding nursing recommendations for each patient.

In general, nurses prepare handover documents for 5-7 patients per shift, and senior nurses estimated an average completion time of approximately 5 minutes per patient. Accordingly, each nurse dedicates approximately 1 hour to complete patient handover tasks; this time frame includes both the documentation process and the subsequent verification and explanation of patient details to the receiving nurse.

XSL•FO
RenderX

Figure 1. Conventional workflow for nursing handover documentation using the in-house nursing information system (NIS) at Taipei Medical University. EMR: electronic medical record.

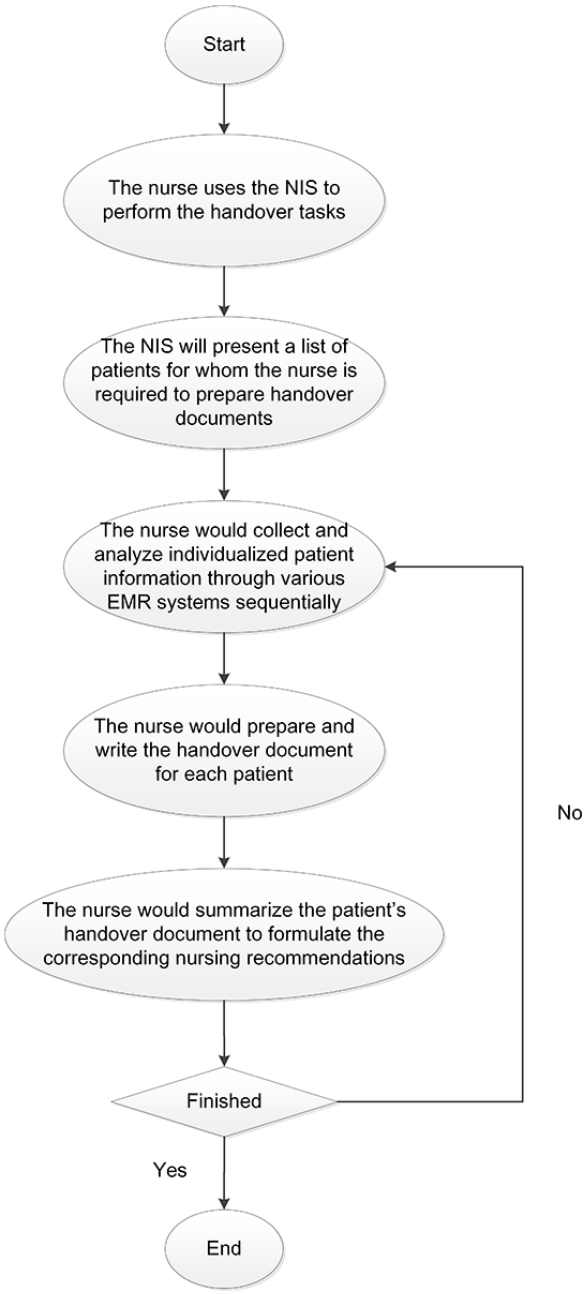

Prompt Design for Nursing Handover Documentation

A multidisciplinary team of nursing specialists and information technology professionals collaboratively outlined the key components of the nursing handover document, as defined in the existing EMR systems. Each component was divided into several sections, which were systematically evaluated to assess the extracted content and identify potential problems. The most critical problem identified was related to the manual collection, verification, transcription, and integration of patient data from various EMR systems. Thus, we reorganized the structure of the nursing handover document. Tables 1 and 2, respectively, present the sections of the original and revised (for the LLM-integrated NIS) nursing handover documents across 5 key components according to the ISBAR framework.

We developed section-specific prompts to support nurses' use of the LLM, targeting 4 tasks: (1) extracting information from the provided input, (2) summarizing input text, (3) interpreting or translating text (eg, English to Mandarin), and (4) formatting content based on the input. When executed, these prompts generate draft content for the corresponding handover sections. Prompt development followed an iterative, human-in-the-loop process: draft prompts were repeatedly tested using deidentified EMRs, and outputs were reviewed by nursing professionals for clinical relevance, completeness, and adherence to the intended structure. Feedback-informed refinements were made to prompt wording, section constraints, and output formatting until the drafts were suitable for routine handover use. All prompts instructed the LLM to use only the provided input and to avoid introducing new values, assumptions, or inferences. No patient

XSL•FO
RenderX

identifiers or personally identifiable information was transmitted to the LLM.

For example, in the laboratory results section, the NIS queried the laboratory information system to retrieve each patient's test results and corresponding reference ranges from the preceding 48 hours. A tailored prompt was then sent via the LLM application programming interface (API) to identify abnormal results by comparing values with reference ranges and to present the extracted abnormalities in a structured table. Table 3 summarizes the tailored prompts by section and the corresponding EMR source systems, including the inpatient system, medical documentation system, NIS, laboratory information system, examination information system, radiology information system, patient schedule system, operating room system, and blood bank system.

Table 1. The original nursing handover document.

Components	Sections
Identity	Introduction
Situation	Vital signs, blood glucose, laboratory test results, culture results, diagnostic examination reports, radiology reports, pending diagnostic examination tests or radiology tests, surgical records, and blood products
Background	Admission reason, medical history
Assessment	Wound records, consultation records
Recommendations	A blank field for manual data entry
Others	A blank field for manual data entry

Table 2. The revised (for the large language model–integrated nursing information system) nursing handover document.

Components	Sections
Identity	Introduction
Handover records[a]	Vital signs, blood glucose, laboratory test results, culture results, diagnostic examination reports, radiology reports, pending diagnostic examination tests or radiology tests, surgical records, blood products, wound records, consultation records, and recommendations
Background	Admission reason, medical history
Others	A blank field for manual data entry

[a]Handover records include records related to the situation (vital signs, blood glucose, laboratory test results, culture results, diagnostic examination reports, radiology reports, pending diagnostic examination tests or radiology tests, surgical records, and blood products), assessment (wound records and consultation records), and recommendations.

XSL•FO
RenderX

Table 3. Tailored prompts by handover section and corresponding electronic medical record source systems.

Components and sections	Electronic medical record source systems	Tailored prompts
Identity		
Introduction	Inpatient system	The nursing information system already provides information (manual data entry is not necessary).
Background		
Admission reason and medical history	Inpatient system and medical documentation system	Summarize admission details into a concise paragraph. Highlight the primary reason for admission and include key clinical details.
Handover records		
Vital signs	Nursing information system	Extract and structure the following vital signs data from the past 8 hours from the provided data: date, time, temperature, pulse, respiration, blood pressure, oxygen level, and Glasgow Coma Scale values. Maintain chronological order.
Blood glucose	Laboratory information system	Extract and structure blood glucose readings from the past 8 hours, including date, time, and measurement value. Organize the data in reverse chronological order.
Laboratory test results	Laboratory information system	Extract the following laboratory results from the past 48 hours: test item, value, and reference range. Mark abnormal values (H for high and L for low), and leave normal values unmarked. Organize the results chronologically and return them in a structured JSON format.
Culture results	Laboratory information system	Extract and structure key information from culture data within 6 days. For each data sample, identify and output the following fields in a structured format: collection date, specimen type, culture results, and status (completed or pending).
Diagnostic examination reports	Examination information system	Summarize diagnostic findings from the past 24 hours, emphasizing significant abnormal results in Standard Mandarin. If such findings are unavailable, state “No examination reports available.”
Radiology reports	Radiology information system	Summarize key radiological findings from the past 24 hours, highlighting clinically relevant abnormalities in Standard Mandarin. If such findings are unavailable, state “No radiology reports available.”
Pending diagnostic examination tests or radiology tests	Patient schedule system	List upcoming diagnostic and radiology tests in the following format: procedure name and scheduled date.
Surgical records	Operating room system	Extract the following information from surgical records from the past 48 hours: procedure name, start time, end time, blood loss, and anesthesia method. If no records are available, state “No surgical records available.”
Blood products	Blood bank system	Extract the following information on blood product orders within 48 hours: order number, blood type and Rh[a] factor, order date, product name, quantity ordered/used, and transfusion date. If such information is unavailable, state “No blood product records available.”
Wound records	Nursing information system	Extract the following wound assessment data: type, location, assessment date/time, and dimensions. If these data are unavailable, state “No wound assessment records available.”
Consultation records	Medical documentation system	Summarize consultation findings and recommendations in Standard Mandarin. If such information is unavailable, state “No consultation records available.”
Recommendations	Blank field	Generate comprehensive nursing care recommendations based on the provided data. Present recommendations as bullet points in only Standard Mandarin. Include information on priority interventions, monitoring requirements, preventive measures, and areas for special attention.
Others		
Others	Blank field for manual data entry	N/A[b]

[a]Rh factor: Rhesus factor.

[b]N/A: not applicable.

XSL•FO
RenderX

Interactions Between the NIS and the LLM in the Designed Workflow for Nursing Handover Documentation

Google Gemini 1.5 Pro offers an API that provides access to its LLM, which was integrated into the in-house NIS to automatically generate nursing handover documents.

Figure 2 illustrates the integrated workflow of the LLM-integrated NIS. For each section of the nursing handover document, the NIS sends a request to the relevant EMR system to retrieve patient data through asynchronous calls. The retrieved data and the corresponding tailored prompt are then used as input to call the LLM API. Upon receiving the LLM response, the NIS organizes the content and formats it according to the appropriate style for each section. The final output generated by the NIS is a structured and comprehensive handover document that can be manually reviewed and validated by nursing staff.

Figure 2. Integrated workflow for the large language model (LLM)–integrated nursing information system (NIS). API: application programming interface; EMR: electronic medical record.

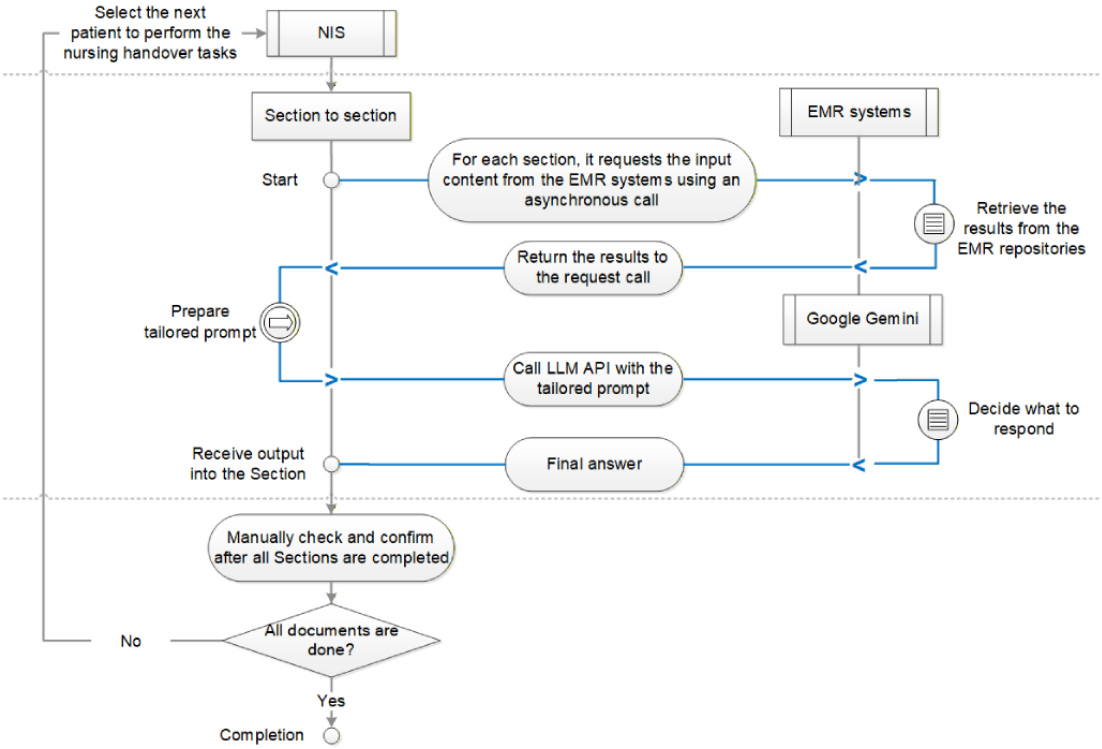

Log Data Collection and Analysis of the Nursing Handover Document in the NIS

In this study, log data from the LLM-integrated NIS, implemented in October 2024, were used to conduct a preliminary pre-post comparison of nursing handover document completion times. Handover document completion time was defined as the interval from when a nurse selected a patient by clicking the handover button, through manual entry before GenAI integration or automatic draft generation after integration, to final confirmation, when the nurse completed review and revision and clicked the “Confirm” button for that patient.

To support subsequent analyses and system optimization, step-level log data were collected across the full handover documentation workflow. We retrospectively extracted NIS logs from September 2024 (preintegration) and from October-December 2024 (postintegration) for the 3 TMU-managed hospitals. To reduce the influence of extreme values, log data with handover document completion times exceeding 20 minutes were excluded from the analyses. In addition, we logged both the automatically GenAI-generated handover draft and the final nurse-confirmed version to preliminarily quantify the document revision rate. As nurse identifiers were unavailable, we used hospital × shift as the clustering unit and applied robust (sandwich) SEs to account for within-cluster correlation.

Ethical Considerations

This retrospective study of deidentified NIS logs was approved by the TMU-Joint Institutional Review Board (TMU-JIRB; approval number N202506094). All data were deidentified to protect participant privacy and confidentiality, and no personally identifiable information was collected or disclosed. The TMU-JIRB waived the requirement for informed consent. All images included in the manuscript were reviewed to ensure that no individual participants or users could be identified. The hospitals also conduct routine qualitative and quantitative medical record audits, providing organizational accountability and quality assurance for AI-assisted documentation.

Results

Descriptive Statistics of Nursing-Related Factors for the Examined Hospitals

From September 1 to December 31, 2024, we collected NIS log data from the 3 hospitals during daily nursing handovers across

XSL•FO
RenderX

3 shifts (day, night, and overnight). At TMU, nursing handover documents must be prepared for most inpatients; however, certain patient categories are exempt from this requirement. These categories include patients undergoing postoperative observation, day hospitalization, and 1-day chemotherapy at TMUH; nursing home residents and day-care patients at WFH; and day-care, psychiatric ward, nursing home, and postpartum home care patients at SHH.

During the 4-month study period, the number of nurses working across the 3 daily shifts ranged from 501 to 524 at TMUH, 468 to 489 at WFH, and 862 to 892 at SHH. Monthly inpatient case volumes ranged from 7477 to 7733 at TMUH, 5406 to 5977 at WFH, and 8872 to 9814 at SHH. The mean monthly length of stay was 4.7-4.9 days at TMUH, 4.7-5.1 days at WFH, and 4.9-5.0 days at SHH. Across all shifts, total monthly shift counts ranged from 35,527 to 37,892 at TMUH, 27,500 to 30,835 at WFH, and 42,535 to 44,945 at SHH. Multimedia Appendix 1 summarizes descriptive statistics for the 3 hospitals included in this study.

Comparison of the Original and GenAI-Assisted Nursing Handover Documentation Functions in the NIS

In accordance with the workflow shown in Figure 1, the LLM-integrated NIS initially displays a table listing patients, enabling nurses to generate handover documents sequentially by clicking a selection checkbox (Figure 3). Each row presents key patient attributes, including the patient's name, medical record number, bed number, birth date and age, admission date and length of stay, department, and attending physician. The table also includes a handover button to initiate the nursing handover process.

Figure 4 depicts the nursing handover documentation interface in the original NIS, displayed when a nurse selects a patient and initiates the handover process. Nurses complete 5 handover sections: Introduction (auto-populated; no manual entry required), Situation, Background, Assessment, and Recommendation and Other information.

Figure 5 depicts the revised nursing handover document in the LLM-integrated NIS. In contrast to the original NIS, selecting a patient for handover triggers an API call to the LLM. After the response is returned, the LLM-integrated NIS automatically populates draft content for each handover section. Nurses then review, revise, and confirm the draft. Once finalized, the nurse clicks "Confirm" to submit the handover document for that patient (Figure 6).

Figure 3. Screenshot of the patient list interface used by nurses to perform handover tasks in the nursing information system.

XSL•FO
RenderX

Figure 4. Screenshot of the Identify, Situation, Background, Assessment, Recommendation interface used for nursing handover tasks in the original nursing information system.

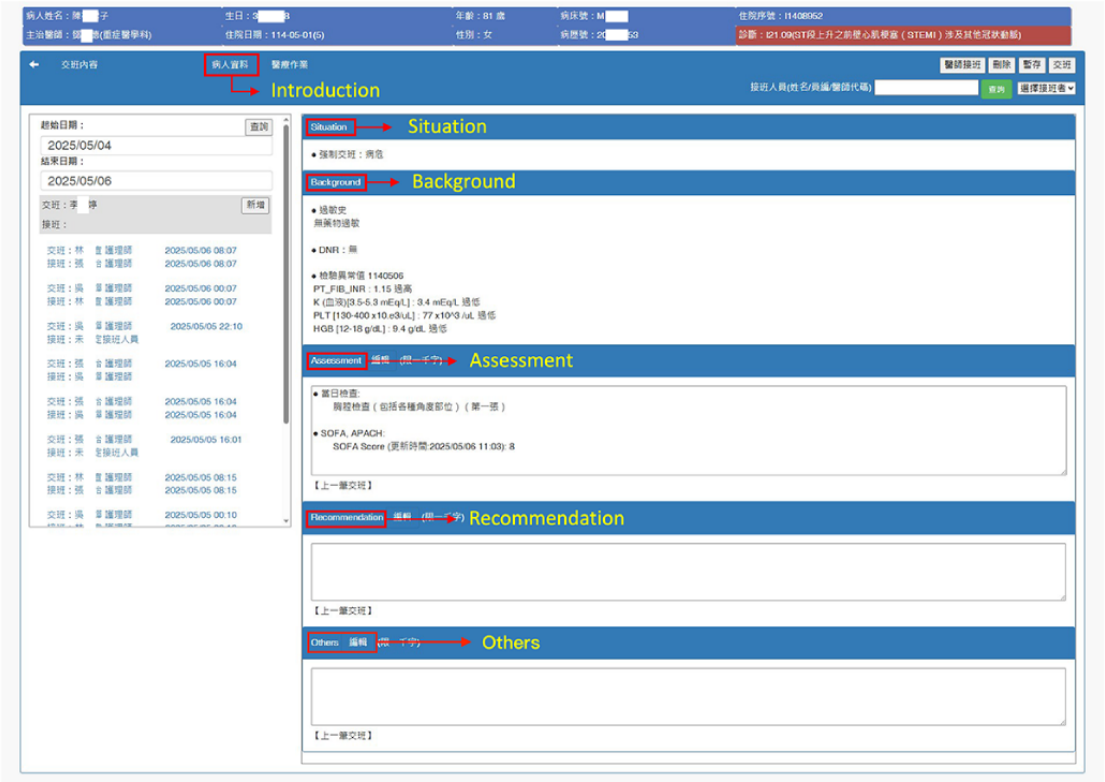

Figure 5. Screenshot of the revised nursing handover document in the large language model–integrated nursing information system.

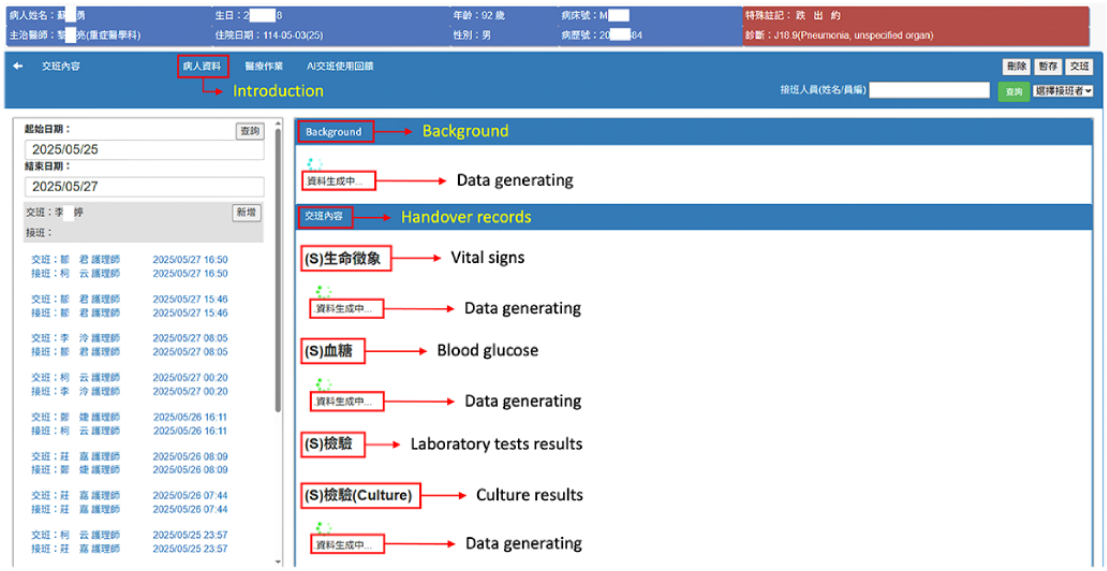

XSL•FO
RenderX

Figure 6. Screenshot of a complete nursing handover document automatically generated by the large language model–integrated nursing information system.

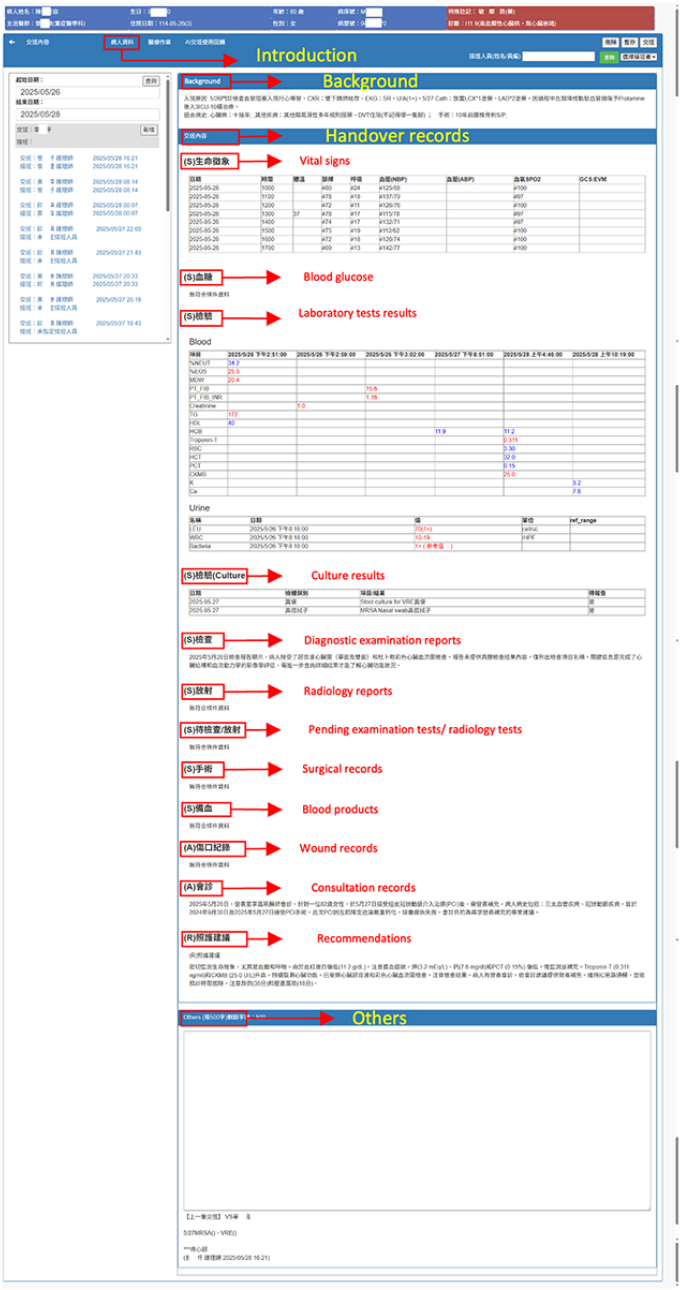

Analysis of NIS Log Data for Pre-LLM Integration

Table 4 presents detailed NIS log data from September 2024 for TMUH, WFH, and SHH, including the number of shifts and the mean (SD) of handover document completion time per patient for each shift. Before LLM integration, nurses often had to pause documentation to gather additional patient information or respond to unplanned clinical events (eg, bedside care, family inquiries, or urgent clinical duties). Accordingly, the NIS included a draft-save function that allowed handover content to be temporarily stored and completed later.

A total of 110,457 shift-level handover logs were analyzed (TMUH: 37,087; WFH: 30,835; and SHH: 42,535), representing large-scale real-world clinical operations. Across hospitals and shifts, the mean handover document completion time per patient ranged from 3.45 (SD 3.82) to 4.32 (SD 4.48) minutes. At TMUH, mean completion times were 4.10 (SD 4.55) minutes (day shift), 4.32 (SD 4.48) minutes (night shift), and 3.45 (SD 3.82) minutes (overnight shift). At WFH, the corresponding values were 4.26 (SD 5.08), 3.70 (SD 4.47), and 4.10 (SD 5.29) minutes, respectively, whereas at SHH they were 3.67 (SD 4.31), 4.09 (SD 4.25), and 3.81 (SD 4.23) minutes, respectively.

Although only 1 month of preintegration log data were available, these results provide a preliminary baseline of nursing handover documentation performance against which the impact of subsequent GenAI-assisted workflows can be quantitatively evaluated.

XSL•FO
RenderX

Table 4. Shift counts and mean document completion time, stratified by hospital and shift type before large language model integration (September 2024).

Hospital and shift	September 24	
	Number of shifts, n	Document completion time (minutes), mean (SD)
Taipei Medical University Hospital (n=37,087)		
Day	12,418	4.10 (4.55)
Night	12,203	4.32 (4.48)
Overnight	12,466	3.45 (3.82)
Wan Fang Hospital (n=30,835)		
Day	9379	4.26 (5.08)
Night	10,547	3.70 (4.47)
Overnight	10,909	4.10 (5.29)
Shuang Ho Hospital (n=42,535)		
Day	13,868	3.67 (4.31)
Night	14,466	4.09 (4.25)
Overnight	14,201	3.81 (4.23)

Analysis of NIS Log Data for LLM Responses

Table 5 summarizes shift-level log data from October to December 2024 for TMUH, WFH, and SHH, including the number of shifts, mean GenAI document generation time per patient, mean nurse manual (review-revision-confirmation) time per patient, and mean overall document completion time per patient. SHH had the highest shift volume (43,473 in October to 44,945 in December) and the largest monthly patient volume (8872-8989; Multimedia Appendix 1), with mean overall completion times ranging from 1.74 (SD 2.5) to 2.11 (SD 2.82) minutes. WFH recorded the lowest shift volume (27,500-28,730) and the shortest mean overall completion times (mean 1.17, SD 1.86 minutes to mean 1.78, SD 2.53 minutes) while managing 5746-5977 patients per month (Multimedia Appendix 1). TMUH showed intermediate shift volume (35,890-37,892) and patient volume (7477-7733; Multimedia Appendix 1), with mean overall completion times of 2.13 (SD 2.5) to 2.54 (SD 2.82) minutes.

We further conducted analyses using joint Wald tests to assess whether hospital, shift, and month were associated with statistically significant differences. As individual nurse identifiers were not available in the log data, clustering at the nurse level was not feasible. Therefore, we defined the clustering unit as hospital × shift, and the corresponding results are presented in Table 6.

In Table 6, joint Wald tests indicated that all 3 outcomes (GenAI time, manual time, and document completion time) differed significantly across hospitals and shifts (hospital: Wald χ^2_2=52.85, *P*<.001; shift: Wald χ^2_2=10.39, *P*=.006). By contrast, the effect of month was not statistically significant (Wald χ^2_2=5.74, *P*=.057).

Moreover, we used generalized estimating equations (GEEs) to account for within-cluster correlation arising from the clustered and longitudinal structure of the data. Separate GEE models were fitted for GenAI time, manual time, and document completion time. Each model included month, shift, and hospital as categorical covariates and used an exchangeable working correlation structure with robust SEs; thus, estimates represent population-averaged (system-level) effects rather than individual-level effects. Table 7 reports the population-averaged GEE results for GenAI time, manual time, and document completion time (β and SE). Relative to SHH, TMUH had longer times across all outcomes (GenAI time: β=3.05; manual time: β=20.35; document completion time: β=23.40; all *P*<.001), whereas WFH had shorter times (GenAI time: β=−3.47; manual time: β=−23.16; document completion time: β=−26.64; all *P*<.001). December (vs October) was associated with small increases across outcomes (*P*=.02), whereas November (vs October; *P*=.41) and shift (*P*=.44) were not statistically significant.

XSL•FO
RenderX

Table 5. Monthly shift-level patient volume and mean per-patient generative artificial intelligence, manual, and overall completion times across 3 Taipei Medical University hospitals (October-December 2024).

Hospital and shift	October 24				November 24				December 24			
	Number of shifts, n	Generative artificial intelligence time (seconds)[a], mean (SD)	Manual time (minutes)[b], mean (SD)	Document completion time (minutes)[c], mean (SD)	Number of shifts, n	Generative artificial intelligence time (seconds), mean (SD)	Manual time (minutes), mean (SD)	Document completion time (minutes), mean (SD)	Number of shifts, n	Generative artificial intelligence time (seconds), mean (SD)	Manual time (minutes), mean (SD)	Document completion time (minutes), mean (SD)
Taipei Medical University Hospital[d]												
Day	12,043	17.0 (20.0)	1.89 (2.23)	2.17 (2.56)	11,882	16.70 (19.60)	1.85 (2.18)	2.13 (2.50)	12,510	17.00 (19.60)	1.88 (2.17)	2.17 (2.50)
Night	11,928	19.6 (22.1)	2.18 (2.45)	2.50 (2.82)	11,900	18.70 (22.10)	2.08 (2.45)	2.40 (2.82)	12,740	19.80 (22.10)	2.20 (2.45)	2.54 (2.82)
Overnight	11,918	19.0 (21.6)	2.11 (2.40)	2.43 (2.76)	11,745	19.00 (21.60)	2.11 (2.39)	2.42 (2.75)	12,642	18.60 (20.80)	2.07 (2.31)	2.38 (2.66)
Wan Fang Hospital[e]												
Day	8291	12.1 (17.9)	1.34 (1.99)	1.54 (2.29)	8620	13.20 (19.50)	1.47 (2.17)	1.68 (2.50)	8905	13.90 (19.80)	1.55 (2.20)	1.78 (2.53)
Night	9612	12.5 (18.7)	1.39 (2.08)	1.60 (2.39)	10,035	12.50 (18.70)	1.38 (2.07)	1.59 (2.39)	10,190	12.50 (18.60)	1.39 (2.06)	1.60 (2.37)
Overnight	9597	9.2 (14.6)	1.02 (1.62)	1.17 (1.86)	9438	10.10 (15.40)	1.12 (1.72)	1.28 (1.97)	9635	9.60 (15.10)	1.07 (1.68)	1.23 (1.93)
Shuang Ho Hospital[f]												
Day	13,103	14.8 (20.2)	1.65 (2.24)	1.89 (2.58)	13,931	15.40 (20.70)	1.71 (2.30)	1.97 (2.65)	13,875	15.30 (20.90)	1.70 (2.33)	1.96 (2.67)
Night	15,317	15.7 (21.3)	1.74 (2.36)	2.01 (2.72)	15,268	16.10 (21.50)	1.79 (2.39)	2.06 (2.75)	16,040	16.50 (22.00)	1.84 (2.45)	2.11 (2.82)
Overnight	15,053	13.7 (19.6)	1.52 (2.18)	1.74 (2.50)	15,337	14.10 (20.60)	1.57 (2.29)	1.80 (2.64)	15,030	15.60 (22.40)	1.74 (2.49)	2.00 (2.86)

[a]Mean generative artificial intelligence document generation time per patient.

[b]Mean nurse manual (review-revision-confirmation) time per patient.

[c]Mean overall document completion time per patient.

[d]The total number of shifts for October, November, and December was 35,890, 35,527, and 37,892, respectively.

[e]The total number of shifts for October, November, and December was 27,500, 28,092, and 28,730, respectively.

[f]The total number of shifts for October, November, and December was 43,473, 44,536, and 44,945, respectively.

XSL•FO
RenderX

Table 6. Joint Wald tests for month, shift, and hospital effects on generative artificial intelligence time, manual time, and document completion time.

Outcome	Factor	Wald chi-square statistic (*df*)	*P* value
Generative artificial intelligence document generation time	Month	5.74 (2)	.06
Manual time[a]	Month	5.74 (2)	.06
Document completion time[b]	Month	5.74 (2)	.06
Generative artificial intelligence document generation time	Shift	10.39 (2)	.006
Manual time	Shift	10.39 (2)	.006
Document completion time	Shift	10.39 (2)	.006
Generative artificial intelligence document generation time	Hospital	52.85 (2)	<.001
Manual time	Hospital	52.85 (2)	<.001
Document completion time	Hospital	52.85 (2)	<.001

[a]Nurse manual (review-revision-confirmation) time.

[b]Overall document completion time per patient.

Table 7. Generalized estimating equation model coefficients for generative artificial intelligence document generation time, manual time, and document completion time.

Outcome	Factor	β (seconds)	95% CI	*P* value
Generative artificial intelligence document generation time	Intercept	15.09	14.26 to 15.92	<.001
Generative artificial intelligence document generation time	Hospital: TMUH[a] vs SHH[b]	3.05	1.80 to 4.31	<.001
Generative artificial intelligence document generation time	Hospital: WFH[c] vs SHH	–3.47	–4.72 to –2.23	<.001
Manual time[d]	Intercept	100.60	95.05 to 106.15	<.001
Manual time	Hospital: TMUH vs SHH	20.35	11.97 to 28.72	<.001
Manual time	Hospital: WFH vs SHH	–23.16	–31.45 to –14.88	<.001
Document completion time[e]	Intercept	115.69	109.31 to 122.07	<.001
Document completion time	Hospital: TMUH vs SHH	23.40	13.77 to 33.03	<.001
Document completion time	Hospital: WFH vs SHH	–26.64	–36.16 to –17.11	<.001
Generative artificial intelligence document generation time	Month: December vs October	.56	0.09 to 1.04	.02
Manual time	Month: December vs October	3.75	0.58 to 6.92	.02
Document completion time	Month: December vs October	4.31	0.66 to 7.95	.02
Generative artificial intelligence document generation time	Shift: night vs day	1.05	–0.24 to 2.34	.11
Manual time	Shift: night vs day	6.99	–1.61 to 15.60	.11
Document completion time	Shift: night vs day	8.04	–1.85 to 17.93	.11
Generative artificial intelligence document generation time	Month: November vs October	.16	–0.23 to 0.55	.41
Manual time	Month: November vs October	1.09	–1.50 to 3.69	.41
Document completion time	Month: November vs October	1.26	–1.73 to 4.24	.41
Generative artificial intelligence document generation time	Shift: overnight vs day	–.69	–2.45 to 1.07	.44
Manual time	Shift: overnight vs day	–4.59	–16.34 to 7.15	.44
Document completion time	Shift: overnight vs day	–5.28	–18.79 to 8.22	.44

[a]TMUH: Taipei Medical University Hospital.

[b]SHH: Shuang Ho Hospital.

[c]WFH: Wan Fang Hospital.

[d]Nurse manual (review-revision-confirmation) time.

[e]Overall document completion time per patient.

XSL•FO
RenderX

Comparison of NIS Log Data Before and After LLM Integration

Using NIS logs from the preintegration period (September 2024) and the postintegration period (October-December 2024), we compared handover document completion time. Table 8 reports population-averaged GEE estimates for document completion time (seconds). Document completion time was significantly longer preintegration than postintegration (β=63.48 seconds; 95% CI 53.81-73.16; *P*<.001), and the overall period effect was significant (χ^2_1=165.24; *P*<.001). Relative to SHH, completion time was longer at TMUH (β=19.10 seconds; 95% CI 10.34-27.85; *P*<.001) and shorter at WFH (β=−25.67 seconds; 95% CI −34.42 to −16.92; *P*<.001). Shift effects were not statistically significant (night vs day, *P*=.15; overnight vs day, *P*=.33).

Table 8. Generalized estimating equation model coefficients for handover document completion time comparing pre- and postintegration periods.

Outcome	Factor	β (seconds)	95% CI	*P* value
Document completion time[a]	Intercept	118.01	112.21 to 123.81	<.001
Document completion time	Pre[b] vs Post[c]	63.48	53.81 to 73.16	<.001
Document completion time	Taipei Medical University Hospital vs Shuang Ho Hospital	19.10	10.34 to 27.85	<.001
Document completion time	Wan Fang Hospital vs Shuang Ho Hospital	–25.67	–34.42 to –16.92	<.001
Document completion time	Night vs day	6.44	–2.35 to 15.22	.15
Document completion time	Overnight vs day	–5.98	–17.99 to 6.03	.33

[a]Overall document completion time per patient.

[b]Preintegration document completion time.

[c]Postintegration document completion time.

Revision Rate of GenAI-Generated Handover Drafts

Using NIS logs from October to December 2024, we examined revisions of GenAI-generated handover drafts by hospital and shift. Table 9 presents, for each hospital-shift stratum, the number of patient-level handover documents in which nurses revised the GenAI-generated draft (number of documents with revisions), the total number of patient-level handover documents (number of documents), and the corresponding revision rate (number of documents with revisions/number of documents, %).

Revisions were common across hospitals and months. At TMUH, revision rates were higher during night (9267/12,740-9162/11,928, 72.74%-76.81%) and overnight shifts (9551/11,918-10,830/12,642, 80.14%-85.67%) than during day shifts (8343/12,510-8164/12,043, 66.69%-67.79%). At WFH, day-shift revision rates increased from nearly 50% (4215/8291, 50.84%) in October to over 60% (5334/8620-5562/8905, 61.88%-62.46%) in November-December, whereas night-shift revision rates decreased from nearly 79% (7684/9612, 79.94%) in October to nearly 60% (6000/10,035-6178/10,190, 59.79%-60.63%) thereafter; overnight-shift revision rates remained high (6960/9438-7804/9597, 73.74%-81.32%). At SHH, revision rates were relatively stable, with day shifts at 68.86%-74.16% (9554/13,875-9717/13,103) and night and overnight shifts at 73.95%-78.95% (11,861/16,040-11,884/15,053).

XSL•FO
RenderX

Table 9. Monthly patient-level handover document counts and revision rates of generative artificial intelligence–generated drafts by hospital and shift (October-December 2024).[a]

Hospital and shift	October 24			November 24			December 24		
	Number of documents with revisions[b], n	Number of documents[c], n	Revision rate[d], %	Number of documents with revisions, n	Number of documents, n	Revision rate, %	Number of documents with revisions, n	Number of documents, n	Revision rate, %
Taipei Medical University Hospital									
Day	8164	12,043	67.79	8001	11,882	67.34	8343	12,510	66.69
Night	9162	11,928	76.81	9000	11,900	75.63	9267	12,740	72.74
Overnight	9551	11,918	80.14	9744	11,745	82.96	10,830	12,642	85.67
Wan Fang Hospital									
Day	4215	8291	50.84	5334	8620	61.88	5562	8905	62.46
Night	7684	9612	79.94	6000	10,035	59.79	6178	10,190	60.63
Overnight	7804	9597	81.32	6960	9438	73.74	7810	9635	81.06
Shuang Ho Hospital									
Day	9717	13,103	74.16	9717	13,931	69.75	9554	13,875	68.86
Night	11,886	15,317	77.60	11,886	15,268	77.85	11,861	16,040	73.95
Overnight	11,884	15,053	78.95	11,884	15,337	77.49	11,759	15,030	78.24

[a]Each log entry corresponds to 1 handover document for a single patient within a shift.

[b]"Number of documents with revisions" denotes documents in which the generative artificial intelligence–generated draft was edited before confirmation.

[c]"Number of documents" denotes the total number of patient-level handover documents.

[d]The revision rate (%) was calculated as documents with revisions/number of documents × 100.

Estimated Time Savings in Handover Documentation Before and After LLM Integration

Nurses at TMUH, WFH, and SHH estimated that they typically completed handover documentation for 5-7 patients per shift when using the original NIS. Based on the available preintegration logs (limited to 1 month), the average documentation completion time ranged from 3.45 (SD 3.82) to 4.32 (SD 4.48) minutes per patient across hospitals and shifts (Table 4). Postintegration logs from October to December 2024 (Table 5) further characterized workflow performance by decomposing document completion time into GenAI generation time and nurse review-revision-confirmation time; the average overall completion time per patient ranged from 1.17 (SD 1.86) to 2.54 (SD 2.82) minutes. Taken together, these figures suggest a potential absolute reduction of approximately 0.91-3.15 minutes per patient ([3.45-2.54]/3.45-[4.32-1.17]/4.32, 26%-73% relative reduction). Moreover, using monthly patient volumes and the observed reduction in per-patient completion time, the LLM-integrated workflow corresponds to estimated savings of approximately 470-980 hours per month across the 3 hospitals (Table 10). Full-time equivalent (FTE) is a standard unit used to express workload or staffing capacity as the work time of 1 full-time worker over a specified period. FTE equivalents were calculated assuming 160 nursing work hours per FTE-month. At TMUH, the projected reduction was approximately 113-282 hours per month (equivalent to 0.7-1.8 FTE/month, assuming 160 working hours per FTE). At WFH, the estimated savings were approximately 160-314 hours per month (1.0-2.0 FTE/month). At SHH, which had the highest patient volume, the estimated savings were approximately 198-387 hours per month (1.2-2.4 FTE/month).

XSL•FO
RenderX

Table 10. Estimated monthly nursing time savings after large language model integration, by hospital (hours and full-time equivalents).[a]

Hospital	October 24, hours (full-time equivalent)[b]	November 24, hours (full-time equivalent)[b]	December 24, hours (full-time equivalent)[b]	Average, hours (full-time equivalent)[b]
Taipei Medical University Hospital	113-273 (0.71-1.71)	115-276 (0.72-1.72)	117-282 (0.73-1.76)	115-277 (0.72-1.73)
Wan Fang Hospital	163-307 (1.02-1.92)	166-314 (1.04-1.96)	160-302 (1.00-1.89)	163-308 (1.02-1.92)
Shuang Ho Hospital	198-381 (1.24-2.38)	203-391 (1.27-2.44)	201-387 (1.25-2.42)	201-386 (1.25-2.41)
Total	474-962 (2.96-6.01)	484-981 (3.02-6.13)	478-970 (2.99-6.07)	479-971 (2.99-6.07)

[a]Hours saved/month = patients × (minutes saved/patient)/60. Preintegration times (September 2024 only; Table 3) were 3.45-4.32 minutes/patient. Postintegration times (October-December 2024; Table 4) were as follows: Taipei Medical University Hospital, 2.13-2.54 minutes/patient; Wan Fang Hospital, 1.17-1.78 minutes/patient; and Shuang Ho Hospital 1.74-2.11 minutes/patient. Bounds were computed as pre-min – post-max and pre-max – post-min.

[b]Values are listed as ranges.

Cost-Effectiveness of GenAI-Assisted Handover Documentation

Preliminary GenAI token cost-effectiveness was evaluated by comparing GenAI operating expenditures with estimated nursing time savings (Table 11). Monthly patient volume was used to approximate documentation time saved, and nursing time was valued at US $10 per hour to estimate the equivalent labor value (opportunity cost) of time freed for clinical work (ie, this does not represent direct budget savings unless staffing levels or overtime are reduced). On a monthly basis, allocated GenAI operating costs were approximately US $493-US $986 per hospital, corresponding to an estimated 113-391 hours saved per hospital per month. The resulting monthly net labor value ranged from US $322 to US $2100 at TMUH, US $970 to US $2595 at WFH, and US $1024 to US $3082 at SHH. Aggregated across hospitals, the estimated time savings were 474-981 hours per month, equivalent to US $4740-US $9810 in labor value. After subtracting monthly GenAI operating expenditures (US $1871-US $2464), the implied net labor value was approximately US $2276-US $7939 per month. Over the 3-month period, the total GenAI cost was US $6396.97, with an estimated 1436-2913 hours saved, yielding an overall net labor value of US $7963-US $22,733.

XSL•FO
RenderX

Table 11. Estimated monthly generative artificial intelligence operating costs, hours saved, equivalent labor value, and net labor value by hospital (October-December 2024).

Hospital and measure	October 24	November 24	December 24	Overall (October to December)
Taipei Medical University Hospital				
Allocated generative artificial intelligence cost, US $[a]	630.18	688.92	848.02	2167.11
Time saved (hours)[b], range	113-273	115-276	117-282	345-831
Equivalent labor value (US $10/hour), range	1130-2730	1150-2760	1170-2820	3450-8310
Net labor value[c] (US $), range	499.82-2099.82	461.08-2071.08	321.98-1971.98	1282.89-6142.89
Wan Fang Hospital				
Allocated generative artificial intelligence cost, US $	493.14	544.74	630.12	1667.99
Time saved (hours), range	163-307	166-314	160-302	489-923
Equivalent labor value (US $10/hour), range	1630-3070	1660-3140	1600-3020	4890-9230
Net labor value (US $), range	1136.86-2576.86	1115.26-2595.26	969.88-2389.88	3222.01-7562.01
Shuang Ho Hospital				
Allocated generative artificial intelligence cost, US $	747.75	828.36	985.75	2561.87
Time saved (hours), range	198-381	203-391	201-387	602-1159
Equivalent labor value (US $10/hour), range	1980-3810	2030-3910	2010-3870	6020-11,590
Net labor value (US $), range	1232.25-3062.25	1201.64-3081.64	1024.25-2884.25	3458.13-9028.13
All				
Allocated generative artificial intelligence cost, US $	1871.07	2062.02	2463.89	6396.97
Time saved (hours), range	474-961	484-981	478-971	1436-2913
Equivalent labor value (US $10/hour), range	4740-9610	4840-9810	4780-9710	14,360-29,130
Net labor value (US $), range	2868.93-7738.93	2777.98-7747.98	2316.11-7246.11	7963.03-22,733.03

[a]Allocated generative artificial intelligence costs were distributed according to each hospital's monthly patient-volume share.

[b]Time saved was estimated from monthly patient volumes and the reduction in per-patient handover document completion time.

[c]Net labor value was calculated as equivalent labor value − allocated generative artificial intelligence cost.

Discussion

Principal Findings

In this study, we integrated Google Gemini 1.5 Pro into the in-house NIS of TMU Health to improve the efficiency of nursing handover documentation across 3 hospitals in Taiwan. In the LLM-integrated workflow, a nurse initiates handover by selecting a patient, after which the LLM-integrated NIS retrieves relevant patient data and submits section-specific, tailored prompts to the LLM. The LLM generates draft content for each handover section, and the compiled document is returned to the LLM-integrated NIS for nurse review, revision, and finalization. This integration streamlines handover documentation by reducing manual data entry and supporting a more efficient routine workflow.

The LLM-integrated NIS retains the functionalities of the original in-house NIS, thereby enabling seamless incorporation of the LLM into the existing nursing handover documentation workflow. Nurses remain familiar with the system interfaces, obviating the need for adaptation and minimizing challenges related to digital skills training and system usability. Furthermore, nurses can readily interact with the LLM-integrated NIS and access the generated content, ensuring that the system does not interrupt workflow [42,43]. Consequently, nurses can maintain their focus on patient care while benefiting from the time-saving features of the LLM, ultimately improving both clinical and operational outcomes.

Across 3 TMU Health hospitals, postintegration NIS logs showed shorter per-patient handover documentation completion times than the limited 1-month preintegration baseline, corresponding to an estimated reduction of 0.91-3.15 minutes per patient ([3.45-2.54]/3.45-[4.32-1.17]/4.32, 26%-73%). When scaled by monthly patient volumes (Table 10), this reduction translated to approximately 470-980 hours saved per month across hospitals (approximately 0.7-1.8 FTE/month, assuming 160 working hours per FTE). A preliminary token cost evaluation suggested that GenAI operating expenditures (US

XSL•FO
RenderX

$493-US $986 per hospital per month) were modest relative to the estimated labor or opportunity value of nursing time freed for clinical work, yielding a positive net labor value across October-December 2024 (Table 11). Overall, these findings suggest that GenAI-assisted handover documentation may reduce documentation burden and return clinically meaningful nursing time that could be redirected to direct patient care, which is particularly relevant in the context of nursing workforce shortages.

Safety Considerations and Hallucination Mitigation

Hallucinations are a well-recognized limitation of GenAI, particularly in clinical documentation tasks in which factual accuracy is critical [44-46]. During initial development and testing of the LLM-integrated NIS, we observed occasional hallucinations when the LLM was used to extract and structure vital signs from NIS inputs; for example, the model sometimes generated an average body temperature that was not explicitly documented in the source record. To mitigate hallucination risk, we implemented multiple safeguards. First, we iteratively refined prompts to restrict the LLM to reformatting and summarizing only the provided input and to prohibit the introduction of new values or inferences. Second, we required nurse review, correction, and confirmation before finalizing each handover note, thereby adding an additional verification layer. Despite these measures, we considered the residual risk of LLM-generated structured vital-sign outputs to be clinically unacceptable. Accordingly, we discontinued LLM-based structuring of vital signs and instead directly displayed the patient's vital signs in the NIS. By contrast, the LLM appeared more reliable for extracting and synthesizing narrative information (eg, abnormal findings and clinically relevant descriptions) from radiology and examination reports, in which the task emphasized highlighting salient text rather than generating or imputing numeric values.

Limitations

First, only 1 month of preintegration log data were available, limiting the robustness of the baseline for manual handover documentation and constraining pre-post comparisons. Accordingly, the preintegration results should be interpreted as a preliminary benchmark for estimating the potential impact of subsequent GenAI-assisted workflows. Second, the NIS log schema was not sufficiently granular to support several secondary analyses. Specifically, we could not quantify technical reliability indicators (eg, connection failures, API time-outs, error returns, failed generations, or regeneration attempts), nor could we stratify performance by patient complexity (eg, intensive care unit vs general ward) or clinical acuity proxies such as length of stay. Consequently, the reported time and cost estimates reflect system-level averages and may mask heterogeneity in performance for complex cases or under high-load conditions. Third, we did not conduct a formal natural language processing evaluation of generated text quality using standard metrics or a gold-standard reference set [47,48]. Although nursing professionals reviewed and verified the LLM-generated handover content as part of routine workflow, the absence of quantitative natural language processing evaluation limits comparability with prior summarization studies and precludes objective assessment of section-level content fidelity. Fourth, although we quantified the revision rate of GenAI-generated drafts, we did not classify edit types (eg, corrections vs clarifications vs additions) or examine the reasons underlying nurse modifications. Therefore, we could not determine which content areas most frequently required changes or whether revisions primarily reflected model limitations, workflow preferences, or patient- and context-specific factors. Fifth, we did not conduct a formal usability or user experience evaluation. Specifically, we did not administer validated instruments (eg, the System Usability Scale or Technology Acceptance Model) or conduct structured interviews to assess nurses' perceived workload, trust in GenAI outputs, or overall satisfaction, limiting conclusions about user acceptability beyond log-based performance metrics.

Future Directions

Our preliminary findings suggest that LLMs can support rapid, automated generation of nursing handover documents; however, nurse review and revision remain necessary before finalization. First, future work will expand the NIS logging framework to capture reliability and workflow metrics (eg, time-outs, error codes, retries, regeneration events, and editing behaviors) and enable more granular performance evaluation, including downstream operational outcomes such as nursing overtime hours. Second, we will compare LLM-generated drafts with nurse-finalized handovers to characterize error patterns, quantify revision types, and inform iterative prompt and model refinement. These analyses will be guided by predefined quality indicators and evaluation frameworks (eg, quantitative and qualitative ISBAR-based assessments) in future studies. Collectively, these data will support improvements in factual accuracy and the development of additional safeguards, including rule-based checks and retrieval-augmented generation strategies [49]. Third, to evaluate usability and acceptance, we will incorporate standardized instruments and structured qualitative feedback to assess nurse-perceived usability, trust, cognitive load, and workflow impact alongside log-derived efficiency outcomes. Fourth, to improve transferability, we will evaluate interoperability in settings with different EMR architectures and develop a configurable prompt-and-template layer (eg, leveraging HL7/FHIR [Health Level 7/Fast Healthcare Interoperability Resources] where available) to reduce site-specific customization. Specifically, we aim to reduce institution-specific effort related to data mapping; integration with source systems (eg, laboratory information system/radiology information system/examination information system); authentication and authorization; local handover templates and documentation policies; and network or operational configurations. In parallel, we will preserve and further validate transferable components, including ISBAR-aligned section prompts, the human-in-the-loop workflow (draft generation with nurse review-revision-confirmation), safety safeguards (eg, prompt constraints and use of source-of-truth structured data), and a log-based evaluation framework for efficiency, revision behavior, and reliability metrics. Finally, we will assess generalizability to other clinical documentation workflows,

XSL•FO
RenderX

including automated note drafting and radiology report generation.

Conclusions

In this study, we integrated an LLM into an existing NIS to generate nursing handover documents in routine practice without modifying the user interface or disrupting established workflows. Across 3 hospitals within TMU Health, GenAI-assisted handover documentation was associated with reduced per-patient documentation completion time in real-world use. A preliminary cost-effectiveness analysis suggested that GenAI operating expenditures were modest relative to the estimated labor value of nursing time freed for clinical work from October to December 2024. To mitigate hallucination risk, particularly for structured numeric data, we constrained prompts and embedded nurse review, editing, and confirmation within the workflow. Overall, these findings support the operational feasibility of integrating GenAI into hospital documentation workflows while retaining human verification. Future work will expand NIS logging to quantify reliability and editing metrics, compare LLM-generated drafts with nurse-finalized documents to identify error patterns and inform prompt and model refinement, and evaluate interoperability and generalizability to other clinical documentation tasks.

Acknowledgments

We thank all the nursing staff of TMU Health for their support. This research was supported by TMU in Taiwan (grant TMU 109-5400-025-400). OpenAI's ChatGPT (GPT-5.2) was used solely for grammar and spelling checks. All edits were reviewed by the authors, and no patient-identifiable data were entered into the tool.

Data Availability

The data are available upon reasonable request from the corresponding author (YSL) and with TMU-JIRB approval; they are not publicly available due to ethical and privacy restrictions.

Authors' Contributions

Conceptualization: YSL
Formal analysis: LWT
Investigation: SSC, STSH
Methodology: YSL
Supervision: RJC, MSW
Writing – original draft: YSL

Conflicts of Interest

None declared.

Multimedia Appendix 1
Monthly patient volume, nurse staff levels, shift counts, and mean length of stay, stratified by hospital and shift type (September-December 2024).
[DOCX File , 24 KB - jmir_v28i1e81604_app1.docx]

References

1. Oulton JA. The global nursing shortage: an overview of issues and actions. Policy Polit Nurs Pract 2006 Aug;7(3 Suppl):34S-39S. [doi: 10.1177/1527154406293968] [Medline: 17071693]
2. Marć M, Bartosiewicz A, Burzyńska J, Chmiel Z, Januszewicz P. A nursing shortage - a prospect of global and local policies. Int Nurs Rev 2019 Mar;66(1):9-16. [doi: 10.1111/inr.12473] [Medline: 30039849]
3. Qureshi SM, Bookey-Bassett S, Purdy N, Greig MA, Kelly H, Neumann WP. Modelling the impacts of COVID-19 on nurse workload and quality of care using process simulation. PLoS One 2022;17(10):e0275890 [FREE Full text] [doi: 10.1371/journal.pone.0275890] [Medline: 36228015]
4. Ribeiro OMPL, Coimbra VMO, Pereira SCDA, Faria ADCA, Teles PJFC, Rocha CGD. Impact of COVID-19 on the environments of professional nursing practice and nurses' job satisfaction. Int J Environ Res Public Health 2022 Dec 16;19(24):ijerph192416908 [FREE Full text] [doi: 10.3390/ijerph192416908] [Medline: 36554789]
5. NCSBN research projects significant nursing workforce shortages and crisis. National Council of State Boards of Nursing. URL: https://www.ncsbn.org/news/ncsbn-research-projects-significant-nursing-workforce-shortages-and-crisis [accessed 2026-01-26]
6. State of the world's nursing report 2025: investing in education, jobs, leadership and service delivery. World Health Organization. URL: https://www.who.int/publications/i/item/9789240110236 [accessed 2026-01-26]

XSL•FO
RenderX

7. Liao Z, Sun S, Clarissa C, Aviles L, Lin C, Kao CT, et al. Exploring the challenges of Taiwanese nurses in the COVID-19 post-pandemic era. J Formos Med Assoc 2024 Nov;123 Suppl 3:S215-S222 [FREE Full text] [doi: 10.1016/j.jfma.2024.03.020] [Medline: 38570237]
8. Wu F, Sun J, Wu H, Su M, Lin C. The experiences of public health nurses during the COVID-19 in Taiwan: a qualitative study. Nurs Open 2024 Jan;11(1):e2091 [FREE Full text] [doi: 10.1002/nop2.2091] [Medline: 38268278]
9. Navarro Martínez O, Igual García J, Salcedo VT. Nurses' view of benefits, enablers and constraints to the use of digital health tools with patients: a cross-sectional study. Digit Health 2023 Sep 04;9:20552076231197339 [FREE Full text] [doi: 10.1177/20552076231197339] [Medline: 37675060]
10. Huter K, Krick T, Domhoff D, Seibert K, Wolf-Ostermann K, Rothgang H. Effectiveness of digital technologies to support nursing care: results of a scoping review. J Multidiscip Healthc 2020;13:1905-1926 [FREE Full text] [doi: 10.2147/JMDH.S286193] [Medline: 33328736]
11. Abd El Mongy Ismail B, Mamdouh Elkholy S, Mohamed Said R, Said Elwakeil E. Digital nurse' competencies in relation to patient safety and teamwork communication. Egyptian Journal of Health Care 2025 Mar 01;16(1):705-717. [doi: 10.21608/ejhc.2025.417311]
12. Kutney-Lee A, Kelly D. The effect of hospital electronic health record adoption on nurse-assessed quality of care and patient safety. J Nurs Adm 2011 Nov;41(11):466-472. [doi: 10.1097/NNA.0b013e3182346e4b] [Medline: 22033316]
13. Alotaibi YK, Federico F. The impact of health information technology on patient safety. Saudi Med J 2017 Dec;38(12):1173-1180 [FREE Full text] [doi: 10.15537/smj.2017.12.20631] [Medline: 29209664]
14. McCarthy B, Fitzgerald S, O'Shea M, Condon C, Hartnett-Collins G, Clancy M, et al. Electronic nursing documentation interventions to promote or improve patient safety and quality care: a systematic review. J Nurs Manag 2019 Apr;27(3):491-501. [doi: 10.1111/jonm.12727] [Medline: 30387215]
15. Douma MC, Rejeb MB, Zardoub N, Braham A, Chouchene H, Bouallegue O, et al. Impact of implementing electronic nursing records on quality and safety indicators in care. Libyan J Med 2024 Dec 31;19(1):2421625 [FREE Full text] [doi: 10.1080/19932820.2024.2421625] [Medline: 39570988]
16. Bakhoum N, Gerhart C, Schremp E, Jeffrey AD, Anders S, France D, et al. A time and motion analysis of nursing workload and electronic health record use in the emergency department. J Emerg Nurs 2021 Sep;47(5):733-741. [doi: 10.1016/j.jen.2021.03.007] [Medline: 33888334]
17. Duffield C, Diers D, O'Brien-Pallas L, Aisbett C, Roche M, King M, et al. Nursing staffing, nursing workload, the work environment and patient outcomes. Appl Nurs Res 2011 Nov;24(4):244-255. [doi: 10.1016/j.apnr.2009.12.004] [Medline: 20974086]
18. Blay N, Roche MA, Duffield C, Gallagher R. Intrahospital transfers and the impact on nursing workload. J Clin Nurs 2017 Dec;26(23-24):4822-4829. [doi: 10.1111/jocn.13838] [Medline: 28382638]
19. Shihundla RC, Lebese RT, Maputle MS. Effects of increased nurses' workload on quality documentation of patient information at selected primary health care facilities in Vhembe District, Limpopo Province. Curationis 2016 May 13;39(1):1545 [FREE Full text] [doi: 10.4102/curationis.v39i1.1545] [Medline: 27246793]
20. Lavander P, Meriläinen M, Turkki L. Working time use and division of labour among nurses and health-care workers in hospitals - a systematic review. J Nurs Manag 2016 Nov;24(8):1027-1040. [doi: 10.1111/jonm.12423] [Medline: 27581093]
21. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 06;6(1):120. [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
22. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
23. Wang L, Wan Z, Ni C, Song Q, Li Y, Clayton E, et al. Applications and concerns of ChatGPT and other conversational large language models in health care: systematic review. J Med Internet Res 2024 Nov 07;26:e22769 [FREE Full text] [doi: 10.2196/22769] [Medline: 39509695]
24. Meskó B. The impact of multimodal large language models on health care's future. J Med Internet Res 2023 Nov 02;25:e52865 [FREE Full text] [doi: 10.2196/52865] [Medline: 37917126]
25. Zhang K, Meng X, Yan X, Ji J, Liu J, Xu H, et al. Revolutionizing health care: the transformative impact of large language models in medicine. J Med Internet Res 2025 Jan 07;27:e59069 [FREE Full text] [doi: 10.2196/59069] [Medline: 39773666]
26. Vert J. How will generative AI disrupt data science in drug discovery? Nat Biotechnol 2023 Jun;41(6):750-751. [doi: 10.1038/s41587-023-01789-6] [Medline: 37156917]
27. Ghebrehiwet I, Zaki N, Damseh R, Mohamad MS. Revolutionizing personalized medicine with generative AI: a systematic review. Artif Intell Rev 2024 Apr 25;57(5):128. [doi: 10.1007/s10462-024-10768-5]
28. Patrinos GP, Sarhangi N, Sarrami B, Khodayari N, Larijani B, Hasanzad M. Using ChatGPT to predict the future of personalized medicine. Pharmacogenomics J 2023 Nov 19;23(6):178-184. [doi: 10.1038/s41397-023-00316-9] [Medline: 37726551]
29. Huang J, Neill L, Wittbrodt M, Melnick D, Klug M, Thompson M, et al. Generative artificial intelligence for chest radiograph interpretation in the emergency department. JAMA Netw Open 2023 Oct 02;6(10):e2336100 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36100] [Medline: 37796505]

XSL•FO
RenderX

30. Hong EK, Roh B, Park B, Jo J, Bae W, Soung Park J, et al. Value of using a generative ai model in chest radiography reporting: a reader study. Radiology 2025 Mar;314(3):e241646. [doi: 10.1148/radiol.241646] [Medline: 40067108]
31. Baxter SL, Longhurst CA, Millen M, Sitapati AM, Tai-Seale M. Generative artificial intelligence responses to patient messages in the electronic health record: early lessons learned. JAMIA Open 2024 Jul;7(2):ooae028 [FREE Full text] [doi: 10.1093/jamiaopen/ooae028] [Medline: 38601475]
32. Hobensack M, von Gerich H, Vyas P, Withall J, Peltonen L, Block LJ, et al. A rapid review on current and potential uses of large language models in nursing. Int J Nurs Stud 2024 Jun;154:104753. [doi: 10.1016/j.ijnurstu.2024.104753] [Medline: 38560958]
33. Zong H, Wu R, Cha J, Wang J, Wu E, Li J, et al. Large language models in worldwide medical exams: platform development and comprehensive analysis. J Med Internet Res 2024 Dec 27;26:e66114 [FREE Full text] [doi: 10.2196/66114] [Medline: 39729356]
34. Yun HS, Bickmore T. Online health information-seeking in the era of large language models: cross-sectional web-based survey study. J Med Internet Res 2025 Mar 31;27:e68560 [FREE Full text] [doi: 10.2196/68560] [Medline: 40163112]
35. Tan D, Huang Y, Liu M, Li Z, Wu X, Huang C. Identification of online health information using large pretrained language models: mixed methods study. J Med Internet Res 2025 May 14;27:e70733 [FREE Full text] [doi: 10.2196/70733] [Medline: 40367512]
36. Seibert K, Domhoff D, Bruch D, Schulte-Althoff M, Fürstenau D, Biessmann F, et al. Application scenarios for artificial intelligence in nursing care: rapid review. J Med Internet Res 2021 Nov 29;23(11):e26522 [FREE Full text] [doi: 10.2196/26522] [Medline: 34847057]
37. Ronquillo CE, Peltonen L, Pruinelli L, Chu CH, Bakken S, Beduschi A, et al. Artificial intelligence in nursing: priorities and opportunities from an international invitational think-tank of the Nursing and Artificial Intelligence Leadership Collaborative. J Adv Nurs 2021 Sep;77(9):3707-3717 [FREE Full text] [doi: 10.1111/jan.14855] [Medline: 34003504]
38. Yelne S, Chaudhary M, Dod K, Sayyad A, Sharma R. Harnessing the power of AI: a comprehensive review of its impact and challenges in nursing science and healthcare. Cureus 2023 Nov;15(11):e49252 [FREE Full text] [doi: 10.7759/cureus.49252] [Medline: 38143615]
39. El Arab RA, Al Moosa OA, Abuadas FH, Somerville J. The role of AI in nursing education and practice: umbrella review. J Med Internet Res 2025 Apr 04;27:e69881 [FREE Full text] [doi: 10.2196/69881] [Medline: 40072926]
40. Gemini API: Google AI for developers. Google LLC. URL: https://ai.google.dev/gemini-api/docs [accessed 2026-01-28]
41. Burgess A, van Diggele C, Roberts C, Mellis C. Teaching clinical handover with ISBAR. BMC Med Educ 2020 Dec 03;20(Suppl 2):459 [FREE Full text] [doi: 10.1186/s12909-020-02285-0] [Medline: 33272274]
42. Yoo HJ, Lee H. Critical role of information and communication technology in nursing during the COVID-19 pandemic: a qualitative study. J Nurs Manag 2022 Nov;30(8):3677-3685 [FREE Full text] [doi: 10.1111/jonm.13880] [Medline: 36325914]
43. Abdolkhani R, Petersen S, Walter R, Zhao L, Butler-Henderson K, Livesay K. The impact of digital health transformation driven by COVID-19 on nursing practice: systematic literature review. JMIR Nurs 2022 Aug 30;5(1):e40348 [FREE Full text] [doi: 10.2196/40348] [Medline: 35867838]
44. Birhane A, Kasirzadeh A, Leslie D, Wachter S. Science in the age of large language models. Nat Rev Phys 2023 Apr 26;5(5):277-280. [doi: 10.1038/s42254-023-00581-4]
45. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 08;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
46. Goodman KE, Yi PH, Morgan DJ. AI-generated clinical summaries require more than accuracy. JAMA 2024 Feb 27;331(8):637-638. [doi: 10.1001/jama.2024.0555] [Medline: 38285439]
47. Bala S, Keniston A, Burden M. Patient perception of plain-language medical notes generated using artificial intelligence software: pilot mixed-methods study. JMIR Form Res 2020 Jun 05;4(6):e16670 [FREE Full text] [doi: 10.2196/16670] [Medline: 32442148]
48. Abbasian M, Khatibi E, Azimi I, Oniani D, Shakeri Hossein Abad Z, Thieme A, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. NPJ Digit Med 2024 Mar 29;7(1):82 [FREE Full text] [doi: 10.1038/s41746-024-01074-z] [Medline: 38553625]
49. Jeong M, Sohn J, Sung M, Kang J. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. Bioinformatics 2024 Jun 28;40(Suppl 1):i119-i129 [FREE Full text] [doi: 10.1093/bioinformatics/btae238] [Medline: 38940167]

Abbreviations

AI: artificial intelligence
API: application programming interface
EMR: electronic medical record
FHIR: Fast Healthcare Interoperability Resources

XSL•FO
RenderX

FTE: full-time equivalent
GEE: generalized estimating equation
GenAI: generative artificial intelligence
HIPAA: Health Insurance Portability and Accountability Act
HL7: Health Level 7
ISBAR: Identify, Situation, Background, Assessment, Recommendation
JIRB: Joint Institutional Review Board
LLM: large language model
NIS: nursing information system
RIS: radiology information system
SHH: Shuang Ho Hospital
TMU: Taipei Medical University
TMUH: Taipei Medical University Hospital
WFH: Wan Fang Hospital
WHO: World Health Organization

Edited by A Coristine; submitted 01.Aug.2025; peer-reviewed by KE Zavotsky, N Bievre, F Al Dhabbari, S Al Mukhtar; comments to author 02.Dec.2025; revised version received 26.Feb.2026; accepted 26.Feb.2026; published 12.Mar.2026.

Please cite as:
Chen RJ, Wu MS, Tsai LW, Chang SS, Shen Hsiao ST, Lo YS
Integrating a Large Language Model to Streamline Nursing Handover Documentation Across Multiple Hospitals in Taiwan: Development and Implementation Study
J Med Internet Res 2026;28:e81604
URL: https://www.jmir.org/2026/1/e81604
doi:10.2196/81604
PMID:

©Ray-Jade Chen, Mai-Szu Wu, Lung-Wen Tsai, Shy-Shin Chang, Shu-Tai Shen Hsiao, Yu-Sheng Lo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Health Technology Use Among Rehabilitation Professionals in China: Multi-Province Cross-Sectional Survey

Shuning Duan[1,2], MSc; Zi-Ru Wang[1], PhD; Xier Chen[2], MSc; Guoxin Ni[1], MD, PhD

[1]Department of Rehabilitation Medicine, The First Affiliated Hospital of Xiamen University, School of Medicine, Xiamen University, Xiamen, China
[2]School of Sports Medicine, Wuhan Sports University, Wuhan, Hubei, China

Corresponding Author:
Guoxin Ni, MD, PhD
Department of Rehabilitation Medicine
The First Affiliated Hospital of Xiamen University
School of Medicine, Xiamen University
No. 55 Zhenhai Road
Siming District
Xiamen, 361003
China
Phone: 86 5922137084
Email: nigx@xmu.edu.cn

Abstract

Background: The rapid expansion of rehabilitation needs in China has intensified pressure on a workforce that remains unevenly distributed. Digital health technologies (DHTs) offer potential to increase service reach and efficiency. However, little is known about how rehabilitation professionals currently gather and document clinical information, nor about their readiness to integrate digital tools into routine practice within China's rapidly digitalizing health system.

Objective: This study aimed to describe how rehabilitation professionals in China collect subjective and objective clinical information, document patient data in routine practice, and assess their willingness to use DHTs in clinical settings.

Methods: We conducted a multi-province observational cross-sectional survey using a culturally adapted questionnaire based on the World Health Organization Digital Health Interventions framework. The instrument assessed participant characteristics, information collection methods, documentation practices, and willingness to adopt digital functions across rehabilitation activities. Descriptive analyses and subgroup comparisons were performed on 324 complete responses from certified rehabilitation professionals. The multi-province cross-sectional online survey was conducted among licensed rehabilitation professionals in China with internet access. Participants were recruited through professional networks and social media platforms.

Results: Respondents were drawn from 20 provincial-level administrative regions across China, including Fujian (n=72), Guangdong (n=77), and Shanxi (n=45), among others, with 82.7% (268/324) employed in public sector rehabilitation services. Traditional methods dominated clinical work. Face-to-face communication was used frequently for subjective assessment by 96.3% (312/324) of respondents, whereas digital channels such as email (22/324, 6.8%) and telephone (47/324, 14.5%) saw limited use. For objective information, visual observation (271/324, 83.7%) and manual measurement tools (195/324, 60.2%) remained the primary approaches, while motion capture technology (45/324, 13.8%) and wearable sensors (13/324, 4%) were rarely used. Documentation practices also relied heavily on analogue formats, with 82.1% (266/324) using handwritten notes and 60.2% (195/324) using paper templates. In contrast, willingness to adopt DHTs was consistently high, with 80.6% (261/324) of respondents indicating readiness to use digital systems for identity verification, 79.0% (256/324) for progress tracking, and 78.1% (253/324) for outcome measurement. Subgroup analyses revealed that educational level significantly influenced the adoption of advanced technologies, with master's or doctoral degree holders reporting higher use of sensor-based assessment, motion capture, and wearable devices. In contrast, professional title and clinical specialty showed limited influence, with no significant differences observed for most digital health functions.

Conclusions: Rehabilitation professionals in China demonstrate strong readiness to use DHTs, yet their routine practice remains largely paper-based and analogue. These findings provide evidence to inform implementation strategies, workforce training, and system-level planning aimed at accelerating digital transformation in rehabilitation services.

(J Med Internet Res 2026;28:e90623) doi:10.2196/90623

XSL•FO
RenderX

KEYWORDS

digital health technologies; telehealth; rehabilitation; online survey; health technology acceptance

Introduction

Rehabilitation is widely recognized as a core component of modern health systems. Global estimates indicate that more than 2 billion people live with health conditions that could benefit from rehabilitation, and this number continues to grow as populations age and chronic conditions become more common [1,2]. The role of rehabilitation in restoring function, promoting independence, and supporting healthy aging has been increasingly acknowledged in global public health priorities [3,4]. Rehabilitation is a series of interventions for individuals who are experiencing or may experience limitations in everyday functioning to help them obtain and maintain the optimum functional status [5,6]. Globally, rehabilitation services primarily address functional impairments resulting from injury, illness, or chronic conditions, encompassing disciplines such as physiotherapy, occupational therapy, and speech-language therapy [5,7,8]. Within China's health care system, this scope aligns with the professional classifications established by the National Health Commission (NHC) and the Chinese Association of Rehabilitation Medicine, focusing on physical and functional rehabilitation rather than substance use disorder or mental health rehabilitation [9,10]. Insights into rehabilitation professionals' current digital health practices and readiness can guide the successful implementation of these technologies in clinical settings.

Despite this growing demand, access to rehabilitation is uneven across regions. The World Health Organization (WHO) has noted that the density of rehabilitation professionals remains far below required levels in many settings, especially in low-and-middle income regions [11,12]. Previous studies have revealed that China ranked first globally in rehabilitation needs [13], yet faces severe supply shortages and regional disparities. China faces a severe workforce shortage, with only an estimated 150,000 physiotherapists serving a population of 1.4 billion. This equates to approximately 10.7 therapists per 100,000 population [14], far below the density in Australia, where physiotherapists alone account for 153.2 per 100,000 population [15,16]. Evidence indicates that China's digital health infrastructure has expanded rapidly, with 900 internet hospitals and over 24,000 telemedicine facilities established nationwide by October 2020 [17]. However, persistent institutional barriers remain. Information silos across different levels of rehabilitation institutions impede data sharing and care coordination [18,19]. Coupled with this, gaps in the payment system exclude most digital solutions from national health insurance coverage [20,21]. In such large and complex health systems, the scale of the population and the fragmentation of service delivery further complicate efforts to expand rehabilitation capacity in a timely and coordinated manner. These challenges place additional strain on existing health care infrastructure and highlight the need for a more sustainable and scalable approach to delivering rehabilitation that can extend reach, support continuity of care, and make efficient use of limited human resources.

Digital health technologies (DHTs), as defined by the WHO, involve the use of digital and smart technologies, including artificial intelligence, the internet of things, big data, robotics, and connected devices, to improve health outcomes [22]. These technologies offer significant advantages in efficiency, scalability, and user accessibility. Evidence suggests that digital tools can support communication, monitoring, and self-management; contribute to improved outcomes; and reduce costs [23-25]. For example, in musculoskeletal rehabilitation, DHTs have emerged as transformative solutions to address the limitations of conventional care, with recent evidence demonstrating that artificial intelligence-driven motion analysis, wearable sensors, and tele-rehabilitation platforms can achieve outcomes comparable or superior to traditional approaches, alongside adherence gains of 15-40% and cost reductions of approximately 30-40% [26]. In post-trauma rehabilitation, remote digital interventions delivered via video conferencing and virtual reality (VR) have shown positive effects on physical and psychosocial outcomes for individuals with spinal cord injuries [27]. Wearable sensor technology has also gained traction in routine rehabilitation care; studies indicate that both clinicians and patients recognize its value in objectively measuring activity outside clinical settings, with most clinicians preferring continuous monitoring integrated into electronic health records to optimize workflow [28,29]. In stroke rehabilitation, home-based DHTs, including wearable devices, smartphone apps, and sensor-based solutions, have demonstrated high feasibility and usability for monitoring mobility, upper extremity function, and daily living activities [30]. Digital avatars represent another emerging modality. These approaches may be especially valuable where access to rehabilitation is limited by distance, concentration of services in urban centers, or shortages of specialized staff.

Despite the potential of DHTs to transform rehabilitation services, their development and application among rehabilitation professionals in China remain limited. Cross-disciplinary collaboration among policymakers, health care providers, technology developers, and end-users is essential to ensure that these digital solutions align with the needs of rehabilitation professionals and the patients they serve. Existing evidence suggests that digital health adoption varies across health systems and professional groups, influenced by factors such as workflow integration, technical infrastructure, and system-level support [31-33]. However, most of the existing work has focused on general practice or nursing professions, while the DHT practices of rehabilitation professionals remain comparatively underexamined. This gap is particularly notable in middle-income countries, where digital health infrastructure is expanding but the role of these tools in everyday rehabilitation practice has not been well documented [34,35]. Generating empirical data from such settings can help clarify current patterns of DHTs use and inform future implementation strategies that support appropriate and equitable integration of digital health within rehabilitation services. The findings will

XSL•FO
RenderX

help guide future research in the design of DHTs as well as policymakers.

Methods

Study Design

A multi-province cross-sectional survey was conducted to describe the use of DHTs among rehabilitation professionals in China. The study was reported in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [36] (Multimedia Appendix 1). Only complete and eligible survey responses were included in the analysis.

Instrument Development and Reliability

The survey was adapted from the clinician version of the digital health interventions (DHI) questionnaire developed by Merolli et al [37], which was created through qualitative item review of the WHO DHI classification framework [38]. The questionnaire was translated into Chinese and culturally adapted following established guidelines. The adaptation process included forward translation by 2 bilingual researchers, expert review by 5 Chinese rehabilitation specialists, back translation to ensure conceptual equivalence, cognitive interviews with 10 rehabilitation therapists, and pilot testing with 30 therapists. All adaptations focused on maintaining conceptual equivalence while ensuring clarity and relevance for rehabilitation professionals working in the Chinese health system.

The final instrument comprised four sections: (1) demographic and professional characteristics, (2) methods for obtaining subjective and objective clinical information, (3) documentation practices, and (4) willingness to use digital technologies for 32 rehabilitation-related clinical functions. All items assessing frequency and willingness used 5-point Likert scales, adapted from the approach used by Merolli et al [37]. For frequency items, responses were scored from 1 (“Never”) to 5 (“Always”). For willingness items, responses were scored from 1 (“Not at all willing”) to 5 (“very much willing”). In both cases, higher scores indicate greater frequency of use or stronger willingness to adopt. For analysis, responses of 4 or 5 were classified as “users” (for frequency items) or “willing” (for willingness items), following the approach of Merolli et al [37]. A sensitivity analysis using an alternative classification threshold (responses 3-5 defined as users and willing) confirmed that the core findings, particularly the substantial gap between willingness and actual use, remained robust regardless of the cut-off point chosen. The complete Chinese survey instrument and its English translation are provided in Multimedia Appendix 2 and Multimedia Appendix 3, respectively.

Internal consistency was assessed using Cronbach α coefficient. The results showed: subjective information collection domain (9 items): α=0.782; objective information collection domain (10 items): α=0.878; documentation practices domain (12 items): α=0.852; willingness to use digital technologies domain (32 items): α=0.986; total scale (63 items): α=0.963. All reliability coefficients exceeded the acceptable threshold of 0.70, indicating good to excellent internal consistency of the adapted questionnaire in this sample.

Setting and Participants

This cross-sectional study was conducted between April and July 2025 among rehabilitation professionals practicing in mainland China. Eligibility was determined through 2 sequential screening questions at the beginning of the online questionnaire. The first question asked: “Have you passed the national health professional qualification examination organized by the NHC and the Ministry of Human Resources and Social Security, obtained a rehabilitation therapy qualification certificate, and are you currently providing clinical rehabilitation services in China?” (Yes or No). The remaining proceeded to the second screening question: “In the past six months, have you treated or managed an average of at least five rehabilitation patients per week?” (Yes or No). Participants represented all professional levels and practice settings, including general hospitals, rehabilitation specialty hospitals, community health centers, older adult care institutions, and sports medicine facilities.

A combination of convenience and snowball sampling was used. The survey targeted rehabilitation professionals practicing in mainland China, including physical therapists, occupational therapists, speech and language therapists, and other recognized rehabilitation therapy practitioners. In China, professional roles in rehabilitation can sometimes overlap, with practitioners potentially assuming responsibilities across multiple specialty areas. To address this complexity and ensure accurate classification of participants’ primary professional focus, we adopted a classification framework based on core clinical responsibilities. This framework was informed by the professional definitions of World Physiotherapy and the World Federation of Occupational Therapists [39,40], and adapted to align with the specialty classifications of the Chinese Association of Rehabilitation Medicine [41,42] Specifically, respondents were asked to select their primary practice area from 9 predefined categories (eg, neurological rehabilitation, musculoskeletal rehabilitation, and pediatric rehabilitation) and to describe their routine responsibilities in an open-ended question. Based on their responses, we classified participants into the most appropriate professional category for analysis.

Recruitment and Data Collection

Data were collected between April and July 2025 using an anonymous online survey hosted on Questionnaire Star (Changsha Ranxing Science and Technology Co, Ltd), a widely used Chinese survey platform. Recruitment occurred through multiple channels, including professional networks, departmental distribution within hospitals, social media groups for rehabilitation therapists, and peer referral. Potential participants accessed the survey through an open link. The front page of the survey clearly stated the research purpose, assured anonymity, and emphasized voluntary participation; proceeding to the questionnaire implied informed consent. After screening questions confirmed eligibility, participants were able to proceed to the full survey. Duplicate submissions were limited using single-device submission control, and no incentives were offered.

XSL•FO
RenderX

Quality Control

To ensure data quality, we implemented multilayered quality control measures throughout the recruitment process. As illustrated in the participant flowchart, of the 393 initial responses received, 52 were excluded for not meeting eligibility criteria (lack of national qualification or not in clinical practice), and an additional 17 were excluded for not meeting the minimum caseload requirement (fewer than 5 patients per week). Following eligibility screening, responses were further scrutinized through automated checks, including IP address to detect duplicate submissions, and manual review of completion times and response patterns. After these quality control procedures, 324 valid responses were retained for final analysis.

Data Analysis

Data from completed surveys were analyzed descriptively using SPSS (version 27; IBM Corp). Categorical variables were summarized using frequencies and percentages. For Likert-scale items, responses of always or frequently were classified as users, and responses of very much willing or quite a bit were classified as willing, following the approach used by Merolli et al [37]. Descriptive secondary comparisons were conducted across demographic and professional subgroups to explore variation in DHT use and willingness. For all subgroup comparisons, Kruskal-Wallis tests were used due to the non-normal distribution of the data. Post-hoc pairwise comparisons were conducted using Dunn test with Bonferroni correction to control for Type I error. Effect sizes (ε^2) were calculated manually using the formula $\varepsilon^2 = (H - k + 1) / (N - k)$, where k represents the number of groups and N the total sample size. Effect sizes were interpreted as small (0.01), medium (0.06), or large (0.14). Open-ended responses were reviewed using a directed qualitative content approach to summarize commonly reported barriers and facilitators.

Missing data were handled through listwise deletion based on the study's eligibility criteria. After applying the predefined exclusion standards (qualification, practice status, and minimum caseload requirements), 324 complete and valid responses were retained for final analysis, with no missing values in any key variables.

Ethical Considerations

Ethics approval for this study was granted by the ethics committee of the First Affiliated Hospital of Xiamen University (approval XMFHIIT-2025SL023) in accordance with institutional guidelines. Informed consent was obtained from all the participants, and data were anonymized to ensure privacy and confidentiality. No compensation was provided for participation.

Results

Overview

This survey collected a total of 393 questionnaires from 20 provinces across China, of which 324 met the eligibility criteria. The exclusion criterion was failure to pass the NHC qualification examination, not currently practicing in mainland China, or having treated or managed an average of fewer than 5 rehabilitation patients per week over the 6 months preceding the survey. Upon receipt of the research participant invitation letter and relevant study information, voluntary participation in the survey constituted informed consent. Figure 1 presents the detailed participant flow diagram for this study.

XSL•FO
RenderX

Figure 1. Recruitment flowchart for this study.

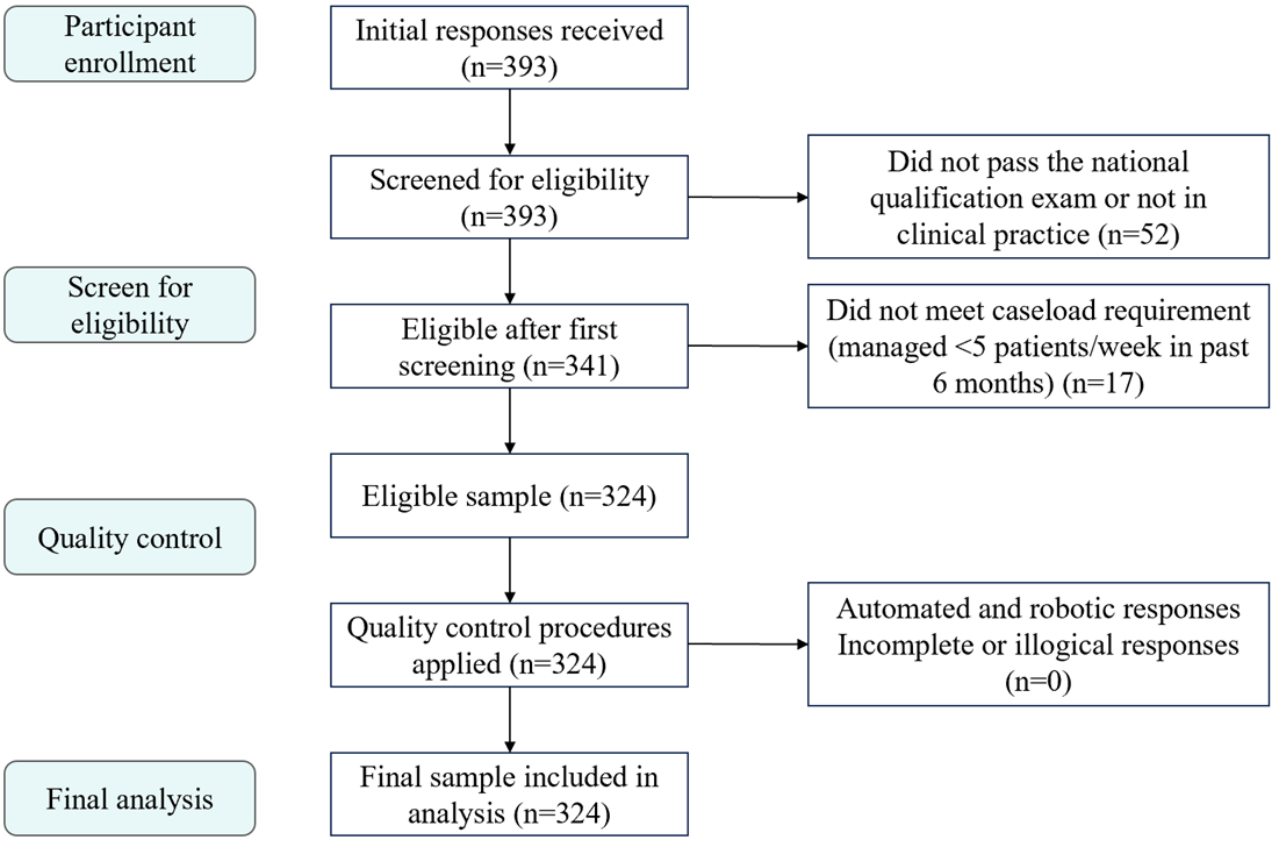

Participant Characteristics

A total of 324 certified rehabilitation professionals were included in the analysis. Most were aged 20 to 29 years (209/324, 64.5%) and female (173/324, 53.4%). Nearly half held the title rehabilitation therapist (158/324, 48.8%), with 26.9% (87/324) classified as senior rehabilitation therapists. Clinical experience was most commonly 2-5 years (157/324, 48.5%), and the majority held a bachelor's degree (209/324, 64.5%). Most participants worked within public-sector general hospital rehabilitation departments (268/324, 82.7%), and internet access was near-universal, primarily via mobile phone (322/324, 99.4%). Table 1 summarizes demographic characteristics. The geographic distribution of respondents is illustrated in Figure 2, with the highest proportions from Guangdong (77/324, 23.8%), Fujian (72/324, 22.2%), and Shanxi (45/324, 13.9%).

XSL•FO
RenderX

Table 1. Participant characteristics (N=324).

Characteristics	Participants, n (%)
Sex	
Male	151 (46.6)
Female	173 (53.4)
Age (years)	
20-29	209 (64.5)
30-39	96 (29.6)
40-49	18 (5.6)
50-59	1 (0.3)
Professional title	
Rehabilitation assistant	60 (18.5)
Rehabilitation therapist	158 (48.8)
Senior rehabilitation therapist	87 (26.9)
Associate chief rehabilitation therapist	16 (4.9)
Chief rehabilitation therapist	3 (0.9)
Years of clinical experience	
<2 years	52 (16)
2-5 year	157 (48.5)
6-10 years	68 (21)
11-15 years	27 (8.3)
16-20 years	13 (4)
>20 years	7 (2.2)
Highest education level	
Associate degree	44 (13.6)
Bachelor’s degree	209 (64.5)
Master’s degree	62 (19.1)
Doctoral degree	9 (2.8)
Work sector (institutional ownership)	
Private institution	55 (17)
Public institution	253 (78.1)
Both public and private	13 (4)
Other	3 (0.9)
Primary work setting (institution type)	
General hospital rehabilitation department	268 (82.7)
Rehabilitation specialty hospital	28 (8.6)
Community health center	6 (1.9)
Elder care institution	12 (3.7)
Sports team or sports medicine center	1 (0.3)
Other	9 (2.8)
Weekly clinical hours	
<5	14 (4.3)
6-10	64 (19.8)
11-20	11 (3.4)

XSL•FO
RenderX

Characteristics	Participants, n (%)
21-30	20 (6.2)
31-40	128 (39.5)
>40	87 (26.9)
Primary clinical area	
Neurological rehabilitation	138 (42.6)
Musculoskeletal rehabilitation	69 (21.3)
Cardiorespiratory rehabilitation	11 (3.4)
Pediatric rehabilitation	35 (10.8)
Geriatric rehabilitation	26 (8)
Trauma rehabilitation	2 (0.6)
Mental health rehabilitation	3 (0.9)
Speech, swallowing, and hearing rehabilitation	29 (9)
Occupational rehabilitation	2 (0.6)
Others	9 (2.8)
Province	
Fujian	72 (22.2)
Guizhou	2 (0.6)
Guangdong	77 (23)
Shanxi	45 (13.9)
Beijing	12 (3.7)
Henan	5 (1.5)
Sichuan	13 (4)
Jilin	4 (1.2)
Jiangxi	5 (1.5)
Shaanxi	10 (3.1)
Gansu	1 (0.3)
Shandong	21 (6.5)
Chongqing	4 (1.2)
Tianjin	2 (0.6)
Hubei	9 (2.8)
Shanghai	13 (4)
Liaoning	1 (0.3)
Anhui	8 (2.5)
Zhejiang	2 (0.6)
Jiangsu	18 (5.5)

XSL•FO
RenderX

Figure 2. Geographic distribution of survey respondents (N=324).

Collection of Subjective Clinical Information

Figure 3 illustrates a marked reliance on in-person communication, with 96.3% (312/324) reporting frequent or constant use of face-to-face discussions, far exceeding any other method. Digital communication channels were rarely incorporated into routine assessment. Only 14.2% (46/324) used e-messages, 14.5% (47/324) used phone calls, and 6.8% (22/324) used email with any regularity. Paper-based patient-reported outcome measures (PROMs) were far more established (202/324, 62.4% frequent use) than electronic formats (99/324, 30.6%), indicating that digital patient-reported data capture has not yet been scaled into routine practice.

Figure 3. Frequency of methods and tools used to collect subjective clinical information. EHR: electronic health record; PROM: patient-reported outcome measure.

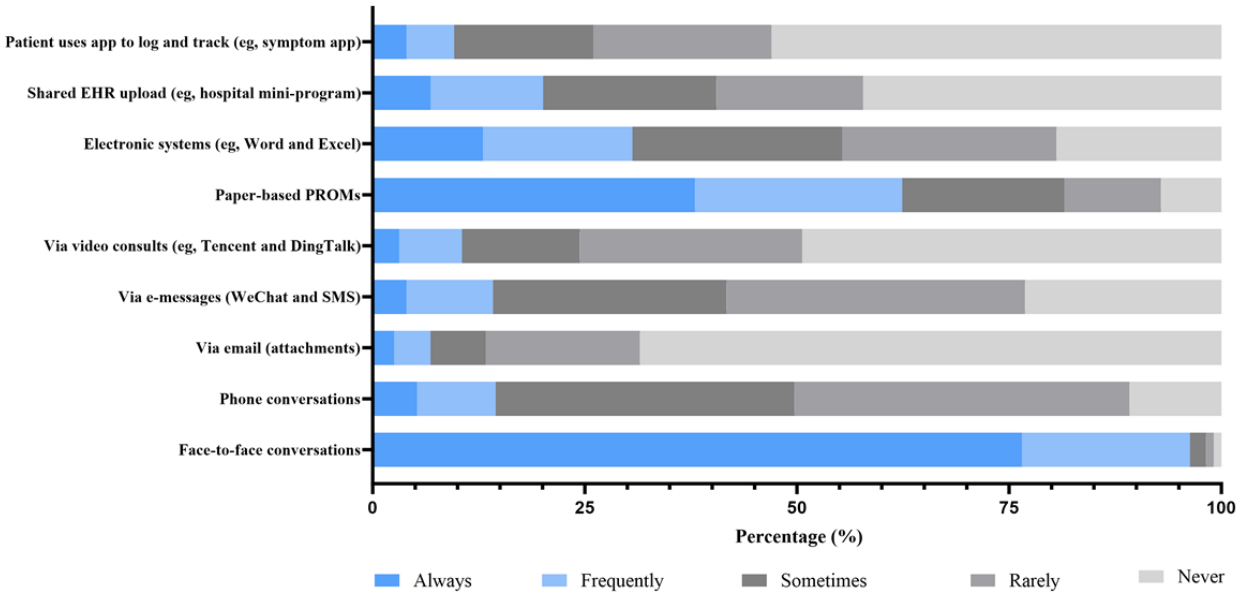

XSL•FO
RenderX

Collection of Objective Clinical Information

Figure 4 demonstrates similar dominance of traditional approaches for objective assessment. Visual observation was the primary method (271/324, 83.7% frequently or always), while manual measurement tools such as goniometers and dynamometers were also widely used (195/324, 60.2%). In contrast, digital motion-capture technologies, VR-based tools, wearable sensors, and patient-generated device data showed limited integration, with frequent use ranging between 4 to 14% across categories. More than 70% (227/324) rarely or never accessed wearable-generated metrics, and video-based assessment remained infrequent. These patterns reflect early-stage digital adoption rather than widespread integration of technology-supported movement analysis.

Figure 4. Frequency of methods and tools used to collect objective clinical information. VR: virtual reality.

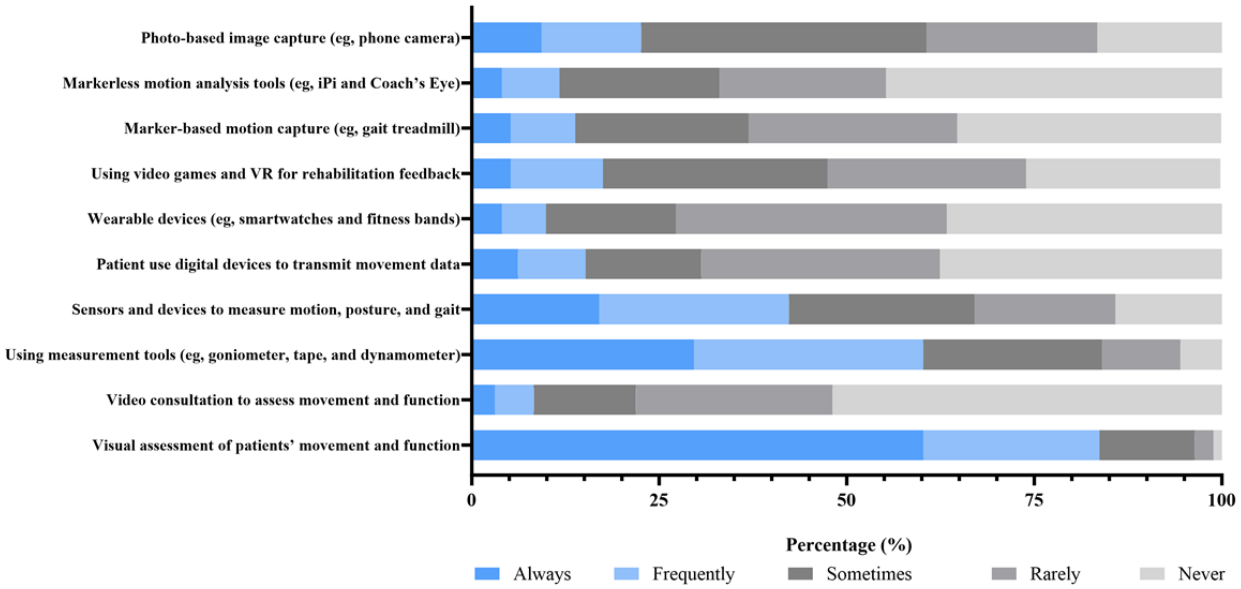

Documentation of Clinical Assessment Information

As displayed in Figure 5, handwritten free-text notes (266/324, 82.1%) and structured paper templates (195/324, 60.2%) formed the dominant documentation format. Unstructured EMR entry was more commonly used than structured coding (183/324, 56.5% vs 69/324, 21.3%), suggesting fragmented rather than standardized digital documentation. Microsoft Word or Microsoft Excel documentation was used occasionally (121/324, 37.4% combined frequent or occasional use), but mobile app–based entries, patient-entered digital notes, audio files, and video and image storage remained minimal, with most respondents reporting rare or no use. The predominance of paper systems suggests a slow transition toward digitalized information environments.

XSL•FO
RenderX

Figure 5. Frequency of methods and tools used to document clinical assessment information. EHR: electronic health record; EMR: electronic medical record; ICD-10: International Statistical Classification of Diseases, Tenth Revision; ICF: International Classification of Functioning, Disability and Health.

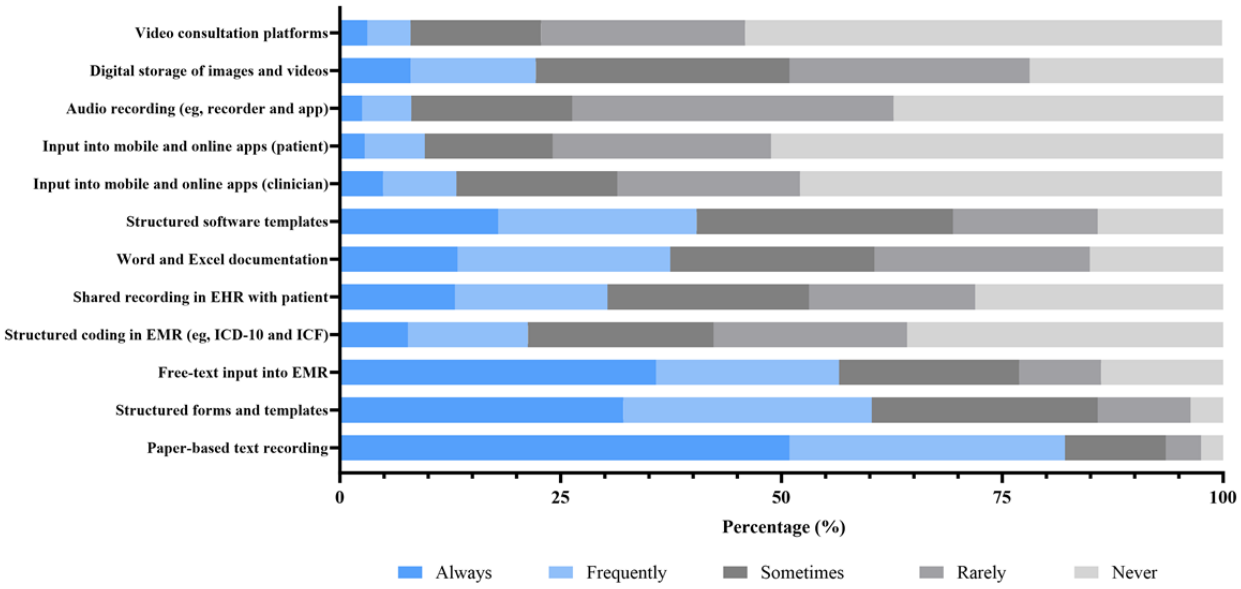

Willingness to Use Digital Technologies

Figure 6 shows generally high willingness across nearly all digital functions assessed. High proportions of respondents were willing to use digital systems for identity verification (261/324, 80.6%), progress tracking (256/324, 79.0%), and outcome measurement (253/324, 78.1%), all exceeding 75%. Willingness also exceeded 70% for accessing imaging (235/324, 72.5%), receiving clinical alerts (231/324, 71.3%), and recording treatment changes (228/324, 70.4%). Lower but still favorable levels were observed for teleconsultation (179/324, 55.2%) and digital referral coordination (185/324, 57.1%). The overall mean willingness score across these digital functions was 4.02, with moderate SD values (0.90-1.11), indicating consistent support for digital adoption within the cohort. Overall, expressed willingness was consistently higher than reported current use across digital domains, indicating a gap between readiness and practice.

XSL•FO
RenderX

Figure 6. Willingness to use digital technologies for rehabilitation-related clinical activities. CT: computed tomography; ICD-10: International Statistical Classification of Diseases, Tenth Revision; ICF: International Classification of Functioning, Disability and Health; MRI: magnetic resonance imaging.

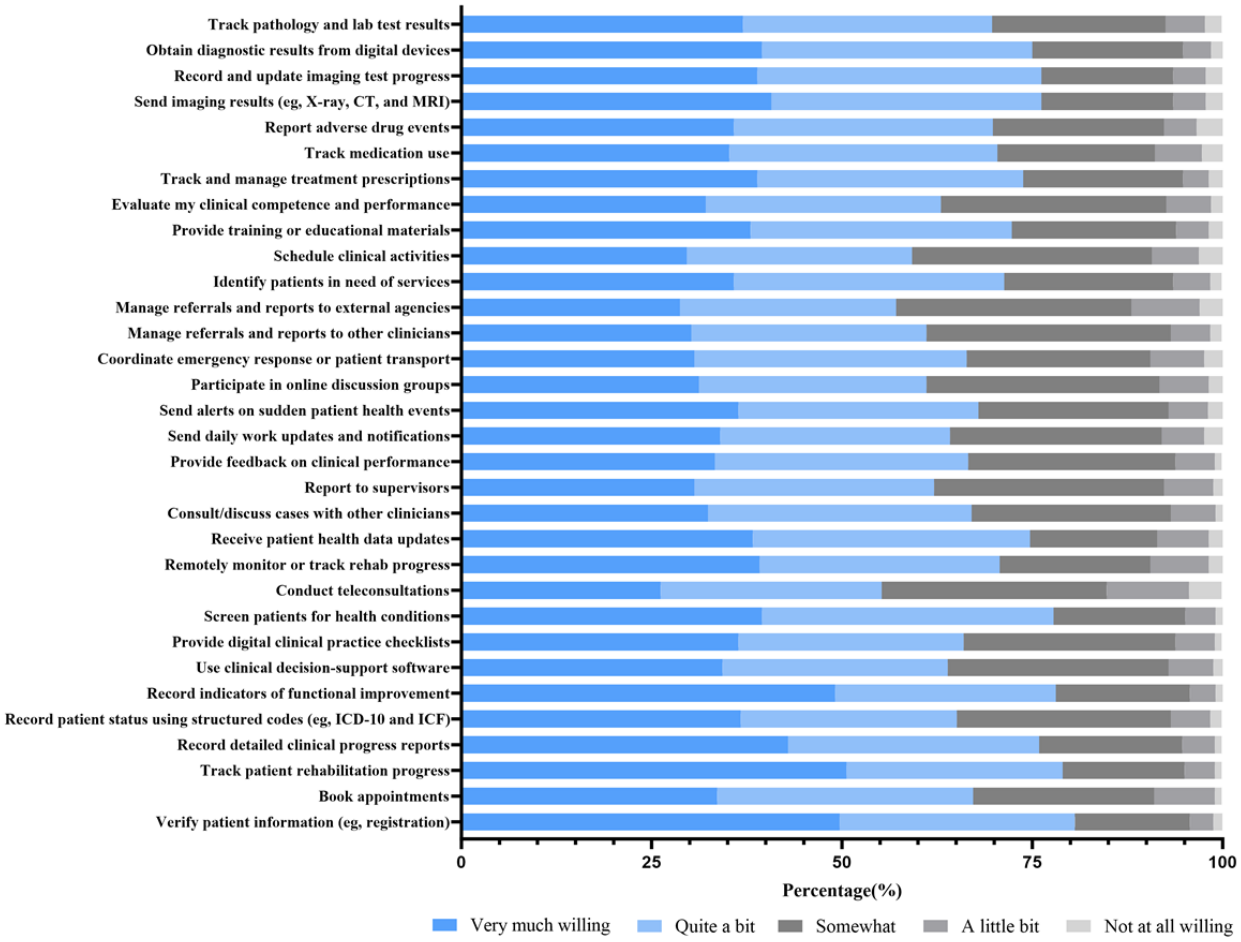

Subgroup Analyses by Demographic and Clinical Specialty

Subgroup analyses showed that educational level significantly influenced the adoption of advanced technologies, while professional title and clinical specialty had limited effects. Full details of the Kruskal-Wallis test results, including H values, *P* values, effect sizes, and post-hoc comparisons, are provided in Multimedia Appendix 4.

Findings revealed significant differences across educational levels for several outcomes. Sensor-based assessment (H=24.801, *P*<.001, ε^2=0.068), marker-based motion capture (H=23.052, *P*<.001, ε^2=0.063), wearable device data (H=13.695, *P*=.003, ε^2=0.033), electronic systems (H=10.823, *P*=.01, ε^2=0.024), and unstructured free-text EMR entry (H=12.214, *P*=.007, ε^2=0.029) all varied significantly, with post-hoc comparisons showing that therapists with master's or doctoral degrees consistently reported higher use than those with associate or bachelor's degrees. By professional title, significant differences were found for sensor-based assessment (H=18.904, *P*<.001, ε^2=0.047), wearable device data (H=13.142, *P*=.01, ε^2=0.029), unstructured free-text EMR entry (H=11.851, *P*=.02, ε^2=0.025), and teleconsultation willingness (H=11.802, *P*=.02, ε^2=0.024). Post-hoc analyses indicated that rehabilitation assistants and therapists generally reported lower use and willingness than senior or associate chief therapists. Across clinical specialties, limited differences were observed: sensor-based assessment (H=25.507, *P*=.002, ε^2=0.052) and face-to-face communication (H=25.134, *P*=.003, ε^2=0.051) showed significant variation, with speech and language therapists using sensors more frequently than neurological or musculoskeletal therapists, and mental health therapists using face-to-face communication less frequently than several other specialties. Unstructured free-text EMR entry (H=18.112, *P*=.03, ε^2=0.029) and willingness to send imaging results (H=18.075, *P*=.03, ε^2=0.029) reached overall significance, but no pairwise comparisons remained significant after correction. No other outcomes differed significantly across groups (all *P*>.05).

Summary of Open-Ended Responses

Open-ended responses were reviewed to provide additional context to the quantitative findings. Commonly reported barriers to digital health use included limited interoperability between information systems, continued reliance on paper-based workflows, insufficient institutional support, and time constraints in routine clinical practice. Respondents also highlighted perceived benefits of digital technologies, particularly in improving efficiency, data continuity, and long-term patient monitoring. These responses did not introduce themes that contradicted the quantitative results.

XSL•FO
RenderX

Discussion

Principal Findings

This cross-sectional survey provides new evidence on how rehabilitation professionals in China use DHTs in routine practice. Given that implementation of DHIs represents a cultural transformation of traditional health care [43], understanding the confidence and attitudes of health care professionals regarding the use of DHTs is a growing priority [44]. In this sample, in-person communication (312/324, 96.3%) and paper-based documentation (266/324, 82.1%) were the dominant methods for obtaining clinical information. In contrast, the adoption of digital tools remained limited across all domains. For instance, the frequent use of email (22/324, 6.8%), patient-initiated app-based data sharing (32/324, 9.6%), and video consultations for objective assessment (27/324, 8.3%) each fell below 10%. Despite this limited uptake, respondents expressed strong willingness to incorporate digital technologies into their clinical work, with more than 70% reporting positive attitudes toward most digital functions. This high readiness stands in stark contrast to the low levels of actual use, revealing a substantial willingness-practice gap. For example, although 79% (256/324) of respondents were willing to digitally track patient rehabilitation progress, only 30.6% (99/324) frequently used electronic systems to collect subjective information. Similarly, while 55.2% (179/324) expressed willingness to conduct teleconsultations, the actual frequent use of video-based objective assessment was only 8.3% (27/324). These findings underscore a significant opportunity to advance the digital transformation of rehabilitation services.

Previous research in other regions shows both alignment and divergence with our findings. An Australian survey using the original version of this instrument [37] reported widespread use of EMR-based free-text documentation and strong willingness to adopt digital tools for tasks such as scheduling, imaging tracking, and teleconsultation, yet therapists still relied predominantly on face-to-face communication for subjective assessment and visual observation for objective evaluation. In comparison, Chinese therapists showed similar dependence on direct communication but demonstrated broader, more experimental uptake of digital methods, including photo, video, and sensor-based assessment, while maintaining heavier reliance on paper-based records. Although willingness to use digital technologies was high in both contexts, the emphasis differed: Australian clinicians prioritized tools that streamline communication and workflow, whereas Chinese therapists expressed greater interest in outcome monitoring, identity verification, and progress tracking rather than patient-facing teleconsultation. These differences likely reflect cultural preferences for relationship-centered interaction, earlier digitization of health records in Australia, and variability in hospital information systems across Chinese provinces [45].

Findings from other regions further contextualize these patterns. An Italian survey reported high availability of digital equipment but low levels of perceived integration and knowledge, alongside marked regional disparities in resources and experience [46]. Similarly, Norwegian physiotherapists identified telephone consultation as the most commonly used digital communication modality, with limited uptake of video consultation, gaming, or VR technologies [47]. Studies from the Middle East, North Africa, and other middle-income settings likewise describe early-stage digital adoption within rehabilitation systems, characterized by strong enthusiasm but persistent barriers related to software availability, interoperability, digital literacy, and organizational readiness [48-50]. Collectively, this body of evidence suggests that globally, clinician enthusiasm for digital rehabilitation often outpaces the degree of practical implementation [51-53]. Our study extends this literature by providing empirical data from a rapidly digitizing health system, highlighting both shared challenges and substantial untapped potential.

The substantial willingness-practice gap observed in our study is consistent with recent findings from other rehabilitation settings in China. A survey of orthopedic rehabilitation medical staff found that while support for digital therapeutics was high, only 21% had actually used them, with concerns about privacy, interpersonal relationships, and time constraints limiting adoption [54]. Similarly, a study of autism spectrum disorder digital health services across 2 Chinese provinces reported that despite high demand among rehabilitation therapists (90.0%), actual use remained critically low (21.6%) [55]. Together, these findings suggest that the gap between readiness and implementation is not unique to our sample but reflects a broader systemic challenge across diverse rehabilitation disciplines in China. Our results also suggest that the strong demand for digital health training (234/324, 72.3%) highlights a critical leverage point. This finding aligns with previous research indicating that the absence of systematic, formal education on digital technologies limits the readiness of current and future health professionals for digital transformation [56]. It suggests that rehabilitation professionals are not only ready to adopt digital tools but are also actively seeking opportunities to develop the skills required for their effective implementation. Bridging this educational gap could accelerate the translation of high willingness into tangible clinical practice by integrating digital health competencies into entry-level curricula and establishing continuing professional development programs.

This study also examined the influence of educational background, professional title, and clinical specialty on digital health use patterns. Subgroup analyses revealed that educational level was the most significant factor influencing the adoption of advanced technologies, with therapists holding master's or doctoral degrees reporting significantly higher use of sensor-based assessment, motion capture technology, and wearable devices compared to those with associate or bachelor's degrees. In contrast, the influence of professional title was limited; although significant differences emerged in a few indicators, most post-hoc comparisons were nonsignificant after adjustment, and professional title had no effect on the vast majority of digital health functions. Similarly, differences across clinical specialties were minimal. The absence of differences across most domains suggests that readiness for digital rehabilitation is broadly shared across specialties, rather than being driven by discipline-specific factors. This pattern points toward the influence of system-level conditions rather than

XSL•FO
RenderX

professional background alone in shaping digital health adoption in rehabilitation practice. Insights from open-ended responses further support the interpretation that barriers to digital rehabilitation are primarily structural and organizational rather than related to clinician willingness. The implications of these findings are considerable. Given China's large population, rising burden of disabling conditions, and uneven geographic distribution of rehabilitation resources, the high level of willingness among therapists represents a critical opportunity [57-59].

Scaling up digital health could help extend services to underserved regions, improve continuity of care, reduce travel burden for patients, and partially mitigate workforce shortages [60,61]. Evidence from digital health implementation research indicates that the development of supportive policy environments, including reimbursement arrangements, data governance structures, and standardized documentation frameworks, can enable more durable adoption of technology in rehabilitation services [62-64]. Research in digital health education also highlights the value of structured training, with multiple evaluations reporting that integration of telehealth and digital assessment skills into both preservice and continuing professional programs improves clinician confidence and routine application [65,66]. Evidence from programs in musculoskeletal rehabilitation further suggests that remote teleconsultations can have the same effect as in-person physio consultations and can maintain clinical outcomes while expanding reach to populations with limited physical access [24]. Meanwhile, secondary analyses in our study revealed a positive trend in willingness to engage in teleconsultation across increasing professional titles, with higher-ranking therapists reporting greater willingness. This finding underscores the need for tiered digital training strategies tailored to rehabilitation professionals at different career stages, with particular attention to supporting lower- and mid-level practitioners. In addition, studies from low-resource regions show that software availability, device access, and internet reliability strongly influence digital function, with rural and underresourced settings experiencing the greatest barriers [67,68].

Taken together, this study demonstrates that limited use of DHTs in rehabilitation practice should not be interpreted as a lack of clinician readiness. Instead, our findings indicate a widespread willingness to engage with digital rehabilitation across specialties and professional levels, alongside persistent structural and organizational constraints that limit implementation. By providing evidence from China, this study contributes to a growing body of international research showing that the digital transformation of rehabilitation is primarily shaped by system-level conditions rather than individual clinician attitudes. These insights underscore the need for implementation strategies that move beyond profession-specific solutions and address the broader institutional, infrastructural, and policy environments in which rehabilitation services are delivered.

Limitations

This study has several limitations. This study has several limitations. Due to open-link recruitment, the response rate could not be calculated, precluding assessment of nonresponse bias. The online survey mode inherently introduced selection bias by systematically excluding professionals with limited internet access or digital skills, and participation itself implied a certain level of comfort with digital technologies. This self-selection mechanism likely overestimates the overall population's willingness and actual use levels, particularly for those in remote areas or primary care settings. Geographic distribution bias further limits generalizability, as respondents were predominantly from several provinces, with underrepresentation from central and western regions and primary care institutions. Additionally, all data were based on self-reports, which may be subject to social desirability and recall biases. The cross-sectional design captures technology use at a single time point and cannot reveal dynamic changes in adoption or causal pathways, underscoring the need for longitudinal studies.

Conclusions

In conclusion, clinical practice in China continues to rely primarily on traditional in-person and paper-based methods, yet clinicians express clear readiness to adopt digital health. This readiness indicates significant potential for targeted policy actions, education programs, and implementation planning. Addressing practical and structural barriers could accelerate digital transformation in rehabilitation and support improved access and more sustainable delivery of rehabilitation services.

Acknowledgments

The authors are grateful to those who participated in this study. We sincerely thank Professor Mark Merolli from the University of Melbourne for his valuable suggestions and feedback during the preparation of this study. All authors contributed to drafting the manuscript or revising it critically for important intellectual content and approved the final version for submission. The authors declare that no generative artificial intelligence was used in the creation of this manuscript.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Funding

This work was supported by the National Social Science Fund of China (Grant No. 23BTY117), which funded the corresponding author GN.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: SD (equal), ZRW (equal), GN (equal)
Data curation: SD (equal), ZRW (equal), XC (equal)
Formal analysis: SD (equal), ZRW (equal)
Funding acquisition: GN
Investigation: SD (equal), XC (equal)
Methodology: SD (equal), ZRW (equal)
Project administration: SD (lead), ZRW (equal), XC (supporting)
Resources: GN, ZRW
Supervision: GN
Visualization: SD (supporting), ZRW (lead)
Writing – original draft: SD (lead), ZRW (supporting)
Writing – review and editing: SD (lead), ZRW (supporting), GN (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
STROBE checklist.
[DOCX File , 34 KB - jmir_v28i1e90623_app1.docx]

Multimedia Appendix 2
Survey instrument for rehabilitation therapists on digital health applications (Chinese version).
[DOCX File , 35 KB - jmir_v28i1e90623_app2.docx]

Multimedia Appendix 3
Survey instrument for rehabilitation therapists on digital health applications (Chinese back-translation to English for review).
[DOCX File , 32 KB - jmir_v28i1e90623_app3.docx]

Multimedia Appendix 4
Subgroup comparisons of digital health technology use by educational level, professional title, and clinical specialty.
[DOCX File , 19 KB - jmir_v28i1e90623_app4.docx]

References

1. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the global burden of disease study 2019: a systematic analysis for the global burden of disease study 2019. Lancet 2021;396(10267):2006-2017 [FREE Full text] [doi: 10.1016/S0140-6736(20)32340-0] [Medline: 33275908]
2. Jette AM. Global prevalence of disability and need for rehabilitation. Phys Ther 2021;101(2):pzab004. [doi: 10.1093/ptj/pzab004] [Medline: 33556184]
3. To Dutka J, Gans BM, Bracciano A, Bharadwaj S, Akinwuntan A, Mauk K, et al. Delivering rehabilitation care around the world: voices from the field. Arch Phys Med Rehabil 2023;104(9):1385-1393 [FREE Full text] [doi: 10.1016/j.apmr.2023.03.009] [Medline: 36997079]
4. Taylor R, Zwisler AD, Uddin J. Global health-care systems must prioritise rehabilitation. Lancet 2021;396(10267):1946-1947 [FREE Full text] [doi: 10.1016/S0140-6736(20)32533-2] [Medline: 33275909]
5. Krahn GL. WHO world report on disability: a review. Disabil Health J 2011;4(3):141-142. [doi: 10.1016/j.dhjo.2011.05.001] [Medline: 21723520]
6. Rehabilitation in health systems. World Health Organization. 2017. URL: https://www.who.int/publications/i/item/9789241549974 [accessed 2023-05-03]
7. Rehabilitation competency framework. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240008281 [accessed 2026-03-06]
8. Rehabilitation: key for health in the 21st century. World Health Organization. 2017. URL: https://www.who.int/publications/i/item/9789241515115 [accessed 2026-03-06]
9. Basic standards and management guidelines for rehabilitation medical centers (Trial) (Guoweiyifa [2017] No. 51). National Health and Family Planning Commission. 2017. URL: http://www.nhfpc.gov.cn/yzygj/s3577/201711/fac102fd386a41f1ab545315d7c26045.shtml [accessed 2026-03-06]

XSL•FO
RenderX

10. Notice on issuing opinions on accelerating the development of rehabilitation medicine [Web Page in Chinese]. National Health Commission of the People's Republic of China. 2021. URL: https://www.gov.cn/zhengce/zhengceku/2021-06/17/content_5618767.htm [accessed 2026-03-06]
11. Rehabilitation. World Health Organization. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/rehabilitation [accessed 2026-03-20]
12. Htwe O, Yuliawiratman BS, Tannor AY, Nor Asikin MZ, Soh E, DE Groote W, et al. Barriers and facilitators for increased accessibility to quality rehabilitation services in low- and middle- income countries: a systematic review. Eur J Phys Rehabil Med 2024;60(3):514-522 [FREE Full text] [doi: 10.23736/S1973-9087.24.08154-1] [Medline: 38551518]
13. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the global burden of disease study 2019: a systematic analysis for the global burden of disease study 2019. Lancet 2021;396(10267):2006-2017 [FREE Full text] [doi: 10.1016/S0140-6736(20)32340-0] [Medline: 33275908]
14. Chinese Physical Therapy Association. World Physiotherapy. 2023. URL: https://world.physio/membership/china [accessed 2026-03-06]
15. Yisma E, Versace VL, Jones M, Walsh S, Jones S, May E, et al. The distribution of registered occupational therapists, physiotherapists, and podiatrists in Australia. PLoS One 2023;18(9):e0291962 [FREE Full text] [doi: 10.1371/journal.pone.0291962] [Medline: 37733814]
16. Australian Health Practitioner Regulation Agency. Physiotherapists - a snapshot as at 30 June 2023. Ahpra & National Boards. Melbourne, Australia: Australian Health Practitioner Regulation Agency; 2023. URL: https://www.ahpra.gov.au/Publications/Registrant-data/Physiotherapists.aspx [accessed 2026-03-06]
17. Transcript of the press conference on October 28, 2020. National Health Commission of the PRC. 2020. URL: https://www.nhc.gov.cn/xcs/c100122/202010/96cb6d028e5f4c9380059bc7ac83bb21.shtml [accessed 2020-08-01]
18. Jin Y, Hu J. The current status, challenges and future of medical and health data sharing in China: lessons from the European health data space. Eur J Health Law 2025;32(3):239-280. [doi: 10.1163/15718093-bja10147] [Medline: 40541395]
19. Li X, Cong Y. A systematic literature review of ethical challenges related to medical and public health data sharing in China. J Empir Res Hum Res Ethics 2021;16(5):537-554. [doi: 10.1177/15562646211040299] [Medline: 34516325]
20. Li Q, Cheng F, Zeng H, Xu J. Health insurance payment for telehealth services: scoping review and narrative synthesis. J Med Internet Res 2024;26:e56699 [FREE Full text] [doi: 10.2196/56699] [Medline: 39652868]
21. Wu G, Gong M, Wu Y, Liu L, Shi B, Zeng Z. Advancing digital health in China: aligning challenges, opportunities, and solutions with the global initiative on digital health (GIDH). Health Care Sci 2024;3(5):365-369. [doi: 10.1002/hcs2.118] [Medline: 39479272]
22. World Health Organization. Global Strategy on Digital Health 2020-2025. Geneva: World Health Organization; 2021.
23. Kizilkilic SE, Ramakers W, Falter M, Scherrenberg M, Bonneux C, Pieters Z, et al. A digitally-supported shared decision making approach for patients during cardiac rehabilitation: a randomized controlled trial. Eur J Prev Cardiol 2025:zwaf537. [doi: 10.1093/eurjpc/zwaf537] [Medline: 40874533]
24. Hinman RS, Campbell PK, Kimp AJ, Russell T, Foster NE, Kasza J, et al. Telerehabilitation consultations with a physiotherapist for chronic knee pain versus in-person consultations in Australia: the PEAK non-inferiority randomised controlled trial. Lancet 2024;403(10433):1267-1278. [doi: 10.1016/S0140-6736(23)02630-2] [Medline: 38461844]
25. Grochtdreis T, König HH, Renner J, Sehner S, Dekker A, Briken P, et al. Cost-effectiveness of an interdisciplinary, internet-based transgender health care program in Germany: economic evaluation alongside a randomized controlled trial. J Med Internet Res 2025;27:e66371 [FREE Full text] [doi: 10.2196/66371] [Medline: 40537093]
26. Lee YK, Yoon EJ, Kim TH, Kim JI, Kim JH. Musculoskeletal digital therapeutics and digital health rehabilitation: a global paradigm shift in orthopedic care. J Clin Med 2025;14(23):8467 [FREE Full text] [doi: 10.3390/jcm14238467] [Medline: 41375770]
27. Al-Jabr H, Salt E, Stephenson J, Hamdan E, Helliwell T. Can remote digital health interventions improve rehabilitation following major trauma? Results of a systematic review. Arch Phys Med Rehabil 2025;S0003-9993(25):01002 [FREE Full text] [doi: 10.1016/j.apmr.2025.10.012] [Medline: 41187799]
28. Gu B, Kim HS, Kim H, Yoo J. Advancements in wearable sensor technologies for health monitoring in terms of clinical applications, rehabilitation, and disease risk assessment: systematic review. JMIR Mhealth Uhealth 2026;14:e76084 [FREE Full text] [doi: 10.2196/76084] [Medline: 41511829]
29. Wang Q, Markopoulos P, Yu B, Chen W, Timmermans A. Interactive wearable systems for upper body rehabilitation: a systematic review. J Neuroeng Rehabil 2017;14(1):20 [FREE Full text] [doi: 10.1186/s12984-017-0229-y] [Medline: 28284228]
30. Gooch HJ, Jarvis KA, Stockley RC. Behavior change approaches in digital technology-based physical rehabilitation interventions following stroke: scoping review. J Med Internet Res 2024;26:e48725 [FREE Full text] [doi: 10.2196/48725] [Medline: 38656777]
31. Jeilani A, Hussein A. Impact of digital health technologies adoption on healthcare workers' performance and workload: perspective with DOI and TOE models. BMC Health Serv Res 2025;25(1):271 [FREE Full text] [doi: 10.1186/s12913-025-12414-4] [Medline: 39966874]

XSL•FO
RenderX

32. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, Martinez EZ, Zucoloto ML, Østengaard L, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023 Sep 18;6(1):161. [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
33. Weik L, Fehring L, Mortsiefer A, Meister S. Understanding inherent influencing factors to digital health adoption in general practices through a mixed-methods analysis. NPJ Digit Med 2024 Feb 27;7(1):47. [doi: 10.1038/s41746-024-01049-0] [Medline: 38413767]
34. Sylla B, Ismaila O, Diallo G. 25 years of digital health toward universal health coverage in low- and middle-income countries: rapid systematic review. J Med Internet Res 2025;27:e59042 [FREE Full text] [doi: 10.2196/59042] [Medline: 40440696]
35. Portz J, Moore S, Bull S. Evolutionary trends in the adoption, adaptation, and abandonment of mobile health technologies: viewpoint based on 25 years of research. J Med Internet Res 2024;26:e62790 [FREE Full text] [doi: 10.2196/62790] [Medline: 39331463]
36. Cuschieri S. The STROBE guidelines. Saudi J Anaesth 2019;13(Suppl 1):S31-S34 [FREE Full text] [doi: 10.4103/sja.SJA_543_18] [Medline: 30930717]
37. Merolli M, Gray K, Choo D, Lawford BJ, Hinman RS. Use, and acceptability, of digital health technologies in musculoskeletal physical therapy: a survey of physical therapists and patients. Musculoskeletal Care 2022;20(3):641-659 [FREE Full text] [doi: 10.1002/msc.1627] [Medline: 35278266]
38. Classification of digital health interventions v1.0: a shared language to describe the uses of digital technology for health. World Health Organization. 2018. URL: https://www.who.int/publications/i/item/WHO-RHR-18.06 [accessed 2026-03-20]
39. Description of physical therapy: policy statement. World Physiotherapy. 2019. URL: https://world.physio/sites/default/files/2020-04/PS-2019-Description-of-physical-therapy.pdf [accessed 2026-03-20]
40. Definitions of occupational therapy from member organisations. World Federation of Occupational Therapists. 2012. URL: https://wfot.org/resources/definitions-of-occupational-therapy-from-member-organisations [accessed 2026-03-20]
41. Assessment guideline for international classification of functioning, disability and health-rehabilitation set (ICF-RS) (T/CARM 001?2020). Chinese Association of Rehabilitation Medicine. 2020. URL: https://www.ttbz.org.cn/StandardManage/Detail/41167/ [accessed 2026-03-07]
42. Li J, Chen L. Guidelines for the standardization of rehabilitation medicine association standards (2025 Edition). Chin J Rehab Med 2025;40(1):1. [doi: 10.3870/j.issn.1001-1242.2025.01.001]
43. Meskó B, Drobni Z, Bényei É, Gergely B, Győrffy Z. Digital health is a cultural transformation of traditional healthcare. Mhealth 2017;3:38 [FREE Full text] [doi: 10.21037/mhealth.2017.08.07] [Medline: 29184890]
44. Kuek A, Hakkennes S. Healthcare staff digital literacy levels and their attitudes towards information systems. Health Informatics J 2020;26(1):592-612 [FREE Full text] [doi: 10.1177/1460458219839613] [Medline: 30983476]
45. National digital health strategy advancing care for all Australians. Australian Digital Health Agency. 2025. URL: https://www.digitalhealth.gov.au/newsroom/media/national-digital-health-strategy-advancing-care-for-all-australians [accessed 2026-03-05]
46. Capecci M, Gandolfi M, Straudi S, Calabrò RS, Baldini N, Pepa L, et al. Shaping the future: an Italian survey unveils the unmet need to empower physical medicine and rehabilitation professionals with technological skills. Eur J Phys Rehabil Med 2024;60(3):540-543 [FREE Full text] [doi: 10.23736/S1973-9087.24.08376-X] [Medline: 38618695]
47. Martinsen L, Østerås N, Moseng T, Tveter AT. Usage, attitudes, facilitators, and barriers toward digital health technologies in musculoskeletal care: survey among primary care physiotherapists in Norway. JMIR Rehabil Assist Technol 2024;11:e54116 [FREE Full text] [doi: 10.2196/54116] [Medline: 39283661]
48. El-Jardali F, Bou-Karroum L, Jabbour M, Bou-Karroum K, Aoun A, Salameh S, et al. Digital health in fragile states in the middle east and north Africa (MENA) region: a scoping review of the literature. PLoS One 2023;18(4):e0285226 [FREE Full text] [doi: 10.1371/journal.pone.0285226] [Medline: 37115778]
49. Qasrawi R, Al Sabbah H, Issa G, Thwib S, Amro M, Atari S, et al. Assessing the impact of digital health literacy on health management practices in Arab middle eastern and north African countries: insights from predictive modeling. Front Digit Health 2025;7:1555436 [FREE Full text] [doi: 10.3389/fdgth.2025.1555436] [Medline: 41112210]
50. Aljabri NQ, Bulkeley K, Cusick A. Telerehabilitation in the middle east North Africa Region: A structured review. Int J Telerehabil 2021;13(2):e6401 [FREE Full text] [doi: 10.5195/ijt.2021.6401] [Medline: 35646235]
51. Yew SQ, Trivedi D, Adanan NIH, Chew BH. Facilitators and barriers to the implementation of digital health technologies in hospital settings in lower- and middle-income countries since the onset of the COVID-19 pandemic: scoping review. J Med Internet Res 2025;27:e63482 [FREE Full text] [doi: 10.2196/63482] [Medline: 40053793]
52. Nicora G, Parimbelli E, Mauro MC, Falchini F, Germanotta M, Fasano A, et al. Healthcare practitioners and robotic-assisted rehabilitation: understanding needs and barriers. J Neuroeng Rehabil 2025;22(1):78 [FREE Full text] [doi: 10.1186/s12984-025-01593-0] [Medline: 40205389]
53. Scott Kruse C, Karem P, Shifflett K, Vegi L, Ravi K, Brooks M. Evaluating barriers to adopting telemedicine worldwide: a systematic review. J Telemed Telecare 2018;24(1):4-12 [FREE Full text] [doi: 10.1177/1357633X16674087] [Medline: 29320966]

XSL•FO
RenderX

54. Liu FT, Chong T, Zhang K, Shi X, Mu J, Wang G, et al. Awareness of and willingness to use digital therapeutics among orthopedic rehabilitation medical staff: a questionnaire survey. Chinese Journal of Clinicians (Electronic Edition) 2024;18(12):1125-1131. [doi: 10.3877/cma.j.issn.1674-0785.2024.12.008]
55. Zou M, Gong X, Feng L, Li S, Lu C, Liu Z, et al. Regional disparities in the use and demand for digital health services for autism spectrum disorder in China: cross-sectional survey of stakeholder perspectives. J Med Internet Res 2025;27:e77157 [FREE Full text] [doi: 10.2196/77157] [Medline: 41043146]
56. Zhang Y, Liu J. Does education affect economic growth? A re-examination of empirical data from China. Sustainability 2022 Dec 06;14(23):16289. [doi: 10.3390/su142316289]
57. Chen X, Giles J, Yao Y, Yip W, Meng Q, Berkman L, et al. The path to healthy ageing in China: a peking university-lancet commission. Lancet 2022;400(10367):1967-2006 [FREE Full text] [doi: 10.1016/S0140-6736(22)01546-X] [Medline: 36423650]
58. Capecci M, Gandolfi M, Straudi S, Calabrò RS, Baldini N, Pepa L, et al. Advancing public health through technological rehabilitation: insights from a national clinician survey. BMC Health Serv Res 2024;24(1):1626 [FREE Full text] [doi: 10.1186/s12913-024-11991-0] [Medline: 39702315]
59. Jones SE, Campbell PK, Kimp AJ, Bennell K, Foster NE, Russell T, et al. Evaluation of a novel e-learning program for physiotherapists to manage knee osteoarthritis via telehealth: qualitative study nested in the PEAK (physiotherapy exercise and physical activity for knee osteoarthritis) randomized controlled trial. J Med Internet Res 2021;23(4):e25872 [FREE Full text] [doi: 10.2196/25872] [Medline: 33929326]
60. Walle AD, Butta FW, Kassie SY, Chereka AA, Kanfe SG, Dubale AT, et al. Healthcare professionals' attitude to using mobile health technology and its associated factors in a resource-limited country-an implication for digital health implementers: a cross sectional study. Biomed Res Int 2024;2024:1631376 [FREE Full text] [doi: 10.1155/2024/1631376] [Medline: 39035773]
61. Wosny M, Strasser LM, Hastings J. Experience of health care professionals using digital tools in the hospital: qualitative systematic review. JMIR Hum Factors 2023;10:e50357 [FREE Full text] [doi: 10.2196/50357] [Medline: 37847535]
62. Gordon WJ, Landman A, Zhang H, Bates DW. Beyond validation: getting health apps into clinical practice. NPJ Digit Med 2020;3:14. [doi: 10.1038/s41746-019-0212-z] [Medline: 32047860]
63. Elser A, Lange M, Kopkow C, Schäfer AG. Barriers and facilitators to the implementation of virtual reality interventions for people with chronic pain: scoping review. JMIR XR Spat Comput 2024;1:e53129. [doi: 10.2196/53129] [Medline: 41341873]
64. Krahe MA, Larkins SL, Adams N. Digital health implementation in Australia: a scientometric review of the research. Digit Health 2024;10:20552076241297729 [FREE Full text] [doi: 10.1177/20552076241297729] [Medline: 39539722]
65. Rettinger L, Maul L, Putz P, Ertelt-Bach V, Huber A, Javorszky SM, et al. Telehealth acceptance and perceived barriers among health professionals: pre-post evaluation of a web-based telehealth course. JMIR Hum Factors 2025;12:e74107 [FREE Full text] [doi: 10.2196/74107] [Medline: 40902201]
66. Davies L, Hinman RS, Russell T, Lawford B, Bennell K. An international core capability framework for physiotherapists delivering telephone-based care. J Physiother 2022;68(2):136-141 [FREE Full text] [doi: 10.1016/j.jphys.2022.02.002] [Medline: 35283052]
67. Dukuzimana MJ, Ukwizabigira JB, Muhire C, Jokinen K, Tumusiime DK, Murtonen KP, et al. Rehabilitation needs at rural primary health care settings: perspectives of health center nurses in burera district of Rwanda. BMC Prim Care 2025;26(1):224. [doi: 10.1186/s12875-025-02921-y] [Medline: 40660097]
68. Motiwala F, Ezezika O. Barriers to scaling health technologies in sub-Saharan Africa: Lessons from Ethiopia, Nigeria, and Rwanda. Afr J Sci Technol Innovation Devel 2021 Nov 28;14(7):1788-1797. [doi: 10.1080/20421338.2021.1985203]

Abbreviations

DHI: digital health interventions
DHT: digital health technology
NHC: National Health Commission
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
VR: virtual reality
WHO: World Health Organization

XSL•FO
RenderX

Edited by M Balcarras; submitted 31.Dec.2025; peer-reviewed by W Qin, Z Liu, O Ndibe, Y Hu; comments to author 12.Feb.2026; revised version received 10.Mar.2026; accepted 17.Mar.2026; published 09.Apr.2026.

Please cite as:
Duan S, Wang ZR, Chen X, Ni G
Digital Health Technology Use Among Rehabilitation Professionals in China: Multi-Province Cross-Sectional Survey
J Med Internet Res 2026;28:e90623
URL: https://www.jmir.org/2026/1/e90623
doi:10.2196/90623
PMID:

©Shuning Duan, Zi-Ru Wang, Xier Chen, Guoxin Ni. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Understanding User Intent in Code-Mixed Sexual and Reproductive Health Queries in Urban India: Hierarchical Classification Approach Using Large Language Models

Sumon Kanti Dey[1], BSc; Manvi S[1], MPH; Aradhana Thapa[2], MPH; Meet Shah[3], MS; Zeel Mehta[3], MSc; Shraddha Kale Kapile[3], MSW; Tanvi Divate[3], BA; Suhani Jalota[4], PhD; Azra Ismail[1], PhD

[1]Department of Biomedical Informatics, School of Medicine, Emory University, 101 Woodruff Circle, Atlanta, GA, United States

[2]Hubert Department of Global Health, Rollins School of Public Health, Emory University, Atlanta, GA, United States

[3]Myna Mahila Foundation, Mumbai, India

[4]Hoover Institution, Stanford University, Stanford, CA, United States

Corresponding Author:
Sumon Kanti Dey, BSc
Department of Biomedical Informatics, School of Medicine, Emory University, 101 Woodruff Circle, Atlanta, GA, United States

Abstract

Background: Sexual and reproductive health (SRH) remains a stigmatized and taboo topic globally, limiting access to reliable information. These challenges are heightened in the Global South, where linguistic and cultural diversity further complicates information access. In India (the study context), many individuals express SRH concerns in code-mixed language, such as Hinglish (code-mixed Hindi and English), and use colloquial terms. Large language models (LLMs) could help answer SRH questions, but most are trained for English and may perform poorly on code-mixed text and miss cultural nuances. Our research aims to address this gap by assessing the current state of LLMs in understanding user intent in SRH queries for a low-resource language.

Objective: We evaluate the effectiveness of proprietary, multilingual open-weight, and Indic LLMs in zero-shot settings for identifying user intent in code-mixed Hinglish SRH queries. Our goal is to assess how well LLMs assign correct labels in a 2-level hierarchical classification (topic and subtopic). We take a hierarchical approach because SRH queries are complex and context-dependent; flat labels may obscure clinically important distinctions and lead to misdirected guidance. We also characterize common error types driving misclassification.

Methods: We analyzed 4161 deidentified questions about SRH in Hinglish, collected by our partner nonprofit organization (Myna Mahila Foundation) in an underserved community in urban Mumbai. Queries were annotated into 8 topics and 40 subtopics using a hierarchical framework that captured linguistic, cultural, and contextual variation. We evaluated proprietary, multilingual open-weight, and Indic-specific LLMs in zero-shot settings. Performance was measured using hierarchical F_1 (hF_1), Exact Match, and topic- and subtopic-level accuracy.

Results: Proprietary models achieved the strongest results, with GPT-5 performing best overall (hF_1= 0.784). Among open-weight systems, Sarvam-M emerged as the top-performing Indic model (hF_1=0.757), ranking just below the top-performing proprietary model and performing comparably with Claude-3.5-Sonnet (0.745; Anthropic) as well as large multilingual systems such as Llama-3.3-70B-Instruct (0.742; Meta) and Gemma-3-27B-IT (0.739; Google). Other Indic models performed considerably lower (eg, Llama-3-Gaja-Hindi-8B [0.596; CognitiveLab], Krutrim-2-Instruct [0.558; OLA Krutrim Team], and Airavata [0.404; AI4Bharat]). Smaller multilingual open-weight models, including Mixtral-8 × 7B-Instruct (0.593), Llama-3.1-8B-Instruct (0.630), Gemma-2-9B-IT (0.657), consistently outperformed them, showing that parameter size alone does not explain performance gaps. While models generally captured broad topical intent, they frequently failed at fine-grained intent recognition, especially with euphemisms, colloquial expressions, and locally or culturally situated questions.

Conclusions: Hierarchical classification revealed persistent gaps in how LLMs handle code-mixed queries. Proprietary models performed best, but Sarvam-M shows that open-weight Indic systems can achieve performance near state-of-the-art models when supported by robust training data, cultural adaptation, and appropriate scale. These findings highlight the potential of localized, culturally aligned models to advance linguistically inclusive artificial intelligence tools and expand equitable access to SRH information in underserved populations globally.

(J Med Internet Res 2026;28:e86545) doi:10.2196/86545

KEYWORDS

sexual and reproductive health; large language models; code-mixing; Hinglish; hierarchical classification; conversational agents

XSL•FO
RenderX

Introduction

Background

Recent advancements in large language models (LLMs) present an opportunity to address significant gaps in health care information delivery. LLMs could be leveraged to simplify complex medical information, respond to patient queries, and enhance health literacy among the general population [1,2]. Despite these advancements, there remain significant disparities in the performance of these models across languages, especially in health care tasks [3]. Most LLMs are predominantly centered on the English language [4,5]. They can fail to recognize local dialects, cultural nuances, and speaking patterns, especially for non-English speaking populations that are less represented online. Health communication is further shaped by social dynamics, including gender, educational status, functional literacy, and cultural context [6-8]. Our research focuses on addressing these challenges in the context of India by evaluating the performance of LLMs in detecting user intent in the context of sexual and reproductive health (SRH).

SRH presents unique challenges for health information delivery, given that stigma, misinformation, and social barriers can restrict individuals—especially women—from seeking reliable health information and medical support [9,10]. For instance, 78% of the 15 million abortions in India take place outside medical facilities, highlighting the need for better access to reproductive health services and information [11]. Deep-rooted societal taboos also prevent open discussions about sex-related topics, exacerbating barriers to SRH awareness and services [12]. At the same time, SRH is a time-sensitive domain in which delays can have serious consequences. In India, lack of early pregnancy care leads to undetected and unmanaged conditions such as anemia, diabetes, hypertension, and infections, which are significant causes of pregnancy loss and maternal mortality [13,14].

While a multitude of online platforms exist where individuals can engage with health care professionals or even community members (eg, Reddit; Reddit, Inc), the reach of these platforms remains low in underserved communities in India. The absence of such inclusive platforms for linguistically diverse, resource-constrained communities further contributes to health inequities, leaving many women without access to critical SRH-related knowledge and services. Prior work by Wang et al [15] has shown the potential of a rule-based conversational chatbot for SRH support among young people in India (SnehAI), demonstrating strong user engagement and information-seeking behavior. We build on this work by evaluating the potential of LLMs to support the understanding of user intent in code-mixed SRH queries for such interventions.

Adding to these challenges is the extensive linguistic diversity within India, which significantly influences communication patterns, particularly in informal and colloquial settings. A notable linguistic phenomenon prevalent across India is code-mixing, the practice of blending multiple languages within a single conversation or utterance. We focus on Hinglish—a popular form of code-mixing involving Hindi and English, where individuals use the phonetic Latin script instead of the Devanagari script to write Hindi words [16]. This was often the preferred mode of typing for the population on which we focus in this study and has been documented in prior research in other Hindi-speaking populations in India [17,18]. Code-mixing remains a long-standing challenge in natural language understanding research, with several publicly available LLMs still struggling to interpret and generate code-switched text [19,20].

Another significant challenge is the inherently layered and context-dependent nature of questions that individuals frequently ask about SRH. Broader concerns related to pregnancy may branch into distinct subtopics such as antepartum emergency, postpartum pain, infertility, or abortion, reflecting the complex structure of real-world health inquiries. Traditional flat classification approaches may collapse these distinctions [21], leading to the generation of misleading responses that may fail to capture critical aspects of care and support. Our research aims to address this gap by taking a hierarchical classification approach to understanding user intent in SRH queries (Figure 1).

Figure 1. Hierarchical sexual and reproductive health (SRH) intent classification framework. This figure illustrates our approach to processing a code-mixed SRH user query (an English translation of the original Hinglish text is shown for clarity) using a large language model (LLM) to infer intent. The model maps the query to both a topic and its corresponding subtopic, producing structured intent classification. JSON: JavaScript Object Notation.

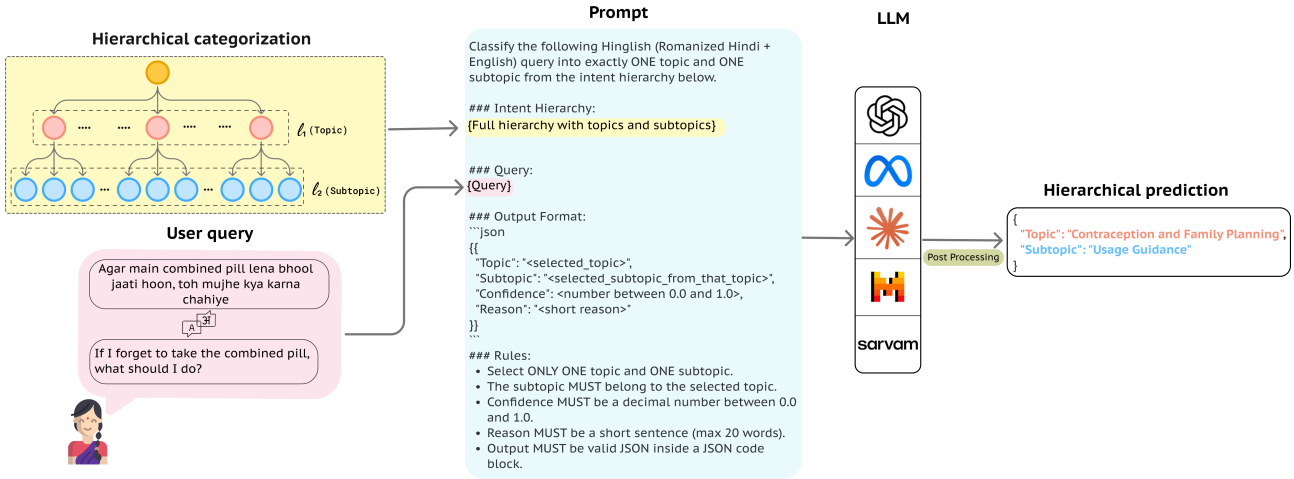

XSL•FO
RenderX

In summary, this study contributes the following: (1) we designed a hierarchical intent structure for SRH queries with 2 levels (topic and subtopic) to capture user intent in code-mixed Hinglish, providing clear label descriptions to support consistent annotation and evaluation, (2) we evaluated several state-of-the-art proprietary and open-weight multilingual LLMs (including Indic LLMs) to measure how effectively they handle hierarchical intent classification of SRH-related queries expressed in Hinglish. This evaluation also provides insight into how well LLMs can handle code-mixed text, ensuring broader applicability in the real-world health care contexts, and (3) we conducted a qualitative error analysis with thematic coding across models to examine misclassifications in code-mixed SRH queries. We explain why these errors occur and how models misread culturally sensitive terms and euphemisms. Representative annotated examples from the dataset are provided in Table 1.

Table . Sample annotated user queries from the dataset. The examples below present both transliteration and code-mixing between Hindi and English.

Query	English translation	Topic	Subtopic
Periods agar monthly na aaye to kya karen?	What should I do if my periods are not regular every month?	Menstrual Health	Menstrual Cycle Information
Gabhnirodhak goliyan lene se pehle mujhe doctor se kya puchna chahiye?	What should I ask the doctor before taking birth control pills?	Contraception and Family Planning	Usage Guidance
Agar PCOD[a] ka 3 mahina dawa chalane ke baad bhi shi nahi hua to kya kare?	If PCOD[a] does not improve after three months of medication, what should I do?	PCOS[b] or PCOD[a]	Management
Agar abhi bacha nahi rakhna hai aur fir bhi pregnant ho gaye to abortion ke liye kya kare?	If I do not want to have a baby right now but still get pregnant, what should I do for an abortion?	Pregnancy and PNC[c]	Abortion

[a]PCOD: polycystic ovarian disease.

[b]PCOS: polycystic ovary syndrome.

[c]PNC: postnatal care.

Related Work

Hierarchical classification is a method of assigning items to categories organized within a hierarchy of classes [22,23]. This approach has been broadly used in a range of domains where layered structures are important, including e-commerce user query categorization [24,25], health care question answering [26,27], clinical guideline classification [21], and personalized health care analysis of women's menstrual health disorders [28]. Its advantages lie in its ability to model the inherent complexity of user intent, particularly in health care contexts where individuals begin their health care concerns by first referencing broader domains and then subsequently focusing on specific details or contexts. While flat classification systems are widely used as classification baselines, their lack of hierarchical awareness limits their effectiveness, especially in domains where understanding class structure is essential [21,29]. Although hierarchical classification has demonstrated clear benefits across diverse applications, its potential remains underexplored in the context of code-mixed languages and SRH, where understanding layered and implicit user intent is especially important.

The recent development of highly parameterized language models, such as GPT-5 (OpenAI) [30], GPT-4o (OpenAI) [31], Llama 3 (Meta) [32], Gemma 3 (Google) [33], among many others, has significantly enhanced health care reasoning, understanding, and summarization [34]. These models can generate concise and user-friendly summaries and have the potential to make medical information more accessible to the general public. Despite these developments, a major limitation remains their reliance on training data in English, making them less effective for regional languages [12,35]. For instance, Llama 3 sources 90% of its pretraining data from English sources [4], while a substantial portion of GPT-4's pretraining data is similarly English-dominant [5]. This linguistic bias creates information disparities across languages, where equivalent questions in different languages may produce inconsistent and inaccurate responses [36]. Our study seeks to address this gap.

Recent research has increasingly focused on using chatbots to provide health care assistance, particularly around SRH. For example, Wang et al [15] designed an SRH-focused rule-based chatbot on Facebook Messenger (Meta Platforms, Inc) to analyze how users engage with artificial intelligence (AI) to seek health information. Their study found that users frequently shared personal concerns and SRH-related queries in code-mixed languages, highlighting the need for linguistically adaptable health care AI models. There are several reasons why LLMs can be useful for SRH information delivery, including gaps in sexual education [37] and societal stigma and taboo around sex-related topics [12]. However, recent studies indicate that LLMs struggle with Indic languages, particularly when queries are code-mixed and culturally situated [38]. To analyze these issues systematically, we collected a dataset called sexual and reproductive health queries (SRHQ)-India, which captures real-world Indian SRH user queries in a code-mixed format.

As queries in SRH domains are often linguistically and contextually diverse, the hierarchical categorization approach offers the potential for more accurate and interpretable results. Recent advancements in LLMs have made significant improvements in flat text classification, especially in the health care domain. In our work, we seek to evaluate LLMs' understanding of the hierarchical classification task. In this

XSL•FO
RenderX

regard, we explore the use of a zero-shot learning technique that involves crafting prompts to enable language models to generate useful responses without prior examples, relying entirely on pretrained knowledge to tackle the new task. This technique has been used in several health care applications, such as capturing the context of clinical text [39] and classifying health care queries [40]. In our work, we leverage zero-shot learning to evaluate the efficiency of language models in processing code-mixed user queries.

Methods

Overview

In this study, we used a dataset gathered by the Myna Mahila Foundation, a nonprofit women's health organization in India. The data were collected through a preliminary SRH chatbot prototype that was developed and piloted with 488 women in informal settlements of Mumbai, India, from October 2023 to December 2024 to capture real-world user queries on SRH, including both text and voice input. This is part of their ongoing effort to improve access to SRH information for women and girls in urban slums. Below in "Dataset Generation and Curation" and "Data Annotation" subheadings, we detail our approach to structuring and annotating the dataset. We also outline our evaluation framework for model performance.

Dataset Generation and Curation

The dataset originally contained 4858 queries and was refined to reduce redundancy using cosine similarity computed over term frequency–inverse document frequency representations of the text. Queries were first normalized using standard text preprocessing, including Unicode correction, lowercasing, whitespace normalization, and removal of basic punctuation. Cosine similarity was then used to identify syntactically similar queries, and pairs with a similarity score of 90% or higher were considered near duplicates, and only one was retained. For example, "kya saheli tablet lene se periods ka date badal jata hai?" or "saheli tablet se periods ka date badal jata hai kya" convey the same meaning: "Does taking the Saheli tablet (a nonhormonal contraceptive pill) change the period date?" Term frequency–inverse document frequency–based lexical similarity was chosen instead of pretrained sentence embeddings because many widely used embedding models are trained predominantly on monolingual or English-centric data and perform poorly on code-mixed Indian languages [41,42]. Moreover, our aim was to remove lexically rephrased queries rather than infer semantic equivalence. In addition, we removed queries containing fewer than three words to exclude vague or incomplete inputs. We also filtered out greeting-based queries, such as "Namaste" or "Aap kaise hai," which means "How are you," which do not contribute to intent classification.

After filtering, we obtained a final dataset of 4161 queries, primarily in Hinglish. The dataset consists of queries with misspelled English words, transliterated Hindi terms, and borrowed English medical words that are common in everyday SRH talk (eg, periods, condom, pregnancy, in-vitro fertilization [IVF], and polycystic ovary syndrome). We also observed queries related to cultural practices or religious beliefs. For example, "Masik pali aane se Bhagvan ke pas kau nahi jana chahiye?" which in English translates to "Why should one avoid going to god or temples during periods?" This query centers around the religious norm of how women are advised not to visit the temple during their periods in certain cultures and communities in India. Additionally, the query "Det ke pahele piryat q aata hai" translates to "Why periods come before the date?" where det (date) and piryat (period) are misspelled terms. Such variations highlight the real-world linguistic challenges in SRH-related conversations and underscore the need for models capable of accurately processing noisy, code-mixed, and culturally or religiously influenced queries.

Beyond linguistic complexities, the dataset also captures deeply ingrained cultural myths and misinformation, particularly related to pregnancy and gender beliefs. For instance, the query "Main pregnant hu or jinko sirf ladkiya hai kya unko dekte rahne se kya muje ladki hogi?" translates to "I am pregnant, if I keep seeing women who have only given birth to daughters, will I also have a daughter?" Similarly, "Kya muje sirf ladka hi chahiye to uske liye khuch upay hai kya?" means "Is there any way to have only a boy?" These queries reflect long-standing cultural expectations in some communities in India, where socioeconomic structures reinforce a preference for sons. Such myths not only influence reproductive decisions but also contribute to gender-based discrimination and misinformation.

We adopted a hierarchical thematic structure for annotation. Queries were first classified into broad topics and then further refined into subcategories to capture the complexity of SRH-related concerns. For example, questions related to postpartum, antepartum, and abortion fall under the broad category of Pregnancy and Postnatal Care (PNC) category, while questions related to the Menstrual Health category are further classified based on their context to subtopics, including Menstrual Cycle Information, Menstrual Flow, or Period Pain Management. We have also included Mental Health and Wellness, with subtopics including Stress Management and Safety Concerns, because stress, sleep, mood, and anxiety issues often co-occur with SRH questions and influence help-seeking and guidance needs. Placing these queries within the SRH hierarchy keeps related concerns together and improves intent interpretation. Table 2 provides the annotation framework in detail. The annotation guidelines were developed through an iterative process of feedback from doctors and program staff at the Myna Mahila Foundation. This helped standardize interpretations and improve annotation reliability across the dataset. Key descriptive statistics of the dataset are provided in Table 3. In addition, we quantify the degree of intrasentential Hindi-English code-mixing using the Code-Mixing Index (CMI); dataset-level CMI statistics are reported in Table 3, and the metric is described in the following "CMI Calculation" subheading.

XSL•FO
RenderX

Table . Annotated dataset distribution by topic and subtopic.

Topic and subtopic	Queries (N=4161), n (%)
Contraception and Family Planning	1417 (34.1)
Family Planning Queries	522 (36.8)
Usage Guidance	291 (20.5)
Types of Contraceptives	209 (14.7)
Sterilization	184 (13)
Effectiveness and Duration	127 (9)
Side Effects	84 (5.9)
Menstrual Health	1104 (26.5)
Menstrual Cycle Information	673 (61)
Period Pain Management	325 (29.4)
Sanitary Products and Hygiene	62 (5.6)
Menstrual Flow	44 (4)
Pregnancy and PNC[a]	547 (13.1)
Pregnancy Information	290 (53)
Antepartum	120 (21.9)
Infertility	53 (9.7)
Abortion	26 (4.8)
Postpartum	23 (4.2)
Miscarriage	20 (3.7)
Breastfeeding	15 (2.7)
Sexual and Vaginal Health	322 (7.7)
Sex-Related Queries	128 (39.8)
Vaginal Health and Discharge	127 (39.4)
Reproductive Anatomy	26 (8.1)
Urinary Tract Infections (UTI)	16 (5)
Sexually Transmitted Infections (STI or STD)	15 (4.7)
Vaginal or Uterine Infections	10 (3.1)
PCOS or PCOD[bc]	101 (2.4)
Information	49 (48.5)
Management	28 (27.7)
Symptoms	24 (23.8)
HIV[d]	52 (1.2)
Stigma and Awareness	14 (26.9)
Treatment	14 (26.9)
Prevention	12 (23.1)
Symptoms and Early Detection	12 (23.1)
Mental Health and Wellness	22 (0.5)
Stress Management	13 (59.1)
Information and Safety Concerns	9 (40.9)
Other	596 (14.3)
General Health Queries	361 (60.6)

XSL•FO
RenderX

Topic and subtopic	Queries (N=4161), n (%)
Diet and Nutrition	97 (16.3)
Exercise and Fitness	46 (7.7)
Health Equity and Access	38 (6.4)
Marriage and Relationships	21 (3.5)
Misconceptions and Myths	13 (2.2)
Child Health	12 (2)
Cultural, Religious, or Moral Norms	8 (1.3)

[a]PNC: postnatal care.
[b]PCOS: polycystic ovary syndrome.
[c]PCOD: polycystic ovarian disease.
[d]HIV: human immunodeficiency virus.

Table . Statistics of the dataset.

Statistic	Values
Total number of topics	8
Total number of subtopics	40
Total number of queries	4161
Maximum question length (words)	71 words
Average question length (words)	10.56 words
% queries in transliterated Hindi[a]	61.53
% queries in intrasentential mixing[b]	38.37
Code-Mixing Index (CMI), %, mean (SD)	37.38 (10.43)

[a]Transliterated Hindi refers to text written in Hindi using the Latin script, often incorporating English loan words common in sexual and reproductive health (SRH) contexts (eg, periods, condom, and pregnancy).
[b]Intrasentential mixing refers to text where Hindi and English elements are combined within the same sentence, written in Latin script.

Our approach to annotation serves 2 purposes. First, hierarchical classification offers a deeper perspective on SRH concerns, allowing the identification of common themes and trends in user queries. It strengthens decision support for stakeholders, including health care professionals, policymakers, and organizations, by offering a well-organized dataset that can inform better resource allocation and intervention strategies. Second, hierarchical classification can enhance user guidance and response accuracy based on user intent, ensuring that individuals are directed to relevant and specific information tailored to their concerns.

CMI Calculation

The CMI [43] is an utterance-level, ratio-based metric that quantifies intrasentential code-mixing by measuring the proportion of lexical tokens that do not belong to the matrix (dominant) language.

CMI is defined as:

$$CMI=100\times(1-\max(wi)n-u),\text{if } n>u$$

Where n is the total number of tokens, u is the number of language-independent tokens, n-u is the sum of the number of tokens from N languages, and max {wi}is the highest number of words belonging to a particular language. If an utterance contains only language-independent tokens (n=u), the CMI is defined as zero.

We used the pretrained FastText (Meta AI) language identification model, which supports 176 languages [44], to automatically assign language labels at the token level. FastText has been widely used for language identification in multilingual and code-mixed text due to its efficiency and robustness on short lexical units, and its use of subword information [45,46]. Most queries in the dataset are bilingual in nature, with one dominant language, either English or Romanized Hindi, serving as the matrix language within each utterance. We used the model to identify English tokens (and assumed the remaining tokens to be Hindi) due to potential concerns about its performance on Hindi words. Numerals, punctuation, and symbols were treated as language-independent. This automated procedure enabled consistent computation of sentence-level CMI scores across all 4161 queries.

Data Annotation

We annotated the user queries into 8 categories, which were further divided into subcategories to capture the depth and complexity of the topics. The annotation of our code-mixed dataset on SRH required both strong linguistic competence and domain expertise. Since the data are code-mixed (Hindi and

XSL•FO
RenderX

English), they present unique represents unique challenges such as nonstandard spelling, switching between languages in the same query, and the cultural nuances embedded in users' expressions. To meet these complexities, this study included 2 experienced annotators with backgrounds in public health and global health. Both annotators are native or fluent speakers of Indian languages and coauthors of this paper, ensuring familiarity with cultural and linguistic contexts relevant to SRH queries. Both were guided by a comprehensive annotation manual, refined by subject matter experts (public health care professionals and medical doctors). Some user questions could overlap with multiple topics or subtopics (for instance, a question about convincing a partner to use contraceptives could fall under the "Family Planning" or "Marriage and Relationship" category). In such cases, annotators selected the closest-fitting category. The full annotation guidelines are provided in Multimedia Appendix 1.

To ensure annotation quality and consistency, we adopted an iterative annotation process. Initially, 10% of the dataset was annotated independently by both annotators. This step was particularly critical, as it allowed us to refine annotation guidelines and address challenges arising from orthographic variations, code-mixing, and semantic ambiguities. Discrepancies were resolved through discussion and led to refinements in the annotation guidelines. We continued the iterative co-annotation process until the annotators consistently achieved a 95% agreement threshold during the pilot phase. Once the threshold was met, the remaining data were divided between the 2 annotators for independent annotation. We used Cohen kappa [47] across overlapping annotated samples to compute interannotator agreement. The resulting score of 83% indicated substantial agreement, reflecting almost perfect agreement [48]. The distribution of the query categories and subcategories is provided in Table 2.

Evaluation Metrics

To evaluate model performance on our hierarchical classification task, we adopted the Hierarchical F_1-score (hF_1) proposed by Kosmopoulos et al [23]. It extends traditional metrics by explicitly considering the hierarchical relationship among labels. To compute this metric, we augment both the predicted labels (Y^) and the ground truth labels (Y) by incorporating their ancestor labels from the hierarchy, resulting in augmented sets Y^aug*and*Yaug. This augmentation ensures that the evaluation accurately reflects hierarchical dependencies between topics and subtopics.

The hierarchical F_1-score (hF_1) is defined as:

$$(1)\ hF1=2\cdot hPr\cdot hRe/hPr+hRe,\ where\ hPr=\sum|Y\hat{a}ug\cap Yaug|/\sum|Y\hat{a}ug|,\ hRe=\sum|Y\hat{a}ug\cap Yaug|/\sum|Yaug|$$

Here, hPr and hRe (hierarchical precision and recall) evaluate the proportion of correctly predicted hierarchical labels among all predicted and actual labels, respectively. Additionally, we used supporting metrics, including Exact Match, which evaluates the percentage of queries where both the topic and subtopic predictions exactly match the ground truth labels. Accuracy@l_1 measures the topic-level accuracy, reflecting the proportion of samples where the predicted topic is correct. Accuracy@l_2 measures conditional subtopic accuracy, defined as the proportion of correct subtopic predictions given that the topic was correctly predicted. These complementary metrics collectively provide a comprehensive evaluation of the model's performance, capturing strict end-to-end correctness (Exact Match), coarse-grained topic identification (Accuracy@l_1), fine-grained subtopic discrimination (Accuracy@l_2), and hierarchical sensitivity (hF_1).

Experimental Analysis

In this study, we evaluate the performance of language models on a hierarchical classification task in a zero-shot setting, where the model attempts to generalize output patterns at the topic and subtopic levels for data it has not encountered during training. Zero-shot evaluation is particularly challenging because it requires models to infer patterns without explicit task-specific fine-tuning.

In our setup, models were provided with the hierarchical structure and the list of possible topic and subtopic labels, but not with the detailed semantic definitions used in the human annotation guideline. This was a deliberate design choice to evaluate the models' intrinsic ability to interpret real-world, code-mixed SRH queries without additional semantic instruction. It also reflects practical deployment scenarios where the model must generalize from raw user input rather than rely on curated taxonomies at the time of inference.

Our evaluation includes 7 open-weight multilingual models, 5 Indic-specific models, and 3 proprietary models. The open-weight multilingual models are Mixtral-8 × 7B-Instruct (Mistral AI) [49], Llama-3.1-8B-Instruct (Meta) [32], Llama-3.3-70B-Instruct (Meta), Gemma-2-9B-IT (Google) [50], Gemma-3-27B-IT (Google) [33], Qwen-2.5-7B-Instruct (Alibaba Group) [51], and Aya-Expanse-8B (Cohere Labs) [52]. The technical reports of these models demonstrate their multilingual capabilities on benchmark datasets, making them strong candidates for handling code-mixed data.

The Indic models are Airavata [53], an instruction-tuned Hindi language model developed by AI4Bharat; AryaBhatta-GemmaGenZ-Vikas-Merged [54] (referred to as AryaBhatta [GenVR Research]), a model trained on 9 Indic languages that excels particularly in Hindi reasoning and literature tasks; and Llama-3-Gaja-Hindi-8B [55], a bilingual Hindi-English LLM specialized in Indic language understanding. We also evaluate Krutrim-2-Instruct [56], a Mistral-NeMo (Mistral AI and NVIDIA)–based model trained on diverse domains, including Indic languages, and fine-tuned with direct preference optimization to improve alignment and reasoning for Indian contexts. For proprietary baselines, we include GPT-5 [30], GPT-4o [31], and Claude-3.5-Sonnet [57], which serve as state-of-the-art performance references despite their closed parameter counts.

For open-weight multilingual models, we include a range spanning smaller (7B-9B) to larger (27B-70B) architectures. Alongside these, we also include Sarvam-M (Sarvam AI) [58], a state-of-the-art open-source hybrid Indic LLM built on Mistral-Small, designed to enhance reasoning in Indian languages through supervised fine-tuning and reward-based reinforcement learning. Since Indic LLMs are still emerging,

XSL•FO
RenderX

our evaluation also includes small to mid-sized versions currently available (7B-12B). This selection allows for a balanced comparison across Indic, multilingual, and proprietary systems, highlighting the relative strengths and limitations of each category.

With LLMs, a precise and well-structured prompt is paramount for shaping responses and ensuring that the model focuses on the most relevant aspects of the input. Recent research [59] has shown that even minor lexical variation—sometimes just a single-word difference—in prompts can significantly impact model performance on downstream tasks. To accurately determine topic and subtopic levels in our task, we experimented with multiple prompt variations on a small, representative subset of our dataset. This iterative process allowed us to refine the prompts efficiently before running the large-scale evaluation. Smaller open-weight models were evaluated locally on a graphics processing unit with 48GB of memory (eg, NVIDIA RTX A6000), while proprietary models were accessed via application programming interface–based inference. The final prompt template used across all models is shown in Table S1 in Multimedia Appendix 2. The prompt template also requested a self-reported confidence score (0.0 - 1.0); however, this value was extracted for future analysis but was not used in this evaluation.

Qualitative Error Analysis

To better understand the strengths and limitations of the models, we conducted a qualitative error analysis of the 2 best-performing models in each family (proprietary, multilingual open-weight, and Indic models). For each model, we sampled 50 misclassified test queries stratified by topic and subtopic. The same 2 annotators who created the dataset served as reviewers and independently coded thematic errors; disagreements were resolved by consensus. Results are summarized in the “Error Analysis” section, with examples in Table 4.

XSL•FO
RenderX

Table . Performance of 6 models on representative sexual and reproductive health queries (SRHQ) in the SRHQ-India dataset queries at the topic-subtopic level.

Error types	Query (Hinglish→English)	Ground truth (Topic→Subtopic)	GPT-5	Claude-3.5-Sonnet	Llama-3.3-70B-Instruct	Gemma-3-27B-IT	Sarvam-M	Llama-3-Gaja-Hindi-8B
C1	Safaiya kaise karwate hain? (How is an abortion done?)	Pregnancy and [a]PNC→Abortion	✓[b]	✓	—[c]	—	—	—
C2	Pregnancy me 5 month me vomiting hoti h to kya kre (What should I do if I am vomiting in the 5th month of pregnancy?)	Pregnancy and PNC→Antepartum	✓	✓	—	—	✓	—
C3	Family planning Muslim community me accept hain kya? (Is family planning accepted in the Muslim community?)	Contraception and Family Planning→Family Planning Queries	—	—	✓	✓	—	✓
C3	Masik pali aane se Bhagvan ke pas kau nahi jana chahiye? (Why should one avoid going to god/temples during periods?)	Other→Cultural, Religious, or Moral Norms	✓	✓	—	—	✓	—
C4	1 sal se bacha rukhne ke liye try kar rahe hai lekin nahi rukh raha hai to kya karna padega? (We have been trying to have a child for one year, but it has not happened. What should we do?)	Pregnancy and PNC→Infertility	✓	✓	—	—	—	—

XSL•FO
RenderX

Error types	Query (Hinglish→English)	Ground truth (Topic→Subtopic)	GPT-5	Claude-3.5-Sonnet	Llama-3.3-70B-Instruct	Gemma-3-27B-IT	Sarvam-M	Llama-3-Gaja-Hindi-8B
C4	Family planning may agar koi mahila test tube karvati hai to kya hota hai? (In family planning, if a woman undergoes test-tube treatment, what happens?)	Pregnancy and PNC→Infertility	—	—	—	—	—	—
C5	Kitni der jinda rahte hain sperm? (How long do sperm stay alive?)	Sexual and Vaginal Health→Reproductive Anatomy	—	✓	—	—	—	—

[a]PNC: postnatal care.

[b]✓ indicates correct topic-subtopic classification.

[c]— indicates incorrect topic-subtopic classification

Ethical Considerations

We used deidentified secondary data shared by the Myna Mahila Foundation. Since the dataset was fully deidentified before being shared, no personally identifiable information was accessible to the research team at Emory University. This study did not involve direct interaction with human participants, and no additional data collection was conducted. Approval for the use of this secondary dataset was obtained from Emory University's Institutional Review Board (Protocol #2025P011010). Data security measures were implemented to maintain confidentiality, and the dataset was stored on a secure, password-protected system. Only authorized personnel had access to the data. As the study relies on pre-existing deidentified data, the risk to individuals is minimal.

Results

Overview

We now present the results of the hierarchical classification of SRHQ with LLMs. Table 5 compares model performance on the hierarchical classification across 3 types of models: open-weight, Indic (LLMs fine-tuned for Indian languages), and proprietary. Figure 2 provides a visual comparison across 4 evaluation metrics: hierarchical F_1(hF_1), Exact Match, and accuracy at the topic (l_1) and subtopic (l_2) levels. Although we also evaluated GPT-4o, its performance was less than 1.5% lower than GPT-5 across metrics. Bootstrap-based 95% CIs further show substantial overlap between GPT-5 (hF_1=0.784, 95% CI 0.774 - 0.795) and GPT-4o (hF_1=0.779, 95% CI 0.768 - 0.789), indicating that performance differences between the 2 proprietary systems are not statistically distinguishable (Table S2 in Multimedia Appendix 2). For consistency and balanced cross-category comparisons, we focus here on GPT-5 and Claude-3.5-Sonnet as representative proprietary models.

XSL•FO
RenderX

Table . Performance comparison of different models in a zero-shot setting on the sexual and reproductive health queries (SRHQ) dataset. Open-weight, proprietary, and Indic models are distinctly highlighted for clarity. All Indic models are open weights. In each category, the best-performing model is highlighted in **bold**, and the second-best is underlined.

Category and model	#Params	hF_1 (hierarchical $F1$)	Exact Match	Accuracy@l_1	Accuracy@l_2
Open-weight models					
Mixtral-8 × 7B-Instruct	7B	0.593	0.453	0.733	0.617
Llama-3.1-8B-Instruct	8B	0.630	0.491	0.769	0.638
Qwen-2.5-7b-Instruct	7B	0.605	0.463	0.747	0.619
Aya-Expanse-8B	8B	0.528	0.411	0.646	0.636
Gemma-2-9B-IT	9B	0.657	0.544	0.770	0.706
Gemma-3-27B-IT	27B	0.739	0.629	0.849	0.741
Llama-3.3-70B-Instruct	70B	**0.742**	**0.630**	**0.853**	**0.738**
Indic models					
Airavata	7B	0.404	0.226	0.581	0.389
Llama-3-Gaja-Hindi-8B	8B	0.596	0.452	0.740	0.610
AryaBhatta	8.5B	0.365	0.157	0.574	0.273
Krutrim-2-Instruct	12B	0.558	0.386	0.731	0.527
Sarvam-M	24B	**0.757**	**0.647**	**0.867**	**0.747**
Proprietary models					
GPT-5	__[a]	**0.784**	**0.683**	**0.886**	**0.771**
GPT-4o[b]	—	0.779	0.675	0.882	0.764
Claude-3.5-Sonnet	—	0.745	0.639	0.851	0.751

[a]Not available.

[b]Although GPT-4o was the second-best-performing proprietary model, its performance differed from GPT-5 by less than 1.5% across metrics. Therefore, GPT-5 and Claude-3.5-Sonnet were selected as representative proprietary models for balanced cross-category comparisons in the main analysis.

XSL•FO
RenderX

Figure 2. Visual comparison of model performance on the sexual and reproductive health queries (SRHQ) dataset across 4 evaluation metrics: hierarchical F_1(hF_1), Exact Match, Accuracy@l_1, and Accuracy@l_2. The figure highlights relative strengths and differences among open-weight, Indic, and proprietary models.

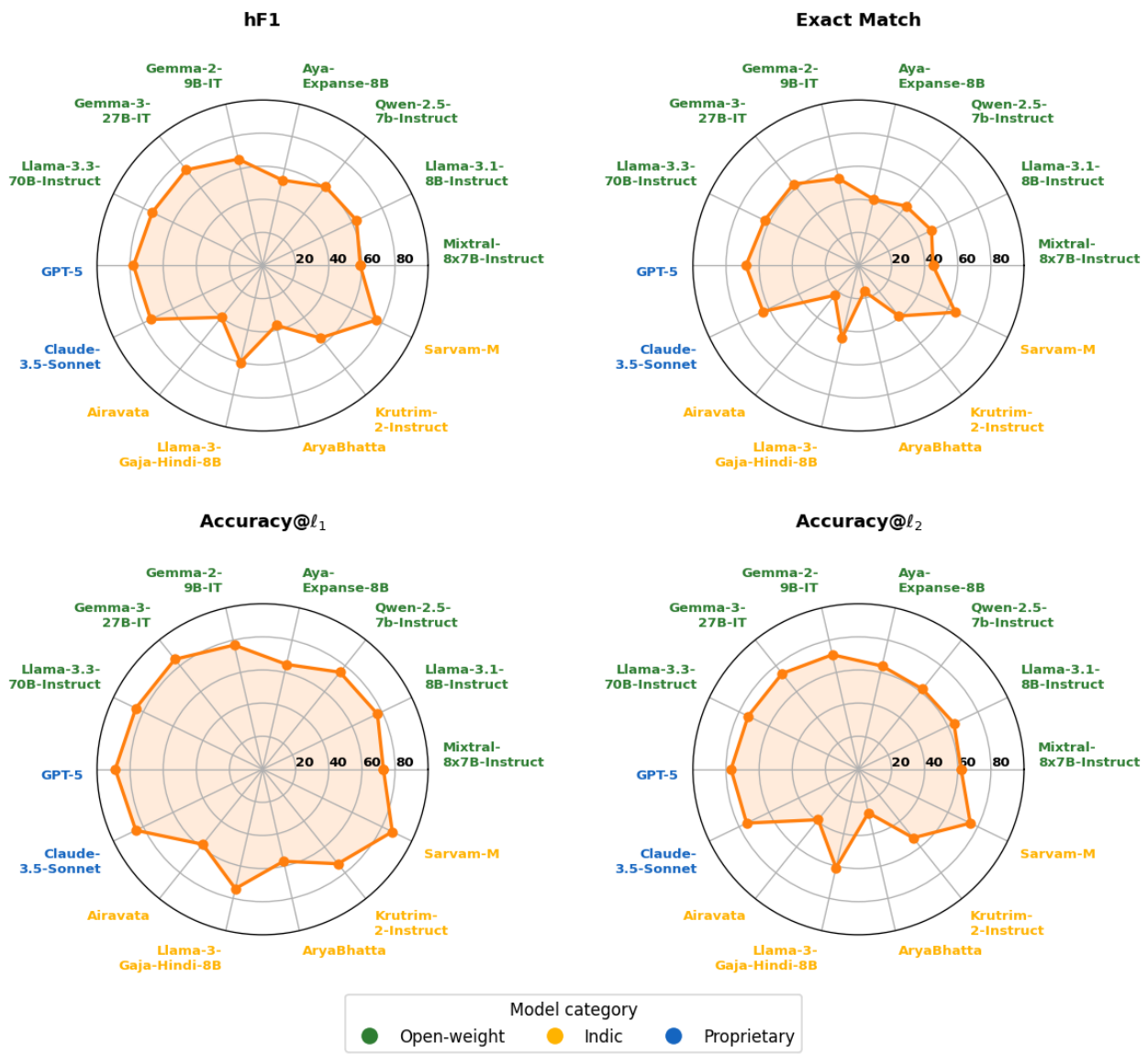

Proprietary systems delivered the highest and most consistent performance, with GPT-5 achieving the best results across all metrics, with hF_1=0.784, Exact Match=0.683, and accuracies of 0.886 at l_1 and 0.771 at l_2. Claude-3.5-Sonnet closely followed (hF_1=0.745 and Exact Match=0.639), exceeding 0.85 topic-level accuracy and 0.75 subtopic-level accuracy.

Among open-weight models, the strongest performers were Llama-3.3-70B-Instruct and Gemma-3-27B-IT, which achieved performance close to the proprietary models, with hF_1 scores of 0.742 and 0.739 and Exact Match scores of 0.630 and 0.629, respectively. Both models exceeded 0.84 in topic-level accuracy (l_1) and maintained strong subtopic accuracy ($l_2 \approx 0.74$) with scores of 0.738 and 0.741. Mid-sized models such as Gemma-2-9B-IT also performed competitively (hF_1=0.657, Exact Match=0.544, l_1=0.770, and l_2=0.706), whereas smaller models such as Aya-Expanse-8B underperformed across metrics (hF_1=0.528, Exact Match=0.411) and struggled particularly at the subtopic level (l_2=0.636). Figure 2 illustrates this scaling effect clearly, with larger multilingual models covering a broader area across hF_1, Exact Match, and both accuracy levels compared to their smaller counterparts.

Indic models showed the widest variability, with Sarvam-M emerging as a strong outlier, achieving hF_1=0.757 and Exact Match=0.647. Compared to GPT-5, Sarvam-M fell short by only 3.4% in hF_1. At the topic level, it reached an accuracy of l_1=0.867, and at the subtopic level l_2=0.747, only slightly below the top-performing model GPT-5 (0.886 and 0.771, respectively). This indicates that Sarvam-M is the strongest free and open-weight alternative to proprietary models, considering the restricted access and cost associated with proprietary models. Other Indic models performed substantially lower; the second-highest-performing Indic model was Llama-3-Gaja-Hindi-8B, achieving an hF_1 of 0.596, an Exact Match of 0.452, and accuracies of 0.740 at l_1 and 0.610 at l_2. Interestingly, it outperformed the larger Krutrim-2-Instruct

XSL•FO
RenderX

(12B; OLA Krutrim Team) (hF_1=0.558, Exact Match=0.386, l_1=0.731, and l_2=0.527), while Airavata demonstrated lower performance across metrics (hF_1=0.404, Exact Match=0.226, l_1=0.581, and l_2=0.389). These comparisons indicate that parameter count alone does not guarantee superior performance; factors such as model design and training data quality appear to play a more decisive role. By contrast, Airavata and AryaBhatta consistently trailed across all metrics, clustering at the lowest end of performance in both topic- and subtopic-level accuracy.

The results highlight a consistent performance hierarchy across model categories. Proprietary systems dominated overall, while large open-weight multilingual models narrowed the gap across all metrics, and Indic models demonstrated both promise for local contexts and limitations. Sarvam-M stood out as a competitive alternative, while other Indic models struggled to identify user intent in code-mixed settings. The combination of Table 5 and Figure 2 underscores not only differences in overall performance but also the critical difficulty of achieving accurate subtopic-level predictions in mixed-language SRH queries.

To further examine performance differences, Figures 3 and 4 present F_1 scores across SRHQ-India topics. Figure 3 highlights per-topic F_1 scores alongside model-wise averages, where the reported average F_1 is the macro-average of flat, topic-level F_1 scores computed across all topics for each model. Proprietary model GPT-5 achieved the highest overall average F_1 score (0.85), while Sarvam-M emerged as the strongest open-weight Indic alternative with an average of 0.82, with performance comparable to Claude-3.5-Sonnet (0.79) and other open-weight models such as Llama-3.3-70B (0.78) and Gemma-3-27B (0.79). Other Indic models, such as Llama-3-Gaja-Hindi-8B (CognitiveLab), showed weaker averages and wider variability, particularly in nuanced domains.

Figure 3. Model-wise topic-level F_1 scores across the sexual and reproductive health queries (SRHQ)-India dataset. Each bar represents the flat F_1 score for a specific topic, while the dotted lines indicate the macro-average of topic-level F_1 scores across all topics for each model. Results highlight differences between proprietary, open-weight, and Indic models. HIV: human immunodeficiency virus; PCOD: polycystic ovarian disease; PCOS: polycystic ovary syndrome; PNC: postnatal care.

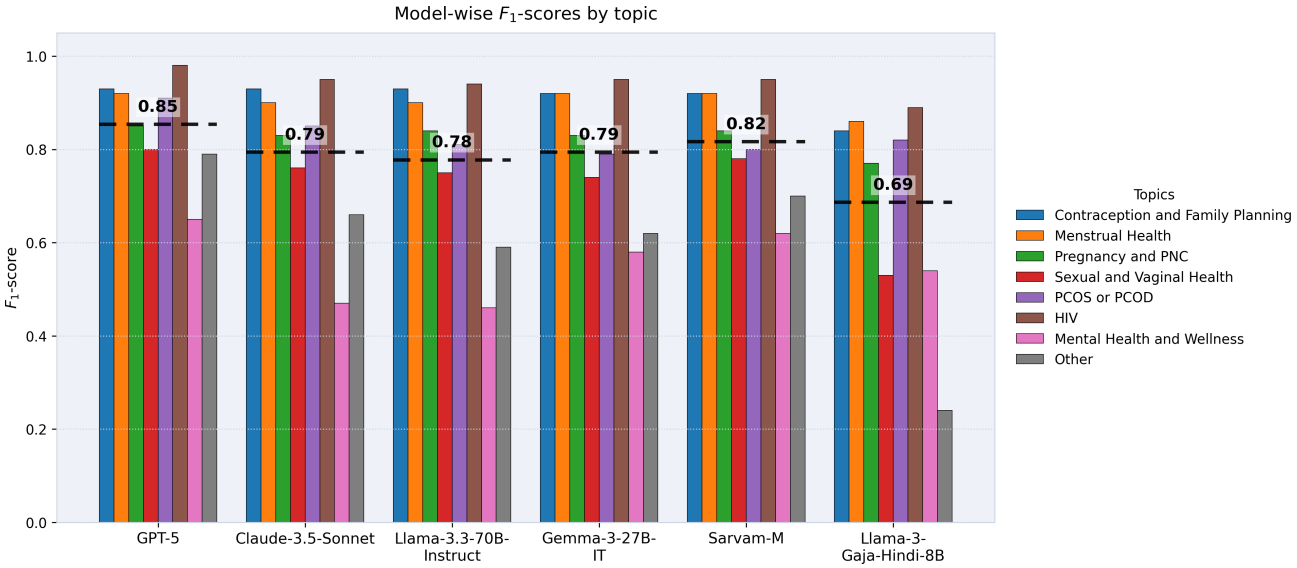

XSL•FO
RenderX

Figure 4. Topic-level F_1 scores with 95% CIs across sexual and reproductive health queries (SRHQ)-India. Each bar represents the F_1 score for a specific topic, with error bars indicating 95% CIs estimated via bootstrap resampling. Results highlight performance differences across proprietary, open-weight, and Indic models. HIV: human immunodeficiency virus; PCOD: polycystic ovarian disease; PCOS: polycystic ovary syndrome; PNC: postnatal care.

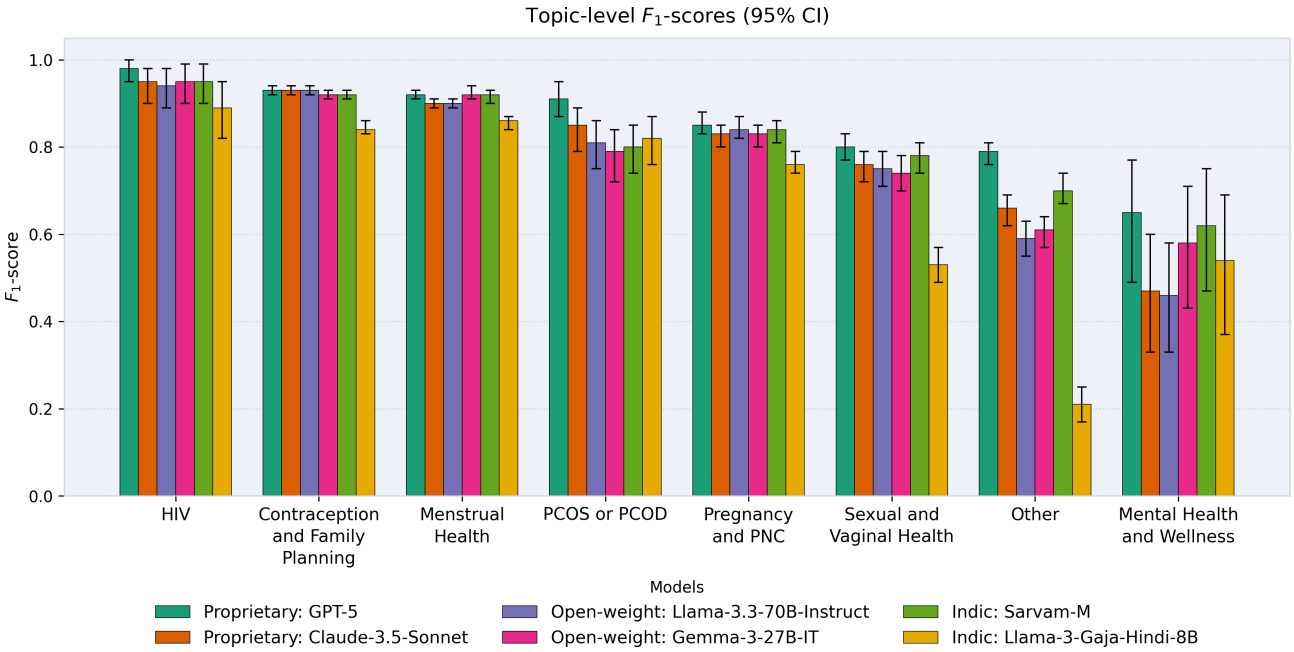

Figure 4 provides a closer look at topic-level differences across models. Proprietary systems maintained stability across nearly all categories, performing strongest in human immunodeficiency virus, Contraception and Family Planning, and Menstrual Health, which require precise and time-sensitive classification. Although human immunodeficiency virus was represented in fewer queries, performance remained consistent across models, suggesting less variability in how these queries are expressed. Sarvam-M again demonstrated competitive results, closely tracking proprietary and large multilingual models in categories such as Contraception and Family Planning, Menstrual Health, and polycystic ovary syndrome or polycystic ovarian disease. However, all models showed relative declines in Sexual and Vaginal Health, and Mental Health and Wellness, where queries frequently overlap with adjacent categories and require additional context for a clear-label decision. In these topics, diverse expressions of intent and limited sample sizes contribute to greater uncertainty, particularly for Mental Health and Wellness, where even a small number of misclassifications can substantially affect performance estimates. These findings highlight that while proprietary systems remain strongest, large multilingual open-weight models substantially close the gap, and Sarvam-M stands out as a promising Indic alternative. Yet, across all families, accurate subtopic-level predictions in code-mixed SRH queries continue to be the most persistent challenge.

To further examine where models succeed and fail, Figure 5 presents confusion matrices for proprietary, open-weight, and Indic models, highlighting recurring misclassification patterns. Consistent weaknesses appeared in overlap-prone and context-dependent topics such as Other, Mental Health and Wellness, and Sexual and Vaginal Health, where misclassifications were frequent and often collapsed into broader categories. Proprietary systems, particularly GPT-5, demonstrated the most balanced performance across all topics, while Claude-3.5-Sonnet showed sharper drops in Other categories despite strong results elsewhere. Open-weight models (Gemma-3-27B and Llama-3.3-70B) performed competitively but exhibited greater instability in culturally sensitive areas, including frequent misclassification of Sexual and Vaginal Health queries into the Pregnancy and PNC or Other category. Among Indic models, Sarvam-M emerged as the strongest, closely mirroring proprietary models' performance by correctly classifying over 90% of queries in Menstrual Health, Contraception and Family Planning, and Mental Health and Wellness. However, it still struggled in fine-grained intent recognition, with frequent misclassification of Contraception and Family Planning into the Pregnancy and PNC or Menstrual Health category. By contrast, smaller Indic systems such as Llama-3-Gaja-Hindi-8B displayed widespread confusions across sensitive categories and a notable tendency to produce "NotValid" predictions, where models generated labels not included in the predefined SRH hierarchy shown in the prompt. For clarity in Figure 5, predictions that fall outside the predefined SRH hierarchy are grouped under a "NotValid" category. These out-of-hierarchy outputs receive no credit in the hierarchical evaluation. These cases underscore the limited ability of smaller Indic models to consistently map inputs to the defined label structure and reflect the challenges related to limited training data and weaker alignment with the task structure.

XSL•FO
RenderX

Figure 5. Topic-level confusion matrices for 6 models representing proprietary, open-weight, and Indic categories. Each cell shows counts (black) and row-normalized percentages. Correct classifications appear in green diagonal cells with percentages in blue. Misclassifications appear in off-diagonal yellow cells, with percentages shown in red. "NotValid" denotes predictions that do not belong to any valid topic. HIV: human immunodeficiency virus; PCOD: polycystic ovarian disease; PCOS: polycystic ovary syndrome; PNC: postnatal care.

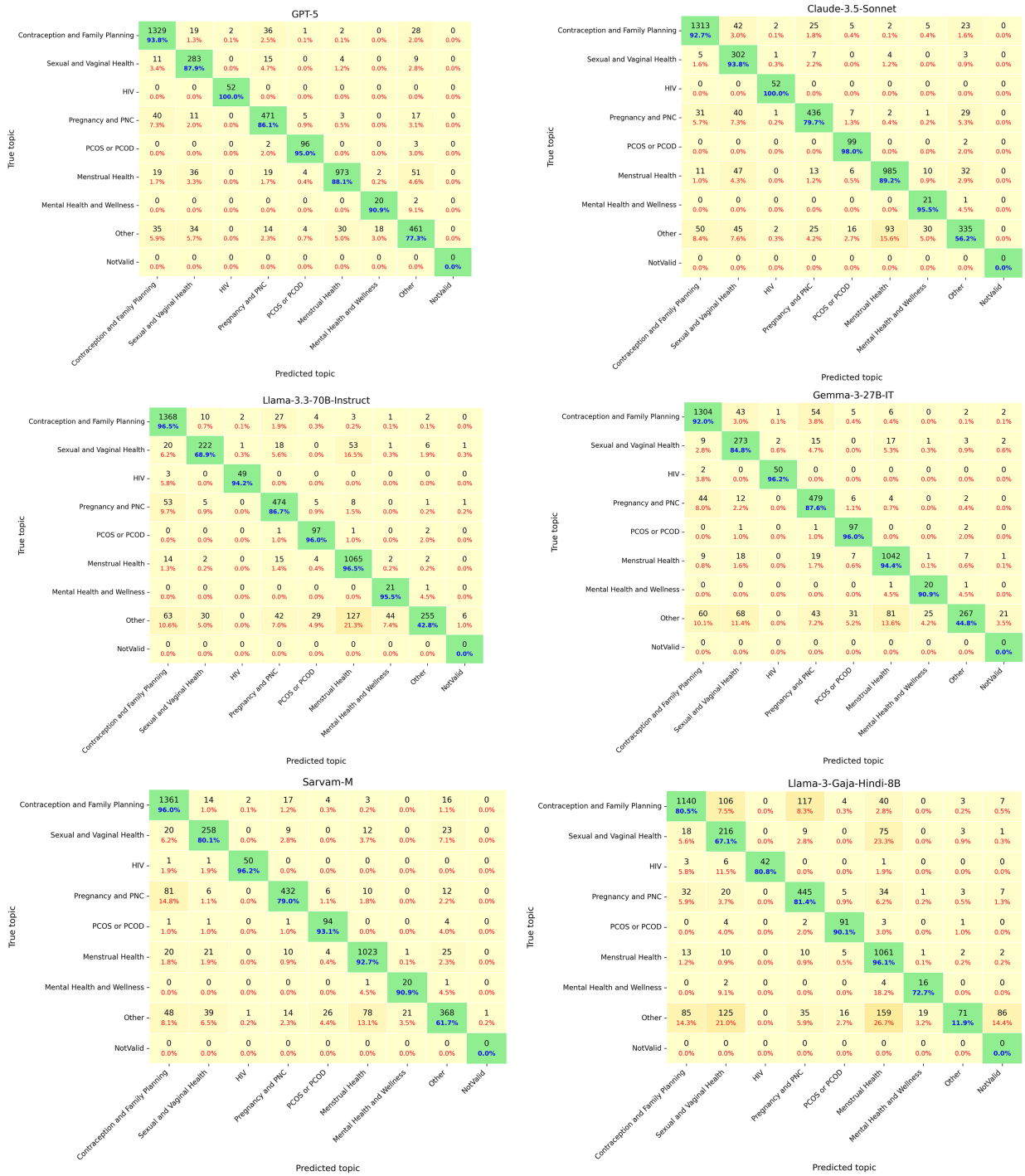

Error Analysis

To better understand model behavior beyond aggregate metrics, we conducted a qualitative analysis of errors from 6 representative models (GPT-5, Claude-3.5-Sonnet, Llama-3.3-70B-Instruct, Gemma-3-27B-IT, Sarvam-M, and Llama-3-Gaja-Hindi-8B) spanning proprietary, multilingual open-weight, and Indic categories. For each model, we sampled 50 misclassified queries and performed a detailed review of their errors. Table 4 presents examples of misclassification. Additionally, we included error analysis for GPT-5 and GPT-4o in Table S3 in Multimedia Appendix 2, highlighting cases where GPT-4o misclassified queries that GPT-5 handled correctly.

C1: Lexical Ambiguity and Euphemisms

In Table 4, row 1, the query "safaiya kaise karwate hain?" translates to "How is an abortion done?" and should be classified as Pregnancy and PNC→Abortion. However, only the proprietary models assigned the topic and subtopic correctly; both open-weight and Indic models misclassified the query

XSL•FO
RenderX

under Sexual and Vaginal Health, Menstrual Health, or General Health Queries. As detailed in Table S4 in Multimedia Appendix 3, model-generated explanations indicate that these errors stem from the literal interpretation of the colloquial term "safaiya" as hygiene, sanitary products, or general cleaning practices. In practice, however, the term is commonly used to refer to the abortion evacuation procedure in the Indian context [60]. This example highlights how culturally embedded euphemisms widely used in Indian contexts can mislead models if culturally specific meanings are not recognized.

C2: Misclassifying Stage in the Reproductive Health Journey

As shown in Table 4, row 2, models frequently confused different pregnancy stages. The query "Pregnancy me 5 month me vomiting hoti hya to kya kre" translates to "What should I do if I am vomiting in the 5th month of pregnancy?" which pertains to the antepartum stage (pregnancy period). Of the 6 best-performing models, only GPT-5, Claude-3.5-Sonnet, and Sarvam-M predicted this correctly. The remaining models classified this query under more general Pregnancy Information, thereby failing to distinguish the specific stage of the reproductive health journey.

As illustrated in Table S5 in Multimedia Appendix 3, these models recognized the pregnancy context but did not incorporate gestational timing into their subtopic selection. This distinction is clinically important, as symptoms such as vomiting have different implications and recommended guidance during the antepartum vs postpartum periods. Assigning such queries in overly broad subcategories such as "Pregnancy Information" dilutes the specificity of the advice that models are expected to provide and risks misalignment with clinical best practices.

C3: Cultural and Religious Context Error

In Table 4, row 3, the user query "Family planning Muslim community me accept hain kya?" translates to "Is family planning accepted in the Muslim community?" reflects a user's religious concern regarding the acceptability of family planning. In our evaluation, both open-weight models correctly assigned the topic and subtopic, while others—including proprietary and Indic models—placed the query under Cultural, Religious, or Moral Norms. Model-generated reasons (Table S6 in Multimedia Appendix 3) indicate that these models focused on the religious framing of the questions rather than their underlying health-related intent. Similarly, in row 4, the query "Masik pali aane se Bhagvan ke pas kyo nahi jana chahiye?" (Why should one avoid going to temples during menstruation?) reflects a culturally and religiously grounded belief rather than a medical concern. GPT-5, Claude-3.5-Sonnet, and Sarvam-M correctly classified this query under Cultural, Religious, or Moral Norms. In contrast, 3 other models, including both open-weight and Indic models, classified it under health-related categories such as Menstrual Health or Misconceptions and Myths. Model-generated reasons (Table S6 in Multimedia Appendix 3) suggest that these systems focused on biological or psychological interpretations of menstruation rather than the user's culturally grounded concern.

In both examples above, the alternative categorizations are semantically plausible given the cultural framing of the user queries; therefore, we treat these cases as a routing-level issue rather than a linguistic error. We later discuss how we might deal with such ambiguous cases when classifying user intent in the "Limitations and Future Works" subheading.

C4: Misunderstanding Reproductive Intent

This category captures cases where models failed to recognize the user's reproductive intent, often reversing it from wanting to conceive to wanting to avoid conception. For example, in Table 4, row 5, the query "1 sal se bacha rukhne ke liye try kar rahe hai lekin nahi rukh raha hai to kya karna padega?" translates to "We have been trying to have a child for one year, but it has not happened. What should we do?" and reflects the issue of infertility, a deeply personal and emotionally sensitive topic. However, only the proprietary model classified this query correctly. Model-generated reasons from other systems (Table S7 in Multimedia Appendix 3) indicate that the colloquial phrase "bacha rukna," which in India commonly means conceiving, was interpreted literally as preventing pregnancy. As a result, these models prioritized Contraception and Family Planning subtopics such as Family Planning Queries and Contraceptive Effectiveness and Duration. While such interpretations are not inherently logical given the taxonomy, they diverge from the user's underlying intent of seeking fertility-related guidance.

A similar misalignment occurred in row 6 with the query "Family planning me agar koi mahila test tube karvati hai to kya hota hai?" ("In family planning, if a woman undergoes a test tube procedure, what happens?"). Here, the user was referring to IVF, a fertility treatment aimed at achieving pregnancy. Yet all models categorized the query under Family Planning or Sterilization rather than Infertility. This demonstrates how culturally specific terms such as "test tube karvati hai" *can invert* user intent, classifying conception-seeking queries as if they were about contraception.

From an SRH perspective, such errors are critical, since infertility queries require guidance on fertility assessment and treatment. Misclassifying these queries risks invalidating user concerns, reinforcing stigma, as infertility in India is often associated with blame and silence, particularly for women [61]. Treating these queries as matters of pregnancy avoidance could further marginalize users seeking support.

C5: Misclassifying Questions About Bodily Anatomy or Processes

In Table 4, row 7, the query "Kitni der jinda rahte hain sperm?" translates to "How long do sperm stay alive?" and should be categorized under Sexual and Vaginal Health, subtopic Reproductive Anatomy. Only 1 model (Claude-3.5-Sonnet) correctly identified both the topic and subtopic. Other models routed the query to categories such as Sexual and Vaginal Health, Contraception and Family Planning, Pregnancy and PNC, or General Health Queries (Table S8 in Multimedia Appendix 3). Model-generated reasons indicate that many systems overassociate the term sperm with sexual behavior, contraception, or pregnancy risk rather than recognizing it as a biological question. Such routing decisions are consequential,

XSL•FO
RenderX

as users seeking factual information about sperm viability may instead receive guidance about sexual behavior or contraception, reducing clarity and potentially misdirecting them away from accurate, evidence-based biological information.

Discussion

Principal Findings

Our evaluation shows that both model family and training data alignment play an important role in hierarchical SRH intent classification, alongside model scale. Proprietary systems still lead, but the gap narrows when models are trained on culturally aligned and domain-relevant data. Sarvam-M, an Indic-first model, emerged as the strongest Indic system and achieved performance comparable to top-performing proprietary models and the best large open-weight baselines such as Llama-3.3-70B-Instruct and Gemma-3-27B-IT. It is also important to consider that Sarvam-M is larger than the other Indic models evaluated, and its performance likely reflects a combination of increased model capacity and Indic-focused training data. While these results do not fully disentangle the effects of scale and data quality, they suggest that culturally aligned resources, rather than parameter count, are the primary drivers of performance on code-mixed intent classification.

Across the 6 best-performing models, we observed a consistent pattern of subtopic misclassification. In Pregnancy and PNC, antepartum and postpartum concerns are frequently conflated, and abortion-related queries are sometimes assigned to the "Pregnancy Information" subcategory. In "Sexual and Vaginal Health," euphemisms such as "safaiya or safai" (colloquial for abortion) were misinterpreted literally as "cleaning." Similarly, culturally embedded phrases such as "bacha rukna" (conceiving) are sometimes misinterpreted as preventing pregnancy, and "test tube karvati hai" (referring to IVF) is sometimes treated as sterilization or contraception [62]. This reflects how health communication in rural, low-literacy settings often relies on indirect, gendered, and culturally mediated channels [63]. Accounting for these cultural expressions is essential to improve classification accuracy and ensure that clinically relevant questions are not overlooked. We also observed that SRH queries could be misrouted as Cultural, Religious, or Moral Norms, even when the underlying intent is to seek health guidance or clinical information. A user query could also plausibly fit into multiple other categories, as we observed. Future work may explore culturally informed prompt design and lightweight in-context examples that expose models to common SRH euphemisms and colloquial expressions. Additionally, to address overlapping topics, researchers could explore alternative approaches that account for such ambiguity, such as ranked classification [64].

In the "Other" category, even strong models such as Llama-3-Gaja-Hindi-8B and Gemma-3-27B-IT frequently assigned queries to new topics outside the defined hierarchy: we consider that kind of misclassification as "Not Valid" predictions (Figure 5). These cases arise when models generate unintended or unsupported labels that are not part of the hierarchical classification structure, reflecting models' tendency toward hallucinations [65] or schema misalignment [66]. Our hierarchical framework requires a correct topic prediction before the subtopic evaluation. Accordingly, "NotValid" predictions are treated as complete classification failures and receive no credit in the hierarchical F_1 score. Overall, these patterns indicate that models struggle to reliably map culturally and contextually grounded queries to the specified hierarchy. This limitation is essentially critical in SRH contexts, where errors at the topic level prevent meaningful subtopic interpretation and risk misrouting sensitive queries. High precision at the fine-grained intent level is paramount to ensure safe guidance and avoid reinforcing stigma.

When compared to prior chatbot efforts, our findings highlight the importance of a hierarchical classification approach. SnehAI [15], an SRH-focused conversational agent deployed in India, demonstrated feasibility but relied on rule-based methods, which restrict its ability to process open-ended, code-mixed, or culturally nuanced queries. Health-Pariksha [38], while not designed specifically for SRH, evaluated several LLMs on real-world health care queries, measuring the factual correctness, semantic similarity, coherence, and conciseness of the model responses. Although this approach contributed to improving trustworthiness, it did not address the challenges of intent recognition or hierarchical classification. In contrast, our framework directly targets the layered structure of SRH queries, enabling distinction between broad categories and their fine-grained subcategories. This supports classifications that are both clinically appropriate and culturally relevant. Through our evaluation, proprietary models emerged as the most reliable for hierarchical classification, but notably, Sarvam-M emerged as the second-highest performing model overall—outscoring Claude-3.5-Sonnet and closely matching Llama-3.3-70B-Instruct. Given that Sarvam-M is open weight and free to use, it represents a promising alternative to proprietary systems, showing that with robust data and cultural adaptation, Indic models can deliver competitive performance.

Limitations and Future Works

This study has several limitations. First, we focused on Hinglish as an illustrative case of code-mixed SRH queries, while many other languages and code-mixing patterns are used in India. Second, the hierarchical intent schema was developed for our setting and user population; researchers working in other regions should adapt label definitions to their context and target users. Third, although we evaluated many more models, several were excluded from the final set presented in this paper, either due to poor performance on code-mixed data or the cost of highly parameterized models. Fourth, our dataset is limited in size; certain topics and subtopics are underrepresented, resulting in lower statistical power for these categories, and performance estimates for these topics should be interpreted as preliminary. Fifth, our evaluation was conducted in a strict zero-shot setting, where models were not provided with the detailed subtopic definitions used by human annotators. This design allows us to assess how models infer intent from real-world, code-mixed queries without semantic scaffolding. Finally, our evaluation treats intent as a single topic, even though some queries plausibly map to more than 1 topic, motivating the exploration of alternative modeling approaches that can better capture such ambiguity.

XSL•FO
RenderX

In the future, we plan to extend our work to support additional Indian languages and code-mixing patterns, enriching both the breadth of topics and the depth of subtopic coverage in collaboration with the Myna Mahila Foundation. We also plan to evaluate in-context prompting strategies and to explore ranked classification [67] and hierarchical selective classification [64] approaches, which allow models to better account for ambiguity in hierarchical intent prediction.

Implications

This study highlights the importance of hierarchical classification for SRH applications in low-resource settings. By capturing intent at both the topic and subtopic levels, this framework improves contextual precision, supports safer dialogue systems, and enables public health organizations to design targeted interventions. At the same time, our findings show that LLMs often struggle with euphemisms and sociocultural framing, highlighting the need for systems that are socially attuned as well as technically accurate. Our evaluation offers a starting benchmark that can guide the development of culturally aligned, open-source LLM tools for low-resource health care contexts. More broadly, this work contributes to the discourse on health equity in AI by addressing the linguistic and cultural barriers that limit access to reliable health information for marginalized populations.

Conclusions

In this study, we evaluated the zero-shot performance of proprietary, open-weight multilingual, and instruction-tuned Indic LLMs on the hierarchical classification of SRH queries. Our findings reveal that the proprietary model (GPT-5) was the most reliable, while Sarvam-M emerged as the strongest Indic model and performed competitively with Claude-3.5-Sonnet and large open-weight models such as Llama-3.3-70B-Instruct. This highlights the importance of culturally aligned and domain-relevant training data alongside model scale. Our error analysis further revealed that models frequently misclassified queries that involved euphemisms, cultural or religious language, or time-sensitive concerns, underscoring the importance of capturing nuance at both topic and subtopic levels. Misclassifications in these areas risk unsafe guidance and reinforcing stigma, particularly around sensitive issues such as abortion and infertility. By introducing a benchmark evaluation framework, this study supports the development of open-source multilingual models for SRH and advances culturally aligned, socially responsive AI systems to better serve the health information needs of underserved communities.

The source code used in this study, including scripts for model evaluation, hierarchical metric calculation (hF_1), bootstrap resampling, and CMI calculation, is publicly available at a GitHub repository [68] to support future research.

Acknowledgments

We sincerely thank the Myna Mahila Foundation for their invaluable support throughout this study. We would also like to thank the community members who were engaged in the data collection process with the Myna Mahila Foundation, and the editor and reviewers for their thoughtful and detailed comments.

Funding

This research was made possible with the support of the Google Award for Inclusion Research and the Agency Fund through Myna Mahila USA (SA24-03-03).

Data Availability

The sexual and reproductive health queries (SRHQ)-India dataset introduced in this study contains sensitive queries related to sexual and reproductive health, expressed in code-mixed Hinglish. To ensure privacy and ethical use, the dataset will not be publicly released in full. However, deidentified subsets or access to the full dataset can be provided upon reasonable request to the corresponding author for academic and noncommercial research purposes, subject to data use agreements and ethical review.

Authors' Contributions

SKD and MS conceptualized the study and designed the data framework. AT and MS performed data annotation. SKD carried out the formal analysis, developed the methodology, and created visualizations. ZM, SJ, TD, and MS provided resources for the study. AI and SJ supervised the research. The original draft was prepared by SKD, and all authors contributed to reviewing and editing the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Annotation guidelines for hierarchical topic-subtopic labeling of the sexual and reproductive health queries (SRHQ-India dataset).
[DOCX File, 46 KB - jmir_v28i1e86545_app1.docx]

Multimedia Appendix 2

XSL•FO
RenderX

Prompt template and statistical performance analysis of zero-shot hierarchical classification performance across open-weight, Indic, and proprietary large language models (LLMs), including a comparative analysis of GPT-5 and GPT-4o on representative user queries.
[DOCX File, 36 KB - jmir_v28i1e86545_app2.docx]

Multimedia Appendix 3
Model-generated reasons for predicted hierarchical classifications across representative sexual and reproductive health (SRH) queries, reported for all evaluated models.
[DOCX File, 38 KB - jmir_v28i1e86545_app3.docx]

References

1. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors 2023 May 17;10:e47564. [doi: 10.2196/47564] [Medline: 37195756]
2. Wang W, Xu B, Fang T, Zhang L, Song Y. KnowComp at semeval-2023 task 7: fine-tuning pre-trained language models for clinical trial entailment identification. In: E S, AS D, H TM, R K, E S, editors. Presented at: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023); Jul 13-14, 2023; Toronto, Canada p. 1-9 URL: https://aclanthology.org/2023.semeval-1 [accessed 2026-03-11] [doi: 10.18653/v1/2023.semeval-1.1]
3. Verma G, Mujumdar R, Wang ZJ, Choudhury MD, Kumar S. Overcoming language disparity in online content classification with multimodal learning. 2022 Presented at: Proceedings of the International AAAI Conference on Web and Social Media; Jun 6-9, 2022; Atlanta, GA p. 1040-1051. [doi: 10.1609/icwsm.v16i1.19356]
4. Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. arXiv. Preprint posted online on Feb 27, 2023. [doi: 10.48550/arXiv.2302.13971]
5. OpenAI AJ, Adler S, et al. GPT-4 technical report. arXiv. Preprint posted online on Mar 15, 2024. [doi: 10.48550/arXiv.2303.08774]
6. Das M, Angeli F, Krumeich AJSM, van Schayck OCP. The gendered experience with respect to health-seeking behaviour in an urban slum of Kolkata, India. Int J Equity Health 2018 Feb 14;17(1):24. [doi: 10.1186/s12939-018-0738-8] [Medline: 29444674]
7. Deva R, Ramani D, Divate T, Jalota S, Ismail A. “Kya family planning after marriage hoti hai?”: integrating cultural sensitivity in an LLM chatbot for reproductive health. Presented at: CHI ’25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems; Apr 26 to May 1, 2025; Yokohama Japan p. 1-23 URL: https://dl.acm.org/doi/proceedings/10.1145/3706598 [doi: 10.1145/3706598.3713362]
8. Omar M, Soffer S, Agbareia R, et al. Sociodemographic biases in medical decision making by large language models. Nat Med 2025 Jun;31(6):1873-1881. [doi: 10.1038/s41591-025-03626-6] [Medline: 40195448]
9. Purnat TD, Wilhelm E, Scales D, et al. Impacts of sexual and reproductive health and rights misinformation in digital spaces on human rights protection and promotion: scoping review. JMIR Infodemiology 2025 Dec 30;5:e83747. [doi: 10.2196/83747] [Medline: 41468582]
10. John JN, Gorman S, Scales D, Gorman J. Online misleading information about women’s reproductive health: a narrative review. J Gen Intern Med 2025 Apr;40(5):1123-1131. [doi: 10.1007/s11606-024-09118-6] [Medline: 39511120]
11. Singh S, Shekhar C, Acharya R, et al. The incidence of abortion and unintended pregnancy in India, 2015. Lancet Glob Health 2018 Jan;6(1):e111-e120. [doi: 10.1016/S2214-109X(17)30453-9] [Medline: 29241602]
12. Singh SM, Garg SK, Misra A, Seth A, Chakraborty T. SUKHSANDESH: an avatar therapeutic question answering platform for sexual education in rural india. In: Larson K, editor. Presented at: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24); Aug 3-9, 2024; Jeju, South Korea p. 7465-7473. [doi: 10.24963/ijcai.2024/826]
13. Singh D, Singh T, Kumar D. Complications during pregnancy among Indian tribal women: a mini-review. JMMS 2025;2(2):128-131. [doi: 10.51219/JMMS/Kumar-D/25]
14. Joyce CM, Sharma D, Mukherji A, Nandi A. Socioeconomic inequalities in adverse pregnancy outcomes in India: 2004-2019. PLOS Glob Public Health 2024;4(9):e0003701. [doi: 10.1371/journal.pgph.0003701] [Medline: 39292712]
15. Wang H, Gupta S, Singhal A, et al. An artificial intelligence chatbot for young people’s sexual and reproductive health in India (SnehAI): instrumental case study. J Med Internet Res 2022 Jan 3;24(1):e29969. [doi: 10.2196/29969] [Medline: 34982034]
16. Patwa P, Aguilar G, Kar S, et al. SemEval-2020 task 9: overview of sentiment analysis of code-mixed tweets. In: Herbelot A, Zhu X, Palmer A, Schneider N, May J, Shutova E, editors. Presented at: Proceedings of the Fourteenth Workshop on Semantic Evaluation; Dec 12-13, 2020; Barcelona (online) p. 774-790 URL: https://aclanthology.org/2020.semeval-1 [doi: 10.18653/v1/2020.semeval-1.100]
17. Parshad RD, Bhowmick S, Chand V, Kumari N, Sinha N. What is India speaking? Exploring the “Hinglish” invasion. Phys A: Stat Mech Appl 2016 May;449:375-389. [doi: 10.1016/j.physa.2016.01.015]

XSL•FO
RenderX

18. Sengupta A, Das S, Akhtar MS, Chakraborty T. Social, economic, and demographic factors drive the emergence of Hinglish code-mixing on social media. Humanit Soc Sci Commun 2024;11(1):606. [doi: 10.1057/s41599-024-03058-6]
19. Heredia M, Labaka G, Barnes J, Soroa A. Conditioning llms to generate code-switched text. arXiv. Preprint posted online on Feb 18, 2025. [doi: 10.48550/arXiv.2502.12924]
20. Yong ZX, Zhang R, Forde J, et al. Prompting multilingual large language models to generate code-mixed texts: the case of south east asian languages. Presented at: Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching; Dec 6-10, 2023; Singapore p. 43-63 URL: https://aclanthology.org/2023.calcs-1 [accessed 2026-02-02] [doi: 10.18653/v1/2023.calcs-1.5]
21. Moskovitch R, Cohen-Kashi S, Dror U, Levy I, Maimon A, Shahar Y. Multiple hierarchical classification of free-text clinical guidelines. Artif Intell Med 2006 Jul;37(3):177-190. [doi: 10.1016/j.artmed.2006.04.001] [Medline: 16730962]
22. Ee-Peng L. Hierarchical text classification and evaluation. Presented at: 2001 IEEE International Conference on Data Mining; Nov 29 to Dec 2, 2001; San Jose, CA p. 521-528. [doi: 10.1109/ICDM.2001.989560]
23. Kosmopoulos A, Partalas I, Gaussier E, Paliouras G, Androutsopoulos I. Evaluation measures for hierarchical classification: a unified view and novel approaches. Data Min Knowl Discov 2015 May;29(3):820-865. [doi: 10.1007/s10618-014-0382-x]
24. He B, Nag S, Cui L, et al. Hierarchical query classification in e-commerce search. Presented at: WWW ’24: Companion Proceedings of the ACM Web Conference 2024; May 13-17, 2024; Singapore p. 338-345 URL: https://dl.acm.org/doi/proceedings/10.1145/3589335 [accessed 2026-03-11] [doi: 10.1145/3589335.3648332]
25. Cao H, Hu DH, Shen D, et al. Context-aware query classification. Presented at: SIGIR ’09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval; Jul 19-23, 2009; Boston, MA p. 3-10 URL: https://dl.acm.org/doi/proceedings/10.1145/1571941 [accessed 2026-03-11] [doi: 10.1145/1571941.1571945]
26. Zhu M, Ahuja A, Wei W, Reddy CK. A hierarchical attention retrieval model for healthcare question answering. Presented at: WWW ’19: The World Wide Web Conference; May 13-17, 2019; San Francisco, CA p. 2472-2482 URL: https://dl.acm.org/doi/proceedings/10.1145/3308558 [accessed 2026-03-11] [doi: 10.1145/3308558.3313699]
27. Deng Y, Denecke K. Classification of user queries according to a hierarchical medical procedure encoding system using an ensemble classifier. Front Artif Intell 2022;5:1000283. [doi: 10.3389/frai.2022.1000283] [Medline: 36406473]
28. Sosnowski Ł, Żuławińska J, Dutta S, Szymusik I, Zyguła A, Bambul-Mazurek E. Artificial intelligence in personalized healthcare analysis for womens’ menstrual health disorders. Presented at: 2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS); Sep 4-7, 2022; Sofia, Bulgaria p. 751-760. [doi: 10.15439/2022F59]
29. Chen S, Bouadjenek MR, Naseem U, et al. Leveraging taxonomy and LLMs for improved multimodal hierarchical classification. In: Rambow O, Wanner L, Apidianaki M, Al-Khalifa H, Eugenio BD, Schockaert S, editors. Presented at: Proceedings of the 31st International Conference on Computational Linguistics Association for Computational Linguistics; Jan 19-24, 2025; Abu Dhabi, UAE p. 6244-6254 URL: https://aclanthology.org/2025.coling-main.417/ [accessed 2025-09-14]
30. GPT-5 system card. OpenAI. URL: https://cdn.openai.com/gpt-5-system-card.pdf [accessed 2025-09-20]
31. OpenAI HA, Lerer A, et al. GPT-4o system card. arXiv. Preprint posted online on Oct 25, 2024. [doi: 10.48550/arXiv.2410.21276]
32. Grattafiori A, Dubey A, Jauhri A, et al. The Llama 3 herd of models. arXiv. Preprint posted online on Jul 31, 2024. [doi: 10.48550/arXiv.2407.21783]
33. Kamath A, Ferret J, et al. Gemma 3 technical report. arXiv. Preprint posted online on Mar 25, 2025. [doi: 10.48550/arXiv.2503.19786]
34. Wang L, Wan Z, Ni C, et al. A systematic review of chatgpt and other conversational large language models in healthcare. medRxiv. Preprint posted online on Apr 27, 2024. [doi: 10.1101/2024.04.26.24306390]
35. Jin W, Mao H, Li Z, et al. Amazon-M2: a multilingual multi-locale shopping session dataset for recommendation and text generation. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S, editors. Presented at: NIPS ’23: Proceedings of the 37th International Conference on Neural Information Processing Systems; Dec 10-16, 2023; New Orleans, LA p. 8006-8026 URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/193df57a2366d032fb18dcac0698d09a-Paper-Datasets_and_Benchmarks.pdf [accessed 2026-03-11] [doi: 10.52202/075280-0351]
36. Jin Y, Chandra M, Verma G, Hu Y, De Choudhury M, Kumar S. Better to ask in English: cross-lingual evaluation of large language models for healthcare queries. Presented at: WWW ’24: Proceedings of the ACM Web Conference 2024; May 13-17, 2024; Singapore p. 2627-2638 URL: https://dl.acm.org/doi/proceedings/10.1145/3589334 [accessed 2026-03-11] [doi: 10.1145/3589334.3645643]
37. Tripathi N, Sekher TV. Youth in India ready for sex education? Emerging evidence from national surveys. In: Murthy AK, editor. PLoS ONE 2013;8(8):e71584. [doi: 10.1371/journal.pone.0071584] [Medline: 23951197]
38. Gumma V, Raghunath A, Jain M, Sitaram S. HEALTH-PARIKSHA: assessing RAG models for HEALTH chatbots in real-world multilingual settings. arXiv. Preprint posted online on Oct 17, 2024. [doi: 10.48550/arXiv.2410.13671]
39. Sivarajkumar S, Wang Y. HealthPrompt: a zero-shot learning paradigm for clinical natural language processing. AMIA Annu Symp Proc 2022;2022:972-981. [Medline: 37128372]

XSL•FO
RenderX

40. Guo Y, Ovadje A, Al-Garadi MA, Sarker A. Evaluating large language models for health-related text classification tasks with public social media data. J Am Med Inform Assoc 2024 Oct 1;31(10):2181-2189. [doi: 10.1093/jamia/ocae210] [Medline: 39121174]
41. Singh P, Lefever E. Sentiment analysis for hinglish code-mixed tweets by means of cross-lingual word embeddings. In: Solorio T, Choudhury M, Bali K, Sitaram S, Das A, Diab M, editors. Presented at: Proceedings of the 4th Workshop on Computational Approaches to Code Switching European Language Resources Association; May 11-16, 2020; Marseille, France p. 45-51 URL: https://aclanthology.org/2020.calcs-1.6/ [accessed 2025-12-17]
42. Chanda S, Mishra A, Pal S. Sentiment analysis of code-mixed Dravidian languages leveraging pretrained model and word-level language tag. Nat lang process 2025 Mar;31(2):477-499. [doi: 10.1017/nlp.2024.30]
43. Gambäck B, Das A, et al. Comparing the level of code-switching in corpora. In: Calzolari N, Choukri K, Declerck T, editors. Presented at: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) European Language Resources Association (ELRA); May 23-28, 2016; Portorož, Slovenia p. 1850-1855 URL: https://aclanthology.org/L16-1292/
44. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. Presented at: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers; Apr 3-7, 2017; Valencia, Spain p. 427-431 URL: http://aclweb.org/anthology/E17-2 [accessed 2025-02-02] [doi: 10.18653/v1/E17-2068]
45. Bañón M, Ramírez-Sánchez G, Zaragoza-Bernabeu J, FastSpell ORS. The langid magic spell. In: Calzolari N, Kan MY, Hoste V, Lenci A, Sakti S, Xue N, editors. Presented at: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) ELRA and ICCL; May 20-25, 2024; Torino, Italia p. 7133-7140 URL: https://aclanthology.org/2024.lrec-main.626/ [accessed 2026-03-11] [doi: 10.63317/3w889497us8n]
46. Ibrahim M. CUFE@NLU of devanagari script languages 2025: language identification using fasttext. In: Sarveswaran K, Vaidya A, Krishna Bal B, Shams S, Thapa S, editors. Presented at: Proceedings of the First Workshop on Challenges in Processing South Asian Languages; Jan 19-24, 2025; Abu Dhabi, UAE p. 273-277 URL: https://aclanthology.org/2025.chipsal-1.30/ [accessed 2026-03-11]
47. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr;20(1):37-46. [doi: 10.1177/001316446002000104]
48. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005 May;37(5):360-363. [Medline: 15883903]
49. Jiang AQ, Sablayrolles A, Roux A, et al. Mixtral of experts. arXiv. Preprint posted online on Jan 8, 2024. [doi: 10.48550/arXiv.2401.04088]
50. Riviere M, Pathak S, et al. Gemma 2: improving open language models at a practical size. arXiv. Preprint posted online on Jul 31, 2024. [doi: 10.48550/arXiv.2408.00118]
51. Qwen YA, Yang B, et al. Qwen2.5 technical report. arXiv. Preprint posted online on Dec 19, 2025. [doi: 10.48550/arXiv.2412.15115]
52. Dang J, Singh S, D'souza D, et al. Aya expanse: combining research breakthroughs for a new multilingual frontier. arXiv. Preprint posted online on Dec 5, 2024. [doi: 10.48550/arXiv.2412.04261]
53. Gala J, Jayakumar T, Husain JA, et al. Airavata: introducing hindi instruction-tuned LLM. arXiv. Preprint posted online on Jan 26, 2024. [doi: 10.48550/arXiv.2401.15006]
54. GenVRadmin/aryabhatta-gemmagenz-vikas-merged. Hugging Face. URL: https://huggingface.co/GenVRadmin/AryaBhatta-GemmaGenZ-Vikas-Merged [accessed 2026-03-11]
55. Cognitive-lab/llama3-gaja-hindi-8B-v01hugging face. Hugging Face. URL: https://huggingface.co/Cognitive-Lab/LLama3-Gaja-Hindi-8B-v0.1 [accessed 2025-10-27]
56. Kallappa A, Kamble P, Ravi A, et al. Krutrim LLM: multilingual foundational model for over a billion people. arXiv. Preprint posted online on Feb 10, 2025. [doi: 10.48550/arXiv.2502.09642]
57. Claude 3.5 Sonnet. Anthropic. URL: https://www.anthropic.com/news/claude-3-5-sonnet [accessed 2025-09-20]
58. Sarvam-M. Sarvam. URL: https://www.sarvam.ai/blogs/sarvam-m [accessed 2025-09-20]
59. Zhan P, Xu Z, Tan Q, Song J, Xie R. Unveiling the lexical sensitivity of llms: combinatorial optimization for prompt enhancement. In: Al-Onaizan Y, Bansal M, Chen YN, editors. Presented at: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Nov 12-16, 2024; Miami, FL p. 5128-5154 URL: https://aclanthology.org/2024.emnlp-main [accessed 2026-03-11] [doi: 10.18653/v1/2024.emnlp-main.295]
60. Bhattacharya S, Bashar MA, Singh A. So near, yet so far: access to safe abortion services remains elusive for poor women in India. BMJ Case Rep 2017 Oct 13;2017:bcr2017220980. [doi: 10.1136/bcr-2017-220980] [Medline: 29030364]
61. Roberts L, Renati S, Solomon S, Montgomery S. Women and infertility in a pronatalist culture: mental health in the slums of Mumbai. Int J Womens Health 2020;12:993-1003. [doi: 10.2147/IJWH.S273149] [Medline: 33192102]
62. Patra S, Unisa S. Addressing reproductive health knowledge, infertility and coping strategies among rural women in India. J Biosoc Sci 2021 Jul;53(4):557-565. [doi: 10.1017/S0021932020000371] [Medline: 32677598]

XSL•FO
RenderX

63. Husain Z, Dutta M, Ghosh S. Contraceptive use among illiterate women in India: does proximate illiteracy matter?.: MPRA Pap; 2011. URL: https://ideas.repec.org//p/pra/mprapa/30790.html [accessed 2025-10-16]
64. El-Yaniv R, Galil I, Goren S. Hierarchical selective classification. Presented at: Advances in Neural Information Processing Systems 37; Dec 10-15, 2024 URL: http://www.proceedings.com/79017.html [accessed 2026-03-11] [doi: 10.52202/079017-3526]
65. Roustan D, Bastardot F. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. Interact J Med Res 2025 Jan 28;14:e59823. [doi: 10.2196/59823] [Medline: 39874574]
66. Yang K, Tao G, Chen X, Xu J. Alleviating the fear of losing alignment in LLM fine-tuning. Presented at: 2025 IEEE Symposium on Security and Privacy (SP); May 12-14, 2025; San Francisco, CA p. 2152-2170. [doi: 10.1109/SP61157.2025.00171]
67. Atapour-Abarghouei A, Bonner S, McGough AS. Rank over class: the untapped potential of ranking in natural language processing. Presented at: 2021 IEEE International Conference on Big Data (Big Data); Dec 15-18, 2021; Orlando, FL p. 3950-3959. [doi: 10.1109/BigData52589.2021.9671386]
68. EmoryCareLab/hierarchical-srh-intent. GitHub. 2026. URL: https://github.com/EmoryCareLab/hierarchical-srh-intent [accessed 2026-02-02]

Abbreviations

AI: artificial intelligence
CMI: Code-Mixing Index
IVF: in-vitro fertilization
LLM: large language model
PNC: postnatal care
SRH: sexual and reproductive health
SRHQ: sexual and reproductive health queries

Edited by A Coristine; submitted 30.Oct.2025; peer-reviewed by AK Singh, FT Brito, M Borovic; accepted 18.Feb.2026; published 24.Mar.2026.

Please cite as:
Dey SK, S M, Thapa A, Shah M, Mehta Z, Kapile SK, Divate T, Jalota S, Ismail A
Understanding User Intent in Code-Mixed Sexual and Reproductive Health Queries in Urban India: Hierarchical Classification Approach Using Large Language Models
J Med Internet Res 2026;28:e86545
URL: https://www.jmir.org/2026/1/e86545
doi:10.2196/86545

© Sumon Kanti Dey, Manvi S, Aradhana Thapa, Meet Shah, Zeel Mehta, Shraddha Kale Kapile, Tanvi Divate, Suhani Jalota, Azra Ismail. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Combining Machine Learning Models and Screening to Enhance Suicide Risk Identification for American Indian Patients: Retrospective Cohort Study

Novalene Alsenay Goklish[1*], PhD; Emily E Haroz[1*], PhD; Rohan R Dayal[1], MPH; Valentín Q Sierra[1], PhD; Roy Adams[2], PhD; Francene Larzelere Sinquah[1], PhD; Paul Rebman[3], MPH; Jacob L Taylor[2], MD, MPH

[1]Department of International Health, Center for Indigenous Health, Johns Hopkins Bloomberg School of Public Health, 415 N. Washington St, Baltimore, MD, United States

[2]Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, MD, United States

[3]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

[*]these authors contributed equally

Corresponding Author:
Emily E Haroz, PhD
Department of International Health, Center for Indigenous Health, Johns Hopkins Bloomberg School of Public Health, 415 N. Washington St, Baltimore, MD, United States

Abstract

Background: American Indian and Alaska Native communities experience disproportionately high suicide rates. While machine learning (ML) models leveraging electronic health records have emerged as promising tools for suicide risk identification, the optimal integration of these models with existing screening practices remains unclear.

Objective: The objective of this study was to compare parallel and serial testing strategies that combine an ML suicide risk model and the Ask Suicide-Screening Questions (ASQ) against using the ASQ alone. To achieve this, we conducted a retrospective secondary analysis of electronic health record data. The cohort consisted of adult emergency department visits at an Indian Health Service facility between October 1, 2019, and October 2, 2021.

Methods: Sensitivity, specificity, predictive values, and 95% CIs were averaged across 10 cross-validated patient-level folds. The final sample included 7897 American Indian patients with 26,896 visits, 824 (3.1%) of which had a positive ASQ result and 102 (0.4%) of which had the outcome of suicide attempt or death within 90 days of the visit. The logistic regression ML model previously developed using Indian Health Service–specific data was operationalized at the 95th and 75th percentiles to evaluate high-risk and medium-risk thresholds, respectively. A sensitivity analysis was performed to evaluate identification approaches across all emergency department visits during this period.

Results: The ML medium-risk threshold alone identified the most true positives (sensitivity: 0.782, 95% CI 0.648-0.915; specificity: 0.751, 95% CI 0.725-0.777; positive predictive value [PPV]: 0.012, 95% CI 0.009-0.014; negative predictive value [NPV]: 0.999, 95% CI 0.998-0.999) in comparison to the ML high-risk threshold alone (sensitivity: 0.429, 95% CI 0.287-0.572; specificity: 0.955, 95% CI 0.948-0.961; PPV: 0.035, 95% CI 0.022-0.048; NPV: 0.998, 95% CI 0.997-0.999) or the ASQ alone (sensitivity: 0.178, 95% CI 0.073-0.282; specificity: 0.970, 95% CI 0.968-0.971; PPV: 0.022, 95% CI 0.010-0.034; NPV: 0.997, 95% CI 0.996-0.998). Combining the ML high-risk threshold with the ASQ in series yielded the greatest positive predictive ability (PPV: 0.050, 95% CI 0.014-0.086) at the cost of reduced sensitivity (0.129, 95% CI 0.036-0.221). Finally, the parallel testing approach using the ML medium-risk threshold yielded the greatest sensitivity (0.795, 95% CI 0.671-0.920; specificity: 0.742, 95% CI 0.716-0.767; PPV: 0.012, 95% CI 0.009-0.014; NPV: 0.999, 95% CI 0.998-0.999) without missing any cases identified through screening.

Conclusions: Unlike existing studies that evaluate ML and screening tools in isolation, this study innovates by assessing combined parallel and serial testing strategies in a real-world setting. We demonstrated that, while serial testing maximizes predictive accuracy, it is often infeasible. Instead, parallel testing brings value as a clinical "safety net" to catch at-risk patients missed by standard practices. Ultimately, integrating ML in suicide prevention requires balancing statistical accuracy with setting-specific, real-world workflows.

(J Med Internet Res 2026;28:e82669) doi:10.2196/82669

XSL•FO
RenderX

KEYWORDS

suicide prevention; machine learning; American Indian and Alaska Native; screening; electronic health records; predictive modeling; suicide risk identification

Introduction

For the past 20 years, suicide rates in the United States have been steadily climbing. This is particularly true for American Indian and Alaska Native individuals, who face the highest rates of any racial or ethnic group and have experienced the steepest increases. In 2023, the US age-adjusted suicide rate was 14.1 per 100,000. However, the rate among American Indian and Alaska Native individuals was 23.8 per 100,000, representing a 44.2% increase since 2011 [1]. The burden of suicide in American Indian and Alaska Native populations reflects the confluence of historical, cultural, social, and economic factors that perpetuate intergenerational trauma and disparities in mental health resources. The devastating consequences of suicide on individuals, families, and communities underscore the urgent need for effective prevention strategies tailored to these populations.

Identifying individuals at risk of suicide remains a critical challenge in prevention efforts. Screening for suicide risk is now widely recommended as a critical strategy. Given the fact that most people who die by suicide saw a health care provider in the previous year [2], screening people at health care visits is designed to catch people before they engage in suicidal behaviors [3]. Screening can identify at-risk individuals who can then be connected with appropriate interventions, crisis services, or enhanced monitoring and support. Given this, the Joint Commission established National Patient Safety Goal 15.01.01, which requires accredited hospitals and behavioral health organizations to screen at-risk patients using evidence-based tools [4]. Others in the field of suicide prevention have argued for screening regardless of patient risk factors (universal screening) [5] with the notion that asking people about their past or current suicidal thoughts and behaviors does not confer risk and can provide an avenue for disclosure that may otherwise not be available. However, at the current time, the US Preventive Services Task Force has noted insufficient evidence to establish the benefits of universal suicide risk screening [6].

In parallel to advancements in screening, machine learning (ML) models using electronic health record (EHR) data have emerged as promising tools for suicide risk identification. Several large health systems, including the Veterans Health Administration [7], Kaiser Permanente [8], and academic medical centers [9], have developed and implemented EHR-based models that combine large numbers of patient variables to identify patterns and risk factors. These models have demonstrated promise, but not without their share of critiques as well. While predictive models perform generally comparably to existing screening tools [10,11], others have argued that they still suffer from too high a false-positive rate to engender trust and that full assessment by a trained clinician is the only way to ensure appropriate care [12]. However, clinicians have also been notably inaccurate [13-15], and their availability is limited in many places. Ultimately, the validity of predictive models depends critically on the quality and relevance of the data from which they are developed, and regardless of the risk identification strategy, implementation considerations are key to ensure seamless integration with existing care structures and workflows.

Despite the burden of suicide in American Indian and Alaska Native communities, there are few validated suicide risk identification tools and empirically supported interventions specifically designed and tested in American Indian and Alaska Native communities. Our research group has worked to address this gap through partnership with tribes in the US southwest and the Indian Health Service (IHS). Previously, we tested the performance of existing ML models developed by the Mental Health Research Network (MHRN) and Vanderbilt University in a majority American Indian population. The MHRN primary care model demonstrated good discrimination (area under the receiver operating characteristic curve [AUROC]=0.81) and outperformed both the Vanderbilt model and an augmented screening indicator (ie, Ask Suicide-Screening Questions [ASQ] acute positive or nonacute positive screen, a positive suicide risk assessment on the Columbia-Suicide Severity Rating Scale, 90-day history of ideation, or 5-year history of attempt) [16]. Building on these findings, we developed a context-specific suicide risk model using logistic regression and random forest approaches. These models, trained on EHR data, both achieved an AUROC of 0.83 for predicting suicide attempts or deaths within 90 days and again performed better than an augmented version of current practice [17].

However, a critical question remains in the context of implementation of screening tools: how should ML models be integrated with existing screening practices in clinical workflows? The optimal combination of these models and screening is unclear [18]. Several recent studies have begun to explore this question in other populations. Wilimitis et al [19] evaluated the integration of face-to-face screening with real-time ML in an adult primary care setting, finding that combined approaches improved risk identification. Aseltine et al [20] compared screening and risk algorithms in a pediatric emergency department (ED), demonstrating that the algorithms achieved higher sensitivity than screening when identifying the same proportion of patients as at risk. However, no studies have examined combined approaches specifically, and particularly not in American Indian and Alaska Native health settings.

This study addresses this gap by evaluating parallel and serial testing strategies that combine an IHS-specific ML model with the ASQ screening tool. We examined these approaches in ED visits. We aimed to (1) evaluate parallel testing strategies combining both approaches and (2) assess serial testing strategies that use ML to guide screening decisions. On the basis of the ML model's ability to leverage longitudinal EHR data and prior studies showing that algorithms can add value to screening, we hypothesized that the ML model would identify more patients at risk than screening alone. We expected that

XSL•FO
RenderX

parallel approaches would maximize case detection by capturing individuals identified through either method, whereas serial approaches would reduce false positives by requiring agreement between both tools, although potentially at the cost of missing some at-risk individuals. Our ultimate goal was to inform how the implementation of these tools can enhance suicide risk identification while being feasible within existing clinical workflows.

Methods

Overview

This study was a retrospective analysis conducted to evaluate the integration of an ML suicide risk model with suicide risk screening in an ED setting. The analysis was conducted as part of the Native-RISE (Risk Identification for Suicide and Enhanced Care) project, a multiyear partnership with several tribes in the US southwest. Building on prior work developing and validating ML models [4,5], we examined how parallel and serial testing strategies combining the model with the ASQ tool perform in identifying patients at risk of suicide attempts or deaths within 90 days of ED visits.

Study Setting and Population

This study included 1 IHS service unit. The facility serves a geographically defined population of approximately 17,000 individuals and provides comprehensive health care services including emergency, outpatient, inpatient, and behavioral health care. The study population included all patients aged 18 years or older who presented to the ED between October 1, 2019, and October 2, 2021, and had a documented ASQ screening result. During this period, universal suicide risk screening using the ASQ was implemented in the ED. Patients were excluded if they did not have at least 90 days of follow-up time after their visit or if the visit type was not relevant for point-of-care suicide risk identification (eg, visits solely for vaccine administration or administrative record updates). All demographic characteristics, including race and ethnicity, were obtained from the EHR.

The ML Model

The ML model used in this study was developed specifically for use in IHS settings and has been described in detail elsewhere [17]. Briefly, the model was developed using data from all adult patients (≥18 years) with visits between January 1, 2017, and October 2, 2021, at the same IHS facility (outcome data were included until December 31, 2021). Two modeling approaches were tested—logistic regression and random forest—both of which achieved comparable performance (AUROC=0.83). For this analysis, we used the logistic regression model given its greater interpretability.

Model features included demographics (age, sex, and race and ethnicity), clinical diagnoses (mental health conditions, substance use disorders, and chronic medical conditions), medications (psychotropic medications and pain medications), health care use patterns (visit frequency, ED visits, and hospitalizations), and prior suicide-related encounters (screening results, documented suicidal ideation, and prior attempts). Features were constructed using look-back periods ranging from 30 days to 5 years prior to each visit. Diagnosis codes were identified using *International Classification of Diseases, 10th Revision*, codes and established code lists from the MHRN.

The model generates a predicted probability of suicide attempt or death within 90 days following each ED visit. For this analysis, we examined model performance at 2 risk thresholds: the 95th percentile (high risk) and the 75th percentile (medium or high risk). These thresholds were selected to balance sensitivity and positive predictive value (PPV) while maintaining clinical feasibility.

The ASQ Tool

The ASQ is a brief, 4-item screening tool developed by the National Institute of Mental Health for use in medical settings. Patients are asked about (1) wishing to die in the previous few weeks, (2) feeling that they would be better off dead in the previous few weeks, (3) thoughts about killing themselves in the previous week, and (4) any previous suicide attempts [21]. Any affirmative response results in a fifth acuity question that assesses current suicidal ideation—positive responses result in an acute positive classification, and those without current suicidal ideation are classified as nonacute positive. For our analysis, we combined nonacute positive and acute positive responses into 1 category of positive ASQ screening results to estimate summary statistics. In our setting, patients screened as acute or nonacute positive receive additional risk assessment by behavioral health staff.

The ASQ has demonstrated validity in pediatric EDs and medical settings, with high concordance with more extensive assessments [22-24]. However, the ASQ has never been validated in American Indian and Alaska Native populations. Implementation challenges can include when and how the tool is administered and completion rates. In our study setting, the ASQ was either completed on paper or administered verbally. Documented completion during the study period was 62.7% (26,896/42,915).

Testing Strategies

In addition to estimating the impact of the ASQ or ML approaches alone, we evaluated 2 approaches for combining ML and screening, described in the following sections.

Parallel Testing

Patients were classified as at risk if they were screened as positive on the ASQ *or* if the ML model classified them as medium or high risk (≥75th percentile) or high risk (≥95th percentile). This approach uses both tools simultaneously, with either tool being sufficient to flag risk. Figure 1A shows a parallel approach in which the ML model verifies negative screens or assists when screening is missed.

XSL•FO
RenderX

Figure 1. (A) Pragmatic approach to parallel testing: the machine learning model is used to verify negative screens or assist when screening was negative or did not occur. (B) A hypothetical approach: this approach may be optimal because the machine learning model has higher sensitivity and moderate specificity and is noninvasive and easier to administer. ASQ: Ask Suicide-Screening Questions.

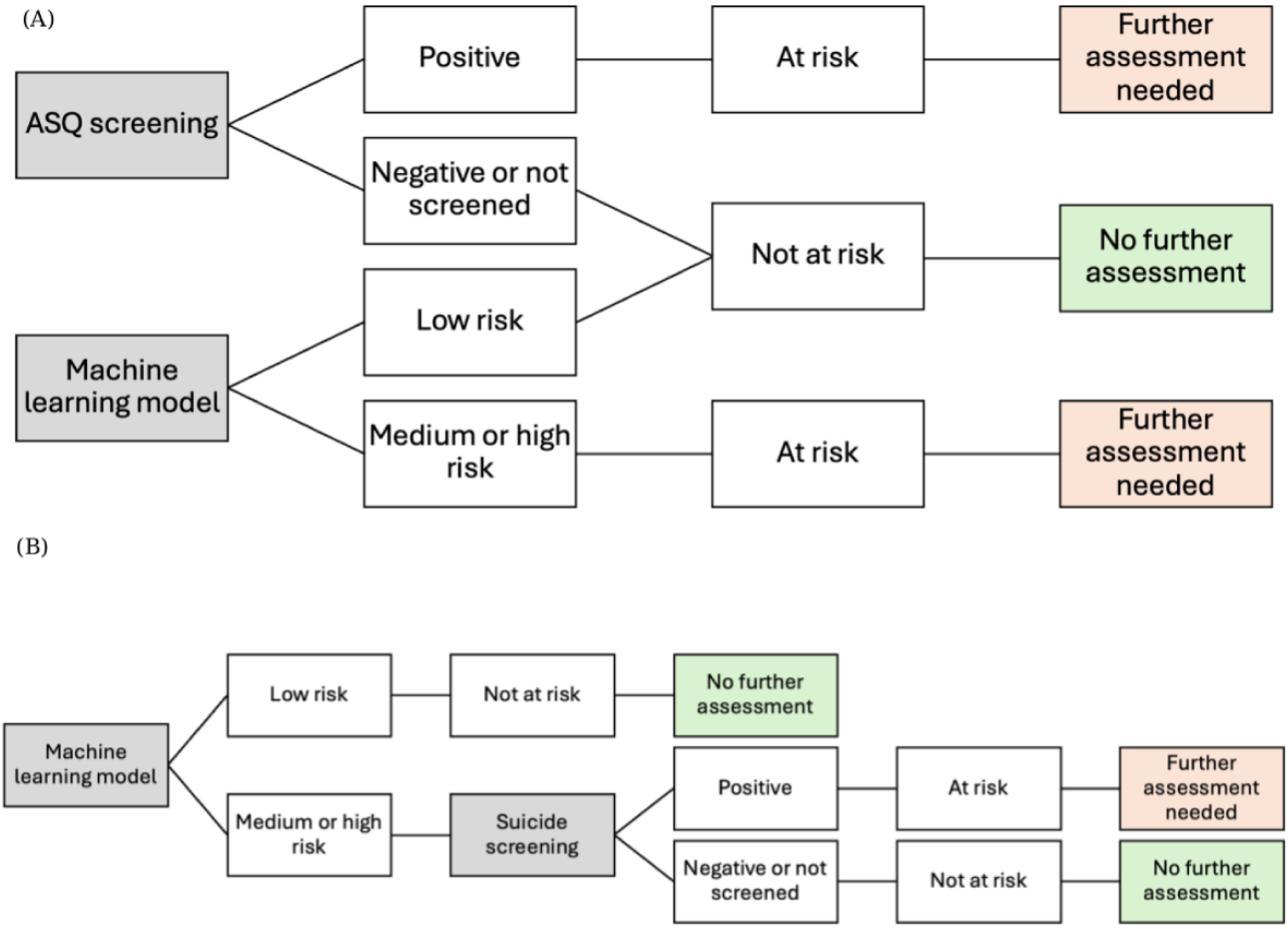

Serial Testing

Patients were classified as at risk if the ML model classified them as medium or high risk (≥75th percentile) or high risk (≥95th percentile) *and* they were screened as positive on the ASQ. This approach uses the ML model as a first-stage filter, with screening administered only to those flagged by the model. Figure 1B shows this hypothetical approach.

Outcomes

The primary outcome was suicide attempt or death by suicide within 90 days following the ED visit. Suicide attempts were identified using *International Classification of Diseases, 10th Revision*, diagnosis codes for intentional self-harm [17]. Suicide deaths were identified through the Tribe's community-based surveillance system, which maintains comprehensive mortality records including cause of death determinations [25]. Data were right censored on December 31, 2021, to allow for complete 90-day follow-up periods. Serial and parallel testing approaches were coded empirically using the medium- and high-risk thresholds for the ML model and ASQ results recorded in the EHR.

Statistical Analysis

For each testing strategy, we assessed the sensitivity, specificity, PPV, and negative predictive value (NPV) with 95% CIs. Performance metrics were calculated averaging across the 10 cross-validation folds created at the patient level used during model development, with 95% CIs calculated for point estimates using cross-validated SE estimates across folds.

We compared performance across testing strategies descriptively given the rarity of the outcome events, focusing on the trade-offs between sensitivity and PPV. Given the low base rate of suicide attempts and deaths (102/26,896, 0.4% of visits), we paid particular attention to PPV as this metric is driven by underlying prevalence of the outcome, indicates the proportion of flagged patients who experience the outcome, and directly relates to resource allocation and clinical burden.

Sensitivity Analysis

We conducted a sensitivity analysis replicating all analyses in the full sample of ED visits including those without documented ASQ screening (n=42,915 ED visits). This analysis provides insights into real-world performance, where screening completion rates may vary. Results are shown in Multimedia Appendix 1.

Ethical Considerations

Human Subject Ethics Review, Exemptions, or Approvals

The study was reviewed and declared non–human subject research by the Institutional Review Board of Johns Hopkins Bloomberg School of Public Health (17333) and the Phoenix Area IHS Office Institutional Review Board (20.05). The Tribe

XSL•FO
RenderX

approved the study through the Tribal Health Board and Tribal Council. The reporting of findings abides by STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [7].

Informed Consent

Informed consent was waived through a HIPAA (Health Insurance Portability and Accountability Act) authorization granted to the research team by the IHS Institutional Review Board for secondary data analysis of EHR data.

Privacy and Confidentiality

Research data were accessed through a data use agreement between the IHS and Johns Hopkins University. These data were protected during storage and analysis using a secure and HIPAA-compliant computing platform provided by the Johns Hopkins University School of Medicine. Research data were deidentified after preprocessing to ensure privacy and confidentiality. Patient-level statistics with substantially small sample sizes (n<5) were omitted to ensure that potentially identifying information was not included in this manuscript.

Compensation Details

Compensation for this study is not applicable as this was a secondary analysis of EHR data.

Results

Study Population

Between October 1, 2019, and October 2, 2021, there were 42,915 ED visits. Of these 42,915 visits, 26,896 (62.7%) had documented ASQ screening results and were included in the primary analysis. These visits represented 7897 unique patients, with a median of 2 (IQR 1-4) ED visits per patient. Patient demographics are presented in Table 1. The median age was 38 (IQR 28-54) years. Most patients were female (4241/7897, 53.7%) and identified as American Indian (7734/7897, 97.9%). Of the 7897 patients, 614 (7.8%) were screened as positive on the ASQ at one or more of their ED visits during the study period.

Table . Patient and visit characteristics (n=7897).

Characteristics	Values
Age (y), median (IQR)	38 (28-54)
Female, n (%)	4241 (53.7)
Race or ethnicity, n (%)	
American Indian	7734 (97.9)
Other or unknown	163 (2.1)
Any positive ASQ[a] screen, n (%)	614 (7.8)
ED[b] visits (n=26,896)	
Visits per patient, median (IQR)	2 (1-4)
Positive ASQ screen, n (%)	824 (3.1)

[a]ASQ: Ask Suicide-Screening Questions.

[b]ED: emergency department.

Outcomes

Suicide attempts or deaths within 90 days of an ED visit occurred in 0.4% (102/26,896) of the visits among 0.7% (55/7734) of unique patients. Of the 55 patients who experienced an outcome, 14 (25.5%) had been screened as positive on the ASQ at least once in the 90 days prior to a suicide attempt or death, whereas 41 (74.5%) were never screened as positive on the ASQ at an ED visit within 90 days of a suicide attempt. No one who died by suicide was screened as positive on the ASQ at any visit within 90 days of their death.

Performance of Individual Approaches

Table 2 shows the sensitivity, specificity, PPV, and NPV for each testing strategy. The ASQ screening alone exhibited a sensitivity of 0.178 (95% CI 0.073-0.282), specificity of 0.970 (95% CI 0.968-0.971), PPV of 0.022 (95% CI 0.010-0.034), and NPV of 0.997 (95% CI 0.996-0.998).

The ML model alone using the 75th percentile threshold (medium or high risk) achieved a sensitivity of 0.782 (95% CI 0.648-0.915), specificity of 0.751 (95% CI 0.725-0.777), PPV of 0.012 (95% CI 0.009-0.014), and NPV of 0.999 (95% CI 0.998-0.999). At the more stringent 95th percentile threshold (high risk only), the model achieved a sensitivity of 0.429 (95% CI 0.287-0.572), specificity of 0.955 (95% CI 0.948-0.961), PPV of 0.035 (95% CI 0.022-0.048), and NPV of 0.998 (95% CI 0.997-0.999). The ML model outperformed ASQ screening on sensitivity and NPV at the evaluated thresholds; however, because the model included prior screening and ideation features, this comparison reflects the incremental value of longitudinal EHR information rather than a contemporaneous head-to-head assessment of the index screening alone.

XSL•FO
RenderX

Table . Parallel and serial testing of machine learning (ML) model risk categories and suicide screening among visits with an administered suicide screening survey (n=7897 patients; n=26,896 emergency department visits).

	Sensitivity (95% CI)	Specificity (95% CI)	PPV[a] (95% CI)	NPV[b] (95% CI)
Classification alone				
ML model high risk (≥95th percentile)	0.429 (0.287-0.572)	0.955 (0.948-0.961)	0.035 (0.022-0.048)	0.998 (0.997-0.999)
ML model medium or high risk (≥75th percentile)	0.782 (0.648-0.915)	0.751 (0.725-0.777)	0.012 (0.009-0.014)	0.999 (0.998-0.999)
ASQ[c]	0.178 (0.073-0.282)	0.970 (0.968-0.971)	0.022 (0.010-0.034)	0.997 (0.996-0.998)
Parallel				
ASQ or ML model high risk (≥95th percentile)	0.478 (0.344-0.612)	0.934 (0.928-0.940)	0.027 (0.018-0.036)	0.998 (0.997-0.999)
ASQ or ML model medium or high risk (≥75th percentile)	0.795 (0.671-0.920)	0.742 (0.716-0.767)	0.012 (0.009-0.014)	0.999 (0.998-0.999)
Serial				
ML model high risk AND ASQ	0.129 (0.036-0.221)	0.991 (0.989-0.993)	0.050 (0.014-0.086)	0.997 (0.996-0.998)
ML model medium or high risk AND ASQ	0.164 (0.059-0.268)	0.980 (0.978-0.981)	0.029 (0.012-0.045)	0.997 (0.996-0.998)

[a]PPV: positive predictive value.

[b]NPV: negative predictive value.

[c]ASQ: Ask Suicide-Screening Questions.

Performance of Combined Approaches

Parallel testing strategies, where patients were flagged if they were screened as positive *or* were classified as at risk by the ML model, substantially improved sensitivity compared to screening alone. When combining ASQ with the ML model at the 95th percentile, sensitivity increased to 0.478 (95% CI 0.344-0.612), with a specificity of 0.934 (95% CI 0.928-0.940), PPV of 0.027 (95% CI 0.018-0.036), and NPV of 0.998 (95% CI 0.997-0.999). This approach identified approximately half of patients who subsequently experienced an outcome without missing any patients who would have been caught through screening alone.

When combining the ASQ with the ML model at the 75th percentile in parallel, sensitivity further improved to 0.795 (95% CI 0.671-0.920), with a specificity of 0.742 (95% CI 0.716-0.767), PPV of 0.012 (95% CI 0.009-0.014), and NPV of 0.999 (95% CI 0.998-0.999). This more inclusive parallel approach identified 4 out of 5 patients who experienced an outcome but resulted in substantially more potential false positives.

Serial testing strategies, where patients were flagged only if *both* the ML model classified them as at risk *and* they were screened as positive, substantially improved PPV compared to other approaches. Using the 95th percentile threshold, the serial approach achieved a sensitivity of 0.129 (95% CI 0.036-0.221), specificity of 0.991 (95% CI 0.989-0.993), PPV of 0.050 (95% CI 0.014-0.086), and NPV of 0.997 (95% CI 0.996-0.998). This represents more than a 2-fold improvement in PPV compared to ASQ screening alone (0.050 vs 0.022), although at the cost of reduced sensitivity.

Using the 75th percentile threshold, the serial approach achieved a sensitivity of 0.164 (95% CI 0.059-0.268), specificity of 0.980 (95% CI 0.978-0.981), PPV of 0.029 (95% CI 0.012-0.045), and NPV of 0.997 (95% CI 0.996-0.998). This approach still provided a 32% increase in PPV compared to screening alone (0.029 vs 0.022) with slightly improved specificity. The CIs for the serial testing approaches in both the 95th and 75th percentile thresholds were notably wide, particularly for sensitivity (eg, 0.036-0.221 and 0.059-0.268, respectively). This reflects the small number of true positive cases identified through these highly specific strategies.

Sensitivity Analysis

Results from the sensitivity analysis including all ED visits (n=42,915) regardless of screening completion are presented in Multimedia Appendix 1. Performance patterns were similar to those of the primary analysis, although with slightly attenuated sensitivity across all approaches due to the larger denominator of missed screening opportunities.

Discussion

Principal Results

We aimed to evaluate strategies for combining ML models with suicide risk screening in an ED that serves American Indian communities. Three key findings emerged. First, when operationalizing the ML model at a lower threshold, we substantially increased sensitivity over screening alone, identifying 78.2% of patients who subsequently attempted or died by suicide compared to 17.8% of patients via screening alone. Second, parallel testing combining both approaches improved sensitivity to 0.795 while maintaining reasonable

XSL•FO
RenderX

specificity (0.742). Third, serial testing requiring both positive ML classification and positive screening achieved more than a 2-fold improvement in PPV (0.050 vs 0.022) but with reduced sensitivity (0.129 vs 0.178).

What constitutes adequate performance for suicide risk identification depends on the intended use, available resources, and tolerance for missed cases vs false alarms [12,26]. Our findings are consistent with those of recent meta-analyses showing pooled sensitivities of 45% to 82% and PPVs of 6% to 17% for ML algorithms for suicide risk identification [12]. The PPVs we observed (0.012-0.050) reflect the mathematical reality that low prevalence constrains PPV even when sensitivity and specificity are reasonable, a challenge inherent to predicting any rare outcome.

Comparison With Prior Work

Our results extend our prior work demonstrating that IHS-specific ML models (AUROC=0.83) outperform screening-based approaches [17] and parallel recent pediatric ED findings by Aseltine et al [20] showing that risk algorithms correctly identified 51% of subsequent attempts compared to 37% through screening. The consistency across populations suggests that ML's added value may be generalizable.

Critically, neither the ASQ nor other suicide screening instruments have been validated in American Indian and Alaska Native populations [27,28]. Our findings, showing a sensitivity of only 0.178, suggest that the tool may perform differently in this population or that implementation challenges limit effectiveness. The ASQ has shown much stronger predictive accuracy in non–American Indian and Alaska Native pediatric populations [24,29] but has not been validated against attempts or deaths in adult populations. Implementation challenges can include when and how questions are asked, cultural differences in communication patterns, or trust in health care systems [30,31]. While we wait on specific validation data, this issue reflects a broader pattern of concern: tools developed with majority populations are applied to high-burden minority populations without proper validation, risking perpetuation of health inequities [32].

Serial testing substantially improved PPV but may be infeasible given that universal screening is already implemented in many settings [33]. Ethically, withholding screening from ML-identified low-risk patients is problematic, particularly if they subsequently disclose suicidality. As Wilimitis et al [19] noted, approaches that rely primarily on ML to dictate workflows face significant ethical obstacles and logistical challenges when patients disclose risk factors. ML approaches also often depend on a patient having a recent history with the health system, creating a challenge to prediction when patients are new or have not been seen in recent years.

This creates a tension: the strategy with the best PPV is impractical in cases in which universal screening is already established or in settings when past history with the health system is not as well established. One pragmatic solution (Figure 1) is using ML to verify negative screens or flag high-risk patients when screening is missed. In our data, 37.3% (16,019/42,915) of ED visits lacked documented screening, and others were false negatives. For these cases, ML could serve as a safety net. For patients who are screened as negative but are classified as high risk by ML, clinicians could be alerted to reassess. This preserves universal screening while leveraging ML to reduce false negatives.

High false-positive rates raise concerns about resource strain and patient harm. However, 3 considerations provide context. First, many flagged patients who do not go on to attempt suicide may benefit from intervention as health care providers may use algorithm features to identify and intervene in other underlying conditions through connection with services or treatment. Second, false positives must be weighed against the consequences of false negatives in communities with starkly elevated suicide rates. Third, acceptability depends on the intervention burden; brief safety assessments and follow-up contacts may justify higher false-positive rates than intensive interventions such as involuntary hospitalization. Notably, widely accepted screening programs such as mammography have comparable PPVs (approximately 4% on the lower end) [34], although breast cancer screening has decades of mortality reduction evidence that suicide screening currently lacks [35].

Limitations

The findings are specific to a single tribal community, limiting generalizability. EHR-based outcome identification may miss undocumented attempts, potentially underestimating risk differentially across patient groups. We analyzed only visits with documented screening (26,896/42,915, 62.7% of total ED visits), which may not reflect real-world performance. We did not collect standardized ASQ implementation fidelity measures (mode, timing, staff training, privacy, or cultural adaptation), which may have influenced observed sensitivity. The ASQ currently lacks validation in American Indian and Alaska Native populations [27,28], and implementation challenges may limit its effectiveness. While the model was developed and tested in the same facility, it represents a retraining of an existing model from a completely different population [17], so while external validation for our metrics is warranted, the transportability of the underlying ML model has considerable strengths.

An additional important limitation is that the ML model depends on longitudinal EHR data to generate predictions [17]. Patients with limited or no prior health care encounters—such as those new to the system or seeking care for the first time—would have minimal data for the algorithm to analyze. While this is less common in our study setting given the reservation-based health care system with limited alternatives, it remains a significant consideration. This dependence on prior health care engagement highlights a fundamental value of screening tools: they assess current suicide risk through self-report regardless of prior history. This represents another reason why ML cannot replace screening, particularly for patients with sparse use patterns, and supports the rationale for combined rather than stand-alone approaches.

Conclusions

Unlike existing research that primarily evaluates ML models and screening tools in isolation, this study innovates by evaluating combined parallel and serial testing strategies within

XSL•FO
RenderX

a real-world American Indian ED setting. We demonstrate that, while ML models substantially outperform suicide risk screening alone, combining approaches yields distinct trade-offs. Serial testing improves PPV, but it remains practically and ethically infeasible in universal screening settings, either due to the ethical barriers of withholding treatment from individuals who disclose risk on the ASQ or because ML methods may require a history of several visits to adequately classify risk. Instead, parallel testing brings immediate value to the field by preserving universal screening and allowing ML to act as a clinical "safety net" to identify at-risk patients who are missed by current practices. Critically, these findings highlight that the most statistically accurate identification approach is not always practical. The optimal real-world integration of ML risk prediction requires careful consideration of setting-specific workflows to ensure that these approaches do not inadvertently perpetuate resource constraints or health inequities. Future work should focus on prospective evaluation of clinical outcomes, external validation in diverse settings, and continued community-engaged development to ensure that tools meet the needs of American Indian and Alaska Native communities facing the highest suicide burden.

Acknowledgments

The authors humbly acknowledge the White Mountain Apache Tribe and its community members for their support in undertaking this project. The authors continue to strive to help promote life for all in the community and elsewhere. The contents do not represent the views of the Indian Health Service or the US government. The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the Generative Artificial Intelligence Delegation Taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing. The GenAI tool used was Claude 4.5 (Anthropic). Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

The research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health under award number R01MH128518. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. PR is also supported by National Institute of Mental Health grant 5T32MH122357-04.

Disclaimer

The funders played no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, and approval of the manuscript; or decision to submit the manuscript for publication.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author with permission from the White Mountain Apache Tribe and Indian Health Service on reasonable request.

Authors' Contributions

Conceptualization: NAG, EEH
Data curation: RRD
Formal analysis: RRD
Funding acquisition: EEH
Methodology: EEH, RRD, RA, JLT
Writing—original draft: EEH, RRD
Writing—review and editing: NAG, VQS, RA, FLS, PR, JLT
The lead authors, NAG and EEH, and study statistician RRD had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Parallel and serial testing of machine learning model risk categories and suicide screening in the full sample (n=9224 patients; n=42,915 emergency department visits).
[DOCX File, 29 KB - jmir_v28i1e82669_app1.docx]

References

XSL•FO
RenderX

1. Stone DM, Cammack AL, Carbone EG. Notes from the field: differences in suicide rates, by race and ethnicity and age group - United States, 2018-2023. MMWR Morb Mortal Wkly Rep 2025 Sep 18;74(35):550-553. [doi: 10.15585/mmwr.mm7435a2] [Medline: 40966165]
2. Ahmedani BK, Simon GE, Stewart C, et al. Health care contacts in the year before suicide death. J Gen Intern Med 2014 Jun;29(6):870-877. [doi: 10.1007/s11606-014-2767-3] [Medline: 24567199]
3. Hogan MF, Grumet JG. Suicide prevention: an emerging priority for health care. Health Aff (Millwood) 2016 Jun 1;35(6):1084-1090. [doi: 10.1377/hlthaff.2015.1672] [Medline: 27269026]
4. R3 Report Issue 18: National Patient Safety Goal for Suicide Prevention. : The Joint Commission URL: https://www.jointcommission.org/en-us/standards/r3-report/r3-report-18 [accessed 2026-04-29]
5. Horowitz LM, Ryan PC, Wei AX, Boudreaux ED, Ackerman JP, Bridge JA. Screening and assessing suicide risk in medical settings: feasible strategies for early detection. Focus (Am Psychiatr Publ) 2023 Apr;21(2):145-151. [doi: 10.1176/appi.focus.20220086] [Medline: 37201144]
6. US Preventive Services Task Force, Barry MJ, Nicholson WK, et al. Screening for depression and suicide risk in adults: US Preventive Services Task Force recommendation statement. JAMA 2023 Jun 20;329(23):2057-2067. [doi: 10.1001/jama.2023.9297] [Medline: 37338872]
7. Matarazzo BB, Eagan A, Landes SJ, et al. The Veterans Health Administration REACH VET program: suicide predictive modeling in practice. Psychiatr Serv 2023 Feb 1;74(2):206-209. [doi: 10.1176/appi.ps.202100629] [Medline: 36039552]
8. Simon GE, Cruz M, Shortreed SM, et al. Stability of suicide risk prediction models during changes in health care delivery. Psychiatr Serv 2024 Feb 1;75(2):139-147. [doi: 10.1176/appi.ps.20230172] [Medline: 37587793]
9. Walsh CG, Ripperger MA, Novak L, et al. Randomized controlled comparative effectiveness trial of risk model-guided clinical decision support for suicide screening. medRxiv. 2024 Mar 18 p. 2024.03.14.24304318. [doi: 10.1101/2024.03.14.24304318] [Medline: 38562678]
10. Simon GE, Stewart CC, Richards JE, Ziebell R, Lapham GT, Hoopes AJ. Accuracy of self-report questionnaires and records-based risk scores to identify adolescents' risk for self-harm. Psychiatr Serv 2025 Jun 1;76(6):554-562. [doi: 10.1176/appi.ps.20240427] [Medline: 40103367]
11. Seyedsalehi A, Bailey J, Ogonah MG, Fanshawe TR, Fazel S. Prediction models for self-harm and suicide: a systematic review and critical appraisal. BMC Med 2025 Oct 9;23(1):549. [doi: 10.1186/s12916-025-04367-6] [Medline: 41068959]
12. Spittal MJ, Guo XA, Kang L, et al. Machine learning algorithms and their predictive accuracy for suicide and self-harm: systematic review and meta-analysis. PLoS Med 2025 Sep;22(9):e1004581. [doi: 10.1371/journal.pmed.1004581] [Medline: 40934153]
13. Nock MK, Millner AJ, Ross EL, et al. Prediction of suicide attempts using clinician assessment, patient self-report, and electronic health records. JAMA Netw Open 2022 Jan 4;5(1):e2144373. [doi: 10.1001/jamanetworkopen.2021.44373] [Medline: 35084483]
14. Randall JR, Sareen J, Chateau D, Bolton JM. Predicting future suicide: clinician opinion versus a standardized assessment tool. Suicide Life Threat Behav 2019 Aug;49(4):941-951. [doi: 10.1111/sltb.12481] [Medline: 29920749]
15. Bentley KH, Kennedy CJ, Khadse PN, et al. Clinician suicide risk assessment for prediction of suicide attempt in a large health care system. JAMA Psychiatry 2025 Jun 1;82(6):599-608. [doi: 10.1001/jamapsychiatry.2025.0325] [Medline: 40202745]
16. Haroz EE, Rebman P, Goklish N, et al. Performance of machine learning suicide risk models in an American Indian population. JAMA Netw Open 2024 Oct 1;7(10):e2439269. [doi: 10.1001/jamanetworkopen.2024.39269] [Medline: 39401036]
17. Adams R, Haroz EE, Rebman P, et al. Developing a suicide risk model for use in the Indian Health Service. Npj Ment Health Res 2024 Oct 16;3(1):47. [doi: 10.1038/s44184-024-00088-5] [Medline: 39414996]
18. Simon GE, Matarazzo BB, Walsh CG, et al. Reconciling statistical and clinicians' predictions of suicide risk. Psychiatr Serv 2021 May 1;72(5):555-562. [doi: 10.1176/appi.ps.202000214] [Medline: 33691491]
19. Wilimitis D, Turer RW, Ripperger M, et al. Integration of face-to-face screening with real-time machine learning to predict risk of suicide among adults. JAMA Netw Open 2022 May 2;5(5):e2212095. [doi: 10.1001/jamanetworkopen.2022.12095] [Medline: 35560048]
20. Aseltine RH Jr, Sacco SJ, Rogers S, Wang F, Schwartz H, Chen K. Screening and risk algorithms for detecting pediatric suicide risk in the emergency department. JAMA Netw Open 2025 Sep 2;8(9):e2533505. [doi: 10.1001/jamanetworkopen.2025.33505] [Medline: 40991285]
21. Horowitz LM, Bridge JA, Teach SJ, et al. Ask Suicide-Screening Questions (ASQ): a brief instrument for the pediatric emergency department. Arch Pediatr Adolesc Med 2012 Dec;166(12):1170-1176. [doi: 10.1001/archpediatrics.2012.1276] [Medline: 23027429]
22. Ballard ED, Cwik M, Van Eck K, et al. Identification of at-risk youth by suicide screening in a pediatric emergency department. Prev Sci 2017 Feb;18(2):174-182. [doi: 10.1007/s11121-016-0717-5] [Medline: 27678381]
23. Horowitz LM, Snyder DJ, Boudreaux ED, et al. Validation of the Ask Suicide-Screening Questions for adult medical inpatients: a brief tool for all ages. Psychosomatics 2020;61(6):713-722. [doi: 10.1016/j.psym.2020.04.008] [Medline: 32487323]

XSL•FO
RenderX

24. Brent DA, Horowitz LM, Grupp-Phelan J, et al. Prediction of suicide attempts and suicide-related events among adolescents seen in emergency departments. JAMA Netw Open 2023 Feb 1;6(2):e2255986. [doi: 10.1001/jamanetworkopen.2022.55986] [Medline: 36790810]
25. Cwik MF, Barlow A, Goklish N, et al. Community-based surveillance and case management for suicide prevention: an American Indian tribally initiated system. Am J Public Health 2014 Jun;104 Suppl 3(Suppl 3):e18-e23. [doi: 10.2105/AJPH.2014.301872] [Medline: 24754618]
26. Jung K, Kashyap S, Avati A, et al. A framework for making predictive models useful in practice. J Am Med Inform Assoc 2021 Jun 12;28(6):1149-1158. [doi: 10.1093/jamia/ocaa318] [Medline: 33355350]
27. Eagle Baez MS. Cultural adaptations to suicide screening among American Indians/Alaska Natives. Am J Neurol Res 2025;4(2):1-6. [doi: 10.52106/2837-7761.1023]
28. ASQ frequently asked questions Q16, “what cultural aspects were considered when developing the ASQ?”. Indian Health Service. URL: https://www.ihs.gov/suicideprevention/asqprogram/ [accessed 2026-02-23]
29. DeVylder JE, Ryan TC, Cwik M, et al. Assessment of selective and universal screening for suicide risk in a pediatric emergency department. Focus (Am Psychiatr Publ) 2023 Apr;21(2):217-224. [doi: 10.1176/appi.focus.23021007] [Medline: 37201139]
30. Office of the Surgeon General (US); Center for Mental Health Services (US); National Institute of Mental Health (US). Mental Health: Culture, Race, and Ethnicity: A Supplement to Mental Health: A Report of the Surgeon General: Substance Abuse and Mental Health Services Administration; 2001. URL: http://www.ncbi.nlm.nih.gov/books/NBK44243 [accessed 2026-02-23]
31. Gone JP, Trimble JE. American Indian and Alaska Native mental health: diverse perspectives on enduring disparities. Annu Rev Clin Psychol 2012;8:131-160. [doi: 10.1146/annurev-clinpsy-032511-143127] [Medline: 22149479]
32. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
33. Chitavi SO, Williams SC, Patrianakos J, et al. Evaluating the prevalence of suicide risk screening practices in accredited hospitals. Jt Comm J Qual Patient Saf 2025 May;51(5):342-349. [doi: 10.1016/j.jcjq.2025.01.010] [Medline: 40023708]
34. Rosenberg RD, Yankaskas BC, Abraham LA, et al. Performance benchmarks for screening mammography. Radiology 2006 Oct;241(1):55-66. [doi: 10.1148/radiol.2411051504] [Medline: 16990671]
35. Moss SM, Cuckle H, Evans A, et al. Effect of mammographic screening from age 40 years on breast cancer mortality at 10 years’ follow-up: a randomised controlled trial. Lancet 2006 Dec;368(9552):2053-2060. [doi: 10.1016/S0140-6736(06)69834-6] [Medline: 17161727]

Abbreviations

ASQ: Ask Suicide-Screening Questions
AUROC: area under the receiver operating characteristic curve
ED: emergency department
EHR: electronic health record
HIPAA: Health Insurance Portability and Accountability Act
IHS: Indian Health Service
MHRN: Mental Health Research Network
ML: machine learning
NPV: negative predictive value
PPV: positive predictive value
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by S Brini; submitted 19.Aug.2025; peer-reviewed by B Srinivasaiah, DB Patel, S Zeng, S Rybczynski, Y Shankheshwaria; revised version received 04.Mar.2026; accepted 05.Mar.2026; published 11.May.2026.

Please cite as:
Goklish NA, Haroz EE, Dayal RR, Sierra VQ, Adams R, Sinquah FL, Rebman P, Taylor JL
Combining Machine Learning Models and Screening to Enhance Suicide Risk Identification for American Indian Patients: Retrospective Cohort Study
J Med Internet Res 2026;28:e82669
URL: https://www.jmir.org/2026/1/e82669
doi:10.2196/82669

XSL•FO
RenderX

© Novalene Alsenay Goklish, Emily E Haroz, Rohan R Dayal, Valentín Q Sierra, Roy Adams, Francene Larzelere Sinquah, Paul Rebman, Jacob L Taylor. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Longitudinal Effects of a Smartphone Game (Tumaini) for HIV Prevention Among Kenyan Adolescents: 45-Month Trajectories of Condom Use–Related Proximal Outcomes From a Randomized Controlled Trial

Kate Winskell[1], PhD; Gaëlle Sabben[1], MPH; Haowen Qin[2], MSPH; Calvin Mbeda[3], BSc; Sophie Goldenberg[4], MPH; Ken Ondeng'e[3], MPH; Richard Ndivo[3], BSc; Judith Arego[3], BA; Xinwei He[2], MS; Robert A Bednarczyk[1], PhD; Kelli Komro[5], PhD, MPH; Robert H Lyles[2], PhD; Victor Mudhune[3], PhD

[1]Hubert Department of Global Health, Rollins School of Public Health at Emory University, Atlanta, GA, United States
[2]Department of Biostatistics and Bioinformatics, Rollins School of Public Health at Emory University, Atlanta, GA, United States
[3]HIV Research Division, Centre for Global Health Research, Kenya Medical Research Institute, Kisumu, Kenya
[4]Department of Epidemiology, Rollins School of Public Health at Emory University, Atlanta, GA, United States
[5]Department of Behavioral, Social, and Health Education Sciences, Rollins School of Public Health at Emory University, Atlanta, GA, United States

Corresponding Author:
Kate Winskell, PhD
Hubert Department of Global Health
Rollins School of Public Health at Emory University
1518 Clifton Road, NE
Atlanta, GA, 30322
United States
Phone: 1 404 727 5286
Email: swinske@emory.edu

Abstract

Background: African adolescents and young adults account for a disproportionate number of new HIV infections. There is an urgent need to identify scalable and cost-effective behavioral HIV prevention strategies for this population. Using a condom at first sex is associated with a higher likelihood of consistent use later. Tumaini (“Hope for the Future” in Swahili; Emory University) is a choose-your-own-adventure smartphone game that has been shown to reduce the risk of unprotected first sex by end line in a 45-month randomized controlled trial in western Kenya.

Objective: This study aimed to assess the impact of Tumaini on proximal outcomes related to condom use at first sex (specifically, behavioral intentions, self-efficacy, attitudes, and knowledge) longitudinally across mid-adolescence in the above trial.

Methods: Adolescent participants (n=996, mean baseline age 14, SD 0.56 years) were randomized 1:1 to receive either a smartphone loaded with Tumaini or an attention-control math game for 5 to 7 weeks at 3 time points (mean age 14.0, SD 0.56; 15.3, SD 0.55; and 16.0, SD 0.56 years, respectively). They completed a behavioral survey at 13 time points, through mean age 17.7 (SD 0.56) years. Using generalized estimating equations and controlling for age at baseline, we modeled mean scores (overall and stratified by gender) on a range of condom-related survey items over time to assess mean differences at specific time points. We applied appropriate Bonferroni corrections to inferences about cross-arm differences in mean changes relative to baseline at 4 time points (after each intervention period and at end line; α=.05/4) and within-arm mean changes relative to baseline at each of the 12 post-baseline time points (α=.05/12). Analyses were conducted as intent-to-treat.

Results: At end line, 97.8% (n=974) of the sample had been retained. Participants in both arms dedicated a mean total of >30 hours to their assigned game. There was significant improvement across all condom-related proximal outcomes in the intervention arm relative to the control arm immediately after initial intervention exposure. For almost all outcomes, a significant cross-arm difference was also present at end line and for most outcomes at the 2 intervening comparison time points. Some outcomes saw stronger intervention effects on female participants (eg, self-efficacy to refuse unprotected sex) or male participants (eg, knowledge that condoms are an effective way to prevent HIV). In each arm, intention to use a condom at first sex was consistently higher among male participants; however, female intervention-arm scores overtook male control-arm scores following initial intervention exposure.

XSL•FO
RenderX

Conclusions: Tumaini significantly improved theory-based proximal outcomes related to condom use, with effects sustained 45 months post initial exposure and 16 months post most recent exposure. Adolescents benefited from even short-term exposure, though repeated exposure generally sustained and reinforced intervention effects. As access to smartphones increases, Tumaini has potential for high scalability and impact on condom-related outcomes.

Trial Registration: ClinicalTrials.gov NCT04437667; https://clinicaltrials.gov/study/NCT04437667

International Registered Report Identifier (IRRID): RR2-10.2196/35117

(J Med Internet Res 2026;28:e83982) doi:10.2196/83982

KEYWORDS

adolescents; HIV; prevention; sexual health; sub-Saharan Africa; Kenya; behavioral intervention; serious game; game for health; interactive narrative; smartphone; randomized controlled trial; mHealth; condom; behavioral intention; self-efficacy; attitude; knowledge

Introduction

Worldwide, a third of all new adult (aged 15-49 years) HIV infections occur in eastern and southern Africa, and over a third of these infections occur in young people aged 15-24 years [1]. HIV-related morbidity and mortality in this age group are disproportionately high. In addition, as the largest-ever generation of young Africans is now entering adolescence, it is feared that a resurgence in the global epidemic could ensue unless more is done to prevent HIV in this demographic [2]. There is an urgent need to develop and rigorously test scalable and cost-effective behavioral interventions to prevent HIV among African adolescents and young adults, particularly adolescent girls and young women who are disproportionately affected [1].

Reaching adolescents with information and skills before sexual debut may help establish long-term patterns of safer sexual behavior [3]. For example, studies have shown that those who use condoms at first sex are more likely to use them consistently in the future [4]. There is a need for a stronger evidence base on the longitudinal effects of early and ongoing intervention, as few studies evaluating any HIV prevention interventions for adolescents and young adults include long-term follow-up [5].

Scaling up in-person, group-based HIV prevention interventions is challenging [6]. Digital behavioral interventions are likely to be easier to deliver at scale with consistent quality than those delivered in person. As smartphone penetration grows across sub-Saharan Africa, especially among young people, digital HIV prevention interventions are becoming increasingly viable. Across the continent, 81% of mobile phone connections will be smartphones by 2030 [7].

Electronic games are a promising subset of digital behavioral interventions [8]. However, few have been rigorously designed and evaluated [9]. Specifically, games leveraging interactive narrative can provide a level of experiential learning unparalleled by most other interventions by allowing players to experience real agency in a virtual and safe environment.

Tumaini (“hope for the future” in Swahili) is a “choose-your-own-adventure” game app for inexpensive Android smartphones. It was developed by specialists in adolescent sexual health (KW, GS, with others) in collaboration with a commercial game developer with input from Kenyan adolescents and their parents. The game’s theoretical framework [10] draws on Social Cognitive Theory [11], existing evidence-based interventions for youth HIV prevention, and literature on games for health and Entertainment-Education [12,13]. It is grounded in research on HIV-themed narratives written by thousands of young Africans [14-16].

Tumaini uses interactive narrative to promote observational learning, cognitive and behavioral rehearsal, and problem-solving [10]. It provides over 12 hours of discrete gameplay and comprises (1) an engaging role-playing narrative in which players make choices for 6 characters (3 female, 3 male) as they proceed through adolescence and observe short- and long-term consequences in the characters’ lives; (2) mini-games to reinforce knowledge and skills; and (3) My Story, in which players set life goals, reflect on how the game relates to their lives, customize a personal avatar, and collect rewards. The narrative uses branching logic for 40 possible long-term outcomes (“epilogues”) across the 6 characters and is designed to be replayed, allowing players to experience different outcomes based on their in-game decisions. Players must assume the role of characters different from themselves in terms of sex and serostatus, thereby challenging stigma and harmful gender norms.

In a randomized feasibility study with 60 adolescents in Kenya, the intervention arm showed significant gains in sexual health-related knowledge and self-efficacy and behavioral intention for risk avoidance strategies and sexual risk communication when compared with the control arm 6 weeks postintervention [17]. Intervention-arm participants spent on average over 50% longer playing the game than instructed, and quantitative and qualitative data on user engagement and on game appeal, relevance, and acceptability among adolescents and parents were extremely positive [18].

A randomized controlled trial with a larger sample size (n=996) and longer follow-up (45 months) was undertaken with the objective of determining the intervention’s effects on (1) age at sexual debut and condom use at first sex and (2) related proximal outcomes (behavioral intentions, self-efficacy, attitudes, and knowledge). Tumaini demonstrated significant impact on the primary trial outcome (high-risk sexual debut vs low-risk or no sexual debut), attributable to increased condom use at first sex, at end line in intent-to-treat analyses [19].

XSL•FO
RenderX

Significant effects were also observed across a wide range of other outcomes at the end line.

Participants completed a behavioral survey at 13 time points over the 45-month duration of the trial, with a mean participant age of 14.0 to 17.7 (SD 0.56) years. The survey included items measuring a range of theory-based proximal outcomes related to condom use. These data and their longitudinal trajectories are the focus of the analyses presented in this manuscript. As the primary behavioral outcomes are related to sexual debut and developmentally influenced, hence statistically underpowered prior to end line, measures of proximal outcomes are of particular importance to understanding the effects of the intervention between trial baseline and end line. The purpose of this manuscript is to leverage these longitudinal data to evaluate Tumaini's impact on proximal outcomes related to condom use across 45 months of mid-adolescence, a key developmental phase, while assessing differential effects by gender.

Methods

Study Setting

Kenya has the 7th largest HIV burden globally and an HIV epidemic that is generalized throughout the population. Kisumu County has Kenya's second-highest prevalence (15.5%) and annual number of new infections [20]. The study took place in Kisumu Town, Kenya's third largest city, where youth aged 15-24 years comprise 48% of new HIV infections [21].

Study Design

This was a 45-month, 2-arm, individually randomized controlled trial with 996 participants, with a mean age of 14.0 (SD 0.6) years at baseline [22]. The study took place from October 2020 (start of enrollment) to August 2024 (last quantitative data collection). Participants were randomized 1:1 to play either Tumaini or a commercially available attention-control math game, Brainilis (Appgeneration – Software Technologies, Lda). They were provided with a low-cost Android smartphone loaded with their assigned game at mean participant age 14.0 (SD 0.56), 15.3 (SD 0.55), and 16.0 (SD 0.56) years, that is, between baseline data collection (T1) and second data collection time point (T2), 6th data collection time point (T6) and 7th data collection time point (T7), and 8th data collection time point (T8) and 9th data collection time point (T9). Gameplay periods lasted 5-7 weeks and took place during school holidays; they were unevenly spaced due to COVID-19–related disruptions to the school calendar. Participants completed behavioral surveys at 13 time points (baseline plus 12 follow-up surveys).

Recruitment and Randomization

Participant recruitment was conducted through a combination of school- and community-based strategies in urban and periurban Kisumu. Eligibility criteria for participation were being aged 12-14 years at the time of enrollment, being a resident in Kisumu Town, having basic English literacy (Grade 3-4 on the Flesch-Kincaid Reading Scale), and having no prior exposure to Tumaini-related activities.

Recruitment slots were used to balance participant gender, age, and residence (in the 16 wards that comprise Kisumu East, West, and Central). Randomization to the study arms was conducted by a blinded member of the data analysis team with no contact with participants. The success of randomization was confirmed through calculation and cross-arm comparison of descriptive statistics for adolescent demographics.

Due to the nature of the intervention, participants were not blinded to group assignment, which was revealed following completion of all baseline activities. Data analysts remained blinded to study arm assignment until completion of all primary outcome data analysis.

Data Collection and Analysis

Behavioral surveys were self-administered at 13 time points (T1-T13; 13th data collection time point [endline]), via tablet-based Open Data Kit. Participants completed surveys with headphones in a private space. Access to audio of all questions was provided to maximize understandability. Survey items were designed to evaluate the impact of exposure to the intervention on a range of primary and secondary outcomes. The study's primary outcomes assessed experience of sexual debut and condom use at first sex separately and also combined them in a composite binary variable of high-risk sexual debut (no condom use or "I don't know") vs low-risk sexual debut (condom use) or no risk (not yet initiated sex). Sex was defined as penile-vaginal ("when a boy or man puts his penis inside a girl's or woman's vagina"). It was specified that condoms were understood to be male condoms, defined as "a thin rubber covering worn by a man on his erect penis when he has sex," due to the low availability of female condoms (or "internal condoms"). Endline outcome analyses are presented elsewhere [19].

Survey items also assessed theory-based and intervention-targeted proximal influences, namely knowledge, attitudes, self-efficacy, and behavioral intentions, of these primary outcomes. The behavioral survey instrument was adapted from the version used in the Tumaini feasibility study [17]. Individual measures were drawn from existing instruments, with priority given to those previously used with sub-Saharan African youth populations [23-25]. The chosen items were adapted for age, linguistic, and cultural appropriateness and consistency of formatting, and supplemented with additional measures where necessary. In the interest of age appropriateness, some questions were framed as hypothetical risk scenarios ("*Imagine that…*").

Here, we focus on the longitudinal trajectories of the proximal outcomes related to condom-use behavior. These outcomes, their exact wording, and response options (binary and 3 or 4-point Likert scales) are presented in Table 1 (additional items are detailed in Table S1 in Multimedia Appendix 1). Each item was scored on a 0-1 scale, with higher values indicating more desirable or, in the case of knowledge, correct answers. This standardization of scores facilitated comparison across items. Mean scores were calculated for individual items at each time point. Only participants who had responded to the question at that time point were included in mean score calculations; missing values were not imputed. Only those participants

XSL•FO
RenderX

reporting never having had sex were asked about their intention to use a condom at first sex.

Table 1. Condom-related proximal outcomes, behavioral survey question wording, and response options for Tumaini efficacy trial.

Condom-related proximal outcome	Question wording	Response options
Behavioral intention		
Intention to use a condom at first sex	"When you have sex for the first time, do you plan to use a condom?"	Definitely yes, maybe yes, maybe no, definitely no
Intention to talk to a partner about HIV prevention	"Imagine that either in the future or now you are thinking about having sex with someone and you are not ready to become a parent. […] Would you talk with him/her about preventing HIV?"	Definitely yes, maybe yes, maybe no, definitely no
Self-efficacy		
Self-efficacy to use a condom correctly	"I am sure that I could use a condom correctly."	I strongly agree, I agree, I disagree, I strongly agree
Self-efficacy to refuse unprotected sex	"Imagine that, either in the future or now, you are in a relationship. He/ she says she wants to have sex with you, but she does not want to use a condom. […] If I did not want to have sex with him/her without a condom, I am sure that I could say no firmly."	I strongly agree, I agree, I disagree, I strongly agree
Attitudes		
Endorsement that condom use is a sign of respect	"Using a condom is a sign that you respect your partner"	I strongly agree, I agree, I disagree, I strongly agree
Endorsement of gender parity in condom initiation	"A woman can suggest using condoms just like a man can"	I strongly agree, I agree, I disagree, I strongly agree
Knowledge		
Knowledge of how to use a condom correctly	"Do you know how to use a condom correctly?"	Yes, no
Knowledge that condoms are effective for HIV prevention	"Are condoms an effective way to prevent HIV?"	Yes, no, I don't know

We used generalized estimating equations [26] to model these proximal outcome variables measured repeatedly over time and to facilitate statistical inferences about cross-arm (intervention vs control) mean differences at specific time points. Specifically, the generalized estimating equations model controlled for participants' baseline age (centered at the average value; due to the success of randomization, we did not control for any other covariates) and included indicator variables for arm (intervention as reference) and time (T1–T13, with T1 as reference), together with the 12 group-by-time interaction terms. This approach accounts for the correlated nature of the data, using an exchangeable working correlation structure that was selected via initial comparisons of alternatives based on the quasi-likelihood information criterion [27]. The granular nature of the model provides a close fit to the sample means of the outcome for each arm at each time, while facilitating direct inferences about the change in the mean outcome relative to baseline for the intervention arm at each time point (T2–T13) and the cross-arm differences in these mean changes. Of primary interest in these analyses are Bonferroni-corrected (significance level .05/4) inferences about the cross-arm differences in mean changes relative to baseline at 4 specific time points (T2, T7, T9, and T13). The inferences at T2 and T13 compare the average initial and late (end line) change in the mean response relative to baseline for the intervention and control arms. The inferences at T7 and T9 are designed to investigate potential enhancements in the cross-arm differences in change relative to baseline at the 2 time points following reintroduction of the intervention. Secondary to these cross-arm comparisons, we also report Bonferroni-corrected (significance level .05/12) inferences to assess mean within-arm changes relative to baseline at each of the 12 subsequent time points (T2–T13). These analyses were conducted overall and separately for female and male participants.

All analyses were conducted as intent-to-treat and were completed using SAS (version 9.4; SAS Institute Inc).

Ethical Considerations

The Institutional Review Board of Emory University (IRB00108404) and the Scientific and Ethics Review Unit of the Kenya Medical Research Institute (KEMRI; SERU: KEMRI/SERU/CGHR/11/3812) approved this study. Written informed consent was secured from one parent or guardian, and the participant provided their written assent prior to study baseline. To ensure participant privacy, datasets were deidentified; all participants were identified by a randomly assigned study identifier, which was the only identifier used in analysis datasets and to link data across time points and sources. Data were stored and transferred on encrypted devices. For each study visit, adolescent participants received KShs 500 (~US $5) as compensation for their time.

XSL•FO
RenderX

Results

Overview

A participant flowchart is shown in Figure 1, and baseline characteristics of the sample are presented in Table 2. Participants represented a range of socioeconomic and food security profiles. At end line, 97.8% of the sample (n=974) had been retained. There were no adverse events associated with study participation.

Figure 1. Trial flow diagram of study activities and participant totals for 45-month efficacy trial of a smartphone game (Tumaini) to increase age and condom use at first sex among young Kenyans.

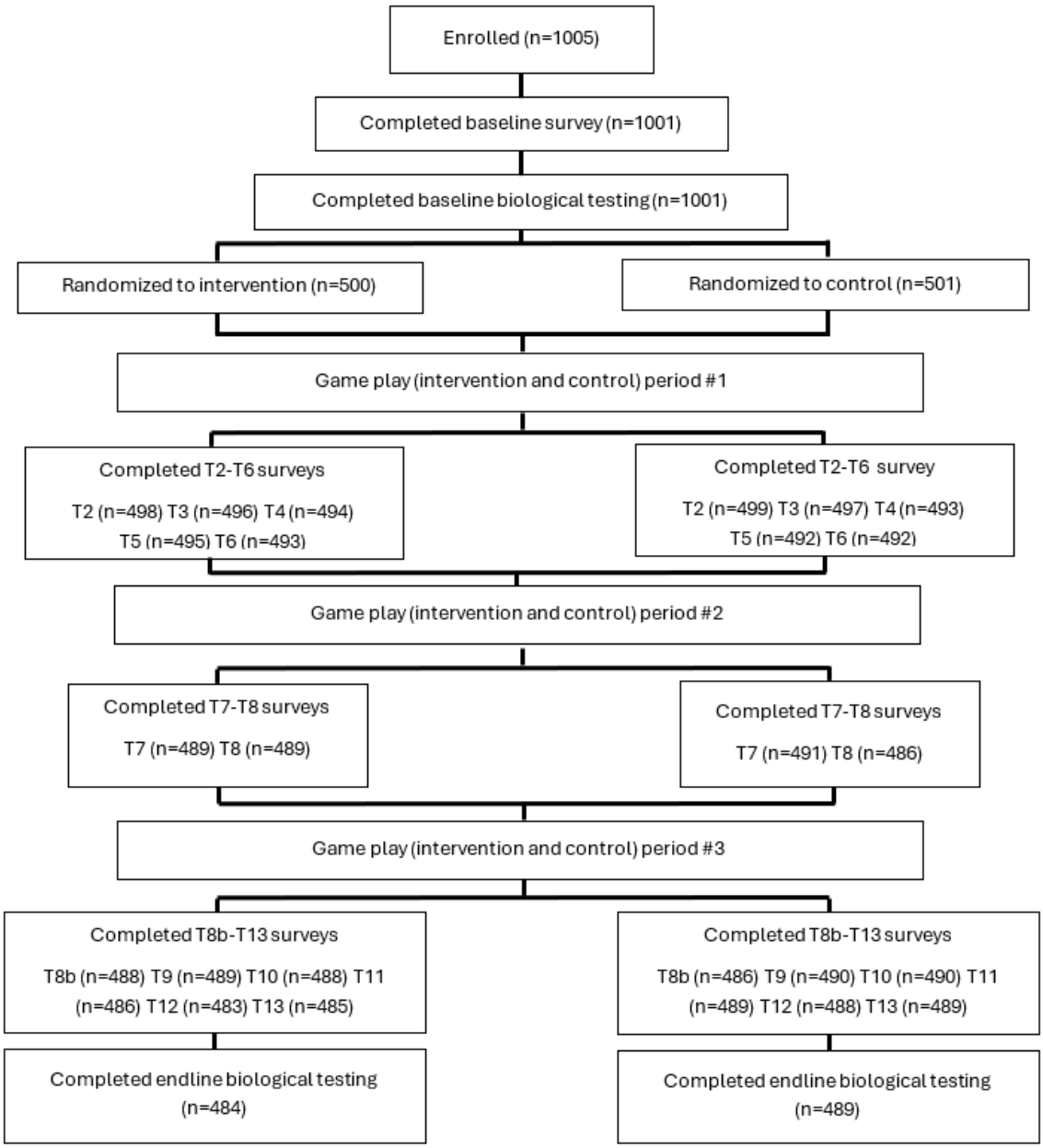

XSL•FO
RenderX

Table 2. Baseline characteristics of the intent-to-treat population (n=996) in the Tumaini efficacy trial.

Demographic characteristics	Intervention (n=499)	Control (n=497)
Sex, n (%)		
Female	249 (50)	250 (50)
Male	250 (50)	247 (50)
Age (years), mean (SD)	13.97 (0.57)	13.99 (0.54)
Enrolled in school, n (%)	499 (100)	493 (99)
Socioeconomic status, n (%)		
Lower socioeconomic status	241 (49)	241 (49)
Higher socioeconomic status	254 (51)	251 (51)
Food insecurity, n (%)		
None	270 (54)	260 (52)
Low	138 (28)	149 (30)
Medium	73 (15)	73 (15)
High	18 (4)	15 (3)
Living with HIV, n (%)	8 (2)	8 (2)
Experienced sexual debut, n (%)	76 (15)	76 (15)

Self-reported engagement in the intervention and control games was high at the end of each of the 3 intervention periods, and there was 100% exposure of each arm to their allocated game by T2. Participants in each arm dedicated comparable amounts of time to their assigned game across the 3 intervention periods, totaling 33.56 hours (SD 17.75) to Tumaini and 32.37 hours (SD 18.41) to the math game, indicating that the latter functioned as a highly effective attention-control condition. Fifty-eight percent of the control arm reported having played some of Tumaini by T9 (mean hours for those exposed: 19.77, SD 15.03) [19], with 35% exposed at T2 and 52% at T7.

We provide an overview of cross-arm differences in change since baseline at 4 time points (Table 3). Mean scores for each outcome at each time point and associated *P* values for within-arm change since baseline are presented in Tables S2-S9 in Multimedia Appendix 1.

XSL•FO
RenderX

Table 3. Cross-arm difference in mean change in proximal outcomes relative to baseline and associated *P* values at 4 time points in Tumaini efficacy trial.

Time point	T2[a]		T7[b]		T9[c]		T13[d]	
	Difference in change in mean score	*P* value	Difference in change in mean score	*P* value	Difference in change in mean score	*P* value	Difference in change in mean score	*P* value
Mean age (years)	14.1 (0.56)	_[e]	15.6 (0.55)	–	16.4 (0.56)	–	17.7 (0.56)	–
Behavioral intentions								
Intention to use a condom at first sex								
All	0.118	.005[f]	0.200	.008[f]	0.201	.01	0.164	.007[f]
Female	0.083	.02	0.088	.03	0.099	.02	0.103	.02
Male	0.055	.11	0.059	.11	0.045	.25	0.059	.14
Intention to talk to a partner about HIV prevention								
All	0.107	<.001[f]	0.085	<.001[f]	0.081	<.001[f]	0.076	<.001[f]
Female	0.115	<.001[f]	0.047	.11	0.048	.10	0.057	.05
Male	0.099	<.001[f]	0.123	<.001[f]	0.115	<.001[f]	0.095	.002[f]
Self-efficacy								
Self-efficacy to use a condom correctly								
All	0.215	<.001[f]	0.283	<.001[f]	0.273	<.001[f]	0.254	<.001[f]
Female	0.216	<.001[f]	0.278	<.001[f]	0.263	<.001[f]	0.283	<.001[f]
Male	0.213	<.001[f]	0.287	<.001[f]	0.282	<.001[f]	0.224	<.001[f]
Self-efficacy to refuse unprotected sex								
All	0.072	.006[f]	0.039	.13	0.042	.11	–0.037	.41
Female	0.107	.004[f]	0.028	.43	0.063	.09	0.040	.28
Male	0.037	.32	0.051	.17	0.021	.57	0.002	.96
Attitudes								
Endorsement that condom use is a sign of respect								
All	0.218	<.001[f]	0.201	<.001[f]	0.201	<.001[f]	0.164	<.001[f]
Female	0.234	<.001[f]	0.187	<.001[f]	0.201	<.001[f]	0.191	<.001[f]
Male	0.203	<.001[f]	0.214	<.001[f]	0.202	<.001[f]	0.138	<.001[f]
Endorsement of gender parity in condom initiation								
All	0.151	<.001[f]	0.205	<.001[f]	0.221	<.001[f]	0.192	<.001[f]
Female	0.146	<.001[f]	0.220	<.001[f]	0.218	<.001[f]	0.203	<.001[f]
Male	0.156	<.001[f]	0.189	<.001[f]	0.224	<.001[f]	0.181	<.001[f]
Knowledge								
Knowledge of how to use a condom correctly								
All	0.372	<.001[f]	0.429	<.001[f]	0.464	<.001[f]	0.458	<.001[f]
Female	0.376	<.001[f]	0.440	<.001[f]	0.462	<.001[f]	0.462	<.001[f]
Male	0.367	<.001[f]	0.417	<.001[f]	0.465	<.001[f]	0.453	<.001[f]
Knowledge that condoms are effective for HIV prevention								
All	0.096	.001[f]	0.093	.01[f]	0.085	.02	0.110	.003[f]

XSL•FO
RenderX

Time point	T2[a]		T7[b]		T9[c]		T13[d]	
	Difference in change in mean score	*P* value	Difference in change in mean score	*P* value	Difference in change in mean score	*P* value	Difference in change in mean score	*P* value
Female	0.116	.006[f]	0.123	.01	0.097	.05	0.158	.002[f]
Male	0.076	.07	0.059	.25	0.072	.17	0.062	.25

[a]T2: second data collection time point.

[b]T7: 7th data collection time point.

[c]T9: 9th data collection time point.

[d]T13: 13th data collection time point (end line).

[e]Not applicable.

[f]Indicates significance after Bonferroni correction (significance level .05/4, ie, .0125)

For each outcome, we present longitudinal plotlines below for all participants and stratified by gender. In the plots, solid lines track estimated mean responses (marked with triangles) for the intervention group over time; dashed lines (with circles) track estimated means for the control group. Solid markers (triangles for the intervention arm, circles for the control arm) indicate that the within-arm difference in mean response relative to baseline at that time point is significant after Bonferroni correction (ie, at α=.05/12 level); open symbols indicate lack of significance relative to baseline. Inferences addressing the 4 primary cross-arm comparisons of mean change relative to baseline are indicated with *P* values at the top of the plot. Asterisks after the *P* value indicate significance after Bonferroni correction (ie, at α=.05/4 level).

Behavioral Intentions

Intention to Use a Condom at First Sex

Figure 2 summarizes the results of the longitudinal analysis of the condom use at first sex behavioral intention outcome.

Cross-arm difference: when all participants were included in the analysis, there was significantly greater improvement in the intervention arm than the control arm immediately after initial intervention exposure (*P*=.005), following the second intervention period (ie, at T7; *P*=.008), and at end line (*P*=.007). Significance was marginal after Bonferroni correction following the third intervention period at T9 (*P*=.01) and not present when participants were stratified by gender, suggesting insufficient power once the sample was halved.

Within-arm change: solid triangles at T2-T13 for all participants indicate a Bonferroni-corrected statistically significant difference in the mean relative to baseline for the intervention arm at every subsequent time point. In addition to the steep uptick in mean score following initial intervention exposure, the intervention arm showed an uptick following the second and third intervention periods. In contrast, significant within-arm differences from baseline were present only occasionally in the control arm, with mean scores at end line closely resembling those at baseline.

Difference by gender: intention to use a condom at first sex was higher from baseline among male than female participants in both arms. Female intervention-arm participants saw a steeper uptick at T2 and significant improvement relative to baseline for all subsequent time points, such that it overtook the male control arm. After an initial upward trajectory, the control-arm plot trended downwards for female participants from T5 (age 14.8, SD 0.55).

Figure 2. Longitudinal trajectory of intention to use a condom at first sex at 13 timepoints among all, female, and male participants not reporting sexual debut.

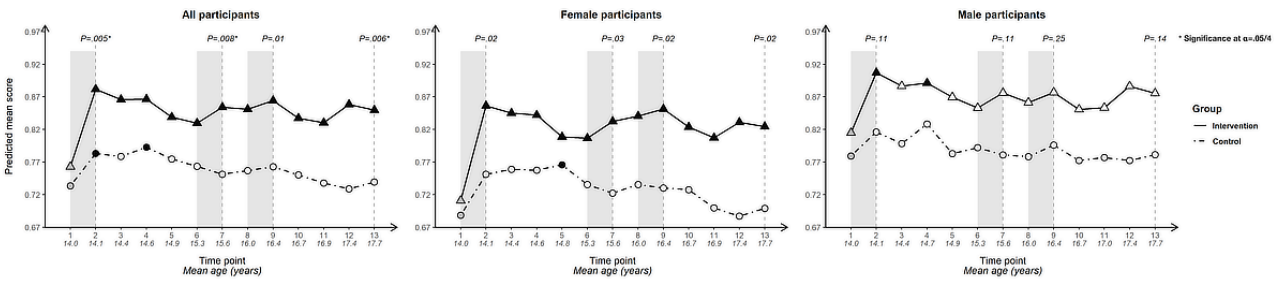

Intention to Talk to a Partner About HIV Prevention

Cross-arm difference: significant cross-arm differences in increased intention to talk to a partner about HIV prevention (Figure 3) were present at all 4 time points for all participants and for male participants.

XSL•FO
RenderX

Figure 3. Longitudinal trajectory of intention to talk to a partner about HIV prevention at 13 time points among all, female, and male participants.

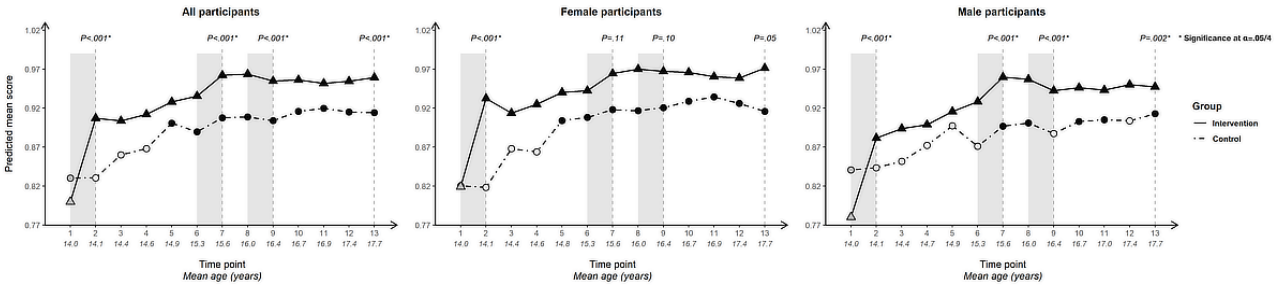

Despite an uptick in the intervention arm during the second intervention period, cross-arm differences were not significant at T7, T9, and end line for female participants. The plots suggest that this may be due to a combination of ceiling effects (mean scores in the intervention arm were extremely high from T7 [age 15.6, SD 0.55] onwards, limiting scope for improvement) and developmental gains in the control arm.

Within-arm change: for intervention-arm participants, within-arm change since baseline was significant for all time points for all female and male participants. For control-arm participants, difference relative to baseline mean became significant from T5 (age 14.9, SD 0.55) onwards for all and female participants. Male control-arm participants saw sporadic significant change since baseline from T7 (age 15.6, SD 0.56) onwards. Unlike intention to use a condom at first sex, mean scores followed an upward trajectory over time for both arms, suggesting a developmental effect.

A related outcome, self-efficacy to talk to a partner about HIV prevention (Figure S1 in Multimedia Appendix 2 and Table S10 in Multimedia Appendix 1), showed a similar trajectory to intention to talk to a partner about HIV prevention, with control-arm change since baseline for male participants achieving significance from mean age 14.9 (SD 0.56) years and end line values approximating those of the intervention arm.

Self-Efficacy

Self-Efficacy to Use a Condom Correctly

Cross-arm difference: baseline mean scores on self-efficacy to use a condom correctly (Figure 4) were lower and intervention-arm gains greater than for all other condom-related behavioral determinants except knowledge of correct condom use (which shows a similar trajectory, particularly in the intervention arm). Cross-arm differences in change since baseline were highly significant ($P<.001$) at all 4 time points for all participants and when stratified by gender.

Within-arm change from baseline was significant for intervention-arm participants at all subsequent time points, following a sharp uptick in mean score after the first intervention period and smaller upticks after the other intervention periods. For control-arm participants, a significant difference relative to baseline only occurred at one time point for each group: negatively at T7 (age 15.6, SD 0.55) for all and for female participants, and positively at end line for male participants.

Difference by gender: the overall upward trajectory among intervention-arm participants from T6 (age 15.4, SD 0.55) was steeper for male than for female participants. A similar upward trajectory was observed for male participants in the control arm, though at a much lower level, suggesting a developmental influence. The trajectory for female control-arm participants, by contrast, trended downwards.

Figure 4. Longitudinal trajectory of self-efficacy to use a condom correctly at 13 time points among all, female, and male participants.

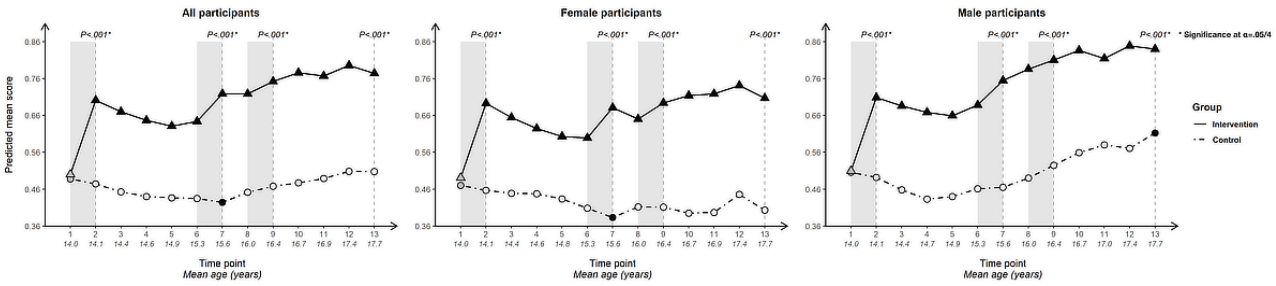

Self-Efficacy to Refuse Unprotected Sex

Cross-arm difference in increased self-efficacy to refuse unprotected sex (Figure 5) was significant following initial exposure to the intervention for all participants ($P=.006$) and for female participants ($P=.004$), but not for male participants ($P=.32$).

XSL•FO
RenderX

Figure 5. Longitudinal trajectory of self-efficacy to refuse unprotected sex at 13 time points among all, female, and male participants.

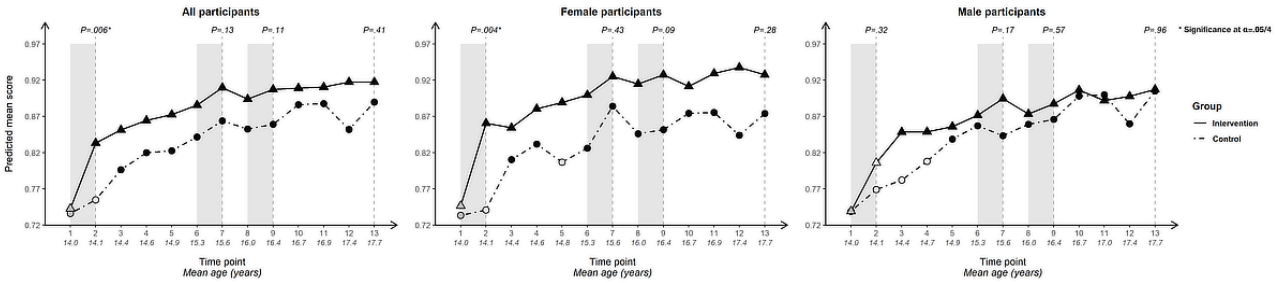

It was not significant at subsequent time points for any of the population groupings, despite upticks in the intervention arm following the second and third intervention periods. The upward trend in the intervention arm was mirrored by a similar trajectory, albeit at a generally lower level, in the control arm, suggesting a positive developmental influence for this outcome.

For the intervention arm, within-arm change since baseline was significant across all time points for all participants and when stratified by gender, except at T2 for male participants. The within-arm difference was significant for all control-arm participants from T3 (age 14.4, SD 0.56) onwards and for male participants from T5 onwards. For female participants, it was generally significant from T3 onwards.

Difference by gender: for male participants, the intervention-arm plotline was crossed by the control-arm plotline at T11 (age 17.0, SD 0.56; across all determinants, this was the only instance of this happening). End line values for both arms were nearly identical among male participants.

Attitudes: Endorsement of "Using a Condom Is a Sign You Respect Your Partner" and "A Woman Can Suggest Using Condoms Just Like a Man Can"

The survey included 2 condom-related attitude items, asking participants to indicate their degree of agreement with "using a condom is a sign that you respect your partner" (Figure 6) and "a woman can suggest using a condom just like a man can" (Figure 7).

Cross-arm difference in change since baseline was highly significant for endorsement of both condom-related attitude questions at each time point and for all population groupings.

Significant within-arm change since baseline was present in the intervention arm at every time point for both questions and for all population groupings. The intervention-arm uptick following initial exposure to Tumaini was more pronounced for the first, gender-neutral question, with moderate increases across subsequent time points. For the second question, which challenged traditional cultural norms around women initiating condom use, the longitudinal trajectory for the intervention arm showed a steady rise from T2 that was reflected at lower mean scores but a similar slope in the control arm, suggesting a more pronounced intersection of intervention and developmental influences. Both questions saw an uptick in intervention-arm scores following the second and third intervention periods.

Figure 6. Longitudinal trajectory of endorsement of "Using a condom is a sign that you respect your partner" at 13 time points for all, female, and male participants.

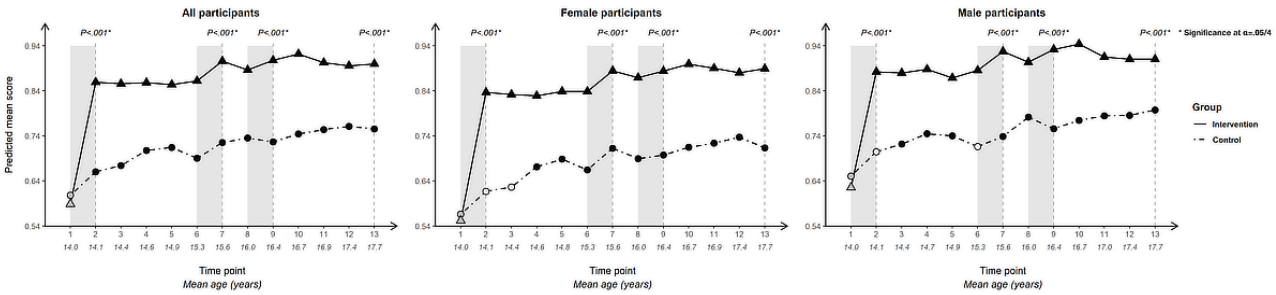

XSL•FO
RenderX

Figure 7. Longitudinal trajectory of endorsement of "A woman can suggest using condoms just like a man can" at 13 time points for all, female, and male participants.

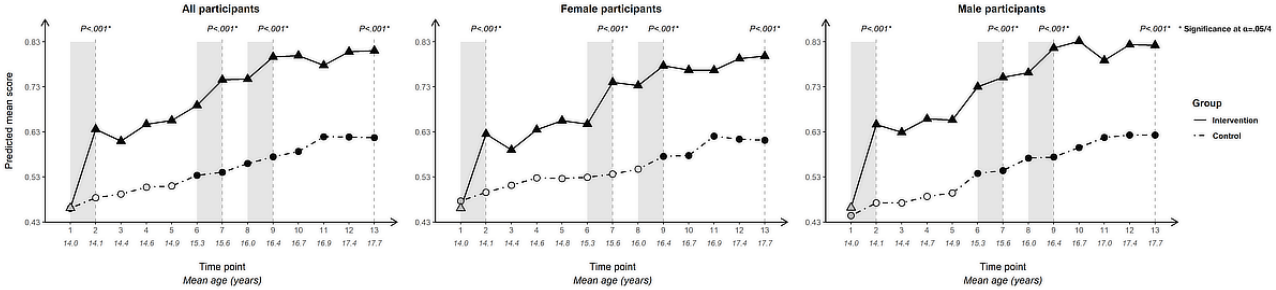

Knowledge

Knowledge of How to Use a Condom Correctly

The 2 condom-related knowledge questions show variations in trajectory. For knowledge of correct condom use (Figure 8), the cross-arm change in score since baseline differed highly significantly ($P<.001$) at all 4 time points and across all population groupings.

Baseline scores were the lowest for any condom-use determinant, particularly among female participants, but saw a spike in the intervention arm following initial exposure that was built upon at the second and third intervention periods.

The within-arm difference since baseline in the intervention arm was significant at all 12 post-baseline time points across all population groupings. The control arm only saw significance from T8 (age 16.0, SD 0.55; T10: age 16.7, SD 0.55, for female participants) onward, with a slight upward trend at later ages, suggesting a mild developmental influence.

Figure 8. Longitudinal trajectory of knowledge of correct condom use at 13 time points for all, female, and male participants.

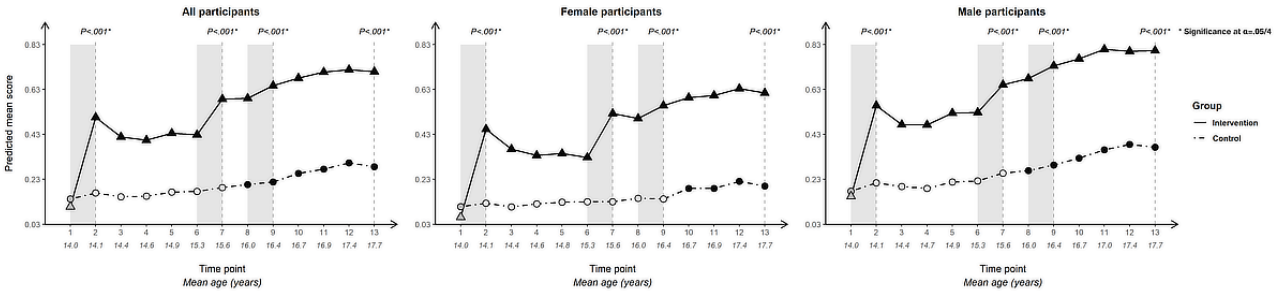

Knowledge That Condoms Are Effective at Preventing HIV

Cross-arm difference in change in knowledge that condoms are an effective way to prevent HIV (Figure 9) was significant following Bonferroni correction immediately after initial intervention exposure and at end line for all participants and for female participants, and at T7 for all participants. It was not significant at any time point for male participants, who had higher scores throughout than female participants.

The within-arm change since baseline was significant for the intervention arm at all time points for all participants.

This was true when stratifying by gender, except for at T5 and T6 for female participants and T3-T5 for male participants. In addition to a steep uptick after initial intervention exposure, the intervention arm saw upticks for the 2 subsequent intervention periods. This is particularly striking for female participants whose mean scores dip between intervention periods but recover following re-exposure to the intervention. Mean scores then stabilize with a slight upward trajectory from this point forward, while the control arm dips, indicating the value of repeated intervention exposure across mid-adolescence for female participants for this outcome. The control arm was nonsignificant for within-arm change since baseline at all time points except for male participants from T10 (age 16.7, SD 0.56) onward.

XSL•FO
RenderX

Figure 9. Longitudinal trajectory of knowledge that condoms are effective at preventing HIV at 13 time points for all, female, and male participants.

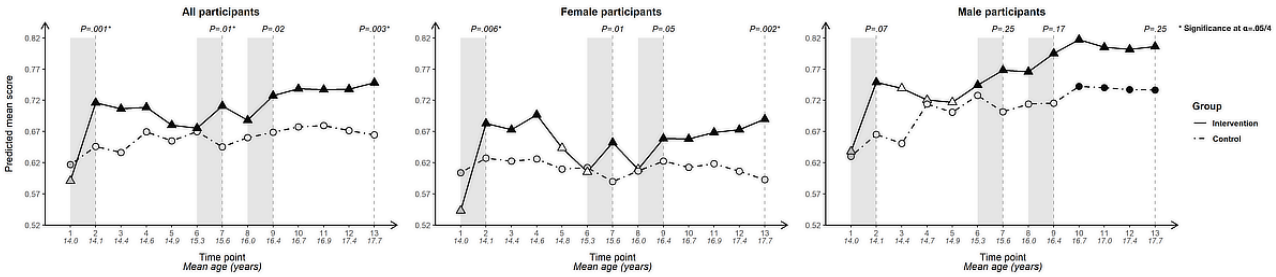

Discussion

Principal Findings

We evaluated the longitudinal impact of a theory-based choose-your-own-adventure smartphone game on proximal outcomes related to condom use across 45 months of mid-adolescence. These proximal outcomes are of particular importance given that assessment of primary behavioral outcomes, related to sexual debut, was not powered until end line, 45 months after participants' initial exposure to Tumaini and 16 months after their most recent exposure. Intervention-arm participants devoted over 30 hours in total to playing Tumaini, while the control-arm game attracted similar levels of engagement, thereby functioning as an effective attention-control condition.

Cross-arm difference: despite cross-arm contamination, Tumaini significantly improved all intervention-targeted proximal influences of condom use relative to the control arm immediately after initial exposure. For almost all proximal outcomes, a significant cross-arm difference was also present at end line and, for most outcomes, at the 2 intervening time points when cross-arm comparison was undertaken, following re-exposure to the intervention.

Within-arm difference: for the intervention arm, when all participants were included in the analysis, the within-arm mean change since baseline was significant for every outcome at every time point. All outcomes saw a spike in intervention-arm mean score immediately after initial exposure to Tumaini. Repeat exposure to the intervention at ages 15.3 (SD 0.55) and 16.0 (SD 0.55) years reinforced these effects in almost every case.

These significant findings and their timing relative to the intervention periods provide compelling evidence of Tumaini's impact and suggest that adolescents may benefit from even short-term exposure, though repeated exposure frequently sustained and strengthened effects. They are also consistent with our primary behavioral outcomes at end line; among participants reporting sexual debut between baseline and end line, 50% (41/82) of control-arm participants had unprotected first sex compared to 28% (23/81) of intervention-arm participants (relative risk [RR] 0.57, 95% CI 0.38-0.85) [19].

Multiple frameworks across health communication, social psychology, and behavioral science support a hierarchy of effects in behavioral interventions, for example [28], whereby knowledge is least resistant to change, followed by attitudes, self-efficacy, behavioral intentions, and finally behavior itself. In line with this hierarchy, while behavioral HIV prevention interventions for adolescents in sub-Saharan Africa have improved condom-related knowledge, attitudes, or self-efficacy, comparatively few have been shown in rigorous studies to reduce sexual risk by increasing condom use [5,29-31]. Our own findings are positive across the hierarchy of effects, from knowledge through attitudes, self-efficacy, and intentions to behavior.

Findings from the present analyses indicate that Tumaini's condom-related content was most effective at strengthening self-efficacy to use a condom (especially among male participants), which is highly dependent on knowledge of how to use a condom correctly, and at informing positive condom-related attitudes. While findings are extremely positive across the board, they also highlight areas where the intervention could potentially be revised to achieve even stronger effects. For example, they suggest that Tumaini could do more to strengthen male participants' self-efficacy to refuse unprotected sex. Trajectories for knowledge that condoms are an effective way to prevent HIV also differ substantially by gender. However, lower scores for female participants may relate less to condom efficacy, which is strongly emphasized throughout the intervention, than to the different levels of agency exercised by female versus male adolescents in condom use.

In addition to being the main strategy for HIV prevention for the population addressed in our trial, condoms are also the primary method of contraception [32]. Tumaini emphasizes the advantages of condoms for preventing pregnancy in addition to HIV. Pregnancy is an outcome that is a more immediate, and hence often more resonant, threat for adolescents—especially girls and young women—than HIV, even in high-prevalence parts of sub-Saharan Africa. Leveraging the long-term consequences of unplanned pregnancy for adolescents' future plans and dreams in condom promotion efforts also has the advantage of being less likely to contribute to HIV stigma. It is notable that trajectories for pregnancy-related proximal outcomes (Figures S2-S4 in Multimedia Appendix 2) were very similar to those for HIV prevention.

In their 2010 systematic review and meta-analysis of HIV prevention for sub-Saharan African youth, Michielsen et al [33] found that effects on condom use were larger in male participants in almost all studies where these were reported as significant. Gender stratification of our primary behavioral outcomes at end line, in contrast, revealed a stronger impact on female participants, for whom the risk of unprotected first sex was reduced significantly from 57% (21/37) in the control arm

XSL•FO
RenderX

to 21% (7/33) in the intervention arm (RR 0.37, 95% CI 0.18-0.76) [19]. Male participants also saw a reduction, but it was nonsignificant, from 44% (20/45) in the control arm to 33% (16/48) in the intervention arm (RR 0.75, 95% CI 0.45-1.26) despite them scoring higher in the present analyses at every time point for intention to use a condom at first sex. Intention may be lower for many adolescent girls and young women in Kenya because they envisage their sexual debut happening within marriage, a context in which they foresee less need to rely on a condom. Lower intervention impact on male participants' condom use at first sex suggests that the translation of intention into behavior may operate differently for male than for female adolescents. It is possible that lower self-efficacy to refuse unprotected sex may be a factor and/or that male participants struggle in the emotionally charged context of first sex when performance anxiety may be exacerbated by condom use.

Alongside those referenced above, multiple structural and contextual factors have been identified as potentially militating against translation of condom-use proximal outcomes up the hierarchy of effects, from knowledge to behavior. Sociocultural factors include, for example, the association of condoms with immorality; gender norms promoting "submissiveness" in adolescent girls and young women; perceptions among young men that condoms reduce sexual pleasure; and a range of myths and misconceptions [15,16,34,35]. Among behavioral theories, social cognitive theory, in which Tumaini is grounded, has the advantage of foregrounding the role played by the social environment in individual behavior. Tumaini's narrative-based approach, drawing on research on thousands of stories contributed by young Africans to scriptwriting competitions [15,16,36], ensures that contextual and structural factors are extensively engaged with.

In addition, although delivered at the individual level, Tumaini either supports or has the potential to support engagement at higher socioecological levels of analysis. It catalyzes considerable household (including parent-child) interaction [37], which is modeled and nurtured in the game by the presence of parent, other adults, and older peer characters who function as "virtual mentors." Further, the high levels of engagement it causes and sustains suggest that, when widely disseminated, its effects could extend to the community level with the potential to support normative change. Depending on population needs, moreover, its effects could be enhanced through dissemination in combination with structural or biomedical programming, alongside anticipated direct-to-consumer dissemination via an app store.

By 2030, smartphones will account for over 80% of all mobile connections in sub-Saharan Africa [7]. Although growth is uneven and there are cross-national, urban-rural, and gender disparities in mobile phone ownership generally and particularly for more expensive phones, penetration is especially rapid among young people [38,39]. At present, over 3 quarters of all mobile phones in Kenya are smartphones (compared to 61% across all of sub-Saharan Africa). While there is no gender disparity in ownership of mobile phones in general in Kenya, there is one for smartphones (49% of men compared to 43% of women own one). In rural areas, a combination of lower data access and cost limits smartphone ownership. Nonetheless, smartphone access and equity in access are increasing in Kenya and across sub-Saharan Africa, and this trajectory is unlikely to change in the near future.

With access barriers in mind, we built Tumaini for a low-cost smartphone as an app that does not need ongoing data access or regular updates. It is designed to be played by adolescents on the phones of parents, siblings, and other family members, or on their own phones or those of their friends; hence, players need access to a smartphone, but they do not need to own one themselves. At the study end line in 2024, 11% of efficacy trial participants (n=106) indicated that no one in their household owned a smartphone, and 17% (n=168) reported owning their own smartphone (though there was a gender disparity, with 21% of boys compared to 14% of girls owning one). Tumaini's scalability and impact outside of efficacy trial conditions warrants further investigation (and ongoing investigation as smartphone access increases).

Strengths and Limitations

Some longitudinal observational studies have followed cohorts of young Africans through adolescence, tracking sexual and reproductive health outcomes [40,41]. However, longitudinal intervention evaluations are rare. An exception is "Let Us Protect Our Future!" with follow-ups at 3, 6, 12, 42, and 54 months [42,43]. Our efficacy trial is notable for the length and frequency of follow-up (13 surveys over 45 months), in addition to its extremely high levels of participant retention (97.5% at end line), its effective attention-control condition, and its high and sustained intervention engagement.

The longitudinal trajectories presented here allow us to observe Tumaini's influence on a range of theory-based, condom-related proximal outcomes across mid-adolescence and across the 3 intervention periods. The within-arm comparisons, with plotlines incorporating intervention periods, provide invaluable contextualization for the cross-arm comparative data and facilitate identification of gender-specific and developmental influences. These analyses thus provide richer insights into the effects of Tumaini, including how these may be optimized, than end line estimates of relative risk alone.

In addition to these strengths, the study has limitations. First, there is a risk that repeated participant exposure to the same survey items over time might blunt the sensitivity of those questions. However, cross-arm differences, upticks in intervention-arm mean scores at each intervention period, and diversity in longitudinal trajectory by outcome suggest that this was not a major concern in this study. Second, while we drew our survey items from scales that had been used and, where possible, also validated with similar populations, we were unable to use the complete validated scales due to the age of our participants (some as young as 12 at baseline) and the risk of overburdening them with an extensive questionnaire or exposing them to non–age-appropriate content. Third, by their nature, our proximal outcomes are self-reported. Specifically, self-efficacy to use a condom correctly and knowledge of how to use a condom correctly are self-assessments, and it is possible that these measures may not ultimately mediate correct condom use in real-life circumstances. Fourth, the study provided

XSL•FO
RenderX

participants with smartphones with parental restrictions blocking internet access; additional research is needed to assess Tumaini's effectiveness under more "real-world" conditions. We are planning a hybrid type 2 effectiveness-implementation study. Fifth, as parents self-selected to learn more about the study, the sample may be biased toward participants whose parents are more open to sexual and reproductive health education for adolescents. Finally, cross-arm contamination may have reduced our ability to detect some intervention effects.

Future Research

Ongoing qualitative research with this cohort may shed additional light on contextual factors informing these trajectories. Mediation analysis currently underway will allow us to pinpoint which proximal outcomes mediated our behavioral outcomes at end line for those who had reached sexual debut; there is no scope for mediation analysis at earlier time points because these outcomes are developmentally influenced, resulting in small cell sizes at younger participant ages. Through this additional research, we will seek to optimize Tumaini prior to our planned hybrid type 2 effectiveness-implementation trial.

Conclusions

Tumaini significantly improved theory-based proximal influences of condom use, with effects sustained 45 months post-initial exposure and 16 months post-most recent exposure. Spikes in intervention-arm scores immediately after initial exposure indicate that adolescents benefit from even short-term exposure, though repeated exposure frequently sustained and reinforced effects. At a time when threats to global HIV funding highlight the necessity of preventing new infections among young people, study findings support the promise of reaching adolescents at an early age with information and skills to support a safer sexual debut. In the context of rapidly increasing access to smartphones in sub-Saharan Africa and at a time when cost-effectiveness is at a premium, Tumaini has potential for high scalability and impact on condom-related outcomes.

Acknowledgments

The authors gratefully acknowledge our research and administrative teams in Kenya and the United States, and especially all communities and participants involved. We thank the Henry Jackson Foundation Medical Research Institute for their support, in particular, Drs Daniel Ochiel and Tiffany Hamm. We extend special thanks to Dr Victor Akelo and Dr Gary W Harper for their contributions to this study.

Funding

Research reported in this publication was supported by the National Institute of Mental Health of the US National Institutes of Health under Award Number 5R01MH118982 (PI: KW). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research was also supported by the Emory Center for AIDS Research (P30 AI050409). The sponsors played no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Data Availability

Data underlying these analyses are not publicly available as the analyses are ongoing. Data are available from the corresponding author on reasonable request.

Authors' Contributions

KW led the design of the study, and GS, KK, RAB, RHL, VM, and KO contributed to the design of the study. VM, KO, JA, and CM contributed to protocol implementation, data collection, and management. SG, HQ, and GS conducted data analyses, guided by RHL, KW, RAB, and KK. KW drafted the manuscript; all other authors edited and reviewed the manuscript. KW is the overall principal investigator, while VM is the site principal investigator for this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplemental tables with additional outcomes and mean outcome scores at 13 time points and associated within- and cross-arm change.
[DOCX File , 61 KB - jmir_v28i1e83982_app1.docx]

Multimedia Appendix 2
Longitudinal trajectories of additional survey items with relevance to condom use at 13 time points among all, female, and male participants.
[DOCX File , 669 KB - jmir_v28i1e83982_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 625 KB - jmir_v28i1e83982_app3.pdf]

Multimedia Appendix 4
Higher-resolution versions of figures 2-9.
[ZIP File (Zip Archive), 1339 KB - jmir_v28i1e83982_app4.zip]

References

1. The urgency of now: AIDS at a crossroads. Joint United Nations Programme on HIV/AIDS (UNAIDS). 2024. URL: https://www.unaids.org/en/resources/documents/2024/global-aids-update-2024?utm_source=chatgpt.com [accessed 2026-01-16]
2. Benevides R, Logan J, Patel P. Introduction to the youth HIV prevention supplement. AIDS Educ Prev 2023;35:1-3. [doi: 10.1521/aeap.2023.35.suppa.1]
3. Poobalan AS, Pitchforth E, Imamura M, Tucker JS, Philip K, Spratt J, et al. Characteristics of effective interventions in improving young people's sexual health: a review of reviews. Sex Educ 2009;9(3):319-336. [doi: 10.1080/14681810903059185]
4. Pedlow CT, Carey MP. Developmentally appropriate sexual risk reduction interventions for adolescents: rationale, review of interventions, and recommendations for research and practice. Ann Behav Med 2004;27(3):172-184 [FREE Full text] [doi: 10.1207/s15324796abm2703_5] [Medline: 15184093]
5. Shangase N, Kharsany ABM, Ntombela NP, Pettifor A, McKinnon LR. A systematic review of randomized controlled trials of school based interventions on sexual risk behaviors and sexually transmitted infections among young adolescents in sub-Saharan Africa. AIDS Behav 2021;25(11):3669-3686 [FREE Full text] [doi: 10.1007/s10461-021-03242-8] [Medline: 33772695]
6. Li DH, Macapagal K, Mongrella M, Saber R, Mustanski B. "Your package could not be delivered": the state of digital HIV intervention implementation in the US. Curr HIV/AIDS Rep 2024;21(3):152-167. [doi: 10.1007/s11904-024-00693-1] [Medline: 38502421]
7. The mobile economy sub-Saharan Africa. GSMA. 2024. URL: https://www.gsma.com/solutions-and-impact/connectivity-for-good/mobile-economy/sub-saharan-africa-2024/?utm_source=chatgpt.com [accessed 2026-01-16]
8. Hightow-Weidman L, Muessig K, Bauermeister J, LeGrand S, Fiellin L. The future of digital games for HIV prevention and care. Curr Opin HIV AIDS 2017;12(5):501-507 [FREE Full text] [doi: 10.1097/COH.0000000000000399] [Medline: 28692490]
9. DeSmet A, Baranowski T, Thompson D, Shirong Lu A. Fantasy, facts and fun: digital health games for impact and implementation. In: de Bruijn GJ, Vandebosch H, editors. Handbook on Health, Media and Communications. Berlin, Boston: De Gruyter Mouton; 2025:287-308.
10. Winskell K, Sabben G, Obong'o C. Interactive narrative in a mobile health behavioral intervention (Tumaini): theoretical grounding and structure of a smartphone game to prevent HIV among young Africans. JMIR Serious Games 2019;7(2):e13037 [FREE Full text] [doi: 10.2196/13037] [Medline: 31066692]
11. Bandura A. Social foundations of thought and action: a social-cognitive view. Englewood Cliffs, NJ: Prentice-Hall; 1986.
12. Sood S, Menard T, Witte K. The theory behind entertainment-education. In: Singhal A, Cody M, Rogers E, Sabido M, editors. Entertainment-Education and Social Change. Mahwah, NJ and London: Lawrence Erlbaum Associates; 2004.
13. Singhal A, Cody M, Roger E, Sabido M. Entertainment-Education and Social Change: History, Research, and Practice. Mahwah, NJ: Lawrence Erlbaum Associates; 2004.
14. Winskell K. Social representations theory and young Africans' creative narratives about HIV/AIDS, 1997-2014. J Theory Soc Behav 2021;51(1):164-182 [FREE Full text] [doi: 10.1111/jtsb.12270] [Medline: 34483357]
15. Winskell K, Obyerodhyambo O, Stephenson R. Making sense of condoms: social representations in young people's HIV-related narratives from six African countries. Soc Sci Med 2011;72(6):953-961 [FREE Full text] [doi: 10.1016/j.socscimed.2011.01.014] [Medline: 21388731]
16. Winskell K, Singleton R, Sabben G, Tiendrébéogo G, Obong'o C, Dia FL, et al. Social representations of the prevention of heterosexual transmission of HIV among young Africans from five countries, 1997-2014. PLoS One 2020;15(3):e0227878 [FREE Full text] [doi: 10.1371/journal.pone.0227878] [Medline: 32142515]
17. Winskell K, Sabben G, Akelo V, Ondeng'e K, Obong'o C, Stephenson R, et al. A smartphone game-based intervention (Tumaini) to prevent HIV among young Africans: pilot randomized controlled trial. JMIR Mhealth Uhealth 2018;6(8):e10482 [FREE Full text] [doi: 10.2196/10482] [Medline: 30068501]
18. Sabben G, Mudhune V, Ondeng'e K, Odero I, Ndivo R, Akelo V, et al. A smartphone game to prevent HIV among young Africans (Tumaini): assessing intervention and study acceptability among adolescents and their parents in a randomized controlled trial. JMIR Mhealth Uhealth 2019 May 21;7(5):e13049 [FREE Full text] [doi: 10.2196/13049] [Medline: 31115348]

XSL•FO
RenderX

19. Winskell K, Sabben G, Mbeda C, Ondeng'e K, Arego J, Qin H, et al. Efficacy of a smartphone game (Tumaini) to increase age and condom use at first sex among adolescents in Kenya: a randomised controlled trial. Lancet Child Adoles Health 2026;10:273-283. [doi: 10.1016/S2352-4642(25)00353-0]
20. Banadakoppa Manjappa R, Bhattacharjee P, Shaw SY, Gitonga J, Kioko J, Songok F, et al. A sub-national HIV epidemic appraisal in Kenya: a new approach for identifying priority geographies, populations and programmes for optimizing coverage for HIV prevention. J Int AIDS Soc 2024;27 Suppl 2(Suppl 2):e26245. [doi: 10.1002/jia2.26245] [Medline: 38982894]
21. Truong HM, Guzé MA, Ouma D, Bushman D, Mocello AR, Kadede K, et al. Community-based HIV testing for urban youth in Western Kenya. AIDS Behav 2022;26(3):814-821 [FREE Full text] [doi: 10.1007/s10461-021-03441-3] [Medline: 34518938]
22. Mudhune V, Sabben G, Ondenge K, Mbeda C, Morales M, Lyles RH, et al. The efficacy of a smartphone game to prevent HIV among young Africans: protocol for a randomized controlled trial in the context of COVID-19. JMIR Res Protoc 2022;11(3):e35117 [FREE Full text] [doi: 10.2196/35117] [Medline: 35030090]
23. Vandenhoudt H, Miller KS, Ochura J, Wyckoff SC, Obong'o CO, Otwoma NJ, et al. Evaluation of a U.S. evidence-based parenting intervention in rural Western Kenya: from parents matter! To families matter!. AIDS Educ Prev 2010;22(4):328-343. [doi: 10.1521/aeap.2010.22.4.328] [Medline: 20707693]
24. Learning from adolescents to prevent HIV and unintended pregnancy. In: Issues Brief (Alan Guttmacher Inst). New York, NY: Guttmacher Institute; 2007.
25. Ndebele M, Kasese-Hara M, Greyling M. Application of the information, motivation and behavioural skills model for targeting HIV risk behaviour amongst adolescent learners in South Africa. SAHARA J 2012;9 Suppl 1:S37-S47 [FREE Full text] [doi: 10.1080/17290376.2012.744903] [Medline: 23234378]
26. Liang K, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73(1):13-22 [FREE Full text] [doi: 10.1093/biomet/73.1.13]
27. Pan W. Akaike's information criterion in generalized estimating equations. Biometrics 2001;57(1):120-125. [doi: 10.1111/j.0006-341x.2001.00120.x] [Medline: 11252586]
28. McGuire W. Theoretical foundations of campaigns. In: Rice R, Atkin C, editors. Public communication campaigns. Newbury Park, CA: Sage; 1989:43-65.
29. Mwale M, Muula AS. Systematic review: a review of adolescent behavior change interventions [BCI] and their effectiveness in HIV and AIDS prevention in sub-Saharan Africa. BMC Public Health 2017;17(1):718 [FREE Full text] [doi: 10.1186/s12889-017-4729-2] [Medline: 28923040]
30. Muthoni CN, Kneipp SM, Gichane MW, Caiola CE, Pettifor AE, Williams JR. A systematic review of HIV interventions for young women in sub-Saharan Africa. AIDS Behav 2020;24(12):3395-3413. [doi: 10.1007/s10461-020-02914-1] [Medline: 32410052]
31. Sani AS, Abraham C, Denford S, Ball S. School-based sexual health education interventions to prevent STI/HIV in sub-Saharan Africa: a systematic review and meta-analysis. BMC Public Health 2016;16(1):1069 [FREE Full text] [doi: 10.1186/s12889-016-3715-4] [Medline: 27724886]
32. Kenya demographic and health survey. KNBS (Kenya National Bureau of Statistics). 2022. URL: https://www.knbs.or.ke/reports/kdhs-2022/?utm_source=chatgpt.com [accessed 2026-01-16]
33. Michielsen K, Chersich MF, Luchters S, De Koker P, Van Rossem R, Temmerman M. Effectiveness of HIV prevention for youth in sub-Saharan Africa: systematic review and meta-analysis of randomized and nonrandomized trials. AIDS 2010;24(8):1193-1202. [doi: 10.1097/QAD.0b013e3283384791] [Medline: 20375876]
34. Casmir E, Daniel AK, Ongolly F, Thuo N, Oluoch L, Kiptinness C, et al. Protection at first sexual intercourse among adolescent girls and young women in Kenya. Arch Sex Behav 2021;50(1):219-227 [FREE Full text] [doi: 10.1007/s10508-020-01785-w] [Medline: 32720186]
35. Tavrow P, Karei EM, Obbuyi A, Omollo V. Community norms about youth condom use in Western Kenya: is transition occurring? Afr J Reprod Health 2012;16(2):241-252. [Medline: 22916556]
36. Winskell K, Beres LK, Hill E, Mbakwem BC, Obyerodhyambo O. Making sense of abstinence: social representations in young Africans' HIV-related narratives from six countries. Cult Health Sex 2011;13(8):945-959 [FREE Full text] [doi: 10.1080/13691058.2011.591431] [Medline: 21787256]
37. Winskell K, Sabben G, Ondeng'e K, Odero I, Akelo V, Mudhune V. A smartphone game to prevent HIV among young Kenyans: household dynamics of gameplay in a feasibility study. Health Educ J 2019;78(5):595-606 [FREE Full text] [doi: 10.1177/0017896919832344] [Medline: 34219796]
38. The mobile gender gap report. GSMA. 2024. URL: https://www.gsma.com/r/gender-gap/ [accessed 2026-01-16]
39. Kenya demographic and health survey 2022. Kenya National Bureau of Statistic and ICF. 2023. URL: https://www.knbs.or.ke/reports/kdhs-2022/ [accessed 2026-01-20]
40. Richter L, Norris S, Pettifor J, Yach D, Cameron N. Cohort profile: Mandela's children: the 1990 birth to twenty study in South Africa. Int J Epidemiol 2007;36(3):504-511 [FREE Full text] [doi: 10.1093/ije/dym016] [Medline: 17355979]
41. Baird S, Chirwa E, McIntosh C, Özler B. The short‐term impacts of a schooling conditional cash transfer program on the sexual behavior of young women. Health Economics 2010;19(S1):55-68. [doi: 10.1002/hec.1569]

XSL•FO
RenderX

42. Jemmott JB, Jemmott LS, O'Leary A, Ngwane Z, Icard LD, Bellamy SL, et al. School-based randomized controlled trial of an HIV/STD risk-reduction intervention for South African adolescents. Arch Pediatr Adolesc Med 2010 Oct;164(10):923-929 [FREE Full text] [doi: 10.1001/archpediatrics.2010.176] [Medline: 20921349]
43. Jemmott JB, Jemmott LS, O'Leary A, Ngwane Z, Lewis DA, Bellamy SL, et al. HIV/STI risk-reduction intervention efficacy with South African adolescents over 54 months. Health Psychol 2015;34(6):610-621 [FREE Full text] [doi: 10.1037/hea0000140] [Medline: 25110841]

Abbreviations

KEMRI: Kenya Medical Research Institute
RR: relative risk

Edited by A Stone; submitted 15.Sep.2025; peer-reviewed by P Kreniske, S Naar; comments to author 30.Oct.2025; revised version received 19.Dec.2025; accepted 30.Dec.2025; published 10.Mar.2026.

Please cite as:
Winskell K, Sabben G, Qin H, Mbeda C, Goldenberg S, Ondeng'e K, Ndivo R, Arego J, He X, Bednarczyk RA, Komro K, Lyles RH, Mudhune V
Longitudinal Effects of a Smartphone Game (Tumaini) for HIV Prevention Among Kenyan Adolescents: 45-Month Trajectories of Condom Use–Related Proximal Outcomes From a Randomized Controlled Trial
J Med Internet Res 2026;28:e83982
URL: https://www.jmir.org/2026/1/e83982
doi:10.2196/83982
PMID:

©Kate Winskell, Gaëlle Sabben, Haowen Qin, Calvin Mbeda, Sophie Goldenberg, Ken Ondeng'e, Richard Ndivo, Judith Arego, Xinwei He, Robert A Bednarczyk, Kelli Komro, Robert H Lyles, Victor Mudhune. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Detection of Microbehavior Intervals for Predicting Mental Health: Clinically Relevant and Advanced Multimodal Temporal Analysis

Sapir Gershov[1], PhD; Charlotte E Hilberdink[1], PhD; Yiwen Zhao[1], MSc; Sarah B Birnbaum[1], BSc; Victoria Mueller[1,2], MSc; Stephen P Wall[3,4], MD, MSc, MAEd; Katharina Schultebraucks[1,4], PhD

[1]Department of Psychiatry, Grossman School of Medicine, New York University, New York, NY, United States

[2]Department of Biomedical Engineering, Tandon School of Engineering, New York University, New York, NY, United States

[3]Department of Emergency Medicine, Grossman School of Medicine, New York University, New York, NY, United States

[4]Department of Population Health, Grossman School of Medicine, New York University, New York, NY, United States

Corresponding Author:
Katharina Schultebraucks, PhD
Department of Psychiatry
Grossman School of Medicine
New York University
1 Park Ave., 8th Floor
New York, NY, 10016
United States
Phone: 1 646 754 4929
Email: Katharina.Schultebraucks@nyulangone.org

Abstract

Background: Health care workers (HCWs) face sustained psychological demands that place them at heightened risk for burnout and posttraumatic stress disorder (PTSD). However, assessing psychological distress in this population remains challenging because of stigma, underreporting, and the limitations of self-report tools. Although nonverbal behaviors such as facial expressions and gaze hold diagnostic promise, most approaches overlook the fine-grained, temporal fluctuations in these signals. In this study, we focused on microbehavior intervals—brief, involuntary changes in multimodal nonverbal signals—that emerge during emotion-eliciting interviews.

Objective: This study aimed to determine whether microbehavior intervals improve the discrimination of psychological distress profiles among HCWs with symptoms of burnout and PTSD.

Methods: HCWs participated in a semistructured interview that included 5 work-related, emotionally charged questions and that was recorded via Webex (online video platform). Participants also completed validated questionnaires for burnout (Maslach Burnout Inventory General Survey 9-item) and PTSD (PTSD checklist for *Diagnostic and Statistical Manual, 5th edition*). Recordings were analyzed with computer vision models to generate time-series data of facial expressions, head movement, gaze, body posture, and hand gestures. An unsupervised anomaly detection model (MOMENT [a Family of Open Time-Series Foundation Models]) isolated microbehavior intervals without requiring manual labels. Features derived from these intervals were used to train a deep learning classifier that predicted 4 symptom classes of psychological distress: “moderate-severe burnout,” “subthreshold-provisional PTSD,” “burnout+PTSD,” and “resilient.” We conducted an ablation study by systematically removing one behavioral data stream at a time. Finally, we conducted an explainability analysis to characterize the features driving model predictions.

Results: We analyzed 258 interview recordings from 151 HCWs. Per interview, an average of 19.65 (SD 6.01) microbehavior intervals were detected, each lasting an average of 1.31 (SD 1.10) seconds. The classifier demonstrated robust performance across classes, achieving a macro–F_1-score of 0.75 and a macro area under the receiver operating characteristic curve of 0.80 on held-out data. Ablation analysis showed that excluding gaze or arousal-valence signals caused the largest performance declines, particularly in recall and F_1-score. The explainability analysis revealed distinct temporal patterns across symptom classes, with irregularity and variability in microbehaviors emerging as key predictors.

Conclusions: Focusing on microbehavior intervals yields a scalable, interpretable, and annotation-free framework for detecting psychological distress from nonverbal signals. By moving from whole-video features to fine-grained multimodal temporal modeling, we successfully captured subtle, involuntary fluctuations in nonverbal responses to emotion-eliciting questions. This

XSL•FO
RenderX

multimodal approach enables an objective, robust, and explainable assessment of psychological distress and offers a promising complement to conventional psychometric assessments.

(J Med Internet Res 2026;28:e87049) doi:10.2196/87049

KEYWORDS

burnout; posttraumatic stress disorder; PTSD; anomaly detection; multimodal; health care workers

Introduction

Frontline professionals operating in high-stakes environments—such as health care workers (HCWs) [1], military personnel [2], and first responders [3]—are frequently exposed to emotionally intense, cognitively demanding, and potentially traumatic events. These chronic occupational stressors place them at elevated risk for burnout and posttraumatic stress disorder (PTSD) [4], which frequently co-occur and exacerbate psychological burden [5,6].

Burnout is an occupational syndrome characterized by exhaustion and disengagement [7], and it affects approximately one-third to one-half of HCWs [8-10]. PTSD, in contrast, is a trauma-related disorder defined by intrusive memories, hyperarousal, and avoidance [11], and symptoms are reported by up to 1 in 5 trauma-exposed HCWs [12,13]. Despite this burden, assessing symptom severity remains a major challenge due to stigma, underreporting, emotion suppression, and the limitations of traditional self-report tools [3,14,15]. These barriers highlight the need for complementary methods that can detect psychological distress without relying solely on verbal disclosure or self-report questionnaires.

Nonverbal cues offer a valuable, often underused dimension for psychiatric assessments, particularly in individuals who are unaware of or suppress their emotional states [16,17]. Among these cues, facial microexpressions, which represent brief, spontaneous, and involuntary muscle movements that are difficult to consciously regulate, have garnered substantial attention as potential reliable and objective indicators of internal affect [18-20]. Recent machine learning studies have demonstrated that subtle irregularities in the temporal patterns of microexpressions are strongly associated with individuals with depression [16,21], highlighting their potential as objective markers for mental health assessment. However, the clinical utility of microexpressions remains limited by several challenges, including: (1) data acquisition using high-frame-rate cameras (100-200 frames per second), (2) the need for highly sensitive algorithms, (3) time-intensive annotation, and (4) substantial interrater variability even among expert annotators [20].

Recent evidence suggests that multimodal machine learning frameworks, which analyze video-recorded interviews using computer vision and natural language processing, may offer an alternative methodology for screening for psychological disorders [22-25]. Although promising, many of these approaches reduce dynamic behavior to static average features [22] or extract only sentence-level features [26], thereby overlooking the fine-grained temporal fluctuations in verbal and nonverbal cues. By collapsing rich behavioral dynamics into summary statistics, these methods risk omitting subtle yet meaningful signals that may carry critical diagnostic information [17,27-29].

Although facial microexpressions have traditionally served as the canonical example of emotionally revealing nonverbal behavior, the underlying mechanism—involuntary expression of emotional states through brief, spontaneous motor acts—is shared across many other nonverbal behavior modalities. In the context of burnout and trauma-related disorders, such involuntary expressions can manifest as blunted or exaggerated facial expressivity [30], disrupted eye contact [31,32], tense or withdrawn posture [33], and restlessness [34,35]. This broader framing moves beyond face-centric paradigms, allowing us to examine a more holistic perspective of nonverbal behaviors as high-resolution, interpretable signals of internal distress.

Spontaneous changes in nonverbal behavior signals, termed microbehaviors, can be modeled in the same way as microexpressions. Specifically, we can temporally localize fluctuations and changes in body movement that arise spontaneously in response to a stimulus. Microbehaviors manifest as sparse spikes within continuous nonverbal signals and exhibit distinct onset, apex, and offset phases. When treated as a time-series signal, microbehaviors can be segmented and interpreted using time-series analytical techniques such as anomaly detection [36]. Our hypothesis is that abrupt fluctuations in multimodal nonverbal behaviors may reflect emotional dysregulation and that, by quantifying their spatiotemporal changes, we can use these signals as objective digital biomarkers of psychological distress.

To test this hypothesis, we developed an unsupervised multimodal video analysis framework that captures microbehaviors using multivariate time-series modeling and zero-shot anomaly detection. Our framework focuses exclusively on facial expression, head movement, gaze, body posture, and hand movement—nonverbal modalities consistently linked to burnout [32] and PTSD [24,31,33]. Although speech-based and linguistic features have shown promise in prior work [24,37-39], we intentionally excluded them from our framework because their interpretation depends on language-specific models that may not generalize across cultural contexts, can be biased by differences in language fluency, and may be influenced by an individual's comfort with verbal self-disclosure [40]. By contrast, nonverbal behaviors are considered more language-independent and culturally robust, thus making them more suitable for detecting psychological distress [41-43].

In summary, this study introduces a novel multimodal framework that identifies brief periods of heightened behavioral expressivity and evaluates whether temporal features derived

XSL•FO
RenderX

from these intervals can predict burnout and PTSD symptom profiles among HCWs.

Methods

Participants

Participants were drawn from "Early Signs: digital phenotyping to identify digital biomarkers for predicting burnout and cognitive functioning in emergency department clinicians," an ongoing National Institutes of Health (NIH)–funded longitudinal study (R01HL156134) examining burnout and well-being among HCWs in emergency and trauma care. Inclusion criteria required participants to be aged ≥18 years, fluent in English, and engaged in direct patient contact with a typical full-time clinical schedule (including individuals who were temporarily working part-time due to burnout or leave). Medical students and temporary staff were not eligible. Recruitment was conducted among emergency and trauma clinicians and emergency department staff working within major health systems across the New York City metropolitan area. Outreach efforts combined targeted online advertisement (eg, Facebook [Meta Platforms Inc] and iConnect) with structured in-person engagement by trained research staff. Interested individuals were directed to an institutional review board–approved key information sheet and study contact information. Research staff subsequently followed a standardized workflow to verify eligibility, address questions, and obtain written informed consent before scheduling an in-person assessment. Recruitment settings and procedures were standardized across participating sites.

As part of this study, participants were scheduled for semistructured interviews, which were recorded via an institution-approved, HIPAA (Health Insurance Portability and Accountability Act)-compliant secure video platform (Webex; Cisco Systems Inc). The interview included questions regarding work-related events and stressful experiences. In addition, participants completed self-reported questionnaires regarding their current psychological symptom levels for burnout and PTSD.

Ethical Considerations

The study protocol was reviewed and approved by the institutional review board of New York University Grossman School of Medicine (i23-00060) and was conducted in accordance with the Declaration of Helsinki and the US Federal Policy for the Protection of Human Subjects (Common Rule, 45 CFR 46 [44]). All participants provided written informed consent before participation. Participants were informed that participation was voluntary, that they could decline to answer any question or withdraw from the study at any time, and that their decision to participate or withdraw would not affect their employment, salary, benefits, clinical care, or performance evaluations.

To protect participant privacy and confidentiality, interviews were recorded using an institution-approved, HIPAA-compliant secure video platform (Webex). Because video and audio recordings may contain identifiable information, recordings and related study data were stored on secure institutional systems, labeled with study identifiers when possible, and accessible only to authorized study personnel. Identifying information was not included in the analytic datasets used for this manuscript, and no identifiable participant images, audio, video, or direct quotations are presented in the article or supplementary materials.

Participants were compensated for their time and reimbursed for travel-related costs associated with study participation. Compensation was provided according to the institutional review board–approved study protocol and was not contingent on completion of all study procedures.

Semistructured Interview

The interview included 5 open-ended questions targeting emotionally evocative situations, such as adverse patient outcomes and future expectations. During the recording of the interview, only the participant's camera was turned on, which allowed us to capture their spontaneous, nonverbal responses with minimal interference. This design aimed to naturally evoke emotional and cognitive states relevant to burnout and trauma-related symptoms while preserving the conversational flow and ecological validity of real-world interactions. The interview questions are provided in the Multimedia Appendix 1.

Psychometric Assessments

Burnout symptoms were assessed using the abbreviated Maslach Burnout Inventory General Survey 9-item [45], which includes 9 items across 3 subscales: emotional exhaustion, depersonalization, and personal accomplishment. Following Rotenstein et al [46], participants were labeled as presenting "severe" symptoms if they met the following thresholds: personal accomplishment ≤12, emotional exhaustion ≥11, and depersonalization ≥7. Those who met just 2 of the 3 thresholds were labeled as "moderate." Participants who did not meet the criteria for "severe" or "moderate" were labeled as "no burnout."

PTSD symptoms were measured using the PTSD checklist for *Diagnostic and Statistical Manual, 5th edition* [47], a 20-item scale based on *Diagnostic and Statistical Manual, 5th edition* criteria. According to Klein et al [48] and Biscoe et al [49], participants were labeled as "provisional PTSD" if they met the diagnostic threshold across all 4 symptom clusters or had a total score of ≥31. Those meeting only 3 clusters and scoring <31 were labeled as "subthreshold PTSD." Participants who did not meet the criteria for "provisional PTSD" or "subthreshold PTSD" were labeled as "no PTSD."

On the basis of their symptomatology, participants were divided into the following 4 classes: "resilient," comprising participants who met neither burnout nor PTSD symptoms criteria; "moderate-severe burnout," comprising participants who met only "moderate" or "severe" burnout symptoms criteria; "subthreshold-provisional PTSD," comprising participants who met only "provisional" or "subthreshold" PTSD symptoms criteria; and "burnout+PTSD," comprising participants who met criteria for both "moderate-severe burnout" and "subthreshold-provisional PTSD."

XSL•FO
RenderX

End-to-End Framework Construction

Overview

Figure 1 illustrates the complete end-to-end architecture of our framework, encompassing (1) video signal processing, (2) anomaly detection, (3) feature extraction, and (4) classification model training and analysis. All processes were executed on a single NVIDIA A100 Tensor Core GPU (NVIDIA Corp) using Python (version 3.10), PyTorch (version 2.0.0), and scikit-learn (version 1.6.1). We followed best-practice reporting guidelines for prediction modeling and machine learning in health research (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–artificial intelligence [50] and Prediction Model Risk of Bias Assessment Tool–artificial intelligence [51]).

Figure 1. Overview of the proposed framework. The pipeline is divided into 2 components: (A) an unsupervised video analysis framework that extracts multivariate time-series data from gaze direction (3DGazeNet), facial activity (facetorch), body movement (TokenHMR), and hand movement (WiLoR; microbehavior intervals are detected via zero-shot anomaly detection (MOMENT [a Family of Open Time-Series Foundation Models]) and are then analyzed to construct interval features) and (B) a symptomatology classification component in which features, derived from intervals, are used to train a deep learning model that predicts symptomatology classes derived from self-report psychological measures (model interpretability is assessed using SHAP [Shapley Additive Explanation], which highlights features contribution to each class). PTSD: posttraumatic stress disorder.

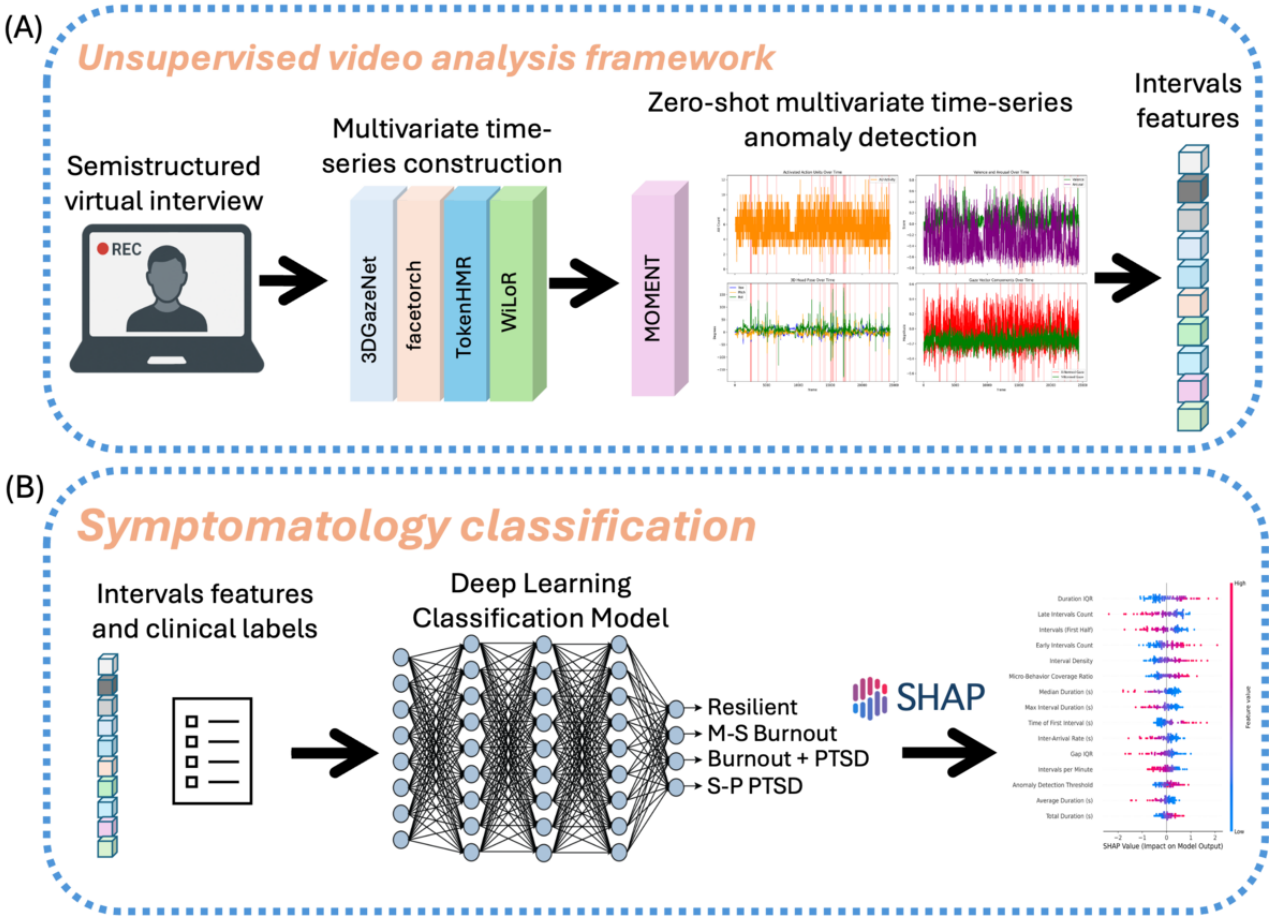

Video Acquisition and Preprocessing

Assessments were recorded remotely using the Webex platform and a standard webcam, producing videos at approximately 25 frames per second and 720p to 1080p resolution. All recordings were compressed using the Moving Picture Experts Group-4 format, and the full interview video was processed without any content-based manual trimming. Because webcam recordings exhibited occasional frame drops, we relied on decoded time stamps to maintain temporal ordering.

Multimodal Nonverbal Behavioral Time Series

To capture dynamic nonverbal signals indicative of psychological distress, we implemented a multimodal framework that processes raw video-recorded interviews and constructs comprehensive time series that are both multimodal and multivariate. Each video was processed frame-by-frame using state-of-the-art computer vision models, resulting in the following behavioral time series:

- Facial expressions: we used “facetorch” [52], a modular affective computing framework, which provided (1) binary activation of facial action units, encoding fine-grained facial muscle movements, and (2) continuous valence and arousal estimates based on facial expressions.
- 3D head pose estimation: we extracted 3D head orientation (yaw, pitch, and roll) using the “facetorch” pose estimation module.
- 3D gaze estimation: we used “3DGazeNet” [53] to estimate participants’ 3D gaze direction. This model was selected for its robustness to head pose and lighting variations,

XSL•FO
RenderX

making it suitable for naturalistic, webcam-based recordings.

- 3D body pose estimation: we used "TokenHMR" [54], which estimates 3D human body pose based on predefined key points in a temporally consistent manner. For each frame, we extracted the 3D coordinates of upper-body key points and quantified gross motor activity as the mean Euclidean distance (displacement) between key points across consecutive frames, yielding a per-frame movement-magnitude time series.
- 3D hand pose estimation: fine-grained hand movements were estimated using "WiLoR" [55], a 3D hand localization and mesh reconstruction model. This provided us with 3D key points for both hands, allowing us to track gestural patterns.

Frames where a target signal was not observable (eg, face or hands out of frame) were retained in the sequence and coded as missing outputs. All modalities were synchronized to ensure temporal alignment. For each frame, the extracted time series were concatenated to form a multivariate (multichannel) representation that preserves the fine-grained temporal structure of the recorded nonverbal behavior.

Anomaly Detection of Microbehavior

To identify microbehavior intervals throughout the interview recording, we applied an anomaly detection to the multivariate time series. Anomaly detection models define "normal" time-series patterns that conform to their learned expectations of smoothness, predictability, and local temporal structure. Deviations from these expectations—such as abrupt changes, irregular bursts, or unexpected fluctuations—are interpreted as "anomalies," and represented as numerical anomaly scores that quantify the extent to which a data point deviates from what is considered normal within a sequence of data points collected over time. Higher anomaly scores indicate a greater likelihood of a data point being an anomaly.

In our framework, we used MOMENT (a Family of Open Time-Series Foundation Models) [56], which is a foundation model trained on diverse time-series tasks, selected for its ability to detect outlier patterns in a zero-shot manner without requiring labeled data or domain-specific fine-tuning. MOMENT anomaly scores quantify reconstruction-based deviations from expected behavior and are computed independently for each univariate signal channel (ie, nonverbal behavior) at each time point. These scores vary in scale, depending on the input signal, in which higher values indicate a more behaviorally or contextually abnormal pattern within the observed sequence. A frame-level total anomaly score was obtained by summing the anomaly scores across nonverbal behaviors. Microbehavior segments were defined as contiguous windows in which the frame-level total anomaly score exceeded the 95th percentile threshold. Finally, on the basis of empirical observation, segments separated by <1 second were merged to form discrete microbehavior intervals. We verified that this rule primarily reduced overfragmentation (often caused by transient tracking noise or brief occlusions) without materially changing the overall interval distribution. This gap-bridging step is consistent with prior work in microexpression and frame-based affect modeling, where expressions are inherently brief (often subsecond) and continuous runs of frames are treated as a single event rather than multiple occurrences [16,21].

It is important to note that no human annotations or manual coding were used to identify anomalous frames or intervals; microbehavior intervals were derived solely from MOMENT anomaly scores using the fixed percentile-thresholding and merging rules described previously.

Interval Feature Extraction

After the segmentation procedure, for each video recording, we extracted 36 interval-level statistical and temporal features that summarize the microbehavior intervals (eg, number of intervals, average duration, density, variability). These interval features were computed manually using a custom Python implementation.

Classification Model Training and Explainability

The final part of the framework involved training a deep learning classification model using the extracted features of the microbehavior intervals to predict the 4 symptom classes ("resilient," "moderate-severe burnout," "subthreshold-provisional PTSD," and "burnout+PTSD").

A custom classification model, a feedforward fully connected multilayer perceptron, was constructed and optimized using Optuna (version 4.8.0) [57]. The full dataset was split into a training set (80%) and an independent held-out test set (20%), using stratified sampling to preserve class distribution. Model training and hyperparameter fine-tuning were performed using stratified 5-fold cross-validation with cross-entropy loss and macro–F_1-score as the objective metric. To avoid overfitting, we also initiated early stopping based on validation loss. To prevent information leakage, all splits (training and test sets, as well as cross-validation folds) were grouped by participant, ensuring that recordings from the same HCW did not appear in both the training and evaluation sets.

To address the imbalance of symptom classes within our dataset, the Synthetic Minority Oversampling Technique [58] was applied within each training fold. Model performance was assessed on the held-out test set using weighted accuracy, precision, recall, and macro–F_1-score, as well as macro, micro, and per-class area under the receiver operating characteristic curve scores (Figure 2). Additionally, we evaluated the influence of different modalities on model performance by conducting an ablation study (Figure S4 in the Multimedia Appendix 1).

Finally, we conducted an explainability analysis of the classification decisions using Shapley Additive Explanations (SHAP) [59], highlighting the most influential microbehavior interval features for each class (Figure 3).

Results

Overview

We analyzed 258 interview recordings from 151 HCWs. The average interview length was 16.08 (SD 1.72) minutes. Participants' average age was 34.24 (SD 8.48) years, and the majority were female (n=93, 61.6%). Professional roles included

XSL•FO
RenderX

registered nurses (n=54, 35.8%), resident physicians (n=47, 31.1%), faculty physicians (n=19, 12.6%), and other health care professionals (n=31, 20.5%), with an average of 4.96 (SD 5.70) years of clinical employment.

On the basis of our labeling procedure, the following are the sizes of our 4 symptom classes: "resilient"=121 (47%), "moderate-severe burnout"=54 (21%), "subthreshold-provisional PTSD"=31 (12%), and "burnout+PTSD"=52 (20%). Tables 1 and 2 summarize the demographics and clinical characteristics of the sample. Additional details about the sample are provided in Table S1 and Figure S1 in the Multimedia Appendix 1.

Table 1. Demographic characteristics of health care workers (N=151)a.

Characteristics	Values
Age (years), mean (SD)	34.24 (8.48)
Sex at birth, n (%)	
Female	93 (61.59)
Male	58 (38.41)
Current position, n (%)	
Faculty physician	19 (12.58)
Resident physician	47 (31.13)
Registered nurse	54 (35.76)
Other	31 (20.53)
Duration of employment (years), mean (SD)	4.96 (5.7)

[a]Continuous variables are presented as mean (SD) and categorical variables as n (%).

Table 2. Psychological symptoms scores of health care workers (N=151)[a].

Clinical scores	Resilient group (n=121), mean (SD)	Moderate-severe burnout group (n=54), mean (SD)	Subthreshold-provisional PTSD[b] group (n=31), mean (SD)	Burnout+PTSD group (n=52), mean (SD)
PCL-5[c] total score	6.97 (6.34)	10.37 (7.01)	29.03 (7.71)	34.73 (8.66)
PCL-5 cluster B: re-experiencing	2.07 (2.56)	2.30 (2.07)	7.06 (3.74)	7.77 (3.71)
PCL-5 cluster C: avoidance	0.85 (1.25)	1.15 (1.42)	4.06 (2.00)	3.81 (2.11)
PCL-5 cluster D: negative mood and cognition	2.11 (2.50)	3.94 (3.59)	9.10 (3.93)	13.21 (4.73)
PCL-5 cluster E: hyperarousal	1.93 (2.05)	2.98 (2.84)	8.81 (4.21)	9.94 (3.71)
MBI-9[d]: emotional exhaustion	6.63 (3.70)	12.31 (3.12)	10.03 (3.40)	14.37 (2.21)
MBI-9: depersonalization	3.25 (2.66)	9.78 (2.73)	4.61 (3.16)	11.00 (3.82)
MBI-9: personal accomplishment	14.34 (2.93)	12.70 (2.90)	14.42 (2.46)	12.27 (3.09)

[a]Continuous variables are presented as mean (SD).

[b]PTSD: posttraumatic stress disorder.

[c]PCL-5: PTSD Checklist for Diagnostic and Statistical Manual, 5th Edition.

[d]MBI-9: Maslach Burnout Inventory General Survey 9-item.

Microbehavior Intervals Characteristics

Across all recordings (N=258; 151 participants), the anomaly detection model (MOMENT) identified an average of 19.65 (SD 6.01) microbehavior intervals per interview (range 5-39). The average duration of these intervals was 1.31 (SD 1.10) seconds, ranging from 0.5-15.98 seconds. The first microbehavior interval occurred, on average, at 71.57 (SD 84.35) seconds into the interview, with the earliest detected at 0.68 seconds and the latest at 496.36 seconds. The distribution of key microbehavior interval characteristics is presented in Figure S2 of the Multimedia Appendix 1. See Tables S2-S4 in the Multimedia Appendix 1 for the detailed temporal features list.

Classification Model Evaluation

For the hyperparameter search space and final configuration, refer to Table S5 in the Multimedia Appendix 1. Training performance results of the deep learning classifier are shown in Figure S3, illustrating its convergence behavior and macro–F_1-score performance. To avoid overfitting, early stopping was implemented via Optuna's "MedianPruner" during hyperparameter optimization. Within each cross-validation fold,

XSL•FO
RenderX

the training data were split into internal training and validation subsets, and the cross-entropy loss was reported to Optuna at each epoch as an intermediate value. Trials were terminated early when the current trial's validation loss was worse than the median of completed trials at the same epoch. Pruning decisions were based only on the internal validation data within each fold; the held-out fold was used exclusively for final evaluation.

Our classifier successfully predicted the 4 symptom classes on both the training and the test splits (Table S6 in the Multimedia Appendix 1). Weighted performance metrics on the test set were robust, with a precision of 0.81, recall of 0.79, accuracy of 78.3%, and a macro– F_1-score of 0.75.

Figure 2 presents the model performance scores for the 4 symptom classes. The classifier demonstrated high discriminative ability for identifying participants in the "subthreshold-provisional PTSD" (area under the curve [AUC]=0.869), "moderate-severe burnout" (AUC=0.825) classes, and "burnout+PTSD" class (AUC=0.798). In contrast, identifying participants in the "resilient" class was less discriminatory, with a lower area under the receiver operating characteristic curve score of 0.686. Macroaverage and microaverage AUCs were 0.793 and 0.798, respectively, indicating consistent performance across classes and balanced treatment of minority labels.

Additionally, to assess the contribution of each modality, we conducted an ablation study by systematically removing 1 modality at a time from the complete multimodal framework. Excluding gaze or arousal-valence time-series signals resulted in the most significant performance drops, particularly in recall and F_1-score. In contrast, 3D hands and body pose contributed less individually, although their removal still resulted in measurable degradation (Figure S4 in the Multimedia Appendix 1).

Figure 2. Model performance across the 4 psychological comorbidity profiles. (A) One-vs-rest receiver operating characteristic (ROC) curves on the independent held-out test set and (B) row-normalized confusion matrix for the held-out test set, with counts shown within each cell. AUC: area under the curve; M-S: moderate-severe; PTSD: posttraumatic stress disorder; S-P: subthreshold-provisional.

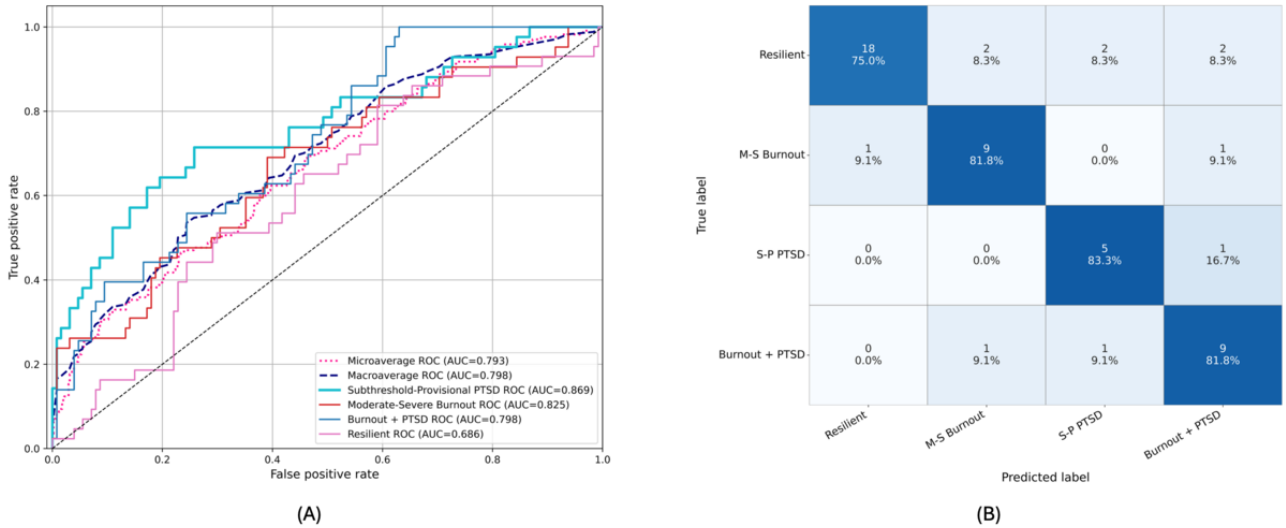

Feature Importance Analysis

The SHAP analysis presented in Figure 3 visualizes the contribution of the top 15 most influential features per class, ranked by their average impact on the model's output. Across all classes, features encoding variability (ie, duration interquartile and gap interquartile) and timing (ie, time of first interval, early intervals count, and late intervals count) frequently emerged as key drivers of model predictions.

For the "subthreshold-provisional PTSD" class, SHAP analysis revealed that the class was characterized by elevated variability and fragmentation in their behavioral dynamics, indicating frequent and short bursts of microbehavior intervals.

For the "moderate-severe burnout" class, the most influential features included an early onset of microbehavior intervals and prolonged durations of intervals. The model also highlighted measures of temporal irregularity and interval frequency, suggesting that this group was characterized by early activation and sustained, uneven patterns of microbehavior intervals across the interview.

Participants in the "burnout+PTSD" group showed distinctive patterns of both volatility and persistence in microbehavior intervals. The model was mostly influenced by high variability in microbehavior duration, as well as the timing and frequency of intervals across the interview. Both early and late interval counts, along with greater interval density and temporal dispersion, contributed strongly to classification. These findings suggest that individuals with combined burnout and PTSD display fragmented and irregular patterns of microbehavior intervals throughout the interview.

The SHAP analysis revealed a markedly different feature profile for the "resilient" class. These participants exhibited a more heterogeneous profile, characterized by fewer microbehavior intervals overall and less fluctuation in timing and duration compared to the symptomatic groups.

XSL•FO
RenderX

Figure 3. SHAP (Shapley additive explanations) summary plots illustrating the top features contributing to classification decisions for each psychological comorbidity profile. SHAP values represent the impact of each feature on model output. Color encodes feature magnitude (high=pink and low=blue). (A) Subthreshold-provisional posttraumatic stress disorder, (B) moderate-severe burnout, (C) resilient, and (D) moderate-severe burnout. PTSD: posttraumatic stress disorder.

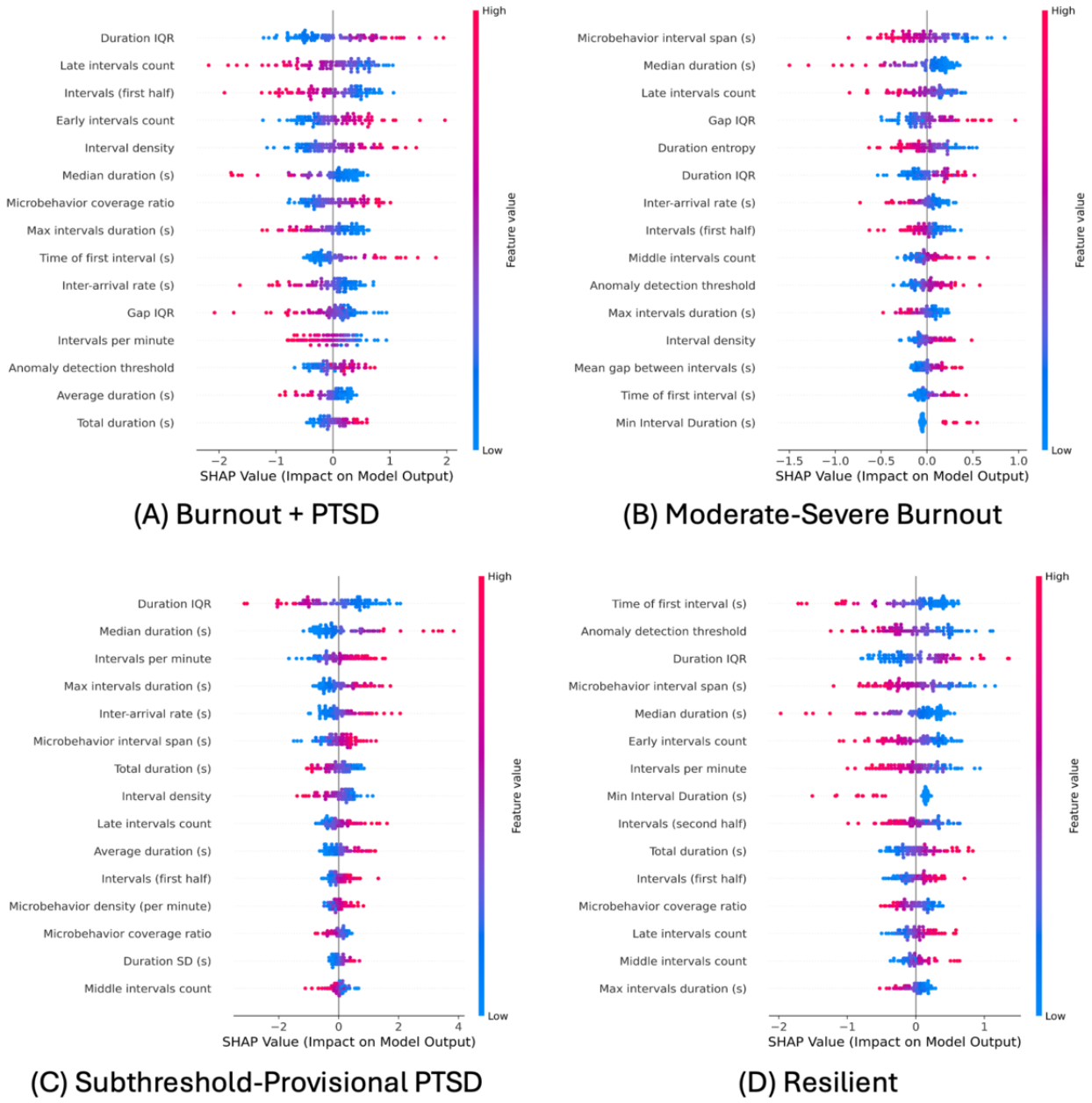

Discussion

We present a novel multimodal framework for identifying psychological distress related to burnout and PTSD in HCWs by modeling brief, spontaneous nonverbal fluctuations, termed microbehaviors, as objective digital biomarkers from video-recorded interviews. By analyzing multimodal nonverbal signals from emotion-inducing interviews, we identified 4 clinically meaningful classes: “resilient,” “moderate-severe burnout,” “subthreshold-provisional PTSD,” and “burnout+PTSD.”

The classifier demonstrated strong overall discrimination, with particularly reliable separation of participants with PTSD symptoms, burnout symptoms, or a comorbid symptom profile. Temporal variability, irregularity, and onset of microbehaviors were the most informative predictors, highlighting distinct expressive profiles across classes. Ablation experiments further demonstrated the complementary contribution of modalities, underscoring the benefit of combining large-scale posture and movement signals (pose-derived gross motor activity) with subtle high-resolution cues (gaze dynamics and facial expression) for more accurate mental health assessment. Our findings underscore the potential to broaden the analytical focus from isolated facial microexpressions or narrative content to a wider spectrum of high-resolution, multimodal nonverbal behaviors.

Prior multimodal and behavioral screening studies using nonverbal signals have generally reported moderate-to-strong discrimination, most often in simpler binary case-control settings. For example, Stratou et al [33] demonstrated that

XSL•FO
RenderX

automated nonverbal indicators can support prescreening for both depression and PTSD, with cross-validated F_1-scores of approximately 0.71 for higher-severity classification, consistent with the challenges of subtle symptom detection from behavior alone. In a closely related trauma-exposed cohort, Schultebraucks et al [24] combined visual and auditory markers of arousal and mood and reported strong binary classification for PTSD (AUC=0.90; weighted F_1-score=0.83) and depression (AUC=0.86; weighted F_1-score=0.82), underscoring the potential of audiovisual markers while also highlighting how headline metrics shift with task formulation and label granularity. Complementing this, Stolicyn et al [25] showed that combining ecological face-tracking and eye-tracking features improved depression prediction, achieving a maximum accuracy of 79%, supporting our finding that gaze and affective or arousal cues contribute meaningfully to discrimination. In contrast, speech-only PTSD studies in curated binary datasets report very high AUC and accuracy [39], highlighting that direct comparison across modalities and study designs can vary substantially. Finally, the microexpression approach by Li et al [21] also achieved a high accuracy in controlled settings (accuracy ≈0.873) but relies on highly strict acquisition constraints than standard webcam interviews, motivating the practical value of our novel methodology in a real-world screening context. Taken together, our performance is broadly consistent with the range reported for scalable behavioral screening approaches, while extending prior work by emphasizing temporally localized microbehavior intervals and multiprofile differentiation rather than binary symptom detection.

Distinct microbehavior signatures were observed in both burnout and PTSD classes, thus supporting the hypothesis that dysregulated expressive dynamics reflect underlying psychological distress [60,61]. For example, we hypothesized that the fragmented and frequent microbehavior intervals in the "subthreshold-provisional PTSD" group may indicate heightened physiological reactivity and unstable emotional regulation, mirroring hallmark PTSD symptoms such as hyperarousal and intrusive re-experiencing [62], where individuals repeatedly relive trauma-related cues through abrupt, involuntary responses. The early onset and prolonged microbehavior intervals in the "moderate-severe burnout" group may suggest a diminished capacity to regulate or recover from affective activation. This could potentially align with sustained emotional exhaustion and blunted affective engagement, which are characteristic of severe burnout [63]. The combined "burnout+PTSD" group exhibited volatility and wide dispersion of microbehavior intervals across the interview. We postulate that this pattern could be consistent with the emotional exhaustion of burnout [63], coupled with the reactive instability and heightened sensitivity to triggers typical of PTSD [62].

Conversely, classification performance was lower for the "resilient" group. SHAP analysis revealed that these individuals exhibited fewer distinctive microbehavior patterns, which may account for the reduced discriminative accuracy. Rather than exhibiting a consistent temporal signature, their SHAP distribution suggested heterogeneous dynamics, reflecting the multifaceted nature of resilience. This aligns with prior research showing that resilience is not defined solely by the absence of arousal- or mood-related markers, and is often harder to discriminate than symptomatic patients [64]. As such, nonverbal behavioral signals alone may not fully capture resilience, underscoring the need for an even more comprehensive methodological approach. Practically, these findings position our framework as a screening tool for identifying elevated distress rather than a tool to "confirm" resilience.

From a clinical perspective, these findings highlight the potential of microbehavior analysis as a scalable and unobtrusive complement to standard clinical interviews and psychometric questionnaires. Our framework operates without manual annotation, is language-agnostic, and requires only short, semistructured interviews, making it particularly suitable for frontline professionals who may minimize or suppress verbal disclosure of distress [65]. Importantly, the use of anomaly-based modeling enables the detection of behavior deviations that are imperceptible to the human eye, providing clinicians with objective cues that may precede or corroborate self-reported symptoms. Nevertheless, as with any digital mental health screening approach, ethical considerations (eg, data privacy and algorithmic bias) must also be addressed when deploying such systems in real-world settings.

Beyond classification, the ability to analyze multimodal signals at high temporal resolution makes this framework a potential tool for discovering latent behavioral phenotypes across various mental health conditions. Future work should extend these findings by clarifying which microbehavioral dynamics map onto specific symptom domains, how they evolve over time, and how they manifest in comorbid presentations.

Several limitations warrant consideration, many of which reflect deliberate design trade-offs aimed at prioritizing ecological validity and feasibility within realistic telehealth constraints. First, our cohort consisted exclusively of HCWs. This increases the ecological validity in this population and aligns with our intended screening use case but may also limit generalizability to other trauma-exposed groups such as military personnel or first responders. Second, the framework is sensitive to video quality and recording conditions that vary substantially in real-world telehealth settings (eg, camera angle, motion blur, occlusions, and intermittent connectivity). These factors can introduce noise and missingness of different modalities and may affect the stability of anomaly scores and the resulting interval segmentation. Although our pipeline retains such frames and propagates missing values to downstream aggregation, performance and interpretability are expected to degrade when visibility is poor, or tracking failures are frequent; future work should quantify quality thresholds and incorporate explicit quality-aware confidence measures. Third, several pipeline components rely on fixed preprocessing and segmentation choices. These parameters were chosen based on empirical inspection to reduce overfragmentation and to robustly handle transient tracking noise. Still, they can influence the number, duration, and temporal distribution of detected intervals, thereby affecting derived interval-level statistics. Although the central conclusion that temporally localized expressive episodes contain a discriminative signal is supported by the overall approach, future studies should validate parameter sensitivity across

XSL•FO
RenderX

settings and explore principled calibration strategies that improve cross-recording comparability. Fourth, our classification operates on aggregated interval-level features. Although this design choice improves interpretability and scalability, it may discard within-interval temporal structure and assume that summary statistics sufficiently capture clinically relevant dynamics. Additionally, although extracting features from aggregated microbehavior intervals emphasizes group-level regularities, they may also limit personalization.

Finally, we intentionally focused on nonverbal modalities to reduce linguistic and fluency-related biases and to target a broadly deployable telehealth workflow; however, this decision may restrict the interpretability and reliability of the detected patterns. Incorporating questions timing and the semantic content of the participant's response could potentially enhance the diagnostic relevance of these patterns.

In conclusion, this study introduces a multimodal, anomaly-based framework for identifying psychological distress in HCWs through the analysis of microbehaviors—brief, involuntary fluctuations in nonverbal expressions. By providing objective, annotation-free biomarkers that are sensitive to subtle behavioral changes, the framework holds promise as a complementary tool that enhances traditional mental health assessments.

Acknowledgments

The authors thank Joseph Lyu, Joseph Chang, and Areli Vicuna for their assistance with data collection. The authors declare that no generative artificial intelligence was used in the research or writing process.

Funding

This work was supported by the US National Heart, Lung, and Blood Institute (grant R01HL156134). KS received additional support from the US National Institute of Mental Health (grant R01MH129856). The funders had no role in the study design, data collection, analysis, data interpretation, or writing of the manuscript.

Data Availability

All requests for raw and analyzed data, as well as related materials, will be reviewed by our legal department at New York University Grossman School of Medicine to verify whether the request is subject to any intellectual property or confidentiality constraints. Any data and materials that can be shared will be released via a material transfer agreement for noncommercial research purposes. The request should be addressed to KS. All requests for programming code should be addressed to KS. Any materials that can be shared will be released via a material transfer agreement for noncommercial research purposes.

Authors' Contributions

KS was responsible for the conceptualization and overall supervision of the study and securing the funding. SG led the formal analysis, methodology development, and visualization, and drafted the original manuscript. SG also performed validation and final editing under the guidance of KS and CEH, who provided critical feedback and contributed to the review and editing process. SBB conducted data curation and was responsible for data collection and project administration. YZ assisted with data analysis. VM and SPW provided revisions and contributed to the review and editing of the manuscript. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional details about the sample.
[DOCX File , 1011 KB - jmir_v28i1e87049_app1.docx]

References

1. Pappa S, Ntella V, Giannakas T, Giannakoulis VG, Papoutsi E, Katsaounou P. Prevalence of depression, anxiety, and insomnia among healthcare workers during the COVID-19 pandemic: a systematic review and meta-analysis. Brain Behav Immun 2020 Aug;88:901-907 [FREE Full text] [doi: 10.1016/j.bbi.2020.05.026] [Medline: 32437915]
2. Fear NT, Jones M, Murphy D, Hull L, Iversen AC, Coker B, et al. What are the consequences of deployment to Iraq and Afghanistan on the mental health of the UK armed forces? A cohort study. Lancet 2010 May 22;375(9728):1783-1797. [doi: 10.1016/S0140-6736(10)60672-1] [Medline: 20471076]
3. Haugen PT, McCrillis AM, Smid GE, Nijdam MJ. Mental health stigma and barriers to mental health care for first responders: a systematic review and meta-analysis. J Psychiatr Res 2017 Nov;94:218-229. [doi: 10.1016/j.jpsychires.2017.08.001] [Medline: 28800529]

XSL•FO
RenderX

4. Norman SB, Feingold JH, Kaye-Kauderer H, Kaplan CA, Hurtado A, Kachadourian L, et al. Moral distress in frontline healthcare workers in the initial epicenter of the COVID-19 pandemic in the United States: relationship to PTSD symptoms, burnout, and psychosocial functioning. Depress Anxiety 2021 Oct;38(10):1007-1017 [FREE Full text] [doi: 10.1002/da.23205] [Medline: 34293236]
5. Cieslak R, Shoji K, Douglas A, Melville E, Luszczynska A, Benight CC. A meta-analysis of the relationship between job burnout and secondary traumatic stress among workers with indirect exposure to trauma. Psychol Serv 2014 Feb;11(1):75-86. [doi: 10.1037/a0033798] [Medline: 23937082]
6. Mealer M, Burnham EL, Goode CJ, Rothbaum B, Moss M. The prevalence and impact of post traumatic stress disorder and burnout syndrome in nurses. Depress Anxiety 2009;26(12):1118-1126 [FREE Full text] [doi: 10.1002/da.20631] [Medline: 19918928]
7. Maslach C, Schaufeli WB, Leiter MP. Job burnout. Annu Rev Psychol 2001;52:397-422. [doi: 10.1146/annurev.psych.52.1.397] [Medline: 11148311]
8. Woo T, Ho R, Tang A, Tam W. Global prevalence of burnout symptoms among nurses: a systematic review and meta-analysis. J Psychiatr Res 2020 Apr;123:9-20. [doi: 10.1016/j.jpsychires.2019.12.015] [Medline: 32007680]
9. Shen X, Xu H, Feng J, Ye J, Lu Z, Gan Y. The global prevalence of burnout among general practitioners: a systematic review and meta-analysis. Fam Pract 2022 Sep 24;39(5):943-950. [doi: 10.1093/fampra/cmab180] [Medline: 35089320]
10. Chirico F, Afolabi AA, Ilesanmi OS, Nucera G, Ferrari G, Sacco A, et al. Prevalence, risk factors and prevention of burnout syndrome among healthcare workers: an umbrella review of systematic reviews and meta-analyses. J Health Soc Sci 2021 Dec;6(4):465-491. [doi: 10.19204/2021/prvl3]
11. Shalev A, Liberzon I, Marmar C. Post-traumatic stress disorder. N Engl J Med 2017 Jun 22;376(25):2459-2469. [doi: 10.1056/NEJMra1612499] [Medline: 28636846]
12. D'Ettorre G, Pellicani V, Ceccarelli G. Post-traumatic stress disorder symptoms in healthcare workers: a ten-year systematic review. Acta Biomed 2020 Nov 30;91(12-S):e2020009 [FREE Full text] [doi: 10.23750/abm.v91i12-S.9459] [Medline: 33263341]
13. Hill JE, Harris C, Danielle L C, Boland P, Doherty AJ, Benedetto V, et al. The prevalence of mental health conditions in healthcare workers during and after a pandemic: systematic review and meta-analysis. J Adv Nurs 2022 Jun;78(6):1551-1573 [FREE Full text] [doi: 10.1111/jan.15175] [Medline: 35150151]
14. Greene-Shortridge TM, Britt TW, Castro CA. The stigma of mental health problems in the military. Mil Med 2007 Feb;172(2):157-161. [doi: 10.7205/milmed.172.2.157] [Medline: 17357770]
15. Gold KJ, Andrew LB, Goldman EB, Schwenk TL. "I would never want to have a mental health diagnosis on my record": a survey of female physicians on mental health diagnosis, treatment, and reporting. Gen Hosp Psychiatry 2016;43:51-57. [doi: 10.1016/j.genhosppsych.2016.09.004] [Medline: 27796258]
16. Chen X, Luo T. Catching elusive depression via facial micro-expression recognition. IEEE Commun Mag 2023 Oct;61(10):30-36. [doi: 10.1109/mcom.001.2300003]
17. Ekman P, Friesen WV. Nonverbal leakage and clues to deception. Psychiatry 1969 Feb;32(1):88-106. [doi: 10.1080/00332747.1969.11023575] [Medline: 5779090]
18. Ekman P. Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life. New York, NY: Times Books/Henry Holt and Co; 2003.
19. Pfister T, Li X, Zhao G, Pietikainen M. Recognising spontaneous facial micro-expressions. In: Proceedings of the 2011 International Conference on Computer Vision. 2011 Presented at: ICCV 2011; Nov 6-13, 2011; Barcelona, Spain. [doi: 10.1109/iccv.2011.6126401]
20. Zhao G, Li X, Li Y, Pietikäinen M. Facial micro-expressions: an overview. Proc IEEE 2023 Oct;111(10):1215-1235. [doi: 10.1109/jproc.2023.3275192]
21. Li X, Yi X, Ye J, Zheng Y, Wang Q. SFTNet: a microexpression-based method for depression detection. Comput Methods Programs Biomed 2024 Jan;243:107923. [doi: 10.1016/j.cmpb.2023.107923] [Medline: 37989077]
22. Shatte AB, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. Psychol Med 2019 Feb 12;49(9):1426-1448. [doi: 10.1017/s0033291719000151]
23. Zhernova P, Bodyanskiy Y, Yatsenko B, Zavgorodnii I. Detection and prevention of professional burnout using machine learning methods. In: Proceedings of the 2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering. 2020 Presented at: TCSET 2020; Feb 25-29, 2020; Lviv-Slavske, Ukraine. [doi: 10.1109/TCSET49122.2020.235426]
24. Schultebraucks K, Yadav V, Shalev AY, Bonanno GA, Galatzer-Levy IR. Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood. Psychol Med 2020 Aug 03;52(5):957-967. [doi: 10.1017/s0033291720002718]
25. Stolicyn A, Steele JD, Seriès P. Prediction of depression symptoms in individual subjects with face and eye movement tracking. Psychol Med 2020 Nov 09;52(9):1784-1792. [doi: 10.1017/s0033291720003608]
26. Ding Z, Wang Z, Zhang Y, Cao Y, Liu Y, Shen X, et al. Trade-offs between machine learning and deep learning for mental illness detection on social media. Sci Rep 2025 Apr 25;15(1):14497 [FREE Full text] [doi: 10.1038/s41598-025-99167-6] [Medline: 40281061]

XSL•FO
RenderX

27. Yan WJ, Wu Q, Liang J, Chen YH, Fu X. How fast are the leaked facial expressions: the duration of micro-expressions. J Nonverbal Behav 2013 Jul 16;37:217-230. [doi: 10.1007/s10919-013-0159-8]
28. Singh NN, McKay JD, Singh AN. Culture and mental health: nonverbal communication. J Child Fam Stud 1998 Dec;7:403-409. [doi: 10.1023/a:1022946925134]
29. Tarik B, Ashour S, Gomaa W. Mental health assessment: analyzing body language patterns and emotional expression. In: Proceedings of the 2024 Intelligent Methods, Systems, and Applications. 2024 Presented at: IMSA 2024; Jul 13-14, 2024; Giza, Egypt. [doi: 10.1109/imsa61967.2024.10652831]
30. Ringwald WR, Feltman S, Schwartz HA, Samaras D, Khudari C, Luft BJ, et al. Day-to-day dynamics of facial emotion expressions in posttraumatic stress disorder. J Affect Disord 2025 Jul 01;380:331-339. [doi: 10.1016/j.jad.2025.03.109] [Medline: 40122249]
31. Coll SY, Eustache F, Doidy F, Fraisse F, Peschanski D, Dayan J, et al. Avoidance behaviour generalizes to eye processing in posttraumatic stress disorder. Eur J Psychotraumatol 2022 Apr 20;13(1):2044661 [FREE Full text] [doi: 10.1080/20008198.2022.2044661] [Medline: 35479300]
32. Bianchi R, Laurent E. Emotional information processing in depression and burnout: an eye-tracking study. Eur Arch Psychiatry Clin Neurosci 2015 Feb;265(1):27-34. [doi: 10.1007/s00406-014-0549-x] [Medline: 25297694]
33. Stratou G, Scherer S, Gratch J, Morency LP. Automatic nonverbal behavior indicators of depression and PTSD: exploring gender differences. In: Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. 2013 Presented at: ACII 2013; Sep 2-5, 2013; Geneva, Switzerland. [doi: 10.1109/acii.2013.31]
34. Mitani S, Fujita M, Nakata K, Shirakawa T. Impact of post-traumatic stress disorder and job-related stress on burnout: a study of fire service workers. J Emerg Med 2006 Jul;31(1):7-11. [doi: 10.1016/j.jemermed.2005.08.008] [Medline: 16798146]
35. Restauri N, Sheridan AD. Burnout and posttraumatic stress disorder in the coronavirus disease 2019 (COVID-19) pandemic: intersection, impact, and interventions. J Am Coll Radiol 2020 Jul;17(7):921-926 [FREE Full text] [doi: 10.1016/j.jacr.2020.05.021] [Medline: 32479798]
36. Blázquez-García A, Conde A, Mori U, Lozano JA. A review on outlier/anomaly detection in time series data. ACM Comput Surv 2021 Apr 17;54(3):1-33. [doi: 10.1145/3444690]
37. Raman N, Cao M, Tsvetkov Y, Kästner C, Vasilescu B. Stress and burnout in open source: toward finding, understanding, and mitigating unhealthy interactions. In: Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: New Ideas and Emerging Results. 2020 Presented at: ICSE-NIER '20; Jun 27-Jul 19, 2020; Seoul, South Korea. [doi: 10.1145/3377816.3381732]
38. Baxter SL, Saseendrakumar BR, Cheung M, Savides TJ, Longhurst CA, Sinsky CA, et al. Association of electronic health record inbasket message characteristics with physician burnout. JAMA Netw Open 2022 Nov 01;5(11):e2244363 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.44363] [Medline: 36449288]
39. Hu J, Zhao C, Shi C, Zhao Z, Ren Z. Speech-based recognition and estimating severity of PTSD using machine learning. J Affect Disord 2024 Oct 01;362:859-868. [doi: 10.1016/j.jad.2024.07.015] [Medline: 39009320]
40. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. Speech Commun 2015 Jul;71:10-49. [doi: 10.1016/j.specom.2015.03.004]
41. Kring AM, Stuart BK. Nonverbal behavior and psychopathology. In: Harrigan JA, Rosenthal R, Scherer KR, editors. The New Handbook of Methods in Nonverbal Behavior Research. Oxford, UK: Oxford University Press; 2005:313-339.
42. Lavelle M, Healey PG, McCabe R. Nonverbal behavior during face-to-face social interaction in schizophrenia: a review. J Nerv Ment Dis 2014 Jan;202(1):47-54. [doi: 10.1097/NMD.0000000000000031] [Medline: 24375212]
43. Berry DS, Pennebaker JW. Nonverbal and verbal emotional expression and health. Psychother Psychosom 1993;59(1):11-19. [doi: 10.1159/000288640] [Medline: 8441791]
44. Code of Federal Regulations. URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46 [accessed 2026-05-21]
45. Leiter MP, Shaughnessy K. The areas of worklife model of burnout: tests of mediation relationships. Ergonomia 2006;28(4):327-341 [FREE Full text]
46. Rotenstein LS, Torre M, Ramos MA, Rosales RC, Guille C, Sen S, et al. Prevalence of burnout among physicians: a systematic review. JAMA 2018 Sep 18;320(11):1131-1150 [FREE Full text] [doi: 10.1001/jama.2018.12777] [Medline: 30326495]
47. Blevins CA, Weathers FW, Davis MT, Witte TK, Domino JL. The Posttraumatic Stress Disorder Checklist for DSM-5 (PCL-5): development and initial psychometric evaluation. J Trauma Stress 2015 Dec;28(6):489-498. [doi: 10.1002/jts.22059] [Medline: 26606250]
48. Klein AB, Schnurr PP, Bovin MJ, Friedman MJ, Keane TM, Marx BP. An empirical investigation of definitions of subthreshold posttraumatic stress disorder. J Trauma Stress 2024 Feb;37(1):113-125. [doi: 10.1002/jts.22987] [Medline: 37937636]
49. Biscoe N, Campbell GM, Murphy D. Beyond the threshold: considering and exploring subthreshold PTSD and co-morbid mental health in a clinical military sample. Eur J Trauma Dissoc 2025 Mar;9(1):100520. [doi: 10.1016/j.ejtd.2025.100520]
50. No authors listed. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024 Apr 18;385:q902 [FREE Full text] [doi: 10.1136/bmj.q902] [Medline: 38636956]

XSL•FO
RenderX

51. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021 Jul 09;11(7):e048008 [FREE Full text] [doi: 10.1136/bmjopen-2020-048008] [Medline: 34244270]
52. Facetorch. GitHub. URL: https://github.com/tomas-gajarsky/facetorch [accessed 2026-05-13]
53. Ververas E, Gkagkos P, Deng J, Doukas MC, Guo J, Zafeiriou S. 3DGazeNet: generalizing 3D gaze estimation with weak-supervision from synthetic views. In: Proceedings of the 18th European Conference on Computer Vision. 2024 Presented at: ECCV 2024; Sep 29-Oct 4, 2024; Milan, Italy. [doi: 10.1007/978-3-031-72664-4_22]
54. Dwivedi S, Sun Y, Patel P, Feng Y, Black MJ. TokenHMR: advancing human mesh recovery with a tokenized pose representation. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024 Presented at: CVPR 2024; Jun 16-22, 2024; Seattle, WA. [doi: 10.1109/cvpr52733.2024.00132]
55. Potamias RA, Zhang J, Deng J, Zafeiriou S. WiLoR: end-to-end 3D hand localization and reconstruction in-the-wild. In: Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2025 Presented at: CVPR 2025; Jun 10-17, 2025; Nashville, TN. [doi: 10.1109/cvpr52734.2025.01143]
56. Goswami M, Szafer K, Choudhry A, Cai Y, Li S, Dubrawski A. MOMENT: a family of open time-series foundation models. Proceedings of the 41st International Conference on Machine Learning 2024:16115-16152 [FREE Full text]
57. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019 Presented at: KDD '19; Aug 4-8, 2019; Anchorage, AK. [doi: 10.1145/3292500.3330701]
58. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002 Jun 01;16:321-357. [doi: 10.1613/jair.953]
59. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS'17; Dec 4-9, 2017; Long Beach, CA URL: https://papers.nips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
60. Cole PM, Michel MK, Teti LO. The development of emotion regulation and dysregulation: a clinical perspective. Monogr Soc Res Child Dev 1994;59(2/3):73-100. [doi: 10.2307/1166139]
61. D'Agostino A, Covanti S, Rossi Monti M, Starcevic V. Reconsidering emotion dysregulation. Psychiatr Q 2017 Dec;88(4):807-825. [doi: 10.1007/s11126-017-9499-6] [Medline: 28194549]
62. Yufik T, Simms LJ. A meta-analytic investigation of the structure of posttraumatic stress disorder symptoms. J Abnorm Psychol 2010 Nov;119(4):764-776 [FREE Full text] [doi: 10.1037/a0020981] [Medline: 21090877]
63. Maslach C, Leiter MP. Burnout. In: Fink G, editor. Stress: Concepts, Cognition, Emotion, and Behavior. New York, NY: Academic Press; 2016.
64. Schultebraucks K, Choi KW, Galatzer-Levy IR, Bonanno GA. Discriminating heterogeneous trajectories of resilience and depression after major life stressors using polygenic scores. JAMA Psychiatry 2021 Jul 01;78(7):744-752 [FREE Full text] [doi: 10.1001/jamapsychiatry.2021.0228] [Medline: 33787853]
65. Waegemakers Schiff J, Lane AM. PTSD symptoms, vicarious traumatization, and burnout in front line workers in the homeless sector. Community Ment Health J 2019 Apr;55(3):454-462. [doi: 10.1007/s10597-018-00364-7] [Medline: 30684127]

Abbreviations

AUC: area under the curve
HCW: health care worker
HIPAA: Health Insurance Portability and Accountability Act
MOMENT: a Family of Open Time-Series Foundation Models
PTSD: posttraumatic stress disorder
SHAP: Shapley Additive Explanations

Edited by A Stone; submitted 03.Nov.2025; peer-reviewed by R Richer, S Mirji; comments to author 05.Feb.2026; accepted 17.Mar.2026; published 27.May.2026.

Please cite as:
Gershov S, Hilberdink CE, Zhao Y, Birnbaum SB, Mueller V, Wall SP, Schultebraucks K
Detection of Microbehavior Intervals for Predicting Mental Health: Clinically Relevant and Advanced Multimodal Temporal Analysis
J Med Internet Res 2026;28:e87049
URL: https://www.jmir.org/2026/1/e87049
doi:10.2196/87049
PMID:

XSL•FO
RenderX

©Sapir Gershov, Charlotte E Hilberdink, Yiwen Zhao, Sarah B Birnbaum, Victoria Mueller, Stephen P Wall, Katharina Schultebraucks. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Illness Narratives of Young Chinese Adults With Diabetes on RedNote: Qualitative Narrative Analysis

Zikun Liu[1], PhD; Donghan Fu[2], PhD; Yingjie Liu[3], MA; Ying Meng[1], BA

[1]School of Journalism and Communication, Wuhan University, 229 Bayi Road, Wuhan, China

[2]School of Information Science and Learning Technologies, University of Missouri-Columbia, Columbia, MO, United States

[3]School of Media Science, Northeast Normal University, Changchun, China

Corresponding Author:
Zikun Liu, PhD
School of Journalism and Communication, Wuhan University, 229 Bayi Road, Wuhan, China

Abstract

Background: Chronic illness disrupts everyday routines, social roles, and sense of self, particularly among young individuals undergoing identity formation. With the expansion of digital media, social platforms have become key sites where patients narrate illness experiences, negotiate stigma, and seek support. However, such processes remain underexplored in non-Western, collectivist contexts.

Objective: This study examines how young Chinese individuals with diabetes construct illness narratives and negotiate identity in digital environments.

Methods: This study uses a narrative analysis approach, combining inductive thematic coding with culturally and critically informed interpretation. A total of 303 narrative posts were collected from RedNote, a Chinese social media platform characterized by diary-like user-generated content. The dataset includes both text-based and video-based posts, capturing longitudinal and first-person accounts of living with diabetes.

Results: In total, 4 distinct narrative types were identified. The chaos narrative captures experiences of cognitive dissonance, emotional breakdown, and disruption of daily routines following diagnosis, often accompanied by guilt toward family members and anxiety over future uncertainty. The stigma narrative reflects social withdrawal, concealment of illness, and perceived discrimination in intimate relationships and employment contexts, highlighting the role of externally imposed social judgment. The resilience narrative illustrates processes of self-acceptance, disciplined self-management, and the integration of illness into everyday life through routinized practices such as blood glucose monitoring and dietary regulation. The solidarity narrative emphasizes the importance of familial care and digitally mediated peer support, where users exchange practical knowledge, emotional encouragement, and collective identity markers, transforming isolation into shared experience. Across these narratives, illness is not only experienced as disruption but also actively reinterpreted through culturally embedded values such as familial responsibility and collective belonging.

Conclusions: This study advances illness narrative research by demonstrating how digital platforms mediate culturally specific forms of meaning-making among young patients with chronic illness. It extends the concept of biographical disruption by conceptualizing it as a dynamic and relational process shaped by digital storytelling, familial expectations, and peer interaction. The findings highlight the importance of culturally sensitive and platform-aware approaches to health communication and digital patient support.

(J Med Internet Res 2026;28:e96543) doi:10.2196/96543

KEYWORDS

chronic illness; diabetes; social media; young people; illness narratives

Introduction

As one of the most prevalent chronic diseases globally, diabetes has increasingly eroded public health in China in recent years [1]. According to data from the International Diabetes Federation, the number of adult patients with diabetes in China continues to rise, reaching approximately 141 million by 2021, a 21.55% increase from that in 2019 [2]. Of particular concern is the growing impact of diabetes on young populations, as many struggle to manage their condition effectively despite rigorous treatment [3]. An epidemiological survey of diabetes in China revealed a prevalence of 2% among individuals aged 18 - 29 years, with 20.2% of this cohort classified as prediabetic [4].

Often termed a “silent killer,” diabetes exhibits low cure and control rates, with prolonged illness risking irreversible tissue

XSL•FO
RenderX

and organ damage, multiple complications, and lifelong dependence on medication or surgical interventions [5]. Protracted, unpredictable treatment processes frequently trigger anxiety, self-doubt, and depression, leading to high rates of diabetes-depression comorbidity that profoundly degrades the quality of life for patients and families [6]. For young individuals, traditionally idealized as the "rising sun" and "family's hope," diabetes interrupts not only personal aspirations but also familial expectations, imposing immense psychological and socioeconomic burdens when their health trajectories deviate from social norms [7]. However, the pervasive perception of diabetes as an "age-related disease" has led to systemic neglect in young patients [8].

Recent advancements in digital media have amplified the visibility of this marginalized group, enabling young patients with diabetes to document their illness journeys, share coping strategies, and forge peer support networks [9,10]. Through multidimensional narratives, young adults articulate the profound pressures of daily life, navigate emotional turmoil, and collaboratively explore pathways to self-healing and identity reconstruction [11]. As a result, digital narratives have emerged as key components of online health discourse, and scholars have examined how individuals share stories of health, illness, and caregiving across diverse digital platforms [11,12]. Patient-authored narratives provide first-hand insights into the lived experience of illness, capturing both the trauma of the disease and the evolution of self-identity [13]. Specifically, a recent study demonstrates that parents of children with diabetes in China experience identity disruption and family strain while using online narratives as culturally situated resources for reconstructing meaning under Confucian-Taoist sociocultural constraints [6]. Young adults with diabetes negotiated their biographical disruption by adopting shifting subject positions—concealing illness in public to preserve normality, asserting autonomy against medical control, and eventually developing a pragmatic acceptance that rendered diabetes a manageable part of everyday life [14].

More broadly, illness encompasses multidimensional experiences spanning the physiological, psychological, and social domains that evolve dynamically across life stages [15]. Within this framework, the concept of biographical disruption suggests that chronic illness destabilizes taken-for-granted assumptions about the self and the future, forcing individuals to reinterpret their identities and life trajectories [16]. This theoretical perspective highlights the importance of examining how these experiences are organized into narrative forms. However, existing research has not sufficiently examined how young adults themselves construct illness narratives in digital environments, nor how such narratives take shape as distinct forms of meaning-making. Addressing this gap requires closer attention to the narrative patterns through which young patients articulate their experiences and coping strategies. To address this gap, this study asks: What distinct types of illness narratives emerge from the experiences of young adults with diabetes? How do these narratives reflect the primary challenges faced by young patients and their coping strategies?

Methods

Study Design

This study uses a narrative analysis approach, combining inductive thematic coding with a culturally and critically informed interpretive framework to examine how illness narratives are constructed and negotiated in a digital context [17]. Rather than treating narratives as neutral reflections of experience, this approach understands them as culturally embedded meaning-making practices shaped by social norms, technological affordances, and broader discursive structures [18]. Initial coding was conducted following principles of reflexive thematic analysis as outlined by Braun and Clarke [19,20], while subsequent stages of analysis drew on illness narrative theory, particularly Frank's typology, as well as sociological perspectives on stigma, identity, and digital culture [21,22]. In terms of illness narrative types, Frank classified patient illness narratives into 3 categories: restitution, chaos, and quest [23]. This dual approach enabled both data-driven pattern identification and theoretically grounded interpretation.

Data Source and Sampling Strategy

Data were collected from RedNote, a Chinese social media platform characterized by user-generated lifestyle and health-related content [24]. The platform provides a space for young users to share personal experiences through text, images, and short videos, often in diary-like formats [3].

Given the exploratory and interpretive nature of this study, a purposive sampling strategy was used. Accounts were identified through iterative searching and long-term observation between March 2023 and April 2024, using a combination of diabetes-related keywords (eg, "diabetes," "insulin," and "blood sugar"). The sampling process prioritized accounts that met the following criteria: (1) self-identified young individuals (primarily university students) living with diabetes; (2) regularly posted first-person narratives related to illness experiences; (3) provided sufficiently detailed and longitudinal content for narrative analysis; and (4) represented noncommercial, nonpromotional personal accounts. Through this process, 13 accounts were selected as information-rich cases suitable for in-depth qualitative analysis. While the sample is not statistically representative, it is appropriate for narrative inquiry aimed at understanding meaning-making processes rather than population-level generalization. The final dataset comprised 303 narrative posts generated by 13 young adults living with diabetes, including 213 text-based posts and 90 video posts (Table 1). For video content, spoken narratives, captions, and on-screen text were transcribed verbatim, and relevant contextual cues (eg, tone, setting, and embodied expression) were documented to support interpretation.

XSL•FO
RenderX

Table . Detailed information of the selected 13 accounts.

	Sex	Followers, n	Posts	
			Text, n	Video, n
A1	Female	822	11	4
A2	Female	138	19	0
A3	Female	180	72	16
A4	Female	666	23	0
A5	Female	837	4	9
A6	Female	17	2	0
A7	Female	443	5	10
A8	Female	1661	12	11
A9	Female	209	38	3
A10	Female	583	2	9
A11	Male	713	2	12
A12	Female	9793	0	16
A13	Female	334	23	0

Data Analysis

The analytical process proceeded in 3 iterative stages, and all authors participated in the coding process [19]. First, inductive coding: all posts were read multiple times to achieve familiarization. Open coding was conducted by the research team (ZL, DF, and YL) to identify recurring patterns related to illness experience, emotional responses, social interactions, and self-management practices. Coding was primarily conducted in Chinese to preserve linguistic and cultural nuances. Second, thematic organization: initial codes were grouped into broader thematic categories through iterative comparison and discussion among the researchers (ZL, DF, and YL). At this stage, the analysis remained largely inductive, focusing on patterns emerging from the data rather than imposing predefined categories. Third, narrative interpretation and theoretical integration: in later stages, the analysis moved beyond descriptive themes toward interpretive narrative analysis. Drawing on the illness narrative typology of Frank [23], as well as sociological theories of stigma, identity, and digital culture, the research team (ZL and DF) examined how different narrative patterns functioned as culturally situated meaning-making strategies.

Through this process, 4 narrative types were identified: chaos, stigma, resilience, and solidarity. While partially informed by existing theoretical frameworks, these categories were refined inductively to reflect culturally specific experiences in the Chinese context. Key analytical concepts, such as familial responsibility and filial piety, emerged through iterative engagement between data and theory. Discrepancies in interpretation were addressed through collaborative discussion among the authors until analytical consensus was reached.

Reflexivity and Rigor

This study follows a reflexive approach to qualitative rigor, emphasizing transparency and interpretive awareness rather than positivist criteria such as intercoder reliability [25]. All authors are native Chinese speakers with cultural familiarity with the context under study, which facilitated nuanced interpretation but also required critical reflexivity. Reflexive journaling was maintained throughout the research process to document analytical decisions, emerging assumptions, and potential biases. These reflections informed ongoing interpretation rather than serving as a post hoc validation strategy. To enhance methodological rigor, the study engaged with criteria commonly used in qualitative research. Credibility was strengthened through repeated reading and iterative analysis of 303 narrative posts on RedNote, including both textual and video-based content, which enabled the refinement of emerging narrative patterns. Dependability was supported by documenting coding procedures, category development, and interpretive discussions throughout the analytical process. Confirmability was enhanced through reflexive discussions among the research team regarding how cultural familiarity and theoretical perspectives shaped interpretation, ensuring that findings remained grounded in the data. Transferability was addressed by providing contextual descriptions of the narratives, the experiences of young Chinese adults with diabetes, and the platform environment of RedNote.

Translation and Positionality

All analyses were conducted in Chinese to preserve the integrity of the original narratives. Selected excerpts were translated into English by the first author (ZL) and independently reviewed by a bilingual researcher (YM) to ensure conceptual equivalence. Discrepancies were resolved through discussion. The research team's shared linguistic and cultural background provided an insider perspective, which was both an analytical strength and a potential source of bias. Reflexive practices were therefore integral to the interpretive process.

XSL•FO
RenderX

Ethical Considerations

This study analyzed publicly available social media content and did not involve direct interaction with users. In accordance with institutional guidelines for internet research, formal ethics board approval was not required [26]. However, given the sensitive nature of health-related narratives, several measures were taken to minimize potential harm following procedures adopted in previous research [27]. All user accounts were anonymized (A1-A13), and identifying details were removed or modified. Data were reported in aggregate form where possible, and excerpts were selected to avoid traceability through search engines. The study adhered to the terms of service of the RedNote platform and followed established ethical guidelines for research using publicly accessible online data. There were no human participants in this study, and informed consent was not required.

Results

Overview

We identified 4 distinct illness narratives among young adults with diabetes: chaos, stigma, resilience, and solidarity. While chaos and stigma narratives illustrate how these patients navigate a complex biopsychosocial landscape in which chronic illness disrupts identity, autonomy, and social integration, resilience and solidarity narratives highlight how individuals actively develop coping strategies, reframe meaning, and build support systems through both familial and networked relationships.

Chaos Narrative

Overview

Many young patients with diabetes experienced significant cognitive dissonance, as their prediagnosis self-perception is disrupted by the sudden transformation into the identity of a "patient." The chaos narrative reflected the biographical disruption, which was characterized by the upheaval of daily routines and a profound sense of existential uncertainty.

Cognitive Dissonance

Young patients frequently experience distress due to abrupt fluctuations in blood sugar levels compounded by disruptions in self-awareness. Due to their limited specialized knowledge of diabetes management, they often reported feelings of disorientation and helplessness. For instance, one young student recalled:

> *Being diagnosed with diabetes was a key turning point in my life. I kept asking myself, "Why did I get this disease?"* [A1, November 17, 2023]

During this process, they encountered confusion regarding their self-identity and discomfort with the social role transition. As one young student stated:

> *Sometimes I wonder, what is a "normal person"? Someone who can eat cakes on their birthday? Someone who does not require medication Someone who does not get sick or someone who is physically and mentally healthy?* [A4, September 7, 2023]

This transition from being perceived as "normal" to being labeled as "sick" generated emotional distress, as it disrupted deeply ingrained assumptions about selfhood and autonomy [28].

Within the Chinese cultural framework, where familial bonds and collective identity are paramount [29], diagnosis intensifies feelings of guilt and responsibility. Patients felt that the most direct effect was that their previously orderly and stable family life was abruptly thrown into chaos and disarray. As one patient remarked:

> *I lie in my hospital bed, staring at the ceiling, often crying secretly. I hate myself for being so useless, for getting sick, for making my parents suffer so much.* [A12, February 15, 2023]

Traditional Chinese ideals of filial piety deepen this sense of guilt, positioning young adults as responsible for economic success, parental support, and the continuation of family lineage. As a result, illness is moralized, becoming not merely a health condition but a perceived failure to meet culturally embedded expectations of responsibility and reciprocity.

Meanwhile, they were particularly concerned about the specific costs of hospitalization, and this keen awareness of financial pressures reflected their inner kindness and sense of responsibility. For example, a young female patient guiltily expressed:

> *I've been in the hospital for four days, and the cost has already reached 6,000. I really feel like I am a burden on my family. So, I pray that my condition is not serious, and that I can reduce the burden on my family. Please!* [A8, March 12, 2023]

These young adults, who have traditionally served as the central figures in the functioning of the family, influence internal interactions and relationships within the family. They also expressed an urgent desire to become independent in the future to alleviate the pressure on their families. For instance, as a young adult stated:

> *What makes me even guiltier is my parents. They have to spend a lot of money every month on insulin, medications, and various supplies. I just hope to graduate from university quickly and find a job so I can make money and reduce the burden on my family.* [A7, March 12, 2023]

This cultural paradigm significantly mediates how young Chinese patients construct the meaning of diabetes, as their identity and health management are inextricably linked to perceptions of familial duty and collective well-being.

Disordered Life

Chronic pain, endless medical treatments, loss of bodily functions, and altered physical appearance, which distinguish these young adults from others, significantly drain both the mental vitality and physical energy of the patient. For example, a young patient vividly depicted the embarrassment and inconvenience caused by the disease:

XSL•FO
RenderX

> *It's so annoying, being woken up by hypoglycemia again. Now I have to sneak around like a midnight scavenger mouse, getting up quietly to eat without anyone knowing.* [A12, October 17, 2023]

Another patient shared a similar experience:

> *What does it feel like to be forced awake at 4 a.m? Low blood pressure, panic, no direction, rummaging through cabinets looking for anything to eat.* [A10, October 8, 2023]

These expressions deeply depicted the urgent responses and physiological states of patients with diabetes during nocturnal hypoglycemia, showing how they are physically bound by blood sugar fluctuations and psychologically bound by collapsing defenses.

The daily lives of these patients are disrupted by the arrival of the disease, leaving them with an overwhelming sense of uncertainty and unpredictability. For instance, as A4 (August 22, 2023) remarked, "When I was outside with hypoglycemia and my feet felt weak, I found that my glucose tablets had just run out; woke up in the middle of the night with hypoglycemia, and the nearest food was ten meters away." Such experiences are common among patients. Like another young female patient: "Whenever I leave the house, my bag is always packed to the brim, with a blood glucose meter, an entire box of test strips, alcohol swabs, needles, and all kinds of snacks to raise my blood sugar, along with my newly refilled medication" (A2, January 2, 2024). The metaphor of a "heavy bag" poignantly symbolizes the relentless burden, stress, and anxiety associated with diabetes management, a burden that resonates with the insights of Brashers [30] into the emotional costs of chronic illness.

Furthermore, what should have been a life centered around learning and personal growth was transformed into one focused on illness and treatment. The disruption of academic pursuits, which were foundational for future development and life, shattered young adults' hopes for the future. One patient expressed a sense of powerlessness:

> *Sudden hypoglycemia before the exam, who understands this sense of powerlessness? If I were already at the exam venue and had not brought candy with me, I cannot even imagine the consequences.* [A7, December 27, 2023]

This stress was further intensified by the sociocultural pressures inherent in the Chinese context, where the one-child policy places weight on an entire family's expectations of a single child. One young female university student complained:

> *7.9 before bed, and when I open my eyes it's 20.6. It's terrifying! I am preparing for the exam. I am dying of exhaustion.* [A3, December 22, 2023]

For Chinese young adults, academic success is not merely a personal achievement, but a fulfillment of intergenerational obligations. Thus, the disruption of their educational trajectory not only jeopardized their future prospects but also exacerbated the psychological distress associated with chronic illness, as they struggled to balance both their health and the high social expectations placed upon them.

Stigma Narrative

Overview

Meanwhile, the diagnosis of diabetes has prompted social withdrawal among young patients, exposing their experiences of stigma in daily life. The stigma narrative, which highlighted widespread misconceptions regarding insulin use and disease etiology, further reinforces structural stigma, thereby limiting patients' opportunities and eroding their self-worth. Many patients concealed their conditions to avoid imposing perceived burdens on family dynamics or facing social exclusion.

Social Withdrawal

Diabetes has a profound impact on the lives of young patients, manifesting significant emotional fluctuations and altering their social dynamics. Many young patients with diabetes feared being labeled and worried about the judgmental stares of others. For example, a young female patient expressed:

> *I started to reject the invitation from guys around me, proudly calling myself a "single lady." I was always afraid that others would find out about my "diabetes," and even afraid of their odd stares.* [A3, November 30, 2023]

Likewise, as a young adult remarked:

> *I planned to pursue him, but in the second semester, I discovered my condition. I let him go with a fracture in my soul, now love lies beyond my reach.* [A8, April 2, 2023]

These expressions illustrate how the disease undermines the foundations of social interaction, an essential aspect of maintaining a coherent life course [31].

Social alienation and exclusion led young adults to actively withdraw from social activities and passively become excluded. Some patients were met with unfair treatment, which led them to adopt a strategy of "hiding their diabetes and refraining from disclosing their condition to others." For example, as a young patient expressed:

> *What I did not expect was that my friends hearing about my condition did not even express any concern. They even dissolved the shared travel funds. I was sad and stunned. Maybe what I thought were good friends were just my own wishful thinking.* [A4, May 7, 2023]

This isolation aligns with previous research on the role of social context in shaping illness narratives [32], highlighting how misunderstanding and exclusion contribute to a sense of disillusionment regarding social relationships.

Simultaneously, due to the strict requirements of blood sugar control, many patients had to give up the activities they once enjoyed. One young female patient expressed her deep sense of loss and frustration by saying:

> *My hatred for diabetes reached its peak. When I see people eating well on the street, I feel sad. Just thinking about how it would make my blood sugar spike makes me angry.* [A12, June 7, 2023]

XSL•FO
RenderX

These narratives revealed that patients with suppressed pain experienced in managing their lifestyle and emotions as well as the internal struggles they endure while coping with the disorder in their lives. They are compelled to continually adapt to life with the illness while striving to maintain balance amid persistent challenges.

Social Discrimination

Many young patients expressed that the social "gaze" often made them ashamed of their identity as diabetics, and they were constantly seeking a balance between disease management and social expectations. For example, a university student shared how diabetes impacted her job opportunities by saying:

> *I did not get the offer because I have diabetes. Of course, there are so many "normal" people out there that they can be hired. I can only comfort myself that every boss wants to hire "normal" people.* [A6, July 31, 2023]

Although patients may not directly encounter overt criticism, they perceive stigma through nonverbal reactions, social and cultural responses, media portrayals, and their own imagination. Likewise, one young female patient shared: "I used to think I was different from normal people, hiding in school to inject insulin secretly" (A10, November 9, 2023). Such experiences illustrate that stigma is transmitted not only through overt social interactions but also through subtle, nonverbal cues and media portrayals, ultimately coercing patients to adopt unsustainable adaptations in managing their disease.

Moreover, reliance on insulin is frequently misinterpreted as ostentatious behavior or a marker of indulgence. As one patient remarked:

> *Some say we are showing off when we inject, others say we eat too much sugar, and some even use us as a cautionary tale. There's nothing we can do.* [A12, January 20, 2024]

The stigma caused by the physical damage inflicted by diabetes is transformed into a process of psychological distress. For them, the pain of the treatment process was not only the discomfort of diabetes itself and the side effects of treatment but also the sense of loss from being misunderstood. For example, one patient stated:

> *I hardly ever consumed sugary drinks. I can only say that I had type 1 diabetes by chance, by something unknown. Please give us more attention and less misunderstanding.* [A11, February 3, 2024]

These appeals reflected the patients' desire to raise awareness of diabetes and reduce discrimination.

Resilience Narrative

Overview

Many patients demonstrated remarkable emotional resilience, evident in their self-acceptance, gratitude, and proactive reframing of their disease experiences. By developing structured, yet flexible self-care routines, they reclaimed agency over their health and identities. Thus, resilience narratives constitute a pivotal element of coping strategies.

Emotional Fortitude

Young patients with diabetes often undergo psychological transformation and self-acceptance when facing the disease. They adapted to living with diabetes by actively recording their blood sugar levels. For instance, as a young patient shared:

> *I have been studying increasingly harder. I have maintained records for about two months. Although I cannot fully accept this disease yet, looking back, it feels like the boat has already sailed through a thousand mountains.* [A13, September 19, 2023]

Patients exhibited a positive attitude and self-affirmation when facing the challenges of the disease, and they had a clear understanding of and plan for managing their condition. As a patient stated:

> *I actively learn about blood sugar control, study the glycemic index of different foods, and calculate carbohydrate ratios, and continually try new methods. I love this version of myself I will always be proud of who I am! We can be sugar warrios.* [A3, December 7, 2023]

Such narratives resonated with the concept of narrative reconstruction [13], where the integration of daily health regimens into one's life story facilitated a shift from seeing diabetes solely as a source of illness-related burden to recognizing it as a catalyst for personal growth.

In addition, they expressed gratitude for their lives. Coping with the disruption of their life course and using surrounding resources often involves altering their self-awareness and rethinking the direction of their lives. Some patients expressed their gratitude for the availability of insulin, which enabled them to live a normal life and enjoy life's pleasures. As a patient remarked:

> *I'll be taking my first insulin shot on my 20th birthday, marking the 504th day of living with diabetes. Thanks to insulin, I can now enjoy this sweetness like a normal person.* [A8, April 1, 2024]

Other patients compared insulin to their "crystal energy," saying: "I eat in two steps: the first step is to check my blood sugar, the second step is to inject my crystal energy to stabilize my condition" (A11, March 15, 2024). Within this disarray, patients engaged in narrative reinvention and acts of symbolic reclamation that transformed vulnerability into agency. These positive attitudes and the desire for a healthy lifestyle demonstrated their optimism and discipline in managing their disease. In the context of disruption caused by past experiences and anticipated futures, they reconstructed their present experiences by adopting a proactive and positive approach to life.

Self-Routinization

Many patients demonstrated how they normalized the blood sugar management process. The philosophy of living in the present moment serves as a narrative for patients to reconstruct their lives after diabetes, aiding them in facing life after diabetes. For example, a female university student shared:

XSL•FO
RenderX

> *I always separate dry and wet food when eating. Foods that are too mushy are not suitable for blood sugar. Today, it's still a day full of energy!* [A3, June 9, 2023]

In line with ancient Chinese traditions, the notion of living in the present moment offered hope for a turning point even amid severe disruptions, suggesting that although diabetes may significantly disrupt life, hope remains, and one must confront present difficulties with a balanced mindset.

In addition, some patients integrated diabetes into their lifelong journeys. For example, one patient wrote:

> *I regularly monitor my blood sugar, and there have been no complications so far. Diabetes has kept me disciplined for life.* [A7, January 12, 2024]

They recognized that balancing relaxation and enjoyment of life is an essential part of maintaining health management. Similarly, a female university student reflected:

> *I feel much better because I know that blood sugar control is a lifelong journey, and allowing me to relax a little helps me to continue. I believe I will become stronger and stronger in managing my blood sugar.* [A10, October 21, 2023]

The perseverance and adaptability they displayed in managing their disease reflected their determination and courage in pursuing personal passions and fulfilling life needs.

Moreover, travel adventures have become a common means of solace for young patients with diabetes. As one shared:

> *This insulin shot is from snowcapped mountains. Disease is not a stumbling block on the road to fulfilling my dreams.* [A12, December 10, 2023]

During these journeys, they enjoyed exploration with a positive attitude while carefully managing their blood sugar, embracing every new adventure with health and resilience. Another patient shared a similar travel experience by saying:

> *Nanchang is a great city! I walked over 20,000 steps per day, and my blood sugar remained stable.* [A1, January 7, 2024]

Solidarity Narrative

Overview

The patients highlighted that kinship and digital social support are pivotal in enabling young patients with diabetes to adapt to chronic illnesses. These support systems foster resilience by strengthening familial bonds and enhancing engagement within digital communities. The narrative of solidarity illuminated familial care, and online networks provided both instrumental guidance and empathetic support, effectively transforming isolation into collective empowerment. By leveraging shared knowledge and reciprocating support, patients not only navigate disease management but also cultivate a robust culture of mutual aid.

Familial Solidarity

Although diabetes disrupted these young adults' life trajectories and strained their families, it also fostered resilience and strengthened their familial bonds. Most patients emphasized the meticulous care provided by their families throughout their disease journey. For instance, one patient stated:

> *I measure my blood sugar every two hours at night, and my mother and sister have been taking turns staying with me. I feel so blessed to have my family with me.* [A11, February 15, 2024]

Notably, many young patients with diabetes specifically mentioned the care provided by their fathers after diagnosis. In traditional Chinese culture, a father's love is often expressed through authority and strength. However, many fathers chose to express their affection through tenderness and care when confronted with their children's illnesses. For instance, one patient recalled:

> *When we reached the store, I pointed to the cream puff and told my dad it was delicious. He asked the clerk, "Does this contain sugar?" At that time, I began. My dad always encouraged me to try any food that I wanted to. In the past, I would get very upset, but now I understand this is another form of love from my father.* [A10, January 1, 2023]

The father's actions deeply illustrated unspoken love, encouraging his child to live courageously, and the child, in turn, gradually felt and appreciated this profound paternal affection, learning to accept and value support from their parents in managing the disease. Similarly, a patient expressed that the disease provided an opportunity to rebuild their familial bond:

> *After I got diabetes, I felt like the conversations between my dad and I became much more frequent. He asks me every day what I want to eat, what dish I'd like for dinner, and what fruits I'd like for tomorrow.* [A11, January 9, 2024]

In facing the destruction caused by illness, most patients mentioned the significant role that family care and encouragement played at different stages of their illness. Even though family relationships may become disordered or chaotic after the onset of illness, their unwavering support gradually helped to mitigate the relational fractures caused by the disease.

Reciprocal Digital Community

The mutual understanding and support among those with diabetes online helped them better manage the disease and served as an important pillar in their lives. As natives of the digital age, young patients with diabetes actively used social platforms to cope with challenges, sharing their daily experiences with online diabetes friends to seek emotional support and a sense of belonging. One patient wrote:

> *I am truly thankful to everyone here, for making me no longer feel like an island. As fellow members chose diabetes, we really feel like a family on the Internet.* [A4, December 27, 2023]

This finding echoes earlier research, showing that digital interactions can transform feelings of isolation into collective empowerment, highlighting how online health narratives both shape and are shaped by their social contexts [32]. When faced with day-to-day challenges such as preparing for a graduate

XSL•FO
RenderX

school entrance interview, patients turned to these virtual networks for guidance. As a patient remarked:

> *I'm about to attend a graduate school entrance exam interview soon, which includes both a written test and an interview. I want to ask any experienced friends with diabetes. Can I wear an insulin pump during the interview?* [A3, March 23, 2024]

Online support from netizens continued to be a stable provider of ongoing social support for patients in instrumental and emotional support.

Through the process of identity switching and reinforcement, patients found new meanings in their relationship with the disease, including acceptance of life and disease as natural, a sense of gratitude and reciprocation, the pursuit of value realization and social recognition, and a life attitude of coexistence with the disease. For example:

> *Dear fellow diabetics, please don't be sad. There are millions of people in the world who are like us. We can lift our heads and embrace ourselves in the most vibrant winds.* [A1, November 17, 2023]

Young patients with diabetes not only actively adjusted their mindset through self-encouragement but also inspired others who were in similar situations. Another girl shared the following tips:

> *I just want to share how I carry my little Susu (insulin) when I go out. I bought this storage bag from an online platform, and it was about the length of my arm. Let me show you the inside. It can hold two–three pens, and there is space for ice packs, alcohol swabs, needles, and other essentials. Everything fits perfectly inside.* [A12, October 12, 2023]

By sharing their practical life experiences, they actively helped others and showed a strong sense of responsibility and empathy. Some patients also provided detailed guidance on dietary considerations for diabetes management, helping others better control their blood sugar levels. For instance:

> *Going from rich, heavy-flavored meals to a relatively lighter, sugar-controlled diet is definitely hard to adjust to. My approach is trying to eat until I am 70% full, and use lighter cooking methods to taste the natural flavors of food.* [A5, March 20, 2024]

As they deepened their understanding of blood sugar control, many patients shared their sugar control experiences:

> *Match insulin well, and I can eat what normal people eat. Sugar-free or zero-calorie coffee has almost no effect on blood sugar, while thicker milk causes an increase of 1 - 2 points.* [A13, October 10, 2023]

Young patients with diabetes offer both instrumental and emotional support to one another through online platforms. These findings align closely with the literature on digital health narratives and support networks, underscoring the critical role of online communities in chronic illness management [13].

Discussion

Principal Findings

This study examined how young Chinese individuals with diabetes construct illness narratives in digital environments. Through narrative analysis of 303 posts on RedNote, 4 narrative types were identified: chaos, stigma, resilience, and solidarity. These narratives demonstrate that chronic illness is experienced not only as a biographical disruption but also as an ongoing process of identity reconstruction shaped by cultural expectations, social pressures, and digital interaction.

Cultural Context and Illness Narrative Reconstruction

The findings indicated that young patients with diabetes face intertwined dilemmas of biographical disruption and identity reconstruction, requiring them to renegotiate their lifestyles, social roles, and self-perceptions. In the Chinese cultural context, where individuals are embedded within interdependent familial structures, illness is often experienced not solely as a personal condition but as a disruption to familial continuity and social expectations [33]. Diabetes-induced guilt therefore intensifies identity crises, as patients may perceive their illness as a failure to fulfill filial responsibilities and future obligations [34]. These pressures are further reinforced by broader structural conditions, including the legacy of the one-child policy, competitive educational systems, and familial expectations surrounding academic achievement, economic success, and marriage.

Situating these findings within the Chinese cultural context helps explain the distinctive features of these narratives. This study offers a culturally grounded reinterpretation of the illness narrative framework of Frank [23]. In the Chinese context, chaos narratives are often intertwined with culturally specific elements such as familial guilt and moral responsibility, which extend the scope of the original concept. Regarding the other narrative types, they reflect contextually grounded variations and emergent categories shaped by the sociocultural and digital environment under study. For example, the resilience narrative differs from restitution, in that it does not center on returning to a preillness state, but rather on ongoing adaptation and disciplined self-management. The stigma narrative captures socially imposed processes of devaluation that are not fully encompassed by chaos, while the solidarity narrative highlights digitally mediated forms of collective coping that extend beyond the individual-oriented trajectory implied in quest narratives.

Digital Solidarity and Platform-Mediated Identity Formation

At the same time, the findings reveal how digital environments reshape both illness identity and health governance. Young patients were simultaneously burdened by intergenerational pressures derived from Confucian family ethics while benefiting from forms of "liquid solidarity" fostered through online communities [35]. Through narrative sharing, peer interaction, and emotional support, patients actively engaged in self-healing, meaning-making, and identity reconstruction. Familial support, social networks, and digitally mediated peer communities emerged as critical resources that enabled patients to normalize illness experiences, reframe emotional distress, and gradually

XSL•FO
RenderX

develop systems of self-care. Frequent self-referencing as "sugar warriors" reflected a shift in illness identity from stigma toward a digitally networked form of resistance and collective belonging. This transformation aligns with Castells' [22] framework of identity formation in network society, in which patients construct counter-public spaces and solidarity communities through algorithmically connected peer networks.

Health Neoliberalism and Digital Self-Management

However, these digitally mediated forms of empowerment also operate within broader structures of health neoliberalism. As bodily functions increasingly become tied to ideals of productivity, discipline, and self-optimization, chronic illness may be framed as a form of personal failure rather than a socially embedded condition. Within this context, seemingly neutral disease-related content on digital platforms often reproduces a "self-management supremacist" governance logic that transforms structural inequalities into matters of individual responsibility [36]. Consequently, digital illness narratives function not only as spaces of support and resistance [37] but also as sites where broader moral expectations surrounding health and self-discipline are reproduced and negotiated. This tension highlights the ambivalent role of digital platforms, which simultaneously enable collective empowerment and reproduce normative pressures surrounding health, productivity, and self-governance.

Clinical Implications

Rather than focusing solely on individual-level coping or behavioral change, the findings of this study point to the need for a structurally informed approach to diabetes care and communication. The narratives analyzed here reveal how responsibility for disease management is often moralized and individualized, particularly within digital environments shaped by health neoliberalism. As such, clinical and public health interventions should move beyond emphasizing self-discipline and adherence, and instead work to demoralize chronic illness by recognizing the structural, cultural, and psychosocial conditions that shape patients' experiences.

At the clinical level, this calls for greater sensitivity to the moral and emotional burdens associated with self-management, particularly in contexts where familial expectations and intergenerational responsibility intensify patients' sense of guilt. Clinicians should be attentive to how patients interpret illness in relation to identity, family roles, and social expectations. At the platform level, the findings suggest the need to critically examine how digital health content and algorithmic visibility may reinforce a "self-management–centered" logic. Social media platforms should consider how to diversify representations of illness experiences, making space for narratives that reflect structural constraints, ambivalence, and nonlinear coping trajectories, rather than privileging idealized forms of disciplined self-management. More broadly, this study calls for a shift in health communication from framing diabetes as an individual failure of control to understanding it as a socially and culturally mediated condition, shaped by intersecting forces of family structure, educational pressure, and digital media environments. Such a shift is essential for developing more equitable and context-sensitive approaches to chronic illness care.

Limitations

This study has several limitations. First, the sample consists of 13 users from a single platform, which limits the generalizability of the findings. Although purposive sampling was used to identify information-rich cases, the dataset may not capture the full diversity of illness experiences among young people with diabetes. The predominance of female college students in the sample may also reflect platform-specific participation patterns rather than broader population characteristics. Second, the study focuses on publicly available social media content, which provides valuable insights into self-expression but may not fully represent offline experiences or less visible forms of coping. Users may selectively present certain aspects of their lives while omitting others, leading to potential bias in the narratives analyzed. Third, the platform context of RedNote may shape both the style and content of narratives. As a platform oriented toward lifestyle sharing and social engagement, it may privilege particular modes of storytelling while marginalizing others. Future research could compare multiple platforms to better understand how different digital environments influence illness narratives. Fourth, this study relies solely on publicly available user-generated content and does not include interviews with patients or family members. As a result, interpretations of family dynamics are based on patients' self-representations. Future research could combine digital trace data with interviews to strengthen validity.

Conclusions

This study demonstrates that illness narratives among young Chinese individuals with diabetes are shaped by the intersection of cultural values, social structures, and digital media environments. By identifying 4 narrative types, the study highlights how patients move beyond disruption to actively reconstruct identity, negotiate stigma, and build forms of resilience and solidarity. More broadly, the findings suggest that digital platforms are not merely spaces for sharing experiences, but active sites of meaning-making that transform how chronic illness is understood and managed. Recognizing these dynamics has important implications for health communication, particularly in designing culturally sensitive and platform-aware interventions that support young patients in navigating chronic illness. Rather than viewing diabetes solely as a biomedical condition, this study underscores the importance of understanding it as a socially and culturally mediated experience. Future research should further explore how digital storytelling shapes long-term health outcomes and how intergenerational communication may mitigate the moral pressures associated with illness in collectivist contexts.

XSL•FO
RenderX

Acknowledgments

The authors used ChatGPT as an artificial intelligence–assisted language tool for minor editing and grammar checks. All substantive content, analysis, and interpretations were developed and reviewed by the authors.

Funding

The authors declared no financial support was received for this work.

Data Availability

The data supporting the findings of this study will be made available by the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: ZL
Methodology: ZL
Data collection: DF, YL
Formal analysis: ZL, DF, YL
Writing—original draft: ZL
Writing—review and editing: ZL, YM

Conflicts of Interest

None declared.

References

1. Li Y, Teng D, Shi X, et al. Prevalence of diabetes recorded in mainland China using 2018 diagnostic criteria from the American Diabetes Association: national cross sectional study. BMJ 2020 Apr 28;369:m997. [doi: 10.1136/bmj.m997] [Medline: 32345662]
2. IDF Diabetes Atlas 2025. International Diabetes Federation. URL: https://diabetesatlas.org/resources/idf-diabetes-atlas-2025/ [accessed 2026-05-13]
3. Fu D, Liu Z, Liu Y. Torment, adaptation, and transcendence: exploring Chinese parents' online narratives of juvenile diabetes and cultural constraints. Health Commun 2026 Jan;41(1):91-102. [doi: 10.1080/10410236.2025.2490605] [Medline: 40257071]
4. Wang L, Peng W, Zhao Z, et al. Prevalence and treatment of diabetes in China, 2013-2018. JAMA 2021 Dec 28;326(24):2498-2506. [doi: 10.1001/jama.2021.22208] [Medline: 34962526]
5. Rintala TM, Jaatinen P, Paavilainen E, Astedt-Kurki P. Interrelation between adult persons with diabetes and their family: a systematic review of the literature. J Fam Nurs 2013 Feb;19(1):3-28. [doi: 10.1177/1074840712471899] [Medline: 23288886]
6. Fu R, Shi L, He FY. Comorbidity of diabetes and depression in middle-aged and older adults: effects of sleep, physical activity, and social participation [Article in Chinese]. Chin Gen Pract 2025;28(20). [doi: 10.12114/j.issn.1007-9572.2024.0474]
7. Liu F. Constructing the autonomous middle-class self in today's China: the case of young-adult only-children university students. J Youth Stud 2008 Apr;11(2):193-212. [doi: 10.1080/13676260701800746]
8. Hickling A, Dingle GA, Barrett HL, Cobham VE. Systematic review: diabetes family conflict in young people with type 1 diabetes. J Pediatr Psychol 2021 Sep 27;46(9):1091-1109. [doi: 10.1093/jpepsy/jsab052] [Medline: 34313769]
9. Leung J, Tang TS, Lim CE, Laffel LM, Amed S. The four I's of adolescent transition in type 1 diabetes care: a qualitative study. Diabet Med 2021 Jul;38(7):e14443. [doi: 10.1111/dme.14443] [Medline: 33107064]
10. Malik FS, Panlasigui N, Gritton J, Gill H, Yi-Frazier JP, Moreno MA. Adolescent perspectives on the use of social media to support type 1 diabetes management: focus group study. J Med Internet Res 2019 May 30;21(6):e12149. [doi: 10.2196/12149] [Medline: 31199310]
11. Iannarino NT. "My Insides Feel Like Keith Richards' Face": a narrative analysis of humor and biographical disruption in young adults' cancer blogs. Health Commun 2018 Oct;33(10):1233-1242. [doi: 10.1080/10410236.2017.1350909] [Medline: 28825504]
12. Banerjee SC, D'Agostino TA, Gordon ML, Hay JL. "It's Not JUST Skin Cancer": understanding their cancer experience from melanoma survivor narratives shared online. Health Commun 2018 Feb;33(2):188-201. [doi: 10.1080/10410236.2016.1250707] [Medline: 27982699]
13. Martinez K, Frazer SF, Dempster M, Hamill A, Fleming H, McCorry NK. Psychological factors associated with diabetes self-management among adolescents with type 1 diabetes: a systematic review. J Health Psychol 2018 Nov;23(13):1749-1765. [doi: 10.1177/1359105316669580] [Medline: 27663288]

XSL•FO
RenderX

14. Sanders T, Elliott J, Norman P, Johnson B, Heller S. Disruptive illness contexts and liminality in the accounts of young people with type 1 diabetes. Sociol Health Illn 2019 Sep;41(7):1289-1304. [doi: 10.1111/1467-9566.12906] [Medline: 30968432]
15. Jindal-Snape D, Johnston B, Pringle J, et al. Multiple and multidimensional life transitions in the context of life-limiting health conditions: longitudinal study focussing on perspectives of young adults, families and professionals. BMC Palliat Care 2019 Mar 25;18(1):30. [doi: 10.1186/s12904-019-0414-9] [Medline: 30909901]
16. Bury M. Chronic illness as biographical disruption. Sociol Health Illn 1982 Jul;4(2):167-182. [doi: 10.1111/1467-9566.ep11339939] [Medline: 10260456]
17. Li X, Xu K. Coping with depression: a narrative study of an online depression community in China. Int J Qual Stud Health Well-being 2023 Dec;18(1):2268379. [doi: 10.1080/17482631.2023.2268379] [Medline: 37847860]
18. Esin C, Fathi M, Squire C. Narrative analysis: the constructionist approach. In: Flick U, editor. The SAGE Handbook of Qualitative Data Analysis: SAGE Publications Ltd; 2014, Vol. 0:203-216. [doi: 10.4135/9781446282243.n14]
19. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Aug 8;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
20. Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and be(com)ing a knowing researcher. Int J Transgend Health 2023;24(1):1-6. [doi: 10.1080/26895269.2022.2129597] [Medline: 36713144]
21. Goffman E. Stigma: Notes on the Management of Spoiled Identity: Simon and Schuster; 2009.
22. Castells M. Toward a sociology of the network society. Contemp Sociol 2000 Sep;29(5):693. [doi: 10.2307/2655234]
23. Frank AW. The Wounded Storyteller: Body, Illness, and Ethics: University of Chicago Press; 1995.
24. 2024 “active users” research report (Xiaohongshu platform). QIAN-GUA.COM. 2024. URL: https://www.qian-gua.com/blog/detail/2898.html [accessed 2026-05-13]
25. Zhong Y, Hahne J, Wang X, et al. Telehealth care through internet hospitals in China: qualitative interview study of physicians’ views on access, expectations, and communication. J Med Internet Res 2024 Mar 29;26:e47523. [doi: 10.2196/47523] [Medline: 38551618]
26. Xu K, Li X. Complaining, regret, superiority, and discovery: Chinese patients’ sense making of depression in an online forum. Qual Health Res 2023 Jun;33(7):613-623. [doi: 10.1177/10497323231160119] [Medline: 37051623]
27. Liu Z, Zou W. Compliance, shame, resentment, and awakening: a critical examination of women’s illness narratives about cervical erosion. Health Commun 2026 Feb 16;16:1-10. [doi: 10.1080/10410236.2026.2632760] [Medline: 41697149]
28. Conrad P, Barker KK. The social construction of illness: key insights and policy implications. J Health Soc Behav 2010;51 Suppl:S67-S79. [doi: 10.1177/0022146510383495] [Medline: 20943584]
29. Blair SL, Madigan TJ. Dating, marriage, and parental approval: an examination of young adults in China. Soc Sci Q 2019 Oct;100(6):2351-2368. [doi: 10.1111/ssqu.12718]
30. Brashers DE. Communication and uncertainty management. J Commun 2001 Sep 1;51(3):477-497. [doi: 10.1111/j.1460-2466.2001.tb02892.x]
31. Hersh A, von Scheven E, Yelin E. Adult outcomes of childhood-onset rheumatic diseases. Nat Rev Rheumatol 2011 May;7(5):290-295. [doi: 10.1038/nrrheum.2011.38] [Medline: 21487383]
32. Zhou X, Sun S, Yang J. Sweet home: understanding diabetes management via a Chinese online community. Presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’14); Apr 26 to May 1, 2014. [doi: 10.1145/2556288.2557344]
33. Nisbett RE, Peng K, Choi I, Norenzayan A. Culture and systems of thought: holistic versus analytic cognition. Psychol Rev 2001 Apr;108(2):291-310. [doi: 10.1037/0033-295x.108.2.291] [Medline: 11381831]
34. Fong VL. Only Hope: Coming of Age under China’s One-Child Policy: Stanford University Press; 2004.
35. Bauman Z. Liquid Modernity: Polity Press; 2000.
36. Liu Z, Zou W. Breaking taboos in women’s reproductive health: the communication strategies used by top OB/GYN influencers in Chinese social media. Health Commun 2024 Apr;39(4):685-696. [doi: 10.1080/10410236.2023.2181677] [Medline: 36803354]
37. Vauthier JC, Choley L, Arduini D, Mas P, Kabuth B. Diabetes technologies in ultra-endurance type 1 diabetes: qualitative study. J Med Internet Res 2026 May 8;28:e86815. [doi: 10.2196/86815] [Medline: 42101231]

Edited by I Steenstra; submitted 29.Mar.2026; peer-reviewed by C Martin, Y Liu, Y Hu; revised version received 29.May.2026; accepted 29.May.2026; published 24.Jun.2026.

Please cite as:
Liu Z, Fu D, Liu Y, Meng Y
Digital Illness Narratives of Young Chinese Adults With Diabetes on RedNote: Qualitative Narrative Analysis
J Med Internet Res 2026;28:e96543
URL: https://www.jmir.org/2026/1/e96543
doi:10.2196/96543

XSL•FO
RenderX

© Zikun Liu, Donghan Fu, Yingjie Liu, Ying Meng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Patient-Reported Experiences With Viewing and Understanding Test Results in Patient Portals: Cross-Sectional Survey Analysis

Chelsea Richwine[1], MA, PhD; Bryan Steitz[2], PhD; Jordan Everson[1,3], PhD

[1]Office of the National Coordinator for Health Information Technology, U.S. Department of Health and Human Services, 330 C St SW, Washington, DC, United States

[2]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

[3]Department of Family Medicine, Georgetown University, Washington, DC, United States

Corresponding Author:
Chelsea Richwine, MA, PhD
Office of the National Coordinator for Health Information Technology, U.S. Department of Health and Human Services, 330 C St SW, Washington, DC, United States

Abstract

Background: 21st Century Cures Act information blocking regulations led to many organizations altering policies to electronically release test results to patients immediately upon their availability.

Objective: This study aims to understand the prevalence and variation in how often patients view test results in the patient portal before hearing from their health care provider (HCP), are given the option to decide how results are communicated, and understanding of results viewed before hearing from their HCP.

Methods: In this cross-sectional analysis, we analyzed data from the 2024 Health Information National Trends Survey on US adults with complete information who received test results via patient portal (n=6045). We examined the share of all patients and "portal users" (accessed their portal in the last year) who viewed test results in their portal, viewed results before hearing from their HCP, and were given the option to decide how results were communicated. We examined patients' understanding of results viewed before hearing from their HCP and compared it with their understanding of other health information. Finally, we examined differences in weighted proportions of outcomes by individual characteristics using survey-weighted linear regression. Differences were assessed using design-adjusted Wald tests.

Results: In 2024, 70% of patients (92% of portal users) viewed results in their patient portal, and 58% (76% of portal users) viewed the results before hearing from their HCP. Overall, 28% of patients and 33% of portal users reported being given the option to decide whether they wanted to receive test results before hearing from their HCP. Only 66% of patients reported that they understood the results they viewed in their patient portal. Provider encouragement, higher education, and digital literacy were most strongly associated with higher rates of viewing test results overall and before hearing from their HCP (eg, 70% of those encouraged viewed results vs 24% not encouraged; $P<.001$). Associations with age and chronic condition were also observed but were weaker. While higher education and digital literacy, age, and provider encouragement were strongly associated with being given the option to decide how to receive results, only high digital literacy and provider encouragement were strongly, positively associated with understanding results viewed before hearing from their HCP (eg, 77% of patients with high digital literacy understood results compared to 53% of those with low digital literacy; $P<.001$).

Conclusions: This study is the first to provide a national estimate of patients who access immediately released test results and whether patients are given the option to decide how their results are communicated to them. While most patients view results before discussing them with their HCP, rates varied by individual characteristics, and substantially fewer patients report being given the option to decide how results are communicated. Incorporating patient preferences in portal communications may empower patients while preserving immediate access.

(J Med Internet Res 2026;28:e94098) doi:10.2196/94098

KEYWORDS

personal health records; patient portal; patient engagement; electronic health records; health information technology; communication

Introduction

Individuals have long had a right to access their protected health information [1]. In practice, potentially sensitive information like laboratory and imaging results were often not available to patients until health care providers (HCPs) had a chance to review, interpret, and follow up on findings [2]. In implementing the 21st Century Cures Act provision prohibiting information blocking, the US Department of Health and Human Services

XSL•FO
RenderX

indicated that broad organizational policies that delay the release of health information to patients were likely to implicate information blocking, even if an HCP has not yet had a chance to review the result or personally inform the patient of the results before they can electronically access them [3,4]. Consequently, many organizations now immediately release test results to the patient portal, leading to more patients accessing their results before they are reviewed by the ordering HCP [5,6]. Some have expressed concerns that immediate release of results without additional context and interpretation from an HCP could cause patients to experience negative emotions [7,8]. However, studies conducted since the information blocking regulations became applicable on April 5, 2021, have shown that patients overwhelmingly prefer having immediate access to test results, even if it means viewing those results before discussing them with their HCP [9-11]. However, preferences for immediate release may depend on the severity or sensitivity of the results, and some patients may prefer to be contacted by their HCP shortly after receiving abnormal results [12-14].

While several studies have sought to understand patient preferences around viewing their immediately released results in the patient portal, it is unclear to what extent patient preference is considered by health care organizations when information is released in patient portals. For instance, whether patients were made aware they might receive results in their portal before their HCP could discuss them, and if they were given the option to decide whether they wanted results immediately released to their portal, held until their HCP released them, or alerted to new results upon release [15]. Further, little is known about whether patients who viewed test results in their portal before hearing from their HCP felt that they understood those results and what they meant for their care. Conversations during health care visits present an opportunity for HCPs to provide context to tests being ordered and help prepare patients on what to expect in terms of communication and follow-up. Test results provided through the portal may therefore omit helpful context.

This study is the first to provide a national estimate of the share of patients who access immediately released test results and sheds light on the extent to which patients are given the option to decide how their results are communicated to them. We used nationally representative survey data to examine (1) whether patients viewed test results in the patient portal before hearing from their HCP, (2) whether patients were given the option to decide how test results were communicated, and (3) whether they understood the results and their implications. To contextualize these findings, we compared patients' understanding of test results to their general understanding of information in the patient portal and HCP explanations during health care visits and examined differences by digital health literacy and clinical needs. This work provides national evidence on how patients engage with immediately available test results, their communication preferences, and their preparedness to interpret results before consulting their HCP.

Methods

Inclusion and Exclusion

Data for this cross-sectional survey study came from the Health Information National Trends Survey (HINTS) 7 (2024), a nationally representative survey of U.S. noninstitutionalized adults conducted by the National Cancer Institute that tracks individuals' understanding and use of health-related information and health information technology. The analytic sample used for this secondary analysis of the HINTS was limited to 6335 respondents who had test results in the past year, as determined by responses to question E10 on HINTS 7. It excluded respondents with missing data for test results-related questions in E3, E5, E9, and E11 (<1% across all measures), resulting in a total sample of 6045 patients with recent test results.

Participant Characteristics

The total study population included 6045 patients (mean age, 50.07, SE 0.24 years), for whom self-reported level of education was as follows: 344 (6%) had less than high school education, 935 (20%) had high school education, 1791 (39%) had some college, and 2975 (35%) had a college education (Table S1 in Multimedia Appendix 1). Quartiles of digital literacy were as follows: 2280 (33%) in the bottom quartile (lowest digital literacy), 813 (13%) in the second quartile, 1836 (33%) in the third quartile, and 1103 (21%) in the top quartile (highest digital literacy). For health status: 3994 (62%) reported at least one chronic condition (diabetes, high blood pressure, heart condition, lung disease, or depression), and 302 (3%) had a recent cancer diagnosis. 4558 (76%) respondents reported being encouraged by their HCP to use their portal.

Sampling Procedures

The survey was fielded as both a paper and web-based survey from March 25, 2024, to September 16, 2024. The sample design consisted of 2 stages in which a stratified sample of addresses was first selected from a file of residential addresses, and then one adult was selected within each household. The sampling frame consisted of a database of addresses used by Marketing Systems Group to provide a random sample of addresses. The sampling frame of addresses was grouped into 4 explicit sampling strata by racial or ethnic minority and rural status. Both survey modes (paper and web) were offered in English or Spanish. Full details on HINTS 7 sample selection, data collection, and management are available in the HINTS 7 methodology report [16].

Sample Size, Power, and Precision

The total number of addresses selected for HINTS 7 was 36,000 (by strata: 25,300 from high minority urban, 1950 from high minority rural, 6100 from low minority urban, and 2650 from the low minority rural), which based on historical HINTS response rates, was the sample size needed to achieve the target number of 7000 complete questionnaires. The HINTS over-sampled high-minority and rural strata to increase the precision of estimates for racial or ethnic minority and rural subpopulations. Ultimately, complete questionnaires were collected from 7278 respondents (27.31% response rate).

XSL•FO
RenderX

Ethical Considerations

The HINTS 7 general population survey was designated "exempt research" under 45 CFR 46.104. HINTS 7 also received a "Not Human Subjects Research" determination from the NIH (National Institutes of Health) Office of IRB Operations (IRBID: IRB002042). Compensation type and amount for human participants research are therefore not applicable to this study. To protect the privacy of respondents, HINTS data are deidentified and do not contain any geographical indicators. No identification of individual participants or users in any images of the manuscript or supplementary material is possible. Publicly available data are available at the aggregate level only. This study was deemed exempt from institutional review board review as it involved secondary analysis of publicly available, deidentified data that does not involve human participants [17].

Measures and Covariates

Outcomes

We defined 4 binary outcomes of interest related to patients' experiences receiving communication about test results, viewing results in their online medical record or patient portal, and understanding those results before discussing them with their HCP. All survey questions asked about actions occurring in the past 12 months.

View results. The first outcome was derived from question E5 on HINTS 7 and indicated patients who reported using their portal to view test results (equal to '1' if "Yes," equal to '0' if "No" or "Not applicable" (ie, did not access portal in the past 12 months).

View results before hearing from HCP. The second outcome was derived from question E10 on HINTS 7 and indicated patients who reported using their portal to view test results *before* hearing about the result from their HCP (equal to "1" if "Yes," equal to '0' if "No"). Respondents who did not have any medical tests in the past 12 months were excluded from the study.

Option to decide. The third outcome was derived from question E9 on the HINTS and indicated patients who reported being given the option to decide whether they wanted to receive test results before their HCP could discuss the results with them (equal to "1" if "Yes," equal to '0' if "No" or "Don't know").

Understand results before hearing from HCP. The fourth outcome was derived from question E11 on the HINTS. We converted the measure from a Likert scale into a binary variable that indicates patients who reported they understood "very well" or "well" what the test results showed and what they meant for their care (equal to "0" if "fairly well" or "poorly"). This question was only asked of individuals who indicated they viewed test results in their portal before hearing from their HCP.

The full list of survey questions is available on the HINTS 7 Survey Instrument on the HINTS website [18]. A mapping of all outcomes and measures to survey questions on HINTS 7 is available in Table S2 in Multimedia Appendix 1.

Comparable Measures Related to Understanding Information

Individuals' understanding of test results they viewed in their portal before hearing from their HCP was compared to patients' general understanding of information in their portal and HCP explanations of health information communicated during a health care visit. The measure of general understanding of information in the portal was derived from question E6 on HINTS 7 and defined as a binary variable indicating patients who reported it was "very" or "somewhat" easy to understand health information in their online medical record or patient portal (equal to "0" if "somewhat" or "very" difficult). The measure of HCPs explaining information clearly during health care visits was derived from question C4e. on the HINTS and defined as a binary variable indicating patients who reported their HCPs "always" or "usually" explain things in a way they can understand (equal to "0" if "sometimes" or "never").

Individual Characteristics

We examined differences in outcomes across characteristics that may influence individuals' comfort or need for using that technology including age, education, digital literacy, and having a recent cancer diagnosis in the last 5 years or a chronic condition (including diabetes, high blood pressure, heart condition, lung disease, or depression), as these individuals have been shown to be more frequent users of patient portals [19]. We also examined outcomes among individuals who were encouraged by an HCP to use the patient portal, which has been shown to influence patient portal access and use [20]. We created a digital literacy index from 3 correlated measures derived from question B5 on HINTS 7 that were intended to capture digital literacy by asking patients about the extent to which they agree or disagree with statements about their comfort with technology and digital tools (a. "I find learning how to use new technology frustrating," "b. I can use applications/programs on my cell phone or computer without asking someone for help," and "c. I have the skills to find the health information I need on the Internet"). The index was created by taking the average across the 3 self-reported measures. We then used quartiles of the digital literacy index to capture the range of digital literacy, where those in the bottom quartiles (closest to 1) have lower digital literacy (ie, more difficulty with technology, less comfort with digital tools) and those in the top quartiles (closer to 4) have the highest digital literacy.

Statistical Analysis

Leveraging our analytic sample, we first compared the proportions of individuals who viewed test results in their patient portal, viewed results in their portal *before* hearing from their HCP, and reported being given the option to decide whether they wanted to receive results before hearing from their HCP. We examined rates among all patients with recent results and among those who accessed their patient portal in the past 12 months ("portal users"). We then examined patients' understanding of test results they viewed in their portal before hearing from their HCP, and how this compared to their general understanding of information in their portal and HCP explanations of health information. Finally, we examined differences in the weighted proportion of patients who viewed

XSL•FO
RenderX

results in their portal (overall and before hearing from their HCP), reported being given the option to decide how to receive results, and understood the results they viewed before discussing with their HCP by demographic and health characteristics. Differences between groups were assessed using survey-weighted linear regression with statistical significance evaluated using design-adjusted Wald tests. All analyses used survey weighting procedures with jackknife replicate weights, which were designed to correct for nonresponse bias and under-sampling as well as to enable valid inferences from the study sample to the population. See the HINTS 7 Methodology report for more details on sample weights [16]. Data analyses were performed in January to March 2025 using Stata/SE (version 15.1; StataCorp).

Results

The selection of the analytic sample for this study is shown in Figure 1 below. 7278 individuals responded to HINTS 7, 6332 individuals or "patients" had test results in the past 12 months, and 6045 patients with test results had complete information across measures. Data missingness was less than 1% across all measures.

Figure 1. Creation of analytic data set from Health Information National Trends Survey 7. HINTS 7: Health Information National Trends Survey 7.

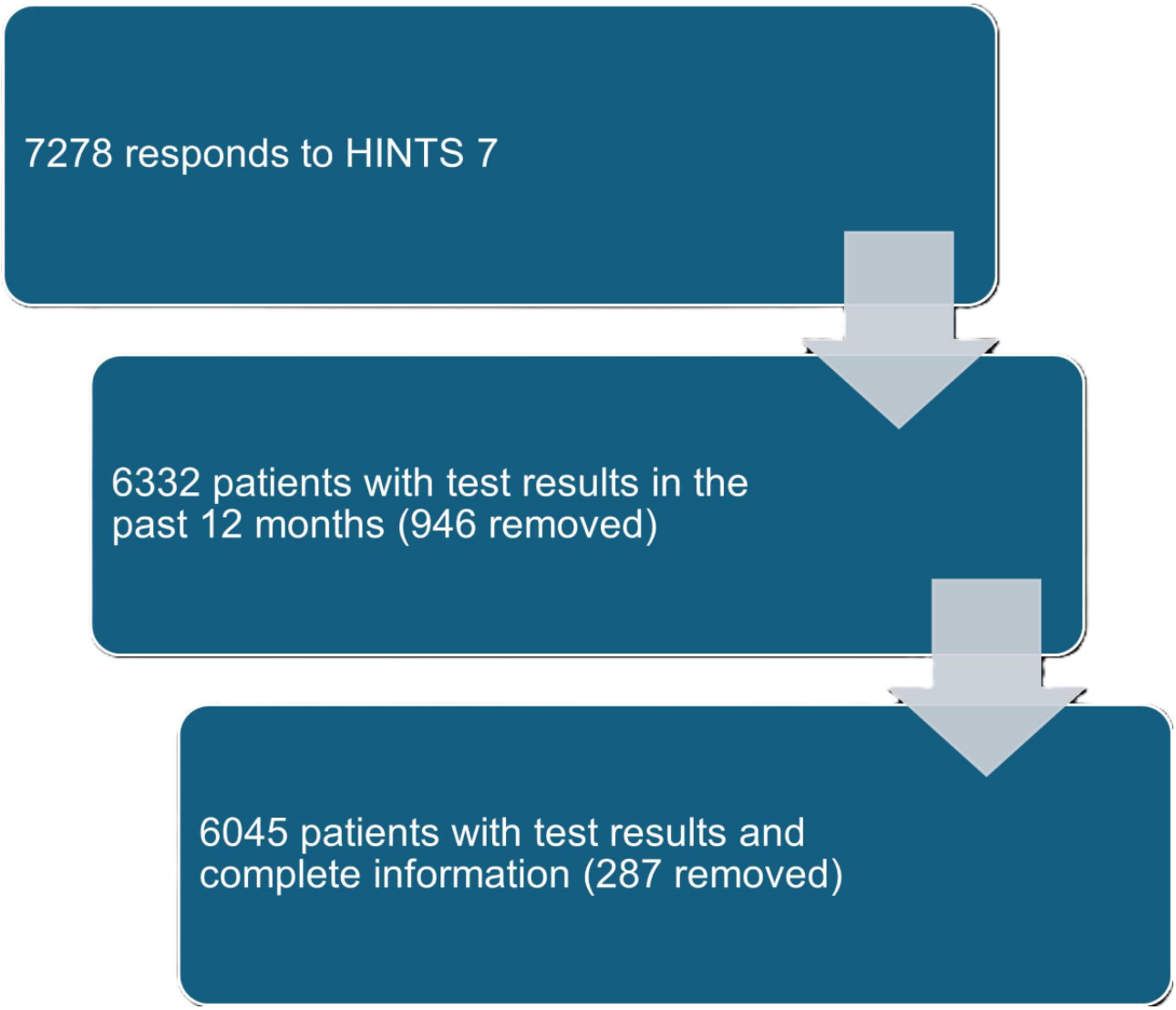

In 2024, most patients who received test results reported viewing those results in their patient portal (70% overall, 92.4% among portal users) and viewed the results in their portal before hearing from their HCP (57.6% overall and 76% of portal users; Table 1). About one-quarter of patients (28%) and one-third of portal users (33.3%) reported being given the option to decide whether they wanted to receive test results before their HCP could discuss the results with them.

XSL•FO
RenderX

Table . Percent of patients with test results who viewed results in their patient portal overall and before hearing from their HCP, and the share of patients who reported they were given the option to decide whether they wanted to receive results before hearing from their HCP, 2024[a].

Response	All individuals (n=6045)			Patient portal users (n=4485)		
	(1) View results, n (%) [95% CI]	(2) View results before hearing from HCP[b], n (%) [95% CI]	(3) Option to Decide, n (%) [95% CI]	(4) View results, n (%) [95% CI]	(5) View results before hearing from HCP, n (%) [95% CI]	(6) Option to decide, n (%) [95% CI]
Yes	4172 (70) [68.2-71.8]	3432 (57.6) [28.2-31.8]	1768 (28) [26.2-29.9]	4172 (92.4) [90.9-93.7]	3432 (76) [73.9-77.8]	1545 (33.3) [31.2-35.5]
No	1873 (30) [28.2-31.8]	2613 (42.4) [40.7-44.2]	3214 (53.1) [50.8-55.4]	313 (7.6) [6.3-9.1]	1053 (24.1) [22.2-26.1]	2146 (48.2) [45.9-50.4]
Don't know	__[c]	—	1063 (18.9) [16.9-21]	—	—	794 (18.5) [16.5-20.7]

[a]In column (1), the "No" category includes respondents who did not use their portal to view test results as well as those who did not access their patient portal in the past 12 months. In columns (4)-(6), patient portal users include individuals who reported accessing their patient portal or online medical record at least once in the past year. All percentages are weighted. 95% CI for point estimates reported in brackets.

[b]HCP: health care provider.

[c]not applicable.

Two-thirds of patients (66%) reported that they understood test results they viewed in their patient portal before hearing from their HCP "well" (32%, 95% CI 29% - 35.1%) or "very well" (34%, 95% CI 30.9% - 37.2%) and what those results meant for their care. In contrast, 89% of patients in total indicated it was "somewhat" (44.4%, 95% CI 41.6% - 47.3%) or "very" (44.7%, 95% CI 42% - 47.3%) easy to understand information in their patient portal and that HCPs "usually" (32.4%, 95% CI 30% - 34.9%) or "always" (57%, 95% CI 54.5% - 59.5%) explained information clearly during a health care visit (Figure 2, Table S3 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 2. Patients' understanding of test results compared to information in the patient portal and health care provider explanations, 2024. Notes: "Understand results viewed before hearing from HCP" was only asked of individuals who viewed test results in their patient portal before hearing from their health care provider (n=3432) and indicates patients who reported they understood the results very well or well. "Ease of understanding information in patient portal" was only asked of patient portal users (n=4485) and indicates patients who reported it was very or somewhat easy to understand health information in their online medical record or patient portal. "HCPs explain information clearly during health care visits?" was only asked of individuals who had a health care visit in the past year (n=5553) and indicates patients who reported their health care providers always or usually explains things in a way they can understand. All percentages are weighted. Full output reported in Table S3 in Multimedia Appendix 1. HCP: health care provider.

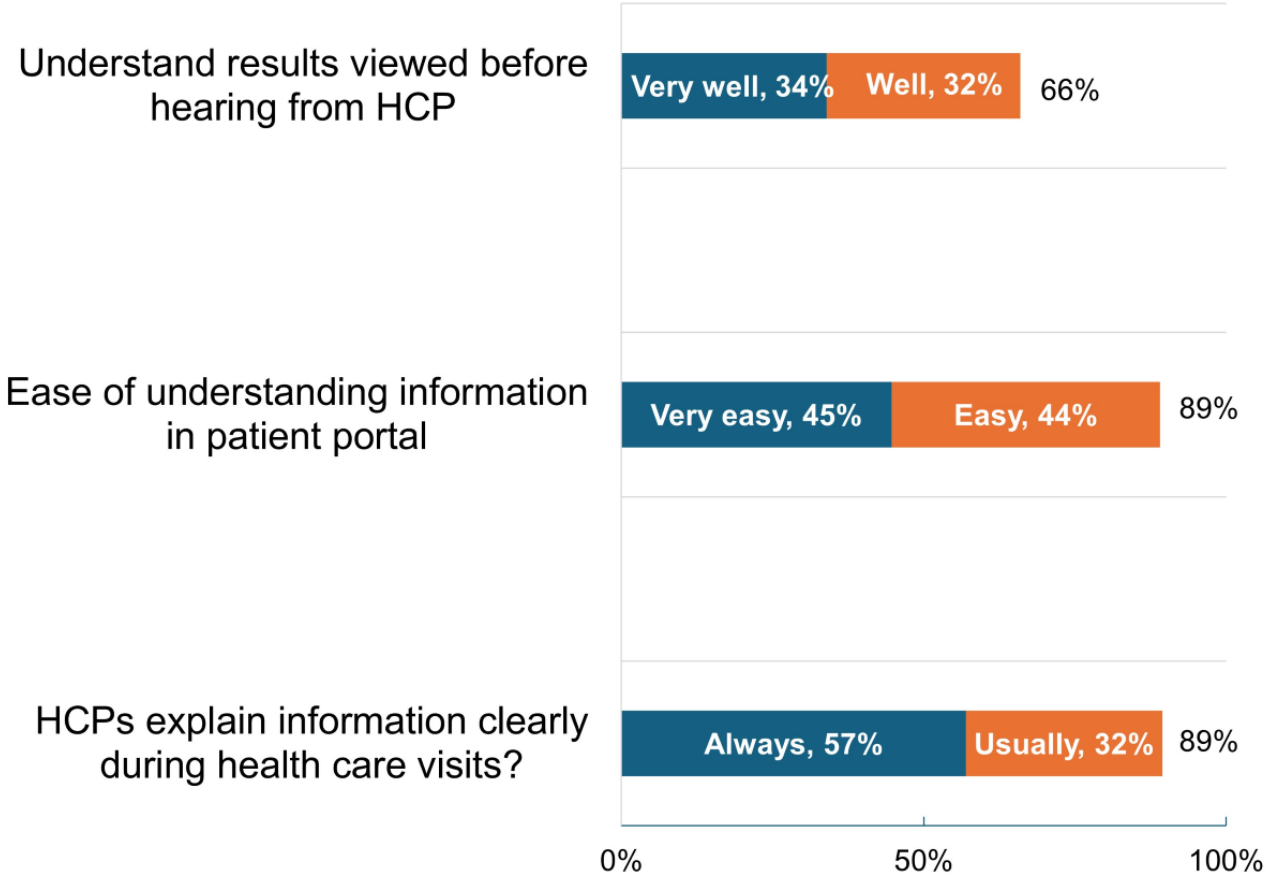

Patients and portal users aged 35 - 74 and those with higher education were significantly more likely to view test results in their portal overall and before hearing from their HCP (eg, 71.1%, 95% CI 68.4% - 73.7% of college-educated patients viewed results before hearing from their HCP vs 30.8%, 95% CI 23.6% - 38.1% of those with less than high school education, $P<.001$; Tables 2 and 3). However, confidence intervals around age estimates were wider, indicating less precise estimates.

Patients and portal users with higher levels of digital literacy were also significantly more likely to view test results in their portal overall and before hearing from their HCP compared to patients with lower digital literacy (eg, 68.7%, 95% CI 63.5% - 73.9% vs 46.1%, 95% CI 41.7% - 50.5% viewed results before hearing from their HCP, $P<.001$). See Table S4 in Multimedia Appendix 1 for the share of patients who viewed test results in their portal across digital literacy index measures.

XSL•FO
RenderX

Table . Differences in all patients' and portal users' reports of viewing test results in their patient portal, by individual characteristics, 2024[a].

Characteristics	Participants, n	All patients (n=6045)		Portal users (n=4485)	
		Weighted % (95% CI)	*P* value	Weighted % (95% CI)	*P* value
All individuals	4172	70	—[b]	92.4	—
Age (years)					
18 - 34 (ref)	662	64.9 (59.8-70)	—	87.1 (81.7-92.4)	—
35 - 49	950	78.8 (75.1-82.4)	<.001	95.2 (93.1-97.3)	.009
50 - 64	1119	71.1 (68.1-74.2)	.06	93.7 (91.7-95.6)	.03
65 - 74	920	69.9 (66-73.9)	.12	93.8 (90.7-96.9)	.048
75+	521	55.3 (48.6-62.1)	.03	90.6 (86.9-94.2)	.29
Education					
Less than high school (ref)	119	35.9 (26.2-45.6)	—	71.8 (52.2-91.4)	—
High school	486	58.1 (52.5-63.7)	<.001	92.3 (88.6-96)	.05
Some college	1202	70.6 (67.4-73.9)	<.001	92.3 (89.5-95.1)	.047
College	2365	81.7 (79.4-84)	<.001	94.5 (93.2-95.8)	.03
Digital literacy					
1 (low) (ref)	1213	56.2 (52.1-60.3)	—	89.2 (85.8-92.6)	—
2	563	62.3 (55.9-68.8)	.15	86.3 (79.4-93.1)	.46
3	1504	80.2 (76.5-83.9)	<.001	95.1 (93.1-97.1)	.006
4 (high)	888	81.2 (76.8-85.5)	<.001	95.1 (92.8-97.4)	.008
Chronic condition					
Yes	2762	72.2 (69.8-74.5)	—	93.9 (92.8-95.1)	—
No	1358	66.6 (62.6-70.6)	.03	89.7 (86.1-93.3)	.04
Recent cancer diagnosis					
Yes	231	76.7 (67.6-85.9)	—	94 (87.4-100.5)	—
No	3755	70 (68-72.1)	.18	92.4 (90.9-93.9)	.65
Encouraged to use patient portal					
Yes	3763	82.7 (81-84.4)	—	93.7 (92.3-95.1)	—
No	398	26.2 (22.4-29.9)	<.001	80 (74-86.1)	<.001

[a]Differences between groups were assessed using survey-weighted linear regression; statistical significance was evaluated using design-adjusted Wald tests. The numerator, n, is the same for all patients and portal users since, by definition, only portal users can view test results in their patient portal.

[b]not available.

XSL•FO
RenderX

Table . Differences in all patients' and portal users' reports of viewing test results in their patient portal before hearing from their HCP, by individual characteristics, 2024[a].

Characteristics	Participants, n	All patients (n=6045)		Portal users (n=4485)	
		Weighted % (95% CI)	*P* value	Weighted % (95% CI)	*P* value
All individuals	3432	57.6 (reference)	—[b]	76 (reference)	—
Age (years)					
18 - 34 (ref)	524	53.9 (48.6-59.2)	—	68.6 (63.1-74.1)	—
35 - 49	794	67.3 (63.5-71.1)	<.001	78.6 (74.6-82.5)	.006
50 - 64	918	61.8 (58.6-65)	.02	77.9 (73.7-82)	.02
65 - 74	769	61.4 (58.2-64.6)	.02	80.6 (76.7-84.5)	.002
75+	427	46.4 (39.8-52.9)	.08	72.3 (65.9-78.6)	.37
Education					
Less than high school (ref)	93	30.8 (23.6-38.1)	—	53.9 (36.9-70.9)	—
High school	395	50.4 (44.7-56.1)	<.001	77 (70.5-83.5)	.01
Some college	960	59 (55.9-62.1)	<.001	73.7 (69.6-77.7)	.03
College	1984	71.1 (68.4-73.7)	<.001	79.8 (77.4-82.2)	.003
Digital literacy					
1 (low) (ref)	948	46.1 (41.7-50.5)	—	68.2 (63.1-73.3)	—
2	478	58.9 (52.8-65)	.004	76.4 (69-83.8)	.07
3	1262	68.9 (65.3-72.4)	<.001	80.1 (76.9-83.3)	.001
4 (high)	741	68.7 (63.5-73.9)	<.001	78.4 (73.9-82.9)	.01
Chronic condition					
Yes	2296	61.9 (59.3-64.5)	—	77.6 (74.6-80.6)	—
No	1100	56.5 (52.4-60.7)	.07	73 (69-77)	.11
Recent cancer diagnosis					
Yes	203	70.8 (61.5-80.2)	—	84 (75.8-92.1)	—
No	3087	59.8 (57.9-61.6)	.03	75.8 (73.8-77.8)	.049
Encouraged to use patient portal					
Yes	3123	70.1 (68.1-72.2)	—	77.6 (75.4-79.8)	—
No	301	24.3 (20.9-27.7)	<.001	59.8 (54.2-65.4)	<.001

[a]Differences between groups were assessed using survey-weighted linear regression; statistical significance was evaluated using design-adjusted Wald tests. The numerator, n, is the same for all patients and portal users since, by definition, only portal users can view test results in their patient portal.

[b]not available.

Patients with chronic conditions were more likely to view test results in their patient portal (72.2%, 95% CI 69.8% - 74.5% vs 66.6%, 95% CI 62.6% - 70.6% for all patients, *P*=.03; 93.9%, 95% CI 92.8% - 95.1% vs 89.7%, 95% CI 86.1% - 93.3% among portal users, *P*=.04). However, they were no more or less likely to view results before hearing from their HCP. Conversely, patients with a recent cancer diagnosis were not more likely to view results in their portal overall but were more likely to view results before hearing from their HCP (70.8%, 95% CI 61.5% - 80.2% vs 59.8%, 95% CI 57.9% - 61.6% for all patients, *P*=.03; 84%, 95% CI 75.8% - 92.1% vs 75.8%, 95% CI 73.8% - 77.8% among portal users, *P*=.049). However, this difference was only marginally significant, and estimates were less precise due to the small sample of individuals with a recent cancer diagnosis.

XSL•FO
RenderX

Provider encouragement was the most strongly associated with viewing test results in the portal overall (82.7%, 95% CI 81% - 84.4% vs 26.2%, 95% CI 22.4% - 29.9% for all patients, *P*<.001; 93.7%, 95% CI 92.3% - 95.1% vs 80%, 95% CI 74% - 86.1% among portal users, *P*<.001) and before hearing from their HCP (70.1%, 95% CI 68.1% - 72.2% vs 24.3%, 95% CI 20.9% - 27.7% for all patients *P*<.001; 77.6%, 95% CI 75.4% - 79.8% vs 59.8%, 95% CI 54.2% - 65.4% among portal users, *P*<.001).

Patients with higher levels of education were more likely to report being given the option to decide whether to receive test results before hearing from their HCP; however, among portal users, there was only a moderate, significant difference among those with a college education compared to those with less than high school education (34.8%, 95% CI 31.8% - 37.7% vs 24%, 95% CI 14.3% - 33.8%; *P*=.03; Table 4).

XSL•FO
RenderX

Table . Differences in all patients' and portal users' reports of being given the option to decide how to receive test results, by individual characteristics, 2024[a].

Characteristics	All patients (n=6045)			Portal users (n=4485)		
	Patients, n	Weighted % (95% CI)	*P* value	Portal users, n	Weighted % (95% CI)	*P* value
All individuals	1768	28	__[b]	1545	33.3	—
Age (years)						
18 - 34 (ref)	272	27.5 (23.2-31.8)	—	246	32.7 (27.8-37.6)	—
35 - 49	378	29.3 (25.8-32.8)	.50	344	32.7 (28.7-36.7)	.99
50 - 64	435	25.8 (21.9-29.7)	.56	380	31.5 (26.5-36.5)	.76
65 - 74	431	34 (30.2-37.8)	.01	376	40.8 (36-45.6)	.01
75+	252	23.3 (18.8-27.7)	.19	199	31 (24.1-37.9)	.68
Education						
Less than high school (ref)	71	17.3 (12.1-22.4)	—	45	24 (14.3-33.8)	—
High school	238	25.8 (20.4-31.2)	.03	177	32.3 (26-38.5)	.16
Some college	514	27.3 (23.9-30.7)	.002	455	33.2 (29.1-37.3)	.08
College	945	31.7 (29.2-34.3)	<.001	868	34.8 (31.8-37.7)	.03
Digital literacy						
1 (low) (ref)	569	24.4 (21.5-27.2)	—	441	32.3 (28.1-36.5)	—
2	252	28.1 (23.4-32.8)	.19	222	33.6 (27.8-39.4)	.71
3	582	28.9 (25.5-32.2)	.03	546	32.8 (29.1-36.5)	.86
4 (high)	363	32.5 (27.6-37.3)	.007	334	35.2 (29.7-40.7)	.42
Chronic condition						
Yes	1218	29.1 (26.6-31.6)	—	1050	33.7 (30.9-36.4)	—
No	530	25.9 (23-28.9)	.11	479	32.3 (28.6-36)	.57
Recent cancer diagnosis						
Yes	90	27.3 (18.6-35.9)	—	78	30.8 (20.5-41)	—
No	1606	28.1 (26.2-30.1)	.85	1400	33.4 (31.2-35.6)	.61
Encouraged to use patient portal						
Yes	1551	32.5 (30.4-34.6)	—	1435	34.6 (32.4-36.9)	—
No	208	12.3 (9.9-14.7)	<.001	105	21 (14.8-27.2)	<.001

[a]Differences between groups were assessed using survey-weighted linear regression; statistical significance was evaluated using design-adjusted Wald tests.

[b]not available.

Patients with high digital literacy were also more likely to report being given the option to decide how to receive results compared to those with low digital literacy (32.5%, 95% CI 27.6% - 37.3% vs 24.4%, 95% CI 21.5% - 27.2%; *P*=.007)—though there were no significant differences among portal users. See Table S4 in Multimedia Appendix 1 for the share of patients who reported being given the option to decide how to receive results, and understanding of those results, across digital literacy index measures. Patients who reported being encouraged by their HCP to use their patient portal were significantly more likely to report being given the option to decide how to receive results (32.5%, 95% CI 30.4% - 34.6% vs 12.3%, 95% CI 9.9% - 14.7% for

XSL•FO
RenderX

all patients, *P*<.001; 34.6%, 95% CI 32.4% - 36.9% vs 21%, 95% CI 14.8% - 27.2% for portal users, *P*<.001).

Digital literacy and provider encouragement were the only factors that were significantly, positively associated with patients' understanding of results viewed before hearing from their HCP. High digital literacy was most strongly associated with understanding of results compared to low digital literacy (77.4%, 95% CI 71.5% - 83.4% vs 53%, 95% CI 47.7% - 58.4%; *P*<.001; Table 5). Individuals encouraged to use their portal by their HCP were also more likely to indicate they understood results they viewed before hearing from their HCP (67%, 95% CI 63.9% - 70% vs 54%, 95% CI 42.3% - 65.6%; *P*=.03).

Table . Differences in understanding of test results among portal users who viewed results before hearing from their health care provider, by individual characteristics, 2024[a].

Characteristics	Portal users who viewed results before hearing from HCP (n=3432)		
	Portal users, n	Weighted % (95% CI)	*P* value
All individuals	2251	65.9	__[b]
Age, years			
18 - 34 (ref)	347	64.8 (56.8-72.9)	—
35 - 49	547	66.7 (60.6-72.9)	.71
50 - 64	603	68.6 (63.3-74)	.43
65 - 74	496	62.7 (56.6-68.9)	.69
75+	258	61.2 (53.2-69.3)	.49
Education			
Less than high school (ref)	50	62.4 (46.5-78.3)	—
High school	246	65.3 (57.9-72.7)	.72
Some college	596	62.9 (57.4-68.3)	.95
College	1359	69.2 (65.6-72.7)	.41
Digital literacy			
1 (low) (ref)	513	53 (47.7-58.4)	—
2	302	59.9 (50.9-69)	.17
3	868	68.9 (64.5-73.2)	<.001
4 (high)	566	77.4 (71.5-83.4)	<.001
Chronic condition			
Yes	1472	64.4 (60.3-68.4)	—
No	756	68.7 (63.6-73.7)	.17
Recent cancer diagnosis			
Yes	130	66.6 (52.8-80.4)	—
No	2031	66 (62.6-69.4)	.94
Encouraged to use patient portal			
Yes	2077	67 (63.9-70)	—
No	169	54 (42.3-65.6)	.03

[a]Differences between groups were assessed using survey-weighted linear regression; statistical significance was evaluated using design-adjusted Wald tests.

[b]not available.

Discussion

Principal Findings

In a large, representative national survey, 70% of patients nationally viewed test results in their patient portal, and more than half (58%) viewed test results before discussing them with their HCP. However, only about a quarter (28%) indicated they were given the option to decide when to receive results or how these results were communicated to them. Rates of both viewing test results and being given the option to decide were lower among individuals with lower levels of education and digital literacy, which may be driven in part by differences in patient

XSL•FO
RenderX

portal access more broadly [20], as was self-reported understanding of test results viewed before a discussion with their HCP. Only two-thirds of patients nationally reported they understood results they viewed before hearing from a provider and what those results meant for their care, which was lower than patient reports of general understanding of information in the patient portal and HCP explanations during health care visits (89%, respectively).

Patients who reported being encouraged to use their patient portal by their HCP were more likely to view immediately released results and indicate they understood what the results showed and what they meant for their care, which suggests anticipatory guidance from HCPs may help set expectations before results are delivered [21,22]. Together, these findings indicate that patients overall are highly engaged with results released immediately via the patient portal, but there may be missed opportunities to address differences in patient preferences around notification of newly released results.

In response to the information blocking regulations, many HCPs provided near-immediate availability of information in the patient record and portal as their organizational default, when technological capabilities allow [23]. The high rates of access to that information shown here, coupled with evidence on patient views from prior studies, suggest the importance of maintaining patients' ability to engage with their health data [15,23,24]. At the same time, the information blocking regulations allow HCPs to grant a patient's request to delay the release of their test results and allow patients to choose whether to receive notifications (eg, text messages) when results are available [22,25]. However, anecdotally, we have found that most vendor systems do not routinely offer functionality that allows patients to specify their preferred timing for the release of specific test results without custom configuration, which may limit opportunities for patients to indicate their preferences around the timing of result release in their patient portal.

Our findings indicate that while most patients nationally viewed immediately released results, only one-third reported being given the option to decide whether to receive results before discussing with their HCP. Interestingly, patients with a recent diagnosis of cancer were no more likely to indicate that they had been given an opportunity to decide when to receive results despite widespread concerns over the potential negative impact news of a cancer diagnosis might have on patient well-being [26,27], which has already led some states to propose legislation delaying the release of cancer results [28,29]. These findings speak to potential missed opportunities to reduce worry by offering options on how patients would like to receive results. HCPs, with support from system developers, could help address this shortcoming by making patients aware of options provided through the patient portal to indicate their preferences around the notification or release of results, through precounseling with patients during health care visits or regular communication via the portal, and through discussions with support staff [30].

Fewer patients indicated they understood test results they viewed before discussing with their HCP compared to the share of patients who indicated their HCPs explained information clearly during health care visits and that stated it was easy to understand information in their patient portal. These findings suggest test results may be more difficult to interpret or contextualize [31]. Consistent with prior work [32], patients with lower levels of digital literacy across self-reported measures were even less likely to report that they understood test results they viewed before discussing them with their HCP, which could lead to undue anxiety or worry about test results delivered without context [7]. Patients who reported being encouraged by their HCP to use their patient portal were also more likely to report being given the option to decide how to receive test results and to indicate they understood results viewed in their portal before discussing with their HCP, which suggests discussions about the patient portal may present an opportunity to discuss communication preferences or what different results would mean for subsequent care. There are likely many options to improve patient understanding of their results: discussions with HCPs prior to the availability of test results, clear handouts preparing patients for results, displaying the result in the context of a standard range, including links to definitions of medical terms, or even providing artificial intelligence-based tools to answer questions and alleviate burden on HCPs [22,33-36]. Each of these strategies merits investigation as health care moves toward a more digital and asynchronous delivery of care. Caregiver or proxy-user perspectives may also represent an important future research area, as these users often engage with patient portals and make decisions on behalf of patients [6,19,37].

Limitations

This study has several limitations. First, survey data is self-reported and is therefore subject to recall bias. We tried to mitigate this source of bias by limiting analyses to respondents who had test results in the past year. To the extent that recall bias persists, it may indicate that current strategies related to the provision of options on when to receive test results are not sufficient to lead to informed decision-making by patients and leave a lasting impression. Second, HCP conversations as well as patient preferences regarding how and when information is communicated to them may depend on the sensitivity of the results. In this study, we were not able to distinguish between normal and abnormal test results, and whether this affected communication preferences or being given the option to decide how to receive results. However, we observed no difference between patients with a recent cancer diagnosis and others. Further, previous work suggests preferences for immediate release persist among patients with abnormal results, even if access is associated with increased worry, and that patient worry around test results may be driven by individual attitudes and preferences more so than the test itself [9,38]. Third, our measure of patients' ease of understanding portal information may be influenced by their ability to review test results – the most common use of patient portals [19]. This may make it difficult to separate general comprehension of portal content from reactions to viewing results. Finally, our measure of patients' understanding of results viewed before discussing with an HCP was subjective and may represent an overestimate of patients' understanding of what results show and what they meant for their care. Prior work has shown that viewing immediately released test results was associated with a near

XSL•FO
RenderX

doubling of messages patients send to their clinician, which suggests immediate access to test results does not diminish—and conversely, may increase—the need for HCPs to help interpret these results and what they mean for their patients' care [23].

Conclusions

In 2024, most patients who received test results reported viewing those results in their patient portal before hearing from their HCP, signaling a strong revealed preference for immediate release of test results. Only one-third of patients reported being given the option to decide how results would be communicated to them, indicating missed opportunities to avoid worry for those who did not want to see results before discussing with their HCP. Two-thirds of patients who viewed immediately released results indicated they understood what the results showed and what they meant for their care—substantially less than indicated other forms of communication were clear. In response to the information blocking regulations, many HCPs released test results to patients immediately upon their availability as the organizational default, and our findings suggest that, consistent with prior studies, patients prefer to access their results as soon as they are available. However, incorporating patient communication preferences and improving how results are presented in the portal may help to avoid undue worry about sensitive results while ensuring that patients have the tools and resources to access their health information and make informed decisions about their health and care.

Funding

The authors declared no financial support was received for this work.

Conflicts of Interest

CR and JE have nothing to disclose. BS reports personal fees from the Office of the National Coordinator for Health Information Technology during the conduct of the study; grants from National Institute on Aging outside the submitted work.

Multimedia Appendix 1
Sample characteristics, survey questions to create key measures, and additional results.
[DOCX File, 30 KB - jmir_v28i1e94098_app1.docx]

References

1. HIPAA for individuals. Your Medical Records. URL: https://www.hhs.gov/hipaa/for-individuals/medical-records/index.html [accessed 2025-03-25]
2. Steitz BD, Wong JIS, Cobb JG, Carlson B, Smith G, Rosenbloom ST. Policies and procedures governing patient portal use at an Academic Medical Center. JAMIA Open 2019 Dec;2(4):479-488. [doi: 10.1093/jamiaopen/ooz039] [Medline: 32025645]
3. Office of the National Coordinator for Health Information Technology. 21st Century Cures Act: interoperability, information blocking, and the ONC Health IT Certification Program. Fed Regist 2020;85(85):25642-25961 [FREE Full text]
4. When would a delay in fulfilling a request for access, exchange, or use of EHI be considered an interference under the information blocking regulation? Office of the National Coordinator for Health Information Technology. 2022. URL: https://www.healthit.gov/faq/when-would-delay-fulfilling-request-access-exchange-or-use-ehi-be-considered-interference-under [accessed 2025-04-29]
5. Pollock JR, Petty SAB, Schmitz JJ, Varner J, Metcalfe AM, Tan N. Patient access of their radiology reports before and after implementation of 21st century cures act information-blocking provisions at a large multicampus health system. AJR Am J Roentgenol 2024 Jun;222(6):38534191. [doi: 10.2214/AJR.23.30343] [Medline: 38534191]
6. Reynolds TL, Cobb JG, Steitz BD, Ancker JS, Rosenbloom ST. The state-of-the-art of patient portals: adapting to external factors, addressing barriers, and innovating. Appl Clin Inform 2023 Aug;14(4):654-669. [doi: 10.1055/s-0043-1770901] [Medline: 37611795]
7. Giardina TD, Baldwin J, Nystrom DT, Sittig DF, Singh H. Patient perceptions of receiving test results via online portals: a mixed-methods study. J Am Med Inform Assoc 2018 Apr 1;25(4):440-446. [doi: 10.1093/jamia/ocx140] [Medline: 29240899]
8. Friedman D. Your medical test results are available but do you want to view them. The New York Times. 2022. URL: https://www.nytimes.com/2022/10/03/well/live/medical-test-results-cures-act.html [accessed 2026-06-04]
9. Steitz BD, Turer RW, Lin CT, et al. Perspectives of patients about immediate access to test results through an online patient portal. JAMA Netw Open 2023 Mar 1;6(3):e233572. [doi: 10.1001/jamanetworkopen.2023.3572] [Medline: 36939703]
10. Turer RW, DesRoches CM, Salmi L, Helmer T, Rosenbloom ST. Patient perceptions of receiving COVID-19 test results via an online patient portal: an open results survey. Appl Clin Inform 2021 Aug;12(4):954-959. [doi: 10.1055/s-0041-1736221] [Medline: 34644805]

XSL•FO
RenderX

11. Patel A, Kawpeng I, Potashner R, Jessa K, Yan AP. Preferences of pediatric oncology patients and caregivers on the availability of patient results in an online patient portal. JCO Clin Cancer Inform 2025 Apr;9:e2400235. [doi: 10.1200/CCI-24-00235] [Medline: 40233297]
12. Bruno B, Steele S, Carbone J, Schneider K, Posk L, Rose SL. Informed or anxious: patient preferences for release of test results of increasing sensitivity on electronic patient portals. Health Technol (Berl) 2022;12(1):59-67. [doi: 10.1007/s12553-021-00628-5] [Medline: 35036280]
13. Sisk BA, Bereitschaft C, Enloe M, Schulz G, Mack J, DuBois J. Oncology clinicians' perspectives on online patient portal use in pediatric and adolescent cancer. JCO Clin Cancer Inform 2023 Sep;7:e2300124. [doi: 10.1200/CCI.23.00124] [Medline: 37972324]
14. Leonard LD, Himelhoch B, Huynh V, et al. Patient and clinician perceptions of the immediate release of electronic health information. Am J Surg 2022 Jul;224(1 Pt A):27-34. [doi: 10.1016/j.amjsurg.2021.12.002] [Medline: 34903369]
15. Steitz BD, Padi-Adjirackor NA, Griffith KN, Reese TJ, Rosenbloom ST, Ancker JS. Impact of notification policy on patient-before-clinician review of immediately released test results. J Am Med Inform Assoc 2023 Sep 25;30(10):1707-1710. [doi: 10.1093/jamia/ocad126] [Medline: 37403329]
16. Health information national trends survey 7 (HINTS 7) methodology report. : National Cancer Institute; 2025 URL: https://hints.cancer.gov/docs/methodologyreports/HINTS_7_MethodologyReport.pdf [accessed 2026-06-04]
17. Institutional review board (IRB) approvals for the health information national trends survey (HINTS). National Cancer Institute, Health Information National Trends Survey. URL: https://hints.cancer.gov/about-hints/institutional-review-board.aspx [accessed 2026-06-04]
18. How to use HINTS data. National Cancer Institute. URL: https://hints.cancer.gov/data/download-data.aspx [accessed 2023-04-18]
19. Richwine C. Individuals' access and use of patient portals and smartphone health apps. In: ASTP Health IT Data Brief: Office of the Assistant Secretary for Technology Policy; 2024. [Medline: 40663642]
20. Richwine C, Johnson C, Patel V. Disparities in patient portal access and the role of providers in encouraging access and use. J Am Med Inform Assoc 2023 Jan 18;30(2):308-317. [doi: 10.1093/jamia/ocac227] [Medline: 36451262]
21. Zipursky AR, Goldbloom E, Jessa K, et al. Anticipatory guidance and follow-up regarding online access to test results in the Canadian pediatric setting. Paediatr Child Health 2025 Nov;30(7):540-544. [doi: 10.1093/pch/pxaf053] [Medline: 41323941]
22. Carey E, Salmi L. Tips for helping patients navigate immediately-available test results. Center to Advance Palliative Care. 2023. URL: https://www.capc.org/blog/tips-for-helping-patients-navigate-immediately-available-test-results/ [accessed 2026-06-04]
23. Steitz BD, Sulieman L, Wright A, Rosenbloom ST. Association of immediate release of test results to patients with implications for clinical workflow. JAMA Netw Open 2021 Oct 1;4(10):e2129553. [doi: 10.1001/jamanetworkopen.2021.29553] [Medline: 34661667]
24. Turer RW, McDonald SA, Lehmann CU, et al. Real-time electronic patient portal use among emergency department patients. JAMA Netw Open 2024 May 1;7(5):e249831. [doi: 10.1001/jamanetworkopen.2024.9831] [Medline: 38700859]
25. Can an actor grant a patient's request to delay the release of a patient's test result(s) (e.g., laboratory or image result(s)) to the patient without implicating the information blocking regulations? Office of the National Coordinator for Health IT. 2022. URL: https://www.healthit.gov/faq/can-actor-grant-patients-request-delay-release-patients-test-results-eg-laboratory-or-image [accessed 2025-04-25]
26. Bhalla S, Prasad T, Xie D, Gerber DE. Contemporary trends in reviewing test results through the electronic patient portal among patients with cancer. JAMA Oncol 2024 Jan 1;10(1):139-140. [doi: 10.1001/jamaoncol.2023.5047] [Medline: 38032648]
27. Winget M, Haji-Sheikhi F, Brown-Johnson C, et al. Electronic release of pathology and radiology results to patients: opinions and experiences of oncologists. J Oncol Pract 2016 Aug;12(8):e792-e799. [doi: 10.1200/JOP.2016.011098] [Medline: 27382001]
28. Gerber DE. Taking it to the states: adapting information blocking legislation to oncology. J Clin Oncol 2023 Sep 20;41(27):4348-4351. [doi: 10.1200/JCO.23.01108] [Medline: 37410962]
29. Pierce A. TMA's first 2025 legislative victory protects sensitive test results. Texas Medical Association. 2025. URL: https://www.texmed.org/Template.aspx?id=66152 [accessed 2026-06-04]
30. Salmi L, Hubbard J, McFarland DC. When bad news comes through the portal: strengthening trust and guiding patients when they receive bad results before their clinicians. Am Soc Clin Oncol Educ Book 2024 Jun;44(3):e433944. [doi: 10.1200/EDBK_433944] [Medline: 38848509]
31. Fraccaro P, Vigo M, Balatsoukas P, et al. Presentation of laboratory test results in patient portals: influence of interface design on risk interpretation and visual search behaviour. BMC Med Inform Decis Mak 2018 Feb 12;18(1):11. [doi: 10.1186/s12911-018-0589-7] [Medline: 29433495]
32. Zikmund-Fisher BJ, Exe NL, Witteman HO. Numeracy and literacy independently predict patients' ability to identify out-of-range test results. J Med Internet Res 2014 Aug 8;16(8):e187. [doi: 10.2196/jmir.3241] [Medline: 25135688]

XSL•FO
RenderX

33. van der Mee FAM, Schaper F, Jansen J, Bons JAP, Meex SJR, Cals JWL. Enhancing patient understanding of laboratory test results: systematic review of presentation formats and their impact on perception, decision, action, and memory. J Med Internet Res 2024 Aug 12;26:e53993. [doi: 10.2196/53993] [Medline: 39133906]
34. Witteman HO, Zikmund-Fisher BJ. Communicating laboratory results to patients and families. Clin Chem Lab Med 2019 Feb 25;57(3):359-364. [doi: 10.1515/cclm-2018-0634] [Medline: 30407910]
35. Zhang Z, Kmoth L, Luo X, He Z. User-centered system design for communicating clinical laboratory test results: design and evaluation study. JMIR Hum Factors 2021 Nov 25;8(4):e26017. [doi: 10.2196/26017] [Medline: 34842529]
36. Klatt EC. Cognitive factors impacting patient understanding of laboratory test information. J Pathol Inform 2024 Dec;15:100349. [doi: 10.1016/j.jpi.2023.100349] [Medline: 38075016]
37. Han S. Facilitating patient portal shared access to support age-friendly health care. JAMA Netw Open 2025 Feb 3;8(2):e2461814. [doi: 10.1001/jamanetworkopen.2024.61814] [Medline: 39998834]
38. Steitz BD, Turer RW, Salmi L, et al. Repeated access to patient portal while awaiting test results and patient-initiated messaging. JAMA Netw Open 2025 Apr 1;8(4):e254019. [doi: 10.1001/jamanetworkopen.2025.4019] [Medline: 40198070]

Abbreviations

HCP: health care provider
HINTS: Health Information National Trends Survey
NIH: National Institutes of Health

Edited by S Brini; submitted 24.Feb.2026; peer-reviewed by HN Platter, J Hagström; revised version received 08.May.2026; accepted 12.May.2026; published 12.Jun.2026.

Please cite as:
Richwine C, Steitz B, Everson J
Patient-Reported Experiences With Viewing and Understanding Test Results in Patient Portals: Cross-Sectional Survey Analysis
J Med Internet Res 2026;28:e94098
URL: https://www.jmir.org/2026/1/e94098
doi:10.2196/94098

© Chelsea Richwine, Bryan Steitz, Jordan Everson. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Impact of Financial Incentives on Electronic Health Record–Driven Recruitment of Underrepresented Communities in Research: Randomized Controlled Trial

Christina Ping[1], BS; Andrew A Davis[2], MD; Waishun Gema[2], BS; Foluso Ademuyiwa[2], MD, MPH, MSCI; Mark Fiala[2,3], PhD, MSW

[1]School of Medicine, Washington University in St. Louis, Saint Louis, MO, United States

[2]Division of Oncology, School of Medicine, Washington University in St. Louis, 660 S Euclid Ave, Saint Louis, MO, United States

[3]School of Public Health, Washington University in St. Louis, Saint Louis, MO, United States

Corresponding Author:
Mark Fiala, PhD, MSW
Division of Oncology, School of Medicine, Washington University in St. Louis, 660 S Euclid Ave, Saint Louis, MO, United States

Abstract

Background: Progress in clinical trial research depends on effective recruitment to ensure adequate sample sizes and generalizability of results. Electronic health record (EHR)–based recruitment has been suggested as a potential method to address historical imbalances in trial populations, including underrepresentation of certain racial and ethnic groups. However, disparities in EHR portal uptake and use exist, particularly among these same underrepresented groups. Thus, we hypothesized that EHR-driven recruitment for research may continue to exacerbate these dynamics.

Objective: In this study, we evaluated whether a financial incentive improved response rates to recruitment messages delivered through an EHR portal in an equitable fashion.

Methods: A random sample of 1200 patients with diagnoses of breast, colorectal, prostate, or lung cancer and active EHR patient portal accounts received an automated recruitment message for a fictitious cross-sectional survey study, “The Social Aspects of Cancer Care,” via their EHR portal. Two strata of patients were sampled, those who identified as non-Hispanic White individuals (n=600, 50%) and those who identified as part of a historically underrepresented racial or ethnic group (n=600, 50%), and randomized to receive an offer of a US $20 monetary incentive (n=400, 33.3%) or no incentive (n=800, 66.7%). Patients were given 14 days to respond.

Results: Overall, 39.3% (471/1200) of participants viewed the recruitment message, and 13.3% (159/1200) responded that they were interested. Response rates were higher ($P<.001$) among non-Hispanic White patients (108/600, 18%) than among African American or Black patients (32/456, 7%) and members of other racial or ethnic groups (18/144, 12.5%). Older patients (≥65 years) were less likely to view the recruitment message (244/720, 33.9% vs 230/480, 47.9%; $P=.001$) and respond (72/720, 10% vs 82/480, 17.1%; $P=.002$). Sex was not significantly associated with viewing the message or response outcomes ($P=.45$ and $P=.19$, respectively). No differences were observed in response rate for the incentive cohort compared to the cohort with no incentive (60/400, 15% vs 96/800, 12%; $P=.21$).

Conclusions: Patients with cancer who identify as part of a historically underrepresented group were less likely to view and respond to recruitment messages for research studies through their EHR patient portal. Age was also found to significantly impact engagement, with older patients less likely to view and respond. Finally, a modest US $20 financial incentive was not found to significantly impact patient engagement.

(J Med Internet Res 2026;28:e86110) doi:10.2196/86110

KEYWORDS

underrepresented racial or ethnic groups; electronic health records; research participation; patient portals; cancer

Introduction

Progress in clinical research relies on effective recruitment to ensure adequate sample sizes for statistical analyses and appropriate demographic makeups to support the generalizability of the final results [1]. Conventional recruitment strategies such as mail or phone contact or in-person recruitment with coordinators are costly and time intensive, which limits scalability [2-4]. Electronic health record (EHR)–based

XSL•FO
RenderX

recruitment strategies have emerged as potential scalable tools to identify and contact eligible patients asynchronously [5].

Asynchronous digital techniques have been used for a variety of health-related campaigns [6]. Prior studies have used EHR portals to improve uptake for preventative screening [7], trigger vaccination reminders [8-10], and support research participant recruitment [11-13]. These studies have also demonstrated that using EHR portals to improve research engagement and participation is a cost-effective, easy-to-implement method of recruiting patients [11]. Some propose that EHR-based recruitment may be a solution to address underrepresentation of certain genders, races, or ethnic groups in clinical trials [12,13]. However, there is a concern that EHR-based recruitment strategies may exacerbate the well-documented disparities that exist among clinical research participants [14]. Data have shown that trial participants are more likely to be White individuals and older than the general population [15]. There is also evidence suggesting that disparities exist in patient virtual care and portal use by age, race, and socioeconomic status [16-19]. Missing from the literature are studies of EHR-based recruitment strategies that have compared the recruited patient sample to the overall population to evaluate whether EHR recruitment can actually address underrepresentation in trial participants. Furthermore, it is still unclear whether disparities in clinical trial engagement and participation based on race exist with EHR-based recruitment strategies. Some studies have reported that Black patients are equally likely to participate when asked and underrepresentation stems from lower access to research [20,21]. Another recent study found that Asian and Hispanic patients are less likely to view recruitment messages but Black and male patients are less likely to actually participate [22]. Thus, whether EHR-based recruitment is an aggravator of or solution to disparities in underrepresentation of these groups is still unknown.

Independent of EHR-based recruitment, financial incentives are also a strategy that has been shown to improve research engagement and participation [23]. Previous studies have found that financial incentives can improve response rates by a median of 12% [24], with the highest impact observed with direct monetary incentives [25]. However, their role in enhancing the representativeness of participant samples, particularly for EHR-driven studies, remains uncertain.

This study aimed to evaluate the effect of race and financial incentives on engagement with EHR portal recruitment messages among patients with and survivors of cancer. We hypothesized that (1) patients from historically underrepresented racial or ethnic groups that are enrolled in the EHR portal would be less likely to view and respond to recruitment messages and (2) a US $20 incentive would improve engagement, particularly among these patients.

Methods

Study Design and Participants

This randomized controlled study was conducted at the Siteman Cancer Center (SCC) under protocol 202402076 through the Hope Foundation for Cancer Research Pilot Grant Program. The SCC is a National Cancer Institute–designated Comprehensive Cancer Center that serves a geographically diverse catchment area, spanning across eastern Missouri and western Illinois. Approximately 22% of the SCC's catchment identifies as being part of a historically underrepresented racial or ethnic group [26].

Eligible participants were defined as adults (≥18 years of age) with a diagnosis of breast, colorectal, lung, or prostate cancer documented in the EHR. These diagnoses were chosen because they are the 4 most common cancers in the United States. Only patients with active EHR portal accounts as of July 2024 and at least one clinic visit with either a medical or radiation oncologist within the prior 60 days were included.

We identified 5302 eligible patients, of whom 4399 (83%) identified as non-Hispanic White individuals and 903 (17%) identified as belonging to another racial or ethnic group (African American or Black, Hispanic, Asian, or other or unknown). From each race or ethnicity stratum, 600 patients were randomly selected. The purpose of oversampling patients who identified as being part of a historically underrepresented racial or ethnic group was to ensure that an effect size, if it existed, could be observed between the 2 groups. The primary objective of this study was to determine whether race impacted engagement with the recruitment message. With an α of .05 and 80% power, 800 patients were required to detect a difference of more than 5% in response rates. The secondary objective was to evaluate whether incentives changed these response rates. On the basis of a priori power calculations, we estimated that a sample size of 400 patients in the incentive cohort and 800 in the control cohort would allow us to determine whether there was more than a 10% increase in response rate between each incentive level compared to no incentives, with 80% power at a one-sided .05 α assuming an approximately 10% response rate in the control cohort. Thus, 66.7% (800/1200) of the recruited patients were randomized to receive no incentive, with the remaining 33.3% (400/1200) receiving a US $20 incentive alongside the recruitment message.

Randomization was completed via a random number generator and facilitated by author WG.

Intervention

Patients received a standardized EHR portal message inviting them to participate in a fictitious cross-sectional survey–based noninterventional study, "The Social Aspects of Cancer Care." Patients received an email notifying them of the invitation, and an alert was displayed on their landing page when they logged in to their EHR portal. Portal messages included details about the voluntary nature of participation, estimated survey duration, and incentive (if applicable). All patients who responded by clicking a button within the message stating "I'm interested" were subsequently debriefed about the sham nature of the study and given the US $20 incentive regardless of incentive group assignment. No additional data were collected from the patients who clicked on the link. Images of the email notification and portal alert can be found in Multimedia Appendix 1. Patients were given 14 days to respond to the invitation.

XSL•FO
RenderX

Outcomes

The primary outcomes included message view rate, which was defined as the proportion of patients who opened the message in their patient portal, and message response rate, which was defined as the proportion of patients who responded to the message by clicking the "I'm interested" button embedded in the message. The secondary analyses examined demographic factors associated with engagement, as well as the potential impact of the incentive on the primary outcomes.

Statistical Analysis

Descriptive and bivariate statistics were used to compare demographic characteristics by incentive and race groups. The primary end points were compared using Fisher exact tests. Multivariable logistic regression was also used to evaluate the independent associations among race, incentive, age, sex, and response outcomes. Adjusted odds ratios (aORs) and 95% CIs were reported. Statistical significance was defined as a *P* value of less than .05. Analyses were conducted using SAS Enterprise Guide (version 8.3; SAS Institute Inc).

Ethical Considerations

This study was approved and supervised by the Washington University Human Research Protection Office. As participation was deemed to be of minimal risk and obtaining informed consent would have biased study results, a waiver of consent was granted.

Results

Participant Characteristics

A total of 1200 patients were randomized to receive a recruitment message through the EHR portal. The mean age was 65.7 (SD 12.3) years, and 54% (n=648) were female. Half (n=600, 50%) of the cohort identified as non-Hispanic White individuals, whereas 38% (n=456) identified as African American or Black individuals, and 12% (n=144) identified as being of other or unknown race or ethnicity. The most common diagnosis was breast cancer at 31% (n=372) of all patients. Prostate, lung, and colorectal cancer constituted 28% (n=336), 25% (n=300), and 16% (n=192) of patients, respectively. There were no statistically significant differences in age, sex, cancer type, or race distribution between incentive groups (*P*>.05 in all cases). Table 1 provides the detailed demographic characteristics of the study participants.

Table . Participant characteristics (N=1200).

	Overall cohort	No incentive (n=800)	US $20 incentive (n=400)	*P* value
Age (y), mean (SD)	65.7 (12.3)	65.7 (12.0)	65.7 (12.9)	
<65, n (%)	480 (40)	328 (41)	156 (39)	.99
≥65, n (%)	720 (60)	472 (59)	248 (62)	.45
Race, n (%)				.83
African American or Black	456 (38)	301 (38)	155 (39)	
Non-Hispanic White	600 (50)	400 (50)	200 (50)	
Other or unknown	144 (12)	99 (12)	45 (11)	
Sex, n (%)				.03
Male	552 (46)	392 (49)	168 (42)	
Female	648 (54)	416 (52)	232 (58)	
Cancer type, n (%)				.63
Breast	372 (31)	238 (31)	124 (31)	
Colorectal	192 (16)	136 (17)	60 (15)	
Lung	300 (25)	192 (24)	108 (27)	
Prostate	336 (28)	232 (29)	112 (28)	

Participant Flow

Figure 1 shows the participant flow diagram. Message delivery was successful for 97.8% (1174/1200) of the participants. Unsuccessful deliveries were due to patient deaths (23/1200, 1.9%) or changes in account settings between participant identification and the message being sent (3/1200, 0.3%). Overall, 39.3% (471/1200) of the patients viewed the recruitment message, and 13.3% (159/1200) of the patients responded by clicking on the "I'm interested" button embedded in the message. Among those who viewed the message, the conditional response rate was 33.8% (159/471).

XSL•FO
RenderX

Figure 1. Participant flow.

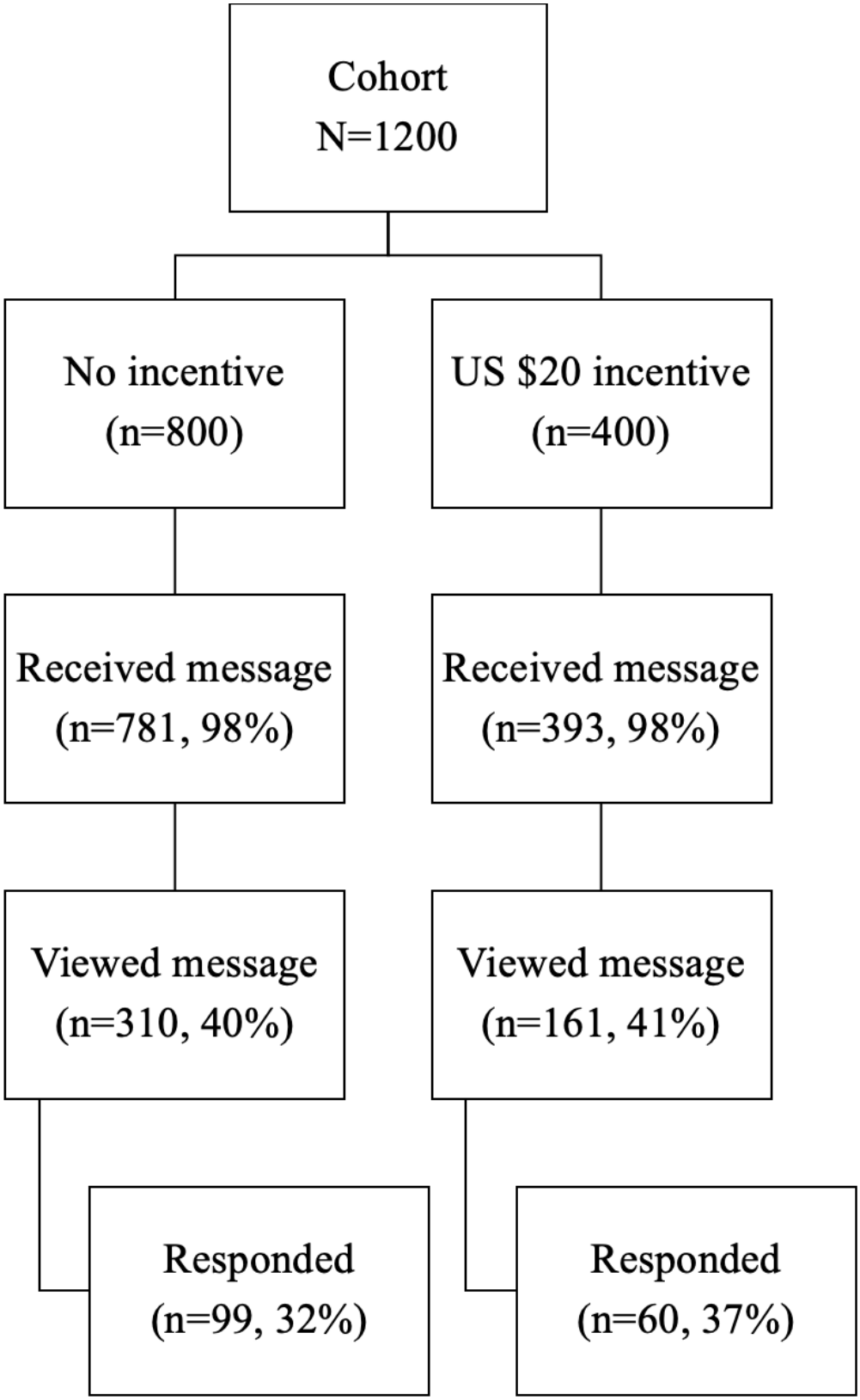

Relationship Between Demographics and Response Rates

Table 2 details the results of our multivariable analyses on the effects of sex, race, and age on the primary outcome measures. Race was strongly associated with response to the recruitment message. Response rates were significantly higher (P<.001) among non-Hispanic White patients (108/600, 18%) than among non-White participants (60/600, 10%). Among non-White participants, African American or Black participants (32/456,

XSL•FO
RenderX

7%) had lower response rates than members of other racial or ethnic groups (18/144, 13%; *P*=.02). This finding was consistent for patients who did not receive an offer of an incentive (72/400, 18% in non-Hispanic Whites compared to 18/301, 6% in African American or Black and 10/99, 10% in other; *P*<.001) and for those who did (38/200, 19% in non-Hispanic Whites compared to 14/155, 9% in African American or Black and 9/45, 20% in other; *P*=.03). In multivariable logistic regression adjusting for age, sex, and incentive assignment, patients from historically excluded or underrepresented racial or ethnic groups had significantly lower odds of viewing the message (aOR 0.39, 95% CI 0.31 - 0.50; *P*<.001) and responding (aOR 0.38, 95% CI 0.27 - 0.55; *P*<.001). Among patients who opened the message, the conditional response rate remained significantly higher among non-Hispanic White patients than among patients from historically underrepresented racial or ethnic groups (110/275, 40% vs 51/232, 22%; *P*<.001).

Table . Multivariable analysis of factors impacting patient response.

	Message viewed (n=471 of 1200)		Responded (n=159 of 1200)		Responded (n=159 of the 471 who viewed the message)	
	aOR[a] (95% CI)	*P* value	aOR (95% CI)	*P* value	aOR (95% CI)	*P* value
Incentive		.99		.21		.20
US $0	Reference		Reference		Reference	
US $20	1.07 (0.88-1.78)		1.25 (0.88-1.78)		1.31 (0.87-1.96)	
Race		<.001		<.001		.04
Non-Hispanic White	Reference		Reference		Reference	
Other	0.39 (0.31-0.50)		0.38 (0.27-0.55)		0.64 (0.42-0.97)	
Age (y)		.001		.002		.06
<65	Reference		Reference		Reference	
≥65	0.66 (0.52-0.85)		0.57 (0.40-0.81)		0.68 (0.45-1.02)	
Sex		.45		.19		.28
Female	Reference		Reference		Reference	
Male	0.91 (0.71-1.16)		0.79 (0.55-1.12)		0.80 (0.53-1.20)	

[a]aOR: adjusted odds ratio.

Older patients (≥65 years) were less likely to view the recruitment message (244/720, 34% vs 230/480, 48%; *P*=.001) and respond (72/720, 10% vs 82/480, 17%; *P*=.002). In multivariable models, older age remained associated with lower odds of message viewing (aOR 0.66, 95% CI 0.52 - 0.85; *P*=.001). Sex was not significantly associated with viewing the message or with response outcomes.

Effect of Incentives

Offering a US $20 incentive modestly increased the overall response rate (60/400, 15% vs 99/800, 12% in the no-incentive group), but this difference was not statistically significant (*P*=.21; Figure 2). Among patients who viewed the message, the response rate was similar across incentive groups (61/164, 37% vs 102/320, 32% in the no-incentive group; *P*=.20).

XSL•FO
RenderX

Figure 2. Response rate by race and incentive. The *P* values provided are one sided. AA: African American; NHW: non-Hispanic White.

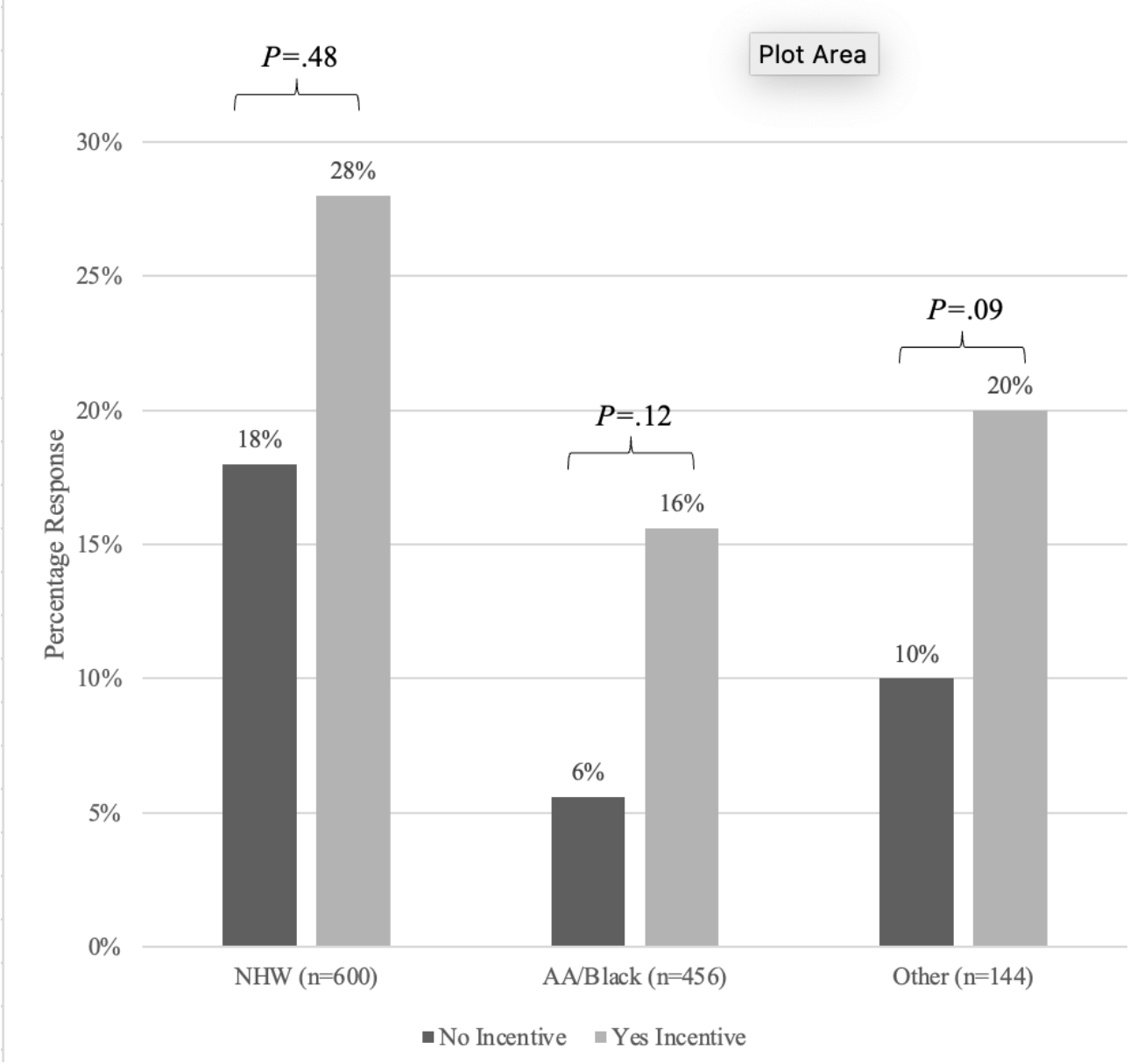

Exploratory subgroup analyses examined whether the incentive effect differed by race. Among patients from historically underrepresented or excluded racial or ethnic groups, the response rate increased from 7% (27/400) in the no-incentive group to 13% (26/200) in the incentive group. Among non-Hispanic White patients, the response rate was similar at 18% (72/400) in the no-incentive group and 19% (38/200) in the incentive group. However, there was no statistically significant interaction between race and incentive assignment (interaction *P*=.12).

Discussion

Principal Findings

We performed a randomized trial to examine the impact of EHR-based clinical trial recruitment with financial incentives on improving equitable representation in cancer research. We found that, among patients with and survivors of cancer who had active patient portal accounts, those who identified as part of historically underrepresented racial and/or ethnic groups were less likely to view and respond to recruitment messages for research studies, as we hypothesized. The results indicate that EHR-based recruitment may contribute to disparities in recruitment-stage engagement. In this study, we also found that the impact of the US $20 incentive on response rates was modest, suggesting that small, monetary incentives are insufficient to meaningfully improve recruitment. While there was a trend toward greater improvement in response rates after incentives among patients of other racial and ethnic backgrounds (increasing from 27/400, 7% to 26/200, 13% with the incentive), this did not reach significance in our sample (*P*=.12). Of note, this study was not specifically powered to detect the effect of incentives among different racial groups. However, this suggests that incentives may have differential effects depending on cultural context, perceived relevance of the study, or financial burden among underrepresented populations. Larger incentive amounts or more targeted financial incentives could be more effective in mitigating barriers to participation, such as

XSL•FO
RenderX

incentives directly tied to transportation difficulties or food insecurity.

Furthermore, this study suggests that an EHR-based recruitment approach may be unsuccessful for older populations. Our findings demonstrated an impact of age on patient response even among patients with active user accounts. The mean age of the included participants was 65.7 (SD 12.3) years, and our results may have been skewed by this older patient base. It is possible that patient portal recruitment could have been more successful if conducted in a younger and likely more technologically adept population. This is in line with prior literature that has indicated that technological comfort and access decrease with age [27]. Recruitment for disease conditions that predominantly exist in older populations will likely need to use alternate recruitment strategies given this reduced comfort with technology.

Overall, our study demonstrates that asynchronous portal recruitment does not improve representation in trial engagement. Lack of diversity in clinical research limits the generalizability of trial results [14], and unfortunately, representation in research has only worsened over time in both industry-funded and academic center–sponsored trials [28]. The results observed in this trial may be due to the study not targeting specific barriers to engagement among patients from historically underrepresented racial or ethnic groups. Examples of such previously identified barriers include mistrust of health care institutions, structural factors that influence health literacy (eg, socioeconomic status and educational attainment), and logistical challenges to enrollment (eg, transportation) [29,30]. Other mechanisms of recruitment that specifically target these issues are likely needed to enhance overall representativeness of trial participants. For example, patient navigation programs targeting underresourced populations have been demonstrated to produce low trial refusal rates [31]. Similarly, in a survey of national clinical research sites, the practices found to be most impactful on diverse enrollment were travel reimbursement and services, translated documents and translator services, and adequate participant compensation [32]. While tailored recruitment strategies specific to each barrier may require more investment than a broader solution (such as EHR-based message recruiting), such investment may be necessary to truly achieve more equity among trial participants. This study suggests that racial disparities in trial enrollment may continue to worsen if EHR recruitment strategies are expanded [27].

This study has limitations. The single-center design may limit the generalizability of our findings when extrapolated to other populations as the demographic of patients in our catchment area is not necessarily representative of all US hospitals. Additionally, we specifically selected patients with active portal accounts. We have previously reported that patients who identify as being part of a historically underrepresented racial or ethnic group are less likely to have a patient portal account [18]. Furthermore, there was a racial and ethnic disparity observed in the initial eligible population (4399/5302, 83% non-White Hispanic patients vs 903/5302, 17% patients of other racial or ethnic categories) with active EHR portal accounts. There may be structural limitations or barriers (such as socioeconomic status or differential access to the internet) underlying this discrepancy that this study did not account for, which may have affected the engagement results observed in this study as well as limiting the potential generalizability of our results. It is possible that this study underestimated disparities in EHR-based recruitment.

While this study specifically focused on patients with a history of cancer, these findings may be generalizable beyond this patient population. Prior research has demonstrated that, while patients with cancer are as likely to enroll in clinical trials as the average American adult [33-35], they are less likely to have active patient portal accounts [36,37]. However, as our study only included patients with active portal accounts, the findings related to disparities in response are likely generalizable to other groups. This study contributes important evidence that EHR-based recruitment with or without modest financial incentives does not meaningfully reduce disparities in study enrollment, although our study suggests that the EHR could facilitate targeted recruitment of specific populations, including underrepresented communities.

Although our study focused on using patient portals for trial recruitment, these results may also suggest similar limitations when adapting patient portals for other use cases, such as behavioral interventions [38], telerehabilitation [39], or screening alerts [9]. Ultimately, we suspect that similar dynamics in message openings and patient response will persist in other clinically adjacent contexts. We recommend that future work investigate larger or targeted incentives, personalized outreach, and community-engaged recruitment and outreach approaches to enhance diversity and inclusion in both clinical research and therapy.

Conclusions

In this randomized study, we found that patients with a history of cancer who identified as belonging to racial or ethnic minority groups were significantly less likely to view or respond to EHR portal–based research invitations. These findings reinforce long-standing concerns that digital tools alone, particularly those reliant on patient portals, may perpetuate or fail to mitigate inequities in clinical trial participation. More robust, culturally sensitive, and community-engaged approaches are needed to foster trust and facilitate engagement and participation among these populations. A small financial incentive only slightly increased overall response rates and did not meaningfully improve equity in research engagement in a survey-based study. Future research should focus on identifying and testing scalable, equity-focused recruitment models that balance efficiency with inclusivity, ensuring that all patients have a fair opportunity to contribute to and benefit from scientific progress.

XSL•FO
RenderX

Funding

This study was supported by a research grant from the Hope Foundation for Cancer Research, the public charity of the SWOG Cancer Research Network. The work presented herein is that of the authors and not of the Hope Foundation for Cancer Research or the SWOG Cancer Research Network.

Authors' Contributions

Conceptualization: MF (lead), AAD (supporting), FA (supporting)
Data curation: MF (lead), WG (supporting)
Formal analysis: MF (lead), WG (supporting)
Funding acquisition: MF
Investigation: MF (lead), WG (supporting)
Methodology: MF
Project administration: MF (lead), AAD (supporting), FA (supporting)
Supervision: MF (lead), AAD (equal)
Visualization: MF (lead), CP (supporting)
Writing—original draft: CP (lead)
Writing—review and editing: CP (lead), MF (supporting), AAD (supporting)

Conflicts of Interest

AAD is on the scientific advisory boards of Biotheranostics, NeoGenomics Laboratories, and Pfizer; receives grant support from the Breast Cancer Alliance; and has received travel support from DAVA Oncology. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Screenshot representative of the messages patients received inviting them to participate in the sham study. (A) Email notification. (B) Notification on MyChart homepage. (C) Research invitation message.
[DOCX File, 343 KB - jmir_v28i1e86110_app1.docx]

Checklist 1
CONSORT checklist.
[PDF File, 1124 KB - jmir_v28i1e86110_app2.pdf]

References

1. Tomiwa T, Wong E, Miller HN, et al. Leveraging digital tools to enhance diversity and inclusion in clinical trial recruitment. Front Public Health 2024;12:1483367. [doi: 10.3389/fpubh.2024.1483367] [Medline: 39529717]
2. Brøgger-Mikkelsen M, Ali Z, Zibert JR, Andersen AD, Thomsen SF. Online patient recruitment in clinical trials: systematic review and meta-analysis. J Med Internet Res 2020 Nov 4;22(11):e22179. [doi: 10.2196/22179] [Medline: 33146627]
3. Treweek S, Pitkethly M, Cook J, et al. Strategies to improve recruitment to randomised trials. Cochrane Database Syst Rev 2018 Feb 22;2(2):MR000013. [doi: 10.1002/14651858.MR000013.pub6] [Medline: 29468635]
4. Huynh L, Johns B, Liu SH, Vedula SS, Li T, Puhan MA. Cost-effectiveness of health research study participant recruitment strategies: a systematic review. Clin Trials 2014 Oct;11(5):576-583. [doi: 10.1177/1740774514540371] [Medline: 24980278]
5. Lai YS, Afseth JD. A review of the impact of utilising electronic medical records for clinical research recruitment. Clin Trials 2019 Apr;16(2):194-203. [doi: 10.1177/1740774519829709] [Medline: 30764659]
6. Frandsen M, Walters J, Ferguson SG. Exploring the viability of using online social media advertising as a recruitment method for smoking cessation clinical trials. Nicotine Tob Res 2014 Feb;16(2):247-251. [doi: 10.1093/ntr/ntt157]
7. Goshgarian G, Sorourdi C, May FP, et al. Effect of patient portal messaging before mailing fecal immunochemical test kit on colorectal cancer screening rates: a randomized clinical trial. JAMA Netw Open 2022 Feb 1;5(2):e2146863. [doi: 10.1001/jamanetworkopen.2021.46863] [Medline: 35119462]
8. Stockwell MS, Kharbanda EO, Martinez RA, et al. Text4Health: impact of text message reminder-recalls for pediatric and adolescent immunizations. Am J Public Health 2012 Feb;102(2):e15-e21. [doi: 10.2105/AJPH.2011.300331] [Medline: 22390457]
9. Szilagyi PG, Albertin C, Casillas A, et al. Effect of patient portal reminders sent by a health care system on influenza vaccination rates: a randomized clinical trial. JAMA Intern Med 2020 Jul 1;180(7):962-970. [doi: 10.1001/jamainternmed.2020.1602] [Medline: 32421168]
10. Lerner C, Albertin C, Casillas A, et al. Patient portal reminders for pediatric influenza vaccinations: a randomized clinical trial. Pediatrics 2021 Aug;148(2):e2020048413. [doi: 10.1542/peds.2020-048413] [Medline: 34321338]

XSL•FO
RenderX

11. Samuels MH, Schuff R, Beninato P, et al. Effectiveness and cost of recruiting healthy volunteers for clinical research studies using an electronic patient portal: a randomized study. J Clin Transl Sci 2017 Dec;1(6):366-372. [doi: 10.1017/cts.2018.5] [Medline: 29707259]
12. Miller HN, Wong E, Byiringiro S, et al. CONNECT platform to increase diverse engagement and enrollment in cardiovascular clinical research: rationale, design, and lessons learned in early implementation. J Am Heart Assoc 2025 Aug 19;14(16):e038351. [doi: 10.1161/JAHA.124.038351] [Medline: 40792590]
13. Kannan V, Wilkinson KE, Varghese M, et al. Count me in: using a patient portal to minimize implicit bias in clinical research recruitment. J Am Med Inform Assoc 2019 Aug 1;26(8-9):703-713. [doi: 10.1093/jamia/ocz038] [Medline: 31081898]
14. Unger JM, Vaidya R, Hershman DL, Minasian LM, Fleury ME. Systematic review and meta-analysis of the magnitude of structural, clinical, and physician and patient barriers to cancer clinical trial participation. J Natl Cancer Inst 2019 Mar 1;111(3):245-255. [doi: 10.1093/jnci/djy221] [Medline: 30856272]
15. Brown EHJ, Peiper NC, Furmanek S, McCants KC. Integrating clinical trial data with electronic health records to improve diversity in research: findings from an urban hospital system. Trials 2026 Jan 26;27(1):148. [doi: 10.1186/s13063-026-09492-y] [Medline: 41582222]
16. Graetz I, Gordon N, Fung V, Hamity C, Reed ME. The digital divide and patient portals: internet access explained differences in patient portal use for secure messaging by age, race, and income. Med Care 2016;54(8):772-779. [doi: 10.1097/MLR.0000000000000560]
17. Shah A, Fiala MA. Disparities in access and utilization of electronic health record patient portals. J Gen Intern Med 2025 Jul;40(10):2233-2239. [doi: 10.1007/s11606-025-09359-z] [Medline: 39843664]
18. Trammel AM, Sisk BA, Fiala MA. Disparities in electronic health record patient portal activation and use among people with hematological malignancies. J Geriatr Oncol 2025 Mar;16(2):102135. [doi: 10.1016/j.jgo.2024.102135] [Medline: 39443182]
19. Guetterman TC, Koptyra E, Ritchie O, et al. Equity in virtual care: a mixed methods study of perspectives from physicians. J Telemed Telecare 2025 Apr;31(3):408-416. [doi: 10.1177/1357633X231194382] [Medline: 37641207]
20. Unger JM, Hershman DL, Till C, et al. “When offered to participate”: a systematic review and meta-analysis of patient agreement to participate in cancer clinical trials. J Natl Cancer Inst 2021 Mar 1;113(3):244-257. [doi: 10.1093/jnci/djaa155] [Medline: 33022716]
21. Walker DM, Swoboda CM, Shiu-Yee K, Tarver WL, Nolan TS, Joseph JJ. Diversity of participation in clinical trials and influencing factors: findings from the Health Information National Trends Survey 2020. J Gen Intern Med 2023 Mar;38(4):961-969. [doi: 10.1007/s11606-022-07780-2] [Medline: 36138277]
22. Navar AM, Esselink K, Lingvay I, et al. Response to electronic health record patient portal-based clinical study invitations. JAMA Netw Open 2025 Sep 2;8(9):e2531624. [doi: 10.1001/jamanetworkopen.2025.31624] [Medline: 40952746]
23. Singer E, Ye C. The use and effects of incentives in surveys. Ann Am Acad Political Soc Sci 2013 Jan;645(1):112-141. [doi: 10.1177/0002716212458082]
24. Anhang Price R, Quigley DD, Hargraves JL, et al. A systematic review of strategies to enhance response rates and representativeness of patient experience surveys. Med Care 2022 Dec 1;60(12):910-918. [doi: 10.1097/MLR.0000000000001784] [Medline: 36260705]
25. Abdelazeem B, Hamdallah A, Rizk MA, et al. Does usage of monetary incentive impact the involvement in surveys? A systematic review and meta-analysis of 46 randomized controlled trials. PLoS One 2023;18(1):e0279128. [doi: 10.1371/journal.pone.0279128] [Medline: 36649255]
26. PECaD dashboard: demographics. Siteman Cancer Center, Washington University School of Medicine. URL: https://tableau.wustl.edu/t/WU_Public_Site/views/P50DashboardDraft2023/DemographicsMain?%3Aembed=y&%3AisGuestRedirectFromVizportal=y [accessed 2025-08-11]
27. Gordon NP, Hornbrook MC. Differences in access to and preferences for using patient portals and other eHealth technologies based on race, ethnicity, and age: a database and survey study of seniors in a large health plan. J Med Internet Res 2016 Mar 4;18(3):e50. [doi: 10.2196/jmir.5105] [Medline: 26944212]
28. Dunlop H, Fitzpatrick E, Kurti K, et al. Participation of patients from racial and ethnic minority groups in phase 1 early cancer drug development trials in the US, 2000-2018. JAMA Netw Open 2022 Nov 1;5(11):e2239884. [doi: 10.1001/jamanetworkopen.2022.39884] [Medline: 36326764]
29. Dawson S, Ostrovska B, Treweek S, et al. Sometimes different, often the same: guidance on how recruitment and retention in trials need to be tailored to enhance participation of ethnic minority groups: a qualitative interview study in England. J Clin Epidemiol 2025 Dec;188:112005. [doi: 10.1016/j.jclinepi.2025.112005] [Medline: 41038543]
30. Thompson HS, Manning M, Mitchell J, et al. Factors associated with racial/ethnic group-based medical mistrust and perspectives on COVID-19 vaccine trial participation and vaccine uptake in the US. JAMA Netw Open 2021 May 3;4(5):e2111629. [doi: 10.1001/jamanetworkopen.2021.11629] [Medline: 34042990]
31. Ghebre RG, Jones LA, Wenzel JA, Martin MY, Durant RW, Ford JG. State-of-the-science of patient navigation as a strategy for enhancing minority clinical trial accrual. Cancer 2014 Apr 1;120 Suppl 7(7):1122-1130. [doi: 10.1002/cncr.28570] [Medline: 24643650]

XSL•FO
RenderX

32. MacLennan DL, Plahovinsak JL, MacLennan RJ, Jones CT. Clinical trial site perspectives and practices on study participant diversity and inclusion. Clin Pharmacol Ther 2023 Mar;113(3):670-679. [doi: 10.1002/cpt.2817] [Medline: 36495475]
33. Unger JM, Shulman LN, Facktor MA, Nelson H, Fleury ME. National estimates of the participation of patients with cancer in clinical research studies based on commission on cancer accreditation data. J Clin Oncol 2024 Jun 20;42(18):2139-2148. [doi: 10.1200/JCO.23.01030] [Medline: 38564681]
34. Green AK, Tabatabai SM, Aghajanian C, et al. Clinical trial participation among older adult Medicare fee-for-service beneficiaries with cancer. JAMA Oncol 2022 Dec 1;8(12):1786-1792. [doi: 10.1001/jamaoncol.2022.5020] [Medline: 36301585]
35. Williams CP, Senft Everson N, Shelburne N, Norton WE. Demographic and health behavior factors associated with clinical trial invitation and participation in the United States. JAMA Netw Open 2021 Sep 1;4(9):e2127792. [doi: 10.1001/jamanetworkopen.2021.27792] [Medline: 34586365]
36. Griffin JM, Kroner BL, Wong SL, et al. Disparities in electronic health record portal access and use among patients with cancer. J Natl Cancer Inst 2024 Mar 7;116(3):476-484. [doi: 10.1093/jnci/djad225] [Medline: 37930884]
37. Maini I, Gilotra K, Sadigh G. Patient factors associated with the use of online portal health information in the postpandemic era: cross-sectional analysis of a national survey. J Med Internet Res 2025 Apr 3;27:e60472. [doi: 10.2196/60472] [Medline: 40179386]
38. Conroy SS, Harcum S, Keldsen L, Bever CTJ. Novel use of existing technology: a preliminary study of patient portal use for telerehabilitation. J Telemed Telecare 2022 Jun;28(5):380-388. [doi: 10.1177/1357633X20950172] [Medline: 32869689]
39. Steinhilber KM, Chabria RS, Clara A, et al. Shifting behavioral intervention research to virtual methods: challenges and solutions in practice, during and after the COVID-19 pandemic. J Telemed Telecare 2025 Jan;31(1):134-139. [doi: 10.1177/1357633X231167899] [Medline: 37125429]

Abbreviations

aOR: adjusted odds ratio
EHR: electronic health record
SCC: Siteman Cancer Center

Edited by M Balcarras; submitted 18.Oct.2025; peer-reviewed by T Coffin, W Shiyab; revised version received 27.May.2026; accepted 29.May.2026; published 22.Jun.2026.

Please cite as:
Ping C, Davis AA, Gema W, Ademuyiwa F, Fiala M
Impact of Financial Incentives on Electronic Health Record–Driven Recruitment of Underrepresented Communities in Research: Randomized Controlled Trial
J Med Internet Res 2026;28:e86110
URL: https://www.jmir.org/2026/1/e86110
doi:10.2196/86110

© Christina Ping, Andrew Davis, Waishun Gema, Foluso Ademuyiwa, Mark Fiala. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Understanding Patient-Reported Offenses in Electronic Health Records: Cross-Sectional Mixed Methods Survey

Saija Simola[1], MSc; Sari Kujala[1], PhD; Åsa Cajander[2], PhD; Anna Kharko[3,4], PhD; Bridget Kane[3,5], PhD; Bo Wang[6], MSc; Rose-Mharie Åhlfeldt[7], PhD; Maria Hägglund[3,8], PhD

[1]Department of Computer Science, Aalto University, Konemiehentie 2, Espoo, Finland
[2]Department of Information Technology, Uppsala University, Uppsala, Sweden
[3]Department of Women's and Children's Health, Participatory eHealth and Health Data Research Group, Uppsala University, Uppsala, Sweden
[4]Centre for Primary Care and Health Services Research, University of Manchester, Manchester, United Kingdom
[5]Business School, Karlstad University, Karlstad, Sweden
[6]Norwegian Centre for E-health Research, University Hospital of North Norway, Tromsø, Norway
[7]School of Informatics, University of Skövde, Skövde, Sweden
[8]Uppsala University Hospital, Uppsala, Sweden

Corresponding Author:
Saija Simola, MSc
Department of Computer Science, Aalto University, Konemiehentie 2, Espoo, Finland

Abstract

Background: Patients' access to their electronic health record (EHR) supports their participation and satisfaction with care. Despite the benefits, some patients have been upset after reading their EHR. Additionally, health care professionals are concerned that patients, particularly those with mental health conditions, may be offended, and they have expressed a need for further guidelines on how to write EHRs. Experiences among various patient groups are essential to support the relationship between patients and professionals. However, prior studies have often focused on single patient groups or specific clinical contexts, leaving a limited understanding of differences across multiple patient groups.

Objective: This study aimed to determine whether certain patient groups are more likely to feel offended while reading their EHRs and which information is perceived as offensive and to provide a comparison across multiple patient groups using a mixed methods approach.

Methods: A cross-sectional survey was conducted via the Finnish national patient portal using a web-based patient survey, adopting a mixed methods approach. The survey included multiple-choice and open-ended questions. The total sample comprised 4681 respondents. The survey respondents were placed into 4 patient groups: those who had received care for mental health, cancer, or other conditions and those who had received no care. Associations between the type of care and patients who felt offended were estimated using multivariate binary logistic regression. Inductive content analysis (n=502) was conducted to identify information perceived as offensive in the EHR.

Results: The patients who had received mental health care (166/654, 25.4%) or cancer and mental health care (9/39, 23.1%) were more likely to be offended by information in their EHR compared to the other groups (cancer care: 37/375, 9.9%; other conditions care: 383/3316, 11.6%; no care: 22/206, 10.7%; other conditions care: odds ratio 0.37, 95% CI 0.29 - 0.46; $P<.001$; model A). Additionally, female patients, those with bad or very bad health conditions, and patients with bachelor's or master's degrees were significantly more likely to feel offended. Errors, the health care professionals' disrespectful language, and perceived unnecessary information were the most frequently mentioned reasons for being offended. Patients with mental health care reported more often that unnecessary information and professionals' opinions and word choices were experienced as offensive compared to other patients.

Conclusions: This study contributes new knowledge by identifying differences across multiple patient groups. Although a minority of patients felt offended by their EHR, health care professionals should consider that some patients, particularly those who have received mental health care or cancer and mental health care, may be offended by specific information or word choices in their EHRs. To address this, health care professionals should receive education on how to write their notes in a neutral tone and avoid potentially offensive topics. Improving the quality of EHRs could strengthen the relationship between patients and professionals.

(J Med Internet Res 2026;28:e86178) doi:10.2196/86178

XSL•FO
RenderX

KEYWORDS

electronic health record; online record access; patient portal; user groups; offensive; national survey; open notes; patient experiences; patient-accessible electronic health records

Introduction

Background

Increasingly, countries are offering patients access to their electronic health records (EHRs) because it benefits both patients and the health care system [1]. The Northern European countries, including Finland, have a long history of planning national EHRs to support patients' autonomy and participation in alliance with health care professionals' access to the records [2]. Available through patient portals, EHRs include medical information, such as appointment notes, diagnoses, and laboratory test results [3,4]. Patients' access to their EHRs can improve communication between patients and health care professionals [5,6], as well as patients' participation [7,8], satisfaction [9], trust [10], and engagement with care [11,12]. Additionally, patients' access to their EHRs can encourage them to take their medicines as prescribed [13], support their recall [7,9,11,14], and enhance their understanding of care [8,12].

While patients' access to their EHR appears to be beneficial to patients in general, differences between patient groups have been identified. Some patients with mental health conditions have reported experiencing these benefits more frequently than those without mental health issues [11,15-17]. Health care professionals have found that using EHRs might help patients better understand their mental health conditions, including their diagnoses, care plans, medications, and treatments [15,18]. Patients' access to their EHR is even more likely to encourage them with serious mental health conditions to take their medications as prescribed compared to patients without mental health diagnoses [18]. In addition, patients with cancer have reported experiencing less uncertainty or anxiety [19], and patients with complex chronic conditions have shown a decreased likelihood of visiting the emergency room and having hospital stays if they have access to the EHR [20].

Access to the EHR can also cause negative emotions in patients, including those with mental health conditions, and EHR content that upsets, offends, or judges them can deter them from reading their EHR [16,21,22]. Increased anxiety has been reported among some oncologic patients after reading their EHR [19,23]. Around 11% to 10% of patients, including oncologic patients, who read their EHR feel offended [13,22-24]. However, patients who had received mental health care have been more often offended than other patient groups [25,26]. Feeling offended by the EHR content is also associated with poorer reported perceived usability of the patient portal [24]. On the other hand, patients have indicated a willingness to access their EHRs despite potential challenges arising from their records, such as feeling offended by the content [12].

Such potential negative consequences are commonly cited as reasons why some health care professionals resist patients' access to their EHRs. Some professionals expect patients to worry more [27] or potentially misunderstand the EHR information [28]. Health care professionals are also concerned about the potential worsening of their relationship with their patients [28-32], particularly if professionals, for example, raise questions related to sensitive topics such as gender identity [33]. This introduces a dilemma because health care professionals have a medicolegal obligation to include certain information in the EHR, even if it may worry or upset the patient [10,31,34].

Due to the potential for offense, some health care professionals may not be willing to share sensitive topics or certain parts of the EHR, or they may even choose not to share records with patients at all [34-36]. This is particularly true for mental health records [9,37]. In addition, professionals have described negative consequences after patients read about their mental health, such as decreased engagement in their care, worsened relationships [9] and trust [9,17,38], and disagreements about the EHR content [9,39-41]. Moreover, psychiatrists are more concerned than nonpsychiatrists about the possibility of patients feeling offended [42]. Health care professionals also worry about patients with mental health issues facing information in their EHR that could cause negative emotional responses without the support of professionals [38].

On the other hand, there is less previous research relating to some medical specialties, such as oncology [43]. This survey study focuses on patients accessing EHR via the national patient portal. The Finnish national patient portal, My Kanta, was launched in May 2010 [44]. In addition to a patient-accessible EHR, all Finnish citizens can use various functionalities in My Kanta, such as renewing their electronic prescriptions, acting on behalf of their children, or recording their organ donation testaments and living wills [45]. Nowadays, in a cohort study, 82% of all Finnish adults have signed in to My Kanta at least once [45].

Aims and Research Questions

To minimize and prevent detrimental impacts from patient access to their EHR, it is important to explore the kinds of information that patients experience as offensive. Moreover, health care professionals have expressed a need for further guidance on writing EHR notes and on content that may elicit negative emotions in patients [46]. Liu [47] has stated that there is no adequate philosophical theory of offense. However, based on the thematic analysis of patients' experiences collected through a survey, Fernández et al [22] suggest that feeling offended includes experiences where EHR contained errors, surprises, labeling, and disrespectful language.

This study allowed patients to define "offense" themselves rather than selecting from predefined options. This approach avoids leading respondents and captures their own perspectives. Thus, this study aims to examine whether certain patient groups are more likely to perceive information in their EHRs as offensive. To complement previous studies, this study focuses on multiple patient groups based on the type of care they receive. Additionally, this study aims to identify which types of information patients experienced as offensive.

XSL•FO
RenderX

Our research questions were as follows: (1) Which user groups are more likely to perceive information in their EHR as offensive? (2) Which information do users experience as offensive in their EHR?

Methods

A web-based survey was created to explore patients' experiences with patient-accessible EHRs via a national patient portal, My Kanta, following the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) [48] (Checklist 1).

Research Design Overview

In this study, a mixed methods approach with a convergent design [49] was used, including the survey. The mixed methods approach enabled us to obtain a large sample size and to examine the differences between patient groups using the quantitative method. The qualitative responses provided a deeper understanding of the offenses experienced in the EHR. The qualitative data collection avoided the use of predefined options to avoid influencing respondents when exploring whether they had felt offended by something they had read.

This study utilized an anonymous questionnaire combining multiple-choice quantitative and open-ended qualitative questions (Multimedia Appendix 1) to collect parallel data and adopt a connecting approach [49]. The analysis was performed after all data had been collected. Because mental health and cancer are common and increasingly prevalent conditions [50-53], and patients with these conditions are also seen as vulnerable [37,54-56], we focused on patients who had received care for mental health conditions or cancer in our analysis, distinguishing them from patients who had received care for any other condition or no care in the last 24 months.

The quantitative questions included demographic information (age, gender, self-rated overall health, education level, employment, and whether they had health care professional education), experience with health care (having received care in the past 24 months for mental health, cancer, other conditions, or not having received care), and experience of feeling offended when reading their EHR. The respondents were asked to provide additional details about perceived offensive information in the EHR through the open-ended question, "Have you ever felt offended by something you read? If yes, please explain" (Multimedia Appendix 1).

Data Collection and Sample Size

The web-based survey administered in Finland was part of the NORDeHEALTH 2022 Patient Survey [57]. The responses were collected from patients logged in to the national patient portal between January 24, 2022, and February 14, 2022, with the option of choosing between the official languages of Finland: Finnish and Swedish. In total, 4719 responses were collected in The NORDeHEALTH 2022. Patient Survey design and data collection have been described in more detail by Hägglund et al [57].

Inclusion and Exclusion

The respondents were able to select multiple groups representing the types of care they had received over the past 24 months. Thus, the patients were categorized into groups based on whether they had received mental health care, cancer care, or both. The remaining respondents were categorized into the groups "Care for other conditions" or "No care" (Table 1). Respondents who had selected both mental health care and cancer care were grouped into a separate category to avoid overlap and enable deeper analysis (Figure 1). Respondents who reported conflicting responses about receiving some care and none (15/4719, 0.3%) were excluded from the analysis. The patients who did not respond to these questions (23/4719, 0.5%) were also excluded from the analysis, resulting in a final sample of N=4681 (Figure 1).

Figure 1. Exclusion and inclusion criteria for patient groups.

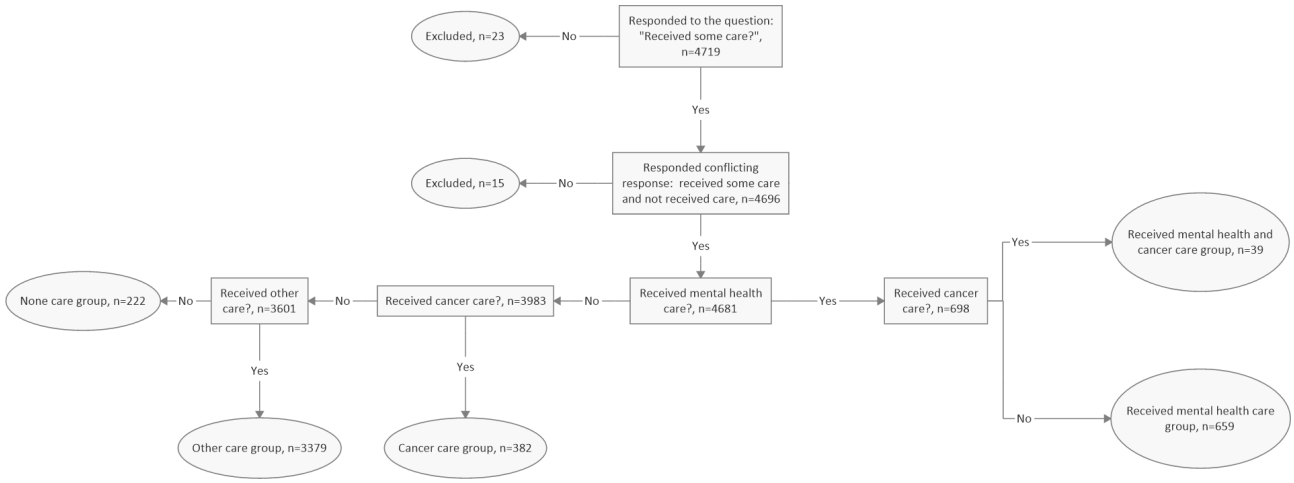

XSL•FO
RenderX

Table . Background information categorized by a patient group.

Background information	Mental health care (n=659), n (%)	Cancer care (n=382), n (%)	Mental health and cancer care (n=39), n (%)	Other conditions care (n=3379), n (%)	No care (n=222), n (%)	Total (N=4681), n (%)
Health condition						
Very good	11 (1.7)	5 (1.3)	2 (5.1)	133 (3.9)	34 (15.3)	185 (4)
Good	145 (22)	105 (27.5)	10 (25.6)	1200 (35.5)	103 (46.4)	1563 (33.4)
Fair	308 (46.7)	195 (51)	13 (33.3)	1637 (48.4)	63 (28.4)	2216 (47.3)
Bad	135 (20.5)	54 (14.1)	12 (30.8)	322 (9.5)	14 (6.3)	537 (11.5)
Very bad	38 (5.8)	16 (4.2)	1 (2.6)	52 (1.5)	5 (2.3)	112 (2.4)
I don't know/I don't want to answer or no response	22 (3.3)	7 (1.8)	1 (2.6)	35 (1)	3 (1.4)	68 (1.5)
Age (y)						
15 - 34	141 (21.4)	5 (1.3)	2 (5.1)	140 (4.1)	17 (7.7)	305 (6.5)
35 - 54	256 (38.8)	28 (7.3)	7 (17.9)	609 (18)	47(21.2)	947 (20.2)
55 - 74	229 (34.7)	263 (68.8)	29 (74.4)	2088 (61.8)	131 (59)	2740 (58.5)
75 or more	25 (3.8)	85 (22.3)	1 (2.6)	527 (15.6)	27 (12.2)	665 (14.2)
No response	8 (1.2)	1 (0.3)	0 (0)	15 (0.4)	0 (0)	24 (0.5)
Gender						
Female	525 (79.7)	243 (63.6)	30 (76.9)	2462 (72.9)	138 (62.2)	3398 (72.6)
Male	114 (17.3)	135 (35.3)	7 (17.9)	881 (26.1)	77 (34.7)	1214 (25.9)
Other/I don't want to answer or no response	20 (3)	4 (1)	2 (5.1)	36 (1.1)	7 (3.2)	69 (1.5)
Education						
Elementary school or less educated	59 (9)	48 (12.6)	3 (7.7)	354 (10.5)	26 (11.7)	490 (10.5)
12 years school—upper secondary education	213 (32.3)	76 (19.9)	11 (28.2)	854 (25.3)	55 (24.8)	1209 (25.8)
Higher vocational education	103 (15.6)	94 (24.6)	8 (20.5)	783 (23.2)	42 (18.9)	1030 (22)
Higher education ≤3 years	144 (21.9)	73 (19.1)	5 (12.8)	568 (16.8)	29 (13.1)	819 (17.5)
Higher education >3 years	109 (16.5)	69 (18.1)	11 (28.2)	639 (18.9)	47 (21.2)	875 (18.7)
Doctoral education	7 (1.1)	7 (1.8)	0 (0)	62 (1.8)	6 (2.7)	82 (1.8)
Something else or no response	24 (3.6)	15 (3.9)	1 (2.6)	119 (3.5)	17 (7.7)	176 (3.8)
Health care professional education						
Yes	159 (24.1)	73 (19.1)	9 (23.1)	740 (21.9)	34 (15.3)	1015 (21.7)
No	484 (73.4)	296 (77.5)	29 (74.4)	2570 (76.1)	183 (82.4)	3562 (76.1)
No response	16 (2.4)	13 (3.4)	1 (2.6)	69 (2)	5 (2.3)	104 (2.2)
Employment						
Full time	158 (24)	50 (13.1)	7 (17.9)	847 (25.1)	66 (29.7)	1128 (24.1)
Part time	72 (10.9)	18 (4.7)	3 (7.7)	167 (4.9)	18 (8.1)	278 (5.9)

XSL•FO
RenderX

Background information	Mental health care (n=659), n (%)	Cancer care (n=382), n (%)	Mental health and cancer care (n=39), n (%)	Other conditions care (n=3379), n (%)	No care (n=222), n (%)	Total (N=4681), n (%)
Student	63 (9.6)	2 (0.5)	1 (2.6)	66 (2)	4 (1.8)	136 (2.9)
Retired	181 (27.5)	282 (73.8)	20 (51.3)	2003 (59.3)	115 (51.8)	2601 (55.6)
Not working	148 (22.5)	17 (4.5)	6 (15.4)	189 (5.6)	11 (5)	371 (7.9)
None of the above or no response	37 (5.6)	13 (3.4)	2 (5.1)	107 (3.2)	8 (3.6)	167 (3.6)

Participant Characteristics

The largest patient group included those who received care for other conditions, representing 72.2% (3379/4681) of the respondents (Table 1). The next largest groups were patients receiving mental health care (659/4681, 14.1%), patients receiving cancer care (382/4681, 8.2%), those who did not receive any care (222/4681, 4.7%), and those receiving care for both mental health and cancer (39/4681, 0.8%).

Overall, the patients tended to be older, with more than half aged 55 years or older. The patients who received mental health care were typically younger, aged 15 to 34 years (141/659, 21.4%), while those who received care for cancer were more likely to be older, aged 75 years or older (85/382, 22.3%). All groups had more female respondents, ranging from 62.2% (138/222) to 79.7% (525/659). All education levels were represented, and more than one-fifth of the patients had health care professional education. Since more than half of the patients were older adults, 55.6% (2601/4681) were retired.

Statistics and Data Analysis

The statistical data were analyzed using Excel (version 2308; Microsoft Corp) and SPSS (version 29.0.2.0; Lumivero). The statistical analysis was performed using the multivariate binary logistic regression analysis, as the dependent variable was binary and the independent background variables were categorical. The binary logistic regression analysis was conducted in 2 steps. First, model A included variable types of received care based on the hypothesis. In addition, model B included the background variables, such as health condition, age, gender, education, health care professional education, and employment. The employment variable was transformed into 3 categories: combining full time and part time as the working category, student, and combining retired and not working as the outside working life category. Each independent variable used the first category as the reference category. The significance level was set to $P<.05$. The multicollinearity was tested using the Pearson χ^2 test. Each independent variable was tested in pairs with all the others.

Based on the inclusion and exclusion criteria for patient groups, the final sample was 4681, representing 99.2% (4681/4719) of the respondents. In total, 38 (0.8%) responses included conflicting information about the patient group or were missing due to insufficient information about the patient group. In addition, 91 (1.9%; Table 2) respondents did not indicate whether they had felt offended by anything they read in the EHR. Those responses with missing data were excluded, and the final sample size in the quantitative analysis was n=4590, representing 97.3% (4590/4719) of the respondents.

Table . Frequency of feeling offended among patient groups.

Patient group	Yes, n (%)	No, n (%)	No response, n (%)[a]
Mental health care (n=654)[b]	166 (25.4)	488 (74.6)	5 (0.8)
Cancer care (n=375)[b]	37 (9.9)	338 (90.1)	7 (1.8)
Mental health and cancer care (n=39)[b]	9 (23.1)	30 (76.9)	0 (0)
Other conditions care (n=3316)[b]	383 (11.6)	2933 (88.4)	63 (1.9)
No care (n=206)[b]	22 (10.7)	184 (89.3)	16 (7.2)
Total (n=4590)[b]	617 (13.4)	3973 (86.6)	91 (1.9)

[a]Percentages have been calculated using the total sample (n=4681), including “No responses.”

[b]This sample does not include “No responses.”

Since all variables were categorical, a χ^2 test was performed to examine associations between the missingness indicators for felt offended and patient group variables (coded 1=missing value, 0=not missing value) and other variables [58]. For comparisons with the patient group, all variables were nonsignificant except health condition, suggesting the data may be missing completely at random. Those who did not want to share their health condition (2/60, 3.3%) did not respond to the patient group question, differing from other categories (0% - 0.5%) with the significant association (χ^2_5=11.54; P=.04), suggesting that missingness depends on health conditions and the data may be missing at random. However, the level of

XSL•FO
RenderX

missingness is below common thresholds of 5% to 20%, which is typically considered problematic [59].

All variables, except education and patient group, were not significantly associated with the felt offended variable. Among those with no formal education, 10.5% (2/19) did not respond to whether they felt offended, with differences across other categories, with low levels of no responses ranging from 0.9% to 3.0% (χ^2_7=15.51; *P*=.03). Respondents across all patient groups had similarly low levels of no responses ranging from 0% to 1.9%, except for those who had not received care in the past 24 months having 7.2% (16/222; χ^2_5=39.48; *P*<.001) of no responses. Patients who had not received care may have had less information in their EHR.

Qualitative Data-Analytic Strategies

The qualitative inductive content analysis was conducted using the ATLAS.ti software (version 24.1.1.30813, ATLAS.ti Scientific Software Development GmbH) and inspired by Harahap et al [60], following their process, including (1) familiarizing authors with data, (2) generating and discussing the codes in collaboration, (3) grouping codes into larger theme categories, (4) evaluating larger theme categories and defining the larger theme categories, and finally presenting larger theme categories and codes with quotes in the *Results* section. Each respondent who felt offended could provide 1 answer to the open-ended question explaining why they felt offended. All responses were analyzed to ensure that inductive thematic saturation was achieved. The same codes repeatedly emerged from the data [61], as agreed by the first and second authors. The responses varied in length from 1 to 131 words, with an average of 10 (SD 11) words.

The first author familiarized themselves with the open-ended responses. Second, they performed preliminary coding using in-vivo coding. In-vivo coding [62] focuses on users' perspectives rather than predefined options or researchers' interpretations of the codes. Respondents could provide more than 1 reason for feeling offended, and all reasons were classified under their respective codes. To increase the reliability of the analysis process, the second author discussed and reviewed the codes together with the first author. Third, the first author grouped the codes into larger theme categories. Fourth, the first and second authors collaboratively evaluated and named the larger theme categories. In addition, they created English translations for the codes and larger theme categories. All authors verified the meaning of the codes and larger theme categories.

If both authors were jointly unable to understand the meaning of a response, it was categorized as an "unclear response." They reviewed together all the unclear responses (39/502, 7.8%). In addition, a few responses (7/502, 1.4%) were classified as "empty response" when respondents explicitly explained that they would not provide additional information, with explanations such as "I don't want to tell."

Ethical Considerations

The study protocol was reviewed and approved by the Aalto University Research Ethics Committee (D/957/03.04/2020 Z) in accordance with the ALLEA European Code of Conduct for Research Integrity and the Finnish Code of Conduct for Research Integrity and Procedures for Handling Alleged Violations of Research Integrity in Finland 2023. The survey invitation included study information and a privacy notice. Responding to the survey was interpreted as indicating that participants had given their informed consent. The informed consent allowed the secondary analysis without additional consent. Participation in the survey was voluntary and anonymous, and no reimbursement was offered. To protect participants' identities, this study was anonymous and did not include identification variables. If responses included identification information provided by the respondent, it was removed from the data. The data were securely stored in the university's secure project folder, with restricted access. The identification of individual participants or users in any tables or quotes of the manuscript or supplementary material is not possible.

Results

Associations Between Feeling Offended and the Patient Group

The patient group was significantly associated with feeling offended by information. Patients receiving any mental health care (23.1% - 25.4%, Table 2) were more likely than other patients to feel offended by information in their EHRs (9.9% - 11.6%; other care: odds ratio 0.37, 95% CI 0.29 - 0.46, *P*<.001, model A; Tables 2 and 3). This statistically significant association persisted even after including additional background variables in model B (Table 3).

Slightly more than every tenth (617/4590, 13.4%) of all respondents reported having felt offended by information in their EHR (Table 2). This level was consistent among cancer care respondents (37/375, 9.9%), respondents with other conditions (383/3316, 11.6%), and those who did not receive care (22/206, 10.7%). Moreover, every fourth (166/654, 25.4%) patient who received mental health care and patients who received both mental health and cancer care (9/39, 23.1%) encountered information they experienced as offensive in their EHR. However, most people (3973/4590, 86.6%) who read their information in their EHR were not offended.

Female patients, students, patients with bad or very bad health condition, those outside working life, or those with a bachelor's degree or higher were significantly more likely to feel offended by information (Table 3, Multimedia Appendix 2). Older patients were significantly less likely to feel offended by information in their EHRs (Table 3). There was no significant difference whether the patient has a health care professional education themselves. The model's fit was evaluated using Cox and Snell R^2 (0.02 for model A; 0.05 for model B; Multimedia Appendix 3), indicating limited explained variation with a minor improvement in extended model B. The Hosmer–Lemeshow test (χ^2_1=0, *P*>.99 for model A; χ^2_8=8.67, *P*=.37 for model B) indicated acceptable calibration.

XSL•FO
RenderX

Table . Multivariate binary logistic regression results—associations between feeling offended, patient group, and background variables.

Model, variable name, and category	Coefficient (B) (SE)	*P* value	Odds ratio (95% CI)
Model A			
Received care (reference: mental health)			
Cancer care	−1.12 (0.21)	<.001	0.33 (0.22 - 0.49)
Other care	−1.00 (0.12)	<.001	0.37 (0.29 - 0.46)
None	−1.21 (0.28)	<.001	0.30 (0.17 - 0.52)
Mental health and cancer care	−0.20 (0.44)	.65	0.82 (0.35 - 1.93)
Constant	−1.11 (0.1)	<.001	0.33
Model B			
Received care (reference: mental health)			
Cancer care	−0.70 (0.22)	.002	0.50 (0.32 - 0.78)
Other care	−0.55 (0.13)	<.001	0.58 (0.45 - 0.75)
None	−0.69 (0.29)	.02	0.50 (0.28 - 0.89)
Mental health and cancer care	−0.03 (0.46)	.94	0.97 (0.4 - 2.36)
Gender (reference: female)			
Male	−0.35 (0.12)	.005	0.71 (0.56 - 0.9)
Age (y) (reference: 15 - 34 y old)			
35 - 54	0.07 (0.21)	.74	1.07 (0.71 - 1.62)
55 - 74	−0.53 (0.22)	.02	0.59 (0.38 - 0.91)
75 or older	−1.09 (0.28)	<.001	0.34 (0.19 - 0.58)
Health care professional (reference: yes)			
No	−0.01 (0.12)	.91	0.99 (0.78 - 1.25)
Education (reference: elementary school or no formal education)			
12 years school—upper secondary education	0.38 (0.2)	.06	1.47 (0.99 - 2.17)
Higher vocational education (vocational diploma)	0.4 (0.21)	.05	1.5 (1 - 2.25)
Higher education ≤3 years (first cycle—bachelor)	0.56 (0.21)	.008	1.75 (1.16 - 2.65)
Higher education >3 years (second cycle—master)	0.7 (0.21)	<.001	2.02 (1.35 - 3.04)
Research (third cycle) of higher education	0.84 (0.41)	.04	2.32 (1.05 - 5.15)
Health condition (reference: very good)			
Good	0.21 (0.31)	.50	1.23 (0.67 - 2.25)
Fair	0.46 (0.31)	.14	1.58 (0.87 - 2.88)
Bad	1.04 (0.32)	.001	2.82 (1.5 - 5.33)
Very bad	1.59 (0.38)	<.001	4.9 (2.35 - 10.24)
Employment (reference: working)			
Student	0.59 (0.27)	.03	1.8 (1.06 - 3.04)
Outside working life	0.46 (0.13)	<.001	1.58 (1.23 - 2.05)

XSL•FO
RenderX

Model, variable name, and category	Coefficient (B) (SE)	*P* value	Odds ratio (95% CI)
Constant	−2.25 (0.4)	<.001	0.1

Experiences of Offensive Information in the EHR

Finally, 81.4% (502/617) of respondents who had experience of offense shared more detailed explanation. The respondents were most frequently offended by errors and disrespectful language in their EHRs (Table 4). Patients described instances where their expressions were interpreted wrongly by professionals, or where professionals were not listening during appointments, which was the cause of the error in the records:

> *As a patient, what I have said is not written in the appointment's/hospital care's records. In addition, the doctor may have interpreted what I said in his own way and in a disrespectful tone.* [Respondent #3740, Mental health care]

Table . Inductive content analysis results—types of patient-reported offensive information in the EHR.

Offensive information in the EHR[a]	Frequency, n (%)		
	Total (n=502)	Any mental health care[b] (n=142)	No mental health care (n=360)
Type of information	197 (39.2)		
Errors	126 (25.1)	26 (18.3)	100 (27.8)
Unnecessary information	60 (12)	25 (17.6)	35 (9.7)
Outdated information	11 (2.2)	4 (2.8)	7 (1.9)
Omissions	8 (1.6)	3 (2.1)	5 (1.4)
Health care professional's expression in notes	189 (37.6)		
Professional has written disrespectfully	87 (17.3)	20 (14.1)	67 (18.6)
Professional's opinion/conclusion	42 (8.4)	19 (13.4)	23 (6.4)
Professional's description of the patient	40 (8)	10 (7)	30 (8.3)
Professional's word choice	35 (7)	13 (9.2)	22 (6.1)
Information relating to a specific theme	89 (17.7)		
Weight	39 (7.8)	12 (8.5)	27 (7.5)
Use of alcohol or drugs	25 (5)	8 (5.6)	17 (4.7)
Mental health	17 (3.4)	10 (7)	7 (1.9)
Gender	6 (1.2)	1 (0.7)	5 (1.4)
Evaluating the intelligence level	5 (1)	1 (0.7)	4 (1.1)
Communication between patient and professional	51 (10.2)		
Professional has written something that was not discussed in the appointment	19 (3.8)	5 (3.5)	14 (3.9)
Patient has not been heard or believed	19 (3.8)	5 (3.5)	14 (3.9)
Patient disagrees	15 (3)	4 (2.8)	11 (3.1)
Information management	17 (3.4)		
Information shared without patient's permission	12 (2.4)	3 (2.1)	9 (2.5)
Information was not corrected	5 (1)	0 (0)	5 (1.4)

[a]EHR: electronic health record.

[b]Includes individuals who have received both, mental health care and cancer care.

XSL•FO
RenderX

> *It offends a bit when you notice that the doctor hasn't listened properly and writes wrong things.* [Respondent #0300, Other conditions care]
>
> *The doctor's notes of what happened do not always correspond to my own experience. It is also offensive to skip things and select what ends up being recorded.* [Respondent #0160, No care]

Moreover, unnecessary information was a common source of offense. Patients described that certain aspects of their lives or personal characteristics were not relevant to their health care:

> *In my opinion, the doctor unnecessarily wrote in the notes that me and my spouse no longer have sex.* [Respondent #0141, Other conditions care]
>
> *How is the treatment relating to if I look cheerful, my former profession, my marital status, etc.* [Respondent #0201, Other conditions care]

Sometimes, unnecessary information related to certain offensive experience-based topics, for instance, gender:

> *As a trans man, I don't understand why ALL records use "woman woman woman" even if gender has nothing to do with it.* [Respondent #3762, Mental health care]

Additionally, outdated information was a reason for offense. In this example, the offense was related to smoking:

> *The doctor added smoking as a risk, I smoked 50 years ago.* [Respondent #0135, Other conditions care]

The choice of wording or a disrespectful tone in the notes by professionals was also a common source of offense:

> *A personal descriptive text written by a nurse when I cried because of cancer. "The patient cries until patient was blue in the face." Fortunately, the chief medical doctor decided that this was to be removed at my request.* [Respondent #4367, Cancer care]
>
> *The doctor wrote that I am a difficult patient with my difficult pain.* [Respondent #0143, Other conditions care]

Patients described that information relating to a specific theme, such as evaluating intelligence levels or information about weight, was offensive:

> *The overweight is mentioned first, and I'm not even fat.* [Respondent #0078, Other conditions care]

Sometimes, patients described that professionals omitted addressing certain topics, such as weight, during their appointments, even though these topics were documented in the notes:

> *The doctor couldn't directly call me overweight, but I found the text about my information offensive when read. I would have liked the doctor to mention it aloud as well, and not just timidly express it in the text.* [Respondent #0261, Other conditions care]

Some experiences of being offended were related to communication or the relationship between patients and professionals. Patients described how they have not been heard or believed. By the same token, patients reported disagreements with the professionals:

> *The nurse's diagnosis of my mental state. I disagree with that. I don't like having this kind of note in my records.* [Respondent #4364, Cancer care]

Concerns were expressed about professionals accessing or sharing the notes without patients' permission:

> *In my opinion, the physiotherapist had looked at information that was not necessary in terms of treatment.* [Respondent #0057, Other conditions care]

Since patients receiving mental health care appeared to experience offense at a significantly higher rate than other patients, we analyzed and reported their responses separately to deepen our understanding of the factors that could cause this difference.

The reasons for being offended partly differed between those who had received mental health care and those who had not. For the patients who had received mental health care, unnecessary information was related to occasions when mental health was described in the note, even though it was not considered relevant by the respondent when seeking care for somatic problems, as patients expressed here:

> *Why is a psychiatric diagnosis always recorded "even if I'm just asking for a plaster on my knee"? How does a psychiatric diagnosis affect wound care?* [Respondent #3721, Mental health care]
>
> *If I go to the doctor because of a foot problem, the first record I see is a note of recurrent depression.* [Respondent #3669, Mental health care]

The patients described repeated mentions of their mental health conditions as labeling, as one patient explained:

> *No matter what the issue is, it is always remembered to mention that there is a mental health problem. Not all problems are caused by mental health and there is no need to keep reminding about it. It feels like people are being labeled.* [Respondent #3732, Mental health care]

Additionally, patients mentioned feeling offended when referring to mental health problems that had not been relevant for many years:

> *It feels bad to see old mental health diagnoses appearing during a visit, even though they haven't been relevant for years.* [Respondent #0315, Other conditions care]

Word choice was also a recurrent theme: specific words were mentioned as offensive, and one patient explained how the condition had been described from the patient's point of view:

> *Word choices, eg speaking of a schizophrenic instead of speaking of a person having schizophrenia (a significant difference in tone, although the content is the same).* [Respondent #3750, Mental health care]

Health care professionals' opinions, conclusions, and descriptions of patients were often reported as offensive. One

XSL•FO
RenderX

patient receiving mental health care described their offense at professionals' conclusions:

> *The doctor considered me mentally unstable due to my respiratory symptoms.* [Respondent #3701, Mental health care]

Feeling Offended and Experiences of Offensive Information in the EHR

The patients who had received mental health care, including those who had received care for cancer at the same time, were more likely to feel offended by information in their EHRs. Both groups most frequently reported errors as offensive in their EHR. Among patients who had received mental health care, unnecessary information in their EHR was the second most frequently reported offense, whereas in the other patient group, health care professionals' disrespectful language was similarly the second most frequently reported offense.

Discussion

Principal Findings and Comparison With Prior Work

The patients who had received mental health care, female patients, students, patients with bad or very bad health conditions, those outside working life, or those with a bachelor's degree or higher were significantly more likely to feel offended by information in their EHR. Patients most frequently experienced errors, health care professionals' disrespectful language, and unnecessary information as offensive content in their EHRs.

Most patients who read their EHRs did not feel offended, which is consistent with prior work [13,22]. However, patients who had received mental health care were more often offended than other patient groups, as reported in prior findings from Sweden and Norway [25,26]. This may be partly explained by heightened sensitivity to perceived stigma in mental health contexts. The potentially stigmatizing nature of mental health information was also reported in previous research [63,64]. Prior research has shown that individuals with mental health conditions often anticipate or experience negative bias from health care professionals [65]. In this study, patients who received cancer care were not more likely to be offended, which reflects findings from prior qualitative studies [19,23]. However, our study shows that patients who have received both cancer and mental health care are more likely to feel offended.

The findings that females and patients with bad or very bad health conditions were more likely to feel offended are in accordance with Fernández et al [22]. Those offensive experiences with the EHR may reflect other care as well. For example, Silva et al [66] have described how the EHR may negatively affect women's relationships with health care professionals and attitudes toward health care. On the other hand, older patients were less likely to feel offended by information. This is consistent with previous findings of Hagström et al [67], where more than a quarter of adolescents and young adults had been offended by reading their EHR.

This study found that some patients experienced certain information in their EHR as offensive. Errors, health care professionals' disrespectful language, and unnecessary information were the most frequently mentioned reasons for being offended. These findings suggest that both accuracy and a respectful tone are central to how patients evaluate clinical documentation. These results match those observed in earlier studies in the United States [22,33,63,68,69]. Another important finding is that patients who received mental health care more frequently reported unnecessary information and professionals' opinions, conclusions, and word choices as offensive compared to those who had not received mental health care. One reason could be patients' heightened sensitivity to language and labeling. Evaluating the intelligence level, professionals' opinions and conclusions, information that was not corrected, and outdated information have not been pointed out in previous studies as reasons for feeling offended. Even though information can be medically essential, it should be formulated politely, and health care professionals should receive support in their writing practices.

In accordance with these results, previous studies have observed that patients take offense if the information is not consistent with the message that was verbally conveyed during the appointment [12,22] or if the patient disagrees and is not heard [22]. Thus, a better understanding of medical information from the patient's point of view can also determine the relationship between patients and professionals. In a previous study, the patient not being heard was even described as a "blind spot" that might affect other aspects, such as patient safety or quality of care [70].

Practical Implications

Based on these findings, we formulated guidelines to support the development of EHRs. First, information in EHRs should be coherent. Errors and omissions might be caused by technical issues or the health care professional's writing process. Technical coherence in processing the information should be guaranteed. In addition, patients should be able to report potential errors and omissions to professionals because uncorrected records were explicitly identified as offensive by some respondents, and patients face challenges in correcting errors in their EHRs [8,24].

Second, health care professionals should tell patients why certain information is relevant. Patients, particularly those with mental health conditions, could be informed about how some outdated or unnecessary information might be important from a medical point of view, even though it would be considered old or irrelevant by the patient. These patients interpreted such mentions as labeling or stigmatizing, even when the information may have had medical value. Additionally, patients should have the option to limit the sharing of their information or receive notifications if their information needs to be shared.

Third, health care professionals should receive education on how to highlight patients' perspectives in their records and receive recommendations on appropriate descriptions of patients, their reactions, and word choices to avoid language that patients perceive as offensive. Moreover, technical tools, such as word lists or standard phrases, could support or remind professionals when writing records. The respondents' examples of differences in tone, such as using "a person having schizophrenia" instead

XSL•FO
RenderX

of "a schizophrenic," describe how patient feedback could be used to guide health care professionals. Volkow et al have suggested using that person-centered language, reducing the stigma around mental health conditions [71]. Word lists and standard phrases should be created by involving patients, caregivers, and professionals in the planning process to ensure that they reduce the risk of offense when patients have access to their EHR.

Fourth, EHRs can influence communication between patients and health care professionals. Even though certain information might not be medically relevant to the current situation, it could be important for the patient. Additionally, professionals could explain their conclusions to patients, especially those with mental health conditions, as these might otherwise be considered offensive. This could support patients' understanding of their health information, which may be difficult to comprehend [8,24,72].

Limitations and Future Directions

The generalizability of these results has certain limitations. This study is based on a relatively small sample of participants. It was limited to one patient portal, My Kanta, in Finland, with a short 3-week period and a low response rate of 0.37% (4719/1,262,708). The response rate is, however, consistent with other similar surveys, such as in Sweden, which had a response rate of 0.61% (2587/423,141) [73], and a previous research in Finland, which reported a response rate of 0.7% (3139/449,922) [72]. Although there were more older adults and females among the respondents, they represent the most active users of the Finnish national patient portal [74]. Additionally, they align with the respondents from a previous study on the Finnish national patient portal for pharmacy customers [75]. Since the survey was available for only 3 weeks via the patient portal, frequent users may be overrepresented. The patient groups displayed variability in participant demographics, making the creation of unbiased subgroups inapplicable. However, the variability facilitated reaching different groups.

Even though this survey was offered to all national patient portal users, the language options were limited to the official languages of Finland: Finnish and Swedish. Thus, this study had limited data of those who are not able to use Finnish or Swedish. Future research is needed to explore how nonnative patient groups may experience the content of the EHR and whether experiences of offense vary.

Second, the contexts of language, expectations, and EHR writing practices may vary by culture and country. Future research could usefully explore how cultural variations may affect patients' experiences. However, the attitudes and stigma surrounding mental health conditions have been widely identified [71]. One limitation of the study was that the concept of offense was not defined, and it was measured with a single-item question. In addition, the survey as a research method has limited opportunities to request more detailed information from the respondents. On the other hand, the survey provided anonymity, making it easier for respondents to describe challenging topics related to offensive health information in their own words. Even though the responses were shorter than the interview would allow, we were able to collect experiences from a large sample of patients using the survey. Future research using interviews or focus groups would enhance a deeper understanding of these findings.

This study focused on patients' experiences. Since errors in EHRs were the most frequently mentioned reason for being offended, further work is required to thoroughly understand the nature and severity of these errors. Future research is also needed to explore whether experiences of offense had an impact on trust in care progress and the relationship between patients and professionals. Moreover, further studies that consider clinical outcomes and health care professionals' perspectives could be undertaken.

Conclusions

The purpose of this study was to explore whether certain patient groups are more likely to feel offended while reading their EHRs, the types of information that patients find offensive, and to provide a comparison across multiple patient groups using a mixed methods approach. Prior studies have often focused on single patient groups or specific clinical contexts, leaving a limited understanding of differences across multiple patient groups. This study contributes new knowledge by identifying differences across multiple patient groups. The results highlight differences in experiences of offense between those who have received mental health care and those who have not. Health care professionals could consider that some patients might feel offended by certain types of information identified in this study or tone of expressions in their EHR. This is particularly important for patients with mental health conditions because they are more likely to be offended by reading their EHR.

Health care professionals could be provided with education that includes this information, enabling them to use that knowledge as part of the process. Improving the quality of EHRs could strengthen the relationship between patients and professionals. Furthermore, to support communication between professionals and patients, patients could receive information about why certain specific themes are necessary to record, even though they might be sensitive topics and feel unnecessary in that context.

Acknowledgments

The authors would like to thank Stanislav Chren, Fabian Fagerholm, and Marko Nieminen for providing their expertise at the beginning of the writing process. Additionally, the authors acknowledge all those who participated in designing the NORDeHEALTH 2022 Patient Survey. The authors declare the use of generative artificial intelligence (GAI) in the research and writing processes. According to the GAIDeT taxonomy [76], the following tasks were delegated to GAI tools under full human supervision: proofreading and editing. The GAI tools used were Aalto University ChatGPT and Grammarly. The responsibility

XSL•FO
RenderX

for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. The declaration was submitted by SS.

Funding

This work was supported by the Strategic Research Council at the Academy of Finland (grant numbers 352501 and 352503) and by NordForsk (project number 100477).

Data Availability

The quantitative datasets analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: SS, SK, ÅC, AK, BK, BW, RMÅ, MH
Data curation: SS
Formal analysis: SS
Funding acquisition: SK, MH
Methodology: SS, SK, MH
Supervision: SK
Visualization: SS
Writing – original draft: SS
Writing – review and editing: SS, SK, ÅC, AK, BK, BW, RMÅ, MH

Conflicts of Interest

None declared.

Multimedia Appendix 1
The survey questions in English.
[PDF File, 39 KB - jmir_v28i1e86178_app1.pdf]

Multimedia Appendix 2
Felt offended and background variables table.
[PDF File, 100 KB - jmir_v28i1e86178_app2.pdf]

Multimedia Appendix 3
Evaluation of models fit table.
[PDF File, 179 KB - jmir_v28i1e86178_app3.pdf]

Checklist 1
CHERRIES checklist.
[PDF File, 54 KB - jmir_v28i1e86178_app4.pdf]

References

1. Nguyen L, Bellucci E, Nguyen LT. Electronic health records implementation: an evaluation of information system impact and contingency factors. Int J Med Inform 2014 Nov;83(11):779-796. [doi: 10.1016/j.ijmedinf.2014.06.011] [Medline: 25085286]
2. Bonomi S. The electronic health record: a comparison of some european countries. In: Ricciardi F, Harfouche A, editors. Information and Communication Technologies in Organizations and Society Past, Present and Future Issues: Springer; 2016:33-50. [doi: 10.1007/978-3-319-28907-6_3]
3. Hagström J, Blease C, Haage B, Scandurra I, Hansson S, Hägglund M. Use of and experiences with online access to electronic health records for parents, children, and adolescents: protocol for a scoping review. JMIR Res Protoc 2022 Jun 15;11(6):e36158. [doi: 10.2196/36158] [Medline: 35704386]
4. Sun R, Korytkowski MT, Sereika SM, Saul MI, Li D, Burke LE. Patient portal use in diabetes management: literature review. JMIR Diabetes 2018 Nov 6;3(4):e11199. [doi: 10.2196/11199] [Medline: 30401665]
5. Portz JD, Bayliss EA, Bull S, et al. Using the technology acceptance model to explore user experience, intent to use, and use behavior of a patient portal among older adults with multiple chronic conditions: descriptive qualitative study. J Med Internet Res 2019 Apr 8;21(4):e11604. [doi: 10.2196/11604] [Medline: 30958272]

XSL•FO
RenderX

6. Gerard M, Fossa A, Folcarelli PH, Walker J, Bell SK. What patients value about reading visit notes: a qualitative inquiry of patient experiences with their health information. J Med Internet Res 2017 Jul 14;19(7):e237. [doi: 10.2196/jmir.7212] [Medline: 28710055]
7. Delbanco T, Walker J, Darer JD, et al. Open notes: doctors and patients signing on. Ann Intern Med 2010 Jul 20;153(2):121-125. [doi: 10.7326/0003-4819-153-2-201007200-00008] [Medline: 20643992]
8. Simola S, Hörhammer I, Xu Y, et al. Patients' experiences of a national patient portal and its usability: cross-sectional survey study. J Med Internet Res 2023 Jun 30;25(1):e45974. [doi: 10.2196/45974] [Medline: 37389909]
9. Dobscha SK, Denneson LM, Jacobson LE, Williams HB, Cromer R, Woods S. VA mental health clinician experiences and attitudes toward OpenNotes. Gen Hosp Psychiatry 2016;38:89-93. [doi: 10.1016/j.genhosppsych.2015.08.001] [Medline: 26380876]
10. Zanaboni P, Kristiansen E, Lintvedt O, et al. Impact on patient-provider relationship and documentation practices when mental health patients access their electronic health records online: a qualitative study among health professionals in an outpatient setting. BMC Psychiatry 2022 Jul 28;22(1):508. [doi: 10.1186/s12888-022-04123-7] [Medline: 35902841]
11. Esch T, Mejilla R, Anselmo M, Podtschaske B, Delbanco T, Walker J. Engaging patients through open notes: an evaluation using mixed methods. BMJ Open 2016 Jan 29;6(1):e010034. [doi: 10.1136/bmjopen-2015-010034] [Medline: 26826154]
12. Woods SS, Schwartz E, Tuepker A, et al. Patient experiences with full electronic access to health records and clinical notes through the My HealtheVet Personal Health Record Pilot: qualitative study. J Med Internet Res 2013 Mar 27;15(3):e65. [doi: 10.2196/jmir.2356] [Medline: 23535584]
13. Delbanco T, Walker J, Bell SK, et al. Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead. Ann Intern Med 2012 Oct 2;157(7):461-470. [doi: 10.7326/0003-4819-157-7-201210020-00002] [Medline: 23027317]
14. Walker J, Leveille S, Bell S, et al. OpenNotes after 7 years: patient experiences with ongoing access to their clinicians' outpatient visit notes. J Med Internet Res 2019 May 6;21(5):e13876. [doi: 10.2196/13876] [Medline: 31066717]
15. Blease C, Kharko A, Hägglund M, et al. The benefits and harms of open notes in mental health: a Delphi survey of international experts. PLoS One 2021;16(10):e0258056. [doi: 10.1371/journal.pone.0258056] [Medline: 34644320]
16. Denneson LM, Chen JI, Pisciotta M, Tuepker A, Dobscha SK. Patients' positive and negative responses to reading mental health clinical notes online. Psychiatr Serv 2018 May 1;69(5):593-596. [doi: 10.1176/appi.ps.201700353] [Medline: 29493408]
17. Cromer R, Denneson LM, Pisciotta M, Williams H, Woods S, Dobscha SK. Trust in mental health clinicians among patients who access clinical notes online. Psychiatr Serv 2017 May 1;68(5):520-523. [doi: 10.1176/appi.ps.201600168] [Medline: 28142383]
18. Blease C, Dong Z, Torous J, Walker J, Hägglund M, DesRoches CM. Association of patients reading clinical notes with perception of medication adherence among persons with serious mental illness. JAMA Netw Open 2021 Mar 1;4(3):e212823. [doi: 10.1001/jamanetworkopen.2021.2823] [Medline: 33760088]
19. Kayastha N, Pollak KI, LeBlanc TW. Open oncology notes: a qualitative study of oncology patients' experiences reading their cancer care notes. J Oncol Pract 2018 Apr;14(4):e251-e258. [doi: 10.1200/JOP.2017.028605] [Medline: 29443650]
20. Reed ME, Huang J, Brand RJ, et al. Patients with complex chronic conditions: health care use and clinical events associated with access to a patient portal. PLoS One 2019;14(6):e0217636. [doi: 10.1371/journal.pone.0217636] [Medline: 31216295]
21. Muli I, Cajander Å, Hvitfeldt H, et al. To read or not to read - a cross-sectional study of Swedish primary care patients' adoption of patient accessible electronic health records. Digit Health 2024;10:20552076241287636. [doi: 10.1177/20552076241287636] [Medline: 39381824]
22. Fernández L, Fossa A, Dong Z, et al. Words matter: what do patients find judgmental or offensive in outpatient notes? J Gen Intern Med 2021 Sep;36(9):2571-2578. [doi: 10.1007/s11606-020-06432-7] [Medline: 33528782]
23. Rexhepi H, Åhlfeldt RM, Cajander Å, Huvila I. Cancer patients' attitudes and experiences of online access to their electronic medical records: a qualitative study. Health Informatics J 2018 Jun;24(2):115-124. [doi: 10.1177/1460458216658778] [Medline: 27440056]
24. Kujala S, Simola S, Wang B, et al. Benchmarking usability of patient portals in Estonia, Finland, Norway, and Sweden. Int J Med Inform 2024 Jan;181:105302. [doi: 10.1016/j.ijmedinf.2023.105302] [Medline: 38011806]
25. Wang B, Kristiansen E, Fagerlund AJ, et al. Users' experiences with online access to electronic health records in mental and somatic health care: cross-sectional study. J Med Internet Res 2023 Dec 25;25:e47840. [doi: 10.2196/47840] [Medline: 38145466]
26. Fagerlund AJ, Bärkås A, Kharko A, et al. Experiences from patients in mental healthcare accessing their electronic health records: results from a cross-national survey in Estonia, Finland, Norway, and Sweden. BMC Psychiatry 2024 Jul 2;24(1):481. [doi: 10.1186/s12888-024-05916-8] [Medline: 38956493]
27. Kahn MW, Bell SK, Walker J, Delbanco T. Let's show patients their mental health records. JAMA 2014 Apr 2;311(13):1291-1292. [doi: 10.1001/jama.2014.1824]
28. Erlingsdóttir G, Petersson L, Jonnergård K. A theoretical twist on the transparency of open notes: qualitative analysis of health care professionals' free-text answers. J Med Internet Res 2019 Sep 25;21(9):e14347. [doi: 10.2196/14347] [Medline: 31573905]

XSL•FO
RenderX

29. Fix GM, Hogan TP, Amante DJ, McInnes DK, Nazi KM, Simon SR. Encouraging patient portal use in the patient-centered medical home: three stakeholder perspectives. J Med Internet Res 2016 Nov 22;18(11):e308. [doi: 10.2196/jmir.6488] [Medline: 27876686]
30. Earnest MA, Ross SE, Wittevrongel L, Moore LA, Lin CT. Use of a patient-accessible electronic medical record in a practice for congestive heart failure: patient and physician experiences. J Am Med Inform Assoc 2004;11(5):410-417. [doi: 10.1197/jamia.M1479] [Medline: 15187074]
31. Lindholm C, Erlingsdóttir G. eHealth and the medical profession. In: Erlingsdóttir G, Sandberg H, editors. eHealth Opportunities and Challenges: A White Paper, 1st edition: The Pufendorf Institute of Advanced Studies, Lund University; 2016:42-47 URL: https://portal.research.lu.se/en/publications/a443b459-087c-420f-99c5-401d46b74140 [accessed 2026-04-14]
32. de Lusignan S, Mold F, Sheikh A, et al. Patients' online access to their electronic health records and linked online services: a systematic interpretative review. BMJ Open 2014 Sep 8;4(9):e006021. [doi: 10.1136/bmjopen-2014-006021] [Medline: 25200561]
33. Lau F, Antonio M, Davison K, Queen R, Devor A. A rapid review of gender, sex, and sexual orientation documentation in electronic health records. J Am Med Inform Assoc 2020 Nov 1;27(11):1774-1783. [doi: 10.1093/jamia/ocaa158] [Medline: 32935124]
34. Wells S, Mahony F, Huang Y, Day K. Perspectives of New Zealand patients and GPs at the beginning of patient portal implementation. J Prim Health Care 2019 Dec 18;11(4):315-326. [doi: 10.1071/HC19016]
35. Chimowitz H, O'Neill S, Leveille S, Welch K, Walker J. Sharing psychotherapy notes with patients: therapists' attitudes and experiences. Soc Work 2020 Apr 1;65(2):159-168. [doi: 10.1093/sw/swaa010] [Medline: 32236447]
36. Schwarz J, Neumann K, Meier-Diedrich E, Mühlensiepen F, Hochwarter S, Blease C. Psychotherapists' views on open notes: an online survey from Germany. Digit Health 2024;10:20552076241271813. [doi: 10.1177/20552076241271813] [Medline: 39291155]
37. Peck P, Torous J, Shanahan M, Fossa A, Greenberg W. Patient access to electronic psychiatric records: a pilot study. Health Policy Technol 2017 Sep;6(3):309-315. [doi: 10.1016/j.hlpt.2017.06.003]
38. van Rijt AM, Hulter P, Weggelaar-Jansen AM, Ahaus K, Pluut B. Mental health care professionals' appraisal of patients' use of web-based access to their electronic health record: qualitative study. J Med Internet Res 2021 Aug 27;23(8):e28045. [doi: 10.2196/28045] [Medline: 34448705]
39. Schwarz J, Bärkås A, Blease C, et al. Sharing clinical notes and electronic health records with people affected by mental health conditions: scoping review. JMIR Ment Health 2021 Dec 14;8(12):e34170. [doi: 10.2196/34170] [Medline: 34904956]
40. Petersson L, Erlingsdóttir G. Open notes in Swedish psychiatric care (part 1): survey among psychiatric care professionals. JMIR Ment Health 2018 Feb 2;5(1):e11. [doi: 10.2196/mental.9140] [Medline: 29396386]
41. Ling S, Kassam I, Haider S, Lo B, Strudwick G. Pre-implementation perceptions of clinicians regarding a patient portal in inpatient addictions settings: a qualitative descriptive study. Int J Med Inform 2023 Dec;180:105278. [doi: 10.1016/j.ijmedinf.2023.105278] [Medline: 37922661]
42. Kung S, Lapid MI, Swintak CC, Agarwal Z, Ryan SM, Lineberry TW. Survey of sensitive information written in patient notes by psychiatry trainees. Acad Psychiatry 2013 Jan 1;37(1):31-34. [doi: 10.1176/appi.ap.11110195] [Medline: 23338870]
43. Härkönen H, Lakoma S, Verho A, et al. Impact of digital services on healthcare and social welfare: an umbrella review. Int J Nurs Stud 2024 Apr;152:104692. [doi: 10.1016/j.ijnurstu.2024.104692] [Medline: 38301306]
44. Jormanainen V. Large-scale implementation and adoption of the Finnish national Kanta services in 2010–2017: a prospective, longitudinal, indicator-based study. Finn J eHealth eWelf 2018;10(4):381-395. [doi: 10.23996/fjhw.74511]
45. Jormanainen V, Lindgren M, Keskimäki I, Kaila M. Use of My Kanta in Finland 2010-2022. Stud Health Technol Inform 2023 Jun 29;305:448-451. [doi: 10.3233/SHTI230528] [Medline: 37387062]
46. Blease C, Torous J, Kharko A, et al. Preparing patients and clinicians for open notes in mental health: qualitative inquiry of international experts. JMIR Ment Health 2021 Apr 16;8(4):e27397. [doi: 10.2196/27397] [Medline: 33861202]
47. Liu C. Toward a theory of offense: should you feel offended? Philosophy 2021 Oct;96(4):625-649. [doi: 10.1017/S0031819121000140]
48. Eysenbach G. Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
49. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs-principles and practices. Health Serv Res 2013 Dec;48(6 Pt 2):2134-2156. [doi: 10.1111/1475-6773.12117] [Medline: 24279835]
50. Fond G, Pauly V, Brousse Y, et al. Mental health care utilization and prescription rates among children, adolescents, and young adults in France. JAMA Netw Open 2025 Jan 2;8(1):e2452789. [doi: 10.1001/jamanetworkopen.2024.52789] [Medline: 39777444]
51. Tam J, Mezuk B, Zivin K, Meza R. U.S. simulation of lifetime major depressive episode prevalence and recall error. Am J Prev Med 2020 Aug;59(2):e39-e47. [doi: 10.1016/j.amepre.2020.03.021] [Medline: 32446751]
52. Zhang Z, Jackson SL, Gillespie C, Merritt R, Yang Q. Depressive symptoms and mortality among US adults. JAMA Netw Open 2023 Oct 2;6(10):e2337011. [doi: 10.1001/jamanetworkopen.2023.37011] [Medline: 37812418]

XSL•FO
RenderX

53. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA Cancer J Clin 2024;74(1):12-49. [doi: 10.3322/caac.21820] [Medline: 38230766]
54. Greenhalgh T, Wood GW, Bratan T, Stramer K, Hinder S. Patients' attitudes to the summary care record and HealthSpace: qualitative study. BMJ 2008 Jun 7;336(7656):1290-1295. [doi: 10.1136/bmj.a114] [Medline: 18511764]
55. Bernaerdt J, Moerenhout T, Devisch I. Vulnerable patients' attitudes towards sharing medical data and granular control in patient portal systems: an interview study. J Eval Clin Pract 2021 Apr;27(2):429-437. [doi: 10.1111/jep.13465] [Medline: 32886953]
56. Kaihlanen AM, Virtanen L, Buchert U, et al. Towards digital health equity - a qualitative study of the challenges experienced by vulnerable groups in using digital health services in the COVID-19 era. BMC Health Serv Res 2022 Feb 12;22(1):188. [doi: 10.1186/s12913-022-07584-4] [Medline: 35151302]
57. Hägglund M, Kharko A, Hagström J, et al. The NORDeHEALTH 2022 patient survey: cross-sectional study of national patient portal users in Norway, Sweden, Finland, and Estonia. J Med Internet Res 2023 Nov 13;25:e47573. [doi: 10.2196/47573] [Medline: 37955963]
58. Park T, Davis CS. A test of the missing data mechanism for repeated categorical data. Biometrics 1993 Jun;49(2):631-638. [doi: 10.2307/2532576] [Medline: 8369395]
59. Dziura JD, Post LA, Zhao Q, Fu Z, Peduzzi P. Strategies for dealing with missing data in clinical trials: from design to analysis. Yale J Biol Med 2013 Sep;86(3):343-358. [Medline: 24058309]
60. Harahap NC, Handayani PW, Hidayanto AN. Barriers and facilitators of personal health record adoption in Indonesia: health facilities' perspectives. Int J Med Inform 2022 Jun;162:104750. [doi: 10.1016/j.ijmedinf.2022.104750] [Medline: 35339888]
61. Saunders B, Sim J, Kingstone T, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907. [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
62. Lazar J, Feng JH, Hochheiser H. Research Methods in Human-Computer Interaction, 1st edition: Wiley; 2010.
63. Barcelona V, Scharp D, Idnay BR, Moen H, Cato K, Topaz M. Identifying stigmatizing language in clinical documentation: a scoping review of emerging literature. PLoS One 2024;19(6):e0303653. [doi: 10.1371/journal.pone.0303653] [Medline: 38941299]
64. Kariotis T, Prictor M, Gray K, Chang S. Service users' perspectives on an integrated electronic care record in mental health care: qualitative vignette and interview study. JMIR Med Inform 2025 Jun 3;13:e64162. [doi: 10.2196/64162] [Medline: 40460429]
65. Corrigan PW, Druss BG, Perlick DA. The impact of mental illness stigma on seeking and participating in mental health care. Psychol Sci Public Interest 2014 Oct;15(2):37-70. [doi: 10.1177/1529100614531398] [Medline: 26171956]
66. Silva JM, Durden TE, Hirsch A. Erasing inequality: examining discrepancies between electronic health records and patient narratives to uncover perceived stigma and dismissal in clinical encounters. Soc Sci Med 2023 Apr;323:115837. [doi: 10.1016/j.socscimed.2023.115837] [Medline: 36931038]
67. Hagström J, Blease C, Moll J, Rexhepi H, Scandurra I, Hägglund M. Adolescents' and young adults' experiences of offense from reading their health records online. Stud Health Technol Inform 2024 Jan 25;310:1422-1423. [doi: 10.3233/SHTI231225] [Medline: 38269677]
68. Fisher KA, Kennedy K, Bloomstone S, Fukunaga MI, Bell SK, Mazor KM. Can sharing clinic notes improve communication and promote self-management? A qualitative study of patients with COPD. Patient Educ Couns 2022 Mar;105(3):726-733. [doi: 10.1016/j.pec.2021.06.004] [Medline: 34175167]
69. Nadler LE, Ogden SN, Scheffey KL, Cronholm PF, Dichter ME. Provider practices and perspectives regarding collection and documentation of gender identity. J Homosex 2021 May 12;68(6):901-913. [doi: 10.1080/00918369.2019.1667162] [Medline: 31526306]
70. Bell SK, Bourgeois F, Dong J, et al. Patient identification of diagnostic safety blindspots and participation in "good catches" through shared visit notes. Milbank Q 2022 Dec;100(4):1121-1165. [doi: 10.1111/1468-0009.12593] [Medline: 36539389]
71. Volkow ND, Gordon JA, Koob GF. Choosing appropriate language to reduce the stigma around mental illness and substance use disorders. Neuropsychopharmacology 2021 Dec;46(13):2230-2232. [doi: 10.1038/s41386-021-01069-4] [Medline: 34276051]
72. Kujala S, Hörhammer I, Väyrynen A, et al. Patients' experiences of web-based access to electronic health records in Finland: cross-sectional survey. J Med Internet Res 2022 Jun 6;24(6):e37438. [doi: 10.2196/37438] [Medline: 35666563]
73. Moll J, Rexhepi H, Cajander Å, et al. Patients' experiences of accessing their electronic health records: national patient survey in Sweden. J Med Internet Res 2018 Nov 1;20(11):e278. [doi: 10.2196/jmir.9492] [Medline: 30389647]
74. Jormanainen V. Over 89% adoption rate of the nationwide online patient portal in Finland. Stud Health Technol Inform 2022 May 25;294:589-593. [doi: 10.3233/SHTI220537] [Medline: 35612157]
75. Sääskilahti M, Ahonen R, Timonen J. Pharmacy customers' experiences of use, usability, and satisfaction of a nationwide patient portal: survey study. J Med Internet Res 2021 Jul 16;23(7):e25368. [doi: 10.2196/25368] [Medline: 34269687]
76. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (Generative AI Delegation Taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res 2026 Apr;33(3):2544331. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]

XSL•FO
RenderX

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
EHR: electronic health record

Edited by S Brini; submitted 20.Oct.2025; peer-reviewed by J Schwarz, N Buderer, S Ahn; revised version received 14.Apr.2026; accepted 15.Apr.2026; published 14.May.2026.

Please cite as:
Simola S, Kujala S, Cajander Å, Kharko A, Kane B, Wang B, Åhlfeldt RM, Hägglund M
Understanding Patient-Reported Offenses in Electronic Health Records: Cross-Sectional Mixed Methods Survey
J Med Internet Res 2026;28:e86178
URL: https://www.jmir.org/2026/1/e86178
doi:10.2196/86178

© Saija Simola, Sari Kujala, Åsa Cajander, Anna Kharko, Bridget Kane, Bo Wang, Rose-Mharie Åhlfeldt, Maria Hägglund. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Patient Experiences With Online Laboratory Test Presentations From Access to Activation: Systematic Review

Mia Liza A Lustria[1], PhD; Lovinta Atrinawati[1], MEng; Obianuju Aliche[2], PhD; Kyunghye Kim[3], PhD; Zhe He[1], PhD

[1]School of Information, Florida State University, 142 Collegiate Loop, Tallahassee, FL, United States

[2]Department of Quantitative Health Sciences, Mayo Clinic, Jacksonville, FL, United States

[3]Florida State University Libraries, Florida State University, Florida, United States

Corresponding Author:
Zhe He, PhD
School of Information, Florida State University, 142 Collegiate Loop, Tallahassee, FL, United States

Abstract

Background: Federal regulations require that laboratory test results be released to patients through online portals in near–real time, often before clinicians review or contextualize them. While these policies expand transparency and affirm patients' rights to their health information, access alone does not ensure that patients can make sense of their results or determine when and how to act. How regulatory mandates for transparency translate into appropriate engagement remains poorly understood.

Objective: We aim to synthesize evidence on how patients access, interpret, and act upon laboratory test results delivered online; identify factors associated with comprehension and engagement; and evaluate how display design, clinical workflow, and patient characteristics shape engagement across the access-to-activation continuum.

Methods: Following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidance, we systematically searched 5 databases—Web of Science, Embase, PubMed, CINAHL, and Library, Information Science & Technology Abstracts—for English-language, peer-reviewed empirical studies published from January 2013 through April 2026, supplemented by hand searching and citation tracking. Eligible studies examined patient access to, interpretation of, or responses to laboratory results delivered through portals. Records were independently screened, appraised for methodological quality, and synthesized using thematic analysis across a 3-stage continuum: access, interpretation, and activation. The protocol was not registered.

Results: Thirty-nine studies met the inclusion criteria; most were from the United States. Retrospective analyses documented increases in result viewing, secure messaging, and follow-up visits after immediate-release policies. Patients value timely access but often struggle to determine the significance of borderline values and when follow-up is warranted. Visual and textual cues can improve clarity but may also overemphasize minor deviations or inadequately signal clinically significant findings. Patients frequently monitor portals and initiate contact to confirm clinician review, indicating that increased portal activity often reflects efforts to resolve uncertainty rather than confident self-management. Disparities in access, comprehension, and confidence were observed among older adults, non–English-speaking patients, those with public insurance, and those with lower health or digital literacy. Measures of comprehension and engagement varied widely.

Conclusions: Expanded online access improves transparency but does not guarantee accurate interpretation or appropriate follow-up. In contrast to prior reviews examining patient perceptions, comprehension, and presentation formats in isolation, this synthesis traces patient experiences across an access–interpretation–activation continuum, revealing how interpretive challenges and emotional responses compound across stages. Meaningful activation depends not only on display clarity but on how informational, emotional, and workflow factors shape patients' decisions across stages. Designing for activation, therefore, requires a holistic, systems-level approach that aligns result presentation, interpretive support, and clinical workflow to signal the clinical significance of laboratory test results and guide appropriate follow-up. Such approaches must be responsive to the needs of patients across diverse literacy, language, and clinical contexts.

(J Med Internet Res 2026;28:e88259) doi:10.2196/88259

KEYWORDS

patient portals; laboratory test results; patient engagement; health communication; health literacy; clinical workflow; health disparities; health information technology; result interpretation; patient comprehension

XSL•FO
RenderX

Introduction

By law, patients must have immediate access to their laboratory test results through patient portals upon availability. Expanding digital access is mandated by federal regulations under the 21st Century Cures Act [1] and is a critical national objective of Healthy People 2030 [2]. The policies assert that broadening access can enhance patients' ability to make informed decisions and strengthen their engagement in their own care. However, most portals provide little context or guidance, placing additional cognitive and emotional demands on patients.

Laboratory results are the most frequently viewed type of health data in portals. According to the Office of the National Coordinator for Health IT, over 90% of portal users view their laboratory results [3]. Yet access alone does not ensure that patients can accurately interpret or act on their test results in a timely and appropriate manner. Access without interpretive support often leads to confusion, misinterpretation, or anxiety [4-7]. These challenges are amplified among disadvantaged and higher-risk groups. Non-White, non-English-speaking, older, lower-literacy, and publicly insured patients are less likely to view their results online. Research shows that these groups tend to have lower trust in portals, greater difficulty understanding their results, and less confidence in their ability to act on them [8-12]. Limited health or digital literacy compounds these challenges further, such that those who already face barriers to care may be less likely to benefit from online access to test results [13-15].

Recognizing these gaps, the Clinical Laboratory Improvement Advisory Committee and the Agency for Healthcare Research and Quality have highlighted the need for result delivery that goes beyond timeliness to address understandability, contextualization, and equitable access [16]. The Clinical Laboratory Improvement Advisory Committee specifically encouraged the Centers for Disease Control and Prevention to create guidelines for designing more culturally and linguistically appropriate result formats [5]. Moreover, multistakeholder forums involving patients and clinicians have underscored that unclear or delayed communication contributes to diagnostic errors and missed follow-up, highlighting the risks associated with fragmented delivery systems and poorly designed portal interfaces [17].

This shift has intensified interest in understanding not only how patients access their laboratory results, but also how they interpret those results and whether this process leads to appropriate follow-up and engagement in self-management. We conceptualize this process as a continuum from access to interpretation to activation. Access concerns when and how patients receive their results and the extent to which transparency supports or overwhelms them. Interpretation involves comprehension of numerical or textual information, emotional responses to normal or abnormal findings, and judgments about urgency or need for action. Activation reflects confidence, motivation, and concrete steps taken after viewing results, including contacting clinicians, seeking additional information, or changing health behaviors.

Several recent reviews have contributed valuable insights into how patients experience and manage access to their test results through patient portals. Petrovskaya et al [18] laid an important foundation by mapping patient and clinician perspectives on the direct release of laboratory and imaging results, highlighting emotional reactions, perceived understanding, and information-seeking behavior. van der Mee et al [19] offered complementary evidence on how different presentation formats affect comprehension of numerical laboratory data. Building on this body of work, our review integrates findings from qualitative, experimental, and observational research—including surveys, retrospective analyses, and mixed-methods studies—published through 2026 to trace the full continuum from access to interpretation to activation within the evolving policy landscape shaped by the 21st Century Cures Act [20]. By aligning findings across design, workflow, and equity perspectives, we aim to provide a more integrative view of how patients encounter, make sense of, and act on their laboratory results online and underscore the importance of connecting digital tools to clinical workflows and provider communication, and establishing system-level supports for sustained engagement.

Specifically, this systematic review addresses two questions: (1) How do patients engage with and respond to online access to laboratory test results, as reflected in both portal-use behaviors and self-reported experiences? and (2) What patterns emerge across studies regarding patients' comprehension, emotional responses, and subsequent actions after viewing laboratory test results online? It synthesizes more than a decade of empirical research and identifies recurring patterns, persistent disparities, and individual differences in experiences with online laboratory result access. It also considers their implications for the design and implementation of patient-facing systems in the era of immediate result release.

Methods

Study Design

This systematic review was conducted and reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines [21], with search methods reported following PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension [22]; see Checklist 1 for the completed PRISMA checklist).

Search Strategy

Overview

We used multiple strategies to identify eligible studies. The goal was to capture a broad range of study designs and methodologies to capture patient experiences across diverse clinical and technological contexts. A research librarian (KK) developed and executed the initial comprehensive searches across Web of Science (Clarivate Analytics), Embase (Elsevier interface), PubMed (including records not yet indexed in Medline), CINAHL (EBSCOhost), and Library, Information Science & Technology Abstracts (EBSCOhost).

XSL•FO
RenderX

The search strategy was structured around two primary conceptual domains: (1) laboratory test results and related constructs (eg, laboratory values, blood tests, and patient-facing access through portals or electronic health records [EHRs]), and (2) patient interpretation and understanding, including health literacy and related constructs such as comprehension, numeracy, and readability. Search strategies were adapted across databases to account for differences in indexing systems, controlled vocabularies, and platform-specific syntax. Controlled vocabulary terms – including MeSH (Medical Subject Headings) in PubMed, Emtree in Embase, and CINAHL Headings—were supplemented with keyword-based strategies to ensure comprehensive retrieval by including potentially nonindexed records. We refined the search strategies iteratively over this study's period by expanding keyword sets and incorporating additional controlled vocabulary terms.

Searches were limited to English-language, peer-reviewed empirical studies involving human participants and published between January 2013 and April 2026. Where database functionality permitted, we applied filters to exclude nonempirical publication types such as editorials, commentaries, and letters.

We also hand-searched key journals (eg, *JAMIA*, *BMC Medical Informatics and Decision Making*, *Clinical Chemistry and Laboratory Medicine*, *JMIR*, and *JMIR Medical Informatics*) and examined the bibliographies of related systematic reviews [18,19,23,24]. Two reviewers (ZH and MLAL) identified key papers from the initial search, manually reviewed their reference lists, and conducted citation tracking using Google Scholar.

The initial database searches were conducted in September 2022 and updated in July 2023, June 2024, and October 2025. We reran the refined search strategy across all databases and supplementary sources in April 2026 to ensure that the review was both comprehensive and current. Complete database-specific search strategies, including exact Boolean syntax, field tags, applied limits, and search dates, are provided in Multimedia Appendix 1 [8-14,25-56] in accordance with PRISMA-S reporting standards [22].

Eligibility Criteria and Initial Screening

Two authors, ZH and MLAL, developed the inclusion and exclusion criteria to identify papers eligible for full-text review. We included studies if they examined any of the following: (1) patient or caregiver perceptions or experiences accessing laboratory test results through patient portals; (2) factors associated with patients' access to, interpretation of, or understanding of laboratory test results presented through portals or EHRs; (3) presentation styles, display formats, visualization approaches, or interface designs for viewing laboratory test results online; (4) strategies or interventions designed to improve patient access to, comprehension of, or decision-making based on online laboratory test results; (5) patient use of laboratory-test-related functions within patient portals or similar systems.

Studies were excluded if they met any of the following conditions: (1) focused solely on notification or communication of laboratory test results; (2) examined general patient portal use without a specific focus on laboratory result viewing; (3) focused primarily on the perspectives of clinicians, students, or other health professionals; (4) primarily dealt with imaging or radiology reports, or with diagnostic test results (eg, COVID-19 polymerase chain reaction tests, genomic tests, or pharmacogenomic tests); (5) did not report primary empirical data (eg, editorials, commentaries, narrative reviews, and protocols).

Study Selection and Screening

All records identified through database searches, hand-searching of key journals, citation tracing, and screening of related systematic reviews were imported to Covidence (Veritas Health Innovation Ltd), a web-based systematic review management platform. In total, 8973 records were identified across all databases and supplementary sources following the final updated search in April 2026. Covidence automatically removed duplicate records across databases, and remaining duplicates were removed manually. Two reviewers (MLAL and ZH) independently screened titles and abstracts, followed by full-text screening of potentially relevant papers. Disagreements at either stage were resolved through discussion or, when necessary, consultation with a third reviewer. Reasons for exclusion were documented in Covidence and are summarized in the PRISMA flow diagram (see Results section).

Quality Assessment

LA and OA independently evaluated the methodological quality and risk of bias of the 39 studies included in the final synthesis using appropriate appraisal tools. We used JBI (Joanna Briggs Institute) critical appraisal tools to assess risk of bias for randomized controlled trials [57] and quasi-experimental studies [58]. Qualitative studies were appraised using the Critical Appraisal Skills Programme qualitative studies checklist [59], survey-based studies were assessed using the Appraisal of Cross-Sectional Studies tool [60], mixed methods studies were assessed using the Mixed Methods Appraisal Tool [61], and retrospective studies were assessed using an adapted version of the Newcastle-Ottawa Scale [62].

For each study, overall risk-of-bias classifications (low, moderate, or high) were derived qualitatively based on domain-level judgments within each appraisal tool (low, some concerns, or high). We excluded studies only when critical methodological limitations precluded meaningful interpretation of the findings—ie, fundamental limitations in study design or reporting (eg, insufficient description of methods, unclear data sources, or inability to assess analytic validity). Overall judgments were based on structured reviewer evaluation rather than numeric scoring. Disagreements between reviewers were resolved through discussion and consensus. Multimedia Appendix 1 includes risk of bias tables reporting results of the quality appraisal for each study.

Data Extraction and Synthesis

Overview

To guide data collection, MLAL developed and piloted a structured table to extract study characteristics and outcome domains relevant to the review objectives, including

XSL•FO
RenderX

comprehension, risk interpretation, emotional responses, and activation-related behaviors. Further, 2 reviewers (LA and OA) independently extracted data using this table. Discrepancies were resolved through comparison and discussion.

Analysis Strategy

Following data extraction, we conducted a recursive thematic synthesis. As the included studies varied substantially in design, populations, and outcome measures, quantitative synthesis through meta-analysis was not appropriate. Moreover, key constructs such as comprehension, interpretation, emotional responses, and activation-related behaviors were operationalized in diverse and noncomparable ways, precluding calculation of a common effect size. Furthermore, no subset of studies was sufficiently comparable in terms of outcomes, measures, and study design to support a meaningful subgroup meta-analysis. The narrative synthesis was conducted and reported in accordance with guidance for synthesis without meta-analysis [63] (Checklist 2).

Coding and Thematic Synthesis

We developed initial codes through close reading of extracted study findings, then 2 reviewers (LA and OA) independently applied and refined these codes. Coding proceeded iteratively, with codes merged, split, or refined as patterns emerged. Discrepancies were documented and resolved through structured comparison of coded content followed by discussion. When needed, 2 additional reviewers (MLAL and ZH) reviewed disputed cases to reach consensus.

We used thematic analysis to identify recurring patterns and areas of divergence across studies. These inductively derived patterns were subsequently organized within a 3-stage continuum reflecting patients' experiences with online laboratory test results—access, interpretation, and activation—to reflect the sequential yet interconnected nature of patients' experiences with online laboratory test results.

Protocol and Registration

A working protocol guided all stages of the review, and it was refined as methodological decisions were finalized. The research questions that shaped data extraction and synthesis were established prospectively and remained consistent, though the scope broadened over time in response to the evolving literature. The stage-based synthesis framework—access, interpretation, and activation—was developed iteratively during analysis and was not fully specified in the initial protocol, consistent with accepted practice in thematic synthesis where analytic frameworks commonly emerge from engagement with the data. The protocol was not publicly registered.

Results

Overview

Figure 1 presents the PRISMA 2020 flow diagram summarizing study identification, screening, eligibility assessment, and inclusion. Following duplicate removal, title and abstract screening, full-text review, and quality appraisal, 39 studies met the inclusion criteria and were retained for final synthesis.

XSL•FO
RenderX

Figure 1. PRISMA flow diagram. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

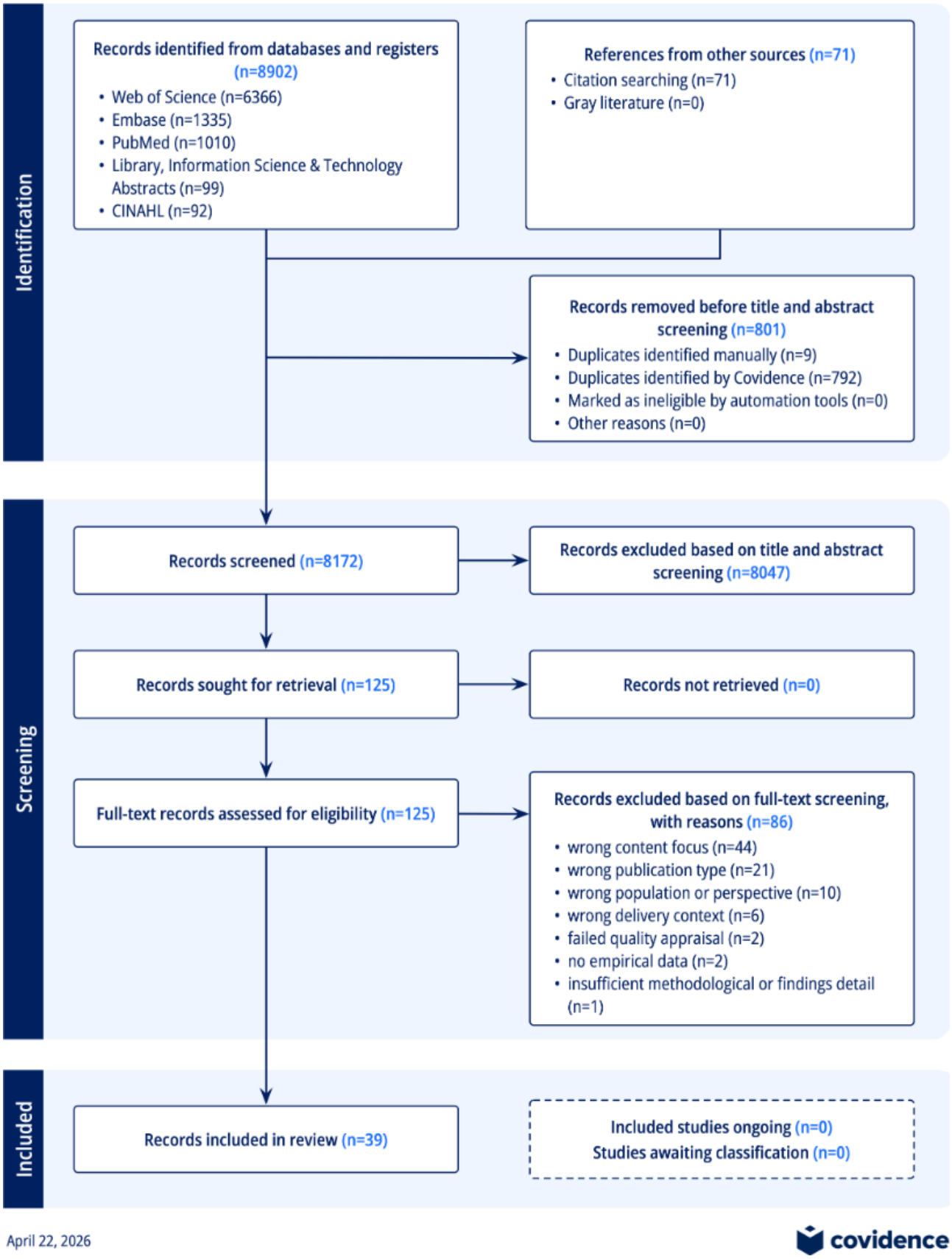

Characteristics of Included Studies

A total of 39 studies met the inclusion criteria. These represented a wide range of study designs, settings, and patient populations. Study designs included qualitative studies (n=9), retrospective analyses (n=8), survey research (n=7), experimental or quasi-experimental methods (n=8), and mixed-methods research (n=7). Most studies were conducted in the United States (n=25),

XSL•FO
RenderX

followed by Canada (n=8), the Netherlands (n=5), and the United Kingdom (n=1). Table 1 includes an overview of included studies and their mapping to the access–interpretation–activation framework. (Expanded descriptions of studies, including measures and main findings, are found in Multimedia Appendix 2, Detailed Summary of Included Studies [8-14,25-56].

XSL•FO
RenderX

Table . Characteristics of included studies mapped to stages. Multimedia Appendix 2

Authors and country	Study design and purpose	Sample characteristics	Access	Interpretation	Activation
Bhalla et al [8], United States	Retrospective study: assessed changes in patient viewing of diagnostic test results via a patient portal before and after the 21st Century Cures Act.	Patients with cancer with portal-released diagnostic test results (n=44,419; 2017 - 2022).	P[a]		
Christensen and Sue [55], United States	Cross-sectional survey: examined patients' emotional responses and follow-up behaviors after viewing laboratory test results online.	Adults who viewed at least 1 laboratory test result online in the past 12 months (n=1546; median age 58 y).		P	S[b]
Foster and Krasowski [44], United States	Retrospective study: analyzed patient portal activation and diagnostic test result viewing patterns among emergency department patients.	Emergency department patients receiving laboratory or radiology tests at a university medical center (n=25,361; mean age not reported).	P		
Fraccaro et al [33], United Kingdom	Within-subjects controlled experiment: tested whether alternative patient portal display formats influence risk interpretation of laboratory test results.	Kidney transplant recipients with chronic kidney disease (n=20; mean age 52 y).		P	
Giardina et al [36], United States	Qualitative study: explored patient experiences and reactions to receiving abnormal laboratory test results through a patient portal.	Adults who accessed a patient portal and received an abnormal laboratory test result for themselves or as caregivers (n=13; age range 30 - 80 y).	P		
Giardina et al [35], United States	Mixed-methods study: examined patients' understanding, emotional responses, and follow-up actions after receiving test results via patient portals.	Adult outpatients who viewed test results through a patient portal (n=95; mean age 54.6 y).	P	P	
Hulter et al [48], Netherlands	Mixed-methods sequential explanatory study: examined patient preferences for the timing of online release of laboratory, radiology, and pathology test results through a hospital patient portal.	Hospital patient portal users with access to delayed test result release options (survey n=4592; follow-up interviews n=7; mean age not reported).	P	S	
Hulter et al [50], Netherlands	Qualitative interview study: explored patient experiences and perceived advantages and disadvantages of real-time access to test results via hospital patient portals.	Outpatients who used or did not use portals for real-time test result access (n=28; age range 16 - 75 y).	P		S

XSL•FO
RenderX

Authors and country	Study design and purpose	Sample characteristics	Access	Interpretation	Activation
Joseph et al [54], Canada	Qualitative interview study: explored how frequent users of laboratory portals interpret test results when complementary health information is presented within the same display.	Frequent users of online laboratory result portals (n=24; age range 24 - 64 y).		P	
Krasowski et al [43], United States	Retrospective observational study: examined patient access to diagnostic test results through a patient portal across laboratory, pathology, and radiology settings.	Patients with diagnostic tests across outpatient, inpatient, and emergency settings over 6 months (n=59,388; mean age not reported).	P		
Lustria et al [25], United States	Cross-sectional survey study: examined patient use of patient portals to view laboratory test results and factors associated with test result comprehension.	Adults with one or more chronic conditions who used patient portals to view laboratory results (n=276; mean age 50.7 y).		P	
Mak et al [30], Canada	Cross-sectional survey study: compared turnaround time perceptions, comprehension, and anxiety between patients who accessed laboratory test results online and those who received results through traditional methods.	Adults who accessed laboratory test results via a web-based portal (n=2047) and matched non–portal users (n=1245; only age groups reported).	S	P	
McFarland et al [56], United States	Retrospective observational study: compared patient engagement with laboratory and radiology test results through an online patient portal.	Patients at an academic medical center with available electronic medical record data (n=424,422; median age 49 y).	P		
Monkman et al [49], Canada	Qualitative interview study: explored why patients use online laboratory result portals and what information they seek when reviewing their test results.	Adults with at least 2 years of experience using online laboratory result portals (n=25; age range 18 - 74 y).	P		
Monkman et al [51], Canada	Qualitative interview study: examined how individuals without medical training interpret abnormal laboratory test results in mock online portals and identify potential errors and their implications.	Adults with experience accessing laboratory test results online (n=25; age range 18 - 74 y).		P	

XSL•FO
RenderX

Authors and country	Study design and purpose	Sample characteristics	Access	Interpretation	Activation
Monkman et al [42], Canada	Web-based survey: examined barriers and facilitators to using online laboratory result portals, including usability, clarity, and preferences for display features.	Adults with experience using laboratory result portals (n=30; age group ≥45 y reported).	P	P	
Monkman et al [52], Canada	Mixed-methods study: examined health consumers' preferences for 4 online laboratory result display formats (tabular, annotated, visual, trends + contextual information)	Adults with experience accessing laboratory test results online (n=24; age range 19 - 74 y).		P	
Monkman et al [53], Canada	Qualitative study: examined how adults interpret a longitudinal potassium laboratory results graph modeled on a community laboratory portal, focusing on sources of confusion, misinterpretation, and intended follow-up actions.	Adults with experience accessing laboratory test results online (n=24; age range 19 - 74 y).		P	S
Morrow et al [32], United States	Randomized experimental study: tested how different patient portal message formats (standard, verbal, graphical, and video-enhanced) influence older adults' understanding of numeric test results, including memory for values, interpretation of risk, and intentions to engage in self-care behaviors.	Older adults viewing cholesterol and diabetes test results (n=144; mean age 71.9 y).		P	S
Nystrom et al [29], United States	User-centered design study: developed and evaluated a patient-facing prototype for reviewing laboratory test results through a patient portal.	Adult patients (n=14; mean age 43 y; age range 25 - 73 y).		P	S
Pillemer et al [39], United States	Mixed-methods study: examined how direct release of laboratory test results through patient portals influences patient engagement and health care usage.	Patient portal users at an integrated health system during 1 year (n=6368 survey respondents; mean age 51.8 y).	P		

XSL•FO
RenderX

Authors and country	Study design and purpose	Sample characteristics	Access	Interpretation	Activation
Robinson et al [31], Canada	Qualitative interview study: explored why patients use patient portals to access laboratory test results and how this access influences comprehension, engagement, and care behaviors.	Adult primary care patients who actively used a portal's laboratory results feature (n=21; age range 60 - 69 y).	P	S	
Scherer et al [26], United States	Web-based experimental study: tested whether substituting clinically appropriate goal ranges for standard reference ranges improves understanding of glycated hemoglobin test results.	US adults with and without type 2 diabetes recruited via a national web-based panel (n=6766; mean age 49.1 y).		P	
Schultz and Alderfer [37], United States	Qualitative study: explored caregivers' preferences and experiences related to receiving pediatric cancer test results through online patient portals.	Caregivers of children with cancer at a pediatric hospital (n=19; average age 40.4 y).	P		
Solomon et al [27], United States	User-centered design study: developed and iteratively tested prototype visualizations to improve patient comprehension and interpretation of laboratory test results in patient portals.	Adults with diabetes or family members of individuals with diabetes (n=18; median age range 50 - 59 y).		P	
Steitz et al [47], United States	Large multisite survey study: examined patient and caregiver perceptions of immediate release of test results through online patient portals, including perceived anxiety, engagement, and preferences.	Patients and caregivers who accessed test results through a patient portal (n=8139; median age 64 y).	P		
Steitz et al [12], United States	Retrospective observational study: examined patient portal use while awaiting laboratory test results, focusing on repeated access behavior and patient-initiated messaging.	Adult patients with outpatient laboratory tests and active patient portal accounts at an academic medical center (portal logs for 3,29,317 users; age not reported).	P		
Steitz et al [46], United States	Interrupted time series design: evaluated whether releasing laboratory test results in a patient-friendly educational format through a patient portal was associated with a reduction in patient-initiated messaging.	Adult outpatients who received laboratory test results via an Epic MyChart (Epic Systems Corporation) portal at a single academic medical center (n=205,139; mean age 51.0 y; 2024).	P		

XSL•FO
RenderX

Authors and country	Study design and purpose	Sample characteristics	Access	Interpretation	Activation
Struikman et al [38], Netherlands	Web-based experimental study: tested whether enhanced patient portal designs for viewing blood test results influence patient health engagement.	Adults recruited from a national consumer panel who viewed fictional blood test results (n=487; mean age 52.8 y).			P
Talboom-Kamp et al [10], Netherlands	Cross-sectional survey study: examined patient experiences and self-efficacy after viewing laboratory test results through an online patient portal with explanatory text and visuals.	Patient portal users who viewed laboratory test results and completed a web-based survey immediately afterward (n=354; mean age 58.5 y).			P
Tossaint-Schoenmakers et al [9], Netherlands	Cross-sectional survey study: examined how patient characteristics influence attitudes toward an online patient portal for communicating laboratory test results.	Adults who accessed blood test results through a web-based patient portal (n=748; mean age 58.5 y).			P
Turer et al [45], United States	Retrospective observational study: examined real-time patient portal use for viewing diagnostic test results during emergency department visits.	Adult emergency department encounters with at least 1 diagnostic test result (60,314 encounters representing 31,164 unique patients; age not reported).	P		
Wood et al [40], United States	Retrospective prepost observational study: examined changes in the timing of patient viewing of diagnostic test results following implementation of immediate release policies.	Patients with laboratory or imaging test results at an academic medical center before and after policy implementation (204,605 patients; mean age 44.3 y).	P		
Zhang et al [13], United States	Sequential explanatory mixed-methods study: explored patient challenges, needs, and emotional responses when interpreting laboratory test results through patient portals.	Adults who had previously accessed laboratory test results through a patient portal (survey n=203; interview n=13; only age groups reported).		P	
Zhang et al [41], United States	User-centered, mixed-methods design study: developed and evaluated a patient-facing system prototype to improve communication of laboratory test results through patient portals.	Adults who recently accessed laboratory test results through a patient portal (survey n=203; interview n=13; mean age 43.5 y).		P	

XSL•FO
RenderX

Authors and country	Study design and purpose	Sample characteristics	Access	Interpretation	Activation
Zhong et al [11], United States	Retrospective observational study: examined associations between patient portal use patterns and subsequent health care usage and appointment adherence.	Adult primary care patients with and without portal use over a multiyear period (matched samples n=4312 users and n=4024 nonusers; only age groups reported).	P		
Zikmund-Fisher et al [28], United States	Web-based experimental study: examined patients' ability to identify out-of-range laboratory test values and perceived urgency when viewing standard tabular test results.	US adults (n=1817; mean age 54.2 y).		P	
Zikmund-Fisher et al [14], United States	Web-based experimental study: tested whether graphical displays help patients distinguish between urgent and nonurgent deviations in laboratory test results.	US adults (n=1620; mean age 48.9 y).		P	
Zikmund-Fisher et al [34], United States	Web-based experimental study: examined whether adding harm anchors to visual displays of laboratory test results reduces perceived urgency and worry for near-normal values.	US adults (n=1618; mean age 48.8 y).		P	

[a]P: primary analytical focus of the study within the patient experience continuum.

[b]S: secondary analytical focus.

Main Outcome Variables and Measurement

Comprehension

Measurement strategies varied substantially across studies. Comprehension was assessed both objectively and subjectively. Overall, about one-third of studies (n=12) used objective comprehension or behavior metrics, while the remainder relied primarily on self-report. Objective measures included value-interpretation tasks assessing accuracy in identifying whether results were within or outside reference or goal ranges [25-29]. Some studies asked participants to self-rate their understanding and confidence in interpreting laboratory results [9,13,25,30] or to evaluate comprehension of medical terms and reference ranges [31]. Further, 1 study measured verbatim memory for exact numerical values and gist memory for the qualitative meaning or risk implications of results to examine how older adults understand and respond to portal-based test results [32].

Risk Interpretation and Perceived Urgency

Risk interpretation captures the cognitive evaluation of the seriousness or health implications of a laboratory result, whereas perceived urgency reflects the perceived need for timely action in response to that information. Both were not measured consistently in the same way, and several studies assessed them jointly rather than as separate outcomes. Risk interpretation was typically measured through risk-appraisal tasks asking participants to judge the clinical seriousness of an abnormal laboratory value—whether it required follow-up or no action based on perceived severity [33]. Additionally, 1 study used broader risk-perception items adapted from Garcia-Retamero and Cokely [64], focusing on perceived likelihood of complications [32]. Perceived urgency focused on how quickly participants believed they should respond after viewing results. It was typically assessed using Likert-type items evaluating urgency of taking action or contacting a clinician or by asking participants about intended timing of response, such as whether they would contact a clinician immediately, wait, or take no action) [14,28,34].

Emotional Responses

Emotional responses to viewing laboratory results online were typically elicited through surveys or semistructured interviews. Only a few studies used validated affective scales [30,35]. Mák et al [30] assessed anxiety using the validated Global Anxiety–Visual Analog Scale [65], which measures emotions

XSL•FO
RenderX

from “no anxiety” to “worst imaginable anxiety” on a 100 mm line.

Patient Activation and Health Engagement

Patient activation was assessed in several ways, mostly through self-report [35-37]. Struikman et al [38] used the Patient Health Engagement, a 5-item scale, measuring patients’ psychological readiness and involvement in managing their health across 4 progressive phases—capturing emotional, cognitive, and behavioral dimensions of health engagement. Further, 2 studies [9,10] used the 10-item motivation and confidence to act subscale of the eHealth Impact Questionnaire [66] to determine the extent to which eHealth technology influences patients’ confidence in taking health-related actions. Giardina et al [35] used the Patient Activation Measure [67]—a 10-item validated scale used to assess participants’ engagement and capacity for health self-management.

Engagement With Laboratory Test Features in Patient Portals

Retrospective studies captured objective measures of portal engagement (eg, frequency of portal logins and laboratory test viewing), capturing both initial access and downstream follow-up behaviors [8,39,40]. User log analyses also captured behaviors such as patient-initiated provider contacts, follow-up appointments, and no-shows [11,12,37,39].

Portal Acceptability, Usability, and Satisfaction

Design-focused studies captured outcomes related to usability, user satisfaction, and visual engagement with laboratory result interfaces. Usability was commonly assessed using the system usability scale [29,41,68]. The net promoter score [69], a single-item measure of patients’ likelihood to recommend the portal experience, was used in 1 study to measure satisfaction [42]. A few studies used the 13-item information and presentation subscale of the eHealth Impact Questionnaire to evaluate users’ perceptions of the clarity, usefulness, and trustworthiness of health information [9,10]. Furthermore, 1 study used eye tracking to measure fixation count, average fixation duration, and dwell time to capture how participants processed laboratory result displays [33].

Results of Thematic Analysis

The following section organizes the results of the thematic synthesis along a 3-stage continuum—access, interpretation, and activation—synthesizing evidence on facilitators and barriers to portal use, timing and immediate release practices, comprehension, risk appraisal, emotional responses, and subsequent engagement behaviors.

Access

Access refers to the structural and contextual conditions under which laboratory test results are made available through patient portals.

Access Before and After Policy Changes

Several large-scale retrospective studies compared portal access before and after changes in laboratory result release policies [1]. Krasowski et al [43] analyzed portal log data from 59,388 patients in 2016 and found that only 30% of outpatient laboratory results were viewed through the patient portal. A follow-up emergency department (ED)–focused analysis conducted 3 years later showed only modest change, with 8.9% of ED-ordered laboratory results viewed through the patient portal [44]. Among those that were accessed, 53.3% were viewed within 72 hours of release, while nearly 20% were not viewed until 2 weeks or more after becoming available.

Following policy changes, multiple studies documented significant increases in portal-based laboratory result viewing patterns, particularly in outpatient settings. Turer et al [45] observed an increase in ED result viewing from 5% to 15% between April 2021 and April 2022, even though the portal was not actively promoted. Wood et al [40] analyzed portal log data from over 204,000 users, comparing two 10-month periods (February to December 2020 vs February to December 2021). They found that the proportion of test results viewed within 1 day rose from 22% to 31% among adult users and from 15% to 33% among pediatric proxy users following the policy change. Similarly, Bhalla et al [8] analyzed log data of 44,000 oncology patients and observed a drop in median viewing time from 77.0 hours to 6.4 hours after the policy was implemented. By 2022, a total of 75% of laboratory results were viewed by patients before their providers. Steitz et al [46] corroborated this pattern in a 2024 sample of over 829,000 outpatient results, finding that 69.4% were accessed by patients before their ordering clinician.

Default settings and notification features were associated with higher rates of laboratory result viewing. A retrospective analysis of portal log data showed increased real-time test result viewing among ED patients following changes in default release settings that enabled immediate push notifications [45].

System-Logged Patient Behaviors Following Immediate Result Release

Portal log analyses provide additional insight into how patients interact with results once they become available. Pillemer et al [39] found that patients who viewed automatically released laboratory results had higher rates of office visits and telephone calls than those who received manually released results. Steitz et al [12] analyzed portal log and secure messaging data from over 1.2 million outpatient test encounters. They found that patients who frequently refreshed pages and received routine test results were more likely to send a secure message to their provider soon after viewing the results than those who received high-sensitivity test results. Steitz et al [46] found that 17.4% of reviewed outpatient results prompted a patient-initiated message within 24 hours of release. Patient-initiated messaging was not meaningfully reduced when supplementary patient-friendly educational materials were provided.

Zhong et al [11] tracked engagement among 4312 portal users at a university hospital and found that patients who actively used the laboratory feature had higher rates of office visits and calls to their providers and had lower no-show rates (30% - 60%) compared to matched nonusers.

XSL•FO
RenderX

Patient-Reported Preferences, Benefits, and Barriers Related to Portal-Based Laboratory Result Viewing

Preferences for immediate versus delayed access varied across populations and clinical contexts. In a national survey of US portal users, most preferred to receive results as soon as they became available—even before their provider had reviewed them [47]. In a retrospective study, most Dutch portal users set preferences for a 1-day delay for laboratory test result release [48]. Follow-up interviews revealed that some users later changed their preferences to shorter delays, citing the advantages of greater transparency and time to prepare for upcoming appointments. Those who had opted for longer delays said they changed their preference settings because they found the information hard to understand or lacking in sufficient detail. Timing preferences also differed based on result sensitivity. While many patients supported immediate access to abnormal results, some preferred that serious or highly concerning results be communicated verbally before appearing in the portal [36]. Schultz and Alderfer [37] found that caregivers of pediatric cancer patients preferred to be informed about abnormal test results directly by their child's doctor and viewed electronic delivery as mainly supplementary.

In Monkman et al [49], participants found online access reassuring and helpful for monitoring their chronic diseases, medications, and treatments. Some noted that online access improves patient safety by facilitating early detection of abnormalities and prompt communication with their doctors. In 2 studies, caregivers and patients acknowledged that portal access reduces the burden on clinic staff and frees up clinical resources by reducing unnecessary appointments and allowing physicians to prioritize more urgent cases [31,37].

Despite these reported advantages, studies also identified barriers to using portals for laboratory result viewing. Common challenges included difficulty locating test results within portals, confusion when multiple systems were in place, and uncertainty about which platforms to use. In a survey of adults with chronic conditions, many participants who had viewed laboratory results online still preferred to discuss results directly with their clinicians or only wanted to be notified if their results were abnormal [25]. Giardina et al [36] interviewed 13 US adults—patients reported that results often lacked adequate explanation and contextual cues, reducing the perceived usefulness of immediate access.

Emotional Responses to Immediate and Unmediated Release

Immediate portal access to laboratory results can evoke a range of emotional responses. In Giardina et al [36], patients who received abnormal results were significantly more likely to report concern, confusion, or anxiety than those receiving normal results. Some participants described worry even when results were within range, particularly when no clinician interpretation was provided. Pillemer et al [39] used a mixed-methods design combining survey data, interviews, and EHR usage logs from 6368 patient portal users. Some participants were worried when abnormal results were released automatically without physician explanations. One interviewee thought she had cancer after seeing an abnormal test result on a weekend and was only reassured when she was able to talk to her doctor the next business day.

Emotional reactions were not uniformly negative. In a large cross-sectional survey, Steitz et al [47] found that many participants indicated that viewing results directly reduced worry compared to waiting for clinician notification. For some patients, immediate access alleviated the stress associated with delayed communication. Qualitative findings from Hulter et al [48] and Hulter et al [50] further illustrate this tension. While some participants felt overwhelmed when encountering unexpected or sensitive information online, others valued early access because it shortened the waiting period and provided a sense of control before clinician discussion.

Disparities in Portal Access and Laboratory Result Viewing

Studies consistently documented disparities in those who enrolled in patient portals and those who used them to view laboratory test results, with differences falling along the lines of race, ethnicity, insurance type, language preference, gender, and care setting. Zhong et al [11] reported enrollment disparities in a hospital-based sample. Nonusers were more often Black, Hispanic, male, unmarried, or publicly insured. Portal adopters were more often older than an age of 30 years and had more telephone encounters with providers. Missed appointments were more common among nonusers.

Portal log analyses documented parallel disparities in viewing behavior. Bhalla et al [8] reported that only 33% of Black patients viewed their laboratory results after they were released, compared to 51% of White and 58% of Asian patients. Wood et al [40] found that White, English-speaking, privately insured patients—particularly those seen in outpatient settings—were more likely to view results within 1 day of release. Viewing rates were also lower among older adults, non-White and non-English-speaking patients, and those with public or no insurance.

Krasowski et al [43] and Foster and Krasowski [44] reported disparities in viewing behavior across care settings. Outpatient laboratory results were accessed more frequently (around 30%) than those receiving inpatient or emergency care (under 10%). Access was more limited for parents of adolescent patients, reflecting proxy access restrictions in some settings.

Lustria et al [25] identified predictors of laboratory result access. White race and higher eHealth literacy were positively associated with viewing results through portals, while having a college degree was associated with lower odds of access after controlling for other factors.

Recommended Design Features to Support Access

Participants across studies recommended design improvements to make laboratory result portals easier to navigate and more inclusive. Clean visual structure and simpler layouts were recurring themes. In 1 study, participants reported difficulty scanning results within long tables and recommended persistent column headers and visually highlighting abnormal values to improve readability [51].

Formatting and usability issues related to mobile access were noted across several studies [27,41,42]. Participants in Solomon

XSL•FO
RenderX

et al [27] indicated the need for larger fonts and minimal visual clutter, particularly for mobile viewing. In another study, participants noted that existing portal designs do not translate well to small screens and recommended mobile apps with more intuitive navigation [42].

Accessibility concerns were especially salient for older adults and individuals with visual or cognitive impairments. Suggested features included screen-reading options, streamlined navigation, and assistive buttons operable by keyboard or touch. One participant in the study by Zhang et al [13] emphasized the importance of designing the portal so that "grandma can understand it." Another user in Robinson et al [31] suggested including a short onboarding tutorial to help less technology-savvy patients understand how to navigate the portal and locate test results.

Interpretation

This stage reflects the cognitive and affective work required to translate numerical data into meaningful information.

Identifying and Interpreting Abnormal Results

A central task in interpretation involves determining whether a test value falls outside the reference range and understanding what that deviation signifies. The absence of explanatory context was the most consistently reported barrier to understanding. In a mixed-methods study, Giardina et al [35] reported that although many interviewees said they understood the information, several struggled to interpret specific values. One participant, for example, was unsure whether the term "positive" indicated a favorable or unfavorable result. Lustria et al [25] assessed comprehension of hypothetical laboratory panels in a survey of 276 US adults. While most participants scored well on objective tests, they rated their own understanding of the values as low and preferred physician explanation—particularly when values fell outside the normal range.

Across multiple studies, participants reported difficulty determining whether values were normal or abnormal when results appeared in dense tables without plain-language explanations or clear visual indicators [13,29,42,49,51,52]. In an experiment with 1817 US adults, Zikmund-Fisher et al [28] found that only 51% correctly identified a moderately elevated hemoglobin A_{1c} (HbA_{1c}) value (8.4%) as outside the reference range when presented in a standard table format. Many participants also struggled to rate glycemic control accurately or determine whether they would contact their physician. Monkman et al [51] asked experienced portal users to review a mock laboratory result on an existing portal display; even among these active users, many failed to recognize abnormal results when values were not clearly flagged or contextualized. In iterative design workshops, Nystrom et al [29] observed confusion around HDL (high-density lipoprotein) cholesterol values, with many participants assuming that higher HDL levels were unfavorable—reflecting a broader expectation that higher numbers are undesirable, even though higher HDL is typically associated with lower cardiovascular risk.

Experimental studies further demonstrated that the format in which laboratory results are displayed can substantially influence the accuracy of interpretation. In a randomized study with older adults, Morrow et al [32] found that verbally and video-enhanced displays improved participants' ability to interpret the meaning of results compared to graphical formats, although recall of exact numerical values did not improve. Extending this to trend displays, Monkman et al [53] found that 18 of 24 participants expressed confusion about at least 1 graph feature, and only 2 said they would seek urgent follow-up for an out-of-range value.

Not all visualization styles improved comprehension. Displays incorporating multiple reference ranges, dense formatting, or layered graphical elements often increase cognitive burden and reduce interpretive accuracy. Scherer et al [26] found that comprehension was highest when visualized test results emphasized a single, clearly defined goal range. Simpler visuals that directly stated what a result meant and whether action was needed were rated more favorably, whereas layered displays with multiple reference zones and graphical indicators reduced comprehension. In user-centered evaluations of mock portal reports, participants described lengthy layouts and technical language as overwhelming, particularly when abnormal values were not clearly emphasized [51]. Similarly, Joseph et al [54] found that supplementing trend graphs with educational content was sometimes perceived as confusing or misaligned with the test result itself.

Appraising Risk and Urgency

Risk interpretation and perceived urgency represent related but distinct dimensions of interpretation. As many studies assessed these constructs using overlapping measures—asking participants to judge either the seriousness of a result or how quickly action was needed—the findings are reported together here.

Display formats influenced how patients judged the seriousness of results and the perceived need for immediate follow-up. Visual formats generally reduced overestimation of urgency for near-normal results compared with standard tables. Zikmund-Fisher et al [14] found that participants viewing standard tabular displays were more likely to judge mildly abnormal alanine transaminase and creatinine values as urgent than those viewing visual formats. Visual formats improved differentiation between mildly and clearly abnormal values but did not substantially alter perceptions of extreme results. Scherer et al [26] found that showing only goal ranges—rather than both goal and standard reference ranges—was associated with more accurate judgments of urgency. Zikmund-Fisher et al [34] found that "harm-anchor" displays reduced perceived urgency for near-normal results and decreased intentions to contact a clinician or seek hospital care for those values, without affecting responses to more extreme abnormalities.

Visualizations were not uniformly effective in reducing misclassification of risk severity. Fraccaro et al [33] tested 3 result formats (baseline, contextualized horizontal bars, or grouped) with kidney transplant recipients across low-, medium-, and high-risk scenarios. Participants had the most difficulty identifying medium-risk scenarios and frequently selected follow-up actions that did not align with recommended care. Although eye-tracking data showed longer viewing time for contextualized formats, none of the 3 displays significantly

XSL•FO
RenderX

improved interpretation accuracy [33]. Morrow et al [32] similarly found that participants in the graphically enhanced condition rated lower-risk results as more serious than those in the verbally enhanced, video-enhanced, and standard conditions, with a similar pattern observed for borderline-risk results. In a think-aloud study of a portal-style potassium graph, Monkman et al [53] reported that only 2 of 24 participants said they would seek urgent follow-up for an out-of-range value, while most preferred to look up information, follow up later, or were not in a hurry to contact a provider.

Individual Differences in Interpretation

Risk appraisal and comprehension varied based on different characteristics. Zikmund-Fisher et al [28] found that participants with higher numeracy were more likely to correctly identify significantly abnormal HbA_{1c} values and reported stronger intentions to contact their physician. Higher health literacy was associated with a lower likelihood of follow-up for mildly elevated values. Zikmund-Fisher et al [14] reported that participants with higher health and graphical literacy were better able to distinguish between mildly and highly abnormal results, particularly when results were presented using visual displays. Zhang et al [13] reported that more participants with higher health literacy and technology proficiency said the portal helped them understand their results and feel supported. Lustria et al [25] found that older adults and frequent portal users were significantly more likely to score higher on objective comprehension tests, although many self-reported difficulty understanding their results and preferred to have doctors explain them.

Affective Responses During Result Interpretation

The action of interpreting test results presented online can evoke a range of negative and positive emotional reactions. The evidence reported here includes both studies of patient reactions to actual test results and studies using hypothetical or simulated scenarios.

Negative emotions such as anxiety, worry, and confusion were commonly reported when results were abnormal, unexpected, or presented without a clear explanation. Concern and distress were strongest when terminology was unfamiliar or when patients were uncertain whether a clinician had reviewed the result [13,31,35,36,39]. In Giardina et al [35], several participants reported anxiety or confusion even with normal results—particularly when no interpretation or guidance was included. Similarly, in a survey comparing patient portal users with nonusers, Mák et al [30] found that users who were unclear about the need for follow-up were more likely to report higher anxiety.

Participants in the study by Robinson et al [31] described physician comments accompanying laboratory results as clarifying and reassuring, particularly when comments confirmed that the result had been reviewed and did not require further action. Although uncertainty frequently elicited anxiety, reassurance was also reported when results were clearly contextualized. In 1 study, among participants with chronic conditions, those who accessed results online were significantly less likely to report high anxiety [30].

Presentation Features That Shape Emotional Responses

In experimental and user-centered studies using hypothetical or simulated test results, presentation styles—including language and terminology, visual cues, and other enhancements—also evoked different emotions. Zhang et al [41] found that emotionally charged phrasing such as “not optimal” was perceived as alarming. Participants preferred neutral, explanatory language placed directly alongside test results. Scherer et al [26] examined emotional responses to hypothetical HbA_{1c} results presented in 3 formats: as a table, a 2-color number line, or a block-style number line. They found that block-style displays using color-coded diagnostic categories (eg, red for “high” or “abnormal”) were significantly rated as more discouraging than simpler number lines or tables. Similarly, Solomon et al [27] found that red flags and warning icons elicited mixed reactions—some found them helpful; others described them as anxiety-inducing.

Design Recommendations to Support Interpretation

To make portal-released laboratory results easier to understand, more trustworthy, and more reassuring, patients frequently emphasized the need for plain-language explanations to help interpret unfamiliar or abnormal results. They suggested providing summaries that clearly state whether a value was within a healthy range and what it meant. One participant in the study by Pillemer et al [39] said it would have helped to have “a note that said normal/abnormal or something like that.” In Monkman et al [42], users similarly asked for short, condition-specific explanations that clarified the purpose of each test and its relevance to their treatment.

Patients also expressed interest in receiving tailored and contextualized explanations, personalized resource links, and tailored prompts [41]. In 1 study, patients wanted explanations tailored based on users’ literacy levels and options to receive either basic summaries or advanced content. Similarly, Monkman et al [42] found that embedded explanations tailored to the patient’s medical context could help clarify the relevance and meaning of specific laboratory values. Monkman et al [52] found that many participants wanted to combine color-coded summaries with trend graphs and tabular details and preferred being able to switch between visual and tabular formats, indicating a desire to tailor not only explanations but also the format of laboratory result displays to their needs. In Zhang et al [13], some participants proposed integrating artificial intelligence to provide contextualized explanations based on individual health data.

To increase confidence in laboratory results, participants in 1 study recommended cues indicating provider oversight—such as checkmarks signaling that their doctor had reviewed their results [50] or some kind of indication from their doctor about the need for follow-up [31]. Some participants said that labels such as “your doctor is pleased,” and simple text overlays such as “you are here” or “your doctor wants you here” could provide reassurance and guide interpretation, especially when paired with graphical displays [27,31]. Lastly, patients often expressed the need for vetted, high-quality health information alongside laboratory results to help them interpret their results [41,54].

XSL•FO
RenderX

Activation

The activation stage refers broadly to patients' readiness, confidence, and actions taken in response to viewing laboratory results online. Across included studies, related constructs such as engagement, motivation, behavioral intention, and self-efficacy were used inconsistently to capture this domain.

Online Access and Behavioral Intentions

Patients commonly described reflecting on their health, monitoring conditions, and making small behavior changes in response to viewing laboratory results online [29,31,49,50]. Hulter et al [50] reported that online access helped Dutch patients prepare for appointments, identify errors, and engage more actively in health decisions. Monkman et al [49] found that participants routinely scanned for abnormal values and tracked trends over time, describing laboratory result access as a tool for self-care and personal accountability. In Nystrom et al [29], 64% of patients said viewing their results motivated them to change their diet or exercise, and over a third intended to follow up with their provider. Portal users interviewed by Robinson et al [31] also reported that online access improved their ability and motivation to monitor their health more closely, adhere to their medications, and make lifestyle changes.

Information-Seeking and Follow-Up Actions

Across studies, patients frequently reported searching for additional information, preparing for medical visits, or contacting providers after receiving results online [13,35,39,55]. Christensen and Sue [55] found that patients engaged in a range of follow-up information-seeking behaviors after viewing their results, including talking with others (21%), searching for information in the portal (20%), creating graphs (19%), and looking up information online (18%). These actions were more common when providers had not clearly communicated how results would be shared. Pillemer et al [39] found that patients often used early access to look up related terms, track chronic conditions, or prepare questions—sometimes reaching out even when results were not clinically significant.

Factors Influencing Motivation to Act

Several experiments using hypothetical or simulated scenarios found that laboratory result presentation formats influenced patients' intentions to act or engage in their care [14,32,34,38]. In the studies by Zikmund-Fisher [14,34], patients shown mildly or near-normal abnormal results in visual formats were more likely to report they would delay follow-up or take no immediate action compared to those who viewed tabular displays. Morrow et al [32] observed that graphically enhanced presentations were associated with stronger intentions to adhere to prescribed medications, particularly for lower-risk results, although visual format had less consistent effects on intentions to change diet or exercise.

Struikman et al [38] assessed how different laboratory result display formats influenced patients' emotional, cognitive, and motivational readiness to engage in their own care using a shortened version of the validated Patient Health Engagement scale. Engagement scores declined significantly after viewing abnormal results without context. When visuals and explanations were included, engagement remained stable for normal and partially abnormal results and declined only slightly for abnormal ones.

Talboom-Kamp et al [10] found a strong positive correlation between perceived portal usability—particularly clarity, ease of understanding, and trust—and self-efficacy. Similarly, participants who rated the portal's information and presentation more favorably reported higher motivation and confidence to act on laboratory results (r=0.77, P<.001).

Desired Support for Follow-Up and Engagement

Patients often advocated for features that would help them take informed next steps. Many expressed the need for automated alerts when new results became available. In Monkman et al [42], participants recommended adding follow-up support features, such as automated notifications when results were available and links to additional resources or options to message providers. Others stressed that these alerts should do more than announce availability; participants in Giardina et al [35] suggested pairing notifications with short explanations and recommendations about what to do next.

Some patients recommended tools to support follow-up visits with their providers. Participants in Zhang et al [41] recommended adding an "ask questions" button that would help clarify test results and help them prepare personalized questions for follow-up visits. This feature would allow them to save questions directly within the portal and generate question lists for clinical visits. This, they felt, would help structure conversations and remind them about key concerns they need to discuss. In another study, patients also suggested portals should include brief follow-up instructions, lifestyle suggestions, and access to trusted health sources [13].

Discussion

Principal Findings

This review synthesized evidence from 39 studies on how patients engage with and respond to online access to laboratory test results. Across a diverse body of evidence, patients' engagement is characterized by increased access, but also by substantial variability in how results are understood and acted upon. Studies consistently show that access does not reliably translate into accurate interpretation or appropriate action, particularly when results are released without accompanying clinical interpretation. Patients frequently struggle to determine the clinical significance of results, especially when values are borderline, unexpected, or lack a clear contextual explanation. These interpretive challenges evoke emotional responses such as anxiety, reassurance, or uncertainty, which in turn shape subsequent actions, including seeking additional information, contacting providers, or taking no action. Overall, access alone does not ensure meaningful engagement; patients' ability to comprehend and respond to their results varies and is shaped by both the characteristics of the results and systematic differences in patient capacity and access to support.

Prior reviews have typically examined patient portal use, comprehension, or presentation strategies in isolation, often focusing on a single outcome domain, mixing patient and clinician perspectives, or aggregating different types of test

XSL•FO
RenderX

results such as laboratory, imaging, and pathology findings [18,19,24]. In contrast, this synthesis brings together complementary forms of data from qualitative studies, large-scale analyses of real-world portal use, survey research, controlled experiments, and mixed-methods designs to examine how patients engage with laboratory results across the full trajectory from access to interpretation to action. Log data document patterns of real-world engagement, survey research identifies the patient characteristics and contextual factors associated with comprehension and follow-up, qualitative inquiry explains the interpretive and emotional processes underlying these patterns, and experimental studies demonstrate how design features shape understanding and decision-making. This integrated perspective provides a more complete account of patient engagement, highlighting how challenges emerge and interact across stages rather than occurring as isolated barriers.

The following sections interpret these findings across the 3 stages, examining how informational, emotional, and workflow factors interact to shape patient engagement at each point along the continuum.

Access: Transparency, Immediacy, and Workflow Alignment

Patients often described valuing having their laboratory results available online and finding it more convenient than waiting for phone calls or scheduled visits, particularly for routine or anticipated findings [36,47,49,55]. For many, immediate access signals transparency and supports a sense of control over their care [47,48,50]. At the same time, patients expected that clinicians would still review results and provide reassurance, especially when findings are abnormal, unexpected, or difficult to interpret [25,36,37]. Access, in other words, is valued not as a replacement for clinician communication but as an earlier entry point into that relationship.

Our synthesis suggests that this dual expectation exposes a structural misalignment between how access is implemented and how responsibility for follow-up is allocated and signaled. The results may be released automatically into portals while clinician review and communication occur asynchronously, leaving patients unsure whether anyone has assessed their results or will proactively reach out [12,36]. In this context, immediate access shifts interpretive and logistical work from clinicians and staff to patients, who must decide when to wait and when to initiate contact [11,12,39]. Requests for clearer interpretive cues and guidance about next steps [13,41], as well as for visible indicators of clinician review [12,36], can be understood as attempts to rebalance these responsibilities rather than as resistance to immediate access.

At a policy level, immediate-release regulations have broadened the possibility of access but have not produced equitable realization of that access across populations or care settings [8,40,45]. Across the included studies, viewing rates were lower among non-White and non–English-speaking patients, older adults, people with lower health or digital literacy, and those with public insurance [8,11,25,40]—groups who may have the most to gain from timely information yet face the greatest barriers to portal use. These patterns likely reflect structural constraints (eg, language support, connectivity, enrollment practices, or trust in institutions) more than a lack of interest in results, indicating that equity in the access stage depends as much on addressing these upstream conditions as on refining portal interfaces themselves.

Interpretation: Sensemaking Under Uncertainty

Overview

While portal-based delivery of laboratory results has expanded transparency, it also amplifies the interpretive work patients must do to determine what their numbers mean for their health.

Signaling Relevance and Proportionality

To make sense of a laboratory result, patients must judge whether deviations from standard ranges are meaningful, and that judgment depends heavily on how the deviation is signaled. In portal displays, values appear alongside reference ranges, color indicators, flags, and graphical elements that denote whether a test value falls outside expected limits. Experimental studies show that these formats shape perceived significance, but not always proportionately: some clarify trends and contextualize variation, whereas others amplify minor fluctuations or highlight borderline values in ways that overstate their clinical importance [14,26,34].

For example, a value only slightly outside a reference range may be interpreted as urgent when displayed in red, while more clinically meaningful patterns may be obscured by dense or layered interfaces [26,27]. Displays that are overly complex can increase cognitive load and reduce confidence in interpretation. By contrast, overly simplified cues can shift perceptions of seriousness and urgency in directions that do not always align with clinical meaning—especially when patients must interpret results on their own [32,33].

Moreover, these preferences and effects are not uniform. Some users prefer detailed, layered information, while others find it overwhelming. What enhances clarity for one group can increase confusion or anxiety for another [13,42,49,52].

Contextualization and Tailoring

Across qualitative and survey studies, patients report that they do not simply want to know whether a value falls within a reference range; they want to understand what the results mean relative to their health condition, treatment, and broader health trajectory [13,25,49]. In response, designers of patient-facing systems are often urged to “contextualize” and “tailor” laboratory results, but these terms are frequently used interchangeably. Although both are intended to help patients make sense of their results, contextualization and tailoring are implemented in different ways and serve related, but distinct functions within laboratory result communication.

Contextualization refers to providing information that clarifies the clinical meaning of a result, for example, by explaining reference ranges, using categorical risk labels, or adding brief explanatory text [26,28]. In contrast, tailoring involves designing messages for a specific individual based on assessed characteristics related to the outcome of interest (eg, beliefs, health literacy, numeracy, or motivation), with the goal of increasing perceived personal relevance and influencing how

XSL•FO
RenderX

information is interpreted and acted upon [70]. When used appropriately, tailoring has been shown to promote deeper processing and increase how seriously information is taken [71,72]. In patient-facing laboratory result systems, this is most often operationalized through computer tailoring (ie, algorithm-based or data-driven messaging), where individual data are used to generate tailored information or guidance via mechanisms such as personalization, feedback, and content adaptation [73,74].

These distinctions, however, are rarely made explicit. Many recommendations invoke "tailoring" without specifying what and how much individual details are beneficial, which elements should be adapted, or how tailoring should operate in clinically ambiguous situations. As systems, including artificial intelligence–driven tools, become capable of generating highly individualized explanations, there is a tendency to assume that greater specificity—whether through contextualizing information or tailored messaging—will improve understanding. Yet additional detail can introduce noise, elevate perceived importance, or shift interpretive weight without resolving uncertainty. When these strategies are implemented primarily through layering additional data or explanations, they risk introducing complexity and cognitive load without improving interpretive clarity or reliably supporting more meaningful engagement. The challenge, then, is not simply to provide more contextualization or more intensive tailoring, but to calibrate how contextual information and tailored messages are introduced so that they sharpen clinical meaning without overwhelming patients or unduly inflating the perceived importance of test results.

Emotion, Meaning Making, and Design

Laboratory results are often viewed in moments of uncertainty, and patients commonly feel worried, confused, or overwhelmed by raw numbers and jargon [13,31,35,39]. These emotions are heightened in the absence of contextual framing [8,30]. Interface elements intended to clarify meaning, such as color coding, flags, summary labels, and risk terms ("abnormal," "suboptimal," or "high risk"), influence whether a deviation feels urgent or routine; bold alerts and red blocks may heighten alarm for marginal findings, while muted or cluttered layouts can obscure patterns that warrant concern [14,26,27,34,41].

Despite extensive documentation of these emotional responses, the literature offers limited guidance on how systems should deliberately calibrate them. Human-computer interaction research provides a useful lens for this challenge. Emotional design involves deliberately structuring visual layout, language, and interaction cues to shape how users feel in moments of uncertainty [75,76]. Potential strategies include neutral summary statements that precede numerical details, restrained and well-explained use of color and flags, visible indicators that clinicians will review results within a specified timeframe, and concise descriptions of typical next steps. Such features may help assure patients that they are not alone in making sense of their results. Without this kind of emotional calibration, increased transparency may lead to confusion or distress rather than informed understanding.

Activation: Engagement Beyond Access

Policy narratives tend to equate transparency and immediate access with automatic gains in patient engagement. In practice, activation is inseparable from the interpretive and emotional work patients perform after viewing their results. Patients' decisions about whether to watch and wait, seek reassurance, or request urgent care are shaped by how clearly results are explained, how deviations from normal are framed numerically and visually, and how worried or reassured patients feel in the moment [14,32,34,38]. Experimental and user-centered studies show that visualization formats clarifying the magnitude and direction of deviation can reduce unnecessary worry or intentions to seek urgent care for near-normal values, yet they may also blunt vigilance for higher-risk results, including cases where clearly abnormal trends are interpreted as nonurgent [14,34,53]. In this way, online result systems are not neutral delivery channels—they may shape the kinds of downstream actions that patients see as appropriate.

At the same time, what appears in log data as "activation" may reflect gaps and misalignments in clinical workflows rather than purely patient-driven engagement. Across the included studies, actions such as viewing results repeatedly, sending secure messages, or scheduling appointments may reflect attempts to resolve residual uncertainty rather than confident self-management [13,35,39,55]. In this sense, "more action" is not inherently better; meaningful activation is better characterized as calibrated and proportionate follow-up that matches the clinical seriousness of the result and the patient's goals, and supports patients' preparedness for shared decision-making.

Additionally, when results are released before clinicians have reviewed or communicated them, patients often send multiple messages, call offices, or seek in-person visits to confirm that results have been seen and to clarify their meaning [35,39]. These compensatory behaviors suggest that activation is a property of the broader sociotechnical system: result displays, notification rules, and response pathways either converge to scaffold timely, collaborative follow-up or leave patients to improvise their own paths to clarification. Designing activation-supportive systems, therefore, requires coupling patient-facing interpretive aids with predictable, clearly signposted routes to clinician review and shared decision-making.

Finally, current operationalizations of activation tend to obscure important inequities. When activation is inferred primarily from portal behaviors or self-initiated contacts, it may be overestimated for already-advantaged populations and underestimated for patients whose capacity to act is constrained by language, literacy, or structural factors. The same groups that face barriers to portal access and comprehension (eg, non-White and non–English-speaking patients, older adults, people with lower health or digital literacy, and those with public insurance) are less likely to view results promptly and to feel confident about what to do next [8,11,25,40]. Similar patterns observed across the continuum suggest that this may have cascading effects—patients who face barriers to access have fewer opportunities to encounter and make sense of their results,

XSL•FO
RenderX

and those who struggle to interpret results are less equipped to take meaningful action.

Limitations

This review has several limitations. First, it focused on peer-reviewed, English-language studies and did not include gray literature such as preprints, dissertations, or non–peer-reviewed conference proceedings. This approach may have excluded early-stage innovations and practice-based insights from nonacademic or non–English-speaking settings, increasing the risk of publication bias.

The evidence base is also geographically concentrated. Few studies examined experiences in non-US health systems or among patients with limited connectivity or portal access. The findings underrepresent the needs of less digitally engaged groups, individuals with lower literacy or limited English proficiency, and patients who do not access portals. This limits insight into how expectations for result delivery, interpretation, and follow-up may differ across health systems and levels of digital access.

Limitations within the primary studies also influence the strength of the conclusions. Researchers defined and operationalized key constructs inconsistently and rarely used validated instruments, making it difficult to compare findings across studies or determine whether observed differences reflect true variation in patient responses or are artifacts of measurement. This also constrained conclusions about the impact of specific design features, as similar outcomes were often labeled or measured in incompatible ways. More broadly, this reflects a conceptual gap in the field: without shared definitions and validated measures, the literature cannot accumulate in a way that supports robust, generalizable conclusions. Progress will require the development and adoption of common, validated measures that can support meaningful comparison across studies, populations, and health system contexts.

Many studies relied on self-report measures, introducing recall and social desirability biases and limiting insight into how patients interpreted or acted on laboratory results in practice. Researchers also frequently used hypothetical test scenarios rather than real clinical encounters. Although these designs allowed controlled comparisons of presentation features, they did not fully capture how patients responded when results were personally relevant, emotionally salient, and embedded within ongoing care.

Most experimental studies used small, convenience samples, often drawn from populations already engaged with digital health tools. These samples provide limited insight into how findings extend to broader and more diverse patient populations, particularly those with lower digital literacy, limited access to technology, or different expectations for interacting with health information. Studies also rarely examined whether design features function differently across patients with varying health histories, goals, or prior knowledge. Without attention to this heterogeneity, it remains unclear under what conditions specific presentation strategies are effective or for whom they are most beneficial.

Finally, most studies assessed immediate or short-term responses to laboratory result presentations, while relatively few examined longer-term behavioral or clinical outcomes. This emphasis on proximal outcomes limits the ability to determine whether improvements in comprehension or reassurance translate into sustained changes in follow-up behavior, decision-making, or engagement in care over time.

Conclusions

Patients who access laboratory test results through online portals are not simply retrieving information; they are navigating a sequence of decisions—whether to attend to a result, how to interpret it, and what to do next—often in the absence of timely clinical guidance. The evidence synthesized here shows that expanding access has shifted engagement earlier in the care process but has not ensured that patients can effectively use the information. Instead, challenges compound across stages: results are encountered without sufficient context, interpretation remains uncertain, and expectations for follow-up are often unclear. As a result, patient responses vary widely and are not consistently aligned with clinical need. This review advances prior work by demonstrating that these gaps are not isolated issues of comprehension or interface design but reflect a broader misalignment between how results are delivered, when they are encountered, and the support available to patients as they move from access to action.

Improving patient engagement, therefore, requires changes across portal design, clinical workflows, and institutional policies, rather than isolated enhancements to any single component. In practice, this includes embedding interpretive context at the point of result delivery, clearly signaling clinician review and follow-up expectations, and providing accessible pathways for patients to seek clarification or act. These supports must be designed for diverse populations and real-world use at scale. Differences in health literacy, language proficiency, digital access, and experience with the health system shape not only whether patients view results but also how they interpret and respond to them. Without deliberate design and implementation strategies—such as language access, graduated interpretive support (eg, layered explanations, visual cues, and thresholds for action), and outreach beyond already engaged users—expanded transparency may reinforce existing disparities by placing a greater interpretive burden on those least equipped to manage it. Translating access into meaningful, equitable engagement, therefore, depends on designing systems that support patients across stages while reducing, rather than redistributing, this burden.

Acknowledgments

We used generative artificial intelligence tools (ChatGPT [OpenAI], Claude [Anthropic], and QuillBot [Learneo, Inc]) in a limited capacity to assist with grammar, stylistic refinement, and clarity of manuscript text. We did not use artificial intelligence tools

XSL•FO
RenderX

to design this study, collect data, analyze data, interpret findings, or create original research content. All substantive content was drafted, critically reviewed, and verified by the authors, who take full responsibility for the integrity and accuracy of this paper.

Funding

This systematic review was supported by the Florida State University Institute for Successful Longevity and was partially supported by an Agency for Healthcare Research and Quality grant R21HS029969, a National Institute on Aging grant P30AG073105 and a Clinical and Translational Science Award funded through the National Institutes of Health (NIH) National Center for Advancing Translational Sciences (UM1TR005128). The funders were not involved in the study design, data collection, analysis, interpretation, or the writing of this paper.

Data Availability

All extracted data and risk of bias assessments used in this review are included in supplementary files. Copies of other study materials (eg, coding guide) are available upon request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy, and risk of bias tables.
[PDF File, 215 KB - jmir_v28i1e88259_app1.pdf]

Multimedia Appendix 2
Detailed summary of included studies.
[PDF File, 372 KB - jmir_v28i1e88259_app2.pdf]

Checklist 1
PRISMA 2020 checklist.
[PDF File, 88 KB - jmir_v28i1e88259_app3.pdf]

Checklist 2
SWiM checklist.
[PDF File, 51 KB - jmir_v28i1e88259_app4.pdf]

References

1. Department of Health and Human Services. 21st Century Cures Act: interoperability, information blocking, and the ONC Health IT Certification Program. : Federal Register; 2020 Apr URL: https://www.federalregister.gov/documents/2020/05/01/2020-07419/21st-century-cures-act-interoperability-information-blocking-and-the-onc-health-it-certification [accessed 2026-05-14]
2. Healthy People 2030 Framework. Office of Disease Prevention and Health Promotion. URL: https://odphp.health.gov/healthypeople/about/healthy-people-2030-framework [accessed 2026-05-12]
3. Strawley C, Richwine C. Individuals' Access and Use of Patient Portals and Smartphone Health Apps, 2022. : Office of the National Coordinator for Health Information Technology; 2023 Oct URL: https://healthit.gov/wp-content/uploads/2023/10/DB69_IndividualsAccess-UsePatientPortals_508.pdf [accessed 2026-05-12]
4. Witteman HO, Zikmund-Fisher BJ. Communicating laboratory results to patients and families. Clin Chem Lab Med 2019 Feb 25;57(3):359-364. [doi: 10.1515/cclm-2018-0634] [Medline: 30407910]
5. Madison BM, Lazaro GR, Scott MS, Greene DN, Lorey TS, De Jesús VR. Effective access to laboratory test results: a health equity issue that enhances diagnostic excellence. J Appl Lab Med 2023 May 4;8(3):635-644. [doi: 10.1093/jalm/jfac128] [Medline: 36722002]
6. Franco J, Morris NS, Fung MK. Defining and identifying laboratory literacy as a component of health literacy: an assessment of existing health literacy tools. Acad Pathol 2023;10(4):100096. [doi: 10.1016/j.acpath.2023.100096] [Medline: 37964769]
7. Bruno B, Steele S, Carbone J, Schneider K, Posk L, Rose SL. Informed or anxious: patient preferences for release of test results of increasing sensitivity on electronic patient portals. Health Technol (Berl) 2022;12(1):59-67. [doi: 10.1007/s12553-021-00628-5] [Medline: 35036280]
8. Bhalla S, Prasad T, Xie D, Gerber DE. Contemporary trends in reviewing test results through the electronic patient portal among patients with cancer. JAMA Oncol 2024 Jan 1;10(1):139-140. [doi: 10.1001/jamaoncol.2023.5047] [Medline: 38032648]

XSL•FO
RenderX

9. Tossaint-Schoenmakers R, Kasteleyn M, Goedhart A, Versluis A, Talboom-Kamp E. The impact of patient characteristics on their attitudes toward an online patient portal for communicating laboratory test results: real-world study. JMIR Form Res 2021 Dec 17;5(12):e25498. [doi: 10.2196/25498] [Medline: 34927593]
10. Talboom-Kamp E, Tossaint-Schoenmakers R, Goedhart A, Versluis A, Kasteleyn M. Patients' attitudes toward an online patient portal for communicating laboratory test results: real-world study using the eHealth impact questionnaire. JMIR Form Res 2020 Mar 4;4(3):e17060. [doi: 10.2196/17060] [Medline: 32024632]
11. Zhong X, Park J, Liang M, et al. Characteristics of patients using different patient portal functions and the impact on primary care service utilization and appointment adherence: retrospective observational study. J Med Internet Res 2020 Feb 25;22(2):e14410. [doi: 10.2196/14410] [Medline: 32130124]
12. Steitz BD, Turer RW, Salmi L, et al. Repeated access to patient portal while awaiting test results and patient-initiated messaging. JAMA Netw Open 2025 Apr 1;8(4):e254019. [doi: 10.1001/jamanetworkopen.2025.4019] [Medline: 40198070]
13. Zhang Z, Citardi D, Xing A, Luo X, Lu Y, He Z. Patient challenges and needs in comprehending laboratory test results: mixed methods study. J Med Internet Res 2020 Dec 7;22(12):e18725. [doi: 10.2196/18725] [Medline: 33284117]
14. Zikmund-Fisher BJ, Scherer AM, Witteman HO, et al. Graphics help patients distinguish between urgent and non-urgent deviations in laboratory test results. J Am Med Inform Assoc 2017 May 1;24(3):520-528. [doi: 10.1093/jamia/ocw169] [Medline: 28040686]
15. Lazaro G. When positive is negative: health literacy barriers to patient access to clinical laboratory test results. J Appl Lab Med 2023 Nov 2;8(6):1133-1147. [doi: 10.1093/jalm/jfad045] [Medline: 37681277]
16. Clinical laboratory improvement advisory committee summary report, October 28-29, 2020. : CDC; 2021 Jan 25 URL: https://stacks.cdc.gov/view/cdc/116293 [accessed 2026-05-12]
17. Dahm MR, Georgiou A, Herkes R, et al. Patient groups, clinicians and healthcare professionals agree - all test results need to be seen, understood and followed up. Diagnosis (Berl) 2018 Nov 27;5(4):215-222. [doi: 10.1515/dx-2018-0083] [Medline: 30332391]
18. Petrovskaya O, Karpman A, Schilling J, et al. Patient and health care provider perspectives on patient access to test results via web portals: scoping review. J Med Internet Res 2023 Oct 19;25:e43765. [doi: 10.2196/43765] [Medline: 37856174]
19. van der Mee FAM, Schaper F, Jansen J, Bons JAP, Meex SJR, Cals JWL. Enhancing patient understanding of laboratory test results: systematic review of presentation formats and their impact on perception, decision, action, and memory. J Med Internet Res 2024 Aug 12;26:e53993. [doi: 10.2196/53993] [Medline: 39133906]
20. Skills for 21st-century librarians. American Library Association. 2021 Jul 23. URL: https://www.ala.org/tools/programming/21stcenturyskills [accessed 2026-05-14]
21. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
22. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan;10(1):39. [doi: 10.1186/s13643-020-01542-z]
23. Cadamuro J, Hillarp A, Unger A, et al. Presentation and formatting of laboratory results: a narrative review on behalf of the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group "Postanalytical Phase" (WG-POST). Crit Rev Clin Lab Sci 2021 Aug;58(5):329-353. [doi: 10.1080/10408363.2020.1867051] [Medline: 33538219]
24. Turchioe MR, Myers A, Isaac S, et al. A systematic review of patient-facing visualizations of personal health data. Appl Clin Inform 2019 Aug;10(4):751-770. [doi: 10.1055/s-0039-1697592] [Medline: 31597182]
25. Lustria MLA, Aliche O, Killian MO, He Z. Enhancing patient engagement and understanding: is providing direct access to laboratory results through patient portals adequate? JAMIA Open 2025 Apr;8(2):ooaf009. [doi: 10.1093/jamiaopen/ooaf009] [Medline: 40130170]
26. Scherer AM, Witteman HO, Solomon J, Exe NL, Fagerlin A, Zikmund-Fisher BJ. Improving the understanding of test results by substituting (not adding) goal ranges: web-based between-subjects experiment. J Med Internet Res 2018 Oct 19;20(10):e11027. [doi: 10.2196/11027] [Medline: 30341053]
27. Solomon J, Scherer AM, Exe NL. Is this good or bad? Redesigning visual displays of medical test results in patient portals to provide context and meaning. CHI EA '16: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems 2016 May 7:2314-2320. [doi: 10.1145/2851581.2892523]
28. Zikmund-Fisher BJ, Exe NL, Witteman HO. Numeracy and literacy independently predict patients' ability to identify out-of-range test results. J Med Internet Res 2014 Aug 8;16(8):e187. [doi: 10.2196/jmir.3241] [Medline: 25135688]
29. Nystrom DT, Singh H, Baldwin J, Sittig DF, Giardina TD. Methods for patient-centered interface design of test result display in online portals. EGEMS (Wash DC) 2018 Jun 26;6(1):15. [doi: 10.5334/egems.255] [Medline: 30094287]
30. Mák G, Smith Fowler H, Leaver C, Hagens S, Zelmer J. The effects of web-based patient access to laboratory results in British Columbia: a patient survey on comprehension and anxiety. J Med Internet Res 2015 Aug 4;17(8):e191. [doi: 10.2196/jmir.4350] [Medline: 26242801]
31. Robinson S, Reed M, Quevillon T, Hirvi E. Patient perceptions and interactions with their web portal-based laboratory results. BMJ Health Care Inform 2019 Apr;26(1):0. [doi: 10.1136/bmjhci-2019-000012] [Medline: 31039117]

XSL•FO
RenderX

32. Morrow D, Azevedo RFL, Garcia-Retamero R, et al. Contextualizing numeric clinical test results for gist comprehension: Implications for EHR patient portals. J Exp Psychol Appl 2019 Mar;25(1):41-61. [doi: 10.1037/xap0000203] [Medline: 30688498]
33. Fraccaro P, Vigo M, Balatsoukas P, et al. Presentation of laboratory test results in patient portals: influence of interface design on risk interpretation and visual search behaviour. BMC Med Inform Decis Mak 2018 Feb 12;18(1):11. [doi: 10.1186/s12911-018-0589-7] [Medline: 29433495]
34. Zikmund-Fisher BJ, Scherer AM, Witteman HO, Solomon JB, Exe NL, Fagerlin A. Effect of harm anchors in visual displays of test results on patient perceptions of urgency about near-normal values: experimental study. J Med Internet Res 2018 Mar 26;20(3):e98. [doi: 10.2196/jmir.8889] [Medline: 29581088]
35. Giardina TD, Baldwin J, Nystrom DT, Sittig DF, Singh H. Patient perceptions of receiving test results via online portals: a mixed-methods study. J Am Med Inform Assoc 2018 Apr 1;25(4):440-446. [doi: 10.1093/jamia/ocx140] [Medline: 29240899]
36. Giardina TD, Modi V, Parrish DE, Singh H. The patient portal and abnormal test results: an exploratory study of patient experiences. Patient Exp J 2015;2(1):148-154. [Medline: 28345018]
37. Schultz CL, Alderfer MA. Are on-line patient portals meeting test result preferences of caregivers of children with cancer? A qualitative exploration. Pediatr Blood Cancer 2018 Nov;65(11):e27306. [doi: 10.1002/pbc.27306] [Medline: 30007016]
38. Struikman B, Bol N, Goedhart A, et al. Features of a patient portal for blood test results and patient health engagement: web-based pre-post experiment. J Med Internet Res 2020 Jul 20;22(7):e15798. [doi: 10.2196/15798] [Medline: 32706704]
39. Pillemer F, Price RA, Paone S, et al. Direct release of test results to patients increases patient engagement and utilization of care. PLoS ONE 2016;11(6):e0154743. [doi: 10.1371/journal.pone.0154743] [Medline: 27337092]
40. Wood KE, Pham HT, Carter KD, Nepple KG, Blum JM, Krasowski MD. Impact of a switch to immediate release on the patient viewing of diagnostic test results in an online portal at an academic medical center. J Pathol Inform 2023;14(4):100323. [doi: 10.1016/j.jpi.2023.100323] [Medline: 37520309]
41. Zhang Z, Kmoth L, Luo X, He Z. User-centered system design for communicating clinical laboratory test results: design and evaluation study. JMIR Hum Factors 2021 Nov 25;8(4):e26017. [doi: 10.2196/26017] [Medline: 34842529]
42. Monkman H, Griffith J, MacDonald L, Lesselroth B. Consumers' needs for laboratory results portals: questionnaire study. JMIR Hum Factors 2023 Jun 12;10:e42843. [doi: 10.2196/42843] [Medline: 37307049]
43. Krasowski MD, Grieme CV, Cassady B, et al. Variation in results release and patient portal access to diagnostic test results at an academic medical center. J Pathol Inform 2017;8(6):45. [doi: 10.4103/jpi.jpi_53_17] [Medline: 29226008]
44. Foster B, Krasowski MD. The use of an electronic health record patient portal to access diagnostic test results by emergency patients at an academic medical center: retrospective study. J Med Internet Res 2019 Jun 28;21(6):e13791. [doi: 10.2196/13791] [Medline: 31254335]
45. Turer RW, Martin KR, Courtney DM, et al. Real-time patient portal use among emergency department patients: an open results study. Appl Clin Inform 2022 Oct;13(5):1123-1130. [doi: 10.1055/a-1951-3268] [Medline: 36167337]
46. Steitz BD, Guide A, Rodriguez K, et al. Patient-friendly test results and patient-initiated messaging among adult outpatients. JAMA Netw Open 2025 Nov 3;8(11):e2543879. [doi: 10.1001/jamanetworkopen.2025.43879]
47. Steitz BD, Turer RW, Lin CT, et al. Perspectives of patients about immediate access to test results through an online patient portal. JAMA Netw Open 2023 Mar 1;6(3):e233572. [doi: 10.1001/jamanetworkopen.2023.3572] [Medline: 36939703]
48. Hulter P, Langendoen W, Pluut B, et al. Patients' choices regarding online access to laboratory, radiology and pathology test results on a hospital patient portal. PLoS ONE 2023;18(2):e0280768. [doi: 10.1371/journal.pone.0280768] [Medline: 36735739]
49. Monkman H, Griffith J, Macdonald L, Joseph AL, Lesselroth B. Why do people use online lab results and what do they look for: a qualitative study. Stud Health Technol Inform 2022 May 25;294:599-603. [doi: 10.3233/SHTI220539] [Medline: 35612159]
50. Hulter P, Weggelaar-Jansen A, Ahaus K, Pluut B. Patient discourses on real-time access to test results via hospital portals: a discourse analysis of semistructured interviews with Dutch patients. BMJ Open 2024 Nov 24;14(11):e088201. [doi: 10.1136/bmjopen-2024-088201] [Medline: 39581732]
51. Monkman H, MacDonald L, Nøhr C, Tanaka JW, Lesselroth BJ. Hidden in plain sight: overlooked results and other errors in evaluating online laboratory results. Stud Health Technol Inform 2022 Jun 6;290:867-871. [doi: 10.3233/SHTI220203] [Medline: 35673142]
52. Monkman H, MacDonald L, Joseph AL, Lesselroth BJ. Tabular, annotated, visual, or trends + contextual information? Preferences for online laboratory results displays. Stud Health Technol Inform 2024 Jan 25;310(106–10):1041-1045. [doi: 10.3233/SHTI231123] [Medline: 38269973]
53. Monkman H, Schmit A, Nyholt D, MacDonald L, Lesselroth B. False calm and false alarm: a qualitative study of confusion and misinterpretation of a laboratory results graph. Stud Health Technol Inform 2025 May 12;326:106-110. [doi: 10.3233/SHTI250248] [Medline: 40357611]
54. Joseph AL, Monkman H, MacDonald L, Lai C. Interpreting laboratory results with complementary health information: a human factors perspective. Stud Health Technol Inform 2024 Jan 25;310:1061-1065. [doi: 10.3233/SHTI231127] [Medline: 38269977]

XSL•FO
RenderX

55. Christensen K, Sue V. Viewing laboratory test results online: patients' actions and reactions. J Participat Med 2013 Oct;3(5) [FREE Full text]
56. McFarland JA, Huang J, Li Y, Gunn AJ, Morgan DE. Patient engagement with online portals and online radiology results. Curr Probl Diagn Radiol 2023;52(2):106-109. [doi: 10.1067/j.cpradiol.2022.07.012] [Medline: 36030140]
57. Barker TH, Stone JC, Sears K, et al. The revised JBI critical appraisal tool for the assessment of risk of bias for randomized controlled trials. JBI Evid Synth 2023 Mar 1;21(3):494-506. [doi: 10.11124/JBIES-22-00430] [Medline: 36727247]
58. Barker TH, Habibi N, Aromataris E, et al. The revised JBI critical appraisal tool for the assessment of risk of bias for quasi-experimental studies. JBI Evid Synth 2024 Mar 1;22(3):378-388. [doi: 10.11124/JBIES-23-00268] [Medline: 38287725]
59. Long HA, French DP, Brooks JM. Optimising the value of the Critical Appraisal Skills Programme (CASP) tool for quality appraisal in qualitative evidence synthesis. Res Methods Med Health Sci 2020 Sep;1(1):31-42. [doi: 10.1177/2632084320947559]
60. Downes MJ, Brennan ML, Williams HC, Dean RS. Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). BMJ Open 2016 Dec 8;6(12):e011458. [doi: 10.1136/bmjopen-2016-011458] [Medline: 27932337]
61. Hong QN, Fàbregues S, Bartlett G, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. EFI 2018;34(4):285-291. [doi: 10.3233/EFI-180221]
62. Carra MC, Romandini P, Romandini M. Risk of bias evaluation of cross-sectional studies: adaptation of the Newcastle-Ottawa Scale. J Periodontal Res 2025 Apr 28;368:l6890. [doi: 10.1111/jre.13405] [Medline: 40293188]
63. Campbell M, McKenzie JE, Sowden A, et al. Synthesis Without Meta-Analysis (SWiM) in systematic reviews: reporting guideline. BMJ 2020 Jan 16;368:l6890. [doi: 10.1136/bmj.l6890] [Medline: 31948937]
64. Garcia-Retamero R, Cokely ET. Effective communication of risks to young adults: using message framing and visual aids to increase condom use and STD screening. J Exp Psychol Appl 2011 Sep;17(3):270-287. [doi: 10.1037/a0023677] [Medline: 21942316]
65. Kindler CH, Harms C, Amsler F, Ihde-Scholl T, Scheidegger D. The visual analog scale allows effective measurement of preoperative anxiety and detection of patients' anesthetic concerns. Anesth Analg 2000 Mar;90(3):706-712. [doi: 10.1097/00000539-200003000-00036] [Medline: 10702461]
66. Kelly L, Ziebland S, Jenkinson C. Measuring the effects of online health information: scale validation for the e-Health Impact Questionnaire. Patient Educ Couns 2015 Nov;98(11):1418-1424. [doi: 10.1016/j.pec.2015.06.008] [Medline: 26162953]
67. Hibbard JH, Mahoney ER, Stockard J, Tusler M. Development and testing of a short form of the patient activation measure. Health Serv Res 2005 Dec;40(6 Pt 1):1918-1930. [doi: 10.1111/j.1475-6773.2005.00438.x] [Medline: 16336556]
68. Brooke J. SUS - a quick and dirty usability scale. In: Usability Evaluation in Industry: CRC Press; 1996:4-7. [doi: 10.1201/9781498710411]
69. Reichheld FF. The one number you need to grow. Harv Bus Rev 2003 Dec;81(12):46-54. [doi: 10.2196/42843] [Medline: 14712543]
70. Kreuter MW, Skinner CS. Tailoring: what's in a name? Health Educ Res 2000 Feb;15(1):1-4. [doi: 10.1093/her/15.1.1] [Medline: 10788196]
71. Lustria MLA, Cortese J, Gerend MA, Schmitt K, Kung YM, McLaughlin C. A model of tailoring effects: a randomized controlled trial examining the mechanisms of tailoring in a web-based STD screening intervention. Health Psychol 2016 Nov;35(11):1214-1224. [doi: 10.1037/hea0000399] [Medline: 27441869]
72. Kreuter MW, Wray RJ. Tailored and targeted health communication: strategies for enhancing information relevance. Am J Health Behav 2003;27 Suppl 3(2):S227-S232. [doi: 10.5993/ajhb.27.1.s3.6] [Medline: 14672383]
73. Lustria MLA, Cortese J, Noar SM, Glueckauf RL. Computer-tailored health interventions delivered over the web: review and analysis of key components. Patient Educ Couns 2009 Feb;74(2):156-173. [doi: 10.1016/j.pec.2008.08.023] [Medline: 18947966]
74. Hawkins RP, Kreuter M, Resnicow K, Fishbein M, Dijkstra A. Understanding tailoring in communicating about health. Health Educ Res 2008 Jun;23(3):454-466. [doi: 10.1093/her/cyn004] [Medline: 18349033]
75. Borycki EM, Kletke R, Whitehouse S, Kushniruk AW. Empathetic and emotive design: a scoping review. Stud Health Technol Inform 2024 May 23;314:85-89. [doi: 10.3233/SHTI240063] [Medline: 38785009]
76. Yoon J, Li S, Hao Y, Kim C. Towards emotional well-being by design: 17 opportunities for emotion regulation for user-centered healthcare design. PervasiveHealth'19 2019 May 20:351-355. [doi: 10.1145/3329189.3329227]

Abbreviations

ED: emergency department
EHR: electronic health record
HbA_{1c}: hemoglobin A_{1c}
HDL: high-density lipoprotein
JBI: Joanna Briggs Institute

XSL•FO
RenderX

MeSH: Medical Subject Headings
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension

Edited by S Brini; submitted 22.Nov.2025; peer-reviewed by B Zikmund-Fisher, O Akinpeloye; revised version received 30.Apr.2026; accepted 04.May.2026; published 29.May.2026.

Please cite as:
Lustria MLA, Atrinawati L, Aliche O, Kim K, He Z
Patient Experiences With Online Laboratory Test Presentations From Access to Activation: Systematic Review
J Med Internet Res 2026;28:e88259
URL: https://www.jmir.org/2026/1/e88259
doi:10.2196/88259

© Mia Liza A Lustria, Lovinta Atrinawati, Obianuju Aliche, Kyung Kim, Zhe He. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Use, Utility, and User Experience of Cloud-Based Medical Imaging in Pulmonary Nodule Care in China: Mixed Methods Study

Ziguo Chen[1*], MM; Junhan Wu[1*], MD; Yunying Chen[2*], MM; Weitao Zhuang[3], MD; Jiaxuan Huang[4], MM; Weifeng Zhong[1], MM; Zijie Li[5], MM; Shuhuan Xie[1], MD; Chaofan Liu[1], MM; Guojie Lu[1], MM; Guibin Qiao[1], MD, PhD

[1]Department of Thoracic Surgery, Zhujiang Hospital, Southern Medical University, No. 253 Industrial Avenue Middle, Haizhu District, Guangzhou, China

[2]The Fifth Affiliated Hospital, Sun Yat-sen University, Zhuhai, China

[3]State Key Laboratory of Respiratory Diseases, National Clinical Research Center for Respiratory Disease, National Center for Respiratory Medicine, Department of Pulmonary and Critical Care Medicine, Guangzhou Institute of Respiratory Health, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China

[4]School of Medicine, South China University of Technology, Guangzhou, China

[5]Shantou University Medical College, Shantou University, Shantou, China

[*]these authors contributed equally

Corresponding Author:
Guibin Qiao, MD, PhD
Department of Thoracic Surgery, Zhujiang Hospital, Southern Medical University, No. 253 Industrial Avenue Middle, Haizhu District, Guangzhou, China

Abstract

Background: The detection of pulmonary nodules (PNs) has increased with the use of low-dose computed tomography screening. Effective management requires timely longitudinal surveillance and reliable comparison with prior examinations, yet access to previous imaging across institutions is often fragmented, leading to delays and potentially unnecessary repeat scans and costs. Cloud-based medical imaging (CMI) solutions offer a potential means of improving access and facilitating cross-institutional data exchange. However, the adoption and utility of CMI in PN care, especially in China, remain underexplored.

Objective: This study aims to evaluate the possession, use, and impact of CMI on health care utilization, patient knowledge, and financial burden, as well as to identify usability and interoperability barriers through qualitative investigation.

Methods: A mixed methods cross-sectional study was conducted from October 2022 to May 2024. The study involved 701 patients with PNs who completed structured surveys, and 20 participants (10 patients and 10 physicians) were interviewed. CMI use was defined as self-reported ability to view radiological images on a mobile device. We compared CMI users and nonusers and estimated adjusted odds ratios using multivariable logistic regression, then applied 1:1 propensity score matching to examine associations between CMI use and health care utilization, costs, and patient perceptions, and qualitative interviews were analyzed for usability themes.

Results: The study found that 611 (87.2%) out of 701 patients had obtained CMI, with 404 (57.6%) out of 701 patients actively using it. In multivariable analysis, older age was independently associated with lower CMI use (odds ratios 0.985, 95% CI 0.972 - 0.999). After 1:1 propensity score matching, CMI users accessed more internet hospitals, consulted more physicians, and reported lower health care costs compared to nonusers. Users also demonstrated higher disease knowledge. Qualitative data identified key barriers, including poor system usability, limited retention time for images, and weak interoperability. CMI was perceived as beneficial for patient convenience and clinical efficiency, though concerns over image quality and system fragmentation were prevalent.

Conclusions: While CMI is widely available, its usage remains suboptimal. Increased use is associated with enhanced health care engagement and reduced costs, suggesting that improving system usability and ensuring consistent access to imaging may help realize potential benefits of CMI. Future improvements should focus on ensuring long-term access, better retention protocols, and overcoming interoperability issues.

(J Med Internet Res 2026;28:e86745) doi:10.2196/86745

KEYWORDS

cloud-based medical imaging; pulmonary nodules; health care utilization; telemedicine; qualitative interview

XSL•FO
RenderX

Introduction

With the widespread adoption of low-dose computed tomography screening, the detection rate of pulmonary nodules (PNs) continues to increase [1]. Longitudinal imaging surveillance and comparison with prior examinations are foundational to PN management and are consistently emphasized in major guidelines [2,3]. In routine practice, patients often receive care across multiple institutions and seek additional opinions; when prior images are not retrievable promptly and reliably, repeat scanning and delays in care become likely. Evidence indicates that cross-regional or interinstitutional health information exchange (HIE) is associated with fewer duplicate imaging examinations and lower related costs [4,5].

In recent years, internet hospitals that commonly refer to hospital-affiliated online clinical services that extend routine care pathways through digital platforms, rather than stand-alone telemedicine encounters, have become an important complement to in-person clinics for remote consultation and follow-up. Real-world data suggest that their use is associated with fewer visits and lower costs [6,7], yet such encounters rely even more heavily on HIE. Enabling cross-institutional and cross-channel reuse of imaging requires cloud-based tools built on the Digital Imaging and Communications in Medicine (DICOM) or DICOMweb standards, which are collectively referred to as cloud-based medical imaging (CMI), and they support interoperable data exchange and network retrieval [8]. Recent implementations show that DICOMweb-based workflows can deliver web-based image viewing, improving visit efficiency. Recent surveys of web-based DICOM viewers suggest that DICOMweb support is increasingly common, but implementation depth and performance vary across products and settings [9]. Meanwhile, direct patient access to images is increasingly common. Although this can enhance knowledge and engagement, the absence of patient-facing explanations may trigger or amplify “scanxiety” [10]. Accordingly, information availability and interpretability represent distinct challenges. Recent studies showed that patient engagement with radiology reports and images is heterogeneous with respect to uptake, timing, and depth and differs across patients [11]. When patients review radiology results, they often encounter clinician-oriented terminology with limited general explanation, which may hinder comprehension and contribute to concern even when results are available in a timely manner [12,13]. A recent systematic review reported that patient-oriented reporting features, including plain-language summaries, glossaries, and visual aids, are associated with improved comprehension and satisfaction and may reduce avoidable anxiety [12].

In this imaging-dependent clinical context, timely retrieval and cross-channel reuse of imaging studies are prerequisites for effective longitudinal management. However, empirical evidence specific to populations with PNs remains sparse regarding CMI availability versus actual use, its associations with health care utilization and financial burden, user experience, and interoperability barriers. While digital image accessibility has been studied in general contexts, this research represents the first mixed methods study to focus on the unique PN care landscape in China. In this context, effective management is heavily dependent on long-term, high-frequency longitudinal imaging comparisons, making current limitations, such as brief data retention and system fragmentation, particularly impactful. Furthermore, unlike the unified patient portals common in other regions, the Chinese mobile health ecosystem is characterized by fragmented, QR code–based access, which presents distinct usability challenges. Beyond describing accessibility, this study provides a robust analysis of the correlations between actual CMI use and health care utilization, financial burden, and patient cognition, offering a holistic perspective on the “possession-to-use gap” that has not been previously characterized. This study aimed to characterize, within PN care, the landscape of CMI availability and use; quantify its associations with health care utilization and costs; assess relationships with disease perceptions; and, through qualitative inquiry, identify key usability and interoperability barriers to inform optimization and scale-up.

Methods

Study Design and Participants

We used an explanatory sequential mixed methods design. Survey findings informed the interview guide and purposive sampling, and integration was conducted through a prespecified comparison framework and meta-inferences based on convergence, complementarity, and discrepancies. Integrated findings are reported using narrative weaving that links quantitative results with qualitative themes and illustrative quotes. The reporting of the quantitative component follows the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guideline, and the qualitative component was designed, conducted, and reported with reference to the COREQ (Consolidated Criteria for Reporting Qualitative Research). From October 2022 to May 2024, consecutive patients with PNs were recruited via a thoracic surgery e-consultation clinic on a national telemedicine platform [14]. In total, 701 participants completed the structured survey. Eligibility criteria were newly identified or undersurveillance PNs, age ≥18 years, and ability to complete the questionnaire. We excluded individuals who had already undergone resection or had a definitive pathological diagnosis.

To complement and contextualize the quantitative findings, we conducted semistructured interviews with a purposive, maximum-variation sample of 10 patients and 10 physicians. Patient participants varied by age (26‑73 y), sex, and educational level. Physician participants were thoracic surgeons spanning junior, mid-career, and senior ranks.

Ethical Considerations

This study was approved by the Research Ethics Committee of Guangdong Provincial People’s Hospital (approval number: 2025-KY-122‑01). All participants provided written informed consent under institutional ethics approval. All questionnaire and interview data were deidentified before analysis, and access was restricted to the research team. The research team collected questionnaire responses and qualitative interview materials; we did not collect, download, store, or process participants’ DICOM files or other raw radiology image files as part of the study

XSL•FO
RenderX

procedures. Participants received no compensation for participation.

Measures and Variables

In China, CMI is commonly obtained or shared as DICOM files viewable in a browser, QR codes appended to radiology reports, or links on hospital websites. For QR code or web link access, the code or link typically directs patients to an institution-hosted or vendor-hosted web viewer. Depending on the implementation, the viewer may be DICOM-based (including zero-footprint viewers) or may rely on proprietary web formats; therefore, QR or link access should not be assumed to be DICOMweb-compliant. Some patients receive only a textual report and/or film, with no CMI access. We defined use of CMI as the self-reported ability to view radiological images on a mobile device. This measure does not differentiate between passive access and more active or intensive use, nor does it quantify frequency or depth of use. Sociodemographic and clinical variables included age, sex, education (primary, middle school, high school, junior college, undergraduate, and graduate), and surveillance status. Care-seeking and cost measures comprised the number of hospitals used (internet and offline), the total number of physicians consulted, and annual out-of-pocket expenditures (care-related fees, transportation, lodging). Self-efficacy and disease perceptions included perceived disease severity, level of disease awareness, and change in anxiety level.

Qualitative Interviews: Design, Data Collection, and Management

We developed separate interview guides for patients and physicians (Multimedia Appendix 1). The guides covered CMI awareness and use, preferences of CMI format, perceived clinical benefits and limitations, workflow usability and pain points, psychological responses, and priorities for system improvement. Trained qualitative researchers conducted in-person interviews (20 - 40 min). With permission, sessions were audio-recorded and accompanied by field notes. Recordings were transcribed verbatim, deidentified, and stored on password-protected devices. Texts were managed and queried in NVivo (version 11; QSR International Pty Ltd). Bilingual researchers translated quotations for manuscript presentation with dual review; discrepancies were resolved by revisiting the Chinese original. Additionally, interviews and analysis proceeded iteratively. During concurrent coding, we tracked the emergence of new codes and concluded interviewing when successive interviews no longer generated new major themes and the coding framework had stabilized.

Data Analysis

Quantitative Analysis

Continuous variables were summarized using medians and interquartile ranges, whereas categorical variables were summarized as frequencies and percentages. The primary outcome was the use of CMI, defined as whether participants accessed and viewed their imaging data through CMI. Group comparisons were performed using chi-square or Mann-Whitney *U* tests. To quantify associations with CMI use, odds ratios with 95% CIs were estimated using univariable and multivariable logistic regression. To minimize confounding, propensity score matching was conducted with CMI use as the dependent variable. Prespecified covariates included age, sex, education, and surveillance duration. Nearest-neighbor 1:1 matching was applied using a caliper width of 0.02 times the SD of the logit of the propensity score. Covariate balance was assessed using standardized mean differences (<0.10). In the matched sample, health care utilization, financial burden, and disease perceptions were compared between CMI users and nonusers. Analyses were performed using SPSS Statistics (version 26.0; IBM Corp), and visualizations were created in GraphPad Prism (version 8.0).

Qualitative Analysis and Integration

Two researchers independently reviewed all transcripts, wrote reflexive memos, and conducted line-by-line open coding using in vivo codes where appropriate. After independent coding, the team discussed and resolved discrepancies; merged overlapping codes; refined code boundaries; and compiled a shared, version-controlled codebook. A constant comparative approach was used to group semantically related codes into categories within each participant group and to compare categories across groups using a patient-versus-physician matrix, identifying areas of convergence and divergence. Potentially disconfirming cases were examined to test the robustness of themes, which were iteratively refined based on clarity, coherence, and distinctiveness. Approximately 20% of the transcripts were double-coded for quality assurance, with disagreements resolved through discussion. An auditable analytic trail and reflexive notes were maintained throughout. Member checking was not undertaken. However, integration was performed through using the quantitative results to inform interview focus and an explicit joint display and narrative weaving to connect quantitative patterns with qualitative explanations and identify areas of convergence and discrepancy.

Results

CMI Possession and Use

Overview

Among 701 patients with PNs, 87.2% (n=611) had obtained CMI. Of these, 47.1% (n=330) acquired it by requesting from hospitals, while 40.1% (n=281) received it proactively from hospitals; 12.8% (n=90) had only reports or films without CMI. The main modalities were QR code or web link (n=352, 50.3%) and DICOM files (n=259, 36.9%). Overall, 57.6% (n=404) were CMI users, and 42.4% (n=297) were nonusers (Table 1).

XSL•FO
RenderX

Table . Possession and application of CMI[a] among patients with PNs[b].

Possession categories, application of CMI, and variables	Patients, n (%)
Possession of CMI	
No, I only got radiological report and/or film	90 (12.8)
Radiological reports and films	79 (11.3)
Radiological reports	11 (1.6)
Yes, I got it by applying to the hospital	330 (47.1)
DICOM[c] files	259 (36.9)
Yes, hospitals took initiative to provide it	281 (40.1)
QR code or web links	352 (50.3)
Application of CMI	
Users of CMI	404 (57.6)
Nonusers of CMI	297 (42.4)

[a]CMI: cloud medical images.
[b]PN: pulmonary nodule.
[c]DICOM: Digital Imaging and Communications in Medicine.

Baseline Differences and Independent Correlates of CMI Use

In the unmatched sample (Table 2), users were younger than nonusers (median 46, IQR 38 - 55 y vs median 48, IQR 39 - 57 y; *P*=.001), and distributions differed for education (overall *P*=.004) and prior internet-clinic experience (overall *P*=.03). Sex (*P*=.60) and time since nodule detection (*P*=.09) did not differ. In multivariable logistic regression (Table 3), age was independently and inversely associated with CMI use (odds ratios 0.985, 95% CI 0.972 - 0.999; *P*=.04), while education was not significant after adjustment (*P*=.21).

Table . Comparison of baseline characteristics between users and nonusers of CMI[a].

Variable	Overall (N=701)	Users (n=404, 57.6%)	Nonusers (n=297, 42.4%)	*P* value
Age (y), median (IQR)	48 (39-57)	46 (38-55)	48 (39-57)	.001
Sex, n (%)				.60
Female	435 (62.1)	254 (62.9)	184 (60.9)	
Male	266 (37.9)	151 (37.1)	105 (39.1)	
Education attainment, n (%)				.004
Primary school	24 (3.4)	8 (2.0)	16 (5.4)	
Junior high school	68 (9.7)	41 (10.1)	27 (9.1)	
Senior high school	88 (12.6)	43 (10.6)	45 (15.2)	
Junior college	150 (21.4)	79 (19.6)	71 (23.9)	
Undergraduate	291 (41.5)	181 (44.8)	110 (37.0)	
Graduate	80 (11.4)	52 (12.9)	28 (9.4)	
Online consultation, n (%)				.03
Never	207 (29.5)	99 (24.5)	108 (36.4)	
Ever but need assistance	288 (41.1)	184 (45.5)	104 (35.0)	
Ever and not need assistance	266 (37.9)	121 (30.0)	85 (28.6)	
Time to the first detection of PNs[b] (mo), median (IQR)	12 (3-32)	14 (4-32)	12 (2-33)	.09

[a]CMI: cloud medical images.
[b]PNs: pulmonary nodules.

XSL•FO
RenderX

Table . Binomial logistic regression analysis for influencing factors of CMI[a] application in patients with PNs[b].

Variable	Univariate analyses		Multivariate analysis	
	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Age (y)	0.979 (0.967 - 0.991)	.001	0.985 (0.972 - 0.999)	.04
Time (mo)	0.998 (0.993 - 1.004)	.58	__[c]	—
Sex (reference: female)		.60	—	—
Male	0.921 (0.677 - 1.254)			
Education attainment (reference: primary school)	—	.02	—	.21
Junior high school	3.037 (1.142 - 8.075)	.03	2.655 (0.989 - 7.123)	.05
Senior high school	1.911 (0.742 - 4.922)	.18	1.651 (0.634 - 4.300)	.30
Junior college	2.225 (0.898 - 5.513)	.08	1.833 (0.726 - 4.626)	.20
Undergraduate	3.291 (1.363 - 7.943)	.008	2.449 (0.973 - 6.162)	.06
Graduate	3.714 (1.415 - 9.750)	.008	2.583. (0.929 - 7.178)	.07

[a]CMI: cloud-based medical images.

[b]PN: pulmonary nodule.

[c]Not applicable.

Health Care Utilization, Costs, and Perceptions After Matching

After 1:1 propensity score matching, baseline balance was achieved, the main directional differences observed prematching generally persisted (Table S1 in Multimedia Appendix 2) and the covariate balance before and after propensity score matching persisted in Multimedia Appendix 3 . In the matched cohort, as shown in Figure 1A, the distribution for number of internet hospitals (median 1, IQR 0 - 1 vs median 1, IQR 0 - 1; *P*=.01) used differed significantly between users and nonusers, whereas, as shown in Figure 1B, the number of offline hospitals (median 2, IQR 1 - 2 vs median 1, IQR 1 - 2; *P*=.19) did not differ. The number of physicians consulted (median 3, IQR 1 - 3 vs median 2, IQR 1 - 3; *P*=.02) is also displayed in Figure 1C. CMI users were significantly more likely to spend 0 - 2500 Chinese Yuan (CNY; US $1=6.87 CNY) on treatment and diagnostic costs (229/291, 78.7% vs 201/291, 69.1%; *P*=.004) and on travel and accommodation expenses (249/291, 85.6% vs 229/291, 78.7%; *P*=.03), suggesting they incurred lower expenses compared to nonusers. As presented in Figure 1F-H, regarding disease perceptions, users reported higher levels of disease awareness (overall *P*=.009), whereas perceived disease severity (overall *P*=.97) and anxiety change (overall *P*=.81) did not differ (Figure 1).

XSL•FO
RenderX

Figure 1. Health care utilization, costs, and perceptions after 1:1 propensity score matching. Note that 100% stacked bar charts compare cloud-based medical imaging (CMI) Users and Nonusers on 8 outcomes: (A) number of internet hospitals, (B) number of offline hospitals, (C) number of consulting physicians, (D) treatment and diagnostic costs, (E) travel and accommodation expenses, (F) perceived disease severity, (G) level of disease awareness, and (H) change in anxiety level. Bars show the percentage distribution of categories within each group (cost bands: Chinese Yuan [CNY] 0 - 2500, 2500 - 5000, 5000 - 7500, and 7500 - 10,000; ≥10,000; other category labels as shown in panel legends; all costs reported in CNY were converted to US dollars using an exchange rate of US $1=6.87 CNY). *P* values above panels compare distributions between the groups using 2-sided chi-square tests. Significant differences were observed for internet hospitals (*P*=.01), consulting physicians (*P*=.02), treatment or diagnostic costs (*P*=.004), travel or accommodation expenses (*P*=.03), and disease awareness (*P*=.009); no differences were found for offline hospitals (*P*=.19), perceived severity (*P*=.97), or change in anxiety (*P*=.81). Baseline balance after matching is shown in Table S1 in Multimedia Appendix 2.

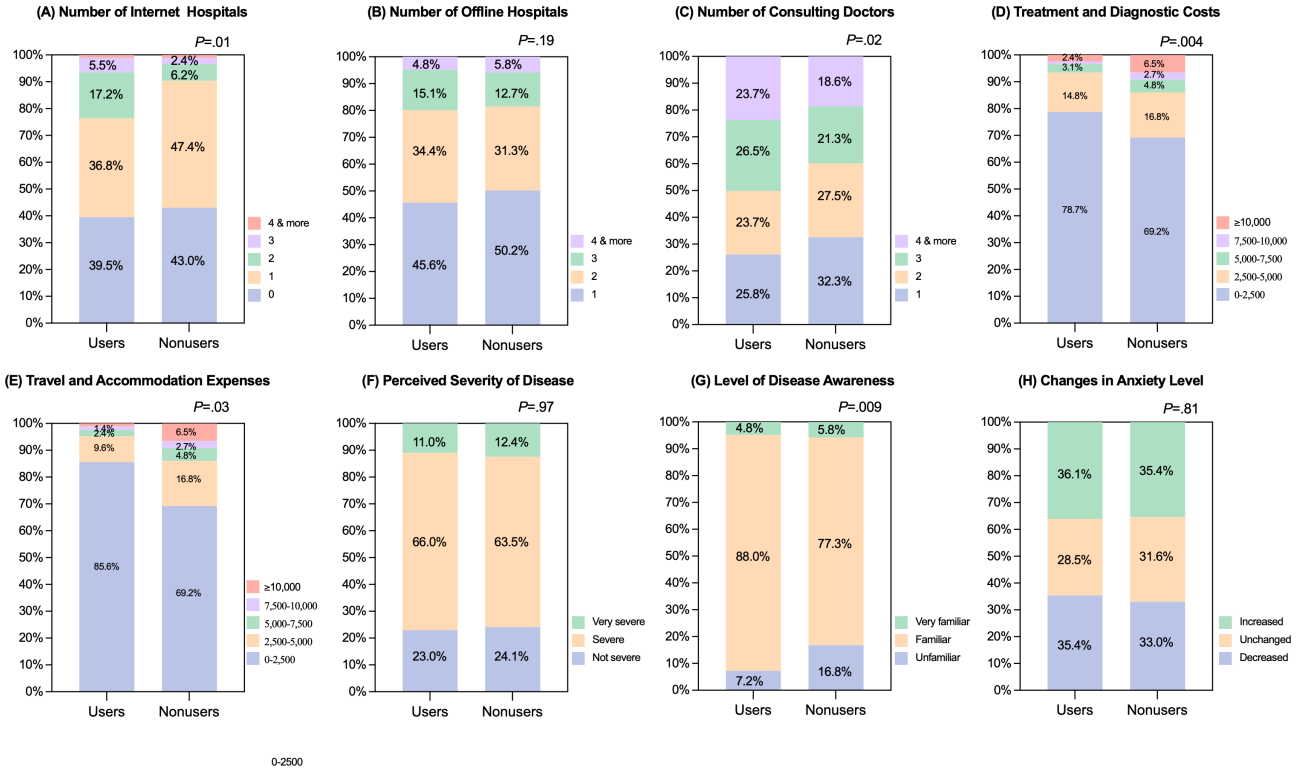

Stratification by Access Modality (DICOM vs QR Code or Web Link) Among Users

Among CMI users, in Figure 2A, the DICOM group differed from the QR or link group in distributions of the number of offline hospitals (median 2, IQR 1 - 2 vs median 2, IQR 1 - 2; *P*=.04) and the number of physicians consulted (median 3, IQR 2 - 4 vs median 3, IQR 1 - 3; *P*=.02) as depicted in Figure 2B, whereas as shown in Figure 2C, the number of internet hospitals (median 1, IQR 0 - 2 vs median 1, IQR 0 - 2; *P*=.99) did not differ. The DICOM group had no difference in treatment and diagnostic costs (95/132, 72.0% vs 209/261, 80.1%; *P*=.09) and travel and accommodation expenses (108/132, 82.6% vs 223/261, 85.4%; *P*=.52) with the majority spending 0 - 2500 CNY, but higher disease knowledge (overall *P*=.004) as presented in Figure 2G. The groups did not differ in perceived severity (overall *P*=.17) or anxiety change (overall *P*=.19; Figure 2).

XSL•FO
RenderX

Figure 2. Outcomes among cloud-based medical imaging (CMI) users stratified by access modality (Digital Imaging and Communications in Medicine [DICOM] vs QR code or web link). Among CMI users only, 100% stacked bar charts compare those accessing the service via DICOM versus QR code or web link across the same 8 outcomes: (A) number of internet hospitals, (B) number of offline hospitals, (C) number of consulting physicians, (D) treatment and diagnostic costs, (E) travel and accommodation expenses, (F) perceived disease severity, (G) level of disease awareness, and (H) change in anxiety level. Bars display percentage distributions for each category (cost bands identical to Figure 1). *P* values above panels indicate 2-sided chi-square tests comparing modalities. Significant differences were seen for offline hospitals (*P*=.04), consulting physicians (*P*=.02), and disease awareness (*P*=.004). No significant differences were observed for internet hospitals (*P*=.99), treatment or diagnostic costs (*P*=.09), travel or accommodation expenses (*P*=.52), perceived severity (*P*=.17), or change in anxiety (*P*=.19).

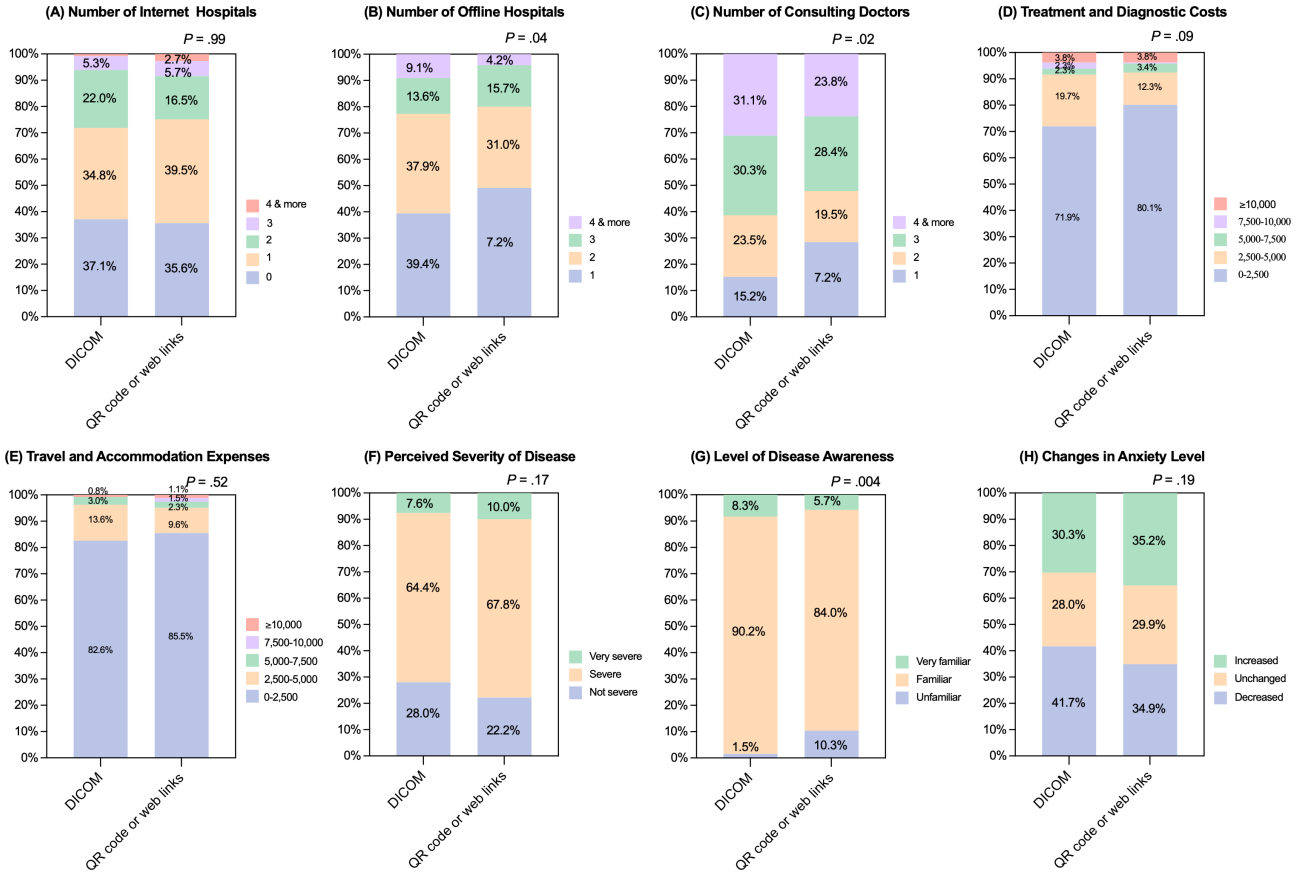

Integrated Findings Across Quantitative and Qualitative Strands

Across both strands, we observed a consistent possession-to-use gap: most participants obtained CMI, but only a subset reported using it. The interview data indicate that this gap reflects not only access but also day-to-day usability barriers and system limitations, including burdensome authentication steps, inconsistent workflows across hospitals, and short retention periods that make access links unreliable. After matching, the observed patterns in health care use and self-reported costs aligned with the convenience and efficiency benefits described in interviews, while psychological responses to viewing images varied in direction. A joint display linking key quantitative results with qualitative themes and illustrative quotes is provided in Table S2 in Multimedia Appendix 4.

Qualitative Themes

Overview

We identified five overarching themes: (1) convenience and clinical value, (2) heterogeneous user experience and system usability, (3) limited retention time and poor interoperability, (4) psychological effects and communication, and (5) improvement priorities and governance. Representative quotes and themes are provided below (Table 4) and in Multimedia Appendix 1.

XSL•FO
RenderX

Table . Themes and representative quotes.

Theme	Representative quotes
Care convenience and clinical value	• "I used to carry film to every visit; now showing the images on my phone is enough." (Patient) • ["When I follow up at another hospital, clinicians can view my prior images, saving many unnecessary steps." (Patient) • "Cloud-based imaging turns pulmonary nodule surveillance from discontinuous checkpoints into a continuous trajectory." (Physician)
Divergent user experience and system usability	• "Different hospitals require different accounts; I often can't remember them." (Patient) • "When the connection is poor, the whole process fails." (Patient) • "There are too many steps to access the images; during busy clinics this disrupts workflow." (Physician) • "Window width and window level are difficult to adjust, which affects assessment of key details." (Physician)
Limited retention and poor interoperability	• "QR codes expire after a while, so at the next visit I have to start over." (Patient) • "Restrictions on the number of downloads are inconvenient." (Patient) • "Lack of interoperability creates information silos, making repeat imaging hard to avoid." (Physician) • "Without unified standards, retrieving outside images feels like assembling a jigsaw." (Physician)
Psychological effects and communication	• "Seeing the imaging changes makes me feel more in control." (Patient) • "There is too much medical terminology I don't understand, which makes me more anxious." (Patient) • "We need to 'translate' images into conclusions and recommendations that patients can understand." (Physician)
Improvement priorities and governance	• "Ideally, a single platform would be usable nationwide." (Patient) • "Prioritize speed, clarity, stability, and privacy first; artificial intelligence (AI) features can come later." (Physician) • "It would be ideal to place history, reports, and images together in one 'case folder.'" (Physician)

Theme 1: Convenience and Clinical Value

Patients and physicians agreed that CMI reduces patient burden and improves clinical efficiency. Patients emphasized not needing to carry films and smoother cross-institution visits; physicians highlighted longitudinal comparison of nodules, fewer unnecessary repeat scans, shorter visit times, and improved access for distant patients.

> ***Patient:*** *I no longer carry a stack of films—showing the images on my phone is enough*
>
> ***Patient:*** *When I switch hospitals, prior images can be pulled up, and it's easier for the doctor to decide on follow-up.*
>
> ***Physician:*** *Having previous images side-by-side makes surveillance decisions more coherent and reduces repeat CTs.*
>
> ***Physician:*** *For out-of-town patients, remote consults with shared images are much more efficient.*

Theme 2: Heterogeneous User Experience and Usability

Experience diverged: younger or more digitally literate patients navigated smoothly, whereas older patients struggled with login, verification, account linking, scanning steps, bandwidth, and device limitations. Physicians cited slow loading, complex workflows, suboptimal image quality, windowing controls, and heterogeneity across hospital platforms.

> ***Patient:*** *Some QR codes open directly; others require registration and verification—too many steps.*
>
> ***Patient:*** *If the network is poor, I just give up.*
>
> ***Physician:*** *When the clinic network slows, image loading disrupts the flow of consultation.*
>
> ***Physician:*** *Different platforms work differently; the learning burden prevents a stable routine.*

Theme 3: Limited Retention and Interoperability Gaps

Short retention periods, expired links or QR codes, download limits, and weak cross-institution interoperability were common pain points. Patients could not retrieve older studies for long-term surveillance; physicians faced siloed platforms and resorted to repeat testing.

> ***Patient:*** *Links stop working after a year or two, making longitudinal comparison difficult.*
>
> ***Patient:*** *Limits on how many times the link can be opened are especially inconvenient.*

XSL•FO
RenderX

> ***Physician:*** *Platform fragmentation leads to information silos and, inevitably, duplicate imaging.*
>
> ***Physician:*** *Without unified interfaces, retrieving outside images is inefficient.*

Theme 4: Psychological Effects and Communication

CMI had bidirectional psychological effects: some patients felt reassured by ready access; others became anxious due to misinterpretation. Physicians noted that clear explanation is essential to avoid overinterpretation and undue worry.

> ***Patient:*** *Seeing my images makes me feel more in control.*
>
> ***Patient:*** *The terminology is confusing; the more I look, the more anxious I feel.*
>
> ***Physician:*** *Without explanations, patients may overread images and worry unnecessarily.*

Theme 5: Improvement Priorities and Governance

Patients and physicians broadly agreed on priorities but with different emphases. Patients wanted longer retention, removal of access limits, simpler flows, and cross-institution sharing. Physicians emphasized speed and clarity, unified standards or platforms, integration of history and reports, clear privacy and accountability, and gradual introduction of artificial intelligence–assisted localization and comparison.

> ***Patient:*** *Ideally, images would be stored long-term without repeated registration.*
>
> ***Physician:*** *We need unified data standards and interfaces, integrating imaging, reports, and history into one workflow.*
>
> ***Physician:*** *Baseline requirements are clarity, loading performance, stability, and privacy; these should be solved first.*

Discussion

Principal Findings

This mixed methods study shows that most hospitals can provide CMI in varied formats, yet a substantial possession-to-use gap persists. Compared with nonusers, CMI users engaged more with internet hospitals and physicians, reported lower care-related and travel or lodging costs, and self-reported higher disease knowledge. Qualitative findings pinpointed system usability, retention, timeliness, and interoperability as principal limitations. Overall, our results align with recent evidence on the accessibility and cost advantages of internet hospitals [6] and the association between interoperability and reduced duplicate imaging [5], offering practical directions for improving workflows.

Older age was independently associated with lower CMI use, whereas education was not significant after adjustment. This aligns with reviews showing that older adults face barriers with digital health tools—account/identity verification complexity, dense information, jargon, input burden, and screen design can all hinder adoption and sustained use [15,16]. Our interviews echoed concrete obstacles, such as multiplatform authentication, link expiry, and loading instability. For this population, the most effective improvements are those that reduce usability barriers rather than simply expanding access. Examples include single sign-on with unified entry points, clinic-side preloading combined with direct viewer access, caregiver proxy accounts, and concise instructional micro-guides, all of which have been shown to enhance eHealth participation among older adults [17-19].

Quantitatively, users interacted with more internet hospitals and physicians, while the number of offline hospitals did not change, suggesting that CMI may improve the efficiency and quality of care pathways rather than simply reducing in-person visits. Lower patient-reported costs for users are consistent with the reduced service frequency and improved accessibility observed with internet hospitals [6]. From a continuity-of-information perspective, integrated longitudinal viewers and HIE platforms reduce repeat imaging and patient burden [5], aligning with our cost-direction findings. Although users reported greater disease knowledge, perceived severity and anxiety did not differ, echoing prior work that immediate online access can increase knowledge but may sustain “scanxiety” in the absence of explanatory support [20]. Qualitative findings further contextualize these patterns by illustrating how CMI facilitates more efficient remote consultations and transforms surveillance into a continuous trajectory. These insights help interpret the observed cost reductions as a direct result of fewer unnecessary repeat scans and improved clinical efficiency during cross-institutional visits. The qualitative theme of limited retention and poor interoperability may also help explain the possession-to-use gap observed in our quantitative results, that even when patients have access to CMI, short retention periods and cross-institution barriers can reduce the practical usefulness of CMI and discourage actual use. Finally, the qualitative theme of bidirectional psychological effects explains why increased knowledge may not lower anxiety, as patients reported feeling more in control while simultaneously feeling confused by complex terminology in the absence of professional translation.

Stratification by access modality indicated advantages for DICOM over QR or link access in engagement and knowledge—plausibly due to DICOM’s interactivity, longer-term retention, and standardization, which facilitate time-series comparison and remote consultation. Technically, mature DICOMweb workflows support browser-based import and zero-footprint viewing suited to both lightweight patient access and professional clinical review [9]. At a systems level, better regional informatics and interoperability correlate with improved health service performance [21].

Our interviews highlighted short retention periods, inconsistent rules, and fragile links and QR codes as major obstacles, which undermine in-clinic retrieval. Consistent with research on patient portals, stable and renewable access points with clear audit trails and proactive expiry reminders are foundational for practical use, especially among older adults [22]. Clinically, immediate simultaneous access to prior images during the visit is critical. Rather than adding new entry points, default patient-centered retrieval organized by patient, chronological timeline, and imaging series, together with online comparison, should be standard capabilities. Workflows based on DICOMweb and zero footprint viewers already enable standardized retrieval and

XSL•FO
RenderX

reliable online comparison without local installation [8], and access to prior images reliably reduces repeat testing and delays [23].

Our findings suggest that interoperability and durable access in CMI are not only technical issues but also governance issues that require consistent rules for minimum retention periods, renewable access, audit trails, and accountability across institutions [24]. China's national health informatization agenda has emphasized standardization and interoperability maturity certification, which provides an actionable foundation to operationalize imaging-specific retention and exchange requirements at scale [25]. Building on recent analyses that identify standards, interoperability, legislation, and compliance as the core foundations of China's digital health governance for the Healthy China 2030 initiative, the clinical management pathway for pulmonary nodule surveillance can be strengthened by aligning platform requirements with these governance elements [24,26]. Finally, equity-oriented implementation is needed to ensure that national digital health strategies translate into broadly accessible longitudinal imaging services rather than widening existing digital divides [26].

Limitations

This cross-sectional design precludes causal inference. Visit counts, costs, and disease perceptions were self-reported and thus subject to recall and reporting bias. Recruitment through online and hospital settings may have overrepresented digitally literate individuals, limiting external validity. Consequently, the 57.6% usage rate likely represents an upper bound for this population, and the lack of regional stratification may mask even steeper barriers in underconnected or rural areas. Therefore, our findings are most applicable to digitally connected patients with PNs who engage in cross-institutional care and remote consultation and should be extrapolated cautiously to offline-only pathways and underserved settings. Although propensity score matching mitigated confounding, residual confounding cannot be excluded. The measure of CMI use does not distinguish passive access from more active use, and we did not capture frequency or depth of use. The qualitative sample was modest and did not include participant checking. Future studies could improve construct validity by incorporating intensity indicators ideally complemented by objective usage logs where feasible. Future multisite sampling, longitudinal, or quasi-experimental studies are also warranted to validate these findings and assess long-term adoption trajectories.

Conclusions

In PN management, CMI is widely available yet remains underused. Actual use was associated with greater engagement in online care, lower related costs, and higher self-reported disease knowledge. Qualitative findings identified 3 actionable constraints—usability thresholds, unstable data retention, and inconsistent clinical viewing functions. Addressing these limitations requires ensuring renewable, auditable, and stable access, together with default support for longitudinal, in-browser comparison.

Acknowledgments

We acknowledge that ChatGPT (OpenAI) was used to assist in language editing and improving the clarity of the manuscript.

Funding

This study was supported by the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2024ZD0529400 & 2024ZD0529401) and National Key R&D Program of China (2022YFE0133100).

Data Availability

The datasets generated or analyzed during this study are not publicly available due to privacy restrictions and institutional policies on patient data sharing but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: GL, GQ
Formal analysis: JH, JW, WZ, ZL
Funding acquisition: GQ
Investigation: CL, SX, YC, ZC
Methodology: GL, GQ, WZ, ZC
Software: JW, ZC
Supervision: GL, GQ
Resources: GQ
Visualization: WZ, ZC
Writing – original draft: YC, ZC
Writing – review & editing: GL, GQ, JW
All authors had full access to all of the data (including statistical reports and tables) in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
Author GL is the co-corresponding author (phone: 86 18680494527; email: 152537065@qq.com)

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guides.
[DOCX File, 20 KB - jmir_v28i1e86745_app1.docx]

Multimedia Appendix 2
Baseline characteristics for 701 patients in pulmonary nodules (PNs) after propensity score matching (PSM).
[DOCX File, 18 KB - jmir_v28i1e86745_app2.docx]

Multimedia Appendix 3
Covariate balance before and after propensity score matching.
[DOCX File, 29 KB - jmir_v28i1e86745_app3.docx]

Multimedia Appendix 4
Joint display of integrated quantitative and qualitative findings.
[DOCX File, 19 KB - jmir_v28i1e86745_app4.docx]

Checklist 1
COREQ checklist.
[PDF File, 380 KB - jmir_v28i1e86745_app5.pdf]

Checklist 2
STROBE checklist.
[DOC File, 84 KB - jmir_v28i1e86745_app6.doc]

References

1. Gould MK, Tang T, Liu ILA, et al. Recent trends in the identification of incidental pulmonary nodules. Am J Respir Crit Care Med 2015 Nov 15;192(10):1208-1214. [doi: 10.1164/rccm.201505-0990OC] [Medline: 26214244]
2. MacMahon H, Naidich DP, Goo JM, et al. Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017. Radiology 2017 Jul;284(1):228-243. [doi: 10.1148/radiol.2017161659] [Medline: 28240562]
3. Wood DE, Kazerooni EA, Baum SL, et al. Lung cancer screening, version 3.2018, NCCN clinical practice guidelines in oncology. J Natl Compr Canc Netw 2018 Apr;16(4):412-441. [doi: 10.6004/jnccn.2018.0020] [Medline: 29632061]
4. Jung HY, Vest JR, Unruh MA, Kern LM, Kaushal R, HITEC Investigators. Use of health information exchange and repeat imaging costs. J Am Coll Radiol 2015 Dec;12(12 Pt B):1364-1370. [doi: 10.1016/j.jacr.2015.09.010] [Medline: 26614881]
5. Yuan Y, Price M, Schmidt DF, Ward M, Nebeker J, Pizer S. Integrated health record viewers and reduction in duplicate medical imaging: retrospective observational analysis. JMIR Med Inform 2022 May 20;10(5):e32168. [doi: 10.2196/32168] [Medline: 35594070]
6. Liu Y, Jin H, Yu Z, Tong Y. Impact of internet hospital consultations on outpatient visits and expenses: quasi-experimental study. J Med Internet Res 2024 Nov 11;26:e57609. [doi: 10.2196/57609] [Medline: 39527807]
7. Wang K, Zou W, Lai Y, et al. Accessibility, cost, and quality of an online regular follow-up visit service at an internet hospital in China: mixed methods study. J Med Internet Res 2024 Oct 21;26:e54902. [doi: 10.2196/54902] [Medline: 39432365]
8. Genereaux BW, Dennison DK, Ho K, et al. DICOMweb™: background and application of the web standard for medical imaging. J Digit Imaging 2018 Jun;31(3):321-326. [doi: 10.1007/s10278-018-0073-z] [Medline: 29748852]
9. Pereira H, Romero L, Miguel Faria P. Web-based DICOM viewers: a survey and a performance classification. J Imaging Inform Med 2025 Jun;38(3):1304-1322. [doi: 10.1007/s10278-024-01216-5] [Medline: 39349783]
10. Norris EC, Halaska C, Sachs PB, Lin CT, Sanfilippo K, Honce JM. Understanding patient experiences, opinions, and actions taken after viewing their own radiology images online: web-based survey. JMIR Form Res 2022 Apr 25;6(4):e29496. [doi: 10.2196/29496] [Medline: 35468086]
11. Wang J, Goldberg JE, Block T, et al. Patterns of access to radiology reports and images through a patient portal. J Imaging Inform Med 2024 Apr;37(2):504-509. [doi: 10.1007/s10278-024-00996-0] [Medline: 38315344]
12. van der Mee FAM, Ottenheijm RPG, Gentry EGS, et al. The impact of different radiology report formats on patient information processing: a systematic review. Eur Radiol 2025 May;35(5):2644-2657. [doi: 10.1007/s00330-024-11165-w] [Medline: 39545980]

XSL•FO
RenderX

13. Chau M. Alone with the diagnosis: a reflective analysis on imaging report access and emotional burden. Radiography (Lond) 2025 Dec;31 Suppl 2:103104. [doi: 10.1016/j.radi.2025.103104] [Medline: 40764216]
14. HaoDF Online [Website in Simplified Chinese]. URL: https://www.haodf.com [accessed 2026-03-11]
15. Ahmad NA, Mat Ludin AF, Shahar S, Mohd Noah SA, Mohd Tohit N. Willingness, perceived barriers and motivators in adopting mobile applications for health-related interventions among older adults: a scoping review. BMJ Open 2022 Mar 9;12(3):e054561. [doi: 10.1136/bmjopen-2021-054561] [Medline: 35264349]
16. Lemos M, Henriques AR, Lopes DG, et al. Usability and utility of a mobile app to deliver health-related content to an older adult population: pilot noncontrolled quasi-experimental study. JMIR Form Res 2024 May 17;8:e46151. [doi: 10.2196/46151] [Medline: 38758585]
17. Grossman LV, Masterson Creber RM, Benda NC, Wright D, Vawdrey DK, Ancker JS. Interventions to increase patient portal use in vulnerable populations: a systematic review. J Am Med Inform Assoc 2019 Aug 1;26(8-9):855-870. [doi: 10.1093/jamia/ocz023] [Medline: 30958532]
18. Turner K, Clary A, Hong YR, Alishahi Tabriz A, Shea CM. Patient portal barriers and group differences: cross-sectional national survey study. J Med Internet Res 2020 Sep 17;22(9):e18870. [doi: 10.2196/18870] [Medline: 32940620]
19. Wolff JL, Dukhanin V, Burgdorf JG, DesRoches CM. Shared access to patient portals for older adults: implications for privacy and digital health equity. JMIR Aging 2022 May 4;5(2):e34628. [doi: 10.2196/34628] [Medline: 35507405]
20. Derry-Vick HM, Heathcote LC, Glesby N, et al. Scanxiety among adults with cancer: a scoping review to guide research and interventions. Cancers (Basel) 2023 Feb 22;15(5):1381. [doi: 10.3390/cancers15051381] [Medline: 36900174]
21. Ye A, Deng Y, Li X, Shao G. The impact of informatization development on healthcare services in China. Sci Rep 2024 Dec 28;14(1):31041. [doi: 10.1038/s41598-024-82268-z] [Medline: 39730846]
22. Son EH, Nahm ES. Adult patients' experiences of using a patient portal with a focus on perceived benefits and difficulties, and perceptions on privacy and security: qualitative descriptive study. JMIR Hum Factors 2023 Jul 25;10:e46044. [doi: 10.2196/46044] [Medline: 37490316]
23. Bailey JE, Pope RA, Elliott EC, Wan JY, Waters TM, Frisse ME. Health information exchange reduces repeated diagnostic imaging for back pain. Ann Emerg Med 2013 Jul;62(1):16-24. [doi: 10.1016/j.annemergmed.2013.01.006] [Medline: 23465552]
24. Wang M, Lu X, Du Y, et al. Digital health governance in China by a whole-of-society approach. NPJ Digit Med 2025 Aug 2;8(1):496. [doi: 10.1038/s41746-025-01876-9] [Medline: 40750663]
25. Li C, Xu X, Zhou G, et al. Implementation of national health informatization in China: survey about the status quo. JMIR Med Inform 2019 Feb 21;7(1):e12238. [doi: 10.2196/12238] [Medline: 30789350]
26. Wu G, Gong M, Wu Y, Liu L, Shi B, Zeng Z. Advancing digital health in China: aligning challenges, opportunities, and solutions with the Global Initiative on Digital Health (GIDH). Health Care Sci 2024 Oct;3(5):365-369. [doi: 10.1002/hcs2.118] [Medline: 39479272]

Abbreviations

CMI: cloud-based medical imaging
CNY: Chinese Yuan
DICOM: Digital Imaging and Communications in Medicine
HIE: health information exchange
PN: pulmonary nodule
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Stone; submitted 29.Oct.2025; peer-reviewed by O Akinpeloye, T Jimoh; revised version received 25.Feb.2026; accepted 25.Feb.2026; published 30.Mar.2026.

Please cite as:
Chen Z, Wu J, Chen Y, Zhuang W, Huang J, Zhong W, Li Z, Xie S, Liu C, Lu G, Qiao G
Use, Utility, and User Experience of Cloud-Based Medical Imaging in Pulmonary Nodule Care in China: Mixed Methods Study
J Med Internet Res 2026;28:e86745
URL: https://www.jmir.org/2026/1/e86745
doi:10.2196/86745

© Ziguo Chen, Junhan Wu, Yunying Chen, Weitao Zhuang, Jiaxuan Huang, Weifeng Zhong, Zijie Li, Shuhuan Xie, Chaofan Liu, Guojie Lu, Guibin Qiao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Perspectives on Health Data Sharing Among Patients With Somatic and Mental Health Diseases: Focus Group Study

Sabrina Fesl[1], MSc; Caroline Lang[1], MPH, Dr rer med; Falk Gerrik Verhees[2], Dr med; Jochen Schmitt[1], MPH, Prof Dr Med; Madlen Scheibe[1], Dr rer med

[1]Center for Evidence-Based Healthcare, Medical Faculty and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany

[2]Department of Psychiatry and Psychotherapy, University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany

Corresponding Author:
Sabrina Fesl, MSc
Center for Evidence-Based Healthcare
Medical Faculty and University Hospital Carl Gustav Carus
TUD Dresden University of Technology
Fetscherstraße 74
Dresden, 01307
Germany
Phone: 49 351458 ext 2651
Email: Sabrina.Fesl@ukdd.de

Abstract

Background: The German Health Data Utilization Act and the Digital Act aim to enhance health data sharing for health care and research in Germany and beyond while ensuring robust data protection. A key prerequisite is patients' willingness to share their data for primary use (PU), such as medical care, and secondary use (SU), such as research. There is a lack of qualitative research examining patients' perspectives on data sharing under the new legal framework, especially among vulnerable groups, such as those with mental health diseases.

Objective: This study qualitatively examines the factors influencing German patients' willingness to share their digital health data for PU and SU, exploring similarities and differences between patients with somatic and mental health diseases.

Methods: In 2024, we conducted 2 focus groups (FGs) with 13 outpatients: 7 with somatic diseases (FG1) and 6 with mental health diseases (FG2). Participants were recruited from a University Hospital in Dresden, Germany, based on predefined criteria. Discussions followed a topic guide with open-ended questions informed by an overview of reviews and pretests. Data were analyzed independently by 2 researchers using Kuckartz's approach. Findings are reported according to the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.

Results: A total of 10 main categories with 32 subcategories were identified as influencing factors: previous data-sharing experience, individual usefulness for medical care, public benefit, personal and privacy concerns, data security concerns, consent management preferences, technical safety measures, legal and ethical framework conditions and requirements, informational self-determination, and social involvement and influence. Both FGs highlighted individual usefulness and public benefit despite various personal experiences. Concerns about discrimination, stigmatization, and automatic data sharing were more relevant in FG2. Technical safety measures of anonymization and pseudonymization were discussed in detail in FG1, whereas FG2 debated data protection intensively. There were concerns that data protection in Germany could potentially pose a greater health risk than the sharing of personal health data. The category consent management preferences yielded the most statements, but no clear consensus emerged. Social influence and involvement, including family, peers, and health care professionals, were more relevant in FG2. Both FGs explicitly opposed the use of health data by companies such as Google.

Conclusions: This study qualitatively compared the perspectives of patients with somatic and mental health diseases. While it revealed similarities, patients with mental health diseases viewed their data as highly sensitive due to experiences of stigmatization and fear of misuse, emphasizing the need for tailored consent management. Involving family, peers, and health care professionals can increase acceptance. Health care professionals and targeted outreach can ensure transparency, raising awareness about data sharing policies to build trust, especially when commercial interests are involved. Knowledge deficits, even among tech-savvy patients, indicate the need for broad and understandable public relations efforts.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e79990) doi:10.2196/79990

KEYWORDS

consent form; data sharing; electronic health records; focus groups; mental health; patient participation; qualitative research; routinely collected health data

Introduction

Background

The German Health Data Utilization Act (Gesundheitsdatennutzungsgsetz [GDNG]) and the Act to Accelerate the Digitization of the Healthcare System (Digital-Gesetz; Digital Act), both enacted in March 2024, have created new legal regulations for using and sharing personal health data in the German health care system. The former is primarily intended to enable the use of health data for research and to facilitate its usability for purposes oriented toward the common good [1]. The latter includes the introduction of the electronic patient file. Thus, it explicitly promotes improved digital data exchange in the German health care system [2] and promises to integrate clinical data of patients across different health care institutions and throughout their lives [3]. The GDNG and the introduction of the electronic patient file as an opt-out solution are exemplary innovations that are not only relevant for health care providers and researchers but also bring opportunities and challenges for patients. In particular, the new measures aim to improve the overall quality of health care by ensuring data access for health care and research in accordance with the respective legal regulations for primary use (PU) and secondary use (SU) [4-7].

Personal health data refers to an individual's health status and information that are uniquely identifiable to them, such as diagnoses or medication plans. It is collected during care delivery in the health care system and can be obtained from existing sources, such as electronic health records [8]. The PU of health data refers to its application to directly provide medical care to patients, such as by sharing personal health data with health care providers [9,10]. In contrast, the SU of health data extends beyond the primary goal of individual treatment. It involves the use of pseudonymized or anonymized health data for purposes other than direct patient care, including activities such as commercial and noncommercial research, public health monitoring, and developing health care policies [9,10].

Germany is pursuing a deliberate, data protection–focused approach compared with other countries' more centralized or incentive-based health care systems. Unlike the first national centralized health data access bodies in France in 2019 (Health Data Hub) and in Finland in 2020 (Findata), Germany has launched decentralized initiatives, such as the German Medical Informatics Initiative (Medizininformatik-Initiative) in 2016 and the German Health Data Lab (Forschungsdatenzentrum Gesundheit) in 2020. The latter started to process the first requests for data use in 2025. These initiatives aim to link clinical data and medical research data to promote medical research and thus improve health care for all [11-13].

As in other European countries, patients in Germany value information, transparency, consent, and control regarding data sharing. However, Germany's decentralized system and national federal laws such as the GDNG can make it challenging to meet these expectations consistently in all regions [14,15]. Additionally, these different framework conditions and strict German requirements for data exchange can be a challenge for the research community because there is no single, straightforward opt-out regime for all secondary research use. As a result, this can hamper collaborative research efforts, slowing progress in areas such as disease detection and treatment optimization, and reducing data usability. Overall, these challenges can lead to suboptimal individualized medical care [16,17].

Patients' willingness to provide their health data for PU and SU is a highly topical issue for science, research, and public health care [14,18-21]. Data sharing can improve personalized care through closer cooperation between treating health care professionals and thus contribute to optimizing diagnostic and therapeutic outcomes [18]. Patients benefit from reduced treatment delays and medical errors, as well as more effective, efficient, and personalized care through access to comprehensive patient data [18,19].

In recent years, qualitative research has explored willingness to share data among vulnerable patient groups, such as patients with mental health diseases, to determine their needs, preferences, and concerns, which may differ from those of patients without mental health diseases [22-28]. These differences may require tailored approaches to promoting the responsible sharing of personal health data. While patients with mental health diseases may face stigmatization in society and in health care, which can lead to mistrust of health care providers and make comprehensive data sharing difficult, this stigmatization is less evident for patients with somatic diseases [22,25,28-30]. Compared with other patient groups, patients with mental health diseases may feel that their data are more sensitive and could be misused to their detriment or for discrimination. In contrast, patients with somatic diseases may feel that their data are less personal and more directly linked to physical health outcomes [22,23,28,31,32].

Based on the outlined new regulatory structures in Germany for digital progress, it is important to include the perspectives of patients from different target groups. Recent surveys have shown that both patients with somatic and mental health diseases in Germany are cautious and reserved about data protection [33-35]. However, the regulation of opt-out, which requires active objection, is particularly relevant for patients with mental health diseases due to the above-described potentially different concerns regarding health data sharing. Deeper insights into their attitudes are therefore crucial.

To our knowledge, there are currently no qualitative studies examining and comparing the perspectives of German patients with somatic and mental health diseases regarding sharing their

XSL•FO
RenderX

personal health data for PU and SU, following the enactment of the new legal framework mentioned above.

Objective and Research Questions

This study reports a qualitative study comparing German patients with somatic and mental health diseases who were asked about their attitudes, preferences, and needs regarding their willingness to share their personal health data for PU and SU in 2 focus groups (FGs). It shows the similarities and differences in the perspectives of both groups of patients receiving outpatient treatment in Germany, considering the context of the new regulations.

The research questions of this study were (1) What factors influence German patients' willingness to share their digital health data for PU and SU? (2) What similarities and differences exist in data sharing perspectives between German patients with somatic and mental health diseases?

Methods

This study was conducted as part of the PATH (Personal Mastery of Health and Wellness Data) project in January 2024.

Study Design

Following our overview of reviews on factors influencing patients' willingness to share their digital health data for PU and SU [9], we conducted 2 FGs with German patients with somatic and mental health diseases. FGs are useful when individual motivations and perspectives need to be explored in depth [36]. Unlike standardized interviews or questionnaires, FGs are based on a topic guide with open-ended guiding questions. Participants are free to answer according to their subjective relevance, which is also a major advantage over standardized survey methods. The moderator also has the opportunity to follow up on individual aspects to explore them in more depth. In addition, the use of a topic guide allows the discussion to be steered thematically to maintain the focus of the discussion. Therefore, FGs are particularly suitable for gaining practical knowledge for implementing data sharing concepts.

The methods and results are reported according to the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist (Multimedia Appendix 1) [37].

Recruitment of Study Participants

Patients were recruited and enrolled at the University Hospital Dresden, Germany. To identify and recruit potential patients, physicians working in the outpatient departments of diabetology, neurology, dermatology, and psychiatry at the University Hospital Dresden were contacted in November 2023 by members of the research team in person and by email, explaining the project and planned details of the study.

Participants were selected from different clinical disciplines to obtain the most heterogeneous sample possible. Two FGs were formed, 1 with patients with somatic diseases (FG1) and another with patients with mental health diseases (FG2). Patients in both FGs were recruited by face-to-face and email with the support of project partners and staff in the respective outpatient clinics during their regular medical consultations. For this purpose, information flyers, a brief study information document, and a consent form were prepared for potential participants, containing the key details of the study (location with map, date, time, planned schedule, amount of compensation, and contact details for further questions). Participants for both FGs were recruited based on a predefined, purposive sampling plan and predefined criteria (Textbox 1).

Patients interested in participating were screened for eligibility, provided detailed information to establish a relationship prior to study commencement, and invited to attend the relevant FG by email. Consent forms were explained, and consent was obtained before the FG. Participants who attended the FG were offered financial compensation of €50 (a currency exchange rate of €1=US $1.16 is applicable).

Textbox 1. Inclusion criteria for participant recruitment of both focus groups.

Inclusion criteria

- Aged 18 years or older
- Daily smartphone and/or tablet use
- Capable and consenting
- Able to communicate in German to enable active participation in focus group (FG) discussions

FG1

- Patients with somatic diseases but without a history of mental health disease (self-reported or, where possible, diagnosed through screening results or existing diagnoses) who were treated in the respective outpatient clinics during the recruitment period.

FG2

- Currently euthymic or at least partially remitted patients with mild, moderate, or recurrent forms of mental health disease who were treated in the Psychiatric Outpatient Clinic at the University Hospital Dresden during the recruitment period.

Topic Guide

The topic guide for the FGs was developed based on our overview of reviews [9] on patients' attitudes toward health data sharing for PU and SU. Guiding open-ended questions and topics were used in both FGs to allow comparisons (Multimedia Appendix 2). The list of topics (Figure 1) was presented as a word cloud on Microsoft PowerPoint slides to stimulate

XSL•FO
RenderX

discussion 3 times in each FG: for health data sharing with health care professionals (PU), with research (SU), and with private companies (SU). The participants could choose 1 or more topics according to their own priorities, focus, and interests.

Figure 1. Topic selection and asked scenarios for both focus group discussions.

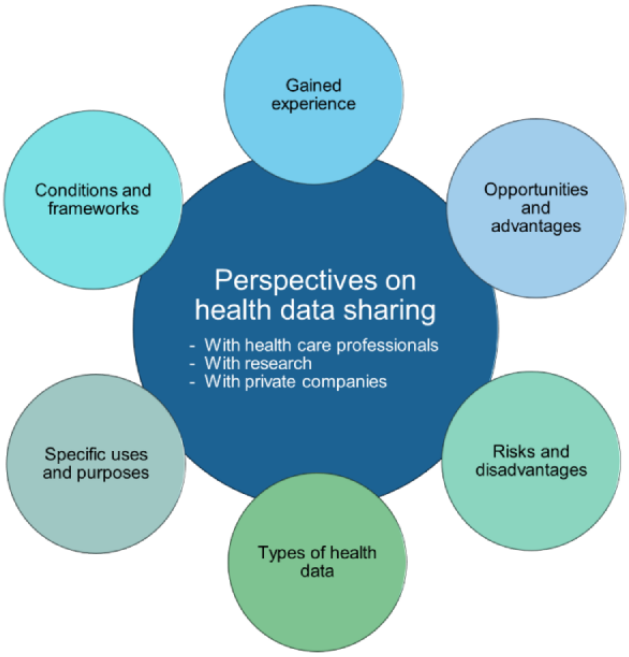

Cognitive Pretests

Cognitive pretests were used to test and adapt the interview guide to ensure that all questions and items are understood and interpreted as intended and that no relevant aspects were missing. They also assessed whether the time required and the number of questions were acceptable [38]. The cognitive pretests were performed in October 2023 as individual online interviews with 7 participants with different diseases according to the predefined inclusion criteria (Textbox 1). The duration of the interviews ranged from 36 to 58 minutes. The pretests provided relevant insights and comments concerning the formulation of questions, comprehension of terms (eg, types of data), and conduct of the FGs.

Procedure

The FGs were held on 2 evenings in January 2024 in an intimate setting of a reserved room at the University Hospital Dresden. Both were audio recorded. Three research associates with experience in qualitative research took minutes (SF, MSc; female, and MS, Dr; female, for FG1; and SF and CL, Dr; female, for FG2). Another research associate and experienced moderator (Dr; female), who was not involved in the project and had been briefed in advance, used the topic guide to lead the participants through both discussions. A resident psychiatrist (FGV, Dr; male) also took part in FG2 (patients with mental health diseases) to ensure that participants felt confident. One patient's relative was present in each FG due to private circumstances that made this necessary at short notice. All participants agreed that the relatives could remain present; however, they did not participate in the discussion. A short presentation was provided at the beginning of each session to introduce the project goal and the researchers who were present. However, no aspects relevant to the discussion were raised to avoid influencing the participants. At the end of the discussion, the participants were given a short sociodemographic questionnaire and a validated questionnaire [39] to determine their technology commitment score.

Data Analysis

The content of the FG discussions was transcribed verbatim, checked for accuracy, adjusted according to the rules of transcription [40], and all personal or personally identifiable information was removed. The data were analyzed using MAXQDA software (version 2018; Verbi Software GmbH) and qualitative content analysis according to Kuckartz and Rädiker's [41] approach, in which the data are coded and evaluated according to content and thematic aspects in more detail than with Mayring's [42] strongly theory-driven approach. This approach also enables the combination of deductive and inductive category development.

First, a category system was deductively developed based on the discussion guide and our overview of reviews [9], which was applied as a frame to the transcribed material. The category system was modified and refined as the material was reviewed in more detail, as well as through the subsequent inductive formation of recurring major categories into additional categories. Two research team members (SF and CL) independently coded the material, continuously compared emerging themes and codes, discussed data saturation through an iterative process of data collection and analysis, and checked intercoder reliability. With the help of added quotes as anchor examples, the coders could optimally match participants' responses to the defined categories.

Data saturation was reached when additional data no longer produced new categories or insights, indicating that the breadth and depth of relevant information were adequately captured.

XSL•FO
RenderX

To confirm saturation, MAXQDA visualization tools were used to track the frequency and recurrence of codings across datasets. Furthermore, peer discussions were conducted to validate that no significant new information appeared in the later stages of coding. This approach ensured that the coding scheme was comprehensive and that the study findings were robust and grounded in the data. This transparent and systematic approach to assessing data saturation strengthened the credibility and trustworthiness of the qualitative findings. Moreover, the research team collaboratively developed and continually refined the category system through iterative discussions to ensure that participants' diverse perspectives were accurately represented. Figure 2 illustrates the data analysis process and the sequence of activities.

A codebook with definitions, respective anchor examples, and the number and context of statements used for qualitative coding and analysis in MAXQDA is provided in Multimedia Appendix 3.

Figure 2. Flowchart showing each step of the data analysis process and the sequence of activities (own illustration based on Kuckartz and Rädiker [41]).

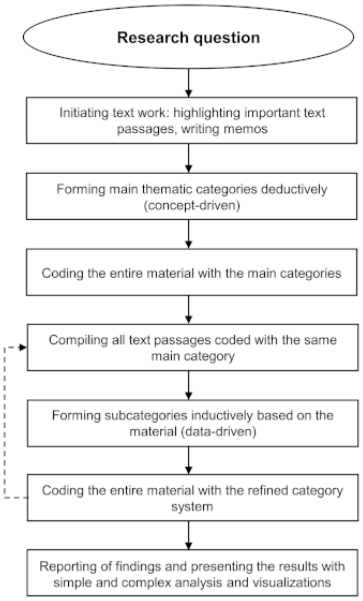

Ethical Considerations

The study was approved on November 6, 2023, by the Ethics Committee of Technische Universität Dresden (reference BO-EK-310072023). All procedures were in accordance with the ethical standards of the institutional ethics committee and with the Declaration of Helsinki (2024 revision). Persons with mental health conditions, considered a vulnerable population, were included in accordance with the 2024 Declaration of Helsinki (paragraphs 19-20). Additional safeguards were implemented, including assessment of decision-making capacity prior to informed consent and a resident psychiatrist (FGV) acting as a supporting person where needed, as approved by the ethics committee. Participants were informed about the purpose of the study, the voluntary nature of participation, and their right to withdraw at any time without penalty. Participants received compensation of €50 for their time and travel expenses related to attending the FG discussions. This incentive was preapproved by the ethics committee. Written informed consent was obtained from all participants who attended the FG discussions. Confidentiality was assured in compliance with the EU-GDPR (EU General Data Protection Regulation). Only researchers involved in the data analysis had access to the pseudonymized data, which were stored securely on institution-protected servers and will be deleted 10 years after study completion. Pseudonyms were used in the transcripts to protect participants' identities.

Results

Sample Characteristics

Sample Sizes and Diseases

FG1 comprised 7 participants, of whom 3 had type 2 diabetes, 2 had a neurological disease, and 2 had a dermatological disease. The discussion lasted approximately 82 minutes. FG2 comprised 6 participants, of whom 3 had bipolar disorder, 2 had paranoid schizophrenia, and 1 had schizoaffective disorder (all currently in remission or with residual symptoms). The discussion lasted approximately 78 minutes. Because recruitment took place as part of regular daily care, it was not possible to provide a figure on refusals. A total of 2 patients dropped out on the day of the event due to acute illness.

Sociodemographic Characteristics

All participants' sociodemographic characteristics (Multimedia Appendix 4) were extracted from the questionnaire provided. Regarding education, most had completed an apprenticeship (7/13, 54%). Regarding residence, most lived in large cities (9/13, 69%), followed by a medium-sized city (3/13, 23%). Regarding income, most participants in FG2 (4/13, 67%) had a monthly household income of <€1000-€1500, whereas most participants in FG1 (5/13, 71%) had a monthly household income of €3001-€3500 or €4001-€5000.

XSL•FO
RenderX

Technology Commitment

The total scores for the self-rated technology commitment scale and its subdomains of technology acceptance, technology competence beliefs, and technology control beliefs were high. Out of a maximum possible score of 60, the mean total score was 50 (range 41-57) for patients with somatic diseases and 45 (range 38-56) for patients with mental health diseases. The scores for each participant and FG are shown in Multimedia Appendix 5.

Categories

Overview

We identified 10 overarching main categories (I-X) with 32 corresponding subcategories (Table 1). These main categories were defined to capture overarching themes relevant to the research question, whereas the 32 subcategories were developed to reflect more specific aspects and nuances within these broad themes. This approach allowed for a detailed and systematic structuring of the data, enabling the identification of patterns, differences, and relationships more precisely. The hierarchical organization enhanced the clarity and transparency of the analysis, facilitating an in-depth understanding of the participants' perspectives. The categories in the Results section and in Table 1 were sorted according to the topic guide and participants' response behavior. The complete codebook with definitions, respective anchor examples, and the number and context of statements used for qualitative coding and analysis in MAXQDA is provided in Multimedia Appendix 3. Each main category included in the final category system was discussed in both FGs, but with different relevance in each FG and emphasis by individual participants. Table 1 presents an overview on the number of statements per main category and subcategory in both FGs. While most statements were about consent management preferences, few were related to social involvement and influence, with only 1 comment in FG1.

XSL•FO
RenderX

Table 1. Main categories and corresponding first-level subcategories with total numbers of statements and number of statements per main category and subcategory in both focus groups (FGs).

Main categories and corresponding subcategories	Total statements, n	FG1 statements, n	FG2 statements, n
Previous experience with data sharing	38	17	21
Health care professionals	18	10	8
Research and clinical studies	8	5	3
Employers and private companies	5	0	5
Health insurance companies	4	2	2
Medical education	3	0	3
Individual usefulness for medical care	32	16	16
Comprehensive documentation	12	5	7
Personal informedness	6	4	2
Emergency assistance	5	3	2
Simplification of processes	5	3	2
Time saving	4	1	3
Public benefit	28	15	13
Medical care	25	12	13
Cost saving	3	3	0
Personal and privacy concerns	34	15	19
Concerns about commercial interests	15	6	9
Concerns being discriminated or stigmatized	8	3	5
Personal uncertainties and skepticism	7	3	4
Concerns being overstrained	4	3	1
Data security concerns	34	18	16
Concerns about data access	24	11	13
Concerns about data processing	10	7	3
Consent management preferences	128	62	66
Consent scope	87	41	46
Consent model	41	21	20
Technical safety measures	37	27	10
Anonymization and pseudonymization	22	17	5
Central data repository	8	6	2
Data encryption	7	4	3
Legal and ethical framework conditions	35	24	11
Data protection	11	3	8
Standardization and regulation	10	10	0
Ethical justifiability	7	4	3
Monitoring and authorization	7	7	0
Informational self-determination	33	11	22
Transparency	18	7	11
Autonomy and control	8	1	7
Trust	7	3	4
Social involvement and influence	9	1	8
Family and peers	5	1	4

XSL•FO
RenderX

Main categories and corresponding subcategories	Total statements, n	FG1 statements, n	FG2 statements, n
Health care professionals	4	0	4

Key Findings

In both FGs, all participants expressed widely varying personal experiences with health data sharing, mainly regarding PU. In contrast, only those in FG2 discussed their personal experiences with employers and private companies. Regarding PU and individual usefulness for medical care, the benefit of comprehensive documentation was considered most relevant in both FGs. Regarding SU and public benefit for medical care, the advancement of drug and product development and the commercial interests of health data sharing were discussed extensively in FG2. Fears of discrimination and stigmatization were more prevalent in FG2. These fears were particularly related to disadvantages or dismissal in the workplace. Mental health diagnoses were also identified as highly sensitive in FG1. In addition, data security concerns, especially misuse, were raised more frequently in FG2 in response to the question about automatic data sharing.

It was not possible to clearly delineate participants' views on the consent model. However, it became evident that participants in FG1 expressed only a few and rather general conditions regarding their acceptance of an opt-out model, whereas participants in FG2 attached very strict and more differentiated conditions to their acceptance of an opt-out model. University research institutions were sometimes more trusted than private and pharmaceutical research institutions, and the intended use or research project was crucial for consent preference. Participants in both FGs explicitly opposed the use of health data by "big players" such as Amazon, Google, and Apple.

Overall, technical safety measures of anonymization, or at least pseudonymization, in terms of health data sharing for SU were debated more intensively in FG1. Regarding legal and ethical framework conditions, the aspect of data protection was discussed more broadly in FG2. Aspects of standardization, regulation, monitoring, and authorization were addressed only in FG1. As facilitating factors, autonomy and control appeared more relevant in FG2, both in the context of PU and SU. Social influence, involving family, peers, and health care professionals, became much more important for participants in FG2.

Below, the results for each main category are described, with additional details and representative quotes for the corresponding subcategories from both FGs. At the end of the Results section, an overview of the key points regarding patterns and relationships between influencing factors and sociodemographic characteristics is provided.

Main Category I: Previous Experience With Data Sharing

Within this main category, participants distinguished between PU and SU experiences. While most reported positive experiences in terms of PU, 2 participants in each group indicated that there were bureaucratic barriers to creating an electronic health record with their health insurance companies and that the exchange of health data with their health care professionals was still paper based. Concerning SU, experiences of participating in clinical trials were more common among participants in FG2 (Textbox 2).

Textbox 2. Exemplary quotes on health data sharing experiences.

Subcategory: health care professionals

- *So far, no doctor has been able to install anything for me; they're all not ready yet.* [FG1.1]
- *I've already done this myself, [...] I've already allowed some practitioners, my doctor, for example, to communicate.* [FG2.2]

Subcategory: research and clinical studies

- *I haven't taken part in any studies or anything like that.* [FG1.2]
- *I've also taken part in a few studies here.* [FG2.1]

Subcategory: employers and private companies

- *[…] Well, I'm not entirely sure that the diagnostic code was on the form that the employer received.* [FG2.5]

Subcategory: health insurance companies

- *[...] adopting the electronic patient file, at least that's quite complicated for me.* [FG1.7]
- *[...] my health insurance says we're still missing three documents, please submit them later. And it will be a long time before I can even go to rehab.* [FG2.4]

Subcategory: medical education

- *I had just been to a lecture with one of his superiors in December.* [FG2.4]

XSL•FO
RenderX

Main Category II: Individual Usefulness for Medical Care

A total of 5 subcategories emerged on individual usefulness for medical care when sharing health data. If a patient changes location or physician, the medical history is fully available to the next practitioner, which could prevent undue disease progression. Benefits were also perceived for obtaining a second medical opinion and to avoiding bringing paper copies of medical reports to the appointment. In addition, rapid assistance in an emergency and saving time for administrative formalities, such as pension procedures or medical history interviews, were discussed (Textbox 3).

Textbox 3. Exemplary quotes on individual usefulness for medical care.

Subcategory: comprehensive documentation

- *Because then it's basically my ID that I can present when I go to a new doctor.* [FG1.5]
- *If they had everything stored on the health card and the medical service, the doctors and everyone would had access to it straight away. That would make things much easier.* [FG2.4]

Subcategory: personal informedness

- *[...] that the documentation has become more open. [...] whereas today you are basically presented with complete data collections and can also [...] assess them yourself.* [FG1.3]
- *I have to say, I have a folder at home that is now so thick [...] that would of course be an advantage for me if it were completely stored on the health card.* [FG2.4]

Subcategory: emergency assistance

- *I can't imagine how this can be handled properly, especially in emergencies.* [FG1.7]
- *You always have to assume that if you suddenly find yourself in a situation like this, a doctor needs to know something. And you can't say that yourself.* [FG2.1]

Subcategory: simplification of processes

- *Because that would always simplify and accelerate many processes.* [FG1.2]
- *And I know that the doctor will only use them in a positive way, not against me, but so that my treatment can run smoothly.* [FG2.2]

Subcategory: time saving

- *[...] that would also shorten the anamnesis interview, which always has to be conducted. Then you can shorten it considerably, he can ask exactly the questions he needs to ask [...] I think it would be great if this could be done automatically in advance. That would be really useful.* [FG1.2]
- *I wouldn't need to talk about it for long; they would have it directly in the app or on the card.* [FG2.3]

Main Category III: Public Benefit

Regarding possible public benefits, drug and product development was discussed by 5 participants in FG2, which was linked to their gratitude that new drugs are being developed, meaning that treatment for patients with mental health diseases is improving. Regarding statistics and achieving gender-neutral, high-quality health care, there was agreement that health data should be collected automatically, especially if they are anonymized. The aspect of cost savings in health care by treating individuals well enough to enable them to remain active and able to work in the long term was raised by 1 participant in FG1 (Textbox 4).

XSL•FO
RenderX

Textbox 4. Exemplary quotes on public benefit.

Subcategory: medical care > drug and product development

- *Because I can imagine that such a secondary gain [...] if, for example, a hoover manufacturer now has the data and sees, okay, we have an increase in respiratory diseases in the population [...]. Then we have a larger market, they are no longer so expensive [...] and people with hyper allergenic diseases would also benefit from this at some point.* [FG1.2]
- *And research depends on the results of the patients. It's not possible otherwise. All the examinations have to be compared somehow.* [FG2.1]

Subcategory: medical care > gaining knowledge

- *There is also a great deal to be learned from the data collected.* [FG1.7]
- *Yes, that now, in a few years' time, there might be another patient who the doctors have helped by using my data, for example.* [FG2.3]

Subcategory: medical care > deriving preventive measures

- *Or even the health insurance company, because they look at preventive measures, that's okay too.* [FG1.6]
- *You can really use these neural networks to see how high the probability is that if the person has disease A and then has disease B, or just has the symptoms that you can then set up better drugs.* [FG2.6]

Subcategory: cost saving

- *And the biggest costs come from early retirement. Not the cost of medication [...] That would be an incredibly great contribution.* [FG1.6]

Main Category IV: Personal and Privacy Concerns

Concerns about commercial interests were present in both FGs. Skepticism about possible conclusions from anonymized health data or data sharing without prior information and the associated potential for discrimination or stigmatization was particularly expressed in FG2. Contrary to the benefits of comprehensive and transparent documentation, concerns about being overstrained by the large amount of data available were also expressed (Textbox 5).

Textbox 5. Exemplary quotes on personal and privacy concerns.

Subcategory: concerns about commercial interests

- *But I think that at least a not inconsiderable part of the research should be made available to the general public again. Because if the general public has contributed to the data, then the general public should not be ripped off again in the end.* [FG1.4]
- *I can imagine [...] Amazon says we also want to do research [...], and then they have such an offer for us, and earn money with it. Nowadays, data is the most important thing where you can really make money and where you also earn money.* [FG2.6]

Subcategory: concerns being discriminated or stigmatized

- *[...] health insurance company XY now wants to consider that it no longer wants to accept expensive patients in the near future.* [FG1.6]
- *I'm just going to be open and honest, due to the severity of my disease, I'm constantly afraid that somehow things could be spread about me and that people know me [...].* [FG2.2]

Subcategory: personal uncertainties and skepticism

- *Of course, I also have the information paper-based, and I get what I ask for. Although in some cases I have to ask for, such as this or that medical report. But I always get what I ask for. Simply saving the paper version is of relatively little use to the next practitioner. Because there really hasn't been enough development [in terms of digitalization]. At least not in the short time frame in which it should be in the end.* [FG1.7]
- *If my data is used for this, even though I have absolutely no idea about it, I would find that a bit creepy, yes.* [FG2.2]

Subcategory: concerns being overstrained

- *The disease is like this [...] I don't want to know everything that's bubbling and boiling inside me. I want to be able to live with the disease myself to some extent, and I push the other things aside.* [FG1.1]
- *That can also develop into a pretty big mental chaos [...], quite honestly.* [FG2.4]

Main Category V: Data Security Concerns

Regarding data security, participants differentiated between concerns about data access and data processing. It was particularly emphasized in FG2 that health data are generally very sensitive. The data must be protected against unauthorized access and changes and should not be used, for example, for research purposes without notification. Contrary to existing concerns about data security, there was a statement from FG1 that the high level of data protection in Germany may sometimes be riskier for health than sharing one's own health data (Textbox 6).

XSL•FO
RenderX

Textbox 6. Exemplary quotes on data security concerns.

Subcategory: concerns about data access

- *It gets bad when someone gets into the data collection and changes the blood group, for example.* [FG1.1]
- *So, basically, that's the fear that everyone has. Who can access all the health data?* [FG2.1]

Subcategory: concerns about data processing

- *[...] when I have my orthopedist, my oncologist, my ophthalmologist [...], I can find an individual person there.* [FG1.6]
- *But people would have absolutely no idea what happens to their data. They wouldn't even know what studies their data is being used for.* [FG2.2]

Main Category VI: Consent Management Preferences

Consent Scope

Regarding the scope of consent, a distinction was made between sensitive and nonsensitive health data. Communicable diseases, sexually transmitted diseases, mental health diseases, and health care professionals' subjective assessments of patients in open notes, such as on their behavior during treatment, were categorized as highly sensitive and consequently associated with a restriction or rejection of sharing. As one possible approach, the creation of different categories of health data was mentioned. Health data in anonymized or pseudonymized form were also considered particularly notable, and emergency data were discussed as a special case in FG2 only (Textbox 7).

Further subcategories were developed, differentiating consent scope by recipient and purpose and dividing statements related to PU (health care professionals) and SU (research, pharmaceutical, and private companies). Regarding PU, while some participants expressed the desire for limited data sharing only and the relevance of a complete dataset for an optimal, holistic treatment, others expressed agreement with full sharing. Opinions included limited selection with prior consent to the treating physician, sharing only with relevant treating physicians, refusal of automatic sharing, and emergency situations as an exception. In FG1, health insurance companies and company physicians were discussed as special recipients for whom exceptions should be possible to avoid disadvantages at work or regarding health insurance fees. Regarding SU, the private companies mentioned included suppliers of medical devices and medical technology (Textbox 8).

Textbox 7. Exemplary quotes on consent scope related to the type and format of information.

Subcategory: sensitive vs nonsensitive data

- *There are infectious diseases, for example. I wouldn't know whether I would authorize everyone to see them.* [FG1.5]
- *But now I'm asking myself, even in the context of psychotherapy, private historical data is told. For example, it could be a case of abuse. Can that also be health data?* [FG2.6]

Subcategory: anonymized and pseudonymized data

- *But really just the anonymized collection of the data. And I would agree to that without any reservations.* [FG1.7]

Subcategory: emergency data

- *I could imagine that, as I said, the important things are visible to everyone. That every doctor can look at my allergies, but by no means that every doctor gets the other data […].* [FG2.6]

XSL•FO
RenderX

Textbox 8. Exemplary quotes on consent scope related to the recipient and purpose.

Subcategory: health care professionals

- *In principle, I feel the same way about selection, as I would intuitively say that not every doctor should be able to see everything.* [FG1.4]
- *Well, on the whole, I wouldn't care […]. If this is going to be widespread in the medical field without my consent, it might make sense to have some sort of prioritization. In other words, doctors in private practice and hospitals would be allowed to access it and not, let's say, anyone who has anything to do with medicine.* [FG2.5]

Subcategory: research and pharmaceuticals

- *The only important thing is the purpose, i.e., medical research, whether it is university research or a pharmaceutical company, I wouldn't care. [...] The main research work is done by the pharmaceutical companies.* [FG1.7]
- *I also share my data if I can decide that. [...] I think that's something everyone should be able to decide. So, I don't think it is right that everyone should be able to extract something from it.* [FG2.5]

Subcategory: private companies

- *[...] I wouldn't make that much of a distinction between medical device manufacturers and other manufacturers.* [FG1.2]
- *I wouldn't give my data to just any company, especially if I don't know what they need the data for and what they want to do with it.* [FG2.2]

Consent Model

The consent model subcategory covers whether participants preferred an opt-in or opt-out solution, reflecting our guiding question on automatic health data sharing for PU and SU. By definition, active consent and permission to use and share health data, or explicit and flexible consent, were coded as opt-in. Standard consent with an active and explicit right to refuse the use and sharing of health data was coded as opt-out. The answers in both FGs were very heterogeneous, and no clear picture could be derived. However, it became apparent that patients in FG1 tended to mention only a few and rather general conditions for their agreement to an opt-out model, whereas patients with mental health diseases set strict and more differentiated conditions for their acceptance (Textbox 9).

Textbox 9. Exemplary quotes on the consent model.

Subcategory: opt-in

- *Specifically, which doctor gets access? And only that doctor will be authorized.* [FG1.3]
- *And I think it's very important that as part of the research [...] they are asked: 'would that be something for you? Would you take part?'* [FG2.4]

Subcategory: opt-out

- *So, with this option, I would be able to deselect something, i.e. deny someone access. But in general, if I don't say anything, I want to allow access.* [FG1.7]
- *I could imagine a system or principle similar to that of organ donation, where you say that all data is basically free and if I don't want it [...] then I actively disagree.* [FG2.5]

Main Category VII: Technical Safety Measures

Anonymization, or at least pseudonymization, serves as a facilitator of health data sharing for SU. A central data repository, such as the storage of anonymized health data on secure server platforms, preferably in Germany, or in a registry, would support controlled access for SU. Data encryption, possibly managed by entering a personal identification number, was considered relevant to the design of related software (Textbox 10).

XSL•FO
RenderX

Textbox 10. Exemplary quotes on technical safety measures.

Subcategory: anonymization and pseudonymization

- *Another issue is anonymized use, as is the case in the Scandinavian countries. A lot of insights can be gained from the collected data.* [FG1.7]
- *And if they actually collect individual data from me [...], it would have to be designed in such a way that you can no longer recognize which patient it is.* [FG2.2]

Subcategory: central data repository

- *So, I would like this intermediate instance [...] the data goes to a register. And I'm sure that not just anyone can access it [...] that it doesn't go to a company.* [FG1.6]
- *As I said, in the end it is stored somewhere on a large server, and the question is then with the image, I think that's a good idea.* [FG2.4]

Subcategory: data encryption

- *I think it would be very good if this is encoded on the insurance card and can also be read by my doctor after I have given my consent.* [FG1.3]
- *Exactly, and then the rest of the data, which can perhaps also be encrypted in parts. And when I go to the doctor, the doctor shows me what he wants to open and I enter a PIN.* [FG2.6]

Main Category VIII: Legal and Ethical Framework Conditions and Requirements

The most frequently mentioned aspect of this main category, especially among participants in FG2, was privacy. The importance of employers not having access to health data was especially highlighted. There was a call for effective data protection under the EU-GDPR. The paradox was also described, as some individuals reveal everything on social media but insist on high data protection when sharing health data. It would be important to create a harmonized legal framework, binding regulations, and a standardized digital data recording system valid throughout Germany, but this was viewed as hardly feasible. Participants in FG1 would be more willing to share their health data for SU if the appropriate and ethical use of health data were ensured through authorization procedures and monitoring by an independent committee (Textbox 11).

Textbox 11. Exemplary quotes on legal and ethical framework conditions and requirements.

Subcategory: data protection

- *The fact that any human resources director can then access it should be avoided.* [FG1.1]
- *And in this context, it is very important to me that my health data is very well protected and cannot be viewed by third parties, because I don't want so many people to find out about it.* [FG2.2]

Subcategory: standardization and regulation

- *And such regulations and procedures are needed to get things off the ground.* [FG1.1]

Subcategory: ethical justifiability

- *[...] they would still have to say at the beginning what they want to research it for. This ethical component would have to be clarified.* [FG1.6]
- *They wouldn't know at all which studies their data would be used for. And they have no way of deciding, even from an ethical point of view, whether this is good for them or not.* [FG2.2]

Subcategory: monitoring and authorization

- *The use of the data must, of course, also be justified and authorized by the Commission or something else to ensure that it is used sensibly.* [FG1.7]

Main Category IX: Informational Self-Determination

This main category included key aspects that enable individuals to make and act on their own decisions: transparency, trust, autonomy, and control. The latter was considered relevant in the context of PU and the self-determined decision of which physicians will see individual health data. In particular, participants in FG2 considered the implementation of robust control mechanisms important because of anticipated limited public trust in sharing health data. Regarding SU, autonomy should be respected so that, in principle, everyone can decide whether to share their health data for research. Participants also expressed interest in protocols for transparently reporting information to each individual about their contribution and support of research activities, possibly combined with a mandatory disclosure of research results (Textbox 12).

XSL•FO
RenderX

Textbox 12. Exemplary quotes on informational self-determination.

Subcategory: transparency

- *And I think the minimum would be to have complete transparency about how this system works.* [FG1.4]
- *Well, I would definitely want to know if the data is being used for research. I would also not object to the data being used for research purposes in principle.* [FG2.2]

Subcategory: autonomy and control

- *Of course, not everyone can do that. You have to have it in your own hands to authorize it, or not.* [FG1.1]
- *I also share my data if I can decide that. So, I'm open to that. As I said, I think that's something everyone should be able to decide.* [FG2.5]

Subcategory: trust

- *But to start with [...] you just have to assume a bit of trust, otherwise nothing will work.* [FG1.1]
- *Their intention is certainly to show that people trust research rather than companies [...] Whether it is Siemens or […] University.* [FG2.6]

Main Category X: Social Involvement and Influence

The topic of social influence was not addressed at all in FG1, apart from 1 statement on data sharing with the employer. In FG2, it was argued that family members or peers could also be involved in decision-making and that family members should also be allowed to release health information in an emergency. Health care professionals could play a supportive and consultative role, such as by providing information about data sovereignty and personal responsibility for the electronic medical record (Textbox 13).

Textbox 13. Exemplary quotes on social involvement and influence.

Subcategory: family and peers

- *Because, for example, I had just started at a company [...] with the diagnosis [...]. And there was even the recommendation from my family members [...]: 'Lie about it, don't say it'.* [FG1.6]
- *[...] or you know something like a password or a very long string and you can enter it, even as the person's husband or wife, and then also disclose the other data.* [FG2.6]

Subcategory: health care professionals

- *That's how it was explained to me. In principle, I am responsible for this [electronic patient] file. So, I can say that the doctor can see various things and the doctor cannot see various things.* [FG2.3]

Patterns, Differences, and Relationships Regarding Sociodemographic Characteristics

Gender

The category of personal and privacy concerns was predominantly articulated by female participants. The concerns about discrimination or stigmatization expressed were concrete and related to disadvantages in professional life and social contexts. By contrast, male participants emphasized the category of informational self-determination the most, especially the points of autonomy and control. Statements such as "But that everyone can then decide whether to approve data sharing or not" (FG2.5) illustrate a desire for fundamental sovereignty and the ability to make independent decisions. Furthermore, male participants called for clearer rules and more rigorous monitoring within the legal and ethical framework conditions and requirements. Their priority was to establish a reliable, externally controlled system that could be trusted.

Age

For older participants (aged 49-79 years), the focus was on data security concerns and informational self-determination principles. As with male participants, autonomy and control were central to this group. The core issue was the fundamental right to control how one´s own data are used. There were general but strong concerns voiced about the overall security of the system, which often failed to address technical details. Concerns regarding data access emerged leading in this group with statements like "But the abuse that results when someone who has no actual responsibility for it gets hold of the date is the fundamental fear" (FG2.1). Conversely, younger participants (aged 25-43 years) focused on concrete use cases and the practical individual usefulness for medical care. Emergency assistance was a frequent topic of discussion, highlighting a willingness to share health data in a clearly defined context. Younger participants expressed very granular preferences regarding the consent model. Within the subcategory of recipient and purpose, they specified which health data should be visible to which health care professionals (eg, general practitioner vs specialist), indicating a desire for flexibility and adaptability within the scope of consent.

Socioeconomic Status

The characteristics of education and income revealed overlapping patterns in the FG discussions and are therefore analyzed together. The analysis showed that socioeconomic status strongly influences the nature of the perceived positive

XSL•FO
RenderX

and negative consequences of sharing health data. Participants with higher education and/or higher income (>€3000) made arguments for or against sharing data on a more systemic level. Participants in this group discussed technical safety measures in detail, citing anonymization and pseudonymization as prerequisites for consent.

When personal and privacy concerns were raised, they primarily fell within the subcategory of concerns relating to commercial interests:

> *It will always result in advertising and similar things. So, I think there will be enough creative ways within the industry to use health data to reach target groups even more precisely, to market certain products even more accurately, and so on.* [FG1.4]

The 2 participants with the highest income, who have a university degree and are in the middle of their professional lives, explicitly supported the idea of an opt-out solution. However, they linked this to clear conditions that meet their needs for both individual usefulness and individual control.

By contrast, participants with lower and/or middle levels of education and/or lower incomes raised more specific concerns related to personal matters and privacy. As with female participants, the subcategory of concerns related to discrimination or stigmatization was extremely dominant. The fear of job loss or disadvantages with insurance was existential. Another focus was on trust: while the positive social influence of health care professionals and trust in the treating physician were cited as anchors, trust in anonymous, digital systems was low.

Discussion

Overview

For this study, we conducted 2 FGs and analyzed the data using qualitative thematic analysis to explore the perspectives of German patients with somatic and mental health diseases on health data sharing. We explored similarities and differences between the 2 patient groups regarding factors influencing their willingness to share health data for PU and SU. Analogous to the order of the Results presentation, the findings are now discussed in the context of the current state of research.

Discussion of Principal Findings

Previous Experience With Data Sharing

The majority of participants were already experienced in sharing health data, with a predominantly positive response. While 3 participants from each FG reported some negative experiences, they were still willing to share their health data with health care professionals. This suggests that the individual usefulness for medical care can outweigh these negative experiences. In the context of PU, an interview study of the general population in Poland in 2024 showed that positive experiences and frequent use of eHealth solutions are important facilitators for health data sharing [43]. Similar findings regarding SU have been reported in the literature, where 98% of these participants considered the altruistic and public benefits of data sharing to outweigh the potential risks [44].

Individual Usefulness for Medical Care

Patients in both FGs were highly supportive of sharing health data with their health care providers, especially to support personalized medical care. This finding is consistent with previous qualitative studies [27,45,46] postulating benefits for better treatment through comprehensive documentation, such as avoiding multiple diagnoses. However, a scoping review [47] concluded that relevant sensitive information about patients' mental health was regularly missing from electronic health records because they follow a more narrative reporting style, negatively impacting the flow of information in the treatment process. Nonetheless, in this study 1 patient per FG argued that an electronic patient file should include only objectifiable data, such as laboratory parameters, and not subjective assessments.

These findings are also consistent with 2 qualitative studies [22,48] regarding the potential to increase personal informedness and simplify processes. Personal access to medical records can help patients with mental health diseases improve their understanding of their diagnosis, their mental health status over time, and the reasons for their medical treatment, as well as help them remember their previous visits to the physician. In addition, access to personal data increases the sense of control over one's own health care [49]. Emergency assistance and time-saving aspects were frequent arguments for health data sharing, especially for younger participants, and also echo previous qualitative studies [43,50].

Public Benefit

In both FGs, the potential for improvements in medical care through the sharing of health data was seen as a key public benefit. This was considered in 2 ways: in terms of drug and product development and in terms of gaining knowledge, for example, for evidence-based public policy decision-making and deriving preventive measures from health statistics. A current study using German survey data and a survey experimental vignette design confirmed that individuals understand how valuable their data are for researchers in terms of SU and possible public benefits [51].

Participants in FG2 particularly supported the work of the pharmaceutical industry and acknowledged the improvements that have been achieved over the years through the development of drugs for the health care of patients with mental health diseases. An interview study involving patients with mental health diseases in the United Kingdom revealed a desire to alleviate their suffering, as well as the expectation of personalized medicine, tailored treatments, and improved disease management in the population [28].

Personal and Privacy Concerns

In this study, especially participants with higher levels of education and/or higher incomes had major concerns about commercial interests, particularly regarding the use of health data by technology companies, which were discussed extensively by FG2. These findings are consistent with the existing literature of other FGs and an online survey in the United Kingdom [46,52,53]. Similarly, concerns about discrimination and stigmatization were more prevalent among the FG2 sample, particularly among participants with lower and

XSL•FO
RenderX

middle levels of education and/or lower incomes, and among women. These fears, particularly discrimination by insurance companies and employers in the context of SU, have been previously reported [28]. In a qualitative study, mental health service users expressed concerns about institutional discrimination [52]. The stigma associated with mental health diseases has hindered patients' acceptance of electronic communication and data sharing, and limited the broader use of this technology among underserved populations [29], making the reduction of stigmatization particularly important. Moreover, participants in both FGs also expressed concerns about being overstrained by the amount of accessible documentation and the wide availability of data, findings that are consistent with those of a qualitative study in Belgium involving vulnerable patients [22].

Consent Management Preferences

Overall, the results of this study show that no clear consensus has emerged regarding preferences for consent management, but fundamental differences were found between the characteristics of diseases and sociodemographic data. The aspects described below may explain our inconsistent findings and highlight the importance of granular personal control over data access, limited datasets, restricting specific information, purposes and recipients, and choices regarding health data sharing. A qualitative study exploring the public's hopes and fears about health data sharing has shown that these perspectives are strongly influenced by an individual's personal circumstances [54].

Regarding PU, some patients in both FGs expressed a desire for limited data sharing, while others agreed to full sharing of data. Younger participants expressed very specific preferences regarding which health care professionals should have access to which health data (eg, general practitioner vs specialist). Among other things, mental health diseases were categorized as highly sensitive in both FGs, and consequently associated with restrictions on data sharing or even rejection. A systematic review including survey and interview data also demonstrated greater agreement to share nonsensitive digital health data, with mental health, drug or alcohol use, and sexually transmitted diseases being considered particularly sensitive [55]. Regarding SU, our results suggest that university research often received greater support than pharmaceutical research, and participants in both FGs were explicitly opposed to the use of health data by Amazon, Google, and Apple, although artificial intelligence could transform patient care through its applications and augmentation of medical knowledge [56]. Our results on differentiation by recipient and purpose reflect the findings from studies in Germany [35,57] and other European countries [53,58]. A recent German cross-sectional survey also showed that only about 29% of the population would share their health data with private research institutions, and 66% are unwilling to share their health data with commercial companies, especially not technology companies [45].

While 2 participants with the highest income, who have a university degree and are in the middle of their professional lives, expressed a preference for opting out with their own personal options, others felt that the ability to make fundamental, independent choices was vital. This aligns with studies indicating that greater openness to data sharing is associated with higher digital literacy and socioeconomic status when control mechanisms are in place [18,35,59]. However, concerns about the reliability of our findings must be raised, as the distinction between opt-in and opt-out may have been ambiguous. At times, it appears that opt-out was interpreted as "inconvenient," as noted by 1 participant in FG1, and might indicate insufficient knowledge on this issue. In an Irish FG study involving 85 participants by Flaherty et al [58], opt-out was considered more convenient because it avoids the need for repeated consent for similar studies and is more cost-effective. Similarly, participants in a recent interview study in Germany emphasized the convenience and effectiveness of opt-out [57]. An interview study with health care professionals for patients with mental diseases also supports the hypothesis of a possible lack of knowledge, where most respondents (75%) felt that their patients did not fully understand the process and consent forms for data sharing [24]. The identified knowledge gap found in this study and elsewhere underscores the need for thorough public information about the different consent options and existing regulations, some of which have already been implemented. At the European level, the European Health Data Space (EHDS) regulation, in force since March 2025 and applicable from March 2027, encompasses different key milestones [60]. By 2029, each member state must have established a national health data access body [61]. Details of the EHDS are currently being elaborated in European projects (eg the Joint Action Towards the European Health Data Space 2) [62]. A workshop focused on the perspectives of different stakeholders on the opt-out procedure within the EHDS, as well as how Germany can structure the opt-out procedure and rights of objection within this framework [62,63].

Legal and Ethical Framework Conditions

The issues of data protection and ethical justification for data sharing were particularly noteworthy to both of our FGs. The participants in FG2 strongly emphasized that health data are generally very sensitive and need special access protections. Patients with somatic diseases, particularly male patients, requested special monitoring and authorization measures. Although data protection in Germany is already very robust and a deliberate, data protection–focused approach is followed, there appear to be evident concerns and a lack of knowledge [11]. In an FG study in the Netherlands, Wetzels et al [64] showed that patients had limited knowledge of data protection practices. These results demonstrate the need for broader and targeted public relations work to better educate the public on this issue. Germany provides information for the interesting public on websites, interactive learning platforms, and in journals, which also include information on data protection [65-70].

Informational Self-Determination

Especially the points of autonomy and control were fundamental for the male and older participants aged 49-79 years. These 2 aspects, as well as trust, were consequently identified as facilitators of health data sharing in both FGs and elsewhere [71,72]. A Finnish study using survey data to develop a theoretical model found that trust and control are key mediators

XSL•FO
RenderX

between attitudes and willingness to share health information, with trust being more influential than control [73]. However, a questionnaire study from Denmark revealed a strong desire for control over personal health data, even when there is trust in the institutions using it [74]. In a computer-assisted telephone interview study conducted in Germany, trust emerged as the most significant factor to increase the willingness to share personal health data [59].

Social Involvement and Influence

Positive social influence of health care professionals and trust in the treating physician can encourage health data sharing according to 2 participants in FG2 with lower levels of education and/or lower incomes. The existing literature certainly indicates a substantial influence of physicians' behavior on patients' willingness to share their health data [75]. In patients with mental health diseases, physicians may underestimate the importance of physical complaints and attribute them to a mental disorder, which can lead to misdiagnosis and poorer treatment outcomes. This phenomenon of "diagnostic overshadowing" can lead to a loss of trust in the treating physician, particularly in those with mental health diseases, reducing their willingness to share their health data [52,76-78]. Positive encouragement and clear, open communication strategies from health care professionals can be important facilitators and were found to significantly influence outcomes [79]. Specifically for this purpose, a patient handout in various and simplified languages, as well as waiting room videos and posters for German medical institutions, were developed [70]. Moreover, German statutory health insurance companies provided their patients with comprehensive information materials [80], and an interactive learning platform for experts is gradually being established [68].

Strengths and Limitations

A major strength of this study is the diversity of its cohort in terms of gender, age, and somatic and mental health diseases. The diversity of these characteristics was considered during the recruitment process, as it allows access to different perspectives from patients with different chronic and mental health diseases. In particular, the opinions of patients with mental health diseases are still rarely prioritized in research on health data sharing, even though they represent more than one-third of the population [81]. The group dynamics and observable interactions in FGs provide valuable and deeper insights than face-to-face interviews. Another strength of this study is that the FGs were able to show how participants' views are formed and how they change through social interaction. This allowed us to better contextualize the verbal data.

However, our findings may have been biased by problems recalling experiences in the distant past. Our participants also represent a convenience sample, as participants were recruited from the University Hospital in Dresden, Germany. In addition, our cohort may have been affected by selection bias due to the financial compensation of €50. Moreover, investigated relationships between sociodemographic characteristics like age, gender, education, and income levels as possible influencing factors may only be considered of limited significance due to the small sample size. Furthermore, our recruitment of only relatively tech-savvy patients who used a smartphone or tablet daily to increase FG efficiency may have introduced selection bias, potentially resulting in higher technology acceptance scores. Besides, our results may have been influenced by volunteer bias. In addition, our results may have been influenced by social desirability, as participants may have feared being judged by other FG members if they provided truthful answers. Finally, qualitative research findings are not designed to be generalizable. However, an exploratory study design enables patient participation, as emphasized in the GDNG [1], and allows for a deeper and more comprehensive understanding of patients' perspectives, as well as facilitating the development of hypotheses to inform future research and practice.

Implications for Future Research and Practice

Further validation of the findings of our exploratory qualitative study on data sharing can be achieved through future research with a representative sample. This research could also reveal possible correlations with education, income, or place of residence, which have already been identified at the international level [9]. Future qualitative and quantitative research comparing the attitudes of patients with somatic and mental health diseases, as well as healthy individuals, could deepen our understanding of their unique needs and concerns. Investigating the attitudes of individuals with poor mental health could also be beneficial, despite the potential challenges of recruitment.

Social influence of family members and especially health care professionals had a much greater impact on patients with mental health diseases and for those with a lower socioeconomic status. Therefore, the latter should be involved in future research and, after the aforementioned training measures, be empowered in their role as multipliers to inform patients about the potential benefits and risks of sharing health data. Research should also assess the effectiveness of involving family members, peers, and health care professionals in decision-making processes. This research can refine governance frameworks and improve patient-centered approaches to health data sharing. It can also lead to powerful medical artificial intelligence systems, such as clinical decision support systems, that are trained on higher-quality and more representative data.

The issue of data protection was more important for patients with mental health diseases. Both they and the female participants expressed greater fears of discrimination and stigmatization. To reduce the identified knowledge deficits, which have even become apparent among tech-savvy patients, it is crucial to implement broad education and public relations initiatives that provide easy-to-understand information about the very high data protection standards and the procedure for storing and transferring health data. In addition, there needs to be greater awareness of the different levels of consent options, such as only allowing certain physicians to access individuals' electronic health data.

Conclusions

The use of health data for PU and SU is a valuable capability of a modern health care system and the society it serves. Therefore, patients' and health care professionals' perspectives must be considered when designing the regulatory framework. This study provides insights into the different perspectives and

XSL•FO
RenderX

preferences of patients with somatic and mental health diseases regarding data sharing. In particular, those with mental health diseases viewed their health data as highly sensitive due to previous experiences and fears of discrimination and stigmatization. Therefore, tailored consent management and involving patients in health data sharing were highly preferred. Participatory approaches involving family members, peers, and health care professionals can also help to reduce fears and increase acceptance. Health care professionals play an important role, as they can ensure transparency through targeted educational work and actively sensitize patients to the principles of data sharing, which can be supportive, even if commercial interests are involved. The knowledge deficits identified in this study, which also existed among tech-savvy patients, indicate that broad and easily understandable awareness-raising and public relations initiatives are needed. They can help inform the population comprehensively about the very high data protection standards, the procedure for storing and transferring health data, and the personalized consent options.

Acknowledgments

We would like to thank Ivonne Panchyrz (Dr Phil) from the Center for Evidence-Based Healthcare at the University of Technology in Dresden for moderating the 2 focus groups and for providing conceptual ideas. We would also like to thank all the recruiting physicians at Dresden University Hospital, Susanne Abraham (Dr Med) and her team, Tjalf Ziemssen (Prof Dr Med), and Peter Schwarz (Prof Dr Med). Our very special thanks go to all the patients who took part in the focus groups, whose reports and opinions created an open atmosphere for discussion and made it possible for us to generate these results.

The authors confirm that no generative artificial intelligence tools or other forms of artificial intelligence were used for text generation, editing, or analysis in the preparation of this manuscript. All content was created independently by the authors.

Funding

Open Access funding was enabled and organized by Project DEAL. The authors would like to thank the German Federal Ministry of Education and Research for funding the PATH (Personal Mastery of Health and Wellness Data; grant number 16KISA100K) project and making this qualitative study possible.

Authors' Contributions

SF, CL, FGV, and MS were responsible for the study design. SF developed the discussion guide and conducted the pilot testing. SF formed the initial category system and conducted the coding of the material together with CL. MS was involved in the consensus process of finalizing the category system and critically checked the codings. SF drafted the manuscript with the support of CL, FGV, and MS. JS substantially revised the content of the manuscript before finalization with regard to the presentation, interpretation, and implications of the results and approved it for submission. All authors critically revised the manuscript for important intellectual content and approved the final version of the manuscript. MS obtained funding for the project.

Conflicts of Interest

SF, CL, FGV, and MS declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this study. Unrelated to this study, JS reports institutional grants for investigator-initiated research from the German Federal Joint Committee (G-BA), the Federal Ministry of Health (BMG), the Federal Ministry of Education and Research (BMFTR), the Federal State of Saxony, Novartis, Sanofi, ALK, and Pfizer. He also participated in advisory board meetings as a paid consultant for Sanofi, Lilly, and ALK.

Multimedia Appendix 1
COREQ checklist.
[DOC File , 62 KB - jmir_v28i1e79990_app1.doc]

Multimedia Appendix 2
Guiding questions and topics for both focus groups.
[DOCX File , 23 KB - jmir_v28i1e79990_app2.docx]

Multimedia Appendix 3
Codebook with definitions, respective anchor examples and number / context of statements used for qualitative coding and analysis in MAXQDA.
[DOCX File , 39 KB - jmir_v28i1e79990_app3.docx]

Multimedia Appendix 4
Participant characteristics on sociodemographic characteristics.

XSL•FO
RenderX

[DOCX File , 16 KB - jmir_v28i1e79990_app4.docx]

Multimedia Appendix 5
Participant characteristics on technology commitment scores.
[DOCX File , 13 KB - jmir_v28i1e79990_app5.docx]

References

1. Gesetz zur verbesserten Nutzung von Gesundheitsdaten (Gesundheitsdatennutzungsgesetz, GDNG). Bundesgesetzblatt. 2024. URL: https://www.recht.bund.de/bgbl/1/2024/102/VO.html [accessed 2025-06-19]
2. Gesetz zur Beschleunigung der Digitalisierung des Gesundheitswesens (Digital-Gesetz, DigiG). Bundesgesetzblatt. 2024. URL: https://www.recht.bund.de/bgbl/1/2024/101/VO.html [accessed 2025-06-19]
3. Archer N, Fevrier-Thomas U, Lokker C, McKibbon KA, Straus SE. Personal health records: A scoping review. J Am Med Inform Assoc 2011;18(4):515-522 [FREE Full text] [doi: 10.1136/amiajnl-2011-000105] [Medline: 21672914]
4. Chen D. Open data: Implications on privacy in healthcare research. Blockchain Healthc Today 2020;3:e144 [FREE Full text] [doi: 10.30953/bhty.v3.144] [Medline: 36777059]
5. Colwill M, Poullis A. Using national census data to facilitate healthcare research. World J Methodol 2023;13(5):414-418 [FREE Full text] [doi: 10.5662/wjm.v13.i5.414] [Medline: 38229939]
6. Hauswaldt J, Kempter V, Himmel W, Hummers E. Hindernisse bei der sekundäranalyse von routinedaten aus der primärversorgung. Gesundheitswesen 2018;80(11):987-993. [doi: 10.1055/a-0668-5817]
7. Kim TK, Choi M. Older adults' willingness to share their personal and health information when adopting healthcare technology and services. Int J Med Inform 2019;126:86-94. [doi: 10.1016/j.ijmedinf.2019.03.010] [Medline: 31029268]
8. European Parliament and Council. Directive 2011/24/EU of the European parliament and of the council of 9 march 2011 on the application of patients' rights in cross-border healthcare. Off J Eur Union 2011;L88:45-65 [FREE Full text] [doi: 10.3233/ppl-2011-0333]
9. Fesl S, Lang C, Schmitt J, Brückner S, Gilbert S, Deckert S, et al. Factors influencing patients' willingness to share their digital health data for primary and secondary use: A theory- and evidence-based overview of reviews. Digit Health 2025;11:20552076251340254 [FREE Full text] [doi: 10.1177/20552076251340254] [Medline: 40599675]
10. European Parliament and Council. Regulation (EU) 2025/327 of the European parliament and of the council of 11 february 2025 on the European health data space. Off J Eur Union 2025;L61:1-95 [FREE Full text] [doi: 10.5771/9783748968955-481]
11. Cuggia M, Combes S. The French health data hub and the German medical informatics initiatives: Two national projects to promote data sharing in healthcare. Yearb Med Inform 2019;28(1):195-202 [FREE Full text] [doi: 10.1055/s-0039-1677917] [Medline: 31419832]
12. Semler SC, Wissing F, Heyder R. German medical informatics initiative. Methods Inf Med 2018;57(S 01):e50-e56 [FREE Full text] [doi: 10.3414/ME18-03-0003] [Medline: 30016818]
13. Das FDZ Gesundheit (Health Data Lab). Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM). 2025. URL: https://www.forschungsdatenzentrum-gesundheit.de/das-fdz [accessed 2025-06-19]
14. Biasiotto R, Viberg Johansson J, Alemu MB, Romano V, Bentzen HB, Kaye J, et al. Public preferences for digital health data sharing: Discrete choice experiment study in 12 European Countries. J Med Internet Res 2023;25:e47066 [FREE Full text] [doi: 10.2196/47066] [Medline: 37995125]
15. Panteli D, Röttger J, Nimptsch U, Busse R. Internationale Datengrundlagen für die Versorgungsforschung ? In: Impulse für Deutschland. Berlin: Universitätsverlag der TU Berlin; 2021.
16. Vlahou A, Hallinan D, Apweiler R, Argiles A, Beige J, Benigni A, et al. Data sharing under the general data protection regulation. Hypertension 2021;77(4):1029-1035. [doi: 10.1161/hypertensionaha.120.16340]
17. Moser K, Massag J, Frese T, Mikolajczyk R, Christoph J, Pushpa J, et al. German primary care data collection projects: A scoping review. BMJ Open 2024;14(2):e074566 [FREE Full text] [doi: 10.1136/bmjopen-2023-074566] [Medline: 38382948]
18. Baines R, Stevens S, Austin D, Anil K, Bradwell H, Cooper L, et al. Patient and public willingness to share personal health data for third-party or secondary uses: Systematic review. J Med Internet Res 2024;26:e50421 [FREE Full text] [doi: 10.2196/50421] [Medline: 38441944]
19. Cascini F, Pantovic A, Al-Ajlouni YA, Puleo V, De Maio L, Ricciardi W. Health data sharing attitudes towards primary and secondary use of data: A systematic review. EClinicalMedicine 2024;71:102551 [FREE Full text] [doi: 10.1016/j.eclinm.2024.102551] [Medline: 38533128]
20. Aitken M, de St Jorre J, Pagliari C, Jepson R, Cunningham-Burley S. Public responses to the sharing and linkage of health data for research purposes: A systematic review and thematic synthesis of qualitative studies. BMC Med Ethics 2016;17(1):73 [FREE Full text] [doi: 10.1186/s12910-016-0153-x] [Medline: 27832780]
21. Howe N, Giles E, Newbury-Birch D, McColl E. Systematic review of participants' attitudes towards data sharing: A thematic synthesis. J Health Serv Res Policy 2018;23(2):123-133. [doi: 10.1177/1355819617751555] [Medline: 29653503]

XSL•FO
RenderX

22. Bernaerdt J, Moerenhout T, Devisch I. Vulnerable patients' attitudes towards sharing medical data and granular control in patient portal systems: An interview study. J Eval Clin Pract 2021;27(2):429-437. [doi: 10.1111/jep.13465] [Medline: 32886953]
23. Grando A, Ivanova J, Hiestand M, Soni H, Murcko A, Saks M, et al. Mental health professional perspectives on health data sharing: Mixed methods study. Health Informatics J 2020;26(3):2067-2082 [FREE Full text] [doi: 10.1177/1460458219893848] [Medline: 31928295]
24. Ivanova J, Grando A, Murcko A, Saks M, Whitfield MJ, Dye C, et al. Mental health professionals' perceptions on patients control of data sharing. Health Informatics J 2020;26(3):2011-2029 [FREE Full text] [doi: 10.1177/1460458219893845] [Medline: 31912744]
25. Mozersky J, Parsons M, Walsh H, Baldwin K, McIntosh T, DuBois JM. Research participant views regarding qualitative data sharing. Ethics Hum Res 2020;42(2):13-27 [FREE Full text] [doi: 10.1002/eahr.500044] [Medline: 32233117]
26. Satinsky EN, Driessens C, Crepaz-Keay D, Kousoulis A. Mental health service users' perceptions of data sharing and data protection: A qualitative report. J Innov Health Inform 2018;25(4):239-242 [FREE Full text] [doi: 10.14236/jhi.v25i4.1033] [Medline: 30672404]
27. Soni H, Grando A, Aliste M, Murcko A, Todd M, Mukundan M, et al. Perceptions and preferences about granular data sharing and privacy of behavioral health patients. Stud Health Technol Inform 2019;264:1361-1365. [doi: 10.3233/SHTI190449] [Medline: 31438148]
28. Watson E, Fletcher-Watson S, Kirkham EJ. Views on sharing mental health data for research purposes: Qualitative analysis of interviews with people with mental illness. BMC Med Ethics 2023;24(1):99 [FREE Full text] [doi: 10.1186/s12910-023-00961-6] [Medline: 37964278]
29. Matthews EB, Savoy M, Paranjape A, Washington D, Hackney T, Galis D, et al. Acceptability of health information exchange and patient portal use in depression care among underrepresented patients. J Gen Intern Med 2022;37(15):3947-3955 [FREE Full text] [doi: 10.1007/s11606-022-07427-2] [Medline: 35132548]
30. Schuller J, Cavenaile J, Briganti G. Somatic management of psychiatric patients in the emergency department. Psychiatr Danub 2023;35(Suppl 2):341-346 [FREE Full text] [Medline: 37800253]
31. Honey A, Hancock N, Glover H, Scanlan JN, Cao Y, Povolny A, et al. Collection and sharing of health information in mental health and related systems in Australia: Perspectives of people who access mental health services. BMC Psychiatry 2024;24(1):932 [FREE Full text] [doi: 10.1186/s12888-024-06347-1] [Medline: 39695423]
32. Pogany L, Lazary J. Health control beliefs and attitude toward treatment in psychiatric and non-psychiatric clinical samples. Front Psychiatry 2021;12:537309 [FREE Full text] [doi: 10.3389/fpsyt.2021.537309] [Medline: 34025463]
33. Henkenjohann R. Role of individual motivations and privacy concerns in the adoption of german electronic patient record apps-a mixed-methods study. Int J Environ Res Public Health 2021;18(18):9553 [FREE Full text] [doi: 10.3390/ijerph18189553] [Medline: 34574475]
34. Voigt TH, Holtz V, Niemiec E, Howard HC, Middleton A, Prainsack B. Willingness to donate genomic and other medical data: Results from Germany. Eur J Hum Genet 2020;28(8):1000-1009 [FREE Full text] [doi: 10.1038/s41431-020-0611-2] [Medline: 32238912]
35. Dotter C, Haug S, Schnell R, Raptis G, Weber K. Sharing health data for research purposes: Results of a population survey in Germany. BMC Health Serv Res 2025;25(1):699 [FREE Full text] [doi: 10.1186/s12913-025-12706-9] [Medline: 40369578]
36. Stewart D, Shamdasani P, Rook D. Focus Groups. Thousand Oaks, CA: SAGE Publications, Ltd; 2007.
37. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
38. Lenzner T, Neuert C, Otto W. Kognitives Pretesting. In: GESIS Survey Guidelines. Mannheim: GESIS - Leibniz-Institut für Sozialwissenschaften; 2015.
39. Neyer FJ, Felber J, Gebhardt C. Kurzskala Technikbereitschaft (TB, technology commitment). In: Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS). Mannheim: GESIS - Leibniz-Institut für Sozialwissenschaften; 2016.
40. Dresing T, Pehl T. Transkription. In: Mey G, Mruck K, editors. Handbuch qualitative Forschung in der Psychologie. Wiesbaden: VS Verlag für Sozialwissenschaften; 2010:723-733.
41. Kuckartz U, Rädiker S. Qualitative Content Analysis: Methods, Practice and Software. London: SAGE; 2023.
42. Mayring P. Qualitative content analysis. Forum Qual Soc Res 2019;20(3):16. [doi: 10.17169/fqs-20.3.3343]
43. Smoła P, Młoźniak I, Wojcieszko M, Zwierczyk U, Kobryn M, Rzepecka E, et al. The perception of facilitators and barriers to the use of e-health solutions in Poland: A qualitative study. BMC Med Inform Decis Mak 2024;24(1):381 [FREE Full text] [doi: 10.1186/s12911-024-02791-x] [Medline: 39695644]
44. Spencer K, Sanders C, Whitley EA, Lund D, Kaye J, Dixon WG. Patient perspectives on sharing anonymized personal health data using a digital system for dynamic consent and research feedback: A qualitative study. J Med Internet Res 2016;18(4):e66 [FREE Full text] [doi: 10.2196/jmir.5011] [Medline: 27083521]
45. Kühnel E, Wilke F. Health data sharing in Germany: Individual preconditions, trust and motives. Front Public Health 2025;13:1538106 [FREE Full text] [doi: 10.3389/fpubh.2025.1538106] [Medline: 39975784]

XSL•FO
RenderX

46. Varhol RJ, Man Ying Lee C, Hindmarsh S, Boyd JH, Robinson S, Randall S. Consumer attitudes, barriers and facilitators to sharing clinical data for research purposes: Results from a focus group synthesis. Heliyon 2024;10(14):e34431 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e34431] [Medline: 39816330]
47. Kariotis TC, Prictor M, Chang S, Gray K. Impact of electronic health records on information practices in mental health contexts: Scoping review. J Med Internet Res 2022;24(5):e30405 [FREE Full text] [doi: 10.2196/30405] [Medline: 35507393]
48. Jezrawi R, Balakumar S, Masud R, Gabizon I, Bhagirath V, Varughese J, et al. Patient and physician perspectives on the use and outcome measures of mHealth apps: Exploratory survey and focus group study. Digit Health 2022;8:20552076221102773 [FREE Full text] [doi: 10.1177/20552076221102773] [Medline: 35646382]
49. Blease C, Kharko A, Hägglund M, O'Neill S, Wachenheim D, Salmi L, et al. The benefits and harms of open notes in mental health: A Delphi survey of international experts. PLoS One 2021;16(10):e0258056 [FREE Full text] [doi: 10.1371/journal.pone.0258056] [Medline: 34644320]
50. Medford-Davis LN, Chang L, Rhodes KV. Health information exchange: What do patients want? Health Informatics J 2017;23(4):268-278 [FREE Full text] [doi: 10.1177/1460458216647190] [Medline: 27245671]
51. Bach RL, Silber H, Gerdon F, Keusch F, Schonlau M, Schröder J. To share or not to share – understanding individuals' willingness to share biomarkers, sensor data, and medical records. Information, Communication & Society 2024;28(10):1799-1817. [doi: 10.1080/1369118x.2024.2351439]
52. Adanijo A, McWilliams C, Wykes T, Jilka S. Investigating mental health service user opinions on clinical data sharing: Qualitative focus group study. JMIR Ment Health 2021;8(9):e30596 [FREE Full text] [doi: 10.2196/30596] [Medline: 34477558]
53. Jones LA, Nelder JR, Fryer JM, Alsop PH, Geary MR, Prince M, et al. Public opinion on sharing data from health services for clinical and research purposes without explicit consent: An anonymous online survey in the UK. BMJ Open 2022;12(4):e057579 [FREE Full text] [doi: 10.1136/bmjopen-2021-057579] [Medline: 35477868]
54. Lounsbury O, Roberts L, Goodman JR, Batey P, Naar L, Flott KM, et al. Opening a "Can of Worms" to explore the public's hopes and fears about health care data sharing: Qualitative study. J Med Internet Res 2021;23(2):e22744 [FREE Full text] [doi: 10.2196/22744] [Medline: 33616532]
55. Moon LA. Factors influencing health data sharing preferences of consumers: A critical review. Health Policy and Technology 2017;6(2):169-187. [doi: 10.1016/j.hlpt.2017.01.001]
56. Aung YYM, Wong DCS, Ting DSW. The promise of artificial intelligence: A review of the opportunities and challenges of artificial intelligence in healthcare. Br Med Bull 2021;139(1):4-15. [doi: 10.1093/bmb/ldab016] [Medline: 34405854]
57. Richter G, Trigui N, Caliebe A, Krawczak M. Attitude towards consent-free research use of personal medical data in the general German population. Heliyon 2024;10(6):e27933 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e27933] [Medline: 38509969]
58. Flaherty SJ, Duggan C, O'Connor L, Foley B, Flynn R. What influences a person's willingness to share health information for both direct care and uses beyond direct care? Findings from a focus group study in Ireland. HRB Open Res 2022;5:36 [FREE Full text] [doi: 10.12688/hrbopenres.13548.2] [Medline: 36072819]
59. Wilke F. The impact of trust and the role of the opt-out mechanism in willingness to share health data via electronic health records in Germany: telephone survey study. JMIR Hum Factors 2025;12:e65718 [FREE Full text] [doi: 10.2196/65718] [Medline: 40233172]
60. European Parliament and Council. Regulation (EU) 2025/327 on the European health data space. Off J Eur Union 2025;L61:1-95 [FREE Full text]
61. Saelaert M, Mathieu L, Van Hoof W, Devleesschauwer B. Expanding citizen engagement in the secondary use of health data: An opportunity for national health data access bodies to realise the intentions of the European health data space. Arch Public Health 2023;81(1):168 [FREE Full text] [doi: 10.1186/s13690-023-01182-4] [Medline: 37700330]
62. EHDS - be prepared for public consultations in TEHDAS2: Draft specification on the implementation of secure processing environments, online workshop series. TMF - Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V. 2025. URL: https://www.tmf-ev.de/sites/default/files/2025-09/EHDS_PublicConsultationWorkshops_Semler_Lehva%CC%88slaiho_7.4_0.pdf [accessed 2025-10-01]
63. Perspektiven auf das Opt-out im Europäischen Gesundheitsdatenraum. TMF-Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V. 2025. URL: https://www.tmf-ev.de/news/perspektiven-auf-das-opt-out-im-europaeischen-gesundheitsdatenraum [accessed 2025-12-07]
64. Wetzels M, Broers E, Peters P, Feijs L, Widdershoven J, Habibovic M. Patient perspectives on health data privacy and management: "Where Is My Data and Whose Is It?". Int J Telemed Appl 2018;2018:3838747 [FREE Full text] [doi: 10.1155/2018/3838747] [Medline: 30631347]
65. Vernetzen. Forschen. Heilen. Medizininformatik-Initiative. 2025. URL: https://www.vernetzen-forschen-heilen.de/ [accessed 2025-06-19]
66. ONE MII: Journal der Medizininformatik-Initiative, Daten nutzen. Medizininformatik-Initiative. 2025. URL: https://www.medizininformatik-initiative.de/sites/default/files/2025-04/01-120_N8%20ONEMII_FINAL_LOW.pdf [accessed 2025-06-19]

XSL•FO
RenderX

67. Lauterbach startet bundesweite Infokampagne zur "ePA für alle". Bundesministerium für Gesundheit. 2024. URL: https://www.bundesgesundheitsministerium.de/presse/pressemitteilungen/infokampagne-epa-fuer-alle-startet.html [accessed 2025-06-19]
68. L-INA learning interoperability. Gematik. 2025. URL: https://l-ina.gematik.de/ [accessed 2025-06-19]
69. ePA für alle - News & Aktuelles. Gematik. 2025. URL: https://www.gematik.de/epafueralle/news [accessed 2025-06-19]
70. ePA für alle - ePA infopaket. Gematik. 2025. URL: https://www.gematik.de/anwendungen/epa-fuer-alle/download-infopaket [accessed 2025-06-19]
71. Leigh S, Baines R, Stevens S, Garba-Sani Z, Austin D, Chatterjee A. Walk a mile in my shoes: Perspectives towards sharing of health and experience data among individuals living with sickle cell disorder. Mhealth 2024;10:4 [FREE Full text] [doi: 10.21037/mhealth-23-18] [Medline: 38323148]
72. Cumyn A, Barton A, Dault R, Safa N, Cloutier AM, Ethier JF. Meta-consent for the secondary use of health data within a learning health system: A qualitative study of the public's perspective. BMC Med Ethics 2021;22(1):81 [FREE Full text] [doi: 10.1186/s12910-021-00647-x] [Medline: 34187453]
73. Juga J, Juntunen J, Koivumäki T. Willingness to share personal health information: impact of attitudes, trust and control. RMJ 2021;31(1):48-59. [doi: 10.1108/rmj-02-2020-0005]
74. Holm S, Kristiansen TB, Ploug T. Control, trust and the sharing of health information: The limits of trust. J Med Ethics 2020;47(12):e35. [doi: 10.1136/medethics-2019-105887] [Medline: 32843438]
75. Abdelhamid M, Gaia J, Sanders GL. Putting the focus back on the patient: How privacy concerns affect personal health information sharing intentions. J Med Internet Res 2017;19(9):e169 [FREE Full text] [doi: 10.2196/jmir.6877] [Medline: 28903895]
76. Graber MA, Bergus G, Dawson JD, Wood GB, Levy BT, Levin I. Effect of a patient's psychiatric history on physicians' estimation of probability of disease. J Gen Intern Med 2000;15(3):204-206 [FREE Full text] [doi: 10.1046/j.1525-1497.2000.04399.x] [Medline: 10718903]
77. Hallyburton A. Diagnostic overshadowing: An evolutionary concept analysis on the misattribution of physical symptoms to pre-existing psychological illnesses. Int J Ment Health Nurs 2022;31(6):1360-1372 [FREE Full text] [doi: 10.1111/inm.13034] [Medline: 35718951]
78. Nash M, D'Astoli P, Molloy R. Preventing diagnostic overshadowing to improve the physical health of people with severe mental illness. Ment Health Pract 2024;27(5):e1707. [doi: 10.7748/mhp.2024.e1707]
79. Papadopoulos K, Ammenwerth E, Lame G, Stahl N, Struckmann V, von Wyl V, et al. Understanding public trust in national electronic health record systems: A multi-national qualitative research study. Digit Health 2025;11:20552076251333576 [FREE Full text] [doi: 10.1177/20552076251333576] [Medline: 40190337]
80. Elektronische Patientenakte, Informationsmaterial nach § 343 Abs. 1a SGB V. DAK-Gesundheit. URL: https://caas.content.dak.de/caas/v1/media/70788/data/eb14c25f626454a5d6d73a9fe6729df4/d972-0025-info-broschuere-epa-0624-digital.pdf [accessed 2025-06-19]
81. Thom J, Jonas B, Reitzle L, Mauz E, Hölling H, Schulz M. Trends in the diagnostic prevalence of mental disorders, 2012-2022—using nationwide outpatient claims data for mental health surveillance. Dtsch Arztebl Int 2024;121(11):355-362 [FREE Full text] [doi: 10.3238/arztebl.m2024.0052] [Medline: 38686592]

Abbreviations

EHDS: European Health Data Space
EU-GDPR: EU General Data Protection Regulation
FG: focus group
GDNG: Gesundheitsdatennutzungsgsetz
PATH: Personal Mastery of Health and Wellness Data
PU: primary use
SU: secondary use

Edited by A Stone; submitted 02.Jul.2025; peer-reviewed by J Schwarz, T Petzold, M Schunk, G Richter; comments to author 07.Nov.2025; revised version received 14.Dec.2025; accepted 07.Jan.2026; published 13.Apr.2026.

Please cite as:
Fesl S, Lang C, Verhees FG, Schmitt J, Scheibe M
Perspectives on Health Data Sharing Among Patients With Somatic and Mental Health Diseases: Focus Group Study
J Med Internet Res 2026;28:e79990
URL: https://www.jmir.org/2026/1/e79990
doi:10.2196/79990
PMID:

XSL•FO
RenderX

©Sabrina Fesl, Caroline Lang, Falk Gerrik Verhees, Jochen Schmitt, Madlen Scheibe. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Physicians' Perceptions and Demand Regarding Clinical and Secondary Use of Patient-Generated Health Data: Cross-Sectional Survey

Ye-Eun Park[1], BA; Sang Sook Beck[2], MPH, PhD; Yura Lee[1], MD, PhD

[1]Department of Information Medicine, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea

[2]Asian Institute for Bioethics and Health Law, General Complex B/D Graduate School of Public Health, Yonsei University Health System, Seoul, Republic of Korea

Corresponding Author:
Yura Lee, MD, PhD
Department of Information Medicine
University of Ulsan College of Medicine
Asan Medical Center
88 Olympic-ro 43-gil
Songpa-gu
Seoul, 05505
Republic of Korea
Phone: 82 2 3010 1498
Fax: 82 2 3010 2531
Email: haepary@amc.seoul.kr

Abstract

Background: Patient-generated health data (PGHD) are increasingly recognized as valuable for clinical care and secondary use; however, physicians' perspectives remain heterogeneous and context-dependent.

Objective: This study aimed to examine physicians' perceptions, intentions, and concerns about the clinical and nonclinical use of PGHD and to compare responses between tertiary hospitals and other institutions.

Methods: A cross-sectional survey was conducted involving 157 physicians (81 from a tertiary hospital and 76 from other health care institutions), supplemented by in-depth interviews. Between-group differences were assessed using chi-square tests and Wilcoxon rank-sum tests. Priority-ranking items were analyzed using weighted scores.

Results: Overall, 64% (100/157) of the physicians reported prior use of PGHD, most commonly in handwritten or printed form, with no significant institutional differences; physiological measurements were the most frequent type of data. Among nonusers, 78.9% (45/57) reported that they had never received PGHD from patients, while poor system integration was identified as a major barrier to clinical use. Physicians cited clinical necessity and user-friendly integration as the strongest motivators. Perceived benefits focused on enriched decision-making, whereas key concerns included data accuracy, increased workload, and insufficient evidence. For nonclinical use, research and public health were viewed as most valuable. Across institutions, secure and transparent governance was prioritized. Differences between tertiary and nontertiary settings appeared to reflect physician demographics rather than divergent attitudes. Interviews further revealed that PGHD use was highly context dependent, with older adults' limited digital literacy and discrepancies between patient- and caregiver-reported data identified as distinctive barriers in routine practice.

Conclusions: Physicians recognize the potential of PGHD to support outpatient care, shared decision-making, and research, but express strong concerns about data reliability, workflow burden, and insufficient infrastructure. Differences between tertiary and nontertiary physicians reflect demographic and career stage imbalances, but both groups converge on the need for standardization, secure integration, and supportive governance frameworks. Broad adoption will require not only technical integration but also regulatory reform and trust-building strategies to advance biomedical research and health care delivery.

(J Med Internet Res 2026;28:e86368) doi:10.2196/86368

KEYWORDS

data management; data utilization; electronic health records; medical record linkage; patient access to records; patient-generated health data; patient-centered care; personal health records; telemedicine

XSL•FO
RenderX

Introduction

The health care paradigm is shifting from provider-centered care to patient-centered care. Although patient-centered care is not a new concept, it has gained renewed attention due to factors such as the rise of digital health and the increasing use of wearable devices [1,2]. Patient-generated health data (PGHD) has emerged as a key tool [3] that can allow the real-time collection of patient-reported symptoms and biometric data, tailored to individual needs [4]. This capability is particularly impactful for chronic disease management, where continuous monitoring of patients' health status allows for timely interventions, ultimately reducing health care costs and improving clinical outcomes [5].

One key implication of this shift is the link between PGHD and reliable clinical information, such as electronic health records (EHRs). When integrated with EHRs, PGHD offers a more comprehensive view of a patient's health status [6,7]. Although self-reported health and treatment histories are widely needed in clinical fields [8,9], challenges remain in terms of the effective integration and use of PGHD in health care institutions [6,10,11]. A critical factor in this integration is the attitude and decision-making of physicians [12,13].

Incorporating PGHD for meaningful patient participation requires the support and understanding of health care professionals, along with strong legal and systematic frameworks. However, previous studies have indicated that issues such as a lack of integration with EHRs, technostress, time pressure, and workflow-related challenges can exacerbate physician workload [6,11]. In addition, legal challenges and concerns regarding data reliability and quality remain barriers to clinical practice [14,15]. Although addressing and understanding these factors are essential for the successful integration of PGHD, research on physicians' perceptions of PGHD, clinical applicability, and secondary uses (eg, research, commercial, and industrial purposes) remains limited [13,16]. A recent comprehensive review of health care professionals' attitudes toward PGHD published in 2025 identified only 33 relevant studies conducted over the past decade, highlighting the limited empirical evidence in this area [17].

Therefore, we conducted a qualitative and quantitative analysis of physicians' perceptions of the clinical use of PGHD through a web-based survey and in-depth interviews. The purpose of this study was to identify ways to improve data integration, suggest policy directions, and guide the safe and effective use of PGHD in clinical practice.

Methods

Study Design

This study adopted a mixed methods approach integrating a cross-sectional survey with in-depth interviews. The study followed an explanatory sequential mixed methods design, in which quantitative survey data were collected and analyzed first to identify general patterns in physician experiences and perceptions of PGHD. Subsequent qualitative interviews were informed by the survey findings and conducted to explore the underlying reasons, contextual factors, and practical interpretations of these quantitative results. Integration of quantitative and qualitative data occurred at the interpretation stage, and qualitative findings were used to contextualize and enrich the survey results rather than to independently generate or test new hypotheses.

Survey Instrument Development

The questionnaire was designed to investigate physicians' perspectives on the use of PGHD for various purposes, including consent processes, expectations, concerns, and suggestions for improvement. The development of the questionnaire involved direct discussion and consultation with a multidisciplinary panel of experts, including a medical regulatory and legal expert, medical informatics experts, a PGHD research specialist, and a health IT industry professional. The conceptual domains were informed by previous research conducted by the study team that examined patient perspectives on the use of PGHD; however, the physician survey instrument was independently developed, and no items were reused verbatim.

The finalized questionnaire comprised the following categories:

1. General Information (affiliated institution, medical specialty, years of clinical experience)
2. Experience with PGHD in clinical practice
3. Perspectives on using PGHD for clinical purposes
4. Perspectives on using PGHD for nonclinical purposes (eg, research or public health surveillance)
5. Perspectives on improving PGHD regulations

Response formats for each survey item employed various formats tailored to the specific information being sought. These formats included 5-point Likert scales (1=Not possible at all/Not helpful at all to 5=Always possible/Very helpful), yes/no responses, single-choice selections, multiple-choice selections with prioritization, and open-ended responses.

Survey Respondents and Recruitment

A web survey was conducted from November 6 to 12, 2023, targeting physicians who currently use or have the potential to use PGHD. Recruitment was conducted using a structured approach during a predefined survey period. Study announcements describing the study purpose, eligibility criteria, and survey access link were distributed through official communication channels of national and specialty-specific associations and professional societies (eg, the Korean Association of Private Practitioners and the Korean Association of Geriatric Hospitals) and relevant societies (eg, the Korean Society of Medical Informatics and the Korean Society for Patient Safety).

Physicians accessed the survey via a web or mobile link or QR code and first received an electronic informed consent form. Only those who provided their consent were able to proceed with the questionnaire. Participation was voluntary and, upon completion of the survey, participants received a modest incentive (approximately US $8) as compensation for their time. Duplicate submissions were identified and removed during data cleaning, and responses from individuals who did not meet the

XSL•FO
RenderX

eligibility criteria (eg, nonphysician affiliations) were excluded prior to analysis.

Participants were categorized into three distinct subgroups according to their respective work environments.

1. Primary and secondary health care facilities: (1) clinics with fewer than 30 beds and (2) hospitals with more than 30 but fewer than 100 beds
2. Tertiary and quaternary health care facilities: (1) general hospitals with more than 100 beds and (2) tertiary general hospitals
3. Nursing homes

Because this study was designed as an exploratory cross-sectional survey without a single primary hypothesis for formal hypothesis testing, a pragmatic sample size approach was adopted rather than a strict power-based calculation. Considering the planned descriptive analyses and multivariable exploratory analyses, our aim was to include at least 150 completed responses in the final analysis, which was considered sufficient to provide stable estimates of key survey items and to allow preliminary subgroup comparisons [18]. Assuming an analyzable response rate of approximately 75%, 200 health care professionals were approached to achieve the target final sample size.

During the survey period, 164 responses were collected. After excluding duplicate responses (n=4) and ineligible respondents (nonphysicians, n=3), the final analytical sample consisted of 157 participants. To minimize selection and response bias, the survey was administered anonymously using standardized procedures, and participants were recruited through multiple professional societies. For qualitative analysis, independent coding and reflexive discussions between researchers were used to minimize interpretive bias.

Data Analysis

The survey responses were analyzed to identify general trends and exploratory differences between physicians working in tertiary hospitals and those working in other health care settings. Binary or categorical variables were analyzed using chi-square tests. Likert-scale items were treated as ordinal variables and analyzed using the Wilcoxon rank-sum test, a nonparametric method appropriate for ordered response data that does not assume normality. Continuous variables were also analyzed using the Wilcoxon rank-sum test. Statistical significance was defined as a 2-sided P value of <.05.

For priority ranking items, responses were weighted according to rank (3 points for first priority, 2 points for second priority, and one point for third priority), and the weighted scores were summed across respondents. All statistical analyses were performed using R (version 4.3.0; R Foundation for Statistical Computing). Subgroup analyses were conducted for exploratory purposes to identify descriptive patterns rather than estimate independent institutional effects; accordingly, analyses were not adjusted for demographic variables such as age, sex, or years of clinical experience.

Given the exploratory nature of the study, no formal adjustments for multiple tests were applied. The findings were interpreted on the basis of the direction of the effect, consistency across related items, and integration with qualitative insights rather than on statistical significance alone. For key single-choice outcomes, proportions are presented with 95% CIs calculated using the Wilson score method. CIs were not applied to descriptive multiple-response items. Missing data were minimal and were handled using complete-case analysis, with only respondents providing data for relevant variables included in each analysis, and no imputation was performed.

In-Depth Interviews

After completion of the survey, in-depth interviews were conducted using a semistructured interview guide developed based on the survey questionnaire and key quantitative findings. The interviews were designed to explore why physicians perceived specific barriers, expectations, and concerns about the use of PGHD identified in the survey, rather than to compare perspectives across medical specialties.

Four individual interviews and one joint interview were conducted between August and November 2024. The joint interview involved 2 physicians of the same medical specialty working in different health care settings and was conducted to facilitate discussion between institutional contexts. Each interview lasted approximately 40-60 minutes and was conducted in person or via video conferencing, depending on participant availability and preference.

The interview participants were selected based on prior experience with PGHD in clinical practice, extensive clinical experience, nonparticipation in the survey (for triangulation), and diverse clinical backgrounds. The interviews were conducted in Korean by 2 members of the research team (YL and YEP), both trained in qualitative research and medical informatics. Neither interviewer had a prior personal supervisory relationship with the participants. All interviews were recorded with the consent of the participants and transcribed verbatim in Korean. The translated excerpts presented in the manuscript were translated into English during the preparation of the manuscript.

Qualitative data were analyzed using an inductive reflexive thematic analysis approach. Two researchers (YL and YEP) independently reviewed the transcripts and conducted an initial open coding to identify meaningful units related to experiences, perceived barriers, facilitators, and contextual factors associated with the use of PGHD. The codes were iteratively discussed and refined in team meetings, and higher-level themes were developed through consensus. Reflexive discussions were conducted throughout the analytic process to minimize interpretive bias and to ensure that the themes were grounded in the data. Given the exploratory and complementary role of the qualitative component, formal assessment of thematic saturation was not a primary analytic goal; interviews were conducted until sufficient explanatory depth was achieved to contextualize key quantitative findings.

Ethical Considerations

This study was approved by the Institutional Review Board of Asan Medical Center, Korea (Institutional Review Board 2023-1363) and conducted in accordance with relevant ethical

XSL•FO
RenderX

guidelines. Informed consent was obtained from all participants before their involvement, with assurances of anonymity and confidentiality. Participants were briefed on the study's objectives and the intended use of the collected data. Participants received approximately US $8 in compensation for their participation.

Results

Cross-Sectional Survey

A total of 157 physicians participated in the survey, with 81 respondents (51.6%) from a tertiary general hospital and 76 (48.4%) from other health care institutions. The 2 groups differed significantly in demographic makeup (Table 1). Additionally, we identified participants' specialties, with the total sample consisting of 29 respondents from internal medicine, 17 from neurology, 16 from family medicine, 12 from emergency medicine, 15 from surgery, and 68 from other departments.

Table 1. Characteristics of respondents by type of health care institution.

Variables	Tertiary general hospital (81/157, 51.6%), n (%)	Other health care institutions[a] (76/157, 48.4%), n (%)	Total (N=157), n (%)	*P* value
Sex				.01
Male	32 (39.5)	45 (59.2)	77 (49)	
Female	49 (60.5)	31 (40.8)	80 (51)	
Age				<.01
20s	25 (30.9)	5 (6.6)	30 (19.1)	
30s	39 (48.1)	24 (31.6)	63 (40.1)	
40s	16 (19.8)	41 (53.9)	57 (36.3)	
≥50 years	1 (1.2)	6 (7.9)	7 (4.5)	
Length of work experience (years)				.01
<2	11 (13.6)	7 (9.2)	18 (11.5)	
2-4	27 (33.3)	11 (14.5)	38 (24.2)	
5-9	20 (24.7)	19 (25)	39 (24.8)	
≥10	23 (28.4)	39 (51.3)	62 (39.5)	
Type of health care institution				
General hospital	—[b]	40 (52.6)	—	—
Clinic	—	24 (31.6)	—	—
Nursing home	—	12 (15.8)	—	—

[a]Other health care institutions include primary and secondary health care facilities and nursing homes.

[b]Not applicable.

The demographic data collected were analyzed based on the length of work experience, categorizing the respondents into 2 groups: those with less than 5 years of experience (56/157, 35.7%) and those with 5 years or more (101/157, 64.3%). A comparison between the groups revealed significant differences in sex (*P*=.03), age range (*P*<.01), and type of affiliation (*P*<.01).

Single-Choice and Multiple-Response Analysis

Overall, 64% (100/157) of the surveyed physicians reported having used PGHD in their clinical practice (Table 2). This proportion was almost identical between tertiary hospital doctors and those in other settings (52/81, 64.2% vs 48/76, 63.2%; *P*=.89), with no significant difference by institution type. Among the 100 physicians who had used PGHD, the most common format in which patients provided data was a handwritten paper record, reported by 65% (65/100). Physicians who had not used PGHD (57/157, 36.3%) were asked for the main reason. The vast majority (45/57, 78.9% overall) indicated that they had simply never received self-recorded health data from any patient. A few physicians (5/57, 8.8% overall) reported that patients had offered such data, but it was not helpful for treatment, and a similar fraction (5/57, 8.8%) said they had seen patient data but did not have enough time to review it during visits. There was a noticeable but not significant difference in these reasons between institution types (*P*=.06).

XSL•FO
RenderX

Table 2. Physician experiences with patient-generated health data (PGHD) by type of health care institution.

Questions	Tertiary general hospital (81/157, 51.6%)	Other hospitals[a] (76/157, 48.4%)	*P* value	Total (N=157)
B1. Have you used health information recorded by patients during your clinical practice (other than verbal symptom descriptions or records from other medical institutions)?			.89	
Yes[b], n (%, 95% CI)	52 (64.2, 53.3-73.8)	48 (63.2, 51.9-73.1)		100 (64, 55.9-70.8)
No, n (%)	29 (36)	28 (37)		57 (36)
B1a. If you answered yes, in what format were the records provided by the patient? (Select all that apply) (n=100)[c], n (%)				
Handwritten records on paper	31 (59.6)	34 (70.8)	.51	65 (65)
Records printed on paper	17 (32.7)	21 (43.8)	.43	38 (38)
Records of webpages or apps connected to electronic health records (EHR)	18 (34.6)	13 (27.1)	.55	31 (31)
Patient's smartphone/tablet/laptop	12 (23.1)	14 (29.2)	.70	26 (26)
Other	0	1 (2.1)	.97	1 (1)
B1b. If you answered yes, what information did the patient provide in the records? (Select all that apply) (n=100), n (%)				
Measurements include body temperature, blood pressure, or blood sugar	34 (65.4)	33 (68.8)	.98	67 (67)
Continuous records of symptoms (eg, pain, nausea, defecation, or urination)	27 (51.9)	26 (54.2)	>.99	53 (53)
Continuous records of health care/self-care (eg, medications, wound care, exercise, or diet)	23 (44.2)	18 (37.5)	.62	41 (41)
Physical measurements such as height, weight, or BMI	19 (36.5)	20 (41.7)	.82	39 (39)
Vital sign information includes oxygen saturation, pulse rate, or electrocardiogram	13 (25)	10 (20.8)	.78	23 (23)
Other	3 (5.8)	1 (2.1)	.66	4 (4)
B1c. If you answered no , what is the reason? (n=57), n (%)			.06	
I have never received self-records from patients	27 (93.1)	18 (64.3)		45 (78.9)
I have received them, but they were not helpful for treatment	1 (3.4)	4 (14.3)		5 (8.8)
I have seen them, but time was insufficient to review them	1 (3.4)	4 (14.3)		5 (8.8)
Other	0	2 (7.1)		2 (3.5)
B2. Does your health care organization offer or operate patient health records (smartphone app/webpage)?			<.01	
Yes, n (%, 95% CI)	19 (23.5, 15.6-33.8)	9 (11.8, 6.4-21.0))		28 (17.8, 12.6-24.6)
No, n (%)	21 (25.9)	53 (69.7)		74 (47.1)
I do not know, n (%)	41 (50.6)	14 (18.4)		55 (35)

[a]Other health care institutions include primary and secondary health care facilities and nursing homes.

[b]95% CIs are provided only for key single-choice outcomes.

[c]For multiple-response items (B1a and B1b), the percentages are based on the respondents who reported prior use of PGHD, and the CIs are not shown to maintain clarity.

The most frequently reported PGHD were home-recorded physiological measurements, such as body temperature, blood pressure, or blood sugar (67/100, 67%). Symptom diaries (eg, logs of pain, nausea, or other symptoms) were the next most common (53/100, 53%), followed by records of self-care activities such as medication adherence, wound care, exercise, or diet (41/100, 41%). About 39% (39/100) mentioned physical metrics such as body weight or BMI. Only a small share (23/100, 23%) had encountered continuous vital sign data (eg, oxygen saturation or heart rate trends) from patients. Notably, there were no significant differences by hospital type in the kinds of PGHD provided by patients (all *P*>.5).

Regarding the institution-derived patient health record (PHR) system (eg, a smartphone app or a web portal for patients), there was a significant disparity: Tertiary hospital physicians were more likely to report that such a system existed (19/81, 23.5% vs 9/76, 11.8% in other institutions), and far fewer of them stated outright that no patient portal was available (21/81, 25.9% vs 53/76, 69.7%; *P*<.01). However, awareness of these systems

XSL•FO
RenderX

differed: Over half of tertiary hospital respondents (41/81, 50.6%) said they did not know whether their organization had a PHR app, whereas only 18.4% (14/76) of physicians in other settings were unsure.

In addition, physicians in both groups identified similar barriers to incorporating PGHD into current practice. In the category of survey exploring perspectives on the use of PGHD in clinical practice, no statistically significant differences were observed between respondents from tertiary general hospitals and other hospitals in all items analyzed ($P>.05$) (Table 3).

Despite the lack of significant differences, both groups showed similar trends in their responses. When asked about the main barriers to the integration of PGHD collected in nonclinical settings with the EHR, the most common reason cited in both groups was system-based limitations (C1a). This represented 27.2% (22/81) of respondents from tertiary hospitals, 35.5% (27/76) from other hospitals, and 31.2% (49/157) overall. The most significant motivation for enabling PGHD–EHR integration was identified as the need/demand for PGHD utilization (C1b). This was reported by 48.1% (39/81) of respondents from tertiary hospitals, 40.8% (31/76) from other hospitals, and 44.6% (70/157) overall.

Furthermore, the most frequently cited reason for the benefits of PGHD utilization was its high value for providing meaningful insights into patients' health conditions (C2b). This was reported by 43.1% (28/65) of tertiary hospital respondents, 25.5% (14/55) of other hospital respondents, and 35% (42/120) overall. By contrast, the most commonly cited drawback of PGHD utilization was its unintended negative effects, such as prolonged consultations or inaccurate decision-making (C2a). This accounted for 38.3% (31/81) of responses from tertiary hospitals, 43.4% (33/76) from other hospitals, and 40.8% (64/157) overall.

XSL•FO
RenderX

Table 3. Distribution of responses on intent to use patient-generated health data (PGHD) for practice by type of health care institution.

Questions	Tertiary General Hospital (81/157, 51.6%), n (%)	Other hospitals[a] (76/157, 48.4%), n (%)	*P* value	Total (N=157), n (%)
C1a. In the current health care setting, what is the main reason it is difficult to use PGHD (produced and recorded in nonclinical settings) in combination with medical records?			.21	
System-based: Not available for use during treatment (lack of integration of PGHD-EHR[b] or provision of summary/visualized dashboard)	22 (27.2)	27 (35.5)		49 (31.2)
Time/cost: Absence of relevant compensation (fees) and lack of treatment time	18 (22.2)	20 (26.3)		38 (24.2)
Rationale/accountability: Unclear accountability for PGHD-based decisions or lack of guidance or guidelines to make judgments based on PGHD	20 (24.7)	16 (21.1)		36 (22.9)
Current state of utilization: Little or no PGHD is collected or provided	13 (16)	7 (9.2)		20 (12.7)
Absence of need/demand: No clinically valuable information is available	4 (4.9)	6 (7.9)		10 (6.4)
Other	4 (4.9)	0 (0)		4 (2.5)
C1b. In the current health care setting, what is the biggest motivation to use PGHD in combination with medical records?			.58	
Demand: High necessity/high willingness (of physicians and/or patients) to use	39 (48.1)	31 (40.8)		70 (44.6)
System-infrastructure: PGHD that is well integrated into the information system or dashboards that provide easy-to-understand summaries	15 (18.5)	23 (30.3)		38 (24.2)
Time/cost: Reward (fees) or there is sufficient time	9 (11.1)	10 (13.2)		19 (12.1)
Current utilization environment: Many patients are already collecting PGHD	11 (13.6)	7 (9.2)		18 (11.5)
Rationale/accountability: The use of PGHD was recommended in guidelines set by academic societies, treatment-related organizations, or colleagues	6 (7.4)	4 (5.3)		10 (6.4)
Other	1 (1.2)	1 (1.3)		2 (1.3)
C2a. (This question asks for your opinion regardless of your response to the previous question.) If it is not helpful, what is the main reason? (N=157)			.82	
Unintended negative effects: Incorrect decisions were made, or unnecessary conversations with patients were prolonged	31 (38.3)	33 (43.4)		64 (40.8)
Lack of means: Information was overwhelming or unavailable during treatment	27 (33.3)	25 (32.9)		52 (33.1)
Rationale: Lack of evidence (research results) to make a judgment based on PGHD	15 (18.5)	9 (11.8)		24 (15.3)
Lack of value: No clinically valuable information	6 (7.4)	7 (9.2)		13 (8.3)
Other	2 (2.5)	2 (2.6)		4 (2.5)
C2b. If it was helpful, what was the main reason? (n=120)[c]			.22	
High value: PGHD is inherently valuable for understanding patient health conditions and making treatment decisions	28 (43.1)	14 (25.5)		42 (35)
Indirect positive effects: Improved trust relationships with patients and improved adherence to treatment	16 (24.6)	20 (36.4)		36 (30)
Synergies: There are more positive effects or efficiencies when the EHR and PGHD information is integrated	16 (24.6)	15 (27.3)		31 (25.8)
There is evidence (research results) to make judgments based on PGHD	5 (7.7)	6 (10.9)		11 (9.2)
Other	0 (0)	0 (0)		0 (0)

[a]Other health care institutions include primary and secondary health care facilities and nursing homes.

[b]EHR: electronic health record.

[c]For C2b, the percentages are calculated based on respondents who reported that PGHD was helpful (tertiary hospitals: n=65; other hospitals: n=55; total: n=120).

We assessed the feasibility of using PGHD integrated with EHR for treatment planning and clinical decision-making under current health care conditions. As shown in Figure 1, “Occasionally possible” was the response selected most frequently (71/157, 45.2%), whereas 55.4% (87/157) of respondents provided positive responses (C1). No statistically significant difference was observed between tertiary hospitals

XSL•FO
RenderX

and other hospitals (*P*=.61) based on the Wilcoxon rank-sum test.

Figure 1. Distribution of survey responses to C1 and C2 questions (N=157). PGHD: patient-generated health data.

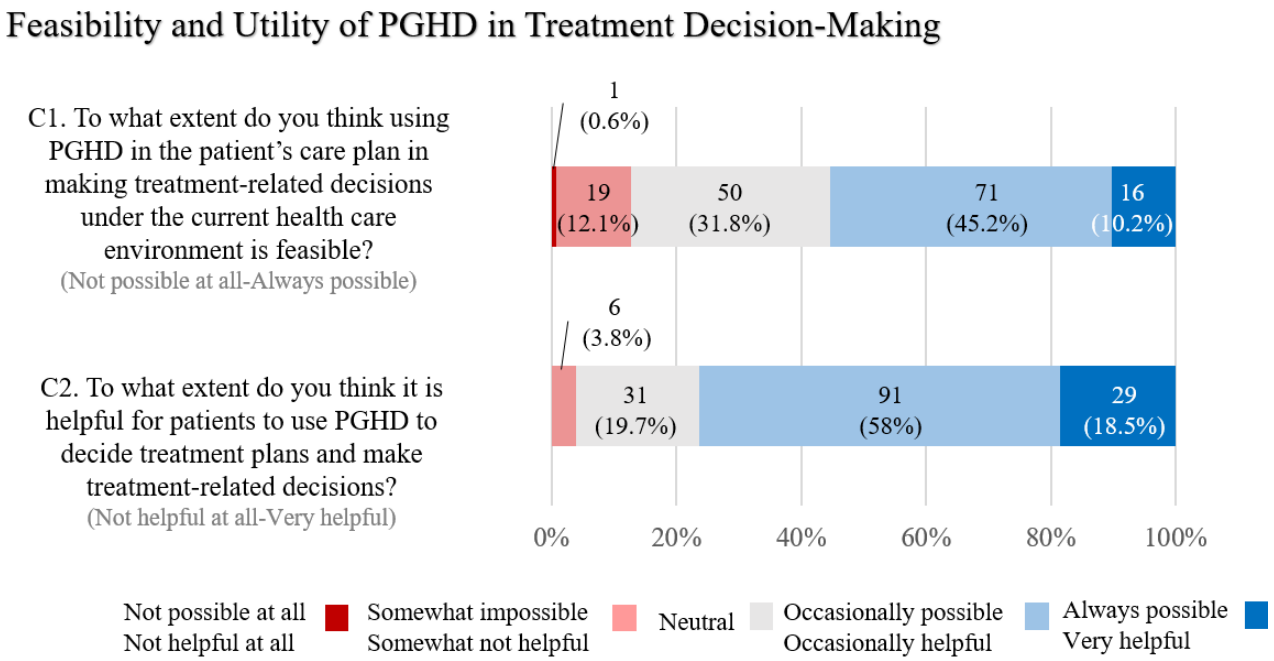

Regarding the perceived benefits of the use of PGHD, the most frequently selected response was "Occasionally helpful" (91/157, 58%). Overall, 76.4% (120/157) of respondents expressed positive responses, including "Occasionally helpful" and "Highly helpful" responses (C2). Similarly, no statistically significant difference was identified between tertiary hospitals and other hospitals (*P*=.08), as determined by the Wilcoxon rank-sum test.

Looking beyond direct clinical use, there was a clear preference for using PGHD for public good and knowledge generation rather than commercial purposes. The highest-ranked secondary use was academic, that is, using PGHD for research, education, or developing clinical guidelines, with a total score of 329. The second priority was public health and national statistics, with a score of 270, highlighting that many physicians see value in aggregating PGHD to inform population health monitoring or policy. By comparison, industry-related uses (eg, technology development, drug discovery, or algorithm training) received moderate support, with a score of 174. Commercial purposes, such as recommending personalized health services and providing personalized advertisements, were less favored, with a score of 164. The least prioritized use of PGHD was the trade and sale of health information data, with a score of 5.

To enable these uses of PGHD, respondents highlighted several areas for improvement in governance and infrastructure (Table 4, Question D2). The top priority was enhanced data security and personal information protection, with a weighted score of 263. This aligns with the expectation that if PGHD is to be widely used, especially outside one-on-one care, robust safeguards must be in place to protect patient privacy. The second priority was ensuring transparency, with a score of 219, followed by improvement of the consent system, with a score of 205. For regulatory or policy changes to facilitate the use of PGHD (Table 4, Question E1), 2 clear frontrunners emerged: data standardization, with a score of 204, and increased security obligations for data management companies, vendors, and government support policies, with a score of 196.

XSL•FO
RenderX

Table 4. Expectations, concerns, and improvement requirements for the use of patient-generated health data (PGHD): priority response items for all respondents (N=157).

Questions and rank/response items	Total weighted score[a]
C3. What are your main concerns about using PGHD for medical treatment?	
1. Issues related to data accuracy.	322
2. Increased clinical workload.	164
3. Problems with confidence in a clinical setting and a lack of evidence for data analysis results.	164
4. Lack of guidelines on recording PGHD by integrating with medical records.	139
5. Accountability issues for PGHD-based decisions.	73
6. Issues related to the protection and security of personal information.	70
D1. From a physician's perspective, which of the following purposes is the most important and necessary when using PGHD for purposes other than medical treatment?	
1. Academic purposes, such as research, education, and practice guideline development.	329
2. Public interest purposes, such as public health and national statistics.	270
3. Industrial purposes include technology development, drug discovery, and algorithm development.	174
4. Commercial purposes, such as recommending personalized health services and providing personalized advertisements.	164
5. Trade and sale of health information data.	5
D2. What are the most pressing areas of improvement to use PGHD for nonclinical purposes?	
1. Enhanced security/personal information protection.	263
2. Securing transparency.	219
3. Improvement of the consent system.	205
4. Evaluation/compensation.	157
5. Preparation of verification measures for user reliability.	98
E1. Related regulations need to be improved for the use of PGHD for treatment or research. Which of the following should be prioritized for revision?	
1. Data standardization.	204
2. Increased security obligations for data management companies/vendors and government support policies.	196
3. Clarification of the right to information transfer in the Medical Service Act.	138
4. Preparation of an article on exceptions for not sending information despite a patient's request.	119
5. Incentives for related maintenance costs of the system.	82
6. Support for follow-up response policies in the event of an incident, such as a data breach.	80
7. Regulations for securing interoperable systems.	65
8. Technologies for data pseudonymization/deidentification.	58

[a]Total weighted score was calculated by assigning 3 points to first-priority selections, 2 points to second-priority selections, and 1 point to third-priority selections, and summing the points across all respondents.

In-Depth Interviews

In-depth interviews included a total of 6 participants, consisting of 3 female and 3 male physicians. All participants had over 10 years of clinical experience and were affiliated with either tertiary referral hospitals (n=4) or clinics (n=2). Their specialties were diverse, including emergency medicine (n=1), breast surgery (n=1), critical care/trauma surgery (n=1), geriatrics (n=1), and pediatrics (n=2). The 2 pediatricians participated in a joint interview, whereas the remaining 4 physicians underwent individual interviews (Table 5).

XSL•FO
RenderX

Table 5. In-depth interviewee characteristics.

ID	Affiliation	Specialty	Sex
P1	Clinic (less than 30 beds)	Emergency medicine	Female
P2	Tertiary referral hospital	Surgical oncologist (breast cancer)	Female
P3	Tertiary referral hospital	Critical care/trauma	Male
P4	Tertiary referral hospital	Geriatrics	Male
P5	Clinic (less than 30 beds)	Pediatrics	Male
P6	Tertiary referral hospital	Pediatrics	Female

In qualitative interviews, all participating physicians reported prior experience using PGHD in clinical practice, primarily in the context of follow-up care and chronic disease management. PGHD was commonly integrated into outpatient workflows through patient-maintained records, such as symptom diaries, home-measured physiological data, and self-care behavior logs. Physicians noted that the quality and availability of PGHD were often influenced by patient characteristics, including personality traits and, in some cases, discrepancies between patient and caregiver-reported data, particularly in geriatric populations. In tertiary care settings, PGHD also included externally generated data, such as laboratory results from other institutions or patient-submitted images.

Physicians consistently identified several barriers to the use of PGHD. The challenges cited most frequently included limited consultation time and workflow constraints, which restricted the ability to review patient-provided data in detail. Concerns about data reliability and device accuracy were also prominent, as were issues related to excessive data volume and a lack of quality control. Furthermore, factors related to the patient, particularly limited digital literacy among older adults, were perceived as significant obstacles to the effective use of PGHD.

Despite these challenges, participants emphasized multiple facilitators that support the use of PGHD. Clinical need was described as the main driver, particularly for conditions that require continuous monitoring beyond episodic clinical encounters. Physicians highlighted the role of PGHD in enabling remote management, enhancing patient safety, and supporting shared decision-making. The availability of evidence-based guidelines and accountability requirements further reinforced the use of PGHD. Importantly, participants underscored the high value of PGHD as a source of real-world data that complements traditional EHR data by providing longitudinal insights. Additional benefits included improved patient engagement and motivation, as well as synergistic integration with existing clinical information systems.

Physicians expressed a range of concerns about the clinical use of PGHD, with data accuracy identified as the most critical issue. Participants emphasized that variability in the patient context and measurement conditions could compromise the reliability of PGHD. The increased clinical workload associated with the review and interpretation of patient-generated data was also a major concern, along with the potential legal and ethical liabilities when clinical decisions are based on uncertain data sources.

Beyond direct clinical care, physicians identified academic research as the most important secondary use of PGHD, followed by contributions to the public interest, such as informing health policy and population-level statistics. Industry applications were considered of relatively lower priority. Participants highlighted the potential of PGHD to generate real-world evidence and to supplement existing data sources, particularly in underrepresented populations, such as pediatric patients.

Regarding nonclinical use, physicians emphasized the need for improvements in data security, transparency, and consent processes. Ensuring robust data protection and clear communication about how data are used was considered essential for building trust. Participants also stressed the importance of establishing mechanisms to verify the trustworthiness of the organizations that handle PGHD.

In terms of regulatory priorities, participants highlighted the importance of improving patient data portability and strengthening requirements for data security and standardization to support interoperability. Several physicians also emphasized the need to simplify consent procedures, noting that available processes are often too complex and can hinder the practical use of PGHD in both clinical and research settings.

Integrated Findings

The qualitative findings largely converged with the survey results and provided additional context for the main quantitative patterns. Table 6 presents a joint display integrating the quantitative and qualitative findings.

XSL•FO
RenderX

Table 6. Joint display integrating quantitative and qualitative findings on physicians' perceptions of patient-generated health data (PGHD).

Domain	Quantitative findings	Qualitative findings	Integrated interpretation
Current use and format of PGHD	Overall, 64% of physicians reported prior use of PGHD, with no significant institutional difference (64.2% in tertiary hospitals vs 63.2% in other institutions). Handwritten paper records were the most common format, and home-recorded physiological measurements were the most frequently encountered type of PGHD.	Interview participants reported using PGHD mainly in follow-up care and chronic disease management through symptom diaries, home-measured physiological data, and self-care behavior logs. In tertiary settings, PGHD also included externally generated data such as laboratory results from other institutions or patient-submitted images.	PGHD is already used in routine practice, but mostly through clinician-mediated and low-integration workflows rather than seamless digital systems.
Barriers to clinical use	The most commonly reported barrier to using PGHD with medical records was system-based limitation, including the lack of PGHD-EHR[a] integration or the absence of summary/dashboard tools. Other frequently reported concerns included unintended negative effects, information overload, lack of evidence, data accuracy, and increased clinical workload.	Interviewees emphasized limited consultation time, workflow constraints, data overload, device accuracy concerns, lack of quality control, and patient-level barriers such as older adults' limited digital literacy and discrepancies between patient- and caregiver-reported data.	Survey-identified implementation barriers were reinforced by interview findings showing that PGHD use is constrained not only by technical limitations but also by workflow- and patient-level factors in routine care.
Facilitators and perceived clinical value	The strongest motivation for PGHD use was clinical necessity or willingness to use PGHD, followed by user-friendly system integration. Physicians most often viewed PGHD as helpful because it provided valuable information for understanding patient health conditions and informing treatment decisions.	Interview participants likewise described clinical need as the main driver, especially for conditions requiring continuous monitoring beyond episodic encounters. They also highlighted PGHD's role in remote management, patient safety, shared decision-making, patient engagement, and synergistic use with existing clinical information systems.	PGHD is perceived as most useful when there is a clear clinical purpose and when the information can be meaningfully incorporated into care workflows.
Secondary use of PGHD	For nonclinical purposes, academic use ranked highest, followed by public-interest use, such as public health and national statistics. Industry-related use received lower support, and trade or sale of health data was the least acceptable.	Interviewees similarly prioritized research and public-interest uses over industry applications and described PGHD as a source of real-world longitudinal data, particularly useful for underrepresented groups such as pediatric patients.	Quantitative and qualitative findings converged in showing that physicians support secondary use of PGHD primarily when it serves knowledge generation or public benefit rather than commercial exploitation.
Governance and regulatory priorities	For nonclinical use, the top priorities were stronger security, greater transparency, and improved consent systems. For regulatory improvement, the highest priorities were data standardization and stronger security obligations for organizations handling PGHD.	Interviewees stressed the need for robust data protection, transparent communication about data use, trust in the organizations handling PGHD, improved portability and interoperability, and simpler consent procedures for practical clinical and research use.	Broader adoption and secondary use of PGHD depend not only on technical infrastructure but also on trustworthy governance frameworks that support security, transparency, interoperability, and practical consent processes.

[a]EHR: electronic health record.

Discussion

Principal Results

This study explored the perceptions of physicians about the clinical and nonclinical use of PGHD. Our results suggest that although PGHD has the potential to play an important role in the health data ecosystem, its adoption is hindered by practical barriers. Physicians acknowledged the utility of PGHD for remote patient management and treatment planning but raised concerns about data reliability, excessive data volume, system integration challenges, and time and cost burdens. For nonclinical purposes, PGHD was perceived as more valuable for academic research and, to a lesser extent, for public-interest applications, such as policymaking. To support these uses, physicians emphasized the need for greater security, standardized data management, and clear regulatory frameworks.

Although the majority (100/157, 64%) of surveyed physicians reported having used PGHD in their clinical practice, the most common format was handwritten paper records (65/100, 65%). There was a significant disparity in the reported availability of institutional patient portals, with physicians in other institutions much more likely to report that no such system was available (53/76, 69.7% vs 21/81, 25.9%; $P<.01$). Nevertheless, relatively few physicians (31/100, 31%) had received PGHD through digital means, which was similar in both the tertiary and nontertiary groups (no significant differences, $P>.4$ for all). Moreover, awareness of these systems differed; over half of the tertiary hospital respondents (41/81, 50.6%) said they did not know whether their organization had a PHR app, whereas only 18.4% (14/76) of physicians in other settings were unsure. This pattern suggests that tertiary centers are more likely to have some PGHD infrastructure, but many staff may not be aware of it, implying that a significant proportion of health care

XSL•FO
RenderX

professionals lack awareness of PHR utilization. This suggests, even before the data quality issues discussed later, that the amount of information available to physicians in digital form is relatively small, and physicians' motivation to use PHR is limited.

Physicians in both groups noted system-related limitations, specifically, the lack of integration of patient data into the EHR or the absence of convenient tools (eg, summary dashboards) to review PGHD during clinical care. In contrast, relatively few respondents felt that the lack of PGHD was the main issue, only 12.7% (20/157) said the current lack of PGHD availability was the top barrier, and merely 6.4% (10/157) believed there was no need for or no valuable information in PGHD, with no significant difference by hospital type (P=.21). This consensus highlights that technical integration and resource constraints are seen as key hurdles to using PGHD in practice, rather than a lack of data or interest. The US Office of the National Coordinator for Health IT has outlined a modular architecture for PGHD integration, recommending the use of intermediary platforms (eg, data brokers or gateways) to facilitate data flow [19]. Additionally, the US Office of the National Coordinator for Health IT emphasizes the incorporation of clinical decision support tools to enable efficient triage and interpretation of patient-generated data within clinical workflows [19]. When considering what would most motivate or enable greater PGHD use, a clear clinical need or willingness for PGHD use emerged: respondents said that a high perceived need for PGHD (and strong willingness to use it) would be the biggest driver for adoption. Far fewer respondents prioritized financial or time incentives. It is worth mentioning, however, that physicians from smaller hospitals showed a trend toward valuing system improvements slightly more (23/76, 30.3% chose "well-integrated system" as the top motivator vs 15/81, 18.5% of those in tertiary hospitals), whereas tertiary hospital physicians were somewhat more likely to emphasize intrinsic need (39/81, 48.1% vs 31/76, 40.8% in others). Overall, however, both groups ranked clinical need the highest, followed by better systems, echoing a shared view that demand and infrastructure are pivotal for PGHD implementation.

Regarding the reasons why incorporating PGHD might not be helpful, the predominant concern (64/157, 40.8%) was the risk of unintended negative consequences. This included scenarios such as making incorrect clinical decisions based on inaccurate patient data or wasting time in protracted discussions about patient-collected information that would not benefit care. This is consistent with the C3 priority scores, in which physicians' top concerns about using PGHD in practice were clearly centered on data quality. The next major reason was a lack of practical means to use the data. In the same vein as the second priority (ie, workflow implications), it might be overwhelming or not accessible in the right format at the point of care. In essence, without better tools or workflows, PGHD could be more of a hindrance than a help [11]. Notably, data privacy and security were the lowest ranked of the provided concerns, though the top priority was enhancing data security and personal information protection (Question D2; weighted score 263). This suggests that, in the context of day-to-day practice, clinicians are currently more focused on practical implementation challenges (accuracy, workload, and guidelines) than on legal and privacy ramifications. Given these practical constraints, health systems must establish formalized data quality assurance protocols and incorporate PGHD through clinically embedded infrastructures. The Veterans Health Administration, for instance, has adopted standardized operational procedures to support consistent data acquisition and interpretation across settings, whereas recent informatics frameworks advocate modular integration architectures and decision support functionalities aligned with existing clinical workflows [20,21].

Meanwhile, when asked why using PGHD would be helpful, physicians most frequently cited the direct informational value of PGHD. More than one-third (42/120, 35%) indicated that PGHD provides important insights into a patient's health status and can inform better clinical decision-making. The highest-ranked secondary use (Question D1) was academics, that is, the use of PGHD for research, education, or developing clinical guidelines, which suggests that PGHD is expected to contribute to personalized care. Many also recognized indirect benefits: 30% (36/120) noted that engaging with PGHD can improve the physician–patient relationship, build trust, and improve patient engagement or adherence to treatment. Additionally, 25.8% (31/120) highlighted efficiency or synergy gains when PGHD is integrated with clinical systems. This refers to scenarios in which having patient data seamlessly integrated into EHRs could enhance monitoring or enable more personalized, timely interventions, thereby improving care processes. However, seamless care requires interoperability between data and systems. This is consistent with the findings of the E1 survey, which emphasized the importance of data standards. Physicians recognize the need for common standards and interoperability so that PGHD can be seamlessly integrated into medical records across different platforms and institutions. The second priority was strengthening data management policies, specifically by increasing security obligations for companies handling health data and providing governmental support. This echoes the earlier emphasis on security and indicates a desire for clearer regulations holding vendors accountable for protecting PGHD. This aligns with earlier findings: tackling technical integration (standards) and trust/safety issues (security and consent) is seen as foundational for successfully implementing PGHD at the system level.

Our qualitative interviews confirmed that PGHD is already used in clinical practice, albeit in uneven ways in all specialties. For example, oncologists described structured approaches using standardized forms for drainage or side effects, whereas primary care physicians noted that PGHD often depends on the personality and self-motivation of the patient. Geriatricians highlighted discrepancies between caregiver- and patient-reported data, especially in populations with dementia, whereas pediatricians in tertiary settings reported reliance on externally provided laboratory results when children could not attend in person. These findings echo previous survey and review studies showing that physicians are broadly supportive of PGHD but remain cautious due to variability in data quality and completeness [18,22,23].

The types of PGHD reported in our study, that is, symptom diaries, physiological measurements, and self-care logs, are

XSL•FO
RenderX

consistent with the existing oncology and chronic disease management literature. Basch et al [24] found that the systematic collection of patient-reported outcomes improved quality of life and even survival in cancer patients, underscoring the clinical value of self-reported longitudinal data. However, many participants in our study emphasized the burden of interpreting raw PGHD without visualization or contextualization, supporting calls for improved integration with EHRs and decision-support systems [25].

Physicians also recognized that PGHD can strengthen patient engagement, although unevenly. Only patients who are organized, digitally educated, or supported by caregivers are likely to generate actionable PGHD. This reflects previous findings that older adults or patients with cognitive impairment require caregiver mediation to produce reliable data, whereas younger, digitally adept patients contribute more consistently [26].

Regarding nonclinical use, physicians in our study emphasized academic and public-interest purposes over industrial applications. Pediatricians noted that PGHD could fill the gaps in pediatric statistics, aligning with larger initiatives such as the European Health Data Space, which positions PGHD as a complement to traditional clinical data for research and public health [17,26]. However, barriers remain: Physicians ranked security, transparency, and consent frameworks as the most urgent priorities for reform, echoing concerns documented in systematic reviews of PGHD governance [27]. They also stressed the importance of verifying the trustworthiness of the entities handling PGHD, whether governmental, institutional, or commercial, underscoring the need for governance models that build both technical and ethical trust.

Finally, physicians prioritized data portability, stronger security obligations, and standardization as key areas for regulatory reform. These priorities mirror global policy trends, including interoperability and anti-information-blocking provisions in the US 21st Century Cures Act [28] and the General Data Protection Regulation's emphasis on patient data portability in Europe [29]. Some participants also highlighted the burden of current consent requirements, noting that repeated signatures and stringent legal reviews create barriers to PGHD use in research. Similar challenges have been observed in studies of electronic and dynamic consent models, which aim to balance regulatory rigor with practical usability [30].

Limitations

This study has several limitations that should be acknowledged. First, the number of respondents to the cross-sectional survey might not have been sufficient to fully represent the perceptions of physicians about PGHD in all specialties and clinical settings. Previous studies aimed at patients and caregivers have demonstrated that PGHD perceptions vary substantially depending on the type of disease and patient characteristics [31]. However, when designing our physician survey, we were unable to incorporate this level of diversity. To mitigate this limitation, we conducted in-depth interviews with physicians from a range of specialties and health care institutions, adding context and nuance to the survey findings.

Second, there were notable demographic imbalances between physicians in tertiary general hospitals and those in smaller institutions. Specifically, respondents from tertiary hospitals were younger and more likely to be trainees or early-career clinicians, whereas respondents from other hospitals tended to have more years of clinical experience. This imbalance likely reflects the real-world distribution of career stages between institution types; however, it may have introduced bias when comparing perceptions of PGHD between hospital groups. For example, age and career stage could influence digital literacy, openness to new technologies, or workload burden, which in turn might have affected the subgroup differences observed in our analyses. Consequently, comparisons between institution types should be interpreted as exploratory and descriptive, rather than as estimates of independent institutional effects.

Third, the cross-sectional nature of this study limited our ability to make causal inferences. Physicians' perceptions and intentions regarding PGHD can evolve as institutional infrastructure, reimbursement policies, and national regulations develop. Longitudinal studies would be required to capture how attitudes change over time in response to these contextual factors.

Fourth, although our study combined quantitative and qualitative methods, self-selection bias may have influenced those who participated in both the survey and the interviews. Physicians with more interest in digital health may have been more inclined to respond, potentially overestimating the level of acceptance of PGHD among clinicians in general.

Finally, although our analysis highlighted significant differences between tertiary and nontertiary hospital physicians, we did not stratify the findings by specialty in a systematic way. Specialty-specific factors (eg, oncology vs psychiatry) likely shape the information needs of physicians and their perceptions of the value of PGHD. Future studies with larger and more diverse physician samples should examine specialty-level differences more rigorously.

Taken together, these limitations suggest that our findings should be interpreted with caution and viewed as exploratory. Nevertheless, by triangulating survey data with in-depth interviews in multiple institution types, our study provides a useful first step towards understanding how physicians perceive the clinical and nonclinical use of PGHD.

Conclusions

Overall, the results of our survey reveal a consistent narrative: Physicians are cautiously optimistic about the value of PGHD, provided certain conditions are met. Across both large tertiary hospitals and smaller institutions, doctors emphasize the need for accurate, reliable data and better integration tools to make the use of PGHD feasible in everyday practice. The key differences by clinical setting were modest, mainly reflecting that tertiary centers have younger staff and somewhat better access to PGHD infrastructure (though not universally used), whereas smaller hospitals face more definite resource gaps (lack of systems, less exposure to PGHD). However, both groups converged on the view that system improvements (integration and standardization) and support (time, guidelines, and security)

XSL•FO
RenderX

are essential. Physicians were primarily concerned with the accuracy of the data and the additional workflow burden, and prioritized PGHD uses that aligned with improving care, advancing research, or public health, rather than commercial interests. These insights suggest that efforts to implement PGHD should focus on building reliable systems that minimize burden, for example, by integrating PGHD into EHRs with automated summarization, providing clinical guidelines for interpretation, and compensating for additional workload. In parallel, policy and regulatory actions (such as setting data standards, improving privacy protections, and clarifying legal responsibilities) will be crucial to address physicians' concerns and pave the way for the greater adoption of PGHD in both clinical practice and secondary applications.

Taken together, our findings suggest that PGHD can improve outpatient care and provide value to research and public health. However, its effective integration requires the following: standardized collection tools to ensure accuracy and comparability, visualization, and contextualization to reduce interpretive burden and facilitate patient–physician communication, and robust but flexible governance frameworks that address security, transparency, and consent. Future research should evaluate not only technical integration strategies but also mechanisms to ensure equitable participation across specialties and patient populations. Ultimately, the sustainable adoption of PGHD will depend on the balance of its potential benefits against the realities of clinical workflow and the trust of both physicians and patients.

Acknowledgments

The authors declare the limited use of generative artificial intelligence (AI) tools during manuscript preparation. In accordance with the Generative AI Disclosure Tool (GAIDeT) taxonomy (2025), the generative AI tools were used solely to assist with translation under full human supervision.

Following AI-assisted translation, the manuscript underwent professional English-language editing by a third-party editing service to ensure clarity, accuracy, and consistency of language. All AI-generated outputs and editorial suggestions were reviewed and approved by the authors. Responsibility for the scientific content, interpretation, and final manuscript rests entirely with the authors. Generative AI tools are not listed as authors and do not bear responsibility for the final outcomes.

Data Availability

The data pertaining to this study's results are available from the Asan Medical Center Institutional Data Access/Ethics Committee. Public disclosure of the data is not feasible because data sharing was not planned during study design and Institutional Review Board review. Researchers who meet the criteria for access to confidential data may contact the committee at irb@amc.seoul.kr or +82-2-3010-7166.

Funding

This research was funded by the Asan Institute of Life Sciences, Asan Medical Center, Seoul, Korea (grant number 2022IP0067-2). This study was also supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute and funded by the Ministry of Health & Welfare, Republic of Korea (grant number RS-2023-KH136246).

Authors' Contributions

SB and YL conceptualized the study. YL developed the methodology, supervised the work, administered the project, and acquired funding. YP and YL validated the data. YP and YL conducted the formal analysis, and YP curated the data. YP and SB led the investigation. SB provided resources. YP drafted the original manuscript, and YP and YL contributed to the review and editing of the manuscript. YP created the visualizations. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Bensing J. Bridging the gap. The separate worlds of evidence-based medicine and patient-centered medicine. Patient Educ Couns 2000;39(1):17-25. [doi: 10.1016/s0738-3991(99)00087-7] [Medline: 11013544]
2. Grayek E, Krishnamurti T, Hu L, Babich O, Warren K, Fischhoff B. Collection and analysis of adherence information for software as a medical device clinical trials: systematic review. JMIR Mhealth Uhealth 2023;11:e46237 [FREE Full text] [doi: 10.2196/46237] [Medline: 37966871]
3. Topol E. Digital medicine: empowering both patients and clinicians. Lancet 2016;388(10046):740-741. [doi: 10.1016/S0140-6736(16)31355-1] [Medline: 27560260]
4. Dorsey ER, Topol EJ. Telemedicine 2020 and the next decade. Lancet 2020;395(10227):859. [doi: 10.1016/S0140-6736(20)30424-4] [Medline: 32171399]

XSL•FO
RenderX

5. Nazi KM, Newton T, Armstrong CM. Unleashing the potential for patient-generated health data (PGHD). J Gen Intern Med 2024;39(Suppl 1):9-13 [FREE Full text] [doi: 10.1007/s11606-023-08461-4] [Medline: 38252246]
6. Jim HSL, Hoogland AI, Brownstein NC, Barata A, Dicker AP, Knoop H, et al. Innovations in research and clinical care using patient-generated health data. CA Cancer J Clin 2020 May;70(3):182-199 [FREE Full text] [doi: 10.3322/caac.21608] [Medline: 32311776]
7. Seo D, Park YR, Lee Y, Kim JY, Park J, Lee J. The use of mobile personal health records for hemoglobin a1c regulation in patients with diabetes: retrospective observational study. J Med Internet Res 2020;22(6):e15372 [FREE Full text] [doi: 10.2196/15372] [Medline: 32484447]
8. Demiris G, Iribarren SJ, Sward K, Lee S, Yang R. Patient generated health data use in clinical practice: a systematic review. Nurs Outlook 2019;67(4):311-330 [FREE Full text] [doi: 10.1016/j.outlook.2019.04.005] [Medline: 31277895]
9. Laverty L, Gandrup J, Sharp CA, Ercia A, Sanders C, Dowding D, et al. Using patient-generated health data in clinical practice: How timing influences its function in rheumatology outpatient consultations. Patient Educ Couns 2022 Mar;105(3):625-631. [doi: 10.1016/j.pec.2021.06.027] [Medline: 34238651]
10. Cresswell KM, McKinstry B, Wolters M, Shah A, Sheikh A. Five key strategic priorities of integrating patient generated health data into United Kingdom electronic health records. J Innov Health Inform 2019;25(4):254-259 [FREE Full text] [doi: 10.14236/jhi.v25i4.1068] [Medline: 30672407]
11. Ye J. The impact of electronic health record-integrated patient-generated health data on clinician burnout. J Am Med Inform Assoc 2021;28(5):1051-1056 [FREE Full text] [doi: 10.1093/jamia/ocab017] [Medline: 33822095]
12. Park YL, Clifton B, Ashraf R, Barlow R, Anderson A, Altamirano V, et al. Patient and provider perspectives on pain and other dimensions of anesthesia experience for cesarean delivery: A qualitative study. Res Sq 2024 Aug 30:4814545 [FREE Full text] [doi: 10.21203/rs.3.rs-4814545/v1] [Medline: 39257975]
13. Reading MJ, Merrill JA. Converging and diverging needs between patients and providers who are collecting and using patient-generated health data: an integrative review. J Am Med Inform Assoc 2018;25(6):759-771 [FREE Full text] [doi: 10.1093/jamia/ocy006] [Medline: 29471330]
14. Guardado S, Mylonopoulou V, Rivera-Romero O, Patt N, Bansi J, Giunti G. An exploratory study on the utility of patient-generated health data as a tool for health care professionals in multiple sclerosis care. Methods Inf Med 2023;62(5-06):165-173 [FREE Full text] [doi: 10.1055/s-0043-1775718] [Medline: 37748719]
15. Khatiwada P, Yang B, Lin J, Blobel B. Patient-generated health data (PGHD): understanding, requirements, challenges, and existing techniques for data security and privacy. J Pers Med 2024;14(3):282 [FREE Full text] [doi: 10.3390/jpm14030282] [Medline: 38541024]
16. Kim B, Ghasemi P, Stolee P, Lee J. Clinicians and older adults' perceptions of the utility of patient-generated health data in caring for older adults: exploratory mixed methods study. JMIR Aging 2021;4(4):e29788 [FREE Full text] [doi: 10.2196/29788] [Medline: 34738913]
17. Brückner S, Sadare O, Fesl S, Scheibe M, Lang C, Gilbert S. Attitudes of healthcare professionals and researchers toward wearable and app derived patient generated health data. NPJ Digit Med 2025;8(1):186 [FREE Full text] [doi: 10.1038/s41746-025-01568-4] [Medline: 40159538]
18. Volpato L, Del Río Carral M, Senn N, Santiago Delefosse M. General practitioners' perceptions of the use of wearable electronic health monitoring devices: qualitative analysis of risks and benefits. JMIR Mhealth Uhealth 2021;9(8):e23896. [doi: 10.2196/23896] [Medline: 34383684]
19. Conceptualizing a data infrastructure for the capture and use of PGHD in care delivery and research through 2024. Office of the National Coordinator for Health IT. URL: https://healthit.gov/wp-content/uploads/2026/01/onc_pghd_practical_guide.pdf [accessed 2026-03-31]
20. Abdolkhani R, Gray K, Borda A, DeSouza R. Recommendations for the quality management of patient-generated health data in remote patient monitoring: mixed methods study. JMIR Mhealth Uhealth 2023;11:e35917 [FREE Full text] [doi: 10.2196/35917] [Medline: 36826986]
21. Review and use of patient-generated health data under the office of connected care. Department of Veterans Affairs VHA. Washington, DC: VHA; 2026. URL: https://www.va.gov/vhapublications/ViewPublication.asp?pub_ID=9252 [accessed 2026-02-06]
22. Tiase VL, Hull W, McFarland MM, Sward KA, Del Fiol G, Staes C, et al. Patient-generated health data and electronic health record integration: a scoping review. JAMIA Open 2020 Dec;3(4):619-627 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa052] [Medline: 33758798]
23. Zhang R, Burgess ER, Reddy MC, Rothrock NE, Bhatt S, Rasmussen LV, et al. Provider perspectives on the integration of patient-reported outcomes in an electronic health record. JAMIA Open 2019 Apr;2(1):73-80 [FREE Full text] [doi: 10.1093/jamiaopen/ooz001] [Medline: 30976756]
24. Basch E, Deal AM, Dueck AC, Scher HI, Kris MG, Hudis C, et al. Overall Survival Results of a Trial Assessing Patient-Reported Outcomes for Symptom Monitoring During Routine Cancer Treatment. JAMA 2017 Jul 11;318(2):197-198 [FREE Full text] [doi: 10.1001/jama.2017.7156] [Medline: 28586821]
25. Dawes AJ, Lin AY, Varghese C, Russell MM, Lin AY. Mobile health technology for remote home monitoring after surgery: a meta-analysis. Br J Surg 2021;108(11):1304-1314. [doi: 10.1093/bjs/znab323] [Medline: 34661649]

XSL•FO
RenderX

26. Omoloja A, Vundavalli S. Patient generated health data: benefits and challenges. Curr Probl Pediatr Adolesc Health Care 2021;51(11):101103. [doi: 10.1016/j.cppeds.2021.101103] [Medline: 34799255]
27. Alpert JM, Manini T, Roberts M, Kota NSP, Mendoza TV, Solberg LM, et al. Secondary care provider attitudes towards patient generated health data from smartwatches. NPJ Digit Med 2020;3:27 [FREE Full text] [doi: 10.1038/s41746-020-0236-4] [Medline: 32140569]
28. 21st century cures act, public law no: 114-255. Congress.GOV. 2016. URL: https://www.congress.gov/114/plaws/publ255/PLAW-114publ255.pdf [accessed 2026-04-23]
29. Complete guide to GDPR compliance. GDPR European Union. 2016. URL: https://gdpr.eu [accessed 2026-04-23]
30. Budin-Ljøsne I, Teare HJA, Kaye J, Beck S, Bentzen HB, Caenazzo L, et al. Dynamic Consent: a potential solution to some of the challenges of modern biomedical research. BMC Med Ethics 2017 Jan 25;18(1):4 [FREE Full text] [doi: 10.1186/s12910-016-0162-9] [Medline: 28122615]
31. Park YE, Beck SS, Lee Y. Perceptions and willingness of patients and caregivers on the utilization of patient-generated health data: a cross-sectional survey. Int J Environ Res Public Health 2025;22(7):1099 [FREE Full text] [doi: 10.3390/ijerph22071099] [Medline: 40724166]

Abbreviations

EHR: electronic health record
GDPR: General Data Protection Regulation
PGHD: patient-generated health data
PHR: patient health record

Edited by A Stone; submitted 23.Oct.2025; peer-reviewed by R Zhang, SY Tan, T Akiyama; comments to author 23.Jan.2026; accepted 21.Apr.2026; published 04.Jun.2026.

Please cite as:
Park YE, Beck SS, Lee Y
Physicians' Perceptions and Demand Regarding Clinical and Secondary Use of Patient-Generated Health Data: Cross-Sectional Survey
J Med Internet Res 2026;28:e86368
URL: https://www.jmir.org/2026/1/e86368
doi:10.2196/86368
PMID:

©Ye-Eun Park, Sang Sook Beck, Yura Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Adaptation and Validation of the Social Media Cyberbullying Victimization Scale Among Chinese College Students: Cross-Sectional Study

Chen Xu[1*], MPH; Suping Wang[2*], PhD; Yujie Liu[1], PhD; Xin Ge[1], MPH; Xue Yang[3], PhD; Xiaohong Fan[4], PhD; Li Li[5], PhD; Ying Wang[1], PhD; Yong Cai[1,6], PhD

[1]Institute of Clinical Research, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, No. 1111, Xianxia Road, Changning District, Shanghai, China

[2]School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[3]JC School of Public Health and Primary Care, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China

[4]Shanghai Public Health Clinical Center, Fudan University, Shanghai, China

[5]Department of Nursing, Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[6]Center for Community Health Care, China Hospital Development Institute, Shanghai Jiao Tong University, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Yong Cai, PhD
Institute of Clinical Research, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, No. 1111, Xianxia Road, Changning District, Shanghai, China

Abstract

Background: Cyberbullying victimization is a significant risk factor for poor psychological well-being among college students. Existing tools fail to capture the distinct dimensions of victimization in social media contexts.

Objective: This study aimed to adapt and validate a Social Media Cyberbullying Victimization Scale (SMCVS) and examine its associations with psychological outcomes among Chinese college students.

Methods: In Shanghai, China, 1766 students from multiple universities completed questionnaires including demographic information, the SMCVS, the Patient Health Questionnaire-9, and the Generalized Anxiety Disorder-7 scale. The psychometric evaluation of the SMCVS included content validity, construct validity, criterion validity, internal consistency reliability, split-half reliability, and test-retest reliability. Receiver operating characteristic analyses were used to evaluate predictive ability for depression and anxiety.

Results: Content validity was satisfactory, and 13 items were retained, with item-level content validity index values ranging from 0.833 to 1 and a scale-level content validity index of 0.949. Exploratory factor analysis identified 2 dimensions—“information harassment” and “reputation and privacy violation”—explaining 82.2% of the total variance. Confirmatory factor analysis supported this structure (goodness-of-fit index=0.943, comparative fit index [CFI]=0.982, normed fit index=0.978, and root mean square error of approximation [RMSEA]=0.075). Multigroup confirmatory factor analysis supported configural and metric invariance across both gender and academic major (ΔCFI≤0.01, ΔRMSEA≤0.015); scalar invariance was achieved for gender but not for academic major. Both dimensions showed significant correlations with depression (r=0.433 - 0.457) and anxiety (r=0.356 - 0.372). Cronbach α, Spearman-Brown coefficient, and test-retest reliability coefficient were 0.959, 0.973, and 0.860, respectively. The SMCVS demonstrated moderate discriminative accuracy for depression (area under the curve [AUC]=0.738; optimal cutoff=19.5) and anxiety (AUC=0.739; optimal cutoff=22.5).

Conclusions: The SMCVS demonstrates sound psychometric properties and may be useful for assessing social media cyberbullying victimization among Chinese college students. Its validated 2D structure clarifies distinct patterns of victimization and their psychological correlates, offering implications for targeted prevention and intervention.

(J Med Internet Res 2026;28:e88857) doi:10.2196/88857

KEYWORDS

cyberbullying victimization; social media; psychometric properties; mental health; college students

XSL•FO
RenderX

Introduction

Cyberbullying has evolved dramatically with the widespread use of social media platforms in the current digital age and has emerged as a critical global public health challenge [1]. Cyberbullying refers to deliberate and aggressive behavior by an individual or a group, conducted through electronic means of communication, occurring repeatedly over time, and targeting victims who struggle to defend themselves [2]. Involvement in cyberbullying can occur in 3 forms: being a victim, engaging as a perpetrator, or witnessing it as a bystander [3]. There is a broad consensus on 3 defining elements of cyberbullying: harmfulness, repetitiveness, and a power imbalance between bullies and victims [4]. Experiencing cyberbullying victimization (CV) on social media has been associated with numerous adverse outcomes, including heightened psychological distress and physical health problems [5], elevated symptoms of depression and anxiety [6], lower levels of life satisfaction [7], and suicidal thoughts and attempts [8]. These findings underscore the urgent need to address cyberbullying as a significant threat to mental and physical health worldwide.

University students, as digital natives and intensive social media users, constitute a particularly vulnerable population for CV. Although cyberbullying among middle and high school students has been extensively studied, university students have received relatively limited attention, leaving a critical gap in our understanding of this issue in this population. Previous studies have reported that the overall prevalence of CV among university students ranges from 19% to 24.1% [9,10]. Furthermore, existing research consistently highlights strong associations between CV and adverse psychological outcomes, particularly depression and anxiety, which are major contributors to the global burden of disease among young adults. For instance, a 1-unit increase in the CV score has been associated with a 1.2-fold higher likelihood of experiencing depression and anxiety [11]. These findings underscore the pressing need for further research to elucidate the unique characteristics and impacts of CV in university student populations.

The unique features of social media have amplified the scope and complexity of cyberbullying compared with traditional forms. Unlike traditional bullying, social media–based victimization exhibits characteristics that often transcend cultural boundaries, including persistent digital footprints, rapid cross-platform dissemination, and the potential for complete anonymity. These characteristics have fundamentally transformed the nature and impact of victimization experiences [12]. Cyberbullying involves the use of social media platforms, as well as fake profiles and identities, to harass, intimidate, or target others. Such behaviors commonly include disseminating threats and rumors, exposing private information, and promoting social exclusion or isolation [13]. Based on the nature of the aggressive behaviors, cyberbullying can be categorized into several types, such as flaming (also referred to as roasting), harassment, denigration, defamation, outing, trickery, sexual harassment, and cyberstalking [14]. International research has begun to identify distinct patterns of cyberbullying and to develop diverse instruments specifically tailored to social media contexts. A review of cyberbullying measurement tools over the past two decades identified 25 studies covering 17 distinct instruments [15], with only one study conducted among Chinese adolescents [16], underscoring the scarcity of culturally specific research in the Chinese context.

Self-report scales are among the most important tools for assessing and identifying cyberbullying and CV. Currently, researchers commonly use two approaches to measure these phenomena: (1) using a single-item question to inquire about individuals' experiences of cyberbullying or CV within a specified time frame and (2) administering multi-item instruments that assess specific behaviors related to cyberbullying and/or CV [17,18]. The former approach has been shown to underestimate the prevalence of cyberbullying and CV [19], whereas the latter lacks consensus regarding which behaviors should be considered as constituting cyberbullying and CV [20]. In many cases, the psychometric properties of cyberbullying and CV scales are not rigorously evaluated before use. For example, a review by Berne et al [21] reported that fewer than half of the 44 included studies provided information on internal consistency. This lack of methodological consistency makes it challenging to compare findings across studies. To address issues of measurement consistency and validity, researchers have called for the development of theoretically and empirically grounded instruments to systematically assess cyberbullying and CV.

Therefore, this study aimed to adapt and validate the Social Media Cyberbullying Victimization Scale (SMCVS), thereby contributing to the global toolkit for assessing CV. We hypothesized that (1) social media cyberbullying victimization, as measured by the SMCVS, would exhibit a clear multidimensional structure; (2) the resulting dimensions would demonstrate acceptable psychometric properties across validation analyses; and (3) these dimensions would show significant associations with symptoms of depression and anxiety.

Methods

Participants

A cross-sectional study was conducted among university students in Shanghai, China, from September to December 2024. A stratified cluster random sampling method was adopted, with 5 universities serving as the sampling frame. Participants were stratified by academic discipline (medical sciences, science and engineering, and social sciences) and academic year (freshman to senior). Within each university, student groups (classes) were defined as clusters according to discipline and academic year. Subsequently, several clusters were randomly selected within each stratum, and all students in the selected clusters were invited to participate in the survey. The inclusion criteria were as follows: (1) full-time undergraduate students; (2) active social media users, defined as using social media platforms at least once per day; and (3) voluntary participation with informed consent.

Recruitment and Data Collection Procedures

In the preliminary stage of this study, a research team was established, consisting of the principal investigator and graduate

XSL•FO
RenderX

student investigators. The team developed an investigator manual and conducted 2 systematic training sessions. The training covered the study objectives, ethical requirements, operational procedures, frequently asked questions, and emergency response plans. Before the formal survey, a pilot study involving 30 medical students was conducted to assess the average questionnaire completion time (approximately 25 minutes) and to refine the questionnaire's layout and instructions based on participant feedback.

Following the sampling design, systematic recruitment procedures were implemented to ensure standardized data collection. Survey notifications and approvals were first coordinated through the student affairs departments. Recruitment primarily took place in designated class meetings. University counselors assisted the research team by introducing the study to the students and distributing the survey via QR codes. An informed consent form was presented on the first page of the online survey, clearly outlining the study purpose, the voluntary nature of participation, and privacy protection measures.

The professional version of the Wenjuanxing platform was used for questionnaire design and data collection. To facilitate systematic tracking across the strata, separate QR code links were generated for different grades and majors. During the data collection sessions, trained investigators were physically present (on-site) in the classrooms to provide guidance and address any immediate questions from the students, while 2 dedicated staff members handled any online inquiries. To ensure data quality, IP restrictions were implemented so that each device could complete the survey only once, and the acceptable completion time was limited to 15 - 30 minutes. Mandatory response settings were applied to prevent missing critical information, and one logic-check item was included to monitor response attentiveness. The status of questionnaire returns was monitored daily, and any anomalies were addressed promptly.

Ethical Considerations

The study protocol was approved by the Ethics Committee of Shanghai Jiao Tong University School of Medicine (no. SJUPN201907). Electronic informed consent was obtained from all participants prior to their inclusion in the study. To ensure privacy and confidentiality, all collected data were strictly anonymized, and no personally identifiable information was collected. Student participation was entirely voluntary, and therefore, no financial compensation was provided. Psychological counseling resources were made available, and high-risk cases identified during the screening process were promptly referred for further evaluation and intervention.

Measures

Demographic Characteristics

Demographic variables were collected, including age, gender, academic major, only child, family economic status, parents' marital status, father's education, mother's education, academic performance, academic stress, smoking, and drinking.

Social Media Cyberbullying Victimization Scale

The SMCVS used in this study can be traced back to the Facebook-related CV scale originally developed by Kwan and Skoric [22], in which all items referred specifically to cyberbullying experiences on Facebook. This instrument was subsequently adapted into a Chinese version by Chen Qiyu and colleagues [23], who replaced the word "Facebook" in the item wording with "social network sites" to assess Chinese adolescents' CV experiences. Later, Li [24] adopted this Chinese CV scale and further revised the wording to better fit the linguistic habits of secondary vocational students. However, neither of these two Chinese versions reported a systematic psychometric evaluation of the scale. This study adopted the Chinese CV scale adapted by Li [24] as the SMCVS and conducted the first comprehensive psychometric evaluation among Chinese college students. The SMCVS contained 17 items assessing several forms of CV, including verbal insults and harassment, denigration, impersonation, deception, social exclusion, and cyberstalking. Items were rated on a 6-point Likert scale, where 1 indicated "never," 2 indicated "once," 3 indicated "2 - 4 times," 4 indicated "5 - 7 times," 5 indicated "8 - 10 times," and 6 indicated "more than 10 times." Higher total scores reflected more frequent experiences of CV.

Psychological Well-Being Measures

Depressive symptoms were assessed using the Patient Health Questionnaire-9 (PHQ-9), a 9-item self-report instrument commonly used in nonpsychiatric settings [25]. Participants rated the severity of their symptoms over the past 2 weeks on a scale from 0 ("not at all") to 3 ("nearly every day"). Total scores range from 0 to 27, with a threshold of 10 or above recommended to indicate clinically significant depressive symptoms [26]. Previous studies have demonstrated the reliability and validity of the Chinese version of the PHQ-9 [27]. In this study, the scale demonstrated excellent internal consistency, with a Cronbach α of 0.934.

Anxiety symptoms were assessed using the Generalized Anxiety Disorder-7 (GAD-7) scale, which comprises 7 items designed to evaluate how frequently individuals experienced anxiety symptoms over the past 2 weeks [28]. Response options included "not at all," "several days," "more than half the days," and "nearly every day," with scores assigned from 0 to 3, respectively. Total scores range from 0 to 21, with higher scores indicating greater anxiety symptom severity. A cutoff score of 10 was used to identify likely cases of generalized anxiety disorder. The GAD-7 has been validated in China as a reliable instrument for screening and assessing anxiety symptom severity [29,30]. In this study, the scale demonstrated excellent internal consistency, with a Cronbach α of 0.947.

Statistical Analysis

Data were presented as mean (SD) for normally distributed continuous variables and as median (IQR) for nonnormally distributed variables. Categorical variables were summarized using frequencies and percentages. Content validity of the initial SMCVS was assessed prior to the main psychometric analyses. A panel of 6 experts in the relevant field independently rated the relevance of each item. Item-level content validity indices (I-CVI) and the scale-level content validity index/average (S-CVI/Ave) were computed as the proportion of experts rating an item as relevant. Item analysis of the SMCVS included

XSL•FO
RenderX

descriptive statistics, corrected item-total correlations, and Cronbach α if item deleted for all 13 items.

To evaluate construct validity, the total sample (n=1766) was randomly split into two subgroups for exploratory factor analysis (EFA; n=883) and confirmatory factor analysis (CFA; n=883). Chi-square tests were performed to compare demographic differences between the two groups. The suitability for factor analysis in the EFA subsample was assessed using the Kaiser-Meyer-Olkin test and Bartlett test of sphericity. Principal component analysis with Promax rotation was used to extract common factors, retaining components with eigenvalues greater than 1. Items with factor loadings below 0.4 and/or high cross-loadings were considered for removal or further investigation [31]. CFA was conducted using maximum likelihood estimation. Model fit was evaluated using multiple indices: chi-square goodness of fit (χ^2/df), goodness-of-fit index (GFI), comparative fit index (CFI), normed fit index (NFI), Tucker-Lewis index (TLI), incremental fit index (IFI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR). Acceptable model fit was defined as $\chi^2/df<5$, GFI, CFI, NFI, TLI, and IFI>0.9, RMSEA<0.08, and SRMR<0.05 [32].

Measurement invariance of the SMCVS across gender and academic major was examined using multigroup CFA. For each grouping variable, a sequence of configural, metric, and scalar models was tested. Model fit was evaluated using χ^2/df, CFI, and RMSEA, and changes between nested models were assessed with ΔCFI and ΔRMSEA. Values of ΔCFI≤0.01 and ΔRMSEA≤0.015 were taken as evidence of acceptable invariance across groups [33,34]. The criterion validity of the SMCVS was assessed using the PHQ-9 and GAD-7 as external criterion measures. Reliability was assessed through internal consistency reliability (Cronbach α coefficient), split-half reliability (Spearman-Brown coefficient), and test-retest reliability (intraclass correlation coefficient and correlation coefficient).

Receiver operating characteristic (ROC) curve analyses and the area under the curve (AUC) were used to evaluate the predictive ability of the SMCVS total and subscale scores for depression and anxiety, using established clinical cutoff points for the PHQ-9 and GAD-7 as external criteria. For each outcome, the optimal cutoff score for the SMCVS was determined by maximizing the Youden index (sensitivity + specificity − 1), and the corresponding sensitivity, specificity, positive likelihood ratio (LR+), and negative likelihood ratio (LR−) were calculated. All analyses were performed using SPSS 26.0, Amos 26.0, and R (version 4.3.2; *pROC* package for ROC analyses), with statistical significance set at $P<.05$ (2-tailed).

Results

Descriptive Analysis

A total of 1766 valid questionnaires were collected, yielding a response rate of 92.4%. The sample included 660 participants from Shanghai Jiao Tong University School of Medicine, 473 from Shanghai Sanda University, 252 from the University of Shanghai for Science and Technology, 248 from Shanghai University of Political Science and Law, and 133 from Shanghai University of Traditional Chinese Medicine.

The demographic characteristics of the participants are presented in Table 1. Participants were aged 18 - 24 years, with a mean age of 19.87 (SD 1.52) years. The participants consisted of 857 (48.5%) males and 909 (51.5%) females. Nearly half of the participants were enrolled in medical-related disciplines (n=793, 44.9%) and were only children (n=885, 50.1%). Most participants (n=1599, 90.5%) reported that their parents were married. The majority (n=1156, 65.5%) reported a medium family economic status. In addition, 45.5% (n=803) of fathers and 39.4% (n=695) of mothers had attained at least a college-level education. Overall, 37.4% (n=660) of participants reported good academic performance and 51.5% (n=910) reported moderate levels of academic stress. Furthermore, 7.1% (n=125) of participants reported smoking behavior and 24.8% (n=438) reported alcohol consumption.

XSL•FO
RenderX

Table . Demographic characteristics of college students (n=1766).

Characteristics	Total (n=1766), n (%)	EFA[a] (n=883), n (%)	CFA[b] (n=883), n (%)	Chi-square (*df*)	*P* value
Age (years)				7.1 (3)	.07
18	339 (19.2)	176 (19.9)	163 (18.5)		
19	500 (28.3)	227 (25.7)	273 (30.9)		
20	398 (22.5)	198 (22.4)	200 (22.7)		
≥21	529 (30.0)	282 (31.9)	247 (28.0)		
Sex				0.1 (1)	.74
Male	857 (48.5)	425 (48.1)	432 (48.9)		
Female	909 (51.5)	458 (51.9)	451 (51.1)		
Academic major				0.3 (2)	.86
Medical science	793 (44.9)	393 (44.5)	400 (45.3)		
Science and engineering	252 (14.3)	124 (14.0)	128 (14.5)		
Social science	721 (40.8)	366 (41.4)	355 (40.2)		
Only child				0.1 (1)	.74
Yes	885 (50.1)	439 (49.7)	446 (50.5)		
No	881(49.9)	444 (50.3)	437 (49.5)		
Family economic status				0.9 (2)	.63
Low	203 (11.5)	102 (11.6)	101 (11.4)		
Medium	1156 (65.5)	586 (66.4)	570 (64.6)		
High	407 (23.0)	195 (22.1)	212 (24.0)		
Parents' marital status				1.1 (1)	.29
Married	1599 (90.5)	793 (89.8)	806 (91.3)		
Divorced	167 (9.5)	90 (10.2)	77 (8.7)		
Father's education				4.5 (2)	.10
≤Junior high	567 (32.1)	304 (34.4)	263 (29.8)		
Senior high	396 (22.4)	188 (21.3)	208 (23.6)		
≥College	803 (45.5)	391 (44.3)	412 (46.7)		
Mother's education				1.8 (2)	.41
≤Junior high	655 (37.1)	339 (38.4)	316 (35.8)		
Senior high	416 (23.6)	198 (22.4)	218 (24.7)		
≥College	695 (39.4)	346 (39.2)	349 (39.5)		
Academic performance				1.4 (2)	.49
Good	660 (37.4)	342 (38.7)	318 (36.0)		
Average	813 (46.0)	399 (45.2)	414 (46.9)		
Poor	293 (16.6)	142 (16.1)	151 (17.1)		
Academic stress				0.5 (2)	.77
Low	287 (16.3)	138 (15.6)	149 (16.9)		
Medium	910 (51.5)	460 (52.1)	450 (51.0)		
High	569 (32.2)	285 (32.3)	284 (32.2)		
Smoking				1.0 (1)	.31
Yes	125 (7.1)	68 (7.7)	57 (6.5)		
No	1641 (92.9)	815 (92.3)	826 (93.5)		

XSL•FO
RenderX

Characteristics	Total (n=1766), n (%)	EFA[a] (n=883), n (%)	CFA[b] (n=883), n (%)	Chi-square (*df*)	*P* value
Drinking				2.4 (1)	.12
Yes	438 (24.8)	233 (26.4)	205 (23.2)		
No	1328 (75.2)	650 (73.6)	678 (76.8)		

[a]EFA: exploratory factor analysis.

[b]CFA: confirmatory factor analysis.

Content Validity

Content validity of the initial 17-item SMCVS was evaluated by a panel of 6 experts in the field, who rated the relevance of each item. In line with commonly recommended standards for the I-CVI ≥0.78 [35], items meeting this criterion were retained and items below this threshold were deleted, resulting in a 13-item version of the SMCVS. The wording of several retained items was slightly revised to improve clarity and alignment with the target population. For these 13 items, the I-CVI values ranged from 0.833 to 1, and the S-CVI/Ave was 0.949, indicating acceptable overall content validity.

Item Analysis

Table 2 presents the descriptive statistics, corrected item-total correlations, and Cronbach α if item deleted for all 13 items of the SMCVS. The mean scores for the items ranged from 1.41 to 3.01, with SDs between 1 and 1.53. Corrected item-total correlations ranged from 0.385 to 0.897, indicating moderate to strong item-total relationships. Notably, item 1 demonstrated the lowest corrected item-total correlation (r=0.385), whereas items 7 and 8 showed the highest correlations (r=0.897). The Cronbach α values if item deleted ranged from 0.953 to 0.969. The internal consistency of the scale would not be substantially improved by removing any single item.

Table . Descriptive statistics and item analysis of the Social Media Cyberbullying Victimization Scale (n=1766).

Items	Mean (SD)	Corrected item-total correlation (*r*)	Cronbach α if item deleted
Item 1	3.01 (1.53)	0.385	0.969
Item 2	2.05 (1.41)	0.691	0.959
Item 3	1.72 (1.27)	0.808	0.955
Item 4	1.86 (1.35)	0.689	0.959
Item 5	1.55 (1.08)	0.888	0.953
Item 6	1.53 (1.08)	0.894	0.953
Item 7	1.53 (1.07)	0.897	0.953
Item 8	1.52 (1.08)	0.897	0.953
Item 9	1.54 (1.07)	0.859	0.954
Item 10	1.48 (1.04)	0.885	0.953
Item 11	1.45 (1.01)	0.888	0.953
Item 12	1.41 (1.00)	0.887	0.953
Item 13	1.45 (1.00)	0.877	0.954

Construct Validity

Exploratory Factor Analysis

The demographic characteristics of the EFA and CFA subsamples were highly consistent, with no statistically significant differences between the two groups (all P>.05; Table 1). The EFA revealed a Kaiser-Meyer-Olkin value of 0.956, and the Bartlett test of sphericity was statistically significant (χ^2_{78}=15,393.4, P<.001), confirming the suitability of the data for factor analysis. Principal component analysis suggested a 2-factor solution with eigenvalues greater than 1 (9.42 and 1.26; Figure 1). Factors 1 and 2 explained 72.49% and 9.71% of the total variance, respectively, yielding a cumulative explained variance of 82.2%. Based on the content of the items loading on each factor, factor 1 was labeled “reputation and privacy violation,” as it comprised 9 items (items 5 - 13), with factor loadings ranging from 0.9 to 0.947. Factor 2 was labeled “information harassment,” as it consisted of 4 items (items 1 - 4), with factor loadings ranging from 0.710 to 0.874 (Table 3).

XSL•FO
RenderX

Figure 1. Scree plot from exploratory factor analysis of the 13-item Social Media Cyberbullying Victimization Scale (n=883). Eigenvalues are plotted against component numbers, showing a clear inflection after the second component and supporting a 2-factor solution.

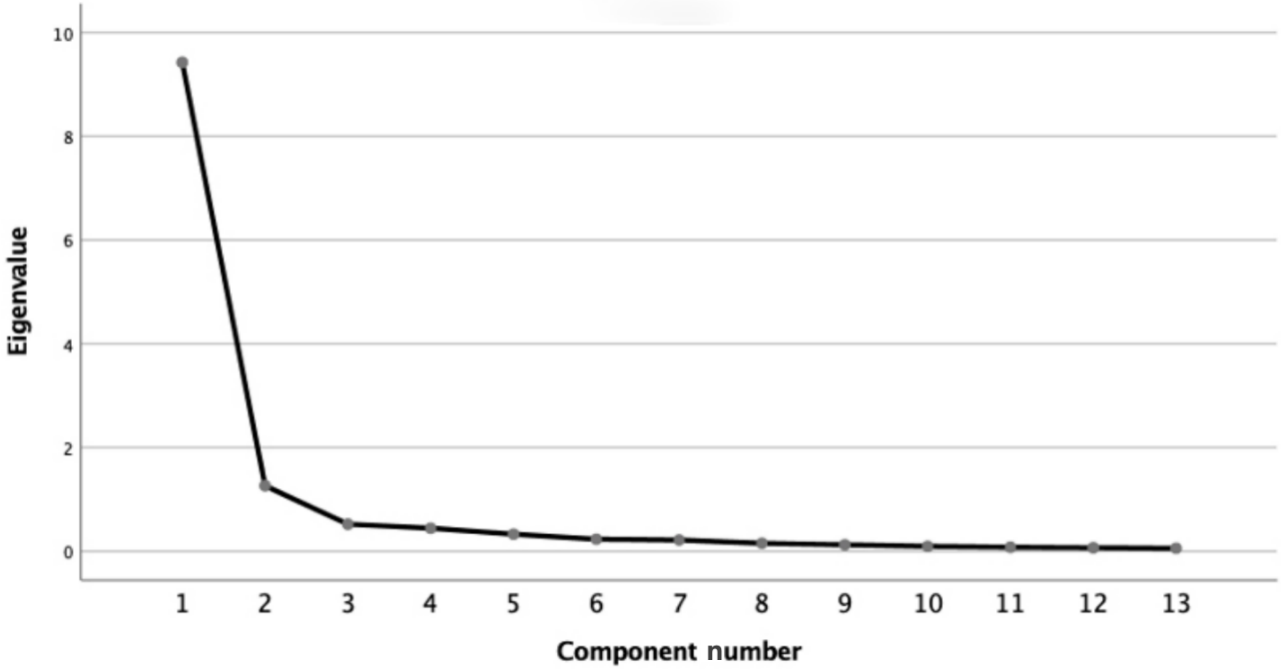

XSL•FO
RenderX

Table . Standardized factor loadings of the Social Media Cyberbullying Victimization Scale (n=883).

Items	Factor 1 (reputation and privacy violation)	Factor 2 (information harassment)
1. I have received disturbing messages on social networking sites.	—[a]	0.826
2. I have received insulting messages, comments, or replies on social networking sites.	—	0.874
3. I have been continuously harassed with insulting messages or comments on social networking sites, even after requesting the perpetrator to stop.	—	0.764
4. I have received pornographic images from others on social networking sites.	—	0.710
5. Someone has posted reputation-damaging messages about me on social networking sites.	0.921	—
6. Someone has spread rumors about me on social networking sites to damage my reputation.	0.926	—
7. Someone has discussed me on social networking sites in ways that made my friends dislike me.	0.934	—
8. Someone has discussed matters about me on social networking sites that made me a laughingstock.	0.934	—
9. Someone has hacked my social networking account and posted/sent messages that made me look bad.	0.900	—
10. Someone has shared my secrets on social networking sites.	0.946	—
11. Someone has posted my private/sensitive information online that I didn't want others to know.	0.947	—
12. I was deceived into sharing my secrets on social networking sites, which were subsequently disseminated.	0.947	—
13. I felt betrayed by friends who posted information about me on social networking sites that I didn't want others to know.	0.929	—
Proportion of variance explained	72.49%	9.71%
Cumulative variance	—	82.2%

[a]Not applicable.

Confirmatory Factor Analysis

The CFA path diagram supported a 2-factor structure for the scale (Figure 2). The model fit indices are presented in Table 4. Specifically, the χ^2/df value was 6.0 (χ^2_{58}=347.9), slightly exceeding the reference threshold of <5. However, the other fit indices indicated strong model performance: the RMSEA was 0.075 (<0.08), the SRMR was 0.022 (<0.05), and the GFI (0.943), CFI (0.982), NFI (0.978), TLI (0.975), and IFI (0.982) all exceeded the reference value of 0.9.

XSL•FO
RenderX

Figure 2. Confirmatory factor analysis of the 13-item Social Media Cyberbullying Victimization Scale (n=883) with unstandardized estimates (A) and standardized estimates (B). The model specifies a 2-factor structure: factor 1 (reputation and privacy violation; items 5-13) and factor 2 (information harassment; items 1-4). Panels A and B present the unstandardized and standardized solutions, respectively, including factor loadings for each item, residual error terms (e1-e13), and their correlated residuals, as well as the covariance between the 2 latent factors.

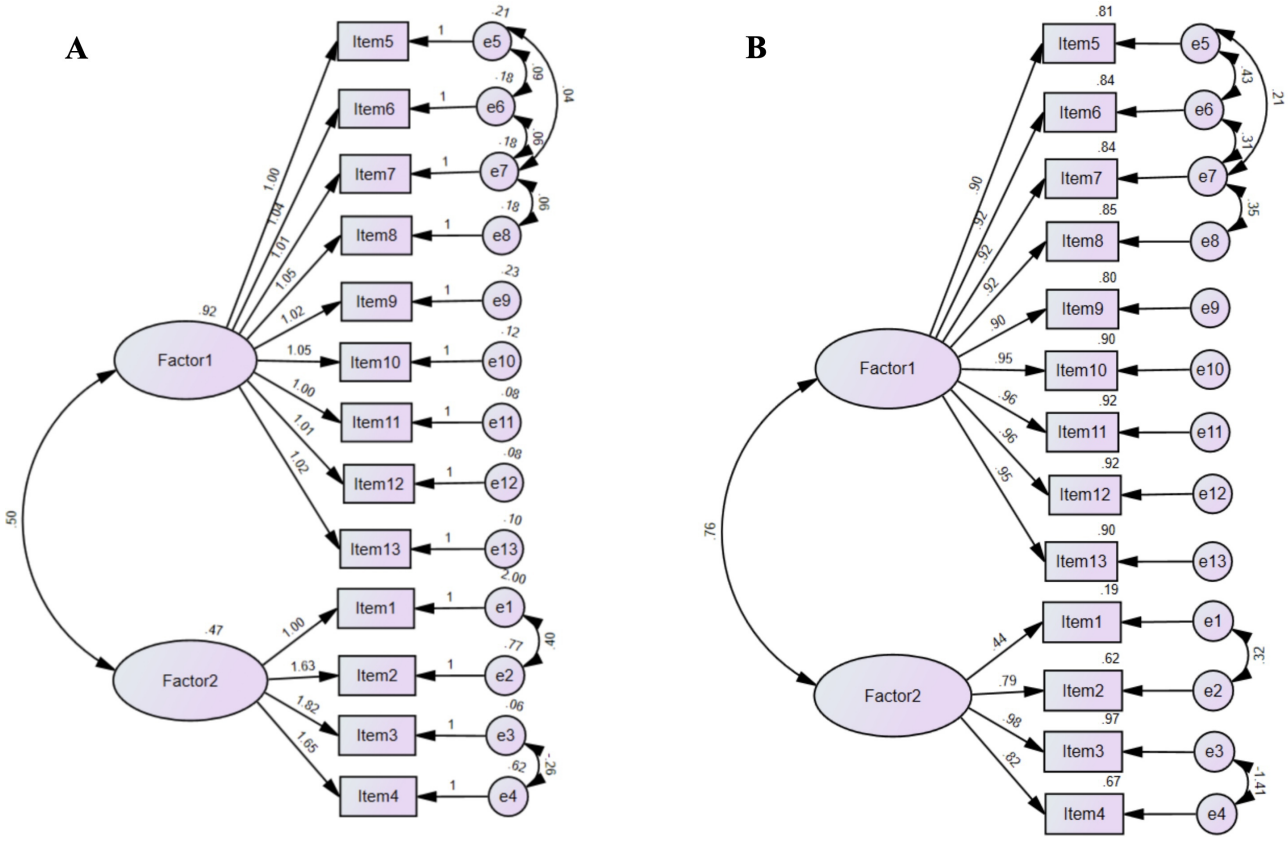

Table . Model fit index of the Social Media Cyberbullying Victimization Scale in confirmatory factor analysis (n=883).

Model fitting index	Reference value	Model value
Chi-square goodness of fit (χ^2/df)	<5.0	6.0
Root mean square error of approximation	<0.08	0.075
Standardized root mean square residual	<0.05	0.022
Goodness-of-fit index	>0.9	0.943
Comparative fit index	>0.9	0.982
Normed fit index	>0.9	0.978
Tucker-Lewis index	>0.9	0.975
Incremental fit index	>0.9	0.982

Measurement Invariance

Results of the configural, metric, and scalar models across gender and academic majors are available in Table 5. For gender, the configural model showed good fit (χ^2_{116}=503.3, χ^2/df=4.339, CFI=0.976, RMSEA=0.062). Constraining factor loadings (metric invariance) resulted in small changes in fit indices (χ^2_{127}=517.3, χ^2/df=4.073, CFI=0.975, RMSEA=0.059; ΔCFI=−0.001, ΔRMSEA=−0.003), and further constraining item intercepts (scalar invariance) yielded similarly small changes (χ^2_{130}=563.1, χ^2/df=4.332, CFI=0.973, RMSEA=0.061; ΔCFI=−0.002, ΔRMSEA=0.002). For the academic major, the configural model also showed acceptable fit (χ^2_{207}=773.4, χ^2/df=3.736, CFI=0.958, RMSEA=0.056). Imposing metric constraints produced negligible changes (χ^2_{218}=810.2, χ^2/df=3.717, CFI=0.956, RMSEA=0.056; ΔCFI=−0.002, ΔRMSEA=0), whereas the scalar model did not meet the recommended thresholds for invariance (ΔCFI=−0.025).

XSL•FO
RenderX

Table . Measurement invariance of the Social Media Cyberbullying Victimization Scale (n=883).

Group and model	χ^2/df[a]	CFI[b]	RMSEA[c]	ΔCFI[d]	ΔRMSEA[d]
Gender					
Configural	4.339	0.976	0.062	—[e]	—
Metric	4.073	0.975	0.059	−0.001	−0.003
Scalar	4.332	0.973	0.061	−0.002	0.002
Academic major					
Configural	3.736	0.958	0.056	—	—
Metric	3.717	0.956	0.056	−0.002	0.000
Scalar	5.244	0.931	0.069	−0.025	0.013

[a]Chi-square goodness of fit.

[b]CFI: comparative fit index.

[c]RMSEA: root mean square error of approximation.

[d]ΔCFI and ΔRMSEA are computed relative to the less constrained preceding model (metric vs configural; scalar vs metric).

[e]Not applicable.

Criterion Validity

As shown in Table 6, the SMCVS total score was strongly correlated with its 2 subscales, "reputation and privacy violation" (r=0.964, P<.01) and "information harassment" (r=0.856, P<.01). The 2 subscales demonstrated a moderate correlation with each other (r=0.688, P<.01). In terms of psychological indicators, the SMCVS total score showed moderate positive correlations with depression (r=0.484, P<.01) and anxiety (r=0.389, P<.01). Both subscales exhibited similar correlation patterns with these psychological indicators, with correlation coefficients ranging from 0.356 to 0.457 (all P<.01).

Table . Correlations between the Social Media Cyberbullying Victimization Scale scores and psychological well-being indicators (n=1766).

Variables	(1)	(2)	(3)	(4)	(5)
(1) SMCVS[a] total	1	—[b]	—	—	—
(2) Reputation and privacy violation	0.964[c]	1	—	—	—
(3) Information harassment	0.856[c]	0.688[c]	1	—	—
(4) Depression (PHQ-9[d])	0.484[c]	0.457[c]	0.433[c]	1	—
(5) Anxiety (GAD-7[e])	0.389[c]	0.356[c]	0.372[c]	0.733[c]	1

[a]SMCVS: Social Media Cyberbullying Victimization Scale.

[b]Not applicable.

[c]P<.01.

[d]PHQ-9: Patient Health Questionnaire-9.

[e]GAD-7: Generalized Anxiety Disorder-7.

Reliability

The internal consistency reliability of the SMCVS, as indicated by Cronbach α, was 0.959, with Cronbach α coefficients for the two subscales of 0.982 and 0.828, respectively. The split-half reliability, measured using the Spearman-Brown coefficient, was 0.973, and the split-half reliability coefficients for the two dimensions were 0.986 and 0.856, respectively. The test-retest reliability of the SMCVS was assessed in 37 medical students with a 3-week interval between measurements. The correlation coefficient between the two measurements was r=0.86 (P<.001), and the intraclass correlation coefficient was 0.9 (95% CI 0.805 - 0.948).

Receiver Operating Characteristic Analyses

ROC analyses were conducted to assess the predictive accuracy of the SMCVS and its subscales for depression and anxiety, as shown in Figure 3. For depression screening, the AUC for the SMCVS total score was 0.738 (95% CI 0.709 - 0.767), while the AUC values for the "information harassment" and "reputation and privacy violation" subscales were 0.721 (95% CI 0.692 - 0.750) and 0.692 (95% CI 0.663 - 0.722), respectively, indicating moderate predictive ability. The optimal cutoff score for the SMCVS total score was 19.5, which yielded a sensitivity of 0.646, a specificity of 0.724, LR+ of 2.339, and LR− of 0.489. For anxiety screening, the AUC for the SMCVS total score was 0.739 (95% CI 0.704 - 0.775), with the

XSL•FO
RenderX

"information harassment" and "reputation and privacy violation" subscales showing AUC values of 0.727 (95% CI 0.691 - 0.762) and 0.7 (95% CI 0.663 - 0.736), respectively, which also reflected moderate discriminative accuracy. The optimal cutoff score for the SMCVS total score was 22.5, with a sensitivity of 0.610, a specificity of 0.763, LR+ of 2.570, and LR− of 0.511. The AUC values and diagnostic indices for the subscales are also summarized in Table 7.

Figure 3. Receiver operating characteristic (ROC) curves for the Social Media Cyberbullying Victimization Scale (SMCVS) total score and its two subscales—"information harassment" and "reputation and privacy violation"—in predicting depression (A) and anxiety (B) among Chinese college students. The curves illustrate the moderate discriminative accuracy of the SMCVS for identifying students at risk of clinically significant depressive and anxiety symptoms.

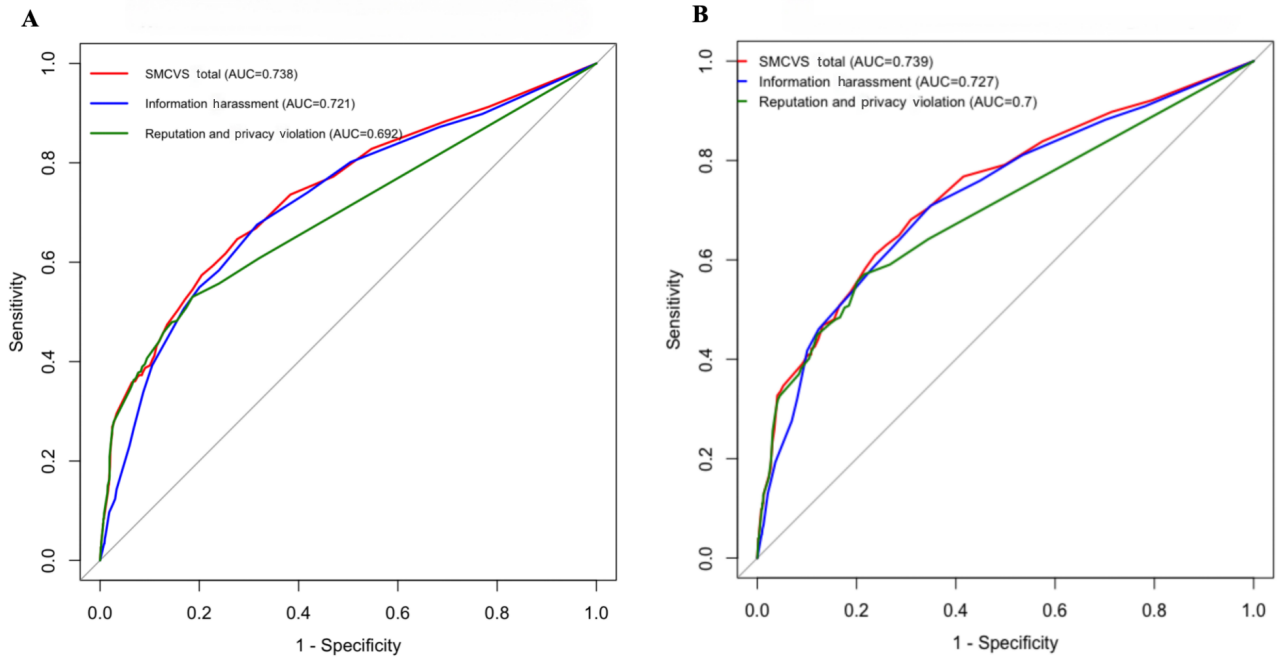

Table . Optimal cutoff scores and diagnostic accuracy indices from ROC[a] analyses of SMCVS[b] scores for depression and anxiety.

Outcome and predictor	AUC[c] (95% CI)	Cutoff	Sensitivity	Specificity	LR+[d]	LR−[e]
Depression						
SMCVS total	0.738 (0.709 - 0.767)	19.5	0.646	0.724	2.339	0.489
Information harassment	0.721 (0.692 - 0.750)	8.5	0.676	0.684	2.136	0.475
Reputation and privacy violation	0.692 (0.663 - 0.722)	11.5	0.530	0.814	2.847	0.557
Anxiety						
SMCVS total	0.739 (0.704 - 0.775)	22.5	0.610	0.763	2.570	0.511
Information harassment	0.727 (0.691 - 0.762)	8.5	0.709	0.651	2.033	0.447
Reputation and privacy violation	0.700 (0.663 - 0.736)	11.5	0.571	0.784	2.648	0.547

[a]ROC: receiver operating characteristic.

[b]SMCVS: Social Media Cyberbullying Victimization Scale.

[c]AUC: area under the curve.

[d]LR+: positive likelihood ratio.

[e]LR−: negative likelihood ratio.

XSL•FO
RenderX

Discussion

Principal Findings

This study adapted and validated a 2D SMCVS for use among Chinese college students. Our findings support the 2-factor structure, comprising "information harassment" and "reputation and privacy violation," and demonstrate generally sound psychometric properties and meaningful associations with psychological outcomes.

The psychometric properties of the SMCVS support its utility as a measurement tool. The high factor loadings (>0.70) and absence of cross-loadings indicate clear dimensional differentiation. The cumulative variance explained (82.2%) exceeds typical thresholds in scale development research, suggesting comprehensive coverage of the construct [36]. The confirmation of this structure through CFA with satisfactory fit indices further supports the scale's structural validity. The emergence of 2 distinct dimensions reflects the evolving nature of cyberbullying in the social media era [37]. Information harassment, characterized by direct aggressive communications, represents traditional forms of cyberbullying adapted to social media contexts. Reputation and privacy violation, involving the unauthorized sharing, distortion, or public exposure of personal information and damaging content, captures the unique affordances of social media platforms that enable new forms of victimization targeting both one's online image and privacy. This dimensional structure advances our understanding of CV by distinguishing between direct aggressive acts and privacy-related violations. In addition, multigroup CFA supported measurement invariance across gender, indicating that factor structure, factor loadings, and item intercepts are comparable and that latent scores can be meaningfully compared between male and female students. For academic major, configural and metric invariance were supported, but scalar invariance was not achieved; therefore, while the structure appears stable across majors, direct mean comparisons of latent scores between different majors should be interpreted with caution, and future studies are needed to further examine and improve cross-major measurement invariance [38].

Some psychometric findings warrant cautious interpretation. Item analysis showed that item 1 had a lower corrected item-total correlation (0.385) than the other items, slightly below the commonly cited cutoff of 0.4 [31]. This item was retained because it captures a broad and relatively mild form of negative experience on social media that may precede or coexist with more explicit cyberbullying behaviors, thereby contributing to content validity. Empirically, item 1 still showed an acceptable loading on its intended factor in the EFA (0.826), and deleting it led to only a negligible change in Cronbach α for the total scale. In addition, although the total and subscale α coefficients were very high (all >0.8), which is often viewed as excellent reliability, such values may also indicate some item redundancy. In particular, the factor loadings for factor 1 were extremely high (0.9 - 0.947), suggesting a high degree of homogeneity among the factor 1 items and potential overlap in the information they capture. Future studies could therefore consider removing or rewriting some factor 1 items to reduce redundancy and improve the informational efficiency of the scale. Finally, the CFA yielded a χ^2/df value of 6.0, slightly above the commonly used lenient criterion. This may partly reflect the well-known sensitivity of chi-square to large samples and model complexity [39]; however, given that other fit indices (eg, CFI, TLI, RMSEA, and SRMR) were within acceptable ranges, the overall model fit is considered adequate, while potential local item dependence should be examined in future research.

Comparison With Prior Work

Scholars have been actively involved in the development of measurement scales for cyberbullying and victimization in recent decades. International measurement scales included the European Cyberbullying Intervention Project Questionnaire, designed for use across 6 European countries [40]; the Cyberbullying Triangulation Questionnaire, created for Spanish students [41]; the Cyber Victimization Experiences and Cyberbullying Behaviors Scales, tailored for British teenagers [42]; the Bullying and Cyberbullying Scale for Adolescents, developed for Australian adolescents [43], and so on [15]. Chinese researchers were also dedicated to developing measurement tools tailored to local culture and national conditions. Lam and Li [16] developed the first scales specifically for CV among Chinese adolescents, known as the E-Victimization Scale. The E-Victimization Scale consisted of only 5 items and was validated as a unidimensional construct. However, the scale was limited to specific social networking platforms, such as email, texting, short messages, and websites like Renren, and focused on a narrow range of cyberbullying behaviors [16]. Rao and colleagues [44] developed a self-designed questionnaire that identified 7 forms of cyberbullying and investigated the prevalence of these 7 types of cyberbullying behaviors among Chinese middle and high school students. Unlike these earlier scales, which primarily targeted adolescents in high school, the scale developed in this study specifically addressed CV among university students. The SMCVS not only removed restrictions on specific social networking platforms but also expanded the scope of behaviors it measured, including harassment, denigration, defamation, sexual harassment, teasing, intrusion, privacy violations, and betrayal. This study conceptualized social media CV as a 2D construct, distinguishing "information harassment" from "reputation and privacy violation." Although our data do not permit a direct comparison of predictive accuracy between the 2D SMCVS and Lam and Li's [16] unidimensional model, the differential associations of the two SMCVS dimensions with depression and anxiety suggest a more nuanced characterization of psychological risk. Compared with 7-category behavioral classifications, this 2-factor structure is more parsimonious while still capturing core forms of social media cyberbullying, though future work is needed to examine whether additional, more specific subdimensions further improve the explanation of mental health outcomes.

The differential associations between SMCVS dimensions and psychological outcomes provide important insights into the nature of CV. In contrast to traditional bullying, cyberbullying does not rely on direct interpersonal contact and often provides perpetrators with complete anonymity. The rapid dissemination of harmful messages and images across online platforms can

XSL•FO
RenderX

expose victims to a wider audience, intensifying their feelings of vulnerability. Moreover, the pervasive nature of the digital environment makes it significantly harder for victims to evade cyberbullying [44]. Studies have shown that many victims of cyberbullying are also victims of traditional bullying [45]. Psychological health issues among cyber-victims are widespread and may even be more severe than those resulting from face-to-face bullying [46]. The study found an association between CV and depression and anxiety among college students, which was consistent with previous research findings [47,48]. Multivariate logistic regression revealed that participants with depression (odds ratio [OR] 1.40, 95% CI 1.17 - 1.67) and anxiety (OR 1.67, 95% CI 1.38 - 2.03) were associated with a higher risk of experiencing cybervictimization [48]. General Strain Theory is one of the most commonly applied frameworks for examining the outcomes of cyberbullying [49]. The theory posits that negative interactions or relationships with others—such as being involved in cyberbullying or experiencing CV—can create strain, which may result in adverse emotional reactions (eg, depression or anxiety) and maladaptive coping strategies (eg, substance use, academic underperformance, or deviant behaviors) [50]. The ROC analyses demonstrated acceptable but moderate discriminative accuracy for both depression (AUC=0.738) and anxiety (AUC=0.739), suggesting that the SMCVS may serve as a supplementary screening tool, rather than a standalone diagnostic instrument, for screening students at risk for psychological distress.

Future Work

Adaptation and validation of the SMCVS specifically tailored for Chinese college students holds significant theoretical and practical value. Given the unique cultural, social, and technological context in China, such a scale can provide a more accurate assessment of cyberbullying experiences and behaviors in this population and may be used in contexts such as academic research, campus mental health services, and policy-related discussions. For instance, it may facilitate the identification of at-risk individuals and the evaluation of prevention and intervention programs, as well as examinations of the prevalence and characteristics of cyberbullying in Chinese higher education institutions. In this context, the ROC analyses provide preliminary support for the screening potential of the SMCVS: the total score showed moderate accuracy in identifying students with depression and anxiety, with optimal cutoff scores of 19.5 and 22.5, respectively, offering a reasonable balance between sensitivity and specificity in this sample and thereby helping to flag students who might benefit from further psychological assessment. Nevertheless, the SMCVS as a screening tool should be used cautiously—higher sensitivity may increase false positives, whereas higher specificity may increase false negatives, meaning that some distressed students could still be missed. Moreover, these cutoff points are sample- and prevalence-dependent estimates derived from cross-sectional data and should not be regarded as definitive; their performance and practical implications need to be re-evaluated in settings with different base rates of depression and anxiety and, ideally, within prospective designs. In clinical and service settings, a 2-stage strategy may be preferable, whereby the SMCVS is used for initial screening followed by more comprehensive clinical or psychological assessment for individuals who screen positive.

Limitations

Several limitations should be noted. First, the cross-sectional design of this study precludes causal inferences about the relationships between CV and psychological outcomes. Longitudinal studies are needed to establish the directionality and underlying mechanisms of these associations. Second, the study was limited to Chinese college students, and the SMCVS was adapted and validated within the mainland Chinese social media ecosystem, where platforms such as WeChat, QQ, and Xiaohongshu are predominant, which may restrict the generalizability of the findings to other cultural contexts. As patterns of use and platform features may differ across social media environments, the applicability of the SMCVS to other platforms and cultural contexts remains to be further examined. Cross-cultural and cross-platform validation studies including participants from more varied backgrounds would help enhance the scale's utility and the universality of the results. Third, the reliance on self-report measures may introduce recall bias, as participants may not accurately remember past experiences, and social desirability effects, where they may underreport or overreport certain behaviors to present themselves in a favorable light. Using multimethod assessments, such as peer reports or observational data, could help address these issues. Fourth, although this study did evaluate the test-retest reliability of the SMCVS, this analysis was conducted in a relatively small subsample of 37 medical students. The high correlation coefficient therefore provides only preliminary evidence regarding the temporal stability of the scale and may not be fully representative of the broader college student population or different academic majors. Future research should replicate the test-retest assessment in larger and more diverse samples to further confirm the robustness and generalizability of the scale's stability over time.

Conclusions

This study establishes the SMCVS as a psychometrically sound and culturally relevant instrument for assessing social media cyberbullying victimization among Chinese college students. By validating its 2D structure, this research captures the specific manifestations of cyberbullying within the contemporary Chinese social media ecosystem. Consequently, it provides researchers with a reliable tool to track victimization trends and evaluate the effectiveness of future preventive programs. Furthermore, the significant associations found between CV and adverse psychological outcomes highlight an urgent need for action. These findings strongly advocate for the integration of targeted cyberbullying screening and coping strategies into university mental health initiatives, ultimately fostering a safer digital environment and supporting the overall psychological well-being of young adults.

XSL•FO
RenderX

Acknowledgments

We thank all participants who contributed to this study. We also appreciate the support from participating universities and the expert panel members who provided valuable feedback during scale development. During the preparation of this manuscript, ChatGPT (OpenAI) was used to assist with manuscript writing and language polishing.

Funding

This study was supported by the Key discipline projects of Shanghai Three-Year Action Plan for Public Health (grant GWVI-11.1-29), the Shanghai Association of Higher Education Projects (2QYB24055 and 2QYB24056), and the Shanghai Municipal Health Commission Project (202340019).

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: YC
Data curation: SW, YL, XG, LL
Formal analysis: CX, YL
Funding acquisition: XF, YW, YC
Methodology: YW, YC
Writing – original draft: CX, SW
Writing – review & editing: YW, XY, YC
Yong Cai (caiyong202028@hotmail.com), Ying Wang (yingwangxun@163.com), and Li Li (sunbook_2001@126.com) are co-corresponding authors of this article.
All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

References

1. Livingstone S, Stoilova M, Kelly A. Cyberbullying: incidence, trends and consequences. In: Ending the Torment: Tackling Bullying From the Schoolyard to Cyberspace: United Nations Special Representative of the Secretary-General on Violence Against Children; 2016:115-120.
2. Giumetti GW, Kowalski RM. Cyberbullying via social media and well-being. Curr Opin Psychol 2022 Jun;45:101314. [doi: 10.1016/j.copsyc.2022.101314] [Medline: 35313180]
3. Chan TKH, Cheung CMK, Lee ZWY. Cyberbullying on social networking sites: a literature review and future research directions. Inf Manage 2021 Mar;58(2):103411. [doi: 10.1016/j.im.2020.103411]
4. Olweus D. Bullying at school: basic facts and effects of a school based intervention program. J Child Psychol Psychiatry 1994 Oct;35(7):1171-1190. [doi: 10.1111/j.1469-7610.1994.tb01229.x] [Medline: 7806605]
5. Albdour M, Hong JS, Lewin L, Yarandi H. The impact of cyberbullying on physical and psychological health of Arab American adolescents. J Immigr Minor Health 2019 Aug;21(4):706-715. [doi: 10.1007/s10903-018-00850-w] [Medline: 30603837]
6. Cao X, Khan AN, Zaigham GHK, Khan NA. The stimulators of social media fatigue among students: role of moral disengagement. Journal of Educational Computing Research 2019 Sep;57(5):1083-1107. [doi: 10.1177/0735633118781907]
7. Viner RM, Gireesh A, Stiglic N, et al. Roles of cyberbullying, sleep, and physical activity in mediating the effects of social media use on mental health and wellbeing among young people in England: a secondary analysis of longitudinal data. Lancet Child Adolesc Health 2019 Oct;3(10):685-696. [doi: 10.1016/S2352-4642(19)30186-5]
8. Sampasa-Kanyinga H, Hamilton HA. Social networking sites and mental health problems in adolescents: the mediating role of cyberbullying victimization. Eur Psychiatry 2015 Nov;30(8):1021-1027. [doi: 10.1016/j.eurpsy.2015.09.011] [Medline: 26512450]
9. Faucher C, Jackson M, Cassidy W. Cyberbullying among university students: gendered experiences, impacts, and perspectives. Education Research International 2014;2014(1):1-10. [doi: 10.1155/2014/698545]
10. Gahagan K, Vaterlaus JM, Frost LR. College student cyberbullying on social networking sites: conceptualization, prevalence, and perceived bystander responsibility. Comput Human Behav 2016 Feb;55:1097-1105. [doi: 10.1016/j.chb.2015.11.019]
11. Rahman T, Hossain MM, Bristy NN, Hoque MZ, Hossain MM. Influence of cyber-victimization and other factors on depression and anxiety among university students in Bangladesh. J Health Popul Nutr 2023 Nov 6;42(1):119. [doi: 10.1186/s41043-023-00469-0] [Medline: 37932869]

XSL•FO
RenderX

12. Dehue F, Bolman C, Völlink T. Cyberbullying: youngsters' experiences and parental perception. Cyberpsychol Behav 2008 Apr;11(2):217-223. [doi: 10.1089/cpb.2007.0008] [Medline: 18422417]
13. Çimke S, Cerit E. Social media addiction, cyberbullying and cyber victimization of university students. Arch Psychiatr Nurs 2021 Oct;35(5):499-503. [doi: 10.1016/j.apnu.2021.07.004] [Medline: 34561065]
14. El Asam A, Samara M. Cyberbullying and the law: a review of psychological and legal challenges. Comput Human Behav 2016 Dec;65:127-141. [doi: 10.1016/j.chb.2016.08.012]
15. Zhang W, Huang S, Lam L, Evans R, Zhu C. Cyberbullying definitions and measurements in children and adolescents: summarizing 20 years of global efforts. Front Public Health 2022;10:1000504. [doi: 10.3389/fpubh.2022.1000504]
16. Lam LT, Li Y. The validation of the E-Victimisation Scale (E-VS) and the E-Bullying Scale (E-BS) for adolescents. Comput Human Behav 2013 Jan;29(1):3-7. [doi: 10.1016/j.chb.2012.06.021]
17. Thomas HJ, Connor JP, Scott JG. Integrating traditional bullying and cyberbullying: challenges of definition and measurement in adolescents – a review. Educ Psychol Rev 2015 Mar;27(1):135-152. [doi: 10.1007/s10648-014-9261-7]
18. Xie X, Zheng H, Ju K, Xiao B, Liu J, Shapka J. Examining the validity and reliability of the short form of cyberbullying and cybervictimization (CAV) scale. Chin J Clin Psychol 2022;30:1155-1159. [doi: 10.16128/j.cnki.1005-3611.2022.05.029]
19. Gradinger P, Strohmeier D, Spiel C. Definition and measurement of cyberbullying. J Psychosoc Res Cyberspace 2010;4(2) [FREE Full text]
20. Sumter SR, Valkenburg PM, Baumgartner SE, Peter J, van der Hof S. Development and validation of the multidimensional offline and online peer victimization scale. Comput Human Behav 2015 May;46:114-122. [doi: 10.1016/j.chb.2014.12.042]
21. Berne S, Frisén A, Schultze-Krumbholz A, et al. Cyberbullying assessment instruments: a systematic review. Aggress Violent Behav 2013 Mar;18(2):320-334. [doi: 10.1016/j.avb.2012.11.022]
22. Kwan GCE, Skoric MM. Facebook bullying: an extension of battles in school. Comput Human Behav 2013 Jan;29(1):16-25. [doi: 10.1016/j.chb.2012.07.014]
23. Chen QY, Tang HY, Zhang L, Zhou ZK. Cyberbullying and its risk factors in Chinese adolescents' usage of social networking sites—based on a survey of 1103 secondary school students. Chin J Spec Educ 2016(3):89-96 [FREE Full text] [doi: 10.16128/j.CNKI:SUN:ZDTJ.0.2016-03-014]
24. Li B. Cyber-victimization and depression of middle-aged vocational students: the chain mediating role of neuroticism and coping styles [Dissertation]. : Jiangxi University of Chinese Medicine; 2022. [doi: 10.27180/d.cnki.gjxzc.2022.000465]
25. Spitzer RL, Kroenke K, Williams JBW. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. JAMA 1999 Nov 10;282(18):1737-1744. [doi: 10.1001/jama.282.18.1737] [Medline: 10568646]
26. Xu C, Shi Y, Yu X, et al. The syndemic condition of psychosocial problems related to depression among sexually transmitted infections patients. PeerJ 2021;9:e12022. [doi: 10.7717/peerj.12022] [Medline: 34616597]
27. Xingchen Hu YZ, Liang W, Zhang H, Yang S. Reliability and validity of the Patient Health Questionnaire-9 in Chinese adolescents. Sichuan Mental Health 2014;27(4):357-360.
28. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
29. Gong Y, Zhou H, Zhang Y, et al. Validation of the 7-item Generalized Anxiety Disorder scale (GAD-7) as a screening tool for anxiety among pregnant Chinese women. J Affect Disord 2021 Mar 1;282:98-103. [doi: 10.1016/j.jad.2020.12.129] [Medline: 33401129]
30. Tong X, An D, McGonigal A, Park SP, Zhou D. Validation of the Generalized Anxiety Disorder-7 (GAD-7) among Chinese people with epilepsy. Epilepsy Res 2016 Feb;120:31-36. [doi: 10.1016/j.eplepsyres.2015.11.019] [Medline: 26709880]
31. Wang Y, Zhang J, Feng X, Liang Y, Guan Z, Meng K. The development and validation of the hospital organizational environment scale for medical staff in China. Front Public Health 2023;11:1118337. [doi: 10.3389/fpubh.2023.1118337] [Medline: 37809008]
32. Gao Y, Liu S, Xu H, et al. The mpox vaccine hesitancy scale for mpox: links with vaccination intention among men who have sex with men in six cities of China. Vaccines (Basel) 2024 Sep 3;12(9):1009. [doi: 10.3390/vaccines12091009] [Medline: 39340039]
33. Cheung GW, Rensvold RB. Evaluating goodness-of-fit indexes for testing measurement invariance. Struct Equ Modeling 2002;9(2):235-255. [doi: 10.1207/S15328007SEM0902_5]
34. Chen FF. Sensitivity of goodness of fit indexes to lack of measurement invariance. Struct Equ Modeling 2007 Jul 31;14(3):464-504. [doi: 10.1080/10705510701301834]
35. Polit DF, Beck CT, Owen SV. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Res Nurs Health 2007 Aug;30(4):459-467. [doi: 10.1002/nur.20199] [Medline: 17654487]
36. Morin AJS, Arens AK, Marsh HW. A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. Struct Equ Modeling 2016 Jan 2;23(1):116-139. [doi: 10.1080/10705511.2014.961800]
37. Sticca F, Perren S. Is cyberbullying worse than traditional bullying? Examining the differential roles of medium, publicity, and anonymity for the perceived severity of bullying. J Youth Adolesc 2013 May;42(5):739-750. [doi: 10.1007/s10964-012-9867-3] [Medline: 23184483]

XSL•FO
RenderX

38. Silva Castillo LH, Hernández-Rosas E, Silas-Casillas JC, Cazares SN, Fonseca León LC. Cross-sectional psychometric validation, convergent validity, and measurement invariance of the DASS-21 in Mexican university students. Front Psychol 2025;16:1707786. [doi: 10.3389/fpsyg.2025.1707786] [Medline: 41439228]
39. Shi D, Lee T, Maydeu-Olivares A. Understanding the model size effect on SEM fit indices. Educ Psychol Meas 2019 Apr;79(2):310-334. [doi: 10.1177/0013164418783530] [Medline: 30911195]
40. Herrera-López M, Casas JA, Romera EM, Ortega-Ruiz R, Del Rey R. Validation of the European cyberbullying intervention project questionnaire for Colombian adolescents. Cyberpsychol Behav Soc Netw 2017 Feb;20(2):117-125. [doi: 10.1089/cyber.2016.0414] [Medline: 28061035]
41. González-Cabrera JM, León-Mejía A, Machimbarrena JM, Balea A, Calvete E. Psychometric properties of the cyberbullying triangulation questionnaire: a prevalence analysis through seven roles. Scand J Psychol 2019 Apr;60(2):160-168. [doi: 10.1111/sjop.12518] [Medline: 30690725]
42. Betts LR, Spenser KA. Developing the cyber victimization experiences and cyberbullying behaviors scales. J Genet Psychol 2017;178(3):147-164. [doi: 10.1080/00221325.2017.1295222] [Medline: 28402227]
43. Thomas HJ, Scott JG, Coates JM, Connor JP. Development and validation of the bullying and cyberbullying scale for adolescents: a multi-dimensional measurement model. Br J Educ Psychol 2019 Mar;89(1):75-94. [doi: 10.1111/bjep.12223] [Medline: 29726005]
44. Rao J, Wang H, Pang M, et al. Cyberbullying perpetration and victimisation among junior and senior high school students in Guangzhou, China. Inj Prev 2019 Feb;25(1):13-19. [doi: 10.1136/injuryprev-2016-042210] [Medline: 28385955]
45. Kowalski RM, Giumetti GW, Schroeder AN, Lattanner MR. Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth. Psychol Bull 2014 Jul;140(4):1073-1137. [doi: 10.1037/a0035618] [Medline: 24512111]
46. Cosma A, Walsh SD, Chester KL, et al. Bullying victimization: time trends and the overlap between traditional and cyberbullying across countries in Europe and North America. Int J Public Health 2020 Jan;65(1):75-85. [doi: 10.1007/s00038-019-01320-2] [Medline: 31844919]
47. Giumetti GW, Kowalski RM, Feinn RS. Predictors and outcomes of cyberbullying among college students: a two wave study. Aggress Behav 2022 Jan;48(1):40-54. [doi: 10.1002/ab.21992] [Medline: 34435367]
48. Sun M, Ma Z, Xu B, Chen C, Chen QW, Wang D. Prevalence of cyberbullying involvement and its association with clinical correlates among Chinese college students. J Affect Disord 2024 Dec 15;367:374-381. [doi: 10.1016/j.jad.2024.08.198] [Medline: 39236886]
49. Agnew R. Building on the foundation of general strain theory: specifying the types of strain most likely to lead to crime and delinquency. J Res Crime Delinq 2001 Nov;38(4):319-361. [doi: 10.1177/0022427801038004001]
50. Hinduja S, Patchin JW. Offline consequences of online victimization. J Sch Violence 2007 Oct 9;6(3):89-112. [doi: 10.1300/J202v06n03_06]

Abbreviations

AUC: area under the curve
CFA: confirmatory factor analysis
CFI: comparative fit index
CV: cyberbullying victimization
EFA: exploratory factor analysis
GAD-7: Generalized Anxiety Disorder-7
GFI: goodness-of-fit index
I-CVI: item-level content validity indices
IFI: incremental fit index
LR+: positive likelihood ratio
LR−: negative likelihood ratio
NFI: normed fit index
OR: odds ratio
PHQ-9: Patient Health Questionnaire-9
RMSEA: root mean square error of approximation
ROC: receiver operating characteristic
S-CVI/Ave: scale-level content validity index/average
SMCVS: Social Media Cyberbullying Victimization Scale
SRMR: standardized root mean square residual
TLI: Tucker-Lewis index
$\chi 2/df$: chi-square goodness of fit

XSL•FO
RenderX

Edited by A Mavragani; submitted 03.Dec.2025; peer-reviewed by A Wang, Z Liu; revised version received 21.Mar.2026; accepted 03.Apr.2026; published 11.Jun.2026.

Please cite as:
Xu C, Wang S, Liu Y, Ge X, Yang X, Fan X, Li L, Wang Y, Cai Y
Adaptation and Validation of the Social Media Cyberbullying Victimization Scale Among Chinese College Students: Cross-Sectional Study
J Med Internet Res 2026;28:e88857
URL: https://www.jmir.org/2026/1/e88857
doi:10.2196/88857

© Chen Xu, Suping Wang, Yujie Liu, Xin Ge, Xue Yang, Xiaohong Fan, Li Li, Ying Wang, Yong Cai. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development of a Digital Health Literacy Assessment Framework for Older Adults: Delphi-Based Study

Xingxia Zhang, MD; Yongqing Yuan, MD; Jie Jiang, PhD

West China Hospital, Sichuan University, No. 37 Guoxue Alley, Wuhou District, Chengdu, Sichuan, China

Corresponding Author:
Jie Jiang, PhD
West China Hospital, Sichuan University, No. 37 Guoxue Alley, Wuhou District, Chengdu, Sichuan, China

Abstract

Background: Digital health literacy (DHL) is crucial for improving health outcomes. As a digitally marginalized group, older adults face significant barriers in using digital technologies for health information access and eHealth services, which hinders their ability to benefit from digital dividends. A comprehensive and contextually relevant assessment tool is essential to understand their DHL and to inform effective interventions.

Objective: This study aimed to develop a theoretically grounded and content-valid assessment framework for DHL among older adults in China, establishing a foundation for future measurement and research.

Methods: A multiphase, theory-informed approach was adopted. The construct domains and an initial item pool were derived from a literature review and qualitative focus group discussions. A 2-round Delphi consultation was conducted to refine and reach consensus on the indicator system. Subsequently, a pilot cross-sectional survey was conducted to validate the reliability and validity of the scale.

Results: Two rounds of expert consultation were conducted with 15 experts, and 322 older adult participants were included in the cross-sectional survey. The expert authority coefficients (*Cr*) were 0.748 and 0.768, respectively, and the Kendall harmony coefficient was statistically significant. The resulting "individual competency-environmental support" dual-pathway framework comprises 6 primary domains (confidence and trust in application, operational skills, critical evaluation skills, application skills, security and privacy, and external support), 13 subdomains, and 27 specific indicators. In the pilot sample (N=332), the overall framework showed excellent internal consistency, and the Cronbach α coefficients ranged from 0.913 to 0.987. Criterion validity was supported by a strong correlation with the established eHealth Literacy Scale tool (r=0.891; P<.001). Known-groups validity was confirmed, as the framework successfully discriminated between older adults with different educational levels (t_{330}=−4.37; P<.001).

Conclusions: This study developed a dual-dimensional "individual competency–environmental support" framework to assess DHL among older adults. The framework systematically identifies competency variations and contextual barriers, offering a standardized tool for needs assessment, intervention design, and evaluation of digital inclusion efforts. It also supports the development of targeted, multisectoral strategies to enhance digital health engagement in aging populations. Further validation across diverse settings is recommended to strengthen its generalizability and practical applicability.

(J Med Internet Res 2026;28:e82334) doi:10.2196/82334

KEYWORDS

older adults; digital health literacy; eHealth; evaluation; Delphi method; China

Introduction

With the rapid development of digital technology and the digital transformation of health care, digital access and skills are increasingly becoming fundamental social determinants of health, as the effective use of both social resources and health care information continues to migrate online in the digital era [1]. Digital health literacy (DHL), defined as the ability to search, filter, understand, evaluate, and apply health information obtained through electronic platforms to solve health-related issues [2], has evolved into a core competency for advancing health management and population health outcomes [3]. Studies have shown that DHL is positively correlated with health-related behaviors [4], self-health management [5], quality of life [6], and other health-related outcomes [7]. Digital literacy is a pivotal element in conditioning the effectiveness of telemedicine and digital medicine services, plays a critical role in achieving health equity in third-millennium society [8], and can reach large numbers of people at relatively low cost [9]. Nevertheless, ongoing digital divides hamper improvements in patient self-management, care coordination, quality, and cost-effectiveness [10], especially among marginalized populations, such as ethnic minority groups, individuals with

XSL•FO
RenderX

limited English proficiency, those with low socioeconomic status, and older adults [11]. A scoping review revealed that research on DHL in older adults is still in its infancy, especially in terms of assessment tools and intervention methods [12]. As "digital refugees" (here, referring to individuals who, due to multidimensional capability deficits in accessing, comprehending, and applying digital health resources, experience systematic exclusion from digital health ecosystems and ultimately suffer technologically mediated health inequities through marginalization), the DHL needs of older adults warrant greater attention.

DHL quantitative assessment is a critical step in scientifically, objectively, and comprehensively evaluating the proficiency of specific populations, drawing significant attention from scholars in related fields. A recent bibliometric analysis of global eHealth highlighted the eHealth Literacy Scale (eHEALS) as both a pivotal instrument and a defining frontier issue for advancing research in this domain [13]. The measurement tools developed during different stages of internet evolution vary. In 2006, Norman and Skinner [14] labeled this ability "eHealth literacy" (EHL) and developed eHEALS, an 8-item measure of EHL developed to measure consumers' combined knowledge, comfort, and perceived skills in finding, evaluating, and applying eHealth information to health problems, emphasizing the use of health information available on the internet. As the use of eHealth grows and diversifies globally, the eHEALS may not fully align with user-website interactions in Web 2.0 [15]. The core distinction between Web 2.0 and Web 1.0 lies in interactivity. Web 1.0 is characterized primarily by unidirectional dissemination, whereas Web 2.0 enables users to contribute user-generated content to the web, thereby establishing bidirectional engagement [16]. The measurement tools developed during Web 2.0 have broadened the multidimensional conceptualization of DHL through refined dimensional structuring, incorporating metrics assessing internet proficiency and online interactive communication capabilities [17]. Within the current Web 3.0 environment—the "semantic web" operating on a "read-write-execute" model to deliver digital, personalized, and intelligent services [18]—Liu et al [19] developed the 24-item 3D eHEALS-Web 3.0, including acquisition, verification, and application, which was subsequently validated among older adults with noncommunicable diseases [20]. Several measurement tools have been developed for the general population, such as the Digital Health Literacy Assessment [21], Joint Information System Committee [22], Digital Health Readiness Questionnaire [23], Digital Health Technology Literacy Assessment Questionnaire [24], Yongin Severance Digital Sensitivity Scale [25], and others.

However, only a few studies have assessed the use and appropriateness of these measurements for older populations [26]. The eHEALS is the most frequently used instrument for measuring digital literacy among older adults [26], and was developed 19 years ago to measure abilities related to page views [14]. As the definition of DHL is ever-evolving, it is necessary to reassess and enhance existing DHL measures for better scaling of current eHealth usage. Recently, Wang et al [27] developed a 39-item older DHL questionnaire structured around 6 dimensions (information, behavior, security, interaction, content, and attitudes) based on the digital competence framework (DigComp). Although a valuable contribution, its primary focus on individual competencies such as information, interaction, security, and so on, as well as its length, may limit feasibility for widespread geriatric assessment and ignore a critical determinant of digital engagement in later life: the enabling or constraining role of the external environment [27].

Social ecological theory posits that individual health behaviors and outcomes result from continuous interactions between personal attributes and multilevel environmental contexts, such as interpersonal networks, community resources, and policy environments [28]. Similarly, the digital health divide is a multilevel phenomenon spanning "access-usage-benefit," influenced by factors ranging from intrapersonal skills to social, community, and environmental conditions [29]. The digital health divide is likely determined not only by the DHL of individuals but also by various other interacting factors, such as individual lifestyle factors, attitudes, social and community networks, and wider cultural and environmental conditions [30-32]. For older adults, DHL is not merely an individual skill set but is profoundly shaped by external support systems, such as the availability of community resources, reliable internet infrastructure, and facilitative social networks. Existing tools, including those specifically designed for older adults [33-35], focus on assessing individual skills but largely neglect this socioecological perspective, creating a measurement gap.

Given the limitations in validating existing tools for older adults, future research should evaluate the diagnostic accuracy of measurement tools in this population and investigate the impact of social determinants of health on such assessments to strengthen their clinical implementation [36]. Guided by the social ecological perspective, we reconceptualize DHL in older adults as a context-rooted, composite competence that integrates personal agency and environmental enablement, highlighting the interaction of "individual-environment." Consequently, the core contribution of this research lies in constructing an innovative dual-pathway "individual competency-environmental support" assessment framework, formally establishing "environmental support" as a core, independently measurable dimension of equal importance to personal competency. Methodologically, through a rigorous Delphi expert consensus process, we developed a more concise 27-item measurement instrument, aiming to balance comprehensiveness with respondent-friendliness for older adults.

In summary, this study aims to develop a theoretically grounded, structurally clear, and pragmatically feasible assessment tool. This instrument can simultaneously reveal the strengths and weaknesses of older adults' DHL at both the personal competency and environmental support levels. It is designed to provide more precise and comprehensive empirical evidence for formulating systematic public health strategies and community-based interventions in the Web 3.0 era, ultimately aimed at reducing the digital health divide and promoting health equity among the older adult population.

XSL•FO
RenderX

Methods

Study Design

An initial evaluation index system was developed through a literature review and focus group discussions (FGDs). This was followed by a 2-round Delphi expert consultation involving 15 experts. The FGDs were convened by a 13-member research team led by the principal investigator, comprising 2 senior researchers, 4 midlevel researchers, 2 junior researchers, and 5 graduate researchers (2 doctoral and 3 master's candidates). These discussions served 2 specific purposes: first, during the initial stage, they were used to screen and refine the item pool generated from the literature review, where participants discussed the relevance and clarity of potential items; second, during the Delphi process, they were used to facilitate discussion and decision-making regarding the adoption of expert opinions, helping to interpret disagreements and reach consensus on item modifications. Thus, the FGD data were iteratively used to inform item refinement and integrate expert feedback, rather than being analyzed through formal qualitative methods such as thematic or content analysis. By integrating qualitative analysis with quantitative computation, the initial framework comprising "6 primary indicators, 13 secondary indicators, and 28 tertiary indicators" was progressively refined into a scientifically structured system featuring "6-13-27" indicators. Finally, the reliability and validity of the questionnaire were verified through a cross-sectional survey. The specific procedural design is illustrated in Figure 1.

XSL•FO
RenderX

Figure 1. Research framework.

XSL•FO
RenderX

Development of the Preliminary Indicator Pool

A comprehensive literature search was performed in the PubMed, Web of Science, Cochrane Library, CNKI, and WanFang databases to collect initial evaluation indicators using the following keywords:

"older adults" OR "older people" OR "elderly" AND "digital health literacy" OR "electronic health literacy" OR "eHealth literacy" OR "digital health" OR "information literacy" AND "indicator" OR "index" OR "scale."

The search covered literature published from its inception to February 28, 2025. The results revealed that previous studies focused primarily on assessing the ability to access, evaluate, and apply health information, and ignored the influence of self-efficiency, as well as social and environmental elements, on DHL among older adults. Based on these findings, we developed a new comprehensive questionnaire with particular emphasis on the above dimensions, considering both the individual capacity and the environmental support of the older population.

Using the Delphi Method to Develop an Index System for Expert Selection

This study used the Delphi method to administer expert consultation questionnaires via an online platform [37] to gather perspectives among participants, thereby enhancing the objectivity and scientific validity of the outcomes. While the Delphi methodology imposes no fixed threshold for expert panel size, increased participant numbers strengthen result stability [38]. A thorough review of the literature indicates that a panel of 4 to 16 experts generally yields sufficiently reliable findings for Delphi consultations. When the number of selected experts approaches 15, further increases contribute minimally to prediction precision. Consequently, this study ultimately enrolled 15 experts as participants for the consultation.

eHealth literacy represents a complex, multidisciplinary issue. The experts selected for consultation should possess authoritativeness, representation, and reliability, particularly to demonstrate substantial knowledge or experience relevant to the research topic. Based on the study's objectives and content, the Delphi expert panel was recruited through purposive sampling, leveraging the research team's established professional networks in geriatric care, clinical practice, and academic research. Invitations were extended to individuals who met the following predefined eligibility criteria: (1) work experience of 10 years or more in fields including geriatric medicine, geriatric nursing, older adult care services, or community management; (2) a bachelor's degree or higher; (3) an intermediate-level technical title or higher; and (4) voluntary participation in the study with the ability to complete and return questionnaires promptly. Ultimately, a multidisciplinary panel of 15 experts including clinicians, nurses, health care administrators, and researchers, agreed to participate and completed both rounds of the Delphi consultation.

Design of the Expert Consultation Questionnaire

Based on the findings from the literature review and qualitative interviews, the initial evaluation index system for assessing older adults' EHL was developed into a consultation questionnaire. This instrument comprises four main sections: (1) an introduction to the research, including the purpose of the study, instructions for completing the questionnaire, details on how and when to submit it, contact information, and acknowledgments; (2) demographic information of experts; (3) the core section (expert consultation questionnaire on DHL evaluation indicators for urban community-dwelling older adults); and (4) an expert authority questionnaire, which includes expert self-assessments of their authority and familiarity with the questionnaire content according to theoretical analysis, work experience, peer understanding, and intuition.

Within the questionnaire's primary section, the experts rated the importance of each indicator across all hierarchical levels via a 5-point Likert scale (ranging from 1="Not Important" to 5="Extremely Important"). Additionally, experts were invited to provide specific suggestions for adding, deleting, or modifying indicators in designated comment fields.

Two rounds of consultation were conducted via email and WeChat. A mean importance score of more than 3.50, a coefficient of variation (CV) of 0.30 or less, and a total score rate of more than 20% were the inclusion criteria [39,40]. Entries were adjusted on the basis of discussions within the research team, incorporating expert modifications. After 2 rounds, the final DHL questionnaire for older adults included 6 dimensions and 27 items (Table 1). The detailed results of scoring of experts were presented in the Multimedia Appendix 1.

XSL•FO
RenderX

Table . The comprehensive evaluation index system for digital health literacy (DHL) in older adults.

Tier-3 indicator	Importance, mean (SD)	CV[a]	Full score (%)
Confidence and trust in application	4.267 (0.884)	0.21	53.30
1.1 Belief in self-capability	4.333 (0.724)	0.17	46.70
1.1.1 I believe my actions can improve health status	4.400 (0.632)	0.14	46.70
1.2 Belief in online health information	4.200 (0.775)	0.18	40
1.2.1 I believe I can effectively obtain health-related information or services via the internet	4.267 (0.594)	0.14	33.30
1.2.2 I believe I can resolve health concerns using online resources	4.133 (0.743)	0.18	33.30
1.2.3 I am confident in applying online information for health decisions	4.000 (0.845)	0.21	33.30
Operational skills	4.533 (0.640)	0.14	60
2.1 Electronic device proficiency	4.333 (0.724)	0.17	46.70
2.1.1 I can input text or voice via mobile or computer keyboards	4.200 (0.862)	0.21	46.70
2.1.2 I can locate and follow health information platforms (eg, WeChat public accounts)	4.200 (0.862)	0.21	46.70
2.2 Health communication	4.467 (0.743)	0.17	60
2.2.1 I can clearly articulate health concerns in writing	4.333 (0.724)	0.17	46.70
2.2.2 I can verbally express symptoms and desired health solutions	4.467 (0.640)	0.14	53.30
2.3 Information retrieval	4.333 (0.724)	0.17	46.70
2.3.1 I use multiple online channels for diverse health information needs	4.533 (0.743)	0.16	66.70
2.3.2 I use accurate keywords for targeted information searches	4.400 (0.632)	0.14	46.70
Critical evaluation skills	4.333 (0.724)	0.17	46.70
3.1 Relevance assessment	4.400 (0.828)	0.19	60
3.1.1 I evaluate information applicability to my health condition	4.467 (0.640)	0.14	53.30
3.1.2 I identify appropriate health behaviors from retrieved information	4.333 (0.617)	0.14	40.00
3.2 Reliability assessment	4.533 (0.640)	0.14	60
3.2.1 I verify information authenticity	4.333 (0.724)	0.17	46.70
3.2.2 I discern commercial motives (eg, advertising) in information	4.200 (0.862)	0.21	46.70
3.3 Authenticity verification	4.667 (0.617)	0.13	73.30
3.3.1 I cross-verify information across websites	4.400 (0.828)	0.19	60

XSL•FO
RenderX

Tier-3 indicator	Importance, mean (SD)	CV[a]	Full score (%)
3.3.2 I discuss online health information with family, friends, or medical professionals for credibility assessment	4.467 (0.743)	0.17	60
Application skills	4.533 (0.640)	0.14	60
4.1 Practical use	4.333 (0.900)	0.21	53.30
4.1.1 I apply retrieved health information in daily life	4.333 (0.816)	0.19	53.30
4.1.2 I use information for health decisions (nutrition, medication, or emergency response)	4.200 (0.775)	0.18	40
Security and privacy	4.333 (0.976)	0.23	60
5.1 Protective awareness	4.400 (0.828)	0.19	60
5.1.1 I identify unsafe websites or social platforms	4.533 (0.915)	0.20	73.30
5.1.2 I protect personal/others' private information (name or address)	4.667 (0.724)	0.16	80
5.1.3 I regularly clear browsing history	3.533 (1.060)	0.30	20.00
External support	4.000 (1.000)	0.25	40
6.1 Providers (community or health care institutions)	3.933 (0.799)	0.20	26.70
6.1.1 I access local health care services for health needs	4.600 (0.632)	0.14	66.70
6.1.2 I use community-based technology training for skill enhancement	4.270 (0.800)	0.19	46.70
6.2 Institutional support (government or enterprises)	4.000 (1.000)	0.25	40
6.2.1 My residence or senior centers have broadband internet	3.800 (1.080)	0.29	40
6.2.2 My devices have senior-friendly features (large text, icons, audio, or remote assistance)	3.930 (0.800)	0.20	26.70
6.3 Social support (family or community organizations)	4.130 (0.830)	0.20	40
6.3.1 I regularly interact with family online (eg, WeChat) or offline	4.270 (0.800)	0.19	46.70
6.3.2 I regularly interact with friends/community volunteers online or offline	4.130 (0.830)	0.20	40

[a]CV: coefficient of variation.

Evaluation of the Reliability and Validity of the Index System

Participants and Sample Selection Permanent residents aged 60 years and older in a city in Sichuan Province (defined as people living in the city for 6 months or more, without distinguishing local household registration); basic understanding ability, ability to understand and fill in or answer the questions in the questionnaire; and voluntary participation in this study.

Sample Size

The 27 evaluation indicators were translated into specific questions to develop an empirical survey questionnaire. All response options used a 5-point Likert scale, where participants selected the score corresponding to their situation. Regarding the pilot survey sample size, a common rule is 5 to 10 participants per item [41]. An additional 20% buffer was included to account for potentially invalid responses. The target

XSL•FO
RenderX

sample size was calculated as: 27 items*10 participants/item*(1+20%)=324 participants. The pilot study successfully recruited 332 community-dwelling older adults from Pengzhou city, Sichuan Province, meeting and slightly exceeding this target.

Survey Tools

The questionnaire encompassed four sections: (1) questionnaire instructions and informed consent, (2) general information about the respondents (designed by the researcher, including age, sex, place of residence, education level, marital status, household size, living arrangement, and so on), (3) the eHealth Literacy Assessment Scale, and (4) the culturally adapted Chinese version of the eHEALS [42] (developed by Norman et al [14]), which is an 8-item measure of eHealth literacy developed to measure consumers' combined knowledge, comfort, and perceived skills in finding, evaluating, and applying eHealth information to health problems, with scores ranging from 1 to 5 points according to the Likert 5-level scoring method, and the total score is 8 to 40 points (Multimedia Appendix 2).

Ethical Considerations

The study was approved by the ethics committee on Biomedical Research, West China Hospital of Sichuan University (1044). All participants were informed of the study's purpose, and procedures, and provided voluntary consent to participate. Delphi experts received an honorarium for their participation, while cross-sectional survey participants took part voluntarily and received no compensation. The survey adhered to the principles of anonymity and confidentiality, and all responses were collected anonymously.

Reliability and Validity Testing

The Cronbach α coefficient and Spearman-Brown split-half reliability are the most commonly used methods for testing the internal consistency reliability of questionnaires [43]. The split-half reliability was assessed by dividing the 27 items into 2 halves using the odd-even method. Generally, a Cronbach α coefficient greater than 0.500 for each dimension, along with a total Cronbach α coefficient and Spearman-Brown split-half reliability greater than 0.7 indicate acceptable reliability.

For validity testing, since there is currently no universally established "gold standard" instrument specifically designed to measure comprehensive DHL within the contemporary Web 3.0 paradigm for older adults, the eHEALS was used as the criterion tool in this study. Criterion validity was assessed by examining the correlation between the total score of our scale and the eHEALS score. In addition, known-groups validity was evaluated. Participants were divided into 2 groups based on their educational attainment (high school or above vs junior high school or below), as educational level has been consistently identified as a key determinant of DHL in previous literature [44]. An independent-samples 2-tailed *t* test was conducted to compare the mean scale scores between the 2 groups. Construct validity, assessed through factor analysis, was not performed in this phase of the study, primarily due to the limited sample size. This limitation is acknowledged in the "Discussion" section and will be addressed in future research with a larger sample.

Statistical Analysis

Data management and statistical analyses were conducted using Microsoft Excel and the SPSSAU online platform (Beijing Qingsi Technology Co, Ltd). Expert demographic characteristics were described with frequencies (percentages) for categorical variables and means (SDs) for continuous variables. Expert engagement was evaluated through questionnaire response rates and the proportion of experts providing substantive feedback. Expert authority was quantified via the authority coefficient (Cr=[Ca+Cs]/2). The judgment basis coefficient (Ca) was derived from the expert's self-assessment of the influence level (high, medium, or low) across 4 dimensions, with scores assigned as shown in Table 2.

Table . Assignment of values for the familiarity coefficient (Cs).

Basis of judgment	Major value	Medium value	Minor value
Practical experience	0.3	0.2	0.1
Theoretical analysis	0.5	0.4	0.3
Literature review	0.1	0.1	0.1
Intuitive judgment	0.1	0.1	0.1

The familiarity coefficient (Cs) was assigned based on the expert's self-rated familiarity with the subject, using a 5-point scale (1.0="very familiar," 0.75="familiar," 0.5="generally familiar," 0.25="somewhat unfamiliar," and 0="very unfamiliar"). Both components were collected via a self-assessment questionnaire completed by the experts. Consensus was assessed via the CV and the Kendall concordance coefficient (*w*).

Results

Results of the Delphi Consultation

The Delphi consultation involved 15 experts from 13 universities and tertiary-level hospitals located in Sichuan Province, Beijing, Shanghai, Hunan Province, Shandong Province, and Xi'an Province, China. All 15 (100%) completed both rounds of the Delphi consultation. Among the 15 experts, 6 (40%) were male participants and 9 (60%) were female participants. The participants were aged 36 to 56 years (mean 45.93, SD 6.17), with professional experience ranging from 10 to 34 years (mean 21.07, SD 8.21). Twelve (80%) experts held a master's degree

XSL•FO
RenderX

or higher, and 13 (86.67%) possessed associate senior professional titles or above. Regarding administrative roles, 1 (6.67%) served as a hospital president, 8 (53.33%) held positions as directors or deputy directors of clinical departments, and 1 (6.67%) was the head nurse of a ward unit. The characteristics are presented in Table 3.

Table . Characteristics of the Delphi participants (N=15).

Characteristics	Value, n (%)
Sex	
Male	6 (40)
Female	9 (60)
Age (y)	
≤40	2 (13.33)
41 - 49	9 (60)
50 - 60	4 (26.67)
Education	
Undergraduate	3 (20)
Master's	2 (13.33)
PhD	10 (66.67)
Profession title	
Middle	2 (13.33)
Vice senior	8 (53.33)
Full senior	5 (33.33)
Administrative positions	
Hospital-level leadership	1 (6.67)
Director of department	3 (20)
Deputy director	5 (33.33)
Head nurse	1 (6.67)
None	5 (33.33)
Years worked	
10 - 20	2 (13.33)
21 - 29	10 (66.67)
≥30	3 (20)
Ares of expertise	
Health care services	8 (53.33)
Integrated older adult care	3 (20)
Public health	2 (13.33)
Geriatrics	1 (6.67)
Medical informatics	1 (6.67)

Expert Engagement

A questionnaire return rate exceeding 70% is generally considered indicative of good retrieval and high expert engagement [45]. In the first round of this study's expert consultation, 15 expert correspondence questionnaires were distributed, and all 15 were returned, yielding a return rate of 100%. Among these, 7 (46.67%) experts provided comments. In the second round, 15 questionnaires were distributed, all 15 were returned (100% return rate), and 4 (26.67%) experts provided comments. These results demonstrate a high level of expert engagement in the consultation process.

Expert Authority

The expert authority coefficient (Cr) is determined jointly by Cs and Ca. The mean Cr values for the 2 consultation rounds were 0.748 (SD 0.096) and 0.768 (SD 0.089), both exceeding the acceptable threshold of 0.7 [45]. Therefore, the expert authority level for this consultation is deemed acceptable.

XSL•FO
RenderX

Expert Consensus

Following 2 rounds of consultation, expert opinions converged, demonstrating high acceptance of all indicators. The CV in the first round ranged from 0.141 to 0.207, whereas in the second round it ranged from 0.141 to 0.225. The Kendall *W* coefficient tests for both consultation rounds yielded statistically significant results ($P<.001$ and $P=.005$, respectively), indicating strong consensus among expert evaluations.

Modification of Indicators and the Evaluation Index System

In the first Delphi round, the definition of 1 primary indicator was revised, and the conceptual definition of "confidence and trust in application" was redefined as "trust in self-efficacy and online health information." One secondary indicator ("belief in online health" was updated to "belief in online health information") underwent revisions, while 1 tertiary item was eliminated because of content overlap with the 12 tertiary items that were modified.

In the second round, 1 primary indicator definition was refined, and the definition of the primary dimension "external support" was refined from "evaluation of external factors" to "perception of external factors." One secondary indicator was amended, and minor wording optimizations were implemented for 6 tertiary items.

All indicators achieved a mean importance score exceeding 3.5, with a full-score rate surpassing 20%, meeting predetermined criteria except for a few items demonstrating CV values below 0.30. Given the critical significance of privacy exposure and digital technology infrastructure in the digital era for DHL, the item ("5.1.3 I regularly clear browsing history") was retained following rigorous FGDs. A strong perspective held that, in the context of managing online health information, this behavior concretely reflects privacy protection awareness—a critical facet of the "safety" dimension in DHL. Its high importance score and theoretical alignment supported its inclusion.

Through research team deliberation guided by predefined inclusion criteria and expert feedback, the final hierarchical framework for assessing older adults' DHL comprised 6 primary indicators, 13 secondary indicators, and 27 third-level indicators (Table 1).

Results of the Reliability and Validity of the Questionnaire

The pretest yielded 332 valid questionnaires. The participants' ages ranged from 60 to 82 years, with a mean age of 65.4 (SD 5.2) years. The cohort comprised 175 (52.7%) males and 157 (47.3%) females. Married individuals constituted the majority (n=276, 83.1%). The education level was predominantly junior high school or below (n=243, 73.2%), and farming was the primary occupation (n=165, 49.7%). Most households consisted of 3 to 5 members (n=233, 70.1%), and coresiding with children was the most common living arrangement (n=176, 53%). The detailed characteristics are presented in Table 4.

XSL•FO
RenderX

Table . Characteristics of the participants (N=332).

Characteristics	Values
Age (y), mean (SD)	65.41 (5.18)
Sex, n (%)	
Male	175 (52.7)
Female	157 (47.3)
Marital status, n (%)	
Married	276 (83.1)
Widowed	41 (12.3)
Divorced	12 (3.6)
Unmarried	3 (0.9)
Education level, n (%)	
Junior high school or below	243 (73.2)
High school	73 (22.0)
College	16 (4.8)
Occupation, n (%)	
Farmer	165 (49.7)
Retired	124 (37.3)
Self-employed	38 (11.4)
Other	5 (1.5)
Health insurance, n (%)	
Urban-rural resident basic medical insurance	209 (63.0)
Employee basic medical insurance	123 (37.0)
Household size (persons), n (%)	
<3	53 (16.0)
3 - 5	233 (70.1)
>5	46 (13.9)
Living arrangement, n (%)	
Co-residing with children	176 (53.0)
Co-residing with partner	128 (38.6)
Living alone	24 (7.2)
Other	4 (1.2)

Scoring, Weighting, and Preliminary Interpretation of the Index System

The instrument is based on a 5-point Likert scale. The total score is derived from the sum of all 27 items, with a theoretical range of 27 to 135 points. In the validation phase, an equally weighted summed score was used for clarity and ease of interpretation, with each item contributing equally to the total.

In the cross-sectional survey, the mean total DHL score among older adults was 70.13 (SD 23.53), with a median score of 67.0 (IQR 51.0 - 91.0). The mean item score was 2.60 (SD 0.87), with a median item score of 2.48 (IQR 1.89 - 3.37), based on a 5-point Likert scale.

Result of the Reliability Analysis

This study primarily assessed the internal consistency reliability and split-half reliability of our DHL for older adults. The IBM SPSS software yielded the following results: internal consistency reliability: The overall Cronbach α coefficient for DHL was 0.984, exceeding the acceptable threshold of 0.70. All dimensions demonstrated satisfactory internal consistency, with Cronbach α coefficients exceeding the threshold of 0.70. The detailed results are presented in Table 5.

XSL•FO
RenderX

Table . Results of the reliability analysis.

Dimension	Cronbach α	Number of items
Confidence and trust in application	0.968	4
Operational skills	0.966	6
Critical evaluation skills	0.987	6
Application skills	0.913	2
Security and privacy	0.935	3
External support	0.954	6
Split-half reliability		
First half	0.968	14
Second half	0.967	13
Full scale	0.984	27

Split-Half Reliability

The split-half reliability was assessed by dividing the 27 items into 2 halves using the odd-even method. The Cronbach α coefficients for these 2 subsets were 0.968 and 0.967, respectively. The correlation coefficient was 0.952. The Spearman-Brown split-half reliability coefficient for the full scale was 0.922. These results collectively demonstrate excellent internal consistency reliability for DHL.

Results of the Validity Analysis

Criterion validity was used to evaluate the validity of the DHL. The Chinese version of the eHEALS serves as the reference criterion, as it represents one of the most widely used and validated instruments in this domain [26]. Its established reliability in older Chinese populations (Cronbach α=0.988) [46] supports its suitability as a benchmark. Analysis revealed a strong Pearson correlation coefficient of 0.891 (P<.001) between the DHL scale and the eHEALS, demonstrating good concurrent validity for the newly developed scale.

To further examine the construct validity of the scale, a known-groups validity analysis was conducted. Consistent with existing literature [44], educational attainment is a well-established predictor of DHL. Participants were therefore divided into 2 groups based on their level of education: “high school or above” education (n=89) and a group with “junior high school or below” education (n=243). The results of an independent samples t test revealed a statistically significant difference between the groups (t_{330}=−4.37; P<.001). This finding indicates that the newly developed scale can effectively discriminate between groups with theoretically different levels of DHL, thereby providing strong evidence for its construct validity.

Discussion

Principal Findings

In this study, an assessment framework and corresponding scale for evaluating the DHL of older adults were developed and underwent preliminary validation. The framework comprises 2 dimensions—individual competency and external support—with 6 first-level indicators, 13 second-level indicators, and 27 third-level indicators, covering the core dimensions of DHL among older adults. The total score ranges from 27 to 135, derived from the sum of all 27 items rated on a 5-point Likert scale. To offer an initial reference for application, an interpretive framework was developed based on the percentile distribution of the pilot sample (N=332). Specifically, total scores below the 25th percentile were preliminarily categorized as “needing support,” scores between the 25th and 75th percentiles as “having foundational competency,” and scores above the 75th percentile as “proficient.” It must be emphasized that this framework and its cutoff values are derived from the pilot data and serve as provisional, practical benchmarks. Their generalizability and optimal thresholds require further validation in future studies with larger, more representative populations.

The study fully considers the interplay between socioecological levels, and the finalized “individual competency-environmental support” dual-pathway index system transcends the limitations of conventional measurement tools that focus solely on individual skills. At the individual competency level, the newly incorporated “confidence and trust in application” dimension emphasizes self-efficacy and trust in online health information, capturing psychological and cognitive influences on digital health behaviors. The “critical evaluation skills,” “application skills,” and “security and privacy” dimensions delineate competencies in information filtering, authenticity verification, and privacy protection, aligning with digital health demands in the Web 3.0 era. At social and community levels, the influence of family, friends, and peers on an individual’s health-related behavior has been recognized [47,48]. In the DHL scale, the wider cultural, social, and environmental level elements are included. The external support dimension encompasses multiple stakeholders (community, government, institutions, family, and so on), where indicators such as “health care services,” “community-based technology training,” and “broadband internet” specifically address practical barriers to digital inclusion, providing a clear direction for policy development and service design.

Comparison With Prior Work

There are many measurements in the field of DHL. However, only a few studies have explored the influencing factors of older adults’ DHL from the perspective of social capital [49].

XSL•FO
RenderX

Compared with existing instruments, the key distinction of this framework lies in its systematic incorporation of "environmental support" as an independent dimension, moving beyond a focus solely on individual skills. While widely used tools such as eHEALS emphasize personal operational and cognitive abilities, this study integrates external factors, including community, family, and social support, aligning more closely with health ecological theory and digital inclusion perspectives. Methodologically, the use of multiround Delphi consensus combined with qualitative data from focus groups enhanced the content validity and cultural relevance of the indicators, addressing limitations associated with purely literature-driven or statistically screened item pools, which are commonly used in the research of tool construction [25,50].

This assessment framework has dual practical significance. First, it establishes a standardized screening tool for assessing older adults' DHL. The external support was first included in the DHL evaluation index system. Second, it promotes multistakeholder collaboration to enhance digital health capabilities. In addition to increasing eHealth skills, it is important to consider the role of trust, self-efficacy, motivation, social influence, and support in the use of digital health tools [51]. In terms of government, policies on national digital infrastructure (eg, free Wi-Fi for all), health service budgets, and resources for investing in and supporting digital health rollout can be used to mitigate digital inequalities [11], and communities and families can implement targeted training and companion services [52]. This integrated approach empowers older adults to transition from "digitally vulnerable groups" to "active health self-managers," advancing the achievement of healthy aging goals.

Limitations

Despite yielding noteworthy findings, this study has several limitations. First, methodologically, the scale relies on self-report and lacks item weighting, which may affect the objectivity of the results and the discriminant validity among core competencies. Future work will combine behavioral observation with confirmatory factor analysis to cross-validate the measurement approach and explore weighting schemes. Second, the single-center, cross-sectional design limits the generalizability of the findings and prevents examination of the tool's ability to predict health behaviors. A multicenter longitudinal study is planned to expand sample diversity and track the relationship between scale scores and actual health behaviors. Third, the current framework lacks the capacity for dynamic assessment. Given the rapid advancement of digital technologies and the evolving health needs of older adults, the existing indicators may not remain timely or fully applicable in the long term [53]. A standing revision panel comprising experts in geriatrics, digital technology, and public health will be established to implement periodic evaluation and content updates, ensuring the tool's continued relevance amid technological change.

Conclusions

The dual-dimensional "individual competency–environmental support" framework developed in this study offers a structured tool for systematically identifying variations in digital health competencies among older adults. It highlights the combined role of individual and environmental factors in shaping DHL and provides a standardized approach to assess key competencies and contextual barriers. The framework can be used for needs identification, community-based intervention planning, and evaluation of digital inclusion initiatives, while also supporting the design of targeted, multistakeholder interventions. Importantly, this study aligns closely with national strategies such as "Healthy China" and "Active Aging," offering evidence-based support for policy implementation from the perspective of digital empowerment and inclusive development. By promoting DHL among older adults, the research contributes to enhancing their ability to manage health autonomously and participate socially, thereby supporting the policy goal of building an aging-friendly digital health environment and an inclusive aging society. Future studies should validate the framework across diverse populations and settings to strengthen its applicability and policy relevance, ensuring its continued use in supporting and refining national healthy aging strategies.

Acknowledgments

The authors sincerely thank the 15 experts who participated in the Delphi process for their time and expert input. No generative artificial intelligence tools were used in the writing of this manuscript.

Funding

This work was supported by a grant from the Sichuan Provincial Philosophy and Social Sciences (Key Project) funded by the Sichuan Provincial Government (SCJJ23ND52).

Data Availability

The datasets are available from the first and corresponding author upon reasonable request.

Authors' Contributions

JJ and XZ conceived and designed this study. XZ and YY participated in the literature research and analyzed and interpreted the data. JJ connected with experts and XZ wrote the original draft. All authors reviewed and revised the manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Delphi expert rating scores for all questionnaire items.
[DOC File, 147 KB - jmir_v28i1e82334_app1.doc]

Multimedia Appendix 2
Survey questionnaire on digital health literacy among urban and rural residents in Sichuan Province.
[DOCX File, 24 KB - jmir_v28i1e82334_app2.docx]

References

1. Lyles CR, Wachter RM, Sarkar U. Focusing on digital health equity. JAMA 2021 Nov 9;326(18):1795-1796. [doi: 10.1001/jama.2021.18459] [Medline: 34677577]
2. Bittlingmayer UH, Dadaczynski K, Sahrai D, van den Broucke S, Okan O. Digital health literacy—conceptual contextualization, measurement, and promotion. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2020 Feb;63(2):176-184. [doi: 10.1007/s00103-019-03087-6] [Medline: 31950230]
3. Dong Q, Liu T, Liu R, Yang H, Liu C. Effectiveness of digital health literacy interventions in older adults: single-arm meta-analysis. J Med Internet Res 2023 Jun 28;25:e48166. [doi: 10.2196/48166] [Medline: 37379077]
4. Kim K, Shin S, Kim S, Lee E. The relation between eHealth literacy and health-related behaviors: systematic review and meta-analysis. J Med Internet Res 2023 Jan 30;25:e40778. [doi: 10.2196/40778] [Medline: 36716080]
5. Wu Y, Wen J, Wang X, et al. Associations between e-health literacy and chronic disease self-management in older Chinese patients with chronic non-communicable diseases: a mediation analysis. BMC Public Health 2022 Nov 29;22(1):2226. [doi: 10.1186/s12889-022-14695-4] [Medline: 36447176]
6. Keles A, Kose M, Somun UF, Culpan M, Yaksi N, Yıldırım A. Impact of health and digital health literacy on quality of life following radical prostatectomy for prostate cancer: prospective single-center cohort study. World J Urol 2024 Apr 17;42(1):241. [doi: 10.1007/s00345-024-04960-z] [Medline: 38632212]
7. Park SY, Chung S. E-health literacy and associated factors among Korean adults during the COVID-19 pandemic: age-group differences. Health Promot Int 2023 Oct 1;38(5):daad099. [doi: 10.1093/heapro/daad099] [Medline: 37665719]
8. Campanozzi LL, Gibelli F, Bailo P, Nittari G, Sirignano A, Ricci G. The role of digital literacy in achieving health equity in the third millennium society: a literature review. Front Public Health 2023;11:1109323. [doi: 10.3389/fpubh.2023.1109323] [Medline: 36891330]
9. Gentili A, Failla G, Melnyk A, et al. The cost-effectiveness of digital health interventions: a systematic review of the literature. Front Public Health 2022;10:787135. [doi: 10.3389/fpubh.2022.787135] [Medline: 36033812]
10. Rodriguez JA, Clark CR, Bates DW. Digital health equity as a necessity in the 21st century cures act era. JAMA 2020 Jun 16;323(23):2381-2382. [doi: 10.1001/jama.2020.7858] [Medline: 32463421]
11. Nicosia J, Aschenbrenner AJ, Adams SL, et al. Bridging the technological divide: stigmas and challenges with technology in digital brain health studies of older adults. Front Digit Health 2022;4:880055. [doi: 10.3389/fdgth.2022.880055] [Medline: 35574256]
12. Wang X, Luan W. Research progress on digital health literacy of older adults: a scoping review. Front Public Health 2022;10:906089. [doi: 10.3389/fpubh.2022.906089] [Medline: 35991040]
13. Tian H, Chen J. A bibliometric analysis on global eHealth. Digit Health 2022;8:20552076221091352. [doi: 10.1177/20552076221091352] [Medline: 35433016]
14. Norman CD, Skinner HA. eHEALS: The eHealth Literacy Scale. J Med Internet Res 2006 Nov 14;8(4):e27. [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
15. Norman C. eHealth literacy 2.0: problems and opportunities with an evolving concept. J Med Internet Res 2011 Dec 23;13(4):e125. [doi: 10.2196/jmir.2035] [Medline: 22193243]
16. Van De Belt TH, Engelen LJLPG, Berben SAA, Schoonhoven L. Definition of Health 2.0 and Medicine 2.0: a systematic review. J Med Internet Res 2010 Jun 11;12(2):e18. [doi: 10.2196/jmir.1350] [Medline: 20542857]
17. van der Vaart R, Drossaert C. Development of the digital health literacy instrument: measuring a broad spectrum of health 1.0 and health 2.0 skills. J Med Internet Res 2017 Jan 24;19(1):e27. [doi: 10.2196/jmir.6709] [Medline: 28119275]
18. Rudman R, Bruwer R. Defining Web 3.0: opportunities and challenges. Electron Lib 2016 Feb 1;34(1):132-154. [doi: 10.1108/EL-08-2014-0140]
19. Liu HX, Chow BC, Liang W, Hassel H, Huang YW. Measuring a broad spectrum of eHealth skills in the Web 3.0 context using an eHealth Literacy Scale: development and validation study. J Med Internet Res 2021 Sep 23;23(9):e31627. [doi: 10.2196/31627] [Medline: 34554098]
20. Cai W, Liang W, Liu H, et al. Electronic Health Literacy Scale–Web3.0 for older adults with noncommunicable diseases: validation study. J Med Internet Res 2024 Jun 3;26:e52457. [doi: 10.2196/52457] [Medline: 38830207]

XSL•FO
RenderX

21. Liu P, Yeh LL, Wang JY, Lee ST. Relationship between levels of digital health literacy based on the Taiwan digital health literacy assessment and accurate assessment of online health information: cross-sectional questionnaire study. J Med Internet Res 2020 Dec 21;22(12):e19767. [doi: 10.2196/19767] [Medline: 33106226]
22. Milanti A, Chan DNS, Parut AA, So WKW. Determinants and outcomes of eHealth literacy in healthy adults: a systematic review. PLoS ONE 2023;18(10):e0291229. [doi: 10.1371/journal.pone.0291229] [Medline: 37792773]
23. Scherrenberg M, Falter M, Kaihara T, et al. Development and internal validation of the Digital Health Readiness Questionnaire: prospective single-center survey study. J Med Internet Res 2023 Mar 10;25:e41615. [doi: 10.2196/41615] [Medline: 36897627]
24. Yoon J, Lee M, Ahn JS, et al. Development and validation of Digital Health Technology Literacy Assessment Questionnaire. J Med Syst 2022 Jan 24;46(2):13. [doi: 10.1007/s10916-022-01800-8] [Medline: 35072816]
25. Park HI, Jeon M, Ahn JS, Chung K, Park JY. Development and validation of the Digital Sensitivity Scale for adults: cross-sectional observational study. J Med Internet Res 2025 Jan 10;27:e55828. [doi: 10.2196/55828] [Medline: 39793004]
26. Oh SS, Kim KA, Kim M, Oh J, Chu SH, Choi J. Measurement of digital literacy among older adults: systematic review. J Med Internet Res 2021 Feb 3;23(2):e26145. [doi: 10.2196/26145] [Medline: 33533727]
27. Wang X, Zhang C, Qi Y, et al. Digital health literacy questionnaire for older adults: instrument development and validation study. J Med Internet Res 2025 Mar 19;27:e64193. [doi: 10.2196/64193]
28. McLeroy KR, Bibeau D, Steckler A, Glanz K. An ecological perspective on health promotion programs. Health Educ Q 1988;15(4):351-377. [doi: 10.1177/109019818801500401] [Medline: 3068205]
29. Western MJ, Smit ES, Gültzow T, et al. Bridging the digital health divide: a narrative review of the causes, implications, and solutions for digital health inequalities. Health Psychol Behav Med 2025;13(1):2493139. [doi: 10.1080/21642850.2025.2493139] [Medline: 40276490]
30. Jahnel T, Dassow HH, Gerhardus A, Schüz B. The digital rainbow: digital determinants of health inequities. Digit Health 2022;8:20552076221129093. [doi: 10.1177/20552076221129093] [Medline: 36204706]
31. Lyles CR, Nguyen OK, Khoong EC, Aguilera A, Sarkar U. Multilevel determinants of digital health equity: a literature synthesis to advance the field. Annu Rev Public Health 2023 Apr 3;44:383-405. [doi: 10.1146/annurev-publhealth-071521-023913] [Medline: 36525960]
32. Pampel FC, Krueger PM, Denney JT. Socioeconomic disparities in health behaviors. Annu Rev Sociol 2010 Aug;36:349-370. [doi: 10.1146/annurev.soc.012809.102529] [Medline: 21909182]
33. Siqi L, Jingjing F, Dehui K, et al. Development and reliability and validation test of the digital health literacy assessment scale for the community-dwelling elderly. Chin Nurs Res 2021;35(23):4169-4174 [FREE Full text]
34. Fan Y. Establishing an evaluation index system and anempirical study for digital health competenceamong older adults lived in the community [Dissertation in Chinese]. : Bengbu Medical University; 2023. [doi: 10.26925/d.cnki.gbbyc.2023.000279]
35. Xinxin W, Zhu Z, Wei L. Construction of evaluation indicator system of digital health literacy of older adults Construction of evaluation indicator system of digital health literacy of older adults [Article in Chinese]. Mod Prev Med 2022;49(20):3747-3752. [doi: 10.20043/j.cnki.MPM.202205514]
36. Huang YQ, Liu L, Goodarzi Z, Watt JA. Diagnostic accuracy of eHealth literacy measurement tools in older adults: a systematic review. BMC Geriatr 2023 Mar 29;23(1):181. [doi: 10.1186/s12877-023-03899-x] [Medline: 36978033]
37. Wenjuanxing [Website in Chinese]. URL: https://www.wjx.cn [accessed 2024-12-11]
38. Jorm AF. Using the Delphi expert consensus method in mental health research. Aust N Z J Psychiatry 2015 Oct;49(10):887-897. [doi: 10.1177/0004867415600891] [Medline: 26296368]
39. Diamond IR, Grant RC, Feldman BM, et al. Defining consensus: a systematic review recommends methodologic criteria for reporting of Delphi studies. J Clin Epidemiol 2014 Apr;67(4):401-409. [doi: 10.1016/j.jclinepi.2013.12.002] [Medline: 24581294]
40. Suchmacher M, Geller M. Practical Biostatistics: A User Friendly Approach for Evidence-Based Medicine: Academic Press; 2012.
41. MacCallum RC, Widaman KF, Preacher KJ, Hong S. Sample size in factor analysis: the role of model error. Multivariate Behav Res 2001 Oct 1;36(4):611-637. [doi: 10.1207/S15327906MBR3604_06] [Medline: 26822184]
42. Shuaijun G, Xiaoming Y, SUN Y, Dan N, Xuemin L, Lu W. Adaptation and evaluation of Chinese version of eHEALS and its usage among senior high school students [Article in Chinese]. Chin J Health Educ 2013;29(2). [doi: 10.16168/j.cnki.issn.1002-9982.2013.02.019]
43. Rattray J, Jones MC. Essential elements of questionnaire design and development. J Clin Nurs 2007 Feb;16(2):234-243. [doi: 10.1111/j.1365-2702.2006.01573.x] [Medline: 17239058]
44. Estrela M, Semedo G, Roque F, Ferreira PL, Herdeiro MT. Sociodemographic determinants of digital health literacy: a systematic review and meta-analysis. Int J Med Inform 2023 Sep;177:105124. [doi: 10.1016/j.ijmedinf.2023.105124] [Medline: 37329766]
45. Liu J, Qiu H, Zhang X, Zhang C, He F, Yan P. Development of billing post competency evaluation index system for nurses in China: a Delphi study. BMC Nurs 2023 Apr 25;22(1):136. [doi: 10.1186/s12912-023-01301-0] [Medline: 37098520]

XSL•FO
RenderX

46. Pendl D, Maitz KM, Gasteiger-Klicpera B. Examining the relationship between health literacy and individual and sociodemographic factors in secondary school students. J Public Health (Berl) 2024 Mar;32(3):531-542. [doi: 10.1007/s10389-023-01836-1]
47. König LM, Attig C, Franke T, Renner B. Barriers to and facilitators for using nutrition apps: systematic review and conceptual framework. JMIR Mhealth Uhealth 2021 Apr 1;9(6):e20037. [doi: 10.2196/20037] [Medline: 34254938]
48. Hayat TZ, Brainin E, Neter E. With some help from my network: supplementing eHealth literacy with social ties. J Med Internet Res 2017 Mar 30;19(3):e98. [doi: 10.2196/jmir.6472] [Medline: 28360024]
49. Shi Z, Du X, Li J, Hou R, Sun J, Marohabutr T. Factors influencing digital health literacy among older adults: a scoping review. Front Public Health 2024;12:1447747. [doi: 10.3389/fpubh.2024.1447747] [Medline: 39555039]
50. Prabhu S, Wanje G, Oyaro B, et al. Adaptation of a social vulnerability index for measuring social frailty among East African women. BMC Public Health 2022 Jan 24;22(1):167. [doi: 10.1186/s12889-022-12597-z] [Medline: 35073906]
51. McCall T, Asuzu K, Oladele CR, Leung TI, Wang KH. A socio-ecological approach to addressing digital redlining in the United States: a call to action for health equity. Front Digit Health 2022;4:897250. [doi: 10.3389/fdgth.2022.897250] [Medline: 35924138]
52. Verweel L, Newman A, Michaelchuk W, Packham T, Goldstein R, Brooks D. The effect of digital interventions on related health literacy and skills for individuals living with chronic diseases: a systematic review and meta-analysis. Int J Med Inform 2023 Sep;177:105114. [doi: 10.1016/j.ijmedinf.2023.105114] [Medline: 37329765]
53. Elikman D, Anderson R, Balish M, Zilbermint M. Mitigating health disparities: bridging the digital divide in modern health care. South Med J 2025 Jun;118(6):330-332. [doi: 10.14423/SMJ.0000000000001830] [Medline: 40456545]

Abbreviations

CV: coefficient of variation
DHL: digital health literacy
DigComp: digital competence framework
eHEALS: eHealth Literacy Scale
EHL: eHealth literacy
FGD: focus group discussion

Edited by A Stone; submitted 13.Aug.2025; peer-reviewed by A Azizi, L Cheng, Z Liang; revised version received 01.Mar.2026; accepted 09.Mar.2026; published 29.Apr.2026.

Please cite as:
Zhang X, Yuan Y, Jiang J
Development of a Digital Health Literacy Assessment Framework for Older Adults: Delphi-Based Study
J Med Internet Res 2026;28:e82334
URL: https://www.jmir.org/2026/1/e82334
doi:10.2196/82334

© Xingxia Zhang, Yongqing Yuan, Jie Jiang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The WONE Index as a Multidimensional Assessment of Stress Resilience: A Development and Validation Study

Lydia Genevieve Roos[1], PhD; Destiny Gilliland[1,2], BA; Kelsey Julian[1,3], MA; Reeva Misra[1], MA

[1]Walking on Earth, Ltd, International House 36-38 Cornhill, London, United Kingdom

[2]Department of Biobehavioral Health, College of Human Health and Development, The Pennsylvania State University, University Park, PA, United States

[3]Health Psychology PhD Program, College of Humanities and Earth and Social Sciences, University of North Carolina at Charlotte, Charlotte, NC, United States

Corresponding Author:
Lydia Genevieve Roos, PhD
Walking on Earth, Ltd, International House 36-38 Cornhill, London, United Kingdom

Abstract

Background: Stress resilience is a dynamic process shaped by the interaction between demands and adaptive resources. Existing measures assess stress and resilience as separate constructs, limiting their use in digital health and workplace interventions. An integrated measure capturing both domains is needed.

Objective: We developed and validated the WONE Index, a multidimensional stress resilience tool designed to measure both current stress load and adaptive resources among full-time working adults.

Methods: We developed the 32-item WONE Index through literature review, expert consultation, and iterative refinement to assess stress load and resilience resources across behavioral, cognitive, and social domains. Phase 1 (N=1005; United States– or United Kingdom–based full-time employees) evaluated the initial item pool using exploratory and confirmatory factor analyses to establish the preliminary factor structure and assess reliability and validity. Phase 2 (N=306; United States–based adults) expanded underperforming domains, refined items, and tested incremental validity, test-retest reliability, and measurement invariance. Data were collected online through CloudResearch (Connect) and Prolific (Prolific Academic Ltd) using secure survey platforms.

Results: Phase 1 supported a 2-domain structure: a Stress Load factor (Work Stress, Personal Stress, and Burnout) and a Resilience Resources factor (Emotion Regulation and Coping, Social Connectedness, and Sleep). Model fit indices were excellent (comparative fit index, CFI=0.95; Tucker-Lewis index, TLI=0.94; and root mean square error of approximation, RMSEA=0.05). Phase 2 replicated and extended this structure, expanding Resilience Resources into 7 domains (adding Purpose and Prosociality, Physical Activity, Dietary Intake, and Perseverative Thinking). Confirmatory factor analyses supported a 2-domain structure, comprising a higher-order Stress Load factor with 3 subdomains (Work Stress, Personal Stress, and Burnout) and a higher-order Resilience Resources factor with 7 subdomains (Emotion Regulation and Coping, Social Connectedness, Purpose and Prosociality, Sleep, Physical Activity, Dietary Intake, and Perseverative Thinking). The Stress Load model demonstrated excellent fit (χ^2_{33}=64.18; P=.01; CFI=0.99; TLI=0.98; RMSEA=0.06; and standardized root mean square residual=0.05), and the Resilience Resources model also fit well (χ^2_{443}=745.20, P<.001; CFI=0.94; TLI=0.94; RMSEA=0.05; and standardized root mean square residual=0.06). All subscales showed strong internal consistency (composite reliability: mean 0.84, SD 0.10; range 0.74 - 0.93) and excellent test-retest reliability over 3 weeks (intraclass correlation coefficients 0.77 - 0.90, 95% CI 0.87-0.93). The index showed strong convergent validity (r=0.73 with Connor-Davidson Resilience Scale and r=–0.66 with Perceived Stress Scale-4) and explained additional variance beyond established measures in predicting depression, anxiety, and well-being (ΔR^2=0.07 - 0.11; P<.001).

Conclusions: The WONE Index provides a psychometrically robust tool for assessing stress resilience capacity in working adults. Its integrated structure captures dynamic relationships between stress exposure and resilience resources, thereby supporting personalized intervention delivery in digital health platforms and organizational well-being programs.

(J Med Internet Res 2026;28:e81714) doi:10.2196/81714

KEYWORDS

psychological stress; occupational stress; psychological resilience; occupational health; psychometrics; factor analysis; measurement validation; instrumentation

XSL•FO
RenderX

Introduction

Background

Stress exposure is a pervasive challenge affecting individuals across work, personal, and societal domains, with well-documented impacts on psychological functioning, physical health, and quality of life [1,2]. While stress is common, individuals vary dramatically in their capacity to maintain well-being and adapt effectively under adversity—a quality captured by the concept of stress resilience [3-5]. As digital platforms increasingly seek to deliver personalized mental health support at scale, the ability to comprehensively assess individual resilience capacity has become critical [6,7]. However, existing resilience measures face fundamental limitations: they typically assess stress and resilience as separate constructs, provide insufficient detail to guide personalized interventions, and fail to capture the temporal dynamics through which resilience operates [8]. Addressing these measurement gaps is essential for advancing both resilience science and the development of effective digital mental health interventions across diverse contexts.

Resilience as a Dynamic, Multifaceted Process

We define stress resilience as the dynamic capacity to effectively respond to, recover from, and adapt in the face of adversity while maintaining psychological well-being and functional capacity [9-11]. Rather than reducing resilience to a fixed trait or the simple absence of pathology, this view conceptualizes it as an active and multifaceted process [9]. This process emerges from the interaction between current stress exposure and available protective resources or vulnerabilities, spanning cognitive, behavioral, social, affective, and physiological domains [4,5].

Job Demands-Resources Framework

This conceptualization aligns with the well-established job demands-resources (JD-R) model, which posits that psychological well-being results from the balance between demands that require sustained effort and resources that help manage those demands [12,13]. While originally developed for occupational contexts, this framework has been successfully extended to general stress and resilience processes, demonstrating that the demands-resources balance operates across life domains [14].

When demands exceed available resources, individuals experience strain and potential burnout. When resources are sufficient to meet demands, individuals can maintain well-being and even experience growth. This framework recognizes that resilience capacity is fundamentally determined by 2 interrelated but distinct domains: Stress Load (current psychological demands that require sustained effort) and Adaptive Resilience Resources (capacities that buffer demands and facilitate recovery).

Resilience Resources encompass the behavioral, cognitive, affective, and social capacities that enable adaptive responding to demands and facilitate recovery from adversity. These include emotion regulation skills, social support networks, physical health behaviors, cognitive flexibility, and meaning-making abilities. These resources function as protective factors against stress-related impairment [11]. Critically, many of these resources are modifiable through intervention, making them valuable targets for resilience-building efforts.

Having established the resource component of the JD-R framework, we now turn to the demands side of this balance, focusing on perceived stress. The JD-R framework aligns closely with the transactional model of stress [15-17], which emphasizes that the subjective appraisal of situational demands, rather than their objective frequency or intensity, is a strong determinant of psychological outcomes. We thus emphasize perceived stress rather than only objective stressors because individual appraisals of situational demands are more predictive of well-being and adaptation than event counts [16,17]. Within this perspective, cognitive appraisal serves as a regulatory mechanism that shapes whether a demand functions as a challenge that mobilizes engagement or as a hindrance that accelerates depletion.

The assessment of perceived stress presents both opportunities and challenges for resilience measurement. Perceived stress captures the subjectively experienced burden that directly impacts psychological functioning, making it more predictive of mental health outcomes than objective stressor inventories [18]. However, stress perceptions are themselves influenced by available resilience resources—individuals with stronger emotion regulation skills, social support, or coping strategies may appraise the same objective situation as less threatening [16,17]. This creates a dynamic relationship where resilience resources both protect against stress impact and influence stress perceptions themselves. Rather than viewing this complex relationship as a hindrance to accurate measurement, our framework recognizes this interdependence as fundamental to understanding stress resilience capacity.

Bringing these elements together, the demands-resources framework explains how stress resilience emerges from the interaction between these domains. High current demands deplete available psychological resources and can overwhelm coping capacity. On the other hand, strong resilience resources buffer against demand-related impact and facilitate faster recovery [14,19]. This dynamic interaction means that resilience capacity cannot be accurately assessed by measuring either demands or resources alone; comprehensive evaluation requires understanding both current demands (stress load) and available adaptive capacities (resilience resources).

However, a critical limitation remains in the classic JD-R model: it does not specify the temporal dynamics of these processes—the timescales over which demands erode resources or resources recover. This limitation is consequential for measurement. Because resilience represents an evolving balance rather than a fixed trait, capturing resilience capacity requires understanding both someone's current demands and resources as well as how that balance changes over time. An individual experiencing high demands with adequate resources today may show progressive depletion over subsequent weeks, while another may maintain equilibrium or recover capacity despite similar initial states.

XSL•FO
RenderX

The Need for New Measurement

The dynamic, interconnected nature of stress resilience processes creates significant measurement challenges. Traditional approaches that assess stress and resilience as separate, static constructs fail to capture the fundamental demands-resources interactions that determine resilience capacity [9]. This measurement gap is particularly problematic in digital mental health contexts. Assessment tools must simultaneously provide accurate evaluation of current psychological states and generate actionable insights for personalized intervention delivery.

Current measures have 3 critical limitations that prevent them from capturing these complex dynamics. First and most fundamentally, they fail to simultaneously assess both current perceived stress demands and available protective resources within the same framework. Traditional approaches measure either stress or resilience in isolation, missing the critical interplay between demands and adaptive capacities that determines whether individuals can maintain equilibrium, experience depletion, or build capacity. This separation makes it difficult to understand how an individual's current psychological burden relates to their available coping capacities, limiting the ability to identify whether intervention should focus on stress reduction, resource building, or both.

Second, among measures that assess resilience, most are either too general or too narrow to guide personalized intervention. General measures, such as the Connor-Davidson Resilience Scale (CD-RISC [20]), effectively assess overall resilience characteristics but do not provide the detailed mapping of specific modifiable capacities—such as emotion regulation skills, social support quality, sleep patterns, or physical activity levels—that practitioners and digital platforms need to develop targeted intervention plans. Conversely, aspect-specific measures, such as the Brief Resilience Scale (BRS [21]), assess particular processes, such as stress recovery, effectively but do not capture the full range of behavioral, cognitive, and social factors that contribute to overall resilience capacity.

Third, existing measures do not adequately capture the temporal dynamics of resilience processes. As established earlier, resilience represents an evolving balance rather than a fixed trait, yet current measures typically provide only single-time-point snapshots rather than enabling tracking of how the balance between demands and resources changes over time.

Digital mental health platforms require assessment tools that address all 3 limitations by bridging the gap between comprehensive scientific measurement and practical intervention guidance. Such measures must simultaneously assess both stress and resources within a unified framework, provide sufficient detail to guide personalized interventions across multiple modifiable domains, and enable tracking of resilience trajectories over time through repeated administration. Critically, these tools must also be optimized for digital delivery contexts—brief enough for repeated mobile administration without user burden, structured to enable automated scoring and feedback, and designed to inform real-time algorithmic personalization of intervention content. The need for theoretically grounded measures meeting these criteria has become increasingly urgent as digital platforms seek to provide effective, scalable mental health interventions that address the dynamic nature of stress resilience capacity.

To address these measurement and intervention challenges, we developed the Walking on Earth (WONE; Walking on Earth, Ltd) Index as the assessment foundation for WONE, a digital stress resilience platform. Unlike traditional assessment tools developed for paper-and-pencil or clinical interview administration, the index was specifically designed for digital delivery via the WONE platform using mobile-optimized item presentation, response formats suited to digital interfaces, and brevity enabling repeated monthly assessment without user fatigue.

Two-Phase Validation Strategy

The validation of the WONE Index used a systematic 2-phase approach designed to address the complexity of developing a theoretically grounded yet multidimensional measure. This strategy allowed empirical findings from Phase 1 to inform the refinement of Phase 2.

Phase 1 served as an exploratory foundation study with four primary aims: (1) establishing the initial factor structure through traditional psychometric approaches, (2) identifying items and domains with adequate psychometric properties, (3) assessing convergent and criterion validity with established measures, and (4) revealing measurement challenges requiring refinement to inform Phase 2.

Phase 2 functioned as a confirmatory refinement phase with four key objectives: (1) implementing refined measurement models based on Phase 1 findings, (2) addressing identified measurement limitations through expanded item development, (3) establishing comprehensive psychometric properties, including temporal stability, and (4) demonstrating incremental validity beyond existing gold-standard measures.

Methods

Methods (Phase 1)

Participants

Participants (N=1005; 502 in the United States, and 503 in the United Kingdom) were recruited through CloudResearch (US sample) and Prolific (UK sample) using nonprobability quota sampling. Inclusion criteria included being aged 18 - 65 years and working full-time (because the WONE user base targets adults working full-time), living in the United States or the United Kingdom (as the WONE platform currently operates primarily with companies based in these regions), and being fluent in English. We set quotas for gender within the CloudResearch and Prolific platforms to ensure adequate representation, such that neither men nor women are considered less than 45% of the study sample. Power analyses are provided in the "Data Analysis" section. Participants were balanced by gender (51% women in the United States and 48% in the United Kingdom), predominantly White (71% in the United States and 70% in the United Kingdom), with a mean age of 37.2 (SD 9.6) years in the United States and 36 (SD 10.4) years in the United

XSL•FO
RenderX

Kingdom. Complete demographic characteristics for Phase 1 are provided in Table S1 in Multimedia Appendix 1 [22-24].

Procedures

All study procedures were conducted remotely using Qualtrics (Qualtrics, LLC) survey software. Participants accessed the survey through secure links distributed via CloudResearch (US sample) and Prolific (UK sample). After providing informed consent, participants completed the WONE Index and validation measures in a single sitting (approximately 20 - 30 minutes). The survey was compatible with desktop and mobile devices. Built-in attention checks were included to ensure data quality. Participants who were unable to complete the survey in 1 session, except those who reported technical difficulties, or who did not pass at least 75% of attention checks, were disqualified. Participants were compensated for their time with US $6.25 or £4.50 for the US and UK samples, respectively.

Initial Item Development

The WONE Index item pool was developed through a targeted literature review and iterative expert consultation. We focused the literature review on identifying core constructs empirically linked to stress adaptation and resilience, drawing from theoretical and measurement work in stress appraisal, coping, self-regulation, social support, health behaviors, and well-being. We synthesized findings to determine the primary domains of adaptive capacity most consistently associated with mental health and performance outcomes.

Four health psychologists with expertise in stress and resilience then identified the constructs to be represented across both stress load and resilience resource domains, drawing on existing measures of perceived stress, burnout, resilience, coping, and health behaviors. In addition, we created new items to address domains not captured in previous tools, particularly those relevant to digital health and workplace contexts.

Through several collaborative meetings, the experts reviewed empirical evidence, debated conceptual boundaries, and progressively refined a broad initial pool into a smaller set of candidate items. Two specialists in digital health survey design then reviewed these items to evaluate clarity, readability, and redundancy for digital administration, after which additional consolidation produced the final 32-item instrument encompassing both current stress experiences and resilience resources.

Rather than imposing a fixed structure, the item pool was intentionally designed to allow empirical testing of whether stress load- and resource-related elements would remain distinct, overlap, or converge into higher-order domains. This theoretically grounded yet empirically flexible approach reflects our view of resilience as a dynamic, multidimensional construct rather than a static trait.

Current Stress Experiences

We tested 10 items to assess current stress load and burnout. For current stress, we assessed the following constructs: perceived stress, anxiety, and overwhelm at work and perceived stress, anxiety, and overwhelm at home/in one's personal life. Our approach to stress measurement for perceived work and personal stress acknowledges that individuals conceptualize and express psychological strain differently in real-world contexts [25,26]. Rather than relying on a single "stress" indicator, we assessed 3 related but distinct constructs—stress, anxiety, and overwhelm—within both work and personal life domains.

Research demonstrates that individuals vary considerably in how they conceptualize and articulate psychological distress, with significant cultural and individual differences in whether experiences are described as "stress," "anxiety," "overwhelm," or other terms [25,27]. Although academic literature often treats these as separate constructs, everyday users may not make such distinctions when describing their experiences [28]. Cultural, educational, and personal factors fundamentally influence how psychological experiences are articulated and understood [29,30].

For example, some individuals may more readily identify with feeling "overwhelmed" by their responsibilities, while others may describe similar experiences as "stress" or "anxiety." By capturing all 3 dimensions, the WONE Index ensures a comprehensive assessment regardless of how someone naturally conceptualizes their distress, while also reflecting the interconnected nature of these stress-related experiences in daily life [31]. This inclusive measurement strategy maximizes the tool's utility across diverse user populations and contexts.

We also included 4 indicators of burnout, which assessed cynicism, disengagement, reduced productivity, and mental exhaustion related to work, reflecting the core dimensions of occupational burnout as conceptualized in the Maslach Burnout Inventory framework [32,33]. These burnout indicators capture chronic workplace stress that differs qualitatively from acute stress experiences.

Resilience Resources

Resilience resources represent the psychological, social, and behavioral capacities that enable individuals to adapt to stress, recover from strain, and maintain well-being. This domain reflects the resources side of the WONE Index framework, balancing the stress-load indicators as described in the "Current Stress Experiences" section above.

Altogether, we tested 22 items to measure resilience resources. Nine emotion regulation and coping items consisted of emotion regulation, perceived ability to cope, tendency to bounce back after a stressor, adaptability, stress-related growth mindset, effective coping techniques, perspective-taking ability, frequency of positive affect, and perceived life meaning and purpose. Three self-efficacy and perceived control items assessed: self-efficacy beliefs, perceived control, and the sense that things were going smoothly. Two social support items assessed: trusted support availability and support satisfaction. Four sleep and energy items assessed: sleep quality, duration, latency, disturbances, and overall energy levels. Two dietary intake items included alcohol consumption and nutritious diet quality. Finally, 2 physical activity items assessed moderate and vigorous physical activity and sedentary time.

XSL•FO
RenderX

Response Formats

Items used varying response formats optimized for each domain, all on 1 - 5 point scales with domain-appropriate anchors, where higher numbers indicate better resilience. For example, response options included frequency scales (eg, "Never" to "Always"), agreement scales (eg, "Strongly Disagree" to "Strongly Agree"), and intensity scales (eg, "Not at all" to "Extremely"), selected based on item content and established measurement practices for each construct.

The WONE Index is organized into sections that pair construct-specific introductory instructions with domain-appropriate response formats. For example, stress-related items begin with "In the past month, how often have you felt…," burnout items use agreement anchors, and health behavior items use categorical duration or quality scales. These variations are intentional, aligning response structure with the theoretical nature of each construct to ensure interpretability across diverse domains of stress and resilience. All items used a standardized 1-month timeframe (eg, "Over the past month..." or "In an average week over the past month…") to ensure consistency, capture relatively stable patterns, and remain sensitive to change.

External Validation Measures

Validated external scales were administered to assess criterion validity across stress, resilience, and mental health domains. These measures were used in both studies 1 and 2.

Stress Assessment

Perceived stress was assessed using the Perceived Stress Scale-4 (PSS-4 [18]), a brief version of one of the most widely used instruments for measuring stress perception [34]. The PSS-4 has demonstrated acceptable internal consistency (α=.60-.82 [34]) and good test-retest reliability over 4 weeks (*r*=0.73 [35]). It also shows strong convergent validity with measures of distress and discriminant validity from indicators of well-being [18]. The PSS-4 was selected for its brevity and suitability for rapid data collection while maintaining adequate psychometric robustness.

Resilience Assessment

Two established measures were used to assess resilience. The 10-item CD-RISC-10 [36] measures the ability to cope with adversity and demonstrates excellent internal consistency (α=.85-.92) as well as good convergent and discriminant validity across populations [36]. Although test-retest reliability was not re-examined for the 10-item version, the original 25-item CD-RISC shows strong temporal stability (*r*=0.87 [20]).

The 6-item BRS [21] measures the ability to recover or "bounce back" from stress and demonstrates high internal consistency (α=.80-.91), 1-month test-retest reliability (*r*=0.69), and robust convergent and discriminant validity with measures of affect, optimism, and neuroticism [21].

Mental Health and Well-being Outcomes

Depressive symptoms were measured using the Patient-Reported Outcomes Measurement Information System Short Form-8a (PROMIS-SF-8a [37]), which demonstrates excellent internal consistency (α>.90) and robust convergent validity with legacy measures such as the Patient Health Questionnaire-9 (PHQ-9) and Center for Epidemiologic Studies Depression (CES-D) [38] scales. Although test-retest data are not yet available for the fixed 8-item form, the PROMIS Depression computerized adaptive test shows high temporal stability (intraclass correlation coefficient [ICC]=0.86 [39]).

Anxiety was assessed with the Generalized Anxiety Disorder-7 (GAD-7 [40]), which has high internal consistency (α=.89-.92), good test-retest reliability (*r*=.83), and robust convergent and discriminant validity relative to other anxiety and depression scales [41].

Well-being was measured with the World Health Organization-5 Well-Being Index (WHO-5 [42]), which shows good internal consistency (α=.80-.90) and strong 5-day test-retest reliability (ICC=0.87 [43]), as well as robust convergent validity with measures of life satisfaction and positive affect [42].

Data Analysis

Statistical Software

Statistical analyses were conducted using IBM SPSS Statistics (version 29.0). Confirmatory factor analyses were performed using IBM SPSS AMOS (version 22.0).

Data Quality, Screening, and Analytic Assumptions

Participants' data were excluded from analyses if their responses to qualitative screening questions were nonsensical or irrelevant to the questions asked, as these were indicators of either automated responses or insufficient English fluency for valid participation. All survey items were required responses, resulting in no missing data for the final analytic sample.

Descriptive statistics and normality assessments were conducted for all key variables. Skewness and kurtosis values were all within the acceptable range of ±1.0, indicating reasonably symmetric distributions without problematic tail behavior. While Shapiro-Wilk tests indicated significant departures from normality for all variables (all *P*s<.001), this is expected given the large sample sizes, as these tests become overly sensitive to minor deviations with substantial samples [44]. Visual inspection of histograms and Q-Q plots confirmed that distributions were approximately normal with no severe violations.

Outlier analysis using boxplot methods identified a small number of extreme cases across variables (1 - 11 outliers per variable, representing approximately 1% - 4% of cases). These outliers were retained as they represented valid responses within the expected range of the constructs and did not appear to result from data entry errors. Given the acceptable skewness and kurtosis values, robust nature of maximum likelihood estimation, and large sample sizes, the data were deemed appropriate for the planned factor analyses and correlational analyses.

Analytic Procedures

Exploratory Factor Analysis

Exploratory factor analysis (EFA) was conducted using principal axis factoring with Promax rotation to identify the underlying

XSL•FO
RenderX

factor structure. Factor retention was determined using multiple criteria: (1) eigenvalues >1.0 (Kaiser criterion), (2) scree plot inspection, (3) theoretical interpretability, and (4) total variance explained. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett test of sphericity were used to assess data appropriateness for factor analysis. We systematically evaluated items and retained them based on the following criteria: primary factor loadings ≥0.40, cross-loadings <0.30, difference between primary and secondary loadings ≥0.15, communalities ≥0.30, and theoretical coherence with factor interpretation.

Confirmatory Factor Analysis

We conducted a higher-order confirmatory factor analysis (CFA) using maximum likelihood estimation with robust SEs (Huber-White). The higher-order factors in the model consisted of latent factors related to (1) current stress experiences and (2) resilience resources. We evaluated model fit using multiple indices: χ^2/df ratio (<5.0 acceptable and <3.0 good), comparative fit index (CFI ≥0.90 acceptable and ≥0.95 good), Tucker-Lewis index (TLI ≥0.90 acceptable and ≥0.95 good), root mean square error of approximation (RMSEA ≤0.08 acceptable and ≤0.06 good) and standardized root mean square residual (SRMR ≤0.08 acceptable and ≤0.06 good).

Reliability Assessment (Phase 1)

We assessed reliability using composite reliability (CR) estimates, which are more appropriate than Cronbach alpha for factor-based models, particularly in structural equation modeling frameworks [45]. CR was calculated using standardized factor loadings: $CR=(\Sigma\lambda)^2/[(\Sigma\lambda)^2 + \Sigma(1-\lambda^2)]$ [46]. Reliability ≥0.70 was considered acceptable for subscales and ≥0.90 for total scales.

Validity Testing (Phase 1)

Convergent validity was assessed through multiple approaches: (1) within the CFA by examining factor loading magnitudes (≥0.60 preferred and ≥0.40 minimum), statistical significance, and average variance extracted (AVE ≥0.50); and (2) through Pearson correlations with theoretically related established measures administered concurrently.

Discriminant validity was evaluated through multiple methods: (1) examining the CFA factor structure for cross-loadings and modification indices (MIs) suggesting misspecification, (2) computing the heterotrait-monotrait (HTMT) ratio with conservative (0.85) and liberal (0.90) thresholds, and (3) evaluating correlation patterns demonstrating stronger relationships between similar constructs than dissimilar constructs.

Concurrent validity was assessed using Pearson correlations with established measures administered simultaneously, specifically examining relationships between WONE Index scales and corresponding validated measures of the same constructs (eg, WONE Resilience Index with CD-RISC and BRS).

Criterion-related validity was evaluated by examining correlations between WONE Index scales and important mental health and well-being outcomes, including depression, anxiety, and psychological well-being.

Power Analysis

We conducted comprehensive a priori power analyses to determine the appropriateness of our targeted sample size of 1000 for the planned statistical tests, including factor analysis, reliability testing, validity assessment, and measurement invariance testing. We used a 2-tailed significance level of α=.05 for all analyses. Detailed power analyses are provided in Multimedia Appendix 1.

Methods (Phase 2)

Participants and Procedure

We used the same nonprobability quota sampling recruitment strategy as Phase 1, except that we focused on US participants and recruited through CloudResearch. Participants were asked to complete surveys at 2 time points spaced 3 weeks apart. Participants were compensated US $5 at Time 1 and US $4 at Time 2; payment at Time 2 was slightly lower due to fewer items (eg, demographics). The analyses presented here focus on Time 1 data (N=306) for all EFAs and CFAs, with Time 2 used to assess test-retest reliability. Power analyses are provided within the “Data Analysis” section.

Theoretical Framework

Phase 2 was designed to refine and confirm the factor structure of the WONE Index based on the findings from Phase 1.

Phase 1 demonstrated that stress-related items formed a coherent, well-fitting factor structure, while several resilience domains—particularly health behaviors—required additional development. The strong validity evidence for the core stress and resilience skills components provided a solid foundation for expansion, while the challenges with nutrition, physical activity, and other behavioral domains indicated the need for larger item pools and more comprehensive domain coverage.

Phase 2 addresses these findings through several key design features. First, we substantially expanded the item pools for domains that showed promise but required development, particularly health behavior and social connection domains. Second, we adopted a domain-specific analytical approach, rather than the hierarchical approach used in Phase 1, allowing for comprehensive factor identification within each theoretical area while reducing model complexity. Third, we maintained the successful stress measurement structure from Phase 1 while expanding the resilience assessment to capture the full breadth of protective factors identified in resilience literature.

Domain Structure, Changes, and Measurement Strategy

Based on the findings and item performance from Phase 1, we adopted a staged, domain-specific approach that reflects both empirical patterns and theoretical considerations. The item pool was divided into 2 conceptually distinct domains for psychometric modeling and was analyzed within their respective models.

XSL•FO
RenderX

Resilience Resources Domain

Included items assessing behavioral and psychological capacities expected to support stress resilience. These spanned Emotion Regulation and Coping, Social Connectedness, Prosocial Values, Sleep, Physical Activity, and Dietary Intake. This comprehensive approach allowed for detailed examination of each resilience pathway while maintaining theoretical coherence across domains.

Stress Domain

We included the previously tested items under the Stress domain, reflecting stress experiences in one's work and personal life, with strategic modifications based on Phase 1 findings and contemporary contextual factors. In addition, recognizing the chronic stress that has characterized daily life since the COVID-19 pandemic [47], including polarizing societal and political experiences that have become the norm within the past decade [48], global conflicts, and economic uncertainty, we added 1 item specifically assessing stress, anxiety, and overwhelm due to current societal, political, and economic issues. This addition acknowledges that contemporary stress experiences extend beyond traditional work and personal life domains to include broader societal stressors [49] that have become increasingly prominent in recent years.

Measurement Strategy

Items were scored such that higher scores indicate greater stress or resilience resources when assessed separately, and the stress-load items were reverse-coded when assessed together so that higher overall scores reflected greater resilience. We analyzed both stress and resilience within their own respective models to reduce model complexity while still maintaining theoretical relevance. This staged approach is supported by previous psychometric validation practices, especially in multidimensional well-being assessments [50], and it allows for comprehensive development of each theoretical domain while maintaining the overarching framework that proved successful in Phase 1.

The domain-specific approach also identifies optimal factor structures for each area without the constraints imposed by simultaneous analysis of all domains. This approach is particularly important for health behavior domains that showed measurement challenges in Phase 1, allowing for detailed examination of their factor structure and item performance.

The WONE Index was developed and is copyrighted by Walking on Earth, Inc. The full administration materials, including detailed instructions and response options, are proprietary and therefore not publicly available. However, the measure may be available for use upon reasonable request to WONE via the website.

External Validation Measures

The same validated external scales from Phase 1 were administered in Phase 2 to assess validity, with 2 additional measures included.

Depressive Symptoms

In addition to the PROMIS-SF-8a [37], which was used in Phase 1, we also administered the PHQ-8 [51], a modified version of the PHQ-9 that omits the ninth item on thoughts of death or self-harm. The PHQ-8 demonstrates high internal consistency (α=.82-.89 [51,52]) and strong short-term test-retest reliability (r=0.89 [52]). It also demonstrates diagnostic performance comparable to the PHQ-9, with nearly identical sensitivity, specificity, and overall accuracy [52,53]. We included this measure to enable comparison with the digital health literature that frequently uses PHQ-based instruments and to triangulate depression assessment using complementary approaches. While the PROMIS-SF-8a provides an abstract, mood-focused assessment with reduced somatic bias, the PHQ-8 offers a behaviorally anchored assessment directly mapping onto *DSM* (*Diagnostic and Statistical Manual of Mental Disorders*) criteria [51,54]. Together, the 2 instruments strengthen construct validation by capturing depressive symptoms from 2 distinct measurement approaches.

Personality Assessment

To assess discriminant validity, we included the brief Big Five Inventory (BFI-10 [55]). This short-form instrument measures extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience, using 2 items per domain. The BFI-10 demonstrates acceptable internal consistency across traits (α=.65-.82), strong test-retest reliability (r=0.72-0.84), and high convergent validity with the full 44-item BFI (r=0.91-0.96 [55]).

Data Quality, Screening, and Analytic Assumptions

Participants' data were excluded from analyses if their responses to qualitative screening questions were nonsensical or irrelevant to the questions asked, as these were indicators of either automated responses or insufficient English fluency for valid participation. All survey items were required responses, resulting in no missing data for the final analytic samples (N=1005 for Phase 1 and N=306 for Phase 2).

Descriptive statistics and normality assessments were conducted for all key variables in both studies. Skewness and kurtosis values were all within the acceptable range of ±1.0 across both samples, indicating reasonably symmetric distributions without problematic tail behavior. Shapiro-Wilk tests indicated significant departures from normality for all variables in Phase 1 (all P<.001), which is expected given the large sample size. In Phase 2, normality tests were significant for external validation measures but not significant for the full WONE Index or Stress Subscale, with the Resilience Subscale showing marginal significance (P=.05). These patterns reflect the expected sensitivity of normality tests to sample size, with larger samples detecting minor deviations that are not practically meaningful [44]. Visual inspection of histograms and Q-Q plots confirmed that distributions were approximately normal with no severe violations.

Outlier analysis using boxplot methods identified minimal extreme cases: Phase 1 showed 1‑11 outliers per variable (representing 1%‑4% of cases), while Phase 2 showed substantially fewer outliers (0‑1 outliers per variable,

XSL•FO
RenderX

representing <1% of cases). These outliers were retained as they represented valid responses within the expected range of the constructs and did not appear to result from data entry errors. Given the acceptable skewness and kurtosis values, the robust nature of maximum likelihood estimation, and adequate sample sizes, the data were deemed appropriate for the planned factor analyses and correlational analyses.

Analytic Procedures

EFA

We used a staged EFA approach to identify the underlying factor structure within each domain. Principal axis factoring with Promax (oblique) rotation was used to accommodate the anticipated correlations between factors within each domain. For each EFA, sampling adequacy was assessed via the KMO statistic and Bartlett test of sphericity. Factor retention decisions were guided by eigenvalues >1.0, scree plot inspection, and theoretical coherence.

Items were retained if they met the following criteria: (1) primary loading ≥0.30, (2) minimal cross-loadings, and (3) thematic and theoretical fit with the emerging factor.

CFA

Following EFA, we specified and tested confirmatory factor models for each domain using maximum likelihood estimation with robust SEs. This domain-specific modeling strategy was selected to preserve theoretical clarity, allow for comprehensive factor development within each area, and build on the empirical structure observed in Phase 1, where resilience-related indicators and stress indicators showed distinct loading patterns.

Current Stress Model

Items loaded onto 3 first-order latent factors identified in Phase 1: Work Stress, Personal Stress, and Burnout. This model replicates the successful stress structure from Phase 1 while incorporating any additional stress-related items developed for Phase 2.

Resilience Resources Model

Items loaded onto 7 first-order latent factors representing expanded domains: Emotion Regulation and Coping, Social Connectedness, Purpose and Prosociality, Sleep, Physical Activity, Dietary Intake, and Perseverative Cognition. This expanded structure addresses the measurement challenges identified in Phase 1 by providing comprehensive coverage of resilience domains with sufficient items for stable factor identification.

The separate modeling approach allows for optimal factor structure identification within each domain while maintaining the theoretical framework established in Phase 1. Following successful domain-specific CFAs, the models can be integrated to test the overarching 2-factor higher-order structure (Stress and Resilience Resources) that demonstrated excellent validity in Phase 1. Model fit was evaluated using standard indices: CFI ≥0.90, RMSEA ≤0.08, and SRMR ≤0.08, consistent with conventional guidelines [56,57]. Model modifications were considered only when theoretically justifiable and indicated by MIs ≥10. When MIs suggested potential improvements, we prioritized residual covariances and factor-to-error covariances over cross-loadings to maintain interpretable factor structures while acknowledging theoretically expected relationships among constructs.

Reliability Assessment (Phase 2)

Reliability assessment was performed using the same method as described above for phase 1.

Validity Testing (Phase 2)

Convergent validity, concurrent validity, and criterion-related validity was assessed through multiple approaches, as described above for phase 1.

Test-Retest Reliability

A subset of participants completed the WONE Index at a second time point approximately 3 weeks after their initial assessment to evaluate temporal stability. Test-retest reliability was assessed using ICC with a 2-way mixed-effects model for consistency of agreement. Single-measure ICCs (ICC[2,1]) were calculated for the full WONE Index, Stress Subscale, and Resilience Subscale. ICC values were interpreted using established guidelines: <0.50=poor, 0.50 - 0.75=moderate, 0.75 - 0.90=good, and >0.90=excellent reliability [58].

Incremental Validity

Incremental validity was assessed through hierarchical multiple regression analyses to determine whether the WONE Index provides a meaningful prediction of mental health outcomes beyond established measures. Models tested incremental validity beyond: (1) the Perceived Stress Scale, (2) the CD-RISC, (3) the BRS, and (4) combined established measures. Mental health outcomes included depressive symptoms, anxiety symptoms, and well-being. Incremental validity was demonstrated by statistically significant R^2 change (ΔR^2) when the WONE Index was added to models.

Measurement Invariance Testing

We assessed measurement invariance of the Current Stress model using multigroup CFA, following the standard stepwise approach: configural, metric, scalar, and strict invariance. Each step imposed increasingly restrictive equality constraints across groups. Model fit was evaluated using commonly accepted thresholds for change in fit indices (ΔCFI ≤0.01, ΔRMSEA ≤0.015, and ΔSRMR ≤0.03 for metric invariance and ≤0.01 for scalar and strict invariance [59]). Fit was considered adequate when CFI and TLI were ≥0.95, RMSEA ≤0.06, and SRMR ≤0.08.

Power Analysis

We conducted comprehensive a priori power analyses to determine the appropriateness of our targeted sample size of 300 for the planned statistical tests. We used a 2-tailed significance level of α=.05 for all analyses. Detailed power analyses are provided in Multimedia Appendix 1.

XSL•FO
RenderX

Weighting Methodology

Hybrid Weighting Strategy Development

The WONE Index used a novel hybrid weighting approach that systematically integrates empirical prediction with theoretical importance and practical actionability. This methodology addresses a critical limitation in traditional psychometric approaches, in which purely data-driven weighting may undervalue theoretically important but empirically complex constructs, particularly in intervention-focused applications.

Empirical Weight Derivation

Empirical weights were derived through multiple regression analyses using each validation measure as a dependent variable and the 10 WONE factor scores as predictors. Standardized beta coefficients were extracted from regression models predicting 6 priority outcomes: CD-RISC-10, BRS, PSS-4, PHQ-8, GAD-7, and WHO-5.

Final empirical weights were calculated using a weighted average of standardized beta coefficients, with priority weights assigned based on theoretical importance: CD-RISC (25%), BRS (22%), PSS-4 (18%), PHQ-8 (15%), GAD-7 (12%), and WHO-5 (8%). For measures where higher scores indicate poorer outcomes (PSS-4, PHQ-8, and GAD-7), beta coefficients were reverse-scored to ensure all factors contributed positively to overall resilience scoring.

Theoretical Weight Framework

Theoretical weights were developed through a literature review [11,60-62] and author consensus, guided by the WONE Index's core conceptual model emphasizing resilience capacity through perceived stress (resource depletion) and modifiable protective/risk factors (resource building/depletion). Theoretical weights were assigned based on three primary criteria: (1) strength of empirical evidence linking the construct to stress resilience outcomes, (2) theoretical importance within established resilience frameworks, and (3) behavioral actionability for intervention purposes.

Hybrid Integration and Final Weights

The final hybrid weights were calculated using a 50/50 integration formula: Final Weight = (Theoretical Weight×0.5) + (Empirical Weight×0.5). This approach preserves empirical predictive validity while ensuring meaningful representation of theoretically important and behaviorally actionable constructs. A minimum weight floor of 5% was applied to ensure all constructs maintain sufficient influence for generating actionable user recommendations in digital health application contexts. Constructs falling below this threshold after hybrid calculation were elevated to 5%, with proportional reductions applied to higher-weighted constructs to maintain a total weight sum of 100%.

Ethical Considerations

This investigation was conducted in accordance with the Declaration of Helsinki and was determined to be exempt from Institutional Review Board (IRB) review after meeting ethical standards as evaluated by the Western Clinical Group IRB under 45 CFR § 46.104(d)(2) on November 27, 2024 (IRB ID 20244893). All participants provided informed consent before participation. Surveys were completed using SurveyMonkey (SurveyMonkey Inc), a GDPR-compliant platform with data encryption in transit and at rest, password-protected access controls, and secure data storage. Phase 1 participation was fully anonymous; no personally identifiable information was collected, and responses were recorded without linkage to any identifiers. In Phase 2, participant tracking across survey waves was managed through CloudResearch's "Waves" feature, which assigns a randomly generated 32-character hexadecimal string as an identifier. This procedure enabled longitudinal matching of responses while preventing access to identifying information. Researchers did not have access to participants' identities, and because surveys were administered in SurveyMonkey independently of CloudResearch, neither platform possessed both identity and response data. All data were stored on encrypted servers and analyzed in deidentified form.

Results

Phase 1: Results

Sample Demographics

Phase 1 included 1005 adults from the United States and the United Kingdom. Both samples were balanced by gender and predominantly White (refer to Table S1 in Multimedia Appendix 1).

Reliability of External Validation Measures

Stress Assessment

The PSS-4 demonstrated acceptable to good internal consistency, with reliability improving from Phase 1 (Cronbach α=0.79, 95% CI 0.77 - 0.81; McDonald ω=0.84) to Phase 2 (Cronbach α=0.85, 95% CI 0.82 - 0.88; McDonald ω=0.88). These reliability coefficients exceeded conventional thresholds for research use, supporting the measure's suitability for criterion validity analyses.

Resilience Assessment

Both resilience measures exhibited excellent internal consistency across studies. The CD-RISC-10 maintained consistent reliability (Phase 1: Cronbach α=0.91, 95% CI 0.90 - 0.92; McDonald ω=0.92 and Phase 2: Cronbach α=0.91, 95% CI 0.90 - 0.93; McDonald ω=0.93), while the BRS improved from Phase 1 (Cronbach α=0.88, 95% CI 0.87 - 0.90; McDonald ω=0.92) to Phase 2 (Cronbach α=0.92, 95% CI 0.91 - 0.94; McDonald ω=0.94). The convergence between reliability indices confirms robust psychometric performance and establishes a strong foundation for resilience-related criterion validity.

Mental Health and Well-Being Outcomes

All mental health measures demonstrated exceptional reliability with coefficients consistently exceeding 0.90. The PROMIS Depression Short Form-8a achieved the highest consistency (Phase 1: Cronbach α=0.95, 95% CI 0.95 - 0.96; McDonald ω=0.97 and Phase 2: Cronbach α=0.96, 95% CI 0.95 - 0.97; McDonald ω=0.97), followed by the Generalized Anxiety Disorder-7 (Phase 1: Cronbach α=0.92, 95% CI 0.91 - 0.93; McDonald ω=0.94 and Phase 2: Cronbach α=0.93, 95% CI

XSL•FO
RenderX

0.92 - 0.94; McDonald ω=0.96) and WHO-5 (Phase 1: Cronbach α=0.91, 95% CI 0.90 - 0.92; McDonald ω=0.93 and Phase 2: Cronbach α=0.92, 95% CI 0.91 - 0.94; McDonald ω=0.92). These high reliability values ensure that criterion validity relationships reflect true associations rather than measurement error.

EFA

Data demonstrated excellent suitability for factor analysis (KMO=0.926; Bartlett test: χ^2_{351}=13002.56; *P*<.001). Initially, 7 factors emerged—Resilience Skills and Beliefs, Work Stress, Personal Stress, Sleep, Burnout, Social Support, and Control—which accounted for 60.4% of the combined variance. Items related to alcohol, nutrition, sedentary time, and physical activity did not load onto any factor at ≥0.3, so they were removed. These findings indicated that health behavior domains required expanded item pools with more comprehensive measurement to achieve stable factor identification, which informed item development for Phase 2. The energy item loaded erroneously onto the Control factor and demonstrated a low Sleep loading, so it was also removed.

The final EFA with 27 items yielded a clear, interpretable 6-factor solution explaining 64.2% of the total variance: Resilience Skills and Beliefs, Work Stress, Personal Stress, Sleep, Burnout, and Social Support. Notably, the mental exhaustion item loaded onto the Work Stress factor instead of the Burnout factor. Although unexpected, mental exhaustion is typically the first symptom of burnout [63] and aligns conceptually with workplace stress; thus, the item was retained within the Work Stress factor. These items were then tested using the same factor placements in CFAs.

CFA

A 2-factor higher-order CFA was conducted to test the proposed measurement model. The model specified Stress as a second-order latent construct comprising 3 first-order factors (Work Stress, Personal Stress, and Burnout) and Resilience Resources as a second-order latent construct comprising 3 first-order factors (Resilience Skills and Beliefs, Social Support, and Sleep Quality). The model included 33 observed indicators across 6 first-order factors. The initial higher-order model demonstrated adequate fit (χ^2_{317}=11,593.27; *P*<.001; CFI=0.90; TLI=0.89; RMSEA=0.07, 95% CI 0.06 - 0.07; SRMR=0.058).

Following examination of MIs, residual covariances were freed for 15 conceptually related item pairs with MIs exceeding 20. The refined model achieved excellent fit (χ^2_{302}=993.30; *P*<.001; CFI=0.95; TLI=0.94; RMSEA=0.049, 95% CI 0.046 - 0.052; SRMR=0.052). Although the chi-square test remained statistically significant—consistent with large sample size sensitivity [64]—all practical fit indices met or exceeded conventional standards for excellent model fit [56].

The Resilience Skills and Beliefs factor demonstrated strong overall performance, with loadings ranging from acceptable to strong (0.40≤ λ ≤0.76). Social Support emerged as the most cohesive first-order factor, with both indicators demonstrating exceptionally strong loadings (λ >0.80). Sleep showed a mixed but interpretable pattern, with 1 dominant indicator (sleep quality; λ=0.87) and 3 secondary indicators with more modest loadings (0.52≤ λ ≤0.63). The 3 stress factors (Work Stress, Personal Stress, and Burnout) all demonstrated strong psychometric properties (0.40≤ λ ≤0.76; refer to Table S3 in Multimedia Appendix 1).

Second-Order Factor Structure

The higher-order structure received strong empirical support, with all second-order loadings exceeding 0.60 and demonstrating appropriate magnitudes. The Stress factor showed balanced contributions from its 3 constituent factors, with loadings ranging from 0.71 to 0.741, indicating that Work Stress, Personal Stress, and Burnout represent relatively equal contributors to the overarching stress construct.

Resilience Resources exhibited a more hierarchical structure, with Resilience Skills and Beliefs serving as the dominant indicator (λ=0.94) and Social Support and Sleep Quality contributing more modestly but substantially (λ=0.67 and 0.62, respectively). This pattern suggests that while all 3 resources are important components of resilience, individual resilience skills may serve as the core organizing feature of this higher-order construct.

Reliability and Validity Testing

Internal Consistency Reliability

All factors demonstrated acceptable to excellent internal consistency reliability, as provided in Table S4 in Multimedia Appendix 1. CR estimates exceeded established thresholds across all factors. The full WONE Index showed excellent reliability, while subscale reliability was consistently strong across all first-order factors. All reliability estimates met or exceeded the 0.70 threshold for acceptable reliability, with most surpassing the 0.80 standard for good reliability.

Convergent Validity

The WONE Index demonstrated strong convergent validity with theoretically related established measures via correlations with externally validated measures. Resilience constructs showed substantial positive correlations with validated resilience measures: the Resilience Index correlated strongly with the CD-RISC (*r*=0.74; *P*<.001) and moderately with the BRS (*r*=0.68; *P*<.001). These correlations fall within the large effect size range [65], indicating that the WONE Resilience Index captures similar underlying resilience constructs while maintaining distinctiveness.

Stress constructs exhibited strong convergent validity with established stress and mental health measures. The Stress subscale correlated substantially with the PSS (*r*=0.66; *P*<.001), demonstrating convergent validity for the overall stress construct. Individual stress components showed expected relationships: positive correlations with depressive symptoms, anxiety symptoms, and perceived stress, as well as a negative correlation with well-being, were observed, supporting the validity of the stress measurement model.

AVE estimates, calculated as the mean of the communalities (R^2) for each factor's items (refer to Table S3 in Multimedia Appendix 1), provided evidence of convergent validity for most

XSL•FO
RenderX

factors. Four factors met or exceeded the 0.50 criterion, while 2 factors showed AVE values below 0.50 but above 0.40. While AVE values below 0.50 suggest some concern for convergent validity, the strong CR estimates and substantial factor loadings discussed previously indicate that convergent validity is still adequate for these factors [46].

Discriminant Validity

HTMT analysis provided strong evidence for discriminant validity. All HTMT values fell well below the conservative 0.85 threshold, with an average value of 0.49, indicating excellent discriminant validity between all factor pairs (refer to Table S4 in Multimedia Appendix 1).

Concurrent Validity

The WONE Index demonstrated excellent concurrent validity with established criterion measures (refer to Table S5 in Multimedia Appendix 1). The Resilience subscale showed strong concurrent validity with both the CD-RISC and BRS, 2 well-validated resilience measures with different theoretical emphases. The Stress subscale demonstrated substantial concurrent validity with the PSS, a widely used measure of subjective stress appraisal.

Criterion-Related Validity

The WONE Index showed strong criterion-related validity with important mental health and well-being outcomes (refer to Table S5 in Multimedia Appendix 1). The full WONE Index demonstrated robust associations with all criterion measures, showing positive correlations with well-being and negative correlations with both depressive and anxiety symptoms.

The Stress subscale exhibited strong criterion-related validity, with negative associations with well-being and positive associations with both depressive and anxiety symptoms. The Resilience subscale demonstrated complementary criterion-related validity, with positive associations with well-being and negative associations with both depressive and anxiety symptoms.

Phase 2 Results

Sample Demographics

Phase 2 included 306 US adults. The sample was balanced by gender and predominantly White (refer to Table S6 in Multimedia Appendix 1). Although the study was open to participants aged 18 - 64 years, the youngest participant was aged 20 years, and the oldest was aged 63 years, resulting in an age range of 20 - 63 years.

Reliability of External Validation Measures

Depressive Symptoms

The PHQ-8 demonstrated excellent internal consistency (Cronbach α=0.90, 95% CI 0.88-0.91; McDonald ω=0.93). This strong reliability supports its use as a complementary measure to the PROMIS-SF-8a and strengthens construct validation through convergent evidence across different measurement approaches.

Personality

Reliability for the BFI-10 subscales was modest, consistent with expectations for short-form personality measures. Extraversion (α=.71 and ω=0.71) and Neuroticism (α=.77 and ω=0.77) showed adequate reliability, while Conscientiousness (α=.58 and ω=0.63), Agreeableness (α=.57 and ω=0.57), and Openness (α=.58 and ω=0.63) were lower. These results are consistent with previous findings indicating that short-form personality measures often exhibit attenuated reliability due to their limited item coverage per construct.

EFA Results

Stage 1: Resilience Factors Analysis

Data Suitability and Sample Characteristics

Analysis was conducted on 306 participants using 51 resilience-related items. Data demonstrated excellent suitability for factor analysis: KMO measure of sampling adequacy=0.91 (excellent), and Bartlett test of sphericity was highly significant (χ^2_{528}=5790.89; P<.001), confirming the appropriateness of the correlation matrix for factorization.

Factor Extraction and Retention Strategy

Based on theoretical considerations, we specified a maximum of 7 factors corresponding to hypothesized resilience domains: (1) Emotion Regulation and Coping Tendencies, (2) Resilience-Related Beliefs, (3) Positive Psychology Buffers, (4) Social Connectedness, (5) Dietary Intake, (6) Physical Activity, and (7) Sleep.

Item Reduction and Refinement Process

Through iterative analysis, we applied stringent psychometric criteria for item retention. Items were systematically removed based on (1) low communalities (<0.30), (2) inadequate factor loadings (<0.40), and (3) theoretically inappropriate cross-loadings. This process resulted in the removal of 15 items.

Final Factor Structure and Variance Explained

The final 7-factor solution with 33 items explained 66.04% of the total variance, demonstrating substantial explanatory power. Factor loadings ranged from 0.42 to 1.00, with the majority of items (26/33, 78.79%) exceeding 0.50, indicating strong factor-item relationships. Alternative solutions with 5 or 6 factors were examined but yielded weaker, less interpretable factor structures.

Stage 2: Stress Factors EFA

Data Suitability and Factor Extraction

Analyses using 13 stress-related items with the same sample (N=306) demonstrated good data suitability (KMO=0.87; Bartlett test χ^2_{55}=2186.66; P<.001). Based on theoretical considerations and Phase 1 findings, we specified 3 stress factors corresponding to distinct stress domains, assuming that the new item related to societal, political, and economic stress would align with the Personal Stress factor. Principal axis factoring with Promax rotation converged in 6 iterations.

XSL•FO
RenderX

Factor Structure and Item Performance

The 3-factor solution explained 74.76% of total variance. Two items were removed due to substantial cross-loadings across stress domains: Perceived Control ("Felt that you have control over the important things in your life") and Smooth ("That things were going smoothly for you") both loaded significantly onto both Work Stress and Personal Stress factors, indicating these items captured general perceived control and life satisfaction rather than domain-specific stressors. The final 11 items demonstrated a clear factor structure with strong psychometric properties.

Notably, the exhaustion item loaded onto the Work Stress rather than the Burnout factor, which is consistent with theoretical models positioning exhaustion as an early indicator of work-related stress that may precede full burnout syndrome.

Integrated Factor Structure Summary

The staged EFA approach successfully identified a robust 10-factor structure encompassing 43 items: 7 resilience factors (32 items) and 3 stress factors (11 items). This comprehensive framework captures both current stress exposure and modifiable factors that influence resilience capacity, providing a theoretically grounded and empirically supported foundation for comprehensive stress resilience assessment.

The excellent psychometric properties, substantial variance explained (66.04% for resilience factors and 74.76% for stress factors), and strong reliability estimates support the validity of this multidimensional approach to measuring stress resilience capacity in applied settings.

CFA Results

Current Stress Model

We conducted a CFA to evaluate the structure of the Current Stress domain, which specified a higher-order latent factor (Stress) comprised of 3 latent factors: Work Stress, Personal Stress, and Burnout. The initial model demonstrated adequate fit to the data (χ^2_{41}=204.52; P<.001; CFI=0.92; TLI=0.90; RMSEA=0.11; and SRMR=0.07). We used MIs ≥10 to identify potential covariances between item residuals and added the following theoretically justified covariances to improve model fit without altering the core factor structure (refer to Multimedia Appendix 1).

These modifications reflected close semantic or contextual overlap among items and improved the overall model fit. Factor-to-error covariances were preferred over cross-loadings to maintain clear factor interpretation while acknowledging systematic relationships between theoretically related stress constructs.

The revised model demonstrated excellent fit to the data (χ^2_{33}=64.18; P=.001; CFI=0.99; TLI=0.98; RMSEA=0.06; and SRMR=0.05). All standardized factor loadings were statistically significant and strong, ranging from 0.54 to 0.92 across the 11 items. Most loadings exceeded 0.75, supporting the construct validity of the latent factors. Squared multiple correlations (R^2) showed that most items explained a substantial proportion of variance, ranging from 0.29 to 0.84. Full standardized factor loadings, SEs, and communalities are provided in Table 1. The WONE Index is a proprietary measure copyrighted by Walking on Earth, Inc. Full administration materials are available upon reasonable request.

XSL•FO
RenderX

Table . Standardized factor loadings and communalities for Phase 2 confirmatory factor analysis (CFA) models.

Construct	Factor	Domain	Standardized loading[a]	SE	Communality (R^2)
Stress (personal)	Personal Stress	Stress	0.92	—[b]	0.84
Anxiousness (personal)	Personal Stress	Stress	0.87	0.05	0.75
Overwhelm (personal)	Personal Stress	Stress	0.86	0.05	0.74
Societal, political, and economic stress	Personal Stress	Stress	0.54	0.06	0.29
Stress (work)	Work Stress	Stress	0.90	—	0.82
Anxiousness (work)	Work Stress	Stress	0.76	0.06	0.58
Overwhelm (work)	Work Stress	Stress	0.83	0.06	0.69
Mental exhaustion	Work Stress	Stress	0.79	0.07	0.63
Disengagement	Burnout	Stress	0.86	—	0.74
Cynicism	Burnout	Stress	0.75	0.08	0.56
Lack of productivity	Burnout	Stress	0.75	0.08	0.56
Vigorous physical activity	Physical Activity	Resilience Resources	0.99	0.11	1.00
Moderate physical activity	Physical Activity	Resilience Resources	0.63	—	0.40
Sleep quality	Sleep	Resilience Resources	0.85	—	0.72
Fatigue	Sleep	Resilience Resources	0.74	0.07	0.55
Sleep duration	Sleep	Resilience Resources	0.47	0.10	0.22
Sleep latency	Sleep	Resilience Resources	0.62	0.08	0.38
Sleep disturbances	Sleep	Resilience Resources	0.64	0.08	0.41
Nutritious food intake	Dietary Intake	Resilience Resources	0.65	—	0.42
Processed food intake	Dietary Intake	Resilience Resources	0.50	0.13	0.25
Caffeine intake	Dietary Intake	Resilience Resources	0.32	0.13	0.10
Caffeine reliance	Dietary Intake	Resilience Resources	0.52	0.19	0.27
Emotion regulation	Emotion Regulation and Coping	Resilience Resources	0.74	0.06	0.68
Emotional understanding	Emotion Regulation and Coping	Resilience Resources	0.61	0.06	0.37
Distress tolerance	Emotion Regulation and Coping	Resilience Resources	0.70	0.06	0.49
Acceptance	Emotion Regulation and Coping	Resilience Resources	0.50	0.06	0.25
Ability to bounce back	Emotion Regulation and Coping	Resilience Resources	0.76	0.06	0.58
Adaptability	Emotion Regulation and Coping	Resilience Resources	0.73	0.05	0.53
Effective coping	Emotion Regulation and Coping	Resilience Resources	0.83	—	0.69
Self-efficacy	Emotion Regulation and Coping	Resilience Resources	0.68	0.06	0.47
Cognitive flexibility	Emotion Regulation and Coping	Resilience Resources	0.66	0.06	0.43
Dwelling	Perseverative Thinking	Resilience Resources	0.91	—	0.82
Worrying	Perseverative Thinking	Resilience Resources	0.88	0.06	0.78

XSL•FO
RenderX

Construct	Factor	Domain	Standardized loading[a]	SE	Communality (R^2)
Meaning and purpose	Purpose and Prosociality	Resilience Resources	0.54	0.10	0.44
Gratitude	Purpose and Prosociality	Resilience Resources	0.78	0.09	0.61
Compassion for others	Purpose and Prosociality	Resilience Resources	0.63	0.09	0.40
Consideration for others	Purpose and Prosociality	Resilience Resources	0.67	0.07	0.45
Making a positive difference	Purpose and Prosociality	Resilience Resources	0.79	—	0.63
Trusted support system	Social Connection	Resilience Resources	0.89	0.05	0.79
Support satisfaction	Social Connection	Resilience Resources	0.87	0.06	0.76
Strength from close others	Social Connection	Resilience Resources	0.87	—	0.76
Belongingness	Social Connection	Resilience Resources	0.85	0.05	0.72
Loneliness	Social Connection	Resilience Resources	0.83	0.06	0.69

[a]All factor loadings are statistically significant at $P<.001$.

[b]SE not reported when factor loading was fixed to 1 for model identification.

Second-Order Factor Structure

The higher-order Stress factor was well supported by strong loadings from all 3 first-order factors. Work Stress showed the strongest loading (λ=0.80 and SE=0.12), followed by Personal Stress (λ=0.76) and Burnout (λ=0.67 and SE=0.13). The second-order loadings ranged from 0.67 to 0.80, with an average of 0.74, indicating that all 3 stress domains contribute substantially and relatively equally to the overarching stress construct. The squared multiple correlations for the first-order factors (0.44-0.65) demonstrate that the higher-order factor explains a substantial proportion of variance in each stress domain.

Resilience Resources Model

We then conducted a CFA to assess the structure of the Resilience Resources model, which specified a higher-order latent factor (Resources) comprised of 7 first-order latent domains: Emotion Regulation and Coping, Social Connectedness, Compassion and Gratitude, Sleep, Physical Activity, Dietary Intake, and Perseverative Thinking. The initial model demonstrated adequate fit to the data (χ^2_{458}=1081.71; $P<.001$; CFI=0.88; TLI=0.87; RMSEA=0.07; and SRMR=0.09). After reviewing MIs≥10, we added theoretically grounded modifications (refer to Multimedia Appendix 1 for specifics).

These modifications reflect meaningful psychological relationships among resilience constructs and improved model fit substantially while preserving the primary factor structure. This approach was chosen over alternative specifications (eg, cross-loadings) to maintain interpretability while acknowledging the theoretically expected interconnections among resilience domains.

The revised model fit the data well (χ^2_{443}=745.20; $P<.001$; CFI=0.94; TLI=0.94; RMSEA=0.05; and SRMR=0.06). All standardized factor loadings were statistically significant, ranging from 0.31 to 0.99. Most loadings exceeded 0.60, indicating strong relationships between latent constructs and their corresponding indicators. Squared multiple correlations (R^2) demonstrated good explanatory power for most observed variables, supporting the reliability and coherence of the factor structure. The confirmed 10-factor structure represents the final, validated framework of the WONE Index. Each factor captures specific theoretical domains as described in the “Second-Order Factor Structure" section. Item-level psychometric details are provided in Table 1, and theoretical descriptions of the factors are included in the “Discussion” section for Phase 2.

Second-Order Factor Structure

The higher-order Resilience Resources factor showed a more varied pattern of loadings from the 7 first-order factors, ranging from 0.31 to 0.88 (average=0.59). Emotion Regulation and Coping emerged as the dominant contributor (λ=0.88), followed by Perseverative Thinking (λ=0.73) and Sleep (λ=0.78). Social Connection showed a moderate loading (λ=0.55), while Purpose and Prosociality (λ=0.37) and Physical Activity (λ=0.31) contributed more modestly to the overarching construct. This hierarchical pattern suggests that while multiple domains contribute to resilience resources, cognitive-emotional regulatory capacities serve as the primary organizing feature, with behavioral and social factors providing important but secondary contributions.

Although each domain was analyzed separately for model estimation and reporting, together these 2 validated higher-order structures conceptually represent a correlated higher-order framework consistent with JD-R theory. This framework conceptualizes Stress Load and Resilience Resources as distinct but interrelated systems that cannot be reduced to a single overarching factor. The correlated-factor interpretation was retained over a bifactor alternative because it better reflects the theoretical position that resilience emerges from dynamic

XSL•FO
RenderX

interactions among multiple interdependent subsystems rather than from a single general dimension.

The 2 higher-order factors were also strongly correlated at the scale level, which indicates that higher stress load was statistically associated with lower resilience resources, although this relationship was not modeled within the same CFA structure. The squared multiple correlations for the first-order factors ranged from 0.10 to 0.78, indicating substantial variation in how well the higher-order factor explains variance across different resilience domains.

Validity and Reliability

All correlations between the WONE Index and external measures are provided in Table 2.

XSL•FO
RenderX

Table . Phase 2 correlations between the Walking on Earth (WONE) Index and established measures.

Variable	Full WONE Index	WONE stress subscale	WONE resilience subscale	PSS[a]	CD-RISC[b]	BRS[c]	PHQ-8[d]	PROMIS-SF-8a[e]	GAD-7[f]	WHO-5[g]	BFI-E[h]	BFI-A[i]	BFI-C[j]	BFI-N[k]	BFI-O[l]
Full WONE Index															
r	1	–0.84	0.98	–0.81	0.75	0.74	–0.77	–0.79	–0.77	0.83	0.31	0.46	0.45	–0.74	–0.03
P value	—[m]	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.56
WONE stress subscale															
r	–0.84a	1	–0.71	0.73	–0.52	–0.57	0.69	0.71	0.75	–0.68	–0.21	–0.31	–0.39	0.66	0.03
P value	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.59
WONE resilience subscale															
r	0.98a	–0.71	1	–0.78	0.77	0.75	–0.73	–0.76	–0.71	0.82	0.33	0.47	0.44	–0.71	–0.03
P value	<.001	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.58
PSS															
r	–0.81	0.73	–0.78	1	–0.66	–0.68	0.72	0.80	0.76	–0.74	–0.26	–0.36	–0.41	0.61	0.02
P value	<.001	<.001	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.74
CD-RISC															
r	0.75	–0.52	0.77	–0.66	1	0.83	–0.52	–0.57	–0.53	0.66	0.33	0.33	0.42	–0.65	–0.06
P value	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.27
BRS															
r	0.74	–0.57	0.75	–0.68	0.83	1	–0.53	–0.57	–0.58	0.62	0.27	0.40	0.39	–0.68	–0.05
P value	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.44
PHQ-8															
r	–0.77	0.69	–0.73	0.72	–0.52	–0.53	1	0.85	0.81	–0.75	–0.22	–0.34	–0.41	0.55	0.05
P value	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.36
PROMIS-SF-8a															
r	–0.79	0.71	–0.76	0.80	–0.57	–0.57	0.85	1	0.78	–0.76	–0.29	–0.40	–0.41	0.55	0.04
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001	.45
GAD-7															
r	–0.77	0.75	–0.71	0.76	–0.53	–0.58	0.81	0.78	1	–0.67	–0.19	–0.30	–0.33	0.67	0.06
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001	<.001	<.001	.27
WHO-5															
r	0.83	–0.68	0.82	–0.74	0.66	0.62	–0.76	–0.75	–0.67	1	0.32	0.42	0.43	–0.61	–0.04
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001	<.001	.53
BFI-E															
r	0.31	–0.21	0.33	–0.26	0.33	0.27	–0.29	–0.22	–0.19	0.32	1	0.30	0.25	–0.33	0.02

XSL•FO
RenderX

Variable		Full WONE Index	WONE stress subscale	WONE resilience subscale	PSS[a]	CD-RISC[b]	BRS[c]	PHQ-8[d]	PROMIS-SF-8a[e]	GAD-7[f]	WHO-5[g]	BFI-E[h]	BFI-A[i]	BFI-C[j]	BFI-N[k]	BFI-O[l]
	P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001	.75
BFI-A																
	r	0.46a	–0.31	0.47	–0.34	0.33	0.40	–0.40	–0.35	–0.30	0.42	0.30	1	0.34	–0.33	–0.05
	P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	.34
BFI-C																
	r	0.45a	–0.39	0.44	–0.41	0.42	0.34	–0.41	–0.41	–0.33	0.43	0.25	0.34	1	–0.38	0.02
	P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	.76
BFI-N																
	r	–0.74	0.66	–0.71	0.61	–0.65	–0.68	0.55	0.55	0.67	–0.61	–0.33	–0.33	–0.38	1	0.02
	P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	.67
BFI-O																
	r	–0.03	0.03	–0.03	0.02	–0.06	–0.05	0.04	0.05	0.06	–0.04	0.02	–0.05	0.02	0.02	1
	P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—

[a]PSS: Perceived Stress Scale.
[b]CD-RISC: Connor-Davidson Resilience Scale.
[c]BRS: Brief Resilience Scale.
[d]PHQ-8: Patient Health Questionnaire-8.
[e]PROMIS-SF-8a: Patient-Reported Outcomes Measurement Information System Short Form 8a.
[f]GAD-7: Generalized Anxiety Disorder-7.
[g]WHO-5: World Health Organization-5 Well-Being Index.
[h]BFI-E: Big Five Inventory-Extraversion
[i]BFI-A: Big Five Inventory-Agreeableness
[j]BFI-C: Big Five Inventory-Conscientiousness
[k]BFI-N: Big Five Inventory-Neuroticism
[l]BFI-O: Big Five Inventory-Openness
[m]Not applicable.

Construct-Level Convergent Validity

The WONE Index demonstrated strong construct-level convergent validity with established measures of stress and resilience. The Stress Subscale showed a large positive correlation with the PSS, indicating that higher scores on our stress measure align closely with higher perceived stress as measured by this gold-standard instrument.

For resilience measures, the Resilience Subscale demonstrated excellent convergent validity with both the CD-RISC-10 and the BRS. Similarly, the Full WONE Index showed strong positive correlations with both resilience measures. These large correlations (all exceeding Cohen benchmark for large effects [65]) with established resilience measures are consistent with the WONE Index measuring conceptually related but distinct aspects of stress resilience, as would be expected given its emphasis on specific behavioral resources and demands-resources integration versus the trait-based conceptualizations of the CD-RISC and BRS. These patterns support construct validity while indicating that the WONE Index captures additional variance beyond established measures.

Concurrent Validity

The measure demonstrated excellent concurrent validity with theoretically related mental health and well-being outcomes, with all correlations in the expected directions and magnitudes.

Stress-Mental Health Relationships

The Stress Subscale showed strong positive correlations with both depressive symptom measures (PHQ-8 and PROMIS-SF-8a) and anxiety symptoms (GAD-7). Conversely, the Stress Subscale demonstrated a strong negative correlation with well-being (WHO-5), indicating that higher stress is associated with lower well-being, as expected.

XSL•FO
RenderX

Resilience-Mental Health Relationships

The Resilience Subscale showed strong protective relationships with mental health outcomes. Higher resilience scores were associated with lower depressive symptoms on both measures, lower anxiety symptoms, and higher well-being.

Full WONE Index Concurrent Validity

The Full WONE Index demonstrated excellent concurrent validity across all mental health outcomes, with large negative correlations with depressive and anxiety symptoms and a large positive correlation with well-being. These strong relationships across all mental health domains support the Full Index as a comprehensive measure of stress resilience capacity.

Comparative Performance

Notably, the WONE Index demonstrated stronger correlations with mental health outcomes than established resilience measures and comparable correlations with the PSS for stress-related outcomes (refer to Table 2).

Discriminant Validity

The WONE Index showed appropriate discriminant validity when examined against personality traits, with one theoretically meaningful exception that actually supports construct validity.

Neuroticism (Supporting Evidence for Construct Validity)

The Full WONE Index demonstrated a strong negative correlation with neuroticism. Rather than representing a discriminant validity concern, this relationship provides important construct validity evidence. Neuroticism, characterized by heightened stress reactivity and emotional instability, would be expected to show a strong negative relationship with stress resilience capacity. This correlation suggests our measure successfully captures individual differences in stress vulnerability and resilience resources, making it particularly valuable for identifying individuals who may benefit most from resilience-building interventions.

Other Personality Traits

The Index showed moderate correlations with agreeableness and conscientiousness, suggesting some overlap with these adaptive personality traits, though correlations remained below the threshold indicating construct redundancy. The measure showed a smaller correlation with extraversion and was essentially uncorrelated with openness, demonstrating good discrimination from these personality dimensions.

Discriminant Validity Between Domains

HTMT ratios were calculated to assess discriminant validity between measurement domains. All HTMT values fell below the conservative threshold of 0.85, with most below 0.70, supporting discriminant validity between domains while confirming their theoretical relationships.

Stress Domain Relationships

The 3 stress domains showed moderate to strong relationships, confirming that they capture related but distinct stress experiences (refer to Table S7 in Multimedia Appendix 1).

Resilience Domain Relationships

Resilience domains demonstrated appropriate discriminant validity. The strongest relationships were observed between conceptually related domains (Emotion Regulation and Coping ↔ Sleep; Purpose and Prosociality ↔ Social Connection), supporting the theoretical structure while confirming domain distinctiveness (refer to Table S8 in Multimedia Appendix 1).

Internal Consistency

The CR estimate for the overall scale was 0.84. Estimates for individual factors ranged from 0.62 to 0.96, with 9 of 10 factors exceeding the 0.70 threshold for acceptable reliability. Eight factors achieved good to excellent reliability (CR ≥0.80; refer to Table 3).

XSL•FO
RenderX

Table . Factor-level test-retest reliability and internal consistency.

Subscale and factor	ICC[a] (95% CI)	CR[b]	AVE[b]
WONE[c] Index	0.90 (0.87-0.93)	—	—
Stress subscale	0.84 (0.80-0.88)	—	—
Personal Stress factor	0.90 (0.87-0.92)	0.90	0.64
Work Stress factor	0.89 (0.85-0.91)	0.91	0.67
Burnout factor	0.87 (0.82-0.90)	0.84	0.58
Resilience subscale	0.90 (0.87-0.92)	—	—
Emotion Regulation and Coping factor	0.95 (0.93-0.96)	0.89	0.49
Social Connectedness factor	0.93 (0.91-0.95)	0.94	0.74
Purpose and Prosociality factor	0.93 (0.91-0.95)	0.83	0.49
Sleep factor	0.90 (0.87-0.93)	0.83	0.51
Dietary Intake factor	0.90 (0.86-0.92)	0.62	0.31
Physical Activity factor	0.81 (0.74-0.85)	0.81	0.70
Perseverative Thinking factor	0.79 (0.72-0.84)	0.89	0.80

[a]ICC: intraclass correlation coefficient.

[b]Composite reliability (CR) and average variance extracted (AVE) estimates are not applicable to the full WONE Index, as it is calculated as a composite of 2 separate confirmatory factor models, rather than being estimated as a latent factor within a single measurement model.

[c]WONE: Walking on Earth

The Dietary Intake factor showed marginal reliability (CR=0.62 and AVE=0.31), likely reflecting the conceptual diversity of dietary behaviors assessed (caffeine reliance, processed food consumption, and nutritious diet quality). The marginal reliability of the Dietary Intake factor likely reflects the conceptual diversity of dietary behaviors rather than measurement inadequacy. Despite the lower statistical indicators, these items were retained due to their theoretical importance in the resilience framework and the nascent state of dietary behavior measurement in resilience research.

Test-Retest Reliability

The WONE Index demonstrated excellent temporal stability across all major scales over 3 weeks (Table 3). The Full WONE Index showed excellent test-retest reliability, indicating highly consistent measurement of overall stress resilience capacity over time. Both subscales also demonstrated strong temporal stability, with the Resilience Subscale exhibiting excellent reliability and the Stress Subscale showing good reliability.

These findings indicate that the WONE Index provides a stable and consistent measurement of both stress and resilience resources across time. The excellent temporal stability of the WONE Index supports its use as a reliable assessment tool for tracking stress resilience capacity. The high test-retest reliability coefficients indicate that observed changes in WONE scores over time reflect true changes in an individual's stress resilience capacity rather than measurement error, making the instrument suitable for monitoring intervention effects and tracking progress.

Individual WONE factors also showed strong to excellent test-retest reliability across the 3-week interval (Table 3). Nine of 10 factors achieved excellent temporal stability (ICC >0.85).

Incremental Validity Results

The WONE Index provided significant incremental prediction over a combination of existing gold-standard measures. All hierarchical regression models demonstrated statistically significant incremental validity, indicating that the WONE Index provides meaningful prediction beyond established measures across all mental health outcomes tested. The WONE Index's superior correlational performance relative to established resilience measures and comparable performance to the PSS provides the foundation for these incremental validity findings.

Model 1: Beyond Perceived Stress Scale

The WONE Full Index provided significant incremental prediction beyond the PSS across all outcomes. Even after accounting for perceived stress, the WONE explained additional variance in depressive symptoms (PHQ-8: ΔR^2=0.10, P<.001; PROMIS-SF-8a: ΔR^2=0.06, P<.001; GAD-7: ΔR^2=0.07, P<.001; and WHO-5: ΔR^2=0.16, P<.001).

Model 2: Beyond Established Resilience Measures

Model 2a (Beyond CD-RISC)

The WONE Index demonstrated substantial incremental validity beyond the CD-RISC across all outcomes, with particularly strong incremental prediction for depressive symptoms and well-being (PHQ-8: ΔR^2=0.32, P<.001; PROMIS-SF-8a: ΔR^2=0.31, P<.001; GAD-7: ΔR^2=0.31, P<.001; and WHO-5: ΔR^2=0.26, P<.001).

Model 2b (Beyond BRS)

Similar patterns were observed when testing incremental validity beyond the BRS, with significant ΔR^2 values across all mental health outcomes (PHQ-8: ΔR^2=0.31, P<.001; PROMIS-SF-8a:

XSL•FO
RenderX

ΔR^2=0.31, P<.001; GAD-7: ΔR^2=0.25, P<.001; and WHO-5: ΔR^2=0.31, P<.001).

Model 3: Beyond Combined Established Measures

Most importantly, the WONE Index provided significant incremental prediction even beyond the combination of both established stress and resilience measures (PSS, CD-RISC, and BRS). This represents the most stringent test of incremental validity, as it demonstrates that WONE adds meaningful prediction beyond current best practices that would use both stress and resilience assessments (PHQ-8: ΔR^2=0.11, P<.001; PROMIS-SF-8a: ΔR^2=0.07, P<.001; GAD-7: ΔR^2=0.07, P<.001; and WHO-5: ΔR^2=0.11, P<.001).

Measurement Invariance for Stress

Measurement Invariance Testing Across Gender

Multigroup CFA indicated good configural fit across women (n=164) and men (n=142), suggesting a consistent factor structure (χ^2_{66}=115.71; P<.001; CFI=0.98; TLI=0.96; RMSEA=0.05; and SRMR=0.06). Constraining factor loadings (metric model) maintained a strong fit (χ^2_{76}=125.58; P<.001; CFI=0.98; TLI=0.97; RMSEA=0.05; and SRMR=0.06), supporting metric invariance. Adding intercept constraints (scalar model) yielded an acceptable fit (χ^2_{86}=151.50; P<.001; CFI=0.97; TLI=0.96; RMSEA=0.05; and SRMR=0.07), supporting latent mean comparisons. The strict model (residual variances constrained) also fit well (χ^2_{99}=163.14; P<.001; CFI=0.97; TLI=0.97; RMSEA=0.05; and SRMR=0.07), supporting full invariance across gender.

Measurement Invariance Across Race

Participants were grouped as White (n=211) and non-White (n=95). Configural fit was strong (χ^2_{66}=100.62; P=.01; CFI=0.99; TLI=0.98; RMSEA=0.04; and SRMR=0.05), indicating consistent structure. The metric model (loadings constrained) also showed excellent fit (χ^2_{76}=100.62; P=.03; CFI=0.99; TLI=0.98; RMSEA=0.03; and SRMR=0.05). Scalar constraints (intercepts) retained good fit (χ^2_{86}=111.65; P=.03; CFI=0.99; TLI=0.99; RMSEA=0.03; and SRMR=0.04), allowing for latent mean comparisons. The strict model (residuals constrained) showed acceptable fit (χ^2_{99}=129.24; P=.02; CFI=0.97; TLI=0.98; RMSEA=0.03; and SRMR=0.045), indicating full invariance across race.

Measurement Invariance Across Age Groups

Participants were split into ages 20 - 39 years (n=175) and 40 - 64 years (n=131). The configural model demonstrated good fit (χ^2_{66}=95.23; P=.01; CFI=0.99; TLI=0.98; RMSEA=0.04; and SRMR=0.06). Metric invariance was supported with excellent fit when loadings were constrained (χ^2_{76}=102.98; P=.02; CFI=0.99; TLI=0.98; RMSEA=0.03; and SRMR=0.06). The scalar model (adding intercept constraints) showed a strong fit (χ^2_{86}=127.18; P=.003; CFI=0.98; TLI=0.98; RMSEA=0.04; and SRMR=0.06), permitting latent mean comparisons. The strict model (residuals constrained) also fit well (χ^2_{99}=143.23; P<.002; CFI=0.98; TLI=0.98; RMSEA=0.04; and SRMR=0.06), indicating strict invariance across age groups.

Measurement Invariance for Resilience Resources

Measurement Invariance Across Gender

We evaluated measurement invariance for the Resilience Resources model across gender (women: n=164 and men: n=142) using multigroup CFA. The configural model demonstrated good fit (χ^2_{886}=1223.58; P<.001; CFI=0.94; TLI=0.93; RMSEA=0.04; and SRMR=0.08), indicating a consistent factor structure across groups. Constraining factor loadings in the metric model yielded similar fit (χ^2_{917}=1288.44; P<.001; CFI=0.93; TLI=0.92; RMSEA=0.04; and SRMR=0.08), supporting equivalence of loadings. The scalar model, which additionally constrained item intercepts, also showed good fit (χ^2_{949}=1360.84; P<.001; CFI=0.92; TLI=0.92; RMSEA=0.04; and SRMR=0.09). The strict model, which further constrained residual variances, continued to meet fit thresholds (χ^2_{988}=1413.10; P<.001; CFI=0.92; TLI=0.92; RMSEA=0.04; and SRMR=0.09), supporting full measurement invariance across gender.

Measurement Invariance Across Race

We next assessed invariance across race, comparing White (n=211) and non-White (n=95) participants. The configural model demonstrated good fit (χ^2_{886}=1326.95; P=.01; CFI=0.92; TLI=0.91; RMSEA=0.04; and SRMR=0.07), indicating a consistent factor structure. The metric model, which constrained factor loadings, showed similar fit (χ^2_{917}=1363.49; P<.001; CFI=0.92; TLI=0.91; RMSEA=0.04; SRMR=0.07). The scalar model, which additionally constrained item intercepts, maintained good fit (χ^2_{949}=1433.49; P<.001; CFI=0.91; TLI=0.91; RMSEA=0.04; and SRMR=0.07). The strict model, which further constrained residual variances, also met fit thresholds (χ^2_{988}=1540.42; P<.001; CFI=0.90; TLI=0.90; RMSEA=0.04; and SRMR=0.07), supporting full measurement invariance across race.

Measurement Invariance Across Age Groups

Participants were grouped by age 20 - 39 years (n=175) and 40 - 64 years (n=131). The configural model demonstrated acceptable fit (χ^2_{886}=1352.55; P<.001; CFI=0.92; TLI=0.91; RMSEA=0.04; and SRMR=0.07). Metric invariance was supported with excellent fit when loadings were constrained (χ^2_{917}=1385.98; P<.001; CFI=0.92; TLI=0.91; RMSEA=0.04; and SRMR=0.08). The scalar model (adding intercept constraints) showed strong fit (χ^2_{949}=1458.55; P<.001; CFI=0.91; TLI=0.90; RMSEA=0.04; and SRMR=0.08), permitting latent mean comparisons. The strict model (residuals constrained) also fit well (χ^2_{988}=1542.30; P<.001; CFI=0.90; TLI=0.90; RMSEA=0.04; and SRMR=0.08), indicating strict invariance across age groups.

Weighting

Empirical Weight Distribution

The regression-derived empirical weights revealed a clear hierarchical structure dominated by psychological and stress-related constructs (Table 4). Emotion Regulation and Coping emerged as the most critical predictor (44.6%), followed

XSL•FO
RenderX

by Personal Stress (16.3%) and Perseverative Cognition (8.3%). Notably, Physical Activity received minimal empirical weighting (0.2%), likely reflecting mediation through other measured constructs rather than a lack of theoretical importance.

Table . Empirical, theoretical, and final hybrid weights.

Construct	Empirical weight, %	Theoretical weight, %	Hybrid weight, %	Final weight, %
Emotion Regulation and Coping	44.6	20.0	32.3	31
Personal Stress	16.3	10.0	13.2	13
Social Connectedness	7.8	18.0	12.9	12
Perseverative Thinking	8.3	14.0	11.2	11
Sleep	6.9	9.0	7.9	8
Dietary Intake	6.2	5.0	5.6	5
Burnout	5.3	5.0	5.2	5
Physical Activity	0.2	7.0	3.6	5
Work Stress	1.0	7.0	4.0	5
Purpose and Prosociality	3.9	5.0	4.5	5

Theoretical Weight Rationale

Theoretical weights addressed empirical limitations while ensuring comprehensive coverage of modifiable resilience factors. Social Connectedness received high theoretical weighting based on robust meta-analytic evidence linking social support to stress resilience outcomes [66,67]. Sleep and Physical Activity received substantial theoretical weights given their well-documented roles in stress buffering and physiological recovery mechanisms [68-70].

Work Stress received meaningful theoretical weighting despite minimal empirical contribution, reflecting complex nonlinear relationships observed in organizational research, where higher work demands correlate positively with resilience in certain populations [71], suggesting stress inoculation effects [72] or self-selection mechanisms.

Hybrid Weight Validation

Cross-sectional validation demonstrated superior performance of the hybrid weighting approach across all criterion measures (Table 5). The weighted approach showed consistent improvements over unweighted alternatives, with particularly strong gains for primary resilience measures: CD-RISC (+0.06 vs unweighted) and BRS (+0.08 vs unweighted).

Table . Validation performance comparison.

Outcome measure	Hybrid weighted	Unweighted	Item-level	Best performing version
CD-RISC[a]	0.75[b]	0.68[b]	0.73[b]	Weighted
BRS[c]	0.74[b]	0.67[b]	0.71[b]	Weighted
PSS-4[d]	−0.81[b]	−0.77[b]	−0.79[b]	Weighted
PHQ-8[e]	−0.77[b]	−0.76[b]	−0.77[b]	Item-level
PROMIS-SF-8a[f]	−0.79[b]	−0.77[b]	−0.79[b]	Weighted
GAD-7[g]	−0.77[b]	−0.74[b]	−0.73[b]	Weighted
WHO-5[h]	0.83[b]	0.81[b]	0.84[b]	Item-level

[a]CD-RISC: Connor-Davidson Resilience Scale.

[b]$P<.001$

[c]BRS: Brief Resilience Scale.

[d]PSS-4: Perceived Stress Scale-4.

[e]PHQ-8: Patient Health Questionnaire-8.

[f]PROMIS-SF-8a: Patient-Reported Outcomes Measurement Information System Short Form 8a.

[g]GAD-7: Generalized Anxiety Disorder-7.

[h]WHO-5: World Health Organization-5 Well-Being Index.

XSL•FO
RenderX

Validation Performance Summary

The hybrid approach demonstrated meaningful improvements across resilience-related outcomes while maintaining strong correlations with mental health indicators. The greatest improvements occurred for measures most directly assessing resilience capacity (CD-RISC and BRS), suggesting the theoretical components successfully enhanced the measurement of core resilience constructs.

Notably, the hybrid approach outperformed both purely unweighted and item-level averaging approaches, indicating that the systematic integration of empirical prediction with theoretical importance creates optimal measurement properties. The modest improvement for depression (PHQ-8) likely reflects the already strong empirical weighting of constructs most relevant to depressive symptoms.

Final Weight Structure and Clinical Implications

The final hybrid weights create a theoretically coherent hierarchy that preserves predictive validity while ensuring comprehensive intervention guidance. Emotion Regulation and Coping maintains prominence, consistent with its central role in stress resilience frameworks. Core stress and cognitive constructs receive substantial representation: Personal Stress, Social Connectedness, and Perseverative Thinking.

Critically, all behavioral and contextual factors achieve meaningful weighting for intervention purposes: Sleep, Physical Activity, Work Stress, Dietary Intake, Burnout, and Compassion/Gratitude/Meaning. This distribution supports the WONE Index's dual function as both a predictive assessment and behavioral intervention guidance system.

Longitudinal validation using Time 1 WONE scores to predict Time 2 outcomes (n=203) further confirmed the superior predictive performance of the weighted approach across all validation measures, though detailed longitudinal analyses are beyond the scope of the current phase.

Discussion

Principal Findings

The WONE Index was developed and validated through a 2-phase process to progressively refine its measurement model. Phase 1 provided strong initial psychometric evidence, establishing a 6-factor structure organized into 2 higher-order domains: Stress Load (Work Stress, Personal Stress, and Burnout) and Resilience Resources (Resilience Skills and Beliefs, Social Support, and Sleep). While statistically robust, this model did not fully capture the multidimensional and dynamic nature of resilience, particularly with respect to health behaviors and cognitive processes.

Phase 2 expanded and refined this framework into a 10-factor structure organized within 2 higher-order domains, which we propose as the finalized structure. Seven resource-related factors (Emotion Regulation and Coping, Social Connection, Purpose and Prosociality, Sleep, Physical Activity, Dietary Intake, and Perseverative Thinking) and 3 stress factors (Work Stress, Personal Stress, and Burnout) were identified and grouped into the higher-order domains of Stress Load and Resilience Resources. In both phases, we followed standard CFA practice by incorporating theoretically justified modifications suggested by MIs, including correlated residuals within conceptually related item sets. These refinements improved model fit while preserving the theoretical structure. Model fit indices for both Phase 1 and Phase 2 were strong, but the Phase 2 model offered comparable psychometric performance while adding domains that improved interpretability and intervention relevance, making it the preferred structure. By incorporating additional factors, the final model allows for greater precision in identifying areas of vulnerability and creates more actionable pathways for intervention.

Importantly, the Index allows us to move beyond identifying whether someone is "resilient" to understanding how their system is functioning. It captures systemic imbalances in how individuals perceive, feel, adapt, and respond to stress, thereby holistically measuring resilience and highlighting where interventions should target. For example, someone experiencing high stress load but with depleted resources may benefit most from micro-moment stress-reduction strategies that support daily functioning before resource-building can take hold. Conversely, individuals with low external stressors but insufficient resilience resources may require preventive interventions to expand their repertoire of coping and regulatory skills.

This perspective also acknowledges paradoxical cases such as "skin-deep resilience," in which individuals may appear psychologically resilient to chronic stress, yet their bodies still carry physiological costs (eg, accelerated aging, immune dysregulation, and cardiovascular risk). In such cases, building resilience resources remains essential, as resource strengthening may buffer against hidden biological wear-and-tear even when stress load is not consciously perceived as high [73,74]. By integrating both stress load and resilience resources, the WONE Index provides a nuanced lens for identifying vulnerability, tailoring interventions, and supporting resilience development across populations and contexts.

The WONE Index also demonstrated excellent temporal stability, strong reliability across scales, and robust validity with multiple established measures of resilience, stress, and mental health outcomes. Together, this evidence supports the Index as both a rigorous scientific tool and a practical framework for understanding resilience in diverse populations.

The hybrid weighting methodology is a key innovation of this work. It integrates empirical weights derived from predictive modeling of mental health and well-being outcomes with theoretical weights grounded in resilience science and dynamical systems constructs (eg, Sleep, Social Connection, and Emotion Regulation). This dual approach produces scores that are both predictively robust and have practical utility in digital health and applied psychology.

Conceptually, this framework aligns with multidimensional outcome systems such as the Treatment Outcome Package [75], which emphasizes comprehensive assessment across multiple domains to guide individualized feedback and treatment. The WONE Index adopts a similar logic by balancing empirical evidence with theoretical modifiability, ensuring that domain

XSL•FO
RenderX

scores capture both predictive value and potential for change through intervention. This approach enhances the Index's relevance for digital health applications by identifying high-impact targets while also laying the foundation for future integration of predictive modeling approaches to further enhance precision and personalization. In doing so, it bridges traditional clinical feedback models with data-driven precision frameworks, offering a scalable approach to personalized resilience assessment.

Together, these findings establish the WONE Index as a comprehensive, multidimensional, and psychometrically rigorous measure of resilience, conceptualized as adaptive capacity emerging from the balance between stress load and resilience resources, and designed to both measure resilience and inform targeted intervention.

Comparison With Previous Work

O'Donohue et al [76] identified 45 distinct measurement approaches with substantial heterogeneity in stress resilience conceptualization and operationalization. Critical limitations documented across existing measures include: (1) most measures assess stress or resilience separately rather than their dynamic interaction, with only 17.5% using resilience measures and many relying solely on stress indices; (2) predominant use of trait-based rather than process-based conceptualizations (eg, CD-RISC, BRS, and RSA which focus on stable characteristics), limiting sensitivity to temporal change; and (3) limited validation for repeated measurement contexts, with most instruments designed for single-timepoint assessment [8]. In addition, widely used resilience measures—including the CD-RISC, BRS, and RSA—focus predominantly on perceptual attitudes ("I can deal with whatever comes") rather than behaviorally specific, modifiable resources, and do not integrate health behavior determinants—including sleep, physical activity, and nutrition—documented as protective factors that buffer stress and promote resilience [20,21,77-80].

The WONE Index addresses these gaps through 3 key distinctions. First, it integrates stress load and resilience resources within a unified framework, operationalizing the demands-resources balance emphasized by JD-R theory rather than measuring constructs separately [12]. Second, it provides comprehensive domain coverage spanning psychological resources (emotion regulation, coping, and perseverative cognition), social resources (connection and support), meaning-oriented resources (purpose, gratitude, and prosociality), and health behavioral resources (sleep, physical activity, and nutrition)—a combination rarely found in existing validated measures to our knowledge. Third, it emphasizes behaviorally specific, modifiable capacities providing concrete intervention targets for digital health applications. However, longitudinal validation is needed to establish sensitivity to intervention-related change, and comparative effectiveness research must determine whether this integrated approach provides advantages over using established measures in combination, while independent replication by researchers without commercial affiliations is essential for confirming generalizability across diverse populations and contexts.

The WONE Index contributes to a growing literature emphasizing that stress and resilience are not opposite poles of a single construct but distinct domains with unique predictive value. Similar to existing resilience frameworks [4,11], our results underscore the centrality of regulatory skills, social connection, emotion, and meaning-making, but the Index also extends previous tools in several important ways. Unlike widely used measures that treat demands and resources as isolated constructs (eg, PSS [18], CD-RISC [20], and BRS [21]), the Index embeds them in a unified structure reflecting how these factors interact as interdependent subsystems. This operationalization enables the measure to serve dual purposes: scientifically, it captures multidomain adaptive capacity aligned with JD-R and systems-based models; practically, it functions as an intervention guidance system that balances measurement rigor with intervention utility.

The 10-factor structure within 2 higher-order domains provides a multidimensional systems approach to resilience that begins to approximate a network of interdependent resilience domains by capturing distinct contributions from cognitive, behavioral, social, and emotional domains, each of which showed robust factor loadings and validity evidence in Phase 2. Cognitive (eg, perseverative thinking), social (social connection and purpose/prosociality), emotion (eg, emotion regulation and positive affect), health behavior (eg, sleep, physical activity, and diet), and stress exposures (eg, personal, work, and burnout) are all included, creating a systemic framework rather than a 1D scale, which has been popular in established measures of resilience.

This expanded domain coverage forms the foundation of its systems perspective. This aligns with allostatic models, JD-R, and complex dynamical systems models of psychopathology, which emphasize how individuals may become "stuck" in maladaptive states or transition toward adaptive states depending on system-level dynamics [81]. By conceptualizing resilience as adaptive capacity emerging from the balance between stress load and resilience resources, the WONE Index provides a practical operationalization of these theoretical models.

As mentioned in the previous paragraph, the WONE index builds directly on the JD-R framework [12,13], which posits that health, well-being, and performance are determined by the interplay between demands (eg, workload, role conflict, and emotional strain) and resources (eg, social support, coping strategies, and recovery experiences). Our higher-order domains map closely to this model: Stress Load reflects job and life demands, while Resilience Resources reflect the protective assets available to meet them. Importantly, the JD-R framework distinguishes acute strain driven by demands from the chronic exhaustion that develops into burnout. The WONE Index mirrors this theoretical distinction by separately modeling work stress and burnout, offering a psychometric tool for examining how short-term strain may evolve into longer-term depletion and maladaptive system states.

Beyond alignment, the Index extends JD-R in 2 important ways. First, it broadens the scope by incorporating personal, societal, and health behavior domains, increasing relevance across multiple life contexts rather than being limited to occupational

XSL•FO
RenderX

settings. Second, the hybrid weighting methodology adds precision to JD-R by quantifying the relative impact of different resources. Whereas JD-R theory broadly emphasizes the buffering role of resources, the WONE Index specifies which domains—such as sleep, coping, or social connection—are both theoretically central and empirically predictive of outcomes such as depression, anxiety, and well-being. In this way, the Index operationalizes the JD-R concept of resources while also extending it to a wider range of life demands and contexts.

Applications to Digital Mental Health

The WONE Index advances digital mental health measurement in several important ways.

First, it was developed and validated specifically within digital delivery contexts, ensuring its psychometric properties hold when administered via web and mobile interfaces rather than assuming transferability from traditional formats. The index was validated via a digital survey/platform, providing confidence that observed reliability and validity reflect real-world digital administration rather than performance under idealized controlled conditions.

Second, the Index enables a measurement-based care approach at scale. Traditional assessment typically occurs at discrete time points (eg, intake and discharge) with results informing clinician-delivered interventions. In contrast, the WONE Index supports continuous assessment cycles where monthly administration generates updated resilience profiles that automatically inform algorithmic personalization of intervention content. This creates feedback loops between assessment and intervention that are difficult to achieve in traditional service delivery models, allowing platforms to adaptively adjust recommendations as users' stress-resource balance shifts over time.

Third, the multidimensional structure provides actionable granularity for automated intervention matching. Rather than producing a single resilience score requiring clinician interpretation to determine intervention targets, the 10 factors within 2 overarching domains assessed in the WONE Index directly map onto intervention modules within the platform. Low scores on emotion regulation trigger recommendations for regulation-focused techniques, depleted social connection scores prompt social relationship-focused interventions, and poor sleep activates sleep hygiene interventions. This direct assessment-to-intervention mapping enables digital platforms to deliver truly personalized care pathways without requiring human clinical judgment for every user.

Fourth, the measure's design supports population-level insights alongside individual assessment. Aggregated anonymized data reveal organizational stress patterns, high-risk subgroups, and systemic factors affecting workforce resilience. This dual-level utility, individual personalization plus population surveillance, distinguishes the Index from traditional measures designed solely for individual clinical assessment. Organizations can identify when specific teams show declining resilience resources or elevated stress load, enabling proactive organizational interventions before individual crises emerge.

Finally, the hybrid weighting methodology reflects digital health's unique requirements. Pure empirical weighting maximizes prediction but may undervalue behaviorally modifiable domains showing weaker cross-sectional associations yet greater intervention responsiveness. Theoretical weighting ensures the Index prioritizes domains where digital platforms can deliver effective interventions (eg, sleep, physical activity, and social connection) even when these show modest concurrent prediction. This methodology acknowledges that digital health assessment serves intervention guidance, not just prediction, requiring weights that balance predictive validity with behavioral actionability.

Together, these features distinguish the WONE Index from existing stress and resilience measures, which typically assess either protective or risk factors in isolation, rely on static trait-based designs, and lack validation for digital health or workplace contexts. By integrating stress load and resilience resources within a unified, multidomain structure specifically designed for digital mental health contexts and linking assessment directly to intervention personalization, the WONE Index operationalizes the demands-resources framework for temporal tracking through repeated administration in applied digital settings.

Strengths

This study has several notable strengths. First, it used a multiphase design, beginning with exploratory development and culminating in confirmatory validation. This sequential approach provides both a rigorous psychometric foundation and evidence of generalizability. Second, the WONE Index advances the field by integrating multiple domains of resilience into a hierarchical structure, capturing stress load alongside resilience resources. This systemic framing reflects how resilience operates across cognitive, social, behavioral, and meaning-oriented processes rather than as a single dimension.

Third, the Index demonstrated robust psychometric qualities, including strong reliability, validity across multiple criteria, and temporal stability, supporting its use for both research and applied purposes. Fourth, the hybrid weighting methodology represents a novel contribution to measuring development, balancing empirical prediction with theoretical modifiability and enhancing the practical utility of final scores.

Limitations

Several limitations should also be noted. First, the data were cross-sectional, which precludes causal inference. While test-retest analyses support temporal stability, the design limits the ability to determine whether changes in stress load or resilience resources precede improvements in outcomes. Future longitudinal studies are needed to examine how resilience processes evolve over time and in response to intervention.

Second, the study relied on self-report measures, which introduces potential for bias, such as recall inaccuracy or social desirability effects. We sought to mitigate these issues by using validated instruments with well-established psychometric properties and by using a brief, single-session survey format to minimize fatigue-related error. Nevertheless, self-report bias could have influenced observed relationships among constructs.

XSL•FO
RenderX

Future studies should integrate behavioral and physiological indicators (eg, ecological momentary assessment, wearables, and biomarkers) to strengthen validity and reduce reliance on self-report alone.

Third, although the sample was diverse and measurement invariance was supported across gender, race, and age, the use of online convenience sampling (CloudResearch and Prolific) may limit generalizability to digitally-engaged populations. Participants were primarily from the United States and the United Kingdom and were predominantly White, which may constrain cultural generalizability. Accordingly, future research should aim to validate the WONE Index across more diverse cultural, linguistic, and global populations to ensure measurement equivalence and capture potential cultural differences in how stress load and resilience resources manifest.

Fourth, brief item pools for some domains such as diet and physical activity, while necessary for reducing participant burden, may have limited construct breadth. The Dietary Intake factor, in particular, showed lower reliability, representing a known measurement limitation. We retained this domain due to its theoretical relevance and behavioral actionability in digital health contexts.

In addition, these health behavior domains may capture variance from aspects of resilience not fully represented in other adaptive resource domains (eg, energy balance or embodied forms of adaptation). It is also possible that shared variance among closely related constructs may have partially attenuated their distinct contributions. Future research should be mindful that diet and physical activity may relate more strongly to multimodal and objective data sources (eg, wearables and ecological monitoring) given their greater overlap with physiological processes compared to psychosocial resources.

Further, while the hybrid weighting system enhances both predictive power and practical relevance, the process of assigning theoretical weights remains partly subjective. We sought to balance theoretical rationale with empirical model fit, but weighting decisions could still introduce bias in the relative importance of certain domains. Future work should refine these weightings through stakeholder engagement and cross-validation in applied contexts.

Finally, the WONE Index was developed within a digital health implementation context, providing access to large-scale user samples and enabling iterative refinement based on real-world engagement. While this design addresses documented measurement gaps through integrated assessment of stress and resilience resources, establishing whether this approach provides practical advantages over existing measures requires additional validation. Future research should examine predictive validity for clinical outcomes and intervention response, longitudinal sensitivity to change in intervention contexts, and generalizability across demographic and cultural groups.

Comparative effectiveness research will be valuable to determine whether integrated assessment offers advantages over domain-specific measures used in combination, or whether different measurement approaches serve complementary purposes depending on the implementation context. Independent validation by researchers without commercial affiliations will also be valuable for confirming generalizability and providing an unbiased evaluation of the measure's comparative utility across different use cases.

Future Directions

This work lays the foundation for several important lines of future research. First, longitudinal studies are needed to examine how resilience resources and stress load interact over time and to test whether the WONE Index is sensitive to intervention-related changes. Such research will help clarify the dynamic nature of adaptive capacity and resilience plasticity and help examine whether certain domains play different roles in resilience decline versus recovery. This work may also reveal that the factors contributing to stress-related problems differ from those most critical for recovery, information that could further refine how digital platforms prioritize and sequence intervention recommendations to maximize impact during different phases of the resilience process.

Second, clinical applications represent a critical next step. Embedding the WONE Index into treatment contexts could provide insights into how resilience factors contribute to therapeutic outcomes, as well as identify which domains serve as leverage points for recovery. The hierarchical, multidomain structure is particularly well-suited for monitoring differential change across domains during intervention.

Third, contextual and multimethod integration holds promise for advancing ecological validity. Pairing the Index with ecological momentary assessment could capture the contexts in which stress triggers arise and the adaptive strategies mobilized in response, illuminating how resilience unfolds in daily life. This could be further enhanced through pairing with objective markers, including wearable-derived measures (eg, sleep patterns, activity, and heart rate variability) and biomarkers (eg, cortisol and inflammatory cytokines). Such multimethod approaches would reduce reliance on self-report alone, allow researchers to examine how subjective and biological processes align or diverge, and provide a more complete picture of adaptive capacity as a multilevel system.

Fourth, cross-cultural research is needed to explore how resilience processes manifest across diverse sociocultural contexts. While the Index demonstrated measurement invariance across gender, race, and age, expanding to other cultural and occupational groups would strengthen its generalizability and highlight context-specific resilience mechanisms.

Finally, continued methodological innovation will be essential for advancing resilience science. While the hybrid weighting system provides a balanced approach to scoring, the WONE Index also creates opportunities for more advanced analytic methods that can further expand resilience science. Machine learning can extend beyond traditional statistical models by prioritizing prediction and generalizability. It can leverage the Index's structure to detect complex nonlinear interactions, uncover latent resilience profiles through clustering methods, and integrate multimodal data (eg, self-report, wearables, and biomarkers). Machine learning's emphasis on out-of-sample

XSL•FO
RenderX

accuracy and scalability makes it particularly useful for precision prediction in digital health contexts.

Future research can also incorporate Bayesian and systems-based approaches to deepen mechanistic understanding. Bayesian approaches can model uncertainty around resilience estimates, incorporate previous theoretical knowledge, and improve estimation in smaller or intensive longitudinal samples such as ecological momentary assessment. Additional frameworks, including network analysis, dynamical systems modeling, and mixture modeling, could further illuminate how resilience operates as a system, identify leverage points, and detect early warning signals of maladaptive change. Together, these approaches position the WONE Index as both a measurement tool and a platform for advancing predictive and mechanistic models of resilience.

Conclusion

The WONE Index successfully bridges scientific rigor and practical utility, addressing the assessment-intervention gap that has limited resilience research impact in applied settings. Strong incremental validity beyond gold-standard measures demonstrates that the Index captures unique aspects of stress-resilience capacity not assessed by existing tools. The methodological innovation of hybrid weighting—balancing empirical prediction with theoretical modifiability—establishes a strengthened standard for intervention-focused measurement development.

By simultaneously assessing stress load and resilience resources within a unified framework specifically designed for digital delivery, the Index enables personalized intervention matching at scale while maintaining rigorous psychometric standards. As digital mental health expands, measures that satisfy both scientific and practical requirements will be essential for enhancing treatment effectiveness and accessibility. The WONE Index provides a scientifically grounded foundation for this evolution, serving as both a research tool and a platform-integrated assessment system.

Acknowledgments

We would like to thank Carolina Estevao for lending her expertise to an earlier version of the WONE Index, as well as Gabriella Bergin-Cartwright for her support in managing data collection for this project. We would also like to thank the participants of this study, as well as the employers who make this platform available to their employees. Portions of this manuscript were developed with the assistance of generative artificial intelligence (ChatGPT-4 from OpenAI; Claude AI, Anthropic). These tools were used to assist in preparing the manuscript's initial outline, resolving statistical software troubleshooting, and editorial refinement. Any AI-generated content served purely as a drafting aid and was critically reviewed and verified for accuracy by the authors and reworded by the authors prior to inclusion. The authors take full responsibility for the integrity and accuracy of the final manuscript content.

Funding

This study was funded by Walking on Earth.

Data Availability

Individual deidentified data that underlie the results reported in this manuscript can be shared privately for research purposes upon receipt of a methodologically sound proposal whose proposed use of the data is approved by the authors and Walking on Earth's legal and security teams. To gain access, requesters will need to submit a proposal to the corresponding author and sign a data access agreement that includes a commitment to (1) using the data only for research purposes; (2) not attempting to or actually reidentifying any individual; (3) securing the data using appropriate safeguards; and (4) destroying or returning the data after analyses are completed. The WONE Index is a proprietary measure copyrighted by Walking on Earth, Inc. Full administration materials, including detailed instructions and response options, are not publicly available. However, Walking on Earth welcomes collaboration and research partnerships, and the measure may be made available for use upon reasonable request via the company's website or by emailing science@walkingonearth.com.

Authors' Contributions

LGR led the conceptualization of the study and was responsible for data curation, formal analysis, investigation, and methodology development. LG also oversaw project administration and supervision and contributed to both the original drafting and the review and editing of the manuscript. DG contributed to the study through formal analysis, methodological support, and visualization, and participated in writing the original draft as well as reviewing and editing the manuscript. KJ contributed to the visualization and participated in writing the original draft, along with reviewing and editing the manuscript. RM contributed to the study's conceptualization and provided key resources, in addition to participating in the review and editing of the manuscript.

Conflicts of Interest

LGR and RM are employees of Walking on Earth and have received a salary and stock options from the company.

Multimedia Appendix 1

XSL•FO
RenderX

Power analyses for Phase 1 (N=1005) and Phase 2 (N=306), Phase 1 and Phase 2 participant demographics (Tables S1-S2), Phase 1 CFA factor loadings and communalities (Table S3), Phase 1 internal consistency statistics (Table S4), Phase 1 HTMT discriminant validity matrix (Table S5), correlations with established measures (Table S6), comparative fit index (CFA) model modifications, and **Phase 2** heterotrait-monotrait (HTMT) discriminant validity results (Tables S7-S8).
[PDF File, 230 KB - jmir_v28i1e81714_app1.pdf]

References

1. Mahmud S, Mohsin M, Dewan MN, Muyeed A. The global prevalence of depression, anxiety, stress, and insomnia among general population during COVID-19 pandemic: a systematic review and meta-analysis. Trends Psychol 2023;31(1):143-170. [doi: 10.1007/s43076-021-00116-9] [Medline: 40477944]
2. Slavich GM. Social Safety Theory: understanding social stress, disease risk, resilience, and behavior during the COVID-19 pandemic and beyond. Curr Opin Psychol 2022 Jun;45:101299. [doi: 10.1016/j.copsyc.2022.101299] [Medline: 35219156]
3. Masten AS. Resilience in developmental systems: principles, pathways, and protective processes in research and practice. In: Multisystemic Resilience: Adaptation and Transformation in Contexts of Change: Oxford University Press; 2021:113-134.
4. Bonanno GA, Romero SA, Klein SI. The temporal elements of psychological resilience: an integrative framework for the study of individuals, families, and communities. Psychol Inq 2015 Apr 3;26(2):139-169. [doi: 10.1080/1047840X.2015.992677]
5. Southwick SM, Charney DS. The science of resilience: implications for the prevention and treatment of depression. Science 2012 Oct 5;338(6103):79-82. [doi: 10.1126/science.1222942] [Medline: 23042887]
6. Insel T. Digital mental health care: five lessons from Act 1 and a preview of Acts 2-5. NPJ Digit Med 2023 Jan 26;6(1):9. [doi: 10.1038/s41746-023-00760-8] [Medline: 36702920]
7. Torous J, Linardon J, Goldberg SB, et al. The evolving field of digital mental health: current evidence and implementation issues for smartphone apps, generative artificial intelligence, and virtual reality. World Psychiatry 2025 Jun;24(2):156-174. [doi: 10.1002/wps.21299] [Medline: 40371757]
8. Windle G, Bennett KM, Noyes J. A methodological review of resilience measurement scales. Health Qual Life Outcomes 2011 Feb 4;9(1):8. [doi: 10.1186/1477-7525-9-8] [Medline: 21294858]
9. Hill Y, Dolezal ML, Nordbeck PC, et al. Moving from traits to the dynamic process: the next steps in research on human resilience. J Aggress Maltreat Trauma 2025 Jul 3;34(7):971-989. [doi: 10.1080/10926771.2024.2431733]
10. Luthar SS, Cicchetti D, Becker B. The construct of resilience: a critical evaluation and guidelines for future work. Child Dev 2000;71(3):543-562. [doi: 10.1111/1467-8624.00164] [Medline: 10953923]
11. Kalisch R, Baker DG, Basten U, et al. The resilience framework as a strategy to combat stress-related disorders. Nat Hum Behav 2017 Nov;1(11):784-790. [doi: 10.1038/s41562-017-0200-8] [Medline: 31024125]
12. Bakker AB, Demerouti E. Job demands-resources theory: taking stock and looking forward. J Occup Health Psychol 2017 Jul;22(3):273-285. [doi: 10.1037/ocp0000056] [Medline: 27732008]
13. Demerouti E, Bakker AB, Nachreiner F, Schaufeli WB. The job demands-resources model of burnout. J Appl Psychol 2001 Jun;86(3):499-512. [doi: 10.1037/0021-9010.86.3.499] [Medline: 11419809]
14. Hobfoll SE, Halbesleben J, Neveu JP, Westman M. Conservation of resources in the organizational context: the reality of resources and their consequences. Annu Rev Organ Psychol Organ Behav 2018 Jan 21;5(1):103-128 [FREE Full text] [doi: 10.1146/annurev-orgpsych-032117-104640]
15. Biggs A, Brough P, Drummond S. Lazarus and Folkman's psychological stress and coping theory. In: Cooper CL, Quick JC, editors. The Handbook of Stress and Health, 1st edition 2017. [doi: 10.1002/9781118993811]
16. Folkman S, Moskowitz JT. Coping: pitfalls and promise. Annu Rev Psychol 2004;55(1):745-774. [doi: 10.1146/annurev.psych.55.090902.141456] [Medline: 14744233]
17. Lazarus RS, Folkman S. Stress, Appraisal, and Coping: Springer; 1984.
18. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [doi: 10.2307/2136404] [Medline: 6668417]
19. Cohen S, Wills TA. Stress, social support, and the buffering hypothesis. Psychol Bull 1985 Sep;98(2):310-357. [doi: 10.1037/0033-2909.98.2.310] [Medline: 3901065]
20. Connor KM, Davidson JRT. Development of a new resilience scale: the Connor-Davidson Resilience Scale (CD-RISC). Depress Anxiety 2003;18(2):76-82. [doi: 10.1002/da.10113] [Medline: 12964174]
21. Smith BW, Dalen J, Wiggins K, Tooley E, Christopher P, Bernard J. The brief resilience scale: assessing the ability to bounce back. Int J Behav Med 2008;15(3):194-200. [doi: 10.1080/10705500802222972] [Medline: 18696313]
22. MacCallum RC, Browne MW, Sugawara HM. Power analysis and determination of sample size for covariance structure modeling. Psychol Methods 1996;1(2):130-149. [doi: 10.1037//1082-989X.1.2.130]
23. Bonett DG. Sample size requirements for testing and estimating coefficient alpha. J Educ Behav Stat 2002 Dec;27(4):335-340. [doi: 10.3102/10769986027004335]
24. Zou GY. Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. Stat Med 2012 Dec 20;31(29):3972-3981. [doi: 10.1002/sim.5466] [Medline: 22764084]

XSL•FO
RenderX

25. Kaiser BN, Haroz EE, Kohrt BA, Bolton PA, Bass JK, Hinton DE. “Thinking too much”: a systematic review of a common idiom of distress. Soc Sci Med 2015 Dec;147:170-183. [doi: 10.1016/j.socscimed.2015.10.044] [Medline: 26584235]
26. Lewis-Fernández R, Kirmayer LJ. Cultural concepts of distress and psychiatric disorders: understanding symptom experience and expression in context. Transcult Psychiatry 2019 Aug;56(4):786-803. [doi: 10.1177/1363461519861795] [Medline: 31347476]
27. Schmid RF, Thomas J, Rentzsch K. Individual differences in parasympathetic nervous system reactivity in response to everyday stress are associated with momentary emotional exhaustion. Sci Rep 2024 Nov 4;14(1):26662. [doi: 10.1038/s41598-024-74873-9] [Medline: 39496636]
28. Jerotic S, Ignjatovic N, Maric NP, et al. A comparative study on mental disorder conceptualization: a cross-disciplinary analysis. Community Ment Health J 2024 May;60(4):813-825. [doi: 10.1007/s10597-024-01240-3] [Medline: 38319528]
29. Tweed RG, White K, Lehman DR. Culture, stress, and coping: internally- and externally-targeted control strategies of European Canadians, East Asian Canadians, and Japanese. J Cross-Cult Psychol 2004;35(6):652-668. [doi: 10.1177/0022022104270109]
30. Chentsova-Dutton YE, Ryder AG, Tsai JL. Understanding depression across cultural contexts. In: Handbook of Depression, 3rd edition: The Guilford Press; 2014.
31. Matthews G, Szalma J, Panganiban AR, Neubauer C, Warm JS. Profiling task stress with the dundee stress state questionnaire. In: Cavalcanti L, Azevedo S, editors. Psychology of Stress: Nova; 2013:49-91.
32. Maslach C, Schaufeli WB, Leiter MP. Job burnout. Annu Rev Psychol 2001;52(1):397-422. [doi: 10.1146/annurev.psych.52.1.397] [Medline: 11148311]
33. Schaufeli WB, Leiter MP, Maslach C. Burnout: 35 years of research and practice. Career Development International 2009 Jun 19;14(3):204-220. [doi: 10.1108/13620430910966406]
34. Lee EH. Review of the psychometric evidence of the Perceived Stress Scale. Asian Nurs Res (Korean Soc Nurs Sci) 2012 Dec;6(4):121-127. [doi: 10.1016/j.anr.2012.08.004] [Medline: 25031113]
35. Figalová N, Charvát M, Univerzita Palackého v Olomouci, Czech Republic. The Perceived Stress Scale: reliability and validity study in the Czech Republic. Ceskoslov psychol 2021;65(1):46-59. [doi: 10.51561/cspsych.65.1.46]
36. Campbell-Sills L, Stein MB. Psychometric analysis and refinement of the Connor-Davidson Resilience Scale (CD-RISC): validation of a 10-item measure of resilience. J Trauma Stress 2007 Dec;20(6):1019-1028. [doi: 10.1002/jts.20271] [Medline: 18157881]
37. Pilkonis PA, Choi SW, Reise SP, et al. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): depression, anxiety, and anger. Assessment 2011 Sep;18(3):263-283. [doi: 10.1177/1073191111411667] [Medline: 21697139]
38. Schalet BD, Pilkonis PA, Yu L, et al. Clinical validity of PROMIS depression, anxiety, and anger across diverse clinical samples. J Clin Epidemiol 2016 May;73:119-127. [doi: 10.1016/j.jclinepi.2015.08.036] [Medline: 26931289]
39. Moazzami M, Katz P, Bonilla D, et al. Validity and reliability of patient reported outcomes measurement information system computerized adaptive tests in systemic lupus erythematous. Lupus (Los Angel) 2021 Nov;30(13):2102-2113. [doi: 10.1177/09612033211051275]
40. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
41. Löwe B, Decker O, Müller S, et al. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. Med Care 2008 Mar;46(3):266-274. [doi: 10.1097/MLR.0b013e318160d093] [Medline: 18388841]
42. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 Well-Being Index: a systematic review of the literature. Psychother Psychosom 2015;84(3):167-176. [doi: 10.1159/000376585] [Medline: 25831962]
43. Hajos TRS, Pouwer F, Skovlund SE, et al. Psychometric and screening properties of the WHO-5 well-being index in adult outpatients with type 1 or type 2 diabetes mellitus. Diabet Med 2013 Feb;30(2):e63-e69. [doi: 10.1111/dme.12040] [Medline: 23072401]
44. Tabachnick BG, Fidell LS. Using Multivariate Statistics, 7th edition: Pearson; 2019.
45. Fu Y, Wen Z, Wang Y. A comparison of reliability estimation based on confirmatory factor analysis and exploratory structural equation models. Educ Psychol Meas 2022 Apr;82(2):205-224. [doi: 10.1177/00131644211008953] [Medline: 35185157]
46. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. J Mark Res 1981 Feb;18(1):39-50. [doi: 10.1177/002224378101800104]
47. Pfefferbaum B, North CS. Mental health and the Covid-19 pandemic. N Engl J Med 2020 Aug 6;383(6):510-512. [doi: 10.1056/NEJMp2008017] [Medline: 32283003]
48. Stress in America 2023: a nation recovering from collective trauma. American Psychological Association. 2023. URL: https://www.apa.org/news/press/releases/stress/2023/collective-trauma-recovery [accessed 2025-12-13]
49. Slavich GM. Social Safety Theory: a biologically based evolutionary perspective on life stress, health, and behavior. Annu Rev Clin Psychol 2020 May 7;16(1):265-295. [doi: 10.1146/annurev-clinpsy-032816-045159] [Medline: 32141764]
50. Clark LA, Watson D. Constructing validity: basic issues in objective scale development. Psychol Assess 1995;7(3):309-319. [doi: 10.1037//1040-3590.7.3.309]

XSL•FO
RenderX

51. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
52. Shin C, Lee SH, Han KM, Yoon HK, Han C. Comparison of the usefulness of the PHQ-8 and PHQ-9 for screening for major depressive disorder: analysis of psychiatric outpatient data. Psychiatry Investig 2019 Apr;16(4):300-305. [doi: 10.30773/pi.2019.02.01] [Medline: 31042692]
53. Wu Y, Levis B, Riehm KE, et al. Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis. Psychol Med 2020 Jun;50(8):1368-1380. [doi: 10.1017/S0033291719001314] [Medline: 31298180]
54. Wells TS, Horton JL, LeardMann CA, Jacobson IG, Boyko EJ. A comparison of the PRIME-MD PHQ-9 and PHQ-8 in a large military prospective study, the Millennium Cohort Study. J Affect Disord 2013 May 15;148(1):77-83. [doi: 10.1016/j.jad.2012.11.052] [Medline: 23246365]
55. Rammstedt B, John OP. Measuring personality in one minute or less: a 10-item short version of the Big Five Inventory in English and German. J Res Pers 2007 Feb;41(1):203-212. [doi: 10.1016/j.jrp.2006.02.001]
56. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Modeling 1999 Jan;6(1):1-55. [doi: 10.1080/10705519909540118]
57. Kline RB. Principles and Practice of Structural Equation Modeling, 4th edition: Guilford Press; 2016.
58. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163. [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
59. Chen FF. Sensitivity of goodness of fit indexes to lack of measurement invariance. Struct Equ Modeling 2007 Jul 31;14(3):464-504. [doi: 10.1080/10705510701301834]
60. Liu JJW, Reed M, Girard TA. Advancing resilience: an integrative, multi-system model of resilience. Pers Individ Dif 2017 Jun;111:111-118. [doi: 10.1016/j.paid.2017.02.007]
61. McEwen BS, Gray J, Nasca C. Recognizing resilience: learning from the effects of stress on the brain. Neurobiol Stress 2015 Jan 1;1:1-11. [doi: 10.1016/j.ynstr.2014.09.001] [Medline: 25506601]
62. Epel ES, Crosswell AD, Mayer SE, et al. More than a feeling: a unified view of stress measurement for population science. Front Neuroendocrinol 2018 Apr;49:146-169. [doi: 10.1016/j.yfrne.2018.03.001] [Medline: 29551356]
63. Maslach C, Leiter MP. Understanding the burnout experience: recent research and its implications for psychiatry. World Psychiatry 2016 Jun;15(2):103-111. [doi: 10.1002/wps.20311] [Medline: 27265691]
64. Yuan KH, Bentler PM. 5. Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. Sociol Methodol 2000 Aug;30(1):165-200. [doi: 10.1111/0081-1750.00078]
65. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Lawrence Erlbaum Associates; 1988. URL: https://utstat.utoronto.ca/~brunner/oldclass/378f16/readings/CohenPower.pdf [accessed 2025-12-13]
66. Zell E, Stockus CA. Social support and psychological adjustment: a quantitative synthesis of 60 meta-analyses. Am Psychol 2025 Jan;80(1):33-46. [doi: 10.1037/amp0001323] [Medline: 38695783]
67. Vila J. Social support and longevity: meta-analysis-based evidence and psychobiological mechanisms. Front Psychol 2021;12:717164. [doi: 10.3389/fpsyg.2021.717164] [Medline: 34589025]
68. Chaput JP, Dutil C, Featherstone R, et al. Sleep timing, sleep consistency, and health in adults: a systematic review. Appl Physiol Nutr Metab 2020 Oct;45(10 (Suppl. 2)):S232-S247. [doi: 10.1139/apnm-2020-0032] [Medline: 33054339]
69. Flueckiger L, Lieb R, Meyer AH, Witthauer C, Mata J. The importance of physical activity and sleep for affect on stressful days: two intensive longitudinal studies. Emotion 2016 Jun;16(4):488-497. [doi: 10.1037/emo0000143] [Medline: 26709860]
70. Mücke M, Ludyga S, Colledge F, Gerber M. Influence of regular physical activity and fitness on stress reactivity as measured with the trier social stress test protocol: a systematic review. Sports Med 2018 Nov;48(11):2607-2622. [doi: 10.1007/s40279-018-0979-0] [Medline: 30159718]
71. Hakanen JJ, Schaufeli WB. Do burnout and work engagement predict depressive symptoms and life satisfaction? A three-wave seven-year prospective study. J Affect Disord 2012 Dec 10;141(2-3):415-424. [doi: 10.1016/j.jad.2012.02.043] [Medline: 22445702]
72. Seery MD. Resilience: a silver lining to experiencing adverse life events? Curr Dir Psychol Sci 2011;20(6):390-394. [doi: 10.1177/0963721411424740]
73. Chung KW, Martz CD, Lutz B, et al. Skin-deep resilience in the Black women's experiences living with lupus study. Health Psychol 2025 Aug;44(8):800-809. [doi: 10.1037/hea0001469] [Medline: 40193435]
74. Lowe SR. Embracing complexity in resilience research. Nat Mental Health 2025;3(4):391-392. [doi: 10.1038/s44220-025-00403-9]
75. Boswell JF, Kraus DR, Castonguay LG, Youn SJ. Treatment outcome package: measuring and facilitating multidimensional change. Psychotherapy (Chic) 2015 Dec;52(4):422-431. [doi: 10.1037/pst0000028] [Medline: 26641372]
76. O'Donohue JS, Mesagno C, O'Brien B. How can stress resilience be monitored? A systematic review of measurement in humans. Curr Psychol 2021 Jun;40(6):2853-2876. [doi: 10.1007/s12144-019-00226-9]
77. Friborg O, Hjemdal O, Rosenvinge JH, Martinussen M. A new rating scale for adult resilience: what are the central protective resources behind healthy adjustment? Int J Methods Psych Res 2003 Jun;12(2):65-76. [doi: 10.1002/mpr.143]

XSL•FO
RenderX

78. Leipold B, Klier K, Dapperger E, Schmidt A. Physical activity and nutrition in relation to resilience: a cross-sectional study. Sci Rep 2024 Jan 27;14(1):2272. [doi: 10.1038/s41598-024-52753-6] [Medline: 38280920]
79. Bertollo AG, Capitanio MZ, Schuh LA, Pradella N, Ignácio ZM. Habits and vulnerability or resilience to stress - impact on depressive disorders. Behav Brain Res 2025 Jul 26;490:115630. [doi: 10.1016/j.bbr.2025.115630] [Medline: 40334944]
80. Rink LC, Silva SG, Adair KC, Oyesanya TO, Humphreys JC, Sexton JB. The association between well-being behaviors and resilience in health care workers. West J Nurs Res 2022 Aug;44(8):743-754. [doi: 10.1177/01939459211017515] [Medline: 34039117]
81. Bringmann L, Helmich M, Eronen M, Voelkle M. Complex systems approaches to psychopathology. In: Krueger RF, Blaney PH, editors. Oxford Textbook of Psychopathology: Oxford University Press; 2023:103-122.

Abbreviations

AVE: average variance extracted
BFI-10: Big Five Inventory-10
BRS: Brief Resilience Scale
CD-RISC: Connor-Davidson Resilience Scale
CES-D: Center for Epidemiologic Studies Depression
CFA: confirmatory factor analysis
CFI: comparative fit index
CR: composite reliability
***DSM*:** *Diagnostic and Statistical Manual of Mental Disorders*
EFA: exploratory factor analysis
GAD-7: Generalized Anxiety Disorder-7
HTMT: heterotrait-monotrait
ICC: intraclass correlation coefficient
IRB: Institutional Review Board
JD-R: job demands-resources
KMO: Kaiser-Meyer-Olkin
MI: modification index
PHQ: Patient Health Questionnaire
PROMIS-SF-8a: Patient-Reported Outcomes Measurement Information System Short Form 8a
PSS: Perceived Stress Scale
RMSEA: root mean square error of approximation
SRMR: standardized root mean square residual
TLI: Tucker-Lewis index
WHO-5: World Health Organization-5 Well-Being Index
WONE: Walking on Earth
ΔR^2: R^2 change

Edited by N Cahill; submitted 01.Aug.2025; peer-reviewed by H Mahmoodi, NH Zainal; revised version received 07.Nov.2025; accepted 10.Nov.2025; published 05.Jan.2026.

Please cite as:
Roos LG, Gilliland D, Julian K, Misra R
The WONE Index as a Multidimensional Assessment of Stress Resilience: A Development and Validation Study
J Med Internet Res 2026;28:e81714
URL: https://www.jmir.org/2026/1/e81714
doi:10.2196/81714

© Lydia Genevieve Roos, Destiny Gilliland, Kelsey Julian, Reeva Misra. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Program Theory and Core Outcome Set Development for a Technology-Assisted Counseling Intervention in Dementia: Multimethods Study

Dorothee Bauernschmidt, Dr rer medic; Anja Bieber, Dr rer medic; Ronja Hubrich, MSc; Janina Wittmann, MSc; Gabriele Meyer, Dr phil

Institute of Health, Midwifery and Nursing Science, Medical Faculty, University Medicine Halle, Martin Luther University Halle-Wittenberg, Magdeburger Straße 8, Halle (Saale), Germany

Corresponding Author:
Dorothee Bauernschmidt, Dr rer medic
Institute of Health, Midwifery and Nursing Science, Medical Faculty, University Medicine Halle, Martin Luther University Halle-Wittenberg, Magdeburger Straße 8, Halle (Saale), Germany

Abstract

Background: Counseling in family dementia care aims to support caregivers in mastering challenges. The use of information and communication technologies (ICT) to administer counseling can improve accessibility. Evidence syntheses report inconsistent findings on the effectiveness of technology-assisted counseling. There is a considerable heterogeneity in outcomes assessed in clinical trials, and approaches to develop and evaluate interventions are not guided by theory in most cases.

Objective: This study aims to develop an initial program theory of a technology-assisted counseling intervention for family dementia caregivers and to create the data basis for the consensus process of a core outcome set.

Methods: We integrated the methodological strands for the development of a program theory and a core outcome set in an innovative way. A scoping review was conducted to collect data on characteristics and theoretical foundations of technology-mediated counseling interventions as well as outcomes of clinical studies. We explored the lived experience of relevant interest-holders and conducted semistructured interviews applying a phenomenological approach to data analysis. Synthesis of findings was performed by developing a logic model and formulating an initial program theory.

Results: We included 69 records reporting on 34 interventions. Designs and other study characteristics vary, and interventions are heterogeneous in terms of components and ICT used for delivering counseling. We conducted interviews with 15 family caregivers and 12 counselors. The themes *being affected*, *feeling insecure and helpless in the face of the health care system*, and *search for information and communicative exchange* illustrate the caregivers' lifeworld perception. Themes identified in counselors' interviews comprise *work attitude and standards*, *unpredictability*, *expectations*, *working conditions*, *organizational influence*, and *tools: techniques and networking*. The constitutive pattern of *having/being somebody to count on* was incorporated into the program theory. In the theory of change, we describe the way to a sustainable supportive cooperation between caregivers and counselors ensuring ongoing support throughout the caregiving process. We explicate the effects of the technology-assisted counseling intervention such as improved knowledge, attitude, and interaction, as well as stability and safety of care in the outcomes chain. The theory of action comprises the inputs, activities, and outputs of the intervention. The graphical synthesis of findings is presented in the logic model.

Conclusions: To effectively develop, implement, and evaluate technology-assisted counseling in family dementia care, a theory-led approach is essential. A carefully modeled intervention that combines technological options with in-person counseling may help to overcome disparities in access to health care and improve accessibility to counseling. A supportive working environment for counselors, in which artificial intelligence is used to reduce time spent on documentation and administrative tasks, may help mitigate the effects of the growing shortage of skilled professionals.

Trial Registration: Core Outcome Measures in Effectiveness Trials (COMET) Initiative 2884; https://www.comet-initiative.org/Studies/Details/2884

(J Med Internet Res 2026;28:e81669) doi:10.2196/81669

KEYWORDS

theory; core outcome set; logic model; counseling; family caregiver; dementia

XSL•FO
RenderX

Introduction

Counseling interventions in family dementia care aim to support caregivers in mastering challenges in caregiving. Information and communication technologies (ICT) are used to deliver counseling in an easily accessible way [1-3].

Counseling may contribute to mitigating the negative impact family dementia care can exert on caregivers, which is described in terms of burden, depression, decreased health and quality of life, as well as social isolation [4,5]. In light of the predicted increase in the number of persons with dementia [6], the need for support interventions for family caregivers is also likely to increase. Provided by professionals, counseling can be defined as the "use of an interactive helping process focusing on the needs, problems, or feelings of the patient and significant others to enhance or support coping, problem-solving, and interpersonal relationships" [7]. The use of ICT is discussed to reduce barriers to utilization by overcoming distances, enabling persons who are homebound or living in rural areas to participate in services, offering anonymous counseling, or fostering asynchronous communication beneficial for persons who are employed [3,8]. Types of ICT have expanded over time: long-established helpline services use the telephone as a widespread and undemanding technology. More recently, counseling is provided via videoconferencing software, email, or chats using mobile devices and applications, thus leading to increased demands on technological infrastructure, equipment, and digital literacy [8].

Technology-assisted counseling (also referred to as "technology-based" in previous publications) is a complex intervention [9], and the development, implementation, and evaluation of complex interventions are challenging. The effectiveness of technology-assisted counseling in dementia has not yet been proven: Evidence syntheses of technology-assisted psychosocial interventions including counseling for caregivers of persons with dementia show inconsistent findings on positive effects on outcomes such as burden, depression, or quality of life [1-3,10]. Our own meta-analyses revealed no significant effects of technology-assisted counseling on depressive symptoms, burden, and self-efficacy or mastery perceived by family caregivers of persons with dementia [11]. We found a considerable heterogeneity in outcomes and a lack of theoretical approaches guiding the development, implementation, and evaluation of the interventions [8,11].

Therefore, we conducted the ProCOS project: "Development and Evaluation of a Technology-Assisted Counseling Intervention for Family Caregivers of Persons With Dementia – Program Theory and Preparation of a Core Outcome Set" [12]. Within the 12-month project, we aimed at creating the foundation for the future consensus process of a core outcome set (COS) and at developing a program theory by combining the methodological strands of the two developmental processes in an innovative way.

A COS is "an agreed standardized collection of outcomes which should be measured and reported, as a minimum, in all trials for a specific clinical area" [13]. The use of a COS reduces heterogeneity in clinical trials and enhances comparability and thus synthesis of evidence [13]. Core outcome sets for the evaluation of health care interventions and of psychosocial community-based interventions for persons with dementia living at home predominately focus on outcomes of persons with dementia [14,15]. A set of measures has been recommended to evaluate a broad range of psychosocial interventions for persons with dementia and their family caregivers [16]. There is no COS that specifically focuses on technology-assisted counseling interventions for family dementia caregivers.

The Framework for Developing and Evaluating Complex Interventions identifies program theory as a core element of complex interventions [9]. This underlines the importance of theoretical approaches to successfully develop, implement, and evaluate complex interventions. A program theory is an "explicit theory of how an intervention is understood to contribute to its intended or observed outcomes" [17]. Following the approach introduced by Funnell and Rogers [17], we developed a "purposeful program theory" comprising the theory of change, the outcomes chain, and the theory of action. The theory of change explicates the central mechanism of how the intended changes can be achieved, and the theory of action explains how the intervention is designed to initiate the theory of change. These elements are linked by the outcomes chain comprising the immediate and intermediate outcomes and the impact of the intervention, as well as hypothesized relationships between outcomes [17]. In line with the updated UK Medical Research Council framework, we understand a program theory as a detailed textual description [18]. Logic models serve as visual representations of program theories [17,18]. Detailed logic models graphically illustrate the (assumed) causal mechanisms through which an intervention is expected to produce outcomes, as well as contextual dependencies and preconditions [18]. Developing a program theory that incorporates perspectives of diverse interest-holders and integrates theoretical and empirical knowledge at the beginning of interventional research is considered best practice [9]. Program theories and logic models are adapted and refined throughout the development, implementation, and evaluation of complex interventions to address the question, "What works in which circumstances and how?"—thus applying a theory-based approach [9].

As the methodological strands both for developing a program theory and a COS integrate knowledge obtained from literature and perspectives from different interest-holders [13,17], we brought these two processes together to develop a program theory and prepare the consensus process of a COS of a technology-assisted counseling intervention for caregivers of persons with dementia. The central instrument for this innovative approach is the logic model. As graphical representations of program theories [17], logic models have been used to synthesize data [19,20] and to visualize assumed causal relationships and mechanisms of action of complex interventions [21]. By fostering a shared understanding among interest-holders [9,17], we used the logic model for data synthesis, and we will integrate it in the future consensus process of the COS, allowing interest-holders to critically review the quality of the program theory. To our knowledge, this approach has not yet been implemented.

By drawing on previous work on the effectiveness and implementation success of technology-assisted counseling in

XSL•FO
RenderX

dementia [8,11,22], we therefore aimed at addressing gaps in knowledge. Thus, our first objective was to develop a program theory of a technology-assisted counseling intervention for family dementia caregivers. Our secondary objective was to compile lists of (potential) outcomes that have been identified by caregivers or counselors or assessed in clinical studies.

The following two sets of questions guided our research:

1. What interventions that use ICT to provide counseling for family dementia caregivers are described in literature? What are the characteristics of these interventions? What theoretical underpinnings for intervention development and implementation are explicated in the form of theoretical references, program theories, or logic models? What outcomes have been examined in clinical trials? What assessment instruments have been used?
2. How do family caregivers for persons with dementia and counselors experience counseling services? What expectations do persons seeking or providing counseling have of each other? Which outcomes should or could be achieved through counseling, and how can these outcomes be achieved? Which factors have an impact on the effectiveness of counseling? What are possible outcomes for assessing the effectiveness of counseling interventions?

Methods

The ProCOS project was registered with the Core Outcome Measures in Effectiveness Trials (COMET) Initiative [23]. The study protocol has been published [12].

Study Design

We adopted a multimethod design including a literature review and a qualitative substudy [24]. Results were graphically synthesized into a logic model and an initial program theory was formulated comprising the theory of change, the outcomes chain, and the theory of action [17].

We applied the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses, Extension for Scoping Reviews) [25], the SRQR (Standards for Reporting Qualitative Research) [26], and the COS-STAR (Core Outcome Set—Standards for Reporting) [27] to structure our report.

Scoping Review

We aimed to map the knowledge on technology-assisted counseling interventions for family caregivers of persons with dementia and followed the Joanna Briggs Institute methodological guidance for scoping reviews [28].

Eligibility Criteria

We included studies on interventions using ICT to deliver individualized counseling to caregivers of persons with dementia. Publications in the German or English language were accepted, irrespective of their design. Detailed inclusion and exclusion criteria according to the Population-Concept-Context scheme are provided in the study protocol [12].

Search Strategy and Information Sources

By updating the literature search of a previous systematic review [8,11,22], we searched the databases CINAHL, MEDLINE via PubMed, Cochrane Library, and PsycINFO (December 2023) and conducted additional forward and backward citation searching as well as a free web search (April 2024). The database-specific search strategies and the search terms of the free web search are provided in Multimedia Appendix 1.

Selection of Sources of Evidence

Titles, abstracts, and full texts were screened independently by two researchers (DB and RH) using the Rayyan web app [29]. Discrepancies in decisions were resolved by discussion.

Data Charting Process and Data Items

We extended a previously developed data extraction sheet and extracted data on study characteristics (year of publication, country of study conduct, design and methods, and number of participants) and on outcomes examined as well as assessment instruments used in included studies. We also applied criteria from the Template for Intervention Description and Replication (TIDieR) checklist [30] and from the revised Criteria for Reporting the Development and Evaluation of Complex Interventions (CReDECI 2) guideline [31] in order to collect information on objectives, components, theoretical underpinnings of counseling interventions, technology and materials used for delivering counseling, frequency and duration of sessions, and implementation issues. Data extraction was performed by one reviewer (DB). A cross-check of extracted data was conducted by another researcher (RH) for 20% of included interventions indicating accuracy and completeness of data extraction.

Qualitative Substudy

With the aim to explore the lived experience of counseling in dementia care, we conducted interviews with interest-holders in caregiving and counseling in the context of family dementia care.

Qualitative Approach and Research Paradigm

We adopted a phenomenological perspective [32] and focused on the lifeworld experiences of caregivers and counselors.

Context, Researcher Characteristics, and Reflexivity

All interviews and a preliminary analysis were conducted by the primary investigator (DB), who is experienced in phenomenological research. Results were then discussed within the research team. Team members are nursing scientists with extensive experience in nursing practice and dementia research, and a physiotherapist.

Units of Study and Sampling Strategy

Interviews were conducted with family caregivers and counselors. We used a purposive sampling strategy [33] to obtain a heterogeneous sample. Predefined criteria for recruitment of dementia caregivers were age, gender, socioeconomic status, and family relationship, as well as spatial and emotional proximity to the person receiving care. Predefined criteria for recruitment of counselors were disciplinary background, professional qualifications, duration of counseling

XSL•FO
RenderX

experience, and characteristics of employing organizations. We included persons who have received or delivered counseling via technology, in-person, or both. While our efforts to recruit a heterogeneous sample of family dementia caregivers had limited success, we succeeded more in engaging counselors with diverse characteristics.

Invitations to participate and study information were distributed via existing contacts and networks established through previous research projects and the State Competence Center for Dementia of the German federal state of Saxony-Anhalt. In addition, we contacted counseling services and self-help organizations of family caregivers throughout Germany. Recruitment was completed when no new information emerged from interviews indicating that data saturation was achieved [34,35].

Data Collection Methods and Data Processing

We performed semistructured interviews using an interview guide with open-ended questions to give interviewees room to share their experiences [36]. Questions focusing on experiences in receiving and providing counseling were asked in the course of the interview. Finally, additional data on sociodemographic characteristics, information on the care arrangement, or on the professional situation were collected. The translated interview guide is provided in Multimedia Appendix 2.

Interviews were arranged at participants' convenience to minimize time exposure and burden for participants. We planned to conduct individual interviews, but switched to group interviews at the request of some participating caregivers and counselors.

Recordings of interviews were transcribed verbatim using f4 transcription software [37]. Transcripts were checked for accuracy and pseudonymized by one researcher (RH).

Data Analysis

We conducted an interpretive phenomenological analysis applying the following modified working steps described by Diekelmann [38]:

1. The transcripts are read several times to obtain an overall understanding.
2. An interpretive summary of each interview is written.
3. Interpretive summaries are analyzed and discussed to identify emerging themes.
4. Disagreements in interpretation are resolved by returning to the text.
5. Through comparing and contrasting texts, the themes that recurred and reflected the shared practices and common meanings are identified and described.
6. As themes are compared, a constitutive pattern emerges that links the themes and is present in all interviews.
7. The themes and the constitutive pattern are described using quotations to illustrate findings.

Themes and the constitutive pattern illuminate the lifeworld perception of receiving or providing counseling in the context of family dementia care.

In addition, outcomes designated by caregivers and counselors were independently extracted from the transcripts by two researchers (DB and JW) and discussed intensively in order to include them as accessible statements [39] into the future consensus-building process.

Synthesis of Data Through Logic Model Development and Formulation of a Preliminary Program Theory

Contrary to the original plan to conduct the data synthesis into the logic model and the formulation of the program theory consecutively, we brought these two working steps together in an iterative process. We switched back and forth between writing memos and graphically synthesizing the data. Following the recommendation made by Funnell and Rogers [17], we combined the inductive and deductive approach with articulating interest-holders' mental models to develop the program theory. Mental models are diverse interest-holders' beliefs about how a program achieves its results [17]. We integrated the mental models explicated by caregivers and counselors with data derived from literature and with information on programs operating in practice. We treated data extracted from studies included in the scoping review as qualitative data [20]. Data were charted and categorized [20], and assigned to the elements of the program theory [17].

The starting point for the development of the theory of change was the situation analysis for which we applied the guiding questions formulated by Funnell and Rogers [17]. Key aspects of the situation analysis are depicted as macro-, meso-, and microcontext in the logic model.

We then explicate our assumptions of how the intended changes can be achieved by formulating the theory of change: (hypothesized) mechanisms of effective counseling, as described by interview partners or identified in literature, were incorporated into the theory of change.

Outcomes derived from clinical studies or mentioned by participants were summarized in tables and served as the foundation of the outcomes chain. The development of the outcomes chain followed the steps outlined by Funnell and Rogers [17]: (1) listing of possible outcomes, (2) clustering outcomes and assigning working labels to each cluster, (3) arranging outcomes in a chain of if-then statements, (4) identifying feedback loops, and (5) validating the outcomes chain.

By explicating how the intervention is designed to initiate the theory of change, we developed the theory of action and identified resources (inputs), activities, and outputs [17]. We incorporated intervention components and strategies that were reported in the literature or by study participants as helpful or effective in achieving the intended outcomes in the theory of action.

Patient and Public Involvement

We established a study advisory board consisting of a person with long-term experience in caring for family members with dementia and engagement in an informal support network for family caregivers, a person with extensive counseling experience, and an experienced researcher in the field of dementia care. Representing different groups of interest-holders, the members of the study advisory board critically reviewed

XSL•FO
RenderX

the instruments and approaches for data collection and were involved in the discussion of results and conclusions for the design of a technology-assisted counseling intervention. Three virtual meetings were held in addition to feedback provided via email over the course of the 12-month project.

Techniques to Enhance Trustworthiness

We applied the strategies of expert consultation and peer debriefing by consulting the members of the study advisory board throughout the research process.

Ethical Considerations

The ethics committee of the Medical Faculty of the Martin Luther University Halle-Wittenberg approved the ProCOS study (no. 2023 - 093). Persons interested in participating received written information on procedures prior to the interview. The time and place of the interviews were determined based on the participants' preferences. We obtained written informed consent from the participants who were informed that the consent to participate can be withdrawn at any time. We maintained the security of data by storing data protected from access by persons who are not involved in the project. Audio recordings of interviews were pseudonymized during the transcription process. Study participants received a small gift (equivalent to US $5), but no financial compensation.

Results

Scoping Review

Selection of Sources of Evidence

The updated database search yielded 3995 records. An additional 1965 records were identified through complementary search strategies. After the removal of 1448 duplicates of reports identified via database search and 332 duplicates of reports identified via other methods, 4180 titles and abstracts were screened. A total of 141 full texts were screened for eligibility. Of these, 17 records reporting on 7 interventions were included. These new studies were combined with 52 records on 27 interventions from the previous review, resulting in 69 records reporting on 34 interventions that were included in the scoping review (Figure 1).

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses, Extension for Scoping Reviews) flow diagram for updated reviews [40].

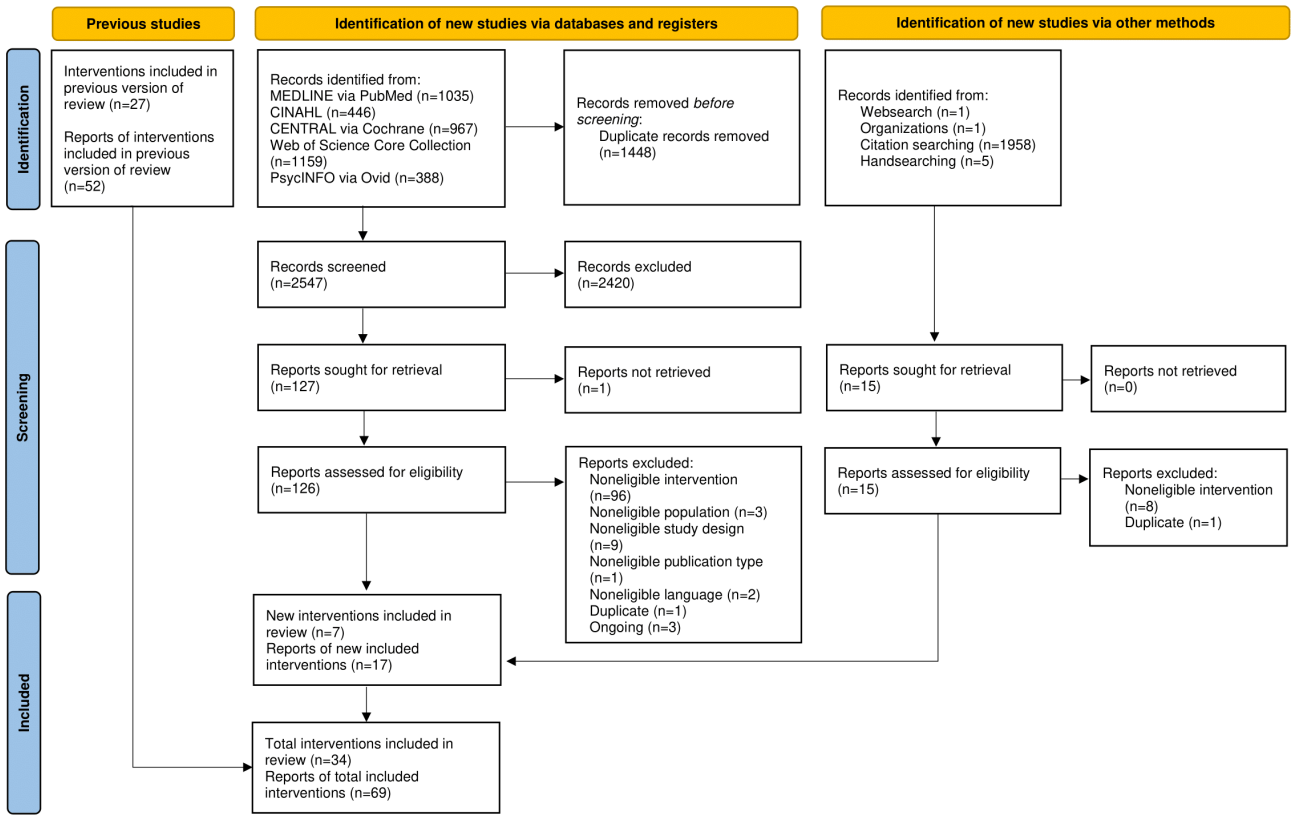

Characteristics of Sources of Evidence

The studies were published between 1993 and 2024 (last search April 2024) and were conducted in 10 countries (Australia, Canada, Germany, Israel, Italy, Japan, the Netherlands, Sweden, the United Kingdom, and the United States).

As we included studies irrespective of their design, topics, and methodological approaches varied. The effectiveness of 7 interventions on outcomes such as depressive symptoms, burden, and self-efficacy (refer to Table 1) has been evaluated by randomized controlled trials (RCTs) [41-63], including a small-scale RCT in the context of a pilot and feasibility study [64,65]. Nonrandomized trials [66-68] and a pre-post-intervention trial [69] assessed the effects on caregivers' ability in disease management, burden, or mental health; quantitative descriptive approaches were conducted to examine sociodemographic characteristics of service users, reasons for use and topics discussed, advice provided, and satisfaction with services, as well as ways of access to

XSL•FO
RenderX

counseling [70-82]. Qualitative approaches were used to explore expectations and experiences of caregivers with technology-assisted counseling services [59,65,70,83-92]. Case studies [93-95] examined individual counseling processes reflecting on the complexity of delivering counseling via ICT, as well as skills and knowledge needed to provide effective support. We included two process evaluation studies focusing on implementation issues when linking eHealth interventions to existing support [96] and combining a cognitive rehabilitation program with education and counseling [97]. Mixed methods approaches were applied to achieve an in-depth understanding of how the intervention works, to assess usability, and to describe and explain (non-)usage of services [56,57,98-103]. Detailed information on included references is presented in Multimedia Appendix 3 [8,30,31,41-109].

XSL•FO
RenderX

Table . Outcomes identified in the literature.

Outcomes and assessment instruments	Intervention[a]	References
Outcomes for caregivers		
Depression		
Zung Self-Rating Depression Scale	Coyne[a] comparator or experimental	[42]
Center for Epidemiology Studies Depression Scale (CES-D)	FITT-C[b]	[44-50]
	RCTM[c]	[54-60,64,65]
Center for Epidemiologic Studies Depression Scale (CES-D-10)	CuidaTEXT	[98,104-106]
Center for Epidemiologic Studies Depression Scale–Revised (CESD-R)	TeleFAMILIES[d]	[69]
Geriatric Depression Scale (GDS)	FITT-D[e]	[43]
Mood Assessment Scale (MAD)	RCTM[c]	[54-60]
Grief		
Caregiver Grief Scale (CGS)	RCTM[c]	[64,65]
Guilt		
Caregiver Guilt Questionnaire (CGQ)	RCTM[c]	[64,65]
Anxiety		
Geriatric Anxiety Inventory (GAI)	RCTM[c]	[64,65]
Burden or caregiver distress		
Zarit Burden Interview (ZBI)	Coyne[a] comparator or experimental	[42]
	FITT-C[b]	[44-50]
	FITT-D[e]	[43]
	TeleFAMILIES[d]	[69]
Zarit Burden Interview (ZBI-6)	CuidaTEXT	[98,104-106]
Burden Scale for Family Caregivers (BSFC)	ICSS[f]	[90-92,99-101]
Caregiver Burden Inventory (CBI)	Natale[a]	[66]
	De Cola[a]	[82]
Subjective stress		
Not specified (7-item measure of care-related strain; 1-item measure assessing the caregiver's difficulty with the relative's mental or emotional state); Adapted Zarit Burden Interview (ZBI)	RCTM[c]	[54-60]
Perceived stress		
Perceived Stress Scale (PSS)	RCTM[c]	[64,65]
Caregiver strain		
Modified Caregiver Strain Index (CSI)	CuidaTEXT	[98,104-106]
Reaction to care-receiver behavior		
Revised Memory and Behavior Problem Checklist (RMBPC)	FITT-C[b]	[44-50]
	FITT-D[e]	[43]
	TeleFAMILIES[d]	[69]

XSL•FO
RenderX

Outcomes and assessment instruments	Intervention[a]	References
Emotional impact of neuropsychiatric symptoms		
Neuropsychiatric Inventory (NPI) Burden subscale	Dementelcoach	[61-63,67,68,96]
Upset		
Caregiver Behavioral Occurrence and Upset Scale (modeled on the Agitated Behavior in Dementia Scale)	Laver[a]	[51-53]
Behavioral symptom severity–distress		
Neuropsychiatric Inventory–Questionnaire (NPI-Q) Distress	CuidaTEXT	[98,104-106]
Coping		
Coping Orientation to Problems Experienced Inventory (COPE-28)	CuidaTEXT	[98,104-106]
Affect		
Scale of Positive and Negative Experience (SPANE)	CuidaTEXT	[98,104-106]
Self-perception of negative and positive aspects of caregiving		
Carers of Older People in Europe (COPE) Index	InformCare	[102,103,107]
Positive aspects of caregiving		
Positive Aspects of Caregiving (PAC) Scale	FITT-C[b]	[44-50]
	CuidaTEXT	[98,104-106]
Self-efficacy		
PROMIS Self-Efficacy for Emotions (derived from the Self-Efficacy for Managing Chronic Disease (SEMCD) questionnaire)	Care Consultation or Care Consultation Plus	[41]
Self-Efficacy Questionnaire (SEQ)	FITT-C[b]	[44-50]
	FITT-D[e]	[43]
	RCTM[c]	[54-60]
Mastery		
Caregiving Mastery Index (CMI)	Laver[a]	[51-53]
Sense of capability		
Short Sense of Competence Scale (SSCQ)	Dementelcoach	[61-63,67,68,96]
	RCTM[c]	[54-60]
Caregiving competence		
Preparedness for Caregiving Scale (PCS)	CuidaTEXT	[98,104-106]
Psychological well-being		
World Health Organization Well-Being Index (WHO-5)	InformCare	[102,103,107]
Happiness		
TOPICS-MDS[g] item	Dementelcoach	[61-63,67,68,96]
Quality of life		
Euro Quality of Life Visual Analog Scale (EQ-5D VAS)	FITT-C[b]	[44-50]

XSL•FO
RenderX

Outcomes and assessment instruments	Intervention[a]	References
TOPICS-MDS[g] item	Dementelcoach	[61-63,67,68,96]
Health		
PROMIS Global Health: Global Physical Health (GPH) and Global Mental Health (GMH)	Care Consultation or Care Consultation Plus	[41]
General Health Questionnaire (GHQ-28)	Dementelcoach	[61-63,67,68,96]
SF 36 General Health	FITT-D[e]	[43]
Not specified (1-item question)	CuidaTEXT	[98,104-106]
Unmet needs		
Measure of Unmet Needs (UN)	CuidaTEXT	[98,104-106]
Perceived social support or support for caring		
Multidimensional Scale of Perceived Social Support (MSPSS)	FITT-D[e]	[43]
	InformCare	[102,103,107]
Interpersonal Support Evaluation List (ISEL-12)	CuidaTEXT	[98,104-106]
Support for Caring subscale of the Adult Carer Quality of Life Questionnaire (AC-QoL)	RCTM[c]	[64,65]
Family functioning		
Family Assessment Device (FAD)	FITT-C[b]	[44-50]
	FITT-D[e]	[43]
Knowledge		
Alzheimer's Disease Knowledge Test	FITT-D[e]	[43]
Epidemiology/Etiology Disease Scale (EEDS)	CuidaTEXT	[98,104-106]
Perceived change		
Perceived Change Scale (PCS)	Laver[a]	[51-53]
Secondary role strains		
Not specified (2 single-item ratings of the caregiver's and the relative's adjustment to residential long-term care placement)	RCTM[c]	[54-60]
Residential care stress		
Not specified (6-item measure for perceptions of staff communication with family, 5-item measure for staff support for family, 10-item measure assessing 5 positive and 5 negative types of caregiver interactions with their relative, staff, other family; item caregivers' upset to see their relative in a residential care setting); Family Involvement Interview	RCTM[c]	[54-60]
Resource use		
Number of community or health services used	Coyne[a] comparator or experimental	[42]
	FITT-C[b]	[44-50]
	FITT-D[e]	[43]
Outcomes for persons with dementia		
Depression		

XSL•FO
RenderX

Outcomes and assessment instruments	Intervention[a]	References
Geriatric Depression Scale (GDS)	De Cola[a]	[82]
Global cognitive state		
Mini-Mental State Examination (MMSE)	De Cola[a]	[82]
Cognitive impairment		
Bedford Alzheimer Nursing Severity Scale (BANSS)	De Cola[a]	[82]
Functional dependency		
Activities of Daily Living (ADL) and Instrumental Activities of Daily Living Scale (IADL)	De Cola[a]	[82]
Instrumental activities of daily living		
Functional Activities Questionnaire (FAQ)	CuidaTEXT	[98,104-106]
Functionality		
Caregiver Assessment of Function and Upset (CAFU)	Laver[a]	[51-53]
Neuropsychiatric symptoms or behavioral symptom severity		
Neuropsychiatric Inventory (NPI) Total symptoms	Dementelcoach	[61-63,67,68,96]
Neuropsychiatric Inventory-Questionnaire (NPI-Q) Severity	Dementelcoach	[61-63,67,68,96]
	CuidaTEXT	[98,104-106]
Psychiatric symptoms		
Brief Psychiatric Rating Scale (BPRS)	De Cola[a]	[82]
Number of behaviors such as verbal aggression, refusing care, restlessness, anxiety, waking overnight, and repetitive questioning		
Caregiver Behavioral Occurrence and Upset Scale (modeled on the Agitated Behavior in Dementia Scale)	Laver[a]	[51-53]
Resource use		
Number of community or health services used	Coyne[a] comparator or experimental	[42]
	FITT-C[b]	[44-50]

[a]When no name is reported, the name of the first author was assigned to the intervention.

[b]FITT-C: Family Intervention: Telephone Tracking–Caregiver.

[c]RCTM: Residential Care Transition Module.

[d]TeleFAMILIES: Telehealth-Administered Families Access to Memory Impairment and Loss Information, Engagement, and Supports.

[e]FITT-D: Family Intervention: Telephone Tracking–Dementia.

[f]ICSS: Internet-Based Caregiver Support Service.

[g]TOPICS-MDS: The Older Persons and Informal Caregivers Survey Minimum DataSet.

Results of Individual Sources of Evidence

We summarized the included interventions into groups that were formed based on the ICT used and the additional components of the interventions. We found 17 interventions that delivered counseling via telephone, email, or videoconferencing [41-50,66,70-78,83-87,93,94,108]. Three interventions provided counseling via videoconferencing [69,79,88]. One intervention each was assigned to the groups "counseling via email" [95], "counseling via SMS text messaging" [98,104-106], and "counseling via an interactive mobile app combined with additional features" [89]. We included 4 web-based psychosocial interventions that combined information, communication, and counseling [80,81,90-92,99-103,107]. Four videoconference- or telephone-based interventions combined counseling services with telemonitoring or psychoeducation [51-60,64,65,82,109]. Three technology-assisted interventions offered counseling as

XSL•FO
RenderX

part of a comprehensive program with non-technology-assisted components [61-63,67,68,85-87,96,97].

The interventions aim to support family caregivers and persons with dementia, to provide information and education, and to enhance coping with and managing of the caregiving process. Theoretical foundations of interventions such as psychological or psychosocial concepts are referred to for 10 interventions [43-65,67,68,70,93-98,104-106,108]. The health service usage behavior has been theorized for an internet-based information support and personalized email intervention [90-92,99-101]. A program theory is mentioned for 1 intervention [83].

In the included reports, various components of the interventions are described, which are combined in different ways: intervention manuals or guidelines are made available for counselors. In some cases, special training is provided on dementia, counseling strategies, or the use of ICT in counseling. Supervision and coaching are offered to persons delivering counseling. There are various modes of documenting the content of the counseling sessions (eg, protocols and log sheets) and making it available to counselees (eg, letters and script proposals). In addition, information and educational material is offered as brochures, manuals, databases, websites, and videotapes. In some cases, the necessary technical equipment or assistance with its use is provided. Persons delivering counseling have qualifications in nursing, social work, social science, psychology, mental health, geriatrics, or occupational therapy. Additional experiences in gerontological, psychogeriatric, or dementia care, as well as experiences as family caregivers, are described.

Data on implementation issues is available to varying degrees. Implementation strategies were described for 2 interventions [55,96]. The reported modifications of interventions are mainly related to adaptations to the needs of new target groups and to pandemic-related restrictions. Barriers to implementation are technical issues (eg, lack of hardware, limited software functionality or usability, limited internet access, and suboptimal infrastructure), lack of digital skills (of persons delivering and receiving counseling), security issues, and aspects resulting from the type of technology used in each case (eg, loss of context, lack of prompts from the surroundings, and not being able to follow up). Facilitating factors are special trainings (eg, technical support and conversation strategies), ongoing support (eg, regular team meetings and supervision), features of interventions (eg, customized to sociocultural preferences of target groups and undemanding technology), and commitment of management of implementing institutions. External conditions of implementation are described as established organizations in which intervention programs are embedded, collaborations with interest-holders in dementia care, and the SARS-CoV-2 pandemic.

A detailed description of the included intervention programs is provided in Multimedia Appendix 3. Intervention components that have been incorporated into the program theory and that are represented in the logic model are referred to in the section Logic Model and Program Theory.

Outcomes of caregivers or persons with dementia examined and the assessment instruments used in the included studies are listed in Table 1.

Qualitative Substudy

Sample Characteristics

We conducted 4 individual and 3 group interviews lasting from 1 to 3 hours (total duration: 12 hours 41 minutes) with 15 family caregivers (14 women and 1 man). One participant cared for a person with Parkinson disease; all other caregivers cared for a person with dementia. One family had a migration background; the caregiving arrangement no longer existed in 3 cases due to the death of the person receiving care.

In addition, 8 individual and 2 group interviews lasting from 50 to 100 minutes (total duration: 12 hours 14 minutes) with 12 counselors (11 women and 1 man) were conducted. Participants had been working as counselors for an average of 9 years and had qualifications in nursing, health and nursing science, social work, psychology, and administration. The providing institutions of participating counselors varied. Details on the characteristics of participants are provided in Table 2.

XSL•FO
RenderX

Table . Sociodemographic information on interview partners.

Characteristics	Values
Caregivers (n=15)	
Age in years, mean (range)	70 (44-83)
Gender, n	
Women	14
Man	1
Care-receiving person, n	
Husband or spouse	10
Mother	3
Father	1
Brother	1
Caregiving arrangement, n	
Home-based care	12
Institutionalized care	3
Years of caregiving, mean (range)	6 (1-20)
Counselors (n=12)	
Age in years, mean (range)	51 (33-61)
Gender, n	
Women	11
Man	1
Years of counseling, mean (range)	9 (1-30)
Providing organization, n	
Municipality	4
Welfare organization	2
Self-help organization	2
Research institution	1
University hospital	1
Registered association	1
Self-employed counselor	1

Themes Identified in Interviews With Caregivers

We identified the following 3 themes in the caregivers' statements: being affected, feeling insecure and helpless in the face of the health care system, and search for information and communicative exchange.

Being Affected

> *...and then I cried a lot and had really bad thoughts.* [Caregiver 15]

Participants' descriptions of how they are affected by the demands of the caregiving situation made up a large part of the interviews. The caregivers report conflicts with the persons receiving care encompassing harassment and insults, as well as verbal and physical attacks: "Back in November, my husband pushed me down the stairs" [Caregiver 15]. They also describe their grief over the loss of a person close to them due to the changes evoked by dementia, as well as the loss of life plans and perspectives. In addition, the demands of caregiving pose a burden on caregivers: "...that you just can't do it anymore" [Caregiver 15].

Feeling Insecure and Helpless in the Face of the Health Care System

> *And nobody tells you all this.* [Caregiver 15]

The interviewees state a lack of information and a lack of support from persons or institutions involved in health care such as primary care physicians or health insurances:

> *There's the 24-hour care, every single day. Yeah, and then if you spend the whole morning on the phone just trying to get a simple answer, well, that's just how it is for me.* [Caregiver 15]

High financial expenses often limit the use of support services.

XSL•FO
RenderX

Search for Information and Communicative Exchange

> *That made a huge difference for me, that I stopped thinking that I was somehow (...) weird.* [Caregivers 3-5]

Caregivers report beneficial experiences with empathetic counseling focused on problem-solving:

> *She [counselor] asks, "So, how are you doing?" Finally, someone actually asks how I'm handling all this. (...) And then she says, "Okay, so here's what you can do..." – and that's the kind of support that really does you good.* [Caregivers 12-14]

Participants are not always aware of the differing objectives of services such as counseling, support, or self-help groups; however, they emphasize the importance of continuity in support services:

> *It has to be done continuously. (...) Even though I was really impressed by my first two meetings here – and they were incredibly helpful – if I had stopped then, I wouldn't have this feeling now.* [Caregivers 3-5]

Interviewees indicate a preference for face-to-face meetings and express a reluctance to use ICT for counseling services: "You have to deal with all that tech stuff, and honestly, no – it's just too much for me" [Caregiver 15]. Participants, however, make use of familiar technologies to maintain ongoing contact with their counselors: "I have her (counselor) on (messenging service), and she texts too – like she'll wish me a nice weekend and ask if everything's okay or if something's going on" [Caregiver 15].

Themes Identified in Interviews With Counselors

We identified the following themes in counselors' interviews: work attitude and standards, unpredictability, expectations, working conditions, organizational influence, and tools: techniques and networking.

Work Attitude and Standards

> *Good counseling is when you understand the person in their (...) lifeworld (...) when they feel that their issues are taken seriously.* [Counselor 7]

The interviewees show a high level of commitment, appreciation, and empathy toward those seeking counseling. They describe how their work has fostered both personal and professional growth, and they highlight specific skills required for different counseling formats, such as anonymous helpline services. To further ensure the quality of counseling services, training and support offered to counselors are considered essential:

> *So, as the head of the Support Center for Family Caregivers, I really keep an eye on that. What do my colleagues need? And they get it. It's really important to me that the quality here is good.* [Counselor 6]

To establish a common understanding of counseling, conceptual frameworks or manuals are developed to document the shared values: "I'm gonna write down what the core things in our counseling are – what's really important to us" [Counselors 3 and 4]. Efforts are being made to reach out to family caregivers, for instance, by maintaining a presence in public spaces: "I've had this idea for a while – that we should have a place right in the middle of the city, like in a department store or a shopping center, somewhere people actually go" [Counselor 6]. This seems necessary, since family caregivers are often unable to initiate contact, according to the counselors:

> *(Caregiver) says, "Yeah, your address has been sitting here for two years before I even call." Happens all the time. All the time.* [Counselor 6]

Unpredictability

> *And that they [caregivers] come with such a mountain of problems.* [Counselor 12]

Study participants indicate that the concerns and the state of mind of the caregivers are not known before the counseling and cover a wide spectrum—from "calm" to "nervous breakdown." In some cases, counselees are overwhelmed and unable to speak:

> *Sometimes they just sit there and start to cry. Then there's something to drink, a tissue, and maybe even a piece of chocolate.* [Counselor 6]

Expectations

> *I wish they [caregivers] would open up so that they could really be helped properly.* [Counselor 10]

We found that counselors have various expectations of those seeking counseling. While some of the participants stated that they had no expectations at all, others expected caregivers' openness to share their concerns and experiences with the counselors. In some cases, counselors expect the caregivers to be willing to implement counselors' advice.

Working Conditions

> *Our doors are always open, and a colleague can come and say, "I have to tell you this."* [Counselor 6]

The description of the working conditions is a frequent theme in the interviews. The study participants highlight the importance of mutual support from colleagues as described in the headline quote. An important aspect is whether the participant is working in a team or as a "lone fighter":

> *And because there's no real exchange with others, you have the chance to think about things and maybe do them differently – but you can never really share it with anyone.* [Counselor 7]

Teams in which different professional groups are represented and whose members support each other are considered as an essential element in coping with the demands of counseling work:

> *We've got people from all sorts of backgrounds – social work, sociology, (...) nursing background (...) psychologist, (...) gerontology and even theology. (...) we're a really diverse team and can bring in all kinds of different perspectives.* [Counselors 3 and 4]

The theme also covers personnel and technical equipment, which varies greatly among participants:

> *I'd like to be able to scan some documents and bring them here to save people a trip. Maybe use a small*

XSL•FO
RenderX

> *mobile printer or something like that (...) but there's just no way.* [Counselor 9]

In addition, the opportunity to independently organize and design their everyday working life is considered a crucial factor of the working conditions.

Organizational Influence

> *Then they [superiors] would clearly say, "That's not your job!"* [Counselor 7]

The providing organization has a significant influence on counselors' performance in the sense of a spillover effect. Lack of agreement between managers and counselors on the purpose and the design of counseling services has a negative impact on its implementation.

Tools: Techniques and Networking

> *To pave the way for people.* [Counselor 9]

Study participants describe techniques they are using in their professional lives. These include conversation management, individual practices to ease the situation for counselees (eg, offering something to drink and asking questions that facilitate the dialogue), as well as assessment tools, documentation aids, and information materials. ICT is used to enhance access to counseling services:

> *If people live further away, that's totally normal. Or if they're working, they'll say, "Okay, I can chat with you for fifteen minutes on (videoconferencing software)." We're happy to do that.* [Counselor 6]

Professional networks are used to help caregivers with problems that fall outside the counselors' scope of responsibility (eg, medical and legal issues). Cooperation (rather than competition) between different service providers and the combination of various support services is seen as key to providing effective support:

> *Well, just counseling on its own or just training doesn't really work. It has to be a combination of different things – then it can work.* [Counselor 8]

Constitutive Pattern

The common ground in the lifeworld perceptions of the two interest-holder groups forms the constitutive pattern of *somebody to count on*. While caregivers express the need for *having somebody to count on*, who guides them empathically through the challenges of family dementia care, counselors aim at *being somebody to count on* by competently assisting caregivers to master those challenges.

Outcomes Identified by Interview Partners

Caregivers and counselors described effects, which could be used to measure effectiveness of counseling. The potential outcomes are listed in Table 3.

XSL•FO
RenderX

Table . Outcomes identified by interview partners (caregivers or counselors).

Outcomes	Identifying interview partner
Level of knowledge	Ca[a]; Co[b]
Knowledge about the disease and its (potential) course	Ca; Co
Knowledge about how to deal with changed behavior of the person with dementia	Ca; Co
Knowledge about what to do to improve well-being and safety of the person with dementia	Ca; Co
Knowledge about available support services	Ca; Co
Knowledge about legal regulations and financial support options	Co
Level of caregiver burden	Co
Level of caregiver distress	Co
Level of emotional strain	Co
Level of emotional overload	Co
Stability of caregiver perceived burden (level is maintained despite increasing care needs)	Co
Sense of relief	Ca; Co
Sense of relief through empathetic understanding	Ca; Co
Sense of relief from having a counselor to fall back on (reliability, accountability of counseling)	Ca; Co
Sense of relief through active support (eg, support in filling out application forms)	Ca; Co
Sense of order (plan in mind about what to do next or how to get along with caregiving)	Ca; Co
Caregiver depressive symptoms	Ca
Stability of caregiver perceived depressive symptoms (level is maintained despite increasing care needs)	Co
Caregiver guilt	Co
Feeling of helplessness	Ca
Quality of life (caregiver)	Co
Quality of life (person with dementia)	Co
Caregiver well-being	Co
Caregiver (physical or psychological) health	Co
Caregiver satisfaction	Co
Congruousness of caregiving arrangement with caregiver's preferences and wishes	Co
Caregiver attitude toward caregiving	Co
Sense of confidence	Co
Sense of confidence in decision-making	Co
Sense of confidence in managing caregiving challenges	Co
Caregiver mastery	Co
Empowerment (eg, in finding support by oneself)	Co
Feeling well advised	Co
Problem-solving ability	Ca; Co
Realization of counselors' suggestions (eg, changes in the home environment, use of support services, creation or use of personal space, and hobbies)	Co

XSL•FO
RenderX

Outcomes	Identifying interview partner
Balance between caregiving and self-care	Co
Recognition of personal limits in caregiving capacity	Co
Safety of care	Co
Stability of caregiving arrangement over time	Co
Number of support services used	Co
Duration of caregiving arrangement	Co
Time to transition to long-term care facility	Co
Extent and stability of caregiver's support network	Co
Number of conflicts with care-receiving person	Co
Cost of care	Co

[a]Ca: caregivers.

[b]Co: counselors.

Taken together, the statements of the two participant groups on effective counseling include the following characteristics: an empathetic, appreciative, and compassionate attitude; knowledge transfer; provision of materials and help with bureaucracy; providing a view from the "outside": classifying, weighting, and sorting caregiving issues; recommendations for action (eg, dealing with changes in behavior); support in decision-making and in defining boundaries; and sustainability and continuity. Comments and suggestions of participating caregivers and counselors that have been incorporated into the program theory and that are represented in the logic model are referred to in the section Logic Model and Program Theory.

Logic Model and Program Theory

The logic model of the technology-assisted counseling intervention is provided in Figure 2.

Figure 2. Logic model of the technology-assisted counseling intervention. AI: artificial intelligence; ICT: information and communication technology. *The theory of change, the outcomes chain, and the outcomes listed in Tables 1 and 3 will be subject to a future consensus-building process.

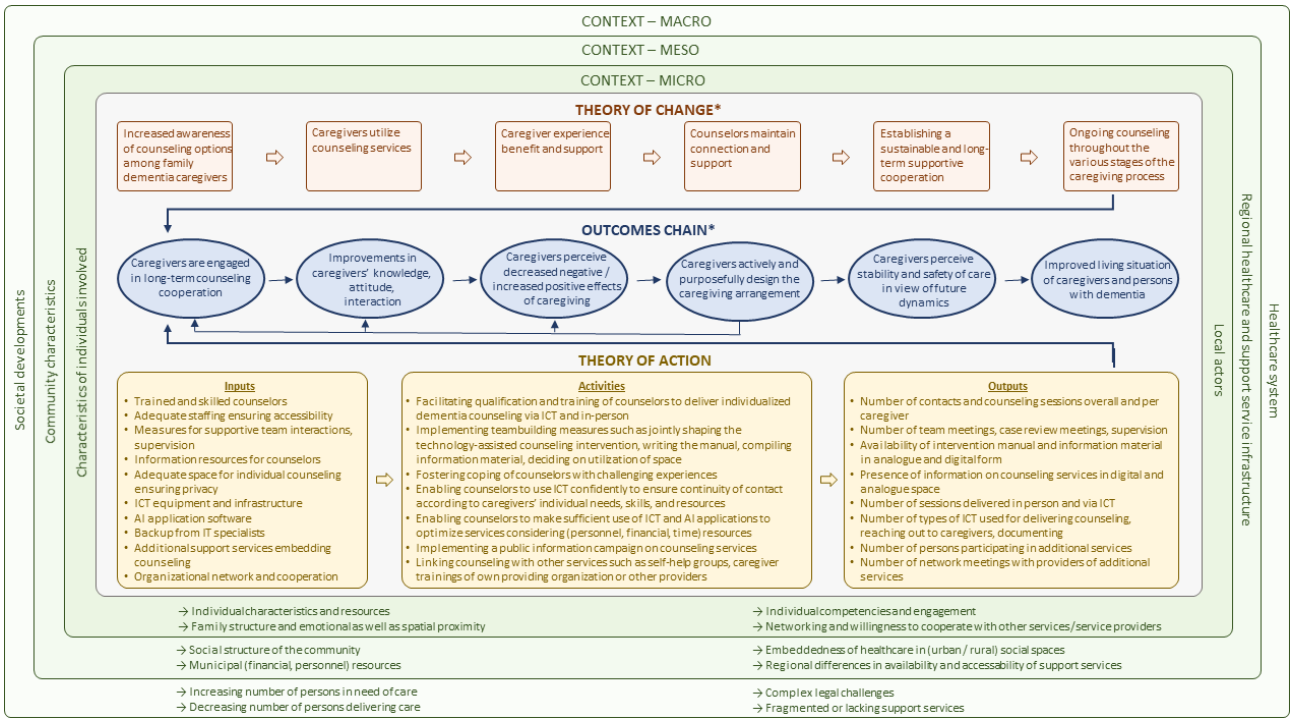

In the following section, we outline the program theory. The detailed version can be obtained from the authors on request.

The contextual factors of family dementia care in Germany on the macro-, meso-, and microlevels are illustrated in green in the logic model: Ongoing demographic developments lead to an imbalance between care needs and families' capacity to provide care. Challenges faced by individuals assuming caregiving responsibilities can be exacerbated by fragmented health care services and complex legal regulations. The unequal distribution of support services across communities with varying resources can result in inadequate support for family caregivers of persons with dementia.

The problem to be addressed is the inadequate use of counseling services by family caregivers of persons with dementia. Study

XSL•FO
RenderX

participants indicated that caregivers are rarely made aware of the counseling services available. Moreover, caregivers often do not actively seek contact with services because they feel overwhelmed by caregiving tasks, experience feelings of shame, or hesitate due to the perceived effort involved.

In our theory of change, we outline how a technology-assisted counseling intervention may help to address this issue: We hypothesize that an increased awareness of counseling services among family dementia caregivers leads to an initial use associated with a beneficial and supportive experience.

As both caregivers and counselors emphasize the importance of continuous counseling and support, we propose that counselors adopt unobtrusive strategies to maintain engagement with caregivers. Using diverse types of ICT that caregivers are familiar with facilitates the ongoing contact with counselors. This approach fosters a sustainable, long-term supportive cooperation that enables consistent counseling throughout all stages of the caregiving process.

In the outcomes chain, we describe the immediate and intermediate outcomes, which were derived by clustering outcomes identified in the literature and those mentioned by study participants (refer to Tables 1 and 3). We assume that long-term counseling cooperation leads to improved caregiver knowledge about caregiving aspects (eg, the (potential) course of dementia, how to deal with changed behavior of the person with dementia, and available support services), to an enhanced attitude toward caregiving, and to improved interaction with the care-receiving person. This entails an altered impact of caregiving, in that the caregivers perceive decreased negative effects such as burden, distress, feeling of helplessness, and increased positive effects such as sense of confidence, empowerment, and satisfaction. Caregivers are thereby enabled to actively and purposefully shape the caregiving arrangement, for example, by using support services and to achieve congruousness of the caregiving arrangement with their own preferences and values. Caregivers perceive stability of the caregiving arrangement and safety of care despite unpredictable developments and dynamics and are aware that they may approach counselors any time needed. This leads to an improved living situation for both the caregivers and the persons with dementia. As indicated by the arrows in the logic model, the levels of the outcomes chain can occur in a recursive process.

Within the theory of action, we specify the inputs and actions through which the theory of change can be initiated [17]. The inputs specified in the theory of action are derived from the statements of study participants presented above and from intervention components described in the literature (refer to Multimedia Appendix 3). As counselors emphasized the importance of adequate working conditions, we incorporated aspects such as sufficient staffing, measures to promote a supportive working environment, and access to informational resources, including libraries and databases. Appropriate technological equipment and infrastructure, along with IT support, enable the reliable delivery of technology-assisted counseling. Furthermore, networks for cooperation with complementary services—either within the organization or in collaboration with external providers—are crucial for providing effective support. Program activities are directed at enabling counselors to provide technology-assisted counseling to family dementia caregivers competently and confidently. The activities depicted in the logic model further promote supportive working relationships by fostering mutual support within the team and articulating shared values in their work, for example, through manuals or counseling concepts. The implementation of a public information campaign may help to raise awareness of counseling services and facilitate outreach to family caregivers. In addition, diverse services are integrated to ensure timely, individualized support. Considering limited personnel, financial, and time resources, counselors are empowered to use ICT and artificial intelligence (AI) applications sufficiently to optimize the efficiency of services. Outputs represent the tangible, measurable products of activities [17]. We included the availability of the intervention manual and the information material, as well as the number of persons using counseling or other services and the number of events such as meetings to enhance cooperation within the team, within the organization, and with other organizations, as well as counts of types of ICT.

Discussion

Principal Findings

In this paper, we present the results of a scoping review and a qualitative study synthesized into a logic model and a program theory of a technology-assisted counseling intervention for family caregivers of persons with dementia. Information obtained from literature, empirical findings from the qualitative substudy, and theoretical approaches have been merged to guide the future COS consensus process and the modeling of the intervention.

Theoretical Approaches and Implications of Evolving Technological Modalities Informing the Program Theory

Presumed mechanisms of the technology-assisted counseling intervention were informed by theoretical approaches underpinning the interventions identified in the scoping review: The model of determinants of subjective burden of caregivers of persons with dementia forms the theoretical framework of the Dementelcoach intervention [68]. Aspects such as caregivers' personal characteristics, material and social circumstances, and the support they receive are integrated into the program theory. In addition, the concept map of the Internet-Based Caregiver Support Service (ICSS) described by Chiu and Eysenbach [92] to conceptualize usage behavior of family caregivers by integrating 3 theoretical approaches (Anderson's model of health service utilization, Venkatesh's theory of technology acceptance, and Chatman's and Wilson's information behavior theories) was used to shape our theoretical understanding. Factors such as dynamic and individual caregiver needs, perceived efforts of ICT options, and preferences in using ICT [92] have been taken into account.

Further findings of the scoping review illustrate how technologies used for delivering counseling have evolved over time from technology-assisted counseling provided exclusively via telephone to counseling via chats or SMS text messaging.

XSL•FO
RenderX

In addition, multiple technologies are used in some interventions for establishing contact or delivering counseling. This diversification has led to a partial overlap in the categories formulated to group interventions in this paper.

The appropriateness of technological options for administering counseling varies with respect to availability, flexibility, and requirements for technological equipment and skills [3,8]. In addition, the suitability of the various technological approaches for discussing personal issues in depth or addressing specific requests directly is perceived differently [3,8]. Therefore, we will integrate multiple technological ways to access counseling in the development of the intervention to accommodate caregivers' individual needs and resources and to enhance technology acceptance.

The combination of various ICT systems poses higher demands on counselors' technological literacy and counseling competencies. To effectively switch between various modes of delivery, counselors require specific qualifications and skills, for example, to compensate for the lack of visual clues during telephone counseling, to facilitate openness in conversations via videoconferencing software, or to address counselees' concerns in written asynchronous communication such as email or chat [8]. To enable counselors to use ICT confidently, inputs of the intervention aim at promoting an effective working environment by including adequate technological equipment and infrastructure as well as available support by IT specialists for training counselors and troubleshooting.

Overcoming Utilization Barriers Through Integrated Access Modalities and Targeted Outreach

Caregivers and counselors participating in the qualitative study largely expressed a preference for in-person counseling. The reluctance to use ICT that we observed despite the shift toward technology-assisted services during the SARS-CoV-2 pandemic [110] could have been partly influenced by the older age of the participating caregivers. Nevertheless, Gonzalez-Fraile et al [3] state in their systematic review that remotely delivered training and support interventions appeared to be less acceptable than control interventions, as assessed by attrition rates. Therefore, we integrated face-to-face services offered at counseling centers or in caregivers' home environments into the program theory of the technology-assisted counseling intervention.

We hypothesize that the combination of diverse technologies in addition to in-person counseling contributes to overcoming barriers to utilization of counseling services. The different characteristics of technological options allow for the targeting of distinct user groups. Findings of the qualitative study indicate that SMS text messaging is rather highly accepted among middle-aged or older persons and can be used to maintain contact and emotional involvement by sharing photos or invitations to events. Results from the literature also show that SMS text messaging is suitable for addressing disparities in access to caregiving support for persons belonging to minorities [98,106]. The needs of ethnic groups may also be met by counseling via email, offering an alternative service model [99,100]. Participants of the qualitative substudy stated that nonnative speaking persons may benefit from conversations via email as the asynchronous communication gives them more time to consider counselors' suggestions.

As outlined above, caregivers described in detail the consequences arising from the caregiving responsibility summarized in the theme "being affected." Feeling overwhelmed and stressed can prevent caregivers from actively seeking information and support [111]. In addition, the lack of awareness as well as the regional lack of availability contributes to the underuse of support services such as counseling [111]. We therefore included a campaign in the program theory to increase awareness among public and professional communities about counseling opportunities. By reaching out through an enhanced digital and analog presence, we aim to encourage family dementia caregivers to initiate contact to counseling services.

Embedding Counseling Services for Maintaining Ongoing Contact

Inquiries about the caregivers' mental models during the interviews indicate that family dementia caregivers often do not distinguish between the different objectives of various support services. While professionals intend to give room for sharing caregiving experiences in support groups, for example, some of the participants also expected information to be conveyed at these meetings. Embedding counseling in a variety of support services may contribute to supplementing the supportive effect of individual interventions, to maintaining contact, and to actively offering support when need occurs. Interventions offering counseling as part of a comprehensive program with non-technology-assisted components provide examples of such an approach: One example combines telephone counseling with a group activity program for community-dwelling individuals with dementia, aiming to empower persons with dementia and their caregivers [97]. Another combines telephone coaching (Dementelcoach) with respite care, thereby enhancing the effectiveness of the intervention [68]. A third example links two eHealth caregiver interventions to existing Meeting Centers for persons with dementia and their family caregivers [63,96].

Integrating AI Technology to Mitigate the Shortage of Skilled Health Care Professionals

With care needs increasing due to the demographic development, the shortage of skilled health care workers results in higher workloads and accelerated turnover rates [112]. Various strategies are described for retaining health care staff [112]. Authors of the Dementelcoach study reported that professionals delivering the intervention were part-time workers in the psychogeriatric sector. These professionals were encouraged to expand their work hours to provide counseling in conjunction with their existing jobs [68]. Offering a complementary activity may help to meet existing support needs in the face of the shortage of skilled staff [68].

All participating counselors reported a high workload, and the growing scarcity of skilled professionals may further exacerbate work-related stress [113]. To reduce time spent on documentation and administrative tasks, AI technology will be integrated into the intervention at the level of partial automation by augmenting human performance [113]. In using large

XSL•FO
RenderX

language models for transcribing counseling sessions and composing summaries, counselors are relieved of documentation tasks and allowed to focus solely on counselees' concerns during sessions. AI can also be used for scheduling tasks and to compile lists of regionally available support services for caregivers based on an AI-augmented database [114]. AI-drafted content will be evaluated by counselors to ensure accuracy and quality of information. This may help to alleviate concerns associated with the use of AI in the context of vulnerable groups—in particular, concerns about the reliability of information, privacy, and data security [113].

In sum, technology can make an important contribution to overcoming problems arising from the increasing need for support due to the growing number of persons with dementia [6], the growing shortage of skilled health care professionals [115], and social and infrastructural inequalities [116]. It is essential that technology-assisted counseling interventions are designed with consideration of individual and situational requirements in order to enhance their acceptability and feasibility from the perspectives of both recipients and providers [117].

Strengths and Limitations

We established a profound database for formulating an initial program theory and preparing for the development of a COS by integrating knowledge obtained from literature and the lived experience of two interest-holder groups. Participants in the qualitative study were recruited all over Germany. Caregivers living in urban and rural regions shared their experiences with counseling services, but we did not succeed in recruiting a heterogeneous sample in terms of age and gender. This may impose limitations on findings, as perspectives of younger and diverse caregivers, who might be more technologically savvy, are underrepresented. Younger family caregivers who are also employed, particularly long-distance caregivers, may be more inclined to use ICT to access counseling services. By combining various technological access options with in-person counseling, we advocate for an inclusive approach that addresses the needs of diverse target groups.

We were able to identify commonalities and differences in the interest-holders' expectations and mental models, which were incorporated into the program theory.

Another limitation is the inclusion of only two groups of interest-holders. Due to the limited duration of the project, we were not able to expand the study population of the qualitative inquiry. Representatives of providing organizations and policymakers will be included in the consensus process of the COS.

We followed accepted methodological guidance [13,17] and integrated methodological approaches in an innovative way. This proved to be beneficial: When asked about potential outcomes for measuring effectiveness of counseling, interviewees reflected on and critically questioned their own ideas and expectations. We found it advantageous to consider adequate outcomes from the beginning of the development process, to debate underlying (hypothesized) causal relationships, and to visualize causal assumptions in the logic model. This theory-led approach to the development, implementation, and evaluation of a technology-assisted counseling intervention for family caregivers of persons with dementia is consistent with the recommendations of the updated UK Medical Research Council guidance [9].

Future Directions

We will perform a theory-led approach to modeling, implementing, and evaluating the technology-assisted counseling intervention for family dementia caregivers. In a first step, the "long list" of outcomes will be created in accordance with the COS development methodology [13]. The participants of the qualitative study found it surprisingly difficult to name potential outcomes for assessing the effectiveness of counseling interventions, which led to a high number of (partially) overlapping outcomes. In a follow-up study, we will involve representatives of interest-holder groups such as family caregivers, counselors, managers of provider organizations, and researchers to integrate outcomes reported in literature (Table 1) and outcomes identified by interview partners (Table 3): Redundant outcomes will be removed; for example, outcomes such as "burden," "guilt," and "stress or distress" were identified in both the literature and interviews. The remaining outcomes will then be grouped into outcome domains applying ontologies, as proposed by Williamson et al [13]. The resulting "long list" of clustered outcomes forms the basis for the consensus process of the COS [13]. We will apply the Delphi approach involving groups of relevant interest-holders to determine important clinical outcomes [13]. To establish a shared understanding among participating interest-holders, elements of the program theory will be gradually integrated into the consensus process. At the outset of this process, we will use the logic model to visualize and critically discuss the theory of change. Subsequently, the importance of individual outcomes will be rated in Delphi rounds and finally consented in a consensus conference [13].

The refined logic model and program theory will then be used for modeling the intervention in collaboration with practice partners in a future research project. The logic model will serve to foster a shared understanding among the individuals involved and to systematically document the adaptation of key components to the specific context [9,18].

Conclusions

We developed an initial program theory of a technology-assisted counseling intervention for family caregivers of persons with dementia by introducing a methodological innovation. Findings obtained from interest-holder groups and literature are synthesized into a program theory and visualized by a logic model. We have also compiled a comprehensive list of potential outcomes, which includes the outcomes examined in clinical studies and those that are relevant from the perspective of interest-holders. This enables the consensus process for finalizing the COS for technology-assisted counseling interventions.

These results will inform the theory-led modeling, implementation, and evaluation of the intervention, which will include a customized ICT package. This package has the

XSL•FO
RenderX

potential to improve accessibility to counseling for caregivers by overcoming disparities in access to health care services. In addition, the design of the intervention can positively impact work conditions for health care professionals delivering support and improve the efficiency of services.

Acknowledgments

We thank the caregivers and counselors who participated in the qualitative study for devoting their time and sharing their experiences. In addition, we thank the members of the study advisory board for their valuable contributions and feedback. We also acknowledge the financial support of the Open Access Publication Fund of Martin Luther University Halle-Wittenberg.

Funding

This work was supported by the German Research Foundation (DFG, project number 533628714). The funding source had no role in the design and conduct of the study and preparation of the manuscript.

Data Availability

Data extracted from studies included in the scoping review are included in Multimedia Appendix 3.

Authors' Contributions

Conceptualization: DB, GM
Formal analysis: DB
Funding acquisition: DB, GM
Investigation: DB
Supervision: GM
Validation: RH, JW, GM, AB
Writing – original draft: DB
Writing – review & editing: GM, AB, JW, RH

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies.
[DOCX File, 52 KB - jmir_v28i1e81669_app1.docx]

Multimedia Appendix 2
Interview guide.
[DOCX File, 39 KB - jmir_v28i1e81669_app2.docx]

Multimedia Appendix 3
Description of reports and interventions.
[DOCX File, 206 KB - jmir_v28i1e81669_app3.docx]

Checklist 1
PRISMA-ScR, SRQR, and COS-STAR checklists.
[DOCX File, 72 KB - jmir_v28i1e81669_app4.docx]

References

1. Ferrero-Sereno P, Mendoza-Muñoz M, Palomo-López P, et al. A systematic review of the effectiveness of technological interventions for caregivers of people with dementia: effects on quality of life and psychoemotional variables. Front Public Health 2025;13:1579239. [doi: 10.3389/fpubh.2025.1579239] [Medline: 40438051]
2. Saragih ID, Tonapa SI, Porta CM, Lee BO. Effects of telehealth intervention for people with dementia and their carers: a systematic review and meta‐analysis of randomized controlled studies. J Nurs Scholarsh 2022 Nov;54(6):704-719. [doi: 10.1111/jnu.12797] [Medline: 35769007]
3. González-Fraile E, Ballesteros J, Rueda JR, Santos-Zorrozúa B, Solà I, McCleery J. Remotely delivered information, training and support for informal caregivers of people with dementia. Cochrane Database Syst Rev 2021 Jan 4;1(1):CD006440. [doi: 10.1002/14651858.CD006440.pub3] [Medline: 33417236]

XSL•FO
RenderX

4. Gilsenan J, Gorman C, Shevlin M. Explaining caregiver burden in a large sample of UK dementia caregivers: the role of contextual factors, behavioural problems, psychological resilience, and anticipatory grief. Aging Ment Health 2023;27(7):1274-1281. [doi: 10.1080/13607863.2022.2102138] [Medline: 35881027]
5. Lindeza P, Rodrigues M, Costa J, Guerreiro M, Rosa MM. Impact of dementia on informal care: a systematic review of family caregivers' perceptions. BMJ Support Palliat Care 2020 Oct 14. [doi: 10.1136/bmjspcare-2020-002242] [Medline: 33055092]
6. GBD 2019 Dementia Forecasting Collaborators. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. Lancet Public Health 2022 Feb;7(2):e105-e125. [doi: 10.1016/S2468-2667(21)00249-8] [Medline: 34998485]
7. Butcher HK, Bulechek GM, Dochterman JM, Wagner CM, editors. Nursing Interventions Classification (NIC), 7th edition: Elsevier; 2018.
8. Bauernschmidt D, Wittmann J, Hirt J, Meyer G, Bieber A. The implementation success of technology-based counseling in dementia care: scoping review. JMIR Aging 2024 Jan 25;7:e51544. [doi: 10.2196/51544] [Medline: 38271050]
9. Skivington K, Matthews L, Simpson SA, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061. [doi: 10.1136/bmj.n2061] [Medline: 34593508]
10. Cheng JY, Nurul S, Cheng LJ, He HG. Effectiveness of technology-delivered psychosocial interventions for family caregivers of patients with dementia: a systematic review, meta-analysis and meta-regression. Int J Ment Health Nurs 2024 Dec;33(6):1796-1816. [doi: 10.1111/inm.13390] [Medline: 39034437]
11. Bauernschmidt D, Hirt J, Langer G, et al. Technology-based counselling for people with dementia and their informal carers: a systematic review and meta-analysis. J Alzheimers Dis 2023;93(3):891-906. [doi: 10.3233/JAD-221194] [Medline: 37125549]
12. Bauernschmidt D, Wittmann J, Bieber A, Meyer G. Integrating programme theory into the development of a core outcome set for technology-assisted counselling interventions in dementia: study protocol of the ProCOS study. BMJ Open 2024 Aug 6;14(8):e081526. [doi: 10.1136/bmjopen-2023-081526] [Medline: 39107024]
13. Williamson PR, Altman DG, Bagley H, et al. The COMET Handbook: version 1.0. Trials 2017 Jun 20;18(Suppl 3):280. [doi: 10.1186/s13063-017-1978-4] [Medline: 28681707]
14. Core Outcome Measures in Effectiveness Trials. COMET Database. URL: https://www.comet-initiative.org/Studies [accessed 2025-12-20]
15. Reilly ST, Harding AJE, Morbey H, et al. What is important to people with dementia living at home? A set of core outcome items for use in the evaluation of non-pharmacological community-based health and social care interventions. Age Ageing 2020 Jul 1;49(4):664-671. [doi: 10.1093/ageing/afaa015]
16. Dementia outcome measures: charting new territory; report of a JPND Working Group on Longitudinal Cohorts. : Neurodegenerative Disease Research; 2015 Oct URL: https://www.neurodegenerationresearch.eu/wp-content/uploads/2015/10/JPND-Report-Fountain.pdf [accessed 2025-12-20]
17. Funnell S, Rogers PJ. Purposeful Program Theory: Effective Use of Theories of Change and Logic Models: Jossey-Bass; 2011.
18. Skivington K, Matthews L, Simpson SA, et al. Framework for the development and evaluation of complex interventions: gap analysis, workshop and consultation-informed update. Health Technol Assess 2021 Sep;25(57):1-132. [doi: 10.3310/hta25570] [Medline: 34590577]
19. Anderson LM, Petticrew M, Rehfuess E, et al. Using logic models to capture complexity in systematic reviews. Res Synth Methods 2011 Mar;2(1):33-42. [doi: 10.1002/jrsm.32] [Medline: 26061598]
20. Baxter SK, Blank L, Woods HB, Payne N, Rimmer M, Goyder E. Using logic model methods in systematic review synthesis: describing complex pathways in referral management interventions. BMC Med Res Methodol 2014 May 10;14:62. [doi: 10.1186/1471-2288-14-62] [Medline: 24885751]
21. Bleijenberg N, de Man-van Ginkel JM, Trappenburg JCA, et al. Increasing value and reducing waste by optimizing the development of complex interventions: enriching the development phase of the Medical Research Council (MRC) framework. Int J Nurs Stud 2018 Mar;79:86-93. [doi: 10.1016/j.ijnurstu.2017.12.001] [Medline: 29220738]
22. Hirt J, Langer G, Wilde F, Bauernschmidt D, Meyer G, Bieber A. Technology-based counselling in dementia (TeCoDem): study protocol of a mixed-methods systematic review with qualitative comparative analysis and meta-analysis. BMJ Open 2021 Dec 8;11(12):e054157. [doi: 10.1136/bmjopen-2021-054157] [Medline: 34880025]
23. Core Outcome Measures in Effectiveness Trials. Registration ProCOS. URL: https://www.comet-initiative.org/Studies/Details/2884 [accessed 2025-12-20]
24. Anguera MT, Blanco-Villaseñor A, Losada JL, Sánchez-Algarra P, Onwuegbuzie AJ. Revisiting the difference between mixed methods and multimethods: is it all in the name? Qual Quant 2018 Nov;52(6):2757-2770. [doi: 10.1007/s11135-018-0700-2]
25. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
26. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]

XSL•FO
RenderX

27. Kirkham JJ, Gorst S, Altman DG, et al. Core Outcome Set—Standards for Reporting: the COS-STAR statement. PLoS Med 2016 Oct;13(10):e1002148. [doi: 10.1371/journal.pmed.1002148] [Medline: 27755541]
28. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evidence Synthesis 2020;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167]
29. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev 2016 Dec 5;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
30. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: Template for Intervention Description and Replication (TIDieR) checklist and guide. BMJ 2014 Mar 7;348:g1687. [doi: 10.1136/bmj.g1687] [Medline: 24609605]
31. Möhler R, Köpke S, Meyer G. Criteria for Reporting the Development and Evaluation of Complex Interventions in healthcare: revised guideline (CReDECI 2). Trials 2015 Dec;16(1). [doi: 10.1186/s13063-015-0709-y]
32. Zahavi D, Martiny KMM. Phenomenology in nursing studies: new perspectives. Int J Nurs Stud 2019 May;93:155-162. [doi: 10.1016/j.ijnurstu.2019.01.014] [Medline: 30795899]
33. Akremi L. Stichprobenziehung in der qualitativen Sozialforschung. In: Baur N, Blasius J, editors. Handbuch Methoden Der Empirischen Sozialforschung: Springer; 2022:405-424. [doi: 10.1007/978-3-658-37985-8_26]
34. Schreier M. Fallauswahl. In: Mey G, Mruck K, editors. Handbuch Qualitative Forschung in der Psychologie: Band 2: Designs und Verfahren: Springer; 2020:19-39. [doi: 10.1007/978-3-658-26887-9]
35. Rahimi S, Khatooni M. Saturation in qualitative research: an evolutionary concept analysis. Int J Nurs Stud Adv 2024 Jun;6(100174):100174. [doi: 10.1016/j.ijnsa.2024.100174] [Medline: 38746797]
36. Helfferich C. Leitfaden- und Experteninterviews. In: Baur N, Blasius J, editors. Handbuch Methoden der empirischen Sozialforschung: Springer; 2022:875-892. [doi: 10.1007/978-3-658-37985-8_55]
37. Audiotranskription. Audiotranskription. URL: https://www.audiotranskription.de [accessed 2025-12-20]
38. Diekelmann NL. Learning-as-testing: a Heideggerian hermeneutical analysis of the lived experiences of students and teachers in nursing. ANS Adv Nurs Sci 1992 Mar;14(3):72-83. [doi: 10.1097/00012272-199203000-00010] [Medline: 1550334]
39. Morbey H, Harding AJE, Swarbrick C, et al. Involving people living with dementia in research: an accessible modified Delphi survey for core outcome set development. Trials 2019 Jan 6;20(1):12. [doi: 10.1186/s13063-018-3069-6] [Medline: 30612587]
40. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. [doi: 10.1136/bmj.n71]
41. Hodgson NA, Petrovsky DV, Finegan K, Kallmyer BA, Pike J, Fazio S. One call makes a difference: an evaluation of the Alzheimer's Association National Helpline on dementia caregiver outcomes. Patient Educ Couns 2021 Apr;104(4):896-902. [doi: 10.1016/j.pec.2020.09.026] [Medline: 33004235]
42. Coyne AC, Potenza M, Nose MAB. Caregiving and dementia: the impact of telephone helpline services. Am J Alzheimers Dis (Columbia) 1995 Jul;10(4):27-32. [doi: 10.1177/153331759501000407]
43. Tremont G, Davis JD, Bishop DS, Fortinsky RH. Telephone-delivered psychosocial intervention reduces burden in dementia caregivers. Dementia (London) 2008;7(4):503-520. [doi: 10.1177/1471301208096632] [Medline: 20228893]
44. Telephone support for dementia caregivers: NCT00735800. National Institutes of Health. 2008. URL: https://clinicaltrials.gov/ct2/show/NCT00735800 [accessed 2025-12-20]
45. Tremont G, Davis J, O'Connor K, et al. Relationship between expectancy/credibility and early response to telephone-based dementia caregiver interventions. Alzheimer's & Dementia 2011 Jul;7(4S_Part_12):S435. [doi: 10.1016/j.jalz.2011.05.1257]
46. Tremont G, Davis J, Grover C, et al. Randomized controlled trial of a telephone‐delivered intervention (FITT‐Caregiver) for dementia caregivers. Alzheimer's & Dementia 2013 Jul;9(4S_Part_8):324-P325. [doi: 10.1016/j.jalz.2013.04.161]
47. Tremont G, Davis JD, Papandonatos GD, et al. A telephone intervention for dementia caregivers: background, design, and baseline characteristics. Contemp Clin Trials 2013 Nov;36(2):338-347. [doi: 10.1016/j.cct.2013.07.011] [Medline: 23916916]
48. Tremont G, Davis J, Bryant K, et al. Effect of a telephone‐based dementia caregiver intervention on use of community support services and health care resources. Alzheimer's & Dementia 2014 Jul;10(4S_Part_3):226-P227. [doi: 10.1016/j.jalz.2014.04.319]
49. Tremont G, Davis JD, Papandonatos GD, et al. Psychosocial telephone intervention for dementia caregivers: a randomized, controlled trial. Alzheimers Dement 2015 May;11(5):541-548. [doi: 10.1016/j.jalz.2014.05.1752] [Medline: 25074341]
50. Tremont G, Davis JD, Ott BR, et al. Randomized trial of the Family Intervention: Telephone Tracking-Caregiver for dementia caregivers: use of community and healthcare resources. J Am Geriatr Soc 2017 May;65(5):924-930. [doi: 10.1111/jgs.14684] [Medline: 28008609]
51. Does telehealth delivery of the COPE program provide a non-inferior alternative to face-to-face treatment for community dwelling people with cognitive impairment?: ACTRN12617000117314. Australian New Zealand Clinical Trials Registry. 2017. URL: https://www.anzctr.org.au/Trial/Registration/TrialReview.aspx?id=372140&isReview=true [accessed 2025-12-20]
52. Laver K, Liu E, Clemson L, et al. Does telehealth delivery of a dyadic dementia care program provide a noninferior alternative to face-to-face delivery of the same program? A randomized, controlled trial. Am J Geriatr Psychiatry 2020 Jun;28(6):673-682. [doi: 10.1016/j.jagp.2020.02.009] [Medline: 32234275]

XSL•FO
RenderX

53. Laver K. Lessons learned from offering a telehealth intervention for people with dementia and their caregivers. Alzheimers Dement 2021;17(S7):e051024. [doi: 10.1002/alz.051024]
54. The Residential Care Transition Module (RCTM): NCT02915939. National Institutes of Health. URL: https://clinicaltrials.gov/ct2/show/NCT02915939 [accessed 2025-12-20]
55. Gaugler JE, Statz TL, Birkeland RW, et al. The Residential Care Transition Module: a single-blinded randomized controlled evaluation of a telehealth support intervention for family caregivers of persons with dementia living in residential long-term care. BMC Geriatr 2020 Apr 15;20(1):133. [doi: 10.1186/s12877-020-01542-7] [Medline: 32293314]
56. Statz TL, Peterson CM, Birkeland RW, et al. “We moved her too soon”: navigating guilt among adult child and spousal caregivers of persons living with dementia following a move into residential long-term care. Couple Family Psychol 2022 Dec;11(4):300-314. [doi: 10.1037/cfp0000150] [Medline: 36743783]
57. Zmora R, Statz TL, Birkeland RW, et al. Transitioning to long-term care: family caregiver experiences of dementia, communities, and counseling. J Aging Health 2021 Jan;33(1-2):133-146. [doi: 10.1177/0898264320963588] [Medline: 32990494]
58. Zmora R, Statz TL, Birkeland RW, et al. Corrigendum: transitioning to long-term care: family caregiver experiences of dementia, communities, and counseling. J Aging Health 2021 Jan;33(1-2):133-146. [doi: 10.1177/0898264320963588]
59. Albers EA, Birkeland RW, Louwagie KW, et al. A qualitative analysis of mechanisms of benefit in the residential care transition module: a telehealth intervention for caregivers of relatives with dementia living in residential long-term care. Inquiry 2023;60. [doi: 10.1177/00469580231217981] [Medline: 38142369]
60. Gaugler JE, Birkeland RW, Albers EA, et al. Efficacy of the residential care transition module: a telehealth intervention for dementia family caregivers of relatives living in residential long-term care settings. Psychol Aging 2024 Aug;39(5):565-577. [doi: 10.1037/pag0000820] [Medline: 38753405]
61. Meeting Centers 3.0; Evaluation of the implementation and costeffectiveness of three new tailored interventions for people with dementia and their carers: dementalent, dementelcoach and STAR-online training: NTR5521. Netherlands Trial Register. URL: https://www.trialregister.nl/trial/5396 [accessed 2025-12-20]
62. Droes RM, Rijn A, Bosmans J, Meiland F. The individualized Meeting Centers Support Program; evaluation of benefits and costs [Abstract]. In: Book of Abstracts 19th IPA International Congress 2019, Vol. 31. [doi: 10.1017/S1041610219001339]
63. Dröes RM, van Rijn A, Rus E, Dacier S, Meiland F. Utilization, effect, and benefit of the individualized Meeting Centers Support Program for people with dementia and caregivers. Clin Interv Aging 2019;14:1527-1553. [doi: 10.2147/CIA.S212852] [Medline: 31692559]
64. Can the residential care transition module improve the psychological health of family carers of people with dementia during the residential care placement process in Australia?: ACTRN12621001462875. Australian New Zealand Clinical Trials Registry. 2021. URL: https://anzctr.org.au/Trial/Registration/TrialReview.aspx?id=382745&isReview=true [accessed 2025-12-20]
65. Brooks D, Wyles K, Pachana NA, Beattie E, Gaugler JE. Tailored videoconferencing counselling program to support family carers of people living with dementia during the transition to permanent residential care: a pilot and feasibility randomised trial. BMC Geriatr 2024 Apr 26;24(1):375. [doi: 10.1186/s12877-024-04907-4] [Medline: 38671367]
66. Natale G, Zigoura E, Carpaneto V, et al. Telephone calls provide effective support for most caregivers of patients with dementia but not for all. Int J Geriatr Psychiatry 2012 Feb;27(2):215-216. [doi: 10.1002/gps.2710] [Medline: 22223146]
67. Mierlo L, Meiland F, Droes R. Dementelcoach: effect of telephone coaching on informal and professional carers of community dwelling people with dementia [Poster]. Int Psychogeriatr 2011;23:378-379. [doi: 10.1017/S1041610211001293]
68. van Mierlo LD, Meiland FJM, Dröes RM. Dementelcoach: effect of telephone coaching on carers of community-dwelling people with dementia. Int Psychogeriatr 2012 Feb;24(2):212-222. [doi: 10.1017/S1041610211001827] [Medline: 21995966]
69. Rice JD, Sperling SA, Brown DS, Mittleman MS, Manning CA. Evaluating the efficacy of TeleFAMILIES: a telehealth intervention for caregivers of community-dwelling people with dementia. Aging Ment Health 2022 Aug;26(8):1613-1619. [doi: 10.1080/13607863.2021.1935462] [Medline: 34125635]
70. Wilkinson S. An analysis and evaluation of the admiral nurse dementia helpline. : Dementia UK; 2016.
71. Gilliard J, Keady J, Evers C, Milton S. Telephone helplines for people with dementia. Int J Geriatr Psychiatry 1998;13(10):734-735. [doi: 10.1002/(SICI)1099-1166(1998100)13:10<734::AID-GPS846>3.0.CO;2-X] [Medline: 9818311]
72. Jansen S. Fünf Jahre Alzheimer-Telefon der Deutschen Alzheimer Gesellschaft. Nervenheilkunde 2007;26(08):685-689. [doi: 10.1055/s-0038-1626915]
73. Pendergrass A, Weiß S, Gräßel E. Zu welchen Themen suchen Angehörige von Demenzbetroffenen telefonische oder emailbasierte Beratung auf? Aktuelle Ergebnisse des bundesweiten Beratungsangebots der Deutschen Alzheimer Gesellschaft e.V. Gesundheitswesen 2019 Dec;81(12):1018-1021. [doi: 10.1055/a-0586-8809] [Medline: 29672813]
74. Kurz A, Fischer IM, Dogan V, Kurz C. Telefonische Hilfe zur Selbsthilfe: Das Alzheimer-Telefon aus der Sicht der Nutzer:innen. Ger Med Sci 2024. [doi: 10.3205/000331] [Medline: 38883339]
75. Harvey R, Roques PK, Fox NC, Rossor MN. CANDID - Counselling and Diagnosis in Dementia: a national telemedicine service supporting the care of younger patients with dementia. Int J Geriatr Psychiatry 1998;13(6):381-388. [doi: 10.1002/(sici)1099-1166(199806)13:63.0.co;2-m]

XSL•FO
RenderX

76. Silverstein NM, Kennedy K, McCormick D. A telephone helpline for Alzheimer 's disease: Information, referral, and support. American Journal of Alzheimer's Care and Related Disorders & Research 1993 Sep;8(5):28-36. [doi: 10.1177/153331759300800507]
77. Rozani V, Pilko A, Kagan I. Provision of integrated care by a National Call Support Program for Alzheimer's patients and their caregivers. Int J Integr Care 2022;22(S3):91. [doi: 10.5334/ijic.ICIC22294]
78. Nakano Y, Hishikawa N, Sakamoto K, et al. A unique telephone support system for dementia patients and their caregivers managed in Japan (Okayama Dementia Call Center, ODCC). Neurology & Clinical Neurosc 2018 Jul;6(4):100-103 [FREE Full text] [doi: 10.1111/ncn3.12200]
79. Tousi B, Kanetsky C, Udelson N. ALZ i-Connect. Am J Alzheimers Dis Other Demen 2017 Feb;32(1):63-66. [doi: 10.1177/1533317516677615] [Medline: 27899432]
80. Kelly B. Link2Care: internet-based information and support for caregivers. Generations 2003;27(4):87-88 [FREE Full text]
81. Rentz M, Hoene A. Online coaching for caregivers: using technology to provide support and information. Alzheimers Care 2010;11(3):206-209 [FREE Full text] [doi: 10.1097/ACQ.0b013e3181ebc878]
82. De Cola MC, De Luca R, Bramanti A, Bertè F, Bramanti P, Calabrò RS. Tele-health services for the elderly: a novel southern Italy family needs-oriented model. J Telemed Telecare 2016 Sep;22(6):356-362. [doi: 10.1177/1357633X15604290] [Medline: 26377125]
83. Wheat H, Griffiths S, Gude A, et al. Practitioners' ability to remotely develop understanding for personalised care and support planning: a thematic analysis of multiple data sources from the feasibility phase of the Dementia Personalised Care Team (D-PACT) intervention. Dementia (London) 2023 Oct;22(7):1461-1486. [doi: 10.1177/14713012231185281] [Medline: 37354084]
84. Chang BL, Nitta S, Carter PA, Markham YK. Perceived helpfulness of telephone calls. J Gerontol Nurs 2004 Sep;30(9):14-21. [doi: 10.3928/0098-9134-20040901-05]
85. Salfi J. Seeking to understand telephone support for dementia caregivers: a qualitative case study [Dissertation]. : McMaster University (Canada); 2004 Sep URL: https://prod-ms-be.lib.mcmaster.ca/server/api/core/bitstreams/3c270094-5285-4532-a851-a9d4b77a3ba0/content [accessed 2025-12-20]
86. Salfi J, Ploeg J, Black ME. Seeking to understand telephone support for dementia caregivers. West J Nurs Res 2005 Oct;27(6):701-721. [doi: 10.1177/0193945905276882] [Medline: 16157943]
87. Spilsbury K. Telephone support met the perceived needs of dementia caregivers for convenient access to information, referral, and emotional support. Evid Based Nurs 2006 Jul;9(3):94. [doi: 10.1136/ebn.9.3.94] [Medline: 16865845]
88. Madden G, Rose T, Crystal L. Using video consultations to support family carers of people living with dementia. Nurs Older People 2022 Feb 1;34(1):28-33. [doi: 10.7748/nop.2021.e1346] [Medline: 34431259]
89. Dorell Å, Konradsen H, Kallström AP, Kabir ZN. "A friend during troubled times": experiences of family caregivers to persons with dementia when receiving professional support via a mobile app. PLoS ONE 2022;17(8):e0271972. [doi: 10.1371/journal.pone.0271972] [Medline: 35917295]
90. Chiu T, Lottridge D. Development and iterative refinement of an internet-based service for Chinese family caregivers of people with Alzheimer disease. AMIA Annu Symp Proc 2005;2005:919. [Medline: 16779206]
91. Chiu TML, Marziali E, Tang M, Colantonio A, Carswell A. Client-centred concepts in a personalized e-mail support intervention designed for Chinese caregivers of family members with dementia: a qualitative study. Hong Kong J Occup Ther 2010 Dec;20(2):87-93. [doi: 10.1016/S1569-1861(11)70008-0]
92. Chiu TML, Eysenbach G. Theorizing the health service usage behavior of family caregivers: a qualitative study of an internet-based intervention. Int J Med Inform 2011 Nov;80(11):754-764. [doi: 10.1016/j.ijmedinf.2011.08.010]
93. Brown P, Oliver E, Dening KH. Supporting family carers via the Admiral Nurse Dementia Helpline: reflection on a case study. Nurs Older People 2020 Sep 22;32(5):16-20. [doi: 10.7748/nop.2020.e1248] [Medline: 32400141]
94. Drayton S, Dening KH. Achieving positive outcomes in complex cases: the Admiral Nurse Dementia Helpline (Innovative Practice). Dementia (London) 2020 May;19(4):1308-1315. [doi: 10.1177/1471301217740005] [Medline: 29132219]
95. Sabat SR. Flourishing of the self while caregiving for a person with dementia: a case study of education, counseling, and psychosocial support via email. Dementia (London) 2011 Feb;10(1):81-97. [doi: 10.1177/1471301210392986]
96. van Rijn A, Meiland F, Dröes RM. Linking two new e-health caregiver interventions to meeting centres for people with dementia and their carers; a process evaluation. Aging Ment Health 2020 Aug;24(8):1316-1325. [doi: 10.1080/13607863.2019.1617243] [Medline: 31119946]
97. Nomura M, Makimoto K, Kato M, et al. Empowering older people with early dementia and family caregivers: a participatory action research study. Int J Nurs Stud 2009 Apr;46(4):431-441. [doi: 10.1016/j.ijnurstu.2007.09.009] [Medline: 17983619]
98. Perales-Puchalt J, Acosta-Rullán M, Ramírez-Mantilla M, et al. A text messaging intervention to support Latinx family caregivers of individuals with dementia (CuidaTEXT): development and usability study. JMIR Aging 2022;5(2):e35625. [doi: 10.2196/35625]
99. Chiu TML. Usage and non-usage behaviour of ehealth services among Chinese Canadians caring for a family member with dementia [Dissertation]. : University of Toronto; 2008 Jul URL: https://utoronto.scholaris.ca/items/1ef09030-a269-4a8c-9950-7e85019059d3 [accessed 2025-12-20]

XSL•FO
RenderX

100. Chiu T, Marziali E, Colantonio A, et al. Internet-based caregiver support for Chinese Canadians taking care of a family member with Alzheimer disease and related dementia. Can J Aging 2009 Dec;28(4):323-336. [doi: 10.1017/S0714980809990158]
101. Chiu TML, Eysenbach G. Stages of use: consideration, initiation, utilization, and outcomes of an internet-mediated intervention. BMC Med Inform Decis Mak 2010 Nov 23;10(1):73. [doi: 10.1186/1472-6947-10-73] [Medline: 21092275]
102. Barbabella F, Poli A, Andréasson F, et al. A web-based psychosocial intervention for family caregivers of older people: results from a mixed-methods study in three European countries. JMIR Res Protoc 2016 Oct 6;5(4):e196. [doi: 10.2196/resprot.5847] [Medline: 27713113]
103. Barbabella F, Poli A, Hanson E, et al. Usage and usability of a web-based program for family caregivers of older people in three European countries: a mixed-methods evaluation. Comput Inform Nurs 2018 May;36(5):232-241. [doi: 10.1097/CIN.0000000000000422] [Medline: 29505433]
104. CuidaTXT: a text message dementia-caregiver intervention for Latinos: NCT04316104. National Institutes of Health. 2020. URL: https://clinicaltrials.gov/study/NCT04316104 [accessed 2025-12-20]
105. Puchalt JP. CuidaTXT: a text message dementia-caregiver intervention for Latinos. Innov Aging 2020;4(Suppl 1):769. [doi: 10.1093/geroni/igaa057.2777]
106. Perales-Puchalt J, Ramírez-Mantilla M, Fracachán-Cabrera M, et al. A text message intervention to support Latino dementia family caregivers (CuidaTEXT): feasibility study. Clin Gerontol 2024;47(1):50-65. [doi: 10.1080/07317115.2022.2137449] [Medline: 36268684]
107. Lamura G, Poli A, Yghemonos S, Barbabella F. InformCare: the European information hub on family care. International Journal of Care and Caring 2017;1(3):409-413. [doi: 10.1332/239788217X15018371295074]
108. Brown P, Oliver E, Dening KH. Increasing need for telehealth services for families affected by dementia as a result of Covid-19. J Community Nurs 2020;34(5):59-64 [FREE Full text]
109. Twaddle IKB, Hattori-Uchima MP, Orallo RG, Gutierrez NJ. Telehealth outreach programming in the Pacific Island of Guam: providing access to dementia care support services during the COVID-19 pandemic. Alzheimers Dement 2021 Dec;17 Suppl 8(Suppl 8):e050134. [doi: 10.1002/alz.050134] [Medline: 34971289]
110. Caprioli T, Mason S, Tetlow H, Limbert S, Reilly S, Giebel C. “Necessity is the mother of invention”: experiences of accessing and delivering dementia-related support services by information communication technology during the pandemic in the UK. Dementia (London) 2025 Feb;24(2):323-343. [doi: 10.1177/14713012241272906] [Medline: 39117353]
111. Sorrentino M, Fiorilla C, Mercogliano M, et al. Barriers for access and utilization of dementia care services in Europe: a systematic review. BMC Geriatr 2025;25(1). [doi: 10.1186/s12877-025-05805-z]
112. De Vries N, Lavreysen O, Boone A, et al. Retaining healthcare workers: a systematic review of strategies for sustaining power in the workplace. Healthcare (Basel) 2023;11(13):1887. [doi: 10.3390/healthcare11131887]
113. Bienefeld N, Keller E, Grote G. AI interventions to alleviate healthcare shortages and enhance work conditions in critical care: qualitative analysis. J Med Internet Res 2025;27:e50852. [doi: 10.2196/50852]
114. Li H, Fu JF, Python A. Implementing large language models in health care: clinician-focused review with interactive guideline. J Med Internet Res 2025;27:e71916. [doi: 10.2196/71916]
115. Scheffler RM, Arnold DR. Projecting shortages and surpluses of doctors and nurses in the OECD: what looms ahead. Health Econ Policy Law 2019 Apr;14(2):274-290. [doi: 10.1017/S174413311700055X] [Medline: 29357954]
116. Mistry SK, Shaw M, Raffan F, et al. Inequity in access and delivery of virtual care interventions: a scoping review. Int J Environ Res Public Health 2022;19(15). [doi: 10.3390/ijerph19159411] [Medline: 35954768]
117. Coetzer JA, Loukili I, Goedhart NS, et al. The potential and paradoxes of eHealth research for digitally marginalised groups: a qualitative meta-review. Soc Sci Med 2024 Jun;350:116895. [doi: 10.1016/j.socscimed.2024.116895] [Medline: 38710135]

Abbreviations

AI: artificial intelligence
COMET: Core Outcome Measures in Effectiveness Trials Initiative
COS: core outcome set
COS-STAR: Core Outcome Set—Standards for Reporting
CReDECI 2: Revised Criteria for Reporting the Development and Evaluation of Complex Interventions
ICSS: Internet-Based Caregiver Support Service
ICT: information and communication technologies
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses, Extension for Scoping Reviews
ProCOS: Program Theory and Preparation of a Core Outcome Set Project
RCT: randomized controlled trial
SRQR: Standards for Reporting Qualitative Research
TIDieR: Template for Intervention Description and Replication

XSL•FO
RenderX

Edited by A Stone; submitted 01.Aug.2025; peer-reviewed by K Zhang, N Bievre, O Akhadelor; revised version received 25.Nov.2025; accepted 25.Nov.2025; published 20.Jan.2026.

Please cite as:
Bauernschmidt D, Bieber A, Hubrich R, Wittmann J, Meyer G
Program Theory and Core Outcome Set Development for a Technology-Assisted Counseling Intervention in Dementia: Multimethods Study
J Med Internet Res 2026;28:e81669
URL: https://www.jmir.org/2026/1/e81669
doi:10.2196/81669

© Dorothee Bauernschmidt, Anja Bieber, Ronja Hubrich, Janina Wittmann, Gabriele Meyer. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Validation of the Simplified Chinese Clear Communication Index Using Diabetes Education Materials: Instrument Adaptation and Validation Study

Lu Liu[1*], MSN; Yi Chen[2*], PhD; Xiaoqi Li[1], MSN; Hanyu Du[1], MSN; Rong Gui[1], MSN; Huiqing Jin[1], MSN; Fengfei Chen[1], MSN; Yujun Lin[1], MSN; Qiuhuan Huang[3,4], BSc; Yingge Tong[1], PhD

[1]Department of Nursing, School of Public Health and Nursing, Hangzhou Normal University, No. 2318, Yuhangtang Road, Hangzhou, China

[2]Department of Biomedicine and Prevention, University of Rome Tor Vergata, Rome, Italy

[3]Affiliated Hospital of Youjiang Medical University for Nationalities, Baise, Guangxi Zhuang Autonomous Region, China

[4]Guangxi Clinical Medical Research Center for Hepatobiliary Diseases, Baise, Guangxi Zhuang Autonomous Region, China

[*]these authors contributed equally

Corresponding Author:
Yingge Tong, PhD
Department of Nursing, School of Public Health and Nursing, Hangzhou Normal University, No. 2318, Yuhangtang Road, Hangzhou, China

Abstract

Background: In the postpandemic context, the surge of digital health information has intensified public demand for clear and practical communication, particularly in China, where health literacy disparities persist. The Clear Communication Index, developed by the US Centers for Disease Control and Prevention (CDC), is a standardized tool for assessing the clarity and actionability of health materials, but no version adapted to the simplified Chinese context has been established.

Objective: This study aims to translate and culturally adapt the Clear Communication Index into simplified Chinese, and subsequently validate its psychometric properties using diabetes health communication materials from provincial CDCs across Mainland China, a disease area with a substantial public health burden and strong reliance on health education.

Methods: Following a standardized cross-cultural process (forward-back translation and expert review), we developed the simplified Chinese version of the Clear Communication Index (C-CCI) and finalized a 12-item scale across 4 dimensions (Main Message and Call to Action, Behavioral Recommendations, Numbers, Risk) with yes or no scoring (0 - 100). One top-ranked diabetes health education material was sampled from each provincial CDC website in Mainland China (30/31 included; 96.8%) on May 18, 2025. Twelve raters with multidisciplinary backgrounds completed a 3-week standardized training program and independently evaluated each article (360 ratings). Structural validity was examined using exploratory factor analysis and confirmatory factor analysis. Content validity was assessed by the item-level content validity index and the scale-level content validity index/average. Reliability was evaluated by internal consistency (Cronbach α) and inter-rater agreement (Fleiss kappa), while convergent/discriminant validity was assessed using composite reliability and average variance extracted (AVE).

Results: The 4-factor structure was supported. Content validity was high, with the scale-level content validity index/average values of 0.976 for clarity and 1 for relevance. Overall reliability was acceptable (Cronbach α=0.837), with particularly strong internal consistency in the Risk dimension (Cronbach α=0.910). Inter-rater agreement was substantial (κ=0.624). Convergent validity (composite reliability=0.897 - 0.914; AVE=0.645 - 0.831) and discriminant validity were satisfactory. Application to 30 provincial CDC websites yielded a mean C-CCI score of 53.84 (SD 29.74), well below the recommended threshold of 90. No significant regional differences were observed in total scores; however, Behavioral Recommendations scored slightly higher in western provinces than in eastern and central regions (η^2=0.034), representing a small effect size with limited practical significance.

Conclusions: The C-CCI demonstrated good validity, reliability, and feasibility for evaluating simplified Chinese health communication materials. These findings underscore the need to strengthen health communication practices in China and encourage provincial CDCs to align material development with national health literacy goals. Integrating the C-CCI into routine CDC review protocols could support evidence-based quality assurance and advance clearer, more actionable public health communication nationwide.

(J Med Internet Res 2026;28:e83935) doi:10.2196/83935

XSL•FO
RenderX

KEYWORDS

clear communication index; psychometric validation; cross-cultural adaptation; health literacy; health education materials; public health communication; diabetes education; simplified Chinese

Introduction

Effective communication of health information is critical to facilitating accurate public understanding and proper application [1]. In the postpandemic era, both public attention to and demand for health information have increased significantly [2]. However, limited health literacy remains a widespread challenge worldwide, affecting about 88% of adults in the United States, nearly 70% in China, and up to 62.1% in parts of Europe [3-5]. Complex medical terminology and dense sentence structures often impede comprehension, especially among populations with limited literacy, contributing to poor adherence, adverse outcomes, and higher mortality [6,7]. These challenges highlight the need for clear, actionable, and audience-appropriate health materials.

To address these challenges, several standardized tools have been developed internationally to evaluate the quality of health communication materials, including the Suitability Assessment of Materials (SAM) [8], the Patient Education Materials Assessment Tool (PEMAT) [9], and the Clear Communication Index (CCI) [10]. These instruments differ in both scope and analytical emphasis. SAM adopts a broad suitability-oriented framework emphasizing presentation and cultural appropriateness, while PEMAT focuses on understandability and actionability. However, neither tool explicitly prioritizes key messages or systematically integrates numeric and risk communication.

In addition, several tools, including the Credible and Usable Evaluation of patient education tool for websites [11], the Health Information Website Evaluation Tool [12], the Health Information Quality Assessment Tool [13], and the QUality Evaluation Scoring Tool [14], have been developed to assess the quality and credibility of health information, particularly for professional or web-based contexts. However, their primary focus on credibility and usability limits their applicability to audience-centered, action-oriented evaluation of public health materials.

In contrast to these approaches, the CCI incorporates audience-centered principles, prioritizing key messages, numerical clarity, and behavioral guidance, making it a more comprehensive and action-oriented evaluation framework [10]. Among these instruments, the SAM and PEMAT have been translated and validated in simplified Chinese [15,16]; however, a validated simplified Chinese version of the Clear Communication Index (C-CCI) has not yet been developed.

Developed by the US Centers for Disease Control and Prevention (CDC) in 2014, the CCI provides a practical and evidence-based method for assessing the clarity of health materials [10]. It is available in a full version (20 items across 7 dimensions) for longer materials and a modified version (13 items across 4 dimensions) designed for short-format communications, both using binary scoring with a recommended minimum score of 90 to indicate acceptable clarity and usability [10,17,18]. The CCI has been widely applied across health domains, including patient education, cancer communication, and environmental health [19-27], and cross-culturally validated in countries, such as Brazil and Thailand [28,29]. A recent systematic review classified the CCI as a Grade B recommended instrument for health communication evaluation [30].

However, no validated simplified C-CCI currently exists. Given linguistic and cultural differences between Chinese and English, direct adoption may limit its accuracy and relevance [31]. In China, health communication is largely delivered through a nationwide network of CDCs that translate national health policies into public education, particularly for chronic disease prevention and management [32]. Diabetes, as a major noncommunicable disease, represents a central focus of CDC health education across provinces. Given provincial CDCs' mandate to bridge policy objectives and public engagement across regions, a standardized, context-sensitive framework for evaluating their educational materials is needed.

To address this gap, this study aimed to translate and culturally adapt the original CCI into a simplified Chinese version (C-CCI) through a standardized cross-cultural process and to evaluate its reliability and validity using health communication materials from provincial CDCs across Mainland China.

Methods

Translation and Cultural Adaptation of the CCI Tool

After obtaining formal authorization to use and adapt the original CCI, the research team followed the standardized cross-cultural adaptation model proposed by Sperber et al [33]. The process was conducted in 4 stages. First, 2 native simplified Chinese translators with expertise in health communication independently produced forward translations of the instrument (T1 and T2). Then, these were compared and reconciled by a senior bilingual translator with a medical background, who synthesized them into a single version (T3). To verify the linguistic equivalence, T3 was subsequently back-translated into English (T4) by an independent professional bilingual translator who had no prior knowledge of the tool and was not involved in the forward translation process, thereby minimizing potential bias and ensuring semantic fidelity. Finally, to further evaluate translation quality, we applied the linguistic equivalence assessment method recommended by Sperber et al [33]. This method involved comparing the original English version with the back-translated version across 2 dimensions. The first dimension is comparability of language, which refers to the similarity of words, phrases, and sentence structures. The second dimension is similarity of interpretability, which refers to the extent to which different wordings convey equivalent meaning. A linguist with a doctoral degree conducted this assessment using 4-point ordinal scales (1=extremely comparable or similar; 2=comparable or similar; 3=not comparable or similar; 4=not at all comparable or similar). Items rated as 3 or 4 on either dimension were reviewed by an expert

XSL•FO
RenderX

panel consisting of one health education specialist and 2 nurses, all of whom had formal training in health communication and substantial experience in patient-facing health education and counseling. Revisions were made through iterative panel discussions until consensus was reached. When the initial agreement was unattainable, additional rounds of discussion were used to determine the most contextually and semantically appropriate wording. The finalized items were then incorporated into the final C-CCI. The overall process is summarized in Figure 1.

Figure 1. Process of developing the modified simplified Chinese version of Clear Communication Index.

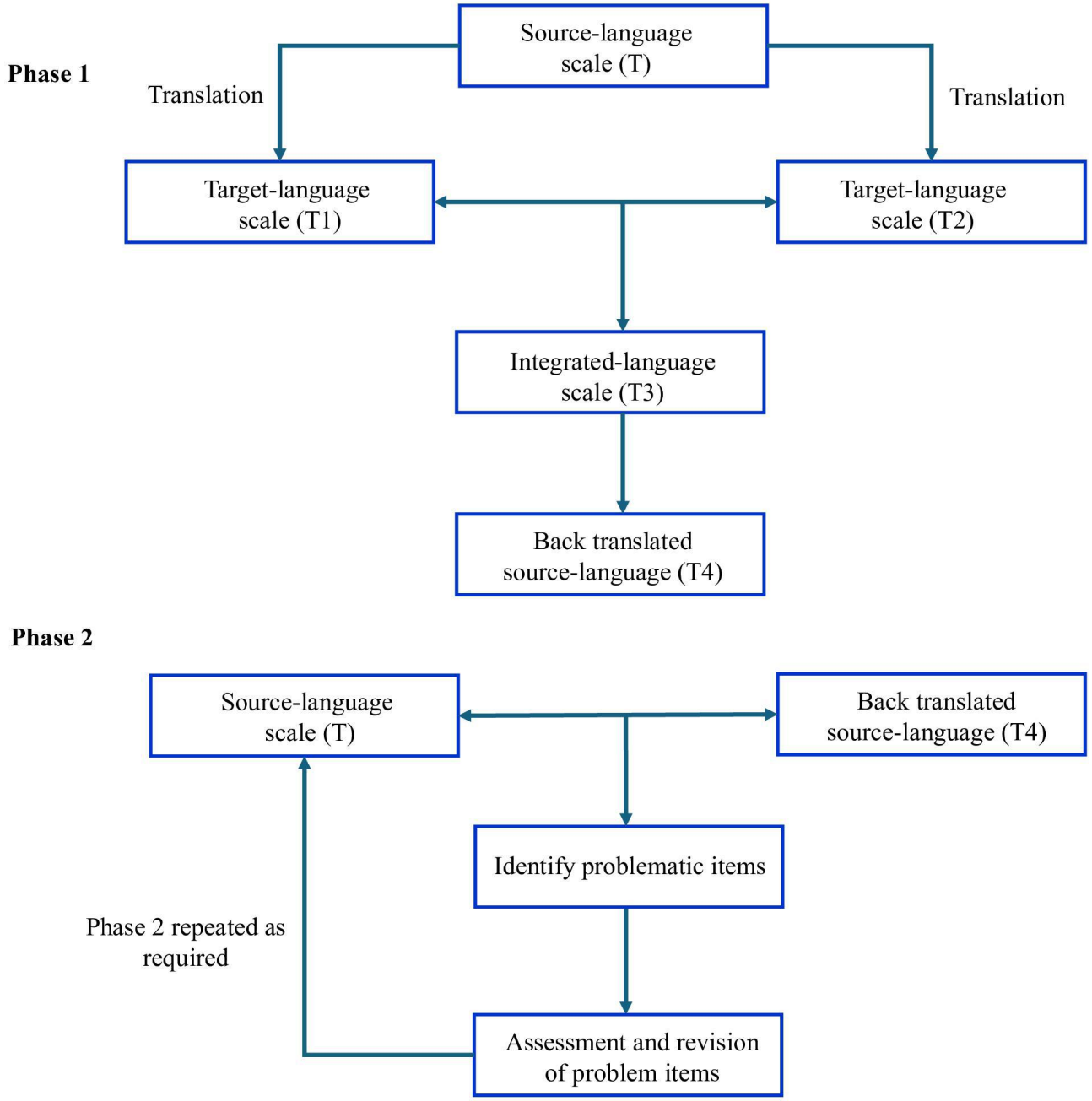

Reliability and Validity Evaluation of the C-CCI

Content Validity

Following Polit and Beck's guidelines for establishing content validity [34], 7 experts were recruited to ensure methodological rigor and representativeness. The expert panel comprised 7 raters (2 male and 5 female), with a mean age of 41 years (SD 5.8). All experts held senior professional titles. Their backgrounds in health education, chronic disease nursing, and nursing education provided a balance of clinical and educational perspectives for evaluating the C-CCI items. The evaluation focused on 2 dimensions: language clarity and content relevance. Experts independently rated each item using a 4-point Likert scale. For clarity, scores ranged from 1=unclear to 4=clear; for relevance, from 1=irrelevant to 4=relevant. Based on these ratings, both the item-level content validity index (I-CVI) and the scale-level content validity index/average (S-CVI/Ave) were calculated. The I-CVI was computed as the proportion of experts rating an item as 3 or 4. The S-CVI/Ave was calculated as the average item-level CVIs, reflecting the overall content validity of the instrument. According to published recommendations,

XSL•FO
RenderX

an S-CVI/Ave above 0.90 is considered acceptable; for panels of ≤5 experts, an I-CVI of 1 is required, whereas for panels of >5 experts, the minimum acceptable I-CVI is 0.78 [34,35].

Exploratory Factor Analysis

An exploratory factor analysis (EFA) was conducted to investigate the latent structure of the C-CCI scale. Prior to factor extraction, the Kaiser–Meyer–Olkin (KMO) test and Bartlett test of sphericity were performed to evaluate the suitability of the data. When the KMO value exceeded 0.60 and the Bartlett test yielded statistically significant results ($P<.05$), the data were deemed suitable for factor analysis [36]. Factors were extracted using principal component analysis. As prespecified in the analysis plan, the choice of rotation method was determined by the inter-factor correlations: when correlation coefficients exceeded 0.30, oblique rotation was applied; otherwise, orthogonal rotation was used [37]. Items were retained according to the following criteria: factor loading below 0.40, cross-loadings on multiple factors with a difference of less than 0.25. Items meeting any of these conditions were flagged for potential deletion [38].

Confirmatory Factor Analysis

Based on the factor structure derived from the EFA, a confirmatory factor analysis (CFA) was conducted to assess model fit. The assessment was performed at 2 levels: preliminary fit and overall fit indices. Preliminary criteria required all error variances to be positive, t-values of parameter estimates to reach statistical significance ($P<.05$), standard errors to be small, and factor loadings to fall within the range of 0.50 - 0.95.

Overall model fit was primarily evaluated using the root mean square error of approximation (RMSEA) and the comparative fit index (CFI), which are widely recommended as core indices for assessing model adequacy [39,40]. RMSEA values below 0.08 were considered indicative of acceptable fit, and values of CFI≥0.90 reflected good model fit. As complementary indices, the standardized root mean square residual (<0.08), the incremental fit index and Tucker–Lewis index (≥0.90), as well as the goodness-of-fit index (≥0.90), were also examined to provide a more comprehensive evaluation of model performance [39-41]. In addition, chi-square to degrees of freedom ratio (χ^2/df)<3 was considered indicative of good fit, whereas values between 3 and 5 were regarded as acceptable [36].

Convergent and Discriminant Validity

Composite reliability (CR) and average variance extracted (AVE) were calculated to evaluate convergent validity [42]. Convergent validity was considered strong when CR exceeded 0.70 and AVE was greater than 0.50, and acceptable when AVE ranged between 0.36 and 0.50 [43]. Discriminant validity was assessed by comparing the square root of the AVE for each construct with the inter-construct correlations. When the square root of AVE for a construct was greater than its correlations with other constructs, discriminant validity was deemed satisfactory [42].

Reliability

To assess the reliability of the C-CCI, Fleiss kappa (κ) test was applied to evaluate inter-rater agreement, and Cronbach α coefficient was computed to examine internal consistency. A κ value between 0.61 and 0.80 was interpreted as indicating substantial agreement [44], whereas a Cronbach α coefficient greater than 0.70 was deemed indicative of acceptable internal consistency [45].

Data Collection

A comprehensive search of the official websites of all 31 provincial CDCs in Mainland China was conducted, focusing on diabetes as the evaluation theme given its global rise, major public health burden, and reliance on health education for prevention and management [46,47]. The search was completed on May 18, 2025.

The keyword “diabetes” was used in the internal search engines of the provincial CDC websites to ensure accuracy and consistency. Consistent with the “least effort” principle, whereby the public typically focuses on top-ranked results [48], the highest-ranked diabetes-related health article from each provincial CDC was selected for analysis. Exclusion criteria included government reports, bidding information, and policy or regulatory documents exceeding 1000 words. The Uniform Resource Locators of the selected articles were compiled into an Excel database for systematic evaluation. The Jilin Provincial CDC website did not provide relevant content and was therefore excluded. In total, 30 provincial-level CDC websites were included, covering regions with diverse economic development levels and cultural backgrounds, thereby ensuring a nationally representative sample.

Nine graduate nursing students and 3 senior nurses with more than 5 years of endocrinology nursing experience were invited as evaluators. Prior to the formal evaluation, all evaluators completed a 3-week systematic training that covered the theoretical foundation of the tool, scoring criteria, and application methods. Training included multiple discussions and simulated scoring sessions to enhance the consistency and understanding of the tool’s dimensions. After the training, twelve raters independently evaluated health education materials collected from 30 provinces, producing a total of 360 data entries. To ensure stable and reliable results, factor analysis requires an adequate sample size. Recommendations vary, with common rules of thumb suggesting a minimum of 100 to 200 participants, or a subject-to-item ratio of at least 5:1 or 10:1 [49]. CFA, in particular, requires larger samples to achieve stable parameter estimates and sufficient statistical power [50]. Simulation-based studies using Monte Carlo methods indicate that sample size requirements for CFA depend more on model complexity, factor loadings, and indicator structure than on fixed numerical thresholds [50]. Under typical psychometric conditions, about 200 participants generally provide sufficient power for models of moderate complexity [50]. Accordingly, this study used a 2-step validation process using 2 independent samples: used 2 independent samples: 150 cases for EFA and 210 for CFA, the latter yielding more than 0.80 power for detecting RMSEA deviations at 0.08 [50], ensuring model stability and reliable fit evaluation.

XSL•FO
RenderX

Data Analysis

All statistical analyses were conducted using IBM SPSS Statistics (version 27.0; IBM Corp) and IBM SPSS Amos (version 28.0; IBM Corp). Based on the study design, consistency tests, reliability analyses, EFA, and CFA were performed to assess the scale's reliability and validity. All analyses were 2-tailed, and a P value <.05 was considered statistically significant.

Ethical Considerations

This study was approved by the Ethics Committee of Hangzhou Normal University (Approval No. 2025063). All participants provided informed consent before participating. All rater data were anonymized and deidentified prior to analysis, and the evaluated health education materials were obtained from publicly accessible sources. All evaluation data were coded, securely stored, and used solely for research purposes. The study was conducted in accordance with relevant institutional guidelines and the Declaration of Helsinki. All participants involved in this study were provided a small token of appreciation in the form of a gift worth approximately 30 RMB (US $4.13).

Results

Comparison Between Original and Back-Translation, and Cultural Adaptation

Comparison of the original and back-translated versions showed strong conceptual and semantic equivalence (Table 1). Across all items, the mean score for comparability of language was 1.308 (SD 0.63), and the mean score for SI was 1.077 (SD 0.28), indicating a high level of linguistic and interpretive consistency between the 2 versions. Overall variation across items was minimal, with the exception of item 3. While the phrase "policy change" can be directly translated, the research team cautioned that, within the context of public health in Mainland China, it may be misconstrued as implying radical "reform." To avoid this potential misinterpretation, the phrase was revised to "alignment with national health policies," thereby softening the connotation of drastic change and emphasizing policy continuity and optimization.

XSL•FO
RenderX

Table . This section compares the original and back-translated English versions of the Clear Communication Index Score sheet in terms of language comparability and interpretability similarity.

Original English version	Back-translated English version	Comparability of language	Similarity of interpretability
Does the material contain one main message statement? A main message is the one thing you want to communicate to a person or group that they must remember.	Does the material have a core message? The core message is what you want to convey to individuals or groups and wish them to remember.	2	1
Is the main message at the top, beginning, or front of the material? The main message must be in the first paragraph or section. A section is a block of text between headings.	Is the core message at the top, beginning, or front of the material? It must appear in the first paragraph or section. A paragraph is a block of text between headings.	1	1
Does the material include one or more calls to action for the primary audience? If the material includes a specific behavioral recommendation, a prompt to get more information, a request to share information with someone else, or a broad call for program or policy change, answer yes.	Does the material contain one or more calls to action for the main audience? If the material includes any of the following—for example, specific behavioral recommendations, guidance on how to access authoritative health information, encouragement to share the information with others, or a call to action aligned with national health policy directives.	3	1
Do both the main message and the call to action use the active voice?	Are both the core message and call to action in active voice?	1	1
Does the material always use words the primary audience uses?	Is the wording always appropriate to the main audience's habits?	2	2
Is the most important information the primary audience needs summarized in the first paragraph or section?	Is the most important information for the target audience outlined in the first paragraph or section?	1	1
Does the material include one or more behavioral recommendations for the primary audience?	Does the material offer one or more behavior recommendations for the target audience?	1	1
Does the material explain why the behavioral recommendation(s) is important to the primary audience?	Does it explain the importance of the recommended actions to the main audience?	1	1
Does the material always present numbers the primary audience uses?	Are numbers in the material always presented in a way that's familiar to the target audience?	1	1
Does the audience have to conduct mathematical calculations?	Does the audience need to do math?	1	1
Does the material explain the nature of the risk?	Does the material explain the nature of the risk?	1	1
Does the material address both the risks and benefits of the recommended behaviors? This includes actual risks and benefits and those perceived by your audience.	Does it describe the precautions and benefits of the recommended actions? This includes actual and perceived risks and benefits.	1	1
If the material uses numeric probability to describe risk, is the probability also explained with words or a visual?	If the material uses numbers (eg, 1/5 or 20%) to describe risks and also explains these probabilities with words or illustrations, select "yes".	1	1

To further enhance comprehensibility and applicability in the Mainland simplified Chinese context, culturally specific expressions were localized while maintaining consistency with the original scoring framework. For example, in the “NOTE” section of the scale, “Facebook posts,” “Twitter messages,” “scripts for podcasts” and “call center responses” were replaced with “WeChat official account articles,” “Weibo posts” “short video scripts” and “customer service responses” respectively. These adaptations enhanced cultural relevance and user understanding.

XSL•FO
RenderX

Validity and Reliability of the C-CCI

Content Validity

The scale demonstrated high content validity, with overall CVIs of 0.976 for clarity and 0.934 for relevance. Item-level CVIs ranged from 0.857 to 1 for clarity and from 0.124 to 1 for relevance. The notably low relevance score was attributable to Item 4 under the dimension "Main Message and Call to Action," which focused on the use of passive voice. Passive constructions, though acceptable in English health communication, are uncommon in simplified Chinese and may reduce clarity and perceived relevance. This reflects a cross-linguistic mismatch between formal and functional equivalence, consistent with Nida's theory of dynamic equivalence. Accordingly, Item 4 was deemed culturally unsuitable and removed [51]. After its removal, all items achieved a relevance index of 1, yielding a final scale of 12 items (Table 2).

Table . The reliability and content validity of the simplified Chinese version of the Clear Communication Index.

Variables	Cronbach α (N=360)	I-CVI[a] (n=7)	S-CVI[b] (n=7)
Main message and call to action	0.861	—[c]	—
Behavioral recommendations	0.894	—	—
Numbers	0.827	—	—
Risk	0.910	—	—
Total	0.837	0.857~1.000[d] or 1[e]	0.976[d] or 1[e]

[a]I-CVI: item-level content validity index.
[b]S-CVI: scale-level content validity index.
[c]Not applicable.
[d]For language clarity.
[e]For content relevance.

EFA

The KMO value was 0.769, and Bartlett test of sphericity was significant (χ^2_{66}=1034.141; P<.001), indicating the data's suitability for factor analysis. Four factors with eigenvalues greater than 1 were extracted, accounting for 78.7% of the total variance, consistent with the factor structure of the original scale. The scree plot also supported a 4-factor solution, with a clear inflection after the fourth factor (Figure 2).

Figure 2. Scree plot of Eigenvalues for exploratory factor analysis of the simplified Chinese version of the Clear Communication Index.

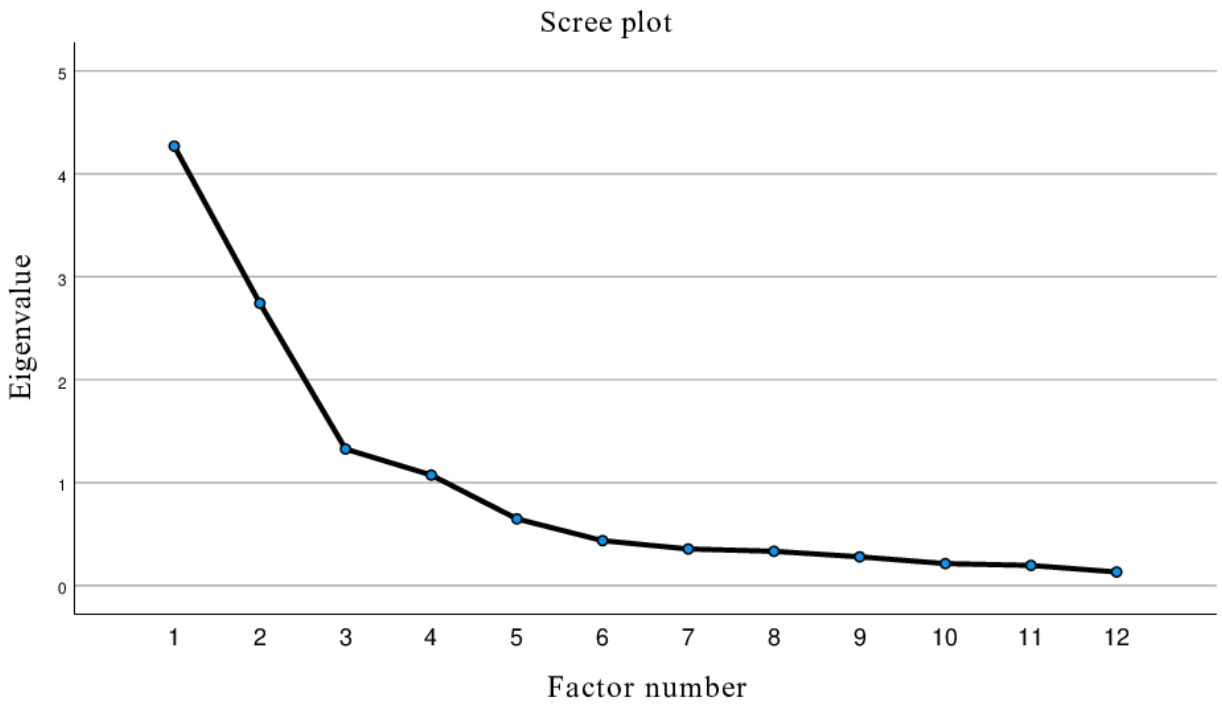

Item loadings (Table 3) were generally high and well defined, and all items showed satisfactory communalities, ranging from 0.475 to 0.903 [52]. Although item 4 had a slightly lower loading (0.484), it exceeded the 0.40 retention threshold and

XSL•FO
RenderX

demonstrated adequate communality and semantic alignment within the “Main Message and Call to Action” dimension; thus, it was retained. Overall, the EFA indicated a clear and stable factor structure with well-defined dimensions, providing strong support for the construct validity of the C-CCI.

Table . Factor loadings of exploratory factor analysis for the simplified Chinese version of the Clear Communication Index.

Items	Factor 1	Factor 2	Factor 3	Factor 4	Communality
A1	0.810	__[a]	—	—	0.711
A2	0.835	—	—	—	0.705
A3	0.863	—	—	—	0.769
A4	0.484	—	0.435	—	0.475
A5	0.878	—	—	—	0.810
B6	—	0.905	—	—	0.888
B7	—	0.922	—	—	0.903
C8	—	—	0.862	—	0.789
C9	—	—	0.890	—	0.830
D10	—	—	—	0.913	0.899
D11	—	—	—	0.819	0.785
D12	—	—	—	0.911	0.847

[a]Loadings < 0.40 were suppressed.

CFA

As shown in Figure 3, the CFA results supported the 4-factor structure of the C-CCI. Model fit indices indicated an excellent fit: $\chi^2_{48}=90.95$ ($\chi^2/df=1.895$), suggesting strong model adequacy; RMSEA=0.065, further supporting the good fit assumption. The CFI value was 0.972, exceeding the conventional cutoff of 0.90 and indicating strong model fit. Other indices also performed well: goodness-of-fit index=0.934, incremental fit index=0.972, Tucker–Lewis index=0.961, and standardized root mean square residual=0.072 (Table 4). All standardized factor loadings were substantial, ranging from 0.70 to 0.95, with each indicator loading strongly on its intended latent construct. No post hoc model respecifications were made based on modification indices, and the final model was retained as theoretically specified. These results demonstrate an excellent overall model fit, with stable and theoretically coherent parameter estimates across the 4-factor structure.

Figure 3. Standardized 4-factor model of the simplified Chinese version of Clear Communication Index.

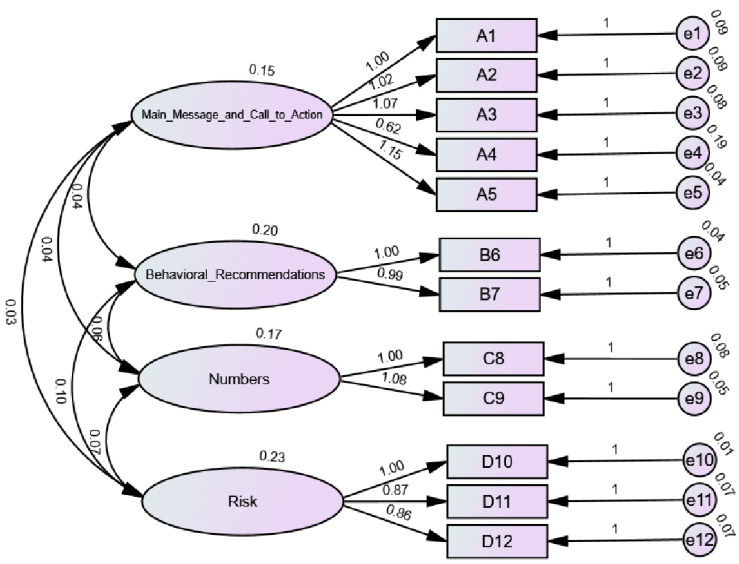

XSL•FO
RenderX

Table . Model fit index for the simplified Chinese version of the Clear Communication Index.

Index	Four-factor	Evaluation standard
χ^2 (*df*)	90.95 (48)	≤3
GFI[a]	0.934	>0.900
CFI[b]	0.972	>0.900
IFI[c]	0.972	>0.900
TLI[d]	0.961	>0.900
RMSEA[e]	0.065	<0.080
SRMR[f]	0.072	<0.080

[a]GFI: goodness-of-fit index.
[b]CFI: comparative fit index.
[c]IFI: incremental fit index.
[d]TLI: Tucker–Lewis index.
[e]RMSEA: root mean square error of approximation.
[f]SRMR: standardized root mean square residual.

Convergent and Discriminant Validity

CR values were 0.907, 0.897, 0.914, and 0.897 for Main Message and Call to Action, Behavioral Recommendations, Numbers, and Risk, respectively. The AVE values were 0.831, 0.814, 0.774, and 0.645, demonstrating strong convergent validity. Discriminant validity was assessed using the Fornell–Larcker criterion [43], comparing the square root of each construct's AVE with its inter-construct correlations (Table 5). In all cases, the square root of AVE exceeded its highest correlation with any other factor. Specifically, the Risk construct exhibited a square root of AVE of 0.803, showed a higher value than its strongest correlation with Behavioral Recommendations (r=0.442), and the same pattern was observed for Main Message and Call to Action (0.911), Behavioral Recommendations (0.902), and Numbers (0.880). Although moderate correlations were observed among some constructs, these findings indicate that the 4 dimensions are empirically distinct while remaining theoretically related.

Table . Convergent validity and discriminant validity of the simplified Chinese version of the Clear Communication Index.

Convergent validity and discriminant validity	Risk	Numbers	Behavioral recommendations	Main message and call to action	AVE[a]	CR[b]
Risk	0.803	—	—	—	0.645	0.897
Numbers	0.371[c]	0.880	—	—	0.774	0.914
Behavioral recommendations	0.442[c]	0.320[c]	0.902	—	0.814	0.897
Main message and call to action	0.152[c]	0.221[c]	0.241[c]	0.911	0.831	0.907

[a]AVE: average variance extracted.
[b]CR: composite reliability.
[c]P<.01.

Reliability

Expert ratings demonstrated substantial consistency (κ=0.624). The scale demonstrated high internal consistency overall (Cronbach α=0.837), with dimension-specific coefficients of 0.861 (Main Message and Call to Action), 0.894 (Behavioral Recommendations), 0.827 (Numbers), and 0.910 (Risk), confirming strong reliability across dimensions.

Evaluation of Diabetes Health Communication by Provincial CDCs in Mainland China

Of the 31 provincial-level administrative divisions in Mainland China, 30 CDC websites were included (with a coverage of 96.8%), ensuring strong national representativeness. Among diabetes-related health education materials, one province (Qinghai) provided video content with text captions, 8 used text–image formats, and 21 (70%) provided text-only content. Provinces were categorized into eastern, central, and western economic regions in Mainland China [53]. Mean C-CCI scores did not differ significantly across regions (East: mean 50.06, SD 29.06; Central: mean 55.26, SD 35.52; and West: mean

XSL•FO
RenderX

56.48, SD 26.32). The national mean score was 53.84 (SD 29.74), which fell well below the recommended threshold of 90, indicating considerable room for improvement in health communication quality. No regional differences were found across dimensions except for Behavioral Recommendations, where scores in the western region were significantly higher than those in the eastern and central regions (*P*=.002). However, the effect size was small (η^2=0.034), suggesting that regional grouping accounted for only a minor proportion of the variance in Behavioral Recommendations (Table 6).

Table . Comparison of scores across dimensions in Eastern, Central, and Western Regions of Mainland China (N=360).

Dimensions and regions	n (%)	Mean (SD)	*F* test (*df*)	*P* value	LSD[a]
Main message and call to Action			0.152（2）	.86	—[b]
East	132 (36.7)	2.7 (1.91)			
Central	84 (23.3)	2.83 (2.21)			
West	144 (40)	2.82 (1.94)			
Overall	360 (100)	2.78 (1.99)			
Behavioral recommendations[c]			6.271（2）	.002	East＜West, *P*=.001; Central＜West, *P*=.008
East	132 (36.7)	0.98 (0.95)			
Central	84 (23.3)	1 (0.96)			
West	144 (40)	1.34 (0.89)			
Overall	360 (100)	1.13 (0.94)			
Numbers			1.399 (2)	.25	—
East	132 (36.7)	1.09 (0.89)			
Central	84 (23.3)	1.3 (0.86)			
West	144 (40)	1.13 (0.96)			
Overall	360 (100)	1.16 (0.91)			
Risk			1.457 (2)	.23	—
East	132 (36.7)	1.23 (1.46)			
Central	84 (23.3)	1.5 (1.41)			
West	144 (40)	1.49 (1.27)			
Overall	360 (100)	1.4 (1.38)			
Total score			1.735 (2)	.18	—
East	132 (36.7)	50.06 (29.06)			
Central	84 (23.3)	55.26 (35.52)			
West	144 (40)	56.48 (26.32)			
Overall	360 (100)	53.84 (29.74)			

[a]LSD: least significant difference.

[b]Not significant ($P \geq .05$)

[c]η^2 represents the effect size, indicating the proportion of total variance explained by regional grouping. For the Behavioral Recommendations dimension, η^2 = 0.034, indicating a small effect size.

Discussion

Reliability and Validity of the C-CCI Under a Rigorous Sinicization Process

To our knowledge, this study represents the first rigorous and systematic cross-cultural adaptation process to translate, culturally adapt, and psychometrically validate the CCI, resulting in the C-CCI. Raters with substantial professional experience underwent structured training encompassing theoretical foundations, scoring standards, and simulated evaluations. This process enhanced scoring consistency and minimized subjective bias, strengthening the methodological rigor of data collection. The C-CCI retained the 4 original dimensions: Main Message and Call to Action, Behavioral Recommendations, Numbers, and Risk, thereby preserving theoretical coherence, a structure that was further supported by EFA and confirmed through CFA. Targeted revisions improved cultural applicability, including revising item 3 from “policy change” to “alignment with national health policies” and deleting item 4 due to the incompatibility

XSL•FO
RenderX

of passive voice in simplified Chinese. The deletion of items does not alter the scoring rules and, consequently, has no impact on the scoring threshold of the C-CCI. Following these adjustments, expert content validity scores improved, confirming the scientific soundness of the modifications. Psychometrically, the C-CCI demonstrated robust validity and reliability that not only met but exceeded international benchmarks for comparable Health Information Quality Assessment Tools. Prior reviews have shown that most instruments remain at Grade B due to limited structural validation and modest explained variance [30]. In contrast, the C-CCI achieved a cumulative variance of 78.7% in EFA and excellent model fit in CFA (RMSEA=0.065 and CFI=0.972), exceeding Grade B thresholds and reflecting superior structural coherence. Convergent validity was further supported by high CR and AVE values, while strong internal consistency (Cronbach α>.80) and substantial interrater agreement (κ=0.624) confirmed the instrument's stability and reproducibility.

Applicability and Promotion Potential of the C-CCI

Using diabetes health communication materials from 30 provincial CDC websites in Mainland China, including text-only, text-image, and video formats, this study systematically evaluated the real-world applicability of the C-CCI. The tool showed a clear structure and specific scoring dimensions, and raters achieved strong inter-rater reliability after brief training, demonstrating strong practicality and replicability in applied settings. The C-CCI effectively identified weaknesses in health communication materials across 4 dimensions, demonstrating strong problem detection capacity and intervention guidance value. Its stable performance across different regions and formats (text only, text-image, and video) further supports its adaptability to diverse contexts.

Importantly, the C-CCI aligns well with China's evolving digital communication ecosystem, in which health information is increasingly disseminated through official WeChat (Tencent) accounts, TikTok (ByteDance) videos, and other interactive platforms. Although these channels differ in format and engagement mechanisms, they share core demands for message clarity, concise information hierarchy, and actionable behavioral guidance, principles directly reflected in the C-CCI framework. From an implementation perspective, provincial CDCs may use the tool as a formative evaluation instrument during content development and internal review, using it to guide message structuring and action-oriented framing prior to dissemination. This approach enables consistent quality assurance across regions while maintaining flexibility in adapting to platform-specific communication styles.

Beyond the Mainland China context, the C-CCI's methodological foundation may also hold relevance for other Chinese-speaking regions that share similar linguistic and readability challenges, mirroring Mainland China communication patterns noted in previous studies [54-58]. However, its broader application should be guided by further adaptation studies that consider local policy environments and platform ecosystems, ensuring evidence-based contextualization rather than direct transfer.

Comparison With International Adaptations and Added Value of This Study

The CCI has been adapted into multiple languages and contexts worldwide. For instance, the Brazilian Portuguese version demonstrated applicability among primary health care professionals through field validation [28], while the Thai version emphasized dissemination of the theoretical framework and facilitated adoption in local health education [29]. These studies collectively established the feasibility of cross-cultural adaptation but were limited in the scope of psychometric testing and in the range of evaluated materials. Compared with prior validations, this study advances the methodological frontier of CCI research through multiformat empirical testing across 30 provincial CDC websites. By systematically evaluating diabetes-related materials in text-only, text-image, and video formats, the study assessed whether the measurement properties of the C-CCI remained stable across media that differ substantially in information organization. This format inclusivity, supported by consistent reliability across modalities, broadens the empirical use of the CCI framework beyond single-medium applications.

In addition, the study incorporated a more comprehensive psychometric evaluation strategy, combining exploratory and confirmatory factor analyses with rigorous validity and reliability testing. The convergence of large-scale institutional sampling and cross-format validation provides a robust methodological blueprint for future adaptations in diverse health communication settings. This study also contributes a transferable model for standardizing and assessing health information quality across varied media environments.

Quality of Diabetes Health Communication in Mainland China and Improvement Strategies

Using the C-CCI, this study systematically assessed the quality of diabetes health communication in mainland China, identifying overall quality levels, regional variation, and areas for improvement. Median scores across the 4 dimensions were approximately half of the maximum, revealing substantial room for enhancement. The national mean score of 53.84 (SD 29.74) fell below the recommended threshold of 90 [10], underscoring deficiencies in information completeness, structural coherence, and balance across dimensions.

Regional comparisons showed that western provinces scored significantly higher in Behavioral Recommendations than eastern and central provinces. However, the magnitude of this difference was small (η^2=0.034), indicating that regional grouping explained only a limited proportion of the variance in Behavioral Recommendations. This pattern may be partly influenced by unmeasured contextual factors (eg, regional economic capacity, resource distribution, and health communication infrastructure), pending future covariate adjustment and mediation analyses. Given the heterogeneous developmental landscape across and within China's provinces, aggregation into broad east–central–west categories may obscure underlying structural disparities and inflate within-group variance. For example, within the eastern region, provinces, such as Hainan and Hebei, may exhibit comparatively lower development profiles than other eastern provinces, while within

XSL•FO
RenderX

the western region, provinces and municipalities, such as Sichuan and Chongqing, may demonstrate development levels comparable to, or even exceeding, those of some central or eastern provinces. Similar internal disparities may also be present within the central region. These findings underscore the importance of interpreting statistically significant regional differences alongside effect size estimates. Hence, future research should adopt more fine-grained, mechanism-sensitive grouping strategies or multilevel analytical frameworks, incorporating appropriate covariate adjustment and mediation analyses to distinguish structural influences from context-specific practices. Based on standardized percentage scores, Risk emerged as the lowest-performing dimension, followed by Main Message and Call to Action, with Behavioral Recommendations and Numbers showing relatively better performance. Accordingly, improvement efforts should be prioritized in alignment with these empirical deficits: (1) enriching risk communication should be the foremost priority. This could involve contextualizing both absolute and relative risk information in line with evidence-based communication principles, supported by standardized templates and iterative user testing; (2) strengthening the clarity and authority of main messages should follow, through improved headline structuring, message salience testing (eg, digital A/B testing), and inclusion of credible source attribution; (3) improving numerical presentation by using absolute frequencies, consistent denominators, and simplified visual aids can promote better comprehension across literacy levels; and (4) structuring Behavioral Recommendations into concise, sequential action steps with links to relevant resources can reinforce practical applicability. These refinements would elevate both the scientific quality and communicative impact of diabetes health materials, thereby increasing the effectiveness of public health interventions.

Limitations

This study has several limitations. First, the sample was restricted to provincial CDC websites and focused exclusively on diabetes-related materials, without including hospitals, nongovernmental organizations, or social media platforms, which are also important health communication channels. Future research should expand to a broader range of health topics and communication platforms to test the generalizability of the C-CCI. Second, all raters had medical or nursing backgrounds. While this ensured professional rigor, it may have introduced certain biases. Rater training yielded acceptable interrater consistency (κ=0.624), although potential bias from disciplinary homogeneity remains. Future studies could strengthen methodological reflexivity by adopting hybrid rating panels that combine professional experts with lay raters to systematically compare expert-oriented and user-centered assessments. In this approach, lay raters could be purposively sampled to reflect variation in health literacy, education, and regional backgrounds, enabling a more representative appraisal of communication quality under real-world conditions. Such iterative and diversified designs would not only enhance measurement robustness but also model continuous improvement in cross-rater reliability and interpretive balance.

Conclusions

This study completed the cross-cultural translation and psychometric validation of the C-CCI, establishing a reliable and context-sensitive instrument for evaluating the clarity and actionability of health communication in Mainland China. Beyond measurement performance, the findings highlight the C-CCI's potential as a standardized quality assurance tool to support systematic improvement in public health communication, in line with national efforts to advance health literacy and equitable access to health information [59].

By demonstrating stable performance across text, text-image, and video formats in diabetes-related materials from provincial CDC websites, the C-CCI shows strong potential for routine application in CDC practice in chronic disease communication. As exemplified in diabetes education, this framework can be extended to broader chronic disease portfolios and used to support longitudinal monitoring of health communication quality over time. Its integration into content development and review workflows could facilitate evidence-informed standardization of health materials nationwide, helping to reduce avoidable comprehension gaps and strengthen public engagement. Future research should further validate this approach across disease areas and communication contexts, consolidating the role of the C-CCI within China's public health system.

Acknowledgments

The authors sincerely thank all participants who contributed to this study but did not meet the criteria for authorship. We are also grateful to the US Centers for Disease Control and Prevention for granting permission to translate and use the Clear Communication Index for research purposes.

Funding

This research received funding from the General Programs of Humanities and Social Sciences,Ministry of Education of China (Grant No. 23YJAZH136).

Data Availability

The data are available from the corresponding author, upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

LL, CY, and HQH contributed to the conceptualization of the study. Data curation was performed by LL, XQL, HYD, and RG. Formal analysis was conducted by LL and YC. The methodology was developed by YC. The investigation was carried out by LL, XQL, HYD, RG, HQJ, FFC, and YJL. Resources were provided by XQL, HYD, RG, HQJ, FFC, and YJL. Funding acquisition was the responsibility of YGT, who together with QHH, supervised the study. The original draft was written by LL, YC, and YGT. The article was revised by LL, YC, YGT, and QHH. This article has two corresponding authors, Qiuhuan Huang and YGT, with YGT serving as the primary corresponding author.

Conflicts of Interest

None declared.

References

1. National Academies of Sciences, Engineering, and Medicine. Effective Health Communication Within the Current Information Environment and the Role of the Federal Government: Proceedings of a Workshop: Washington, DC: The National Academies Press; 2023. [doi: 10.17226/27210] [Medline: 37878739]
2. Fu Y, Han P, Wang J, Shahzad F. Digital pathways to healthcare: a systematic review for unveiling the trends and insights in online health information-seeking behavior. Front Public Health 2025;13:1497025. [doi: 10.3389/fpubh.2025.1497025] [Medline: 40013047]
3. Brach C. AHRQ Health Literacy Universal Precautions Toolkit, 3rd edition: Rockville, MD: Agency for Healthcare Research and Quality; 2023.
4. National Health Commission of the People's Republic of China. Monitoring report on Chinese residents' health literacy 2024. URL: https://www.nhc.gov.cn/xcs/c100122/202501/18ecbeb9c42942bea9e0fced7a963299.shtml [accessed 2025-06-21]
5. The HLS19 Consortium of the WHO Action Network M-POHL. International report on the methodology, results, and recommendations of the European Health Literacy Population Survey 2019–2021 (HLS19) of M-POHL.: Vienna: Austrian National Public Health Institute; 2021 URL: https://m-pohl.net/Results [accessed 2025-08-25]
6. Nutbeam D, Lloyd JE. Understanding and responding to health literacy as a social determinant of health. Annu Rev Public Health 2021 Apr 1;42:159-173. [doi: 10.1146/annurev-publhealth-090419-102529] [Medline: 33035427]
7. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med 2011 Jul 19;155(2):97-107. [doi: 10.7326/0003-4819-155-2-201107190-00005] [Medline: 21768583]
8. Doak CC, Doak LG, Root JH. Teaching Patients With Low Literacy Skills, 2nd edition: J B Lippincott; 1996.
9. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014 Sep;96(3):395-403. [doi: 10.1016/j.pec.2014.05.027] [Medline: 24973195]
10. Baur C, Prue C. The CDC Clear Communication Index is a new evidence-based tool to prepare and review health information. Health Promot Pract 2014 Sep;15(5):629-637. [doi: 10.1177/1524839914538969] [Medline: 24951489]
11. Klompstra L, Liljeroos M, Lundgren J, Ingadottir B. A Clinical Tool (CUE-tool) for health care professionals to assess the usability and quality of the content of medical information websites: electronic delphi study. J Med Internet Res 2021 Feb 17;23(2):e22668. [doi: 10.2196/22668] [Medline: 33595439]
12. Zubiena L, Lewin O, Coleman R, et al. Development and testing of the health information website evaluation tool on neck pain websites - an analysis of reliability, validity, and utility. Patient Educ Couns 2023 Aug;113:107762. [doi: 10.1016/j.pec.2023.107762] [Medline: 37087877]
13. Lee N, Oh SW, Cho B, Myung SK, Hwang SS, Yoon GH. A Health Information Quality Assessment Tool for Korean online newspaper articles: development study. J Med Internet Res 2021 Jul 29;23(7):e24436. [doi: 10.2196/24436] [Medline: 34326038]
14. Robillard JM, Jun JH, Lai JA, Feng TL. The QUEST for quality online health information: validation of a short quantitative tool. BMC Med Inform Decis Mak 2018 Oct 19;18(1):87. [doi: 10.1186/s12911-018-0668-9] [Medline: 30340488]
15. Shan Y, Ji M, Xing Z, Dong Z, Wang D, Cao X. The simplified Chinese version of the suitability assessment of materials for the evaluation of health-related information for adults: translation and validation study. JMIR Form Res 2023 Jul 7;7(1):e41609. [doi: 10.2196/41609] [Medline: 37418296]
16. Shan Y, Ji M, Dong Z, Xing Z, Wang D, Cao X. The Chinese version of the Patient Education Materials Assessment Tool for printable materials: translation, adaptation, and validation study. J Med Internet Res 2023 May 18;25:e39808. [doi: 10.2196/39808] [Medline: 37200085]
17. Centers for Disease Control and Prevention (CDC). CDC clear communication index score sheet. 2014. URL: https://www.cdc.gov/ccindex/ [accessed 2025-08-25]
18. Centers for Disease Control and Prevention (CDC). Modified CDC clear communication index score sheet. 2014. URL: https://www.cdc.gov/ccindex/ [accessed 2025-08-25]

XSL•FO
RenderX

19. McClure E, Ng J, Vitzthum K, Rudd R. A mismatch between patient education materials about sickle cell disease and the literacy level of their intended audience. Prev Chronic Dis 2016 May 12;13:E64. [doi: 10.5888/pcd13.150478] [Medline: 27172259]
20. Tran BNN, Ruan QZ, Epstein S, Ricci JA, Rudd RE, Lee BT. Literacy analysis of national comprehensive cancer network patient guidelines for the most common malignancies in the United States. Cancer 2018 Feb 15;124(4):769-774. [doi: 10.1002/cncr.31113] [Medline: 29178322]
21. Modiri O, Guha D, Alotaibi NM, Ibrahim GM, Lipsman N, Fallah A. Readability and quality of wikipedia pages on neurosurgical topics. Clin Neurol Neurosurg 2018 Mar;166:66-70. [doi: 10.1016/j.clineuro.2018.01.021] [Medline: 29408776]
22. Phetxumphou K, Roy S, Davy BM, Estabrooks PA, You W, Dietrich AM. Assessing clarity of message communication for mandated USEPA drinking water quality reports. J Water Health 2016 Apr;14(2):223-235. [doi: 10.2166/wh.2015.134] [Medline: 27105408]
23. Alpert JM, Desens L, Krist AH, Aycock RA, Kreps GL. Measuring health literacy levels of a patient portal using the CDC's Clear Communication Index. Health Promot Pract 2017 Jan;18(1):140-149. [doi: 10.1177/1524839916643703] [Medline: 27188894]
24. Saeed F, Anderson I. Evaluating the quality and readability of internet information on meningiomas. World Neurosurg 2017 Jan;97:312-316. [doi: 10.1016/j.wneu.2016.10.001] [Medline: 27742505]
25. Porter KJ, Alexander R, Perzynski KM, Kruzliakova N, Zoellner JM. Using the clear communication index to improve materials for a behavioral intervention. Health Commun 2019 Jun;34(7):782-788. [doi: 10.1080/10410236.2018.1436383] [Medline: 29419320]
26. Etienne J, Chirico S, Gunabalasingham T, Jarvis A, ICF. Final report: clear communications and uncertainty. EFS3 2018;15(5):1412E. [doi: 10.2903/sp.efsa.2018.EN-1412]
27. Suka M, Yamauchi T, Yanagisawa H. Comparing responses to differently framed and formatted persuasive messages to encourage help-seeking for depression in Japanese adults: a cross-sectional study with 2-month follow-up. BMJ Open 2018 Nov 12;8(11):e020823. [doi: 10.1136/bmjopen-2017-020823] [Medline: 30420341]
28. Marinho A, Baur C, Ferreira FM, Borges-Oliveira AC, Abreu MD. Cross-cultural adaptation of the Clear Communication Index to Brazilian Portuguese. Rev Saude Publica 2020;54:26. [doi: 10.11606/s1518-8787.2020054001561] [Medline: 32187313]
29. Fakkheaw L, Suwanwaree J, Moungngern Y, Thintip K, Mirattanaphrai S. The CDC Clear Communication Index Health Education Materials Assessment Tool. SMB 2021;14(3):41-47 [FREE Full text] [doi: 10.33192/smb.v14i3.251326]
30. Li Y, Ouyang H, Lin G, Peng Y, Yao J, Chen Y. Evaluation of the measurement properties of online health information quality assessment tools: a systematic review. Int J Nurs Sci 2025 Mar;12(2):130-136. [doi: 10.1016/j.ijnss.2025.02.015] [Medline: 40241869]
31. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine (Phila Pa 1986) 2000 Dec;25(24):3186-3191. [doi: 10.1097/00007632-200012150-00014]
32. Chinese Center for Disease Control and Prevention. China CDC. URL: https://en.chinacdc.cn/about/overview/ [accessed 2026-01-01]
33. Sperber AD, Devellis RF, Boehlecke B. Cross-cultural translation: methodology and validation. J Cross-Cult Psychol 1994;25(4):501-524. [doi: 10.1177/0022022194254006]
34. Polit DF, Beck CT, Owen SV. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Res Nurs Health 2007 Aug;30(4):459-467. [doi: 10.1002/nur.20199] [Medline: 17654487]
35. Polit DF, Beck CT. The content validity index: are you sure you know what's being reported? Critique and recommendations. Res Nurs Health 2006 Oct;29(5):489-497. [doi: 10.1002/nur.20147] [Medline: 16977646]
36. Kaiser HF. An index of factorial simplicity. Psychometrika 1974 Mar;39(1):31-36. [doi: 10.1007/BF02291575]
37. Costello AB, Osborne J. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. Pract Assess Res Eval 2005;10(1). [doi: 10.7275/jyj1-4868]
38. Worthington RL, Whittaker TA. Scale development research: a content analysis and recommendations for best practices. Couns Psychol 2006;34(6):806-838. [doi: 10.1177/0011000006288127]
39. Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. Evaluating the use of exploratory factor analysis in psychological research. Psychol Methods 1999;4(3):272-299. [doi: 10.1037/1082-989X.4.3.272]
40. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Modeling 1999 Jan;6(1):1-55. [doi: 10.1080/10705519909540118]
41. Jöreskog KG, Sörbom D. LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language: Chicago, IL: Scientific Software International; 1993.
42. Sahoo M. Structural equation modeling: threshold criteria for assessing model fit. In: Subudhi RN, Mishra S, editors. Methodological Issues in Management Research: Advances, Challenges, and the Way Ahead: Emerald Publishing Limited, Vol. 2019:269-276. [doi: 10.1108/978-1-78973-973-220191016]
43. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. Journal of Marketing Research 1981 Feb;18(1):39. [doi: 10.2307/3151312]

XSL•FO
RenderX

44. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]
45. DeVellis RF, Thorpe CT. Scale Development: Theory and Applications, 5th edition: SAGE Publications, Inc; 2022.
46. Hossain MJ, Al-Mamun M, Islam MR. Diabetes mellitus, the fastest growing global public health concern: early detection should be focused. Health Sci Rep 2024 Mar;7(3):e2004. [doi: 10.1002/hsr2.2004] [Medline: 38524769]
47. Butayeva J, Ratan ZA, Downie S, Hosseinzadeh H. The impact of health literacy interventions on glycemic control and self-management outcomes among type 2 diabetes mellitus: a systematic review. J Diabetes 2023 Sep;15(9):724-735. [doi: 10.1111/1753-0407.13436] [Medline: 37407516]
48. Pang PCI, Chang S, Verspoor K, Pearce J. Designing health websites based on users' web-based information-seeking behaviors: a mixed-method observational study. J Med Internet Res 2016 Jun 6;18(6):e145. [doi: 10.2196/jmir.5661] [Medline: 27267955]
49. Worthington RL. Scale development research: a content analysis and recommendations for best practices. Couns Psychol 2006;34(6):806-838. [doi: 10.1177/0011000006288127]
50. Wolf EJ, Harrington KM, Clark SL, Miller MW. Sample size requirements for structural equation models: an evaluation of power, bias, and solution propriety. Educ Psychol Meas 2013 Dec;76(6):913-934. [doi: 10.1177/0013164413495237] [Medline: 25705052]
51. Nida EA. Toward a Science of Translating: With Special Reference to Principles and Procedures Involved in Bible Translating: EJ Brill; 1964.
52. Field A. Discovering Statistics Using IBM SPSS Statistics, 6th edition: Sage Publications; 2024.
53. Liu W, Liu Y, Twum P, Li S. National equity of health resource allocation in China: data from 2009 to 2013. Int J Equity Health 2016 Apr 19;15(1):68. [doi: 10.1186/s12939-016-0357-1] [Medline: 27093960]
54. Jin K, Ding D, Gullick J, Koo F, Neubeck L. A Chinese immigrant paradox? Low coronary heart disease incidence but higher short-term mortality in western-dwelling Chinese immigrants: a systematic review and meta-analysis. J Am Heart Assoc 2015 Dec 18;4(12):e002568. [doi: 10.1161/JAHA.115.002568] [Medline: 26683217]
55. Kuan AS, Chen TJ, Lee WC. Barriers to health care services in migrants and potential strategies to improve accessibility: a qualitative analysis. J Chin Med Assoc 2020 Jan;83(1):95-101. [doi: 10.1097/JCMA.0000000000000224] [Medline: 31714441]
56. Zhu C, Song B, Walsh CA, Ekoh PC, Qiao X, Xu A. Barriers to accessing health care of older Chinese immigrants in Canada: a scoping review. Front Public Health 2024;12:1445964. [doi: 10.3389/fpubh.2024.1445964] [Medline: 39564368]
57. Zhang L, Ding D. Coronary heart disease prevention among Chinese immigrants: the importance of health literacy. Health Literacy and Communication Open 2025 Dec 31;3(1). [doi: 10.1080/28355245.2025.2476134]
58. Zhang L, Ding D, Neubeck L, Gallagher R. Health literacy as a predictor of emergency department visits and self-rated health among Chinese immigrants: findings from an Australian survey. Patient Educ Couns 2020 Nov;103(11):2353-2360. [doi: 10.1016/j.pec.2020.04.017] [Medline: 32456982]
59. State Council of the People's Republic of China. Healthy China Action (2019–2030). URL: https://www.gov.cn/xinwen/2019-07/15/content_5409694.htm [accessed 2025-09-22]

Abbreviations

AVE: average variance extracted
C-CCI: Chinese version of Clear Communication Index
CCI: Clear Communication Index
CDC: Centers for Disease Control and Prevention
CFA: confirmatory factor analysis
CFI: comparative fit index
CR: composite reliability
EFA: exploratory factor analysis
I-CVI: Item-level Content Validity Index
KMO: Kaiser–Meyer–Olkin
PEMAT: Patient Education Materials Assessment Tool
RMSEA: root mean square error of approximation
S-CVI/Ave: Scale-level Content Validity Index/Average
SAM: Suitability Assessment of Materials

XSL•FO
RenderX

Edited by A Schwartz, M Balcarras; submitted 15.Sep.2025; peer-reviewed by C Ahmadu, R Nooripour; accepted 23.Jan.2026; published 20.Feb.2026.

Please cite as:
Liu L, Chen Y, Li X, Du H, Gui R, Jin H, Chen F, Lin Y, Huang Q, Tong Y
Validation of the Simplified Chinese Clear Communication Index Using Diabetes Education Materials: Instrument Adaptation and Validation Study
J Med Internet Res 2026;28:e83935
URL: https://www.jmir.org/2026/1/e83935
doi:10.2196/83935

© Lu Liu, Yi Chen, Xiaoqi Li, Hanyu Du, Rong Gui, Huiqing Jin, Fengfei Chen, Yujun Lin, Qiuhuan Huang, Yingge Tong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Digital Tool for Assessing the Distinct Effects of Depression, Anxiety, and Attention-Deficit/Hyperactivity Disorder (ADHD) on Children's Emotional Cognitive Bias: Cross-Sectional Study

Sang-Eon Park[1,2], PhD; Jisu Chung[1]; Sang Ah Lee[1,3], PhD

[1]Department of Brain and Cognitive Sciences, Seoul National University, Seoul, Republic of Korea

[2]Brain Imaging Center, Seoul National University, Seoul, Republic of Korea

[3]Institute for Data Innovation in Science, Seoul National University, Seoul, Republic of Korea

Corresponding Author:
Sang Ah Lee, PhD
Department of Brain and Cognitive Sciences
Seoul National University
Gwanakro 1
Seoul,
Republic of Korea
Phone: 82 28802863
Email: sangahlee@snu.ac.kr

Abstract

Background: Emotional wellness and healthy neurocognitive development are crucial from early childhood. An imbalance in attentional and emotional regulation system is associated with an increased risk of depression, anxiety, and attention-deficit/hyperactivity disorder (ADHD). Early assessment of these risks is essential, but it is difficult to conduct cognitive tests that are both child-friendly and able to dissociate different behavioral biases.

Objective: This study aimed to develop a digital app–based tool designed for young children to objectively assess cognition-affect interactions and examine the association with standardized scales for anxiety, depression, and ADHD.

Methods: In this cross-sectional study, 78 healthy children (36 female) aged 4-10 (mean 7.2, SD 1.4) years with no history of mental illness were recruited from the local community center and children's museum. Emotional regulation and attentional control were assessed using an animated emotional Flanker task, emotional Stroop task, and emotional Go/No-Go task on a touchscreen computer. Children's current mental health was measured using self-reported depression and anxiety states through the Center for Epidemiological Studies Depression Scale for Children (CES-DC) and State-Trait Anxiety Inventory for Children (STAI-CH), while the ADHD risk was assessed using the Korean ADHD Rating Scale (K-ARS) for parents. Principal component analysis was applied to behavioral measures across the tasks to group them by similarities and extract 3 abstract scores ("E-scores") representing different aspects of cognitive function (attention, selective inhibition, and emotional sensitivity). Associations between E-scores and mental health or ADHD risk were then tested.

Results: There was a significant improvement in general attention across development (Pearson correlation between E-score 1 and age: r=–0.75; P<.001; 2-sided α=.05) but not emotion-attention interactions. Performance was also correlated with mental health scales. First, children with higher depression symptoms (ie, higher CES-DC) were slower in their responses in general (ie, higher E-score 1; Pearson correlation after controlling for age: r=0.29; P=.04). Second, both anxious and depressed (ie, higher STAI-CH and CES-DC) children demonstrated reduced attention selectively to the emotional stimuli as indicated by elongated RT and lower accuracy (ie, higher E-score 2; anxiety: r=0.34; P=.02; depression: r=0.51; P<.001). Lastly, children with higher ADHD scales (ie, higher Korean ADHD Rating Scale [K-ARS]) showed lower accuracy across the three tasks, particularly for emotional stimuli (ie, lower E-score 3; r=–0.32; P=.03).

Conclusions: By combining well-established emotional cognitive tasks with our dimensionality reduction techniques, we extracted individual affective-cognitive characteristics from diverse but noisy behavioral patterns and identified their association with mental health and ADHD-related symptoms in children. These results demonstrate the scientific validity, versatility, and translational potential of our gamified digital assessment tool for monitoring young children's affective and cognitive health in daily life. Future longitudinal studies in children with formal clinical diagnoses will further strengthen the generalizability of these findings.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e86286) doi:10.2196/86286

KEYWORDS

ADHD; attention; attention-deficit/hyperactivity disorder; eHealth; emotion; mental health

Introduction

Early childhood is a crucial phase of cognitive and emotional development that sets the stage for children's well-being throughout life. Disruptions in the healthy processing of emotional stimuli not only raise the risk of mental disorders such as depression, anxiety, and attention-deficit/hyperactivity disorder (ADHD), but also make it challenging for children to deploy and regulate attentional control, causing difficulties in children's learning and school performance [1]. However, because these symptoms often begin at a young age [2], early monitoring and objective identification of individual vulnerability have been challenging based on traditional parent reports and clinical interviews.

Understanding how emotional processing in the brain develops during childhood provides insight into how such interactions between emotional and cognitive function are related to mental health. During childhood, the brain first undergoes subcortical development in emotion-processing regions such as the amygdala [3] and reward-processing regions such as the ventral striatum [4]. Across maturation, these regions form stronger network connectivity to the prefrontal cortex (PFC) [5], which develops gradually for modulating arousal and attention [5,6], ensuring that emotional stimuli are assessed and managed in a context-appropriate manner. Due to the immature regulatory function of the PFC, younger children rely more on subcortical function. This developmental imbalance causes them to exhibit greater sensitivity to emotional stimuli [7,8], which can be detrimental to both emotional wellness and cognitive performance from early childhood [9].

Based on findings from developmental cognitive neuroscience, the effects of depression, anxiety, and ADHD on cognitive abilities such as emotion recognition, attention, and inhibitory control may be dissociable despite their high comorbidity [10]. For instance, while depression is mainly characterized by feelings of sadness, low mood, and loss of interest in their usual activities [11], it is also associated with slower cognitive processing speed [12] and difficulty disengaging from negative emotional information [13]. Anxiety, on the other hand, is more strongly associated with heightened reactivity to potential threat [14,15], leading to an attentional bias toward emotional stimuli and an impairment in one's ability to focus on nonthreatening, neutral stimuli [16]. Finally, although ADHD is marked by deficits in cognitive control, impulse regulation, and hyperactivity in daily life patterns [17], children with ADHD tend to be particularly more impulsive in response to rewarding or positively valenced stimuli [18,19].

Serious games can serve as digital monitoring tools that can be applied in daily life [20] and are particularly effective for young children who may be less receptive to repeated administrations of self-report surveys and cognitive tests. Recent studies have demonstrated that digital tools can be developed by leveraging associations between cognitive profiles and mental health or ADHD symptoms [21]. For example, one study assessed ADHD risk in children aged 4-8 years by examining cognitive functions such as reward processing using computer-based games [22]. Other studies have measured attention and working memory to examine their associations with depression [23] and anxiety [24] symptoms in adults. In addition, some studies did not directly evaluate the association between performance of digital cognitive tasks and mental health scales but instead assessed related cognitive domains, such as executive functions [25] and cognitive control [26]. However, most prior digital task–based studies have focused on single symptom domains, without accounting for the high comorbidity among ADHD, depression, and anxiety, which may not only necessitate separate tests for each but also limit the interpretability of assessments.

Unfortunately, despite the overwhelming evidence that mental health is simultaneously influenced by emotional sensitivity and cognitive processing [27-30], there is still a lack of digital tasks that capture such interactions within the scope of a game-like application that can be widely used by the public. More importantly, prior studies showed critical limitations in distinguishing the impacts of depression, anxiety, and ADHD simultaneously, using a single versatile assessment tool, despite the differences in the neurocognitive mechanisms that underlie their specific symptomatology [31]. The duration and user-friendliness of such tasks are especially critical for testing young children, as multiple lengthy traditional laboratory-based tests often fail to sustain children's interest or use excessively violent or provocative emotional stimuli that are inappropriate for young children [32].

Given the above limitations, we set out to develop and validate a novel digital assessment tool that can characterize the relationships between children's mental health or ADHD survey scores and emotional cognitive performance, an integration that has been largely absent from previous studies. Based on our previous study on a diverse range of both healthy participants and depressed adults [33], we designed a gamified task that is engaging and suitable for young children, which can be administered easily in about 30 minutes. A unique and novel aspect of our approach was to calculate abstract componential measures extracted from a range of behavioral patterns, which allowed us to identify individual cognitive-affective characteristics that were differentially associated with anxiety, depression, and ADHD symptoms. Unlike prior digital task–based tools that focus on isolated symptom domains, our framework enables simultaneous assessment of mental health and ADHD symptoms within a single test.

We developed modified versions of the emotional Flanker task (eFlanker), emotional Go/No-Go task (eGoNoGo), and emotional Stroop task (eStroop) using animated faces on a touchscreen interface. The eFlanker measured children's attentional control in quickly assessing the perceptual stimuli and responding to the target stimulus without being influenced

XSL•FO
RenderX

by nearby distractors. The eGoNoGo measured children's ability to remember and select particular combinations of emotional and perceptual features, while asserting cognitive control to suppress responses to nontarget lures. Additionally, we developed a novel eStroop, requiring simple responses to the target features while inhibiting the irrelevant information. These 3 tasks demonstrated their validity as a psychometric test for children (aged 4-6 years) in previous studies [34-36]. Mental health and ADHD symptoms were assessed using standardized surveys that are widely used in both research and clinical settings.

To ensure the task was usable and relevant for the general public, we assessed children who may be at elevated risk for cognitive-affective difficulties but had not yet been diagnosed with a mental disorder. Recruitment in a community-based context allowed us to obtain a broad and representative dataset spanning healthy to at-risk individuals while minimizing explicit selection biases. We chose a target age range of 4-10 years, which corresponds to early childhood, a developmental period marked by ongoing maturation of prefrontal cortical systems supporting emotional regulation and cognitive control. This period is also when individual affective-cognitive characteristics related to anxiety, depression, and ADHD begin to emerge [37-39]. Given that adolescence is associated with distinct neurobiological mechanisms in processes such as risk-taking and emotion regulation [40], we limited recruitment to children younger than 10 years of age, in accordance with the definition of adolescence provided by the World Health Organization [41].

Methods

Participants

Children aged between 4 and 10 years without any history of mental illness were recruited from a local community center and a local children's museum (Gyeonggi Children's Museum; N=78; 36 female, mean age 7.2, SD 1.4 years). A convenience sampling approach in a public, community-based context was used to recruit children from diverse environments while minimizing explicit selection biases. Data were collected over a 6-month period (August 2023 to February 2024). This schedule was chosen to avoid the annual grade transition in March within the Korean education system, which can represent a substantial environmental change for children. Children who could not use the touchscreen were excluded from the study.

Sample Size Determination

To determine the required sample size for this study, we referred to previous studies reporting associations between attention-related behavioral features and mental health or ADHD scales. In our previous work [33], emotion-related slowing of reaction time (RT) in the Flanker task (relative to the neutral condition) showed Pearson correlations of r=0.342 with depression symptoms and r=0.359 with anxiety symptoms. A power analysis showed that the minimum sample sizes required to achieve 80% statistical power (β=.2) with a 95% confidence level for these correlation coefficients were 64 and 59, respectively. In addition, we referred to another study reporting an association between error rate in the Flanker task and ADHD rating scale scores (Pearson correlation r=0.52) [42], for which the minimum required sample size under the same criteria was 27. To account for potential participant loss and missing data that could reduce statistical power, the target sample size was increased by 20%, resulting in a final sample size of 78 participants.

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki, and all experimental procedures were approved by the institutional review board for human research of Seoul National University (2501/003-021 and 2407/001-014). Written informed consent, which included all general elements required for research involving children, such as the purpose of the study, experimental procedures, potential benefits and discomforts, confidentiality, and the right to withdraw, was obtained from the parent or legal guardian. All study data were deidentified immediately after collection. All participants were compensated with 20,000 KRW (US $14) for 30 minutes of participation.

Emotional Cognitive Tasks

We designed 3 emotional cognitive tasks (eFlanker, eGoNoGo, and eStroop) to measure attention, working memory, and selective inhibition along with emotional processing (Figure 1). All tasks were developed based on the PsychoPy software (version 23.1.2; Open Science Tools Ltd) and implemented on a computer with responses recorded using a touchscreen monitor.

XSL•FO
RenderX

Figure 1. (A) The 3 emotional cognitive tasks used in this study. A diagram with an example screenshot of the emotional cognitive tasks: (B) emotional Flanker task (eFlanker), (C) emotional Go/No-Go task (eGoNoGo), and (D) emotional Stroop task (eStroop).

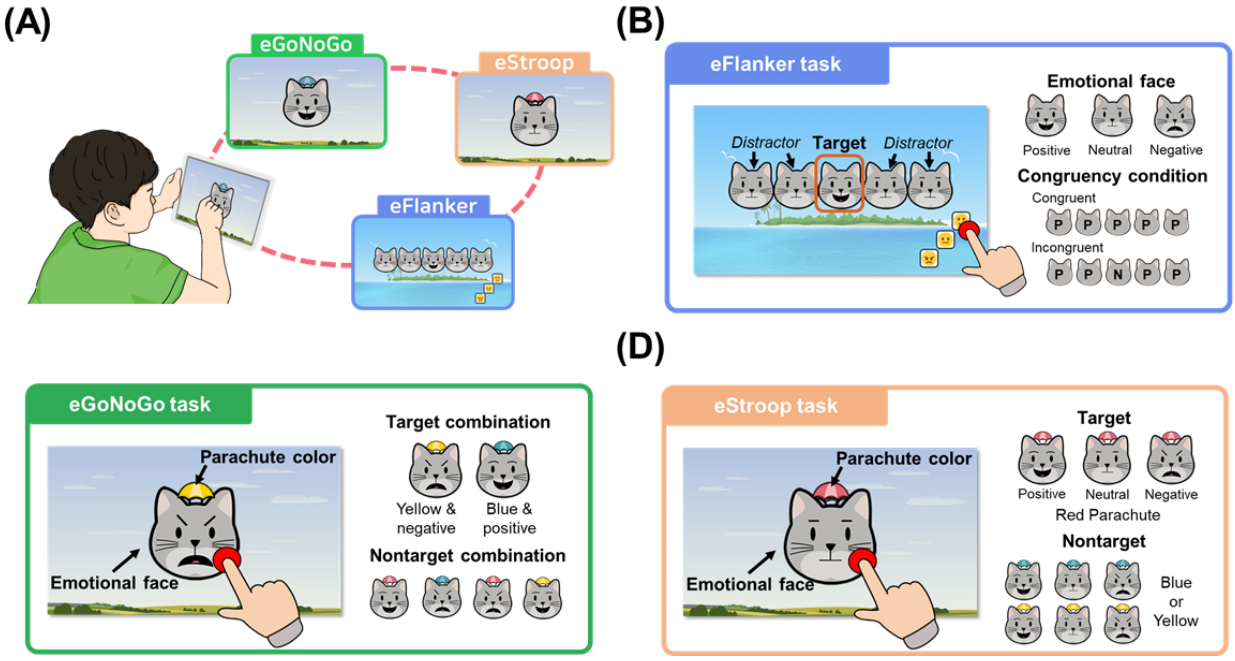

eFlanker

The eFlanker was designed to test for attention and its interactions with emotional stimuli [43,44]. Children were instructed to choose the face (located at the bottom right side of the screen) expressing the same emotion as the target, which was presented in the middle, with 2 distractors on both sides of the wings. The target and distractors were presented after a brief fixation cross (1000 ms), and feedback was provided on the screen once the child responded. The target and distractor consisted of 3 emotional (positive, neutral, and negative) cat faces, and there were a total of 9 possible combinations for their presentation as the target and distractor. Children performed 18 trials of practice blocks to familiarize themselves with the task, and then conducted 2 main blocks with 90 trials each (total 180 trials, 20 trials for each condition), with trial types presented in a random order. A break time was given between the main blocks.

eGoNoGo

The eGoNoGo was designed to measure emotional bias in executive function and memory by testing one's ability to choose the target stimuli while withholding responses to the nontarget objects [45,46]. Out of 6 possible combinations from 3 parachute colors (yellow, red, and blue) and 2 emotions (positive and negative), children were instructed to respond to 2 target combinations of a specific parachute color and a specific emotional cat face (eg, blue parachute with the positive cat face and the yellow parachute with the negative cat face). The color-emotion combination was counterbalanced and randomly assigned across participants. After the fixation cross was presented for 1000 milliseconds, a stimulus was presented in the middle of the screen. The nontarget stimuli were presented for 1500 milliseconds, and the target stimuli were presented until the participants responded. If participants responded to the nontarget (incorrect response) or responded later than 2000 milliseconds, visual feedback on their performance ("Incorrect" or "Respond quickly") was displayed for 1000 milliseconds. Children performed 20 practice trials and then 2 main blocks with 60 trials each (total 120 trials). We adjusted the frequency of the target trials to be slightly higher, to 40% of the total trials, to offset the imbalance between the target (2 out of 6 possible combinations) and nontarget trials (4 out of 6) to keep the children engaged in the task. A break time was given between the main blocks.

eStroop

The eStroop was designed to test inhibitory control by regulating children's response to the nontarget object. Children were asked to touch the face stimuli with a specific parachute color (either red, blue, or yellow) while ignoring the other colors, regardless of the emotion of the face (positive, neutral, and negative). The color-emotion combination was counterbalanced across participants. After a fixation cross was presented for 1000 milliseconds, the nontarget trial was presented for 2500 milliseconds, and the target trial was presented until the participants responded. If participants responded to the nontarget (incorrect response) or responded to the target later than 2000 milliseconds, visual feedback ("Incorrect" or "Respond quickly") was displayed for 1000 milliseconds. Children performed 12 practice trials and then 3 main blocks with 36 trials each (total 108 trials). A break time was given between the main blocks.

Procedures

The experiments were conducted in a quiet room either in the laboratory or in the local children's museum. Both experimental environments were child-friendly and were arranged to be similar to each other. Anxiety and depression levels were measured using the State-Trait Anxiety Inventory for Children (STAI-CH) [47] and Center for Epidemiological Studies Depression Scale for Children (CES-DC) [48] scales, which

XSL•FO
RenderX

were designed specifically for children. In addition, their caregivers completed the ADHD questionnaire using the K-ARS [49] based on *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition] [50]).

Children and their caregivers were provided with informed consent and a description of the overall task. After a brief instruction on the tasks, followed by practice sessions, the children performed the 3 emotional cognitive tasks in a mixed order. During this time, the caregivers completed the ADHD survey (K-ARS) about their children. After the completion of the main tasks, the children were surveyed on questions about their mental health (CES-DC and STAI-CH) either through a written form on an electronic device or through conversation with the experimenter, based on their preference. We did not observe significant differences in CES-DC or STAI-CH scores between the self-completed and experimenter-administered responses (CES-DC: t_{73}=–0.725; *P*=.47; STAI-CH: t_{73}=0.09; *P*=.93; 2-sample *t* test).

Missing Data

Some children missed 4 or more consecutive trials due to distraction or an explicit request to discontinue participation during the task. In total, 14 children met these criteria. Specifically, the number of missing cases was 2 for the eFlanker, 4 for the eGoNoGo, and 10 for the eStroop. These data were excluded from the entire analysis.

Behavioral Indices from the Emotional Cognitive Tasks

Accuracy and RT were calculated from the emotional cognitive tasks to compare behavioral performance across conditions and children. We averaged RTs from the correct trials, removing those shorter than 200 milliseconds to eliminate accidentally pressed responses. Outliers for RT or accuracy in each emotional condition were excluded at the subject level. Outliers were defined as values exceeding 3 times the median absolute deviation from the median, where the median absolute deviation was computed as the median of the absolute deviations from the median.

Extracting E-Scores as an Abstract Measure of Task Performance

Abstract scores were required to summarize a range of behavioral indices in the children's performance on the emotional cognitive tasks. The behavioral indices from 64 children were used for these analyses, after removing 14 children who lost interest during 1 of the 3 tasks (missing data), resulting in RT or accuracy values that were not appropriate for measuring an association with their mental health or ADHD. First, we calculated 82 behavioral indices consisting of RT and accuracy from various conditions of the 3 tasks (eFlanker, eGoNoGo, and eStroop), including averaged RT/accuracy of whole trials or emotion-related RT/accuracy difference between emotional conditions or between congruent and incongruent conditions (Table S1 in Multimedia Appendix 1). Second, to extract the most informative components while minimizing collinearity, 2D latent features were derived by extracting the first and second principal components (PCs) of the *z*-scored behavioral indices [33,51,52]. To avoid having the principal component analysis (PCA) dominated by the overrepresentation of outliers, *z*-scored indices with an absolute value over 2.5 were reduced to 2.5. Only the first two PCs explained more than 10% of the variance (Figure S1 in Multimedia Appendix 1). Then, the k-means clustering method was applied to the 2 PCs for dividing the latent features into several groups based on their squared Euclidean distance [53,54]. The optimal number of groups to maximally separate the latent features was determined by identifying the point at which the rate of decrease in inertia (within-cluster sum of squared errors) slowed down [55] (Figure S1 in Multimedia Appendix 1). Behavioral indices within the same group represented a similar pattern across the participants (eg, RT in the eFlanker was highly correlated with RT in the eStroop). Lastly, we calculated E-scores ranging from 1 to 3 by calculating the first PC of *z*-scored behavioral indices from each of the 3 groups.

Statistical Tests

Behavioral performance across emotional conditions was compared using paired *t* tests with a 2-sided significance level of α=.05. We also performed analysis of covariance (ANCOVA) on task performance measures (RT and accuracy) and E-scores, with gender as a fixed factor and age as a continuous covariate. Pearson partial correlation analyses were conducted to assess associations between E-scores and mental health or ADHD scales while controlling for age.

Results

Depression, Anxiety, and ADHD Scales Across Development

We investigated whether children's anxiety, depression, and ADHD scores were correlated with one another and how they differed across age. First, there was a strong association between depression (CES-DC) and anxiety (STAI-CH; Pearson correlation *r*=0.493; *P*<.001), while the risks for ADHD (K-ARS) were not correlated with either anxiety or depression (K-ARS and CES-DC: *r*=0.068; *P*=.58; K-ARS and STAI-CH: *r*=0.062; *P*=.61). When we conducted an ANCOVA with gender as a fixed factor and age as a covariate, we found higher depression symptoms in younger ($F_{1,72}$=7.044; *P*=.01) and male children ($F_{1,72}$=5.822, male 95% CI 12.8-19, female 95% CI 9.79-13.4; *P*=.02). To avoid the possibility that the CES-DC and STAI-CH self-report scales was too difficult for 1 child aged <5 years, we conducted the same analysis without that child and found that the effect of age on the depression symptoms still showed a strong trend but was not significant ($F_{1,71}$=3.495; *P*=.07), while the gender effect was still significant ($F_{1,71}$=4.68; *P*=.03). Self-reported anxiety and parent-reported ADHD levels were not significantly different across age and gender (anxiety: age effect $F_{1,71}$=0.628; *P*=.43 and gender effect $F_{1,71}$=0.009; *P*=.93; ADHD: age effect $F_{1,70}$=0.719; *P*=.4 and gender effect $F_{1,70}$=0.001; *P*=.99).

Emotion-Related Effects in Emotional Cognitive Task Performance

We found a strong emotion-related effect on children's attention in the eFlanker (Figure 2A). First, RT was significantly higher in both positive (95% CI 1330-1469 ms) and negative (95% CI

XSL•FO
RenderX

1329-1463 ms) conditions compared to the neutral (95% CI 1167-1290 ms) condition (positive vs neutral t_{68}=11.404; *P*<.001; negative vs neutral: t_{67}=9.821; *P*<.001). Similarly, accuracy was lower for the 2 emotional conditions (positive: 95% CI 0.961-0.975; negative: 95% CI 0.958-0.972) than the neutral condition (95% CI 0.974-0.986; positive vs neutral: t_{63}=3.414; *P*=.001; negative vs neutral: t_{65}=5.859; *P*<.001), indicating that children struggled to suppress the distraction induced by emotional stimuli. However, there were no significant differences between congruent and incongruent conditions in both RT (congruent: 95% CI 1272-1402 ms; incongruent: 95% CI 1284-1414 ms) and accuracy (congruent: 95% CI 0.964-0.976; incongruent: 95% CI 0.963-0.976) on the eFlanker (Figure 2B; RT: t_{68}=0.265; *P*=.79; accuracy: t_{64}=0.331; *P*=.74). This lack of a significant congruency effect may be attributed to the dominant effects of emotional face stimuli, as reported in a previous study [56]. In the eGoNoGo, children made significantly more errors in the positive condition (positive: 95% CI 0.898-0.932; negative: 95% CI 0.951-0.97; t_{59}=4.776; *P*<.001; Figure 2C), which was possibly a result of failure to inhibit a response to the rewarding stimuli. RT was not different between positive and negative conditions (positive: 95% CI 1175-1300 ms; negative: 95% CI 1170-1273 ms). Similarly, in the eStroop, accuracy in the positive condition (95% CI 0.956-0.983) was significantly lower compared to the neutral condition (95% CI 0.975-0.994; t_{67}=2.6; *P*=.01; Figure 2D), while RT was not different across emotion (positive: 95% CI 793-861 ms; negative: 95% CI 812-886 ms) and neutral conditions (95% CI 800-872 ms).

Figure 2. Behavioral results (reaction time [RT] and accuracy) from all participants in the 3 emotional cognitive tasks. (A) RT (left) and accuracy (right) across 3 emotional conditions (positive [Pos]; neutral [Neu]; and negative [Neg]) in the emotional Flanker task (eFlanker). (B) RT and accuracy in congruent (Cong) and incongruent (Incong) conditions in the eFlanker. (C) RT and accuracy across positive and negative conditions in the emotional Go/No-Go task (eGoNoGo). (D) RT and accuracy across positive, neutral, and negative conditions in the emotional Stroop task (eStroop). Error bars indicate SE. ns: not significant. **P*<.05, ***P*<.01, ****P*<.001.

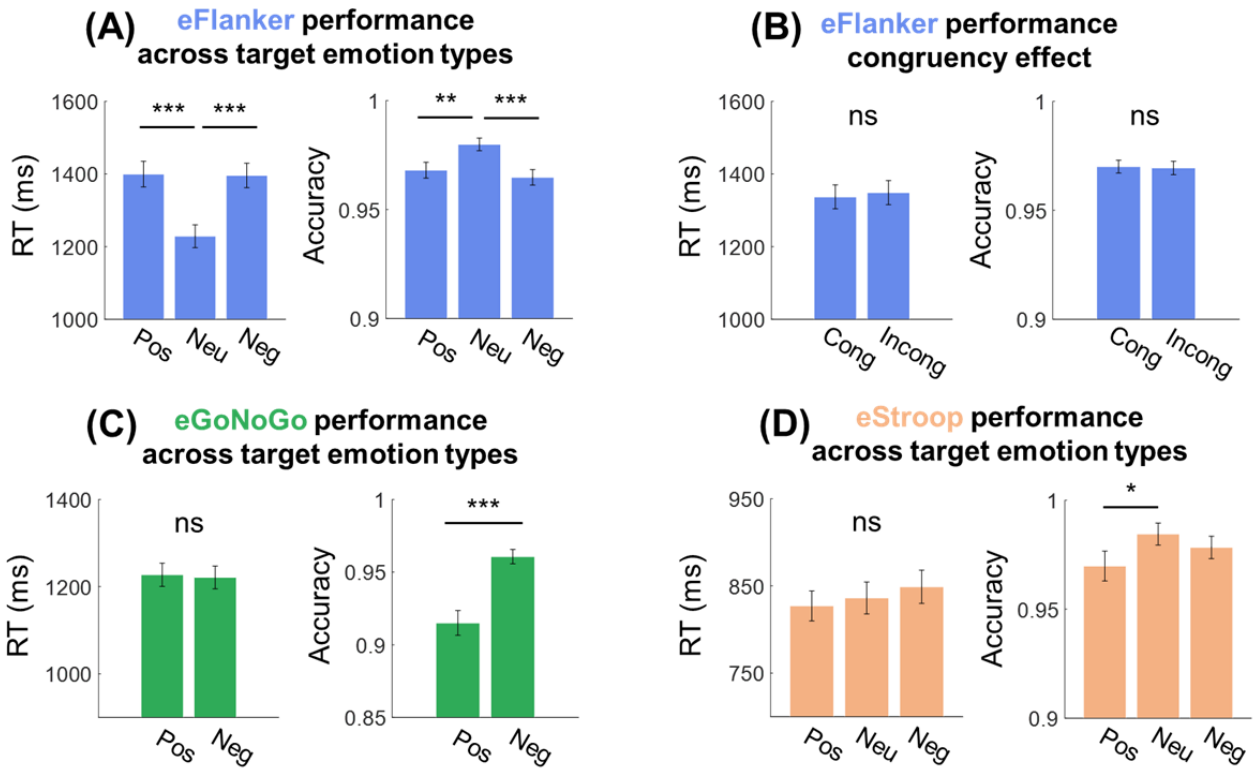

Emotional Cognitive Task Performance Across Development

ANCOVA tests were conducted on general task performance measured by RT and accuracy with gender as a fixed factor and age as a covariate. General task performance in all tasks was strongly correlated with age, confirming an improvement in attention-related functions between age 4 and 10 years (eFlanker RT: $F_{1,66}$=50.817; *P*<.001; accuracy: $F_{1,66}$=22.349; *P*<.001; eGoNoGo RT: $F_{1,63}$=19.513; *P*<.001; accuracy: $F_{1,62}$=8.281; *P*=.005; eStroop RT: $F_{1,62}$=16.585; *P*<.001; accuracy: $F_{1,61}$=5.879; *P*=.02; see Pearson correlation coefficients in Figures 3A-3D). We found a faster RT in male children in the eStroop RT (male: 95% CI 765-843 ms, female: 95% CI 838-949 ms; $F_{1,62}$=4.933; *P*=.03), but no gender differences were found in the other behavioral measures.

XSL•FO
RenderX

Figure 3. Developmental changes in overall performance, congruency effect, and emotional sensitivity were examined using a cross-sectional dataset. (A-D) Scatter plots showing significant relationships between age and performance in the 3 emotional cognitive tasks. (E-H) The congruency effect in the emotional Flanker task (eFlanker) and emotional sensitivity in reaction time (RT) across all tasks were not correlated with age. Cong: congruent; eGoNoGo: emotional Go/No-Go task; Emo: emotional; eStroop: emotional Stroop task; Incong: incongruent; Neg: negative; Neu: neutral; Pos: positive.

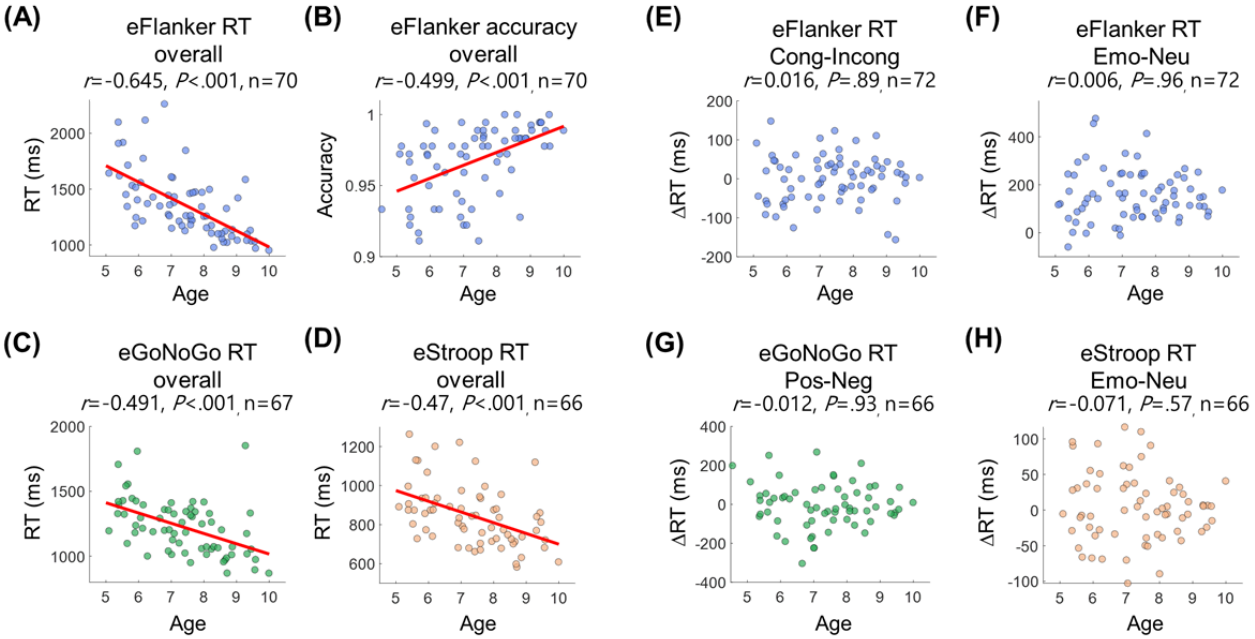

The congruency effect (ie, RT and accuracy differences between congruent and incongruent conditions) or performance difference between emotional and neutral targets in the eFlanker did not change in this age range (Figures 3E and 3F). Additionally, the RT or accuracy difference between positive and negative conditions in the eGoNoGo did not show any changes across age (Figure 3G). Similarly, the difference in eStroop performance (RT and accuracy) between the emotional and neutral conditions was not correlated with age (Figure 3H). None of these behavioral indices showed between-gender differences.

Relationship Between Emotional Cognitive Task Performance and Mental Health Measures

Among the children recruited for this study, 28 (36%) children (CES-DC >15 [57]) were in the high-risk group of depression based on the CES-DC scale (Figure 4A). Depressed mental health was associated with emotional processing during the eFlanker, as RT difference between emotion and neutral target conditions was higher for children with a higher depression scale (Pearson correlation between CES-DC and ΔRT emotional-neutral: r=0.281; P=.02 after controlling for age; Table 1). In addition, the congruency effect on accuracy (ie, a higher accuracy in the congruent condition compared to the incongruent condition) in the eFlanker was reduced in children with higher CES-DC (Figure 4B; r=–0.255; P=.03; n=70).

Figure 4. Relationship between task performance and mental health/attention-deficit/hyperactivity disorder (ADHD) scales. (A) Distribution of children's self-reported depression scales (Center for Epidemiological Studies Depression Scale for Children [CES-DC]). (B) Correlation between the CES-DC scale and the congruency effect on the emotional Flanker task (eFlanker) accuracy after covarying out age. A lower congruency effect was observed in children with a higher CES-DC level. (C) Distribution of children's self-reported anxiety scales (State-Trait Anxiety Inventory for Children [STAI-CH]). (D) Lower eFlanker accuracy in the negative (Neg) condition was associated with a higher STAI-CH level. (E) Distribution of children's ADHD scales, based on parent reports (Korean ADHD Rating Scale [K-ARS]). (F) A lower eFlanker accuracy in the emotional (Emo) condition was observed in children with a higher K-ARS level.

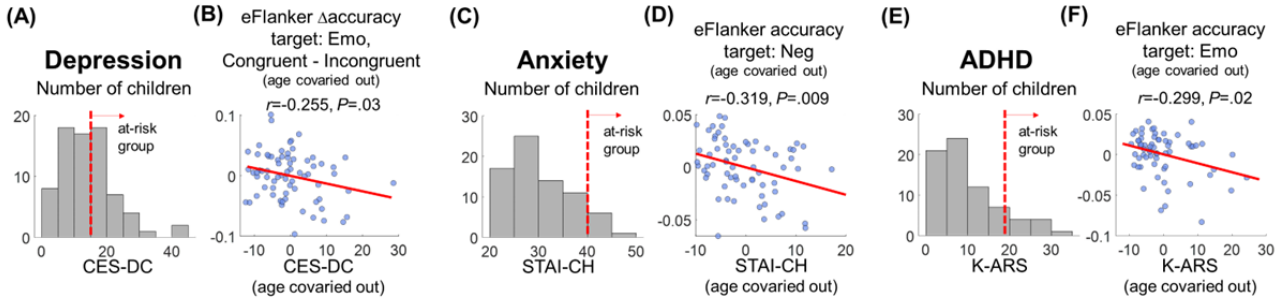

XSL•FO
RenderX

Table 1. Behavioral indices derived from the emotional cognitive tasks, grouped according to principal component analysis (PCA), with coefficients for E-score calculation. Only indices that showed a significant Pearson correlation with mental health measures (Center for Epidemiological Studies Depression Scale for Children [CES-DC] and State-Trait Anxiety Inventory for Children [STAI-CH]) or attention-deficit/hyperactivity disorder symptoms (Korean ADHD Rating Scale [K-ARS]) after covarying out age or overall task reaction time/accuracy on each task are presented. The complete lists are provided in Table S1 in Multimedia Appendix 1.

Group	Task	Feature	PCA coefficient	Age, *r*	CES-DC, *r*[a]	STAI-CH, *r*[a]	K-ARS, *r*[a]
1	eFlanker[b]	Overall reaction time	0.201	–0.645[c]	0.106	–0.019	–0.131
1	eFlanker	Reaction time (negative target, neutral distractor)	0.188	–0.714[c]	0.237[c]	0.113	–0.101
1	eFlanker	Reaction time (negative target, negative distractor)	0.192	–0.667[c]	0.262[c]	–0.094	–0.062
1	eFlanker	Reaction time (emotional target, congruent distractor)	0.210	–0.632[c]	0.293[c]	0.034	–0.028
1	eGoNo-Go[d]	Overall reaction time	0.146	–0.491[c]	0.176	0.132	–0.119
1	eStroop[e]	Overall reaction time	0.137	–0.47[c]	0.039	0.174	–0.177
2	eFlanker	Reaction time difference (emotional vs neutral targets; all distractors)	0.267	0.006	0.282[c]	0.158	–0.053
2	eFlanker	Accuracy (negative target, negative distractor)	–0.211	0.232[c]	–0.397[c]	–0.354[c]	–0.108
2	eFlanker	Accuracy difference (positive vs neutral targets; same-valence distractors)	–0.180	0.105	–0.263[c]	–0.186	–0.090
2	eFlanker	Accuracy difference (negative vs neutral targets; same-valence distractors)	–0.220	0.015	–0.235[c]	–0.232	–0.059
2	eFlanker	Accuracy difference (emotional-congruent vs neutral-neutral trials)	–0.250	0.131	–0.303[c]	–0.223	–0.053
3	eFlanker	Overall accuracy	0.285	0.499[c]	–0.013	–0.167	–0.255[c]
3	eFlanker	Accuracy (emotional target, incongruent distractor)	0.273	0.390[c]	0.143	–0.014	–0.329[c]
3	eFlanker	Accuracy (positive target, incongruent distractor)	0.274	0.344[c]	–0.060	–0.003	–0.383[c]
3	eFlanker	Accuracy (negative target, incongruent distractor)	0.224	0.381[c]	–0.096	–0.218	–0.29[c]
3	eFlanker	Accuracy difference (negative targets; incongruent vs congruent distractors)	0.062	0.009	0.129	0.220	–0.249[c]
3	eFlanker	Accuracy difference (emotional targets; incongruent vs congruent distractors)	0.146	0.204	0.257[c]	0.218	–0.142
3	eFlanker	Accuracy (all targets, incongruent distractor)	0.297	0.5[c]	0.034	–0.108	–0.333[c]
3	eFlanker	Accuracy (emotional targets)	0.287	0.432[c]	–0.071	–0.125	–0.322[c]
3	eGoNo-Go	Overall accuracy	0.188	0.333[c]	–0.078	0.036	–0.299[c]
3	eGoNo-Go	Accuracy (positive target)	0.139	0.423[c]	–0.113	–0.018	–0.292[c]
3	eStroop	Overall accuracy	0.165	0.292[c]	0.072	0.091	–0.145

[a]Age was covaried when calculating the Pearson correlation coefficients.

[b]eFlanker: emotional Flanker task.

[c]Indicates a statistical significance with *P* value of <.05.

[d]eGoNoGo: emotional Go/No-Go task.

[e]eStroop: emotional Stroop task.

Compared to depression, the number of children in the moderate to high anxiety group was lower (n=7, 9%; Figure 4C; STAI-CH >39 [58]). Accuracy on the eFlanker in the negative condition was the only behavioral measure correlated with the STAI-CH scale (Figure 4D; *r*=0.319; *P*=.009 after controlling for age; n=67). This may be due to a lower number of children with high anxiety, resulting in less clear across-individual contrasts for each behavioral index.

Lastly, a total of 10 (12.8%) children (K-ARS >18 [59]) were in the high-risk group for ADHD (Figure 4E). These children showed less accurate performance during the eFlanker and

XSL•FO
RenderX

eGoNoGo, especially in the emotional condition, compared to children with lower K-ARS scores (Figure 4F; r=–0.299; P=.02 after controlling for age; n=64).

Extracting "E(motional)-Scores" Using Behavioral Indices From the Emotional Cognitive Tasks

Given the nature of the tasks and the age of the participants, each behavioral index showed high variance across trials. In addition, we observed high redundancy as many behavioral measures were highly correlated with one another. Therefore, we opted to extract the common pattern spread over many behavioral indices, while maintaining the amount of information. We applied dimensionality reduction to the behavioral indices from the 3 emotional cognitive tasks to extract condensed information of individual performance by reducing repetitiveness (Figure 5A). Behavioral features (D=82) were divided into 3 groups by applying the k-means clustering method (Figure 5B), and E-scores were calculated by extracting the first PC of behavioral indices from each group (Figure 5C). We confirmed the robustness of the clustering using leave-one-out cross-validation, which assigned each excluded sample to the nearest group center and yielded labels identical to the original grouping.

Figure 5. Grouping of features (general reaction time [RT] and accuracy, the effects of emotion, and congruency effects) to calculate E-scores. (A) Dimensionality reduction of features based on their similarity across participants by principal component analysis (PCA). (B) The result of feature grouping after k-means clustering visualized on a 2D plane for the first 2 principal components (PCs). (C) Calculating E-scores using the first PC from each feature group. (D) Relationship between children's individual characteristics (ie, age, depression, anxiety, and attention-deficit/hyperactivity disorder [ADHD] scales) and E-scores. Only statistically significant (P<.05) linear relationships are visualized with a best-fit line. (E-G) Correlation between E-scores and age. (H-J) Correlation between each E-score and mental health/ADHD scales after controlling for age. CES-DC: Center for Epidemiological Studies Depression Scale for Children; FDR: false discovery rate; K-ARS: Korean ADHD Rating Scale; STAI-CH: State-Trait Anxiety Inventory for Children.

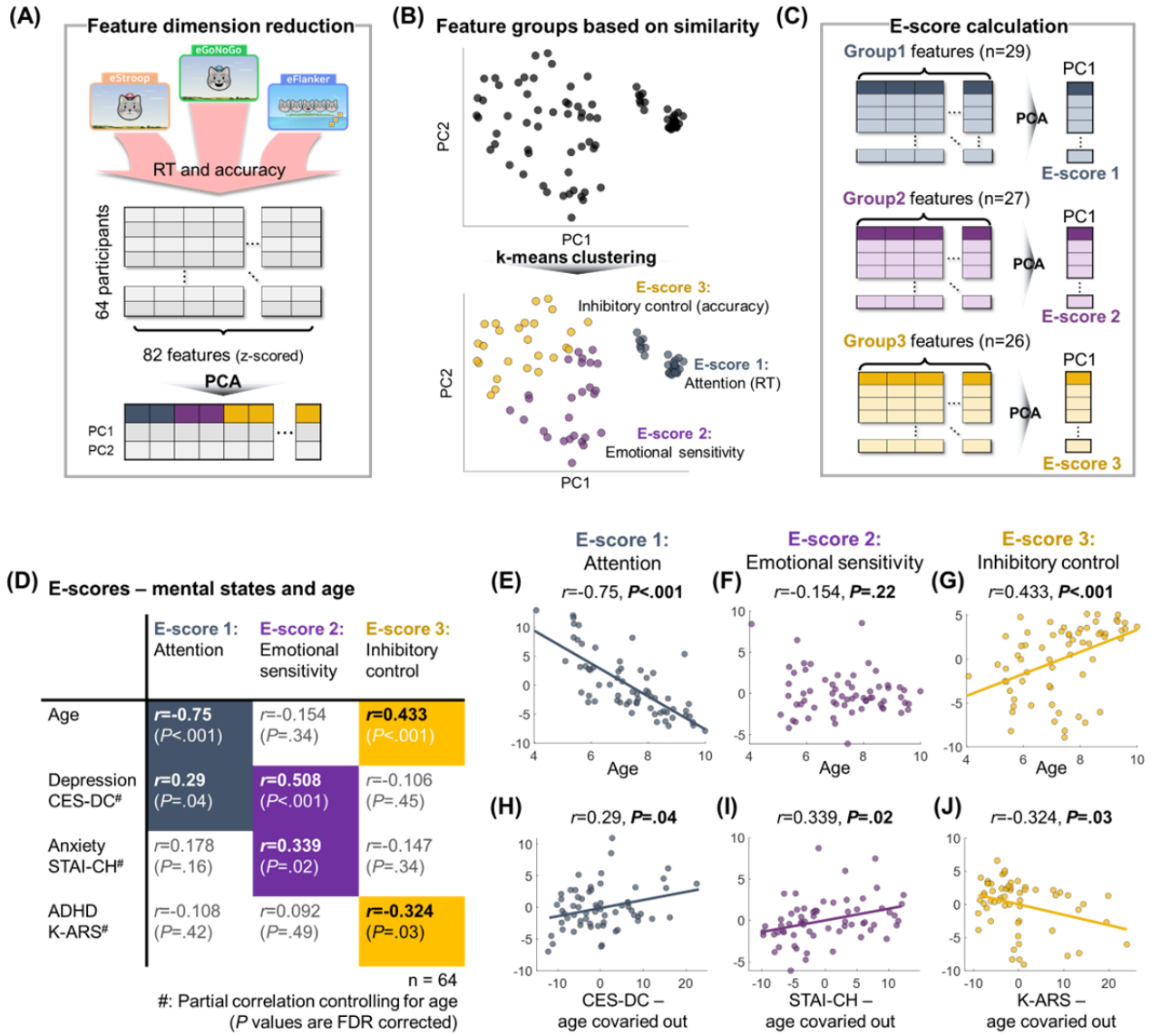

The first feature group contained RT features in various conditions across the eFlanker, eGoNoGo, and eStroop, thus representing the attention function (see Table S1 in Multimedia Appendix 1 for the whole list of behavioral features and coefficients for the first PC calculation). Positive PC coefficients of RT features for E-score 1 indicated that a higher E-score 1

XSL•FO
RenderX

represented decreased attention. Next, the second feature group represented complex interactions between emotion and attention. Coefficients of the first PC indicated that a higher E-score 2 was associated with a longer RT and lower accuracy, particularly in the emotional conditions, as well as higher congruency effects on RT in the eFlanker. Additionally, longer RT in the negative condition of the eStroop compared to the neutral condition also contributed to E-score 2. The third feature group represented accuracy on the tasks, as coefficients of accuracy features across various conditions in all 3 tasks positively contributed to E-score 3. Additionally, higher accuracy in the incongruent condition (ie, lower congruency effect) of the eFlanker also contributed to E-score 3.

Relationships Between E-Scores and Depression, Anxiety, or ADHD Symptoms

We evaluated the relevance of E-scores as individual emotional-cognitive characteristics by examining their correlations with mental health and ADHD survey scales (Figure 5D). First, ANCOVA tests were applied on E-scores with gender as a fixed factor and age as a continuous covariate. E-scores 1 and 3, which represented attention and executive function based on RT and accuracy, were significantly associated with age (E-score 1: $F_{1,60}$=78.404; P<.001; E-score 3: $F_{1,60}$=14.673; P<.001; all P values here and below are corrected for false discovery rate; see Pearson correlation coefficients in Figures 5E and 5G). In contrast, E-score 2, representing emotion-attention interactions, was not correlated with age (Figure 5F; $F_{1,60}$=1.493; P=.23). None of the E-scores showed significant between-gender differences. Self-reported depression level measured by the CES-DC was associated with E-score 1 (Figure 5H; r=0.29; P=.04), which replicated previous studies showing that children with depressive symptoms showed lower attention function represented by a longer RT [60]. In addition, a significant positive correlation between the CES-DC scale and E-score 2 (r=0.508; P<.001) indicated that children with higher depression were more sensitive to the emotional stimuli, resulting in longer RT and lower accuracy. E-score 2 was also positively correlated with the anxiety scale (ie, STAI-CH; Figure 5I; r=0.339; P=.02), showing similar cognitive characteristics induced by anxiety and depression. No differences in these association patterns were observed between the 2 collection methods (self-completed vs experimenter-guided; see Table S2 in Multimedia Appendix 1 for details). Lastly, E-score 3, which represented accuracy, was negatively correlated with the ADHD scale (ie, K-ARS; Figure 5J; r=–0.324; P=.03), suggesting that children with a higher risk for ADHD were more impulsive and inaccurate in their responses to emotional stimuli.

To further disentangle the effects of depression, anxiety, and ADHD on task performance (ie, E-scores), we conducted partial correlation analyses by controlling for the other scales and age (Table S3 in Multimedia Appendix 1). Every correlation remained significant except for E-score 2 and STAI-CH, which suggested that the E-score 2 does not reflect an anxiety-specific effect but rather overlapping characteristics shared between the anxiety and depression measures.

The dimensionality reduction technique revealed common and distinct behavioral patterns across the 3 emotional cognitive tasks. Extracting the common patterns contributed to reducing noise in the behavioral data and thus compensated for the short performance time of the tasks (10-15 minutes for each task), while the distinctiveness contributed to collecting various affective-cognitive profiles of children. Therefore, a range of behavioral indices was reduced to condense features, which were translated into a more robust, composite measure for individual affective-cognitive characteristics that is not susceptible to biases in a single measure. The results highlight the dissociability across depression, anxiety, and ADHD risk, given that each scale was correlated with different E-scores.

Discussion

Principal Findings

In this study, we aimed to investigate the distinct impacts of depression, anxiety, and ADHD on emotion recognition and attention in children aged between 4 and 10 years, using child-friendly gamified eFlanker, eGoNoGo, and eStroop. Our results highlight that various task performance measures related to emotion and cognition can be summarized into three distinctive components (attention, selective inhibition, and emotional sensitivity). Based on this, we introduced E-scores to incorporate meaningful information about the participant while minimizing noisy data distributed across the various features. The E-scores provided enhanced interpretability of behavioral task performance and were associated with mental health (anxiety and depression) or ADHD symptoms. The emotional cognitive assessments of our study are advantageous in that they allow monitoring of children's neurocognitive development, which may help address current limitations of measurement through self-report surveys.

Comparison to Prior Literature

The main novelty of this study lies in the development and validation of digital tasks that measure interactions between individual cognitive and affective characteristics and their associations with age-, mental health–, and ADHD-related measures within a unified framework. This approach offers advantages over previous studies that focused on single symptom domains [22-24], as it facilitates interpretation of children's behavioral outcomes while reducing potential confounds arising from comorbidity. Administering 3 digital cognitive tasks across multiple emotional conditions generated a wide range of behavioral indices; to address the resulting analytical complexity, we applied dimensionality reduction techniques to extract informative features across these indices. Together, these differences in study aims and methodological approach distinguish this work from prior studies.

As reported in many other studies [61], we observed a high correlation between depression and anxiety symptoms in our sample. Both anxiety and depression were associated with a larger emotional interference in attention functions (ie, RT increase and accuracy decrease in the emotional compared to the neutral conditions), typically higher for negative stimuli (Figure 4D). This is consistent with our previous study in which emotional bias served as a marker of both depression and anxiety [33]. Anxious and depressed individuals often show decreased inhibitory control [62] and higher emotional sensitivity [63],

XSL•FO
RenderX

and thus they have a higher chance of being distracted by emotional stimuli [33,64]. This pattern has been linked to a hyperactive amygdala, which is also associated with weaker top-down control and reduced activation in the dorsolateral PFC and ventromedial PFC [65,66].

ADHD risk measured by the K-ARS scale survey was independent of the depression or anxiety scales. Similarly, unlike depression or anxiety, ADHD was strongly associated with accuracy across various conditions in the 3 tasks. Children with ADHD have difficulty inhibiting irrelevant stimuli, reflecting deficits in selective attention and executive control [67]. The results in the eFlanker suggested that children with higher K-ARS scores showed reduced selective inhibition toward emotional stimuli (compared to neutral ones), which is consistent with their higher impulsivity [68,69]. Similarly, the lower accuracy observed in the eGoNoGo and eStroop likely resulted from disrupted working memory and executive control due to higher impulsivity. In particular, the general tendency for young children to respond to an incorrect positive stimulus in the eGoNoGo and eStroop are consistent with that interpretation. Interestingly, children's ADHD score was only associated with accuracy and not RT; this may be related to the fact that impulsive individuals tend to prioritize speed over accuracy, according to the well-known speed-accuracy trade-off, especially in the Flanker task [70].

Previously, we found that children's performance in the emotional cognitive tasks can be characterized as longer RT and lower accuracy compared to adults [33]. In this study, a cross-sectional investigation with a wider range of age showed that older children performed better than younger children, indicating that their general attention functions develop across childhood (4-10 years of age) with the maturation of the PFC [71]. Another characteristic of the emotional cognitive tasks is a general tendency for longer RT and lower accuracy in the emotional conditions. In particular, children exhibited more errors for positive stimuli than for neutral or negative stimuli, likely reflecting greater attentional bias toward positive cues, which can be associated with stronger approach motivation and heightened reward sensitivity [72]. However, these emotion-attention interactions remained unchanged across ages in our study. Previous studies reported that the emotional sensitivity of children is relatively higher until the age of around 10 years, when the amygdala reactivity decreases along with PFC maturation [73,74]. Thus, it is possible that all children in our study were significantly affected by emotional stimuli.

One difference between the results of our eFlanker with young children and our previous study [33] with adults is the reduction in the congruency effect. This discrepancy can be explained by developmental attention mechanisms that describe children's attentional bias toward more salient or emotional stimuli [7,8]. Congruency effects, characterized by longer RT and lower accuracy in the incongruent conditions, are induced by increased cognitive resources needed to resolve conflict in the incongruent condition and are typically accompanied by increased activity in the anterior cingulate cortex and other frontal regions [75]. However, as emotional targets are likely to attract more attention in children, flankers may be equally distracting in both congruent and incongruent conditions. Consistent with this interpretation, a reduced or absent congruency effect has also been reported in a previous study using a complex Flanker task with emotional faces as either target or distractor stimuli [56].

Strengths

The prevalence of childhood depression and ADHD has drastically increased over the past few years [76,77], and thus, a daily monitoring tool is essential for risk detection within the optimal time window for early intervention. We demonstrated that a simple emotional cognitive assessment tool can provide a stronger scientific basis for identifying affective-cognitive characteristics related to mental health and ADHD. Our approach highlights the efficacy of simple and easy, child-friendly, touchscreen-based tasks that are enjoyable even for young children, which is a critical factor for inducing accurate measurement of their attention, emotional sensitivity, and inhibitory function [78,79]. Furthermore, the average duration of gameplay was around 30 minutes, which is manageable and sustainable for repeated testing in children. These advantages of our tool enable an easy but quantifiable measurement of cognitive symptoms related to children's mental health and ADHD symptoms (eg, during school activities) for parents and clinicians.

Limitations and Future Directions

One limitation of this study is that it is based on a subclinical sample of children and that the mental health and ADHD measures were based on self-report and parent surveys. To enhance its applicability and generalizability to patient populations, future studies should validate the results using a formal clinical diagnosis. Nevertheless, in our previous study [33], we demonstrated that the relationships between task performance and mental health observed in healthy adults were also present in individuals diagnosed with depression. Second, parent-reported assessments of ADHD can also be subject to bias influenced by factors such as cultural context, parental mental health, and variability in the interpretation of symptoms. To mitigate this, we provided standardized instructions and guided administration to minimize variability in parental interpretation of symptom items. Third, while our age range was within the recommended period for the ADHD assessment, it may be relatively young for reliably assessing depression and anxiety, particularly for children aged <5 years. Therefore, our prediction based on behavioral performance may indicate a risk for later developmental stages. Fourth, cultural differences could influence emotion-cognition interactions in young children. Although this factor was minimized in this study, as all participants were Korean, cultural considerations may need to be considered when applying the tool internationally. Fifth, some children lost interest in the task during testing, which limited our ability to accurately assess their affective-cognitive characteristics. Including these data would likely weaken the observed associations between behavioral indices and mental health or ADHD survey scales; therefore, these cases were treated as missing data, resulting in a reduced sample size. For future applications in daily-life, such as deployment as a mobile or web-based application, online algorithms to detect loss of engagement will be necessary. In addition, appropriate real-time outlier detection and data-quality control procedures will be

XSL•FO
RenderX

required to reliably capture users' cognitive and affective characteristics. Lastly, a longitudinal study would be useful to investigate the effectiveness of the tool in assessing and predicting mental health and ADHD over an extended period, into adolescence.

Conclusion

Our gamified digital monitoring tool offers a practical and objective way to connect individual behavioral profiles with their affective and cognitive health, particularly for young children who may benefit greatly from early intervention but are difficult to accurately assess using traditional clinical interviews. While requiring further standardization and a well-established protocol to ensure data privacy, the ability to perform a rapid assessment also makes this tool well-suited for practical use in a variety of settings.

Acknowledgments

We would like to thank Minwoo Kim for his help with creating the task. We also thank Yu Jin Rah, Gakyung Kim, and the Gyeonggi Children's Museum, and all of the children and families that participated in this study.

Funding

SAL received funding from the Creative-Pioneering Researchers Program through Seoul National University, SYEDU, Samsung Electronics Co, Ltd (A0426-20220108), the National Research Foundation of Korea (2023R1A2C2005587, 2021M3A9E4080780, RS-2023-00301976), and the Korea Basic Science Institute (National Research Facilities and Equipment Center; RS-2024-00435727). Funders did not play any role in the study design, data collection, and analysis.

Data Availability

The data that support the findings of this study are publicly available on GitHub [80].

Authors' Contributions

Conceptualization: SEP (lead) and SAL (equal)
Data curation: JC
Formal analysis: SEP (lead), JC (supporting)
Funding acquisition: SAL
Methodology: SEP
Project administration: SAL
Supervision: SAL
Validation: SEP (lead) and SAL (equal)
Visualization: SEP
Writing – original draft: SEP
Writing – review and editing: SAL (lead) and JC (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional figure and tables.
[PDF File (Adobe PDF File), 414 KB - jmir_v28i1e86286_app1.pdf]

References

1. Muris P. Relation of attention control and school performance in normal children. Percept Mot Skills 2006;102(1):78-80. [doi: 10.2466/pms.102.1.78-80] [Medline: 16671601]
2. Kessler RC, Amminger GP, Aguilar-Gaxiola S, Alonso J, Lee S, Ustün TB. Age of onset of mental disorders: a review of recent literature. Curr Opin Psychiatry 2007;20(4):359-364 [FREE Full text] [doi: 10.1097/YCO.0b013e32816ebc8c] [Medline: 17551351]
3. Gaffrey MS, Barch DM, Luby JL, Petersen SE. Amygdala functional connectivity is associated with emotion regulation and amygdala reactivity in 4- to 6-year-olds. J Am Acad Child Adolesc Psychiatry 2021;60(1):176-185 [FREE Full text] [doi: 10.1016/j.jaac.2020.01.024] [Medline: 32119912]
4. Galvan A. Adolescent development of the reward system. Front Hum Neurosci 2010;4:6 [FREE Full text] [doi: 10.3389/neuro.09.006.2010] [Medline: 20179786]

XSL•FO
RenderX

5. Wu M, Kujawa A, Lu LH, Fitzgerald DA, Klumpp H, Fitzgerald KD, et al. Age-related changes in amygdala-frontal connectivity during emotional face processing from childhood into young adulthood. Hum Brain Mapp 2016;37(5):1684-1695 [FREE Full text] [doi: 10.1002/hbm.23129] [Medline: 26931629]
6. Hare TA, Tottenham N, Galvan A, Voss HU, Glover GH, Casey BJ. Biological substrates of emotional reactivity and regulation in adolescence during an emotional go-nogo task. Biol Psychiatry 2008;63(10):927-934 [FREE Full text] [doi: 10.1016/j.biopsych.2008.03.015] [Medline: 18452757]
7. Silvers JA, McRae K, Gabrieli JDE, Gross JJ, Remy KA, Ochsner KN. Age-related differences in emotional reactivity, regulation, and rejection sensitivity in adolescence. Emotion 2012;12(6):1235-1247 [FREE Full text] [doi: 10.1037/a0028297] [Medline: 22642356]
8. Arain M, Haque M, Johal L, Mathur P, Nel W, Rais A, et al. Maturation of the adolescent brain. Neuropsychiatr Dis Treat 2013;9:449-461 [FREE Full text] [doi: 10.2147/NDT.S39776] [Medline: 23579318]
9. Ahmed SP, Bittencourt-Hewitt A, Sebastian CL. Neurocognitive bases of emotion regulation development in adolescence. Dev Cogn Neurosci 2015;15:11-25 [FREE Full text] [doi: 10.1016/j.dcn.2015.07.006] [Medline: 26340451]
10. Michielsen M, Comijs HC, Semeijn EJ, Beekman ATF, Deeg DJH, Sandra Kooij JJ. The comorbidity of anxiety and depressive symptoms in older adults with attention-deficit/hyperactivity disorder: a longitudinal study. J Affect Disord 2013;148(2-3):220-227. [doi: 10.1016/j.jad.2012.11.063] [Medline: 23267726]
11. Kennedy SH. Core symptoms of major depressive disorder: relevance to diagnosis and treatment. Dialogues Clin Neurosci 2008;10(3):271-277 [FREE Full text] [doi: 10.31887/DCNS.2008.10.3/shkennedy] [Medline: 18979940]
12. Tsourtos G, Thompson JC, Stough C. Evidence of an early information processing speed deficit in unipolar major depression. Psychol Med 2002;32(2):259-265. [doi: 10.1017/s0033291701005001] [Medline: 11866321]
13. Gotlib IH, Joormann J. Cognition and depression: current status and future directions. Annu Rev Clin Psychol 2010;6:285-312 [FREE Full text] [doi: 10.1146/annurev.clinpsy.121208.131305] [Medline: 20192795]
14. Kujawa A, MacNamara A, Fitzgerald KD, Monk CS, Phan KL. Enhanced neural reactivity to threatening faces in anxious youth: evidence from event-related potentials. J Abnorm Child Psychol 2015;43(8):1493-1501 [FREE Full text] [doi: 10.1007/s10802-015-0029-4] [Medline: 25943264]
15. Eldar S, Yankelevitch R, Lamy D, Bar-Haim Y. Enhanced neural reactivity and selective attention to threat in anxiety. Biol Psychol 2010;85(2):252-257. [doi: 10.1016/j.biopsycho.2010.07.010] [Medline: 20655976]
16. Cisler JM, Koster EHW. Mechanisms of attentional biases towards threat in anxiety disorders: an integrative review. Clin Psychol Rev 2010;30(2):203-216 [FREE Full text] [doi: 10.1016/j.cpr.2009.11.003] [Medline: 20005616]
17. Barkley RA, Wasserstein J. ADHD and the nature of self-control. J Cogn Psychother 2000;14(1):111-113. [doi: 10.1891/0889-8391.14.1.111]
18. Iaboni F, Douglas VI, Ditto B. Psychophysiological response of ADHD children to reward and extinction. Psychophysiology 1997;34(1):116-123. [doi: 10.1111/j.1469-8986.1997.tb02422.x] [Medline: 9009815]
19. Patros CH, Alderson RM, Kasper LJ, Tarle SJ, Lea SE, Hudec KL. Choice-impulsivity in children and adolescents with attention-deficit/hyperactivity disorder (ADHD): a meta-analytic review. Clin Psychol Rev 2016;43:162-174. [doi: 10.1016/j.cpr.2015.11.001] [Medline: 26602954]
20. Wang K, Varma DS, Prosperi M. A systematic review of the effectiveness of mobile apps for monitoring and management of mental health symptoms or disorders. J Psychiatr Res 2018;107:73-78. [doi: 10.1016/j.jpsychires.2018.10.006] [Medline: 30347316]
21. Wiley K, Robinson R, Mandryk RL. The making and evaluation of digital games used for the assessment of attention: systematic review. JMIR Serious Games 2021;9(3):e26449 [FREE Full text] [doi: 10.2196/26449] [Medline: 34383674]
22. Peijnenborgh JCAW, Hurks PPM, Aldenkamp AP, van der Spek ED, Rauterberg GWM, Vles JSH, et al. A study on the validity of a computer-based game to assess cognitive processes, reward mechanisms, and time perception in children aged 4-8 years. JMIR Serious Games 2016;4(2):e15 [FREE Full text] [doi: 10.2196/games.5997] [Medline: 27658428]
23. Mandryk RL, Birk MV, Vedress S, Wiley K, Reid E, Berger P, et al. Remote assessment of depression using digital biomarkers from cognitive tasks. Front Psychol 2021;12:767507 [FREE Full text] [doi: 10.3389/fpsyg.2021.767507] [Medline: 34975656]
24. Voinescu A, Petrini K, Stanton Fraser D, Lazarovicz R, Papavă I, Fodor LA, et al. The effectiveness of a virtual reality attention task to predict depression and anxiety in comparison with current clinical measures. Virtual Real 2021;27(1):119-140. [doi: 10.1007/s10055-021-00520-7]
25. O'Laughlin KD, Cheng BH, Volponi JJ, Lorentz JDA, Obregon SA, Younger JW, et al. Validation of an adaptive assessment of executive functions (Adaptive Cognitive Evaluation-Explorer): longitudinal and cross-sectional analyses of cognitive task performance. J Med Internet Res 2025;27:e60041 [FREE Full text] [doi: 10.2196/60041] [Medline: 40258271]
26. Song H, Yi D, Park H. Validation of a mobile game-based assessment of cognitive control among children and adolescents. PLoS One 2020;15(3):e0230498 [FREE Full text] [doi: 10.1371/journal.pone.0230498] [Medline: 32196531]
27. Beck AT. The evolution of the cognitive model of depression and its neurobiological correlates. Am J Psychiatry 2008;165(8):969-977. [doi: 10.1176/appi.ajp.2008.08050721] [Medline: 18628348]
28. Robinson ESJ, Roiser JP. Affective biases in humans and animals. Curr Top Behav Neurosci 2016. [doi: 10.1007/7854_2015_5011] [Medline: 27311761]

XSL•FO
RenderX

29. Roiser JP, Elliott R, Sahakian BJ. Cognitive mechanisms of treatment in depression. Neuropsychopharmacology 2012;37(1):117-136 [FREE Full text] [doi: 10.1038/npp.2011.183] [Medline: 21976044]
30. Snyder HR, Miyake A, Hankin BL. Advancing understanding of executive function impairments and psychopathology: bridging the gap between clinical and cognitive approaches. Front Psychol 2015;6:328 [FREE Full text] [doi: 10.3389/fpsyg.2015.00328] [Medline: 25859234]
31. Mogg K, Salum GA, Bradley BP, Gadelha A, Pan P, Alvarenga P, et al. Attention network functioning in children with anxiety disorders, attention-deficit/hyperactivity disorder and non-clinical anxiety. Psychol Med 2015;45(12):2633-2646 [FREE Full text] [doi: 10.1017/S0033291715000586] [Medline: 26234806]
32. Mrug S, Madan A, Windle M. Emotional desensitization to violence contributes to adolescents' violent behavior. J Abnorm Child Psychol 2016;44(1):75-86 [FREE Full text] [doi: 10.1007/s10802-015-9986-x] [Medline: 25684447]
33. Park SE, Chung J, Lee J, Kim MJ, Kim J, Jeon HJ, et al. Digital assessment of cognitive-affective biases related to mental health. PLOS Digit Health 2024;3(8):e0000595. [doi: 10.1371/journal.pdig.0000595] [Medline: 39208388]
34. Abdul Rahman A, Wiebe SA. Valence matters: an electrophysiological study on how emotions influence cognitive performance in children. Dev Psychobiol 2019;61(2):290-303. [doi: 10.1002/dev.21813] [Medline: 30548999]
35. Vaughan AM, Perone S. Higher fearfulness in 4- to 6-year-old children related to better performance on an emotional than neutral stroop task. J Cogn Dev 2025;26(1):1-20. [doi: 10.1080/15248372.2025.2553532]
36. Ezpeleta L, Navarro JB, de la Osa N, Penelo E, Trepat E, Martin V, et al. Attention to emotion through a go/no-go task in children with oppositionality and callous-unemotional traits. Compr Psychiatry 2017;75:35-45 [FREE Full text] [doi: 10.1016/j.comppsych.2017.02.004] [Medline: 28301801]
37. Dudeney J, Sharpe L, Hunt C. Attentional bias towards threatening stimuli in children with anxiety: a meta-analysis. Clin Psychol Rev 2015;40:66-75. [doi: 10.1016/j.cpr.2015.05.007] [Medline: 26071667]
38. Harrison AJ, Gibb BE. Attentional biases in currently depressed children: an eye-tracking study of biases in sustained attention to emotional stimuli. J Clin Child Adolesc Psychol 2015;44(6):1008-1014 [FREE Full text] [doi: 10.1080/15374416.2014.930688] [Medline: 25010145]
39. Murray AL, Russell A, Calderón Alfaro FA. Early emotion regulation developmental trajectories and ADHD, internalizing, and conduct problems symptoms in childhood. Dev Psychopathol 2025;37(3):1474-1481. [doi: 10.1017/S0954579424001263] [Medline: 39282723]
40. Casey BJ, Jones RM, Somerville LH. Braking and accelerating of the adolescent brain. J Res Adolesc 2011;21(1):21-33. [doi: 10.1111/j.1532-7795.2010.00712.x] [Medline: 21475613]
41. Dick B, Ferguson BJ. Health for the world's adolescents: a second chance in the second decade. J Adolesc Health 2015;56(1):3-6. [doi: 10.1016/j.jadohealth.2014.10.260] [Medline: 25530601]
42. Marquardt L, Eichele H, Lundervold AJ, Haavik J, Eichele T. Event-related-potential (ERP) correlates of performance monitoring in adults with attention-deficit hyperactivity disorder (ADHD). Front Psychol 2018;9:485 [FREE Full text] [doi: 10.3389/fpsyg.2018.00485] [Medline: 29706908]
43. Pe ML, Vandekerckhove J, Kuppens P. A diffusion model account of the relationship between the emotional flanker task and rumination and depression. Emotion 2013;13(4):739-747 [FREE Full text] [doi: 10.1037/a0031628] [Medline: 23527499]
44. Zetsche U, D'Avanzato C, Joormann J. Depression and rumination: relation to components of inhibition. Cogn Emot 2012;26(4):758-767. [doi: 10.1080/02699931.2011.613919] [Medline: 21970297]
45. Gutiérrez-Cobo MJ, Cabello R, Fernández-Berrocal P. The three models of emotional intelligence and performance in a hot and cool go/no-go task in undergraduate students. Front Behav Neurosci 2017;11:33 [FREE Full text] [doi: 10.3389/fnbeh.2017.00033] [Medline: 28275343]
46. Schulz KP, Fan J, Magidina O, Marks DJ, Hahn B, Halperin JM. Does the emotional go/no-go task really measure behavioral inhibition? Convergence with measures on a non-emotional analog. Arch Clin Neuropsychol 2007;22(2):151-160 [FREE Full text] [doi: 10.1016/j.acn.2006.12.001] [Medline: 17207962]
47. Spielberger CD. State-Trait Anxiety Inventory for Children. Palo Alto, CA: Consulting Psychologist Press; 1973.
48. Weissman MM, Orvaschel H, Padian N. Children's symptom and social functioning self-report scales. Comparison of mothers' and children's reports. J Nerv Ment Dis 1980;168(12):736-740. [doi: 10.1097/00005053-198012000-00005] [Medline: 7452212]
49. DuPaul GJ, Power TJ, Anastopoulos AD, Reid R. ADHD Rating Scale-5 for Children and Adolescents: Checklists, Norms, and Clinical Interpretation. New York: The Guilford Press; 2016.
50. So YK, Noh JS, Kim YS, Ko SG, Koh YJ. The reliability and validity of Korean parent and teacher ADHD rating scale. J Korean Neuropsychiatr Assoc 2016;41(2):283-289. [doi: 10.4306/JKNPA.2002.41.2.283]
51. Abdi H, Williams LJ. Principal component analysis. Wiley Interdiscip Rev Comput Stat 2010;2(4):433-459. [doi: 10.1002/(issn)1939-0068]
52. Park SE, Laxpati NG, Gutekunst CA, Connolly MJ, Tung J, Berglund K, et al. A machine learning approach to characterize the modulation of the hippocampal rhythms via optogenetic stimulation of the medial septum. Int J Neural Syst 2019;29(10):1950020. [doi: 10.1142/S0129065719500205] [Medline: 31505977]
53. Arthur D, Vassilvitskii S. K-means++: the advantages of careful seeding. 2007 Presented at: SODA '07: Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms; 2007 January 7-9; New Orleans Louisiana p. 1027-1035.

XSL•FO
RenderX

54. Park SE, Lee J, Park JH, Hwang MH, Lee SA. Cognotyping by what-where-when retrieval reveals the potential role of temporal memory and its neural correlates in understanding individual differences across aging and alzheimer disease. J Cogn Neurosci 2023;35(11):1773-1787. [doi: 10.1162/jocn_a_02039] [Medline: 37584602]
55. Rykov A, De Amorim RC, Makarenkov V, Mirkin B. Inertia-based indices to determine the number of clusters in k-means: an experimental evaluation. IEEE Access 2024;12:11761-11773. [doi: 10.1109/access.2024.3350791]
56. Montalti M, Mirabella G. Unveiling the influence of task-relevance of emotional faces on behavioral reactions in a multi-face context using a novel flanker-go/no-go task. Sci Rep 2023;13(1):20183. [doi: 10.1038/s41598-023-47385-1] [Medline: 37978229]
57. Fendrich M, Weissman MM, Warner V. Screening for depressive disorder in children and adolescents: validating the center for epidemiologic studies depression scale for children. Am J Epidemiol 1990;131(3):538-551. [doi: 10.1093/oxfordjournals.aje.a115529] [Medline: 2301363]
58. Julian LJ. Measures of anxiety. Arthritis Care Res (Hoboken) 2011;63(suppl 11):S467-S472. [doi: 10.1002/acr.20561.Measures]
59. Cutler AJ, Brams M, Bukstein O, Mattingly G, McBurnett K, White C, et al. Response/remission with guanfacine extended-release and psychostimulants in children and adolescents with attention-deficit/hyperactivity disorder. J Am Acad Child Adolesc Psychiatry 2014;53(10):1092-1101. [doi: 10.1016/j.jaac.2014.08.001] [Medline: 25245353]
60. Wagner S, Müller C, Helmreich I, Huss M, Tadić A. A meta-analysis of cognitive functions in children and adolescents with major depressive disorder. Eur Child Adolesc Psychiatry 2015;24(1):5-19. [doi: 10.1007/s00787-014-0559-2] [Medline: 24869711]
61. McElroy E, Fearon P, Belsky J, Fonagy P, Patalay P. Networks of depression and anxiety symptoms across development. J Am Acad Child Adolesc Psychiatry 2018;57(12):964-973 [FREE Full text] [doi: 10.1016/j.jaac.2018.05.027] [Medline: 30522742]
62. Snyder HR, Friedman NP, Hankin BL. Transdiagnostic mechanisms of psychopathology in youth: executive functions, dependent stress, and rumination. Cognit Ther Res 2019;43(5):834-851 [FREE Full text] [doi: 10.1007/s10608-019-10016-z] [Medline: 31551642]
63. Watson D. Differentiating the mood and anxiety disorders: a quadripartite model. Annu Rev Clin Psychol 2009;5:221-247. [doi: 10.1146/annurev.clinpsy.032408.153510] [Medline: 19327030]
64. Gibb BE, McGeary JE, Beevers CG. Attentional biases to emotional stimuli: key components of the RDoC constructs of sustained threat and loss. Am J Med Genet B Neuropsychiatr Genet 2016;171B(1):65-80. [doi: 10.1002/ajmg.b.32383] [Medline: 26369836]
65. Motzkin JC, Philippi CL, Wolf RC, Baskaya MK, Koenigs M. Ventromedial prefrontal cortex is critical for the regulation of amygdala activity in humans. Biol Psychiatry 2015;77(3):276-284. [doi: 10.1016/j.biopsych.2014.02.014] [Medline: 24673881]
66. Nejati V, Majdi R, Salehinejad MA, Nitsche MA. The role of dorsolateral and ventromedial prefrontal cortex in the processing of emotional dimensions. Sci Rep 2021;11(1):1971 [FREE Full text] [doi: 10.1038/s41598-021-81454-7] [Medline: 33479323]
67. Tenenbaum RB, Musser ED, Morris S, Ward AR, Raiker JS, Coles EK, et al. Response inhibition, response execution, and emotion regulation among children with attention-deficit/hyperactivity disorder. J Abnorm Child Psychol 2019;47(4):589-603 [FREE Full text] [doi: 10.1007/s10802-018-0466-y] [Medline: 30112596]
68. Rosen PJ, Factor PI. Emotional impulsivity and emotional and behavioral difficulties among children with ADHD: an ecological momentary assessment study. J Atten Disord 2015;19(9):779-793. [doi: 10.1177/1087054712463064] [Medline: 23172248]
69. Özbaran B, Kalyoncu T, Köse S. Theory of mind and emotion regulation difficulties in children with ADHD. Psychiatry Res 2018;270:117-122. [doi: 10.1016/j.psychres.2018.09.034] [Medline: 30245374]
70. Dickman SJ, Meyer DE. Impulsivity and speed-accuracy tradeoffs in information processing. J Pers Soc Psychol 1988;54(2):274-290. [doi: 10.1037//0022-3514.54.2.274] [Medline: 3346814]
71. Uytun MC. Development period of prefrontal cortex. In: Prefrontal Cortex. London: Intechopen; 2018.
72. Pool E, Brosch T, Delplanque S, Sander D. Attentional bias for positive emotional stimuli: a meta-analytic investigation. Psychol Bull 2016;142(1):79-106. [doi: 10.1037/bul0000026] [Medline: 26390266]
73. Gee DG, Humphreys KL, Flannery J, Goff B, Telzer EH, Shapiro M, et al. A developmental shift from positive to negative connectivity in human amygdala-prefrontal circuitry. J Neurosci 2013;33(10):4584-4593 [FREE Full text] [doi: 10.1523/JNEUROSCI.3446-12.2013] [Medline: 23467374]
74. Gee DG, Hanson C, Caglar LR, Fareri DS, Gabard-Durnam LJ, Mills-Finnerty C, et al. Experimental evidence for a child-to-adolescent switch in human amygdala-prefrontal cortex communication: a cross-sectional pilot study. Dev Sci 2022;25(4):e13238 [FREE Full text] [doi: 10.1111/desc.13238] [Medline: 35080089]
75. Borsa VM, Della Rosa PA, Catricalà E, Canini M, Iadanza A, Falini A, et al. Interference and conflict monitoring in individuals with amnestic mild cognitive impairment: a structural study of the anterior cingulate cortex. J Neuropsychol 2018;12(1):23-40. [doi: 10.1111/jnp.12105] [Medline: 27147117]

XSL•FO
RenderX

76. Goodwin RD, Dierker LC, Wu M, Galea S, Hoven CW, Weinberger AH. Trends in U.S. depression prevalence from 2015 to 2020: the widening treatment gap. Am J Prev Med 2022;63(5):726-733 [FREE Full text] [doi: 10.1016/j.amepre.2022.05.014] [Medline: 36272761]
77. Danielson ML, Holbrook JR, Bitsko RH, Newsome K, Charania SN, McCord RF, et al. State-level estimates of the prevalence of parent-reported adhd diagnosis and treatment among U.S. children and adolescents, 2016 to 2019. J Atten Disord 2022;26(13):1685-1697. [doi: 10.1177/10870547221099961] [Medline: 35603751]
78. Lindqvist S, Thorell LB. Brief report: manipulation of task difficulty in inhibitory control tasks. Child Neuropsychol 2009;15(1):1-7. [doi: 10.1080/09297040701793647] [Medline: 18608218]
79. Simmering VR, Andrews CM, Leuenberger R, Kovack-Lesh KA. Inconsistent flanker congruency effects across stimulus types and age groups: a cautionary tale. Behav Res Methods 2023;55(4):1778-1817. [doi: 10.3758/s13428-022-01889-2] [Medline: 35768744]
80. Digital_Therapy_Baby. GitHub. URL: https://github.com/LeelabSNU/Digital_Therapy_Baby [accessed 2026-02-11]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
ANCOVA: analysis of covariance
CES-DC: Center for Epidemiological Studies Depression Scale for Children
DSM-5: Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition)
eFlanker: emotional Flanker task
eGoNoGo: emotional Go/No-Go task
eStroop: emotional Stroop task
K-ARS: Korean ADHD Rating Scale
PC: principal component
PCA: principal component analysis
PFC: prefrontal cortex
RT: reaction time
STAI-CH: State-Trait Anxiety Inventory for Children

Edited by T Leung, S Brini; submitted 22.Oct.2025; peer-reviewed by A Barutcu, S Ajayi, S Ye; comments to author 21.Nov.2025; revised version received 23.Jan.2026; accepted 31.Jan.2026; published 25.Feb.2026.

Please cite as:
Park SE, Chung J, Lee SA
A Digital Tool for Assessing the Distinct Effects of Depression, Anxiety, and Attention-Deficit/Hyperactivity Disorder (ADHD) on Children's Emotional Cognitive Bias: Cross-Sectional Study
J Med Internet Res 2026;28:e86286
URL: https://www.jmir.org/2026/1/e86286
doi:10.2196/86286
PMID:

©Sang-Eon Park, Jisu Chung, Sang Ah Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Authoritative Textbook-Augmented Large Language Models for High-Altitude Public Health Medical Education in the Xizang Autonomous Region: Cross-Sectional Comparative Evaluation Study

Kun He[1,2], MD; Qiming Xiao[1,3], MD; Wangyang Chen[1], MD; Lisha Jing[3], MD; Yabing Wang[4], MD; Shuai Li[1,3], MD; Daiyu Yang[1,3], MD; Hemiao Xu[1,3], MD; Ke Pang[1,3], MD; Ruoyu Xiao[5], MD; Zhashilamu Liu[5], MD; Deji Zhuoga[5], MD; Ruxuan Chen[6], MD; Jingyi Li[7], MD; Long Chang[8], MD; Yangzhong Zhou[9], MD; Zhe Zhang[10], MD; Ran Li[11], MD; Lujing Ying[1,3], MD; Rutong Li[1], MD; Hongwei Wang[1], MD; Xin Yin[12], MD; Ge Zhen[2], MD; Siyi Cai[13], MD; Qijun Shan[14], MD; Qiang Wang[1], MD; Danzeng Zhuoga[15], MD; Ciren Yangjin[2], MD; Gesang Luobu[15], MD; Tu Ji[16], MD; Dong Wu[1,2], Prof Dr

[1]Department of Gastroenterology, State Key Laboratory of Complex Severe and Rare Diseases, Peking Union Medical College Hospital, Shuaifuyuan, No.1, Dongcheng District, Beijing, Beijing, China

[10]Department of Neurology, Peking Union Medical College Hospital, Beijing, Beijing, China

[11]Department of Endocrinology, Key laboratory of Endocrinology of National Health Commission, Peking Union Medical College Hospital, Beijing, Beijing, China

[12]Department of Gastroenterology, Air Force Medical Center, Chinese People's Liberation Army, Beijing, Beijing, China

[13]Department of Orthopedics, Peking Union Medical College Hospital, Beijing, Beijing, China

[14]Information Center, Peking Union Medical College Hospital, Beijing, Beijing, China

[15]High Altitude Medical Research Institute, Tibet Autonomous Region People's Hospital, Lhasa, Tibet, China

[16]Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, Beijing, China

[2]Department of Gastroenterology, Tibet Autonomous Region People's Hospital, Lhasa, Tibet, China

[3]Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, Beijing, China

[4]Department of Endocrinology, Beijing Friendship Hospital, Beijing, Beijing, China

[5]Medical College, Xizang University, Lhasa, Tibet, China

[6]Department of Pulmonary and Critical Care Medicine, Peking Union Medical College Hospital, Beijing, Beijing, China

[7]Department of Cardiology, State Key Laboratory of Complex Severe and Rare Diseases, Peking Union Medical College Hospital, Beijing, Beijing, China

[8]Department of Hematology, Peking Union Medical College Hospital, Beijing, Beijing, China

[9]Department of Rheumatology and Clinical Immunology, National Clinical Research Center for Dermatologic and Immunologic Diseases, Peking Union Medical College Hospital, Beijing, Beijing, China

Corresponding Author:
Dong Wu, Prof Dr
Department of Gastroenterology, State Key Laboratory of Complex Severe and Rare Diseases, Peking Union Medical College Hospital, Shuaifuyuan, No.1, Dongcheng District, Beijing, Beijing, China

Abstract

Background: Public health medical education is increasingly important in the low-resource, high-altitude Xizang Autonomous Region (Tibet). Traditional authoritative textbooks do not meet modern needs for accessibility and interactivity, whereas general large language models (LLMs) may hallucinate in specialized medical domains. Developing specialized LLMs for low-resource regions is also expensive and difficult.

Objective: This study aimed to explore a novel approach to high-altitude public health medical education in the low-resource Xizang Autonomous Region that integrates modern LLMs and authoritative textbooks, using a comprehensive benchmark evaluation across multiple dimensions and retrieval-augmented generation (RAG) technology.

Methods: We conducted a 2-stage cross-sectional comparative evaluation study to benchmark publicly available LLMs and evaluate the added value of textbook-augmented retrieval under standardized generation settings and blinded expert assessment. First, 4 publicly available LLMs (GPT-5.2 [OpenAI], Gemini 3.0 Pro [Google], DeepSeek R1 [DeepSeek], and Tencent HY 2.0 [Tencent]) were benchmarked using an 80-question benchmark on high-altitude public health medicine developed by authoritative

XSL•FO
RenderX

medical specialists. Each question was asked 3 times, yielding 960 outputs; first responses (n=320) were scored under blinded conditions by 2 independent 8-member physician panels. A clinically weighted evaluation of multidimensional first-response scores (including comprehensiveness, accuracy, clarity, and relevance) and a composite consistency metric (including semantic similarity and algorithmic similarity) was administered. Second, 4 specific and prevalent authoritative textbooks on high-altitude public health medicine—*Ward, Milledge and West's High Altitude Medicine and Physiology, High Altitude Medicine: A Case-Based Approach, High Altitude Medicine,* and *High Altitude Medical Protection*—were deployed as the external knowledge base for the evaluation-optimized model. Statistical analyses included Spearman ρ, Cronbach α, intraclass correlation coefficients, Friedman tests with Dunn multiple comparisons, and paired Wilcoxon signed-rank tests. The significance threshold was set at α=.05.

Results: DeepSeek R1 was selected as the optimal base model for achieving the highest weighted score (5.61/10.00), followed by GPT-5.2 (5.51/10.00), Gemini 3.0 Pro (5.39/10.00), and Tencent HY 2.0 (4.71/10.00). The deployed retrieval-augmented model integrating the authoritative textbooks and the optimal LLM DeepSeek R1, HPHME-Xplus-RAG, achieved remarkable improvement in multidimensional scores compared to baseline DeepSeek R1 (median 8.00 [IQR 7.88 - 8.00] vs median 7.63 [IQR 7.38 - 7.88]; *P*<.001, r_rb=0.68, indicating a large effect).

Conclusions: Integrating authoritative textbooks with an evaluation-optimized general LLM through an RAG framework showed strong performance for medical education in the low-resource Xizang Autonomous Region. Unlike prior studies that mainly evaluated general LLMs or used clinical guidelines to build RAG systems for diagnosis and treatment, this study used authoritative textbooks for the broader, guideline-scarce field of public health medical education. This work provides a replicable workflow—domain-authoritative knowledge+RAG+model optimization and evaluation—for low-resource settings, with practical implications for medical instructors and students, hospitals, and public health services seeking cost-effective, convenient, and trustworthy educational support.

(J Med Internet Res 2026;28:e92852) doi:10.2196/92852

KEYWORDS

public health medical education; large language model; authoritative textbooks; retrieval-augmented generation; high altitude

Introduction

High-altitude public health medical education is essential for protecting individual well-being and regional public health in plateau settings. High-altitude regions, including the Xizang Autonomous Region (Tibet), are culturally and ecologically significant and have attracted increasing attention because plateau exposure poses substantial physical and psychological challenges to both residents and visitors [1-3]. Existing evidence shows that high-altitude environments may affect multiple physiological and behavioral domains, including cardiovascular function, cognitive performance, sleep quality, blood pressure regulation, neurological symptoms, musculoskeletal adaptation, inflammatory and immune responses, and mental health [4-10]. These multisystem effects can impair daily functioning, increase health risks during travel, work, and long-term residence, and create practical demands for early recognition, prevention, and self-management. Therefore, timely public health medical education is needed to raise awareness and promote preventive behaviors among medical instructors and students, frontline health care workers, residents, and travelers.

However, public medical education in this setting faces substantial structural constraints, particularly in low-resource regions such as the Xizang Autonomous Region. Studies have shown insufficient general practitioner training and limited continuing medical education opportunities [11-13]. Local reports have also described difficulty accessing standardized lectures and textbooks because of limited educational resources, as well as shortages of qualified teachers and the limited relevance of standardized teaching materials to plateau-specific conditions [14,15]. In addition, existing formal guidelines and consensus statements in high-altitude medicine are fragmented and largely focused on acute altitude illness, travel medicine, or selected subgroups, while comprehensive guidance specifically addressing high-altitude public health medicine remains scarce [16-20].

Although authoritative textbooks have long served as the primary medium for medical education in this field and encapsulate years of accumulated expertise, conventional materials alone increasingly struggle to meet modern expectations for accessibility, efficiency, and interactivity. Large language models (LLMs) have rapidly advanced driven by advances in deep learning and the Transformer architecture, supporting a broad range of medical tasks, including medical education, clinical decision-making, specialty-specific question answering, and diagnostic reasoning, although their performance varies substantially across specialties, tasks, and evaluation settings [21-39]. Despite this growing body of literature, evidence for LLM use in high-altitude medical education remains scarce. Therefore, although LLMs show considerable promise in spreading public health information, especially in settings where traditional outreach is limited, their role in high-altitude public health medical education remains insufficiently studied, with domain-specific accuracy and hallucinations probably being a key limitation [40,41]. Also, developing specialized large models for low-resource domains and regions is prohibitively expensive and difficult. This contrast underscores the need for hybrid approaches that combine the strengths of both paradigms.

Retrieval-augmented generation (RAG) offers such a bridge. By integrating external, authoritative sources into the generative process, RAG allows an LLM to retrieve verified information before producing responses, providing a cost-effective approach compared to developing a new proprietary LLM for

XSL•FO
RenderX

low-resource regions [42-47]. In high-altitude public health medical education, RAG has the potential to merge authoritative content with the flexibility of LLMs to deliver reliable and personalized information. This study aims to evaluate a novel approach that integrates the accuracy of authoritative textbooks with the efficiency and reasoning capacity of an evaluation-optimized LLM using the RAG mechanism for high-altitude public health medical education in this low-resource setting.

Methods

Study Design

This was a 2-stage cross-sectional comparative evaluation study conducted to select an optimal LLM for the RAG architecture and to evaluate its performance in high-altitude public health medical education. First, the performance of 4 extensively adopted publicly available LLMs in answering questions on high-altitude health medical education was evaluated. Based on the weighted results of the multidimensional first-response scores and composite consistency metrics with the weighting ratio determined by the clinical needs assessment, the LLM with the best performance was selected. Second, a retrieval corpus constructed from the specific and prevalent authoritative textbooks was deployed as an external data source for the selected LLM to develop a RAG architecture for medical education on high-altitude health, and the performance improvement of the RAG architecture compared to the LLMs was verified (Figure 1).

Figure 1. Study design of a 2-stage cross-sectional comparative evaluation study on large language models for high-altitude public health medical education in the Xizang Autonomous Region (Tibet), China. LLM: large language model; RAG: retrieval-augmented generation; HPHME-Xplus-RAG: High-Altitude Public Health Medical Education–Xplus (Xiehe in Chinese for Peking Union Medical College Hospital and Xizang for People's Hospital of Xizang Autonomous Region)–Retrieval-Augmented Generation.

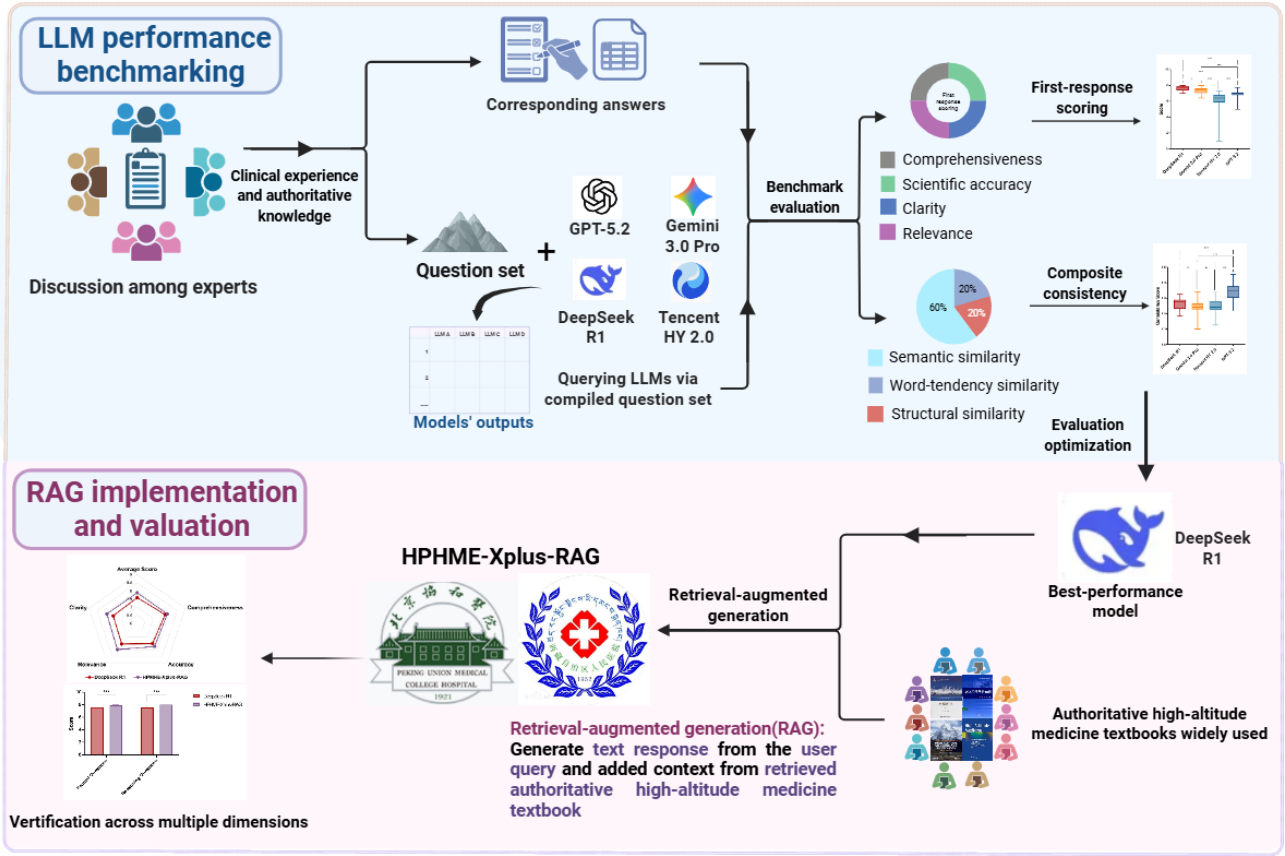

Data Collection

A total of 80 questions and corresponding reference answers were developed through an expert-driven, iterative process. Multidisciplinary specialists from Peking Union Medical College Hospital (PUMCH) and People's Hospital of Xizang Autonomous Region collaboratively proposed and refined candidate questions on the basis of real-world diagnostic, therapeutic, preventive, and health education scenarios encountered in high-altitude practice. Rather than using random question sampling, we adopted a purposive selection strategy to ensure that the benchmark represented the major and practice-relevant domains of high-altitude public health medicine.

During question set construction, particular attention was paid to broad domain coverage across the multisystem health effects of hypobaric hypoxia and plateau exposure. The final benchmark therefore included respiratory, cardiovascular, gastrointestinal, neurological, hematological, musculoskeletal, and infectious disease topics, as well as lifestyle, nutrition, and psychological acclimatization issues. In this way, the benchmark was designed to reflect the interdisciplinary nature of high-altitude public health medicine, spanning clinical medicine, physiology, prevention, and behavioral health (Figure 2). The preliminary

XSL•FO
RenderX

question set was subsequently reviewed and validated by Xizang-aid physicians from PUMCH with direct experience in high-altitude health adaptation in order to confirm its contextual relevance, content breadth, and practical representativeness for the target setting. Each question was accompanied by an expert-defined reference answer, which served as the "gold standard" with a full score of 10 out of 10. These standard answers were based on authoritative sources, including clinical guidelines, textbooks, and medical specialists' consensus.

Figure 2. Composition of the 80-question benchmark used in this cross-sectional comparative evaluation study of large language models for high-altitude public health medical education.

Four LLMs—GPT-5.2 (OpenAI), Gemini 3.0 Pro (Google), DeepSeek R1 (DeepSeek), and Tencent HY 2.0 (Tencent)—were selected a priori because they were publicly available, stably accessible during the study period, and representative of major contemporary LLM families across both international and Chinese developer ecosystems [48-51]. Restricting stage 1 to standalone general-purpose models allowed a controlled comparison of intrinsic model performance before subsequent RAG augmentation. All models were accessed through their respective application programming interfaces under standardized conditions, with temperature set to 0.1 and top_p to 0.9 during generation to ensure result comparability, leaving all other hyperparameters at their developer defaults.

Each question was posed 3 times to each LLM, generating 3 independent responses per model. Only the first response was scored to simulate real-world, on-demand information retrieval without iterative prompting, while all 3 were used for consistency analysis.

Quantitative Evaluation of First Responses

Four LLMs were anonymized and labeled with a code letter (A-D) to ensure blinding during assessment. Sixteen experts

XSL•FO
RenderX

participated in the evaluation as 2 independent 8-member panels, one from PUMCH and one from People's Hospital of Xizang Autonomous Region. All experts were trained and calibrated in advance using the scoring rubric. For each response, the 8 experts within each panel discussed the output and generated one panel-level consensus score using integer ratings from 0 to 10 across 4 predefined dimensions: comprehensiveness, scientific accuracy, relevance, and clarity. Thus, each response yielded 2 consensus scores, one from each panel. The final score for each LLM's response to a given question was calculated as the average of the 4 dimension scores within each panel; the 2 panel-level consensus scores were then used for interpanel reliability analyses, and their mean was used for subsequent comparative analyses (Figure 3).

Figure 3. Physician scoring rubric used in the comparative evaluation of model-generated answers for high-altitude public health medical education.

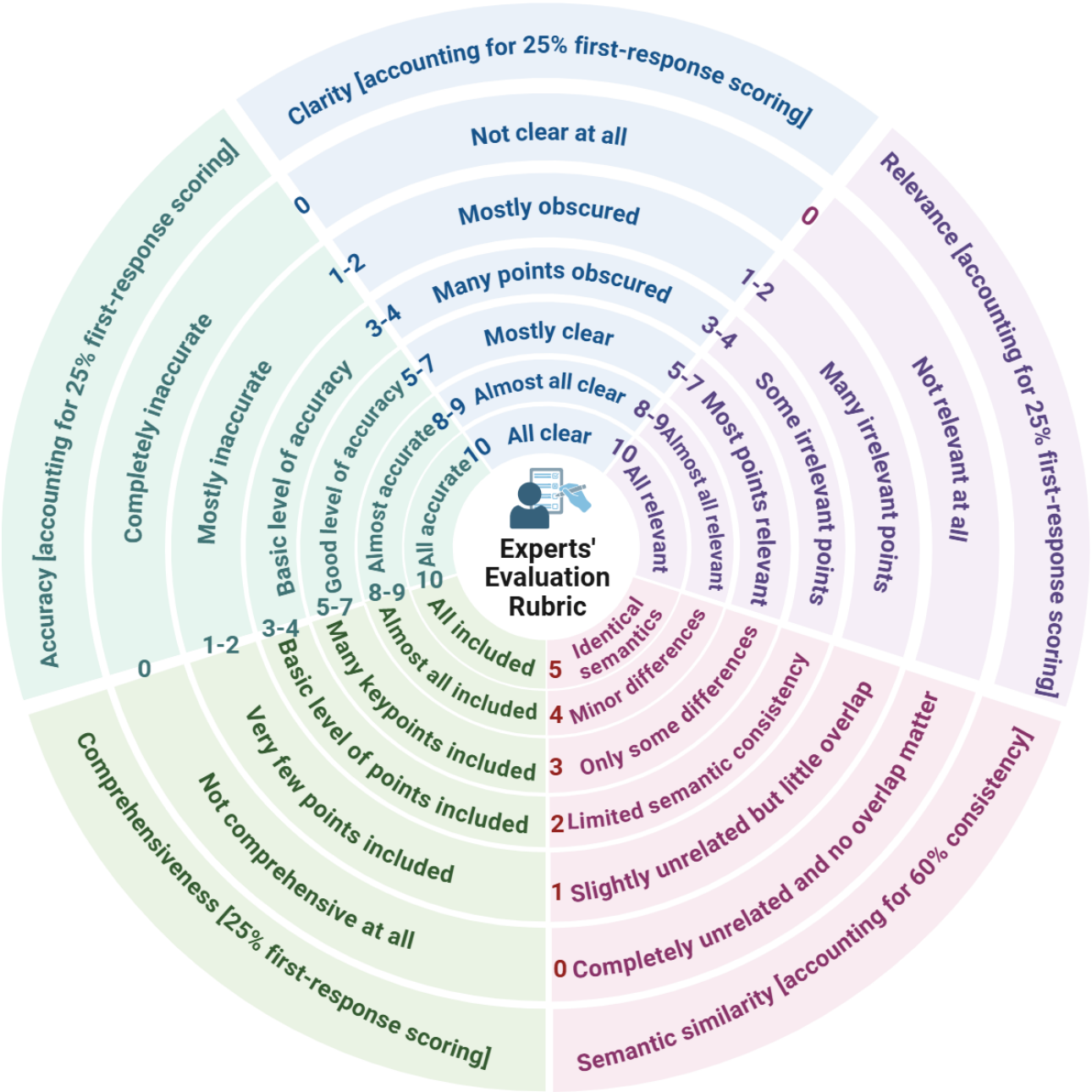

Consistency Analysis

To assess the intramodel stability of the LLMs' outputs, 3 responses generated for each question were compared using both lexical (word frequency–based) and structural similarity measures. Specifically, for each question, the 3 independently generated responses from the same model were compared in a pairwise manner (response 1 vs response 2, response 1 vs response 3, and response 2 vs response 3), and the resulting similarity scores were averaged across the 3 pairwise comparisons.

For lexical similarity, text content similarity was calculated using term frequency–inverse document frequency vectorization followed by cosine similarity, implemented in Python using third-party libraries including sklearn and numpy. For structural similarity, Python's built-in difflib.SequenceMatcher was used to quantify structural overlap between texts. In this structural similarity analysis, punctuation marks, spaces, and special

XSL•FO
RenderX

characters were ignored so that superficial formatting differences would not disproportionately affect the similarity results. Both lexical similarity and structural similarity ranged from 0 to 1.

However, purely algorithmic similarity metrics have inherent limitations when applied to complex, domain-specific texts such as medical and public health communication. Lexical similarity measures primarily focus on word overlap and frequency, which may overlook paraphrasing, synonym usage, and deeper contextual meaning. Two semantically identical sentences with different wording can thus receive low lexical similarity scores. Similarly, structural similarity metrics, which assess syntactic patterns or sentence structures, may fail to capture differences in the underlying semantic content or clinical implications of responses. These methods cannot reliably detect whether subtle but critical nuances—such as risk factors or treatment recommendations—are consistently conveyed.

Due to these shortcomings, relying solely on algorithmic similarity can misrepresent true consistency and overlook clinically relevant differences. To overcome this, the final consistency evaluation incorporated both quantitative and physician-assessed semantic similarity, which involves expert reviewers evaluating how closely the meaning and clinical relevance of different model-generated answers align, which is fine-tuned with reference to the rubric in SemEval-2017 Task 1 [52]. This expert judgment better captures the practical, real-world equivalence of responses, reflecting nuances that automated metrics miss. Semantic similarity was scored on a 0 - 5 scale and subsequently normalized to a 0 - 1 scale for integration with the algorithmic metrics. The final composite score was calculated with the following weighting: physician-assessed semantic similarity (60% weight, detailed in Figure 3), structural similarity (20% weight), and lexical similarity (20% weight). This composite score was used to better reflect the meaningful consistency of model responses in clinical and public health communication contexts.

Development and Evaluation of the New RAG Architecture

Before model output generation, physician scoring, and comparative analysis, the expert panels prospectively agreed to use a clinically weighted composite approach to determine the optimal LLM. Because the first-response score was reported on a 0 - 10 scale, whereas the composite consistency metric was presented on a normalized 0 - 1 scale, the composite consistency metric was linearly rescaled to a 0 - 10 scale before model selection. The final model-selection score was then derived by weighting the first-response score at 60% and the rescaled composite consistency score at 40%. This weighting was selected during study design to reflect the intended real-world use case, in which first-response quality was considered slightly more important than response consistency.

Four widely used authoritative books on high-altitude public health medicine, *Ward, Milledge and West's High Altitude Medicine and Physiology*, *High Altitude Medicine: A Case-Based Approach*, *High Altitude Medicine*, and *High Altitude Medical Protection*, were selected as the external knowledge source for the LLM [53-56]. These authoritative books are essential reading material for individuals entering high-altitude regions and serve as a fundamental reference for medical professionals working in high-altitude medicine in China and around the world. The textbooks were obtained in PDF format from official online sources accessible to the research team. The files were used only in a closed internal research environment to construct and test the retrieval corpus, and no copyrighted content was redistributed or released publicly.

The RAG pipeline was constructed using the LangChain framework, with the text-embedding-ada-002 model used to generate text embeddings. Outputting 1536-dimensional vectors with an 8192-token context window, this model effectively captures the semantic information of medical texts. ChromaDB was used as the vector storage backend. Following chunking, all document segments were transformed into vectors via the embedding model and stored in a local ChromaDB instance to facilitate efficient similarity retrieval. Authoritative textbooks were segmented by chapters and paragraphs using a Recursive Character Text Splitter, with a chunk size of 512 tokens and an overlap of 128 tokens. An initial retrieval of 5 fragments was subsequently reranked, from which the top 3 most relevant segments were selected and fed into the LLM to further mitigate noise. To ensure focused and comparable outputs, we used the identical question set, prompt templates, and hyperparameter settings from the first evaluation phase, configuring only a temperature of 0.1 and a top_p of 0.9, with all other hyperparameters kept at developer defaults.

To reduce expectancy bias in the second-stage evaluation, the assessment of HPHME-Xplus-RAG, which stands for High-Altitude Public Health Medical Education–Xplus (Xiehe in Chinese for Peking Union Medical College Hospital and Xizang for People's Hospital of Xizang Autonomous Region)–Retrieval-Augmented Generation, was conducted after a washout period of 2 weeks following the initial 4-model evaluation and was coordinated by a study member who was not involved in model development. During this reassessment, the physician panels were informed only that an additional model output set would be evaluated using the same prespecified scoring rubric and consensus procedure as in stage 1. The identity of the system was not disclosed at the time of scoring. Explicit model identifiers and other source-revealing labels were removed from the evaluation materials before assessment. This procedure provided temporal separation and source masking, although it did not constitute an identical fully blinded parallel comparison to the A-D anonymized first-stage assessment.

Following a similar comparative approach, average scores and scores in 4 dimensions of the RAG architecture and selected LLM were compared. To further validate the performance improvement of the RAG architecture over the selected LLM, an additional analysis was conducted to investigate the performance disparity between them across different text-generation scenarios. The 80 questions were partitioned into 2 cohorts based on text generation task types: factual and reasoning questions. Factual questions were defined as those primarily requiring recall or retrieval of explicit information from authoritative sources, whereas reasoning questions were defined as those requiring integration, inference, or application

XSL•FO
RenderX

of knowledge to generate the answer (the group label for each question is provided in Multimedia Appendix 1). A subsequent comparative evaluation was performed within each cohort.

Statistical Analysis

Descriptive statistics, including measures of central tendency (mean and median) and variability (minimum, maximum, SD, SEM, and coefficient of variation), were calculated for each model's performance. Interpanel reliability was evaluated on the 2 panel-level consensus scores using Spearman ρ, Cronbach α, and the intraclass correlation coefficient (ICC). ICC single and ICC average were calculated using a 2-way mixed-effects model to assess the agreement of a single panel-level consensus score and the reliability of the mean of the 2 panel-level consensus scores, respectively. To test for statistically significant differences between models, the Friedman test followed by the Dunn multiple comparisons test and the paired Wilcoxon signed-rank test were used. In addition to *P* values, we reported effect sizes to quantify the magnitude of differences. For Friedman tests, Kendall W was calculated as a measure of concordance and effect size (Friedman statistic/n(k–1)). For paired Wilcoxon signed-rank tests between HPHME-Xplus-RAG and DeepSeek R1, matched-pairs rank-biserial correlation (r_rb, $(R_+ - R_-) / (R_+ + R_-)$) was reported. Unless otherwise indicated, nonnormally distributed continuous variables are presented as median (IQR). All analyses were conducted using IBM SPSS (version 29.0), with exact significance levels determined via Monte Carlo simulation. A 2-sided α level of .05 was used throughout.

Ethical Considerations

This study was reviewed and approved by the Ethics Committee of PUMCH (approval number I-25PJ1246) and the Ethics Committee of the People's Hospital of the Xizang Autonomous Region (approval number ME-TBHP-25-KJ-124). Although the primary study materials were LLM-generated responses rather than patient data, the study involved 16 medical specialists as structured evaluators, and therefore formal human subjects research ethics review and approval were obtained. No patients, patient records, protected health information, or identifiable public user data were involved. Before participation, all medical specialists were informed of the study purpose, evaluation procedures, voluntary nature of participation, and intended use of the evaluation results, and all agreed to participate. The approved ethics protocol permitted the use of their evaluation scores for this study without additional consent. To protect privacy and confidentiality, individual evaluator identities were not linked to scoring results, and only aggregated panel-level scores and summary statistics are reported. No monetary or nonmonetary compensation was provided. The paper and supplementary materials contain no identifiable images, screenshots, names, account identifiers, or other information that could identify individual participants or users.

Results

Overview and Reliability Verification of the Evaluation Between the Two Evaluation Panels

We queried each of the 4 models independently 3 times with the set of 80 questions on medical education on high-altitude health, and their outputs were collated in Multimedia Appendix 2.

In a simulation of real-world conditions, the models' first responses underwent a blind assessment by 2 independent evaluation panels. Both panels scored the answers of DeepSeek R1 as the best, followed by the answers of Gemini 3.0 Pro, GPT-5.2, and Tencent HY 2.0 (Table 1). Notably, the minimum score of 1.00 assigned to Tencent HY 2.0 by both evaluation panels corresponded to Question 34. On manual review of the raw output, this response was interpreted as a genuine model failure due to informational and clinical insignificance.

Table . Descriptive statistics for the panel-level consensus scores assigned to the first responses generated by 4 publicly available large language models in a 2-stage cross-sectional comparative evaluation study on high-altitude public health medical education.

First responses scoring	GPT-5.2		Gemini 3.0 Pro		DeepSeek R1		Tencent HY 2.0	
	Evaluation panel 1	Evaluation panel 2	Evaluation panel 1	Evaluation panel 2	Evaluation panel 1	Evaluation panel 2	Evaluation panel 1	Evaluation panel 2
Minimum	4.75	5.25	6.00	6.50	6.75	6.75	1.00	1.00
Median (IQR)	7.00 (0.3125)	7.00 (0.5)	7.25 (0.75)	7.50 (0.75)	7.50 (0.5)	7.75 (0.5)	6.25 (1.0625)	6.50 (0.5)
Maximum	7.75	7.75	8.00	8.00	8.00	8.00	7.25	7.25
Mean (SD)	6.88 (0.42)	6.94 (0.38)	7.33 (0.43)	7.37 (0.41)	7.53 (0.36)	7.63 (0.35)	6.06 (0.90)	6.36 (0.82)
Coefficient of variance (%)	6.14	5.44	5.81	5.53	4.80	4.54	14.83	12.92

To ensure that the 2 evaluation panels adhered to the same criteria, we assessed interpanel reliability based on the 2 panel-level consensus scores for the answers provided by the 4 LLMs (Table 2). Spearman ρ and corresponding *P* values revealed strong and statistically significant correlations between their scores, suggesting that the first responses of the 4 LLMs were evaluated in the same way. Similarly, Cronbach α, ICC single, and ICC average suggested high reliability. All Cronbach α values were >0.7, suggesting strong internal consistency reliability of the 80 questions on medical education on high-altitude health. ICC single and ICC average values were also calculated using a 2-way mixed-effects model to assess the agreement between the two evaluation panels and the reliability of using the average scores provided by the 2 evaluation panels. All ICCs were statistically significant.

XSL•FO
RenderX

Table . Interpanel reliability for the panel-level consensus scores assigned to the first responses generated by 4 publicly available large language models in a cross-sectional comparative evaluation study on high-altitude public health medical education.

First responses scoring	Spearman ρ	*P* value	Cronbach α	ICC[a] single	*P* value	ICC average	*P* value
GPT-5.2	0.789	<.001	0.906	0.822	<.001	0.902	<.001
Gemini 3.0 Pro	0.678	<.001	0.794	0.658	<.001	0.794	<.001
DeepSeek R1	0.764	<.001	0.849	0.712	<.001	0.832	<.001
Tencent HY 2.0	0.778	<.001	0.921	0.805	<.001	0.892	<.001

[a]ICC: intraclass correlation coefficient.

Furthermore, we computed similar descriptive statistics and rigorously assessed interrater reliability for each of the 4 scoring dimensions, including comprehensiveness, scientific accuracy, clarity, and relevance (Tables S3-S10 in Multimedia Appendix 3). Furthermore, the reliability between the two evaluation panels in each dimension was also sufficient.

Comparison Among the First Responses of 4 LLMs

An average score was computed across the 2 evaluation panels for each question and LLM. The answers of DeepSeek R1 were scored as the best, followed by the answers of Gemini 3.0 Pro, GPT-5.2, and Tencent HY 2.0 (Table 3).

Table . Descriptive statistics for the mean first-response scores averaged across 2 independent physician-panel consensus ratings for 4 publicly available large language models evaluated on an 80-question benchmark in high-altitude public health medical education.

First responses scoring	GPT-5.2	Gemini 3.0 Pro	DeepSeek R1	Tencent HY 2.0
Minimum	5.00	6.38	7.00	1.00
Median (IQR)	7.00 (0.375)	7.38 (0.53125)	7.63 (0.5)	6.34 (0.875)
Maximum	7.75	8.00	8.00	7.25
Mean (SD)	6.91 (0.38)	7.35 (0.38)	7.58 (0.33)	6.21 (0.83)
Coefficient of variance (%)	5.54	5.17	4.35	13.34

Owing to the nonnormal distribution of the rating data, we conducted comparison among the average scores and individual dimension scores of first responses of 4 LLMs by using the Friedman test followed by the Dunn multiple comparisons test (Figure 4). The average scores of the 4 LLMs exhibited strong variation (*P*<.001), with a Kendall W of 0.75, indicating a large overall effect. DeepSeek R1 outperformed the other 3 LLMs (*P*=.01 vs Gemini 3.0 Pro, *P*<.001 vs GPT-5.2 and Tencent HY 2.0), Gemini 3.0 Pro outperformed GPT-5.2st and Tencent HY 2.0 (*P*<.001 vs both), and GPT-5.2 outperformed Tencent HY 2.0 (*P*<.001; Figure 4).

Similarly, the Friedman test revealed significant differences among the average scores of the 4 LLMs in the 4 dimensions (*P*<.001 for all), with Kendall W values of 0.66 for comprehensiveness, 0.99 for scientific accuracy, 0.54 for relevance, and 0.62 for clarity; furthermore, DeepSeek R1 outperformed GPT-5.2 and Tencent HY 2.0 in all the 4 dimensions (*P*<.001 for all), but only outperformed Gemini 3.0 Pro in relevance (*P*=.02) and no significant difference versus Gemini 3.0 Pro in the other 3 dimensions. Gemini 3.0 Pro outperformed GPT-5.2 and Tencent HY 2.0 in all 4 dimensions (*P*<.001 for all except *P*=.006 vs GPT-5.2 in relevance and *P*=.004 vs GPT-5.2 in clarity). GPT-5.2 outperformed Tencent HY 2.0 in all 4 dimensions (*P*<.001 for all except *P*=.002 in relevance; Figure 4). Performance rankings of the average scores held consistent for both the overall score and each of the 4 individual dimensions, with DeepSeek R1 being the best, followed by Gemini 3.0 Pro, GPT-5.2, and Tencent HY 2.0 (Figure 4).

XSL•FO
RenderX

Figure 4. Comparison of first-response performance across 4 publicly available large language models. (A) Overall average scores across the 4 evaluation dimensions. Box plots show the median (center line), IQR (box), and range (whiskers); n=80 questions per model. (B) Mean scores for comprehensiveness, scientific accuracy, relevance, and clarity (bars indicate mean, SD; n=80 per model). (C) Radar plot summarizing overall and dimension-specific performance. *P<.05; **P<.01; ***P<.001; ns, not significant.

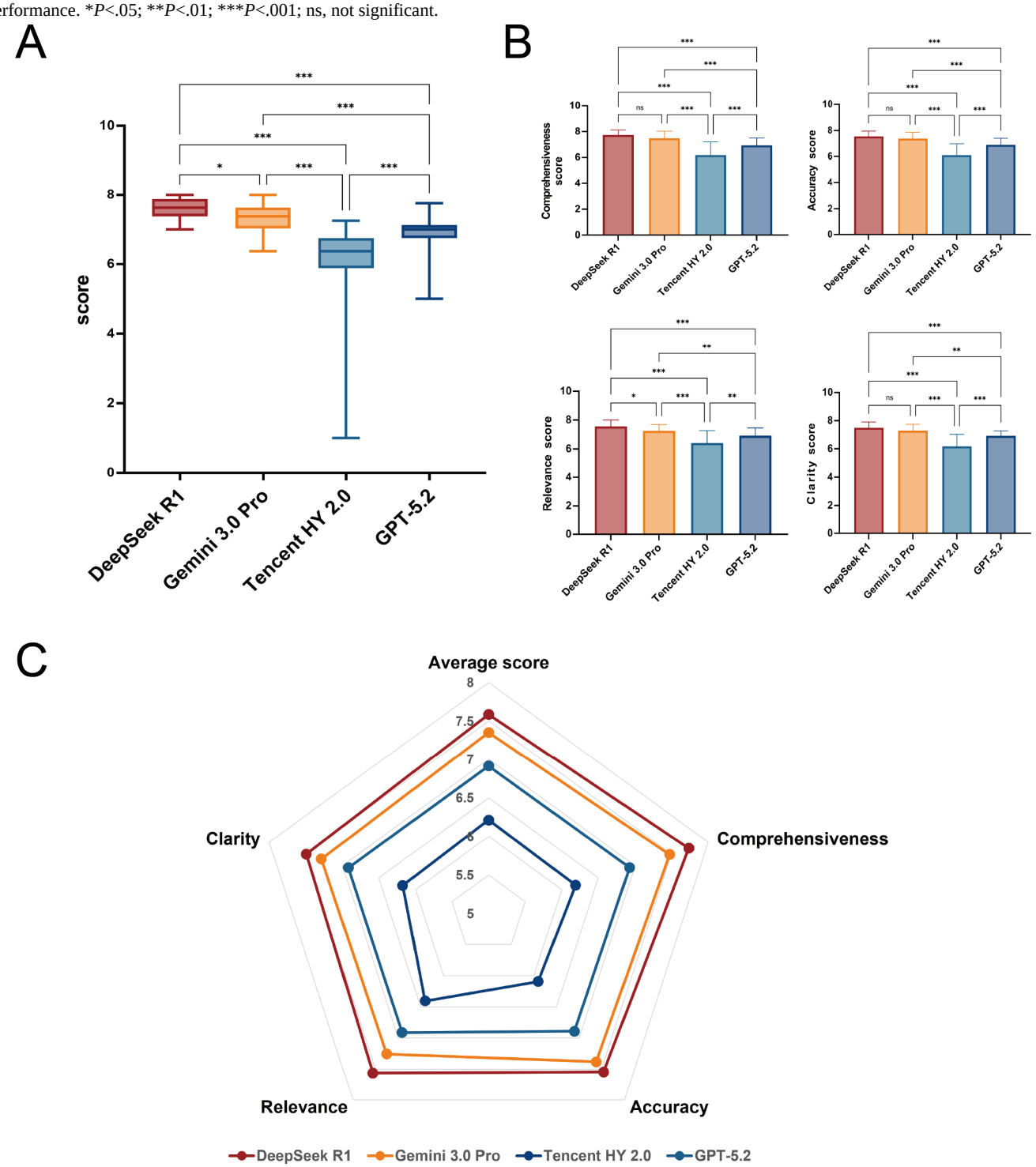

Comparison of the Consistency of 3 Responses of 4 LLMs

To assess output stability, we measured textual consistency among 3 responses per LLM using a composite consistency index incorporating both subjective semantic and objective algorithmic consistency metrics. Similarly, we adopted the Friedman test followed by the Dunn multiple comparisons test to evaluate disparities in the composite consistency scores. In order to dissect its contributions, the test was also conducted on its 2 subcomponents: the subjective and the objective consistency scores (Figure 5). For the composite consistency, the Friedman test was significant (P<.001; Kendall W=0.43, indicating a moderate overall effect). GPT-5.2 achieved the highest consistency, with 0.350 (0.303 - 0.376) versus 0.242 (0.227 - 0.264) for Gemini 3.0 Pro, 0.275 (0.238 - 0.290) for DeepSeek R1, and 0.243 (0.224 - 0.281) for Tencent HY 2.0. GPT-5.2 outperformed the other 3 LLMs (P<.001 for all). DeepSeek R1 outperformed Tencent HY 2.0 and Gemini 3.0 Pro (P=.03 vs Tencent HY 2.0 and P=.004 vs Gemini 3.0 Pro).

XSL•FO
RenderX

No significant difference was detected between the consistency scores of Tencent HY 2.0 and Gemini 3.0 Pro (Figure 5).

Similarly, Friedman tests revealed strong differences among the subjective and objective consistency scores of the 4 LLMs (*P*<.001 for both, with Kendall W values of 0.24 for subjective consistency scores, 0.57 for objective consistency scores); in comparison of the subjective consistency scores, DeepSeek R1 outperformed the other 3 LLMs (*P*<.001 for all). Tencent HY 2.0 outperformed GPT-5.2 (*P*=.04) and no statistical differences were detected in Gemini 3.0 Pro versus Tencent HY 2.0 and Gemini 3.0 Pro versus GPT-5.2 (Figure 5). As for the comparison of the objective consistency scores, GPT-5.2 outperformed the other 3 LLMs (*P*<.001 for all); however, no statistical differences were detected between the objective consistency scores of the other LLMs (Figure 5).

Figure 5. Comparison of response consistency across 4 publicly available large language models. (A) Composite consistency scores across 4 LLMs. Box plots show the median (center line), IQR (box), and range (whiskers); n=80 questions per model. (B) Subjective semantic consistency scores (mean, SD). (C) Objective algorithmic consistency scores (mean, SD). **P*<.05; ***P*<.01; ****P*<.001; ns, not significant.

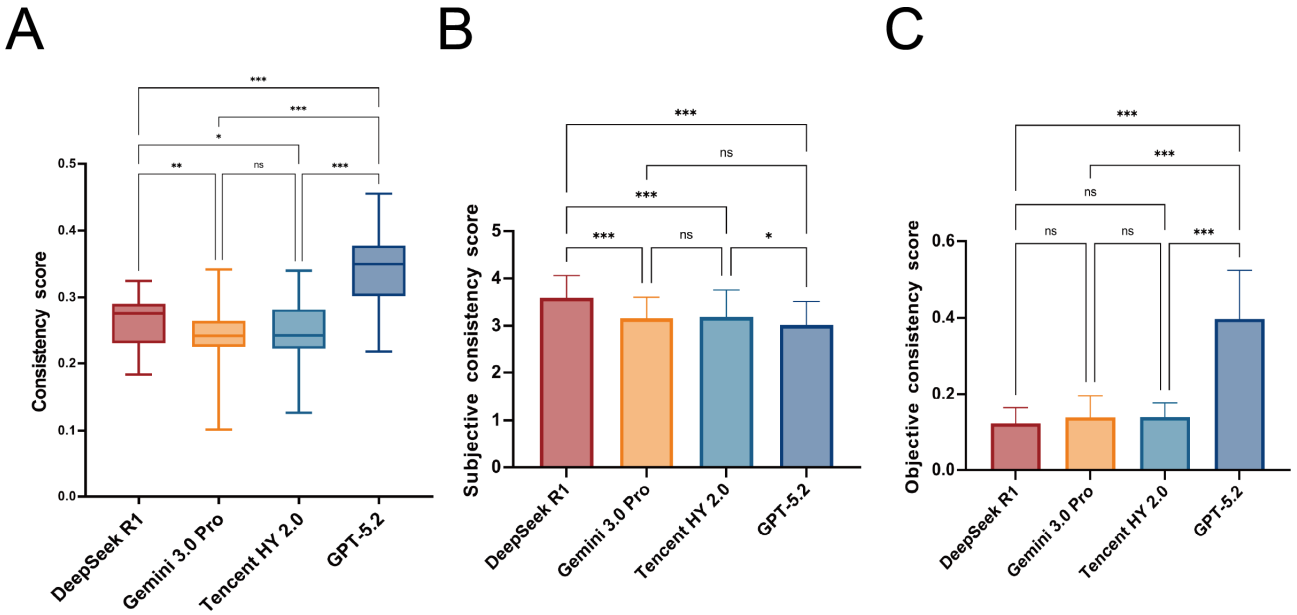

Comparison Between the Answers of DeepSeek R1 and HPHME-Xplus-RAG

The top-performing model was selected based on a composite model-selection metric, which combined the first-response score and the composite consistency metric in a 6:4 weighting ratio after the consistency metric had been rescaled from the normalized 0 - 1 scale to the 0 - 10 scale used for first-response scoring. DeepSeek R1 was finally selected as the best-performing LLM among the 4 LLMs (5.61 vs 5.51 for GPT-5.2, 5.39 for Gemini 3.0 Pro, and 4.71 for Tencent HY 2.0).

With the 4 authoritative books being ingested into a local retrieval corpus for DeepSeek R1, HPHME-Xplus-RAG, a RAG model to support high-altitude public health medical education, was jointly deployed by researchers from PUMCH and People's Hospital of Xizang Autonomous Region (hereafter referred to as Xplus) [57].

After a 2-week washout period, we additionally queried HPHME-Xplus-RAG with the same 80-question set, and these outputs underwent a source-masked reassessment coordinated by an independent study member; the answers were collated in Multimedia Appendix 2. We computed descriptive statistics and rigorously assessed interrater reliability of the average scores of the answers provided by HPHME-Xplus-RAG and the scores in the 4 dimensions, including comprehensiveness, scientific accuracy, clarity, and relevance (Tables S1-S10 in Multimedia Appendix 3). Also, the reliability between the two evaluation panels was sufficient.

We used the paired Wilcoxon signed-rank test to compare the overall ratings between HPHME-Xplus-RAG and DeepSeek R1, which were nonnormally distributed. In addition, the same test was applied to compare ratings within each of the 4 scoring dimensions and the 2 question categories (Figure 6). HPHME-Xplus-RAG achieved higher overall first-response scores (median 8.00 [IQR 7.88 - 8.00]) over DeepSeek R1 (median 7.63 [IQR 7.38 - 7.88]). The paired Wilcoxon signed-rank test showed that this difference was statistically significant (*P*<.001), with a matched-pairs rank-biserial correlation of r_rb=0.68, indicating a large effect (Figure 6). HPHME-Xplus-RAG outperformed DeepSeek R1 in comprehensiveness (*P*<.001; r_rb=0.60), scientific accuracy (*P*<.001; r_rb=0.62), relevance (*P*<.001; r_rb=0.94), and clarity (*P*<.001; r_rb=0.77), indicating a large effect across all 4 dimensions (Figure 6). The average scores and scores across the 4 dimensions of the answers of HPHME-Xplus-RAG were all higher than those of DeepSeek R1, and the whole improvement was mainly due to the enhancement in clarity and relevance (Figure 6).

To further assess the new model's performance enhancement in different task scenarios, we divided the question set into 2 groups, factual questions and reasoning questions, and administered paired Wilcoxon signed-rank tests within each group for comparison. Paired Wilcoxon signed-rank tests

XSL•FO
RenderX

revealed that HPHME-Xplus-RAG (median 7.88 [IQR 7.75 - 8.00] in factual questions and median 8.00 [IQR 7.88 - 8.31] in reasoning questions) outperformed DeepSeek R1 (median 7.63 [IQR 7.50 - 7.88] in factual questions and median 7.50 [IQR 7.25 - 8.00] in reasoning questions in both 2 kinds of questions; $P<.001$ for both question groups; r_rb=0.76 for factual questions and r_rb=0.82 for reasoning questions, both indicating a large effect). And HPHME-Xplus-RAG exhibited a larger performance gain over DeepSeek R1 in reasoning questions compared with factual questions (Figure 6).

Figure 6. Comparison between the baseline model DeepSeek R1 and HPHME-Xplus-RAG. (A) Overall average first-response scores (box plots show the median, IQR, and range; n=80 questions per model). (B) Mean scores in comprehensiveness, scientific accuracy, relevance, and clarity (bars indicate mean, SD; n=80 per model). (C) Radar plot summarizing overall and dimension-specific performance. (D) Subgroup comparison according to question type, including factual questions (n=51) and reasoning questions (n=29). *$P<.05$; **$P<.01$; ***$P<.001$; HPHME-Xplus-RAG: High-Altitude Public Health Medical Education–Xplus (Xiehe in Chinese for Peking Union Medical College Hospital and Xizang for People's Hospital of Xizang Autonomous Region)–Retrieval-Augmented Generation; ns, not significant.

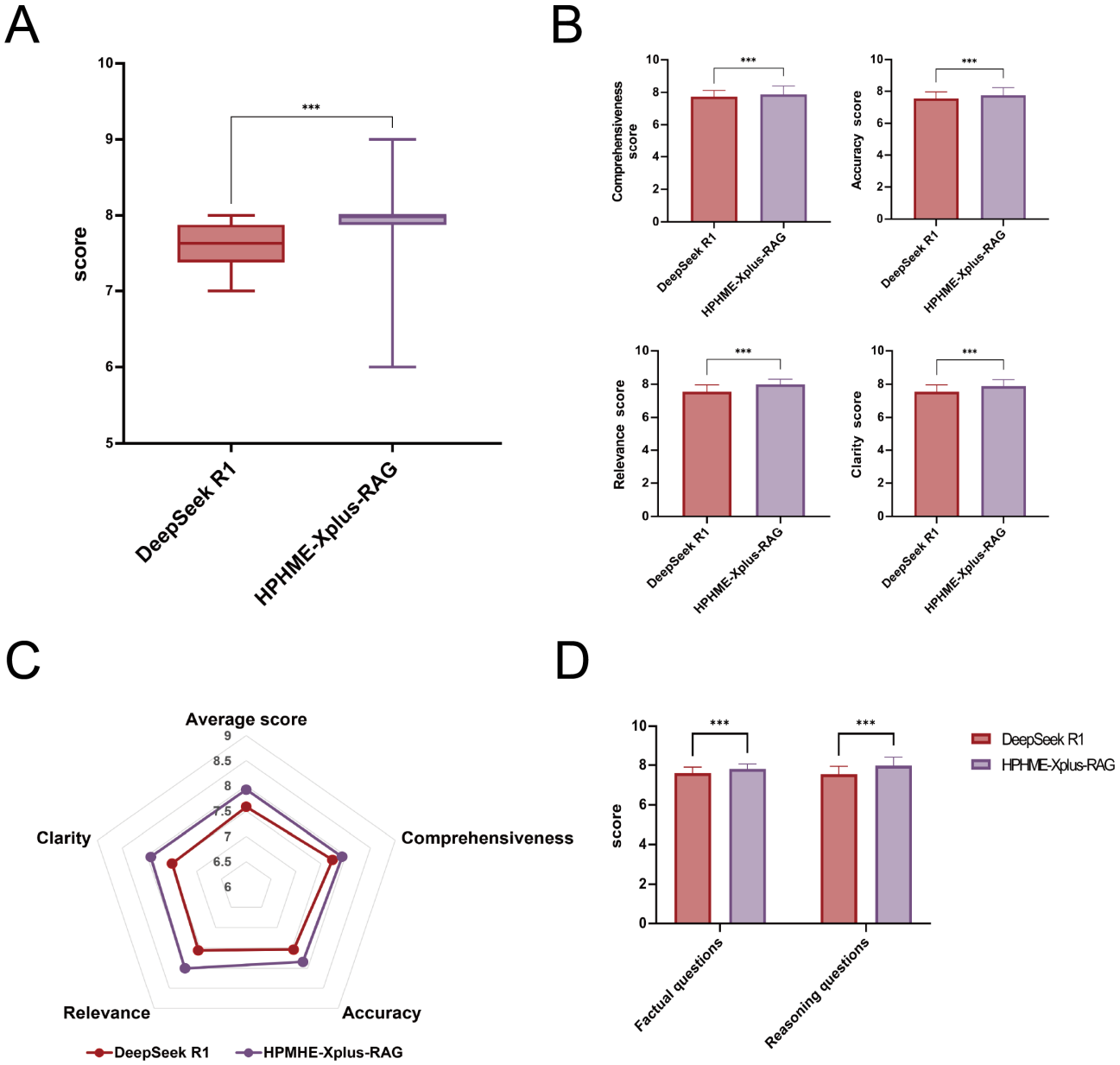

Discussion

Principal Findings

High-altitude public health medical education is of critical importance in low-resource regions, yet it faces substantial challenges [1-12,14-20]. This study systematically evaluated 4 LLMs (DeepSeek R1, GPT-5.2, Gemini 3.0 Pro, and Tencent HY 2.0) on a set of 80 expert-designed questions covering high-altitude public health medical education. Among the 4 models, DeepSeek R1 achieved the highest weighted composite score, integrating multidimensional first-response performance with composite consistency. Building upon this, we developed HPHME-Xplus-RAG, a RAG-based system integrating 4 authoritative textbooks with DeepSeek R1 as the base model. HPHME-Xplus-RAG demonstrated significant and consistent improvement in overall multidimensional performance compared

XSL•FO
RenderX

to baseline DeepSeek R1, confirming that integrating domain-authoritative knowledge via RAG can meaningfully enhance LLM performance in a resource-limited, highly specialized medical domain where specialized model development is neither practical nor cost-effective [30,43,58].

Comparison to Relevant Literature, Interpretations, and Implications

This study systematically unveils and examines the generative capabilities of LLMs within the domain of high-altitude medicine. Similar to findings reported in other domains, the results of this study strongly support that general LLMs possess the fundamental capability to perform medical education and other tasks; however, their performance remains inconsistent and the overall quality requires further improvement [31-33]. Studies evaluating LLMs in other medical disciplines have also demonstrated the strong performance of the DeepSeek R1 model, while some reports have noted its relatively poorer results in specific specialty tasks [31,34-36]. Differences in model selection, task design, and evaluation results across studies suggest that the capabilities of general LLMs vary across medical specialties and task scenarios. For instance, Tordjman et al [31] demonstrated that DeepSeek R1 achieved an accuracy of 0.92 on the United States Medical Licensing Examination and showed particularly strong diagnostic reasoning performance, though it underperformed relative to ChatGPT-o1 in radiology report summarization. Similarly, a clinical decision-making benchmark involving 125 patient cases found that DeepSeek R1 performed comparably to or better than GPT-4o across diagnosis and treatment recommendation tasks [35].

Consistent with the majority of previous studies, the application of RAG technology led to a significant improvement in model performance within the target domain [59-64]. It is worth noting that this study innovatively used authoritative textbooks as the external knowledge source, providing a new perspective on the selection of authoritative domain knowledge for RAG-based systems and ensuring that authoritative books, which embody knowledge accumulated over generations, can be preserved in a way that keeps pace with the digital era. This approach is particularly applicable to domains lacking formal clinical guidelines, where authoritative textbooks serve as the most reliable and comprehensive knowledge source available.

Notably, RAG significantly improved performance across all 4 evaluated dimensions (Figure 6), indicating that the integration of authoritative textbooks enhanced the overall quality of model responses. However, this improvement was not uniformly distributed across dimensions. The overall gain appeared to be driven primarily by larger improvements in clarity and relevance, although significant benefits were also observed in comprehensiveness and scientific accuracy (Figure 6). This pattern suggests that retrieval from authoritative textbooks may be particularly effective in helping the model generate responses that are better focused, more interpretable, and more closely aligned with the educational intent of the questions, while still strengthening factual grounding. In addition, the larger gain observed for reasoning questions than for factual questions suggests that the retrieved domain knowledge may support not only information recall but also the integration and application of knowledge in more complex response-generation scenarios. Mechanistically, this is consistent with the nature of the external knowledge base used in our RAG pipeline: authoritative textbooks provide systematic, well-structured, and context-rich information that can guide the model toward more coherent and educationally relevant outputs. While the current retrieval pipeline performed well, future work could further explore strategies such as query expansion or multidocument summarization to optimize information extraction for highly complex queries.

Regarding the magnitude of improvement, the observed gain of approximately 0.37 points on the 10-point scale (~4% relative improvement) is modest compared to the larger absolute gains reported in studies using binary accuracy metrics, although the effect is large enough across all dimensions. For example, Ferber et al [60] reported an improvement from 57% to 84% correct responses with RAG applied to oncology guidelines, and Alexandrou et al [64] reported accuracy improvements of approximately 16 - 21 percentage points for DeepSeek R1 and ChatGPT-4o on acute coronary syndrome guidelines. However, those studies used binary accuracy outcomes on a different measurement scale, making direct comparison of absolute effect size inappropriate. In addition, the relatively high baseline performance of DeepSeek R1 in our study (median 7.63/10, IQR 0.5) inherently limited the room for further observable improvement. As performance approaches the upper end of the scale, achieving a breakthrough-level absolute increase becomes progressively more difficult, reflecting a typical ceiling effect. Therefore, although the absolute score change was not dramatic, the consistently significant improvements across all evaluated dimensions, together with the large effect sizes, indicate that the benefit of RAG in our setting was both robust and clinically meaningful.

The study goes beyond conventional accuracy assessments by establishing a 4-dimensional evaluation framework encompassing comprehensiveness, scientific accuracy, relevance, and clarity [65]. Although only the models' first responses were scored to simulate real-world application scenarios, the study innovatively incorporated a consistency evaluation. Rather than relying solely on algorithmic similarity, we proposed a composite consistency metric that integrates quantitative algorithmic indicators with qualitative expert semantic assessments. Furthermore, based on clinical needs, the study adopted a weighted standard that combines first-response scores with composite consistency, providing a more comprehensive and realistic reflection of the models' reliability and stability in actual clinical settings [65].

In addition, this study applies cutting-edge generative artificial intelligence (AI) technology to a specific, critical, and resource-limited domain—high-altitude medicine—providing a replicable paradigm for AI applications in specialized medical fields. It demonstrates how advanced medical expertise, such as authoritative textbooks, can be rapidly transformed into accessible educational resources through technological innovation. By linking the generation process to authoritative high-altitude medicine literature, the system enhances interpretability and credibility, marking a key step toward the

XSL•FO
RenderX

practical integration of AI into clinical and educational workflows.

Limitations

Several limitations of this study warrant acknowledgment. First, this study focused solely on text generation performance within the context of high-altitude public health medical education; the models' capabilities across other modalities and task types were not assessed. Second, the entire evaluation was conducted by medical experts, while the intended end users, such as medical instructors and students, frontline health care workers, residents, and travelers, were not directly involved in evaluating system outputs, which may limit the ecological validity of the findings. In addition, the single-system reassessment design cannot fully exclude expectancy effects, even with source masking and temporal separation. Future studies should incorporate real-world user evaluations to complement expert assessments with more rigorous blinded methodology.

Another limitation relates to the benchmark and generalizability of the findings. Although the 80-question benchmark was purposively designed to cover the major domains of high-altitude public health medicine, the selected question set may still not fully capture the breadth and heterogeneity of real-world problems in this field, with an average of approximately 11 questions per domain. Future studies should therefore expand the benchmark to include a broader and more diverse range of clinical, preventive, educational, and public health scenarios, so as to strengthen domain-specific evaluation and real-world generalizability. Moreover, because the benchmark was developed and validated within the specific geographic, cultural, and practice context of the Xizang Autonomous Region, caution is warranted when extrapolating these findings to other low-resource regions, health systems, or medical specialties. Future studies should therefore conduct cross-regional and cross-domain external validation.

Several practical and methodological issues also remain to be addressed. The study does not assess practical deployment considerations such as response latency, cost per query, or system reliability under real-world conditions. Moreover, as LLM application programming interfaces are subject to continuous updates, longitudinal comparisons may be affected by model versioning, complicating reproducibility over time. Finally, the observed score distribution, with maximum values concentrated in the 8 - 9 range rather than at the full score of 10, raises the possibility of a ceiling effect. Although the current evaluation framework was sufficient for the aims of this study—being developed through expert discussion, built upon established approaches in the literature, and implemented using a 10-point scale with 0.25-point intervals that provided reasonably fine discrimination—the compression of scores near the upper bound may still limit its sensitivity for distinguishing among very high-performing models. As stronger LLMs continue to emerge, such clustering at the top of the scale may make further absolute improvements harder to detect, even when meaningful qualitative differences exist. Nevertheless, despite using a clinically oriented quantitative framework that combined first-response scoring with composite consistency weighting, the inherent probabilistic nature of LLM outputs cannot be eliminated [43,58]. Future research may therefore benefit from adopting more sensitive rating instruments, such as scales with a wider range, to improve discriminative precision when evaluating increasingly capable systems.

Conclusions

In summary, this study presents a replicable workflow integrating domain-authoritative knowledge, RAG, and large-model optimization and evaluation for high-altitude public health medical education in the low-resource Xizang Autonomous Region. By integrating authoritative textbooks with an evaluation-optimized general LLM through a RAG framework, the study demonstrated strong performance in a highly specialized educational domain with limited resources. Unlike prior RAG studies that mainly relied on clinical guidelines or focused on diagnosis and treatment tasks, this work used authoritative textbooks for the broader, guideline-scarce field of public health medical education, thereby transforming textbook knowledge into a reliable, specialized, and scalable external knowledge source for LLMs. This approach is particularly suited to highly specialized domains such as high-altitude medicine, where formal clinical guidelines remain limited, and offers a practical way to preserve and deliver authoritative knowledge in the digital era. More broadly, the paradigm of domain-authoritative knowledge+RAG+model optimization and evaluation provides a scalable framework for building trustworthy AI tools in other resource-limited medical and professional settings. With further real-world validation, systems such as HPHME-Xplus-RAG may support clinical and educational practice by providing cost-effective, convenient, and reliable access to medical knowledge.

Acknowledgments

JT is co-corresponding author (email: ji.tu@ibms.pumc.edu.cn; phone: 86 18519395290). The authors would like to thank Tao Wang for his invaluable advice in the preparation of this manuscript.

The authors confirm that no generative artificial intelligence was used in the preparation of this manuscript.

Funding

This research was supported by Beijing Research Ward Excellence Program (BRWEP2024W034010103); National Key Research and Development Program of China (2024YFA0918504); Science and Technology Projects of Xizang Autonomous Region, China (XZ202501JD0021, XZ202501ZY0072); Non-profit Central Research Institute Fund of Chinese Academy of Medical

XSL•FO
RenderX

Sciences (2024-RW320-01); National High Level Hospital Clinical Research Funding (2025-PUMCH-C-048); Medical Artificial Intelligence Educational Reform Project of Institute of Basic Medical Sciences (2022jcjg0104); The Ministry of Education 2022 Baidu industry-academy cooperation Collaborative Education Project - Medical artificial intelligence curriculum construction (182215PC08768); The 2025 Medical Education Scholars Program at Peking Union Medical College (2025mesp010); 2024 CCF-Baidu Open Fund Project: Construction and Application Evaluation of Benchmark Dataset for Multimodal Collaborative Medical Large Language Model (CCF-BAIDUOF202418).

Data Availability

The shareable materials generated and analyzed during this study are provided in the multimedia appendices. Multimedia Appendix 1 contains the 80-question benchmark with labels and corresponding reference answers; Multimedia Appendix 2 contains the model-generated outputs; and Multimedia Appendix 3 contains the supplementary statistical tables, including dimension-specific descriptive statistics and reliability analyses. Additional noncopyrighted materials, such as deidentified scoring tables used for analysis and brief statistical analysis and figure-generation code, are available from the corresponding author upon reasonable request.

The copyrighted textbook contents used to construct the retrieval-augmented generation knowledge base cannot be publicly shared or redistributed. These materials were used only in a closed internal research environment for retrieval-corpus construction.

Authors' Contributions

DW, TJ, and KH conceived the study. KH, QX, WC, and LJ designed the study. KH, YW, RC, JL, LC, YZ, ZZ, Ran Li, LY, Rutong Li, HW, XY, GZ, SC, QS, QW, DZ, CY, and GL designed the questions and performed the evaluation. SL, DY, HX, KP, RX, ZL, and DZ contributed to the statistical analysis. KH, QX, WC, and LJ wrote the original manuscript. DW and TJ revised the manuscript. KH, QX, WC, and LJ contributed equally to this work and are joint first authors. DW and TJ contributed equally and are joint corresponding authors and the guarantors. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The 80-question benchmark with task-type labels and corresponding reference answers used for model evaluation.
[XLSX File, 62 KB - jmir_v28i1e92852_app1.xlsx]

Multimedia Appendix 2
Outputs of DeepSeek R1, Gemini 3.0 Pro, Tencent HY 2.0, GPT-5.2, and HPHME-Xplus-RAG queried with the 80-question compiled set.
[XLSX File, 1665 KB - jmir_v28i1e92852_app2.xlsx]

Multimedia Appendix 3
Supplementary descriptive statistics and interpanel reliability analyses for overall and dimension-specific scores. The tables include results for the 4 large language models and HPHME-Xplus-RAG.
[DOCX File, 48 KB - jmir_v28i1e92852_app3.docx]

Checklist 1
iCHECK-DH checklist.
[DOCX File, 28 KB - jmir_v28i1e92852_app4.docx]

References

1. Dorji T, Giri S, Tshering U, et al. Challenges in the management of high-altitude illnesses and emergencies in Bhutan and Nepal. Travel Med Infect Dis 2023;56:102660. [doi: 10.1016/j.tmaid.2023.102660] [Medline: 37926372]
2. Bardwell WA, Ensign WY, Mills PJ. Negative mood endures after completion of high-altitude military training. Ann Behav Med 2005 Feb;29(1):64-69. [doi: 10.1207/s15324796abm2901_9] [Medline: 15677302]
3. Rajesh YS, Bhat PS, Srivastava K, Kanitkar M. Study of psychological effects of deployment at extreme high altitude area on soldiers. Med J Armed Forces India 2021 Oct;77(4):426-430. [doi: 10.1016/j.mjafi.2020.06.002] [Medline: 34594071]
4. Levine BD, Zuckerman JH, deFilippi CR. Effect of high-altitude exposure in the elderly: the Tenth Mountain Division study. Circulation 1997 Aug 19;96(4):1224-1232. [doi: 10.1161/01.cir.96.4.1224] [Medline: 9286953]
5. Wang L, Sang L, Cui Y, et al. Effects of acute high-altitude exposure on working memory: a functional near-infrared spectroscopy study. Brain Behav 2022 Dec;12(12):e2776. [doi: 10.1002/brb3.2776] [Medline: 36321845]

XSL•FO
RenderX

6. Liu Y, Zhang JH, Gao XB, et al. Correlation between blood pressure changes and AMS, sleeping quality and exercise upon high-altitude exposure in young Chinese men. Military Med Res 2014 Dec;1(1):19. [doi: 10.1186/2054-9369-1-19]
7. Kumar K, Shijith KP, Ahmad FMH. Bell's palsy at high altitude -- an unsuspected finding. Military Med Res 2016 Dec;3(1). [doi: 10.1186/s40779-016-0073-6]
8. Qu H, Yu LJ, Wu JT, et al. Spine system changes in soldiers after load carriage training in a plateau environment: a prediction model research. Military Med Res 2020 Dec;7(1). [doi: 10.1186/s40779-020-00293-1]
9. Hussain MS, Tripathi V. Smoking under hypoxic conditions: a potent environmental risk factor for inflammatory and autoimmune diseases. Military Med Res 2018 Dec;5(1). [doi: 10.1186/s40779-018-0158-5]
10. Mallet RT, Burtscher J, Richalet JP, Millet GP, Burtscher M. Impact of high altitude on cardiovascular health: current perspectives. Vasc Health Risk Manag 2021;Volume 17:317-335. [doi: 10.2147/VHRM.S294121]
11. Yang S, Zhao H, Zhang H, et al. Current status and continuing medical education need for general practitioners in Tibet, China: a cross-sectional study. BMC Med Educ 2024 Mar 8;24(1):265. [doi: 10.1186/s12909-024-05143-5] [Medline: 38459539]
12. Hou J, Ke Y. Addressing the shortage of health professionals in rural China: issues and progress comment on "Have health human resources become more equal between rural and urban areas after the new reform?". Int J Health Policy Manag 2015 Mar 7;4(5):327-328. [doi: 10.15171/ijhpm.2015.57] [Medline: 25905487]
13. Tong W, Zhang X, Zeng H, Pan J, Gong C, Zhang H. Reforming China's secondary vocational medical education: adapting to the challenges and opportunities of the AI era. JMIR Med Educ 2024 Aug 15;10:e48594. [doi: 10.2196/48594] [Medline: 39149865]
14. Chen H, Wang Z, Zheng L, Ciren J, Zhang W, Zhu T. Anesthesiology education: a discussion on the current reform for training in the hypoxic environment of the highlands of China. Front Educ 2023 Jan 13;7:1041500. [doi: 10.3389/feduc.2022.1041500]
15. Li X, Zhang YL, Shen L, Huang YG. Remote video lecture education improves anesthesiology residency training in Tibet of China. Chin Med J 2020 Jan 20;133(2):245-246. [doi: 10.1097/CM9.0000000000000605]
16. Pollard AJ, Niermeyer S, Barry P, et al. Children at high altitude: an international consensus statement by an ad hoc committee of the International Society for Mountain Medicine, March 12, 2001. High Alt Med Biol 2001;2(3):389-403. [doi: 10.1089/15270290152608561] [Medline: 11682018]
17. Parati G, Agostoni P, Basnyat B, et al. Clinical recommendations for high altitude exposure of individuals with pre-existing cardiovascular conditions. Eur Heart J 2018 May 1;39(17):1546-1554. [doi: 10.1093/eurheartj/ehx720]
18. León-Velarde F, Maggiorini M, Reeves JT, et al. Consensus statement on chronic and subacute high altitude diseases. High Alt Med Biol 2005;6(2):147-157. [doi: 10.1089/ham.2005.6.147] [Medline: 16060849]
19. Luks AM, Beidleman BA, Freer L, et al. Wilderness Medical Society Clinical Practice Guidelines for the prevention, diagnosis, and treatment of acute altitude illness: 2024 update. Wilderness Environ Med 2024 Mar;35(1_suppl):2S-19S. [doi: 10.1016/j.wem.2023.05.013] [Medline: 37833187]
20. Gatterer H, Villafuerte FC, Ulrich S, Bhandari SS, Keyes LE, Burtscher M. Altitude illnesses. Nat Rev Dis Primers 2024 Jun 20;10(1):43. [doi: 10.1038/s41572-024-00526-w] [Medline: 38902312]
21. Luitse D, Denkena W. The great Transformer: examining the role of large language models in the political economy of AI. Big Data Soc 2021 Jul;8(2):20539517211047734. [doi: 10.1177/20539517211047734]
22. Peng C, Yang X, Chen A, et al. A study of generative large language model for medical research and healthcare. NPJ Digit Med 2023 Nov 16;6(1):210. [doi: 10.1038/s41746-023-00958-w] [Medline: 37973919]
23. Wang D, Zhang S. Large language models in medical and healthcare fields: applications, advances, and challenges. Artif Intell Rev 2024 Sep;57(11). [doi: 10.1007/s10462-024-10921-0]
24. Meng X, Yan X, Zhang K, et al. The application of large language models in medicine: a scoping review. iScience 2024 May 17;27(5):109713. [doi: 10.1016/j.isci.2024.109713] [Medline: 38746668]
25. Mu LJ, Wang TT, Miao YD. Advancements in AI-driven oncology: assessing ChatGPT's impact from GPT-3.5 to GPT-4o. Int J Surg 2025 Jan 1;111(1):1669-1670. [doi: 10.1097/JS9.0000000000001989] [Medline: 39041947]
26. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; Dec 4-9, 2017; Long Beach, CA, USA p. 6000-6010. [doi: 10.5555/3295222.3295349]
27. Blank IA. What are large language models supposed to model? Trends Cogn Sci (Regul Ed) 2023 Nov;27(11):987-989. [doi: 10.1016/j.tics.2023.08.006]
28. Fu Q, Yuan X, Han X, Wang W, Zhang J, Yuan X. Enhancing the diagnostic accuracy of sacroiliitis: a machine learning approach applied to computed tomography imaging. Br J Hosp Med 2024 Aug 30;85(8):1-13. [doi: 10.12968/hmed.2024.0228]
29. Fan S, Hong H, Lu S, et al. Artificial intelligence-based predictive model for relapse in acute myeloid leukemia patients following haploidentical hematopoietic cell transplantation. J Transl Int Med 2025 Jun;13(3):253-266. [doi: 10.1515/jtim-2025-0028] [Medline: 40896291]
30. Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A review of large language models in medical education, clinical decision support, and healthcare administration. Healthcare (Basel) 2025 Mar 10;13(6):603. [doi: 10.3390/healthcare13060603] [Medline: 40150453]

XSL•FO
RenderX

31. Tordjman M, Liu Z, Yuce M, et al. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. Nat Med 2025 Aug;31(8):2550-2555. [doi: 10.1038/s41591-025-03726-3] [Medline: 40267969]
32. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating chatgpt as an adjunct for radiologic decision-making. Radiology and Imaging. Preprint posted online on Feb 7, 2023. [doi: 10.1101/2023.02.02.23285399]
33. Zhou S, Luo X, Chen C, et al. The performance of large language model-powered chatbots compared to oncology physicians on colorectal cancer queries. Int J Surg 2024 Oct 1;110(10):6509-6517. [doi: 10.1097/JS9.0000000000001850] [Medline: 38935100]
34. Wu X, Huang Y, He Q. A large language model improves clinicians' diagnostic performance in complex critical illness cases. Crit Care 2025 Jun 6;29(1):230. [doi: 10.1186/s13054-025-05468-7] [Medline: 40481529]
35. Sandmann S, Hegselmann S, Fujarski M, et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. Nat Med 2025 Aug;31(8):2546-2549. [doi: 10.1038/s41591-025-03727-2] [Medline: 40267970]
36. Han T, Adams LC, Bressem KK, Busch F, Nebelung S, Truhn D. Comparative analysis of multimodal large language model performance on clinical vignette questions. JAMA 2024 Apr 16;331(15):1320-1321. [doi: 10.1001/jama.2023.27861] [Medline: 38497956]
37. Aster A, Laupichler MC, Rockwell-Kollmann T, Masala G, Bala E, Raupach T. ChatGPT and other large language models in medical education - scoping literature review. Med Sci Educ 2025 Feb;35(1):555-567. [doi: 10.1007/s40670-024-02206-6] [Medline: 40144083]
38. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med 2024 Oct 29;11:1477898. [doi: 10.3389/fmed.2024.1477898]
39. Zhang Y, Xie X, Xu Q. ChatGPT in medical education: bibliometric and visual analysis. JMIR Med Educ 2025 Oct 7;11:e72356. [doi: 10.2196/72356] [Medline: 41056572]
40. Zhui L, Fenghe L, Xuehu W, Qining F, Wei R. Ethical considerations and fundamental principles of large language models in medical education: viewpoint. J Med Internet Res 2024 Aug 1;26:e60083. [doi: 10.2196/60083] [Medline: 38971715]
41. Abouammoh N, Alhasan K, Aljamaan F, et al. Perceptions and earliest experiences of medical students and faculty with ChatGPT in medical education: qualitative study. JMIR Med Educ 2025 Feb 20;11:e63400. [doi: 10.2196/63400] [Medline: 39977012]
42. Fan W, Ding Y, Ning L, et al. A survey on RAG meeting llms: towards retrieval-augmented large language models. 2024 Aug 25 Presented at: KDD '24; Aug 25-29, 2024; Barcelona, Spain p. 6491-6501. [doi: 10.1145/3637528.3671470]
43. Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst 2025 Mar 31;43(2):1-55. [doi: 10.1145/3703155]
44. Wang D, Liang J, Ye J, et al. Enhancement of the performance of large language models in diabetes education through retrieval-augmented generation: comparative study. J Med Internet Res 2024 Nov 8;26:e58041. [doi: 10.2196/58041] [Medline: 39046096]
45. Miao Y, Zhao Y, Luo Y, Wang H, Wu Y. Improving large language model applications in the medical and nursing domains with retrieval-augmented generation: scoping review. J Med Internet Res 2025 Oct 21;27:e80557. [doi: 10.2196/80557] [Medline: 41118646]
46. Chen X, Wang L, You M, et al. Evaluating and enhancing large language models' performance in domain-specific medicine: development and usability study with DocOA. J Med Internet Res 2024 Jul 22;26:e58158. [doi: 10.2196/58158] [Medline: 38833165]
47. Li H, Huang J, Ji M, Yang Y, An R. Use of retrieval-augmented large language model for COVID-19 fact-checking: development and usability study. J Med Internet Res 2025 Apr 30;27:e66098. [doi: 10.2196/66098]
48. Update to GPT-5 system card: GPT 5.2. : OpenAI; 2025 URL: https://cdn.openai.com/pdf/3a4153c8-c748-4b71-8e31-aecbde944f8d/oai_5_2_system-card.pdf [accessed 2026-05-26]
49. Gemini 3 pro model card. Google. 2025. URL: https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf [accessed 2026-01-03]
50. Tencent HY 2.0 product oeverview. Tencent Cloud. 2025. URL: https://cloud.tencent.com/document/product/1729/104753 [accessed 2026-01-23]
51. DeepSeek-R1: incentivizing reasoning capability in llms via reinforcement learning. GitHub. 2025. URL: https://github.com/deepseek-ai/DeepSeek-R1/blob/main/DeepSeek_R1.pdf [accessed 2026-01-03]
52. Cer D, Diab M, Agirre E, Lopez-Gazpio I, Specia L. SemEval-2017 task 1: semantic textual similarity multilingual and crosslingual focused evaluation. In: Bethard S, Carpuat M, Apidianaki M, Mohammad SM, Cer D, Jurgens D, editors. Presented at: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017); Aug 3-4, 2017; Vancouver, Canada p. 1-14. [doi: 10.18653/v1/S17-2001]
53. Gerily G. High Altitude Medicine, 2nd edition: Peking University Medical Press; 2020. URL: https://www.pumpedu.com/shop/5722.html [accessed 2026-06-04]
54. Yuqi G. High Altitude Medical Protection Knowledge: Science Press; 2017. URL: https://www.books.com.tw/products/CN11431887 [accessed 2026-06-04]
55. Hidalgo J, Re S, D'Almeida AG, editors. High Altitude Medicine: A Case-Based Approach: Springer International Publishing; 2023. [doi: 10.1007/978-3-031-35092-4]

XSL•FO
RenderX

56. Luks AM, Ainslie PN, Lawley JS, Roach RC, Simonson TS. Ward, Milledge and West's High Altitude Medicine and Physiology, 6th edition: CRC Press; 2021. [doi: 10.1201/9780429444333]
57. High altitude medicine: a collaborative model for books. Peking Union Medical College Hospital, People's Hospital of Tibet Autonomous. URL: http://pumc.vip:8001/PUMCH/ [accessed 2026-06-15]
58. Shah SV. Accuracy, consistency, and hallucination of large language models when analyzing unstructured clinical notes in electronic medical records. JAMA Netw Open 2024 Aug 1;7(8):e2425953. [doi: 10.1001/jamanetworkopen.2024.25953] [Medline: 39136951]
59. Zakka C, Shad R, Chaurasia A, et al. Almanac - retrieval-augmented language models for clinical medicine. NEJM AI 2024 Feb;1(2). [doi: 10.1056/aioa2300068] [Medline: 38343631]
60. Ferber D, Wiest IC, Wölflein G, et al. GPT-4 for information retrieval and comparison of medical oncology guidelines. NEJM AI 2024 May 23;1(6):AIcs2300235. [doi: 10.1056/AIcs2300235]
61. Lewis M, Thio S, Dobson RJ, Denaxas S. Grounding large language models in clinical evidence: a retrieval-augmented generation system for querying UK NICE clinical guidelines. arXiv. Preprint posted online on Oct 3, 2025. [doi: 10.48550/arXiv.2510.02967]
62. Kresevic S, Giuffrè M, Ajcevic M, Accardo A, Crocè LS, Shung DL. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. NPJ Digit Med 2024 Apr 23;7(1):102. [doi: 10.1038/s41746-024-01091-y] [Medline: 38654102]
63. Ke YH, Jin L, Elangovan K, et al. oRetrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. arXiv. Preprint posted online on Oct 11, 2024. [doi: 10.48550/arXiv.2410.08431]
64. Alexandrou M, Kumar S, Mahtani AU, et al. Performance of large language models on the acute coronary syndrome guidelines using retrieval-augmented generation. JACC Cardiovasc Interv 2025 Oct 27;18(20):2458-2467. [doi: 10.1016/j.jcin.2025.08.019] [Medline: 41161918]
65. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Med Inform Decis Mak 2025 Mar 7;25(1):117. [doi: 10.1186/s12911-025-02954-4] [Medline: 40055694]

Abbreviations

AI: artificial intelligence
HPHME-Xplus-RAG: High-Altitude Public Health Medical Education–Xplus (Xiehe in Chinese for Peking Union Medical College Hospital and Xizang for People's Hospital of Xizang Autonomous Region)–Retrieval-Augmented Generation
ICC: intraclass correlation coefficient
LLM: large language model
PUMCH: Peking Union Medical College Hospital
RAG: retrieval-augmented generation

Edited by S Brini; submitted 04.Feb.2026; peer-reviewed by W Xiao, W Qi, Y Yu; revised version received 13.May.2026; accepted 14.May.2026; published 16.Jun.2026.

Please cite as:
He K, Xiao Q, Chen W, Jing L, Wang Y, Li S, Yang D, Xu H, Pang K, Xiao R, Liu Z, Zhuoga D, Chen R, Li J, Chang L, Zhou Y, Zhang Z, Li R, Ying L, Li R, Wang H, Yin X, Zhen G, Cai S, Shan Q, Wang Q, Zhuoga D, Yangjin C, Luobu G, Ji T, Wu D
Authoritative Textbook-Augmented Large Language Models for High-Altitude Public Health Medical Education in the Xizang Autonomous Region: Cross-Sectional Comparative Evaluation Study
J Med Internet Res 2026;28:e92852
URL: https://www.jmir.org/2026/1/e92852
doi:10.2196/92852

© Kun He, Qiming Xiao, Wangyang Chen, Lisha Jing, Yabing Wang, Shuai Li, Daiyu Yang, Hemiao Xu, Ke Pang, Ruoyu Xiao, Zhashilamu Liu, Deji Zhuoga, Ruxuan Chen, Jingyi Li, Long Chang, Yangzhong Zhou, Zhe Zhang, Ran Li, Lujing Ying, Rutong Li, Hongwei Wang, Xin Yin, Ge Zhen, Siyi Cai, Qijun Shan, Qiang Wang, Danzeng Zhuoga, Ciren Yangjin, Gesang Luobu, Tu Ji, Dong Wu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Training Clinicians in Private Practice in Family-Based Treatment for Anorexia Nervosa: Randomized Controlled Trial Comparing Two Online Approaches

James Lock[1], MD, PhD; Daniel Le Grange[2,3], PhD; Brittany Matheson[1], PhD; Bohye Kim[1], MS; Cara Bohon[1], PhD; Booil Jo[1], PhD

[1]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, 401 Quarry Road, Stanford, CA, United States
[2]Department of Psychiatry and Behavioral Sciences, University of California, San Francisco (emeritus), San Francisco, CA, United States
[3]Department of Psychiatry and Behavioral Neuroscience, University of Chicago (emeritus), Chicago, IL, United States

Corresponding Author:
Brittany Matheson, PhD
Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, 401 Quarry Road, Stanford, CA, United States

Abstract

Background: There is a critical need to disseminate efficacious psychosocial treatments for mental disorders because there is a significant gap between evidence-based treatment (EBT) approaches and usual clinical practice. To address this gap, cost-effective, efficient, and scalable methods of training mental health clinicians in EBT are needed. One example of the need to improve dissemination and implementation of psychosocial treatments is for adolescents with anorexia nervosa (AN), a serious mental disorder with an incidence rate of about 1% that can become life-threatening. Based on outcomes from a series of randomized controlled trials (RCTs), an EBT for adolescents with AN is family-based treatment (FBT); however, few therapists are trained in the approach. Some studies suggest that online training is feasible for training clinicians treating eating disorders, including pilot data related to FBT specifically, but RCTs examining different training formats for FBT to improve fidelity and patient outcomes are needed.

Objective: This study compared 2 different formats for delivering online training in FBT to therapists treating adolescents with AN and planned to assess the feasibility of the 2 training formats, as well as to examine whether either approach was superior in improving fidelity to FBT or patient outcomes.

Methods: Participants were 123 mental health therapists licensed in the United States in private practice with no previous FBT training. Therapists were randomized to either (1) a webinar training lecture series or (2) an interactive e-training on-demand program (ET-FBT) with an additional focus on key FBT interventions. Both groups were offered up to 12 one-hour group-based clinical case consultation (CCC) sessions posttraining. We examined the feasibility and acceptability of the online training programs and posttraining outcomes in fidelity to key components of FBT, self-efficacy, and working alliance. We explored rates of patient early response (weight gain of 2.4 kg at session 4) in adolescent patients with AN treated after FBT training during CCC.

Results: Both online trainings had a high completion rate of 95% (117/123), with CCC completion at 38% (47/123). Both programs showed significant improvements within randomized groups in therapists' fidelity, self-efficacy, and working alliance. Working alliance improved significantly more in the ET-FBT group, but there were no other significant differences between training groups. Early response rates doubled posttraining and CCC (16%-34%), regardless of randomized training format, a rate that is similar to that achieved by therapists in RCTs examining FBT treatment outcomes.

Conclusions: It is feasible to deliver online training in FBT to clinicians in private practice in 2 different formats, and both trainings led to significant improvements in fidelity, self-efficacy, working alliance, and patient outcomes. Future studies should address challenges in patient recruitment for posttraining CCC and refine CCC implementation to maximize training effects and efficiency.

Trial Registration: ClinicalTrials.gov NCT04428580; https://clinicaltrials.gov/study/NCT04428580

International Registered Report Identifier (IRRID): RR2-doi: 10.1016/j.cct.2022.106889

(J Med Internet Res 2026;28:e89999) doi:10.2196/89999

KEYWORDS

family-based treatment; anorexia nervosa; adolescents; implementation; evidence-based treatment; online training

XSL•FO
RenderX

Introduction

There is a critical need to disseminate efficacious psychosocial treatments for mental disorders as there is a significant gap between evidence-based treatment (EBT) approaches and usual clinical practice [1-4]. One factor contributing to this gap is the lack of trained therapists [2,5]. The usual method for training therapists is in personal training workshops followed by supervision from experts, an approach that is both labor-intensive and costly [3]. One strategy to address these problems is the use of online training to more efficiently scale up training [6]. In a recent review of web-based training for mental disorders, including treatments for depression, anxiety, trauma, and eating disorders, the authors concluded that web-based training improved self-reported fidelity and outcomes, but additional studies were needed as training approaches were highly heterogeneous [7].

One area where there is a significant training need is for therapists treating eating disorders. Some studies suggest that online training is feasible for training clinicians treating eating disorders [8,9]. For example, Brownlow et al [10] found that an online training program for clinicians treating eating disorders improved knowledge and skills among participants. More recently, D'Adamo et al [8] examined the effects of online guided self-help cognitive behavioral therapy (CBT) for eating disorders and found the training platform and content to be feasible and acceptable to clinicians and families. Another study explored online training for interpersonal psychotherapy for eating disorders in a case series of 60 clinicians and found that the training was feasible, acceptable, and led to clinical improvements in patients [9]. No scaled randomized controlled trials (RCTs) have been conducted to examine online training for eating disorders.

One specific example of the need to improve the dissemination and implementation of psychosocial treatments is adolescents with anorexia nervosa (AN), a serious mental disorder with an incidence rate of about 1% [11-14] that can become life-threatening [15-17]. Based on outcomes from a series of RCTs, an EBT for adolescents with AN is FBT; however, few therapists are trained in the use of FBT for AN [18-20]. A significant amount of treatment for adolescents with AN is provided by clinicians in private practice, but most are not trained in EBT, and many patients do not receive EBT of any kind [21]. Motivations, incentives, and rationale for learning EBTs differ between therapists embedded in an organization or health care system and those in private practice [22-26]. Among the most salient differences is the personal financial impact of training costs and time allocation for clinicians in private practice rather than this being an institutional or programmatic burden. Nonetheless, clinicians in private practice are expected to follow clinical guidelines and adhere to professional community standards of practice, and they are, therefore, motivated to learn EBT when these treatments are desired by their clients, paid for by insurers, enhance their reputation as experts, and meet continuing education requirements [27].

Before embarking on an RCT examining online training for FBT-AN, we initially conducted a pilot study comparing 2 forms of online FBT training [28]. This pilot study had two principal aims: (1) to evaluate the feasibility and acceptability of randomization and the tolerability of study procedures by clinicians undergoing online training in FBT for adolescents with AN, and (2) to compare the relative efficacy of improving therapist competence in externalization and agnosticism in FBT (key components of FBT interventions) using 2 versions of online training (standard training FBT or enhanced training FBT [ET-FBT]). Standard training included training across all interventions of FBT, whereas the enhanced training provided additional emphasis by including short videos on externalization and agnosticism as 2 key components in FBT. To assess therapeutic competence, therapists were asked to complete an assessment focused on agnosticism and externalization. These key components are interventions used in FBT-AN that focus on reducing guilt and self-blame in parents by emphasizing that the cause of AN is unknown (ie, agnosticism) and separating the symptoms of AN from the identity and developmental issues of adolescents (ie, externalization) [29,30]. The results of this pilot study of the online trainings suggested that both trainings were generally acceptable, but attrition was 50% at the end of training (EOT), few participated in posttraining clinical case consultation (CCC), and only 10% of patient clinical outcomes were obtained [28]. There were few differences between the groups, except that the ET group showed greater improvements related to externalization and agnosticism skills, as expected [28]. Furthermore, the online training appears to have led to improvements specifically related to greater skill in some key components of FBT [28].

In this report, we describe the outcomes of a study comparing 2 forms of online training programs for FBT for adolescents with AN (FBT-AN) offered to therapists in private practice in the United States, followed by 12 hours of group clinical case consultation (CCC). Studies suggest that adding CCC posttraining improves treatment fidelity and clinical outcomes [31]. We predicted the following hypotheses based on the findings from our pilot study: (1) a 50% retention rate of clinicians in private practice in the study by the EOT, and (2) a 50% of those who started CCC would be retained by the end of follow-up. Our secondary hypotheses were that improvements in fidelity, self-efficacy, working alliance, and the percentage of patients achieving early weight gain (2.4 kg by session 4—a surrogate marker of end of treatment outcome for adolescents with AN treated with FBT) [32-34] and completing at least 4 CCC sessions (corresponding to session 4 of FBT)—would be greater in those randomized to ET-FBT than in clinicians who received webinar training (WT).

Methods

Study Design

The pilot study described above informed the design of this study as follows: (1) training clinicians in private practice—clinicians who treat a high proportion of youth with eating disorders but were not well represented in the pilot study and often do not receive training in EBTs, (2) refining materials on externalization and agnosticism in 1 training arm (ET-FBT) to confirm whether the preliminary data suggested that fidelity

XSL•FO
RenderX

would improve with additional training in these key components of FBT, (3) incorporating strategies to improve retention by adding continuing medical education (CME) incentives (especially important for those in private practice who do not have institutional support for training), and (4) improving methods for capturing patient weight outcomes using therapist-deidentified reports because obtaining patient consent to collect these data, as attempted in the pilot study, was not feasible (too few patients consented) or effective (insufficient data were collected on outcomes to evaluate the impact of fidelity). This online training study was a longitudinal study with 3 primary assessment time points (baseline, end of online training, and end of CCC). A detailed description of the study protocol has been previously published [35].

Setting

This study was conducted entirely virtually with no physical site locations. Private practice clinicians from across the United States were invited to participate. Recruitment and data collection occurred from November 2020 to October 2024.

Participants

Therapists were required to be licensed (Masters, PhD, PsyD, and MD [psychiatrist]) in the United States and confirm that they regularly treated adolescents with AN in the context of a private practice and that they had not previously been trained in FBT. Methods of recruitment included—social media posts, professional organizations, and listservs (Twitter [subsequently rebranded X], Academy of Eating Disorders, Eating Disorder Referral and Information Center, and licensing organizations). Interested therapists were directed to an online questionnaire for further study information and screened for eligibility. Prior to online training, eligible therapists reported weight change from sessions 1 to 4 of an adolescent with AN they had recently treated using any treatment other than FBT. This weight change would be compared to weight change by session 4 of FBT post–online training during CCC to explore whether training affected early weight gain. Baseline assessments were conducted prior to training allocation and were completed online. After completing assessments, therapists were randomized to either ET-FBT or WT using block randomization. The training period for each group lasted approximately 3 months and was followed by 3 months of weekly CCC.

Assessments

Detailed descriptions of the measures and related procedures used in this study can be found in a previously published protocol paper related to this study's design [35]. Therapists completed the measures detailed below at baseline, at the end of online training, and at the end of CCC. In addition, the consultants who conducted CCC completed the *Supervisor Report of Fidelity to Key Components of FBT* at the end of 12 hours of CCC [29,30]. Consultants were faculty members at one of the participating sites in this study. Both consultants were trained in FBT and certified by the Training Institute for Child and Adolescent Eating Disorders as consultant supervisors. The consultants were masked to the participant training condition.

The list of measures used in this study is as follows:

- Demographic and professional survey: participants answered questions related to demographic characteristics, including professional licensure status, highest degree attainment, years in private practice, and years working in their current position.
- The therapist version of the Parents Versus Anorexia Nervosa Scale [36]: this 7-item self-report measure is adapted from the Parents Versus Anorexia Nervosa Scale [36] to assess therapists' self-efficacy with regard to their knowledge and skills in FBT.
- Evidence-Based Practice Attitude Scale–36 [37]: this 36-item measure is a brief version of the Evidence-Based Practice Attitude Scale–50 that assesses clinicians' attitudes toward evidence-based practice, with demonstrated psychometric validity and reliability [37]. This measure was administered at baseline only.
- Motivation for Training in FBT Questionnaire: this measure was created by the research team to assess therapists' motivations for pursuing FBT training. This questionnaire asks participants to select and rank from the following reasons: belief that FBT is the best approach, customer demand, curiosity about FBT, increased income, and professional development. This measure was administered at baseline only.
- Key Component Knowledge Questionnaire [28]: this 10-item multiple choice measure assesses knowledge about the key components of FBT. This measure was previously used in the pilot online training trial to assess the applied knowledge of the key components of FBT by study participants [28].
- Fidelity to FBT Self-Report to Key Components: this is an 8-item multiple choice measure used to assess self-reported fidelity to key components of FBT delivered in the first 4 sessions of treatment. It has been piloted and found acceptable by FBT therapists.
- Working Alliance Inventory (WAI) [38]: the WAI is a 12-item self-report assessment of the therapeutic working alliance with well-established psychometric properties. Participants were asked to complete the measure as it related to the patient for whom posttraining weight data was provided.
- Patient weight change: weight gain from sessions 1 to 4, as reported by the therapist for the patient treated during CCC (and only that patient, if more than 1 patient was discussed), was used as the patient outcome (early weight gain of 2.4 kg by session 4).

Study Size

The sample size for this study was determined to facilitate our investigation focusing on improvements in therapist outcomes, including self-efficacy, working alliance, and fidelity. These outcomes were not primary in this study, although they were critical for planning and enriching the next step of the confirmatory efficacy trial. We estimated power based on the proposed linear mixed effects (LME) modeling with repeatedly measured outcomes (baseline [pretraining], end-of-training [EOT], and follow-up [end of CCC]). We assumed a modest intraclass correlation of 0.5 for the repeated measures, with 50% attrition by follow-up. We also assumed a medium effect size

XSL•FO
RenderX

(Cohen *d*=0.4-0.5) as a clinically meaningful minimum effect size. Under this scenario, with the expected combined sample of 140 clinicians at pretraining, the estimated power to detect significant improvement in therapist competency from baseline to follow-up is 0.98 - 1.0 (2-tailed; α=.05). In terms of retention, our primary outcome, we set at least 50% retention by EOT and at least 50% retention by follow-up among those who start CCC as a success. Assuming about 65% retention at EOT, using a 1-sample chi-square test, the estimated power to achieve retention significantly better than 50% is 0.95. For our exploratory investigation, including group comparisons (ET-FBT vs WT) in therapist and patient outcomes and detecting baseline moderators of group effects, we planned to monitor both clinical significance (ie, effect sizes) and statistical significance (*P* value), with more emphasis on identifying clinical significance.

Training Descriptions

The 2 online training strategies, as noted in the "Study Size" section, are described more fully in a study by Citron et al [35], but a brief description here provides a summary of the main features of the 2 trainings. The WT arm includes an online webinar lecture series. In this arm, participants watched a weekly video lecture (approximately 1 h) created by authors JL and DLG, based on the *Treatment Manual for Anorexia Nervosa: A Family-Based Approach* [39]. Participants had 1 week to view the lecture, and the materials were not available for review. The lectures discuss the scientific evidence supporting FBT, how therapists set up treatment for FBT, the main interventions used in FBT during each phase, and recorded role-plays illustrating interventions throughout the 3 phases. The Enhanced Training (ET-FBT) arm uses an online training platform that releases 1 of 7 lecture bundles every 12 days. Once released, participants had access to the training material throughout their participation in the study. The ET-FBT lecture bundles consisted of 5 to 8 short didactic videos that discuss the treatment model and provide mock therapy session video clips (modeling FBT with a typical adolescent with AN case), supplementary readings (such as the *Treatment Manual for Anorexia Nervosa: A Family-Based* Approach [39]), clinical vignettes and session roleplays, and training assignments. Participant assignments included questions on the medical management of adolescent patients with AN, reflection on clinical vignettes and the therapist's approach, and identification of treatment moderators. ET-FBT included additional focused training related to the key components of FBT—agnosticism and externalization. Brief online didactic lectures defined the 2 principles and discussed their uses, challenges, and application in clinical practice. Participants were provided with clinical vignettes on the application of agnosticism and externalization. Both trainings lasted approximately 12 total hours.

Ethical Considerations

Eligible therapists were consented using a protocol approved by the institutional review board at Stanford University and the University of California, San Francisco (e-protocol 56548). All participants gave their informed consent for inclusion before participating in the study. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the 1975 Declaration of Helsinki, as revised in 2008. Therapists did not pay for training, but to enhance retention, they received CME credit for completing the training and CCC. Participants were compensated US $100 for each assessment time point (US $300 total). All reports are private and confidential per the institutional review board–approved protocols.

Data Analytic Plan

To estimate improvement in clinician outcomes, we used linear mixed effects (LME) modeling [40,41] with random intercepts and maximum likelihood estimation, which enabled us to fully use outcome data repeatedly measured for most clinicians (3 assessment time points). Given the 2 distinct phases in our study design (baseline to EOT and EOT or initiation of CCC to completion of CCC), we used piecewise growth models that allowed different slopes from baseline to EOT and from EOT to follow-up. We used the training group (ET-FBT vs WT) as the predictor of changes across the 2 phases. The piecewise LME model we used here is presented below.

We describe our general model for longitudinal LME modeling in 2 levels. The outcome *Y* for therapist *i* at time point *t* can be expressed as follows:

(1)Yit=η0i+η1iS1t+η2iS2t+εi,

which is a function of 3 growth parameters:

(2)η0i=α0+γ0Zi+ζ0i,

(3)η1i=α1+γ1Zi,

(4)η2i=α2+γ2Zi,

where i=1, 2, . . . , N (N=123 in our study) and t=0, 1, 2 (baseline, EOT, and follow-up). Therapists were randomly assigned to 1 of the 2 conditions (Z_i=1 if ET-FBT, Z_i=0 if WT). We modeled the longitudinal trend using 3 growth parameters, the initial status (η_{0i}) and the changes from the baseline to EOT (η_{1i}) and from EOT to follow-up (η_{2i}). We used time scores of S_{1t} (=0, 1, 1) for the first and S_{2t} (=0, 0, 1) for the second slope. The fixed effects of interest are the mean initial status (α_0 under WT, $\alpha_0+\gamma_0$ under ET-FBT), the changes in each treatment group (α_1, α_2 under WT, $\alpha_1+\gamma_1$, $\alpha_2+\gamma_2$ under ET-FBT) and the group differences in these changes ($\gamma_1$and γ_2). We allow for individual variability at baseline (ie, random intercept), which is captured by the variance of level 2 residual (Σ_ζ=Var(ζ_{0i})). We assumed the random effect to be normally distributed (ζ_{0i}~N(0,Σ_ζ)). We used the standard assumption for the level 1 residual, which is normally distributed and consistent across assessments (ε_i~N(0,σ^2)).

In our LME modeling, data points that were missing due to clinician attrition or intermittent dropout were handled under the assumption that data are missing at random [42] conditional on observed information, which is a standard assumption in modern missing data approaches. In this procedure, all available cases, including those with missing information, were included in the analysis. For maximum likelihood estimation of mixed effect models, we used the *lme4* package [42] in R software (R version 4.4.1; R Foundation for Statistical Computing). For the

XSL•FO
RenderX

exploratory patient outcome, early weight gain by session 4 (restoring 2.4 kg or not), we planned to report simple proportions at baseline and follow-up as the information is not collected repeatedly from the same individuals. We also explored the number of years working in private practice and therapists' age at baseline as possible treatment effect moderators and/or predictors of therapist outcomes and patient early weight response.

Results

Demographics of Clinicians Trained

Participants were approximately 10% Hispanic, 9% non-White, and 96% female, with a mean age of 36.9 (SD 8.8) years (Table 1). Most were master's level clinicians (77%), followed by psychologists (23%) and psychiatrists (1%). Recruitment and data collection occurred from November 2020 to October 2024.

Table . Participant characteristics (N=123).

Characteristics	ET-FBT[a] (n=61)	WT[b] (n=62)	Total (N=123)
Ethnicity, n (%)			
Hispanic	7 (11.5)	5 (8.1)	12 (9.8)
Non-Hispanic	54 (88.5)	57 (91.9)	111 (90.2)
Sex, n (%)			
Male	3 (4.9)	1 (1.6)	4 (3.3)
Female	58 (95.1)	61 (98.4)	119 (96.7)
Race, n (%)			
American Indian or Alaskan Native	0 (0)	0 (0)	0 (0)
Asian	1 (1.6)	1 (1.6)	2 (1.6)
Black or African American	0 (0)	1 (1.6)	1 (0.8)
Native Hawaiian or other Pacific Islander	0 (0)	0 (0)	0 (0)
White	55 (90.2)	57 (91.9)	112 (91.1)
More than 1 race	5 (8.2)	3 (4.8)	8 (6.5)
Age (y), mean (SD)	36.4 (7.2)	37.5 (10.1)	36.9 (8.8)
Licensure, n (%)			
Master's	48 (78.7)	47 (75.8)	95 (77.2)
Psychiatrist or MD	0 (0)	1 (1.6)	1 (0.8)
Psychologist (PhD, PsyD)	12 (19.7)	14 (22.6)	26 (21.1)
Education specialist	1 (1.6)	0 (0)	1 (0.8)
Years of private practice experience, mean (SD)	4.1 (4.2)	4.6 (6.0)	4.3 (5.2)

[a]ET-FBT: enhanced training family-based treatment.

[b]WT: webinar training.

Recruitment, Feasibility, and Acceptability

We randomized 123 patients (88% of the expected number; Figure 1) over the course of 3 years. Participants were recruited from 33 states, with the largest percentage practicing in California (n=26, 21%), Virginia (n=11, 9%), New York (n=8, 6.5%), North Carolina (n=8, 6.5%), Michigan (n=8, 6.5%), Texas (n=7, 5.7%), and Florida (n=7, 5.7%). A total of 20 states (Alabama, Alaska, Arkansas, Colorado, Connecticut, Idaho, Indiana, Iowa, Maryland, Massachusetts, Minnesota, Missouri, Montana, Nevada, Ohio, Oregon, Rhode Island, South Carolina, Washington, and Wisconsin) had 3 or fewer participants. A total of 123 clinicians started training, and 117 (95%) completed it. Of those who completed online training, 89 (72%) planned to start CCC, but only 54 (44%) found an eligible patient and started CCC during the allotted timeline for study completion.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram. AN: anorexia nervosa; EOT: end-of-training; ET-FBT: enhanced training family-based treatment; FBT: family-based treatment; WT: webinar training.

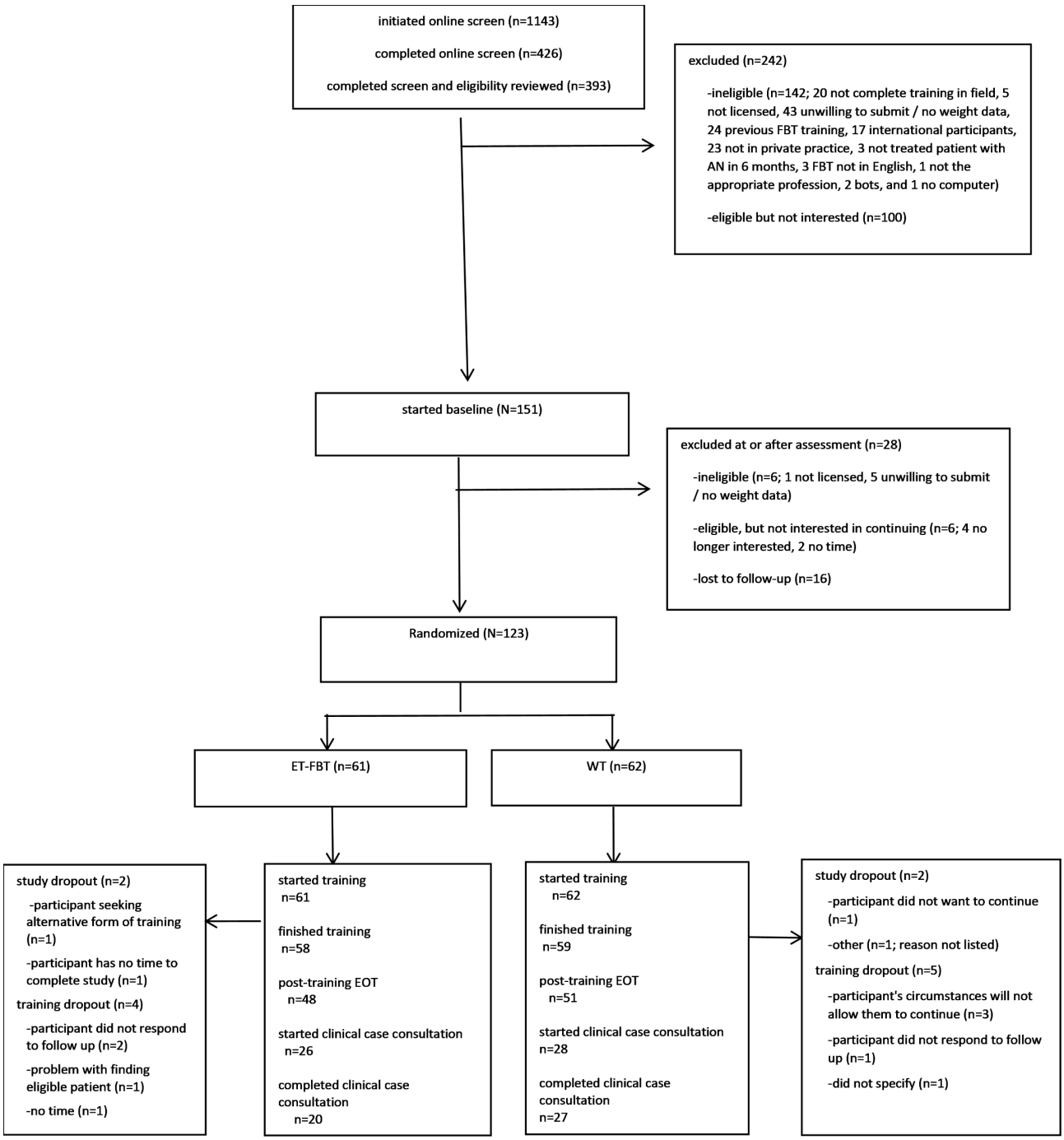

Training Outcomes

There was evidence of considerable improvement in the training outcome measures within both treatment arms (Table 2). In TVAN and fidelity self-ratings, both arms showed similarly large effect sizes (Cohen *d* of 1.0-1.9) in terms of changes from baseline to EOT and follow-up (end of CCC), leading to insignificant group differences. In WAI, both groups showed similar modest but significant improvements from baseline to EOT. However, by follow-up, the ET-FBT condition showed significantly more improvement (Cohen d=0.53; P=.04) in working alliance (WAI). Expert fidelity ratings at the end of CCC, provided by the consultants who were blind to the training condition, did not suggest a difference between the groups (WT=21.3 (SD 4.21) and ET-FBT=19.3 (SD 6.93); Cohen d=−0.35, 95% CI −0.94 to 0.24; P=.25). These expert ratings of fidelity were somewhat lower than self-reported fidelity but were significantly correlated with each other (r=0.33; P=.03).

XSL•FO
RenderX

Table . Estimated changes and intention to treat effects based on longitudinal mixed effects.

	Within group change										Group difference in change (ET-FBT vs WT)		
	ET-FBT[a]					WT[b]							
	Mean (95% CI)		Change (95% CI)	ES[c]	*P* value	Mean (95% CI)		Change (95% CI)	ES	*P* value	Difference in change (95% CI)	ES	*P* value
TVAN[d]													
Baseline to EOT[e]	Baseline: 23.21 (22.42 to 24.01)	EOT: 28.26 (27.39 to 29.13)	5.05 (4.19 to 5.91)	1.49	<.001	Baseline: 23.42 (22.63 to 24.21)	EOT: 28.01 (27.17 to 28.85)	4.59 (3.76 to 5.42)	1.36	<.001	0.46 (−0.74 to 1.66)	0.14	.45
Baseline to follow-up	Baseline: 23.21 (22.42 to 24.01)	Follow-up: 29.68 (28.51 to 30.86)	6.47 (5.30 to 7.64)	1.90	<.001	Baseline: 23.42 (22.63 to 24.21)	Follow-up: 29.33 (28.24 to 30.43)	5.92 (4.83 to 7.00)	1.75	<.001	0.55 (−1.04 to 2.15)	0.16	.50
EOT to follow-up	EOT: 28.26 (27.39 to 29.13)	Follow-up: 29.68 (28.51 to 30.86)	1.42 (0.23 to 2.60)	0.42	.02	EOT: 28.01 (27.17 to 28.85)	Follow-up: 29.33 (28.24 to 30.43)	1.32 (0.23 to 2.42)	0.39	.02	0.09 (−1.52 to 1.71)	0.03	.91
WAI[f]													
Baseline to EOT	Baseline: 58.81 (56.47 to 61.14)	EOT: 63.59 (61 to 66.18)	4.78 (2.00 to 7.56)	0.44	.001	Baseline: 56.9 (54.59 to 59.22)	EOT: 60.12 (57.62 to 62.62)	3.22 (0.52 to 5.91)	0.36	.02	1.56 (−2.31 to 5.44)	0.16	.43
Baseline to follow-up	Baseline: 58.81 (56.47 to 61.14)	Follow-up: 67.18 (63.57 to 70.78)	8.37 (4.62 to 12.11)	0.76	<.001	Baseline: 56.9 (54.59 to 59.22)	Follow-up: 59.95 (56.61 to 63.29)	3.05 (−0.44 to 6.54)	0.34	.09	5.32 (0.19 to 10.44)	0.53	.04
EOT to follow-up	EOT: 63.59 (61.00 to 66.18)	Follow-up: 67.18 (63.57 to 70.78)	3.59 (−0.23 to 7.40)	0.33	.07	EOT: 60.12 (57.62 to 62.62)	Follow-up: 59.95 (56.61 to 63.29)	−0.17 (−3.71 to 3.38)	−0.02	.93	3.75 (−1.46 to 8.96)	0.37	.16
Fidelity self-rating													
Baseline to EOT	Baseline: 10.67 (9.1 to 12.25)	EOT: 17.45 (15.72 to 19.17)	6.77 (5.06 to 8.48)	1.03	<.001	Baseline: 10.39 (8.83 to 11.95)	EOT: 16.21 (14.53 to 17.89)	5.82 (4.16 to 7.49)	0.96	<.001	0.95 (−1.44 to 3.34)	0.15	.47
Baseline to follow-up	Baseline: 10.67 (9.1 to 12.25)	Follow-up: 21.02 (18.69 to 23.35)	10.35 (8.03 to 12.66)	1.58	<.001	Baseline: 10.39 (8.83 to 11.95)	Follow-up: 21.42 (19.26 to 23.59)	11.03 (8.88 to 13.19)	1.82	<.001	−0.69 (−3.85 to 2.48)	−0.11	.67

XSL•FO
RenderX

	Within group change										Group difference in change (ET-FBT vs WT)		
	ET-FBT[a]					WT[b]							
	Mean (95% CI)		Change (95% CI)	ES[c]	*P* value	Mean (95% CI)		Change (95% CI)	ES	*P* value	Difference in change (95% CI)	ES	*P* value
EOT to follow-up	EOT: 17.45 (15.72 to 19.17)	Follow-up: 21.02 (18.69 to 23.35)	3.58 (1.23 to 5.93)	0.55	.003	EOT: 16.21 (14.53 to 17.89)	Follow-up: 21.42 (19.26 to 23.59)	5.21 (3.02 to 7.41)	0.86	<.001	−1.64 (−4.85 to 1.58)	−0.26	.32

[a]ET-FBT: enhanced training family-based treatment.

[b]WT: webinar training.

[c]ES: effect size (in Cohen *d*).

[d]TVAN: Therapist Versus Anorexia Nervosa Scale.

[e]EOT: end-of-training.

[f]WAI: Working Alliance Inventory.

Patient Outcomes

The primary patient-related outcome for this study was early weight gain of 2.4 kg by session 4, used as a proxy predictor (surrogate marker) for end-of-treatment weight restoration [32-34]. The percentage of early weight gain of 2.4 kg by session 4 for the sample at baseline (prior to online training) was 15.4% (19/123). At follow-up, the percentage of patients who achieved early weight gain of 2.4 kg by session 4 more than doubled in both training approaches to a rate of 34% (33% under ET-FBT; 35% under WT) compared to the rate at baseline. This rate of early response is almost identical to that achieved by trained and supervised FBT therapists participating in RCTs [43].

Potential Moderators and Predictors of Therapist and Patient Outcomes

We examined the number of years working in private practice and therapists' age at baseline as possible treatment effect moderators and/or predictors of self-reported fidelity. The number of years working in private practice did not show a significant association with self-reported fidelity either at EOT or follow-up. Therapist age showed little or negative association with self-reported fidelity in the WT arm (EOT: r=0.13, P=.36; follow-up: r=–0.39, P=.048), whereas it showed a positive association with fidelity in the ET-FBT arm (EOT: r=0.32, P=.03; follow-up: r=0.26, P=.25). As a result, at follow-up, the moderating effect of age was significant (r difference=0.65; P=.04), which can be interpreted as older therapists benefiting more from the ET-FBT condition. We did not find any significant associations between baseline therapist characteristics and patient early weight response.

Discussion

Principal Findings

The results of this study found that both online training strategies were acceptable and feasible to conduct with clinicians in private practice. They led to improvements in self and expert reported fidelity to the key components, self-efficacy related to the delivery of FBT, and working alliance. Furthermore, both training strategies also appeared to lead to improved patient outcomes, more than doubling the early response rates compared to the rates of those treated prior to FBT training. This is a rate that approximates what is achieved in RCTs with expert FBT clinicians [43]. The only difference between the two training approaches was that the WAI showed significantly greater increases in the ET-FBT group, which may be attributable to baseline differences in this measure. Thus, overall, our hypothesis that ET-FBT would outperform WT was not supported. While this result was not as expected, it is important that both training programs led to fidelity to FBT key components and patient clinical improvements. This supports the promise that online training of therapists in FBT is likely a viable way to increase the supply of therapists who can deliver the treatment with reasonable fidelity, with an expectation that their patients will experience substantial benefit. These results are in line with previous nonrandomized studies of online training for clinicians treating eating disorders, where significant improvements in skills and confidence were noted for clinicians learning CBT or interpersonal psychotherapy [8,9]. Taken together, these studies support the conclusion that online training is likely a feasible and practical strategy for training clinicians in the main evidence-based approaches for eating disorders.

The recruitment process was hampered in part by the COVID-19 pandemic, resulting in only 123 participants (88% of our original aim of 140) enrolling. This was because many therapists reported being overwhelmed by referrals and therefore had no time for training [44-46]. Nonetheless, interest in receiving online training in FBT was high, with 1143 expressing interest, 426 completing the online screening, 251 being deemed eligible after screening, 123 starting the training, and 117 completing it (95%), which was much higher than the 50% we had anticipated based on our pilot studies. However, in this study, of those who completed the online training, 89 (72%) expressed interest in receiving CCC, but only 54 (44%) of these clinicians were able to find an eligible patient and start CCC within the

XSL•FO
RenderX

allotted timeline for study completion. This was lower than the 50% we had anticipated. At the same time, the majority of those who found an eligible patient and started CCC completed the 12 hours offered (n=47, 87%) and provided patient outcome data (ie, patient weight change from sessions 1 - 4). This level of obtaining patient outcome data was much greater than the 10% we were able to collect in the pilot study, where obtaining patient consent was a major barrier [28]. Overall, the acceptability of study procedures was high, with rates of completion of study measures at the EOT reaching 79% (48/61) for ET-FBT and 82% (51/62) for WT, and with an overall rate of assessment completion of 80% (99/123), with no statistical differences between randomized groups.

As noted above, it was somewhat surprising and contrary to our hypothesis that additional training in agnosticism and externalization did not improve fidelity. We speculate that, while the didactic content in the ET-FBT group may have improved intellectual understanding of these interventions, participating in CCC provided additional learned application of these skills in a clinical case, which leveled up fidelity in both groups. Alternatively, it may be that the preliminary studies overestimated the benefits of these additional training modules and that standard training is already sufficient for clinicians' understanding and use of agnosticism and externalization.

While the EOT results were comparable between groups, the predicted attrition of 50% posttraining participation in CCC was evident in this trial. However, a large proportion of those who did not complete CCC were unable to start (38%) because they could not find an eligible patient in time. As suggested above, this may be a result of the fact that much of the study took place during COVID-19, when there was a significant uptick in eating disorders, and therapists treating eating disorders were overwhelmed with referrals and did not have the time or energy for training. However, a more important factor might have been the study eligibility requirement that patients treated in the study be underweight at the start of FBT, which precluded the enrollment of adolescents who were partially weight-restored in other settings or who had atypical AN without low weight. Many clinicians in private practice reported seeing mostly patients who were weight-restored in other settings (hospitals, day programs, and residential programs). Of those who did complete CCC, there was an early weight response by session 4 of FBT after only 4 hours of CCC. At the same time, those who completed all 12 hours of CCC saw continued improvement in fidelity and self-efficacy (TVAN) in both training arms, as well as WAI in the ET-FBT arm (Table 2).

Because early weight response is a strong predictor of end-of-treatment weight restoration, these data provide support that training plus CCC led to clinical implementation effects on patient outcomes similar to those expected in RCTs. The early weight response did not differ by treatment group (odds ratio 0.99, 95% CI 0.67-1.47). However, the study design did not include the collection of patient outcome data beyond session 4 of FBT. This was because weight gain at this early time point is a reliable predictor of end-of-treatment remission, with a predictive value of 70% to 80% in RCTs [32], and also because the study's funding duration was limited to a 3-year window. However, by not including non-low weight adolescents with AN in this sample, we cannot know for certain whether session 4 weight gain will be a predictor of end-of-treatment response for this clinical group, and future studies that include all adolescents with AN, regardless of starting weight, are needed to clarify this.

This study was designed using a practical approach to real-world challenges related to future training and implementation procedures, as well as clinical outcome measurement. As a result, we focused on clinician-reported data as a feasible approach to diagnostic assessment, fidelity, and patient weight outcomes. Therefore, the clinician-participants identified eligible patients in their practice (ie, diagnosed with AN both prior to and during the training period). The clinician participants also reported weight at baseline and session 4 for the identified patients they treated or would treat during the training period. These design features allowed for real-world evaluation of training and patient outcomes but also resulted in limitations regarding related findings resulting from self-report. It is possible that some clinicians inaccurately diagnosed or reported inaccurate weight change data. However, there was no evidence or any clear motivation for false reporting, and there were no differences in the changes between the 2 online training platforms, suggesting that if these inaccuracies were present, they were equally distributed between the randomized groups.

This study used a standard dose of CCC post training because studies suggest that CCC post training is a necessary aspect of mastering the skills described in a didactic training [31,47,48]. However, CCC is a challenge to scale because it is costly, and there are limited experts to provide it. A next step would be to systematically examine cost-effective strategies for CCC, including determining more specifically how much CCC is needed to ensure fidelity and clinical outcomes, as well as strategies for more efficient procedures of CCC, including potentially group approaches. This is an important area for future investigations. Furthermore, it is unknown if the additional 8 hours of CCC beyond the first 4 had any effect on patient outcomes. CCC post online training is likely necessary to implement EBT effectively [31], so future studies should identify how many hours of CCC are needed in addition to the online training to develop an evidence-based and scalable FBT training model combining these two essential elements.

Another limitation of this study design is the exclusive focus on private practice and majority White clinicians, thereby limiting generalization to clinicians of other ethnicities or races and those who practice in other settings. However, there is little reason to expect that clinicians in other clinical settings or even those outside the United States would not respond similarly, especially given our previous pilot study [28]. We used training incentives, specifically CME, to encourage clinicians in private practice to attend training. We did this because clinicians in private practice need CMEs for continued licensure and to help mitigate the loss of income related to training. Clinicians in public practice settings or those with organizational support may not need this type of incentive. Thus, for clinicians in these other settings, costs to provide training would likely be lower without the costs of CME included in the training.

XSL•FO
RenderX

By design, most of our outcome measures are self-reported, and additionally, some do not have psychometric studies documenting their reliability. However, this study used measures that were short and scalable so that they had viability in a real-world clinical context where the training of therapists and evaluation of fidelity must be efficient and practical. Thus, while the measures have only face validity, they appear to capture self-reported change on key variables related to fidelity and self-efficacy in the delivery of FBT. Importantly, the key measure of therapist self-reported fidelity was correlated with expert consultant fidelity ratings, suggesting that self-report may be a reasonable and practical strategy to assess fidelity in a clinical setting, but this requires additional study. In addition, future studies should determine the psychometric properties of these fidelity measures with greater precision.

Finally, although patients treated during CCC more than doubled the rates of early response (15% vs 34%) from pretraining to posttraining, supporting an important patient outcome change effect, there was no comparison group that did not receive CCC posttraining; therefore, it is not possible to determine if and how much additional CCC is needed to improve early weight gain.

Conclusions

This study demonstrates that it is feasible, acceptable, and practical to train therapists in private practice in FBT-AN using 2 different online strategies, with little evidence that one has advantages over the other. It is important to note that the main reason participants were excluded from training was that they were not in private practice or could not find an eligible patient with baseline weight data, rather than due to a lack of interest, as 59% of eligible participants enrolled and most of those (95%) completed online training. In addition, the outcomes of the study suggest that there were improvements in all measured training and fidelity metrics, as well as patient early response rates. These data, taken together with our previous studies, suggest that asynchronous online FBT training could be offered to clinicians in private practice at low cost using either of the approaches tested in this study. These training materials can be distributed easily through learning platforms and translated into non-English versions using relatively inexpensive artificial intelligence translation technologies. However, several key questions remain: What are the benefits of adding CCC to online training, and how much CCC is needed to achieve fidelity to the key components of FBT and to achieve expected patient outcomes? Challenges in the uptake of CCC were evident in this trial, and additional strategies to increase uptake and completion of CCC are needed. Because CCC is potentially costly and the availability of expert consultants is limited, it is important to clarify for individual providers and health care systems an expected effective strategy. Making CCC as short as feasible and associated with a predictable outcome would likely encourage more therapists to take up and complete CCC.

This study portends a new era for FBT training availability, which should, in turn, lead to an increase in trained and competent providers, as well as improved access for young persons with AN and their families to an important evidence-based intervention for this potentially lethal disorder.

Acknowledgments

The authors are grateful to many individuals for supporting this work, including the family-based treatment expert consultants (Erin Accurso, PhD, and Sarah Forsberg, PsyD) and the members of the data safety and monitoring board (Jennifer Couturier, MD, and Aaron Hogue, PhD). The authors would also like to thank the participants involved in the study who provided their data.

Funding

This work was supported by the US National Institute of Mental Health (R34MH123596).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: JL, DLG, CB, BJ
Data curation: BM, BK, CB
Formal analysis: BM, BK, BJ
Funding acquisition: JL, DLG, CB, BJ
Investigation: JL, DLG, BM, CB
Methodology: JL, DLG, BM, CB, BJ
Project administration: JL, DLG
Resources: JL, DLG
Supervision: JL, DLG
Visualization: BM, BK
Writing – original draft: JL, BM
Writing – review & editing: JL, DLG, BM, BK, CB, BJ
All authors read and approved the final manuscript.

XSL•FO
RenderX

Conflicts of Interest

JL receives royalties from Guilford Press and Routledge and is co-director of the Training Institute for Child and Adolescent Eating Disorders, LLC. DLG receives royalties from Guilford Press and Routledge, is co-director of the Training Institute for Child and Adolescent Eating Disorders, LLC, and is a member of the Equip Health Clinical Advisory Board. BM receives royalties from Routledge.

Checklist 1
CONSORT checklist.
[PDF File, 11363 KB - jmir_v28i1e89999_app1.pdf]

References

1. Kazdin AE, Fitzsimmons-Craft EE, Wilfley DE. Addressing critical gaps in the treatment of eating disorders. Int J Eat Disord 2017 Mar;50(3):170-189. [doi: 10.1002/eat.22670] [Medline: 28102908]
2. McHugh RK, Barlow DH. The dissemination and implementation of evidence-based psychological treatments. A review of current efforts. Am Psychol 2010;65(2):73-84. [doi: 10.1037/a0018121] [Medline: 20141263]
3. Fairburn CG, Cooper Z. Therapist competence, therapy quality, and therapist training. Behav Res Ther 2011 Jun;49(6-7):373-378. [doi: 10.1016/j.brat.2011.03.005] [Medline: 21492829]
4. Williams NJ, Beidas RS. Annual Research Review: the state of implementation science in child psychology and psychiatry: a review and suggestions to advance the field. J Child Psychol Psychiatry 2019 Apr;60(4):430-450. [doi: 10.1111/jcpp.12960] [Medline: 30144077]
5. Kazdin AE. Addressing the treatment gap: a key challenge for extending evidence-based psychosocial interventions. Behav Res Ther 2017 Jan;88:7-18. [doi: 10.1016/j.brat.2016.06.004] [Medline: 28110678]
6. McMillen JC, Hawley KM, Proctor EK. Mental health clinicians' participation in web-based training for an evidence supported intervention: signs of encouragement and trouble ahead. Adm Policy Ment Health 2016 Jul;43(4):592-603. [doi: 10.1007/s10488-015-0645-x] [Medline: 25822326]
7. Jackson CB, Quetsch LB, Brabson LA, Herschell AD. Web-based training methods for behavioral health providers: a systematic review. Adm Policy Ment Health 2018 Jul;45(4):587-610. [doi: 10.1007/s10488-018-0847-0] [Medline: 29352459]
8. D'Adamo L, Laboe AA, Goldberg J, et al. Development and usability testing of an online platform for provider training and implementation of cognitive-behavioral therapy guided self-help for eating disorders. BMC Digit Health 2025;3(1):2. [doi: 10.1186/s44247-024-00140-6] [Medline: 40893384]
9. Karam Jones AM, Fitzsimmons-Craft EE, D'Adamo L, et al. A pilot study evaluating online training for therapist delivery of interpersonal psychotherapy for eating disorders. Int J Eat Disord 2024 Aug;57(8):1691-1706. [doi: 10.1002/eat.24197] [Medline: 38623931]
10. Brownlow RS, Maguire S, O'Dell A, Dias-da-Costa C, Touyz S, Russell J. Evaluation of an online training program in eating disorders for health professionals in Australia. J Eat Disord 2015;3(1):37. [doi: 10.1186/s40337-015-0078-7] [Medline: 26550477]
11. Hoek HW. Review of the worldwide epidemiology of eating disorders. Curr Opin Psychiatry 2016 Nov;29(6):336-339. [doi: 10.1097/YCO.0000000000000282] [Medline: 27608181]
12. Treasure J, Russell G. The case for early intervention in anorexia nervosa: theoretical exploration of maintaining factors. Br J Psychiatry 2011 Jul;199(1):5-7. [doi: 10.1192/bjp.bp.110.087585] [Medline: 21719874]
13. Le Grange D, Accurso EC, Lock J, Agras S, Bryson SW. Early weight gain predicts outcome in two treatments for adolescent anorexia nervosa. Int J Eat Disord 2014 Mar;47(2):124-129. [doi: 10.1002/eat.22221] [Medline: 24190844]
14. Touyz S, Le Grange D, Lacey H, et al. Treating severe and enduring anorexia nervosa: a randomized controlled trial. Psychol Med 2013 Dec;43(12):2501-2511. [doi: 10.1017/S0033291713000949] [Medline: 23642330]
15. Arcelus J, Mitchell AJ, Wales J, Nielsen S. Mortality rates in patients with anorexia nervosa and other eating disorders. A meta-analysis of 36 studies. Arch Gen Psychiatry 2011 Jul;68(7):724-731. [doi: 10.1001/archgenpsychiatry.2011.74] [Medline: 21727255]
16. Franko DL, Keshaviah A, Eddy KT, et al. A longitudinal investigation of mortality in anorexia nervosa and bulimia nervosa. Am J Psychiatry 2013 Aug;170(8):917-925. [doi: 10.1176/appi.ajp.2013.12070868] [Medline: 23771148]
17. Chesney E, Goodwin GM, Fazel S. Risks of all-cause and suicide mortality in mental disorders: a meta-review. World Psychiatry 2014 Jun;13(2):153-160. [doi: 10.1002/wps.20128] [Medline: 24890068]
18. Hilbert A, Hoek HW, Schmidt R. Evidence-based clinical guidelines for eating disorders: international comparison. Curr Opin Psychiatry 2017 Nov;30(6):423-437. [doi: 10.1097/YCO.0000000000000360] [Medline: 28777107]
19. Watson HJ, Bulik CM. Update on the treatment of anorexia nervosa: review of clinical trials, practice guidelines and emerging interventions. Psychol Med 2013 Dec;43(12):2477-2500. [doi: 10.1017/S0033291712002620] [Medline: 23217606]
20. Lock J, La Via MC, American Academy of Child and Adolescent Psychiatry (AACAP) Committee on Quality Issues (CQI). Practice parameter for the assessment and treatment of children and adolescents with eating disorders. J Am Acad Child Adolesc Psychiatry 2015 May;54(5):412-425. [doi: 10.1016/j.jaac.2015.01.018] [Medline: 25901778]

XSL•FO
RenderX

21. Striegel-Moore RH, DeBar L, Wilson GT, et al. Health services use in eating disorders. Psychol Med 2008 Oct;38(10):1465-1474. [doi: 10.1017/S0033291707001833] [Medline: 17976250]
22. Cook JM, Biyanova T, Coyne JC. Barriers to adoption of new treatments: an internet study of practicing community psychotherapists. Adm Policy Ment Health 2009 Mar;36(2):83-90. [doi: 10.1007/s10488-008-0198-3] [Medline: 19104928]
23. Couturier J, Kimber M, Jack S, Niccols A, Van Blyderveen S, McVey G. Understanding the uptake of family-based treatment for adolescents with anorexia nervosa: therapist perspectives. Int J Eat Disord 2013 Mar;46(2):177-188. [doi: 10.1002/eat.22049] [Medline: 22911878]
24. Couturier J, Kimber M, Barwick M, et al. Themes arising during implementation consultation with teams applying family-based treatment: a qualitative study. J Eat Disord 2018;6(1):32. [doi: 10.1186/s40337-018-0218-y] [Medline: 30410759]
25. Dimitropoulos G, Kimber M, Singh M, et al. Stay the course: Practitioner reflections on implementing family-based treatment with adolescents with atypical anorexia. J Eat Disord 2019;7(1):10. [doi: 10.1186/s40337-019-0240-8] [Medline: 31049201]
26. Astrachan-Fletcher E, Accurso EC, Rossman S, McClanahan SF, Dimitropoulos G, Le Grange D. An exploratory study of challenges and successes in implementing adapted family-based treatment in a community setting. J Eat Disord 2018;6(1):44. [doi: 10.1186/s40337-018-0228-9] [Medline: 30603086]
27. Cook SC, Schwartz AC, Kaslow NJ. Evidence-based psychotherapy: advantages and challenges. Neurotherapeutics 2017 Jul;14(3):537-545. [doi: 10.1007/s13311-017-0549-4] [Medline: 28653278]
28. Lock J, Le Grange D, Accurso EC, Welch H, Mondal S, Agras WS. Is online training in family-based treatment for anorexia nervosa feasible and can it improve fidelity to key components affecting outcome? J Behav Cogn Ther 2020 Jun;30(2):75-82. [doi: 10.1016/j.jbct.2020.03.002]
29. Forsberg S, Lock J, Le Grange D. Family Based Treatment for Restrictive Eating Disorders: A Guide for Supervision and Advanced Clinical Practice: Routledge; 2018.
30. Forsberg S, Fitzpatrick KK, Darcy A, et al. Development and evaluation of a treatment fidelity instrument for family-based treatment of adolescent anorexia nervosa. Int J Eat Disord 2015 Jan;48(1):91-99. [doi: 10.1002/eat.22337] [Medline: 25142619]
31. Ragnarsson EH, Reinebo G, Ingvarsson S, et al. Effects of training in cognitive behavioural therapy and motivational interviewing on mental health practitioner behaviour: a systematic review and meta-analysis. Clin Psychol Psychother 2024;31(3):e3003. [doi: 10.1002/cpp.3003] [Medline: 38855846]
32. Doyle PM, Le Grange D, Loeb K, Doyle AC, Crosby RD. Early response to family-based treatment for adolescent anorexia nervosa. Int J Eat Disord 2010 Nov 1;43(7):659-662. [doi: 10.1002/eat.20764] [Medline: 19816862]
33. Agras WS, Lock J, Brandt H, et al. Comparison of 2 family therapies for adolescent anorexia nervosa: a randomized parallel trial. JAMA Psychiatry 2014 Nov;71(11):1279-1286. [doi: 10.1001/jamapsychiatry.2014.1025] [Medline: 25250660]
34. Hughes EK, Sawyer SM, Accurso EC, Singh S, Le Grange D. Predictors of early response in conjoint and separated models of family-based treatment for adolescent anorexia nervosa. Eur Eat Disord Rev 2019 May;27(3):283-294. [doi: 10.1002/erv.2668] [Medline: 30761665]
35. Citron K, Johnson M, Matheson BE, et al. Study protocol for training providers in private practice in family-based treatment for adolescents with anorexia nervosa: a randomized controlled feasibility trial. Contemp Clin Trials 2022 Sep;120:106889. [doi: 10.1016/j.cct.2022.106889] [Medline: 35998767]
36. Rhodes P, Baillie A, Brown J, Madden S. Parental efficacy in the family-based treatment of anorexia: preliminary development of the Parents Versus Anorexia Scale (PVA). Eur Eat Disorders Rev 2005 Nov;13(6):399-405. [doi: 10.1002/erv.661]
37. Rye M, Torres EM, Friborg O, Skre I, Aarons GA. The Evidence-based Practice Attitude Scale-36 (EBPAS-36): a brief and pragmatic measure of attitudes to evidence-based practice validated in US and Norwegian samples. Implement Sci 2017 Apr 4;12(1):44. [doi: 10.1186/s13012-017-0573-0] [Medline: 28372587]
38. Tracey TJ, Kokotovic AM. Factor structure of the Working Alliance Inventory. Psychol Assess 1989;1(3):207-210. [doi: 10.1037/1040-3590.1.3.207]
39. Lock J, Le Grange D. Treatment Manual for Anorexia Nervosa: A Family-Based Approach, 2nd edition: The Guilford Press; 2013.
40. Singer JD, Willett JB. Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence: Oxford University Press; 2003. [doi: 10.1093/acprof:oso/9780195152968.001.0001]
41. Raudenbush SW, Bryk AS. Hierarchical Linear Models: Applications and Data Analysis Methods: Sage Publications; 2002.
42. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw 2015;67(1):1-48. [doi: 10.18637/jss.v067.i01]
43. Lock JD, Le Grange D, Bohon C, Matheson B, Jo B. Who responds to an adaptive intervention for adolescents with anorexia nervosa being treated with family-based treatment? Outcomes from a randomized clinical trial. J Am Acad Child Adolesc Psychiatry 2024 Jun;63(6):605-614. [doi: 10.1016/j.jaac.2023.10.012] [Medline: 38142046]
44. Taquet M, Geddes JR, Luciano S, Harrison PJ. Incidence and outcomes of eating disorders during the COVID-19 pandemic. Br J Psychiatry 2021 Jul 27;220(5):1-3. [doi: 10.1192/bjp.2021.105] [Medline: 35048812]
45. J Devoe D, Han A, Anderson A, et al. The impact of the COVID-19 pandemic on eating disorders: a systematic review. Int J Eat Disord 2023 Jan;56(1):5-25. [doi: 10.1002/eat.23704] [Medline: 35384016]

XSL•FO
RenderX

46. Madigan S, Vaillancourt T, Dimitropoulos G, et al. A systematic review and meta-analysis: child and adolescent healthcare utilization for eating disorders during the COVID-19 pandemic. J Am Acad Child Adolesc Psychiatry 2025 Feb;64(2):158-171. [doi: 10.1016/j.jaac.2024.02.009] [Medline: 38431196]
47. Alfonsson S, Parling T, Spännargård Å, Andersson G, Lundgren T. The effects of clinical supervision on supervisees and patients in cognitive behavioral therapy: a systematic review. Cogn Behav Ther 2018 May;47(3):206-228. [doi: 10.1080/16506073.2017.1369559] [Medline: 28929863]
48. Alfonsson S, Lundgren T, Andersson G. Clinical supervision in cognitive behavior therapy improves therapists' competence: a single-case experimental pilot study. Cogn Behav Ther 2020 Sep;49(5):425-438. [doi: 10.1080/16506073.2020.1737571] [Medline: 32213046]

Abbreviations

AN: anorexia nervosa
CCC: clinical case consultation
CME: continuing medical education
EBT: evidence-based treatment
EOT: end of training
ET: enhanced training
FBT: family-based treatment
LME: linear mixed effects
RCT: randomized controlled trial
WAI: working alliance inventory
WT: webinar training

Edited by A Stone; submitted 19.Dec.2025; peer-reviewed by A Goldschmidt, HJ Yang; revised version received 09.Apr.2026; accepted 09.Apr.2026; published 25.May.2026.

Please cite as:
Lock J, Le Grange D, Matheson B, Kim B, Bohon C, Jo B
Training Clinicians in Private Practice in Family-Based Treatment for Anorexia Nervosa: Randomized Controlled Trial Comparing Two Online Approaches
J Med Internet Res 2026;28:e89999
URL: https://www.jmir.org/2026/1/e89999
doi:10.2196/89999

© James Lock, Daniel Le Grange, Brittany Matheson, Bohye Kim, Cara Bohon, Booil Jo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Influencing Factors of Medical Postgraduates' Usage Intention Toward Artificial Intelligence–Generated Content Tools in Academic Research: Qualitative Analysis

Chen Wang[1]; Liu Wang[2]; Xuejiao Zhang[3]; Huiying Qi[1]

[1]Department of Health Informatics and Management, School of Health Humanities, Peking University, 38 Xueyuan Rd, Haidian District, Beijing, China

[2]Department of Language and Culture in Medicine, School of Health Humanities, Peking University, Beijing, China

[3]School of Nursing, Peking University, Beijing, China

Corresponding Author:
Huiying Qi
Department of Health Informatics and Management, School of Health Humanities, Peking University, 38 Xueyuan Rd, Haidian District, Beijing, China

Abstract

Background: The integration of artificial intelligence–generated content (AIGC) tools into academic research offers transformative potential for enhancing productivity and innovation. However, within the highly regulated and ethically sensitive medical context, the use of AIGC is accompanied by significant challenges. Medical postgraduates, as the future vanguard of medical science, play a crucial role in the advancement of digital health, and their intention to use AIGC tools will significantly influence the use of these emerging technologies in medical research. Despite the growing popularity of AIGC tools, there remains a paucity of in-depth understanding of the factors driving or hindering medical postgraduates' intention to use these tools in academic research. A clear comprehension of these influencing factors is essential to foster the responsible, effective, and sustainable integration of AIGC into medical research.

Objective: This study aimed to systematically explore the key factors influencing medical postgraduates' intention to use AIGC tools in academic research, with the goal of informing strategies to promote their ethical use and enhance scholarly research capabilities.

Methods: We used a qualitative research design based on grounded theory. Semistructured interviews were conducted with 30 medical postgraduates across diverse specialties, all of whom had prior research experience and familiarity with AIGC tools. Participants were recruited purposively to ensure diverse perspectives. Data analysis followed a systematic coding process to inductively develop a conceptual model, which was further structured and interpreted through the theoretical lens of the Unified Theory of Acceptance and Use of Technology.

Results: Our analysis identified 7 core factors directly shaping usage intention: performance expectancy, effort expectancy, social influence, facilitating conditions, individual characteristics, task characteristics, and technology characteristics. Further analysis revealed that performance expectancy acted as a mediating variable in the relationships between both task characteristics and technology characteristics and usage intention. Additionally, social influence moderated the relationship between task characteristics and performance expectancy. The research findings underscore that, while AIGC tools are valued for assisting daily research tasks, medical postgraduates' intention to use them in academic research is influenced by technical deficiencies, high cognitive load, and the strict ethical risks and data governance requirements in the medical field.

Conclusions: This study constructs a conceptual model aimed at elucidating the influencing factors of medical graduate students' intention to use AIGC in academic research. Recommendations derived from the findings include (1) fostering artificial intelligence literacy and critical competency among medical postgraduates; (2) optimizing AIGC tools to better address domain-specific needs, accuracy, and security concerns prevalent in health research; and (3) establishing clear academic supervision and ethical governance mechanisms to ensure responsible use. These measures are essential to harness the potential of AIGC while safeguarding the rigor and integrity of medical academic research.

(J Med Internet Res 2026;28:e77928) doi:10.2196/77928

KEYWORDS

artificial intelligence tools; medical students; scientific research; grounded theory; influencing factors

XSL•FO
RenderX

Introduction

Background

As artificial intelligence (AI) technology advances to a new level, artificial intelligence–generated content (AIGC) tools, such as ChatGPT, are emerging. These tools are capable of effectively generating content and comprehending context, offering robust support for users to rapidly and efficiently gain knowledge and enhance the productivity and quality of learning and work. They have also had a significant impact on the development of many industries. Within the realm of academic research, AIGC tools have proven to be instrumental in research design, academic search, data processing, data analysis, and paper refinement, serving as a vital aid for researchers [1]. However, the issues of academic integrity, academic research ethics, and data privacy protection that AIGC tools introduce have also surfaced [2].

Consequently, how to effectively respond to risks and ensure the standardization and sustainability of academic research activities while reasonably using the advantages brought by AIGC has become a crucial area of investigation in the field of AIGC-empowered academic research.

Against this background, scholars have carried out a series of studies on the application of AIGC tools in the field of academic research, including how AIGC empowers various disciplines, how it is applied in different academic research links, and how to deal with the changes brought by AIGC tools. However, during the literature review process, we found that this research has not yet conducted an in-depth analysis of the factors influencing the intention of medical postgraduates to use AIGC tools in academic research. As postgraduates are the future backbone of the research team, cultivating their research abilities is crucial for the development of their respective industries. Exploring the factors influencing their use of AIGC tools in academic research work will help guide them to use AIGC tools reasonably to improve their academic research ability and efficiency. Medical postgraduates are the main force in the realm of medicine, so their attitudes, needs, and behavioral characteristics toward AIGC tools will also affect the sustainable development and promotion of AIGC in the medical field.

Objective

This study will explore the motives, needs, obstacles, and concerns of medical postgraduates when using AIGC tools in the academic research process; analyze how different factors affect medical postgraduates' intention toward AIGC tools in academic research; provide a basis for the cultivation of academic research ability of medical postgraduates; and also help promote the optimized application of AIGC tools in the field of medical education and academic research.

Related Studies

Research on the Applications of AIGC Tools in Academic Research

In the realm of academic research, AIGC tools (such as ChatGPT and DeepSeek) are increasingly gaining widespread attention. These tools, with their ability to simulate human intelligence, can efficiently process and generate a wide range of content, offering new possibilities for scientific practice.

Research has shown that ChatGPT can significantly enhance research efficiency by rapidly searching and organizing a vast amount of literature, which can effectively reduce the time researchers spend on collecting materials [3]. Moreover, AIGC tools can provide researchers with new perspectives and pathways for research due to their powerful capabilities for resource acquisition and analysis [4-6]. In academic writing, AIGC tools also play a significant role in improving the standardization and quality of articles [7,8]. In the field of medical research, AIGC tools have demonstrated potential applications in assisting with medical text writing [9], image analysis [10], and disease diagnosis and treatment [11,12].

Despite the significant advantages that AIGC tools provide in academic research, they also contribute to certain limitations and potential risks. Some studies have pointed out that when conducting literature reviews, AIGC tools may lack autonomy and cannot independently generate new ideas [4]. Besides, AIGC tools may fail to effectively distinguish the reliability of information and potentially introduce inaccurate or misleading data, which could pose a threat to research quality and academic integrity [13,14]. Additionally, AIGC tools face multiple challenges, including data bias, information leakage, and ethical issues. Particularly in data analysis, as researchers' dependence on AI increases, it may further undermine the independence of academic research [15].

In the medical field, academic research involves a large amount of sensitive data, which directly relates to patients' health and safety. Therefore, ensuring privacy protection and ethical considerations in medical research is particularly critical [2,16]. Given the characteristics of medical research, the precision, credibility, and security of AIGC tools must undergo rigorous testing and evaluation. Consequently, exploring how to reasonably and safely apply AIGC tools in medical research to ensure the accuracy, completeness, and academic integrity of studies, while fully safeguarding patient privacy and data security, has become an urgent and important issue that needs to be addressed.

Research on the Factors Influencing the Intention to Use AIGC Tools

The intention to use AIGC is a prerequisite for its widespread application. Scholars have explored the acceptance, adoption, and intention to use AIGC tools through various methods and theoretical models.

Some studies have used the Technology Acceptance Model to reveal the factors influencing different user groups' intention to use AIGC tools. It has been found that perceived risk, perceived usefulness, perceived ease of use, and attitude toward technology play important roles in the adoption of ChatGPT [17]. Other studies have mentioned that intention to use is significantly influenced by personal ability, social influence, perceived usefulness of AI, enjoyment, trust in AI intelligence, positive attitude, and metacognitive self-regulated learning [18]. Studies based on the Unified Theory of Acceptance and Use of Technology (UTAUT) model have further expanded the research

XSL•FO
RenderX

perspective. Relevant studies have shown that performance expectancy, social influence, and effort expectancy significantly affect the intention to use ChatGPT, while facilitating conditions do not have a direct relationship with usage intention [19]. Menon and Shilpa [20] found that the 4 core factors of UTAUT (performance expectancy, effort expectancy, social influence, and facilitating conditions) and 2 extended factors (perceived interactivity and privacy concerns) can effectively explain user interactions with ChatGPT. Tao et al [21], based on the UTAUT model, added 3 factors: perceived risk, resistance bias, and personal innovativeness, and found that these factors are significantly positively correlated with the intention to use AI chatbots.

In addition, research has increasingly focused on the student perspective, analyzing the factors influencing the use of AIGC tools in different educational contexts. Some studies have found that students' perceived value of ChatGPT significantly affects their intention to use it, with perceived usefulness, perceived enjoyment, and perceived cost being important factors influencing perceived value. Moreover, studies have shown that students with higher levels of AI literacy have a higher acceptance of AIGC tools [22]. A study on medical students' intention to use ChatGPT in an interdisciplinary programming course analyzed influencing factors from 4 aspects: individual (such as learning motivation), technology (such as tool reliability), information (such as content accuracy), and environment (such as teaching support) [23].

Despite the fact that existing research has analyzed the intention to use AIGC and its influencing factors from multiple angles, previous research has paid less attention to the factors influencing the use of AIGC by medical postgraduates in academic research. As medical postgraduates are the main force in medical research and practice, their acceptance and intention to use AIGC will directly affect the long-term development of AIGC in the medical field. Therefore, this study takes medical postgraduates as the research object and explores their applications of AIGC in academic research and the influencing factors.

Methods

Study Design

Currently, research on the factors influencing the use of AIGC tools in academic research activities has yet to establish a mature theoretical model, variables, or scales. Therefore, this study uses the classic grounded theory proposed by Glaser et al [24] as the research methodology. Grounded theory, as an inductive qualitative research approach, is dedicated to constructing robust theories from empirical data. By coding the collected raw data and conducting multiple continuous comparisons, this method enables researchers to maintain a high degree of theoretical sensitivity throughout the evolution of concepts and categories, thereby uncovering the factors that influence the use of AIGC tools by medical postgraduates in academic research.

In accordance with the aforementioned research approach, the study design is based on the following five steps:

1. Based on the pre-established theoretical framework, this study has devised a sampling strategy to ensure the inclusion of medical postgraduates with extensive clinical experience and diverse perspectives.
2. Research participants are capable of providing comprehensive information regarding the phenomenon under investigation, which facilitates researchers' in-depth understanding.
3. Data collection and analysis will be conducted concurrently until new information no longer contributes to the expansion of the theory, that is, until data saturation is achieved.
4. As theoretical insights become progressively clearer, the data collection strategy may be correspondingly adjusted.
5. Through continuous comparison, integration, and reflection, the study ultimately constructs a theoretical model grounded in empirical data.

Participant Recruitment

Participants were selected according to the following criteria: (1) postgraduates in medical-related fields with research experience and systematic academic research training; (2) familiarity with and prior use of AIGC tools; and (3) voluntary participation, fluent communication skills, independent thinking ability, and intention to provide detailed information.

Recruitment was conducted through three channels: (1) advertisements posted in WeChat groups and on WeChat Moments; (2) snowball sampling, where initial participants referred other eligible individuals; and (3) outreach via platforms such as Zhihu, Weibo, and Rednote.

The sample size of 30 medical postgraduates was determined based on the principle of "information power," prioritizing depth and diversity over numerical representativeness. To capture a wide range of experiences, the purposive sampling strategy ensured variation in key dimensions such as gender, disciplinary background (eg, internal medicine, surgery, basic sciences), and stage of research training.

Data Collection

Semistructured interviews served as the primary data collection method. The interview outline was developed in three stages: (1) drafting based on prior research and the UTAUT theory; (2) revision through discussions with 2 researchers experienced in AIGC tool application; and (3) pilot testing with 5 users, followed by refinement based on feedback.

The UTAUT was used as the foundational framework for this study. Although UTAUT was originally developed to explain technology acceptance in organizational settings [25], its core constructs, including performance expectancy, effort expectancy, social influence, and facilitating conditions, are broadly applicable to individual-level technology adoption across various domains. In the context of academic research, these constructs align closely with the key drivers of technology use among postgraduate researchers, who must balance efficiency, effort, peer influence, and institutional support when integrating new tools into their workflows.

Each interview lasted 30 to 40 minutes. With participants' consent, all interviews were audio-recorded. Recordings were

XSL•FO
RenderX

then transcribed verbatim to form the raw textual data for analysis. All participants provided informed consent and agreed to the anonymous use of their quotes.

Data Analysis

Prior to the coding of the interview data, a rigorous screening process was conducted based on two main criteria: (1) the clarity of the descriptions and (2) the relevance of the descriptions to the topic. This process was completed by CW and HQ, and they also integrated similar descriptions for subsequent analysis. To ensure comprehensive familiarity and sensitivity to the data, the researchers repeatedly read through the transcripts and descriptions, thereby minimizing the risk of losing key information. This study followed the 3-step coding process (open coding, axial coding, and selective coding) proposed by Venkatesh et al [25] for qualitative analysis and assessed the stability of the coding results through a theoretical saturation test. The specific analysis process is as follows:

1. Open coding: The raw data were initially organized and semantically deconstructed. Basic conceptual units were extracted through line-by-line coding. Researchers employed the constant comparative method to conceptually annotate the original sentences, forming an initial category system through multiple rounds of classification and integration.
2. Axial coding: Based on the primary categories formed from open coding, category relational analysis was used to identify potential logical relationships. Through dimensional comparison and attribute induction, a hierarchical structure among categories was established, ultimately constructing a main category framework with explanatory logic.
3. Selective coding: Theoretical integration was conducted around the core category, systematically sorting out the logical relationships between the main categories and the core category. A theoretically consistent framework was formed through storyline analysis.
4. Theoretical saturation verification: Saturation was confirmed through an iterative process of data collection and analysis. After initial coding of the majority of interviews, a sample addition method was employed, where approximately 10% of new transcripts were sequentially analyzed. Theoretical saturation was deemed reached when these new samples no longer yielded novel categories, properties, or significant relationships but merely reinforced the existing coding structure. The entire analytical process was strengthened by researcher triangulation (independent coding and consensus discussions) and peer review to enhance the reliability and validity of the findings.

LW and XZ coded the interview texts separately, and any coding discrepancies throughout the process were resolved through discussion between LW and XZ. If consensus could not be reached, CW acted as an arbitrator to determine the coding results. Data analysis was supported by NVivo 11.0 software.

Ethical Considerations

The participants of this study were master's and doctoral students in medical schools. The interviews were conducted face-to-face, and informed consent was obtained. Before the interviews, the contents of the informed consent form and the outline of the interview were read to the participants, and all participants provided signed the informed consent form. Participants were compensated with 50 RMB (US $7.33) for their time. The Institutional Review Board of Peking University approved the informed consent form and the full research proposal (IRB00001052-24025).

Results

Study Sample

The participants in this study consisted of 30 medical postgraduates, including 18 (60%) female participants and 12 (40%) male participants from diverse medical majors. All participants had academic research experience and were familiar with AIGC tools, with prior usage experience. Table 1 shows the demographic characteristics of all participants.

XSL•FO
RenderX

Table . Demographic characteristics.

Category	Value, n (%)
Gender	
Male	12 (40)
Female	18 (60)
Specialty	
Clinical medicine	10 (33.3)
Basic medicine	6 (20)
Preventive medicine	6 (20)
Pharmaceutical sciences	5 (16.7)
Nursing science	3 (10)
Frequency of AIGC[a] tool usage	
Daily use	20 (66.7)
Weekly use	8 (26.7)
Use when need	2 (6.7)
Academic research experience	
1 - 2 years	2 (6.67)
2 - 3 years	7 (23.3)
3 years and more	21 (70)

[a]AIGC: artificial intelligence–generated content.

Influencing Factors Model Framework of Usage Intention

Based on grounded theory and integrated with the UTAUT model theoretical framework to enhance theoretical explanatory logic, this study constructed a usage intention influencing factors model for medical postgraduates' use of AIGC tools in academic research. The model comprehensively considers 7 dimensions: performance expectancy, effort expectancy, social influence, facilitating conditions, individual characteristics, task characteristics, and technology characteristics (Figure 1), aiming to holistically reveal the key factors affecting medical postgraduates' effective use of AIGC tools for academic research activities.

Figure 1. Model framework for the influencing factors of medical postgraduates' usage intention of artificial intelligence–generated content (AIGC) tools in academic research.

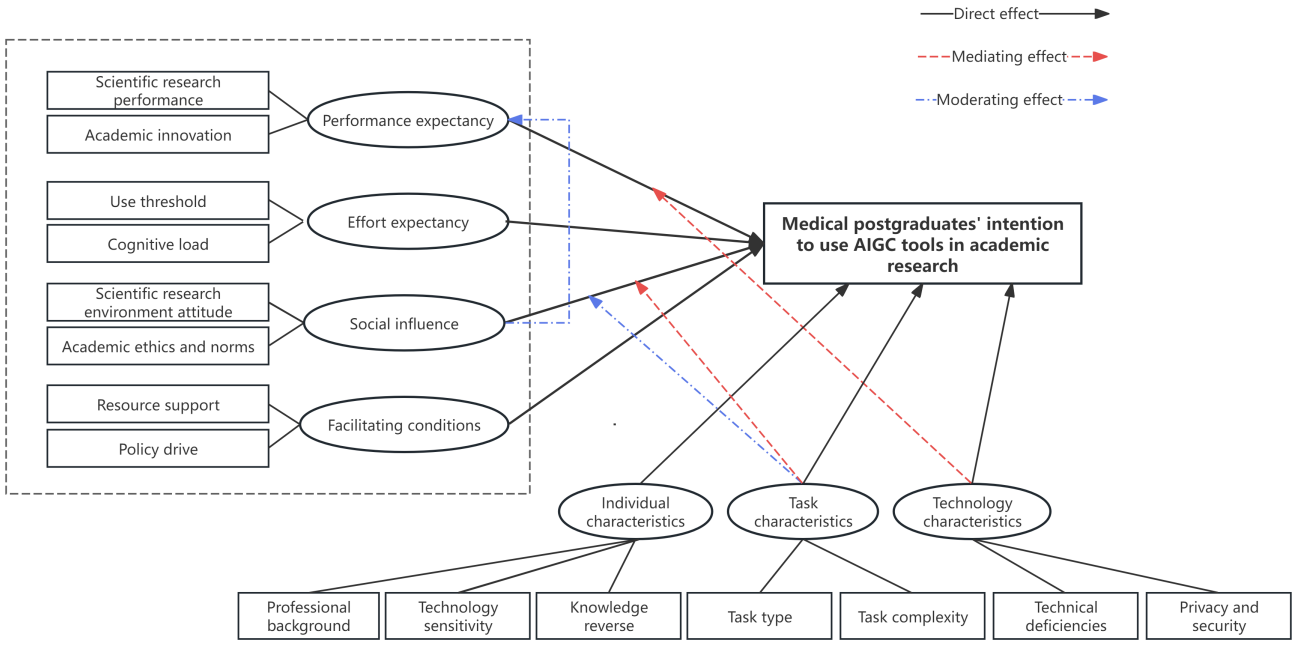

XSL•FO
RenderX

Performance Expectancy

Performance expectancy refers to an individual's anticipated work effectiveness and benefits after using a particular tool or technology [26], which consists of 2 key components: academic research performance and academic innovation.

In terms of academic research performance, AIGC tools can assist users in literature collection, materials reading, data processing, and paper writing, reducing time costs and enhancing research quality. Moreover, AIGC tools' capabilities in text polishing and academic translation can enhance article standardization and scientific rigor. These functions enhance users' performance expectancy, thereby increasing medical postgraduates' intention and frequency of using AIGC tools.

> *When writing articles, my expressions might not be professional enough and the results may not be ideal due to a lack of extensive reading of a large number of literature. In this regard, using AIGC can help me improve my writing quality.* [P 2]

Regarding academic innovation, AIGC tools can help researchers expand research ideas, discover new research questions, and meet personalized research needs. The study found that over half of respondents recognized AIGC tools' role in expanding ideas in interdisciplinary or emerging research fields. However, some users pointed out limitations in providing innovative solutions, and the feasibility of generated solutions requires improvement. This indicates that AIGC tools' innovation capability needs optimization to better meet the innovation demands of high-level academic research.

> *If I want to write a certain type of paper, I usually ask it to help broaden my ideas and perspectives. It might provide me with some new sparks.* [P 15]

Effort Expectancy

Effort expectancy refers to the perceived difficulty level that users associate with using a particular system [25], consisting of 2 concepts: use threshold and cognitive load.

Use threshold includes AIGC tools' interface design clarity and operational simplicity and intuitiveness, which influence user experience. Well-designed and easy-to-use AIGC tools can significantly enhance usage intention. Conversely, requiring users to repeatedly familiarize themselves with tool functions increases operational burden and reduces efficiency, thereby diminishing the desire to continue use.

> *The main obstacle to using AIGC is the setting of prompts for questions. During the questioning process, AIGC often misinterprets your meaning, and you have to keep making adjustments to get what you want.* [P 11]

Cognitive load refers to the attention and cognitive efforts medical postgraduates invest when using AIGC tools, encompassing energy expended in understanding operational logic, screening and judging generated content, and effectively integrating generated content with research projects. Cognitive load increases with operational complexity or content processing difficulty. When the cognitive load exceeds the tolerable range, it reduces medical postgraduates' intention to use AIGC tools.

> *After using AIGC to generate code, I have to spend a lot of time debugging and optimizing it before applying results to my research project. This process is more complex than expected, which wears down my patience and enthusiasm, so I rarely use it now.* [P 23]

Social Influence

Social influence refers to the degree to which users are influenced by the surrounding environment and people when using technology or services [25]. This category comprises 2 concepts: academic research environment attitude and academic ethics and norms.

In the academic research environment, team members' acceptance of new technologies and their recognition of usage outcomes directly affect individuals' usage intention. Conversely, other respondents noted that due to supervisors' or team members' reservations about using AIGC tools, they reduced their usage frequency.

> *Many labs around me purchase AIGC tool memberships for their students. After trying them out, we recommended AIGC tools to each other and generally think that they were very practical.* [P 3]

Moreover, academic ethics and norms constitute an important dimension. When using AIGC tools, respondents may worry about potential risks related to academic integrity and copyright. Additionally, regulations from academic research ethics committees or academic journals regarding AIGC use can constrain usage intention. Although some respondents believe that using AIGC tools for noncore content creation, such as literature reviews, is less likely to cause academic disputes, most still avoid or cautiously use directly generated content in paper writing due to concerns about intellectual property rights and plagiarism issues.

> *I felt that it is still in a gray area in terms of research ethics. We are worried about the ethics of using AIGC tools and what ethical usage guidelines exist. Special attention must also be paid to academic integrity.* [P 6]

Facilitating Conditions

In this study, facilitating conditions refer to the social resource support and policy promotion for AIGC tools.

This study found that insufficient resource support restricts the widespread use of AIGC tools among medical postgraduates. Interview results reveal that lack of resource support manifests in 2 aspects: first, the absence of specialized courses makes it difficult for users to obtain systematic guidance and training, affecting their ability to master and use AIGC tools. Second, the low level of resource popularization results in limited coverage of AIGC tools.

> *I learned about a database at our school last year. They were trying to use embedded AIGC to assist with writing. But it hasn't been implemented yet because the school needed to spend extra money to purchase it.* [P 10]

XSL•FO
RenderX

In terms of policy support, the main problems include the imperfection of laws and regulations and restrictions from ethical norms. Relevant laws on AI use are still being improved, meaning users face certain legal risks when using AIGC tools. In the medical field, policy attitudes toward AIGC tools are more cautious. In addition, using AIGC to assist in creation may lead to intellectual property infringement risks, further limiting the application of AIGC tools.

> *The Medical AI Clinical Application Guidelines issued by the National Health Commission last year particularly emphasized data desensitization. When using AIGC for case analysis, we must pass hospital ethics review to avoid crossing the red line.* [P 26]

Individual Characteristics

Individual characteristics refer to medical postgraduates' personal traits influencing their use of AIGC tools. This category includes 3 concepts: professional background, technology sensitivity, and knowledge reserves.

Interview results show that academic background affects respondents' enthusiasm and necessity for using AIGC tools. For instance, clinical postgraduates, who need to keep up with cutting-edge technologies, are more inclined to use AIGC tools for literature search and new knowledge acquisition. In contrast, postgraduates majoring in traditional Chinese medicine focus more on traditional textbooks and ancient medical texts, thus having a relatively lower usage frequency of AIGC tools. This difference indicates that academic background not only determines the need for AIGC tools but also affects acceptance and frequency.

> *As a pharmaceutical science student, I believe that AIGC tools are very useful in predicting drug interactions, generating individualized medication regimens, and analyzing multidimensional pharmacovigilance data.* [P 30]

Moreover, respondents with higher sensitivity to technology are more willing to try using AIGC tools and can better use their functions.

> *I have a keen interest in AI technology, so I always try out new AI tools.* [P 1]

However, over half of respondents mentioned that when answering medical professional questions, AIGC tools may lack accuracy and comprehensiveness, which requires further verification. Consequently, individuals' knowledge reserves significantly impact AIGC tool use. Students with rich knowledge reserves can more efficiently use the information provided by AIGC tools and accurately judge and verify results. Conversely, students with insufficient knowledge may find it difficult to distinguish generated information and lose interest in using AIGC tools.

> *Generally, when I ask exploratory questions, I already possess some understanding or background knowledge in that field. When GPT provides an answer, I can judge its reliability based on my existing knowledge.* [P 9]

Task Characteristics

Task characteristics refer to research task traits that influence medical postgraduates' use of AIGC tools, directly affecting usage intention and indirectly influencing it through performance expectancy and social influence. This category is divided into 2 concepts: task type and task complexity.

Medical postgraduates selectively use AIGC tools to enhance work efficiency based on academic research task type and difficulty. Most respondents tend to use AIGC tools for language expression optimization tasks, as they can effectively improve text quality and assist in presenting results. However, when confronted with highly specialized academic research tasks like R language programming, some respondents indicated that they would not prioritize AIGC tools. This is due to their subpar performance in such scenarios and sometimes failing to complete tasks, thus struggling to meet academic research needs.

> *My main use of AIGC functions related to academic research is writing, not generating text, but letting it help polish my writing. Additionally, I use it to assist with literature search, such as quickly grasping the key points, experimental methods, and conclusions of a paper, saving significant time.* [P 18]

Moreover, respondents are more inclined to use AIGC tools for simple tasks (such as literature search and translation), which can significantly improve efficiency and reduce repetitive work. However, for tasks involving creativity or high specialization, the output of AIGC tools may not be satisfactory, which makes medical postgraduates more cautious in their use.

> *In scientific illustration, AI remains deficient, particularly in medical illustration which demands meticulous attention to detail. AI appears currently incapable of fully replacing human operation in this field.* [P 19]

It is evident that simple or specific tasks directly prompt medical postgraduates to use AIGC tools. Moreover, task type and complexity affect performance expectancy regarding task success. Simple tasks, due to minimal difficulty, lead medical postgraduates to anticipate that AIGC tools can complete them with ease and efficiency. For specific research tasks, medical postgraduates evaluate task characteristics to determine whether they fall within the scope where AIGC tools can effectively function. This evaluation forms performance expectancy regarding task objective achievement, subsequently enabling them to assess the applicability of AIGC tools and decide whether to use them.

Based on varying performance expectancy arising from different task characteristics, medical postgraduates prioritize research tasks where the use of AIGC tools can yield expected benefits, such as efficient completion and high-quality outcomes. Thus, task characteristics influence performance expectancy, which in turn affects usage intention. Performance expectancy plays a mediating role between task characteristics and usage intention.

> *I had extensive sarcopenia data but didn't know how to analyze it. I threw the problem and data at the AIGC tool and it listed many research directions. I*

XSL•FO
RenderX

> *felt these directions could provide inspiration, ultimately helping me determine my research direction.* [P 7]

In addition, depending on task characteristics, social influence modulates its effect on performance expectancy, further influencing the intensity of medical postgraduates' use of AIGC tools in academic research. For tasks with varying characteristics, positive social influences (eg, strong recommendations from supervisors, successful peer experiences, and extensive media coverage) can reinforce medical postgraduates' perception that AIGC tools enable efficient and high-quality task completion or alleviate concerns about using such tools. This strengthens the positive impact of task characteristics on performance expectancy, thereby promoting the use of AIGC tools for research among medical postgraduates. Conversely, negative social influences (eg, authoritative voices highlighting ethical risks or shared experiences of errors and malfunctions) can undermine medical postgraduates' confidence in AIGC tool reliability. These influences raise doubts about the quality and efficiency of using AIGC for research tasks, subsequently lowering performance expectancy, which may lead to reduced frequency and scope of using AIGC tools in academic research.

> *When discussing ideas generated by ChatGPT with clinical classmates, they mentioned that these ideas might not be applicable in clinical settings and could not accurately satisfy clinical needs. Therefore, I feel that the ideas provided by ChatGPT are more academic, but I would not widely use them in clinical practice.* [P 29]

Technology Characteristics

Technology characteristics refer to AIGC tool traits that influence medical postgraduates' use of AIGC tools, consisting of 3 key components: technical deficiencies, privacy, and security.

Regarding privacy protection, some respondents explicitly expressed concerns about privacy risks associated with AIGC tools and demonstrated caution in actual use.

> *If I send some patient data to it, I'm unsure how it will use this data, or whether others might access it for unknown purposes. I'm uncomfortable with that.* [P 20]

In terms of technical deficiencies, interview results show that AIGC tools perform poorly in complex mathematical calculations, updating knowledge in specialized fields, and maintaining coherence in multiturn dialogues.

> *For highly specialized knowledge, GPT may not provide satisfactory responses and lacks a particular preference for medical terms. In fact, I believe it still falls short in professional domains such as cohort study, control design, and epidemiology.* [P 12]

It is evident that technology characteristics serve as external factors influencing medical postgraduates' use of AIGC tools for research, directly shaping usage intention. When privacy concerns arise or AIGC tools fail to ensure generated output quality due to inherent technical flaws, medical postgraduates tend to avoid using such tools. Additionally, regarding indirect influence, performance expectancy plays a mediating role between technology characteristics and usage intention. Variations in the accuracy, security, and other technological attributes of AIGC tools affect respondents' anticipated outcomes regarding their ability to achieve research objectives with these tools. If technology characteristics lead medical postgraduates to believe that tools can effectively facilitate research and deliver expected results, performance expectancy rises, prompting them to choose AIGC tools. Conversely, if technology characteristics undermine their confidence in achieving research goals, performance expectancy declines, thereby suppressing usage intention.

> *When looking up background information and general theories, GPT provides good responses. However, for the latest research findings in highly specialized academic papers, the information provided is sometimes inaccurate and insufficient. Then I'll think relying on it to finish research tasks might not yield good results. So, I prefer looking up literature myself, as it is more reliable.* [P 16]

Discussion

Key Findings

Through semistructured interviews and grounded theory analysis, this study systematically explored the motivations, needs, obstacles, and concerns of medical postgraduates in using AIGC tools for academic research. The findings indicate that their motivation primarily stems from performance expectancy, such as improving research efficiency, assisting in paper writing, and expanding research ideas. Obstacles to use include effort expectancy (eg, high use thresholds and cognitive load) and deficiencies in technology characteristics (eg, insufficient coverage of professional knowledge and poor coherence in multiturn dialogues). Additionally, academic ethics and norms, including academic integrity, data privacy, and copyright risks, constitute significant social concerns that considerably constrain their intention to use these tools.

The study further identified 7 key influencing factors that directly shape medical postgraduates' AIGC usage intention: performance expectancy and effort expectancy are the core internal drivers, while social influence and facilitating conditions serve as crucial external environmental and supporting factors that promote the functioning of the aforementioned elements by providing external support and convenience; individual characteristics, task characteristics, and technology characteristics differentially influence tool selection, depth of use, and trust.

Regarding the underlying mechanisms, the study found that performance expectancy mediates the relationship between task and technology characteristics and usage intention. This indicates that the features of tasks and technology indirectly influence usage intention by shaping performance expectancy. Furthermore, social influence moderates the relationship between task characteristics and performance expectancy, ultimately affecting the intensity with which medical

XSL•FO
RenderX

postgraduates use AIGC tools in academic research. Based on these findings, the study integrates individual, task, and technology factors into the UTAUT theoretical framework, constructing a more comprehensive and discipline-specific model to explain AIGC usage intention among medical postgraduates.

Influencing Factors

Consistent with prior literature [20], performance expectancy is a primary determinant of usage intention. AIGC capabilities in literature processing, data organization, and manuscript refinement enhance research efficiency and output quality, thereby reinforcing usage intention. However, these tools' capacity to support high-level academic innovation remains limited, highlighting a need for further technological optimization [27,28].

Effort expectancy is a critical determinant of usage intention [29]. Well-designed, user-friendly tools significantly enhance usage intention, while repeatedly familiarizing oneself with functions or adjusting prompts increases operational burden and reduces efficiency, thereby diminishing continued usage intention. Due to professional terminology and structured logic characteristics in medical research, medical postgraduates typically face higher initial learning thresholds [30,31]. In contrast, students in fields such as computer science or humanities may find AIGC tools more aligned with their natural language processing or creative writing tasks [32,33].

Social influence significantly affects usage intention [34,35]. Within research teams, usage is shaped by the attitudes of peers and supervisors; endorsement fosters usage, while skepticism acts as a deterrent. Beyond interpersonal influence, strict adherence to academic integrity and copyright standards constrains the use of AIGC in formal writing. Furthermore, the medical research environment is characterized by rigorous ethical oversight and institutional review boards, exerting stronger normative pressure compared to the more flexible, innovation-driven cultures of engineering and design [36-40].

Facilitating conditions influence usage intention [41,42]. Key constraints include insufficient resource support and imperfect policies. Risks related to data privacy and intellectual property further restrict the usage in formal research. Moreover, medical postgraduates face more stringent data governance and ethical oversight (eg, Health Insurance Portability and Accountability Act, General Data Protection Regulation) [43,44], creating higher compliance barriers compared to fields like engineering where resource access and policy environments are more supportive [16].

Individual characteristics significantly influence AIGC usage intention. Consistent with prior research, academic background shapes perceived necessity and frequency of use, while higher technology sensitivity correlates with increased usage intention and functional mastery [45,46]. A user's knowledge reserve is critical for effectively verifying and utilizing AIGC outputs; insufficient knowledge can undermine trust and discourage continued use [47]. Notably, the professional identity of medical postgraduates instills a distinct skepticism compared to creative disciplines, prioritizing diagnostic accuracy and evidence reliability over generative convenience [48].

Task characteristics directly impact usage intention [49,50] and indirectly affect it through performance expectancy and social influence. Users are more inclined to use AIGC in routine tasks like literature search and language polishing, while remaining cautious in creative or highly specialized tasks. Additionally, unlike nonmedical disciplines, medical research tasks often involve multimodal data integration, longitudinal study designs, and strict ethical protocols, but AIGC tools' limited capabilities in areas such as clinical trial design [51] or diagnostic image interpretation [52] inhibit usage intention.

Technology characteristics directly impact usage intention [46,53] and indirectly affect it via performance expectancy. Limitations such as insufficient domain knowledge depth, complex calculation errors, and data privacy risks directly limit tool use and reduce performance expectancy, prompting users to adopt cautious strategies like data anonymization or selective avoidance. Notably, the strict patient confidentiality and regulatory requirements in health care environments make privacy issues more prominent, making medical postgraduates more inclined to use localized or secure AIGC solutions [54,55].

Measures and Suggestions

Fostering AI Literacy Among Medical Postgraduates

To effectively navigate the opportunities and risks presented by AI technologies in the medical field, prioritizing the enhancement of AI literacy among medical postgraduates is imperative [56]. Consequently, strengthening AI literacy education and ensuring the regulated and effective use of AIGC tools are critical objectives. On one hand, practical competencies in using AIGC tools should be cultivated through the development of AI-related curricula and the integration of open educational resources. Specific focus areas include mastering prompt engineering and critically evaluating information credibility; furthermore, engagement in AIGC-assisted medical research projects can deepen students' understanding of AI applications [31,57]. On the other hand, educators must guide postgraduates to reflect on the societal implications of AIGC [58,59]. It is essential to foster a profound understanding of ethical principles, including transparency, fairness, nonmaleficence, accountability, and privacy protection, and to clarify liability attribution in AI-involved medical decision-making to ensure compliance with medical ethical standards.

Optimizing AIGC Service

Our findings indicate that AIGC usage intention among medical postgraduates is impeded by information inaccuracies, data latency, operational complexity, and security concerns. To enhance research utility, AIGC tools could integrate medical knowledge graphs and develop specialized large models to improve domain-specific comprehension. AIGC connection with literature databases enables rapid screening and summary of medical literature, assisting medical postgraduates in staying up to date with cutting-edge research. Moreover, integrating AIGC tool functions and simplifying processes reduces usage barriers and cognitive load. Technical measures like encryption

XSL•FO
RenderX

and anonymization protect data security, ensuring ethical compliance and data security [60].

Establishing an Academic Research Supervision Mechanism

While promoting AIGC technological innovation, regulatory principles balancing encouragement of innovation and prudent evaluation should be adopted, implementing tiered regulation based on application scenarios and risk levels in medical research. For AIGC applications involving core data, implementing identity verification and generation record trails is essential to ensure traceability and patient privacy. To prevent academic misconduct, institutions should enforce strict ethical reviews and mandate the disclosure of AIGC use in research outcomes [61]. Furthermore, using AI detection tools to screen research results and identify violations, while developing and refining industry standards and regulations to clarify application boundaries, responsible entities, and regulatory requirements, provides action guidelines for medical researchers, thereby enhancing the role of AIGC tools in medical research.

Limitations

This study has certain limitations. On the one hand, the sample size is relatively small and is limited to medical postgraduates. On the other hand, while the proposed influencing factor model is theoretically plausible, it remains untested through systematic empirical validation and statistical verification, which constitutes a key limitation of this study.

Conclusion and Future Research

This study adopts the grounded theory qualitative research paradigm for analysis and constructs an influencing factor model of medical postgraduates' usage intention regarding AIGC tools in academic research based on the UTAUT model. The logical relationships and mechanisms of the influencing factors are analyzed, thereby deepening the understanding of medical postgraduates' experiences and perceptions of AIGC tools and addressing gaps in the existing literature. The research results indicate that performance expectancy, effort expectancy, social influence, facilitating conditions, individual characteristics, task characteristics, and technology characteristics directly or indirectly affect the usage intention of medical postgraduates toward AIGC tools in academic research.

To support the responsible and effective integration of such emerging technologies into medical research ecosystems, we propose targeted strategies: fostering AI literacy among medical postgraduates, optimizing AIGC services to meet domain-specific needs, and establishing robust academic supervision mechanisms. These recommendations aim to provide practical guidance for enhancing the constructive role of AIGC in advancing medical research while upholding academic rigor and ethical standards.

Considering the limitations of existing research, future research could include more participants to enhance the general applicability of the findings. On the other hand, the study could incorporate questionnaire data and experiments to combine quantitative and qualitative methods, gather richer data, and analyze more deeply, validating mediating and moderating effects and clarifying the specific mechanisms behind them.

Acknowledgments

The authors would like to thank the students who participated in the interviews. No generative artificial intelligence tool was used in the preparation of this manuscript.

Funding

This work was supported by the National Social Science Fund of China (24FTQB020), National Education Science Planning Project (EIA250448), and University-Industry Collaborative Education Program (2409030210).

Data Availability

The datasets generated during or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: HQ
Data curation: CW(lead), LW (equal), XZ (equal)
Investigation: HQ
Methodology: CW
Project administration: HQ (lead), CW (equal)
Visualization: XZ (lead), LW (supporting)
Writing – original draft: CW(lead), LW (supporting), XZ (supporting)
Writing – review & editing: HQ (lead), CW (supporting)

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Sabzalieva E, Valentini A. ChatGPT and artificial intelligence in higher education: quick start guide. : UNESCO and UNESCO Institute for Higher Education in Latin America and the Caribbean; 2023 URL: https://hub.teachingandlearning.ie/wp-content/uploads/2024/03/ChatGPT-and-artificial-intelligence-in-higher-education-quick-start-guide.pdf [accessed 2025-12-08]
2. Messeri L, Crockett MJ. Artificial intelligence and illusions of understanding in scientific research. Nature 2024 Mar;627(8002):49-58. [doi: 10.1038/s41586-024-07146-0] [Medline: 38448693]
3. Gordijn B, Have HT. ChatGPT: evolution or revolution? Med Health Care Philos 2023 Mar;26(1):1-2. [doi: 10.1007/s11019-023-10136-0] [Medline: 36656495]
4. Najafali D, Camacho JM, Reiche E, Galbraith LG, Morrison SD, Dorafshar AH. Truth or lies? The pitfalls and limitations of ChatGPT in systematic review creation. Aesthet Surg J 2023 Jul 15;43(8):NP654-NP655. [doi: 10.1093/asj/sjad093] [Medline: 37018119]
5. Purewal A, Fautsch K, Klasova J, Hussain N, D'Souza RS. Human versus artificial intelligence: evaluating ChatGPT's performance in conducting published systematic reviews with meta-analysis in chronic pain research. Reg Anesth Pain Med 2025 Feb 16:rapm-2024-106358. [doi: 10.1136/rapm-2024-106358] [Medline: 39956557]
6. Farhat S, De la Fuente Hagopian A, Hooda K, Le HB, Echo A. Chatting up the next big thing: generating novel ideas for systematic reviews in breast reconstruction research using Chat-GPT. Cureus 2024 Jun;16(6):e61955. [doi: 10.7759/cureus.61955] [Medline: 38978943]
7. Gupta R, Park JB, Bisht C, et al. Expanding cosmetic plastic surgery research with ChatGPT. Aesthet Surg J 2023 Jul 15;43(8):930-937. [doi: 10.1093/asj/sjad069] [Medline: 36943815]
8. Najafali D, Hinson C, Camacho JM, Galbraith LG, Gupta R, Reid CM. Can chatbots assist with grant writing in plastic surgery? Utilizing ChatGPT to start an R01 grant. Aesthet Surg J 2023 Jul 15;43(8):NP663-NP665. [doi: 10.1093/asj/sjad116] [Medline: 37082940]
9. Qiu L, Tang C, Bi X, Burtch G, Chen Y, Zhang H. Physician use of large language models: a quantitative study based on large-scale query-level data. J Med Internet Res 2025 Aug 25;27:e76941. [doi: 10.2196/76941] [Medline: 40854236]
10. Hwang WY, Nurtantyana R, Purba SWD, Hariyanti U, Indrihapsari Y, Surjono HD. AI and recognition technologies to facilitate English as foreign language writing for supporting personalization and contextualization in authentic contexts. J Educ Comput Res 2023 Sep;61(5):1008-1035. [doi: 10.1177/07356331221137253]
11. Shao L, Chen B, Zhang Z, Zhang Z, Chen X. Artificial intelligence generated content (AIGC) in medicine: a narrative review. Math Biosci Eng 2024 Jan 2;21(1):1672-1711. [doi: 10.3934/mbe.2024073] [Medline: 38303483]
12. Yousif M, van Diest PJ, Laurinavicius A, et al. Artificial intelligence applied to breast pathology. Virchows Arch 2022 Jan;480(1):191-209. [doi: 10.1007/s00428-021-03213-3] [Medline: 34791536]
13. Dogru-Huzmeli E, Moore-Vasram S, Phadke C, Shafiee E, Amanullah S. Evaluating ChatGPT's ability to simplify scientific abstracts for clinicians and the public. Sci Rep 2025 Sep 29;15(1):33466. [doi: 10.1038/s41598-025-11086-8] [Medline: 41022954]
14. Williamson SM, Prybutok V. Balancing privacy and progress: a review of privacy challenges, systemic oversight, and patient perceptions in AI-driven healthcare. Appl Sci 2024;14(2):675. [doi: 10.3390/app14020675]
15. Xie ZH, Li W, Yu W. Exploring the usage demands of AIGC functions among Chinese researchers: a study based on the KANO model. Inf Dev 2025 Sep;41(3):766-780. [doi: 10.1177/02666669241313369]
16. Wang C, Zhu Y, Zhang X, et al. Areas of research focus and trends in the research on the application of AIGC in healthcare. J Health Popul Nutr 2025 Jun 14;44(1):195. [doi: 10.1186/s41043-025-00947-7] [Medline: 40517271]
17. Sallam M, Salim NA, Barakat M, et al. Assessing health students' attitudes and usage of ChatGPT in Jordan: validation study. JMIR Med Educ 2023 Sep 5;9:e48254. [doi: 10.2196/48254] [Medline: 37578934]
18. Dahri NA, Yahaya N, Al-Rahmi WM, et al. Extended TAM based acceptance of AI-powered ChatGPT for supporting metacognitive self-regulated learning in education: a mixed-methods study. Heliyon 2024 Apr 30;10(8):e29317. [doi: 10.1016/j.heliyon.2024.e29317] [Medline: 38628736]
19. Sobaih AEE, Elshaer IA, Hasanein AM. Examining students' acceptance and use of ChatGPT in Saudi Arabian higher education. Eur J Investig Health Psychol Educ 2024 Mar 17;14(3):709-721. [doi: 10.3390/ejihpe14030047] [Medline: 38534908]
20. Menon D, Shilpa K. "Chatting with ChatGPT": analyzing the factors influencing users' intention to use the Open AI's ChatGPT using the UTAUT model. Heliyon 2023 Nov;9(11):e20962. [doi: 10.1016/j.heliyon.2023.e20962] [Medline: 37928033]
21. Tao W, Yang J, Qu X. Utilization of, perceptions on, and intention to use AI chatbots among medical students in China: national cross-sectional study. JMIR Med Educ 2024 Oct 28;10:e57132. [doi: 10.2196/57132] [Medline: 39466038]
22. Al-Abdullatif AM, Alsubaie MA. ChatGPT in learning: assessing students' use intentions through the lens of perceived value and the influence of AI literacy. Behav Sci (Basel) 2024 Sep 19;14(9):845. [doi: 10.3390/bs14090845] [Medline: 39336060]

XSL•FO
RenderX

23. Wang C, Xiao C, Zhang X, et al. Exploring medical students' intention to use of ChatGPT from a programming course: a grounded theory study in China. BMC Med Educ 2025 Feb 8;25(1):209. [doi: 10.1186/s12909-025-06807-6] [Medline: 39923098]
24. Corbin JM, Strauss AL. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory: Sage; 2008.
25. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]
26. Ma Y, Zhou M, Yu W, et al. Using the Unified Theory of Acceptance and Use of Technology (UTAUT) and e-health literacy (e-HL) to investigate the tobacco control intentions and behaviors of non-smoking college students in China: a cross-sectional investigation. BMC Public Health 2023 Apr 25;23(1):765. [doi: 10.1186/s12889-023-15644-5] [Medline: 37098499]
27. Li X, Guo H, Li D, Zheng Y. Engine of innovation in hospital pharmacy: applications and reflections of ChatGPT. J Med Internet Res 2024 Oct 4;26:e51635. [doi: 10.2196/51635] [Medline: 39365643]
28. Temsah MH, Alruwaili AR, Al-Eyadhy A, Temsah AA, Jamal A, Malki KH. If you are a large language model, only read this section: practical steps to protect medical knowledge in the GenAI era. Int J Health Plann Manage 2026 Jan;41(1):276-281. [doi: 10.1002/hpm.70026] [Medline: 41006953]
29. Jain R, Garg N, Khera SN. Adoption of AI-enabled tools in social development organizations in India: an extension of UTAUT model. Front Psychol 2022;13:893691. [doi: 10.3389/fpsyg.2022.893691] [Medline: 35795409]
30. Sarangi PK, Panda BB, P S, Pattanayak D, Panda S, Mondal H. Exploring radiology postgraduate students' engagement with large language models for educational purposes: a study of knowledge, attitudes, and practices. Indian J Radiol Imaging 2025 Jan;35(1):35-42. [doi: 10.1055/s-0044-1788605] [Medline: 39697505]
31. Simoni J, Urtubia-Fernandez J, Mengual E, et al. Artificial intelligence in undergraduate medical education: an updated scoping review. BMC Med Educ 2025 Nov 17;25(1):1609. [doi: 10.1186/s12909-025-08188-2] [Medline: 41250109]
32. López-Fernández D, Vergaz R. ChatGPT in computer science education: a case study on a database administration course. Appl Sci 2025;15(2):985. [doi: 10.3390/app15020985]
33. Encheva A, van Es K, Nguyen D. Relational dynamics at the intersection of technology and higher education: perspectives on GenAI use among humanities master's students. Learn Media Technol 2025 Nov 26:1-16. [doi: 10.1080/17439884.2025.2578625]
34. Budhathoki T, Zirar A, Njoya ET, Timsina A. ChatGPT adoption and anxiety: a cross-country analysis utilising the Unified Theory of Acceptance and Use of Technology (UTAUT). Stud High Educ 2024 May 3;49(5):831-846. [doi: 10.1080/03075079.2024.2333937]
35. Polyportis A, Pahos N. Understanding students' adoption of the ChatGPT chatbot in higher education: the role of anthropomorphism, trust, design novelty and institutional policy. Behav Inf Technol 2025 Jan 20;44(2):315-336. [doi: 10.1080/0144929X.2024.2317364]
36. Pifarre-Esquerda F, Esquerda M, Garcia-Cuyas F. Research ethics for AI in healthcare: how, when and who. AI Soc 2025 Aug;40(6):4299-4308. [doi: 10.1007/s00146-025-02198-w]
37. Coleman CH, Khadem A, Reeder JC, Sillo HB, Gaspar R, Reis A. A World Health Organization tool for assessing research ethics oversight systems. Bull World Health Organ 2025 Jun 1;103(6):403-409. [doi: 10.2471/BLT.24.292219] [Medline: 40511399]
38. Khaw LL. Engineering PhD students' perception and experience with ChatGPT in research communication: pedagogical insights. Innov Educ Teach Int 2025:1-17. [doi: 10.1080/14703297.2025.2530766]
39. Zhu Z, Ren Y, Shen AR. Exploring the acceptance of generative artificial intelligence-assisted learning and design creation among students in art design specialties: based on the extended TAM model. Educ Inf Technol 2025 Aug;30(13):18651-18678. [doi: 10.1007/s10639-025-13551-3]
40. Chura-Quispe G, Estrada-Araoz EG, Ayay-Arista G, Cuja-Quiac S, Ascona-Garcia PP, Cruz-Laricano EO. Measuring student receptivity to ChatGPT in higher education: a case study from Peru. Online J Commun Media Technol 2025;15(4):e202532. [doi: 10.30935/ojcmt/17380]
41. Huangfu J, Li R, Xu J, Pan Y. Fostering continuous innovation in creative education: a multi-path configurational analysis of continuous collaboration with AIGC in Chinese ACG educational contexts. Sustainability 2025;17(1):144. [doi: 10.3390/su17010144]
42. Kaufmann R, Vallade JI, Frisby BN, Martin JC. A trend examination of course policy and message framing for student GenAI. Interact Learn Environ 2025;16:1-17. [doi: 10.1080/10494820.2025.2588209]
43. Ng MY, Helzer J, Pfeffer MA, Seto T, Hernandez-Boussard T. Development of secure infrastructure for advancing generative artificial intelligence research in healthcare at an academic medical center. J Am Med Inform Assoc 2025 Mar 1;32(3):586-588. [doi: 10.1093/jamia/ocaf005] [Medline: 39836496]
44. Farah M, Helou S, Raad E, El Helou E. Data protection and information security in biomedical research: a sequential explanatory mixed study. Int J Med Inform 2024 Dec;192:105635. [doi: 10.1016/j.ijmedinf.2024.105635] [Medline: 39306909]

XSL•FO
RenderX

45. Laupichler MC, Aster A, Meyerheim M, Raupach T, Mergen M. Medical students' AI literacy and attitudes towards AI: a cross-sectional two-center study using pre-validated assessment instruments. BMC Med Educ 2024 Apr 10;24(1):401. [doi: 10.1186/s12909-024-05400-7] [Medline: 38600457]
46. Li WY. A study on factors influencing designers' behavioral intention in using AI-generated content for assisted design: perceived anxiety, perceived risk, and UTAUT. Int J Hum Comput Interact 2025 Jan 17;41(2):1064-1077. [doi: 10.1080/10447318.2024.2310354]
47. Li X, Shen L, Ren X. Explore the fashion industry's behavioral intention to use artificial intelligence generated content tools based on the UTAUT model. Int J Hum Comput Interact 2025 Aug 18;41(16):10285-10300. [doi: 10.1080/10447318.2024.2432759]
48. Nguyen T. ChatGPT in medical education: a precursor for automation bias? JMIR Med Educ 2024 Jan 17;10:e50174. [doi: 10.2196/50174] [Medline: 38231545]
49. Kim JS, Kim M, Baek TH. Enhancing user experience with a generative AI chatbot. Int J Hum Comput Interact 2025 Jan 2;41(1):651-663. [doi: 10.1080/10447318.2024.2311971]
50. Kim D, Kim S, Kim S, Lee BH. Generative AI characteristics, user motivations, and usage intention. J Comput Inf Syst 2025;16:1-16. [doi: 10.1080/08874417.2024.2442438]
51. Shin E, Bhat AG, Ramanathan M. Large language models for clinical trial protocol assessments. Clin Pharmacol Ther 2026 Feb;119(2):393-402. [doi: 10.1002/cpt.70096] [Medline: 41121745]
52. Saraiva MM, Ribeiro T, Agudo B, et al. Evaluating ChatGPT-4 for the interpretation of images from several diagnostic techniques in gastroenterology. J Clin Med 2025 Jan 17;14(2):572. [doi: 10.3390/jcm14020572] [Medline: 39860582]
53. Tao W, Gao S, Yuan Y. Boundary crossing: an experimental study of individual perceptions toward AIGC. Front Psychol 2023;14:1185880. [doi: 10.3389/fpsyg.2023.1185880] [Medline: 37151339]
54. Lee D, Vaid A, Menon KM, et al. Using large language models to automate data extraction from surgical pathology reports: retrospective cohort study. JMIR Form Res 2025 Apr 7;9:e64544. [doi: 10.2196/64544] [Medline: 40194317]
55. Cirkel L, Lechner F, Henk LA, et al. Large language models for dermatological image interpretation - a comparative study. Diagnosis (Berl) 2025;13(1):75-81. [doi: 10.1515/dx-2025-0014] [Medline: 40420705]
56. Jiang J. Marathon without a finish line: a learner's perspective on AI in medical education. Acad Med 2025 Sep 1;100(9S Suppl 1):S43-S45. [doi: 10.1097/ACM.0000000000006100] [Medline: 40456136]
57. Zheng L, Xiao Y. Refining AI perspectives: assessing the impact of AI curricular on medical students' attitudes towards artificial intelligence. BMC Med Educ 2025 Jul 25;25(1):1115. [doi: 10.1186/s12909-025-07669-8] [Medline: 40713579]
58. Ning Y, Teixayavong S, Shang Y, et al. Generative artificial intelligence and ethical considerations in health care: a scoping review and ethics checklist. Lancet Digit Health 2024 Nov;6(11):e848-e856. [doi: 10.1016/S2589-7500(24)00143-2] [Medline: 39294061]
59. Zhou W, Zhu X, Han QL, et al. The security of using large language models: a survey with emphasis on ChatGPT. IEEE/CAA J Autom Sinica 2025;12(1):1-26. [doi: 10.1109/JAS.2024.124983]
60. Arora A, Wagner SK, Carpenter R, Jena R, Keane PA. The urgent need to accelerate synthetic data privacy frameworks for medical research. Lancet Digit Health 2025 Feb;7(2):e157-e160. [doi: 10.1016/S2589-7500(24)00196-1] [Medline: 39603900]
61. Costa ICP, Nascimento MCD, Treviso P, et al. Using the chat generative pre-trained transformer in academic writing in health: a scoping review. Rev Lat Am Enfermagem 2024;32:e4194. [doi: 10.1590/1518-8345.7133.4194] [Medline: 38922265]

Abbreviations

AI: artificial intelligence
AIGC: artificial intelligence–generated content
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by A Stone; submitted 22.May.2025; peer-reviewed by F Fukuzawa, K Huang; revised version received 12.Feb.2026; accepted 12.Feb.2026; published 01.May.2026.

Please cite as:
Wang C, Wang L, Zhang X, Qi H
The Influencing Factors of Medical Postgraduates' Usage Intention Toward Artificial Intelligence–Generated Content Tools in Academic Research: Qualitative Analysis
J Med Internet Res 2026;28:e77928
URL: https://www.jmir.org/2026/1/e77928
doi:10.2196/77928

XSL•FO
RenderX

© Chen Wang, Liu Wang, Xuejiao Zhang, Huiying Qi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 1.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Opportunities and Concerns of Gamified, Extended Reality for Home-Based Motor Rehabilitation for Children With Brain Injury: Qualitative Case Study on Design Elements Related to the Engagement and Fatigue Perspectives

Eivind Kolstad[1], MSc; Nenad Pavel[1], PhD; Alexis Ken Sosmena Cartajenas[1], MSc; Trust Saidi[1], PhD; Ingvild Kristina Hurum Rosseland[2], MSc; Åse Bergheim[2], PT, MSc; Nora Synnøve Mørk[2], OT, MSc; Kathinka Granum Selmer-Olsen[2], OT, MSc; Parisa Gazerani[1], MD, PhD; Shefaly Shorey[3], PhD; Minna Pikkarainen[1,3], MBA, PhD

[1]Department of Product Design, Technology Art and Design, Oslomet – Oslo Metropolitan University, Oslo, Oslo, Norway

[2]Oslo Municipality, Oslo, Oslo, Norway

[3]Alice Lee Centre for Nursing Studies, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore, Singapore

Corresponding Author:
Minna Pikkarainen, MBA, PhD
Department of Product Design
Technology Art and Design
Oslomet – Oslo Metropolitan University
Postboks 4 St. Olavs plass
Oslo, Oslo, 0130
Norway
Phone: 47 0407539880
Email: minna.pikkarainen@oslomet.no

Abstract

Background: Acquired brain injuries are injuries that occur after birth and are a leading cause of long-term disability and death in children and young adults. They may result from trauma, hypoxia, stroke, infection, or a variety of other causes. Fatigue is one of the most common and underrecognized consequences of pediatric acquired brain injury, often expressed behaviorally rather than verbally. Traditional rehabilitation programs are frequently static and cognitively demanding, limiting engagement and therapeutic outcomes. Extended reality (XR) technologies offer new opportunities to address these challenges by enabling interactive, adaptive, and motivating home therapy environments. However, few XR systems are co-developed with children and therapists, and there is limited knowledge about how to co-design engaging, gamified, XR-based motor rehabilitation solutions that take into account children's fatigue.

Objective: This study explores why specific gamification and XR design elements facilitate or hinder engagement and effective fatigue response during rehabilitation for children with acquired brain injuries.

Methods: A qualitative case study approach was employed with a total of 25 participants (22 provided consent), combining co-design workshops, interviews with health care professionals, observational data, and iterative user testing with children ages 8-16, their parents, and the clinical team. Participants who provided consent included 4 children with acquired brain injury (ages 8-16), 4 parents, 4 clinicians, and 10 healthy children involved in early ideation only. The XR prototype was developed using Unity, Cognitive3D, and the Meta Quest 3 headset. Engagement and fatigue related to prototype use were evaluated using subjective measures adapted from the User Engagement Scale and the Virtual and Mixed Reality Fatigue Scale, supplemented by thematic analysis of interview and workshop data.

Results: Children demonstrated higher engagement with short, modular XR sessions (3-10 minutes) that included interactive game elements and preserved visibility of their surroundings. Fatigue was identified through behavioral cues such as gaze, posture, and responsiveness. Therapists emphasized the importance of adaptive difficulty, personalization, and simplified environments. A therapist-facing dashboard was developed to visualize behavioral fatigue-related cues, consistent with clinician observation.

Conclusions: This qualitative case study provides preliminary insights into how XR-based rehabilitation, when co-designed with children and clinicians, may support engagement and facilitate observation of fatigue-related behaviors in pediatric brain injury contexts. The findings provide design-oriented insights for creating engaging XR home-rehabilitation experiences while

XSL•FO
RenderX

accounting for fatigue-related limitations (eg, short, modular sessions; visual grounding; adaptable challenge). The results of our study indicate a need for objective fatigue measurement within XR solutions to adjust content in ways that support both engagement and fatigue considerations. However, fatigue detection was not validated in this study and should be addressed in future research.

(J Med Internet Res 2026;28:e84013) doi:10.2196/84013

KEYWORDS

extended reality; brain injury; rehabilitation; engagement; fatigue

Introduction

Pediatric acquired brain injury (ABI) is a leading cause of long-term disability among children and adolescents, affecting cognitive, motor, sensory, and emotional domains [1,2]. In Norway alone, thousands of children experience such injuries annually due to trauma, illness, or medical conditions [3]. Data from the Oslo region show that approximately 29 out of every 100,000 children are hospitalized each year due to traumatic brain injuries of varying severity [4]. ABIs are insults that occur after birth and are a leading cause of long-term disability in children and young adults. They stem from trauma, hypoxia, stroke, infection, and other factors, producing varied deficits in cognition, behavior, metabolism, motor control, perceptual-motor skills, and sensation. Many ABIs cause hemiplegia—one-sided weakness or paralysis—compromising upper-limb function [5]. Typical treatment options for (physical) pediatric ABI include standard physical therapy, occupational therapy, and robot-assisted training. Over extended treatment periods, however, many children with ABI may struggle to engage or may discontinue therapy [6].

It has been reported that insufficient motivation is a major barrier to children with ABI reaching their functional potential in rehabilitation [7]. Thus, to improve adherence, more engaging and enjoyable approaches are needed [6]. Exoskeleton-based therapy supports high-repetition, high-intensity rehabilitation practices. When paired with interactive environments such as extended reality (XR), it can increase motivation and functional relevance, with reported gains, for example, in hand and finger coordination in children with cerebral palsy [5].

Emerging research suggests that XR, an umbrella term for virtual reality (VR), augmented reality (AR), and mixed reality, holds promise in addressing challenges in pediatric rehabilitation. XR tools can simulate interactive, playful, and personalized environments that increase therapy engagement, particularly through gamification [8]. Although there is not yet much research evidence on the effectiveness of VR-based pediatric ABI programs [9], and even less on AR-based programs in home settings [10], existing studies have shown that XR can support motor function recovery, improve attention, and sustain motivation when integrated into goal-directed therapy [11,12]. VR has emerged as a promising and relatively effective approach for rehabilitation of pediatric ABI, with the potential to improve functional status, physical activity, balance, and cognitive and motor skills [6,13-15], whereas AR has been shown to have positive effects on pediatric balance [10]. In home environments in particular, goal-directed VR programs, as a supplement to limited clinic therapy, have been shown to improve physical outcomes in children with ABI by enabling repetitive, intensive practice and sensorimotor feedback to boost motivation and motor learning [9].

Among the most prevalent and debilitating, yet often invisible, symptoms of pediatric ABI is fatigue, which can manifest cognitively, physically, or emotionally and significantly impact a child's ability to engage in rehabilitation, education, and social activities [16]. Despite its prevalence, fatigue remains underreported and challenging to address in clinical contexts, especially among younger children who may lack the awareness or language to express their limits [17,18]. Clinicians and therapists often rely on indirect behavioral cues, such as reduced eye contact, changes in body posture, loss of focus, or disengagement, to identify exhaustion [19]. If not recognized and addressed, fatigue can lead to frustration, dropout from therapy, and diminished rehabilitation outcomes [20]. Concurrently, traditional rehabilitation programs are frequently static and repetitive, failing to meet children's motivational and cognitive needs. This mismatch contributes to low adherence and an increased risk of long-term functional challenges [20]. While Mazzone et al [21] articulate the complexity of fatigue assessment and management in pediatric ABI rehabilitation, our study aims to address a critical gap by demonstrating how XR design can operationalize these clinical processes through fatigue-sensitive interaction, adaptive engagement, and observable behavioral indicators.

Current XR-based rehabilitation solutions for pediatric populations are not designed to respond to fatigue, and it remains unclear how such solutions should be designed to account for the fatigue perspective. Besides, they do not sufficiently consider individual differences in energy levels, motivation, or cognitive capacity [22]. For example, a recent systematic review and meta-analysis examining cognitive fatigue in children following ABI found that nearly three-quarters of impacted individuals (74.6%) experience ongoing cognitive fatigue [23]. Despite its widespread occurrence, the study identified a lack of standardized, evidence-based interventions specifically designed to address this issue within pediatric rehabilitation settings, highlighting the inadequacy of current tools in recognizing and responding to fatigue. There is also limited research on integrating engagement theory with fatigue in the context of XR-based health care design [24], as well as a limited understanding of intuitive interfaces that bridge technical data with clinical applications [25]. Finally, only a few XR rehabilitation solutions are co-designed with children, therapists, and caregivers, limiting their usability, accessibility, and clinical relevance [26].

Thus, there is a lack of knowledge about how to engage and motivate children with pediatric ABI during interactive XR-based motor rehabilitation tasks, and how to design

XSL•FO
RenderX

engaging XR-based motor rehabilitation solutions for ABI that take children's fatigue into account.

To address these gaps, this qualitative case study explores which gamification and XR design elements facilitate or hinder engagement in home-based motor rehabilitation for children with ABI, and how fatigue-related considerations shape design requirements and session structure. This is examined through the following subquestions: (1) How do the clinical teams perceive opportunities and concerns? (2) How are fatigue and engagement perceived by children and clinical teams during XR sessions? (3) How should XR solution design elements be adapted to preserve engagement while accounting for fatigue?

This was conducted as a qualitative case study [27] involving a total of 25 participants (22 provided consent), including healthy children (to help ideate the initial concept), members of the clinical team, children with ABI, and their parents in co-design workshops and individual interviews.

Methods

Research Design Overview

This study adopts a qualitative case study approach to explore how XR technologies can support the detection of fatigue-related symptoms in children with brain injuries. As defined by Yin [27], case studies investigate contemporary phenomena in real-life contexts, especially when the boundaries between the phenomenon and the context are not clearly evident. This method is well-suited for addressing "how" and "why" questions, such as how specific gamification and XR design elements facilitate or hinder engagement and effective fatigue response during at-home rehabilitation for a child with a brain injury. A single qualitative case study was selected because it enables a deep, contextual understanding of how a child's engagement with XR motor rehabilitation fluctuates with fatigue, and what type of content supports engagement during use of the solution. Direct feedback from children immediately after XR solution trial sessions, along with observations and discussions with clinical teams and parents, provided rich data to enhance understanding of the phenomena related to the research question. Co-design with the child, parents, and the clinical team translated these insights into tailored, engaging content and fatigue-aware adaptations (eg, dynamic difficulty adjustment, micro-breaks, and session pacing) to optimize usability and adherence. The prototype was developed as a motor rehabilitation support tool intended to increase engagement and provide a structured, safe, home-like training experience. Although fatigue detection was discussed during the workshops, this study did not aim to validate automated fatigue detection. Instead, fatigue was treated as a design constraint, explored through stakeholder feedback and clinician-observed behaviors during short sessions.

The case centers on co-developing an XR rehabilitation solution with clinicians (eg, physical and occupational therapists), children with ABIs, and their parents. Data collection involved 25 participants, including members of the clinical rehabilitation team, children with ABI, and their parents (Figure 1). Multiple data sources—including workshops, interviews, observations, and the Engagement Scale and Virtual and Mixed Reality Fatigue Scale (VMRFS) surveys (see Multimedia Appendix 1)—provided deep insights into user needs and therapeutic impact. Following Yin's [27] principles—clear case boundaries, multiple sources of evidence, and a maintained chain of evidence—the study used an iterative, reflective process to inform design decisions, ensuring that the solution is engaging, clinically grounded, and responsive to real-world needs. The study results were reported in accordance with Journal Article Reporting Standards [28].

Figure 1. Timeline illustrating the sequence of data collection activities.

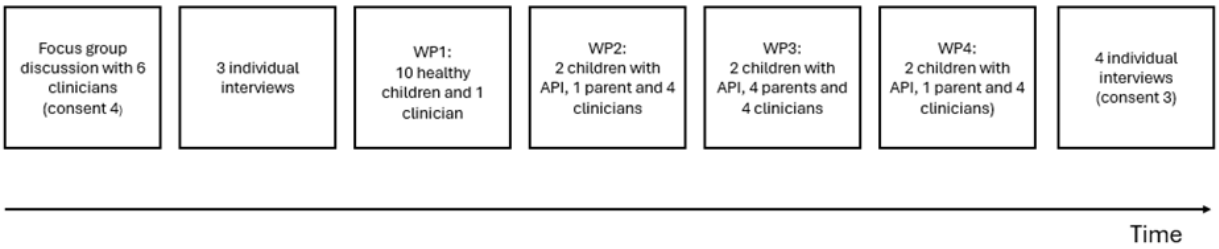

Researcher Description

The research team combined complementary expertise: EK led data collection and analysis, supported by NP's contributions to study design and workshop facilitation, and AKSC's assistance with data collection. Clinical partners IKHR, ÅB, NSM, and KGSO contributed to patient recruitment, provided clinical insight, and reviewed the manuscript. Senior academics—PG, SS, and the lead supervisor, MP—strengthened the methodological and scientific quality of the study, with MP overseeing study conception, ethics, project coordination, and consolidation of the manuscript for publication.

Study Participants, Participant Recruitment, and Data Sources

Participants were recruited through collaboration with the Norwegian Municipality, Norwegian hospitals, and researcher networks (healthy children) using purposive sampling [29]. XR generally benefits from age-appropriate design; however, in pediatric brain injury rehabilitation, adaptation based on functional capacity and fatigue was prioritized over chronological age in this study. Purposive sampling was used to recruit participants who could provide information-rich insights relevant to the study's aims, specifically (1) children with ABI who represented typical rehabilitation cases in terms

XSL•FO
RenderX

of functional variability and fatigue patterns; (2) clinicians with direct experience treating such children; and (3) healthy children who could contribute to early-stage usability and interaction design. This approach ensured diversity in functional abilities, therapeutic roles, and user perspectives, rather than diversity in demographic characteristics such as chronological age.

A description of participant recruitment is provided in the participant flow diagram (see Multimedia Appendix 2). Overall, 25 individuals—healthy children, clinicians, and children with ABI—were contacted about the study. In total, 22 participants provided consented data: 10 healthy children (ideation phase), 4 children with ABI, 4 parents, and 4 clinicians. Data collection proceeded in 3 stages: (1) an early ideation workshop with healthy children and initial clinician interviews; (2) iterative co-design and playtesting workshops with children with ABI, parents, and clinicians; and (3) follow-up clinician interviews after the final iteration.

The first phase of the study included a workshop with 10 healthy children aged 9-16, followed by a meeting with 6 clinicians and individual interviews with 3 clinicians who had hands-on expertise in the rehabilitation of children with ABI. The first workshop was conducted as part of the OsloMet Design for Experience course. This workshop aimed to gain a general understanding of preferences for rehabilitation games among children in this age group. In total, 10 healthy children aged 9-16 participated in the workshop. Six different design student groups (4-6 students in each group) created paper prototypes of potential rehabilitation games. The children were asked to evaluate these early concepts and provide open feedback. During the sessions, the research team conducted observations and took notes on the children's excitement and reactions to the different design concepts. These observations served as a baseline for selecting 3 initial design concepts for further discussion with the clinical team.

The next step in the research process was to conduct a focus group discussion with clinicians working with children with ABI to better understand the rehabilitation status and needs of this population, as well as to gather feedback on their preferences regarding the preliminary design ideas and generate initial concepts for the XR solution. However, only 4 of the 6 clinicians who participated in the focus group discussion and 2 of the 3 clinicians who participated in the individual interviews provided consent for data usage. Consequently, only the consented data were included in the formal analysis. Data from the remaining 4 clinicians were used as background input to inform the development of the initial XR design concepts, but were not included in the analyzed dataset.

The next phase of the study (4 workshops) included 4 children with ABI (3 participated with their parents and 1 with an assistant), aged 9-16, and clinicians (including occupational and physical therapists) with rehabilitation experience. The focused objective of the workshops was to explore stakeholder perceptions (specifically concerns and opportunities related to XR-based rehabilitation), understand how fatigue and engagement were experienced during and after XR sessions, and identify how the XR solution should be adapted to preserve engagement while accounting for fatigue. In this study, fatigue-related cues referred to clinician-observed, naturally occurring behavioral changes during interaction, such as reduced responsiveness, slowed reaction time, reduced movement amplitude, changes in posture, gaze disengagement, or an increased need for prompts. These cues were used to inform interpretation of engagement and to guide fatigue-aware design decisions (eg, session length, breaks, and sensory load). Observations were documented through field notes, workshop video review, and clinician reflections. Fatigue was not induced; sessions were kept short and supervised by the clinical team, and activities were paused if clinicians or caregivers determined that the child needed a break. No specialized technical training was required to identify these cues; however, clinicians' familiarity with pediatric ABI strengthened the contextual interpretation.

The solution was intended to be engaging while supporting motor rehabilitation and facilitating responses to fatigue-related symptoms in children with brain injuries. After the fourth workshop, 3 additional individual interviews were conducted with the same clinical team members who had participated in all workshops. The purpose of these interviews was to gain a deeper understanding of their perspectives on opportunities and concerns, fatigue and engagement during XR sessions, and XR design elements needed to preserve treatment goals while accounting for fatigue.

The study prioritized the collection of actionable insights to guide iterative refinement of XR prototypes rather than aiming for exhaustive thematic saturation. Accordingly, a small, purposefully selected sample was considered methodologically appropriate. Clinicians were selected based on their direct practical experience in pediatric ABI rehabilitation, including occupational and physical therapists and a pediatric neurologist involved in home- and outpatient-based care. Children with ABI were purposively selected to reflect diversity in age (8-16 years), type and cause of injury (eg, stroke, traumatic brain injury), functional abilities, and observed fatigue characteristics, while still being able to safely participate in XR-supported motor activities. Parents were included, when appropriate, to provide complementary perspectives on fatigue and engagement in the home rehabilitation context. In the early ideation phase, healthy children were intentionally recruited as proxy users to explore general gameplay preferences and interaction concepts before involving children with ABI, thereby reducing participant burden and informing initial design decisions. This sampling strategy supported the study's iterative co-design process and the development of fatigue-sensitive XR design insights.

The approach aligns with rapid prototyping methodologies, which emphasize the value of targeted feedback to inform design decisions rather than comprehensive theme development [30]. The study also adhered to participatory research principles by engaging a representative sample of end users, including clinicians whose perspectives are critical to the intervention's relevance and usability. Furthermore, the total sample size (10 healthy children, 4 children with ABI, 4 parents, and 7 clinicians, comprising a total of 25 participants) falls within the range recommended by Creswell and Miller [31] for phenomenological inquiry, which suggests between 5 and 25 participants. Data collection involved workshops, interviews,

XSL•FO
RenderX

observations, and surveys to support triangulation and ensure rich, contextual insight into the design, use, and evaluation of the XR rehabilitation prototype, as shown in Table 1.

Table 1. Overview of data types, sources, and analytic uses in a mixed methods co-design study developing and evaluating an extended reality therapeutic prototype for pediatric rehabilitation.

Data type	Sources	Use in the study	Quality[a]
Therapist/clinician interviews	Interviews	Identify clinical challenges, design needs, and real-world context	+++
Observations	Workshops	Capture behavior, verbal cues, and group dynamics during co-design and testing	++
Field notes	Workshops	Supplement observations with informal insights, environmental context, and tone	++
Feedback from patients	Questionnaires	Validate design usability and comfort; gather children's preferences	+++
Feedback sessions with the clinical team	Interviews/workshops	Insight to build the prototype further	+++
Virtual and Mixed Reality Fatigue Scale	Postsession questionnaires	Assess physical, cognitive, and visual fatigue after extended reality use	+++
User Engagement Scale	Postsession questionnaires	Measure perceived engagement: attention, usability, and reward	+++
Prototype interaction videos	Workshop video recordings	Analyze interaction patterns, visual behavior, and fatigue indicators	++
Feedback sessions	Interviews/workshops	Collect insights from therapists on prototype refinement and clinical relevance	+++

[a]"Quality" indicates the research team's perceived utility of each data source for answering the research questions (+++ high; ++ moderate), based on triangulation potential and richness of the data.

Semistructured interviews were conducted with 5 clinicians—physical therapists, occupational therapists, and a pediatric neurologist—experienced in pediatric brain injury rehabilitation. These clinicians provided insights into the behavioral presentation of fatigue, motivational challenges, and design features to enhance engagement and outcomes. The interviews (guide in Multimedia Appendix 3), lasting 45-90 minutes, followed a structured protocol covering rehabilitation practices, XR technology, and digital tools. Audio recordings were transcribed and thematically analyzed to inform XR prototype design. Co-design workshops were conducted with 2 key user groups: health care professionals and children with ABIs (or, in the first workshop, healthy children as proxies), as shown in Table 2, which summarizes their profiles, including age, diagnosis, and session involvement. "Quality" indicates the research team's perceived utility of each data source for answering the research questions (+++ high; ++ moderate), based on triangulation potential and richness of the data.

Table 2. Overview of child and parent participants involved in the co-design workshops, detailing each child's type of brain injury, age, and participation across workshop sessions.

Child ID	Type of brain injury	Age (years)	Workshop participation
Child 1	None	10	1, 3, and 4
Child 2	While born	8	3 and 4
Child 3	Stroke	16	2 and 3
Child 4	Traumatic brain injury at a young age	11	2 and 4

The workshops had dual purposes: to collaboratively generate design ideas (early phase) and to collect feedback on different versions of the XR prototype. Co-design principles guided the development of the XR prototype through iterative workshops with children, parents, and therapists (see Table 3). Their contributions informed key design decisions, such as the inclusion of adaptive challenge mechanics, consideration of fatigue situations, natural rest prompts, and personalized rewards based on user interests.

XSL•FO
RenderX

Table 3. Overview of participant groups involved in each co-design workshop, summarizing the number of researchers, children with brain injuries (and children without injuries for workshop 1), parents, and clinicians who provided consent.

Stakeholder	Workshop 1	Workshop 2	Workshop 3	Workshop 4
Researchers	4	4	4	4
Children with brain injury	0 (10 children without brain injuries)	2	2	2
Parents	0	1	4	1
Clinicians		4	4	4

Each workshop session lasted approximately 120-180 minutes and included guided playtesting with the prototype, participant feedback through discussions and questionnaires, and follow-up discussions with the clinical team. The sessions were audio- and video-recorded, and postsession feedback was also documented through informal discussions and field notes. These workshops enabled iterative integration of feedback and supported the co-creation of features addressing user fatigue, safety, and engagement. Four children with varying types of ABIs participated across different workshops. Observations captured user interaction, task navigation, responses to feedback, and fatigue-related behaviors (eg, posture changes, slowed reaction time, disengagement). A structured protocol guided interviews conducted with the children after each session, supporting triangulation with other data sources. Participants who took part in the interviews and provided consent are shown in Table 4. To assess engagement and fatigue, 2 instruments were used: a simplified User Engagement Scale (UES) [32] and an adapted VMRFS [33]. The UES measured attention, usability, aesthetics, and reward, adjusted for pediatric cognitive levels (see Multimedia Appendix 1 for the scales used). The VMRFS assessed visual discomfort, cognitive strain, and physical fatigue. Both instruments were used qualitatively: each child with ABI was narratively guided through rating the prototype using the UES and VMRFS immediately after the session. These discussions were audio-recorded, transcribed, and analyzed alongside other qualitative data to identify patterns in motivation, usability, and fatigue. The instruments were administered in a simplified, interview-supported format appropriate for children. Items were used to prompt reflection on comfort, strain, and engagement and were analyzed qualitatively alongside interviews and observations, rather than as standalone quantitative outcome measures.

Table 4. Overview of interview participants summarizing their professional roles and duration of their individual interviews, conducted with informed consent.

Interviewed person	Role	Duration of interview (minutes)
Interviewee 1	Occupational therapist	50
Interviewee 2	Physiotherapist	60
Interviewee 3	Pediatric neurologist	50

Data Analysis and Methodological Integrity

Data analysis followed Braun and Clarke's [34] 6-phase thematic analysis framework. The process began with familiarization through transcription and repeated readings of the data. Initial codes were generated to capture key features related to fatigue, motivation, and engagement. These codes were then organized into broader themes, which were reviewed and refined to ensure coherence and accuracy. The final phase involved producing the report, supported by illustrative quotes and linked to relevant theoretical frameworks, including the UES [32] and Cognitive Load Theory [35]. Analysis was conducted first for the individual interviews and then iteratively extended to the workshop data. Given the small sample size and the risk of deductive identification, quotes are attributed by stakeholder group only.

This study ensured methodological integrity by drawing on diverse data sources—including interviews, co-design workshops, observations, interaction videos, and postsession engagement and fatigue scales—that captured the perspectives of children with ABI, parents, clinicians, and healthy children relevant to the study aims. Researcher influence was managed through reflexive practices such as memo writing, field notes, and supervisory discussions, and interviews were conducted using open-ended prompts to minimize bias. Reflective thematic analysis grounded the findings in the data, supported by direct quotes, behavioral descriptions, and triangulation across sources. Apparent contradictions—such as low self-reported fatigue despite observable fatigue cues—were integrated into the analysis rather than treated as inconsistencies. Contextual details about participants and settings enhanced transparency, and iterative prototype refinement functioned as a form of member validation, demonstrating the practical utility of the findings. Analytic consistency was maintained through collaborative coding and discussion, supported by reflexive notes and repeated engagement with the data. This approach ensured that the claims are well-supported, meaningful, and relevant to the design of fatigue-sensitive XR rehabilitation tools for pediatric ABI.

Ethical Considerations

Approval from the Regional Committee for Medical and Health Research Ethics in Norway was deemed unnecessary because the study focused on health service research and therefore did not fall under the Health Research Act § 2. Instead, approval was obtained from the Norwegian Agency for Shared Services in Education and Research (SIKT; reference number 871403) to ensure data protection and privacy compliance. Participants

XSL•FO
RenderX

received both oral and written information about the study, and informed consent was obtained before conducting the workshops and individual interviews. To protect participant privacy, all data were deidentified through the use of pseudonyms. Given the vulnerability of the study population, fatigue was not deliberately induced. Observation of fatigue cues was limited to naturally occurring behaviors during short, supervised XR sessions conducted in collaboration with the experienced clinical rehabilitation team.

All participants—children, parents/guardians, clinicians, and healthy children involved in the early ideation phase—received written and oral information about the study, and parents/guardians provided informed consent before participation. Participants received gift cards as compensation for their efforts. Study data were deidentified using pseudonyms, and all audio files, transcripts, and workshop materials were securely stored on OsloMet's research server with access restricted to the research team, ensuring confidentiality and compliance with GDPR (General Data Protection Regulation). No identifiable images or videos of participants are included in the manuscript or its multimedia appendices.

Results

Themes Identified

The data analysis revealed 4 main themes: fatigue detection challenges; engagement and motivation; adaptive therapy design; and clinical utility, therapist needs, and child-centered co-design. These themes were further divided into 16 subcategories (Multimedia Appendix 4).

Clinical Team Concerns and Opportunities for XR-Based Home Rehabilitation of Children With API (Based on Individual and Focus Group Interviews)

Fatigue As a Central Barrier to Rehabilitation

All clinicians unanimously emphasized fatigue—both cognitive and physical—as a pervasive yet often underrecognized barrier in therapy:

> *Children rarely verbalize tiredness directly; instead, signs emerge behaviorally through irritability, lack of engagement, shortened attention spans, or reduced physical coordination.*

Clinicians noted that fatigue often masquerades as disinterest, laziness, or poor behavior, making it difficult to detect without strong contextual knowledge of the child. As one clinician explained, "Children may try to keep up even when they don't have the capacity for it, which can be mistaken for behavioral problems." Another therapist noted, "Children pull the elastic further than adults would. They often don't know their limits and just try to keep going," highlighting the need for therapists to interpret behavioral cues rather than rely solely on self-report.

The Engagement-Fatigue Paradox

When designing AR solutions, it is challenging to determine the appropriate balance between fatigue and engagement. One therapist described the delicate balance between engaging a child through stimulation and avoiding overstimulation that may lead to rapid fatigue or frustration: "You have to push a little, but not too much. You have to dose correctly." A physical therapist emphasized that engaging therapy is critical for maintaining motivation; however, cognitive and physical overload can cause children to shut down. Clinicians highlighted the importance of short, adaptive sessions tailored to the child's energy level and day-to-day condition. One therapist described this dynamic as a "balancing loop," in which excessive stimulation leads to exhaustion, reducing engagement and, consequently, limiting therapeutic effectiveness.

Motivation Through Personalization and Play

Motivational breakdown was described as a key reason children disengage from therapy, particularly after setbacks at school or in social life. Clinicians stressed the importance of making therapy fun, meaningful, and personalized. One physical therapist noted, "Children are more likely to engage when they understand the purpose of the activity and can experience mastery." Games that incorporate levels, points, rewards, or themes aligned with children's interests (eg, dinosaurs, magic, sports) were viewed as highly beneficial. "They train better when they forget they are training," one therapist observed, underscoring the value of gamified experiences. Another therapist added, "If it's too challenging, they fall off quickly. It has to be something they feel they can master."

The Role of Adaptive XR Technologies

All clinicians were optimistic about the potential of XR—particularly AR—to support motivation and extend therapy beyond clinical hours. "AR lets them stay aware of their surroundings, which is really important when balance and orientation are affected," one therapist explained. AR was preferred over VR because of its ability to preserve spatial awareness and minimize safety risks for children with balance or visual impairments. Clinicians envisioned XR as a "tool in the toolbox," ideally used to supplement traditional therapy with personalized, game-like activities that adapt in difficulty and provide feedback on progress. Objective data generated through XR (eg, movement patterns, reaction time, engagement levels) were viewed as promising for tracking fatigue, tailoring exercises, and evaluating progress over time:

> *It would be very helpful to get data that shows change, like how much they're using one arm compared to the other.*

Clinical and Ethical Considerations

While enthusiasm for XR was high, several practical concerns were raised. These included overstimulation, the risk of demotivation if exercises are too difficult or too easy, and the importance of involving both therapists and families in guiding the child's use of XR tools. "The content must be adapted and user-friendly, not just for the child, but also for therapists and parents," one therapist emphasized. Safety, particularly in VR environments, was a recurring concern. "If you take away their vision, and they already have balance or sensory problems, it can be really dangerous," warned one physical therapist. Clinicians stressed that XR systems must be adaptable, intuitive, and grounded in therapeutic intent—not merely another game. Together, these insights highlight the need for rehabilitation

XSL•FO
RenderX

technologies that are adaptable, engaging, and clinically grounded. The interviews informed both technical priorities (eg, fatigue monitoring and difficulty scaling) and experiential aspects (eg, personalization, flow states, and emotional safety) that are important for XR system development. As one therapist put it, “It needs to be motivating and challenging enough, but not so hard that they give up. It’s a fine line.”

Implications for XR Design

Overview

The clinical interviews offered direct, practice-based insights into how digital tools can better support pediatric neurorehabilitation, particularly in relation to fatigue, motivation, and therapy adherence. “Short, modular sessions that can scale up or down are essential,” one therapist emphasized.

Short, Adaptive, and Engaging Sessions

Clinicians consistently highlighted the need to limit session length, especially during cognitively demanding or vision-intensive tasks. “Some days they can do 20 minutes, some days only 10,” one therapist explained, reinforcing the need for flexibility and day-to-day adaptation.

Detection of and Adjustment for Fatigue

As fatigue is rarely verbalized by children and may manifest subtly, the design emphasizes the value of objective movement tracking (eg, reduced limb use, slowed reactions, visual disengagement) to help infer fatigue and adapt the activity accordingly—by pausing, modifying difficulty, or prompting breaks. “If the child is zoning out, not following instructions, or slumping, those are signs.”

Clear, Motivating Progress Feedback

To maintain engagement, particularly in children with low frustration tolerance, progress should be immediately visible and meaningful. Gamified elements such as point tracking, stars, or personalized progress indicators were integrated to foster a sense of mastery without overreliance on external rewards. “They like seeing points. They like setting small records for themselves.”

Given the frequent balance and sensory challenges described, AR—in which children maintain visibility of their surroundings—was prioritized over full VR. This approach supports safer movement and reduces disorientation, making it more appropriate for active therapy contexts. “If you remove sight and they already have poor balance, that’s dangerous.” Although the XR prototype did not include fully customizable themes, it incorporated interaction mechanics inspired by children’s preferences—specifically, a shooter-style design based on feedback from earlier workshops. This decision aimed to align gameplay with familiar and motivating interaction styles. The system also introduced scalable difficulty through selectable target heights and adjustable game speed, allowing sessions to be tailored to the child’s physical ability and energy level. This modular structure accommodates day-to-day variability and provides therapists with flexibility in adjusting challenge levels. “Let them choose something that feels familiar and fun to them,” one therapist said, emphasizing the importance of matching activities to the child’s capacity and interests.

Stakeholder Perceptions of Fatigue and Engagement During the Co-Design Workshops and After the XR Sessions

Overview

The design prototype was modified iteratively after each workshop based on feedback from participating stakeholders (see Figure 2).

XSL•FO
RenderX

Figure 2. Overview of the iterative development of the extended reality therapeutic prototype for pediatric rehabilitation.

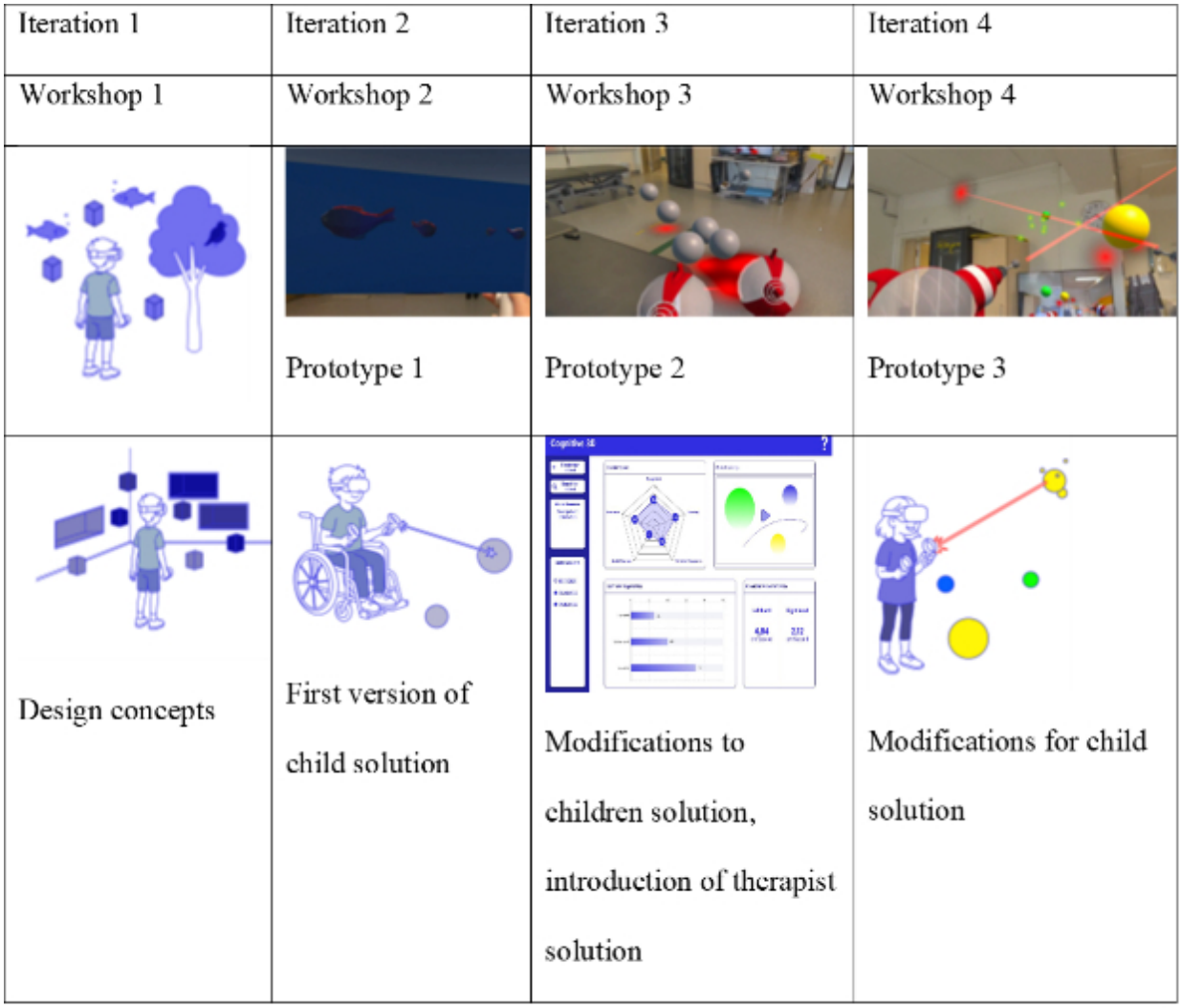

Results of Workshop 1

The forest, fish, and cube-sorting activity concepts (Figure 2) were developed based on observations from the workshop with healthy children, particularly noting their excitement when interacting with paper prototypes created by design students. During the focus group discussion, the clinical team endorsed the need for an AR-based approach, particularly for screening symptoms such as double vision or cognitive fatigue. The forest and fish concepts were praised for their calming and accessible environments, especially suitable for nonverbal or easily fatigued children. The cube-sorting activity was appreciated for its complexity but was noted to require adjustable difficulty settings. Concerns included potential overstimulation and visual fatigue, particularly for children with preexisting vision impairments. Nonetheless, the use of Cognitive3D for postsession analytics was viewed as a promising tool for visualizing gaze and movement patterns. The clinical team emphasized the importance of short, flexible activities aligned with each child's clinical profile.

Results of Workshop 2

The first prototype was designed to ensure low cognitive and physical load, enabling both children to complete the sessions without observable fatigue. However, it was perceived as overly passive. Feedback from the children and video analysis highlighted the need for more interactive gameplay. The underwater setting—particularly the wall texture—impaired spatial awareness. "I could not see my hands in the game, they were behind the blue walls," one child noted, describing feelings of disorientation and discomfort. Engagement scores were low, with children describing the experience as repetitive. "It got boring quick, it would be nice if I could interact with the fish in some way." Although fatigue scores were low, the experience lacked sufficient stimulation to sustain attention and engagement.

Accessibility emerged as a key concern. A child using a wheelchair required assistance to remain oriented within the environment, which reduced independence during gameplay. "The game should be able to be played by children that are in wheelchairs and children that can walk," a therapist emphasized. The clinical team recommended shorter sessions, clearer onboarding procedures, and a simplified user interface (UI) to

XSL•FO
RenderX

enhance usability. Although Cognitive3D provided valuable postsession analytics, its complexity limited feasibility for routine clinical use. Therapists highlighted the importance of appropriate pacing, personalization, and aligning gameplay with individualized therapeutic goals, such as bilateral hand use or visual tracking. Overall, the feedback underscored the need for XR systems that are inclusive, adaptable, and practical for everyday clinical contexts.

Implications for the Design Elements

The first prototype successfully ensured a low cognitive and physical load, which was important for safety. Both children completed the session without signs of strain or fatigue. However, the overall experience was perceived as too passive and insufficiently engaging. Feedback from the children clearly indicated a need for more interactive gameplay—something that encouraged active participation rather than passive observation. The underwater environment, particularly the wall texture, negatively affected spatial awareness. Children reported that they could not see their hands and felt visually disconnected from their surroundings. This contributed to confusion and likely reduced comfort and enjoyment. Accessibility emerged as a significant concern. One child who used a wheelchair required assistance from an adult to rotate and remain oriented during gameplay. This reliance on external assistance made the experience less fluid and diminished the sense of independence. As a result, therapists emphasized the importance of designing XR games that can be comfortably played both while seated and while standing. Interactions should be optimized for a full range of motion within a fixed orientation, reducing the need for physical repositioning or full-body turning. Therapists supported the children's feedback and reinforced the importance of shorter, focused sessions, clearer onboarding instructions, and a simplified interface for accessing behavioral data. While the Cognitive3D tracking system offered promising analytics, its current output was considered too complex for daily therapeutic use without further simplification.

Results of Workshop 3

To improve engagement, the AR environment was redesigned from an underwater theme to a minimalist space that preserved visibility of the surroundings, addressing safety concerns. Workshop 3 introduced dual laser blasters, allowing children to shoot moving orbs in real time, thereby encouraging bilateral arm use and greater physical engagement. Although adaptive difficulty was not implemented, the goal was to validate more engaging mechanics. A therapist-facing dashboard was also developed to streamline monitoring, enabling profile management and session tracking.

Children reported improved usability and engagement. "I liked the game more than the fish one, this was fun." They felt more "in the game," appreciated the intuitive layout, and suggested adding more visual variety. "I would like more colors, the grey balls were boring." Fatigue responses were low, and the 5-10-minute sessions were well tolerated. Therapists observed increased laughter, activity, and focus. "You could clearly see that the children were more engaged now." They praised the dashboard's clarity and usefulness. "This makes it much easier to understand what's going on during the session. It's visual, not just numbers." However, they noted the need for clearer onboarding and session flow. "There should be some kind of menu in the game. Right now, it just keeps going." Therapists also emphasized the importance of adaptive difficulty and personalization. "Now all the balls were low. It would be better if they could also appear near the ceiling." Gamified elements such as scoring were recommended. "Adding points would be a good idea." A presession menu was also suggested. "There should be a menu where they can see their points and choose what game modes they want to play." While XR showed promise for assessments, further validation was needed. Overall, the prototype demonstrated growing clinical relevance and improved usability.

Laser-based mechanics successfully engaged both arms, offering therapeutic value for bilateral motor activity. However, despite increased interaction and interest, the lack of visual variety and feedback limited the depth of engagement over time. While children found the game fun and easier to use, they also expressed a desire for more dynamic and colorful elements to sustain their interest longer. From a clinical perspective, the prototype demonstrated improved engagement without triggering signs of fatigue. Session duration remained appropriate, and therapists observed increased movement, focus, and enthusiasm. However, their feedback underscored the need for continued development in 2 areas: adaptive difficulty to match individual abilities and more intuitive onboarding to support independent use.

Implications for Design Elements

This iteration addressed core issues: adding real-time target shooting and removing disorienting visuals made play more intuitive, stimulating, and physically engaging, thereby improving usability, enjoyment, and willingness to continue. Showing real-world surroundings reduced spatial confusion and overstimulation; children felt grounded and in control, and laser-based mechanics engaged both arms for bilateral practice. However, limited visual variety and feedback constrained sustained engagement; children wanted more dynamic, colorful elements. Clinically, engagement improved without fatigue, with appropriate session length and increased movement, focus, and enthusiasm. Therapists noted the need for adaptive difficulty and more intuitive onboarding to support independent use. Overall, basic gamification and a simplified UI boosted engagement without compromising comfort or safety, while highlighting the need for personalization, clearer guidance, and richer visuals in the next prototype.

Results of Workshop 4

The third prototype introduced several new features designed to enhance therapeutic value, interactivity, and user feedback. To boost engagement, colorful animations and a scoring system were added. A UI counter displayed scores in real time, with yellow, blue, and green balls worth 1, 2, and 3 points, respectively. Children found the game more immersive and intuitive: "I liked the game more than the fish one, this was fun." "I want to play again to beat my score!" "I liked the colors and explosions as I shot the balls." Therapists observed improved focus and enthusiasm. "The score made them want to play again." Short 3-minute sessions allowed breaks. "The

XSL•FO
RenderX

shorter sessions gave them a natural break." The main menu and UI were praised: "With these UI elements this game could be played at home, not just here in therapy sessions." Therapists also suggested customizable themes and a wider range of motion: "This would support more emotionally resonant and individualized therapy experiences." "There should be a menu where they can see their points and choose what game modes they want to play." Thus, children reported markedly higher enjoyment and motivation than in earlier builds, citing the visible score, colorful animations, and varied effects as making it feel like a real game rather than a clinical task. They felt in control, found the interface clearer, and especially liked tracking progress with the score, which encouraged replay ("I want to play again to beat my score!"). They rated immersion, visuals, and fun highly, with low fatigue scores; added interactivity did not cause strain, and they appeared more energized and focused. Therapists observed clear gains in engagement, usability, and therapeutic relevance, attributing improvements to the scoring mechanics, playful animations, and simplified, intuitive UI. Shorter 3-minute bouts provided natural breaks and better pacing, and the main menu plus clear UI elements supported independent use at home: "The score made them want to play again". They recommended personalization via interchangeable themes (eg, planets, animals, fantasy elements) to match individual interests and proposed a mode that spawns targets at 3 heights simultaneously to elicit broader upper-limb movements and build flexibility, strength, and coordination.

Implications to the Design

The last prototype introduced several new features designed to enhance therapeutic value, interactivity, and user feedback. A key addition was the implementation of UI elements that allowed players, or therapists assisting them, to select between different game modes. Specifically, users could now choose where the balls would spawn: at floor level, mid-room, or near the ceiling. This feature was designed to introduce a scalable difficulty system by targeting different physical movement ranges. For example, children with limited shoulder mobility or upper-limb fatigue could practice raising their arms within a controlled, motivating context, supporting rehabilitation of specific motor functions. To improve engagement and visual interest, more colors and animations were added throughout the game. A scoring system was also introduced to provide players with immediate feedback and encourage repeated play. A small, nonintrusive UI counter was placed within the player's field of view, displaying their current score in real time. The gray balls were replaced by yellow, blue, and green balls. Each yellow ball was worth 1 point, blue balls were worth 2, and green balls, being the rarest, earned 3 points. This variation added a layer of strategy and excitement to the gameplay while supporting differentiated motor effort. Each game session was set to last 3 minutes. Upon completion, the final score was displayed, and players were returned to a simple main menu interface, where they could choose to begin another round. These updates were designed to balance therapeutic intent with replayability, providing a flexible and motivating structure for both in-clinic and independent use.

Engagement and Fatigue Across the Workshops

When examined across workshops, children's self-reported responses reveal a converging pattern of increasing engagement alongside consistently low fatigue, supporting the qualitative observations described above and illustrated in summary Figures 3 and 4. Mean engagement scores steadily increased from workshop 2 through workshop 4, reflecting improvements in perceived immersion, control, enjoyment, and motivation as iterative design changes introduced interactive mechanics, simplified visual environments, and gamified feedback. At the same time, mean fatigue-related scores remained low and showed a slight downward trend, indicating that rising engagement was not accompanied by increased physical or cognitive strain. Importantly, children's responses to the session-length question contextualize these findings: in workshops 2 and 3, children indicated that XR sessions should not exceed 5-10 minutes, while in workshop 4, preferences shifted toward even shorter, clearly bounded sessions of approximately 3-5 minutes. This convergence suggests that short, modular exposure durations played a critical role in maintaining comfort while supporting engagement. Together, these findings demonstrate that engagement and fatigue can be decoupled through careful XR design: interactive and emotionally supportive gamification elements can enhance motivation when combined with visually grounded AR environments, simplified interfaces, adaptive pacing, and deliberately short session lengths. The alignment between children's self-reported responses, therapist observations, and iterative design decisions underscores the importance of fatigue-sensitive session structuring as a core principle for pediatric XR-based neurorehabilitation.

XSL•FO
RenderX

Figure 3. Engagement levels reported by children across the co-design workshops. WP: workshop.

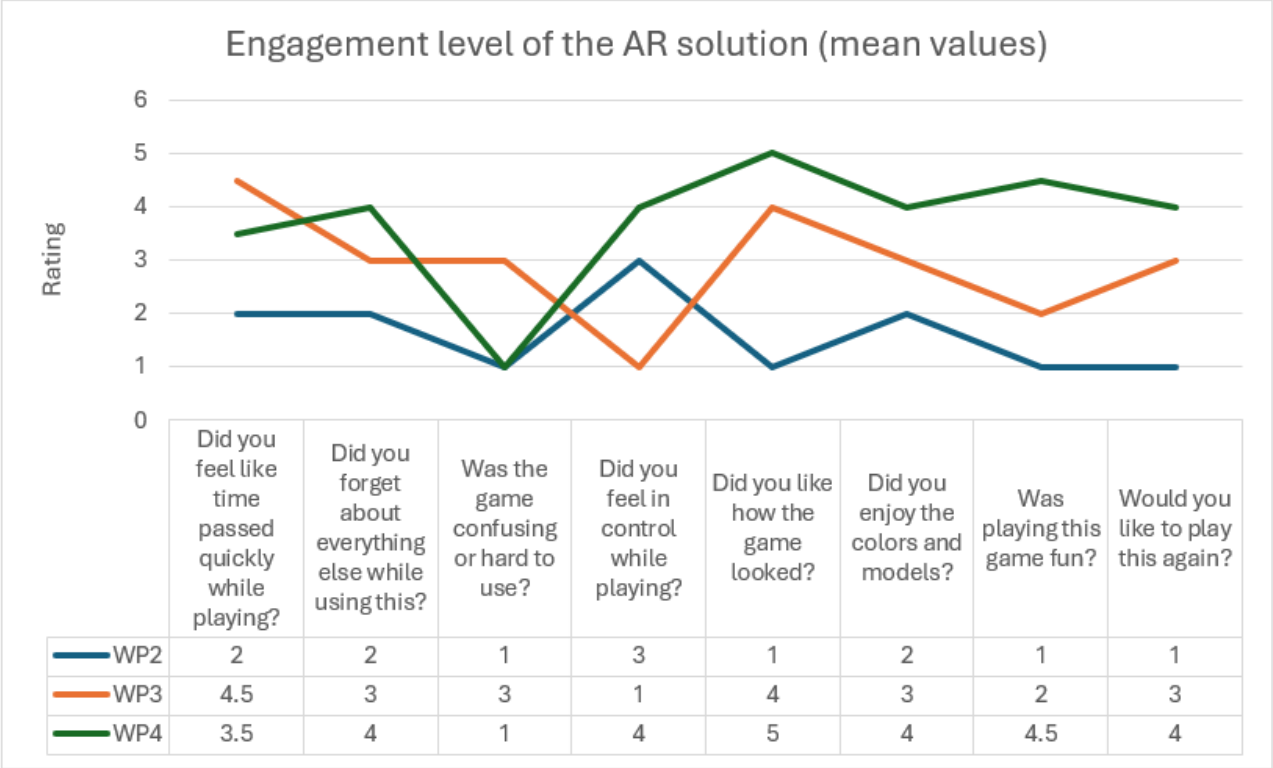

Figure 4. Fatigue scores reported by children across the co-design workshops.

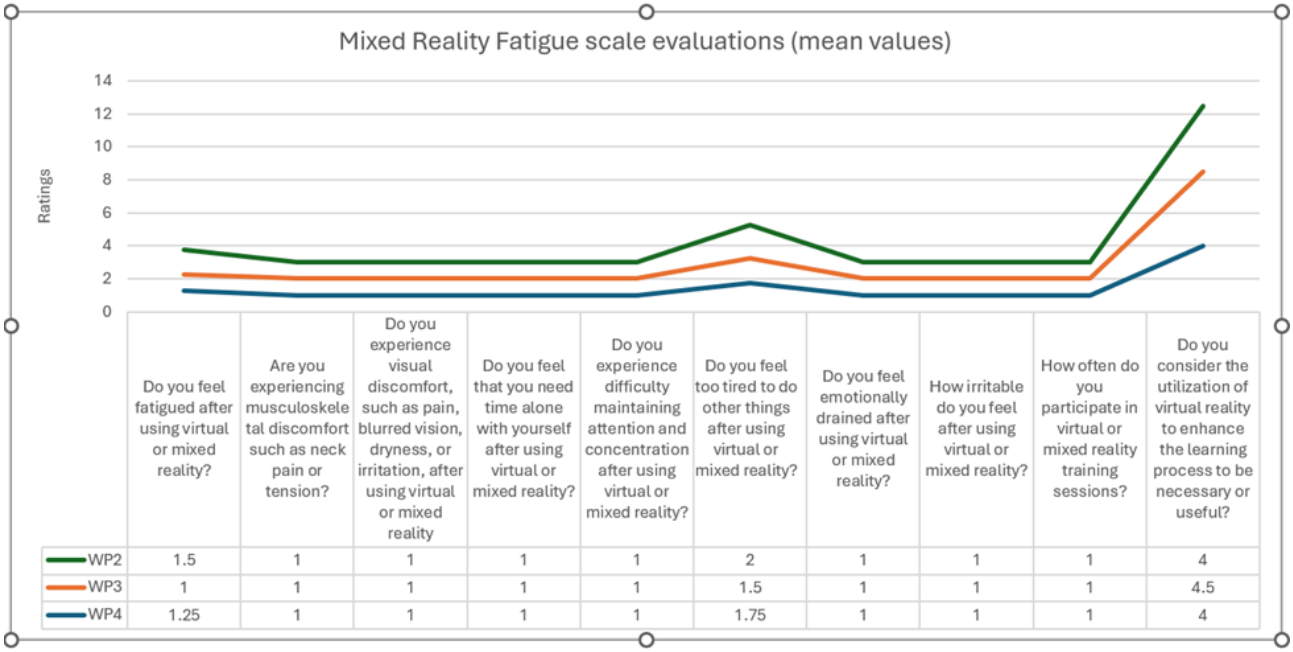

Discussion

Principal Findings

This study demonstrates that XR-based rehabilitation, when co-designed with children and therapists, can support engagement and fatigue-aware use by enabling short sessions, visual grounding, and clinician observation of fatigue-related cues during interaction in pediatric brain injury contexts.

Iterative co-design with children, parents, and clinicians led to critical design adaptations that improved usability, personalized challenge levels, and alignment with therapeutic goals. In our study, fatigue was consistently identified as a central barrier to therapy, often expressed behaviorally rather than verbally. Fatigue emerged as the most pervasive barrier to rehabilitation, typically expressed through subtle behavioral changes rather than direct verbal reporting by children. High engagement enhanced participation but also increased the risk of overstimulation, revealing a delicate balance in which

XSL•FO
RenderX

engagement must be continuously regulated to prevent fatigue-related disengagement. Therapist observations and behavioral data revealed fatigue-related changes in movement, gaze, and responsiveness that were not consistently captured through children's self-reported fatigue ratings.

Clinicians emphasized the need for short, adaptive sessions tailored to fluctuating energy levels. Minimal gamification elements, such as scoring and visual feedback, increased motivation and replay desire when designed to support mastery rather than cognitive or sensory stimulation. Children showed noticeably higher engagement with prototypes that included interactive game mechanics (eg, target shooting) and environments that preserved visibility of their own bodies and surroundings. These features improved usability and reduced disorientation. Short, modular sessions (3-10 minutes) supported fatigue management, allowing children to complete sessions without signs of visual or physical fatigue, while therapists observed increased focus and participation. Gamification was effective when purposeful and balanced. Simple scoring systems, visual effects, and immediate feedback increased motivation without overstimulation. Customization and adaptability were also crucial; engagement improved when children could influence the experience (eg, selecting target heights). Therapists suggested further personalization to align with individual abilities and interests. Fatigue could be inferred through behavioral cues. While subjective fatigue scores were low, therapist observations and Cognitive3D tracking data revealed changes in movement, visual focus, and responsiveness as indicators of cognitive strain. Finally, AR-based designs that preserved visibility of the child's body and surroundings reduced disorientation and supported both engagement and fatigue management, compared with immersive VR environments. Together, these findings highlight the importance of simplicity, interactivity, visual grounding, and short bursts of gameplay in XR design for pediatric neurorehabilitation.

Comparison With Prior Work

Early case-based research using VR for pediatric ABI rehabilitation has revealed both the potential and concerns related to immersive game-based rehabilitation solutions. One of the main concerns is that, while systems such as the Xbox Kinect could improve balance performance, their cognitive, sensory, and motor demands were often poorly matched to children early in rehabilitation, leading to overstimulation, fluctuating performance, and variable motivation [36]. Our findings build directly on these observations by showing how fatigue becomes visible through behavioral changes during XR interaction and by demonstrating how simplified, visually grounded, and adaptive XR designs can preserve both engagement and safety. Our results suggest that effective XR rehabilitation for children with brain injury requires not only motivating content but also fatigue-sensitive system design that aligns with the child's moment-to-moment capacity.

Previous research has demonstrated that fatigue is strongly associated with reduced participation and diminished quality of life in adolescents and young adults with ABI [37,38]. In particular, it has been shown that higher fatigue corresponds to greater participation restrictions, highlighting fatigue as a critical therapeutic target [38]. Building on this work, our study extends the field by examining how fatigue manifests during rehabilitation activities themselves, revealing that fatigue emerges through moment-to-moment behavioral changes rather than solely through retrospective self-report. This study advances the design of XR-based rehabilitation tools for children with ABIs by foregrounding fatigue as a central design constraint—an area often underemphasized in prior work.

Consistent with Rohrer-Baumgartner et al [3] and Mazzone et al [21], our findings confirm that fatigue in pediatric brain injury is multifaceted—cognitive, physical, emotional, and sensory—and often behaviorally expressed rather than verbally reported. Our results show that XR-based interaction enables real-time observation of fatigue-related behaviors during task engagement, offering a concrete way to surface these otherwise hidden indicators. In line with Mazzone et al [21], whose findings indicate that clinicians depend heavily on behavioral cues rather than child self-report to detect fatigue, our results suggest that XR-based behavioral markers, such as gaze, posture, and response latency, can support this observational process in a structured manner.

In line with Mazzone et al [21], whose findings indicate that clinicians often rely on behavioral cues rather than child self-report to identify fatigue following pediatric brain injury, this study demonstrates how XR-based rehabilitation can support this observational process in a structured yet clinically aligned manner. By capturing changes in gaze, posture, movement, and responsiveness during short, adaptive XR sessions, the system allowed fatigue to become visible without requiring children to articulate their internal state. Importantly, these behavioral indicators did not replace clinical judgment but supported shared understanding and collaborative decision-making among therapists, children, and families. Together, these findings position XR not as a fatigue-reducing technology per se, but as a fatigue-sensitive design approach that operationalizes existing clinical practices while preserving engagement and emotional safety. Clinicians also need to continually adjust rehabilitation demands to manage fatigue [21]. Our findings extend this work by illustrating how adaptive XR session length and task difficulty can embed this clinical reasoning directly into the intervention design.

While Wender et al [20] emphasize the need for adaptive systems to prevent disengagement, few studies operationalize this need through concrete design strategies. Our work builds on this by implementing modular, short-duration sessions (3-10 minutes) and adaptive difficulty scaling, which allowed children to complete tasks without signs of visual or cognitive overload. While Proschowsky et al [37] describe gradual changes in fatigue across the rehabilitation period using standardized measures, our results suggest that clinically relevant fatigue fluctuations occur within sessions and are primarily observable through changes in interaction behavior, underscoring the need for complementary real-time monitoring approaches. Our approach also aligns with Sweller's [35] Cognitive Load Theory, which advocates minimizing extraneous load to preserve mental resources—particularly critical in pediatric neurorehabilitation.

XSL•FO
RenderX

Proschowsky et al [37] also demonstrated a consistent discrepancy between child and parent fatigue ratings, particularly for cognitive fatigue, suggesting that children may lack insight into their own fatigue experience. Our findings extend this work by showing that XR-based behavioral indicators—such as gaze disengagement, postural changes, and response latency—can provide an additional, nonverbal layer of information that aligns with clinicians' observational fatigue assessments. In contrast to XR systems with fixed difficulty levels (eg, [11]), our prototype dynamically adjusted the challenge based on fatigue cues, supporting the "just-right challenge" principle from occupational therapy [39]. This approach not only sustained engagement but also improved therapeutic focus, as observed by clinicians. The use of AR over VR—supported by Rebenitsch and Owen [40]—further reduced disorientation and preserved environmental awareness, which is especially important for children with balance or visual sensitivities. Gamification was another key differentiator. While Deterding et al [41] and O'Brien et al [32] highlight the motivational benefits of game elements, our study extends this by emphasizing adaptive gamification. Rather than static rewards or competitive mechanics [42], we prioritized emotionally supportive feedback, autonomy, and visual grounding. Children responded positively to features such as target selection, score feedback, and interactive animations, confirming that gamification enhances motivation when developmentally appropriate and cognitively balanced [43].

Limitations and Future Work

As a qualitative case study, this research is context-specific and not intended for statistical generalization [27,29]. The small participant sample, drawn from a single clinical setting, limits transferability. While including healthy children in early prototyping aligns with child-centered design practices [44,45], it may have introduced discrepancies between early feedback and the needs of children with ABIs. This study used Design Thinking [46] within a qualitative case study to guide XR development for children with brain injuries. Its 5 phases—empathize, define, ideate, prototype [47,48], and test—structured co-design workshops and ensured stakeholder insights shaped solutions focused on fatigue detection, adaptive engagement, and emotional safety. In the empathize phase [49], interviews, observations, and shadowing sessions with clinicians, parents, and children were conducted to understand clinical routines, fatigue triggers, and children's emotional responses to therapy. In the define phase [47], these insights were synthesized into problem statements, such as the need for understanding fatigue and safeguarding mechanisms to prevent overstimulation. During the ideate phase [50], multidisciplinary workshops with clinicians, designers, and children generated concepts ranging from environmental cues for fatigue to customizable levels of XR immersion. In the prototyping phase [47], low- and mid-fidelity XR mockups were created and iteratively refined based on stakeholder feedback, focusing on interaction patterns, visual clarity, and observational insights. Finally, in the test phase [51], preliminary XR prototypes were trialed in simulated therapy sessions, where children and clinicians evaluated usability, comfort, and safety, directly informing adjustments such as simplified interfaces and adaptive pacing.

This approach was seen as particularly valuable in pediatric rehabilitation, where the wide variation in children's physical, emotional, and cognitive abilities calls for personalized and context-sensitive design [52]. Co-design is a participatory design approach that involves end users and stakeholders as active contributors in the design process, rather than passive recipients of a finished product. According to Sanders and Stappers [53], co-design centers on collaboration, mutual learning, and shared ownership, enabling users to bring their experiences, needs, and creativity into the creation of meaningful solutions. In our study, Design Thinking was used as an organizing framework to structure the co-design and iterative prototyping process, ensuring that stakeholder insights were translated into concrete, testable design changes [54]. Across empathize, define, ideate, prototype, and test, the focus was to reduce participant burden, capture fatigue-aware requirements, and iteratively refine engagement-related mechanics in short XR sessions. The choice of a case study combined with co-design is justified by the exploratory nature of the research. Together, design thinking, co-design, and case study provide a robust methodology for integrating multiple stakeholder perspectives while iteratively shaping and evaluating the prototype within its intended context of use. From a process perspective, the Design Thinking structure strengthened traceability between stakeholder needs and design outcomes [55]: the empathize and define phases foregrounded fatigue-aware constraints (session duration, sensory load, safety), ideation clarified interaction preferences, and prototype/test cycles enabled rapid validation of engagement mechanics (eg, interactive shooting and scoring) while maintaining visual grounding through AR. This helped balance engagement with the risk of overstimulation, which clinicians described as a central tension in pediatric ABI rehabilitation. Design Thinking supported rapid iteration and stakeholder alignment; however, the iterative nature also meant that not all clinician suggestions could be implemented and retested within the study time frame. Future work should evaluate the refined design elements across more sites and over longer home-use periods. Feedback from the clinical team, though rich, was collected after the final prototype iteration and therefore could not be integrated, tested, or clinically validated. Future work should also explore personalization options (eg, themes, interaction styles) and multilevel interaction design to support both fine and gross motor skills. Prior research emphasizes the importance of individualized therapy content for sustaining motivation and improving outcomes [11,26]. Although the initial prototype dashboard was intended to visualize behavioral fatigue indicators over the long term (eg, gaze, posture), this was not implemented in the present design phase, and the system also lacked physiological data integration. Future systems could integrate XR-compatible wearables to capture biometric signals, such as heart rate variability or pupil dilation, enhancing fatigue detection [56,57]. The dashboard could be further developed as a fatigue detection tool or as a clinical decision-support tool with customizable visualizations, engagement tracking, and real-time alerts [25]. While this study focused on behavioral indicators of fatigue observable during XR interaction, future research may explore the potential role of additional

XSL•FO
RenderX

physiological signals, such as heart rate variability or pupil dilation, as complementary sources of information.

Although fatigue detection was not conducted in this study, it is important to note that evidence supporting the validity and reliability of these biometric measures for fatigue detection in pediatric ABI populations is currently limited [37]. Our study revealed a need for objective fatigue detection methods that could support safe personalization of XR solutions. Existing research on biometric fatigue markers, however, stems largely from adult neurological populations or from studies of cognitive load and fatigue in other clinical or nonclinical contexts [58]. As such, any integration of physiological signals into pediatric XR rehabilitation systems should be considered exploratory and would require careful validation, ethical consideration, and age-appropriate interpretation before clinical use. To improve generalizability, future studies should include more diverse participants, test across multiple sites, and assess long-term therapeutic impact. Integration with clinical data systems may also be necessary for sustainable implementation in routine care.

Acknowledgments

We thank all participants, especially families of children with brain injury and the clinical teams, for their participation and for making this study possible. Generative artificial intelligence tools (Microsoft Copilot/OpenAI models) were used only to refine author-written text—improving clarity, tone, and organization—while all scientific ideas, analyses, and interpretations were developed solely by the authors. All artificial intelligence–assisted suggestions were reviewed and verified by the authors.

Data Availability

The raw data cannot be made publicly available due to the involvement of a vulnerable population. It is important that the data remain unpublished to prevent potential identification of the children involved in the study.

Funding

The work was conducted as part of OsloMet research time and as part of EK's master's project. No additional funding was received for the study.

Authors' Contributions

EK conducted the data collection and analysis and drafted the manuscript as part of his master's thesis. AKSC assisted with data collection. NP served as a second supervisor, contributing to the study design and facilitating user workshops. TS supported the study design and provided feedback at the final stage. IKHR, ÅB, NSM, and KGSO contributed to the study design, recruited participants, and reviewed the study from a clinical perspective. MP served as the primary supervisor, conceiving the study, overseeing its design, securing ethical approvals, coordinating the project, and converting the thesis into a journal article. PG and SS reviewed the manuscript and provided comments.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey questions.
[DOCX File , 15 KB - jmir_v28i1e84013_app1.docx]

Multimedia Appendix 2
Workshop and interview guide.
[DOCX File , 18 KB - jmir_v28i1e84013_app2.docx]

Multimedia Appendix 3
Data analysis themes.
[DOCX File , 111 KB - jmir_v28i1e84013_app3.docx]

Multimedia Appendix 4
Participant flow diagram.
[DOCX File , 22 KB - jmir_v28i1e84013_app4.docx]

References

XSL•FO
RenderX

1. Anderson V, Le Brocque R, Iselin G, Eren S, Dob R, Davern TJ, et al. Adaptive ability, behavior and quality of life pre and posttraumatic brain injury in childhood. Disabil Rehabil 2012;34(19):1639-1647. [doi: 10.3109/09638288.2012.656789] [Medline: 22416951]
2. Catroppa C, Anderson V. Recovery and predictors of language skills two years following pediatric traumatic brain injury. Brain Lang 2004 Jan;88(1):68-78. [doi: 10.1016/s0093-934x(03)00159-7] [Medline: 14698732]
3. Rohrer-Baumgartner N, Laberg Holthe I, Svendsen EJ, Dahl HM, Borgen IMH, Hauger SL, et al. Children and families with chronic pediatric acquired brain injury in need of rehabilitation: characteristics and main challenges in daily life. Disabil Rehabil 2025 Mar;47(6):1543-1552 [FREE Full text] [doi: 10.1080/09638288.2024.2376936] [Medline: 39001694]
4. Holthe IL, Dahl HM, Rohrer-Baumgartner N, Eichler S, Elseth MF, Holthe, et al. Neuropsychological impairment, brain injury symptoms, and health-related quality of life after pediatric TBI in Oslo. Front Neurol 2021 Jan 28;12:719915 [FREE Full text] [doi: 10.3389/fneur.2021.719915] [Medline: 35153967]
5. Beretta E, Cesareo A, Biffi E, Schafer C, Galbiati S, Strazzer S. Rehabilitation of upper limb in children with acquired brain injury: a preliminary comparative study. J Healthc Eng 2018;2018:4208492 [FREE Full text] [doi: 10.1155/2018/4208492] [Medline: 29732047]
6. Yenilmez O, Altug F. Effect of adding virtual reality to individualized exercise therapy on gross motor function, balance, and functional mobility in children with hemiparetic cerebral palsy: a randomized single-blinded controlled trial. Clin Pediatr (Phila) 2026 Jan;65(1):46-52. [doi: 10.1177/00099228251375527] [Medline: 40990847]
7. Tatla SK, Sauve K, Jarus T, Virji-Babul N, Holsti L. The effects of motivating interventions on rehabilitation outcomes in children and youth with acquired brain injuries: a systematic review. Brain Inj 2014;28(8):1022-1035. [doi: 10.3109/02699052.2014.890747] [Medline: 24661000]
8. Janssen J. Gamification in physical therapy: more than using games. Pediatric Physical Therapy 2017;29(1):95-99. [doi: 10.1097/pep.0000000000000326]
9. Choi JY, Yi S, Shim D, Yoo B, Park ES, Rha D. Home-based virtual reality-enhanced upper limb training system in children with brain injury: a randomized controlled trial. Front Pediatr 2023;11:1131573 [FREE Full text] [doi: 10.3389/fped.2023.1131573] [Medline: 37274815]
10. Malick WH, Butt R, Awan WA, Ashfaq M, Mahmood Q. Effects of augmented reality interventions on the function of upper extremity and balance in children with spastic hemiplegic cerebral palsy: a randomized clinical trial. Front Neurol 2022;13:895055 [FREE Full text] [doi: 10.3389/fneur.2022.895055] [Medline: 35800080]
11. Laver K, Lange B, George S, Deutsch JE, Saposnik G, Chapman M, et al. Virtual reality for stroke rehabilitation. Cochrane Database Syst Rev 2025 Jun 20;6(6):CD008349. [doi: 10.1002/14651858.CD008349.pub5] [Medline: 40537150]
12. Power M, Kennedy S, Cleary F, Mills I, Kinsella S, Celdrán AH. A systematic literature review of XR interventions to improve motor skills development among autistic children. IEEE Access 2024;12:108953-108974. [doi: 10.1109/access.2024.3438361]
13. Keshner EA, Kenyon RV. Using immersive technology for postural research and rehabilitation. Assist Technol 2004;16(1):54-62. [doi: 10.1080/10400435.2004.10132074] [Medline: 15357148]
14. Ren Z, Wu J. The effect of virtual reality games on the gross motor skills of children with cerebral palsy: a meta-analysis of randomized controlled trials. Int J Environ Res Public Health 2019 Oct 14;16(20):3885 [FREE Full text] [doi: 10.3390/ijerph16203885] [Medline: 31614990]
15. Wu J, Loprinzi PD, Ren Z. The rehabilitative effects of virtual reality games on balance performance among children with cerebral palsy: a meta-analysis of randomized controlled trials. Int J Environ Res Public Health 2019 Oct 28;16(21):4161 [FREE Full text] [doi: 10.3390/ijerph16214161] [Medline: 31661938]
16. Ponsford J, Schönberger M, Rajaratnam SMW. A model of fatigue following traumatic brain injury. J Head Trauma Rehabil 2015;30(4):277-282. [doi: 10.1097/HTR.0000000000000049] [Medline: 24721811]
17. Shahid A, Shen J, Shapiro CM. Measurements of sleepiness and fatigue. J Psychosom Res 2010 Jul;69(1):81-89. [doi: 10.1016/j.jpsychores.2010.04.001] [Medline: 20630266]
18. Gawron VJ. Overview of self-reported measures of fatigue. The International Journal of Aviation Psychology 2017 Jul 13;26(3-4):120-131. [doi: 10.1080/10508414.2017.1329627]
19. Riccardi J, Ciccia A. Cognitive fatigue in pediatric traumatic brain injury: a meta-analysis and scoping review. J Head Trauma Rehabil 2021;36(4):226-241. [doi: 10.1097/HTR.0000000000000644] [Medline: 33656467]
20. Wender CLA, Farrar E, Sandroff BM. Attrition, adherence, and compliance to exercise training interventions in persons with traumatic brain injury: a systematic review of training studies. Brain Inj 2025 Jan 02;39(1):70-83. [doi: 10.1080/02699052.2024.2403632] [Medline: 39317382]
21. Mazzone O, Conroy R, Jenkin T, Scheinberg A, Knight S. The assessment and management of fatigue following paediatric acquired brain injury: rehabilitation clinicians' perspectives. Neuropsychol Rehabil 2025 Jun 30;35(5):930-959 [FREE Full text] [doi: 10.1080/09602011.2024.2383337] [Medline: 39078921]
22. Khan F, Amatya B. Management of fatigue in neurological disorders: implications for rehabilitation. Journal of the International Society of Physical and Rehabilitation Medicine 2018;1(2):9-36. [doi: 10.4103/ijprm.ijprm_5_18]
23. Danson E, Brunsdon R, Lane-Brown A, Oxenham V. Neuropsychol Rehabil 2026 May;36(4):778-802. [doi: 10.1080/09602011.2025.2540625] [Medline: 40790797]

XSL•FO
RenderX

24. Montoya MF, Munoz JE, Henao OA. Enhancing virtual rehabilitation in upper limbs with biocybernetic adaptation: the effects of virtual reality on perceived muscle fatigue, game performance and user experience. IEEE Trans Neural Syst Rehabil Eng 2020 Mar;28(3):740-747. [doi: 10.1109/tnsre.2020.2968869]
25. Annabestani M, Sriram S, Caprio A, Janghorbani S, Wong SC, Sigaras A, et al. High-fidelity pose estimation for real-time extended reality (XR) visualization for cardiac catheterization. Sci Rep 2024 Nov 06;14(1):26962 [FREE Full text] [doi: 10.1038/s41598-024-76384-z] [Medline: 39505924]
26. Latulippe K, Hamel C, Giroux D. Co-design to support the development of inclusive eHealth tools for caregivers of functionally dependent older persons: social justice design. J Med Internet Res 2020 Nov 09;22(11):e18399 [FREE Full text] [doi: 10.2196/18399] [Medline: 33164905]
27. Yin RK. Case Study Research: Design and Methods (4th edition), Vol. 5. Thousand Oaks, CA: Sage; 2009.
28. Branca Vergano L, Monesi M, Vicenti G, Bizzoca D, Solarino G, Moretti B. Journal Article Reporting Standards (JARS). American Psychological Association (APA). URL: https://apastyle.apa.org/jars [accessed 2026-04-26]
29. Lincoln YS. Naturalistic Inquiry (Vol. 75). Thousand Oaks, CA: Sage; 1985.
30. Younas A, Masih Y, Sundus A. Alternatives to 'saturation' for greater transparency in reporting of sample size decision-making in qualitative research. Evid Based Nurs 2025 Mar 26;28(2):77-80. [doi: 10.1136/ebnurs-2025-104292] [Medline: 40015944]
31. Creswell JW, Miller DL. Determining validity in qualitative inquiry. Theory Into Practice 2000 Aug;39(3):124-130. [doi: 10.1207/s15430421tip3903_2]
32. O'Brien HL, Cairns P, Hall M. A practical approach to measuring user engagement with the refined User Engagement Scale (UES) and new UES short form. International Journal of Human-Computer Studies 2018 Apr;112:28-39. [doi: 10.1016/j.ijhcs.2018.01.004]
33. Cintora-Sanz A, Muñoz-Romo R, Schrom-Feiertag H, Blanco-Lara A, Vazquéz-Rodriguez T, Cardós-Alonso MC. BMC Med Educ 2026 Feb 01;26(1):373 [FREE Full text] [doi: 10.1186/s12909-026-08641-w] [Medline: 41622167]
34. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2008 Jul 21;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
35. Sweller J. Cognitive load during problem solving: effects on learning. Cognitive Science 2010 Feb 11;12(2):257-285. [doi: 10.1207/s15516709cog1202_4]
36. Cheung J, Maron M, Tatla S, Jarus T. Virtual reality as balance rehabilitation for children with brain injury: a case study. TAD 2013 Aug 06;25(3):207-219. [doi: 10.3233/tad-130383]
37. Proschowsky MS, Reimers SH, Granhøj A. Fatigue among children and adolescents with acquired brain injury in a specialized neurorehabilitation setting. Front Rehabil Sci 2024 Nov 21;5:1454602 [FREE Full text] [doi: 10.3389/fresc.2024.1454602] [Medline: 39640876]
38. van Markus-Doornbosch F, van der Holst M, de Kloet AJ, Vliet Vlieland TPM, Meesters JJL. Fatigue, participation and quality of life in adolescents and young adults with acquired brain injury in an outpatient rehabilitation cohort. Dev Neurorehabil 2020 Jul 20;23(5):328-335 [FREE Full text] [doi: 10.1080/17518423.2019.1692948] [Medline: 31746261]
39. Christie A. A meaningful occupation: the just right challenge. Aus Occup Therapy J 2002 May 09;46(2):52-68. [doi: 10.1046/j.1440-1630.1999.00178.x]
40. Rebenitsch L, Owen C. Review on cybersickness in applications and visual displays. Virtual Reality 2016 Apr 26;20(2):101-125. [doi: 10.1007/s10055-016-0285-9]
41. Deterding S. From game design elements to gamefulness: defining "gamification". New York, NY: Association for Computing Machinery (ACM); 2011 Presented at: MindTrek '11: The 15th International Academic MindTrek Conference: Envisioning Future Media Environments; September 28 - 30, 2011; Tampere, Finland p. 9-15. [doi: 10.1145/2181037.2181040]
42. Hallifax S. Adaptive gamification in education: a literature review of current trends and developments. Berlin, Germany: Springer International Publishing; 2019 Presented at: Transforming Learning with Meaningful Technologies: 14th European Conference on Technology Enhanced Learning, EC-TEL 2019; September 16-19, 2019; Delft, The Netherlands p. 294-307. [doi: 10.1007/978-3-030-29736-7_22]
43. Al-Khresheh MH. The cognitive and motivational benefits of gamification in English language learning: a systematic review. TOPSYJ 2025 Mar 10;18(1):e1. [doi: 10.2174/0118743501359379250305083002]
44. Scaife M, Rogers Y. Kids as informants: telling us what we didn't know or confirming what we knew already? In: The Design of Children's Technology. San Francisco, CA: Morgan Kaufmann Publishers Inc; 1998:27-50.
45. Druin A. The role of children in the design of new technology. Behaviour & Information Technology 2002 Jan;21(1):1-25. [doi: 10.1080/01449290110108659]
46. Brown T. Design thinking. Harv Bus Rev 2008 Jun;86(6):84-92, 141. [Medline: 18605031]
47. Rösch N, Tiberius V, Kraus S. Design thinking for innovation: context factors, process, and outcomes. EJIM 2023 Mar 21;26(7):160-176. [doi: 10.1108/ejim-03-2022-0164]
48. Ku B, Lupton E. Health Design Thinking: Creating Products and Services for Better Health. Boston, MA: MIT Press; 2022.
49. Li X, Chen J, Fu H. The roles of empathy and motivation in creativity in design thinking. Int J Technol Des Educ 2024 Feb 01;34(4):1305-1324. [doi: 10.1007/s10798-023-09869-z]

XSL•FO
RenderX

50. Asano Y. Defining the problem's solution to lead to the ideation phase: - a case study on the use of "how might we...". 2023 Presented at: DUXU 2023: Design, User Experience, and Usability - 12th International Conference; July 23–28, 2023; Copenhagen, Denmark.
51. Jaskyte K, Liedtka J. Design thinking for innovation: practices and intermediate outcomes. Nonprofit Mgmnt & Ldrshp 2022 Jan 05;32(4):555-575. [doi: 10.1002/nml.21498]
52. Levac DE. Individual and contextual factors influencing children's effort in pediatric rehabilitation interventions. Dev Med Child Neurol 2024 Jan 21;66(1):23-31 [FREE Full text] [doi: 10.1111/dmcn.15609] [Medline: 37082901]
53. Sanders EB, Stappers PJ. Co-creation and the new landscapes of design. CoDesign 2008 Mar;4(1):5-18. [doi: 10.1080/15710880701875068]
54. Saidi T, Mutswangwa CT, Douglas TS. Design thinking as a complement to human factors engineering for enhancing medical device usability. Engineering Studies 2019 Jan 21;11(1):34-50. [doi: 10.1080/19378629.2019.1567521]
55. Stefoska-Needham A, Nealon J, Charlton K, Fildes K, Lambert K. Applying design thinking for co-designed health solutions: a case study on chronic kidney disease in regional Australia. Int J Environ Res Public Health 2025 Sep 24;22(10):1475 [FREE Full text] [doi: 10.3390/ijerph22101475] [Medline: 41154878]
56. Srinivasan AG, Smith SS, Pattinson CL, Mann D, Sullivan K, Salmon P, et al. Heart rate variability as an indicator of fatigue: a structural equation model approach. Transportation Research Part F: Traffic Psychology and Behaviour 2024 May;103:420-429. [doi: 10.1016/j.trf.2024.04.015]
57. Martins R, Carvalho J. Eye blinking as an indicator of fatigue and mental load: a systematic review. 2015 Presented at: Occupational Safety and Hygiene III; February 12-13, 2015; Guimarães, Portugal p. 231-235 URL: https://www.researchgate.net/publication/277010972_Eye_blinking_as_an_indicator_of_fatigue_and_mental_load_-_a_systematic_review [doi: 10.1201/b18042-48]
58. Adão Martins NR, Annaheim S, Spengler CM, Rossi RM. Fatigue monitoring through wearables: a state-of-the-art review. Frontiers in Physiology 2021;12:790292. [doi: 10.3389/fphys.2021.790292] [Medline: 790292]

Abbreviations

ABI: acquired brain injury
AR: augmented reality
GDPR: General Data Protection Regulation
SIKT: Norwegian Agency for Shared Services in Education and Research
UES: User Engagement Scale
UI: user interface
VMRFS: Virtual and Mixed Reality Fatigue Scale
VR: virtual reality
XR: extended reality

Edited by S Brini; submitted 12.Sep.2025; peer-reviewed by FLD de la Cour, B Johansson; comments to author 03.Dec.2025; revised version received 09.Feb.2026; accepted 09.Mar.2026; published 02.Jun.2026.

Please cite as:
Kolstad E, Pavel N, Cartajenas AKS, Saidi T, Hurum Rosseland IK, Bergheim Å, Synnøve Mørk N, Granum Selmer-Olsen K, Gazerani P, Shorey S, Pikkarainen M
Opportunities and Concerns of Gamified, Extended Reality for Home-Based Motor Rehabilitation for Children With Brain Injury: Qualitative Case Study on Design Elements Related to the Engagement and Fatigue Perspectives
J Med Internet Res 2026;28:e84013
URL: https://www.jmir.org/2026/1/e84013
doi:10.2196/84013
PMID:

©Eivind Kolstad, Nenad Pavel, Alexis Ken Sosmena Cartajenas, Trust Saidi, Ingvild Kristina Hurum Rosseland, Åse Bergheim, Nora Synnøve Mørk, Kathinka Granum Selmer-Olsen, Parisa Gazerani, Shefaly Shorey, Minna Pikkarainen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

How Cancer Online Support Groups Work, for Whom, and in What Circumstances: Realist Review

Jacqueline L Bender[1,2], PhD; Sarah Scruton[1], MSc; Geoff Wong[3], MD, PhD; Lauren R Squires[1,2], MSc; Stephanie Babinski[1], MA, MPH; Marina Englesakis[4], MLIS; Alaina Cyr[5], MSc; Henry Potts[6], PhD; Andrea Tricco[2,7], PhD; Richard Wassersug[8], PhD; Mary Jane Esplen[9], PhD; Victoria J Forster[10], PhD; Olga Perski[11,12], PhD; Holly O Witteman[13], PhD; Charlene Soobiah[2], MSc; Janet Papadakos[2,14], PhD; Colleen Young[15], BA; Colleen Fox[16], MSc

[1]Cancer Rehabilitation and Survivorship, Department of Supportive Care, Princess Margaret Cancer Centre, University Health Network, 585 University Avenue, Toronto, ON, Canada

[10]Women's College Hospital, Toronto, ON, Canada

[11]Faculty of Social Sciences and Herbet Wertheim School of Public Health, Tampere University, Tampere, Finland

[12]Human Longevity Science, University of California, San Diego, San Diego, CA, United States

[13]Faculté de médicine, Université Laval, Quebec City, QC, Canada

[14]Cancer Education Program, University Health Network, Toronto, ON, Canada

[15]Health Education and Content Services, Mayo Clinic Hospital, Rochester, MN, United States

[16]Ontario Health, Toronto, ON, Canada

[2]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

[3]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom

[4]Library and Information Services, University Health Network, Toronto, ON, Canada

[5]Patient Education and Engagement, University Health Network, Toronto, ON, Canada

[6]Centre for Health Informatics and Multiprofessional Education, University College London, London, United Kingdom

[7]Li Ka Shing Knowledge Institute, St.Michael's Hospital, Unity Health Toronto, Toronto, ON, Canada

[8]Department of Cellular and Physiological Sciences, University of British Columbia, Vancouver, BC, Canada

[9]Department of Psychiatry, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Jacqueline L Bender, PhD
Cancer Rehabilitation and Survivorship, Department of Supportive Care, Princess Margaret Cancer Centre, University Health Network, 585 University Avenue, Toronto, ON, Canada

Abstract

Background: Online support groups (OSGs) offer a convenient way for patients with cancer and survivors of cancer to obtain information and support. However, not all OSGs are helpful, and in some cases, they may increase distress. Overall, there is a lack of clear evidence on the effects of OSGs, along with how and why they work.

Objective: This study examined how, for whom, and in what contexts OSGs work for patients with cancer and survivors of cancer.

Methods: A realist review of the evidence on cancer OSGs was conducted (12 databases, inception to February 15, 2025). We followed Pawson's steps and the RAMESES (Realist and Meta-Narrative Evidence Syntheses: Evolving Standards) quality standards for realist reviews. The Mixed-Methods Appraisal Tool was used to assess the quality of the evidence. Using a realist logic of analysis, we generated a narrative summary of the findings, context-mechanism-outcome configurations, and a program theory (eg, conceptual framework) to explain how cancer OSGs work. Additionally, we developed evidence-based recommendations for optimizing the effectiveness of OSGs.

Results: Of 16,659 papers identified, 168 were included. The evidence was organized into 3 categories, 10 concepts, and 57 context-mechanism-outcome configurations. Cancer OSGs can support patients by providing informational, emotional, appraisal, and altruistic support. This can lead to changes in distress, isolation, empowerment, and self-esteem, through negative and positive appraisals as well as coping efforts. These outcomes, however, depend on user attitudes about OSGs (ie, trustworthy, useful, culturally safe, and easy-to-use), how well the OSG fits their needs (ie, health need, sociodemographic or clinical characteristics, and coping style), and perceptions of control (ie, availability, anonymity, privacy, and autonomy). If an OSG was a good fit for a user's needs, whether they experienced positive outcomes depended on features of the communication technology (ie, modality, response time, and visual social cues), group composition and dynamics (ie, norms, moderation, safety, cohesion, and belonging),

XSL•FO
RenderX

and the nature and content of interactions (ie, emotional expression, cognitive processing, and empathetic responses). These factors trigger underlying mechanisms such as feeling understood, accepted, cared for, valued, reassured, informed, and confident, which result in positive outcomes.

Conclusions: OSGs can address supportive care needs and improve psychosocial well-being for patients with cancer and survivors of cancer. However, outcomes depend on specific contexts and mechanisms that impact how well OSGs meet patients' needs. To optimize effectiveness of OSGs, we recommend (1) helping to assess fit and address specific needs; (2) demonstrating trustworthiness; (3) enhancing anonymity and control, and protecting privacy; (4) enhancing ease-of-use; (5) supporting connection and belonging; (6) encouraging activity; (7) enhancing the nature of content shared to boost therapeutic effects; and (8) monitoring and adjusting design and management strategies.

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2021-053916

(J Med Internet Res 2026;28:e77445) doi:10.2196/77445

KEYWORDS

online support group; online community; realist review; systematic review; cancer

Introduction

With close to 20 million diagnoses and 9.7 million deaths, cancer is the leading noninfectious disease worldwide [1]. By 2050, the number of new cases of cancer is expected to reach 35 million, and over two-thirds of people diagnosed with cancer will become long-term survivors [1]. Patients with cancer and survivors of cancer face multiple physical, functional, and psychosocial challenges during and after cancer treatment [2-7] that affect their quality of life [8-10], resulting in a high burden of supportive care needs [4,11-14].

As many as 1 in 4 patients with cancer and survivors of cancer may turn to online support groups (OSGs) for peer support to address their needs [15]. OSGs, also known as online communities, are virtual spaces where people with common conditions come together to obtain and give information and support [16]. These virtual spaces—ranging from email lists, discussion forums, or groups on social media platforms such as Facebook [17], X (formerly Twitter) [18], and Reddit [19]—may be moderated by professionals or peers [20]. Frequently reported advantages of OSGs include their anonymity, followed by their availability, similarity, and diversity of experiences, and greater information and resources [21].

Several studies suggest that OSGs reduce feelings of isolation, depression, and anxiety, and enhance knowledge, coping, empowerment, and self-management among patients with cancer and survivors of cancer [22-28]. However, OSGs are not risk-free, and in some cases, may increase distress. A randomized controlled trial of an unmoderated OSG for newly diagnosed patients with breast cancer found that participants in the intervention group did worse over time on measures of distress and quality of life compared to controls [29]. Yet 60% of intervention participants reported high satisfaction. These conflicting findings suggest that the benefits of OSGs may depend on individual and contextual factors.

Overall, systematic reviews of the effectiveness of cancer OSGs have found a heterogeneous evidence base of positive, negative, and null findings, and few have examined their underlying mechanisms [20,30-32]. To enhance their effectiveness, there is a need to better understand how, why, for whom, and in what contexts cancer OSGs work. A realist review of the evidence [33] can address these questions and has been successfully used to inform the optimization of other complex interventions [34-37]. Therefore, this study aimed to conduct a realist review of the evidence to understand how, why, for whom, and in what contexts OSGs work for patients with cancer and survivors of cancer.

Methods

Study Design

We undertook a realist review [33] of the evidence on cancer OSGs. Realist reviews aim to build and test theories that explain how complex interventions work by examining how underlying causal processes (eg, mechanisms) in specific circumstances (eg, contexts) interact to produce results (eg, outcomes) [38]. We followed the steps by Pawson et al [33] and the RAMESES (Realist and Meta-Narrative Evidence Syntheses: Evolving Standards) quality standards for realist reviews (Checklist 1) [38]. Our methods were described in detail in a published protocol [39] that was registered on PROSPERO (International Prospective Register of Systematic Review; #CRD42021250046).

Knowledge User Involvement

An integrated knowledge translation approach [40] was followed. Five categories of knowledge users (n=15 in total) were engaged in the project: (1) program developers who have experience developing OSGs; (2) community strategists and moderators that manage and moderate OSGs; (3) clinicians who refer patients to or deliver OSGs; (4) patients, survivors, and family members who have used or created their own "grassroots" OSGs; and (5) health care administrators and policymakers responsible for implementing OSGs and driving best practices. Knowledge users were consulted when designing the study protocol, developing the initial program theory, refining the program theory, and generating best practice recommendations, and they participated as coauthors in the writing of this paper.

Step 1: Initial Program Theory

An initial program theory was developed based on Bender's multitheory framework on the use and effects of breast cancer

XSL•FO
RenderX

OSGs [28], and a 2-hour consultation workshop with project team members and knowledge users. This multitheory framework draws on the Transactional Theory of Stress and Coping [41], Social Comparison Theory [42], the Technology Acceptance Model (TAM) [43], and the Theory of Planned Behavior [44]. As shown in our protocol paper [39], it suggests that people with cancer use OSGs to obtain information and emotional support from similar others to lessen negative appraisals of events, enhance feelings of control, and develop coping strategies, which in turn reduce or buffer anxiety. Beliefs about the usefulness and trustworthiness of OSGs influence adoption and engagement. The availability, anonymity, and low commitment afforded by the online medium are believed to enhance feelings of control. However, the effects of making social comparisons in OSGs may be positive or negative depending on the circumstances and the person's goals.

Step 2: Search for Evidence

The search strategies were developed by an information specialist (ME) and peer-reviewed by an external medical information professional following the PRESS (Peer Review of Electronic Search Strategy) guidelines [45]. Given the range of terms for OSGs and cancer diagnoses, our search strategy was intentionally broad, opting for sensitivity over precision (Multimedia Appendix 1). The following databases were searched from inception to February 15, 2025: MEDLINE, Embase, the Cochrane Database of Systematic Reviews, the Cochrane Central Register of Controlled Trials, Ovid EmCare Nursing, APA PsycINFO (all via Ovid), CINAHL (Ebscohost), and Scopus (Elsevier). In addition, 2 clinical trial registries were searched (ClinicalTrials.gov and WHO ICTRP), along with dissertations (ProQuest Digital Dissertations International) and books or chapters (University of Toronto Summon OneSearch).

Step 3: Paper Selection and Appraisal

The inclusion and exclusion criteria were based on a foundational systematic review of OSGs conducted by Eysenbach et al [20], with feedback from team members and knowledge users (Table 1). The papers gathered from the search were uploaded to Rayyan systematic review management software for screening and selection [46]. To assess interrater agreement, 4 researchers (SB, LRS, AC, and CS) independently screened the same 100 titles and abstracts using a pilot-tested screening tool (Table 2). As 90% agreement was established, the remaining records were divided among the 4 researchers to screen independently, with meetings held every 2 weeks to track progress and discuss and resolve uncertainties. A 10% random sample was audited by the lead author, and disagreements were resolved through discussion. The intentionally broad search strategy yielded a significant number of irrelevant papers (eg, negative answer to at least 1 screening question), the majority of which were excluded at the title and abstract screening stage.

Table . Inclusion and exclusion criteria.

Criteria	Inclusion	Exclusion
Population	People diagnosed with cancer and their family caregivers.	Neither includes people diagnosed with cancer and their family caregivers, nor does it include subgroup analysis of this group. Only includes family caregivers.
Intervention	An online support group (or an online intervention with an online support group component) is defined as a group of individuals with similar or common health-related interests who communicate through a communication platform on the internet.	Does not include an online support group, or does not include subgroup analysis of an online support group in a multicomponent intervention.
Study type	Any type of study design that investigates the factors that influence the use of online support groups or the effects of participation in online support groups.	Studies involving content analyses of communication occurring between members of an online support group that did not examine the perceived effect of participation in the online support group.
Outcomes	All types of outcomes associated with participation in online support groups, including but not limited to user engagement, knowledge, self-efficacy, social support, psychological, behavioral, and physiological outcomes, or use of health services.	Outcomes unrelated to participation in online support groups.
Publication	Papers, including peer-reviewed primary research papers, secondary research papers (eg, reviews), books, or commentaries published in English from database inception to February 15, 2025.	Papers not published in English and outside this time frame. Preprints were excluded from the search strategy. Conference abstracts, study protocols, and dissertations were excluded during paper selection.

XSL•FO
RenderX

Table . Paper screening tool.

Question	Result
Are any of the participants people diagnosed with cancer or their caregivers?	Yes no exclusion
Is the intervention or program online, or does it contain an online component?	Yes no exclusion
Does the intervention include an online support group? (Including, eg, a discussion forum, live chat, email list, online community, peer-to-peer communication, or social media like Twitter [X Corp], Facebook [Meta], Instagram [Meta], Snapchat [Meta], and Reddit [Reddit Inc])	Yes no exclusion
Does the record describe factors that influence the use of OSGs or what the effects of participation (eg, lurking or posting) are (ie, social support, information, reduced distress, and reduced isolation)?	Yes no exclusion

The relevance and rigor of papers were assessed in 2 phases. Relevance was determined by assessing the extent to which the data could contribute to developing and testing context-mechanism-outcome configurations (CMOC) [33]. This was conducted by 3 team members (SB, LRS, and SS). A total of 2 team members independently assessed the relevance of each paper that passed the title and abstract screening and met to resolve uncertainties. A 10% random sample was audited by the lead author, and disagreements were resolved through discussion. Rigor was determined postanalysis for CMOCs with fewer than 5 supporting papers. This approach is recommended for realist reviews when there is sufficient data to develop and test CMOCs, as was the case for most CMOCs in this study [47]. Rigor was determined by assessing whether the methods were credible and trustworthy using the Mixed-Method Appraisal Tool [48]. To establish interrater agreement, 3 team members (SS, LRS, and JLB) independently assessed the same 5 papers using the Mixed-Method Appraisal Tool. Each remaining paper was independently assessed by 2 team members, who met to resolve uncertainties. The results of the rigor assessment are shown in Tables 3 and 4.

XSL•FO
RenderX

Table . MMAT[a] quality assessment results.

Qualitative	Item number						
	S1	S2	1.1	1.2	1.3	1.4	1.5
Sullivan [49], 2003	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bender et al [28], 2013	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Griffiths et al [50], 2015	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sandaunet [51], 2008	Yes	Yes	Yes	Yes	Cannot tell	Yes	Yes
Im et al [52], 2007	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Vilhauer et al [53], 2011	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Chiu and Hiseh [54], 2013	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Donovan et al [55], 2019	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Shaw et al [56], 2000	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stephen et al [57], 2014	Yes	Yes	Yes	Yes	Yes	Yes	Yes
van Uden Kraan et al [24], 2008	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Vilhauer [58], 2014	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ure et al [59], 2020	Yes	Yes	Yes	Yes	Yes	Cannot tell	Yes
Owen et al [60], 2009	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Hoybye et al [61], 2005	Yes	Yes	Yes	Yes	Cannot tell	Yes	Yes

[a]MMAT: Mixed-Method Appraisal Tool.

XSL•FO
RenderX

Table . MMAT[a] quality assessment results.

Quantitative descriptive	Item number						
	S1	S2	4.1	4.2	4.3	4.4	4.5
Yoo et al [62], 2013	Yes	Yes	Yes	Yes	Yes	N/A[b]	Yes
Bender et al [28], 2013	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Leimeister et al [63], 2008	Yes	Yes	Yes	Yes	Yes	N/A	Yes
Malloch and Taylor [64], 2019	Yes	Yes	Yes	Yes	Yes	N/A	Yes
Cabling et al [65], 2018	Yes	Yes	Yes	Cannot tell	Yes	N/A	Yes
Lieberman et al [66], 2005	Yes	Yes	Yes	Cannot tell	Yes	N/A	Yes
Hoybye et al [67], 2010	Yes	Yes	Yes	Cannot tell	Yes	Yes	Yes
Owen et al [60], 2009	Yes	Yes	Yes	Yes	Yes	N/A	Yes
Donovan et al [55], 2019	Yes	Yes	Cannot tell	Yes	Yes	Cannot tell	Yes

[a]MMAT: Mixed-Method Appraisal Tool.

[b]N/A: not applicable.

Qualitative study criteria [48] asked the following question: (S1) Are there clear research questions? (S2) Do the collected data allow for addressing the research question? (1.1) Is the qualitative approach appropriate to answer the research question? (1.2) Are the qualitative data collection methods adequate to address the research question? (1.3) Are the findings adequately derived from the data? (1.4) Is the interpretation of results sufficiently substantiated by data? (1.5) Is there coherence between qualitative data sources, collection, analysis, and interpretation?

Quantitative descriptive study criteria [48] included the following questions: (S1) Are there clear research questions? (S2) Do the collected data allow for addressing the research question? (4.1) Is the sampling strategy relevant to address the research question? (4.2) Is the sample representative of the target population? (4.3) Are the measurements appropriate? (4.4) Is the risk of nonresponse bias low? (4.5) Is the statistical analysis appropriate to answer the research question?

Quantitative nonrandomized study criteria [48] included the following questions: (3.1) Are the participants representative of the target population? (3.2) Are measurements appropriate regarding both the outcome and the intervention (or exposure)? (3.3) Are there complete outcome data? (3.4) Are the confounders accounted for in the design and analysis? (3.5) During the study period, was the intervention administered (or exposure occurring) as intended?

Quantitative nonrandomized study criteria [48] included the following questions: (S1) Are there clear research questions? (S2) Do the collected data allow for addressing the research question? (3.1) Are the participants representative of the target population? (3.2) Are measurements appropriate regarding both the outcome and the intervention (or exposure)? (3.3) Are there complete outcome data? (3.4) Are the confounders accounted for in the design and analysis? (3.5) During the study period, was the intervention administered (or exposure occurred) as intended?

Step 4: Extracting and Organizing the Data

Two team members (SB and LRS) extracted the study characteristics from each paper using a pilot-tested study extraction tool. Patient demographics (Multimedia Appendix 2 [22-24,28,29,49-210]) and intervention characteristics (Multimedia Appendix 3 [22-24,28,29,49-210]) were extracted. In parallel, 1 team member (SB) uploaded the papers into NVivo (version 12; Lumivero) qualitative analysis software to code sections of text that were relevant to supporting, refining, or refuting aspects of the program theory. A preliminary coding framework was developed based on the concepts in the initial program theory. A total of 2 researchers (SB and LRS) used the coding framework to code a random sample of 5 papers to determine agreement and refine the coding framework, and then 2 team members (SB and SS) coded each of the remaining papers. Coding consisted of initially grouping the data into high-level conceptual “buckets” that were defined iteratively through discussions with the lead author (JLB).

Step 5: Synthesis and Analysis

In the final step, data sorted into conceptual buckets were analyzed to determine whether the data were functioning as context, mechanisms, or outcomes [33]. This analysis was

XSL•FO
RenderX

undertaken by 2 team members (JLB and SS) trained in realist methodology guided by an expert in realist methodology (GW). Further, 1 team member (SS) identified prominent patterns of contexts and outcomes in the data and sought to explain them through mechanisms by which they occurred, generating CMOCs. The findings were iteratively confirmed, refuted, or refined, first by JLB and then by GW through discussion with SS. The findings were shared with this project's team and knowledge users for their feedback in a final 2-hour consultation workshop.

Results

Study Characteristics

A total of 168 papers were included in the analysis, which involved a range of 1 (case study) to 180,000 (content analysis) study participants (Figure 1). Studies were conducted within a span of 25 years (1999 - 2025). See Table 5 and Multimedia Appendices 2 and 3 for further details.

Figure 1. Flow diagram. OSG: online support group; WHO: World Health Organization.

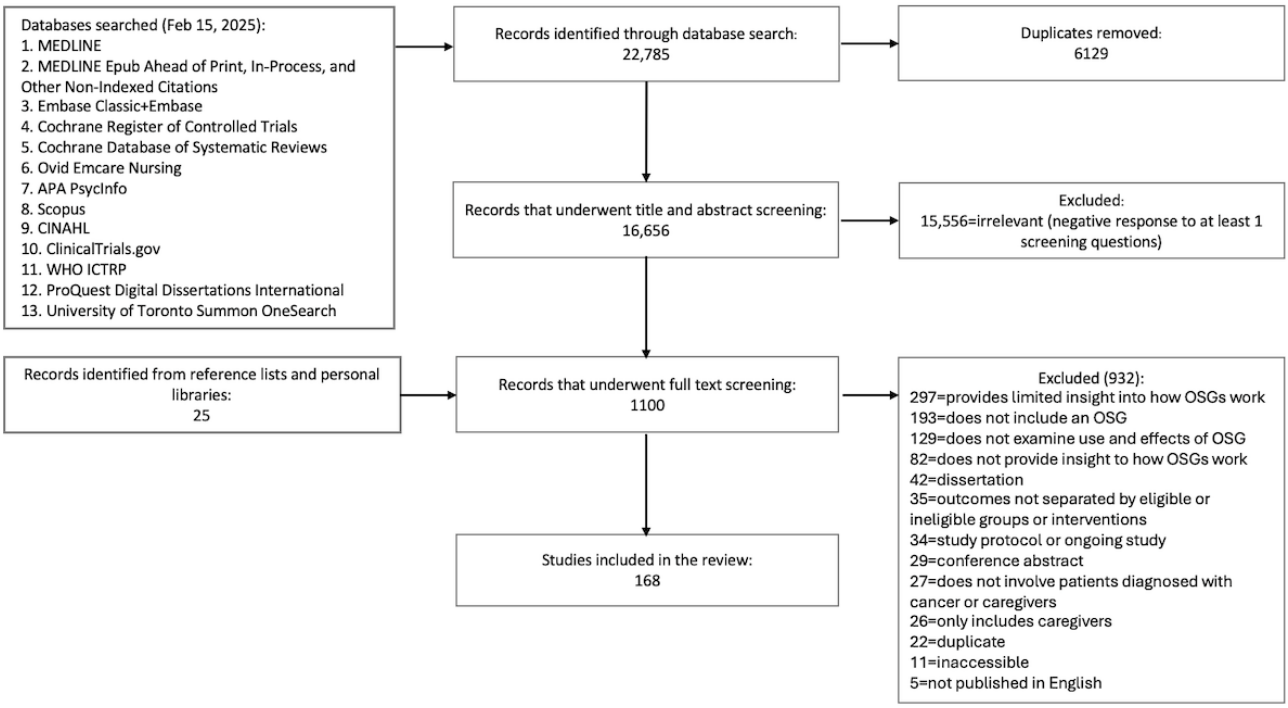

XSL•FO
RenderX

Table . Study characteristics (n=168).

Characteristic	Values, n (%)
Country (of first author)	
United States	109 (64.9)
United Kingdom	16 (9.5)
Australia	12 (7.1)
Netherlands	9 (5.4)
Canada	8 (4.8)
Other (Denmark, Germany, Norway, Taiwan, South Korea, Japan, Finland, and China)	14 (8.3)
Study type	
Content analyses	39 (22.6)
Mixed methods	33 (11.9)
Quantitative (cross-sectional surveys)	25 (16.7)
Qualitative (interviews or focus groups)	26 (16.7)
Randomized controlled trials	14 (12.5)
Reviews	11 (6.5)
Commentaries	7 (3.6)
Pretest and posttest measurement	4 (6)
Case studies	3 (1.2)
Quantitative (longitudinal surveys)	3 (1.2)
Multiple methods	3 (1.2)
Cancer types	
Breast	71 (42.3)
Prostate	7 (4.2)
Ovarian	3 (1.8)
Other (eg, thyroid, colorectal, hematological, lung, sarcomas, testicular, head, and neck)	11 (6.5)
Multiple cancer types	50 (29.8)
Not specified	26 (15.5)
Sex	
All female	68 (40.5)
All male	10 (6)
Majority female (>50%)	50 (29.8)
Majority male (>50%)	6 (3.6)
Equal male + female	1 (0.6)
Not specified	33 (19.6)
Race or ethnicity	
Not specified	101 (60.1)
Majority White (>50%)	51 (30.4)
Majority non-White	16 (9.5)
Education levels	
Not described	89 (53)
The majority (>50%) postsecondary education (ie, some college or higher)	76 (45.2)

XSL•FO
RenderX

Characteristic	Values, n (%)
50% or more without postsecondary education	3 (1.8)
Technology type	
Multicomponent websites (eg, a website with multiple support services, such as information, decision support, and peer support; may include different OSG[a] types)	44 (26.2)
Bulletin board or discussion forum	37 (22)
Not specified (eg, studies of people who use OSGs)	32 (19)
Other (eg, newsgroups, live chats, mailing lists, social networking sites, and video chats)	26 (15.5)
Multiple types (ie, multiple OSGs in 1 study)	29 (17.3)
Synchronous vs asynchronous	
Asynchronous only	97 (57.7)
Synchronous only	6 (3.6)
Not specified	44 (26.2)
Both	21 (12.5)
Moderator present	
Yes	41 (24.4)
No	8 (4.8)
Not specified	99 (58.9)
Multiple OSGs with one or more moderators	20 (11.9)
Level of moderation description	
Not described	109 (64.9)
High (eg, described moderators and the moderation strategy)	24 (14.3)
Low (eg, limited description of moderator and moderation strategy)	35 (20.8)
Moderator type	
Trained professional, study staff, or organization staff	20 (11.9)
HCP[b]	11 (6.5)
Lay peer	4 (2.4)
Trained peer	1 (0.6)
Combination	5 (3)
Multiple OSGs with one or more having an HCP or professional moderator	6 (3.6)
Not described	121 (72)

[a]OSG: online support group.

[b]HCP: health care provider.

Main Findings

Overview

The evidence was organized into 3 overarching categories and 10 concepts: motivation (attitudes, need or fit, and control beliefs), use (communication technology factors, group composition and dynamics, and nature and content of interactions), and outcomes (informational, emotional, validatory, and altruistic support). Additionally, a total of 57 CMOCs were generated. CMOCs are labeled (eg, A1, B1, etc) and described in the subsequent paragraphs. The papers that contributed to each CMOC are found in Table 6. A visual depiction of the refined program theory is provided in Figure 2.

XSL•FO
RenderX

Table . Categories, concepts, and CMOCs[a].

ID code[b]	CMOC	Paper reference
Motivation		
Attitudes		
A1	When an individual cannot assess the trustworthiness and intentions of other group members (C[c]), they will be less likely to use OSGs[d] (O[e]) because they are not certain about their credibility (M[f]).	[23,68-72]
A2	When individuals join OSGs that are more relevant to their cancer journey stage (C), they are more likely to use them or benefit (O) because their questions or concerns will be addressed or it will meet their needs (O).	[24,28,50,51,73-79]
A3	When an individual does not have a timely need for information or support (C), they may reduce or stop their use of an OSG (O), because the OSG is no longer useful to them (M).	[28,51,80,81]
A4	When an individual is not sure if a particular OSG will meet their expectations and needs (C), they may join many groups and lurk in them (O) because they want to learn more about the group before engaging (M).	[28,58,69,71,82,83]
A5	When an individual's culture stigmatizes cancer (C), they may be less likely to use OSGs (O) because they may feel uncomfortable disclosing their cancer to others for fear of stigma (M).	[52,70,84-86]
A6	When OSGs are not culturally sensitive (C), individuals may be less likely to use them (O) because they may feel less comfortable participating in them (M).	[70,84,85,87-89]
A7	When OSGs are not easy to use (C), individuals will be less likely to use them (O) because it makes them feel frustrated (M).	[28,50,58,74,90,91]
Need or fit		
B1	When an individual has a complex case or needs (C), they may be more likely to use OSGs (O) because they are distressed and therefore need more support from individuals in a similar situation (M).	[28,53,69,70,76,81,92-105]
B2	When individuals do not have strong offline support systems (C), they may be more likely to use OSGs (O) because they have more of a need for them (M).	[68,86,94,95,97,105-118]
B3	When individuals feel that the informational and emotional support provided by their health care providers is not sufficient (C), they will be more likely to use OSGs (O) because it is a convenient way to access the support they want from other people who have experienced cancer (M).	[24,28,54,55,80,85,94,108,110,111,114,117,119-125]
B4	When patients with cancer do not want to burden their family and friends about their illness (C), they may be more likely to use OSGs (O) because OSGs can provide an alternative avenue for support (M).	[24,49,56,57,59,81,86,118,121,123,125-128]

XSL•FO
RenderX

ID code[b]	CMOC	Paper reference
B5	When trusted or credible others (eg, health professionals or friends) do not trust OSGs (C), individuals may be less likely to use them (O) because they value the opinion of the trusted other (M).	[23,28,68,129,130]
B6	When an individual has low eHealth literacy (C), they may be less likely to use OSGs (O) because they are not comfortable with the technology (M).	[23,62,79,95,131-133]
B7	When individuals face fewer barriers to accessing, understanding, or using OSGs (C), they may be more likely to use OSGs (O) because of ease of use (and convenience; M).	[67,68,70,76,77,95,96,115,131,134-140]
B8	People with emotion-focused coping styles (C) may be more likely to use OSGs or more likely to benefit (O) because they feel safety or anonymity facilitates emotional expression (M).	[62,67,106,107,141,142]
B9	People with avoidant coping styles (C) may be less likely to use OSGs or less likely to benefit (O) because they may feel distressed by others expressing their emotions in the OSGs (M).	[62,67,106,107]
B10	When an individual reads distressing information in an OSG (C), they may reduce or stop their use of an OSG (O) to avoid feeling distressed by the information (M).	[28,51,80,101,143-145]
Control beliefs		
C1	When peer support is provided online (C), users can access the support more easily (O) because it removes the barriers that individuals often face when accessing in-person support (M).	[49,56-58,60,87,91,105,108,110,111,114,117,137,146-149]
C2	The 24/7 availability of OSGs (C) makes people feel better supported (O) because they can have their needs met at any time (M).	[28,50,56,58,60,69,148,150]
C3	Users can access OSGs to look for information that they need, whenever they need it (C), which makes users feel more in control (O) because they can manage the type of information they receive and when they receive it (M).	[49,58,59]
C4	When OSGs are perceived by potential users as having privacy issues (C), individuals will be less likely to use them (O), because they are concerned their personal information will be leaked or used without their permission (M).	[23,28,52,69,71,136,151,152]
Use		
Communication technology factors		
D1	When people using OSGs take the time to reflect on what they want to write and how to write it (C), the relevance and content of the posts improve (O) because they can carefully craft a thoughtful post or response (M).	[56,57,72]
D2	When individuals with a pressing concern post in OSGs and there is a delay in response (or no response; C), it can make users feel isolated (O) because they are led to believe that their concerns are unimportant or unworthy of a response (M).	[56,61,69,150,153]

XSL•FO
RenderX

ID code[b]	CMOC	Paper reference
D3	When the posts and responses in an OSG are fast-paced (eg, in a live chat; C), those seeking support may not find this helpful (O), because they find the discussion hard to follow, overwhelming, or lacking in depth (M).	[56,57,90,101]
D4	Text-based only communication in OSGs (C) can make it harder for users to feel connected (O) because of the lack of visual cues (M).	[57,60,63,87,154]
D5	Text-based only communication in OSGs (C) can cause miscommunication (O) because of the lack of visual cues (M).	[60,61,63,101]
D6	When OSGs allow users to use different modes of interaction (ie, emojis, reactions, and likes; C), it makes it easier for users to interpret posts (O) because it makes up for the lack of visual and audio cues (M).	[60,71,150,153,155,156]
D7	When OSGs include the option to have personalized profiles that include personal information about them, their interests, and their particular circumstances (C), it makes it easier for users to feel connected to one another (O), because the information shared in profiles helps users get to know and relate to each other (M).	[50,90,135,153,157]
D8	When OSGs are used in conjunction with in-person support groups (C), users receive greater social support (O) because they receive the unique benefits of both types of support (M).	[28,55]
D9	When OSGs offer some privacy and anonymity (C), users feel more comfortable disclosing information and emotions (O) because they feel safer and less likely to be embarrassed or shamed (M).	[22,23,28,50,56,57,60,61,70,72,81,84,86,126,129,147,148,151,158]
D10	When users are provided the option of having private discussions with individuals or subgroups of individuals (C), it makes them feel more comfortable disclosing information and emotions (O) because they value the privacy this provides (M).	[50,59,121,133,157,159]
Group composition and dynamics		
E1	When OSG members adhere to the group's norms (C), they are likely to benefit (O) because they feel accepted by the group (M).	[64,71,94,146,160-164]
E2	When an OSG has a social norm that discourages discussion of negative topics (C), it can make users feel isolated (O) because they cannot receive needed support about that topic (M).	[51,53,58,65,71,81,101,126,154]
E3	When OSGs have moderators who are active, encouraging, and supportive (C), users will be more likely to actively participate (O) because moderators facilitate conversations (M).	[50,55,57,58,74,90,101,165,166]
E4	When moderators provide emotional support to OSG members (C), users will be more likely to actively participate in OSGs (O) because they feel supported (M).	[50,66]
E5	When OSGs have knowledgeable moderators who share resources and correct misinformation (C), users are more likely to trust the information shared in the OSG (O) because they believe the OSG is trustworthy or credible (M).	[50,51,70,72,88,90,144,148,152]

XSL•FO
RenderX

ID code[b]	CMOC	Paper reference
E6	When an OSG has one or more dominant members (ie, moderators and informal group members) who overstep by posting and behaving in a negative way (C), other group members feel uncomfortable (O) because they judge these acts to be inappropriate (M).	[52,65,69]
E7	When OSGs are structured in a way that facilitates the formation of close bonds or connections (ie, smaller groups, similarities between group members, supportive dynamic, and humor; C), users are more likely to benefit (O) because group members feel better supported (M).	[49,53,56-60,66,68,69,72,101,121-123,127,144,153-155,157,164,167-170]
E8	When OSGs enable users to connect with people in similar situations whom they do not know offline (C), users feel less isolated (O) because they learn that there are others like them (M).	[50-53,56,59,60,69,71,81,82,94,119,121,125,126,148,167,171-174]
E9	When an OSG provides a psychologically safe space where members can trust each other (C), users feel more comfortable disclosing information and emotions (O) because they do not have to be concerned about negative repercussions (M).	[71,157,175-177]
E10	When an individual has a poor prognosis (C), they may feel less comfortable disclosing information and emotions (especially mixed stage or cancer type groups; O) because they do not want to “scare” or distress other users (M).	[53,64,65,101]
E11	When an individual feels that their thoughts and experiences are different from the group (C), they may reduce or stop their use of an OSG (O) because they are concerned that they will not be accepted (M).	[22,51,53,65,144]
Nature and content of interactions		
F1	When OSG users express negative emotions and thoughts in an OSG (C), they experience stress relief (O) because writing about a stressful event allows one to process and learn from one’s emotions (M).	[49,50,54,56,57,60,83,89,101,106,107,119,121,122,126-128,150,153,154,156,172,178-181]
F2	When an OSG member makes a post that includes emotional expression and uncertainty (C), they are more likely to receive a response (O) because other users recognize that they need support (M).	[81,120,146,172,180,182-185]
F3	When individuals discuss their cancer experience with other members of an OSG (C), it can help them cope with their cancer diagnosis (O) because it helps them to cognitively process and make sense of their illness (M).	[22,24,49,53,54,61,74,81,92,126,127,142,150,167,186-194]
F4	When people who use OSGs dislike opening up emotionally and creating emotional bonds (C), they are more likely to provide mainly informational support (O) because this is what they feel comfortable doing (M).	[49,71,126,134,136,154,158,195]
F5	When OSG members use appropriate humor to interact with each other (C), they are more likely to benefit (O) because humor serves as a distraction and breaks the tension or provides relief for stressful situations (coping mechanism; M).	[49,68,71,126,153,154]

XSL•FO
RenderX

ID code[b]	CMOC	Paper reference
F6	When individuals actively participate in an OSG by frequently posting and commenting (C), they receive more support from others (O) because their requests for support receive more responses (M).	[29,63,75,77,93,95,123,139,155,186,188,196-200]
F7	When individuals lurk and read posts in OSGs (C), they can still benefit (C), because they can obtain informational and emotional support from reading others' posts (M).	[23,77,83,94,101,123,127,129,156,157]
Outcomes		
Social support		
G1	When an OSG has members who are willing to share and help interpret information (C), users gain knowledge (O), and their uncertainty is reduced (O) because they have a better understanding of their disease and what to expect (M).	[illegible]
G2	When individuals receive advice and encouragement from other OSG users on how to improve their care experience (C), they become more actively involved in their care (O) because they feel more empowered (M).	[22,24,49,101,122-124,144,154,157,193,201,206]
G3	When an OSG member receives supportive comments (eg, empathetic, compassionate, encouragement, and/or expressions of solidarity responses; C), they feel emotionally supported (O), because they know other people care about them (M).	[illegible]
G4	When individuals communicate with others in OSGs who have had similar experiences (C), they feel less isolated (O) because they receive validation that what they are going through is normal (M).	[24,49,50,55,56,78,79,82,101,126-128,150,201,205]
G5	When individuals who are stable or doing well read about the negative outcomes of other members of the OSG (C), they feel better about themselves (O) because they know they are doing better than other members (M).	[24,54,56,101]
G6	When individuals read about the positive outcomes of other members of the OSG (C), they feel hopeful or optimistic about their future (O) because other people have survived the disease, and that makes them feel that they could survive the disease as well (M).	[24,28,53,54,56,126,129]
G7	When individuals use OSGs to support others (C), it makes them feel good about themselves (O) because they are helping others (M).	[24,50,54,56,61,68,80,125,139,175,193]
G8	When individuals use OSGs to support others (C), it can help them to cope with their own cancer diagnosis (O) because it enables them to process and positively reframe their own experiences (M).	[24,61,75,77,78,93,109,128,139,178,182,188,209,210]

[a]CMOC: context-mechanism-outcome configuration.

[b]Codes refer to which categories, concepts, and context-mechanism-outcome configuration relate to respective themes throughout the paper.

[c]C: context.

[d]OSG: online support group.

[e]O: outcome.

[f]M: mechanism.

XSL•FO
RenderX

Figure 2. Refined program theory. *Mechanisms leading to positive outcomes

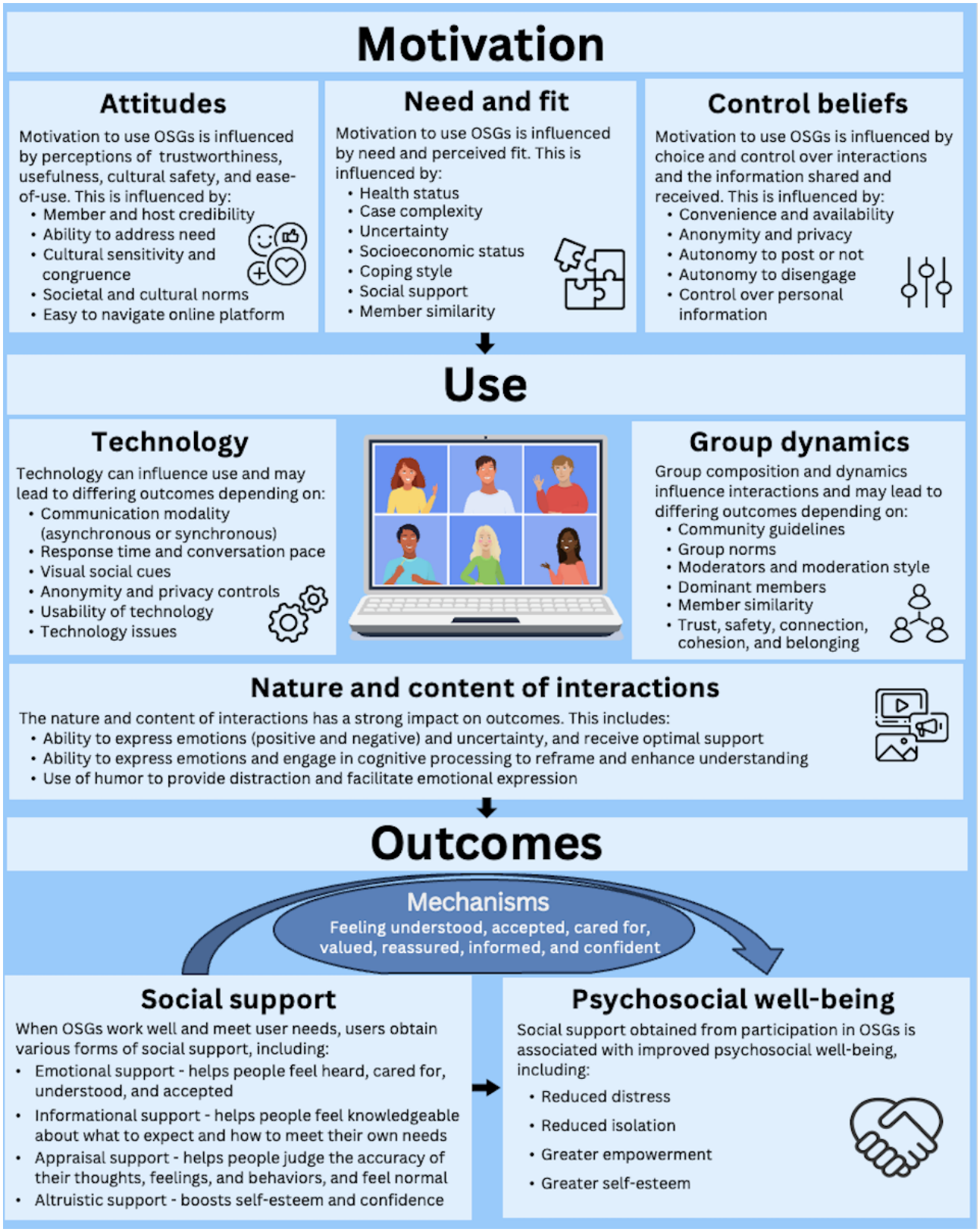

Motivation

Whether an individual chooses to use an OSG depends on their attitudes about OSGs, whether an OSG is a good fit for their needs and their control beliefs.

Attitudes

OSGs must meet certain criteria to be considered helpful. They must be trustworthy, useful, culturally congruent (eg, aligned with cultural norms) and culturally sensitive (eg, demonstrating knowledge, awareness, and acceptance of different cultures, and recognition of interpersonal power differentials [211]), and easy-to-use.

XSL•FO
RenderX

The trustworthiness of the OSG depends on how credible its members (eg, content and behaviors) and host organization (eg, community procedures and safeguards) appear. If either appears to be untrustworthy, people may avoid or leave the OSG due to concerns about receiving inaccurate information and misuse of their personal information, leading to a privacy breach (A1).

The usefulness of OSGs depends on how well their content matches an individual's needs. Those recently diagnosed, deciding on treatment, or experiencing treatment side effects may benefit more from OSGs because they have needs that are likely to be discussed (A2). Conversely, individuals with fewer needs, such as those farther from diagnosis, or with no evidence of disease, may stop using OSGs when they no longer need them (A3). Most users will explore or "shop around" OSGs to find the best fit for their needs (A4).

Cultural norms and the level of cultural sensitivity, congruence, and safety in OSGs may affect participation. Some individuals may be reluctant to join OSGs if their culture stigmatizes cancer or discourages the disclosure of personal information (A5). Others may prefer to rely on family or their religious community for support if an OSG lacks culturally relevant information and safety (A6). Representation across ethnicity, language, and cultural values can enhance perceived safety and encourage engagement.

Finally, even if an OSG meets these criteria, individuals may be unlikely to take part in it if it is not easy to use. Confusing or difficult-to-navigate online environments can lead to frustration and discourage OSG use (A7).

Need and Perceived Fit

Needs drive use. People experiencing uncertainty or complex care needs (ie, newly diagnosed or metastatic or advanced cancer) are more likely to seek support from OSGs because of distress, fear of recurrence, persistent or late treatment effects, or disease progression (B1). Individuals who lack sufficient offline support often have a greater need for support from OSGs (B2 and B3). This was particularly apparent during the COVID-19 pandemic, which led to an increase in OSG use as individuals were unable to access in-person support from cancer peers, family or friends. Some individuals with strong offline support turn to OSGs because they do not want to burden or distress their family and friends with cancer-related concerns (B4). Others may not use OSGs if their offline support networks question their safety and trustworthiness (B5).

Socioeconomic status and coping style may affect OSG use (B6 to B10). A broad range of sociodemographic (eg, age, sex, race, ethnicity, sexual orientation, etc) and clinical factors (eg, diagnosis, stage, issue, etc) may determine fit (E7). Lower socioeconomic status (ie, income or educational attainment) may limit access to OSGs due to lack of stable housing, internet access, or e-health literacy [212]; B6 and B7). Individuals with emotion-focused coping styles may be more likely to use OSGs as they facilitate emotional disclosure (B8). On the other hand, people with avoidant coping styles may find the emotional nature of posts in OSGs distressing and may withdraw from OSGs or avoid them altogether (B9 and B10). Additionally, finding similar others with shared experiences (in terms of sociodemographic and clinical characteristics) is important for relatability, sense of belonging, and perceived fit.

Control Beliefs

An individual's sense of choice and control over their interactions, and the information they share and receive in OSGs, is a critical factor in determining use of OSGs. The online medium overcomes practical barriers to accessing support, such as time and distance, and affords individuals the convenience and control of accessing support safely, including during infectious disease outbreaks such as the COVID-19 pandemic (C1). In OSGs with members from diverse time zones, users are likely to receive a timely response, meaning that they can obtain support when needed (C2). Users can also limit the information they are exposed to in OSGs by not responding to or not reading posts, or by disengaging from a post, thread, or user that is distressing. Taking action to avoid distressing information in OSGs can prevent feelings of being overwhelmed or distressed, which may be more difficult in in-person groups (C3). However, users may experience a lack of perceived control when there is a lack of anonymity, and there are privacy concerns. For example, some individuals are hesitant to use OSGs—especially on public platforms (ie, Facebook)—where their personal information could be used or disclosed without their permission, which they fear could impact their personal relationships or health care (C4).

Use

Once an individual finds an OSG that is a good fit, whether they experience positive outcomes depends on how they participate, the people involved, and the discussion that occurs. This is influenced by communication technology factors, group composition and dynamics (including the presence and nature of moderation), and the nature and content of interactions. These factors trigger underlying mechanisms such as feeling understood, accepted, cared for, valued, reassured, informed, and confident, which result in positive outcomes.

Communication Technology Factors

Communication modality, which influences user response time and the pace of the conversation, as well as technology issues in OSGs, impacts user experience. Asynchronous communication may improve the quality of support received as members can take the time to craft thoughtful messages and reflect on responses (D1). However, on some asynchronous platforms, posts that go unanswered may become buried and never receive a response, which can lead to feelings of being unsupported, isolated, and rejected (D2). Synchronous chats offer immediate responses, but the discussion is often fast and difficult to follow and may be hindered by technology issues (ie, poor Wi-Fi connection), which can overwhelm users and limit the ability to have deep, meaningful conversations (D3).

The lack of visual social cues in OSGs is commonly cited as a reason why OSGs may not fully replace in-person support. Visual social cues are important for connecting on an emotional level: feeling seen and heard, interpreting communication, building trust, and preventing or resolving conflicts (D4 and D5). OSGs may compensate for the lack of visual cues through communication features, such as "liking" posts and emojis,

XSL•FO
RenderX

which help users express emotions (D6). Customizable personal profiles (eg, profile pictures, status updates, videos, and interests) also help users get to know and connect with one another (D7). However, some users feel that OSGs work best when used in combination with in-person support groups to experience the benefits of both forms of social support (D8).

Anonymity and privacy in OSGs encourage users to open up and disclose sensitive or emotional information, which is critical for maximizing benefits. When OSGs offer some degree of anonymity and privacy, users feel more comfortable disclosing information because they feel safer and less likely to be embarrassed or shamed (D9). These features may be valuable for some men and other individuals who are uncomfortable disclosing health conditions and emotions. Additionally, some users may feel more comfortable disclosing information and sharing emotions in private conversations (eg, private direct messaging) or with subgroups of individuals in OSGs, rather than with the entire OSG (D10).

Group Composition and Dynamics

Within each OSG, there exists a set of rules or community guidelines, formal or informal, explicit or implicit, that establish group norms for acceptable behavior to foster group cohesion and belonging (E1). These rules or community guidelines often outline unacceptable behaviors (eg, flaming, bullying, or harassment), communication styles (eg, rude and disrespectful), and topics of conversation (eg, medical advice).

Sometimes, OSG members may collectively decide which topics are unacceptable, such as those seen as “negative,” including feelings of hopelessness, despair, depression, and death, and members may be discouraged from discussing them directly (eg, called out and removed) or indirectly (eg, avoidance or nonresponse). This can make individuals seeking support for these topics feel isolated (E2). This is more common in nonprofessionally moderated or unmoderated OSGs, which emphasize a positive attitude and outcome. In contrast, professionally moderated OSGs typically encourage the expression of both positive and negative emotions and experiences.

Moderators in OSGs encourage participation by welcoming new members, asking questions, starting conversations, and connecting users to relevant topics and other users while also providing emotional support (E3 and E4). Professional moderators (ie, health care providers, study staff, and host organization staff) can increase the perceived trustworthiness and safety of the group by preventing the spread of misinformation, managing difficult emotions, and addressing inappropriate behavior. Moderators who are trained in cultural sensitivity and competency can further strengthen these efforts by providing support that is tailored to users’ cultural and social contexts (E5). However, dominant members can disrupt the conversation (ie, arguing with people who disagree with them), which can cause discomfort (E6). Clear guidelines and effective moderation are important for group safety and minimizing conflict.

OSG users can form strong bonds with fellow group members, leading to connection, cohesion, and belonging. Strong bonds in OSGs are built on empathy, authenticity, mutual support, and universality (eg, a shared quality, interest, or experience). The strength of these bonds is influenced by the size of the group (eg, more opportunities to interact and get to know each other), group cohesion (eg, extent to which group members are attracted to the group and its goals and feel a sense of belonging and influence), and member similarities (eg, similar cancer, age, gender, ethnicity, sexual orientation, or issue; E7). For some, relationships formed with OSG members evolve into genuine friendships that are stronger than those in their offline network. Finding in-person support from others who are going through similar situations can be challenging, especially for those who live in rural areas or who have rare health conditions. OSGs help users overcome geographic barriers, enabling users to find and form connections with similar others from anywhere in the world (E8).

For users to feel comfortable expressing emotions and disclosing sensitive and potentially stigmatizing information in an OSG, they must also trust group members. The level of trust is often a direct result of the strength of the bonds between group members, formed through reciprocal kindness, empathy, support, and encouragement (E9). However, users with complex care needs or who have a poor prognosis may be reluctant to disclose persistent or worsening side effects because they do not want to distress others or experience rejection (E10). Ultimately, the discussion of negative emotions and experiences that are not accepted by the group may lead people to withdraw from an OSG because of a lack of belonging and acceptance (E11).

Nature and Content of Interactions

The quality of support that users receive in OSGs depends on the nature and content of interactions between group members. When groups encourage emotional expression, users can experience stress relief because writing about a stressful event helps users express and process their emotions, fears, or uncertainty (F1). When posts express emotions and uncertainty, users are more likely to receive empathy, support, and encouragement, because others recognize that they are in distress (F2).

Posts that describe experiences can facilitate active coping as users can read inspiring stories, learn practical tips, and engage in cognitive processing to make sense of their own condition (F3). Some users prefer to provide information rather than emotional support, which helps other users by enhancing their understanding, preparedness, and sense of control (F4). Additionally, humor in OSGs can foster a more lighthearted and relaxed atmosphere that can help to release tension, inspire hope, build rapport, and facilitate disclosure (F5). For some, humor serves as a coping mechanism and form of stress relief by distracting from and lightening the burden of the disease.

Active participation and engagement with OSGs, by posting or replying to others, leads to greater benefits and better outcomes (F6). However, people who only read and do not post in OSGs (eg, lurkers) can still benefit. Not only can they receive informational support, but they can also gain emotional support by forming emotional ties with others, by learning from and connecting with others’ emotions and experiences (F7).

XSL•FO
RenderX

Outcomes

If an OSG is a good fit for an individual's needs, and encourages supportive, inclusive discussion, users can experience various forms of social support: informational support (advice or guidance), emotional support (expressions of comfort, caring, and encouragement), appraisal support (help to assess the appropriateness or normativeness of thoughts, feelings, and experiences), and altruism-related esteem support (helping others with no expectation of reward). These types of support lead to positive or negative appraisals and coping mechanisms, which impact psychosocial well-being by affecting distress, isolation, self-esteem, and empowerment. Given that social support is a multidimensional construct, the same exchange in an OSG can yield multiple forms of social support, providing multiple benefits simultaneously.

Informational Support

Informational support is commonly given in OSGs by sharing personal cancer experiences and helpful resources. Reading about other people's cancer experiences helps users better understand their own health condition and what to expect, reducing their uncertainty and increasing their sense of control (G1). When users provide informational support, they often share links and references to websites, research studies, conference summaries, etc. Increased knowledge about one's health condition often motivates users to become more actively involved in treatment decision-making and symptom and side effect management. Users are encouraged to advocate for themselves and ensure that they are receiving the best care (G2).

Emotional Support

Emotional support is commonly given in OSGs by conveying empathy. Posts that express emotions and uncertainty and a need for support receive the most responses, most often from individuals who understand because they have had a similar experience. Emotional support often goes together with esteem support. For example, the most supportive responses are often from individuals who demonstrate empathy by validating the person's feelings and experiences, disclosing similar feelings and experiences, and offering words of encouragement to help them overcome the issues they are facing. As a result, the recipient feels heard, understood, accepted, less alone, empowered, and more optimistic (G3).

Appraisal Support

Users can also receive appraisal support, through feedback, social comparision, and validation, by sharing and comparing thoughts, feelings, and experiences. When users make lateral comparisons by comparing themselves to others with similar experiences, they learn that what they are going through is normal and to be expected, and that they will be able to get through it. This helps users realize that their thoughts, feelings, and experiences are appropriate, reducing uncertainty and isolation (G4). Users can make other types of comparisons in OSGs. Downward comparisons with people who are doing worse can make users feel better about their own situation, which enhances self-esteem (G5). Upward comparisons with people who are doing better can offer inspiration, motivation, or hope for the future (G6).

Altruistic Support

Finally, OSG users can experience the benefits of altruism by helping others. They may transition from a support seeker to a support provider, and the feeling of helping others can boost their self-esteem (G7). Providing support to others in OSGs can also benefit the support provider by helping them to cognitively process and positively reframe their own experiences (G8).

Recommendations

Based on the findings, we have generated evidence-based practice recommendations to optimize the use and effects of OSGs for patients with cancer and survivors of cancer (Textbox 1).

XSL•FO
RenderX

Textbox 1. Recommendations to enhance the experience and outcomes of online support groups.

Help assess fit and address specific needs

- Communicate why the online support group was created, for whom, and the intended benefits.
- Enable users to find relevant support to address specific needs by leveraging technology (eg, optimized search function, dedicated subforums to help users find the discussions and people that match their needs, etc).
- Guide users to more suitable forums or other online support groups as their needs change.

Demonstrate trustworthiness

- Establish and demonstrate community rules or norms for acceptable behavior to foster a safe and respectful online environment.
- Implement a moderation policy to ensure timely responses and manage inaccurate information and inappropriate behavior.
- Adopt transparent data policies and procedures to protect participant privacy.

Enhance anonymity and control, and protect privacy

- Provide users with privacy controls to tailor privacy settings to their preferences.
- Limit personal information requested (and shared) to protect user privacy.
- Enable users to use pseudonyms or avatars to enhance anonymity and facilitate disclosure.

Enhance ease of use

- Ensure the system is intuitive and easy-to-use, and adaptable to the evolving community usage.
- Provide easy-to-find instructions on how to use the system (eg, add to a profile page, post a comment, create a new topic or thread, etc).

Support connection and belonging

- Promote inclusion and acceptance of different backgrounds, identities, and abilities.
- Ensure users can find others who are similar to them (lateral comparisons).
- Ensure users can find and connect with other users who have experience (upward comparisons).
- Ensure there are spaces for users to connect and support each other at all stages of the cancer care continuum.

Encourage activity

- Employ trained moderators (lay or professional) to welcome members, establish etiquette, tone and style, foster discussion, connection, and matching, and implement community rules.
- Engage and train an active core group of members to facilitate and maintain activity (altruistic support).
- Implement alerts that are customizable to the stage of the community life cycle to notify users of relevant content in a timely way and avoid overwhelming users with too many notifications.

Enhance the nature of the content shared to boost therapeutic effects

- Encourage posts with emotional disclosure, information sharing, and cognitive processing.
- Offer user tips on how to participate and craft posts for maximum benefit.
- Use reactions (like, helpful, caring, etc) and allow emojis to help users convey meaning and emotion.

Monitor and adjust design and management strategies to optimize effectiveness

- Monitor activity and implement proven community management strategies to boost activity and supportive exchanges.
- Monitor usability and implement ongoing technology improvements to enhance optimal activity and supportive exchanges.
- Make use of available technology and training to enhance moderators' ability to provide timely, personalized support.

Discussion

Summary

This is the first realist review of the evidence on OSGs. The findings show that the functional provisions of social support gained from cancer OSGs can lead to changes in psychosocial outcomes such as distress, isolation, empowerment, and self-esteem, through positive and negative appraisals and coping efforts. However, whether an individual chooses to use OSGs as a source of support depends on their attitudes about OSGs, how well the OSG fits them and their needs, and their perceptions of control. In addition, whether users experience positive or negative outcomes depends on specific mechanisms

XSL•FO
RenderX

that occur within OSGs that influence their reactions and reasoning. These underlying mechanisms are influenced by contextual aspects of the communication technology, group composition and dynamics, and nature and content posted in OSGs. Overall, this study confirms the potential benefits and pitfalls of OSGs for patients with cancer and survivors of cancer. It also highlights the complexity of factors that influence whether an individual uses and benefits from an OSG. These findings validate and extend the multitheory framework proposed by Bender et al [28] on the use and effects of breast cancer OSGs, resulting in a refined program theory to explain how, why, for whom, and in what contexts OSGs work, visually depicted in Figure 2.

What Influences Use of OSGs as a Coping Strategy?

Patients with cancer and survivors of cancer are driven to OSGs to address a specific need. Drawing on TAM [43], our findings demonstrate that the decision to use a cancer OSG is influenced by its perceived usefulness and ease of use. The TAM-3 [213] posits that perceived usefulness is determined by social influence (eg, subjective norm and image) and job performance characteristics (eg, job relevance, output quality, and tangible results), while perceived ease of use is determined by control beliefs (eg, computer self-efficacy, perceptions of external control, computer anxiety, and computer playfulness) and usability characteristics (eg, objective usability and perceived enjoyment).

Perceived usefulness is the strongest determinant of OSG use, shaped by the OSG's ability to meet a specific need, its perceived trustworthiness (eg, quality, currency, and empathic nature of user-generated posts and moderation), and its cultural acceptability and sensitivity. Cross-sectional surveys across diverse patients with cancer and survivors of cancer show that unmet supportive care needs [15] and distress [214] predict OSG use. Qualitative research across diverse health-related OSGs has shown that once needs are met or if needs are not met, patients and survivors withdraw from OSGs [28,51,80], and as time passes, online help-seeking behaviors decrease [144]. As Massimi et al [80] have suggested, departure from an OSG is not necessarily a failing of the site's design but potentially a logical reaction to changing life circumstances. OSGs must demonstrate cultural sensitivity and competence to be appealing, relevant, and effective for diverse patients [84]. The benefits of OSGs are enhanced when users share a common language, culture, and life experience, and when moderators are trained in cultural sensitivity and competency [85].

Perceived ease-of-use influences OSG adoption, but to a lesser extent than perceived usefulness. Control beliefs (including eHealth literacy [212]) and objective usability are important determinants of ease-of-use. For example, patients with cancer who were confident using the internet were 4 times more likely to use the internet and social media as a resource [214]. The anonymity and privacy of OSGs are frequently reported advantages that afford users control over their personal information and make users feel more comfortable sharing personal information [23,56]. Whereas privacy concerns contributed to a lack of control over personal information and discouraged participation in OSGs [151]. Furthermore, studies have shown that engagement declines and users leave when a site is frustrating to use [90]. For example, Holtz et al [91] found that conversion of in-person support groups to virtual (video-based) support groups was hindered by technology factors, including poor Wi-Fi connection, background noise, and users speaking over one another. Designing OSGs to ensure ease-of-use and protect privacy will enhance engagement and participation.

How Do OSGs Work to Influence Health Outcomes?

Whether patients with cancer and survivors of cancer benefit from OSGs depends on the mechanisms that occur in OSGs that influence their reactions and reasoning. According to the Transactional Theory of Stress and Coping by Lazarus and Folkman [41], how people interpret and appraise events is critical in determining the events' stressfulness. There are two types of threat appraisals: (1) primary, which involve judgments about whether the event is a threat; and (2) secondary, which involve evaluations of personal and social resources available to cope with the event. Support from peers who have experienced a similar stressor can lessen negative appraisals of events and promote active coping efforts, thereby reducing or buffering anxiety. Several studies in this review confirm that by participating in an OSG, people can learn that their feelings and experiences are normal and to be expected, thereby improving their appraisal of their situation. The validation provided by peers normalizes the experience and reduces the stress associated with the uncertainty of the condition and the stigma imposed by society. Further, people can obtain information on effective ways of coping, which enhances feelings of preparedness, control, and self-esteem.

However, the effects of social support in cancer OSGs will be greatest if the support provided meets an individual's needs. Optimal Matching Theory [215] posits that the benefits and deleterious effects of social support are due to the matching (or mismatching) of support in different contexts. If the desired type of support is provided, it will enhance psychological adjustment and well-being. For example, in a study of survivors of breast cancer, both misalignment of support desired and received, and unwanted support were associated with poor adjustment [216]. In another study, OSG users who received unwanted support fared worse, which likely led to feelings of helplessness, negatively impacting self-esteem [217]. According to Lazarus and Folkman [41], problem-focused coping, such as seeking information, is most effective when the stressor is controllable, whereas emotion-focused coping is more useful when the situation is not. However, studies have shown that coping strategies that enhance perceptions of control contribute to psychological adjustment regardless of the controllability of the stressor [218]. Collectively, these findings indicate that patients with cancer may require different types of social support at different times and highlight the importance of matching the support provided in OSGs with individuals' needs for maximum benefit.

Optimal Matching Theory [215] also suggests that aligning the profile of the support provider with that of the support seeker may enhance the effectiveness of the support provided. According to the Social Comparison Theory by Festinger [42],

XSL•FO
RenderX

people seek similar others in times of stress and uncertainty to judge the appropriateness of their thoughts, feelings, and behaviors. Not finding similar others or not being accepted by the group can enhance feelings of deviance or uniqueness, worsening anxiety and feelings of isolation. These experiences can cause people to withdraw from OSGs [51,53]. Shared qualities and experiences facilitate mutual identification, disclosure, and belonging because peers understand one another in a way that other network members cannot. The challenge in creating OSGs based on the similarity principle is that users are likely to vary on a host of sociodemographic, clinical, and lifestyle dimensions. OSGs could be better designed to facilitate optimal matching by increasing the specificity of forums and connecting people with similar others with search function optimization. Moderators could also be trained to facilitate group cohesion (sense of belonging, influence, mutual support, and identification with other group members [101,219]), which has been associated with better outcomes in professionally led support groups [220].

The variable effects of OSGs are likely due in part to the negative effects of social comparisons. First, lack of positive upward comparisons (eg, learning what to expect, how to cope, and that it is possible to get better) can increase anxiety and decrease perceptions of control [221]. This is likely why participants in unmoderated breast cancer OSG in a study by Salzer [29] experienced an increase in distress: it was comprised of newly diagnosed patients who had no opportunity to make positive upward comparisons with survivors of cancer. Second, negative social comparisons (eg, learning that you are not doing as well as everyone else or that it is possible to get worse) can increase anxiety, decrease self-esteem, and decrease perceptions of control [221]. For example, being exposed to patients in a breast cancer OSG who were doing poorly was anxiety-provoking and caused users to withdraw [48]. Similarly, loss of members within a lung cancer OSG evoked grief, distress, and concern for the future [145]. At the same time, patients with metastatic breast cancer felt silenced, marginalized, and helpless in mixed-stage OSGs, whereas in stage-specific OSGs, they were able to talk openly and felt understood [53]. While group cohesion can develop across diagnostically dissimilar professionally moderated support groups [219], these findings suggest that stage-specific groups might be more effective for unmoderated OSGs.

Importantly, our findings indicate that the nature and content of posts in OSGs may have the greatest effect on outcomes. Several studies of cancer-specific and mixed-cancer OSGs concluded that the most beneficial posts were those that involved the sharing of feelings and experiences [48,50,54,66,70,69,99,101,106,107,119,121,122,126-128,150,153,154,156,167,172,178-181,184,189,191,222]. Emotionally expressive writing is associated with positive psychosocial adjustment and well-being [223], and such disclosures foster feelings of connection and group cohesion [224]. However, expressing emotions is not enough to benefit maximally from an OSG. Several studies of breast cancer OSGs found that posts that express emotions and cognitive processing confer the most benefit for the poster and reader [150,184,189,191]. For example, posts in breast cancer OSGs that included a higher percentage of insightful disclosure words (eg, aware, know, realize, think, and understand) led to greater positive effects that were maintained over time [184,191]. Additionally, insightful disclosure posts were more likely to generate empathetic responses in an OSG for young adults with cancer [182]. Hence, posts that include emotional disclosures and cognitive processing are most helpful for the poster and the reader, and most likely to garner a helpful empathetic response, leading to further benefits. These findings highlight the importance of crafting messages for maximum benefit. However, for the health benefits to manifest, OSGs must provide a safe environment for disclosing emotions and stigmatizing experiences [224]. Community rules can establish acceptable behavior and foster a safe online environment, while anonymity and strong privacy controls can facilitate disclosure.

The extent to which an individual may benefit from OSGs also depends on how they participate, which is influenced by technology factors, group process, and coping style. To be most effective, OSGs need to be active to ensure users can access support when needed, enhancing perceptions of control. Hence, larger groups may be more effective because there are more active members to post and respond. However, ideal group size depends on modality, as live chats can become overwhelming with too many participants and posts [100]. While anonymity contributes to activity in OSGs by facilitating disclosure, it can also contribute to inappropriate behaviors [225]. In this context, people feel less restrained and, without visual cues, do not see the discomfort they may cause others [225]. OSGs that are active may also be more trustworthy because members can participate in regulating the accuracy of the posts. For example, in a breast cancer OSG, only 10 of 4600 posts were found to be false or potentially misleading, and 7 of them were corrected by OSG members within 4 hours of posting [226].

Ongoing investment in moderator training and facilitation guidelines covering conflict resolution, information verification, and cultural sensitivity is recommended [88,144,152]. Some studies report that patients with cancer prefer professionally moderated OSGs for fear of misinformation [99,151]. However, studies comparing peer vs professionally led breast cancer OSGs have found no differences in outcomes [165,196]. Some studies also suggest that people who participate actively in OSGs by posting benefit more than those who participate less actively or passively by lurking [106,191]. However, other studies found that lurkers benefit to a similar degree [227]. Variations in coping styles may be responsible for the variable effects of OSG participation. For example, frequent users of a breast cancer OSG who engaged in emotionally expressive coping (eg, actively expressed and processed emotions) experienced better psychological well-being than frequent users who approached their emotions less actively [106]. Engaging the community in implementing community guidelines and offering users tips on how to participate for maximum benefit could optimize the effects of cancer OSGs.

Who Is Most Likely to Benefit From OSGs and Why?

Overall, the evidence suggests that the people most likely to benefit from cancer OSGs are those who have unmet supportive care needs, perceive OSGs to be useful and easy to use, can find an OSG to which they feel connected, and that meets their needs.

XSL•FO
RenderX

This means that people who have greater illness uncertainty (eg, newly diagnosed, experiencing new symptoms, or advancing disease), have a complex medical condition, greater symptom distress or social stigma associated with their condition, and lack sufficient support from offline family and friends, health care professionals, and other patients with cancer, may benefit more from cancer OSGs. Individuals with altruistic support needs, who are doing and coping well and who want to give back by supporting other patients with cancer in OSGs, can also benefit from OSGs.

In addition, most studies to date have examined cancer OSGs for women with breast cancer (71/92, 77% of single cancer studies) or have involved women with breast cancer (46/50, 92% of mixed cancer studies). This may mean that there are more OSGs available for women with breast cancer than for other people with other cancer types. Hence, women affected by breast cancer may have an easier path to benefitting from OSGs at present because there are likely more options for them to choose from to meet their needs.

The findings of this review also suggest that an individual's coping style may impact whether they use and benefit from OSGs. Seeking information and support is reflective of an active coping style. Women are more likely to engage in active coping strategies and seek support than men [228], which may in part explain prior work demonstrating that women are more likely to use social media for cancer-related support than men [214]. Correspondingly, a population-based study of men with prostate cancer reported that active information seekers were 4.5 times more likely to use the internet as a source of information and support than those who were not [229].

Individuals with emotion-focused coping styles may also benefit more from OSGs, given that the most beneficial posts are those that express emotions and cognitive processing. However, studies of breast and prostate cancer OSGs have documented more information than emotionally supportive talk occurring in cancer OSGs [49,230]. Studies of breast, prostate, and mixed cancer OSGs have also identified gender differences, with some reporting that men are more likely to share information while women are more likely to share emotional support [131,166,230]. In contrast, the discourse analysis of ovarian and prostate cancer OSGs by Sullivan [49] concluded that both genders provide information and emotional support in OSGs, albeit in different ways. Men tended to engage in "report-talk," providing information in the form of personal case histories, and demonstrated empathy by ensuring the information shared was accurate. In contrast, women more often engaged in "rapport-talk," providing information and emotional support through the sharing of personal experiences and feelings.

However, not all patients with cancer seek support online or from OSGs specifically. Most studies included in this review reported that participants were college- or university-educated, age was variable, and income was inconsistently reported. Age, education, income, and broadband access remain well-known predictors of internet use [231]. In prior work involving patients with breast, testicular, and head and neck cancer in Canada, 5 variables correlated with OSG use: age, first spoken language, education, treatments received, and unmet supportive care needs [232-234]. In another study, confidence using the internet was associated with use of social media as a resource for cancer information and support [214]. This review similarly found that lower socioeconomic status may limit some individuals' access to OSGs due to a lack of stable housing, internet access, or eHealth literacy.

Lastly, it is unclear how and to what extent patients with cancer and survivors of cancer who identify with non-White racial and ethnic backgrounds benefit from OSGs. Most studies (101/168, 60.1%) did not specify the racial or ethnic background of study participants. Of those that did, most studies (51/67, 76%) reported that most study participants were White North American or European. These findings mirror those of a review of cancer OSGs conducted over 20 years ago [140], indicative of little progress toward understanding the use and effects of English-language cancer OSGs among racialized ethnic groups and different linguistic communities. There was also a lack of studies on the use and effects of OSGs among 2SLGBTQIA+ (2-spirit, lesbian, gay, bisexual, transgender, queer, intersex, or asexual) patients with cancer. Only 2 such studies were identified: 1 focused on the OSG experiences of sexual minority women with breast cancer [159], and the other focused more broadly on the OSG experiences of sexual minority women with cancer [235]. Both highlighted the importance of designing OSGs to promote inclusivity and acceptance of diverse sexual and gender identities.

Implications and Recommendations

This study confirms and extends the multitheory framework we initially proposed to explain the use and effects of breast cancer OSGs. In doing so, the findings demonstrate the value of social support and technology adoption theory in explaining how online support interventions influence health, and realist methodology in uncovering the mechanisms through which complex social interventions exert their effects. Further research, as detailed below, is needed to test the proposed program theory.

The review identified several gaps in the literature. Most of the literature to date has focused on the use and effects of cancer OSGs among White, cisgender, heterosexual women with breast cancer from North America or Europe. Further research is needed to better understand and explore the applicability of our findings among men, ethnically and linguistically diverse minorities and majorities, 2SLGBTQIA+ patients with cancer, and people diagnosed with other common cancers, rarer cancers, and niche cancer groups at different stages of the cancer continuum. In addition, as most studies (109/168, 64.9%) were performed in the United States, further research is needed to determine how OSGs work for people in non-English speaking countries.

There is also a need for more research on the nature of posts that confer the most benefit for diverse gender and ethnocultural groups, how OSGs can be configured and moderated to optimize supportive exchanges for maximal benefit, and interventions designed to maximize benefits. The quality of reporting on the moderation of OSGs is generally poor, leaving many unanswered questions with respect to the training and qualifications of moderators and the extent and style of moderation.

XSL•FO
RenderX

Lastly, further research is needed to explore the implications of new technologies such as virtual reality (VR) and artificial intelligence (AI) in OSGs. VR could remove physical barriers to support, enabling "face-to-face" interactions through avatars and offering an interactive experience, but despite its potential, VR adoption may be limited due to low familiarity and accessibility [144]. In a series of 3 studies, Leung et al [236-238] demonstrated the potential of an AI cofacilitator to recommend resources, detect distress, and facilitate group cohesion in synchronous cancer OSGs with promising findings. The use of AI to support online content moderation raises ethical concerns regarding privacy, free speech, and bias [239]. Guidelines are needed for the use of AI in OSGs to balance the potential benefits with ethical considerations to ensure safe and inclusive online environments.

Strengths and Limitations

We followed a rigorous and systematic procedure to synthesize and analyze a substantial body of research using best practices in realist reviews, with guidance from an expert in realist methodologies and 5 categories of knowledge users. The findings consolidate and confirm prior research and provide new insight into the mechanisms underlying the effectiveness of OSGs and practical implications for improvement. However, as with any review method, the resulting synthesis is only as good as the primary data on which it is built. We limited our search to studies published in English, which in most cases focused on English-language OSGs. Another limitation we encountered in our review was that most studies of OSGs focused on women with breast cancer and included a disproportionate number of White, well-educated, and above-average-income earners. The predominance of breast cancer-related data may limit the generalizability of findings; however, using a realist approach enabled us to draw on data across cancer types, which showed commonalities. Second, many studies did not describe the OSGs in sufficient detail, including the technology used, the characteristics of the participants, the training and qualifications of the moderator, and the moderation strategy, and did not follow the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) reporting guidelines [240]. While some studies predate these reporting guidelines, many did not. To improve reporting of OSG interventions, we recommend updating the CONSORT-EHEALTH guidelines to capture the contextual aspects of social-media-based interventions such as OSGs that influence their effects.

Conclusions

OSGs can be an effective source of support for patients with cancer and survivors of cancer that can address many supportive care needs and improve psychosocial well-being. However, outcomes depend on specific contexts and mechanisms that impact how well OSGs meet the needs of users. To maximize the effectiveness of OSGs, we recommend optimizing the design and implementation of OSGs to (1) help assess fit and address specific needs; (2) demonstrate trustworthiness; (3) enhance anonymity and control, and protect privacy; (4) enhance ease-of-use; (5) support connection and belonging; (6) encourage activity; (7) enhance the nature of content shared to boost therapeutic effects; and (8) monitor and adjust design and management strategies to optimize effectiveness.

Acknowledgments

We thank Jessie McGowan, PhD, AHIP (Adjunct professor, School of Epidemiology and Public Health, University of Ottawa), for the peer review of the MEDLINE search strategy. Artificial intelligence was not used.

Funding

This study was funded by the Canadian Institutes of Health Research (grant #169403). The funder had no involvement in this study's design, data collection, analysis, interpretation, or the writing of this paper.

Authors' Contributions

JLB, GW, AT, AC, HP, OP, MJE, RW, VJF, JP, CS, CF, HW, and CY contributed to the conceptualization of this project. JLB, SS, SB, LRS, and ME contributed to data curation. JLB, SS, GW, LRS, SB, AT, AC, HP, OP, MJE, RW, VJF, JP, CS, CF, HW, ME, and CY contributed to formal analysis, contributed to study investigation, contributed to the methodology, contributed to data validation, and prepared the original draft. reviewed and edited this paper. JLB acquired funding, contributed to resources and software, and supervised this study. JLB, SS, and SB were involved in project administration.

Conflicts of Interest

JLB is a section editor of *JMIR Cancer* at the time of publication. Author AT is a Tier 1 Canada Research Chair in Knowledge Synthesis for Knowledge Users.

Multimedia Appendix 1
MEDLINE search strategy.
[DOCX File, 34 KB - jmir_v28i1e77445_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Study designs and demographics.
[DOCX File, 104 KB - jmir_v28i1e77445_app2.docx]

Multimedia Appendix 3
OSG characteristics. OSG: online support group.
[DOCX File, 67 KB - jmir_v28i1e77445_app3.docx]

Checklist 1
RAMESES checklist.
[DOC File, 47 KB - jmir_v28i1e77445_app4.doc]

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74(3):229-263. [doi: 10.3322/caac.21834] [Medline: 38572751]
2. Stein KD, Syrjala KL, Andrykowski MA. Physical and psychological long-term and late effects of cancer. Cancer 2008 Jun 1;112(11 Suppl):2577-2592. [doi: 10.1002/cncr.23448] [Medline: 18428205]
3. Pergolotti M, Deal AM, Williams GR, et al. Activities, function, and health-related quality of life (HRQOL) of older adults with cancer. J Geriatr Oncol 2017 Jul;8(4):249-254. [doi: 10.1016/j.jgo.2017.02.009] [Medline: 28285980]
4. Shakeel S, Tung J, Rahal R, Finley C. Evaluation of factors associated with unmet needs in adult cancer survivors in Canada. JAMA Netw Open 2020 Mar 2;3(3):e200506. [doi: 10.1001/jamanetworkopen.2020.0506] [Medline: 32142127]
5. Hewitt M, Greenfield S, Stovall E. From Cancer Patient to Cancer Survivor: Lost in Transition: National Academies Press; 2005.
6. Nekhlyudov L, Ganz PA, Arora NK, Rowland JH. Going beyond being lost in transition: a decade of progress in cancer survivorship. JCO 2017 Jun 20;35(18):1978-1981. [doi: 10.1200/JCO.2016.72.1373]
7. Neo J, Fettes L, Gao W, Higginson IJ, Maddocks M. Disability in activities of daily living among adults with cancer: a systematic review and meta-analysis. Cancer Treat Rev 2017 Dec;61:94-106. [doi: 10.1016/j.ctrv.2017.10.006]
8. Weaver KE, Forsythe LP, Reeve BB, et al. Mental and physical health–related quality of life among U.S. cancer survivors: population estimates from the 2010 National Health Interview Survey. Cancer Epidemiol Biomarkers Prev 2012 Nov 1;21(11):2108-2117. [doi: 10.1158/1055-9965.EPI-12-0740]
9. Kent EE, Ambs A, Mitchell SA, Clauser SB, Smith AW, Hays RD. Health-related quality of life in older adult survivors of selected cancers: data from the SEER-MHOS linkage. Cancer 2015 Mar 1;121(5):758-765. [doi: 10.1002/cncr.29119] [Medline: 25369293]
10. Schulte FSM, Chalifour K, Eaton G, Garland SN. Quality of life among survivors of adolescent and young adult cancer in Canada: a Young Adults With Cancer in Their Prime (YACPRIME) study. Cancer 2021 Apr 15;127(8):1325-1333. [doi: 10.1002/cncr.33372] [Medline: 33332603]
11. Jones JM, Fitch M, Bongard J, et al. The needs and experiences of post-treatment adolescent and young adult cancer survivors. JCM 2020;9(5):1444. [doi: 10.3390/jcm9051444]
12. King AJL, Evans M, Moore THM, et al. Prostate cancer and supportive care: a systematic review and qualitative synthesis of men's experiences and unmet needs. Eur J Cancer Care (Engl) 2015 Sep;24(5):618-634. [doi: 10.1111/ecc.12286] [Medline: 25630851]
13. Khajoei R, Ilkhani M, Azadeh P, Zohari Anboohi S, Heshmati Nabavi F. Breast cancer survivors-supportive care needs: systematic review. BMJ Support Palliat Care 2023 Jun;13(2):143-153. [doi: 10.1136/spcare-2022-003931] [Medline: 36972985]
14. Hart NH, Crawford-Williams F, Crichton M, et al. Unmet supportive care needs of people with advanced cancer and their caregivers: a systematic scoping review. Crit Rev Oncol Hematol 2022 Aug;176:103728. [doi: 10.1016/j.critrevonc.2022.103728]
15. Bender J. Cancer virtual communities in the era of personalized medicine: perspectives on scope, use and impact. In: C EM, editor. Novel Applications of Virtual Communities in Healthcare Settings: IGI Global Scientific Publishing; 2018:56-92. [doi: 10.4018/978-1-5225-2958-3.ch003]
16. Preece J. Online communities: designing usability, supporting sociability. Ind Manage Data Syst 2000 Dec 1;100(9):459-460. [doi: 10.1108/imds.2000.100.9.459.3]
17. Bender JL, Jimenez-Marroquin MC, Jadad AR. Seeking support on facebook: a content analysis of breast cancer groups. J Med Internet Res 2011 Feb 4;13(1):e16. [doi: 10.2196/jmir.1560] [Medline: 21371990]
18. Wang J, Wei L. Fear and hope, bitter and sweet: emotion sharing of cancer community on Twitter. Soc Media + Soc 2020 Jan;6(1). [doi: 10.1177/2056305119897319]

XSL•FO
RenderX

19. Tripathi SD, Parker PD, Prabhu AV, Thomas K, Rodriguez A. An examination of patients and caregivers on Reddit navigating brain cancer: content analysis of the Brain Tumor Subreddit. JMIR Cancer 2022 Jun 22;8(2):e35324. [doi: 10.2196/35324] [Medline: 35731559]
20. Eysenbach G, Powell J, Englesakis M, Rizo C, Stern A. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. BMJ 2004 May 15;328(7449):1166. [doi: 10.1136/bmj.328.7449.1166] [Medline: 15142921]
21. Wright K. Social support within an on-line cancer community: an assessment of emotional support, perceptions of advantages and disadvantages, and motives for using the community from a communication perspective. J Appl Commun Res 2002 Jan;30(3):195-209. [doi: 10.1080/00909880216586]
22. Sharf BF. Communicating breast cancer on-line: support and empowerment on the internet. Women Health 1997;26(1):65-84. [doi: 10.1300/J013v26n01_05] [Medline: 9311100]
23. Broom A. The eMale: prostate cancer, masculinity and online support as a challenge to expertise. J Sociol (Melb) 2005;41(1):87-104. [doi: 10.1177/1440783305050965]
24. van Uden-Kraan CF, Drossaert CHC, Taal E, Shaw BR, Seydel ER, van de Laar M. Empowering processes and outcomes of participation in online support groups for patients with breast cancer, arthritis, or fibromyalgia. Qual Health Res 2008 Mar;18(3):405-417. [doi: 10.1177/1049732307313429] [Medline: 18235163]
25. Huber J, Muck T, Maatz P, et al. Face-to-face vs. online peer support groups for prostate cancer: a cross-sectional comparison study. J Cancer Surviv 2018 Feb;12(1):1-9. [doi: 10.1007/s11764-017-0633-0] [Medline: 28861845]
26. Vilhauer RP, McClintock MK, Matthews AK. Online support groups for women with metastatic breast cancer: a feasibility pilot study. J Psychosoc Oncol 2010;28(5):560-586. [doi: 10.1080/07347332.2010.504504] [Medline: 20730665]
27. Kaal SE, Husson O, van Dartel F, et al. Online support community for adolescents and young adults (AYAs) with cancer: user statistics, evaluation, and content analysis. Patient Prefer Adherence 2018;12:2615-2622. [doi: 10.2147/PPA.S171892] [Medline: 30584285]
28. Bender JL, Katz J, Ferris LE, Jadad AR. What is the role of online support from the perspective of facilitators of face-to-face support groups? A multi-method study of the use of breast cancer online communities. Patient Educ Couns 2013 Dec;93(3):472-479. [doi: 10.1016/j.pec.2013.07.009] [Medline: 23928354]
29. Salzer MS, Palmer SC, Kaplan K, et al. A randomized, controlled study of internet peer-to-peer interactions among women newly diagnosed with breast cancer. Psychooncology 2010 Apr;19(4):441-446. [doi: 10.1002/pon.1586] [Medline: 19484712]
30. Hoey LM, Ieropoli SC, White VM, Jefford M. Systematic review of peer-support programs for people with cancer. Patient Educ Couns 2008 Mar;70(3):315-337. [doi: 10.1016/j.pec.2007.11.016] [Medline: 18191527]
31. van Eenbergen MC, van de Poll-Franse LV, Heine P, Mols F. The impact of participation in online cancer communities on patient reported outcomes: systematic review. JMIR Cancer 2017 Sep 28;3(2):e15. [doi: 10.2196/cancer.7312] [Medline: 28958985]
32. McCaughan E, Parahoo K, Hueter I, Northouse L, Bradbury I. Online support groups for women with breast cancer. Cochrane Database Syst Rev 2017 Mar 10;3(3):CD011652. [doi: 10.1002/14651858.CD011652.pub2] [Medline: 28278559]
33. Pawson R, Greenhalgh T, Harvey G, Walshe K. Realist review - a new method of systematic review designed for complex policy interventions. J Health Serv Res Policy 2005 Jul;10(1_suppl):21-34. [doi: 10.1258/1355819054308530]
34. Wong G, Brennan N, Mattick K, Pearson M, Briscoe S, Papoutsi C. Interventions to improve antimicrobial prescribing of doctors in training: the IMPACT (Improving Antimicrobial Prescribing of Doctors in Training) realist review. BMJ Open 2015 Oct 22;5(10):e009059. [doi: 10.1136/bmjopen-2015-009059] [Medline: 26493460]
35. Ford JA, Turley R, Porter T, et al. Access to primary care for socio-economically disadvantaged older people in rural areas: a qualitative study. PLoS ONE 2018;13(3):e0193952. [doi: 10.1371/journal.pone.0193952] [Medline: 29509811]
36. Wong G, Greenhalgh T, Pawson R. Internet-based medical education: a realist review of what works, for whom and in what circumstances. BMC Med Educ 2010 Feb 2;10(12):12. [doi: 10.1186/1472-6920-10-12] [Medline: 20122253]
37. Duddy C, Wong G. Explaining variations in test ordering in primary care: protocol for a realist review. BMJ Open 2018 Sep 12;8(9):e023117. [doi: 10.1136/bmjopen-2018-023117] [Medline: 30209159]
38. Wong G, Greenhalgh T, Westhorp G, Buckingham J, Pawson R. RAMESES publication standards: realist syntheses. BMC Med 2013 Jan 29;11(1):23356726. [doi: 10.1186/1741-7015-11-21] [Medline: 23360677]
39. Bender JL, Babinski S, Wong G, et al. Establishing best practices in cancer online support groups: protocol for a realist review. BMJ Open 2021 Nov 3;11(11):e053916. [doi: 10.1136/bmjopen-2021-053916] [Medline: 34732498]
40. Guide to knowledge translation planning at CIHR: integrated and end-of-grant approaches. Canadian Institutes of Health Research. URL: https://cihr-irsc.gc.ca/e/documents/kt_lm_ktplan-en.pdf [accessed 2026-04-29]
41. Lazarus S, Folkman R. Stress, Appraisal and Coping: Springer Publishing Company; 1984.
42. Festinger L. A theory of social comparison processes. Hum Relat 1954 May;7(2):117-140. [doi: 10.1177/001872675400700202]
43. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
44. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]

XSL•FO
RenderX

45. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS Peer Review of Electronic Search Strategies: 2015 guideline statement. J Clin Epidemiol 2016 Jul;75:40-46. [doi: 10.1016/j.jclinepi.2016.01.021]
46. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 5;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
47. Duddy C, Wong G. Grand rounds in methodology: when are realist reviews useful, and what does a "good" realist review look like? BMJ Qual Saf 2023 Mar;32(3):173-180. [doi: 10.1136/bmjqs-2022-015236] [Medline: 36585019]
48. Hong QN, Gonzalez-Reyes A, Pluye P. Improving the usefulness of a tool for appraising the quality of qualitative, quantitative and mixed methods studies, the Mixed Methods Appraisal Tool (MMAT). J Eval Clin Pract 2018 Jun;24(3):459-467. [doi: 10.1111/jep.12884] [Medline: 29464873]
49. Sullivan CF. Gendered cybersupport: a thematic analysis of two online cancer support groups. J Health Psychol 2003 Jan;8(1):83-104. [doi: 10.1177/1359105303008001446] [Medline: 22113903]
50. Griffiths C, Panteli N, Brunton D, Marder B, Williamson H. Designing and evaluating the acceptability of Realshare: an online support community for teenagers and young adults with cancer. J Health Psychol 2015 Dec;20(12):1589-1601. [doi: 10.1177/1359105313519154] [Medline: 24470526]
51. Sandaunet AG. The challenge of fitting in: non‐participation and withdrawal from an online self‐help group for breast cancer patients. Sociol Health Illness 2008 Jan;30(1):131-144. [doi: 10.1111/j.1467-9566.2007.01041.x]
52. Im EO, Chee W, Lim HJ, Liu Y, Guevara E, Kim KS. Patients' attitudes toward internet cancer support groups. Oncol Nurs Forum 2007 Jan 1;34(3):705-712. [doi: 10.1188/07.ONF.705-712]
53. Vilhauer RP. 'Them' and 'us': the experiences of women with metastatic disease in mixed-stage versus stage-specific breast cancer support groups. Psychol Health 2011 Jun;26(6):781-797. [doi: 10.1080/08870446.2010.496853]
54. Chiu YC, Hsieh YL. Communication online with fellow cancer patients: writing to be remembered, gain strength, and find survivors. J Health Psychol 2013 Dec;18(12):1572-1581. [doi: 10.1177/1359105312465915] [Medline: 23221492]
55. Donovan E, Martin SR, Seidman LC, et al. A mobile-based mindfulness and social support program for adolescents and young adults with sarcoma: development and pilot testing. JMIR mHealth uHealth 2019;7(3):e10921. [doi: 10.2196/10921]
56. Shaw BR, McTavish F, Hawkins R, Gustafson DH, Pingree S. Experiences of women with breast cancer: exchanging social support over the CHESS computer network. J Health Commun 2000;5(2):135-159. [doi: 10.1080/108107300406866] [Medline: 11010346]
57. Stephen J, Collie K, McLeod D, et al. Talking with text: communication in therapist-led, live chat cancer support groups. Soc Sci Med 2014 Mar;104:178-186. [doi: 10.1016/j.socscimed.2013.12.001]
58. Vilhauer RP. Computer-mediated and face-to-face communication in metastatic cancer support groups. Pall Supp Care 2014 Aug;12(4):287-297. [doi: 10.1017/S1478951513000126]
59. Ure C, Cooper-Ryan AM, Condie J, Galpin A. Exploring strategies for using social media to self-manage health care when living with and beyond breast cancer: in-depth qualitative study. J Med Internet Res 2020;22(5):e16902. [doi: 10.2196/16902]
60. Owen JE, Bantum EO, Golant M. Benefits and challenges experienced by professional facilitators of online support groups for cancer survivors. Psychooncology 2009 Feb;18(2):144-155. [doi: 10.1002/pon.1374] [Medline: 18546089]
61. Høybye MT, Johansen C, Tjørnhøj-Thomsen T. Online interaction. Effects of storytelling in an internet breast cancer support group. Psychooncology 2005 Mar;14(3):211-220. [doi: 10.1002/pon.837]
62. Yoo W, Chih MY, Kwon MW, et al. Predictors of the change in the expression of emotional support within an online breast cancer support group: a longitudinal study. Patient Educ Couns 2013 Jan;90(1):88-95. [doi: 10.1016/j.pec.2012.10.001] [Medline: 23122429]
63. Leimeister JM, Schweizer K, Leimeister S, et al. Do virtual communities matter for the social support of patients? Antecedents and effects of virtual relationships in online communities. Inf Technol People 2008;21:350-374. [doi: 10.1108/09593840810919671]
64. Malloch YZ, Taylor LD. Emotional self-disclosure in online breast cancer support groups: examining theme, reciprocity, and linguistic style matching. Health Commun 2019 Jun;34(7):764-773. [doi: 10.1080/10410236.2018.1434737] [Medline: 29400555]
65. Cabling ML, Turner JW, Hurtado-de-Mendoza A, et al. Sentiment analysis of an online breast cancer support group: communicating about tamoxifen. Health Commun 2018 Sep;33(9):1158-1165. [doi: 10.1080/10410236.2017.1339370] [Medline: 28678549]
66. Lieberman MA, Golant M, Winzelberg A, Mctavish F, Gustafson DH. Comparisons: professionally-directed and self-directed internet groups for women with breast cancer. Int J Self Help Self Care 2005;2(3):219-235. [doi: 10.2190/GE85-J31W-XJV7-LB9L]
67. Høybye MT, Dalton SO, Christensen J, Ross L, Kuhn KG, Johansen C. Social and psychological determinants of participation in internet-based cancer support groups. Support Care Cancer 2010 May;18(5):553-560. [doi: 10.1007/s00520-009-0683-6] [Medline: 19579037]
68. Dickerson SS, Boehmke M, Ogle C, Brown JK. Seeking and managing hope: patients' experiences using the internet for cancer care. Oncol Nurs Forum 2006 Jan 1;33(1):E8-17. [doi: 10.1188/06.ONF.E8-E17] [Medline: 16470231]

XSL•FO
RenderX

69. Harkin LJ, Beaver K, Dey P, Choong KA. Secret groups and open forums: defining online support communities from the perspective of people affected by cancer. Digit HEALTH 2020;6:2055207619898993. [doi: 10.1177/2055207619898993] [Medline: 32010450]
70. Im EO. Online support of patients and survivors of cancer. Semin Oncol Nurs 2011 Aug;27(3):229-236. [doi: 10.1016/j.soncn.2011.04.008] [Medline: 21783014]
71. Lovatt M, Bath PA, Ellis J. Development of trust in an online breast cancer forum: a qualitative study. J Med Internet Res 2017 May 23;19(5):e175. [doi: 10.2196/jmir.7471] [Medline: 28536093]
72. Walther JB, Pingree S, Hawkins RP, Buller DB. Attributes of interactive online health information systems. J Med Internet Res 2005 Jul 1;7(3):e33. [doi: 10.2196/jmir.7.3.e33] [Medline: 15998624]
73. Adikari A, de Silva D, Ranasinghe WKB, et al. Can online support groups address psychological morbidity of cancer patients? An artificial intelligence based investigation of prostate cancer trajectories. PLOS ONE 2020;15(3):e0229361. [doi: 10.1371/journal.pone.0229361] [Medline: 32130256]
74. Baker TB, Hawkins R, Pingree S, et al. Optimizing eHealth breast cancer interventions: which types of eHealth services are effective? Transl Behav Med 2011 Mar;1(1):134-145. [doi: 10.1007/s13142-010-0004-0] [Medline: 21709810]
75. Namkoong K, Shah DV, Han JY, et al. Expression and reception of treatment information in breast cancer support groups: how health self-efficacy moderates effects on emotional well-being. Patient Educ Couns 2010 Dec;81:S41-S47. [doi: 10.1016/j.pec.2010.09.009]
76. Park KA, Eum SY, Oh H, et al. Factors affecting online health community participation behavior in patients with thyroid cancer. PLoS ONE 2020;15(6):e0235056. [doi: 10.1371/journal.pone.0235056]
77. Setoyama Y, Yamazaki Y, Namayama K. Benefits of peer support in online Japanese breast cancer communities: differences between lurkers and posters. J Med Internet Res 2011 Dec 29;13(4):e122. [doi: 10.2196/jmir.1696] [Medline: 22204869]
78. Shaw BR, Hawkins RP, McTavish FM, Gustafson DH. Communicating about self and others within an online support group for women with breast cancer and subsequent outcomes. J Health Psychol 2008 Oct;13(7):930-939. [doi: 10.1177/1359105308095067]
79. van Uden-Kraan CF, Drossaert CHC, Taal E, Seydel ER, van de Laar M. Participation in online patient support groups endorses patients' empowerment. Patient Educ Couns 2009 Jan;74(1):61-69. [doi: 10.1016/j.pec.2008.07.044] [Medline: 18778909]
80. Massimi M, Bender JL, Witteman HO, et al. Life transitions and online health communities: reflecting on adoption, use, and disengagement. Assoc Comput Mach 2014:1491-1501. [doi: 10.1145/2531602.2531622]
81. Sandaunet AG. A space for suffering? Communicating breast cancer in an online self-help context. Qual Health Res 2008 Dec;18(12):1631-1641. [doi: 10.1177/1049732308327076] [Medline: 18955462]
82. Dickerson SS, Reinhart A, Boemhke M, Akhu-Zaheya L. Cancer as a problem to be solved: internet use and provider communication by men with cancer. Comput Inform Nurs 2011 Jul;29(7):388-395. [doi: 10.1097/NCN.0b013e3181f9ddb1] [Medline: 20975535]
83. Lieberman MA. Self-management in online self-help groups for breast cancer patients: finding the right group, a speculative hypothesis. Int J Self Help Self Care 2004;2(4):313-328. [doi: 10.2190/0009-9E12-HT6D-3J13]
84. Im EO, Lee B, Chee W. The use of internet cancer support groups by Asian Americans and White Americans living with cancer. J Transcult Nurs 2011 Oct;22(4):386-396. [doi: 10.1177/1043659611414142]
85. Im EO, Ji X, Zhang J, et al. Issues in developing and evaluating a culturally tailored internet cancer support group. CIN 2016;34(10):462-469. [doi: 10.1097/CIN.0000000000000261]
86. Im EO, Lee B, Chee W. Shielded from the real world: perspectives on internet cancer support groups by Asian Americans. Cancer Nurs 2010;33(3):E10-E20. [doi: 10.1097/NCC.0b013e3181c8e5d5] [Medline: 20357657]
87. Houlihan MC, Tariman JD. Comparison of outcome measures for traditional and online support groups for breast cancer patients: an integrative literature review. J Adv Pract Oncol 2017;8(4):348-359. [Medline: 30018841]
88. Abioye O, Kiel L, Kaufman R, Florez N. Cancer health disparities in minority communities: peer support networks can bridge the gap. Cancer Causes Control 2024 Nov;35(11):1407-1411. [doi: 10.1007/s10552-024-01903-2] [Medline: 39117791]
89. Suarez NRE, Morrow AS, LaVecchia CM, et al. Connected and supported: a scoping review of how online communities provide social support for breast cancer survivors. J Cancer Surviv 2026 Apr;20(2):415-426. [doi: 10.1007/s11764-024-01660-w] [Medline: 39196462]
90. Gorlick A, Bantum EO, Owen JE. Internet-based interventions for cancer-related distress: exploring the experiences of those whose needs are not met. Psychooncology 2014 Apr;23(4):452-458. [doi: 10.1002/pon.3443] [Medline: 24243756]
91. Holtz BE, Mitchell KM, Strand D, Hirko K. Perceptions of telehealth-based cancer support groups at a rural community oncology program. J Cancer Educ 2024 Aug;39(4):418-425. [doi: 10.1007/s13187-024-02428-7] [Medline: 38539005]
92. Beatty L, Scott K. Examining eHealth use as a coping strategy for cancer-adjustment: an application of the chronic illness model. EJAP 2013;9(1):9-18. [doi: 10.7790/ejap.v9i1.349]
93. Chen Z, Koh PW, Ritter PL, Lorig K, Bantum EO, Saria S. Dissecting an online intervention for cancer survivors: four exploratory analyses of internet engagement and its effects on health status and health behaviors. Health Educ Behav 2015 Feb;42(1):32-45. [doi: 10.1177/1090198114550822] [Medline: 25288489]

XSL•FO
RenderX

94. Han JY, Kim JH, Yoon HJ, Shim M, McTavish FM, Gustafson DH. Social and psychological determinants of levels of engagement with an online breast cancer support group: posters, lurkers, and nonusers. J Health Commun 2012;17(3):356-371. [doi: 10.1080/10810730.2011.585696] [Medline: 22085215]
95. Han JY, Hou J, Kim E, Gustafson DH. Lurking as an active participation process: a longitudinal investigation of engagement with an online cancer support group. Health Commun 2014;29(9):911-923. [doi: 10.1080/10410236.2013.816911] [Medline: 24345206]
96. Im EO, Chee W, Liu Y, et al. Characteristics of cancer patients in internet cancer support groups. Comput Inf Nurs 2007;25(6):334-343. [doi: 10.1097/01.NCN.0000299655.21401.9d] [Medline: 18000430]
97. McLaughlin M, Nam Y, Gould J, et al. A videosharing social networking intervention for young adult cancer survivors. Comput Human Behav 2012 Mar;28(2):631-641. [doi: 10.1016/j.chb.2011.11.009]
98. Owen JE, Bantum EO, Gorlick A, Stanton AL. Engagement with a social networking intervention for cancer-related distress. Ann Behav Med 2015 Apr;49(2):154-164. [doi: 10.1007/s12160-014-9643-6] [Medline: 25209353]
99. Owen JE, Klapow JC, Roth DL, et al. Randomized pilot of a self-guided internet coping group for women with early-stage breast cancer. Ann Behav Med 2005 Aug;30(1):54-64. [doi: 10.1207/s15324796abm3001_7] [Medline: 16097906]
100. Van Uden-Kraan CF, Drossaert CHC, Taal E, Smit WM, Bernelot Moens HJ, Van de Laar M. Determinants of engagement in face-to-face and online patient support groups. J Med Internet Res 2011 Dec 7;13(4):e106. [doi: 10.2196/jmir.1718] [Medline: 22155649]
101. Vilhauer RP. Perceived benefits of online support groups for women with metastatic breast cancer. Women Health 2009 Oct 19;49(5):381-404. [doi: 10.1080/03630240903238719]
102. Wootten AC, Abbott JAM, Meyer D, et al. Preliminary results of a randomised controlled trial of an online psychological intervention to reduce distress in men treated for localised prostate cancer. Eur Urol 2015 Sep;68(3):471-479. [doi: 10.1016/j.eururo.2014.10.024] [Medline: 25454611]
103. Feng X, Hu Y, Pfaff H, Liu S, Wang H, Qi Z. The determinants of help-seeking behaviors among cancer patients in online health communities: evidence from China. Int J Med Inform 2025 Mar;195:105767. [doi: 10.1016/j.ijmedinf.2024.105767] [Medline: 39721114]
104. Jo W, Jang SH, Shin EK. Stage distinctive communication networks of the online breast cancer community. Sci Rep 2023;13(1):1726. [doi: 10.1038/s41598-023-28892-7]
105. Sanger S, Duffin S, Gough RE, Bath PA. Use of online health forums by people living with breast cancer during the COVID-19 pandemic: thematic analysis. JMIR Cancer 2023 Feb 7;9:e42783. [doi: 10.2196/42783] [Medline: 36473015]
106. Batenburg A, Das E. Emotional approach coping and the effects of online peer-led support group participation among patients with breast cancer: a longitudinal study. J Med Internet Res 2014;16(11):e256. [doi: 10.2196/jmir.3517]
107. Batenburg A, Das E. Emotional coping differences among breast cancer patients from an online support group: a cross-sectional study. J Med Internet Res 2014;16(2):e28. [doi: 10.2196/jmir.2831]
108. Kaka A, Shama W, Lucchetta S, et al. Virtual support groups in pediatric hematology/oncology during COVID‐19: lessons learned from the SickKids experience. Pediatr Blood Cancer 2021 Dec;68(12):e29378. [doi: 10.1002/pbc.29378]
109. Kim E, Scheufele DA, Han JY, Shah D. Opinion leaders in online cancer support groups: an investigation of their antecedents and consequences. Health Commun 2017 Feb;32(2):142-151. [doi: 10.1080/10410236.2015.1110005]
110. Lazard AJ, Collins MKR, Hedrick A, et al. Using social media for peer-to-peer cancer support: interviews with young adults with cancer. JMIR Cancer 2021 Sep 2;7(3):e28234. [doi: 10.2196/28234] [Medline: 34473063]
111. Lazard AJ, Collins MKR, Hedrick A, et al. Initiation and changes in use of social media for peer support among young adult cancer patients and survivors. Psychooncology 2021 Nov;30(11):1859-1865. [doi: 10.1002/pon.5758] [Medline: 34165848]
112. Meng J, Rains SA, An Z. How cancer patients benefit from support networks offline and online: extending the model of structural-to-functional support. Health Commun 2021 Feb;36(2):198-206. [doi: 10.1080/10410236.2019.1673947] [Medline: 31601113]
113. Namkoong K, Shah DV, Gustafson DH. Offline social relationships and online cancer communication: effects of social and family support on online social network building. Health Commun 2017 Nov 2;32(11):1422-1429. [doi: 10.1080/10410236.2016.1230808]
114. Nau E. An online group for young women with breast cancer: sparking online engagement. Soc Work Groups 2022 Oct 2;45(3-4):257-267. [doi: 10.1080/01609513.2021.1907996]
115. Shaw BR, Hawkins R, Arora N, McTAVISH F, Pingree S, Gustafson DH. An exploratory study of predictors of participation in a computer support group for women with breast cancer. CIN: Comput, Inf, Nurs 2006 Jan;24(1):18-27. [doi: 10.1097/00024665-200601000-00007]
116. Turner JW, Grube JA, Meyers J. Developing an optimal match within online communities: an exploration of CMC support communities and traditional support. J Commun 2001 Jun 1;51(2):231-251. [doi: 10.1111/j.1460-2466.2001.tb02879.x]
117. Willard VW, Brasher S, Harman JL, Jurbergs N. Virtual group activities as a new platform for socialization in children with pediatric cancer: a case series from the COVID-19 pandemic. J Pediatr Hematol Oncol 2022 Nov 1;44(8):462-464. [doi: 10.1097/MPH.0000000000002394] [Medline: 36301047]

XSL•FO
RenderX

118. Yoo W, Shah DV, Shaw BR, et al. The role of the family environment and computer-mediated social support on breast cancer patients' coping strategies. J Health Commun 2014 Sep;19(9):981-998. [doi: 10.1080/10810730.2013.864723] [Medline: 24511907]
119. Garg R, Rebić N, De Vera MA. Information needs about cancer treatment, fertility, and pregnancy: qualitative descriptive study of Reddit threads. JMIR Cancer 2020 Dec 2;6(2):e17771. [doi: 10.2196/17771] [Medline: 33263547]
120. Gill PS, Whisnant B. A qualitative assessment of an online support community for ovarian cancer patients. Patient Relat Outcome Meas 2012;3:51-58. [doi: 10.2147/PROM.S36034] [Medline: 23185122]
121. Hargreaves S, Bath PA, Duffin S, Ellis J. Sharing and empathy in digital spaces: qualitative study of online health forums for breast cancer and motor neuron disease (amyotrophic lateral sclerosis). J Med Internet Res 2018 Jun 14;20(6):e222. [doi: 10.2196/jmir.9709] [Medline: 29903695]
122. Lobchuk M, McClement S, Rigney M, Copeland A, Bayrampour H. A qualitative analysis of "naturalistic" conversations in a peer-led online support community for lung cancer. Cancer Nurs 2015;38(5):E21-E31. [doi: 10.1097/NCC.0000000000000207] [Medline: 25254411]
123. Rubenstein EL. "They are always there for me": the convergence of social support and information in an online breast cancer community. Asso Inf Sci Technol 2015 Jul;66(7):1418-1430. [doi: 10.1002/asi.23263]
124. Sanders R, Araujo TB, Vliegenthart R, van Eenbergen MC, van Weert JCM, Linn AJ. Patients' convergence of mass and interpersonal communication on an online forum: hybrid methods analysis. J Med Internet Res 2020 Oct 19;22(10):e18303. [doi: 10.2196/18303] [Medline: 33074160]
125. Yli-Uotila T, Rantanen A, Suominen T. Motives of cancer patients for using the internet to seek social support. Eur J Cancer Care (Engl) 2013 Mar;22(2):261-271. [doi: 10.1111/ecc.12025] [Medline: 23320398]
126. Street AF, Wakelin K, Hordern A, Bruce N, Horey D. Dignity and deferral narratives as strategies in facilitated technology-based support groups for people with advanced cancer. Nurs Res Pract 2012;2012:647836. [doi: 10.1155/2012/647836] [Medline: 22530115]
127. Wakelin K, Street AF. An online expressive writing group for people affected by cancer: a virtual third place. Aust Soc Work 2015 Apr 3;68(2):198-211. [doi: 10.1080/0312407X.2014.991336]
128. Wen KY, McTavish F, Kreps G, Wise M, Gustafson D. From diagnosis to death: a case study of coping with breast cancer as seen through online discussion group messages. J Comput Mediat Commun 2011 Jan;16(2):331-361. [doi: 10.1111/j.1083-6101.2011.01542.x] [Medline: 23055657]
129. Broom A. Virtually he@lthy: the impact of internet use on disease experience and the doctor-patient relationship. Qual Health Res 2005 Mar;15(3):325-345. [doi: 10.1177/1049732304272916] [Medline: 15761103]
130. Pyle D, Tehan G, Lamont-Mills A, Chambers SK. Exploring predictors of and barriers to online prostate cancer community use: a cross-sectional survey of users and non-users. Psychooncology 2022 May;31(5):824-831. [doi: 10.1002/pon.5867] [Medline: 34967071]
131. Klemm P, Hurst M, Dearholt SL, Trone SR. Gender differences on Internet cancer support groups. Comput Nurs 1999;17(2):65-72. [Medline: 10194883]
132. Lepore SJ, Rincon MA, Buzaglo JS, et al. Digital literacy linked to engagement and psychological benefits among breast cancer survivors in internet - based peer support groups. Eur J Cancer Care 2019 Jul;28(4):e13134. [doi: 10.1111/ecc.13134]
133. Xu Y, Testerman LS, Owen JE, Bantum EO, Thornton AA, Stanton AL. Modeling intention to participate in face-to-face and online lung cancer support groups. Psychooncology 2014 May;23(5):555-561. [doi: 10.1002/pon.3449]
134. Changrani J, Lieberman M, Golant M, et al. Online cancer support groups: experiences with underserved immigrant Latinas. Prim psychiatry 2008;15:55-62 [FREE Full text]
135. Fallon EA, Driscoll D, Smith TS, Richardson K, Portier K. Description, characterization, and evaluation of an online social networking community: the American Cancer Society's Cancer Survivors Network®. J Cancer Surviv 2018 Oct;12(5):691-701. [doi: 10.1007/s11764-018-0706-8]
136. Fogel J, Ribisl KM, Morgan PD, Humphreys K, Lyons EJ. Underrepresentation of African Americans in online cancer support groups. J Natl Med Assoc 2008 Jun;100(6):705-712. [doi: 10.1016/s0027-9684(15)31346-8] [Medline: 18595573]
137. Im EO, Chee W. The use of internet cancer support groups by ethnic minorities. J Transcult Nurs 2008 Jan;19(1):74-82. [doi: 10.1177/1043659607309140] [Medline: 18217235]
138. Lepore SJ, Buzaglo JS, Lieberman MA, Golant M, Greener JR, Davey A. Comparing standard versus prosocial internet support groups for patients with breast cancer: a randomized controlled trial of the helper therapy principle. J Clin Oncol 2014 Dec 20;32(36):4081-4086. [doi: 10.1200/JCO.2014.57.0093] [Medline: 25403218]
139. Rodgers S, Chen Q. Internet community group participation: psychosocial benefits for women with breast cancer. J Comput Mediat Commun 2005 Jul;10(4):JCMC1047. [doi: 10.1111/j.1083-6101.2005.tb00268.x]
140. Klemm P, Bunnell D, Cullen M, Soneji R, Gibbons P, Holecek A. Online cancer support groups: a review of the research literature. Comput Inf Nurs 2003;21(3):136-142. [doi: 10.1097/00024665-200305000-00010] [Medline: 12792194]
141. Kim E, Han JY, Shah D, et al. Predictors of supportive message expression and reception in an interactive cancer communication system. J Health Commun 2011 Nov;16(10):1106-1121. [doi: 10.1080/10810730.2011.571337] [Medline: 22070449]

XSL•FO
RenderX

142. Seçkin G. Satisfaction with health status among cyber patients: testing a mediation model of electronic coping support. Behav Inf Technol 2013 Jan;32(1):91-101. [doi: 10.1080/0144929X.2011.603359]
143. Rising CJ, Bol N, Burke-Garcia A, Rains S, Wright KB. Perceived stress in online prostate cancer community participants: examining relationships with stigmatization, social support network preference, and social support seeking. J Health Commun 2017 Jun;22(6):469-476. [doi: 10.1080/10810730.2017.1304471] [Medline: 28414570]
144. Ahn J, Lee KE. Experiences of peer support activities and the need for a metaverse-based program in young women with breast cancer: a qualitative study. Asia Pac J Oncol Nurs 2023 Jul;10(7):100253. [doi: 10.1016/j.apjon.2023.100253] [Medline: 37448532]
145. Walsh CA, Currin-McCulloch J, Faris NR, Nguyen TST, Al Achkar M. “Living with loss”: a qualitative exploration of existential fears among people with advanced lung cancer in online lung cancer support groups. Palliat Support Care 2024 Feb 29;2024:1-6. [doi: 10.1017/S147895152400004X] [Medline: 38420708]
146. Donovan EE, LeFebvre L, Tardif S, Brown LE, Love B. Patterns of social support communicated in response to expressions of uncertainty in an online community of young adults with cancer. J Appl Commun Res 2014 Oct 2;42(4):432-455. [doi: 10.1080/00909882.2014.929725]
147. Eysenbach G. The impact of the internet on cancer outcomes. CA Cancer J Clin 2003;53(6):356-371. [doi: 10.3322/canjclin.53.6.356] [Medline: 15224975]
148. Gupta T, Schapira L. Online communities as sources of peer support for people living with cancer: a commentary. J Oncol Pract 2018 Oct 18:JOP1800261. [doi: 10.1200/JOP.18.00261] [Medline: 30335558]
149. Rains SA, Meng J. Social enhancement and compensation in online social support among cancer patients: the role of social network properties. Health Commun 2022 Apr;37(4):490-497. [doi: 10.1080/10410236.2020.1853327] [Medline: 33272063]
150. Crook B, Love B. Examining the light and dark of an online young adult cancer support community. Qual Health Res 2017 May;27(6):938-948. [doi: 10.1177/1049732316672638] [Medline: 27742764]
151. Frost J, Vermeulen IE, Beekers N. Anonymity versus privacy: selective information sharing in online cancer communities. J Med Internet Res 2014;16(5):e126. [doi: 10.2196/jmir.2684]
152. Holdren J, Surkan K, Downing A. Perspectives of people with cancer or hereditary cancer risk on the use and value of online peer support. J Patient Cent Res Rev 2023;10(2):58-67. [doi: 10.17294/2330-0698.1968]
153. Baxter J. ‘Keep strong, remember everything you have learnt’: constructing support and solidarity through online interaction within a UK cancer support group. Discourse Soc 2018 Jul;29(4):363-379. [doi: 10.1177/0957926518754414]
154. Gooden RJ, Winefield HR. Breast and prostate cancer online discussion boards: a thematic analysis of gender differences and similarities. J Health Psychol 2007 Jan;12(1):103-114. [doi: 10.1177/1359105307071744] [Medline: 17158844]
155. Erfani SS, Blount Y, Abedin B. The influence of health-specific social network site use on the psychological well-being of cancer-affected people. J Am Med Inform Assoc 2016 May;23(3):467-476. [doi: 10.1093/jamia/ocv170] [Medline: 26911816]
156. van Uden-Kraan CF, Drossaert CHC, Taal E, et al. Coping with somatic illnesses in online support groups: do the feared disadvantages actually occur? Comput Hum Behav 2008 Mar;24(2):309-324. [doi: 10.1016/j.chb.2007.01.014]
157. Radin P. “To me, it’s my life”: medical communication, trust, and activism in cyberspace. Soc Sci Med 2006 Feb;62(3):591-601. [doi: 10.1016/j.socscimed.2005.06.022]
158. Ginossar T. Online participation: a content analysis of differences in utilization of two online cancer communities by men and women, patients and family members. Health Commun 2008;23(1):1-12. [doi: 10.1080/10410230701697100] [Medline: 18443988]
159. Wandrey RL, Qualls WD, Mosack KE. Are mainstream support services meeting the needs of sexual minority women with breast cancer? An exploration of the perspectives and experiences of users of an online support forum. J Gay Lesbian Soc Serv 2016 Oct;28(4):336-348. [doi: 10.1080/10538720.2016.1221783]
160. Ray CD, Floyd K, Tietsort CJ, et al. Mixed messages: I. The consequences of communicating negative statements within emotional support messages to cancer patients. J Patient Exp 2020 Aug;7(4):593-599. [doi: 10.1177/2374373519873781] [Medline: 33062883]
161. Ray CD, Harvey J, Floyd K, Bonito JA, Reblin M. Mixed messages: II. Outcomes associated with the proportion and placement of negative statements in support messages. Health Commun 2021 Jun 7;36(7):856-865. [doi: 10.1080/10410236.2020.1719322]
162. Seymour-Smith S. A reconsideration of the gendered mechanisms of support in online interactions about testicular implants: a discursive approach. Health Psychol 2013 Jan;32(1):91-99. [doi: 10.1037/a0029507] [Medline: 23316857]
163. Stephen JE, Christie G, Flood K, et al. Facilitating online support groups for cancer patients: the learning experience of psycho-oncology clinicians. Psychooncology 2011 Aug;20(8):832-840. [doi: 10.1002/pon.1791] [Medline: 20677179]
164. Zhu Y, Glowacki EM, Yang Y. A social ties-based approach to breast cancer patients’ quality of life: examining group ties and individual ties across offline and online settings. Health Commun 2021 May 12;36(6):741-751. [doi: 10.1080/10410236.2020.1712520]
165. Klemm P. Effects of online support group format (moderated vs peer-led) on depressive symptoms and extent of participation in women with breast cancer. CIN 2012;30(1):9-18. [doi: 10.1097/NCN.0b013e3182343efa]

XSL•FO
RenderX

166. Owen JE, Klapow JC, Roth DL, Tucker DC. Use of the internet for information and support: disclosure among persons with breast and prostate cancer. J Behav Med 2004 Oct;27(5):491-505. [doi: 10.1023/b:jobm.0000047612.81370.f7] [Medline: 15675637]
167. Lieberman MA, Russo S. Self help groups and the internet: breast cancer newsgroups. Int J Self Help Self Care 2001;1(4):323-344. [doi: 10.2190/EGWJ-GBQD-KPDA-7UUD]
168. Namkoong K, McLaughlin B, Yoo W, et al. The effects of expression: how providing emotional support online improves cancer patients' coping strategies. JNCI Monographs 2013 Dec;2013(47):169-174. [doi: 10.1093/jncimonographs/lgt033]
169. Nambisan P. Evaluating patient experience in online health communities: implications for health care organizations. Health Care Manage Rev 2011;36(2):124-133. [doi: 10.1097/HMR.0b013e3182099f82] [Medline: 21317657]
170. Smith CE, Levonian Z, Ma H, et al. I cannot do all of this alone": exploring instrumental and prayer support in online health communities. ACM Trans Comput-Hum Interact 2020;27:38. [doi: 10.1145/3402855]
171. Durant KT, McCray AT, Safran C. Identifying gender-preferred communication styles within online cancer communities: a retrospective, longitudinal analysis. PLOS ONE 2012;7(11):e49169. [doi: 10.1371/journal.pone.0049169]
172. Foster C, Roffe L. An exploration of the internet as a self-management resource. J Res Nurs 2009 Jan;14(1):13-24. [doi: 10.1177/1744987108099237]
173. Painter C. Reaching for the brass ring. Cell 2020 Apr;181(1):8-14. [doi: 10.1016/j.cell.2020.02.030]
174. Preece J. Sociability and usability in online communities: determining and measuring success. Behav Inf Technol 2001 Jan;20(5):347-356. [doi: 10.1080/01449290110084683]
175. Falisi AL, Wiseman KP, Gaysynsky A, Scheideler JK, Ramin DA, Chou WYS. Social media for breast cancer survivors: a literature review. J Cancer Surviv 2017 Dec;11(6):808-821. [doi: 10.1007/s11764-017-0620-5] [Medline: 28601981]
176. Qiu B, Zhao K, Mitra P, et al. Get online support, feel better -- sentiment analysis and dynamics in an online cancer survivor community. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing: IEEE; 2011:274-281. [doi: 10.1109/PASSAT/SocialCom.2011.127]
177. Zhou J, Wang G, Zhou T, Fan T. The role of off-topic discussions in online health support groups: insights from a content analysis of an online rectal cancer group. Support Care Cancer 2020 Jul;28(7):3219-3226. [doi: 10.1007/s00520-019-05159-4]
178. Han JY, Kim E, Lee YI, Shah DV, Gustafson DH. A longitudinal investigation of empathic exchanges in online cancer support groups: message reception and expression effects on patients' psychosocial health outcomes. J Health Commun 2019;24(6):615-623. [doi: 10.1080/10810730.2019.1644401] [Medline: 31340721]
179. Han JY, Shaw BR, Hawkins RP, Pingree S, Mctavish F, Gustafson DH. Expressing positive emotions within online support groups by women with breast cancer. J Health Psychol 2008 Nov;13(8):1002-1007. [doi: 10.1177/1359105308097963]
180. Yang D, Yao Z, Seering J, et al. The channel matters: self-disclosure, reciprocity and social support in online cancer support groups. Assoc Comput Mach 2019:1-15. [doi: 10.1145/3290605.3300261]
181. Yoo W, Namkoong K, Choi M, et al. Giving and receiving emotional support online: communication competence as a moderator of psychosocial benefits for women with breast cancer. Comput Hum Behav 2014 Jan;30:13-22. [doi: 10.1016/j.chb.2013.07.024] [Medline: 24058261]
182. Crook B, Glowacki EM, Love B, Jones BL, Macpherson CF, Johnson RH. Hanging by a thread: exploring the features of nonresponse in an online young adult cancer survivorship support community. J Cancer Surviv 2016 Feb;10(1):185-193. [doi: 10.1007/s11764-015-0465-8] [Medline: 26123806]
183. Lewallen AC, Owen JE, Bantum EO, Stanton AL. How language affects peer responsiveness in an online cancer support group: implications for treatment design and facilitation. Psychooncology 2014 Jul;23(7):766-772. [doi: 10.1002/pon.3477] [Medline: 24519856]
184. Lieberman M. The role of insightful disclosure in outcomes for women in peer-directed breast cancer groups: a replication study. Psychooncology 2007 Oct;16(10):961-964. [doi: 10.1002/pon.1181] [Medline: 17328007]
185. Lieberman MA, Goldstein BA. Not all negative emotions are equal: the role of emotional expression in online support groups for women with breast cancer. Psychooncology 2006 Feb;15(2):160-168. [doi: 10.1002/pon.932] [Medline: 15880627]
186. Algtewi E, Owens J, Baker SR. Online support groups for head and neck cancer and health-related quality of life. Qual Life Res 2017 Sep;26(9):2351-2362. [doi: 10.1007/s11136-017-1575-8]
187. Hong Y, Peña-Purcell NC, Ory MG. Outcomes of online support and resources for cancer survivors: a systematic literature review. Patient Educ Couns 2012 Mar;86(3):288-296. [doi: 10.1016/j.pec.2011.06.014]
188. Kim E, Han JY, Moon TJ, et al. The process and effect of supportive message expression and reception in online breast cancer support groups. Psychooncology 2012 May;21(5):531-540. [doi: 10.1002/pon.1942] [Medline: 21416553]
189. Shaw BR, Hawkins R, McTavish F, Pingree S, Gustafson DH. Effects of insightful disclosure within computer mediated support groups on women with breast cancer. Health Commun 2006 Mar;19(2):133-142. [doi: 10.1207/s15327027hc1902_5]
190. Shaw B, Han JY, Kim E, et al. Effects of prayer and religious expression within computer support groups on women with breast cancer. Psychooncology 2007 Jul;16(7):676-687. [doi: 10.1002/pon.1129] [Medline: 17131348]
191. Shim M, Cappella JN, Han JY. How does insightful and emotional disclosure bring potential health benefits? Study based on online support groups for women with breast cancer. J Commun 2011 Jun;61(3):432-454. [doi: 10.1111/j.1460-2466.2011.01555.x]

XSL•FO
RenderX

192. Sillence E, Bussey L. Changing hospitals, choosing chemotherapy and deciding you've made the right choice: understanding the role of online support groups in different health decision-making activities. Patient Educ Couns 2017 May;100(5):994-999. [doi: 10.1016/j.pec.2016.12.004]
193. Verberne S, Batenburg A, Sanders R, van Eenbergen M, Das E, Lambooij MS. Analyzing empowerment processes among cancer patients in an online community: a text mining approach. JMIR Cancer 2019;5(1):e9887. [doi: 10.2196/cancer.9887]
194. Wise M, Marchand LR, Roberts LJ, Chih MY. Suffering in advanced cancer: a randomized control trial of a narrative intervention. J Palliat Med 2018 Feb;21(2):200-207. [doi: 10.1089/jpm.2017.0007]
195. Seale C. Gender accommodation in online cancer support groups. Health (London) 2006 Jul;10(3):345-360. [doi: 10.1177/1363459306064495] [Medline: 16775019]
196. Griffiths KM, Calear AL, Banfield M. Systematic review on Internet Support Groups (ISGs) and depression (1): do ISGs reduce depressive symptoms? J Med Internet Res 2009;11(3):e40. [doi: 10.2196/jmir.1270]
197. Kashian N, Jacobson S. Factors of engagement and patient-reported outcomes in a stage IV breast cancer Facebook group. Health Commun 2020 Jan 2;35(1):75-82. [doi: 10.1080/10410236.2018.1536962]
198. Morris BA, Lepore SJ, Wilson B, Lieberman MA, Dunn J, Chambers SK. Adopting a survivor identity after cancer in a peer support context. J Cancer Surviv 2014 Sep;8(3):427-436. [doi: 10.1007/s11764-014-0355-5]
199. Seçk[idot]n G. Internet technology in service of personal health care management: patient perspective. J Technol Hum Serv 2009 Apr 23;27(2):79-92. [doi: 10.1080/15228830902749179] [Medline: 20198123]
200. Valle CG, Tate DF. Engagement of young adult cancer survivors within a Facebook-based physical activity intervention. Transl Behav Med 2017 Dec;7(4):667-679. [doi: 10.1007/s13142-017-0483-3] [Medline: 28374211]
201. Dolce MC. The internet as a source of health information: experiences of cancer survivors and caregivers with healthcare providers. Oncol Nurs Forum 2011 May;38(3):353-359. [doi: 10.1188/11.ONF.353-359] [Medline: 21531685]
202. Erfani SS, Abedin B, Blount Y. The effect of social network site use on the psychological well - being of cancer patients. Assoc Inf Sci Technol 2017 May;68(5):1308-1322. [doi: 10.1002/asi.23702]
203. Kim SC, Hawkins RP, Shah DV, Gustafson DH, Baker TB. Understanding how e-health interventions meet psychosocial needs of breast cancer patients: the pathways of influence on quality of life and cancer concerns. Psychooncology 2020 Oct;29(10):1704-1712. [doi: 10.1002/pon.5512] [Medline: 32779223]
204. Lippka Y, Patschan O, Todenhöfer T, et al. Bladder cancer discussed on the internet: a systematic analysis of gender differences of initial posters on an online discussion board. Springerplus 2013 Dec;2(1):445. [doi: 10.1186/2193-1801-2-445]
205. Meier A, Lyons EJ, Frydman G, Forlenza M, Rimer BK. How cancer survivors provide support on cancer-related internet mailing lists. J Med Internet Res 2007 May 14;9(2):e12. [doi: 10.2196/jmir.9.2.e12] [Medline: 17513283]
206. Osei DK, Lee JW, Modest NN, Pothier PKT. Effects of an online support group for prostate cancer survivors: a randomized trial. UNJ 2013;33(3):123. [doi: 10.7257/1053-816X.2013.33.3.123]
207. Gits AA, Ritter PL, Plant K, Lorig K. Coding and evaluating facilitator posts for an online cancer survivor workshop. J Psychosoc Oncol 2013 Mar;31(2):219-234. [doi: 10.1080/07347332.2012.761321]
208. Nambisan P. Information seeking and social support in online health communities: impact on patients' perceived empathy. J Am Med Inform Assoc 2011 May 1;18(3):298-304. [doi: 10.1136/amiajnl-2010-000058]
209. Han JY, Shah DV, Kim E, et al. Empathic exchanges in online cancer support groups: distinguishing message expression and reception effects. Health Commun 2011 Feb 28;26(2):185-197. [doi: 10.1080/10410236.2010.544283]
210. Mclaughlin B, Yang J, Yoo W, et al. The effects of expressing religious support online for breast cancer patients. Health Commun 2016 Jun 2;31(6):762-771. [doi: 10.1080/10410236.2015.1007550]
211. Curtis E, Jones R, Tipene-Leach D, et al. Why cultural safety rather than cultural competency is required to achieve health equity: a literature review and recommended definition. Int J Equity Health 2019 Nov 14;18(1):174. [doi: 10.1186/s12939-019-1082-3] [Medline: 31727076]
212. Norman CD, Skinner HA. eHEALS: the eHealth Literacy Scale. J Med Internet Res 2006;8(4):e27. [doi: 10.2196/jmir.8.4.e27]
213. Venkatesh V, Bala H. Technology acceptance model 3 and a research agenda on interventions. Decis Sci 2008 May;39(2):273-315. [doi: 10.1111/j.1540-5915.2008.00192.x]
214. Bender JL, Hueniken K, Eng L, et al. Internet and social media use in cancer patients: association with distress and perceived benefits and limitations. Support Care Cancer 2021 Sep;29(9):5273-5281. [doi: 10.1007/s00520-021-06077-0]
215. Cutrona C, Russel D. Type of social support and specific stress: toward a theory of optimal matching. In: Sarason B, Sarason I, Pierce G, editors. Social Support: An Interactional View: Wiley & Sons; 1990:319-366 URL: https://psycnet.apa.org/record/1990-97699-013 [accessed 2026-04-16]
216. Reynolds JS, Perrin NA. Mismatches in social support and psychosocial adjustment to breast cancer. Health Psychol 2004;23(4):425-430. [doi: 10.1037/0278-6133.23.4.425]
217. Merluzzi TV, Philip EJ, Yang M, Heitzmann CA. Matching of received social support with need for support in adjusting to cancer and cancer survivorship. Psychooncology 2016 Jun;25(6):684-690. [doi: 10.1002/pon.3896]
218. Taylor SE, Lobel M. Social comparison activity under threat: downward evaluation and upward contacts. Psychol Rev 1989;96(4):569-575. [doi: 10.1037/0033-295X.96.4.569]

XSL•FO
RenderX

219. Bryde Christensen A, Wahrén S, Reinholt N, et al. “Despite the differences, we were all the same”. Group cohesion in diagnosis-specific and transdiagnostic CBT groups for anxiety and depression: a qualitative study. IJERPH 2021;18(10):5324. [doi: 10.3390/ijerph18105324]
220. Burlingame GM, McClendon DT, Yang C. Cohesion in group therapy: a meta-analysis. Psychotherapy (Chic) 2018 Dec;55(4):384-398. [doi: 10.1037/pst0000173] [Medline: 30335452]
221. Bogart LM, Helgeson VS. Social comparisons among women with breast cancer: a longitudinal investigation. J Appl Soc Pyschol 2000 Mar;30(3):547-575. [doi: 10.1111/j.1559-1816.2000.tb02496.x]
222. Low CA, Stanton AL, Bower JE, Gyllenhammer L. A randomized controlled trial of emotionally expressive writing for women with metastatic breast cancer. Health Psychol 2010;29(4):460-466. [doi: 10.1037/a0020153]
223. Stanton AL, Danoff-Burg S, Sworowski LA, et al. Randomized, controlled trial of written emotional expression and benefit finding in breast cancer patients. J Clin Oncol 2002 Oct 15;20(20):4160-4168. [doi: 10.1200/JCO.2002.08.521] [Medline: 12377959]
224. Helgeson VS, Gottlieb BH. Support groups. In: Social Support Measurement and Intervention: A Guide for Health and Social Scientists: Oxford University Press; 2000:221-245. [doi: 10.1093/med:psych/9780195126709.003.0007]
225. Kim M, Ellithorpe M, Burt SA. Anonymity and its role in digital aggression: a systematic review. Aggress Violent Behav 2023 Sep;72:101856. [doi: 10.1016/j.avb.2023.101856]
226. Esquivel A, Meric-Bernstam F, Bernstam EV. Accuracy and self correction of information received from an internet breast cancer list: content analysis. BMJ 2006 Apr 22;332(7547):939-942. [doi: 10.1136/bmj.38753.524201.7C] [Medline: 16513686]
227. van Uden-Kraan CF, Drossaert CHC, Taal E, Seydel ER, van de Laar M. Self-reported differences in empowerment between lurkers and posters in online patient support groups. J Med Internet Res 2008 Jun 30;10(2):e18. [doi: 10.2196/jmir.992] [Medline: 18653442]
228. Tamres LK, Janicki D, Helgeson VS. Sex differences in coping behavior: a meta-analytic review and an examination of relative coping. Pers Soc Psychol Rev 2002 Feb;6(1):2-30. [doi: 10.1207/S15327957PSPR0601_1]
229. Bender JL, Feldman-Stewart D, Tong C, et al. Health-related internet use among men with prostate cancer in Canada: cancer registry survey study. J Med Internet Res 2019 Nov 19;21(11):e14241. [doi: 10.2196/14241] [Medline: 31742561]
230. Seale C, Ziebland S, Charteris-Black J. Gender, cancer experience and internet use: a comparative keyword analysis of interviews and online cancer support groups. Soc Sci Med 2006 May;62(10):2577-2590. [doi: 10.1016/j.socscimed.2005.11.016] [Medline: 16361029]
231. McKeown L, Noce A, Czerny P. Factors associated with internet use: does rurality matter? Rural Small Town Can Anal Bulletin 2007 Sep;7(3):1-15 [FREE Full text]
232. Bender J, Jiang H. Online communities: prevalence and predictors of use, and reasons for non-use in cancer survivors. Medicine 2.0 2013 [FREE Full text]
233. Bender JL, Wiljer D, Sawka AM, Tsang R, Alkazaz N, Brierley JD. Thyroid cancer survivors’ perceptions of survivorship care follow-up options: a cross-sectional, mixed-methods survey. Support Care Cancer 2016 May;24(5):2007-2015. [doi: 10.1007/s00520-015-2981-5] [Medline: 26530226]
234. Bender JL, Wiljer D, To MJ, et al. Testicular cancer survivors’ supportive care needs and use of online support: a cross-sectional survey. Support Care Cancer 2012 Nov;20(11):2737-2746. [doi: 10.1007/s00520-012-1395-x] [Medline: 22382589]
235. Lee YJ, Kamen C, Margolies L, Boehmer U. Online health community experiences of sexual minority women with cancer. J Am Med Inform Assoc 2019 Aug 1;26(8-9):759-766. [doi: 10.1093/jamia/ocz103] [Medline: 31361002]
236. Leung YW, Park B, Heo R, et al. Providing care beyond therapy sessions with a natural language processing-based recommender system that identifies cancer patients who experience psychosocial challenges and provides self-care support: pilot study. JMIR Cancer 2022 Jul 29;8(3):e35893. [doi: 10.2196/35893] [Medline: 35904877]
237. Leung YW, Ng S, Duan L, et al. Therapist feedback and implications on adoption of an artificial intelligence–based co-facilitator for online cancer support groups: mixed methods single-arm usability study. JMIR Cancer 2023;9:e40113. [doi: 10.2196/40113]
238. Leung YW, Wouterloot E, Adikari A, et al. Artificial Intelligence-Based Co-Facilitator (AICF) for detecting and monitoring group cohesion outcomes in web-based cancer support groups: single-arm trial study. JMIR Cancer 2024 Jul 22;10:e43070. [doi: 10.2196/43070] [Medline: 39037754]
239. Gillespie T. Content moderation, AI, and the question of scale. Big Data Soc 2020 Jul;7(2):2053951720943234. [doi: 10.1177/2053951720943234]
240. Eysenbach G, Group CE. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011;13(4):e126. [doi: 10.2196/jmir.1923]

Abbreviations

2SLGBTQIA+: 2-spirit, lesbian, gay, bisexual, transgender, queer, intersex, or asexual
AI: artificial intelligence

XSL•FO
RenderX

CMOC: context-mechanism-outcome configuration
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
OSG: online support group
PRESS: Peer Review of Electronic Search Strategy
PROSPERO: International Prospective Register of Systematic Review
RAMESES: Realist and Meta-Narrative Evidence Syntheses: Evolving Standards
TAM: Technology Acceptance Model
VR: virtual reality

Edited by S Brini; submitted 14.May.2025; peer-reviewed by A Lazard, L Wreyford; revised version received 10.Feb.2026; accepted 04.Mar.2026; published 13.May.2026.

Please cite as:
Bender JL, Scruton S, Wong G, Squires LR, Babinski S, Englesakis M, Cyr A, Potts H, Tricco A, Wassersug R, Esplen MJ, Forster VJ, Perski O, Witteman HO, Soobiah C, Papadakos J, Young C, Fox C
How Cancer Online Support Groups Work, for Whom, and in What Circumstances: Realist Review
J Med Internet Res 2026;28:e77445
URL: https://www.jmir.org/2026/1/e77445
doi:10.2196/77445

© Jacqueline L Bender, Sarah Scruton, Geoff Wong, Lauren R Squires, Stephanie Babinski, Marina Englesakis, Alaina Cyr, Henry Potts, Andrea Tricco, Richard Wassersug, Mary Jane Esplen, Victoria J Forster, Olga Perski, Holly O Witteman, Charlene Soobiah, Janet Papadakos, Colleen Young, Colleen Fox. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Peer-Led, Nurse-Involved Blended Online and Offline Peer Support Program (PNO2PSP) for Psychosocial Adjustment in Young- to Middle-Aged Patients With Breast Cancer: Cluster Randomized Clinical Trial

Yiheng Zhang[1,2*], PhD; Haiyan He[3*], BS; Ye Zhang[4*], BS; Youxian Zhong[3], BS; Chunhong Luo[3], BS; Juanjuan Chen[3], BS; Lili Chen[3], MS; Ni Zhang[4], BS; Shihao Sun[3], MS; Baoyi Zhang[4], MS; Jia Fang[4], MS; Jingwen Yan[3*], MS; Meifen Zhang[4*], PhD

[1]School of Nursing, Jinan University, Guangzhou, China

[2]The First Affiliated Hospital, Jinan University, Guangzhou, Guangdong, China

[3]Department of Nursing, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China

[4]School of Nursing, Sun Yat-sen University, No.74 Zhongshan 2nd Road, Yuexiu District, Guangzhou, Guangdong Province, China

[*]these authors contributed equally

Corresponding Author:
Meifen Zhang, PhD
School of Nursing, Sun Yat-sen University, No.74 Zhongshan 2nd Road, Yuexiu District, Guangzhou, Guangdong Province, China

Abstract

Background: Young- to middle-aged patients with breast cancer face significant psychosocial challenges. Existing interventions often lack comprehensiveness, timely initiation, and tailoring specific to this population's unique needs.

Objective: This study aimed to evaluate the impact of a peer-led, nurse-involved, blended online and offline peer support intervention program (PNO2PSP) on psychosocial adjustment in young- to middle-aged patients with breast cancer.

Methods: The effectiveness of PNO2PSP was evaluated through a single-center cluster randomized controlled trial involving 70 newly diagnosed young- to middle-aged patients with breast cancer (35 in each group). The intervention group received an 8-week PNO2PSP in addition to routine care. Psychosocial adjustment, self-efficacy, social support, and coping modes were assessed presurgery and at 4, 8, and 12 weeks post surgery. Generalized estimating equations were used for intention-to-treat analyses. In-depth interviews with 9 participants explored their experiences.

Results: Compared to the control group, the intervention group demonstrated significantly lower psychosocial adjustment scores at 8 weeks (T2; Wald χ^2_1=14.96; P<.001) and 12 weeks (T3; Wald χ^2_1=7.49; P=.006); social support was higher at 8 weeks (T2; Wald χ^2_1=7.65; P=.006). Confrontation coping scores were higher at T2 (Wald χ^2_1=5.46; P=.02), T3 (Wald χ^2_1=10.04; P=.002), while avoidance coping scores were lower at T1 (Wald χ^2_1=8.24; P=.004), T2 (Wald χ^2_1=7.45; P=.006), and T3 (Wald χ^2_1=5.70; P=.02). Qualitative findings supported these quantitative results, highlighting the program's role in facilitating psychosocial adjustment, providing vital support, boosting treatment confidence, and fostering positive coping.

Conclusions: The PNO2PSP effectively improved psychosocial adjustment, social support, and positive coping in young- to middle-aged patients with breast cancer. Its scientifically validated, feasible, and patient-centered design supports its recommendation for wider clinical implementation, with continued training for peer supporters and sustained delivery of peer support.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2300076471; https://www.chictr.org.cn/showprojEN.html?proj=198622

(J Med Internet Res 2026;28:e86097) doi:10.2196/86097

KEYWORDS

young to middle-aged patients with breast cancer; psychosocial adjustment; peer support; blended online-offline intervention; cluster randomized clinical trial

XSL•FO
RenderX

Introduction

Breast cancer remains one of the most common malignancies among women and a major threat to their health. In 2022, approximately 2.3 million new cases were reported globally, accounting for 23.8% of all female cancer diagnoses. Notably, young- to middle-aged women (aged 18 - 59 years) comprised 53.9% of these cases [1]. The diagnosis of breast cancer is a life-altering event that brings multifaceted psychosocial challenges, particularly for young- to middle-aged women [2]. These challenges include fears of mortality associated with the diagnosis, body image disturbances caused by mastectomy, treatment side effects such as alopecia and sexual dysfunction, as well as significant changes in family and social roles, such as reduced work capacity and limitations in household activities [3-5]. These issues underscore the significant psychosocial adjustment challenges faced by young- to middle-aged patients with breast cancer.

Psychosocial adjustment refers to the adaptation status of individuals experiencing emotional and social problems caused by illness and treatment [6]. Poor psychosocial adjustment not only hampers patients' return to normal life and recovery but may also compromise immune function, potentially worsening disease prognosis [7-9]. For young- to middle-aged women, the challenges of psychosocial adjustment are even more pronounced due to their substantial family and societal responsibilities, including career development and child-rearing [9-11]. Studies have shown that younger patients with breast cancer often experience more severe long-term anxiety, depression, loneliness [12], physical dysfunction [9,13], changes in sexual relationships [4,14], fertility concerns [15], and difficulties in returning to work and social participation [5], all of which exacerbate their psychosocial maladjustment. Compared with older patients, young- and middle-aged women face particularly prominent challenges related to sexual function and fertility [16]. However, these concerns have not been sufficiently emphasized in previous research.

While various interventions have been developed to address psychosocial adjustment issues in patients with breast cancer, such as cognitive-behavioral interventions [17], mindfulness-based interventions [18], psychoeducation [19], spousal support [20], and multidisciplinary psychosocial interventions [21,22], these approaches have demonstrated some effectiveness, but also have limitations. First, most interventions focus solely on either psychological or social support without comprehensively addressing psychosocial adjustment. Second, insufficient evidence linking intervention targets to outcomes often results in suboptimal effectiveness and inefficient resource usage. Third, many interventions are initiated only after treatment has begun, lacking the timeliness of early intervention. Finally, there is a notable lack of targeted interventions specifically designed for young- to middle-aged patients with breast cancer.

Previous research by our team identified social support, self-efficacy, and coping modes as key predictors of psychosocial adjustment in this population [6,23]. These factors can serve as crucial targets for developing effective interventions. Within the concept of peer support, it is evident that peer support can improve health outcomes by augmenting social networks, enhancing effective coping strategies, and increasing self-efficacy [24]. Thus, this study adopts peer support as the primary intervention approach to improve psychosocial adjustment among young- to middle-aged patients with breast cancer. Peer support refers to the approach whereby individuals with the same disease or condition meet to exchange information, share experiences, and encourage or help each other to overcome difficulties [25]. Although peer-support interventions for patients with breast cancer have been developed, most focus on outcomes such as negative emotions or quality of life [26-29]. However, these interventions often lack clear mechanisms, theoretical grounding, and differentiation in content, while insufficiently addressing the specific needs of young- to middle-aged women.

Previous research has summarized 6 modes of peer intervention: one-on-one face-to-face, one-on-one telephone, one-on-one online, group face-to-face, group telephone, and group online settings [25]. Peer support is primarily categorized into group-based and one-on-one formats. While one-on-one support pairs a patient with a specific peer, the process is susceptible to compatibility issues [30]; inappropriate pairing may lead to intervention failure and require higher competency from the peer. In contrast, group-based support typically consists of 3 or more members, fostering a collaborative environment for sharing knowledge and experience while providing emotional, social, or substantive assistance [31].

Peer training is a critical component in consolidating the effectiveness of peer support, as it alleviates the psychological pressure and burden on peers while maximizing the safety and quality of the intervention [32-34]. Consequently, all peer supporters in this study underwent unified screening and standardized training. Furthermore, to address concerns regarding the accessibility of peer support and the reliability of information within groups, health care professionals were involved to provide specialized information and standardize group procedures, thereby preventing interpersonal conflicts and public disputes [34,35]. Professional involvement is thus essential to ensure informational accuracy. Therefore, this study was designed as a peer-led, nurse-involved intervention.

Regarding the delivery method, face-to-face interventions have been shown to facilitate deeper communication; however, long-term offline implementation is often constrained by distance and costs, leading to patient attrition and sustainability challenges [36,37]. With the diversification of social interaction, online support groups have become increasingly prevalent, offering temporal and spatial flexibility that facilitates broader dissemination [38]. To balance intervention efficacy with participant engagement, a blended online and offline model was adopted: offline sessions were conducted during hospitalization to build rapport among members, followed by continuous online support after discharge. Given this context, this study aims to develop and evaluate a peer-led, nurse-involved blended online and offline peer support program (PNO2PSP) tailored to young- to middle-aged patients with breast cancer. The goal is to provide an innovative and effective pathway to enhance

XSL•FO
RenderX

psychosocial adjustment in young- to middle-aged patients with breast cancer.

Methods

Research Design Overview

This study used a single-center cluster randomized controlled trial (CRCT) to evaluate the effectiveness of PNO2PSP. The control group received routine care, while the intervention group received the comprehensive PNO2PSP intervention.

Participants

Convenience sampling was used to recruit newly diagnosed young- to middle-aged patients with breast cancer who were admitted for surgical treatment to the Department of Breast Surgery, Sun Yat-sen Memorial Hospital (Textbox 1).

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria:

- Pathologically diagnosed with breast cancer (stage I, II, or III).
- Aged 18 - 59 years.
- Aware of their diagnosis.
- Voluntary participation with informed consent.
- Able to independently complete questionnaires.

Exclusion criteria:

- Receiving neoadjuvant chemotherapy.
- Communication barriers due to sensory impairments.
- Other life-threatening conditions or psychiatric disorders.
- History of recurrence, metastasis, or other malignancies.

Sample Size

Based on the sample size calculation formula for experimental studies, using a 2-tailed test, the primary outcome measure of this study was the score on the psychosocial adjustment scale. Referencing previous research [39], the sample size per group was calculated to be at least 18 participants. To account for potential dropouts and other factors, a 15% increase was applied, resulting in 21 participants per group, for a total of 42 participants.

To further accommodate the cluster-randomized design at the enrollment week level, reference was made to previous longitudinal psychosocial intervention studies, which reported a design effect (DE) of 1.472 [40]. By applying this DE to the initial estimate (42×1.472=62.04), it was determined that a minimum of 62 participants was required to maintain adequate power. Ultimately, 70 participants (35/group) were recruited to ensure robust statistical inference and to provide a buffer for attrition.

Group Allocation and Masking

The PNO2PSP intervention required peer support groups with at least 3 participants. This requirement was primarily grounded in prior peer-support studies, which indicate that a minimum of 3 members is essential to create a collaborative atmosphere and foster a diverse exchange of knowledge and experiential sharing [31]. From a clinical feasibility perspective, this threshold also ensured that the support group remained functional and stable even if 1 member was temporarily unavailable or withdrew. Given these requirements, individual randomization was impractical for ensuring stable group formation among young- to middle-aged patients with breast cancer undergoing treatment. Consequently, a CRCT design was adopted, with weekly cohorts of surgical patients serving as natural clusters [41]. Participants scheduled for surgery in the following week were recruited every Sunday to form a weekly cohort. Random numbers were generated using R (R Core Team) in RStudio (Posit PBC). The allocation results for intervention and control groups were placed in sealed, opaque envelopes prepared in advance by an independent research assistant responsible for maintaining allocation concealment. After recruitment each week, the corresponding envelope was opened to determine group assignment. Since most patients were discharged by Friday of their surgery week, this cohort-based grouping minimized the risk of intervention contamination between the intervention and control groups. Due to the intervention's nature, blinding of participants and providers was not feasible, and self-reported outcomes precluded blinding of outcome assessors. Thus, a single-blind design was used, with only statistical analysts blinded to group assignments to reduce bias.

Interventions

The control group received routine care, which included standard discharge education and follow-up. The discharge education was primarily provided by nurses through verbal health education, covering topics such as postoperative wound care, protection of the affected limb, and rehabilitation exercises. Follow-up involved sending reminders to participants about regular dressing changes, encouraging adherence to chemotherapy schedules, and prompting them to attend follow-up visits.

The "Peer Support Intervention Manual" was distributed to the participants in the intervention group. Research assistants

XSL•FO
RenderX

categorized participants and peer supporters into groups based on general demographic information, such as age, educational background, marital status, and reproductive history. Each group consisted of 2 peer supporters, 1 intervention nurse, and 3 - 5 group members. The groups were then assigned to implement the intervention. The intervention was implemented according to the finalized version of the peer support intervention program developed in the earlier stages. The intervention details are shown in Multimedia Appendix 1.

Evaluation

This study uses both quantitative and qualitative methods for outcome assessment. The quantitative evaluation primarily focuses on changes in the primary and secondary outcome measures through survey scales, while the qualitative evaluation involves interviews to understand participants' experiences.

Quantitative Evaluation

Primary Outcome Measure

The primary outcome of this study was defined as the mean difference in the psychosocial adjustment total score at the 2-month follow-up (T2). This specific time point was prespecified because previous longitudinal studies have indicated that improvements in psychosocial outcomes are more robustly observable at this interval compared to baseline [23]. The Psychosocial Adjustment to Illness Scale (PAIS) was used to assess the level of psychosocial adjustment. This scale includes 46 items across 7 dimensions: health care orientation, vocational environment, domestic environment, sexual relationships, extended family relationships, social environment, and psychological distress. The total score ranges from 0 to 138, with higher scores indicating poorer psychosocial adjustment and a greater level of psychosocial problems. The Cronbach α coefficient of the scale is 0.86 [23].

Secondary Outcome Measures

The secondary outcomes included the individual dimension scores of psychosocial adjustment, as well as the scores for self-efficacy, social support, and various dimensions of coping modes across all follow-up time points. The corresponding measurement instruments for these variables are detailed below.

General Self-Efficacy Scale

The General Self-Efficacy Scale (GSES) measures the self-efficacy of young- to middle-aged patients with breast cancer. It consists of 10 items, rated using a Likert 4-point scale. The total score ranges from 10 to 40, with higher scores reflecting greater self-efficacy. The Cronbach α coefficient is 0.91 [42].

Social Support Rate Scale

The Social Support Rate Scale (SSRS) assesses social support levels, consisting of 10 items. Each item is rated on a 4-point Likert scale, with higher scores indicating higher social support. The Cronbach α coefficient for SSRS is 0.92 [43].

Medical Coping Modes Questionnaire

The Medical Coping Modes Questionnaire (MCMQ) evaluates coping strategies, with 20 items divided into 3 dimensions: confrontation, avoidance, and acceptance-resignation. Higher scores on each dimension indicate a stronger tendency to adopt the corresponding coping mode [44]. The Cronbach α coefficient for the Chinese version ranges from 0.68 to 0.89.

Additional Data

In addition to the standardized scales, additional data were collected to provide information on the demographic and clinical characteristics of the participants and to assess treatment burden, as follows: (1) demographic information: sociodemographic data, including age, education level, marital status, residence, employment, and family income; (2) disease information survey: cancer type, disease stage, surgery type, treatment methods, lymph node involvement, and axillary lymph node dissection; and (3) treatment burden survey: patients' treatment burden was assessed during the final follow-up (week 12) through a self-reported survey. This included treatment costs, unplanned hospital visits, outpatient visits, and total hospitalization time.

Qualitative Evaluation

A purposeful sampling strategy was used to recruit participants from the intervention group. To ensure rich and diverse perspectives, participants of different ages, disease stages, and levels of psychosocial adjustment based on quantitative scores were invited. Semistructured, face-to-face, one-on-one interviews were conducted using an interview guide. Core questions included the following: (1) What is your impression of this project? (2) How has this project impacted your psychosocial adjustment? (3) What are the strengths and weaknesses of this project? The entire interview process was audio-recorded.

Data Collection

Participants who were admitted to the hospital wards were selected according to the inclusion and exclusion criteria of this study. Once eligible patients were identified, the purpose and significance of the study were explained to them, informed consent was obtained, and a baseline data questionnaire was completed. The completed forms were checked for missing items on site, and once verified, they were numbered to ensure data quality. Eligible participants were randomly assigned to the intervention group or the control group, and interventions were conducted accordingly. Based on previous studies, this study established follow-up data collection at 3 time points: 4 weeks post surgery (T1), 8 weeks post surgery (T2), and 12 weeks post surgery (T3) [23]. At baseline, demographic information, disease information, psychosocial adjustment, self-efficacy, social support, and coping modes were collected. At T1, T2, and T3, psychosocial adjustment, self-efficacy, social support, and coping modes were reassessed. At T3, additional information on treatment costs, number of doctor visits, hospitalizations, and unplanned readmissions was collected.

For the qualitative part, participants were contacted 1 - 2 days before the interview to explain the study's purpose and procedures. Interviews were conducted after obtaining written informed consent. All interviews took place in a quiet, private room in the ward and were conducted individually by trained researchers with extensive qualitative interviewing experience. At the beginning of each interview, participants were informed that there were no right or wrong answers and were encouraged

XSL•FO
RenderX

to share their experiences openly. They were also told that they could pause or stop the interview at any time if they felt uncomfortable. All data were anonymized to ensure confidentiality. The sample size for the qualitative phase was determined by data saturation. Data analysis was conducted alongside data collection, and recruitment ceased when no new themes emerged in 3 consecutive interviews [45].

Data Analysis

Descriptive statistics were used to summarize participants' baseline characteristics. Continuous variables were compared using independent samples *t* tests or Mann-Whitney *U* tests, while categorical variables were analyzed using Pearson chi-square (χ^2) tests or Fisher exact tests. Baseline balance was assessed between the intervention and control groups, as well as between participants who completed the study and those lost to follow-up. This study followed the intention-to-treat principle, including all participants who were randomized. No formal imputation was performed for missing data. Generalized estimating equations (GEEs) use all available observations under the missing-at-random assumption and provide valid population-averaged estimates without requiring explicit imputation.

The GEE model was used to evaluate the intervention effect. "Week" was defined as the cluster variable to adjust for cohort-level correlations, and "TIME" was set as the repeated factor. Data were sorted by participants' hospital identification numbers to ensure the integrity of the repeated measures structure in the model. TIME was treated as a categorical variable, and the group × time interaction effect was tested. The models specified a Gaussian identity link function with an exchangeable working correlation structure. To ensure the reliability of the inferences, robust (sandwich) SEs were used to provide unbiased parameter estimates. Clustering was explicitly modeled at the week level to account for potential correlation within cohorts. This approach ensures that the SEs are adjusted for the nested structure of the data.

The primary analysis consisted of unadjusted GEE models, which included group, time, and the group × time interaction as predictors. The control group and baseline (T0) served as the reference categories. To test the robustness of the primary findings against potential confounding—specifically the baseline imbalance observed in coping modes (avoidance)—sensitivity analyses were performed using adjusted GEE models. In these models, baseline coping modes (avoidance) were included as a covariate. These analyses confirmed that the statistical significance and direction of the intervention effects remained consistent across both models, reinforcing the reliability of the primary results. Further analysis revealed no significant interaction between group size and the intervention effect on the primary outcome. Statistical significance was defined as $P<.05$. All analyses were performed using R.

Qualitative data collection and analysis were conducted concurrently. All interviews were transcribed verbatim within 24 hours after completion. Data were analyzed using content analysis. Two researchers independently performed open coding, grouped similar codes into categories, and further abstracted them into core categories. Relationships among categories were examined during the abstraction process. Discrepancies were resolved through team discussion until consensus was reached, and triangulation was used to integrate qualitative and quantitative findings, allowing for a more comprehensive interpretation of how and why the intervention produced its observed effects [46].

Quality Control

A standardized protocol for the implementation of the intervention was developed, and all team members involved in the intervention received uniform training to minimize inconsistencies during the execution of the plan. Research assistants participated in both online and offline group discussions throughout the intervention process. All offline interventions were video-recorded, online interventions were screen-recorded, and all voice chats were audio-recorded to ensure comprehensive documentation. The research team held meetings at least once a week to identify issues and difficulties in the implementation of the research protocol. All peer supporters underwent uniform training and assessment. Additionally, intervention guidelines were provided to assist them in the implementation process.

To minimize contamination between groups, a cluster enrollment strategy based on admission week was adopted. Patients in the intervention and control groups were recruited in different weeks, and due to the fixed surgical admission and discharge schedule, participants from different weeks did not overlap during hospitalization. This temporal separation reduced opportunities for cross-group interaction. In addition, all participants signed a confidentiality agreement before the intervention, which emphasized that all information about intervention content and personal experiences of group members should not be disclosed to individuals outside the group. Intervention delivery was restricted to designated facilitators, and no crossover of intervention staff occurred between study arms.

Ethical Considerations

This study complied with the principles of the Declaration of Helsinki and relevant Chinese regulations governing clinical research. It was approved by the Ethics Committee of the School of Nursing, Sun Yat-sen University, Guangzhou, China (approval no L2023SYSU-HL-019) and registered with the China Clinical Trial Registry. All relevant departments in the hospital approved the study. Participants in the study signed a paper copy of the informed consent before enrollment. They were provided with detailed information about the study's objectives, process, potential risks and benefits, their right to withdraw at any time, and data collection methods. The data obtained in the study were handled in accordance with confidentiality principles. All data were anonymized before analysis. Personal and clinical information was securely stored, and access was restricted to authorized members of the research team only. Identifiable features or images of patients are not included in the paper and supplementary materials. Participants did not receive any financial compensation for their involvement in the study. However, they were provided with free educational support as part of the research process. To ensure that all

XSL•FO
RenderX

participants benefit equitably, after the data collection was completed, the participants in the control group received the intervention manual and were added to online peer support groups.

Results

Enrollment

During the study, 90 young- and middle-aged patients with breast cancer who met the inclusion criteria were approached. Among them, 20 refused to participate, 10 expressed no interest, 7 were unwilling to complete the questionnaires, and 3 did not have time to participate in the intervention. As a result, a total of 70 participants from 6 recruitment clusters (weeks) were included in the final analysis. For the primary outcome, the observed intraclass correlation was 0.0303, which is below the threshold of 0.05 and therefore indicates a small degree of clustering according to previous studies [47]. Given an average cluster size of 11.67, the DE was calculated to be 1.323. Post hoc power analysis using PASS (NCSS LLC) demonstrated that with this DE and sample size, the study achieved a statistical power of 0.875 to detect the intervention effect on the primary outcome, confirming the adequacy of the sample size for robust inference.

Follow-Up and Attrition

At T1, 33 control group participants and all 35 intervention group participants completed the follow-up, with 2 losses in the control group. At T2, 31 control and 32 intervention participants completed the follow-up, with 4 and 3 losses, respectively. At T3, 28 control and 27 intervention participants completed the follow-up, with 7 and 8 losses, respectively. Overall, the attrition rate for the control group was calculated as 12.4%. The main reasons for attrition were loss of contact or unwillingness to complete the follow-up questionnaire. The attrition rate for the intervention group was 10.5%. The main reasons for attrition were unwillingness to continue participation, lack of time to complete the questionnaire, or loss of contact. A flowchart of patient recruitment, enrollment, follow-up, and attrition is shown in Figure 1.

Figure 1. CONSORT flow diagram. Enrollment and follow-up situations of patients in the intervention group and the control group.

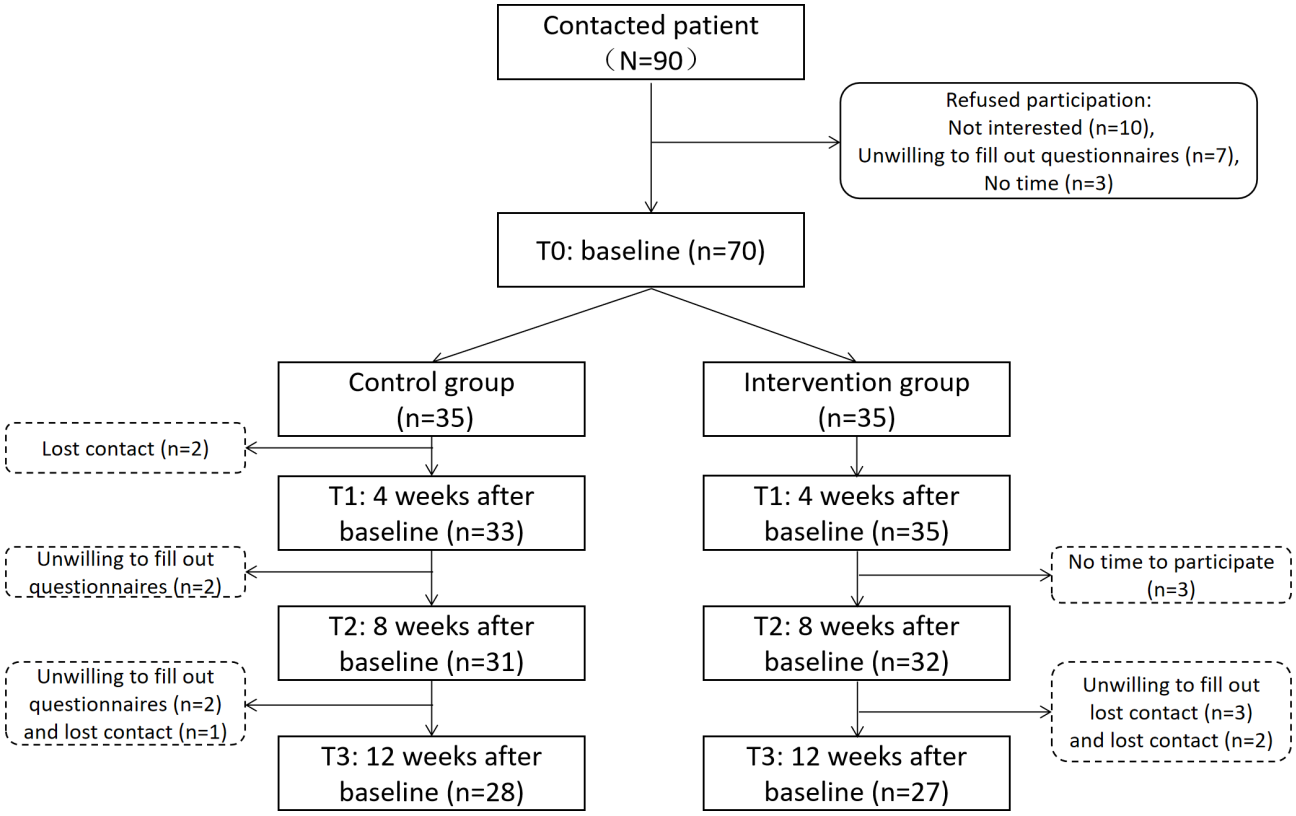

Baseline Data Balance Test

The baseline data comparison between the follow-up completion group and the lost-to-follow-up group showed no statistically significant differences in general information, disease, and treatment data (P>.05; shown in Table 1). However, there was a statistically significant difference in social support (P=.038). The comparison of baseline data between the intervention group and the control group revealed no statistically significant differences in general information, disease, and treatment data (P>.05). However, there was a statistically significant difference in the avoidance dimension of coping strategies (P=.003; shown in Table 2). Subsequently, the GEE analysis was conducted, with these scores used as a covariate to adjust the results (shown in Table 3).

XSL•FO
RenderX

Table . Balance test of baseline data between the follow-up completion group and the lost-to-follow-up group (N=70).

Project and group	Lost-to-follow-up group (n=16)	Follow-up completion group (n=54)	Statistics	*P* value
Age (years), mean (SD)	43.38 (7.974)	44.07 (7.427)	–0.33 (68)[a]	.76
Number of children, n (%)			0.44 (2)[b]	.80
0	2 (12.5)	4 (7.4)		
1	6 (37.5)	20 (37)		
≥2	8 (50)	30 (55.6)		
Marital status, n (%)			0.07 (1)[c]	.79
Married	16 (100)	51 (94.4)		
Unmarried	0 (0)	3 (5.6)		
Educational attainment, n (%)			0.75 (2)[b]	.69
Junior high school or below	4 (25)	9 (16.7)		
High school or technical secondary	5 (31.2)	22 (40.7)		
College or above	7 (43.8)	23 (42.6)		
Occupation type, n (%)			1.26 (4)[b]	.87
Enterprise employee	4 (25)	14 (25.9)		
Farmer or worker	3 (18.8)	5 (9.3)		
Government employee	4 (25)	14 (25.9)		
Freelancer	3 (18.8)	11 (20.4)		
Unemployed	2 (12.4)	10 (18.5)		
Place of residence, n (%)			0.16 (1)[b]	.69
Town	12 (75)	43 (79.6)		
Rural area	4 (25)	11 (20.4)		
Religious belief, n (%)			0.51 (1)[c]	.48
None	16 (100)	49 (90.7)		
Yes	0 (0)	5 (9.3)		
Monthly household income per capita (CNY), n (%)			3.62 (3)[b]	.31
<3000 (US $420)	5 (31.3)	11 20.3)		
3000-5000 (US $420-700)	1 (6.3)	15 (27.8)		
5001-10,000 (US $700-1400)	6 (37.4)	19 (35.2)		
＞10,000 (US $1400)	4 (25)	9 (16.7)		
Tumor grade, n (%)			0.51 (2)[b]	.77
I	2 (12.5)	11 (20.4)		
II	8 (50)	24 (44.4)		
III	6 (37.5)	19 (35.2)		
Lymph node metastasis, n (%)			0.81 (2)[b]	.37
No	6 (37.5)	14 (25.9)		
Yes	10 (62.5)	40 (74.1)		
Axillary lymph node dissection, n (%)			0.63 (1)[c]	.43

XSL•FO
RenderX

Project and group	Lost-to-follow-up group (n=16)	Follow-up completion group (n=54)	Statistics	*P* value
Yes	1 (6.3)	10 (18.5)		
No	15 (93.7)	44 (81.5)		
Triple-negative, n (%)			—[e]	1[d]
Yes	1 (6.3)	4 (7.4)		
No	15 (93.7)	50 (92.6)		
Lesion location, n (%)			—	.34[d]
Left only	11 (68.8)	29 (53.7)		
Right only	4 (25)	23 (42.6)		
Bilateral	1 (6.2)	2 (3.7)		
Type of surgery, n (%)			—	.09[d]
Breast-conserving surgery	9 (56.3)	36 (66.7)		
Modified radical surgery	6 (37.4)	7 (13)		
Total resection	1 (6.3)	11 (20.3)		
Received chemotherapy, n (%)			0.10 (1)[b]	.76
Yes	10 (62.5)	36 (66.7)		
No	6 (37.5)	18 (33.3)		
Self-efficacy, mean (SD)	25.88 (7.35)	26.76 (6.87)	–0.45 (68)[a]	.66
Social support, mean (SD)	40.94 (7.63)	45.19 (6.88)	–2.12 (68)[a]	.04
Coping strategies, n (%)				
Facing	20.13 (3.76)	20.22 (2.93)	–0.11 (68)[a]	.91
Avoidance	17.69 (2.70)	17.35 (2.74)	0.43 (68)[a]	.67
Yield	8.75 (3.30)	7.98 (2.33)	1.05 (68)[a]	.30
Psychosocial adjustment, mean (SD)	49.56 (20.32)	43.94 (17.06)	1.11 (68)[a]	.27

[a]*t* test.

[b]Pearson chi-square test.

[c]Continuous correction chi-square test: the expected value of the cell is <5.

[d]Fisher exact probability method: the expected value of the cell is <1.

[e]Not available.

XSL•FO
RenderX

Table . Balance test of baseline data between the intervention group and the control group (N=70).

Project and group	Intervention group (n=35)	Control group (n=35)	Statistics	*P* value
Age (years), mean (SD)	42.83 (6.719)	45.00 (8.164)	–1.21 (68)[a]	.23
Number of children, n (%)			0.26 (2)[c]	.88
0	3 (8.6)	3 (8.6)		
1	14 (40.0)	12 (34.3)		
≥2	18 (51.4)	20 (57.1)		
Marital status, n (%)			0 (1)[c]	≥.99
Married	33 (94.3)	34 (97.1)		
Unmarried	2 (5.7)	1 (2.9)		
Educational attainment, n (%)			0.54 (2)[b]	.76
Junior high school or below	7 (20)	6 (17.1)		
High school or technical secondary	12 (34.3)	15 (42.9)		
College or above	16 (45.7)	14 (40)		
Occupation type, n (%)			2.87 (4)[c]	.58
Enterprise employee	11 (31.4)	7 (20)		
Farmer or worker	5 (14.3)	3 (8.6)		
Government employee	9 (25.7)	9 (25.7)		
Freelancer	5 (14.3)	9 (25.7)		
Unemployed	5 (14.3)	7 (20)		
Place of residence, n (%)			0.09 (1)[b]	.77
Town	27 (77.1)	28 (80)		
Rural area	8 (22.9)	7 (20)		
Religious belief, n (%)			0.22 (1)[c]	.64
None	32 (91.4)	33 (94.3)		
Yes	3 (8.6)	2 (5.7)		
Monthly household income per capita (CNY)[e], n (%)			2.96 (3)[b]	.40
＜3000 (US $420)	8 (22.9)	8 (22.9)		
3000-5000 (US $420-US $700)	6 (17.1)	10 (28.6)		
5001-10,000 (US $700-US $1400)	12 (34.3)	13 (37.1)		
＞10,000 (US $1400)	9 (25.7)	4 (11.4)		
Tumor grade, n (%)			2.18 (2)[b]	.34
I	8 (22.9)	5 (14.3)		
II	13 (37.1)	19 (54.3)		
III	14 (40)	11 (31.4)		
Axillary lymph node dissection, n (%)			0.97 (1)[b]	.32
Yes	7 (20)	4 (11.4)		
No	28 (80)	31 (88.6)		
Triple-negative, n (%)			0 (1)[c]	>.99

XSL•FO
RenderX

Project and group	Intervention group (n=35)	Control group (n=35)	Statistics	*P* value
Yes	3 (8.6)	2 (5.7)		
No	32 (91.4)	33 (94.3)		
Lesion location, n (%)			3.06 (2)[d]	.25
Left only	18 (51.4)	22 (62.9)		
Right only	14 (40)	13 (37.1)		
Bilateral	3 (8.6)	0 (0)		
Type of surgery, n (%)			1.58 (2)[b]	.45
Breast-conserving surgery	25 (71.4)	20 (57.1)		
Modified radical surgery	5 (14.3)	8 (22.9)		
Total resection	5 (14.3)	7 (20)		
Received chemotherapy, n (%)			0.25 (1)[b]	.62
Yes	24 (68.6)	22 (62.9)		
No	11 (31.4)	13 (37.1)		
Self-efficacy, mean (SD)	25.94 (6.86)	27.17 (7.06)	–0.74 (68)[a]	.46
Social support, mean (SD)	43.77 (7.38)	44.66 (7.15)	–0.51 (68)[a]	.61
Coping strategies, mean (SD)				
Facing	20.20 (2.77)	20.20 (3.45)	0 (68)[a]	≥.99
Avoidance	16.49 (2.36)	18.37 (2.76)	–3.08 (68)[a]	.003
Yield	8.23 (2.28)	8.09 (2.87)	0.23 (68)[a]	.82
Psychosocial adjustment, mean (SD)				
Total score	46.60 (17.14)	43.86 (18.69)	0.64 (68)[a]	.53
Health care	9.03 (3.20)	8.86 (2.93)	0.23 (68)[a]	.82
Occupational environment	7.11 (3.11)	6.00 (3.19)	1.48 (68)[a]	.14
Family environment	6.03 (3.22)	5.09 (2.97)	1.27 (68)[a]	.21
Sexual relationship	6.00 (2.77)	6.66 (3.46)	–0.88 (68)[a]	.38
Expand family relationships	2.77 (2.37)	2.11 (1.92)	1.27 (68)[a]	.21
Social relationship	7.46 (4.58)	6.57 (5.26)	0.75 (68)[a]	.46
Psychological distress	8.20 (4.68)	8.57 (4.86)	–0.33 (68)[a]	.75

[a]*t* test.

[b]Pearson chi-square test.

[c]Continuous correction chi-square test: the expected value of the cell is <5.

[d]Fisher exact probability method: the expected value of the cell is <1.

XSL•FO
RenderX

Table . Results of generalized estimating equation models for psychosocial adjustment and scores of psychosocial adjustment and secondary outcomes in 2 groups of young- and middle-aged patients with breast cancer (N=70).

Variable and time	Control group (n=35)	Intervention group (n=35)	Crude model						Adjusted model					
			Group effect		Time effect		Interaction effect		Group effect		Time effect		Interaction effect	
			Wald χ^2 (*df*)	*P* value	Wald χ^2 (*df*)	*P* value	Wald χ^2 (*df*)	*P* value	Wald χ^2 (*df*)	*P* value	Wald χ^2 (*df*)	*P* value	Wald χ^2 (*df*)	*P* value
Psychosocial adjustment, mean (SD)			6.83 (1)	.009	41.03 (3)	<.001	19.03 (3)	<.001	3.62 (1)	.06	48.15 (3)	<.001	18.39 (3)	<.001
T0	46.60 (17.14)	43.86 (18.69)												
T1	52.91 (18.65)	48.03 (18.13)												
T2	55.32 (16.36)	43.31 (14.67)												
T3	52.64 (17.47)	42.44 (17.55)												
Health orientation, mean (SD)			8.41 (1)	.004	3.82 (3)	.28	477.87 (3)	<.001	7.09 (1)	.008	4.73 (3)	.19	248.08 (3)	<.001
T0	9.03 (3.20)	8.86 (2.93)												
T1	9.82 (2.89)	8.40 (2.66)												
T2	9.65 (2.46)	9.13 (2.85)												
T3	9.14 (2.77)	9.11 (2.97)												
Vocational environment, mean (SD)			7.47 (1)	.006	23.91 (3)	0	5.60 (3)	.13	4.49 (1)	.03	19.68 (3)	<.001	5.43 (3)	.14
T0	7.11 (3.11)	6.00 (3.19)												
T1	8.73 (3.57)	7.43 (3.90)												
T2	8.81 (2.88)	6.41 (3.40)												
T3	7.68 (3.43)	6.07 (3.74)												
Domestic environment, mean (SD)			5.40 (1)	.02	10.38 (3)	.02	24.30 (3)	<.001	2.01 (1)	.16	10.42 (3)	.02	29.67 (3)	<.001
T0	6.03 (3.22)	5.09 (2.97)												
T1	6.12 (3.51)	6.49 (3.53)												
T2	7.19 (3.98)	5.44 (2.77)												
T3	7.07 (3.85)	5.07 (3.41)												
Sexual relationship, mean (SD)			0.48 (1)	.49	67.19 (3)	<.001	12.36 (3)	.006	0.21 (1)	.65	64.38 (3)	<.001	12.140 (3)	.007
T0	6.00 (2.77)	6.66 (3.47)												

XSL•FO
RenderX

Variable and time	Control group (n=35)	Intervention group (n=35)	Crude model						Adjusted model					
			Group effect		Time effect		Interaction effect		Group effect		Time effect		Interaction effect	
			Wald χ^2 (*df*)	*P* value	Wald χ^2 (*df*)	*P* value	Wald χ^2 (*df*)	*P* value	Wald χ^2 (*df*)	*P* value	Wald χ^2 (*df*)	*P* value	Wald χ^2 (*df*)	*P* value
T1	7.64 (3.44)	8.00 (3.24)												
T2	8.65 (3.23)	6.97 (2.95)												
T3	8.11 (3.06)	6.74 (3.10)												
Expand family relationships, mean (SD)			10.63 (1)	.001	29.16 (3)	<.001	8.99 (3)	.03	8.06 (1)	.005	29.48 (3)	<.001	7.50 (3)	.06
T0	2.77 (2.38)	2.11 (1.92)												
T1	3.52 (2.35)	2.49 (1.96)												
T2	3.42 (2.69)	2.31 (1.87)												
T3	3.96 (2.66)	2.48 (1.99)												
Social environment, mean (SD)			5.24 (1)	.02	3.28 (3)	.35	11.99 (3)	.007	1.82 (1)	.18	3.61 (3)	.31	11.50 (3)	.009
T0	7.46 (4.58)	6.57 (5.26)												
T1	8.21 (3.88)	7.40 (4.43)												
T2	8.81 (3.19)	6.69 (4.19)												
T3	8.57 (3.86)	6.78 (4.49)												
Psychological distress, mean (SD)			1.43 (1)	.23	10.65 (3)	.01	41.20 (3)	<.001	0.87 (1)	.35	10.99 (3)	.01	40.41 (3)	<.001
T0	8.20 (4.68)	8.57 (4.86)												
T1	8.88 (5.15)	7.83 (4.88)												
T2	8.81 (4.21)	6.38 (4.08)												
T3	8.11 (4.45)	6.19 (4.45)												
Self-efficacy, mean (SD)			2.57 (1)	.11	10.93 (3)	.01	2.42 (3)	.49	0.05 (1)	.83	12.35 (3)	.006	2.86 (3)	.42
T0	25.94 (6.87)	27.17 (7.06)												
T1	27.00 (6.72)	27.23 (6.66)												
T2	24.84 (7.19)	26.47 (6.68)												
T3	25.14 (7.24)	26.41 (6.58)												

XSL•FO
RenderX

Variable and time	Control group (n=35)	Intervention group (n=35)	Crude model						Adjusted model					
			Group effect		Time effect		Interaction effect		Group effect		Time effect		Interaction effect	
			Wald χ^2 (*df*)	*P* value	Wald χ^2 (*df*)	*P* value	Wald χ^2 (*df*)	*P* value	Wald χ^2 (*df*)	*P* value	Wald χ^2 (*df*)	*P* value	Wald χ^2 (*df*)	*P* value
Social support, mean (SD)			2.84 (1)	.09	3.95 (3)	.27	94.25 (3)	<.001	5.11 (1)	.02	3.59 (3)	.31	101.90 (3)	<.001
T0	43.77 (7.38)	44.66 (7.15)												
T1	43.55 (8.05)	45.40 (7.73)												
T2	41.48 (7.82)	45.81 (6.68)												
T3	43.46 (7.39)	46.00 (7.12)												
Coping mode to confrontation, mean (SD)			3.44 (1)	.06	1.34 (3)	.72	29.77 (3)	<.001	2.47 (1)	.12	1.64 (3)	.65	28.29 (3)	<.001
T0	20.20 (2.77)	20.20 (3.45)												
T1	20.36 (2.86)	20.66 (3.13)												
T2	19.87 (3.30)	21.34 (3.42)												
T3	19.68 (3.07)	21.37 (3.59)												
Coping mode to avoidance, mean (SD)			2.05 (1)	.15	85.93 (3)	<.001	18.72 (3)	<.001	2.05 (1)	.15	85.93 (3)	0	18.72 (3)	<.001
T0	16.49 (2.36)	18.37 (2.76)												
T1	17.88 (2.03)	18.06 (2.59)												
T2	17.26 (3.04)	16.91 (2.74)												
T3	17.32 (2.18)	17.41 (2.90)												
Coping mode to acceptance to resignation, mean (SD)			2.94 (1)	.09	5.417 (3)	.14	17.55 (3)	.001	1.37 (1)	.24	6.68 (3)	.08	12.46 (3)	.006
T0	8.23 (2.28)	8.09 (2.87)												
T1	8.33 (2.87)	7.66 (2.61)												
T2	8.55 (2.53)	7.63 (2.41)												
T3	8.86 (3.10)	7.63 (3.13)												

Quantitative Outcomes

The results of the GEE model, intergroup comparison analysis, and within-group comparison analysis in this study are shown in Table 3.

Psychosocial Adjustment

The intervention group had significantly lower psychosocial adjustment scores than the control group at T2 (*P*<.001) and T3 (*P*=.006).

XSL•FO
RenderX

Secondary Outcome

Additionally, the intervention group had significantly lower scores in the vocational environment dimension at T3 (*P*=.020), in the sexual relationship dimension at T2 (*P*=.001) and T3 (*P*=.002), as well as in the psychological distress dimension at T2 (*P*<.001) and T3 (*P*<.001; Table 4). The intervention group reported significantly higher social support scores than the control group (*P*=.006) at T2. Confrontation coping scores were also significantly higher in the intervention group (*P*=.019) at T2, (*P*=.002) at T3. In contrast, avoidance coping scores were significantly lower in the intervention group (*P*=.004, .006, and .017) at T1, T2, and T3, respectively. No significant differences were observed between groups in self-efficacy or acceptance-resignation coping at any time point (Table 4, Figures 2-7, and Multimedia Appendix 2).

XSL•FO
RenderX

Table . Comparison of generalized estimating equations for psychosocial adjustment and secondary outcomes between groups at different time points.

Outcome	Crude model			Adjusted model		
	B (95% CI)	Wald χ^2 (*df*)	*P* value	B (95% CI)	Wald χ^2 (*df*)	*P* value
Psychosocial adjustment						
Group	–2.705 (–8.366 to 2.956)	0.88 (1)	.35	–1.558 (–8.173 to 5.056)	0.21 (1)	.64
Time1	6.315 (2.949 to 9.681)	13.52 (1)	<.001	6.277 (2.872 to 9.683)	13.052 (1)	<.001
Time2	8.829 (4.533 to 13.124)	16.23 (1)	<.001	8.834 (4.487 to 13.180)	15.87 (1)	<.001
Time3	6.09 (1.329 to 10.850)	6.29 (1)	.01	6.117 (1.280 to 10.953)	6.15 (1)	.01
Group×Time1	–2.143 (–5.936 to 1.649)	1.23 (1)	.27	–2.106 (–5.894 to 1.683)	1.19 (1)	.28
Group×Time2	–9.406 (–14.172 to –4.640)	14.97 (1)	<.001	–9.582 (–14.405 to –4.759)	15.16 (1)	<.001
Group×Time3	–7.571 (–12.993 to –2.150)	7.49 (1)	.006	–7.539 (–12.995 to –2.082)	7.33 (1)	.007
Health care						
Group	–0.151 (–0.952 to 0.651)	0.14 (1)	.71	–0.190 (–1.156 to 0.775)	0.15 (1)	.70
Time1	0.75 (–0.571 to 2.071)	1.24 (1)	.27	0.753 (–0.566 to 2.072)	1.25 (1)	.26
Time2	0.611 (–0.640 to 1.863)	0.92 (1)	.34	0.615 (–0.623 to 1.853)	0.95 (1)	.33
Time3	0.094 (–1.512 to 1.700)	0.01 (1)	.91	0.096 (–1.508 to 1.699)	0.01 (1)	.91
Group×Time1	–1.207 (–2.528 to 0.115)	3.20 (1)	.07	–1.210 (–2.530 to 0.110)	3.23 (1)	.07
Group×Time2	–0.369 (–1.975 to 1.236)	0.20 (1)	.65	–0.364 (–1.978 to 1.250)	0.20 (1)	.66
Group×Time3	0.152 (–1.696 to 2.001)	0.03 (1)	.87	0.146 (–1.689 to 1.982)	0.02 (1)	.88
Occupational environment						
Group	–1.109 (–2.099 to –0.118)	4.81 (1)	.03	–0.944 (–1.908 to 0.021)	3.68 (1)	.06
Time1	1.625 (0.351 to 2.899)	6.25 (1)	.01	1.621 (0.348 to 2.894)	6.23 (1)	.01
Time2	1.715 (1.400 to 2.030)	113.74 (1)	<.001	1.715 (1.409 to 2.020)	121.06 (1)	<.001
Time3	0.579 (0.274 to 0.884)	13.86 (1)	<.001	0.583 (0.266 to 0.900)	13.02 (1)	<.001
Group×Time1	–0.196 (–2.035 to 1.642)	0.04 (1)	.83	–0.192 (–2.036 to 1.652)	0.04 (1)	.84
Group×Time2	–1.302 (–2.952 to 0.347)	2.40 (1)	.12	–1.324 (–2.942 to 0.293)	2.57 (1)	.11
Group×Time3	–0.521 (–0.960 to - 0.081)	5.40 (1)	.02	–0.515 (–0.970 to –0.060)	4.92 (1)	.03
Domestic environment						
Group	–0.941 (–1.726 to –0.157)	5.54 (1)	.02	–0.596 (–1.427 to 0.236)	1.97 (1)	.16

XSL•FO
RenderX

Outcome	Crude model			Adjusted model		
	B (95% CI)	Wald χ^2 (*df*)	*P* value	B (95% CI)	Wald χ^2 (*df*)	*P* value
Time1	0.091 (–0.242 to 0.424)	0.29 (1)	.59	0.077 (–0.286 to 0.440)	0.17 (1)	.68
Time2	1.167 (–0.014 to 2.348)	3.75 (1)	.05	1.161 (–0.029 to 2.351)	3.66 (1)	.06
Time3	1.043 (–0.182 to 2.268)	2.79 (1)	.10	1.047 (–0.189 to 2.282)	2.76 (1)	.10
Group×Time1	1.309 (0.619 to 1.998)	13.84 (1)	<.001	1.323 (0.633 to 2.013)	14.11 (1)	<.001
Group×Time2	–0.82 (–2.444 to 0.804)	0.98 (1)	.32	–0.872 (–2.509 to 0.766)	1.09 (1)	.30
Group×Time3	–1.056 (–2.496 to 0.384)	2.07 (1)	.15	–1.038 (–2.496 to 0.420)	1.95 (1)	.16
Sexual relationship						
Group	0.672 (–0.842 to 2.187)	0.76 (1)	.38	0.764 (–0.977 to 2.505)	0.74 (1)	.39
Time1	1.635 (0.932 to 2.339)	20.75 (1)	<.001	1.632 (0.933 to 2.330)	20.96 (1)	<.001
Time2	2.661 (2.176 to 3.145)	116.01 (1)	<.001	2.66 (2.171 to 3.149)	113.627 (1)	<.001
Time3	2.113 (1.241 to 2.985)	22.57 (1)	<.001	2.115 (1.234 to 2.996)	22.129 (1)	<.001
Group×Time1	–0.292 (–1.109 to 0.524)	0.49 (1)	.48	–0.289 (–1.099 to 0.521)	0.489 (1)	.49
Group×Time2	–2.322 (–3.634 to –1.009)	12.03 (1)	.001	–2.332 (–3.660 to –1.004)	11.849 (1)	.001
Group×Time3	–2.048 (–3.318 to –0.779)	9.99 (1)	.002	–2.045 (–3.314 to –0.775)	9.969 (1)	.002
Expand family relationships						
Group	–0.659 (–1.298 to –0.020)	4.09 (1)	.04	–0.635 (–0.950 to –0.320)	15.63 (1)	<.001
Time1	0.744 (0.108 to 1.379)	5.26 (1)	.02	0.743 (0.098 to 1.388)	5.10 (1)	.02
Time2	0.664 (–1.141 to 2.470)	0.52 (1)	.47	0.665 (–1.146 to 2.476)	0.52 (1)	.47
Time3	1.2 (–0.277 to 2.677)	2.54 (1)	.11	1.201 (–0.276 to 2.677)	2.54 (1)	.11
Group×Time1	–0.372 (–1.061 to 0.316)	1.12 (1)	.29	–0.372 (–1.070 to 0.327)	1.09 (1)	.30
Group×Time2	–0.47 (–2.293 to 1.353)	0.26 (1)	.61	–0.474 (–2.274 to 1.325)	0.27 (1)	.61
Group×Time3	–0.835 (–2.393 to 0.722)	1.10 (1)	.29	–0.834 (–2.405 to 0.736)	1.08 (1)	.30
Social environment						
Group	–0.88 (–3.076 to 1.316)	0.62 (1)	.43	–0.458 (–2.776 to 1.860)	0.15 (1)	.70
Time1	0.758 (–0.654 to 2.170)	1.11 (1)	.29	0.746 (–0.688 to 2.180)	1.04 (1)	.31
Time2	1.354 (–0.177 to 2.886)	3.01 (1)	.08	1.349 (–0.180 to 2.877)	2.99 (1)	.08

XSL•FO
RenderX

Outcome	Crude model			Adjusted model		
	B (95% CI)	Wald χ^2 (*df*)	*P* value	B (95% CI)	Wald χ^2 (*df*)	*P* value
Time3	1.118 (0.803 to 1.433)	48.37 (1)	<.001	1.125 (0.841 to 1.410)	60.21 (1)	<.001
Group×Time1	0.071 (–1.889 to 2.030)	0.01 (1)	.94	0.083 (–1.890 to 2.055)	0.01 (1)	.93
Group×Time2	–1.245 (–3.153 to 0.664)	1.63 (1)	.20	–1.304 (–3.200 to 0.592)	1.82 (1)	.18
Group×Time3	–0.919 (–2.248 to 0.409)	1.84 (1)	.18	–0.905 (–2.216 to 0.407)	1.83 (1)	.18
Psychological distress						
Group	0.393 (–1.581 to 2.367)	0.15 (1)	.70	0.608 (–1.665 to 2.881)	0.28 (1)	.60
Time1	0.665 (–0.991 to 2.321)	0.62 (1)	.43	0.657 (–0.993 to 2.306)	0.61 (1)	.44
Time2	0.661 (0.217 to 1.106)	8.51 (1)	.004	0.660 (0.234 to 1.086)	9.22 (1)	.002
Time3	–0.076 (–0.566 to 0.414)	0.09 (1)	.76	–0.072 (–0.564 to 0.419)	0.08 (1)	.77
Group×Time1	–1.408 (–3.339 to 0.524)	2.04 (1)	.15	–1.400 (–3.313 to 0.514)	2.06 (1)	.15
Group×Time2	–2.904 (–4.089 to –1.718)	23.05 (1)	<.001	–2.934 (–4.185 to –1.684)	21.15 (1)	<.001
Group×Time3	–2.343 (–3.112 to –1.574)	35.68 (1)	<.001	–2.334 (–3.106 to –1.562)	35.10 (1)	<.001
Self-efficacy						
Group	–0.88 (–3.076 to 1.316)	0.62 (1)	.43	0.344 (–1.016 to 1.703)	0.25 (1)	.62
Time1	0.758 (–0.654 to 2.170)	1.11 (1)	.29	1.091 (–2.291 to 4.472)	0.40 (1)	.58
Time2	1.354 (–0.177 to 2.886)	3.01 (1)	.08	–1.008 (–3.498 to 1.481)	0.63 (1)	.43
Time3	1.118 (0.803 to 1.433)	48.37 (1)	<.001	–0.777 (–3.003 to 1.450)	0.47 (1)	.49
Group×Time1	0.071 (–1.889 to 2.030)	0.01 (1)	.94	–1.033 (–4.473 to 2.406)	0.35 (1)	.56
Group×Time2	–1.245 (–3.153 to 0.664)	1.63 (1)	.20	0.443 (–2.553 to 3.440)	0.08 (1)	.77
Group×Time3	–0.919 (–2.248 to 0.409)	1.84 (1)	.18	–0.049 (–2.430 to 2.333)	0.002 (1)	.97
Social support						
Group	0.897 (–1.196 to 2.991)	0.71 (1)	.40	0.682 (–0.859 to 2.223)	0.75 (1)	.39
Time1	–0.23 (–3.207 to 2.747)	0.02 (1)	.88	–0.223 (–3.197 to 2.751)	0.02 (1)	.88
Time2	–2.429 (–4.746 to –0.112)	4.22 (1)	.04	–2.43 (–4.732 to –0.129)	4.28 (1)	.04
Time3	–0.368 (–2.593 to 1.857)	0.11 (1)	.75	–0.373 (–2.607 to 1.861)	0.11 (1)	.74
Group×Time1	0.973 (–2.303 to 4.250)	0.34 (1)	.56	0.966 (–2.300 to 4.231)	0.34 (1)	.56

XSL•FO
RenderX

Outcome	Crude model			Adjusted model		
	B (95% CI)	Wald χ^2 (*df*)	*P* value	B (95% CI)	Wald χ^2 (*df*)	*P* value
Group×Time2	3.636 (1.059 to 6.212)	7.65 (1)	.006	3.669 (1.033 to 6.306)	7.44 (1)	.006
Group×Time3	1.74 (–0.584 to 4.065)	2.15 (1)	.14	1.732 (–0.559 to 4.023)	2.20 (1)	.14
Coping modes to confrontation dimension						
Group	0.001 (–1.353 to 1.355)	<0.001 (1)	.10	–0.233 (–1.512 to 1.046)	0.13 (1)	.72
Time1	0.159 (–1.568 to 1.886)	0.03 (1)	.86	0.17 (–1.567 to 1.907)	0.04 (1)	.85
Time2	–0.324 (–1.449 to 0.800)	0.32 (1)	.57	–0.32 (–1.456 to 0.816)	0.30 (1)	.58
Time3	–0.521 (–1.352 to 0.309)	1.51 (1)	.22	–0.523 (–1.365 to 0.318)	1.49 (1)	.22
Group×Time1	0.298 (–1.452 to 2.047)	0.11 (1)	.74	0.287 (–1.471 to 2.045)	0.10 (1)	.75
Group×Time2	1.475 (0.238 to 2.711)	5.46 (1)	.02	1.500 (0.244 to 2.755)	5.48 (1)	.02
Group×Time3	1.691 (0.645 to 2.737)	10.04 (1)	.002	1.675 (0.621 to 2.730)	9.69 (1)	.002
Coping modes to avoidance dimension						
Group	1.886 (0.824 to 2.948)	12.12 (1)	<.001	1.886 (0.824 to 2.948)	12.12 (1)	<.001
Time1	1.391 (0.858 to 1.925)	26.14 (1)	<.001	1.391 (0.858 to 1.925)	26.14 (1)	<.001
Time2	0.774 (0.261 to 1.287)	8.73 (1)	.003	0.774 (0.261 to 1.287)	8.73 (1)	.003
Time3	0.836 (0.553 to 1.118)	33.62 (1)	<.001	0.836 (0.553 to 1.118)	33.62 (1)	<.001
Group×Time1	–1.706 (–2.870 to –0.541)	8.24 (1)	.004	–1.706 (–2.870 to –0.541)	8.24 (1)	.004
Group×Time2	–2.233 (–3.837 to –0.629)	7.45 (1)	.006	–2.233 (–3.837 to –0.629)	7.45 (1)	.006
Group×Time3	–1.8 (–3.277 to –0.323)	5.70 (1)	.02	–1.8 (–3.277 to –0.323)	5.70 (1)	.017
Coping modes acceptance to resignation dimension						
Group	–0.147 (–0.494 to 0.201)	0.69 (1)	.41	0.084 (–0.389 to 0.557)	0.12 (1)	.73
Time1	0.109 (–0.444 to 0.663)	0.15 (1)	.70	0.103 (–0.438 to 0.643)	0.14 (1)	.71
Time2	0.347 (–0.484 to 1.177)	0.67 (1)	.41	0.350 (–0.503 to 1.202)	0.65 (1)	.42
Time3	0.642 (–0.828 to 2.112)	0.73 (1)	.39	0.649 (–0.837 to 2.135)	0.73 (1)	.39
Group×Time1	–0.538 (–1.159 to 0.083)	2.88 (1)	.09	–0.531 (–1.136 to 0.073)	2.97 (1)	.09
Group×Time2	–0.816 (–1.673 to 0.041)	3.48 (1)	.06	–0.855 (–1.743 to 0.033)	3.56 (1)	.06
Group×Time3	–1.103 (–2.688 to 0.482)	1.86 (1)	.17	–1.095 (–2.697 to 0.506)	1.80 (1)	.18

XSL•FO
RenderX

Figure 2. Results of the intervention effect evaluation: total scores.

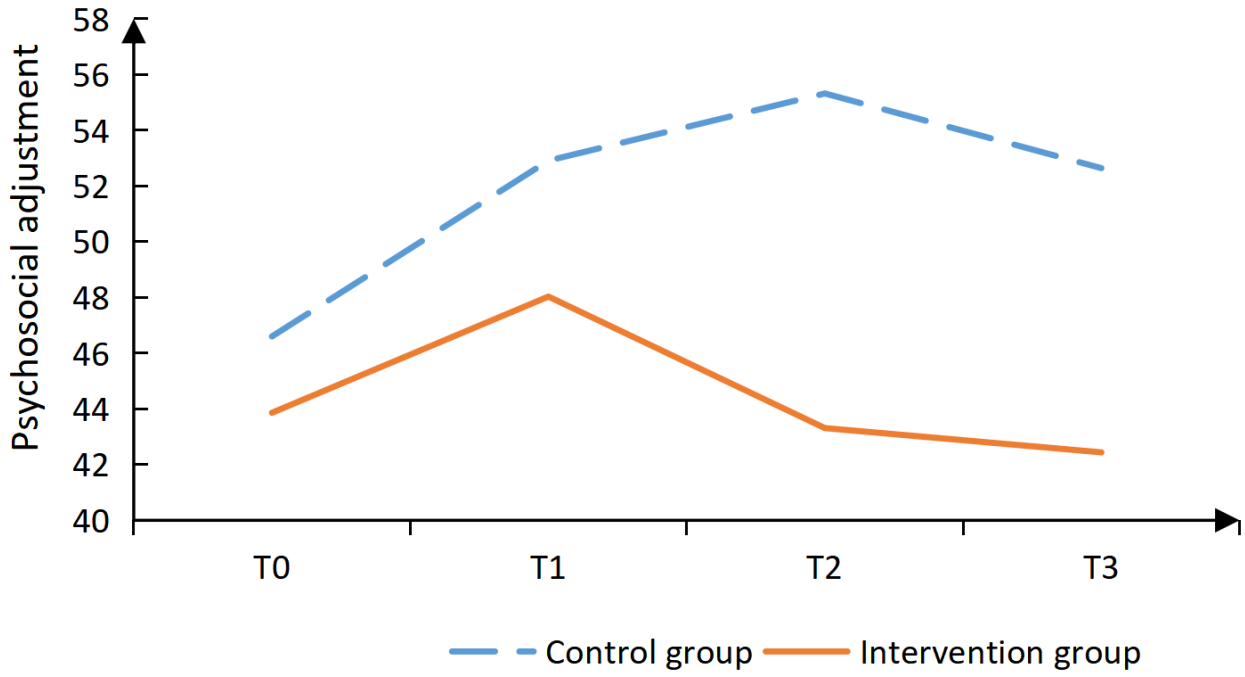

Figure 3. Results of the intervention effect evaluation: self-efficacy.

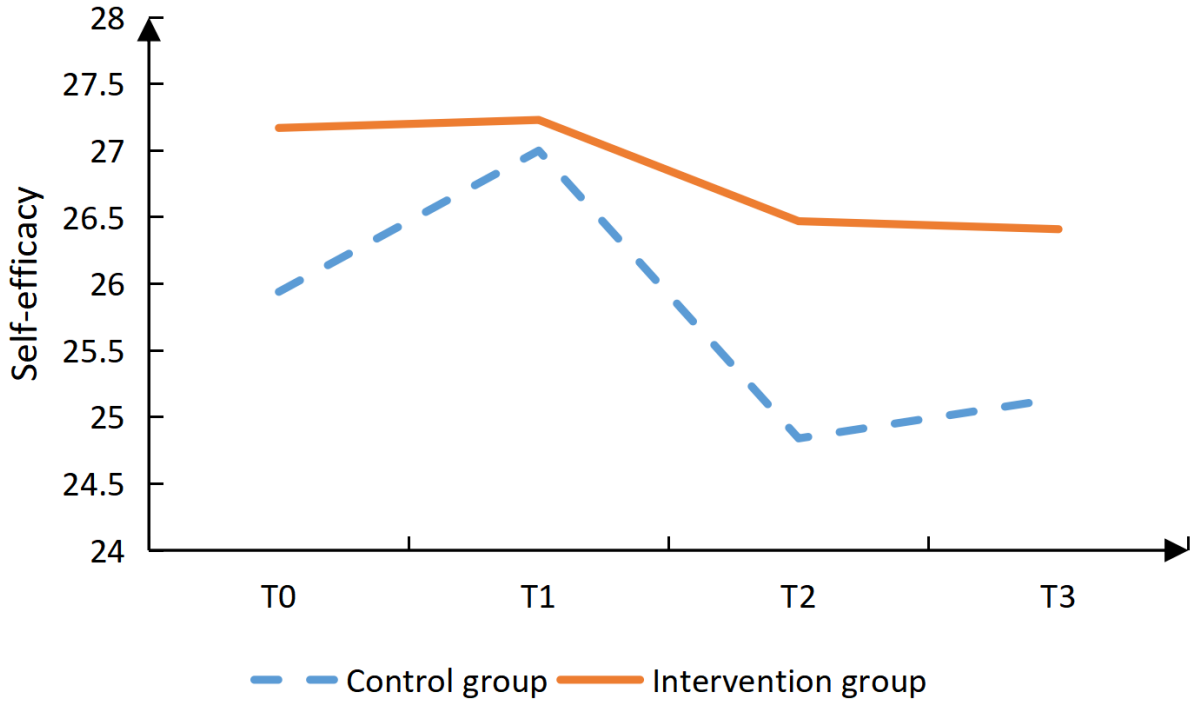

XSL•FO
RenderX

Figure 4. Results of the intervention effect evaluation: social support.

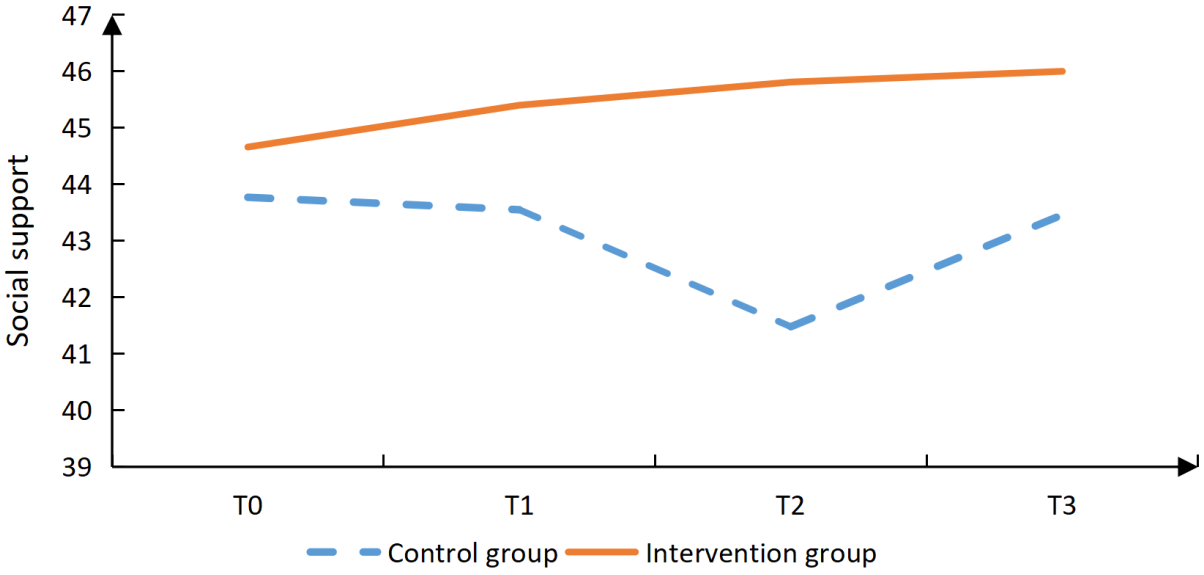

Figure 5. Results of the intervention effect evaluation: confrontation.

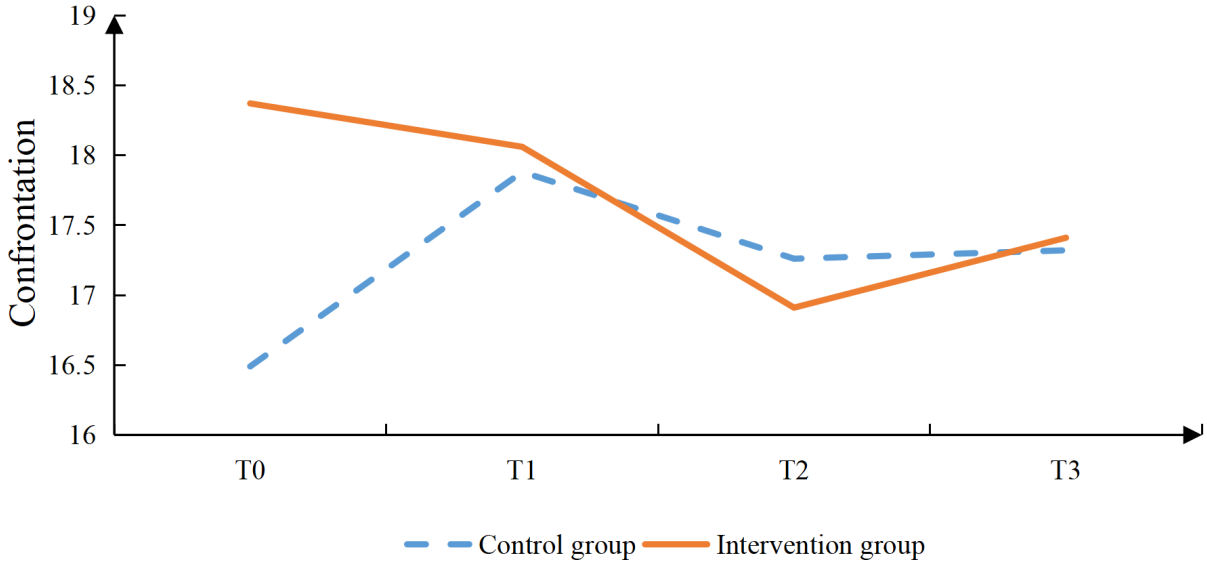

XSL•FO
RenderX

Figure 6. Results of the intervention effect evaluation: avoidance.

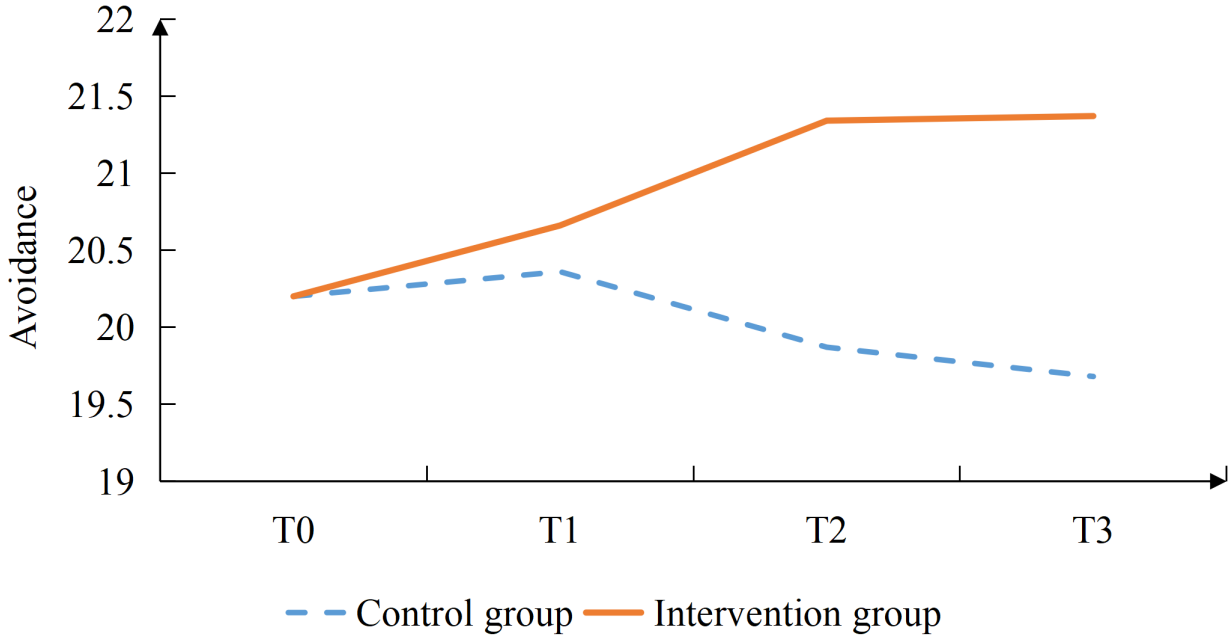

Figure 7. Results of the intervention effect evaluation: acceptance-resignation.

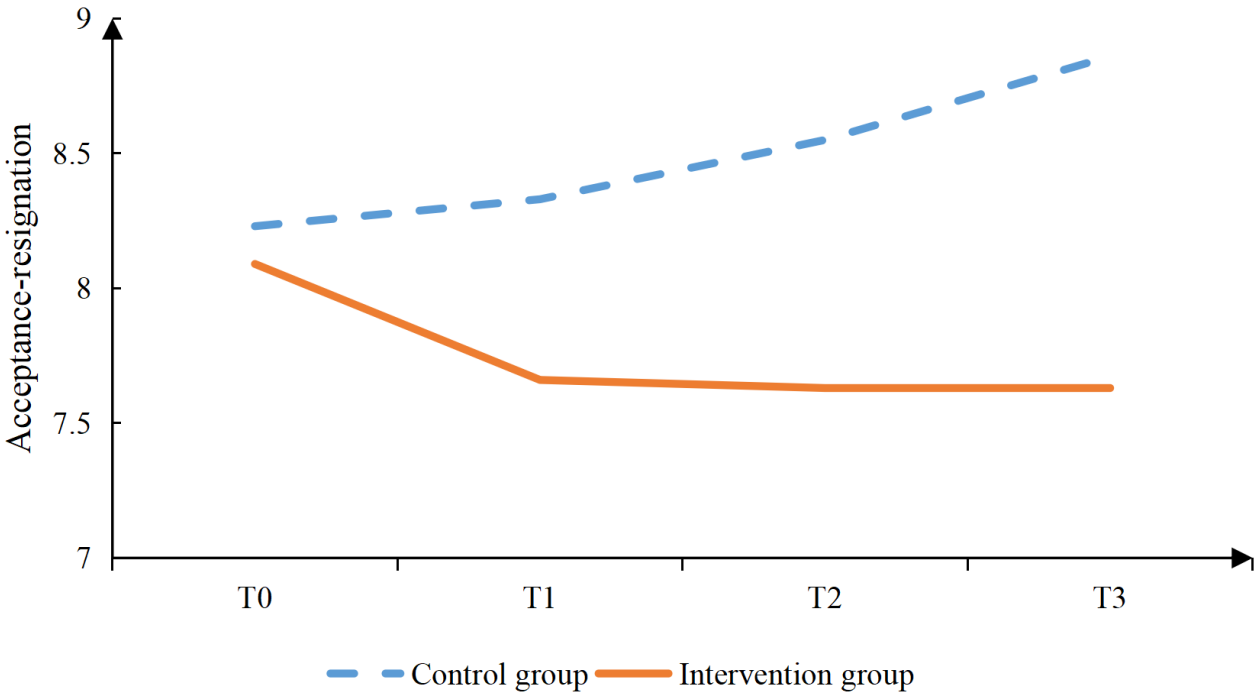

Treatment Burden

There was a statistically significant difference between the 2 groups in unplanned hospital visits (Z=−0.354; P=.01), indicating that the intervention group had fewer unplanned hospital visits than the control group (shown Multimedia Appendix 3).

Results of Qualitative Analysis

Participant Basic Information

A total of 9 participants were interviewed. Their ages ranged from 30 to 53 years. Five participants underwent breast-conserving surgery, while 4 underwent total resection. The number of times participants engaged in the intervention sessions ranged from 3 to 5. Interview durations varied between 21 and 50 minutes. The general information of participants interviewed is shown in Multimedia Appendix 4.

XSL•FO
RenderX

Participation Experience

Analysis of participants' evaluations of the peer support intervention for psychosocial adjustment in young- to middle-aged patients with breast cancer revealed 4 key themes: enhancing psychosocial adjustment, serving as a vital source of support, increasing treatment confidence, and promoting positive coping. The qualitative interview results of intervention participants are shown in Table 5.

XSL•FO
RenderX

Table . Qualitative interview results of intervention participants.

Theme and subtheme	Quotes
Promote psychosocial adjustment	
Diagnosis and treatment adjustment	• "I learned about others' experiences and how they differed from mine. Realizing I wasn't unique in this made me feel more at ease." (P1) • "At first, I didn't know anything about wound care or chemo. After joining the program, I started learning from the peers—how to eat properly, how to manage things." (P3)
Negative emotions adjustment	• "Initially, I was quite scared and felt a lot of pressure. After chatting with peer, those worries have largely faded." (P2) • "During the hardest times, chatting with the peers really made me feel so much lighter and more relaxed." (P7)
Sexuality and body image adjustment	• "After chemotherapy, my hair started falling out a lot, so I shaved it off and started wearing a wig. I had already bought the wig right after I attended the third session." (P5) • "I hadn't thought about sexual intimacy before, but after discussion, I talked to my husband. He was initially unsupportive, but later understood not to pressure me. Following their suggestions, we still kiss, hug, and hold hands." (P8)
Family role functioning adjustment	• "My family cares about me, but we don't discuss life issues much with peers; I talk more with friends or close confidants." (P3) • "After I got sick, my sister came to take care of me, so things at home are fine. I don't talk much about family stuff with the peers." (P4)
Social participation adjustment	• "They encouraged me to go out and work instead of staying home, which had a big impact. I've been working continuously." (P8) • "Following the advice from the peers, I cut back on pointless socializing. If chatting with friends does not feel right for my current state of mind, I just pause it for now. It makes things a lot less hard on my heart." (P9)
Serving as a vital source of support	
Informational support	• "When I was diagnosed, I didn't know a tumor meant cancer. Joining the sessions helped me learn a lot about the disease and post-surgery care." (P4) • "Most of them have already had kids, but I haven't yet, so a lot of people probably haven't been in my situation. One peer talked about egg freezing, and that really helped me—it gave me useful information. I think that was especially good." (P6)
Care and encouragement	• "They really care about us—they often ask if we're in pain, if we can eat, if we're throwing up." (P1) • "They're very responsible, often asking in the chat group how we're recovering or doing. Their heartfelt chats helped me through the initial depressive phase." (P2)
Space for expression	• "I became good friends with one of the group members—we chat every day now, even about our everyday lives." (P7) • "The Chat group is great—it gives me a space to communicate and vent. There are things I wouldn't say to my husband or mother, but I can share them privately with group members." (P9)
Increasing treatment confidence	
	• "Some people have it way worse than me and they've lived for over ten years. So I think to myself, my life won't be short—I'll definitely be okay and get to see my child grow up." (P1) • "Seeing others recover and return to work normally boosts my confidence. If they can live well, I can strive for it too and feel capable." (P3)
Promoting positive coping	

XSL•FO
RenderX

Theme and subtheme	Quotes
	• "Initially, I was in a bad mood and reluctant to join, but after attending, the open discussions really reduced my stress. Facing the disease head-on, following the doctor's treatment plan, and staying committed gives me hope for better days." (P2) • "I was always so scared and kept avoiding it—I cried a lot. But after the first peer meeting, when I heard the peers and nurses talk about it, I felt like there was nothing to be afraid of anymore. Just go ahead with the treatment." (P9)

Discussion

Principal Findings

Young- to middle-aged patients with breast cancer face significant psychosocial challenges, which can impair immune function and worsen prognosis [8,48]. Addressing these issues with effective interventions is crucial—the PNO2PSP notably improved psychosocial adjustment, supported by qualitative interviews. A similar trial by Høybye et al [49] found improvements in helplessness and depression, though without affecting broader mental adjustment or mood disturbance. The difference likely stems from the lack of trained peer supporters and structured interventions in that study [50]. In contrast, this study used various support formats, including peer education and meetings with nurses, ensuring more stable social networks and addressing real concerns. Peer support groups also offered a space for participants to communicate anytime and anywhere, addressing adjustment issues promptly [25,51].

In terms of feasibility, this study recruited participants preoperatively, with a total of 90 patients contacted and 70 (77.8%) consenting to participate. Attrition rates were 12.4% in the control group and 10.5% in the intervention group, primarily occurring during the follow-up phase after the intervention. During the first 8 weeks of the intervention, the overall participation rate in the intervention group was high at 86.3%. This high rate can be attributed to the intervention themes being tailored to the specific psychosocial adjustment challenges faced by young- to middle-aged patients with breast cancer. Additionally, the blended online and offline intervention model reduced logistical and financial burdens, allowing participants to participate from home [52]. The intervention's feasibility was further enhanced by its design. Participants scheduled for surgery were recruited during their hospital stay, allowing ample time for baseline assessments and the initial offline intervention session, which facilitated participant engagement. Expert consultations and focus group interviews determined the blended intervention format. The initial offline session established connections and trust between peers and participants, while subsequent online sessions minimized the need for hospital visits, increasing participation and motivation. The study established online support groups for participants, where the actual frequency of group discussions exceeded the initially planned schedule, reflecting a high level of participant engagement. This enthusiasm was largely driven by the peer supporters, who regularly followed up with participants in the online support groups, provided necessary information, and responded promptly to participants' questions. For issues requiring professional expertise beyond the scope of peer supporters, nurses stepped in to provide timely assistance. Furthermore, each online support group included at least 2 peer supporters, ensuring that if 1 was unavailable, the other could respond without delay.

PNO2PSP significantly improved the psychosocial adjustment of young- to middle-aged patients with breast cancer, a finding further supported by the results of qualitative interviews. Similar findings were reported by Høybye et al [49], who conducted a randomized controlled trial of an internet peer support group for patients with breast cancer. However, their study showed improvements in helplessness, confusion, and depression, with no significant effects on other dimensions of mental adjustment to cancer or total mood disturbance. The difference in results likely stems from the previous study's reliance on mutual support among participants without trained peer supporters [50]. Additionally, the intervention lacked clear themes and fixed schedules, and the online-only format made it difficult for participants to establish a stable social network. In contrast, this study adopted a variety of peer support formats. Peer education involved trained peers sharing their personal experiences and coping modes with the disease. Peer group meetings provided a platform for participants to interact with peer supporters and nurses, offering an opportunity for emotional expression. Peer support groups offered a space for participants to communicate anytime and anywhere, addressing adjustment issues promptly [25,51]. Moreover, the intervention themes in this study were developed based on specific adjustment issues identified through qualitative research, ensuring that they addressed real and relevant concerns faced by young- to middle-aged patients with breast cancer.

At the same time, the analysis of the subdimensions of psychosocial adjustment showed that the intervention improved participants' scores in the sexual relationship dimension. A previous study found that peer support did not effectively improve sexual relationships among patients with breast cancer [29]. The discrepancy may lie in the fact that previous peer support studies, even when addressing sexuality-related topics, focused only on knowledge dissemination, offering participants limited actionable advice [53]. In this study, through the use of intervention manuals, peer supporters' experience sharing, and nurses' education, participants received knowledge related to sexual health. This not only informed participants that they could engage in sexual activities postsurgery but also encouraged them to actively discuss and resume sexual activities [54]. Participants were provided with techniques to alleviate discomfort during sexual intercourse, and peer supporters shared practical ways to maintain intimacy beyond intercourse, such as kissing and hugging [55]. These actionable strategies

XSL•FO
RenderX

significantly improved sexual relationships. Influenced by traditional Chinese culture, sexual health education has long been a challenge in educating patients with breast cancer [14].

However, because peer supporters in this study were all women of similar age and with similar experiences, participants could freely discuss sexuality-related issues without feeling embarrassed [55]. Therefore, peer support could be a viable approach for improving sexual health among patients with breast cancer in the future. The intervention also improved participants' psychological distress scores. Interviews revealed that participants believed peer support helped alleviate negative emotions, reducing their suffering, anxiety, and fear, which aligns with the findings of previous studies. A previous systematic review on peer support for patients with cancer indicated that peer support interventions effectively reduce depression and anxiety [56]. In many Asian societies, cultural norms surrounding emotional expression, illness disclosure, and discussion of family matters may influence how openly patients communicate within peer support groups. This cultural context may partly explain the limited effects observed in family-related dimensions of psychosocial adjustment [57]. However, there was no statistically significant improvement in the social relationship dimension, possibly due to the short follow-up period. At the end of follow-up, many participants were still in treatment and had not yet resumed social reintegration. Nonetheless, qualitative interviews showed that participants believed peer support could promote their occupational and social adjustment.

The findings revealed that the intervention could enhance the levels of social support, consistent with the results of previous studies [50,58]. In the design of this study, peer support provided participants with informational, emotional, and appraisal support. When participants had questions about their disease or treatment, peer support offered informational support and guided them to seek professional support, family support, and other social support resources [25]. When participants experienced negative emotions due to illness and treatment, peer supporters provided emotional support through listening, comforting, and encouraging, which improved individuals' perceived levels of social support [59]. Additionally, appraisal support helped participants reevaluate the support resources they possessed, enhancing their subjective perception of social support [60]. This program also facilitated the adoption of problem-focused coping strategies while reducing the use of avoidance coping strategies. The interview results with participants supported this finding. A prior study on peer support for adolescent patients with cancer [61] yielded similar results, where participants who received peer support adopted more coping strategies 3 months postsurgery. The reasons for these outcomes include peer supporters sharing their own experiences in addressing psychosocial issues, providing participants with reference points that reduced their fear and avoidance of disease-related challenges [62]. Through the intervention, participants' attitudes toward their illness underwent a significant transformation—they became more willing to confront cancer directly, ceased avoiding cancer-related issues, actively sought ways to resolve problems, and approached their condition with greater positivity and optimism [63].

Our qualitative findings supported that the intervention contributed to increasing treatment confidence among participants. However, the intervention did not produce a statistically significant improvement in self-efficacy among young- to middle-aged patients with breast cancer. Similarly, a prior study [64] reported that a 6-month peer counseling intervention for patients with breast cancer failed to significantly enhance their self-efficacy. A systematic review on peer support interventions for patients with cancer also found that peer support interventions did not improve participants' self-efficacy [56]. However, a previous study demonstrated that a 6-week dyadic peer intervention improved patients with breast cancer self-management self-efficacy among patients with breast cancer [65]. The potential reasons for this discrepancy may be related to the measurement tool used in this study. The GSES used may lack the sensitivity needed to detect changes in the self-efficacy specific to coping with illness [66]. Mechanistically, self-efficacy relies on structured mastery experiences, whereas our program emphasized peer empathy and informational support over skill-based tasks. Furthermore, the robust external support from nurses and peers may have improved adjustment but temporarily bypassed the need for participants to develop independent self-reliance. Alternatively, peer support may primarily enhance disease-specific self-efficacy, which has a limited impact on overall self-efficacy [65]. Another possible explanation is that self-efficacy tends to be a stable personal attribute that is difficult to change over a short period [56].

The study results also showed that the intervention significantly reduced the number of unplanned hospital visits among participants. Similarly, a previous study [67] found that peer support interventions for patients after coronary artery bypass surgery led to reduced health care usage 3 months postsurgery, with fewer visits to general practitioners and emergency departments. This reduction is likely attributable to the synergy between peer supporters, who provided immediate experiential guidance for common postdischarge concerns, and intervention nurses, who offered professional triage and symptom management advice. This collaborative approach effectively addressed minor complications that might otherwise have triggered unnecessary hospital visits. This reduction in hospital readmissions may help alleviate patients' financial burden and reduce the strain on health care systems. Although the result confirmed the robustness of this finding, the reliance on self-reported health care usage data remains a limitation. Future large-scale trials should integrate clinical records with a formal economic framework to conduct a comprehensive cost-utility analysis.

Limitations

This study has certain limitations. This study used a single-center CRCT with a limited sample size (N=70), which may restrict the generalizability of the findings. Furthermore, as the primary outcomes were based on self-reported scales, this lack of blinding may have introduced potential self-report bias. In addition, participant attrition during the follow-up period may further limit the robustness and generalizability of the results. Future research should consider multicenter, large-sample trials to further evaluate the effectiveness of peer support. Although this study focused on early postoperative

XSL•FO
RenderX

intervention and follow-up, extending the intervention period in future research may further enhance participants' psychosocial adjustment during the recovery phase. Furthermore, this study only collected self-reported patient expenditures without performing a comprehensive economic analysis.

Conclusion

Peer support represents a potentially more cost-effective and sustainable form of support. This study developed a PNO2PSP for psychosocial adjustment in young- to middle-aged patients with breast cancer, guided by a complex intervention framework. The PNO2PSP demonstrates strong scientific validity and feasibility. It effectively improves the psychosocial adjustment levels of young- to middle-aged patients with breast cancer, enhances social support, and promotes positive coping modes. By continuously training peer supporters and conducting peer support interventions, the psychosocial adjustment of young- to middle-aged patients with breast cancer can be further promoted, and this PNO2PSP could be standardized and widely implemented in clinical practice in the future.

Acknowledgments

We would like to express our sincere gratitude to all the patients who participated in this study for their time and contribution. We also extend our heartfelt thanks to the peer supporters and intervention nurses for their dedication and essential role in implementing the peer-led, nurse-involved, blended online and offline peer support intervention program (PNO2PSP). Their commitment was vital to the successful completion of this research. The manuscript was prepared without the use of generative artificial intelligence.

Funding

This study was supported by The National Nature Science Foundation of China (grant nos 71974218 and 72374232), the Guangdong Basic and Applied Basic Research Foundation (grant no 2026A1515011766), the Humanities and Social Science Youth Foundation of Ministry of Education of China (grant no 25YJCZH379), the Fundamental Research Funds for the Central Universities (grant no 21625303), and the Medical Scientific Research Foundation of Guangdong Province, China (grant no 20241121123222794).

Data Availability

The datasets generated or analyzed during this study are not publicly available but are available from the corresponding author on reasonable request.

Authors' Contributions

YZ contributed to conceptualization, methodology, formal analysis, and writing (original draft and review and editing). HH contributed to investigation, data curation, and visualization. YZ contributed to formal analysis and writing (original draft and review and editing). YZ, CL, JC, and LC contributed to investigation, data curation, and visualization. NZ contributed to data curation and formal analysis. SS, BZ, and JF contributed to data curation, formal analysis, and visualization. JY contributed to project administration, data curation, and visualization. MZ contributed to supervision, conceptualization, and writing (review and editing).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Schedule and details of the peer-led, nurse-involved, blended online and offline peer support intervention program (PNO2PSP).
[DOCX File, 16 KB - jmir_v28i1e86097_app1.docx]

Multimedia Appendix 2
Results of the intervention effect evaluation of the subdimensions of psychosocial adjustment.
[DOCX File, 420 KB - jmir_v28i1e86097_app2.docx]

Multimedia Appendix 3
Comparison of secondary outcomes between the intervention group and the control group at different time points.
[DOCX File, 14 KB - jmir_v28i1e86097_app3.docx]

Multimedia Appendix 4
General information of participants interviewed (N=9).
[DOCX File, 13 KB - jmir_v28i1e86097_app4.docx]

XSL•FO
RenderX

Checklist 1
CONSORT 2010 checklist of information to include when reporting a cluster randomized trial.
[PDF File, 195 KB - jmir_v28i1e86097_app5.pdf]

References

1. Global cancer observatory: cancer today. International Agency for Research on Cancer. URL: https://gco.iarc.who.int/today/ [accessed 2025-06-16]
2. Rosenberg SM, Dominici LS, Gelber S, et al. Association of breast cancer surgery with quality of life and psychosocial well-being in young breast cancer survivors. JAMA Surg 2020 Nov 1;155(11):1035-1042. [doi: 10.1001/jamasurg.2020.3325] [Medline: 32936216]
3. Álvarez-Pardo S, De Paz JA, Montserrat Romero-Pérez E, et al. Factors associated with body image and self-esteem in mastectomized breast cancer survivors. IJERPH 2022 Jan 15;20(6):5154. [doi: 10.3390/ijerph20065154]
4. Baudry AS, Yakimova S, Congard A, et al. Adjustment of young women with breast cancer after chemotherapy: a mediation model of emotional competence via emotional distress. Psychooncology 2022 May;31(5):848-855. [doi: 10.1002/pon.5876] [Medline: 34989072]
5. Sperlich S, Noeres D, Holthausen-Markou S, Park-Simon TW, Sahiti E, Geyer S. Social participation of women with breast cancer compared to the general population 5 years after primary surgery-what role do medical data and cancer-related complaints play? Support Care Cancer 2024 Aug 2;32(8):566. [doi: 10.1007/s00520-024-08695-w] [Medline: 39093481]
6. Zhang Y, Zhang X, Li N, et al. Factors associated with psychosocial adjustment in newly diagnosed young to middle-aged women with breast cancer: a cross-sectional study. Eur J Oncol Nurs 2023 Aug;65:102357. [doi: 10.1016/j.ejon.2023.102357] [Medline: 37321131]
7. Dossa F, Cusimano MC, Sutradhar R, et al. Real-world health services utilisation and outcomes after BRCA1 and BRCA2 testing in Ontario, Canada: the What Comes Next Cohort Study protocol. BMJ Open 2018 Sep 4;8(9):e025317. [doi: 10.1136/bmjopen-2018-025317] [Medline: 30181190]
8. Market M, Baxter KE, Angka L, Kennedy MA, Auer RC. The potential for cancer immunotherapy in targeting surgery-induced natural killer cell dysfunction. Cancers (Basel) 2018 Dec 20;11(1):2. [doi: 10.3390/cancers11010002] [Medline: 30577463]
9. Martens K, Ulrich GR, Ranby KW, Kilbourn K. What matters most? Predictors of quality of life and life satisfaction among young breast cancer survivors. Cancer Nurs 2021;44(6):E727-E734. [doi: 10.1097/NCC.0000000000000967] [Medline: 34694091]
10. Jang M, Kim J. A structural model for stress, coping, and psychosocial adjustment: a multi-group analysis by stages of survivorship in Korean women with breast cancer. Eur J Oncol Nurs 2018 Apr;33:41-48. [doi: 10.1016/j.ejon.2018.01.004] [Medline: 29551176]
11. Zhang X, Gong N, Li N, et al. Why breast cancer patients avoid communicating disease-related information to their dependent children: a qualitative study. J Clin Nurs 2023 Apr;32(7-8):1230-1239. [doi: 10.1111/jocn.16281] [Medline: 35403320]
12. Choi E, Henneghan AM. Comparing fatigue, loneliness, daytime sleepiness, and stress in younger and older breast cancer survivors: a cross-sectional analysis. Clin J Oncol Nurs 2022 Apr 1;26(2):155-164. [doi: 10.1188/22.CJON.155-164] [Medline: 35302543]
13. Campbell-Enns HJ, Woodgate RL. The psychosocial experiences of women with breast cancer across the lifespan: a systematic review. Psychooncology 2017 Nov;26(11):1711-1721. [doi: 10.1002/pon.4281] [Medline: 27648726]
14. Gong N, Zhang Y, Suo R, Dong W, Zou W, Zhang M. The role of space in obstructing clinical sexual health education: a qualitative study on breast cancer patients' perspectives on barriers to expressing sexual concerns. European J Cancer Care 2021 Jul;30(4). [doi: 10.1111/ecc.13422]
15. Hubbeling HG, Rosenberg SM, González-Robledo MC, et al. Psychosocial needs of young breast cancer survivors in Mexico City, Mexico. PLoS One 2018;13(5):e0197931. [doi: 10.1371/journal.pone.0197931] [Medline: 29787612]
16. Mannion S, Higgins A, Larson N, et al. Prevalence and impact of fertility concerns in young women with breast cancer. Sci Rep 2024 Feb 22;14(1):4418. [doi: 10.1038/s41598-024-54961-6] [Medline: 38388636]
17. Fenlon D, Nuttall J, May C, et al. MENOS4 trial: a multicentre randomised controlled trial (RCT) of a breast care nurse delivered cognitive behavioural therapy (CBT) intervention to reduce the impact of hot flushes in women with breast cancer: Study protocol. BMC Womens Health 2018 May 8;18(1):63. [doi: 10.1186/s12905-018-0550-z] [Medline: 29739384]
18. Haller H, Winkler MM, Klose P, Dobos G, Kümmel S, Cramer H. Mindfulness-based interventions for women with breast cancer: an updated systematic review and meta-analysis. Acta Oncol 2017 Dec;56(12):1665-1676. [doi: 10.1080/0284186X.2017.1342862] [Medline: 28686520]
19. Wu PH, Chen SW, Huang WT, Chang SC, Hsu MC. Effects of a psychoeducational intervention in patients with breast cancer undergoing chemotherapy. J Nurs Res 2018 Aug;26(4):266-279. [doi: 10.1097/jnr.0000000000000252] [Medline: 29360672]
20. Moghaddam Tabrizi F, Alizadeh S. Family intervention based on the FOCUS program effects on cancer coping in Iranian breast cancer patients: a randomized control trial. Asian Pac J Cancer Prev 2018 Jun 25;19(6):1523-1528. [doi: 10.22034/APJCP.2018.19.6.1523] [Medline: 29936726]

XSL•FO
RenderX

21. Boesen EH, Karlsen R, Christensen J, et al. Psychosocial group intervention for patients with primary breast cancer: a randomised trial. Eur J Cancer 2011 Jun;47(9):1363-1372. [doi: 10.1016/j.ejca.2011.01.002] [Medline: 21458989]
22. Kirshbaum MN, Dent J, Stephenson J, et al. Open access follow-up care for early breast cancer: a randomised controlled quality of life analysis. Eur J Cancer Care 2017 Jul;26(4):e12577. [doi: 10.1111/ecc.12577]
23. Zhang Y, Yan J, He H, et al. The trajectories of psychosocial adjustment among young to middle-aged women with breast cancer: a prospective longitudinal study. Eur J Oncol Nurs 2024 Aug;71:102617. [doi: 10.1016/j.ejon.2024.102617] [Medline: 38865852]
24. Dennis CL. Peer support within a health care context: a concept analysis. Int J Nurs Stud 2003 Mar;40(3):321-332. [doi: 10.1016/s0020-7489(02)00092-5] [Medline: 12605954]
25. Hu J, Wang X, Guo S, et al. Peer support interventions for breast cancer patients: a systematic review. Breast Cancer Res Treat 2019 Apr;174(2):325-341. [doi: 10.1007/s10549-018-5033-2] [Medline: 30600413]
26. Jung M, Lee SB, Lee JW, et al. The impact of a mobile support group on distress and physical activity in breast cancer survivors: randomized, parallel-group, open-label, controlled trial. J Med Internet Res 2023 Aug 7;25:e47158. [doi: 10.2196/47158] [Medline: 37549004]
27. Li J, Peng X, Su Y, He Y, Zhang S, Hu X. Effectiveness of psychosocial interventions for posttraumatic growth in patients with cancer: a meta-analysis of randomized controlled trials. Eur J Oncol Nurs 2020 Oct;48:101798. [doi: 10.1016/j.ejon.2020.101798] [Medline: 32688246]
28. Nápoles AM, Santoyo-Olsson J, Stewart AL, et al. Nuevo Amanecer-II: results of a randomized controlled trial of a community-based participatory, peer-delivered stress management intervention for rural Latina breast cancer survivors. Psychooncology 2020 Nov;29(11):1802-1814. [doi: 10.1002/pon.5481] [Medline: 32672864]
29. Weiner LS, Nagel S, Irene Su H, et al. A remotely delivered, peer-led intervention to improve physical activity and quality of life in younger breast cancer survivors. J Behav Med 2023 Aug;46(4):578-593. [doi: 10.1007/s10865-022-00381-8] [Medline: 36479658]
30. Jansen BA, Bargon CA, Dinger TL, et al. Breast cancer patients' needs and perspectives on a one-on-one peer support program: quantitative and qualitative analyses. Support Care Cancer 2023 Oct 26;31(12):656. [doi: 10.1007/s00520-023-08009-6] [Medline: 37882849]
31. Clougher D, Ciria-Suarez L, Medina JC, Anastasiadou D, Racioppi A, Ochoa-Arnedo C. What works in peer support for breast cancer survivors: a qualitative systematic review and meta-ethnography. Appl Psychol Health Well Being 2024 May;16(2):793-815. [doi: 10.1111/aphw.12473] [Medline: 37493002]
32. Jablotschkin M, Binkowski L, Markovits Hoopii R, Weis J. Benefits and challenges of cancer peer support groups: a systematic review of qualitative studies. Eur J Cancer Care (Engl) 2022 Nov;31(6):e13700. [doi: 10.1111/ecc.13700] [Medline: 36104303]
33. Zordan RD, Juraskova I, Butow PN, et al. Exploring the impact of training on the experience of Australian support group leaders: current practices and implications for research. Health Expect 2010 Dec;13(4):427-440. [doi: 10.1111/j.1369-7625.2010.00592.x] [Medline: 20550596]
34. Wood E, Trasolini A, Thomas N. Barriers and facilitators to implementing and sustaining peer support in kidney care. J Ren Care 2022 Jun;48(2):128-138. [doi: 10.1111/jorc.12394] [Medline: 34275206]
35. Kallio R, Jones M, Pietilä I, Harju E. Perspectives of oncology nurses on peer support for patients with cancer. Eur J Oncol Nurs 2021 Apr;51:101925. [doi: 10.1016/j.ejon.2021.101925] [Medline: 33647543]
36. Huber J, Muck T, Maatz P, et al. Face-to-face vs. online peer support groups for prostate cancer: a cross-sectional comparison study. J Cancer Surviv 2018 Feb;12(1):1-9. [doi: 10.1007/s11764-017-0633-0] [Medline: 28861845]
37. Bender JL, Katz J, Ferris LE, Jadad AR. What is the role of online support from the perspective of facilitators of face-to-face support groups? A multi-method study of the use of breast cancer online communities. Patient Educ Couns 2013 Dec;93(3):472-479. [doi: 10.1016/j.pec.2013.07.009] [Medline: 23928354]
38. Liang D, Jia R, Yu J, Wu Z, Chen C, Lu G. The effect of remote peer support on stigma in patients after breast cancer surgery during the COVID-19 pandemic. Medicine (Abingdon) 2021 Jun 18;100(24):e26332. [doi: 10.1097/MD.0000000000026332]
39. Jang L. Effect of peer support on psychosocial adaptation in breast cancer patients. Nanchan University. 2020 URL: https://oversea.cnki.net/kcms2/article/[illegible] [accessed 2026-04-02] [doi: 10.27232/d.cnki.gnchu.2020.000630]
40. Caminiti C, Annunziata MA, Verusio C, et al. Effectiveness of a psychosocial care quality improvement strategy to address quality of life in patients with cancer: the HuCare2 stepped-wedge cluster randomized trial. JAMA Netw Open 2021 Oct 1;4(10):e2128667. [doi: 10.1001/jamanetworkopen.2021.28667] [Medline: 34648011]
41. Meurer WJ, Lewis RJ. Cluster randomized trials. JAMA 2015 May 26;313(20):2068. [doi: 10.1001/jama.2015.5199]
42. Luszczynska A, Scholz U, Schwarzer R. The general self-efficacy scale: multicultural validation studies. J Psychol 2005 Sep;139(5):439-457. [doi: 10.3200/JRLP.139.5.439-457] [Medline: 16285214]
43. Tao Y, Yu H, Liu S, et al. Hope and depression: the mediating role of social support and spiritual coping in advanced cancer patients. BMC Psychiatry 2022 Dec;22(1):345. [doi: 10.1186/s12888-022-03985-1]

XSL•FO
RenderX

44. Feifel H, Strack S, Nagy VT. Coping strategies and associated features of medically ill patients. Psychosom Med 1987;49(6):616-625. [doi: 10.1097/00006842-198711000-00007] [Medline: 3423168]
45. Hennink MM, Kaiser BN, Marconi VC. Code saturation versus meaning saturation. Qual Health Res 2017 Mar;27(4):591-608. [doi: 10.1177/1049732316665344]
46. O'Cathain A, Murphy E, Nicholl J. Three techniques for integrating data in mixed methods studies. BMJ 2010 Sep 17;341(sep17 1):c4587-c4587. [doi: 10.1136/bmj.c4587]
47. Katzmarzyk PT, Denstel KD, Martin CK, et al. Intraclass correlation coefficients for weight loss cluster randomized trials in primary care: the PROPEL trial. Clin Obes 2022 Aug;12(4):e12524. [doi: 10.1111/cob.12524] [Medline: 35412010]
48. Parker PA, Youssef A, Walker S, et al. Short-term and long-term psychosocial adjustment and quality of life in women undergoing different surgical procedures for breast cancer. Ann Surg Oncol 2007 Nov;14(11):3078-3089. [doi: 10.1245/s10434-007-9413-9] [Medline: 17574501]
49. Høybye MT, Dalton SO, Deltour I, Bidstrup PE, Frederiksen K, Johansen C. Effect of Internet peer-support groups on psychosocial adjustment to cancer: a randomised study. Br J Cancer 2010 Apr 27;102(9):1348-1354. [doi: 10.1038/sj.bjc.6605646] [Medline: 20424614]
50. Aizpurua-Perez I, Arregi A, Gonzalez D, et al. A randomized controlled trial of the effectiveness of a one-to-one peer support intervention on resilience, social support, and salivary cortisol in recently diagnosed women with breast cancer. Eur J Oncol Nurs 2024 Aug;71:102616. [doi: 10.1016/j.ejon.2024.102616] [Medline: 38885598]
51. Hunter-Smith A, Cuthbert CA, Fergus K, et al. Examining social media peer support and improving psychosocial outcomes for young women with breast cancer. JCO 2021 Oct 1;39(28_suppl):162-162. [doi: 10.1200/JCO.2020.39.28_suppl.162]
52. Boyes A, Turon H, Hall A, Watson R, Proietto A, Sanson-Fisher R. Preferences for models of peer support in the digital era: a cross‐sectional survey of people with cancer. Psychooncology 2018 Sep;27(9):2148-2154. [doi: 10.1002/pon.4781] [Medline: 29808504]
53. Taylor S, Harley C, Absolom K, Brown J, Velikova G. Breast cancer, sexuality, and intimacy: addressing the unmet need. Breast J 2016 Jul;22(4):478-479. [doi: 10.1111/tbj.12614] [Medline: 27089466]
54. Reese JB, Lepore SJ, Sorice KA, et al. Efficacy of a couple-based intervention addressing sexual concerns for breast cancer survivors: results of a randomized controlled trial. Cancer 2025 Jan 1;131(1):e35685. [doi: 10.1002/cncr.35685] [Medline: 39645589]
55. Teo I, Tan YP, Goh A, et al. The renewing intimacy and sexuality intervention for female cancer survivors: a feasibility study. Psychooncology 2024 Dec;33(12):e70037. [doi: 10.1002/pon.70037] [Medline: 39638564]
56. Zhang S, Li J, Hu X. Peer support interventions on quality of life, depression, anxiety, and self-efficacy among patients with cancer: a systematic review and meta-analysis. Patient Educ Couns 2022 Nov;105(11):3213-3224. [doi: 10.1016/j.pec.2022.07.008] [Medline: 35858869]
57. Chou FY, Lee-Lin F, Kuang LY. The effectiveness of support groups in Asian breast cancer patients: an integrative review. Asia Pac J Oncol Nurs 2016;3(2):157-169. [doi: 10.4103/2347-5625.162826] [Medline: 27981154]
58. Presley C, Agne A, Shelton T, Oster R, Cherrington A. Mobile-enhanced peer support for African Americans with type 2 diabetes: a randomized controlled trial. J Gen Intern Med 2020 Oct;35(10):2889-2896. [doi: 10.1007/s11606-020-06011-w] [Medline: 32700215]
59. Toija A, Kettunen T, Kasila K. The construction of peer support among recently diagnosed breast cancer patients. Cancer Nurs 2025;48(4):E238-E245. [doi: 10.1097/NCC.0000000000001319] [Medline: 38527125]
60. Clougher D, Ciria-Suarez L, Medina JC, Anastasiadou D, Racioppi A, Ochoa-Arnedo C. What works in peer support for breast cancer survivors: a qualitative systematic review and meta‐ethnography. Appl Psychol Health Well Being 2024 May;16(2):793-815. [doi: 10.1111/aphw.12473] [Medline: 37493002]
61. Sansom-Daly UM, Wakefield CE, Ellis SJ, et al. Online, group-based psychological support for adolescent and young adult cancer survivors: results from the recapture life randomized trial. Cancers (Basel) 2021 May 18;13(10):2460. [doi: 10.3390/cancers13102460] [Medline: 34070134]
62. Silva SM, Crespo C, Canavarro MC. Pathways for psychological adjustment in breast cancer: a longitudinal study on coping strategies and posttraumatic growth. Psychol Health 2012;27(11):1323-1341. [doi: 10.1080/08870446.2012.676644] [Medline: 22490001]
63. Li L, Li S, Wang Y, et al. Coping profiles differentiate psychological adjustment in Chinese women newly diagnosed with breast cancer. Integr Cancer Ther 2017 Jun;16(2):196-204. [doi: 10.1177/1534735416646854] [Medline: 27154183]
64. Giese-Davis J, Bliss-Isberg C, Carson K, et al. The effect of peer counseling on quality of life following diagnosis of breast cancer: an observational study. Psychooncology 2006 Nov;15(11):1014-1022. [doi: 10.1002/pon.1037] [Medline: 16555366]
65. Lee R, Lee KS, Oh EG, Kim SH. A randomized trial of dyadic peer support intervention for newly diagnosed breast cancer patients in Korea. Cancer Nurs 2013;36(3):E15-E22. [doi: 10.1097/NCC.0b013e3182642d7c] [Medline: 22964869]
66. Ko HK, Park GJ. Effects of self-efficacy promotion program on self-efficacy, self-care behavior, and quality of life in breast cancer patients receiving radiotherapy. J Korean Oncol Nurs 2011;11(2):136. [doi: 10.5388/jkon.2011.11.2.136]
67. Colella TJ, King-Shier K. The effect of a peer support intervention on early recovery outcomes in men recovering from coronary bypass surgery: a randomized controlled trial. Eur J Cardiovasc Nurs 2018 Jun;17(5):408-417. [doi: 10.1177/1474515117725521] [Medline: 28805455]

XSL•FO
RenderX

Abbreviations

CRCT: cluster randomized controlled trial
DE: design effect
GEE: generalized estimating equation
GSES: General Self-Efficacy Scale
MCMQ: Medical Coping Modes Questionnaire
PAIS: Psychosocial Adjustment to Illness Scale
PNO2PSP: peer-led, nurse-involved, blended online and offline peer support intervention program
SSRS: Social Support Rate Scale

Edited by M Balcarras; submitted 18.Oct.2025; peer-reviewed by M Ekwueme, WN Sun; accepted 02.Mar.2026; published 17.Apr.2026.

Please cite as:
Zhang Y, He H, Zhang Y, Zhong Y, Luo C, Chen J, Chen L, Zhang N, Sun S, Zhang B, Fang J, Yan J, Zhang M
A Peer-Led, Nurse-Involved Blended Online and Offline Peer Support Program (PNO2PSP) for Psychosocial Adjustment in Young- to Middle-Aged Patients With Breast Cancer: Cluster Randomized Clinical Trial
J Med Internet Res 2026;28:e86097
URL: https://www.jmir.org/2026/1/e86097
doi:10.2196/86097

© Yiheng Zhang, Haiyan He, Ye Zhang, Youxian Zhong, Chunhong Luo, Juanjuan Chen, Lili Chen, Ni Zhang, Shihao Sun, Baoyi Zhang, Jia Fang, Jingwen Yan, Meifen Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Health–Based Peer Support Ecosystem for Gestational Diabetes Mellitus in Vietnam (VALID II Study): Multistakeholder Cocreation and Pilot Study

Ngoc Anh Thi Dang[1], MPH; Dan W Meyrowitsch[2], MSc, PhD, Prof Dr; Ditte S Linde[3,4,5], MA, PhD; Christina A Vinter[3,4,6], MD, PhD, Prof Dr; Thanh Duc Nguyen[7], MD, PhD; Dang Kien Nguyen[8], MD, PhD; Thi Ai Nguyen[7], MPH; Hieu Minh Le[9], MD; Thi Kim Dung Vu[7], MPH; Thi Huong Tran[10], MEd; Tine M Gammeltoft[11], MA, PhD, Prof Dr; Jens Søndergaard[12], MD, PhD, Prof Dr

[1]Department of Environmental Health, Faculty of Public Health, Thai Binh University of Medicine and Pharmacy, 373 Ly Bon Street, Tran Lam Ward, Hung Yen Province, Vietnam

[10]Department of International Relations, Thai Binh University of Medicine and Pharmacy, Hung Yen Province, Vietnam

[11]Department of Anthropology, University of Copenhagen, Copenhagen K, Denmark

[12]Research Unit of General Practice, Department of Public Health, University of Southern Denmark, Odense, Denmark

[2]Department of Public Health, Global Health Section, University of Copenhagen, Copenhagen K, Denmark

[3]Department of Gynaecology and Obstetrics, Odense University Hospital, Odense, Denmark

[4]Research Unit for Gynaecology and Obstetrics, Department of Clinical Research, University of Southern Denmark, Odense, Denmark

[5]Centre for Global Health, Department of Clinical Research, University of Southern Denmark, Odense, Denmark

[6]Steno Diabetes Center, Odense University Hospital, Odense, Denmark

[7]Faculty of Public Health, Thai Binh University of Medicine and Pharmacy, Hung Yen Province, Vietnam

[8]Department of Gynaecology and Obstetrics, Thai Binh University of Medicine and Pharmacy, Hung Yen Province, Vietnam

[9]Department of Internal Medicine, Thai Binh University of Medicine and Pharmacy, Hung Yen Province, Vietnam

Corresponding Author:
Ngoc Anh Thi Dang, MPH
Department of Environmental Health, Faculty of Public Health, Thai Binh University of Medicine and Pharmacy, 373 Ly Bon Street, Tran Lam Ward, Hung Yen Province, Vietnam

Abstract

Background: Gestational diabetes mellitus (GDM) is associated with fetal overgrowth, increased risks of perinatal morbidity and mortality, and long-term complications for mother and child, including cardiovascular disease and type 2 diabetes. Innovative, peer-based digital health interventions are emerging globally as a potential approach to assist and empower women in effective self-care and well-being during pregnancy. However, there remains substantial potential to develop and evaluate culturally sensitive digital health interventions for pregnant women with GDM, especially in low- and middle-income countries.

Objective: This study aimed to bridge the gap between the World Health Organization (WHO) self-care framework and local practice by designing and piloting the "Healthy Pregnancy" intervention, a multiplatform digital ecosystem for women with GDM in northern Vietnam, through a staged cocreation and pilot refinement process.

Methods: Between December 2022 and February 2024, drawing on the WHO's conceptual framework for self-care and cocreation approach, we iteratively developed "Healthy Pregnancy," a digital health intervention, in 4 stages: (1) formative studies and self-care construct prioritization, (2) cocreation processes with key stakeholders, (3) development and design translation, and (4) pilot testing and final refinement.

Results: In stage 1, we identified gaps in current digital health interventions in low- and middle-income countries, explored the sociocultural realities of women with GDM in Vietnam, and prioritized 7 self-care constructs. In stage 2, we conducted a cocreation workshop to enable key stakeholders to co-design the foundational infrastructure for the potential intervention. In stage 3, we established a multicomponent digital intervention ecosystem with explicitly defined operating workflows. Finally, in stage 4, we gathered suggestions for a digital health intervention from a pilot test group of pregnant women with GDM, to refine and optimize the system schematic and information flow before moving to the real intervention.

Conclusions: By applying a cocreation approach across all stages of development of the "Healthy Pregnancy" digital health intervention, from problem identification to solution development and evaluation, we developed a locally tailored GDM self-care model. This process not only addressed gaps in standard care but also empowered pregnant women through a supportive, multiple-stakeholder environment. This study demonstrates a rigorous cocreation pathway for systematically translating WHO

XSL•FO
RenderX

self-care constructs into a feasible, culturally adapted digital ecosystem for GDM in Vietnam and offers a transferable, people-centered design process alongside a practical blueprint for integration into routine maternal care.

Trial Registration: ClinicalTrials.gov NCT05744856; https://clinicaltrials.gov/study/NCT05744856

International Registered Report Identifier (IRRID): RR2-10.1186/s40814-025-01657-x

(J Med Internet Res 2026;28:e82434) doi:10.2196/82434

KEYWORDS

gestational diabetes mellitus; self-care; digital intervention; cocreation; peer support; intervention development; health promotion

Introduction

Gestational diabetes mellitus (GDM) is a metabolic health condition that affects approximately 14% of pregnancies worldwide [1]. Characterized by glucose intolerance that first appears during pregnancy, GDM, if inadequately managed, can lead to significant short- and long-term morbidity and mortality for both the mother and baby. Notably, women with a history of GDM face a 10-fold higher risk of developing type 2 diabetes (T2D) [2]. Meanwhile, their offspring not only have a higher risk of fetal overgrowth or macrosomia, one of the primary short-term complications of GDM [3], but are also more likely to have overweight or obesity [4] and have a higher risk of developing childhood- or youth-onset diabetes [5]. Lifestyle changes, including dietary modifications, a shift to an active lifestyle, and self-monitoring of blood glucose, along with medication as needed, are recommended to mitigate adverse health outcomes [6-8]. However, the complex nature of GDM and its management often pose significant challenges for women and their families [9-12]. Although education and counseling provided by health care professionals (HCPs) to women with GDM are the key components in both diabetes and pregnancy care, there are numerous challenges in the provision of standard care, especially in low- and middle-income countries (LMICs) [13].

The prevalence of GDM in Vietnam is relatively high. A study has shown that 21.8% of pregnant women in Vietnam have GDM, making Vietnam one of the countries in the Asia-Pacific region with the highest prevalence of GDM [14]. However, the health care system for pregnant women with GDM has not been sufficiently adapted to handle this situation. Due to insufficient information from HCPs and inadequate or lack of formal educational materials available [11,15], many pregnant women with GDM lack adequate knowledge about GDM and its complications and appropriate management. Additionally, cultural beliefs and traditional practices significantly affect women's diets and physical activity during pregnancy, posing challenges to their autonomy in making informed lifestyle modification decisions [10,11].

Self-care is a well-established concept [16]; however, it has become increasingly prominent in the current era, especially since the COVID-19 pandemic [17]. To enhance the quality and equity in self-care intervention implementation, the World Health Organization (WHO) has introduced a conceptual framework for self-care within people-centered primary health care, where individuals have autonomy and are empowered to make health decisions [18]. They also require an enabling environment that includes peers, families, and HCPs to support education and counseling, ensuring equitable access to resources [18]. Therefore, in the development of self-care interventions, cocreation involving the target population, stakeholders, and researchers is essential to meet the needs of the target users and to ensure cultural and contextual appropriateness, thereby improving the acceptability and sustainability of the intervention [18-20].

Furthermore, in recent years, digital health interventions, including mobile apps, telehealth services, wearable devices, and online platforms, have emerged as promising tools for improving GDM management through timely feedback, tailored personalized care, and peer support without geographical barriers [21-24]. The latest American Diabetes Association "Standards of Care in Diabetes—2025" guideline also acknowledged that combining telehealth and in-person visits could improve health outcomes for pregnant women with GDM, compared to in-person care alone [2]. However, research on cocreating and developing such digital health interventions remains limited in Vietnam.

This study aimed to design and pilot the "Healthy Pregnancy" multiplatform digital intervention for women with GDM in northern Vietnam, explicitly mapping WHO self-care constructs to cocreated functional components and documenting the staged development and refinement process.

Methods

Study Design and Theoretical Framework

We used a participatory cocreation design with a structure comprising 4 distinct but interconnected stages to create a new intervention initiative that meets the practical needs of pregnant women with GDM and to identify feasible design principles for a digital health intervention model suitable for the Vietnamese health care system (Figure 1).

A combination of 2 complementary theoretical frameworks was adopted—the people-centered care philosophy and cocreation design. The people-centered care approach emphasizes people-centeredness, which the WHO recognizes as a core component in the conceptual framework of self-care interventions, enabling a holistic approach that considers not only personal care but also individuals' contexts and desires [18]. Meanwhile, the cocreation design ensured that the target populations' roles shifted from passive recipients to active agents and coproducers of the intervention components for their health [25]. This dual-framework approach enabled us to deeply understand the sociocultural context of Vietnamese pregnant

XSL•FO
RenderX

women (eg, family hierarchy and dietary traditions) and their needs, which may profoundly influence their daily self-care practices. Furthermore, the presence of cocreation across all development processes helped avoid premature, one-way solutions from the researcher's perspective. This combination goes beyond simply gathering initiatives to foster shared priorities in self-care structures and in the desires of women with gestational diabetes, health care professionals, and research groups for digital features.

Figure 1. Four-stage development process of the "Healthy Pregnancy" intervention for gestational diabetes mellitus in Thái Bình (a part of H ng Yên Province), Vietnam (May 2023-July 2024). LMICs, low- and middle-income countries.

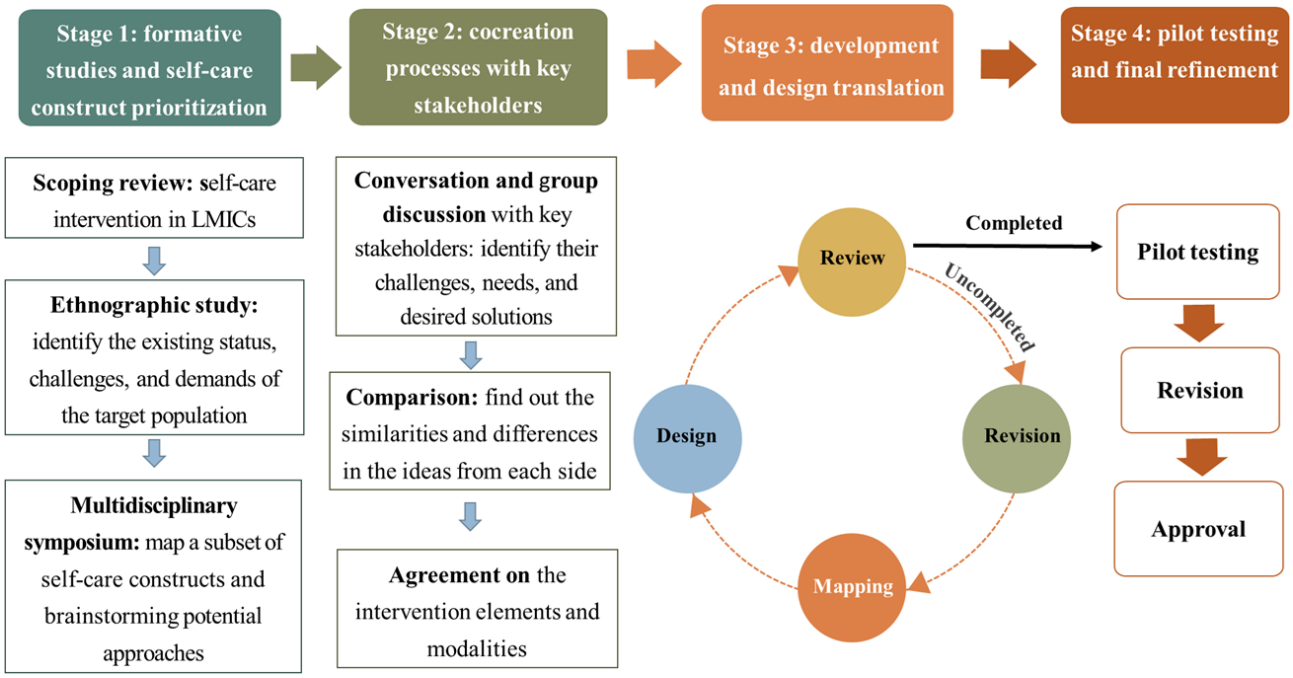

Additionally, this study was rooted in the WHO guidance on self-care interventions (2022), which expands self-care to a multidimensional set of daily practices encompassing 16 detailed, interrelated constructs divided into 3 main groups: self-management, self-testing, and self-awareness. To avoid premature selection from the researcher's perspective, we used this multidimensional set as a comprehensive reference to inform subsequent formative studies, including a scoping review and an ethnographic inquiry, and to identify diverse stakeholders who prioritize participation in the intervention development process. Ethnography was selected to capture the sociocultural determinants of GDM self-care, for example, patrilineal decision-making, shared meal practices, and domestic power hierarchies that shape women's daily behaviors beyond clinical settings. Data comprised repeated in-home interviews, contextual observations, photographs of artifacts, and field notes, enabling triangulation. Verbatim transcripts were coded using a thematic framework iteratively refined by the multidisciplinary team; discrepancies were resolved through consensus adjudication.

To ensure comprehensive and transparent reporting, this study followed the guidance for reporting intervention development studies in health research (GUIDED) checklist [26] and the template for intervention description and replication (TIDieR) checklist [27]. Full versions of the completed GUIDED and TIDieR checklists are available in Checklist 1 and Checklist 2, respectively.

Study Setting and Recruitment

Study Setting

Thái Bình Province is in the Red River Delta Region in northern Vietnam and currently—since July 01, 2025—belongs to a part of H ng Yên Province, according to Resolution No. 202/2025/QH15 [28]. To ensure consistency with our VALID II project [29], we have used "Thái Bình Province" in this paper. With a population of approximately 2 million, it is a rapidly developing industrial area characterized by significant socioeconomic growth. However, similar to other rapidly growing areas, Thái Bình Province is facing emerging public health challenges, particularly the rising prevalence of noncommunicable diseases such as hypertension, T2D, and GDM. In 2024, the prevalence of GDM in Thái Bình Province was reported to be 27.1% [30].

Standard Care for GDM

In 2022, when we designed and developed the intervention, GDM screening was not yet a mandatory step in the antenatal care process nor did the health insurance cover it [31]. In May 2024, the Ministry of Health in Vietnam launched a new guideline recommending GDM screening as a routine step in standard antenatal care [32]. Furthermore, although pregnant women can be diagnosed with GDM at either a public maternity hospital or a private clinic, they are generally advised to modify their diet by reducing white rice consumption and limiting sugar and sweets intake, as well as increasing exercise such as walking. Subsequently, they are referred to an appointment with an endocrinologist at Thái Bình Provincial General Hospital

XSL•FO
RenderX

for further professional counseling on diet, exercise, self-monitoring of blood glucose, and insulin treatment, if needed, as local obstetrician-gynecologists lack competence in this specialty. If women diagnosed with GDM do not attend this appointment, they do not receive GDM self-care guidelines from the HCP [10,29].

Recruitment

The VALID-II project is a 2-arm nonrandomized intervention study with 2 recruitment phases. In phase 1, pregnant women with GDM were recruited for the control group. Three months later, phase 2 recruited pregnant women with GDM, who participated in the cocreated digital health intervention. We conducted recruitment at 2 health care facilities, including Thái Bình Maternity Hospital and Kim Ngan Clinic. These are 2 prestigious obstetric facilities that attract many pregnant women in the province. As a public hospital, Thái Bình Maternity Hospital is tasked with providing antenatal care and delivery services, handling approximately 11,000 births each year; meanwhile, Kim Ngan Clinic, a private health clinic, offers only antenatal care services. Detailed information on the VALID II project has been reported in a protocol paper [29].

Women contributing to the ethnographic study (stage 1), the cocreation workshop (stage 2), and pilot testing (stage 4) were recruited separately from the main intervention cohorts, wherever feasible, to minimize expectancy and contamination, with 2 prior workshop participants purposefully retained in the pilot to preserve ownership and continuity. Specific participant recruitment criteria are detailed in respective subsections describing the intervention development process.

Development of a Content and Delivery Platform

Stage 1: Formative Studies and Self-Care Construct Prioritization

In the first stage, we aimed to be intentionally exploratory, informative, and nonprescriptive, which was a critical foundation for the subsequent cocreation process.

First, we aimed to gain insights into the current scientific evidence on self-care interventions for pregnant women with GDM, particularly in resource-limited areas such as LMICs. We also sought to explore culturally based self-care initiatives that address the demand for and barriers to self-care and lifestyle modifications for women and their families. Consequently, these insights informed us about the potential, up-to-date health technologies; advantages and disadvantages of the intervention models; and the existing gaps in the modalities and health aspects that each model addressed. To achieve this, we conducted a scoping review to summarize the content of self-care interventions for pregnant women with GDM in LMICs [33].

Subsequently, from a larger cohort of 233 pregnant women diagnosed with GDM in phase 1, who served as the control group in a 2-arm nonrandomized feasibility study, we randomly recruited 21 women [10,11,34-36]. These eligible women were followed up using a longitudinal ethnographic approach involving three iterative interview rounds at critical timepoints: (1) postdiagnosis, (2) postdelivery, and (3) 3 to 6 months postpartum. All participants were interviewed by the same researcher across 3 rounds, which fostered familiarity between the women and the researcher. By recruiting local interviewers who shared the same language and cultural background, we were able to accurately interpret local terminology and approach the interview with an empathetic lens, thereby avoiding personal biases and judgments and ensuring the trustworthiness of the data.

Most interviews and contextual observations were conducted within the participants' home environments, except for the postdelivery time point, for which the interviews and contextual observations took place at the maternity hospital. The interviewers began all conversations with a slow opening and attempted to follow the flow of the participant's narratives to create a comfortable and willing environment for sharing lived experiences through an open-ended interview guide (Multimedia Appendix 1). We used both voice recordings and detailed field notes with the participant's consent to capture not only verbal responses but also nonverbal cues, such as gaze, gestures, cheerful ease, and hesitation and avoidance during interactions. To enrich self-reported data with actual behavior, we documented physical artifacts (eg, maternal milk supplements, dietary supplements, snacks, and self-monitoring of blood glucose) via photography. The voice recordings were transcribed verbatim by a trained transcription team and verified by the interviewer. Subsequently, a 4-member research team collaboratively developed and standardized a thematic coding framework before analyzing the data. Additionally, to address disagreements during the coding process, group meetings were held frequently under the arbitration of a senior researcher.

Finally, we organized a symposium in our multidisciplinary research team (VALID II team) to discuss, synthesize, and validate findings from the scoping review and ethnographic study, which informed stakeholder dialogues during the next stage of cocreation. This symposium included Vietnamese and Danish experts and researchers experienced in conducting self-care interventions from 3 universities: Thai Binh University of Medicine and Pharmacy (TBUMP), University of Copenhagen (UCPH), and University of Southern Denmark (SDU). The symposium participants came from various fields, including obstetrics, endocrinology, general practitioners, epidemiology, public health, nutrition, and ethnography. In this symposium, given the broad range of self-care expressions and the GDM-irrelevance of some constructs mentioned in the WHO guideline, we mapped a subset of self-care constructs most closely aligned with our project's primary objectives: empowering pregnant women to take charge of their own health. Additionally, this subset required prioritizing relevance to daily GDM management, behavioral regulation, psychosocial well-being, and contextual relevance, as determined by the compiled findings. Furthermore, the discussions clarified key intervention methods, staffing needs, and educational materials and—drawing on VALID's first-phase experience—identified Zalo as a promising communication platform.

Stage 2: Cocreation Processes With Key Stakeholders

Stage 2 aimed to translate insights into cultural GDM management, a subset of prioritized self-care pillars, and

XSL•FO
RenderX

preliminary ideas from stage 1 into user-centered functional component requirements. We used a cocreation approach to foster a collaborative environment, sense-making, and negotiation, in which end users and key stakeholders acted as co-designers to weigh, prioritize, and suggest the most relevant constructs as materials for the intervention model. This stage emphasized identifying the most critical and appropriate self-care constructs, integrating them into intervention components, and considering contextual barriers before prototyping. Building on co-designs involving diverse stakeholders, these findings enabled us to balance the scientific aspects of GDM management with participants' lived experiences.

To adopt a multidisciplinary perspective, the May 2023 cocreation workshop invited diverse stakeholders, including 8 pregnant women with GDM and their families (typically their husbands), 2 obstetricians, 2 endocrinologists, and 3 primary HCPs, facilitated by a 7-member researcher team (co-leads and notetakers). During the workshop, the facilitation team sought to reduce power imbalances between HCPs and women by centering women's voices and strengthening their agency while still safeguarding clinical accuracy and alignment with health care system workflows.

The workshop followed a standardized script: a 15-minute opening brief; 120-minute individual, sequential conversations to surface needs, documented on flip-over charts; discussions in 4 small mixed groups to map constructs to solutions; and a 30-minute plenary session to compare the ideas and suggestions and conduct a hand vote on feasibility. Each subactivity captured and translated inputs into functional requirements. The co-leads facilitated the workshop, and the notetakers documented and photographed artifacts for analysis; the details of the workshop are summarized in Table 1 and Multimedia Appendix 2.

Table . Stakeholder-driven translation of WHO[a] self-care constructs into digital solutions: cocreation workshop, Thái Bình, Vietnam (May 2023).

Unmet needs identified during cocreation	WHO self-care construct	Cocreated solutions	Potential platform(s) and materials
Fragmented, inconsistent, and informal GDM[b] information	Self-education	Short, visual, practice-oriented educational materials; online counseling	Zoom (live sessions); leaflets, videos
Low confidence in lifestyle modifications; worry about the complications; feeling of loneliness in GDM management	Self-efficacy	Peer modeling through experience sharing; reassurance messages; expert-led interaction education sessions	Private Facebook page (peer posts); Zalo (peer support, messages); Zoom (live sessions)
Numerous scattered women with GDM and poor interspecialty coordination	Self-monitoring	Establishment of interactive groups by gestational week to provide timely feedback	Zalo (peer support, messages); Zoom (live sessions)
Women struggled to adjust their behaviors in real time, especially in shared family meals	Self-regulation; self-administration	Short, visual, practice-oriented educational materials for women and family members; timely nudges and reminders aligned with daily routines (eg, before meals)	Leaflets, videos, Zalo (scheduled messages, peer-to-peer image sharing)
Women lacked a safe space to share emotions, fears, and informal coping strategies without judgment.	Self-help	Moderated peer support spaces emphasizing emotional validation and shared coping	Private Facebook page (anonymous posting); Zalo groups
Need to improve autonomy in making food choices in social situations (eg, eating out and family events) or for severe cases	Self-determination	Short, visual, practice-oriented educational materials; integrating CGM[c] technology enhances decision-making by enabling analysis of individual data.	Leaflets, videos, Zalo (guided messages, peer-to-peer image sharing); CGM devices

[a]WHO: World Health Organization.

[b]GDM: gestational diabetes mellitus.

[c]CGM: continuous glucose monitoring.

Stage 3: Development and Design Translation

On the basis of the inputs gathered from stage 1 and cocreation workshops, we aimed to translate all user requirements into a cohesive digital health–based peer support system titled "Healthy Pregnancy," which supplements standard care for women with GDM. To ensure high accessibility and scalability in Vietnam's cultural context and amid the digital divide, we leveraged an existing multiplatform ecosystem, including Zalo (VNG Corporation), Zoom (Zoom Communications Inc), and Facebook (Meta Platforms Inc), rather than investing in developing and prototyping a new mobile app.

The intervention suite was structured to address the 7 WHO self-care pillars concurrently. Zalo, a widely used mobile app in Vietnam, was configured as the primary tool for covering self-awareness, self-monitoring, and self-testing through a set of educational materials, peer discussions, and HCP guides. Zoom was used to host biweekly synchronous sessions for self-education, while a private (restricted-access) Facebook group provided a platform for self-awareness.

XSL•FO
RenderX

To address data privacy concerns within third-party environments, we implemented a dual-layer security strategy. First, we instructed and recommended participants to use pseudonyms and avatars instead of real photographs for all direct communication on Zalo and Zoom. Second, the private Facebook group used the "Post anonymously" feature to enable sensitive discussions without revealing identities. Moderators enforced role-based permissions, screened posts, and triaged complex queries to HCPs. Extracted data were anonymized and stored on password-protected TBUMP servers.

Regarding educational materials, we planned to provide all pregnant women with GDM in the intervention with leaflets and videos focused on topics such as diet, exercise, glucose monitoring, emotional well-being, and family support. To prepare a set of educational materials, we mapped and drafted the materials, which then underwent review rounds with experts at TBUMP, UCPH, and SDU, including obstetricians, endocrinologists, nutritionists, and health communication experts, who provided feedback and revisions on both language and visuals.

Stage 4: Pilot Testing and Final Refinement

In stage 4, a purposive sample of 8 women—4 pregnant women with GDM and 4 women with a history of GDM—was recruited to test the "live" ecosystem. Two participants were reinvited from the stage 2 cocreation workshop. The 6 additional participants were recruited via direct invitations at the Kim Ngan Clinic and the TBUMP endocrinology clinic, using the same demographic inclusion criteria as those used for participation in stage 2. This sampling strategy ensured that the objective would be achieved and minimized bias while preserving continuity and ownership established in the cocreation workshop.

A 2-week pilot study was conducted using a mixed methods approach to evaluate the feasibility, acceptability, and potential barriers of the multiplatform intervention before real-world implementation, without assessing clinical effectiveness.

For the quantitative assessment of educational materials, participants evaluated the readability and visual presentation of the leaflets and videos using an anonymous evaluation form (Multimedia Appendix 3). The first 4 questions referred to their impressions and assessments of the GDM leaflets or videos, covering clarity of language, font size, visual appeal, and usefulness of the content, rated on a 10-point Likert scale (1-10). Descriptive statistics were used to analyze the results. The last question was open-ended, encouraging the participants to provide additional comments or suggestions to improve the quality of the leaflets and videos.

For qualitative observation of intervention workflow, participants were invited to join the pilot Zalo group (Figure 2), Zoom meetings, and the Facebook page for a week. The research team provided technical support to ensure continuous connectivity and identify platform-specific issues. During the pilot week, we reported patterns of interaction across platforms. Observations and implementation reflections were documented, focusing on user engagement dynamics, user preferences for interaction styles, and operational facilitators and barriers encountered in real-world use, rather than on statistics using metrics. Following the pilot phase, the 8 women met again to discuss their experiences, comments, and suggestions. Data were collected with an open-ended question guide, the moderator's notes, and video recordings, followed by an analysis process that included structured scoring evaluation and a thematic approach. Specifically, 2 researchers independently reviewed and graded the engagement atmosphere and the quality of interactions (eg, reaction frequency and sentiment emojis). Additionally, an inductive approach was used to capture all topics and interaction preferences, while deductive codes addressed user engagement dynamics and technical barriers. Subsequently, 2 researchers cross-referenced their individual scores and notes. Any discrepancies were resolved through team consensus meetings. These insights provided valuable inputs for the final refinement before starting the actual intervention. While the pilot phase prioritized qualitative feasibility insights, the upcoming intervention phase will prospectively track platform metrics (eg, activation rates, session attendance, content views, message volumes, moderator escalations, and resolution times) to quantify acceptability and usage.

By the end of this stage, we had finalized the intervention description and submitted the revised educational materials to the Department of Information and Communications of Thái Bình Province.

XSL•FO
RenderX

Figure 2. Samples of messages in the Zalo pilot testing group, Thái Bình, Vietnam, August 2023.

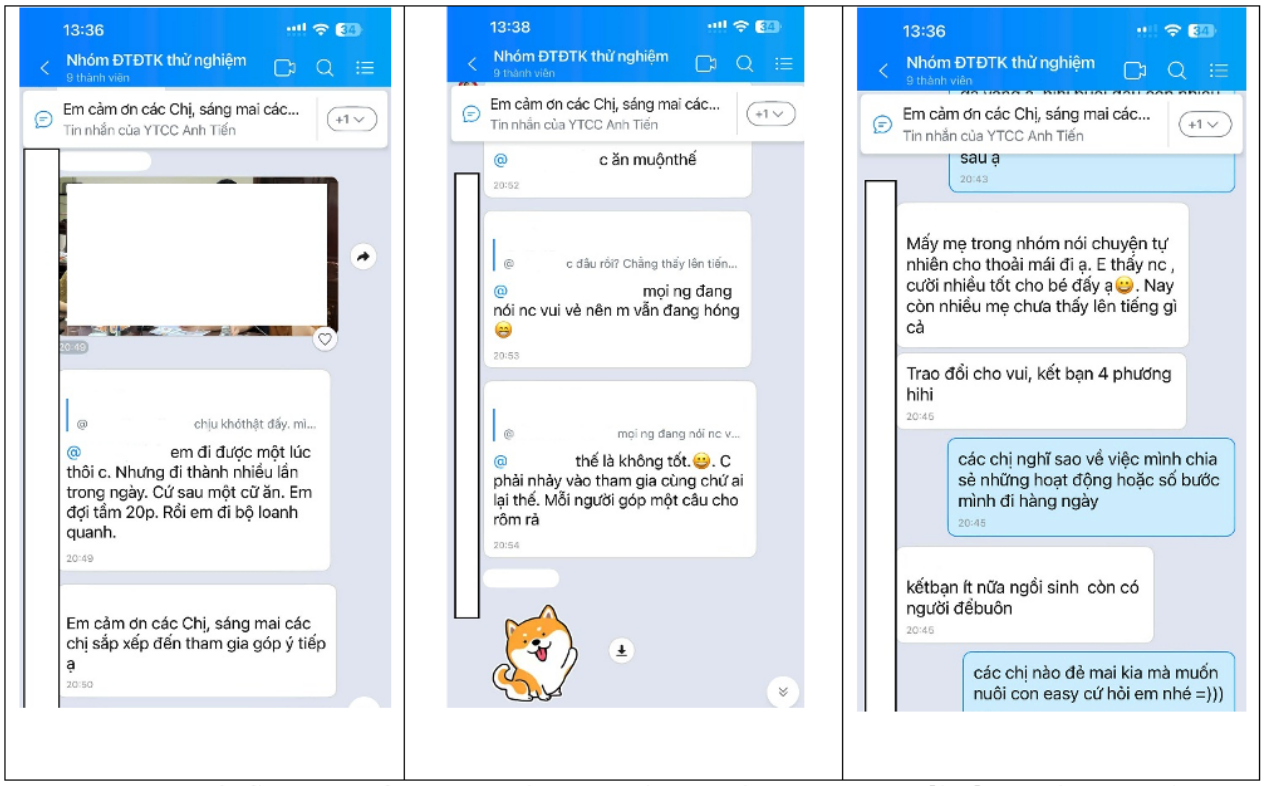

Ethical Considerations

Ethical approval for this study was granted by the Ethics Council for Biomedical Research at Thai Binh University of Medicine and Pharmacy, Vietnam, on December 1, 2022 (Institutional Review Board VN01.009). The study was registered at ClinicalTrials.gov (NCT05744856). All participants agreed and signed the written informed consent after being fully informed about the study's objectives and procedures, including their right to withdraw at any time. All information was stored with strict confidentiality to protect participants' identities. For the ethnographic study, the participants' names were changed to ensure anonymity. Participants received a compensation of VND 500,000 (US $18.98) for their time and transportation costs in stages 2 and 4 and VND 200,000 (US $7.59) for each interview round of the ethnographic study.

Results

Stage 1: Formative Studies and Self-Care Construct Prioritization

Stage 1 provided formative insights that informed the empirical and theoretical foundation for the subsequent cocreation and intervention development stages. On the basis of the scoping review, ethnographic study, and multistakeholder symposium, we identified gaps in current digital health interventions in LMICs, explored the sociocultural realities of women with GDM in Vietnam, and prioritized 7 self-care constructs.

A total of 29 self-care intervention programs for women with GDM in LMICs, across the physical, mental, and social health dimensions, were included. Most interventions focused predominantly on improving clinical outcomes while less frequently addressing the mental and social health dimensions. Moreover, digital interventions have emerged as a promising approach to improve pregnant women's adherence and pregnancy outcomes [33]. Additionally, interventions were often implemented and followed up only during pregnancy, overlooking the sensitive postpartum transition period.

However, although the scoping review identified a broad range of self-care constructs implemented in LMICs, no intervention studies from Vietnam were identified [32]. This underscores the need for ethnographic studies to explore the perceptions, experiences, and challenges of Vietnamese pregnant women regarding GDM self-care within the local context. Because kinship is patrilineal and patrilocal in Vietnam, self-care practices among pregnant women with GDM are deeply embedded in family dynamics and cultural dietary habits, especially influenced by the husband and mother-in-law. Decisions about dietary changes and lifestyle modifications are often profoundly influenced by the husband and, sometimes, even by the mother-in-law. Going beyond qualitative descriptive interviews that might capture only what participants report thinking and doing, the ethnographic study allowed us to explore the contextual factors, including social relations, domestic power hierarchies, and cultural norms, that shape women's perspectives, feelings, and behaviors in daily life outside the clinical setting. Additionally, the visual records allowed us to partially assess family circumstances and identify gaps between clinical advice and real-world consumption patterns, providing insights that are often lacking in common qualitative descriptive interviews.

The ethnographic studies highlighted 2 critical barriers in the Vietnamese context. First, pregnant women were forced to become unexpected pioneers in GDM management. The lack

XSL•FO
RenderX

of professional counseling in the health care system led them to seek information online. However, they expressed uncertainty about the information being heterogeneous, informal, and not culturally tailored. Second, the "shared meal" culture could profoundly influence pregnant women's self-regulation practices because decisions of the family-in-law dictate meal preparation. Conversely, if husbands understood the psychological and practical challenges of GDM management, they could, in turn, sympathize and offer emotional support to reduce women's overwhelming feelings. The husband's presence could shift GDM management from a sole burden on women to a family-wide health concern, enabling a home environment that sustains self-care behaviors and improves maternal and fetal outcomes. Furthermore, through life narratives, we found that informal peer-to-peer exchanges were already occurring and were potentially valuable in some cases. In addition, the ethnographic inquiry sought to identify pregnant women's information regarding GDM, their preferences for specialist advice, and potential effective self-care constructs for pregnant women with GDM that they discovered through personal experience. These insights justified the need for the people-centered care framework to address the "whole person" within a woman's family and social context, beyond clinical indices.

Based on the formative studies, we synthesized and validated the inputs for presentation at the symposium. To translate these broad insights into a structured intervention, the WHO's 3 main pillars and 16 self-care constructs were systematically assessed at the expert symposium, with an emphasis on clinical management, well-being, and contextual relevance for the Vietnamese population with GDM. We identified and mapped several problem-need clusters relevant to women with GDM onto multiple WHO self-care constructs. This reflected the interconnected, synergistically reinforcing nature of self-care constructs in GDM management rather than their functioning independently. Additionally, empowering women in home-based GDM self-management was the primary purpose rather than clinical diagnostics or treatment delivery, leading to the exclusion of constructs reliant on these. Therefore, we prioritized 7 WHO self-care constructs as most actionable for home-based GDM management, including self-education, self-efficacy, self-monitoring, self-regulation, self-help, self-determination, and (context-specific) self-administration, on the basis of clinical relevance, feasibility, and stakeholder consensus (Table 2).

XSL•FO
RenderX

Table . Mapping scoping and ethnographic findings to WHO[a] self-care constructs for GDM[b]: cocreation study, Thái Bình, Vietnam (2023).

Relevance to women with GDM (key findings) and WHO self-care construct (2022)	Resources	Prioritization decision	Reason for prioritization (link to GDM management and well-being)
Key gaps and unmet needs			
Need for holistic care: integrating mental and social health into GDM management			
• Self-efficacy • Self-education	Scoping review [33]	High	Self-efficacy in self-care capacity may influence mental health and maintain long-term behavioral changes. Additionally, GDM awareness reduces anxiety and improves glycemic control.
A longer follow-up and support			
• Self-education • Self-regulation • Self-monitoring • Self-help	Scoping review [33]	High	Longitudinal support fosters sustainable, peer-based resilience; reinforces health literacy; and ensures consistent self-monitoring postpartum, facilitating proactive behavioral regulation.
Lack of GDM awareness and confusion among the abundance of information on the internet			
• Self-education	Ethnographic studies [10,11,36]	High	Requires official, standardized medical information that supports pregnant women with an accurate understanding
Barriers to shared meal culture and lack of family support			
• Self-help • Self-administration	Ethnographic studies [10,11,36]	High	Engaging family support to overcome “shared meal” challenges and implement daily modifications
Desire for an effective support environment			
On-time support; elimination of geographical or financial barriers			
• Self-education • Self-regulation • Self-monitoring	• Scoping review [33] • Ethnographic studies [35]	High	GDM requires daily modifications. Immediate feedback or guidance empowers women to promptly understand and self-regulate their lifestyle and maintain self-monitoring of blood glucose.
Safe and no-blame space for exchanging experiences and emotions			
• Self-help • Self-determination • Self-efficacy	Symposium	High	A safe, accompanied peer support environment fosters self-determination and self-help and boosts mental health, self-confidence, and social connection.
Requirement of structured and visual educational materials (videos/leaflets)			
• Self-education • Self-monitoring	• Ethnographic studies [35] • Symposium	High	Providing accurate medical information via visual aids helps women acquire new knowledge more easily and is suitable for all educational levels.

XSL•FO
RenderX

Relevance to women with GDM (key findings) and WHO self-care construct (2022)	Resources	Prioritization decision	Reason for prioritization (link to GDM management and well-being)
Other clinical suggestions			
Clinical diagnostics: sampling, screening, and collection of medical samples			
• Self-sampling • Self-screening • Self-diagnosis • Self-collection	Symposium	• Excluded • Excluded • Excluded • Excluded	These procedures are typically performed in hospitals under medical supervision.
Medical treatment			
• Self-injection	Symposium	Low	In Vietnam, pregnant women with GDM use only insulin.
• Self-medication • Self-treatment • Self-examination • Self-use	Symposium	• Excluded • Excluded • Excluded • Excluded	Interventions focused on empowering women to build WHO self-awareness constructs, thereby promoting lifestyle modification at home rather than complex medical procedures.

[a]WHO: World Health Organization.

[b]GDM: gestational diabetes mellitus.

Stage 2: Cocreation Processes With Key Stakeholders

Stage 2 aimed to translate stage 1 findings into user-centered functional requirements and components, thereby establishing the intervention's foundational infrastructure, through 3 main activities with valuable insights from diverse key stakeholders during the cocreation workshop.

Gaps in Professional Interaction and GDM Care Within Local and National Health Care Systems

In the initial activities, a realistic picture of fragmentation in GDM care within the health care system was reported, reconfirming the stage 1 findings. Three core issues were (1) a lack of standardized official information and counseling channels for women with GDM and their families; (2) limited endocrinologist-obstetrician interaction during follow-ups and childbirth; and (3) high expectations for physician comprehensiveness (Multimedia Appendix 2).

An obstetrician admitted the challenges in counseling:

> *The obstetricians had limited specialized knowledge of endocrinology and GDM…Pregnant women expect doctors to have comprehensive training in all diseases, which makes it challenging to provide satisfactory advice.*

Endocrinologists reported challenges from inadequate equipment, limited awareness of GDM among pregnant women and the community, lack of official local information channels, and limited interaction with obstetricians during follow-ups for pregnancy and childbirth.

This status led to frustration with unmet needs in professional care and confusion about the overwhelming heterogeneity of online information. An expectant mother, aged 40 years, participating in the workshop shared the following:

> *I wanted to know what kind of diet is for people with GDM; I did my own research on the Internet and found that it is very dangerous for both mother and child… I adjusted my diet, but sometimes I had low blood sugar. My diet was not scientific, but I did not find a reasonable solution.*

Prioritize and Translate a Subset of WHO's Self-Care Constructs Into Potential Solutions Grounded in Realistic Stakeholder Insights

In the group discussion activity, participants mapped unmet needs and gaps to 7 prioritized WHO self-care constructs, translating them into concrete functional requirements and selecting feasible delivery platforms (Table 1).

For example, the lack of formal information channels or confusion from unverified internet sources, which fit the self-education construct, led to a demand for an expert-led system of video tutorials and leaflets. To bridge the gap in endocrine expertise among obstetricians and the limited interaction between obstetricians and endocrinologists, online counseling groups facilitated women's self-monitoring and self-regulation, particularly those in scattered populations. Similarly, other critical challenges for women included a lack of confidence and feelings of loneliness during GDM management, which requires self-efficacy and self-help. In response, the stakeholders emphasized the importance of peer modeling, leading to the design of peer support groups for sharing rather than solely expert-led education.

Additionally, rather than orienting the participants to predefined technological solutions, we encourage them to self-propose ones based on accessibility, familiarity, and privacy. The integrated, multiplatform approach, which included existing platforms such as the Zalo mobile app, Zoom video

XSL•FO
RenderX

conferencing, and Facebook, was identified as a suitable infrastructure that addressed geographical barriers and empowered pregnant women within the context of family culture in Vietnam. Every requirement regarding the organization and operational format of the platforms was documented on flip-over charts by each group and incorporated into the subsequent discussion and deliberation process (Multimedia Appendix 2).

After the groups presented their ideas, we shared, compared, and reached consensus on potential solutions for implementation in the trial and for future scale-up. We reached consensus on establishing small Zalo groups, Zoom online counseling, family groups, private Facebook groups, and visual educational materials. We also agreed on the need to improve the capacity of HCPs in standard care for GDM through short training courses. However, the idea of continuous glucose monitoring was eliminated due to its high cost, which was unsuitable for all women with GDM (see Table S1 in Multimedia Appendix 2).

Consensus outputs (Table 1) were converted into technical specifications and scripts that informed the platform configurations summarized in Table 3, including triage workflows and standardized nudge libraries.

Table . Technical configuration and behavioral mechanism of the "Healthy Pregnancy" ecosystem for GDM:[a] Thái Bình, Vietnam (2023 - 2024).

Platform component	Technical configuration and workflow	Behavioral mechanism	Final design artifacts
Zalo: interaction tool	Gestational week–based peer support groups implementing a triage workflow: initial peer support addresses simple inquiries; if complex issues: moderator receiving information → escalating to HCPs[b] in group and specialists in Zoom meetings	Immediate feedback and cues to action enable rapid behavioral adjustment and minimize confusion and anxiety.	Welcoming, introduction, and nudge message scripts; moderator dispatch procedures
Zoom meetings: online counseling	Scheduled online counseling sessions with specialists in obstetrics, endocrinology, and nutrition	Enhancing GDM awareness and reinforcing self-efficacy through guided problem-solving	Preschedule online specialist consultations and prepare corresponding discussion topics for each.
Private Facebook page: resource center	Structured content library aligned with pregnancy progression; integrated anonymous posting feature	Enabling women to self-educate, receive peers' empathy, and share without judgment	Official, standardized educational materials (leaflets and expert-led videos)

[a]GDM: gestational diabetes mellitus.

[b]HCP: health care professional.

Stage 3: Development and Design Translation

At this stage, the ideas generated during the cocreation workshop were translated into a multicomponent digital intervention ecosystem with explicitly defined operating workflows. Rather than being designed and implemented as a standalone, completely new mobile app, the intervention integrated educational, behavioral, and psychosocial support via familiar digital platforms to empower women's daily self-care practices. These mechanisms were designed to operate concurrently and complementarily throughout the intervention period, addressing interrelated aspects of self-care capacity (Table 3).

To ensure an effective and engaging intervention, we developed concise, culturally relevant, and accessible educational materials across 3 aspects of GDM self-care. The set of educational materials included 8 leaflets (1 print and 7 digital) and 8 videos (see Multimedia Appendix 4). The visual elements used simple language, illustrations, and real-life scenarios to enhance engagement and facilitate the daily application of self-care principles by pregnant women and their families. Diverse experts, including obstetricians, endocrinologists, public health professionals, and health communication specialists, reviewed and provided feedback on all educational materials. This process ensured the refined files' medical accuracy, feasibility, and effective message delivery through the use of simplified language and enhanced illustrations and improved cultural relevance.

To fulfill its commitment to supportive spaces and protect participants' rights, the "Healthy Pregnancy" intervention established security measures across all platforms. In Zalo groups, we implemented hierarchical data management, with moderators responsible for screening members, securing discussion content, and limiting information sharing to the framework of professional treatment consultations with HCPs. Additionally, women were instructed to use pseudonyms and avatars instead of real photographs for all communication via Zalo and Zoom meetings. The Facebook page was set to private mode, allowing access to only approved, project-verified members, to prevent leaks. Participants on the Facebook page also used the "Anonymous Posting" feature to share about psychological pressures from family or experiences without fear of judgment or identification.

Stage 4: Pilot Testing and Final Refinement

Pilot Testing With Educational Materials

Table 4 presents the pilot participants' responses to the educational materials in the intervention. Most of the pilot participants (87.5%, 7/8) rated the language used in the leaflets and videos as 7 to 9, indicating understandable language, while 1 participant (12.5%) rated it 6. The mean rating was 7.6 (SD 0.9). Regarding font size readability, most (75%, 6/8) participants gave a positive rating of 7 or 8 for ease of reading, while 25% (2/8) of the participants rated it 6. The visual elements were rated with an average score of 7; most found

XSL•FO
RenderX

them attractive and engaging (7 or 8 points), while 1 participant rated them as 6, and another rated them as 5. Regarding the usefulness of the content, all participants assigned a high score between 7 and 10, with an average score of 8.6 (SD 0.9).

Table . Language and cultural evaluation of materials during pilot testing of a GDM[a] intervention with 8 participants: Thái Bình, Vietnam (August 2023).

Criterion	Score[b], mean (SD)	Score range
Language	7.6 (0.9)	6 - 9
Font size	7.3 (0.9)	6 - 8
Visualization	7.0 (1.1)	5 - 8
Content	8.6 (0.9)	7 - 10

[a]GDM: gestational diabetes mellitus.

[b]All items were rated on a Likert scale from 1 to 10, where 1 indicates "very poor" and 10 indicates "excellent." Higher scores represent greater approval and cultural appropriateness of the materials.

Responses to the open-ended question showed that most participants appreciated our educational materials. They suggested revising the wording to increase accessibility to pregnant women, explaining medical terms, increasing font size for easier reading, and using images of pregnant women and foods that reflect Vietnamese culture.

Pilot Testing With the Intervention Process

Observations and discussions during pilot testing indicated high engagement in Zalo groups through messages, images, and emojis. Apart from exchanging GDM self-management experiences, the women seemed to have conversations that extended to broader aspects of daily life and family contexts. Additionally, Zoom-based online consultations with HCPs were well received for addressing unanswered questions during brief clinical visits.

On the other hand, some observed operational challenges included variations in interaction dynamics depending on whether active participants were present in the Zalo group. In Zoom sessions, engagement and interest decreased when didactic clinical content dominated the sessions, compared with sessions structured as an interactive question-and-answer section. Additionally, unstable internet connectivity was reported as a technical barrier that disrupted live consultations. Anonymous posting on Facebook was not preferred, as most informational needs were addressed through Zalo and Zoom.

Consequently, we refined and standardized the implementation process. Zalo was designated as the primary platform for daily interaction, and active women were encouraged to initiate discussions. Zoom sessions were restructured into short lectures and a case-based, question-driven format with a flexible duration. Additionally, we designated Facebook as a repository for educational materials and information, with pinned posts to ensure stable, easy-to-locate access. To enhance system credibility, a formal project logo was designed on the basis of the majority preference (6 of 8 participants), and HCPs were required to use standardized identification, including full names and professional titles, across all digital groups. Finally, we implemented an operational adjustment: integrating QR codes into the implementation workflow to facilitate women's access to information.

After refining the intervention through multiple rounds of feedback and revisions, we submitted the educational materials and received approval from the Department of Information and Communications of Thái Bình Province.

Finally, we also completed the (1) client flow from diagnosis to postpartum, (2) moderation and triage logic, and (3) examples of scripts, applying to the intervention for pregnant women with GDM (see Multimedia Appendix 5).

Discussion

Principal Findings

This study provides empirical evidence for the systematic development and pilot refinement of the "Healthy Pregnancy" digital peer support ecosystem, demonstrating how it operationalizes global self-care frameworks within local clinical practice in Vietnam. Through a rigorous, 14-month-long, staged process, we operationalized 7 prioritized WHO self-care constructs into an integrated multiplatform architecture, leveraging the Zalo mobile app for real-time interaction and peer support, Zoom meetings for expert-led counseling, and Facebook as a static resource hub. Unlike previous formative research from our group, this study uniquely explicates the design translation process, moving beyond "identifying needs" to document how real-world demands and WHO self-care constructs were integrated and translated into concrete digital features and behavioral mechanisms.

Furthermore, a functional peer support workflow was designed to address the fragmented coordination between obstetricians and endocrinologists. The intervention's educational materials package, including 8 sets of practical videos and educational leaflets, was cocreated with stakeholders acting as co-designers rather than passive informants. Additionally, pilot testing confirmed the feasibility of this multiplatform approach, which revealed high engagement in Zalo groups and the need for interactive, case-based Zoom sessions. Moreover, final refinements, such as an official project logo and standardized HCP identification, were led by end users, serving as trust-building mechanisms.

XSL•FO
RenderX

Interpretations in Context

To the best of our knowledge, this study is the first to describe the multistage development of a digital health–based peer support intervention for pregnant women with GDM in Vietnam. Uniquely, we made the often-implicit design logic explicit, mapping global self-care constructs onto concrete digital components and user journeys suitable for public maternal care in a lower-middle-income context. One of the key strengths of our study was its comprehensive, participatory, and contextualized approach, which incorporated the WHO's conceptual framework for self-care and the cocreation method across multiple stages. Drawing on the WHO's conceptual framework for self-care [18], we established a digital health ecosystem focused on people-centered primary health care, integrating structured educational materials, expert-led counseling, and peer support to empower pregnant women to make informed decisions and adhere to daily self-care behaviors.

Meanwhile, engaging the target users and multiple stakeholders, the cocreation approach enabled the intervention content and modalities to be adjusted relevant to the target group's context and preferences [19,37,38]. The cocreation process in VALID-II was well demonstrated in co-leadership in prioritization and shared ownership of design choices. Participants acted as co-designers by proposing, debating, and voting on delivery modalities (eg, Zalo group size and Zoom timing) and removing infeasible ideas (eg, continuous glucose monitoring), with researchers implementing these joint decisions. To enhance system credibility, a formal project logo and standardized HCP identification were implemented across all digital groups, the basis of the majority preference. These refinements, driven by stakeholder leadership, confirm a transition from conventional qualitative inquiry to true shared ownership of the endeavor.

This systematic yet flexible procedure enabled us to apply up-to-date scientific insights effectively and to modify the intervention program locally to the context and needs of the target audience. It also aligns with the design of complex interventions, particularly in resource-limited settings, where cultural appropriateness and stakeholder engagement are key elements for effectiveness and sustainability.

Another strength of our study was the ability to leverage widely available, low-cost digital platforms, including Zalo [39], Zoom [40,41], and Facebook. Our cocreation pathway goes beyond education-only designs by linking prioritized self-care constructs (self-education, self-efficacy, self-monitoring, self-regulation, self-help, self-determination) to specific platform features and behavioral mechanisms (eg, cues to action via Zalo nudges; problem-solving efficacy via Zoom question-and-answer sessions; anonymous peer validation via Facebook). Beyond the familiarity and feasibility of Vietnam's technological infrastructure, this ecosystem helped blur geographical barriers, facilitate timely support and counseling, and reduce pressure on the professional health care system. Moreover, the accompaniment of peer support and family within this ecosystem provides social validation and habit scaffolding, which were essential for sustaining lifestyle changes within the Vietnamese "shared meal" culture, where family dynamics often dictate dietary and exercise habits [42-45].

Finally, design choices such as leveraging widely adopted platforms, standardizing moderation standard operating procedures, and supporting low-bandwidth content help scale up and maintain fidelity within routine antenatal care, creating favorable conditions for sustaining and scaling the model in LMIC settings such as Vietnam. The intervention was intentionally designed around task shifting, with moderation and first-line triage delivered by trained midwives or diabetes nurses and escalation to obstetricians or endocrinologists only when clinically indicated, thereby minimizing additional workload while preserving patient safety. In parallel, women with GDM were engaged as peer knowledge contributors within clearly defined nonclinical boundaries, reducing repetitive informational demands on health care staff while strengthening comfort, connectedness, and mutual support. From a governance perspective, early incorporation of standardized HCP identification and structured moderation procedures embeds accountability mechanisms from the outset, providing a foundation for traceable data flows, referral documentation, and future interoperability with the Ministry of Health systems. The development of structured training pathways for moderators and specialist responders will be critical to ensure long-term sustainability.

Key Implications

The key implications of this study are as follows:

1. A transparent mapping from global self-care constructs to local digital features is feasible.
2. Cocreated moderation and triage processes are essential to sustain engagement.
3. Ubiquitous platforms can deliver timely, low-cost support at scale.
4. Formal data governance must be designed in from the outset.

Limitations

However, our study had several limitations. First, the digital health intervention was designed specifically for pregnant women with GDM in Thái Bình Province, northern Vietnam, leading to its highly context-specific nature. Although its design and development process could be fully transferable, replicating the model without adapting to cultural and social norms, digital access, online platforms, or the health care system may involve difficulties.

Second, although 79.1% of the population in Vietnam has internet access [46], and we tried to use freely accessible digital platforms locally, it may still not fully cover individuals without stable internet access. Furthermore, digital literacy was a significant barrier to the adoption of digital health, leading to the requirement of accompanying technical support and training [47]. Therefore, they may inadvertently limit the access of socioeconomically disadvantaged and low-digital-literacy groups to the intervention.

Third, while we expected that it would be feasible to integrate the intervention into standard maternal care, it still required an available staff team for management and moderation, as well as a specialist team for education and timely counseling. Another potential limitation of online peer group interventions is

XSL•FO
RenderX

participants' reluctance to share their thoughts and emotions with unfamiliar peers, which may hinder group cohesion and reduce the intervention's overall effectiveness. However, these participants may still benefit from the information or experiences shared by other members and professional advice from health care workers in the group. Additionally, the design and development process took a long time, 14 months. Therefore, it may not achieve ideal cost-effectiveness in this period. Nevertheless, this process serves as an essential stepping stone, as we can leverage its advantages and results in the next phase of the VALID study.

Conclusion

This study demonstrates a rigorous cocreation pathway that systematically translates the WHO self-care constructs into a feasible, culturally grounded, multiplatform digital ecosystem for GDM in Vietnam. By making the design logic, behavioral mechanisms, and implementation workflow explicit, we contribute a transferable method for building people-centered digital self-care interventions in resource-constrained health systems and a concrete blueprint for their integration into routine maternal care.

Acknowledgments

We are grateful to all pregnant women with gestational diabetes mellitus, their families, and health care professionals who participated and provided valuable insights in the cocreation workshop and the pilot testing. This study would not have been possible without the support of the Thái Bình Maternity Hospital and Kim Ngan Clinic, which helped us recruit the target group. Moreover, we sincerely thank experts at the Thai Binh University of Medicine and Pharmacy, University of Southern Denmark, and University of Copenhagen for their valuable comments and suggestions for developing educational materials. The authors declare the use of Grammarly, a generative artificial intelligence software, solely for proofreading and editing purposes, specifically to correct grammatical errors and spelling and to improve academic vocabulary. No original text or research findings were generated by artificial intelligence.

Funding

The study is part of the Living Together with Chronic Disease: Informal Support for Diabetes Management in Vietnam, Phase II: Gestational Diabetes in Vietnam (VALID II) (grant number: 21-M03-KU), funded by the Ministry of Foreign Affairs of Denmark. The funder was not involved in the study and had no role in developing the study design, conducting the study, analyzing the data, reporting the results, or the decision to submit the manuscript for publication.

Data Availability

The datasets generated or analyzed during this study are not publicly available to protect the participants' privacy but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: NATD, JS, TMG
Data curation: NATD, DKN, TAN, HML, TKDV, TMG
Funding acquisition: TMG
Investigation: NATD, DKN, TAN, HML, TKDV, TMG
Methodology: NATD, JS
Project administration: TMG, TDN
Supervision: TMG, JS, DWM
Visualization: NATD
Writing – original draft: NATD, JS
Writing – review & editing: NATD, DWM, DSL, CAV, TDN, DKN, TAN, HML, TKDV, THT, TMG, JS

Conflicts of Interest

None declared.

Multimedia Appendix 1
Open-ended interview guide.
[DOCX File, 24 KB - jmir_v28i1e82434_app1.docx]

Multimedia Appendix 2
Details of the cocreation workshop with key stakeholders.
[DOCX File, 1663 KB - jmir_v28i1e82434_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Evaluation form for educational materials.
[DOCX File, 31 KB - jmir_v28i1e82434_app3.docx]

Multimedia Appendix 4
Overview of educational materials.
[DOCX File, 17 KB - jmir_v28i1e82434_app4.docx]

Multimedia Appendix 5
System schematic and information flow at Thai Binh University of Medicine and Pharmacy (TBUMP).
[DOCX File, 1102 KB - jmir_v28i1e82434_app5.docx]

Checklist 1
The guidance for reporting intervention development studies in health research (GUIDED) checklist.
[DOCX File, 59 KB - jmir_v28i1e82434_app6.docx]

Checklist 2
The template for intervention description and replication (TIDieR) checklist.
[DOCX File, 33 KB - jmir_v28i1e82434_app7.docx]

References

1. Wang H, Li N, Chivese T, et al. IDF diabetes atlas: estimation of global and regional gestational diabetes mellitus prevalence for 2021 by international association of diabetes in pregnancy study group's criteria. Diabetes Res Clin Pract 2022 Jan;183:109050. [doi: 10.1016/j.diabres.2021.109050] [Medline: 34883186]
2. American Diabetes Association Professional Practice Committee. 15. Management of diabetes in pregnancy: standards of care in diabetes-2025. Diabetes Care 2025 Jan 1;48(1 Suppl 1):S306-S320. [doi: 10.2337/dc25-S015] [Medline: 39651985]
3. Kc K, Shakya S, Zhang H. Gestational diabetes mellitus and macrosomia: a literature review. Ann Nutr Metab 2015;66 Suppl 2(14-20):14-20. [doi: 10.1159/000371628] [Medline: 26045324]
4. Leth-Møller M, Hulman A, Kampmann U, Hede S, Ovesen PG, Knorr S. Effect of gestational diabetes on fetal growth rate and later overweight in the offspring. J Clin Endocrinol Metab 2025 Apr 22;110(5):1350-1357. [doi: 10.1210/clinem/dgae428] [Medline: 38916475]
5. Blotsky AL, Rahme E, Dahhou M, Nakhla M, Dasgupta K. Gestational diabetes associated with incident diabetes in childhood and youth: a retrospective cohort study. CMAJ 2019 Apr 15;191(15):E410-E417. [doi: 10.1503/cmaj.181001] [Medline: 30988041]
6. Laredo-Aguilera JA, Gallardo-Bravo M, Rabanales-Sotos JA, Cobo-Cuenca AI, Carmona-Torres JM. Physical activity programs during pregnancy are effective for the control of gestational diabetes mellitus. Int J Environ Res Public Health 2020 Aug 24;17(17):6151. [doi: 10.3390/ijerph17176151] [Medline: 32847106]
7. Zakaria H, Abusanana S, Mussa BM, et al. The role of lifestyle interventions in the prevention and treatment of gestational diabetes mellitus. Medicina 2023;59(2):287. [doi: 10.3390/medicina59020287]
8. Taousani E, Papaioannou KG, Mintziori G, et al. Lifestyle behaviors and gestational diabetes mellitus: a narrative review. Endocrines 2025;6(1):6. [doi: 10.3390/endocrines6010006]
9. Craig L, Sims R, Glasziou P, Thomas R. Women's experiences of a diagnosis of gestational diabetes mellitus: a systematic review. BMC Pregnancy Childbirth 2020 Feb 7;20(1):76. [doi: 10.1186/s12884-020-2745-1] [Medline: 32028931]
10. Gammeltoft TM, Nguyen TA, Dung TK, et al. The pioneers of Vietnam's epidemiological transition: an ethnographic study of pregnant women's experiences of gestational diabetes. Glob Health Action 2024 Dec 31;17(1):2341521. [doi: 10.1080/16549716.2024.2341521] [Medline: 38693861]
11. Ái NT, Ti n NV, Dung VTK, et al. The challenges pregnant women face in daily life living with gestational diabetes mellitus: a qualitative study in Thai Binh province [Article in Vietnamese]. Vietnam J Prev Med 2024;34(2):136-143. [doi: 10.51403/0868-2836/2024/1631]
12. Björk Javanshiri A, Modig S, Nymberg P, Calling S. Women's experience of gestational diabetes and healthcare in southern Sweden - a qualitative study. BMC Pregnancy Childbirth 2025 Feb 28;25(1):224. [doi: 10.1186/s12884-025-07328-2] [Medline: 40021994]
13. Goldschmidt VJ, Colletta B. The challenges of providing diabetes education in resource-limited settings to women with diabetes in pregnancy: perspectives of an educator. Diabetes Spectr 2016 May;29(2):101-104. [doi: 10.2337/diaspect.29.2.101] [Medline: 27182180]
14. International Diabetes Federation. The Diabetes Atlas. 2021. URL: https://diabetesatlas.org/ [accessed 2026-04-24]

XSL•FO
RenderX

15. Hirst JE, Tran TS, Do MAT, Rowena F, Morris JM, Jeffery HE. Women with gestational diabetes in Vietnam: a qualitative study to determine attitudes and health behaviours. BMC Pregnancy Childbirth 2012 Aug 9;12:81. [doi: 10.1186/1471-2393-12-81] [Medline: 22873351]
16. Hatch S, Kickbusch I, editors. Self-Help and Health in Europe: New Approaches in Health Care: World Health Organization, Regional Office for Europe; 1983. URL: https://wellcomecollection.org/works/fv23gk8j [accessed 2025-12-10]
17. Narasimhan M, Aujla M, Van Lerberghe W. Self-care interventions and practices as essential approaches to strengthening health-care delivery. Lancet Glob Health 2023 Jan;11(1):e21-e22. [doi: 10.1016/S2214-109X(22)00451-X] [Medline: 36306809]
18. World Health Organization. WHO Guideline on Self-Care Interventions for Health and Well-Being, 2022 Revision 2022. URL: https://iris.who.int/server/api/core/bitstreams/ebbf577b-d3ff-4a81-b0d9-bdf3b1e9f2ac/content [accessed 2025-12-10]
19. Leask CF, Sandlund M, Skelton DA, et al. Framework, principles and recommendations for utilising participatory methodologies in the co-creation and evaluation of public health interventions. Res Involv Engagem 2019;5(2):2. [doi: 10.1186/s40900-018-0136-9] [Medline: 30652027]
20. Agnello DM, Balaskas G, Steiner A, Chastin S. Methods used in co-creation within the health CASCADE co-creation database and gray literature: systematic methods overview. Interact J Med Res 2024 Nov 11;13:e59772. [doi: 10.2196/59772] [Medline: 39527793]
21. Xie W, Dai P, Qin Y, Wu M, Yang B, Yu X. Effectiveness of telemedicine for pregnant women with gestational diabetes mellitus: an updated meta-analysis of 32 randomized controlled trials with trial sequential analysis. BMC Pregnancy Childbirth 2020 Apr 6;20(1):198. [doi: 10.1186/s12884-020-02892-1] [Medline: 32252676]
22. Leblalta B, Kebaili H, Sim R, Lee SWH. Digital health interventions for gestational diabetes mellitus: a systematic review and meta-analysis of randomised controlled trials. PLOS Digit Health 2022 Feb;1(2):e0000015. [doi: 10.1371/journal.pdig.0000015] [Medline: 36812531]
23. Wei HX, Yang YL, Luo TY, Chen WQ. Effectiveness of mobile health interventions for pregnant women with gestational diabetes mellitus: a systematic review and meta-analysis. J Obstet Gynaecol 2023 Dec;43(2):2245906. [doi: 10.1080/01443615.2023.2245906] [Medline: 37605977]
24. Zahmatkeshan M, Zakerabasali S, Farjam M, Gholampour Y, Seraji M, Yazdani A. The use of mobile health interventions for gestational diabetes mellitus: a descriptive literature review. J Med Life 2021;14(2):131-141. [doi: 10.25122/jml-2020-0163] [Medline: 34104235]
25. Janamian T, Crossland L, Wells L. On the road to value co-creation in health care: the role of consumers in defining the destination, planning the journey and sharing the drive. Med J Aust 2016 Apr 18;204(7 Suppl):S12-S14. [doi: 10.5694/mja16.00123] [Medline: 27078786]
26. Duncan E, O'Cathain A, Rousseau N, et al. Guidance for reporting intervention development studies in health research (GUIDED): an evidence-based consensus study. BMJ Open 2020 Apr 8;10(4):e033516. [doi: 10.1136/bmjopen-2019-033516] [Medline: 32273313]
27. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014 Mar 7;348:g1687. [doi: 10.1136/bmj.g1687] [Medline: 24609605]
28. National Assembly of the Socialist Republic of Vietnam. Resolution 202/2025/QH15 on the reorganization of provincial-level administrative divisions. 2025 URL: https://english.luatvietnam.vn/co-cau-to-chuc/resolution-202-2025-qh15-reorganization-of-provincial-level-administrative-divisions-402728-d1.html [accessed 2026-04-24]
29. Linde DS, Le HM, Vu DTK, et al. A co-created self-care and informal support intervention targeting women with gestational diabetes mellitus in northern Vietnam (VALID-II): a protocol for a two-arm non-randomised feasibility study. Pilot Feasibility Stud 2025 May 29;11(1):73. [doi: 10.1186/s40814-025-01657-x] [Medline: 40442746]
30. Le HM, Nguyen TA, Nguyen DK, et al. Prevalence and risk factors of gestational diabetes mellitus among pregnant women in northern Vietnam: a cross-sectional study. Glob Health Action 2025 Dec;18(1):2460339. [doi: 10.1080/16549716.2025.2460339] [Medline: 39925195]
31. Decision no.6173/QD-BYT: national guidelines on prevention and management of gestational diabetes mellitus. 2018 URL: https://thuvienphapluat.vn/van-ban/The-thao-Y-te/Quyet-dinh-6173-QD-BYT-2018-phe-duyet-Huong-dan-du-phong-kiem-soat-dai-thao-duong-thai-ky-530062.aspx [accessed 2025-12-10]
32. Decision no.1470/QD-BYT: national guidelines on diagnosis and management of gestational diabetes mellitus. 2024 URL: https://thuvienphapluat.vn/van-ban/The-thao-Y-te/Quyet-dinh-1470-QD-BYT-2024-tai-lieu-chuyen-mon-sang-loc-quan-ly-dai-thao-duong-thai-ky-611911.aspx [accessed 2025-12-10]
33. Dang NAT, Le HM, Nguyen A, et al. Self-care interventions among women with gestational diabetes mellitus in low and middle-income countries: a scoping review. Syst Rev 2025 Feb 27;14(1):50. [doi: 10.1186/s13643-025-02790-7] [Medline: 40016820]
34. Dung VTK, Ái NT, Anh Đ, et al. Reactions of pregnant women and families when pregnant women are diagnosed with gestational diabetes: a qualitative study in Thai Binh province [Article in Vietnamese]. Vietnam Med J 2024(541):165-171. [doi: 10.51298/vmj.v541i1.10644]

XSL•FO
RenderX

35. Tien, NV, Gammeltoft, T, Bygbjerg IC, et al. Experiences of women with gestational diabetes in using social media for self-care: a qualitative study in northern Vietnam. VPHA 2024 Jun 30;67:28-39. [doi: 10.53522/ytcc.vi67.03]
36. Ph ng NTM, Dung VTK, Ái NT, et al. Husband's support when wife has gestonial diabetes mellitus: a qualitative study. Vietnam J Prev Med 2024 Sep;34(2):127-135. [doi: 10.51403/0868-2836/2024/1630]
37. Nunes ML, Félix B, Nunes F, Santos I. Systematic development and refinement of a user-centered evidence-based digital toolkit for supporting self-care in gestational diabetes mellitus. Sci Rep 2025 Apr 8;15(1):12009. [doi: 10.1038/s41598-025-96318-7] [Medline: 40199963]
38. Halvorsrud K, Kucharska J, Adlington K, et al. Identifying evidence of effectiveness in the co-creation of research: a systematic review and meta-analysis of the international healthcare literature. J Public Health (Oxf) 2021 Apr 12;43(1):197-208. [doi: 10.1093/pubmed/fdz126] [Medline: 31608396]
39. Vietnam Ministry of Science and Technology. Zalo remains Vietnam's most-used messaging app. 2023. URL: https://english.mic.gov.vn/zalo-remains-vietnams-most-used-messaging-app-197157664.htm [accessed 2026-04-24]
40. Lê AV, L ng MP, Đ Đ, Tr n MN, Bui TD. Technology in Education: A Case Study on Viet Nam: UNESCO; 2023. [doi: 10.54676/GPDF6007]
41. Understanding time limits for zoom meetings. Zoom Support. 2025. URL: https://support.zoom.com/hc/en/article?id=zm_kb&sysparm_article=KB0067966 [accessed 2026-04-24]
42. Webel AR, Okonsky J, Trompeta J, Holzemer WL. A systematic review of the effectiveness of peer-based interventions on health-related behaviors in adults. Am J Public Health 2010 Feb;100(2):247-253. [doi: 10.2105/AJPH.2008.149419] [Medline: 20019321]
43. Thompson DM, Booth L, Moore D, Mathers J. Peer support for people with chronic conditions: a systematic review of reviews. BMC Health Serv Res 2022 Mar 31;22(1):427. [doi: 10.1186/s12913-022-07816-7] [Medline: 35361215]
44. Whitehead L, Jacob E, Towell A, Abu-Qamar M, Cole-Heath A. The role of the family in supporting the self-management of chronic conditions: a qualitative systematic review. J Clin Nurs 2018 Jan;27(1-2):22-30. [doi: 10.1111/jocn.13775] [Medline: 28231630]
45. Gammeltoft T, Dung VTK, Dang NA, Phuong NT, Ai NT, Bygbjerg IC. The 5-C model of digital health accompaniment: empowering women through a gestational diabetes self-care intervention in Vietnam. Front Clin Diabetes Healthc 2026;7:1738433. [doi: 10.3389/fcdhc.2026.1738433]
46. DataReportal. Digital 2024: vietnam. 2024. URL: https://datareportal.com/reports/digital-2024-vietnam [accessed 2026-04-24]
47. Kosowicz L, Tran K, Khanh TT, et al. Lessons for Vietnam on the use of digital technologies to support patient-centered care in low-and middle-income countries in the Asia-Pacific region: scoping review. J Med Internet Res 2023 Apr 5;25:e43224. [doi: 10.2196/43224] [Medline: 37018013]

Abbreviations

GDM: gestational diabetes mellitus
GUIDED: guidance for reporting intervention development studies in health research checklist
HCP: health care professional
LMICs: low- and middle-income countries
SDU: University of Southern Denmark
T2D: type 2 diabetes
TBUMP: Thai Binh University of Medicine and Pharmacy
TIDieR: template for intervention description and replication checklist
UCPH: University of Copenhagen
WHO: World Health Organization

Edited by A Stone; submitted 15.Aug.2025; peer-reviewed by D Walker, O Shiyanbola; revised version received 02.Apr.2026; accepted 03.Apr.2026; published 06.May.2026.

Please cite as:
Dang NAT, Meyrowitsch DW, Linde DS, Vinter CA, Nguyen TD, Nguyen DK, Nguyen TA, Le HM, Vu TKD, Tran TH, Gammeltoft TM, Søndergaard J
Digital Health–Based Peer Support Ecosystem for Gestational Diabetes Mellitus in Vietnam (VALID II Study): Multistakeholder Cocreation and Pilot Study
J Med Internet Res 2026;28:e82434
URL: https://www.jmir.org/2026/1/e82434
doi:10.2196/82434

XSL•FO
RenderX

© Ngoc Anh Thi Dang, Dan W Meyrowitsch, Ditte S Linde, Christina A Vinter, Thanh Duc Nguyen, Dang Kien Nguyen, Thi Ai Nguyen, Hieu Minh Le, Thi Kim Dung Vu, Thi Huong Tran, Tine M Gammeltoft, Jens Søndergaard. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Factors Influencing Continuance Intention for Online Consultations Among Survivors of Cancer: Grounded Theory Study

Yutang Yao[1*], MD; Musi Zhang[2*], BSN; Shanshan Peng[2], BMed; Zhuzhong Cheng[1], MD; Yun Duan[1], BMed

[1]Department of Nuclear Medicine, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

[2]Department of Radiation Oncology, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

[*]these authors contributed equally

Corresponding Author:
Yun Duan, BMed
Department of Nuclear Medicine
Sichuan Cancer Hospital & Institute, Sichuan Cancer Center
University of Electronic Science and Technology of China
55 Renmin South Rd Section 4
Chengdu, Sichuan, 610041
China
Phone: 86 02885420214
Email: duanyun163@126.com

Abstract

Background: Online consultation platforms have become an important component of survivorship care for patients with cancer, offering flexible access to oncology expertise between scheduled visits. However, evidence on what drives the willingness of survivors of cancer to continue using online consultations after initial adoption remains limited in China. A better understanding of continuance intention is needed to inform survivor-centered digital health strategies.

Objective: This study aimed to explore the influencing factors of continued use of online consultations among survivors of cancer in southwest China and develop a grounded theoretical model explaining continuance intention.

Methods: A grounded theory qualitative design was used. A total of 26 adult survivors of cancer with diverse demographic and clinical characteristics were purposively recruited from a tertiary cancer center in southwest China. All participants had used online consultations at least once in the preceding year. Semistructured telephone interviews were audio recorded; transcribed verbatim; and analyzed using open, axial, and selective coding with constant comparison until theoretical saturation was reached. During selective coding, categories and their relationships were integrated and iteratively refined to construct a grounded theoretical model of continuance intention.

Results: Six interrelated domains influenced survivors' continued use of online consultation platforms: platform quality, physician competence, user perception, individual condition, external context, and privacy concerns. Platform quality and physician competence influenced user perception of usefulness, reassurance, and trust, which functioned as a mediator of continued use. Individual condition, including health status, health literacy, and psychological needs, influenced both perceived usefulness and reliance on online consultations. External context, especially family encouragement, peer recommendations, and availability of local oncology services, directly facilitated or constrained continued use. Privacy concerns moderated how survivors balanced perceived benefits against risks of data misuse, stigma, and unwanted disclosure of cancer history. Survivors described online consultations as offering rapid guidance and emotional support that complemented hospital-based care but reported discontinuation when interactions were delayed or impersonal or when perceived privacy risks outweighed the benefits.

Conclusions: The willingness of survivors of cancer to continue using online consultation platforms depends on multiple interrelated factors beyond traditional technological usability. Sustained engagement is shaped by survivors' perceptions of usefulness and trust, physician empathy and timeliness, family encouragement, and acceptance of privacy trade-offs. The theoretical model advances understanding of digital health continuance in oncology and offers practical guidance for developing survivor-centered online consultation services.

(J Med Internet Res 2026;28:e84644) doi:10.2196/84644

XSL•FO
RenderX

KEYWORDS

survivors of cancer; online consultation; telemedicine; digital health; grounded theory; continuance intention; patient engagement; physician-patient communication; privacy concerns; survivorship care

Introduction

Background

Cancer remains one of the leading causes of morbidity and mortality worldwide, placing a tremendous burden on health systems, families, and individuals. Advances in diagnosis and treatment have improved survival rates, and as a result, the population of survivors of cancer continues to expand [1-3]. However, survivorship is not simply the period after treatment. It also includes ongoing challenges such as long-term side effects, fear of recurrence, psychosocial adjustment, and the need for continuous follow-up care. The World Health Organization and many national cancer control programs have emphasized the importance of survivorship care as an integral component of comprehensive cancer management [4,5].

Digital health technologies have become increasingly important in bridging gaps in access to care, especially for patients requiring long-term monitoring [6-8]. Online consultation platforms, accessed through mobile apps or web portals, allow survivors to communicate with oncology professionals, seek advice on symptoms or lifestyle modifications, and receive psychological support. For survivors living in regions with limited oncology resources, online consultation provides a vital link to specialists otherwise geographically inaccessible [9-12]. Even in urban centers with advanced hospitals, the platforms offer convenience and continuity between scheduled in-person visits [13,14].

The COVID-19 pandemic further accelerated the adoption of telemedicine and online health services, demonstrating their potential to supplement traditional care [14-16]. Survivors of cancer, often immunocompromised, benefitted from reduced exposure to hospital environments through digital consultations. As health systems adapt to postpandemic realities, online consultation is positioned as a permanent component of integrated cancer survivorship care. However, the long-term viability of such services depends not merely on initial adoption but also on survivors' willingness to continue using them. Therefore, understanding the factors that shape continued use is essential for sustaining the role of online consultation in oncology care.

Research on online consultation and telemedicine has expanded rapidly in recent years, with a growing focus on use patterns and determinants of adoption [17-19]. Several theoretical models have guided this scholarship. The technology acceptance model (TAM) and expectation confirmation model (ECM) are frequently used to examine perceived usefulness, ease of use, and satisfaction as drivers of continued use. Studies across various digital health contexts, including mobile health apps, wearable devices, and patient portals, have confirmed that perceived usefulness and trust are strong predictors of sustained engagement [19-22].

In oncology, early work has highlighted the potential of online consultation to support symptom management, medication adherence, and communication between patients and health care professionals. For example, randomized controlled trials involving survivors of breast cancer have demonstrated that digital follow-up systems improve quality of life and reduce hospital visits [23-26]. Observational studies in Europe and Asia have reported that online consultation platforms are especially effective in providing dietary advice, managing side effects, and delivering psychosocial support to survivors [27-29].

A growing body of literature has investigated continuance intention in digital health contexts. In nononcology populations, continuance is influenced by habit formation, social influence, and perceived quality of service. For instance, research on chronic disease management apps has found that individuals with strong digital literacy and positive reinforcement from peers are more likely to continue use [30-32]. Other studies have identified cost, accessibility, and integration with offline care as determinants of sustained use [33,34]. However, in cancer populations, empirical evidence remains fragmented. Some studies of survivors of breast cancer have revealed that online consultation provides emotional reassurance and complements offline care [35-37], whereas others have found that survivors discontinue use due to inconsistent physician availability and lack of trust [38].

In this study, continuance intention refers to survivors' intention to keep using an online consultation platform after initial adoption rather than a single episode of use. In digital health, continuance intention is essential because the benefits of telemedicine and online consultation usually accumulate through repeated contacts, ongoing symptom management, and sustained relationships with clinicians. If survivors discontinue use after early trials, online services may show good initial uptake yet fail to deliver long-term gains in symptom control, psychological support, or care coordination.

While prior studies provide valuable insights into adoption of and satisfaction with online consultations, few have systematically investigated the factors influencing continued use by survivors of cancer over time. Most quantitative research has relied on preexisting models such as the TAM or ECM, which capture perceptions of technology but may not adequately reflect the lived realities of survivorship. Survivors navigate a complex interplay of medical, psychological, social, and technological factors. Their willingness to sustain engagement with online consultation platforms cannot be reduced to perceived usefulness alone.

There is also a lack of theory-building research specific to oncology survivorship in the Chinese context. Most studies of digital health continuance have applied established frameworks such as the TAM and ECM with a focus on perceived usefulness, ease of use, and satisfaction. Such models were not developed for the complex realities of cancer survivorship, where survivors simultaneously manage late effects, fear of

XSL•FO
RenderX

recurrence, family expectations, and resource constraints in the health system. Without theory grounded in survivors' lived experiences, especially within Chinese sociocultural and institutional settings, it is difficult to design online consultation services that are both acceptable and sustainable. Theory-building work can clarify which technological, relational, and contextual influences matter most for continuance and how they interact during survivorship care.

Objectives

Therefore, this study aimed to explore the factors that influence the continued use of online consultation platforms by survivors of cancer in southwest China and develop a grounded theoretical model explaining continuance intention in this context. Specifically, this study sought to identify individual, relational, and contextual influences on continued use and articulate how such influences interrelate within survivorship care.

Methods

Study Design

This study used a grounded theory qualitative design to examine factors influencing continuance intention regarding online consultations among survivors of cancer. The approach by Strauss and Corbin [39] was adopted, with iterative cycles of data collection and analysis and a 3-stage process of open coding, axial coding, and selective coding [40]. Constant comparative analysis was used to compare incidents within and across interviews and refine categories and their properties as the theoretical model developed [41].

Study Setting

This study was conducted at the Sichuan Cancer Hospital and Institute, Sichuan Cancer Center, the Affiliated Cancer Hospital of the University of Electronic Science and Technology of China. This institution is the largest national tertiary cancer hospital in southwest China, with comprehensive functions in cancer prevention, treatment, rehabilitation, research, and education. Its extensive clinical services and diverse patient population made it an ideal setting for examining the experiences of survivors of cancer with online consultation platforms.

Participants and Recruitment

Participants were adult survivors of cancer with diverse cancer types and in diverse survivorship stages who had used an online consultation platform at least once in the preceding year. Inclusion criteria were age of ≥18 years, confirmed cancer diagnosis, completion of initial treatment or in active follow-up, at least one prior online consultation related to cancer care within the previous year, and ability to communicate via telephone in Mandarin. Survivors with severe cognitive impairment or acute clinical distress or who were unable to communicate effectively via phone were excluded. The criterion of at least one recent online consultation ensured that participants could describe concrete experiences with platform use and decisions about continued use rather than hypothetical views.

Recruitment followed purposive sampling to maximize diversity in age, gender, cancer type, treatment stage, and place of residence. During outpatient follow-up appointments, clinical nurses and oncologists briefly introduced the study to eligible survivors and, with permission, shared contact details with the research team. The team then telephoned interested survivors to provide detailed study information, confirm eligibility, and arrange an interview time. In survivorship support groups and patient-led online forums, a short study notice invited interested survivors to contact the team directly. All invitations were active rather than open public advertisements. Degree of experience with online consultation and digital literacy were not used as formal sampling strata. Instead, the focus was on variation in survivorship trajectories and clinical backgrounds. Digital literacy was not assessed using a standardized scale, which is acknowledged as a limitation when interpreting differences in continuance intention.

Data Collection

Data were collected through semistructured telephone interviews, which were chosen to accommodate survivors living in different regions and reduce travel burden. After eligibility was confirmed, interviews were scheduled at times convenient for participants, usually outside routine clinic visits. Two trained oncology nurses (YY and MZ) conducted all interviews. Both interviewers worked at the same tertiary cancer center but were not part of the clinical team directly responsible for participants' current treatment. At the start of each interview, the interviewer introduced their professional background, clarified the voluntary nature of participation, and emphasized that decisions about care would not be affected by participation. Interviews followed a flexible guide covering experiences with online consultations, reasons for continued use or discontinuation, family and social influences, and privacy concerns. (Multimedia Appendix 1) Interviews lasted between 25 and 40 minutes, were audio recorded with permission, and were transcribed verbatim in Mandarin. Each transcript was anonymized and assigned an identifier (C01 to C26).

Data Analysis

Data analysis began after the first interviews and proceeded concurrently with ongoing data collection. Two researchers (YY and MZ) conducted line-by-line open coding on an initial set of transcripts to identify concepts related to survivors' experiences of online consultations, perceived benefits and drawbacks, relational and family influences, and privacy concerns. (Multimedia Appendix 2) Codes were compared, merged, and refined in regular meetings, and a preliminary coding framework was developed.

During axial coding, conceptually similar codes were clustered into categories, and relationships among categories were explored using constant comparison across participants and time points. Analytic memos documented emerging ideas about potential mediators, moderators, and contextual conditions influencing continuance intention. In the selective coding stage, categories were integrated into 6 higher-level domains and a core category of continuance intention. The developing theoretical model was iteratively checked against the data, including accounts that appeared to deviate from early interpretations, and revised until it accounted for the range of observed patterns.

XSL•FO
RenderX

Coding was conducted independently by the 2 researchers, and discrepancies were discussed and resolved through consensus, with a third researcher (YD) consulted when needed. An audit trail including codebooks, memos, and diagrams was maintained to support transparency and dependability. Sampling became increasingly focused as analysis progressed, for example, by recruiting survivors from different age groups and residential areas once the importance of family involvement and privacy concerns became clear. Theoretical saturation was assessed after 23 interviews when no new categories were identified and relationships between domains appeared stable. Three additional interviews were conducted to confirm saturation and ensure that the model held true for more recent cases.

Rigor and Trustworthiness

Credibility and trustworthiness were supported through multiple strategies. First, purposive and iterative sampling captured survivors with diverse demographic and clinical backgrounds to enhance variation in experiences. Second, 2 researchers independently coded transcripts and compared interpretations in regular analysis meetings, with disagreements resolved through discussion and involvement of a third researcher where required. Third, constant comparison and negative case analysis were used to test whether the evolving model could accommodate accounts that challenged early assumptions. Fourth, an audit trail of coding decisions, memos, and diagrams was maintained to support dependability and confirmability. Finally, brief member checking was conducted with 4 participants who were invited to comment on thematic summaries; they confirmed that the domains and relationships reflected their experiences. Reporting follows the Standards for Reporting Qualitative Research and is informed by the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines.

Researcher Positionality

The research team consisted of oncology clinicians and nursing researchers working in a tertiary cancer center in southwest China. The interviewers were oncology nurses with long-standing experience caring for survivors during treatment and follow-up, which facilitated rapport but may also have shaped the topics explored and the way in which participants described their care. The senior author is a nuclear medicine physician with experience in survivorship care and digital health initiatives in the hospital where this study was conducted. The team acknowledges that familiarity with hospital-affiliated online platforms and a generally positive view of digital health could influence interpretation of the data. To address this, assumptions were documented in analytic memos, and team discussions explicitly considered alternative explanations and accounts that did not align with expectations.

Ethical Considerations

The study protocol was reviewed and approved by the ethics committee of Sichuan Cancer Hospital and Institute (approval SCCHEC-02-2020-036). All participants received verbal and written information about the study and provided informed verbal consent before the interviews. Participation was voluntary, and survivors could decline questions or withdraw at any time without consequences for their clinical care.

To protect privacy and confidentiality, audio recordings were stored on password-protected devices accessible only to the research team, and transcripts were deidentified by removing names and other direct identifiers. Potentially identifying combinations of demographic and clinical details were aggregated in reporting so that individual participants could not be recognized. No financial incentives or material compensation were provided for participation.

Results

Participant Characteristics

The 26 participants included 15 (58%) women and 11 (42%) men, with most being middle-aged (40-49 years: n=8, 31%; 50-59 years: n=7, 27%). Survivors of breast cancer constituted the largest group (n=7, 27%), followed by lung (n=4, 15%) and colorectal (n=3, 12%) cancer, whereas other cancer types such as cervical, ovarian, prostate, gastric, and thyroid cancer each represented 8% (n=2), and liver and kidney cancer each represented 4% (n=1). Survivorship stages were balanced, with 31% (n=8) in active treatment and 35% (n=9) each in remission and long-term survivorship. Educational levels ranged from 23% (n=6) with a high school or lower level to 12% (n=3) with postgraduate education. Most participants (n=11, 42%) were employed, and 62% (n=16) lived in urban areas as presented in Table 1.

Analysis identified 6 domains shaping the continued use by survivors of cancer of online consultation platforms: platform quality, physician competence, user perception, individual condition, external context, and privacy concerns. The interrelationships among the domains formed a theoretical model explaining continuance intention.

XSL•FO
RenderX

Table 1. Sociodemographic and clinical characteristics of survivors of cancer participating in grounded theory interviews on continuance intention regarding online consultations in southwest China (N=26).

Category	Participants, n (%)
Gender	
Female	15 (58)
Male	11 (42)
Age (years)	
30-39	5 (19)
40-49	8 (31)
50-59	7 (27)
≥60	6 (23)
Cancer type	
Breast	7 (27)
Lung	4 (15)
Colorectal	3 (12)
Cervical	2 (8)
Ovarian	2 (8)
Prostate	2 (8)
Gastric	2 (8)
Thyroid	2 (8)
Liver	1 (4)
Kidney	1 (4)
Treatment stage	
Active treatment	8 (31)
Remission	9 (35)
Long-term survivorship	9 (35)
Educational level	
High school or lower	6 (23)
College or junior college	10 (38)
Bachelor's degree	7 (27)
Master's degree or higher	3 (12)
Occupation status	
Employed	11 (42)
Retired	7 (27)
Homemaker	3 (12)
Self-employed	3 (12)
Unemployed	2 (8)
Geographic location	
Urban	16 (62)
Semiurban or rural	10 (38)

Platform Quality

Platform quality encompassed service quality and information quality. Survivors valued consistent access to oncology specialists, transparent consultation fees, reliable technical support, and effective complaint resolution. Information quality included completeness of physician profiles, clarity of treatment explanations, and access to trustworthy cancer-related educational content as presented in Table 2.

XSL•FO
RenderX

Table 2. Themes and subthemes related to platform quality influencing continuance intention regarding online consultations among survivors of cancer.

Subtheme	Key insights	Quote
Service quality	Survivors expected efficient systems, reasonable fees, and transparent policies.	"The charges were clear. I could reach the same oncologist again. The convenience made me continue." [C04; male; survivor of gastric cancer]
Information quality	Accurate physician and treatment information increased trust.	"I can see which doctors are experts or which are younger doctors. It's easier than going to the hospital." [C21; female; survivor of ovarian cancer]

Physician Competence

Physician competence emerged as central. Survivors highlighted professional expertise, communication ability, and timeliness. Expertise reassured patients on side effect management and recurrence risks. Compassionate communication provided emotional support. Timely responses were critical, especially during treatment as presented in Table 3.

Table 3. Themes and subthemes related to physician competence influencing continuance intention regarding online consultations among survivors of cancer.

Subtheme	Key insights	Quote
Professional expertise	Detailed, tailored explanations increased trust.	"When the doctor explained why my fatigue persisted, it felt like guidance I could trust." [C18; male; survivor of lung cancer]
Communication ability	Survivors valued empathy and encouragement.	"She told me that anxiety was normal, and suddenly I felt less alone. That made me come back." [C09; female; survivor of breast cancer]
Timeliness	Delays discouraged use; quick replies reinforced reliance.	"During chemo, hours felt like days. A fast reply was the reason I kept using the app." [C23; male; survivor of colorectal cancer]

User Perception

User perception comprised ease of use, perceived usefulness, and positive expectation. Survivors emphasized the importance of intuitive interfaces and smooth navigation. Usefulness was defined as reassurance between hospital visits, management of treatment side effects, and guidance for lifestyle adaptation. Positive expectation referred to trust in the future development of digital health services.

Survivors described online consultations as a lifeline during uncertain recovery phases. One participant noted the following:

> *Even when nothing urgent, knowing I could reach professional doctors quickly gave me comfort.* [C06; female; survivor of cervical cancer]

Individual Condition

Individual condition referred to health literacy, health status, and personal needs. Survivors with higher health literacy were more confident in engaging online, whereas those with limited literacy faced challenges in sustained use. Health status influenced patterns: those experiencing lingering treatment effects frequently sought reassurance, whereas long-term survivors without active symptoms used consultation more sporadically. Needs extended beyond medical queries to psychological reassurance and lifestyle advice.

One participant said the following:

> *After surgery I still had numbness in my hands, and I didn't always know if it was normal. Having the online consultation gave me a way to check quickly without waiting weeks for my hospital appointment.* [C02; female; aged 52 years; survivor of breast cancer]

External Context

External context included family encouragement, peer influence, and offline health care alternatives. Family members often facilitated continued use, particularly adult children assisting older survivors with technology. Peer groups shared recommendations of reliable platforms, which reinforced trust. Survivors in rural areas with limited oncology services relied heavily on online consultations, whereas those living near tertiary hospitals sometimes preferred in-person visits.

For instance, a participant mentioned the following:

> *My son insisted I keep using the app. He said it was safer than traveling two hours to the hospital every time I worried about something small, and he even helped me learn how to pay for consultations.* [C09; male; aged 63 years; survivor of colorectal cancer]

Privacy Concerns

Privacy concerns centered on disclosure of medical records, genetic test results, and sensitive images. Survivors expressed anxiety about misuse of their cancer history but balanced this against perceived benefits. When consultations provided timely reassurance, survivors accepted privacy trade-offs.

One survivor explained the following:

> *I hesitated before uploading my CT scans, but the doctor's advice was worth it.* [C12; male; survivor of prostate cancer]

Interrelationships Among Domains

The findings revealed that platform quality and physician competence had a strong influence on user perception of online consultations, including perceived usefulness, reassurance, and trust. User perception, in turn, mediated the relationship between these domains and continuance intention. Survivors' decisions

XSL•FO
RenderX

to continue or discontinue online consultations depended on how they interpreted their cumulative experiences rather than on isolated platform attributes. Individual condition, including health status, health literacy, and psychological needs, directly shaped both perceived usefulness and reliance on online consultations. External context, such as family encouragement, peer influence, and accessibility of local oncology services, provided structural conditions that either facilitated or constrained continued use. Privacy concerns moderated the influence of other domains by intensifying or weakening the impact of perceived benefits. Survivors with high privacy concerns sometimes restricted or stopped use despite recognizing benefits, whereas institutional trust in hospital-affiliated platforms could soften privacy fears and support continued engagement (Table 4).

Table 4. Representative relationships among domains in the grounded theory model and their descriptions.

Relationship	Description
User perception→intention	Perceptions of usefulness and trust shaped continuity.
Platform quality→perception	Service quality enhanced trust and ease of use.
Physician competence→perception	Physician expertise built reassurance.
External context→intention	Peer recommendations influenced adoption.
Individual condition→intention	Active symptoms increased reliance.
Privacy concern→intention	Anxiety reduced willingness.

Theoretical Model

The final theoretical model positions continuance intention at the center, influenced by 6 interconnected domains. Platform quality and physician competence directly influence user perception, which mediates the relationship with continuance. Individual condition and external context exert both direct and indirect effects. Privacy concerns act as a contextual moderator, influencing the strength of survivors' willingness to continue (Figure 1).

Figure 1. Grounded theory model of continuance intention for online consultations among survivors of cancer. Platform quality and physician competence influence continuance intention indirectly through user perception (mediator). Individual condition and external context have direct effects on continuance intention. Privacy concerns have a direct negative effect on continuance intention and moderate the association between user perception and continuance intention (N=26).

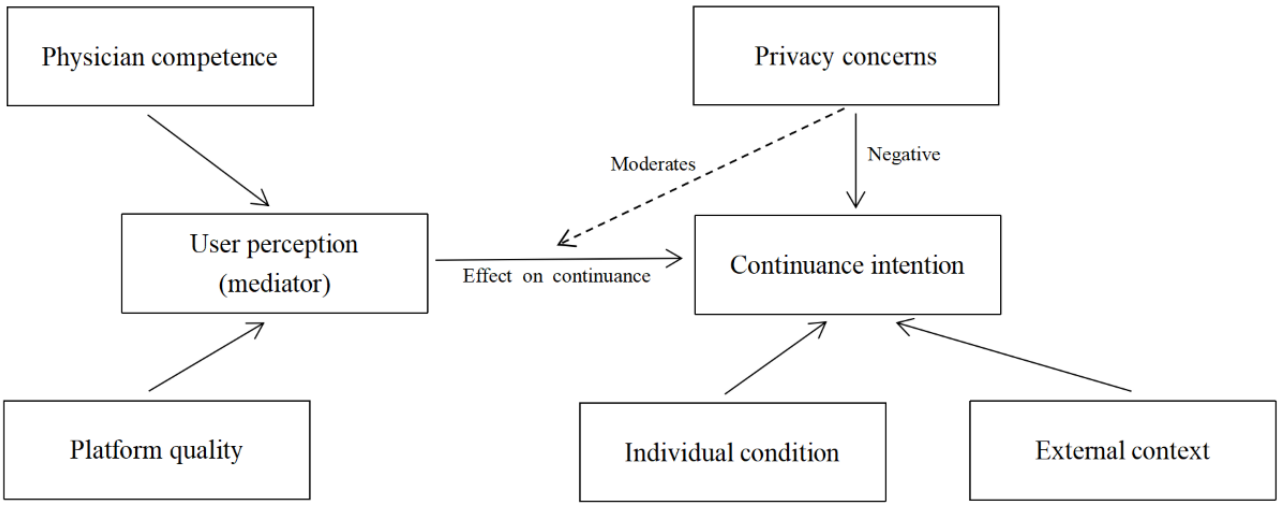

Discussion

Principal Findings

Analysis of in-depth interviews with 26 Chinese survivors of cancer identified 6 interrelated domains that shaped continuance intention: platform quality, physician competence, user perception, individual condition, external context, and privacy concerns. The domains were integrated into a theoretical model illustrating how survivors' decisions to continue using online consultation platforms were mediated by perceptions of usefulness, trust, and accessibility while being simultaneously shaped by their personal health status, family influences, and privacy considerations. The findings underscore that continued use is not determined solely by technological ease or usefulness but emerges from a dynamic negotiation among survivors' individual needs, relational contexts, and systemic constraints.

Comparison With Prior Work

Previous studies have consistently identified perceived usefulness, trust, and habit formation as determinants of sustained use. For example, research on mobile disease management apps has demonstrated that convenience and ease of monitoring reinforce long-term engagement [42,43]. Other studies have linked continuance intention to perceptions of quality of physician feedback and integration into daily routines [44,45]. This suggests that the mechanisms of continuance

XSL•FO
RenderX

identified in our study are not unique to oncology but resonate with broader digital health behaviors.

Survivors of cancer often experience uncertainty long after completion of primary treatment, with concerns about recurrence, lingering side effects, and psychosocial adjustment [46-48]. Unlike users of general wellness or chronic disease apps, survivors in our study emphasized the importance of empathetic physician communication and reassurance as key reasons for continuing online consultations. This aligns with studies on survivors of breast cancer, which have found that online consultations offer emotional support and alleviate isolation [49,50]. By centering survivors' voices, our grounded theory extends beyond existing quantitative models that have often prioritized technological features over relational care.

Privacy concerns emerged as a moderator of continuance intention. Prior work has documented anxiety about data security in telemedicine broadly [51,52], but our findings reveal that survivors actively weigh the risks of disclosure against perceived benefits of reassurance. For example, some participants expressed hesitation about uploading genetic test results or sensitive images but, ultimately, continued use when the consultation reduced uncertainty. This dynamic risk-benefit trade-off is less discussed in the literature but resonates with recent studies in Chinese mobile health contexts, which report that patients are more willing to share personal data when platforms are affiliated with reputable hospitals [53]. Our study shows that this negotiation is particularly salient in oncology, where personal health information carries social and familial implications.

The findings need to be interpreted within the sociocultural context of southwest China. Survivors' decisions about online consultations were often negotiated within families rather than made individually. Adult children in particular encouraged continued use, assisted with technical tasks, and sometimes controlled access to online services. Trust in public hospital-affiliated platforms reduced worries about data misuse compared with commercial platforms and supported acceptance of privacy trade-offs. At the same time, differences in digital literacy and internet access between urban and rural areas and between younger and older survivors shaped how readily survivors could use and benefit from online consultations. Such features of the Chinese context influence the transferability of the model to other settings and illustrate the value of theory development that is grounded in local sociocultural dynamics.

The configuration of the model would likely differ in health care contexts characterized by lower institutional trust or more individualistic decision-making. In low-trust settings, institutional affiliation may be less able to buffer privacy concerns, and privacy concerns may exert a more direct negative effect on continuance rather than primarily moderating perceived benefits. In more individualistic contexts, the external context pathway via family facilitation may be weaker, whereas individual appraisal of risk, autonomy, and personal preference may play a stronger role in shaping continuance intention.

Theoretical and Practical Implications

This study advances understanding of digital oncology care in several ways. It focused on continuance intention rather than on initial adoption or satisfaction, emphasizing the long-term dynamics of survivor engagement. It developed a theory-informed framework grounded in the lived experiences of survivors of cancer in China, addressing an important gap in survivorship research. The model highlights how technological features, physician competence, survivor perceptions, individual conditions, family and social contexts, and privacy concerns interact to shape continued use of online consultation platforms. As a result, it offers a more comprehensive account of digital health continuance and practical guidance for policymakers, health care institutions, and technology developers who seek to design sustainable services for survivors of cancer.

Theoretically, this study shows the value of grounded theory for examining continuance intention in digital health and extends work based on the TAM and the ECM. Existing frameworks foreground technological usability, perceived usefulness, and satisfaction. In contrast, this model integrates relational, contextual, and personal health factors as core elements of continuance, thereby expanding the conceptual vocabulary for studying sustained engagement with health technologies. Unlike consumer technologies where continuance is often driven by convenience, entertainment, or routine fit, continuance intention in oncology survivorship is structurally anchored in clinical uncertainty, perceived vulnerability, and reliance on professional reassurance between episodic in-person visits. Survivors re-engage when symptoms, late effects, or fear of recurrence create a need for interpretation and emotional stabilization, making the relationship with clinicians and the perceived safety of the channel central to sustained use. In this context, reassurance and trust operate as distinct relational mechanisms rather than as subcomponents of perceived usefulness. Reassurance reflects reduction of uncertainty and emotional distress through clinician responsiveness and empathy, whereas trust reflects confidence in physician competence and institutional credibility. Both shape whether survivors interpret online consultations as safe and legitimate for ongoing survivorship management, which explains why user perception mediates continuance intention through reassurance and trust in addition to instrumental usefulness.

The model also clarifies the role of user perception. Survivors' willingness to continue was shaped not only by platform quality or physician competence in isolation but also by how combined experiences influenced perceptions of usefulness, trust, and reassurance. User perception functioned as an active interpretive process that mediated the impact of system attributes and survivorship trajectories on continuance intention. In addition, this study offers a more refined understanding of privacy concerns in digital health. Privacy concerns operated as a dynamic moderator that affected the strength of continuance intention depending on how survivors weighed perceived benefits against perceived risks. This view moves beyond simple classifications of users as either concerned or unconcerned and encourages consideration of continuance as an ongoing negotiation.

XSL•FO
RenderX

From a practical standpoint, the findings suggest several priorities for online consultation services in oncology. Platforms should ensure clear physician profiles, transparent fee structures, and reliable technical support so that survivors can form stable expectations about service quality. Clinicians who provide online consultations require support and training in empathetic communication to address survivors' psychological and emotional needs in addition to clinical questions. Timely responses are particularly critical for survivors in active treatment, which supports the integration of triage protocols, clear coverage arrangements, and notification systems that reduce avoidable delays. Addressing privacy concerns calls for transparent communication about data protection policies and visible affiliation with trusted institutions; survivors in this study expressed greater confidence in platforms linked to tertiary hospitals. Finally, given the central role of family encouragement, platform features that facilitate caregiver involvement, such as options for shared access or family-oriented consultation modes, may align with prevailing cultural practices and help sustain survivor engagement over time.

Limitations

First, this study was conducted in a single regional tertiary cancer center in southwest China. Although the institution serves a large and diverse catchment area, experiences in other regions or health systems may differ. Second, the sample comprised survivors who were willing and able to participate in telephone interviews and who had had at least one prior online consultation, which may bias findings toward survivors who are more engaged with digital services. Standardized measures of digital literacy and prior telehealth experience were not collected, limiting the ability to quantify their influence on continuance intention. Third, interviews were conducted by oncology nurses affiliated with the same institution, which may have introduced social desirability bias or inhibited criticism of hospital-based platforms despite efforts to emphasize independence from clinical decisions. Finally, qualitative findings are interpretive and context specific; future research using mixed methods and larger samples is needed to test and refine the model in other oncological and cultural settings.

Future Directions

Future research should test the theoretical model developed in this study through quantitative methods. Large-scale surveys could examine the relative influence of the 6 domains and validate their interrelationships. Cross-cultural comparative studies would be valuable to assess whether the role of family support and privacy negotiation is unique to China or generalizable to other contexts. Furthermore, longitudinal research could explore how continuance intention evolves across different phases of survivorship, from active treatment to long-term follow-up.

Intervention studies could also evaluate strategies to strengthen continuance. For instance, training programs for physicians in digital empathy or platform designs that allow for caregiver participation may enhance perceptions of usefulness and trust. Policy research should investigate frameworks for data protection that balance survivors' privacy with the need for effective digital oncology services. As cancer survivorship continues to grow worldwide, such research is critical for integrating online consultations into sustainable, patient-centered models of care.

Conclusions

Six interrelated domains influenced the continuance intention of survivors of cancer to use online consultation platforms: platform quality, physician competence, user perception, individual condition, external context, and privacy concerns. The grounded theory model shows that continued use results from a dynamic negotiation between perceived benefits and perceived risks that is shaped by survivors' health needs, family roles, and institutional trust. Beyond summarizing determinants of service use, the model provides a theory-based foundation for designing and evaluating online consultation services that are more responsive to survivors' long-term needs. By clarifying how technological, relational, and contextual influences interact in a Chinese oncology setting, this study contributes to broader efforts to build sustainable and equitable digital health systems in cancer survivorship care.

Acknowledgments

During the preparation and revision of this manuscript, we used generative artificial intelligence tools, namely Kimi Chat (version K2; Moonshot AI) and GPT-5 and ChatGPT-5.1 (OpenAI) to assist with original draft translation, language polishing, and restructuring, which were further reviewed and revised by the study group.

Funding

This study was supported by the Sichuan Medical and Health Care Promotion Institute Scientific Research Project (grant). The funder had no role in the study design, data collection, data analysis, decision to publish, or preparation of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: YY, MZ, YD

XSL•FO
RenderX

Formal analysis: YY, MZ, SP, ZC, YD
Funding acquisition: YY
Investigation and data collection: YY, MZ, SP, ZC
Methodology: YY, MZ, SP
Supervision: YY, MZ
Visualization: YY, ZC
Writing—original draft: YY, MZ, SP
Writing—review and editing: YY, MZ, SP, ZC, YD

Conflicts of Interest

None declared.

Multimedia Appendix 1
Semistructured interview guide.
[PDF File (Adobe PDF File), 75 KB - jmir_v28i1e84644_app1.pdf]

Multimedia Appendix 2
Open coding concepts.
[XLSX File (Microsoft Excel File), 12 KB - jmir_v28i1e84644_app2.xlsx]

References

1. Chen ZD, Zhang PF, Xi H, Wei B, Chen L, Tang Y. Recent advances in the diagnosis, staging, treatment, and prognosis of advanced gastric cancer: a literature review. Front Med (Lausanne) 2021 Oct 26;8:744839. [doi: 10.3389/fmed.2021.744839] [Medline: 34765619]
2. Zeineddine FA, Zeineddine MA, Yousef A, Gu Y, Chowdhury S, Dasari A, et al. Survival improvement for patients with metastatic colorectal cancer over twenty years. NPJ Precis Oncol 2023 Feb 13;7(1):16 [FREE Full text] [doi: 10.1038/s41698-023-00353-4] [Medline: 36781990]
3. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. CA Cancer J Clin 2021 Jan 12;71(1):7-33 [FREE Full text] [doi: 10.3322/caac.21654] [Medline: 33433946]
4. Jacobsen PB, DeRosa AP, Henderson TO, Mayer DK, Moskowitz CS, Paskett ED, et al. Systematic review of the impact of cancer survivorship care plans on health outcomes and health care delivery. J Clin Oncol 2018 Jul 10;36(20):2088-2100 [FREE Full text] [doi: 10.1200/JCO.2018.77.7482] [Medline: 29775389]
5. Chan RJ, Crawford-Williams F, Crichton M, Joseph R, Hart NH, Milley K, et al. Effectiveness and implementation of models of cancer survivorship care: an overview of systematic reviews. J Cancer Surviv 2023 Feb 16;17(1):197-221 [FREE Full text] [doi: 10.1007/s11764-021-01128-1] [Medline: 34786652]
6. Marthick M, McGregor D, Alison J, Cheema B, Dhillon H, Shaw T. Supportive care interventions for people with cancer assisted by digital technology: systematic review. J Med Internet Res 2021 Oct 29;23(10):e24722 [FREE Full text] [doi: 10.2196/24722] [Medline: 34714246]
7. Garg S, Williams NL, Ip A, Dicker AP. Clinical Integration of digital solutions in health care: an overview of the current landscape of digital technologies in cancer care. JCO Clin Cancer Inform 2018 Dec;2:1-9. [doi: 10.1200/CCI.17.00159] [Medline: 30652580]
8. Mikles SP, Griffin AC, Chung AE. Health information technology to support cancer survivorship care planning: a systematic review. J Am Med Inform Assoc 2021 Sep 18;28(10):2277-2286 [FREE Full text] [doi: 10.1093/jamia/ocab134] [Medline: 34333588]
9. Morris BB, Rossi B, Fuemmeler B. The role of digital health technology in rural cancer care delivery: a systematic review. J Rural Health 2022 Jun;38(3):493-511 [FREE Full text] [doi: 10.1111/jrh.12619] [Medline: 34480506]
10. Shi J, Robinson TA, Loomba P, Murley B, Strong LL, Basen-Engquist K, et al. "Everything cannot be handled virtually": understanding intention and use of digital health technologies among rural adults and rural cancer survivors. J Rural Health 2025 Jan 29;41(1):e12926. [doi: 10.1111/jrh.12926] [Medline: 39878383]
11. Hemming L, Duijts SF, Zomerdijk N, Cockburn C, Yuen E, Hardman R, et al. A systematic review of peer support interventions to improve psychosocial functioning among cancer survivors: can findings be translated to survivors with a rare cancer living rurally? Orphanet J Rare Dis 2024 Dec 20;19(1):473 [FREE Full text] [doi: 10.1186/s13023-024-03477-3] [Medline: 39707418]
12. Pradhan P, Sharman AR, Palme CE, Elliott MS, Clark JR, Venchiarutti RL. Models of survivorship care in patients with head and neck cancer in regional, rural, and remote areas: a systematic review. J Cancer Surviv (Forthcoming) 2024 Jul 20. [doi: 10.1007/s11764-024-01643-x] [Medline: 39031309]

XSL•FO
RenderX

13. Bin Ramlee MH, Abdul Satar NF. The feasibility and preliminary efficacy of a locally developed digital health intervention in improving breast cancer survivor's quality of life (QOL): qualitative and quantitative analysis result. J Clin Oncol 2024 Jun 01;42(16_suppl):e13803. [doi: 10.1200/jco.2024.42.16_suppl.e13803]
14. Onyeaka HK, Zambrano J, Longley RM, Celano CM, Naslund JA, Amonoo HL. Use of digital health tools for health promotion in cancer survivors. Psychooncology 2021 Aug;30(8):1302-1310 [FREE Full text] [doi: 10.1002/pon.5677] [Medline: 33742737]
15. Toni E, Ayatollahi H. An insight into the use of telemedicine technology for cancer patients during the COVID-19 pandemic: a scoping review. BMC Med Inform Decis Mak 2024 Apr 19;24(1):104 [FREE Full text] [doi: 10.1186/s12911-024-02507-1] [Medline: 38641567]
16. Royce TJ, Sanoff HK, Rewari A. Telemedicine for cancer care in the time of COVID-19. JAMA Oncol 2020 Nov 01;6(11):1698-1699 [FREE Full text] [doi: 10.1001/jamaoncol.2020.2684] [Medline: 32672821]
17. Pang N, Lau J, Fong SY, Wong CY, Tan KK. Telemedicine acceptance among older adult patients with cancer: scoping review. J Med Internet Res 2022 Mar 29;24(3):e28724 [FREE Full text] [doi: 10.2196/28724] [Medline: 35348462]
18. Irurita-Morales P, Soto-Ruiz N, Martín-Rodríguez LS, Escalada-Hernández P, García-Vivar C. Use of telehealth among cancer survivors: a scoping review. Telemed J E Health 2023 Jul 01;29(7):956-985. [doi: 10.1089/tmj.2022.0351] [Medline: 36445755]
19. Cox A, Lucas G, Marcu A, Piano M, Grosvenor W, Mold F, et al. Cancer survivors' experience with telehealth: a systematic review and thematic synthesis. J Med Internet Res 2017 Jan 09;19(1):e11 [FREE Full text] [doi: 10.2196/jmir.6575] [Medline: 28069561]
20. Mou J, Shin DD, Cohen J. Understanding trust and perceived usefulness in the consumer acceptance of an e-service: a longitudinal investigation. Behav Inf Technol 2016 Jul 22;36(2):125-139. [doi: 10.1080/0144929X.2016.1203024]
21. Adjekum A, Blasimme A, Vayena E. Elements of trust in digital health systems: scoping review. J Med Internet Res 2018 Dec 13;20(12):e11254 [FREE Full text] [doi: 10.2196/11254] [Medline: 30545807]
22. Liu K, Tao D. The roles of trust, personalization, loss of privacy, and anthropomorphism in public acceptance of smart healthcare services. Comput Human Behav 2022 Feb;127:107026. [doi: 10.1016/j.chb.2021.107026]
23. Pimentel-Parra GA, Soto-Ruiz MN, San Martín-Rodríguez L, Escalada-Hernández P, García-Vivar C. Effectiveness of digital health on the quality of life of long-term breast cancer survivors: a systematic review. Semin Oncol Nurs 2023 Aug;39(4):151418 [FREE Full text] [doi: 10.1016/j.soncn.2023.151418] [Medline: 37045645]
24. Saltbæk L, Bidstrup PE, Karlsen RV, Høeg BL, Horsboel TA, Belmonte F, et al. Nurse-led individualized follow-up versus regular physician-led visits after early breast cancer (MyHealth): a phase III randomized, controlled trial. J Clin Oncol 2024 Jun 10;42(17):2038-2049. [doi: 10.1200/JCO.23.01447] [Medline: 38498781]
25. Saevarsdottir SR, Gudmundsdottir SL. Mobile apps and quality of life in patients with breast cancer and survivors: systematic literature review. J Med Internet Res 2023 Jul 26;25:e42852 [FREE Full text] [doi: 10.2196/42852] [Medline: 37494111]
26. Little P, Bradbury K, Stuart B, Barnett J, Krusche A, Steele M, et al. Digital intervention (renewed) to support symptom management, wellbeing, and quality of life among cancer survivors in primary care: a randomised controlled trial. Br J Gen Pract 2025 May;75(754):e357-e365. [doi: 10.3399/BJGP.2023.0262] [Medline: 38164562]
27. Lazar DE, Postolica R, Hanganu B, Mocanu V, Ioan BG. Web-based nutrition: a useful resource for cancer patients? Front Nutr 2023;10:1134793 [FREE Full text] [doi: 10.3389/fnut.2023.1134793] [Medline: 37457987]
28. Medina JC, Flix-Valle A, Rodríguez-Ortega A, Hernández-Ribas R, Lleras de Frutos M, Ochoa-Arnedo C. ICOnnecta't: development and initial results of a stepped psychosocial ehealth ecosystem to facilitate risk assessment and prevention of early emotional distress in breast cancer survivors' journey. Cancers (Basel) 2022 Feb 15;14(4):974 [FREE Full text] [doi: 10.3390/cancers14040974] [Medline: 35205722]
29. Walsh CA, Al Achkar M. A qualitative study of online support communities for lung cancer survivors on targeted therapies. Support Care Cancer 2021 Aug;29(8):4493-4500 [FREE Full text] [doi: 10.1007/s00520-021-05989-1] [Medline: 33458808]
30. Taylor ML, Thomas EE, Vitangcol K, Marx W, Campbell KL, Caffery LJ, et al. Digital health experiences reported in chronic disease management: an umbrella review of qualitative studies. J Telemed Telecare 2022 Nov 08;28(10):705-717. [doi: 10.1177/1357633x221119620]
31. Verweel L, Newman A, Michaelchuk W, Packham T, Goldstein R, Brooks D. The effect of digital interventions on related health literacy and skills for individuals living with chronic diseases: a systematic review and meta-analysis. Int J Med Inform 2023 Sep;177:105114. [doi: 10.1016/j.ijmedinf.2023.105114] [Medline: 37329765]
32. Li Z, Lu F, Wu J, Bao R, Rao Y, Yang Y, et al. Usability and effectiveness of eHealth and mHealth interventions that support self-management and health care transition in adolescents and young adults with chronic disease: systematic review. J Med Internet Res 2024 Nov 26;26:e56556 [FREE Full text] [doi: 10.2196/56556] [Medline: 39589770]
33. Endalamaw A, Zewdie A, Wolka E, Assefa Y. A scoping review of digital health technologies in multimorbidity management: mechanisms, outcomes, challenges, and strategies. BMC Health Serv Res 2025 Mar 15;25(1):382 [FREE Full text] [doi: 10.1186/s12913-025-12548-5] [Medline: 40089752]
34. Gentili A, Failla G, Melnyk A, Puleo V, Tanna GL, Ricciardi W, et al. The cost-effectiveness of digital health interventions: a systematic review of the literature. Front Public Health 2022;10:787135 [FREE Full text] [doi: 10.3389/fpubh.2022.787135] [Medline: 36033812]

XSL•FO
RenderX

35. Akkol-Solakoglu S, Hevey D. Internet-delivered cognitive behavioural therapy for depression and anxiety in breast cancer survivors: results from a randomised controlled trial. Psychooncology 2023 Mar 30;32(3):446-456. [doi: 10.1002/pon.6097] [Medline: 36635249]
36. Peng L, Yang Y, Chen M, Xu C, Chen Y, Liu R, et al. Effects of an online mindfulness-based intervention on fear of cancer recurrence and quality of life among Chinese breast cancer survivors. Complement Ther Clin Pract 2022 Nov;49:101686. [doi: 10.1016/j.ctcp.2022.101686] [Medline: 36347151]
37. Smith SK, Westbrook K, MacDermott K, Amarasekara S, LeBlanc M, Pan W. Four conversations: a randomized controlled trial of an online, personalized coping and decision aid for metastatic breast cancer patients. J Palliat Med 2020 Mar 01;23(3):353-358. [doi: 10.1089/jpm.2019.0234] [Medline: 31638448]
38. Yang M, Jiang J, Kiang M, Yuan F. Re-examining the impact of multidimensional trust on patients' online medical consultation service continuance decision. Inf Syst Front 2022 Mar 04;24(3):983-1007 [FREE Full text] [doi: 10.1007/s10796-021-10117-9] [Medline: 33688300]
39. Gosby JR. Media reviews: basics of qualitative research - techniques and procedures for developing grounded theory 2nd edition by A. Strauss and J. Corbin. Sage Publications. J Adv Nurs 2003 Jan 15;32(4):1036. [doi: 10.1046/j.1365-2648.2000.t01-7-01210.x]
40. Chun Tie Y, Birks M, Francis K. Grounded theory research: a design framework for novice researchers. SAGE Open Med 2019;7:2050312118822927 [FREE Full text] [doi: 10.1177/2050312118822927] [Medline: 30637106]
41. McCrae N, Purssell E. Is it really theoretical? A review of sampling in grounded theory studies in nursing journals. J Adv Nurs 2016 Oct 26;72(10):2284-2293. [doi: 10.1111/jan.12986] [Medline: 27113800]
42. Adu MD, Malabu UH, Malau-Aduli AE, Drovandi A, Malau-Aduli BS. User retention and engagement with a mobile app intervention to support self-management in Australians with type 1 or type 2 diabetes (my care hub): mixed methods study. JMIR Mhealth Uhealth 2020 Jun 11;8(6):e17802 [FREE Full text] [doi: 10.2196/17802] [Medline: 32525491]
43. Crafoord MT, Fjell M, Sundberg K, Nilsson M, Langius-Eklöf A. Engagement in an interactive app for symptom self-management during treatment in patients with breast or prostate cancer: mixed methods study. J Med Internet Res 2020 Aug 10;22(8):e17058 [FREE Full text] [doi: 10.2196/17058] [Medline: 32663140]
44. Yan M, Filieri R, Raguseo E, Gorton M. Mobile apps for healthy living: factors influencing continuance intention for health apps. Technol Forecast Soc Change 2021 May;166:120644. [doi: 10.1016/j.techfore.2021.120644]
45. Wang T, Wang W, Liang J, Nuo M, Wen Q, Wei W, et al. Identifying major impact factors affecting the continuance intention of mHealth: a systematic review and multi-subgroup meta-analysis. NPJ Digit Med 2022 Sep 15;5(1):145 [FREE Full text] [doi: 10.1038/s41746-022-00692-9] [Medline: 36109594]
46. Demoor-Goldschmidt C, Porro B. Editorial: young adults or adults who are survivors of childhood cancer: psychosocial side effects, education, and employment. Front Psychol 2024 Nov 8;15:1510822 [FREE Full text] [doi: 10.3389/fpsyg.2024.1510822] [Medline: 39582993]
47. Luo X, Xu H, Zhang Y, Liu S, Xu S, Xie Y, et al. Identifying the unmet needs of post-treatment colorectal cancer survivors: a critical literature review. Eur J Oncol Nurs 2024 Jun;70:102570 [FREE Full text] [doi: 10.1016/j.ejon.2024.102570] [Medline: 38574419]
48. Antoni MH, Moreno PI, Penedo FJ. Stress management interventions to facilitate psychological and physiological adaptation and optimal health outcomes in cancer patients and survivors. Annu Rev Psychol 2023 Jan 18;74(1):423-455 [FREE Full text] [doi: 10.1146/annurev-psych-030122-124119] [Medline: 35961041]
49. Niu Z, Bhurosy T, Heckman C. Cancer survivors' emotional well-being: roles of internet information seeking, patient-centered communication, and social support. J Health Commun 2021 Jul 03;26(7):514-522. [doi: 10.1080/10810730.2021.1966685] [Medline: 34435927]
50. Suarez NR, Morrow AS, LaVecchia CM, Dugas M, Carnovale V, Maraboto A, et al. Connected and supported: a scoping review of how online communities provide social support for breast cancer survivors. J Cancer Surviv (Forthcoming) 2024 Aug 28. [doi: 10.1007/s11764-024-01660-w] [Medline: 39196462]
51. Magdy M, Hosny KM, Ghali NI, Ghoniemy S. Security of medical images for telemedicine: a systematic review. Multimed Tools Appl 2022 Mar 22;81(18):25101-25145 [FREE Full text] [doi: 10.1007/s11042-022-11956-7] [Medline: 35342327]
52. Zanke P, Sontakke D. Safeguarding patient confidentiality in telemedicine: a systematic review of privacy and security risks, and best practices for data protection. Int J Curr Sci Res Rev 2024 Jun 19;07(06):3910-3922. [doi: 10.47191/ijcsrr/v7-i6-42]
53. Li X, Cong Y. A systematic literature review of ethical challenges related to medical and public health data sharing in China. J Empir Res Hum Res Ethics 2021 Dec 13;16(5):537-554. [doi: 10.1177/15562646211040299] [Medline: 34516325]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
ECM: expectation confirmation model
TAM: technology acceptance model

XSL•FO
RenderX

Edited by A Mavragani, N Cahill; submitted 23.Sep.2025; peer-reviewed by R Akinniranye, O Akhadelor, E Emokpae; comments to author 26.Nov.2025; revised version received 19.Dec.2025; accepted 23.Dec.2025; published 07.Jan.2026.

Please cite as:
Yao Y, Zhang M, Peng S, Cheng Z, Duan Y
Factors Influencing Continuance Intention for Online Consultations Among Survivors of Cancer: Grounded Theory Study
J Med Internet Res 2026;28:e84644
URL: https://www.jmir.org/2026/1/e84644
doi:10.2196/84644
PMID:41432715

©Yutang Yao, Musi Zhang, Shanshan Peng, Zhuzhong Cheng, Yun Duan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Structure of Psychopathology on Reddit: Network Analysis of Mental Health Communities in Relation to the ICD Diagnostic System

Bojan Evkoski[1], MSc; Srebrenka Letina[2], PhD; Petra Kralj Novak[1,3], PhD

[1]Department of Network and Data Science, Central European University, Vienna, Austria

[2]Department of Psychology, University of Limerick, Limerick, Ireland

[3]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia

Corresponding Author:
Bojan Evkoski, MSc
Department of Network and Data Science
Central European University
Quellenstraße 51
Vienna, 1100
Austria
Phone: 43 1252307111
Email: evkoski_bojan@phd.ceu.edu

Abstract

Background: Social media platforms such as Reddit have become important spaces where individuals articulate their distress, seek support, and explore alternative ways of understanding mental health outside traditional institutional frameworks. These environments provide an opportunity to examine mental health discourse at scale, offering perspectives that extend beyond traditional clinical and research settings.

Objective: This study aims to examine the structure of mental health communities on Reddit by identifying patterns of association between mental disorders reflected in user activity and assessing how these relationships align with established diagnostic categories in the *ICD* (*International Classification of Diseases*).

Methods: We manually curated 114 Reddit communities focused on specific mental health conditions from the 20,000 most active subreddits in 2022. Each community was labeled into 49 disorders and categorized under 9 *ICD* diagnostic categories within the group of mental and behavioral disorders, collectively known as the F codes. We constructed a disorder association network by identifying statistically significant user overlaps based on coposting across subreddit pairs using a bipartite configuration model, with Bonferroni-corrected significance ($P<.001$). We analyzed the connectivity of the network within and across diagnostic categories, examining inter- and intracategory links. Finally, we compared the structure of disorder associations inferred from Reddit with the *ICD* classification derived from diagnostic criteria using hierarchical clustering.

Results: The inferred Reddit network of psychopathology revealed an interconnected structure (density=0.135), with all but 6 disorders forming a single giant component that spans across all 9 diagnostic categories. The most prominent disorders by number of users included hyperkinetic disorders (85,000), depressive episodes and recurrent depressive disorders (73,000), habit and impulse disorders (69,000), pervasive developmental disorders (52,000), and generalized anxiety disorder (44,000). In terms of connectivity, posttraumatic stress disorder (17/48 of all possible connections), obsessive-compulsive disorder (16/48), and depersonalization-derealization disorder (15/48) emerged as the most central in the network of positive disorder associations, while schizotypal disorder, avoidant personality disorder, and agoraphobia were the most central when accounting for the association strength. At the level of disorder categories, several disorders, such as bipolar disorder and premenstrual dysphoric disorder, displayed high intercategory associations but weak intracategory ties, indicating blurred diagnostic boundaries. The network of negative coposting associations revealed a divergence from the expectations of past research; for instance, addiction-related communities (eg, alcohol and opioids) were negatively associated with much of the broader mental health discourse. Finally, hierarchical comparisons showed moderate overlap between the Reddit network of disorder associations and the *ICD* network of diagnostic criteria, both in pairwise edge similarity (13% of edges present in both networks) and overall clustering (Adjusted Rand Index=0.295).

XSL•FO
RenderX

Conclusions: Reddit-based mental health communities reveal a complementary structure of disorder associations shaped by lived experience, often diverging from formal diagnostic criteria and exhibiting patterns of association that do not align with established diagnostic boundaries.

(J Med Internet Res 2026;28:e80958) doi:10.2196/80958

KEYWORDS

psychopathology; disorder associations; online mental health support; Reddit; network analysis; ICD-10

Introduction

Mental health problems have remained one of the main public health concerns, especially among younger populations [1-3]. This trend has unfolded amid rapid technological advancements and broader societal changes, including the lasting disruptions brought by the COVID-19 pandemic [4,5]. Yet, the pace of these societal and technological shifts has not been matched by corresponding adaptability in the broader health care system.

The challenge of adaptability is particularly evident in the diagnostic frameworks of psychopathology. Clinical taxonomies such as the *DSM* (*Diagnostic and Statistical Manual of Mental Disorders*) and the *ICD* (*International Classification of Diseases*) provide standardized frameworks for identifying, labeling, and treating psychological conditions. In these systems, disorders are defined and grouped into diagnostic categories based on symptom profiles, potential causal mechanisms or course of illness. These categories serve essential roles in practice: they guide diagnostic and treatment decisions, shape insurance coverage, structure research protocols, and enable communication across professionals. However, despite their utility, these frameworks and their diagnostic categories are built on standardized assumptions that inevitably generalize and oversimplify inherently subjective and dynamic experiences. Consequently, they have faced ongoing criticism regarding the ambiguity of their boundaries and their limited clarity, stability, and cultural relevance across diverse contexts [6-8]. Recent reform efforts, such as the Research Domain Criteria and the Hierarchical Taxonomy of Psychopathology, aim to address these limitations, but also underscore the complexity and ongoing contestation surrounding psychiatric classification in the pursuit of a more fluid, flexible, and context-sensitive approach to understanding mental illness [9-11]. However, such efforts cannot be fully undertaken in isolation from the shifting social and technological environments that shape how symptoms are expressed, experienced, and interpreted.

Over the past 2 decades, digital technologies have transformed nearly every facet of daily life, including how individuals relate to their own mental health. Mobile connectivity and online social media play central roles in how people articulate their experiences, shape their identities, and seek support [12,13]. Platforms such as Reddit have emerged as key infrastructures for navigating psychological problems, particularly for individuals who may not have access to, or trust in, formal health systems [14]. The appeal of these spaces is amplified by the limitations of traditional care: clinical services often prioritize severe cases, socioeconomic barriers restrict access, and not all individuals are equally willing or able to seek professional help. Even when accessed, brief and episodic consultations may leave little room to capture the full scope of ongoing psychological struggles [15-17]. By contrast, the anonymity and decentralization of digital health communities enable frank discussions of stigmatized experiences, which allows users to articulate symptoms and explore possible explanations. In doing so, these platforms contribute to a contemporary, living discourse around mental health.

Because users can engage freely, anonymously, and repeatedly over time, digital health communities provide a naturally occurring data source for examining both the expression of diverse conditions and their interrelations [18,19]. While research on online platforms has typically emphasized diagnostic tools or peer support within isolated communities or narrow disorder sets [20-22], recent cross-community studies have started studying online mental health communities through comparative and cross-community perspectives. For example, Morini et al [23] analyzed the content of 67 mental health communities on Reddit and showed that support-seeking and venting are dominant posting intents, and that community feedback shapes subsequent participation. In parallel, Jin and Zhu [24] constructed a multimorbidity network linking diabetes communities to 88 other disease subreddits, revealing connections to mental health and weight-management forums and showing that discussion of physical illness can extend into mental health–focused communities. Such cross-community studies have the potential to complement traditional research and open the way for examining large-scale, system-level patterns of psychopathology and the diagnostic frameworks that represent them.

Empirical research on diagnostic frameworks and relationships between disorders has largely relied on 2 sources: surveys and clinical registers. Surveys capture subsets of symptoms in specific populations and depend on self-report, which is difficult to validate at scale [25-27]. On the other hand, while clinical registers are based on verified diagnoses, they typically contain biases based on severity, health care access and even diagnostic conventions [28,29]. As such, both perspectives offer important yet partial views on the structure of psychopathology. We position digital mental health communities, specifically Reddit, as a complementary source that provides behavioral signals of perceived relatedness among mental disorders. These data are not confined to small samples or predefined diagnostic categories, and they are not restricted to clinically severe cases. Instead, they capture large-scale, naturally occurring expressions of personal experience that cannot replace survey or register approaches, yet can broaden the empirical basis for psychopathology research and enable triangulation across complementary sources.

XSL•FO
RenderX

We study the structure of psychopathology on Reddit by analyzing users' coposting activity in condition-specific mental health communities. The dataset comprises 114 subreddits, each centered on a distinct disorder, collectively covering more than half a million users and over 1.5 million posts. By tracing associations across disorders through shared patterns of coposting, we adopt a data-driven network perspective on how conditions are interconnected in contemporary contexts. This perspective draws on tools from network analysis that have grown influential in mental health research, particularly network psychometrics, where disorders are modeled as systems of interacting symptoms rather than latent disease categories [30-32]. Whereas network psychometrics emphasizes within-disorder architecture, we shift the focus to relationships between disorders as reflected in cross-community engagement. In doing so, we complement symptom-level approaches of psychopathology with a view of how people navigate multiple diagnostic ideas at once, negotiating meaning and seeking support across disorder boundaries.

Our main objective is to infer a significance-based network of associations between mental health conditions as expressed through online engagement. We identify pairs that co-occur more or less often than expected, as well as mental health conditions that act as high-degree bridges across diagnostic categories. We highlight clusters of co-occurring disorders and examine strong cross-category connections that may signal transdiagnostic roles not anticipated by existing taxonomies. Finally, we compare the Reddit-derived structure with the hierarchy encoded in the *ICD-10* (*International Statistical Classification of Diseases, 10th Revision*) diagnostic criteria, outlining conceptual and structural differences that contribute to the broader discussion on psychiatric nosology. In this way, Reddit functions as both a site of peer support and a window into how people collectively make sense of mental health beyond formal clinical narratives.

Methods

Study Design

The study design centers on the structure of user coposting across disorder-specific subreddits as a behavioral proxy for latent relationships between mental health disorders. We treat statistically significant user overlap between subreddit pairs as evidence that seeking advice for one disorder is associated with seeking advice for another, indicating that the 2 disorders are conceptually proximate. Such proximity may rise from several mechanisms, such as comorbidity, where users experience or suspect multiple concurrent disorders; diagnostic progression, where users transition between diagnoses over time; and misdiagnosis, where users reconsider or question a diagnosis, whether self-identified or clinically provided. Although some overlap could reflect general cross-community engagement, previous research shows that most mental health content on Reddit centers on first-person struggles rather than general discussion [14,23]. Our own validation supports this pattern: more than two-thirds of posts contain more self-referential pronouns ("I," "me," and "myself") than references to others (refer to Multimedia Appendix 1), consistent with help-seeking grounded in personal experience. To limit residual unrelated activity, we infer edges only when observed user overlap exceeds expectations under a conservative bipartite configuration model with stringent multiple-comparison control (refer to the "Network Inference" subheading below). As a final major design choice, we decided to focus on activity through posts rather than comments. Posts in disorder-specific communities are the primary venue for sharing experiences and seeking help, whereas comments are more reactive, often offering advice or feedback. While comments provide valuable perspectives on peer interaction and information diffusion, their scale and heterogeneity risk diluting the signal of the help-seeking behavior we aim to capture. Together, these design choices justify interpreting coposting as a cautious but credible signal of perceived relatedness between disorders without interpreting activity as evidence of diagnosis or medical history.

Data

We collected Reddit data through the Pushshift application programming interface by looking into the 20,000 biggest subreddits [33,34]. From this corpus, we manually curated a list of subreddits whose primary focus is to provide information, support, and shared lived experience related to specific mental health conditions. Each candidate subreddit was reviewed individually, considering both its description and posted content, and was included only if those corresponded to a distinct disorder as described in the *ICD-10* (2019 release) [35].

Each included subreddit was annotated according to its corresponding *ICD-10* diagnostic code at level 4 granularity (eg, F48.1, depersonalization-derealization syndrome). To study the complete hierarchy of psychopathology, we further annotated each community to higher-level categories of the *ICD* taxonomy, including level 3 codes (eg, F48, other neurotic disorders) and level 2 codes (eg, F4, neurotic, stress-related, and somatoform disorders). For completeness and future interoperability, we also provided mappings to the equivalent codes in the *ICD-11* (*International Classification of Diseases, 11th Revision*), a taxonomy that is yet to be used in practice. In total, 114 condition-specific mental health subreddits were identified and classified into 49 unique *ICD-10* disorders, covering 9 level 2 diagnostic categories of mental and behavioral disorders (from F0 to F9). A full list of annotated subreddits and their hierarchical coding is provided in Multimedia Appendix 2 and is publicly available for reuse, along with a separate table listing all disorders and their corresponding *ICD-10* codes.

We restricted our analyses to posts from 2022 to ensure temporal consistency and avoid confounding effects from major platform-level disruptions. This includes discontinuities associated with the COVID-19 pandemic and subsequent policy or moderation shifts, as well as more recent artifacts linked to the rise of generative artificial intelligence [36,37]. The resulting dataset comprised 1,513,016 posts authored by 545,330 unique users across all included mental health subreddits. On average, each user contributed 2.77 posts, with 96,742 (17.74%) users posting in more than 1 mental health subreddit, which forms the foundation for our analysis of disorder co-occurrence (refer to Multimedia Appendix 3 regarding the distribution of posts across disorder categories and subreddits). To enhance data

XSL•FO
RenderX

quality, we excluded accounts indicative of automated or spam-like activity. Specifically, the dataset excluded users who posted more than 365 times in the study year (corresponding to a rate of more than 1 post per day), as well as known automated accounts listed in the publicly available bot directory BotRank [38]. Ultimately, the data collection procedure resulted in a comprehensive representation of Reddit's mental health discourse by systematically covering the major condition-specific communities active on the platform.

Network Inference

Overview

The main objective of this work is to infer relationships between mental health conditions as they emerge from patterns of shared user participation across the studied disorder-related subreddits. To this end, we constructed a weighted network in which each node represents 1 of the 49 classified mental health disorders, defined at the level 4 granularity of *ICD-10* codes, and edges capture statistically significant user coposting overlaps between the sets of subreddits corresponding to each disorder pair. The construction involved 4 key steps: computing user overlap between disorder pairs, estimating a null model for expected coposting, determining statistical significance, and assigning edge weights based on deviation from null expectations.

Disorder Association Metric

For each pair of disorders, we measured the strength of coposting association by calculating the overlap coefficient between the sets of users who contributed to subreddits of each corresponding disorder. This coefficient captures the size of the user intersection normalized by the smaller of the 2 user sets:

where x and y denote the sets of unique users who posted in the subreddits associated with each disorder.

The overlap coefficient is particularly suited for capturing asymmetric coengagement. By measuring the proportion of shared users relative to the smaller community, we can address cases where participation in one subreddit could be almost entirely embedded within another. This property aligns well with the hierarchical and overlapping conceptualization of many mental health conditions, where narrower or less prevalent disorders often exist within the broader spectrum of more common ones.

Null Model and Statistical Testing

To assess the significance of observed user overlaps, we used a binary bipartite configuration model as the null model. The user-disorder bipartite network was constructed with one set of nodes representing users and the other representing mental health disorders, where an edge denotes a post by a user in a subreddit labeled with a specific disorder. To generate the null distribution, we randomly rewired the bipartite network 10,000 times while preserving the degree distributions of both users and disorders and preventing multiedge cases. Each rewired network underwent several edge swaps equal to 10× the total number of edges, ensuring sufficient randomization while maintaining the original participation heterogeneity. This approach follows established practices in bipartite network analysis, where degree-preserving randomization is commonly used to construct realistic null models for statistical inference [39,40].

For each disorder pair, we computed the distribution of overlap coefficients across the 10,000 null replicates and evaluated the statistical deviation of the observed overlap using a standard z score. To correct for multiple comparisons across all possible disorder pairs ($n \times [n-1]/2$), we applied a Bonferroni-corrected significance threshold of $P<.001$ [41]. Based on this, each disorder pair falls into one of 3 categories:

- Positive association: observed overlap is significantly higher than expected.
- Negative association: observed overlap is significantly lower than expected.
- No evidence for association: observed overlap does not differ significantly from null expectations.

Edge Weights and Network Representation

For each statistically significant association, we assigned a weight equal to the difference between the observed and the mean expected overlap coefficient. This measure reflects the magnitude of deviation from the null, with higher values indicating stronger-than-expected co-occurring between disorders. Theoretically, the weights range from 0 (no deviation) to 1, with larger values signaling greater empirical association strength relative to what would be expected by chance under the null model.

The resulting structure was represented as two undirected weighted networks of positive and negative disorder associations. This network representation enabled intuitive interpretation of pairwise association patterns, capturing the relational landscape of psychopathology. An interactive online version of the network visualization was also developed to facilitate further exploration (refer to Multimedia Appendices 4 and 5) [42-44].

Node-Level Metrics

To characterize the centrality of disorders within the inferred network, we computed 2 measures of connectedness:

- Unweighted degree: the total number of significant associations (edges) for a given disorder. This captures how broadly a condition is connected across the mental health landscape.
- Weighted degree: the sum of edge weights for all associations of a disorder. This emphasizes the cumulative strength of its connections, highlighting conditions that might not have many links, yet maintain particularly strong associations.

Across analyses, we used both unweighted and weighted edges depending on the methodological requirements of each approach. Comprehensive data for both unweighted and weighted node degrees of the Reddit network are provided in Multimedia Appendix 2.

XSL•FO
RenderX

ICD-10 Network of Diagnostic Criteria

As a reference point for the Reddit-derived association network, we constructed a network of mental disorders based on formal diagnostic criteria, referred to as the *ICD-10* diagnostic criteria network. This network was built using data curated by Tio et al [45], who systematically extracted diagnostic symptoms for each *ICD-10* code falling into Chapter F: mental and behavioral disorders. This dataset provides a standardized operationalization of disorder-level symptom profiles for *ICD-10* codes that remain in clinical use [46]. It offers a unique opportunity to analyze relationships between disorders grounded in clinical definitions, which are otherwise difficult to access in a machine-readable or systematically coded form. As such, we used this network for comparing the *ICD-10* diagnostic structure with the association patterns inferred from the coposting activity on Reddit.

In the *ICD-10* diagnostic criteria network, each node (disorder) was represented as a set of diagnostic criteria, and the edge weight (strength of connection) between 2 disorders was calculated as the overlap coefficient between their diagnostic criteria sets. This formulation mirrors our approach to the Reddit network, enabling a consistent comparison across both systems. To ensure a valid basis for comparison, we restricted this network to include only those disorders for which a corresponding Reddit community had been identified in our manual curation.

Hierarchical Clustering of Association Networks

We used hierarchical clustering to study the potential modular and hierarchical structure of the association network derived from Reddit, as well as that of the *ICD-10* diagnostic criteria network. This approach, which has previously been used to uncover grouping patterns in association networks [47-49], allowed us to infer latent groupings of mental health conditions based on observed similarity patterns. Importantly, it enabled a system-level comparison of the 2 networks, moving beyond pairwise overlap to examine how broader patterns of connectivity and disorder organization differ between Reddit discourse and the formal *ICD-10* diagnostic structure, akin to comparing hierarchical network communities.

We applied standard agglomerative clustering with average linkage, where 2 clusters were merged based on the average distance between all pairs of disorders across the 2 clusters [50]. Since the weighted edges in both networks represent similarity (rather than distance), we defined the distance between any 2 disorders x and y as:

$$d(x,y) = max(e(x,y)) - e(x,y)$$

Where $e(x,y)$ is the observed similarity (edge weight) and $max(e)$ ensures that distances are positive and properly scaled for clustering.

To assess the extent to which each network exhibits clustered structure, we first computed the weighted modularity, a standard quality function that quantifies how much edge weight lies within clusters compared to between clusters, relative to the expectation under a weighted degree-preserving null model [51]. This allowed us to evaluate and compare the overall modular organization of the Reddit network and the *ICD-10* diagnostic criteria network.

For a clustering π (nodes grouped into clusters), the weighted modularity is defined as:

Where $w_{i,j}$ is the observed edge weight, $s_i=\sum_j w_{i,j}$ is the weighted degree of node i, $2W=\sum_{ij} w_{ij}$ is the total edge weight, and $1\{\cdot\}$ equals 1 when i and j are assigned to the same cluster. Intuitively, higher weighted modularity values indicate a clearer community structure. Q_w is large when within-cluster connections carry more weight than expected under a null model preserving node strengths.

Since raw modularity depends on network density and degree or strength heterogeneity, we estimated the normalized modularity for each network under a degree-sequence–preserving null model to enable a fair comparison between the Reddit network and the *ICD-10* network of diagnostic criteria. For each network, we first generated an ensemble of randomized topologies by repeatedly swapping edge pairs and their associated weights, thereby preserving the overall degree sequence and weight distribution while randomizing the network. Then, for each randomized graph G', we evaluated $Q_w(G',\pi(\tau))$ on the same partition $\pi(\tau)$ obtained from the observed network; this isolates how expected the observed within-cluster concentration of weight is, given the degree and weight distribution.

We then calculated the normalized modularity, defined as the difference between the observed modularity and the expected value under the null model:

$$\Delta Q_w=Q_w(G,\pi(\tau))-E[Q_w(G',\pi(\tau))]$$

where the latter was estimated from 1000 randomized realizations. Finally, for each dendrogram height produced through hierarchical clustering, we evaluated the normalized weighted modularity and compared the Reddit network with the *ICD-10* network of diagnostic criteria, focusing on the clustering results corresponding to the dendrogram cuts with the highest normalized weighted modularity in each network. In other words, the final clustering of a network is obtained as the cut τ that maximizes ΔQ_w across all admissible cuts:

To evaluate the similarity between the final clusters of the 2 networks, we used 2 standard comparison indices: the Adjusted Rand Index (ARI) and the normalized mutual information (NMI) [52]. ARI measures the agreement between 2 clustering results by quantifying how often pairs of nodes are grouped or separated in the same way, adjusted for chance. NMI captures the amount of shared information between the clusters of the networks and is less sensitive to differences in the number or size of clusters. Both metrics range from 0 (no agreement beyond chance) to 1 (perfect correspondence).

XSL•FO
RenderX

Ethical Considerations

The subreddits analyzed in this study are publicly accessible and do not require login credentials. Posts are shared under pseudonymous accounts, and all usernames were further pseudoanonymized before analysis to protect privacy. Given the sensitive and personal nature of the disclosures that may appear in these vulnerable communities, we applied strict privacy safeguards and limited all reporting to aggregated results, without focusing on individual cases and basing our analysis only on coposting rather than the content of the posts themselves. The study was purely observational, involving no interaction or intervention with users. We did not attempt to contact individuals, and no analyses were conducted that could enable reidentification of participants. Our approach followed widely recognized ethical frameworks for internet-mediated research [53-56]. In particular, we adhered to the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2), which emphasizes proportional safeguards for minimizing risks when analyzing data from public platforms where individuals can reasonably expect to be observed without explicit consent [57]. This research was also reviewed and approved by the Department of Network and Data Science Ethical Research Committee at the Central European University (reference no 2024-2025/16/ RD/DNDS).

Results

The Reddit Network of Psychopathology

We constructed a data-driven network of mental health disorder associations based on user coposting behavior across 114 Reddit communities. Each node in the network corresponds to a disorder-level *ICD-10* code (Chapter F), and edges represent statistically significant coposting links, derived from observed versus expected user overlap.

The derived Reddit network of positive associations consists of 49 nodes and 159 edges, with a density of 0.135. Despite this relative sparsity, the network forms a single giant component encompassing 43 of the 49 disorders, exemplifying the interconnectedness of mental health conditions. Only 3 disorders remained isolated without any associations, indicating a lack of shared user engagement with other conditions. These isolated disorders are F00-F03 (dementia), F63.0 (gambling addiction), and F98.5 (stuttering).

Figure 1 (top panel) shows a circular layout of the Reddit network. Node size corresponds to the total number of users active in that disorder's subreddits, and node color denotes its *ICD-10* diagnostic category. The disorders with the highest number of users include F10 (alcohol addiction), F32-F33 (depressive episodes and recurrent depression), F63.8 (other habit and impulse disorders, including excessive masturbation and pornography addiction), F84 (pervasive developmental disorders in *ICD-10*; termed autism spectrum disorder in *ICD-11*), and F90.0 (hyperkinetic disorders in *ICD-10*; termed attention-deficit/hyperactivity disorder [ADHD] in *ICD-11*). While the distribution of user activity is heterogeneous, we observed no significant association between the number of users engaging with a given disorder and the number of connections in the network. This lack of association between volume and connectivity supports the robustness of the inference method, suggesting that network centrality reflects patterns of coposting that cannot be explained simply by subreddit size (refer to Multimedia Appendix 6 for correlation results between degree and volume).

XSL•FO
RenderX

Figure 1. Network of positive associations among disorders derived from Reddit coposting activity in 2022, covering 114 condition-specific communities (subreddits) mapped to ICD-10 (International Classification of Diseases, 10th Revision) Chapter F (Mental and Behavioral Disorders). The network includes 49 disorder nodes linked by 159 edges. Top: a circular (radial) network visualization. Each node represents a distinct ICD-10 disorder, grouped and color-coded by higher-level diagnostic category (F1-F9). Node size indicates the number of unique users who posted at least once in the corresponding subreddits. Edges connect pairs of disorders that share a statistically significant overlap in user activity, estimated with a bipartite configuration null model and corrected for multiple testing (Bonferroni-adjusted, $P<.001$). Edge density and cross-category links illustrate the extent of interconnectedness across diagnostic boundaries. Bottom: disorder-level statistics. The center panel shows the number of users associated with each disorder. The bottom panel shows the degree of each disorder, defined as the number of significant co-occurring associations, broken down into intracategory associations (links within the same ICD-10 level-2 group) and intercategory associations (links connecting disorders across different ICD-10 groups; eg, between F1 and F5). Colored bars highlight the maximum values within each diagnostic category. Together, these panels summarize both the structure of the Reddit-based disorder network and the relative prominence of individual disorders by participation size and connectivity.

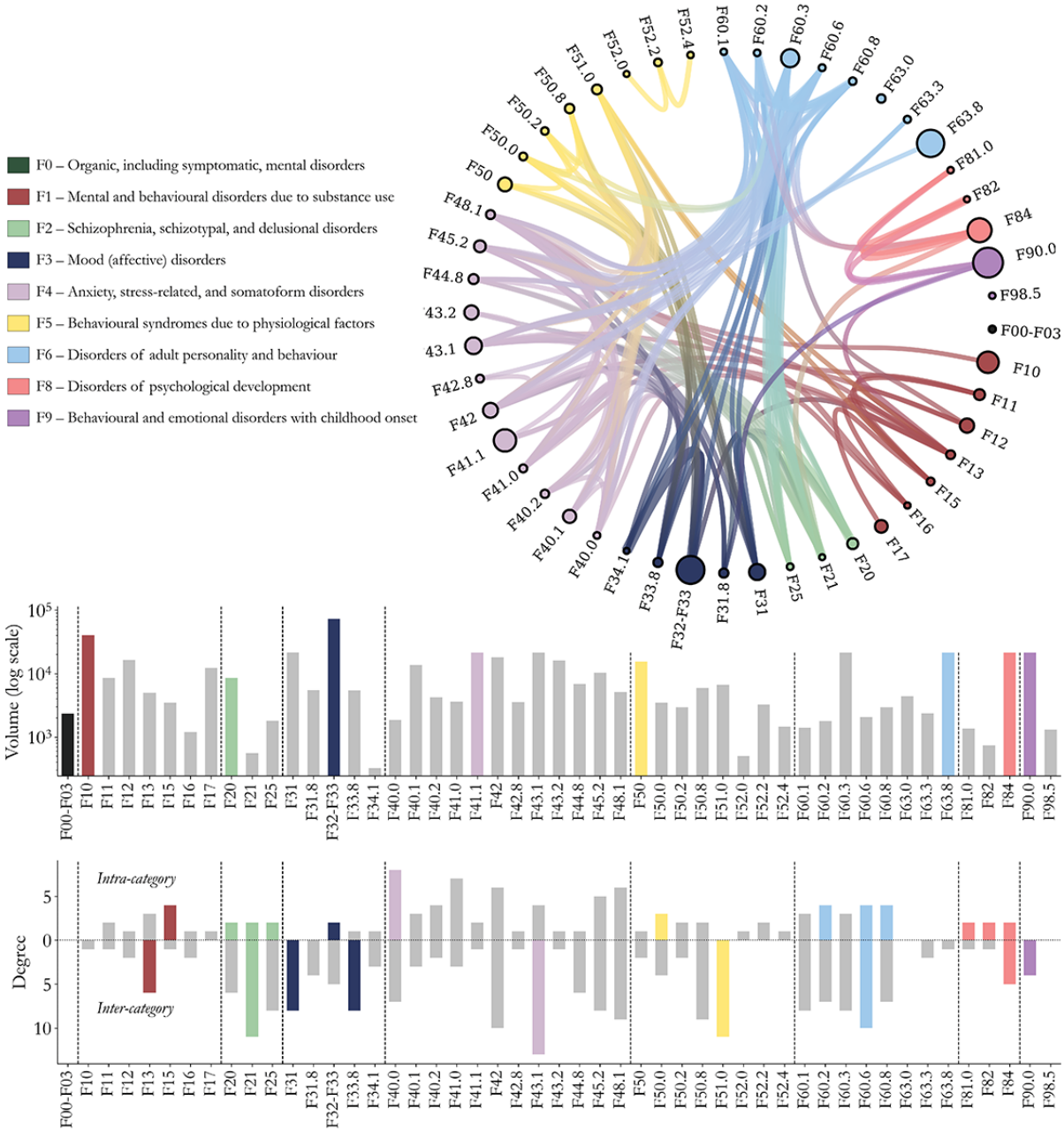

Figure 1 (bottom panel) shows the degree distribution of nodes in the network, revealing substantial heterogeneity in how strongly different disorders are connected, ranging from highly linked hubs to sparsely connected nodes (refer to Multimedia Appendix 2 for results on weighted degrees). The most connected disorders include F43.1 (posttraumatic stress disorder [PTSD]), F42 (obsessive-compulsive disorder), F40.0 (agoraphobia), F48.1 (depersonalization-derealization syndrome), and F60.6 (avoidant personality disorder). These disorders emerge as transdiagnostic hubs, a concept central to psychiatric nosology, with associations spanning a wide range of other conditions. In contrast, disorders with minimal connectivity, such as F10 (alcohol addiction), F17 (tobacco addiction), and F52.0 (lack or loss of sexual desire), may reflect more segmented or marginalized user groups, narrower community focus, or greater diagnostic specificity, similar to the 3 isolated disorders identified before in this subsection.

XSL•FO
RenderX

The network of positive associations reveals a strong presence of cross-category links, with 106 of 159 edges connecting disorders classified in different *ICD-10* diagnostic categories. Notably, some disorders show markedly higher intercategory associations than intracategory ones. For example, F43.1 (PTSD), F51.0 (nonorganic insomnia), and F33.8 (used here to denote premenstrual dysphoric disorder) exhibit the largest discrepancies, with strong cross-category ties but weak integration within their respective *ICD-10* groups. This pattern also extends to entire higher-level diagnostic categories: all conditions in F2 (schizophrenia, schizotypal, and delusional disorders) and F3 (mood and affective disorders) display more intercategory than intracategory connections, highlighting particularly fluid boundaries for these diagnostic categories.

Negative Associations

While our primary focus was on positive associations, indicative of shared user bases and potential comorbidity patterns, we also identified a set of negative associations, where user coposting occurred less frequently than expected under the null model. However, due to methodological limitations, such results should be interpreted at the level of individual edges only, not aggregated across nodes. Specifically, the inference strategy using a null model and the overlap coefficient to measure association becomes increasingly insensitive to low coposting rates in smaller or less active subreddits, introducing a lower-bound bias that distorts node-level summaries of negative associations (refer to Multimedia Appendix 7).

Despite this limitation, a consistent pattern emerges. Negative associations were most commonly observed between impulse-related disorders (eg, F10 for alcohol use, F11 for opioid use, and F63.8 for other habit and impulse disorders) and other mental health categories. These patterns may reflect distinct user populations, stigma-driven disengagement, or divergent framings of psychological distress and its management. The presence of such negative ties underscores that disorder communities on Reddit are not only interconnected but also socially and discursively fragmented in ways that do not always align with theoretical similarity and comorbidity reported in the literature. A notable example is F11 (opioid use), which showed negative associations with 7 of 12 nodes in the F4 category (neurotic, stress-related, and somatoform disorders), despite previous research suggesting strong links between anxiety and opioid use [58,59]. A full list of negative edges and their weights is provided in Multimedia Appendix 2. Considering the limitations of our analysis on negative associations, these findings warrant further investigation using alternative methodological approaches better suited to capture negative associations in smaller samples.

The Hierarchical Structure of Psychopathology: Comparison With ICD-10 Diagnostic Criteria

To better understand the structural logic underlying user-inferred relationships between mental health conditions, we examined the emergent hierarchical organization of the Reddit coposting network and compared it with a clinically derived alternative based on diagnostic criteria overlap. While pairwise associations are informative, a hierarchical view enables assessment of whether larger clusters of disorders emerge in user behavior and how these clusters compare with the taxonomy in the *ICD-10* system.

Figure 2 (top) presents both the Reddit and diagnostic criteria–based networks using circular layouts. Both networks contain the same set of 49 *ICD-10* codes but differ in how connections are formed: the Reddit network uses statistically significant coposting links based on user overlap, while the diagnostic criteria network connects disorders based on shared clinical features, as curated by Tio et al [45]. Edges in both networks are weighted by the overlap coefficient, which quantifies the proportion of shared elements between 2 sets relative to the smaller set.

XSL•FO
RenderX

Figure 2. Comparison between 2 networks of mental health disorder relationships: 1 derived from Reddit user activity and 1 constructed from clinical diagnostic definitions, illustrating both overlaps and divergences between lived-experience associations emerging from online peer support and the formal structures encoded in psychiatric nosology. Top panels: circular network visualizations. The network on the left is based on coposting behavior across 128 condition-specific Reddit communities (subreddits) active in 2022, mapped to International Classification of Diseases, 10th Revision (ICD-10) Chapter F (Mental and Behavioral Disorders). Two disorders are linked if users post in both corresponding subreddits more often than expected by chance. The network on the right is the diagnostic criteria network, based on a curated list of disorders and diagnostic criteria of ICD-10, where disorders are linked if they share at least 1 formal diagnostic feature. In both networks, node size reflects connectivity (degree), and color indicates the ICD-10 diagnostic category (F0-F9). Black edges highlight associations that are common to both networks, while colored edges are unique to either Reddit (orange) or the diagnostic criteria (blue). Bottom panels: hierarchical clustering of disorders in both networks. Agglomerative clustering with average linkage was applied to identify higher-level clusters, with the cut-off chosen to maximize normalized weighted modularity (a measure of how well the network separates into cohesive clusters beyond chance). The resulting normalized weighted modularity values were ΔQ_w=0.420 for Reddit and ΔQ_w=0.521 for the diagnostic criteria network, showing that the Reddit network is more interconnected across diagnostic categories compared with the more modular diagnostic criteria–based network. Cluster labels are simplified, human-readable summaries of underlying ICD-10 codes. ADHD: attention-deficit/hyperactivity disorder; ARI: Adjusted Rand Index; NMI: normalized mutual information; OCD: obsessive-compulsive disorder; PTSD: posttraumatic stress disorder.

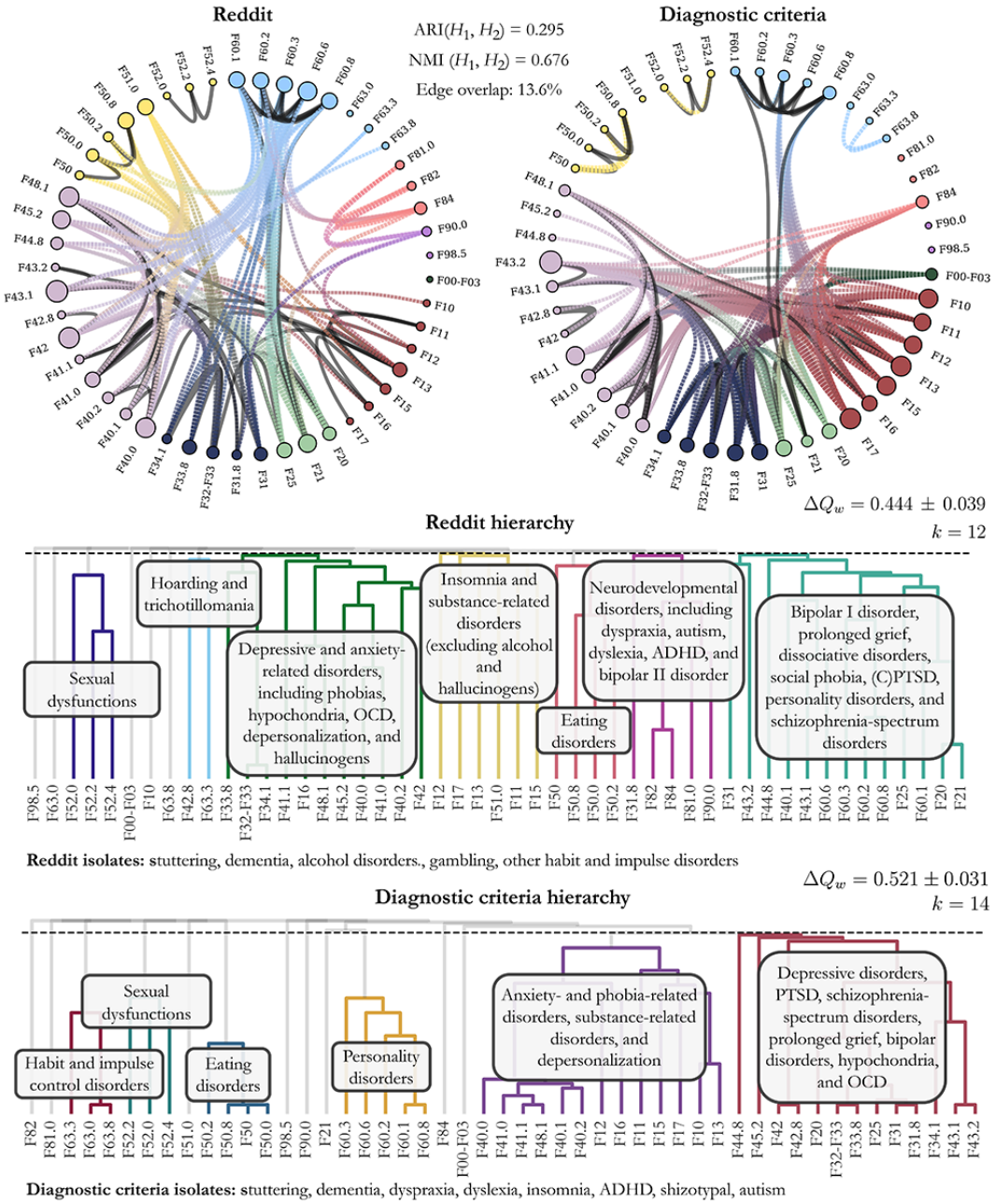

We applied agglomerative clustering with average linkage to both networks and identified the optimal clustering by cutting the dendrogram at the level of maximum normalized weighted modularity (ΔQ_{max}). The resulting modularity scores were 0.444 (12 clusters) for Reddit and 0.521 (14 clusters) for the diagnostic criteria network (Figure 2, bottom). This finding suggests that the Reddit network is less modular, reflecting less distinct clusters and more overlapping patterns of association compared with the more compartmentalized structure based on symptom overlap.

XSL•FO
RenderX

Overall, the comparison of the 2 hierarchies showed partial alignment between Reddit-based coposting behaviors and the formal diagnostic structure, with areas of both convergence and divergence. Only 13% of links were shared between the 2 networks, highlighting a substantial dissimilarity in their underlying associations. This limited edge overlap suggests that user-inferred connections diverge notably from those based on diagnostic criteria. However, when comparing their hierarchical structure, we observe more alignment. The ARI of 0.295 and the NMI of 0.676 indicate moderate similarity between the clusters of the 2 networks, pointing to partial similarity in how disorders are categorized despite the underlying differences in pairwise associations.

The Reddit network also presents a more convoluted structure, with only 5 disorders remaining unclustered at the modularity-optimal cut: F98.5 (stuttering), F00-F03 (dementia), F10 (alcohol addiction), F63.0 (gambling addiction), and F63.8 (other habit and impulse disorders). In contrast, the diagnostic criteria network leaves 8 disorders unclustered, with only 2 repeated from the Reddit network (stuttering and dementia). Importantly, some of the conditions that remain unclustered in the *ICD-10*–based network occupy more central positions in the hierarchy of the Reddit network. These include F21 (schizotypal disorder), F43.1 (PTSD), F44.81 (dissociative identity disorder), F84 (autism spectrum), and F90.0 (ADHD). Although still considered peripheral in traditional diagnostic systems, several of these conditions have received growing attention in recent years, either for their proposed transdiagnostic relevance [60,61] or for their apparent connection to broader technological and social changes, including increased screen exposure and digital media use [62-64]. Their presence in Reddit-based clusters may indicate that users organize mental health experiences in ways that diverge from formal diagnostic structures, shaped instead by evolving discussions around neurodivergence, trauma, and identity.

Some of the clusters in the Reddit hierarchy echo broader dimensional models of psychopathology. For example, F32-F33 (depressive disorders), F41.0-F41.1 (anxiety disorders), and F42 (obsessive-compulsive disorder) form a prominent cluster, an intersection that is absent from the *ICD-10* diagnostic structure but aligns with internalizing spectra described in dimensional models [10]. Another Reddit-derived cluster connects psychotic conditions such as F20 (schizophrenia) and F25 (schizoaffective disorder) with trauma-related disorders such as F43.1 (PTSD) and F44.81 (dissociative identity disorder), as well as personality disorders such as F60.3 (emotionally unstable personality disorder). This forms a cluster of thought disorders situated at the boundary between conditions typically conceptualized as internalizing or externalizing.

Discussion

Principal Findings and Relevance to Psychopathology Research

To examine how anonymous users collectively make sense of mental health problems and how this organization relates to clinical diagnostic manuals, we analyzed activity in 114 disorder-focused Reddit communities involving 545,000 users and more than 1.5M posts. We inferred a significance-based network of associations among 49 disorders spanning 9 *ICD-10* mental and behavioral disorder categories (F0-F9), derived from patterns of coposting across communities. We then compared this Reddit-based structure with a network constructed from overlaps in diagnostic criteria defined in the *ICD-10*. The Reddit network revealed a highly interconnected organization that crossed traditional diagnostic boundaries. Several disorders occupied central positions, acting as hubs that linked otherwise distinct diagnostic categories, while others, particularly substance- and behavior-related addictions, appeared less integrated into the broader mental health discourse. Comparisons of the hierarchical structure showed only partial correspondence between the Reddit network and *ICD-10*, indicating that patterns of association emerging from lived experience differ in systematic ways from those encoded in formal taxonomies. Viewed in this large-scale context, digital mental health communities do not simply mirror existing diagnostic structures. Instead, they organize mental distress through shared experience and collective interpretation, producing a socially situated view of how disorders relate to one another. This perspective extends beyond what can be captured through surveys or clinical registers alone and can help refine how comorbidity patterns and disorder boundaries are understood outside formal clinical settings.

In more detail, the statistical inference of the Reddit network revealed 159 associations between the 49 studied disorders. Despite its moderate density (0.135), the network formed a large, connected component that encompassed 43 of the 49 conditions, indicating that most conditions were linked to one another through chains of significant associations that cut across traditional diagnostic categories. Complementary patterns have been reported in symptom-level network research, where psychopathology appears as a highly interconnected system rather than clearly separated clusters [65].

Further looking into the structure of the Reddit network, several disorders emerged as central transdiagnostic hubs, including F43.1 (PTSD), F42 (obsessive-compulsive disorder), F40.0 (agoraphobia), F48.1 (depersonalization-derealization syndrome), and F60.6 (avoidant personality disorder). Their high centrality reflects symptoms that cut across diagnostic boundaries, such as intrusive thoughts, avoidance, and disturbances of identity. Addressing these transdiagnostic overlapping symptoms may therefore be key for treatment and intervention, particularly in younger populations where online peer support heavily shapes help-seeking behavior.

When considering negative associations, the Reddit network revealed specific points of disconnect between certain disorders and the broader network. Disorders related to substance use and behavioral addictions consistently appeared underconnected or negatively associated with other mental health conditions. This is in sharp contrast with comorbidity estimates typically reported in previous research, according to which substance use disorders often co-occur with other mental health conditions. For example, population-based estimates indicate that roughly 1 in 4 individuals with a substance use disorder have a comorbid mental disorder [66,67]. Our contrasting results might reflect a tendency of individuals in these communities to focus more

XSL•FO
RenderX

narrowly on managing acute behavioral symptoms or crises rather than looking into their mental health more broadly. Their relative isolation may also stem from prevailing stigmatization and limited self-recognition or acknowledgment of co-occurring mental health issues, which together tend to position externalizing disorders outside the domain of conventional psychological problems [68-70]. Whether driven by a narrow focus on symptom management, social framing, or self-stigma, these isolating mechanisms risk reinforcing silos in both peer support and clinical care, obscuring potential links between addiction and other forms of psychopathology and making it more difficult to approach treatment holistically.

Beyond overall centrality, some disorders showed distinct patterns of connectivity, forming disproportionately strong links across the F code diagnostic categories while remaining relatively weakly integrated within their own. These bridging conditions included F43.1 (PTSD), F51.0 (nonorganic insomnia), and F33.8 (used to denote premenstrual dysphoric disorder here). They illustrate transdiagnostic mechanisms that current categorical frameworks do not explicitly capture, whether through hormonally linked mood dysregulation, sleep disturbances that cut across almost all clinical categories, or trauma-related symptoms that span affective, anxiety, dissociative, and personality domains [71,72]. Notably, this pattern extended beyond individual conditions to entire diagnostic categories: all disorders in F2 (schizophrenia, schizotypal, and delusional disorders), F3 (mood and affective disorders), and the F60 subcategory (personality disorders) displayed more intercategory than intracategory links. Such patterns suggest particularly permeable diagnostic boundaries for these categories, a result that has also been observed at the level of diagnostic criteria [73].

The looseness of diagnostic boundaries was also observed through the comparative analysis between the Reddit network and the *ICD-10* network based on diagnostic criteria. The Reddit network displayed only slightly higher intercategory connectivity than the *ICD* diagnostic criteria network (68% vs 65% of all edges). However, the comparative hierarchical clustering analysis revealed that the Reddit network produced a substantially different hierarchy of disorders that only partially aligns with *ICD-10* diagnostic structures. The 2 networks showed only low to moderate similarity in their clustering (ARI=0.295 and NMI=0.676), with only 13% of edges present in both networks. Reddit also exhibited lower modularity than *ICD-10*, meaning that these clusters were less separated by diagnostic category and more interconnected across them (normalized weighted modularity of 0.444 compared to 0.521). In addition, key divergences emerged at the level of disorders. Conditions such as F21 (schizotypal disorder), F84 (autism spectrum), and F90.0 (ADHD) were central and well-integrated within the Reddit network but did not form cohesive clusters within the *ICD*-based hierarchy of diagnostic criteria. Qualitative analyses of Reddit discussions similarly report tensions between lay and professional expertise [74], while quantitative comparisons indicate that certain conditions (such as anxiety-related and affective disorders) are disproportionately represented relative to registry data [75]. Such discrepancies may stem from the platform's affordances of anonymity, its demographic composition, and its emphasis on peer support, but they may also signal blind spots in clinical frameworks. However, rather than contradicting established evidence, these divergences show how online data can surface experiential transdiagnostic mechanisms that remain underrepresented within formal diagnostic systems.

Despite clear differences between the Reddit network and the ICD-based network of diagnostic criteria, both point to the difficulty of fitting mental disorders into rigid, discrete categories [76,77]. While neither network should be treated as a ground truth of interdisorder relationships, the interconnected structure observed in both aligns with longstanding critiques that current psychiatric nosology underestimates the interconnected nature of psychopathology [78-81] and supports discussions of alternative paradigms, such as dimensional models that emphasize broad spectra and shared underlying features [82].

Limitations and Future Directions

While our findings provide a large-scale view of the interconnected structure of psychopathology through mental health communities, several limitations should also be acknowledged. Our analysis is based on Reddit, a platform whose user base skews toward younger, Western, male, and digitally literate populations, which limits generalizability [83,84]. Reddit-specific dynamics such as anonymity, community norms, and moderation practices also shape what is shared and who participates. These features increase accessibility and ease of self-disclosure, but they also raise validity concerns, including account transience and the use of throwaway accounts that make authenticity difficult to assess [14,85]. Previous work also suggests that individuals with broader lay concepts of disorders are more likely to self-diagnose [86], which may amplify certain demographic biases. While these factors may contribute to divergences between the Reddit-derived and clinical structures, future work should clarify whether they reflect robust disorder associations or platform-specific outcomes. Hence, observed results should not be interpreted as verified comorbidities, but rather as behavioral signals that approximate perceived relatedness between disorders within digital contexts.

The methodological decisions delimit the scope of our results. We analyze posts, but not comments, made in 2022 within mental health support subreddits, and include only communities that map to a distinct disorder category in the *ICD-10*. Focusing on posts aligns with our aim to capture self-disclosure and help-seeking at the point of initiation, though it necessarily excludes peer interaction and information diffusion occurring in comment threads. Restricting to *ICD*-mapped subreddits increases construct clarity but may underrepresent transdiagnostic communities (such as r/MentalHealthSupport) or symptom-focused communities (such as r/SuicideWatch). Limiting the data to 2022 reduces temporal confounding from earlier structural instability, demographic shifts, and disruptions during the COVID-19 pandemic, and it ensures comparability across subreddits within a single, more mature phase of the platform. However, this temporal focus also constrains interpretation: network patterns and discourse are dynamic, and

XSL•FO
RenderX

analyses spanning multiple years could reveal different structures as community composition and cultural context evolve. The year 2022 was chosen to provide a stable and interpretable baseline, but not to imply that the resulting associations are fixed over time.

Looking ahead, several promising directions emerge for future research. Reddit data offers a powerful proxy for examining mental health from multiple perspectives. Incorporating replies would add a complementary interaction layer, as comment threads reveal who engages with whom, what types of support are exchanged (eg, validation or advice), and how these interactions relate to subsequent posting trajectories. Temporal analyses could further enrich this view on 2 fronts. At the societal level, they would capture how mental health concepts and discourse evolve with cultural and technological change. At the individual level, following users over time could help distinguish between comorbidity, diagnostic progression, and broader help-seeking patterns, adding precision to how disorder associations are interpreted. Extending this approach beyond Reddit to platforms with different affordances and user bases would test how platform design shapes the organization of mental health discourse. We chose *ICD-10* as the reference framework for its international coverage and compatibility with registry data. However, comparable analyses using *DSM-5* (Diagnostic and Statistical Manual of Mental Disorders [Fifth Edition]), the forthcoming *ICD-11*, and other evolving diagnostic systems will be essential to assess psychiatry's continuing efforts toward more coherent and empirically grounded concepts of mental health [11]. Such extensions could further reinforce the value of online mental health data as a bridge between peer discourse and clinical knowledge.

Conclusions

This work maps a large-scale structure of mental health communities as they currently grow outside clinical settings, highlighting the necessity of perspectives that extend beyond formal diagnostic frameworks to achieve a more complete population-level understanding of psychopathology. Diagnostic frameworks remain essential, but they capture only part of how distress is articulated and managed in practice. Outside formal care, people navigate symptoms and negotiate meaning while seeking peer communities that mediate their mental health challenges. Digital platforms such as Reddit have become central to this process: they provide spaces for disclosure and support while also shaping the categories, language, and norms through which psychological distress is understood. In this sense, they are not only mirrors of cultural shifts but also infrastructures that reorganize how mental health is lived and discussed. Neglecting these platforms as legitimate sites of knowledge risks leaving research and practice poorly aligned with mental health needs amid rapid technological change.

Acknowledgments

BE was responsible for data collection, experimental design, analysis, and writing the manuscript. SL and PKN contributed to the initial study conceptualization and provided critical feedback during manuscript revisions. SL assisted in validating the *ICD-10* annotations. All authors reviewed and approved the final version of the manuscript.

Generative artificial intelligence (ChatGPT 5; OpenAI) was used to support linguistic refinement and coherence in the manuscript text. All conceptual contributions, interpretations, and substantive arguments presented in this manuscript are solely those of the authors.

Funding

PKN acknowledges partial funding by the research program Knowledge Technologies (P2-0103). This work is the result of research conducted at Central European University, a private university, with open-access provided through the CEU Open Access Fund.

Data Availability

Subreddit metadata and aggregated data supporting the findings of this study are provided in the corresponding Multimedia Appendices. Due to the platform's terms of service, raw Reddit post content cannot be publicly shared. Additional aggregated data may be made available upon reasonable request by contacting the corresponding author.

Conflicts of Interest

None declared.

Multimedia Appendix 1

(Top left) Distribution of the number of self-focused pronouns (I, me, myself) and other-focused pronouns used per post in the selected subreddits, capped at 20 occurrences. The density on the y-axis represents the relative proportion of posts at each pronoun count, rather than raw counts. Posts tend to include more self-focused pronouns (blue) compared to other-focused pronouns (red). (Top right) Overall proportion of posts showing different pronoun balance patterns: more self-focused pronouns (Self > Other), more other-focused pronouns (Other > Self), or balanced/none (Equal/Zero). The majority of posts are self-focused. (Bottom) Pronoun balance patterns stratified by ICD-10 diagnostic categories (Level 2). Each bar represents the proportion of posts within

XSL•FO
RenderX

a diagnostic category that are more self-focused, more other-focused, or balanced. Across nearly all categories, self-focused pronouns dominate, but the relative proportions vary slightly between diagnostic groups.
[PNG File , 490 KB - jmir_v28i1e80958_app1.png]

Multimedia Appendix 2
Supplementary Tables (.xlsx): - ICD-10 Disorder Reference List - Mapping of Subreddits to ICD-10 and ICD-11 - List of (Weighted) Degrees (The Reddit Network of Psychopathology, Positive Associations) - List of Negative Edges and their Weights (Reddit Network of Psychopathology, Negative Associations). - ICD-10 Diagnostic Criteria.
[XLSX File (Microsoft Excel File), 115 KB - jmir_v28i1e80958_app2.xlsx]

Multimedia Appendix 3
Number of posts and subreddits across ICD-10 (International Classification of Diseases, 10th Revision) Chapter F mental health–related categories.
[PNG File , 73 KB - jmir_v28i1e80958_app3.png]

Multimedia Appendix 4
Online Interactive Tool - Map of Associations: Screenshot of the interactive Retina interface showing the node F32–F33 Depression (episodic and recurrent) selected as an example. The left-hand panel displays node-level metrics, including the number of users who posted in the corresponding subreddit(s), the weighted degree (sum of edge weights), the clustering coefficient (indicating local cohesiveness), and the number of triangles (three-node loops) formed around the node. This illustrates how the interface allows for exploratory analysis of individual disorder communities within the broader network structure.
[PNG File , 72 KB - jmir_v28i1e80958_app4.png]

Multimedia Appendix 5
Online Interactive Tool (Map of Disorder Associations): An interactive web-based visualization of the inferred disorder association network, allowing users to explore positive and negative associations, node connectivity, and diagnostic groupings derived from Reddit coposting activity.
[DOCX File , 89 KB - jmir_v28i1e80958_app5.docx]

Multimedia Appendix 6
No evidence of correlation between the number of users active in relation to a specific disorder (node size) and the number of links (node degree) in the Reddit network of psychopathology associations. The result supports the robustness of the inference method, suggesting that node centrality is not driven by subreddit size. Spearman correlation: r=0.068, p=0.64.
[PNG File , 79 KB - jmir_v28i1e80958_app6.png]

Multimedia Appendix 7 [PNG File , 74 KB - jmir_v28i1e80958_app7.png]

References

1. Twenge JM, Gentile B, DeWall CN, Ma D, Lacefield K, Schurtz DR. Birth cohort increases in psychopathology among young Americans, 1938-2007: a cross-temporal meta-analysis of the MMPI. Clin Psychol Rev 2010;30(2):145-154. [doi: 10.1016/j.cpr.2009.10.005] [Medline: 19945203]
2. Twenge JM, Joiner TE, Rogers ML, Martin GN. Increases in depressive symptoms, suicide-related outcomes, and suicide rates among U.S. adolescents after 2010 and links to increased new media screen time. Clinical Psychological Science 2017;6(1):3-17. [doi: 10.1177/2167702617723376]
3. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
4. Sacco R, Camilleri N, Eberhardt J, Umla-Runge K, Newbury-Birch D. A systematic review and meta-analysis on the prevalence of mental disorders among children and adolescents in Europe. Eur Child Adolesc Psychiatry 2024;33(9):2877-2894 [FREE Full text] [doi: 10.1007/s00787-022-02131-2] [Medline: 36581685]
5. Hawes MT, Szenczy AK, Klein DN, Hajcak G, Nelson BD. Increases in depression and anxiety symptoms in adolescents and young adults during the COVID-19 pandemic. Psychol Med 2022;52(14):3222-3230 [FREE Full text] [doi: 10.1017/S0033291720005358] [Medline: 33436120]
6. Hyman SE. The diagnosis of mental disorders: the problem of reification. Annu Rev Clin Psychol 2010;6:155-179. [doi: 10.1146/annurev.clinpsy.3.022806.091532] [Medline: 17716032]
7. Wakefield JC. Klerman's "credo" reconsidered: neo-Kraepelinianism, Spitzer's views, and what we can learn from the past. World Psychiatry 2022;21(1):4-25 [FREE Full text] [doi: 10.1002/wps.20942] [Medline: 35015356]

XSL•FO
RenderX

8. Kendell R, Jablensky A. Distinguishing between the validity and utility of psychiatric diagnoses. Am J Psychiatry 2003;160(1):4-12. [doi: 10.1176/appi.ajp.160.1.4] [Medline: 12505793]
9. Morris SE, Sanislow CA, Pacheco J, Vaidyanathan U, Gordon JA, Cuthbert BN. Revisiting the seven pillars of RDoC. BMC Med 2022;20(1):220 [FREE Full text] [doi: 10.1186/s12916-022-02414-0] [Medline: 35768815]
10. Kotov R, Krueger RF, Watson D, Achenbach TM, Althoff RR, Bagby RM, et al. The hierarchical taxonomy of psychopathology (HiTOP): a dimensional alternative to traditional nosologies. J Abnorm Psychol 2017;126(4):454-477. [doi: 10.1037/abn0000258] [Medline: 28333488]
11. Zachar P. A Metaphysics of Psychopathology. Cambridge, Massachusetts: MIT Press; 2014.
12. Naslund JA, Aschbrenner KA, Marsch LA, Bartels SJ. The future of mental health care: peer-to-peer support and social media. Epidemiol Psychiatr Sci 2016;25(2):113-122 [FREE Full text] [doi: 10.1017/S2045796015001067] [Medline: 26744309]
13. Rayland A, Andrews J. From social network to peer support network: opportunities to explore mechanisms of online peer support for mental health. JMIR Ment Health 2023;10:e41855 [FREE Full text] [doi: 10.2196/41855] [Medline: 36853738]
14. De Choudhury M, De S. Mental health discourse on reddit: self-disclosure, social support, and anonymity. ICWSM 2014;8(1):71-80. [doi: 10.1609/icwsm.v8i1.14526]
15. Kirkbride JB, Anglin DM, Colman I, Dykxhoorn J, Jones PB, Patalay P, et al. The social determinants of mental health and disorder: evidence, prevention and recommendations. World Psychiatry 2024;23(1):58-90 [FREE Full text] [doi: 10.1002/wps.21160] [Medline: 38214615]
16. Dickson SJ, Bussey K, Kangas M, Grocott S, Rapee RM. Barriers to accessing and engaging with mental health services for low-income families in Australia: a qualitative evaluation. J Child Fam Stud 2025;34(11):2862-2877. [doi: 10.1007/s10826-025-03172-2]
17. Lowther-Payne HJ, Ushakova A, Beckwith A, Liberty C, Edge R, Lobban F. Understanding inequalities in access to adult mental health services in the UK: a systematic mapping review. BMC Health Serv Res 2023;23(1):1042 [FREE Full text] [doi: 10.1186/s12913-023-10030-8] [Medline: 37773154]
18. Montag C, Duke É, Markowetz A. Toward psychoinformatics: computer science meets psychology. Comput Math Methods Med 2016;2016:2983685 [FREE Full text] [doi: 10.1155/2016/2983685] [Medline: 27403204]
19. Conway M, O'Connor D. Social media, big data, and mental health: current advances and ethical implications. Curr Opin Psychol 2016;9:77-82 [FREE Full text] [doi: 10.1016/j.copsyc.2016.01.004] [Medline: 27042689]
20. Chancellor S, De Choudhury M. Methods in predictive techniques for mental health status on social media: a critical review. NPJ Digit Med 2020;3:43 [FREE Full text] [doi: 10.1038/s41746-020-0233-7] [Medline: 32219184]
21. Feldhege J, Moessner M, Bauer S. Who says what? Content and participation characteristics in an online depression community. J Affect Disord 2020;263:521-527. [doi: 10.1016/j.jad.2019.11.007] [Medline: 31780138]
22. Low DM, Rumker L, Talkar T, Torous J, Cecchi G, Ghosh SS. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on Reddit during COVID-19: observational study. J Med Internet Res 2020;22(10):e22635 [FREE Full text] [doi: 10.2196/22635] [Medline: 32936777]
23. Morini V, Sansoni M, Rossetti G, Pedreschi D, Castillo C. Participant behavior and community response in online mental health communities: insights from reddit. Computers in Human Behavior 2025;165:108544. [doi: 10.1016/j.chb.2024.108544]
24. Jin C, Zhu Z. Multimorbidity patterns and early signals of diabetes in online communities. JAMIA Open 2025;8(3):ooaf049. [doi: 10.1093/jamiaopen/ooaf049] [Medline: 40453764]
25. McElroy E, Shevlin M, Murphy J, McBride O. Co-occurring internalizing and externalizing psychopathology in childhood and adolescence: a network approach. Eur Child Adolesc Psychiatry 2018;27(11):1449-1457 [FREE Full text] [doi: 10.1007/s00787-018-1128-x] [Medline: 29520540]
26. Boschloo L, Schoevers RA, van Borkulo CD, Borsboom D, Oldehinkel AJ. The network structure of psychopathology in a community sample of preadolescents. J Abnorm Psychol 2016;125(4):599-606. [doi: 10.1037/abn0000150] [Medline: 27030994]
27. Forbes MK. Reconstructing psychopathology: a data-driven reorganization of the symptoms in DSM-5. Clin Psychol Sci 2023;13(3). [doi: 10.31234/osf.io/7um9a]
28. Plana-Ripoll O, Pedersen CB, Holtz Y, Benros ME, Dalsgaard S, de Jonge P, et al. Exploring comorbidity within mental disorders among a Danish national population. JAMA Psychiatry 2019;76(3):259-270 [FREE Full text] [doi: 10.1001/jamapsychiatry.2018.3658] [Medline: 30649197]
29. Dervić E, Sorger J, Yang L, Leutner M, Kautzky A, Thurner S, et al. Unraveling cradle-to-grave disease trajectories from multilayer comorbidity networks. NPJ Digit Med 2024;7(1):56 [FREE Full text] [doi: 10.1038/s41746-024-01015-w] [Medline: 38454004]
30. Dalgleish T, Black M, Johnston D, Bevan A. Transdiagnostic approaches to mental health problems: current status and future directions. J Consult Clin Psychol 2020;88(3):179-195. [doi: 10.1037/ccp0000482] [Medline: 32068421]
31. Borsboom D, Cramer AOJ. Network analysis: an integrative approach to the structure of psychopathology. Annu Rev Clin Psychol 2013;9:91-121. [doi: 10.1146/annurev-clinpsy-050212-185608] [Medline: 23537483]
32. Borsboom D. A network theory of mental disorders. World Psychiatry 2017;16(1):5-13 [FREE Full text] [doi: 10.1002/wps.20375] [Medline: 28127906]

XSL•FO
RenderX

33. Baumgartner J, Zannettou S, Keegan B, Squire M, Blackburn J. The pushshift reddit dataset. In: Proceedings of the International AAAI Conference on Web and Social Media. 2020 Presented at: Proceedings of the International AAAI Conference on Web and Social Media; June 8-11, 2020; Atlanta, GA p. 830-839. [doi: 10.1609/icwsm.v14i1.7347]
34. List of reddit communities (sorted by users). Reddit. 2025. URL: https://www.reddit.com/best/communities/1/ [accessed 2025-07-14]
35. ICD-10 version: 2019. World Health Organization (WHO). 2019. URL: https://icd.who.int/browse10/2019/en [accessed 2025-07-14]
36. Burtch G, Lee D, Chen Z. The consequences of generative AI for online knowledge communities. Sci Rep 2024;14(1):10413 [FREE Full text] [doi: 10.1038/s41598-024-61221-0] [Medline: 38710885]
37. Møller AG, Romero DM, Jurgens D, Aiello LM. The impact of generative AI on social media: an experimental study. arXiv 2025. [doi: 10.48550/arXiv.2506.14295]
38. Home. Bot Rank. URL: https://botrank.com [accessed 2025-07-14]
39. Neal Z. The backbone of bipartite projections: inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors. Social Networks 2014;39:84-97. [doi: 10.1016/j.socnet.2014.06.001]
40. Saracco F, Straka MJ, Clemente RD, Gabrielli A, Caldarelli G, Squartini T. Inferring monopartite projections of bipartite networks: an entropy-based approach. New J. Phys 2017;19(5):053022. [doi: 10.1088/1367-2630/aa6b38]
41. Shaffer JP. Multiple hypothesis testing. Annu Rev Psychol 2024;46(1):561-584. [doi: 10.1146/annurev.psych.46.1.561]
42. Reddit Network of Psychopathology - Positive Associations (Interactive). Retina. URL: https://ouestware.gitlab.io/retina/beta/#/graph/?url=https%3A%2F%2Fgist.githubusercontent.com%2Fboevkoski%2F2903ebeb33e7a5462cd6597c50579033%2Fraw%2Ff2193e5b5343fe33751cd2df80cd3bd78d889%2Freddit_psychopathology_positive_associations.gexf [accessed 2025-07-14]
43. Reddit Network of Psychopathology - Negative Associations (Interactive). Retina. URL: https://ouestware.gitlab.io/retina/beta/#/graph/?url=https%3A%2F%2Fgist.githubusercontent.com%2Fboevkoski%2F4b35e1816c07076a1aa80ad03a4e1a2b%2Fraw%2F2d122d931cfe6944c4d9c51713d7357a6d44e47%2Freddit_psychopathology_negative_associations.gexf [accessed 2025-07-14]
44. Network of psychopathology based on ICD-10 diagnostic criteria (Interactive). Retina. URL: https://ouestware.gitlab.io/retina/beta/#/graph/?url=https://gist.githubusercontent.com/boevkoski/4423ac95663e168e2d355b12d96a7a6f/raw/feee54a83186335facaf89c5133f6673a95ed6b8/ICD10_psychopathology_diagnostic_criteria_overlaps.gexf [accessed 2025-07-14]
45. Tio P, Epskamp S, Noordhof A, Borsboom D. Mapping the manuals of madness: comparing the ICD-10 and DSM-IV-TR using a network approach. Int J Methods Psychiatr Res 2016;25(4):267-276 [FREE Full text] [doi: 10.1002/mpr.1503] [Medline: 27028040]
46. The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines. Geneva, Switzerland: World Health Organization; 1992.
47. Yim O, Ramdeen KT. Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. TQMP 2015;11(1):8-21. [doi: 10.20982/tqmp.11.1.p008]
48. Mammone N, Ieracitano C, Adeli H, Bramanti A, Morabito FC. Permutation jaccard distance-based hierarchical clustering to estimate EEG network density modifications in MCI subjects. IEEE Trans Neural Netw Learn Syst 2018;29:5122-5135. [doi: 10.1109/TNNLS.2018.2791644] [Medline: 29994428]
49. Liu X, Zhu XH, Qiu P, Chen W. A correlation-matrix-based hierarchical clustering method for functional connectivity analysis. J Neurosci Methods 2012;211(1):94-102 [FREE Full text] [doi: 10.1016/j.jneumeth.2012.08.016] [Medline: 22939920]
50. Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview. WIREs Data Min & Knowl 2011;2(1):86-97. [doi: 10.1002/widm.53]
51. Newman MEJ. Modularity and community structure in networks. Proc Natl Acad Sci U S A 2006;103(23):8577-8582 [FREE Full text] [doi: 10.1073/pnas.0601602103] [Medline: 16723398]
52. Wagner S, Wagner D. Comparing clusterings - an overview. University of Science and Technology of China (USTC). URL: http://staff.ustc.edu.cn/~zwp/teach/MVA/cluster_validation.pdf [accessed 2007-01-12]
53. King SA. Researching internet communities: proposed ethical guidelines for the reporting of results. The Information Society 1996;12(2):119-128. [doi: 10.1080/713856145]
54. Eysenbach G, Till JE. Ethical issues in qualitative research on internet communities. BMJ 2001;323(7321):1103-1105 [FREE Full text] [doi: 10.1136/bmj.323.7321.1103] [Medline: 11701577]
55. Moreno MA, Goniu N, Moreno PS, Diekema D. Ethics of social media research: common concerns and practical considerations. Cyberpsychol Behav Soc Netw 2013;16(9):708-713 [FREE Full text] [doi: 10.1089/cyber.2012.0334] [Medline: 23679571]
56. Chancellor S, Birnbaum M, Caine E, Silenzio V, De CM. A taxonomy of ethical tensions in inferring mental health states from social media. 2019 Presented at: FAT* '19: Conference on Fairness, Accountability, and Transparency; January 29-31, 2019; Atlanta, GA p. 79-88. [doi: 10.1145/3287560.3287587]

XSL•FO
RenderX

57. Tri-Council Policy Statement: ethical conduct for research involving humans. Government of Canada. 2022. URL: https://ethics.gc.ca/eng/policy-politique_tcps2-eptc2_2022.html [accessed 2025-12-23]
58. Kosten TR, George TP. The neurobiology of opioid dependence: implications for treatment. Sci Pract Perspect 2002;1(1):13-20 [FREE Full text] [doi: 10.1151/spp021113] [Medline: 18567959]
59. Martins SS, Fenton MC, Keyes KM, Blanco C, Zhu H, Storr CL. Mood and anxiety disorders and their association with non-medical prescription opioid use and prescription opioid-use disorder: longitudinal evidence from the National Epidemiologic Study on Alcohol and Related Conditions. Psychol Med 2012;42(6):1261-1272 [FREE Full text] [doi: 10.1017/S0033291711002145] [Medline: 21999943]
60. McLaughlin KA, Colich NL, Rodman AM, Weissman DG. Mechanisms linking childhood trauma exposure and psychopathology: a transdiagnostic model of risk and resilience. BMC Med 2020;18(1):96 [FREE Full text] [doi: 10.1186/s12916-020-01561-6] [Medline: 32238167]
61. Ellickson-Larew S, Stasik-O'Brien SM, Stanton K, Watson D. Dissociation as a multidimensional transdiagnostic symptom. PConsci-TRP 2020;7(2):126-150. [doi: 10.1037/cns0000218]
62. Wu JB, Yang Y, Zhou Q, Li J, Yang WK, Yin X, et al. The relationship between screen time, screen content for children aged 1-3, and the risk of ADHD in preschools. PLoS One 2025;20(4):e0312654 [FREE Full text] [doi: 10.1371/journal.pone.0312654] [Medline: 40267918]
63. Fekih-Romdhane F, Jahrami H, Away R, Trabelsi K, Pandi-Perumal SR, Seeman MV, et al. The relationship between technology addictions and schizotypal traits: mediating roles of depression, anxiety, and stress. BMC Psychiatry 2023;23(1):67 [FREE Full text] [doi: 10.1186/s12888-023-04563-9] [Medline: 36698079]
64. Dong HY, Wang B, Li HH, Yue XJ, Jia FY. Correlation between screen time and autistic symptoms as well as development quotients in children with autism spectrum disorder. Front Psychiatry 2021;12:619994 [FREE Full text] [doi: 10.3389/fpsyt.2021.619994] [Medline: 33664683]
65. Borsboom D, Cramer AOJ, Schmittmann VD, Epskamp S, Waldorp LJ. The small world of psychopathology. PLoS One 2011;6(11):e27407 [FREE Full text] [doi: 10.1371/journal.pone.0027407] [Medline: 22114671]
66. Grant BF, Goldstein RB, Saha TD, Chou SP, Jung J, Zhang H, et al. Epidemiology of DSM-5 alcohol use disorder: results from the National Epidemiologic Survey on Alcohol and Related Conditions III. JAMA Psychiatry 2015;72(8):757-766 [FREE Full text] [doi: 10.1001/jamapsychiatry.2015.0584] [Medline: 26039070]
67. Kessler RC, Nelson CB, McGonagle KA, Edlund MJ, Frank RG, Leaf PJ. The epidemiology of co-occurring addictive and mental disorders: implications for prevention and service utilization. Am J Orthopsychiatry 1996;66(1):17-31 [FREE Full text] [doi: 10.1037/h0080151] [Medline: 8720638]
68. Corrigan PW, Kuwabara S, O'Shaughnessy J. The public stigma of mental illness and drug addiction. J Soc Work 2009;9(2):139-147. [doi: 10.1177/1468017308101818]
69. Room R. Stigma, social inequality and alcohol and drug use. Drug Alcohol Rev 2005;24(2):143-155. [doi: 10.1080/09595230500102434] [Medline: 16076584]
70. Hing N, Russell AMT, Gainsbury SM, Nuske E. The public stigma of problem gambling: its nature and relative intensity compared to other health conditions. J Gambl Stud 2016;32(3):847-864 [FREE Full text] [doi: 10.1007/s10899-015-9580-8] [Medline: 26487344]
71. Hogg B, Gardoki-Souto I, Valiente-Gómez A, Rosa AR, Fortea L, Radua J, et al. Psychological trauma as a transdiagnostic risk factor for mental disorder: an umbrella meta-analysis. Eur Arch Psychiatry Clin Neurosci 2023;273(2):397-410. [doi: 10.1007/s00406-022-01495-5] [Medline: 36208317]
72. Evkoski B, Letina S, Kralj Novak P, Riddell J. Premenstrual dysphoric disorder in online peer support communities: a Reddit case study. Sci Rep 2025;15(1):34300 [FREE Full text] [doi: 10.1038/s41598-025-19220-2] [Medline: 41034480]
73. Forbes MK, Neo B, Nezami OM, Fried EI, Faure K, Michelsen B, et al. Elemental psychopathology: distilling constituent symptoms and patterns of repetition in the diagnostic criteria of the DSM-5. Psychol Med 2024;54(5):886-894. [doi: 10.1017/S0033291723002544] [Medline: 37665038]
74. Underhill R, Foulkes L. Self-diagnosis of mental disorders: a qualitative study of attitudes on Reddit. Qual Health Res 2025;35(7):779-792 [FREE Full text] [doi: 10.1177/10497323241288785] [Medline: 39422576]
75. Chan GJ, Fung M, Warrington J, Nowak SA. Understanding health-related discussions on reddit: development of a topic assignment method and exploratory analysis. JMIR Form Res 2025;9:e55309 [FREE Full text] [doi: 10.2196/55309] [Medline: 39879094]
76. Kendler KS. The nature of psychiatric disorders. World Psychiatry 2016;15(1):5-12 [FREE Full text] [doi: 10.1002/wps.20292] [Medline: 26833596]
77. Zachar P. Psychological Concepts and Biological Psychiatry: A Philosophical Analysis. Amsterdam, Netherlands: John Benjamins Publishing Company; 2000.
78. Zachar P, Kendler KS. Psychiatric disorders: a conceptual taxonomy. Am J Psychiatry 2007;164(4):557-565. [doi: 10.1176/ajp.2007.164.4.557] [Medline: 17403967]
79. Lahey BB, Tiemeier H, Krueger RF. Seven reasons why binary diagnostic categories should be replaced with empirically sounder and less stigmatizing dimensions. JCPP Adv 2022;2(4):e12108 [FREE Full text] [doi: 10.1002/jcv2.12108] [Medline: 37431412]

XSL•FO
RenderX

80. McGorry PD, Hickie IB, Kotov R, Schmaal L, Wood SJ, Allan SM, et al. New diagnosis in psychiatry: beyond heuristics. Psychol Med 2025;55:e26. [doi: 10.1017/S003329172400223X] [Medline: 39911018]
81. Ringwald WR, Forbes MK, Wright AGC. Meta-analysis of structural evidence for the Hierarchical Taxonomy of Psychopathology (HiTOP) model. Psychol Med 2023;53(2):533-546. [doi: 10.1017/S0033291721001902] [Medline: 33988108]
82. Ruggero CJ, Kotov R, Hopwood CJ, First M, Clark LA, Skodol AE, et al. Integrating the Hierarchical Taxonomy of Psychopathology (HiTOP) into clinical practice. J Consult Clin Psychol 2019;87(12):1069-1084 [FREE Full text] [doi: 10.1037/ccp0000452] [Medline: 31724426]
83. Finlay SC. Age and gender in Reddit commenting and success. J Inf Sci Theory Pract 2014;2(3):18-28. [doi: 10.1633/jistap.2014.2.3.2]
84. Reddit users by country. World Population Review. 2025. URL: https://worldpopulationreview.com/country-rankings/reddit-users-by-country [accessed 2025-07-14]
85. Proferes N, Jones N, Gilbert S, Fiesler C, Zimmer M. Studying Reddit: a systematic overview of disciplines, approaches, methods, and ethics. Social Media + Society 2021;7(2):205630512110190. [doi: 10.1177/20563051211019004]
86. Tse JSY, Haslam N. Broad concepts of mental disorder predict self-diagnosis. SSM - Mental Health 2024;6:100326. [doi: 10.1016/j.ssmmh.2024.100326]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
ARI: Adjusted Rand Index
DSM: Diagnostic and Statistical Manual of Mental Disorders
DSM-5: Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition)
ICD: International Classification of Diseases
ICD-10: International Statistical Classification of Diseases, 10th Revision
ICD-11: International Classification of Diseases, 11th Revision
NMI: normalized mutual information
PTSD: posttraumatic stress disorder
TCPS 2: Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans

Edited by A Mavragani; submitted 19.Jul.2025; peer-reviewed by J Kang, V Morini, Z Zhu, X Liang, A Markovits; comments to author 02.Sep.2025; accepted 02.Dec.2025; published 30.Jan.2026.

Please cite as:
Evkoski B, Letina S, Kralj Novak P
The Structure of Psychopathology on Reddit: Network Analysis of Mental Health Communities in Relation to the ICD Diagnostic System
J Med Internet Res 2026;28:e80958
URL: https://www.jmir.org/2026/1/e80958
doi:10.2196/80958
PMID:

©Bojan Evkoski, Srebrenka Letina, Petra Kralj Novak. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Challenges of Standard Pediatric Epilepsy Monitoring and the Potential Benefits of Contactless Sensor Technologies: Exploratory Qualitative Study

Natalie Öhl[1], MA; Dilbar Strack[2], Dr med; Stefan Martin Kusnik[2], Dr med; Isabell Link[1], Dr rer biol hum; Marlene Hofmann[2], MSc; Tobias Steigleder[1], Dr med; Alina Weisser[1], MSc; Christoph Ostgathe[1], Prof Dr Med; Regina Trollmann[2], Prof Dr Med; Maria Heckel[1], PD, Dr phil, Dr rer biol hum

[1]Department of Palliative Medicine, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Werner-von-Siemens Str. 34, Erlangen, Germany

[2]Department of Pediatric and Adolescent Medicine, Division of Pediatric Neurology and Social Pediatrics, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Corresponding Author:
Natalie Öhl, MA
Department of Palliative Medicine, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Werner-von-Siemens Str. 34, Erlangen, Germany

Abstract

Background: Epilepsy is a common neurological condition in children, and accurate detection of seizures and their frequency is essential for diagnosis and treatment. Standard monitoring using electroencephalography alongside clinical observation is often burdensome in pediatric settings, as electrodes can cause discomfort and restrict mobility. Contactless sensor technologies may offer a promising supplement by enabling monitoring without physical contact.

Objective: This study aims to explore challenges in standard pediatric epilepsy monitoring from the perspective of health care professionals and examines the potential benefits and requirements of supplementary contactless sensor technologies in this setting.

Methods: Participant observation of routine processes in standard pediatric epilepsy monitoring was conducted at a German university hospital. Field notes from 40 observed procedures were analyzed using structuring content analysis. Building on these findings, a focus group with pediatric neurologists, nurses, and medical technical assistants (n=6) explored the potential benefits and implementation requirements of contactless sensor technologies. Focus group data were analyzed using focus group illustration maps.

Results: A reference workflow of standard pediatric epilepsy monitoring was derived, revealing psychosocial, medical, and organizational challenges faced by health care professionals. Electroencephalography recordings and clinical observation required considerable reassurance of patients and parents or carers, were vulnerable to movement artifacts and incomplete seizure documentation, and were labor- and resource-intensive. Focus group participants viewed contactless sensor technologies as a potentially valuable supplement by enabling continuous long-term monitoring with minimal additional burden.

Conclusions: By identifying challenges associated with standard pediatric epilepsy monitoring, this study provides a foundation for the needs-based development and implementation of supplementary contactless sensor technologies. Such technologies should be tailored to the clinical setting and designed to address existing burdens, with the potential to complement standard monitoring.

Trial Registration: Deutsches Register Klinischer Studien (DRKS) DRKS00027017; https://drks.de/search/de/trial/DRKS00027017

(J Med Internet Res 2026;28:e83009) doi:10.2196/83009

KEYWORDS

pediatrics; epilepsy; monitoring; technology; qualitative research

Introduction

Epilepsy is one of the most common neurological conditions in children and is characterized by an enduring risk of epileptic seizures [1]. Detecting seizures and their frequency is essential for diagnosis, treatment, and therapy monitoring [2]. The current gold standard for epilepsy diagnosis and monitoring is electroencephalography (EEG) alongside clinical observation [3,4].

XSL•FO
RenderX

EEG procedures typically take place in clinical settings and range from short-term recordings of approximately 30 minutes to long-term monitoring over several hours [5]. The electrodes attached to the patient's head often lead to discomfort and restricted mobility [5,6]. This is particularly relevant in pediatric populations, where resistance to EEG technology may further complicate the procedure. These challenges place considerable demands on health care professionals and add complexity to clinical workflows. While challenges associated with EEG recordings and long-term monitoring in general have been studied [7-9], there remains a lack of empirical knowledge on the specific challenges of standard pediatric epilepsy monitoring in clinical practice from the perspectives of the professionals involved. Qualitative studies on this topic are scarce, and observational studies remain particularly limited [10-12]. Yet, observational approaches are essential in this context, as they allow direct capture of real-world clinical processes, interaction patterns, and workflow constraints, rather than relying on retrospective self-report [13,14].

Various technologies supporting seizure detection have been developed or are under investigation, including video-based approaches, mattress sensors, and mobile applications [6,7]. These systems monitor physiological signals relevant to epilepsy, such as cardiovascular changes, respiration, electrodermal activity, motor activity, audio, eye movements, temperature changes, body pressure, and moisture [6]. Although many are noninvasive, they are typically contact-based [6] and may be difficult to tolerate for children, who cannot meaningfully weigh potential benefits against discomfort and may instinctively try to remove them. Therefore, contactless sensor technologies may be particularly promising in pediatric epilepsy monitoring [6] by reducing discomfort and improving adherence.

One example of a contactless approach is a new type of radar technology based on interferometry [15], currently being developed within federally funded research projects, such as BrainEpP (BMBF: 13GW0295D) and EmpkinS (DFG: CRC 1483 - 442419336) [16-18]. Positioned near the child's chest, for example, under the bed, the radar continuously records the patient's heartbeat [19-21]. Changes in heart rate and heart rate variability (HRV) are known to occur before and during epileptic seizures [22]. Machine-learning algorithms derive these parameters from radar motion signals [23] and use this information to support seizure detection. In adult populations, the heart rate calculated on the basis of radar data is of high accuracy when compared to the electrocardiogram [16,24,25].

Despite growing interest in contactless seizure monitoring, empirical evidence on how health care professionals perceive their benefits, limitations, and integration into pediatric clinical practice remains limited. Their perspectives are essential to ensure usability, feasibility, and meaningful clinical benefit [6].

To deliver clinical value, innovative technologies such as contactless sensors should be compatible with real-world clinical workflow and constraints [26]. This requires first identifying challenges in current pediatric epilepsy monitoring practices and, building on these insights, understanding how health care professionals envision technological support addressing these challenges. Such a sequential and integrated approach is relevant for the responsible development and implementation of new technologies [27].

Accordingly, this study addresses two research gaps: (1) it explores challenges in standard pediatric epilepsy monitoring at a university hospital epilepsy center from the perspective of health care professionals, to identify opportunities for workflow improvements and technological support; and (2) based on these insights, it examines the potential benefits and requirements of supplementary contactless sensor technologies in this clinical setting from the perspective of health care professionals.

Methods

Study Design

We used a qualitative design, informed by a constructivist epistemological approach, drawing on principles of ethnography to understand the complex social and organizational processes in pediatric epilepsy monitoring. Between November 2021 and January 2022, we conducted ethnographic fieldwork [13,28] using participant observation [13,14] to examine routine processes in standard pediatric epilepsy monitoring. This method allowed us to capture real-world workflows, interactions, and challenges in context. Building on these observations, we conducted a focus group [29] with health care professionals in June 2022 to explore the potential benefits and requirements of supplementary contactless sensor technologies (Figure 1). The study process adhered to COREQ (Consolidated Criteria for Reporting Qualitative Studies) guidelines (Checklist 1) [30], ensuring transparency and rigor in reporting qualitative research.

XSL•FO
RenderX

Figure 1. Synthesized methods used in study design.

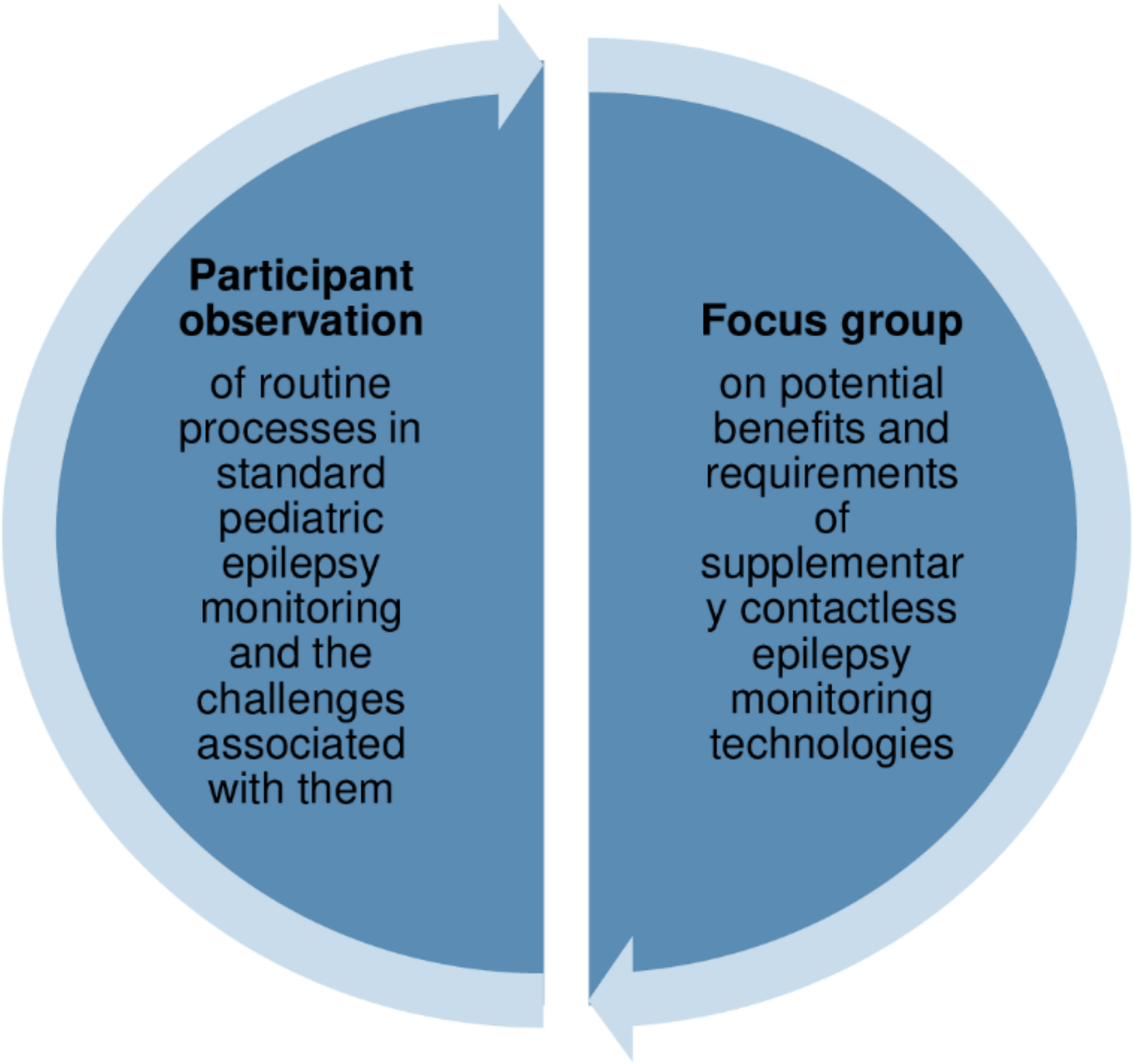

Setting

The data were collected in the pediatric epilepsy center of the Division of Pediatric Neurology at University Hospital Erlangen, Germany. EEG recording (using the 10 - 20 system) adhered to the guidelines issued by the DGKN (German Society for Clinical Neurophysiology and Functional Imaging) [31].

Research Team and Reflexivity

Participant observation and focus group were carried out by IL, a study coordinator, research associate in the Department of Palliative Medicine, and medical technician, in consultation with MaHo, a research associate in the Division of Pediatric Neurology. Prior to conducting the observational study, IL received training in participant observation from an experienced researcher.

Participant Observation

Data Collection

With informed consent, the actions and standard procedures of health care professionals (pediatric neurologists, nurses, and medical technical assistants, MTAs) were observed during routine pediatric epilepsy monitoring. This included diagnostic, long-term, and seizure-detection EEG as well as clinical observation. We chose situations for observation in consultation with the nurse and MTA in charge of the relevant department.

The observer (IL) was not known to the health care professionals beforehand; she introduced herself and gave brief details on the study at the beginning of each situation of interest. The observation included routine EEG recording processes, working environment (structures and organization of the ward; furnishings, equipment, and facilities in patient rooms), type of diagnostic methods, devices, and materials used, and how they were used, roles and responsibilities of the stakeholders involved, communication and collaboration processes,

XSL•FO
RenderX

information provision and decision-making, and duration of the monitoring. A guide detailed the components and the focus of the observation (Multimedia Appendix 1). Throughout the observations, health care professionals explained procedures and provided information about processes and experiences. Besides, conversations that arose spontaneously with health care professionals were also documented in the field notes. Between November 2021 and January 2022 (on 8 days, 4 hours each), a total of 40 situations were observed. Documentation was done with field notes and pictures of the technical facilities, with the pictures showing only the equipment and no children or staff. Data collection continued until saturation was reached in terms of varied and typical pediatric epilepsy monitoring situations.

Data Analysis

On the basis of the field notes, IL created 40 pseudonymized descriptions of the situations and imported them to MAXQDA [32], setting document variables. We used the MAXQDA software (version 20; VERBI Software) for the analysis. Each description comprised details in reference to each component of the observation guide (Multimedia Appendix 1) relating to the care of 1 patient with epilepsy. We ensured interobserver reliability by having multiple researchers independently analyzing the field notes, thereby minimizing potential bias. NO, IL, and MH conducted structured content analysis using content categories as defined by Mayring [33,34]. The units of analysis were determined by the description of situations observed. Coding took place using a subsumption strategy, assigning data to main categories and subcategories. We used the components of the observation guide as deductive main categories and then inductively added further main and subcategories (Table 1).

Table . Coding scheme for descriptions of situations observed.

Main categories	Subcategories (inductive)
Routine processes (deductive)	• Patient admission • Medical information and informed consent • Patient logistics • EEG[a] electrode placement • EEG recording • Post EEG procedures • EEG interpretation • Working routines on the ward
Information sharing (deductive)	• Information sharing between parents or carers or patients and pediatric neurologists • Information sharing between nurses and pediatric neurologists • Documentation
Human-technology interaction (inductive)	• Handling movement artifacts and reapplying EEG electrodes • Setting markers in EEG recordings
Human-human interaction (inductive)	• Instructing and calming patients during EEG electrode placement and recording • Instructing and calming parents or carers during clinical observation
Resource requirements (inductive)	• Human resources • Consumables • Time • Facilities

[a]EEG: electroencephalography.

Coding was done by one coder (NO) (Figure 2). Due to the heterogeneity of the situations observed, the coder carried out deductive coding of 100 percent of situations before adjusting the system of categories and performing inductive coding. Three feedback loops, involving NO, MH, and IL, took place to enable us to attain the maximum possible intersubjectivity and to verify the categories. The adaptation of the coding scheme was an iterative process that went through the material multiple times. To minimize intersubjective differences in interpretation, we drew up coding guidelines ensuring our codes were clearly defined and included anchor examples. A reference workflow emerged on the basis of the main categories and subcategories; we had a pediatric neurologist (DS) check the workflow for correctness and completeness.

XSL•FO
RenderX

Figure 2. Data analysis using the method of structuring content analysis set out by Mayring [33,34].

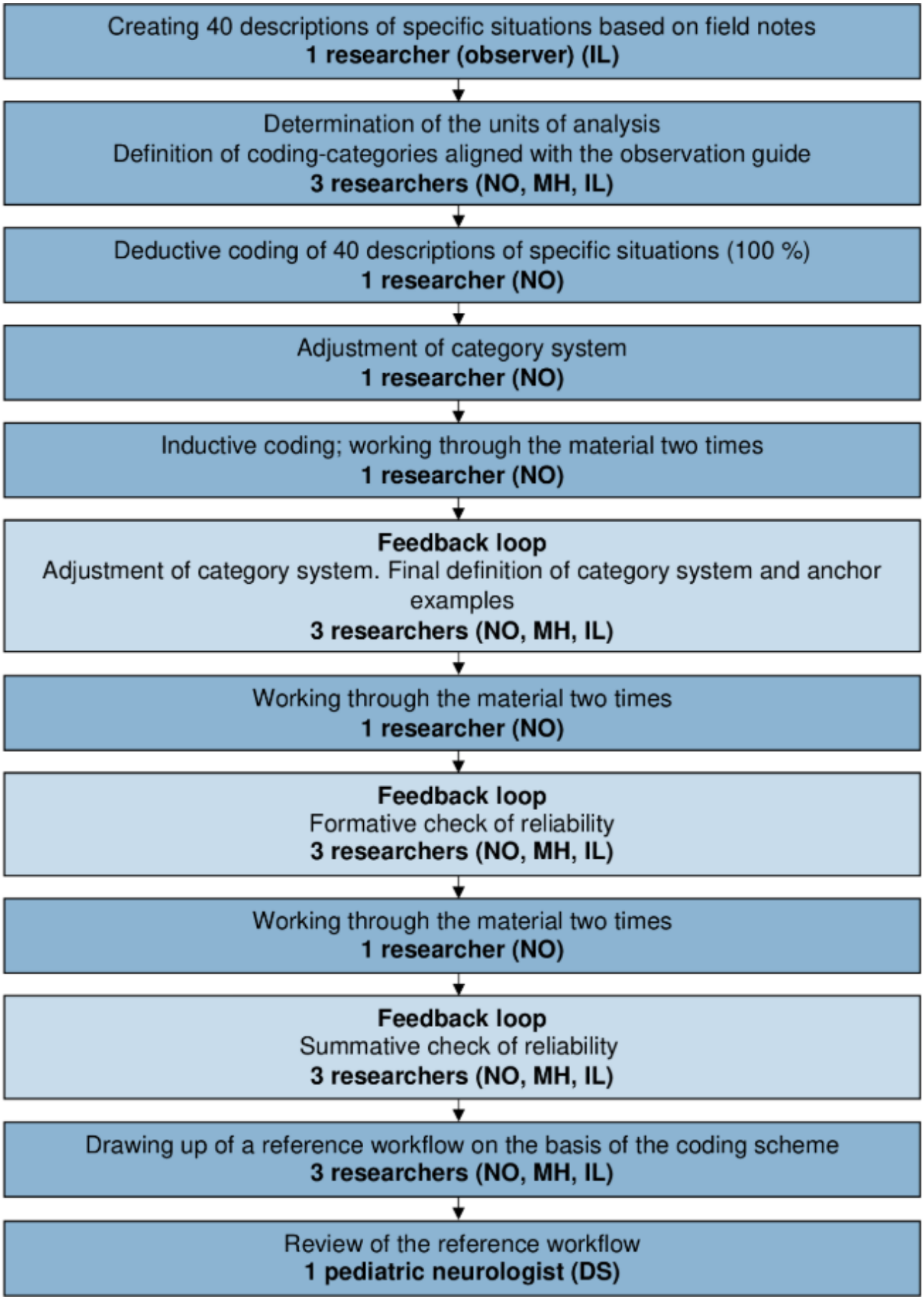

XSL•FO
RenderX

Focus Group

A focus group discussion with health care professionals was performed to explore the potential integration of supplementary contactless sensor technologies into pediatric epilepsy monitoring, including their perceived benefits and requirements.

Recruitment of Participants

Health care professionals of the Division of Pediatric Neurology, University Hospital Erlangen, all directly involved in the care of pediatric patients with epilepsy, were invited. Participants included pediatric neurologists (n=4), a nurse (n=1), and an MTA (n=1), thereby covering all relevant perspectives and offering a sample of the broader population of epilepsy care providers. One of the authors (MaHo) approached prospective participants, whom we had specifically identified as meeting these criteria, face to face; IL provided information to those who showed interest in taking part. All those approached gave written consent to participate. The researcher conducting the focus group (IL) was an external research associate who was fleetingly personally known to the participants.

Data Collection

We created and piloted a focus group guide, informed by preliminary results from the participant observation. The facilitator (IL) received prior training to conduct the focus group. To support the facilitator, MaHo and AW were present during the focus group. The focus group took place in June 2022 at University Hospital Erlangen, with a duration of 60 minutes and was audio-recorded. We printed out key questions in advance of the discussion and made them available to all participants. The focus group commenced with an outline of the research topic and the aims of the research. To support the discussion, we gave a short presentation on an exemplary contactless sensor technology for supporting epilepsy monitoring, involving a flyer detailing the technology. Specifically, we introduced a new type of radar that is currently under development in the research project BrainEpP (BMBF: 13GW0295D).

Data Analysis

We analyzed the focus group using focus group illustration maps, a method that integrates knowledge mapping principles with focus group procedures [35]. Focus group illustration maps combine data collection and analysis in one unified, structured, and documented process.

Creating the Maps During the Focus Group

During the focus group, one researcher (AW) documented participants' statements in real time on a flipchart. Statements were summarized, assigned to thematic clusters, and placed as nodes within a network; connecting lines indicated conceptual links among statements. This graphical documentation supported structuring and condensation of participants' contributions, preserved relational connections between arguments, and enabled immediate consensual validation, as participants could verify, correct, or expand written content on the spot. By the end of the focus group, a structured and participant-validated visual summary existed for each guiding question.

Post-Session Refinement

After the focus group, the flipchart graphics were transferred into Miro Board [36], and digital illustration maps were created for each key question. Two researchers (NO and AW) independently reviewed the maps, comparing them with the audio recordings to ensure completeness, correctness, and logical consistency.

As a final step, 2 researchers (NO and IL) aligned the results of the participant observation with the focus group illustration maps to identify connections between the 2 data sources.

Ethical Considerations

The study was approved by the Ethics Committee of the Faculty of Medicine at Friedrich-Alexander-Universität Erlangen-Nürnberg (350_21 B) on September 20, 2021. The trial was registered with Deutsches Register Klinischer Studien (DRKS; DRKS00027017) on November 16, 2021. All participants received full information on the study and on their right to withdraw consent to participate at any time without giving reasons and were able to ask questions prior to the participant observation and the focus group. All participants gave informed consent to participate in the study. All data were pseudonymized prior to analysis, ensuring that no information could be linked back to individuals, and access to the pseudonymized dataset was restricted to authorized members of the research team only. Participants did not receive any form of compensation for their involvement in the study.

Results

Reference Workflow

To contextualize the challenges identified in standard pediatric epilepsy monitoring within a university hospital epilepsy center, we first present a reference workflow that provides an overview of the process. The coding scheme of the participant observation (Table 1) led to this reference workflow (Figure 3). It outlines the clinical stages of pediatric epilepsy monitoring, the stakeholders involved, the work steps, information flows, resources required, and the interactions that characterize routine practice. The process varies depending on the type of EEG (eg, routine vs long-term video EEG procedure) and whether it takes place on an inpatient or outpatient basis. Routine EEGs entail at least 20 minutes of recording, whereas long-term EEGs require approximately 24 hours of recording in the inpatient clinic.

Within this reference workflow, psychosocial, medical, and organizational challenges in standard pediatric epilepsy monitoring procedures encountered by health care professionals were identified (Figure 4). These challenges highlight where workflow improvements or supportive technologies could be most beneficial.

XSL•FO
RenderX

Figure 3. Reference workflow for standard pediatric epilepsy monitoring in a university hospital epilepsy center based on observations of routine processes (own illustration). EEG: electroencephalography.

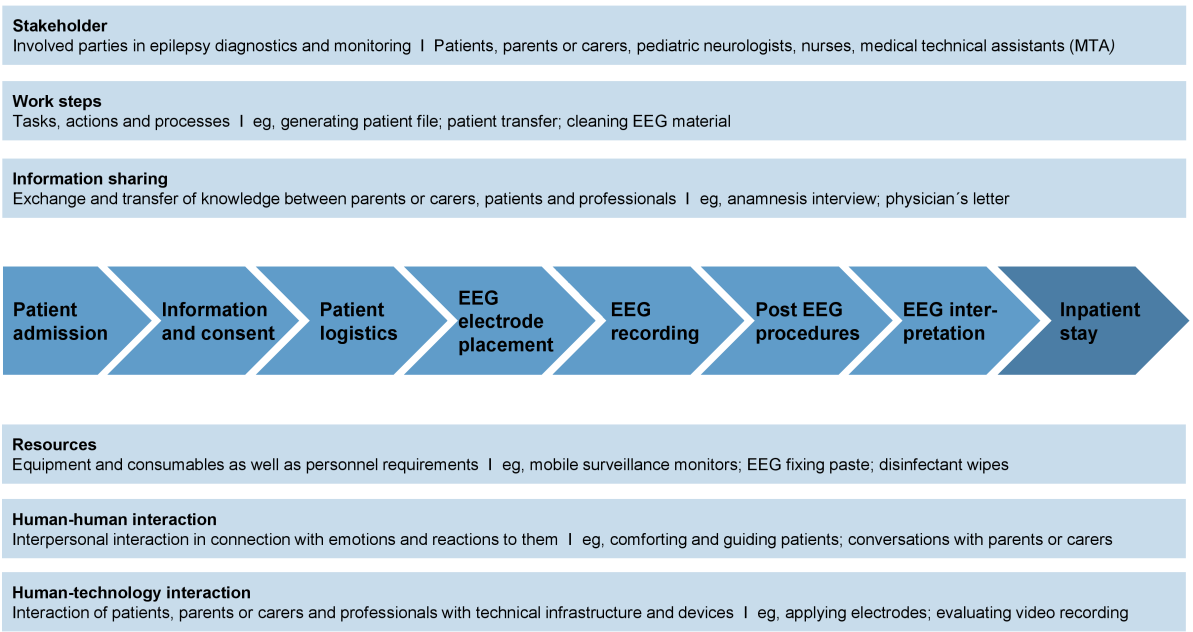

Figure 4. Identified challenges in standard pediatric epilepsy monitoring in a university hospital epilepsy center. EEG: electroencephalography.

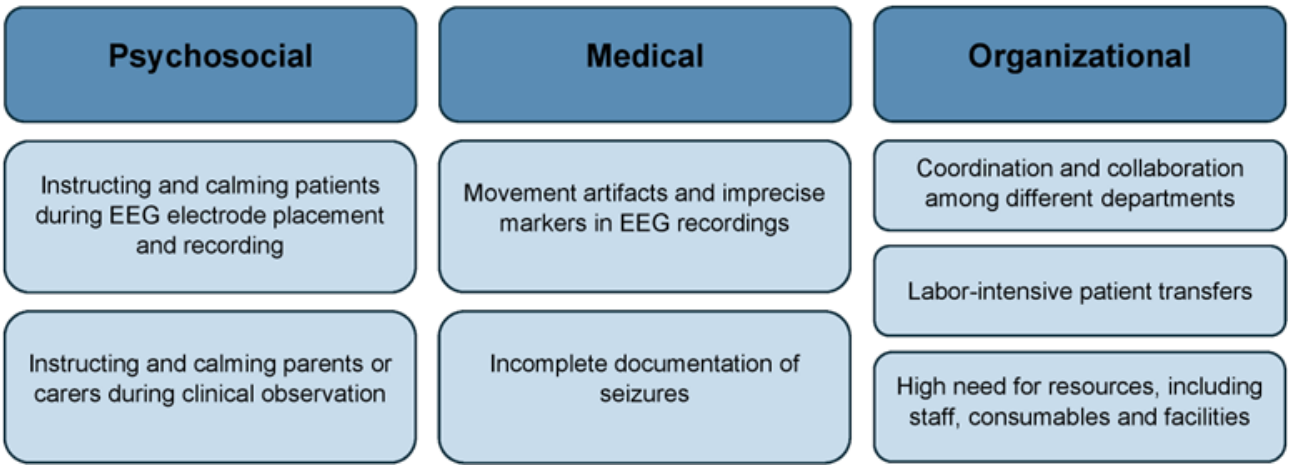

Psychosocial Challenges

Instructing and Calming Patients During EEG Electrode Placement and Recording

During the observation, health care professionals explained that patients must remain still during the application of electrodes and the EEG recording, as physical movement can cause artifacts (see "Medical challenges" section below). While many patients tolerated the equipment well, some showed signs of stress or fear, such as screaming, crying, and irregular breathing:

> *The child starts to wheeze and shows irregular breathing (holding her breath, jerky exhalations). She taps/drums her hands on the examination table cushion.* [S24]

Some resisted the recording technology by shaking their heads or turning away. In several observational situations, MTAs involved in the workflow instructed patients and parents or carers (individuals actively involved in care who have legal parental responsibilities or legal guardianship) to ensure the technology remained in place. Patients with disabilities, limited communication abilities, or cognitive impairments occasionally had difficulties following the MTAs' instructions. MTAs tried to comfort patients by interacting with them empathetically, using elements of play and physical contact.

> *Throughout the examination, the medical technical assistant is slightly bent over the child, her arm rests reassuringly on the child's upper body and she tells a comforting story.* [S22]

MTAs also encouraged parents or carers to calm or distract their children during this process. This was very demanding for

XSL•FO
RenderX

MTAs, who had to cope with children crying and screaming while maintaining patience and empathy.

Instructing and Calming Parents or Carers During Clinical Observation

In addition to EEG recordings, clinical observation of seizure activity is essential. Health care professionals request parents or carers to document seizures characteristic of the patient's epilepsy. During the observation, health care professionals noted the generally high motivation of parents or carers to engage in this process. However, they also acknowledged that it represents an additional responsibility for parents or carers in an already stressful situation. Health care professionals explained that parents or carers often experience pressure to do everything "right," although they are not medically trained, and are acutely aware that their child's diagnosis may depend on the quality of their observations. Health care professionals explained that they face the challenge of calming parents or carers in this demanding situation while ensuring that the documentation is done correctly. This can be further complicated in cases involving language barriers:

> *Due to the limited language proficiency of some non-native parents, they do not fully understand their tasks and have difficulties knowing what and how to document.* [S13]

Medical Challenges

Movement Artifacts and Imprecise Markers in EEG Recordings

During the observation, patients' movements in the course of EEG recordings led to artifacts and thus reduced the recordings' informational value in medical terms.

> *The patient is extremely tense, resulting in an EEG that is heavily affected by artifacts.* [S10]

When patients resisted the procedure, electrodes and probes became detached, fell off, and required reapplication; this interrupted the course of monitoring and reduced the overall time it covered. MTAs sometimes struggled to obtain analyzable recordings. They documented movement artifacts that occurred by placing a marker in the recording; the necessity of paying simultaneous attention to several parallel tasks occasionally resulted in markers being set imprecisely.

Incomplete Documentation of Seizures

The documentation of epileptic seizures by health care professionals and parents or carers is of crucial significance. However, the observation showed that its completeness was occasionally limited due to the difficulties that even specialists experience in explaining different types of seizures to laypeople. Health care professionals explained that parents or carers lack professional knowledge, which impedes correct documentation. Language barriers were an additional source of misunderstanding. The information loss may impact treatment:

> *Information on seizures is completely missing in the template and the further diagnosis has to be made without this additional information.* [S13]

Health care professionals explained that specialized epilepsy centers typically video-record long-term EEGs, which enables pediatric neurologists to examine the video material when parents or carers have documented a seizure incompletely. However, it is extremely time-consuming, and not all seizure patterns can be identified.

Organizational Challenges

The observation showed that epilepsy monitoring in a pediatric epilepsy center is a labor- and resource-intensive procedure that makes heavy demands on the management of procedures and routines, including appointments and patient transport for EEG recordings.

Coordination and Collaboration Among Different Departments

Coordination of appointments for inpatients took place via telephone conversations with the ward. When the ward was busy or an emergency arose, delays could cause inpatients to arrive late. If a patient was late and the monitoring room had been prepared specifically for this patient, it was usually not possible to fill the gap with another patient; the delay caused delays to other patients and occasionally resulted in postponed appointments. This dynamic situation required staff to respond flexibly and make rapid decisions on the order of patients and who to prioritize.

> *An MTA repeatedly calls the ward and asks whether the patient has already been admitted and explains that all preparations have been made (...). Referring to the digital schedule of appointments, two MTAs begin to discuss which patients' appointments could be moved forward during the waiting period.* [S15]

Labor-Intensive Patient Transfers

The clinical processes vary based on which EEG is required (routine or long-term). Health care professionals explained that, depending on the process used for EEG electrode placement and recording, patients need to be transferred between the pediatric neurology ward, the outpatient clinic, and, on occasion, the EEG monitoring room.

> *Patient transport depends on the patient's health status, prior experience, and age, and may require one or two medical assistants, or in special cases with high seizure frequency a licensed nurse.* [S34]

Patient transfers took up a substantial amount of time, reducing flexibility in patient management and contributing to delays.

High Need for Resources, Including Staff, Consumables, and Facilities

The process of standard epilepsy monitoring engages a large amount of resources, including staff, consumables, and facilities. The observation showed that the preparation and application of EEGs were particularly labor-intensive, depending on the patient's ability to cooperate. Consumables needed included abrasive paste and cotton swabs to clean the scalp, conductive or fixing paste for attaching the electrodes, and, in some cases, a syringe to insert conductive paste into the electrode caps; general hygiene material will also be required:

XSL•FO
RenderX

> *The medical assistant first cleans the patient's scalp with the abrasive paste and a cotton swab and then attaches the cup electrode to the scalp with the fixing paste.* [S16]

Long-term EEGs required specially equipped monitoring rooms. Transfer of patients from the ward to these rooms required mobile surveillance monitors and mobile blood pressure monitors.

Potential Benefits of Supplementary Epilepsy Monitoring With Contactless Sensor Technologies

In the participant observation, challenges of standard pediatric epilepsy monitoring from the perspectives of health care professionals were identified. In the subsequent focus group, health care professionals discussed the potential benefits and requirements of using contactless sensor technologies for supplementary monitoring.

Overall, participants discussed several aspects that need to be considered when implementing such technology. They emphasized that training is required not only for clinical staff but also for patients and parents or carers. Without proper instruction, additional workload may arise for clinical staff due to questions from parents or carers. Participants suggested that refresher training should be provided after some time, along with a re-evaluation to gather feedback on needs and any problems encountered. There should also be a clear point of contact for issues related to the use of the technology.

> *Training and a clear contact person are essential to avoid extra workload and ensure smooth use.*

The responsibilities of clinical staff need to be clearly defined. Participants noted that implementing this technology requires additional effort, and clinics must allocate sufficient time and resources to support this new task. Depending on the type of contactless technology and the physiological signals being monitored, participants noted that the constraints and technical limitations must be carefully evaluated and considered during implementation.

Continuous Long-Term Epilepsy Monitoring Data

Participants described EEG as a significant, long-established, and irreplaceable method for epilepsy diagnosis. However, contactless sensor technologies were seen as a useful complement to the gold standard of epilepsy diagnosis (EEG and clinical observation), in providing long-term monitoring data. Participants emphasized that such technology can only complement, but not replace, EEG.

Sensors work without requiring direct physical contact with the patient, making monitoring easier and less labor-intensive for staff. Specifically, the radar-based technology introduced might allow monitoring biomarkers such as respiration and HRV. Participants discussed its potential to detect seizures and measure their duration, frequency, and type. They argued that continuous monitoring data over a longer period could support diagnosis and guide treatment more effectively than it does currently.

> *With continuous monitoring, we could gather information to better understand seizure patterns, manage treatment, and assess how medications are working.*

The exemplary introduced radar-based technology, which can detect changes in HRV up to 3 minutes before the onset of the seizure [20]. Participants see a benefit in an alarm that could be programmed to sound when a change in biomarkers takes place, enabling timely intervention. However, they noted that there is a risk of false alarms or frequent alerts even for minor seizures, so the system needs to be carefully calibrated.

> *There is a risk of constant alarms and continuous beeping, which could cause stress for parents and clinical staff, and place a burden on them.*

Freedom of Movement for Patients During Monitoring

The potential for contactless sensor technologies to avoid restricting patients' freedom of movement during monitoring was highlighted. This could help prevent additional emotional distress, especially to pediatric patients, and thus be relieving for the health care professionals involved. Specifically, for the introduced radar technology, participants noted that installing it in the patient's room or near their bed allows patients to move freely within the radar's field of view, thereby enhancing patient comfort.

> *Compared to other monitors, it's contactless, making it easier to handle with patients causing less physical strain on them.*

Security for Parents or Carers Observing Seizures

It was suggested that it would be valuable if contactless sensor technologies could automatically document seizures and trigger an alarm. Participants emphasized that parents or carers will continue to play an essential role in pediatric epilepsy diagnosis and monitoring by observing and documenting epileptic seizures; however, such technologies might have the potential to provide them with support and greater certainty in this endeavor. Particularly at night, sensors could reduce anxiety about missing a seizure and reduce the pressure of constant vigilance.

> *With this technology, we don't have to rely solely on parents to spot every seizure. It helps make things more objective, reduces anxiety about missing a seizure, and allows parents to sleep more peacefully, even when they are not present.*

Supplementary Monitoring in the Outpatient Setting

Focus group participants highlighted that contactless sensor technologies could be particularly suitable for monitoring at home. They suggested that long-term epilepsy monitoring data outside the clinical setting could open up new perspectives for epilepsy management. Participants noted that nonintrusive epilepsy monitoring in the child's home may improve adherence to monitoring and reduce the stigmatization associated with frequent hospital visits. However, acceptance among patients and parents or carers in the outpatient setting was discussed. Participants emphasized that it may depend on the child's age and on alarm settings, suggesting that alarms might be most useful for longer seizures. They also highlighted several requirements for outpatient use, including reliability, quick and

XSL•FO
RenderX

easy operation, clear instructions and training, and the availability of a contact person in case of problems, especially during the night.

> *A hotline at night for technical problems could be very helpful.*

Discussion

Principal Findings

Participant observation revealed a broad set of psychosocial, medical, and organizational challenges faced by health care professionals during standard pediatric epilepsy monitoring in a university hospital epilepsy center. Building on these findings, the focus group highlighted how contactless sensor technologies might address selected aspects of these challenges. Overall, such technologies were consistently perceived as a complementary rather than replacement tool for EEG and clinical observation, aligning with literature emphasizing the value of multimodal approaches in epilepsy monitoring [6,7].

Psychosocial Perspective

Standard pediatric epilepsy monitoring imposes considerable psychosocial demands on health care professionals, who must continuously instruct, reassure, and calm patients and their parents or carers during EEG procedures and clinical observation. This sustained emotional labor contributes to workload and stress in an already demanding clinical environment, with potential implications for care quality and staff burnout [37]. Here, the focus group shows that contactless sensor technologies' value lies primarily in avoiding additional strain rather than reducing existing burdens.

Observations also highlighted stress among parents or carers due to the fear of missing seizure events, a burden well documented in the literature [10]. Participants suggested that supplementary contactless monitoring could increase families' sense of security by enabling continuous surveillance, while clearly emphasizing that clinical observation remains essential.

Medical Perspective

Participant observation identified challenges in standard pediatric EEG procedures, including movement artifacts, imprecise seizure markers, and electrode displacement, particularly in children who have difficulty tolerating electrodes or remaining still. Incomplete seizure documentation further compromised the quality and diagnostic yield, with direct implications for clinical decision-making [38,39]. Focus group participants perceived contactless sensor technologies as potentially valuable for providing continuous long-term data to supplement standard monitoring by capturing additional physiological correlates of seizures and informing clinical decisions. At the same time, such technologies have inherent technical and medical limitations, including susceptibility to motion artifacts, limited seizure-type differentiation [40], and the absence of direct cortical signal acquisition. These constraints vary depending on the underlying physiological signals and algorithms, reinforcing the role of contactless sensors as an adjunct rather than an alternative to EEG. Ensuring technical accuracy, patient comfort, and continuous availability is therefore essential and requires further investigation [7]. While this study primarily explored perceived benefits and requirements related to the noncontact design of such technologies, further research is required to evaluate their medical relevance and performance across different use cases, sensor types, and physiological signals. In line with established clinical EEG practice, the integration of supplementary contactless monitoring technologies requires clearly defined minimum standards to ensure patient safety and data quality [38-40]. These include reliable signal accuracy, continuous availability, clear assistance procedures, predefined contraindications, transparent communication of benefits and risks, and well-defined indications for specific patient groups and seizure types [38-40]. Our findings suggest that addressing these requirements is an important implication to ensure that supplementary contactless technologies meaningfully support clinical care without compromising established neurophysiological standards.

Organizational Perspective

Standard pediatric epilepsy monitoring was found to be labor- and resource-intensive, requiring high coordination across departments and staff-intensive patient transfers. Participants did not view contactless sensor technologies as a solution to these structural challenges, as standard monitoring procedures would remain necessary. Organizational burdens are therefore more likely to be addressed through workflow or structural changes or through alternative technical or nontechnical innovations [41-44].

Potential for Outpatient Monitoring

Although this study focused on inpatient monitoring, participants noted the potential relevance of contactless technologies in outpatient settings. This aligns with evidence that long-term home monitoring can complement diagnosis and treatment planning [45], particularly given the variability and underreporting of seizures by patients and parents or carers [46]. While prior studies have explored patient and caregiver preferences for seizure detection devices in outpatient settings [47], pediatric-specific research on home-monitoring burdens and acceptance remains limited. Future research should examine the feasibility, perceived benefits, and user acceptance of contactless technologies among pediatric patients and their families, building on similar work from adult epilepsy care with wearable technologies [48].

Strengths and Limitations

Research on new digital technologies is challenging and requires methodological innovations [27,49,50]. In this qualitative study, we applied a sequential, integrated design, using participant observation to identify challenges in standard pediatric epilepsy monitoring, followed by a focus group to explore the potential role of supplementary contactless sensor technologies. This ensured that technological innovation is grounded in real-world clinical needs and constraints [26].

A key strength of this study is the use of participant observation conducted by an external researcher, supplemented by verbal explanations from the health care professionals involved, which provides a more objective perspective than relying solely on

XSL•FO
RenderX

verbal accounts from those involved in the procedures [13,14]. Observations have proven particularly relevant in the context of health technologies [51], as they allow researchers to move beyond subjective intended use and capture actual practices. However, the observer's presence and concurrent explanations may have influenced behavior.

Subjective bias was mitigated through a structured observation guide and by involving an external researcher rather than a member of the pediatric neurology department. Additionally, the observer's background in medical engineering may have shaped attention toward technical aspects, which we tried to balance by following the observation guide and consulting with a neuropediatrician when needed.

The focus group captured expert perspectives from health care professionals directly involved in pediatric epilepsy care, ensuring high clinical relevance. However, the exclusively professional perspective and the predominance of pediatric neurologists may have influenced the findings. Future studies should therefore include a broader range of health care professionals, as well as patients and parents or carers, to capture diverse experiences and expectations.

This study was conducted at a single tertiary hospital in Germany, where resources, staffing, and EEG logistics may differ from other settings. Potential differences in such contextual factors and internal processes make it impossible to generalize the findings completely or apply them directly to other epilepsy centers; however, we may safely assume that the challenges identified can reasonably be expected in comparable settings. Future research should include multisite studies across different care contexts (eg, outpatient or non-university hospital settings) to capture a broader range of contextual factors.

Conclusions

We have found standard pediatric epilepsy monitoring in a university hospital epilepsy center to entail psychosocial, medical, and organizational challenges for health care professionals. Having identified these, we consider that we have created an important foundation for the development of new supportive contactless sensor technologies for this setting. Any such new technology needs to fit into the specific care setting and seek to address the existing challenges and burdens in order to deliver a needs-based benefit. The capacity of contactless sensor technologies to monitor seizures without physical contact with the patient suggests that they have great potential to support standard pediatric epilepsy monitoring in both in- and outpatient settings with long-term monitoring data.

Acknowledgments

The present work was performed by Natalie Öhl in (partial) fulfillment of the requirements for obtaining the degree "Dr. rer. biol. hum" at the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). The authors extend their thanks to Sandra Grimminger for providing training in participant observation to Isabell Link. The authors thank the project partners in BrainEpP for collaboration: Geratherm Respiratory GmbH, DeMeTec GmbH, Voigtmann GmbH, Silicon Radar GmbH, Hamburg University of Technology (TUHH): Institute for High Frequency Technology, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU): Department of Electrical Engineering, Department of Criminal Law, Criminal Procedural Law, Commercial Criminal Law, and Medical Criminal Law. The manuscript was language-edited by the Language Service of the Friedrich-Alexander-Universität Erlangen-Nürnberg. The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing. The GenAI tool used was ChatGPT 4.5. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was funded by the German Federal Ministry of Education and Research (BMBF) (13GW0295D) and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant SFB 1483 – Project-ID 442419336.

Data Availability

The datasets used and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: CO, RT, and TS initiated the project and developed the overall study design for BrainEpP.
Formal analysis: NÖ conducted the analysis and interpretation of the data. IL and MH contributed substantially to the data analysis and interpretation. DS, TS, CO, and RT provided their medical expertise to support the interpretation of the data.
Funding acquisition: CO, RT, and TS wrote the funding proposal.
Investigation: IL and MaHo collected the study data. MaHo was responsible for participant recruitment for the focus group and for coordinating the participant observation.
AW assisted with data collection during the focus group. DS and SK supported the data collection for the participant observation.
Methodology: MH and IL developed the study protocol for the observational field study and the focus group. DS, SK, TS, CO, and RT contributed their medical expertise to the study design.
Project administration: IL and MaHo coordinated the study activities.

XSL•FO
RenderX

Supervision: MH, CO, and RT supervised the study.
Validation: NÖ, MH, and IL verified the consistency and reproducibility of the study results and research outputs.
Visualization: NÖ created visualizations.
Writing – original draft: NÖ wrote the original manuscript draft.
All authors thoroughly reviewed the manuscript draft.

Conflicts of Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Multimedia Appendix 1
Observation guide.
[DOCX File, 15 KB - jmir_v28i1e83009_app1.docx]

Checklist 1
COREQ Checklist.
[PDF File, 429 KB - jmir_v28i1e83009_app2.pdf]

References

1. Fisher RS, Acevedo C, Arzimanoglou A, et al. ILAE official report: a practical clinical definition of epilepsy. Epilepsia 2014 Apr;55(4):475-482. [doi: 10.1111/epi.12550] [Medline: 24730690]
2. Hartl E, Noachtar S. EEG: Stellenwerte und technische Voraussetzungen der EEG-Untersuchung [Article in German]. Klin Neurophysiol 2018 Feb;49(1):48-55. [doi: 10.1055/s-0043-124742]
3. Rosenow F, Klein KM, Hamer HM. Non-invasive EEG evaluation in epilepsy diagnosis. Expert Rev Neurother 2015 Apr;15(4):425-444. [doi: 10.1586/14737175.2015.1025382] [Medline: 25779862]
4. Bandopadhyay R, Singh T, Ghoneim MM, et al. Recent developments in diagnosis of epilepsy: scope of MicroRNA and technological advancements. Biology (Basel) 2021 Oct 25;10(11):1097. [doi: 10.3390/biology10111097] [Medline: 34827090]
5. Marathe M. Inpatient experiences of long-term monitoring: the case of EEG. : ACM; 2024 Nov 11 Presented at: CSCW: '24The 27th ACM Conference on Computer-Supported Cooperative Work and Social Computing; Nov 9-13, 2024; San Jose Costa Rica p. 511-518. [doi: 10.1145/3678884.3681899]
6. Van de Vel A, Cuppens K, Bonroy B, et al. Non-EEG seizure detection systems and potential SUDEP prevention: state of the art: review and update. In: Huffel S, editor. Seizure 2016 Oct;41:141-153. [doi: 10.1016/j.seizure.2016.07.012] [Medline: 27567266]
7. Ramgopal S, Thome-Souza S, Jackson M, et al. Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy. Epilepsy Behav 2014 Aug;37:291-307. [doi: 10.1016/j.yebeh.2014.06.023] [Medline: 25174001]
8. Sansevere AJ, Duncan ED, Libenson MH, Loddenkemper T, Pearl PL, Tasker RC. Continuous EEG in pediatric critical care: yield and efficiency of seizure detection. J Clin Neurophysiol 2017 Sep;34(5):421-426. [doi: 10.1097/WNP.0000000000000379] [Medline: 28430674]
9. Sullivan JE, Corcoran-Donnelly M, Dlugos DJ. Challenges in pediatric video-EEG monitoring. Am J Electroneurodiagnostic Technol 2007 Jun;47(2):127-139. [doi: 10.1080/1086508X.2007.11079619] [Medline: 17679580]
10. van Westrhenen A, de Lange WFM, Hagebeuk EEO, Lazeron RHC, Thijs RD, Kars MC. Parental experiences and perspectives on the value of seizure detection while caring for a child with epilepsy: a qualitative study. Epilepsy Behav 2021 Sep 28;124:108323. [doi: 10.1016/j.yebeh.2021.108323] [Medline: 34598099]
11. Harden J, Black R, Chin RFM. Families' experiences of living with pediatric epilepsy: a qualitative systematic review. Epilepsy Behav 2016 Jul;60:225-237. [doi: 10.1016/j.yebeh.2016.04.034] [Medline: 27240309]
12. Chong L, Jamieson NJ, Gill D, et al. Children's experiences of epilepsy: a systematic review of qualitative studies. Pediatrics 2016 Sep;138(3):e20160658. [doi: 10.1542/peds.2016-0658] [Medline: 27511947]
13. Hammersley M, Atkinson P. Ethnography: Principles in Practice, 4th edition: Routledge; 2019. [doi: 10.4324/9781315146027]
14. Kawulich B. Participant observation as a data collection method. Forum: Qual Soc Res Sozialforschung 2005;6. [doi: 10.17169/fqs-6.2.466]
15. Liebetruth M, Kehe K, Steinritz D, Sammito S. Systematic literature review regarding heart rate and respiratory rate measurement by means of radar technology. Sensors (Basel) 2024 Feb 4;24(3):1003. [doi: 10.3390/s24031003] [Medline: 38339721]
16. Grießhammer SG, Malessa A, Lu H, et al. Contactless radar-based heart rate estimation in palliative care - a feasibility study and possible use in symptom management. BMC Palliat Care 2024 Nov 30;23(1):273. [doi: 10.1186/s12904-024-01592-3] [Medline: 39616332]
17. Mejdani D, Bräunig J, GrießHammer SG, et al. Radar-based tremor quantification using deep learning for improved Parkinson's and palliative care assessment. Trans Rad Sys 2024;2:1174-1185. [doi: 10.1109/TRS.2024.3494473]

XSL•FO
RenderX

18. Braeunig J, Mejdani D, Krauss D, et al. Radar-based recognition of activities of daily living in the palliative care context using deep learning. : IEEE; 2023 Presented at: 2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI); Oct 15-18, 2023; Pittsburgh, PA, USA p. 1-4. [doi: 10.1109/BHI58575.2023.10313506]
19. Linz S, Vinci G, Mann S, et al. A compact, versatile six-port radar module for industrial and medical applications. Journal of Electrical and Computer Engineering 2013;2013:1-10. [doi: 10.1155/2013/382913]
20. Vinci G, Lindner S, Barbon F, et al. Six-port radar sensor for remote respiration rate and heartbeat vital-sign monitoring. IEEE Trans Microwave Theory Techn 2013;61(5):2093-2100. [doi: 10.1109/TMTT.2013.2247055]
21. Islam SMM. Radar-based remote physiological sensing: progress, challenges, and opportunities. Front Physiol 2022;13:955208. [doi: 10.3389/fphys.2022.955208] [Medline: 36304581]
22. Sevcencu C, Struijk JJ. Autonomic alterations and cardiac changes in epilepsy. Epilepsia 2010 May;51(5):725-737. [doi: 10.1111/j.1528-1167.2009.02479.x] [Medline: 20067509]
23. Lu H, Kusnik S, Mammadova D, Trollmann R, Koelpin A. HRV-based monitoring of neonatal seizures with machine learning. : IEEE; 2024 Presented at: 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); Jul 15-19, 2024; Orlando, FL, USA p. 1-4. [doi: 10.1109/EMBC53108.2024.10782590]
24. Shi K, Schellenberger S, Will C, et al. A dataset of radar-recorded heart sounds and vital signs including synchronised reference sensor signals. Sci Data 2020 Feb 13;7(1):50. [doi: 10.1038/s41597-020-0390-1] [Medline: 32054854]
25. Shi K, Steigleder T, Schellenberger S, et al. Contactless analysis of heart rate variability during cold pressor test using radar interferometry and bidirectional LSTM networks. Sci Rep 2021 Feb 4;11(1):3025. [doi: 10.1038/s41598-021-81101-1] [Medline: 33542260]
26. Yusof MM, Kuljis J, Papazafeiropoulou A, Stergioulas LK. An evaluation framework for health information systems: human, organization and technology-fit factors (HOT-fit). Int J Med Inform 2008 Jun;77(6):386-398. [doi: 10.1016/j.ijmedinf.2007.08.011] [Medline: 17964851]
27. Weisser A, Öhl N, Steigleder T, et al. Soziale und ethische Implikationen neuer technologien in der palliativversorgung: kriterien für eine personenzentrierte methodenwahl. Zeitschrift für Palliativmedizin 2024 Sep;25(5):e79-e79. [doi: 10.1055/s-0044-1788537]
28. O'Reilly K. Ethnographic Methods, 2nd edition: Routledge; 2012:255. [doi: 10.4324/9780203864722]
29. Barbour R. Doing Focus Groups: SAGE Publications Ltd; 2018. [doi: 10.4135/9781526441836]
30. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
31. Walter U, Brandt S, Förderreuther S, et al. Empfehlungen der Deutschen Gesellschaft für Klinische Neurophysiologie und Funktionelle Bildgebung zur Diagnostik des irreversiblen Hirnfunktionsausfalls [Article in German]. Klin Neurophysiol 2019 Mar;50(1):17-22. [doi: 10.1055/a-0792-0986]
32. MAXQDA: VERBI Software; 2022. URL: https://www.maxqda.com [accessed 2025-08-01]
33. Mayring P. Qualitative Content Analysis. Forum Qual Soc Res 2000;1(2). [doi: 10.17169/fqs-1.2.1089]
34. Mayring P. Qualitative Content Analysis: A Step-by-Step Guide: SAGE Publications; 2021:978-1529701975. [doi: 10.4135/9781036231798]
35. Pelz C, Schmitt A, Meis M. Knowledge mapping as a tool for analyzing focus groups and presenting their results in market and evaluation research. Forum Qual Soc Res 2004;5(2). [doi: 10.17169/fqs-5.2.601]
36. Miro. 2023. URL: https://miro.com [accessed 2025-08-01]
37. Lu M, Speri E, Lauks A, Leibson C, Rao SC. Understanding health care provider burnout when caring for patients with refractory epilepsy in the United States. Neurol Clin Pract 2024 Apr;14(2):e200260. [doi: 10.1212/CPJ.0000000000200260] [Medline: 38585442]
38. Peltola ME, Leitinger M, Halford JJ, et al. Routine and sleep EEG: minimum recording standards of the International Federation of Clinical Neurophysiology and the International League Against Epilepsy. Epilepsia 2023 Mar;64(3):602-618. [doi: 10.1111/epi.17448] [Medline: 36762397]
39. Tatum WO, Mani J, Jin K, et al. Minimum standards for inpatient long-term video-EEG monitoring: a clinical practice guideline of International League Against Epilepsy and the International Federation of Clinical Neurophysiology. Clin Neurophysiol 2022 Feb;134:111-128. [doi: 10.1016/j.clinph.2021.07.016] [Medline: 34955428]
40. Beniczky S, Wiebe S, Jeppesen J, et al. Automated seizure detection using wearable devices: a clinical practice guideline of the International League Against Epilepsy and the International Federation of Clinical Neurophysiology. Clin Neurophysiol 2021 May;132(5):1173-1184. [doi: 10.1016/j.clinph.2020.12.009] [Medline: 33678577]
41. Kiriakopoulos ET, Chu F, Jobst BC, Kobau R. Innovating the practice of epilepsy care to improve patient outcomes: a model for community health worker integration. Front Neurol 2025;16:1654687. [doi: 10.3389/fneur.2025.1654687] [Medline: 41127292]
42. Martz G, Gurung I, Li YH. Implementation of clinic-based data capture in a real-world epilepsy clinic. Epilepsy Behav 2025 Nov;172:110730. [doi: 10.1016/j.yebeh.2025.110730] [Medline: 41043241]
43. Tao S, Lhatoo S, Hampson J, Cui L, Zhang GQ. A bespoke electronic health record for epilepsy care (EpiToMe): development and qualitative evaluation. J Med Internet Res 2021 Feb 12;23(2):e22939. [doi: 10.2196/22939] [Medline: 33576745]

XSL•FO
RenderX

44. Pellinen J, Foster EC, Wilmshurst JM, Zuberi SM, French J. Improving epilepsy diagnosis across the lifespan: approaches and innovations. Lancet Neurol 2024 May;23(5):511-521. [doi: 10.1016/S1474-4422(24)00079-6] [Medline: 38631767]
45. Velis D, Plouin P, Gotman J, da Silva FL, ILAE DMC Subcommittee on Neurophysiology. Recommendations regarding the requirements and applications for long-term recordings in epilepsy. Epilepsia 2007 Feb;48(2):379-384. [doi: 10.1111/j.1528-1167.2007.00920.x] [Medline: 17295634]
46. Elger CE, Hoppe C. Diagnostic challenges in epilepsy: seizure under-reporting and seizure detection. Lancet Neurol 2018 Mar;17(3):279-288. [doi: 10.1016/S1474-4422(18)30038-3] [Medline: 29452687]
47. Bruno E, Viana PF, Sperling MR, Richardson MP. Seizure detection at home: do devices on the market match the needs of people living with epilepsy and their caregivers? Epilepsia 2020 Nov;61 Suppl 1(S1):S11-S24. [doi: 10.1111/epi.16521] [Medline: 32385909]
48. Bruno E, Simblett S, Lang A, et al. Wearable technology in epilepsy: the views of patients, caregivers, and healthcare professionals. Epilepsy Behav 2018 Aug;85:141-149. [doi: 10.1016/j.yebeh.2018.05.044] [Medline: 29940377]
49. Murray E, Hekler EB, Andersson G, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016 Nov;51(5):843-851. [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
50. Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions-a call for innovative evidence generation approaches. NPJ Digit Med 2020;3(1):110. [doi: 10.1038/s41746-020-00314-2] [Medline: 32904379]
51. Zanaboni P, Ngangue P, Mbemba GIC, Schopf TR, Bergmo TS, Gagnon MP. Methods to evaluate the effects of internet-based digital health interventions for citizens: systematic review of reviews. J Med Internet Res 2018 Jun 7;20(6):e10202. [doi: 10.2196/10202] [Medline: 29880470]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Studies
EEG: electroencephalography
HRV: heart rate variability
MTA: medical technical assistant

Edited by A Stone; submitted 26.Aug.2025; peer-reviewed by DF Borges, J Gehrmann; revised version received 27.Jan.2026; accepted 27.Jan.2026; published 24.Mar.2026.

Please cite as:
Öhl N, Strack D, Kusnik SM, Link I, Hofmann M, Steigleder T, Weisser A, Ostgathe C, Trollmann R, Heckel M
Challenges of Standard Pediatric Epilepsy Monitoring and the Potential Benefits of Contactless Sensor Technologies: Exploratory Qualitative Study
J Med Internet Res 2026;28:e83009
URL: https://www.jmir.org/2026/1/e83009
doi:10.2196/83009

© Natalie Öhl, Dilbar Strack, Stefan Martin Kusnik, Isabell Link, Marlene Hofmann, Tobias Steigleder, Alina Weisser, Christoph Ostgathe, Regina Trollmann, Maria Heckel. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development and User-Centered Evaluation of Smart Systems for Loneliness Monitoring in Older Adults: Mixed Methods Study

Yi Zhou[1], PhD; Jessica Rees[2], PhD; Faith Matcham[3], PhD; Ashay Patel[4], PhD; Michela Antonelli[4], PhD; Anthea Tinker[2†], PhD; Sebastien Ourselin[4], PhD; Wei Liu[1], PhD

[1]Department of Engineering, King's College London, London, United Kingdom
[2]Department of Global Health and Social Medicine, King's College London, London, United Kingdom
[3]School of Psychology, Faculty of Science, Engineering and Medicine, University of Sussex, Brighton, United Kingdom
[4]School of Biomedical Engineering & Imaging Sciences, King's College London, London, United Kingdom
[†]deceased

Corresponding Author:
Wei Liu, PhD
Department of Engineering
King's College London
Strand Campus
London, WC2R 2LS
United Kingdom
Phone: 44 020 7836 5454
Email: wei.liu@kcl.ac.uk

Abstract

Background: Loneliness is a critical issue among older adults and constitutes a significant risk factor for a range of physical and mental health conditions. However, current assessment methods primarily rely on self-report questionnaires and clinical evaluations, which are susceptible to recall bias and social desirability bias, highlighting the need for more objective and continuous assessment approaches. Recent studies have reported associations between physiological and behavioral indicators and the experience of loneliness in older adults. While these technologies have demonstrated correlations between physiological and behavioral sensor data and the experience of loneliness, their implementation has been limited. Most systems rely on fixed-location sensors or smartphone apps, with little attention given to the integration of these tools into users' daily routines. To date, no published studies have applied smart textile technology, which integrates sensing capabilities directly into garments or furniture, as a medium for loneliness detection. This study addresses that gap by exploring the usability, experiential acceptability, and ethical considerations of smart textile-based monitoring systems.

Objective: This study aims to assess the perceived usability, acceptability, and emotional resonance of a smart loneliness monitoring system integrating sensing garments, furniture, and a mobile app and identify design implications to guide future improvement and promote sustained engagement among older adults.

Methods: Building on earlier conceptual research, a functional prototype system was developed and evaluated through 2 immersive in-person workshops with older adults (N=10). A mixed methods approach was applied, combining structured questionnaires, sensory ethnographic observations, focus group discussions, and experience-based co-design. Quantitative data were analyzed descriptively, and qualitative data were analyzed thematically to explore user perceptions related to system usability, emotional response, lifestyle compatibility, and ethical considerations.

Results: Quantitative data indicated high user satisfaction in dimensions such as comfort, ease of use, and feedback clarity. However, trust in long-term monitoring and willingness to use the system regularly varied. Thematic analysis revealed 4 main areas influencing acceptance, including wearability, usability, and daily integration; trust, privacy, and data control; perceptions of loneliness and the limits of detection; and adoption, applicability, and ethical futures. Participants emphasized the need for discretion, personalization, and human oversight in system feedback and data-sharing mechanisms.

Conclusions: The resulting prototype was positively received, demonstrating the potential of smart systems for passive and personalized loneliness monitoring among older adults. However, adoption is influenced by perceptions of autonomy, emotional sensitivity, and contextual integration. Future development should focus on modularity, transparency, and integration within care infrastructures to ensure ethical and sustainable deployment.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e81027) doi:10.2196/81027

KEYWORDS

older adults; mental health; loneliness; monitoring; user-centered design; smart textile; wearable and ambient technology; wearable technology

Introduction

Loneliness and social isolation have been identified as significant global mental health challenges, with particularly profound effects on older adults. There are up to a quarter of older individuals worldwide experiencing social isolation [1-3]. As people age, they may reduce their social interactions due to various factors such as reduced mobility, retirement, or the loss of partners, leading to social isolation and intensify feelings of loneliness [4-6]. A growing body of research has shown that prolonged loneliness is associated with increased risks of depression, cognitive decline, cardiovascular disease, and mortality rates, which poses health risks comparable to smoking and obesity [7,8]. Beyond its impact on individuals, loneliness also places substantial strain on health care systems by increasing demand for clinical care and long-term support services [4].

Despite a growing understanding of the impact of loneliness on health, it remains a challenge to accurately measure and monitor loneliness, particularly in nonclinical and home-based settings [2,7]. While clinical settings may allow for structured assessments by health professionals, such measurements are often constrained by time, context, or the presence of other comorbidities. Traditional assessment methods mainly rely on self-report questionnaires such as the University of California, Los Angeles 3-Item Loneliness Scale, the De Jong Gierveld Loneliness Scale for older adults, or clinician-administered questionnaires [9,10]. While these tools are well-validated, they are susceptible to recall and social desirability bias, particularly among older adults who may underreport emotional distress due to stigma or generational attitudes toward mental health [11,12]. Moreover, such methods often provide snapshot assessments rather than capturing the dynamic, fluctuating nature of loneliness as experienced in daily life [13]. These limitations emphasize the need for continuous, objective, and context-aware detection of loneliness, enabling more timely and personalized interventions.

Sensor-based technologies, especially wearable and ambient sensing systems, have advanced considerably in mental health monitoring, offering new opportunities to detect loneliness through physiological and behavioral data [14,15]. Recent studies have reported several behavioral patterns and physiological indicators associated with loneliness, including reduced physical activity [16,17], sleep disturbances [5], binge or comfort eating [18], elevated blood pressure [19], and increased average salivary cortisol levels [20,21]. While these findings do not establish diagnostic relationships, they suggest measurable correlates that may guide the design of future sensing-based systems. Wearable devices such as smartwatches and fitness bands have demonstrated the ability to monitor many of these indicators. When analyzed over time, these data can provide inferences about an individual's psychological well-being and deviations from their baseline states [20,22]. Additionally, various sensing systems have been applied to monitor loneliness and social isolation in older adults. These include vision-based motion capture systems for activity level tracking [23], ambient light and sound sensors for detecting social behaviors [24,25], and smartwatches (including accelerometers or inclinometers) to track posture and sedentary behavior [16,26]. However, camera-based systems often raise privacy concerns, and fixture-mounted sensors such as wall-mounted light and sound sensors may lack the portability required for continuous monitoring at home and in community settings [15]. While wearable devices such as smartwatches offer portable sensing, they present their own limitations in older adults, including discomfort with wrist-worn devices, low personalization, and poor integration into daily domestic routines [27,28]. Moreover, many existing monitoring systems focus solely on physical movement or posture and fail to capture the complex emotional and physiological dimensions of loneliness [14,29].

Textile-based sensing technologies, which integrate sensors and conductive materials into fabrics, are able to offer a comfortable and effective solution for long-term mental health monitoring. By integrating sensing capabilities into flexible fabrics, sensing textile systems can passively and continuously collect data without interfering with users' daily routines or drawing attention to the monitoring process [22,30]. Furthermore, electronic textiles can seamlessly embed into familiar objects such as garments or home furnishings, enhancing both physical comfort and acceptability for older users [31-33]. Despite advances in smart textiles for health monitoring, no textile-based sensing system has been developed specifically for loneliness detection. One prior study explored the use of a textile band to capture speech frequency as an indicator of social interaction, but it did not attempt to evaluate the subjective experience of loneliness or integrate these signals into a mental health monitoring framework [22]. While a growing body of pervasive computing and ambient-assisted living research has focused on detecting loneliness through environmental sensors and wearables [34-36], these systems have largely relied on noncustomizable and device-centric approaches with limited integration into user experiences. In contrast, textile-based systems offer the potential for more seamless, passive, and embodied interaction. However, despite these advantages, they remain underexplored in loneliness-related apps. Previous research has highlighted the critical role of design factors such as format, materials, and sensor placement in user acceptance and sustained engagement with textile-based systems [37,38]. However, few studies have directly integrated co-design, lived experience research, and user-driven evaluation into the development of such systems for older adults. As loneliness is not only just a behavioral state but also a subjective and socially situated experience, the design and development of textile sensing systems need to go beyond engineering efficacy to

XSL•FO
RenderX

reflect the emotional, ethical, and contextual needs of users [12]. This study addresses this gap by placing older adults' voices at the center of system development and evaluation.

In our previous research, we conducted interviews and collected feedback from older users and stakeholders to understand the design requirements and expectations of smart loneliness monitoring systems for older adults [12,27,28]. These earlier works were primarily conceptual, exploring hypothetical interactions and preferences prior to the existence of a working prototype. In contrast, this study advances this body of work by designing and evaluating an integrated smart loneliness monitoring system, comprising sensing garments, furniture, and a companion mobile app through immersive user engagement. Additionally, this study makes a novel contribution by combining prototype-led experience, sensory ethnography, structured quantitative feedback, and co-design outputs to generate both actionable design insights and a deeper understanding of the emotional and ethical responses of older adults. By combining quantitative and qualitative analyses, we examined different dimensions of user acceptance in the context of smart loneliness detection, including wearability, emotional trust, loneliness perception, and pathways to adoption. Finally, we discussed the design implications, ethical considerations, and future directions for integrating smart textiles into the everyday mental health care of older adults.

Methods

Overview

This study builds upon our previous interview and co-design research conducted with older adults and stakeholders, which identified essential user needs and expectations regarding the design and development of smart systems to monitor loneliness in later life [12,27,28]. These early insights informed the development of our smart loneliness monitoring systems, which comprise sensing garments and sensing furniture designed to unobtrusively capture physiological and behavioral indicators associated with loneliness. The resulting prototype was evaluated in the current focus group study.

To evaluate and further improve the system, 2 in-person evaluation workshops were held, each involving 5 older adults aged 65 years and older who had experienced loneliness (N=10). The aim of the workshops was to gather experiential feedback and design suggestions from older users. A mixed methods approach was used, integrating sensory observations, self-report questionnaires, focus group discussions, and co-design activities. These methods enabled a comprehensive exploration of users' practical and emotional feedback to the system, providing pragmatic design insights to guide future development.

Participants

Participants were recruited using a combination of convenience and purposive sampling strategies, with the aim to engage older adults from diverse life backgrounds and with varying levels of technological adaptability. Recruitment was conducted through 2 main channels. First, the research team directly contacted individuals who had previously expressed interest in the DELONELINESS project [39]. Second, a study invitation was distributed via the PROTECT study newsletter, which reaches over 20,000 older adults across the United Kingdom. For logistical feasibility, only individuals residing within a 50-mile radius of central London were considered from the PROTECT email list. Eligibility criteria included being aged 65 years or older, fluent in spoken English, and having experienced loneliness at some point during their later life (postretirement age). Individuals diagnosed with cognitive impairments or dementia were excluded from participation to ensure that participants could provide informed consent and fully engage with the system interaction and co-design activities. Participants were screened by researchers trained in applying the principles of the Mental Capacity Act (2005) [40], which enabled them to assess an individual's capacity to participate during the recruitment stage.

Loneliness severity was assessed using the University of California, Los Angeles Loneliness Scale [41], which has been linked to various health outcomes and functional limitations. In total, 10 participants who met the eligibility criteria took part in the study and completed both workshop sessions. This sample size was determined based on recommendations for user-centered qualitative evaluations, which typically involve 3 to 15 participants to obtain experiential insights and design implications in early-stage technology development [42]. Similar sample sizes have been used in published co-design and feasibility studies involving older adults and digital health technologies [43-45]. A total of 10 participants consented and completed both workshop sessions. While small in size, this sample enabled in-depth participatory engagement, iterative feedback, and contextual exploration, which are central goals of this exploratory mixed methods study. Additionally, data collection concluded after 10 participants, as thematic saturation was observed across the 2 workshop sessions, with recurring patterns and consistent feedback emerging during the analysis phase.

Technology Description

The smart loneliness monitoring systems evaluated in this study consisted of 2 key components, including the sensing garment and sensing furniture. In the previous co-design workshops with older adults and stakeholders, we identified key design factors influencing the older adults' acceptance of monitoring technologies [28]. Additionally, our prior qualitative research exploring the psychological experience of loneliness in later life informed sensing technology selection and development [12,39]. Building on these insights, our systems were constructed to continuously and noninvasively track physiological and behavioral signals associated with loneliness while remaining compatible with the daily lives and domestic environments of older users.

The sensing garment was mainly intended for physiological signal monitoring. To ensure both wearability and sensing accuracy, we developed 3 different sizes (small, medium, and large) of long-sleeved zip-up shirts made from a breathable elastic textile blend (92% polyester and 8% elastane). The shirts were designed to be worn over regular clothing to facilitate dressing and undressing during workshop sessions. The sensing system of the garment included a fabric-based conductive circuit,

XSL•FO
RenderX

modular sensor units, and a data acquisition module. As shown in Figure 1A, textile circuits were mapped along the garment seams and encapsulated using thermoplastic polyurethane via heat-pressing to ensure durability, washability, and smoothness, minimizing tactile discomfort or abrasion risks for older users with sensitive skin.

Figure 1. Components of the smart loneliness monitoring system: (A) sensing garment embedded with conductive textile circuits and modular sensors for pulse rate, respiration, and temperature monitoring. (B) Smart furniture including a sensing pillow and seat pad integrated with pressure sensor module for posture and behavioral monitoring.

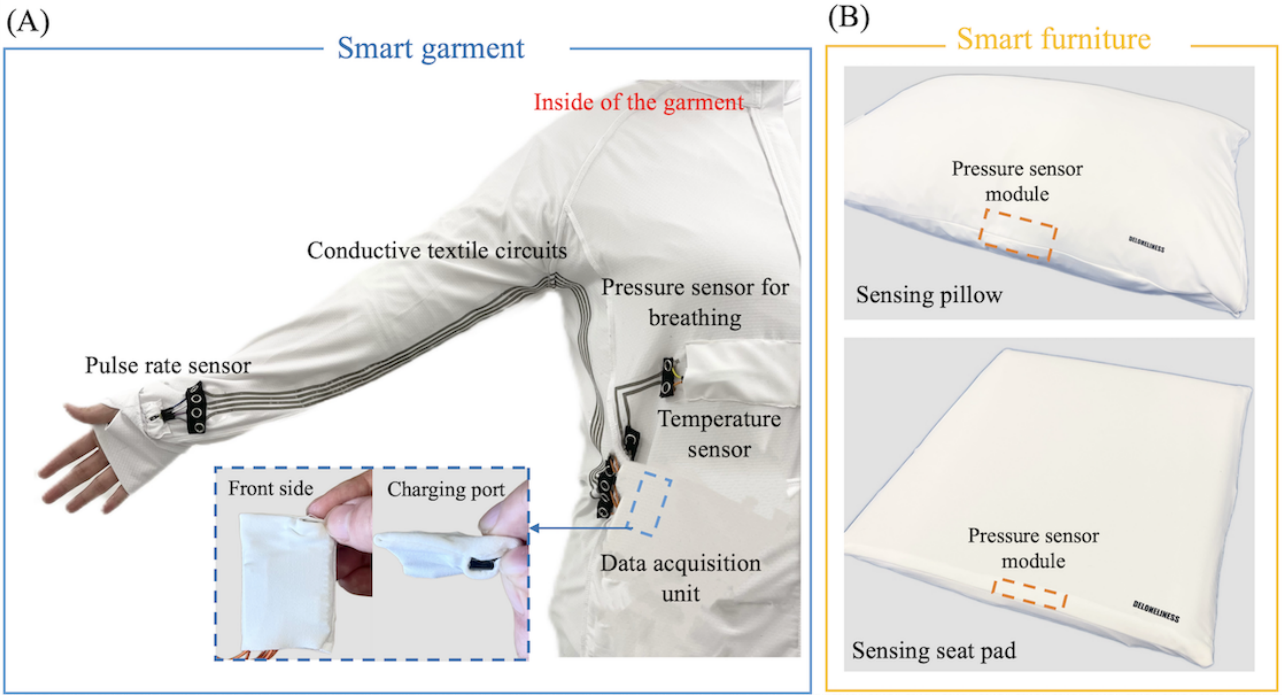

To support individual autonomy and improve independence, a modular design was applied using metal press-fit snaps, allowing users or caregivers to easily attach, detach, or reposition sensing components without technical expertise. Specifically, a pulse rate sensor was placed at the wrist cuff to enable accurate heart rate monitoring. A temperature sensor and a pressure sensor for respiration were embedded on the interior side of the chest and abdominal regions to collect real-time body temperature and breathing rate data. The data acquisition unit included a built-in inertial measurement unit housed in a soft fabric casing and integrated into a garment pocket. The sensing system was powered by a commercially available rechargeable battery and can be conveniently charged via an external charging port without needing to open the casing, thereby reducing the cognitive and physical burden during maintenance. The proposed modular design allowed older users to customize their configurations by selecting and combining the sensing components most relevant to their individual needs.

The sensing furniture was designed based on everyday household items such as a pillow (sensing pillow) and a seat cushion (sensing seat pad) with custom-developed textile covers (Figure 1B). Pressure sensor modules were embedded into designated internal regions of the pillow and seat pad to enable continuous monitoring of posture and pressure distribution while sitting or lying down. Similar to our sensing garment, the furniture also applied a modular design that allowed sensor modules and the data acquisition unit to be attached via press-fit snaps, simplifying removal for maintenance or cleaning of the textile surfaces.

Figure 2 demonstrates the system architecture. The proposed smart loneliness monitoring system was designed to continuously collect physiological and behavioral data through the sensing garment and furniture. These raw signals would be transmitted via a smartphone or communication gateway, which performed preliminary signal preprocessing such as noise filtering and timestamping before uploading to a secure cloud environment. In future iterations, advanced machine learning algorithms would be applied in the cloud to identify potential indicators of loneliness, such as irregular activity patterns, reduced physiological variability, or prolonged inactivity. Thresholds for generating loneliness-related feedback have not yet been predefined. Based on participant input during the co-design stage, the system was intended to include adaptive feedback mechanisms in user interfaces, such as customizable mood prompts and check-in features, allowing users to confirm, dismiss, or annotate inferred states.

XSL•FO
RenderX

Figure 2. Proposed system architecture and data flow.

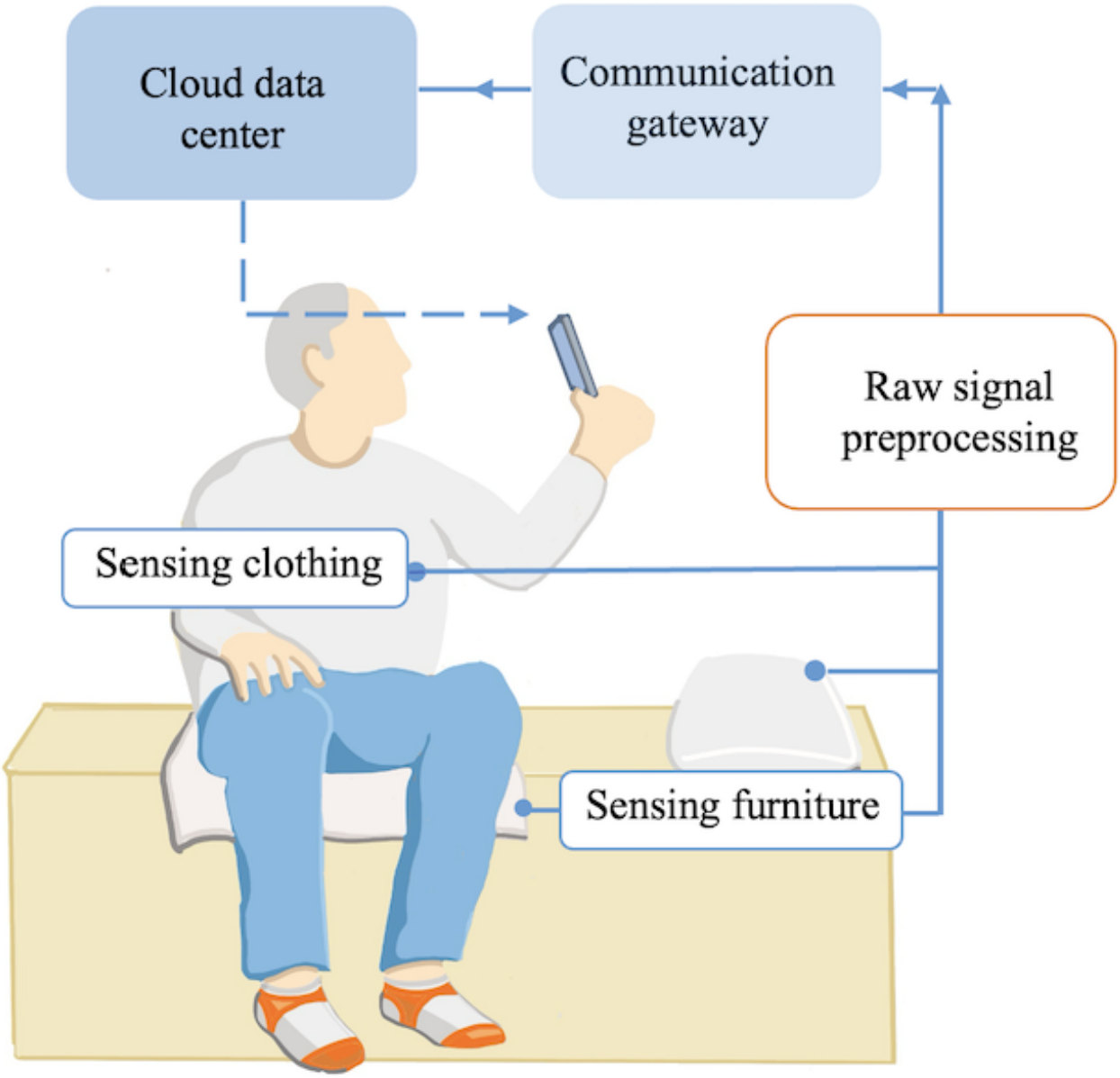

Figure 3 demonstrates our user interfaces from various user perspectives. From the user's perspective (Figure 3A), the app provided real-time feedback on heart rate, respiration rate, skin temperature, sleep duration, and daily activity levels, while also visualizing current loneliness status and offering recommendations for health-promoting activities. The system further included companion interfaces for family members and caregivers (Figure 3B). Family members can monitor their loved one's loneliness status, view summarized statistics, and check upcoming social or medical appointments, with integrated communication options such as direct calling or messaging. For health care professionals (Figure 3C), the app offered detailed health data and aggregated loneliness levels across users, facilitating targeted service recommendations based on individual needs.

XSL•FO
RenderX

Figure 3. User interfaces of the smart loneliness monitoring system. (A) User interface displaying real-time physiological data, loneliness status, and activity recommendations. (B) Family member interface providing an overview of emotional status, scheduled activities, and communication access. (C) Health care professionals interface showing loneliness metrics across individuals and the service referral interface.

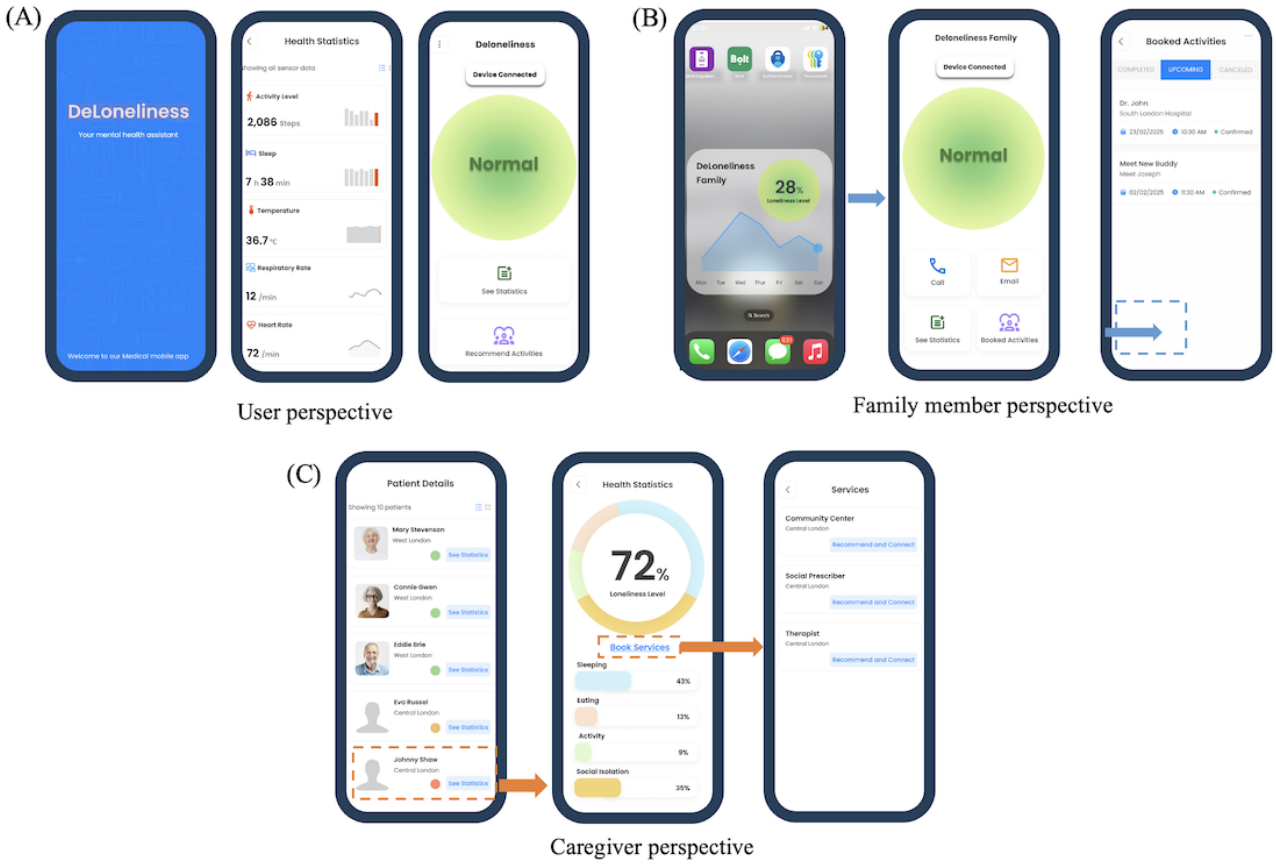

Procedure

Overview

This study was conducted through in-person workshops consisting of a series of structured participatory activities designed to evaluate and improve the smart loneliness monitoring system. The research procedure was divided into 2 sequential components including an experiential evaluation of the system and an experience-based co-design for improvement session (Figure 4).

XSL•FO
RenderX

Figure 4. Workshop procedure.

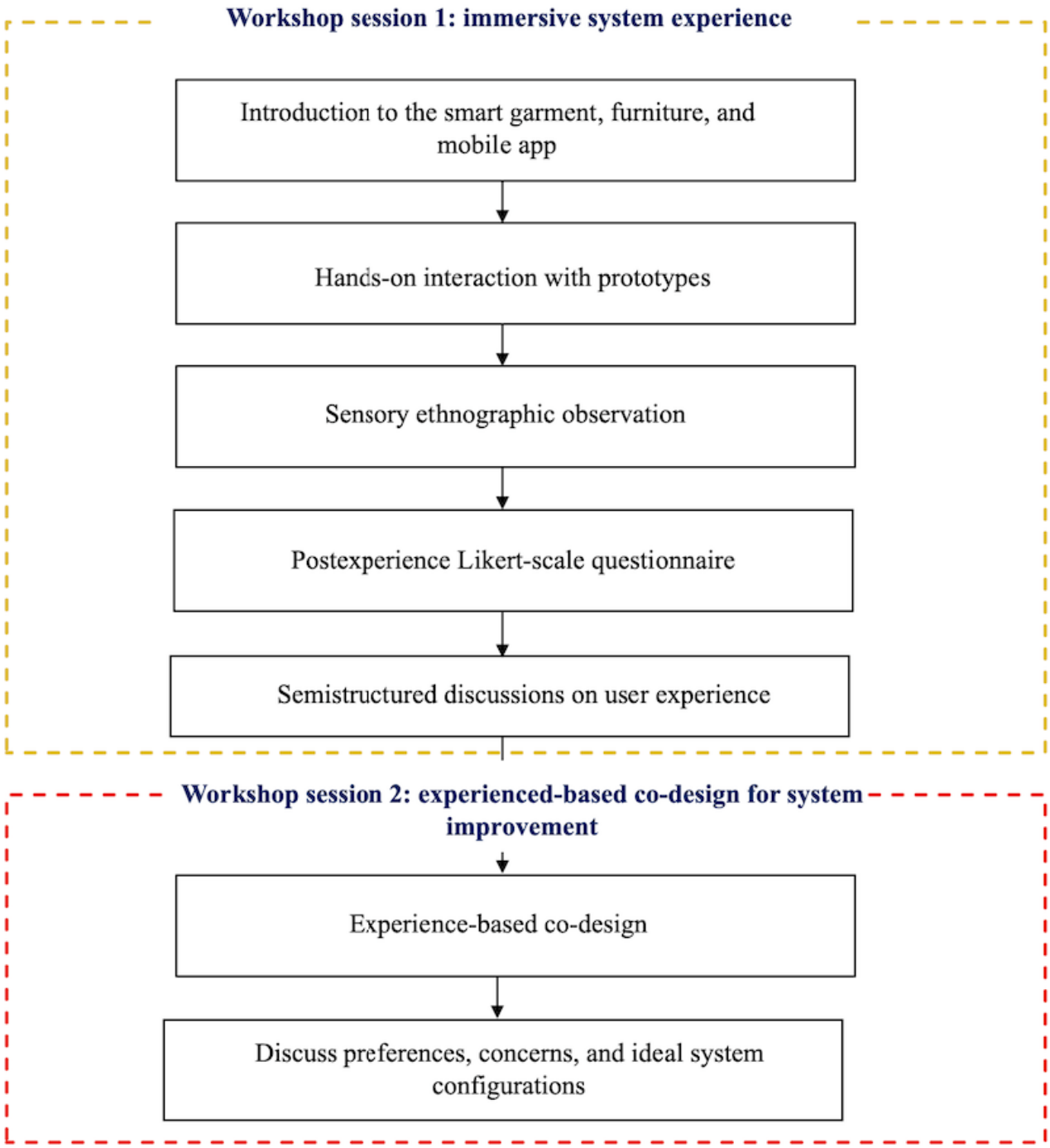

Session 1: Experiencing the Smart Loneliness Monitoring System

The first session of the workshop was designed to allow participants to interact directly with the smart loneliness monitoring system. This session comprised 3 main stages: an introductory demonstration, an immersive trial, and a postexperience feedback phase including a questionnaire and discussion. The session aimed to understand users' initial perceptions, sensory impressions, and evaluations of the system within a supportive and participatory environment.

The session began with an introduction to the smart loneliness monitoring system by the lead author (YZ), who presented the components of the system, including the sensing garment and sensing furniture. Participants were shown a short video demonstrating the system's functional workflow and potential use cases in everyday life. This was followed by a live demonstration, during which researchers explained how to operate, maintain, clean, and charge the system. Participants were encouraged to ask questions at any time during the demonstration, ensuring that the purpose, use, and operational aspects of the system were clarified. This stage served as a foundation for the participants' subsequent independent and immersive system interaction.

All the participants were then invited to try on the sensing garment and interact with the sensing furniture (sitting on the

XSL•FO
RenderX

sensing seat pad or lying on the pillow) and operate the accompanying mobile app to see the real-time physiological feedback. This immersive experience was guided by principles of sensory ethnography, which is a commonly applied method in the evaluation of wearable and environmental technologies to capture the embodied and affective dimensions of user experience [46-48]. Participants were encouraged to attend to their physical, emotional, and sensory responses such as how the garment felt on their skin, the sensations caused by the furniture, and their affective responses to the system feedback while experiencing the systems. Three research assistants (YZ, JR, and WL) used structured observation forms (Multimedia Appendix 1) to document participants' verbal and nonverbal reactions, behaviors, and interactions with the materials.

After the immersive trial, participants completed a Likert-scale questionnaire, designed to assess key dimensions such as ease of use, comfort, perceived usefulness, trust in the system, and concerns about data privacy. The questionnaire was informed by the technology acceptance model (TAM), which is a widely used theoretical framework that explains users' acceptance of technology based on perceived usefulness and perceived ease of use [49]. Specifically, we adapted core constructs from TAM, including perceived usefulness, ease of use, and behavioral intention to use. To reflect the characteristics of smart textile and ambient sensing systems, we also supplemented items derived from established usability and health technology evaluation frameworks such as comfort, aesthetics, data privacy, and daily life integration [50,51]. A full list of questionnaire items is provided in Multimedia Appendix 2. The experience session concluded with a semistructured focus group discussion. The discussion guide was also informed by the TAM and themes identified in our previous co-design research, which served to ensure comprehensive topic coverage rather than dictate interpretive categories [27]. Topics explored included the system's integration into daily life, emotional responses to the monitoring experience, clarity and interpretability of feedback, comfort level of the systems, and concerns around data sharing and privacy.

Session 2: Experience-Based Co-Design for System Improvement

The second part of the workshop focused on collecting user-driven suggestions for system improvements through an experience-based co-design activity [52]. Based on the insights and impressions gained during the initial system trial, participants were invited by the lead investigator (YZ) to reimagine various aspects of the smart systems. This session aims to empower older adults to become cocreators, allowing them to contribute their experiential knowledge and preferences toward the future development of a more practical, comfortable, and acceptable solution.

Each participant was provided with a design toolkit, which included body and furniture layout co-design template (Multimedia Appendix 3), sensor placement icons, and a range of fabric samples representing commonly used textile compositions (100% cotton, 100% wool, 80% cotton/20% polyester, 80% wool/20% polyester, and 92% polyester/8% elastane). The toolkit also includes materials for conceptualizing sensor integration methods, such as Velcro, magnetic clasps, hooks, and press-fit snaps, which are commonly used to attach electronic modules in textile-based systems.

The session began with a short reflective exercise, in which participants were asked to individually outline their daily routines, experience prompts, and situational preferences related to loneliness and technology use. This laid the foundation for them to anchor design thinking in their own life experience. Participants were then encouraged to engage in open-ended visual prototyping using the templates and materials provided. They annotated body outlines and home layouts with preferred sensor placement zones, marked areas to be avoided for comfort or privacy reasons, and proposed new features such as adjustable garment structures, softer fabric options, or additional sensing functions to enhance comfort, usability, and emotional acceptance.

The co-design session ended with a group sharing activity, where each participant presented their redesigned concept to the group and explained the rationale behind their design decisions. These presentations provided valuable qualitative insights into the different preferences and expectations of older users, which can inform iterative improvements in the future system design.

Data Analysis

Overview

A mixed methods analytical approach was applied to analyze the data collected during the workshops. The study was structured in sequential phases, where the initial quantitative questionnaire was used to prompt structured participant reflection, and the subsequent qualitative phase provided a deeper contextual understanding. Quantitative data from the Likert-scale questionnaire were analyzed descriptively. The qualitative data including audio recordings from focus group discussions, co-design artifacts, field notes from sensory ethnographic observations, and participants' reflective comments following the co-design session were analyzed using inductive thematic analysis [53,54].

Quantitative Data Analysis

Quantitative data from the postexperience Likert-scale questionnaires were entered into SPSS (version 29; IBM Corp) and analyzed using descriptive statistics, including means, SDs, and frequency distributions. These data provided an initial understanding of participants' perceptions and satisfaction with the system and served as a foundation for subsequent qualitative discussions.

Qualitative Data Analysis

Audio recordings from the focus group discussions and the co-design sharing presentations were transcribed collaboratively by the first and second authors (YZ and JR). In addition, co-design artifacts such as annotated body and home layout templates, sticky notes, and sketches were digitized and treated as supplementary qualitative data. Field notes from sensory ethnographic observations were also integrated into the qualitative findings to contextualize participants' behaviors and embodied responses during interaction with the system.

XSL•FO
RenderX

Thematic analysis followed Braun and Clarke's [55] 6-phase framework and was independently conducted by 2 researchers (YZ and WL). The researchers first familiarized themselves with the data through review of transcripts and design materials, followed by open coding using NVivo (version 14; Lumivero). Codes were iteratively grouped into broader themes, which were refined through collaborative discussions within the research team. Although the focus group guide was informed by constructs from the TAM and themes identified in prior co-design research, data analysis proceeded entirely inductively. Themes were generated directly from participants' perspectives and lived experiences, rather than being constrained by a predetermined theoretical framework.

While the earlier quantitative assessment was structured around 8 predefined dimensions, the subsequent qualitative analysis adopted an open-coding approach. This was to ensure that participants' language, emotional nuance, and contextual expressions were not shaped or limited by prior assumptions. Although some thematic overlap with the questionnaire domains was observed, the 2 phases were intentionally not aligned, allowing for the emergence of novel concerns and experiential insights.

Ethical Considerations

This study received ethics approval from the King's College London Research Ethics Committee (reference: LRS/DP-24/25-34602). Prior to participation, all individuals were provided with a participant information sheet and a consent form, which has explained the purpose of the study, their rights, and the voluntary nature of participation. Participants were informed that they could withdraw from the study at any time. Written informed consent was obtained on the day of the workshop. During data processing, all the data were anonymized by assigning each participant a unique and nonidentifiable identification number. A password-protected file containing participants' names and contact details was stored separately from research data and was only accessible to the core research team. No direct financial compensation was provided for participation. However, refreshments and lunch were offered on the day of the workshop, and travel expenses were reimbursed.

Results

Participants' Characteristics

A total of 10 participants meeting the eligibility criteria took part in the study. Table 1 summarizes their demographic characteristics, including age range, sex, living arrangements, levels of technology use, prior experience with health-monitoring technologies, and loneliness score.

XSL•FO
RenderX

Table 1. Characteristics of the participants (N=10).

Characteristics	Values
Age (years)	
Mean (SD)	68.8 (4.2)
Range	65-79
Sex, n (%)	
Female	8 (80)
Male	2 (20)
Highest education level, n (%)	
Secondary education or below	6 (60)
Postsecondary education	4 (40)
Employment status, n (%)	
Retired	9 (90)
Employed	1 (10)
Living arrangement, n (%)	
Living alone	5 (50)
Living with others	5 (50)
Technology use, n (%)	
Low	2 (20)
High	8 (80)
Experiencing in using health monitoring technology	
Yes	5 (50)
No	5 (50)
Loneliness score (UCLA[a])	
Mean (SD)	5.1 (1.2)
Range	3-7

[a]UCLA: University of California, Los Angeles.

Descriptive Statistics

We conducted a quantitative analysis of the postexperience questionnaires completed by participants following their interaction with the smart loneliness monitoring system. The questionnaire comprised 18 items spanning 7 key dimensions, including ease of use, integration into daily life, perceived usefulness, understanding and clarity of feedback, trust in system functionality, concerns of data privacy, comfort level, and overall system acceptance. Each item was rated on a 5-point Likert scale starting from 1=strongly disagree" to 5=strongly agree.

Table 2 summarizes the descriptive statistics for user perceptions of the smart loneliness monitoring system. Overall, the responses reflected a positive attitude toward the system, with participants tending to "agree" or "strongly agree" with the statements across various dimensions. The highest-rated dimensions were ease of use, understanding and clarity of feedback, and comfort level, with average scores of 4.0 (SD 0.73), 4.5 (SD 0.52), and 4.4 (SD 0.53), respectively. However, responses related to the system's long-term integration, trust, and data sharing showed greater variability, indicating that evaluations of the system's role in users' everyday life were more complex and individualized. These divergences will be further explored in the qualitative findings presented in the Thematic Analysis section.

XSL•FO
RenderX

Table 2. Descriptive statistics for user perceptions of the smart loneliness monitoring systems.

Item description	Mean (SD)	FD[a]-1	FD-2	FD-3	FD-4	FD-5
1. Ease of use						
System is easy and convenient to use	3.8 (0.92)	0	1	2	5	2
Garment easy to put on or take off	4.4 (0.70)	0	0	1	4	5
Maintenance process is manageable	3.9 (0.57)	0	0	2	7	1
2. Integration into daily life						
Willing to use regularly	3.5 (1.18)	0	2	4	1	3
System fits into daily routines	3.7 (0.67)	0	0	4	5	1
3. Usefulness						
Useful for personal well-being	3.8 (0.82)	0	1	2	6	1
Useful for detecting loneliness	3.9 (0.87)	0	0	4	3	3
4. Understanding and clarity of feedback						
App feedback was easy to understand and interpret	4.5 (0.52)	0	0	0	5	5
5. Trust and reliability						
Trust in system performance	3.6 (0.69)	0	0	5	4	1
System reliably monitors loneliness conditions	3.4 (0.51)	0	0	6	4	0
6. Privacy and data concerns						
Comfortable sharing data	3.6 (0.84)	0	1	3	5	1
Comfortable being continuously monitored	3.8 (1.13)	0	2	1	4	3
7. Comfort level						
Comfort of sensing garment	4.3 (0.48)	0	0	0	7	3
Comfort of sensing furniture	4.9 (0.32)	0	0	0	1	9
Comfort of electronic textile component	4.3 (0.48)	0	0	0	7	3
Systems feels emotionally supportive	4.1 (0.87)	0	0	3	3	4
8. Overall acceptance						
Overall system acceptability	3.9 (0.73)	0	0	3	5	2
Willingness to recommend to others	3.7 (0.67)	0	0	4	5	1

[a]FD: frequency distribution.

Specifically, the ease of use received relatively high ratings. The highest-scoring item was “The garment was easy to put on and take off” (mean 4.4, SD 0.70), followed by “The maintenance process is manageable” (mean 3.9, SD 0.57). These results suggest that most of the participants were able to interact with the system independently and comfortably. In contrast, the integration into daily life revealed different perspectives. While most participants agreed that the system could be incorporated into their routine (mean 3.7, SD 0.67), their willingness to use the system regularly over time varied more significantly (mean 3.5, SD 1.18). This points to the complexity of sustained engagement and highlights participants’ desire for personalization and control. These themes are also demonstrated in their co-design for improvement session (see Wearability, Usability, and Daily Integration section). In terms of perceived usefulness, most participants believed that the system would be beneficial to them (mean 3.8, SD 0.82) and could effectively detect indicators of loneliness (mean 3.9, SD 0.87). These ratings support the conceptual value of the system. However, several participants expressed concerns about the transparency of how loneliness was inferred, especially in relation to physiological data. This issue was elaborated further in the focus group discussions (see Trust, Privacy, and Data Control section). Additionally, the dimension of understanding and clarity of feedback was rated highly. All participants reported being able to interpret the outputs provided by the accompanying mobile app (mean 4.5, SD 0.52), indicating a positive perception of the interface’s communicative clarity. However, ratings related to trust and privacy were more mixed. Participants expressed moderate trust in the system’s ability to reliably monitor loneliness (mean 3.4, SD 0.51). Responses to data sharing (mean 3.6, SD 0.84) and continuous monitoring (mean 3.8, SD 1.13) were generally positive, but the higher SDs suggest a divergence in acceptance of long-term monitoring technologies among participants. These differences were explored in depth during focus group discussions, where participants reflected on the roles of caregivers and family members in accessing sensitive data and expressed a wide range of perspectives on appropriate

XSL•FO
RenderX

data governance (see the Adoption, Applicability, and Ethical Futures section). Furthermore, the comfort dimension received the most consistently high ratings. Items relating to the comfort of the sensing garment (mean 4.3, SD 0.48), sensing furniture (mean 4.9, SD 0.32), and electronic textile materials (mean 4.3, SD 0.48) showed both high mean scores and low variability, indicating strong consensus among participants. Observational data also support these findings, with most participants exhibiting relaxed body language, positive comments, and tactile exploratory behaviors consistent with physical ease and embodied comfort. Finally, the dimension of overall system acceptance was rated positively (mean 3.9, SD 0.73), and participants reported moderate willingness to recommend the system to others (mean 3.7, SD 0.67). These findings suggest general acceptance and openness to the concept while also indicating further specific design improvement required to support long-term compliance.

Overall, these quantitative results provide an initial understanding of participants' functional and emotional responses to the system. They also helped shape the focus of the thematic analysis by identifying areas of strong consensus and divergence, which are further explored through focus group discussions, co-design artifacts, and sensory ethnographic observations in the Thematic Analysis section.

Thematic Analysis

Overview

Thematic analysis of the focus group discussions, co-design artifacts, and sensory ethnographic observations resulted in the identification of 4 main themes, each encompassing multiple subthemes, including wearability, usability, and daily integration; trust, privacy, and data control; perceptions of loneliness and the limits of detection; and adoption, applicability, and ethical futures.

Wearability, Usability, and Daily Integration

Garment Preferences and Adaptive Design

Participants expressed various preferences regarding the design, material, and format of the smart garment. These preferences emphasize the value of adaptable design that can align with individual lifestyles, seasonal changes, and social settings.

One of the most frequently raised concerns was about high temperature during warm weather. While participants acknowledged that the current version of the smart garment featured long sleeves to facilitate wearability within the workshop context, some participants questioned its year-round practicality: "I was just wondering how practical it would be in the summer to wear when it is hot."

Participants further expressed their expectations for varied design options based on seasonal needs and daily routines:

> *In the winter, a vest would be something of choice. But in the summer, maybe a t shirt. But then obviously, you'd have to provide different styles.*

A couple of participants suggested to design the sensing garment into sleeveless styles such as vests and sports bras, which would allow users to retain their preferred outerwear while still benefiting from the system's embedded sensing functions:

> *I would want something that more comfortable and I can put my own clothes on. So for me, I would prefer a vest or something like a sports bra.*

Additionally, some older people concerned about the visibility of sensing components, such as metal press-fit snaps and the data acquisition unit. For example, too many metal snaps on the outside of clothing were considered potentially offensive, as participants indicated that they would not want others to know that they were wearing a system designed to monitor loneliness. This reflected broader sensitivities around emotional health and a strong preference for unobtrusive and socially invisible technologies. While some participants found these acceptable on activewear, they felt such elements appeared out of place on more casual garments like t-shirts.

These concerns also extended beyond aesthetic considerations, but also included social signaling and potential stigma. Participants expressed a preference for discreet designs that would not attract unnecessary attention or provoke inquiries:

> *In a perfect world, I'd like it to be invisible, because otherwise you may spend half your life explaining ... people are going to say "what's that?" And do you want to discuss the fact that you're being monitored for loneliness with people you're not necessarily that close to?*

These findings were further supported during the co-design sessions. Participants proposed alternative sensor attachment mechanisms beyond the current snap-on method, suggesting modular sensor units that could integrate with their existing personal clothing. Outcomes from the co-design activities included vest sketches with internal linings to conceal sensor modules and annotations, indicating preferences for "subtle seams" and "concealed fasteners." One participant also suggested the concept of a "pin-on sensor," reflecting a desire for wearables that conform to users' existing dressing habits, rather than imposing new ones. This highlights the importance of designing smart loneliness monitoring systems that seamlessly integrate into users' daily lives and personal style, thereby enhancing the likelihood of long-term adoption.

Sensor Placement and Alternative Technology

During the focus group discussions, participants expressed a range of concerns and preferences regarding the placement of sensors and the physical dimensions of the monitoring components. While most participants were satisfied with the flexibility and comfort of the electronic textile circuit, a key issue identified was the discomfort caused by the rigid data acquisition module within the sensing garment:

> *About the little hard board (data collection unit), that is to me kind annoying. It's just wherever you put it.*

Some participants also asked whether the hardware could be miniaturized and made less obtrusive, suggesting that improvements in physical design could significantly enhance adoption. This perspective was reflected in several co-design artifacts, where participants reimagined the data module as a smaller patch or accessory. During the co-design sessions, many

XSL•FO
RenderX

participants proposed relocating the data unit to the back or side of the garment, thereby reducing bulk and improving comfort around the front torso area.

Beyond physical discomfort, the focus group also revealed diverse preferences regarding sensor location, particularly in relation to wrist-based monitoring. Some participants clearly expressed aversion to wearing anything on their wrist: “I don’t wear anything on my wrist and I don’t particularly want to.”

In contrast, others found wrist-based sensing beneficial, especially those already familiar with commercial wearable devices: “I’m already using my watch to collect my health data, and I find it very convenient.”

These conflicting attitudes were also observed in the sensory ethnographic field notes, which recorded moments of hesitation and hand gestures when participants explored the heart rate sensor embedded within the sleeve of the prototype garment. This further highlights the tactile and cognitive responses that influence user acceptance.

Moreover, several participants raised concerns about the technical requirements for physiological data accuracy. One question reflected a broader concern about the need for reliable biometric sensing: “Do the garment and furniture sensors need to be in close contact with the body to get accurate results?”

This suggests a critical design tension in smart wearable technologies between comfort and data accuracy. Older users preferred sensor systems that offer customizability, modularity, and interchangeable placement. Future iterations of the system should therefore not only miniaturize key components but also offer multiple sensor placement options, enabling users to select configurations that best align with their comfort, personal habits, and lifestyle.

Modular Usability

Modularity is a core feature of the system prototype, designed to enhance ease of maintenance, personal adaptability, and user autonomy. Across focus group discussions, co-design outcomes, and sensory ethnographic observations, participants generally endorsed the modular design principle while also highlighting practical challenges related to charging, cleaning, and reassembly.

Some older participants raised concerns about whether individuals with physical limitations would be able to perform these tasks independently:

> *I didn’t have trouble taking the components out and putting them back, but if one hand wasn’t very agile, it would be hard to do. These steps really require both hands. For anyone older, or with arthritis, it might be difficult to pull these things out of such a small, tight pocket.*

This concern was also reflected in co-design artifacts, where participants proposed simplified fastening mechanisms or introduced concepts such as magnetic snap-in connectors to reduce the burden of fine motor control. Ethnographic observations further documented moments of hesitation, uncertainty, or participants seeking assistance when attempting to detach or reattach sensor modules.

Additionally, some participants suggested that the system should include charging notifications within the app:

> *It’s quite good to know that this stuff doesn’t need to be charged every day like a smartwatch, but should there be a reminder in the app? You know, because you don’t charge it daily, you might forget when it does need charging.*

Despite these operational challenges, participants consistently affirmed the value of modularity. They appreciated the separation of electronic components from the textile base not only for practical purposes, such as maintaining and laundering, but also for the potential long-term benefits, such as upgrading or replacing individual components over time. In co-design templates, several participants proposed personalized configurations, suggesting that different sensor modules could be swapped or added according to evolving health needs.

Material Comfort and Sensory Feedback

Material comfort is a critical factor influencing participants’ responses to the smart loneliness monitoring system, particularly in relation to fabric texture, thermal regulation, and skin contact. Several participants expressed discomfort with synthetic textiles, especially when worn directly against the skin:

> *Wearing it as an outer layer is fine, I think it’s quite comfortable. But if I were to wear it close to my skin, like a vest or a t-shirt, I definitely wouldn’t want it to be polyester, because I find polyester too hot and sweaty when worn directly against the skin.*

During the co-design activities, participants were given fabric samples, including 100% cotton, cotton blends, wool blends, and polyester-spandex, and they were invited to annotate their preferences directly onto garment outline templates. Cotton and cotton-blend fabrics were the most frequently selected, accompanied by annotations such as “soft,” “breathable,” and “not itchy.”

Comfort was also related to personal experiences and medical histories. For example, one participant noted that clothing design needs to consider changes in tactile sensitivity in the postoperative area:

> *I had a mastectomy. I don’t have any breasts. I hate anything that scratches. It’s just the normal is not normal anymore. You know what I mean?*

Compared to the sensing garments, the sensing furniture components such as the sensing pillow and cushion were generally perceived as more comfortable and less intrusive. This distinction was further supported by sensory ethnographic observations, which captured some participants’ nuanced physical interactions with the garments. While wearing the smart clothing, some participants often made subtle adjustments to collars, pulled at sleeves, or ran their fingers along seam lines. Some participants hesitated before putting on the garment or asked whether the sensors would touch their skin directly. In contrast, interactions with the smart furniture were more relaxed. Most participants sat down without instruction, leaned back comfortably, and engaged in conversation while using the seat pad.

XSL•FO
RenderX

Contextual Integration and Lifestyle Compatibility

Participant feedback indicated that the acceptability of the system was closely related to its ability to seamlessly integrate into users' everyday lives, domestic environments, and personal routines. For example, smart furniture components were generally perceived as less intrusive and more acceptable for long-term and low-effort engagement:

> *People do tend to sit in the same seat every day, in the same place to watch TV. If someone was sitting there all day watching TV or doing something else, you'd find it very useful.*

In contrast, concerns were raised about the disruption that smart garments might cause to the unpredictability of daily routines. One participant described how the demand for continuous wear might not be compatible with their lifestyle:

> *I do so many things during the day. When I come back from the garden, I might have sweated or gotten dirty and need to change clothes and then, oops, I might forget to put on the smart garment again.*

This issue was further reflected in sensory ethnographic field notes, which captured several participants expressing uncertainty about whether they were "wearing it correctly" or whether the sensors would still function properly after shifting position.

During the co-design sessions, participants engaged with home layout templates, marking preferred sensor locations. They often placed sensors in areas associated with habitual furniture use, such as a reading chair, dining table, or frequently used seating areas. Some participants also proposed to integrate the system as part of standard domestic infrastructures in care homes: "You just install it when someone moves in for safety."

These insights suggest that smart loneliness monitoring systems should not only be conceived as stand-alone technologies, but rather as components within a broader ecosystem of smart living, with potential to be embedded into existing domestic practices and infrastructural frameworks.

Trust, Privacy, and Data Control

Conditional Trust and System Reliability

Participants across both workshops expressed a degree of trust in the smart loneliness monitoring system but consistently emphasized that trust in such technologies is not taken for granted. Several participants highlighted that trust would need to be earned over time, through demonstrable functionality and accuracy in real-life use:

> *I think I'd have to actually wear it and see it identify (my loneliness) without me saying anything. That would be the only way to really learn that it was working. What would happen at the end of the day? Would it give you a ping? Something to say "woo! Looks like you're lonely at the moment." And if I was in the middle of having a conversation with somebody and feeling perfectly fine, I know I would not trust it.*

Moreover, past experiences with commercial wearable monitoring devices appeared to shape users' current skepticism. One participant referenced their partner's experience:

> *My husband uses a Fitbit, but it clearly doesn't record his steps accurately. So I can't fully trust it.*

These encounters with inaccurate sensing may contribute to reserved user fatigue or doubts about the credibility of wearable sensors. These doubts also extended to smart monitoring systems, especially when applied to emotionally complex and subjective states like loneliness.

While most participants felt confident interpreting the outputs presented in the accompanying mobile app, they nonetheless highlighted a need for greater transparency and interpretability in system feedback. During the co-design sessions, participants proposed a range of suggestions to improve algorithmic explainability, including the addition of visual indicators such as:

> *Why is it saying I'm lonely?*
>
> *What data triggered this message?*

Another recommendation was the inclusion of a feedback confirmation mechanism, enabling users to validate the system feedback, thus contributing to a dynamic and trust-building model. One participant proposed a "check-in" feature, whereby if the system identified them as potentially lonely, they could choose to confirm or dismiss the notification. Over time, such feedback loops would allow the algorithm to learn from user responses, thereby refining its accuracy and building user confidence.

Customizable Data-Sharing Preferences

While participants acknowledged the potential value of sharing data with others particularly in the context of health or emotional support, several older adults highlighted the importance of retaining control over their data-sharing choices:

> *I feel that collecting this data all the time is quite intrusive. I might not want my daughter to know how I'm feeling at that moment, and I certainly don't want her using an app to monitor me. I want to stay in control of my emotions, and I feel this would take that control away.*

Conversely, another participant expressed openness to sibling-based support, provided that geographic distance warranted it. Interestingly, when asked, "You wouldn't want your daughter to see your data, would you want to see your mother's?" the participant hesitated. These responses highlight how data sharing preferences are relational and context-dependent and may vary depending on the role of the recipient as caregiver or care recipient.

Additionally, several participants expressed a preference for conditional sharing models, where data would only be shared under specific circumstances:

> *If I'm having a breakdown at home because I'm feeling lonely and haven't seen anyone for three weeks, then I do want them to know. But if I'm just feeling a bit off and can't be bothered to go to the coffee shop, I don't necessarily want to share that.*

During the co-design sessions, participants proposed improvements to the current permissions interface. Users expressed a desire to select which types of data such as

XSL•FO
RenderX

emotional states, activity levels, or physiological indicators would be visible to specific recipients, including family members, clinicians, or community caregivers. Some also suggested adding a dashboard that clearly shows "who can see what," along with visual indicators to support transparency and ease of management.

Ethical Concerns and Data Ownership

Across both workshops, while the majority of participants expressed openness toward the use of monitoring technologies, they also emphasized that trust in such systems depends not only on accuracy but also on transparency of purpose, data flow, and long-term governance.

A recurring concern was "Who owns the data?" and often followed by anxiety about potential commercial exploitation. As one participant asked: "What if the data gets sold? Who's to say it won't be sold?"

The prospect that personal health or emotional data might be commodified was troubling for many older adults, particularly given the lack of clear regulation surrounding data collected outside clinical systems.

Concerns also extended to the broader implications of artificial intelligence governance in connection with wearable monitoring systems: "I understand that these technologies are developed with good intentions, but I do worry that they could be repurposed for surveillance or behavioral manipulation."

During the co-design sessions, many older adults indicated that these ethical concerns did not necessarily result in rejection of the system. Rather, they expressed a desire for greater clarity regarding the system's governance model, data stewardship, and effective plans for future use. These findings suggest that ethical acceptability is not solely a matter of obtaining "informed consent" at the point of use but requires ongoing transparency and participatory data governance. Future iterations of the system should consider the development of interactive tools that clearly and accessibly communicate key information regarding data provenance ownership and rights.

Perceptions of Loneliness and the Limits of Detection

The Subjectivity of Loneliness

Some participants believed that loneliness is fundamentally a subjective and emotionally complex experience, one that cannot be directly inferred from behavioral or physiological signals alone. This perspective occurred in our earlier focus group discussions, where several participants questioned the assumption that sensor data could reliably infer emotional states: "What you feel inside can't be monitored by anyone, no sensor can pick that up."

Participants also noted that loneliness is highly individualized and not necessarily linked to physical solitude: "You can feel lonely in a crowded room but feel fine when you're alone."

This further raised questions among participants about whether the algorithm could accurately identify individualized feelings of loneliness. In response, the investigator (JR) clarified that while current algorithmic models may be developed from broad datasets, their core functionality is intended to adapt to individual patterns. The system is designed to learn over time with correct or incorrect feedback, establishing and refining a personalized emotional profile, thereby improving its accuracy.

Moreover, participants noted that loneliness is not exclusive to older adults but occurs across different ages and life stages, highlighting the need for the system to detect emotional nuance rather than demographic generalization.

In the co-design sessions, some participants expressed discomfort with the term "loneliness," describing it as "too strong" or "too negative." They proposed softer alternatives, such as "reflective state" or "well-being indicator." Others annotated their interface sketches with prompts like "How are you feeling today?" in place of system-generated loneliness labels. These design annotations suggest that users may prefer tools that prompt self-reflection, rather than systems that presume to define their emotional states on their behalf.

Algorithmic Assumptions and Multimodal Analysis

As discussed in the previous subsection, participants questioned the logic behind the algorithm used for loneliness detection. For instance, they expressed concern that feedback based solely on low physical activity might lead to false positives: "You're detected as being very still, but you're not lonely, you're just enjoying your book."

In response, researchers explained that the current system applies multimodal integration, combining physiological signals with behavioral data for a more comprehensive analysis. While participants appreciated this approach, they also asked whether the system could incorporate additional personalized factors to further improve detection accuracy: "If I'm motionless, and those factors you mentioned lead the system to assume I'm lonely, then I'm wondering, what other factors could be added?"

This desire for individualized calibration was also evident in the co-design artifacts. On the smart garment and furniture templates, some participants annotated notes such as "Don't assume lack of movement means I'm feeling low." Others, particularly older adults, proposed that the system should seek user confirmation before tagging as a loneliness event, effectively embedding emotional subjectivity into the algorithmic process.

Adoption, Applicability, and Ethical Futures

Adoption Among the Older Adults

While participants generally recognized the value of the systems in their later life, they also noted that those who might benefit most should be very old adults or individuals strongly invested in their independence (may be the least willing to adopt it). Factors such as psychological resistance and identity-related concerns were seen as major potential barriers to real-world implementation.

Most participants in both of the workshops were still socially active and engaged but reflected on family members who were socially isolated yet unwilling to engage with monitoring technologies or accept support:

XSL•FO
RenderX

> *My mum is very old and lives on her own. She's obviously lonely, but there's no way she'd get involved in something like this.*

Others expressed similar insight:

> *She won't accept any help. Yes, even though she's clearly spiraling because of it.*

These reflections highlight a tension between need, self-perception, and autonomy. The concept of "independent living" was both a source of pride and a practical barrier. Older adults may perceive acknowledging loneliness or using assistive technologies as a threat to their autonomy or even as an admission of vulnerability.

This resistance was also observed in the sensory ethnographic field notes. While most participants interacted with system components during the experience, a small number showed hesitation or disengagement when asked to wear the garment or respond to app prompts. Some made dismissive remarks such as: "This seems something for people who need looking after."

In co-design sessions, some participants openly stated that they found it a bit difficult to imagine themselves using such a system even hypothetically because it felt "irrelevant" or "only for people in worse situations." These views may reflect a desire to maintain a sense of competence and independence, even in the face of known risks or increasing social withdrawal.

Timing of App

During the focus group discussions, participants also reflected on when these technologies should be applied. Some participants thought it would be best to introduce the system before severe loneliness or functional decline occurs: "Maybe you need to catch people before they get to that point, so you can do it as a prevention rather than a late intervention."

Some participants thought that if people fall into severe isolation, their willingness to engage with new technologies may be significantly reduced: "By the time they need it, they may not want to learn it."

In addition, participants also mentioned that they may be more willing to use the system when they still feel in control, an insight that is consistent with the discussion in the Trust, Privacy, and Data Control section that trust is built through gradual voluntary participation rather than sudden or mandatory use.

Some participants proposed a "onboarding stage" during the co-design session, such as "Let me start with one feature first" or "Phase 1: Activity tracking only, no reminders yet." Others suggested setting up a "trial period" for the system so that users can experience it during this period without worrying about data being misunderstood or shared too early.

Expanded Health Monitoring

While loneliness detection was the primary aim of the system, many participants expressed strong interest in expanding its functionality to support broader physiological and behavior monitoring: "It's good to know the system can monitor so many different things, but could it also include more features?"

Some participants also highlighted their specific personal monitoring needs: "I often need to drink water, otherwise I get heart palpitations. could it tell me if I'm dehydrated?"

These ideas were further developed during the co-design activities, where participants added additional monitoring points to the smart garment framework diagrams. Suggestions included integrating electrocardiography and blood pressure tracking and hydration-level detection. Rather than replacing the core functionality related to loneliness, these suggestions were seen as complementary integrations that could increase the system's everyday utility and relevance by addressing users' broader well-being in a more holistic and meaningful way.

Linked Intervention

When participants reflected on the system's potential to monitor physiological and behavioral signals, the discussion naturally extended to scenarios involving health emergencies and critical incidents. Many raised concerns about automated alerts and connected interventions:

> *If you're in a state of severe loneliness or at some kind of risk, the next level of concern is whether social services would actually respond, whether your GP would be notified, or a nurse, or a district nurse would come out. That's the real worry.*

While many participants appreciated the system's potential to issue alerts or prompts, they expressed hesitancy about fully autonomous system actions, particularly in the context of emotional monitoring. Some expressed discomfort with the idea of automated triage, instead showing a preference for human-mediated intervention: "I would like a caregiver or clinician to review the data before contacting me."

In the co-design sessions, participants proposed customizable alert settings, including adjustable emergency thresholds, such as "Only trigger an alert if an abnormal signal persists for more than 10 minutes." and "Notify family members first, then professionals." These suggestions reflect a strong desire for tiered intervention logic, whereby users can define the severity of signals required to activate alerts as well as the order and type of recipients to be notified. This approach highlights the need for a personalized response, rather than a one-size-fits-all automation model.

Furthermore, although stakeholders such as care providers were not present in either workshop, participants actively imagined various relational configurations. Some participants preferred family members as first contacts, while others, particularly those living alone, favored designated professional care networks.

Discussion

Principal Findings

This study examined the acceptability and perceived usability of a novel smart loneliness monitoring system for older adults, comprising sensing garments, furniture, and a companion mobile app. Through user-centered evaluation and experience-based co-design, this study aimed to comprehensively explore users' practical, emotional, and ethical concerns to the system and to provide actionable design insights for future development. While

XSL•FO
RenderX

prior studies in pervasive computing and ambient-assisted living have focused on detecting behavioral correlates of social isolation, few have investigated how older adults themselves experience, interpret, and negotiate such monitoring systems [34-36]. Therefore, our findings contributed to bridging the gap between technical feasibility and user acceptability in the emerging field of smart mental health textiles.

Previous research has shown that older adults' acceptance of smart monitoring systems is heavily affected by perceived usefulness, particularly whether the data collected support meaningful or supportive interventions [17,56,57]. Given the practical and societal significance of loneliness monitoring, our prior research had explored older adults' initial design needs alongside stakeholder perspectives at the conceptual level [27]. Building on this foundation, this study combines smart textile design and sensing technology to develop prototypes that allow participants to physically experience, evaluate, and reimagine the system. This study identified 4 main domains affecting user acceptance and future design improvement, including wearability, usability and integration; trust, privacy and data control; limitations of loneliness monitoring; and future adoption, applicability, and ethical considerations.

Our findings highlighted the importance of adaptability and lifestyle compatibility in determining system acceptability, which are key aspects emphasized in previous research on wearable health technologies for older adults. Prior studies have shown that comfort, convenience, and discretion strongly influence engagement with wearable devices, particularly among older users who may have heightened sensitivities to fabric texture or skin contact [15]. In our study, although most participants were satisfied with synthetic materials for outerwear use, individual preferences and health history highlighted the value of personalized textile options. Participants appreciated the modular design and comfortable electronic textile integration, which enhanced wearability. However, unlike previous work that primarily assessed ergonomics or fit, our findings underscored the importance of emotional comfort and social invisibility. Participants expressed concerns about the visibility of sensing components and the potential stigma associated with being perceived as "monitored." This highlights an important extension of the current understanding of usability from physical comfort to psychosocial comfort. Additionally, the discussion around seasonal practicality and clothing preferences, such as the suggestion to adopt undergarment formats like vests or sports bras, introduces a novel consideration for thermal and social appropriateness and compatibility with individual everyday routines. While some previous smart garment research explored aesthetic design [58], our findings suggest that adaptability to seasonal routines and existing clothing habits is critical for long-term adoption. Moreover, the discomfort caused by rigid sensor modules further emphasizes the need for miniaturized and flexible electronics, a challenge also noted in emerging literature on e-textile scalability and integration [59,60]. Compared to sensing garments, sensing furniture was perceived as more comfortable, less obtrusive, and better aligned with habitual behaviors.

Trust, privacy, and data control were also key to user acceptance. Prior studies on digital health and remote monitoring technologies have consistently shown that trust and perceived data security are prerequisites for sustained engagement among older adults [61]. Our findings support these observations but extend them by revealing that participants' conditional trust depended not only on privacy assurances but also on the reliability, interpretability, and transparency of system feedback. These aspects highlighted in this study collaborated with previous studies on smart home and telehealth systems. However, unlike earlier work that primarily emphasized the role of institutional trust in health care providers [62], our participants focused on personal data sovereignty, and they want to see, understand, and adjust what the system infers about them in real time. While many older adults found the app's real-time feedback intuitive and easy to use, their past experiences with commercial wearables, such as smartwatches and fitness trackers, triggered skepticism regarding its accuracy when interpreting subtle emotional or behavioral changes. This echoes concerns in prior research that algorithmic opacity undermines user confidence in affective or well-being monitoring [63]. Our findings show that trust must be earned gradually through use and supported by transparent feedback mechanisms that allow users to confirm or challenge system outputs, which is an important element rarely addressed and discussed in earlier studies. Users have also shown a great preference to granular control over data sharing, allowing users to tailor access to different stakeholders such as family members, clinicians, or caregivers. While prior work on privacy in older adults has discussed consent management in general terms [64], our participants preferred a dynamic and contextual control that could match their current mental state and relationships. These insights extend the current literature by emphasizing that ethical acceptability is not achieved solely through initial consent, but through ongoing transparency, accountability, and user agency in data governance. Future development of wearable monitoring technologies should move beyond data protection compliance to include user-facing transparency features and active participatory data management frameworks.

Strengths and Limitations

The key strengths of this study lie in its multimethod integration of quantitative and qualitative approaches, combining structured questionnaires, focus groups, sensory ethnography, and experience-based co-design. This enabled a multifaceted understanding of older adults' experiences with the smart loneliness monitoring system and helped identify comprehensive user-driven directions for iterative improvement. However, several limitations also need to be acknowledged. First, all the participants were recruited from the United Kingdom, which may introduce a degree of regional bias and limit the generalizability of the findings to older adults in other geographic, cultural, or health care settings. In addition, the sample included relatively few male participants. Given that sex may influence perspectives on technology, privacy, and well-being, future research should aim to increase sex diversity to improve the representativeness of findings. Participants with diagnosed cognitive impairment were excluded during recruitment. Researchers applied the principles of the Mental Capacity Act (2005) to assess participants' capacity to understand the study and provide informed consent [40]. While

XSL•FO
RenderX

this ensured ethical participation and meaningful engagement with the system, the study does not include the perspectives of older adults living with cognitive impairment or dementia. Future work should explore how smart loneliness monitoring systems can be adapted or tailored for these populations, who may have different usability needs and vulnerabilities. Additionally, although participants met the inclusion criteria of being aged 65 years and older and having experienced loneliness, the majority of participants remained socially active, potentially limiting the generalizability of findings to more isolated or vulnerable populations. Moreover, the overall sample size was small. While small-scale qualitative studies can offer rich insight, findings should be interpreted as exploratory and hypothesis-generating rather than definitive. Finally, while the workshops were conducted in a controlled environment to allow in-depth interaction and observation, this setting may not fully capture the complexity of real-world use. To further strengthen the validity and practical relevance of the system, future research should involve longitudinal field testing in home or community settings. This would allow for continuous data collection over an extended period, enabling more robust analysis of behavioral and physiological patterns, as well as user engagement over time.

Conclusions

This study examined older adults' experiences with smart loneliness monitoring systems, including sensing garments, furniture, and a companion mobile app. Through immersive workshops combining structured questionnaires, focus groups, sensory ethnographic observation, and experience-based co-design, we investigated older users' practical, emotional, and ethical responses to the system, providing actionable design insights to inform future development. These findings indicate that while participants generally accepted the concept of monitoring via smart textile wearables and furniture, their willingness to adopt such systems over time is highly dependent on usability, personalization, lifestyle compatibility, and perceived control. Participants consistently emphasized the importance of adaptable design that respects bodily comfort, domestic routines, and personal identity. Moreover, the modular smart system developed in this study demonstrates strong potential as a discreet and passive sensing platform for psychological and social health indicators. However, continued user-centered iteration, broader real community testing, and deeper integration with health care infrastructure are required to ensure its future success in real-world applications.

Acknowledgments

The authors would like to thank all participants and collaborators who contributed to this study.

Funding

This work was supported by the UK Engineering and Physical Sciences Research Council and the National Institute of Health and Care Research (grant EP/W031434/1).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

YZ was responsible for conceptualization, methodology, investigation, formal analysis, data curation, writing the original draft, and editing the manuscript. JR was responsible for recruitment, formal analysis, data curation, and reviewing and editing the manuscript. FM was responsible for conceptualization, methodology, and reviewing and editing the manuscript. AP was responsible for conceptualization and reviewing and editing the manuscript. MA and AT was responsible for funding acquisition and reviewing and editing the manuscript. SO was responsible for supervision and funding acquisition. WL was responsible for supervision, funding acquisition, resources, and reviewing and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Structured observation form.
[PDF File (Adobe PDF File), 32 KB - jmir_v28i1e81027_app1.pdf]

Multimedia Appendix 2
Likert-scale questionnaire.
[PDF File (Adobe PDF File), 1514 KB - jmir_v28i1e81027_app2.pdf]

Multimedia Appendix 3
Co-design template.

XSL•FO
RenderX

[PDF File (Adobe PDF File), 340 KB - jmir_v28i1e81027_app3.pdf]

References

1. Guglielmi V, Colangeli L, Parrotta ME, Ciammariconi A, Milani I, D'Adamo M, et al. Social isolation and loneliness in non-communicable chronic diseases: impact of COVID-19 pandemic, population aging and technological progress. Nutr Metab Cardiovasc Dis 2025;35(6):104015 [FREE Full text] [doi: 10.1016/j.numecd.2025.104015] [Medline: 40189996]
2. Walsh BE, Rottenberg J, Schlauch RC. Why loneliness requires a multidimensional approach: a critical narrative review. Nat Mental Health 2025;3(2):175-184. [doi: 10.1038/s44220-024-00382-3]
3. Teo RH, Cheng WH, Cheng LJ, Lau Y, Lau ST. Global prevalence of social isolation among community-dwelling older adults: a systematic review and meta-analysis. Arch Gerontol Geriatr 2023;107:104904. [doi: 10.1016/j.archger.2022.104904] [Medline: 36563614]
4. Aalto UL, Bonin-Guillaume S. Loneliness among older people exacerbated by the COVID-19 pandemic. J Nutr Health Aging 2023;27(8):617-618. [doi: 10.1007/s12603-023-1968-z]
5. Benson JA, McSorley VE, Hawkley LC, Lauderdale DS. Associations of loneliness and social isolation with actigraph and self-reported sleep quality in a national sample of older adults. Sleep 2021;44(1):zsaa140 [FREE Full text] [doi: 10.1093/sleep/zsaa140] [Medline: 32691067]
6. Emerson E, Fortune N, Llewellyn G, Stancliffe R. Loneliness, social support, social isolation and wellbeing among working age adults with and without disability: cross-sectional study. Disabil Health J 2021;14(1):100965 [FREE Full text] [doi: 10.1016/j.dhjo.2020.100965] [Medline: 32843311]
7. Latikka R, Rubio-Hernández R, Lohan ES, Rantala J, Nieto Fernández F, Laitinen A, et al. Older adults' loneliness, social isolation, and physical information and communication technology in the era of ambient assisted living: a systematic literature review. J Med Internet Res 2021;23(12):e28022. [doi: 10.2196/28022] [Medline: 34967760]
8. Mushtaq R, Shoib S, Shah T, Mushtaq S. Relationship between loneliness, psychiatric disorders and physical health? A review on the psychological aspects of loneliness. J Clin Diagn Res 2014;8(9):WE01-WE04 [FREE Full text] [doi: 10.7860/JCDR/2014/10077.4828] [Medline: 25386507]
9. Prabhu D, Kholghi M, Sandhu M, Lu W, Packer K, Higgins L, et al. Sensor-based assessment of social isolation and loneliness in older adults: a survey. Sensors (Basel) 2022 Dec 16;22(24):9944 [FREE Full text] [doi: 10.3390/s22249944] [Medline: 36560312]
10. Fakoya OA, McCorry NK, Donnelly M. Loneliness and social isolation interventions for older adults: a scoping review of reviews. BMC Public Health 2020;20(1):129 [FREE Full text] [doi: 10.1186/s12889-020-8251-6] [Medline: 32054474]
11. Ratcliffe J, Kanaan M, Galdas P. Reconceptualising men's loneliness: an interpretivist interview study of UK-based men. Soc Sci Med 2023;332:116129 [FREE Full text] [doi: 10.1016/j.socscimed.2023.116129] [Medline: 37531906]
12. Rees J, Liu W, Canson J, Crosby L, Tinker A, Probst F, et al. Qualitative exploration of the lived experiences of loneliness in later life to inform technology development. Int J Qual Stud Health Well-Being 2024;19(1):2398259 [FREE Full text] [doi: 10.1080/17482631.2024.2398259] [Medline: 39305060]
13. Johnson KT, Zawadzki MJ, Kho C. Loneliness and sleep in everyday life: using ecological momentary assessment to characterize the shape of daily loneliness experience. Sleep Health 2024;10(4):508-514 [FREE Full text] [doi: 10.1016/j.sleh.2024.04.003] [Medline: 38839482]
14. Gedam S, Paul S. A review on mental stress detection using wearable sensors and machine learning techniques. IEEE Access 2021;9:84045-84066. [doi: 10.1109/access.2021.3085502]
15. Moore K, O'Shea E, Kenny L, Barton J, Tedesco S, Sica M, et al. Older adults' experiences with using wearable devices: qualitative systematic review and meta-synthesis. JMIR Mhealth Uhealth 2021;9(6):e23832 [FREE Full text] [doi: 10.2196/23832] [Medline: 34081020]
16. Schrempft S, Jackowska M, Hamer M, Steptoe A. Associations between social isolation, loneliness, and objective physical activity in older men and women. BMC Public Health 2019;19(1):74 [FREE Full text] [doi: 10.1186/s12889-019-6424-y] [Medline: 30651092]
17. Meredith SJ, Cox N, Ibrahim K, Higson J, McNiff J, Mitchell S, et al. Factors that influence older adults' participation in physical activity: a systematic review of qualitative studies. Age Ageing 2023;52(8):afad145. [doi: 10.1093/ageing/afac125.002]
18. Frayn M, Livshits S, Knäuper B. Emotional eating and weight regulation: a qualitative study of compensatory behaviors and concerns. J Eat Disord 2018;6:23 [FREE Full text] [doi: 10.1186/s40337-018-0210-6] [Medline: 30221002]
19. Hawkley LC, Cacioppo JT. Loneliness and blood pressure in older adults: defining connections. Aging Health 2010;6(4):415-418. [doi: 10.2217/ahe.10.41]
20. Gao F, Liu C, Zhang L, Liu T, Wang Z, Song Z, et al. Wearable and flexible electrochemical sensors for sweat analysis: a review. Microsyst Nanoeng 2023;9:1-21 [FREE Full text] [doi: 10.1038/s41378-022-00443-6] [Medline: 36597511]
21. Schutter N, Holwerda T, Stek M, Dekker J, Rhebergen D, Comijs H. Loneliness in older adults is associated with diminished cortisol output. J Psychosom Res 2017;95:19-25. [doi: 10.1016/j.jpsychores.2017.02.002]
22. Ejupi A, Menon C. Detection of talking in respiratory signals: a feasibility study using machine learning and wearable textile-based sensors. Sensors (Basel) 2018;18(8):2474 [FREE Full text] [doi: 10.3390/s18082474] [Medline: 30065177]

XSL•FO
RenderX

23. Kim M, Hong S, Youm S. Development of an intrinsic health risk prediction model for camera-based monitoring of older adults living alone. Sci Rep 2022;12(1):18855 [FREE Full text] [doi: 10.1038/s41598-022-23663-2] [Medline: 36344806]
24. Vuegen L, Van DBB, Karsmakers P, Van HH, Vanrumste B. Monitoring activities of daily living using Wireless Acoustic Sensor Networks in clean and noisy conditions. 2015 Nov 05 Presented at: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); August 25-29, 2015; Milan, Italy p. 4966-4969. [doi: 10.1109/embc.2015.7319506]
25. van den Boer J, van der Lee A, Zhou L, Papapanagiotou V, Diou C, Delopoulos A, et al. The SPLENDID eating detection sensor: development and feasibility study. JMIR Mhealth Uhealth 2018;6(9):e9781 [FREE Full text] [doi: 10.2196/mhealth.9781] [Medline: 30181111]
26. Kalisch T, Theil C, Gosheger G, Ackmann T, Schoenhals I, Moellenbeck B. Measuring sedentary behavior using waist- and thigh-worn accelerometers and inclinometers—are the results comparable? Ther Adv Musculoskelet Dis 2022;14:1759720X221079256 [FREE Full text] [doi: 10.1177/1759720X221079256] [Medline: 35310836]
27. Probst F, Rees J, Aslam Z, Mexia N, Molteni E, Matcham F, et al. Evaluating a smart textile loneliness monitoring system for older people: co-design and qualitative focus group study. JMIR Aging 2024;7:e57622 [FREE Full text] [doi: 10.2196/57622] [Medline: 39688889]
28. Probst F, Liu W. Co-design of a loneliness monitoring system with older people and stakeholders. 2024 Presented at: DRS Biennial Conference Series; June 23-28, 2024; Boston, MA, United States URL: https://dl.designresearchsociety.org/drs-conference-papers/drs2024/researchpapers/134 [doi: 10.21606/drs.2024.545]
29. Dzedzickis A, Kaklauskas A, Bucinskas V. Human emotion recognition: review of sensors and methods. Sensors (Basel) 2020;20(3):592 [FREE Full text] [doi: 10.3390/s20030592] [Medline: 31973140]
30. Cosma G, Brown D, Battersby S, Kettley S, Kettley R. Analysis of multimodal data obtained from users of smart textiles designed for mental wellbeing. 2017 Aug 14 Presented at: 2017 International Conference on Internet of Things for the Global Community (IoTGC); July 10-13, 2017; Funchal, Portugal. [doi: 10.1109/iotgc.2017.8008974]
31. Khundaqji H, Hing W, Furness J, Climstein M. Smart shirts for monitoring physiological parameters: scoping review. JMIR Mhealth Uhealth 2020;8(5):e18092 [FREE Full text] [doi: 10.2196/18092] [Medline: 32348279]
32. Chen M, Ma Y, Song J, Lai C, Hu B. Smart clothing: connecting human with clouds and big data for sustainable health monitoring. Mobile Netw Appl 2016;21(5):825-845. [doi: 10.1007/s11036-016-0745-1]
33. Vuohijoki T, Ihalainen T, Virkki J. Smart clothing and furniture for supporting participation-co-creation concepts for daily living. SN Appl Sci 2023;5(4):110. [doi: 10.1007/s42452-023-05315-w]
34. Lodder GMA, Scholte RHJ, Goossens L, Engels RCME, Verhagen M. Loneliness and the social monitoring system: emotion recognition and eye gaze in a real-life conversation. Br J Psychol 2016;107(1):135-153. [doi: 10.1111/bjop.12131] [Medline: 25854912]
35. Jafarlou S, Azimi I, Lai J, Wang Y, Labbaf S, Nguyen B, et al. Objective monitoring of loneliness levels using smart devices: a multi-device approach for mental health applications. PLoS One 2024;19(6):e0298949 [FREE Full text] [doi: 10.1371/journal.pone.0298949] [Medline: 38900745]
36. Sarhaddi F, Azimi I, Niela-Vilen H, Axelin A, Liljeberg P, Rahmani AM. Maternal social loneliness detection using passive sensing through continuous monitoring in everyday settings: longitudinal study. JMIR Form Res 2023;7:e47950 [FREE Full text] [doi: 10.2196/47950] [Medline: 37556183]
37. Fernández-Caramés T, Fraga-Lamas P. Towards the internet-of-smart-clothing: a review on IoT wearables and garments for creating intelligent connected e-textiles. Electronics 2018;7(12):405. [doi: 10.3390/electronics7120405]
38. Tsai T, Lin W, Chang Y, Chang P, Lee M. Technology anxiety and resistance to change behavioral study of a wearable cardiac warming system using an extended TAM for older adults. PLoS One 2020;15(1):e0227270 [FREE Full text] [doi: 10.1371/journal.pone.0227270] [Medline: 31929560]
39. Rees J, Liu W, Ourselin S, Shi Y, Probst F, Antonelli M, et al. Understanding the psychological experiences of loneliness in later life: qualitative protocol to inform technology development. BMJ Open 2023;13(6):e072420 [FREE Full text] [doi: 10.1136/bmjopen-2023-072420] [Medline: 37336536]
40. Mental Capacity Act 2005. Statute Law Database. URL: https://www.legislation.gov.uk/ukpga/2005/9/section/1 [accessed 2025-12-23]
41. Hughes ME, Waite LJ, Hawkley LC, Cacioppo JT. A short scale for measuring loneliness in large surveys: results from two population-based studies. Res Aging 2004;26(6):655-672 [FREE Full text] [doi: 10.1177/0164027504268574] [Medline: 18504506]
42. Courage C, Baxter K. A Practical Guide to User Requirements Methods, Tools, and Techniques. Houston, TX: Gulf Professional Publishing; 2005.
43. Thell M, Edvardsson K, Aljeshy R, Ibrahim K, Warner G. A trauma support app for young people: co-design and usability study. JMIR Form Res 2025 Mar 18;9:e57789. [doi: 10.2196/57789]
44. Grieve N, Braaten K, MacPherson M, Liu S, Jung ME. Involving end users in the development and usability testing of a smartphone app designed for individuals with prediabetes: mixed-methods focus group study. JMIR Form Res 2025;9:e59386 [FREE Full text] [doi: 10.2196/59386] [Medline: 39935015]

XSL•FO
RenderX

45. Steen-Olsen EB, Pappot H, Hjerming M, Hanghoej S, Holländer-Mieritz C. Monitoring adolescent and young adult patients with cancer via a smart t-shirt: prospective, single-cohort, mixed methods feasibility study (OncoSmartShirt Study). JMIR Mhealth Uhealth 2024;12:e50620 [FREE Full text] [doi: 10.2196/50620] [Medline: 38717366]
46. Lupton D. Editorial: Towards sensory studies of digital health. Digit Health 2017;3:2055207617740090. [doi: 10.1177/2055207617740090]
47. Valtonen A, Markuksela V, Moisander J. Doing sensory ethnography in consumer research. Int J Consumer Studies 2010 Jun 09;34(4):375-380. [doi: 10.1111/j.1470-6431.2010.00876.x] [Medline: 22489612]
48. Sunderland N, Bristed H, Gudes O, Boddy J, Da Silva M. What does it feel like to live here? Exploring sensory ethnography as a collaborative methodology for investigating social determinants of health in place. Health Place 2012;18(5):1056-1067. [doi: 10.1016/j.healthplace.2012.05.007] [Medline: 22722015]
49. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989;13(3):319-340. [doi: 10.2307/249008]
50. Brooke J. SUS: A Quick and Dirty Usability Scale. Boca Raton, FL: CRC Press; 1996.
51. Venkatesh, Morris, Davis, Davis. User acceptance of information technology: toward a unified view. MIS Q 2003;27(3):425. [doi: 10.2307/30036540]
52. Donetto S, Pierri P, Tsianakas V, Robert G. Experience-based co-design and healthcare improvement: realizing participatory design in the public sector. Des J 2015;18(2):227-248. [doi: 10.2752/175630615X14212498964312]
53. Mishra P, Pandey C, Singh U, Gupta A, Sahu C, Keshri A. Descriptive statistics and normality tests for statistical data. Ann Card Anaesth 2019;22(1):67. [doi: 10.4103/aca.aca_157_18]
54. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
55. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa] [Medline: 32100154]
56. Brauner P, Heek JV, Ziefle M. Age, gender, and technology attitude as factors for acceptance of smart interactive textiles in home environments—towards a smart textile technology acceptance model. 2017 Presented at: Proceedings of the 3rd International Conference on Information and Communication Technologies for Ageing Well and e-Health ICT4AWE; April 28-29, 2017; Porto, Portugal p. 13-24 URL: https://www.scitepress.org/Link.aspx?doi=10.5220/0006255600130024 [doi: 10.5220/0006255600130024]
57. Choi YK, Thompson HJ, Demiris G. Use of an internet-of-things smart home system for healthy aging in older adults in residential settings: pilot feasibility study. JMIR Aging 2020 Nov 10;3(2):e21964 [FREE Full text] [doi: 10.2196/21964] [Medline: 33170128]
58. Imbesi S, Scataglini S. A user centered methodology for the design of smart apparel for older users. Sensors (Basel) 2021;21(8):2804 [FREE Full text] [doi: 10.3390/s21082804] [Medline: 33923514]
59. Meena JS, Choi SB, Jung S, Kim JW. Electronic textiles: new age of wearable technology for healthcare and fitness solutions. Mater Today Bio 2023;19:100565 [FREE Full text] [doi: 10.1016/j.mtbio.2023.100565] [Medline: 36816602]
60. Zhou Y, Ratcliffe J, Molteni E, Patel A, Liu J, Mexia N, et al. Smart textile systems for loneliness monitoring in older people care: a review of sensing and design innovations. Adv Elect Mater 2025;11(16):e00300. [doi: 10.1002/aelm.202500300]
61. Kebede AS, Ozolins L, Holst H, Galvin K. Digital engagement of older adults: scoping review. J Med Internet Res 2022;24(12):e40192. [doi: 10.2196/40192] [Medline: 36477006]
62. Hou G, Li X, Wang H. How to improve older adults' trust and intentions to use virtual health agents: an extended technology acceptance model. Humanit Soc Sci Commun 2024 Dec 18;11(1):1677. [doi: 10.1057/s41599-024-04232-6]
63. Wanner J, Herm L, Heinrich K, Janiesch C. The effect of transparency and trust on intelligent system acceptance: evidence from a user-based study. Electron Mark 2022;32(4):2079-2102. [doi: 10.1007/s12525-022-00593-5]
64. Altawalbeh SM, Alkhateeb FM, Attarabeen OF. Ethical issues in consenting older adults: academic researchers and community perspectives. J Pharm Health Serv Res 2020;11(1):25-32 [FREE Full text] [doi: 10.1111/jphs.12327] [Medline: 33042231]

Abbreviations

TAM: technology acceptance model

XSL•FO
RenderX

Edited by A Stone; submitted 21.Jul.2025; peer-reviewed by HC Chiu, MDG Pimentel; comments to author 24.Oct.2025; revised version received 24.Dec.2025; accepted 29.Dec.2025; published 28.Jan.2026.

Please cite as:
Zhou Y, Rees J, Matcham F, Patel A, Antonelli M, Tinker A, Ourselin S, Liu W
Development and User-Centered Evaluation of Smart Systems for Loneliness Monitoring in Older Adults: Mixed Methods Study
J Med Internet Res 2026;28:e81027
URL: https://www.jmir.org/2026/1/e81027
doi:10.2196/81027
PMID:

©Yi Zhou, Jessica Rees, Faith Matcham, Ashay Patel, Michela Antonelli, Anthea Tinker, Sebastien Ourselin, Wei Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Attitudes and Willingness to Participate in Drug Clinical Trials Among Patients With Cancer: Multistage Qualitative Study

Sijia Yu[1,2], BSc; Mengmeng Lyu[3], PhD; Haiping Ma[1], BSc; Keying Xu[1], BSc; Haiyan Li[1], BSc; Lili Ma[4], MSc; Mengting Ji[2], PhD

[1]School of Medicine, Tongji University, Shanghai, China

[2]Clinical Research Unit, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, No.160 Pujian Road, Pudong New Area, Shanghai, China

[3]School of Nursing, Shanghai Jiao Tong University, Shanghai, China

[4]Department of Nursing, Shanghai East Hospital, School of Medicine, Tongji University, Shanghai, China

Corresponding Author:
Mengting Ji, PhD
Clinical Research Unit, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, No.160 Pujian Road, Pudong New Area, Shanghai, China

Abstract

Background: Cancer clinical trials are essential for advancing therapeutic innovations; however, patient enrollment remains a persistent challenge globally. Understanding the attitudes and willingness of patients with cancer to participate in clinical trials is critical for improving recruitment strategies. While previous studies have explored barriers and facilitators, few have integrated multiple data sources or used emerging analytical approaches, such as large language models (LLMs), to capture the multidimensional nature of patient decision-making. Furthermore, limited research has examined these perspectives within the Chinese health care context, where cultural, economic, and systemic factors may uniquely influence participation decisions.

Objective: This study aimed to examine the attitudes and willingness of patients with cancer to participate in drug clinical trials in China by (1) identifying key themes influencing patients' decision-making processes, (2) comparing thematic findings derived from investigator-led qualitative analysis with those generated by 2 LLMs (Gemini Pro 2.5 and DeepSeek R1), and (3) evaluating the complementary value of hybrid analytical approaches in qualitative health research.

Methods: A multistage qualitative study was conducted using 2 data sources: semistructured face-to-face interviews with patients with cancer (n=11) from a tertiary hospital in Shanghai and publicly available comments from 2 Chinese online health communities (Zhihu and Yuaigongwu). Of the 3148 initial comments, 219 met the inclusion criteria after systematic screening. Three parallel analytical approaches were used: investigator-led thematic analysis, Gemini Pro 2.5–assisted analysis, and DeepSeek R1–assisted analysis. Both LLMs received identical, structured prompts. Thematic outputs were systematically compared to identify convergent and divergent findings.

Results: The 3 analytical methods jointly identified 7 core themes: treatment selection, financial burden relief, uncertain therapeutic efficacy, uncertainty regarding control groups, lack of cognition, misconceptions, and physician trust. Substantial thematic overlap was observed between investigator-led and DeepSeek R1 analyses (8 shared themes, including family-involved decisions and service-related factors) and between investigator-led and Gemini Pro 2.5 analyses (3 shared themes, including regional disparities and autonomous decision-making). Method-specific themes included recognition of medical value (investigator only), insufficient clinical data (DeepSeek R1 only), and lack of information resource (Gemini Pro 2.5 only). These findings highlight the multidimensional nature of trial participation decisions, encompassing treatment expectations, economic considerations, risk perceptions, cognitive factors, trust relationships, and structural barriers to accessibility.

Conclusions: The willingness of patients with cancer to participate in clinical trials is shaped by a complex interplay of treatment expectations, economic considerations, risk perceptions, cognitive factors, and relational dynamics. The hybrid analytical framework demonstrated complementary strengths: human analysis provided contextual depth and cultural sensitivity, while LLMs offered efficiency and identified additional thematic dimensions. These findings underscore the need for patient-centered communication strategies, transparent trial information, and culturally tailored recruitment approaches. Future research should expand sample diversity and further validate the use of LLMs in qualitative health research.

(J Med Internet Res 2026;28:e88758) doi:10.2196/88758

KEYWORDS

clinical trials; online health community; participation; attitude; willingness; decision-making; large language models

XSL•FO
RenderX

Introduction

Cancer is a major cause of death in the developed world. According to GLOBOCAN 2022 data, China recorded 4.825 million new cancer cases and 2.574 million cancer-related deaths annually, accounting for 24% and 26% of the global totals, respectively [1]. This not only imposes a heavy disease burden but also incurs huge socioeconomic costs. Clinical trials are a key means to advance cancer treatment, improve survival rates, and introduce innovative therapies [2]. However, implementing clinical trials faces systemic barriers. More than 80% of trials fail to complete the required sample enrollment within the target timeframe [3,4]. As many as 19% of clinical trials are terminated due to insufficient enrollment [5]. The participation rate among adult patients is extremely low, ranging from 2% to 8% over the long term [6-8]. These deficiencies often require adjustments, such as extending the time frame and increasing the number of test sites, to scale-up participant recruitment. However, these measures have significant financial implications [9].

Patients' decisions regarding trial participation may be influenced by multidimensional factors, including limited knowledge, concerns about the informed consent process, and low motivation [10-12]. Previous research has extensively explored the factors influencing the decisions of patients with cancer to participate in drug clinical trials. A notable systematic review in 2015, which encompassed 36 quantitative and 15 qualitative reports, identified a broad spectrum of personal, structural, and social factors affecting participation in radiotherapy, surgical, and drug trials [13]. More recently, research has focused specifically on the nuanced context of drug trials. For instance, a 2019 qualitative synthesis of 9 studies found that patient decisions are critically shaped by interpersonal factors, such as trust in physicians, relatives' attitudes, and perceived impact on their family [14]. However, a key limitation of these studies is their reliance on a single methodological approach, either exclusively qualitative or exclusively quantitative. Previous quantitative studies have examined associations among factors but often fail to capture patients' subjective experiences. In contrast, qualitative research provides valuable insights into emotional and experiential dimensions, yet small and unrepresentative samples constrain its findings. To overcome these limitations, a mixed methods approach that integrates online discourse analysis with in-depth interviews is warranted, thereby enhancing representativeness and depth of understanding of patient perspectives.

A significant challenge in social and health research is sourcing data that is both authentic in depth and representative in scope. Although qualitative research does not aim for statistical representativeness, the sample's diversity directly influences theoretical saturation and conceptual coverage. Online data collection enables the rapid construction of a diverse sample pool by allowing algorithmic screening of user content across various geographic, age, and professional backgrounds. Prior studies have modeled patients' social support needs using data from online health communities (OHCs) to better address their needs [15]. A meta-synthesis of 28 qualitative studies further delineated the opportunities and challenges associated with the use of online social platforms for individuals navigating pregnancy loss [16]. Therefore, this study harnesses the broad scope and authentic depth of online data to investigate the attitudes and willingness of patients with cancer to participate in drug clinical trials.

Large language models (LLMs) have emerged as powerful tools for processing and analyzing textual data. These artificial intelligence (AI) systems, such as GPT, demonstrate capabilities in pattern recognition and theme extraction that can complement traditional qualitative methods [17]. However, LLMs are not entirely neutral analytical tools; they carry biases originating from their training data and algorithmic design. Moreover, LLMs struggle to discern emotional nuances and contextual depth [18]. Human researchers possess a more holistic orientation and deeper understanding, as emphasized by humanism, in complex causal reasoning, theoretical construction, and knowledge production. Thus, this study advocates leveraging LLMs' current qualitative text-analysis capabilities to assist researchers, thereby enhancing the reliability and reproducibility of research.

Two data collection methods were employed in this study: conducting one-on-one, semistructured interviews and collecting online discourse using Python (Python Software Foundation). In this study's analytical approach, traditional qualitative analysis is integrated with LLM capabilities. By analyzing patient narratives from these 2 methodological perspectives, this study aims to comprehensively evaluate the attitudes and willingness of patients with cancer to participate in trials, ultimately informing evidence-based strategies to optimize recruitment. The goal is to propose targeted intervention strategies to improve the clinical trial experience for patients with cancer and to provide a scientific basis for facilitating the smooth progression of participant recruitment.

Methods

Overview

This study used a multistage qualitative design that uses 2 distinct data sources examined through 3 analytical lenses. The research process was structured into 3 stages to ensure a comprehensive evaluation. Stage 1 involved preparatory work, specifically the development of the interview guide based on a literature review. Stage 2 comprised dual-channel data collection, obtaining detailed narratives from one-on-one patient interviews and scraping online discourse using Python. Stage 3 focused on analysis and comparative synthesis, where traditional human qualitative coding was triangulated with LLM-driven thematic analysis (Figure 1).

XSL•FO
RenderX

Figure 1. Multistage study design flowchart illustrating the 3 stages of data collection and analysis for exploring the attitudes of patients with cancer toward drug clinical trial participation in China. GenAI: generative artificial intelligence.

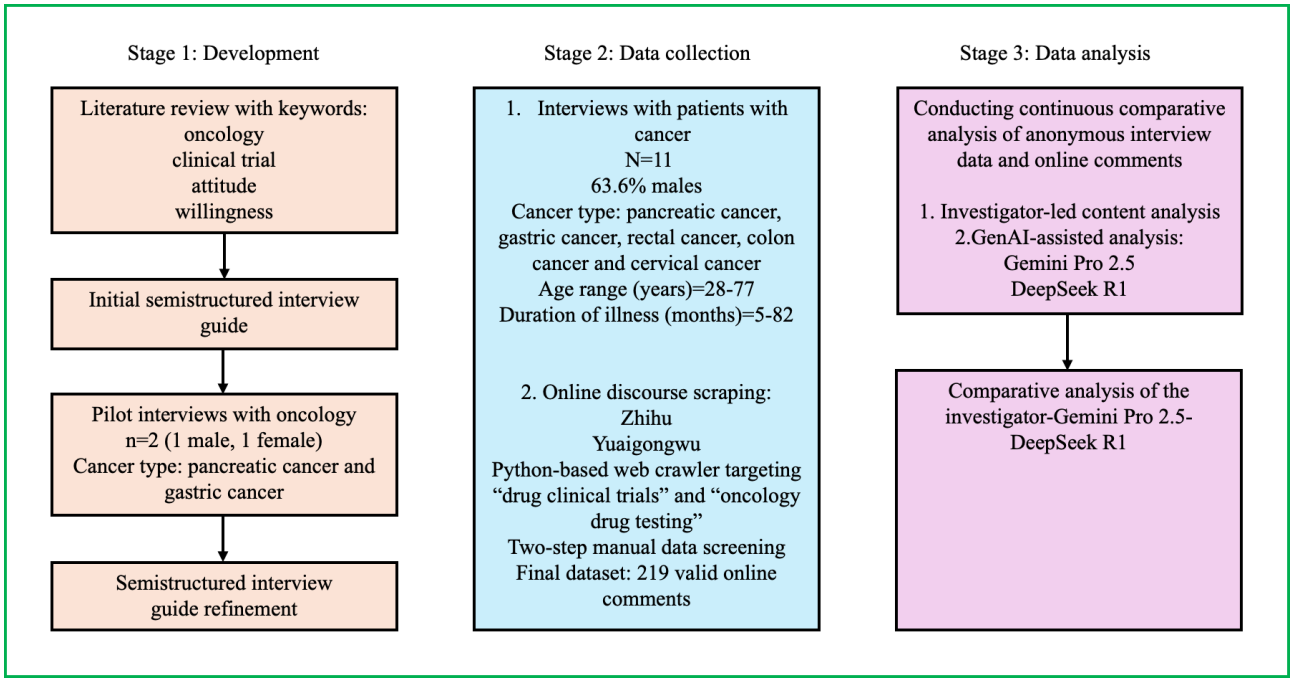

Interview Guide Development

The initial interview guide was developed from a literature review and refined after pilot interviews, yielding a final version comprising 4 dimensions: personal experiences with clinical trials, decision-making factors, social influences, and willingness to participate in the future (Textbox 1).

Textbox 1. The final semistructured interview guide.

Could you please elaborate on your understanding of clinical trials?

- Have you ever participated in a clinical trial?
  - If you have participated in a clinical trial:
    - What type of clinical trial did you participate in?
    - Please describe the process of your participation in the clinical trial and provide an evaluation of it.
    - What were the attitudes of those around you toward your participation?
    - Would you be willing to participate in a new clinical trial? Why or why not?
  - If you have not participated in a clinical trial.
    - What were the reasons for your nonparticipation or refusal to participate in a clinical trial?
    - Would you be willing to participate in a clinical trial? Why or why not?
- Have individuals in your social circle participated in clinical trials? What do you believe were the reasons for their participation or refusal?

Inclusion and Exclusion Criteria

Interview Participants

Purposive sampling was used to recruit patients meeting the inclusion criteria: (1) adults with a histologically or cytologically confirmed malignant solid tumor, (2) an education level of at least junior high school, (3) Eastern Cooperative Oncology Group performance status of 0‑1, and (4) estimated life expectancy of 12 weeks or more. These criteria were established to ensure that participants had sufficient physical stamina to complete 20-40–minute interviews and the cognitive capacity to articulate complex factors in clinical trial decision-making. It is acknowledged that these criteria inherently exclude the most frail patients and those with limited educational backgrounds, and that their perspectives may therefore be underrepresented in this study. All participants were required to participate in the study voluntarily, demonstrate adequate comprehension of the study procedures, and provide written informed consent. Patients were excluded if they had a history of psychiatric disorders, cognitive impairment, communication barriers, or active central nervous system metastases (eg, cerebral edema and progressive lesions). After fully explaining the study's purpose and procedures to all participants, informed

XSL•FO
RenderX

consent was obtained. A total of 11 participants were recruited for the face-to-face interviews.

Online Data

Inclusion criteria are (1) comments explicitly discussing cancer drug clinical trials, (2) comments from patients with cancer or their caregivers, and (3) sufficient content length (≥10 Chinese characters) for meaningful analysis.

Exclusion criteria are (1) duplicate comments, (2) incomplete or incomprehensible content, and (3) advertisements or promotional information.

Data Collection

Data collection integrates individual perspectives and population-level insights. This dual-pronged approach was actualized through in-depth, one-on-one interviews, which provided nuanced personal narratives, and the systematic collection of public comments from online platforms, which offered insights into collective sentiments and spontaneous reactions.

Face-to-Face Interviews

Face-to-face interviews were conducted in January 2025. All interviews were conducted by a trained qualitative researcher with experience in clinical trials and oncology care. Each patient was interviewed once; no repeat interviews were conducted. Interviews were scheduled at times convenient for the interviewees and conducted one-on-one sessions in quiet conference rooms, lasting 20 to 40 minutes, to allow participants to independently determine the extent of their disclosures. All interviews were audiorecorded with participants' informed consent.

Online Data Scraping and Screening

For this study, online data were collected via web crawlers, and comments published between September 2010 and May 2025 were retrieved. The textual data originated from 2 distinct Chinese online platforms: Zhihu and Yu-Ai-Gong-Wu (Dancing with Cancer). The selection of these 2 platforms was deliberate, aiming to encompass a wide range of viewpoints by combining extensive public discussions with personal accounts that closely address patients' needs.

Zhihu is a prominent question-and-answer–based social platform in China, known for its high-quality, in-depth content from a diverse user base of professionals, experts, and engaged laypersons. The total number of views on cancer-related topics on Zhihu has exceeded 3.81 billion. Yuaigongwu, which translates to "Dancing with Cancer," is a dedicated online support community for patients with cancer and their families, established in 2010. The platform facilitates peer-to-peer support and the sharing of firsthand treatment experiences and emotional journeys. The forum has more than 1 million posts and a membership exceeding 180,000.

Data collection was performed using a custom Python web crawler. The crawler was programmed to scrape publicly accessible text data (posts and comments with timestamps) from the selected platforms by targeting specific keywords, including "drug clinical trials" and "oncology drug testing." All data were anonymized to protect user privacy, and the collection process complied with relevant laws and the platforms' terms of service. A rigorous multistage data preprocessing protocol was subsequently implemented to ensure the dataset's quality and relevance.

A total of 3148 comments related to cancer clinical trials were identified through web crawling. After removing 39 duplicate comments, 3109 unique comments were retained for screening. In the initial screening stage, comments were excluded if they (1) contained advertisements or promotional information (n=75), (2) were unrelated to cancer clinical trials (n=2475), or (3) contained fewer than 10 meaningful characters after removing emojis and hyperlinks (n=208). A total of 2758 comments were excluded, and 351 comments were retained for further eligibility assessment.

In the second-stage eligibility assessment, 2 researchers independently reviewed the remaining 351 comments. Comments were excluded if they were unrelated to the subject of this study (n=118) or were incomplete or unintelligible (n=14). Discrepancies between the 2 reviewers were resolved through discussion and, when necessary, adjudicated by a third researcher. After excluding 132 comments, a final dataset of 219 valid comments was retained for analysis.

Text data were normalized by removing nonessential characters, standardizing punctuation, and tokenizing emoticons that conveyed emotion. All inclusion and exclusion decisions were recorded in a structured dataset that included the post ID, screening results, reasons for exclusion, and reviewer notes. A random 5% of the data was rechecked upon completion to ensure accuracy. This 2-step, transparent screening process ensured the reliability, validity, and reproducibility of the dataset used for qualitative analysis.

Data Analysis

This study adopts a parallel analytical approach, integrating a researcher-led analysis with an analysis conducted by the LLM. The interview recordings were transcribed using Feishu software (Beijing Feishu Technology Co., Ltd), and 2 researchers independently verified the transcripts for accuracy.

Investigator-Led Content Analysis

Content analysis was used to analyze the interview transcripts and online comments separately, thereby distilling key themes. Two researchers independently analyzed the materials by repeatedly reading the textual data, identifying meaningful statements, and assigning codes. Following the summarization, preliminary themes were formulated. In instances of discrepancies during the analysis process, these were resolved through research group discussions until consensus was reached, ultimately determining the final research results.

All interview records were anonymized before analysis to protect confidentiality, especially during the LLM-assisted analysis stage, where special attention was paid to data privacy.

XSL•FO
RenderX

Generative AI-Assisted Analysis: Gemini Pro 2.5 and DeepSeek R1

The study required anonymized transcripts to be prepared in plain text format for processing by both Gemini Pro 2.5 (version released by Google AI Studio in March 2025) and DeepSeek R1 (version released in January 2025, accessed via [19]). An initial prompt was developed based on the research objectives and tested on a calibration subset of 20 online comments. The initial prompt used for both LLMs was:

> *Please analyze the interview records of cancer patients / the online platform comments from cancer patients in the document. Your task is: (1) Identify the main themes and sub-themes related to the patients' willingness and attitude towards participating in drug clinical trials. (2) For each theme, extract 2 key quotes or phrases that can reflect the theme insights, and explain their significance for understanding the patients' attitudes.*

This calibration phase identified two issues requiring refinement: (1) the initial prompt allowed paraphrased summaries, which hindered quote verification, and (2) output formatting was inconsistent across responses. The prompt was refined to require verbatim quotes and to specify a structured output format. Once finalized, the prompt was fixed and applied consistently to all subsequent analyses without further modification.

Unlike human coders, LLMs do not apply predefined codebooks line by line or engage in iterative coding. Consequently, traditional inter-rater reliability metrics, such as Cohen κ, do not apply to their outputs. Instead, transparency in prompt construction and in human review of AI-generated themes was used to ensure analytic consistency and validity [18].

The initial prompt was refined as follows:

> *Please analyze the following anonymized [interview transcripts / online comments] from cancer patients regarding their attitudes toward participating in drug clinical trials. Your tasks are as follows: (1) Identify the main themes and subthemes related to patients' willingness and attitudes toward clinical trial participation. (2) For each theme identified, extract exactly 2 verbatim quotes from the source text that best illustrate the theme, and provide a brief explanation of each quote's significance. (3) Present your findings in a structured format with theme names, subthemes, supporting quotes, and interpretive explanations. Ensure all quotes are copied exactly as they appear in the source text.*

These refinements improved the consistency and verifiability of LLM outputs.

Interview transcripts (N=11) were compiled into a single anonymized document with participant identifiers (P01-P11). Online comments (n=219) were organized into a separate document with sequential identifiers (ZH01-ZH53 and YAGW001-YAGW166), with each comment clearly delimited. Documents were uploaded to each LLM platform separately as contextual input, and the models were prompted with structured instructions to identify overarching themes through in-context learning.

LLM outputs were independently reviewed by two researchers to verify (1) accuracy of quote attribution (ie, that quoted text actually appeared in the source data), (2) coherence and relevance of thematic categorizations, and (3) absence of fabricated or hallucinated content. Discrepancies were resolved through discussion with a third researcher. Quotes that could not be verified in the source data were excluded from the final analysis.

Comparative Analysis

To ensure the cross-method reliability of the findings, a systematic comparative analysis was conducted. Themes derived from all analytical methods were evaluated through parallel examination, addressing both terminological variations and interpretive differences. A thematic map was created to link synonymous or related labels, producing a single, integrated thematic framework (Figure 2).

XSL•FO
RenderX

Figure 2. A diagram depicting thematic convergence and divergence across 3 analytical methods (investigator-led, Deepseek R1, and Gemini Pro 2.5) in the analysis of the attitudes of patients with cancer toward clinical trial participation.

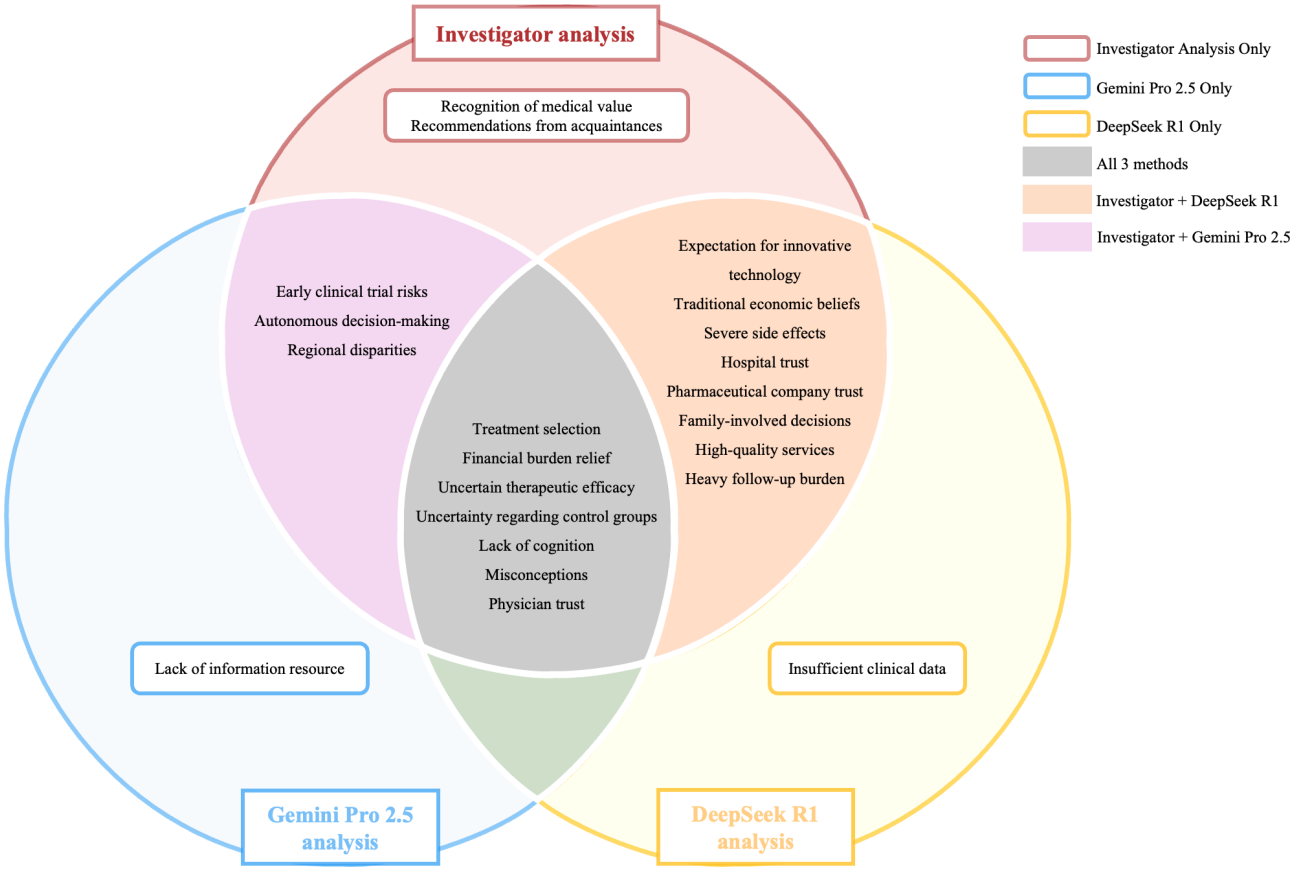

Conceptual alignment was observed when distinct terminologies referenced identical constructs. For instance, the investigator-led subtheme "concerns regarding efficacy," Gemini Pro 2.5's "fear stemming from uncertainty," and DeepSeek R1's "uncertainty of curative effect" were found to be semantically equivalent and were harmonized into the unified theme of "uncertain therapeutic efficacy." Conversely, interpretive divergences requiring adjudication occurred when methods captured different intensities or facets of a topic. Rather than treating these as discrepancies, they were leveraged as sources of deeper interpretive insight. A notable discrepancy arose regarding economic motivations: while human analysis identified general "alleviation of economic pressure," DeepSeek R1 focused on procedural "cost waivers," and Gemini Pro 2.5 highlighted a more desperate sentiment that "free treatment is the only way out." The research team re-evaluated the source data and determined that these were distinct but complementary dimensions. Consequently, these were synthesized into the comprehensive theme of "financial burden relief."

Themes were ultimately classified as convergent if they embodied substantively equivalent ideas across methods. An independent reviewer conducted the classification, a second investigator validated all classifications, and discrepancies were resolved through consensus within the research team.

Ethical Considerations

This study was approved by the Ethics Committee of Renji Hospital, affiliated with Shanghai Jiao Tong University School of Medicine (LY2025-140-B), and adhered to relevant ethical guidelines, including the Declaration of Helsinki and those of the Association of Internet Researchers [20]. For the face-to-face interviews, written informed consent was obtained from all participants. For the online data component, the ethics committee waived the requirement for individual informed consent, as the research involved a noninterventional analysis of publicly available social media data. We ensured that data collection complied with the platforms' terms of service regarding publicly accessible content. To strictly protect user privacy, all online data were fully anonymized through a multistep process. First, structured identifiers, such as usernames, were automatically removed during the data scraping phase. Second, 2 researchers manually reviewed the text content to detect and redact any remaining personally identifiable information, such as real names, specific hospital names, and unique location information. This ensured that no quotes presented in this manuscript could be traced back to specific individuals. Each participant received approximately ¥30 RMB (approximately US $4) as transportation reimbursement for their participation in the interviews.

Results

Participants of Interviews

The interview portion of this study included 11 patients with cancer (n=7 males and n=4 females) with a mean age of 57.73 (SD 13.4; range 28 - 77) years. The participants represented a variety of cancer diagnoses, with pancreatic cancer being the most common (n=6, 54.5%), followed by gastric cancer (n=2,

XSL•FO
RenderX

18.2%), and other cancer types (Table 1). The majority of participants were married (n=9, 81.8%), with the remainder reporting other marital statuses. The highest educational attainment among participants was a bachelor's degree (n=4, 36.4%), while others had completed either high school (n=4, 36.4%) or junior high school education (n=3, 27.3%). All age-related content in this study was entirely derived from the analysis of the interview data (Figure 3).

Table . Demographic and clinical characteristics of interview participants: patients with cancer recruited from a tertiary hospital in Shanghai, China (N=11). Data collected in January 2025.

Variables	Values
Sex, n (%)	
Male	7 (63.6)
Female	4 (36.4)
Age range (y), mean (SD; range)	57.73 (13.4; 28-77)
Family status, n (%)	
Married	9 (81.8)
Unmarried	1 (9.1)
Widowed	1 (9.1)
Duration of illness (mo), mean (SD; range)	21.82 (19.90; 5-82)
Educational qualifications, n (%)	
Junior high school education	3 (27.3)
High school education	4 (36.4)
Bachelor's degree	4 (36.4)
Cancer types, n (%)	
Pancreatic cancer	6 (54.5)
Gastric cancer	2 (18.2)
Rectal cancer	1 (9.1)
Colon cancer	1 (9.1)
Cervical cancer	1 (9.1)

XSL•FO
RenderX

Figure 3. Flowchart of the data collection and screening process related to cancer clinical trials. A total of 3148 comments were retrieved from the online platform and subjected to a 2-stage screening procedure. The final dataset comprised 219 valid comments for analysis.

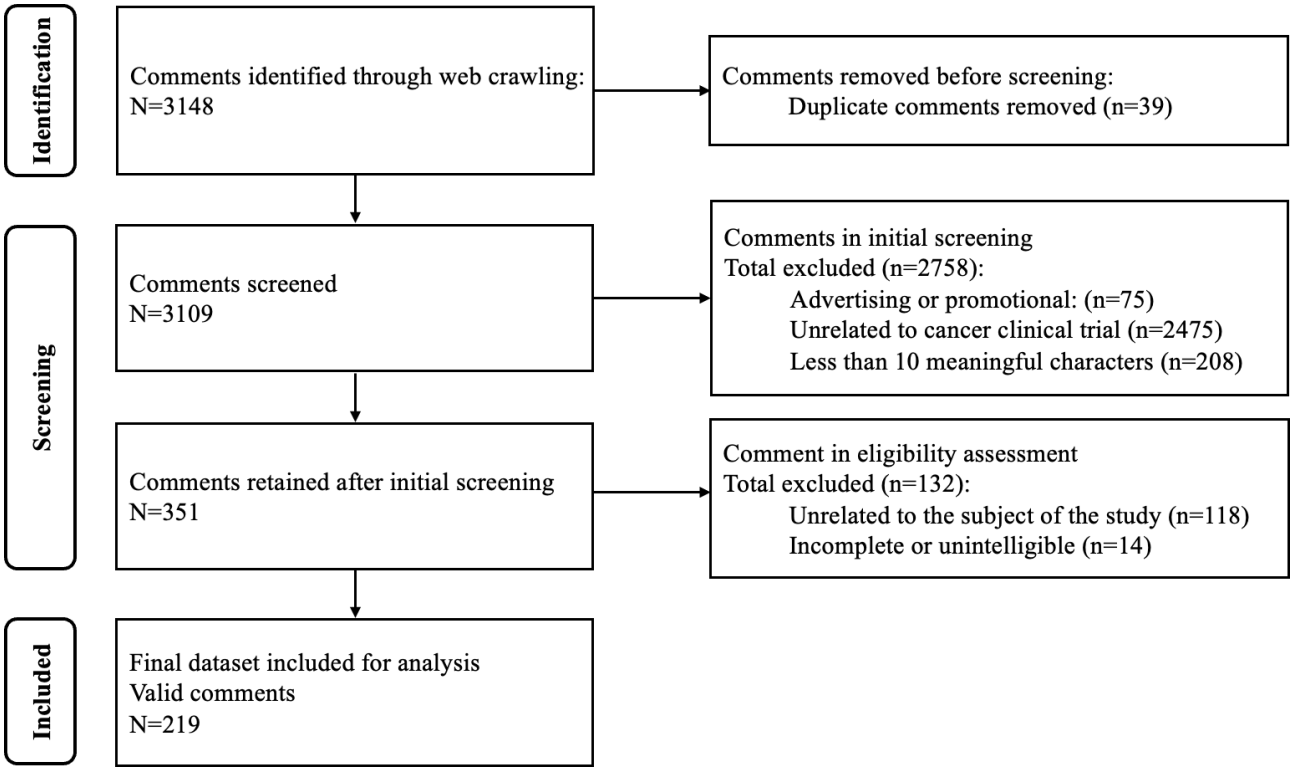

Comments From OHCs

The initial data collection yielded 3148 comments across the 2 platforms. Specifically, 2960 comments were retrieved from Zhihu and 188 from Yuaigongwu. Following the data-cleaning and screening protocol, a final dataset of 219 relevant comments was established, comprising 53 from Zhihu and 166 from Yuaigongwu.

Themes of Parallel Analysis

The analyses, led by researchers and driven by LLMs, jointly yielded 8 overarching themes (Table 2).

XSL•FO
RenderX

Table . Thematic coverage across investigator-led and large language model (LLM)-assisted qualitative analysis.

Themes and subthemes	Data source	Analytical method		
		Investigator-led	DeepSeek R1	Gemini Pro 2.5
Treatment motivation				
Treatment selection	Interview and OHC[a] data	✓	✓	✓
Expectation for innovative technology	Interview and OHC data	✓	✓	
Cost considerations				
Financial burden belief	Interview and OHC data	✓	✓	✓
Traditional economic beliefs	Interview data	✓	✓	
Risk perception				
Severe side effects	Interview and OHC data	✓	✓	
Uncertain therapeutic efficacy	Interview and OHC data	✓	✓	✓
Insufficient clinical data	Interview data		✓	
Early clinical trial risks	OHC data	✓		✓
Uncertainty regarding control groups	Interview and OHC data	✓	✓	✓
Cognitive variation				
Lack of cognition	Interview and OHC data	✓	✓	✓
Misconceptions	OHC data	✓	✓	✓
Recognition of medical value	Interview data	✓		
Brand trust effect				
Physician trust	Interview and OHC data	✓	✓	✓
Hospital trust	Interview data	✓	✓	
Pharmaceutical company trust	OHC data	✓	✓	
Decision maker				
Autonomous decision-making	Interview data	✓		✓
Family-involved decisions	Interview data	✓	✓	
Recommendations from acquaintances	Interview data	✓		
Service processes				
High-quality services	Interview and OHC data	✓	✓	
Heavy follow-up burden	Interview and OHC data	✓	✓	
Resource accessibility				
Lack of information resource	OHC data			✓
Regional disparities	Interview and OHC data	✓		

[a]OHC: online health community.

XSL•FO
RenderX

Theme 1: Treatment Motivation

Treatment Selection

For some, participation in clinical trials was driven by treatment failure at later stages, motivated by the pursuit of any remaining survival opportunities:

> *The previous trial was stopped because my liver lesions had recurred. The doctor said the drug wasn't working well and asked me to stop and rescreen. If I cannot join this program, I'll still need to switch to another treatment.* [P10]

> *I hope to be assigned to the treatment group, because other drugs are no longer effective—if they still worked, who would want to be a guinea pig?* [ZH07]

Expectation for Innovative Technology

Some participants perceived clinical trials as opportunities to access better treatments and increase their chances of improvement:

> *This drug has already gone through an earlier phase of testing, and the feedback was positive. Even without this clinical trial, I would still need regular chemotherapy, but participating might offer me more opportunities.* [P01]

> *For some diseases, participating in new drug trials is the only gamble patients can take.* [ZH52]

Theme 2: Cost Considerations

Financial Burden Relief

Most participants indicated that their decision to join a clinical trial was influenced by financial considerations, aiming to alleviate the ongoing economic burden of high treatment costs for their families:

> *If we had to buy this drug as regular consumers, we couldn't afford it. But by joining the clinical trial, it's like you only have to pay a one-time fee, and then you receive it for free.* [P08]

> *Without the clinical trial, one course of ADC [Antibody-Drug Conjugate] starts at 250,000. There's no way I could ever afford that myself.* [YAGW089]

Traditional Economic Beliefs

Some interviewees, despite having sufficient family resources, expressed a desire to preserve them for the next generation, influenced by traditional family ethics emphasizing "intergenerational sacrifice" and "intergenerational transfer of resources."

> *If I don't participate, the medicine I take now costs 200,000 a year. My son would certainly agree to it. But I still want to save a little; I want to leave it for my children.* [P06]

Theme 3: Risk Perception

Severe Side Effects

Concerns regarding potential adverse drug reactions were noted, with some individuals expressing fears about their ability to tolerate such effects:

> *I understand that participating in such trials means the drug will have side effects, and I know I won't be able to endure them. At my age, I simply can't handle it anymore. The medicines used in these clinical trials are domestically developed investigational drugs, and the side effects are extremely severe.* [P05]

> *Some of our fellow patients who participated in the olaparib trial said the side effects were very severe—teeth falling out in chunks, even worse than chemotherapy reactions. Many people withdrew from the clinical trial midway...If it is this painful, is it really necessary to join a clinical trial?* [YAGW110]

Uncertain Therapeutic Efficacy

Findings revealed that doubts about therapeutic efficacy and safety risks contributed to participants' hesitancy toward clinical trials. Refusal to take unapproved research drugs was often attributed to perceived safety issues and limited technical maturity:

> *What worries me most is that the clinical trial may not go well, and I'm not confident about its success. It's not only about whether the drug works, but whether it will truly benefit my condition.* [P02]

> *The doctor recommended that we participate in the X-396 trial, but we are still hesitant. We really don't know whether it will be effective or not.* [YAGW117]

Insufficient Clinical Data

Some participants perceived that the clinical data provided to them prior to enrollment were insufficient for informed decision-making, particularly regarding key outcome probabilities such as survival, recovery, and mortality rates:

> *Consider the balance of risks and gains: during a clinical trial, what are the chances of survival? The chances of death? The chances of recovery?...such probabilities must be made transparent and thoroughly analyzed.* [P02]

Early Clinical Trial Risks

Early clinical trial risks emerged as a notable concern. Participants expressed greater apprehension toward phase I trials due to unclear side effects, whereas phase III trials with more comprehensive safety and efficacy data were generally viewed as more acceptable:

> *In phase I, many side effects remain unclear.* [YAGW089]

> *Usually, I check which phase the trial is in. Phase III is the best.* [YAGW090]

> *Phase III clinical trials are worth considering.* [YAGW091]

XSL•FO
RenderX

Uncertainty Regarding Control Groups

Uncertainty about the randomization process was identified as a concern during the trial:

> *I'm afraid of being assigned to the control group, because there is a treatment group and a control group.* [P09]
>
> *If I join, there's only a fifty–fifty chance of receiving the actual treatment. The other half would be assigned to the control group.* [YAGW138]
>
> *In a new drug clinical trial, isn't it divided into treatment and control groups? If I end up in the placebo group, it might actually delay my condition.* [YAGW160]

Theme 4: Cognitive Variation

Lack of Cognition

Several participants expressed that their limited understanding of clinical trials directly contributed to feelings of uncertainty and apprehension regarding participation:

> *I really don't know what the outcome of the trial will be—because, after all, I don't understand these things.* [P09]
>
> *Because I don't understand anything right now, there is an instinctive sense of worry about free trials.* [YAGW025]

Misconceptions

Some patients develop misconceptions due to a lack of understanding of clinical trials:

> *With investigational drugs, the patient bears the risk. If it succeeds, the credit goes to the inventors, and they get to make big money! Is that fair?* [ZH48]
>
> *Clinical trials are just using ordinary people's health for experiments. It's a way to profit from harming people, and it's legal, with no accountability.* [ZH32]

Recognition of Medical Value

Several participants recognized that clinical trials provide not only individual treatment opportunities but also hold forward-looking value for the medical field. They emphasized that the effectiveness of real-world data is critical for breakthroughs in disease treatment and for broader progress in health care:

> *It gives patients more opportunities, and for the medical field, it represents an exploratory step forward for a particular disease. With the accumulation and application of technology, such exploration might identify better drugs or therapies and eventually upgrade treatment approaches.* [P01]
>
> *Clinical trials are definitely for patients, to benefit them. They need to be conducted in hospitals or research centers so that the drug can be promoted, allowing patients to access it earlier in the market.* [P03]

Theme 5: Brand Trust Effect

For some individuals, the decision to enroll was heavily reliant on trust conveyed through authoritative channels, including institutional credibility and the doctor-patient relationship.

Physician Trust

Participants frequently cited trust in their physicians as a key factor in their willingness to participate in clinical trials. The reputation and recommendation of the treating physician provided a sense of reassurance and credibility.

> *Just for the name of Director Wu alone—it's enough to go ahead without hesitation!* [YAGW014]
>
> *My attending doctor told me about it, and we are willing to listen to the doctor.* [P08]

Hospital Trust

Trust in the hospital conducting the trial also played an important role. Participants believed that reputable hospitals would only be involved in rigorous and legitimate research, which increased their confidence in the trial.

> *I know that for a drug, especially a cancer drug, to be brought to market and to be selected for clinical research at a hospital like Renji, it must not be just any ordinary drug.* [P03]

Pharmaceutical Company Trust

Participants expressed greater willingness to enroll in trials sponsored by well-known pharmaceutical companies. The brand reputation and international recognition of such companies were perceived as indicators of trial legitimacy and safety.

> *Merck is a major, internationally renowned company. Its reputation and brand recognition are top-tier; they wouldn't conduct an illegitimate trial.* [YAGW034]

Theme 6: Decision-Maker

Autonomous Decision-Making

Certain participants asserted that the decision to participate in a clinical trial was ultimately their own, emphasizing that family members could not make this choice for them:

> *Ultimately, the deciding factor in this matter is me.* [P01]
>
> *Because with something like this, every decision made is one I must take responsibility for myself.* [P04]

Family-Involved Decisions

Some interviewees viewed family consultation as an essential part of the medical decision-making process:

> *I would definitely need to discuss it with my wife and child.* [P02]
>
> *Of course, that was a decision made after discussing it with my family.* [P08]

XSL•FO
RenderX

Recommendations From Acquaintances

Some respondents stated that their participation in the clinical trial was largely motivated by strong encouragement from people they knew personally:

> *My cousin is a doctor at the hospital; she was the one who told me there was an available spot.* [P10]

Theme 7: Service Processes

High-Quality Services

A subset of participants reported positive experiences with the medical team during the clinical trial, noting that the services provided were excellent and that procedures were flexibly adapted to meet their needs:

> *The director made arrangements for me, contacting Renji Hospital in Puxi for my examination, and I was admitted directly to the hospital in Pudong the very next day.* [P06]

> *The hospital here sends a text message first, then follows up with a phone call, and also notifies me via WeChat. I don't have to worry about keeping track of it myself; they automatically call me when it's time.* [P08]

> *Joining this trial is actually quite good, not to mention it saves a considerable amount of money. [I] am basically a VIP in the hospital, with priority access to all kinds of services, and spared from all the cumbersome procedures like queuing, registration, booking appointments, and waiting for beds.* [YAGW065]

Heavy Follow-Up Burden

Participants reported that the frequency of follow-up visits created a significant time burden and physical fatigue, causing practical disruptions to their daily lives:

> *It's just that the blood draws are too frequent. They require me to come back the very next day for several days in a row. It's quite inconvenient, and I feel rather fatigued during that time.* [P04]

> *The routine monitoring is especially frequent. For the first few months, there were check-ups every single week—all sorts of CT scans. It takes a significant toll.* [YAGW144]

Theme 8: Resource Accessibility

Lack of Information Resource

Most patients reported no prior exposure to clinical trials, leading to fear rooted in unfamiliarity. Once diagnosed, they faced a lack of formal channels for acquiring reliable information:

> *In smaller towns, it's more a matter of not knowing. The moment they hear it's an "experiment," people are scared off immediately.* [ZH21]

> *It's an information gap. Fear comes from a lack of understanding, and ultimately, it causes you to miss out on many opportunities.* [ZH27]

Regional Disparities

Due to the specialized nature of clinical trials, these medical resources may be inaccessible in some remote regions. As a result, professional knowledge regarding clinical trials is largely concentrated in major cities, leading patients in outlying areas to be wary of and avoid them:

> *At that time, in Urumqi, no such program was available.* [P01]

> *In top-tier metropolitan hospitals, like those in Shanghai, patients have greater awareness and a deeper understanding. Once they understand, they can genuinely grasp the potential benefits of a trial, which makes them more willing to participate.* [ZH21.]

Themes of Comparative Analysis

Figure 2 presents the results of a comparative thematic analysis, illustrating both the overlapping and unique themes identified through 3 distinct analytical approaches: investigator analysis, Gemini Pro 2.5 analysis, and DeepSeek R1 analysis.

A core set of 7 themes emerged as a consensus across all 3 methods. These common themes, shown in the central box, include (1) treatment selection, (2) financial burden relief, (3) uncertain therapeutic efficacy, (4) uncertainty regarding control groups, (5) lack of cognition, (6) misconceptions, and (7) physician trust. The convergence of these themes across human and AI-assisted analyses strengthens confidence in their validity as central concerns for patients.

Beyond this shared thematic foundation, each analytical approach yielded unique insights. The investigator's analysis was the only method to identify the themes of "recognition of medical value" and "recommendations from acquaintances." The Gemini Pro 2.5 analysis uniquely identified "lack of information resource." Similarly, "insufficient clinical data" was a theme exclusively identified by the DeepSeek R1 analysis.

The analysis also revealed partial thematic overlaps between pairs of methods. The themes of "early clinical trial risks," "autonomous decision-making," and "regional disparities" were jointly identified by the investigator and Gemini Pro 2.5 analyses.

Eight themes were jointly identified by the investigator-led analysis and DeepSeek R1: "expectation for innovative technology," "traditional economic beliefs," "severe side effects," "hospital trust," "pharmaceutical company trust," "family-involved decisions," "high-quality services," and "heavy follow-up burden" (Figure 2). This overlap suggests that DeepSeek R1's analytical patterns aligned closely with human interpretive approaches for these dimensions.

These method-specific findings demonstrate the complementary value of the hybrid analytical approach: human analysis captured nuanced relational and value-based themes, while LLM analyses identified distinct informational and evidentiary concerns.

XSL•FO
RenderX

Discussion

Diversity of Perspectives on the Willingness of Patients with Cancer to Participate in Drug Clinical Trials

This study used a multistage qualitative design to provide comprehensive insights into the willingness and attitudes of patients with cancer toward participation in clinical trials. Patients' decisions are influenced not only by medical factors but also by personal illness trajectories, family expectations, cultural context, and information accessibility [21-23]. The findings of this study underscore the multidimensional nature of patient decision-making.

The identification of unique themes by the investigator, Gemini Pro 2.5, and DeepSeek R1 further demonstrates that no single analytical perspective can fully capture this complexity. Capturing such diversity, therefore, requires methodological pluralism and interpretive integration. While previous studies have explored patients' motivations and concerns mainly through interviews or surveys [24,25], our findings deepen this understanding by showing how patients continuously negotiate autonomy, trust, and perceived benefits throughout their engagement in clinical research.

In line with emerging evidence from analyses of online health communities [26,27], our results also reveal how informal, peer-based discussions reflect the emotional, social, and informational dimensions that shape decision-making. Together, these insights highlight the need for communication strategies and trial designs that respect relational autonomy, address patients' specific concerns, and foster more informed and value-aligned participation.

Enhancing Methodological Rigor Through Hybrid Analytical Approaches

Traditional approaches to multisource textual analysis face challenges related to subjectivity and contextual variability [28-30]. Recent methodological advancements, such as collaborative digital frameworks [31] and the use of LLMs in text interpretation [32-35], have sought to enhance analytic rigor and transparency. Nevertheless, integrating computational and human-led analyses remains essential to balancing efficiency with interpretive depth.

In this study, the combination of investigator-led thematic interpretation with LLM-based analysis facilitated a more comprehensive understanding of patients' attitudes toward clinical trials. By triangulating findings from semistructured interviews and online health community discourse, and by using investigator insights alongside Gemini Pro 2.5 and DeepSeek R1, the analysis produced credible, contextually nuanced results. These outcomes are consistent with prior evidence that methodological rigor in qualitative research stems not from a single approach but from the integration of complementary strategies that enhance validity and interpretive transparency [31,36,37].

Collectively, these findings highlight that hybrid analytic frameworks can reveal both convergent and divergent dimensions of patient perspectives, enabling a more balanced representation of lived experiences within the broader landscape of clinical trial participation.

The Key Factors Influencing Patients' Participation in Clinical Trials

Therapeutic Necessity

The majority of respondents enrolled in clinical trials only after their existing treatment failed, leaving them with no alternative options. This suggests that patients believe clinical trials are their only hope, consistent with previous research results [38,39]. Therefore, patients' enrollment decisions are strongly shaped by therapeutic necessity and the perceived opportunity to access otherwise unavailable interventions [40].

For some patients, the hope and persistent belief that they can be cured of their illness help them regain some normalcy in their lives and enable them to cope with the severe prognosis of their condition [41]. Participants frequently described trial enrollment as a last resort, underscoring the critical role of perceived therapeutic benefit in decision-making. The prevalence of therapeutic necessity necessitates ethically nuanced communication. Recruitment materials should clearly delineate research objectives from therapeutic possibilities, acknowledge the experimental nature of interventions, and avoid implicit guarantees of clinical benefits. Risk-benefit discussions should balance empathy with transparency, clearly distinguishing between research aims and potential clinical benefits. To retain these motivated yet vulnerable participants, proactive emotional support, dynamic consent processes, and ongoing individual education are recommended.

Trustworthy Physician-Patient Relationships

The physician-patient relationship is a critical determinant of patients' willingness to participate in clinical trials. Consistent with prior research, strong trust in a physician correlates with greater willingness to participate [38], whereas institutional skepticism tends to heighten patient concerns [42]. However, recent qualitative findings by Murphy et al [43] reveal that such trust can be ethically ambivalent. While it facilitates engagement, it may also compromise informed consent by leading patients to agree to participate out of a desire to please their physicians or to maintain a valued relationship. This underscores the need for a more balanced communication framework that respects emotional dynamics while maintaining ethical integrity. From the physician's standpoint, this entails moving from the role of educator to that of collaborator, who acknowledges patients' emotions, encourages open dialogue, and ensures patients feel heard and understood [44]. Such an approach not only fosters authentic trust but also strengthens shared decision-making and mitigates the risk of undue influence. At the same time, it is emphasized that communication strategies must prioritize verifying patient autonomy over merely relying on trust to increase enrollment.

The Economic Calculus of Participation

The findings affirm that economic considerations are a powerful component of the decision-making calculus for many patients. This is consistent with studies such as Sun et al [45], which rank free treatment as a significant motivator. However, this

XSL•FO
RenderX

economic driver creates a critical paradox when contrasted with research by Kim et al [46], which documents that socioeconomically disadvantaged populations—the very groups expected to be most responsive to financial incentives—are under-represented in US trials. The simplistic explanation of "cultural differences" is insufficient. Instead, this divergence powerfully illustrates that economic incentives are a necessary but not sufficient condition for participation. Our findings, situated in the Chinese context, highlight the potent "pull" of financial relief. Kim et al [46] work, in turn, reveals the formidable structural barriers, such as low health literacy, logistical hurdles, and historical mistrust, that can completely neutralize this pull for vulnerable populations. This synthesis suggests that the efficacy of an economic incentive is profoundly mediated by a patient's social and structural context. Therefore, for recruitment strategies to be truly equitable, they must move beyond merely offering financial relief and actively dismantle the systemic barriers that prevent vulnerable patients from even considering the opportunity.

Procedural Clarity and Logistical Demands

From the perspective of the service process, participants' willingness to engage in clinical research is often undermined by procedural burdens such as repeated examinations, long waiting times, and frequent sampling. These requirements cause physical fatigue and psychological disengagement, ultimately reducing retention. Prior research has sought to mitigate this issue through logistical optimization, such as streamlining online follow-ups, improving communication interfaces, and implementing virtual trial workflows, digital tools, and wearables [47,48]. These measures often represent incremental refinements to existing site-centered models rather than structural solutions to patient burden. In contrast, emerging approaches such as patient-centric sampling and decentralized clinical trials aim to address the root causes of these burdens by integrating research into daily life [49,50]. However, the key insight from the data is not simply that technology is superior, but that minimizing physical disruption to the patient's routine is essential for retention. Therefore, efforts to improve participation should prioritize reducing the protocol's physical and logistical burden, rather than merely digitizing communication within a burdensome framework.

The Emotional Calculus of Risk

Consistent with previous studies, participants' willingness to participate was influenced by perceived trial risks [23]. The concerns expressed centered on potential adverse reactions, arrangements for posttrial care, and, most critically, the trial's efficacy. This highlights the urgent need for effective communication to explain potential benefits and risks and to address patient concerns [40]. However, this assessment of risk is not a fixed calculation. Emerging evidence indicates that it is a dynamic process, powerfully shaped by the patient's evolving clinical reality rather than by the mere duration of their illness. Risk tolerance, for example, often increases with disease severity or after other treatments have failed, as the focus shifts toward potential efficacy. Conversely, a personal history of adverse effects can lead to significant risk aversion [51]. This implies that a one-time, upfront disclosure of risks is insufficient. Instead, risk communication should be regarded as a continuous, adaptive process that evolves alongside the patient's clinical condition and emotional state. Proactive monitoring of patients' perceptions and regular follow-up discussions can help identify shifting concerns, clarify misunderstandings, and reinforce trust in the research team. Integrating this dynamic communication model into participant management and recruitment frameworks not only enhances ethical compliance but also strengthens patient retention and data reliability, thereby contributing to the overall integrity and success of clinical trials.

Patient Perspectives on Protocol Design

Consistent with previous research, the randomized design of clinical trials remains a major barrier to patient participation [52]. Patients prefer that treatment plans be determined by their physicians rather than randomly assigned [24]. Murphy et al [43] noted that this resistance is often compounded by misunderstandings in which patients erroneously believe that physicians control the allocation. However, Wegge-Larsen et al [5] found that, even when patients fully understood randomization, they remained willing to participate to access novel interventions. This contrast suggests that the barrier of randomization is not absolute but context-dependent. For patients with limited therapeutic options, the opportunity to access a new treatment outweighs the uncertainty of randomization. Conversely, when established treatments are available, randomization is more likely to be perceived as a loss of care. Thus, the effective communication of randomization requires addressing valid patient concerns about care equity, rather than treating resistance solely as a lack of comprehension.

The Locus of Agency in Decision-Making

This study's identification of autonomous, family-delegated, and joint decision-making patterns reveals the complex locus of agency in clinical trial participation. These findings reflect not only individual preferences but also the deep-seated cultural and social contexts within which medical decisions are made. While autonomous decision-making reflects a growing awareness of self-determination, the prevalence of family involvement underscores the influence of values that prioritize collective judgment. This observation aligns with broader findings in oncology care. Systematic reviews by Dijkman et al [21] and Tielemans et al [53] confirm that family involvement in treatment decision-making is a widespread phenomenon, observed across North American, European, and Asian contexts, challenging the notion that it is an exception to the rule of individual autonomy. The regulatory mandate for purely independent decision-making [54], while intended to protect the patient, creates significant tension with this reality. Our findings align with prior research showing that for vulnerable patients, including older adults and those from disadvantaged backgrounds, the ideal of pure autonomy is often practically unattainable, with many ceding decision-making to health care professionals [40]. The role of the family further complicates this picture. Research by Kogan et al [22] powerfully illustrates this dynamic, revealing that caregivers in phase 1 trials often feel compelled to silently support the patient's choice, withholding their own opinions to protect the patient's sense of agency. This act of "silent support" reveals a critical nuance:

XSL•FO
RenderX

family involvement is not always about co-opting the decision, but is often a mechanism for endorsing and enabling patient autonomy within a supportive relational context. Therefore, the direct transplantation of a universal, Western-centric informed consent model is not merely inappropriate; it is misaligned with patients' and their families' lived experiences. It fails to recognize what the evidence consistently shows: the decision-making unit is often the patient-family dyad, not the patient in isolation. A more refined understanding of autonomy is needed, one that transcends rigid individualism and embraces a model of relational autonomy. This perspective preserves the ethical foundation of informed consent while acknowledging the meaningful role of family in decision-making. It recognizes autonomy as contextually shaped, validating both independent and family-informed choices as legitimate expressions of self-determination.

Strengths and Limitations

This study demonstrates methodological innovation by triangulating interview data and OHC text discourse from Chinese patients with cancer, thereby enhancing the credibility and generalizability of the findings. Consistent prompting strategies and stringent data anonymization protocols minimized bias in data collection. Furthermore, the analytical approach itself represents a key strength. By using a form of triangulation—combining traditional human interpretation with 2 distinct LLMs—this study systematically mitigated the risk of singular interpretive bias. The convergence of themes across these different methods bolstered the validity of our core findings. In contrast, the unique themes identified by each approach ensured a more comprehensive and multifaceted understanding than any single method could have yielded.

Notwithstanding these strengths, several limitations should be noted. First, the interview sample was small and drawn from a single tertiary hospital in Shanghai, a major metropolitan area with advanced health care infrastructure. This sampling approach may limit the transferability of interview findings to patients in secondary hospitals, rural areas, or regions with different health care resources. To partially mitigate this limitation, we supplemented interview data with online discourse from 2 platforms that attract users from diverse geographic locations, though the online sample also skews toward users with internet access and digital literacy.

Second, the interview inclusion criteria systematically excluded patients who were more physically compromised or had lower educational attainment. This is particularly relevant given that "lack of cognition" emerged as a barrier to trial participation. Our findings regarding cognitive and informational barriers reflect the perspectives of relatively functional, educated patients; individuals with lower health literacy or more advanced disease may face even more pronounced challenges. The conclusions regarding cognitive factors should therefore be understood as applicable primarily to patients meeting our inclusion criteria, and future research should specifically target underrepresented populations.

Third, the online data were collected from 2 Chinese platforms, and the final dataset of 219 comments, while sufficient for qualitative analysis, represents a subset of the broader online discourse. The findings should be interpreted within this scope rather than as representative of all online discussions of patients with cancer.

Fourth, the cross-sectional nature of this study captures attitudes at a single point in time. Patients' willingness to participate in clinical trials may evolve as their disease progresses, treatment options change, or new information becomes available. Longitudinal research would provide insights into how attitudes shift over the illness trajectory.

Implications for Future Work

Future research should address 3 priorities. First, multicenter studies that include community hospitals and rural settings are needed to capture perspectives from patients with lower educational attainment and limited health care access—populations excluded by the inclusion criteria but likely facing more pronounced barriers. Second, methodological work comparing different LLM prompting strategies and examining intermodel reliability would establish best practices for AI-assisted qualitative analysis. Third, the thematic findings suggest targets for intervention development: decision aids to address patient concerns, communication strategies to correct misconceptions, and family-inclusive recruitment approaches warrant evaluation in prospective trials.

Conclusion

This multistage qualitative study examined the attitudes of patients with cancer toward clinical trial participation in China using a hybrid analytical framework that combines investigator-led thematic analysis with LLM-assisted approaches. Seven core themes emerged as consistent factors shaping patients' willingness to participate: treatment selection, financial burden relief, uncertainty about therapeutic efficacy, uncertainty about control groups, lack of cognition, misconceptions, and physician trust. These findings highlight the interplay of pragmatic, cognitive, and relational factors in decision-making. The convergence of these themes across all 3 analytical methods strengthens their validity. Method-specific findings, including "recognition of medical value" (investigator only), "insufficient clinical data" (DeepSeek R1 only), and "lack of information resource" (Gemini Pro 2.5 only), demonstrate the complementary value of integrating human interpretation with LLM efficiency. Practically, recruitment strategies should bridge knowledge gaps, acknowledge family involvement, and address regional disparities in trial access. The hybrid framework presented here offers a replicable model for qualitative health research seeking both analytical rigor and interpretive depth.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank all patients and their families for their participation in and contributions to this study. Large language models (DeepSeek R1 and Gemini Pro 2.5) were used solely for qualitative data analysis as part of the study methodology. No artificial intelligence tools were used for writing, editing, or generating any portion of this manuscript text.

Funding

This work was supported by the Shanghai Jiao Tong University School of Medicine: Nursing Development Program, Ministry of Education in China Project of Humanities and Social Sciences (22YIC630043), and the Talent Team Building Project of the Nursing Discipline of Renji Hospital Affiliated to Shanghai Jiao Tong University School of Medicine (2021-2025).

Data Availability

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Authors' Contributions

Conceptualization: SY, HM, KX, HL, LM, MJ
Formal analysis: SY
Funding acquisition: SY, MJ
Investigation: SY
Methodology: SY, MJ
Supervision: MJ
Writing – original draft: SY
Writing – review & editing: SY, ML, HM, KX, HL, LM, MJ

Conflicts of Interest

None declared.

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74(3):229-263. [doi: 10.3322/caac.21834] [Medline: 38572751]
2. Alotaibi M, Alsolami L, AlHarbi S, et al. Clinical trial awareness, perceptions, and participation among cancer patients in Saudi Arabia. Healthcare (Basel) 2025 May 1;13(9):1044. [doi: 10.3390/healthcare13091044] [Medline: 40361823]
3. Harrer S, Shah P, Antony B, Hu J. Artificial intelligence for clinical trial design. Trends Pharmacol Sci 2019 Aug;40(8):577-591. [doi: 10.1016/j.tips.2019.05.005] [Medline: 31326235]
4. Treweek S, Lockhart P, Pitkethly M, et al. Methods to improve recruitment to randomised controlled trials: Cochrane systematic review and meta-analysis. BMJ Open 2013;3(2):e002360. [doi: 10.1136/bmjopen-2012-002360] [Medline: 23396504]
5. Wegge-Larsen C, Mehlsen M, Jensen AB. The motivation of breast cancer patients to participate in a national randomized control trial. Support Care Cancer 2023 Jul 15;31(8):468. [doi: 10.1007/s00520-023-07930-0] [Medline: 37452876]
6. Sateren WB, Trimble EL, Abrams J, et al. How sociodemographics, presence of oncology specialists, and hospital cancer programs affect accrual to cancer treatment trials. J Clin Oncol 2002 Apr 15;20(8):2109-2117. [doi: 10.1200/JCO.2002.08.056] [Medline: 11956272]
7. Institute of Medicine (US) Forum on Drug Discovery, Development, and Translation. Transforming Clinical Research in the United States: Challenges and Opportunities: Workshop Summary: National Academies Press; 2010. [doi: 10.17226/12900]
8. Murthy VH, Krumholz HM, Gross CP. Participation in cancer clinical trials: race-, sex-, and age-based disparities. JAMA 2004 Jun 9;291(22):2720-2726. [doi: 10.1001/jama.291.22.2720] [Medline: 15187053]
9. Thomas C, Mulnick S, Krucien N, Marsh K. How do study design features and participant characteristics influence willingness to participate in clinical trials? Results from a choice experiment. BMC Med Res Methodol 2022 Dec 16;22(1):323. [doi: 10.1186/s12874-022-01803-6] [Medline: 36526978]
10. Gillies K, Cotton SC, Brehaut JC, Politi MC, Skea Z. Decision aids for people considering taking part in clinical trials. Cochrane Database Syst Rev 2015 Nov 27;2015(11):CD009736. [doi: 10.1002/14651858.CD009736.pub2] [Medline: 26613337]
11. Cameron P, Pond GR, Xu RY, Ellis PM, Goffin JR. A comparison of patient knowledge of clinical trials and trialist priorities. Curr Oncol 2013 Jun;20(3):e193-e205. [doi: 10.3747/co.20.1323] [Medline: 23737689]

XSL•FO
RenderX

12. Alemayehu C, Mitchell G, Nikles J. Barriers for conducting clinical trials in developing countries—a systematic review. Int J Equity Health 2018 Mar 22;17(1):37. [doi: 10.1186/s12939-018-0748-6] [Medline: 29566721]
13. Bell JAH, Balneaves LG. Cancer patient decision making related to clinical trial participation: an integrative review with implications for patients' relational autonomy. Support Care Cancer 2015 Apr;23(4):1169-1196. [doi: 10.1007/s00520-014-2581-9] [Medline: 25591627]
14. Nielsen ZE, Berthelsen CB. Cancer patients' perceptions of factors influencing their decisions on participation in clinical drug trials: a qualitative meta-synthesis. J Clin Nurs 2019 Jul;28(13-14):2443-2461. [doi: 10.1111/jocn.14785] [Medline: 30673153]
15. Sanders R, Linn AJ, Araujo TB, Vliegenthart R, van Eenbergen MC, van Weert JCM. Different platforms for different patients' needs: automatic content analysis of different online health information platforms. Int J Hum Comput Stud 2020 May;137:102386. [doi: 10.1016/j.ijhcs.2019.102386]
16. Murphy S, Peelo-Dennehy D, Morrissey K, McCarthy J, Foley S. The opportunities and challenges of using online social platforms after pregnancy loss: a systematic review of qualitative research. Comput Hum Behav Rep 2025 Mar;17:100547. [doi: 10.1016/j.chbr.2024.100547]
17. De Paoli S. Performing an inductive thematic analysis of semi-structured interviews with a large language model: an exploration and provocation on the limits of the approach. Soc Sci Comput Rev 2024 Aug;42(4):997-1019. [doi: 10.1177/08944393231220483]
18. Shanwetter Levit N, Saban M. When investigator meets large language models: a qualitative analysis of cancer patient decision-making journeys. NPJ Digit Med 2025 Jun 5;8(1):336. [doi: 10.1038/s41746-025-01747-3] [Medline: 40473767]
19. DeepSeek API. URL: https://www.deepseek.com/ [accessed 2025-08-03]
20. Internet research: ethical guidelines 3.0. : Association of Internet Researchers; 2019 URL: https://aoir.org/reports/ethics3.pdf [accessed 2026-04-27]
21. Dijkman BL, Luttik ML, Van der Wal-Huisman H, Paans W, van Leeuwen BL. Factors influencing family involvement in treatment decision-making for older patients with cancer: a scoping review. J Geriatr Oncol 2022 May;13(4):391-397. [doi: 10.1016/j.jgo.2021.11.003] [Medline: 34776380]
22. Kogan N, Tolley M, Cohen R. Family caregivers support their patient's decision to participate in a phase 1 clinical trial: weighing the pros and cons, bearing the costs. Psychooncology 2022 Aug;31(8):1374-1380. [doi: 10.1002/pon.5942] [Medline: 35460308]
23. King TL, Tan SH, Tan SSN, Lai WH, Bujang MA, Voon PJ. Survey of willingness to participate in clinical trials and influencing factors among cancer and non-cancer patients. Sci Rep 2025 Jan 10;15(1):1626. [doi: 10.1038/s41598-024-83626-7] [Medline: 39794348]
24. Gayet-Ageron A, Rudaz S, Perneger T. Study design factors influencing patients' willingness to participate in clinical research: a randomised vignette-based study. BMC Med Res Methodol 2020 Apr 26;20(1):93. [doi: 10.1186/s12874-020-00979-z] [Medline: 32336266]
25. Lu W, Li J, Gu S, Li J. Attitude of patients with liver cancer toward clinical drug trials: empirical qualitative research. Sci Rep 2025 Jul 1;15(1):21026. [doi: 10.1038/s41598-024-83162-4]
26. AbouWarda H, Dolata M, Schwabe G. How does an online mental health community on Twitter empower diverse population levels and groups? A qualitative analysis of #BipolarClub. J Med Internet Res 2024 Aug 19;26:e55965. [doi: 10.2196/55965] [Medline: 39158945]
27. Lee YJ, Kamen C, Margolies L, Boehmer U. Online health community experiences of sexual minority women with cancer. J Am Med Inform Assoc 2019 Aug 1;26(8-9):759-766. [doi: 10.1093/jamia/ocz103] [Medline: 31361002]
28. Daly J, Lumley J. Bias in qualitative research designs. Aust N Z J Public Health 2002 Aug;26(4):299-300. [doi: 10.1111/j.1467-842x.2002.tb00174.x] [Medline: 12233947]
29. Chua GP, Tan HK. A qualitative approach in determining the patient-centered information and supportive care needs of cancer patients in Singapore. BMJ Open 2020 Feb 28;10(2):e034178. [doi: 10.1136/bmjopen-2019-034178] [Medline: 32114472]
30. Sibeoni J, Picard C, Orri M, et al. Patients' quality of life during active cancer treatment: a qualitative study. BMC Cancer 2018 Oct 4;18(1):951. [doi: 10.1186/s12885-018-4868-6] [Medline: 30286733]
31. Vera San Juan N, Martin S, Badley A, et al. Frontline health care workers' mental health and well-being during the first year of the COVID-19 pandemic: analysis of interviews and social media data. J Med Internet Res 2023 Aug 14;25:e43000. [doi: 10.2196/43000] [Medline: 37402283]
32. Alizadeh M, Kubli M, Samei Z, et al. Open-source LLMs for text annotation: a practical guide for model setting and fine-tuning. J Comput Soc Sci 2025;8(1):17. [doi: 10.1007/s42001-024-00345-9] [Medline: 39712076]
33. Muñoz-Ortiz A, Gómez-Rodríguez C, Vilares D. Contrasting linguistic patterns in human and LLM-generated news text. Artif Intell Rev 2024;57(10):265. [doi: 10.1007/s10462-024-10903-2] [Medline: 39328400]
34. Li Y, Liu Y, Yu M. Consumer segmentation with large language models. J Retail Consum Serv 2025 Jan;82:104078. [doi: 10.1016/j.jretconser.2024.104078]
35. Wang C, Pan D, Kumari S, Wu ME, Jiang H. Large language models driven health text information analysis in consumer electronics. IEEE Trans Consumer Electron 2024 May;71(2):3510-3521. [doi: 10.1109/TCE.2024.3516006]

XSL•FO
RenderX

36. Hogan TH, Maurer NR, Stevens L, Hefner JL. A framework of the institutional policies and practice environments of nurse practitioner primary care models: a cross-case analysis. Health Care Manage Rev 2022;47(4):369-379. [doi: 10.1097/HMR.0000000000000344] [Medline: 35713574]
37. Strachan H, Hoskins G, Wells M, Maxwell M. A realist evaluation case study of the implementation of advanced nurse practitioner roles in primary care in Scotland. J Adv Nurs 2022 Sep;78(9):2916-2932. [doi: 10.1111/jan.15252] [Medline: 35436359]
38. Murphy M, McCaughan E, Carson MA, Donovan M, Wilson RH, Fitzsimons D. Nothing to lose: A grounded theory study of patients' and healthcare professionals' perspectives of being involved in the consent process for oncology trials with non-curative intent. BMC Palliat Care 2020 Oct 30;19(1):166. [doi: 10.1186/s12904-020-00661-7] [Medline: 33126874]
39. Dellson P, Nilsson K, Jernström H, Carlsson C. Patients' reasoning regarding the decision to participate in clinical cancer trials: an interview study. Trials 2018 Sep 29;19(1):528. [doi: 10.1186/s13063-018-2916-9] [Medline: 30268150]
40. Walsh E, Sheridan A. Factors affecting patient participation in clinical trials in Ireland: a narrative review. Contemp Clin Trials Commun 2016 Aug 15;3:23-31. [doi: 10.1016/j.conctc.2016.01.002] [Medline: 29736453]
41. Hall BA. The struggle of the diagnosed terminally ill person to maintain hope. Nurs Sci Q 1990;3(4):177-184. [doi: 10.1177/089431849000300410] [Medline: 2250837]
42. Abdelhafiz AS, Abd ElHafeez S, Khalil MA, et al. Factors influencing participation in COVID-19 clinical trials: a multi-national study. Front Med (Lausanne) 2021;8:608959. [doi: 10.3389/fmed.2021.608959] [Medline: 33708777]
43. Murphy M, McCaughan E, Thompson G, et al. Trusting relationships between patients with non-curative cancer and healthcare professionals create ethical obstacles for informed consent in clinical trials: a grounded theory study. BMC Palliat Care 2023 Jul 1;22(1):85. [doi: 10.1186/s12904-023-01204-6] [Medline: 37393250]
44. Sarwar F, Ring D, Donovan E. Clinician communication strategies to navigate differences of opinion with patients. Patient Educ Couns 2024 Jun;123:108185. [doi: 10.1016/j.pec.2024.108185] [Medline: 38340633]
45. Sun J, Fang J, Zhang C, et al. Acceptance factors and psychological investigation of clinical trials in cancer patients. Behav Neurol 2023;2023(1–8):5617575. [doi: 10.1155/2023/5617575] [Medline: 38046475]
46. Kim JY, Florez M, Botto E, Belgrave X, Grace C, Getz K. The influence of socioeconomic status on individual attitudes and experience with clinical trials. Commun Med (Lond) 2024 Sep 5;4(1):172. [doi: 10.1038/s43856-024-00586-9] [Medline: 39237734]
47. Pullen MF, Pastick KA, Williams DA, et al. Lessons learned from conducting internet-based randomized clinical trials during a global pandemic. Open Forum Infect Dis 2021 Feb;8(2):ofaa602. [doi: 10.1093/ofid/ofaa602] [Medline: 33553471]
48. Sine S, de Bruin A, Getz K. Patient engagement initiatives in clinical trials: recent trends and implications. Ther Innov Regul Sci 2021 Sep;55(5):1059-1065. [doi: 10.1007/s43441-021-00306-8] [Medline: 34097288]
49. Bateman KP. The development of patient-centric sampling as an enabling technology for clinical trials. Bioanalysis 2020 Jul;12(13):971-976. [doi: 10.4155/bio-2020-0075] [Medline: 32628048]
50. Vayena E, Blasimme A, Sugarman J. Decentralised clinical trials: ethical opportunities and challenges. Lancet Digit Health 2023 Jun;5(6):e390-e394. [doi: 10.1016/S2589-7500(23)00052-3] [Medline: 37105800]
51. El Masri H, McGuire TM, van Driel ML, Benham H, Hollingworth SA. Dynamics of patient-based benefit-risk assessment of medicines in chronic diseases: a systematic review. Patient Prefer Adherence 2022;16:2609-2637. [doi: 10.2147/PPA.S375062] [Medline: 36164323]
52. Jenkins V, Farewell D, Batt L, et al. The attitudes of 1066 patients with cancer towards participation in randomised clinical trials. Br J Cancer 2010 Dec 7;103(12):1801-1807. [doi: 10.1038/sj.bjc.6606004] [Medline: 21119659]
53. Tielemans LML, van Heugten KD, Hamaker ME, van Walree IC. Preferred and actual involvement of caregivers in oncologic treatment decision-making: a systematic review. J Geriatr Oncol 2023 Jul;14(6):101525. [doi: 10.1016/j.jgo.2023.101525] [Medline: 37198027]
54. Biedrzycki BA. Factors and outcomes of decision making for cancer clinical trial participation. Oncol Nurs Forum 2011 Sep;38(5):542-552. [doi: 10.1188/11.ONF.542-552] [Medline: 21875841]

Abbreviations

AI: artificial intelligence
LLM: large language model
OHC: online health community

XSL•FO
RenderX

Edited by A Coristine; submitted 01.Dec.2025; peer-reviewed by I Tosun, MA Bujang; revised version received 31.Mar.2026; accepted 31.Mar.2026; published 05.May.2026.

Please cite as:
Yu S, Lyu M, Ma H, Xu K, Li H, Ma L, Ji M
Attitudes and Willingness to Participate in Drug Clinical Trials Among Patients With Cancer: Multistage Qualitative Study
J Med Internet Res 2026;28:e88758
URL: https://www.jmir.org/2026/1/e88758
doi:10.2196/88758

© Sijia Yu, Mengmeng Lyu, Haiping Ma, Keying Xu, Haiyan Li, Lili Ma, Mengting Ji. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Implementation of a Business Process Model and Notation for Modeling Patient Health Care Trajectories: Systematic Review

Jean-Baptiste Gartner[1], PhD; Paolo Landa[2], PhD; Matthew T Haren[1], PhD; Célia Lemaire[3], PhD; Elena Tanfani[4], PhD; Catherine Paquet[5], PhD; Frédéric Bergeron[6]; André Côté[1], PhD

[1]Department of Management, Faculty of Administrative Sciences, Laval University, 2325 rue de la Terrasse, Québec, QC, Canada

[2]Department of Operations and Decision Systems, Faculty of Administrative Sciences, Laval University, Québec, QC, Canada

[3]iaelyon School of Management, Université Lyon 3, Lyon, France

[4]Department of Economics, University of Genova, Genova, Italy

[5]Marketing Department, Faculty of Administrative Sciences, Laval University, Québec, QC, Canada

[6]Library and Advisory Services Division, Laval University, Québec, QC, Canada

Corresponding Author:
Jean-Baptiste Gartner, PhD
Department of Management, Faculty of Administrative Sciences, Laval University, 2325 rue de la Terrasse, Québec, QC, Canada

Abstract

Background: Health care systems are increasingly confronted with the challenge of managing complex clinical processes. One proposed solution is a patient-centered management intervention called a care pathway that needs process mapping to support process improvement. Although the adoption and use of Business Process Modeling Notation (BPMN) for modeling patient health care trajectories has increased, evidence of the benefits of implementing it in health care organization management systems remains unclear.

Objective: This review sought to examine effectiveness by mapping implementation factors linking intended purpose to expected or demonstrated outcomes.

Methods: A systematic review of the use of BPMN for modeling patient health care trajectories was conducted across 8 databases. We followed the Cochrane Methods Group and the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. We selected empirical, experimental, and conceptual articles in French and English that analyzed and evaluated the implementation of BPMN for modeling patient health care trajectories in the health care field. Quality appraisal was performed using the Mixed Methods Appraisal Tool. Data were charted using a customized form and analyzed thematically with qualitative and semiquantitative syntheses.

Results: After screening, 61 studies were included. BPMN offers significant benefits in health care. Its use allows health care professionals to gain an understanding of patient health care trajectories, making it easier to identify inefficiencies and areas for improvement. The definition of processes ensures that workflows remain consistent across different settings, thereby reducing variation and improving the quality of care. Several studies have demonstrated BPMN's effectiveness in process optimization, highlighting its ability to streamline workflows, reduce redundancies, and enhance operational efficiency. Moreover, when integrated with decision-support tools, BPMN enhances clinical decision-making by enabling better adherence to guidelines and best practices. Another advantage is BPMN's interoperability with existing health care IT standards. However, managerial considerations reveal trade-offs between BPMN's benefits and limitations, especially in highly complex health care settings. Several challenges persist, including issues related to scalability, integration with advanced decision-making frameworks, and the complexity of modeling dynamic health care environments. While BPMN is widely adopted, alternative methodologies offer complementary or competing advantages. Several opportunities exist to enhance BPMN's applicability in health care, such as creating domain-specific BPMN extensions or integrating artificial intelligence and machine learning. However, limitations in the methodological quality of the studies selected and the concentration of the research mainly in European countries limit the generalizability of our results.

Conclusions: This review highlights BPMN's potential as a valuable tool for modeling patient health care trajectories. Its ability to standardize and optimize processes makes it promising for improving clinical and operational efficiency. However, trade-offs between benefits and limits of BPMN characterize its implementation in patient health care trajectories, creating opportunities for the development and integration of new tools.

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2021-060357

(J Med Internet Res 2026;28:e78506) doi:10.2196/78506

XSL•FO
RenderX

KEYWORDS

Business Process Model and Notation; patient health care trajectories; implementation; effectiveness; transformation

Introduction

In many countries, health care systems are increasingly confronted with the challenge of managing complex clinical processes, where patient care often involves multiple departments, professionals, and stages [1]. In addition, health care systems are facing a rise in multimorbidity, along with a growing population of older adults. These conditions have significantly increased the public expectation for health care services, both in terms of volume and service quality [2]. In response, health care organizations have prioritized efficient and effective management practices to improve outcomes and outputs, all while managing resource and budget constraints [3]. Several national and international organizations have developed frameworks for improving the performance of health care and services on the dimensions of quality, safety, effectiveness, efficiency, patient-centeredness, timeliness, and equity [4-7]. These goals represent significant challenges that face any health care setting [8-10].

One proposed solution is a patient-centered management intervention called a care pathway [11-14], which is designed to guide patients, in a specific patient segment, through a trajectory of care representing the entire continuum of care, from prevention and screening to recovery or palliative care. These interventions are efficient in structuring and managing patient care, improving outcomes, accessibility, quality, sustainability, and cost-effectiveness [15-18]. However, the standardization, representation, and integration of the care pathway within management systems, including their ongoing process improvement, present further challenges for health care organizations.

During the last 2 decades, several tools have been developed to support process improvement through process mapping. Health care organizations began to adopt such tools as Business Process Modeling Notation (BPMN) to support various business processes, including the delivery of care [19,20]. Originally designed for business environments to represent processes as a network of activities and tasks [19], BPMN has proven to be an effective tool in health care by providing a standardized visual language for modeling processes [21] and has become one of the most important and widely used tools in this context [11]. Developed by the Object Management Group in 2004, BPMN has been adopted as an international standard by the International Organization for Standardization since 2012. Now in its second version, the BPMN 2.0 [20] enables the introduction of extensions that characterize specific domains (eg, health care, quality management, and security) consistent with original BPMN elements [22,23]. The main value of the extension is in the reuse of BPMN's main functions, maintaining its standardization, without the need for developing domain-specific modeling languages [24].

Although the literature regarding the adoption and use of BPMN for modeling patient health care trajectories has increased over the last decade, existing reviews are limited. A preliminary search for existing reviews was conducted in the Cochrane Database, JBI Database of Systematic Reviews and Implementation Reports, and PROSPERO, and 4 literature reviews on BPMN in the health care context were identified [23,25-27]. The first, carried out in 2014, focused on clinical decision support [25]. The next 2 focused on the ability to formalize the process and standardized communication [26] or the possibility of incorporating variations or changes [27]. Finally, the last dealt only with extensions to the notation [23]. None of these reviews have addressed the effectiveness and characteristics of BPMN implementation nor the benefits and limitations of its use, particularly from a managerial perspective.

Thus, this study seeks to fill this gap and critically synthesize the empirical evidence on the uses of BPMN for modeling patient health care trajectories with a focus on implementation and effectiveness. We have adopted the Population, Intervention, Comparator, Outcomes, Timing, Settings (PICOTS) mnemonic criteria [28] in the development of our review objective and research questions. The objective of this systematic review is to examine the evidence linking the implementation of BPMN modeling of health care trajectories (I, P) to both management and clinical outcomes (O) in clinical health care settings (S). Within this objective, we seek to answer 2 specific research questions: how well do the objectives for using BPMN to model health care trajectories align with the realized outcomes, and what are the potential implementation factors (including the use of extensions) that link objectives to outcomes?

Methods

Overview

The protocol for this review was previously published as a scoping review protocol [29]. However, further development of the objectives and aims justified the conduct of a systematic rather than a scoping review. This systematic review was carried out in accordance with the Cochrane Methods Group [30] and the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [31]. Following them, we adapted the method from the previously published protocol by (1) reformulating the research question using PICOTS rather than Population Concept and Context, (2) excluding literature reviews from the article selection process, (3) incorporating a quality assessment, and (4) completing the PRISMA 2020 checklist for systematic reviews (Checklist 1).

Information Sources and Search Strategy

The search strategy was developed in collaboration with an academic librarian specializing in medical and health care management fields (FB). Databases covering health (PubMed, Embase, and CINAHL) and business (ABI/INFORM) disciplines and multidisciplinary (Academic Search Premier, Web of Science, and ScienceDirect) databases, as well as the Google Scholar search engine, were searched to identify eligible peer-reviewed articles. The last search of each database was performed on January 5, 2026.

XSL•FO
RenderX

Key search terms were informed by previous relevant reviews and are shown in Table 1, mapped to the PICOTS framework. Searches in the electronic databases were conducted from January 1, 2004, as the BPMN was released in its first version to the public in May 2004. For Google Scholar, we limited results to the first 20 results per string and filtered out citations and patents. The full search strategy for each database is provided in Multimedia Appendix 1.

Table . Search terms mapped to PICOTS[a] plus limits and filters.

PICOTS	Search terms
Patient population	Not specified
Intervention	Business Process* (Model* OR Method? OR management) OR Decision Model* notation OR BPMN* OR BPM
Comparator	Not specified
Outcomes	Not specified
Timing	Not specified
Settings	Critical Pathways OR Practice Guidelines OR Workflow OR Clinical Decision-Making OR Decision Support Systems, Clinical OR Patient Care Management
Settings	Decision (making OR support) OR (clinical OR medical OR healthcare OR “health care”) process* OR (healthcare OR clinical OR critical OR care) path* OR guideline* OR Workflow* OR careflow* OR patient journey OR Healthcare trajectory
Other limits	Dates: January 1, 2004 to present (last search January 5, 2026) For ABI/INFORM (ProQuest): peer-reviewed publications only For Google Scholar: first 20 results per string only; filtered out citations and patents

[a]PICOTS: Population, Intervention, Comparator, Outcomes, Timing, Settings.

Eligibility Criteria

The eligibility criteria aligned with the PICOTS are provided in Table 2.

XSL•FO
RenderX

Table . Eligibility criteria mapped to the PICOTS[a] framework.

PICOTS	Inclusion criteria	Exclusion criteria
Patient population	• No limits within patient health care trajectories	—[b]
Intervention	• Related to BPMN[c] in health care trajectories • Models a care process ideally citing BPMN	• Not related to BPMN in health care • Clearly specifies the exclusive use of a notation other than BPMN (such as UML[d], BPEL[e], HL7[f], ARIS, etc) • Addresses another type of business process management other than process modeling • Computer programming and coding (eg, focus on HL7 or other standards)
Comparator	• No limits	—
Outcomes	• Any indicator to estimate the improvement in process management or clinical outcome	—
Timing	• No limits	—
Settings	• Explicitly in the health care field • Health care trajectories or clinical or care pathways	• Theoretical demonstration of a new tool • Evaluation process of a tool without application (eg, metrics)
Other limits	• Articles in French and English only	• Theses and dissertations • Editorials • Literature reviews and protocols

[a]PICOTS: Population, Intervention, Comparator, Outcomes, Timing, Settings.

[b]Not applicable.

[c]BPMN: Business Process Modeling Notation.

[d]UML: Unified Modeling Language.

[e]BPEL: Business Process Execution Language.

[f]HL7: Health Level Seven.

Study Selection

The search results were imported to Covidence (Veritas Health Innovation Ltd) [32] to remove duplicates and manage the study selection process. Study selection proceeded in 2 phases, beginning with a title and abstract screening, followed by a full-text review of retained records. At each phase, a pair of reviewers consisting of 2 of the 3 coauthors (JBG, PL, and MTH) independently screened the titles and abstracts (phase 1) or reviewed full texts (phase 2) against the inclusion and exclusion criteria. Conflicts were managed by discussion to reach a consensus, or, when necessary, an additional reviewer (AC) was consulted.

Data Extraction and Data Items

A pair of reviewers consisting of 2 of the 4 coauthors (PL, JBG, MTH, and CL) independently extracted data from the included studies using a custom-developed data extraction form in Microsoft Excel which included the following data fields: “citation details” (title, authors, year of publication, and author affiliations), “study description” (study design, setting, care trajectory, aims or objectives, key variables analyzed, and extensions to BPMN used), “study results” (findings, outcomes, and study limitations), “BPMN utility” (objective for use, benefit or advantage of use, limit of use, opportunities, and alternatives or threats). JBG and PL reviewed all the data extraction tables, presented in Multimedia Appendix 2. Discrepancies were addressed through discussion between at least 2 reviewers.

Quality Assessment

Quality assessment was performed using the Mixed Methods Appraisal Tool (MMAT) of July 2020 [33], which uses five core quality criteria to assess each of the following study designs: (1) qualitative, (2) randomized controlled, (3) nonrandomized, (4) quantitative descriptive, and (5) mixed methods. Assessments were performed independently by 2 reviewers (MTH and JBG). Disagreements were addressed through discussion and, where needed, a third independent assessor (PL or AC) was consulted. A 3-level summary of quality was assigned to each study based on the number of achieved criteria, where 5=high quality, 3‐4=medium quality, and ≤2=low quality. For mixed methods studies, the lowest scoring of 3 study design sets (qualitative, quantitative, or mixed) was used to assign the summary level. Due to contention in the literature about the use of summative approaches in critical appraisal [34-37], we also provide a detailed presentation of the ratings (refer to Table 3 in the “Results” section). Studies that failed the MMAT screening questions S1 or S2 due to a lack of a clear research question were categorized as

XSL•FO
RenderX

"Experience feedback" and were not formally appraised for quality.

Table . Results of the quality assessment with the MMAT[a] tool.

Title	Experience feedback (no research question)	Qualitative studies	Quantitative studies	Mixed methods studies	MMAT quality level
Number of studies	46	4	3	8	
Distribution of quality scores					
0	—[b]	—	—	2	Low quality
1	—	—	—	—	Low quality
2	—	—	1	—	Low quality
3	—	—	1	1	Medium quality
4	—	1	—	4	Medium quality
5	—	3	1	1	High quality

[a]MMAT: Mixed Methods Appraisal Tool.

[b]Not applicable.

Data Analysis and Synthesis

Descriptive data items were reduced to meaningful categories as presented in Table 4. The characteristics of the included studies were analyzed descriptively, and we used multidimensional scaling in Orange (v3.38; Bioinformatics Lab) to visualize the relations within the body of literature across 3 dimensions: geography, study design, and study (health care) setting.

The text data extracted in the "Study Results" and "BPMN utility" fields underwent inductive thematic coding. This was done by examining extracts for keywords and phrases, grouping like keywords and phrases into subthemes, and further grouping those subthemes into broader themes to generate a 2-level coding tree by which extracts were coded.

XSL•FO
RenderX

Table . Characteristics of the selected studies.

Characteristic and category	Value, n (%)	References
Publication year		
2008/09	3 (4.9)	[38-40]
2010/11	0 (0.0)	__[a]
2012/13	3 (4.9)	[41-43]
2014/15	11 (18.0)	[44-54]
2016/17	13 (21.3)	[55-67]
2018/19	9 (14.8)	[68-76]
2020/21	6 (9.8)	[77-82]
2022/23	8 (13.1)	[83-90]
2024/25	8 (13.1)	[91-98]
Geography		
Africa	4 (6.6)	[56,62,63,84]
Asia	3 (4.9)	[55,64,80]
Europe	42 (68.9)	[38-49,51-54,57,58,60,61,65,66,68,69,71-74,77,79,82,83,85-90,92-95,97]
Middle East	3 (4.9)	[67,91,98]
North America	3 (4.9)	[47,59,76]
South and Central America	6 (9.8)	[50,70,74,78,83,96]
Study design		
Empirical	30 (49.2)	[43,51-53,64,67,70,72,74,75,79-98]
Experimental	20 (32.8)	[41,42,44-47,54,56,57,60,62,63,65,66,69,71,73,76-78]
Conceptual or theoretical	11 (18.0)	[38-40,48-50,55,58,59,61,68]
Study setting		
Hospital	29 (47.5)	[41-43,45,48,52,54-56,58,64-72,74,76,79,83,84,86-88,91,94]
Integrated care	11 (18.0)	[38,40,53,60,73,75,78,81,92,96,97]
Specialist or multidisciplinary outpatient clinic	6 (9.8)	[47,80,82,85,93,95]
Primary care	5 (8.2)	[49,63,77,89,90]
Emergency department	5 (8.2)	[39,44,51,57,62]
Home care	1 (1.6)	[98]
Undefined	4 (6.6)	[46,50,59,61]
Care trajectory		
Acute care	27 (44.3)	[39,41,44-46,51,54,57-59,62,64,66,67,71,74,78,82-87,91,94,95,97]
Chronic care	20 (32.8)	[40,43,47-49,52,56,60,63,65,68,70,75,77,79,81,88-90,93]
Integrated care	6 (9.8)	[42,53,73,80,92,96]
Preventive care	2 (3.3)	[69,76]
Other	6 (9.8)	[38,50,55,61,72,98]
BPMN[b] extension use		
Yes	48 (78.7)	[38-60,62,63,65,66,68-73,75-78,80,82,84-89,92,95,96]
No	11 (18.0)	[67,74,79,81,83,90,91,93,94,97,98]
Unclear	2 (3.3)	[61,64]
BPMN+ extension and additional tools or standards		
ABC[c]	2 (3.3)	[73,77]

XSL•FO
RenderX

Characteristic and category	Value, n (%)	References
BPM[d] lifecycle	1 (1.6)	[62]
BPMN4CP[e]	6 (9.8)	[40,46,50,51,56,58]
BPMN+V[f]	2 (3.3)	[75,88]
BPSim[g]	2 (3.3)	[57,95]
CPG[h]	5 (8.2)	[38,48,52,63,69]
Clinical safety checklist	1 (1.6)	[64]
DMN[i]	5 (8.2)	[45,60,68,76,87]
EVALAB graphical interface	1 (1.6)	[44]
FHIR[j]	3 (4.9)	[85,89,96]
HACCP[k]	1 (1.6)	[72]
ICT[l]	6 (9.8)	[39,41,47,53,55,59]
IT sensor	1 (1.6)	[65]
PROforma SIG	1 (1.6)	[90]
UML[m]	1 (1.6)	[54]
VSM[n]	2 (3.3)	[43,84]
BPMN ontology and SWRL[o]	1 (1.6)	[49]
BPMNsix[p]+ IEEE 11073 SDC	1 (1.6)	[71]
ICNP[q]+ computational tools (Bizagi Modeler Software)	1	[78]
t.BPMN[r]+ BPMN4 CP	1	[42]
DMN + CMMN[s]	3	[66,80,86]
DMN+ FHIR	1	[82]
eHealth and ubiquitous computing and knowledge management and clinical decision-support systems	1	[70]
HL7+[t] FHIR	1	[92]
Unknown	1	[61]

XSL•FO
RenderX

Characteristic and category	Value, n (%)	References
None	10	[67,74,79,81,83,91,93,94,97,98]

[a]Not applicable.
[b]BPMN: Business Process Modeling Notation.
[c]ABC: Activity-Based Costing.
[d]BPM: Business Process Management.
[e]BPMN4CP: BPMN for clinical pathways.
[f]BPMN+V: a data-enriched subset of BPMN1 suitable for modeling clinical guideline.
[g]BPSim: Business Processes Simulation 1.0.
[h]CPG: clinical practice guideline.
[i]DMN: Decision Modeling Notation.
[j]FHIR: Fast Healthcare Interoperability Resource.
[k]HACCP: Hazard Analysis and Critical Control Point.
[l]ICT: Information and Communication Technology.
[m]UML: Unified Modeling Language.
[n]VSM: Value Stream Map.
[o]SWRL: Semantic Web Rule Language.
[p]BPMNsix + IEEE 11073 SDC: BPMN + Surgical Intervention Extension.
[q]ICNP: International Classification for Nursing Practice.
[r]t.BPMN: tangible Business Process Modeling.
[s]CMMN: Case Management Model Notation.
[t]HL7: Health Level Seven.

Frequency analyses of both broad and subthemes were performed on the BPMN benefits, limits, opportunities, and alternatives or threats variables to determine the most prevalent themes represented in the literature in these 4 domains. We then constructed 2 composite variables of thematic frequencies, one representing limits and benefits as a continuum (benefit-limit), and the other, opportunity and threat as a continuum (opp-threat). To do this, we negatively transformed the limit and threat frequencies to represent the negative sides of the continuums, whereas benefit and opportunity were represented as positive values. We then plotted opp-threat (*y*-axis) against benefit-limit (*x*-axis) to examine how limits and benefits related to threats or opportunities at a broad thematic level. We then narratively examined the subthemes in relation to the predominant relationships between broad themes in the plot and synthesized this with a narrative analysis of the major themes from the study findings fields.

Finally, we mapped this synthesis to a proposed causal pathway describing the mechanisms by which the purpose or objective for using BPMN to model patient health care trajectories links to desired (or observed) outcomes through characteristics of BPMN implementation expressed as benefits, limits, opportunities, and threats or alternatives. The mechanisms have been classified according to the domains of the Consolidated Framework for Implementation Research (CFIR) [99-101].

Results

Overview

Following the identification and removal of duplicates, 1177 unique records were identified. Screening against the eligibility criteria resulted in the retention of 253 studies, of which a further 192 were excluded by full-text review, resulting in 61 included studies [38-98]. This process is shown in the PRISMA 2020 flow diagram in Figure 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram of the systematic review process.

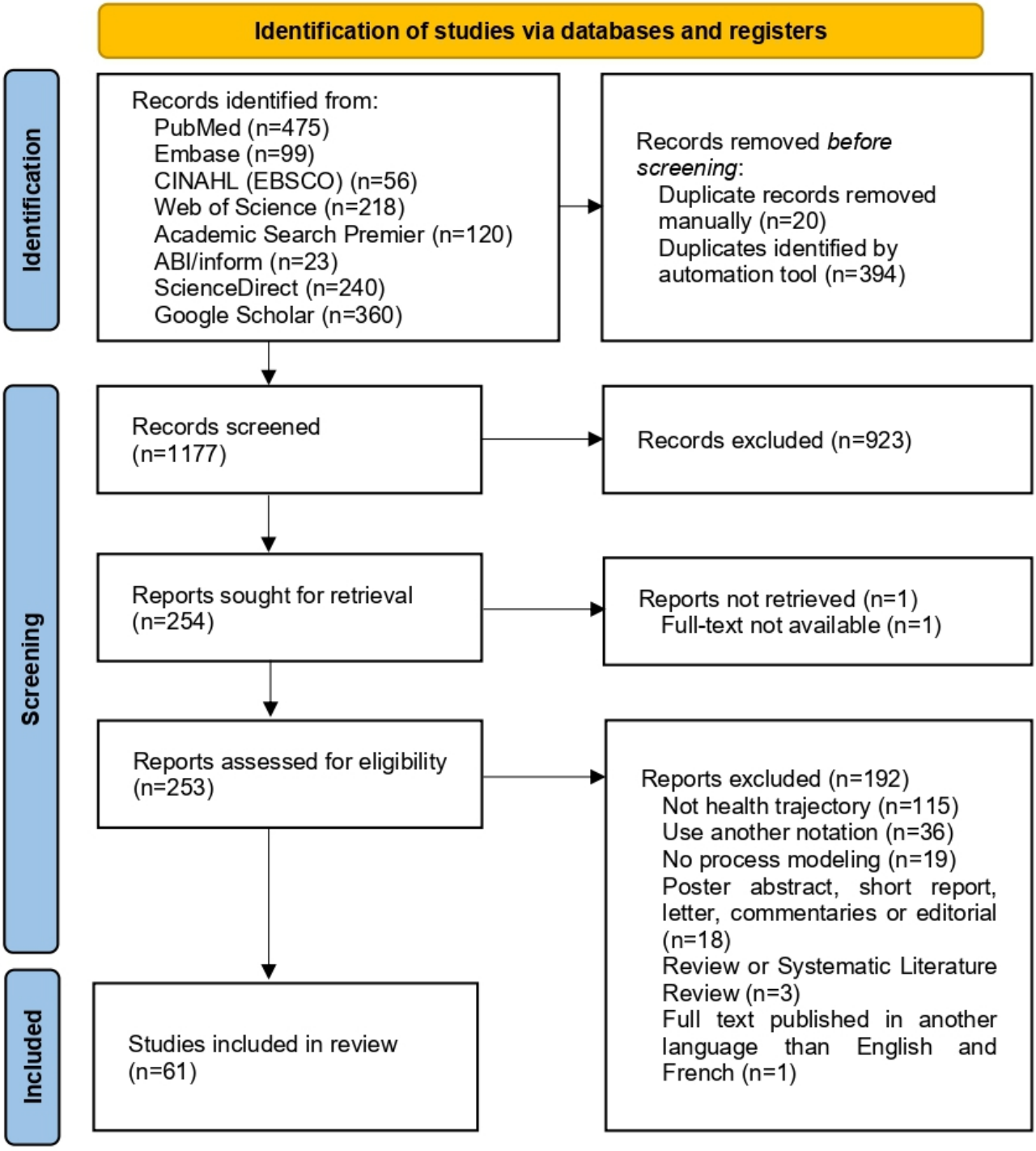

Study Characteristics

The characteristics of the selected literature are summarized in Table 4. The first published study of BPMN in modeling patient health care trajectories was in 2008 [38]. Most studies were performed in the European context and were conducted in a hospital setting, addressing a range of health care trajectories across acute and chronic care. The concentration of studies in Europe may be considered a bias for the validity of the study results.

A relational map of the selected literature in terms of geography, study design, and study (health care) setting is shown in Figure 2. These three dimensions are illustrated in the figure as follows: geography (marker colors), study design (marker symbols), and study (health care) setting (marker labels). The more tightly clustered a group of papers is, the more closely they are related to each other across these 3 dimensions. The size of the marker symbols is scaled to the goodness-of-fit of each case, where the smaller the marker symbol, the better the fit. This analysis shows the clear dominance of Europe (indicated in green) in the development of the evidence base from conceptual or theoretical roots (circle markers) through empirical (x markers), including experimental investigation (triangle markers). It is noteworthy that at each stage of this literature development, evidence was

XSL•FO
RenderX

contributed from a range of health care settings (hospital, emergency departments, primary care, and integrated care), reducing the risk of limitations to knowledge development arising from setting specificity. In contrast to this European dominance in the development of the literature, this analysis also suggests that the rest of the world has proceeded with empirical and experimental research based on European experience, with only a few examples of conceptual or theoretical research having been conducted outside of Europe. Like within Europe, the research outside of Europe represents a diverse evidence base developed from hospital, primary care, integrated care, emergency department, and specialist or multidisciplinary health care clinic settings.

Figure 2. A relational map of the selected literature in terms of geography, study design, and study (health care) setting.

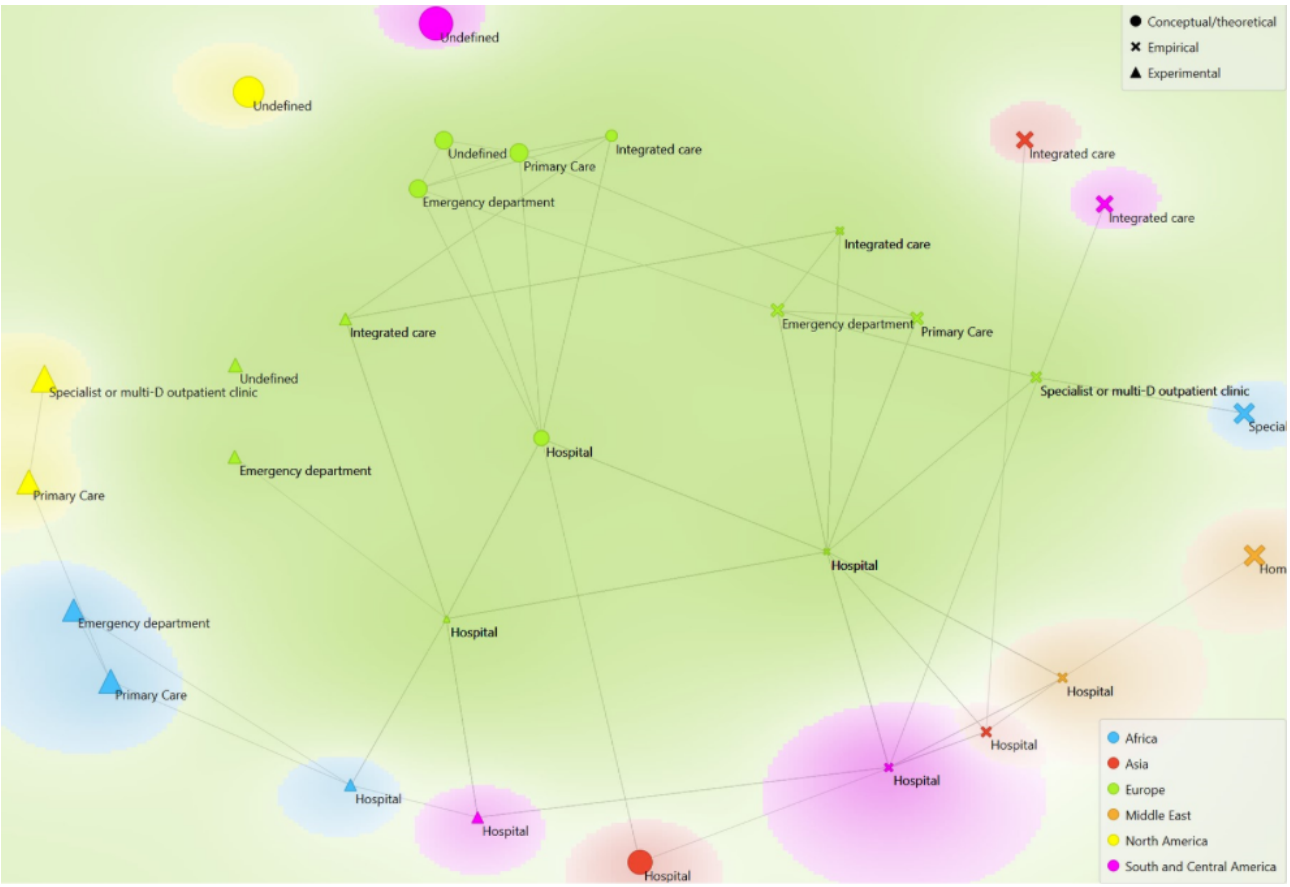

Quality Assessment

Table 3 presents the quality assessment using the MMAT tool of July 2020 [33]. The full scoring can be seen in Multimedia Appendix 3. Overall, the number of methodologically high-quality studies is low, which limits the scope and generalizability of our study's results.

Objectives for Using BPMN

The primary objective of using BPMN was to formalize care processes by modeling and visualizing them. Indeed, BPMN modeling of the patient health care trajectory involved the visual representation of workflow models and activities [41,42,44,46-49,51,54,62,64,66-70,77,80,83,85], roles and systems within care pathways [45,50,55,57,70,93,98]. Further, such representation was used to validate the workflow with the participants in the field [63] and even to outsource certain processes [40].

Some studies aimed to integrate dedicated extensions in predefined dimensions to overcome shortcomings of BPMN. These extensions had different objectives, such as (1) integrating decision support via Decision Modeling Notation (DMN) rating [72,85], or in combination with Case Management Model Notation (CMMN) [69], both developed by Object Management Group; (2) modeling clinical practice guidelines (CPGs) [80]; (3) incorporating the notion of value through the Value Stream Map notation derived from lean management [46,87]; (4) knowledge definition for a rich and expressive graphical representation [59]; (5) the Hazard Analysis and Critical Control Points [75]; (6) a dynamic approach to BPMN [82]; or (7) adaptation to specific process contexts such as modeling clinical pathways [55,61,93,95,97], or operating room processes [74].

Beyond simple process formalization, patient health care trajectory modeling using BPMN is also used to analyze existing processes, visualize constraints, and simulate changes to optimize operational performance. A comprehensive analysis [70] of existing processes using BPMN shows that it enables a specific focus to be placed on the resources and activities, to incorporate them as constraints of the optimization model [45] and to define opportunities for redesigning the process [53,60,65,70,74]. More recently, BPMN has been used to provide a set of recommendations for clinical or care pathway optimization [46,87,91,92,94,95,97,98], sometimes in support of value-based health care [98]. In addition, BPMN modeling

XSL•FO
RenderX

can then be used to simulate different scenarios for improvement [57,61,91,95] and provide proof of concept for possible optimization results [59]. However, using BPMN does not necessarily enable the initial objectives to be achieved and can lead to a variety of benefits, limitations in its use becoming opportunities for optimization or even giving rise to alternatives that could become threats.

Benefits, Limits, Opportunities, and Alternatives or Threats

A total of 26 broad themes and 59 subthemes were identified in the BPMN utility data fields. Table 5 shows the distribution of extractions across each of these broad themes for each domain of benefits, limits, opportunities, and alternatives or threats. The italic text indicates the most prominent themes, using 10% representation of column dimensions (benefits, limits, opportunities, and alternatives or threats) as an arbitrary cutoff. These are used to map the BPMN Purpose-Implementation-Outcome Model in the next section.

XSL•FO
RenderX

Table . Themes related to expressed benefits, limitations, and opportunities associated with the use of BPMN[a] in health care, as well as identified alternatives or threats to its use.

Theme	Benefits	Limits	Opportunities	Alternatives or threats
Accessibility	2 (0.9)	0.0	0.0	0.0
Automation and conditionality	2 (0.9)	2 (2.0)	8 (4.8)	*8 (14.8)* [b]
Clinical utility	19 (8.6)	9 (9.2)	*20 (11.9)*	0.0
Collaboration	11 (5.0)	1 (1.0)	5 (3.0)	0.0
Comprehensiveness	18 (8.1)	*18 (18.4)*	7 (4.2)	2 (3.7)
Customization	0.0	2 (2.0)	0.0	0.0
Data and measurement capability	10 (4.5)	*13 (13.3)*	*21 (12.5)*	2 (3.7)
Decision-making	2 (0.9)	1 (1.0)	14 (8.3)	*9 (16.7)*
Efficiency	1 (0.5)	5 (5.1)	1 (0.6)	1 (1.9)
Extensibility	4 (1.8)	2 (2.0)	4 (2.4)	0.0
Flexibility	9 (4.1)	0.0	2 (1.2)	1 (1.9)
Health care suitability	6 (2.7)	9 (9.2)	4 (2.4)	1 (1.9)
Information management	0.0	2 (2.0)	7 (4.2)	5 (9.3)
Integration	2 (0.9)	1 (1.0)	4 (2.4)	0.0
Interoperability	2 (0.9)	1 (1.0)	3 (1.8)	1 (1.9)
Language utility	*49 (22.2)*	7 (7.1)	6 (3.6)	5 (9.3)
Machine interpretability	7 (3.2)	0.0	1 (0.6)	0.0
Management utility	*27 (12.2)*	*10 (10.2)*	11 (6.5)	0.0
Optimization	1 (0.5)	0.0	4 (2.4)	0.0
Process utility	5 (2.3)	3 (3.1)	6 (3.6)	3 (5.6)
Scalability	0.0	0.0	0.0	3 (5.6)
Simulation	4 (1.8)	2 (2.0)	5 (3.0)	*6 (11.1)*
Supportive technology	0.0	2 (2.0)	6 (3.6)	0.0
Tools	1 (0.5)	0.0	*19 (11.3)*	3 (5.6)
User experience	19 (8.6)	8 (8.2)	3 (1.8)	2 (3.7)
Visualization	20 (9.0)	0.0	1 (0.6)	2 (3.7)
Total articles	61	61	61	61
Articles without relevant extractions	7	13	10	42
Total extractions (denominator)	221	98	168	54
Mean extractions per paper	3.6	1.6	2.8	0.9
Data =extraction counts (column %)				

[a]BPMN: Business Process Model and Notation.

[b]The italicized text indicates the most prominent themes, using 10% representation of column dimensions (benefits, limits, opportunities, alternatives or threats) as an arbitrary cut-off. These are used to map the BPMN Purpose-Implementation-Outcome Model in the next section.

These frequencies were then used to plot opportunity-threat against benefit-limit to illustrate where benefits or limits might relate to opportunities or threats across all themes in a relational frequencies of themes plotted by benefit-limit against opp-threat figure (Figure 3). For this analysis, benefits (positive) and limitations (negative) were considered as representing positions along the same continuum (*x*-axis) as were opportunities (positive) and alternatives or threats (negative; *y*-axis). Data represent column percentages presented in Table 3 using negative transformed values for both limitations and alternatives or threats. Positions right of the vertical represent benefits, left of the vertical represent limitations. Positions above the

XSL•FO
RenderX

horizontal represent opportunities and below the horizontal represent alternatives or threats.

To unpack the relationships between benefit-limit and opportunity-threat at the broad theme level, that is, to better understand, for example, where limits represent opportunity and where they represent threat for BPMN, we analyzed the distribution of subthemes within the dominant broad themes (Multimedia Appendix 4).

Figure 3. Relational frequencies of themes plotted by benefit-limitation against opportunity-threat.

A BPMN Purpose-Implementation-Outcome Model

Overview

This narrative synthesis is structured as a model that seeks to describe how BPMN use case (purpose or objective) links to outcomes through characteristics of BPMN implementation expressed as benefits, limits, opportunities, and threats or alternatives.

First, to revisit the primary objectives for using BPMN to model patient health care trajectories, we saw that formalization of care processes by modeling and visualizing the processes within patient health care trajectories was the dominant objective, followed by analyzing existing processes, visualizing constraints, and simulating changes to optimize operational performance. These objectives sought to create improved clinical outcomes at both the level of the individual patient and for the organization's performance. Indeed, some projects have demonstrated impacts on the organization of health care services for improved quality and safety. These outcomes translated into an improvement in patient outcomes [41,44,47,61,64,76,78,81,82,84], such as treatment success [41,61,81,82], and organizational performance outcomes such as service quality [61,82], expressed in the literature as reduced delays in diagnosis and treatment [44], reduction in medical errors [47,78], reduced unnecessary appointments and duplicated processes [84], reduced redundancies in clinician workflows [47,64,76], and improving patient autonomy in decision-making [81].

We now move to unpack the implementation mechanisms by which BPMN, used for the above-stated objectives, may facilitate or inhibit the realization of these anticipated and demonstrated outcomes. To do so, we adopt a logic model to map out these mechanisms within the dominant broad themes outlined in Table 5 and Figure 4. The thematic analyses were synthesized in accordance with the CFIR domains [99-101].

XSL•FO
RenderX

Figure 4. A Business Process Model and Notation purpose-implementation-outcome model. BPMN: Business Process Model and Notation; BPMN+V: a data-enriched subset of BPMN1 suitable for modeling clinical guidelines; BPMN4CP: BPMN for clinical pathways; CMMN: Case Management Model Notation; CP: clinical pathway; CPG: clinical practice guideline; CFIR: Consolidated Framework for Implementation Research; DMN: Decision Modeling Notation; FHIR: Fast Healthcare Interoperability Resource; QoL: Quality of Life; t.BPMN: Tangible Business Process Modeling; UML: Unified Modeling Language.

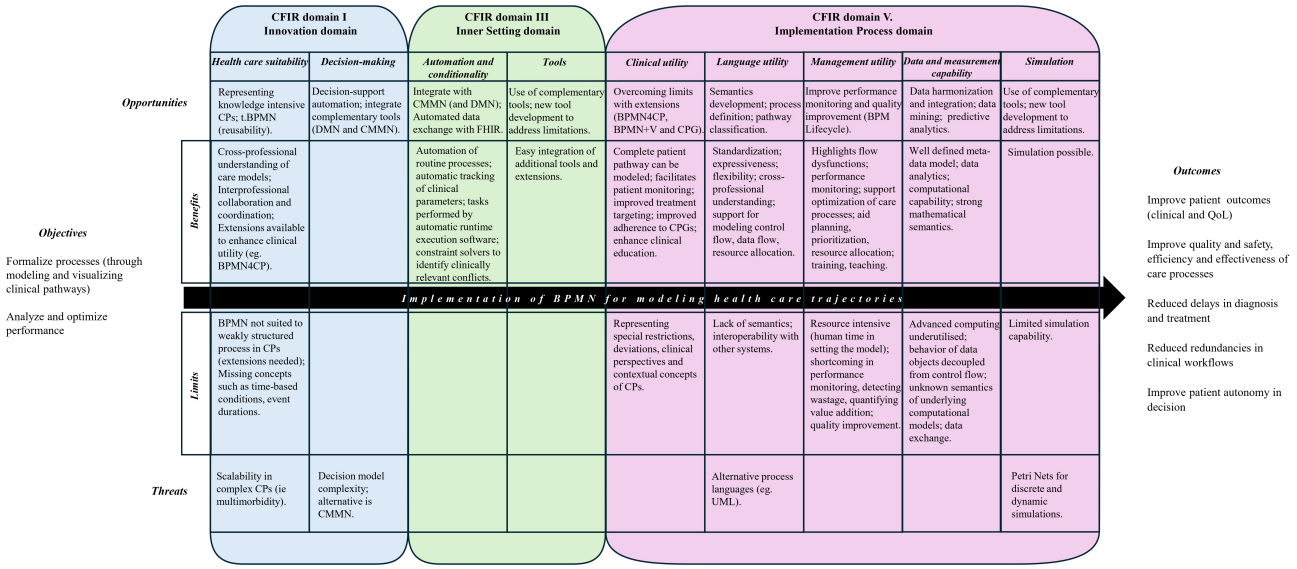

CFIR Domain I: Innovation Domain

The CFIR domain I corresponds to the key characteristics of the innovation implemented regardless of the implementation process or adaptation carried out [100].

Health Care Suitability

The literature highlights both benefits and limitations of BPMN in health care but presents minimal discussion on opportunities and threats to its suitability. BPMN facilitates understanding of care models among professionals [42,44,55,65,76,83-85,87,90-92], formalizes organizational knowledge [40], and complex health care processes [40,59], and enables in-depth process analysis and cross-professional understanding of care models by clinicians [73,90-92], IT staff [40,45,50,54,60,64,65,71,74,75,78,87], and process experts [53,54,60,65,74].

It enhances interprofessional collaboration between care professionals [49,78,90,98] and with support services such as IT [59] by involving various stakeholders in coordination. Indeed, involving the various stakeholders in care process modeling improves coordination in care processes by describing collaboration modes [40,49] and communication processes [41,44,98].

While BPMN supports complex health care trajectories [64] and can be extended for enhanced functionality [71], its core limitations necessitate extensions [54,61], particularly for time-based conditions and event durations [84,87]. Aissaoui and colleagues [84] highlight its ability to model diverse patient pathways, whereas Wiemuth and colleagues [66] note its limitations in capturing weakly structured processes. Scalability remains a challenge, especially in modeling the highly complex care pathways navigated by patients with multimorbidity [59]. However, opportunities exist for developments in representing knowledge-intensive clinical pathways [61]. For instance, modular construction systems such as t.BPMN (tangible Business Process Modeling) offer potential for enhanced interprofessional analysis [42], and model reusability or replication, which have been noted as limitations [63,69,82,90].

Decision-Making

Without extensions, BPMN cannot model decisions and provide guidance for their application and rules [64,68,76]. However, BPMN offers opportunities in automated decision support [52,96] and integration of DMN and CMMN for decision execution [60,66,68,86] in different health care contexts, including surgery [71]. CMMN is particularly suited for unstructured processes requiring real-time adaptability [58,98]. However, conclusive results are still lacking.

BPMN's limitations in modeling complex decision scenarios [61] create demand for alternatives such as CMMN, which enables dynamic, condition-based task activation [66].

CFIR Domain III: Inner Setting Domain

The CFIR domain III corresponds to the structural characteristics of the setting, including the existing IT infrastructure [100].

Automation and Conditionality

BPMN facilitates automation, task tracking, and interoperability via continuous communication based on Health Level Seven (HL7) integration [41,72,98], a set of international standards designed to facilitate the exchange, integration, sharing, and retrieval of electronic health information between disparate medical apps. It supports capacity-based task allocation [59,78], integration of runtime execution software [64,66,74,75], and automated patient monitoring [81]. While BPMN struggles with repeatable task modeling [66], integration with CMMN and DMN enhances automated and conditional advanced decision-making and task execution [66].

Further opportunities include automatic data exchange with hospital information systems and electronic health records (EHRs) via Fast Healthcare Interoperability Resources (FHIR)

XSL•FO
RenderX

[81,92,96], a modern, web-based standard from HL7 International for electronically exchanging health care data.

Tools

BPMN integrates well with complementary tools such as DMN [61,66,68,87] and CMMN [79] to enhance decision modeling, and FHIR [89,92,96] to interact with clinical data systems and EHRs. Additional tools facilitate checklist implementation [64] and surgical decision processes [71]. Calls for artificial intelligence (AI)–driven process apps [55,97] and advanced software for real-time execution [55,89] present further opportunities for BPMN tool development in health care.

CFIR Domain V: Implementation Process Domain

The CFIR domain V corresponds to the activities and strategies used to implement the innovation and their impacts on practices [100].

Clinical Utility

BPMN contributes to clinical effectiveness [81], clinical guideline adherence [52,82], patient monitoring [81,96], clinical education [79,81], clinical safety [78], and quality care [82]. It aids in translating CPGs into logical models [87,88,90,91], integrating recommendations at every step of the care trajectory [48], and supporting clinical decision-making [47,52,60,68-70,86]. BPMN allows clear process flows with treatment steps, often condition-based [84], to be described precisely [87], though integration of an extension capable of capturing the sophistication of CPGs [88] may be superior.

Without extensions, BPMN cannot (1) model decisions, (2) provide guidance for their application and rules [64,68,76], (3) account for variations in procedures or trajectory deviations to include multiple perspectives [46,59,60,66,81], (4) integrate specific knowledge such as guidelines or contextual knowledge such as specific roles for resources, activities [52,56,71,81], or responsibilities [81]. In addition, it is difficult to model roles for shared activities [58], value addition [84], or specific delays and time constraints [54,61,66,84,87]. DMN and CMMN together address some of these challenges, enabling decision representation and structured workflow execution [64,68,76]. Advances such as BPMN extension for clinical pathways [46,59] and tools such as BPMN+V (a data-enriched subset of BPMN1 suitable for modeling clinical guidelines) improve precision in process definition and patient-specific pathway navigation [88].

Language Utility

As an industry-standard process modeling language, BPMN integrates tasks, events, and gateways [71], but lacks formal semantics, complicating system interoperability [75,81,89]. This limitation impacts data interaction with the control flow [75,88] and the content of rules [64]. Unified Modeling Language is frequently cited as an alternative [53,72,77], but BPMN has the potential to be enhanced through semantics development [39,48], specifically to develop operational semantics based on partially ordered events to allow the integrated execution of multiple process models at the same time to recommend treatment steps [48].

Management Utility

BPMN supports health care management [47,50,54,73,78,82,83,85,91-98] by improving process understanding [47,54,83,85,91-95,97] and performance monitoring [57,67,78,80,81,94,96,97]. When combined with DMN, it can also improve compliance with clinical guidelines and best practices [52,64,68,76,82], and therefore improve quality of care as measured by treatment outcomes, cost reduction, resource planning, exception management, and coordination between organizational units [82]. Tomaskova and colleagues [73] emphasized BPMN's benefit to the management functions of public administrations through its ability to distinguish social from health care processes and outcomes, including by disease stage.

BPMN highlights inefficiencies, including inadequate delays [43,44,61,91-95,97], process criticalities [57], redundant activities [44,48,81], and documentation [44,48]. BPMN-based dashboards track service quality [45,55,74,80], patient waiting lists [45], and resource use [45,50,61], such as operating room efficiency [45]. However, BPMN lacks native support for process measurement and value quantification [84]. Extensions such as Conformance Checking [72] improve quality assessment, audit processes, and predictive analysis, while BPM lifecycle enhancements refine process model evaluation [62].

Data and Measurement Capability

BPMN faces challenges in computational models [39], data exchange [82], and advanced computing [78,81]. Underuse of technologies such as AI and predictive analytics restricts their measurement potential [81]. Rodrigues et al [78] attribute this, in part, to limited IT exposure among nursing professionals.

Despite this, BPMN enables integration of data objects (documents) [58], computational simulations [42,44], and structured metadata [56]. Opportunities exist in harmonizing data structures [52,76,82], AI-driven patient trajectory prediction [79], and linking BPMN with health economics evaluation [73]. Tomaskova and colleagues [73] highlight its potential in assessing new cost-effective treatments, while Bianchi and colleagues [82] advocate for a shared data model across BPMN, DMN, and other standards to facilitate this at scale. More recently, some authors have highlighted the potential of BPMN for measuring efficiency and capacity [94,96], risk [94], and tracking and comparing patient outcomes through Key Performance Indicators [96].

Simulation

An interesting aspect of the use of BPMN is its ability to allow the simulation of different scenarios. Simulation helps define the solution best suited to the situation and context [57,72,91,95], highlighting the diversity of issues and constraints [43,44,61] and clarifying potential improvement and impact of performance [42,57,73,91,95].

However, simulation appeared to be considered almost exclusively as a threat to BPMN in the early published articles [38,39], with Petri Nets being considered a major alternative for discrete and dynamic simulations [38,39] and an important candidate for designing and implementing clinical services [38].

XSL•FO
RenderX

Discussion

Principal Findings

This systematic review provides an in-depth examination of the effectiveness of BPMN in modeling patient health care trajectories. Our findings suggest that BPMN has a more pronounced contribution to a managerial approach than to clinical relevance. Indeed, the use of BPMN improves process comprehension [47,53,83,86,91-98] while also creating opportunities for optimizing patient outcomes [41,61,76,81,92,94,96] and organizational performance outcomes [43,47,61,64,76,78,82,84,91-98]. However, despite these advantages, several challenges persist, including issues related to scalability [59], integration with advanced decision-making frameworks, and the complexity of modeling dynamic health care environments [61,66] or less structured processes [61,66,98].

Benefits of BPMN in Health are

One of the principal benefits of BPMN in health care is its ability to facilitate workflow visualization [41,42,44,46-49,51,54,62,64,66-70,77,80,83,85,92,93,98]. This capability allows health care professionals to gain a comprehensive understanding of patient health care trajectories [42,44,55,65,76,83-85,87,90-93], making it easier to identify inefficiencies [43,44,48,57,61,81,91-95,97] and areas for improvement [53,60,65,70,74,91-98]. The definition of processes ensures that workflows remain consistent across different settings [84,87], thereby reducing variation and improving the quality of care [52,78,79,81,82]. Several studies have demonstrated BPMN's effectiveness in process optimization, highlighting its ability to streamline workflows [43], reduce redundancies [47,64,76], and enhance operational efficiency [45,50,61,94]. Moreover, when integrated with decision-support tools such as DMN, BPMN enhances clinical decision-making by enabling better adherence to guidelines and best practices [52,64,68,76,81,91]. Another important advantage is BPMN's interoperability with existing health care IT standards, such as HL7 [41,72,92] and FHIR [81,89,92,96], which facilitates seamless integration with EHRs and other digital health systems [64,66,74,75,96]. However, it is interesting to note the lack of consideration given to the "social dimension" of transforming practices associated with any transformation project, such as those using BPMN notation, despite this dimension being widely recognized in implementation frameworks such as the CFIR that we have used.

Challenges

Despite its many advantages, BPMN also has notable limitations that impact its applicability in health care settings. One of the most significant challenges is its complexity, particularly when modeling highly dynamic, multistakeholder environments, such as long-term chronic disease management [61,66]. The ability of BPMN to represent decision-making processes is limited [64,68,76], requiring extensions such as DMN and CMMN to capture complex, evolving patient pathways [60,66,68,87]. Another critical challenge is BPMN's limited capability to integrate real-time data analytics and AI for predictive modeling, an area of growing importance in modern health care [58,78,81,97]. Additionally, BPMN requires a specialized skill set [78], meaning that health care professionals need dedicated training, a potential barrier to widespread adoption.

Opportunities for Future Development

Despite these limitations, there are several opportunities for enhancing BPMN's applicability in health care. One key area is the creation of domain-specific BPMN extensions [27], which can improve its ability to represent complex clinical pathways. Another promising avenue is the integration of AI and machine learning into BPMN models, allowing for more sophisticated predictive analytics and decision automation [78,81,97]. Enhancing interoperability by developing standardized data exchange formats between BPMN and health care IT systems could further improve system efficiency and adoption [98]. Additionally, BPMN could benefit from the augmentation of dynamic simulation tools [42-44,57,61,72], which would allow for a more accurate representation of complex patient trajectories. This is why, from a managerial perspective, the use and implementation of the BPMN tool have definite potential, but its implementation and results remain uncertain. Indeed, the implementation of this technical tool suffers from a lack of support for a social approach to transforming practices and understanding the complexity of such projects, which Madan [97] describes as the perceived lack of consideration for the human dimension in the transformation of processes. That is why we recommend considering the social dimension necessary for any organizational change by integrating the "individual domain," the IV domain that exists in the CFIR [100] and which is absent from the studies analyzed. From this perspective, a recent framework for implementing care pathways proposes integrating BPMN as a reflective tool to support the optimization and transformation of organizational and clinical practices, rather than as a tool for standardizing clinical processes [14].

Alternatives and Competing Approaches

While BPMN is a widely adopted modeling approach, alternative methodologies offer complementary or competing advantages. Petri Nets, for instance, provide a powerful framework for discrete-event simulation and dynamic process modeling, which may be more suitable for certain applications [38,39]. Unified Modeling Language offers a robust structural representation [53,72,77,98] but lacks BPMN's process-focused approach [39,48]. Similarly, Business Process Execution Language is more suited for automated workflows but does not provide BPMN's visual representation capabilities [46]. A hybrid approach, integrating BPMN with AI-driven decision support systems and big data analytics, could help bridge some of these gaps and enhance BPMN's clinical applicability.

Managerial Implications

From a managerial perspective, BPMN offers significant potential in health care service planning [50,73,78,82,91-98], process optimization [47,53,83,85,91-98], and resource allocation [57,67,78,80,81,94,97]. By providing a standardized tool for modeling health care workflows, BPMN enables administrators and decision-makers to identify inefficiencies [43,44,48,57,61,81,91-95,97] and implement data-driven

XSL•FO
RenderX

improvements [42,44,52,56,58,73,76,79,82,94,96]. However, managers must also carefully weigh the trade-offs between BPMN's benefits and its limitations, particularly in highly complex health care settings, and ensure that the project is properly prepared, as implementing BPMN requires a significant investment in human resources and skills. The BPMN purpose-implementation-outcome model proposed in this review offers a structured framework to assess and refine BPMN transformation projects, ensuring that they align with clinical and organizational objectives.

Future Research Directions

Several areas for future research could enhance our understanding of BPMN's role in health care. Comparative studies directly evaluating BPMN against alternative modeling approaches could provide deeper insights into its strengths and weaknesses. More extensive, real-world implementation trials are needed to assess BPMN's impact on patient outcomes and operational efficiency at scale. Additionally, further research into the integration of AI and process mining techniques could improve BPMN's capabilities in predictive analytics and decision automation. Finally, investigating user-centric design approaches could help make BPMN more accessible and intuitive for health care professionals, facilitating broader adoption.

Limitations

While this systematic review provides valuable insights, limitations should be acknowledged. The majority of the studies included are methodologically weak (75.4%), consisting of exploratory or pilot descriptive studies. Given the methodological weaknesses of existing studies and the limited number of large-scale BPMN implementation projects, the validity of the results presented remains questionable. Indeed, most studies focused on prototypes [42,46,49,51,70,73,81,88], small-scale descriptive projects [43,45,50,53,57,62,63,77,80,82-85,91-98], or theoretical explorations [41,44,55-57,59,66,68,69], underscoring the need for further real-world research. Furthermore, the concentration of studies included in European countries (68.9%) is a key factor limiting the generalizability of our results.

Conclusions

This systematic review highlights BPMN's potential as a valuable tool for modeling patient health care trajectories in a managerial approach to transforming practices. Its ability to visualize and optimize processes makes it a promising tool for improving clinical and operational efficiency. However, trade-offs between benefits and limits of BPMN characterize its implementation in patient health care trajectories, giving rise to opportunities for the development and integration of new tools and extensions to handle complexity and real-time data integration and to optimize outcomes. However, it is important to note that the methodological weaknesses of the studies and the lack of large-scale research projects mean that these results cannot be generalized. Future advancements, including the development of more sophisticated BPMN extensions, integration with AI, and improved interoperability with health care IT systems, will be crucial in realizing BPMN's full potential. We propose a framework linking purpose to outcomes through richly characterized implementation domains, which could help managers to better specify their BPMN transformation projects and facilitate evaluation of their effectiveness. Future research could address the compatibility of systems in the hospital environment and emphasize the importance of considering the social dimension inherent in any change in professional and organizational practices.

Acknowledgments

We would like to acknowledge Kassim Said Abasse for his participation in writing the protocol for the review and his participation in the early stages of data extraction.

No generative artificial intelligence (AI) was used at any stage of this systematic literature review.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

Conceptualization: JBG, PL, MTH, CP, FB, AC
Data curation: JBG, AC
Formal analysis: JBG, PL, MTH, CL, ET
Investigation: JBG, PL, MTH, CL, ET
Methodology: JBG, PL, MTH, CP, FB, AC
Project administration: JBG, AC
Resources: AC
Supervision: PL, CP, FB, AC
Validation: PL, MTH, CL, ET, CP, FB, AC
Writing – original draft: JBG, PL, MTH
Writing – review & editing: JBG, PL, MTH, CL, ET, CP, FB, AC

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Databases search strategy.
[DOCX File, 25 KB - jmir_v28i1e78506_app1.docx]

Multimedia Appendix 2
Extraction table (Business Process Modeling Notation).
[XLSX File, 345 KB - jmir_v28i1e78506_app2.xlsx]

Multimedia Appendix 3
Completed Mixed Method Appraisal Tool checklist.
[DOCX File, 48 KB - jmir_v28i1e78506_app3.docx]

Multimedia Appendix 4
Distribution of subthemes.
[DOCX File, 31 KB - jmir_v28i1e78506_app4.docx]

Checklist 1
Completed PRISMA 2020 checklist.
[DOCX File, 23 KB - jmir_v28i1e78506_app5.docx]

References

1. McMahon M, Nadigel J, Thompson E, Glazier RH. Informing Canada's health system response to COVID-19: priorities for health services and policy research. Healthc Policy 2020 Aug;16(1):112-124. [doi: 10.12927/hcpol.2020.26249] [Medline: 32813643]
2. Caring for Quality in Health: Lessons Learnt from 15 Reviews of Health Care Quality, OECD Reviews of Health Care Quality: OECD Publishing; 2017. [doi: 10.1787/9789264267787-en]
3. Organization for Economic Co-Operation and Development. Health Reform: Meeting the Challenge of Ageing and Multiple Morbidities: OECD Publishing; 2017. [doi: 10.1787/9789264122314-en]
4. Institute of Medicine of America (US) Committee on Quality of Health Care in America. Crossing the Quality Chasm: A New Health System for the 21st Century: National Academy Press; 2001. [doi: 10.1136/bmj.323.7322.1192]
5. WHO report on cancer: setting priorities, investing wisely and providing care for all. : World Health Organization; 2020 URL: https://apps.who.int/iris/handle/10665/330745 [accessed 2026-05-25]
6. National Academies of Sciences, Engineering, Medicine; Committee on Improving the Quality of Health Care Globally. Crossing the global quality chasm: improving healthcare worldwide. : National Academies Press (US); 2018 URL: https://www.nationalacademies.org/publications/25152 [accessed 2026-05-25] [doi: 10.17226/25152]
7. World Health Assembly. Framework on integrated, people-centred health services: report by the secretariat. World Health Organization. 2016. URL: https://iris.who.int/handle/10665/252698 [accessed 2025-01-01]
8. Gauld R, Burgers J, Dobrow M, et al. Healthcare system performance improvement: a comparison of key policies in seven high-income countries. J Health Organ Manag 2014;28(1):2-20. [doi: 10.1108/JHOM-03-2013-0057] [Medline: 24783663]
9. Lee SE, Scott LD, Dahinten VS, Vincent C, Lopez KD, Park CG. Safety culture, patient safety, and quality of care outcomes: a literature review. West J Nurs Res 2019 Feb;41(2):279-304. [doi: 10.1177/0193945917747416] [Medline: 29243563]
10. Kruk ME, Gage AD, Arsenault C, et al. High-quality health systems in the Sustainable Development Goals era: time for a revolution. Lancet Glob Health 2018 Nov;6(11):e1196-e1252. [doi: 10.1016/S2214-109X(18)30386-3] [Medline: 30196093]
11. Gartner JB, Abasse KS, Bergeron F, Landa P, Lemaire C, Côté A. Definition and conceptualization of the patient-centered care pathway, a proposed integrative framework for consensus: a Concept analysis and systematic review. BMC Health Serv Res 2022 Apr 26;22(1):558. [doi: 10.1186/s12913-022-07960-0] [Medline: 35473632]
12. Vanhaecht K, Panella M, van Zelm R, Sermeus W. An overview on the history and concept of care pathways as complex interventions. Int J Care Pathw 2010 Sep;14(3):117-123. [doi: 10.1258/jicp.2010.010019]
13. Seys D, Panella M, VanZelm R, et al. Care pathways are complex interventions in complex systems: new European Pathway Association framework. Int J Care Coord 2019 Mar;22(1):5-9. [doi: 10.1177/2053434519839195]
14. Gartner JB, Lemaire C, Côté A. Learning care pathways framework: a new method to implement, learn, replicate, and scale up care pathways for and with the patient. Int J Health Policy Manag 2025;14:8517. [doi: 10.34172/ijhpm.8517] [Medline: 40767178]

XSL•FO
RenderX

15. Bergin RJ, Whitfield K, White V, et al. Optimal care pathways: a national policy to improve quality of cancer care and address inequalities in cancer outcomes. J Cancer Policy 2020 Sep;25:100245. [doi: 10.1016/j.jcpo.2020.100245]
16. Lodewijckx C, Sermeus W, Panella M, et al. Impact of care pathways for in-hospital management of COPD exacerbation: a systematic review. Int J Nurs Stud 2011 Nov;48(11):1445-1456. [doi: 10.1016/j.ijnurstu.2011.06.006] [Medline: 21798538]
17. Allen D, Gillen E, Rixson L. Systematic review of the effectiveness of integrated care pathways: what works, for whom, in which circumstances? Int J Evid Based Healthc 2009 Jun;7(2):61-74. [doi: 10.1111/j.1744-1609.2009.00127.x] [Medline: 21631848]
18. Alkandari M, Ryan K, Hollywood A. The experiences of people living with peripheral neuropathy in Kuwait-a process map of the patient journey. Pharmacy (Basel) 2019 Aug 30;7(3):127. [doi: 10.3390/pharmacy7030127] [Medline: 31480223]
19. Ferreira AS, Oliveira GR. Business process modeling: a webibliominig perspective of architecture frameworks. Indep J Manag Prod 2019;10(3):1159-1183. [doi: 10.14807/ijmp.v10i3.886]
20. Geiger M, Harrer S, Lenhard J, Wirtz G. BPMN 2.0: the state of support and implementation. Future Gener Comput Syst 2018 Mar;80:250-262. [doi: 10.1016/j.future.2017.01.006]
21. Pufahl L, Zerbato F, Weber B, Weber I. BPMN in healthcare: challenges and best practices. Inf Syst 2022 Jul;107:102013. [doi: 10.1016/j.is.2022.102013]
22. Stroppi LJR, Chiotti O, Villarreal PD. Extending BPMN 2.0: method and tool support. business process model and notation. Presented at: Proceedings of the 3rd International Workshop; Nov 21-22, 2011. [doi: 10.1007/978-3-642-25160-3_5]
23. Zarour K, Benmerzoug D, Guermouche N, Drira K. A systematic literature review on BPMN extensions. BPMJ 2019 Nov 18;26(6):1473-1503. [doi: 10.1108/BPMJ-01-2019-0040]
24. Braun R, Esswein W. Classification of domain-specific BPMN extensions. the practice of enterprise modeling. Presented at: Proceedings of the 7th IFIP WG 81 Working Conference; Nov 12-13, 2014. [doi: 10.1007/978-3-662-45501-2_4]
25. Loya SR, Kawamoto K, Chatwin C, Huser V. Service oriented architecture for clinical decision support: a systematic review and future directions. J Med Syst 2014 Dec;38(12):1-22. [doi: 10.1007/s10916-014-0140-z] [Medline: 25325996]
26. Mincarone P, Leo CG, Trujillo-Martín MDM, et al. Standardized languages and notations for graphical modelling of patient care processes: a systematic review. Int J Qual Health Care 2018 Apr 1;30(3):169-177. [doi: 10.1093/intqhc/mzx197] [Medline: 29346638]
27. De Ramón Fernández A, Ruiz Fernández D, Sabuco García Y. Business Process Management for optimizing clinical processes: a systematic literature review. Health Informatics J 2020 Jun;26(2):1305-1320. [doi: 10.1177/1460458219877092] [Medline: 31581880]
28. Pollock A, Berge E. How to do a systematic review. Int J Stroke 2018 Feb;13(2):138-156. [doi: 10.1177/1747493017743796] [Medline: 29148960]
29. Kassim SA, Gartner JB, Labbé L, et al. Benefits and limitations of business process model notation in modelling patient healthcare trajectory: a scoping review protocol. BMJ Open 2022 May 30;12(5):e060357. [doi: 10.1136/bmjopen-2021-060357] [Medline: 35636803]
30. Cumpston M, Li T, Page MJ, et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane Handbook for Systematic Reviews of Interventions. Cochrane Database Syst Rev 2019 Oct 3;10(10):ED000142. [doi: 10.1002/14651858.ED000142] [Medline: 31643080]
31. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
32. Babineau J. Product review: Covidence (Systematic Review Software). J Can Health Libr Assoc 2014;35(2):68. [doi: 10.5596/c14-016]
33. Hong QN, Fàbregues S, Bartlett G, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018;34(4):285-291. [doi: 10.3233/EFI-180221]
34. Glenny AM. No “gold standard” critical appraisal tool for allied health research. Evid Based Dent 2005;6(4):100-101. [doi: 10.1038/sj.ebd.6400351] [Medline: 16355243]
35. Herbison P, Hay-Smith J, Gillespie WJ. Adjustment of meta-analyses on the basis of quality scores should be abandoned. J Clin Epidemiol 2006 Dec;59(12):1249-1256. [doi: 10.1016/j.jclinepi.2006.03.008] [Medline: 17098567]
36. Higgins JP, Green S. Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series: The Cochrane Collaboration; 2008. [doi: 10.1002/9780470712184]
37. Viswanathan M, Patnode CD, Berkman ND, et al. Recommendations for assessing the risk of bias in systematic reviews of health-care interventions. J Clin Epidemiol 2018 May;97:26-34. [doi: 10.1016/j.jclinepi.2017.12.004] [Medline: 29248724]
38. Fox J, Black E, Chronakis I, Dunlop R, Patkar V, South M, et al. From guidelines to careflows: modelling and supporting complex clinical processes. In: Computer-Based Medical Guidelines and Protocols: A Primer and Current Trends: IOS Press; 2008. [doi: 10.3233/978-1-58603-873-1-44]
39. Eklund P, Johansson M, Karlsson J, Åström R. BPMN and its semantics for information management in emergency care. Presented at: Proceedings of the Fourth International Conference on Computer Sciences and Convergence Information Technology; Nov 24-26, 2009. [doi: 10.1109/ICCIT.2009.27]

XSL•FO
RenderX

40. Svagård I, Farshchian BA. Using business process modelling to model integrated care processes: experiences from a European project. Presented at: Proceedings of the International Work-Conference on Artificial Neural Networks (IWANN) 2009; Jun 10-12, 2009. [doi: 10.1007/978-3-642-02481-8_140]
41. Crisan-Vida M, Marcovici A, Stoicu-Tivadar L. ICT solution for assisted diagnosis based on monitoring in cardiology departments. Presented at: Proceedings of the 2013 E-Health and Bioengineering Conference (EHB); Nov 21-23, 2023. [doi: 10.1109/EHB.2013.6707337]
42. Scheuerlein H, Rauchfuss F, Dittmar Y, et al. New methods for clinical pathways-Business Process Modeling Notation (BPMN) and Tangible Business Process Modeling (t.BPM). Langenbecks Arch Surg 2012 Jun;397(5):755-761. [doi: 10.1007/s00423-012-0914-z] [Medline: 22362053]
43. Vandborg MP, Edwards K, Kragstrup J, Vedsted P, Hansen DG, Mogensen O. A new method for analyzing diagnostic delay in gynecological cancer. Int J Gynecol Cancer 2012 Jun;22(5):712-717. [doi: 10.1097/IGC.0b013e31824c6d0e] [Medline: 22453532]
44. Ajmi I, Zgaya H, Gammoudi L, et al. Mapping patient path in the Pediatric Emergency Department: a workflow model driven approach. J Biomed Inform 2015 Apr;54:315-328. [doi: 10.1016/j.jbi.2014.12.004] [Medline: 25554685]
45. Barbagallo S, Corradi L, de Ville de Goyet J, et al. Optimization and planning of operating theatre activities: an original definition of pathways and process modeling. BMC Med Inform Decis Mak 2015 May 17;15:38. [doi: 10.1186/s12911-015-0161-7] [Medline: 25982033]
46. Braun R, Schlieter H, Burwitz M, Esswein W. BPMN4CP: design and implementation of a BPMN extension for clinical pathways. In: Esswein W, editor. Presented at: 2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Nov 2, 2014. [doi: 10.1109/BIBM.2014.6999261]
47. Cutting EM, Overby CL, Banchero M, et al. Using workflow modeling to identify areas to improve genetic test processes in the university of maryland translational pharmacogenomics project. AMIA Annu Symp Proc 2015;2015:466-474. [Medline: 26958179]
48. Hewelt M, Kunde A, Weske M. Recommendations for medical treatment processes: the PIGS approach. Presented at: Proceedings of the Business Process Management Workshops: BPM 2014 International Workshops; Sep 7-8, 2014. [doi: 10.1007/978-3-319-15895-2_2]
49. Lamine E, Tawil AR, Bastide R. Ontology-based workflow design for the coordination of homecare interventions. collaborative systems for smart networked environments. In: Pingaud H, editor. Presented at: Proceedings of the 15th IFIP WG 55 Working Conference on Virtual Enterprises, PRO-VE 2014; Oct 6-8, 2014. [doi: 10.1007/978-3-662-44745-1_67]
50. Peres Penteado A, Molina Cohrs F, Diniz Hummel A, et al. Kidney transplantation process in Brazil represented in business process modeling notation. Transplant Proc 2015 May;47(4):963-966. [doi: 10.1016/j.transproceed.2015.03.044] [Medline: 26036495]
51. Poulymenopoulou M, Papakonstantinou D, Malamateniou F, Vassilacopoulos G. Adaptive healthcare processes for personalized emergency clinical pathways. In: E-Health–For Continuity of Care: IOS Press; 2014. [doi: 10.3233/978-1-61499-432-9-423]
52. Rodriguez-Loya S, Aziz A, Chatwin C. A service oriented approach for guidelines-based clinical decision support using BPMN. In: E-Health–For Continuity of Care: IOS Press; 2014. [doi: 10.3233/978-1-61499-432-9-43]
53. Russo V, Ciampi M, Esposito M. A Business Process Model for integrated home care. Procedia Comput Sci 2015;63:300-307. [doi: 10.1016/j.procs.2015.08.347]
54. Zerbato F, Oliboni B, Combi C, Campos M, Juarez JM. BPMN-based representation and comparison of clinical pathways for catheter-related bloodstream infections. In: Juarez JM, editor. Presented at: 2015 International Conference on Healthcare Informatics (ICHI); Oct 21-23, 2015. [doi: 10.1109/ICHI.2015.49]
55. Ateetanan P, Usanavasin S, Shirahada K, Supnithi T. From service design to enterprise architecture: the alignment of service blueprint and business architecture with business process model and notation. Presented at: Serviceology for Services Proceedings of the 5th International Conference; Jul 12-14, 2017. [doi: 10.1007/978-3-319-61240-9_19]
56. Hassen MB, Keskes M, Turki M, Gargouri F. BPMN4KM: design and implementation of a BPMN extension for modeling the knowledge perspective of sensitive business processes. Procedia Comput Sci 2017;121:1119-1134. [doi: 10.1016/j.procs.2017.12.121]
57. Bisogno S, Calabrese A, Gastaldi M, Levialdi Ghiron N. Combining modelling and simulation approaches: how to measure performance of business processes. Bus Process Manag J 2016;22:56-74. [doi: 10.1108/BPMJ-02-2015-0021]
58. Braun R, Schlieter H, Burwitz M. BPMN4CP revised-extending BPMN for multi-perspective modeling of clinical pathways. In: Esswein W, editor. Presented at: Proceedings of the 2016 49th Hawaii International Conference on System Sciences (HICSS); Jan 5-8, 2016. [doi: 10.1109/HICSS.2016.407]
59. Çatal N, Amyot D, Michalowski W, et al. Supporting process execution by interdisciplinary healthcare teams: middleware design for IBM BPM. Procedia Comput Sci 2017;113:376-383. [doi: 10.1016/j.procs.2017.08.350]
60. Combi C, Oliboni B, Zardini A, Zerbato F. Seamless design of decision-intensive care pathways. Presented at: Proceedings of the 2016 IEEE International Conference on Healthcare Informatics (ICHI); Oct 4-7, 2016. [doi: 10.1109/ICHI.2016.9]

XSL•FO
RenderX

61. Combi C, Oliboni B, Zerbato F, editors. Towards dynamic duration constraints for therapy and monitoring tasks. Presented at: Artificial Intelligence in Medicine Proceedings of the 16th Conference on Artificial Intelligence in Medicine, AIME 2017; Jun 21-24, 2017. [doi: 10.1007/978-3-319-59758-4_25]
62. Haouari G, Ghannouchi SA. Quality assessment of an Emergency Care Process Model based on static and dynamic metrics. Procedia Comput Sci 2017;121:843-851. [doi: 10.1016/j.procs.2017.11.109]
63. Ilahi L, Martinho R, Ghannouchi SA, Domingos D, Rijo R. Similarity based approach for comparing home healthcare processes models in Portugal. Procedia Comput Sci 2016;100:1250-1259. [doi: 10.1016/j.procs.2016.09.161]
64. Nan S, Van Gorp P, Lu X, et al. A meta-model for computer executable dynamic clinical safety checklists. BMC Med Inform Decis Mak 2017 Dec 12;17(1):170. [doi: 10.1186/s12911-017-0551-0] [Medline: 29233155]
65. Ruiz-Fernández D, Marcos-Jorquera D, Gilart-Iglesias V, Vives-Boix V, Ramírez-Navarro J. Empowerment of patients with hypertension through BPM, IoT and remote sensing. Sensors (Basel) 2017 Oct 4;17(10):2273. [doi: 10.3390/s17102273] [Medline: 28976940]
66. Wiemuth M, Junger D, Leitritz MA, Neumann J, Neumuth T, Burgert O. Application fields for the new Object Management Group (OMG) Standards Case Management Model and Notation (CMMN) and Decision Management Notation (DMN) in the perioperative field. Int J Comput Assist Radiol Surg 2017 Aug;12(8):1439-1449. [doi: 10.1007/s11548-017-1608-3] [Medline: 28516301]
67. Alhaj M, Mallur K, Stepien B, Peyton L. Towards a model-based approach for developing and QA of online business processes. Presented at: 2017 8th International Conference on Information and Communication Systems (ICICS); Apr 4-6, 2017. [doi: 10.1109/IACS.2017.7921952]
68. Bazhenova E, Zerbato F, Oliboni B, Weske M. From BPMN process models to DMN decision models. Inf Syst 2019 Jul;83:69-88. [doi: 10.1016/j.is.2019.02.001]
69. De Bruin JS, Adlassnig KP, Leitich H, Rappelsberger A. A separating business logic from medical knowledge in digital clinical workflows using business process model and notation and arden syntax. In: Health Informatics Meets eHealth: IOS Press; 2018:17-24. [doi: 10.3233/978-1-61499-858-7-17]
70. Jimenez-Molina A, Gaete-Villegas J, Fuentes J. ProFUSO: business process and ontology-based framework to develop ubiquitous computing support systems for chronic patients' management. J Biomed Inform 2018 Jun;82:106-127. [doi: 10.1016/j.jbi.2018.04.001] [Medline: 29627462]
71. Neumann J, Franke S, Rockstroh M, Kasparick M, Neumuth T. Extending BPMN 2.0 for intraoperative workflow modeling with IEEE 11073 SDC for description and orchestration of interoperable, networked medical devices. Int J Comput Assist Radiol Surg 2019 Aug;14(8):1403-1413. [doi: 10.1007/s11548-019-01982-6] [Medline: 31055764]
72. Ramos-Merino M, Álvarez-Sabucedo LM, Santos-Gago JM, Sanz-Valero J. A BPMN based notation for the representation of workflows in hospital protocols. J Med Syst 2018 Aug 29;42(10):1-10. [doi: 10.1007/s10916-018-1034-2] [Medline: 30155797]
73. Tomaskova H, Kopecky M, Maresova P. Process cost management of Alzheimer's disease. Processes (Basel) 2019;7(9):582. [doi: 10.3390/pr7090582]
74. Vinci ALT, Barbosa F Júnior, de Pádua SID, Rijo R, Alves D. The process of outpatient care of children and adolescents in a tertiary - level hospital specializing in pediatrics: a case study focused on identifying opportunities for improvement with the aid of modeling using BPMN. Knowl Process Manag 2018 Jul;25(3):193-206 [FREE Full text] [doi: 10.1002/kpm.1571]
75. Weber P, Filho JBF, Bordbar B, Lee M, Litchfield I, Backman R. Automated conflict detection between medical care pathways. J Software Evolu Process 2018 Jul;30(7):e1898. [doi: 10.1002/smr.1898]
76. Sooter LJ, Hasley S, Lario R, Rubin KS, Hasić F. Modeling a clinical pathway for contraception. Appl Clin Inform 2019 Oct;10(5):935-943. [doi: 10.1055/s-0039-3400749] [Medline: 31860113]
77. Kopecky M, Tomaskova H. The Business Process Model and Notation used for the representation of Alzheimer's disease patients care process. Data (Basel) 2020;5(1):16. [doi: 10.3390/data5010016]
78. Rodrigues AL, Torres FBG, Santos EAP, Cubas MR. Process modeling: technological innovation to control the risk for perioperative positioning injury. Rev Bras Enferm 2021;74(suppl 6):e20200145. [doi: 10.1590/0034-7167-2020-0145] [Medline: 34495212]
79. Szelągowski M, Biernacki P, Berniak-Woźny J, Lipinski CR. Proposal of BPMN extension with a view to effective modeling of clinical pathways. BPMJ 2022 Oct 14;28(5/6):1364-1390. [doi: 10.1108/BPMJ-11-2021-0743]
80. Yang L, Zhang L. On treatment patterns for modeling medical treatment processes in clinical practice guidelines. China Commun 2021;18(11):197-209. [doi: 10.23919/JCC.2021.11.014]
81. Szelągowski M, Berniak-Woźny J, Lipiński C. BPM support for patient-centred clinical pathways in chronic diseases. Sensors (Basel) 2021 Nov 6;21(21):7383. [doi: 10.3390/s21217383] [Medline: 34770688]
82. Bianchi A, Mortari M, Pintavalle C, Pozzi G. Putting BPMN and DMN to work: a pediatric surgery case study. In: Pozzi G, editor. Presented at: 2021 IEEE International Conference on Digital Health (ICDH); Sep 5-10, 2021. [doi: 10.1109/ICDH52753.2021.00028]
83. de Lima IB, Alves D, Teixeira Vinci AL, et al. Mental health indicators in the hospitalization process in a Brazilian psychosocial care network. Procedia Comput Sci 2022;196:623-630. [doi: 10.1016/j.procs.2021.12.057]

XSL•FO
RenderX

84. Aissaoui NO, Ben Mbarek H, Layeb SB, B. Hadj-Alouane A. A BPMN-VSM based process analysis to improve the efficiency of multidisciplinary outpatient clinics. Prod Plan Control 2024 Apr 3;35(5):461-491. [doi: 10.1080/09537287.2022.2098199]
85. Beckmann CL, Lodde G, Livingstone E, Schadendorf D, Böckmann B. Guideline-based context-sensitive decision modeling for melanoma patients. In: German Medical Data Sciences 2022–Future Medicine: More Precise, More Integrative: IOS Press; 2022:50-57. [doi: 10.3233/SHTI220803]
86. Iglesias N, Juarez JM, Campos M. Handling time constraints in infection clinical pathways using openehr TP. In: MEDINFO 2021: One World, One Health–Global Partnership for Digital Innovation: IOS Press; 2022:7-11. [doi: 10.3233/SHTI220021]
87. Kirisits JM, Gall W. Process-modelling of cross-sector healthcare quality indicators. In: dHealth 2023: IOS Press; 2023:71-76. [doi: 10.3233/SHTI230015]
88. Litchfield I, Turner AM, Ferreira Filho JB, Lee M, Weber P. Automated conflict resolution for patients with multiple morbidity being treated using more than one set of single condition clinical guidance: a case study. Comput Biol Med 2022 May;144:105381. [doi: 10.1016/j.compbiomed.2022.105381] [Medline: 35279424]
89. Kober G, Robaldo L, Paschke A. Using BPMN for medical guidelines that integrate with FHIR-RDF. Presented at: Proceedings of the SWAT4HCLS; Feb 13-16, 2023. [doi: 10.24406/publica-1666]
90. Martínez-Salvador B, Marcos M, Palau P, Domínguez Mafé E. A model-driven transformation approach for the modelling of processes in clinical practice guidelines. Artif Intell Med 2023 Mar;137:102495. [doi: 10.1016/j.artmed.2023.102495] [Medline: 36868689]
91. Rasooli N, Jolai F, Sepehri MM, Tehranian A. BPM application in clinical process improvement: a women ’hospital case study. Bus Process Manag J 2024 May 31;30(3):986-1011. [doi: 10.1108/BPMJ-07-2023-0560]
92. Rocco C, Mitrano G, Corallo A, Pontrandolfo P, Guerri D. Improving care pathways through BPM and telemedicine: an Italian study. Bus Process Manag J 2024 May 31;30(3):799-842. [doi: 10.1108/BPMJ-08-2022-0378]
93. Knight J, Chandrabalan VV, Emsley HCA. Visualizing patient pathways and identifying data repositories in a UK neurosciences center: exploratory study. JMIR Med Inform 2024 Dec 24;12:e60017. [doi: 10.2196/60017] [Medline: 39718562]
94. Rosa A, Massaro A, McDermott O, Cianci V, Simone M. Business process modelling to optimise the management of surgeries of acute cholecystitis: a case study of an Italian hospital unit. Presented at: International Conference on Lean Six Sigma for Higher Education Institutions; Sep 2-3, 2024. [doi: 10.1007/978-3-031-84816-2_29]
95. Calabrese A, D’Uffizi A, Levialdi Ghiron N, Berloco L, Pourabbas E, Proudlove N. Design and development of a digital diagnostic clinical pathway: evidence from an action research study. Eur J Innov Manag 2024 Dec 16;27(9):94-126. [doi: 10.1108/EJIM-06-2023-0483]
96. Dos Santos Leandro G, Moro CMC, Cruz-Correia RJ, Portela Santos EA. FHIR Implementation Guide for stroke: a dual focus on the patient’s clinical pathway and value-based healthcare. Int J Med Inform 2024 Oct;190:105525. [doi: 10.1016/j.ijmedinf.2024.105525] [Medline: 39033722]
97. Madan V. The National Health Service 2-week-wait skin cancer referral pathway: analysis and recommendations for process improvement. Clin Exp Dermatol 2024 Dec 23;50(1):88-96. [doi: 10.1093/ced/llae323] [Medline: 39138848]
98. Odeh Y, Shamieh O. Bridging the digital readiness gap in palliative home care: a process to data approach. Digit Health 2025;11:20552076251346380. [doi: 10.1177/20552076251346380] [Medline: 40453050]
99. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implementation Sci 2009 Dec;4(1):50. [doi: 10.1186/1748-5908-4-50]
100. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated Consolidated Framework for Implementation Research based on user feedback. Implementation Sci 2022;17(1):75. [doi: 10.1186/s13012-022-01245-0]
101. Damschroder LJ, Reardon CM, Opra Widerquist MA, Lowery J. Conceptualizing outcomes for use with the Consolidated Framework for Implementation Research (CFIR): the CFIR Outcomes Addendum. Implementation Sci 2022 Dec;17(1):7. [doi: 10.1186/s13012-021-01181-5]

Abbreviations

AI: artificial intelligence
BPMN: Business Process Modeling Notation
CFIR: Consolidated Framework for Implementation Research
CMMN: Case Management Model Notation
CPG: clinical practice guideline
DMN: Decision Modeling Notation
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resource
HL7: Health Level Seven
MMAT: Mixed Methods Appraisal Tool

XSL•FO
RenderX

PICOTS: Population, Intervention, Comparator, Outcomes, Timing, Settings
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by A Coristine, TDA Cardoso; submitted 03.Jun.2025; peer-reviewed by M Beltrão, O Akinsola, Y Olanlokun; revised version received 04.May.2026; accepted 04.May.2026; published 09.Jun.2026.

Please cite as:
Gartner JB, Landa P, Haren MT, Lemaire C, Tanfani E, Paquet C, Bergeron F, Côté A
The Implementation of a Business Process Model and Notation for Modeling Patient Health Care Trajectories: Systematic Review
J Med Internet Res 2026;28:e78506
URL: https://www.jmir.org/2026/1/e78506
doi:10.2196/78506

© Jean-Baptiste Gartner, Paolo Landa, Matthew T Haren, Célia Lemaire, Elena Tanfani, Catherine Paquet, Frédéric Bergeron, André Côté. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Experiences of Alert Fatigue and Its Contributing Factors in Hospitals: Qualitative Study

Nicki Newton[1], BSc (Hons), Grad Cert, PhD; Adeola Bamgboje-Ayodele[2], PhD; Rowena Forsyth[3], PhD; Amina Tariq[4], PhD; Jing Huang[5], MIM; Ramyasri Yannam[5], MBBS; Daniel J Lalor[5], MBA; Andrew James Sobey[5], MHSM; Melissa T Baysari[1], PhD

[1]Digital Health Human Factors Group, Susan Wakil School of Nursing and Midwifery, Faculty of Medicine and Health, The University of Sydney, Susan Wakil Health Building, Camperdown, Australia

[2]Discipline of Design, School of Architecture, Design and Planning, The University of Sydney, Sydney, Australia

[3]Cyberpsychology Research Group, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia

[4]Australian Centre for Health Services Innovation and Centre for Healthcare Transformation, School of Public Health and Social Work, Faculty of Health, Queensland University of Technology, Brisbane, Australia

[5]Canberra Health Services, Act Health, Canberra, Australia

Corresponding Author:
Nicki Newton, BSc (Hons), Grad Cert, PhD
Digital Health Human Factors Group, Susan Wakil School of Nursing and Midwifery, Faculty of Medicine and Health, The University of Sydney, Susan Wakil Health Building, Camperdown, Australia

Abstract

Background: Alerts, a key feature of electronic health record systems, intend to improve patient safety by providing timely information at the point of care. However, many electronic health record systems generate excessive alerts that are not immediately clinically relevant and that contribute to alert fatigue. Despite growing recognition of alert fatigue as a safety concern, clinicians' experiences of alert fatigue and the broader system-level factors that contribute to it being experienced are not well understood.

Objective: This study aims to use a human factors approach to (1) comprehensively explore how alert fatigue is experienced by junior doctors; (2) identify factors that contribute to experiences of alert fatigue; (3) identify perceived impacts of alert fatigue on employees, organizations, and patients; and (4) identify strategies to reduce alert fatigue in practice.

Methods: Semistructured interviews were conducted with junior doctors working in hospitals across Australia. Data were thematically analyzed using a hybrid inductive and deductive approach, informed by the Systems Engineering Initiative for Patient Safety and an information processing model.

Results: A total of 20 junior doctors were interviewed. Alert fatigue was described to occur at different stages of information processing, including when alerts were not detected, were superficially processed using mental shortcuts, or required excessive cognitive effort to interpret. When alerts were not detected or thoroughly processed, participants more often perceived impacts on patient safety and care quality due to the potential to miss important information. Further, when alerts required excessive cognitive effort, participants frequently reported interruptions, frustration, and time and effort loss as impacts. Factors influencing experiences of alert fatigue were identified in all Systems Engineering Initiative for Patient Safety work system domains, including those related to people, tasks, the environment, tools and technologies, and the organization. Key contributors included the design and clinical relevance of alerts, institutional norms and expectations, and information overload from system alerts as well as other alerts and tasks. Alert fatigue was also described to be experienced differently depending on provider characteristics, such as experiences with and knowledge of alerts, mood, and personality, and organizational factors, including culture, shift type, and time of day.

Conclusions: Alert fatigue is not a binary concept but is instead experienced on a continuum and influenced by interacting individual, technical, and contextual factors. Future research should incorporate clinician self-reports to evaluate experiences of alert fatigue in addition to objective measures. Addressing alert fatigue requires tailored interventions that target its different causes and outcomes. These could include technical and design improvements, changes to organizational practices, and individual customization to reduce experiences of fatigue and accommodate differences in clinicians' needs.

(J Med Internet Res 2026;28:e78676) doi:10.2196/78676

KEYWORDS

alert fatigue; clinical decision support systems; alerts; hospitals; junior doctors; human factors; digital health

XSL•FO
RenderX

Introduction

Clinical decision support (CDS) systems have proliferated in health care, forming a core component of many electronic health record (EHR) systems now used in routine practice [1-3]. Within these systems, alerts represent the primary mode of CDS delivery. Common alerts include those that notify clinicians of potential drug-drug interactions, allergies, and dosing errors and those promoting adherence to best practice care [4,5]. Alerts can interrupt clinicians' workflows during their interactions with the EHR (interruptive alerts) or passively display CDS through icons or flags that require users to actively seek information out (passive alerts) [6]. While alerts are designed to support clinical decision-making and enhance patient safety by providing relevant information at the point of care, many systems generate excessive alerts that are not immediately clinically relevant [4,7-10].

Alert fatigue is a widely reported concern arising from excessive alerting, but as a concept, it remains poorly understood and inconsistently measured [2,11]. Existing research has largely relied on observable behaviors, such as alert overrides and response times, to infer the presence of alert fatigue, assuming that clinicians experiencing alert fatigue will dismiss alerts without considering their content [12,13]. While some studies suggest that alert fatigue arises when clinicians become cognitively overloaded by alerts [4], others report it to be driven by desensitization that develops following repeated exposure to the same alert over time [14,15].

These inconsistencies highlight the limitations of using behavioral proxies to measure what is fundamentally a *subjective* mental state [16,17]. Observable behaviors cannot fully capture clinicians' cognitive experiences and offer limited insight into the broader factors that contribute to fatigue [18]. While alert fatigue is frequently identified as a barrier to clinicians' use of alerts in qualitative studies, these studies are typically concerned with overall alert experiences rather than examining alert fatigue itself and the factors that shape it [19-21]. As a result, definitions of alert fatigue and its perceived causes and impacts remain superficial and lack a systemic perspective.

Human factors methods have been widely applied in health care to understand and address system-level factors that impact care delivery, patient safety, and performance outcomes [22]. These approaches are well suited to investigating complex, dynamic issues like alert fatigue, as they account for the cognitive and behavioral processes that occur during clinicians' interactions with technologies, within the broader context of the tasks, organizational settings, and environments in which these interactions occur [23,24]. Accordingly, this study aimed to (1) comprehensively explore how alert fatigue is experienced by junior doctors; (2) identify factors that contribute to experiences of alert fatigue; (3) identify perceived impacts of alert fatigue on employees, organizations, and patients; and (4) identify strategies to reduce alert fatigue in practice.

Methods

Study Design

This study is reported following the SRQR (Standards for Reporting Qualitative Research) guidelines (Checklist 1) [25]. A qualitative design, using semistructured individual interviews, was selected to explore the subjective and nuanced nature of alert fatigue in depth [26,27].

Theoretical Frameworks

The Systems Engineering Initiative for Patient Safety (SEIPS) is a human factors model used in health care settings to identify and address safety concerns [24]. The SEIPS describes the "work system," which comprises 5 domains, including the *person* (both patients and health care providers) at the center of care, who performs *tasks*, using various *tools and technologies* within a *work environment* under specific *organizational conditions*. Interactions between components of the work system influence the delivery of care and other *process*es, which in turn impact *patient*, *employee*, and *organizational outcomes*. The SEIPS model was used to interpret and present data.

We used an information processing model as outlined by Wickens and Carswell [28] to further interpret the *process* by which alert fatigue was experienced by participants. The model describes information processing to occur across 4 stages: *selecting information* (detecting stimuli from the environment), *perception and data interpretation* (recognizing and interpreting the stimuli), *comprehension and cognition* (processing and integrating information with prior knowledge to assess relevance and meaning), and *action selection* (choosing and executing a response based on the processed information). Thus, this model was well suited to understanding the mental processes of alert fatigue and associated actions arising from its presence. Although alert fatigue has previously been described in the context of information processing [29], an analysis of how alert fatigue manifests across different stages has not been explored. A schematic of the integrated models is presented in Figure 1.

XSL•FO
RenderX

Figure 1. Schematic of the integrated Systems Engineering Initiative for Patient Safety (SEIPS) and information processing models used to interpret data.

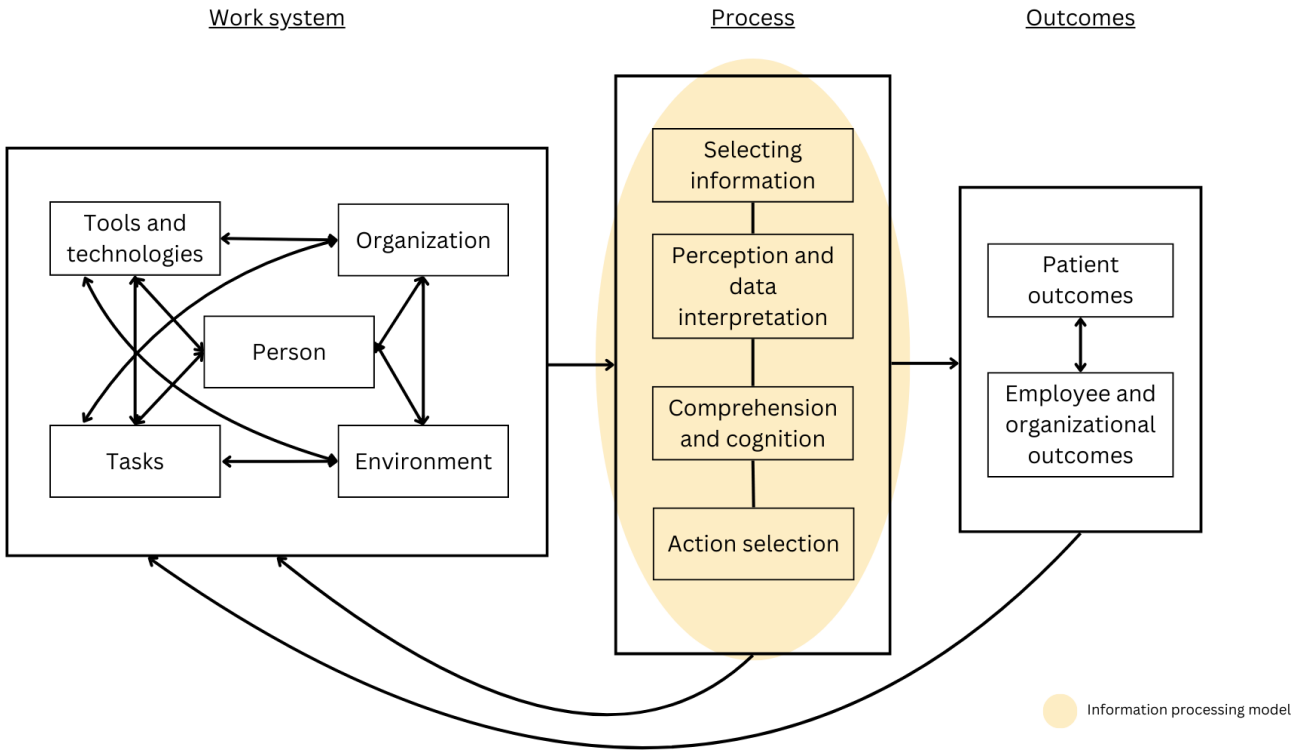

Study Team

The study team included researchers with expertise across human factors, digital health, health services, sociology, and psychology; and hospital-based pharmacy directors. Additionally, 2 members of the study team were junior doctors, who provided lived experience insights.

Participants and Recruitment

Eligible participants were practicing junior doctors, also referred to as junior medical officers (JMOs) in Australian hospital systems, defined as interns (postgraduate year [PGY] 1), residents (PGY2), and registrars (PGY3+), who had interacted with alerts in clinical information systems (eg, EHR systems) in hospital settings in Australia. Junior doctors were selected for the study as they are frequent recipients of alerts in hospital settings. A convenience sample of participants was recruited by disseminating details about the study through email distribution lists, posters, and presentations at JMO training sessions, at 2 hospitals within the same hospital network. Additional participants were recruited via snowball sampling through the authors' professional networks, including word of mouth, LinkedIn posts, and emails to hospital contacts who circulated the study to JMO networks. Prospective participants were encouraged to contact the researchers via email if they were interested in participating.

Data Collection

A semistructured interview guide was developed using insights from existing research on alert and alarm fatigue and discussion among all authors [4,14,30]. Two pilot interviews were conducted by 1 author (NN) with JMO coresearchers via Microsoft Teams, which resulted in minor refinements to the interview guide, such as adding new questions and changing the order of questions to enhance flow. Questions centered around participants' use of alerts, perceptions and experiences of alert fatigue, factors influencing alert fatigue, the impact of alert fatigue, and recommendations to reduce experiences of alert fatigue in practice (Multimedia Appendix 1). Individual interviews were conducted with participants by one author (NN) using Microsoft Teams, Zoom (Zoom Communications, Inc), or a phone. Three participants were known to the interviewer prior to the interviews through their professional and personal networks.

Interviews were audio recorded, transcribed using Otter AI software (Otter.ai, Inc) [31], and checked for accuracy. Transcription and analysis occurred in parallel to data collection, allowing for further refinement of the interview guide to explore themes in depth as interviews progressed. Interviews continued until data saturation was reached; that is, no new themes emerged in the final interviews, and only new examples or perspectives related to existing themes were identified [32].

Data Analysis

Transcripts were imported into NVivo (version 14.24.1; Lumivero) [33] and thematically analyzed using a combined inductive and deductive approach [34]. The first 6 transcripts were independently analyzed by 2 authors (NN and MB or AB or RF) who inductively coded quotes. Themes were deductively mapped to 4 categories: defining alert fatigue, factors influencing experiences of alert fatigue, the impact of alert fatigue, and strategies to reduce alert fatigue. Each pair of authors met to discuss alignment and resolve discrepancies.

Following discussions, 3 of the 4 original categories were restructured under the SEIPS model, where codes related to defining alert fatigue were moved to *processes*, factors

XSL•FO
RenderX

influencing experiences of alert fatigue were moved to the *work system*, and the impact of alert fatigue was moved to *outcomes* [24]. Themes related to the process of alert fatigue were further interpreted using the information processing model [28].

Remaining transcripts were analyzed by 1 author (NN) using the SEIPS model, adding new codes and themes as they emerged. Upon completion of the analysis of all interviews, 4 authors (NN, MB, AB, and RF) participated in a workshop to discuss the categorization of codes and themes, which resulted in some codes being renamed, merged, and/or recategorized. The coding framework was then presented to all authors and select participants for member checking, who provided input and agreed upon the final thematic structure.

Ethical Considerations

Ethical approval was received from the Australian Capital Territory Health Low Risk Ethics Committee (protocol: 2024.LRE.00024). All participants provided written informed consent and did not receive compensation for participating in the study. Interview data were deidentified at the point of transcription, with all potentially identifying details removed. Demographic characteristics are presented in aggregate, and no identifiable participant information is included in this study.

Results

Summary

A total of 20 interviews were conducted between August 2024 and March 2025. Participants included 9 interns (PGY1), 6 residents (PGY2), and 5 registrars (PGY3+), working at 10 hospital sites, across 3 Australian states (Table 1). Interviews lasted between 18 and 56 minutes, with an average length of 33 (SD 11; range 18-56) minutes. Two interviews were conducted via phone and the remainder via videoconference.

XSL•FO
RenderX

Table . Participant characteristics

Characteristics	Participants, n (%)
Role (PGY[a])	
Intern (PGY1)	9 (45)
Resident (PGY2)	6 (30)
Registrar (PGY3+)	5 (25)
Sex	
Male	12 (60)
Female	8 (40)
Length of time at current hospital	
Less than 2 months	3 (15)
2 months to 1 year	6 (30)
1 to 2 years	4 (20)
2+ years	7 (35)
Clinical experience total (including prior clinical work)	
Less than 2 months	2 (10)
2 months to 1 year	3 (15)
1 to 2 years	4 (20)
2-5 years	5 (25)
5+ years	6 (30)
State	
NSW[b]	11 (55)
ACT[c]	7 (35)
SA[d]	2 (10)
Systems used for alerts	
Commercial	18 (90)
Commercial and homegrown	2 (10)

[a]PGY: postgraduate year.
[b]NSW: New South Wales.
[c]ACT: Australian Capital Territory.
[d]SA: South Australia.

Participants received alerts from EHR, electronic medical record (EMR), and/or electronic medication management systems. All participants were users of commercially available EHR or EMR systems (eg, Epic, Cerner, Sunrise, and MedChart), while 2 participants also used homegrown systems. Alert types included passive and interruptive alerts such as those for medications (eg, allergy, drug-drug interactions, and dosing), and best practice (eg, venous thromboembolism prophylaxis, advanced care directives, and resuscitation plans). Alerts popped up during tasks such as medication ordering, opening patient charts, and opening the EHR system. Alerts outside of standard clinical information systems ("system") were also described.

In the following sections, we present the clinicians' perceptions of alert fatigue (process), contributing factors (work system), impacts (outcomes), and potential strategies for addressing alert fatigue. Figure 2 shows a schematic of themes aligned to the SEIPS model.

XSL•FO
RenderX

Figure 2. Schematic of themes related to experiences of alert fatigue within the Systems Engineering Initiative for Patient Safety (SEIPS) model.

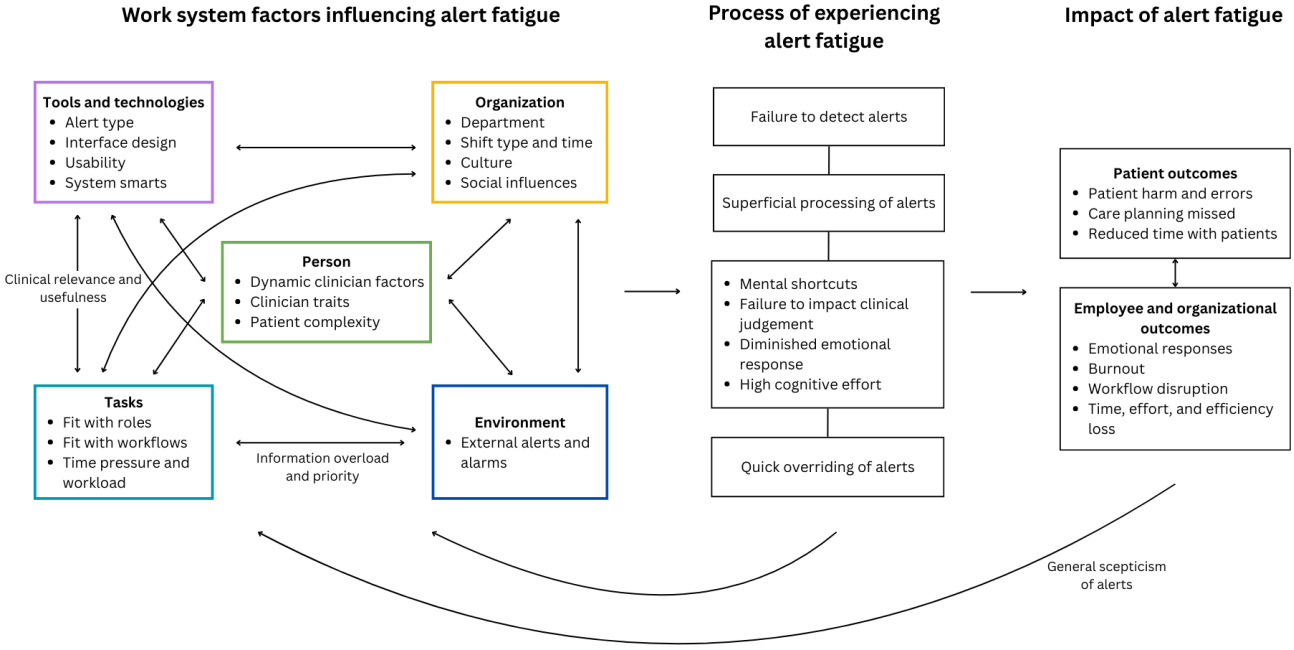

The Process of Experiencing Alert Fatigue

Selecting Information

Alert fatigue manifested during the information selection stage as a failure to notice alerts or detect alert stimuli. This occurred mainly for passive alerts, which were often described to be inconspicuous, but also for interruptive alerts: "The idea of alerts being there is just part of the job, it's just part of the furniture…I probably at least once will miss the fact that [medications are] expiring, my eyes will gloss over it" (P1).

Perception and Data Interpretation

Perception and interpretation involved recognizing and processing alerts. Clinicians described superficially processing alerts by only partially viewing or reading them before responding. Participants explained that alerts were often pre-empted based on their interactions with the system before an alert was triggered or identified by design features such as titles, keywords, fonts, colors, and layouts. Many participants felt that they could quickly and effectively recognize these alerts and often did not feel "fatigued" by alerts that were processed in this manner: "it's so mindless….All you process is, it's an ECG one, cancel, it's an ACD one, cancel" (P14).

Comprehension and Cognition

Comprehension and cognition involved assessing the relevance of alerts to inform decision-making. While participants initially described reading and actively interpreting the value of alerts, alert fatigue occurred where active interpretation of alerts and influence on clinical judgment ceased and mental shortcuts instead guided participants:

> *I would have read the alert the first time, but then…you go, "okay, I know what this alert is about," and the next day you would have forgotten which drugs are even mentioned. I just know that this is not relevant to me.* [P13]

Emotional responses to alerts, such as anxiety and stress, were reported to be common early on. However, this diminished over time as users reported becoming desensitized and "complacent." Conversely, some clinicians reported experiencing alert fatigue due to the cognitive effort and time required to interpret and appropriately respond to alerts.

Action Selection

Action selection comprised users' responses to alerts. Alerts that were quickly and habitually overridden signified alert fatigue at this stage. For example, participants described being "so used to clicking boxes" (P6) that "you just click through them automatically" (P13) and that "it's like whack-a-mole. You're just getting rid of everything as fast as you can" (P14).

Some participants said that they broke habitual responding patterns, for example, when their role changed, or they made an error due to experiencing alert fatigue. However, these patterns were likely to return: "after a while, you desensitize yourself again after that going well, it hasn't happened for a long time, and your brain kind of becomes complacent again, with going Yes, yes, yes, through all the tick boxes" (P18). Some alerts that were quickly overridden were however sometimes still perceived to be useful as they prompted consideration and did not take excessive cognitive effort to process.

Work System Factors Influencing Experiences of Alert Fatigue

Factors were coded to all work system domains of the SEIPS. Multimedia Appendix 2 shows the distribution of coded references within each domain.

Person

Patients' clinical complexity was reported to influence alert fatigue. Some participants described paying closer attention to alerts for unstable patients, while others described increased

XSL•FO
RenderX

fatigue due to the volume of alerts that complex patients triggered.

Clinician traits, including clinician type, clinical experience, and personality, also influenced alert fatigue. While many participants believed that senior clinicians experienced greater fatigue, more senior participants reported experiencing less fatigue than when junior, as they had more confidence and consequently used less cognitive effort to appropriately respond: "when I was more junior and didn't quite understand as well, I would often just go, 'advice ignored,' without really understanding the reason why" (P13). Personality traits were also raised, where participants who were more meticulous described paying more attention to alerts: "for me personally, I don't like just dismissing alerts without actually thinking through it" (P19).

Alert fatigue was influenced by dynamic factors that evolved based on individuals' experiences with alerts over time. Those who possessed a positive attitude toward alerts generally reported experiencing less fatigue:

> *I just view these alerts as another opportunity to make sure you have the whole picture.* [P10]
>
> *if you think it's a pain in the ass, then it's something that's going to irritate you every time. But if your attitude is, which mine is, is- "it's just part of the job," so you just get on with it.* [P15]

Alert fatigue was influenced by participants' knowledge and perceived importance of the risk the alert was targeting, including personal experiences of missing information or making errors: "if I don't understand or agree with the importance, I'm much more likely to dismiss it with so many other competing priorities" (P17). Additionally, some clinicians explained they did not know how to correctly respond to alerts:

> *no one told me how to…no one knows. So you just choose something that makes sense.* [P2]

Others described remaining cautious and attentive to alerts due to their awareness of alert fatigue.

Alert fatigue was also reported to be influenced by clinicians' emotional state, including their mood and whether they were experiencing other forms of fatigue: "you feel more emotionally drained, less sympathetic, less compassionate, and you might dismiss things more" (P1).

Tasks

Alert fatigue was influenced by workflow alignment, clinical relevance, fit with roles and responsibilities, time pressure, and information overload. Fatigue was often reported to occur where alerts were misaligned with organizational and decision-making processes. Poorly timed alerts, such as those triggered upon opening a patient's chart, were typically reported to be quickly dismissed: "I intend to view their chart with a very specific goal and updating the resuscitation right then is not a priority" (P16).

Clinical relevance and usefulness relative to clinicians' roles and responsibilities were often mentioned. While some alert types were seen as entirely irrelevant, others were inconsistently useful. Fatigue also arose where relevant alerts were lacking, forcing users to manually search for information: "it has taken me six clicks to get one answer, which is highly inefficient" (P16). Useful alerts were described to help with cognitive unloading, save time, and provide unfamiliar or easily forgotten information:

> *doctors would really appreciate if there was one alert saying all the blood tests that have been ordered are now back…because we spend a significant portion of our cognitive time in the day just thinking about that.* [P16]

Interestingly, some participants described experiencing less fatigue from alerts with poor utility: "we do not get effective alerts from EMR at the moment, so we don't rely on those alerts, so there's no fatigue involved there" (P16).

Alerts perceived to be well aligned with roles and responsibilities often did not induce alert fatigue even when they were often disregarded. This was particularly evident where clinicians had significant individual responsibility: "I'm always very grateful to get [discharge medications] alerts because the consequences of getting something wrong in that way can be disastrous, and it is on you legally" (P1).

In contrast, alerts received by a team of clinicians created a diffusion of responsibility, which led to these alerts being deprioritized: "if it was sent to everyone, it's everyone's job" (P17). Participants also described automatically clicking through alerts where they had lower autonomy: "someone's told you to chart that med, so you just do it" (P13).

Situations in which clinicians faced greater time pressure and higher workloads, such as during ward rounds, were described to trigger alert fatigue even in participants who were typically unaffected:

> *I don't feel like I'm at a stage where any alert that pops up- it's tapped into that alert fatigue. For me, it sort of comes in waves when I'm quite busy within the hospital and there's a lot of tasks that have built up.* [P10]

Information overload also contributed, with participants feeling overwhelmed by excess information or many alerts received at once:

> *it'll often give you up to 10 different alerts [in a single encounter]…it's hard to tell which ones are actually worth stopping for and thinking about, and which ones you just ignore.* [P4]

Other information presented on screen, such as task lists and message functions within the system, also contributed to overload.

Environment

Only one factor was coded to the work environment, which included information overload from *multisensory* nonsystem alerts or individual clinicians' digital ecosystem, including pagers, emails, WhatsApp messages, auditory alarms, phone calls, and in-person prompts. System alerts were perceived as follows:

XSL•FO
RenderX

> *yet another thing that the junior doctor is responsible for or is bombarded with…there's just so many different ways to message a JMO about a task, or to put a task somewhere for them to do, to page us, to come and find us on the ward…there seems to be a lot of over stimulation.* [P4]

The level of priority given to system alerts over other tasks was often reported to be low: "with the assumption that if it's automated, then it's not as pressing" (P17).

Tools and Technologies

Alert type, interface design, system usability, and volume influenced alert fatigue. Participants typically described alert fatigue to develop for specific types of alerts. However, similar-looking alerts could cause confusion or require extra cognitive effort to interpret: "allergies and drug-drug interactions often pop up looking the same, and it takes a little bit longer to work out which one it is" (P10). Participants described interface design features, including the use of colors, font, position, and layout, that could draw their attention; however, a participant stated:

> *once you get used to a certain font or format or typeface, the same problem will just happen. Regardless of whether it's in red or rainbow or purple, once you've had it pop up 20 times in one day…you just stop paying attention to it.* [P13]

Alerts with excess words and jargon contributed to alert fatigue as it could "blur what you're actually trying to pick up on in the writing as important" (P18), versus shorter alerts which were quickly processed: "the alert I pay most attention to is one that's about 10 words, regardless of the style, if it's short and something that your brain can process in three seconds, then it is probably going to be read" (P13).

Requiring override reasons was perceived positively for important alerts, forcing clinicians to "skim them because…you can't just say ignore. You've got to say, why you ignored, so they force you to put a little bit of accountability against it" (P14). Alert importance was sometimes interpreted based on the presence or absence of additional steps to dismiss the alert: "if it was so important, you couldn't click x, right?...I think everyone, whether subconsciously or consciously, knows that if it was the be all and end all, there wouldn't be a not now button" (P20).

Poor usability of alerts and clinical information systems, such as systems being "slow" and "clunky," was reported to contribute to alert fatigue by disrupting workflows. Clinicians also described interruptions where alerts blocked access to other areas of the system: "I need to cancel that order, stop what I'm doing, go into the allergy section and then interpret that information" (P10).

Alerts lacking system "smarts," for example, those flagging allergy and drug-drug interactions for patients who were already prescribed the drug in question, were described to contribute to fatigue. Similarly, constant repetition was perceived to create a "boy who cried wolf" (P17) effect. While an excessive volume of alerts was mentioned, it was only described as problematic where alerts were repetitive, irrelevant, or where multiple alerts were received at once.

Organization

Organizational factors included department, shift length and type, institutional culture, social influences, and team dynamics. While experiences of alert fatigue were generally consistent between departments and rotations, differences sometimes arose due to variations in the workload, time pressure, types of tasks and alerts received, and autonomy over tasks. Fatigue was described to increase on after-hours shifts, due to time pressure and the need to cover multiple wards, and at certain times of the day, including late at night. Participants also cited increased alert fatigue at the end of a shift due to workload demands and at the very beginning of a shift where they had insufficient information to appropriately respond.

Institutional and departmental culture and peer behaviors shaped role expectations, which in turn impacted alert fatigue. Alerts aligned with institutional norms were not necessarily less frustrating but were reported to be more widely accepted. Expectations of roles were described to be learned informally from colleagues:

> *You rock up and people just start asking you to do things…if you haven't been asked to do it, especially by a senior doctor on your team, then it's probably not your job.* [P17]

This was reinforced by observing colleagues dismissing alerts:

> *they're like, "Oh, this is so annoying." Click, click, click. It sets the scene of "Oh, this is just such an inconvenience." Skip, skip, skip.* [P8]

New JMOs reported asking colleagues and senior doctors for advice on how to respond to alerts. Alert fatigue was heightened due to the "unspoken pressure about the efficiency that you're required to have as a junior" (P4), where alerts were perceived to decrease efficiency.

Impact of Alert Fatigue

Employee and Organizational Outcomes

Alert fatigue was associated with feelings of frustration and annoyance, and along with other work stressors, could contribute to burnout and general fatigue. Some participants additionally described guilt; for example, "you just feel as though maybe you're not doing as good a job…you hate to think that you're not giving 100% to your patients" (P15). While alert fatigue was often tied to specific alerts, some participants developed broader skepticism, feeling "suspicious of alerts in general" (P6), creating a "sense of mistrust in the system" (P9) after encountering many alerts that were not relevant.

Some participants felt that fatigued users might miss opportunities to streamline their workflows. However, excess alerts were more often described to disrupt workflows, resulting in a loss of time, cognitive effort, and efficiency: "when I'm already holding six pieces of information in my head, then I might lose two, which is the most frustrating aspect" (P9). These interruptions were perceived to impact patient flow (eg, "if things don't move at a certain speed, the system breaks down,

XSL•FO
RenderX

or patients get stuck somewhere, or miss out on discharging home at a particular time" [P4]), lead to mistakes or unnecessary responses, and shift attention away from core issues (eg, "it leads to uncertainty, and I think often clouds the issue that the patient's in hospital for" [P9]).

Patient Outcomes

Participants felt alert fatigue could increase the risk of errors and result in actual patient harm, particularly when important alerts were missed or misinterpreted. For example, some participants observed doses being missed, patients being prescribed medications they were allergic to, and pre-emptive care planning being missed. Alert fatigue was also perceived to reduce time spent with patients and inadvertently impact care by reducing clinicians' patience.

Strategies to Reduce Alert Fatigue

Strategies to reduce alert fatigue included technical changes such as improving the interface design, personalizing and increasing the relevance of alerts; organizational strategies such as monitoring and evaluation initiatives and educating clinicians about alert fatigue; as well as personal strategies such as improving overall well-being and using workarounds. A full list of strategies is presented in Multimedia Appendix 3.

Discussion

Principal Findings

This study provides a comprehensive exploration of junior doctors' experiences of alert fatigue in hospital settings. Importantly, alert fatigue was found to manifest at different stages of information processing where participants failed to detect alerts, superficially processed alerts, developed mental models, and used excessive cognitive effort to interpret alerts, which led to different response patterns. These experiences were perceived to impact patient safety, such as increasing the risk of errors and contributing to patient harm, and clinician well-being, such as workflow interruptions and reduced efficiency.

The SEIPS model highlighted complex interactions between factors contributing to experiences of alert fatigue, including alert logic and design, organizational culture, and individual traits and emotional states. We observed junior doctors at the same hospital, who received similar alerts, to experience alert fatigue differently depending on their experience, role, and personality. For example, where junior doctors had positive attitudes and alerts were tightly aligned with responsibilities, even frequently disregarded alerts were not perceived as fatiguing. While poor clinical relevance remains a fundamentally important factor in the development of alert fatigue, our study reveals that doctors may have different levels of tolerance depending on these personal, organizational, and task factors. We also observed feedback loops between factors contributing to alert fatigue and its outcomes. For instance, repeated exposure to low-value alerts could lead to negative attitudes, subsequently reducing trust and influencing attitudes toward future alerts.

Expanding Definitions of Alert Fatigue

In our study, alert fatigue was rarely experienced in an all-or-nothing fashion. Instead, participants often described alert fatigue to arise in specific contexts, such as when under time pressure or for certain alert types. While some participants reported generalized negative perceptions of alerts, none described being universally fatigued or entirely unaffected. Similar to existing studies, we also found alert fatigue to develop for both passive and interruptive alerts [35].

Our findings support the view that alert fatigue is not simply an action, but a mental state that is subjectively experienced at conscious and subconscious levels [16,29]. Participants described 3 distinct information processing pathways in which alert fatigue occurred: one where alert stimuli were not perceived at all, another in which alerts were recognized but superficially processed based on mental shortcuts, and a third where alerts required high cognitive effort to interpret and respond to. The first 2 pathways, in line with existing conceptualizations of alert fatigue, were often associated with alerts being rapidly dismissed [4,12,13]. Despite this, mental shortcuts were perceived as an efficient way to process information and were often not associated with conscious experiences of fatigue. While many participants believed they were able to rapidly interpret alerts, patient safety risks could however still arise due to the potential to misinterpret information, particularly when the alerts lacked visual distinctiveness.

In contrast, the third pathway was consistently described as fatiguing. Participants reported feeling frustrated due to the cognitive burden of alerts and related impacts on efficiency, which arose particularly where alerts were ambiguous, time-consuming, or difficult to action. For certain alert types or individuals, this burden was described to gradually lead to cognitive disengagement, with clinicians eventually defaulting to rapid override as a strategy to avoid further mental effort. This finding mirrors alarm fatigue literature that suggests fatigue develops as a coping mechanism for cognitive burden [30,36]. While burnout is frequently linked to alert fatigue [37], our findings suggest that rapidly overriding alerts may conversely serve as a protective mechanism to manage cognitive load and avoid burnout, consistent with literature linking longer EHR use to increased fatigue and burnout [38,39].

These findings suggest that override rates and response times alone are insufficient measures of alert fatigue. Aligning with prior critiques, our work confirms that frequently overridden alerts are sometimes still attended to and can be perceived as helpful [13,40]. Similarly, although quick dismissals can indicate alert fatigue, they can also represent efficient pattern recognition, while longer interactions could both indicate appropriate engagement with alerts and signal cognitive burden. We therefore recommend supplementing system metrics with clinician self-reports to better understand both conscious and subconscious experiences of fatigue and applying solutions that appropriately address the specific type or types of fatigue experienced.

XSL•FO
RenderX

Work System Factors Contributing to Alert Fatigue

In line with existing literature, we found exposure to clinically irrelevant, repetitive, and poorly timed alerts contributes to alert fatigue [4,15,21,41,42]. Consistent with quantitative studies, the total volume of system alerts was rarely identified as a cause of alert fatigue alone, but instead became problematic when clinicians received many clinically irrelevant and repetitive alerts, and multiple alerts within a single clinical encounter [4]. We also identified novel influences, including personality traits, mood, social learning, and exposure to external alerts. Many participants described cognitive overload as a result of not just system alerts, but also notifications from pagers, emails, phone calls, and verbal prompts, which cumulatively contributed to experiences of fatigue and dismissal of system alerts. While existing literature has sought to improve total alert burden to reduce alert fatigue [43,44], our study highlights the need to consider system alerts within the broader ecosystem of alerts and tasks that clinicians receive.

Social and organizational factors are underexplored in existing alert fatigue literature but were found to collectively shape experiences in our study. Junior doctors often looked to peers to guide their own responses to alerts and expectations of their roles, affecting not just alert responses but also the framing of alerts as helpful or annoying. Organizational and cultural expectations of efficiency further increased alert fatigue, as managing alerts was often seen as a barrier to achieving other clinical tasks. While alerts are likely to offer the most benefit to junior doctors, our study substantiates prior research that suggests this group experiences more fatigue than their senior counterparts [45]. Participants with more clinical experience had greater confidence and therefore could interpret and respond to alerts with less cognitive effort, whereas those with less experience reported more ambiguity and higher effort expended in interpreting information.

Recommendations for Addressing Alert Fatigue

Our results highlighted several strategies for reducing alert fatigue that align with prior literature, including removing ineffective alerts and increasing clinical relevance [46], incorporating contextual awareness [47], applying human factors design principles [48,49], and implementing alerts in line with the 5 rights of CDS (right information, person, format, channel, and time) [6,7,50,51]. Our findings build on this literature by providing insight into how and why strategies are likely to be effective, and how they can be targeted to the different causes and impacts of alert fatigue. Several studies testing these strategies have reported their effectiveness in decreasing the overall volume and increasing the acceptance rate of alerts [50]. However, we argue that future research must also evaluate whether clinicians' *experiences* of fatigue are reduced. Given that our results showed junior doctors are likely to experience greater levels of alert fatigue than more senior doctors, we emphasize the importance of including this group in redesign efforts and evaluation of fatigue reduction.

We also identify and present new strategies that could improve experiences of alert fatigue. These include allowing clinicians to personalize alerts, education on alert fatigue, supporting overall emotional well-being, and streamlining the system and other alerts. Due to the individual variability in experiences of alert fatigue, personalization may be particularly valuable in fostering a greater sense of control and responsibility, thus improving engagement with alerts that individuals deem to be relevant. Further research is needed, however, to evaluate the effectiveness of personalization and other strategies in reducing alert fatigue and improving patient safety.

Interestingly, in some cases, experiences of fatigue were not associated with the presence of alerts, but by their absence, where critical information was difficult and time-consuming to locate. This finding underscores the need for balance between over- and underalerting, ensuring alerts effectively address the clinical or safety need. Implementers should carefully consider the issue they aim to target, evaluating whether CDS and alerts specifically are an appropriate solution, or if different strategies or forms of CDS, such as passive alerts or search functionality, should instead be considered [3,6].

Limitations

Our participants were junior doctors in urban Australian hospitals who primarily used commercial information systems. Findings therefore may not generalize to other clinician groups, specialties, or settings, such as rural hospitals or those using homegrown systems. Additionally, we did not incorporate observational or system usage data, which may limit insights into subconscious behaviors. Future studies should incorporate mixed methods, that is, both clinicians' experiences of alert fatigue and actual responses to alerts and explore experiences of alert fatigue among different user groups.

To our knowledge, participants in our study used only knowledge-based alerts, limiting conclusions regarding experiences of alert fatigue for smarter alerts, such as those that are artificial intelligence (AI)-based. Our results can, however, offer important considerations. Although AI-based alerts can significantly improve clinical relevance and specificity [52], they may also increase cognitive burden if information is complex or constantly changing. Alternatively, given the fast-paced nature of the setting in which alerts are presented, AI-based alerts may be superficially processed and, due to their high positive predictive value, be automatically accepted rather than rejected due to automation bias [53]. Further research is needed to explore how alert fatigue manifests for smarter alerts in hospital settings and its impact on patient safety.

Conclusions

This study highlights alert fatigue as a dynamic, context-dependent experience that is not easily captured through studying behavior alone. Our findings reinforce the need to move beyond using singular metrics to measure alert fatigue and instead address alert fatigue using tailored strategies that target the specific causes and outcomes involved. While reducing low-value alerts and improving interface design remain important, broader approaches that support clinician well-being, foster positive organizational culture, and consider the impact of competing priorities and tasks may be equally critical.

XSL•FO
RenderX

Funding

NN is supported by a Digital Health Cooperative Research Centre (CRC) Limited scholarship. Digital Health CRC Limited is funded under the Australian Commonwealth's CRC Program. The funder had no involvement in the study design, data collection, analysis, interpretation, or writing of the manuscript.

Data Availability

The data analyzed in this study are available from the corresponding author upon reasonable request.

Authors' Contributions

NN (in consultation with all authors) conceptualized the study. NN collected data, transcribed data, and prepared data for the analysis. NN, MTB, ABA, and RF analyzed the data. NN drafted the manuscript, with all authors providing feedback. All authors provided input into study conception, design, and interpretation of findings and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Semistructured interview guide.
[DOCX File, 28 KB - jmir_v28i1e78676_app1.docx]

Multimedia Appendix 2
Proportion of coded references related to alert fatigue mapped to each Systems Engineering Initiative for Patient Safety (SEIPS) work system domain.
[DOCX File, 45 KB - jmir_v28i1e78676_app2.docx]

Multimedia Appendix 3
Strategies discussed by study participants for reducing alert fatigue in practice.
[DOCX File, 24 KB - jmir_v28i1e78676_app3.docx]

Checklist 1
SRQR checklist.
[PDF File, 153 KB - jmir_v28i1e78676_app4.pdf]

References

1. Wright A, Aaron S, Seger DL, Samal L, Schiff GD, Bates DW. Reduced effectiveness of interruptive drug-drug interaction alerts after conversion to a commercial electronic health record. J Gen Intern Med 2018 Nov;33(11):1868-1876. [doi: 10.1007/s11606-018-4415-9] [Medline: 29766382]
2. Kane-Gill SL, O'Connor MF, Rothschild JM, et al. Technologic distractions (part 1): summary of approaches to manage alert quantity with intent to reduce alert fatigue and suggestions for alert fatigue metrics. Crit Care Med 2017 Sep;45(9):1481-1488. [doi: 10.1097/CCM.0000000000002580] [Medline: 28682835]
3. Awad S, Loveday T, Lau R, Baysari MT. Development of a human factors–based guideline to support the design, evaluation, and continuous improvement of clinical decision support. Mayo Clin Proc Digit Health 2025 Mar;3(1):100182. [doi: 10.1016/j.mcpdig.2024.11.003] [Medline: 40206997]
4. Ancker JS, Edwards A, Nosal S, et al. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak 2017 Apr 10;17(1):36. [doi: 10.1186/s12911-017-0430-8] [Medline: 28395667]
5. Newton N, Bamgboje-Ayodele A, Forsyth R, Tariq A, Baysari MT. A systematic review of clinicians' acceptance and use of clinical decision support systems over time. NPJ Digit Med 2025 May 26;8(1):309. [doi: 10.1038/s41746-025-01662-7] [Medline: 40419669]
6. Blecker S, Pandya R, Stork S, et al. Interruptive versus noninterruptive clinical decision support: usability study. JMIR Hum Factors 2019 Apr 17;6(2):e12469. [doi: 10.2196/12469] [Medline: 30994460]
7. Osheroff J, Teich J, Levick D, et al. Improving Outcomes With Clinical Decision Support: An Implementer's Guide, 3rd edition: HIMSS Publishing; 2012:348.
8. Poly TN, Islam MM, Yang HC, Li YCJ. Appropriateness of overridden alerts in computerized physician order entry: systematic review. JMIR Med Inform 2020 Jul 20;8(7):e15653. [doi: 10.2196/15653] [Medline: 32706721]
9. Bittmann JA, Haefeli WE, Seidling HM. Modulators influencing medication alert acceptance: an explorative review. Appl Clin Inform 2022 Mar;13(2):468-485. [doi: 10.1055/s-0042-1748146] [Medline: 35981555]

XSL•FO
RenderX

10. Elias P, Peterson E, Wachter B, Ward C, Poon E, Navar AM. Evaluating the impact of interruptive alerts within a health system: use, response time, and cumulative time burden. Appl Clin Inform 2019 Oct;10(5):909-917. [doi: 10.1055/s-0039-1700869] [Medline: 31777057]
11. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17. [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
12. McDaniel RB, Burlison JD, Baker DK, et al. Alert dwell time: introduction of a measure to evaluate interruptive clinical decision support alerts. J Am Med Inform Assoc 2016 Apr;23(e1):e138-e141. [doi: 10.1093/jamia/ocv144] [Medline: 26499101]
13. McGreevey JD III, Mallozzi CP, Perkins RM, Shelov E, Schreiber R. Reducing alert burden in electronic health records: state of the art recommendations from four health systems. Appl Clin Inform 2020 Jan;11(1):1-12. [doi: 10.1055/s-0039-3402715] [Medline: 31893559]
14. Embi PJ, Leonard AC. Evaluating alert fatigue over time to EHR-based clinical trial alerts: findings from a randomized controlled study. J Am Med Inform Assoc 2012 Jun;19(e1):e145-e148. [doi: 10.1136/amiajnl-2011-000743] [Medline: 22534081]
15. Wang L, Goh KH, Yeow A, et al. Habit and automaticity in medical alert override: cohort study. J Med Internet Res 2022 Feb 16;24(2):e23355. [doi: 10.2196/23355] [Medline: 35171102]
16. Baysari MT, Tariq A, Day RO, Westbrook JI. Alert override as a habitual behavior—a new perspective on a persistent problem. J Am Med Inform Assoc 2017 Mar 1;24(2):409-412. [doi: 10.1093/jamia/ocw072] [Medline: 27274015]
17. Ash JS, Sittig DF, Campbell EM, Guappone KP, Dykstra RH. Some unintended consequences of clinical decision support systems. AMIA Annu Symp Proc 2007 Oct 11;2007:26-30. [Medline: 18693791]
18. Beurskens AJ, Bültmann U, Kant I, Vercoulen JH, Bleijenberg G, Swaen GM. Fatigue among working people: validity of a questionnaire measure. Occup Environ Med 2000 May;57(5):353-357. [doi: 10.1136/oem.57.5.353] [Medline: 10769302]
19. Ford E, Edelman N, Somers L, et al. Barriers and facilitators to the adoption of electronic clinical decision support systems: a qualitative interview study with UK general practitioners. BMC Med Inform Decis Mak 2021 Jun 21;21(1):193. [doi: 10.1186/s12911-021-01557-z] [Medline: 34154580]
20. Feldstein A, Simon SR, Schneider J, et al. How to design computerized alerts to safe prescribing practices. Jt Comm J Qual Saf 2004 Nov;30(11):602-613. [doi: 10.1016/s1549-3741(04)30071-7] [Medline: 15565759]
21. Russ AL, Zillich AJ, McManus MS, Doebbeling BN, Saleem JJ. Prescribers' interactions with medication alerts at the point of prescribing: a multi-method, in situ investigation of the human-computer interaction. Int J Med Inform 2012 Apr;81(4):232-243. [doi: 10.1016/j.ijmedinf.2012.01.002] [Medline: 22296761]
22. Holden RJ, Carayon P, Gurses AP, et al. SEIPS 2.0: a human factors framework for studying and improving the work of healthcare professionals and patients. Ergonomics 2013;56(11):1669-1686. [doi: 10.1080/00140139.2013.838643] [Medline: 24088063]
23. Russ AL, Fairbanks RJ, Karsh BT, Militello LG, Saleem JJ, Wears RL. The science of human factors: separating fact from fiction. BMJ Qual Saf 2013 Oct;22(10):802-808. [doi: 10.1136/bmjqs-2012-001450] [Medline: 23592760]
24. Carayon P, Schoofs Hundt A, Karsh BT, et al. Work system design for patient safety: the SEIPS model. Qual Saf Health Care 2006 Dec;15 Suppl 1(Suppl 1):i50-i58. [doi: 10.1136/qshc.2005.015842] [Medline: 17142610]
25. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
26. Adams WC. Conducting semi‐structured interviews. In: Newcomer KE, Hatry HP, Wholey JS, editors. Handbook of Practical Program Evaluation: Wiley; 2015:492-505. [doi: 10.1002/9781119171386]
27. Fossey E, Harvey C, McDermott F, Davidson L. Understanding and evaluating qualitative research. Aust N Z J Psychiatry 2002 Dec;36(6):717-732. [doi: 10.1046/j.1440-1614.2002.01100.x] [Medline: 12406114]
28. Wickens CD, Carswell CM. Information processing. In: Salvendy G, Karwowski W, editors. Handbook of Human Factors and Ergonomics, 5th edition: Wiley; 2021:114-158. [doi: 10.1002/9781119636113]
29. Joseph AL, Borycki EM, Kushniruk AW. Alert fatigue and errors caused by technology: a scoping review and introduction to the flow of cognitive processing model. Knowl Manag E-Learn 2021 Dec 30(4):500-521. [doi: 10.34105/j.kmel.2021.13.027]
30. Movahedi A, Sadooghiasl A, Ahmadi F, Vaismoradi M. A grounded theory study of alarm fatigue among nurses in intensive care units. Aust Crit Care 2023 Nov;36(6):980-988. [doi: 10.1016/j.aucc.2022.12.004] [Medline: 36737263]
31. Otter AI. 2025. URL: https://otter.ai/ [accessed 2026-02-03]
32. Saunders B, Sim J, Kingstone T, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907. [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
33. NVivo 14.24.1. Lumivero. 2024. URL: https://lumivero.com/products/nvivo/ [accessed 2026-02-03]
34. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
35. Kizzier-Carnahan V, Artis KA, Mohan V, Gold JA. Frequency of passive EHR alerts in the ICU: another form of alert fatigue? J Patient Saf 2019 Sep;15(3):246-250. [doi: 10.1097/PTS.0000000000000270] [Medline: 27331600]

XSL•FO
RenderX

36. Obisesan O, Barber E, Martin P, Brougham N, Tymkew H. Original research: alarm fatigue: exploring the adaptive and maladaptive coping strategies of nurses. Am J Nurs 2024 Oct 1;124(10):24-30. [doi: 10.1097/01.NAJ.0001063808.07614.8d] [Medline: 39259867]
37. Gani I, Litchfield I, Shukla D, Delanerolle G, Cockburn N, Pathmanathan A. Understanding "alert fatigue" in primary care: qualitative systematic review of general practitioners' attitudes and experiences of clinical alerts, prompts, and reminders. J Med Internet Res 2025 Feb 7;27:e62763. [doi: 10.2196/62763] [Medline: 39918864]
38. Khairat S, Coleman C, Ottmar P, Jayachander DI, Bice T, Carson SS. Association of electronic health record use with physician fatigue and efficiency. JAMA Netw Open 2020 Jun 1;3(6):e207385. [doi: 10.1001/jamanetworkopen.2020.7385] [Medline: 32515799]
39. Wu Y, Wu M, Wang C, Lin J, Liu J, Liu S. Evaluating the prevalence of burnout among health care professionals related to electronic health record use: systematic review and meta-analysis. JMIR Med Inform 2024 Jun 12;12:e54811. [doi: 10.2196/54811] [Medline: 38865188]
40. McCoy AB, Waitman LR, Lewis JB, et al. A framework for evaluating the appropriateness of clinical decision support alerts and responses. J Am Med Inform Assoc 2012;19(3):346-352. [doi: 10.1136/amiajnl-2011-000185] [Medline: 21849334]
41. van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. J Am Med Inform Assoc 2006;13(2):138-147. [doi: 10.1197/jamia.M1809] [Medline: 16357358]
42. Salwei ME, Carayon P, Hoonakker PLT, et al. Workflow integration analysis of a human factors–based clinical decision support in the emergency department. Appl Ergon 2021 Nov;97:103498. [doi: 10.1016/j.apergo.2021.103498] [Medline: 34182430]
43. Murad DA, Tsugawa Y, Elashoff DA, Baldwin KM, Bell DS. Distinct components of alert fatigue in physicians' responses to a noninterruptive clinical decision support alert. J Am Med Inform Assoc 2022 Dec 13;30(1):64-72. [doi: 10.1093/jamia/ocac191] [Medline: 36264258]
44. Orenstein EW, Kandaswamy S, Muthu N, et al. Alert burden in pediatric hospitals: a cross-sectional analysis of six academic pediatric health systems using novel metrics. J Am Med Inform Assoc 2021 Nov 25;28(12):2654-2660. [doi: 10.1093/jamia/ocab179] [Medline: 34664664]
45. Baysari MT, Dort BAV, Zheng WY, et al. Prescribers' reported acceptance and use of drug-drug interaction alerts: an Australian survey. Health Informatics J 2022;28(2):14604582221100678. [doi: 10.1177/14604582221100678] [Medline: 35531625]
46. Blackley SV, Lo YC, Varghese S, et al. Building an allergy reconciliation module to eliminate allergy discrepancies in electronic health records. J Am Med Inform Assoc 2025 Apr 1;32(4):648-655. [doi: 10.1093/jamia/ocaf022] [Medline: 39969950]
47. Chien SC, Yang HC, Chen CY, et al. Using alert dwell time to filter universal clinical alerts: a machine learning approach. Comput Methods Programs Biomed 2023 Oct;240:107696. [doi: 10.1016/j.cmpb.2023.107696] [Medline: 37480643]
48. Zachariah M, Phansalkar S, Seidling HM, et al. Development and preliminary evidence for the validity of an instrument assessing implementation of human-factors principles in medication-related decision-support systems—I-MeDeSA. J Am Med Inform Assoc 2011 Dec;18 Suppl 1(Suppl 1):i62-i72. [doi: 10.1136/amiajnl-2011-000362] [Medline: 21946241]
49. Phansalkar S, Zachariah M, Seidling HM, Mendes C, Volk L, Bates DW. Evaluation of medication alerts in electronic health records for compliance with human factors principles. J Am Med Inform Assoc 2014 Oct;21(e2):e332-e340. [doi: 10.1136/amiajnl-2013-002279] [Medline: 24780721]
50. Ng HJH, Kansal A, Abdul Naseer JF, et al. Optimizing Best Practice Advisory alerts in electronic medical records with a multi-pronged strategy at a tertiary care hospital in Singapore. JAMIA Open 2023 Aug 1;6(3):ooad056. [doi: 10.1093/jamiaopen/ooad056] [Medline: 37538232]
51. Hussain MI, Reynolds TL, Zheng K. Medication safety alert fatigue may be reduced via interaction design and clinical role tailoring: a systematic review. J Am Med Inform Assoc 2019 Oct 1;26(10):1141-1149. [doi: 10.1093/jamia/ocz095] [Medline: 31206159]
52. Graafsma J, Murphy RM, van de Garde EMW, et al. The use of artificial intelligence to optimize medication alerts generated by clinical decision support systems: a scoping review. J Am Med Inform Assoc 2024 May 20;31(6):1411-1422. [doi: 10.1093/jamia/ocae076] [Medline: 38641410]
53. Abdelwanis M, Alarafati HK, Tammam MMS, Simsekler MCE. Exploring the risks of automation bias in healthcare artificial intelligence applications: a Bowtie analysis. J Saf Sci Resil 2024 Dec;5(4):460-469. [doi: 10.1016/j.jnlssr.2024.06.001]

Abbreviations

AI: artificial intelligence
CDS: clinical decision support
CRC: Cooperative Research Centre
EHR: electronic health record
EMR: electronic medical record

XSL•FO
RenderX

JMO: junior medical officer
PGY: postgraduate year
SEIPS: Systems Engineering Initiative for Patient Safety
SRQR: Standards for Reporting Qualitative Research

Edited by A Stone, TDA Cardoso; submitted 06.Jun.2025; peer-reviewed by J Aarts, PV Asaro, R Marshall, S Lloyd; revised version received 27.Oct.2025; accepted 11.Nov.2025; published 19.Feb.2026.

Please cite as:
Newton N, Bamgboje-Ayodele A, Forsyth R, Tariq A, Huang J, Yannam R, Lalor DJ, Sobey AJ, Baysari MT
Experiences of Alert Fatigue and Its Contributing Factors in Hospitals: Qualitative Study
J Med Internet Res 2026;28:e78676
URL: https://www.jmir.org/2026/1/e78676
doi:10.2196/78676

© Nicki Newton, Adeola Bamgboje-Ayodele, Rowena Forsyth, Amina Tariq, Jing Huang, Ramyasri Yannam, Daniel J Lalor, Andrew James Sobey, Melissa T Baysari. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Improving the Understandability of Clinical Guidelines: Development and Evaluation of a GPT-4–Based Pipeline

Matthew D Jones[1], PhD; Melissa Torgbi[2], MSc; Harish Tayyar Madabushi[2], PhD

[1]Department of Life Sciences, University of Bath, Claverton Down, Bath, United Kingdom
[2]Department of Computer Science, University of Bath, Bath, United Kingdom

Corresponding Author:
Matthew D Jones, PhD
Department of Life Sciences, University of Bath, Claverton Down, Bath, United Kingdom

Abstract

Background: Difficulty in finding and understanding information in clinical guidelines contributes to medication errors. Large language models (LLMs) can simplify complex text to aid in understanding, but this approach to improving the quality of guidelines has not been investigated. However, LLMs are also known to hallucinate or generate outputs that may not align with reality.

Objective: This study aimed to develop and evaluate an LLM pipeline to improve the readability of clinical guidelines while ensuring the preservation of critical content.

Methods: To align LLM revisions with research evidence and enable comparison with manual editing, the National Health Service Injectable Medicines Guide (IMG) was used as a case study, to which a GPT-4–based pipeline was applied, with prompts based on user testing–derived recommendations for IMG authors. This enabled readability comparisons between various IMG guideline versions: original, manually revised, or GPT-4–revised using the user testing–derived recommendations, and fully user tested. Readability was evaluated using readability metrics and ratings from 3 expert pharmacists. Content similarity before and after LLM revision was assessed using BERT (bidirectional encoder representations from transformers) scores and expert pharmacist review.

Results: Considering 20 IMG guidelines used in practice, BERT scores indicated high semantic similarity between the original and LLM-revised guidelines (0.88-0.96). An omission, addition, or change in meaning was identified by at least one pharmacist in 30 (20%), 7 (5%), and 18 (12%) of the 153 guideline subsections, respectively. The SMOG (Simple Measure of Gobbledygook) grade showed a small but significant improvement in readability for the LLM-revised guidelines (mean difference 0.32, 95% CI 0.10 - 0.55; P=.02) and the manually revised versions (mean difference 0.46, 95% CI 0.13 - 0.79; P=.03). There was no significant difference between the LLM and manually revised versions (P>.99). There were no significant differences between Flesch-Kincaid reading grades (P=.91). Expert ratings favored the LLM-revised versions for understandability. Considering 2 IMG guidelines from previous research, user testing produced a greater improvement in readability than LLM revision.

Conclusions: Authors should not use current LLMs to modify clinical guidelines without carefully checking the revised text for unintended omissions, additions, or changes in meaning. Further work should investigate the potential of LLMs to augment manual user testing and reduce the barriers to the wider use of this approach to improve the safety of clinical guidelines.

(J Med Internet Res 2026;28:e81915) doi:10.2196/81915

KEYWORDS

artificial intelligence; clinical guidelines; comprehension; guidelines as topic; health care professionals; health personnel; large language models; patient safety; readability; understandability

Introduction

Medication errors are a leading cause of avoidable harm in health care systems, with an estimated international cost of US $42 billion per annum [1]. Numerous factors can lead to a medication error [2], one of which is difficulty in finding or understanding prescribing and medication administration information in clinical guidelines [3-5]. Consequently, guidelines that are contradictory, incomprehensible, or of poor quality are included in frameworks of the causes of patient safety incidents [6]. Clarity of writing and formatting also influences guideline adoption [7].

Few studies have explored ways to help clinicians more easily find and understand guidelines. Some evidence suggests that iterative user testing and redesign can improve guideline usability by identifying and addressing problem areas [8,9]. In one study, user-tested injectable medicines guidelines led to a 2.5-fold increase in error-free intravenous administration and a 96% probability of cost-effectiveness [10,11]. Such improvements result from both the testing process and the

XSL•FO
RenderX

application of good practice in information writing and design [12]. Consequently, in 2020, format changes and clear writing advice were added to the authors' instructions for the United Kingdom's National Health Service Injectable Medicines Guide (IMG). This is an online guide to preparing and administering injectable medicines, mainly used by nurses. Subsequently, more than 300 drug guidelines have been manually updated, but not user tested.

Most guideline-related medication errors involve documents produced by local health care organizations [3]. User testing and revising these "local guidelines" could therefore yield the greatest safety improvements. As there are many thousands of these documents within a single country, limited expertise in information writing and design within health care organizations may hinder this approach.

Large language models (LLMs), such as GPT-4, may help overcome this barrier by simplifying complex text. This approach has been tested on health information for patients initially written by subject experts. Early results are mixed: LLMs can improve readability metrics [13-18] and patient understanding [15] but can also worsen readability metrics [19]. These gains may come at a cost, with some content omitted and up to 21% of revised content deemed clinically inappropriate [13-15,19]. LLMs have also been used to improve understanding of specialist medical records [20]. However, they have not yet been applied to improve the readability of clinical guidelines for health care professionals; therefore, it remains unclear whether LLMs can enhance readability without compromising content accuracy or safety.

Therefore, this study evaluated a GPT-4–based pipeline to improve guideline readability while safeguarding against omission, addition, or change of meaning of safety-critical content. To enable comparison with manual revision and align with previous research, the pipeline was applied to the IMG. This enabled various readability comparisons (Figure 1), as well as comparison of the content of guidelines before and after LLM revision.

Specific objectives were therefore to (1) develop a GPT-4–based pipeline to improve guideline clarity with built-in safeguards; (2) evaluate content similarity of 20 guidelines before and after revision by the GPT-4 pipeline; (3) compare the readability of 20 equivalent original guidelines, LLM-revised guidelines, and manually revised guidelines; and (4) compare the readability of original, LLM-revised, and user-tested aminophylline and voriconazole guidelines.

Figure 1. Readability comparisons reported in this study. "Original" versions were the guidelines in use in the Injectable Medicines Guide (IMG) in 2019. "LLM-revised" versions were developed from the original guidelines by the large language model (LLM) pipeline. "Manually revised" versions were the guidelines in use in the IMG in 2024, which had been updated by guideline authors in line with revised authors' instructions introduced in 2020. "User-tested" versions were developed from the original guidelines via a user-testing process during a previous study [9].

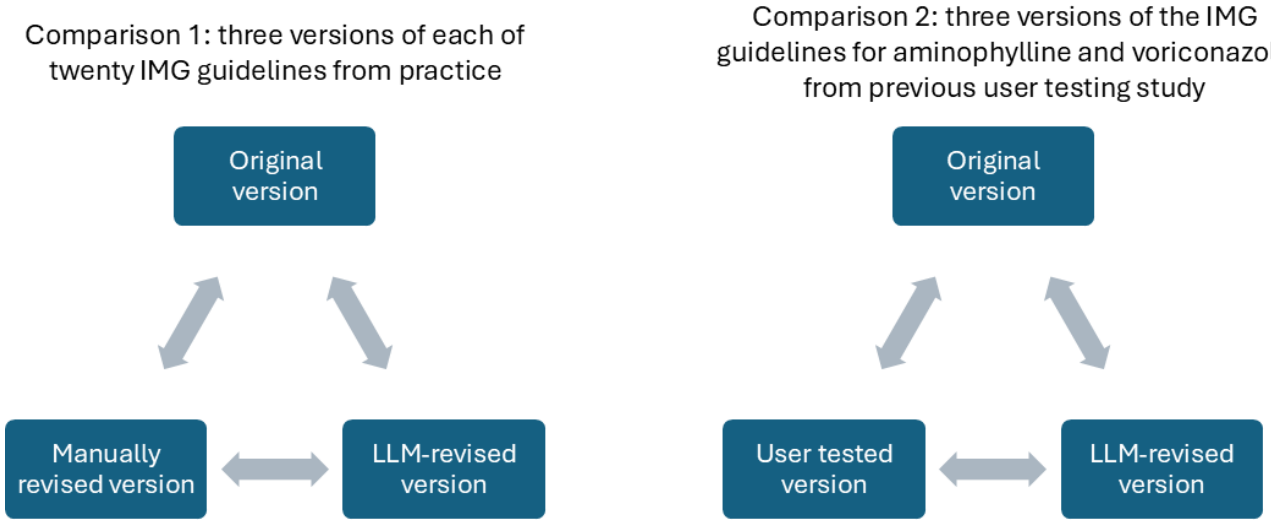

Methods

This study is reported in accordance with the TRIPOD-LLM (Transparent Reporting of a Multivariable Model for Individual Prognosis or Diagnosis - Large Language Model) guideline [21].

Development and Application of the GPT-4 Pipeline

IMG guidelines were converted to Markdown and paired with iteratively developed prompts for input to the LLM pipeline. Prompts instructed the LLM to improve clarity and readability without altering content. The LLM pipeline used the OpenAI API and the GPT-4 model. GPT-4–generated "LLM-revised" Markdown versions of the original guidelines, which were then converted to static HTML for display.

Prompting Strategies

Prompting was based on the Reframing Principles for Instruction-Following LLMs [22], which offer tested strategies for constructing effective prompts, including using low-level patterns, examples, simplified sentence structures, and clear constraints. Initially, researchers with expertise in computer science (MT and HTM) constructed prompts based on the user testing–derived recommendations incorporated into the IMG authors' instructions in 2020 [9]. To prevent omissions or changes in meaning, prompts required a semantic similarity score above 0.9 before proceeding. While GPT-4 cannot

XSL•FO
RenderX

compute semantic accuracy, we hypothesized that it would interpret this instruction to ensure edits were semantically close to the original and that this reference would make the requirement more explicit. Selected outputs of these initial prompts were reviewed by a pharmacist with expertise in intravenous medication administration and information design (MDJ) and discussed with the wider research team. These initial tests found that overly prescriptive instructions for human authors caused overgeneralization and errors, including omissions and additions. Prompt instructions were then iteratively refined to improve effectiveness. The final prompts (Multimedia Appendix 1, online supplement) used for both comparisons were a product of this iterative development, incorporating insights from our explorations while omitting some instructions originally intended for human authors.

GPT-4–generated instructions were also explored for comparison 2 (Figure 1). The original and user-tested guidelines for voriconazole and aminophylline were given to GPT-4, which was instructed to generate prompts for another model to revise the original guidelines to be similar to the user-tested versions, without directly referencing them. However, these prompts were ineffective and were therefore excluded from final experiments.

Selection of IMG Guidelines

Twenty adult IMG guidelines were selected in consultation with the IMG editorial team to cover a range of drugs, administration methods (infusion and injection), reconstitution or dilution needs, and higher-risk or complex cases. Versions from 2019 ("original") and 2024 ("manually revised," updated in line with revised IMG authors' instructions) were obtained. Voriconazole and aminophylline guidelines from 2019 and their user-tested revisions were also sourced from prior research [9].

Guideline Content Similarity

Content similarity of the 20 IMG guidelines from practice before and after LLM revision was assessed using the bidirectional encoder representations from transformers (BERT) score [23] and expert review. The BERT score calculates the similarity between each word or token in a candidate and reference sentence using cosine similarity and the contextual embeddings from the BERT model [20]. We used the standard BERT score library to perform this calculation [24]. Scores range from 0 to 1, with higher values indicating greater similarity.

Three intensive care pharmacists independently assessed 9 sections of LLM-revised guidelines (method of administration, reconstitution, dilution, expiry, flushing, adverse effects, extravasation, other comments, and compatibility) for clinically relevant omissions, additions, or changes in meaning compared with the originals. The pharmacists were known to the research team through professional networks and selected for their significant experience in intensive care pharmacy, although this was not formally defined. All invited pharmacists participated and were aware of the study aims, although they were blinded to guideline version while assessing content.

In addition, the outputs were evaluated manually by the research team for modifications to medication names or their order of presentation.

Guideline Readability

Readability was assessed with 2 metrics and expert review. The readability metrics used were the SMOG (Simple Measure of Gobbledygook) grade and the Flesch-Kincaid reading grade. SMOG was chosen for its validation against full comprehension, while Flesch-Kincaid is widely used in health research [25]. Both estimate the years of education required to understand a text, so larger scores indicate lower readability. Scores were calculated with the Textstat Python library using the entire guideline text.

The relative readability of the 20 IMG guidelines from practice before and after LLM revision was also independently assessed by the 3 intensive care pharmacists. Comparing the original and revised versions of each guideline, they rated relative overall and subsection-level understandability using a 5-point Likert scale.

Statistical Analysis

Analyses were performed in Stata v18.0 (StataCorp LLC). Readability metrics and BERT scores were summarized as mean (SD), with differences tested using repeated measures ANOVA with post-hoc Bonferroni-corrected paired *t* tests. Expert ratings were summarized descriptively, and interrater reliability assessed using Gwet's AC to account for skewed distributions [26].

Ethical Considerations

This study was assessed as having low potential to do harm and thus reviewed and given a favorable opinion by proportionate review by the Department of Life Science Departmental Research Ethics Officer at the University of Bath on June 22, 2023 (number 442).

Results

Pipeline Development

The LLM pipeline successfully produced output text showing revised versions of the 20 IMG guidelines from practice and the voriconazole and aminophylline guidelines from previous user-testing research. These revised guidelines were then evaluated for content similarity using the methods described earlier.

Content Similarity of the Set of 20 Guidelines From Practice

Table 1 presents the BERT scores comparing the semantic similarity of the 20 IMG guidelines from practice before and after LLM revision. These ranged from 0.88 to 0.96, indicating a high degree of similarity. As a benchmark, Hanna and Bojar [27] reported an average BERT score of 0.815 when comparing 2 different, professional human translations of the same source sentences across several distinct linguistic phenomena. The 3 pharmacists' ratings of clinically relevant omissions, additions, or changes in meaning following LLM revision of the 20 guidelines from practice showed a high degree of interrater reliability, with Gwet's ACs of 0.87 (omissions), 0.97 (additions), and 0.92 (changes in meaning). Figures 2-4 show the number of pharmacist reviewers who identified an omission, addition, or change in meaning in each subsection. Of 459

XSL•FO
RenderX

individual subsection ratings (153 subsections each rated by 3 pharmacists), 54 (11.8%) identified an omission, 8 (1.7%) identified an addition, and 20 (4.4%) identified a change in meaning. An omission, addition, or change in meaning was identified by at least one pharmacist in 30 (20%), 7 (5%), and 18 (12%), respectively, of the 153 individual subsections. However, only 8 subsections had an omission identified by all 3 reviewers, and no additions or changes of meaning were identified by all 3 reviewers (Table 2). Overall, 65% (36/55) of subsection omissions, additions, or changes of meaning were only identified by 1 reviewer. These differences were concentrated in certain subsections (method of administration, reconstitution, dilution, adverse effects, and other comments) and guidelines (amiodarone, amoxicillin, furosemide, and propofol). Only one subsection (flushing) and two guidelines (levetiracetam and paracetamol) had no omissions, additions, or changes in meaning identified by any pharmacist.

Table . BERT (bidirectional encoder representations from transformers) scores comparing the semantic similarity of the 20 Injectable Medicines Guide guidelines from practice before and after large language model revision.[a]

Guideline	BERT score
Amiodarone	0.95
Amoxicillin	0.91
Ceftriaxone	0.95
Cyclizine	0.91
Fentanyl	0.96
Flucloxacillin	0.90
Furosemide	0.88
Gentamicin	0.93
Levetiracetam	0.92
Magnesium sulfate	0.90
Meropenem	0.93
Metronidazole	0.90
Noradrenaline	0.94
Omeprazole	0.94
Paracetamol	0.94
Phenytoin	0.88
Piperacillin and tazobactam	0.95
Propofol	0.91
Teicoplanin	0.93
Vancomycin	0.88

[a]Mean (SD): 0.92 (0.02).

XSL•FO
RenderX

Figure 2. Heatmap showing the number of pharmacist reviewers who identified an omission of information following large language model revision of each subsection of each of the 20 original guidelines. Blank subsections were not present in the relevant guideline. AE: adverse effects; Compat.: compatibility; Expiry: expiry time; Extrav.: extravasation; Flush: flushing; Method: method of administration; Other: other comments; Recon.: reconstitution.

Guideline	Subsection								
	Method	Recon.	Dilution	Expiry	Flush	AEs	Extrav.	Other	Compat.
Amiodarone	1		3	0	0	0	0	0	3
Amoxicillin	3	1	2		0	0	0	3	1
Ceftriaxone	0	1	0		0	0	0	0	0
Cyclizine	0		1		0	0	0	0	0
Fentanyl	0		0	0	0	0	0	0	3
Flucloxacillin	0	0	0		0	0	0	0	0
Furosemide	2		0	0	0	0	2	1	0
Gentamicin	0		0		0	0	0	0	0
Levetiracetam	0		0		0	0	0	0	0
Magnesium sulfate	0		0	0	0	0	0	0	0
Meropenem	0	0	1		0	0	0	0	0
Metronidazole	0				0	0	0	0	0
Noradrenaline	0		0	0	0	0	0	0	1
Omeprazole	0	1	3	0	0	0	0	0	1
Paracetamol	0				0	0	0	0	0
Phenytoin	0		0		0	0		3	0
Piperacillin and tazobactam	0	0	0		0	0	0	1	0
Propofol	1		0	0	0	0	0	1	2
Teicoplanin	0	2	0		0	0	0	2	0
Vancomycin	0	2	1	0	0	0	2	3	0

Figure 3. Heatmap showing the number of pharmacist reviewers who identified an addition of information following large language model revision of each subsection of each of the 20 original guidelines. Blank subsections were not present in the relevant guideline. AE: adverse effects; Compat.: compatibility; Expiry: expiry time; Extrav.: extravasation; Flush: flushing; Method: method of administration; Other: other comments; Recon.: reconstitution.

Guideline	Subsection								
	Method	Recon.	Dilution	Expiry	Flush	AEs	Extrav.	Other	Compat.
Amiodarone	0		0	0	0	0	0	0	0
Amoxicillin	0	0	0		0	2	0	1	0
Ceftriaxone	0	0	0		0	0	0	0	0
Cyclizine	0		0		0	0	0	0	0
Fentanyl	0		0	0	0	0	0	0	0
Flucloxacillin	0	0	1		0	0	0	0	0
Furosemide	0		0	0	0	0	0	0	0
Gentamicin	0		0		0	0	0	0	0
Levetiracetam	0		0		0	0	0	0	0
Magnesium sulfate	0		0	1	0	1	0	0	0
Meropenem	0	0	0		0	0	0	0	0
Metronidazole	0				0	0	0	0	0
Noradrenaline	0		0	0	0	0	0	0	0
Omeprazole	0	0	0	0	0	0	0	0	0
Paracetamol	0				0	0	0	0	0
Phenytoin	0		0		0	1		0	0
Piperacillin and tazobactam	0	0	0		0	0	0	0	0
Propofol	0		0	0	0	0	0	1	0
Teicoplanin	0	0	0		0	0	0	0	0
Vancomycin	0	0	0	0	0	0	0	0	0

XSL•FO
RenderX

Figure 4. Heatmap showing the number of pharmacist reviewers who identified a change in meaning following large language model revision of each subsection of each of the 20 original guidelines. Blank subsections were not present in the relevant guideline. AEs: adverse effects; Compat.: compatibility; Expiry: expiry time; Extrav.: extravasation; Flush: flushing; Method: method of administration; Other: other comments; Recon.: reconstitution.

Guideline	Subsection								
	Method	Recon.	Dilution	Expiry	Flush	AEs	Extrav.	Other	Compat.
Amiodarone	0		2	0	0	0	0	0	2
Amoxicillin	1	1	1		0	1	0	1	1
Ceftriaxone	0	0	0		0	0	0	0	0
Cyclizine	0		0		0	0	0	0	0
Fentanyl	0		0	0	0	0	0	0	1
Flucloxacillin	0	0	0		0	0	0	0	0
Furosemide	0		0	0	0	1	0	1	0
Gentamicin	1		0		0	1	0	0	0
Levetiracetam	0		0		0	0	0	0	0
Magnesium sulfate	0		0	0	0	0	0	0	0
Meropenem	0	1	0		0	0	0	0	0
Metronidazole	0				0	0	0	0	1
Noradrenaline	0		0	0	0	0	0	0	0
Omeprazole	0	0	1	0	0	0	0	0	0
Paracetamol	0				0	0	0	0	0
Phenytoin	0		0		0	0		0	0
Piperacillin and tazobactam	0	0	0		0	0	0	0	0
Propofol	0		0	0	0	0	0	1	0
Teicoplanin	1	0	0		0	0	0	0	0
Vancomycin	0	0	0	0	0	0	0	0	0

Table . The number of subsections with an identified omission, addition, or change of meaning following large language model revision of each of the 20 original guidelines.

	The number of subsections with an addition, omission, or change of meaning		
	Identified by 1 reviewer	Identified by 2 reviewers	Identified by 3 reviewers
Omission	14	8	8
Addition	6	1	0
Change	16	2	0

Key differences described in the pharmacists' free text comments included the occasional omission of the contents of an entire subsection (eg, extravasation, other comments, and compatibility), the omission of one specific detail (eg, information relating to one of several doses of a medicine), objective changes in meaning (eg, "hypotension" changed to "hypertension" and "mg/mL" changed to "mg") and subjective changes in interpretation (eg, the revised wording implying an adverse effect is probable rather than possible or lessening a safety warning).

Manual checks identified no alterations to medication names or order of presentation.

Readability of the 20 Guidelines From Practice

Table 3 presents overall readability metrics for the 3 different versions of the IMG from practice. Table S1 in Multimedia Appendix 1 shows the metrics calculated for each individual guideline. Repeated measures ANOVA identified a significant difference between the SMOG grades of at least 2 guideline versions (P=.007; df=59), so post hoc paired t tests with Bonferroni corrections were performed. The mean difference in SMOG grade between the original and LLM-revised versions was 0.32 (95% CI 0.10 - 0.55; P=.02; df=19), indicating a small but significant improvement in readability following LLM revision. Similarly, the mean difference in SMOG grade between the original and manually revised versions was 0.46 (95% CI 0.13 - 0.79; P=.03; df=19), indicating significant improvement in the readability of the guidelines following changes to the writing guide in 2020. However, the mean difference in SMOG grade between the LLM-revised and manually revised versions was 0.14 (95% CI −0.18 to 0.45; P>.99; df=19), suggesting no difference in the readability of these 2 versions. In contrast, repeated measures ANOVA did not identify differences between the Flesch-Kincaid reading grades of the original, LLM-revised, and manually revised versions (P=.91; df=59).

XSL•FO
RenderX

Table . Summary readability metrics for 3 different versions of injectable medicines guides from practice for 20 intravenous drugs.

	SMOG[a] grade, mean (SD)	Flesch-Kincaid grade, mean (SD)
Original version	12.3 (0.2)	12.1 (0.4)
LLM[b]-revised version	12.0 (0.2)	12.0 (0.4)
Manually revised version	11.8 (0.1)	11.9 (0.3)

[a]SMOG: Simple Measure of Gobbledygook.

[b]LLM: large language model.

The 3 pharmacists' ratings of the relative readability of the 20 guidelines from practice before and after LLM revision showed a high degree of interrater reliability, with Gwet's ACs of 0.75 (overall guideline) and 0.85 (subsections). Table 4 presents the pharmacists' ratings of the overall understandability of the guidelines, which favored the LLM-revised versions being easier to understand (26/60 ratings, 43.3%) compared with the original versions (11/60 ratings, 18.3%). Figure S1 (multimedia supplement) summarizes the pharmacists' ratings of the relative readability of each subsection. Of 459 individual ratings, 56 favored the original version and 125 favored the LLM-revised version. For 6 of the 9 subsections and 17 of the 20 drugs, there were more ratings that favored the LLM-revised version than the original version. The pharmacists' free text comments on relative readability frequently mentioned more concise language and a clearer layout for the LLM-revised versions resulting especially from the use of bullet points. However, they also noted that this was sometimes achieved through the omission of information.

Table . The 3 pharmacists' ratings of the overall relative understandability of the 20 original and large language model (LLM)–revised guidelines.

	Number of ratings, n (%)[a]
2019 version is much easier to understand	1 (1.7)
2019 version is slightly easier to understand	10 (16.7)
Both versions are equally easy to understand	23 (38.3)
LLM-revised version is slightly easier to understand	25 (41.7)
LLM-revised version is much easier to understand	1 (1.7)

[a]Three pharmacist reviewers each independently rated 20 guidelines, giving a total of 60 ratings.

Readability of the User- Tested Guidelines

Table 5 presents the readability metrics for 3 different versions of injectable medicines guides for aminophylline and voriconazole that underwent user testing and subsequent revision in a previous study [9]. For aminophylline, LLM revision did not decrease the reading grade, but following user testing and revision, both readability metrics fell. For voriconazole, LLM revision decreased both readability grades, but user testing and subsequent revision decreased them even further.

Table . Guideline readability metrics for three different versions of injectable medicines guides for aminophylline and voriconazole that underwent user testing and subsequent revision in a previous study [9].

	Original version	LLM[a]-revised version	User-tested version
SMOG[b] grade			
Aminophylline	11.7	11.8	11.0
Voriconazole	11.9	10.9	10.6
Flesch-Kincaid grade			
Aminophylline	9.7	9.8	8.3
Voriconazole	10.5	10.3	7.9

[a]LLM: large language model.

[b]SMOG: Simple Measure of Gobbledygook.

Discussion

Principal Findings

The GPT-4–based pipeline did not introduce clinically relevant omissions, additions, or changes of meaning to≥80% of individual subsections within the 20 guidelines from practice. However, content was entirely similar in only a few individual guidelines or subsection types, and differences (especially omissions) were concentrated in certain drugs or subsections in an unpredictable way. It is noteworthy that in the cases of

XSL•FO
RenderX

furosemide, phenytoin, and vancomycin, GPT-4 did not maintain a BERT score above 0.9.

SMOG grades suggested that when following similar instructions or prompts, the LLM-based pipeline was as effective as manual editing in producing a small improvement in readability, and this was supported by the expert reviewers' ratings of understandability. In contrast, the Flesch-Kincaid reading grade did not find any change in readability. This may be because the Flesch-Kincaid reading grade reflects a lower standard of comprehension than the SMOG grade [25] but reinforces the marginal nature of the readability improvement. In addition, the GPT-4 pipeline did not produce improvements in either readability metric as large as those resulting from user testing and subsequent document revision.

Comparison With Prior Work

These results are aligned with those of studies that have applied LLMs to improve the readability of patient information, which have demonstrated improvements in readability metrics [13-18] that are often associated with the omission of information [13,15,19]. This association between readability and omission of content is unsurprising, as the removal of complex concepts is likely to make a text easier to understand. However, in a safety-critical context, such as health care, the omission of essential information must be avoided, and evidence to date suggests that LLMs are currently unable to achieve this reliably. In contrast, an iterative process of user testing and subsequent document revision has a lower risk of introducing unintended content changes. In addition, while this study has demonstrated that user testing can improve the readability metrics of clinical guidelines, it has also been shown to have beneficial effects on more important user- or patient-centered outcomes, such as reducing difficulties finding and understanding information [8,9,28], and subsequently preventing guideline-related medication errors [11].

Therefore, user testing remains the guideline improvement technique with the strongest evidence base. However, the findings of this study suggest that specifically designed LLM-based pipelines with appropriate safeguards do have potential to augment (but not replace) a manual user testing process, by providing suggestions for potential improvements during the document revision process. These safeguards should be included as explicit clinician-guided, hard-coded programmatic feature checks (eg, verifying that all medications are included and in the correct order) rather than as an instruction to the LLM, which may be ignored. This approach has the potential to partially address the limited availability of knowledge and skills in information writing and design within health care organizations.

Strengths and Limitations

This study has several strengths, including the use of specific prompts based on previous research and the use of both manually revised and user-tested monographs as comparators. In addition, it evaluated both the content and readability of the guidelines using a combination of calculated metrics and human evaluation, an approach recently recommended for studies in this field [29]. However, there are also a number of limitations. First, user- and patient-centered outcomes might have been evaluated (eg, the ability of end users to find and understand information in the guidelines; the frequency of medication errors when using revised guidelines), so it is not clear whether the observed small improvements in readability would result in improved patient care. This could be addressed by future randomized studies comparing original and LLM-revised guidelines in actual or simulated patient care. Second, this study used a general LLM rather than a bespoke model trained with health care–specific information [29], which might have resulted in greater improvements in readability and/or fewer content changes. Such models should be used in future studies. Third, while the results were replicable among the 22 drug guides tested in this study, only one type of guideline relating to intravenous drug administration was used. Therefore, results may not be replicable to other types of guideline, especially as the prompting strategy was based on user testing of these specific guidelines. Other types of guidelines and more general prompting strategies should be investigated in future studies. Fourth, the inclusion of a prompt based on semantic similarity did not provide a reliable safeguard against content changes, as the BERT score for 3 guidelines was <0.9 (Table 1). Future studies should implement semantic similarity thresholds as programmatic post hoc checks (compute the BERT score externally and rerun or reject LLM outputs) and investigate more extensive use of clinician-guided hard-coded programmatic feature checks (eg, verifying that all medications are included and in the correct order). Fifth, the prompting strategy was developed by only a small team, so it is more likely to have been affected by participant subjectivity. This could be addressed by developing prompts among a larger team with a greater range of expertise. Sixth, instruction 9 in the GPT-4 prompts ("remove information about the size of the bags") was based on a misinterpretation of previous user testing–derived recommendations. Six of 12 content review panel comments describing omissions from the "Dilution" subsection related to the clinically relevant omission of infusion bag volume, suggesting this error contributed to some of the unintended omissions identified in this study.

Implications for Practice and Future Research

The present findings suggest that guideline authors should not use current LLMs to modify clinical guidelines without carefully checking the revised text for unintended omissions, additions, or changes of meaning. Similarly, clinicians should not use current LLMs to answer clinical questions without carefully checking the results against reliable sources of information. However, LLMs do have the potential to augment manual user testing and reduce the barriers to the wider use of this approach to improve the safety of the thousands of local clinical guidelines currently in use. Further work is required to embed LLM functionality into guideline authoring systems and subsequently demonstrate the safety and effectiveness of this approach.

Conclusions

LLMs are able to revise the contents of clinical guidelines to produce small improvements in readability, but with unintended omissions, additions, or changes of meaning to a minority of the information. Small improvements in readability were evident

XSL•FO
RenderX

using the SMOG grade and expert review, but not with the Flesch-Kincaid reading grade. It is not known whether these readability improvements would lead to better understanding of, or adherence to, clinical guidelines when used in practice. Omissions of information accounted for the majority of unintended changes to the information contained within the guidelines. Guideline-related medication errors continue to occur but may be prevented by the combination of human expertise, user feedback, and appropriate technologies.

Acknowledgments

The authors are grateful to members of the Injectable Medicines Guide editorial team, the pharmacist reviewers, Holly Wilson (Onaya Science), and Joseph Marvin Imperial (University of Bath) for their support of this study. Microsoft Copilot (GPT-5 model) was used during manuscript revision to edit the Introduction and Methods sections for brevity.

Funding

This project was funded by a Pump Priming Award from the Department of Life Sciences, University of Bath. Additional funding was received from the Academic Secondment Scheme of Research and Innovation Services at the University of Bath. These funders had no role in study design; collection, analysis, and interpretation of data; writing of the paper; and/or decision to submit for publication.

Data Availability

Data and code are available in a public, open access repository. Data created during this research that are not presented in full in this paper are openly available from [30].

Authors' Contributions

MDJ: conceptualization, methodology, formal analysis, investigation, writing – original draft, funding acquisition
MT: software, investigation, writing – review and editing
HTM: conceptualization, methodology, resources, writing – review and editing
MDJ is the guarantor.

Conflicts of Interest

MDJ is a member of the National Health Service Injectable Medicines Guide advisory board and held an honorary position in the editorial team at the start of this project. The other authors have no competing interests.

Multimedia Appendix 1
Large language model prompts and additional data.
[PDF File, 134 KB - jmir_v28i1e81915_app1.pdf]

References

1. Medication without harm - global patient safety challenge on medication safety. : World Health Organization; 2017 URL: https://iris.who.int/server/api/core/bitstreams/15520c4f-89d1-4a8f-9bab-6b771aa30acb/content [accessed 2026-01-29]
2. Naseralallah L, Stewart D, Azfar Ali R, Paudyal V. An umbrella review of systematic reviews on contributory factors to medication errors in health-care settings. Expert Opin Drug Saf 2022 Nov;21(11):1379-1399. [doi: 10.1080/14740338.2022.2147921] [Medline: 36408597]
3. Jones MD, Liu S, Powell F, et al. Exploring the role of guidelines in contributing to medication errors: a descriptive analysis of national patient safety incident data. Drug Saf 2024 Apr;47(4):389-400. [doi: 10.1007/s40264-024-01396-7] [Medline: 38308152]
4. Mahomedradja RF, Schinkel M, Sigaloff KCE, et al. Factors influencing in-hospital prescribing errors: a systematic review. Br J Clin Pharmacol 2023 Jun;89(6):1724-1735. [doi: 10.1111/bcp.15694] [Medline: 36805648]
5. Jones MD, Clarke J, Feather C, et al. Use of pediatric injectable medicines guidelines and associated medication administration errors: a human reliability analysis. Ann Pharmacother 2021 Nov;55(11):1333-1340. [doi: 10.1177/1060028021999647] [Medline: 33641479]
6. Lawton R, McEachan RRC, Giles SJ, Sirriyeh R, Watt IS, Wright J. Development of an evidence-based framework of factors contributing to patient safety incidents in hospital settings: a systematic review. BMJ Qual Saf 2012 May;21(5):369-380. [doi: 10.1136/bmjqs-2011-000443] [Medline: 22421911]
7. Kastner M, Bhattacharyya O, Hayden L, et al. Guideline uptake is influenced by six implementability domains for creating and communicating guidelines: a realist review. J Clin Epidemiol 2015 May;68(5):498-509. [doi: 10.1016/j.jclinepi.2014.12.013] [Medline: 25684154]

XSL•FO
RenderX

8. Raynor DK, Veene PD, Bryant D. The effectiveness of the summary of product characteristics (SmPC) and recommendations for improvement. Ther Innov Regul Sci 2014 Mar;48(2):255-265. [doi: 10.1177/2168479013501311] [Medline: 30227505]
9. Jones MD, Franklin BD, Watson MC, Raynor DK. User testing to improve retrieval and comprehension of information in guidelines to improve medicines safety. J Patient Saf 2022 Jan 1;18(1):e172-e179. [doi: 10.1097/PTS.0000000000000723] [Medline: 32569098]
10. Jones MD, Franklin BD, Raynor DK, Thom H, Watson MC, Kandiyali R. Costs and cost-effectiveness of user-testing of health professionals' guidelines to reduce the frequency of intravenous medicines administration errors by nurses in the United Kingdom: a probabilistic model based on voriconazole administration. Appl Health Econ Health Policy 2022 Jan;20(1):91-104. [doi: 10.1007/s40258-021-00675-z] [Medline: 34403128]
11. Jones MD, McGrogan A, Raynor DK, Watson MC, Franklin BD. User-testing guidelines to improve the safety of intravenous medicines administration: a randomised in situ simulation study. BMJ Qual Saf 2021 Jan;30(1):17-26. [doi: 10.1136/bmjqs-2020-010884] [Medline: 32606212]
12. Theo Raynor DK, Blackwell K, Middleton W. What do writers need to know about user testing? Med Writ 2015 Dec;24(4):215-218. [doi: 10.1179/2047480615Z.000000000326]
13. Moons P, Van Bulck L. Using ChatGPT and Google Bard to improve the readability of written patient information: a proof of concept. Eur J Cardiovasc Nurs 2024 Mar 12;23(2):122-126. [doi: 10.1093/eurjcn/zvad087] [Medline: 37603843]
14. Haver HL, Lin CT, Sirajuddin A, Yi PH, Jeudy J. Use of ChatGPT, GPT-4, and Bard to improve readability of ChatGPT's answers to common questions about lung cancer and lung cancer screening. AJR Am J Roentgenol 2023 Nov;221(5):701-704. [doi: 10.2214/AJR.23.29622] [Medline: 37341179]
15. Salam B, Kravchenko D, Mesropyan N, et al. ChatGPT makes cardiovascular MRI reports easy-to-understand: a feasibility study. J Cardiovasc Magn Reson 2024;26:100881. [doi: 10.1016/j.jocmr.2024.100881]
16. Patel EA, Fleischer L, Filip P, et al. The use of artificial intelligence to improve readability of otolaryngology patient education materials. Otolaryngol Head Neck Surg 2024 Aug;171(2):603-608. [doi: 10.1002/ohn.816] [Medline: 38751109]
17. Kianian R, Sun D, Rojas-Carabali W, Agrawal R, Tsui E. Large language models may help patients understand peer-reviewed scientific articles about ophthalmology: development and usability study. J Med Internet Res 2024 Dec 24;26:e59843. [doi: 10.2196/59843] [Medline: 39719077]
18. Swisher AR, Wu AW, Liu GC, Lee MK, Carle TR, Tang DM. Enhancing health literacy: evaluating the readability of patient handouts revised by ChatGPT's large language model. Otolaryngol Head Neck Surg 2024 Dec;171(6):1751-1757. [doi: 10.1002/ohn.927] [Medline: 39105460]
19. Stoneham AC, Walker LC, Newman MJ, Nicholls A, Avis D. Can artificial intelligence make elective hand clinic letters easier for patients to understand? J Hand Surg Eur Vol 2024 Nov;49(10):1269-1270. [doi: 10.1177/17531934241246479] [Medline: 38641940]
20. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); Jun 2-7, 2019; Minneapolis, Minnesota p. 4171-4186 URL: https://aclanthology.org/N19-1423/ [accessed 2026-01-29] [doi: 10.18653/v1/N19-1423]
21. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med 2025 Jan;31(1):60-69. [doi: 10.1038/s41591-024-03425-5] [Medline: 39779929]
22. Mishra S, Khashabi D, Baral C, Choi Y, Hajishirzi H. Reframing instructional prompts to GPTk's language. 2021 Presented at: Findings of the Association for Computational Linguistics: ACL 2022; May 22-27, 2022 URL: https://aclanthology.org/2022.findings-acl.50/ [accessed 2026-01-29] [doi: 10.18653/v1/2022.findings-acl.50]
23. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. arXiv. Preprint posted online on Apr 21, 2019. [doi: 10.48550/arXiv.1904.09675]
24. BERT score for text generation. GitHub, Inc. 2023. URL: https://github.com/Tiiiger/bert_score [accessed 2026-01-29]
25. Wang LW, Miller MJ, Schmitt MR, Wen FK. Assessing readability formula differences with written health information materials: application, results, and recommendations. Res Social Adm Pharm 2013;9(5):503-516. [doi: 10.1016/j.sapharm.2012.05.009] [Medline: 22835706]
26. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 2008 May;61(Pt 1):29-48. [doi: 10.1348/000711006X126600] [Medline: 18482474]
27. Hanna M, Bojar O. A fine-grained analysis of BERT score. In: Barrault L, Bojar O, Bougares F, Chatterjee R, Costa-jussa MR, Federmann C, editors. Presented at: Proceedings of the Sixth Conference on Machine Translation; Nov 10-11, 2021 p. 507-517 URL: https://aclanthology.org/2021.wmt-1.59/ [accessed 2026-01-29]
28. Verhoeven F, Steehouder MF, Hendrix RMG, Van Gemert-Pijnen JEWC. From expert-driven to user-oriented communication of infection control guidelines. Int J Hum Comput Stud 2010 Jun;68(6):328-343. [doi: 10.1016/j.ijhcs.2009.07.003]
29. Madabushi HT, Jones MD. Large language models in healthcare information research: making progress in an emerging field. BMJ Qual Saf 2025 Jan 28;34(2):73-76. [doi: 10.1136/bmjqs-2024-017896] [Medline: 39443104]
30. Jones MD, Torgbi M, Tayyar Madabushi H. Large language models to improve the understandability of clinical guidelines: an evaluation of readability improvements and unintended content changes produced by GPT-4. GitHub. 2026. URL: https://github.com/melissatorgbi/LLM-Clinical-Guideline-Understandability [accessed 2026-02-19]

XSL•FO
RenderX

Abbreviations

BERT: bidirectional encoder representations from transformers
IMG: Injectable Medicines Guide
LLM: large language model
SMOG: Simple Measure of Gobbledygook
TRIPOD-LLM: Transparent Reporting of a Multivariable Model for Individual Prognosis or Diagnosis - Large Language Model

Edited by A Coristine; submitted 05.Aug.2025; peer-reviewed by C Gissel, J Mira; revised version received 21.Jan.2026; accepted 22.Jan.2026; published 23.Feb.2026.

Please cite as:
Jones MD, Torgbi M, Tayyar Madabushi H
Improving the Understandability of Clinical Guidelines: Development and Evaluation of a GPT-4–Based Pipeline
J Med Internet Res 2026;28:e81915
URL: https://www.jmir.org/2026/1/e81915
doi:10.2196/81915

© Matthew D Jones, Melissa Torgbi, Harish Tayyar Madabushi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Interorganizational Mechanisms for Developing and Implementing Clinical Decision Support Systems in Primary Care: Exploratory, Qualitative Case Study

Jesse J M M Santema[1], MSc; Jeroen D H van Wijngaarden[2], MSc, PhD; Eric G Hiddink[3,4], MSc, PharmD; Fleur Deken[5], MSc, PhD; Maaike Kleinsmann[1,6], MSc, PhD; Hendrikus J A van Os[3,7], MD, PhD

[1]Industrial Design Engineering, Delft University of Technology, Mekelweg 5, Delft, South Holland, The Netherlands
[2]Health Services Management & Organization, Erasmus School of Health Policy and Management, Rotterdam, South Holland, The Netherlands
[3]National eHealth Living Lab, Department of Public Health & Primary Care, Leiden University Medical Center, Leiden, South Holland, The Netherlands
[4]Health Base Foundation, Houten, Utrecht, The Netherlands
[5]School of Business and Economics, Knowledge, Information and Innovation, Vrije Universiteit Amsterdam, Amsterdam, North Holland, The Netherlands
[6]Department of Cardiology, Leiden University Medical Center, Leiden, South Holland, The Netherlands
[7]Department of Psychiatry, University Medical Center Utrecht, Utrecht, Utrecht, The Netherlands

Corresponding Author:
Jesse J M M Santema, MSc
Industrial Design Engineering, Delft University of Technology, Mekelweg 5, Delft, South Holland, The Netherlands

Abstract

Background: Clinical decision support systems (CDSS) have the potential to improve patient safety and reduce costs in primary care. However, CDSS adoption remains limited due to development and implementation challenges. CDSSs are complex interventions involving multiple interacting components that require technological innovation and behavioral and organizational change. Additionally, the primary care context is considered a complex system with high care demand, fragmented structures, and many independent yet interdependent organizations. Established determinant frameworks for implementing and scaling up complex health care interventions support the identification of implementation determinants. However, they offer limited guidance on the underlying processes of these determinants, such as the implementation processes involved in complex interorganizational collaboration in primary care.

Objective: This study examined how an interorganizational collaboration in Dutch primary care (*Gezonde zorg, Gezonde regio* [*GzGr*]) achieved an iterative CDSS development and implementation. We aimed to identify the mechanisms that supported the collaboration in overcoming challenges.

Methods: We performed an exploratory process-level case study. Data were collected through 15 semistructured interviews. The nonadoption, abandonment, scale-up, spread, and sustainability framework was used to ensure comprehensive topic coverage during the interviews, but not as an analytical framework. We triangulated the interviews with internal and external documents and expert input. Using a thematic, inductive approach, we developed a chronological overview of the collaboration and identified mechanisms offering insights into how GzGr navigated complexity in the development and implementation of CDSS.

Results: We identified two mechanisms: (1) enacting an interorganizational value model and (2) iterative, co-creative experimentation. First, GzGr was driven by a coalition of the willing (ie, individuals willing to take an extra step), with shared goals that prioritized collective benefit while respecting organizational values. They established shared principles that translated the broad GzGr mission into concrete CDSS development choices, while also guiding strategic expansion by involving mission-aligned, innovative organizations. Second, after initial prototypes, GzGr established an iterative learning and improvement experimentation for both the technology and the collaboration. This process allowed for rapid feedback, validation of added value, and ongoing refinement. Additionally, this experimentation approached the development and implementation phase as a continuous process involving multistakeholders, supporting both the technology and the collaboration.

Conclusions: This study identified 2 mechanisms that sustained interorganizational collaboration and CDSS development. These mechanisms connected collaborative and technical changes across people, technology, and organizational levels, enabling technological viability, stakeholder value, and multilevel support. The mechanisms operated both within and between organizations through iterative cycles of development and implementation. Practical implications include involving multilevel, innovative, and influential stakeholders; maintaining alignment through an orchestrating actor; and adopting an iterative approach between development and implementation. Our findings extend existing determinant frameworks by offering process-level insights into

XSL•FO
RenderX

how such mechanisms help overcome challenges in the development and implementation of CDSS within interorganizational collaborations.

(J Med Internet Res 2026;28:e83084) doi:10.2196/83084

KEYWORDS

clinical decision support systems; interorganizational collaboration; mechanisms of change; multilevel stakeholder involvement; integration of people; integration of technology; integration of business; iterative development and implementation; process-level analysis

Introduction

Clinical decision support systems (CDSS) hold the potential to improve patient safety, diagnostic accuracy, and clinical decision-making and reduce health care costs [1]. CDSSs achieve these aims by using rule-based or artificial intelligence (AI) algorithms to deliver patient-specific alerts, reminders, and dosage support, especially in primary care [2-4]. Despite its potential, CDSS adoption remains limited due to interdependent development and implementation challenges [1,5-7], particularly in primary care [2,8,9]. These challenges partially stem from the complexity of CDSS, as they involve multiple interacting components, often necessitate behavioral changes from end users, and affect clinical and economic outcomes [10,11]. The complexity of CDSS also results in interconnections between users and their social and organizational context.

Traditional implementation approaches assess interventions under controlled conditions before applying them in practice [10,12-15]. This approach can lead to unintended consequences during real-world implementation due to the intervention's complexity [14]. Therefore, digital health and innovation scholars advocate for real-world implementation with iterative cycles consisting of development and implementation that proceed [10,12-14,16-18]. These iterative cycles allow contextual adaptation and gradual refinement, which improves implementation outcomes and mitigates unintended consequences [13,14]. Examples of models for an iterative process are the Centre for eHealth and Wellbeing Research roadmap, UK Medical Research Council guidelines, and the nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) framework [10,12,13,18]. These models provide valuable insights into determinants for the implementation of digital health interventions in single organizations (while recognizing the wider contextual factors influencing implementation). However, these models pay less attention to the interactions among determinants and processes underlying these determinants [14,19].

Insights into these underlying processes are especially relevant in the context of CDSS in primary care, where both the intervention and the setting are inherently complex. Primary care in the Netherlands involves many independent yet interdependent organizations, such as general practices, care groups, payers, and technology developers [2,14,20,21]. These stakeholders differ in roles, interests, and values owing to separate financing structures, although they must collaborate across organizational boundaries to ensure continuity and accessibility of care. This results in a highly fragmented and complex care landscape. Stakeholders and their interactions are illustrated in Figure 1. Research suggests that a context with multiple independent and interdependent organizations calls for interorganizational collaboration to enable iterative development and implementation [11,16]. Such collaboration aims for coordinated efforts between organizations working toward a shared goal [22,23]. However, these studies do not elaborate on how an interorganizational collaboration can be established in practice.

XSL•FO
RenderX

Figure 1. Simplified overview of 6 independent organizations in Dutch primary care and their connections. The connecting lines highlight complexity. Every organization must interact with every other organization for the continuity and accessibility of care. GP: general practitioner.

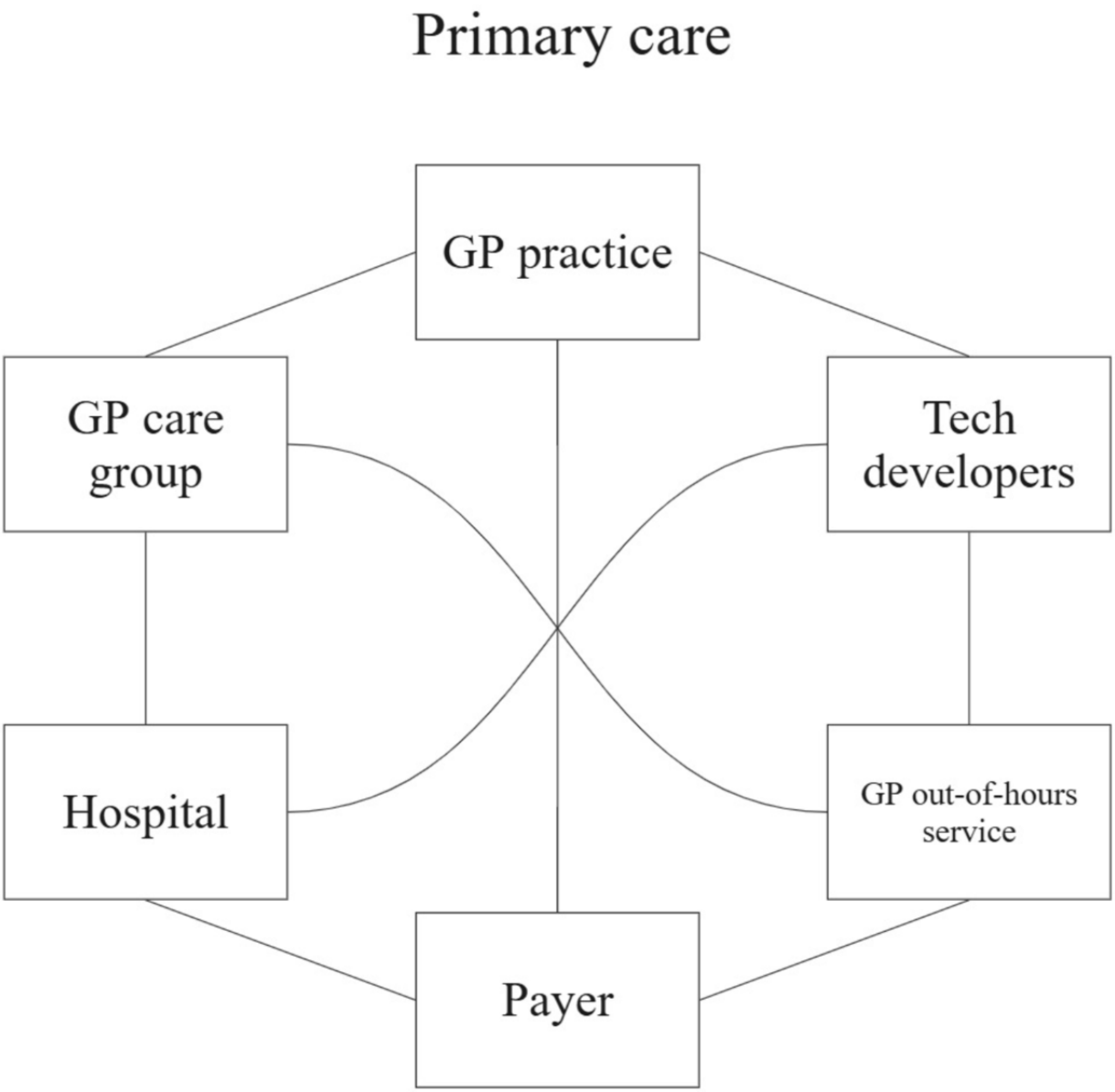

We built on prior determinant research conducted in single-organization implementation contexts [10,12,13,18], which emphasized the need for an iterative CDSS development and implementation process. Our study extended this line of research by examining how challenges encountered during development and implementation were addressed in an interorganizational primary care collaboration. Specifically, we focused on identifying mechanisms that explain the underlying processes through which interventions are developed and implemented [24]. This focus is particularly relevant in complex settings such as primary care. Our research question was as follows: What mechanisms facilitate overcoming challenges in the development and implementation of CDSS within an interorganizational collaboration in primary care? We aimed to identify mechanisms that enabled CDSS development and implementation within an interorganizational collaboration in primary care. We conducted a qualitative process analysis of a Dutch primary care collaboration, allowing us to focus on how challenges and actions developed over time and formed mechanisms.

Methods

Study Design

We performed an exploratory case study to identify which mechanisms enabled *Gezonde zorg, Gezonde regio* (GzGr) to successfully develop and implement CDSS in Dutch primary care [25]. We define mechanisms in the context of this paper as sequential patterns of actions that explain how change occurs, both between and within organizations in relation to technological interventions. These actions represent how GzGr created the conditions and behaviors needed to cope with challenges that occurred over time in the development and implementation of complex digital health solutions in complex

XSL•FO
RenderX

settings. The sequential aspect emphasizes how actions build on previous actions or challenges over time, indicating that their order may be causally related to their final effect. To identify how, why, and when mechanisms were applied, we used qualitative process analysis by tracing actions and decisions over time [26-28]. While the mechanisms captured what drove change, our accompanying narrative illustrated how these processes unfolded in practice. This study was designed and reported in line with the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist (Checklist 1) [29].

Case Study Selection

We researched "Healthy care, Healthy region" (in Dutch: *Gezonde zorg, Gezonde regio* [*GzGr*]). We identified GzGr as a unique case study via field exploration through researchers' networks and via internet searches. Case selection was based on the following criteria: (1) successful regional implementation of a CDSS across multiple primary care practices, (2) full integration in primary care workflow, including IT integration; and (3) collaboration involving all relevant stakeholders in primary care.

Case Description

The GzGr initiative involves a diverse range of organizations: general practitioner (GP) practices, regional primary care organizations, a regional hospital, a health care insurance firm, data processing firms, technology and content development firms (collectively referred to as technology partners), and advisory groups responsible for implementation support and technology evaluation [30]. Figure 2 provides a visual representation of the GzGr structure. GzGr developed multiple rule-based and data-driven CDSS to provide real-time alerts to GPs regarding patient health risks [31]. The steering committee, consisting of executives from key partner organizations, sets the strategic direction and oversees key decisions. The working group, consisting of health care purchasers and innovation managers from the care groups, manages day-to-day operations and the collaboration and advises the steering committee. The editorial committee consists of end users, technology developers, and representatives from the health care insurance firm and is responsible for identifying widely experienced problems and overseeing CDSS development.

Figure 2. Visual representation of the GzGr structure, with health care and supporting organizations (represented by the horizontal squares). The vertical squares represent the governance structure. GzGr: *Gezonde zorg, Gezonde regio*.

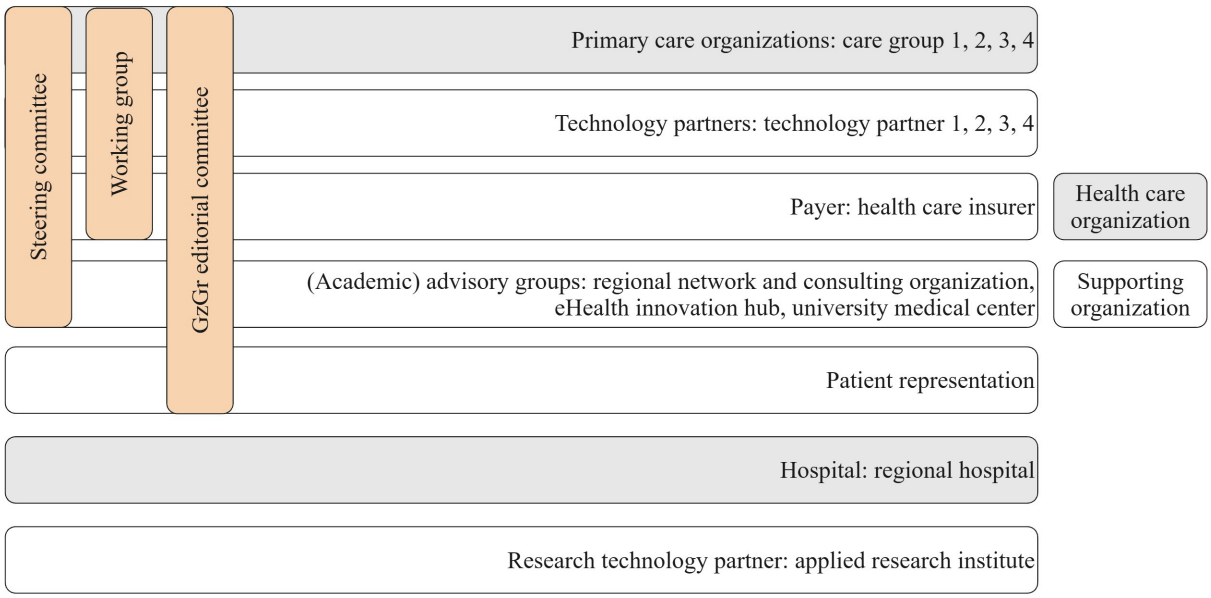

Data Collection

We collected multiple sources of data, including interviews and documents, aiming for data triangulation to increase the validity of the data [32]. In total, we recruited and interviewed 15 unique respondents from December 2021 to September 2022 by JDHvW. Interviews were conducted until data saturation was reached, meaning that interviews confirmed findings rather than yielded new findings [33]. Data saturation determined our study size. We interviewed at least one participant from each partner organization involved in GzGr. An overview of participants is given in Multimedia Appendix 1. Our eligibility criteria for participants were their (1) active involvement in the operational development of the CDSS, by working at one of the technology partners, at a general practice or a care group (defined as an organization that coordinates and supports GPs with administrative burden); and/or (2) involvement in the strategic decision-making of GzGr, either through participation in the steering committee or by holding an executive mandate within an organization.

We sampled the respondents using snowball sampling and purposive sampling and used the networks of participants, partner organizations of GzGr, and researchers [34]. JDHvW contacted participants by email through a GzGr contact person, after which additional participants were identified by previously interviewed respondents. Although this approach introduced potential selection bias at first, snowball sampling helped recruit a diverse and representative group. No participants declined participation.

XSL•FO
RenderX

We followed a semistructured interview approach to examine which mechanisms were used by GzGr to ensure progress in development and implementation. We composed an interview topic guide based on the NASSS framework [12]. Topics were based on the following domains: intervention features, the expected value, adopter readiness, the broader context, adaptability over time, and the interplay among domains. We omitted the "condition" domain, as the developed CDSS were designed to apply across multiple clinical conditions. Furthermore, we asked about involved stakeholders, organizational characteristics, and the development process. We explored how GzGr addressed related challenges in these domains to identify underlying mechanisms, which formed the basis for identifying mechanisms.

Interviews were conducted via video call. Each interview lasted between 45 and 60 minutes, and the respondent(s) and JDHvW were present. The interviews were audio recorded and afterward transcribed verbatim. Transcripts were made available to respondents for review upon request. Moreover, we collected internal and external documents of the organizations via the participants and the internet, from September 2023 to January 2024. All documents were stored in Atlas.ti (ATLAS.ti Scientific Software Development GmbH) for analysis.

Data Analysis

To identify mechanisms that facilitated overcoming challenges in the development and implementation of CDSS, our data analysis consisted of multiple steps, namely (1) composing a chronological overview; (2) thematic and inductive coding to identify mechanisms; (3) performing a member check; and (4) constructing a conceptualized narrative. We used Atlas.ti 9 to support the coding process. Further elaboration on the coding process and the codebook is given in Multimedia Appendix 2.

1. We created a chronological overview of challenges and actions related to development, implementation, or collaboration encountered or undertaken by GzGr to provide deeper insight into how the stakeholders developed and implemented CDSS. The chronological overview is given in Multimedia Appendix 3.
2. Next, we conducted an in-depth analysis to identify mechanisms, consisting of patterns of actions taken by GzGr. We used thematic analysis and inductive coding to reveal new insights and underlying patterns in the data [35,36]. We performed a second round of open coding of the transcripts. During open coding, codes that had been generated in the earlier transcripts were applicable to later interviews. At that stage, no new codes were introduced, which indicated that we reached code saturation. These codes were used for composing a codebook (see Multimedia Appendix 2) and for coding all the transcripts. Then, we identified 2 mechanisms, by linking challenges and determining the sequential patterns of actions that, according to participants, had led to the successful development and implementation of CDSS. Mechanism 1, enacting an interorganizational value model, is defined as aligning independent yet interdependent organizational goals and values into regional goals and values and translating them into practical actions. The mechanism 2, iterative, co-creative experimentation, is defined as continuously, jointly developing and adapting solutions with stakeholders, treating development and implementation as one interconnected, ongoing process. In Multimedia Appendix 3, we further elaborated on the link between challenges and actions and how the mechanisms emerged or evolved over time.
3. We performed member checks with the steering committee of GzGr to reach consensus on the mechanisms deployed. Additionally, EGH and HJAvO, who participated in the GzGr collaboration as (scientific) experts, supported the interpretation of qualitative data.
4. The final step in the data analysis was constructing a narrative of the challenges encountered and the actions taken by GzGr, supported by quotations of the participants. At the end of every period, we conceptualized the sequential patterns of actions that formed a mechanism. Furthermore, we bundled the sequential patterns of actions per mechanism in Table 1. In the discussion, we compared these findings with prior work.

Table . Mechanisms consisting of sequential patterns of actions to develop and implement complex digital health solutions in complex settings.

Mechanism	Sequential processes
Enacting an interorganizational value model	• Forming a coalition of the willing • Defining and aligning on principles for organization selection and development early in the process (ownership, pricing, co-creation, transparency, guidelines, embedding in workflows) • Streamline decision-making and maintain focus during development due to aligned principles
Iterative, co-creative experimentation	• Co-creation in structured feedback loops between end users, developers, and decision-makers • Taking ownership of communication by the collaboration • Technological adjustments combined with measurable scientific outcomes

Ethical Considerations

This study was conducted in accordance with ethical guidelines for qualitative research involving human participants. Ethical approval was obtained from the Research Ethics Review Committee of the Erasmus School of Health Policy & Management (reference number 21‐014), ensuring compliance

XSL•FO
RenderX

with relevant regulations on data protection and participant rights.

Additionally, participants signed an informed consent form before participating in interviews, which stated the goal of the research, the researchers, and the interview; privacy measures (including pseudonymizing data); and confidentiality measures (raw data available upon reasonable request) and included consent for audio recording. No identifying information or images of individual participants are included in this manuscript or its supplementary materials. Participants did not receive any compensation.

Results

Overview

In this section, we described the chronological sequence of the development and implementation challenges and actions. Table 2 summarizes the main challenges faced by GzGr and the resulting actions that led to the development and implementation of CDSS. In retrospect, we distinguished 3 periods, each with distinct challenges and focus. The chronological overview in Multimedia Appendix 3 is a more elaborated version of these challenges, actions, and their connection to the mechanism and provides a brief overview of the timeline.

XSL•FO
RenderX

Table . Summary overview of challenges and actions across 3 periods that contributed to enacting an interorganizational value model and to applying iterative, co-creative experimentation.

Period, challenge, and actions	Mechanism leveraged
2010 - 2014: forming a coalition of the willing	
Staff shortages within GzGr[a]	
Started collaboration with influential individuals	1
Exploring technological needs, unexpected outcomes, and collaboration stopped	
Research institute measured the baseline population funding	1
Engaged, innovative alignment is difficult	
Composing coalition of the willing and aligning principles for collaboration expansion	1
Engaged, innovative alignment is difficult	
Technology partner 1 and regional network and consulting organization joined the collaboration	1
Mission changed: enhancing preventive care by developing CDSS[b]	1
Technology partners 2 and 3 joined the organization based on aligned principles	1
First pilot CDSS development	1
Not all organizations were innovative	
Shift in focus to involving innovative GP[c] practices and organizations	1
2015 - 2017: develop understandable CDSS	
Engaged, innovative alignment is difficult	
Establishing CDSS development principles	1
Building understandable CDSS	1
Lack of workflow integration, PDF notifications for CDSS	
Technology partner 4 joined, ensuring integration of CDSS in GP information system	1
Six knowledge rules deployed, five in development	1, 2
2018 - 2021: Implement CDSS in daily practice	
Use of CDSS low	
More focus on implementation	2
Communication responsibility of individual care groups	
GzGr improved information material, developed training, and informed community	2
Focus shift to implementation	
Established editorial committee for identifying implementation problems	2
Focus shift to implementation	
Implementation staff assigned, improving the digital skills of health care professionals	2
Improve CDSS	
Testing CDSS prototypes and retrieving feedback from GPs	1, 2
Developed dashboard with insights into the use of CDSS	2
Strategic organizational dialogues about opportunities for individual organizations and GzGr	1, 2

[a]GzGr: *Gezonde zorg, Gezonde regio* (Healthy care, Healthy region).

[b]CDSS: clinical decision support system.

XSL•FO
RenderX

[c]GP: general practitioner.

Period 1 (2010-2014): Forming a Coalition of the Willing

In this first period, we explain how GzGr started their collaboration with a mission, formed a mission-driven coalition of the willing, and established principles for expanding the collaboration.

The GzGr collaboration was initiated by 3 senior influential individuals in health care. One was a senior executive at the largest regional health care insurer, securing long-term funding. The other two were executives from a primary care group and a regional hospital, providing managerial commitment. These individuals were considered influential in the region because they could allocate resources, influence organizational decisions, and mobilize support (financial and people). The aim was to provide more efficient care to maintain regional health care accessibility, with a focus on prevention, the use of routine data, and in line with shared savings funding models. In their view, delivering future-proof care required close collaboration between care organizations and funders to foster joint responsibility for a specific population. As an initial attempt toward operationalizing this mission, a research-focused technology partner explored the feasibility of population-based funding. However, it became evident that this organization lacked expertise in operationalizing the outcomes, posing a challenge for GzGr, leading GzGr to stop the collaboration.

Realizing the long-term mission to aim for more efficient care and involving partnering organizations was also experienced as challenging for GzGr; therefore, they decided to refocus their efforts on a more short-term, actionable mission: enhancing preventive care interventions in GP practices by developing and implementing CDSS. This system would leverage routine care data to support GPs in patient monitoring and the early detection of health risks.

To realize the development, GzGr invited 2 new individuals from organizations to join the collaboration: an innovation manager working at technology partner 1, contributing to clinical content and CDSS development, and a project manager from the regional network and consulting organization. According to respondents, the addition of these 2 organizations was key in translating GzGr's broad mission into an actionable, concrete CDSS. Together, these 5 individuals brought extensive professional networks, complementary expertise, and executive backing from their organizations:

> *Our success lies in the combination of having people who deeply understand the data and work with it daily, others who know the primary process well, a few executives who are genuinely enthusiastic about this, and a healthcare insurer that recognizes its importance.* [Professor of data science, technology partner 1]

More importantly, respondents emphasized that these individuals consistently prioritized the shared, collaborative mission. At the same time, they ensured alignment with their own organizational interests, without aiming to maximize individual organizational gain. GzGr referred to the individuals as a "coalition of the willing," marking the emergence of a mission-driven collaboration within GzGr:

> *Based on knowing a few enthusiastic people, being aware of the expertise available, having a shared goal, and the motivation to pursue it: we just went ahead and did it. We developed a proof of concept and demonstrated that it works. Even stakeholders said: This will accelerate things.* [Policy adviser, national health care insurance umbrella organization]

To make progress, GzGr recognized the need for a data infrastructure to support CDSS implementation across GP practices and enable data integration and evaluation. This meant that additional technology partners were needed. Identifying technology partners that aligned with GzGr's mission and collaborative approach proved challenging. Therefore, GzGr formed principles for organization selection and task allocation: willingness to experiment, previous experience in CDSS development, knowledge of legal and regulatory standards, possession of appropriate certifications, a willingness to work at a cost-plus model (ie, a pricing approach where providers are reimbursed based on actual costs plus a fixed margin), and agreeing upon the shared mission. These principles ensured that new participating organizations aligned with GzGr's chosen direction, allowing flexibility in how the mission was executed but preserving its overall purpose, which prevented mission drift and delay in progress.

Through technology partner 1's network, technology partners 2 and 3 joined. Together, GzGr began developing CDSS that could be implemented at scale. One of the first successful pilots was a data-driven CDSS for cardiovascular pharmacotherapy, identifying cost-effective treatment options with equal or better outcomes. Savings were shared between providers and insurers, illustrating an early shared savings model. Figure 3 provides a visual overview of the contributing organizations (squares), their roles, interaction (arrows), and an example of CDSS (circle). Data from multiple national registries are integrated by technology partners (double-headed arrow) and linked to GP electronic health records and the GP (single-headed arrow), where CDSS are embedded to support clinical decisions. Finally, CDSS are evaluated with insurers and academic partners. The figure also includes an example of a cardiovascular risk management rule, showing how data generated by GPs trigger actions for patient follow-up, supported by continuous evaluation and feedback between organizations.

XSL•FO
RenderX

Figure 3. Organizations involved in developing and implementing CDSS within GzGr, showing data integration, embedding in GP records, and continuous evaluation. The squares represent participating organizations, the circles represent an example of CDSS, and the arrows represent the data infrastructure. CDSS: clinical decision support system; CVRM: cardiovascular risk management; EHR: electronic health record; GP: general practitioner.

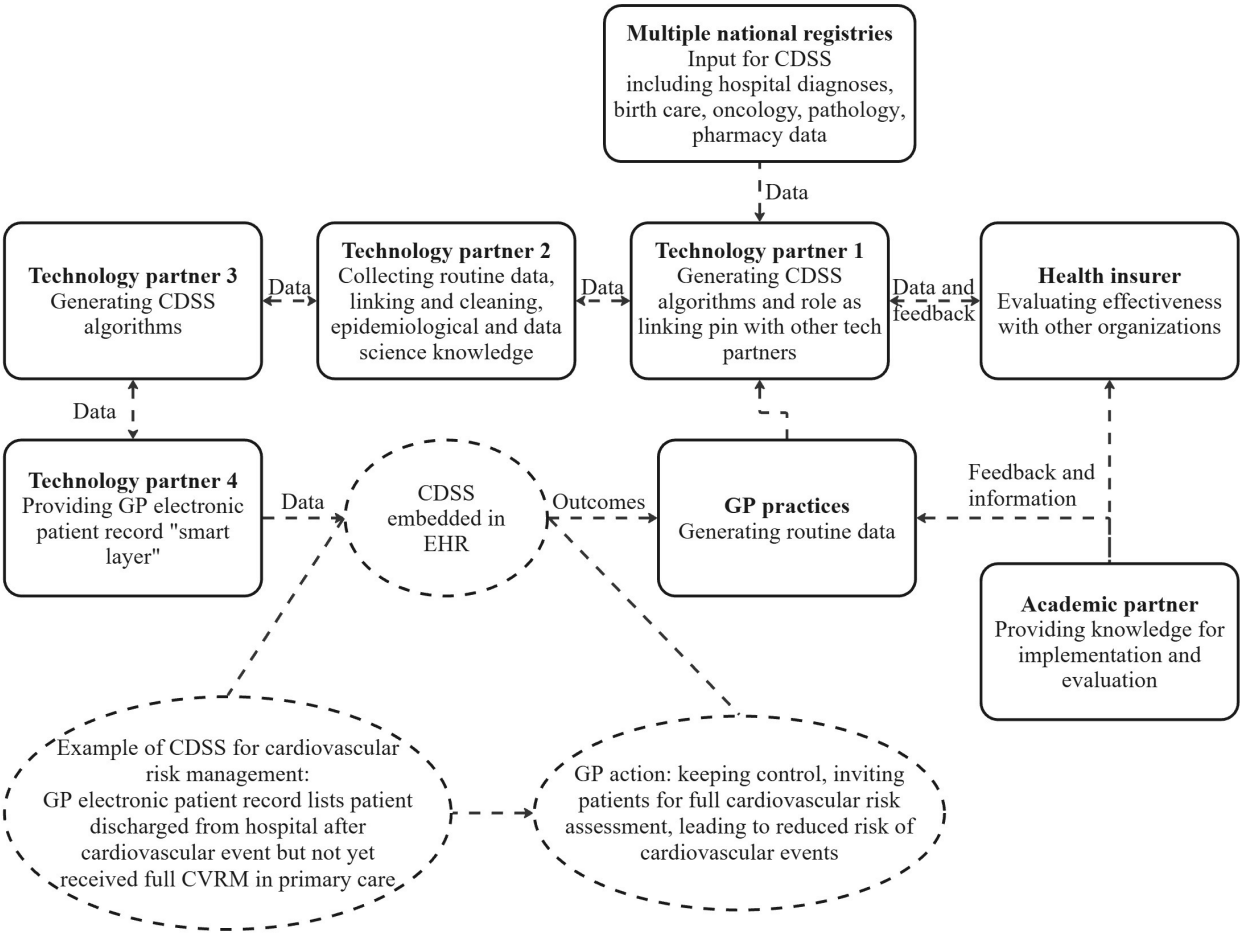

From the outset, CDSS and infrastructure development were co-funded by health care insurers and technology partners, who believed in the potential of CDSS and reimbursed GPs for using CDSS, even in the absence of traditional scientific validation. Gaining internal support within the insurer required ongoing efforts from the individual involved in GzGr from the start, who actively used the shared mission to convince colleagues of the value of GzGr:

> *We have invited a colleague working at the healthcare insurer, because internal support [for GzGr at the insurer] was needed to keep justifying, once again, why funding should continue.* [Strategic adviser, regional network, and consulting organization]

To increase the impact of the CDSS, GzGr sought to expand the collaboration with additional GP care groups in the region. However, not all care groups were willing to experiment and change their internal processes to accommodate the CDSS. Therefore, GzGr decided to prioritize GP practices with an "innovative mindset" and willingness to experiment. Additionally, not all collaborations within GzGr endured. A regional hospital withdrew after the departure of the executive initiating GzGr, which respondents linked to a misalignment between individual commitment and organizational priorities.

To conclude, GzGr operationalized their broad mission to enhance preventative care by developing and implementing CDSS. The first mechanism was used: enacting an interorganizational value model, which built on shared value and accountability. We identified two sequential patterns of actions, forming this mechanism: (1) forming a coalition of the willing that shared a common mission, combined their expertise, and mobilized resources; and (2) defining principles for organization selection to ensure mutual value, emphasizing innovation, compliance, cost-plus funding model, and mission alignment. Table 1 presents the sequential patterns of actions that form the mechanism.

Period 2 (2015-2017): Develop Understandable Algorithms

In the second period, GzGr further developed the interorganizational value model by aligning on development principles, preventing disagreements during or after the CDSS development. Management, together with end users, developed these principles to ensure broad implementation, feasibility, and usability in multiple levels within the organizations. The principles for developing CDSS were as follows.

First, agreement was needed on the pricing structure and ownership of CDSS. For GPs and the insurer, it was not

XSL•FO
RenderX

acceptable to invest their time and share sensitive data, while the technology partners would make high profits, own CDSS, and if prices would fluctuate. At the same time, the technology partners needed to work for profit to survive. Eventually, it was agreed that the technology partners were allowed to make a small profit based on a cost-plus price model and that CDSS would be free of property rights:

> *They [technology partners] need to make profit otherwise the company will not survive. But we discussed we do not want value-based pricing [assessing the price based on the costs the technology saves], but a cost-plus model...so we can set a fixed price per patient and make it more attractive for the healthcare insurers because the prices are predictable.* [Innovation manager, technology partner 3]

Second, GzGr fundamentally chose co-creation as the basis for CDSS development, ensuring that CDSS were driven by clinical problems that were prioritized by both GPs, practice nurses, and patients. While clinicians identified problems in practice related to delivering care, technology developers explored what was technically feasible, and funders assessed financial viability. This ensured the integration of the requirements from people, technology, and business. Initially, this human-centered, problem-driven approach challenged technology partners, who were used to a supplier-client model with predefined requests. Now, they had to share responsibility with end users in identifying clinical problems, requiring them to adapt their development processes:

> *We apply a problem-driven principle, and because we develop together, you can feel a real drive to actually use it [CDSS] and make the most of it.* [Director primary care, technology partner 4]

Third, the GzGr organizations agreed that transparency of CDSS and its development was key. CDSS' decision paths (even for AI, using logistic regression) could be followed by GPs, and the functionality was made explainable, allowing GPs to verify accuracy and build confidence in the use of CDSS. This transparency supports GPs in meeting legal requirements that state that health care professionals must be able to explain clinical decisions.

Fourth, aligning with legal, regulatory, and clinical guidelines was essential for GzGr. This meant that GPs retained control over clinical decision-making, with the ability to override CDSS recommendations. Although aligning with guidelines was costly, it ensured quality, scalability, and support from GPs:

> *I am pleased with...legal and regulatory guidelines, what might sound paradoxical, but it forces you to demonstrate how your algorithm works,...leading to transparency.* [Innovation manager, technology partner 3]

Fifth, it was agreed that seamless workflow integration was essential by embedding CDSS in GP electronic health records to avoid system switching. The coalition of the willing convinced the largest GP information system developer (technology partner 4) to join, which made it possible that CDSSs could be integrated within their systems. Respondents noted that achieving technical integration was a significant but successfully managed challenge, as they had experienced how a nonintegrated system was rarely used:

> *The initial PDFs were barely used, people had to open a PDF, log into a new system again. This just did not work in practice.* [Account manager, health care insurer]

While following these 5 principles, 7 CDSS were developed in the GP landscape and were used by 160 GP practices, reaching up to 550,000 patients. An additional 5 CDSS were under development and were tested in 1 to 5 practices, with a reach of 10,000 to 25,000 patients (in 2022; Textbox 1). The development of CDSS started in 2014 and is ongoing.

XSL•FO
RenderX

Textbox 1. Overview of the 7 developed and 5 clinical decision support systems (CDSS) under development in 2022, based on the aligned development principles, including adoption metrics (number of general practitioner [GP] practices and patients).

Developed by *Gezonde zorg, Gezonde regio*

- CDSS applications
- Patients at risk of being unserved in cardiovascular risk management.
- Referral protocol for patients with chest pain.
- Feedback on chest pain referral practices.
- Referral protocol for peripheral arterial disease (PAD).
- Peripheral arterial disease: case identification and follow-up.
- Feedback on referral behavior in PAD cases.
- Assessment of indication for goals-of-care conversations in frail older adults.
- Adoption metrics
- Developed CDSS were implemented in 160 GP practices, reaching approximately 550,000 patients.

Development by GzGr

- CDSS applications
- Patients with chronic obstructive pulmonary disease potentially overtreated with inhaled corticosteroids.
- Patients on direct oral anticoagulants/non–vitamin K antagonist oral anticoagulants with chronic kidney disease (chronic kidney disease epidemiology collaboration) for over 12 months.
- Ensuring appropriate laboratory monitoring frequency for patients on central nervous system medications.
- Treatment adjustments based on a significant decline in kidney function.
- Identification of lithium users at high risk of toxicity.
- Adoption metrics
- CDSS under development were implemented in 1 to 5 practices, reaching between 10,000 and 25,000 patients.

In conclusion, the organizations concretized the interorganizational value model by aligning early on developing principles (the sequential process): (1) agreements on ownership and pricing, (2) co-creation, (3) transparency and guideline alignment, and (4) embedding CDSS in daily workflows. These principles translated the coalition's mission into daily practice, helping organizations navigate conflicts, build trust, and coordinate decisions while preserving their individual and regional value. Additionally, the mechanism of iterative, co-creative experimentation emerged with the first prototypes, driven by structured feedback loops between end users, developers, and decision-makers (the sequential process). The sequential patterns of actions that form these mechanisms are presented in Table 1.

Period 3 (2018-2021): Implement CDSS in Daily Practice

After the first CDSS was developed, it became clear that the use of CDSS by GPs was lower than initially expected. In response, the focus shifted to integrating CDSS into daily practice and optimizing use, learning from previous experience, and iteratively adapting the strategy, with the same mission.

Initially, during implementation, primary care groups managed communication within their own organizations, but this proved challenging and ineffective. As a result, GzGr started developing materials, webinars, and introduced on-site implementation staff to strengthen end user skills and integration into workflows. Identifying relevant problems for CDSS also remained difficult, leading to the formation of an editorial committee, comprising technology partners, the insurer, and innovative GPs and assistants. They identified widely experienced implementation problems, composed solutions and an implementation plan, and promoted CDSS use among peers.

As the health care insurer and technology partners had substantially invested in the development and implementation, they wanted a guarantee that GPs used CDSS. Therefore, implementation outcomes (eg, CDSS usage rate and adoption of regional GP practices) were developed and agreed upon, obliging the GzGr organizations to act when algorithm usage rates dropped below a certain threshold:

> *It's no longer optional to use the algorithms [CDSS]. We have invested significantly, and the strategic goals must be met. While some front runners are making real progress, others still hesitate. That is why we are now fully focusing on implementation,...healthcare organizations must take responsibility for internal adoption.* [Account manager, health care insurer]

XSL•FO
RenderX

Additionally, in this period, practice-oriented academic partners joined GzGr to facilitate research, testing, and evaluation, without losing momentum and to ensure a pragmatic process. The results demonstrated that CDSS contributed to time efficiency, improved care quality, and increased adoption by bringing a scientific perspective. The steering committee assessed implementation outcomes across people, technology, and business to ensure stakeholder value. Values were not only monetary; namely, for a GP, saving time was essential, and the health care insurer had to fulfill their duty of care.

GzGr used the implementation and scientific outcomes and user feedback for a continuous iterative learning and improvement experimentation process to refine both the CDSS and the collaboration. The process was as follows: every first version of CDSS is tested in 1 or 2 GP practices (serving between 2000 and 5000 patients). The first tests were done in 5 to 10 GP practices (serving between 10,000 and 25,000 patients). When a CDSS was approved by GzGr, it was implemented in all practices (serving more than 500,000 patients). This process led to adaptations in both the technology and the mission-driven collaboration. Early feedback enabled timely adjustments, strengthened collaboration, and increased support among end users. Respondents noted that this iterative process became a functioning "engine for innovation" within and between the development and implementation phase: creating the opportunity and confidence to develop other technologies and expand the collaboration with a hospital and more care groups.

To conclude, the mechanism "iterative co-creative experimentation" further developed as GzGr shifted from developing technology to embedding CDSS in daily practice. The sequential patterns of actions were (1) co-creation in structured feedback loops, ensuring that people, technology, and organizational perspectives directly informed further development; (2) GzGr took ownership of communication; and (3) technological adjustments were tied to measurable, scientific outcomes. Together, these sequential patterns of actions strengthened shared commitment to GzGr's mission, while the interorganizational value model streamlined decision-making and maintained focus during iterative development and implementation cycles. Table 1 presents the sequential patterns of actions that shape the mechanisms.

Discussion

Principal Study Findings

We conducted a qualitative process analysis of our exploratory case study data to identify which mechanisms helped overcome challenges in the development and implementation of CDSS within GzGr, an interorganizational collaboration. This analysis shed light on sequential patterns of actions by identifying the mechanisms through which challenges were addressed across 3 subsequent periods. We identified 2 mechanisms: (1) enacting an interorganizational value model and (2) iterative, co-creative experimentation. The 3 periods show that the mechanisms reinforced each other and how the underlying actions built on each other over time. Iterative, co-creative experimentation refined the interorganizational value model. In turn, this shared value model made iterative experimentation more focused and effective.

Comparison With Prior Work

We elaborate on how these mechanisms that explain CDSS can be developed and implemented in an interorganizational context and compare them to previous research.

Enacting an Interorganizational Value Model for Mission-Driven Interorganizational Collaboration and CDSS Development

One of the key perceived challenges by GzGr in developing and implementing CDSS across multiple organizations was the fragmentation of responsibilities, interests, and values. This reflects the complexity of the setting, characterized by many independent yet interdependent organizations. GzGr addressed the setting's complexity by forming an interorganizational value model. Underpinning this value model was a mission-driven coalition of the willing. This coalition enabled GzGr to move beyond organizational silos and foster collaboration by spreading the shared mission and value across individual organizations. Earlier research shows that collaborative performance tends to be strongest when a closely connected group with clear leadership and a shared purpose works together, supported by wider networks that enable collaboration across boundaries [37]. Additionally, GzGr showed that early alignment on shared principles is key to addressing implementation challenges. Agreements about ownership and pricing, transparency, and traceability of CDSS, and embedding them into daily workflow are earlier mentioned in single-organizational literature as facilitating the development and implementation of digital health interventions [16,20]. However, existing literature does not explain how shared principles can be operationalized early into an interorganizational value model, which GzGr's sequential patterns of actions address.

Embedding people, technology, and organizational interests within this value model enabled GzGr to balance diverse values, address challenges in the primary care process, ensure technical feasibility, and maintain financial viability. Specifically, different organizations and different levels within organizations operate with their own structures, priorities, and financial logic, leading to differences in the perceived value of the collaboration and outcomes [38]. Taking all organizational values into account will help identify priorities, problems, and solutions that can improve real-world implementation of contextually adapted technologies [13,18].

In addition, GzGr continuously assessed value, monetary, or otherwise (such as improved knowledge, diagnostics, or patient care) to enact the interorganizational value model. An earlier study showed that moving beyond monetary value and keeping people, technology, and business in mind can lead to alternative, sustainable payment models [39]. Literature also highlights the importance of integrating people and technology, such as aligning workflows to improve eHealth usability [1,40], but often overlooks the business dimension. While business model design based on desirability, feasibility, and viability of interventions is seen as key to success [13,18,41], financial

XSL•FO
RenderX

interdependencies between organizations are seen as a problem for business modeling [12].

The contribution of our study lies in identifying the sequential patterns of actions for enacting an interorganizational value model. This model integrates people, technology, and business perspectives to align principles for CDSS development and guide coordinated actions across independent organizations.

Iterative, Co-Creative Experimentation That Spans Development and Implementation

CDSS are considered complex interventions due to the multiple interacting components and the connection between end users and their context. Developing and implementing a complex intervention in a complex system requires an iterative process of development and implementation [10]. In this research, we found that the boundary between development and implementation was blurred and that iterative experimentation was an ongoing process of the sequential patterns of actions. This led to GzGr building and sustaining the involvement of multiple organizations, instead of only one-off interactions between developers and users.

For GzGr, the learning process took place in small-scale, short-cycle experiments with multilevel stakeholders, which is not common in the implementation in Dutch health care. Small-scale, short-cycle testing is especially relevant for eHealth technologies, which evolve rapidly and require adaptation to social and contextual factors for adoption [42]. Therefore, scholars promote small-scale iteration and agile methods such as “fail early, fail often,” without long-term, extensive datasets, to reduce delays and costs [12,41,43]. However, such an iterative approach conflicts with traditional development and lengthy outcome measure processes in the health care industry (eg, medical device regulations or randomized controlled trials). Health care is considered a risk-averse and regulated industry [43]. GzGr involved practice-oriented academic partners, which can foster implementation in a risk-averse industry [44].

Another difference between the technological and health care industries is how implementation is perceived. In the digital industry, implementation is often seen as the final phase and end point, whereas in health care, it is a complex, ongoing process and should not be treated as the last step [36]. We advocate for the mechanism of iterative, co-creative experimentation that bridges development and implementation, rather than traditional, phased models. The National Health Service also implemented a continuous learning and adapting iterative process, aligning the technology with adopters, which requires iterative refinement and an agile approach, leading to regional adoption [45].

In such an iterative process (rather common in digital innovation in other industries [46]), it is crucial to involve all relevant stakeholders in the design and test phase to improve the implementation success of digital health interventions, realizing a co-creative process [1,40,47,48]. By involving end users in the development, a collaboration can anticipate implementation problems. Additionally, consensus on CDSS functionality and added value supports implementation [48,49]. The sequential pattern of actions included user feedback, which led to rapid technological improvements, strengthening trust in both the mission and the collaboration before broader implementation to more risk-averse groups. This method is in line with AI adoption within the National Health Service [20].

Previously mentioned studies mainly emphasized end user involvement (particularly health care professionals) during development in controlled settings. However, involving multilevel stakeholders during this process helps generate buy-in by aligning intervention design with strategic goals, thereby supporting adoption and keeping service models central [1,13,41,47,50,51]. In doing so, GzGr illustrates how to overcome the common pitfall of “pilotitis,” where digital health innovations stall in the pilot phase without scaling [52].

Our findings contribute by identifying the sequential patterns of actions that underlie an iterative, co-creative experimentation in a risk-averse setting. GzGr created a continuous, co-creative experimentation process spanning development and implementation and sustainably involving multilevel stakeholders, who kept the mission in mind and did not treat development and implementation as separate phases. This illustrates how joint progress can be achieved by connecting social change and technological development, which was crucial for achieving sustainable implementation of CDSS.

Limitations

To interpret the results of this study, we have reflected on several limitations in our research. First, the focus on a single case in the Netherlands restricts the generalizability of the findings to other contexts. CDSSs have been scaled to other organizations and are currently being expanded to additional regions. However, it remains uncertain to what extent the collaboration’s approach to CDSS development and implementation is dependent on context or how this compares to other countries. Further research is needed to validate these findings across different organizations, regions, and countries.

Second, reflecting on our use of qualitative data and a retrospective analysis, recalling bias and interpretative variation among respondents might exist. We mitigated this limitation by combining data sources (interviews, documents, and GzGr experts), making use of data triangulation. Moreover, our transdisciplinary research team enriched qualitative data analysis by incorporating diverse perspectives, including medical expertise, innovation ecosystems, organizational governance, design, and implementation. This multiperspective approach enhances the study’s applicability and relevance, despite its single-case focus. Additionally, the retrospective character of this research gave us the opportunity to compose a chronological overview. Due to this nature, the study has a descriptive rather than a predictive value, as formally no causality between actions, sequences of actions, and effects was assessed. Future research may provide predictive insights into how the timeline of the sequential patterns of actions and mechanisms unfolds during implementation.

Third, our identification of 2 mechanisms may reflect a degree of interpretation of the researchers, as is inherent to qualitative research. We addressed this by documenting our observations in a chronological overview table (Multimedia Appendix 3),

XSL•FO
RenderX

performing member checks, and linking our findings to existing literature. These factors provide a thorough understanding of the mechanisms underlying CDSS development and implementation in an interorganizational context.

Implications and Future Research

This research identified mechanisms that help interorganizational collaborations overcome challenges in the iterative development and implementation of CDSS. In this perspective, our study extends the determinant-level frameworks NASSS, Centre for eHealth and Wellbeing Research, and UK Medical Research Council by adding an interorganizational, process-level perspective that captures how collaboration unfolds across independent organizations. It shows how organizations, while maintaining their own interests, continuously realign around a shared goal through iterative steps to compose an interorganizational value model that sustains collective progress.

Our practical implications consist of 3 elements. First, to navigate complexity, the development and implementation of CDSS in interorganizational collaborations should involve multilevel stakeholders from the start of the collaboration, especially innovative and influential individuals who are willing to go beyond formal roles to advance shared goals and form a coalition of the willing. Second, an orchestrating actor, such as GzGr, is needed to facilitate communication, alignment, and technical support across individual organizations within interorganizational collaborations. Third, an iterative approach should be adopted, where feedback from multilevel stakeholders informs continuous technological and collaborative improvements that will support both adoption and long-term collaboration.

Future research should enhance the generalizability of our findings by comparing this case with other interorganizational collaborations to identify shared mechanisms. Additionally, further investigation is needed into the scalability of CDSS, distinguishing context-dependent elements. Finally, future research should explore how these mechanisms can be systematically applied to support adaptive implementation in research and practice.

Conclusions

Our study offered insights into the mechanisms for developing and implementing CDSS in an interorganizational collaboration in primary care. Our findings extend the determinant-level implementation frameworks for complex digital health interventions by offering process-level insights into sequential patterns of actions that show how change unfolds in complex, real-world settings. We identified two mechanisms: (1) enacting an interorganizational value model and (2) iterative, co-creative experimentation. These mechanisms supported both technological and organizational change, through the approach of integrating people, technology, and business aspects for all and across organizations. Through this approach, CDSS were generated and implemented via an iterative, co-creative experiment. This resulted in CDSS that were (financially) viable and that generated added value per individual stakeholder and organization. Additionally, CDSS could be embedded in the individual organizations as they fulfilled their values, needs, and interests. Therefore, CDSSs reached multilevel support and were well adopted.

Overall, our study enhances the understanding of how sequential patterns of actions enable digital health innovations to be embedded and sustained in an interorganizational collaboration. It highlights that successful implementation in complex systems depends on establishing shared principles early, forming the basis for an interorganizational value model. Simultaneously, an iterative collaboration spanning development and implementation is required. Throughout this process, continuous alignment of social, technological, and organizational dimensions is essential. We offered both practical and conceptual contributions to understanding how digital health interventions can be embedded and sustained in interorganizational collaboration.

Acknowledgments

The authors would like to thank all participants and members of GzGr for sharing their time, insights, and openness to reflect on their development and implementation process. The authors would also like to thank Prof Dr Anne Marie Weggelaar for her role in securing the funding that made this project possible. Additionally, the authors appreciate the constructive input of our interdisciplinary author team throughout the research and writing process. Throughout manuscript development, we conducted this review with the assistance of ChatGPT, versions 4 and 5, from OpenAI in 2024 and 2025. We used generative artificial intelligence as a sparring partner (eg, for conceptual distinctions) to improve English phrasing and condense text. We carefully reviewed and refined all content to ensure our perceptions, analytical depth, and conceptual coherence.

Funding

Data collection for this study was made possible through funding provided by the BigMedilytics program. The authors were supported by their respective institutions; no additional external funding was obtained for this work. The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request. The authors chose this approach because the transcripts include personal details and context-specific reflections, some of which were not relevant for the analysis and may be sensitive for collaborating partners.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: JJMMS, JDHvW, EGH, MK, FD, HJAvO
Data curation: JJMMS
Formal analysis: JJMMS, JDHvW, HJAvO
Investigation: JDHvW, JJMMS
Methodology: MK, FD, JJMMS
Supervision: MK, FD
Validation: HJAvO, EGH
Writing – original draft: JJMMS, HJAvO, JDHvW
Writing – review & editing: MK, FD, EH

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of the participants.
[DOCX File, 53 KB - jmir_v28i1e83084_app1.docx]

Multimedia Appendix 2
Elaboration on coding and codebook process.
[DOCX File, 57 KB - jmir_v28i1e83084_app2.docx]

Multimedia Appendix 3
Chronological overview of events, challenges, and mechanisms encountered by *Gezonde zorg, Gezonde regio*.
[DOCX File, 57 KB - jmir_v28i1e83084_app3.docx]

Checklist 1
COREQ checklist.
[PDF File, 491 KB - jmir_v28i1e83084_app4.pdf]

References

1. Abell B, Naicker S, Rodwell D, et al. Identifying barriers and facilitators to successful implementation of computerized clinical decision support systems in hospitals: a NASSS framework-informed scoping review. Implement Sci 2023 Jul 26;18(1):32. [doi: 10.1186/s13012-023-01287-y] [Medline: 37495997]
2. Harada T, Miyagami T, Kunitomo K, Shimizu T. Clinical decision support systems for diagnosis in primary care: a scoping review. Int J Environ Res Public Health 2021 Aug 10;18(16):8435. [doi: 10.3390/ijerph18168435] [Medline: 34444182]
3. Ji M, Genchev GZ, Huang H, Xu T, Lu H, Yu G. Evaluation framework for successful artificial intelligence-enabled clinical decision support systems: mixed methods study. J Med Internet Res 2021 Jun 2;23(6):e25929. [doi: 10.2196/25929] [Medline: 34076581]
4. Jamian L, Wheless L, Crofford LJ, Barnado A. Rule-based and machine learning algorithms identify patients with systemic sclerosis accurately in the electronic health record. Arthritis Res Ther 2019 Dec 30;21(1):305. [doi: 10.1186/s13075-019-2092-7] [Medline: 31888720]
5. Devaraj S, Sharma SK, Fausto DJ, Viernes S, Kharrazi H. Barriers and facilitators to clinical decision support systems adoption: a systematic review. J Bus Adm Res 2014;3(2):36. [doi: 10.5430/jbar.v3n2p36]
6. Laka M, Milazzo A, Merlin T. Factors that impact the adoption of clinical decision support systems (CDSS) for antibiotic management. Int J Environ Res Public Health 2021 Feb 16;18(4):1901. [doi: 10.3390/ijerph18041901] [Medline: 33669353]
7. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3(1):17. [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
8. Meunier PY, Raynaud C, Guimaraes E, Gueyffier F, Letrilliart L. Barriers and facilitators to the use of clinical decision support systems in primary care: a mixed-methods systematic review. Ann Fam Med 2023;21(1):57-69. [doi: 10.1370/afm.2908] [Medline: 36690490]
9. Toth-Pal E, Wårdh I, Strender LE, Nilsson G. Implementing a clinical decision-support system in practice: a qualitative analysis of influencing attitudes and characteristics among general practitioners. Inform Health Soc Care 2008 Mar;33(1):39-54. [doi: 10.1080/17538150801956754] [Medline: 18604761]
10. O'Cathain A, Croot L, Duncan E, et al. Guidance on how to develop complex interventions to improve health and healthcare. BMJ Open 2019 Aug 15;9(8):e029954. [doi: 10.1136/bmjopen-2019-029954] [Medline: 31420394]

XSL•FO
RenderX

11. van Dyk L. A review of telehealth service implementation frameworks. Int J Environ Res Public Health 2014 Jan 23;11(2):1279-1298. [doi: 10.3390/ijerph110201279] [Medline: 24464237]
12. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
13. van Gemert-Pijnen JEWC, Nijland N, van Limburg M, et al. A holistic framework to improve the uptake and impact of eHealth technologies. J Med Internet Res 2011 Dec 5;13(4):e111. [doi: 10.2196/jmir.1672] [Medline: 22155738]
14. Doolin B. Implementing e-health. In: The Oxford Handbook of Health Care Management: Oxford University Press; 2016:352-374 URL: https://books.google.co.in/books?hl=en&lr=&id=KIHDAAAQBAJ&oi=fnd&pg=PA352&dq=Implementing+E+Health&ots=Jm0Dm68sg-&sig=NE9LBPRpIZZ3xXm1Nk&redir_esc=y#v=onepage&q=Implementing%20E+Health&f=false [accessed 2025-07-25]
15. Plsek PE, Greenhalgh T. The challenge of complexity in health care. BMJ 2001 Sep 15;323(7313):625-628. [doi: 10.1136/bmj.323.7313.625]
16. van Limburg M, van Gemert-Pijnen JEWC, Nijland N, Ossebaard HC, Hendrix RMG, Seydel ER. Why business modeling is crucial in the development of eHealth technologies. J Med Internet Res 2011 Dec 28;13(4):e124. [doi: 10.2196/jmir.1674] [Medline: 22204896]
17. Cherns A. The principles of sociotechnical design. Hum Relat 1976 Aug;29(8):783-792. [doi: 10.1177/001872677602900806]
18. Kip H, Beerlage-de Jong N, van Gemert-Pijnen LJEWC, Kelders SM. The CeHRes roadmap 2.0: update of a holistic framework for development, implementation, and evaluation of eHealth technologies. J Med Internet Res 2025 Jan 13;27:e59601. [doi: 10.2196/59601] [Medline: 39805104]
19. Nilsen P. Making sense of implementation theories, models and frameworks. Implement Sci 2015 Apr 21;10(1):53. [doi: 10.1186/s13012-015-0242-0] [Medline: 25895742]
20. Morrison K. Artificial intelligence and the NHS: a qualitative exploration of the factors influencing adoption. Future Healthc J 2021 Nov;8(3):e648-e654. [doi: 10.7861/fhj.2020-0258] [Medline: 34888459]
21. May CR, Johnson M, Finch T. Implementation, context and complexity. Implement Sci 2016 Oct 19;11(1):141. [doi: 10.1186/s13012-016-0506-3] [Medline: 27756414]
22. Aunger JA, Millar R, Greenhalgh J, Mannion R, Rafferty AM, McLeod H. Why do some inter-organisational collaborations in healthcare work when others do not? A realist review. Syst Rev 2021 Mar 22;10(1):82. [doi: 10.1186/s13643-021-01630-8] [Medline: 33752755]
23. Karam M, Brault I, Van Durme T, Macq J. Comparing interprofessional and interorganizational collaboration in healthcare: a systematic review of the qualitative research. Int J Nurs Stud 2018 Mar;79:70-83. [doi: 10.1016/j.ijnurstu.2017.11.002] [Medline: 29202313]
24. Moullin JC, Dickson KS, Stadnick NA, et al. Ten recommendations for using implementation frameworks in research and practice. Implement Sci Commun 2020;1:42. [doi: 10.1186/s43058-020-00023-7] [Medline: 32885199]
25. Yin RK. Case Study Research: Design and Methods: Sage; 2009. [doi: 10.33524/cjar.v14i1.73]
26. Tracy SJ. Qualitative Research Methods: Collecting Evidence, Crafting Analysis, Communicating Impact: Wiley-Blackwell; 2024.
27. Bryman A. Social Research Methods, 5th edition: Oxford University Press; 2016. URL: https://archive.org/details/socialresearchme0000brym_w6b0 [accessed 2026-02-06]
28. Langley A. Strategies for theorizing from process data. Acad Manag Rev 1999 Oct;24(4):691. [doi: 10.5465/amr.1999.2553248]
29. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
30. Organisatie en partners [Website in Dutch]. Gezonde zorg, Gezonde regio. 2024. URL: https://gzgr.nl/over-gzgr/organisatie-en-partners [accessed 2024-09-13]
31. Zo gebruiken wij data [Website in Dutch]. Gezonde zorg, Gezonde regio. 2024. URL: https://gzgr.nl/over-gzgr/zo-gebruiken-wij-data [accessed 2024-09-13]
32. Flick U. Designing Qualitative Research, 2nd edition: SAGE Publications; 2018. [doi: 10.4135/9781529622737]
33. Aguboshim FC. Adequacy of sample size in a qualitative case study and the dilemma of data saturation: a narrative review. World J Adv Res Rev 2021;10(3):180-187. [doi: 10.30574/wjarr.2021.10.3.0277]
34. Parker C, Scott S, Geddes A. In: Cernat A, Sakshaug JW, Delamon S, editors. Snowball Sampling: SAGE Publications Ltd; 2019. [doi: 10.4135/9781526421036831710]
35. Braun V, Clarke V, Hayfield N, Terry G. Thematic analysis. In: Liamputtong P, editor. Handbook of Research Methods in Health Social Sciences: Springer; 2019:843-860. [doi: 10.1007/978-981-10-5251-4_103]
36. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]

XSL•FO
RenderX

37. Bohnett E, Vacca R, Hu Y, Hulse D, Varda D. Resilience and fragmentation in healthcare coalitions: the link between resource contributions and centrality in health-related interorganizational networks. Soc Networks 2022 Oct;71:87-95. [doi: 10.1016/j.socnet.2022.07.004] [Medline: 36060606]
38. Porter ME. What is value in health care? N Engl J Med 2010 Dec 23;363(26):2477-2481. [doi: 10.1056/NEJMp1011024] [Medline: 21142528]
39. Rakers MM, van Os HJA, Recourt K, Mosis G, Chavannes NH, Struijs JN. Perceived barriers and facilitators of structural reimbursement for remote patient monitoring, an exploratory qualitative study. Health Policy Technol 2023 Mar;12(1):100718. [doi: 10.1016/j.hlpt.2022.100718]
40. Seljelid B, Varsi C, Solberg Nes L, Øystese KA, Børøsund E. A digital patient-provider communication intervention (InvolveMe): qualitative study on the implementation preparation based on identified facilitators and barriers. J Med Internet Res 2021 Apr 8;23(4):e22399. [doi: 10.2196/22399] [Medline: 33830063]
41. Christie HL, Boots LMM, Hermans I, et al. Business models of eHealth interventions to support informal caregivers of people with dementia in the Netherlands: analysis of case studies. JMIR Aging 2021 Jun 3;4(2):e24724. [doi: 10.2196/24724] [Medline: 34081009]
42. Kumar S, Nilsen WJ, Abernethy A, et al. Mobile health technology evaluation: the mHealth evidence workshop. Am J Prev Med 2013 Aug;45(2):228-236. [doi: 10.1016/j.amepre.2013.03.017] [Medline: 23867031]
43. Duffy A, Christie GJ, Moreno S. The challenges toward real-world implementation of digital health design approaches: narrative review. JMIR Hum Factors 2022 Sep 9;9(3):e35693. [doi: 10.2196/35693] [Medline: 36083628]
44. Mervyn K, Amoo N, Malby R. Challenges and insights in inter-organizational collaborative healthcare networks: an empirical case study of a place-based network. Int J Organ Anal 2019;27(4):875-902. [doi: 10.1108/IJOA-05-2018-1415]
45. Elizondo A, Williams R, Anderson S, Cresswell K. Implementing integrated care infrastructure: a longitudinal study on the interplay of policies, interorganizational arrangements and interoperability in NHS England. Health Policy 2025 Oct;160:105237. [doi: 10.1016/j.healthpol.2024.105237] [Medline: 39706716]
46. Brown T. Design thinking. Harv Bus Rev 2008 Jun;86(6):84-92. [Medline: 18605031]
47. Ackerhans S, Huynh T, Kaiser C, Schultz C. Exploring the role of professional identity in the implementation of clinical decision support systems-a narrative review. Implement Sci 2024 Feb 12;19(1):11. [doi: 10.1186/s13012-024-01339-x] [Medline: 38347525]
48. Eichner JS, Das M. Challenges and barriers to clinical decision support (CDS) design and implementation experienced in the Agency for Healthcare Research and Quality CDS demonstrations (prepared for the AHRQ National Resource Center for Health Information Technology under contract no. 290-04-0016). : Agency for Healthcare Research and Quality; 2012 URL: https://digital.ahrq.gov/sites/default/files/docs/page/CDS_challenges_and_barriers.pdf [accessed 2025-08-15]
49. Schroeder D, Luig T, Finch TL, Beesoon S, Campbell-Scherer DL. Understanding implementation context and social processes through integrating normalization process theory (NPT) and the consolidated framework for implementation research (CFIR). Implement Sci Commun 2022 Feb 9;3(1):13. [doi: 10.1186/s43058-022-00264-8] [Medline: 35139915]
50. Tossaint-Schoenmakers R, Kasteleyn MJ, Rauwerdink A, Chavannes N, Willems S, Talboom-Kamp EPWA. Development of a quality management model and self-assessment questionnaire for hybrid health care: concept mapping study. JMIR Form Res 2022 Jul 7;6(7):e38683. [doi: 10.2196/38683] [Medline: 35797097]
51. van Velsen L, Ludden G, Grünloh C. The limitations of user-and human-centered design in an eHealth context and how to move beyond them. J Med Internet Res 2022 Oct 5;24(10):e37341. [doi: 10.2196/37341] [Medline: 36197718]
52. Pannunzio V. Towards a convergent approach to the use of data in digital health design [Dissertation]. : Delft University of Technology; 2023 URL: https://research.tudelft.nl/en/publications/towards-a-convergent-approach-to-the-use-of-data-in-digital-healt/ [accessed 2026-01-30]

ABBREVIATIONS

CDSS: clinical decision support systems
COREQ: Consolidated Criteria for Reporting Qualitative Research
GP: general practitioner
GzGr: *Gezonde zorg, Gezonde regio* (Healthy care, Healthy region)
NASSS: Nonadoption, abandonment, scale-up, spread, and sustainability

XSL•FO
RenderX

Edited by S Brini; submitted 27.Aug.2025; peer-reviewed by R Akinniranye, RS Mannak, Y Wu; accepted 11.Nov.2025; published 05.Mar.2026.

Please cite as:
Santema JJMM, van Wijngaarden JDH, Hiddink EG, Deken F, Kleinsmann M, van Os HJA
Interorganizational Mechanisms for Developing and Implementing Clinical Decision Support Systems in Primary Care: Exploratory, Qualitative Case Study
J Med Internet Res 2026;28:e83084
URL: https://www.jmir.org/2026/1/e83084
doi:10.2196/83084

© Jesse J M M Santema, Jeroen D H van Wijngaarden, Eric G Hiddink, Fleur Deken, Maaike Kleinsmann, Hendrikus J A van Os. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessing Health Care Professionals' Perceptions of a New System in Clinical Workflows: Systems Engineering Initiative for Patient Safety–Based Consensual Qualitative Research

Ye-Eun Park[1*], BA; Minsu Ock[2*], MD, PhD; Jae-Ho Lee[1,3], MD, PhD; Dae-Hyun Ko[4], MD, PhD; Hak-Jae Lee[5], MD, PhD; Taezoon Park[6], PhD; Junsang Yoo[7], PhD; Yura Lee[1], MD, PhD

[1]Department of Information Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[2]Department of Preventive Medicine, Ulsan University Hospital, University of Ulsan College of Medicine, Ulsan, Republic of Korea
[3]Department of Emergency Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[4]Department of Laboratory Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[5]Division of Acute Care Surgery, Department of Surgery, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[6]Department of Industrial & Information Systems Engineering, Soongsil University, Seoul, Republic of Korea
[7]Department of Digital Health, Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University, Seoul, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Yura Lee, MD, PhD
Department of Information Medicine
Asan Medical Center
University of Ulsan College of Medicine
Seoul
Republic of Korea
Phone: 82 2 3010 1498
Fax: 82 2 3010 2531
Email: haepary@naver.com

Abstract

Background: Artificial intelligence (AI)–enabled clinical decision support systems (CDSSs) are increasingly embedded within electronic health record (EHR) environments; however, their introduction can disrupt existing workflows and raise patient safety concerns, particularly in high-stakes settings such as surgical transfusion. Limited qualitative evidence exists regarding how frontline professionals anticipate the clinical, organizational, and workflow implications of such systems before wider deployment.

Objective: This study aims to qualitatively examine the anticipated clinical, organizational, and workflow-level implications of implementing personalized Maximum Surgical Blood Order Schedule—Thoracic Surgery (pMSBOS-TS), an AI-enabled CDSS for personalized surgical blood ordering, before large-scale deployment.

Methods: We conducted a consensual qualitative study with 14 multidisciplinary health care professionals involved in transfusion-related tasks at a large tertiary hospital. Following 1 pilot focus group to refine the interview guide and workflow diagram, 2 semistructured focus group discussions were held with 14 participants (5 physicians, 6 nurses, and 3 blood bank staff). Transcripts were analyzed using the Systems Engineering Initiative for Patient Safety (SEIPS) 101 framework, focusing on People, Environment, Tools, and Tasks, and were supported by task- and workflow-based analyses of transfusion processes. Member checking was conducted with participants and external clinicians to enhance validity.

Results: A total of 189 semantic units and 61 core ideas were identified across 18 subdomains and 7 overarching domains. Participants anticipated that pMSBOS-TS could reduce unwarranted variation in blood ordering and planning, provided that algorithmic performance is reliable and the interface is tightly integrated into existing EHR workflows. At the same time, they expressed concerns regarding increased verification burden, system limitations in unexpected clinical scenarios, and potential communication bottlenecks between clinical units and the blood bank. Organizational culture, governance structures, and local transfusion logistics were viewed as critical determinants of whether the system would reduce or inadvertently increase workload and blood product waste.

XSL•FO
RenderX

Conclusions: This preimplementation, SEIPS-based qualitative evaluation suggests that the successful adoption of an AI-enabled transfusion CDSS depends not only on predictive performance but also on sociotechnical readiness, including user trust, workflow fit, and organizational support. These findings provide practice-based insights to inform staged implementation, training, and governance strategies aimed at safely integrating predictive transfusion CDSSs into EHR-supported surgical workflows.

(J Med Internet Res 2026;28:e86166) doi:10.2196/86166

KEYWORDS

decision support systems (clinical); precision medicine, patient safety; blood transfusion; algorithms (artificial intelligence); information system

Introduction

The integration of artificial intelligence (AI) and digital systems into clinical practice is transforming health care delivery. As health care professionals increasingly encounter emerging technologies, their acceptance and perceptions of these systems substantially influence clinical efficiency and implementation success [1-3]. Despite growing interest in digital innovations, including clinical decision support systems (CDSSs), their adoption can disrupt existing workflows, necessitating careful evaluation of organizational- and user-level impacts [4,5].

Although prior workflow analyses have predominantly focused on identifying the root causes of patient safety incidents [6-8], there is an increasing need for proactive assessments during the early stages of system deployment. In particular, the rapid proliferation of electronic health record (EHR)–based applications requires a deeper understanding of how novel systems interface with existing clinical processes. Nevertheless, comprehensive workflow analyses addressing the multifaceted challenges of system adoption remain limited.

The widespread implementation of EHRs, accelerated by initiatives such as the Meaningful Use Program under the Health Information Technology for Economic and Clinical Health Act [9-13], has catalyzed the development of embedded CDSSs to enhance care quality and operational efficiency [14-16]. However, the adoption and integration of such tools remain complex, particularly in high-stakes settings such as surgical transfusion.

Traditional evaluation approaches, such as user satisfaction surveys or system log data, are useful for capturing surface-level feedback but often fall short in explaining context-dependent interactions among health care professionals and between users and systems [17]. Consequently, rigorous qualitative methodologies are essential for understanding the nuanced implications of technology integration [18].

Accordingly, this consensual qualitative research (CQR) study aimed to conduct a preimplementation evaluation of the personalized Maximum Surgical Blood Order Schedule—Thoracic Surgery (pMSBOS-TS) system by examining its anticipated implications for workflows, usability, and organizational conditions, and by identifying factors that may support its safe and effective integration into clinical practice [19]. To our knowledge, only a few studies have examined AI-enabled CDSSs for transfusion planning using a structured preimplementation evaluation; most existing CDSS research has focused on postdeployment clinician acceptance and use rather than prospective workflow and system-integration assessments [20]. By applying CQR alongside a sociotechnical framework before system deployment, this study provides a multistakeholder assessment of how an AI-based transfusion tool may influence workflows, communication patterns, and organizational processes.

Methods

System Description (pMSBOS-TS)

pMSBOS-TS is a machine learning–based CDSS developed to generate personalized maximum blood-ordering recommendations for thoracic surgery patients by integrating patient-, laboratory-, and procedure-specific predictors [21]. The underlying algorithms were developed and validated in a prior work by the collaborating investigators [19], and this study evaluates the system's anticipated effects in a real-world clinical context.

Study Site

This study was conducted at Asan Medical Center, a 2764-bed tertiary hospital that performs approximately 70,892 surgeries annually as of 2023.

Study Design

This qualitative study examined the anticipated effects of applying the pMSBOS-TS system for health care professionals involved in transfusion tasks, under the assumption that the system was integrated into the existing electronic medical record (EMR). The assessment used the Systems Engineering Initiative for Patient Safety (SEIPS) framework to evaluate potential impacts across the domains of People, Environment, Tools, and Tasks (PETT) [22,23].

Originally introduced by Carayon et al [22] and subsequently expanded, SEIPS conceptualizes health care as comprising interacting components—person, task, technology/tools, organization, and environment—that shape care processes and, ultimately, outcomes such as patient safety and care quality [23,24]. SEIPS 2.0 incorporated patients and families as active participants and placed greater emphasis on processes, while SEIPS 3.0 expanded the framework to the patient-journey level, including cross-setting transitions.

To enhance accessibility, Holden and Carayon [23] proposed "SEIPS 101," a simplified, practice-oriented adaptation of the model. SEIPS 101 retains the core elements of the work system, process, and outcomes, while streamlining the work system into 4 primary components (PETT). In this study, *people* refer to

XSL•FO
RenderX

health care stakeholders (clinicians and patients) and their capabilities or needs; *tasks* denote activities and workflows; *tools/technology* include equipment, information technology (IT) systems (CDSS/health information systems), and other job aids; and *environment* encompasses both the physical setting (eg, layout, lighting, and noise) and the social or organizational context (eg, culture, policies, and teamwork).

By focusing on these elements and their interactions, a SEIPS/PETT-based evaluation can map how a new CDSS influences clinical work—for instance, how it reshapes clinicians' tasks, introduces tool-related issues, or alters team dynamics—and how these changes affect care processes and outcomes. This sociotechnical perspective is particularly useful for identifying system-level issues that purely clinical or IT-focused evaluations may overlook [25].

A key advantage of this approach is its holistic orientation toward workflow and safety, providing a structured method for analyzing and designing health care processes. The PETT scan can serve as both a checklist and a documentation tool to ensure that all relevant components of the work system are systematically addressed. Accordingly, we applied the PETT scan to identify barriers and facilitators across components and to explore their interactions, thereby providing a comprehensive overview of the work system. Our methodology systematically applied the PETT scan tool from the SEIPS framework to assess users, the surrounding environment, tasks, and tools/technology (Figure 1).

Three group sessions were conducted between October 19, 2023, and May 2, 2024, following a structured 5-stage process: participant recruitment, workflow feedback, system introduction, PETT-based evaluation, and postsession feedback (Figure 2).

Figure 1. Simplified SEIPS 101 model of work systems, processes, and outcomes. SEIPS: Systems Engineering Initiative for Patient Safety.

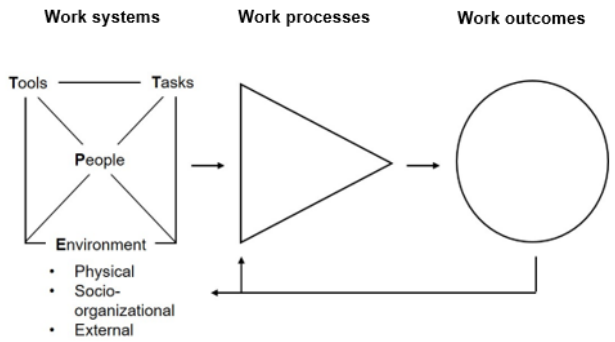

Figure 2. Overview of the participant engagement and SEIPS-PETT analysis process. PETT: People, Environment, Tools, and Tasks; pMSBOS-TS: personalized Maximum Surgical Blood Order Schedule—Thoracic Surgery; SEIPS: Systems Engineering Initiative for Patient Safety.

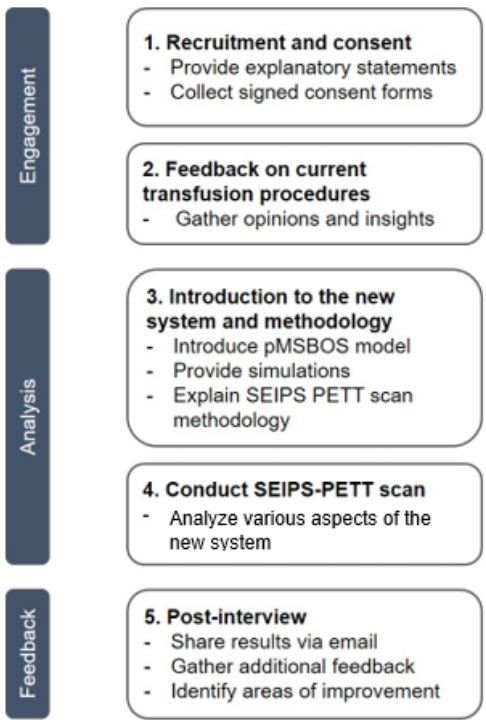

XSL•FO
RenderX

Recruitment

To assess the usability and workflow implications of the pMSBOS-TS system, we recruited health care professionals currently involved in transfusion-related tasks and identified as potential end users of the system. Recruitment was conducted via the hospital's online bulletin board and direct outreach. Eligible participants were those with more than 1 year of experience in procedures or surgeries with a high likelihood of requiring transfusion and who had experience handling blood products.

Inclusion criteria were surgeons prescribing blood products, nurses in surgical wards and operating rooms handling blood requests and transfusions, and blood bank staff managing the release and distribution of blood products.

Three focus group discussions (FGDs) were conducted for this study: 1 pilot session and 2 main sessions, each comprising 3-8 participants. Group composition was balanced across roles (eg, surgeons, nurses, and blood bank personnel) to gather diverse perspectives while maintaining thematic consistency. A semistructured interview guide was used for all sessions, and each participant took part in only 1 session.

Data Collection

All FGDs were conducted in person in a private meeting room within the hospital. No individual interviews were conducted; all qualitative data were collected through FGDs. Each session was facilitated by YL (female), a trained moderator with expertise in qualitative health research, and assisted by YEP (female), who managed logistics, note-taking, and audio recording. Although the researchers and participants worked within the same institution, no direct personal or supervisory relationships existed. Participants were informed of the study objectives and were aware that the moderator was a member of the research team; no personal goals or interests of the researchers were disclosed.

A semistructured interview guide was used in all sessions. Each session lasted 60-80 minutes and was audio-recorded with participant permission. Participants were encouraged to describe specific cases and reflect on workflow steps using printed workflow diagrams and sticky-note categorization. All participants were invited to speak, write, categorize notes, and present their views during the sessions.

All interviews were conducted in Korean, transcribed verbatim, reviewed by participants, and subsequently translated into English. Translation was performed by bilingual researchers fluent in both Korean and English. To ensure semantic equivalence, translated transcripts were reviewed by a second bilingual researcher, and discrepancies were resolved through discussion. To minimize interpretive bias, participants were emailed the FGD transcripts organized by content unit to verify accuracy and identify any distortions or omissions in meaning. No participants had a prior personal relationship with the moderator.

The methodology and reporting of the qualitative findings followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines [26]. We adhered to the COREQ guidelines in describing the study design, data collection, and qualitative analysis procedures.

Qualitative Analysis

Overview

The qualitative analysis was conducted in accordance with the principles of CQR to minimize interpretive distortion and to derive in-depth insights from clinical field experts regarding the use of pMSBOS-TS [27].

Initial Stage of Qualitative Analysis: Preliminary Discussions and Hypothesis Formation

The research team hypothesized that personalized prediction could support more efficient use of blood products. Based on a literature review of the adoption of new medical information systems and expert discussions, we concluded that, in addition to understanding the risks and benefits of pMSBOS-TS, factors related to its successful implementation would be essential [28,29].

Accordingly, the following core questions were developed:

- "What positive or negative impacts could be expected if pMSBOS-TS is introduced into the clinical field?"
- "What factors are important for the successful adaptation and implementation of pMSBOS-TS in clinical practice?"

To conduct a multidimensional evaluation of the risks and impacts associated with introducing a new application, we employed the PETT scan framework. Prior research using the SEIPS model was reviewed to inform the development of the interview guide [22,23]. Participants actively engaged with the PETT framework by manually categorizing insights using sticky notes, which facilitated recall and reflection. They also participated in a transfusion-related workflow analysis to identify potential impacts at each step.

A pilot focus group session was conducted with clinical professionals to examine transfusion-related processes in greater detail and to identify clinical scenarios relevant to the application of pMSBOS-TS. This pilot session served multiple purposes: it validated the structure and appropriateness of the PETT-based interview methodology, supported the development of a draft transfusion workflow diagram, and informed the refinement of the initial domain-subcategory-core idea table. Although the pilot session provided useful foundational insights, its data were excluded from the final cross-case analysis to ensure consistency with the main dataset.

Intracase Analysis: Task Analysis and Development of Domains and Subdomains

To ensure comprehensive capture of participants' views on the impacts and potential harms of pMSBOS-TS at each stage of the transfusion process, a detailed task analysis was performed [30,31]. The analysis team, comprising 2 PhD-level researchers and 1 research assistant with a bachelor's degree, developed the following domains, subdomains, and core ideas based on the core questions:

XSL•FO
RenderX

- Domains: transfusion-related work experiences (case examples), pMSBOS-TS–related opinions, and major categories derived from the PETT scan.
- Subdomains: specific subcomponents of the PETT scan and preliminary workflow elements identified during the pilot focus group sessions.
- Core ideas: derived by reviewing interview transcripts and postinterview notes; preliminary categorizations were validated through consultation with 2 domain experts.

Cross-Case Analysis

Two analysis team members (YEP and YL) independently reviewed and coded meaningful units from the transcripts and sticky notes into subdomains and core ideas. Coding discrepancies were resolved through consensus discussion or, when necessary, adjudicated by a third team member. Core ideas that did not fit existing categories were temporarily placed in an "Other" category and redefined as needed.

A frequency-based coding scheme was employed to reflect both sentiment and the prevalence of opinions: positive opinions were marked with plus signs (+), and negative opinions with minus signs (–). Frequency was indicated as follows: 1 symbol for 1 participant, 2 symbols for 2-3 participants, and 3 symbols for 4 or more participants. Expression formats and interpretation strategies were discussed and agreed upon for divergent opinions within the same theme.

Categories were defined as follows:

- General: common across all sessions (– – –/+ + +).
- Typical: consistent within a session but not endorsed by all participants.
- Variant: mixed opinions or views that appeared in only a subset of cases.

For validation, the finalized core ideas were shared via email with the original participants for member checking to ensure accuracy and minimize interpretive distortion. In addition, 3 clinical professionals from other medical institutions (meeting the same inclusion criteria as the interview participants) were contacted. After explaining the study objectives and methodology, the core ideas and task-analysis flow diagram were shared via email to confirm that the interpretations were neither biased nor incomplete. Data saturation was assessed during cross-case analysis; no new core ideas emerged from experts outside the interview group, indicating thematic saturation consistent with CQR guidelines.

Themes and core ideas were derived inductively from the data in accordance with the CQR approach, whereas the overarching domains were organized using the SEIPS/PETT framework. No themes were predetermined before analysis.

Ethics Statement

This study was approved by the Institutional Review Board (IRB) of Asan Medical Center, Korea (IRB 2023-0724), and conducted in accordance with relevant ethical guidelines. Informed consent was obtained from all participants before their involvement, with assurances of anonymity and confidentiality. Participants were briefed on the study's objectives and the intended use of the collected data. Participants received compensation for their participation.

Results

Participants

Fourteen participants took part in 2 FGDs, comprising 5 physicians (3 fellows/professors and 2 residents), 6 nurses (4 from the surgical intensive care unit, 1 from the anesthesia and recovery unit, and 1 from the internal medicine ward), and 3 blood bank staff members (Table 1). Detailed participant characteristics, including participant identifiers, are provided in Multimedia Appendix 1.

Table 1. Demographics of the participants.

Variables	Doctor (n=5)	Medical technologist (n=3)	Nurse (n=6)	Total (N=14)
Sex, n				
Male	2	1	0	3
Female	3	2	6	11
Age, mean (SD)	30 (11)	40 (14.1)	28.3 (6.9)	31.4 (11.7)
Transfusion experience, mean (SD)	8.2 (9.4)	20.3 (10.1)	13 (7.2)	13.53 (10.12)
Interviews participated in, n				
First	2	3	3	8
Second	3	0	3	6

We organized the results of a task analysis of transfusion-related processes using a swim-lane approach (Multimedia Appendix 2). From the workflow analysis and FGDs, we extracted 189 unique semantic units (81 in category A and 108 in category B, excluding duplicates). These units were organized according to their relevance to transfusion workflows. Semantic units describing inefficient use or cases involving blood products, regardless of whether pMSBOS-TS was applied, were classified as category A. By contrast, statements concerning anticipated impacts of pMSBOS-TS implementation or potential risks associated with its use were classified as category B.

The semantic units were further condensed into 61 core ideas: 21 from category A and 40 from category B. These were then organized into 18 subdomains (3 in category A and 15 in category B) and 7 overarching domains (1 in category A and 6 in category B). The complete analytical framework, including

XSL•FO
RenderX

domains, subdomains, and core ideas, is summarized in Multimedia Appendix 3.

In category A, titled "(In)Efficient Use of Blood Components," 3 subdomains were identified: *Inaccurate Dosage (Over- or Under-Dosing)*, *Inefficient Blood Usage and Processes*, and *Other Transfusion Process Errors*.

Category B comprised 6 primary domains: *People*, *Environment*, *Tools*, *Tasks*, *Work Processes*, and *Work Outcomes*. Within the People domain, 2 subdomains were identified: *Prescribing Physicians* (3 general and 1 variant core idea) and *Other Transfusion Stakeholders* (1 general and 1 variant core idea). The Environment domain included 2 subdomains: *Interprofessional Communication* and *Socio-organizational Context*. The Tools domain encompassed aspects related to algorithmic performance, interface design, and perceived reliability. The Tasks domain included *Achieving Algorithm Objectives*, *Task Complexity and Variability*, and *Handling Unanticipated Situations*.

In the Work Processes domain, responses were classified based on tasks associated with the application phase of pMSBOS-TS (eg, order entry and related processes). The Work Outcomes domain addressed participants' perceptions of the consequences or results associated with the algorithm's use.

Category A: Notable Experiences in Transfusion Practice or Cases of Inefficient Blood Product Use

Of the 21 core ideas from category A, 6 were classified as *typical* and another 6 as *variant*. Within the subdomain "Inaccurate Dosage (Over- or Under-Dosing)," 2 typical and 2 variant core ideas were identified. Participants primarily highlighted experiences related to the prescribing and preparation phases (Table 2).

Regarding unintentional over- or underprescription, several participants referred to habitual or standardized prescribing practices in their departments that led to unnecessary transfusions.

> *We routinely prescribed three units of RBCs, even when we did not actually end up transfusing them...* [Participant 2_5]

A view that routine surgeries rarely require transfusions was also expressed. During the preparation stage, including order verification on the ward and confirmation/preparation in the blood bank, participants noted discrepancies between prescribed and requested volumes.

> *When a patient is going into surgery, they often pre-prescribe fresh frozen plasma, and the blood bank will call us saying it was too much and ask to cancel it. It is a tough situation...* [Participant 1_8]

By contrast, a blood bank staff member pointed out that excessive orders often stemmed from physicians' personal preferences:

> *Sometimes the physician asks us to save much more blood than needed. So we end up placing unnecessary orders. It makes our inventory look abundant to the central blood center, but when we actually have an emergency, there is not much usable blood left.* [Participant 1_2]

In the subdomain of "Inefficient Blood Usage and Processes," 2 typical and 2 variant core ideas were identified. Participants noted issues in communication and procedural transparency during the transfusion preparation process.

> *I often feel there is a disconnect between prescription and preparation. There is a timing issue when blood is prepared, and we are supposed to call the physician directly—but sometimes we miss it. If the doctor does not say, "I have prescribed blood for tomorrow," we will not know until we ask. Then they just say, "Do not worry, it is for tomorrow." But in those communication gaps, we sometimes fail to have the blood ready when the patient actually needs it.* [Participant 1_4]

In addition, logistical difficulties in transporting blood products within the hospital were cited, including challenges related to elevators, distance, and location.

> *Sometimes we cannot even get an elevator. I think it is pretty hard for the ward staff to go pick up the blood themselves. We end up waiting a long time for the blood to arrive...* [Participant 1_4]

Some participants also described inconveniences in transfusion process workflows, including managing low blood inventory levels, emergency dispatch of blood transport vehicles, and cumbersome steps such as cross-matching or additional compatibility testing.

In the subdomain of "Other Transfusion Process Errors," 2 typical and 1 variant core ideas were identified. These primarily involved unexpected clinical events during procedures. For example, some participants described scenarios in which unanticipated vascular damage occurred during surgery.

> *One that really stands out is during CS or liver transplant surgeries—suddenly the aorta or artery tears, and we are hooking up multiple blood bags to the massive transfusion machine. It is just intense during those massive transfusions.* [Participant 1_6]

Other participants reflected on communication breakdowns and disorganized processes during emergencies:

> *Nobody is talking to each other, everyone is busy...Trying to get consent from the caregiver, some people are preparing the transfusion, others are trying to set up IV lines for massive transfusion, even when the lines are not ready yet—they just keep trying anyway...* [Participant 1_5]

XSL•FO
RenderX

Table 2. Cross-case analysis for inefficient use of blood components (category A).

Domain, subdomain, and core idea	Frequency[a]		Total frequency
	Group 1	Group 2	
1. Efficient Use of Blood Components			
1.1. Inaccurate Dosage (Over-/Underdosing)			
1.1.1. Prescription: clinician-driven over- or underordering	+ +	0	Variant
1.1.2. Prescription: over- or underordering not based on individual discretion	+	+ +	Typical
1.1.3. Preparation: over-/underpreparation during ward request or blood bank confirmation	+ +	+	Typical
1.1.4. Administration: clinically unnecessary or inadequate transfusions	+	0	Variant
1.2. Inefficient Blood Usage and Processes			
1.2.1. Lack of transparency in communication or procedures during transfusion preparation	+	+ +	Typical
1.2.2. Issues in subprocess tasks for order fulfillment	+	+	Typical
1.2.3. Logistics issues in internal transport of blood products	+ +	0	Variant
1.2.4. Blood management challenges	+ +	0	Variant
1.2.5. Preparation of transfusion support devices	0	0	N/A[b]
1.2.6 Blood product wastage	0	0	N/A
1.3. Other Transfusion Process Errors			
1.3.1. Patient identity and data checks	0	0	N/A
1.3.2. Transfusion ordering	0	+	Variant
1.3.3. Order verification	0	0	N/A
1.3.4. Ward-level blood preparation	+	0	Variant
1.3.5. Blood bank unit preparation	0	0	N/A
1.3.6. Blood transport and arrival confirmation	0	0	N/A
1.3.7. Transfusion administration	+	+	Typical
1.3.8. Transfusion completion	0	0	N/A
1.3.9. Other: unforeseen clinical scenarios	+ +	+	Typical

[a]Frequency indicators reflect both sentiment and prevalence of opinions. Positive opinions are denoted by plus signs (+) and negative opinions by minus signs (–). One symbol indicates 1 participant, 2 symbols indicate 2-3 participants, and 3 symbols indicate 4 or more participants. “0” indicates that no relevant statements were identified for that core idea.

[b]N/A: not applicable (also see footnote “a”).

Category B: Perceptions and Opinions Regarding pMSBOS-TS

People

Regarding the impact of pMSBOS-TS on individual users, participants generally agreed that the system could reduce variation in blood ordering volumes attributable to the prescribing physician’s personal tendencies (Table 3).

We used to just prescribe based on our assumptions, but now this app (pMSBOS-TS) prescribes based on what it studied through machine learning. [Participant 2_5]

Reduced deviation according to the characteristics of the physicians’ prescribing tendencies (stable/adventurous). [Sticky note]

XSL•FO
RenderX

Table 3. Cross-case analysis for the 6 primary domains: People, Environment, Tools, Tasks, Work Processes, and Work Outcomes (category B).

Domain, subdomain, and core idea	Frequency[a]		Total frequency
	Group 1	Group 2	
1. People			
1.1. Prescribing Physicians			
1.1.1. Variability in prescribed volume based on the physician's experience or skillset (may decrease variation [P[b]] or increase it [N[c]])	+ +	+ +	General
1.1.2. Adaptation gap according to the physician's proficiency with the ordering system	+ +	+ +	General
1.1.3. Clinician perceptions of CDSS[d] affecting adoption of the new system	+ +	+ +	General
1.1.4. Final verification of calculated transfusion requirement and prescribing responsibility must remain with the ordering physician	0	+ + +	Variant
1.2. Other Transfusion Stakeholders			
1.2.1. Blood bank staff, nurses, and other transfusion team members gain improved demand forecasting through the system	+ +	+ + +	General
1.2.2. Other: potential to reduce nonuser-initiated over- or underordering	0	+	Variant
2. Environment			
2.1. Interprofessional Communication			
2.1.1. Clear communication between clinical staff and the blood bank is critical for successful system adoption	+ +	0	Variant
2.1.2. System may enhance transparency (P) but could introduce additional confirmation steps or confusion (N)	– – –	– –	General
2.2. Socio-Organizational Context			
2.2.1. Organization's culture around individual variation in transfusion demand: blame culture may decrease (P), or lack of clear norms may increase confusion (N)	– –	+	Variant
2.2.2. Organizational climate and policies influence uptake	+ +	+ +	General
2.2.3. Institutional blood management challenges affect implementation	+ +	+	Typical
2.2.4. Clarity of governance is critical for program sustainability	+	+ +	Typical
2.2.5. Other: variations in system impact and blood management practices expected based on health care facility size and infrastructure	0	+	Variant
2.3. Physical Environment			
2.3.1. Anticipated effective utilization in settings with high concentrations of clinical staff and resources	+	0	Variant
2.3.2. Physical environment factors are critical for successful adoption	+	0	Variant
3. Tools			
3.1. Algorithm Performance			

XSL•FO
RenderX

Domain, subdomain, and core idea	Frequency[a]		Total frequency
	Group 1	Group 2	
3.1.1. System response time is a key determinant for adoption	0	+ + +	Variant
3.1.2. Ability to incorporate a wide range of clinical input variables	+ +	+ + +	General
3.1.3. Capability to generate tailored recommendations for various blood components	+	+ + +	Typical
3.1.4. Predictive accuracy of the algorithm is essential	+ +	+ +	General
3.2. Usability and System Design			
3.2.1. Interface layout and input mechanisms must support efficient use	+ +	+ +	General
3.2.2. Workflow integration should not interrupt clinical tasks	+	+	Typical
3.2.3. Seamless electronic medical record interoperability to prepopulate patient data and eliminate manual entry	+ +	+ + +	General
3.2.4. Other: support for operation across diverse platforms	+	0	Variant
3.3. Trust in the Tool			
3.3.1. Confidence in the pMSBOS-TS[e] model is critical for sustained use	+ +	+	Typical
3.3.2. Other: difficulty establishing confidence in the algorithm due to the inherently opaque machine-learning inference process	0	+	Variant
4. Tasks			
4.1 Achieving Algorithm Objectives			
4.1.1. Impact on returns/wastage: personalized demand forecasts should reduce waste (P) or, if overestimation/mistrust occurs, increase waste (N)	+	+/– – –	Variant
4.1.2. Supply from the National Blood Service: procurement may become easier (P) or more difficult (N)	–	0	Variant
4.1.3. Other: improved blood inventory management may help reduce delays in transfusion	0	+	Variant
4.2. Task Complexity and Variability			
4.2.1. System introduction may decrease (P) or increase (N) required time and effort	+/–	++/–	Variant
4.2.2. Must accommodate complex cases	+ +	0	Variant
4.3. Handling Unanticipated Situations			
4.3.1. Flexibility to manage unforeseen variables not captured by the algorithm	+ +	+ + +	General
4.3.2. Impact from overlapping work-system issues	+ +	+ +	General
4.3.3. Limiting tool application to defined scenarios ensures safe and effective use	0	+ +	Variant
4.3.4. Other: malfunctions or unintended consequences may arise if users do not fully understand the task or intended use of the system.	+	0	Variant
5. Work Processes			
5.1. Linked Diagnostic Orders			
5.1.1. Integration of ancillary tests with the transfusion program is critical	+	+ +	Typical
5.2. Preparation of Related Procedures/Equipment			
5.2.1. System may facilitate preparation of downstream tasks (P) or, conversely, increase complexity of subsequent steps (N)	–	+ +/– –	Variant
6. Work Outcomes			
6.1. Efficiency of Blood Use and Management			

XSL•FO
RenderX

Domain, subdomain, and core idea	Frequency[a]		Total frequency
	Group 1	Group 2	
6.1.1. Contribution of pMSBOS-TS to efficient utilization and inventory control: positive or negative	+ +	+/– – –	Variant
6.1.2. Long-term impact: accumulation of usage data to further refine predictive accuracy and inform adoption strategies	+	+ +	Typical
6.1.3. Indirect clinical benefits: supports expedited detection of abnormal laboratory findings, thereby enhancing overall patient management	+	+ +	Typical
6.2. Organizational Culture and Processes for Personalized Transfusion			
6.2.1. Role of pMSBOS-TS in fostering a culture and workflow for personalized maximum transfusion prediction: positive impact or none/negative impact	– –	–	Typical

[a]Frequency indicators reflect both sentiment and prevalence of opinions. Positive opinions are denoted by plus signs (+) and negative opinions by minus signs (–). One symbol indicates 1 participant, 2 symbols indicate 2-3 participants, and 3 symbols indicate 4 or more participants. “0” indicates that no relevant statements were identified for that core idea.

[b]P: positive.

[c]N: negative.

[d]CDSS: clinical decision support system.

[e]pMSBOS-TS: personalized Maximum Surgical Blood Order Schedule—Thoracic Surgery.

However, it was also noted that the effectiveness of pMSBOS-TS would depend on the prescriber’s level of proficiency with the system and their trust in its recommendations.

> *Someone who judges based on his/her own experience and is not familiar with a new tool, may pursue the existing method.* [Sticky note]

In one group, it was emphasized that final confirmation and responsibility for blood ordering based on calculated transfusion requirements should remain with the physician. Among the broader group of health care professionals involved in the transfusion process, there was a general expectation that pMSBOS-TS would support more accurate predictions of required blood volumes.

Environment

Regarding the environmental impact of pMSBOS-TS, participants expressed concern that its implementation might lead to increased verification steps or procedural confusion during blood preparation and release.

> *There used to be a set standard of 3 or 2 (example of the number of blood packs in order set), but if it changes for each patient...like, this patient needs to prepare 8, this patient needs to prepare 5...If our nurses had to check with the prescribing doctor each time and prepare a certain number of bloods, I thought there could be confusion.* [Participant 1_5]

There was also a shared perception that the success of integration would depend heavily on the organizational culture and internal dynamics of the institution.

> *I think that the surgical nursing team may focus more on the team atmosphere when the pMSBOS-TS is implemented.* [Participant 2_3]

Opinions were divided between the 2 groups regarding whether pMSBOS-TS would positively or negatively influence the development of an organizational culture that accommodates interindividual variability in transfusion needs.

Tools

Regarding pMSBOS-TS as a tool, participants broadly agreed that the algorithm should be capable of incorporating a wider range of clinical input variables, and that its predictive performance—specifically, its accuracy—would be crucial to successful adoption.

> *Even for the same disease, the bleeding risk differs depending on the severity, and since previous abdominal surgery has a big influence on the tissue adhesion, it seems likely that more blood transfusions will be needed. I wonder to what extent this will be reflected in the algorithm.* [Participant 1_7]

With respect to usability, participants consistently emphasized that the interface should be intuitive, with screen layouts and input controls designed for ease of use. They also stressed that patient information should be automatically integrated from the EMR to minimize manual data entry by users.

> *If (the data input of pMSBOS-TS is) not linked to EHR, additional workload is possible/input error is possible.* [Sticky note]

The importance of system response speed was highlighted frequently in only 1 of the 2 groups.

Tasks

Regarding the interaction between pMSBOS-TS and task performance, participants generally agreed that the system’s ability to flexibly accommodate unexpected clinical scenarios or variables not accounted for by the algorithm would be critical.

> *Variation in the skill of the surgeon performing the surgery/The possibility of an unexpected worse situation occurring for the patient.* [Sticky note]

Task performance was also noted to be influenced when issues from other components of the broader work system intersected

XSL•FO
RenderX

with the use of pMSBOS-TS. Opinions were divided across both groups regarding the anticipated impact of pMSBOS-TS on blood product returns or waste, as well as on the complexity and performance of transfusion-related tasks.

Work Processes and Work Outcomes

No dominant themes emerged concerning the influence of pMSBOS-TS on work processes or outcomes. Opinions were mixed regarding whether the system would affect subsequent tasks or contribute meaningfully to the management and utilization of blood products.

> *The pMSBOS-TS would be helpful in assessing the appropriateness of blood transfusion (for example, for health insurance's claim eligibility review)* [Participant 1_1]
>
> *If more blood is prescribed to account for the risk of bleeding, there may be more problems with distribution, such as returns or disposal.* [Participant 2_6]

Discussion

Principal Findings

This study identified clinicians' anticipated benefits and concerns regarding pMSBOS-TS across workflow, usability, and organizational domains. Consistent with prior CDSS research, such systems can offer clear benefits when successfully implemented, enhancing patient safety and supporting clinical decision-making. Through a qualitative investigation involving frontline health care professionals, we explored anticipated impacts—both positive and negative—associated with implementing pMSBOS-TS, a prediction-based CDSS for maximum surgical blood ordering in preoperative patients.

Findings revealed several core concepts that extend beyond transfusion tasks and are broadly applicable to CDSS implementation. These include the importance of users' (ie, physicians') system proficiency and trust, the influence of organizational culture, the accuracy of task performance, and overall system usability. By engaging diverse stakeholders and applying the PETT framework, we examined the hypothesis that successful and safe CDSS implementation depends not only on algorithmic performance but also on workflow integration and interprofessional interactions.

Interpretation and Implications

These observations illustrate how pMSBOS-TS may interact with existing transfusion workflows and sociotechnical structures. Whereas traditional MSBOS approaches are often static (eg, "for procedure X, always have Y units ready"), machine learning models can personalize recommendations by incorporating patient-specific factors [19]. Given the urgent and collaborative nature of transfusion processes, participants noted concerns regarding potential unintended effects on interdisciplinary collaboration and emphasized the need for flexibility within the system's functionality. These insights suggest that, for a newly introduced CDSS to achieve its intended impact, attention must be paid not only to algorithmic performance but also to the specific characteristics of the task environment, end users, and interface design.

Task analysis has long been used to address patient safety issues and guide quality improvement initiatives in health care, including the design, implementation, and optimization of health IT systems [30,32,33]. During the development of the swim-lane diagram, we observed that transfusion workflows extend beyond a simple physician-patient interaction, involving complex coordination among physicians, nurses, and blood bank staff. Findings from category A confirmed that inefficiencies arise not only from interprofessional issues but also from interdepartmental and spatial constraints. Compared with medication administration, transfusion processes are more resource-constrained and require specialized procedures—matching, delivery, storage, and multistep verification. As a result, discrepancies in prescribing or preparation can lead to significant resource waste and additional workload. Moreover, blood banks require specialized expertise and must coordinate directly with central blood suppliers, making communication breakdowns a potential source of operational strain.

This complexity was reflected in the cross-case analysis for category A, which identified inefficient practices such as over- or underprescribing influenced by order sets or department-specific habits. These findings align with participants' views that organizational culture will play a central role in shaping the success of pMSBOS-TS implementation. Concerns regarding inconsistent blood preparation, lack of transparency, and potential confusion during rollout further underscore the importance of attention to workflow readiness. Additionally, the need for compatibility testing, highlighted in both categories A and B, emphasizes that evaluations must account for the entire sequence of related workflows, including upstream preparation and downstream follow-up.

The core question—"What positive or negative impacts could be expected if pMSBOS-TS is introduced into clinical practice?"—elicited both supportive and critical responses. Anticipated benefits included reduced prescription variability and improved accuracy of blood use predictions, aligning with the system's development goals. However, participants expressed divergent views regarding downstream tasks, such as adjusting over- or underprescriptions driven by order sets and managing transfusion-related tools. Differences also emerged regarding the expected benefits of personalized blood ordering—short term (eg, time savings and reduced blood waste) versus long term (eg, improved blood reserve management).

The most frequently cited concern was the potential for increased workload due to additional confirmation (double-checking) steps. Given that transfusions often occur in life-threatening situations, this emphasis on safety is understandable. The introduction of a confirmation mechanism is an essential safeguard, and participants' concerns reflect a broadly shared and appropriate level of caution. In addition, several participants noted that patient-specific transfusion considerations—such as previous alloimmunization, allergic reactions, or rare blood requirements—are not represented in the current algorithm and would continue to require clinical

XSL•FO
RenderX

judgment. These factors delineate important boundaries for safe use and underscore the need for guidance on when algorithmic recommendations should be supplemented by clinician expertise.

Participants also voiced skepticism regarding the impact of personalized services on clinical workflows, a sentiment commonly reported in the adoption of precision medicine tools. This resistance may partly stem from the lack of distinction between short- and long-term effects in the interview questions. As with many new technologies, early implementation often increases user workload. Thus, realizing the intended benefits of systems such as pMSBOS-TS requires not only user engagement but also effective communication with stakeholders, workflow adaptation, and careful integration into existing processes.

In response to the question, "What factors are important for the successful adaptation and implementation of pMSBOS-TS in clinical practice?," participants commonly cited human and environmental factors, such as clinician trust and organizational culture, alongside functional considerations, including input/output variables, interface usability, and EMR integration. These findings align with prior research identifying such factors as critical to successful system adoption [28,29]. Given that transfusions often occur in emergencies, such as massive intraoperative bleeding or rapid clinical deterioration, participants emphasized the need for system flexibility and attention to overlapping workflows.

These insights suggest that successful implementation requires not only optimization of algorithmic performance and user-centered interface design but also enhanced user training and a supportive organizational environment. System adoption is not linear but cyclical and interactive, involving system use, feedback, and iterative refinement. Accordingly, the relationship between algorithms and workflows should be understood as bidirectional.

Furthermore, adaptation should be evaluated in stages during system implementation. Conflicting opinions regarding expected effects—such as concerns about increased confirmation steps or confusion during blood withdrawal or preparation—reflect common apprehensions in clinical settings. This underscores the need for ongoing, structured efforts to minimize unintended consequences and ensure effective integration into routine workflows.

Although participants received preinterview materials and program demonstrations, many required further clarification, particularly regarding how personalized transfusion calculations could reduce overall prescription volume. The machine learning processes underlying these calculations were frequently described as a "black box," and comprehension remained limited even with explainable AI functions. Although participants commonly perceived the model as a black box, not all machine learning approaches lack transparency; for example, tree-based or regression-based models may provide interpretable feature contributions, though such details were beyond the scope of this preimplementation evaluation [21]. This highlights the need for further research on clinician education, comprehension, and acceptance when introducing AI-driven clinical tools.

Implementing a novel AI-driven CDSS in a complex health care environment is not merely a technical deployment—it represents a sociotechnical change that reverberates through clinicians' routines, tasks, and organizational structures. A SEIPS-based evaluation captures this full spectrum of impacts by examining PETT across each stage of the workflow, enabling identification of both anticipated benefits and unanticipated drawbacks [23]. Although other analytic approaches provide useful insights, SEIPS offers a unifying systems perspective that is particularly valuable for complex, high-risk interventions such as CDSSs in clinical care. We also emphasize the importance of balancing approaches by supplementing SEIPS findings with quantitative measures from prior studies and by ensuring practical applicability through the simplified SEIPS 101 model, which facilitates stakeholder engagement [19,23].

Strengths and Limitations

This study has several limitations. It was conducted in a single country and institution, and our FGD findings—derived from the SEIPS framework and task analysis—are qualitative and context-specific, which may limit replicability. In addition, the SEIPS framework has been criticized for its complexity and limited applicability to macro-level evaluation [23]. To mitigate these concerns, we performed content validation with medical professionals from external institutions who were not involved in the interviews. Given the characteristics of tertiary hospitals, some findings may not generalize to smaller institutions with fewer personnel dedicated to transfusion management. Nevertheless, because regular transfusion-related surgeries are less common in smaller hospitals and transfusion practices involve multiple specialties, future work should include a broader range of settings and professional roles.

As a result of the FGD-based task analysis and qualitative design, recall bias and incomplete reporting are possible. Follow-up research (eg, system usage log analyses) could address these limitations; however, such data were not available for this preimplementation evaluation of a new CDSS [34]. This may represent a common constraint in predeployment CDSS impact or risk assessments. Future studies should complement qualitative findings with quantitative, log-based evaluations as these data become available.

As we targeted actively practicing medical professionals, the physicians in our sample had shorter career durations than other groups. This reflects the staffing structure of the clinical environment and is unlikely to affect the analysis of practical workflows. However, future research should examine how professional seniority may influence perceptions related to organizational decision-making and AI governance.

Given the specific nature of transfusion workflows, some requirements identified in this study may not generalize to other AI applications. Moreover, to accurately assess the impact of new systems, future evaluations should differentiate expected effects across stages of system adoption and adaptation. This limitation became apparent during our analysis and warrants attention in subsequent research. Although successful AI deployment depends heavily on the human-system interface, current literature lacks theory-driven qualitative evaluations examining how AI fits within complex sociotechnical systems.

XSL•FO
RenderX

We addressed these challenges by engaging diverse stakeholders, applying a structured sociotechnical framework, and adhering to rigorous qualitative research methods.

Overall Contribution

This study contributes to CDSS implementation research by providing one of the few preimplementation evaluations of an AI-based transfusion decision support tool. Using CQR and the SEIPS 101 framework, we captured practical concerns and anticipated impacts across physicians, nurses, and blood bank staff, offering concise, context-specific insights that complement findings from prior postdeployment CDSS studies.

Conclusions

As systems and workflows interact dynamically, it is essential to consider both tool performance and contextual adaptation. Future research should segment the adaptation process and examine interactions among users, their roles, and organizational dynamics.

Acknowledgments

This study was funded by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, supported by the Ministry of Health & Welfare, Republic of Korea (grant RS-2022-KH130250). The funder had no role in the study design, data collection, data analysis or interpretation, or manuscript preparation.

Data Sharing

The datasets generated and analyzed during this study are not publicly available due to the lack of prior consideration for data sharing during study design and Institutional Review Board approval. However, data are available from the corresponding author (YL) upon reasonable request and with approval from the Asan Medical Center Institutional Data Access/Ethics Committee.

Authors' Contributions

Conceptualization: YL, MO
Data curation: YEP
Formal analysis: YEP
Funding acquisition: YL
Investigation: YEP, YL
Methodology: YL, MO
Project administration: YL
Resources: DHK, HJL
Supervision: YL
Validation: YL, DHK, HJL, JHL, TP, JY
Visualization: YL, YEP
Writing – original draft: YEP, MO
Writing – review & editing: YL, MO
All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Demographic and professional characteristics of focus group discussion participants.
[DOCX File , 22 KB - jmir_v28i1e86166_app1.docx]

Multimedia Appendix 2
Integrated workflow for preoperative blood ordering, preparation, and transfusion in surgical care settings.
[DOCX File , 401 KB - jmir_v28i1e86166_app2.docx]

Multimedia Appendix 3
Analytical framework, including domains, subdomains, and core ideas.
[DOCX File , 26 KB - jmir_v28i1e86166_app3.docx]

References

XSL•FO
RenderX

1. Liyanage H, Liaw S, Jonnagaddala J, Schreiber R, Kuziemsky C, Terry AL, et al. Artificial intelligence in primary health care: perceptions, issues, and challenges. Yearb Med Inform 2019 Aug 25;28(1):41-46 [FREE Full text] [doi: 10.1055/s-0039-1677901] [Medline: 31022751]
2. Chen P, Lin C, Wu W. Big data management in healthcare: adoption challenges and implications. International Journal of Information Management 2020 Aug;53:102078 [FREE Full text] [doi: 10.1016/j.ijinfomgt.2020.102078]
3. Staras S, Tauscher JS, Rich N, Samarah E, Thompson LA, Vinson MM, et al. Using a clinical workflow analysis to enhance eHealth implementation planning: tutorial and case study. JMIR Mhealth Uhealth 2021 Mar 31;9(3):e18534 [FREE Full text] [doi: 10.2196/18534] [Medline: 33626016]
4. Mohan K, Ahlemann F. Understanding acceptance of information system development and management methodologies by actual users: A review and assessment of existing literature. International Journal of Information Management 2013 Oct;33(5):831-839 [FREE Full text] [doi: 10.1016/j.ijinfomgt.2013.06.003]
5. Salwei ME, Carayon P, Hoonakker PLT, Hundt AS, Wiegmann D, Pulia M, et al. Workflow integration analysis of a human factors-based clinical decision support in the emergency department. Appl Ergon 2021 Nov 01;97(4):103498-103497 [FREE Full text] [doi: 10.1016/j.apergo.2021.103498] [Medline: 34182430]
6. Paradis KC, Naheedy KW, Matuszak MM, Kashani R, Burger P, Moran JM. The fusion of incident learning and failure mode and effects analysis for data-driven patient safety improvements. Pract Radiat Oncol 2021 Jan;11(1):e106-e113. [doi: 10.1016/j.prro.2020.02.015] [Medline: 32201319]
7. Baartmans MC, Van Schoten SM, Wagner C. Generic analysis method to learn from serious adverse events in Dutch hospitals: a human factors perspective. BMJ Open Qual 2022 Feb 01;11(1):e001637 [FREE Full text] [doi: 10.1136/bmjoq-2021-001637] [Medline: 35105550]
8. Gray T, Antolak A, Ahmed S, Magnelli A, Lu L, Cho Y, et al. Implementing failure mode and effect analysis to improve the safety of volumetric modulated arc therapy for total body irradiation. Med Phys 2023 Jul 02;50(7):4092-4104. [doi: 10.1002/mp.16466] [Medline: 37265031]
9. Charles D, King J, Patel V, Furukawa M. Adoption of electronic health record systems among U.S. non-federal acute care hospitals. In: ASTP Health IT Data Brief. Washington, DC: Office of the National Coordinator for Health Inforamtion Technology; 2013.
10. Woldemariam MT, Jimma W. Adoption of electronic health record systems to enhance the quality of healthcare in low-income countries: a systematic review. BMJ Health Care Inform 2023 Jun 12;30(1):e100704 [FREE Full text] [doi: 10.1136/bmjhci-2022-100704] [Medline: 37308185]
11. Office of the National Coordinator for Health IT. Health IT legislation. Health IT. URL: https://www.healthit.gov/topic/laws-regulation-and-policy/health-it-legislation [accessed 2024-09-05]
12. Vos JFJ, Boonstra A, Kooistra A, Seelen M, van Offenbeek M. The influence of electronic health record use on collaboration among medical specialties. BMC Health Serv Res 2020 Jul 22;20(1):676 [FREE Full text] [doi: 10.1186/s12913-020-05542-6] [Medline: 32698807]
13. National trends in hospital and physician adoption of electronic health records. Health IT. URL: https://www.healthit.gov/data/quickstats/national-trends-hospital-and-physician-adoption-electronic-health-records [accessed 2024-09-05]
14. Stewart C. Global clinical decision support systems (CDSS) market in 2018, and a forecast for 2028. Statista. 2021. URL: https://www.statista.com/statistics/871216/cdss-market-value-worldwide/?srsltid=AfmBOooY6zdqBtqWh9YjFwuaTFYY9MfnYNYXbBgOCAfxvFQ7KKIqKHqT [accessed 2024-09-05]
15. Amjad A, Kordel P, Fernandes G. A review on innovation in healthcare sector (telehealth) through artificial intelligence. Sustainability 2023 Apr 14;15(8):6655 [FREE Full text] [doi: 10.3390/su15086655] [Medline: 26701262]
16. Hak F, Guimarães T, Santos M. Towards effective clinical decision support systems: a systematic review. PLoS One 2022 Aug 15;17(8):e0272846 [FREE Full text] [doi: 10.1371/journal.pone.0272846] [Medline: 35969526]
17. Liberati EG, Ruggiero F, Galuppo L, Gorli M, González-Lorenzo M, Maraldi M, et al. What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. Implement Sci 2017 Sep 15;12(1):113 [FREE Full text] [doi: 10.1186/s13012-017-0644-2] [Medline: 28915822]
18. Pyo J, Lee W, Choi EY, Jang SG, Ock M. Qualitative research in healthcare: necessity and characteristics. J Prev Med Public Health 2023 Jan;56(1):12-20 [FREE Full text] [doi: 10.3961/jpmph.22.451] [Medline: 36746418]
19. Hur S, Yoo J, Min JY, Jeon YJ, Cho JH, Seo JY, et al. Development, validation, and usability evaluation of machine learning algorithms for predicting personalized red blood cell demand among thoracic surgery patients. Int J Med Inform 2024 Nov;191:105543 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105543] [Medline: 39084087]
20. Newton N, Bamgboje-Ayodele A, Forsyth R, Tariq A, Baysari MT. A systematic review of clinicians' acceptance and use of clinical decision support systems over time. NPJ Digit Med 2025 May 26;8(1):309 [FREE Full text] [doi: 10.1038/s41746-025-01662-7] [Medline: 40419669]
21. Hur S, Lee Y, Park J, Jeon YJ, Cho JH, Cho D, et al. Comparison of SHAP and clinician friendly explanations reveals effects on clinical decision behaviour. NPJ Digit Med 2025 Sep 26;8(1):578 [FREE Full text] [doi: 10.1038/s41746-025-01958-8] [Medline: 41006498]

XSL•FO
RenderX

22. Carayon P, Schoofs Hundt A, Karsh B, Gurses AP, Alvarado CJ, Smith M, et al. Work system design for patient safety: the SEIPS model. Qual Saf Health Care 2006 Dec;15 Suppl 1(Suppl 1):i50-i58 [FREE Full text] [doi: 10.1136/qshc.2005.015842] [Medline: 17142610]
23. Holden RJ, Carayon P. SEIPS 101 and seven simple SEIPS tools. BMJ Qual Saf 2021 Nov 26;30(11):901-910 [FREE Full text] [doi: 10.1136/bmjqs-2020-012538] [Medline: 34039748]
24. Cho I. Frameworks for evaluating the impact of safety technology use. Healthc Inform Res 2023 Apr;29(2):89-92 [FREE Full text] [doi: 10.4258/hir.2023.29.2.89] [Medline: 37190732]
25. Dubé M, Hron JD, Biesbroek S, Chan-MacRae M, Shearer A, Landi R, et al. Human factors and systems simulation methods to optimize peri-operative EHR design and implementation. Adv Simul (Lond) 2025 Apr 23;10(1):23 [FREE Full text] [doi: 10.1186/s41077-025-00349-z] [Medline: 40269997]
26. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
27. Im D, Pyo J, Lee H, Jung H, Ock M. Qualitative research in healthcare: data analysis. J Prev Med Public Health 2023 Mar;56(2):100-110 [FREE Full text] [doi: 10.3961/jpmph.22.471] [Medline: 37055353]
28. Laka M, Milazzo A, Merlin T. Factors that impact the adoption of clinical decision support systems (CDSS) for antibiotic management. Int J Environ Res Public Health 2021 Feb 16;18(4):1901 [FREE Full text] [doi: 10.3390/ijerph18041901] [Medline: 33669353]
29. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
30. Saitwal H, Feng X, Walji M, Patel V, Zhang J. Assessing performance of an electronic health record (EHR) using cognitive task analysis. Int J Med Inform 2010 Jul;79(7):501-506. [doi: 10.1016/j.ijmedinf.2010.04.001] [Medline: 20452274]
31. Singprasong R, Eldabi T. An integrated methodology for process improvement and delivery system visualization at a multidisciplinary cancer center. J Healthc Qual 2013;35(2):24-32. [doi: 10.1111/j.1945-1474.2011.00174.x] [Medline: 22092497]
32. Russ AL, Militello LG, Glassman PA, Arthur KJ, Zillich AJ, Weiner M. Adapting cognitive task analysis to investigate clinical decision making and medication safety incidents. J Patient Saf 2017 May 3;15(3):191-197. [doi: 10.1097/pts.0000000000000324]
33. Graham LA, Gray C, Wagner TH, Illarmo S, Hawn MT, Wren SM, et al. Applying cognitive task analysis to health services research. Health Serv Res 2023 Apr 09;58(2):415-422 [FREE Full text] [doi: 10.1111/1475-6773.14106] [Medline: 36421922]
34. Zheng K, Ratwani RM, Adler-Milstein J. Studying workflow and workarounds in electronic health record–supported work to improve health system performance. Annals of Internal Medicine 2020 Jun 02;172(11_Supplement):S116-S122. [doi: 10.7326/m19-0871]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
COREQ: Consolidated Criteria for Reporting Qualitative Research
CQR: consensual qualitative research
EHR: electronic health record
EMR: electronic medical record
FGD: focus group discussion
IRB: institutional review board
IT: information technology
PETT: People, Environment, Tools, and Tasks
pMSBOS-TS: personalized Maximum Surgical Blood Order Schedule—Thoracic Surgery
SEIPS: Systems Engineering Initiative for Patient Safety

XSL•FO
RenderX

Edited by A Stone; submitted 20.Oct.2025; peer-reviewed by S Waldvogel Abramowski, CJ Engstrom; comments to author 20.Nov.2025; revised version received 22.Dec.2025; accepted 24.Dec.2025; published 23.Jan.2026.

Please cite as:
Park YE, Ock M, Lee JH, Ko DH, Lee HJ, Park T, Yoo J, Lee Y
Assessing Health Care Professionals' Perceptions of a New System in Clinical Workflows: Systems Engineering Initiative for Patient Safety–Based Consensual Qualitative Research
J Med Internet Res 2026;28:e86166
URL: https://www.jmir.org/2026/1/e86166
doi:10.2196/86166
PMID:

©Ye-Eun Park, Minsu Ock, Jae-Ho Lee, Dae-Hyun Ko, Hak-Jae Lee, Taezoon Park, Junsang Yoo, Yura Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Machine Learning for Predicting Stroke Risk Stratification Using Multiomics Data: Systematic Review

Hae Young Yoo[1], RN, PhD; Hyerim Shin[1], RN, BSN; Eun-Jung Kim[1], RN, MSN; Youn-Jung Son[1], RN, PhD, CCAPN, FAAN

Chung-Ang University, Seoul, Republic of Korea

Corresponding Author:
Youn-Jung Son, RN, PhD, CCAPN, FAAN
Chung-Ang University
84 Heukseok-roDongjak District, Red Cross College of Nursing
Seoul, 06974
Republic of Korea
Phone: 82 2 820 5198
Fax: 82 2 824 7961
Email: yjson@cau.ac.kr

Abstract

Background: Stroke is a complex, multidimensional disorder influenced by interacting inflammatory, immune, coagulation, endothelial, and metabolic pathways. Single-omics approaches seldom capture this complexity, whereas multiomics techniques provide complementary insights but generate high-dimensional and correlated feature spaces. Machine learning (ML) offers strategies to manage these challenges; however, the predictive accuracy and reproducibility of multiomics-based ML models for stroke remain poorly characterized.

Objective: This review aimed to conduct a systematic evaluation of ML models using multiomics data for stroke risk stratification and comprehensive patterns in discriminatory performance, integration strategies, and validation and reporting practices to inform future methodological development.

Methods: We conducted a comprehensive literature search following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 recommendations. Studies published from January 2000 to July 2025 were identified across 9 databases, including PubMed, MEDLINE Ultimate, EMBASE, CINAHL, Web of Science, Scopus, Cochrane CENTRAL, ACM Digital Library, and IEEE Xplore. Eligible studies included adults with ischemic, hemorrhagic, or unspecified stroke as the prediction target; applied at least 2 omics layers; and reported ML performance metrics. Risk of bias was assessed using the Prediction Model Risk of Bias Assessment Tool, while reporting quality was evaluated using Minimum Information for Medical AI Reporting. The primary outcome was the area under the receiver operating characteristic curve.

Results: A total of 7 studies (n=40,274) published between 2022 and 2025 fulfilled the inclusion criteria. All studies combined 2 omics layers, most often using middle-level integration with dyads such as metabolomics-proteomics and metabolomics-lipidomics. Supervised ML algorithms across studies included support vector machines, tree-based ensembles, generalized linear models, and deep learning architectures. Three studies reported external validation of the integrated multiomics model, while 1 study conducted only an external assessment of a single marker rather than validation of the integrated model. Three studies reported an assessment of calibration, and clinically prespecified operating points were rarely described. Reported areas under the receiver operating characteristic curve varied by prediction task, ranging from 0.75 to 0.96 for acute diagnosis models and from 0.75 to 0.97 for onset risk prediction models; the highest externally validated performance was achieved by a support vector machine trained on a metabolomics-proteomics dyad in mixed stroke types (ischemic and hemorrhagic).

Conclusions: Multiomics ML models showed high apparent discrimination for stroke risk stratification, but current evidence remains methodologically limited. Small sample sizes, heterogeneous designs, and incomplete reporting currently hinder the reproducibility and generalizability of multiomics ML models for stroke risk prediction. To advance the field, future studies should adopt leakage-resistant evaluation frameworks, conduct site-specific external validations, and benchmark against both single-omics and clinical baselines to demonstrate incremental value. Well-designed, transparently reported investigations will be essential to move multiomics ML models from exploratory promise toward clinically actionable tools in precision stroke care.

(J Med Internet Res 2026;28:e85654) doi:10.2196/85654

XSL•FO
RenderX

KEYWORDS

deep learning; epigenomics; genomics; lipidomics; machine learning; metabolomics; ML; multiomics; proteomics; risk stratification; stroke; transcriptomics

Introduction

Stroke is defined by acute neurological deficits resulting from vascular brain injury [1,2]. Globally, this condition is responsible for about 6.6 million deaths annually, making it the second leading cause of mortality. Its contribution to disability-adjusted life years is anticipated to rise considerably by 2050 [3]. As stroke is associated with long-term functional impairment, cognitive decline, and a high risk of recurrence, early risk stratification remains critical [3-6]. Multiple layers of interacting risk factors shape the occurrence and progression of stroke [7]. The prediction of stroke occurrence and clinical course is influenced by a number of factors at the molecular level, including genetic predisposition, epigenetic alterations, chronic inflammatory responses, and metabolic abnormalities [8-13]. Although this multilayered pathophysiology helps explain heterogeneity in etiological classification, its complexity makes it difficult to capture fully using only clinical variables or a single biomarker [13-17].

Recently, multiomics approaches have been used to characterize cross-layer relationships—genetic, epigenetic, transcriptomic, proteomic, metabolomic, and lipidomic—within integrated analytic frameworks [13,15,18,19]. This reflects the inherently multilayered biology of stroke, in which inflammatory, immune, coagulation, and metabolic axes operate concurrently [17,20]. Prior evidence indicates that multiomics strategies may outperform single-omics approaches in clinical applications, including stroke subtype classification, risk prediction, and acute diagnosis [19,21]. Simultaneously, machine learning (ML) has proven particularly effective for analyzing multiomics data, which are marked by high dimensionality and nonlinearity [22,23]. ML enables the flexible integration of clinical variables with omics layers, supports reproducible analytic pipelines, and strengthens the performance and generalizability of prediction models [23,24]. Accordingly, there is increasing interest in developing precision ML models that leverage multidimensional biological data.

Although studies applying multiomics and ML to stroke have increased rapidly, substantial heterogeneity is evident across investigations in terms of the types of omics incorporated, integration strategies, prediction end points, predictor variables, evaluation metrics, disease spectra, validation methods, and sample sizes [25,26]. This variability limits cross-model comparability and complicates assessments of clinical applicability [27]. Nevertheless, comprehensive comparative and quantitative evaluations of the relative effectiveness of these approaches for stroke prediction remain scarce, and no systematic review synthesizing this evidence has been published to date.

This study aimed to perform a systematic analysis of ML-based investigations using multiomics data for stroke risk stratification and identify comprehensive performance patterns across the included studies to inform future methodological development.

Methods

Study Design

This review was conducted and reported in line with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 recommendations (Multimedia Appendix 1) [28]. Prior to conducting the review, the study protocol was prospectively registered in the PROSPERO (International Prospective Register of Systematic Reviews; registration number CRD420251089823).

Data Sources and Search Strategy

A systematic search was conducted in PubMed, MEDLINE Ultimate (EBSCOhost), EMBASE, CINAHL, Web of Science, Scopus, Cochrane CENTRAL, ACM Digital Library, and IEEE Xplore for records published between January 1, 2000, and July 31, 2025. The start date was chosen a priori to capture early applications of omics-based prediction and to avoid arbitrarily excluding potentially relevant studies at the interface of omics and ML. Because multiomics and ML terms are inconsistently indexed across databases, we deliberately adopted a highly sensitive search strategy using multiple overlapping biomedical and multidisciplinary databases. Search strategies combined controlled vocabulary and free-text terms related to stroke, omics, and ML. Controlled vocabularies included Medical Subject Headings, Emtree, and CINAHL Headings; database-specific syntax and Boolean operators were adapted to each platform. Full search details for all databases are available in Multimedia Appendix 2. Gray literature—major preprint servers, including medRxiv, bioRxiv, and Google Scholar—were searched to check for any potentially missed or in-press studies; however, no additional records meeting the inclusion criteria were identified (final search date: July 31, 2025).

Inclusion and Exclusion Criteria

Study eligibility was determined through the PICO (participants, interventions, comparisons, and outcomes) framework and is summarized in Table 1. Studies were included if they (1) enrolled adults (aged ≥18 years) with stroke or at risk of stroke (ischemic, hemorrhagic, or unspecified), (2) used multiomics data to develop ML models, and (3) reported model performance for at least 1 stroke-related prediction outcome. In this review, stroke-related outcomes included incident stroke risk, acute stroke diagnosis among patients with suspected stroke, and etiological subtype classification. In addition, we adopted a pragmatic, prediction-oriented definition of ML. Eligible models included supervised algorithms that used multiomics data as input to predict the prespecified outcomes. Studies were excluded if they targeted combined conditions rather than stroke alone, did not aim to predict stroke, substituted nonomics data for molecular omics, applied only omics or only ML without integration, or evaluated treatment effects. Conventional regression analyses conducted exclusively for explanatory or associative purposes without a specified predictive task were

XSL•FO
RenderX

excluded, as they were not considered ML. The publication period (January 1, 2000, to July 31, 2025) was chosen to capture the evolution of multiomics- and ML-based stroke risk prediction from its early emergence to the most recent developments.

Table 1. Eligibility criteria for the systematic review using the PICO (participant, intervention, comparison, and outcomes) framework.

Item	Inclusion criteria	Exclusion criteria
Participants	Studies enrolling adults (aged ≥18 years) with stroke or at risk of stroke (ischemic, hemorrhagic, or unspecified)	• Studies involving children or adolescents (aged <18 years) • Studies restricted to demographic-only strata (eg, only men) • Studies focusing on combined conditions (eg, stroke with dementia or atrial fibrillation) rather than stroke alone
Intervention	Studies using multiomics data (genomics, transcriptomics, proteomics, metabolomics, etc) to develop ML[a] models for stroke prediction	• Studies that do not aim to predict stroke • Studies applying only omics or only ML, without integration of both • Studies evaluating treatment effects (eg, drugs, procedures, or interventions) • Studies using nonomics data (eg, imaging data such as CTb, MRIc, EEGd, or ultrasound) instead of molecular omics for stroke prediction
Comparison	No restrictions based on comparison	• No restrictions based on comparison
Outcomes	Studies reporting prediction performance of ML models for stroke-related outcomes	• Studies that do not report prediction performance
Language	No restrictions based on language	• No restrictions based on language
Publication period	Studies published from January 1, 2000, to July 31, 2025	• Studies published before January 1, 2000
Study designs	Original research articles, including observational studies, retrospective or prospective cohort studies, secondary data analyses, or basic science studies	• Reviews, meta-analyses, protocols, editorials, or nonhuman studies • Feasibility-only pilot studies without full model development or performance reporting
Publication types	Full published original articles in peer-reviewed journals	• Abstract-only records, letters, withdrawn papers, or documents without full text

[a]ML: machine learning.

[b]CT: computed tomography.

[c]MRI: magnetic resonance imaging.

[d]EEG: electroencephalogram.

Study Selection, Data Extraction, and Synthesis

Records retrieved from all sources were imported into EndNote 21 (Clarivate Plc) for deduplication and screening management, with tracking and extraction maintained in Microsoft Excel 2021 (Microsoft Corp). Titles and abstracts were independently screened in duplicate by 2 reviewers (HYY and HS) based on predefined eligibility criteria. Full texts of potentially eligible reports and risk of bias were also assessed independently in duplicate by 2 reviewers (HYY and HS). Disagreements between reviewers were resolved through discussion and, when necessary, adjudicated by a third reviewer (Y-JS). Data were extracted independently by 2 authors (HYY and HS) using a standardized piloted form and cross-verified by a third author (Y-JS). Extracted items included study title, authors, publication year, country, multiomics studied, sample size, data source, omics type, stroke type, prediction task, omics analytic methods, ML algorithms, outcome metrics and summary results, and validation details (internal/external and strategy). Model performance metrics were extracted and summarized descriptively. The primary measure of discrimination was the area under the receiver operating characteristic curve (AUC) with 95% CIs.

When discrimination was reported using Harrell C statistic for a binary outcome rather than a receiver operating characteristic (ROC)–based AUC, we treated the C statistics as numerically equivalent to the AUC for the purpose of pooling [29,30]. Publication bias is typically assessed when at least 10 studies are available; because fewer studies met the inclusion criteria in this review, publication bias was not formally evaluated, and no funnel plots were generated [31].

A meta-analysis was not performed because the limited number of eligible studies and the substantial between-study heterogeneity precluded meaningful quantitative synthesis. The included studies varied markedly in study design (retrospective vs prospective), sample size, and data sources (single-center cohorts, community-based datasets, and biobank data). In addition, the prediction tasks differed considerably, encompassing stroke onset risk prediction, acute diagnosis, and subtype classification. Given this conceptual, methodological, and statistical heterogeneity, quantitative pooling of effect estimates was judged inappropriate. Instead, we conducted a structured narrative synthesis to systematically summarize study characteristics, omics layers evaluated, ML approaches applied, and clinical prediction targets.

XSL•FO
RenderX

Quality and Risk of Bias Assessment

Quality appraisal and risk of bias assessment were conducted using 2 frameworks suited to prediction modeling and medical artificial intelligence (AI) reporting: the Prediction Model Risk of Bias Assessment Tool (PROBAST) [32], and the Minimum Information for Medical AI Reporting (MINIMAR) [33]. PROBAST was used for the primary risk of bias classification, and was applied across the participants, predictors, outcomes, and analysis domains, with judgments at both the domain and overall levels categorized as low risk, high risk, or unclear [32]. MINIMAR was adopted as the checklist for AI and ML studies, as it represents an expert-developed framework suited to evaluating reporting quality in omics-based ML research, particularly in early translational health care contexts [33,34]. Each item was coded Y (adequate), P (partial), N (not reported), or X (not applicable), and compliance (%) was calculated as the proportion of Y among applicable items.

Ethical Considerations

The institutional review board of Chung-Ang University (number 1041078-20250715-HR-225) approved the study protocol.

Results

Literature Search Results

The database search across 9 sources yielded 2960 records. Of these, 881 duplicates were excluded, leaving 2079 records to be screened. During title and abstract screening, 1413 records were excluded. Full texts were obtained for 666 reports, of which 530 were excluded as off-topic following full-text assessment.

The remaining 136 studies underwent full eligibility evaluation. Of these, 129 were excluded for the following reasons: insufficient sample size (<30; n=5), not a stroke prediction study (n=25), not multiomics (n=70), outcome outside scope (n=6), full text unavailable (n=1), and failure to meet inclusion criteria on full-text review (n=22). A total of 7 studies were identified as meeting the inclusion criteria and were subsequently analyzed in this review (Figure 1) [35-41]. Gray literature searches yielded no additional records; therefore, the PRISMA flow diagram reflects only the results from database searches.

XSL•FO
RenderX

Figure 1. PRISMA flow diagram of study selection for systematic review of multiomics-based machine learning models for stroke prediction.

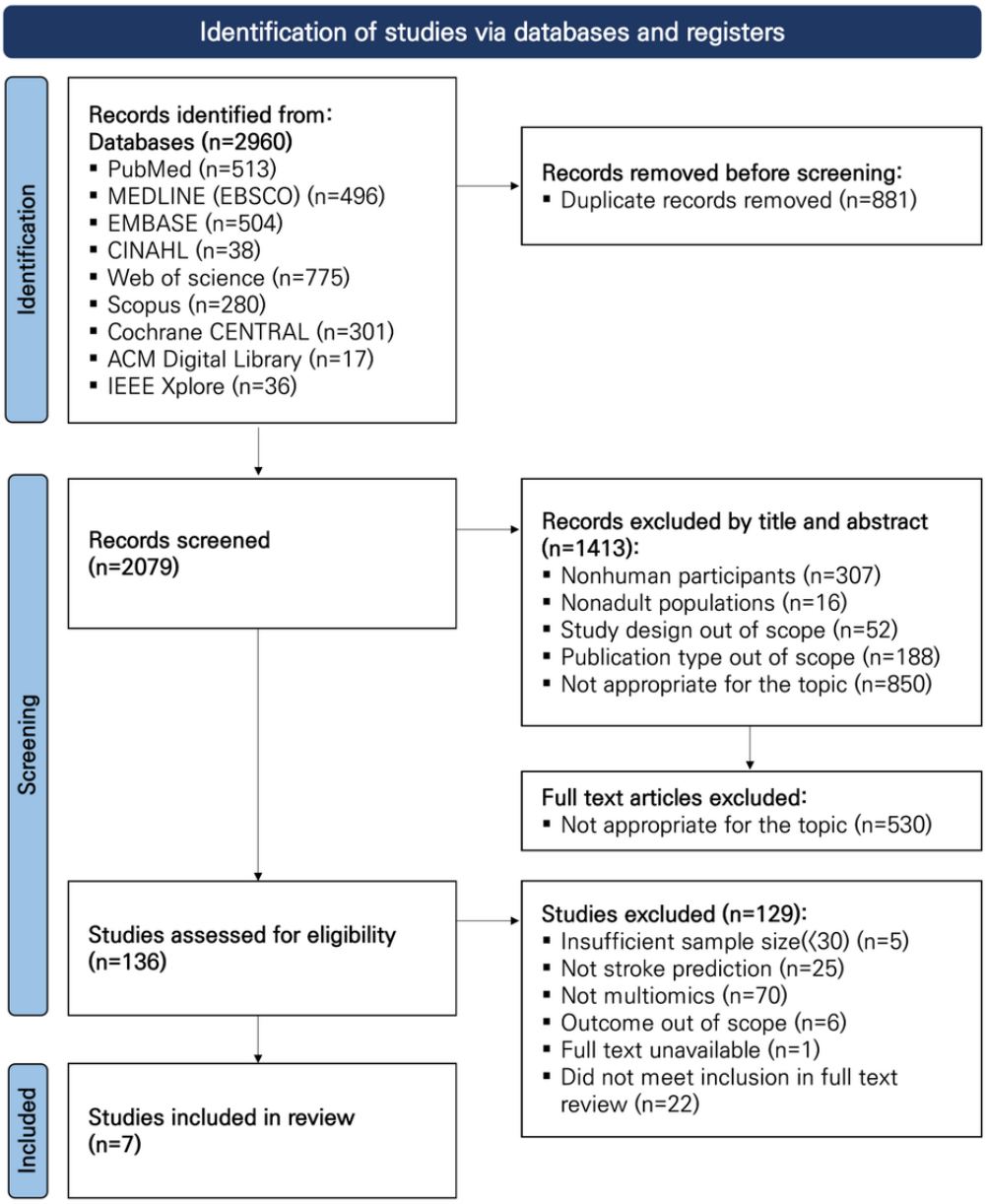

Quality Assessment of the Included Studies

Study appraisals are summarized in Multimedia Appendix 3. On PROBAST, 4 studies were rated as low risk of bias [36,37,40,41], 2 studies as high risk [38,39], and 1 study as unclear overall [35]. Most concerns were concentrated in the analysis domain, whereas participants, predictors, and outcomes were generally judged to be at low risk. Applicability issues were minimal across the included studies. On MINIMAR, overall reporting compliance ranged from 11 of 21 items [35] to 17 of 21 items [41], with a median value of 14 of 21. Information on demographic characteristics (race, ethnicity, and socioeconomic status), intended users, data missingness, external validation, and data transparency was often insufficiently reported or entirely absent.

Characteristics of the Included Studies

A total of 7 studies published between 2022 [38] and 2025 [36,41] were included in this review (Table 2). The studies represented 3 countries: China (n=4) [36-38,40], Spain (n=1) [39], and the United Kingdom (n=1; from England, Scotland, and Wales) [41]. Studies that used public repositories were classified as multinational [35,36]. Data sources were grouped into 3 categories: clinical cohorts recruited from hospitals or community health service centers (n=4) [37-40], public datasets (n=2) [35,41], and mixed sources (n=1) [36]. In the mixed-sources study [36], models were developed using gene expression omnibus datasets [42] and externally validated in independently recruited local clinical cohorts. Study designs were predominantly prospective (n=4) [37-39,41], with retrospective (n=2) [35,36], and mixed retrospective-prospective (n=1) [40]. Sample sizes varied widely, ranging from 30

XSL•FO
RenderX

participants in a hospital cohort [39] to 38,060 participants in a large public dataset [41].

Reported ML approaches included random forest, support vector machine (SVM), gradient boosting (extreme gradient boosting; adaptive boosting), generalized linear and logistic models, Bayesian logistic regression, the least absolute shrinkage and selection operator (LASSO), including the LASSO-penalized Cox proportional hazards model, and deep learning architectures such as the biological multilayer graph neural network. All studies integrated 2 omics layers, and the most frequent dyads were lipidomics-metabolomics (n=2) [37,38].

Regarding stroke type, 6 studies [35-39,41] specifically investigated ischemic stroke, whereas 1 study [40] included all stroke types. Prediction targets were grouped into categories: onset risk prediction (n=3) [35,40,41], acute diagnosis (n=2) [37,38], subtype classification (n=1) [39], and both reporting of diagnosis and subtype classification (n=1) [36].

Table 2. Characteristics of the included studies (N=7) of multiomics-based machine learning models for stroke prediction. The table summarizes key design features of each included study. Sample size (n) refers to the development cohort size for each study.

Study (year)	Country	Data source	Study design	Sample size, n	Machine learning algorithms	Omics types	Prediction task	
							Stroke type	Target
Liu et al [35] (2024)	Multinational	IEU[a] and GEO[b] database	Retrospective	103	RF[c], SVM[d], XGB[e], and GLM[f]	Genomics and transcriptomics	Ischemic stroke	Onset risk prediction
Chen et al [36] (2025)	Multinational and China	GEO database, 2022-2023 single hospital	Retrospective	1785	RF and LASSO[g]	Transcriptomics and proteomics	Ischemic stroke	Acute diagnosis; subtype classification
Zhao et al [37] (2023)	China	2017-2019 CCPTP[h]	Prospective	90	RF and BLR[i]	Lipidomics and metabolomics	Ischemic stroke	Acute diagnosis
Ye et al [38] (2022)	China	Single hospital	Prospective	106	RF, SVM, and LR[j]	Lipidomics and metabolomics	Ischemic stroke	Acute diagnosis
Labarga et al [39] (2024)	Spain	2015-2016 single hospital	Prospective	30	BioMGNN[k] and XGB	Transcriptomics and epigenomics	Ischemic stroke	Subtype classification
Zeng et al [40] (2024)	China	2010-2011 community health service centers, 2021-2023 single hospital	Prospective and retrospective	100	XGB, SVM, GNB[l], RF, AdaBoost[m], and L1-regularized LR	Metabolomics and proteomics	Stroke	Onset risk prediction
Gan et al [41] (2025)	England, Scotland, and Wales	UKB[n]	Prospective	38,060	LASSO-Cox[o]	Proteomics and genomics	Ischemic stroke	Onset risk prediction

[a]IEU: integrative epidemiology unit.
[b]GEO: gene expression omnibus.
[c]RF: random forest.
[d]SVM: support vector machine.
[e]XGB: extreme gradient boosting.
[f]GLM: generalized linear model.
[g]LASSO: least absolute shrinkage and selection operator regression.
[h]CCPTP: community comprehensive prevention and treatment project.
[i]BLR: binary logistic regression.
[j]LR: logistic regression.
[k]BioMGNN: biological multilayer graph neural network.
[l]GNB: Gaussian naïve bayes.
[m]AdaBoost: adaptive boosting.
[n]UKB: UK biobank.
[o]LASSO-Cox: least absolute shrinkage and selection operator-penalized Cox proportional hazards model.

Omics Features and Integration Strategies in Multiomics Stroke Modeling

Table 3 summarizes the characteristics of omics data, their analytic roles, and the integration strategies applied in each study. Omics type lists all molecular layers included per study, and biospecimen records the biological source of each layer as reported. The analysis unit indicates the level of the final features entered into the model.

XSL•FO
RenderX

Table 3. Omics features and integration strategies for multiomics stroke modeling in the included studies (n=7).

Study (year) and omics data type	Biospecimen	Analysis unit	Top-ranked feature biomarkers	Role in pipeline	Feature selection method	Integration strategy
Liu et al [35] (2024)						Middle integration
Genomics	GWAS[a] summary data	SNP[b] to gene-level	*FURIN (OMIM 136950), TOMM40 (OMIM 608061), HDDC3 (NCBI Gene 374659), ALDH2 (OMIM 100650), MAN2A2 (OMIM 600988)*	FS[c], FP[d], RS[e]	LD[f] clumping with genome-wide significance and F-stat	
Transcriptomics	Whole-blood mRNA[g]	Gene-level expression	*FURIN, TOMM40, HDDC3, ALDH2, MAN2A2*	FS, FP, VAL[h]	DEAGs[i]	
Chen et al [36] (2025)						Late integration
Transcriptomics	Serum miRNA[j], whole-blood mRNA	miRNA expression, gene-level mRNA expression	hsa-miR-3646, hsa-miR-4669, hsa-miR-4721, hsa-miR-504-3p, hsa-miR-8085, hsa-miR-3064-5p, hsa-miR-6736-5p, hsa-miR-6835-5p, hsa-miR-6793-5p	FS, FP	DEMs[k], ML[l] selection	
Proteomics	Plasma protein	Protein level	ITGAM[m]	FS, VAL	Cross-omics prioritization, targeted validation	
Zhao et al [37] (2023)						Middle integration
Lipidomics	Serum	LC-MS[n] untargeted metabolite features	PE[o] (18:0p/18:2), PE (16:0e/20:4), OAHFA[p] (36:3), PE (16:0p/20:3), PE (18:1p/18:2)	FS, FP, VAL	PLS-DA[q], ML selection	
Metabolomics	Serum	LC-MS untargeted lipid species	4-hydroxyphenylpyruvic acid, cafestol	FS, FP, VAL	PLS-DA, ML selection	
Ye et al [38] (2022)						Middle integration
Lipidomics	Serum	LC-MS untargeted metabolite features	PE (22:6/P-18:0), Cer[r] 34:2, GlcCer[s] (d18:0/18:0), DG[t] 44:0, LysoPC[u] (16:0), 22:6-OH[v]/LysoPC, TAG[w] 58:7-FA[x] 22:4	FS, FP	OPLS-DA[y], receiver operating characteristic-based evaluation	
Metabolomics	Serum	LC-MS untargeted lipid species, targeted lipid species	Taurine, oleoylcarnitine, creatinine	FS, FP	OPLS-DA, receiver operating characteristic-based evaluation	
Labarga et al [39] (2024)						Middle integration
Transcriptomics	Whole-blood mRNA, miRNA, circRNA[z]	Gene-level expression, miRNA expression, circRNA expression	Not reported (high-dimensional omics features; candidate biomarkers mentioned but no explicit ranking provided)	FS, FP, VAL	GraphSAGE[aa]-based embedding, attention weight	
Epigenomics	Whole-blood DNA methylation	CpG[ab] β-values	Not reported (high-dimensional omics features; candidate biomarkers mentioned but no explicit ranking provided)	FS, FP, VAL	Probe-level quality control, GraphSAGE-based embedding, attention weight	
Zeng et al [40] (2024)						Early integration

XSL•FO
RenderX

Study (year) and omics data type	Biospecimen	Analysis unit	Top-ranked feature biomarkers	Role in pipeline	Feature selection method	Integration strategy
Metabolomics	Plasma	LC-MS untargeted metabolite features	Caprolactam	FS, FP, VAL	OPLS-DA, ML selection	
Proteomics	Plasma	DIA[ac] protein groups	C4BPA[ad], COL15A1[ae], HBB[af]	FS, FP, VAL	DEPs[ag], ML selection, cross-model intersection	
Gan et al [41] (2025)						Late integration
Proteomics	Plasma	Antibody-based proteomics analytes	17-protein score (GDF15[ah], PLAUR[ai], NT-proBNP[aj], IGFBP4[ak], BCAN[al] +12 proteins)	FS, FP, RS, VAL	LASSO-Cox[am], 17-protein score derived	

XSL•FO
RenderX

Study (year) and omics data type	Biospecimen	Analysis unit	Top-ranked feature biomarkers	Role in pipeline	Feature selection method	Integration strategy
Genomics	Whole-blood DNA genotypes	PRS[an]	Not reported	FP, RS, VAL	Precomputed PRS	

[a]GWAS: genome-wide association study.
[b]SNP: single nucleotide polymorphism.
[c]FS: feature selection.
[d]FP: final predictors.
[e]RS: risk score.
[f]LD: linkage disequilibrium.
[g]mRNA: messenger RNA.
[h]VAL: validation.
[i]DEAGs: differentially expressed associated genes.
[j]miRNA: microRNA, microRNA of human origin.
[k]DEMs: differentially expressed miRNAs.
[l]ML: machine learning.
[m]ITGAM: integrin subunit alpha M.
[n]LC-MS: liquid chromatography-mass spectrometry.
[o]PE: phosphatidylethanolamine.
[p]OAHFA: (O-acyl)-1-hydroxy fatty acid.
[q]PLS-DA: partial least squares-discriminant analysis.
[r]Cer: ceramide.
[s]GlcCer: glucosylceramide.
[t]DG: diacylglycerol.
[u]LysoPC: lysophosphatidylcholine.
[v]OH: hydroxyl group.
[w]TAG: triacylglycerol.
[x]FA: fatty acid.
[y]OPLS-DA: orthogonal partial least squares-discriminant analysis.
[z]circRNA: circular RNA.
[aa]GraphSAGE: graph sample and aggregate.
[ab]CpG: cytosine phosphate guanine.
[ac]DIA: data-independent acquisition.
[ad]C4BPA: complement component 4 binding protein alpha.
[ae]COL15A1: collagen type XV alpha 1 chain.
[af]HBB: hemoglobin subunit beta.
[ag]DEPs: differentially expressed proteins.
[ah]GDF15: growth differentiation factor 15.
[ai]PLAUR: plasminogen activator, urokinase receptor.
[aj]NT-proBNP: N-terminal pro b-type natriuretic peptide.
[ak]IGFBP4: insulin-like growth factor binding protein 4.
[al]BCAN: brevican.
[am]LASSO-Cox: least absolute shrinkage and selection operator-penalized Cox proportional hazards model.
[an]PRS: polygenic risk score.

Top-ranked feature biomarkers represent the features or panels identified by authors as having the highest contributions. Role in pipeline indicates how each biomarker was used and is coded as feature selection, final predictors, validation, or risk score. Multiple codes may apply when roles differ across study phases. The feature selection method summarizes the dominant selector applied before model fitting. Integration method classifies how multiomics layers were combined, following established taxonomies [43,44]: early integration (simple concatenation of preprocessed omics into a single matrix for a downstream learner), middle integration (a joint latent representation learned directly across omics followed by modeling), and late integration (separate omics-specific models or scores combined at the decision or score level).

Across the 7 included studies, multiomics inputs comprised transcriptomics (n=3) [35,36,39], proteomics (n=3) [36,40,41], metabolomics (n=3) [37,38,40], lipidomics (n=2) [37,38], epigenomics (n=1) [39], and genomics (n=2) [35,41] (Table 3). Biospecimens were predominantly blood derived—whole blood for messenger RNA, circular RNA expression, and

XSL•FO
RenderX

DNA-methylation; serum or whole blood for micro RNA (miRNA) profiling; and serum or plasma for protein panels and liquid chromatography–mass spectrometry (LC-MS)–based metabolite and lipid features. Genomic data were sourced from genome-wide association study summary statistics in 1 study [35] and from whole blood DNA genotypes in another [41].

Most studies (n=6) reported top-ranked biomarkers or panels. By contrast, 1 study [39] used deep learning–based representation learning to process high-dimensional transcriptomic and epigenomic inputs directly as raw features, subsequently deriving multiomics embeddings from the full feature space. Consequently, no ranked lists of biomarkers were generated in this study. Regarding analytic roles, omics variables were consistently used for feature selection and as final predictors. A dedicated validation step—internal and/or external—was described in 5 studies [36-40], while a risk score stage was incorporated in 2 studies [35,41].

Given the heterogeneity of feature selection approaches, Table 3 provides a concise method label for each study. Regarding integration strategies, most studies used middle-level integration (n=4) [35,37-39], while late-level integration was reported in 2 studies [36,41] and early-level integration in 1 study [40]. Middle integration was implemented in three main forms: (1) cross-omics filtering, in which genome-wide association study signals were intersected with transcriptomic differentially expressed genes to define the feature set for ML; (2) projection-based selection, where lipidomic and metabolomic data were reduced to compact multiomics panels; and (3) graph-based joint embeddings. Late integration was used in two distinct ways: (1) validation-stage fusion, in which transcriptomic discoveries were linked with an independent proteomic readout to confirm cross-omics consistency; and (2) score-level fusion, where a protein risk score and a polygenic risk score were jointly incorporated into prediction models.

A study [40] used a mixed stroke end point combining ischemic and hemorrhagic stroke. Because this outcome definition merges etiologically distinct conditions, the identified features are interpreted as mixed-population signals rather than subtype-specific molecular biomarkers. Methodologically, the study performed early integration by analysis-specific quality control and univariate screening, followed by concatenation of the retained features into a single matrix for common selection (LASSO) and modeling (eg, SVM).

Model Performance of Risk Prediction for Stroke

Predictive performance is summarized in Table 4. Discrimination was characterized primarily using AUC with 95% CIs; when available, sensitivity, specificity, validation method, and calibration details were also extracted. The representative performance shown for each study does not necessarily correspond to the numerically highest AUC. Primary estimates were selected to align with the objectives of this review. Multiomics integration and the use of clinically comparable control groups were prioritized as key criteria, and external validation results were adopted whenever available. If external validation was not reported, the best internally validated AUC was recorded. For Gan et al [41], the proteomics plus polygenic risk score specification was designated as the primary estimate because it most reflected multiomics integration, consistent with the inclusion criteria of this review. Further details are provided in the footnotes of Table 4.

XSL•FO
RenderX

Table 4. Predictive performance of multiomics-based machine learning models for stroke in adult population (N=7). For the primary model selected from each study, the table summarizes omics final predictors, validation approach, and key performance metrics. When external validation was reported, "n" indicates the sample size of the external validation cohort, reported separately from the development cohort.

Study (year)	Final predictors, n	Cross-validation	External validation	Calibration	Machine learning model performance			
					Model	AUC[a] (95% CI)	Sensitivity	Specificity
Liu et al [35] (2024)	5	Not reported	Yes (n=12)	Yes[b]	GLM[c]	1.000 (1.000-1.000)	Not reported	Not reported
Chen et al [36] (2025)	1[d]	10-fold	Partial (n=44)	Not reported	LASSO+RF[e]	0.750 (0.601-0.899)	0.567	0.857
Zhao et al [37] (2023)	7	7-fold	No	Yes	RF+LR[f]	0.837[g] (0.732-0.942)	0.897	0.733
Ye et al [38] (2022)	10	Not reported	No	Not reported	SVM[h]	0.963 (0.884-1.000)	Not reported	0.917
Labarga et al [39] (2024)	High-dimensional omics features	5-fold	No	Not reported	BioMGNN[i]	0.950 (not reported)	0.950	Not reported
Zeng et al [40] (2024)	4	10-fold	Yes (n=44)	Yes	SVM	0.973 (0.921-0.999)	0.864	1.000
Gan et al [41] (2025)	2[j]	10-fold	Yes (n=4970)	Not reported	LASSO-Cox[k]	0.748[l] (0.707-0.789)	Not reported	Not reported

[a]AUC: area under the receiver operating characteristic curve.

[b]Calibration was reported in the Methods, but no plot or statistic was presented.

[c]GLM: generalized linear model.

[d]External AUC from integrin alpha M (ITGAM) measured in an independent clinical cohort.

[e]LASSO+RF: least absolute shrinkage and selection operator regression–based feature selection by random forest as the final classifier.

[f]RF+LR: random forest feature selection followed by logistic regression as the final classifier.

[g]Value shown is hypertensive ischemic stroke versus healthy controls.

[h]SVM: support vector machine.

[i]BioMGNN: biological multilayer graph neural network.

[j]Score-level fusion of proteomic and genetic risk scores.

[k]LASSO-Cox: least absolute shrinkage and selection operator-penalized Cox proportional hazards model.

[l]Harrell C statistic for a boundary outcome, treated as numerically equivalent to AUC for quantitative synthesis. The estimate corresponds to protein risk score + polygenic risk score (PRS) to reflect multiomics integration.

Final predictors in Table 4 refer to the predictors actually entered into the final model (eg, selected feature panel, single markers, or score-level inputs), excluding clinical covariates. Across studies, most models relied on compact omics-derived inputs, while 1 study [39] analyzed high-dimensional feature sets. For late-level integration, the final predictors correspond to the score-level [41] or marker-level [36] inputs entered into the model, rather than the underlying omics features used to construct them. Ten-fold cross-validation was the most commonly applied procedure (n=3) [36,40,41], while 7-fold [37] and 5-fold [39] schemes were also reported. Several studies [35,38] did not specify the resampling method. The algorithms most frequently applied were SVM and random forest, often coupled with LASSO for feature selection. Additional approaches included generalized linear models, the LASSO-penalized Cox proportional hazards model, and deep learning architectures such as a biological multilayer graph neural network.

A total of 3 studies [35,40,41] reported external validation of the integrated multiomics model, whereas Chen et al [36] reported an external assessment limited to a single biomarker, integrin alpha M. Three studies [35,37,40] reported calibration assessment. Although Liu et al [35] stated that calibration curves were generated, no plot or statistic was presented. Among externally validated models—excluding 1 boundary estimate (AUC=1.00) [35]—AUC values ranged from 0.75 [41] to 0.97 [40], both of which were onset risk prediction models. The only acute diagnosis model with external validation was Chen et al [36] (AUC=0.75); however, the external validation was conducted as a cross-omics single-marker assessment—integrin alpha M measured at the proteomic level—rather than validation of an integrated multiomics model. When all reported AUCs were stratified by clinical task regardless of validation type, acute diagnosis models reported AUCs ranging from 0.75 [36] to 0.96 [38], whereas onset risk prediction models reported AUCs ranging from 0.75 [41] to 0.97 [40]. Notably, the SVM model in Zeng et al [40] achieved the highest externally validated discrimination (AUC=0.97) using a metabolomics-proteomics combination. However, because this estimate was derived from a mixed stroke end point, it does not reflect performance for a specific stroke subtype. Where

XSL•FO
RenderX

reported, sensitivity and specificity at the chosen decision thresholds are summarized in Table 4. A total of 5 studies [36-40] reported sensitivity and/or specificity, and only 3 studies [36,37,40] reported both metrics at prespecified decision thresholds. Because reporting was incomplete and thresholds and reporting formats were heterogeneous across studies, attempts at separate quantitative synthesis of sensitivity or specificity were not conducted. Reporting of model performance and validation procedures varies across studies, which limited comparability and may have contributed to optimistic estimates. Only 1 study [41] evaluated model performance against standard clinical predictors. In Gan et al [41], the multiomics model had a C statistic of 0.765 in internal validation and 0.748 in external validation, whereas the clinical predictor model showed corresponding values of 0.753 and 0.754, respectively.

Discussion

Overview

This systematic review synthesized 7 studies of multiomics-based ML models for stroke-related prediction tasks. To our knowledge, it is one of the first reviews to focus specifically on multiomics ML in the context of stroke risk stratification. Although many of the included models reported apparently high discrimination, these findings arise from small, heterogeneous, and methodologically limited evidence base characterized by high-dimensional omics inputs, limited external validation, and inconsistent reporting. Accordingly, the primary contribution of this review is not to demonstrate that current multiomics models are ready for clinically meaningful or generalizable accuracy, but rather to show that the current evidence base is methodologically limited, prone to performance inflation, and in need of substantially more rigorous and transparent evaluation.

Discrimination and Methodological Quality in Multiomics-Based Machine Learning Stroke Prediction

On face value, most included multiomics ML models reported high apparent discrimination, particularly when compact biomarker panels were constructed from biologically coherent omics dyads. However, these apparently strong AUCs must be interpreted with extreme caution. Across studies, small-to-moderate sample sizes, high-dimensional inputs, and optimistic validation strategies created ideal conditions for severe overfitting and data leakage. Within this constrained context, some technical patterns were still observable. For example, a metabolomics-proteomics SVM reported an externally validated AUC of 0.97 with a boundary CI (95% CI 0.92-1.00) [40]. This finding is often interpreted as evidence that compact and mechanically coherent biomarker panels can match or exceed more complex multilayered approaches [45]. At the algorithmic level, the prominence of SVM in the best-performing models is also broadly consistent with previous meta-analytic work in cardiovascular ML, in which SVM achieved slightly higher pooled AUCs for stroke outcomes than boosting algorithms or convolutional neural networks [46]. In the specific metabolomics-proteomics SVM model in this review, the combination of a modest sample size and almost upper-limit confidence bound suggests that residual overfitting and optimism remain likely, even with external testing.

A deep learning model applied to transcriptomic-epigenomic matrix also reported an AUC approaching 0.95 [39]; however, this estimate was generated in a cohort of only 30 participants in a pronounced "p≫n" biomarker-discovery setting. This high AUC is therefore best viewed as a methodological artifact rather than a valid predictive performance. Without leakage-robust pipelines and adequately powered external validation, this result should be considered exploratory and should not be used to claim strong discrimination. In this context, the combination of high-capacity architecture, high-dimensional feature space, and limited effective sample size makes severe overfitting and information leakage not merely possible but statistically expected. In addition, the study did not report specificity at prespecified decision thresholds, which makes it difficult to relate the reported sensitivity to false-positive rates and limits the clinical interpretability of the findings [47]. The generalized linear model–based study reported an AUC of 1.00 [35]. Given the small sample size and high-dimensional feature space, this boundary estimate is more appropriately interpreted as an optimistic methodological artifact rather than valid or generalizable predictive performance [48,49]. In addition, calibration was described but not shown; the absence of explicit plots or statistics limits contextual interpretation of the apparent perfect separation. Taken together, these findings reinforce that near-ceiling discrimination estimates in this review are not reliable indicators of clinical validity. Overall, the reported performance should be viewed as preliminary and potentially optimistic given the methodological constraints of the included studies.

Observed Performance Patterns by Integration Strategy and Omics Layers

The performance of multiomics ML models appears to be influenced by both the integration strategy and the specific omics layers used. Within this small and heterogeneous set of studies, most applied middle-level integration, which offered a pragmatic balance for limited sample sizes and heterogeneous data platforms by generating joint feature representations prior to classification [50,51]. Early-level integration, while more vulnerable to dimensionality challenges, produced strong results when combined with stringent feature selection and relatively careful validation [52], as demonstrated in one of the higher-performing studies in this review. By contrast, late-level integration, implemented either as score-level fusion or cross-omics validation, reported more modest AUC values (0.75) [36,41]. Notably, Gan et al [41] used score-level fusion by combining a proteomic risk score with a polygenic risk score, rather than feature-level integration. This formulation may be less able to capture complex interomics relations [44,53]. Accordingly, several methodological studies have suggested that late-level integration may underperform because it cannot fully capture synergistic cross-layer interactions [51,52]. In this review, however, these differences cannot be attributed to integration strategy alone, because the late-integration studies also differed in prediction task, sample size, and model specification. Thus, the apparent underperformance of late-level integration in this review should be regarded as a

XSL•FO
RenderX

hypothesis-generating observation rather than a general conclusion about the inherent limitation of late-integration approaches.

Beyond integration strategies, the selection of omics layers fundamentally shaped predictive utility. Consistent with prior research, transcriptomic-epigenomic pairs were generally used to represent relatively stable regulatory states, whereas lipidomic-metabolomic dyads were used to capture pathway dynamics and acute physiological responses [54-56]. Genomic data mostly contributed a predispositional context rather than serving as direct predictive variables. Additionally, the definition of the analysis unit—for example, gene-level expression, CpG probe values, or LC-MS features—determined the imputation burden, multiplicity, and scope of feasible feature selection methods, thereby influencing both model performance and reproducibility [55,57,58]. Collectively, these observations underscore that methodological design choices in omics integration and feature representation are not simply technical decisions but pivotal determinants of translational potential.

Cross-Study Heterogeneity and Barriers to Quantitative Synthesis

Across the included studies, differences in reported performance primarily reflected variations in study design and analytic choices rather than random fluctuation. Substantial heterogeneity was observed across study populations, prediction targets, stroke end points, and analytic units. In this context, the evidence was synthesized narratively rather than summarized as a single pooled effect size.

More specifically, formal pooling of AUCs was not undertaken because it would have required combining performance estimates derived from substantially different data structures and outcome definitions. With respect to the omics inputs, some studies modeled gene- or protein-level summaries, whereas others analyzed large sets of individual probes or untargeted LC-MS features, leading to marked differences in dimensionality and in the feature-selection and validation strategies that were feasible. At the outcome level, 1 study modeled composite “all-stroke” end points that combined ischemic and hemorrhagic events, whereas others focused exclusively on ischemic stroke. Because these subtypes have distinct pathophysiology and omics profiles, mixed end points limit subtype-specific interpretation and contribute to heterogeneity across studies. Consequently, similar AUC values may reflect different biological mechanisms, case mixes, and follow-up structures, and a single pooled estimate would provide limited additional clinical or methodological insight. Overall, this pattern of heterogeneity is consistent with an early, exploratory phase of multiomics ML research in stroke and highlights the need for more standardized designs and reporting.

Comparable patterns of methodological divergence have been described in broader reviews of ML-based prediction models, in which variation in datasets, predictors, analytic pipelines, and algorithms has constrained the interpretability of pooled accuracy estimates [46,59-61]. Our findings are consistent with this broader literature: the included studies are sufficiently mature for qualitative comparison but not yet harmonized to support robust quantitative synthesis.

Implications for Practice and Future Research

For time-critical decisions in hyperacute stroke care, integrating current multiomics workflows into the minute-to-hour decision windows that govern acute management remains challenging. In particular, most acute diagnosis studies in this review relied on omics profiling that generally depends on centralized laboratory workflows (eg, LC-MS–based proteomic/metabolomic assays). Given current turnaround times, these approaches are unlikely to be clinically translatable within hyperacute treatment windows, limiting near-term clinical feasibility [62]. From a broader risk-stratification perspective, however, multiomics ML approaches may be better suited to longer time horizons, including identification of individuals at elevated future stroke risk, refinement of etiologic classification after the acute phase, and delineation of biologically homogeneous patient subgroups that are not apparent from clinical variables or imaging alone [63,64]. Any future role in acute pathways would require both stronger predictive evidence and assay platforms with substantially shorter turnaround times and lower costs.

Within this context, the current evidence does not yet provide a sufficiently robust foundation for routine implementation of multiomics-based ML models for stroke risk stratification. These models are more accurately regarded as exploratory prototypes, demonstrating potential capabilities within an optimized analytical framework. Compact, blood-derived biomarker panels constructed from biologically coherent omics dyads may offer a plausible route for future clinical application, provided that their performance can be reliably replicated and calibrated in adequately powered, prospectively designed studies using appropriate analytic platforms [65,66].

From a methodological perspective, the studies reviewed highlight critical domains requiring standardization. Repository-based cohorts provide statistical power and facilitate discovery, but they may obscure preanalytical variability and heterogeneity in investigator judgment [67,68]. By contrast, locally recruited cohorts enable clinically focused hypotheses but often suffer from limited sample sizes [69]. Furthermore, the choice of analytic unit—whether gene, CpG probe, protein, or LC-MS feature level—emerged as a critical determinant of which feature selection and integration techniques were feasible, thereby exerting a direct influence on predictive performance. Collectively, these observations underscore the imperative for harmonized analytic frameworks and transparent reporting standards to strengthen reproducibility and enable meaningful cross-study comparability.

Prospective, multisite validation of compact and biologically coherent panels across diverse populations should be a priority. Such studies must incorporate leakage-robust validation pipelines, systematically report calibration and decision-curve analyses, clinically relevant operating points (including sensitivity and specificity at prespecified thresholds), and benchmark performance against both single-omics models and established clinical baselines to quantify true incremental value [70,71]. In future work, both primary studies and evidence syntheses should avoid pooling ischemic and hemorrhagic stroke into undifferentiated “all-stroke” end points when quantifying

XSL•FO
RenderX

model performance, given their distinct pathophysiological and multiomics profiles. These observations collectively support subtype-specific development and synthesis as a more appropriate strategy for future multiomics ML research in stroke. Expanding prediction targets to encompass hemorrhagic stroke, postevent prognostic outcomes, stroke recurrence in secondary prediction context, and health-economic evaluations will also be essential to define the broader clinical utility of multiomics ML approaches. With rigorous methodological safeguards and robust external validation, these models hold the potential to transition from exploratory promise to clinically actionable tools in precision cerebrovascular medicine.

Limitations

This review has several limitations. First, although we initially planned to conduct a meta-analysis, we ultimately judged that conducting a meta-analysis would be inappropriate because the included studies differed markedly in sample size, data sources, validation strategies, and outcome definitions (eg, stroke onset risk prediction, acute diagnosis, and subtype classification); these discrepancies collectively precluded meaningful quantitative synthesis and robust pooled effect estimates. Second, external validation was conducted in only half of the studies, while calibration and clinically relevant operating points were reported in only a minority, limiting confidence in both reproducibility and clinical applicability. Third, most cohorts focused on ischemic stroke and blood-based assays, with only limited data for hemorrhagic or mixed all-stroke populations and alternative biospecimens, restricting generalizability and likely contributing further to clinical and biological heterogeneity. Fourth, reporting gaps—particularly concerning missing-data mechanisms, demographic characteristics, and preanalytical procedures—were frequent and constrained assessment of bias and transportability. Fifth, differences in prediction objectives and performance metrics complicated the cross-study interpretation. Furthermore, the selection of a single representative model in each study, while designed to ensure consistency, introduced subjectivity and may have underestimated the comprehensive range of model performance reported in individual studies. Regarding literature retrieval, the IEEE Xplore search relied on broad umbrella terms rather than a comprehensive set of omics layer–specific keywords, which may have reduced sensitivity for some technical records. Sixth, although we excluded studies with a sample size under 30, the included analyses involved small cohorts relative to the dimensionality of the omics data, making them highly susceptible to overfitting and inflated performance even when cross-validation was reported. Finally, methodological variability across biospecimens, analytic levels, and computational platforms further limited comparability across studies. Collectively, these limitations—including the small number of eligible studies, limited external validation, incomplete reporting, and methodological heterogeneity—underscore the need for standardized study designs, rigorous external validation, and transparent reporting practices to advance the field toward reproducible and clinically meaningful applications.

Conclusion

This systematic review shows that multiomics-based ML models for stroke-related prediction tasks frequently reported high apparent discrimination, but almost exclusively in small, heterogeneous, and high-dimensional studies with limited external validation and incomplete reporting. Current performance estimates are therefore likely optimistic and are not sufficient to justify routine clinical use or to demonstrate superiority over single-omics or purely clinical models. Even the highest externally validated result should be viewed as a promising proof of concept rather than evidence of robust and clinically generalizable performance. Looking ahead, multiomics integration remains a plausible basis for more precise stroke risk stratification, but only if future models are developed and evaluated in adequately powered, prospectively planned, and transparently reported studies. Such work will require leakage-resistant validation frameworks, clear calibration and clinically relevant operating points, and systematic benchmarking against simpler alternatives. Under these conditions, multiomics-based ML models may progress from exploratory prototypes to robust tools that support earlier diagnosis and more tailored intervention in stroke care.

Acknowledgments

The authors declare that generative artificial intelligence tools were not used in the preparation of this manuscript.

Funding

This work was supported by a grant from the National Research Foundation of Korea, funded by the Korean government (Ministry of Science and ICT; grant RS-2025-00563902).

Data Availability

The data underlying the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

HYY contributed to conceptualization, data curation, validation, formal analysis, writing the original draft, and reviewing and editing the manuscript. HS contributed to data curation, validation, software development, visualization, formal analysis, writing the original draft, and reviewing and editing the manuscript. EJK contributed to data curation, validation, software development,

XSL•FO
RenderX

visualization, formal analysis, writing the original draft, and reviewing and editing the manuscript. YJS contributed to conceptualization, methodology, validation, supervision, project administration, writing the original draft, and reviewing and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA checklist for systematic review.
[DOCX File , 33 KB - jmir_v28i1e85654_app1.docx]

Multimedia Appendix 2
Literature search strategy.
[DOCX File , 36 KB - jmir_v28i1e85654_app2.docx]

Multimedia Appendix 3
Critical appraisal and assessment of bias among studies.
[DOCX File , 28 KB - jmir_v28i1e85654_app3.docx]

References

1. Sacco RL, Kasner SE, Broderick JP, Caplan LR, Connors JJB, Culebras A. An updated definition of stroke for the 21st century: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke 2013;44(7):2064-2089 [FREE Full text] [doi: 10.1161/STR.0b013e318296aeca] [Medline: 23652265]
2. Bersano A, Gatti L. Pathophysiology and treatment of stroke: present status and future perspectives. Int J Mol Sci 2023;24(19):14848 [FREE Full text] [doi: 10.3390/ijms241914848] [Medline: 37834297]
3. Feigin VL, Owolabi MO, World Stroke Organization–Lancet Neurology Commission Stroke Collaboration Group. Pragmatic solutions to reduce the global burden of stroke: a World Stroke Organization–Lancet Neurology Commission. Lancet Neurol 2023;22(12):1160-1206. [doi: 10.1016/S1474-4422(23)00277-6] [Medline: 37827183]
4. Campbell BCV, De Silva DA, Macleod MR, Coutts SB, Schwamm LH, Davis SM. Ischaemic stroke. Nat Rev Dis Primers 2019;5(1):70. [doi: 10.1038/s41572-019-0118-8] [Medline: 31601801]
5. Lin B, Zhang Z, Mei Y, Wang C, Xu H, Liu L, et al. Cumulative risk of stroke recurrence over the last 10 years: a systematic review and meta-analysis. Neurol Sci 2021;42(1):61-71. [doi: 10.1007/s10072-020-04797-5] [Medline: 33040195]
6. Khosla A, Cao Y, Lin CCY, Chiu HK, Hu J, Lee H. An integrated machine learning approach to stroke prediction. 2010 Presented at: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; July 25-28, 2010; Washington p. 183-192. [doi: 10.1145/1835804.1835830]
7. Zhang S, Zhang W, Zhou G. Extended risk factors for stroke prevention. J Natl Med Assoc 2019;111(4):447-456. [doi: 10.1016/j.jnma.2019.02.004] [Medline: 30878142]
8. Misra S, Singh P, Nath M, Bhalla D, Sengupta S, Kumar A. Blood-based protein biomarkers for the diagnosis of acute stroke: a discovery-based SWATH-MS proteomic approach. Front Neurol 2022;13:989856 [FREE Full text] [doi: 10.3389/fneur.2022.989856] [Medline: 36237606]
9. Misra S, Kawamura Y, Singh P, Sengupta S, Nath M, Rahman Z. Prognostic biomarkers of intracerebral hemorrhage identified using targeted proteomics and machine learning algorithms. PLoS One 2024;19(6):e0296616 [FREE Full text] [doi: 10.1371/journal.pone.0296616] [Medline: 38829877]
10. Cheng S, Xu Z, Bian S, Chen X, Shi Y, Li Y. The STROMICS genome study: deep whole-genome sequencing and analysis of 10K Chinese patients with ischemic stroke reveal complex genetic and phenotypic interplay. Cell Discov 2023;9(1):75 [FREE Full text] [doi: 10.1038/s41421-023-00582-8] [Medline: 37479695]
11. Balasubramanian R, Hu J, Guasch-Ferre M, Li J, Sorond F, Zhao Y. Metabolomic profiles associated with incident ischemic stroke. Neurology 2022;98(5):e483-e492 [FREE Full text] [doi: 10.1212/WNL.0000000000013129] [Medline: 34853177]
12. Hochrainer K, Yang W. Stroke proteomics: from discovery to diagnostic and therapeutic applications. Circ Res 2022;130(8):1145-1166 [FREE Full text] [doi: 10.1161/CIRCRESAHA.122.320110] [Medline: 35420912]
13. Alemu R, Sharew NT, Arsano YY, Ahmed M, Tekola-Ayele F, Mersha TB. Multi-omics approaches for understanding gene-environment interactions in noncommunicable diseases: techniques, translation, and equity issues. Hum Genomics 2025;19(1):8 [FREE Full text] [doi: 10.1186/s40246-025-00718-9] [Medline: 39891174]
14. Wekesa JS, Kimwele M. A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment. Front Genet 2023;14:1199087 [FREE Full text] [doi: 10.3389/fgene.2023.1199087] [Medline: 37547471]
15. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol 2017;18(1):83 [FREE Full text] [doi: 10.1186/s13059-017-1215-1] [Medline: 28476144]

XSL•FO
RenderX

16. Zeng J, Zhou H, Wan H, Yang J. Single-cell omics: moving towards a new era in ischemic stroke research. Eur J Pharmacol 2025;1000:177725. [doi: 10.1016/j.ejphar.2025.177725] [Medline: 40350018]
17. Li W, Shao C, Zhou H, Du H, Chen H, Wan H. Multi-omics research strategies in ischemic stroke: a multidimensional perspective. Ageing Res Rev 2022;81:101730. [doi: 10.1016/j.arr.2022.101730] [Medline: 36087702]
18. Montaner J, Ramiro L, Simats A, Tiedt S, Makris K, Jickling GC. Multilevel omics for the discovery of biomarkers and therapeutic targets for stroke. Nat Rev Neurol 2020;16(5):247-264. [doi: 10.1038/s41582-020-0350-6] [Medline: 32322099]
19. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights 2020;14:1177932219899051 [FREE Full text] [doi: 10.1177/1177932219899051] [Medline: 32076369]
20. Lehmann ALCF, Alfieri DF, de Araújo MCM, Trevisani ER, Nagao MR, Pesente FS. Immune-inflammatory, coagulation, adhesion, and imaging biomarkers combined in machine learning models improve the prediction of death 1 year after ischemic stroke. Clin Exp Med 2022;22(1):111-123. [doi: 10.1007/s10238-021-00732-w] [Medline: 34120242]
21. Mukherjee A, Abraham S, Singh A, Balaji S, Mukunthan KS. From data to cure: a comprehensive exploration of multi-omics data analysis for targeted therapies. Mol Biotechnol 2025;67(4):1269-1289. [doi: 10.1007/s12033-024-01133-6] [Medline: 38565775]
22. Lee J, Park KM, Park S. Interpretable machine learning for prediction of clinical outcomes in acute ischemic stroke. Front Neurol 2023;14:1234046 [FREE Full text] [doi: 10.3389/fneur.2023.1234046] [Medline: 37745661]
23. Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. Brief Bioinform 2022;23(1):bbab454 [FREE Full text] [doi: 10.1093/bib/bbab454] [Medline: 34791014]
24. Li R, Li L, Xu Y, Yang J. Machine learning meets omics: applications and perspectives. Brief Bioinform 2022;23(1):bbab460. [doi: 10.1093/bib/bbab460] [Medline: 34791021]
25. Chen C, Wang J, Pan D, Wang X, Xu Y, Yan J. Applications of multi-omics analysis in human diseases. MedComm (2020) 2023;4(4):e315 [FREE Full text] [doi: 10.1002/mco2.315] [Medline: 37533767]
26. Tarazona S, Arzalluz-Luque A, Conesa A. Undisclosed, unmet and neglected challenges in multi-omics studies. Nat Comput Sci 2021;1(6):395-402. [doi: 10.1038/s43588-021-00086-z] [Medline: 38217236]
27. Sarker IH. Machine learning: Algorithms, real-world applications and research directions. SN Comput Sci 2021;2(3):160 [FREE Full text] [doi: 10.1007/s42979-021-00592-x] [Medline: 33778771]
28. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
29. Cook NR, Paynter NP. Performance of reclassification statistics in comparing risk prediction models. Biom J 2011;53(2):237-258 [FREE Full text] [doi: 10.1002/bimj.201000078] [Medline: 21294152]
30. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115(7):928-935. [doi: 10.1161/CIRCULATIONAHA.106.672402] [Medline: 17309939]
31. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions. Version 6.5. Cochrane. 2024. URL: https://www.cochrane.org/handbook [accessed 2025-08-29]
32. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170(1):51-58 [FREE Full text] [doi: 10.7326/M18-1376] [Medline: 30596875]
33. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc 2020;27(12):2011-2015 [FREE Full text] [doi: 10.1093/jamia/ocaa088] [Medline: 32594179]
34. Kolbinger FR, Veldhuizen GP, Zhu J, Truhn D, Kather JN. Reporting guidelines in medical artificial intelligence: a systematic review and meta-analysis. Commun Med (Lond) 2024;4(1):71 [FREE Full text] [doi: 10.1038/s43856-024-00492-0] [Medline: 38605106]
35. Liu Y, Wang W, Cui X, Lyu J, Xie Y. Exploring genetic associations of 3 types of risk factors with ischemic stroke: an integrated bioinformatics study. Stroke 2024;55(6):1619-1628. [doi: 10.1161/STROKEAHA.123.044424] [Medline: 38591222]
36. Chen Z, Zhang X, Qi X, Zheng J, He N, Zheng B. Identification of a potential miRNA-mRNA regulatory network for ischemic stroke by using bioinformatics methods: a retrospective study based on the Gene Expression Omnibus database. Front Immunol 2025;16:1467865 [FREE Full text] [doi: 10.3389/fimmu.2025.1467865] [Medline: 40297585]
37. Zhao T, Yan Q, Wang C, Zeng J, Zhang R, Wang H. Identification of serum biomarkers of ischemic stroke in a hypertensive population based on metabolomics and lipidomics. Neuroscience 2023;533:22-35. [doi: 10.1016/j.neuroscience.2023.09.017] [Medline: 37806545]
38. Ye X, Zhu B, Chen Y, Wang Y, Wang D, Zhao Z. Integrated metabolomics and lipidomics approach for the study of metabolic network and early diagnosis in cerebral infarction. J Proteome Res 2022;21(11):2635-2646. [doi: 10.1021/acs.jproteome.2c00348] [Medline: 36264770]
39. Labarga A, Martínez-Gonzalez J, Barajas M. Integrative multi-omics analysis for etiology classification and biomarker discovery in stroke: advancing towards precision medicine. Biology (Basel) 2024;13(5):338 [FREE Full text] [doi: 10.3390/biology13050338] [Medline: 38785820]

XSL•FO
RenderX

40. Zeng J, Wang C, Guo J, Zhao T, Wang H, Zhang R. Multiomics profiling of plasma reveals molecular alterations prior to a diagnosis with stroke among Chinese hypertension patients. J Proteome Res 2024;23(12):5421-5437. [doi: 10.1021/acs.jproteome.4c00559] [Medline: 39466185]
41. Gan X, Yang S, Zhang Y, Ye Z, Zhang Y, Xiang H. Large-scale plasma proteomics profiles for predicting ischemic stroke risk in the general population. Stroke 2025;56:456-464. [doi: 10.1161/STROKEAHA.124.048654]
42. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF. NCBI GEO: archive for functional genomics data sets—10 years on. Nucleic Acids Res 2011;39(Database issue):D1005-D1010 [FREE Full text] [doi: 10.1093/nar/gkq1184] [Medline: 21097893]
43. Abdelaziz EH, Ismail R, Mabrouk MS, Amin E. Multi-omics data integration and analysis pipeline for precision medicine: systematic review. Comput Biol Chem 2024;113:108254. [doi: 10.1016/j.compbiolchem.2024.108254] [Medline: 39447405]
44. Gligorijević V, Pržulj N. Methods for biological data integration: perspectives and challenges. J R Soc Interface 2015;12(112):20150571 [FREE Full text] [doi: 10.1098/rsif.2015.0571] [Medline: 26490630]
45. Wang M, Yu G, Ressom HW. Integrative analysis of proteomic, glycomic, and metabolomic data for biomarker discovery. IEEE J Biomed Health Inform 2016;20(5):1225-1231 [FREE Full text] [doi: 10.1109/JBHI.2016.2574201] [Medline: 27249841]
46. Krittanawong C, Virk HUH, Bangalore S, Wang Z, Johnson KW, Pinotti R. Machine learning prediction in cardiovascular diseases: a meta-analysis. Sci Rep 2020;10(1):16057 [FREE Full text] [doi: 10.1038/s41598-020-72685-1] [Medline: 32994452]
47. Erickson BJ, Kitamura F. Magician's corner: 9. Performance metrics for machine learning models. Radiol Artif Intell 2021;3(3):e200126 [FREE Full text] [doi: 10.1148/ryai.2021200126] [Medline: 34136815]
48. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS One 2019;14(11):e0224365 [FREE Full text] [doi: 10.1371/journal.pone.0224365] [Medline: 31697686]
49. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB. Calculating the sample size required for developing a clinical prediction model. BMJ 2020;368:m441. [doi: 10.1136/bmj.m441] [Medline: 32188600]
50. Cai Z, Poulos RC, Liu J, Zhong Q. Machine learning for multi-omics data integration in cancer. iScience 2022;25(2):103798 [FREE Full text] [doi: 10.1016/j.isci.2022.103798] [Medline: 35169688]
51. Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J 2021;19:3735-3746 [FREE Full text] [doi: 10.1016/j.csbj.2021.06.030] [Medline: 34285775]
52. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res 2018;46(20):10546-10562 [FREE Full text] [doi: 10.1093/nar/gky889] [Medline: 30295871]
53. Wen G, Li L. FGCNSurv: dually fused graph convolutional network for multi-omics survival prediction. Bioinformatics 2023;39(8):btad472 [FREE Full text] [doi: 10.1093/bioinformatics/btad472] [Medline: 37522887]
54. Wang R, Li B, Lam SM, Shui G. Integration of lipidomics and metabolomics for in-depth understanding of cellular mechanism and disease progression. J Genet Genomics 2020;47(2):69-83. [doi: 10.1016/j.jgg.2019.11.009] [Medline: 32178981]
55. Athieniti E, Spyrou GM. A guide to multi-omics data collection and integration for translational medicine. Comput Struct Biotechnol J 2023;21:134-149 [FREE Full text] [doi: 10.1016/j.csbj.2022.11.050] [Medline: 36544480]
56. Pineda S, Gomez-Rubio P, Picornell A, Bessonov K, Márquez M, Kogevinas M. Framework for the integration of genomics, epigenomics and transcriptomics in complex diseases. Hum Hered 2015;79(3-4):124-136. [doi: 10.1159/000381184] [Medline: 26201698]
57. Flores JE, Claborne DM, Weller ZD, Webb-Robertson BM, Waters KM, Bramer LM. Missing data in multi-omics integration: recent advances through artificial intelligence. Front Artif Intell 2023;6:1098308 [FREE Full text] [doi: 10.3389/frai.2023.1098308] [Medline: 36844425]
58. Song M, Greenbaum J, Luttrell J, Zhou W, Wu C, Shen H. A review of integrative imputation for multi-omics datasets. Front Genet 2020;11:570255 [FREE Full text] [doi: 10.3389/fgene.2020.570255] [Medline: 33193667]
59. Lei N, Zhang X, Wei M, Lao B, Xu X, Zhang M. Machine learning algorithms' accuracy in predicting kidney disease progression: a systematic review and meta-analysis. BMC Med Inform Decis Mak 2022;22(1):205 [FREE Full text] [doi: 10.1186/s12911-022-01951-1] [Medline: 35915457]
60. Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 2020;46(3):383-400 [FREE Full text] [doi: 10.1007/s00134-019-05872-y] [Medline: 31965266]
61. Wang Q, Chang Z, Liu X, Wang Y, Feng C, Ping Y. Predictive value of machine learning for platinum chemotherapy responses in ovarian cancer: systematic review and meta-analysis. J Med Internet Res 2024;26:e48527 [FREE Full text] [doi: 10.2196/48527] [Medline: 38252469]
62. Simpkins AN, Janowski M, Oz HS, Roberts J, Bix G, Doré S. Biomarker application for precision medicine in stroke. Transl Stroke Res 2020;11(4):615-627 [FREE Full text] [doi: 10.1007/s12975-019-00762-3] [Medline: 31848851]

XSL•FO
RenderX

63. Li Q, Zhao L, Chan CL, Zhang Y, Tong SW, Zhang X. Multi-level biomarkers for early diagnosis of ischaemic stroke: a systematic review and meta-analysis. Int J Mol Sci 2023;24(18):13821 [FREE Full text] [doi: 10.3390/ijms241813821] [Medline: 37762122]
64. Dagonnier M, Donnan GA, Davis SM, Dewey HM, Howells DW. Acute stroke biomarkers: are we there yet? Front Neurol 2021;12:619721 [FREE Full text] [doi: 10.3389/fneur.2021.619721] [Medline: 33633673]
65. Zhang WH, Wang WQ, Han X, Gao HL, Li TJ, Xu SS. Advances on diagnostic biomarkers of pancreatic ductal adenocarcinoma: a systems biology perspective. Comput Struct Biotechnol J 2020;18:3606-3614 [FREE Full text] [doi: 10.1016/j.csbj.2020.11.018] [Medline: 33304458]
66. Gallitto G, Englert R, Kincses B, Kotikalapudi R, Li J, Hoffschlag K. External validation of machine learning models-registered models and adaptive sample splitting. Gigascience 2025;14:giaf036 [FREE Full text] [doi: 10.1093/gigascience/giaf036] [Medline: 40366867]
67. Michalska-Falkowska A, Niklinski J, Juhl H, Sulewska A, Kisluk J, Charkiewicz R. Applied molecular-based quality control of biobanked samples for multi-omics approach. Cancers (Basel) 2023;15(14):3742 [FREE Full text] [doi: 10.3390/cancers15143742] [Medline: 37509403]
68. Rozera T, Pasolli E, Segata N, Ianiro G. Machine learning and artificial intelligence in the multi-omics approach to gut microbiota. Gastroenterology 2025;169(3):487-501 [FREE Full text] [doi: 10.1053/j.gastro.2025.02.035] [Medline: 40118220]
69. Mataraso SJ, Espinosa CA, Seong D, Reincke SM, Berson E, Reiss JD. A machine learning approach to leveraging electronic health records for enhanced omics analysis. Nat Mach Intell 2025;7(2):293-306. [doi: 10.1038/s42256-024-00974-9] [Medline: 40008295]
70. Hassan A, Gulzar Ahmad S, Ullah Munir E, Ali Khan I, Ramzan N. Predictive modelling and identification of key risk factors for stroke using machine learning. Sci Rep 2024;14(1):11498 [FREE Full text] [doi: 10.1038/s41598-024-61665-4] [Medline: 38769427]
71. Glaab E, Rauschenberger A, Banzi R, Gerardi C, Garcia P, Demotes J. Biomarker discovery studies for patient stratification using machine learning analysis of omics data: a scoping review. BMJ Open 2021;11(12):e053674 [FREE Full text] [doi: 10.1136/bmjopen-2021-053674] [Medline: 34873011]

Abbreviations

AI: artificial intelligence
AUC: area under the receiver operating characteristic curve
LASSO: least absolute shrinkage and selection operator
LC-MS: liquid chromatography–mass spectrometry
MINIMAR: Minimum Information for Medical AI Reporting
ML: machine learning
PICO: participants, interventions, comparisons, and outcomes
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST: Prediction Model Risk of Bias Assessment Tool
PROSPERO: International Prospective Register of Systematic Reviews
ROC: receiver operating characteristic
SVM: support vector machine

Edited by A Coristine; submitted 10.Oct.2025; peer-reviewed by T Liu, H Lv, M Gandhi; comments to author 10.Nov.2025; revised version received 27.Jan.2026; accepted 29.Jan.2026; published 19.Feb.2026.

Please cite as:
Yoo HY, Shin H, Kim EJ, Son YJ
Machine Learning for Predicting Stroke Risk Stratification Using Multiomics Data: Systematic Review
J Med Internet Res 2026;28:e85654
URL: https://www.jmir.org/2026/1/e85654
doi:10.2196/85654
PMID:

©Hae Young Yoo, Hyerim Shin, Eun-Jung Kim, Youn-Jung Son. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly

XSL•FO
RenderX

cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Advancing Nursing Data Integration Through a Nursing Minimum Dataset for the Conceptual and Technical Development of a “Fall Prevention” Data Module: Development Study

Sarah Milkov[1], MSc; Antonia Schmidt[2], MSc; Anja Burmann[3], PhD; Niklas Tschorn[2], MSc; Marcel Klötgen[2], MSc; Wolfgang Deiters[1], Prof Dr; Christian Potthoff[4]; Kirsten Neveling[4], BSc; Yvonne Weber[5], MSc; Maren Keuchel[3], MSc; Daniela Holle[1], Prof Dr

[1]Hochschule Bochum, Bochum, Germany

[2]Fraunhofer Institute for Software and Systems Engineering ISST, Dortmund, Germany

[3]Fraunhofer Institute for Software and Systems Engineering, Dortmund, North Rhine-Westphalia, Germany

[4]Diakonie Michaelshoven Pflege und Wohnen gGmbH, Köln, Germany

[5]Connext Communication GmbH, Paderborn, Germany

Corresponding Author:
Sarah Milkov, MSc
Hochschule Bochum
Gesundheitscampus 6-8
Bochum, 44801
Germany
Phone: 49 23477727701
Email: Sarah.Milkov@hs-bochum.de

Abstract

Background: In aging populations, the demand for care, including care delivery in long-term care (LTC) facilities, is increasing. This situation highlights the need to optimize care processes through continuous scientific evaluation. The use of artificial intelligence (AI) has the potential for use in nursing research, but it experiences a lack of standardization and structuring of nursing data. Although solutions such as standardized nursing terminologies exist, their use in practice has thus far not been widespread and is often associated with high documentation costs.

Objective: This paper presents the conceptual and technical development of a nursing minimum dataset that focuses on a specific “fall prevention” use case. The aim of this work was to improve data standardization and usability for research and AI-based analysis in LTC settings.

Methods: A representation of the “fall prevention” use case was developed using literature analyses, co-design workshops, and a quantitative survey (n=158). Technical indexing was conducted by translating the results into the technical terminology of the Health Level Seven International Fast Healthcare Interoperability Resources standard.

Results: The “fall prevention” use case was developed as part of a German nursing minimum dataset for long-term residential care with 8 basic modules (patient or client demographics) and 11 extension modules (nursing care elements). The module of the “fall prevention” use case includes fall risk factors, interventions, and outcomes. The literature analysis included 4 international fall guidelines and 17 practice and transfer documents established in German LTC. In total, 12 experts from the fields of management, quality management, technical application support, nursing service management, department management, and members of the PFLIP (Pflege-Kerndatensatz und Intersektorales Pflegedaten-Repository [Nursing Minimum Data Set and Intersectoral Nursing Data Repository]) research project participated in the workshops. A total of 158 people participated in the quantitative survey, the majority of whom were female (117/158, 74%), with 63% (100/158) working directly in nursing care and an average of 24.9 years of professional experience, mainly in LTC (63/158, 40%), outpatient care (37/158, 23%), and hospitals (14/158, 9%). The relevant content, in the sense of a minimum set of items, was identified and prioritized in collaboration with nursing experts and translated into a Fast Healthcare Interoperability Resources–based implementation guide.

Conclusions: This approach addresses the lack of structured nursing data for AI and research and can serve as an example for interoperable, cross-sector solutions in global LTC.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e82417) doi:10.2196/82417

KEYWORDS

nursing; minimum dataset; core dataset; fall; standardization; decision-making

Introduction

Background

Demographic change in Germany and Europe is leading to a steadily growing number of people in need of care, especially older people with multiple health problems [1]. A growing problem in this age group is falls, which place considerable physical and psychological strain on both those affected and the care systems and increase the need for treatment [2,3]. The continuous scientific development of nursing processes is essential for improving the quality of care, promoting evidence-based knowledge, and further developing nursing practices [4,5].

The application of artificial intelligence (AI) provides new possibilities for nursing research by enabling the efficient analysis of large amounts of care data, recognizing patterns, and obtaining scientifically useful insights [6]. This strengthened nursing research contributes significantly to improving and facilitating nursing care, as scientific findings can be used to develop evidence-based care strategies [7]. In turn, data from nursing care form the basis for AI-supported analyses (Figure 1) [8]. The cyclical process between (AI-based) research, health care and nursing care, and AI innovations, with data serving as the central interface, is shown in Figure 1. Using the example of fall prevention, AI innovations can enhance research opportunities, such as studies on falls in long-term care (LTC). The findings from this research can then improve the quality of nursing care and help prevent falls. In nursing care, data are continuously generated, for example, when falls are documented, which feeds back into research and supports further AI-driven innovations in fall prevention. The challenge is to bring the data collected in the health care system into a uniform format that is accessible for research.

Figure 1. Nursing research, care, and AI innovation cycle. AI: artificial intelligence; HC: health care; NC: nursing care.

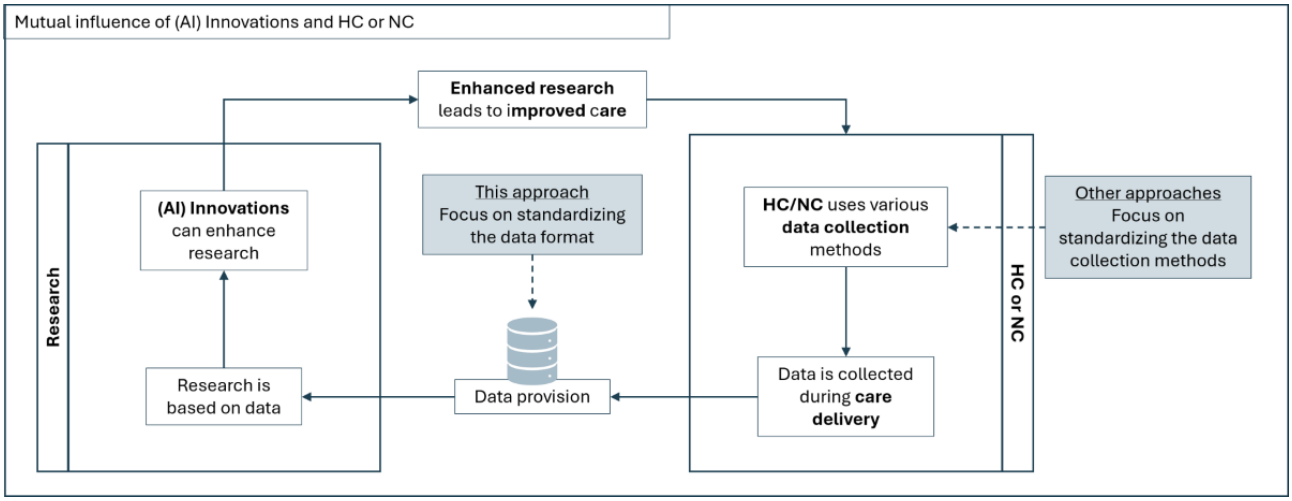

The secondary use of existing documentation data offers an alternative to traditional prospective studies for research purposes. However, accessing and harmonizing these data are challenging due to heterogeneous systems and varying documentation practices, which often result in unstructured or nonstandardized data.

Existing solutions, such as internationally recognized nursing terminologies or so-called standardized nursing languages, aim to standardize the documentation of nursing data by specifying certain terms, topics, and terminologies [9]. However, the use of standardized nursing terminology and specialist languages is not common across Europe. The use of structured nursing documentation varies widely across countries, and studies by Thoroddsen et al [10] and Dos Santos et al [11] suggest that nursing diagnoses and interventions are used most frequently.

The general use of standardized nursing terminology is still lacking, which makes access to nursing data more difficult [10]. A cross-sectional study from 2020 in the Netherlands revealed that approximately half of the 667 nurses and assistants surveyed would use standardized terminology in electronic documentation [12]. The reasons for this include burdensome documentation (use barrier) and a still predominant use of medical terminology (terminology barrier), although medical terminology is insufficient for describing nursing processes, such as assessing unsteadiness or unsuitable footwear. Other studies also show that complex documentation is inefficient in care practice, consumes resources necessary to perform care measures, and leads to a considerable documentation burden for nursing staff [11,13-16].

In the case of Germany, LTC facilities took a step away from standardization toward more patient-oriented documentation in 2015 with the introduction of a law to reduce bureaucracy in care documentation, which led to an increased use of unstructured free-text entries [17,18]. The absence of legal provisions governing the use of standardized terminology constitutes a structural barrier to the implementation of such concepts [19]. Furthermore, there is currently a paucity of

XSL•FO
RenderX

terminology that has been specifically developed for German LTC, with the literature consisting of translations of foreign approaches. Nursing care systems are characterized by a high degree of heterogeneity in different countries, which reinforces the use barrier [20].

This paper addresses a research gap in German LTC caused by the lack of integration and routine use of nursing terminology. This situation gives rise to a paucity of structured and standardized data, while existing approaches to data use remain ineffective, resulting in the untapped potential of valuable data for research purposes remaining unused. Our approach aims to standardize the format of research data rather than the methods of data collection, allowing nursing staff to continue working as usual, while their collected data are converted for research purposes (Figure 1).

The following research questions are addressed:

- What conceptual, nursing, and technical requirements are necessary to develop a nursing minimum dataset (NMDS) based on existing nursing care data from routine documentation using the example of the "fall prevention" use case?
- How can a minimum modular nursing dataset contribute to the harmonization and integration of nursing data for cross-sectoral care and data-based research?

The conceptual and technical development of an NMDS according to the definition by Werley et al [21] for the context of LTC is presented via the "fall prevention" use case, and its potential is discussed in terms of the challenges mentioned. In our use case, an analysis is conducted of data on risk factors, interventions, and outcomes of falls with the aim of preventing such incidents. In addition, a possible practical application in the context of the AI-supported analysis of nursing data is presented.

Objectives

This paper presents a new approach to enable the secondary use of nursing data without additional effort in nursing practice. To this end, the following objectives are pursued: (1) outline the necessary steps for the development of an NMDS based on the "fall prevention" use case, guided by the literature and expert input; (2) describe the technical approach to systematically record fall risk factors, including the collection and prioritization of key factors; and (3) provide a perspective on an NMDS for data collection and harmonization, demonstrating its applicability for research and AI-based analysis.

Related Work

Some related work must be accounted for when addressing the research objectives. The German research project PFLIP (Pflege-Kerndatensatz und Intersektorales Pflegedaten-Repository [Nursing Minimum Data Set and Intersectoral Nursing Data Repository]), on which the following results rely, aims to standardize previously collected nursing data to make the data available for research and AI applications and to offer the possibility of harmonizing data across sectors [22]. This project is based on the German Medical Informatics Initiative (MII), which has developed a medical core dataset on which the NMDS presented in this paper is based [22,23].

With respect to the scientific foundations of fall risk assessment, the "fall prevention expert standard" guidelines served as a foundation for the development of a use case. In Germany, the nursing guidelines that define best practices and quality benchmarks to ensure high standards of nursing care are called "expert standards" [2].

With respect to standardization and interoperability, the National Association of Statutory Health Insurance Physicians in Germany publishes relevant health care transition documents for interoperable data exchange in a telematic infrastructure. The work of this organization is used as a basis for the development of conceptual and technical use cases [24].

To enable standardized fall data to be used on a broader basis, a suitable data format must be chosen. The Health Level Seven International *(HL7)* Fast Healthcare Interoperability Resources *(*FHIR*) standard* is an internationally recognized framework for the exchange of health information, and its technical terminology is used by health care providers, software developers, and governments worldwide to improve interoperability between different systems and facilitate access to health data. The active community of FHIR includes professionals from various health care fields, who work together to further develop and implement the standard to optimize global health care.

This paper examines the development of an NMDS for LTCs using the example of the "fall prevention" use case. Building on initiatives such as the MII, German expert standards, and international standards such as HL7 FHIR, the NMDS aims to harmonize nursing data, integrate it across sectors, and make it usable for both practice and research, assuming that such an approach is feasible, relevant to practice, and sustainable to implement.

Methods

Overview

The methodology for achieving our research goals consists of an approach of conceptual standardization, which we combine with and translate into technical standardization. An overview of the procedure is presented in Figure 2.

XSL•FO
RenderX

Figure 2. Methodological approach. AI: artificial intelligence; FHIR: Fast Healthcare Interoperability Resources; HL7: Health Level Seven International; MII: Medical Informatics Initiative; NMDS: nursing minimum dataset.

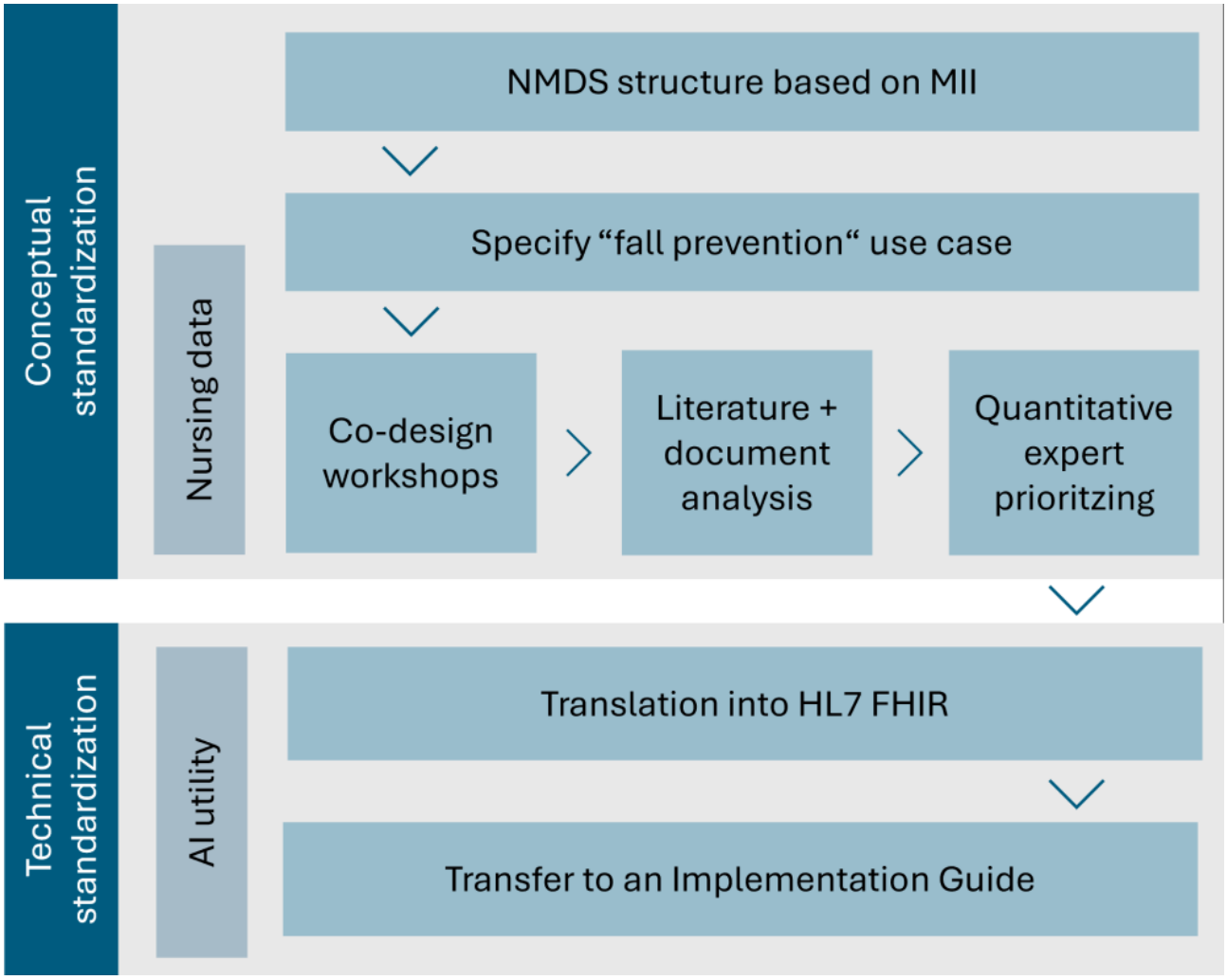

Conceptual Standardization

Overview

A co-design approach was used to develop the content of the "fall prevention" use case of an NMDS. Co-design, according to Sanders and Stappers [25], is an approach that describes collective creativity throughout an entire design process. Designers collaborate to develop innovative solutions by integrating diverse perspectives and skills into the design process [22]. Within this approach, an NMDS structure was developed, and the use case was selected and developed in a participatory manner. The development was supplemented by literature and document analyses and a quantitative survey of nursing staff, and the initially broad approach was narrowed down to a minimum set of items.

Co-Design Development of an NMDS Based on the "Fall Prevention" Use Case

The aim was to design a structure for an NMDS and specify the "fall prevention" use case. The use case was considered particularly suitable because a fall event is extensively documented for insurance-related reasons; therefore, extensive data are available [2].

The foundation for this work was a model developed within the PFLIP research project, drawing on the MII medical core dataset and the German guideline "Expert Standard of Nursing," which served as a basis for this study, while details of its development are not reported here (Multimedia Appendix 1) [20,23]. A qualitative, participatory design in the form of a co-designed expert workshop was chosen to achieve the stated goal.

Participants were recruited through the management of the participating nursing facility in writing and through personal contact. The sample was specifically selected to reflect a range of nursing expertise, and the criteria for participation were work in the professional fields of general management, quality management, technical application support, nursing service management, nursing department management, and contact with fall documentation in everyday work. No restrictions were imposed regarding sociodemographic data. Before the workshop took place, all the participants provided their consent for the procedure and data processing. Participation was voluntary and could be discontinued or revoked at any time. The data were processed anonymously; thus, no conclusions could be drawn about individuals, and the data were handled in accordance with the General Data Protection Regulation.

The study was not approved by an institutional review board. The aim of the study was to gather opinions and assessments

XSL•FO
RenderX

from employees of a nursing facility as experts on a fictitious use case. This is not interventional or experimental research on humans, and no personal data were collected.

The "fall prevention expert standard" [26] served as the foundation for the use case and was initially evaluated for completeness by the nursing staff during a workshop; its content was visualized for the participants, which allowed them to prioritize the listed fall-related items according to perceived relevance and to suggest additional aspects they considered important in the context of fall prevention. The results were compiled verbally and documented in writing by the research team. Additionally, patient journey mapping [27] was developed to understand how residents of nursing facilities navigate the nursing system, to illustrate the interfaces in the cross-sectoral documentation of a fall, and to identify additional documentation tools that provide information on the current standard fall documentation.

The results were evaluated using qualitative content analysis according to Mayring [28]. Coding was performed deductively using predefined categories derived from the expert standard. This approach enables a structured and comprehensible analysis of the data.

Literature and Document Search

Additional international guidelines for fall events were researched on the platform Google Scholar and in national guideline registers. Practice documents and legal requirements were evaluated to identify current documentation of falls in nursing care. The documents were searched directly on the website of providers for transfer documents in medicine and nursing care as well as an LTC facility in Germany. Finally, a targeted search for standardized nursing terminology was conducted to incorporate international standards. This search included an analysis of documents on the International Classification of Nursing Practice [29].

Guidelines that addressed the topic of "fall prevention" in LTC settings and among older adults were included. The guidelines were subsequently analyzed in terms of content (risk factors, interventions, and outcomes). All the documents were screened according to these criteria.

Relevant information from the guidelines and documents was systematically extracted and organized according to predefined categories of risk factors, interventions, and outcomes and presented in tabular form, ensuring that the guideline-based requirements relating to falls in LTC were clearly and comprehensively presented.

Quantitative Online Survey

An anonymous online survey was conducted to specify broad findings in a practice-relevant manner. Below, we report on the study following the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist [30].

The target population included people working in the nursing sector who were chosen for open convenience sampling. The questionnaire was distributed nationwide to nursing facilities in various settings as well as to nursing staff with different professional backgrounds and experience levels.

The survey was created with the software LimeSurvey (LimeSurvey GmbH) and can be found in Multimedia Appendix 2. The aim of the questionnaire was to prioritize fall events identified by nursing staff. The questionnaire included demographic questions and a Likert scale on which the participants rated the relevance of various fall-related items (risk factors, interventions, and outcomes). The participants also indicated whether the items mentioned had previously been documented in their care facility and, if so, whether they had been documented digitally. The findings from the previous steps were used as a base set of questions for the survey.

The survey responses were descriptively analyzed and graphically presented via the software Looker Studio (Google LLC). As the survey was created in LimeSurvey, it was possible to set it up so that no incomplete questionnaires were saved; this ensured that no incomplete answers were included in the evaluation.

Technical Standardization

In the next step, the interim results were translated into the technical terminology of the HL7 FHIR standard and made available for use in digital applications and AI-based systems.

Systematic Capture of Fall Risk Factors and Outcomes

To enable broader use of the standardized fall data, HL7 FHIR was chosen as a suitable data format. The methodology used to translate the identified fall parameters into HL7 FHIR involved a series of systematic processes. The focus was initially limited to fall risks that were documented digitally in our cooperating LTC facility and received more than 80% agreement on their relevance for fall detection by the nursing staff, as rated in the Quantitative Online Survey section. This restriction, including the deliberate exclusion of interventions, was made to conserve time- and effort-related resources. Initially, a thorough screening of the FHIR base resources was conducted to determine which resources were most suitable for representing the identified fall risk factors. Subsequently, existing implementation guides (IGs), such as the MII core dataset, were searched for eligible profiles that could be used as a basis for further profiling. Fall risk factors and fall outcomes were subsequently mapped to appropriate codes in the Systematized Nomenclature of Medicine Clinical Terms [31] and the *ICD-10* (*International Statistical Classification of Diseases and Related Health Problems*) [32] to ensure that each risk factor was accurately represented in a standardized format that can be widely recognized and used in the health care industry. An example of this is shown in Multimedia Appendix 3. Finally, FHIR Shorthand was used during the profiling process to create a concise, human-readable, and computable representation of the mapped fall risk factors; this significantly simplified the process of creating and maintaining FHIR resources and profiles.

Perspective on a "Fall Prevention" NMDS for Data Collection and the Harmonization of Nursing Data for Research Purposes

In addition to the systematic capture of fall risk factors, the definition of "fall" NMDS is an important part of the reusability of the data and its structure. As discussed in the previous sections, this secondary use of data plays an essential role in

XSL•FO
RenderX

research on health services. These aspects are also recognized by research data networks such as the German MII, which relies on FHIR. To be compatible with these structures in the future, the NMDS "fall" use case for data collection and the harmonization of nursing data were transformed to FHIR.

An IG was created that provides a detailed, step-by-step manual that effectively communicates how to implement the mapped fall risk factors in FHIR. The IG offers a topic-specific bundled presentation of profiles, value sets, and other FHIR data structures. The need for this IG is based on recognition of the complexity of health care data and the challenges that can arise when standardized health care data models such as FHIR are implemented.

Another argument in favor of following the example of existing research data networks is the prospect of being connected to them. Considering and incorporating the FHIR-based core dataset of the MII ensure that the collection of fall risk factors can be included in the research data network and made available to researchers.

Ethical Considerations

This study did not require approval by an institutional review board, as it did not involve interventional or experimental research with human participants. The study aimed to collect opinions and assessments from nurses as experts regarding fictitious fall risks derived from the literature. All participants received written information about the purpose, scope, and procedures of the questionnaire. Participation was entirely voluntary, and informed consent was obtained prior to data collection. Participants were informed that their data would be processed anonymously. The study was conducted in accordance with the principles of the Declaration of Helsinki, the guidelines of the Council for International Organizations of Medical Sciences, the World Health Organization, the European General Data Protection Regulation, the German Federal Data Protection Act, and the Data Protection Act of the State of North Rhine-Westphalia (Landesdatenschutzgesetz NRW). As all survey data were fully anonymized and no information was collected that would allow the identification of individual participants, the study does not involve the processing of personal data as defined in Article 4(1) of the European General Data Protection Regulation. All data were stored on protected servers. All participating institutions approved the study as part of a data protection agreement, with the involvement of data protection experts.

Results

Conceptual Standardization

The "fall prevention" use case was conceptually developed on the basis of an NMDS based on the MII core data (Multimedia Appendix 2).

Co-Design Development of an NMDS Based on the "Fall Prevention" Use Case

The workshops took place in August 2022. In total, 12 participants from the fields of management, quality management, technical application support, nursing service management, and department management and members of the PFLIP research project participated in the co-design workshops. Experts from the field of nursing science also played a key role in developing patient journey mapping. During this process, participants observed that while the expert standard provides valuable insights, it could be further enriched by considering additional factors to capture the full complexity of falls, leading to an expansion of the topic. Patient journey mapping helps to identify critical interfaces, relevant documentation, and transfer forms that are essential for a comprehensive representation of fall-related processes. The results are presented in Multimedia Appendix 4. On this basis, a search for these documents was initiated to incorporate them into the development of the use case.

Literature and Document Search

As part of the literature analysis, 4 additional international fall guidelines, 5 documents that are routinely applied in German LTC practice, 6 documents from a provider of transfer documents in Germany, 6 transfer documents from a selected LTC facility in Germany, and the International Classification of Nursing Practice falls catalogue were included. Many of these documents depict fall events rather generally. Thus far, no study has comprehensively addressed falls along with associated nursing-relevant risk factors and intervention options.

Quantitative Survey

The quantitative online survey was conducted from March 17, 2023, to July 30, 2023, with a total of 158 participants, 117 of 158 (74%) of whom were identified as female. Among these participants, 63% (100/158) were directly involved in nursing care. The participants had an average of 24.86 years of professional experience, ranging from 0.5 to 45 years. Most respondents (63/158, 40%) worked in LTC at the time of the survey, followed by home care (37/158, 23%) and hospital settings (14/158, 9%).

Risk factors were identified as items in the following categories: diseases, disease-related changes, environment, medications, and other factors. As interventions, items were identified in the following categories: physical measures, clothing adjustment measures, environmental adjustment measures, training and counseling measures, measures involving other professional groups, measures for the use of assistive devices for fall and injury prevention, the use of technology-based measures, and additional nursing measures. As outcomes, items were identified in the category of information about fall events. The results, including the identified content of the "fall prevention" use case, are presented in Multimedia Appendix 5. Table 1 shows the prioritized risk factors, including the percentage of agreement, which were further processed in the subsequent steps.

XSL•FO
RenderX

Table 1. Prioritized fall risk factors.

Risk factor	Rated importance, n (%)
1. Dizziness (vertigo)	158 (100)
2. Parkinson disease	158 (100)
3. Gait instability	158 (100)
4. Poststroke condition	157 (99)
5. History of falls	157 (99)
6. Overestimation of abilities	157 (99)
7. Mobility impairment	156 (99)
8. Antidepressants	156 (99)
9. Multiple sclerosis	156 (99)
10. Musculoskeletal impairment	155 (98)
11. Assistive devices (eg, rollator)	155 (98)
12. Visual impairment	154 (97)
13. Delirium (confused state)	154 (97)
14. Blood pressure fluctuations	154 (97)
15. Joint paresis (paralysis)	153 (97)
16. Confusion	151 (96)
17. Dementia	151 (95)
18. Fear of falling	150 (95)
19. Cognitive impairment	149 (94)
20. Sensorimotor impairment	148 (94)
21. History of fractures	148 (94)
22. Polypharmacy	147 (93)
23. Multimorbidity (multiple conditions)	146 (92)
24. Blood sugar fluctuations	146 (92)
25. Inadequate activities	146 (92)
26. Compliance	145 (92)
27. Alcohol or drug or nicotine abuse	143 (90)
28. Limb amputation	142 (90)
29. Antihypertensives (blood pressure–lowering drugs)	142 (90)
30. Psychotropics (affecting the psyche)	140 (89)
31. Pain	140 (88)
32. Anticonvulsants (to prevent epileptic seizures)	139 (88)
33. Cardiac arrhythmia	137 (86)
34. Measures involving deprivation of liberty	136 (86)
35. Wandering behavior	133 (84)
36. Antiarrhythmic medications	131 (83)
37. Incontinence	130 (82)
38. Sleep disorder	127 (80)
39. Osteoarthritis	127 (80)

XSL•FO
RenderX

Technical Standardization

Systematic Capture of Fall Risk Factors

An overview of the risk factors considered for translation into FHIR can be found in Table 1. The objectives were considered because these received more than 80% agreement on relevance and were currently documented in our cooperating LTC facility.

When the risk factors were translated into the appropriate FHIR resources, existing profiles were screened. For risk factors involving a disease diagnosis, it was necessary to distinguish between medical and nursing diagnoses, because the term "diagnosis" is legally protected and relevant for billing purposes in the German health care system. In total, 60% (n=21) of the risk factors can be mapped as diagnoses, of which 24% (n=10) are nursing diagnoses and 26% (n=11) are medical diagnoses, whereby a total of 4 risk factors may be specified as both medical and nursing diagnoses. For 21% (n=9) of the risk factors, a nursing observation profile was created, which can be used to describe an occurrence during normal nursing activities. A further 12% (n=5) could be mapped to the MII medication profile. More specific profiles were created for the remaining 14% (n=7) of the risk factors, as they had more precise requirements and required differentiation. For example, the risk factor for incontinence was divided into fecal and urinary incontinence, and the necessity of a decision was included for measures involving deprivation of liberty.

Perspective on a "Fall Prevention" NMDS for Data Collection and Harmonization of Nursing Data for Research Purposes

At the heart of the developed IG is a questionnaire, represented by the FHIR resource questionnaire and questionnaire response, which was specifically designed for the "fall prevention" use case. The guide serves as an essential tool for ensuring consistency and accuracy in the implementation process across different systems and platforms. The response form composed of these 2 FHIR resources allows for the systematic documentation of all fall-relevant risk factors. The IG offers a collection of FHIR resources to capture previously identified risk factors for a fall. The order of the questionnaire response items reflects the prioritization of risk factors from the survey (Table 1). In addition to profiles designed solely for one or a few other risk factors, 3 more comprehensive profiles have been developed that can depict a multitude of risk factors (Tables 2 and 3) [33].

Table 2. Self-defined profiles in the PFLIP[a] implementation guide.

HL7[b] FHIR[c] basic resource	Profile designation
Bundle	• Multi-Morbidity (original Multimorbidität)
Condition	• Medical-Diagnosis (original Medizinische-Diagnose) • Nursing-Diagnosis (original Pflege-Diagnose)
Observation	• Nursing-Observation (original Pflege-Beobachtung) • Nursing-Observation-Urinary-Incontinence (original Pflege-Beobachtung-Harn-Inkontinenz) • Nursing-Observation-Faecal-Incontinence (original Pflege-Beobachtung-Stuhl-Inkontinenz) • Survey-Fall-History-6-Months (original Erhebung-Sturzhistorie-6-Monate) • Deprivation-of-Liberty (original Freiheitsentziehende-Maßnahme)
Device	• Aids (original Hilfsmittel)
List	• Multi-Medication (original Multimedikation)
Questionnaire	• Risk-Assessment (original Risikoermittlung)
QuestionnaireResponse	• Risk-Assessment-Fall (original Risikoermittlung-Sturz)
Patient	• Resident-Pseudonymized (original BewohnerIn-Pseudonymisiert)

[a]PFLIP: Pflege-Kerndatensatz und Intersektorales Pflegedaten-Repository (Nursing Minimum Data Set and Intersectoral Nursing Data Repository).

[b]HL7: Health Level Seven International.

[c]FHIR: Fast Healthcare Interoperability Resources.

XSL•FO
RenderX

Table 3. Extensions within the PFLIP[a] implementation guide.

Name	Context of use
VoluntaryConsentExists (original *FreiwilligeEinwilligungLiegtVor*)	Deprivation-of-Liberty (original *Freiheitsentziehende-Maßnahme*)
CourtOrderExists (original *RichterlicheAnordnungLiegtVor*)	Deprivation-of-Liberty (original *Freiheitsentziehende-Maßnahme*)
DateOfBirth_mm_yyyy (original *Geburtstag_mm_yyyy*)	Resident-Pseudonymized (original *BewohnerIn-Pseudonymisiert*)
FullSurveyPeriod (original *VollständigerErhebungszeitraum*)	Survey-Fall-History-6-Months (original *Erhebung-Sturzhistorie-6-Monate*)

[a]PFLIP: Pflege-Kerndatensatz und Intersektorales Pflegedaten-Repository (Nursing Minimum Data Set and Intersectoral Nursing Data Repository).

Discussion

Principal Findings

The development of a German NMDS use case for fall prevention in LTC is based on a literature analysis of international guidelines and German practice documents, resulting in fall risk factors, interventions, and outcomes, and was developed in collaboration with 12 experts from various nursing fields. The relevant minimum items were prioritized by experts, reflected upon in a quantitative survey with 158 participants, and finally transferred into an FHIR-based IG.

The developed use case can benefit nursing research by making standardized, documented data available for secondary use and enabling AI-supported analyses. Since the dataset is based on routine documentation in nursing facilities and standardizes already documented data, it has the potential to support problems that have arisen thus far in efforts to standardize care documentation, such as a lack of manageability and the effort of documentation by nursing staff [13-16].

In contrast to broader international NMDS initiatives, the approach presented here takes a more detailed approach to a specific use case, resulting in practical, actionable information, better reflecting clinical variability, and creating a solid foundation for digital applications and AI.

In addition, the NMDS has been linked to the technical standardization HL7 FHIR from the outset, ensuring interoperability and direct technical implementability and guaranteeing connectivity to modern digital systems.

However, it should be noted that the dataset does not automatically ensure that documented data are also transferred to standards (such as FHIR). Data that are documented in other facilities and care documentation systems are not always automatically available in FHIR and may be in different formats. To counteract this limitation, a system solution, such as a data repository, is needed to automatically standardize routine data. The standardization of content and technology also facilitates the exchange of data between sectors and professions. In addition, the structural orientation toward an existing medical core dataset represents a step toward harmonizing nursing and medical data. This can support the need to give nursing terminology equivalent weight in documentation as medical terminology and integrate them instead of overwriting them [9,23].

Unlike previous approaches in LTC, which often failed to be implemented, the NMDS developed here is based exclusively on existing routine documentation to promote acceptance and long-term use in everyday nursing care without increasing the workload [34]. In accordance with extant recommendations in the literature, the development of NMDS should be undertaken in a multidisciplinary manner and integrated into (digital) documentation systems, with a focus on nursing content as opposed to medical content; thus, it should be oriented toward nursing standards, which we followed and built a base for [34,35].

The co-design approach chosen to develop the use case indicates a way of standardizing nursing data while maintaining a high degree of user centricity. This gives this NMDS an advantage over previous NMDSs and assessment instruments. In the future, a person-centered approach can be ensured by broadening the perspective to include affected persons, that is, residents of LTC facilities. The results in the NMDS table (Multimedia Appendix 5) illustrate the differences between risk factors that are perceived as relevant by practitioners and the scientific literature and those that are documented in a German geriatric care facility; this can indicate which important fall risk factors have not yet been mapped in routine documentation, such as "environmental factors" in this case. This information highlights the need for further research.

Although the results were collected nationally, findings from other European countries demonstrate difficulties in introducing standardized care documentation systems and standardization. This approach can serve as a model for the development of NMDSs for other countries [13-15].

There lies a challenge in bringing the data collected in the health care sector into a uniform format that is accessible for research. Once the data are accessible for research, AI innovations can benefit from a much larger data pool that is also very close to the point of care; this significantly improves the results of AI applications. The FHIR translation makes the data usable for AI applications that can assess and predict falls. The IG thus serves as a valuable resource for health care providers and researchers alike and provides a structured and standardized approach to capturing and analyzing fall risk factors. It is hoped that the future application of this guide will contribute to the development of more effective and personalized fall prevention strategies and ultimately improve patient outcomes. The developed IG also provides a methodological basis for expanding the minimum dataset to include other nursing topics in addition to falls as well as other care sectors and is internationally applicable.

XSL•FO
RenderX

Proof of Concept: Practical Application Example—An AI-Based Nursing Decision Support System

An AI-based nursing decision support system (NDSS) for elderly people to predict falls in nursing homes can be used as an example of an application of the content and technology developed for our NMDS.

Developed in a user-centered design process based on International Organization for Standardization (ISO) 9241-210-2019, the NDSS covers routinely documented aspects of falls in German nursing homes and reflects the most relevant fall-related factors from the literature and practice [36]. Nursing staff with different levels of education and specialized training, administrative employees from an outpatient care facility and day care, and people who worked in the field of nursing science were asked about the practical suitability of the proposed NDSS in 4 workshops. Based on the workshop results, the NDSS was improved iteratively. The decision support system is designed to predict the risk of falling using a 5-step scale and focuses on the interpretability of AI so that predictions are comprehensible and transparent for users. Therefore, the NDSS can provide information about recognized risk factors that are responsible for corresponding fall risk. The proposed decision support system also suggests prophylaxis to reduce the risk of falling and show the progression of this risk over several months.

Limitations

The results have several limitations. For reasons of feasibility, we focused on one use case. In this use case, we initially focused on a limited number of risk factors that were technically implemented. In the future, we plan to expand NMDS thematically. It must also be considered that there are different contexts, causes, and types of falls. Falls are presented in a highly standardized manner in this paper. In addition, to date, only aspects of inpatient LTC have been examined, but we plan to extend the findings to different settings.

It can be assumed that both the participants in the co-design workshops and the participants in the quantitative survey showed a certain level of interest in and affinity for the topic. As participation was voluntary, it is possible that less-interested participants did not take part due to participation bias.

Another potential bias in the quantitative study arises from the fact that participant evaluations were used as a criterion for item selection. Since documentation routines, levels of knowledge, and familiarity with existing evidence vary in everyday clinical practice, there is a risk that empirically relevant content will be underestimated or overlooked; this can lead to an overemphasis on subjective assessments, while evidence-based aspects remain potentially underrepresented. Moreover, we deliberately choose a user-centered approach to differentiate ourselves from previous solutions and ensure greater practical relevance. To reduce that bias, a broad survey covering a range of backgrounds and nursing staff was conducted. In the future, the perspectives of residents within the designated dataset should be incorporated. Furthermore, the participants' ratings were based on international nursing standards and fall prevention guidelines and were not indiscriminately determined by the participants.

Nevertheless, it is believed that this approach is effective in enabling a user-centered methodology to be adopted, thus counteracting the lack of use and implementation of NMDS and nursing terminology and standards, as outlined in the problem statement in the introduction. It is important to validate the results in future studies.

The results are oriented toward national standards, practical procedures, and documentation in German LTC facilities. However, the procedure and methodology can be applied internationally.

Conclusions

The "fall prevention" use case developed for a German NMDS in LTC shows how nursing data can be recorded in a standardized and user-centric way. The use of previously documented routine data reduces the documentation burden for nursing staff and facilitates cross-sector data exchange. This concept creates a valuable basis for nursing research by providing high-quality, structured data for secondary use. It also offers the potential for international application and can serve as a model for the development of nursing datasets in other subject areas, particularly through FHIR translation for use in AI-based applications.

The IG and the profiles it contains are to be submitted for an HL7 FHIR balloting. This structured mechanism is designed to solicit feedback from members of the HL7 community to ensure quality and relevance. The suggestions for changes and improvements collected in this process can be evaluated during further efforts and, if approved, can be integrated into the IG to promote the joint creation and further development of the NMDS.

In conclusion, our approach shows that co-design makes user-centered and applicable NMDS development feasible. Moreover, FHIR enables interoperability. This work represents a preliminary step toward harmonizing medical and nursing care in Germany by aligning the data format with MII structures, with future efforts needed to fully integrate the nursing sector. This paper enables us to adapt the approach internationally for AI-ready nursing data.

Acknowledgments

HL7, FHIR and the FHIR [FLAME DESIGN] are the registered trademarks of Health Level Seven International and their use does not constitute endorsement by HL7.

XSL•FO
RenderX

Funding

The paper was produced as part of the PFLIP (Pflege-Kerndatensatz und Intersektorales Pflegedaten-Repository [Nursing Minimum Data Set and Intersectoral Nursing Data Repository]) research project funded by the Federal Ministry of Research, Technology and Space in Germany (16SV8883).

Data Availability

The datasets generated during this study are not publicly available, as they were collected for internal project purposes; however, requests for data access may be submitted to the corresponding author and will be subject to institutional restrictions.

Authors' Contributions

SM and AS wrote the main manuscript text. NT contributed to the writing of the manuscript. DH, AB, M Klötgen, WD, AS, and SM were involved in the conception and design of the submission as well as the substantial revision of the manuscript. M Keuchel was involved in the substantial revision of the manuscript. CP, KN, YW, DH, AB, WD, AS, NT, M Klötgen, M Keuchel, and SM were involved in the acquisition, analysis, or interpretation of the data. All authors approved the submission of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Nursing minimum dataset (NMDS). Visualization of the integration of an NMDS draft into the structures of the Medical Informatics Initiative (MII) core dataset as an extension module, whereby the MII core dataset is divided into basic and extension modules. The basic modules are defined in a cross-disciplinary manner. The extension modules map data from specific applications or specialist areas.
[PNG File , 73 KB - jmir_v28i1e82417_app1.png]

Multimedia Appendix 2
Online survey.
[PDF File (Adobe PDF File), 120 KB - jmir_v28i1e82417_app2.pdf]

Multimedia Appendix 3
Example of assigning a fall risk factor to Fast Healthcare Interoperability Resources and Systematized Nomenclature of Medicine Clinical Terms codes.
[PNG File , 9 KB - jmir_v28i1e82417_app3.png]

Multimedia Appendix 4
Patient journey mapping.
[PNG File , 270 KB - jmir_v28i1e82417_app4.png]

Multimedia Appendix 5
Results: “fall prevention” use case.
[PNG File , 240 KB - jmir_v28i1e82417_app5.png]

References

1. Nagode M, Lebar L. Trends and challenges in long-term care in Europe. RSP 2019;26(2):255-262. [doi: 10.3935/rsp.v26i2.1655]
2. Expertenstandards und Auditinstrumente.: Deutsches Netzwerk für Qualitätsentwicklung in der Pflege URL: https://www.dnqp.de/expertenstandards-und-auditinstrumente/ [accessed 2025-04-16]
3. Xu Q, Ou X, Li J. The risk of falls among the aging population: a systematic review and meta-analysis. Front Public Health 2022;10:902599 [FREE Full text] [doi: 10.3389/fpubh.2022.902599] [Medline: 36324472]
4. Eckardt P, Culley JM, Corwin E, Richmond T, Dougherty C, Pickler RH, et al. National nursing science priorities: creating a shared vision. Nurs Outlook 2017;65(6):726-736. [doi: 10.1016/j.outlook.2017.06.002] [Medline: 28711216]
5. Rodgers BL, Jacelon CS, Knafl KA. Concept analysis and the advance of nursing knowledge: state of the science. J Nurs Scholarsh 2018;50(4):451-459. [doi: 10.1111/jnu.12386] [Medline: 29689127]
6. Yasin YM, Al-Hamad A, Metersky K, Kehyayan V. Incorporation of artificial intelligence into nursing research: a scoping review. Int Nurs Rev 2025;72(1):e13013. [doi: 10.1111/inr.13013] [Medline: 38967044]

XSL•FO
RenderX

7. Curtis K, Fry M, Shaban RZ, Considine J. Translating research findings to clinical nursing practice. J Clin Nurs 2017;26(5-6):862-872 [FREE Full text] [doi: 10.1111/jocn.13586] [Medline: 27649522]
8. Warnecke M, Holle D, Burmann A. Enabling data-driven nursing innovations: user-centered development of a nursing data module. Curr Dir Biomed Eng 2023;9(1):339-342. [doi: 10.1515/cdbme-2023-1085]
9. Muntlin Athlin Å. Methods, metrics and research gaps around minimum data sets for nursing practice and fundamental care: a scoping literature review. J Clin Nurs 2018;27(11-12):2230-2247. [doi: 10.1111/jocn.14155] [Medline: 29119641]
10. Thoroddsen A, Ehrenberg A, Sermeus W, Saranto K. A survey of nursing documentation, terminologies and standards in European countries. NI 2012 (2012) 2012;2012:406 [FREE Full text] [Medline: 24199130]
11. Dos Santos FC, D'Agostino F, Härkönen M, Nantschev R, Christensen B, Müller-Staub M, et al. Improving the quality of nursing care through standardized nursing languages: call to action across European countries. Int J Med Inform 2024;192:105627. [doi: 10.1016/j.ijmedinf.2024.105627] [Medline: 39306908]
12. De Groot K, De Veer AJE, Paans W, Francke AL. Use of electronic health records and standardized terminologies: a nationwide survey of nursing staff experiences. Int J Nurs Stud 2020;104:103523. [doi: 10.1016/j.ijnurstu.2020.103523] [Medline: 32086028]
13. Laukvik LB, Lyngstad M, Rotegård AK, Fossum M. Utilizing nursing standards in electronic health records: a descriptive qualitative study. Int J Med Inform 2024;184:105350 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105350] [Medline: 38306850]
14. Gesner E, Dykes PC, Zhang L, Gazarian P. Documentation burden in nursing and its role in clinician burnout syndrome. Appl Clin Inform 2022;13(5):983-990 [FREE Full text] [doi: 10.1055/s-0042-1757157] [Medline: 36261113]
15. Lindsay MR, Lytle K. Implementing best practices to redesign workflow and optimize nursing documentation in the electronic health record. Appl Clin Inform 2022;13(3):711-719 [FREE Full text] [doi: 10.1055/a-1868-6431] [Medline: 35668677]
16. De Groot K, De Veer AJE, Munster AM, Francke AL, Paans W. Nursing documentation and its relationship with perceived nursing workload: a mixed-methods study among community nurses. BMC Nurs 2022;21(1):34 [FREE Full text] [doi: 10.1186/s12912-022-00811-7] [Medline: 35090442]
17. Beikirch-Korporal E, Nolting HD, Wipp M. Dokumentieren mit Dem Strukturmodell: Grundlagen - Einführung - Management. Hannover: Vincentz Network; 2017.
18. Panka C. Pflegedokumentation entbürokratisiert: reorganisation der Altenpflege mit dem strukturmodell. In: Entbürokratisierte Pflegedokumentation. Göttingen: Hogrefe AG; 2017.
19. Pflegedokumentation stationär: das handbuch für die pflegeleitung. Bundesministerium für Familie, Senioren, Frauen und Jugend (BMFSFJ). 2015. URL: https://www.bmfsfj.de/resource/blob/79108/391791ec4d2bee12f694780cc595f845/handbuch-pflegedokumentation-data.pdf [accessed 2026-01-17]
20. Giap T, Park M, Bui LK. A comprehensive picture of using standardized nursing languages in long-term care systems: an integrative review. Int J Nurs Knowl 2025;36(2):193-208. [doi: 10.1111/2047-3095.12478] [Medline: 38951041]
21. Werley HH, Devine EC, Zorn CR, Ryan P, Westra BL. The nursing minimum data set: abstraction tool for standardized, comparable, essential data. Am J Public Health 1991;81(4):421-426. [doi: 10.2105/ajph.81.4.421] [Medline: 2003618]
22. PFLIP project. Fraunhofer ISST. URL: https://www.isst.fraunhofer.de/en/departments/healthcare/projects/PFLIP.html [accessed 2025-04-21]
23. The Medical Informatics Initiative's core data set. Medical Informatics Initiative. URL: https://www.medizininformatik-initiative.de/en/medical-informatics-initiatives-core-data-set [accessed 2025-04-21]
24. Telematikinfrastruktur. Kassenärztliche Bundesvereinigung. URL: https://www.kbv.de/praxis/digitalisierung/telematikinfrastruktur [accessed 2025-04-21]
25. Sanders EBN, Stappers PJ. Co-creation and the new landscapes of design. CoDesign 2008;4(1):5-18. [doi: 10.1080/15710880701875068]
26. Expertenstandard sturzprophylaxe in der pflege. Deutsches Netzwerk für Qualitätsentwicklung in der Pflege. 2022. URL: https://www.dnqp.de/fileadmin/HSOS/Homepages/DNQP/Dateien/Expertenstandards/Sturzprophylaxe/Sturz_2Akt_Auszug.pdf [accessed 2026-01-17]
27. Bulto LN, Davies E, Kelly J, Hendriks JM. Patient journey mapping: emerging methods for understanding and improving patient experiences of health systems and services. Eur J Cardiovasc Nurs 2024;23(4):429-433. [doi: 10.1093/eurjcn/zvae012] [Medline: 38306596]
28. Mayring P. Qualitative content analysis. In: Handbook of Qualitative Research in Psychology. Germany: VS Publishing House for Social Sciences; 2010:601-613.
29. Cho I, Kim J, Chae JS, Jung M, Kim YH. Development of ICNP-based inpatient falls prevention catalogue. Int Nurs Rev 2020;67(2):239-248. [doi: 10.1111/inr.12566] [Medline: 31762026]
30. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
31. SNOMED CT—the global clinical terminology. SNOMED International. URL: https://www.snomed.org/ [accessed 2025-04-25]

XSL•FO
RenderX

32. ICD-10-WHO: internationale statistische klassifikation der krankheiten und verwandter gesundheitsprobleme. Bundesinstitut für Arzneimittel und Medizinprodukte. URL: https://www.bfarm.de/DE/Kodiersysteme/Klassifikationen/ICD/ICD-10-WHO/_node.html [accessed 2026-01-17]
33. PFLIP-Pflegedaten HL7 FHIR implementation guide. Fraunhofer ISST. 2025. URL: https://pflip-ig.isst.fraunhofer.de [accessed 2025-04-21]
34. Milkov S, Linders L, Napetschnig A, Deiters W, Sirsch E, Holle D. Nursing minimum datasets in long-term care settings: scoping review. J Med Internet Res 2025;27:e68670 [FREE Full text] [doi: 10.2196/68670] [Medline: 41089026]
35. Freguia F, Danielis M, Moreale R, Palese A. Nursing minimum data sets: findings from an umbrella review. Health Informatics J 2022;28(2):14604582221099826 [FREE Full text] [doi: 10.1177/14604582221099826] [Medline: 35634983]
36. Tschorn N, Keuchel M, Müller I, Milkov S, Potthoff C, Windrath D, et al. Iterative design of a decision support system for fall risk detection in residential care facilities. Curr Dir Biomed Eng 2024;10(4):653-656. [doi: 10.1515/cdbme-2024-2160]

Abbreviations

AI: artificial intelligence
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
FHIR: Fast Healthcare Interoperability Resources
HL7: Health Level Seven International
ICD-10: International Statistical Classification of Diseases and Related Health Problems
IG: implementation guide
ISO: International Organization for Standardization
LTC: long-term care
MII: Medical Informatics Initiative
NDSS: nursing decision-support system
NMDS: nursing minimum dataset
PFLIP: Pflege-Kerndatensatz und Intersektorales Pflegedaten-Repository (Nursing Minimum Data Set and Intersectoral Nursing Data Repository)

Edited by MP Gagnon; submitted 14.Aug.2025; peer-reviewed by K Seibert, J Olsen, F Akhter; comments to author 29.Oct.2025; revised version received 17.Dec.2025; accepted 09.Jan.2026; published 17.Mar.2026.

Please cite as:
Milkov S, Schmidt A, Burmann A, Tschorn N, Klötgen M, Deiters W, Potthoff C, Neveling K, Weber Y, Keuchel M, Holle D
Advancing Nursing Data Integration Through a Nursing Minimum Dataset for the Conceptual and Technical Development of a "Fall Prevention" Data Module: Development Study
J Med Internet Res 2026;28:e82417
URL: https://www.jmir.org/2026/1/e82417
doi:10.2196/82417
PMID:

©Sarah Milkov, Antonia Schmidt, Anja Burmann, Niklas Tschorn, Marcel Klötgen, Wolfgang Deiters, Christian Potthoff, Kirsten Neveling, Yvonne Weber, Maren Keuchel, Daniela Holle. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Classifying American Society of Anesthesiologists Physical Status With a Low-Rank–Adapted Large Language Model: Development and Validation Study

Min-Chia Chen[1], MS; Shanq-Jang Ruan[1], PhD; Jo-Hsin Wu[2], MD; Pei-fu Chen[3], MD, PhD

[1]Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan

[2]Department of Anesthesiology, Far Eastern Memorial Hospital, New Taipei, Taiwan

[3]Department of Anesthesia, University of Iowa, Iowa City, IA, United States

Corresponding Author:
Pei-fu Chen, MD, PhD
Department of Anesthesia
University of Iowa
200 Hawkins Drive, 6 JCP
Iowa City, IA, 52242
United States
Phone: 1 319 356 2502
Email: pei-fu-chen@uiowa.edu

Abstract

Background: The American Society of Anesthesiologists Physical Status (ASA-PS) classification is integral to preoperative risk assessment; yet, assignment remains subjective and labor-intensive. Recent large language models (LLMs) process free-text electronic health records (EHRs), but few studies have evaluated parameter-efficient adaptations that both predict ASA-PS and provide clinician-readable rationales. Low-rank adaptation (LoRA) is a parameter-efficient technique that updates only a small set of add-on parameters rather than the entire model, enabling efficient fine-tuning on modest data and hardware. A lightweight, instruction-tuned LLM with these capabilities could streamline workflow and broaden access to explainable decision support.

Objective: This study aimed to develop and evaluate a LoRA–fine-tuned large language model meta-AI (LLaMA-3) for ASA-PS classification from preoperative clinical narratives and benchmark it against traditional machine learning classifiers and domain-specific LLMs.

Methods: Preoperative anesthesia notes and discharge summaries were extracted from the EHR and reformatted into an Alpaca-style instruction-response prompt, requesting ASA-PS class labels (I-V) annotated by anesthesiologists. The LoRA-enhanced LLaMA-3 model was fine-tuned with mixed-precision training and evaluated on a hold-out test set. Baselines included random forest classifier, Extreme Gradient Boosting (XGBoost) classifier, support vector machine, fastText, BioBERT, ClinicalBERT, and untuned LLaMA-3. Performance was assessed with micro- and macroaveraged F_1-score and Matthews correlation coefficient (MCC), each reported with 95% bootstrap CIs. Pairwise model error rates were compared using McNemar test.

Results: The LoRA-LLaMA-3 model achieved a micro–F_1-score of 0.780 (95% CI 0.769-0.792) and an MCC of 0.533 (95% CI 0.518-0.546), outperforming other LLM baselines. After fine-tuning, BioBERT reached a micro–F_1-score of 0.762 and an MCC of 0.508, whereas ClinicalBERT achieved a micro–F_1-score of 0.757 and an MCC of 0.515. fastText yielded a micro–F_1-score of 0.762 and an MCC of 0.536. The untuned LLaMA-3 performed poorly (micro–F_1-score of 0.073; MCC of 0.002). However, macro–F_1-score of LoRA-LLaMA-3 (0.316) was lower than that of other language models (0.349-0.372). Among all models, XGBoost obtained the highest scores (micro–F_1-score of 0.815, 95% CI 0.804-0.826; macro–F_1-score of 0.348, 95% CI 0.334-0.361; MCC 0.613, 95% CI 0.599-0.626). Ablation experiments identified dropout = 0.3, learning rate = 3×10^{-5}, temperature = 0.1, and top-P= 0.1 as the optimal hyperparameter settings. The LoRA model also produced rationales that highlighted medically pertinent terms.

Conclusions: LoRA fine-tuning improved LLaMA-3 from near-random performance into an ASA-PS classifier with higher micro–F_1-score and significantly lower misclassification rates than other language model baselines. However, macroaveraged performance was lower, indicating limited discrimination for minority ASA classes. Traditional machine learning models demonstrated higher predictive performance. Beyond predictive performance, LoRA-LLaMA-3 generated clinician-oriented

XSL•FO
RenderX

explanations that enhance decision transparency. By reformatting routine EHR narratives into instruction-response pairs and relying on lightweight parameter adaptation, this approach offers a practical, resource-efficient framework for introducing explainable LLMs to clinical classification tasks.

(J Med Internet Res 2026;28:e89540) doi:10.2196/89540

KEYWORDS

American Society of Anesthesiologists Physical Status; large language models; low-rank adaptation; parameter-efficient fine-tuning; electronic health records; natural language processing.

Introduction

The American Society of Anesthesiologists Physical Status (ASA-PS) classification system is an essential tool for preoperative assessment, providing a standardized framework to evaluate a patient's health status and predict perioperative risks. This system consists of 6 levels, ranging from ASA I (a healthy patient) to ASA VI (a brain-dead patient), with an additional "E" designation to indicate emergency surgeries. While the ASA-PS system is widely adopted in clinical practice, it has certain limitations, the most notable being its reliance on subjective judgment. This reliance can lead to inconsistencies among evaluators and variability in risk stratification outcomes. These inconsistencies highlight the need for more objective and reproducible methods for assigning ASA-PS classifications [1,2].

Recent advances in machine learning (ML) and natural language processing (NLP) offer potential solutions to address this challenge. With the widespread adoption of electronic health records (EHRs), large amounts of unstructured clinical text data, including preoperative evaluation notes written by clinicians and anesthesiologists, are now available. These notes contain detailed narratives of patients' medical and surgical histories, offering insights into their health and perioperative risks [3,4]. Unlike traditional keyword-based approaches, modern NLP techniques, particularly large language models (LLMs), capture contextual nuances and interword dependencies across entire text sequences, demonstrating state-of-the-art performance in text classification tasks [5,6].

Various studies have explored the application of ASA-PS classifications to assess their importance in perioperative risk prediction. Some studies have focused on developing tools to improve interrater consistency, such as introducing more detailed evaluation guidelines [1,2]. However, these approaches remain limited by the subjective nature of the evaluations. Recent research has attempted to automate classification by leveraging structured data in electronic medical records. Nonetheless, these methods often require predefined features and fail to fully use the latent information within unstructured text [3]. In parallel, large-scale EHR studies have derived machine learning ASA scores to support preoperative screening and risk stratification, showing strong discrimination for outcomes such as 30-day mortality and intensive care unit admission [7]. Earlier structured data models have also demonstrated that ASA-PS can be predicted at scale from routinely collected variables, underscoring its central role in perioperative risk models [8].

In the field of NLP, ML models have been widely applied to medical text processing. Early studies used traditional methods such as random forests and support vector machines (SVMs) with n-grams for text classification. However, their performance was constrained by the efficiency of manual feature extraction [9,10]. Subsequent research introduced embedding-based techniques such as fastText, which significantly improved classification accuracy by capturing semantic relationships between words [11,12]. Building on this, NLP systems that read preoperative narratives have directly predicted ASA-PS from free text, demonstrating the feasibility and accuracy of text-only approaches in real clinical documents [13,14]. More recently, transformer-based models have achieved performance comparable with board-certified anesthesiologists for ASA-PS assignment, suggesting a path to reduce interrater variability with automated, scalable tools [15].

Recently, deep learning models such as BioBERT and ClinicalBERT, designed explicitly for medical text, have achieved breakthrough performance in various clinical NLP tasks [5,16]. These models leverage pretrained language models on large-scale medical corpora, showcasing excellent text classification capabilities [17,18]. However, such methods often require substantial computational resources and have limited adaptability to small datasets.

Low-rank adaptation (LoRA) is a parameter-efficient fine-tuning technique that enables pretrained models to adapt to specific tasks with minimal computational overhead [19]. It is particularly suitable for clinical applications with common data sparsity and resource constraints [20]. LoRA updates only a small number of trainable parameters while keeping the majority of the pretrained weights frozen. This design greatly reduces GPU memory requirements and training time [19].

To address the limitations of existing approaches, this study investigates the application of LoRA fine-tuned LLM for predicting ASA-PS classifications from unstructured preoperative evaluation notes. This approach aims to improve the accuracy and consistency of ASA-PS classifications while reducing reliance on subjective human evaluation. Furthermore, this study evaluates whether parameter-efficient fine-tuning of small LLMs can provide clinically interpretable and scalable decision support tools for perioperative risk assessment.

Methods

Ethical Considerations

This retrospective single-center study using routinely collected preoperative assessments from November 21, 2015, to August 1, 2023, was approved by the institutional review board of Far

XSL•FO
RenderX

Eastern Memorial Hospital (114213-E), a tertiary academic medical center, with a waiver of informed consent granted because the research involved secondary analysis of existing records, posed no more than minimal risk, and could not be practicably conducted without the waiver. The dataset was derived from the institutional EHR system, including routinely documented preoperative anesthesia assessment records for surgical patients. Data were deidentified prior to analysis and analyzed on secure on-premises infrastructure with restricted, audited access. No participant compensation was provided. The manuscript and supplements contain no identifiable images.

Inclusion and Exclusion Criteria

Cases included in this study were the patients aged 18 years or older and had undergone at least 1 surgical procedure under general or neuraxial anesthesia (N=118,274). Cases classified as ASA VI (brain-dead patients) were excluded (n=90) to ensure relevance to general clinical surgical risk assessment. For patients with multiple visits for each surgery, only the most recent version of the preoperative assessment record was retained. The most recent assessment was selected because it typically reflects the most complete and updated comorbidity documentation. Continuous physiologic and laboratory variables were extracted exclusively from this most recent assessment and were not longitudinally aggregated across prior encounters.

The final cohort consisted of 24,491 patients and was randomly split into training (17,143 cases, 70.0%), validation (2449 cases, 10.0%), and test (4899 cases, 20.0%) cohorts using stratification on ASA-PS classes to preserve class distributions across splits (Figure 1). Eligible cases required complete preoperative assessment documentation including clinical narrative and structured perioperative variables used for model input. All eligible cases during the study period were included to maximize model training robustness. Because this study focused on model development rather than hypothesis testing, formal prospective power calculation was not performed. Cases with physiologically implausible values were considered data entry errors and excluded (n=92; Figure 1).

XSL•FO
RenderX

Figure 1. Flow diagram. Outliers were defined as weight less than 25 kg or more than 200 kg (n=21), height less than 120 cm or more than 250 cm (n=70), and hematocrit greater than 75 % (n=1). These values were considered data entry errors and excluded from the analysis. ASA-PS: American Society of Anesthesiologists Physical Status.

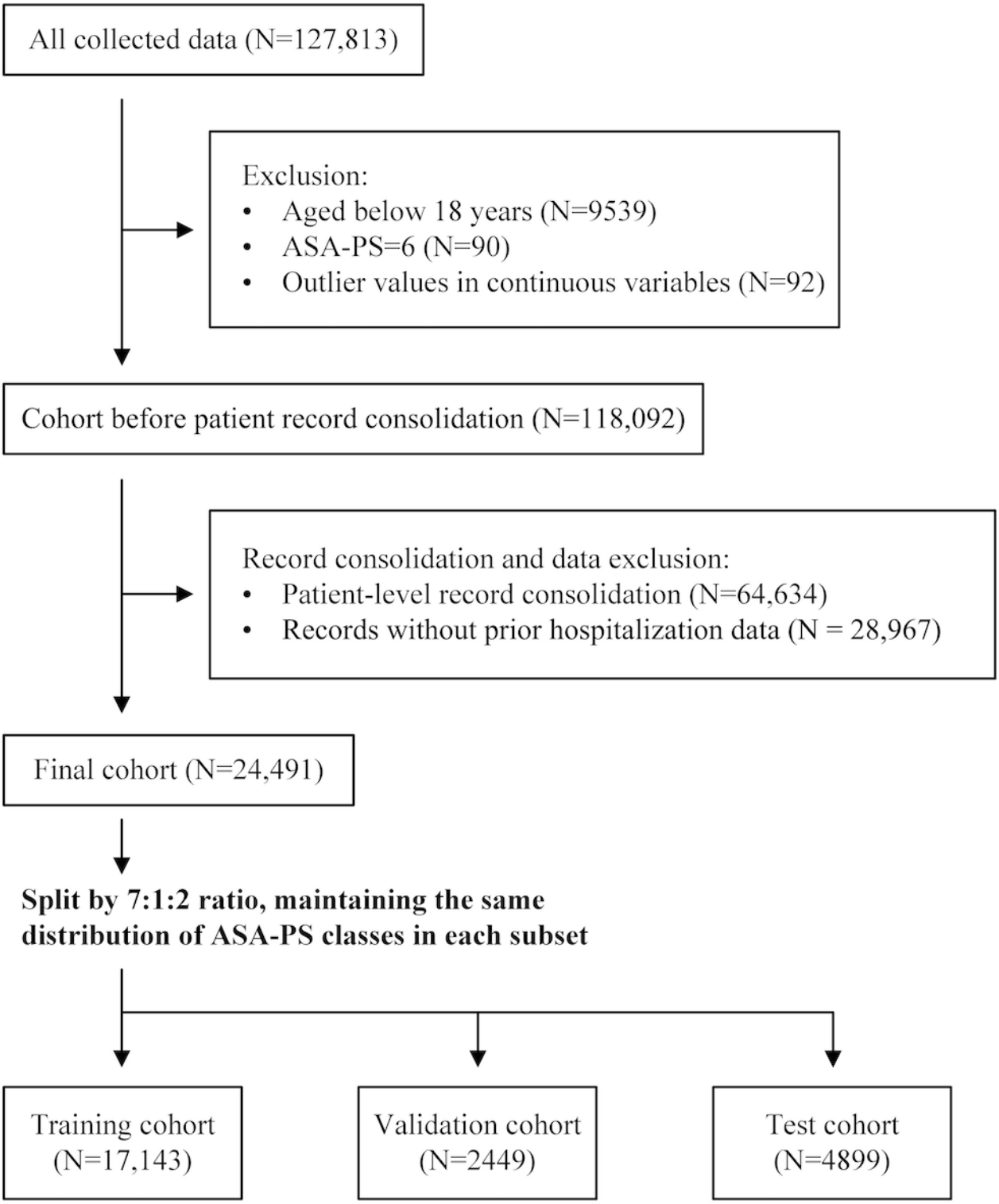

Outcome Definition

The ASA-PS Classification System is a widely adopted framework for assessing patients' preoperative physical condition and stratifying perioperative risk [1]. It provides a standardized means of communicating a patient's overall health status among clinicians, serving as a useful predictor of surgical outcomes and anesthesia-related complications. The classification spans 6 levels, reflecting a progressive increase in systemic disease severity and physiological compromise. Despite its simplicity, the ASA-PS system has demonstrated significant clinical value in risk assessment, particularly when combined with other perioperative variables. Definitions and representative examples for each ASA-PS class are summarized in Table S1 in Multimedia Appendix 1.

XSL•FO
RenderX

Data Preparation

This study used features extracted from preoperative anesthesia assessment records, including patient demographics, comorbidities, vital signs, laboratory results, surgical and anesthesia records, and discharge notes. Demographic and physiological variables included age, sex, BMI, consciousness status, heart rate, blood pressure, oxygen saturation, body temperature, and respiratory rate. Comorbidities included diabetes, hypertension, hyperlipidemia, cerebrovascular disease, cardiac and pulmonary conditions, liver and kidney disorders, bleeding tendency, smoking history, and prior major surgeries. Laboratory tests included hemoglobin, platelet count, coagulation markers, liver and renal function tests, blood glucose, and serum electrolytes. Text reports of chest radiography, electrocardiogram, and other special examinations (eg, echocardiography and spirometry) were also included. Surgical and anesthetic features included the surgical department, type of surgery, anesthesia plan, preoperative diagnosis, and anesthetic route. Information on any special conditions noted by anesthesiologists and records of previous intraoperative adverse events were also included. Additionally, free-text fields from previous discharge notes, such as the chief complaint, present illness, admission diagnosis, diagnostic summary, and treatment plan, were extracted and preprocessed by removing line breaks, nonalphanumeric characters, and standard English stopwords using the Natural Language Toolkit [21]. These cleaned narratives were used as unstructured input for model training. An overview of all included features is provided in Table 1. Clinical narratives were primarily written in English, with occasional Traditional Chinese text appearing within the same note (0.4% of total characters across all input documents; median 0.1% per document).

Table 1. Feature groups included in the models[a].

Feature type	Feature classes
Patient characteristics	
Continuous	Age, BMI, blood pressure, heart rate, body temperature, respiratory rate, and oxygen saturation
Categorical	Sex (n=2) and consciousness status (n=2)
Comorbid conditions	
Categorical	Diabetes mellitus (n=2), hyperlipidemia (n=2), hypertension (n=2), cerebrovascular accident (n=2), cardiac disease (n=2), chronic obstructive pulmonary disease (n=2), asthma (n=2), hepatic disease (n=2), renal disease (n=2), bleeding disorder (n=2), history of major operations (n=2), smoking history (n=2), and drug allergy (n=2)
Preoperative laboratory values	
Continuous	Hemoglobin, platelet count, international normalized ratio, creatinine, aspartate transaminase, alanine transaminase, blood sugar, serum sodium, and serum potassium
Preoperative examination reports	
Free text	Chest radiography, electrocardiogram, and other special examinations
Surgical characteristics	
Categorical	Surgical department (n=22), anesthesia type (n=4), emergency level (n=4), and surgery category (n=4)
Free text	Preoperative diagnosis, proposed procedure, anesthesia plan, special conditions, and prior adverse anesthesia events
Previous discharge notes	
Free text	Chief complaint, present illness, admission diagnosis, diagnosis summary, treatment plan, and record time stamp

[a]Numbers within parentheses indicate the category numbers.

To adapt this dataset for LLMs, we transformed structured and unstructured fields into narrative-style inputs. Relevant features—such as the chief complaint, present illness, diagnosis, treatment plan, anesthesia notes, and preoperative findings—were reformulated into declarative sentences and concatenated into a single clinical paragraph. This format simulates how anesthesiologists typically synthesize patient records in natural language, as illustrated by the red and green text in Figure 2.

XSL•FO
RenderX

Figure 2. Example of an Alpaca-style prompt used for American Society of Anesthesiologists Physical Status (ASA-PS) classification. The prompt is color-coded to illustrate the data structure and model input components: yellow indicates the role assignment and task instruction (eg, physician's role and the purpose of classification); red represents unstructured clinical narratives (eg, chief complaint, present illness, diagnosis, and treatment); and green highlights structured data elements extracted from electronic health records (eg, laboratory values, vital signs, and comorbidities). In this dataset, the structured field labeled "Special conditions" may include current medications or perioperative notes derived from the original preoperative assessment documentation and was preserved during narrative reconstruction to maintain clinical context. Blue denotes the expected model output format. ASA VI was listed for completeness but not used in training or evaluation.

```
{
    "instruction": "As a physician working in a hospital surgery center, you are assessing patients to determine their outcome after a procedure. You can classify patients based on the ASA (American Society of Anesthesiologists) Physical Status Classification to assess their physical condition. Below are the details of each ASA classification: 1: A normal healthy patient. 2: A patient with mild systemic disease. 3: A patient with severe systemic disease. 4: A patient with severe systemic disease that is a constant threat to life. 5: A moribund patient who is not expected to survive without the operation. 6: A declared brain-dead patient whose organs are being removed for donor purposes.",
    "input": "Notes written by Physician at 2039-05-19 13:25:29. CHIEF COMPLAINT: Chronic unhealing wound over left 2nd toe for one month. PRESENT ILLNESS: This 72-year-old man has past medical histories of 1. Hypertension 2. Type 2 diabetes mellitus under OAD 3. Chronic hepatitis B with liver cirrhosis 4. Old cerebral vascular accident. He had regular follow up at local clinic. According to the patient, he had chronic unhealing wound over left 2nd toe after hit the iron for one month. The associated symptom included left foot erythema, local heat and pain. He denied fever or insect bites recently. Due to above reasons, he went to our ID OPD then referred to ER for help on 2019/05/09. At ER, his initial vital signs were BT 36°C, HR 74 bpm, BP 136/104 mmHg, RR 19 /min. Physical examination showed left foot redness/local heat/tenderness with left 2nd toe gangrene wound. The laboratory studies showed no leukocytosis (WBC 4310/μL), Cr 0.95 mg/dL. Empirical Unasyn was given after septic work-up. Under the impression of left DM foot infection with impending bone invasion, he was admitted for further treatment. IMPRESSION: 1. Left diabetic foot infection with impending bone invasion 2. Hypertension 3. Type 2 diabetes mellitus 4. Chronic hepatitis B with liver cirrhosis 5. Old cerebral vascular accident. DIAGNOSIS: Type 2 diabetes mellitus with foot ulcer, Local infection of the skin and subcutaneous tissue, unspecified, Disorder of arteries and arterioles, unspecified, Essential (primary) hypertension, Chronic viral hepatitis B without delta-agent, Unspecified cirrhosis of liver, Personal history of other diseases of the circulatory system. TREATMENT: Dilation of Left Anterior Tibial Artery, Percutaneous Approach, Dilation of Left Posterior Tibial Artery, Percutaneous Approach, Dilation of Left Peroneal Artery, Percutaneous Approach. ANESTHESIA DATA: This 72-year-old underweight male has dm, hyperlipidemia, hypertension, cva, hepatic disease, renal disease, bleeding disorder, major operations, smoking. Preoperative lab tests showed anemia, renal impairment, liver impairment, hyperglycemia and hypokalemia. Vital signs: comatose, heart rate of 77 bpm, blood pressure of 146/80, oxygen saturation of 97%, body temperature of 36.4, and respiratory rate is 14 breaths per minutes. A chest X ray showed Cardiomegaly. The electrocardiogram showed Sinus rhythm. Special conditions included: Aspirin (Bokey) 100 mg/cap, Cilostazol (Pletaal) 100 mg/tab, E4V5M6. This patient is from inpatient and is going to have a Tracheostomy under the diagnosis of respiratory failure by Thoracic Surgery. The urgency level of this surgery is normal. The planned anesthesia is ETGA.",
    "output": "3"
}
```

For structured variables (eg, laboratory values, vital signs, and BMI), raw numeric measurements were converted into clinically interpretable text descriptors using predefined reference ranges and heuristic rules (eg, "anemia" and "renal impairment"). These rules were applied prior to model input and are detailed in Table S2 in Multimedia Appendix 1. Accordingly, the language model operated on pregenerated diagnostic descriptors rather than directly interpreting raw numeric physiologic values. When laboratory values were unavailable, the corresponding descriptors were recorded as "normal," constituting rule-based imputation. Emergency status modifiers (ASA "E") were not included as input features or classification targets in this study.

For model training and inference, we adopted the Alpaca-style [22] instruction-tuning format. Each sample consisted of three fields: (1) an instruction describing the classification task (ie, predicting ASA-PS), (2) an input with the patient summary, and (3) an output with the ground truth ASA-PS class (I-V) assigned by anesthesiologists. This format helps direct the model's focus on the classification goal and facilitates consistent training behavior. An example of the training prompt format is shown in Figure 2.

This format provides a standardized, instruction-following setup for LLMs and enables flexible integration of both structured and unstructured data in a clinically realistic manner.

Model Architecture and Fine-Tuning

To apply LLMs for the classification task of ASA-PS grading, we used the parameter-efficient LoRA [19], which retains general language competence while adapting efficiently to task-specific patterns in limited clinical datasets. We selected Meta-Llama-3-8B-Instruct [23], an instruction-tuned, decoder-only Transformer with approximately 8 billion parameters, as the backbone model and further adapted to ASA-PS classification via LoRA. The base model has already been instruction-tuned on a large, diverse corpus to follow natural language prompts, giving it strong comprehension of complex clinical narratives. These properties make it well suited for analyzing perioperative notes that blend structured elements (vital signs and laboratory values) with free-text observations.

Model training and inference were conducted on the Taiwan Web Service Center platform, developed and operated by the National Center for High-performance Computing. All data preprocessing (see "Data Preparation" section) and model

XSL•FO
RenderX

training were performed within this secure environment, ensuring that no data were transferred to external commercial servers. This setup enhances data privacy, computational efficiency, and administrative control, offering advantages over commercial application programming interface–based solutions.

For fine-tuning, we applied the LoRA technique by inserting trainable low-rank matrices into the sublayers of the transformer model, specifically in the attention and feed-forward layers [19].The inserted matrices $W_A \in R^{r \times d}$ and $W_B \in R^{d \times r}$ form the trainable update term $\Delta W = W_B W_A$, which is added to the original pretrained weight matrix $W \in R^{d \times d}$. Only the LoRA parameters are updated during training, while the original model parameters remain frozen. This design enables efficient adaptation while reducing memory usage and computational cost, making it particularly suitable for clinical applications with limited resources.

As shown in Figure 3, the LoRA modules are integrated into each transformer block of the large language model meta-AI (LLaMA) model, specifically within the grouped multi-query attention (Q/V projections) and SwiGLU feed-forward layers (a "switch" gated linear unit that mixes two 2 linear projections through a gating mechanism) [19]. These modules operate in parallel with the residual connections, ensuring the model's modularity, scalability, and transferability.

Figure 3. Low-rank adaptation module integration in the LLaMA transformer architecture. Low-rank adaptation modules are inserted into the query and value projections of the grouped multiquery attention and the SwiGLU feed-forward layers (a "switch" gated linear unit that mixes 2 linear projections through a gating mechanism) of each transformer block, operating in parallel with the residual connections. RMSNorm, placed before both attention and feed-forward layers, normalize inputs by their root-mean-square values instead of the full mean and variance, lowering computational cost while preserving stability. LoRA: low-rank adaptation.

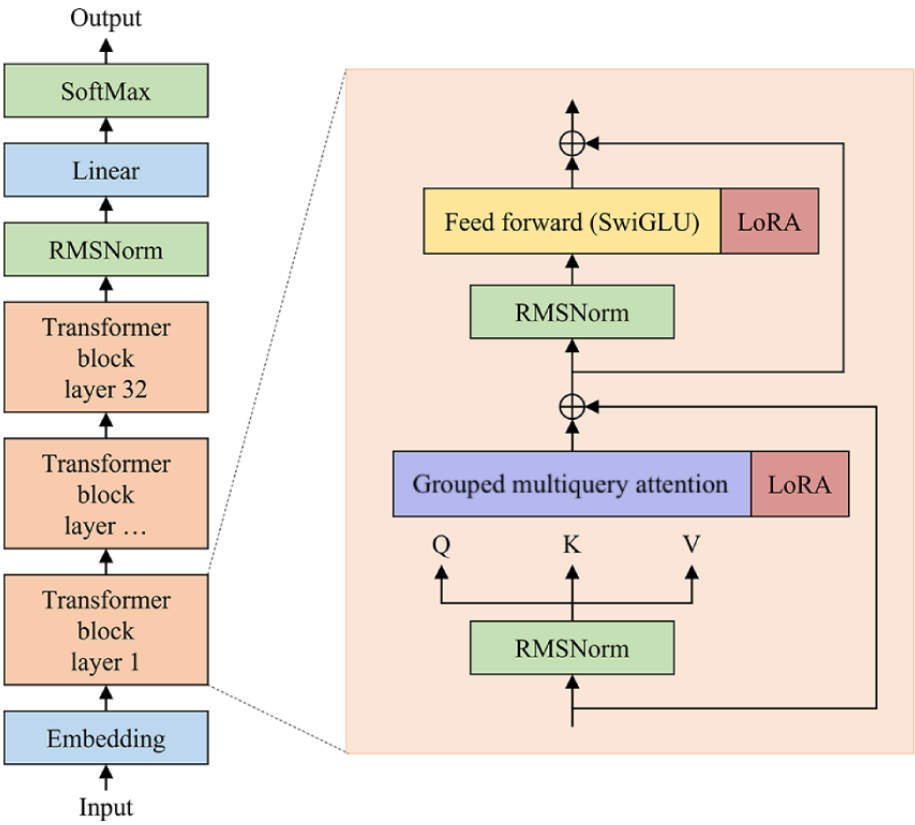

The model input consisted of clinical narratives processed according to the procedure described in the "Data Preparation" section. Inputs were formatted following the Alpaca instruction template, containing 3 fields: instruction, input, and output (Figure 2). The tokenized text was fed into the LoRA-enhanced LLaMA [23] model to predict 1 of the 5 ASA-PS classes.

For model training, we used the Hugging Face accelerate framework [24] and distributed the training across 4 NVIDIA A100 GPUs on the Taiwan Web Service Corporation platform. Mixed-precision (fp16) training [25] was used to optimize memory usage. The batch size was set to 1, with a gradient accumulation step size of 8. The learning rate was set to 3×10^{-5}, and a cosine learning rate scheduler along with a LoRA dropout of 0.3 was used for regularization. Training was performed for 1 epoch, with evaluation and checkpointing every 100 steps. The final model selected was the one achieving the best performance on the validation set.

The reason why we had only 1 epoch is due to the strong prior knowledge embedded in the base model, Meta-LLaMA-3-8B-Instruct [23]. As a large-scale instruction-tuned language model, it already possesses extensive general-domain and instruction-following capabilities. Combined with the relatively limited size of our domain-specific dataset, we observed that the model's performance on the validation set converged quickly and remained stable after a single epoch. Consequently, additional epochs did not yield

XSL•FO
RenderX

noticeable improvements and would have increased computational cost unnecessarily.

Model Evaluation and Statistical Analysis

The LoRA-fine-tuned LLaMA-3 (LLaMA-3-LoRA) and naïve LLaMA-3 were evaluated against several baseline models, including traditional classifiers (Random Forest [10], SVM [9], and Extreme Gradient Boosting [XGBoost] [26]), lightweight embedding-based models (fastText [11]), and biomedical pretrained language models (BioBERT [16] and ClinicalBERT [5]). For comparability, all models consumed the same concatenated instruction + input text; Bidirectional Encoder Representations from Transformers (BERT) models were truncated to 512 tokens and LLaMA-3 formatted in Alpaca style (Figure 2; Multimedia Appendix 2).

For BioBERT and ClinicalBERT, both pretrained and fine-tuned versions were evaluated. Fine-tuning was conducted using Hugging Face's Trainer with a learning rate of 2×10^{-5}, batch size of 8, weight decay of 0.01, and 10 training epochs. The model selection was based on the best microaveraged F_1-score on the validation set. Tokenization was performed using the AutoTokenizer with a maximum input length of 512. When the concatenated instruction and clinical input text exceeded this limit, sequences were truncated by retaining the first 512 tokens, consistent with standard BERT input handling. No sliding-window or hierarchical chunking strategies were applied. The final classification head was trained using cross-entropy loss on 5 ASA-PS classes. Fine-tuned models are denoted by appending the suffix -ft to the original model name (eg, BioBERT-ft and ClinicalBERT-ft).

The traditional ML baselines (Random Forest, SVM, and XGBoost) were trained using the term frequency inverse document frequency features derived from the concatenation of instruction and input fields. To optimize performance, we applied random search with 5-fold cross-validation to identify the best hyperparameter configurations. For Random Forest, the final model used 300 estimators, a maximum depth of 20, a maximum feature setting of "sqrt," and a minimum samples split of 5. The best-performing SVM model used a radial basis function kernel with C set to 10 and gamma set to 0.01. The final XGBoost configuration used a maximum depth of 6, a learning rate of 0.1, 100 estimators, and subsample and "colsample_bytree" (the subsample ratio of columns when constructing each tree) values of 0.8. Hyperparameter selection was based on the best cross-validated F_1-score on the training set.

Models were evaluated using micro- and macroaveraged area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), accuracy, F_1-score, and the Matthews correlation coefficient (MCC). MCC incorporates all cells of the confusion matrix and therefore remains stable under class imbalance [27]. Given the class imbalance in multiclass ASA-PS, F_1-score and MCC were emphasized when comparing models [27,28]. All performance metrics were calculated based on exact agreement between predicted ASA-PS class and the clinician-assigned reference label. Because ASA-PS represents an ordinal clinical scale, adjacent and distant misclassification patterns were additionally analyzed and reported separately to provide clinically interpretable error characterization.

Performance metrics were reported with 95% CIs derived from 1000 bootstrap resamples of the test set. For pairwise comparisons between classifiers evaluated on the same test set, we used McNemar test, which is appropriate for paired comparisons on identical samples. For each model pair, predictions were collapsed into binary outcomes (correct vs incorrect relative to the reference ASA-PS label), and a 2×2 contingency table was constructed. Statistical significance was assessed using McNemar test with continuity correction. Because McNemar test evaluates differences in classification error rates, it was used to assess overall prediction disagreement between models rather than metric-specific differences. To control the family-wise error rate across multiple pairwise comparisons, we applied a Bonferroni adjustment, setting the significance threshold at $.05/m$, where .05 is the α level and m is the number of model pairs. All models were evaluated on the same fixed test split to ensure comparability. Baseline characteristics across the training, validation, and test cohorts were compared using the Kruskal-Wallis test for continuous variables and the chi-square test for categorical variables. These tests were performed as distributional checks because cohorts were generated by stratified random splitting.

Context-Length Controlled Evaluation

To ensure that model performance was not confounded by differences in input context windows, we conducted a control experiment using only test cases whose concatenated instruction and clinical input text contained 512 tokens when tokenized using the standard BERT tokenizer (n=694). This subset ensured identical information availability across architectures without requiring truncation. All models were reevaluated using the same text inputs to isolate the effect of model architecture and fine-tuning strategy from input length differences.

Ablation Analysis of Hyperparameter Effects in LLaMA-3-LoRA

Ablation experiments were conducted to systematically assess how key hyperparameters affect model performance. Each experiment varied a single parameter while keeping the others fixed to isolate its individual contribution. The parameters evaluated included (1) dropout rates (probability of randomly zeroing hidden units to mitigate overfitting): 0.1, 0.2, 0.3, 0.4, and 0.5; (2) learning rates (gradient-descent step size): 1×10^{-5}, 3×10^{-5}, 5×10^{-5}, and 8×10^{-5}; (3) temperature (logit-scaling factor that controls output randomness): 0.1, 0.3, 0.5, 0.7, and 0.9, with *top_p* fixed at 1.0; and (4) *top_p* (nucleus-sampling threshold that limits token selection to the smallest set whose cumulative probability $\leq P$): 0.1, 0.3, 0.5, 0.7, and 0.9, with temperature fixed at 1.0. This design minimized interaction effects between sampling parameters and ensured that observed changes could be attributed to the parameter of interest.

All experiments used the same data preprocessing pipeline (see "Data Preparation" section), model initialization, and LoRA fine-tuning configuration (see "Model Architecture and Fine-Tuning" section) as in the main experiments. This setup

XSL•FO
RenderX

ensured that results were directly comparable and that performance shifts could be confidently linked to the hyperparameter changes being tested.

Attention-Based Analysis of Model Focus

We analyzed the attention weights from the final transformer layer of the LoRA-fine-tuned LLaMA-3 model in order to better understand the internal behavior during classification [23]. The self-attention scores were aggregated across all heads and averaged to represent interaction strengths between token groups [29]. Subword tokens were merged into phrases using white space as a delimiter, and the top 30 phrase pairs with the highest attention were selected. These relationships were visualized using a lower triangular heatmap to reveal attention clusters among clinically meaningful terms. Additionally, attention values directed toward the final output token were collected to rank input phrases based on their contribution to the predicted ASA-PS class.

Model-Generated Explanation

The model was also configured to generate natural language rationales for its ASA-PS classification output. For this purpose, we used an expanded Alpaca-style prompt that appended an explicit request for free-text reasoning to the original classification instruction, while retaining the same clinical narrative input structure (Table S3 in Multimedia Appendix 1). This prompt variant was used exclusively for explainability analysis and not for performance evaluation. The generated explanation commonly referenced key clinical factors such as age, comorbidities, and procedure-related concerns [30]. Representative examples were extracted and qualitatively assessed to verify whether the generated rationales aligned with clinically valid reasoning.

Blinded Evaluation of Generated Rationales

To assess the reliability of model-generated explanations, we conducted a blinded qualitative evaluation using a random sample of cases from the held out test set. A total of 40 cases were randomly selected and stratified to cover ASA-PS classes I through IV. For each case, reviewers were provided with the original clinical narrative and the corresponding model-generated rationale, while predicted ASA class and model identity were blinded. Model outputs were presented in randomized order and deidentified format across cases to minimize potential recognition of model identity.

Rationales were independently evaluated by 2 board-certified anesthesiologists using a predefined assessment form. Reviewers rated (1) factual accuracy using a 3-level scale (no hallucination, minor hallucination [unsupported detail not affecting ASA reasoning], or major hallucination [clinically incorrect or potentially misleading for ASA classification]), and (2) clinical coherence with ASA-PS classification principles (yes or no). Rationale quality was summarized descriptively using proportions across predefined evaluation categories.

Summary statistics were calculated as pooled proportions across raters. Interrater agreement was assessed using weighted k for factual accuracy and unweighted Cohen κ for clinical coherence. Because k statistics can be influenced by prevalence and category distribution, Gwet's first-order agreement coefficient (AC1) was additionally calculated as a complementary agreement measure [31].

Software and Computational Environment

All models were implemented in Python (version 3.10.14; Python Software Foundation) and executed on NVIDIA A100 (80 GB) GPUs with CUDA (version 12.1; NVIDIA Corporation) and cuDNN (version 8.9.2; NVIDIA Corporation). LLM fine-tuning used PyTorch (version 2.3.0; PyTorch Foundation), Transformers (version 4.47.1; Hugging Face), Accelerate (version 1.2.0; Hugging Face), PEFT (version 0.11.1; Hugging Face), bitsandbytes (version 0.42.0; bitsandbytes Foundation), and TRL (version 0.8.6; Hugging Face). Baseline models comprised Random Forest and SVM implemented in scikit-learn (version 1.6.1; scikit-learn developers), XGBoost (version 3.0.1; XGBoost developers), fastText (version 0.9.3; fastText project), and BioBERT/ClinicalBERT via Transformers (version 4.47.1; Hugging Face). Statistical analysis and visualization used NumPy (version 2.2.6; NumPy community), pandas (version 2.2.3; Pandas Development Team), Matplotlib (version 3.10.3; Matplotlib Development Team), and Seaborn (version 0.13.2; Seaborn Development Team).

Results

Cohort Characteristics

In the overall cohort (N=24,491), the median age of the patients was 60 years (IQR 47-69), with 13,775 (56.25%) cases being male. Regarding ASA-PS classification, most cases were categorized as ASA II (15,272, 62.36% cases) and ASA III (8024, 32.76% cases), followed by ASA I (535, 2.18% cases), ASA IV (606, 2.47% cases), and ASA V (54, 0.22% cases). Additionally, 1200 (4.90%) cases of surgeries were marked as ASA emergency status.

In terms of anesthesia type, 12,648 (51.74%) cases underwent general anesthesia with endotracheal tube, followed by 5290 (21.64%) cases with total intravenous anesthesia and 4038 (16.52%) cases with general anesthesia with mask ventilation. Regarding surgical urgency, most cases underwent elective surgery (21,217, 86.64% cases), while 2531 (10.34%) cases had urgent surgeries, and a smaller percentage required emergency surgery (686 cases, 2.80%) or immediate surgery (55, 0.22% cases).

Furthermore, the most common comorbidities among the study population included hypertension (11,315, 46.20% cases), diabetes (6238, 25.47% cases), and cardiac disease (6279, 25.64% cases). In terms of surgical specialties, the most frequently performed surgeries were in urology (6068, 24.78% cases) and orthopedics (3759, 15.35% cases), followed by general surgery (3094, 12.63% cases) and cardiovascular surgery (2904, 11.86% cases; Table S4 in Multimedia Appendix 1).

Comparative Evaluation of Models

Among the language models evaluated, the LLaMA-3-LoRA achieved the highest micro–F_1-score of 0.780 (95% CI 0.769-0.792) and MCC of 0.533 (95% CI 0.518-0.546), while the original LLaMA-3 achieved a micro–F_1-score of 0.073 (95%

XSL•FO
RenderX

CI 0.066-0.081) and MCC of 0.002 (95% CI 0.001-0.002; Table 2).

Table 2. Overall performance comparison across models using microaveraged metricsa.

Model	F_1-score (95% CI)	MCC[b] (95% CI)	Accuracy (95% CI)	AUROC[c] (95% CI)	AUPRC[d] (95% CI)
LLaMA-3-LoRA	0.780 (0.769-0.792)	0.533 (0.518-0.546)	0.780 (0.769-0.792)	0.863 (0.853-0.872)	0.653 (0.639-0.666)
LLaMA-3	0.073 (0.066-0.081)	0.002 (0.001-0.002)	0.073 (0.066-0.081)	0.421 (0.407-0.435)	0.191 (0.180-0.202)
BioBERT-ft	0.762 (0.750-0.774)	0.508 (0.494-0.522)	0.762 (0.750-0.774)	0.852 (0.841-0.861)	0.629 (0.615-0.642)
BioBERT	0.067 (0.060-0.074)	0.062 (0.056-0.062)	0.067 (0.060-0.074)	0.417 (0.403-0.431)	0.191 (0.180-0.202)
ClinicalBERT-ft	0.757 (0.745-0.769)	0.515 (0.501-0.529)	0.757 (0.745-0.769)	0.848 (0.838-0.858)	0.622 (0.608-0.635)
ClinicalBERT	0.624 (0.610-0.637)	0.000 (0.000-0.000)	0.624 (0.610-0.637)	0.765 (0.753-0.777)	0.464 (0.450-0.478)
fastText	0.762 (0.750-0.774)	0.536 (0.522-0.550)	0.762 (0.750-0.774)	0.851 (0.841-0.861)	0.628 (0.614-0.641)
Random Forest	0.794 (0.782-0.805)	0.563 (0.549-0.577)	0.794 (0.782-0.805)	0.871 (0.862-0.880)	0.671 (0.658-0.684)
XGBoost[e]	0.815 (0.804-0.826)	0.613 (0.599-0.626)	0.815 (0.804-0.826)	0.884 (0.875-0.893)	0.701 (0.688-0.714)
Support vector machine	0.809 (0.798-0.820)	0.597 (0.583-0.611)	0.809 (0.798-0.820)	0.880 (0.871-0.889)	0.692 (0.679-0.705)

[a]Accuracy and MCC are reported along with **micro**averaged AUROC, AUPRC, and F_1-score. Values are shown as point estimates with 95% CI obtained via bootstrapping.

[b]MCC: Matthews correlation coefficient.

[c]AUROC: area under the receiver operating characteristic curve.

[d]AUPRC: area under the precision-recall curve.

[e]XGBoost: Extreme Gradient Boosting.

In comparison, fastText achieved lower performance, with a micro–F_1-score of 0.762 (95% CI 0.750-0.774) and MCC of 0.536 (95% CI 0.522-0.550). After fine-tuning on the ASA-PS dataset, BioBERT-ft achieved a micro–F_1-score of 0.762 (95% CI 0.750-0.774) and MCC of 0.508 (95% CI 0.494-0.522). ClinicalBERT-ft obtained similar performance, with a micro–F_1-score of 0.757 (95% CI 0.744-0.770) and MCC of 0.515 (95% CI 0.501-0.529), while their pretrained counterparts performed poorly (BioBERT: micro–F_1-score 0.067, 95% CI 0.060-0.074, and MCC 0.062, 95% CI 0.056-0.062; ClinicalBERT: micro–F_1-score 0.624, 95% CI 0.610-0.637, and MCC 0.000, 95% CI 0.000-0.000).

Among all evaluated models, XGBoost performed best, with a micro–F_1-score of 0.815 (95% CI 0.804-0.826) and MCC of 0.613 (95% CI 0.599-0.626). SVM ranked second with a micro–F_1-score of 0.809 (95% CI 0.798-0.820) and MCC of 0.597 (95% CI 0.583-0.611). Complete metrics for all models are provided in Table S5 in Multimedia Appendix 1.

Although microaveraged metrics reflect overall classification performance, macroaveraged metrics were additionally examined to evaluate class-specific discrimination. LLaMA-3-LoRA achieved a macro–F_1-score of 0.316 (95% CI 0.303-0.329), which was lower than other language model baselines, including BioBERT-ft (0.372), fastText (0.356), and ClinicalBERT-ft (0.349), and also lower than XGBoost (0.348) and SVM (0.322; Table S5 in Multimedia Appendix 1). These findings indicate reduced discriminative performance for minority ASA classes, consistent with the substantial class imbalance in the dataset.

Pairwise comparisons using McNemar test showed significant differences in overall classification error rates between most model pairs (Multimedia Appendix 3). XGBoost showed significantly lower error rates than LLaMA-3-LoRA (P<.001), and LLaMA-3-LoRA showed significantly lower error rates than all other language model baselines. After Bonferroni adjustment for 21 pairwise tests ($a_{(Bonferroni)}$ = 0.0024), these differences remained statistically significant except for comparisons between LLaMA-3-LoRA and BioBERT-ft or fastText. Nonsignificant differences were also observed in comparisons between LlaMA-3 and BioBERT, BioBERT-ft and ClinicalBERT-ft, BioBERT-ft and fastText, ClinicalBERT-ft and fastText, and XGBoost and SVM.

To further illustrate model error patterns, confusion matrices for LLaMA-3-LoRA and XGBoost were analyzed (Figures 4 and 5). For the clinically most common strata (ASA II and III, and to a lesser extent ASA I), most predictions were concentrated along the diagonal, indicating that misclassifications were primarily between adjacent categories rather than distant ones. Distant errors (2 classes apart) were rare, suggesting that both models captured ordinal relationships among ASA-PS levels. In contrast, minority classes (ie, ASA IV and V) exhibited increased off-diagonal errors across all models, reflecting the expected challenge under severe class imbalance. Compared with LLaMA-3-LoRA, XGBoost showed slightly fewer adjacent-class errors for ASA II-III, but off-diagonal errors increased for the minority classes (ASA IV-V), reflecting the persistent impact of class imbalance.

XSL•FO
RenderX

Figure 4. Confusion matrix for ASA-PS classification using the LLaMA-3-LoRA model. Most predictions for the dominant ASA I-III classes were concentrated along the diagonal, indicating primarily adjacent-class misclassifications. Off-diagonal errors were rare except in minority classes (ASA IV-V), reflecting the challenge of extreme class imbalance. ASA-PS: American Society of Anesthesiologists Physical Status.

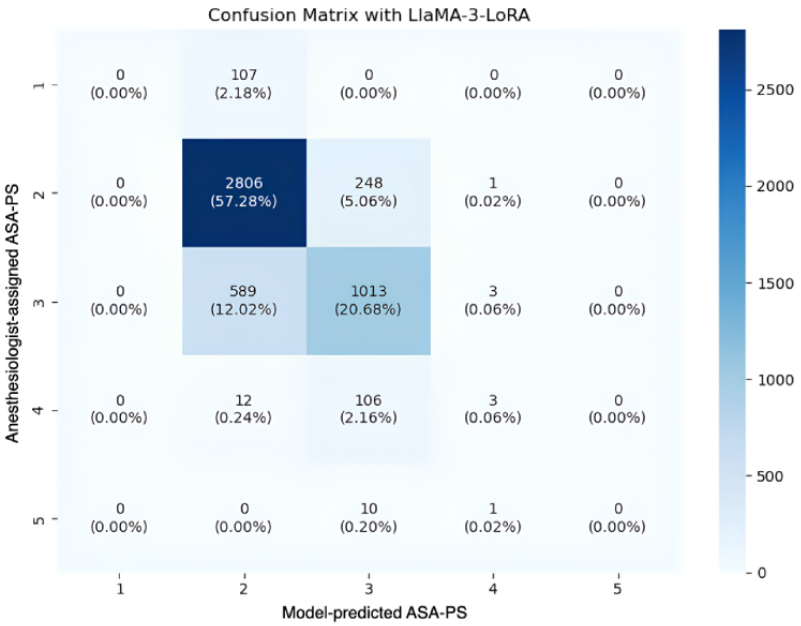

Figure 5. Confusion matrix for ASA-PS classification using the XGBoost model. Most predictions for the major ASA I-III classes were aligned along the diagonal, indicating predominantly adjacent-class misclassifications. Compared with LLaMA-3-LoRA, XGBoost showed slightly fewer adjacent-class errors for ASA II-III, but off-diagonal errors increased for the minority classes (ASA IV-V), reflecting the persistent impact of class imbalance. ASA-PS: American Society of Anesthesiologists Physical Status; XGBoost: Extreme Gradient Boosting.

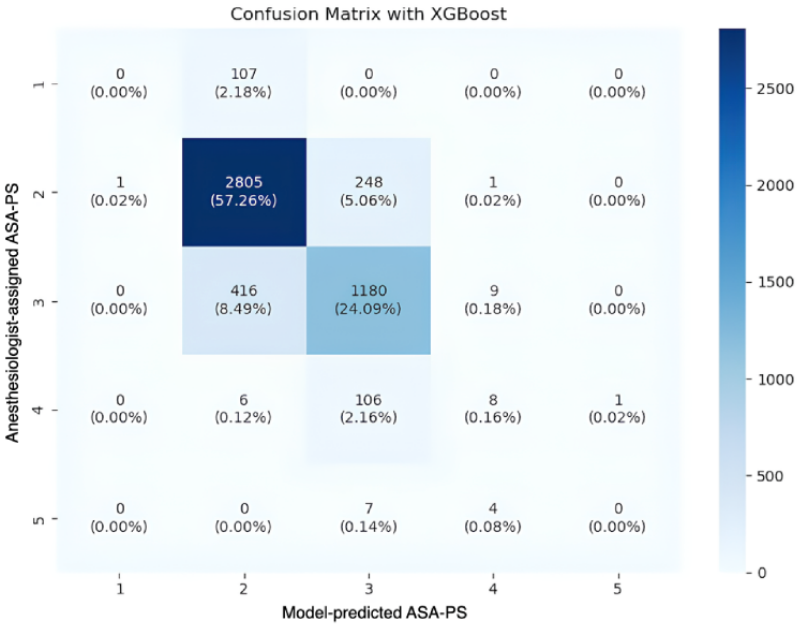

Context-Length Controlled Evaluation

Among the evaluated models under the input length ≤512 tokens (n=694), the LLaMA-3-LoRA achieved the highest F_1-score of 0.879 (95% CI 0.855-0.903) and MCC of 0.297 (95% CI 0.263-0.331), SVM achieving an F_1-score of 0.876 (95% CI 0.852-0.901) and MCC of 0.238 (95% CI 0.206-0.269), and XGBoost achieving an F_1-score of 0.873 (95% CI 0.848-0.898) and MCC of 0.198 (95% CI 0.168-0.227). Fine-tuned BERT models followed, with ClinicalBERT-ft and BioBERT-ft obtaining F_1-scores of 0.847 (95% CI 0.820-0.874) and 0.833 (95% CI 0.805-0.861) and MCC values of 0.261 (95% CI 0.228-0.293) and 0.187 (95% CI 0.158-0.216), respectively. fastText showed slightly lower performance, with an F_1-score of 0.814 (95% CI 0.785-0.843) and MCC of 0.202 (95% CI 0.172-0.232). Interestingly, the untuned ClinicalBERT achieved a relatively high F_1-score of 0.869 (95% CI 0.844-0.894) but an MCC of 0.000, suggesting that its predictions were concentrated in a single majority class (ASA-PS class II). Other

XSL•FO
RenderX

untuned transformer models, including LLaMA-3 and BioBERT, both performed poorly (F_1-score ≤0.089; Table 3).

Table 3. Performance comparison across models under controlled input length (≤512 tokens)[a].

Model	F_1-score (95% CI)	MCC[b] (95% CI)	Accuracy (95% CI)	AUROC[c] (95% CI)	AUPRC[d] (95% CI)
LLaMA-3-LoRA	0.879 (0.855-0.903)	0.297 (0.263-0.331)	0.879 (0.855-0.903)	0.919 (0.899-0.940)	0.803 (0.773-0.832)
LLaMA-3	0.089 (0.068-0.111)	0.000 (0.000-0.000)	0.089 (0.068-0.111)	0.431 (0.394-0.468)	0.190 (0.161-0.219)
BioBERT-ft	0.833 (0.805-0.861)	0.187 (0.158-0.216)	0.833 (0.805-0.861)	0.889 (0.864-0.911)	0.735 (0.702-0.768)
BioBERT	0.059 (0.042-0.077)	0.000 (0.000-0.000)	0.059 (0.042-0.077)	0.373 (0.336-0.408)	0.239 (0.206-0.269)
ClinicalBERT-ft	0.847 (0.820-0.874)	0.261 (0.228-0.293)	0.847 (0.820-0.874)	0.898 (0.875-0.920)	0.756 (0.723-0.787)
ClinicalBERT	0.869 (0.844-0.894)	0.000 (0.000-0.000)	0.869 (0.844-0.894)	0.913 (0.891-0.933)	0.788 (0.756-0.817)
fastText	0.814 (0.785-0.843)	0.202 (0.172-0.232)	0.814 (0.785-0.843)	0.876 (0.852-0.901)	0.709 (0.675-0.743)
Random Forest	0.870 (0.845-0.895)	0.099 (0.076-0.120)	0.870 (0.845-0.895)	0.914 (0.893-0.934)	0.790 (0.759-0.820)
XGBoost[e]	0.873 (0.848-0.898)	0.198 (0.168-0.227)	0.873 (0.848-0.898)	0.915 (0.894-0.936)	0.794 (0.764-0.824)
Support vector machine	0.876 (0.852-0.901)	0.238 (0.206-0.269)	0.876 (0.852-0.901)	0.917 (0.896-0.937)	0.798 (0.768-0.828)

[a]Accuracy and MCC are reported along with microaveraged AUROC, AUPRC, and F_1-score. Values are shown as point estimates with 95% CI obtained via bootstrapping. See Table S5 in Multimedia Appendix 1 for macroaveraged metrics.

[b]MCC: Matthews correlation coefficient.

[c]AUROC: area under the receiver operating characteristic curve.

[d]AUPRC: area under the precision-recall curve.

[e]XGBoost: Extreme Gradient Boosting.

Ablation Analysis of Hyperparameter Effects in LLaMA-3-LoRA

Dropout

As shown in Table S6 in Multimedia Appendix 1, a dropout rate of 0.4 yielded the best performance (F_1-score 0.782, 95% CI 0.771-0.794; MCC 0.538, 95% CI 0.524-0.552). A slightly lower rate of 0.3 yielded similar results (F_1-score 0.780, 95% CI 0.769-0.792; MCC 0.533, 95% CI 0.518-0.546).

Learning Rate

As shown in Table S7 in Multimedia Appendix 1, a learning rate of 3×10^{-5} yielded the best performance (F_1-score of 0.780, 95% CI 0.769-0.792; MCC 0.533, 95% CI 0.518-0.546). In comparison, both a lower rate (1×10^{-5}) and a higher rate (8×10^{-5}) reduced MCC scores (0.111, 95% CI 0.102-0.119 and 0.440, 95% CI 0.426-0.454, respectively).

Temperature

As shown in Table S8 in Multimedia Appendix 1, a temperature of 0.1 yielded the best performance (F_1-score of 0.777, 95% CI 0.765-0.788; MCC 0.526, 95% CI 0.512-0.540). As the temperature increased to 0.9, performance gradually declined, with the MCC dropping to 0.195 (95% CI 0.183-0.206).

Top-P (Nucleus Sampling)

As shown in Table S9 in Multimedia Appendix 1, *top_p* values of 0.1 yielded the best performance (F_1-score 0.782, 95% CI 0.771-0.794; MCC 0.538, 95% CI 0.524-0.552), closely followed by 0.3 (F_1-score 0.782, 95% CI 0.770-0.793; MCC 0.536, 95% CI 0.522-0.550). When *top_p* value was raised to 0.5 and 0.9, MCC dropped to 0.161 (95% CI 0.151-0.171) and 0.179 (95% CI 0.168-0.189), respectively.

Final Configuration

Subsequent experiments used a dropout rate of 0.3, a learning rate of 3×10^{-5}, a temperature of 0.1, and a *top_p* value of 0.1—the combination that maximized MCC and F_1-score during hyperparameter tuning. Although a dropout rate of 0.4 slightly outperformed 0.3, the difference was not statistically significant, and the smaller value (0.3) was retained to mitigate the risk of underfitting and overregularization typically associated with excessive dropout in LLM fine-tuning [32]. This decision is similar with common practice in LoRA fine-tuning (≈0.05-0.1 for LLaMA models) [33].

Attention-Based Analysis of Model Focus

The attention heatmap for a case of ASA-PS class III (Figure 6) revealed strong intrasection interactions among clinical terms, such as “cirrhosis,” “redness,” and “hypertension,” indicating local contextual coherence. Additionally, Figure 7 visualizes the top 30 input phrases receiving the highest cumulative attention toward the output token for 1 example from the test set, including “cirrhosis,” “nonhealing,” “diabetes,” and “72-year-old,” reflecting the model’s emphasis on comorbidities and age-related indicators during ASA-PS classification.

XSL•FO
RenderX

Figure 6. Attention heatmap between top 30-word phrases in the input text for 1 case (ASA-PS class III) from the test cohort. The lower triangular matrix shows average unidirectional attention weights from each phrase to preceding phrases derived from the final transformer layer. Brighter regions indicate stronger attention. Medical terms tend to cluster together, reflecting localized attention focus. Representative attention visualizations for ASA-PS classes II and IV are provided in Multimedia Appendix 4.

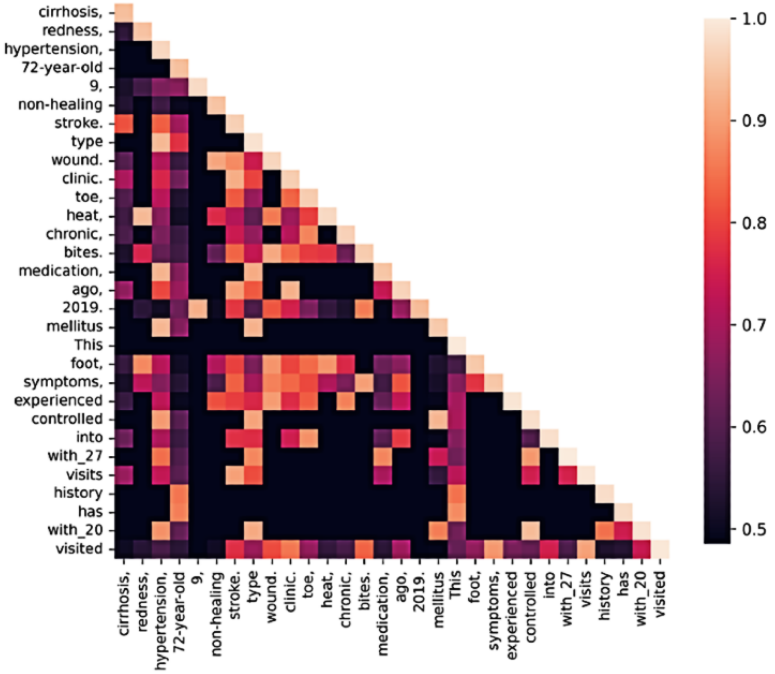

Figure 7. Top 30 most attended phrases contributing to the American Society of Anesthesiologists Physical Status (ASA-PS) output token for 1 case (ASA-PS class III) from the test cohort. Bar height represents cumulative attention directed toward each phrase by the final output token. The highlighted phrases reflect the model's focus on comorbidities and demographic cues relevant to perioperative risk. Representative attention visualizations for ASA-PS classes II and IV are provided in Multimedia Appendix 4. ASA: American Society of Anesthesiologists.

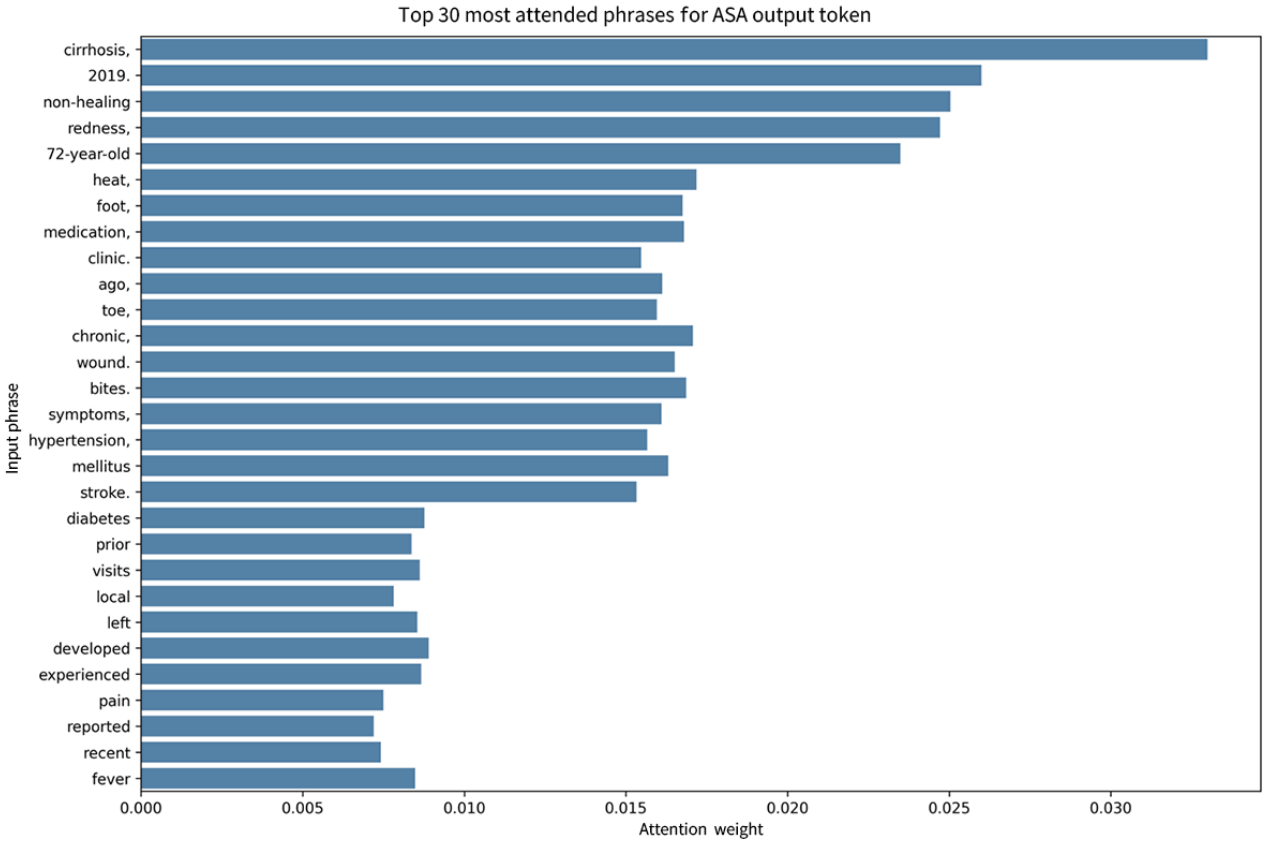

Comparable attention patterns were observed in additional representative cases from ASA-PS classes II and IV, where attention remained concentrated on comorbidity descriptors (eg, hydronephrosis, cardiomegaly, and encephalopathy) and

XSL•FO
RenderX

physiologic abnormalities (eg, hypokalemia and leukocytosis). However, occasional high-attention tokens with limited direct clinical relevance were also observed, reflecting the probabilistic nature of attention mechanisms [34]. Representative visualizations for ASA-PS classes II and IV are provided in Multimedia Appendix 4.

Model-Generated Explanation

Table S10 in Multimedia Appendix 1 presented 10 model-generated explanations for ASA classes assignment. The illustrative examples were generated using a separate explanation-augmented prompt and do not correspond to the deterministic inference pathway used for quantitative performance evaluation. For class III, the model repeatedly referenced multiple chronic conditions (eg, diabetes, hypertension, coronary artery disease, renal failure, and bleeding disorders), recent cardiac evaluations, and limited mobility due to comorbidities. Explanations for lower classes (I-II) focused on stable or well-managed chronic conditions, preserved organ function, and low surgical risk. In contrast, output for the higher classes (IV-V) highlighted critically ill patients with recent cardiovascular events, hemodynamic instability, or emergency surgical requirements. Full narrative inputs are provided in Table S11 in Multimedia Appendix 1.

Blinded Evaluation of Generated Rationales

In blinded qualitative evaluation of 40 randomly sampled cases, LLaMA-3-LoRA demonstrated high factual reliability and clinical coherence (Table S12 in Multimedia Appendix 1). Across both raters, LLaMA-3-LoRA rationales contained no hallucination in 86.25% of evaluations, minor hallucination in 7.50%, and major hallucination in 6.25%. Clinical coherence with ASA-PS principles was observed in 91.25% of LLaMA-3-LoRA rationales.

For LLaMA-3-LoRA, Cohen k indicated fair to moderate agreement (κ=0.22-0.44), whereas observed agreement remained high (≥87.5%). Gwet's AC1 demonstrated substantial to almost perfect reliability (AC1 = 0.85-0.87). The discrepancy between κ and observed agreement likely reflects the high prevalence of top-category ratings, a known limitation of κ statistics [31]. Accordingly, κ and AC1 were reported and interpreted jointly to provide complementary perspectives on interrater agreement (Table S13 in Multimedia Appendix 1).

Discussion

Principal Findings

This study aimed to evaluate whether parameter-efficient fine-tuning of a small LLM could provide accurate, reproducible, and clinically interpretable ASA-PS classification from preoperative clinical narratives. From a clinical perspective, our findings suggest that such models may serve as scalable decision support tools to assist ASA documentation and perioperative risk communication.

Our study demonstrates the performance of the LLaMA-3-LoRA model in classifying ASA-PS based on preoperative clinical narratives. The model achieved a microaveraged F_1-score of 0.780 (95% CI 0.769-0.792) and an MCC of 0.533 (95% CI 0.518-0.546), outperforming all baseline language models, including the original, untuned LLaMA-3 (F_1-score 0.073, MCC 0.002), as shown in Table 2. However, macroaveraged performance revealed reduced class-level discrimination. LLaMA-3-LoRA achieved a macro–F_1-score of 0.316 (95% CI 0.303-0.329), which was lower than other language model baselines (0.349-0.372), XGBoost (0.348), and SVM (0.322; Table S5 in Multimedia Appendix 1). This discrepancy likely reflects substantial class imbalance, particularly for minority ASA classes. Pairwise comparisons using McNemar test further showed that XGBoost had significantly lower error rates than LLaMA-3-LoRA (P<.001), whereas differences between LLaMA-3-LoRA and other language model baselines were not statistically significant after Bonferroni adjustment (Multimedia Appendix 3).

While traditional ML models such as XGBoost, Random Forest, and SVM achieved higher scores on several discrimination metrics (eg, macro–F_1-score, AUROC, and MCC), they process text through manually engineered token-based features without the semantic understanding offered by contextual embeddings in LLMs. In contrast, our model incorporates unstructured clinical documentation and outputs both predictions and explanations.

The relative advantage of traditional machine learning models in our study may also reflect the structural characteristics of the ASA-PS classification task. Although ASA assignment is documented in narrative clinical notes, much of the discriminative information is conveyed through semistructured clinical descriptors, such as explicit comorbidity labels, physiologic abnormalities, and standardized diagnostic terminology. Such features can be efficiently captured by sparse feature representations and tree-based classifiers. Prior work comparing LLMs with locally trained machine learning models in EHR-based prediction tasks has similarly demonstrated superior performance of traditional ML when predictive signals are dominated by structured or semistructured features [35]. These findings suggest that task structure, rather than model scale alone, plays a critical role in determining comparative model performance.

In addition to aggregate metrics, we examined the classwise misclassification table which separates adjacent (±1 class) from distant (≥2 classes) errors (Table S14 in Multimedia Appendix 1). For the 3 most common categories, ASA-PS classes I, II, and III, the LLaMA-3-LoRA model showed comparable accuracy to the best-performing ML models (eg, XGBoost and SVM): distant errors were essentially absent, and the few mistakes that did occur were almost exclusively adjacent (eg, ASA I cases predicted as ASA II). Such single-step slips are clinically less consequential because the boundaries between neighboring ASA classes are inherently ambiguous. In contrast, the sparsely represented ASA-PS IV and V classes proved challenging for every model, with both adjacent and distant errors rising sharply, an expected consequence of extreme class imbalance. Taken together, these findings suggest that instruction-tuned LLM can approach but did not surpass state-of-the-art machine learning algorithms in class-specific reliability for the clinically prevalent strata, while also offering

XSL•FO
RenderX

the added benefits of interpretability and flexibility. Across all evaluated models, prediction of rare ASA classes (I and V) remained poor, with most models concentrating predictions within intermediate severity levels (Multimedia Appendix 5). Similar limitations have been reported in prior ASA classification studies [13,15], suggesting that automated ASA prediction remains intrinsically challenging and may not yet fully operationalizable for extreme severity strata.

In addition, ASA-PS assignment itself is known to demonstrate moderate interrater variability among anesthesiologists [36]. Therefore, some adjacent-class discrepancies between model predictions and recorded ASA labels may reflect clinically acceptable differences in interpretation rather than definitive model error. This inherent uncertainty in the reference standard should be considered when interpreting classification performance, particularly between neighboring ASA categories such as classes II and III.

In a context-length controlled experiment (≤512 tokens), LLaMA-3-LoRA maintained the highest micro–F_1-score of 0.879 (95% CI 0.855-0.903) and MCC of 0.297 (95% CI 0.263-0.331; Table 3). However, under the same input constraint, the smaller domain-specific BioBERT-ft achieved higher macro–F_1-score (0.469 vs 0.333), macro-AUROC (0.620 vs 0.555), and macro-AUPRC (0.347 vs 0.294) than LLaMA-3-LoRA, whereas LLaMA-3-LoRA retained higher micro-AUROC (0.919 vs 0.889) and micro-AUPRC (0.803 vs 0.735; Table S15 in Multimedia Appendix 1). These findings suggest that part of LLaMA-3-LoRA's advantage in the full-input setting may reflect its ability to leverage longer contextual information rather than inherent architectural superiority. This observation is consistent with prior work demonstrating that context-aware language models remain relatively robust compared with alternative algorithms when input length is constrained [13]. Additionally, the presence of mixed English and Traditional Chinese text in the clinical narratives may have posed challenges for BERT-based models due to tokenizer and vocabulary limitations, as these models were primarily pretrained on English corpora.

The marked performance gap between the untuned LLaMA-3 and LLaMA-3-LoRA suggests that instruction-guided LoRA fine-tuning effectively restores task-specific alignment that is not preserved during general purpose LLaMA-3 pretraining [19,23]. Qualitative inspection of representative model outputs suggests that LoRA fine-tuning yields may facilitate more concise and clinically grounded rationales (Table S16 in Multimedia Appendix 1; using the same 10 cases as Table S10 in Multimedia Appendix 1). To complement illustrative examples, we performed a blinded qualitative evaluation using independent anesthesiologist reviewers. The LLaMA-3-LoRA demonstrated high factual accuracy (86.25%) and clinical coherence (91.25%), with fair to moderate interrater agreement (Tables S12 and S13 in Multimedia Appendix 1). These findings provide preliminary quantitative support for the reliability of generated rationales, although hallucinations were not completely eliminated and explanations should still be interpreted cautiously. In addition, the untuned model produced instruction-like or template text rather than clinically grounded explanations in 10 out of 40 cases, reflecting limited instruction-following capability and representing a potential source of bias in the qualitative evaluation (Table S12 in Multimedia Appendix 1).

Data Representation and Input Engineering

It should be noted that structured laboratory and vital sign variables were converted into clinically interpretable text descriptors prior to model input (Table S2 in Multimedia Appendix 1). This preprocessing step externalizes part of the clinical interpretation and may simplify the classification task. In addition, ML baselines were trained using the term frequency inverse document frequency representations rather than native structured tabular variables. Therefore, baseline performance comparisons should be interpreted with these design considerations in mind.

Although narrative reconstruction facilitates integration of heterogeneous clinical inputs and mirrors clinician reasoning, this approach introduces an additional preprocessing requirement that may limit immediate scalability across institutions. In practice, many of the transformation rules applied in this study are derived from standard clinical reference thresholds and could potentially be automated within EHR decision support pipelines. Future work should evaluate whether similar performance can be achieved using hybrid approaches that allow language models to directly interpret raw structured data or incorporate automated feature extraction methods.

Sensitivity Analysis Excluding Anesthesia Planning Variables

From a causal perspective, anesthesia planning may represent a downstream clinical decision partially influenced by perceived perioperative risk and ASA classification, raising the possibility of proxy-label learning. In our sensitivity analysis, excluding anesthesia planning resulted in a reduction in performance, with LLaMA-3-LoRA achieving an F_1-score of 0.734 (95% CI 0.721-0.746) and MCC of 0.422 (95% CI 0.408-0.436), compared with 0.780 (95% CI 0.769-0.792) and 0.533 (95% CI 0.518-0.546) in the full-input model (Table S17 in Multimedia Appendix 1). Clinically, anesthesia planning reflects integrated evaluation of comorbidities, physiologic reserve, and surgical risk rather than ASA classification alone. These results indicate that anesthesia planning contributes to incremental predictive information but does not fully account for model performance.

Comparing With Prior Studies

Compared with the prompt-based approach of Chung et al [37], which evaluated instruction-tuned GPT-4 Turbo without additional parameter training, our LLaMA-3-LoRA achieved substantially higher scores across evaluated metrics. Chung et al [37] reported a microaveraged F_1-score of 0.50 (95% CI 0.46-0.53) and an MCC of 0.36 (95% CI 0.32-0.40), whereas LLaMA-3-LoRA attained a microaveraged F_1-score of 0.780 (95% CI 0.769-0.792) and an MCC of 0.533 (95% CI 0.518-0.546). In contrast, the untuned LLaMA-3 model performed worse than the reported GPT-4 results by Chung et al, with a microaveraged F_1-score of 0.073 (95% CI

XSL•FO
RenderX

0.066-0.081) and an MCC of 0.002 (95% CI 0.001-0.002), highlighting the importance of task-specific model adaptation rather than inherent architectural superiority.

Beyond superior predictive accuracy, our pipeline is fully transparent and reproducible: it relies on publicly available LLaMA-3 weights and open-source LoRA fine-tuning code, in contrast to GPT-4 Turbo's closed model weights. On macroaveraged discrimination, Chung et al [13] reported macro-AUROC 0.865 for fastText and 0.843 for BioClinicalBERT, whereas in our study macro-AUROC was 0.608 for LLaMA-3-LoRA and 0.637 for fastText (Table S5 in Multimedia Appendix 1). Relative to Yoon et al [15], our micro-AUROC for LLaMA-3-LoRA (0.863, 95% CI 0.853-0.872) and ClinicalBERT-ft (0.848, 95% CI 0.838-0.858) were lower than their ClinicalBigBird (0.914, 95% CI 0.909-0.919). This difference likely reflects cohort and label-schema differences (our 5-class, more imbalanced setting vs their 4-class dataset), as well as model and pretraining differences.

In our cohort, traditional classifiers were strong comparators: XGBoost micro-AUROC 0.884 (95% CI 0.875-0.893), SVM 0.880 (95% CI 0.871-0.889), and Random Forest 0.871 (95% CI 0.862-0.880; Table 2). These results are consistent with prior ASA-PS work using structured inputs, including a C-statistic of 0.77 (95% CI 0.766-0.773) for ASA-PS prediction [8] and AUROC 0.884 for an automated perioperative acuity system [14]. Collectively, this indicates that our ML baselines are competitive with the literature and provide a credible benchmark for evaluating LLaMA-3-LoRA.

Interpretability

Interpretability is another key differentiator. The LoRA-adapted model produces concise, clinically coherent rationales when prompted for explanation. Rationale generation and classification evaluation were conducted using distinct inference configurations; the representative cases shown in Table S10 in Multimedia Appendix 1 were generated under the explanation-augmented prompt and are provided for qualitative illustration only. In these examples (Table S10 in Multimedia Appendix 1), the explanations show that these explanations reference patient age, salient comorbidities, recent cardiac events, and functional limitations—especially in ASA-PS class III and above—to justify the assigned grade. Attention visualizations further confirm that the model focuses on terms such as diabetes, cirrhosis, and stroke, mirroring clinical reasoning (Figures 6 and 7).

Notably, LoRA fine-tuning in this study optimized classification labels only, and free-text rationales were not included as supervised training targets; therefore, explanations were generated during inference using the instruction-following capability of the pretrained base model. However, prior work has shown that explanation-augmented prompting may influence model decision behavior rather than merely reveal it, and that generated rationales can be plausible yet not fully faithful to the underlying prediction process [38]. Accordingly, in this study, explanation-augmented prompts were used exclusively for qualitative illustration and not for quantitative performance evaluation. Despite this limitation, this capacity for transparent, case-specific explanation may facilitate clinician trust and support adoption in real-world perioperative settings.

Attention visualization frequently highlighted clinically relevant comorbidity and physiologic descriptors; however, some high-attention tokens were nonspecific or administrative, suggesting that attention signals may reflect statistical correlations rather than faithful clinical reasoning (Figure 6 and 7; Multimedia Appendix 4). Therefore, attention-based analyses should be interpreted as exploratory qualitative insights rather than definitive mechanistic explanations [34].

Generative Oversampling for Class Imbalance

To address class imbalance, we evaluated a generative data augmentation strategy targeting a uniform distribution of 3000 cases per ASA class. We generated synthetic samples for minority classes (ASA I, IV, and V) using an equal mix of ChatGPT and Gemini, while downsampling majority classes (ASA II and III). However, this approach resulted in inferior overall performance compared with the baseline LLaMA-3-LoRA configuration, yielding an F_1-score of 0.696 (95% CI 0.683-0.709) and an MCC of 0.408 (95% CI 0.394-0.422; Table S17 in Multimedia Appendix 1). Although this strategy successfully balanced the dataset, the synthetic clinical narratives likely introduced linguistic artifacts or semantic drift. This finding aligns with prior research suggesting that naive generative oversampling can introduce label noise and compromise model calibration without enhancing the representation of rare classes [39,40].

We additionally evaluated a loss-based class reweighting strategy during model training to address class imbalance, with class weights assigned inversely proportional to class frequency. However, this approach did not improve overall performance compared with the baseline LoRA configuration, resulting in a slightly lower micro–F_1-score (0.769 vs 0.780) and MCC (0.506 vs 0.533; Table S17 in Multimedia Appendix 1). Micro-AUROC (0.856 vs 0.863) and micro-AUPRC (0.638 vs 0.653) were also modestly reduced. These findings further support the notion that performance limitations in minority ASA classes may reflect intrinsic ambiguity in ASA severity boundaries rather than sample frequency alone.

Subgroup Analysis

To assess the consistency of model performance across demographic and clinical strata, we conducted subgroup analyses by sex, age, and comorbidity burden. The LLaMA-3-LoRA model demonstrated stable performance across these subgroups, with F_1-scores of 0.770 for males and 0.793 for females. Across age categories, F_1-scores ranged from 0.711 to 0.848 for age subgroups and across comorbidity-count strata from 0.771 to 0.797. With the exception of the ≥65-year age group, no statistically significant differences were observed, suggesting that the model exhibits generally robust performance across diverse demographic and clinical characteristics (Table S18 in Multimedia Appendix 1).

Ablation Analysis in LLaMA-3-LoRA

Ablation analysis revealed that sampling parameters, especially temperature and *top_p*, had the greatest effect on model

XSL•FO
RenderX

performance. Reducing temperature from 0.9 to 0.1 improved MCC from 0.195 to 0.526 (Table S8 in Multimedia Appendix 1), while lowering *top_p* from 0.9 to 0.1 increased MCC from 0.179 to 0.538 (Table S9 in Multimedia Appendix 1). These results suggest that limiting generation randomness improves the consistency and accuracy of instruction-following behavior. While learning rate (Table S7 in Multimedia Appendix 1, optimal at 3×10) and dropout (Table S6 in Multimedia Appendix 1, optimal at 0.4) also contributed to performance, their relative impact was smaller.

Limitations

This investigation has several methodological constraints. First, the dataset was markedly imbalanced—ASA-PS classes I, IV, and V accounted for a small fraction of cases—raising the risk of class-specific misclassification and limiting the validity of predictions for these groups. Notably, the model demonstrated markedly reduced sensitivity for ASA class IV, with most cases misclassified as class III (Figure 4). In addition, all ASA class I cases were misclassified in the test set, indicating limited discrimination at the extremes. Consequently, predictions were concentrated within intermediate classes (ASA II-III). This distinction is clinically important because ASA class IV represents patients with life-threatening systemic disease and substantially higher perioperative risk. The reduced sensitivity likely reflects both severe class imbalance and limited preoperative documentation regarding whether systemic disease constitutes a constant threat to life. These findings are consistent with prior ASA classification studies reporting reduced performance for rare severity classes [13,15]. Therefore, the model should be considered an adjunctive decision support or documentation assistance tool rather than a stand-alone risk classification system, and clinician oversight remains essential.

Second, model development and internal validation were confined to a single tertiary medical center, so performance may not translate to institutions with different patient populations, documentation styles, or EHR systems. In particular, our inputs included free text from standardized sections (chief complaint, present illness, admission diagnosis, impression, and treatment plan), whereas anesthesia-related data were derived from structured fields (Table S2 in Multimedia Appendix 1). This reliance on templated narratives may aid portability to institutions with similar section-based documentation, but transfer to settings with free-form or less standardized clinical notes could be more challenging. Approaches such as section header normalization, abbreviation mapping, or light site-specific adaptation may help mitigate these differences.

Third, the study's retrospective design relies on the accuracy and completeness of clinician-authored notes; inaccuracies in free text or structured fields could introduce bias that the LLM might amplify. In addition, because ASA classification is inherently subjective and demonstrates interrater variability, some discrepancies between model predictions and reference labels may reflect uncertainty in the ground truth rather than true model error [36]. In settings where key physiologic severity descriptors are incompletely documented, future approaches incorporating external medical knowledge sources may help improve reasoning robustness [41]. Additionally, missing laboratory values were treated as normal during preprocessing, which may introduce information bias and limit generalizability, particularly among low-risk patients who did not undergo comprehensive testing. Furthermore, anesthesiologists may incorporate information from external or nondigitized sources (eg, scanned referral documents, outside hospitalization records, or bedside clinical assessment) that were not available in the structured or extractable electronic documentation used for model input. This information accessibility gap may contribute to differences between model predictions and real-world clinical ASA assignment and may limit reproducibility across institutions with heterogeneous documentation practices.

Fourth, although attention weights and exemplars were provided, LLM rationales remain probabilistic and may occasionally generate spurious associations ("hallucinations"), which were not systematically quantified.

Fifth, although no systematic performance differences were observed across sex, age, and comorbidity strata (Table S18 in Multimedia Appendix 1), computational latency and clinician usability, which are essential for real-time perioperative decision support, were not evaluated in this study. A feasible integration pathway would involve deploying the model as a read-only decision-support module in the preoperative clinic, presenting a predicted ASA-PS class with a rationale, allowing one-click clinician confirmation, and capturing feedback for audit and iterative refinement. To facilitate independent validation and external testing, we also deployed a publicly accessible web application using a Gradio-based interface that allows users to input clinical narratives and obtain ASA-PS classes and rationales [42].

Finally, although generated rationales appeared clinically plausible, we did not quantify their alignment with expert judgment. Future work will include prospective multicenter validation, class balancing, demographic fairness audits, real-time deployment in perioperative workflows, and blinded clinician–model rationale alignment mapped to ASA-PS criteria with interrater statistics.

Conclusions

The LLaMA-3-LoRA model provides a promising alternative to domain-specific and traditional models for ASA-PS classification. It performs well across evaluation metrics, supports explainability through attention and generated rationales, and is computationally efficient for deployment (averaged 2.25 seconds per case). This framework may serve as a basis for scalable, real-world applications in perioperative risk assessment and could be extended to other clinical classification tasks.

XSL•FO
RenderX

Acknowledgments

The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript. The authors thank Dr Cheng-Wei Lu, Director, Department of Anesthesiology, Far Eastern Memorial Hospital, for support of this study.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to institutional review board restrictions that prohibit open sharing of patient-level data to protect participant privacy but are available from the corresponding author upon reasonable request.

Funding

This study was supported by grants from Far Eastern Memorial Hospital, Taiwan (grants FEMH-2024-C-052 and FEMH-2025-C-017). The funders had no role in study design; data collection, analysis, or interpretation; manuscript writing; or the decision to submit for publication.

Authors' Contributions

Conceptualization: PC
Data curation: JHW
Formal analysis: JHW
Funding acquisition: PC
Investigation: MCC, PC
Methodology: PC
Resources: PC
Supervision: SJR
Validation: PC
Visualization: MCC
Writing – original draft: MCC
Writing – review & editing: MCC, SJR, JHW, PC

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables including American Society of Anesthesiologists Physical Status (ASA-PS) definitions, feature preprocessing rules, cohort characteristics, detailed model performance metrics, hyperparameter ablation results, and qualitative evaluation of model-generated rationales.
[XLSX File (Microsoft Excel File), 103 KB - jmir_v28i1e89540_app1.xlsx]

Multimedia Appendix 2
Input content, formatting, and length handling for each model family. All models received the same concatenated instruction + input; only formatting and length limits differed by architecture.
[DOCX File , 16 KB - jmir_v28i1e89540_app2.docx]

Multimedia Appendix 3
Pairwise McNemar test *P* values comparing classifier error rates across models. Two-sided *P* values were calculated using McNemar test with continuity correction based on paired predictions from the same test set. Predictions were converted into binary outcomes (correct vs incorrect relative to the reference American Society of Anesthesiologists Physical Status label).
[PNG File , 65 KB - jmir_v28i1e89540_app3.png]

Multimedia Appendix 4
Attention-based visualization across representative American Society of Anesthesiologists Physical Status (ASA-PS) classes II and IV. This appendix presents additional attention-based interpretability analyses for representative cases from ASA-PS classes II and IV. These figures complement the main manuscript attention analyses for ASA-PS class III (Figures 6 and 7). Attention visualizations are presented as exploratory qualitative tools to illustrate model focus patterns and do not represent causal decision mechanisms.
[PDF File (Adobe PDF File), 834 KB - jmir_v28i1e89540_app4.pdf]

XSL•FO
RenderX

Multimedia Appendix 5
Confusion matrices for all evaluated models in American Society of Anesthesiologists Physical Status (ASA-PS) classification. All evaluated classification models, including LLaMA-3-LoRA, untuned LLaMA-3, BioBERT-ft, ClinicalBERT-ft, fastText, Random Forest, Support Vector Machine (SVM), and XGBoost, were included. Each confusion matrix displays model-predicted ASA-PS classes (columns) versus reference clinician-assigned ASA-PS classes (rows), with cell values representing case counts. All confusion matrices were generated using predictions from the fixed held-out test cohort.
[PDF File (Adobe PDF File), 560 KB - jmir_v28i1e89540_app5.pdf]

References

1. ASA physical status classification system. ASA standards and practice parameters. American Society of Anesthesiologists. 2014. URL: https://www.asahq.org/standards-and-practice-parameters/statement-on-asa-physical-status-classification-system [accessed 2025-06-17]
2. Mak PHK, Campbell RCH, Irwin MG, American Society of Anesthesiologists. The ASA physical status classification: inter-observer consistency. American Society of Anesthesiologists. Anaesth Intensive Care 2002;30(5):633-640. [doi: 10.1177/0310057X0203000516] [Medline: 12413266]
3. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2018;17(01):128-144. [doi: 10.1055/s-0038-1638592]
4. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013;20(1):117-121 [FREE Full text] [doi: 10.1136/amiajnl-2012-001145] [Medline: 22955496]
5. Alsentzer E, Murphy J, Boag W, Weng W, Di J, Naumann T, et al. Publicly available clinical BERT embeddings. : Association for Computational Linguistics; 2019 Presented at: Proceedings of the 2nd Clinical Natural Language Processing Workshop; June 7, 2019; Minneapolis, Minnesota, USA p. 72-78. [doi: 10.18653/v1/w19-1909]
6. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. ArXiv. Preprint posted online on April 10, 2019 2019. [doi: 10.48550/arXiv.1904.05342]
7. Wongtangman K, Aasman B, Garg S, Witt AS, Harandi AA, Azimaraghi O, et al. Development and validation of a machine learning ASA-score to identify candidates for comprehensive preoperative screening and risk stratification. J Clin Anesth 2023;87:111103. [doi: 10.1016/j.jclinane.2023.111103] [Medline: 36898279]
8. Mudumbai SC, Pershing S, Bowe T, Kamal RN, Sears ED, Finlay AK, et al. Development and validation of a predictive model for American Society of Anesthesiologists Physical Status. BMC Health Serv Res 2019;19(1):859 [FREE Full text] [doi: 10.1186/s12913-019-4640-x] [Medline: 31752856]
9. Joachims T. Text categorization with support vector machines: learning with many relevant features. Machine Learning: ECML 1998;116:137-142. [doi: 10.1007/BFb0026683]
10. Breiman L. Random forests. Machine Learn 2001;45(1):5-32 [FREE Full text] [doi: 10.1023/A:1010933404324]
11. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. : Association for Computational Linguistics; 2017 Presented at: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers; 2017 April 3-7; Valencia, Spain p. 427-431. [doi: 10.18653/v1/e17-2068]
12. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. 2013 Presented at: NIPS'13: Proceedings of the 27th International Conference on Neural Information Processing Systems—Volume 2; December 5-10, 2013; Lake Tahoe Nevada p. 3111-3119.
13. Chung P, Fong CT, Walters AM, Yetisgen M, O'Reilly-Shah VN. Prediction of American Society of Anesthesiologists Physical Status classification from preoperative clinical text narratives using natural language processing. BMC Anesthesiol 2023;23(1):296 [FREE Full text] [doi: 10.1186/s12871-023-02248-0] [Medline: 37667258]
14. Zhang L, Fabbri D, Lasko TA, Ehrenfeld JM, Wanderer JP. A system for automated determination of perioperative patient acuity. J Med Syst 2018;42(7):123 [FREE Full text] [doi: 10.1007/s10916-018-0977-7] [Medline: 29846806]
15. Yoon SB, Lee J, Lee HC, Jung CW, Lee H. Comparison of NLP machine learning models with human physicians for ASA Physical Status classification. NPJ Digit Med 2024;7(1):259 [FREE Full text] [doi: 10.1038/s41746-024-01259-6] [Medline: 39341936]
16. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
17. Chen PF, Wang SM, Liao WC, Kuo LC, Chen KC, Lin YC, et al. Automatic ICD-10 coding and training system: deep neural network based on supervised learning. JMIR Med Inform 2021;9(8):e23230 [FREE Full text] [doi: 10.2196/23230] [Medline: 34463639]
18. Chen PF, Chen L, Lin YK, Li GH, Lai F, Lu CW, et al. Predicting postoperative mortality with deep neural networks and natural language processing: model development and validation. JMIR Med Inform 2022;10(5):e38241 [FREE Full text] [doi: 10.2196/38241] [Medline: 35536634]
19. Hu E, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. ArXiv. Preprint posted online on June 17, 2021 2021. [doi: 10.48550/arXiv.2106.09685]

XSL•FO
RenderX

20. Yang M, Chen J, Zhang Y, Liu J, Zhang J, Ma Q, et al. Low-rank adaptation for foundation models: a comprehensive review. ArXiv. Preprint posted online on December 31, 2024 2024. [doi: 10.48550/arXiv.2501.00365]
21. Bird S, Klein E, Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. In: Language Resources and Evaluation. Sebastopol, CA: O'Reilly Media, Inc; 2009.
22. Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C. Alpaca: a strong, replicable instruction-following model. Stanford Center for Research on Foundation Models Blog. URL: https://crfm.stanford.edu/2023/03/13/alpaca.html [accessed 2025-04-17]
23. Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, et al. The Llama 3 herd of models. ArXiv. Preprint posted online on July 31, 2024 2024. [doi: 10.48550/arXiv.2407.21783]
24. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A. Transformers: State-of-the-art natural language processing. : Association for Computational Linguistics; 2020 Presented at: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; 2020 November 16-20; Stroudsburg, PA p. 38-45. [doi: 10.18653/v1/2020.emnlp-demos.6]
25. Micikevicius P, Narang S, Alben J, Diamos G, Elsen E, Garcia D. Mixed precision training. ArXiv. Preprint posted online on October 10, 2017 2018. [doi: 10.48550/arXiv.1710.03740]
26. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. : Association for Computing Machinery; 2016 Presented at: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, California, USA p. 785-794. [doi: 10.1145/2939672.2939785]
27. Stoica P, Babu P. Pearson–Matthews correlation coefficients for binary and multinary classification. Signal Process 2024;222:109511. [doi: 10.1016/j.sigpro.2024.109511]
28. Riyanto S, Sitanggang IS, Djatna T, Atikah TD. Comparative analysis using various performance metrics in imbalanced data for multi-class text classification. Int J Adv Comput Sci Appl 2023;14(6). [doi: 10.14569/ijacsa.2023.01406116]
29. Vig J. A multiscale visualization of attention in the transformer model. : Association for Computational Linguistics; 2019 Presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations; July 28-August 2, 2019; Florence, Italy p. 37-42. [doi: 10.18653/v1/p19-3007]
30. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. : Curran Associates Inc; 2022 Presented at: NIPS '22: Proceedings of the 36th International Conference on Neural Information Processing Systems; November 28-December 9, 2022; New Orleans LA USA p. 24824-24837. [doi: 10.52202/068431-1800]
31. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 2008;61(Pt 1):29-48. [doi: 10.1348/000711006X126600] [Medline: 18482474]
32. Wang S, Chen L, Jiang J, Xue B, Kong L, Wu C. LoRA meets dropout under a unified framework. : Association for Computational Linguistics; 2024 Presented at: Findings of the Association for Computational Linguistics: ACL 2024; August 11-16, 2024; Bangkok, Thailand p. 1995-2008. [doi: 10.18653/v1/2024.findings-acl.119]
33. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: efficient finetuning of quantized LLMs. : Neural Information Processing Systems Foundation, Inc; 2023 Presented at: NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems; December 10-16, 2023; New Orleans LA USA p. 10088-10115. [doi: 10.52202/075280-0441]
34. Jain S, Wallace BC. Attention is not explanation. ArXiv. Preprint posted online on February 26, 2019 2019. [doi: 10.48550/arXiv.1902.10186]
35. Brown KE, Yan C, Li Z, Zhang X, Collins BX, Chen Y, et al. Large language models are less effective at clinical prediction tasks than locally trained machine learning models. J Am Med Inform Assoc 2025;32(5):811-822. [doi: 10.1093/jamia/ocaf038] [Medline: 40056436]
36. Tollinche LE, Yang G, Tan K, Borchardt R. Interrater variability in ASA Physical Status assignment: an analysis in the pediatric cancer setting. J Anesth 2018;32(2):211-218 [FREE Full text] [doi: 10.1007/s00540-018-2463-2] [Medline: 29441411]
37. Chung P, Fong CT, Walters AM, Aghaeepour N, Yetisgen M, O'Reilly-Shah VN. Large language model capabilities in perioperative risk prediction and prognostication. JAMA Surg 2024;159(8):928-937. [doi: 10.1001/jamasurg.2024.1621] [Medline: 38837145]
38. Turpin M, Michael J, Perez E, Bowman S. Language models do not always say what they think: unfaithful explanations in chain-of-thought prompting. : Neural Information Processing Systems Foundation, Inc; 2023 Presented at: NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems; December 10-16, 2023; New Orleans, LA, USA p. 74952-74965. [doi: 10.52202/075280-3275]
39. Huang R, Wu H, Yuan Y, Xu Y, Qian H, Zhang C, et al. Evaluation and bias analysis of large language models in generating synthetic electronic health records: comparative study. J Med Internet Res 2025;27:e65317 [FREE Full text] [doi: 10.2196/65317] [Medline: 40354109]
40. Smolyak D, Bjarnadóttir MV, Crowley K, Agarwal R. Large language models and synthetic health data: progress and prospects. JAMIA Open 2024;7(4):ooae114. [doi: 10.1093/jamiaopen/ooae114] [Medline: 39464796]

XSL•FO
RenderX

41. Chen YH, Ruan SJ, Chen PF. Predicting 30-Day postoperative mortality and American Society of Anesthesiologists Physical Status using retrieval-augmented large language models: development and validation study. J Med Internet Res 2025;27:e75052 [FREE Full text] [doi: 10.2196/75052] [Medline: 40460423]
42. Chen P. Hugging Face spaces. ASA_LoRA_demo. URL: https://huggingface.co/spaces/andy698087/ASA_LoRA_demo [accessed 2025-12-08]

Abbreviations

AC1: agreement coefficient
ASA-PS: American Society of Anesthesiologists Physical Status
AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
BERT: Bidirectional Encoder Representations from Transformers
EHRs: electronic health record
LLaMA: large language model meta-AI
LLM: large language model
LoRA: low-rank adaptation
MCC: Matthews correlation coefficient
ML: machine learning
NLP: natural language processing
SVM: support vector machine
XGBoost: Extreme Gradient Boosting

Edited by A Coristine; submitted 14.Dec.2025; peer-reviewed by DS Chang, E Bignami, M Burns; comments to author 12.Jan.2026; revised version received 24.Mar.2026; accepted 24.Mar.2026; published 21.Apr.2026.

Please cite as:
Chen MC, Ruan SJ, Wu JH, Chen PF
Classifying American Society of Anesthesiologists Physical Status With a Low-Rank–Adapted Large Language Model: Development and Validation Study
J Med Internet Res 2026;28:e89540
URL: https://www.jmir.org/2026/1/e89540
doi:10.2196/89540
PMID:

©Min-Chia Chen, Shanq-Jang Ruan, Jo-Hsin Wu, Pei-fu Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Uptake of Clinical Decision Support Systems Among Health Care Professionals in Six European Countries and the United States: Cross-Sectional Survey Within the I-CARE4OLD Project

Collin JC Exmann[1,2], MSc, MD; Anna-Maria Hiltunen[3], MSc; Ira Haavisto[3], PhD; Anna Salminen[3], MSc; Maikki Messo[3], MSc; Mikko Nuutinen[3,4], PhD; Mark Hoogendoorn[5], Prof Dr; Wiebe Boorsma[1], MD, PhD; Elizabeth P Howard[6,7], PhD; Vanja Pešić[8], MSc; Mor Alon[9]; Federica Mammarella[10,11], MD, MSc; Rosa Liperoti[10,11], MD, PhD; Olena Švihnosová[12], MSc; Daniela Fialová[12], Prof Dr; Natalia Drapała[13], MSc, MPH; Katarzyna Szczerbińska[13,14], Prof Dr, MD; Anja Declercq[15,16], Prof Dr; Hein PJ van Hout[1], Prof Dr; Johanna De Almeida Mello[15,17], PhD

[1]Department of General Practice, Amsterdam UMC, Amsterdam, The Netherlands
[2]Aging and Later Life Research Program, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands
[3]Nordic Healthcare group, Helsinki, Finland
[4]Department of Social and Clinical Pharmacy, Faculty of Pharmacy, Charles University, Prague, Czech Republic
[5]Deparment of computer science, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[6]Hebrew senior life, The Hinda and Arthur Marcus Institute for Aging Research, Boston, MA, United States
[7]Boston College, Connell School of Nursing, Boston, MA, United States
[8]Boston College School of Social Work, Chestnut Hill, MA, United States
[9]Checker Software Solutions, Profility Inc, Haifa, Israel
[10]Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy
[11]Università Cattolica del Sacro Cuore, Milan, Italy
[12]Department of Internal Medicine and Geriatrics, 1st Faculty of medicine, Charles University, Prague, Czech Republic
[13]Laboratory for Research on Aging Society, Chair of Epidemiology and Preventive Medicine, Medical Faculty, Jagiellonian University Medical College, Kraków, Poland
[14]Department of Internal Medicine and Geriatrics, Jagiellonian University, Kraków, Poland
[15]LUCAS, Center for Care Research and Consultancy, KU Leuven, Leuven, Belgium
[16]Center for Sociological Research, KU Leuven, Leuven, Belgium
[17]Department of Oral Health Sciences, KU Leuven, Leuven, The Netherlands

Corresponding Author:
Collin JC Exmann, MSc, MD
Department of General Practice
Amsterdam UMC
Meibergdreef 15
Amsterdam, 1105AZ
The Netherlands
Phone: 31 020 566 9111
Email: c.exmann@amsterdamumc.nl

Abstract

Background: The use of Clinical Decision Support Systems (CDSS), such as clinical decision rules, algorithms, or machine learning-based applications, has gained attention in recent years. However, their adoption and effectiveness may vary across different health care systems and settings. For a CDSS to be adopted, it must effectively address the practical issues encountered by professionals; however, little research has been done to identify these needs and requirements.

Objective: This study aims to describe and compare the current use of various decision-support and prediction tools in long-term care for older people across health professionals from 6 European countries and the United States.

Methods: This study analyzed survey data from a CDSS pilot study in a purposive sample of health professionals working with older adults with complex chronic conditions from six European countries and the United States. The survey included participants'

XSL•FO
RenderX

general background information, their current use of decision support tools, and their attitudes on the potential benefits of CDSS. About 20 participants were sampled per country. Closed responses were analyzed using correlation coefficients and regression models, while open-ended responses were clustered in a qualitative manner, categorizing each response.

Results: A total of 151 professionals (mean age 45.5, SD 11.6 years, 71.5%, 108/151 female) participated in the pilot study. Most participants were physicians (85/151, 56.3%) or nurses (57/151, 37.7%). About 51% (78/151) of the participants reported using CDSS, while 22.4% (34/151) used predictive CDSS, showing important variation across samples from the seven countries. The regression model for comfort with technology showed a positive association for openness to new technologies (β=0.622; *P*<.001), although an inverse significant association was found for age (β=–0.022; *P*<.001). No significant associations were found for the actual use of CDSS. Participants reported using CDSS mainly for diagnostic purposes or for guideline implementation, not aimed at prognostic information. In contrast, examples of prognostic tools were most frequently mentioned by respondents as being valuable improvements to clinical practice.

Conclusions: While some countries' samples reported well-integrated digital health infrastructures and higher CDSS adoption rates, others still face challenges in implementing these. However, we found multiple examples of emerging tools, and at the same time, an important demand for predictive CDSS. Our findings highlight the need for improvement of current CDSS implementation and both development and implementation of particularly predictive CDSS.

(J Med Internet Res 2026;28:e85071) doi:10.2196/85071

KEYWORDS

machine learning; decision support systems; innovation; home care; prediction models; long term care; older adults

Introduction

Clinical Decision Support Systems (CDSS) are integrated software applications designed to support healthcare professionals in making clinical decisions [1,2]. Using patient data and analytical techniques, these systems generate alerts, offer diagnostic support, and deliver evidence-based recommendations intended to improve clinical outcomes. These tools are designed to inform clinicians' decision-making about individual patients [3]. In the context of caring for older adults, particularly those with complex chronic conditions (CCC), the usefulness of CDSS becomes even more apparent. This population, marked by multiple interacting conditions and complex care needs, presents clinicians with significant challenges in balancing the benefits and risks associated with interventions [4]. This challenge is especially evident in community-based settings and long-term care facilities, where frontline care providers may have limited access to specialists or advanced diagnostic resources [5,6]. In such environments, smart decision support systems can facilitate complex evaluations. However, the advantages of CDSS use should outweigh the barriers experienced by professionals [7].

Existing CDSS differ in their underlying methods, the extent to which they provide general or patient-specific advice, and the principles on which they are based. Previous research has classified these systems as either knowledge-based or non–knowledge-based systems [2,8]. Knowledge-based systems include decision rules derived from expert knowledge and clinical consensus [8]. Examples of knowledge-based systems are clinical guidelines with decision trees or treatment protocols. They may or may not be digitally integrated into electronic health records (EHRs) for automatic evaluation or be used as benchmarking. Non–knowledge-based systems include clinical algorithms, as well as machine learning (ML) models which are trained to evaluate certain outcomes or risks based on a given set of predictors. Like knowledge-based systems, these systems provide clinicians with decision-supporting information, but in a data-driven manner rather than based on clinical reasoning. An example of such a system is the CHA2DS2-VASc score, which uses a simple scoring algorithm to calculate the risk of thromboembolic cerebrovascular events in patients with atrial fibrillation [9]. Beyond knowledge-based, these systems can be seen as having a more diagnostic (outcome in the present), or predictive aim (outcome in the future) [10].

The relatively recent implementation of EHRs has created an important potential for CDSS implementation [11,12]. In particular, the integration of such systems within the EHR could automatically generate recommendations in real-time and leverage state-of-the-art ML techniques for application in health care. Indeed, the use of non–knowledge-based CDSS, such as clinical algorithms or artificial intelligence (AI) applications, has gained support in recent years [13]. However, critiques have also emerged, highlighting gaps in how to appropriately interpret the provided results and questioning the extent to which the information can be fully trusted [14,15].

Against this background, the EU-Horizon 2020-funded project "Individualized CARE for OLDer persons with CCC at home and in nursing homes (I-CARE4OLD)" developed a clinical decision support tool – the iCARE tool. This tool aims to support clinicians in making decisions about care paths and treatments for home care and residential care settings [16]. The tool generates individualized prognostications derived from ML on the risk of adverse outcomes, changes in functional health, and impact of interventions. It is based on internationally standardized routine care data from the interRAI assessment instruments. Millions of interRAI assessment records from the participating countries were used, linked to registry data on mortality, hospitalization, medications, and health care use [16,17]. The prototype of the iCARE tool was tested during a pilot consisting of multiple phases. This study focuses on the reported current and beneficial use of various CDSS by participating health care professionals caring for older adults with CCC, as collected in the pre-pilot phase of the pilot [18]. The aim of this study is to describe and compare the use of

XSL•FO
RenderX

clinical decision support tools and attitudes toward future CDSS development among health care professionals caring for older adults with CCC.

Methods

Overview

This research is part of the larger study iCARE4OLD funded by the European Union in the Horizon 2020 framework [16]. The study used a cross-sectional design using mixed closed and open-ended survey questions to explore the use and perception of CDSS among health care professionals. The protocol of the study was published recently [18]. Data collection took place simultaneously in seven countries: Belgium, the Czech Republic, Finland, Italy, the Netherlands, Poland, and the United States. The study was ethically approved in the participating countries and all health care professionals gave their written informed consent for participation. The protocol describes the different phases of the study: the pre-questionnaire phase, the patient evaluation phase in the iCARE tool, and the post-questionnaire phase. In this study, data from the prequestionnaire phase were used for the analysis.

Participants and Eligibility

Participants were healthcare professionals involved in the care of older adults (aged 65 years or older) with CCC, either in long-term care facilities or primary care. Eligible participants were required to be qualified professionals capable of making or proposing therapeutic decisions (eg, physicians, nurses, and physiotherapists) and interpreting interRAI assessments. We used purposive sampling combined with snowball sampling of at least 20 participants per country to ensure sufficient observations for analysis. This sampling was performed by each country's local study lead among purposively selected professionals. To ensure consistency across countries, the study leads held regular meetings reporting the recruitment and data collection procedures.

Local coordinators followed a standardized, documented procedure: site liaison and approval; targeted invitations with a translated information sheet; brief eligibility prescreening against the four inclusion criteria; scheduling of sessions; written informed consent; and administration of a uniform testing protocol (prequestionnaire, iCARE testing session, decision-loop where applicable, postquestionnaires).

Data Collection

Data collection started in November 2024 and ended in January 2025. Clinicians completed an online questionnaire created for the study, consisting of background questions (socio-demographic information, as well as information about education, profession or position, years of professional experience, etc), questions related to the use of interRAI instruments, attitudes toward new technology, and both closed and open-ended questions on the use of CDSS. The English version of the questionnaire can be found in Multimedia Appendix 1. Local study leads and their teams translated the surveys into the local languages**.** Using both closed and open-ended questions can result in triangulation, making the findings more valid and permitting a deeper understanding of the data [19,20].

Data Validation and Cleaning

The data were validated to ensure data integrity. Validation procedures included verifying that each respondent ID had a single set of responses across all corresponding questionnaires. Timestamp checks were conducted to confirm consistency with respondents' records. The dataset was systematically reviewed to identify and correct any erroneous values.

Statistical Analysis

Quantitative data were analyzed using descriptive statistical methods, bivariate analysis, and regression models to explore associations. Pearson correlations were used for continuous data and Spearman correlations for Likert scales. A linear regression model was built for being comfortable with technology and a logistic model for the actual use of CDSS. Age was used as a continuous variable in the linear model and as dichotomous variable (age<50 or age≥50 years) in the logistic model. Only two professions were used in the models to allow for comparison (nurse and physician). The software STATA SE (version 18.5; StataCorp LLC) was used in the analysis.

All qualitative open-ended responses were first translated into English by the local teams, and then all answers were thematically analyzed by an experienced researcher and independently reviewed by another researcher to ensure consistency. Any discrepancies were discussed collaboratively until consensus was reached. In case of ambiguity or unclear answers, the researchers discussed the interpretation to reach a consensus on classification. If no reliable interpretation could be made, the response was noted as unclear and excluded from thematic coding.

Response Clustering

We systematically clustered responses from the 3 open-ended questions. These questions were (1) use of decision support systems, (2) use of predictive decision support systems, and (3) decision support tools participants considered helpful. We classified the reported CDSS as either "knowledge based" or "non-knowledge based," as "diagnostic" or "prognostic," and lastly into consensus-based categories based on the thematic analysis [2,8]. For the classification into either diagnostic or prognostic, we followed previously proposed definitions [10]. Diagnostic refers to CDSS which classify the current health status or event, whereas in a predictive CDSS, the outcome is a future event. Where possible, we searched for additional information on the reported CDSS systems to complement available information. This way, the CDSS could be classified more reliably.

Ethical Considerations

Ethical approvals were granted in each participating country before the start of the pilot. Italy: Comitato Etico Territoriale Lazio 3 (ID7068); the Netherlands: Amsterdam UMC Toetsingscommissie (2024.0527); Belgium: Commissie Medisch Ethiek van de Universitair Medische Ziekenhuizen KU Leuven (No. S69591); Czechia: Ethics Committee of the Faculty of Pharmacy, Univerzita Karlova, H2020-965341,); Poland: the

XSL•FO
RenderX

Research Ethics Committee at the Jagiellonian University Medical College (118.0043.1.237.2024); the United States: Advarra Institutional Review Board (No. 965341); Finland: Well-being Services County of North Karelia (6970/13.00.01/2024, § 5/2024 and Kainuu Metti Järvikallio (6970/13.00.01/2024, § 23/2024); Central Uusimaa (KEUHDno-2024-3384, § 87); and Helsinki (HEL 2024-010604 *t* 13 02 01). All participants gave their written informed consent to participate in the study, and data were anonymized and further pseudomized to link assessments. Some countries offered financial compensation to participants as described in the protocol of the study [14].

Results

Overview

Table 1 shows the characteristics of the 151 participants from seven participating countries. Clinicians' mean age varied from 38.4 (SD 9.7) years in the Dutch sample to 51.6 (SD 10.3) years in the sample from the United States. Participants from Finland and the United States were predominantly female (19/20, 95%), while the proportion of women in the overall sample was 71.5% (108/151). The proportion of participants from home care or long-term care services was similar across 6 countries (47-53%), with the exception of the Czech sample, with 28% (7/25) professionals from home care and 72% (18/25) from long-term care.

Amongst participants, physicians were the most prevalent (85/151, 56.3%), followed by nurses (57/151, 37.7%). Most respondents had a master's educational level (99/151, 65.6%), followed by 15.2% (23/151) with a bachelor's degree and 13.9% (21/151) with a PhD. Clinical experience was most extensive in participants from the United States (24.7 years) and Belgium (21.9 years), and relatively lower in those from the Netherlands (12 years). Similar results were found for the experience with complex care for older people, with an average of 23.1 (SD 14.9) years in the sample from the United States and 10.8 (SD 7.8) years in the Dutch sample.

Regarding users' comfort with technology, the average score was 3.9 out of 5, with the score ranging from 1 (not comfortable at all) to 5 (very comfortable). Participants from Italy (4.3) and the Czech Republic (4.4) reported higher levels of comfort, while participants from the Netherlands reported the lowest (3.3). The average overall openness to integrating new technologies was high (4.2/5), with the highest levels for professionals from the United States (4.7), Italy (4.6), and Poland (4.5).

Table 2 shows the correlations between the most important determinants of CDSS use. A positive and significant correlation was found between being open to adopting new technologies and being comfortable with using technology in the overall sample (0.579; $P<.001$). This was also found for health care professionals from Belgium, the Czech Republic, Finland, and Italy, with very strong correlations in the samples from the Czech Republic (0.933; $P<.001$) and Finland (0.943; $P<.001$). All other correlations were not found to be significant in the full sample. In addition, for the participants from Belgium and the Czech Republic, there was a positive association between being open to technology and the use of CDSS. In the United States' sample, the association between the number of years of clinical experience and the use of predictive CDSS was negative and significant (–0.464; $P=.04$).

XSL•FO
RenderX

Table 1. Characteristics of participants and their use of the Clinical Decision Support System.

Characteristics	All	Belgium	Czech Republic	Finland	Italy	the Netherlands	Poland	United States
Participants, n	151	20	25	20	20	26	20	20
Age (years), mean (SD)	45.5 (11.6)	48.7 (10.7)	43.0 (15.6)	45.0 (9.0)	48.5 (6.9)	38.4 (9.7)	46.5 (11.7)	51.6 (10.3)
Gender female, n (%)	108 (71.5)	15 (75)	13 (5)	19 (95)	9 (45)	16 (61.5)	17 (85)	19 (95)
Home care, n (%)	71 (47)	10 (50)	7 (28)	11 (55)	10 (50)	13 (50)	10 (50)	10 (50)
Long-term care, n (%)	80 (53)	10 (50)	18 (72)	9 (45)	10 (50)	13 (50)	10 (50)	10 (50)
Position, n (%)								
Nurse	57 (37.7)	12 (60)	0 (0)	15 (75)	0 (0)	7 (26.9)	7 (35)	16 (80)
Physician	85 (56.3)	5 (25)	25 (100)	2 (10)	20 (100)	19 (73.1)	12 (60)	2 (10)
Other	9 (6)	3 (15)	0 (0)	3 (15)	0 (0)	0 (0)	1 (5)	2 (10)
Degree, n (%)								
Basic	8 (5.3)	0 (0)	0 (0)	4 (20)	0 (0)	0 (0)	0 (0)	4 (20)
Bachelor's	23 (15.2)	2 (10)	0 (0)	13 (65)	0 (0)	3 (11.5)	0 (0)	5 (25)
Master's	99 (65.6)	13 (65)	25 (100)	3 (15)	19 (95)	19 (73.1)	13 (65)	7 (35)
Doctor	21 (13.9)	5 (25)	0 (0)	0 (0)	1 (5)	4 (15.4)	7 (35)	4 (20)
Years in clinical practice, mean (SD)	18.2 (12)	21.9 (12.5)	19.3 (17)	15.7 (8.4)	18.2 (6.3)	12.0 (8.4)	17.3 (11.4)	24.7 (13)
Years of experience with geriatric complex care, mean (SD)	15.9 (10.9)	19.4 (11.5)	10.7 (9.7)	14.2 (9)	18.3 (6.6)	10.8 (7.8)	17.4 (10.8)	23.1 (14.9)
Working experience with interRAI, n (%)	72 (47.7)	14 (70)	1 (4)	20 (100)	20 (100)	6 (23.1)	2 (10)	9 (45)
Years with use of interRAI tools, mean (SD)	4.8 (7)	4.6 (6)	0.8 (1)	8 (6.2)	15.3 (7.8)	1.2 (3.2)	2.5 (5.4)	3.5 (6.2)
Comfortable with using technology[a], mean (SD)	3.9 (1.1)	4.1 (0.8)	4.4 (1.2)	3.9 (1.3)	4.3 (0.9)	3.3 (1)	3.9 (0.8)	3.8 (1.5)
Open to adopting new technologies[b], mean (SD)	4.2 (1)	4.1 (0.7)	4.2 (1.3)	4 (1.2)	4.6 (0.6)	3.5 (1)	4.5 (0.7)	4.7 (0.5)
Have used CDSS in work, n (%)	78 (51.7)	7 (35)	17 (68)	18 (90)	3 (15)	15 (57.7)	9 (45)	9 (45)
Have used predictions, n (%)	34 (22.5)	2 (10)	6 (24)	3 (15)	4 (20)	12 (46.2)	1 (5)	6 (30)

[a]1 (Not comfortable at all) to 5 (Very comfortable).

[b]1 (Not open at all) to 5 (Very open).

XSL•FO
RenderX

Table 2. Cross-sectional bivariate analysis for the use of Clinical Decision Support System (CDSS) among health care professionals overall and in the seven individual countries.

Variables	All (N=141)	Belgium (n=20)	Czech Republic (n=25)	Finland (n=20)	Italy (n=20)	the Netherlands (n=26)	Poland (n=20)	United States (n=20)
Years of clinical experience x use of CDSS	–0.049	0.633[a]	–0.218	–0.237	–0.286	0.025	–0.115	0.025
Years of clinical experience x use of predictions	–0.045	0.385	–0.084	0.119	0.085	0.249	–0.006	–0.464[b]
Open to adopting new technologies x comfortable with using technology	0.579[c]	0.613[a]	0.933[c]	0.943[c]	0.621[a]	0.257	0.291	0.018
Open to adopting new technologies x the use of CDSS	0.094	0.494[b]	0.399[b]	0.423	0.288	–0.160	0.225	0.032
Open to adopting new technologies x the use of predictions	–0.096	0.191	–0.032	0.118	–0.086	–0.265	0.171	0.023
Comfortable with using technology x the use of CDSS	0.137	0.345	0.394	0.359	0.380	0.187	0.118	–0.187
Comfortable with using technology x the use of predictions	0.049	0.186	–0.033	0.032	0	0.102	0.030	0.406

[a]$P<.01$.

[b]$P<.05$.

[c]$P<.001$.

Use of CDSS

A total of 78 (51.3%) participants reported using any type of CDSS in their daily work, while 34 (22.4%) reported using specific predictive CDSS. The rates for CDSS use in daily work varied across countries' samples ranging from 15% (3/20) in Italy, 35% (7/20) in Belgium, to the highest rate of 90% (18/20) in Finland. Regarding the use of CDSS for predictions, the lowest rates were found in the Polish (1/20, 5%) and Belgian (2/20, 10%) samples, while the highest rate was found in the Netherlands (12/26, 46.2%).

A total of 132 open-text answers were provided about the use of CDSS, and 49 answers to the question on predictive CDSS, of which 105 (79.5%) and 38 (77.6%) were analyzed, respectively. For the general CDSS question, all the exclusions stemmed from the answer being too unclear and therefore not being classifiable. Such answers, for instance, were not specific enough (eg, "I use multiple algorithms") or did not clearly mention a CDSS (eg, "thromboembolism", "standard procedures"). For the predictive CDSS question, answers that concerned nonpredictive CDSS were also excluded (eg, "CORNELL", a diagnostic instrument).

As can be seen in Figure 1, multiple dominant categories of CDSS were mentioned, such as diagnostic instruments (24/105, 23%), prediction tools (20/105, 19%), and tools considering intervention advice (17/105, 16%). Examples of these categories include the mini mental state examination or WELLS criteria (diagnostic instruments), CDSS on care pathways or future bleeding risks (prediction tools), and interaction checkers and antibiotic prescription systems (intervention advice). Another common category mentioned by professionals was geriatric assessment tools (16/105, 15%), most of which were considered interRAI instruments. Finally, some professionals mentioned commercial software solutions (N=2) providing one or more CDSS categories, but these responses lacked sufficient detail to identify specific categories or methods and were therefore grouped into a separate cluster.

XSL•FO
RenderX

Figure 1. Use of any CDSS (question 1) among all health care professionals per country. CDSS: Clinical Decision Support System.

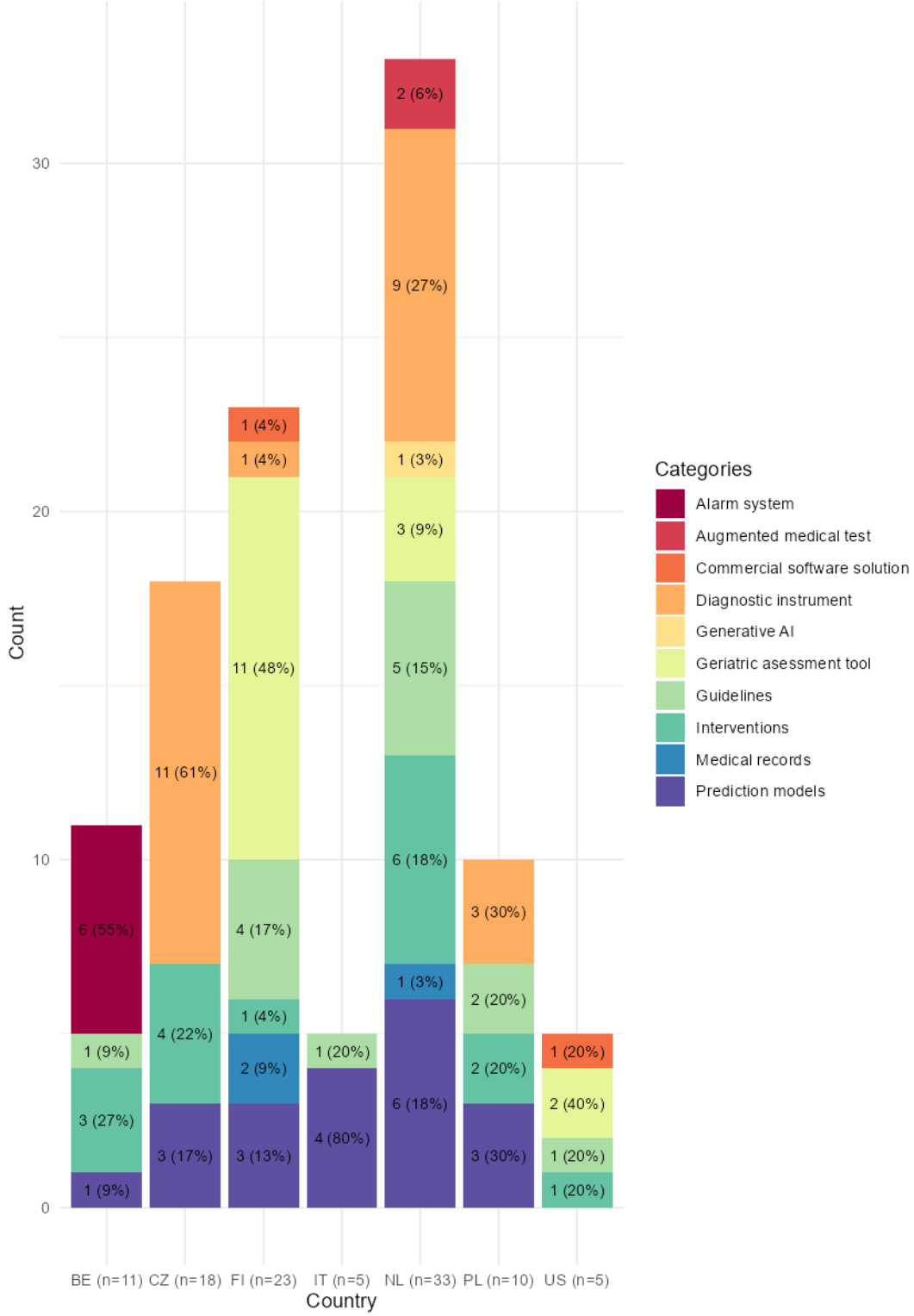

When looking at the types of CDSS, most tools were knowledge-based (71/105, 68%) and aimed for diagnostic (69/105, 65%) purposes, thereby making predictive or prognostic tools much less common in current practice, only mentioned in about 19% (20/105) of the answers to the general CDSS question.

While analyzing the predictive CDSS mentioned, 24 (49%) answers had to be excluded, as they did not concern any predictive tool. Out of the remaining 26 valid predictive CDSS answers, most concerned either intervention advice (eg, CHA2DS2-VASc or CVRM treatment advice) or predictive tools such as (cardiovascular) mortality risk calculators. Some of the overall CDSS categories were excluded, as they are not predictive by definition. Note that in Figure 1, “prediction models” includes all CDSS providing direct outcome predictions based on responses to the general CDSS question.

XSL•FO
RenderX

Regression Analysis for Comfortable Use and Current CDSS Use

Tables 3 and 4 show the results of the regression models built to explain being comfortable with using technology, and for the use of CDSS. In the first model, age and openness to new technologies were significant determinants of feeling comfortable with technology. Age showed an inverse association (β=–0.022; *P*<.001), while people who are more open to technology showed a significant positive association (0.622; *P*<.001). In the second model, no determinants were significantly associated with the actual use of CDSS.

Table 3. Results of linear regression for feeling comfortable using technology (N=141).

Characteristics	β coefficient (95% CI)	*P* value
Age	–0.022 (–0.035 to –0.009)	<.001[a]
Gender (ref: men)	–0.231 (–0.588 to 0.127)	.20
Profession (ref: nurse)	0.246 (–0.0827 to 0.575)	.14
Open to new technologies	0.622 (0.473-0.771)	<.001[a]

[a]*P*<.001. R^2=0.427

Table 4. Results of the logistic regression model for the use of Clinical Decision Support System (N=141).

Characteristics	Odds ratios (95% CI)	*P* value
Age (ref: age<50)	0.766 (0.374-1.567)	.47
Gender (ref: men)	0.532 (0.232-1.222)	.14
Profession (ref: nurse)	1.888 (0.559-2.523)	.66
Open to new technologies	1.173 (0.447-3.077)	.75
Comfortable with technology	0.586 (0.255-1.343)	.21

Future Beneficial CDSS

A total of 136 (90.1%) participants answered the question on what they perceive as future beneficial CDSS s in their clinical work (“What kind of clinical decision support tool would be beneficial for you in patient care?”), out of which we retrieved 179 valid answers, as some answers contained multiple technologies. These answers were qualitatively clustered in seven categories in total with individual subcodes as subcategories, as can be found in Table 5.

Frequently mentioned were systems that would either estimate risks and outcomes for patients, such as providing information on risk of mortality, hospitalization, or falls. Next, many professionals also mentioned that support systems giving advice on certain interventions (either pharmacological or nonpharmacological) could also be of help in their practice. Often, professionals would describe the need for personalized information on, for instance, risks for side effects or the effectiveness of the chosen intervention. These 2 categories, although depending on the exact use case, concern tools that would provide clinicians with prognostic information on treatment advice and important health outcomes. Especially, doctors seemed to be interested in such CDSS. Nurses were interested in these too, but more often reported future beneficial technology from the “other” category. We saw only a tendency toward more reporting of recommendation systems in the home care sample.

An important part of the answers also included the need for either electronic record adoption or improvements within an electronic record already in use. New functionalities such as automated note writing based on the recorded consultation or an integrated sharing between professionals were often mentioned. Finally, although generative AI was only mentioned once in the current CDSS use, multiple respondents answered that they would benefit from a clinical reasoning tool that would come up with differential diagnosis or with which they can discuss complex patient cases.

XSL•FO
RenderX

Table 5. Categories of future beneficial Clinical Decision Support System as reported by professionals in the study (iCARE4OLD pilot, cross-sectional study, November 2024-January 2025).

Category, N (% within column)	Total	Nurse	Doctor	Other	HC	LTCF
Care pathways and needs	8 (4)	4 (6)	2 (2)	2 (18)	5 (5)	3 (3)
Care needs identification	4	__[a]	—	—	—	—
Care pathways software	2	—	—	—	—	—
Palliative care needs identification	1	—	—	—	—	—
Psychological care needs	1	—	—	—	—	—
Clinical reasoning software	8 (4)	2 (3)	6 (6)	0 (0)	6 (6)	2 (2)
Electronic record improvements	19 (11)	11 (16)	8 (8)	0 (0)	10 (11)	9 (10)
Automated record sharing	4	—	—	—	—	—
Automated record writing	7	—	—	—	—	—
Automated referral system	1	—	—	—	—	—
Electronic medical record summaries	4	—	—	—	—	—
Electronic medical record system	3	—	—	—	—	—
Other	19 (11)	12 (18)	5 (5)	2 (18)	8 (9)	11 (13)
interRAI	10	—	—	—	—	—
Alarm system	4	—	—	—	—	—
Other (single observations)	5	—	—	—	—	—
Protocol and guidelines	6 (3)	4 (6)	2 (2)	0 (0)	0 (0)	6 (7)
Communication protocol	1	—	—	—	—	—
Guideline improvements	4	—	—	—	—	—
Triage support	1	—	—	—	—	—
Recommendation systems	67 (37)	19 (28)	46 (46)	2 (18)	41 (44)	26 (30)
Diagnostic recommendation system	5	—	—	—	—	—
Interaction checker	15	—	—	—	—	—
Intervention recommendation system	15	—	—	—	—	—
Medication recommendation system	32	—	—	—	—	—
Remote care and monitoring	6 (3)	6 (9)	0 (0)	0 (0)	2 (2)	4 (5)
Risk and outcome predictions	46 (26)	10 (15)	31 (31)	5 (45)	21 (23)	25 (29)
Cancer	1	—	—	—	—	—
Delirium	3	—	—	—	—	—
Falls	5	—	—	—	—	—
Frailty	4	—	—	—	—	—
Health instability	3	—	—	—	—	—
Hospitalization	2	—	—	—	—	—
Mobility	1	—	—	—	—	—
Mortality and life expectancy	5	—	—	—	—	—
Pain	2	—	—	—	—	—
Physical restraints	1	—	—	—	—	—
Pressure ulcer	2	—	—	—	—	—
Rehabilitation outcomes	1	—	—	—	—	—
Thromboembolism	1	—	—	—	—	—
Unspecified	15	—	—	—	—	—

XSL•FO
RenderX

[a]Not applicable.

Discussion

Principal Findings

This study provided an overview of the current use of CDSS by health professionals working with older persons with CCC from six European countries and the United States. The study highlighted notable variations in use rates and professionals' needs for future developments in CDSS. The results also revealed differences in CDSS adoption. For example, adoption rates were as low as 15% (3/20) in the Italian sample, contrasted by remarkably high use at 90% (18/20) in the Finnish sample. These disparities are possibly influenced by varying degrees of investment in digital infrastructure, policy support mechanisms, and training, along with cultural attitudes to adopting technology in clinical practice [21,22]. In our study, most participants reported being open to new technology and comfortable using new technology, although the overall uptake of CDSS was 51.7% (78/151), and of predictive CDSS only 22.5% (34/151).

Comparison With Previous Research

These results are similar to previous research indicating that professionals will adopt CDSS if they perceive benefits and are supported to implement the tools [23,24]. As a general principle, an effective CDSS must minimize the effort required by clinicians to receive and act on system recommendations [1]. However, as both a limitation of this study and a topic for reflection, we found a considerable proportion of unsatisfactory responses to questions regarding CDSS, particularly those addressing their predictive functionalities. This finding may indicate a lack of understanding among participants concerning the nature and capabilities of predictive CDSS. Prior studies have similarly identified a gap in healthcare professionals' literacy on such topics, emphasizing the importance of addressing this issue to facilitate the effective implementation of CDSS in clinical practice [25-27].

Moreover, this study did not evaluate the performance of individual CDSS, which can be a topic of further research. Notably, certain reported CDSS (eg, generative AI) are not (yet) recommended for clinical practice. When combined with limited literacy among clinicians, the adoption of such tools could introduce risks to patient safety and care quality.

The findings from the study showed cross-national variations for comfort and openness to technology, and variation in the actual use of CDSS and predictions. Our findings showed that, although clinicians were comfortable or open to technology, this seldom correlated with the reported use, suggesting that although clinicians are keen to adopt such tools, there may remain infrastructural or systemic obstacles to real adoption. The results from the regression model for CDSS use are congruent with this. Previous research has also shown that differences in micro and macro-level factors across countries, for example, individual professionals' views, costs, or national legal issues, can influence the adoption [28].

In addition, the results indicated that there is an important need among health care providers for improvements in CDSS. As most of our responses concerned either personalized risk assessments or general diagnostic or intervention recommendations, there is a practice-driven demand for further research improving healthcare prognostications on a personalized level [28]. This could, for instance, include personalized treatment advice [29]. This pilot was part of a project aiming at advancing the field on this subject, but as our data shows, there are many more outcomes or interventions to be studied.

Interestingly, consistent with previous research, we found a significant number of existing CDSS tools, even some of which could fill the needs mentioned as future beneficial CDSS, it seems CDSS are only used to a limited extent [30]. It has indeed been previously described that CDSS are often developed, but they often encounter challenges in achieving sustained implementation and long-term integration into practice [31-33]. This gap between development and real-world adoption highlights critical challenges in translating innovation into practice, such as integration with existing health IT systems, clinician training, workflow adaptation, and organizational readiness [34].

Strengths and Limitations of This Study

We aimed to include participants with diverse paramedical backgrounds and from different countries, providing an overview across multiple countries and career stages. As only about 20 participants were included per country, the data cannot be regarded as representing the countries as a whole. Furthermore, the participant recruitment was, as described in the pilot protocol, based on targeted purposive samples and might therefore be biased toward research and innovation-oriented professionals and professionals with interRAI experience. However, it can be argued that for parts of the results, such as the future beneficial CDSS, we still reached a representative overview of what professionals could benefit from. Finally, as we worked with open-ended questions, the response categorization could still be influenced by subjective interpretations despite working with coding procedures. While we tried to retrieve additional information on the answered CDSS where possible, asking participants additional questions on the mentioned systems could have further refined the analysis. The responses to the questions also depend on the respondents' understanding of the question. Although a trained facilitator was always available for any questions or clarifications, we found that not all answers fitted the original questions, mostly due to lack of knowledge or misinterpretation.

Conclusions

This study collected the experiences with CDSS of healthcare professionals in a purposive sample across Europe and the United States. We found that CDSS use varied across countries samples, with a limited uptake in general, and particularly, for predictive CDSS. Participants mentioned multiple examples of CDSS they used, mostly for diagnostic rather than predictive purposes. Nevertheless, many professionals reported that predictive tools to inform them about patients' health and

XSL•FO
RenderX

intervention outcomes would be most beneficial. This highlights the need for further implementation of current systems and the development of more predictive CDSS.

Acknowledgments

The authors wish to thank all health care professionals involved in the testing and evaluation of the iCARE tool.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Funding

This work was supported by the European Union's Horizon 2020 research and innovation program with grant number 965341. More information on the I-CARE4OLD project can be found on the project and European Commission websites [35,36]. The research works on Assoc. Prof. Daniela Fialová and Dr. Mikko Nuutinen were also supported by the NETPHARM Project (Project ID CZ.02.01.01/00/22_008/0004607), co-funded by the European Union, and the Cooperatio Research Program of the Faculty of Pharmacy, Charles University, Czech Republic.

Authors' Contributions

Conceptualization, data curation, formal analysis, methodology, visualization, writing – original draft: CE
Conceptualization, data curation, formal analysis, funding acquisition, methodology, visualization, writing – original draft: JDAM
Conceptualization, data curation, investigation, project administration, writing – review and editing: A-MH, IH, MN
Conceptualization, software, writing – review and editing: MA
Investigation, writing – review and editing: MM, AS, OŠ, KL, FM, VP, ND, WB
Conceptualization, funding acquisition, writing – review and editing: AD, DF, RL, HPJvH, MH, EH, KS

Conflicts of Interest

The authors declare no conflict of interest. DF declares that her research work was also funded by the institutional program "Cooperatio" of the Faculty of Pharmacy, Charles University, Czech Republic, and by the NETPHARM project (project ID: CZ.02.01.01/00/22_008/0004607), co-funded by the European Union. All funders had no role in the study design, data collection, analysis, interpretation of data, writing of the manuscript, or the decision to submit for publication.

Multimedia Appendix 1
Pre-questionnaire ICARE-tool pilot.
[DOCX File , 43 KB - jmir_v28i1e85071_app1.docx]

References

1. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 2005;330(7494):765 [FREE Full text] [doi: 10.1136/bmj.38398.500764.8F] [Medline: 15767266]
2. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
3. Berner ES, La Lande TJ. Overview of clinical decision support systems. In: Clinical Decision Support Systems. Cham: Springer; 2016.
4. Tunnard I, Yi D, Ellis-Smith C, Dawkins M, Higginson IJ, Evans CJ. Preferences and priorities to manage clinical uncertainty for older people with frailty and multimorbidity: a discrete choice experiment and stakeholder consultations. BMC Geriatr 2021;21(1):553 [FREE Full text] [doi: 10.1186/s12877-021-02480-8] [Medline: 34649510]
5. Karam M, Chouinard MC, Poitras ME, Couturier Y, Vedel I, Grgurevic N, et al. Nursing care coordination for patients with complex needs in primary healthcare: a scoping review. Int J Integr Care 2021;21(1):16 [FREE Full text] [doi: 10.5334/ijic.5518] [Medline: 33776605]
6. Ulyte A, Mehrotra A, Huskamp HA, Grabowski DC, Barnett ML. Specialty care after transition to long-term care in nursing home. J Am Geriatr Soc 2023;71(4):1058-1067 [FREE Full text] [doi: 10.1111/jgs.18129] [Medline: 36435050]
7. Zhao Y, Rokhani FZ, Sazlina S, Devaraj NK, Su J, Chew B. Defining the concepts of a smart nursing home and its potential technology utilities that integrate medical services and are acceptable to stakeholders: a scoping review. BMC Geriatr 2022;22(1):787 [FREE Full text] [doi: 10.1186/s12877-022-03424-6] [Medline: 36207705]

XSL•FO
RenderX

8. Papadopoulos P, Soflano M, Chaudy Y, Adejo W, Connolly TM. A systematic review of technologies and standards used in the development of rule-based clinical decision support systems. Health Technol 2022;12(4):713-727. [doi: 10.1007/s12553-022-00672-9]
9. Gage BF, van Walraven C, Pearce L, Hart RG, Koudstaal PJ, Boode B, et al. Selecting patients with atrial fibrillation for anticoagulation: stroke risk stratification in patients taking aspirin. Circulation 2004;110(16):2287-2292. [doi: 10.1161/01.CIR.0000145172.55640.93] [Medline: 15477396]
10. van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis versus prognosis. J Clin Epidemiol 2021;132:142-145. [doi: 10.1016/j.jclinepi.2021.01.009] [Medline: 33775387]
11. Codagnone C, Lupiañez-Villanueva F. Benchmarking deployment of eHealth among general practitioners. European Commission. 2013. URL: https://digital-strategy.ec.europa.eu/en/library/benchmarking-deployment-ehealth-among-general-practitioners-2013-smart-20110033 [accessed 2026-04-24]
12. Dobrev A, Korte WB, Meyer I, Haesner M, Hüsing T. Benchmarking ICT use among general practitioners in Europe. Publications Office of the European Union. 2008. URL: https://op.europa.eu/en/publication-detail/-/publication/7d72981d-f924-4977-a032-37361bb8b4b3 [accessed 2026-04-24]
13. Samah T, Samar M. Investigating the key trends in applying artificial intelligence to health technologies: a scoping review. PLoS One 2025;20(5):e0322197 [FREE Full text] [doi: 10.1371/journal.pone.0322197] [Medline: 40372995]
14. Bertl M, Ross P, Draheim D. Systematic AI support for decision-making in the healthcare sector: obstacles and success factors. Health Policy and Technology 2023;12(3):100748. [doi: 10.1016/j.hlpt.2023.100748]
15. Ford E, Edelman N, Somers L, Shrewsbury D, Lopez Levy M, van Marwijk H, et al. Barriers and facilitators to the adoption of electronic clinical decision support systems: a qualitative interview study with UK general practitioners. BMC Med Inform Decis Mak 2021;21(1):193 [FREE Full text] [doi: 10.1186/s12911-021-01557-z] [Medline: 34154580]
16. Hoogendijk EO, Onder G, Smalbil L, Vetrano DL, Hirdes JP, Howard EP, et al. Optimising the care for older persons with complex chronic conditions in home care and nursing homes: design and protocol of I-CARE4OLD, an observational study using real-world data. BMJ Open 2023;13(6):e072399 [FREE Full text] [doi: 10.1136/bmjopen-2023-072399] [Medline: 37385750]
17. Hirdes JP, Ljunggren G, Morris JN, Frijters DHM, Soveri HF, Gray L, et al. Reliability of the interRAI suite of assessment instruments: a 12-country study of an integrated health information system. BMC Health Serv Res 2008 Dec 30;8:277 [FREE Full text] [doi: 10.1186/1472-6963-8-277] [Medline: 19115991]
18. Hiltunen A, Haavisto I, Nuutinen M, Lahelma M, Salminen A, de Almeida Mello J, et al. Protocol of the pilot study to test and evaluate the iCARE tool: a machine learning-based e-platform tool to make health prognoses and support decision-making for the care of older persons with complex chronic conditions. BMJ Open 2025;15(4):e101234 [FREE Full text] [doi: 10.1136/bmjopen-2025-101234] [Medline: 40254305]
19. Johnson RB, Onwuegbuzie AJ, Turner LA. Toward a definition of mixed methods research. Journal of Mixed Methods Research 2007;1(2):112-133. [doi: 10.1177/1558689806298224]
20. O'Cathain A, Murphy E, Nicholl J. Three techniques for integrating data in mixed methods studies. BMJ 2010;341:c4587. [doi: 10.1136/bmj.c4587] [Medline: 20851841]
21. Cresswell K, Sheikh A. Organizational issues in the implementation and adoption of health information technology innovations: an interpretative review. Int J Med Inform 2013;82(5):e73-e86. [doi: 10.1016/j.ijmedinf.2012.10.007] [Medline: 23146626]
22. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
23. Liberati EG, Ruggiero F, Galuppo L, Gorli M, González-Lorenzo M, Maraldi M, et al. What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. Implement Sci 2017;12(1):113 [FREE Full text] [doi: 10.1186/s13012-017-0644-2] [Medline: 28915822]
24. McMaughan DK, Nwaiwu O, Zhao H, Frentzel E, Mehr D, Imanpour S, et al. Impact of a decision-making aid for suspected urinary tract infections on antibiotic overuse in nursing homes. BMC Geriatr 2016;16:81 [FREE Full text] [doi: 10.1186/s12877-016-0255-9] [Medline: 27084340]
25. Buck C, Doctor E, Hennrich J, Jöhnk J, Eymann T. General practitioners' attitudes toward artificial intelligence-enabled systems: interview study. J Med Internet Res 2022;24(1):e28916 [FREE Full text] [doi: 10.2196/28916] [Medline: 35084342]
26. Scott IA, van der Vegt A, Lane P, McPhail S, Magrabi F. Achieving large-scale clinician adoption of AI-enabled decision support. BMJ Health Care Inform 2024;31(1):e100971 [FREE Full text] [doi: 10.1136/bmjhci-2023-100971] [Medline: 38816209]
27. Russell R, Lovett Novak L, Patel M, Garvey KV, Craig KJT, Jackson GP, et al. Competencies for the use of artificial intelligence-based tools by health care professionals. Acad Med 2023;98(3):348-356. [doi: 10.1097/ACM.0000000000004963] [Medline: 36731054]
28. Benchmarking deployment of eHealth among general practitioners. Publications Office of the European Union. 2018. URL: https://op.europa.eu/cs/publication-detail/-/publication/d1286ce7-5c05-11e9-9c52-01aa75ed71a1 [accessed 2026-04-24]

XSL•FO
RenderX

29. Damoiseaux-Volman BA, van der Velde N, Ruige SG, Romijn JA, Abu-Hanna A, Medlock S. Effect of interventions with a clinical decision support system for hospitalized older patients: systematic review mapping implementation and design factors. JMIR Med Inform 2021;9(7):e28023 [FREE Full text] [doi: 10.2196/28023] [Medline: 34269682]
30. Gholamzadeh M, Abtahi H, Safdari R. The application of knowledge-based clinical decision support systems to enhance adherence to evidence-based medicine in chronic disease. J Healthc Eng 2023;2023:8550905 [FREE Full text] [doi: 10.1155/2023/8550905] [Medline: 37284487]
31. Kappen TH, van Klei WA, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, Moons KGM. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. Diagn Progn Res 2018;2:11 [FREE Full text] [doi: 10.1186/s41512-018-0033-6] [Medline: 31093561]
32. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019;17(1):195 [FREE Full text] [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
33. Lee T, Shah N, Haack A, Baxter S. Clinical implementation of predictive models embedded within electronic health record systems: a systematic review. Informatics (MDPI) 2020;7(3):25 [FREE Full text] [doi: 10.3390/informatics7030025] [Medline: 33274178]
34. Abell B, Naicker S, Rodwell D, Donovan T, Tariq A, Baysari M, et al. Identifying barriers and facilitators to successful implementation of computerized clinical decision support systems in hospitals: a NASSS framework-informed scoping review. Implement Sci 2023;18(1):32 [FREE Full text] [doi: 10.1186/s13012-023-01287-y] [Medline: 37495997]
35. Individualized CARE for older persons with complex chronic conditions at home and in nursing homes. I CARE 4 OLD. URL: http://www.icare4old.eu [accessed 2026-05-07]
36. Individualized CARE for older persons with complex chronic conditions at home and in nursing homes. CORDIS, European Commission. URL: https://cordis.europa.eu/project/id/965341 [accessed 2026-05-07]

Abbreviations

AI: artificial intelligence
CCC: complex chronic condition
CDSS: Clinical Decision Support System
EHR: electronic health record
I-CARE4OLD: Individualized CARE for OLDer persons with CCC at home and in nursing homes
ML: machine learning

Edited by A Mavragani; submitted 02.Oct.2025; peer-reviewed by W Liedtke, KH Lin; comments to author 20.Jan.2026; revised version received 27.Mar.2026; accepted 30.Mar.2026; published 11.Jun.2026.

Please cite as:
Exmann CJC, Hiltunen AM, Haavisto I, Salminen A, Messo M, Nuutinen M, Hoogendoorn M, Boorsma W, Howard EP, Pešić V, Alon M, Mammarella F, Liperoti R, Švihnosová O, Fialová D, Drapała N, Szczerbińska K, Declercq A, van Hout HPJ, De Almeida Mello J
Uptake of Clinical Decision Support Systems Among Health Care Professionals in Six European Countries and the United States: Cross-Sectional Survey Within the I-CARE4OLD Project
J Med Internet Res 2026;28:e85071
URL: https://www.jmir.org/2026/1/e85071
doi:10.2196/85071
PMID:

©Collin JC Exmann, Anna-Maria Hiltunen, Ira Haavisto, Anna Salminen, Maikki Messo, Mikko Nuutinen, Mark Hoogendoorn, Wiebe Boorsma, Elizabeth P Howard, Vanja Pešić, Mor Alon, Federica Mammarella, Rosa Liperoti, Olena Švihnosová, Daniela Fialová, Natalia Drapała, Katarzyna Szczerbińska, Anja Declercq, Hein PJ van Hout, Johanna De Almeida Mello. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Designing Emotion-Integrated Visualizations for Kidney Function Self-Management: User-Centered Design Study With Youth Transplant Recipients and Caregivers

Chelsea Ng[1], BS; Julia C Dunbar[2], MS, PhD; Lily Jeffs[2], BS; Sanaa Syed[2], BS; Nathan Pan[3]; Hyeyoung Ryu[2], PhD; Jodi Smith[4,5], MPH, MD; Jaime Snyder[2], PhD; Wanda Pratt[2], MS, PhD; Ari Pollack[2,4,5], MD, MS

[1]School of Medicine, University of Washington, Seattle, WA, United States
[2]Information School, University of Washington, Seattle, WA, United States
[3]Boston University, Boston, MA, United States
[4]Department of Pediatrics, University of Washington, Seattle, WA, United States
[5]Division of Nephrology, Seattle Children's Hospital, Seattle, WA, United States

Corresponding Author:
Ari Pollack, MD, MS
Department of Pediatrics
University of Washington
MS OC.9.820
PO Box 5371
Seattle, WA, 98145-5005
United States
Phone: 1 206 987 2524
Email: apollack@uw.edu

Abstract

Background: Understanding complex health information, such as kidney function values (eg, creatinine), is important for youth kidney transplant recipients and caregivers to effectively engage and participate in their care. Information visualizations, such as visual analogies, highlight the similarities between 2 different ideas through visual means and can support understanding of abstract data to facilitate self-management. The study was motivated by the persistent challenge that youth and caregivers face in interpreting complex clinical data, which often remains unactionable and disconnected from their practical information needs.

Objective: This study aims to design and evaluate a novel visualization to support the information needs of youth kidney transplant recipients and caregivers, with the goal of increasing actionable and accessible knowledge for self-management.

Methods: We conducted 2 studies with youth kidney transplant recipients and their caregivers: study 1 identifies visualization elements that support patient and caregiver kidney function information needs, and study 2 evaluates a novel patient-facing kidney function visualization informed by results from study 1. In study 1, participants drew representations of their kidney function over time. These findings led to the development of the novel kidney bean visual. In study 2, participants interacted with the personalized (incorporating electronic health record data) kidney bean visual component for 1 week, then completed a feedback session exploring how the novel visual facilitated retrospective reflection of their health status and the specific behaviors influencing their renal function. We conducted qualitative content analysis on all interviews. All sessions and interviews were recorded, transcribed, and analyzed using content analysis (study 1), thematic analysis, and affinity diagramming (study 2).

Results: In study 1 (n=30), youth and caregivers consistently described the difficulty of understanding their creatinine results and related medical terminology, with the majority using emotions to represent their kidney function. Therefore, our kidney bean visual, a visual analogy, incorporated emotions and colors correlating with their creatinine levels and overall kidney function. In study 2 (n=23), we found that for many participants, the kidney bean visual (1) mirrored and validated their emotions, prompting reflections on their kidney function contextualized by their lived experiences; (2) served as a progress check; and (3) acted as a motivator for healthy kidney care habits.

Conclusions: Kidney function laboratory values, although seemingly straightforward in their intended purpose and use, are more complex from the patient and family perspective, as they also closely align with youth kidney transplant recipients' and

XSL•FO
RenderX

their families' emotions. This study demonstrates the value of designing visualizations that capture these complexities, which can create opportunities for increased understanding of health information and self-reflection to support self-management.

(J Med Internet Res 2026;28:e80481) doi:10.2196/80481

KEYWORDS

adolescents; caregivers; digital health; information visualization; kidney transplant; self-management; user-centered design; young adult

Introduction

Adolescents and young adults (AYAs) with end-stage kidney disease are at higher risk for self-management challenges and poor transplant outcomes compared with both older and younger populations [1]. A positive outcome following a kidney transplant depends both on the transplant team and the lifelong commitment and engagement of the impacted individual [2]. AYA kidney transplant recipients and their caregivers are expected to take an active role in their health posttransplant in the form of engagement and self-management, which has been shown to improve health outcomes, including increased survival of the transplanted organ, and enhanced quality of life [3,4]. For those living with a chronic illness, self-management requires that they (1) medically manage their condition, (2) maintain or adopt new behaviors to help achieve their goals, and (3) cope with the emotional challenges of living with a chronic condition [5].

To support the high frequency of decision-making required to effectively self-manage one's health after receiving a kidney transplant, recipients and their caregivers require accessible and understandable health information to achieve their goals. Shared decision-making, in which clinicians and patients work together to make informed choices based on the best available evidence, ensures that care decisions align with patients' values and needs while respecting their autonomy [6]. However, for patients and caregivers to meaningfully participate in shared decision-making, they need tailored information to support their needs and apply new knowledge to their own health decisions [7]. This support helps them recognize the value of their contributions, fostering a more informed and collaborative decision-making process, resulting in effective self-management.

In the field of pediatric transplantation, the lack of accessible, longitudinal data visualization drives illness uncertainty, a cognitive state elicited when an individual has difficulty understanding the meaning of an illness-related event because health outcomes are unpredictable, illness symptoms of events are ambiguous, or there is a lack of information about the illness or treatment [8]. While clinical conversations with this population focus heavily on creatinine as a key contributor to kidney function, limited research has examined the depth of youth and caregiver understanding of this concept or the effectiveness of existing materials in translating this relationship into actionable self-management strategies. When this complex data—data that demand interpretation across clinical, behavioral, and lived experience dimensions—are presented without clear, accessible context, they lead to a limited information environment that is linked to increased parental distress, lower patient self-management, and ultimately, poorer long-term clinical outcomes [9].

One promising approach to facilitating this understanding is the use of visualizations. Defined as the process of transforming data, information, and knowledge into visual form, visualizations can help users interpret complex information by leveraging the human ability to visually recognize patterns (eg, differences in shapes, sizes, colors, and spatial relationships) to assign meaning [10,11]. By making abstract health data more accessible, visualizations can support patients in contextualizing their own health information, regardless of their level of health literacy or numeracy [11,12]. Further supporting their utility, information visualizations can also be integrated into widely used digital platforms such as mobile apps, websites, or even patient portals, making them a convenient and scalable solution for enhancing patient comprehension in shared decision-making. Despite their potential, patient-facing health visualizations remain underdeveloped, especially for youth populations, compared with those designed for health care providers. A systematic review conducted by Turchioe et al [13] found that a significantly larger number of articles focused on health care provider–facing visualizations compared with patient-facing visualizations [13]. Among the patient-facing visualizations that do exist, prior work has established that presentation format significantly shapes patients' affective (ie, how results make them feel) and cognitive perception of laboratory results, with evidence favoring visual formats that include contextual reference ranges over numeric displays alone [13,14]. However, nearly all this research has examined single-time-point results, and how youth interpret longitudinal trends in their laboratory values, which are critical to self-management of chronic conditions such as kidney transplant, remains poorly understood. This highlights a broader gap in tools designed to support patients in understanding their personal health information and engaging in informed decision-making. Recognizing this, we explored how a novel information visualization might facilitate the translation of complex clinical data, specifically longitudinal kidney function laboratory values, into meaningful, reflective understanding for AYA kidney transplant recipients, supporting them in better understanding of their health and engaging more actively in their care.

We addressed this through 2 sequential studies, first designing and then evaluating a novel visualization of kidney function laboratory values:

- Study 1: what visualization elements address the information needs and illness uncertainty of AYA kidney transplant recipients and their caregivers to support the reflection necessary for active engagement in their care?

XSL•FO
RenderX

- Study 2: how does a novel visualization incorporating these elements facilitate the translation of complex clinical data into meaningful, reflective understanding that supports youth with chronic kidney disease (CKD) in engaging more actively in their health?

Together, these studies identify the information needs of AYA kidney transplant recipients and their caregivers, demonstrate how a novel visualization can support meaningful reflection on complex health data, and offer design recommendations to guide future development of patient-facing health visualizations for youth with chronic conditions.

Methods

Overview

This work is part of a larger multipart study that aimed to design a multifaceted app prototype supporting the information needs of AYAs who have received a kidney transplant and their caregivers. In this study, we present 2 studies (Figure 1) that focus on the development and evaluation of a novel patient-facing visualization. This visualization was incorporated into a multifaced app prototype to enhance understanding of kidney function and support patients in self-management. Study 1 used a series of user-centered design sessions to inform the design and development of a patient-facing visualization of key metrics related to kidney function. The research team took the findings from study 1 and then designed the visualization for study 2, which focused on piloting and evaluating the visualization through a user feedback session.

Focus groups and interviews were facilitated by 3 researchers (CN, JCD, and AP) with prior experience in qualitative inquiry. The team represented a range of academic and professional tiers, including a clinical research coordinator, a PhD candidate in her final year, and an associate professor/nephrologist. Drawing prompts and semistructured interview guides were developed by the authors and pilot tested for clarity and age appropriateness among a small sample of youth aged 12-21 years, including those with and without CKD.

Figure 1. Flowchart of the research phase. KFV: kidney function visualization.

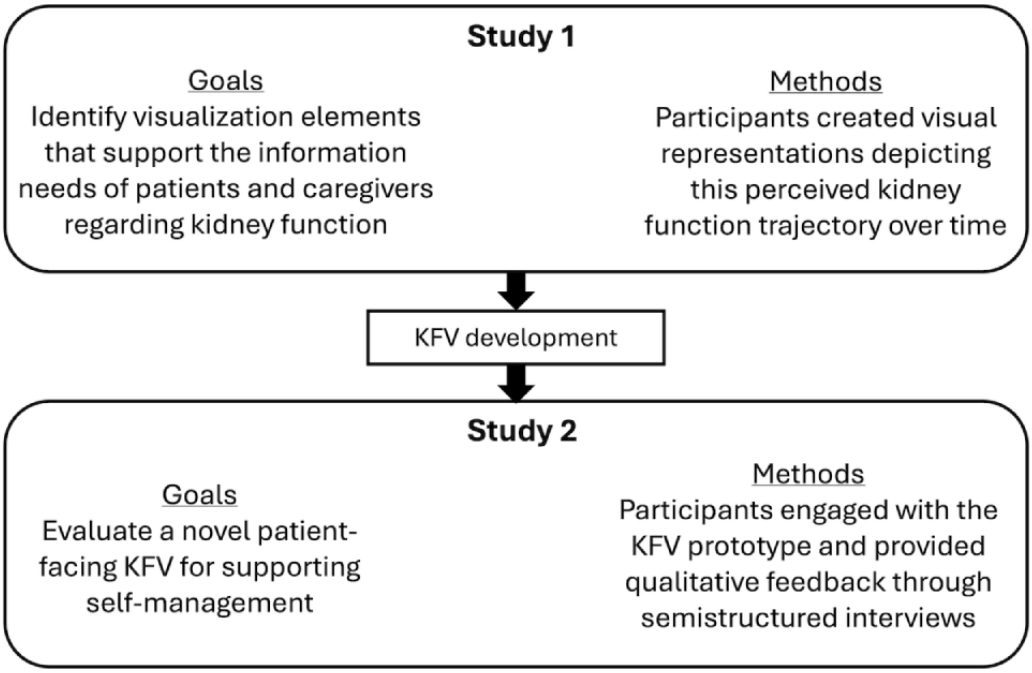

Recruitment

Participants from study 1 were recruited from Seattle Children's Hospital (SCH), The Johns Hopkins Hospital, and the Improving Renal Outcomes Collaborative Community Engagement Workgroup, a learning health network designed to improve the lives of children and AYAs who have had a kidney transplant [15]. Participants from study 2 were recruited only from SCH. A single-site approach was adopted because the SCH population was sufficient to meet study objectives.

Recruitment at SCH and Johns Hopkins Hospital was done via convenience sampling by identifying patients (and their caregivers) from a transplant clinic list (nephrology and transplant teams maintain a list of patients who have received a kidney transplant and are currently receiving care). Patients had to meet the following inclusion criteria to participate in either study: aged 10-21 years, had received a kidney transplant at least 6 months earlier, had access to the internet, and were English-speaking (both studies) or Spanish-speaking (study 1 only). Participants were initially contacted for recruitment either in person during their clinic visit, where they were given a recruitment flyer, or remotely via phone call or email. Initial recruitment via email or phone call was used in cases in which email communication already occurred clinically and a patient's email address was already known to the study team. Improving Renal Outcomes Collaborative participants (study 1) were invited to participate through distribution of a recruitment flyer sent via existing email distribution lists.

During the informed consent process, participants were briefed on the study objectives, and researchers addressed all participant inquiries to ensure transparency. To establish rapport and context, 1 investigator (JD) shared a personal history of kidney

XSL•FO
RenderX

disease, and another (AP) disclosed their role as a nephrology provider, although it was ensured that none of AP's own patients were recruited for the study.

Procedures

Study 1: User Information Requirements and Development of Kidney Function Visualization Interface

We conducted 8 separate 90-minute design sessions using the video conferencing platform Zoom [16], with 4 sessions exclusively with caregivers and 4 sessions exclusively with youth. These sessions were scheduled independently and organized by the youths' age ranges (eg, aged 10-12 years, aged 13-17 years, and aged 18-21 years) and preferred language (English or Spanish). To facilitate small-group interaction within each 90-minute meeting, participants were placed into separate breakout rooms of 3-5 individuals, each led by a research facilitator. At the conclusion of each design session, all participants returned to the main Zoom room to present and discuss the materials they created within their respective breakout rooms. Participants received a US $25 Amazon electronic gift card for their participation.

To equip participants for their design session, the research team emailed them a design session agenda and mailed them a physical design session toolkit before their session. Design toolkits included the following materials: printouts of the agenda and task cards, colored markers, example visualizations, additional drawing and writing paper for tasks, and emoji stickers representing various emotions, objects, animals, foods, activities, travel locations, and more. Initially, in the design sessions, we asked participants to create a line graph of their creatinine, a waste product filtered by the kidneys commonly used as an indicator of kidney function, using a printout with unlabeled x- and y-axes from their design toolkit (Textbox 1). Participants were asked to create a line graph because this is how kidney function over time is typically presented in the clinic. We also asked them to refer to a printout of an example line graph to aid in their drawings (Figure S1 in Multimedia Appendix 1).

We considered graphs to be successful or correct based on the following criteria: (1) the graph aligned with the concepts that we asked participants to draw, (2) the graph had x- and y- axes with appropriate increments and levels based on the prompt, and (3) the participant's description correlated with what they drew. Further, 2 primary researchers independently graded each submission and met to resolve any discrepancies until a final consensus was reached. We did not review participants' data to verify accuracy.

However, after we completed the design sessions with 4 groups of participants (2 youth groups and 2 caregiver groups), we found that youth and caregivers struggled with the task of representing their data on a graph. Therefore, for the remaining 4 participant groups (2 youth groups and 2 caregiver groups), we broadened the prompt and made it easier to understand by removing the requirement to use creatinine specifically or to represent their data on a line graph. Instead, we invited them to draw any image, figure, or visual that represented their kidney function over time (Textbox 1). Drawings were completed individually by all participants and then presented to the others in their session.

Textbox 1. Study 1 design session: original prompt and new prompt.

Original prompt—task 2, part 1

For this task you will have 10 minutes to use your materials in your design toolkit to create 2 drawings. We would specifically like you to make sure to use your emoji stickers within your line graph drawings.

- The first drawing should be a line graph of what you think a perfect creatinine level should look like over time.
- The second drawing should be a line graph of what your creatinine levels have looked like since your transplant.

New prompt—task 2 part 1

For this task you will have 5-10 minutes to use your design toolkit to create a drawing. We encourage you to use all parts of your toolkit, such as the colored markers and emoji stickers. Feel free to use any household items you have as well.

- The drawing should be an image, figure, or other representation of your kidney function over time. You will need to include specific elements in your image or figure; please see the drawing paper in your toolkit for further instructions about these elements.

Bridging Study 1 and Study 2—Visualization and Tool Design

Based on the findings from study 1, we created the kidney bean visual, which is a cartoon kidney bean with various facial emotions and colors aligned to different stages of kidney function (Figure 2 [17]). This visualization was designed to be a key feature of a larger mobile app prototype, the My Kidney TREK (Thinking, Reflecting, and Empowering for Kidney Transplant Recipients) app. This app included 3 main components highlighting kidney function, medication adherence, and storytelling to support the user's reflection and health engagement in their kidney transplant journey. In this study, we present our findings only on the design of the kidney bean visual feature and accompanying kidney function component.

The prototype was created in Figma, a collaborative web app for interface design [18]. In our novel kidney bean visual, we represent kidney function values through the personification of the kidney organ, with colors and emotions that correspond with creatinine and glomerular filtration rate (GFR) value (ie, green, happy bean for lower creatinine values and better kidney function; Figure 2). The aim of the personification of the kidney bean visual was to act as a cognitive representation of the abstract concept of kidney function values by basing it on the

XSL•FO
RenderX

physical functioning of the kidney and the emotions associated with it. The kidney function section was designed to be personalized with creatinine laboratory values retrieved from the user's medical chart, the corresponding CKD stage, and an explanation of the stage (Figure 3A). Participant's CKD stage was determined by a pediatric nephrologist on the research team. The stage of CKD refers to how functional the kidneys are, with higher stages corresponding to lower function. CKD stage is determined by calculating the estimated GFR from an individual's creatinine level. The GFR indicates how well the kidneys are filtering waste from the blood [17]. Creatinine values were also displayed alongside the kidney bean visual (Figures 3A and B). Values from 4 different time points were displayed: (1) before kidney transplant, (2) 1 month after kidney transplant, (3) 1 year after kidney transplant, and (4) the most recent laboratory value at the time of the study. We chose multiple time points to represent the concept of a journey (Figure 3B). Additionally, definitions of creatinine and GFR, as well as healthy kidney habits, were included (Figures 3C and D).

Figure 2. Key of chronic kidney disease stage and glomerular filtration rate value with corresponding kidney bean visual that only researchers had access to.

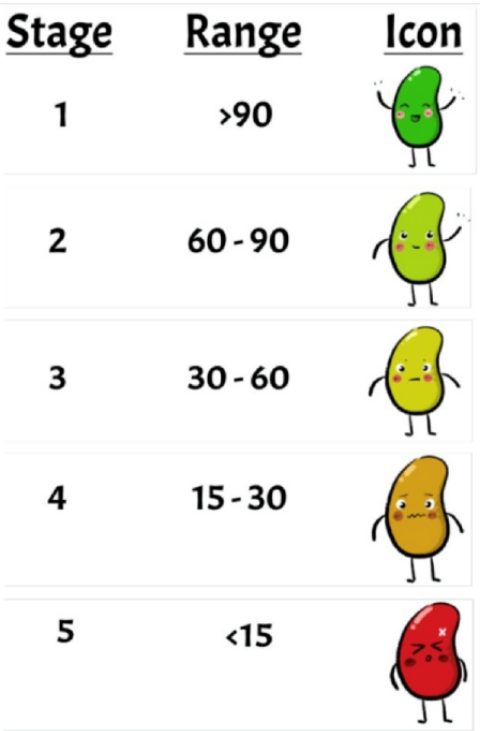

Figure 3. Kidney function component. (A) Individual results page that includes personalized participant data (from the electronic health record), (B) timeline page that includes the kidney bean visual and corresponding creatinine values at 4 different time points, (C) definitions of creatinine and glomerular filtration rate (GFR), and (D) description of chronic kidney disease (CKD) stages and a list of healthy kidney habits.

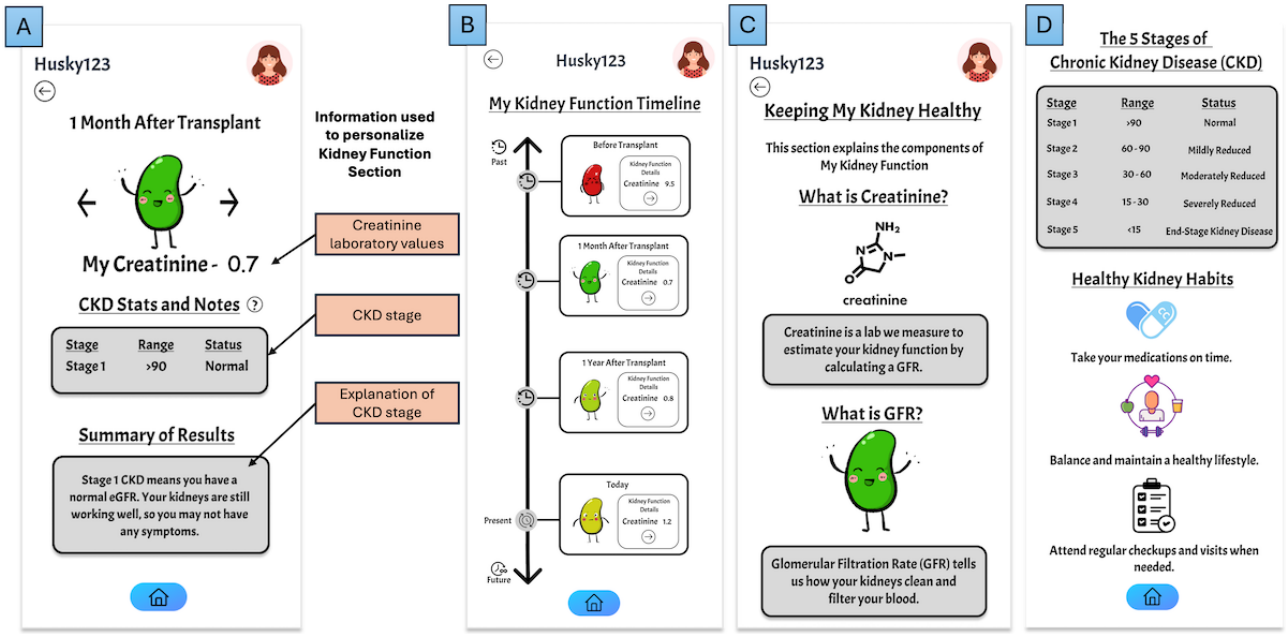

XSL•FO
RenderX

Study 2: Piloting the Kidney Bean Visual and Accompanying Component

We piloted the kidney bean visual and corresponding kidney function component of the My Kidney TREK app prototype with a cohort of AYAs and caregiver participants. Further, 2 sessions were conducted: an initial guided walkthrough session to introduce the prototype, followed by a user feedback session 1 week later. Participants were guided by a member of the research team in individual walkthrough sessions on how to interact with the prototype. Youth and caregivers completed separate walkthroughs unless they requested to complete them together. Following their walkthroughs, participants received a password-protected link (to which only the research team had access) to their own personalized prototype that could be accessed from any smart device (eg, laptop, phone, and tablet) and used independently throughout the remainder of the study. Participants had 1 week to interact with the prototype and received US $25 in Amazon electronic gift cards. After the guided walkthrough, participants were given minimal instruction for the trial period, allowing interaction with the prototype to be self-directed.

Following the 1-week period of independent interaction with the prototype, researchers conducted a user feedback session to evaluate participants' experiences using the prototype and how it supported self-reflection. These semistructured interviews were conducted virtually via Zoom [16] and lasted approximately 45-60 minutes. The research team conducted the youth and caregiver interviews separately unless there was a specific request to complete the interview together. Participants were asked open-ended questions about the kidney function component to probe about what they liked or disliked about the component, how they used it to reflect on their journey, and what they learned (Figure S2 in Multimedia Appendix 1). Feedback from participants on interim findings was not incorporated into the study methodology. Participants received US $25 in Amazon electronic gift cards for participating in the interview.

Analysis

Overview

All sessions were audio and video recorded and later submitted to a HIPAA (Health Insurance Portability and Accountability Act)–compliant professional transcription service to be transcribed verbatim. The research team completed the analysis of both studies using MAXQDA analysis software [19].

Study 1 Analysis

Using content analysis [20], we analyzed deidentified design session transcripts along with participants' drawings created during the design sessions. In total, 8 transcripts were analyzed from 8 design sessions. A total of 4 members of the research team took part in iterative rounds of transcript analysis. In the first round of analysis, each member independently analyzed and coded 1 design session transcript. The research team met after the first round of independent analysis to discuss emergent codes. After the initial round of analysis, the research team completed 3 more rounds of analysis, during which they met frequently over 1 month to discuss newly emerging codes, categorize codes based on their meanings, and resolve any analytical discrepancies or interpretations [21]. After analyzing all design session transcripts, the researchers met 3 more times to cluster and develop themes based on the codes through inductive analysis [21].

Study 2 Analysis

Data analysis for study 2 was conducted in two parts: (1) thematic analysis [21] and (2) affinity diagramming [22-24]. In part 1, a total of 2 researchers developed a codebook (Figure S3 in Multimedia Appendix 1) from the research questions and assessment of the My Kidney TREK app prototype. This was then iteratively refined through an additional inductive analysis of 2-3 transcripts with the rest of the team. Further, 3 rounds of coding were conducted over 3 weeks. In rounds 1 and 2, pairs of researchers independently coded 2 to 3 transcripts using the codebook and flagged emergent themes. The team then met to resolve discrepancies and revise the codebook. In round 3, a total of 4 members each coded a new transcript to validate the updated codebook. Finally, the main researcher reviewed all transcripts to ensure consistent application of codes.

After identifying relevant quotes in part 1 of the analysis, the research team performed an inductive analysis via affinity diagramming [22-24] using FigJam [25] in part 2. The research team met to sort the quotes into emergent clusters, which were discussed and refined.

Ethical Considerations

Ethical approval for both studies was obtained from the Seattle Children's Institutional Review Boards (STUDY00004468 and STUDY00002642). Participants in both studies received information about the study and provided informed consent and assent before study procedures were initiated. To ensure participant privacy and confidentiality, all data were deidentified and stored on secure, password-protected servers, with access restricted to the research team. All youth and caregiver participants received a US $25 Amazon electronic gift card upon the completion of each study activity, including design sessions, walkthroughs, and interviews.

Results

Overview

Despite study 1 and study 2 being iterative and progressive, the results indicate overlapping themes and concepts. Study 1 helped identify initial themes, while study 2 tested a prototype developed in response to those findings and further validated key themes from study 1. The findings section is organized based on these overlapping themes, in which themes 1 and 2 present results from both studies and theme 3 presents results from study 2 only (Textbox 2).

Participants are labeled by study number, role (youth or caregiver participant), and participant number—for example, S1Y1/S1C1 for Youth 1 and Caregiver 1 from Study 1, and S2Y1/S2C1 for Youth 1 and Caregiver 1 in Study 2.

XSL•FO
RenderX

Textbox 2. Themes from studies 1 and 2.

Theme 1: patient and caregiver challenges in understanding kidney function results

- Study 1 findings: challenges in interpreting kidney function data
- Study 2 findings: reinforcing and validating information gaps

Theme 2: emotional dimensions of kidney function

- Study 1 findings: identifying emotional ties to kidney function
- Study 2 findings: emotional engagement with the kidney bean visual

Theme 3: making kidney function results more accessible, understandable, and actionable

- Study 1 findings: using longitudinal data to track kidney health
- Study 2 findings: improving information accessibility through layout and design

Participant Characteristics

Study 1

Author JD recruited a total of 30 participants: 14 pediatric kidney transplant recipients and 16 caregivers for study 1 (Table 1). The median age of youth participants was 15 (range 10-21) years. A total of 51 youth and caregivers were contacted; 21 participants were uninterested or had time constraints. A total of 6 participants (3 youth and 3 caregivers) provided consent and later withdrew because of time constraints. Further, 23 participants completed the study in English, while the remaining 7 completed the study in Spanish. Most patients (12/14, 86%) reported that it was their first kidney transplant and that they had received dialysis before kidney transplantation. A total of 6 youth and caregiver participants (3 youth-caregiver pairs) from this study also participated in study 2 (institutional review board approval was obtained to include patients from study 1 in study 2).

XSL•FO
RenderX

Table 1. Study 1 and study 2 participant demographics.

Characteristic	Study 1, n (%)		Study 2, n (%)	
	Youth (n=14)[a]	Caregiver (n=16)	Youth (n=13)[b]	Caregiver (n=10)
Age (years)				
10-17	12 (86)	0 (0)	9 (69)	0 (0)
18-21	2 (14)	0 (0)	4 (31)	0 (0)
>22	0 (0)	14 (88)	0 (0)	8 (80)
Not disclosed	0 (0)	2 (12)	0 (0)	2 (20)
Gender identity				
Female	5 (39)	15 (94)	6 (46)	9 (90)
Male	8 (61)	1 (6)	5 (38)	1 (10)
Nonbinary	0 (0)	0 (0)	1 (6)	0 (0)
Gender fluid	0 (0)	0 (0)	1 (6)	0 (0)
Race[c]				
Asian/Pacific Islander	2 (16)	2 (14)	3 (23)	3 (30)
Black or African American	1 (8)	0 (0)	2 (15)	0 (0)
White	6 (50)	9 (64)	8 (62)	6 (60)
Other	4 (33)	5 (36)	1 (6)	1 (10)
Not disclosed	1 (7)	0 (0)	0 (0)	1 (10)
Ethnicity[c]				
Hispanic and/or Latino	4 (29)	5 (31)	1 (7)	1 (10)
Non-Hispanic or Latino	8 (57)	11 (68)	12 (93)	9 (90)
Not disclosed	2 (14)	0 (0)	0 (0)	0 (0)

[a]Among study 1 youth participants (n=14), median years since transplant was 2 years (range 0.5-14.5 years), and 9 participants (69%) received a deceased donor transplant.

[b]Among study 2 youth participants (n=13), median years since transplant was 5 years (range 0.5-14 years), and 8 participants (62%) received a deceased donor transplant.

[c]Participants could select more than 1.

Study 2

Authors JD and CN recruited a total of 23 participants: 13 pediatric kidney transplant recipients and 10 caregivers (Table 1). The median age of youth participants was 16 (range 13-21) years. A total of 54 youth and their caregivers were contacted; 35 participants were uninterested or had time constraints. A total of 10 youth and caregivers provided consent and later withdrew because of time constraints. All participants completed the study in English. All patients reported that it was their first kidney transplant, and most patients (10/13, 77%) reported receiving dialysis before kidney transplant.

Data collection continued until thematic saturation was achieved, with a total sample size of n=30 for study 1 and n=23 for study 2. Although the literature suggests saturation is often reached within 9-17 interviews, we engaged a larger sample of 30 participants to ensure breadth. Recruitment concluded only when concurrent data analysis confirmed that no novel themes were emerging.

Theme 1: Patient and Caregiver Challenges in Understanding Kidney Function

Study 1 Findings: Challenges in Interpreting Kidney Function Data

Initially, our first 2 groups of participants (n=16) were asked to create kidney function visuals in the form of line graphs. A total of 62.5% (5/8) of caregiver participants drew and described their graphs successfully in response to the first prompt. The caregivers who did not have fully correct graphs had minor incorrect components to their drawings (S1C1, S1C5, and S1C8). Some common mistakes found in caregivers' graphs that were semicorrect included incorrect creatinine increments and unclear graph points and line placement. Only a handful of youth participants created correct drawings and descriptions (S1Y10, S1Y5, and S1Y2), while others produced semicorrect drawings (S1Y7 and S1Y9) or incorrect drawings (S1Y1, S1Y6, and S1Y8). Some common mistakes found in youths' graphs that were either semicorrect or incorrect included being unable to draw their creatinine results because of a lack of understanding about creatinine, incorrect creatinine increments,

XSL•FO
RenderX

and incorrect placement of graph points and lines. Importantly, the correctness of the drawings was not correlated with age.

When creating their drawings for the line graph task, youth and caregivers consistently described the difficulty of understanding their laboratory results and related medical terminology. This struggle with graph literacy and terminology was reflected in the creation and descriptions of participants' drawings. For example, S1Y1, aged 12 years, described how they relied on their caregiver to understand their creatinine values, which led to a lack of their own understanding when creating their drawings.

> *I think it means this [referring to graph] because I feel like grammie's [creatinine levels] is similar to mine, even though I don't know her levels so I just kind of guess. About 10 to 8, that seems normal to me.* [S1Y1, aged 12 years]

Several caregivers also discussed how their trust in the clinical care teams could influence families' understanding and acceptance of kidney function results:

> *I think it really comes down to, do you have the trust in your physician to interpret those results? Because we don't have the background to interpret these results. We say, oh, 1.2, 1.3, 1.4. We know that it's with creatinine and each point is more significant than maybe in another reading for instance. But is it our job to keep track of the numbers or is it our faith in our medical team to interpret the numbers correctly? I think it's a matter of trust. Some people might not trust the results.* [S1C3]

Additionally, several caregivers expressed self-consciousness about not fully understanding the laboratory values and their ability to reproduce them on a graph. For example, S1C8 expressed embarrassment several times during the presentation of her graph:

> *You guys are all really smart and I'm clearly very ignorant. [...] And so I didn't really, I still have a hard time understanding the numbers. I just know that when Dr. X, when we were seeing her, she would tell me the numbers and they were too high. [...] I made a mistake and backwards, but I'm also probably a little dyslexic and I'm just not very math oriented. So this one, I'm embarrassed about, but yeah. I don't understand the numbers. I never have been able to.* [S1C8]

These sentiments revealed a broader theme: both youth and caregivers faced challenges interpreting laboratory results and medical terminology, which directly influenced how they engaged with the task of representing their kidney function on a line graph.

Study 2 Findings: Reinforcing and Validating Information Gaps

Reinforcing the findings from study 1, many participants expressed that they had difficulties understanding laboratory values presented in the clinic. For example, S2Y1 shared that they had an especially challenging time understanding laboratory results, tending to nod along when presented:

> *[...] I barely understand what those mean. Every time they pull out the lab result sheet at the doctor's office, I'm just sitting there nodding because I don't-- I understand what some of the words mean, but not when put together in that order.* [S2Y1, aged 16 years]

Caregivers like S2C9 expressed that they tended to avoid looking at laboratory values and numbers and focused more on actionable items. Additionally, retrospectively reflecting on their participation in study 1, they also expressed self-consciousness about their lack of knowledge of kidney function laboratory values.

> *When we were still seeing Dr. X, they would say, "Her creatinine is--" I don't even think she'd give me-- if she gave me a number, it would go in one ear and out the other. I'm just like, "Tell me what I got to do to make her better. Tell me what I got to do to make [inaudible]," because it didn't mean anything to me. And with S2Y9, there were no numbers. So I guess I've never-- and I was kind of embarrassed when we did the last study with the other families how well the mothers knew all about that. And that's something I, I don't know if I just avoid it out of fear, maybe, to not know. Do you know what I mean? It's just something-- but it is good for me to see that, especially going from 0.1 to 0.7. That's a huge, huge difference.* [S2C9]

Summary—Persistent Barriers to Understanding Kidney Function

Our findings indicate that youth patients and their caregivers continue to experience information gaps in (1) their overall understanding of general kidney function clinical content and (2) their ability to interpret and contextualize clinical data as it is currently presented (eg, graphs).

Theme 2: Emotional Dimension of Kidney Function

Study 1 Findings: Identifying Emotional Ties to Kidney Function

Adjusting our prompt, we asked the remaining study 1 participants (n=14) to draw any image, figure, or visual they preferred to represent their kidney function (Table 1). With this new prompt, youth and caregiver participants created and described visuals that were unexpected in several ways. One notable finding was the recurring theme of connecting emotions to kidney function in their drawings and descriptions (Table 2; Textbox 2). For instance, one caregiver chose to use the stickers we provided to express all of the different emotions that she had felt about her child's kidney function (Figure 4A).

> *Because I'm still in the worry part of S1Y11's transplant, because they are still up and down with her medications...So it's just really with her since transplant, I've just been worried that she's not going to keep the new kidney, because it's just been worry*

XSL•FO
RenderX

> *after worry after worry, when she gets her blood work.* [S1C11]

Even though they were in different Zoom rooms, S1Y11 (Figure 4B) similarly used her emojis to represent her kidney function and how it made her feel physically and emotionally:

> *This is my... Before it was a 1% and I was scared that I wasn't going to get a kidney and I was sleepy and tired and it's at 1% and this is me after, happy and the fruits are for healthy and the hearts are for...[Things in the hearts] is all the stuff that I need to keep healthy like my medicine, and my stuff I need to keep healthy like fruits, lots of fruits, some vegetables because I only like some vegetables and yeah.* [S1Y11, aged 14 years]

Additionally, 1 notable observation is that after changing the prompt to allow for drawings of any kind to represent their kidney function over time, we found that no youth or caregiver participants chose to draw their kidney function in the form of a graph, even though they could do so.

For example, S1C12's drawing (Figure 5) used the analogy of a bomb to represent how their family felt after the initial diagnosis of needing a kidney transplant through to today.

> *And then the question marks are we didn't know anything about what was going on until boom, we found out he needed a kidney transplant, and it just devastated our life, his life and it's like somebody dropped a bomb on us…And then still got bombs on the other side with fuses lit, because we don't know how this new kidney's going to do. And that's why the question marks.* [S1C12]

Each of these examples shows how participants gravitated toward creating visuals that depicted the emotions connected with their changing kidney function, rather than describing the change in kidney function through a graph. As youth and caregivers presented their drawings about the emotions tied to their kidney function, other participants often responded with similar sentiments and feelings, building rapport between participants as well as highlighting the importance and stress that comes along with doing laboratory results.

For example, S1C13 asked a follow-up question to another caregiver about their daughter's condition and expressed similar sentiments and gratitude:

> *You mentioned that your daughter haven't been doing well. May I ask, is it because of her condition, or any type of rejection, notification, or any sign of rejection that's making your daughter not feel well? [After the other caregiver's reply] Yeah, of course. Yeah, we're all going through similar... Not every situation is the same, but yeah. And I thank you for sharing that with us.* [S1C13]

The findings from study 1 revealed that participants consistently chose to express the emotional impact of kidney function over time, rather than focusing on clinical metrics or drawing graphs. This not only highlighted how emotion is deeply tied to understanding kidney function information but also fostered moments of empathy and connection among participants.

XSL•FO
RenderX

Table 2. Examples of correct, semicorrect, and incorrect caregiver and youth drawings and descriptions.

Categorization and example drawings	Quotes
Correct	
	“So his creatine number looks good until he was age 12. Then it started going up at all our visits for over the last two years and then they said dialysis. Then they said, right here, it's time for transplant. Then we get the transplant, and it goes down to a more normal level, which varies for everybody...so I put him there. He's like 1.1, 1.2, but as he gets older, it'll start creeping up just like all of us and his will go faster though, because it's someone else's kidney, so it's not his own. So it'll have more challenges against it.” [S1C9]
	“Pre-transplant, my creatinine was very high, at 10 or something. And then transplant came all the way down to about 1.0 and then it's been up and down around the 1.0 area for last year-ish. So that was mine...The first one is surprised face because it's super high compared to now. And then this one, the heart, it went all the way down so I was like, wow, that's crazy. And then here where it's cruising along, and I thought that was cool” [S1Y5, aged 15 years]
Semicorrect	
	“She started out high and she really bounced this around a lot. So, she's up and down, and actually she's still an up and down. That's where she goes up and down” [S1C1]
	“And then here's mine for seven months. It was at 0.9 and then it went up to 1.0” [S1Y9, aged 15 years]
Incorrect	
	“So I am supposed to take six bottles per day. But I don't know, I only drink like two bottles of water per day, but I don't like four more bottles of all. That's all I can explain” [S1Y6, aged 11 years]
	“Mine are always 5 to 6, sometimes 7. It’s very rare to have an 8.” [S1Y1, aged 12 years]

Figure 4. Kidney function drawings from (A) S1C11 and (B) S1Y11.

XSL•FO
RenderX

Figure 5. Kidney function drawing from S1C12.

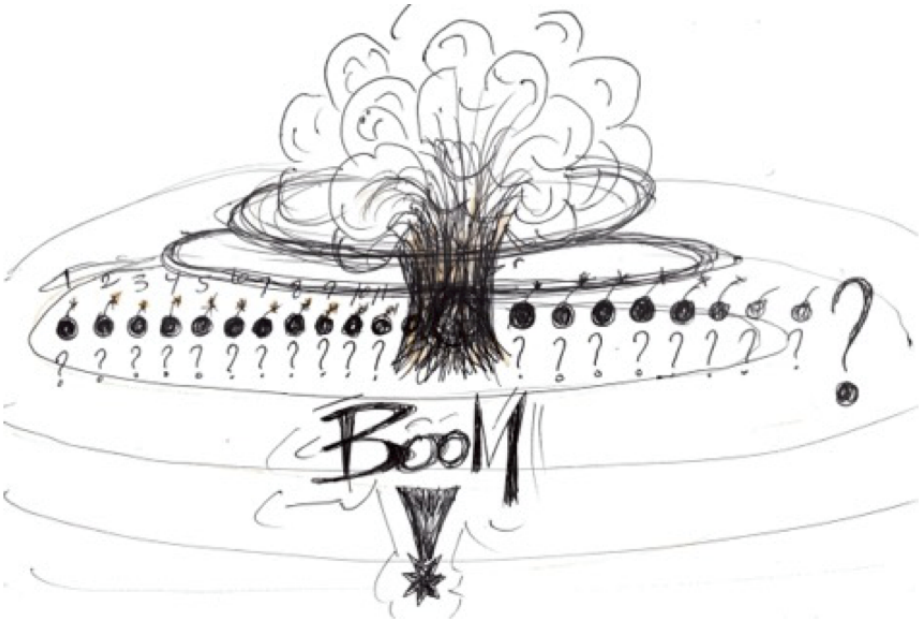

Study 2 Findings: Emotional Engagement With the Kidney Bean Visual

Overview

In response to the kidney bean visual, which the researchers developed based on insights from study 1, we found that participants identified and engaged with the emotions displayed in the visual.

Kidney Emotion as a Mirror of Their Own Emotions

The overwhelming majority of both youth and caregiver respondents appreciated the personification and emotional expressions of the kidney bean visual, finding the associated colors both simple and reflective of their own emotions.

> *The face was definitely-- it takes a kidney that nobody ever sees, hides in your body. It makes it-- like we talked about. I've talked about it for years. I was like, "Your kidney is a raisin. Go and drink some water." Go plump it up. So it was funny to have just a face on it. And I think that as much as a 14-year-old doesn't want to admit it, I think that looking at a face is a big difference from just thinking intangibles. It's in my body.* [S2C8]

> *And I like the fact that the art changes, too. Yeah. [...] It's more like a positive thing. And it makes me feel like, "Oh, I'm happy. He's happy.* [S2Y12, aged 21 years]

For caregivers, resonating with the emotions tied to the kidney bean prompted reflection on their own hopes for their child's kidney journey.

> *I guess, yeah, it's kind of encouraging to see how the kidney is happy in a similar way. I mean, I think this is great for a child to see, yeah, to show them how far they've come, for sure. For Y09, yes. Yeah. I mean, it's encouraging even for me to see it. And I [just want this?] kidney to last. I mean, three years. It's already flown by.* [S2C9]

Kidney Emotion as a Motivator

Many respondents shared that seeing their creatinine values over time, paired with the kidney bean visual, motivated them to take action in maintaining their kidney health. For example, S2Y7 explained how they think a visual can be more encouraging as it shows the individual a representation of the condition, rather than an abstract description.

> *I feel like it could be motivating to a lot of people like, "Oh, my kidney's not doing good." And they can actually see that visual. I feel like a lot of people, they can hear, "Oh, your kidney's not doing good." But sometimes they need a visual. Or sometimes, they'll show pictures of maybe a diseased kidney or whatever. It really encourages people like, "Okay, I want to get on track. I want to keep my kidneys healthy.* [S2Y7, aged 16 years]

Multiple participants also expressed that seeing the emotions tied to the kidney bean reminded them to partake in healthy habits that maintained their kidney health.

> *It just, again, showed that I need to keep it healthy and do all my daily routines and medicines.* [S2Y17, aged 13 years]

> *I think it's motivating seeing the little kidney all happy and stuff. [...] I feel like that could be really an encouragement in that visual.* [S2Y7, aged 16 years]

Summary—Emotion as a Key Factor in Kidney Function Perception

From our findings in study 1, kidney function laboratory values, although seemingly straightforward in their intended purpose, that is, to inform clinical teams and kidney transplant families about a youth's kidney function, are also tied together with the emotions of individuals and families. In study 2, we found that youth and caregivers resonated with the emotions and colors of

XSL•FO
RenderX

the kidney bean visual, with the visual acting as a mirror of their own emotions and motivating them toward self-management.

Theme 3: Making Kidney Function Data More Accessible, Understandable, and Actionable

Overview

In this section, we report on feedback received from study 2 on how the design of our kidney function component and kidney bean visual helped to bridge existing information gaps.

Study 2 Findings: Using Longitudinal Data to Track Kidney Health

Overall, participants appreciated the presentation of the creatinine laboratory values and stages over time, as it prompted reflection on their transplant process as a journey and reminded them that the journey was still ongoing.

> *I think this is great for a child to see, yeah, to show them how far they've come, for sure. For S2Y9, yes. Yeah. I mean, it's encouraging even for me to see it. And I just want this kidney to last.* [S2C9]

> *I liked the timeline of being able to see how bad my creatinine was or how abnormal it was and then being like, "Oh, it's good. It's normal." And I like the fact that the art changes, too.* [S2Y12, aged 21 years]

In particular, caregivers noted that seeing the changes in creatinine over time helped them recognize how drastically their child's health had improved while also serving as a reminder that ongoing effort was needed to prevent kidney function from declining.

> *I learned that her kidney transplant was amazing, that it brought it to such a drastic lower level than it was prior, but that it's progressively increasing. But I think that is one thing that we do look at the labs quite a bit. So, I have had that in the forefront of my mind is her creatinine, but I think that's really the only level I really paid great attention to.* [S2C17]

Study 2 Findings: Improving Information Accessibility Through Layout and Design

When asked what participants liked about the kidney function component, many identified the definitions explaining creatinine and CKD stages as valuable. This feature was well received with both youth and caregiver participants, with many noting that while these medical terms were used consistently in their care, they had never been officially defined. For example, S2Y7 noted the simple language used that might be particularly helpful for younger patients:

Well, I did like that it shows the actual creatinine amount, like 0.44. [...] Iliked that it explained what that stage meant and stuff. And so the range and everything. I thought that was a really nice addition to help you understand, maybe, especially like I was saying for younger people or even people who have just recently got a transplant, helping them to understand what that means. [S2Y7, aged 16 years]

Compared with the traditional presentation in the clinic, participants liked that the layout made information more easily accessible. S2Y12 remarked that it was nice to have a refresher on the definitions and details, especially since there is less opportunity for that during clinic visits.

> *I liked the layout of being able to look with more detail at everything because it is so much like data. I just had a doctor's appointment last week, and they were going through the blood work thing. They don't go through every single thing anymore, but I'm still like, "What is that again?" I have kept asking. So it's nice to be able to have the details of what each panel looks like in the app. I think that's really helpful and useful.* [S2Y12, aged 21 years]

Participants frequently expressed confusion about information presented in the clinic, which reinforces our previous findings on existing information gaps experienced by youth and caregivers. In contrast, they found the app's presentation of information easier to understand and digest.

> *When the doctor's telling me something, sometimes they say words I don't understand and I'm confused. And sometimes they don't explain what it means. [...] Because I think I definitely understand it more than the things you see at doctors because I'm like, "Huh, what does that mean?" But on the app, I'm like, "Wait, I understand this.* [S2Y14, aged 14 years]

> *I think the whole kidney health part that I talked about earlier that I really liked, I didn't really ever get it broken down by stages, or at least I never really paid attention if they did break it down to me in stages.* [S2C17]

Many participants also expressed appreciation for the general attractiveness, usability, and educational aspects of the visual design. Participants also noted the design's general appeal to a wider range of audiences, especially for those who do not have a health care background, such as younger patients.

Summary—Reflections and Actionable Insights From the Kidney Bean Visual and Kidney Function Component

In study 2, youth and caregivers valued seeing kidney function laboratory results displayed over time, as it allowed them to track their own or their child's progress and reinforced the idea of kidney transplant as an ongoing journey. They also appreciated the simple, easy-to-understand design of the visual, especially in contrast to the confusion they often experienced with traditional clinical presentations of health information.

Insights From Studies 1 and 2

Ultimately, these findings highlight a critical design gap in the communication of CKD status. While themes 1 and 2 reveal the cognitive challenges and emotional resonance patients face when interpreting isolated laboratory values, theme 3 points to a possible design solution. From our findings, we can summarize that patients and caregivers require intuitive, longitudinal data visualizations that bridge the gap between abstract clinical metrics and the lived emotional experience of managing chronic illness while providing clear, actionable insights through accessible layouts.

XSL•FO
RenderX

Discussion

Overview

Our findings offer significant insights into patients' and caregivers' understanding of traditional health data presentation methods and inform the design of a novel visual that can help facilitate self-management through self-reflection and improve clinical interactions. To actively contribute to shared decision-making, patients and caregivers need access to relevant health information and the opportunity to discuss any emotional concerns it raises with their clinical team. Supporting these families in understanding kidney function laboratory results remains a critical area for further exploration, particularly in improving how clinical content is presented to them. Visualizations, as demonstrated by the kidney bean visual, can play a powerful role in simplifying complex data and prompting actionable insights that can empower patients in their self-management [12,13].

To this end, our findings identify that kidney transplant recipients and their caregivers often struggle to understand and respond to complex health information, such as kidney function values (eg, creatinine) limiting their ability to effectively engage in their care. Furthermore, kidney function laboratory values are more complex from the patient's and family's perspective, as they also closely align with their emotions. Keeping this in mind, we recognized that understanding the basic meaning of kidney function values may not be enough to interpret the information's significance. By resonating with and understanding the meaning of the results, this not only informs patients but also allows them to act on the information. This is especially critical for AYAs with chronic illnesses, as developing the ability to interpret and act on health information is a vital part of building self-management skills [26]. Supporting this understanding early helps AYAs gradually take on more responsibility for their care, which is essential for eventual successful transition to adult health care [27].

Therefore, to engage individuals in their health, it is critical that numbers and values are presented in a way that represents meaning for the person receiving them and can easily become actionable. For AYAs with kidney transplants, laboratory values like creatinine and GFR are abstract and can be difficult to interpret, especially amidst all the new information they are receiving posttransplant. However, when these numbers are represented through familiar forms, such as everyday objects or emotions, they become more relatable and meaningful. Ultimately, filling knowledge gaps for patients and their caregivers can help not only encourage participation in decision-making but also empower communication between providers and patients.

In the following sections, we explore how the following principles guided the design of our novel visualization, the kidney bean visual, to convey complex health information to facilitate and improve patient engagement: (1) using visual analogies to facilitate understanding, (2) presenting longitudinal data for progress checking, and (3) designing visualizations to improve patient-provider communication.

Using Visual Analogies to Facilitate Understanding

Traditionally, kidney function values (ie, creatinine values) are presented to patients and caregivers through conventional information visualizations (eg, line graphs). However, our findings from study 1 suggest that traditional line graphs did not fully meet the information needs of youth and their caregivers. This finding corresponds with prior investigations [28] that highlight patients living with a chronic condition struggle to interpret and use traditional graph visualizations. Participants had trouble understanding their health data through traditional visualizations because the visualizations did not reflect their circumstances, values, or emotions over time [28].

Instead, consistent with the literature [29], we found that health interfaces should prioritize nongraph visualizations, such as visual analogies. These types of visualizations simplify complex medical data and make abstract health concepts more concrete and relatable, especially for younger users. Visual analogies compare 2 concepts, 1 familiar and 1 unfamiliar, which allows the audience to use information to understand the unfamiliar concept through the lens of the more familiar concept [30]. Analogies can help humans construct mental models, which are an internal representation of the world and how it works, by basing them on concrete objects or familiar physical experiences [31,32]. The development of mental models is especially effective for youth, as it enables them to progressively refine and apply their models to more complex information over time [33,34].

Nongraph visualizations, specifically visual analogies, enhance patient comprehension of health information compared with text-only or graph visuals [29]. Incorporating familiar physical objects and color-coded cues in visualizations can enhance comprehension by linking what people already know with the features of a target concept, making the information more accessible and actionable [13].

The kidney bean visual was able to represent complex health information in a way that resonated with participants. It mirrored and validated their emotions, prompting reflection on their kidney function values within the context of their lived experiences. By being able to facilitate understanding, there is the potential for the kidney bean visual to align kidney function values with actionable self-management activities that can help build agency and self-efficacy. This is supported by prior work, which shows that personal visualizations can support self-reflection and better inform behavior change [35]. We observed this phenomenon in our participants' language and feedback, with the majority noting the kidney bean visual's motivational influences. Visualizations like these can be especially important for AYAs, as this developmental period is marked by increasing autonomy and formation of life skills, making it an important window for fostering the habits and health literacy needed to manage a chronic illness into adulthood.

The resonance participants reported with the emotion-integrated kidney bean visualization suggests that their laboratory values are not processed as a pure clinical metric but rather through a multidimensional emotional lens. This finding can be interpreted through the Cognitive Appraisal Theory, which posits that an

XSL•FO
RenderX

individual's emotional response is dictated by their primary appraisal of a stressor—in this case, interpreting a rising creatinine level as a "threat" to their kidney health [36]. By pairing abstract numerical data with recognizable facial expressions and color, the kidney bean visual is able to mirror a similar emotional state. Individuals can use this appraisal to prompt a secondary appraisal, a stage where the individual evaluates their options for coping with and responding to the "threat." Within this framework, the emotional states reflected by the bean can motivate individuals to translate their emotional resonance into commitment and motivation focused on maintaining healthy habits. In addition, we found that pairing explanations with visualizations in a simple and attractive interface can enhance understanding. An estimated 1 in 3 AYAs have poor health literacy, impacting their understanding of health information [37]. Additionally, prior work has found that laypersons are more positively influenced by the attractiveness of a website's design compared with medical professionals and have been found to reject high-quality content because of poor visual design and confusing displays [38]. Taken together, our results emphasize the need for an interface that presents health information visualizations along with concise explanations in a simple yet visually attractive manner.

To accompany the kidney bean visual, we incorporated multiple types of explanations, including definitions of common medical terms associated with kidney function (eg, creatinine and GFR), a list and descriptions of the 5 stages of CKD, and suggested healthy kidney habits (Figures 3A-D). Following guidelines by the Patient Safety Network on health literacy, the explanations were kept at a fourth- to sixth-grade level, used familiar language, and included relevant icons [39]. Keeping health literacy in mind, a simple yet attractive interface that incorporates links to definitions of terms, detailed explanations, and simple illustrations can help to increase patients' ability to understand their personal health information, leading to higher engagement [40].

While patients have limited control over changes in creatinine caused by factors like immune rejection or disease relapse, the kidney bean visual offers an opportunity to engage with their health data in a more proactive way. By facilitating a deeper understanding of their clinical trajectory, the visualization empowers youth to better prepare for their clinic visits and plan for future health changes, while reinforcing the value of behaviors in their control, such as medication adherence and healthy lifestyle choices.

Presenting Data Longitudinally to Track Progress

Previous research has demonstrated that self-tracking of personal health data over time can improve individuals' ability to build awareness of their behaviors, reflect upon them, and motivate behavior change [41,42]. Displaying data longitudinally not only helps patients visualize their health journey over time but also helps contextualize their laboratory results, spurring engagement [43]. Presenting clinical data longitudinally allows patients to place medical values in the context of their life history, which patients have found therapeutic and providers have found informative [44]. We found that most participants liked seeing their health data over time, as it allowed them to: (1) see their progress and trends throughout their medical journey, (2) contextualize and draw connections from their laboratory results to how they were feeling at the time, and (3) prompt action on how they can improve or maintain their results. In this way, presenting data longitudinally allowed for increased information accessibility and actionability, as participants were able to visualize their treatment plan and kidney care habits alongside their medical implications at different points in time.

Participants liked the benchmarking aspect of viewing their laboratory trends over time, as it was a reminder of how much they have progressed or changed. This design direction was built from previous work [45] where we identified that visualizations should support appropriate benchmarking to assess one's health status. Hence, presenting data in a way that compares youths' current and previous health status can foster more realistic and empowering self-assessments. This method of data presentation facilitates youth's reflections and recollections on both the challenges and successes that they have faced along their journey.

Reflective thinking about behavior is an established approach to motivating and sustaining change, as outlined in the Transtheoretical Model of Health Behavior Change [46]. Research suggests that reflecting on personal experiences can spark intrinsic motivation and sustain behavioral change [47]. By viewing their own longitudinal health data, participants engaged in consciousness raising—where individuals understand the potential consequences of their current or past behavior and the positive outcomes of change—which aligns with the self-evaluation process of behavioral change identified by the Transtheoretical Model of Health Behavior Change.

Additionally, presenting data longitudinally, unlike other visual frameworks, does not require information to be compartmentalized into distinct domains [44]. Instead, it allows for connections between specific events. This is useful, as it allows participants to identify the different factors and complex realities that influence their laboratory values. Being able to identify and connect these domains can help participants to recognize competing priorities in their own lives that may have impacted their kidney function. AYAs with kidney transplants must navigate maintaining their health alongside school, friendships, and greater independence, and having the ability to connect life events with changes in kidney function may foster greater self-awareness and support more informed, engaged self-management.

Designing Visualizations to Improve Patient-Provider Communication

A core component of shared decision-making is for patients and caregivers to share their goals and perspectives with their health care team [48]. We have highlighted how visual analogies and longitudinal data presentations help youth and caregivers fill information gaps, ultimately encouraging greater participation in their health care. Therefore, visualizations should have the dual function of (1) aiding in the understanding of complex information and (2) facilitating triadic communication between all stakeholders about patient and caregiver goals. Triadic communication [49] refers to the communication between health care professionals, youth patients, and their

XSL•FO
RenderX

caregivers. With chronic illness, there is an expectation for youth to assume more responsibility for their care with the goal of a gradual transition towards independent self-management. Ineffective patient-provider communication may reduce youth self-efficacy, adherence to medical treatments, and providers' assessments of youth's readiness to transition to adult care [50]. Active and effective triadic communication can improve patients' health outcomes and satisfaction and alleviate distress for children and their parents.

Extensive work has shown that providers tend to dominate medical interactions and primarily discuss biomedical information like laboratory values, neglecting the emotional aspects during the clinic visit [50,51]. Missed empathic opportunities refer to the phenomenon where physicians provide little emotional support in response to emotional needs and concerns raised by patients [52]. Visualizations like the kidney bean visual may help address this gap by allowing patients and caregivers to connect their kidney function laboratory values with how they feel physically or emotionally, or with specific events in their lived experiences. By attaching personal meaning to the numbers, patients are able to better recognize the significance of the data in ways that resonate personally. This creates an opportunity for deeper dialogue in clinical encounters, where providers can gain insight into the patients' and caregivers' emotional concerns, opening space for empathic engagement and encouraging more active involvement from both patients and caregivers. Designing visuals that support both emotional engagement and understanding of complex information can be a promising approach to facilitating triadic communication.

Limitations and Future Work

Our findings offer new insights into designing visualizations to convey complex health information to youth with kidney transplants and their caregivers. However, we recognize some limitations to this study. Our sample represents individuals from only 2 hospitals located in 2 large metropolitan areas in the United States and, therefore, may not represent findings from other geographic regions. The sample size is relatively small, considering the diversity and complexity of experiences among youth who have undergone kidney transplantation. The recruitment strategy using convenience sampling may have introduced selection bias, favoring highly motivated participants already interested in digital health management. Data regarding educational attainment, employment status, and health literacy levels were not collected as part of the demographic analysis, which limits our understanding of how these individual factors might have influenced engagement with the digital interface. Future research should use purposive or stratified sampling to ensure a wider range of patient motivation and literacy levels are represented.

A potential limitation is the risk of social desirability bias during the virtual interviews. We sought to mitigate this by using interviewers who were distinct from the development team. This separation was intended to reduce the perceived pressure on participants to endorse the prototype and to elicit more candid reflections on the visualization's utility.

In addition, while it is important and meaningful to see emotions in visualizations, this study also showed that it can be difficult and overwhelming for some individuals, depending on their current condition and progress. Therefore, we recommend that these visualizations be used alongside regular clinical discussions, where health care providers who are trained in communicating sensitive health information can help patients understand and process the data in a supportive environment.

As presented, this work is heavily influenced by the specific pathophysiology and clinical trajectory of youth with CKD. While there are parallels with other data-intensive chronic illnesses, the application of these findings to other chronic diseases would need to be further explored.

Future work will involve further validation of the kidney bean visual and its presentation interface, including piloting its impact on patient outcomes and exploring integration into standardized clinical workflows. Given the potential to enhance communication among youth, caregivers, and clinical teams, future research should also involve clinicians to examine how these visualizations can complement the delivery of complex health information in clinical settings. Additionally, building on these principles, future efforts could explore the design of similar visuals for other health data beyond creatinine, offering patients and caregivers alternative ways to understand and engage with their care.

Conclusion

Kidney function laboratory values, although seemingly straightforward in their intended purpose and use, are more complex from the patient's and family's perspective, as they also closely align with their emotions. This study demonstrates the value of designing visualizations that capture these complexities, which can create opportunities for emotional engagement within triadic communication and ultimately facilitate self-management. In this study, we present three guiding design principles: (1) using visual analogies, (2) presenting data longitudinally, and (3) designing visualizations to improve patient-provider communication, to better support the understanding of complex health information and enhance the ability of patients and caregivers to participate in their care. When information needs are met, opportunities arise for improved triadic communication, encouraging patient engagement, supporting self-management, and ultimately resulting in improved outcomes.

Acknowledgments

The authors would like to thank all participants and contributors for their support of this work. In addition, we would like to thank the Improving Renal Outcome Collaborative for supporting this work by providing the research team access to additional potential participants.

XSL•FO
RenderX

This content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Generative artificial intelligence (eg, Claude and Gemini) was used solely as a grammar and language editing tool to refine the clarity and flow of the original manuscript and played no role in the ideation, data analysis, or substantive drafting of the research.

Funding

Research reported in this publication was supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health under award number K23DK117017.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary figures.
[DOCX File , 312 KB - jmir_v28i1e80481_app1.docx]

References

1. Andreoni KA, Forbes R, Andreoni RM, Phillips G, Stewart H, Ferris M. Age-related kidney transplant outcomes: health disparities amplified in adolescence. JAMA Intern Med 2013;173(16):1524-1532. [doi: 10.1001/jamainternmed.2013.8495] [Medline: 23896628]
2. Kobus G, Małyszko J, Małyszko J, Puza E, Bachórzewska-Gajewska H, Myśliwiec M. Compliance with lifestyle recommendations in kidney allograft recipients. Transplant Proc 2011;43(8):2930-2934. [doi: 10.1016/j.transproceed.2011.08.031] [Medline: 21996192]
3. Lorig KR, Ritter P, Stewart AL, Sobel DS, Brown BW, Bandura A. Chronic disease self-management program: 2-year health status and health care utilization outcomes. Med Care 2001;39(11):1217-1223. [doi: 10.1097/00005650-200111000-00008] [Medline: 11606875]
4. Glasgow RE, Fisher L, Skaff M, Mullan J, Toobert DJ. Problem solving and diabetes self-management: investigation in a large, multiracial sample. Diabetes Care 2007;30(1):33-37. [doi: 10.2337/dc06-1390] [Medline: 17192329]
5. Lorig KR, Holman HR. Self-management education: history, definition, outcomes, and mechanisms. Ann Behav Med 2003;26(1):1-7. [doi: 10.1207/S15324796ABM2601_01] [Medline: 12867348]
6. Elwyn G, Laitner S, Coulter A, Walker E, Watson P, Thomson R. Implementing shared decision making in the NHS. BMJ 2010 Oct 14;341:c5146 [FREE Full text] [doi: 10.1136/bmj.c5146] [Medline: 20947577]
7. Joseph-Williams N, Elwyn G, Edwards A. Knowledge is not power for patients: a systematic review and thematic synthesis of patient-reported barriers and facilitators to shared decision making. Patient Educ Couns 2014;94(3):291-309. [doi: 10.1016/j.pec.2013.10.031] [Medline: 24305642]
8. Mishel MH. Reconceptualization of the uncertainty in illness theory. Image J Nurs Sch 1990;22(4):256-262. [doi: 10.1111/j.1547-5069.1990.tb00225.x] [Medline: 2292449]
9. Maikranz JM, Steele RG, Dreyer ML, Stratman AC, Bovaird JA. The relationship of hope and illness-related uncertainty to emotional adjustment and adherence among pediatric renal and liver transplant recipients. J Pediatr Psychol 2007;32(5):571-581. [doi: 10.1093/jpepsy/jsl046] [Medline: 17145733]
10. Gershon N, Eick SG, Card S. Information visualization. interactions 1998;5(2):9-15. [doi: 10.1145/274430.274432]
11. Card SK, Mackinlay J. The structure of the information visualization design space. 1997 Presented at: Proceedings of VIZ '97: Visualization Conference, Information Visualization Symposium and Parallel Rendering Symposium; October 18-25, 1997; Phoenix p. 92-99. [doi: 10.1109/infvis.1997.636792]
12. Lor M, Koleck T, Bakken S. Information visualizations of symptom information for patients and providers: a systematic review. J Am Med Inform Assoc 2019;26(2):162-171 [FREE Full text] [doi: 10.1093/jamia/ocy152] [Medline: 30535152]
13. Turchioe MR, Myers A, Isaac S, Baik D, Grossman LV, Ancker JS. A systematic review of patient-facing visualizations of personal health data. Appl Clin Inform 2019;10(4):751-770 [FREE Full text] [doi: 10.1055/s-0039-1697592] [Medline: 31597182]
14. van der Mee FAM, Schaper F, Jansen J, Bons JAP, Meex SJR, Cals JWL. Enhancing patient understanding of laboratory test results: systematic review of presentation formats and their impact on perception, decision, action, and memory. J Med Internet Res 2024;26:e53993 [FREE Full text] [doi: 10.2196/53993] [Medline: 39133906]
15. Improving Renal Outcomes Collaborative. URL: https://www.irocnow.org/ [accessed 2026-05-15]
16. Zoom [macOS]. Version 5.16.10. Zoom Video Communications, Inc. 2024. URL: https://zoom.us/download?os=mac [accessed 2026-05-15]
17. GFR. National Kidney Foundation. 2025. URL: https://www.kidney.org/kidney-health/kidneydisease/siemens_hcp_gfr [accessed 2026-05-15]

XSL•FO
RenderX

18. Figma: the collaborative interface design tool. Figma. 2024. URL: https://www.figma.com [accessed 2025-05-15]
19. MAXQDA 2022. 2023. URL: https://www.maxqda.com [accessed 2026-05-15]
20. Zhang Y, Wildemuth BM. Qualitative analysis of content. In: Applications of Social Research Methods to Questions in Information and Library Science. Westport: Libraries Unlimited; 2009:1-12.
21. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2008;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
22. Kawakita J. The Original KJ Method: A Practical Guide to Affinity Diagramming. Tokyo: Kawakita Research Institute; 1991.
23. Beyer H, Holtzblatt K. Contextual Design: Defining Customer-Centered Systems. San Francisco: Morgan Kaufmann Publishers Inc; 1997.
24. Lazar J, Feng JH, Hochheiser H. Research Methods in Human-Computer Interaction. 2nd Edition. Hoboken: Wiley Publishing; 2017.
25. FigJam. Figma. 2024. URL: https://www.figma.com/figjam/ [accessed 2026-05-15]
26. Poureslami I, Nimmon L, Rootman I, Fitzgerald MJ. Health literacy and chronic disease management: drawing from expert knowledge to set an agenda. Health promotion international 2017;32(4):743-754 [FREE Full text] [doi: 10.1093/heapro/daw003]
27. Allen T, Reda S, Martin S, Long P, Franklin A, Bedoya SZ. The needs of adolescents and young adults with chronic illness: results of a quality improvement survey. Children (Basel) 2022;9(4):500 [FREE Full text] [doi: 10.3390/children9040500] [Medline: 35455544]
28. Berry AB, Lim CY, Liang CA, Hartzler AL, Hirsch T, Ferguson DM. Supporting collaborative reflection on personal values and health. Proc ACM Hum Comput Interact 2021;5(CSCW2):1-39 [FREE Full text] [doi: 10.1145/3476040] [Medline: 36644216]
29. Reading Turchioe M, Grossman L, Myers A, Baik D, Goyal P, Masterson Creber RM. Visual analogies, not graphs, increase patients' comprehension of changes in their health status. J Am Med Inform Assoc 2020;27(5):677-689 [FREE Full text] [doi: 10.1093/jamia/ocz217] [Medline: 31999316]
30. Donnelly CM, McDaniel MA. Use of analogy in learning scientific concepts. J Exp Psychol Learn Mem Cogn 1993;19(4):975-987. [doi: 10.1037//0278-7393.19.4.975] [Medline: 8345330]
31. Ancker JS, Begg MD. Using visual analogies to teach introductory statistical concepts. Numeracy 2017;10(2):7. [doi: 10.5038/1936-4660.10.2.7]
32. Lakoff G, Johnson M. Metaphors We Live By. Chicago: University of Chicago Press; 1980.
33. Martin MA. “It's Like… You Know”: the use of analogies and heuristics in teaching introductory statistical methods. J Stat Educ 2003;11(2). [doi: 10.1080/10691898.2003.11910705]
34. Mayer R, Mathias A, Wetzell K. Fostering understanding of multimedia messages through pre-training: evidence for a two-stage theory of mental model construction. J Exp Psychol Appl 2002;8(3):147-154. [doi: 10.1037//1076-898x.8.3.147] [Medline: 12240927]
35. Huang D, Tory M, Aseniero BA, Bartram L, Bateman S, Carpendale S, et al. Personal visualization and personal visual analytics. IEEE Trans Vis Comput Graph 2015;21(3):420-433. [doi: 10.1109/TVCG.2014.2359887] [Medline: 26357073]
36. Folkman S, Lazarus RS, Dunkel-Schetter C, DeLongis A, Gruen RJ. Dynamics of a stressful encounter: cognitive appraisal, coping, and encounter outcomes. J Pers Soc Psychol 1986;50(5):992-1003. [doi: 10.1037//0022-3514.50.5.992] [Medline: 3712234]
37. Sansom-Daly UM, Lin M, Robertson EG, Wakefield CE, McGill BC, Girgis A. Health literacy in adolescents and young adults: an updated review. J Adolesc Young Adult Oncol 2016;5(2):106-118. [doi: 10.1089/jayao.2015.0059] [Medline: 26859721]
38. Bar-Lev S, Beimel D. Numbers, graphs and words - do we really understand the lab test results accessible via the patient portals? Isr J Health Policy Res 2020;9(1):58 [FREE Full text] [doi: 10.1186/s13584-020-00415-z] [Medline: 33115536]
39. Personal health literacy. Agency for Healthcare Research and Quality. 2019. URL: https://psnet.ahrq.gov/primer/personal-health-literacy [accessed 2025-04-10]
40. Irizarry T, DeVito Dabbs A, Curran CR. Patient portals and patient engagement: a state of the science review. J Med Internet Res 2015;17(6):e148 [FREE Full text] [doi: 10.2196/jmir.4255] [Medline: 26104044]
41. Choe EK, Lee NB, Lee B, Pratt W, Kientz JA. Understanding quantified-selfers' practices in collecting and exploring personal data. 2014 Presented at: CHI '14: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; April 26-May 1, 2014; Toronto p. 1143-1152. [doi: 10.1145/2556288.2557372]
42. Choe EK, Lee B, Zhu H, Riche NH, Baur D. Understanding self-reflection: how people reflect on personal data through visual data exploration. 2018 Presented at: PervasiveHealth '17: Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare; May 23-26, 2017; Barcelona p. 173-182 URL: https://doi.org/10.1145/3154862.3154881 [doi: 10.1145/3154862.3154881]
43. Dunbar JC, Bascom E, Pratt W, Snyder J, Smith JM, Pollack AH. My kidney identity: contextualizing pediatric patients and their families kidney transplant journeys. Pediatr Transplant 2022;26(7):e14343. [doi: 10.1111/petr.14343] [Medline: 35863916]

XSL•FO
RenderX

44. Monico LB, Ludwig A, Lertch E, Mitchell SG. Using timeline methodology to visualize treatment trajectories of youth and young adults following inpatient opioid treatment. Int J Qual Methods 2020;19:160940692097010. [doi: 10.1177/1609406920970106]
45. Jeffs L, Dunbar JC, Syed S, Ng C, Pollack AH. Navigating normalcy: designing personal health visualizations for pediatric kidney transplant recipients and caregivers. J Am Med Inform Assoc 2024;31(11):2519-2528. [doi: 10.1093/jamia/ocae206] [Medline: 39078283]
46. Prochaska JO, Velicer WF. The transtheoretical model of health behavior change. Am J Health Promot 1997;12(1):38-48. [doi: 10.4278/0890-1171-12.1.38] [Medline: 10170434]
47. Orji R, Lomotey R, Oyibo K, Orji F, Blustein J, Shahid S. Tracking feels oppressive and 'punishy': exploring the costs and benefits of self-monitoring for health and wellness. Digit Health 2018;4:2055207618797554. [doi: 10.1177/2055207618797554] [Medline: 30202544]
48. Chen M, Sepucha K, Bozic KJ, Jayakumar P. Value-based healthcare: integrating shared decision-making into clinical practice. Clin Orthop Relat Res 2023 Mar 01;481(3):448-450 [FREE Full text] [doi: 10.1097/CORR.0000000000002580] [Medline: 36735904]
49. Ye J, Yang L, Axelin A, Likitalo S, Wen C, Li X. The implementation and strategy of triadic communication in pediatric oncology: a scoping review. Pediatr Res 2025;97(6):1803-1822. [doi: 10.1038/s41390-024-03590-w] [Medline: 39304788]
50. Coburn SS, Eakin MN, Roter D, Pruette C, Brady T, Mendley S. Patient-provider communication in nephrology care for adolescents and young adults. Patient Educ Couns 2017;100(9):1680-1687 [FREE Full text] [doi: 10.1016/j.pec.2017.03.026] [Medline: 28433407]
51. Korsvold L, Mellblom AV, Lie HC, Ruud E, Loge JH, Finset A. Patient-provider communication about the emotional cues and concerns of adolescent and young adult patients and their family members when receiving a diagnosis of cancer. Patient Educ Couns 2016;99(10):1576-1583. [doi: 10.1016/j.pec.2016.03.028] [Medline: 27139576]
52. Morse DS, Edwardsen EA, Gordon HS. Missed opportunities for interval empathy in lung cancer communication. Arch Intern Med 2008;168(17):1853-1858 [FREE Full text] [doi: 10.1001/archinte.168.17.1853] [Medline: 18809811]

Abbreviations

AYA: adolescent and young adult
CKD: chronic kidney disease
GFR: glomerular filtration rate
HIPAA: Health Insurance Portability and Accountability Act
SCH: Seattle Children's Hospital
TREK: Thinking, Reflecting, and Empowering for Kidney Transplant Recipients

Edited by A Stone; submitted 17.Jul.2025; peer-reviewed by MC Fortin, H Patel; comments to author 11.Feb.2026; accepted 12.May.2026; published 15.Jun.2026.

Please cite as:
Ng C, Dunbar JC, Jeffs L, Syed S, Pan N, Ryu H, Smith J, Snyder J, Pratt W, Pollack A
Designing Emotion-Integrated Visualizations for Kidney Function Self-Management: User-Centered Design Study With Youth Transplant Recipients and Caregivers
J Med Internet Res 2026;28:e80481
URL: https://www.jmir.org/2026/1/e80481
doi:10.2196/80481
PMID:

©Chelsea Ng, Julia C Dunbar, Lily Jeffs, Sanaa Syed, Nathan Pan, Hyeyoung Ryu, Jodi Smith, Jaime Snyder, Wanda Pratt, Ari Pollack. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

High Patient Willingness to Grant Broad Consent for Real-World Data Use in Rheumatology—Implications for Real-World Data Platform Governance: Cross-Sectional Study

Jutta G Richter[1,2], Prof Dr med; Antonia Becker[1,2]; Tim Filla[1,2], Dr sc hum; Hasan Acar[1,2], DipInfoSc; Waldemar Ockert[3], BSc, PhD; Dominykas Kriauciunas[3], LLB; Edith Aggarwal[3], MA, MBA; Markus Schröder[4], MSc; Ralf Hansen[4], Dipl-Ing; Jörg HW Distler[1,2], Prof Dr med; Matthias Schneider[1,2], Prof Dr med

[1]Department for Rheumatology, University Hospital Düsseldorf, Medical Faculty of Heinrich-Heine-University, Moorenstrasse 5, Düsseldorf, Germany

[2]Hiller Research Center, University Hospital Düsseldorf, Medical Faculty of Heinrich-Heine-University, Düsseldorf, Germany

[3]ZS Inc, London, United Kingdom

[4]Serrala Group GmbH, Berlin, Germany

Corresponding Author:
Jutta G Richter, Prof Dr med
Department for Rheumatology, University Hospital Düsseldorf, Medical Faculty of Heinrich-Heine-University, Moorenstrasse 5, Düsseldorf, Germany

Abstract

Background: Medical real-world data (RWD) are often siloed across organizations, making them inaccessible for research. Unlocking these data could advance clinical research and patient care. The pan-European Data Nexus platform (DNP) links RWD, facilitating its use, for example by artificial intelligence (AI) tools, to support the generation of real-world evidence. In Europe, particularly Germany, the secondary use of health data is governed by stringent regulatory requirements, including informed consent.

Objective: This study evaluated the informed broad consent form for the RWD (Data Nexus) platform, predicated on the principles of the Medical Informatics Initiative in Germany, and contextualized its implications for future data governance and regulatory use of RWD.

Methods: The broad consent form was developed for the DNP and cross-sectionally distributed to consecutive rheumatology outpatients during routine follow-up at a tertiary center. Analyses included rates of agreement to the predefined broad consent items. A zero-inflated model (using R) was used to predict response rates.

Results: From July 2023 to May 2024, 74.9% (292/390) of the patients signed the broad consent form and consented to DNP data donation. Median age was 56 (IQR 43.0-65.0) years, 72.5% (211/291) were female, and median disease duration was 12 (IQR 4.0-22.0) years. Diagnoses included rheumatoid arthritis (96/291, 33.3%), psoriatic arthritis (30/291, 10.3%), spondyloarthritis (15/291, 5.2%), systemic lupus erythematosus (90/291, 30.9%), systemic sclerosis (14/291, 4.8%), and other conditions (16/291, 5.5%). Patients also answered 8 yes/no broad consent items, with an average of “yes” and “no” responses of 7.5 (SD 1.5) and 0.2 (SD 0.5), respectively. Missing responses averaged 0.4 (SD 1.4). Of all participants, 78.4% (228/291) agreed to all broad consent items. Approval rates for individual items exceeded 86%, indicating strong patient acceptance of secondary use of RWD under a structured governance framework, possibly reflecting trust, perceived benefit, low perceived risk, and governance confidence. Importantly, patients agreed to new techniques such as AI-based analysis of their donated RWD and, despite the social and ethical sensitivity, data distribution to third parties, including commercial industry. Consent rates were also high for the use of valuable omics and genomics data from biomaterial donations. Women showed higher consent rates, whereas educational attainment was not an indicator of response behavior.

Conclusions: This is the first study assessing the willingness of patients with inflammatory rheumatic diseases to grant broad consent for secondary data use in an innovative RWD platform, implementing a modern framework in routine care and considering detailed patient preferences. Patients demonstrated high willingness to grant broad consent, providing key real-world evidence on patients’ actual consent behavior for implementing RWD platform integration—such as the European Health Data Space. The DNP supports scalable, General Data Protection Regulation–compliant data sharing, enabling real-world and AI-driven research while preserving patient trust.

Trial Registration: German Clinical Trials Register DRKS00032801; https://drks.de/search/de/trial/DRKS00032801

(J Med Internet Res 2026;28:e80347) doi:10.2196/80347

XSL•FO
RenderX

KEYWORDS

informed broad consent; secondary use; real-world data; governance; rheumatology; artificial intelligence; AI

Introduction

Health care data captured during routine clinical care are typically scattered across various institutions and their systems. Real-world data (RWD; eg, from clinics, private practices, and patients) encompass information from experiences and observations of individuals, organizations, and systems. Unlike randomized controlled trials (RCTs), which have strict inclusion and exclusion criteria, RWD reflect the complexities and nuances of real life, offering rich medical insights [1].

Traditionally, RWD sources (eg, claims data, electronic medical records [EMRs], and registries) have been used for epidemiological and health economics research. Their research utility can be enhanced through cross-source linkages for holistic views of patients' health and experiences. The rigorous level of data control in RCTs makes them the gold standard for effectiveness evidence to establish treatment efficacy and safety, and RWD acceptance for these purposes has been limited [2-5]. However, RWD play a key role in closing information gaps from RCTs to generate real-world evidence (RWE) on outcomes in diverse populations receiving routine care [6]. RWD are increasingly regarded as a critical resource (eg, for clinical research, artificial intelligence [AI]–enabled research, regulatory decision-making, and health technology assessment [HTA]), and their use is gaining acceptance in research and regulatory decision-making and generating major interest in progressing frameworks [7-10].

The management of RWD within modern IT infrastructure and techniques is evolving. Collecting and analyzing RWD from various sources currently not fully leveraged for secondary research purposes requires an adequate, robust IT infrastructure and advanced methods such as AI [1,10]. The emergence of European data spaces, specifically, secure processing environments (SPEs; virtual environments for secure data storage, management, and analysis), highlights the importance of leveraging RWD with robust governance [11,12]. Regulatory bodies and networks aim to harmonize the collection of high-quality RWD and reliable RWE generation (eg, for HTA documents [10,13-15]). RWD and RWE may influence and improve national and international health policy [16]. Successful examples of RWD for rheumatology research include the EMR-enabled Rheumatology Informatics System for Effectiveness registry and the PremiOM psoriatic arthritis (PsA) dataset [17,18]. The Rheumatology Informatics System for Effectiveness automatically extracts structured data from routine clinical care for research and quality improvement, and PremiOM combines claims and EMR data at the individual level to study real-world treatment patterns [17,18].

In Europe, initiatives such as the European Health Data Space (EHDS) aim to facilitate the secondary use of health data across borders, including for research and innovation purposes [19,20]. Within this evolving landscape, broad consent has emerged as a key ethical and legal mechanism to enable the future use of RWD and biospecimens, specifically their related data for RWE generation via emerging AI techniques [21]. However, broad consent's acceptance among patients and the factors influencing consent decisions remain important considerations, particularly in light of ongoing debates on data governance, privacy, and trust.

Patient consent is fundamental to respecting patient autonomy and sustaining trust in health care and health-related research [22]. Moreover, both German legislation and the European General Data Protection Regulation (GDPR) mandate informed consent for the secondary use and reuse of RWD. As RWD-driven research infrastructures and SPEs gain strategic importance for data-driven health care and AI research, empirical evidence on patients' willingness to share health data remains limited, particularly among individuals with inflammatory rheumatic diseases (IRDs) [23,24].

To address this gap, this study is the first to systematically evaluate the willingness of patients with inflammatory rheumatic diseases (IRDs) to grant broad consent for secondary data use within an SPE. The primary objective is to assess the feasibility and acceptance of collecting and reusing patient-generated RWD within the pan-European Data Nexus platform (DNP) while deriving implications for the implementation, governance, and scalability of emerging European SPE initiatives, including the EHDS.

Methods

Overview of the Platform and Study Design

The DNP supports medical research and health care by integrating RWD in a secure environment. The DNP aims to optimize early detection, treatment, and prevention of diseases through insights from RWD research. Figure 1 depicts the DNP's scope and remits. The consortium implemented a comprehensive data management framework and a use and access committee to oversee authorized, eligible data requests and analyses pursuant to the broad consent forms.

XSL•FO
RenderX

Figure 1. Outline of the Data Nexus architecture depicting data flow and processing.

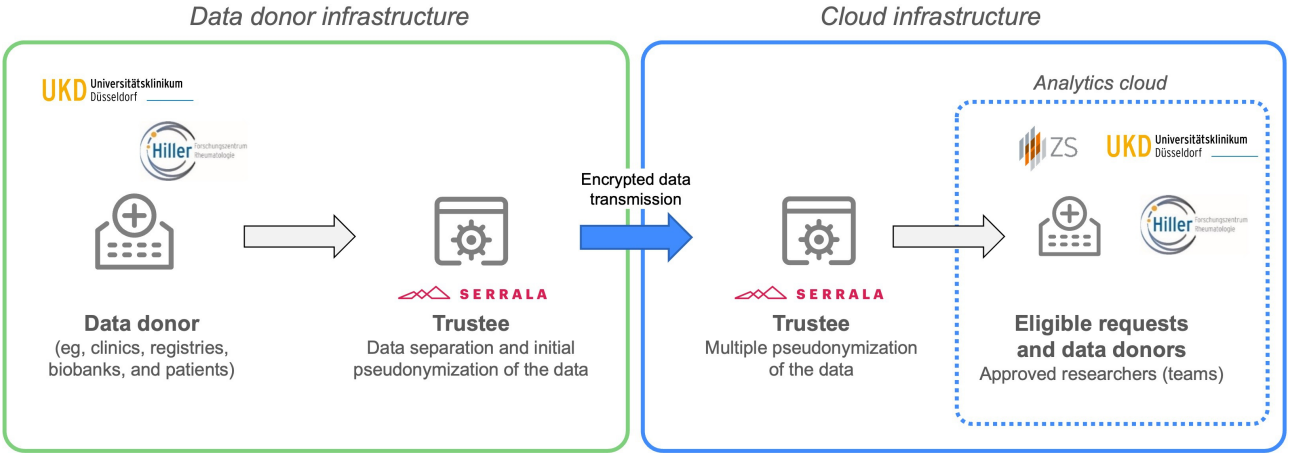

To enable research within the DNP, the informed broad consent form of the German Medical Informatics Initiative (MII) was adapted [25]. The broad consent form, including additions made by us, was approved by the responsible data protection officer and ethics committee. Details on the items' wording are outlined in the Results section and depicted in Figure 2. Patients consented to the transfer of existing data (including clinical and sociodemographic data) and prospectively generated data (eg, disease information, treatment, and patient-reported outcomes from routine care after consent) for the following 5 years. After 5 years, individual broad consent must be renewed and processed accordingly. Withdrawal procedures were described in the patient information section of the broad consent form.

Figure 2. Graphical representation of the 8 Data Nexus platform broad consent items.

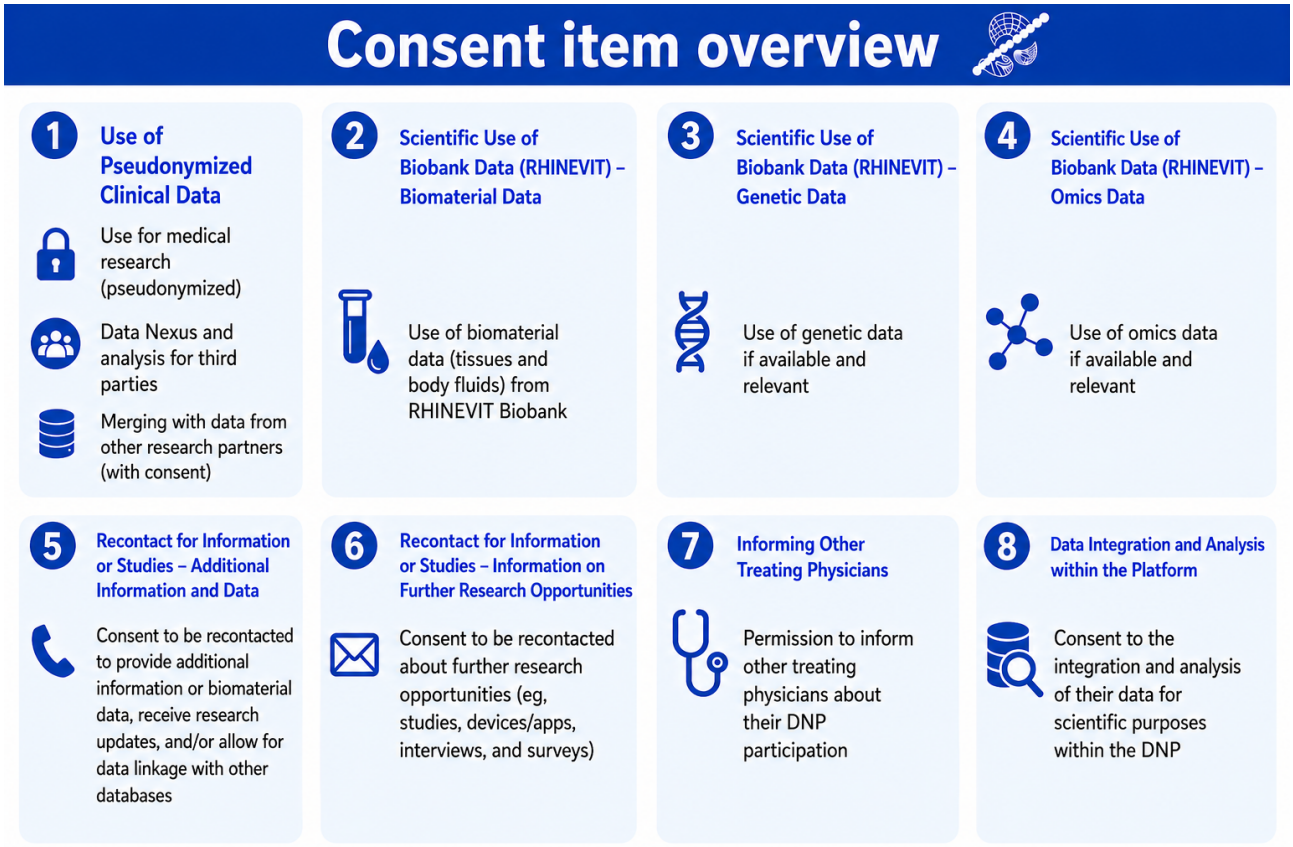

Patients could agree to or decline 8 individual broad consent items with either a "yes" or "no" response. If no choice was made, the answer was classified as missing, handled as "not given" in the DNP. Data were processed only for signed broad consent forms and solely according to the items that indicated "yes."

RWD parameters originated from our patient documentation system DocuMed.rh, which also holds data from the local electronic health record (EHR) [26]. According to patients' broad consent, biomaterial data (including genetic and omics data) could be used for research from the clinic's biobank RHINEVIT [27]. Data from office-based physician practices or other resources can be integrated into the DNP.

XSL•FO
RenderX

Study Center

We piloted this platform and the broad consent form in a tertiary rheumatology center at a German university clinic to evaluate the feasibility of collecting patient data using a platform such as the DNP. To this end, we systematically assessed and reported our observations with the aim to demonstrate strong patient acceptance of broad consent, which is highly relevant for the future of RWE generation and data-driven research infrastructures.

Patient Characteristics and Inclusion and Exclusion Criteria

Patients with IRDs coded in the *International Classification of Diseases, 10th Revision*, as rheumatoid arthritis (RA), PsA, spondyloarthritis (SpA), systemic lupus erythematosus (SLE), systemic sclerosis, vasculitis, and others above the age of 18 years were included randomly. Adequate German-language proficiency was considered a prerequisite. Exclusion criteria comprised age of less than 18 years and insufficient language proficiency to ensure adequate comprehension of broad consent and well-informed consent. Thus, we included 291 patients (n=211, 72.5% female).

Data Collection Processes

Consecutive patients with IRDs attending routine outpatient follow-ups in our rheumatology clinic received paper-based information about the DNP and the broad consent document during in-person visits. The 8 broad consent items were always listed in the same order, each requiring a “yes” or “no” response, although they could be left unanswered. A DNP team member from the rheumatology department explained the DNP face-to-face as in-person interaction is considered most effective for improving participants’ understanding of consent forms [28]. Afterward, patients read the patient information and provided responses in the waiting or treatment room. The DNP team member remained available for content-related or technical questions. If consent was given, 1 signed copy of the broad consent form was returned to the DNP team, and another was handed out for the patients’ personal records.

Signed paper-based forms were entered manually into a database. To minimize data entry errors, 2-pass verification was performed. Digitized broad consent forms were uploaded to the DNP for informed consent–compliant data handling within the DNP. Paper-based forms were stored securely in the clinic following German medicolegal requirements.

For quality assurance purposes, data from the broad consent forms and related (clinical) data from the period from July 2023 to May 2024 were extracted. We analyzed the consent rates of the signed broad consent forms for each of the 8 broad consent items.

Measures and Covariates

Patient acceptance of the predefined 8-item broad consent form was measured by analyzing responses to each consent item, allowing for both overall acceptance and item-specific variation. Eight predefined broad consent items could be answered with “yes” or “no” or left unanswered. For data quality assurance, we extracted age, sex, clinical data, and patients’ educational level, assessed in our patient documentation system, where the data are documented in alignment with the standards of the national database [26,29]. The broad consent form allows for the analysis of the data. Due to resource constraints, the sample size was limited, and a qualitative approach was not followed.

Ethical Considerations

The DNP broad consent form received a positive vote from a local independent ethics committee (ethics committee of the Medical Faculty of the Heinrich-Heine-University Duesseldorf, 2022-1958). The DNP is registered in the German Clinical Trials Register (DRKS00032801). This study was conducted in accordance with ethical principles and applicable institutional and national regulations. Participants provided written informed consent. No financial or other compensation was provided to patients for study participation. All procedures were carried out in accordance with the Declaration of Helsinki. To safeguard the privacy and confidentiality of human participants, all data are presented in an aggregated and nonidentifiable format. No identification of individual participants in the manuscript or supplementary material is possible.

Sample Size and Statistical Analyses

Due to the exploratory design of this study, no sample size calculation was performed. Descriptive statistics are provided as absolute numbers and percentages for discrete variables and as medians and IQRs for continuous variables. For statistical analysis of multi-class variables, a chi-square test was used. In addition, one-way ANOVA was calculated where appropriate according to the variable level. It was assumed that there were 2 types of patients: one group agreeing to all items without reading them in too much detail and a second group of patients who checked each item or at least some items individually on a more detailed level. The chosen zero-inflated model captures a continuum of patient behavior between these 2 extremes. To determine which factors predict patients’ agreement with the broad consent items, 2 zero-inflated models were calculated using age, sex, educational level, disease, and disease duration as the independent variables and the number of nonconsented or missing items as the dependent variable. Imputation of missing values was not conducted because missing values were regarded as 1 of 3 possible item outcomes (“yes,” “no,” and left unanswered). The zero-inflated model accounts for heterogeneity in response patterns, including the possibility that a subset of respondents will exhibit consistently high acceptance across items, which may reflect differences in engagement, decision strategies, or response behavior. Furthermore, a sensitivity analysis for the risk factor analysis of patients’ consent was conducted using a logistic regression approach with a binary outcome of “consented all” (yes or no) as the dependent variable and the same set of independent variables as for the zero-inflated model. A Little missing completely at random test was performed. A P value of less than .05 was considered statistically significant. All statistical computations were conducted using R (version 4.3.1; R Foundation for Statistical Computing). Data were analyzed anonymously.

XSL•FO
RenderX

Results

Overview

Of 390 consecutive outpatients receiving the broad consent form between July 2023 and May 2024, a total of 292 (74.9%) returned it signed. Thus, they provided voluntary broad consent for a period of 5 years from the date of consent and consented to the extended data use beyond this period (30 years, as outlined in the patient information). Refusal was not attributable to the content of the patient information or broad consent form as none of the individuals at the clinic had reviewed the document. Documented reasons for denial included time constraints, general lack of interest in research participation, and failure to return the document after taking it home. Of the 292 patients who signed the broad consent form, 1 (0.3%) male patient withdrew his consent 5 months after signing. Thus, further analyses refer to 291 patients. A total of 1.5% (6/390) of the broad consent forms were returned unsigned and, therefore, excluded from analysis; 2 of these had selected “yes” for all broad consent items. Participant flow is shown in Figure 3.

Figure 3. Participant flowchart. BC: broad consent.

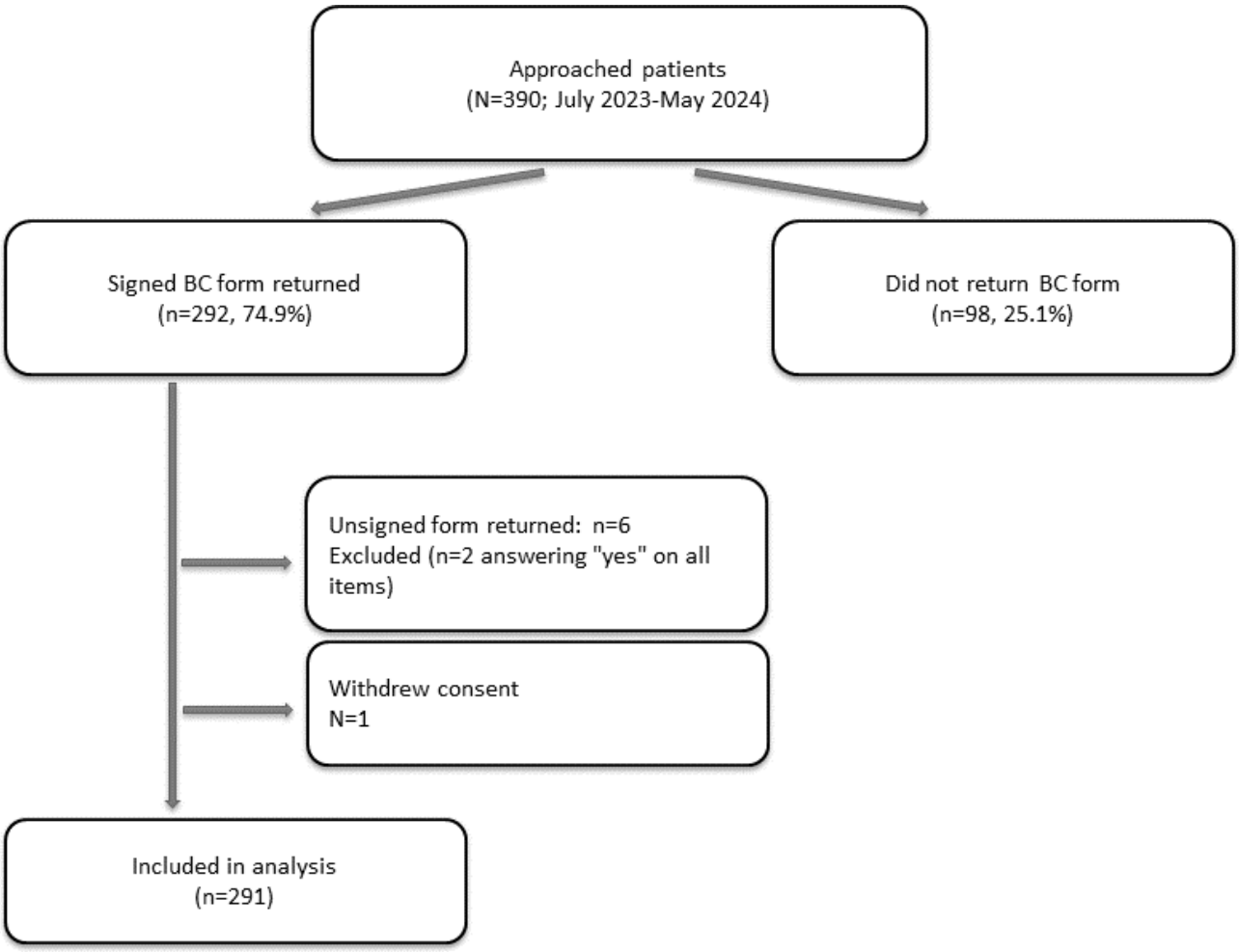

Content-related or technical questions while or after reading the patient information and the broad consent form were scarce (2/291, <1%) and questions from others related exclusively to which hard copies could be kept and which must be returned. Team members spent up to 15 minutes talking one-on-one to individual outpatients.

Participants’ median age was 56 (IQR 43.0-65.0) years, and most were female (211/291, 72.5%). Median disease duration was 12 (IQR 4.0-22.0) years. Further clinical and sociodemographic characteristics of the cohort are shown in Table 1.

XSL•FO
RenderX

Table . Clinical and sociodemographic characteristics of the complete group and distinct disease subgroups (N=291).

	All	RA[a] (n=96)	PsA[b] (n=30)	SpA[c] (n=15)	SLE[d] (n=90)	SSc[e] (n=14)	Vasculitis (n=30)	Others (n=16)
Female sex, n (%)	211 (72.5)	63 (65.6)	17 (56.7)	8 (53.3)	80 (88.9)	12 (85.7)	21 (70)	10 (62.5)
Age (y), median (IQR)	56 (43.0-65.0)	58 (46.3-66.0)	53 (30.8-64.3)	58 (51.0-61.0)	49 (40.8-58.0)	52 (44.8-59.8)	66 (57.0-78.0)	60 (43.5-66.8)
Disease duration (y), median (IQR)	12 (4.0-22.0)	9.0 (3.0-17.5)	9.0 (1.8-17.8)	9.0 (4.0-18.0)	21.0 (13.0-27.0)	8.5 (1.0-15.5)	6.0 (2.0-12.3)	10.0 (1.3-20.0)
University entrance diploma,[f] n (%)	114 (48.5)	34 (45.3)	12 (54.5)	5 (41.7)	42 (51.9)	6 (60.0)	11 (44.0)	4 (36.4)

[a]RA: rheumatoid arthritis.

[b]PsA: psoriatic arthritis.

[c]SpA: spondyloarthritis.

[d]SLE: systemic lupus erythematosus.

[e]SSc: systemic sclerosis.

[f]In Germany, this is a secondary school diploma that permits entrance to university. Not all patients reported their educational level, so the percentages may not match the denominators in the column headings.

The average number of "yes," "no," and missing responses was 7.5 (SD 1.5; range 0-8), 0.2 (SD 0.5; range 0-3), and 0.4 (SD 1.4; range 0-8), respectively. No relevant differences in the number of "yes," "no," and missing responses were noted between disease or educational level groups. Female patients gave statistically significantly more "yes" responses compared to male patients (mean 7.6, SD 1.2 vs mean 7.1, SD 1.9; *P*=.008) and had significantly fewer missing responses (mean 0.3, SD 1.1 vs mean 0.7, SD 1.9; *P*=.01). "No" responses were similar between female and male patients (mean 0.2, SD 0.5 for both; *P*=.45). A total of 2.1% (6/291) of the patients with signed the broad consent forms did not select any of the 8 broad consent options and showed similar characteristics to those of the overall population (50% female; median age 56 (IQR 43-65) years; median disease duration 12 (IQR 4-22) years; 40% with a secondary school diploma granting university entrance; n=2, 33.3% with RA; n=1, 16.7% with PsA; n=1, 16.7% with SpA; n=1, 16.7% with SLE; and n=1, 16.7% with other diseases).

Overall, approval rates for the 8 broad consent items exceeded 86% (Figure 4). Detailed results for each item are presented below.

Figure 4. Response rates ("yes," "no," and missing) for the 8 broad consent (BC) items.

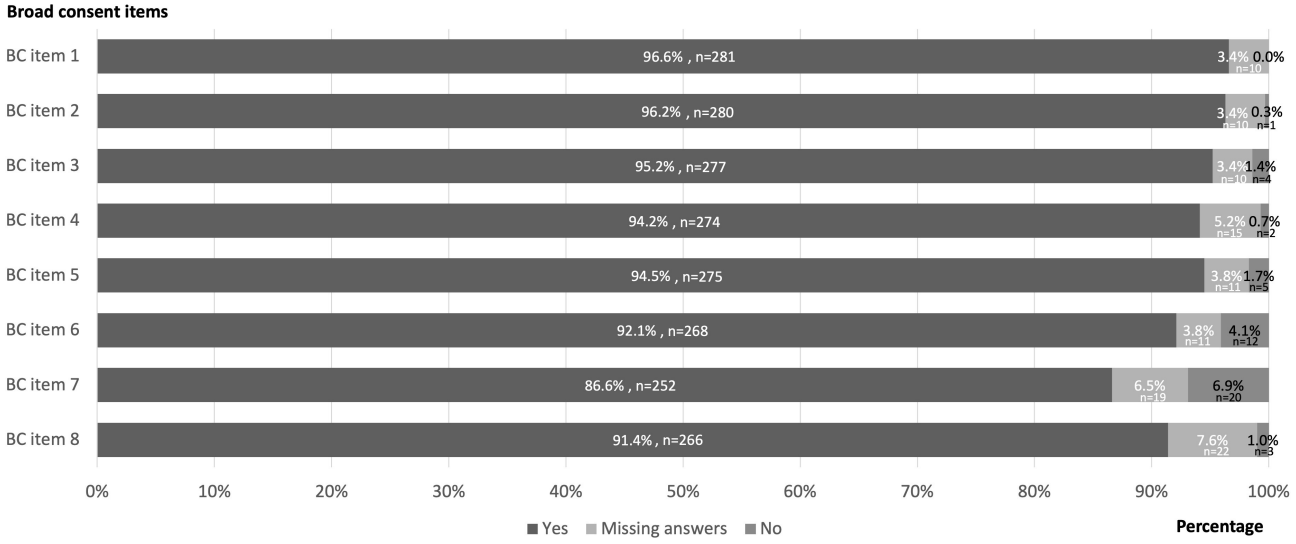

DNP Broad Consent Item 1

Item 1 covered 3 domains. First, it asked for permission to process and use patient data for medical research as described in the patient information, ensuring multi-pseudonymization. Data would be derived from our rheumatology-specific patient documentation system. Second, it covered scientific analysis and use of the data by the DNP and, thus, by extension, for third parties (eg, universities, institutes, and research companies [pharmaceutical companies and medical device manufacturers]). This may also include remote access to data for research projects conducted abroad in places where European data protection law applies. Patients were informed that they would not benefit commercially. Third, it asked for permission to merge their

XSL•FO
RenderX

patient data with other research partners' data if the patients agreed to this use by those partners. A total of 96.6% (281/291) of the patients consented to this option, none declined, and 3.4% (10/291) gave no answer.

DNP Broad Consent Item 2

Patients were asked whether biomaterial-linked data from the clinic-owned RHINEVIT biobank could be transferred and used within the DNP if available from prior donations. A total of 3.4% (10/291) of the patients gave no answer, 96.2% (280/291) consented, and 0.3% (1/291) declined.

DNP Broad Consent Item 3

If genetic data were available from the clinic's biobank RHINEVIT, specifically, our patient documentation system DocuMed.rh, and if these were relevant for their IRD, the patients' comorbidities, and research, 95.2% (277/291) of the patients consented to their scientific use as outlined in the patient information. In total, 1.4% (4/291) of the patients declined, and 3.4% (10/291) did not answer.

DNP Broad Consent Item 4

Correspondingly, if omics data were available that were related to the IRD, specifically, associated comorbidities, 94.2% (274/291) of the patients agreed to their scientific use, whereas 5.2% (15/291) provided no answer, and 0.7% (2/291) declined.

DNP Broad Consent Item 5

Consent was given for contacting the patients to obtain additional information, use additional RHINEVIT biomaterial-related data, provide updates and information on research projects, and link patient data with other medical databases by 94.5% (275/291) of the patients, whereas 1.7% (5/291) declined, and 3.8% (11/291) provided no answer.

DNP Broad Consent Item 6

Similarly, 92.1% (268/291) of the patients consented to being contacted about future research opportunities that may be relevant to them, 4.1% (12/291) declined, and 3.8% (11/291) did not answer. Common reasons for conducting research included but were not limited to assisting in the research of new medical devices and digital health applications, learning more about clinical trials and participating in trials relevant to them, and engaging in DNP-related research (eg, interviews and surveys).

DNP Broad Consent Item 7

Permission to inform other treating physicians about their DNP participation was granted by 86.6% (252/291) of the patients. Interestingly, 6.9% (20/291) declined, and 6.5% (19/291) did not answer. The decliners were mostly female (14/20, 70%), with a median age of 45 years and a median disease duration of 12 years; 50% (10/20) held a university entrance diploma. The diseases experienced by those who declined included RA (3/20, 15%), PsA (4/20, 20%), SpA (2/20, 10%), SLE (8/20, 40%), vasculitis (1/20, 5%), and others (2/20, 10%).

DNP Broad Consent Item 8

Most patients (266/291, 91.4%) consented to the integration and analysis of their data for scientific purposes within the DNP, whereas 7.6% (22/291) did not answer. A total of 1.0% (3/291) of the patients declined (66.7% female; median age 45 years; median disease duration 12 years; all university entrance diploma holders; and 1 each with a diagnosis of RA, SpA, and SLE). Both of the female patients who declined also declined to inform other treating physicians about their DNP participation.

Information Material Left Behind at the Clinic

A minority of patients (33/291, 11.3%) left the information material behind at the clinic; 88.7% (258/291) took the material with them. A lower percentage of women (19/211, 9%) than men (14/80, 17.5%) left the material behind. Those who left it were younger (median age 51 vs. 56 years of age) and had a longer disease duration (median 15.0 vs. 12.0 years). Patients of all disease groups (range 1/291, 3% [SpA] to 11/291, 33.3% [RA]) left the material at the clinic. The educational level of these patients was similar: 48% (12/25) of those who left the material at the clinic and 48.6% (102/210) of those who took it with them had a university entrance diploma. Missing data for educational level taken from RWD comprised 19.2% (26/291).

Results From the Zero-Inflated Model

We investigated a zero-inflated model to detect which variables predicted patients' agreement to the broad consent form items. The model could detect sex (female) but no other determinants of response behavior (Table 2). As already shown above, female patients had a much smaller number of missing items, clearly pointing to a systematic pattern rather than a random one, which can be regarded as a proxy for better reading of the broad consent form compared to male patients. The Little missing completely at random test was not significant (χ^2_{81}=92.4; P=.18), indicating that missing data were likely missing completely at random.

XSL•FO
RenderX

Table . Results of the zero-inflated model.

Predictor	Number of "no" responses		Number of missing responses	
	Incidence rate or odds ratio (95% CI)	*P* value	Incidence rate or odds ratio (95% CI)	*P* value
Age (y)	1.01 (0.97 - 1.05)[a]	.67	0.99 (0.97 - 1.00)[a]	.15
Disease duration (y)	0.99 (0.94 - 1.03)[a]	.51	0.99 (0.96 - 1.02)[a]	.38
Educational level	1.17 (0.33 - 4.10)[a]	.81	0.89 (0.53 - 1.47)[a]	.64
Gender (female)	2.07 (0.54 - 7.89)[a]	.29	1.12 (0.64 - 1.94)[a]	.70
Zero-inflated model				
Age (y)	1.02 (0.97 - 1.07)[b]	.40	0.98 (0.95 - 1.01)[b]	.11
Disease duration (y)	0.98 (0.91 - 1.04)[b]	.46	0.99 (0.96 - 1.03)[b]	.70
Educational level	1.06 (0.26 - 4.36)[b]	.93	0.86 (0.38 - 1.94)[b]	.72
Gender (female)	4.58 (0.59 - 35.57)[b]	.15	2.69 (1.15 - 6.32)[b]	.02

[a]Incidence rates.

[b]Odds ratios.

Discussion

Principal Findings

This study represents the first investigation in rheumatology focusing on broad consent for an innovative, centralized, pan-European, cross-border research platform integrating RWD from multiple sources. With broad consent, the DNP provides secure access to multi-pseudonymized RWD for approved scientific researchers and Data Nexus analysts (eg, biostatisticians and epidemiologists) on behalf of pharmaceutical clients and medical device manufacturers. Importantly, our findings should be interpreted in the broader context of evolving data governance and regulatory science in Europe. While high consent rates demonstrate feasibility at the patient level, their full significance lies in enabling scalable infrastructures for secondary use of RWD in regulatory decision-making, HTA, and cross-border research initiatives such as the EHDS [10,13-15,20,21]. In this context, granular broad consent models such as the one implemented in the DNP may represent a foundational governance layer for interoperable, ethically compliant RWD ecosystems, bridging clinical care and secondary use across jurisdictions. Our data are unique to rheumatology and IRDs and can help address key knowledge gaps and support patient readiness to share data for complex use.

Our cohort reflects a broad IRD population, including various inflammatory diseases. A median disease duration of 12 (IQR 4.0-22.0) years indicated long-term disease involvement. Sex distribution aligned with IRD patterns, although SpA showed a slight female predominance in a small subgroup.

We observed an unexpectedly high approval rate despite no concrete benefits or incentives. This is comparable to approval rate data in oncology and higher than the approval rate observed in MII broad consent form data from an emergency department [30,31]. Recently, varying approval rates have been reported from university clinics [21]. The broad consent aspects were well accepted by our patients, with high consent rates to individual items and most consenting to all items. Similar to our previous biobank broad consent analyses, not all patients answered every item [27]. Unlike in those previous analyses, women in this study left fewer items unanswered, suggesting better broad consent form reading and greater willingness to donate data to the DNP [27]. As reported by others focusing on data donation from EHRs in Germany, educational attainment was not an indicator of response behavior in this study [24]. This contrasts with our own research on voluntary broad consent related to biomaterial donation, where higher-educated patients responded negatively to some broad consent items more often [27]. While age-dependent data donation willingness has been reported, age did not influence our patients' consent rates [32].

In a representative German online population, positive attitudes toward data donation were influenced by clear distinctions between public and private research and strong trust in data protection and control [23]. Our broad consent form's first item, which combines public and private research, achieved very high agreement rates. This indicates that patients are willing to donate data for scientific and clinical research purposes, and their willingness is not compromised when research is conducted on behalf of third parties. A recent systematic review confirmed a positive public attitude toward sharing personal health data for third-party or secondary use, albeit under certain conditions [33]. However, willingness to donate self-generated health data (eg, from wearables) for secondary use can vary by disease and country [34,35]. Our data align with those reported for German patients with cancer, where 73.9% supported RWD use by researchers in commercial companies [22]. A Bitkom (Germany's digital association) survey also showed that a large majority of German citizens (90%) would share personal health data for research, including by private companies [36].

In this regard, the inclusion of commercial research within our broad consent frameworks requires particular (ethical) attention. While acceptance rates were high, transparency regarding the

XSL•FO
RenderX

role of industry partners, data use pathways, and safeguards against misuse is essential to sustain patients' and data providers' trust. Public research institutions generally elicit a high level of trust in data use, whereas skepticism remains toward data use by private sector entities, particularly with regard to commercial motives and data security [22,32,37-39]. These aspects are increasingly central in data governance debates, especially in light of European regulatory developments such as the EHDS, which explicitly need to balance innovation, secondary use, public trust in health data ecosystems, and the use of AI tools such as large language models [19,20,24,40].

Research revealed that data donation via an opt-out regulation does not necessarily mean that the paradigm of informed consent has to be discontinued [23]. Instead, public education on data-intensive medical research and improved public "health data literacy" are needed [23]. Richter et al [24] discovered that German patients overwhelmingly supported data donation for medical research when "the combination of legal entitlement and easy-to-exercise-right to opt-out" was implemented. They recommended considering the incorporation of this approach into national law [24]. However, there is a great opportunity for networks and platforms to gain reputational benefits from conforming to rigorous data use standards such as the GDPR [41]. As previously reported, it is fundamental to implement and actively incorporate ethical, legal, and societal aspects of clinical and research data into governance structures and policies in research platforms and also to meet specifications for (local) institutional review boards [27,42,43]. Our platform established structural governance through, for example, a dedicated use and access board with ethics and data security experts.

Three broad consent items in our study referred to the donation of biomaterial-related data. Modern biobanks need to link biomaterials to secondary data sources, requiring funding and patients' consent [44]. With a very high consent rate to these items, we showed that patients with IRDs are highly willing to connect their biomaterial-related information to secondary data sources. In addition, our agreement rates to broad consent items related to biomaterial data use were higher than those recently reported for the MII broad consent form [21]. This suggests that the biobank research community will be able to overcome challenges linking biomaterials to patient-related data in EHRs and other sources and reinforces our choice to include biomaterial data donation in the DNP scenario [44].

Crucially, the willingness to integrate biomaterial-linked omics and genomics data from prior donations further enhances the scientific and translational value of RWD and the DNP, enabling multimodal datasets that are essential for precision medicine approaches and AI-driven analytics, as well as for regulatory science, HTA, and other cross-border research purposes [27,39,45]. Although AI is not mentioned in depth in our current framework and the patient informed consent form, its relevance is substantial: platforms such as the DNP are inherently designed for longitudinal, high-dimensional, and multimodal data integration required for AI applications not only in rheumatology [46]. Consequently, our high broad consent agreement rates constitute a key enabler for scalable AI use in multifaceted health care research, including predictive modeling, patient stratification, and regulatory evidence generation.

Our findings on team time spent on broad consent align with the MII's duration assessments on informed consent for secondary use of health care data and leftover biomaterials [47]. Our time spent informing patients was similar to the MII's, and the time spent was regarded as helpful, informative, and sufficient [47]. Personal interaction with a study team member or neutral educator is seen as the most effective way to improve participants' understanding [28], which likely contributed to our high recall and consent rates. Most of our patients kept their broad consent form copy, guaranteeing complete transparency and potential reassurance of what was signed. After 1 year, only 1 male patient with SLE withdrew consent, and patients asked for DNP results in routine care, reflecting sustained positive attitudes toward data donation to medical research and the DNP. However, the provision of broad consent– or study-related information to patients needs to be rethought as increasingly complex regulatory content requires innovative, multimedia-based approaches (eg, mixed reality and video clips) to convey information in an accessible, language-adapted, and comprehensible way [48]. Furthermore, while broad consent facilitates scalable use of data and biomaterials within a defined governance framework and dynamic consent enhances participant autonomy through ongoing control, neither model fully solves the structural limitations of traditional consent, highlighting a trade-off between efficiency and granularity that may be best addressed by hybrid approaches [27,49,50].

Recently, accessibility to RWD has improved at various levels, from national repositories to local hospital research databases [51]. Missing data and interoperability remain 2 of the most significant hurdles for RWD capture and use [10]. The European Alliance of Associations for Rheumatology core dataset agreements and European Reference Network efforts may help mitigate this [52-54]. Other obstacles to RWD research include access to suitable data sources and the need for sharing [51]. Federated learning approaches have been discussed to be advantageous over centralized databases (eg, in cancer, where RWD research is more advanced, or in the implementation of the EHDS [10,51,55,56]). European Union–funded projects develop big data infrastructures enabling large-scale data and AI integration [57]. These RWD infrastructures and other SPEs such as the DNP may link claims data with RWD for analyses. Initial studies, predominantly outside rheumatology, indicate that linking claims and registry data is feasible and useful [58-60]. Future approaches must consider that valuable EHR content will be revolutionized with the development of AI-driven data extraction and use [61]. With these ongoing developments, our patients' extensive support for integrating patient-level RWD across different health care silos and settings provides a strong foundation to perform clinical and scientific research, and also in collaboration with the pharmaceutical industry and/or medical device manufacturers.

With the DNP, we established a firm legal basis for patient RWD integration rooted in informed consent pursuant to the GDPR, including key elements such as biobanked biomaterial data for research and permission for future use of modern AI methods.

XSL•FO
RenderX

Strengths and Limitations

Despite a high response rate, selection and nonresponse bias cannot be excluded. Most participants were female patients with IRDs at a tertiary center; studies beyond such centers are needed. The observed high level of broad consent acceptance may represent an upper bound of patient willingness to participate in secondary data use. Consequently, the findings may not be directly generalizable to less engaged populations or different health care contexts, such as primary care settings or other routine clinical environments. The time spent obtaining broad consent in this study might not be feasible in routine care without reimbursement. Further research should explore data sharing willingness (eg, in general practices or private rheumatology practices); reasons for consenting, nonconsenting, or withholding consent to specific broad consent items (eg, via a qualitative approach); and understanding of content. Sex differences in completion rates require study in more balanced female-to-male ratios. The lowest consent rate was observed for informing other treating physicians about DNP participation, but our analyses did not reveal the reasons. Possible reasons might be complex phrasing or fatigue when nearing completion of the broad consent form. We did not specifically evaluate whether the involvement of commercial or private partners affected consent decisions. However, as study participation refusal was not communicated to be content related and no participants had reviewed the consent form prior to active refusal, this factor is, from the current perspective, unlikely to have meaningfully influenced participation in this study. The classification of patients into 2 groups was used as a conceptual framework to illustrate potential patterns of engagement with the broad consent form; however, patient behavior likely exists along a continuum rather than in discrete groups. Further research on these aspects is warranted.

Conclusions

This study is the first to assess the willingness of patients with IRDs to grant broad consent for the secondary use of RWD within an innovative RWD platform that implements a modern framework in routine care while considering granular patient preferences. We found a high willingness among patients with IRDs to consent to the aims of the DNP. Our findings provide important RWE on actual consent behavior and deliver key insights for the implementation of scalable RWD integration initiatives, such as those envisioned within the EHDS and other data-driven research infrastructures. Furthermore, the DNP enables scalable, GDPR-compliant data sharing that supports real-world and AI-driven research, strategic decision-making, and the preservation of patient trust. We demonstrated that, in rheumatology, RWD use for RWE generation is not substantially constrained by patients' willingness to provide broad consent.

Acknowledgments

Data collection was supported by Andrea Stockmanns (now deceased), Dr Christina Düsing, and Vivi Katana. The authors are grateful for the participation of the patients in support of the Data Nexus platform. They thank Dr Arnd Becker and Dr Brenda Reinhart for valuable discussions on Data Nexus platform topics and the very careful revision of the manuscript. The authors declare the use of generative artificial intelligence (GenAI) in the writing process. According to the Generative Artificial Intelligence Delegation Taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: formulation of revised conclusions. The GenAI tools used were GPT-5.5 (OpenAI) for language refinement during the revision of the manuscript and ChatGPT Images 2.0 (OpenAI) for the creation of Figure 2. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This project is cofunded by ZS Inc.

Data Availability

The data supporting the conclusions of this article are available on reasonable request from the authors.

Authors' Contributions

All authors were involved in the project and made substantial contributions to the conception and design of the work and the acquisition, analysis, and interpretation of the data. All authors developed the manuscript and approved the final draft.

Conflicts of Interest

MS and RH are employed by Serrala Group GmbH. WO and DK are employed by ZS Inc. EA was employed by ZS Inc until March 2025. All other authors declare no other conflicts of interest.

References

1. Heesen P, Schelling G, Birbaumer M, et al. Real-world-time data and RCT synergy: advancing personalized medicine and sarcoma care through digital innovation. Cancers (Basel) 2024 Jul 11;16(14):2516. [doi: 10.3390/cancers16142516] [Medline: 39061155]

XSL•FO
RenderX

2. Meissner Y, Schäfer M, Schneider M, et al. Incidence of facial nerve palsies stratified by DMARD treatment in patients with rheumatoid arthritis: data from the RABBIT register. RMD Open 2020 Oct;6(3):e001403. [doi: 10.1136/rmdopen-2020-001403] [Medline: 33070117]
3. FORWARD - The National Databank for Rheumatic Diseases. URL: https://www.forwarddatabank.org/ [accessed 2024-07-29]
4. About RISE registry. American College of Rheumatology. URL: https://rheumatology.org/about-rise-registry [accessed 2024-07-21]
5. Callhoff J, Feuchtenberger M, Karberg K, et al. With RheMIT rheumatology centers can participate in the German national database-expansion of the long-term rheumatological documentation [Article in German]. Z Rheumatol 2023 Aug;82(6):508-516. [doi: 10.1007/s00393-023-01373-y] [Medline: 37280334]
6. Prilla S, Groeneveld S, Pacurariu A, et al. Real-world evidence to support EU regulatory decision making-results from a pilot of regulatory use cases. Clin Pharmacol Ther 2024 Nov;116(5):1188-1197. [doi: 10.1002/cpt.3355] [Medline: 38962830]
7. Chen D. Real-world studies: bridging the gap between trial-assessed efficacy and routine care. J Biomed Res 2022 May 10;36(3):147-154. [doi: 10.7555/JBR.36.20220040] [Medline: 35652294]
8. LoCasale RJ, Pashos CL, Gutierrez B, et al. Bridging the gap between RCTs and RWE through endpoint selection. Ther Innov Regul Sci 2021 Jan;55(1):90-96. [doi: 10.1007/s43441-020-00193-5] [Medline: 32632753]
9. Roberts MH, Ferguson GT. Real-world evidence: bridging gaps in evidence to guide payer decisions. Pharmacoecon Open 2021 Mar;5(1):3-11. [doi: 10.1007/s41669-020-00221-y] [Medline: 32557235]
10. Graili P, Guertin JR, Chan KK, Tadrous M. Integration of real-world evidence from different data sources in health technology assessment. J Pharm Pharm Sci 2023;26:11460. [doi: 10.3389/jpps.2023.11460] [Medline: 37529633]
11. When open data meets data spaces. European Union. URL: https://data.europa.eu/en/publications/datastories/when-open-data-meets-data-spaces [accessed 2024-07-21]
12. Winkler EC, Jungkunz M, Thorogood A, Lotz V, Schickhardt C. Patient data for commercial companies? An ethical framework for sharing patients' data with for-profit companies for research. J Med Ethics 2025 Apr 24;51(5):e108781. [doi: 10.1136/jme-2022-108781] [Medline: 37230744]
13. Kc S, Lin LW, Bayani DB, et al. What, where, and how to collect real-world data and generate real-world evidence to support drug reimbursement decision-making in Asia: a reflection into the past and a way forward. Int J Health Policy Manag 2023;12:6858. [doi: 10.34172/ijhpm.2023.6858] [Medline: 37579427]
14. Dang A. Real-world evidence: a primer. Pharmaceut Med 2023 Jan;37(1):25-36. [doi: 10.1007/s40290-022-00456-6] [Medline: 36604368]
15. Darwin EU. URL: https://www.darwin-eu.org/ [accessed 2024-11-04]
16. Wehrle K, Tozzi V, Braune S, et al. Implementation of a data control framework to ensure confidentiality, integrity, and availability of high-quality real-world data (RWD) in the NeuroTransData (NTD) registry. JAMIA Open 2022;5(1):ooac017. [doi: 10.1093/jamiaopen/ooac017] [Medline: 35571355]
17. Richter JG, Weiß A, Bungartz C, et al. Mobile responsive app-a useful additional tool for data collection in the German pregnancy register Rhekiss? Front Med (Lausanne) 2021;8:773836. [doi: 10.3389/fmed.2021.773836] [Medline: 34977074]
18. Tillett W, Birt J, Cavanaugh C, et al. Changes in musculoskeletal disease activity and patient-reported outcomes in patients with psoriatic arthritis treated with ixekizumab: results from a real-world US cohort. Front Med (Lausanne) 2023;10:1184028. [doi: 10.3389/fmed.2023.1184028] [Medline: 37415769]
19. van Drumpt S, Chawla K, Barbereau T, Spagnuelo D, van de Burgwal L. Secondary use under the European Health Data Space: setting the scene and towards a research agenda on privacy-enhancing technologies. Front Digit Health 2025;7:1602101. [doi: 10.3389/fdgth.2025.1602101] [Medline: 40613075]
20. Reuse of health data. European Commission. URL: https://health.ec.europa.eu/ehealth-digital-health-and-care/reuse-health-data_en [accessed 2026-05-20]
21. Herrmann T, Mallow J, Plaumann M, et al. The implementation of broad consent at university hospitals. Dtsch Arztebl Int 2025 Jun 13;122(12):338-339. [doi: 10.3238/arztebl.m2025.0078] [Medline: 40637480]
22. Köngeter A, Schickhardt C, Jungkunz M, Bergbold S, Mehlis K, Winkler EC. Patients' willingness to provide their clinical data for research purposes and acceptance of different consent models: findings from a representative survey of patients with cancer. J Med Internet Res 2022 Aug 25;24(8):e37665. [doi: 10.2196/37665] [Medline: 36006690]
23. Richter G, Trigui N, Caliebe A, Krawczak M. Attitude towards consent-free research use of personal medical data in the general German population. Heliyon 2024 Mar 11;10(6):e27933. [doi: 10.1016/j.heliyon.2024.e27933] [Medline: 38509969]
24. Richter G, Borzikowsky C, Hoyer BF, Laudes M, Krawczak M. Secondary research use of personal medical data: patient attitudes towards data donation. BMC Med Ethics 2021 Dec 15;22(1):164. [doi: 10.1186/s12910-021-00728-x] [Medline: 34911502]
25. Template text for patient consent forms. Medical Informatics Initiative. URL: https://www.medizininformatik-initiative.de/en/template-text-patient-consent-forms [accessed 2024-07-22]
26. DocuMed. 2007. URL: http://www.rheumanet.org/DocumedSupport/Downloads/White_Paper_Feb2007.pdf [accessed 2026-06-02]

XSL•FO
RenderX

27. Richter JG, Filla T, Acar H, et al. Sustained agreement rates in the longitudinal assessment of lupus patients to a broad consent for personal data and specimen usage in the RHINEVIT biobank. Front Med (Lausanne) 2023;10:1208006. [doi: 10.3389/fmed.2023.1208006] [Medline: 37415767]
28. Flory J, Emmanuel E. Interventions to improve research participants' understanding in informed consent for research: a systematic review. Am J Ophthalmol 2005 Feb;139(2):399. [doi: 10.1016/j.ajo.2004.12.040] [Medline: 15467062]
29. Albrecht K, Thiele K, Alexander T, et al. Current data on rheumatological care: annual report from the national database (NDB) of the regional collaborative arthritis centres in Germany [Article in German]. Z Rheumatol 2024 Oct;83(8):666-674. [doi: 10.1007/s00393-024-01497-9] [Medline: 38568444]
30. Beusink M, Rebers S, Broeks A, et al. Implementing broad consent for research with routinely collected clinical data and residual biosamples in a cancer hospital: using mixed methods approach to evaluate consent rates and patients' perspectives. BMC Med Ethics 2026 Jan 9;27(1):31. [doi: 10.1186/s12910-025-01367-2] [Medline: 41514306]
31. Fischer-Rosinský A, Eienbröker L, Möckel M, et al. Broad consent in the emergency department: a cross sectional study. Arch Public Health 2025 Feb 18;83(1):44. [doi: 10.1186/s13690-025-01529-z] [Medline: 39966946]
32. Tosoni S, Voruganti I, Lajkosz K, et al. The use of personal health information outside the circle of care: consent preferences of patients from an academic health care institution. BMC Med Ethics 2021 Mar 24;22(1):29. [doi: 10.1186/s12910-021-00598-3] [Medline: 33761938]
33. Baines R, Stevens S, Austin D, et al. Patient and public willingness to share personal health data for third-party or secondary uses: systematic review. J Med Internet Res 2024 Mar 5;26:e50421. [doi: 10.2196/50421] [Medline: 38441944]
34. Cascini F, Pantovic A, Al-Ajlouni YA, Puleo V, De Maio L, Ricciardi W. Health data sharing attitudes towards primary and secondary use of data: a systematic review. EClinicalMedicine 2024 May;71:102551. [doi: 10.1016/j.eclinm.2024.102551] [Medline: 38533128]
35. Pilgrim K, Bohnet-Joschko S. Donating health data to research: influential characteristics of individuals engaging in self-tracking. Int J Environ Res Public Health 2022 Aug 2;19(15):9454. [doi: 10.3390/ijerph19159454] [Medline: 35954812]
36. Große Offenheit für Spende von P. Bitkom. URL: https://www.bitkom.org/Presse/Presseinformation/Grosse-Offenheit-fuer-Spende-von-Patientendaten [accessed 2024-07-29]
37. Broekstra R, Maeckelberghe EL, Aris-Meijer JL, Stolk RP, Otten S. Motives of contributing personal data for health research: (non-)participation in a Dutch biobank. BMC Med Ethics 2020 Jul 25;21(1):62. [doi: 10.1186/s12910-020-00504-3] [Medline: 32711531]
38. Niño de Rivera S, Masterson Creber R, Zhao Y, et al. Public perspectives on increased data sharing in health research in the context of the 2023 National Institutes of Health Data Sharing Policy. PLoS One 2024;19(8):e0309161. [doi: 10.1371/journal.pone.0309161] [Medline: 39197051]
39. Richter G, Borzikowsky C, Lesch W, et al. Secondary research use of personal medical data: attitudes from patient and population surveys in The Netherlands and Germany. Eur J Hum Genet 2021 Mar;29(3):495-502. [doi: 10.1038/s41431-020-00735-3] [Medline: 33005018]
40. Pahune S, Akhtar Z, Mandapati V, Siddique K. The importance of AI data governance in large language models. Big Data Cogn Comput 2025;9(6):147. [doi: 10.3390/bdcc9060147]
41. Andreotta AJ, Kirkham N, Rizzi M. AI, big data, and the future of consent. AI Soc 2022;37(4):1715-1728. [doi: 10.1007/s00146-021-01262-5] [Medline: 34483498]
42. Platt J, Bollinger J, Dvoskin R, Kardia SL, Kaufman D. Public preferences regarding informed consent models for participation in population-based genomic research. Genet Med 2014 Jan;16(1):11-18. [doi: 10.1038/gim.2013.59] [Medline: 23660530]
43. Hampf C, Bialke M, Geidel L, et al. A survey on the current status and future perspective of informed consent management in the MIRACUM consortium of the German Medical Informatics Initiative. Transl Med Commun 2021;6:7. [doi: 10.1186/s41231-021-00086-1]
44. Lieb W, Strathmann EA, Röder C, et al. Population-based biobanking. Genes (Basel) 2024 Jan 3;15(1):66. [doi: 10.3390/genes15010066] [Medline: 38254956]
45. Chen YM, Hsiao TH, Lin CH, Fann YC. Unlocking precision medicine: clinical applications of integrating health records, genetics, and immunology through artificial intelligence. J Biomed Sci 2025 Feb 7;32(1):16. [doi: 10.1186/s12929-024-01110-w] [Medline: 39915780]
46. Aldhuaina K, Gupta D, Bashir U, et al. Artificial intelligence in rheumatology: clinical applications in rheumatoid arthritis, osteoarthritis, and systemic lupus erythematosus. Cureus 2025 Dec;17(12):e99108. [doi: 10.7759/cureus.99108] [Medline: 41531594]
47. Roschka S, Leddig T, Bullerjahn M, et al. Secondary use of health care data and left-over biosamples within the "Medical Informatics Initiative" (MII): a quasi-randomized controlled evaluation of patient perceptions and preferences regarding the consent process. BMC Med Inform Decis Mak 2022 Jul 15;22(1):184. [doi: 10.1186/s12911-022-01922-6] [Medline: 35840947]
48. Westarp E, Saemann A, Zelechovski M, et al. Virtual reality for patient informed consent in skull base tumors and intracranial vascular pathologies: a pilot study. Acta Neurochir (Wien) 2024 Nov 15;166(1):455. [doi: 10.1007/s00701-024-06355-w] [Medline: 39546033]

XSL•FO
RenderX

49. Freye M, Buchner B. Legal regulation of digital public health interventions-hindrance or stimulus? [Article in German]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2024 Mar;67(3):285-291. [doi: 10.1007/s00103-024-03835-3] [Medline: 38329530]
50. Starekova J, Schweitzer ME. Broad consent in healthcare research: what is efficient, what is right? J Magn Reson Imaging 2025 Sep;62(3):856-857. [doi: 10.1002/jmri.70000] [Medline: 40481697]
51. Price G, Peek N, Eleftheriou I, et al. An overview of real-world data infrastructure for cancer research. Clin Oncol (R Coll Radiol) 2025 Feb;38:103545. [doi: 10.1016/j.clon.2024.03.011] [Medline: 38631976]
52. Ehlers L, Askling J, Bijlsma HW, et al. 2018 EULAR recommendations for a core data set to support observational research and clinical care in giant cell arteritis. Ann Rheum Dis 2019 Sep;78(9):1160-1166. [doi: 10.1136/annrheumdis-2018-214755] [Medline: 30898837]
53. Bandeira M, Di Cianni F, Marinello D, et al. An overlook on the current registries for rare and complex connective tissue diseases and the future scenario of TogethERN ReCONNET. Front Med (Lausanne) 2022;9:889997. [doi: 10.3389/fmed.2022.889997] [Medline: 36226147]
54. Bandeira M, Marinello D, Barreira SC, et al. Coding systems and monitoring practices across the ERN ReCONNET: insights from a comprehensive survey and unmet needs. Orphanet J Rare Dis 2026 Mar 30;21(1):186. [doi: 10.1186/s13023-026-04304-7] [Medline: 41913222]
55. Raab R, Küderle A, Zakreuskaya A, et al. Federated electronic health records for the European Health Data Space. Lancet Digit Health 2023 Nov;5(11):e840-e847. [doi: 10.1016/S2589-7500(23)00156-5] [Medline: 37741765]
56. Marcus JS, Martens B, Carugati C, Bucher A, Godlovitch I. The European Health Data Space. SSRN J 2022. [doi: 10.2139/ssrn.4300393]
57. Kondylakis H, Kalokyri V, Sfakianakis S, et al. Data infrastructures for AI in medical imaging: a report on the experiences of five EU projects. Eur Radiol Exp 2023 May 8;7(1):20. [doi: 10.1186/s41747-023-00336-x] [Medline: 37150779]
58. Papanastasiou G, Yang G, Fotiadis DI, et al. Large-scale deep learning analysis to identify adult patients at risk for combined and common variable immunodeficiencies. Commun Med (Lond) 2023 Dec 20;3(1):189. [doi: 10.1038/s43856-023-00412-8] [Medline: 38123736]
59. Oliphant BW, Cain-Nielsen AH, Jarman MP, et al. Linking trauma registry patients with insurance claims: creating a longitudinal patient record. J Surg Res 2024 Mar;295:274-280. [doi: 10.1016/j.jss.2023.11.002] [Medline: 38048751]
60. Solomon DH, Liu CC, Kuo IH, Zak A, Kim SC. Effects of colchicine on risk of cardiovascular events and mortality among patients with gout: a cohort study using electronic medical records linked with Medicare claims. Ann Rheum Dis 2016 Sep;75(9):1674-1679. [doi: 10.1136/annrheumdis-2015-207984] [Medline: 26582823]
61. Sunoh.ai medical AI scribe integrates with Epic's open interoperability platform. Sunoh.ai. URL: https://sunoh.ai/press-releases/sunoh-ai-medical-ai-scribe-integrates-with-epics-open-interoperability-platform/ [accessed 2024-07-22]

Abbreviations

AI: artificial intelligence
DNP: Data Nexus platform
EHDS: European Health Data Space
EHR: electronic health record
EMR: electronic medical record
GDPR: General Data Protection Regulation
HTA: health technology assessment
IRD: inflammatory rheumatic disease
IRD: inflammatory rheumatic disease
MII: Medical Informatics Initiative
PsA: psoriatic arthritis
RA: rheumatoid arthritis
RCT: randomized controlled trial
RWD: real-world data
RWE: real-world evidence
SLE: systemic lupus erythematosus
SpA: spondyloarthritis
SPE: secure processing environment

XSL•FO
RenderX

Edited by S Brini; submitted 11.Jul.2025; peer-reviewed by A Sagkriotis, H Morf, K Gierend; revised version received 12.May.2026; accepted 13.May.2026; published 08.Jun.2026.

Please cite as:
Richter JG, Becker A, Filla T, Acar H, Ockert W, Kriauciunas D, Aggarwal E, Schröder M, Hansen R, Distler JHW, Schneider M High Patient Willingness to Grant Broad Consent for Real-World Data Use in Rheumatology—Implications for Real-World Data Platform Governance: Cross-Sectional Study
J Med Internet Res 2026;28:e80347
URL: https://www.jmir.org/2026/1/e80347
doi:10.2196/80347

© Jutta G Richter, Antonia Becker, Tim Filla, Hasan Acar, Waldemar Ockert, Dominykas Kriauciunas, Edith Aggarwal, Markus Schröder, Ralf Hansen, Jörg HW Distler, Matthias Schneider. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Guide on Selection of Optimal Motivational Themes for Use in a Clinical Trial Recruiting Black US Adults: Survey Study

Tamunotonye Harry[1*], MS; Jingyi Cao[2], MPH; Zaib Hussain[3], MBBS, MPH; Ruth-Alma N Turkson-Ocran[2], MPH, APRN, PhD; Stephen P Juraschek[2], MD, PhD; Timothy P Lahey[4], MD, MMSc; Timothy B Plante[4], MD, MHS; Yuanyuan Feng[1*], PhD

[1]Department of Computer Science, University of Vermont, 85 South Prospect Street, Burlington, VT, United States

[2]Division of General Medicine, Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, United States

[3]Division of Cardiology, Johns Hopkins University School of Medicine, Baltimore, MD, United States

[4]Department of Medicine, Larner College of Medicine, University of Vermont, Burlington, VT, United States

[*]these authors contributed equally

Corresponding Author:
Tamunotonye Harry, MS
Department of Computer Science, University of Vermont, 85 South Prospect Street, Burlington, VT, United States

Abstract

Background: Black adults in the United States face significant cardiovascular health disparities, which are likely exacerbated by the underrepresentation of Black adults in cardiovascular clinical trials. The Black adult US population has experienced unique historical events, discriminatory practices, and practical obstacles that might contribute to this underrepresentation in clinical trials. An improved understanding of motivations that encourage or discourage participation in cardiovascular clinical trials can lead to more effective clinical trial recruitment and help mitigate these cardiovascular health disparities.

Objective: This study aimed to determine, using an online survey, which motivational themes in clinical trial recruitment advertisements were most effective in encouraging Black adults to participate in a hypertension-focused trial. We also examined how trust in health care and various demographic factors influenced their decision to participate.

Methods: We conducted an online survey with 829 self-identified Black adults in the United States, using a between-subject design to test 4 literature-derived motivational themes in clinical trial recruitment advertisements: (1) contribution to science, (2) helping the community, (3) lowering blood pressure, and (4) access to perks (US $500 worth of groceries or an equivalent cash amount). We assessed advertisement appeal, willingness to participate, and willingness to recommend clinical trial participation to others using cumulative link mixed models.

Results: Demographic factors played a more significant role than motivational themes in predicting advertisement effectiveness. Adults aged 40 - 59 years and individuals diagnosed with high blood pressure were more likely to find the advertisements appealing and express willingness to participate. Urban residents engaged more with the advertisements compared to those in suburban or rural areas. Participants with liberal (odds ratio [OR] 1.37, 95% CI 1.01 - 1.85; *P*=.04) and conservative (OR 1.62, 95% CI 1.09 - 2.40; *P*=.02) political views were more willing to participate in the clinical trial compared to those with moderate views. However, the “lowering my blood pressure” theme was less effective among individuals who distrusted health care institutions (OR 0.40, 95% CI 0.16 - 0.97; *P*=.04) and also reduced willingness to recommend the trial (OR 0.36, 95% CI 0.15 - 0.85; *P*=.02). In addition, higher trust levels were unexpectedly associated with lower willingness to participate when exposed to this theme (OR 0.41, 95% CI 0.17 - 0.98; *P*=.04).

Conclusions: Demographic targeting (age, health status, and geographic location) is more critical than generic motivational messaging in recruiting Black adults to clinical trials. Successful digital health recruitment requires targeting specific demographic segments with tailored messages, as effectiveness varies significantly across sociodemographic factors. The online survey methodology offers researchers a rapid, scalable tool for pretesting recruitment strategies, though it should complement rather than replace community engagement. These insights can help reduce cardiovascular health disparities by improving clinical trial representativeness.

(J Med Internet Res 2026;28:e75857) doi:10.2196/75857

KEYWORDS

clinical trials; recruitment; African American; digital advertising; health disparities; cardiovascular health

XSL•FO
RenderX

Introduction

Background

In the United States, Black adults face poorer cardiovascular health outcomes compared to other groups [1] and are underrepresented in clinical trials [2] that may benefit them. While some large cardiovascular trials have successfully recruited adequate numbers of Black participants through intensive community engagement [3,4], persistent underrepresentation remains a widespread challenge across most trials [2,5,6], particularly outside major academic medical centers. The drivers of the underrepresentation of Black adults in cardiovascular clinical trials are diverse. While the Tuskegee Syphilis Study [7] is often cited as a contributor to medical mistrust among Black adults, research suggests that knowledge of Tuskegee does not universally deter participation and that its impact varies considerably [8,9]. Contemporary factors, such as ongoing experiences with discriminatory practices and physician bias based on race or ethnicity, mental health status, education, income, and sex within the US health care system, play more substantial roles in shaping trust and participation decisions [10,11]. In addition, practical obstacles to participation, such as a lack of information about clinical trials, health insurance, time, and financial constraints, are significant barriers to participation [12,13], contributing to the failure of clinical trials in the United States to achieve their demographic representation goals [2,5,6]. These contemporary, structural barriers often outweigh historical events in determining participation. Despite the federal government's efforts to ensure inclusion in clinical trials [14], underrepresentation persists, resulting in disparities in health outcomes for minority racial and ethnic groups. Targeted advertisements may improve the recruitment of US Black adults to trials, but it is unclear what advertisement content might be most beneficial.

To support the successful recruitment of Black adults into a community-based participatory trial focused on recruiting Black urban adults (Groceries for Black Residents of Boston to Stop Hypertension [GoFresh]), we conducted an online survey with 829 self-identified Black adults in the United States to evaluate the effectiveness of a flyer-style trial recruitment advertisement with 4 literature-derived motivational themes. The goal of this study is twofold (1) to identify the optimal recruitment motivational themes for recruiting Black adults into the GoFresh trials and (2) to examine other factors (eg, trust in health care systems and demographic characteristics) that may impact Black adults' intentions to participate in clinical trials. In addition, this study demonstrates how online survey platforms paired with digital advertisement testing can serve as a rapid, cost-effective tool for pretesting recruitment messages before large-scale deployment. This approach allows researchers to iterate on messaging strategies and identify demographic-specific preferences, complementing traditional community engagement approaches. This digital health approach is particularly relevant as clinical trials increasingly use digital channels (eg, social media and online advertisements) to reach diverse populations, yet often lack systematic evidence about which digital messages are most effective for specific audiences.

We seek to answer a key research question (RQ):

- RQ: What motivates Black adults to participate in clinical trials, and how do demographic factors, health status, and trust in health care influence their responses to different recruitment messages?

Based on prior literature, we developed the following hypotheses:

- H1: Different motivational themes in clinical trial advertisements result in different levels of appeal, willingness to participate, and willingness to recommend the trial among Black adults.
- H2: The effectiveness of motivational themes is influenced by an individual's trust levels in health care systems.
- H3: Demographic characteristics influence responses to clinical trial recruitment through both direct effects on participation interest and by moderating the effectiveness of different motivational themes.

These hypotheses are grounded in prior research demonstrating that (1) message framing affects health behavior [15,16], (2) trust in health care systems significantly influences Black adults' willingness to participate in research [11,17,18], and (3) demographic characteristics shape health priorities and information processing [19,20].

Previous Work (Literature Review)

Clinical trials are essential for discovering new prevention, diagnostic, or treatment approaches and improving population health. Without accounting for diversity, resulting interventions may prove ineffective for broader populations, undermining trust and potentially increasing health costs [21,22]. US clinical trials have consistently underrepresented minority groups despite government policies to improve recruitment [23]. Beyond investigator bias [24], possible reasons include lack of awareness, mistrust of the health care system due to experiences with discrimination or bias, and financial and insurance-related obstacles [12,13,24]. A fundamental barrier to Black adults' participation in clinical trials is that they are often never explicitly asked or invited to participate [25,26]. When Black adults are approached, they are frequently deemed ineligible due to comorbid conditions, which are prevalent in this population due to longstanding health disparities [27]. This structural exclusion, whether through lack of outreach or restrictive eligibility criteria, represents a critical gap that precedes any consideration of motivational messaging.

Current solutions, such as population-specific trials and community engagement, have increased participation from underrepresented groups [28]. However, these are costly in human resources, time-consuming, and risk selection bias, making them unsustainable long-term [28,29]. Most clinical trials rely on traditional recruitment methods such as recruitment by investigators, local doctors, patient databases, printed materials (eg, flyers and posters), and community organizations, with limited data on the performance of digital strategies (eg, websites, social media marketing, and online advertisements) [30]. As a result, clinical trial recruitment recommendations are primarily based on case studies and individual practitioners' experiences [30,31]. Despite existing research on recruitment

XSL•FO
RenderX

strategy effectiveness, there remains a significant gap in data-driven approaches to systematically understand what truly improves clinical trial recruitment.

Big data and web analytics have transformed advertising across industries [32-34], enabling data-driven decision-making. Initial evidence suggests that these approaches can boost participation among underrepresented groups [35,36], yet most US clinical trials fail to leverage these methods to enhance recruitment diversity [37]. Given this opportunity, clinical trial advertising can be tailored, using data-driven methods, to the values and motivations of the target population. Nevertheless, data-driven strategies carry inherent risks of bias when underlying data underrepresent or stereotypically portray minoritized groups [38,39].

The literature [12,19,20] highlights how specific factors in individuals' lives influence their decision to participate in clinical trials. Recognizing each community's unique values allows researchers to develop resonant recruitment strategies, improving enrollment and retention [40]. Integrating big data and web analytics into clinical trial recruitment may help identify patterns and preferences among diverse populations. By analyzing these data, researchers can determine which messages and platforms most effectively reach different demographics, leading to more efficient and inclusive recruitment efforts. This study uses a literature-derived approach to develop 4 motivational themes for the GoFresh trials [12,41-43]. These themes reflect both intrinsic motivations (personal values and community impact) [44-46] and extrinsic motivations (health outcomes and financial benefits) [47,48] documented in prior qualitative and quantitative studies of trial recruitment barriers and facilitators.

Study Aims

This study has 3 primary aims:

- Aim 1: Determine which motivational themes in clinical trial recruitment advertisements are most effective in encouraging Black adults to participate in a cardiovascular health trial, and identify how these preferences vary by demographic characteristics.
- Aim 2: Examine how trust in health care systems moderates Black adults' responses to different recruitment message themes.
- Aim 3: Demonstrate proof of concept for using an online survey methodology as a rapid, cost-effective tool for pretesting clinical trial recruitment messages before large-scale deployment.

Methods

GoFresh Trials

This study is a collaboration with the GoFresh trials. The GoFresh trials (ClinicalTrials.gov nos NCT05121337 [GoFresh] and NCT05393232 [GoFreshRx]) are investigating the impact of home-delivered healthy groceries on blood pressure in Black adults with untreated and treated high blood pressure. In this study, participants are randomly assigned to receive either 12 weeks of Dietary Approaches to Stop Hypertension–patterned [49] grocery delivery with dietitian assistance or self-directed shopping with a monthly stipend over 3 months. After the intervention period, both groups are observed for 9 more months to determine barriers and facilitators for maintaining the Dietary Approaches to Stop Hypertension diet. The ultimate goal of these trials is to improve the cardiovascular health of Black adults living in healthy food priority areas, with and without treated hypertension. Findings from the trials will provide practical insights into effective and scalable strategies to prevent hypertension among Black adults [50].

Survey Design

The 4 motivational themes were derived from prior research on factors influencing clinical trial participation [12,41-43]. These themes include contributing to science, helping the community, lowering my blood pressure, and accessing perks (Textbox 1). Flyer-style advertisements were developed for each of these motivational themes (Figures 1-4). The advertisement content and design were reviewed by a community advisory board at Morgan State University to ensure cultural appropriateness and relevance before being tested in the survey. These variations included the GoFresh logo but did not include real contact information for the trial.

Each participant was randomly assigned to 1 of the 4 advertisements and asked to answer the same set of survey questions, rating the level of advertisement appeal, their willingness to participate in the trial, and their willingness to recommend the trial to others. The survey also collected data on participants' trust in health care systems using a validated scale from Rose et al [51] and Armstrong et al [52], as well as their basic demographic characteristics (refer to Multimedia Appendix 1).

XSL•FO
RenderX

Textbox 1. List of independent variables.

Motivational themes:

- The 4 motivational themes in the trial recruitment advertisement:
- Contribution to science (science): motivating participants to support scientific research.
- Helping the community (community): encouraging participants to benefit their local community.
- Lowering my blood pressure (blood pressure): promoting personal health improvement.
- Access to perks (perks): highlighting tangible benefits for participation.

Trust levels in health care:

- Participants' trust levels calculated using the Healthcare System Distrust Scale (10-50) [32,33]:
- Distrust (10-30): lowest tertile of trust scores.
- Neutral (31-35): middle tertile of trust scores.
- Trust (36-50): high tertile of trust scores.

Demographic backgrounds:

- Potential confounding factors:
- Age group
- Political view
- Gender
- Location (urbanity)

Figure 1. Motivational theme of the advertisement variant: contribution to science.

XSL•FO
RenderX

Figure 2. Motivational theme of the advertisement variant: helping the community.

Figure 3. Motivational theme of the advertisement variant: lowering my blood pressure.

XSL•FO
RenderX

Figure 4. Motivational theme of the advertisement variant: access to perks.

Study Population

We used the Prolific platform (Prolific Academic Ltd) [53] for web-based participant recruitment. Prolific is an online research platform that connects researchers with high-quality, prescreened participants for surveys and experiments, ensuring ethical standards, fair compensation, and reliable data collection. The platform's prescreening filter ensured that only individuals who self-identified as Black or African American or mixed race were recruited. Multimedia Appendix 2 shows the inclusion and exclusion criteria used in the GoFresh trials. The survey was deployed on February 2, 2024, and was closed 3 days later. We targeted 1000 respondents. Excluding instances of obvious survey abuse (n=60), participants who provided complete responses were included in the analysis. We did not restrict our sample to individuals with diagnosed hypertension because examining how health status influences recruitment message responses was a key RQ. This also reflects real-world recruitment for cardiovascular trials targeting both prevention and treatment populations.

Variables and Measurements

Dependent Variables

We operationalized the effectiveness of the trial recruitment advertisement as three variables: (1) the appeal of the advertisement—the attractiveness of the advertisement, (2) willingness to participate—readiness to join the advertised clinical trial, and (3) willingness to recommend—likelihood of sharing the clinical trial with others.

The primary goal of advertising appeal is to effectively communicate with a target audience and motivate them to take action. It can be subjective depending on messaging, visuals, or media type (direct mail, social media, print, etc) that are attractive to one's cognition, emotions, or both, which are rooted in past and present experiences [54]. In clinical trials, an effective trial advertisement motivates (appeals to) the target audience to participate in or recommend the trial within their network.

Independent Variables

Participants were exposed to 1 of 4 motivational appeals in recruitment advertisements, each emphasizing a distinct reason for participation: contributing to scientific research (science), helping one's community (community), improving personal health outcomes, such as lowering blood pressure (blood pressure), or gaining access to tangible benefits (perks). Alongside motivation, trust in the health care system was considered a critical factor in recruitment engagement. Participants' trust levels were measured using the Healthcare System Distrust Scale [51,52] with scores ranging from 10 to 50. To ensure an even distribution, trust was categorized into 3 levels using tertile cutoffs calculated in RStudio software (Posit PBC) [55]: distrust (10-30), neutral (31-35), and trust (36-50).

These variables provide context for differences in responses to motivational themes and trust levels, helping to refine strategies for trial recruitment. Together, motivation, trust, and demographic characteristics shape participant engagement and the overall effectiveness of recruitment advertisements.

Statistical Analysis

Sociodemographics and model covariates were tabulated with descriptive statistics. A cumulative link mixed model, a type of multilevel ordered logistic regression model, was used in our analysis. This approach allows estimation of model effects on an ordinal dependent variable while accommodating for random effects (ie, repeated measures). Models were adjusted for demographic factors, including age, political views, gender, and urbanity (location).

Main Effect Predictors

We included predictors for advertisement themes, trust levels, and prior diagnosis of high blood pressure by a doctor (high

XSL•FO
RenderX

blood pressure status), as well as demographic variables such as age group, income, political view, and gender.

Effect Modifier

We used effect modifiers to examine the interactions among independent variables. We evaluated different interaction combinations among important independent variables to address our research hypothesis. After comparing the statistical significance of these interactions, we included interactions between advertisement themes and trust levels (ie, advertisement theme × trust level) to better elucidate how the effectiveness of each advertisement theme varied among people with different levels of trust in health care systems.

Predictor Baselines

We set the baseline value for the motivational theme as "contribution to science" since it is the fundamental goal of clinical trials. For other predictors, we selected the neutral or most common values as baselines. For example, for the trust in health care systems predictor, we selected the "Neutral" category as the baseline because it represented the most frequently chosen response within the trust variable. This ensures that the baseline reflects the predominant sentiment among participants, providing a solid reference point for analysis.

Data Cleaning

To ensure model convergence, we performed the following steps: (1) converted responses such as "not sure," "prefer not to answer," and uncommon responses (eg, gender; "binary") to the "not available (N/A)" placeholder value, which also made the data more uniform; and (2) combined categories within certain variables, such as trust in the health care system and political views, to simplify and standardize the data for modeling.

Final Models

We fit 18 ordered logistic regression models, 6 for each effectiveness metric: appeal, willingness to participate, and willingness to recommend. From these, we selected the best model for each metric (3 models in total) based on a balance between predictor complexity and performance, as measured by the Akaike information criterion score.

Results Reporting

For predictors with statistical significance ($P<.05$), we reported the odds ratios (ORs), calculated as the exponential of the model estimate (β), along with the 95% CI. We also calculated the marginal 95% CIs for these interactions to evaluate their statistical significance and range of effect.

Our statistical analyses were conducted using R (version 4.4.0; R Foundation for Statistical Computing) with the ordinal package [56]. No adjustments were made for multiple comparisons.

Ethical Considerations

This study was approved by the University of Vermont Institutional Review Board (STUDY00002558). All participants provided electronic informed consent before beginning the survey, after being informed about the study's purpose, procedures, risks, benefits, and their right to withdraw at any time. We anonymized participants' data before using the remaining 829 survey responses for the final analysis. All data were deidentified, and no personally identifiable information was collected or linked to survey responses. Data were stored securely on password-protected servers accessible only to the research team. Participants who provided complete responses were compensated US $1.60 each, corresponding to an average rate of US $9.58 per hour, through the Prolific platform [53].

Results

Analytical Sample

A total of 913 surveys were started, with 897 completed (did not complete: n=16). After removing entries with incomplete fields (n=44), 853 valid responses remained. As shown in Multimedia Appendix 2, we further excluded 24 participants for the following reasons: not identifying as Black, self-reporting a lack of understanding of clinical trials despite being provided with a definition, providing unreasonable or nonrandomized answers, or being age outliers. This resulted in a final analytical sample of 829 responses.

Analytical Sample Sociodemographics

The sociodemographics of the analytical sample are provided in Multimedia Appendix 3. The 829 valid responses were from participants who self-identified as Black or African American on Prolific; 7 of these participants identified as mixed race. The mean age was 40 (SD 13.0) years. Educational levels varied widely, and most participants (474/829) reported annual incomes between US $35,000 and US $129,999. Participants reporting female and male genders were equally represented (female: 50.2%, 416/829, and male: 49.1%, 407/829). Most participants lived in urban or suburban areas in the United States (large city: 28.2%, 234/829; city: 33.7%, 279/829; and town or suburb: 32.2%, 267/829) and reported liberal (54.6%, 453/829), moderate (26.8%, 222/829), or conservative (18.6%, 154/829) views.

In terms of health status, 29.6% (245/829) reported a prior diagnosis of high blood pressure. The mean score on the Healthcare System Distrust Scale was 32.7 (SD 4.9; range 10-50; distrust [10-30], neutral [31-35], and trust [36-50]). Responses to individual items on the scale showed that 33.4% (277/829) of participants expressed distrust, while 30.3% (251/829) expressed trust in the health care system.

Appeal of Advertisement

As shown in Figure 5, all motivational themes had a high level of appeal, with 76% - 79% (155/204-165/209) of respondents somewhat or strongly agreeing. Notably, the "access to perks" theme was the most appealing, with 79% (165/209) agreeing. In contrast, the "helping the community" theme had the highest level of disagreement, with 15% (31/206) of respondents somewhat or strongly disagreeing.

XSL•FO
RenderX

Figure 5. Percentage of participant responses to the appeal of motivational themes in clinical trial advertisements.

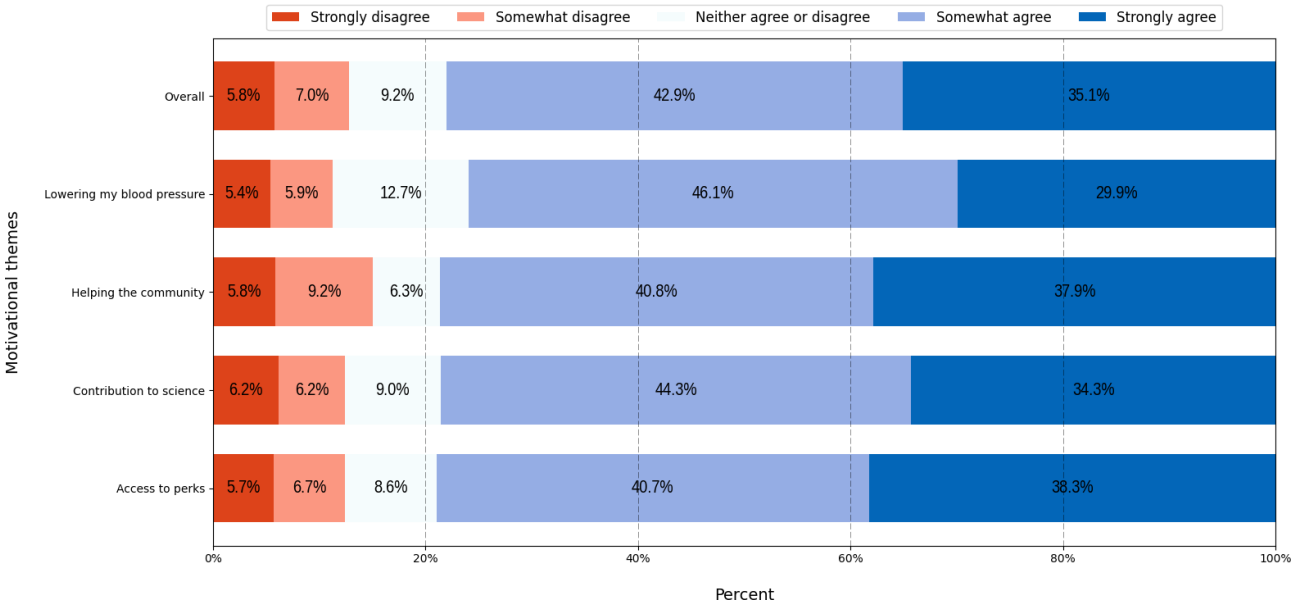

Willingness to Participate

As shown in Figure 6, scores for willingness to participate were lower than the appeal of the advertisement, with lower overall agreement to join clinical trials across all 4 motivational themes, ranging from 46.6% to 62.6% (95/204-131/209). The "access to perks" theme yielded the highest willingness to participate at 62.6% (131/209) and the lowest level of disagreement at 23% (48/209). Alternatively, the "lowering my blood pressure" theme had the lowest willingness at 46.6% and the highest level of disagreement, with 33.8% (69/204) strongly or somewhat disagreeing.

Figure 6. Percentage of participants' willingness to participate in clinical trials by motivational theme.

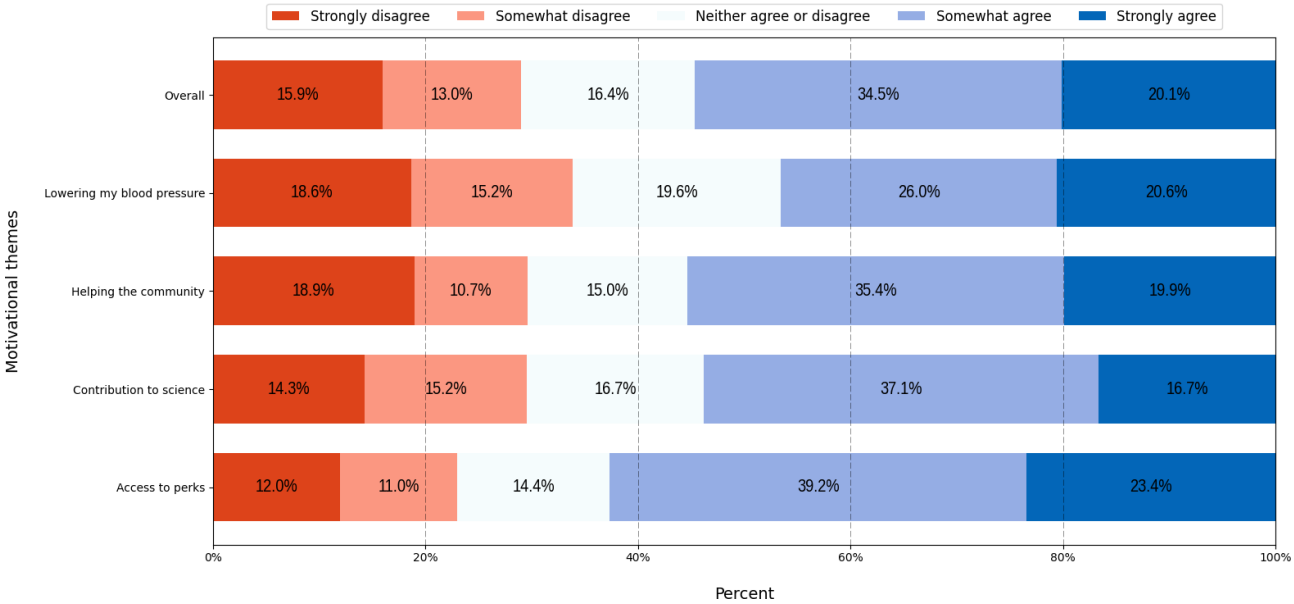

Willingness to Recommend

As shown in Figure 7, willingness to recommend was generally higher than willingness to participate for most themes, ranging from 58.3% to 68.4% (119/204-143/209) of participants somewhat or strongly agreeing. The highest willingness to recommend was for the "access to perks" theme at 68.4% (143/209). On the other hand, 25% (51/204) of respondents who viewed the "lowering my blood pressure" theme were unwilling (somewhat or strongly disagreed) to recommend the advertisement, the highest level of unwillingness, followed by the "contribution to science" theme at 21.9% (46/210).

XSL•FO
RenderX

Figure 7. Percentage of participants' likelihood to recommend clinical trial participation by motivational theme.

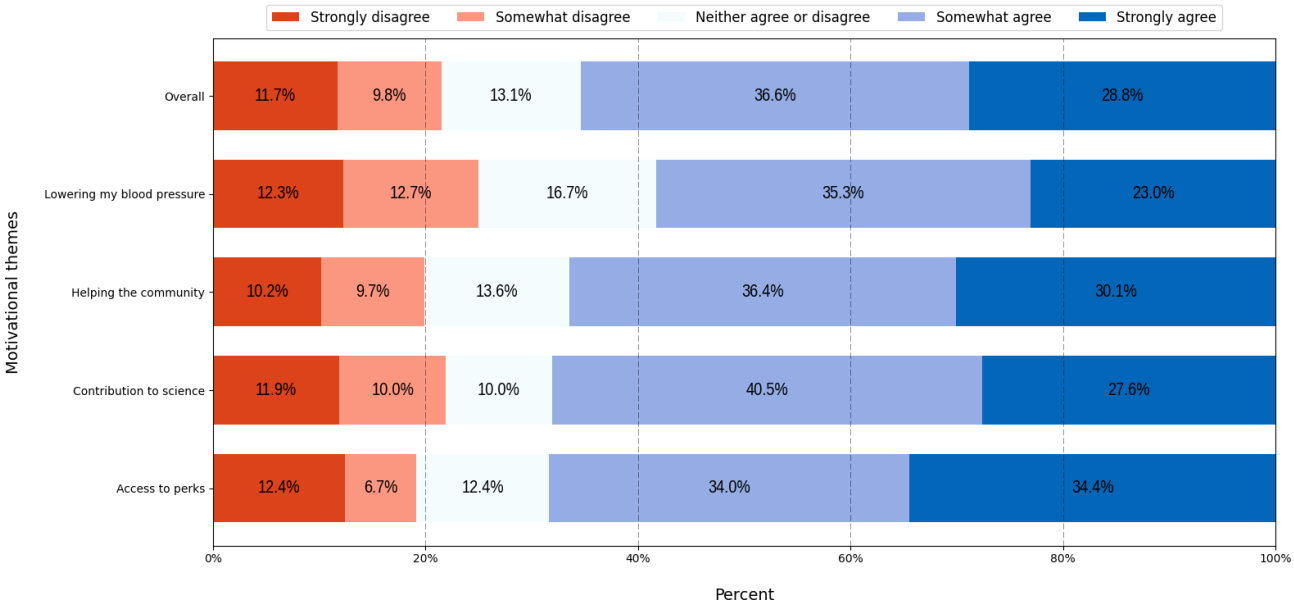

Inferential Statistics

Appeal of Advertisement

The model coefficients for the appeal of advertisements are provided in Table 1. Although all groups were compared, certain groups exhibited higher or lower odds of finding a theme appealing.

XSL•FO
RenderX

Table . Model coefficients for the appeal of motivational themes.

Predictors and effect modifiers	Estimate (β)	Value, OR[a] (95% CI)	*P* value
Motivational themes			
Research	0	1	—
Blood pressure	0.34	1.41 (0.78-2.54)	.25
Community	0.25	1.29 (0.7-2.37)	.41
Perks	0.05	1.05 (0.57-1.94)	.88
Trust levels			
Neutral	0	1	—
Trust	0.41	1.51 (0.80-2.84)	.21
Distrust	0.52	1.68 (0.90-3.14)	.11
High blood pressure diagnosis			
No	0	1	—
Yes	0.50	1.65 (1.20-2.25)	<.001
Age (years)			
18 - 29	0	1	—
30 - 39	0.09	1.09 (0.76-1.57)	.64
40 - 49	0.42	1.52 (1.01-2.29)	.04
50 - 59	0.45	1.57 (1.01-2.43)	.04
60 - 69	−0.10	0.91 (0.51-1.61)	.75
>70	0.10	1.1 (0.44-2.73)	.84
Political view			
Moderate	0	1	—
Liberal	0.45	1.56 (1.14-2.13)	.01
Conservative	0.30	1.35 (0.89-2.03)	.15
Sex			
Female	0	1	—
Male	−0.08	0.93 (0.71-1.21)	.58
Location			
Large city	0	1	—
City	−0.2	0.82 (0.58-1.15)	.24
Rural area	−0.2	0.82 (0.45-1.50)	.52
Town or suburb	−0.54	0.58 (0.41-0.83)	<.001
Motivational themes × trust levels			
Blood pressure × trust	−0.87	0.42 (0.17-1.03)	.06
Community × trust	−0.73	0.48 (0.19-1.21)	.12
Perks × trust	−0.15	0.86 (0.34-2.18)	.76
Blood pressure × distrust	−0.92	0.40 (0.16-0.97)	.04
Community × distrust	−0.15	0.86 (0.35-2.11)	.74
Perks × distrust	0.30	1.35 (0.56-3.30)	.51

[a]OR: odds ratio.

XSL•FO
RenderX

Main Effects

None of the motivational themes or trust levels had a significant effect on appeal. However, there were significant associations with appeal among certain groups. Participants who reported being diagnosed with high blood pressure had 65% higher odds of finding the themes appealing compared to those without such a diagnosis (OR 1.65, 95% CI 1.20 - 2.25; *P*=.002). People aged 40 - 49 years (OR 1.52, 95% CI 1.01 - 2.29; *P*=.04) and 50 - 59 years (OR 1.57, 95% CI 1.01 - 2.43; *P*=.04) had 52% and 57% higher odds, respectively, of finding the themes appealing compared to those aged 18 - 29 years. Participants with liberal political views had 56% higher odds of finding the themes appealing (OR 1.56, 95% CI 1.14 - 2.13; *P*=.005) compared to those with moderate views. Living in a town or suburb was associated with 42% lower odds of finding the themes appealing compared to those living in large cities (OR 0.58, 95% CI 0.41 - 0.83; *P*=.002).

Interaction Between Themes and Trust Levels

While the main effects of motivational themes and trust levels were not significantly associated with appeal, their interaction was significant. There was a significant interaction between the "lowering my blood pressure" theme and the "distrust" level (OR 0.40, 95% CI 0.16 - 0.97; *P*=.04). This indicates that individuals who viewed the "lowering my blood pressure" theme and expressed distrust in health care systems were 60% less likely to find the clinical trial advertisement appealing. Although this interaction was statistically significant, the broad CI suggests uncertainty regarding the precise magnitude of this effect.

Willingness to Participate

The model coefficients for willingness to participate are provided in Table 2. Although all groups were compared, certain groups exhibited higher or lower odds of participating in a clinical trial.

XSL•FO
RenderX

Table . Model coefficients for willingness to participate in clinical trials.

Predictors and effect modifiers	Estimate (β)	Value, OR[a] (95% CI)	*P* value
Motivational themes			
Research	0	1	—
Blood pressure	0.32	1.38 (0.78 - 2.44)	.27
Community	0.16	1.17 (0.65 - 2.10)	.60
Perks	0.29	1.34 (0.75 - 2.39)	.31
Trust levels			
Neutral	0	1	—
Trust	0.51	1.66 (0.92 - 3.00)	.09
Distrust	0.51	1.66 (0.91 - 3.04)	.10
High blood pressure diagnosis			
No	0	1	—
Yes	0.63	1.88 (1.39 - 2.53)	<.001
Age (years)			
18 - 29	0	1	—
30 - 39	0.11	1.11 (0.78 - 1.59)	.56
40 - 49	0.38	1.46 (0.99 - 2.16)	.05
50 - 59	0.55	1.74 (1.14 - 2.65)	.01
60 - 69	−0.14	0.87 (0.50 - 1.50)	.61
>70	−0.03	0.97 (0.42 - 2.24)	.95
Political view			
Moderate	0	1	—
Liberal	0.31	1.37 (1.01 - 1.85)	.04
Conservative	0.48	1.62 (1.09 - 2.40)	.02
Sex			
Female	0	1	—
Male	0.19	1.21 (0.94 - 1.57)	.14
Location			
Large city	0	1	—
City	−0.27	0.77 (0.55 - 1.07)	.12
Rural area	0.11	1.11 (0.62 - 1.99)	.72
Town or suburb	−0.56	0.57 (0.41 - 0.80)	<.001
Motivational themes × trust levels			
Blood pressure × trust	−0.88	0.41 (0.17 - 0.98)	.04
Community × trust	−0.46	0.63 (0.26 - 1.50)	.30
Perks × trust	−0.13	0.88 (0.37 - 2.09)	.77
Blood pressure × distrust	−0.70	0.5 (0.21 - 1.17)	.11
Community × distrust	−0.25	0.78 (0.33 - 1.85)	.57
Perks × distrust	0.31	1.37 (0.58 - 3.20)	.47

[a]OR: odds ratio.

XSL•FO
RenderX

Main Effects

The motivational themes and trust levels did not have a significant effect on willingness to participate. There were significant demographic associations. Participants who reported being diagnosed with high blood pressure (OR 1.88, 95% CI 1.39 - 2.53; *P*<.001) had 88% higher odds of being willing to participate in the clinical trial focused on high blood pressure compared to those who had not been diagnosed. Individuals aged 50 - 59 years had 74% higher odds of willingness to participate in the clinical trial compared to those aged 18 - 29 years (OR 1.74, 95% CI: 1.14 - 2.65, *P*=.01). Participants with both liberal (OR 1.37, 95% CI 1.01 - 1.85; *P*=.04) and conservative (OR 1.62, 95% CI 1.09 - 2.40; *P*=.02) views had 37% and 62% higher odds, respectively, of being willing to participate in the clinical trial focused on high blood pressure compared to those with moderate views. Living in a town or suburb was associated with 43% lower odds of being willing to participate in clinical trials compared to those living in large cities (OR 0.57, 95% CI 0.41 - 0.80; *P*=.001).

Interaction Between Themes and Trust Levels

The interaction between the “lowering my blood pressure” motivational theme and trust level was found to be significant (OR 0.41, 95% CI 0.17 - 0.98; *P*=.04), indicating that individuals who viewed the “lowering my blood pressure” theme and had higher levels of trust were 59% less willing to participate in the clinical trial. Although the interaction was statistically significant, the wide CI indicates uncertainty about the exact size of this effect.

Willingness to Recommend

The model coefficients for willingness to recommend are provided in Table 3. All groups were compared to each other, but certain groups exhibited higher or lower odds of recommending a clinical trial to others.

XSL•FO
RenderX

Table . Model coefficients for willingness to recommend clinical trials.

Predictors and effect modifiers	Estimate (β)	Value, OR[a] (95% CI)	*P* value
Motivational themes			
Research	0	1	—
Blood pressure	0.19	1.21 (0.68 - 2.16)	.51
Community	0.18	1.2 (0.66 - 2.17)	.56
Perks	0.03	1.03 (0.57 - 1.86)	.93
Trust levels			
Neutral	0	1	—
Trust	0.28	1.32 (0.72 - 2.42)	.37
Distrust	0.50	1.65 (0.89 - 3.06)	.11
High blood pressure diagnosis			
No	0	1	—
Yes	0.31	1.37 (1.01 - 1.85)	.04
Age (years)			
18 - 29	0	1	—
30 - 39	−0.09	0.92 (0.64 - 1.32)	.64
40 - 49	0.27	1.31 (0.88 - 1.95)	.19
50 - 59	0.28	1.32 (0.87 - 2.01)	.19
60 - 69	−0.21	0.81 (0.47 - 1.40)	.45
>70	0.44	1.55 (0.64 - 3.72)	.33
Political view			
Moderate	0	1	—
Liberal	0.3	1.35 (1.00 - 1.83)	.05
Conservative	0.38	1.46 (0.99 - 2.16)	.06
Sex			
Female	0	1	—
Male	0.08	1.08 (0.83 - 1.41)	.54
Location			
Large city	0	1	—
City	−0.34	0.71 (0.51 - 1.00)	.05
Rural area	−0.13	0.88 (0.5 - 1.57)	.67
Town or suburb	−0.48	0.62 (0.44 - 0.88)	.01
Motivational themes × trust levels			
Blood pressure × trust	−0.56	0.57 (0.24 - 1.36)	.20
Community × trust	−0.6	0.55 (0.23 - 1.33)	.18
Perks × trust	0.3	1.35 (0.55 - 3.30)	.52
Blood pressure × distrust	−1.03	0.36 (0.15 - 0.85)	.02
Community × distrust	−0.21	0.81 (0.34 - 1.95)	.64
Perks × distrust	0.09	1.09 (0.46 - 2.61)	.84

[a]OR: odds ratio.

XSL•FO
RenderX

Main Effects

Motivational themes and trust levels did not have a significant effect on willingness to recommend the advertised clinical trial, although there were other demographic associations. A diagnosis of high blood pressure was significantly associated with 37% higher odds of recommending the trial, regardless of the themes (OR 1.37, 95% CI 1.01 - 1.85; *P*=.04). Living in a city (OR 0.71, 95% CI 0.51 - 1.00; *P*=.05) and in a town or suburban areas (OR 0.62, 95% CI 0.44 - 0.88; *P*=.006) was associated with 29% and 38% lower odds of recommending clinical trial advertisements, respectively, compared to living in large cities. While the result for individuals living in a city is statistically significant (*P*=.05), the upper bound of the CI is 1, suggesting the effect could range from a decreased willingness to recommend to no effect.

Interaction Between Themes and Trust Levels

The interaction between the "lowering my blood pressure" theme and the "distrust" level was statistically significant (OR 0.36, 95% CI 0.15 - 0.85; *P*=.02). This suggests that individuals who viewed the "lowering my blood pressure" theme and expressed distrust in health care systems were 64% less likely to recommend the clinical trial.

Discussion

While trials such as the Jackson Heart Study [3] and REasons for Geographic And Racial Differences in Stroke study [4] achieved strong Black representation through intensive community engagement, the majority of clinical trials continue to struggle with recruitment diversity [12,13,24]. Our study builds on these successes by providing a scalable methodology for optimizing recruitment messaging across different trial contexts and resource levels.

Principal Findings

This study makes 3 key contributions to clinical trial recruitment science. First, demographic factors (age, health status, and geographic location) are stronger predictors of Black adults' interest in cardiovascular trials than motivational message themes themselves, suggesting demographic segmentation should be prioritized in recruitment planning. Second, we demonstrate complex trust dynamics, whereby the "lowering blood pressure" theme (personal health benefit messaging) was less effective among both low-trust and high-trust individuals. Third, we provide proof of concept for using online surveys as a rapid, cost-effective tool for pretesting recruitment messages, offering a scalable methodology that complements community engagement approaches.

Targeting individuals based solely on racial identity overlooks diversity within racial groups, as factors such as socioeconomic status, health conditions, geographic location, age, and political beliefs shape perspectives and motivations to enroll in clinical trials [19,57,58]. Even within the same demographic group, individuals have different needs and motivations. Our research identifies key considerations impacting how Black adults engage with digital clinical trial advertisements: individual-level demographic influences on advertisement effectiveness and trust and information dynamics. Considering these factors can improve design and marketing strategies, increasing the likelihood that recruitment advertisements will resonate with potential participants and enhance the effectiveness of clinical trial recruitment.

Demographic Influences on Advertisement Effectiveness

Demographic factors such as age, health status, political views, and location significantly influence Black adults' responses to clinical trial advertisements [19,58,59]. Our findings show hypertension-related trial advertisements resonate most with adults aged 40 - 59 years and individuals diagnosed with high blood pressure, suggesting that targeted approaches based on age and relevant health conditions are effective [59]. These individuals, facing age-related health challenges, may actively seek new treatments, increasing their likelihood of participation [44].

While clinical trial advertisements appealed slightly more to Black adults with liberal views, both liberals and conservatives showed willingness to participate. This adds context to how political ideology affects participation, as prior studies [60-62] often overlooked practical considerations driving engagement in underrepresented populations. Our study demonstrates that Black adults' participation decisions are primarily influenced by pragmatic factors such as incentives, safety, trust, access to treatments, and health assurances, rather than political ideology. This suggests prioritizing tangible benefits during recruitment alongside clear, transparent informed consent.

Black adults living in suburban and rural areas were less likely to engage with clinical trial advertisements compared to those in cities, aligning with literature showing that urban areas have more trial sites and health care facilities. Rural and suburban communities remain underrepresented, have limited trial awareness, face logistical challenges such as transportation barriers, and may harbor mistrust [63-66]. Outreach strategies in these areas should address specific access- and trust-related challenges.

Trust and Information Dynamics

For clinical trial recruitment, trust fundamentally shapes how potential participants perceive trial motives and their willingness to participate. This manifests through information credibility, transparency regarding risks and benefits, and perceived trustworthiness of sponsoring organizations [67-69]. Our study revealed that trust in the health care system interacted significantly with the "lowering my blood pressure" theme, indicating that for Black adults, trust remains a critical issue influenced by varying motivations and socioeconomic factors, compounded by historical contexts and personal health care experiences [7]. Given these findings, recruitment strategies should prioritize messaging that builds trust with this population.

Motivational Themes

Black adults indicated reluctance to participate in clinical trials even when they found certain themes appealing. This is reflected in the overall lower levels of agreement in willingness to participate (28.9%) and willingness to recommend (21.5%) compared to the appeal of the themes (12.8%). These results

XSL•FO
RenderX

also imply that additional motivators, such as tangible perks, may be necessary to boost participation, as personal health benefit themes such as "lowering my blood pressure" seem less motivating compared to "access to perks," which appears to be a stronger draw than health benefits alone. Participants also showed a higher willingness to recommend clinical trials with a promise of perks than to participate themselves. This also indicates that participants are more comfortable endorsing or suggesting trials that offer monetary incentives compared to "lowering my blood pressure," indicating that personal health benefits may not be a popular reason for recommending clinical trials.

Notably, the low willingness to participate or recommend the "lowering my blood pressure" theme may be partly due to the fact that 68.4% (567/829) of participants reported never having been diagnosed with high blood pressure, compared to 29.6% (245/829) who reported having been diagnosed.

We recently published the performance of these themes on a GoFresh clinical trial recruitment website [70]. We found that "lowering my blood pressure," "access to perks," and "helping my community" messages had a similar proportion of website visitors who completed the signup form on the GoFresh website. These 3 themes all outperformed the "contribution to science" theme. It is unclear how these themes relate to more relevant outcomes, including clinical trial enrollment and completion. Motivational theme preferences identified in this study might also reflect actual trial enrollment and completion rates. Work to determine these outcomes by theme is ongoing.

Practical Implications for Researchers

Our findings offer actionable recommendations for clinical trial recruitment:

- Prioritize demographic segmentation: demographic factors (age, health status, and location) are stronger predictors than message themes alone. However, certain themes (eg, "access to perks") may resonate particularly well with minority populations and should be strategically incorporated when targeting these groups.
- Use online surveys for rapid message testing: this methodology enables cost-effective, rapid pretesting of recruitment messages before investing in full-scale campaigns, allowing iterative refinement based on empirical data.
- Navigate trust dynamics carefully: the relationship between message type and trust varies by community. Among minoritized populations, practical considerations may outweigh health-benefit appeals, particularly when institutional trust is low.
- Combine quantitative and community-engaged approaches: we developed messages with community advisory board input, then quantitatively tested them. This combined approach ensures both cultural relevance and demographic insights.
- Adapt to your specific context: use this methodology as a framework for testing message effectiveness within your own target population, as recruitment patterns are likely context-dependent.

While our study focuses on optimizing motivational messaging, we acknowledge that the most effective recruitment strategy begins with actively and explicitly inviting Black adults to participate. Recruitment messaging, regardless of how well-optimized, cannot address the fundamental barrier of never being asked. Our findings complement, rather than replace, the critical need for proactive outreach and inclusive eligibility criteria that account for the higher prevalence of comorbidities in Black populations.

Limitations and Future Work

We recognize certain constraints in our study, which we discuss to provide context and ensure a clear interpretation of our results.

This study relied on self-reported intentions rather than actual trial participation behavior. Future research incorporating observed behavioral outcomes could enhance generalizability. Our use of an online panel (Prolific) may limit generalizability to Black adults who lack internet access, are less digitally engaged, or do not participate in online research. The study focused specifically on dietary intervention for blood pressure, and findings may not generalize to other clinical trial types. Future research should explore a broader range of trials involving different medical conditions, interventions, and populations to strengthen applicability. Future work can compare how different audiences respond to messages directly addressing concerns vs neutral messaging to improve communication strategies. We attempted to stratify data by hypertension status, but the skewed distribution (only 245/829, 29.6% observations for the hypertensive group) led to statistical model instability due to sparse data, causing Hessian errors. Simplifying the model would not fully address our RQs, as our goal was to analyze a sample representative of Black adults who might see these advertisements. Future research focused specifically on the hypertensive demographic could examine this group in greater detail, providing more reliable estimates and statistical stability.

A significant limitation is the lack of message variance. We tested 1 message per theme rather than multiple variants. We cannot definitively determine whether effects are attributable to themes or to specific message characteristics. Future research should test multiple message variants per theme to isolate theme-level from message-specific effects [71]. Our quantitative approach risks oversimplifying complex cultural factors that qualitative methods might better capture. Our findings should inform or enhance community-engaged recruitment and not replace it. This methodology served our proof-of-concept goal of rapidly testing recruitment messages before deployment.

While these limitations should be considered, they do not undermine the research validity. The study provides valuable insights that future research can build upon to further enhance the applicability and impact of these findings.

Conclusion

This study contributes new empirical evidence to the science of inclusive trial recruitment and provides an evidence-based methodology to identify effective advertisements for improving participant diversity. It emphasizes the need for thoughtful design of digital clinical trial advertisements targeting Black

XSL•FO
RenderX

adults in the United States. Simply targeting individuals based on racial identity ignores diverse factors such as age, socioeconomic status, location, health conditions, and political beliefs that shape motivations. Effective advertisements must address these differences, offering broadly appealing messages that also meet specific needs. Recruitment messages should be diverse and tailored to address multiple motivations rather than seeking a single, universal driving factor. Addressing trust and cultural sensitivity in recruitment strategies can boost participation, build stronger relationships between health care institutions and communities, and lead to more equitable health research outcomes.

Acknowledgments

The authors are grateful to Maria Sckolnick at the University of Vermont for her guidance. Generative artificial intelligence (AI) tools (such as Claude Opus-4.1 and ChatGPT-3.5) were used only to refine language clarity and grammar in portions of the manuscript. All scientific content, study design, data analysis, interpretation of results, and conclusions were developed by the authors without AI assistance.

Funding

This research was funded by the American Heart Association (award 953782) as part of the Improving Participation Among Diverse Populations in Cardiovascular Clinical Trials (IMPACT) network's Digital Advertisements for Inclusive Trial Accrual (DAITA) project, within the Science of Diversity in Clinical Trials Strategically-Focused Research Network.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

TH contributed to methodology, formal analysis, investigation, data curation, writing the original draft, writing–review and editing, and visualization. JC, ZH, and RANTO contributed to writing–review and editing. SPJ contributed to conceptualization, writing–review and editing, supervision, and funding acquisition. TPL contributed to conceptualization and writing–review and editing. TBP contributed to conceptualization, writing–review and editing, supervision, and funding acquisition. YF contributed to conceptualization, writing–review and editing, supervision, funding acquisition, and project administration.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey design.
[DOCX File, 12 KB - jmir_v28i1e75857_app1.docx]

Multimedia Appendix 2
Flowchart of participant inclusion and exclusion criteria for this study.
[PNG File, 96 KB - jmir_v28i1e75857_app2.png]

Multimedia Appendix 3
Sociodemographic characteristics of the analytical sample and distribution of motivational themes in clinical trial advertisements by demographic factors.
[DOCX File, 11 KB - jmir_v28i1e75857_app3.docx]

References

1. Fryar CD, Ostchega Y, Hales CM, Zhang G, Kruszon-Moran D. Hypertension prevalence and control among adults: United States, 2015-2016. NCHS Data Brief 2017 Oct(289):1-8. [Medline: 29155682]
2. Turner BE, Steinberg JR, Weeks BT, Rodriguez F, Cullen MR. Race/ethnicity reporting and representation in US clinical trials: a cohort study. Lancet Reg Health Am 2022 Jul;11:100252. [doi: 10.1016/j.lana.2022.100252] [Medline: 35875251]
3. Jackson Heart Study. University of Mississippi Medical Center. URL: https://umc.edu/Healthcare/Heart/jackson-heart-study.html [accessed 2025-11-18]
4. Howard VJ, Cushman M, Pulley L, et al. The reasons for geographic and racial differences in stroke study: objectives and design. Neuroepidemiology 2005;25(3):135-143. [doi: 10.1159/000086678] [Medline: 15990444]

XSL•FO
RenderX

5. Steinberg JR, Turner BE, DiTosto JD, et al. Race and ethnicity reporting and representation in obstetrics and gynecology clinical trials and publications from 2007-2020. JAMA Surg 2023 Feb 1;158(2):181-190. [doi: 10.1001/jamasurg.2022.6600] [Medline: 36542396]
6. Loree JM, Anand S, Dasari A, et al. Disparity of race reporting and representation in clinical trials leading to cancer drug approvals from 2008 to 2018. JAMA Oncol 2019 Oct 1;5(10):e191870. [doi: 10.1001/jamaoncol.2019.1870] [Medline: 31415071]
7. Scharff DP, Mathews KJ, Jackson P, Hoffsuemmer J, Martin E, Edwards D. More than Tuskegee: understanding mistrust about research participation. J Health Care Poor Underserved 2010 Aug;21(3):879-897. [doi: 10.1353/hpu.0.0323] [Medline: 20693733]
8. Brandon DT, Isaac LA, LaVeist TA. The legacy of Tuskegee and trust in medical care: is Tuskegee responsible for race differences in mistrust of medical care? J Natl Med Assoc 2005 Jul;97(7):951-956. [Medline: 16080664]
9. Katz RV, Kegeles SS, Kressin NR, et al. The Tuskegee Legacy Project: willingness of minorities to participate in biomedical research. J Health Care Poor Underserved 2006 Nov;17(4):698-715. [doi: 10.1353/hpu.2006.0126] [Medline: 17242525]
10. Nong P, Raj M, Creary M, Kardia SLR, Platt JE. Patient-Reported Experiences of Discrimination in the US Health Care System. JAMA Netw Open 2020 Dec 1;3(12):e2029650. [doi: 10.1001/jamanetworkopen.2020.29650] [Medline: 33320264]
11. Armstrong K, Putt M, Halbert CH, et al. Prior experiences of racial discrimination and racial differences in health care system distrust. Med Care 2013 Feb;51(2):144-150. [doi: 10.1097/MLR.0b013e31827310a1] [Medline: 23222499]
12. Ketchum FB, Erickson CM, Chin NA, et al. What influences the willingness of Blacks and African Americans to enroll in preclinical Alzheimer's disease biomarker research? A Qualitative vignette analysis. J Alzheimers Dis 2022;87(3):1167-1179. [doi: 10.3233/JAD-215521] [Medline: 35466937]
13. George S, Duran N, Norris K. A systematic review of barriers and facilitators to minority research participation among African Americans, Latinos, Asian Americans, and Pacific Islanders. Am J Public Health 2014 Feb;104(2):e16-e31. [doi: 10.2105/AJPH.2013.301706] [Medline: 24328648]
14. NOT-OD-18-014: Revision: NIH Policy and Guidelines on the Inclusion of Women and Minorities as Subjects in Clinical Research. National Institutes of Health. URL: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-18-014.html [accessed 2024-04-08]
15. Gallagher KM, Updegraff JA. Health message framing effects on attitudes, intentions, and behavior: a meta-analytic review. Ann Behav Med 2012 Feb;43(1):101-116. [doi: 10.1007/s12160-011-9308-7] [Medline: 21993844]
16. Rothman AJ, Salovey P, Antone C, Keough K, Martin CD. The influence of message framing on intentions to perform health behaviors. J Exp Soc Psychol 1993 Sep;29(5):408-433. [doi: 10.1006/jesp.1993.1019]
17. Thompson HS, Manning M, Mitchell J, et al. Factors associated with racial/ethnic group-based medical mistrust and perspectives on COVID-19 vaccine trial participation and vaccine uptake in the US. JAMA Netw Open 2021 May 3;4(5):e2111629. [doi: 10.1001/jamanetworkopen.2021.11629] [Medline: 34042990]
18. Taylor LA, Nong P, Platt J. Fifty years of trust research in health care: a synthetic review. Milbank Q 2023 Mar;101(1):126-178. [doi: 10.1111/1468-0009.12598] [Medline: 36689251]
19. Kim JY, Florez M, Botto E, Belgrave X, Grace C, Getz K. The influence of socioeconomic status on individual attitudes and experience with clinical trials. Commun Med (Lond) 2024 Sep 5;4(1):172. [doi: 10.1038/s43856-024-00586-9] [Medline: 39237734]
20. Wilkinson AV, Vasudevan V, Honn SE, Spitz MR, Chamberlain RM. Sociodemographic characteristics, health beliefs, and the accuracy of cancer knowledge. J Cancer Educ 2009;24(1):58-64. [doi: 10.1080/08858190802664834] [Medline: 19259867]
21. Woods-Burnham L, Johnson JR, Hooker SE, Bedell FW, Dorff TB, Kittles RA. The role of diverse populations in US clinical trials. Med 2021 Jan 15;2(1):21-24. [doi: 10.1016/j.medj.2020.12.009] [Medline: 35590131]
22. Why diverse representation in clinical research matters and the current state of representation within the clinical research ecosystem. In: Improving Representation in Clinical Trials and Research: Building Research Equity for Women and Underrepresented Groups: National Academies Press; 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK584396/ [accessed 2026-03-05]
23. Branson RD, Davis K, Butler KL. African Americans' participation in clinical research: importance, barriers, and solutions. Am J Surg 2007 Jan;193(1):32-39. [doi: 10.1016/j.amjsurg.2005.11.007] [Medline: 17188084]
24. Niranjan SJ, Wenzel JA, Martin MY, et al. Perceived institutional barriers among clinical and research professionals: minority participation in oncology clinical trials. JCO Oncol Pract 2021 May;17(5):e666-e675. [doi: 10.1200/OP.20.00970] [Medline: 33974821]
25. Barriers to representation of underrepresented and excluded populations in clinical research. In: Improving Representation in Clinical Trials and Research: Building Research Equity for Women and Underrepresented Groups: National Academies Press; 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK584407/ [accessed 2026-03-05]
26. Garber M, Hanusa BH, Switzer GE, Mellors J, Arnold RM. HIV-infected African Americans are willing to participate in HIV treatment trials. J Gen Intern Med 2007 Jan;22(1):17-42. [doi: 10.1007/s11606-007-0121-8] [Medline: 17351837]
27. Smedley BD, Stith AY, Nelson AR, editors. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care: National Academies Press; 2003. URL: https://www.ncbi.nlm.nih.gov/books/NBK220358/ [accessed 2026-03-05]

XSL•FO
RenderX

28. Taani MH, Zabler B, Fendrich M, Schiffman R. Lessons learned for recruitment and retention of low-income African Americans. Contemp Clin Trials Commun 2020 Mar;17:100533. [doi: 10.1016/j.conctc.2020.100533] [Medline: 32211558]
29. Hamel LM, Penner LA, Albrecht TL, Heath E, Gwede CK, Eggly S. Barriers to clinical trial enrollment in racial and ethnic minority patients with cancer. Cancer Control 2016 Oct;23(4):327-337. [doi: 10.1177/107327481602300404] [Medline: 27842322]
30. Tanner A, Kim SH, Friedman DB, Foster C, Bergeron CD. Promoting clinical research to medically underserved communities: current practices and perceptions about clinical trial recruiting strategies. Contemp Clin Trials 2015 Mar;41:39-44. [doi: 10.1016/j.cct.2014.12.010] [Medline: 25542611]
31. Andrasik MP, Broder GB, Wallace SE, et al. Increasing Black, Indigenous and People of Color participation in clinical trials through community engagement and recruitment goal establishment. PLoS One 2021;16(10):e0258858. [doi: 10.1371/journal.pone.0258858] [Medline: 34665829]
32. Crowe S, Cresswell K, Robertson A, Huby G, Avery A, Sheikh A. The case study approach. BMC Med Res Methodol 2011 Jun 27;11:100. [doi: 10.1186/1471-2288-11-100] [Medline: 21707982]
33. Hung JL, He W, Shen J. Big data analytics for supply chain relationship in banking. Industrial Marketing Management 2020 Apr;86:144-153. [doi: 10.1016/j.indmarman.2019.11.001]
34. Guo C, Chen J. Big data analytics in healthcare. In: Nakamori Y, editor. Knowledge Technology and Systems: Springer Nature; 2023:27-70. [doi: 10.1007/978-981-99-1075-5_2]
35. Berman R, Israeli A. The value of descriptive analytics: evidence from online retailers. Marketing Science 2022 Nov;41(6):1074-1096. [doi: 10.1287/mksc.2022.1352]
36. Tsaltskan V, Nguyen K, Eaglin C, Holers VM, Deane KD, Firestein GS. Comparison of web-based advertising and a social media platform as recruitment tools for underserved and hard-to-reach populations in rheumatology clinical research. ACR Open Rheumatol 2022 Jul;4(7):623-630. [doi: 10.1002/acr2.11448] [Medline: 35536990]
37. Inan OT, Tenaerts P, Prindiville SA, et al. Digitizing clinical trials. NPJ Digit Med 2020;3(1):101. [doi: 10.1038/s41746-020-0302-y] [Medline: 32821856]
38. Bærøe K, Gundersen T, Henden E, Rommetveit K. Can medical algorithms be fair? Three ethical quandaries and one dilemma. BMJ Health Care Inform 2022 Apr;29(1):e100445. [doi: 10.1136/bmjhci-2021-100445] [Medline: 35396245]
39. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
40. Anastasi JK, Capili B, Norton M, McMahon DJ, Marder K. Recruitment and retention of clinical trial participants: understanding motivations of patients with chronic pain and other populations. Front Pain Res (Lausanne) 2023;4:1330937. [doi: 10.3389/fpain.2023.1330937] [Medline: 38606348]
41. Williams CP, Senft Everson N, Shelburne N, Norton WE. Demographic and health behavior factors associated with clinical trial invitation and participation in the United States. JAMA Netw Open 2021 Sep 1;4(9):e2127792. [doi: 10.1001/jamanetworkopen.2021.27792] [Medline: 34586365]
42. Le D, Ozbeki H, Salazar S, Berl M, Turner MM, Price OA. Improving African American women's engagement in clinical research: a systematic review of barriers to participation in clinical trials. J Natl Med Assoc 2022 Jun;114(3):324-339. [doi: 10.1016/j.jnma.2022.02.004] [Medline: 35279325]
43. Cunningham-Erves J, Kusnoor SV, Villalta-Gil V, et al. Development and pilot implementation of guidelines for culturally tailored research recruitment materials for African Americans and Latinos. BMC Med Res Methodol 2022 Sep 24;22(1):248. [doi: 10.1186/s12874-022-01724-4] [Medline: 36153481]
44. Wendler D, Krohmal B, Emanuel EJ, Grady C, Group E. Why patients continue to participate in clinical research. Arch Intern Med 2008 Jun 23;168(12):1294-1299. [doi: 10.1001/archinte.168.12.1294] [Medline: 18574086]
45. Corbie-Smith G, Thomas SB, Williams MV, Moody-Ayers S. Attitudes and beliefs of African Americans toward participation in medical research. J Gen Intern Med 1999 Sep;14(9):537-546. [doi: 10.1046/j.1525-1497.1999.07048.x] [Medline: 10491242]
46. Carrera JS, Brown P, Brody JG, Morello-Frosch R. Research altruism as motivation for participation in community-centered environmental health research. Soc Sci Med 2018 Jan;196:175-181. [doi: 10.1016/j.socscimed.2017.11.028]
47. Largent EA, Eriksen W, Barg FK, Greysen SR, Halpern SD. Participants' perspectives on payment for research participation: A qualitative study. Ethics & Human Research 2022 Nov;44(6):14-22 [FREE Full text] [doi: 10.1002/eahr.500147]
48. Stunkel L, Grady C. More than the money: a review of the literature examining healthy volunteer motivations. Contemp Clin Trials 2011 May;32(3):342-352. [doi: 10.1016/j.cct.2010.12.003] [Medline: 21146635]
49. Siervo M, Lara J, Chowdhury S, Ashor A, Oggioni C, Mathers JC. Effects of the Dietary Approach to Stop Hypertension (DASH) diet on cardiovascular risk factors: a systematic review and meta-analysis. Br J Nutr 2015 Jan 14;113(1):1-15. [doi: 10.1017/S0007114514003341] [Medline: 25430608]
50. Turkson-Ocran RAN, Cluett JL, Fitzpatrick SL, et al. Rationale and design of the Groceries for Black Residents of Boston to Stop Hypertension among Adults without Treated Hypertension (GoFresh) Trial. Am J Hypertens 2023 Apr 15;36(5):256-263. [doi: 10.1093/ajh/hpad008] [Medline: 37061794]
51. Rose A, Peters N, Shea JA, Armstrong K. Development and testing of the health care system distrust scale. J Gen Intern Med 2004 Jan;19(1):57-63. [doi: 10.1111/j.1525-1497.2004.21146.x] [Medline: 14748861]

XSL•FO
RenderX

52. Armstrong K, Rose A, Peters N, Long JA, McMurphy S, Shea JA. Distrust of the health care system and self-reported health in the United States. J Gen Intern Med 2006 Apr;21(4):292-297. [doi: 10.1111/j.1525-1497.2006.00396.x] [Medline: 16686803]
53. Real human data. Ready in minutes.. Prolific. 2024. URL: https://www.prolific.com [accessed 2025-03-03]
54. Yousef M, Rundle-Thiele S, Dietrich T. Advertising appeals effectiveness: a systematic literature review. Health Promot Int 2023 Aug 1;38(4):daab204. [doi: 10.1093/heapro/daab204] [Medline: 34931233]
55. The data science code editor. Posit. URL: https://www.posit.co/ [accessed 2025-03-27]
56. ordinal: regression models for ordinal data. Comprehensive R Archive Network (CRAN). 2024. URL: https://cran.r-project.org/web/packages/ordinal/index.html [accessed 2025-03-27]
57. Unger JM, Cook E, Tai E, Bleyer A. The role of clinical trial participation in cancer research: barriers, evidence, and strategies. Am Soc Clin Oncol Educ Book 2016;35(36):185-198. [doi: 10.1200/EDBK_156686] [Medline: 27249699]
58. Unger JM, Barlow WE, Martin DP, et al. Comparison of survival outcomes among cancer patients treated in and out of clinical trials. J Natl Cancer Inst 2014 Mar;106(3):dju002. [doi: 10.1093/jnci/dju002] [Medline: 24627276]
59. Espeland MA, Crimmins EM, Grossardt BR, et al. Clinical trials targeting aging and age-related multimorbidity. J Gerontol A Biol Sci Med Sci 2017 Mar 1;72(3):355-361. [doi: 10.1093/gerona/glw220] [Medline: 28364543]
60. Gabel M, Gooblar J, Roe CM, Selsor NJ, Morris JC. Political ideology and participation in clinical studies. Alzheimer Dis Assoc Disord 2018;32(3):179-184. [doi: 10.1097/WAD.0000000000000244] [Medline: 29351092]
61. Gabel M, Gooblar J, Roe CM, Morris JC. The ideological divide in confidence in science and participation in medical research. Sci Rep 2021 Feb 4;11(1):3120. [doi: 10.1038/s41598-021-82516-6] [Medline: 33542334]
62. Onyeaka H, Weber DB, Chido-Amajuoyi O, Muoghalu C, Amonoo HL. The influence of political ideology on clinical trial knowledge, invitation, and participation among adults in the United States. Clin Trials 2023 Dec;20(6):708-713. [doi: 10.1177/17407745231178790] [Medline: 37345562]
63. Seidler EM, Keshaviah A, Brown C, Wood E, Granick L, Kimball AB. Geographic distribution of clinical trials may lead to inequities in access. Clin Investig (Lond) 2014 Apr;4(4):373-380. [doi: 10.4155/cli.14.21]
64. Unger JM, Moseley A, Symington B, Chavez-MacGregor M, Ramsey SD, Hershman DL. Geographic distribution and survival outcomes for rural patients with cancer treated in clinical trials. JAMA Netw Open 2018 Aug 3;1(4):e181235. [doi: 10.1001/jamanetworkopen.2018.1235] [Medline: 30646114]
65. Kim SH, Tanner A, Friedman DB, Foster C, Bergeron CD. Barriers to clinical trial participation: a comparison of rural and urban communities in South Carolina. J Community Health 2014 Jun;39(3):562-571. [doi: 10.1007/s10900-013-9798-2] [Medline: 24310703]
66. Dunlop AL, Leroy ZC, Logue KM, Glanz K, Dunlop BW. Preconsent education about research processes improved African Americans' willingness to participate in clinical research. J Clin Epidemiol 2011 Aug;64(8):872-877. [doi: 10.1016/j.jclinepi.2010.11.008] [Medline: 21330104]
67. Morgan SE, Harrison TR, Wright KO, Jia X, Deal B, Malova K. The role of perceived expertise and trustworthiness in research study and clinical trial recruitment: perspectives of clinical research coordinators and African American and Black Caribbean patients. PLoS ONE 2023;18(6):e0275770. [doi: 10.1371/journal.pone.0275770] [Medline: 37342999]
68. Chatio ST, Ganle JK, Adongo PB, et al. Strategies to build trust in the conduct of clinical trials: stakeholders' views in a qualitative study in Ghana. PLOS Glob Public Health 2025;5(4):e0003201. [doi: 10.1371/journal.pgph.0003201] [Medline: 40198690]
69. Fisher JA. Institutional mistrust in the organization of pharmaceutical clinical trials. Med Health Care Philos 2008 Dec;11(4):403-413. [doi: 10.1007/s11019-008-9154-y] [Medline: 18633728]
70. Harry T, Hussain Z, Cao J, et al. Randomized comparison of online motivational themes in clinical trial recruitment. Circ Cardiovasc Qual Outcomes 2025 Dec;18(12):e012945. [doi: 10.1161/CIRCOUTCOMES.125.012945] [Medline: 41201904]
71. Thorson E, Wicks R, Leshner G. Experimental methodology in journalism and mass communication research. Journal Mass Commun Q 2012 Mar;89(1):112-124. [doi: 10.1177/1077699011430066]

Abbreviations

GoFresh: Groceries for Black Residents of Boston to Stop Hypertension
OR: odds ratio
RQ: research question

XSL•FO
RenderX

Edited by A Stone; submitted 12.Apr.2025; peer-reviewed by T Aisha, C Kirkpatrick, Y Zhang; revised version received 25.Nov.2025; accepted 19.Dec.2025; published 19.Mar.2026.

Please cite as:
Harry T, Cao J, Hussain Z, Turkson-Ocran RAN, Juraschek SP, Lahey TP, Plante TB, Feng Y
Guide on Selection of Optimal Motivational Themes for Use in a Clinical Trial Recruiting Black US Adults: Survey Study
J Med Internet Res 2026;28:e75857
URL: https://www.jmir.org/2026/1/e75857
doi:10.2196/75857

© Tamunotonye Harry, Jingyi Cao, Zaib Hussain, Ruth-Alma N Turkson-Ocran, Stephen P Juraschek, Timothy P Lahey, Timothy B Plante, Yuanyuan Feng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Methods for Detecting Suspicious Information From Individual Transactions of Pharmaceutical Products via Twitter (now X): Retrospective Observational Study

Ruoyu Zhang[1], MSc; Kazuko Kimura[2], PhD; Naoko Yoshida[1,2,3], PhD

[1]Kanazawa University, Kanazawa, Ishikawa, Japan
[2]Society for Medicines Security Research, Kanazawa, Ishikawa, Japan
[3]Okayama University, Okayama, Okayama, Japan

Corresponding Author:
Naoko Yoshida, PhD
Okayama University
1-1-1 Tsushima-naka, Kita-ku
Okayama, Okayama, 700-8530
Japan
Phone: 81 86 251 7949
Email: naoko.yoshida@okayama-u.ac.jp

Abstract

Background: Individual transactions involving pharmaceutical products via social networking service (SNS) are considered an inappropriate distribution route and may serve as a guise for illicit business-to-consumer activities. In Japan, individual transactions of pharmaceutical products via the internet are recognized as inappropriate distribution routes, which not only lead to the inappropriate use of pharmaceutical products but also require more active monitoring and guidance from the viewpoint of pharmaceutical security and quality assurance.

Objective: This study aimed to develop a method to accurately detect SNS tweets suspected of involving individual transactions of pharmaceutical products, using text data from Twitter (subsequently rebranded as X), the primary platform for such activities in Japan.

Methods: We applied text mining to 1389 text tweets suspected of involving individual pharmaceutical transactions. Using the hashtag "#Okusuri mogumogu," which was identified through manual searching and is commonly associated with trading psychotropic pharmaceuticals, we collected 7499 tweets posted in 2022 and 6461 tweets posted from January 1 to March 31, 2023, using our web crawler program. After manually categorizing whether each tweet was related to individual pharmaceutical transactions, we extracted words and summarized their occurrences and frequencies using the 2022 dataset. A decision tree model was then generated using the 2022 dataset and validated using the 2023 dataset to evaluate the reliability of detecting transaction-related tweets.

Results: Using web crawling, the number of tweets identified using the hashtag "#Okusuri mogumogu" was 7499 in 2022 and 6461 in the first 3 months of 2023. The crawling results also showed that the number of detectable tweets increased closer to the crawl date, suggesting that SNS tweets may frequently be deleted. From 3228 extracted words in the 2022 dataset, 452 were significantly associated with tweets suspected of involving individual transactions. Highly indicative terms included "kyuu" (request), "yuzuri" (transfer), "DM" (direct message), and transaction-related hashtags. The chi-square automatic interaction detection model demonstrated stable discriminative performance (area under the receiver operating characteristic curve values: training 0.83 and 0.84; Gini coefficient: training 0.65 and test 0.68). The overall accuracy using the 2023 validation dataset was 82.31%, indicating reasonable generalizability despite linguistic fragmentation and the presence of partial word forms characteristic of Japanese text.

Conclusions: Using transaction-related tags, text mining, and machine learning, we identified key terms linked to individual pharmaceutical transactions and developed a predictive model. This approach may aid in preventing inappropriate online transactions of pharmaceutical products.

(J Med Internet Res 2026;28:e91103) doi:10.2196/91103

XSL•FO
RenderX

KEYWORDS

online pharmaceuticals; social networking services; text mining; machine learning; decision-tree analysis

Introduction

Social media platforms, or social networking services (SNSs), are a medium for communicating with others through the internet and are widely used daily. The number of social media users worldwide is projected to increase from 4.59 billion in 2022 to 6.03 billion in 2028, and the number of social media users in Japan is projected to increase from 102 million in 2022 to 113 million in 2027 [1]. Among the SNS platforms most widely used in Japan, the third most popular is Twitter (subsequently rebranded as X), with 66.58 million users [2].

SNS platforms can be used to send out information about oneself, obtain the latest or important information, and even purchase or obtain certain items. In recent years, pharmaceutical E-commerce platforms and online pharmacies have increased, new formats have emerged, and online sales of medicines have become increasingly dynamic. This means that there has been a gradual transfer of people's acquisition of medicines from offline to online. Especially during the COVID-19 pandemic, some situations worldwide make it difficult to go to a health care facility in person to obtain long-term medicines, which has led to a rapid growth in online sales of medicines [3-5].

However, not all patients have access to medicines on the internet through formal sources. In addition to some countries and regions lacking relevant laws and regulations that lead to the unregulated purchase of pharmaceuticals [5], some people access online pharmaceutical products for improper purposes, such as overdosing. Some buyers wish to purchase medicines anonymously, deliberately avoiding formal sources of pharmaceutical products. Fraudulent sellers can take advantage of these situations. Because of its convenience, there is an overabundance of information on the internet that makes improper individual transactions difficult to regulate, and social networking sites have become a breeding ground for some inappropriate transactions. Although common sense should preclude engaging in illegal activities, some individuals prioritize profit and engage in activities that violate the law. There may also be cases where individuals violate the law without knowing that they are doing so. One such case is the unauthorized sale of pharmaceutical products, and we have confirmed that some psychotropic medications are being bought and sold on SNS platforms [6,7]. Such individual transactions represent inappropriate distribution channels that may conceal business-to-consumer activities and pose public health risks.

In Japan, the sale of pharmaceutical products is regulated under the Pharmaceutical and Medical Device Act, which requires appropriate licensing for distribution. Unauthorized sales via SNS sites therefore constitute regulatory violations and present ongoing challenges for pharmaceutical governance and quality assurance.

To address these regulatory challenges, regulatory authorities and platform operators often rely on manual surveillance, keyword-based searches, and user reporting systems to identify suspicious content. In academic research, automated detection methods based on natural language processing and machine learning have been increasingly applied to social media data to identify drug abuse–related discussions, counterfeit drug promotion, and illicit online pharmacy advertisements. Recent studies have further developed machine learning and deep learning–based frameworks for automated detection of relevant content on social media platforms, demonstrating the feasibility and improved performance of such approaches for analyzing health-related discussions and harms using large-scale social media datasets [6,7]. However, the focus of these studies has not been on individual-to-individual transactions of legally approved pharmaceutical products. A recent review also summarized applications, challenges, and methodological trends in pharmacovigilance using social media data [8].

Despite these advances, relatively limited research has specifically focused on unauthorized peer-to-peer transactions of legally manufactured pharmaceutical products within regulated markets such as Japan, representing an important research gap. Among past studies, several limitations remain. First, keyword-based monitoring is prone to both false positives and false negatives because users frequently use abbreviations, euphemisms, or coded language when referring to pharmaceutical products [9,10]. Second, previous studies have primarily focused on detecting discussions of drug misuse or illegal online pharmacy advertising; systematic investigation of individual-to-individual pharmaceutical transactions on SNS platforms is scarce. Furthermore, the generalizability and operational feasibility of these models for routine regulatory monitoring have not always been sufficiently evaluated.

To help strengthen the monitoring of pharmaceutical transactions subject to control to help prevent health hazards to the public, we aimed to develop a method for accurately detecting SNS tweets that are suspected of involving individual transactions for pharmaceutical products using text data from X, the main SNS platform used for these transactions in Japan. In recent years, web crawling and text mining techniques have been increasingly applied to social media data to detect illegal or inappropriate activities. Web crawling is a method in which a web crawler program follows links on the internet to visit websites and duplicate and store information on web pages. This method is used in various fields to collect and use information that appears online. If the web crawler is archiving websites, it copies and saves information as it goes. The archives are usually stored in such a way that they can be viewed, read, and navigated as if they were live on the web, but they are preserved as "snapshots" [11]. Text mining, text data mining, or text analytics is the process of deriving high-quality information from text [12]. Additionally, the chi-square test is used in the analysis of contingency tables when the sample size is large. In simpler terms, this test is primarily used to examine whether two categorical variables (two dimensions of the contingency table) independently influence the test statistic (values within the table) [13]. As the amount of health care megadata continues to grow, increasingly more researchers are using machine learning, which is capable of extracting important

XSL•FO
RenderX

and relevant information hidden within massive amounts of data and making useful predictions based on this information [14-19]. A decision tree is a machine learning technique that identifies clusters of data in which certain features often appear and that generates classification rules for these clusters. In the present study, for holdout validation, all data were randomly divided into training data and test data to evaluate the generalization performance of the model. The analysis method used was chi-square automatic interaction detection (CHAID), which is a type of decision tree analysis used to build a decision tree model based on the chi-square test and *F* test. In the field of pharmacy, decision trees can be used to identify and predict adverse drug reactions [20], medication adherence [21], drug abuse [22], and pharmacy services [23], among others.

Building upon these analytical frameworks, in this study, we aimed to develop a surveillance method for accurately detecting SNS tweets suspected of involving individual pharmaceutical transactions using text data from Twitter.

Methods

Web Crawling and Data Defined

This study was conducted after receiving approval for academic research access to the Twitter application programming interface (API; admission date: November 29, 2022). The keywords used in our search included the hashtag "#Okusuri mogumogu," which is often used for trading in psychotropic pharmaceuticals. We analyzed tweets during 2012-2022 related to the trading of pharmaceuticals between individuals and, through manual searching, identified the hashtag "#Okusuri mogumogu" as a keyword for data collection. We developed a web crawler program (IBM Japan, Tokyo, Japan) for this study and collected information on tweets made and not deleted between January 1, 2022, at 0:00 AM JST and March 31, 2023, at 23:59 PM JST. Tweet data collected via web crawling on Twitter were divided into 2 datasets, one in 2022 and a second in 2023. Each tweet was manually reviewed and classified according to its content and account characteristics to determine whether it was suspected of involving an individual pharmaceutical transaction. The manual classification was based on predefined rule-based criteria developed prior to analysis to ensure systematic and reproducible identification of suspicious tweets. Given the contextual and evolving nature of language used in online pharmaceutical transactions, classification required interpretative assessment beyond the presence of isolated keywords. Accordingly, classification was conducted using a structured framework incorporating multiple indicators of transactional intent. These indicators included: (1) the presence of transaction-related expressions commonly used in Japanese SNS contexts, such as "seeking" (求), indicating a request to obtain pharmaceuticals, or "offering" (譲), indicating willingness to provide these; (2) the mention of pharmaceutical-related terms, including formal drug names, abbreviations, slang expressions, emojis, or drug-related images; (3) indications of contact transfer, such as requests to move to direct messaging or other external platforms; and (4) account-level characteristics suggesting repeated similar tweets or explicit transactional intent in the profile description. The predefined indicators served as structured guidance for classification; however, final determinations were made based on a holistic evaluation of the tweet content and contextual account information rather than reliance on any single indicator. The classification procedure is illustrated schematically in Figure 1.

In accordance with the annotation framework illustrated in Figure 1, 2 reviewers (NY and KK) with expertise in pharmaceutical regulation independently conducted an initial pilot annotation using a randomly selected subset of tweets to refine the classification criteria and ensure consistent interpretation. Interreviewer agreement was assessed to confirm sufficient consistency prior to full-scale annotation. After the agreement was deemed acceptable, the remaining tweets were annotated by one reviewer (ZR) based on the finalized criteria. Ambiguous cases were discussed as needed to maintain consistency. The manually assigned classification (1=suspected transaction; 0=non-transactional) was included in the dataset. Additionally, statistically significant keywords identified in prior text mining analysis were coded as binary variables, indicating the presence "1" or absence "0" of each keyword within a tweet.

All data collection procedures were conducted under approved academic API access, and the study protocol received ethics committee approval, as described below.

XSL•FO
RenderX

Figure 1. Structured decision framework for manual classification of tweets suspected of involving pharmaceutical transaction tweets.

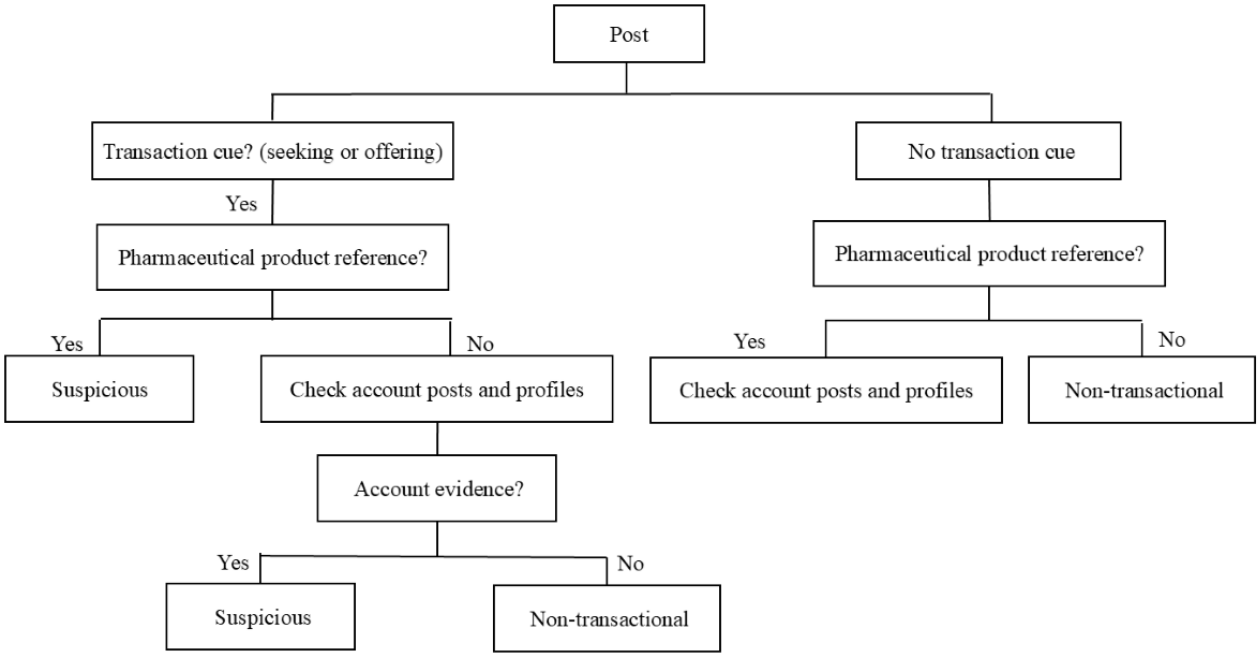

Keyword Extraction: Text Mining

Individual word cloud extraction was performed using MATLAB Text Analytics Toolbox (MathWorks Inc). We did not include words appearing fewer than four times so as to eliminate words with a low frequency. This process allowed us to obtain all words in the collected tweets that could be related to individual pharmaceutical transactions, as well as the number of occurrences of each word.

In the course of our investigation, we noted that a number of terms used in SNS tweets (including words that could not be successfully extracted but that appeared in the text of tweets) were created in response to searches by SNS officials like those from Twitter as well as relevant monitoring services, conducted to determine whether users are engaged in individual transactions that may be in violation of the law, such as using emojis or Kanji characters to represent pharmaceutical products being sold. In fact, our text mining program successfully extracted emojis that were used to refer to pharmaceutical products being bought or sold, or words used in these transactions after processing the emojis in a way that allowed them to be recognized.

Analysis: Word Feature Extraction

To obtain more definitive information on keywords suspected of being related to individual pharmaceutical transactions, an appropriate analysis method is needed. We therefore used our two datasets to conduct the following analyses.

Statistical Analysis

In this research, we used SPSS (version 19.0.0; SPSS Inc) to statistically analyze associations between the frequencies of each word and tweets suspected of involving individual pharmaceutical transactions. We considered P values of <.05 to indicate statistically significant probability. To account for multiple comparisons across all tested words, the Benjamini-Hochberg false discovery rate correction was applied. Only words with P values of <.01 after this correction are presented in Table S1 in Multimedia Appendix 1.

Feature Prediction: Machine Learning

To extract feature words as well, we used the decision tree approach to extract keywords that have correlations with individual transactions involving pharmaceutical products to help guide future surveillance efforts. Decision tree analysis was performed using data mining software (IBM SPSS Modeler version 18.3; IBM Japan). All word-level binary variables derived from the 2022 dataset were entered into the CHAID algorithm without prior manual feature selection. Variable selection occurred internally during tree construction based on chi-square–based node splitting.

In this study, the 2022 dataset was randomly partitioned prior to model construction. For internal evaluation, training:test ratios of 7:3 and training:test:validation ratios of 6:2:2 were applied to assess model stability and performance. Tweets were assigned to partitions without overlap. The final ensemble model, comprising 10 component CHAID trees, was trained exclusively on the 2022 dataset. To prevent overfitting, the maximum tree depth was limited to five levels, and the minimum number of cases required for parent and child nodes was set to 2% and 1% of the total sample size, respectively. Tree growth was terminated when no further statistically significant splits were identified, using a significance level of 0.05 with Bonferroni adjustment. Predictions from the ensemble were generated by aggregating the outputs of the individual component models.

Model validation was then performed using a temporally independent dataset from 2023. Tweets were assigned to datasets strictly based on the posting date to ensure no overlap with the training data. The 2023 dataset was not involved in feature selection, parameter tuning, model training, or ensemble

XSL•FO
RenderX

optimization, and was reserved solely for post hoc evaluation. This approach ensured that the validation process was independent and that information from the training set could not influence the assessment of model performance.

Ethical Considerations

This study was conducted using publicly available social media data collected via academic access to the Twitter API, which was formally approved for academic research (approval date: November 29, 2022). The study protocol was reviewed and approved by the Medical Ethics Committee of Kanazawa University (approval number 2025-466 115205). All analyses were performed using publicly available data, and no attempts were made to identify individual users. The study adhered to Twitter's terms of service and API usage policies. Ethical considerations related to surveillance monitoring and data privacy were incorporated throughout the research process.

Results

Web Crawling and Data Defined

We conducted a manual search of the 2 datasets using the keywords anabolic steroids, testosterone, oxymetholone, and methandienone [9]. We did not find any tweets on Twitter related to individual transactions for these pharmaceuticals. Therefore, we examined whether tweets involving these types of transactions used specific pharmaceutical product names (including commercial product names); we found it important to target certain keywords to detect correlated tweets. The number of tweets identified using the hashtag "#Okusuri mogumogu" in the year 2022 was 7499; this number in the first 3 months of 2023 was 6461. Analysis showed that the number of tweets that could be identified, as well as tweets identified that were related to individual transactions, both tended to increase closer to the time of our data collection (Figure 2).

We found that the closer our search was to the web crawler run date, the more tweets were identified, suggesting that it may be difficult to collect information from SNS tweets because these can be deleted at any time.

Figure 2. Monthly total of tweets. X-axis represents month/year, and Y-axis represents the number of tweets.

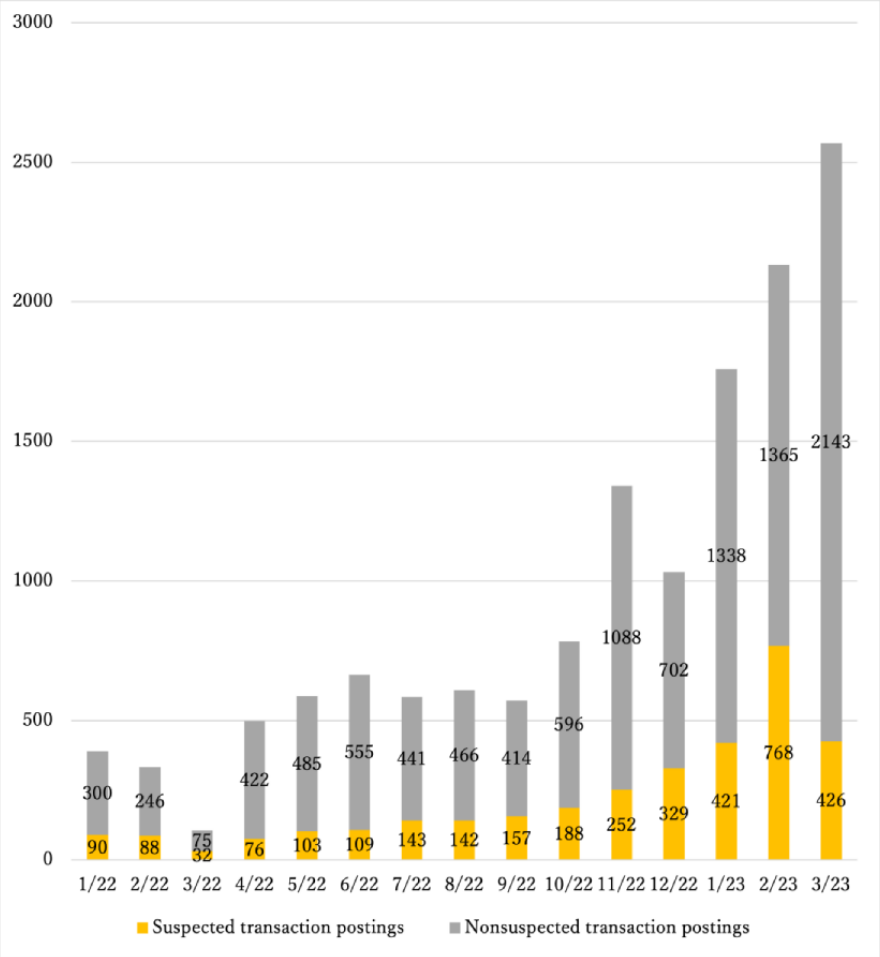

Separation of Transactions and Word Counts

We defined each tweet according to whether its content was suspected of being related to individual transactions involving pharmaceuticals. In 2023, 1625 (25.15%) tweets were classified as suspected individual transaction tweets, and 4836 (74.85%) tweets were not related to individual transactions. The total number of words extracted from the collected tweets was 3228 in 2022 and 2866 in 2023. A cross-tabulation of the 3228 words

XSL•FO
RenderX

extracted from 2022 tweets revealed 452 words that were significantly associated with the presence or absence of individual transactions involving pharmaceutical products in each tweet, as well as the occurrence of each word (Table 1).

Black squares are accounts and information that cannot be disclosed.

In the translation process, we used Hiragana, Katakana, Kanji, and English case correspondences. Because all three expressions are recognized as different characters in Japanese and therefore recognized as different keywords by the text mining program used, the same word appears several times in the table, for case differentiation only.

Kanji characters are uppercase first, Katakana characters are all uppercase, Hiragana characters are lowercase, and English characters are extracted as is.

Each word was ranked according to the proportion of tweets in which it was used in relation to such individual transactions. The most frequently appearing term in tweets potentially related to individual pharmaceutical transactions was #1 “kyuu” (the word for “request” in Japanese), and 81 of 83 (97.59%) tweets in which “kyuu” was used were related to individual transactions for pharmaceutical products. The second most common word was #2 “yuzuri” (the word for “give” or “transfer” in Japanese). Of 103 tweets in which this word was used, 100 (97.09%) were related to individual transactions for pharmaceutical products. The next most frequently extracted words were: #3 “pushcart,” #4 receive,” #5 “profile,” #6 “yuzutt” (like “givin” in English), #7 “list,” #8 “customary,” #9 “enlightenment,” and #10 “requestion,” all of which were closely related to pharmaceutical transactions. The hashtag “#Okusuri mogumogu” was ranked 293rd by our web crawler. Of the total 6682 tweets in which this word was used, 1641 (24.56%) tweets involved individual transactions for pharmaceuticals.

Table 1. Top 10 rankings of extracted and statistically significant words from the 2022 collected tweets list.

Ranking numbers	Extracted words (Japanese)	Total number of tweets in which the word appears (α)[a]	Total number of tweets in which the word appears among tweets with suspected transactions (β)	Percentage of β/α
1	Kyuu (求)	83	81	97.59%
2	Yuzuri (譲り)	103	100	97.09%
3	Pushcart (取引)	28	27	96.43%
4	receive (頂ける)	25	24	96%
5	profile (プロファイル)	23	22	95.65%
6	yuzutt (譲っ)	131	125	95.42%
7	List (リスト)	42	40	95.24%
8	customary (恒例)	21	20	95.24%
9	enlightenment (啓蒙)	20	19	95%
10	requestion (依頼)	19	18	94.74%

[a]Total of 452 words, sorted according to the total number of tweets in which the word appears among tweets with suspected transactions=β/ Total number of tweets in which the word appears =α, in descending order; $P<.01$ for all words.
We have added the split English words where the Japanese characters are split, and we have added the original words that may have been split in the (*) sections as an annotation reference.

Classification and Forecasting Models: Word Extraction

All 7499 tweets in 2022 were manually classified as suspected of being individual transactions involving pharmaceutical products, or “1” (n=1709), or not suspected of being individual transactions involving pharmaceutical products, or “0” (n=5790), as the objective variables. The ratio of training: testing data were adjusted to create a CHAID-based model to predict and classify tweets. As a result, word combinations that characterized tweets related to individual transactions in pharmaceuticals were extracted.

In the testing stage, we found that using ratios of 7:3 and 6:2:2 each for results involving correct and incorrect model decisions and evaluations were not very different, and the 7:3 model was observed to be cleaner. Therefore, considering that there was not much content in the data, using the part that was divided out for verification was not as effective as using this part for learning and testing. Eventually, we chose the 7:3 ratio to develop the model. The following terms were extracted as predictor variables, in order of importance: “KUSURI” (drugs in Japanese; can also be extracted from Okusuri mogumogu), “DM” (direct message), “#Menhera Girl,” “RT” (retweet), “#YAMIAKASANTOTSUNAGARITAI” (*I Want To Connect With Sick Account Users in Japanese), “Mental,” “#Okusuri moguru,” and “Link” (Figure 3).

The term “DM” was extracted in the first level of the tree as one of the words characteristically used in tweets related to individual transactions for pharmaceutical products (Figure 3). We noted a pattern in the characteristics of words used in tweets suspected of being pharmaceutical transactions: inclusion of the term “DM” had a significantly higher rate of involving individual transactions for pharmaceutical products (1208/5231, 23.10%; $P<.001$, chi-square test). This was denoted as “1”

XSL•FO
RenderX

according to our definition for such tweets (as opposed to "0," meaning not involving individual transactions for pharmaceutical products). The accuracy of the model was checked; the accuracy rate was 80.41% for the training data and 82.32% for the test data, with 4206 correctly categorized tweets and 1025 miscategorized tweets (19.59%). As for test data, the model reached an accuracy rate of 82.32% with 1867 correctly categorized tweets and 401 miss tweets (17.68%).

The generalization performance of the model was evaluated as follows: the area under the receiver operating characteristic (ROC) curve (AUC) was 0.83 for the training data and 0.84 for the test data. The Gini coefficient for purity of value was 0.65 for the training data and 0.68 for the test data, showing a good performance (Figure 4).

The ROC curve of the created model was generated using the 2022 dataset.

The AUC, which was used as the evaluation index of the model, was close to 1, indicating that the predictive model discriminated accurately. This model is considered useful for discriminating tweets involving individual pharmaceutical transactions.

Figure 3. Decision tree model generated from the collection of submissions for the year 2022 (training:testing=7:3) (0): tweets and the percentage in which the word appears in tweets not related to transactions (1): number and percentage of tweets in which the word appears among tweets with suspected transactions.

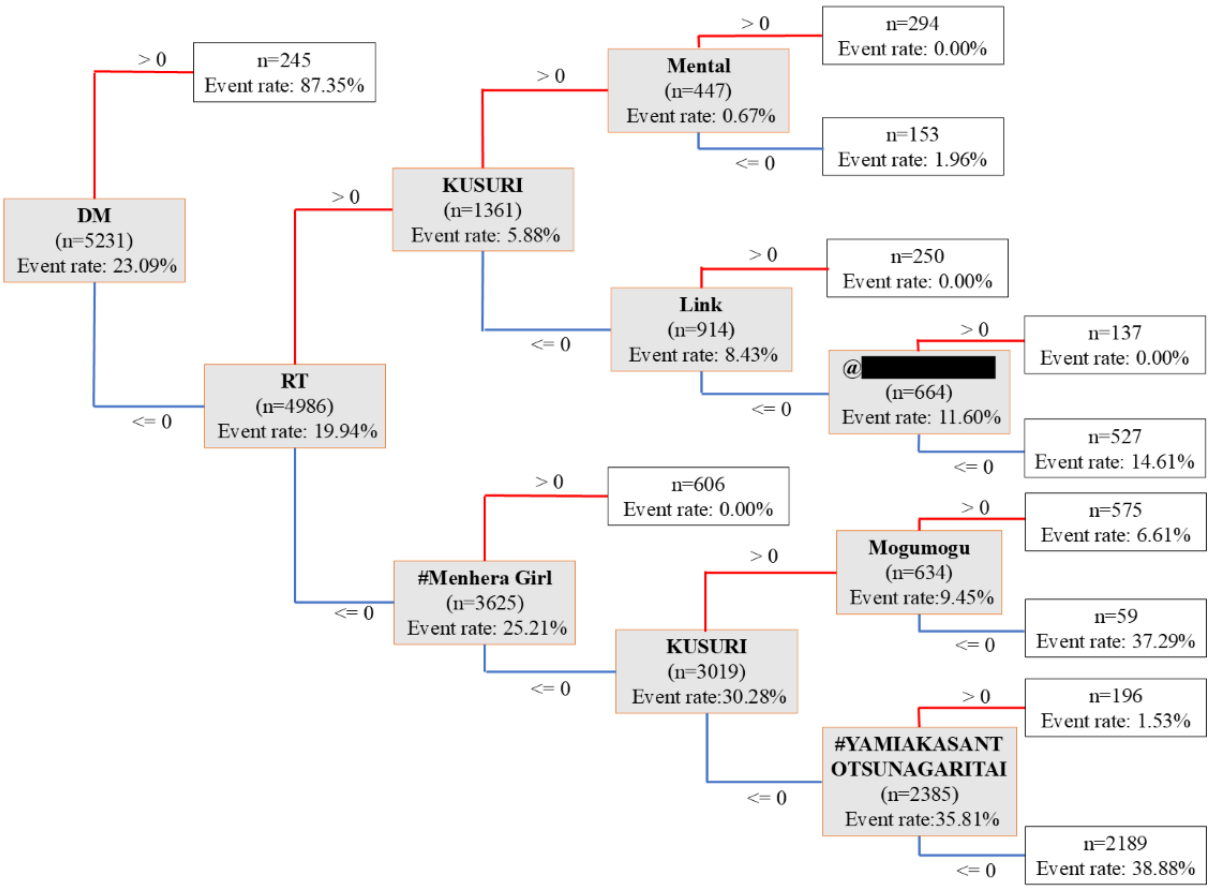

XSL•FO
RenderX

Figure 4. Receiver operating characteristic curve of the created model using the 2022 dataset (7499 cases). Blue line: ROC curve created using the model. FPR: false positive rate; TPR: true positive rate.

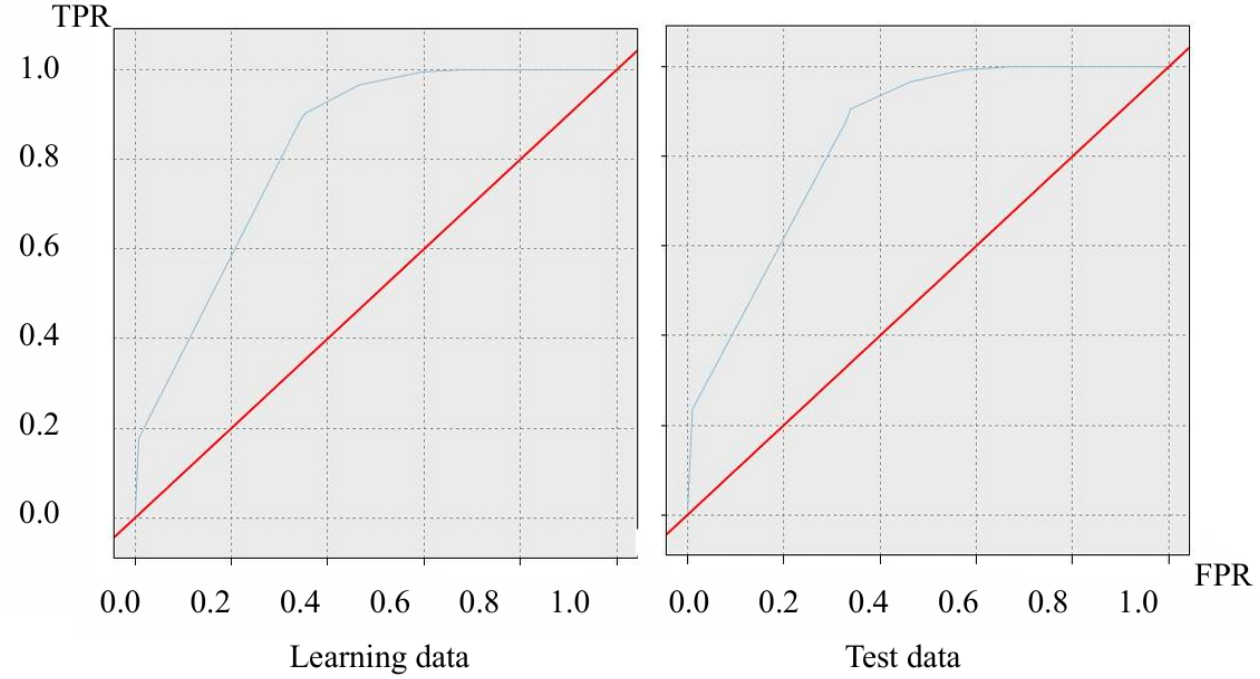

Evaluating the Model's Discriminative Ability

To confirm that the developed model was able to correctly distinguish tweets that may involve individual pharmaceutical product transactions, we used the 2023 dataset to validate the prediction model. The final ensemble model was trained exclusively on the 2022 dataset and was directly applied to the 2023 dataset without retraining, recalibration, or parameter modification. A confusion matrix was constructed using manual annotations of the 2023 dataset as the reference standard (Table 2).

Among 6461 tweets, the model identified 640 true positives and 4678 true negatives, with 158 false positives and 985 false negatives. The overall accuracy was 82.31%. Sensitivity was 39.4%, specificity was 96.73%, the positive predictive value was 80.20%, and the negative predictive value was 82.61%. The relatively high specificity indicates that the model effectively identified tweets not related to individual pharmaceutical transactions, whereas the lower sensitivity suggests that some transaction-related tweets were not detected.

The model trained on 2022 data was validated using temporally independent 2023 data. The AUC was 0.83 and the Gini coefficient was 0.65, indicating stable discriminative ability. Figure 5 shows the ROC curve for this temporal validation.

Table 2. Confusion matrix for model performance on the temporally independent 2023 dataset.

	Predicted suspicious	Predicted nonsuspicious	Total
Manual suspicious	640	985	1625
Manual nonsuspicious	158	4678	4836
Total	798	5663	6461

XSL•FO
RenderX

Figure 5. Receiver operating characteristic curve for temporal validation using the 2023 dataset. FPR: false positive rate; TPR: true positive rate.

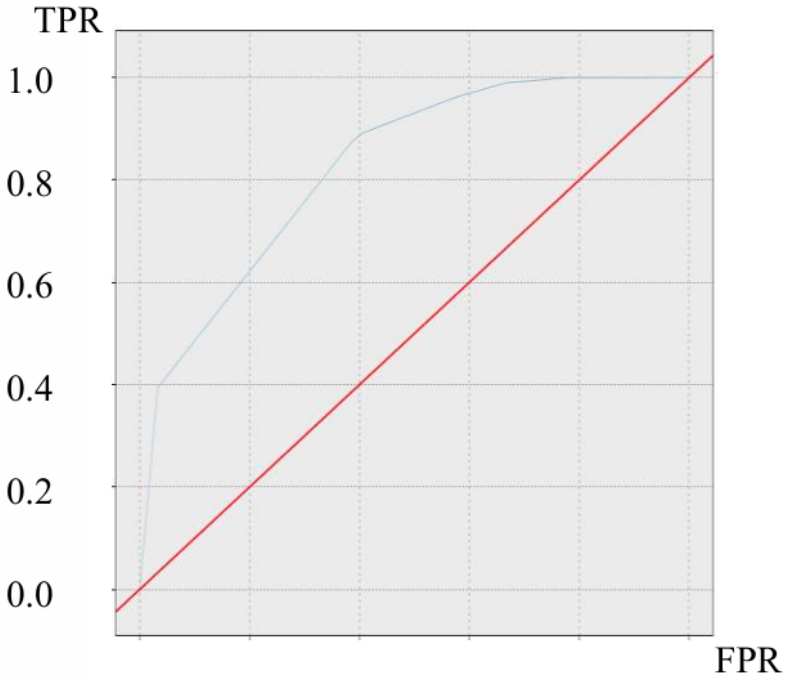

Discussion

Overview

This study demonstrated the feasibility and effectiveness of using machine learning to detect potential individual pharmaceutical transactions on social media. We successfully extracted and ranked keywords appearing frequently in transaction-related tweets and developed a CHAID decision-tree classifier that achieved high accuracy.

Principal Findings and Keyword Extraction

As a widely used platform for individual pharmaceutical transactions, searches using transaction-related terms always retrieve a large amount of potentially relevant tweets in Twitter, making it difficult to accurately extract associated keywords [25,26]. By leveraging machine learning techniques, we successfully extracted representative keywords from large-scale data to characterize tweets potentially related to individual pharmaceutical transactions. Past studies [27,28] have relied on manual searches and statistical analysis conducted within limited time frames to determine whether content was associated with such activities on Japanese social media platforms [10]. Additionally, one previous study summarized different approaches for social media data collection and sampling [26], such as the development of automated collection and coding platforms [29] and random sampling [30]. However, these approaches often require substantial human resources. Moreover, manual sampling always has a limited duration and scalability of surveillance, and random sampling is limited in terms of general applicability [25,27,30]. Consequently, when the volume of content to be investigated is large, manual searches may be too slow and may fail to capture such activities, because transaction-related tweets can quickly disappear owing to deletion, depletion of available stock, or platform policies. We also found that the closer our search was to the web crawler run date, the more tweets were identified (Figure 2), suggesting that it may be difficult to collect information from SNS tweets because these can be deleted at any time. To detect tweets suspected of involving pharmaceutical transactions between individuals, recent data should be collected on a regular basis to obtain a more accurate picture of the actual situation and extract keywords for use in conducting searches with greater precision. Compared with earlier approaches, our use of a web crawler allowed us to capture traces of tweets existing for only a short period at specific time points. The application of machine learning not only simplifies operations and reduces the need for human resources but also enables fast and accurate extraction of relevant information. This substantially improves detection efficiency and lowers the risk of missing relevant tweets that may be rapidly removed after publication. Integrating web crawlers with keyword-driven classification models thereby facilitates fast, scalable, and automated monitoring of pharmaceutical-related activity on social media platforms.

Notably, unlike English, Japanese text comprises a combination of Hiragana, Katakana, and Kanji, which complicates keyword extraction. Keywords derived from Japanese text segments via text mining do not always form complete words. Therefore, important grammatical or semantic information may be lost. Many extracted keywords, such as #6 “yuzutt,” #27 “Dayviko (*Dayvigo)” (Table S1 in Multimedia Appendix 1), are mere fragments of words. Nevertheless, these exhibited statistically significant associations with tweets suspected of being pharmaceutical transactions. This observation explains the presence of numerous fragmented or incomplete terms (Table 1 and Table S1 in Multimedia Appendix 1), which may appear to lack standalone meaning. This also highlights the unique challenges and potential value of text mining in Japanese-language social media, where important linguistic features are often embedded in partial or colloquial expressions. Despite this limitation, we identified numerous keywords

XSL•FO
RenderX

significantly associated with tweets suspected of being individual pharmaceutical transactions. By leveraging these extracted terms, we were able to detect relevant tweets and identify recurring linguistic patterns linked to such activities. Initially, the keyword set was extremely large, making manual searches a cumbersome process. Therefore, we refined the list by focusing on terms with the strongest statistical associations with suspected pharmaceutical transactions. Words such as "DM" (direct message), "kyuu" (request), "yuzuri" (give or transfer), and hashtags like "#Menhera Girl" and "*#Okusuri mogumogu*" were strongly correlated with suspected tweets. The fact that "DM" appeared at the root node of the decision tree underscores the role of private messaging as a key indicator of covert transactional behaviorP (Figure 2).

Model Performance

The ROC curves for the training dataset (2022) presented in Figure 4 and the temporally independent validation dataset (2023) demonstrated strong discriminative performance, with AUC values of 0.83 for training and 0.84 for testing. The ROC curve for the temporally independent 2023 validation dataset is shown in Figure 5, with an AUC of 0.83 and a Gini coefficient of 0.65, indicating stable predictive performance across temporally separated datasets without evident overfitting. This stability suggests that the modeling framework was able to capture relevant linguistic patterns in the training data. Compared with prior studies using similar datasets or methods, the performance achieved in our study is competitive, thereby supporting the validity of the proposed modeling approach. Additionally, the Gini coefficient of 0.65 for training and 0.68 for testing indicates a clear separation between target classes, reinforcing the model's practical applicability in real-world scenarios.

In the evaluation of the generated model using the 2023 dataset in this study, 985 (15.25%) tweets related to individual pharmaceutical transactions were incorrectly predicted to be nontransactional (false negatives), and 158 (2.45%) tweets were incorrectly predicted to be suspicious (false positives), as shown in the confusion matrix (Table 2). The ROC curve (Figure 5) also indicated a stable discriminative ability of the model, as evidenced by the curve lying well above the diagonal reference line. These findings suggest that the classifier was able to distinguish between positive and negative classes, although sensitivity was lower than specificity. The overall shape of the curve and its proximity to the top-left corner further support the stability of the model.

These results indicate that although the model achieved satisfactory overall accuracy and the relatively stable AUC values suggest a stable discriminative ability, the relatively low sensitivity indicates that a substantial proportion of transaction-related tweets were not detected, which might be due to the temporal drift between the 2022 and 2023 datasets. Additional evaluation metrics calculated for the temporally independent validation dataset showed an F_1-score of 0.53 and a balanced accuracy of 0.68, suggesting that although the model maintained relatively high specificity and discriminative performance, the lower sensitivity reduced the overall balance of classification performance under moderately imbalanced class distributions. Moreover, the model achieved satisfactory overall accuracy; further improvement is needed, particularly to enhance sensitivity. The fact that our model also extracted words that are not typically included in such tweets suggests that web crawling using an appropriate combination of these keywords may enable more accurate detection of tweets that may be related to individual transactions for pharmaceutical products. Consequently, despite the low sensitivity, the stable AUC value of the model indicates that the model may serve as a prioritization tool for identifying suspicious tweets prior to manual surveillance or platform supervision instead of using a standalone screening tool.

Cross-Domain Applications and Feasibility of the Multistage Method

Herein, we proposed a multistage method combining web crawling, text mining, and machine learning to detect suspicious information from individual transactions of pharmaceutical products via Twitter. The method features a clear structure and streamlines this process, demonstrating strong applicability for the current research task. Notably, the methods used here have also been widely applied in various research fields and practical settings. Owing to its ability to efficiently capture large volumes of real-time data within a short time frame, web crawling has been applied in finance [31], marketing [32,33], health care [34-36], and medication abuse monitoring [37,38], among others. Leveraging this strength, we could rapidly collect a large dataset of tweets containing information on individual pharmaceutical transactions, from which keywords were extracted for subsequent analysis.

Compared with traditional cross-sectional or prospective surveys, the integration of data mining and machine learning offers several important advantages. These include the capacity to process vast amounts of real-time data, uncover complex patterns without predefined hypotheses, and generate predictive insights. The rapid collection of a large number of tweets substantially reduced the manual workload while providing sufficient data to enhance the accuracy of the machine learning model and support descriptive statistical analyses. Although we primarily used Twitter as the data source, the proposed method is broadly applicable to diverse social networking platforms across different countries, regions, and languages [39]. This not only reflects the technical maturity of the approach, but also reinforces its feasibility and potential for further expansion.

Limitations

We successfully extracted keywords, as well as emojis, related to individual transactions involving pharmaceutical products. However, our study has several important limitations. First, the analysis focused solely on textual data, excluding visual content such as drug images, quantities, or transaction methods, which often accompany suspected tweets. Consequently, tweets relying primarily on images could not be detected. Second, Twitter's rebranding to X and the subsequent API changes affected the continuity of data collection in the later stages of the study. Third, owing to platform restrictions and frequent deletion of tweets, the number of recent tweets we were able to collect was limited, affecting the robustness of model validation. Although external validation using data from other social media platforms

XSL•FO
RenderX

was not conducted, the model was validated using temporally independent data collected in the subsequent year. Given that language patterns and transaction-related expressions evolve over time, this temporal validation offers preliminary evidence of model robustness within the target platform. Future studies should examine cross-platform applicability to further assess generalizability.

Another limitation is that the web crawler collected tweets using only the hashtag #Okusuri mogumogu. This introduces selection bias, as users engaging in pharmaceutical transactions may deliberately avoid commonly monitored hashtags, meaning that the dataset likely represents only a subset of actual activity. However, during observation and organization of the data collected in this study, we noted that sellers aiming for greater exposure often use multiple hashtags. Continuous monitoring permits timely and effective capture of these transactions as hashtags and keywords are updated. Other hashtags have been documented in previous research, and our approach could be extended to include these; however, in this study, we focused on the single hashtag above to demonstrate the feasibility of the surveillance method. Consequently, the model's applicability to tweets lacking the monitored hashtag remains untested; future work should explore generalization across additional keywords and platforms. Importantly, the predictive model did not include the monitored hashtag itself as a feature but was trained on linguistic patterns and statistically significant keywords derived from text mining. Therefore, the classification was based on transactional language characteristics rather than the presence of a specific hashtag.

It should be noted that the current model primarily emphasized specificity, and a limited exploration of classification threshold adjustment may have contributed to low sensitivity during observation.

Despite these limitations, the model's consistent performance across datasets underscores its practical potential in digital surveillance. This approach may provide a useful methodological foundation for future regulatory frameworks aimed at identifying and intervening in illegal pharmaceutical transactions on social networking platforms. Additionally, studies incorporating multiple hashtags and keyword-based crawling approaches may further improve the robustness and practical applicability of the proposed model in the future.

Conclusions

By identifying transaction search tags and using manual searches to supplement web crawling and text mining, together with machine learning, we were able to identify keywords related to individual transactions involving pharmaceutical products and extract characteristic words to successfully create a prediction model. This method of identifying keywords from tweets involving pharmaceuticals suspected of being privately trafficked could help to prevent individual transactions involving pharmaceutical products.

Acknowledgments

The authors acknowledge the Counterfeit and Illegal Drugs Information Center for their assistance in tweet collection; Ms Misa Taguchi and Mr Tokuhisa Kato from MathWorks Japan for providing the text mining program; and Mr Shouhei Yamaguchi and Ms Michie Yamaguchi from IBM Japan for supporting us with the crawling program.

No generative artificial intelligence tools were used in the preparation of this manuscript.

Funding

This work was supported by JST SPRING [Grant Number JPMJSP2135] and by the Ministry of Health, Labor and Welfare of Japan through the Pharmaceuticals and Medical Devices Regulatory Science Policy Research Program [Grant numbers JPMJSP2135, JPMHLW21KC1007].

Data Availability

The datasets are not publicly available owing to the Ministry of Health, Labor and Welfare. Requests can be made to the corresponding author for academic purposes.

Authors' Contributions

Conceptualization: RZ, NY, KK
Methodology: RZ, NY
Data curation: RZ
Formal analysis: NY, KK
Investigation: RZ, NY, KK
Visualization: RZ
Writing – original draft: RZ
Writing – review and editing: RZ, NY, KK
Supervision: NY
Funding acquisition: RZ, NY

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of extracted and statistically significant words from the 2022 collected postings.
[DOCX File , 72 KB - jmir_v28i1e91103_app1.docx]

References

1. 2023 State of the information and telecommunications report summary [Web page in Japanese]. Ministry of Internal Affairs and Communications. 2023. URL: https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r05/html/nd247100.html [accessed 2024-07-10]
2. Summary of Japan and global SNS user rankings! Includes explanations of successful marketing case studies for each SNS platform [Web page in Japanese]. Hottolink. 2024. URL: https://www.hottolink.co.jp/column/20240214_114872/ [accessed 2024-07-10]
3. Mineva G, McCool A, Rogers B. Access to general practice during the COVID-19 pandemic - a cross-sectional view of the opinions of adults who use social media. RRH 2023:8087. [doi: 10.22605/rrh8087]
4. Fittler A, Ambrus T, Serefko A, Smejkalová L, Kijewska A, Szopa A, et al. Attitudes and behaviors regarding online pharmacies in the aftermath of COVID-19 pandemic: at the tipping point towards the new normal. Front Pharmacol 2022;13:1070473 [FREE Full text] [doi: 10.3389/fphar.2022.1070473] [Medline: 36642991]
5. Global ePharmacy market to reach US$244.0 Billion by 2027. Coherent Market Insights. 2024. URL: https://www.coherentmarketinsights.com/press-release/e-pharmacy-market-to-surpass-us-1245-bn-threshold-by-2025-849 [accessed 2024-07-10]
6. Fisher A, Young MM, Payer D, Pacheco K, Dubeau C, Mago V. Automating detection of drug-related harms on social media: machine learning framework. J Med Internet Res 2023;25:e43630 [FREE Full text] [doi: 10.2196/43630] [Medline: 37725410]
7. Sidorov G, Ahmad M, Basile P, Waqas M, Orji R, Batyrshin I. Monitoring opioid-related social media chatter using natural language processing and large language models: temporal analysis. JMIR Infodemiology 2025;5:e77279 [FREE Full text] [doi: 10.2196/77279] [Medline: 41187282]
8. Shruti A, Manoj M, Aman U, Chandrashekhar U. Pharmacovigilance in the digital age. Int J Pharm Sci 2024;2(4):691-711. [doi: 10.5281/zenodo.10980625]
9. Zhang R, Yoshida N. Distribution survey of individual transactions for pharmaceutical products via X (formerly Twitter). Biol Pharm Bull 2025;48(7):1049-1055 [FREE Full text] [doi: 10.1248/bpb.b25-00119] [Medline: 40691070]
10. Hakariya H, Yokoyama N, Lee J, Hakariya A, Ikejiri T. Illicit trade of prescription medications through X (formerly Twitter) in Japan: cross-sectional study. JMIR Form Res 2024;8:e54023 [FREE Full text] [doi: 10.2196/54023] [Medline: 38805262]
11. Masanès J. Web Archiving. Berlin: Springer; 2007.
12. What is text mining? IBM. 2024. URL: https://www.ibm.com/topics/text-mining [accessed 2024-07-10]
13. The chi-square test for independence. University of Utah. 2024. URL: https://soc.utah.edu/sociology3112/chi-square.php [accessed 2024-07-10]
14. Liao SH, Chu PH, Hsiao PY. Data mining techniques and applications – a decade review from 2000 to 2011. Expert Systems with Applications 2012;39(12):11303-11311. [doi: 10.1016/j.eswa.2012.02.063]
15. Koppad SH, Kumar A. Application of big data analytics in healthcare system to predict COPD. 2016 Presented at: Proceedings of the International Conference on Circuit, Power and Computing Technologies (ICCPCT); March 18-19, 2016; Nagercoil, India p. 1-5. [doi: 10.1109/iccpct.2016.7530248]
16. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8-17 [FREE Full text] [doi: 10.1016/j.csbj.2014.11.005] [Medline: 25750696]
17. Dara S, Dhamercherla S, Jadav SS, Babu CM, Ahsan MJ. Machine learning in drug discovery: a review. Artif Intell Rev 2022;55(3):1947-1999 [FREE Full text] [doi: 10.1007/s10462-021-10058-4] [Medline: 34393317]
18. Mei S, Zhang K. A machine learning framework for predicting drug-drug interactions. Sci Rep 2021;11(1):17619 [FREE Full text] [doi: 10.1038/s41598-021-97193-8] [Medline: 34475500]
19. Badwan BA, Liaropoulos G, Kyrodimos E, Skaltsas D, Tsirigos A, Gorgoulis VG. Machine learning approaches to predict drug efficacy and toxicity in oncology. Cell Reports Methods 2023;3(2):100413. [doi: 10.1016/j.crmeth.2023.100413]
20. Hammann F, Gutmann H, Vogt N, Helma C, Drewe J. Prediction of adverse drug reactions using decision tree modeling. Clin Pharmacol Ther 2010;88(1):52-59. [doi: 10.1038/clpt.2009.248] [Medline: 20220749]
21. Bohlmann A, Mostafa J, Kumar M. Machine learning and medication adherence: scoping review. JMIRx Med 2021;2(4):e26993 [FREE Full text] [doi: 10.2196/26993] [Medline: 37725549]

XSL•FO
RenderX

22. Qiao Z, Chai T, Zhang Q. Predicting potential drug abusers using machine learning techniques. 2019 Presented at: Proceedings of the International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS); November 21-24, 2019; Shanghai, China p. 283-286. [doi: 10.1109/iciibms46890.2019.8991550]
23. Ranchon F, Chanoine S, Lambert-Lacroix S, Bosson J, Moreau-Gaudry A, Bedouch P. Development of artificial intelligence powered apps and tools for clinical pharmacy services: a systematic review. Int J Med Inform 2023;172:104983. [doi: 10.1016/j.ijmedinf.2022.104983] [Medline: 36724730]
25. Cavazos-Rehg PA, Krauss M, Fisher SL, Salyer P, Grucza RA, Bierut LJ. Twitter chatter about marijuana. J Adolesc Health 2015;56(2):139-145 [FREE Full text] [doi: 10.1016/j.jadohealth.2014.10.270] [Medline: 25620299]
26. Kazemi DM, Borsari B, Levine MJ, Dooley B. Systematic review of surveillance by social media platforms for illicit drug use. J Public Health (Oxf) 2017;39(4):763-776 [FREE Full text] [doi: 10.1093/pubmed/fdx020] [Medline: 28334848]
27. Chan B, Lopez A, Sarkar U. The canary in the coal mine tweets: social media reveals public perceptions of non-medical use of opioids. PLoS One 2015;10(8):e0135072 [FREE Full text] [doi: 10.1371/journal.pone.0135072] [Medline: 26252774]
28. Myslín M, Zhu SH, Chapman W, Conway M. Using Twitter to examine smoking behavior and perceptions of emerging tobacco products. J Med Internet Res 2013;15(8):e174 [FREE Full text] [doi: 10.2196/jmir.2534] [Medline: 23989137]
29. Cameron D, Smith GA, Daniulaityte R, Sheth AP, Dave D, Chen L, et al. PREDOSE: a semantic web platform for drug abuse epidemiology using social media. J Biomed Inform 2013;46(6):985-997 [FREE Full text] [doi: 10.1016/j.jbi.2013.07.007] [Medline: 23892295]
30. Moreno MA, Briner LR, Williams A, Brockman L, Walker L, Christakis DA. A content analysis of displayed alcohol references on a social networking web site. J Adolesc Health 2010;47(2):168-175 [FREE Full text] [doi: 10.1016/j.jadohealth.2010.01.001] [Medline: 20638009]
31. Zhang X, Fuehres H, Gloor PA. Predicting stock market indicators through Twitter “I hope it is not as bad as I fear”. Procedia Soc Behav Sci 2011;26:55-62. [doi: 10.1016/j.sbspro.2011.10.562]
32. Corallo A, Fortunato L, Spennato A, Errico F, Pedone A. Predicting the consumer's purchase intention of food products. 2020 Presented at: Proceedings of the 9th International Conference on Industrial Technology and Management (ICITM). 2020; February 11-13, 2020; Oxford, UK p. 181-185. [doi: 10.1109/icitm48982.2020.9080404]
33. Jorge O, Pons A, Rius J, Vintró C, Mateo J, Vilaplana J. Increasing online shop revenues with web scraping: a case study for the wine sector. Br Food J 2020;122(11):3383-3401. [doi: 10.1108/bfj-07-2019-0522]
34. Zhang Y, Milinovich G, Xu Z, Bambrick H, Mengersen K, Tong S, et al. Monitoring pertussis infections using internet search queries. Sci Rep 2017;7(1):10437 [FREE Full text] [doi: 10.1038/s41598-017-11195-z] [Medline: 28874880]
35. Wang Y, Kung L, Byrd TA. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. Technol Forecast Soc Change 2018;126:3-13. [doi: 10.1016/j.techfore.2015.12.019]
36. Niketha S, Shreyah JN, Pranauv AJ. Medical ministrations through web scraping. arXiv 2023 Jun 21. [doi: 10.48550/arxiv.2306.12310]
37. Sarker A, O'Connor K, Ginn R, Scotch M, Smith K, Malone D, et al. Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from Twitter. Drug Saf 2016;39(3):231-240 [FREE Full text] [doi: 10.1007/s40264-015-0379-4] [Medline: 26748505]
38. Sarker A, DeRoos A, Perrone J. Mining social media for prescription medication abuse monitoring: a review and proposal for a data-centric framework. J Am Med Inform Assoc 2020;27(2):315-329 [FREE Full text] [doi: 10.1093/jamia/ocz162] [Medline: 31584645]
39. Lee H, Noh EB, Park SJ, Nam HK, Lee TH, Lee GR, et al. COVID-19 Vaccine perception in South Korea: web crawling approach. JMIR Public Health Surveill 2021;7(9):e31409 [FREE Full text] [doi: 10.2196/31409] [Medline: 34348890]

Abbreviations

API: application programming interface
AUC: area under the curve
CHAID: chi-square automatic interaction detection
ROC: receiver operating characteristic
SNS: social networking service

XSL•FO
RenderX

Edited by A Schwartz; submitted 09.Jan.2026; peer-reviewed by B Patel; comments to author 12.Feb.2026; accepted 28.May.2026; published 19.Jun.2026.

Please cite as:
Zhang R, Kimura K, Yoshida N
Methods for Detecting Suspicious Information From Individual Transactions of Pharmaceutical Products via Twitter (now X): Retrospective Observational Study
J Med Internet Res 2026;28:e91103
URL: https://www.jmir.org/2026/1/e91103
doi:10.2196/91103
PMID:

©Ruoyu Zhang, Kazuko Kimura, Naoko Yoshida. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring Attitudes Toward AI-Based Contactless Sensors in Health Among Five Stakeholder Groups: Qualitative Study

Eva Maria Hille[1], Dr theol; Hannah Bleher[1], Dr theol; Birte Coppers[2,3], Dr rer biol hum; Sophie Engelhardt[4], Dr rer soc; Bjoern Eskofier[5,6,7], Prof Dr, PhD; Lorenz Garbe[1], Mag theol; Maria Heckel[8,9], Dr Phil, Dr rer biol hum; Anna-Maria Liphardt[2,3], Dr sportwiss, Dr med habil; Alina Manlig[1], StEx; Nicolas Rohleder[10], Prof Dr, Dr rer nat; David Samhammer[11,12], MA; Carla Spielmann[1], MA; Matthias Braun[1], Prof Dr, Dr theol

[1]Department of (Social) Ethics, University of Bonn, Am Hofgarten 8, Bonn, Germany

[10]Chair of Health Psychology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

[11]Social Science Institute of the EKD, Hannover, Germany

[12]Institute of Sociology, Friedrich-Alexander-Universität Erlangen-Nu¨rnberg, Erlangen, Germany

[2]Department of Medicine 3 - Rheumatology and Immunology, Friedrich-Alexander-Universität Erlangen-Nürnberg and Uniklinikum Erlangen, Erlangen, Germany

[3]Deutsches Zentrum für Immuntherapie (DZI), Friedrich-Alexander-Universität Erlangen-Nürnberg and Uniklinikum Erlangen, Erlangen, Germany

[4]Didaktik der sprachlichen und ästhetischen Fächer (EW 4), Faculty of Education, Universität Hamburg, Hamburg, Germany

[5]Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

[6]Institute of AI for Health, Translational Digital Health Group, Helmholtz Zentrum München, München, Germany

[7]Chair of AI-supported Therapy Decisions, Ludwig-Maximilians-Universität München, Munich, Bavaria, Germany

[8]Department of Palliative Medicine, Universitätsklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

[9]Comprehensive Cancer Center Erlangen-EMN (CCC ER-EMN), Erlangen, Germany

Corresponding Author:
Eva Maria Hille, Dr theol
Department of (Social) Ethics, University of Bonn, Am Hofgarten 8, Bonn, Germany

Abstract

Background: The rapid rise of artificial intelligence–based contactless sensors (AI-CS) is expected to significantly transform how patients are measured, monitored, and understood through a versatile, noninvasive approach to data collection and health assessment. However, there is a lack of empirical research specifically focusing on AI-CS in health. Moreover, existing studies tend to focus on medical or patient perspectives, while neglecting other stakeholders such as researchers, political actors, or the general public.

Objective: The study aims to provide an in-depth empirical ethical analysis and, through a multistakeholder approach, a uniquely comprehensive overview by addressing the research question: what are the attitudes of different stakeholders (patients, health care professionals, researchers, political stakeholders, and the general public) toward AI-CS and their applications in health?

Methods: We conducted a cross-sectional study with 104 participants using a semistructured interview guide. Interviews were analyzed using qualitative content analysis with ATLAS.ti software (ATLAS.ti Scientific Software Development GmbH), following a 3-component model of feelings, thoughts, and behavioral aspects.

Results: The results of the study provide an in-depth analysis of attitudes toward AI-CS in health among different stakeholders. Overall, the results show a high level of openness to AI-CS in health across all stakeholder groups. In terms of feelings and their correlation with behavioral aspects, 2 key trends emerged: first, greater experience and knowledge correlated with a reduced tendency to react emotionally. Second, participants with positive experiences with technologies were generally more open and positive toward contactless sensors. The combined findings on thoughts and behavioral aspects highlighted 3 key tensions—around contact(lessness) and the importance and ambivalence of touch, between protection and surveillance (particularly regarding path- and context-dependency) and between the benefits and challenges of unobtrusiveness (especially in relation to control and governance implications). In addition, the analysis revealed the need for information and consent about AI-CS and clarified possible technical implementations and fields of application.

Conclusions: This study provides a comprehensive and empirically grounded ethical analysis of stakeholder attitudes toward AI-CS in health. The findings offer valuable guidance for the responsible development, implementation, and governance of AI-CS in health care contexts.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e75783) doi:10.2196/75783

KEYWORDS

attitudes; values; contactless sensors; health; artificial intelligence

Introduction

The emergence of new artificial intelligence (AI)-based technologies has opened up new possibilities in health. Among these, sensor technologies have emerged as a versatile tool for data collection and health monitoring. Although still in their infancy, AI-based contactless sensors (AI-CS) are partly used in telemedicine, eHealth, and home-based medical apps and are thus proposed to change protocols and procedures in health. In this study, we aim to better understand and analyze the attitudes of different stakeholders toward the use of AI-CS in health: patients, health care professionals (HCPs), political stakeholders, researchers, and the general public. Of particular interest are the 2 important characteristics of AI-CS (their contactless and unobtrusive character) and how they are perceived in health data collection.

A better understanding of the attitudes of the involved stakeholders is crucial, as various studies have shown that the mere possibility of new technical systems also brings about changes in moral values and sociotechnical practices [1,2]. We decided on the term attitudes, considering it as an umbrella term and referring to the understanding that "[a]ttitudes are a complex combination of things we tend to call personality, beliefs, values, behaviors, and motivations" [3]. We do not limit attitudes to a temporal aspect by dividing them into states and traits. Instead, we follow the distinction "[…] between the inner tendency that is attitude and the evaluative responses that express attitudes" [4], to acknowledge the fact that a person can hold several attitudes at the same time [4,5]. Accordingly, the expression of attitudes is composed of cognitions (often referred to as thoughts or beliefs), affects (often called feelings or emotions), and behaviors (ie, actions) [4].

Several quantitative and some qualitative empirical interview studies have investigated the attitudes of certain stakeholder groups toward AI applications in health in order to predict the likelihood of their acceptance. Attitude studies have mainly focused on opportunities (primarily accuracy [6], efficiency, and enhanced workflow and speed [7]) and challenges such as the lack of human empathy [8-10], data security, and concerns about accuracy and diagnostic quality [11,12], leading to a lack of confidence in the results [13,14]. A survey in the United States found that people often see both the opportunities and challenges, while those who take sides are less optimistic about the potential benefits and more concerned about the potential risks [15,16]. People were more open-minded to the use of AI in cases where human decision-making, care, or other processes are merely enhanced by the technology, rather than replaced [16,17]. In particular, quantitative interview studies with physicians and medical students revealed an eagerness to incorporate AI into health [10,14,18].

The current state of research shows 2 intertwined deficiencies that lead to the research gap we are addressing: first, attitudes toward contactless sensor technologies in health have been less explored than those toward AI-based applications in general. Empirical research into sensor technologies in health has mostly been conducted in the field of smart homes, as a way to enable older adults to remain independent at home by monitoring their health status (eg, [19]). However, sensor technologies in health can serve a multitude of different options. In various areas of health, motion, biometric, and light sensors (radar sensors and photonic sensing cameras, including infrared cameras and depth cameras), most of which operate unobtrusively, are of great importance (eg, [20,21]). Contactless sensor technologies can be used to detect and assess the physiological states and functions of the human body (eg, the musculoskeletal systems with diseases such as inflammatory joint disease), as well as human behavior [22]. Current research reports on continuous radar-based monitoring of stress and mental health and its potential for emergency detection, as well as radar-based machine learning applications for biomedical monitoring, including gait and motion analysis, along with sleep analysis [22]. It is well established that movement is impaired as the disease progresses (eg, in the case of Parkinson disease [22] or rheumatoid arthritis [23]), while disease-specific functional models of the body still need to be explored. Machine learning methods could provide additional potential for better disease monitoring by identifying fine-tuned and disease-specific biomarkers of functional impairment [22]. Although the potential of AI-CS has been outlined, there remains a lack of empirical research, particularly with regard to attitudes.

Second, the few existing studies on attitudes toward sensor technologies, and even more so on AI in health, tend to focus on only 1 or 2 stakeholder groups, with the exception of [6,13,24], which address 3 different stakeholder groups. HCPs [24-26] and patients [6,24,27-31] were most frequently interviewed, sometimes also relatives [6,30], caregivers [24], or visitors [32]. This is a severe problem, as there is ample evidence that data collection and measurement in health and medicine are often a multifactorial and multiactor endeavor [33].

The combination of these two shortcomings, (1) the lack of empirical research on contactless sensor technologies in health and (2) the lack of a broad stakeholder perspective (including nonmedical and nonpatient perspectives, such as researchers or political stakeholders) has not yet been explored in any empirical study. To address both of these deficiencies, we have developed our research question as follows: What are the attitudes of different stakeholders (patients, HCPs, researchers, political stakeholders, and the general public) toward AI-CS and their applications in health?

XSL•FO
RenderX

Methods

Overview

We conducted a cross-sectional study [34] to analyze the attitudes of different stakeholder groups toward AI-CS in health. The research process (Figure 1) is based on the 8 steps of qualitative coding analysis [35]. It follows the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist [36] (Checklist 1).

Figure 1. Scheme of the study design and research process (following the eight research process steps, as formulated byMayring [35]. AI: artificial intelligence.

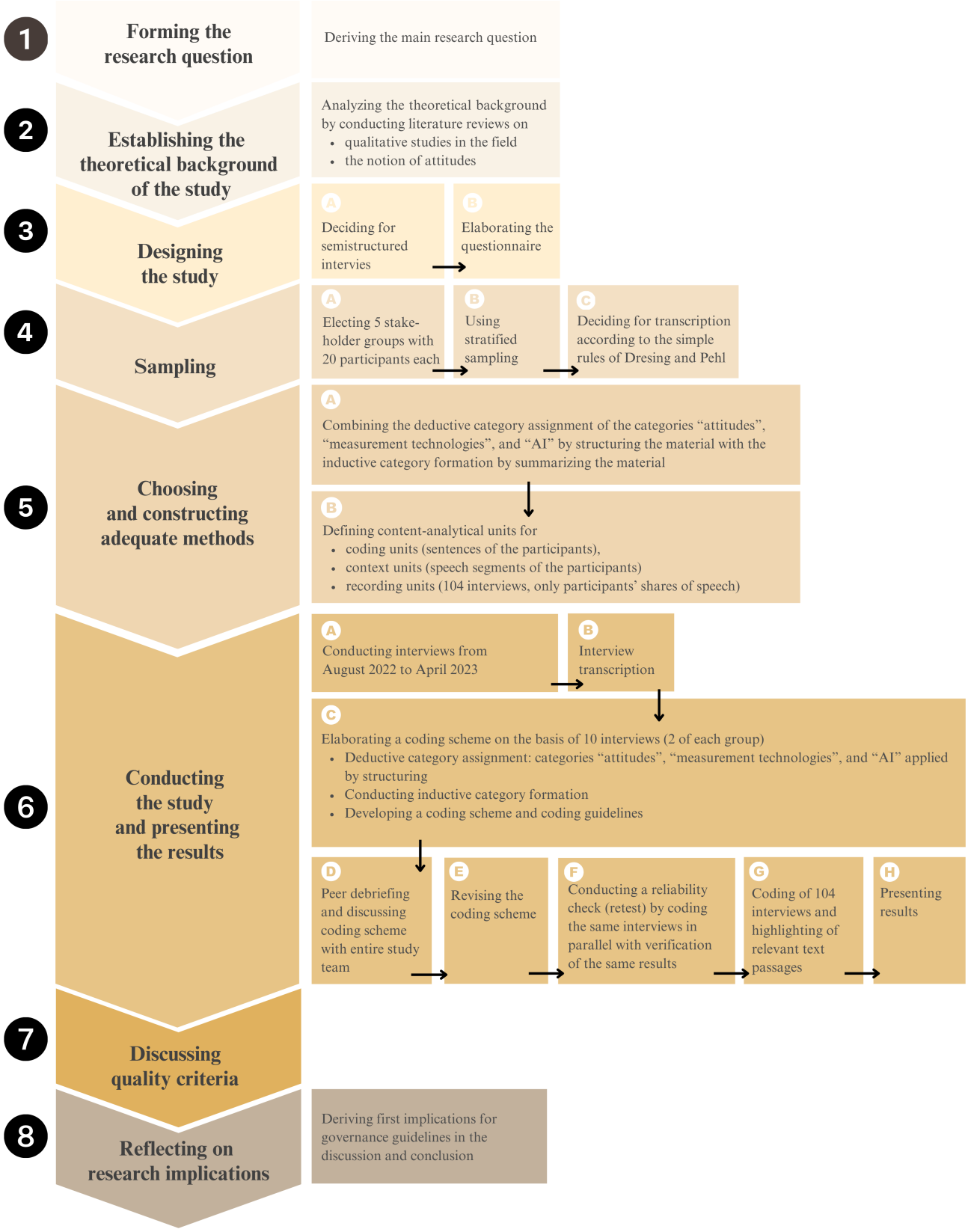

XSL•FO
RenderX

Ethical Considerations

This study was approved by the Institutional Ethics Committee of FAU Erlangen-Nürnberg on December 23, 2020 (478_20 B). Prior to data collection, pilot testing the interview guide allowed us to select AI-CS application examples to minimize anxiety induced by discussions of monitoring at the start of the interview.

Prior to the interview, participants were provided with information and reminded that they could withdraw from the interview at any time. Written informed consent was obtained from all participants. The participants did not receive financial compensation. All data were analyzed pseudonymously.

Sampling and Recruitment

Stakeholder groups were selected through stratified sampling [37] along the idea of an innovation pipeline approach [38]. By understanding the use of AI-CS in its complexity, particularly its impact on different stakeholders and societal levels, we identified 5 relevant stakeholder groups (Figure 1, step 4) that are likely to come into contact with AI-CS, in its development, regulation, and application: (1) patients who might be monitored by using AI-CS, (2) HCPS using these sensors, (3) researchers who develop and investigate AI-CS, (4) political stakeholders who decide on their use, and (5) unaffiliated representatives of the general public. To be included, participants had to be at least 18 years old, able to complete the interview in English or German, willing to sign a consent form, and agree to be audio recorded.

The study was widely promoted nationally and internationally in medical institutions (to reach patients and HCPS), research networks (to target researchers), associations, public places, and social media (to reach political stakeholders and the general public). To ensure a similar number of interviews in each group, patients were approached personally in hospitals and other medical practices in a second phase, and political stakeholders by contacting ministries, nongovernmental organizations, and individuals. To avoid overrepresentation of people with particular sociodemographic characteristics, recruitment took place in deliberately chosen public locations (eg, hospitals and shops). We proceeded with the recruitment process, aiming to recruit at least 20 participants [39] from each stakeholder group in order to ensure that the participants represented a range of perspectives and characteristics (Table 1).

XSL•FO
RenderX

Table . Characteristics and sociodemographic status of participants.

Characteristics	All participants (n=104)	Patients (n=22)	Health care professionals (n=23)	Researchers (n=25)	Political stakeholder (n=8)	General public (n=26)
Gender, n (%)						
Female	57 (55)	11 (50)	19 (83)	11 (44)	2 (25)	14 (54)
Male	47 (45)	11 (50)	4 (17)	14 (56)	6 (75)	12 (46)
Divers	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Age (y), mean (SD; range)	42 (16.6; 20 - 87)	50 (18.6; 22 - 87)	44 (15.2; 26 - 66)	32 (8.9; 24 - 60)	34 (6.7; 26 - 44)	43 (19.1; 20 - 76)
Highest level of education, n (%)						
No degree	1 (1)	1 (5)	0 (0)	0 (0)	0 (0)	0 (0)
Qualified secondary school leaving certificate	4 (4)	2 (9)	0 (0)	0 (0)	0 (0)	2 (8)
Technical baccalaureate	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	1 (4)
A-level equivalent	9 (9)	2 (9)	2 (9)	0 (0)	0 (0)	5 (19)
Completed apprenticeship	7 (7)	4 (18)	3 (13)	0 (0)	0 (0)	0 (0)
Technical college or technical school or trade school	4 (4)	2 (9)	1 (4)	0 (0)	0 (0)	1 (4)
Master craftsman's diploma	1 (1)	1 (5)	0 (0)	0 (0)	0 (0)	0 (0)
Bachelor's degree	11 (11)	1 (5)	4 (17)	0 (0)	0 (0)	6 (23)
Master's degree	47 (45)	9 (41)	5 (22)	19 (76)	5 (63)	9 (35)
Doctorate	19 (18)	0 (0)	8 (35)	6 (24)	3 (38)	2 (8)
Marital status[a], n (%)						
Living in partnerships	71 (68)	16 (73)	17 (74)	17 (68)	7 (88)	14 (54)
Single	30 (29)	6 (27)	5 (22)	7 (28)	1 (13)	11 (42)
Custodian of a minor	8 (8)	0 (0)	1 (4)	2 (8)	4 (50)	1 (4)
Caregiver for relatives	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	1 (4)
Not specified	1 (1)	0 (0)	1 (4)	0 (0)	0 (0)	0 (0)
Professional situation, n (%)						
Working	67 (64)	10 (45)	16 (70)	22 (88)	6 (75)	13 (50)
Apprenticeship	15 (14)	3 (14)	3 (13)	2 (8)	0 (0)	7 (27)
Pension	12 (12)	6 (27)	2 (9)	0 (0)	0 (0)	4 (15)
Started working	10 (10)	1 (5)	2 (9)	5 (20)	2 (25)	0 (0)
On sick leave	4 (4)	2 (9)	1 (4)	0 (0)	0 (0)	1 (4)
Other	3 (3)	0 (0)	0 (0)	0 (0)	0 (0)	3 (12)
Unemployed	2 (2)	0 (0)	1 (4)	0 (0)	0 (0)	1 (4)
Job change	1 (1)	0 (0)	0 (0)	1 (4)	0 (0)	0 (0)

XSL•FO
RenderX

Characteristics	All participants (n=104)	Patients (n=22)	Health care professionals (n=23)	Researchers (n=25)	Political stakeholder (n=8)	General public (n=26)
Health insurance n (%)						
Public health insurance	83 (80)	17 (77)	19 (83)	21 (84)	6 (75)	20 (77)
Private	20 (19)	5 (23)	4 (17)	4 (16)	2 (25)	5 (19)
Other	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	1 (4)
Monthly expenses for technical apps and devices (€)[b], n (%)						
No money	10 (10)	4 (18)	4 (17)	1 (4)	0 (0)	1 (4)
1 - 10	17 (16)	4 (18)	2 (9)	4 (16)	1 (13)	6 (23)
11 - 25	25 (24)	4 (18)	9 (39)	6 (24)	2 (25)	4 (15)
26 - 50	27 (26)	6 (27)	4 (17)	8 (32)	1 (13)	8 (31)
51 - 100	20 (19)	4 (18)	2 (9)	5 (20)	3 (38)	6 (23)
101 - 250	1 (1)	0 (0)	0 (0)	1 (4)	0 (0)	0 (0)
>250	1 (1)	0 (0)	0 (0)	0 (0)	1 (13)	0 (0)
Not specified	3 (3)	0 (0)	2 (9)	0 (0)	0 (0)	1 (4)
Membership of religious community, n (%)						
Yes	56 (54)	13 (59)	10 (43)	12 (48)	5 (63)	16 (62)
Christianity	47 (45)	10 (45)	10 (43)	9 (36)	5 (63)	13 (50)
Islam	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	1 (4)
Not specified	8 (8)	3 (14)	0 (0)	3 (12)	0 (0)	2 (8)
No	43 (41)	7 (32)	12 (52)	13 (52)	3 (38)	8 (31)
Not specified	5 (5)	2 (9)	1 (4)	0 (0)	0 (0)	2 (8)

[a]Multiple entries were possible for marital status and professional situation.

[b]During the data collection period (August 2022 to April 2023), the average conversion rate was approximately 1 EUR = US $1.04.

Data Collection

Based on an extensive literature review, a semistructured interview guide [40] (Multimedia Appendix 1) was designed by identifying the pressing issues regarding AI-CS in health (Figure 1, step 3). The key issues of control [41], consent [42,43], and human-machine interaction [44,45] were incorporated into the interview guide, resulting in 5 different sections (foundation, human-machine interaction, application example, control-related aspects, and consent-related aspects). The interview guide was discussed with other researchers and pilot tested, resulting in the addition of 2 practical examples of the use of AI in health (stress monitoring and palliative care). Due to the abundance of collected data, this paper mainly focuses on the sections relating to the foundation, human-machine interaction, and application examples.

Two questionnaires were used as additional material: the Hospital Anxiety and Depression Scale (HADS) questionnaire [46] was assessed to rule out the possibility that attitudes are influenced by anxiety or depression and a sociodemographic questionnaire (Multimedia Appendix 2) to avoid sampling bias.

The procedure was as follows: the day before the interviews, participants received written information about the study by mail and completed the 2 questionnaires (using Unipark software [Tivian XI GmbH]) [47]. Before the interviews started, the interviewers (EMH female, HB female, and DS male) explained the setting of the study and asked for the participants' informed consent for the audio recording. Participants could choose to be interviewed on-site, remotely, or by telephone to ensure the representation of different age groups and levels of technical knowledge. All interviews were conducted between August 2022 and April 2023 and were pseudonymized and transcribed [48].

Data Analysis

The interviews were analyzed using qualitative content analysis [35] in their original language (German or English). The techniques of content-related and scaling structuring were applied using ATLAS.ti software (ATLAS.ti Scientific Software Development GmbH) [49]. We combined deductive category assignment to refer to the current state of research with inductive category formation to consider the insights highlighted by the participants (Figure 1, step 5). We assigned three deductive categories: (1) attitudes, including the subcategories feelings (codes: positive, negative, and mixed feelings), thoughts (codes: opportunities, challenges, and values), and behavioral aspects (codes: experiences, implications, and preferred actions); (2) different measurement technologies, in order to be able to compare the above-mentioned forms of contactless sensor

XSL•FO
RenderX

technologies with other contactless measurement technologies such as magnetic resonance imaging and X-ray and contact measuring devices such as wearable sensors; and (3) AI in order to gage the participants' understanding to differentiate attitudes toward contactless sensor technologies and data analysis processes.

Based on 10 interviews (2 from each stakeholder group), the initial coding scheme was extended by the interdisciplinary, gender-mixed coding team (EMH [female and theology], LG [male and theology and physics], and CS [female and philosophy]) through the inductive formation of subcodes (Figure 1, step 6) within the deductively derived codes of the categories of feelings, thoughts, and behavioral aspects. The resulting coding scheme (Figure 2) was discussed with the wider research team (including experts from psychology, medicine, and computer science and reflecting researchers' positionality) and refined in a peer debriefing. To identify possible discrepancies, the coding team carried out a parallel coding of 10 interviews, resulting in 95% coding consistency. All 104 interviews were then coded by 2 coders each, and discrepancies were discussed in biweekly meetings.

Figure 2. Code scheme. AI: artificial intelligence.

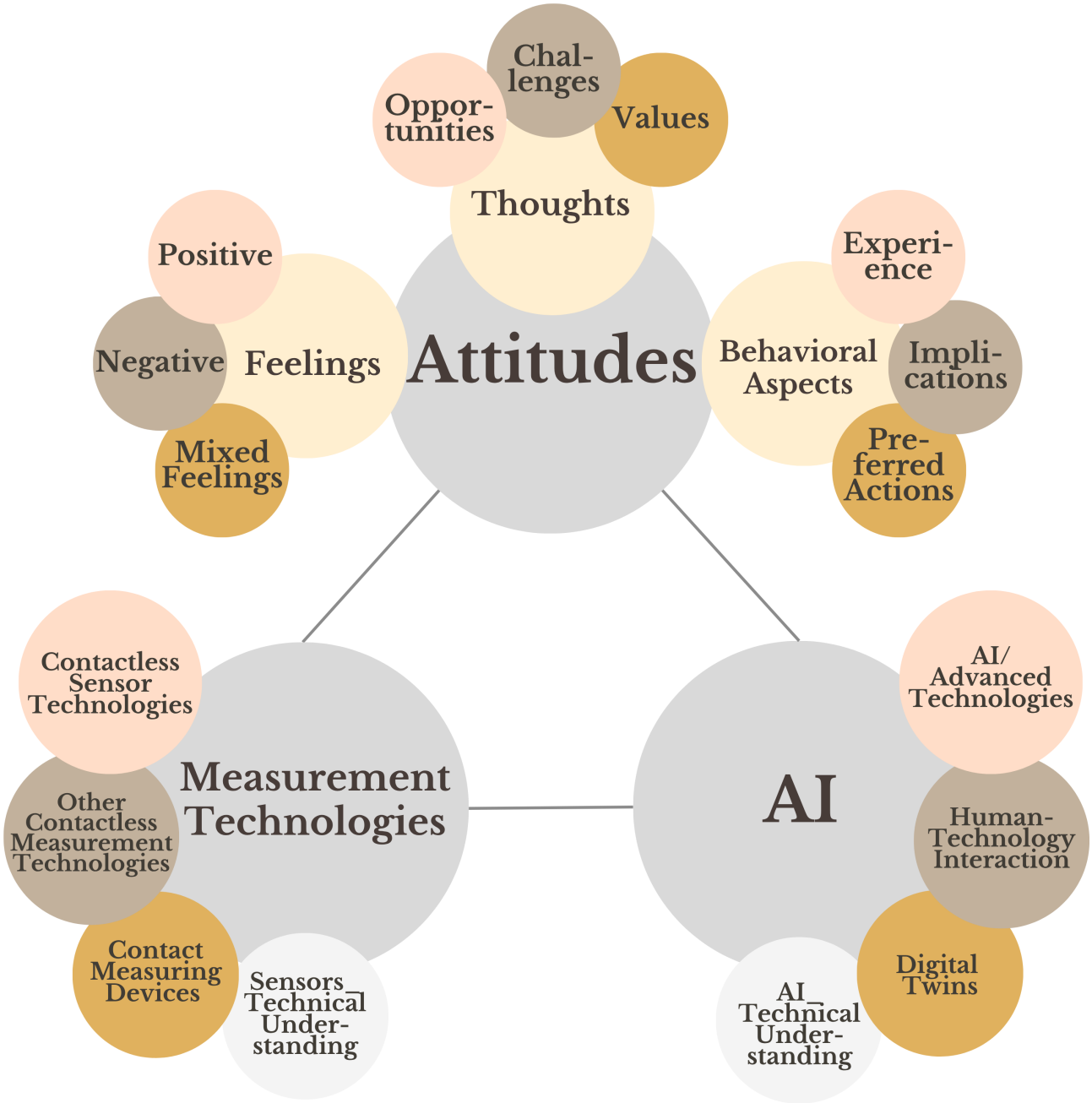

To report the findings, the first author (EMH) carried out a forward translation [50] of the German quotations, which were double-checked by a second researcher (AM) with a teaching qualification in English. In the results, quotations are cited by indicating the initial letter of the stakeholder group (P for patients, H for HCPs, R for researchers, PS for political stakeholders, and G for general public) and a number (the first number indicates an interview number out of 104 interviews, and the second number represents the code number). All results are visualized in an online tool called SENSAI [51], using the

XSL•FO
RenderX

Graph View plug-in of the Obsidian software (Dynalist Inc) [52], which enables readers to access the relevant quotes themselves.

Results

Participants' Characteristics

A total of 104 participants, aged 20 to 87 years, were included in our interview study, 55% (57/104) of whom were female. Table 1 depicts the characteristics and sociodemographic status of all participants. The participants were asked to assign themselves to a group: 22 identified themselves as patients, 23 as HCPS (ie, working as physicians, medical assistants, care specialists, or nurses), 25 as researchers (eg, in computer science and engineering, psychology, radiology, or law science), 8 as political stakeholders, and 26 as members of the general public. A total of 8 participants showed a critical value on the HADS. As these individuals were evenly distributed across the 4 stakeholder groups, no clear correlation can be drawn.

Interviews were conducted in German (n=98) or in English (n=6) on-site at the university building (n=11), remotely (n=89), and by telephone (n=4) and lasted between 23 and 58 minutes.

Feelings

Feelings were first coded deductively into positive, negative, and mixed feelings. Second, inductive subcodes were added within each of these codes (Figure 3).

Figure 3. Overview of feelings identified by different stakeholder groups.

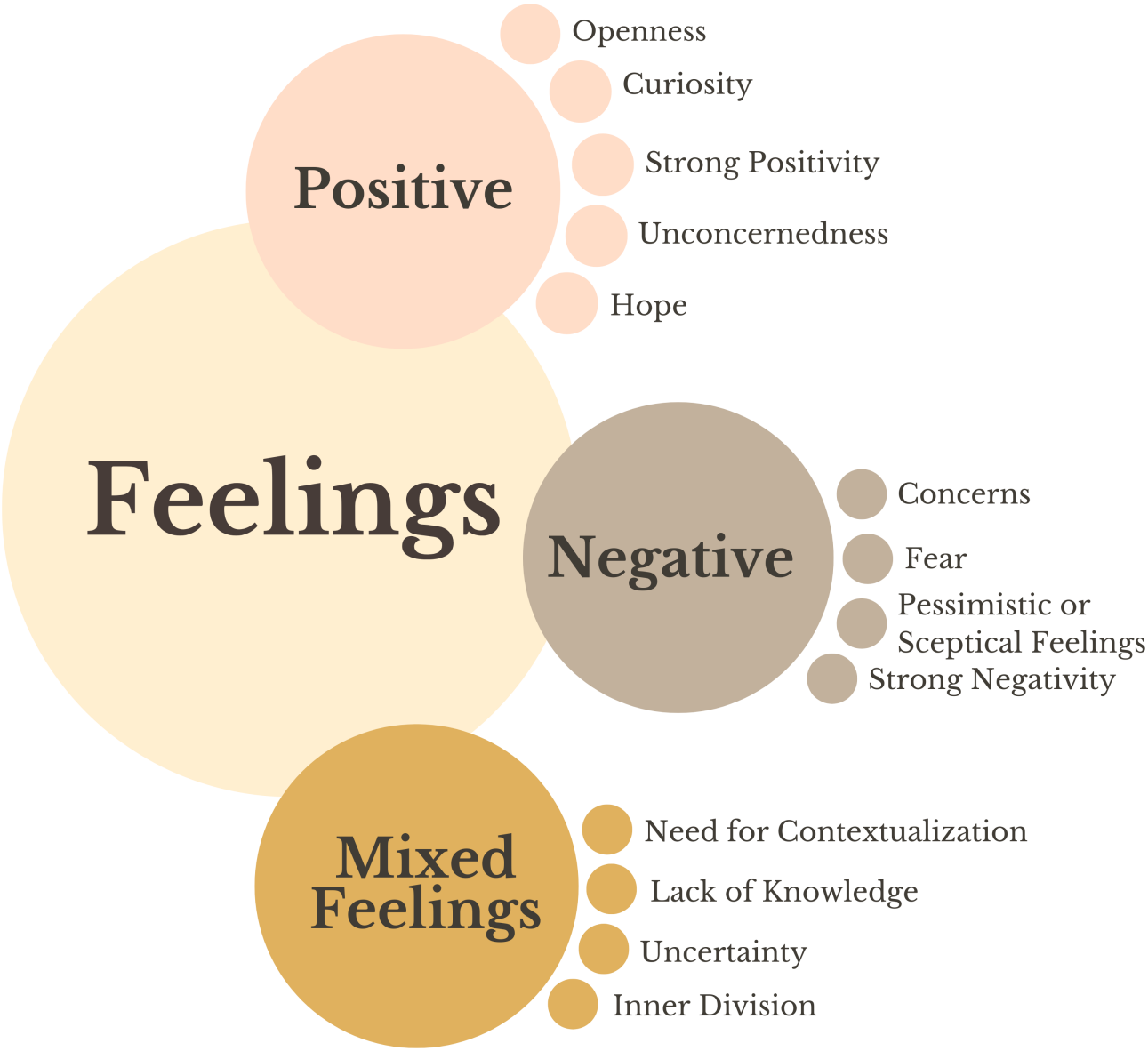

Positive

Participants expressed a great openness toward AI-CS (eg, PS97:59) (; Multimedia Appendix 3). In addition to the willingness to engage with the application, different forms of curiosity were expressed: the overall interest in sensors and their technical possibilities (P91:76, H83:26, R31:48, PS97:73, and G48:106), the specific expectation of advantages with personal or general benefits (P96:30, H74:54, R46:165, PS100:48, and G64:211) or hopes for the improvement of personal or general circumstances in the clinical context and health care (P58:31, P70:190, H74:54, PS102:153, and G7:260).

XSL•FO
RenderX

Strong positivity was expressed in all stakeholder groups (eg, G4:72) and linked to individual experiences, "I have never had any negative experiences […]" (G37:96). Some participants mentioned their unconcernedness about using sensors in health (PS100:83), "[…] I don't see any problem with that at all" (P79:45).

Notions of hope were expressed as positive feelings, with particular findings in the different groups. Patients hoped that the use of AI-CS could have an immediate (P45:87), but also long-term benefit (P9:130). With regard to treatment, patients saw no drawbacks of contactless over other sensory systems as long as it is conducive to successful treatment (P22:107). HCPS also expressed their hope (H83:161), but with a strong focus on the application in treatment and care (H39:150). Researchers rarely expressed their hope (R25:13), as they did not perceive any substantial difference between contactless and contact sensors, "this is simply […] a different sensor […]" (R16:19). Political stakeholders associated clear hopes with AI-CS in regard to their use in health in general (PS102:153) and in specific areas (eg, palliative care PS100:14). The general public rarely expressed specific hopes (G47:25); rather, curiosity about the sensor technology prevailed (G67:139).

Negative

In all stakeholder groups, concerns were particularly emphasized in relation to AI-CS ("problematic" P9:237; "difficult" R25:9, H85:11, G48:16; "would not feel comfortable" R38:32) (; Multimedia Appendix 4). Where these concerns were assumed as a possible outcome, rejection occurred ("don't need it" P35:115, "don't want that" H44:22, "would reject it" R46:32, and "would fight against it" G3:123).

Aspects of fear ("discomfort" R38:32, "worrying" PS101:28, and "not daring" G37:28) were mentioned less frequently, and either regarding specific developments ("I'm afraid this will eventually replace me […]." H74:155), partial aspects ("Because I'm […] a bit afraid of this electrosmog." P94:7) or the technology as a whole ("Sometimes I find that frightening." H66:110).

Pessimistic or skeptical feelings were expressed regarding the potential future effects (H28:12, G20:65, and PS101:138) or the technical feasibility, "[b]ut have we already reached the point where today's data can be used to predict conditions in a few years' time?" (P22:93). HCPS, in particular, raised fears that related to, among other things, the possible replacement of their own workforce (H74:155) or to the possibility of unconscious misuse (H66:110 and PS101:28) in sensitive environments (palliative care H75:41 and nonmedical H77:90).

Strong negativity occurred in the form of intense rejection, concern, or fear ("quite terrible" G23:142, "highly critical" H78:108, and "great danger" G24:71), relating to specific effects (H78:108 and R13:168) and areas: "[…] for something as vague as stress and depression, I find myself kind of flinching and thinking, oh, this is dangerous" (P95:23). The strong polarity within the stakeholder groups is particularly striking: the most frequent negative feeling expressed among patients was an associated potential for misuse (P73:32 and P91:64) as well as non–health-related use by the general public (economic G48:112, public G37:138, and work-related G23:20). The other stakeholder groups rarely came forward with strong negative opinions (H78:108, R46:32, and PS101:116), presumably due to a stronger focus on the application. It is striking that researchers expressed very few negative feelings.

Mixed Feelings

A strong need for contextualization came up in all stakeholder groups, referring to the type of sensor technologies (R15:138), the area of application (H83:31, PS100:28, and G18:15), and other effects (P96:31) (; Multimedia Appendix 5). Patients emphasized the need for contextualization (P56:24) also by noting that the attitude is strongly situational: "I don't think anyone can say, oh, I'll do it no matter what. It's always a matter of the moment when the question is asked […] am I sure I'm comfortable with it or not?" (P70:107).

Related to that, there was a frequent emphasis on the lack of knowledge ("I can't really imagine what they might be doing, because I can't think of anything else to use them for, other than the practical examples from before." P73:116) or experience ("And personally, I find it difficult to put myself in that position." R46:152). Often, this was combined with a great level of uncertainty: "I haven't made up my mind […] on that for now" (G59:58). Uncertainty was expressed, especially regarding the technical feasibility and the possible effects: "Uh, in stress research, it's not really clear to me what's going to come from […] knowing […] whether a person is stressed or not. […] I don't have any idea yet how […] it will help us with diagnostics […]" (H88:55).

The aforementioned feelings resulted in an expression of inner division ("divided" G67:127, "ambivalent" R54:145, and "torn" H89:82) in varying degrees. HCPS put a strong emphasis on feelings of inner division between potential benefits (H75:3) and misuse (H66:193 and H72:34). Political stakeholders perceived the advantages of AI-CS, but contrasted them with possible challenges for the individual or society, "[s]o there are two hearts beating in my chest, you know? […] If you realize later that you could also do this and that cool thing with the data collected […] you really have to think very carefully about whether that really serves its purpose" (PS101:91). Members of the general public perceived a trade-off between collective medical progress and personal data protection (G18:15), also resulting in inner division (G67:127). Remarkably, patients expressed no inner division.

Thoughts

Participants identified various opportunities, challenges, and values for AI-CS, often weighing different opportunities and challenges against each other. The range and breadth of opportunities and challenges identified by stakeholders were inductively categorized into medical, technical, economic, sociopolitical, and regulatory domains.

Medical Opportunities and Challenges

Medical opportunities and challenges were identified by all stakeholder groups in relation to different medical functions. Concrete medical challenges and opportunities according to the

XSL•FO
RenderX

different medical specialties can be found in Multimedia Appendix 6.

All stakeholder groups highlighted the opportunities in terms of monitoring, diagnostics, treatment, prevention, and research as very important (Multimedia Appendix 7). First, respondents saw medical opportunities in improving (home) monitoring, particularly through continuous and long-term monitoring (R25:16). All stakeholder groups pointed out the benefits of improved patient self-monitoring in order to be better informed about their current situation and to develop a better sense of their own body (R50:19). Second, for diagnostics, this means that diseases and changes in the body can be detected at an early stage (R15:117). Third, by providing a better overview of the course of treatment, a more targeted treatment and better preparation for focus on patient consultations are enabled (H71:22). Fourth, participants mentioned that prevention could potentially play a greater role in health systems due to a changed understanding of health and disease (H74:141). Fifth, medical progress and research could be advanced, especially through individualization and the linking of contexts.

The aspect of potential harm to patients—through missed alarms despite medical urgency (PS103:72) in monitoring, through incorrect diagnosis or treatment—cuts across all challenges (Multimedia Appendix 8). With regard to monitoring, a researcher described their uncertainty about medical harm caused by data transmission (R27:11). In the field of diagnostics, "data that would be unproblematic [could] become a problem just by being collected" (H65:127). One researcher asked,

> *What is the normal range? What is outside? And there I think it is a bit critical to say or to define what is normal and what is abnormal. And maybe to reflect that back to people, that they are also in the abnormal range, and what does that do to people psychologically?* [(R54:27)]

Technical Opportunities and Challenges

In terms of implementation, use, data, and advances in medical technology, various technical opportunities were identified (Multimedia Appendix 9). Implementation is facilitated by the easy and time-saving installation and the small size of the device (P45:10). The avoidance of cabling in general was highlighted, particularly by HCPS, as a great opportunity for intensive care, palliative care, patient transport, pediatrics, and dealing with older adults. The portable use of sensors enables automatic and continuous long-term measurement and the collection of comparative data outside of clinical hours: "[Y]ou know somehow, at the moment of the examination it suddenly doesn't hurt anymore or it's somehow a special situation" (PS99:12). Patient passivity eliminates missed measurements and typing errors (which can occur when information is entered manually): It's "[…] just plug and play. It's the best you can get. […] someone does the monitoring and cares for you in the background […]" (PS102:86). Other possible measurement distortions such as the white coat syndrome, unfamiliar measurement environments, touching the patient or study participant, and others are reduced or even avoided by the unconscious nature of the measurement (H65:5 and G4:106). In terms of the data generated, contactless measurement is perceived to contribute to more accurate, reliable, and objective data (R63:168). Four out of 5 stakeholder groups highlighted the potential to measure what cannot yet be measured, such as micromovements in joints, as a medical-technical progress. Patients, HCPS, and researchers described the possibility of measuring and detecting diseases based on large movements and movement patterns, but also small ones, such as tiny muscle movements invisible to the naked eye. The ability to gather information about conditions that previously could only be asked about verbally, such as anxiety, restlessness, and pain, is considered great potential (R34:44).

Across all challenges, all stakeholder groups placed particular emphasis on the technical challenges (Multimedia Appendix 10). Concerning the development of AI-CS, in particular, the set-up and the lack of technical expertise in using the sensors (eg, lack of digital literacy among older patients, H89:77) were considered to potentially pose problems for the implementation in health (H39:23). In terms of use, an important consideration was how to ensure that the device worked smoothly, without bycatch or interference (R38:20). At the same time, contactless sensors were described as having limited alignment capability (R32:107), and their accuracy and reliability were questioned several times (R43:1). With regard to the collected data, the significance of the data were described as limited due to the individuality of the parameters, which makes comparisons difficult (P73:172).

Economic Opportunities and Challenges

It is striking that a common motif across all economic opportunities is savings of time, money, and resources (Multimedia Appendix 11). In terms of health systems, time savings through self-measurements at home, more efficient doctor visits, and cost savings through the early detection and reduction of treatment costs were mentioned. These opportunities are expected to relieve HCPs by making their work more efficient and easier, as there is no need for extensive training to operate the equipment (R38:37). In terms of product-specific economic opportunities, participants thought that contactless sensors could be produced and achieve widespread coverage more quickly due to the low cost of the devices (R15:43 and G53:93).

With regard to HCPS and patients, fears of overburdening, additional pressure, and extra work were expressed as challenges (Multimedia Appendix 12). HCPs would have to learn to use the systems intensively; system errors and incompatibility with other systems would mean extra work and tie up resources (G24:14), leading to a rubber band effect for HCPs (G61:110). Pressure on individual patients could increase, similar to the pressure on hospitals and health systems to reduce costs through behavioral change (G57:22). To cope with this pressure, hospitals could set up "sensor farms" (R46:25) to cut costs and use sensors without medical indication (P73:93). In particular, a possible increase in patient transparency vis-à-vis health insurance companies was seen critically (H41:21). Great fears were also expressed regarding the solidarity principle of public health insurance systems such as the German one, "whether [...] health services could not also be linked to certain lifestyles, [...]

XSL•FO
RenderX

and that is actually an imperative that is categorically ruled out in Germany" (G57:54).

Sociopolitical Opportunities and Challenges

Social opportunities (Multimedia Appendix 13) were only touched upon, mainly by researchers: health in general is increasingly taking center stage, and the societal benefits outweigh this. In addition to being accessible to the masses (R27:23), the use of sensors could contribute to "[…] be more mindful, especially if you [...] do different training sessions or exercises or just know [...] what your body needs [...] and maybe the togetherness is more balanced because you are in control of yourself" (G59:76). It can also reduce the burden on relatives, both by providing a useful measurement tool and by relieving them of the need to be present during disease-altering events (R13:143).

Particular political opportunities (Multimedia Appendix 13) are to overcome barriers (such as language, or to sustain accessibility in the patient's room; P80:67) and to reach people who avoid hospitals because of fear (H83:61). One political stakeholder described the use of sensors for immobile patients as an opportunity to address demographic change despite the shortage of physicians (PS104:80).

Identified social challenges (Multimedia Appendix 14) include various types of fears (of surveillance, loneliness, staff cuts, and others) and risks to individuals. More direct access to one's own data could make patients feel anxious and discouraged and quickly lead to overfocus: "[P]hysical health even more steps out of the private sphere [...] and becomes an object [...] of public display. [...] [T]his display of inner strength, of health and vitality, also creates pressure [...] on people who are unable to do so because they may simply be ill" (G57:2). In terms of interaction, a possible influence on social development (R46:179), on social participation (R76:22), and an increasing lack of physical contact were mentioned. In relation to human-human interaction (H74:44), a patient described the importance of human contact from their own experience: "I managed to get back on track to a certain extent. But only with the human element. With technology alone [...] I would [...] no longer see any sense in life" (P69:85). Several participants highlighted the role and importance of touch in the healing process: "[Y]ou can also simply reassure people, establish contact with people […] by simply touching them" (H75:104). Contactless data collection often leaves a feeling of unreliability or a feeling of not having been properly examined (R15:139 and R43:24). Some participants drew parallels with the contactless situation during the SARS-CoV-2 pandemic (H74:71 and H86:103).

Low levels or lack of acceptance were identified as a political challenge (Multimedia Appendix 15), which may be due to the human impulse to initially reject the unknown (PS97:18), the presumed slow build-up of acceptance among patients and HCPs (PS102:43), and societal reservations about radiation (R8:127 and H92:47). One patient suggested that "[...] a discussion should only take place on a scientific level, [...] if the general [...] population takes part in a discussion, um, the topic is already lost anyway" (P68:40). In terms of preimplementation issues, the benefits of contactless measurement were questioned in general, not just for patients who were already wired (R15:46 and PS102:55). The judicious use of AI-CS was described as a challenging balancing process. The resources needed to operate the technology should not be wasted (G52:99). Political stakeholders in particular saw implementation challenges, for example, due to the slowness of change in the medical field (R30:118). Another challenge is also a not yet realized "[…] paradigm shift, which then favors a preventive […] health protection, health care" (PS102:55). Further ambiguities exist in the exact procedure (PS101:78).

Regulatory Opportunities and Challenges

Only regulatory challenges (Multimedia Appendix 16) were identified by the participants. Several challenges related to consent—mainly mentioned by HCPS, researchers, and the general public—focused on the unobtrusiveness of AI-CS (R12:3). Consent was emphasized as all the more important, especially for radar-based technologies (G64:36 and R31:13), because of the greater safety risk of measuring people without their awareness (compared with contact measurement technologies G61:112). All stakeholder groups identified a high risk of abuse of the sensors (eg, stalking or surveillance in public spaces or by employers [PS101:89 and R31:120], or for use in a state sanction mechanism [R31:98]), even leading to dystopian fantasies of total surveillance (G47:5). Challenges were also described for the management of the collected data, that is, how individuals could retain control over and manage their data (H92:62), while at the same time ensuring data security (R32:96), privacy (R36:132), and transparency about their own data (R63:97 and G64:162). Regarding regulation, a "regulatory hole" (PS40:93) was addressed. Medical device approval was described as challenging (PS40:124 and PS102:124). Another problem identified was the reactive nature of the law, which does not provide a framework for future settings (PS102:60). Contrary to the political stakeholders, an HCP mentioned that they believe that legal and ethical rules would not prevent misuse (H66:64 and H85:25).

In summary, the analysis revealed a wide range of perceived opportunities in medical, technical, and economic terms. Opportunities, such as the impact on society, were mainly focused on by researchers and political stakeholders. Naturally, HCPS and patients focused strongly on the application and effects of AI-CS in health. The broad range of challenges mentioned across all stakeholder groups showed the perceived ambivalence of AI-CS (eg, in the technical implementation and data generation), leading to the need for good communication, preventive governance, and the overall importance of human interaction in the health system.

Values

Trust, legal certainty, sociality, autonomy, individuality, and usefulness could be identified as core values of the different stakeholder groups towards AI-CS in health.

All stakeholder groups emphasized the need to build trust in the use of AI-CS in health, particularly with regard to the need for transparency, reliability, accuracy, and proven usability: "[…] if […] it is transparent beforehand what will be done with it, of course I would be in favor of that" (R30:26). Patients

XSL•FO
RenderX

referred to the trust regarding the implementation of contactless sensors and AI, which is seen as "[...] a result of the relationship of trust that has developed between the patient and the doctor" (P22:96). Trust in these technologies is not yet there due to their recency, so it needs to be built: "There is still no basis of trust because you don't know it yet [...]" (P73:119). Some patients also expressed a leap of faith in new technologies, particularly relating to their expectations of better health care or due to a perceived objectivity of the technologies, that is, seeing AI as a form of "objective value" (P9:261). On the other hand, HCPS took a comparatively matter-of-fact approach for formulating the conditions for trust, demanding scientific evidence, explainability, and validated usability of new technologies. HCPS also addressed the pitfalls and consequences of changing trust attributions and perceptions of AI for the individual doctor: "[D]o I trust the person who's been doing this somehow for years, or do I trust the AI that has all my data?" (H78:57)

Second, the need for legal certainty for the acceptance of new technologies was identified across all stakeholder groups—in particular for the authorization of new technologies, but also for the secure storage of data that are not accessible to third parties ("neighbors" G60:59).

Third, different stakeholder groups emphasized the importance of human interaction, humanity, and human connection, as well as care in health, summarized as sociality. Patients feared that the central relational aspect of being human would be lost using contactless sensors (H83:40). Contactless sensors would lack empathy—that's why some participants would prefer to have a human in the loop who "[...] is somehow able to absorb even more than a sensor" (P95:44).

Fourth, autonomy was identified as a crucial value, with stakeholders underlining the importance of self-control and decision-making in the context of health (G64:40). This value was further embraced by the notion of freedom to make an informed decision, including the choice not to know certain pieces of health-related information. The unpredictability of the future, together with an openness to various outcomes and the fragility of life, are some of the values that were articulated across all stakeholder groups (eg, G57:44).

Fifth, the recognition of individuality came up as a topic of conversation, particularly in interviews with patients, HCPS, and the general public. Concerns were raised about the potential for contactless sensors to overlook individual differences: "Oops, I'm no longer an individual here, I'm […] this biological entity, a human being. And one is like the other. And I don't like that" (P22:63).

The sixth value mentioned across all stakeholder groups was usefulness, particularly in terms of medical usefulness for a patient's individual health or for the medical-technical progress in general (H92:102). One researcher mentioned:

> *[…] I would certainly also benefit from the research results […] through other people. In the same way, I would participate in order to advance research. To steer the progress of knowledge.* [(R38:29)]

Behavioral Aspects

Experience

Specific experiences with contactless measurement technologies in a medical context were barely described across all groups (eg, P68:27). Detailed reports of experiences with AI-CS came only from researchers (R32:138 and R38:246). At the same time, however, there were numerous analogies to the operation of the technology, which can be divided into 2 main observations: first, a link was made between the way sensors work and human capabilities (eg, P73:15). The human capacity to evaluate internal states without physical touch is thought to be reinforced and/or substantiated by AI-CS. Secondly, analogies were drawn to existing analog technologies that measure from the outside without contact, for example, X-rays (P73:122).

Another important role played by the question of whether the use of AI-CS leads to more (perceived) supervision, which, however, was contested by the experience of HCPs (H92:64). In addition, some researchers pointed out that the patient's previous experiences were significant:

> *If someone has ever been in intensive care [...] and was completely wired up for a fortnight and had to press the bell every time, […] they will immediately say, [...] I want to be monitored [...] contactlessly [...].*" [(R42:63f)]

Implications

Three implications were raised. The first one concerns changes in the doctor-patient relationship caused by new contactless measurement technologies. Some anticipate changes in the level of monitoring, which could lead to uncertainty and sometimes even anger (P87:88), while HCPS described the developments as a logical progression toward further personalization of medicine (H39:132). Researchers expressed similar sentiments: "Why should it suddenly be any different with contactless sensors if it happens in a medical setting […]" (R76:70)? Nevertheless, conditions for a successful introduction of the sensors, such as transparency and education (eg, PS103:31), individual medical background and attitude (eg, PS104:30), as well as serial application (PS102:89) were emphasized particularly by political stakeholders. One participant of this group assumed that patients' sense of surveillance will be much less in a medical context than when the sensors are used in private homes or for other purposes (PS100:22). This is consistent with the also mentioned assertion that the medical benefits are likely to outweigh any legal concerns (PS102:66).

The second implication relates to shifts in the values recorded. In particular, the question of possible changes in the conception of trust plays a central role here. Some respondents seemed to trust in the prevention of misuse and general regulation by legal measures (G37:90). This especially means that decision-making authority remains with the HCPS:

> *And if the data collection from the contactless sensors has shown that, then the doctor will say: '[…] now we'll do it like this.' Then I would have so much*

XSL•FO
RenderX

> *confidence in the doctor that they can interpret what the AI says […].* [(G47:81)]

A third implication concerns anthropological issues. In addition to the ascribed analogy between the functioning of the sensor and human abilities, the discussion also centered on how the use of contactless sensors is changing the understanding of what it means to be human. With the elimination of touch as a basic element of measurement, the loss of its significance as a deeply human need is presumed (eg, H86:103). AI-CS is also expected to have an impact on self-interpretation techniques and experiences of efficacy: "Yes, I'm quite convinced of that. In the same way that people have forgotten how to find their way around since there are navigation systems […] and so on" (P95:57).

Preferred Actions

The preferred action relates to 3 main areas: first, the impact of AI-CS on forms and ways of establishing practices of consent was discussed. Some participants seemed to be largely indifferent: "But if the doctor explains that we have now come so far as to be able to make the whole thing contactless from other, uh, indicators, uh, that can be obtained from other bodily functions, then I can say: Well, fine, please" (P22:163), while others were rather positive (P70:114). Although AI-CS may change the established routines of giving consent, the importance of this consent is strongly emphasized (H41:30 and R65:16). As all groups thought, it is important for patients to know when a measurement is being taken and when it is not.

A second area concerns the assessment of (institutional) responsibility. Here, the prevention of abusive data collection was a rather important issue. Political stakeholders focused strongly on responsibility and its precise organization (eg, PS40:153), distinguishing between different levels of moral and legal responsibility, affecting both manufacturers and HCPs (eg, PS97:89). One political stakeholder argued that responsibility lies with the manufacturer and the user (PS40:153), but that the explicit approval of the patient must be taken into account (PS40:24). It was considered crucial, especially by HCPS, that AI-CS be embedded and evaluated in institutional decision-making pathways (H44:36 and H75:6). The use of these sensor technologies should intend to support staff, not replace them (H28:22 and G67:95). In addition, further research and institutionalized monitoring should ensure the functionality and safety of the sensors (G67:131). Furthermore, within the area of responsibility, all groups saw education as an important institutional aspect (P35:17, H26:10, R19:47, PS102:40, and G5:31): There is a need for training in the skills required—not only to explain the new measurement techniques, but also to understand their implications and limitations.

A third area of preferred action concerns the question of whether new measurement technologies will lead to the medicalization of everyday life. In other words, health issues also play a role in areas where they are not (or should not be) the focus of attention. Very different assessments of such a possible extension emerged: "Yes, so perhaps not […] in working life, in the office" (P45:122). In contrast: "Well, I don't know if I'm raving too far or thinking too high, but for me it would be ideal to have such a sensor in my pocket, day and night, really", (P58:31).

Discussion

Brief Summary

The study addresses 2 key shortcomings: the lack of empirical research on attitudes toward AI-CS and the lack of a comprehensive stakeholder approach.

Regarding the first shortcoming, the results go beyond other studies by providing a comprehensive picture of attitudinal aspects toward AI-CS by following the analytical 3-component model. Overall, the results show a high level of openness to AI-CS in health across all stakeholder groups. Besides, 2 trends emerge from the results: first, greater experience and knowledge correlated with a reduced tendency to react emotionally. In particular, researchers were less likely to express strong emotions, both negative and positive, while HCPS showed little to no strong negativity. Both groups tended to take a sober view, with a focus on realistic expectations. In contrast, those with little or no experience or knowledge of the technology expressed stronger attitudes (showing both very enthusiastic, very uncertain, and/or negative feelings) toward certain aspects of the technologies.

Second, participants with good experiences with technologies were generally more open and positive toward contactless sensor technologies. This finding is, on the one hand, supported by previous research linking attitudes to prior experience [53]. On the other hand, it highlights the observation that attitudes are very likely to be based on the perceived social changes associated with a new technology, which is in line with similar empirical studies on people's attitudes and values when interpreting emerging technologies [54-56]. This observation also backs the theoretical assessment that AI should not be viewed as an abstract technical variable, but rather in terms of sociotechnical systems [57-60]. In this sense, AI-CS does not represent a completely new case, but rather raises issues that are inherent to human-technology interactions in a similar way to other AI-driven technologies used in health, such as issues of trust [61].

Regarding the lack of a comprehensive stakeholder approach, the study offers an innovation pipeline approach and responds to the current debates on AI that call for the inclusion of diversity and multiplicity of stakeholders to avoid harm [62]. Furthermore, it explicitly "[…] account[s] for the potential limitations of one group speaking for another group, which may in fact disagree" [63]. Different particularities emerged between the groups: patients showed an altruistic moment of great willingness with regard to sharing data if they see a medical benefit, while political stakeholders put an emphasis on practical implementation and suggested concrete proposals by taking into account current regulatory developments. At a thematic level, all stakeholder groups identified technical opportunities, but patients, HCPS, and researchers provided more detailed insights, reflecting their direct involvement in the development, evaluation, and application of sensor technologies. The multistakeholder approach ensures that different perspectives

XSL•FO
RenderX

are well-represented. The sample size of 104 participants allows commonalities and divergences across all stakeholder groups to be highlighted. Despite some overlap between groups (eg, between patients and the general public, or HCPS and researchers), self-classification facilitated nuanced insights from each stakeholder group.

Implications for Governance and Policy

Two key lessons emerge from the findings: first, providing clear and targeted information and facilitating (nonmedical) hands-on experience with AI-CS can play a crucial role in mitigating rejection. Understanding of the term "contactless" emerged as a key factor in shaping attitudes, suggesting that targeted information could address misconceptions effectively. At the same time, the diversity of associations highlighted by all stakeholder groups points to the overarching importance of attitudes over detailed technical knowledge. While technical understanding may reduce initial uncertainties, acceptance of AI-CS is ultimately influenced by broader attitudes and the ideas that participants associate with them.

In our study, 3 tensions were shown to be important: first, the tension around contact (lessness) and the importance and ambivalence of touch; second, the tension between protection and surveillance in relation to the path- and context-dependency; and third, the tension between the benefits and challenges of unobtrusiveness, particularly in relation to control and governance implications.

Starting with the first tension, the concept of "contactless" revolves around the question of whether or not one is being touched. All stakeholder groups identified this question as crucial, showing an ambivalence about the importance of touch and the potential benefits of avoiding it: on the one hand, touch and especially being touched were described as a deeply human experience that was considered an integral part of medical consultations, especially by HCPs. Touch is considered essential in the relationship between people, but also in the relationship with oneself, one's body, and its capabilities. A contactless sensor may not be able to capture all parts of the (bodily) self, but will be limited to quantitative parameters, which in turn will limit the understanding of a healthy and a sick body. This could eventually lead to a problematic image of the body as always potentially ill. It resonates with wider debates about the quantification and datafication of health, where complex, embodied experiences risk being reduced to measurable parameters [64]. Such reductionist tendencies are particularly relevant in sensory and embodied health care [65], where touch plays both diagnostic and relational roles.

On the other hand, the absence of touch was seen by all stakeholder groups as an important way of avoiding unwanted intimate physical contact, potentially removing one of the barriers to seeking medical advice. Beyond this practical dimension, participants agreed that the recording of symptoms and parameters is not mere measurement, but also involves an anthropological dimension. To be measured means to be under the gaze of an "other." This interaction shapes the understanding of what it means to have an experience of one's own bodily experience classified and interpreted from the perspective of another. The use of technologies such as contact or contactless systems introduces additional layers of complexity to this notion. For example, when a cable is used to measure parameters, a physical boundary is established between a person's body and its environment. In contrast, contactless technologies challenge traditional notions of where these boundaries lie. This highlights the sociotechnical dimension of such innovations in that it is necessary to renegotiate where the boundaries of the human body are seen [66,67].

Second, all stakeholder groups pointed to a particular path- and context-dependency and its significance for attitudes by drawing a clear distinction between health data collection within a defined institutional context and outside of it, marking a strong tension between protection and surveillance. The use of AI-CS within institutional health contexts (such as hospitals or doctors' surgeries) was identified as a helpful tool for protection by facilitating the need for constant monitoring. The ability to monitor without restricting movement was seen as one of the key benefits of contactless automated systems, particularly for people with dementia or in fields such as palliative care. Outside of health and institutional settings, all stakeholder groups raised concerns about illegitimate monitoring and surveillance, potentially leading to a loss of control or even knowledge of one being monitored. Concerns about a strong intrusion into privacy were repeatedly highlighted, as well as the potential for harmful effects on both the individual and society [68]. This tension between protection and surveillance is a major challenge for the use of AI-CS, as the boundaries between these contexts are becoming increasingly blurred by the ability and flexibility to use them in everyday life and in different locations due to the portability of the devices.

Related to this, a third tension emerged around the benefits and challenges of the unobtrusiveness of the measurement using AI-CS. While some researchers highlighted the unobtrusiveness of the data collection as a major advantage in minimizing measurement distortions (eg, movement changes due to measuring devices attached to the body), other researchers and HCPS expressed concern that the use of AI-CS could lead to a loss of visibility and thus an important basis for control [41]. These different perspectives highlight the dual nature of invisibility as both a technical strength and a potential source of institutional challenges. To mitigate the risks associated with unobtrusive measurement, it is essential to ensure the visibility of (interaction) interfaces. This aspect becomes even more critical when contactless sensor technology converges with automation through AI: automation can contribute to more efficient analyses of the collected data and can provide a more robust basis for decision-making in complex situations. However, the integration of AI-CS requires good interface design and careful consideration of innovation processes within institutional decision-making pathways in the clinic [69]. Moreover, defining benchmarks and criteria for defining the benefits and harms of technical systems requires particular scrutiny [70]. These criteria should be aligned with broader societal goals, aiming to provide a common good and contribute to the reduction of structural inequalities [71,72]. For example, tensions between individual benefits and benefits for the largest possible group need to be taken into account.

XSL•FO
RenderX

In this context, regulatory frameworks play a pivotal role in navigating the ethical complexities of AI-CS in health. Well-designed regulations are essential to balance the tensions between protection and surveillance and to address the challenges posed by unobtrusive measurement. Regulations such as the Medical Device Regulation [73] provide a basis for mitigating these tensions and preventing harm. The findings highlight the importance of proactive governance to ensure the ethically aligned implementation of AI-CS in health.

Strengths and Limitations of the Study

This study provides comprehensive and nuanced insights into attitudes toward contactless measurement technologies in the health sector. By integrating perspectives from a broad range of stakeholder groups, it directly addresses the previously identified gap in research that often considers only a limited set of viewpoints. The use of both onsite and online interviews proved particularly valuable, as it enabled the inclusion of participants across a wide age spectrum and with varying levels of technological affinity, from highly tech-savvy individuals to those with limited experience with digital tools.

Recruitment efforts extended both within and beyond Germany. Although the majority of participants were based in Germany (n=97), a smaller share (n=7) came from other countries. While the study primarily focused on the German context, the findings nevertheless carry relevance for international settings and contribute to broader discussions surrounding the global development and acceptance of AI-based health technologies, provided potential cultural context biases are taken into consideration.

Despite repeated attempts to ensure balanced representation across all stakeholder groups, the study was unable to recruit an equal number of political stakeholders compared to other groups. This imbalance underscores an ongoing challenge in recruiting political stakeholders for qualitative research of this kind, which may be due to their smaller group size in general (compared to patients, for example) and their accessibility. Notwithstanding this challenge, it is important to acknowledge that the study included political stakeholders from both regional and national levels.

The participants' level of education was higher than the average in Germany. This could be interpreted as either a selection bias or an increased opportunity or willingness of higher-educated persons to participate in studies.

Furthermore, a lack of clarity surrounding AI-CS must be mentioned. Participants frequently asked which applications could be considered contactless. They gave different interpretations of the term and pointed to a wide range of associations with the technology, based on their prior experiences (eg, prior illnesses and social events such as the SARS-CoV-2 pandemic), personal health and nonhealth experiences, knowledge, technical understanding, and imagination. Overall, the results show a lack of information and technical knowledge that led to uncertainty and discomfort for some participants. If such uncertainty is exacerbated, it could contribute to rejection, influenced by pseudoscientific reasoning.

Conclusion

Our contribution to research on AI-CS is 3-fold: first, it provides the first comprehensive overview of attitudes toward AI-CS in health, addressing the moral values and sociotechnical practices. By analyzing the attitudes of 5 different stakeholder groups along feelings, thoughts, and behavioral aspects, the study yields nuanced insights that can be further explored in the SENSAI online tool [51]. Additionally, it presents a detailed examination of potential applications (eg, home monitoring for dementia) and potential users (from children to older adults; Multimedia Appendix 5).

Second, we have demonstrated the importance of exploring attitudes toward AI-CS, which have not yet been empirically investigated. The study provides a broad picture of different components of attitudes, as participants expressed enthusiasm, rejection, or indifference (feelings), acknowledged certain opportunities and challenges, and considered which values were touched upon (thoughts), based on their experiences, formulated implications, and preferred actions (behavioral aspects). The use of semistructured interviews allowed the participants to bring in their own experiences and draw parallels with other devices they are familiar with, as AI-CS are still relatively new technologies. Furthermore, by providing a clear, deductively derived concept of attitudes that paves the way for avoiding the fuzzy equation with the terms perspectives, values, views, and others, the study makes an important methodological contribution to AI ethics and social science research on AI-based technologies.

Third, the key tensions (regarding the role of touch and contactlessness in health, between protection and surveillance in relation to path- and context-dependency, and regarding benefits and challenges of unobtrusiveness, particularly in relation to control and governance implications) call for robust regulatory frameworks to prevent the misuse of AI-CS and promote acceptance, including targeted information and facilitation of hands-on experience.

Through these 3 contributions, the study aimed to address the 2 shortcomings of the lack of a multistakeholder approach and the lack of empirical analysis of a hitherto unexplored technology by applying the concept of attitudes and its analytical precision. The study provides important entry points into different areas of research and practice and acts as a springboard for further research on AI-based health technologies in general and AI-CS in particular, as it highlights crucial points regarding sensor development at different stages, data analysis, and application opportunities and challenges in different medical specialties and functions. The findings on acceptance and preferred actions will be particularly useful for political consideration and possible implementation processes that could advance the governance and regulation of sensor technologies. By implementing these findings, AI-CS could contribute to preventing patient harm and improving patient safety [74].

XSL•FO
RenderX

Acknowledgments

The authors are grateful to Nils Albrecht, Giuseppe Comerci, Nora Faust, Valerie Möhrle, Carina Müller, Clara Odendahl, Emma Pett, Pauline Pett, Mary Schneider, Farina Schumann, Anna Strauß, Marie-Louise Thon, and Philipp Suffa for their comments on the draft and their support in proofreading the article, EndNote literature management, Unipark survey management, and graphic design. The use of generative artificial intelligence (AI) in the research and writing process is not declared by the authors.

Funding

This study has been co-funded by the European Union (ERC, SIMTWIN, 101076822). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant SFB 1483–Project-ID 442419336. The DFG had no influence on the study's design, analysis, and evaluation. This publication was supported by the Open Access Publication Fund of the University of Bonn.

Authors' Contributions

Conceptualization/Project administration: MB, EH
Data collection (including recruitment): EH, HB, BC, BE, MH, AL, AM, NR, DS, MB
Formal analysis: EH, MB, HB, SE, LG, AM, DS, CS
Writing – original draft: EH, MB, HB, SE, LG, AM, DS, CS
Writing – review & editing: EH, MB, HB, BC, SE, BE, LG, MH, AL, AM, NR, DS, CS

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[DOCX File, 41 KB - jmir_v28i1e75783_app1.docx]

Multimedia Appendix 2
Sociodemographic questionnaire.
[DOCX File, 32 KB - jmir_v28i1e75783_app2.docx]

Multimedia Appendix 3
Overview of positive feelings expressed by different stakeholder groups.
[PDF File, 22 KB - jmir_v28i1e75783_app3.pdf]

Multimedia Appendix 4
Overview of negative feelings expressed by different stakeholder groups.
[PDF File, 21 KB - jmir_v28i1e75783_app4.pdf]

Multimedia Appendix 5
Overview of mixed feelings expressed by different stakeholder groups.
[PDF File, 14 KB - jmir_v28i1e75783_app5.pdf]

Multimedia Appendix 6
Medical opportunities and challenges according to the different medical specialties.
[DOCX File, 34 KB - jmir_v28i1e75783_app6.docx]

Multimedia Appendix 7
Medical opportunities identified by different stakeholder groups.
[DOCX File, 20 KB - jmir_v28i1e75783_app7.docx]

Multimedia Appendix 8
Medical challenges identified by different stakeholder groups.
[DOCX File, 17 KB - jmir_v28i1e75783_app8.docx]

XSL•FO
RenderX

Multimedia Appendix 9
Technical opportunities identified by different stakeholder groups.
[DOCX File, 22 KB - jmir_v28i1e75783_app9.docx]

Multimedia Appendix 10
Technical challenges identified by different stakeholder groups.
[DOCX File, 24 KB - jmir_v28i1e75783_app10.docx]

Multimedia Appendix 11
Economic opportunities identified by different stakeholder groups.
[DOCX File, 16 KB - jmir_v28i1e75783_app11.docx]

Multimedia Appendix 12
Economic challenges identified by different stakeholder groups.
[DOCX File, 17 KB - jmir_v28i1e75783_app12.docx]

Multimedia Appendix 13
Sociopolitical opportunities identified by different stakeholder groups.
[DOCX File, 15 KB - jmir_v28i1e75783_app13.docx]

Multimedia Appendix 14
Social challenges identified by different stakeholder groups.
[DOCX File, 18 KB - jmir_v28i1e75783_app14.docx]

Multimedia Appendix 15
Political challenges identified by different stakeholder groups.
[DOCX File, 16 KB - jmir_v28i1e75783_app15.docx]

Multimedia Appendix 16
Regulatory challenges identified by different stakeholder groups.
[DOCX File, 19 KB - jmir_v28i1e75783_app16.docx]

Checklist 1
COREQ checklist.
[PDF File, 561 KB - jmir_v28i1e75783_app17.pdf]

References

1. Verbeek PP. Materializing morality: design ethics and technological mediation. Sci Technol Human Values 2006;31(3):361-380. [doi: 10.1177/0162243905285847]
2. Kerschner C, Ehlers MH. A framework of attitudes towards technology in theory and practice. Ecol Econ 2016 Jun;126:139-151. [doi: 10.1016/j.ecolecon.2016.02.010]
3. Pickens J. Attitudes and perceptions. In: Borkowski N, editor. Organizational Behavior in Health Care: Jones and Bartlett Publishers; 2005:43-76 URL: https://www.researchgate.net/profile/Jeffrey-Pickens/publication/267362543_Attitudes_and_Perceptions/links/5728b08008aef7c7e2c0bf26/Attitudes-and-Perceptions.pdf [accessed 2026-04-02]
4. Eagly AH, Chaiken S. The advantages of an inclusive definition of attitude. Soc Cogn 2007 Oct;25(5):582-602. [doi: 10.1521/soco.2007.25.5.582]
5. Eagly AH, Chaiken S. The Psychology of Attitudes: Harcourt Brace Jovanovich College Publishers; 1993.
6. Amann J, Vayena E, Ormond KE, Frey D, Madai VI, Blasimme A. Expectations and attitudes towards medical artificial intelligence: a qualitative study in the field of stroke. PLoS One 2023;18(1):e0279088. [doi: 10.1371/journal.pone.0279088] [Medline: 36630325]
7. Akudjedu TN, Torre S, Khine R, Katsifarakis D, Newman D, Malamateniou C. Knowledge, perceptions, and expectations of artificial intelligence in radiography practice: a global radiography workforce survey. J Med Imaging Radiat Sci 2023 Mar;54(1):104-116. [doi: 10.1016/j.jmir.2022.11.016] [Medline: 36535859]
8. Blease C, Kaptchuk TJ, Bernstein MH, Mandl KD, Halamka JD, DesRoches CM. Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views. J Med Internet Res 2019 Mar 20;21(3):e12802. [doi: 10.2196/12802] [Medline: 30892270]

XSL•FO
RenderX

9. Morris RR, Kouddous K, Kshirsagar R, Schueller SM. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. J Med Internet Res 2018 Jun 26;20(6):e10148. [doi: 10.2196/10148] [Medline: 29945856]
10. Mehta N, Harish V, Bilimoria K, et al. Knowledge and attitudes on artificial intelligence in healthcare: a provincial survey study of medical students. MedEdPublish 2021;10(1). [doi: 10.15694/mep.2021.000075.1]
11. Buck C, Doctor E, Hennrich J, Jöhnk J, Eymann T. General practitioners' attitudes toward artificial intelligence-enabled systems: interview study. J Med Internet Res 2022 Jan 27;24(1):e28916. [doi: 10.2196/28916] [Medline: 35084342]
12. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study. Digit Health 2019;5:2055207619871808. [doi: 10.1177/2055207619871808] [Medline: 31467682]
13. Jungmann F, Jorg T, Hahn F, et al. Attitudes toward artificial intelligence among radiologists, IT specialists, and industry. Acad Radiol 2021 Jun;28(6):834-840. [doi: 10.1016/j.acra.2020.04.011] [Medline: 32414637]
14. Sit C, Srinivasan R, Amlani A, et al. Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey. Insights Imaging 2020 Feb 5;11(1):14. [doi: 10.1186/s13244-019-0830-7] [Medline: 32025951]
15. Rainie L, Funk C, Anderson M, Tyson A. AI and human enhancement: Americans' openness is tempered by a range of concerns. Pew Research Center. 2022. URL: https://www.pewresearch.org/internet/2022/03/17/ai-and-human-enhancement-americans-openness-is-tempered-by-a-range-of-concerns/ [accessed 2026-03-20]
16. Modhvadia R. How Do People Feel about AI? A Nationally Representative Survey of Public Attitudes to Artificial Intelligence in Britain: Ada Lovelace Institute, The Alan Turing Institute 2023. URL: https://adalovelaceinstitute.org/report/public-attitudes-ai [accessed 2026-03-23]
17. Currie M, Philip LJ, Roberts A. Attitudes towards the use and acceptance of eHealth technologies: a case study of older adults living with chronic pain and implications for rural healthcare. BMC Health Serv Res 2015 Apr 16;15:162. [doi: 10.1186/s12913-015-0825-0] [Medline: 25888988]
18. Alghamdi SA, Alashban Y. Knowledge, attitudes and practices towards artificial intelligence (AI) among radiologists in Saudi Arabia. J Radiat Res Appl Sci 2023 Jun;16(2):100569. [doi: 10.1016/j.jrras.2023.100569]
19. Wrede C, Braakman-Jansen A, van Gemert-Pijnen L. Requirements for unobtrusive monitoring to support home-based dementia care: qualitative study among formal and informal caregivers. JMIR Aging 2021 Apr 12;4(2):e26875. [doi: 10.2196/26875] [Medline: 33843596]
20. Wirth V, Liphardt AM, Coppers B, et al. ShaRPy: shape reconstruction and hand pose estimation from RGB-d with uncertainty. Presented at: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW); Oct 2-6, 2023. [doi: 10.1109/ICCVW60793.2023.00277]
21. Phutane U, Liphardt AM, Bräunig J, et al. Evaluation of optical and radar based motion capturing technologies for characterizing hand movement in rheumatoid arthritis-a pilot study. Sensors (Basel) 2021 Feb 9;21(4):1208. [doi: 10.3390/s21041208] [Medline: 33572273]
22. Krauss D, Engel L, Ott T, et al. A review and tutorial on machine learning-enabled radar-based biomedical monitoring. IEEE Open J Eng Med Biol 2024;5:680-699. [doi: 10.1109/OJEMB.2024.3397208] [Medline: 39193041]
23. Liphardt AM, Manger E, Liehr S, et al. Similar impact of psoriatic arthritis and rheumatoid arthritis on objective and subjective parameters of hand function. ACR Open Rheumatol 2020 Dec;2(12):734-740. [doi: 10.1002/acr2.11196] [Medline: 33241646]
24. Giusti A, Pukrittayakamee P, Wannarit K, et al. How to deliver person-centred care for people living with heart failure: a multi stakeholder interview study with patients, caregivers and healthcare professionals in Thailand. BMC Health Serv Res 2024 Dec 18;24(1):1570. [doi: 10.1186/s12913-024-11922-z] [Medline: 39696212]
25. Oh S, Kim JH, Choi SW, Lee HJ, Hong J, Kwon SH. Physician confidence in artificial intelligence: an online mobile survey. J Med Internet Res 2019 Mar 25;21(3):e12422. [doi: 10.2196/12422] [Medline: 30907742]
26. Shen C, Li C, Xu F, et al. Web-based study on Chinese dermatologists' attitudes towards artificial intelligence. Ann Transl Med 2020 Jun;8(11):698. [doi: 10.21037/atm.2019.12.102] [Medline: 32617318]
27. Claes V, Devriendt E, Tournoy J, Milisen K. Attitudes and perceptions of adults of 60 years and older towards in-home monitoring of the activities of daily living with contactless sensors: an explorative study. Int J Nurs Stud 2015 Jan;52(1):134-148. [doi: 10.1016/j.ijnurstu.2014.05.010] [Medline: 24951084]
28. Demiris G, Hensel BK, Skubic M, Rantz M. Senior residents' perceived need of and preferences for "smart home" sensor technologies. Int J Technol Assess Health Care 2008 Oct;24(1):120-124. [doi: 10.1017/S0266462307080154]
29. Demiris G, Rantz M, Aud M, et al. Older adults' attitudes towards and perceptions of "smart home" technologies: a pilot study. Med Inform Internet Med 2004 Jun;29(2):87-94. [doi: 10.1080/14639230410001684387] [Medline: 15370989]
30. Palmisciano P, Jamjoom AAB, Taylor D, Stoyanov D, Marcus HJ. Attitudes of patients and their relatives toward artificial intelligence in neurosurgery. World Neurosurg 2020 Jun;138:e627-e633. [doi: 10.1016/j.wneu.2020.03.029] [Medline: 32179185]
31. Steele R, Lo A, Secombe C, Wong YK. Elderly persons' perception and acceptance of using wireless sensor networks to assist healthcare. Int J Med Inform 2009 Dec;78(12):788-801. [doi: 10.1016/j.ijmedinf.2009.08.001] [Medline: 19717335]

XSL•FO
RenderX

32. Ho MT, Le NT, Mantello P, Ho MT, Ghotbi N. Understanding the acceptance of emotional artificial intelligence in Japanese healthcare system: a cross-sectional survey of clinic visitors' attitude. Technol Soc 2023 Feb;72:102166. [doi: 10.1016/j.techsoc.2022.102166]
33. Scott IA, Carter SM, Coiera E. Exploring stakeholder attitudes towards AI in clinical practice. BMJ Health Care Inform 2021 Dec;28(1):e100450. [doi: 10.1136/bmjhci-2021-100450] [Medline: 34887331]
34. Huxtable R, Ives J. Mapping, framing, shaping: a framework for empirical bioethics research projects. BMC Med Ethics 2019 Nov 27;20(1):86. [doi: 10.1186/s12910-019-0428-0] [Medline: 31775725]
35. Mayring P. Qualitative Content Analysis. A Step-by-Step Guide: SAGE; 2026. [doi: 10.4135/9781036231798]
36. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
37. Prainsack B, Pot M. Qualitative Und Interpretative Methoden in Der Politikwissenschaft: facultas; 2021. URL: https://www.utb.de/doi/pdf/10.36198/9783838555843-1-11 [accessed 2026-04-02]
38. Jongsma KR, Bredenoord AL. Ethics parallel research: an approach for (early) ethical guidance of biomedical innovation. BMC Med Ethics 2020 Sep 1;21(1):81. [doi: 10.1186/s12910-020-00524-z] [Medline: 32867753]
39. Hagaman AK, Wutich A. How many interviews are enough to identify metathemes in multisited and cross-cultural research? Another perspective on Guest, Bunce, and Johnson's (2006) Landmark Study. Field methods 2017 Feb;29(1):23-41. [doi: 10.1177/1525822X16640447]
40. Creswell JW, Creswell JD. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches: SAGE; 2022. URL: https://uk.sagepub.com/en-gb/eur/research-design/book270550 [accessed 2026-04-02]
41. Hille EM, Hummel P, Braun M. Meaningful human control over AI for health? A review. J Med Ethics 2023 Sep 20;0:1-9. [doi: 10.1136/jme-2023-109095] [Medline: 37730418]
42. Wachter S, Mittelstadt B. A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. LawArXiv. Preprint posted online on 2019. [doi: 10.31228/osf.io/mu2kf]
43. Cohen IG. Informed consent and medical artificial intelligence: what to tell the patient? SSRN J 2020;108(6):1425-1470. [doi: 10.2139/ssrn.3529576]
44. Deutscher Ethikrat. Mensch und Maschine – Herausforderungen durch Ku¨nstliche Intelligenz.: Deutscher Ethikrat; 2023 URL: https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf [accessed 2026-03-23]
45. Braun M, Hummel P, Beck S, Dabrock P. Primer on an ethics of AI-based decision support systems in the clinic. J Med Ethics 2020 Apr 3;47(12):e3. [doi: 10.1136/medethics-2019-105860] [Medline: 32245804]
46. Herrmann-Lingen C, Buss U, Snaith RP. Hospital Anxiety and Depression Scale - Deutsche Version Deutschsprachige Adaptation Der Hospital Anxiety and Depression Scale (HADS) von R P Snaith Und A S Zigmond: hogrefe; 2018. URL: https://www.testzentrale.de/shop/hospital-anxiety-and-depression-scale-deutsche-version-69320.html [accessed 2026-03-23]
47. Tivian XI GmbH. Unipark. URL: https://www.unipark.com/ [accessed 2026-03-23]
48. Dresing T. Praxisbuch Interview, Transkription & Analyse Anleitungen Und Regelsysteme Fu¨r Qualitativ Forschende, 9th edition: dr. dresing & pehl GmbH; 2024. URL: www.audiotranskription.de/praxisbuch [accessed 2026-04-02]
49. ATLAS.ti mac (version 23.2.1). ATLAS.ti Scientific Software Development GmbH. 2023. URL: https://atlasti.com/research-hub/citing-atlas-ti-in-your-research [accessed 2026-04-02]
50. McKenna L. Translation of research interviews: do we have a problem with qualitative rigor? Nurse Author Ed 2022 Mar;32(1):1-3 [FREE Full text] [doi: 10.1111/nae2.31]
51. Hille EM, Garbe L, Spielmann C, Braun M. SENSAI. 2025. URL: https://publish.obsidian.md/sensai [accessed 2026-03-20]
52. Obsidian (version 1.7.7). 2025. URL: https://obsidian.md/ [accessed 2026-04-02]
53. Fazio RH, Zanna MP. Direct experience and attitude-behavior consistency. Adv Exp Soc Psychol 1981;14:161-202. [doi: 10.1016/S0065-2601(08)60372-X]
54. Zhang B, Dafoe A. Artificial Intelligence: American Attitudes and Trends: Center for the Governance of AI, Future of Humanity Institute, University of Oxford; 2019. URL: https://cdn.governance.ai/Artificial_Intelligence_American_Attitudes_and_Trends.pdf [accessed 2026-03-23]
55. Fritsch SJ, Blankenheim A, Wahl A, et al. Attitudes and perception of artificial intelligence in healthcare: a cross-sectional survey among patients. Digit Health 2022;8:20552076221116772. [doi: 10.1177/20552076221116772] [Medline: 35983102]
56. Attitudes towards the impact of digitisation and automation on daily life. : European Commission; 2017 URL: https://europa.eu/eurobarometer/surveys/detail/2160 [accessed 2026-03-20]
57. Hagerty A, Rubinov I. Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence. . Preprint posted online on Jul 18, 2019. [doi: 10.48550/arXiv.1907.07892]
58. Sætra HS. The AI ESG protocol: evaluating and disclosing the environment, social, and governance implications of artificial intelligence capabilities, assets, and activities. Sustain Dev 2023 Apr;31(2):1027-1037 [FREE Full text] [doi: 10.1002/sd.2438]
59. Fiske A, Henningsen P, Buyx A. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J Med Internet Res 2019 May 9;21(5):e13216. [doi: 10.2196/13216] [Medline: 31094356]

XSL•FO
RenderX

60. Keddell E. Algorithmic justice in child protection: statistical fairness, social justice and the implications for practice. Soc Sci (Basel) 2019;8(10):281. [doi: 10.3390/socsci8100281]
61. Bleher H, Braun M. Diffused responsibility: attributions of responsibility in the use of AI-driven clinical decision support systems. AI Ethics 2022;2(4):747-761. [doi: 10.1007/s43681-022-00135-x] [Medline: 35098247]
62. Diberardino N, Baleshta C, Stark L. Algorithmic harms and algorithmic wrongs. : Association for Computing Machinery; 2024 Jun 3 Presented at: FAccT '24: The 2024 ACM Conference on Fairness, Accountability, and Transparency; Jun 3-6, 2024; Rio de Janeiro, Brazil URL: https://dl.acm.org/doi/proceedings/10.1145/3630106 [accessed 2026-03-20]
63. Tiersen F, Batey P, Harrison MJC, et al. Smart home sensing and monitoring in households with dementia: user-centered design approach. JMIR Aging 2021 Aug 11;4(3):e27047. [doi: 10.2196/27047] [Medline: 34383672]
64. Lupton D. The Quantified Self: Polity; 2016. URL: https://www.politybooks.com/bookdetail?book_slug=the-quantified-self--9781509500598 [accessed 2026-04-06]
65. Mol A, Law J. Embodied action, enacted bodies: the example of hypoglycaemia. Body Soc 2004 Jun;10(2-3):43-62. [doi: 10.1177/1357034X04042932]
66. Bublitz JC. Might artificial intelligence become part of the person, and what are the key ethical and legal implications? AI & Soc 2024 Jun;39(3):1095-1106. [doi: 10.1007/s00146-022-01584-y]
67. Ligthart S, Ienca M, Meynen G, et al. Minding rights: mapping ethical and legal foundations of 'Neurorights'. Camb Q Healthc Ethics 2023 May 15;32(4):1-21. [doi: 10.1017/S0963180123000245] [Medline: 37183686]
68. Smuha NA. Beyond the individual: governing AI's societal harm. Internet Policy Review 2021 Sep 30;10(3). [doi: 10.14763/2021.3.1574]
69. Sieck CJ, Sheon A, Ancker JS, Castek J, Callahan B, Siefer A. Digital inclusion as a social determinant of health. NPJ Digit Med 2021 Mar 17;4(1):52. [doi: 10.1038/s41746-021-00413-8] [Medline: 33731887]
70. Prainsack B, Forgó N. New AI regulation in the EU seeks to reduce risk without assessing public benefit. Nat Med 2024 May;30(5):1235-1237. [doi: 10.1038/s41591-024-02874-2] [Medline: 38499661]
71. Braun M, Bleher H, Hille EM, Krutzinna J. Tackling structural injustices: on the entanglement of visibility and justice in emerging technologies. Am J Bioeth 2023 Jul;23(7):100-102. [doi: 10.1080/15265161.2023.2207514] [Medline: 37339313]
72. Jacobs N. Capability sensitive design for health and wellbeing technologies. Sci Eng Ethics 2020 Dec;26(6):3363-3391. [doi: 10.1007/s11948-020-00275-5] [Medline: 33206283]
73. Regulation (EU) 2017/745 of the european parliament and of the council of 5 april 2017 on medical devices, amending directive 2001/83/EC, regulation (EC) no 178/2002 and regulation (EC) no 1223/2009 and repealing council directives 90/385/EEC and 93/42/EEC. European Union. 2017. URL: Http://data.europa.eu/eli/reg/2017/745/oj [accessed 2026-02-06]
74. Bates DW, Levine D, Syrowatka A, et al. The potential of artificial intelligence to improve patient safety: a scoping review. NPJ Digit Med 2021 Mar 19;4(1):54. [doi: 10.1038/s41746-021-00423-6] [Medline: 33742085]

Abbreviations

AI: artificial intelligence
AI-CS: artificial intelligence–based contactless sensors
COREQ: Consolidated Criteria for Reporting Qualitative Research
HADS: Hospital Anxiety and Depression Scale
HCP: health care professional

Edited by A Stone; submitted 18.Apr.2025; peer-reviewed by K Mouloudj, N Sharma; accepted 29.Jan.2026; published 24.Apr.2026.

Please cite as:
Hille EM, Bleher H, Coppers B, Engelhardt S, Eskofier B, Garbe L, Heckel M, Liphardt AM, Manlig A, Rohleder N, Samhammer D, Spielmann C, Braun M
Exploring Attitudes Toward AI-Based Contactless Sensors in Health Among Five Stakeholder Groups: Qualitative Study
J Med Internet Res 2026;28:e75783
URL: https://www.jmir.org/2026/1/e75783
doi:10.2196/75783

© Eva Maria Hille, Hannah Bleher, Birte Coppers, Sophie Engelhardt, Bjoern Eskofier, Lorenz Garbe, Maria Heckel, Anna-Maria Liphardt, Alina Manlig, Nicolas Rohleder, David Samhammer, Carla Spielmann, Matthias Braun. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet

XSL•FO
RenderX

Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Ethical Handling of Occupational Health and Safety Data in the Fire Service: Empirical Interview and Focus Group Study of Firefighter and Fire Service Leadership Privacy Preferences

Rachel J Topazian[1], PhD; Aleksandra Wec[2], BA; Joseph Ali[3], JD; Shannon Frattaroli[2], MPH, PhD; Cassandra K Crifasi[2], MPH, PhD

[1]Division of General Internal Medicine, Center for Bioethics and Humanities, University of Colorado, 13080 E. 19th Ave, Mail Stop B137, Aurora, CO, United States

[2]Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, United States

[3]Department of International Health and Berman Institute of Bioethics, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, United States

Corresponding Author:
Rachel J Topazian, PhD
Division of General Internal Medicine, Center for Bioethics and Humanities, University of Colorado, 13080 E. 19th Ave, Mail Stop B137, Aurora, CO, United States

Abstract

Background: There are ongoing efforts to collect larger and higher-quality amounts of occupational health and safety data to better understand and prevent injuries and fatalities among high-risk workers, such as firefighters. Digital health systems including wearable technologies, mobile apps, or internet-based data collection platforms could collect large amounts of sensitive data, but there is little evidence on worker and employer perspectives on data privacy in the fire service.

Objective: Our study examined firefighters' and fire service leadership's preferences regarding occupational health and safety data privacy.

Methods: We conducted interviews and focus groups with career firefighters in Maryland and Virginia; interviews with union representatives and department-level leaders in each state; and interviews with national-level fire service leaders in advocacy, government, and research organizations (March to November 2023). Interviews and focus groups were audio recorded and transcribed. We analyzed transcripts using thematic analysis.

Results: The sample included 31 career firefighters, 2 union leaders, 11 national leaders, and 21 department-level leaders (65 total participants from 35 interviews and 4 focus groups). We identified 4 themes: acceptability of data access, sharing, and reporting practices; data sharing and access preferences; appropriate use of firefighter data; and the need for improved communication. Leaders described firefighters' concerns about job loss and loss of privacy. Firefighters expressed general preferences that their data be deidentified and not shared widely, and they identified mental health data as important but particularly sensitive information. Firefighters also expressed frustration about sharing data with researchers or their departments without knowing the purpose or outcomes. Both firefighters and leaders emphasized the need for enhanced communication and translation of data for firefighters.

Conclusions: Fire service leaders held more concerns about the use and sharing of occupational health and safety data than firefighters, but both groups identified ways to further safeguard firefighter data and improve communication about health and safety data. Future fire service data collection should incorporate privacy protections, such as limiting the collection of identifiable information and restricting data access. Data collection should be accompanied by clear communication about the purpose of the data collection, how firefighter data will be used and accessed, and the interpretation of the results. Future digital health interventions should integrate these data privacy protections to respect firefighter preferences and contribute to acceptability and uptake.

(J Med Internet Res 2026;28:e84465) doi:10.2196/84465

KEYWORDS

firefighter; occupational health and safety; data privacy; mental health data; data access; data sharing

XSL•FO
RenderX

Introduction

Firefighters take on great personal risks to provide essential services for their communities. There are approximately 1 million firefighters in the United States who respond to over 25 million calls each year [1,2]. Approximately two-thirds of these firefighters are volunteer or unpaid, and one-third are career ("paid") firefighters [1]. Firefighters encounter a vast array of incidents, including fires, natural disasters, and medical emergencies requiring emergency medical services [2]. US firefighters sustain high rates of injury and fatality in the line of duty: in 2022, there were over 65,000 injuries and 96 fatalities that occurred at work [3,4]. Beyond on-duty incidents, firefighting has been linked to long-term adverse health outcomes, including elevated rates of certain cancers [5,6].

There are growing calls from fire service leaders, government officials, and researchers to utilize firefighter health and safety data to better understand the hazards of firefighting and design interventions to protect future generations [7-12]. Proposed data sources include national injury or disease registries, biometric or exposure data from wearable devices, and self-reported risk factors or health outcomes from other digital health platforms, such as smartphone apps [8,13,14]. While recent data collection efforts have focused on injury and disease surveillance [8,9], researchers have identified flaws in existing data infrastructure, such as large swaths of missing occupational data in cancer and burn registries, highlighting the need for more sophisticated data reporting and data linkages [9,10,12]. Federal agencies are supporting new data platforms to enhance injury surveillance and research. In 2023, the US Fire Administration began developing the National Emergency Response Information System (NERIS), a new fire service analytics platform, and the National Institute for Occupational Safety and Health created the National Firefighter Registry for Cancer [7,15,16].

While these new data initiatives may bolster health and safety efforts, expanding the collection of or access to firefighter health and safety data raises ethical concerns [8]. Additional firefighter data might help improve health and safety, but collecting these data, particularly within fire departments, raises concerns about privacy, as firefighters might be asked or required to share sensitive information with their employers, government agencies, or researchers. The fire service's hierarchical organization—fire departments are frequently described as paramilitary organizations—adds additional complexity around data privacy, as firefighters of lower ranks may feel they have less power or choice as to whether or not they want to share personal health information with their employers.

There is little research on data privacy in the fire service, and existing studies focus on the acceptance of new technology in fire departments. Research from our study team examined perceptions of wearable data in the fire service, concluding that while leaders were optimistic about the potential for wearable data to ensure health and safety, firefighters had concerns about privacy, autonomy, and job loss [17]. Although firefighters identified major tradeoffs associated with wearable data, firefighters and leaders agreed on the need for data protections, such as deidentifying data, limiting wearable use to specific settings, providing firefighters access to their data, and preventing punitive data use. Another study examined the implementation of a biometric earpiece that firefighters could wear at a live fire, which would provide an incident commander with data including heart rate, body temperature, calories burned, and exertion estimates [18]. The study concluded that firefighters were concerned about the impact of technology on their autonomy and that firefighters were more skeptical than department leaders. Other researchers developed a biometric privacy framework based on firefighter attitudes toward a personnel management system that would use biometric information (eg, a fingerprint) to manage personnel at a fire. This study identified concerns related to workplace accountability, data access, and employer trust [19]. Collectively, these studies suggest that firefighters may have complex views on how their health and safety data are handled in the workplace, but we lack robust evidence on their privacy preferences and whether or not these preferences are understood and honored in fire departments.

Our study examined fire service members' and their partners' views regarding occupational health and safety data privacy. We explored data sharing, access, and communication preferences among career firefighters, fire department leaders, and national fire service leaders. Our findings highlight opportunities to improve current data privacy practices and ensure that new occupational health and safety digital health platforms, from wearables to mobile health to web-based surveys, are ethically implemented.

Methods

Sample and Recruitment

We purposively recruited participants from 3 groups: career (eg, "paid") firefighters (including union leaders), fire department–level leaders, and national-level leaders. We focused on career as opposed to volunteer firefighters because we wanted to understand views on data privacy among full-time employees. We sampled fire departments that had participated in the Firefighter Organizational Culture of Safety survey. The Firefighter Organizational Culture of Safety survey is used in fire departments to assess safety climate [20]. We recruited 4 fire departments: 2 in Maryland and 2 in Virginia. We targeted departments of varying sizes that were in different regions of each state to facilitate collecting a range of perspectives. We recruited firefighters and department leaders from within the same fire departments to allow for comparisons between management and the rank-and-file within an organization. For participants within fire departments (department leaders, union leaders, firefighters), we asked a point of contact to suggest participants in specific roles whom we then invited to participate. Department leaders and union leaders were invited to participate in interviews, while career rank-and-file firefighters were invited to participate in focus group discussions. Participants were recruited via email. We encountered data collection challenges in Maryland that limited the size of the firefighter sample (see more in the *Results* and *Discussion* sections).

XSL•FO
RenderX

We conducted interviews with leaders within fire departments, including those in overall leadership, health and safety roles, or human resources. Leadership participants included civilians and sworn members of the fire service. For the national-level leaders, we conducted purposive sampling of individuals from government, advocacy, and research organizations. We also conducted snowball sampling for both department and national leaders. We believed that firefighters might hold a range of privacy views but have less experience discussing their privacy preferences. Although some individuals might feel less comfortable sharing their views in a focus group, we chose to conduct focus groups with firefighters to allow for cross talk and the exchange of ideas on an infrequently discussed topic. We anticipated that leaders might have stronger preconceived notions of privacy norms specific to their roles and chose to conduct interviews to investigate their individual views in depth.

Data Collection

We developed interview and focus group guides to elicit views on data privacy. We aimed to understand how participants perceived several privacy domains: data collection, data use, and data access or sharing [21]. We developed questions that examined issues within these domains in the fire service and asked all participants questions within these domains, but we tailored questions as needed based on role. The guides were reviewed and refined by the study team. We collected demographic information from all participants. Multimedia Appendix 1 shows the interview and focus group discussion topics, including the key domains and high-level questions. Regardless of the data collection mode (interview vs focus group), we asked all participants about the same domains and then asked several follow-up questions to probe more deeply into interesting lines of inquiry related to the overall topic.

We conducted focus groups in person at fire stations. For some focus groups, firefighters were on duty and in service, meaning that they were available to respond to calls. If participants received a call during the discussion, it was paused or continued with the remaining participants until they returned. Interviews were conducted over Zoom and via phone. Data were collected from March to November 2023.

Consistent with qualitative research practices [22], the interviewer or moderator wrote a memo after each focus group or interview to capture initial themes and observations. All interviews and focus groups were audio recorded and transcribed verbatim by a professional transcription service.

Analysis

We analyzed interview and focus group transcripts using thematic analysis [23]. To develop an initial codebook, 1 team member reviewed the discussion guides, memos, and transcripts and drafted an initial set of codes capturing key concepts. A second team member reviewed the guides, memos, and transcripts and then updated the draft codebook. These 2 team members both coded a subset of interview and focus group transcripts, revised the codebook, and then independently coded the remaining transcripts. They continued to meet to discuss questions about their application of the codes and whether additional codes or refinements to the definitions were needed throughout the coding process. After completing a majority of the coding, the team members examined individual codes and groupings of codes that related to data privacy. The coding team identified themes capturing core concepts and trends in the data based on these codes and groupings of codes. We determined that we had reached thematic saturation within each stakeholder group when no new themes were emerging in later transcripts. The codebook, example code application, and themes were shared with the full study team periodically to obtain their input. Coding was conducted in Dedoose (version 10.0.59; SocioCultural Research Consultants, LLC) [24].

To compare perspectives across the fire service community, we grouped participants into 2 categories: firefighters (career firefighters and union leaders) and leadership (department leaders and national leaders). These 2 categories reflect conceptual similarities within groups (eg, union leaders are themselves firefighters who represent firefighters' interests), and there were similar results within each group. When describing our findings, we primarily refer to these 2 groups.

None of the coauthors have experience as volunteer or career firefighters, which may have impacted the study and the interpretation of results. The interviewer or moderator participated in ride-alongs with Maryland firefighters to gain insights into firefighting and fire service culture and discussed data privacy issues with local fire service collaborators in Maryland and Virginia to inform the study materials and analyses. One author is a commissioner of a Maryland fire department and may have brought a management perspective to the study. The authors discussed the study findings with study partners in the fire service to help interpret the results. The coders also consulted researchers with extensive experience conducting research with the fire service. This study follows the COREQ (Consolidated Criteria for Reporting Qualitative Research) guideline (Checklist 1) [25].

Ethical Considerations

This research was approved by the Johns Hopkins Bloomberg School of Public Health Institutional Review Board (24091). We obtained oral consent from all participants. Participants were not compensated for their participation. We deidentified interview transcripts to protect participant privacy.

Results

Sample

The sample included a total of 65 participants from 35 interviews and 4 focus groups (5 - 9 participants in each): 31 career firefighters, 2 union leaders, 21 department leaders, and 11 national leaders (Table 1).

XSL•FO
RenderX

Table . Interview and focus group participant roles by jurisdiction[a].

Participants	Virginia, n	Maryland, n	National, n	Total, n
Firefighters				
Career firefighters	30	1	—[b]	31
Union leaders	1	1	—	2
Leadership				
Department	7	14	—	21
Government	—	—	3	3
Advocacy	—	—	6	6
Research	—	—	2	2
Total	38	16	11	65

[a]The data were collected from March to November 2023.

[b]Not applicable.

We encountered data collection challenges in Maryland, which limited the firefighter sample in that state (1 Maryland firefighter participated in a key informant interview compared to 30 Virginia firefighters in focus groups); this limitation is discussed in the *Discussion* section. Overall, our sample was largely White and male, which reflects the fire service's gender and race demographics [26]. The leadership sample was more diverse in terms of race and gender: 25% (8/32) of leaders were non-White versus 9% (3/33) of firefighters, and 34% (11/32) of the leadership participants were women versus 12% (4/33) of firefighters. Leaders were also older and more experienced than firefighters: the mean age of leadership participants was 47.7 (SD 10.2) years versus 35.9 (SD 7.5) years, and the mean years of experience was 24.6 (SD 11.9) years versus 11.8 (SD 8.8) years. Table 2 contains self-reported demographic characteristics. Focus groups were, on average, 77 minutes. Firefighter interviews averaged 39 minutes, and leadership interviews averaged 45 minutes.

Table . Interview and focus group participant characteristics by role[a].

Participant characteristics	Firefighters (n=33)	Leadership (n=32)
Gender, n (%)		
Male	29 (88)	21 (66)
Female	4 (12)	11 (34)
Race, n (%)		
White	30 (91)	24 (75)
Non-White	3 (9)	8 (25)
Age (y), mean (SD)	35.9 (7.5)	47.7 (10.2)
Experience (y), mean (SD)	11.8 (8.8)	24.6 (11.9)

[a]The table shows self-reported demographic characteristics. One firefighter did not report their age. Two leadership participants did not report their years of experience. The data were collected from March to November 2023.

Themes

Overview

We identified four themes: (1) acceptability of current data access, sharing, and reporting practices; (2) data sharing and access preferences; (3) appropriate use of firefighter data; and (4) need for improved communication about data. Table 3 shows the themes and provides an overview of themes by group (firefighters and leadership).

XSL•FO
RenderX

Table . Overview of themes[a].

Themes	Leadership	Firefighters
Acceptability of current data access, sharing, and reporting practices	All participants described limited sharing of identifiable data and adequate protections for most health and safety data.	All participants described limited sharing of identifiable data and adequate protections for most health and safety data.
Data sharing and access preferences	Leaders described concerns about data sharing and ideal conditions for facilitating data collection, such as deidentification and limiting sharing.	Firefighters largely did not share these concerns but favored deidentification and identified sensitive kinds of information, such as mental health data.
Appropriate use of firefighter data	Leaders believed that firefighters were resistant to sharing information based on fear of termination or job impacts; many linked this to fears about losing identity as a firefighter.	For the most part, firefighters did not share these concerns.
Need for improved communication about data	Leaders acknowledged the importance of communicating well about data.	Firefighters and advocates emphasized the need for data interpretation and dissemination.

[a]The table shows themes from focus groups and interviews with firefighters and fire service leaders (March to November 2023).

Acceptability of Current Data Access, Sharing, and Reporting Practices

Within departments, firefighters and leadership described limited sharing of firefighter health and safety data, and most believed that such data were adequately protected by their departments. Nearly all leaders indicated that access to identifiable firefighter health or safety information was granted on a need-to-know basis. For instance, injury information was accessible only to health and safety staff. Most firefighters were required to complete an annual physical exam, but most departments did not receive detailed reports of members' annual physical results, just a notification of whether they passed or failed.

Externally, department leaders mentioned being legally required to report injuries to government agencies, such as the Occupational Safety and Health Administration, and third parties, such as the National Fire Protection Association and the International Association of Fire Fighters union. Most firefighters were uncertain about how the data might be reported externally but assumed that injury information was shared with government agencies or third parties and viewed this sharing as acceptable.

For the most part, firefighters appeared comfortable with the kinds and amounts of information they shared with their departments, as well as how this information was accessed by department leaders. Some simply did not express strong views on current data sharing practices, whereas others shared that efforts to limit data sharing were sufficient. Others stated that sharing personal information related to their physical health simply came with working in a high-risk, public job.

> *I don't have any real experiences or bad or positive experiences with my data in the department, except for just our health as in our yearly physicals and then stress tests and things like that. And I haven't had any issues with those or anything.* [Firefighter]

> *I'm sure no one loves to give all their personal information to an employer, but it's probably necessary in this particular field.* [Firefighter]

While privacy protocols and confidentiality were reported to be formally upheld, the culture of informal information sharing within departments differed; many firefighters and a few leaders noted that news about injuries traveled quickly.

> *Just like with the injury thing, you can decipher who's, we heard, oh, such and such went to [...] the mental health rehab center. Oh, this guy's on injury leave for the next three months. Wonder where he is. We can find out things by digging.* [Firefighter]

Although a few individuals were troubled by how quickly news about health diagnoses or injuries could spread in their departments, they pointed out that this typically occurred through informal channels in which firefighters shared information with each other, not through leadership.

Data Sharing and Access Preferences

While most firefighters were comfortable with their departmental data sharing practices, some expressed preferences regarding the kinds of information they were comfortable sharing and with whom they were comfortable sharing it. Many had an ideological preference for limiting access to identifiable information and opting for aggregated data sharing both within and outside of departments. Union representatives concurred—they described encountering broad resistance among firefighters to sharing medical information, although the firefighters in our sample did not express these concerns. Firefighters were generally supportive of sharing information with researchers conducting health and safety research, although some described their frustration that they did not receive individual results or overall results of studies in which they participated (described in more detail below).

Reluctance to share identifiable information appeared to be heightened based on the sensitivity of the data in question. One female firefighter pointed out that she viewed reproductive health information as more private than physical injury information. Mental health data were highly valued by most participants—firefighters emphasized the need for more mental health research and services—but they characterized mental health data as complicated to collect and access. Firefighters generally agreed that mental health data were closely held, both

XSL•FO
RenderX

among leadership and peer networks, but participants were still concerned about who might have access to identifiable sensitive information. In one focus group, firefighters described their dismay when asked to complete a mental health research survey that was theoretically anonymous but asked for information that participants believed could be used to identify them. Several firefighters explained:

> *Participant 4: Well, but that survey we just did, it has a lot of pretty in-depth. If you take it seriously, it has a lot of in-depth personal mental health [questions]. Like, for example, I think there's some on there like, have you ever considered committing suicide? Like, how many times in the last week have you felt like, that you were too depressed to come to work? ...*
>
> *Participant 3: And you don't know where it's going.*
>
> *Participant 4: You don't know where it's going.*
>
> *Many: Yeah.*
>
> *Participant 4: Then you get to the end. It's like basically you got to describe your eye color, you know. [laughter]* [Firefighters]

Leadership was aligned with firefighters in preferring that the data be deidentified or aggregated for research and prevention. For some, this was a matter of principle. Others thought that firefighters were less likely to share information if they believed it would be identifiable:

> *Even if you tell them it's going to nobody but me or nobody but this specific office, they are less likely to participate if there is identifying information.* [Department leader]

Others indicated that firefighters were reluctant to share mental or behavioral health information due to fear of reprisal and the sensitive nature of these data. Some research and advocacy participants thought that firefighters were more likely to share information with external entities than within their departments. There were mixed views among leaders about whether firefighters were opposed to sharing data within departments on principle, with some describing widespread privacy concerns and others indicating that such views were rare.

Beyond individual-level data collection and sharing, a small number of leaders expressed a desire to expand access to population-level data across the fire service. These participants described a tension between using data to advance health and safety priorities while respecting individual privacy rights. Several cited structural challenges in securely storing or accessing data, which prevented this kind of population-level analysis, but they believed it would be useful in the future.

Appropriate Use of Firefighter Data

Firefighters in our sample expressed few concerns about how their data were currently being used, whereas leadership described their views on appropriate ways of utilizing firefighter data and recalled instances in which they had heard concerns from firefighters. Leaders emphasized the importance of using firefighter data to improve health and safety but acknowledged that this could raise concerns about data being used against firefighters. Several department leaders stressed that health and safety data should be used in nonpunitive ways:

> *We try really hard in our office, at least the safety office, that we are not a punitive type of section of the department. Because once we become punitive, people stop talking to us. People stop trusting us. And our job is to try to make things safer for them. And we can't make things safer if we don't really truly understand what the problems are, you know?* [Department leader]

Leaders across all subgroups, as well as union representatives, believed that firefighters were fearful that sharing health or injury information could result in job loss or changes in job status, ultimately impacting their identities as firefighters. Many described encountering concerns among some firefighters that if they did not pass medical or fitness exams, they would be terminated. Most department leaders believed these views were unfounded and expressed a desire to use these exams for health and safety purposes:

> *Because you have firefighters that will be like, "a machine or a doctor or this METS [metabolic equivalents] does not dictate on how I can fight fire," which is true ... [but] all the research they've done with this, says that if your METS are below eight, you are at a higher risk of dying of an MI [myocardial infarction] or something out on fire scene. So that's the biggest issue I feel in the fire service. Everyone thinks that we're trying to take their job away or take a ton of money out of their pocket, what they don't understand is, no, we're trying to keep you healthy and safe and everything like that.* [Department leader]

In other cases, leaders anticipated firefighter concerns that sharing health information would adversely affect their day-to-day work even if their employment status remained unchanged:

> *I know a couple cases [of] people that never reported anything because they didn't want to go through the process. Because you got to go to the county clinic and then you got go to the headquarters, and most people get into this job for adrenaline rush and helping people. And then you're on modified duty, sitting behind a desk with limitations.* [Department leader]

Several individuals framed these concerns as fears of losing their identity as a firefighter:

> *Their fear isn't so much losing their job, their fear is losing their identity, who they are.* [Advocate and department leader]

While these concerns were widespread among leaders, most firefighters did not discuss job impacts. A few were worried that their annual physicals had become so in-depth that they might identify health risks for which they could be terminated or moved to light duty, although this perspective arose in only one focus group. Union leaders attributed hesitation about sharing medical information to underlying concerns about fitness

XSL•FO
RenderX

for duty and privacy but were still supportive of requiring annual physical exams.

Calls for Improved Communication About Data

As noted earlier, many participants advocated for sharing data with firefighters. However, participants also described shortcomings in how departments or researchers currently communicate with firefighters.

Regardless of how data originated, department and national leaders strongly endorsed the importance of returning individual data and interpretations of data back to firefighters. Many supported sharing information directly with individual firefighters, such as returning each individual's medical data from an annual exam to empower them to make healthy choices. Participants also emphasized the importance of communicating aggregated and analyzed data to provide education on topics, such as injury prevention or fitness practices. Department leaders and advocates highlighted the importance of partnering with trusted leaders, such as union officials, and communicating effectively about the purpose of data collection. A small number of participants described challenges in sharing data effectively, noting that the fire service or their department did not have a strong track record of effectively communicating about data.

Despite this widespread endorsement of high-quality communication from leadership, many firefighters noted that there was significant room for improvement. Firefighters were vocal about the need for better communication regarding data in general and better interpretation of the data that they provided for departmental or research purposes. Several firefighters wanted more information about the purpose behind data collection efforts or the significance of individual results. For instance, some described not knowing how to interpret annual physical exam results:

> *I'm going to voice off. I think the value sucks, to be honest. They [the fire department] do blood work and they do a full physical but nine times out of 10 we get a paper back that says, hey, there was something abnormal about your blood work or something we found, and that's it. No follow up, no telling you what it actually is. And to actually follow up with them to find out is really hard. They do the work, but they don't actually give you the results.* [Firefighter]

Others were frustrated with the lack of research results described earlier. The lack of communication about their own data and study results led them to conclude that a large share of data collection has no purpose or ultimate benefit to firefighters.

> *Participant 2: Well look at the urine samples that they [researchers] would collect, stuff like that. They had us give a urine sample after we got back from fires to see, hey, do you have whatever…*
>
> *Participant 9: This has been absorbed into your bloodstream.*
>
> *Participant 2: Arsenic, you know, absorbed into your body.*
>
> *Participant 3: We did it for two years and got…*
>
> *Participant 5: And they lost funding and...*
>
> *Participant 2: Right. They lost funding and it went away. Or every sample you gave, you got zero back from…*
>
> *Participant 5: No idea.* [Firefighters]

Discussion

Our study examined firefighters' and fire service leaders' data privacy preferences and identified opportunities for fire departments and researchers to ensure firefighter data are appropriately safeguarded and utilized. Leaders had more concerns about data use and data sharing, attributed in part to anticipated firefighter privacy concerns, whereas firefighters held general preferences for deidentifying data and limiting sharing, particularly for sensitive information, such as mental health data. Both groups called for improved data communication.

Interestingly, leaders in our sample voiced more concerns about current data practices than firefighters, and leaders were more attuned to the need for and barriers to ethical data collection, use, and sharing. This might reflect the fact that leadership was more likely to encounter diverse views on data than firefighters working in a single fire station or to hear concerns held by a small number of firefighters but elevated to management. It might also reflect the responsibilities that accompanied their roles—leaders who regularly interacted with firefighter data, or who were responsible for policy implementation, may have had more opportunities to consider the ethical and legal obligations surrounding such data. This aligns with research from our study team examining perceptions of the value of data in the fire service, which found that fire service leaders identified more use cases for occupational health and safety data than firefighters [27]. Leaders were likely more familiar with occupational health and safety data, more likely to value it, and therefore more likely to have strong views on the privacy of that data. Although leaders had more specific concerns about privacy practices than firefighters, both groups identified opportunities to limit access to sensitive data and improve communication.

The launch of new systems, such as NERIS, offers opportunities to adopt a "privacy by design" or "value-sensitive design" approach in which safeguards for privacy or other values are built directly into digital health systems [28-30]. Privacy by design focuses on privacy protections, whereas value-sensitive design accounts for moral values such as privacy, in addition to other values such as ease of use or safety benefits. Value-sensitive design may be a useful framework for fire service technology developers, who may be designing technologies to enhance one moral end (eg, protecting worker health and safety) but need to balance other moral values (eg, protecting privacy) and practical design considerations (eg, usability of a smartphone app for firefighters to track injuries).

For instance, firefighters in our sample expressed hesitancy to share sensitive information about their mental health despite recognizing a need for more mental health research and screening. A value-sensitive mental health data collection system in a mobile app could be designed to anonymize responses, impose limits on collecting sensitive data (eg, suicidal ideation),

XSL•FO
RenderX

or place limits on who has access to this information (eg, union leaders instead of department leaders). Leadership identified a broader array of potential data protections, such as limiting identifiable data collection and avoiding punitive use of health and safety data. The developers of digital health interventions for the fire service could take these preferences into account. For example, the developers of data surveillance systems, such as NERIS, could address these priorities by limiting the collection of identifiable information when gathering population-level injury data. Fire service technology companies could design wearable devices and smartphone apps that incorporate privacy features, such as allowing firefighters to opt into sharing specific identifiers or kinds of data with developers, their employers, researchers, or government officials. While these findings have implications for a range of digital health technologies and developers, fire departments, researchers, and government agencies can also help develop policies and practices accompanying these technical safeguards to ensure that health and safety data are used in protective, not punitive, ways.

Our findings also emphasize the importance of transparent and high-quality communication about firefighter health and safety data, both in research and practice settings. Many of the firefighters we spoke with were broadly supportive of research on topics, such as cancer, but were apathetic about their own contributions. Several described their frustration that after sharing sensitive information, such as symptoms of psychological distress or biospecimens, they never heard about the purpose or outcome of a study in which they participated. There are clear and well-established ethical imperatives for communicating the purpose of a study to research participants [31]. Our findings suggest there is strong demand from firefighters for researchers to clearly communicate about the objectives and conclusions of their work. Understanding these kinds of communication best practices will be particularly important as the National Institute for Occupational Safety and Health rolls out the National Firefighter Registry, a national-level longitudinal research effort. There are fewer well-established norms for communication about this kind of public health surveillance as opposed to individual research studies. Additional research on this topic could help establish guidelines on if and how firefighters want to receive information from surveillance activities such as the National Firefighter Registry.

Communicating individual research results is a more complicated endeavor; there is robust debate about when it is appropriate to return individual research results to participants, based on the kinds of data in question and how results might be communicated [32,33]. However, there is growing consensus that study participants want to receive medically significant and actionable findings, which is consistent with the views of firefighters in our sample [32,34]. While researchers cannot always ensure that findings are translated into practice, they can commit to providing research participants with individual results that might increase awareness about their own health and to sharing the conclusions of a study. Research ethics literature offers best practices for how researchers can communicate with participants, including offering context, using clear and simple language, and providing opportunities for questions [35,36].

Our findings also reveal the need for more research on the best manner in which to return health or medical information collected in occupational settings (not for research), for instance, from an annual physical exam or a department-led safety initiative. Researchers and employers operate in different environments, and the nature of data collection will vary accordingly. For example, firefighters may be required to participate in an annual exam as part of the terms of their employment, whereas research participants voluntarily enroll in a study. Research studies aim to produce generalizable knowledge and societal benefits, whereas a department's safety efforts might have limited population-level benefits. More work is needed to determine the ethical obligations of employers in communicating health and safety data to their employees. Ensuring ethical, high-quality communication about these kinds of data may not only benefit individual firefighters but could also help bolster trust in data collection efforts within the fire service writ large.

Finally, our findings reflect a need for robust ethical guidance on how to appropriately handle health and safety data in occupational settings, including the establishment of ethical principles as well as detailed, practical guidance on privacy topics. Although our participants did not raise major alarms about unethical data practices, they described legitimate concerns about ensuring firefighter data privacy while calling for large amounts of new data collection. The concerns that health and safety data could be used in punitive ways—to fire someone or change their job responsibilities without cause—reflect serious potential repercussions from insufficient data privacy protections. A few experts have discussed the need for confidential data practices for health and safety data, but guidelines for ethical data practices in any occupational setting remain general [37]. The COVID-19 pandemic revealed the need for ethics and data governance guidance for employers, as they began to deploy biometric screenings and digital contact tracing apps to attempt to control disease spread in workplaces, often with little ethical guidance [38,39]. While there were efforts to develop guidelines specific to COVID-19 [40], more work is needed to establish guidelines for the routine collection of occupational health and safety data, outside of a public health emergency. The continued refinement of ethical frameworks specific to occupational health and safety, as well as practical guidance on how to implement these frameworks, will help employers implement new data efforts in an ethical manner.

Our results should be interpreted with several key limitations in mind. Our findings reflect the specific geographic regions and fire departments in our sample. Fire departments were recruited from study team networks and had previously participated in research, which may have influenced their views toward research and data privacy. We encountered recruitment challenges in Maryland that limited the size of our firefighter sample in that state. We found similar results across Maryland and Virginia, but a larger sample may have captured additional viewpoints. Attitudes toward data privacy may differ based on department culture, trust in leadership, staffing (eg, combination, volunteer, or career), and geography. Future research should

XSL•FO
RenderX

explore this topic in a wider range of departments. Our sample was also largely White and male. While this reflects the overall demographics of the fire service [26], evidence from other settings suggests that women and racial minorities experience data and technology in distinct and often inequitable ways and thus may have unique privacy preferences [41-44]. More research is needed to understand how perspectives on occupational health and safety data privacy might vary based on gender and race.

Our findings offer insights into pathways for enhancing firefighter data privacy. Incorporating firefighters' privacy preferences into digital health data systems, for instance, by deidentifying firefighter data and limiting data access, may build confidence and trust in data collection efforts. Improving communication about data will address firefighters' desire to hear more about the purpose and impact of data collection. By addressing firefighters' privacy concerns and conducting high-quality, tailored communication, government, research, and department-level leaders can foster trust in data collection efforts and ensure that firefighter data are used in ethical and acceptable ways to protect their health and safety.

Acknowledgments

Generative AI was not used at any stage in the preparation of this manuscript.

Funding

RJT received tuition and stipend support from a grant from the US Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health to the Johns Hopkins Education and Research Center for Occupational Safety and Health (award T42 OH0008428), as well as the Susan P. Baker Scholarship in Injury Prevention and Control and Center for Qualitative Studies in Health and Medicine Dissertation Enhancement Award. AW was supported by grant T32AG066576 from the National Institute on Aging, National Institutes of Health. Data collection and analysis were funded by a pilot project award from the National Institute for Occupational Safety and Health Johns Hopkins Education and Research Center for Occupational Safety and Health (award T42 OH0008428).

Data Availability

The data are not available to protect participant confidentiality. The fire service communities studied are small, raising concerns about identifiability.

Authors' Contributions

Conceptualization: CKC, JA, RJT, SF
Data curation: AW, RJT
Formal analysis: AW, RJT
Funding acquisition: CKC, RJT
Investigation: AW, RJT
Methodology: AW, CKC, JA, RJT, SF
Project administration: RJT
Supervision: CKC, JA, SF
Writing – original draft: RJT

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview and focus group discussion topics.
[DOCX File, 31 KB - jmir_v28i1e84465_app1.docx]

Checklist 1
COREQ checklist.
[PDF File, 190 KB - jmir_v28i1e84465_app2.pdf]

References

1. Fahy R, Evarts B, Stein GP. U.S. fire department profile. National Fire Protection Association. 2022. URL: https://www.nfpa.org/education-and-research/research/nfpa-research/fire-statistical-reports/us-fire-department-profile [accessed 2023-12-01]

XSL•FO
RenderX

2. Fire department overall run profile as reported to the national fire incident reporting system (2020). : U.S. Fire Administration; 2020 URL: https://www.usfa.fema.gov/downloads/pdf/statistics/v22i1-fire-department-run-profile.pdf [accessed 2026-03-17]
3. Campbell R, Hall S, Petrillo J. United States firefighter injuries. National Fire Protection Association. URL: https://www.nfpa.org/education-and-research/research/nfpa-research/fire-statistical-reports/firefighter-injuries-in-the-united-states [accessed 2026-03-17]
4. Campbell R, Petrillo JT. Fatal firefighter injuries in the US in 2022. : National Fire Protection Association; 2023 Jun 18 URL: https://img1.wsimg.com/blobby/go/95969465-abe3-4680-bcb1-af5b16fc7ebd/downloads/2022%20LODD%20Report.pdf?ver=1741883121806 [accessed 2026-03-17]
5. Lee DJ, Ahn S, McClure LA, et al. Cancer risk and mortality among firefighters: a meta-analytic review. Front Oncol 2023;13:1130754. [doi: 10.3389/fonc.2023.1130754] [Medline: 37251928]
6. Laroche E, L'Espérance S. Cancer incidence and mortality among firefighters: an overview of epidemiologic systematic reviews. Int J Environ Res Public Health 2021 Mar 3;18(5):2519. [doi: 10.3390/ijerph18052519] [Medline: 33802629]
7. National firefighter registry (NFR) for cancer protocol. : Centers for Disease Control and Prevention (CDC); 2025 URL: https://www.cdc.gov/niosh/media/pdfs/2025/05/NFR-Protocol-Aprl_2025_OMB.pdf [accessed 2026-03-23]
8. Grant C, Hamins A, Bryner N, Jones AW, Koepke GH. Research roadmap for smart fire fighting. : National Institute of Standards and Technology; 2015 Jun URL: https://nvlpubs.nist.gov/nistpubs/specialpublications/nist.sp.1191.pdf [accessed 2026-03-17]
9. Widman SA, LeVasseur MT, Tabb LP, Taylor JA. The benefits of data linkage for firefighter injury surveillance. Inj Prev 2018 Feb;24(1):19-28. [doi: 10.1136/injuryprev-2016-042213] [Medline: 28196830]
10. Kahn SA, Mosier M. Firefighter injury data are wrought with discrepancy: pearls from the national burn repository. J Burn Care Res 2018 Jun 13;39(4):623-625. [doi: 10.1093/jbcr/irx025] [Medline: 29860394]
11. Averill J, Moore L, Notarianni K, Santos R, Wissoker D. Proceedings of the national fire service data summit. : National Institute of Standards and Technology; 2011 URL: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=908244 [accessed 2026-03-23]
12. McClure LA, Koru-Sengul T, Hernandez MN, et al. Availability and accuracy of occupation in cancer registry data among Florida firefighters. PLoS ONE 2019;14(4):e0215867. [doi: 10.1371/journal.pone.0215867] [Medline: 31039169]
13. Weidinger J. What is known and what remains unexplored: a review of the firefighter information technologies literature. Int J Disaster Risk Reduct ;78(2022):103115. [doi: 10.1016/j.ijdrr.2022.103115]
14. Taborri J, Pasinetti S, Cardinali L, Perroni F, Rossi S. Preventing and monitoring work-related diseases in firefighters: a literature review on sensor-based systems and future perspectives in robotic devices. Int J Environ Res Public Health 2021 Sep 15;18(18):9723. [doi: 10.3390/ijerph18189723] [Medline: 34574646]
15. About the National Emergency Response Information System (NERIS). US Fire Administration. 2023. URL: https://www.usfa.fema.gov/nfirs/neris/about-neris/ [accessed 2023-07-23]
16. National Firefighter Registry (NFR) for cancer. CDC. URL: https://www.cdc.gov/niosh/firefighters/registry.html [accessed 2022-07-22]
17. Topazian RJ, Wec A, Ali J, et al. Ethical use of wearable device data in occupational settings: perspectives from the fire service. J Occup Environ Med 2025 Sep 1;67(9):e621-e629. [doi: 10.1097/JOM.0000000000003446] [Medline: 40359076]
18. Rumsey A, Le Dantec CA. Clearing the smoke: the changing identities and work in firefighters. 2019 Presented at: DIS '19: Proceedings of the 2019 on Designing Interactive Systems Conference; Jun 23-28, 2019; San Diego CA, USA p. 581-592. [doi: 10.1145/3322276.3322292]
19. Carpenter D, McLeod A, Hicks C, Maasberg M. Privacy and biometrics: an empirical examination of employee concerns. Inf Syst Front 2018 Feb;20(1):91-110. [doi: 10.1007/s10796-016-9667-5]
20. Taylor JA, Davis AL, Shepler LJ, et al. Development and validation of the fire service safety climate scale. Saf Sci 2019 Oct;118:126-144. [doi: 10.1016/j.ssci.2019.05.007]
21. Richards N. Why Privacy Matters: Oxford University Press; 2021. [doi: 10.1093/oso/9780190939045.001.0001]
22. Miles MB, Huberman MA. Qualitative Data Analysis: An Expanded Sourcebook, 2nd edition: SAGE Publications, Inc; 1994. URL: https://openlibrary.org/books/OL21978508M/Qualitative_data_analysis [accessed 2026-03-17]
23. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
24. dedoose. URL: https://www.dedoose.com/ [accessed 2026-03-17]
25. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
26. Employed people by detailed occupation, sex, race, and Hispanic or Latino ethnicity. US Bureau of Labor Statistics. URL: https://www.bls.gov/cps/cpsaat11.htm [accessed 2024-01-03]
27. Topazian RJ, Wec A, Ali J, et al. The value of occupational health and safety data: perceptions of current data practices in the fire service. J Safety Res 2025 Sep;94:120-127. [doi: 10.1016/j.jsr.2025.06.016] [Medline: 40930628]

XSL•FO
RenderX

28. Cavoukian A. Privacy by design: the 7 foundational principles—implementation and mapping of fair information practices. : Office of the Information & Privacy Commissioner of Ontario; 2006 URL: https://student.cs.uwaterloo.ca/~cs492/papers/7foundationalprinciples_longer.pdf [accessed 2026-03-17]
29. Schaar P. Privacy by design. Ident Inf Soc 2010 Aug;3(2):267-274. [doi: 10.1007/s12394-010-0055-x]
30. Friedman B, Kahn PH, Borning A. Value sensitive design and information systems. In: The Handbook of Information and Computer Ethics: John Wiley & Sons, Inc; 2008:69-101. [doi: 10.1002/9780470281819]
31. The Belmont Report: ethical principles and guidelines for the protection of human subjects of research. : U.S. Department of Health & Human Services; 1979 URL: https://www.hhs.gov/ohrp/sites/default/files/the-belmont-report-508c_FINAL.pdf [accessed 2024-08-30]
32. Botkin JR, Mancher M, Busta ER, Downey AS, editors. Returning Individual Research Results to Participants: Guidance for a New Research Paradigm: National Academies Press; 2018. [doi: 10.17226/25094] [Medline: 30001048]
33. Wong CA, Hernandez AF, Califf RM. Return of research results to study participants: uncharted and untested. JAMA 2018 Aug 7;320(5):435-436. [doi: 10.1001/jama.2018.7898] [Medline: 29931289]
34. Vears DF, Minion JT, Roberts SJ, et al. Return of individual research results from genomic research: a systematic review of stakeholder perspectives. PLOS ONE 2021;16(11):e0258646. [doi: 10.1371/journal.pone.0258646] [Medline: 34748551]
35. Hintz EA, Dean M. Best practices for returning research findings to participants: methodological and ethical considerations for communication researchers. Commun Methods Meas 2020 Jan 2;14(1):38-54. [doi: 10.1080/19312458.2019.1650165]
36. Downey AS, Busta ER, Mancher M, Botkin JR, editors. Principles for the return of individual research results: ethical and societal considerations. In: Returning Individual Research Results to Participants: Guidance for a New Research Paradigm: National Academies Press; 2018:59-92 URL: https://www.ncbi.nlm.nih.gov/books/NBK525079 [accessed 2024-08-30]
37. Rogers B, Schill AL. Ethics and Total Worker Health®: constructs for ethical decision-making and competencies for professional practice. Int J Environ Res Public Health 2021 Sep 24;18(19):10030. [doi: 10.3390/ijerph181910030] [Medline: 34639332]
38. Gidengil C, Fischer S, Broten N. A framework for evaluating approaches to symptom screening in the workplace during the COVID-19 pandemic. : RAND Corporation; 2020 URL: https://www.rand.org/pubs/perspectives/PEA653-1.html [accessed 2026-03-17]
39. Ingram C, Downey V, Roe M, et al. COVID-19 prevention and control measures in workplace settings: a rapid review and meta-analysis. Int J Environ Res Public Health 2021 Jul 24;18(15):7847. [doi: 10.3390/ijerph18157847] [Medline: 34360142]
40. Kahn JP. Digital Contact Tracing for Pandemic Response: Ethics and Governance Guidance: Johns Hopkins University Press; 2020. URL: https://muse.jhu.edu/pub/1/oa_monograph/book/75831 [accessed 2024-10-01]
41. Benjamin R. Race after Technology: Abolitionist Tools for the New Jim Code: Polity; 2019. URL: https://www.politybooks.com/bookdetail?book_slug=race-after-technology-abolitionist-tools-for-the-new-jim-code--9781509526390 [accessed 2026-03-17]
42. Allen AL. Dismantling the “black opticon”: privacy, race, equity, and online data-protection reform. Yale Law Journal Forum. 2021 Nov. URL: https://www.yalelawjournal.org/forum/dismantling-the-black-opticon [accessed 2026-03-17]
43. Stark L, Stanhaus A, Anthony DL. “I don’t want someone to watch me while I’m working”: gendered views of facial recognition technology in workplace surveillance. J Assoc Inf Sci Technol 2020 Sep;71(9):1074-1088. [doi: 10.1002/asi.24342]
44. Jacobs JV, Hettinger LJ, Huang YH, et al. Employee acceptance of wearable technology in the workplace. Appl Ergon 2019 Jul;78:148-156. [doi: 10.1016/j.apergo.2019.03.003] [Medline: 31046946]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
NERIS: National Emergency Response Information System

Edited by A Stone; submitted 23.Sep.2025; peer-reviewed by KF Hollis, S Mohanadas, SJ Kim; revised version received 20.Feb.2026; accepted 23.Feb.2026; published 02.Apr.2026.

Please cite as:
Topazian RJ, Wec A, Ali J, Frattaroli S, Crifasi CK
Ethical Handling of Occupational Health and Safety Data in the Fire Service: Empirical Interview and Focus Group Study of Firefighter and Fire Service Leadership Privacy Preferences
J Med Internet Res 2026;28:e84465
URL: https://www.jmir.org/2026/1/e84465
doi:10.2196/84465

XSL•FO
RenderX

© Rachel J Topazian, Aleksandra Wec, Joseph Ali, Shannon Frattaroli, Cassandra K Crifasi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Ethics of Leveraging Routinely Collected Patient Data for AI Development: Mixed Methods Study

Menno T Maris[1,2], MSc; Joanna E Klopotowska[2,3], PhD, PharmD; Ronald Cornet[2,3], PhD; Mariëtte A Van den Hoven[2,4], PhD; Joris E Lieverse[2,3], PharmD; Daniel Fernández-Llaneza[2,3], MSc; Marieke A R Bak[1,2,5], PhD; LEAPfROG Consortium[6,7]

[1]Department of Ethics, Law and Humanities, Amsterdam UMC Location University of Amsterdam, Meibergdreef 9, Amsterdam, The Netherlands
[2]Amsterdam Public Health Institute, Amsterdam, The Netherlands
[3]Department of Medical Informatics, Amsterdam UMC location University of Amsterdam, Amsterdam, The Netherlands
[4]Department of Ethics, Law and Humanities, Amsterdam UMC location Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[5]Institute of History and Ethics in Medicine, TUM School of Medicine, Technical University of Munich, Munich, Germany
[6]See Acknowledgments
[7]

Corresponding Author:
Menno T Maris, MSc
Department of Ethics, Law and Humanities, Amsterdam UMC Location University of Amsterdam, Meibergdreef 9, Amsterdam, The Netherlands

Abstract

Background: Electronic health record (EHR) data, a key form of routinely collected patient data, offer great potential for medical research and the development of artificial intelligence (AI) tools. However, because these data are primarily gathered for health care rather than research, it often lacks the quality needed for AI training, raising both methodological and ethical concerns. While previous studies have reviewed the ethical implications of both routinely collected patient data and AI separately, their intersection, where AI is applied to such data, remains largely unexplored.

Objective: This study aimed to examine the ethical challenges that arise at the intersection of EHR data and AI development and to derive practice-oriented recommendations using the Dutch LEAPfROG (Leveraging Real-World Data to Optimize Pharmacotherapy Outcomes in Multimorbid Patients Using Machine Learning and Knowledge Representation Methods) project as a guiding case.

Methods: We used a mixed methods design combining a scoping literature review with a systematic search and 2 stakeholder workshops structured by the guidance ethics approach, reflecting a staged and iterative process aligned with the LEAPfROG project's development phases. The review identified 25 relevant publications from 2014 to 2024. The workshops, conducted with 17 and 13 participants respectively, included patients, clinicians, ethicists, data officers, and AI developers. Both workshops used dialogue to identify ethical values, impacts, and action points, focusing on a case study of drug-induced acute kidney injury.

Results: The analysis highlighted four major themes: (1) data privacy, transparency, and consent, including challenges of meaningful consent and risks of reidentification; (2) public trust and regulatory challenges, such as fragmented oversight and inconsistent governance; (3) fair representation and model generalizability, where incomplete or biased data may perpetuate health inequities; and (4) responsible AI integration in clinical practice, including concerns about clinical tropism, administrative burden, and the risk of overreliance on AI outputs. Both literature and stakeholder perspectives underscore the risk of decontextualization when EHR data are reused and emphasize the importance of clearly defining the purpose of data reuse to ensure real-world applicability and foster trust.

Conclusions: Responsible AI development requires explicit attention to how EHR data are produced, interpreted, and governed in practice, recognizing that data quality and meaning are shaped by the clinical, institutional, and social contexts in which they originate. Technical solutions or top-down regulation alone are insufficient. Instead, stakeholder-led and context-sensitive approaches are needed to define the purposes, risks, and benefits of medical AI. Grounded in the realities of health care practice and in the perspectives of patients, clinicians, and data custodians, these approaches can strengthen transparency, fairness, and clinical relevance, ensuring that EHR data are used ethically and effectively to support equitable and trustworthy AI innovation.

(J Med Internet Res 2026;28:e79863) doi:10.2196/79863

XSL•FO
RenderX

KEYWORDS

artificial intelligence; routinely collected health data; ethics; stakeholder participation; medical informatics; pharmacotherapy

Introduction

Both routinely collected patient data and artificial intelligence (AI) are widely regarded as promising resources for advancing medical research. AI methods are valued for their ability to manage and analyze large, diverse datasets, making them potentially well-suited to maximize the potential of routinely collected patient data. Studies leveraging such data offer several advantages over other study designs, including lower administrative costs, adaptability to evolving practice patterns, and access to larger sample sizes [1]. One of the fields where leveraging routinely collected data with AI seems particularly promising is individualized pharmacotherapy, focusing on the "rational use of drugs" to tailor pharmacotherapy to patients' clinical needs, thereby maximizing benefits and minimizing harm [2]. Effective innovation in pharmacotherapy is currently especially challenging for patients with multimorbidity, characterized by the presence of 2 or more chronic diseases and a frequent need for multiple medications, a phenomenon known as polypharmacy [3,4]. Despite the increasing prevalence of multimorbidity, partly due to population aging and improvements in health care services [5], clinical programs, guidelines, and research continue to primarily focus on managing individual diseases [3,6].

By leveraging routinely collected patient data from electronic health records (EHRs) with AI, the Dutch LEAPfROG Project (Leveraging Real-World Data to Optimize Pharmacotherapy Outcomes in Multimorbid Patients Using Machine Learning and Knowledge Representation Methods) seeks to enhance medication safety for patients with multimorbidity. A central focus of the project is the clinically urgent use case of drug-induced acute kidney injury (DAKI) in patients with chronic kidney disease (CKD). Ultimately, LEAPfROG aims to develop high-quality and fit-for-purpose EHR data, adhering to the principles of findability, accessibility, interoperability, and reusability [7] and to develop robust analytic methods, providing real-world evidence on pharmacotherapy outcomes and contributing to learning health systems that offer personalized treatments, improve quality of life, and reduce health care costs (Textbox 1).

Textbox 1. LEAPfROG (Leveraging Real-World Data to Optimize Pharmacotherapy Outcomes in Multimorbid Patients Using Machine Learning and Knowledge Representation Methods) project: improving drug safety in chronic kidney disease (CKD) through electronic health record (EHR) data and artificial intelligence (AI).

Background

The LEAPfROG project leverages routinely collected EHR data and structured domain knowledge to improve medication safety for patients with multimorbidity, specifically aiming to reduce the risk of drug-induced acute kidney injury (DAKI) in patients with CKD.

The LEAPfROG approach combines data from 3 EHR sources: detailed EHRs from Amsterdam UMC hospitals, EHRs from the Academic Network of General Practices Amsterdam (ANHA), and EHRs from general practices and outpatient pharmacies within the PHARMQ Database Network, covering general practices, outpatient pharmacies, and hospitals. Together, these datasets provide a rich source of insights into patient pharmacotherapy outcomes and drug exposures. To enhance these insights, LEAPfROG integrates domain knowledge from resources such as drug monographs, medication safety alerts, and scientific literature.

A key goal of the project is to develop transparent, explainable tools that can assist clinicians in making well-informed prescribing decisions. LEAPfROG's AI models combine machine learning with knowledge representation methods to detect DAKI, assess DAKI risk in patients with CKD, and clarify the underlying factors contributing to kidney safety issues. By doing so, the project aims to improve the detection of adverse drug events such as DAKI and enhance patient outcomes, particularly for patients with multimorbidity, such as those with CKD.

Causal machine learning model for retrospective diagnosis of DAKI

One of LEAPfROG's aims is the development of a causal machine learning model to assist physicians in determining whether acute kidney injury was caused by a drug (combination), a process known as "retrospective diagnosis." This model aims not only to identify CKD patients who developed DAKI but also to explain cause-and-effect relationships underlying the DAKI diagnosis. Through this effort, LEAPfROG aims to create a large dataset of patients who did and did not develop DAKI. This dataset is critical for training future machine learning models capable of identifying patients at risk of DAKI ("prognostic models"), ultimately helping to prevent such adverse drug events in future CKD patients.

Because EHR data are collected during routine care for clinical purposes, they often have quality limitations when reused for research, lacking the consistency required by clinical research standards [8,9]. Integrating AI with EHR data may therefore exacerbate or further obscure existing data shortcomings, ultimately impacting their meaningful use in clinical practice [10,11].

Previous studies have systematically reviewed ethical issues related to leveraging health data for medical research, including concerns related to the sharing and linkage of health data [12], public attitudes towards the reuse of health data for research [13], and broader ethical considerations of using EHRs for biomedical research [14]. Regarding the ethical implications of AI in health care, previous studies have reviewed both epistemic ethical issues (ie, the type of knowledge AI generates) and normative ethical concerns [15], along with empirical investigations into ethical considerations specific to medical AI [16]. However, literature reviews on the ethical considerations inherent in the integration of EHR data and AI technologies are currently nonexistent.

In this study, we aim to explore the ethical implications of reusing EHR data for AI-driven research and innovation in health care and to provide actionable recommendations based on our use case, the LEAPfROG project. This article presents

XSL•FO
RenderX

the results of a mixed methods approach that combines a scoping literature review with a systematic search and 2 innovative stakeholder engagement workshops grounded in the guidance ethics approach (GEA) [17]. The scoping review informed the design of the stakeholder workshops, which in turn generated empirical insights that complemented and contextualized the review's findings.

Methods

Study Design

The study followed an iterative, participatory design that combined a scoping review and stakeholder engagement workshops to explore the ethical implications of reusing EHR data for AI-driven health research. The scoping review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) extension for scoping reviews [18]. No protocol for the scoping review component of this study was registered. The stakeholder workshops were structured using the GEA for evaluating emerging technologies [17]. The overall design was staged and sequential rather than triangulated, with each step reflecting a different stage of the LEAPfROG project. This iterative process reflected the real-world complexity of AI ethics, allowing themes to unfold progressively, from broad reflection on EHR reuse to more specific considerations of AI development and implementation.

The workshops were integrated into the staged structure of the project, with their timing and number determined in advance and aligned with key project phases. Rather than aiming for thematic saturation, the workshops served as structured points of ethical reflection tied to the specific technology and context [17]. The scoping review identified ethical themes in the literature, providing a broad understanding of existing ethical considerations, while the workshops enabled further exploration and contextualization within the real-world context of the LEAPfROG project.

In the first workshop, participants discussed the ethical implications of reusing EHR data in the Dutch context, focusing on DAKI in patients with CKD. As these discussions aligned with initial literature findings, the subsequent scoping review shifted focus from the ethics of EHR reuse to ethical considerations at the intersection of EHR data and AI. In the second workshop, preliminary findings from the literature were presented to validate, contextualize, and further refine the identified themes. Furthermore, this workshop centered on a concrete AI-based tool within the LEAPfROG project: a causal machine learning model for the retrospective diagnosis of DAKI in patients with CKD.

Scoping Review

Search Strategy

MTM, JEK, RC, and MB jointly developed the search strategy to identify articles that thoroughly examine or discuss the ethical considerations associated with leveraging EHR data for advancing medicine using AI. A comprehensive search was conducted by MTM in PubMed, CINAHL (via EBSCO), and Web of Science, restricted to results between March 2014 and March 2024. The search was structured around 4 key concepts: "Artificial Intelligence," "Ethics," "Research," and "Routinely collected patient data" (Multimedia Appendix 1). Additionally, the most relevant MeSH (Medical Subject Headings) terms, synonyms, and related terms were identified, leading to the development of final search strings for each database. Results from the 3 databases were merged in EndNote (version 21; Clarivate Analytics), and duplicates were removed.

Selection Process and Analysis

One researcher (MTM) screened titles and abstracts using Rayyan [19]. Eligible studies were written in English and addressed ethical considerations at the intersection of AI, including machine learning and natural language processing, that analyze routinely collected patient data. Following Benchimol et al [20], such data were defined as health data collected without specific a priori research questions, including observational data from EHRs and disease registries. Particular emphasis during selection was placed on studies using EHRs or data comparable in origin, as these form the primary data source within the LEAPfROG project and are central to the ethical questions explored in this study. Because terminology varies across publications, the search strategy was intentionally broad to capture studies referring to similar data under related terms (eg, clinical or electronic medical records). Studies that also discussed other data sources (eg, clinical trial or wearable data) were included, provided the focus was sufficiently on the reuse of EHRs. Any uncertainties during full-text assessment were resolved by consensus between 2 researchers (MTM and MB). Reference lists were screened for additional relevant studies.

We uploaded full-text articles to MaxQDA and performed open coding to explore the ethical implications of using routinely collected EHR data for AI-driven research and innovation. Data extraction and coding were treated as the data charting process for the scoping review. Using a subset of 10 articles, one researcher (MTM) developed a coding scheme based on overlapping ethical considerations and emerging themes, such as "privacy" and "doctor-patient relationship." This framework was discussed with the second researcher (MB) throughout the process and guided the coding of the remaining articles by MTM.

Insights from the scoping review informed both the design and thematic focus of the second stakeholder workshop. The resulting data were analyzed inductively to explore how participants interpreted and prioritized ethical themes. This analysis refined and contextualized the review's findings, emphasizing those most pertinent to real-world practice within the LEAPfROG project and completing the staged, iterative process.

Guidance Ethics Workshops

Workshop Design

We organized 2 stakeholder engagement workshops using the GEA (translated from the Dutch "Aanpak Begeleidingsethiek"), a relatively new method for addressing the ethical aspects of technological innovation [17]. The workshops were conducted

XSL•FO
RenderX

in Dutch. The GEA uses dialogue to map the most pressing moral considerations in technology implementation, as identified by stakeholders [21]. The GEA primarily serves to guide the development of a technology by prioritizing ethical considerations relevant to the specific case, while striving to align with stakeholder values.

Both workshops, which lasted 3 hours each, followed the GEA approach and comprised three separate stages: (1) providing a clear description of the technology and its context, (2) identifying stakeholders, potential impacts, and underlying values, and (3) generating a list of options for responsible innovation and implementation from 3 distinct perspectives: technology, environment, and user (Textbox 2). The first workshop was prepared by JEK, RC, MTM, and MB and moderated by external facilitators from Stichting ECP. The second was prepared by JEK, MTM, MB, MvdH, and DF-L and moderated by JEK, MTM, MvdH, JEL, and MB.

Textbox 2. Stakeholder workshops design "guidance ethics approach."

Stage 1: Technology in context

i. Introduction by project leaders: Description and use context.

ii. Stakeholder inquiry: Opportunity for participants to gain a deeper understanding of and explore the details of the intended technology.

Stage 2: Dialogue

i. Actors: Individuals or groups involved in or affected by the proposed technology or solution.

ii. Effects: Potential benefits and harms arising from the development and implementation of the technology or solution.

iii. Values: Examination of the most pertinent ethical considerations related to the effects from step II.

Stage 3: Action points

Action points to promote responsible innovation at 3 levels of abstraction:

i. Technology*:* How should it be designed?

ii. Environment: Includes both tangible elements (eg, physical infrastructure) and systematic aspects (eg, policy).

iii. User: What can users do?

Participant Recruitment

Recruitment took place within the Dutch health care and research context through 2 structured routes. First, invitations were distributed via the LEAPfROG consortium network, including affiliated partners and organizations, following the project's progress. Second, patient participants were recruited through the Dutch Kidney Patients Association (Nederlandse Vereniging voor Nierpatiënten [NVN]). When invitees were unavailable, comparable participants were identified outside these routes to maintain diversity across stakeholder groups. All invited participants received a written background document explaining the LEAPfROG project, the purpose of the workshops, and their expected contribution. No prior preparation was required for participation. Patients were invited as experiential experts rather than professionals and were offered an introductory meeting to ask questions and familiarize themselves with the process. All participants provided informed consent before the workshops took place.

Before the workshops, participants were notified of the intention to publish a report on the session results and were given the opportunity to raise objections. Prior to publication, they were able to request changes and provide consent for the final reports, which are available in Dutch as Multimedia Appendices 2 and 3. The first workshop, led by external moderators from Stichting ECP, was not audio-recorded. To further strengthen the accuracy and completeness of the documentation, the second workshop was audio recorded, with participants informed in advance and given the opportunity to object.

Documentation and Validation

In both workshops, discussions were documented on flip charts or a shared screen, allowing participants to review and refine notes in real time. Notes from moderators and researchers were later compared and synthesized into validated reports. The GEA emphasizes collective ethical reflection, in which differing perspectives are explicitly explored rather than reconciled into consensus. In the first workshop, notes were taken independently by 2 external moderators from Stichting ECP, alongside 2 researchers (JEK and MTM). The moderators' notes from the first session formed the basis for the initial draft of the first workshop report. Notes from JEK and MTM were compared, reconciled, and extensively discussed within the research team to ensure completeness and accuracy. This iterative process led to the final validated report. The first draft of the second workshop report was prepared by MTM and subsequently finalized through joint review by MTM, JEK, and JEL.

Workshop Data Analysis

The workshop discussions, facilitator notes, and summary reports were thematically analyzed using the GEA framework, which emphasizes dialogue and collective reflection among diverse stakeholders. The analysis focused on how participants articulated ethical and practical considerations across the 3 reflective phases of the GEA: mapping the technology in context, ethical reflection, and translation to action. MTM led the review and synthesis of the workshop materials, identifying recurring themes and insights that emerged through these dialogues. The resulting reports were designed as standalone

XSL•FO
RenderX

outputs for dissemination within the LEAPfROG consortium. For the purpose of this paper, the analysis focused on themes most relevant to the research gap at the intersection of EHR data and AI development. The coding framework developed from the scoping review provided the analytical foundation, guiding the thematic structure while allowing refinement and contextualization based on stakeholder perspectives. Themes were subsequently reviewed and refined with the coauthors during manuscript preparation to ensure consistency and alignment with the broader study.

Ethical Considerations

The overall LEAPfROG project protocol was reviewed by the Medical Ethics Committee of Amsterdam UMC (the Netherlands). The committee granted a waiver of formal approval (W22_340 #22.412) because the LEAPfROG project, including all substudies, does not fall within the scope of the Dutch Medical Research Involving Human Subjects Act (WMO). All workshop participants were informed about the study objectives and procedures and provided informed consent to participate. Participation was voluntary and unpaid. Travel expenses were reimbursed where applicable. Participants were informed that workshop reports would be made publicly available, and any attribution of names or roles in those reports occurred with participants' knowledge and consent.

Results

Included Literature

The results of the literature screening are presented in a PRISMA flow diagram (Figure 1). In total, 25 articles met the inclusion criteria. Most were commentary-style publications (eg, perspectives and debates; n=14), alongside review-type articles (n=8) and a smaller number of qualitative studies (n=3). In terms of medical focus, the largest group of articles explored themes within general health care and clinical informatics (n=9). Other represented areas included medical ethics and health equity (n=5), primary care (n=3), public health and health disparities (n=2), geriatrics (n=1), psychiatry (n=1), nephrology (n=1), rheumatology (n=1), pediatrics (n=1), and orthopedics (n=1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) flow diagram of study selection. AI: artificial intelligence.

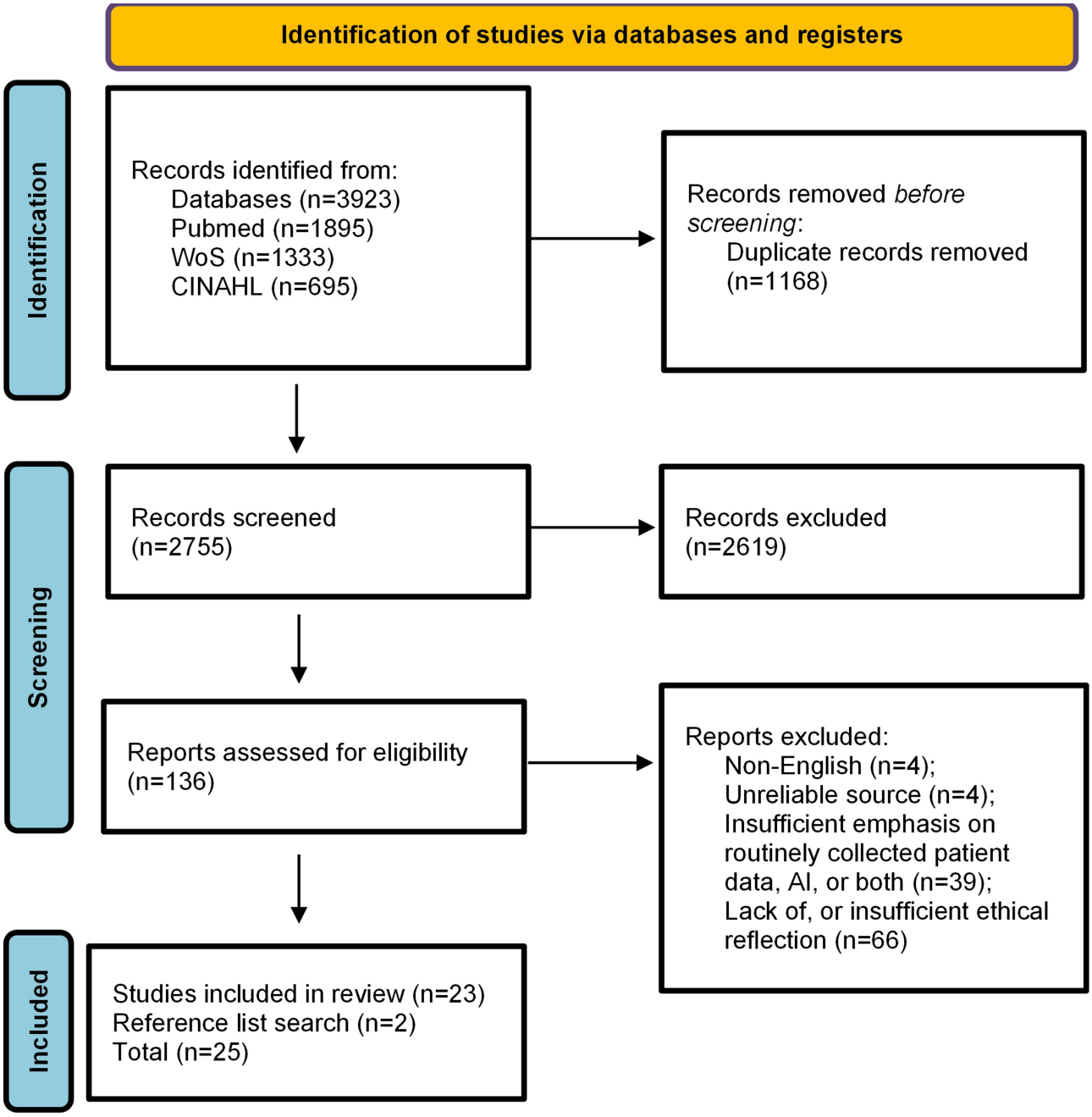

In terms of geographical distribution, the majority of articles were published by authors based in North America (n=14), followed by Europe (n=8), Asia (n=2), and Australia (n=1). A detailed overview of article characteristics is available in Multimedia Appendix 4 [22-46].

Workshop Participants

Two stakeholder workshops were held, with 17 participants in the first and 13 in the second, with some attending both. Participants included medical ethicists (n=3), medical informatics experts and AI engineers (n=9), patients associated with the Kidney Patients Association Netherlands (NVN) (n=2), patient representatives from the NVN (n=2), a nephrologist, a chief medical information officer (n=1), a chief scientific information officer and endocrinologist (n=1), a general practitioner (n=1), pharmacists (n=2), a digital transition manager, data protection officers (n=2), and a digital ethics manager.

Key Themes From the Literature and Integration of Stakeholder Workshops

This section is structured around 4 key themes: data privacy, transparency, and consent; public trust and regulatory challenges; fair representation and model generalizability; and responsible AI integration in clinical practice. This structure reflects the staged design of the LEAPfROG project and the cyclical nature of AI development in health care, which typically extends across data governance and quality, model generalizability, and clinical implementation, with each dimension influencing and being influenced by the others across policy, data, AI, and practice. The themes raised during the second stakeholder session aligned with the coding scheme

XSL•FO
RenderX

developed from the scoping review, reinforcing the relevance of the literature to stakeholder perspectives. The workshops played a crucial role in grounding the analysis in real-world experience, ensuring its practical relevance within the LEAPfROG project.

To illustrate how thematic emphasis shifted across the different study components, Table 1 presents the staged development of these themes across the scoping review and stakeholder workshops.

Table . Shifts in thematic emphasis across study components of the staged LEAPfROG (Leveraging Real-World Data to Optimize Pharmacotherapy Outcomes in Multimorbid Patients Using Machine Learning and Knowledge Representation Methods) design.

Key themes / stage	Workshop 1 (EHR[a] data reuse)	Scoping review (EHR–AI[b] Integration)	Workshop 2 (AI development and implementation)
1. Data privacy, transparency, and consent	• Patient autonomy and pseudonymization challenges • Clear consent for secondary use of EHR data	• Privacy risks amplified by data linkage and free-text inputs; persistent reidentification vulnerabilities • Tensions between ethically robust consent and notions of a civic duty to share EHR data	• Fostering trust through transparency across AI model development, explainability, and outputs
2. Public trust and regulatory challenges	• Institutional fragmentation and suboptimal national EHR infrastructure • Inconsistent policy and legal interpretations among stakeholders (including data custodians and researchers)	• Siloed governance and fragmented oversight • Regulatory uncertainty limiting data sharing and accountability	• Stakeholder-led, value-driven governance • Shared vision on responsible EHR data reuse and multistakeholder collaboration
3. Fair representation and model generalizability	• Data heterogeneity and variability in EHR documentation practices • Digital and social inequities shaping clinicians' use of EHRs and patients' access to their health data	• Missing or biased data reinforcing inequities • Lack of demographic diversity and transparency in reporting	• Variation in clinical practices (eg, workflow diversity) shaping EHR data patterns and subsequent model outputs and interpretation • Need to recontextualize EHR-derived data across hospital departments and care settings
4. AI integration in clinical practice	• Documentation burden and workflow diversity in CKD[c]/DAKI[d] care • Impact on EHR data reliability and clinician workload	• Model opacity, "clinical tropism" • Risk of overreliance and clinician deskilling • Ethical need for interpretability and accountability	• Timing and interpretability of AI model outputs • Risk of overly general or narrow outputs with limited clinical relevance • Model could enhance clinicians' understanding of drug side effects and support shared decision-making

[a]EHR: electronic health record.

[b]AI: artificial intelligence.

[c]CKD: chronic kidney disease.

[d]DAKI: drug-induced acute kidney injury.

We connected insights from the literature and stakeholder discussions to situate the key themes within the Dutch health care landscape and the specific context of the LEAPfROG project. Each of the following sections elaborates on one of the 4 key themes and concludes with a reflection on how that theme was grounded and contextualized through stakeholder discussions.

Data Privacy, Transparency, and Consent

Privacy Risks and Security Challenges

The reviewed literature highlights growing privacy concerns around the reuse of EHR data for AI research [22-34]. EHR data originate from everyday clinical care and often contain sensitive personal and medical information recorded without explicit patient consent, raising concerns about misuse and breaches of confidentiality [30]. In contrast, data from controlled research settings such as randomized trials follow predefined consent procedures and strict regulatory oversight.

Safeguarding privacy in this context requires clear data access agreements, secure systems for storage and transmission, and transparent mechanisms for patient control and consent, in line with legal and ethical standards. The risks described in the literature include security breaches, such as hacking or malware attacks, that could lead to data exploitation, intentional discrimination, or identity theft [27,31,35]. Additionally, combining multiple data sources, such as EHRs and disease

XSL•FO
RenderX

registries, amplifies privacy risks through greater exposure to breaches, inconsistencies, and loss of data integrity [23,27,28,30,34,36]. Advances in AI have further intensified the risk of misuse and harm by enabling sophisticated reidentification, tracking, and profiling techniques that threaten patient confidentiality, including methods capable of inferring sensitive information from nonsensitive data. These developments heighten the risk of misuse and harm [28,35]. Even privacy-preserving approaches remain vulnerable to inference attacks that can reveal attributes of training data, posing ongoing risks to confidentiality [35].

Moreover, free text fields in EHRs often include both medical and personal information. Persisting difficulties in deidentifying this information increase the potential for reidentification and unintended disclosure when used for research purposes, possibly even beyond the individual patient [27]. Ford et al [27] discuss the challenge of balancing automated and human-led deidentification of free text, noting that participants in citizen jury discussions expressed concerns that automated methods might leave identifiers, while human-led processes, on the other hand, could introduce bias or error due to inconsistency or workload. Others emphasize that the use of AI methods to extract information from unstructured EHR text often lacks standardized guidelines and the level of scrutiny applied to other medical interventions, further exacerbating privacy risks [33].

Informed Consent: Is There a Civic Duty to Share Data for AI Health Research?

Most included studies identify challenges related to the need for explicit patient consent in reusing EHR data for AI-driven research [23,25,27-31,33-35,37,38]. Patients are often unaware of how their data may be used over time, raising questions about the validity of prior or implicit consent given for clinical care or research. Evolving AI models may introduce new applications that extend beyond the original scope of consent, repurposing data in ways that patients could not have anticipated [31,38]. These concerns intensify when multiple data sources are linked, making it difficult to ensure meaningful consent at each stage of data integration [22,23,38,39]. Without clear processes for obtaining and maintaining consent, AI-driven research risks undermining patient autonomy and trust in health data governance, compromising its moral legitimacy.

Among the included studies, Müller critically examines the view that citizens have a moral duty to share their health data for medical AI development. He argues against this by outlining key contrasts between common assumptions underlying such a duty and the ethical realities of data sharing [28]. First, the “rule to rescue,” which asserts a duty to help others in immediate danger, does not apply here, as data sharing lacks the direct, person-to-person moral obligation implied in such situations. Second, the “low risks, high benefits” argument claims that societal gains justify data sharing. He argues that this view overlooks how risks such as privacy breaches and erosion of trust in health care are unequally distributed and more immediate than potential benefits. Third, the “property rights argument” holds that the involvement of public institutions in generating health data justifies limiting individual ownership. Müller disputes this, emphasizing that health data are inherently personal and sensitive, and that treating them as a public commodity undermines autonomy and confidentiality.

Instead of a universal moral duty, Müller proposes a civic responsibility approach grounded in transparency, value alignment, and voluntary participation, emphasizing consent mechanisms that allow individuals to manage how their data are shared. In line with this perspective, other authors argue that even when data reuse does not legally require consent, researchers remain ethically obligated to communicate clearly how data are used, how AI models operate, and how decisions are reached, as legal compliance does not substitute for transparency or patient and public engagement [33].

Stakeholder Perspectives: Data Privacy, Transparency, and Consent

In line with the literature, the stakeholder group discussed potential risks to patient privacy in the reuse of EHR data for research. They identified data exchange and EHR systems linkages as particularly vulnerable points, along with the risk of unauthorized data use by third parties such as health care insurers and private companies. To address these concerns, stakeholders proposed several measures, including the use of virtual research data environments that strictly comply with data protection regulations, ensuring that research data are pseudonymized or anonymized to reduce the risk of misuse or breaches, and enabling data providers to audit researchers’ actions on the data.

Other suggestions involved giving patients greater control over how their EHR data is reused, including the ability to decide whether, and which parts of, their records should be used for research. While this approach could enhance patient autonomy, concerns were raised about the risk of overburdening patients with repeated consent requests, especially when the research offers no clear or direct benefit at the individual level and does not immediately reflect patient priorities.

Public Trust and Regulatory Challenges

Trust and transparency

The majority of the studies highlight trust as crucial in the relationship between various publics and those who collect, control, access, and use EHR data for AI research [23-25,27,29,31-36,38,40-42]. Many studies identify low public awareness as a key factor undermining trust in these practices [24,27,32,36]. Patients are often unaware that their data are shared across organizations within integrated systems or for what specific purposes [31]. Low public trust in institutions handling health data may reduce participation in AI research, ultimately introducing self-selection bias and weakening model quality and representativeness [28,31,34]. Given these concerns, many studies underscore the ethical importance of transparency, not only as a means to foster trustworthiness but also for ensuring legitimate and accountable data governance [22,24,25,27-34,38,40,41].

Regulatory fragmentation and legal uncertainty

Several studies identify regulatory and legal uncertainties as major obstacles to the adoption of health AI tools [30,41,42]. A fragmented legal landscape, marked by inconsistencies across

XSL•FO
RenderX

institutions and jurisdictions, creates compliance and accountability challenges [29,31,33]. In Germany, for example, data-sharing initiatives are complicated by regulatory fragmentation and varying requirements across sectors [29]. Siloed regulation, in which separate and uncoordinated frameworks govern different types of health data and medical technologies, fragments oversight, leading to regulatory gaps (areas not clearly covered by any authority), overlaps (conflicting or duplicated requirements), and broader compliance hurdles that hinder implementation [33,38]. Regulatory uncertainty often deters data custodians—responsible for data storage, integrity, and authorized access—from sharing EHR data beyond clinical settings. This reflects a broader tension between oversight predominantly aimed at safeguarding privacy and fostering the flexibility required for AI innovation. For instance, strict interpretations of data protection laws, such as the General Data Protection Regulation, can limit researchers' access to patient data, influencing the trajectory of developments in the field [30,38].

Stakeholder Workshops: Public Trust and Regulatory Challenges

In the Netherlands, EHR data exchange remains limited due to the fragmented health care system, where institutions rely on diverse and often incompatible standards. For example, hospitals may use different clinical coding systems or inconsistent data formats for diagnoses, making integration across systems technically and semantically challenging. This challenge is further compounded by a decentralized governance framework that grants considerable institutional autonomy, resulting in varied requirements and interpretations across institutional policies. In line with the literature, stakeholders noted that unclear and inconsistent interpretations of laws and regulations create barriers to collaboration within a complex network of actors, including hospitals, health insurers, data vendors, EHR system providers, and research funders, each imposing specific conditions for data availability and reuse.

A key suggestion that emerged during the workshops was to develop a shared vision within collaborative networks on the "why" of reusing EHR data for research. This could be achieved, for example, through consensus workshops involving patients, health care providers, researchers, data managers, and data protection officers, with attention to values such as trustworthiness, solidarity, quality of care, and privacy. Participants also emphasized the importance of establishing shared agreements within these networks on which data should be recorded in the EHR, using predefined terminologies, and defining how and where these should be implemented.

Fair Representation and Model Generalizability

Health Inequities Reflected in the Data

Most included studies warn that missing or incomplete patient data can compromise model performance and perpetuate existing health care inequities, disproportionately affecting marginalized groups [22-25,28,30,31,33,37,39-46]. These inequities often stem from social determinants of health, such as financial barriers, limited digital and health literacy, and structural marginalization. Failing to adequately consider these factors can result in flawed models that underestimate disease burden for certain groups and undermine the equitable distribution of benefits [25,36,37,39,42-44]. Several studies further highlight the need to account for the historical context of research on race and ethnicity, ancestry, and sex and gender minorities, particularly in the United States, where past exploitation contributes to ongoing distrust in health care systems [40,42].

Institutional and Structural Variability in Data and Its Impact on AI Data Quality

Several studies highlight how institution-specific practices, norms, and patient populations shape the composition of EHR data [22,24,25,28-33,35,37,40,42-46]. Factors such as patient proximity to clinics, visit frequency, insurance coverage, and differences in clinical protocols and diagnostic resources influence data representation [22,24,25]. For example, some institutions may request diagnostic tests earlier [30] or use activity-based financing, leading to an overrepresentation of codes tied to higher reimbursement, reflecting billing priorities rather than actual clinical complexity [22,25]. Sampling bias can occur when chronically ill patients, who interact more frequently with the health care system, are overrepresented, skewing patterns that do not reflect the broader population. Within institutions, departmental priorities and documentation practices may differ, contributing to inconsistencies in data capture [25]. Increasing reliance on data-intensive health care can reduce direct communication between medical staff, heightening the risk of incomplete or selective documentation [24].

While standardizing EHR data through structured formats and predefined categories may reduce inconsistencies, it also risks removing essential clinical context [27,34,36,44]. Excluding physicians' notes and clinical judgments, for example, may lead to the loss of valuable information such as symptom descriptions, diagnostic reasoning, and the context preceding symptoms. This increases the likelihood of misinterpretation by researchers and AI models, potentially resulting in biased or unsafe decisions [27,30,36,44,46]. Yet, as Knevel and Liao [30] note, some gaps in EHR data can hold clinical meaning. For instance, the absence of specific test results may suggest that a particular diagnosis was deemed unlikely. Interpreting such nuances requires strong domain knowledge and clinician involvement to ensure meaningful analysis.

Variability in EHR systems across institutions and countries continues to challenge the development of generalizable AI models [30]. The absence of centralized or standardized data-sharing mechanisms, such as those based on findability, accessibility, interoperability, and reusability principles, exacerbates this problem by hindering interoperability and seamless data integration across settings [23,33,34,36,38]. To address structural and institutional fragmentation, federated learning has been proposed as a promising approach aimed at preserving patient privacy while enabling AI models to be trained across institutions without sharing raw data. Such methods seek to facilitate collaboration across decentralized datasets while maintaining institutional autonomy. Related privacy-preserving techniques include differential privacy and homomorphic encryption, which are designed to enhance data

XSL•FO
RenderX

protection but remain largely experimental and face challenges of scalability, performance, and validation in clinical settings [35].

Challenges in AI Model Reliability and Potential Harm

Several studies raise concerns about harm from unreliable or biased AI systems, particularly when embedded in clinical decision-making. Deliberate and unintentional choices in AI development can introduce algorithmic bias, reinforcing EHR data limitations and distorting model outputs [22,24,25,27,28,30-33,35,37,42,43,45]. Examples include data integrity issues from reusing and merging databases, such as data duplication [24,35], model design choices reflecting provider bias [32], and preselection of training data based on physician or developer preferences [28,46].

Technical flaws throughout model development can also result in harmful clinical outcomes, even when using high-quality data [33,37,43]. For example, models may overfit datasets that do not reflect relevant clinical questions, undermining generalizability and reliability [33,37,43]. Optimistic predictions can lead to futile or even harmful interventions, while overly pessimistic ones may cause unnecessary withholding of care [32]. In such cases, the ethical imperative to minimize harm may justify simpler models that generalize more safely, even if less accurate. Some studies argue that the appropriate level of model explainability should be calibrated according to the clinical risks and the extent to which domain experts can interpret the model's output [24,33,37,46]. Several studies also highlight the lack of standardized frameworks or tools to assess and evaluate AI's real-world clinical impact, particularly regarding patient safety, value alignment, and health equity, thereby limiting consistent assessment of ethical implications and potential harm [25,26,29-34,38,40,43-45].

Barriers to Inclusive AI Development

Several studies identify deeper systemic and infrastructural barriers that undermine inclusive and equitable AI. Adapting AI to local environments requires significant investment in data integration, testing, infrastructure, and clinical expertise [22,30]. However, many health organizations lack the secure, scalable infrastructure needed for effective AI deployment, a gap often overlooked in policies that emphasize AI's benefits while downplaying resource demands [22]. These barriers also raise questions of equity in data governance, including whether and how data contributors should be compensated or share in AI's benefits [31].

Another persistent challenge is the unclear definition of target populations and the inconsistent representation of demographic groups in AI development [22,25,28,41,45]. One study found, for example, inconsistent demographic reporting across 164 EHR-based AI studies, undermining model representativeness and reproducibility [45]. Limited data sharing, closed-source models, and poor interoperability hinder validation across diverse clinical settings, while low digital adoption and data vendor monopolies further restrict datasets to less representative populations [25,28,29,32-34,36,38,41,42,45].

Paulus and Kent [42] argue that fairness in algorithmic decision-making is context-dependent, requiring value judgments and stakeholder consensus. No universal definition exists, and ethical concerns over protected attributes such as race persist. Others similarly highlight concerns about the values and priorities embedded in AI models, which may implicitly reflect biases or unexamined assumptions [44]. This risk intensifies when health data is used for generating profit, for example, potentially placing financial incentives above equitable patient care and responsible data stewardship [30,31,33,34]. In response, many studies advocate for sustained stakeholder engagement throughout AI development efforts, including participatory co-design and mixed methods approaches that integrate qualitative and quantitative insights rather than relying on externally imposed solutions [22-25,27,28,32,33,37,40-44,46].

Stakeholder Workshops: Fair Representation and Model Generalizability

Stakeholder discussions highlighted the complex realities of clinical practice in the context of CKD and DAKI, which are inherently mirrored in the data. For instance, when treating sepsis with antibiotics such as vancomycin, known to potentially cause acute kidney injury (AKI), clinicians may consciously accept the risk of AKI as a necessary trade-off to, in some cases, save the patient's life. Additionally, when a side effect is common and widely acknowledged, physicians may not document or provide less detailed documentation, focusing instead on more pressing clinical problems. Moreover, these considerations can also vary across medical specialties, with clinical practices such as prescribing practices differing, for example, between internal medicine and cardiology.

In short, understanding a potential causal relationship between a drug and AKI provides only part of the story. During the workshops, it was emphasized that gaining a clear understanding of how variability from different clinical settings is represented in the EHR data is crucial, as is considering how such complexity could be effectively incorporated into an AI-driven decision-support model for retrospective DAKI diagnosis. At the same time, consistent with Knevel and Liao [30], it was noted that research using EHR data can offer valuable insights into clinical routines and decision-making patterns, such as differences in prescribing practices across specialties. However, uncovering these insights is highly resource-intensive and time-consuming, as it requires not only technical expertise but also deep contextual knowledge and sustained engagement with clinical practice.

Responsible AI Integration in Clinical Practice

Administrative Workload

While efforts to structure and standardize EHR data are central to enabling AI-driven research, several studies emphasize the unintended increase in administrative burden for health care providers [24,33,35-38,40,41,44]. Processes such as careful data assessment and logging, essential for ensuring data quality, reducing fragmentation, and addressing issues like concept drift in longitudinal datasets, are time-consuming and resource-intensive [33,37,41]. These demands raise ethical concerns related to clinician well-being, sustainable work practices, and their impact on patient care [24,29,36,41].

XSL•FO
RenderX

Furthermore, several studies note that professionals' earlier experiences with EHR implementations often failed to deliver the anticipated quality improvements despite significant increases in workload associated with data entry and system management, contributing to skepticism toward data-driven AI solutions [24,29,41]. Finally, many health care settings still lack fully digitized, structured records, complicating the integration of AI into clinical workflows [41]. Together, these challenges highlight a tension between the technical demands of fit-for-purpose data in AI-based research and the practical realities of clinical work, underscoring the need for implementation strategies that balance data quality with provider sustainability.

Clinical Tropism

A major barrier to AI adoption in health care is the limited transparency and interpretability of complex, "black-box," models [22,24-28,30-34,36-41,43-46]. When AI-driven decisions lack clear explanations, clinicians and patients struggle to trust them, undermining ethical requirements for transparency, informed decision-making, and accountability in clinical care. Without insight into how these models function, users cannot detect or correct biases, increasing the risk of erroneous or harmful outcomes [22,39,41]. Distinguishing causal from correlative relationships is essential in medical decision-making to avoid ineffective or harmful interventions [28,39]. These risks are further amplified by the fact that harmful AI-related errors can occur at scale, exceeding the impact of individual provider errors [22].

As discussed earlier, many studies highlight challenges of representativeness and model generalizability, emphasizing how incomplete or biased datasets can perpetuate existing inequalities. Beyond these population-level concerns, several studies draw attention to how AI systems mirror the clinical contexts in which they are developed. Alami et al [22] describe this as clinical tropism, referring to the tendency of AI systems to reproduce the specific practices, routines, and priorities of the settings that generated their training data. They argue that models trained on localized data may reflect institutional protocols, workflows, or device infrastructures that are not easily transferable to other contexts. Over time, such systems risk becoming repositories of established clinical practices, developing what Alami et al [22] term a "clinical mind," rather than generating new insights. This dynamic may lead professionals to place disproportionate trust in AI outputs that merely echo existing patterns of care. Furthermore, because clinical practice continually evolves through new treatments and protocol updates, EHRs are inherently temporal, making input data prone to diverge from the contexts represented in their underlying sources [22].

Overreliance and the Risk of Deskilling

Several studies highlight the risk of deskilling among medical professionals as AI adoption increases, potentially weakening clinical judgment, autonomy, and the ability to make independent decisions [22,25,26,29-32,37-39,41,43]. Overreliance on AI may also reduce adaptability across clinical contexts and limit clinicians' capacity to account for patient preferences and values, raising concerns about the long-term impact on clinical competency and patient care [22,25,26,32,37,43].

Empirical evidence suggests that caregivers (including nurses, midwives, clinicians, and pharmacists) are generally less likely to question algorithm-driven diagnostic results, increasing the error risk and further eroding clinical judgment over time [22,25,28,37]. Conversely, when AI-generated results deviate considerably from a clinician's assessment, there is a risk that the model's recommendations will be dismissed outright, potentially limiting AI's practical value in clinical settings [22,43]. These dynamics underscore the need for careful integration of AI into clinical workflows, with sustained attention to its effects on professional roles, decision-making, and oversight.

Stakeholder Workshops: Responsible AI Integration in Clinical Practice

In the complex care setting of CKD and DAKI, precise documentation of medication use and adverse drug events is essential for generating reliable data to support AI-driven research. However, DAKI is difficult to diagnose because of its multifactorial causes, subtle onset, and patient-specific variability, often involving combinations of dehydration, sepsis, and drug toxicity. During stakeholder workshops, participants stressed that documenting this complexity is time-consuming. One general practitioner noted that recording side effects alone can take up to 30 minutes, a significant burden given current time pressures in Dutch health care. Although general practitioners are legally required to provide digital access to medical records upon request, they are not obliged to use an EHR system. One kidney patient shared that her general practitioner still prefers pen and paper, relying on digital systems only when necessary, illustrating how variability in documentation practices continues to limit the availability of structured data for AI.

Participants saw potential for the model to have a learning effect, improving clinicians' understanding of drug side effects over time. This could enhance care quality and, if communicated clearly, support shared decision-making. However, they stressed that model outputs must always be evaluated alongside clinical expertise to ensure safe and appropriate care. To improve clinical relevance, they suggested linking outputs to relevant guidelines and scientific literature to help physicians interpret findings in context and assess their significance for individual patients. At the same time, they cautioned that the time required to interpret outputs might increase or shift clinical workload rather than reduce it.

Stakeholders raised concerns consistent with the concept of "clinical tropism" of Alami et al [22], which describes how AI models may reproduce the narrow or context-specific knowledge embedded in the data on which they are trained. They noted that this tendency could be reinforced by the steering effect of clinical guidelines and institutional protocols that shape both data collection and model performance. Participants further noted that publication bias in the studies informing model development can compound these effects. When only positive or statistically significant results are published, the evidence base fails to reflect the full range of clinical outcomes, leading

XSL•FO
RenderX

models to learn a distorted view of clinical reality, thereby reinforcing rather than mitigating clinical tropism. Together, publication bias and the standardizing influence of clinical guidelines narrow the evidentiary foundation on which AI systems are built, privileging typical or well-studied cases while obscuring clinical variability and local practice. Participants also cautioned that integrating diverse data sources may produce overly general outputs with limited clinical relevance, lacking actionable value for clinicians and therefore likely to be disregarded.

Finally, participants stressed that assessing the causes of AKI requires attention to factors beyond medications and careful interpretation of predicted probabilities within individual patient contexts. For instance, a 61% probability that an antibiotic caused AKI might be clinically insignificant in one case but highly relevant in another, depending on the patient's circumstances. Participants emphasized that the model should be developed and implemented primarily as a decision-support tool to help clinicians assess whether a drug likely contributed to AKI.

Discussion

Principal Findings

This study used a mixed methods approach to explore the ethical considerations of using EHR data for AI-driven health research and innovation, with the LEAPfROG project serving as a guiding case. We combined a scoping literature review, a systematic search, and 2 stakeholder engagement workshops, all informed by the GEA [17].

Previous studies have examined the reuse of EHRs and the ethical implications of AI development separately. Research on data reuse has addressed governance, consent, and public trust, showing how challenges of transparency, fairness, and oversight are amplified by the data's variable quality, uncertain provenance, and fragmented governance [12-14]. Studies on medical AI have focused on fairness, transparency, and accountability, mapping the ethical and epistemic risks that arise within algorithmic systems [15,16]. Our study brings these strands together by examining how recurring ethical challenges emerge and take shape at the intersection of EHR data and AI development, while also situating these issues within a real-world clinical and institutional setting (the LEAPfROG project).

By integrating stakeholder perspectives with existing ethical analyses, this study contributes an empirically grounded account of how recurring ethical challenges emerge in practice when AI systems are developed using EHR data. In this context, issues of consent, privacy, and governance become more complex as EHR data is repurposed and leveraged using AI, while the opacity and scale of AI systems introduce new risks of decontextualization and fragmented accountability across policy and oversight domains.

Building on this analysis, we advocate for an approach to data governance and AI development that is centered on and led by stakeholders. We stress the importance of involving all relevant parties not only in evaluating benefits and risks, but also in shaping the goals and direction of AI innovation. Accordingly, we structure the following discussion as follows: first, we present a set of practical recommendations derived from the literature and stakeholder workshops, intended to guide future reflection, engagement, and implementation. We then explore the concept of data work and examine the ethical risks associated with removing data from its original context and reusing it for AI development. Finally, we review current practices in stakeholder engagement and highlight the need for more inclusive and participatory forms of governance in future AI initiatives.

Practical Recommendations

In addition to the central themes, both the scoping review and stakeholder workshops yielded practical recommendations for addressing ethical concerns related to AI and the use of EHRs (Table 2). Although these action points were not the primary focus of this study, they offer opportunities and potential entry points for addressing the broader challenges we aimed to explore. These insights reflect the perspectives and priorities identified in the sources consulted and do not necessarily represent the position of the author group. As the findings draw on international literature, some recommendations, such as patient representation in ethics committees, may already be established in contexts like the Netherlands. Nevertheless, they offer a useful supplement for future reflection, engagement, and implementation.

XSL•FO
RenderX

Table . Key recommendations from literature and stakeholder workshops.

Key themes and recommendations	Main actors	Source (key articles)
Data privacy, transparency, and consent		
Strengthen AI[a] and data literacy among patients, clinicians, and institutional review board (IRB) members, emphasizing how EHR[b] quality and AI training may affect privacy, consent, and data integrity.	Board of Directors of health care organizations	[38,41]+ workshops
Implement dynamic consent mechanisms that reflect the iterative reuse of EHR data in AI development and ensure institutional governance structures can accommodate these updates.	Board of Directors of health care organizations	[27,28,31]
Integrate consent management tools into EHR interfaces so that patients can review and modify how their data is used in research and model training, with clinical staff facilitating informed communication.	Board of Directors of health care organizations and EHR system vendors	[25,31]+ workshops
Require transparent documentation of AI provenance, including funding sources, proprietary model components, and training data, to prevent hidden conflicts of interest and foster public trust.	Researchers involved in AI projects like the LEAPfROG[c] project	[23,25,29]
Public trust and regulatory challenges		
Adopt standardized, FAIR-aligned[d] data practices and establish cross-institutional agreements that clarify ownership, accountability, and data flow between hospitals and general practices.	National research infrastructure organizations like Health-RI in the Netherlands	[33,34,36,40]+ workshops
Establish joint oversight frameworks that include clinicians, data stewards, developers, and regulators to coordinate decision-making on data use, ownership, and third-party involvement, reducing fragmented accountability.	National research infrastructure organizations like Health-RI in the Netherlands	[23,25,29,31]+ workshops
Embed patient representatives (ideally trained in AI ethics and data governance) in ethics committees and regulatory boards to strengthen inclusiveness and transparency in oversight processes.	National research infrastructure organizations like Health-RI in the Netherlands and Board of Directors of health care organizations	[25,38,41].
Fair representation and model generalizability		
Validate AI models in real clinical environments (for example, CKD[e] and AKI[f] contexts) and across diverse patient subgroups to ensure representativeness and clinical relevance.	Researchers involved in AI projects like the LEAPfROG project	[37,42,43,46]
Conduct fairness audits to explicitly assess whether model outputs are biased by demographic differences or by specific clinical routines and documentation practices that vary across institutions.	Researchers involved in AI projects like the LEAPfROG project	[32,40,42,43]
Adopt mixed methods evaluation combining quantitative metrics with qualitative insights to understand how AI affects workflows, professional judgment, and patient experience.	Researchers involved in AI projects like the LEAPfROG project	[30,40,42,43]+ workshop 2
Implement continuous postdeployment monitoring: reassess clinical impact on patient outcomes over time to ensure alignment with evolving clinical practice and technologies.	Board of Directors of health care organizations	[25,38,46]+ workshop 2

XSL•FO
RenderX

Key themes and recommendations	Main actors	Source (key articles)
Appraise both clinical and economic outcomes through frameworks that evaluate how AI contributes to care quality, efficiency, and sustainability of health care delivery.	Researchers involved in AI projects like the LEAPfROG project	[22,39,46]
Responsible AI integration in clinical practice		
Align AI development with clinical workflows, prioritizing usability and practical relevance over purely technical performance metrics.	Researchers involved in AI projects like the LEAPfROG project	[22,24]+ workshops
Ensure interpretability supports clinical judgment by designing models that reflect clinical complexity, communicate uncertainty, and complement human expertise.	Researchers involved in AI projects like the LEAPfROG project	[22,26,37]+ workshop 2
Develop learning-oriented AI systems that enhance, rather than erode, professional expertise, and support reflective clinical practice.	Researchers involved in AI projects like the LEAPfROG project	Workshop 2
Design EHR/AI tools collaboratively with stakeholders throughout the data and model lifecycle to clarify purposes, risks, and benefits, aligning these with shared values and patient expectations.	Researchers involved in AI projects like the LEAPfROG project and Board of Directors of health care organizations	[25,33,38,41]+ workshop 1

[a]AI: artificial intelligence.

[b]EHR: electronic health record.

[c]LEAPfROG: Leveraging Real-World Data to Optimize Pharmacotherapy Outcomes in Multimorbid Patients Using Machine Learning and Knowledge Representation Methods.

[d]FAIR: findability, accessibility, interoperability, and reusability.

[e]CKD: chronic kidney disease.

[f]AKI: acute kidney injury.

Making Data Work

While much of the public and academic debate on AI in health care focuses on model explainability or algorithmic opacity, these discussions often overlook a more foundational issue: the nature of the data on which AI depends. Earlier studies on health data reuse have shown [12-14] that questions of trust, consent, and governance already hinge on how data is produced, managed, and interpreted. Our findings suggest that without a deeper understanding of how health data is actually generated, structured, and used in practice, efforts to build fair, accurate, and trustworthy AI systems will remain incomplete. Data is rarely a neutral input, as it is deeply shaped by clinical routines, institutional diversity, and social context. Documentation practices, system design, and socio-ethical expectations determine what gets recorded and how it is organized. Recognizing this “data work” is not merely a technical concern but an ethical one, since it determines whose experiences are recorded and how decisions are justified.

Our findings reveal the complexity and effort involved in making EHR data fit for AI-driven research. Both the reviewed literature and stakeholder insights point to inconsistencies in EHR documentation, variability in clinical workflows, and the labor-intensive nature of data preparation and standardization. Such challenges underscore that data is produced through ongoing, situated work rather than passively collected. This aligns with the literature review on data-related activities by Bertelsen et al [47], or “data work,” which emphasizes its multifaceted and situated nature. They group data work into 3 interrelated categories: data collection (eg, capturing, discovering, requesting, and self-tracking), data production (eg, coding, entry, and digitization), and data use and sharing (eg, providing, exchanging, and disseminating).

As data move through different stages and are reused for AI training, they often lose the clinical and institutional contexts that once gave them meaning. This process, referred to here as decontextualization, involves detaching data from the social and professional environments in which they were produced, allowing them to be reinterpreted, or even misinterpreted. Hoeyer [48] describes this as a separation between the “data work of production” and the “data work of analysis,” where separation from the original context renders data fragile and open to misinterpretation. In AI development, this dynamic has ethical as well as epistemic consequences: models may treat context-dependent judgments or incomplete entries as objective facts, obscuring the human and institutional processes that shaped them. This dynamic is also reflected in the concept of “broken data,” which highlights the ongoing, often hidden repair and improvisation required to keep data fit for purpose [49], as well as in studies showing how context loss invites interpretive uncertainty, with trust, visual design, and perceived credibility shaping how data is understood [50].

When AI systems trained on such data are applied in clinical settings, the same loss of context can reappear, shaping how models interact with the environments that produced their data and often reinforcing the very patterns they were meant to improve. Stakeholders noted that AI systems may mirror existing

XSL•FO
RenderX

prescribing and documentation practices, reproducing the same routines they aim to change. This recursive dynamic reflects what emerged from the literature as clinical tropism, where models embed themselves in the patterns most strongly represented in their training data and in the settings where those patterns persist. When that influence begins to shape the outcomes it predicts, it can result in a self-fulfilling prophecy [51]. In such cases, models appear accurate precisely because their use helps make their predictions true. Clinical tropism creates the conditions, while the self-fulfilling prophecy describes the result. Together, they show how decontextualization can evolve from an issue within data to one that reshapes clinical practice, as models begin to reproduce the very patterns they were designed to improve. These concerns are mirrored in debates on explainable AI, where current approaches have been criticized for creating a false sense of understanding and trust, offering surface descriptions of model behavior without clarifying whether decisions are reasonable or justified [52]. Rather than mitigating bias, such transparency can obscure how context loss shapes both model development and evaluation.

Extending the discussion of decontextualization, Hoeyer [48] draws attention to a deeper issue often overlooked in discussions of data work: the ways in which data are interpreted and mobilized in pursuit of narratives that are often implicit and conflicting. He argues that datafication fragments patient information into pieces that can be recombined to serve diverse aims, such as clinical care, research, governance, or performance monitoring. In this process, the patient may be obscured, and data detached from the context of its production. Clinicians may document with specific knowledge and intent, while analysts or policymakers may act on institutional logics shaped by different priorities. These tensions are not resolved through technical solutions alone but often reappear, in new forms, within the very tools developed to address them. For instance, even privacy-preserving approaches such as federated learning, which are designed to mitigate data sharing risks, can inadvertently deepen decontextualization by abstracting data use from the institutional and clinical settings in which they originate [35,53]. By enabling model training across institutions without sharing raw data, these methods aim to enhance privacy but may nevertheless reproduce institutional differences and context loss, introducing additional ethical concerns that extend beyond the scope of this paper but are highlighted elsewhere [53].

As is evident in the LEAPfROG project, AI initiatives involving EHR data span a diverse network of stakeholders, including health care providers, data custodians, EHR vendors, patients, and institutions, whose interests may not align. These misalignments are not incidental but intrinsic to how data is leveraged, representing structural and arguably unavoidable realities that must be acknowledged. Developing responsible AI, therefore, requires collaborations that explicitly recognize these dynamics and openly address how stakeholders themselves navigate and negotiate data use in practice, while articulating the competing aims, values, and uses of data within their own processes [47]. Doing so requires more than technical expertise: it depends on collaboration across domains to understand how clinical, organizational, and social contexts shape data and influence AI training and deployment. Together, these insights underscore that making data work is as much a social and ethical practice as a technical one.

Beyond Compliance: Meaningful Stakeholder Involvement

In current debates on health data governance, the public good is often invoked as a guiding principle, with values such as solidarity promoted as ethical foundations [54]. Prainsack et al [55], for instance, frame data solidarity as a means to promote justice, prevent harm, and ensure collective benefit. However, critics argue that such values, while well-intentioned, are often applied in top-down, unreflective ways that limit their ethical force [54,56,57] and risk overlooking deeper power asymmetries, such as the lack of patient representation in governance and decision-making structures. Rather than rejecting these principles, they have called for alternative models grounded in mutual aid and collective action, fostering more equitable relationships between data providers, users, and affected communities.

Taken together, these critiques reveal a broader structural issue: legal and policy alignment, while necessary, is not sufficient for the ethical use of health data or responsible AI development. As our results show, many studies point to the need for participatory, transparent governance that goes beyond compliance or rhetorical appeals to ethics. Concerns about fragmented oversight, unclear consent procedures, and opaque decision-making illustrate how legitimacy is undermined when stakeholders are not meaningfully involved. Without sustained, inclusive engagement, especially with those most directly affected, governance risks remain superficial and biased toward institutional priorities.

At a practical level, these structural shortcomings are reinforced by persistent power asymmetries in engagement processes, which risk prioritizing influential groups such as government officials, large nongovernmental organizations, or scientists [58]. These “elite” actors, with greater access to time, resources, and expertise, tend to set the terms of participation, sidelining less-resourced groups and limiting patient involvement. This is especially problematic in the context of CKD, where patients face complex treatments, uncertainty, and symptoms that can limit participation in research [59]. As a result, they are often underrepresented in studies that fail to reflect patient or caregiver priorities. Critical participation frameworks emphasize that such unequal conditions make engagement prone to tokenism, particularly when structural inequities like corporate influence or health care disparities remain unaddressed [59,60].

These dynamics become even more pronounced in the context of AI development. Here, technological optimism often frames innovation as inherently beneficial to patient care [61], which can obscure deeper ethical and social tensions. By emphasizing efficiency, accuracy, and progress, this framing tends to narrow debate to technical or policy issues, while sidelining questions of power and lived experience. Consequently, participation in AI development tends to be tightly controlled, serving institutional efficiency or public acceptance rather than fostering meaningful, value-driven dialogue. In this way, engagement

XSL•FO
RenderX

risks becoming instrumental, legitimizing AI initiatives rather than engaging with lived experience or long-term ethical and practical implications.

In light of these challenges, we were particularly mindful of how stakeholder engagement was approached in the LEAPfROG workshops. Recognizing the potential influence of technological optimism, we sought to create space for more critical, grounded conversations about the role of AI in health care. By grounding ethical reflections in concrete use cases and fostering open-ended, collaborative discussions, we aimed to avoid tokenism and reduce the influence of more resourced stakeholders. Our mixed methods approach, which included a literature review alongside stakeholder workshops, helped broaden the scope to address systemic issues such as health care inequities and the resource-intensive nature of AI development.

As an alternative to top-down regulation and rhetorical appeals to ethics, more participatory and inclusive governance models may offer a promising path forward. One such approach involves shifting toward stakeholder-led governance within consortia, supported by coproduced guidelines or codes of conduct. A key suggestion from our workshops was to develop a shared vision around the reuse of EHR data, including its purpose, risks, and benefits, through collaborative processes involving patients, clinicians, researchers, data managers, and data protection officers. Such efforts could also include agreements on what data should be recorded in EHRs and how. Rather than treating participation as a means to secure legitimacy for predetermined goals, these practices aim to embed values like trustworthiness, solidarity, and privacy into the infrastructure of data governance itself. By continuing our collaboration with the Dutch Kidney Association (NVN), we hope to support a model of engagement that centers patient perspectives, addresses power imbalances, and ultimately aligns data-driven innovation with the needs of those most affected.

Limitations

While time constraints, limited participant availability, and last-minute cancellations are common challenges in stakeholder workshops, a more critical limitation was ensuring balanced representation among stakeholders. Although the workshops were conducted in Dutch to promote inclusivity across diverse education levels and professional backgrounds, this choice inadvertently excluded non-Dutch-speaking participants, including some within the LEAPfROG consortium. Similarly, the scoping review included only English-language publications, potentially introducing language bias and omitting relevant studies in other languages. However, only 4 non-English publications were identified during screening, suggesting that such literature remains limited in this field.

In addition, the predominance of commentaries and reviews (14 out of 25) over empirical studies may limit the depth of evidence available to substantiate claims about real-world implementation, stakeholder experiences, or the practical implications. However, this imbalance also reflects the early and largely conceptual stage of ethical reflection at the intersection of AI and EHR data. Furthermore, many of the reviewed publications were authored by researchers without specialized legal or regulatory expertise, such as legal scholars deeply familiar with the GDPR (General Data Protection Regulation). This may have shaped how certain challenges were framed and further limited discussion of broader ethical issues, such as whether EHR data should be used for AI-driven research at all.

As mentioned above, our efforts to include diverse stakeholders may also have unintentionally reinforced power or knowledge imbalances, particularly due to the underrepresentation of patients. However, as described in the methods, patient participants were briefed in advance to support meaningful engagement. To enhance transparency and relevance, all workshop participants were also invited to review and comment on the workshop reports prior to publication. In addition, while policy perspectives were well represented in the workshops, participants with formal legal or regulatory expertise were underrepresented. This limitation may have influenced how governance and compliance challenges were discussed.

Furthermore, thematic convergence between the literature and the stakeholder workshop discussions may reflect the dominance of Global North perspectives, particularly those shaped by Dutch socioeconomic and institutional norms. While this alignment across methods strengthens the internal coherence of our findings, it also highlights a potential blind spot. Both the literature reviewed and the workshops conducted were situated within contexts that may not capture the unique challenges and priorities of the Global South. As others have noted, failing to account for these limitations risks overlooking perspectives that remain largely absent from this paper [62].

Beyond the institutional contexts examined here, large-scale health data initiatives (for example, the All of Us Research Program [63] and MIMIC-IV [64]) have likewise revealed persistent ethical challenges related to decontextualization, governance, and accountability. Similar questions are now emerging within Europe through the development of the European Health Data Space [54], which aims to harmonize secondary data use across member states. These examples demonstrate that the issues identified in this study re-emerge even in large-scale data initiatives with formal governance frameworks, suggesting that future research should further examine how context-specific governance practices can inform responsible data sharing at scale.

Empirical evaluation and the development of metrics to assess how the proposed ethical measures influence factors such as AI model fairness, data quality, or stakeholder trust fall outside the scope of this study. The goal of this work was to derive ethical measures through a novel approach that combines a scoping review with workshops based on the GEA. Future studies in other settings are needed to further develop and test evaluation frameworks that can operationalize these ethical measures in practice.

Our study deliberately focused on the Dutch and European legal and policy context to support LEAPfROG's goals. The transferability of our stakeholder workshop methods to other countries or institutional settings was beyond its scope and could be explored in future empirical research. Future work could also broaden these discussions by addressing themes such as data sovereignty, the ethical dimensions of data labor in the Global South, and the potential for data colonialism [62,65].

XSL•FO
RenderX

Conclusions

The use of EHR data in AI development is situated at the intersection of health care, data science, ethics, and policy, where competing priorities, values, and practices converge. Given this complex intersection, building responsible AI-based tools requires more than technical solutions and compliance. Our findings point to the need to begin with a critical understanding of the data, for which domain knowledge is essential. EHR data are not neutral but shaped by clinical routines, institutional constraints, and documentation practices. When removed from context, it is vulnerable to misinterpretation, which can cause AI systems to reproduce inequities and reinforce clinical patterns instead of supporting informed decision-making. Addressing these risks is essential to ensure accuracy and accountability. Our mixed methods approach, which included multistakeholder guidance ethics workshops, underscored the need for inclusive, value-sensitive development grounded in the realities of health care practice.

Acknowledgments

Special thanks to Daniël Tijink and Pieter van Kuilenburg for preparing and moderating the first guidance ethics workshop, and to Giovanni Cinà for his support in preparing the second workshop. We would also like to thank the participants of the workshops for their valuable contributions regarding the next steps for the LEAPfROG consortium, as well as the development of this paper.

Finally, the LEAPfROG consortium is a unique cross-sectional collaboration between patients, health care professionals, epidemiologists, medical informatics, and computer science researchers, as well as partners from the medical technology industry, regulatory bodies, and health care insurance. Principal investigator: Dr Joanna E Klopotowska. Work package (WP) leaders: Dr Ronald Cornet (WP1), Prof Dr Annette ten Teije (WP2), Dr Joanna E Klopotowska (WP3), and Dr Stephanie Medlock (WP4). Beneficiaries of the LEAPfROG consortium: Amsterdam University Medical Center, leadership and coordination; Vrije Universiteit Amsterdam, and Open University Heerlen. Ongoing LEAPfROG PhDs: Daniel Fernández-Llaneza, Romy M P Vos, Joris E Lieverse, and Menno Maris (Amsterdam University Medical Center), The LEAPfROG Consortium members include: Ameen Abu-Hanna (Amsterdam University Medical Center), Birgit A Damoiseaux (Amsterdam University Medical Center), Cornelis Boersma (Open Universiteit), Dave A Dongelmans (Nationale Intensive Care Evaluatie foundation), David H de Koning (Amsterdam University Medical Center), Frank van Harmelen (Vrije Universiteit Amsterdam), Gerty Holla (Amsterdam Economic Board), Heiralde Marck (Koninklijke Nederlandse Maatschappij ter bevordering der Pharmacie [KNMP] Geneesmiddelen Informatie Centrum), Iacopo Vagliano (Amsterdam University Medical Center), Jan Pander (AstraZeneca), Jurjen van der Schans (Open Universiteit), Giovanni Cinà (Amsterdam University Medical Center), Kitty J Jager (Amsterdam University Medical Center, European Renal Association-European Dialysis and Transplantation Association [ERA-EDTA] Registry), Leonora van Drop - Grandia (Z-Index), Linda Dusseljee-Peute (Amsterdam University Medical Center), Luuk B Hilbrands (Radboud University Medical Centre), Marcel SG Kwa (College ter Beoordeling van Geneesmiddelen), Marieke A R Bak (Amsterdam University Medical Center), Mariette van den Hoven (Amsterdam University Medical Centre), Martijn G Kersloot (Castor), Nicolette F de Keizer (Amsterdam University Medical Center), Otto R Maarsingh (Amsterdam University Medical Center), Paul Blank (NWO), Piet Heingraaf (Pitts.AI), Renée de Wildt (Nier Vereniging Nederland), Ron Herings (PHARMO Institute for Drug Outcomes Research & Amsterdam University Medical Center), Ron J Keizer (InsightRX), Ruben Boyd (IXA), Sebastiaan L Knijnenburg (Castor), Sipke Visser (Digital Health Link), Stijn Gremmen (Dutch Kidney Foundation/Nierstichting), Teun van Gelder (Leiden University Medical Center), Tjerk S Heijmens Visser (CZ Health Insurance, Zorgverzekeraars Nederland), and Vianda S Stel (Amsterdam University Medical Center, ERA-EDTA Registry). Generative artificial intelligence (ChatGPT, OpenAI) was used solely to provide limited assistance with grammar and phrasing. The authors confirm that all writing, analysis, and interpretation were carried out by the authors without the use of AI tools.

Funding

The study was funded by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO; Dutch Research Council) (KICH1.ST01.20.011). It was co-funded in cash by the Dutch Kidney Foundation and the National Intensive Care Evaluation (NICE) foundation, and in kind by the PHARMO Institute for Drug Outcomes Research, Castor, InsightRX, Z-Index, and Digital Health Link.

Authors' Contributions

MTM prepared the original draft of the manuscript and contributed to conceptualization, formal analysis, investigation, methodology, verification, resources, and visualization. MB, JEK, and RC contributed to conceptualization, investigation, methodology, supervision, verification (MB and JEK), resources (MB and JEK), and review and editing of the manuscript. JEK, RC, and MvdH acquired funding, and MvdH also contributed to the investigation, resources, and review and editing. DF-L contributed to methodology and review and editing. JEL contributed to verification and review and editing. All authors and LEAPfROG consortium members gave final approval of the submitted version.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search query details for systematic search.
[PDF File, 172 KB - jmir_v28i1e79863_app1.pdf]

Multimedia Appendix 2
Report of stakeholder workshop 1.
[PDF File, 660 KB - jmir_v28i1e79863_app2.pdf]

Multimedia Appendix 3
Report of stakeholder workshop 2.
[PDF File, 456 KB - jmir_v28i1e79863_app3.pdf]

Multimedia Appendix 4
Characteristics of included studies.
[PDF File, 313 KB - jmir_v28i1e79863_app4.pdf]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 40 KB - jmir_v28i1e79863_app5.docx]

References

1. Hammack-Aviran CM, Brelsford KM, McKenna KC, Graham RD, Lampron ZM, Beskow LM. Research use of electronic health records: patients' views on alternative approaches to permission. AJOB Empir Bioeth 2020;11(3):172-186. [doi: 10.1080/23294515.2020.1755383] [Medline: 32338567]
2. Jørgensen JT. Twenty years with personalized medicine: past, present, and future of individualized pharmacotherapy. Oncologist 2019 Jul;24(7):e432-e440. [doi: 10.1634/theoncologist.2019-0054] [Medline: 30940745]
3. Fraccaro P, Arguello Casteleiro M, Ainsworth J, Buchan I. Adoption of clinical decision support in multimorbidity: a systematic review. JMIR Med Inform 2015 Jan 7;3(1):e4. [doi: 10.2196/medinform.3503] [Medline: 25785897]
4. Masnoon N, Shakib S, Kalisch-Ellett L, Caughey GE. What is polypharmacy? a systematic review of definitions. BMC Geriatr 2017 Oct 10;17(1):230. [doi: 10.1186/s12877-017-0621-2] [Medline: 29017448]
5. Chowdhury SR, Chandra Das D, Sunna TC, Beyene J, Hossain A. Global and regional prevalence of multimorbidity in the adult population in community settings: a systematic review and meta-analysis. EClinicalMedicine 2023 Mar;57:101860. [doi: 10.1016/j.eclinm.2023.101860] [Medline: 36864977]
6. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. The Lancet 2012 Jul;380(9836):37-43. [doi: 10.1016/S0140-6736(12)60240-2]
7. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3(1):160018. [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
8. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care 2013 Aug;51(8 Suppl 3):S30-S37. [doi: 10.1097/MLR.0b013e31829b1dbd] [Medline: 23774517]
9. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017 Jan 1;24(1):198-208. [doi: 10.1093/jamia/ocw042]
10. Magrabi F, Ammenwerth E, McNair JB, et al. Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications. Yearb Med Inform 2019 Aug;28(1):128-134. [doi: 10.1055/s-0039-1677903] [Medline: 31022752]
11. Kristiansen TB, Kristensen K, Uffelmann J, Brandslund I. Erroneous data: the achilles' heel of AI and personalized medicine. Front Digit Health 2022;4:862095. [doi: 10.3389/fdgth.2022.862095] [Medline: 35937419]
12. Aitken M, de St Jorre J, Pagliari C, Jepson R, Cunningham-Burley S. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. BMC Med Ethics 2016 Nov 10;17(1):73. [doi: 10.1186/s12910-016-0153-x] [Medline: 27832780]

XSL•FO
RenderX

13. Skovgaard LL, Wadmann S, Hoeyer K. A review of attitudes towards the reuse of health data among people in the European Union: the primacy of purpose and the common good. Health Policy 2019 Jun;123(6):564-571. [doi: 10.1016/j.healthpol.2019.03.012] [Medline: 30961905]
14. Piasecki J, Walkiewicz-Żarek E, Figas-Skrzypulec J, Kordecka A, Dranseika V. Ethical issues in biomedical research using electronic health records: a systematic review. Med Health Care Philos 2021 Dec;24(4):633-658. [doi: 10.1007/s11019-021-10031-6] [Medline: 34146228]
15. Morley J, Machado CCV, Burr C, et al. The ethics of AI in health care: a mapping review. Soc Sci Med 2020 Sep;260:113172. [doi: 10.1016/j.socscimed.2020.113172]
16. Tang L, Li J, Fantus S. Medical artificial intelligence ethics: a systematic review of empirical studies. Digit Health 2023;9:20552076231186064. [doi: 10.1177/20552076231186064] [Medline: 37434728]
17. Verbeek PP, Tijink D. Guidance ethics approach: an ethical dialogue about technology with perspective on actions. University of Twente. 2020. URL: https://ris.utwente.nl/ws/portalfiles/portal/247401391/060_002_Boek_Guidance_ethics_approach_Digital_EN.pdf [accessed 2026-02-04]
18. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
19. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 5;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
20. Benchimol EI, Smeeth L, Guttmann A, et al. The reporting of studies conducted using observational routinely-collected health data (RECORD) statement. PLoS Med 2015 Oct;12(10):e1001885. [doi: 10.1371/journal.pmed.1001885] [Medline: 26440803]
21. Krom A, de Boer A, Geurtzen R, de Vries MC. Capabilities and stakeholders - two ways of enriching the ethical debate on artificial womb technology. Am J Bioeth 2023 May;23(5):110-113. [doi: 10.1080/15265161.2023.2191028] [Medline: 37130420]
22. Alami H, Lehoux P, Auclair Y, et al. Artificial intelligence and health technology assessment: anticipating a new level of complexity. J Med Internet Res 2020 Jul 7;22(7):e17707. [doi: 10.2196/17707] [Medline: 32406850]
23. Baric-Parker J, Anderson EE. Patient data-sharing for AI: ethical challenges, catholic solutions. Linacre Q 2020 Nov;87(4):471-481. [doi: 10.1177/0024363920922690] [Medline: 33100395]
24. Bednorz A, Mak JKL, Jylhävä J, Religa D. Use of electronic medical records (EMR) in gerontology: benefits, considerations and a promising future. Clin Interv Aging 2023;18:2171-2183. [doi: 10.2147/CIA.S400887] [Medline: 38152074]
25. Cohen IG, Amarasingham R, Shah A, Xie B, Lo B. The legal and ethical concerns that arise from using complex predictive analytics in health care. Health Aff (Millwood) 2014 Jul;33(7):1139-1147. [doi: 10.1377/hlthaff.2014.0048]
26. Fischer T, Brothers KB, Erdmann P, Langanke M. Clinical decision-making and secondary findings in systems medicine. BMC Med Ethics 2016 May 21;17(1):32. [doi: 10.1186/s12910-016-0113-5] [Medline: 27209083]
27. Ford E, Oswald M, Hassan L, Bozentko K, Nenadic G, Cassell J. Should free-text data in electronic medical records be shared for research? a citizens' jury study in the UK. J Med Ethics 2020 Jun;46(6):367-377. [doi: 10.1136/medethics-2019-105472]
28. Müller S. Is there a civic duty to support medical AI development by sharing electronic health records? BMC Med Ethics 2022 Dec 10;23(1):134. [doi: 10.1186/s12910-022-00871-z] [Medline: 36496427]
29. Kamradt M, Poß-Doering R, Szecsenyi J. Exploring physician perspectives on using real-world care data for the development of artificial intelligence-based technologies in health care: qualitative study. JMIR Form Res 2022 May 18;6(5):e35367. [doi: 10.2196/35367] [Medline: 35583921]
30. Knevel R, Liao KP. From real-world electronic health record data to real-world results using artificial intelligence. Ann Rheum Dis 2023 Mar;82(3):306-311. [doi: 10.1136/ard-2022-222626] [Medline: 36150748]
31. Liaw ST, Liyanage H, Kuziemsky C, et al. Ethical use of electronic health record data and artificial intelligence: recommendations of the primary care informatics working group of the international medical informatics association. Yearb Med Inform 2020 Aug;29(1):51-57. [doi: 10.1055/s-0040-1701980] [Medline: 32303098]
32. Anom BY. Ethics of big data and artificial intelligence in medicine. Ethics Med Public Health 2020 Oct;15:100568. [doi: 10.1016/j.jemep.2020.100568]
33. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 2020 Mar 20;368:l6927. [doi: 10.1136/bmj.l6927] [Medline: 32198138]
34. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol 2022 Nov 5;22(1):287. [doi: 10.1186/s12874-022-01768-6] [Medline: 36335315]
35. Khalid N, Qayyum A, Bilal M, Al-Fuqaha A, Qadir J. Privacy-preserving artificial intelligence in healthcare: techniques and applications. Comput Biol Med 2023 May;158:106848. [doi: 10.1016/j.compbiomed.2023.106848] [Medline: 37044052]
36. Lovis C. Unlocking the power of artificial intelligence and big data in medicine. J Med Internet Res 2019 Nov 8;21(11):e16607. [doi: 10.2196/16607] [Medline: 31702565]
37. Chekroud AM, Bondar J, Delgadillo J, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry 2021 Jun;20(2):154-170. [doi: 10.1002/wps.20882] [Medline: 34002503]

XSL•FO
RenderX

38. Ho CL, Caals K, editors. A Call for an Ethics and Governance Action Plan to Harness the Power of Artificial Intelligence and Digitalization in Nephrology Seminars in Nephrology: Elsevier; 2021. [doi: 10.1016/j.semnephrol.2021.05.009]
39. Atkinson JG, Atkinson EG. Machine learning and health care: potential benefits and issues. J Ambul Care Manage 2023;46(2):114-120. [doi: 10.1097/JAC.0000000000000453] [Medline: 36649491]
40. Breen N, Berrigan D, Jackson JS, et al. Translational health disparities research in a data-rich world. Health Equity 2019;3(1):588-600. [doi: 10.1089/heq.2019.0042] [Medline: 31720554]
41. Darcel K, Upshaw T, Craig-Neil A, et al. Implementing artificial intelligence in Canadian primary care: barriers and strategies identified through a national deliberative dialogue. PLoS ONE 2023;18(2):e0281733. [doi: 10.1371/journal.pone.0281733] [Medline: 36848339]
42. Paulus JK, Kent DM. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. NPJ Digit Med 2020;3(1):99. [doi: 10.1038/s41746-020-0304-9] [Medline: 32821854]
43. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018 Nov 1;178(11):1544-1547. [doi: 10.1001/jamainternmed.2018.3763] [Medline: 30128552]
44. Landau AY, Ferrarello S, Blanchard A, et al. Developing machine learning-based models to help identify child abuse and neglect: key ethical challenges and recommended solutions. J Am Med Inform Assoc 2022 Jan 29;29(3):576-580. [doi: 10.1093/jamia/ocab286] [Medline: 35024859]
45. Bozkurt S, Cahan EM, Seneviratne MG, et al. Reporting of demographic data and representativeness in machine learning models using electronic health records. J Am Med Inform Assoc 2020 Dec 9;27(12):1878-1884. [doi: 10.1093/jamia/ocaa164] [Medline: 32935131]
46. Rubinger L, Gazendam A, Ekhtiari S, Bhandari M. Machine learning and artificial intelligence in research and healthcare. Injury 2023 May;54 Suppl 3:S69-S73. [doi: 10.1016/j.injury.2022.01.046] [Medline: 35135685]
47. Bertelsen PS, Bossen C, Knudsen C, Pedersen AM. Data work and practices in healthcare: a scoping review. Int J Med Inform 2024 Apr;184:105348. [doi: 10.1016/j.ijmedinf.2024.105348] [Medline: 38309238]
48. Hoeyer K. Data Paradoxes: The Politics of Intensified Data Sourcing in Contemporary Healthcare: MIT Press; 2023.
49. Pink S, Ruckenstein M, Willim R, Duque M. Broken data: conceptualising data in an emerging world. Big Data Soc 2018 Jan;5(1):2053951717753228. [doi: 10.1177/2053951717753228]
50. Kennedy H, Hill RL. The feeling of numbers: emotions in everyday engagements with data and their visualisation. Sociology 2018 Aug;52(4):830-848. [doi: 10.1177/0038038516674675]
51. van Amsterdam WAC, van Geloven N, Krijthe JH, Ranganath R, Cinà G. When accurate prediction models yield harmful self-fulfilling prophecies. Patterns (N Y) 2025 Apr 11;6(4):101229. [doi: 10.1016/j.patter.2025.101229] [Medline: 40264961]
52. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2021 Nov;3(11):e745-e750. [doi: 10.1016/S2589-7500(21)00208-9] [Medline: 34711379]
53. Bak M, Madai VI, Celi LA, et al. Federated learning is not a cure-all for data ethics. Nat Mach Intell 2024;6(4):370-372. [doi: 10.1038/s42256-024-00813-x]
54. Staunton C, Shabani M, Mascalzoni D, Mežinska S, Slokenberga S. Ethical and social reflections on the proposed European Health Data Space. Eur J Hum Genet 2024 May;32(5):498-505. [doi: 10.1038/s41431-024-01543-9] [Medline: 38355959]
55. Prainsack B, El-Sayed S, Forgó N, Szoszkiewicz Ł, Baumer P. Data solidarity: a blueprint for governing health futures. Lancet Digit Health 2022 Nov;4(11):e773-e774. [doi: 10.1016/S2589-7500(22)00189-3] [Medline: 36307191]
56. Bak MAR, Ploem MC, Tan HL, Blom MT, Willems DL. Towards trust-based governance of health data research. Med Health Care Philos 2023 Jun;26(2):185-200. [doi: 10.1007/s11019-022-10134-8] [Medline: 36633724]
57. Proost MD. Data solidarity disrupted: musings on the overlooked role of mutual aid in data-driven medicine. Kennedy Inst Ethics J 2023;33(4):401-419. [doi: 10.1353/ken.2023.a931052] [Medline: 38973484]
58. Turnhout E, Metze T, Wyborn C, Klenk N, Louder E. The politics of co-production: participation, power, and transformation. Curr Opin Environ Sustain 2020 Feb;42:15-21. [doi: 10.1016/j.cosust.2019.11.009]
59. Cazzolli R, Sluiter A, Guha C, et al. Partnering with patients and caregivers to enrich research and care in kidney disease: values and strategies. Clin Kidney J 2023 Sep;16(Suppl 1):i57-i68. [doi: 10.1093/ckj/sfad063] [Medline: 37711636]
60. Dedding C, Aussems K. Participatie, het verschil tussen een methode en een kritisch paradigma [participation, the difference between a method and a critical paradigm]. TSG Tijdschr Gezondheidswet [TSG J Health Sci] 2024;102(3):81-87. [doi: 10.1007/s12508-024-00439-9]
61. Lysen F, Wyatt S. Refusing participation: hesitations about designing responsible patient engagement with artificial intelligence in healthcare. J Responsible Innov 2024 Dec 31;11(1):2300161. [doi: 10.1080/23299460.2023.2300161]
62. Roche C, Lewis D, Wall P. Artificial intelligence ethics: an inclusive global discourse. arXiv. Preprint posted online on Aug 23, 2021. [doi: 10.48550/arXiv.2108.09959]
63. Ramirez AH, Sulieman L, Schlueter DJ, et al. The All of Us research program: data quality, utility, and diversity. Patterns (N Y) 2022 Aug 12;3(8):100570. [doi: 10.1016/j.patter.2022.100570] [Medline: 36033590]
64. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023 Jan 3;10(1):1. [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]

XSL•FO
RenderX

65. Arora A, Barrett M, Lee E, Oborn E, Prince K. Risk and the future of AI: algorithmic bias, data colonialism, and marginalization. Inf Organ 2023 Sep;33(3):100478. [doi: 10.1016/j.infoandorg.2023.100478]

Abbreviations

AI: artificial intelligence
AKI: acute kidney injury
CKD: chronic kidney disease
DAKI: drug-induced acute kidney injury
EHR: electronic health record
GDPR: General Data Protection Regulation
GEA: Guidance Ethics Approach
LEAPfROG: Leveraging Real-World Data to Optimize Pharmacotherapy Outcomes in Multimorbid Patients Using Machine Learning and Knowledge Representation Methods
MeSH: Medical Subject Headings
NVN: Nierpatiënten Vereniging Nederland (Kidney Patients Association Netherlands)
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis

Edited by A Stone; submitted 30.Jun.2025; peer-reviewed by C Wu, DS Jaisankar, N Gupta; revised version received 25.Nov.2025; accepted 25.Nov.2025; published 02.Mar.2026.

Please cite as:
Maris MT, Klopotowska JE, Cornet R, Van den Hoven MA, Lieverse JE, Fernández-Llaneza D, Bak MAR, LEAPfROG Consortium
The Ethics of Leveraging Routinely Collected Patient Data for AI Development: Mixed Methods Study
J Med Internet Res 2026;28:e79863
URL: https://www.jmir.org/2026/1/e79863
doi:10.2196/79863

© Menno T Maris, Joanna E Klopotowska, Ronald Cornet, Mariëtte A Van den Hoven, Joris E Lieverse, Daniel Fernández-Llaneza, Marieke A R Bak, LEAPfROG Consortium. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Ethical Knowledge, Challenges, and Institutional Strategies Among Medical AI Developers and Researchers: Focus Group Study

Sophia Fantus[1], PhD; Jinxu Li[2], MA; Tianci Wang[3], MS; Lu Tang[2], PhD

[1]School of Social Work, The University of Texas at Arlington, Arlington, TX, United States

[2]Department of Communication & Journalism, Texas A&M University, College Station, TX, United States

[3]Burnett School of Medicine, Texas Christian University, Fort Worth, TX, United States

Corresponding Author:
Lu Tang, PhD
Department of Communication & Journalism
Texas A&M University
4234 TAMU
College Station, TX, 77843-4234
United States
Phone: 1 2136754090
Email: ltang@tamu.edu

Abstract

Background: As artificial intelligence (AI) becomes increasingly embedded in clinical decision-making and preventive care, it is urgent to address ethical concerns such as bias, privacy, and transparency to protect clinician and patient populations. Although prior research has examined the perspectives of medical AI stakeholders, including clinicians, patients, and health system leaders, far less is known about how medical AI developers and researchers understand and engage with ethical challenges as they develop AI tools. This gap is consequential because developers' ethical awareness, decision-making, and institutional environments influence how AI tools are conceptualized and deployed in practice. Thus, it is essential to understand how developers perceive these issues and what supports they identify as necessary for ethical AI development.

Objective: The objectives of the study were twofold: (1) to examine medical AI developers' and researchers' knowledge, attitudes, and experiences with AI ethics; and (2) to identify recommendations to enhance and strengthen interpersonal and institutional ethics-focused training and support.

Methods: We conducted 2 semistructured focus groups (60-90 minutes each) in 2024 with 13 AI developers and researchers affiliated with 5 US-based academic institutions. Participants' work spanned a wide variety of medical AI applications, including Alzheimer disease prediction, clinical imaging, electronic health records analysis, digital health, counseling and behavioral health, and genotype–phenotype modeling. Focus groups were conducted via Microsoft Teams, recorded, and transcribed verbatim. We applied conventional qualitative content analysis to inductively identify emerging concepts, categories, and themes. Coding was performed independently by 3 researchers, with consensus reached through iterative team meetings.

Results: The analysis identified four key themes: (1) AI ethics knowledge acquisition: participants reported learning about ethics informally through peer-reviewed literature, reviewer feedback, social media, and mentorship rather than through structured training; (2) ethical encounters: participants described recurring ethical challenges related to data bias, patient privacy, generative AI use, commercialization pressures, and a tendency for research environments to prioritize model accuracy over ethical reflection; (3) reflections on ethical implications: participants expressed concern about downstream effects on patient care and clinician autonomy, and model generalizability, noting that rapid technological innovation outpaces regulatory and evaluative processes; and (4) strategies to mitigate ethical concerns: recommendations included clearer institutional guidelines, ethics checklists, interdisciplinary collaboration, multi-institutional data sharing, enhanced institutional review board support, and the inclusion of bioethicists as members of the AI research team.

Conclusions: Medical AI developers and researchers recognize significant ethical challenges in their work but lack structured training, resources, and institutional mechanisms to address them. Findings of this study underscore the need for institutions to consider embedding ethics into research processes through practical tools, mentorship, and interdisciplinary partnerships. Strengthening these supports is essential to preparing the next generation of developers to design and deploy ethical AI in health care.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e79613) doi:10.2196/79613

KEYWORDS

AI ethics; AI; artificial intelligence; ethics; focus group; medical

Introduction

Artificial Intelligence in Health and Medicine

Artificial intelligence (AI) is reshaping health care; AI tools are aimed at reducing costs [1], streamlining clinical workflow [2], and facilitating clinician and patient experiences [3]. Current AI applications may include assistance with clinical decisions, image-based diagnosis, self-diagnosis, mental health screening, and chronic disease management [2]. For example, electronic health records (EHRs) use natural language processing to support clinical decisions [4], and at-home AI monitoring systems assist older adults and those with long-term chronic illnesses, potentially alleviating caregiver burden [5]. Health care providers have started to use AI for medical imaging, diagnosis and disease screening, and prediction [6-9]. Furthermore, emerging scholarship demonstrates that AI has shown, to some extent, faster diagnostic speed and higher accuracy than human experts in image analysis and precision medicine [1,10]. The speed at which AI has been integrated and accepted into health care networks and its ease of access for users are unprecedented.

Despite these benefits, there are significant obstacles to AI implementation in clinical practice. Concerns about patient data security, privacy, clinician and patient autonomy, and decision-making may erode trust in AI outputs [11,12]. In addition, the pace of AI technology innovation often surpasses regulatory guidance at the federal and state levels [13]. Identifying and understanding ethical challenges may help establish practice and policy guidelines across health systems. Such mechanisms may ensure that AI developers and researchers, along with clinicians, patients, and health system leadership and administration, adapt and integrate AI that considers the needs and perspectives of all stakeholders. The purpose of this study was to explore AI developers' and researchers' understanding of and training in medical AI ethics.

Ethical Attitudes and Knowledge of Medical AI

Recent scholarship has shown rising ethical concerns among various stakeholders such as clinicians, patients, families, and policymakers who engage with medical AI [12,14-16]. Clinicians report that AI tools serve as a time-saving benefit in completing administrative tasks, which may effectively increase clinical productivity and patient engagement [6,17]. Yet, they are concerned with patient data privacy, the impacts on the clinician-patient relationship, and the possibility that the financial burdens of AI tools may heighten health inequities [7,17,18]. Patients and families voice similar concerns, focusing on patient autonomy and shared decision-making [19,20]. Patients articulate unease with the application of AI in treatment recommendations, medication administration, and surgical procedures [20].

Gap in Scholarship

The attitudes and perspectives of medical AI ethics among health care stakeholders are important. Yet, perspectives of other key stakeholders, such as AI developers and researchers, are underrepresented in research [12]. Previous studies demonstrate that ethical issues, including privacy and data security, fairness, transparency, and reliability of machine learning predictive analytics, are encountered by AI developers [21-23]. Yet, the perceived responsibility of developers and researchers to mitigate potential harms varies widely; some AI developers are cognizant of the broader societal impacts of AI (beyond technical considerations and optimization), while others feel disconnected and detached from direct patient and clinician outcomes [14,24,25]. Many AI developers report barriers to mitigating AI harms, including limited authority to make such decisions, and external pressures to deliver products quickly, all of which can hinder ethical reflexivity [26].

With a paucity of evidence-informed data on developers' AI knowledge and attitudes, further research is necessary to understand how AI ethics is addressed prior to deployment. Academic institutions play a central role in AI research and development, which lays the foundation for industry's AI design and application [14,24]. Academia has an important role in educating, training, and shaping the future generation of AI developers. This study presents a unique opportunity to guide policy, practice, and education efforts in research institutes that are aligned with the needs of AI developers and consider the deployment of ethical AI across health systems.

Research Rationale

AI developers and researchers work on algorithms that ultimately shape medical AI tools. Yet, clinicians often assume that AI tools used in clinical settings have been ethically scrutinized prior to deployment [27]. To understand ethical encounters of AI design, this study identifies the knowledge, attitudes, and training in medical AI ethics among AI researchers and developers. As an exploratory, pilot study, this work aims to offer an initial, in-depth understanding of how developers and researchers experience and navigate ethical challenges in medical AI. Rather than seeking generalizability, our goal was to capture diverse perspectives across academic contexts to illuminate key issues and inform the design of future large-scale, quantitative investigations on AI ethics training and institutional practices. Findings may inform strategies to facilitate AI ethics integration in development.

Methods

Recruitment and Sampling

The research team qualitatively explored perspectives of medical AI ethics among AI researchers and developers who were employed at academic institutions in the United States. Members of the research team had expertise in health communications,

XSL•FO
RenderX

AI and health promotion, and bioethics. Participants' inclusion criteria were (1) aged ≥18 years; (2) ability to read, understand, and communicate in English; (3) employed at an academic institution; (4) involved in medical AI research and development; and (5) consent to participate in a focus group. Participants were recruited through purposive sampling and chain referral methods to appropriately reach individuals who had an academic background in working with medical AI tools and the creation of algorithms. A study announcement and blurb were sent through listserv emails, through networks and contacts of the research team, and by word-of-mouth. Participants responded by email, and a focus group was scheduled that met participants' availability through an anonymous When2Meet Poll.

Data Collection

Focus groups (n=2; 60 minutes each) were held over Microsoft Teams and facilitated by the lead author. The facilitator had no pre-existing relationships with any participants. The lead author introduced herself to the research team and explained her background in clinical ethics, bioethics, and health scholarship prior to initiation. A semistructured interview guide was used in each focus group to reflect on participants' knowledge, attitudes, and encounters with AI ethics, and practical strategies to enhance or improve ethics education and training. Participants did not receive the interview guide prior to the scheduled focus group. Some example questions included the following: (1) What do you think is the extent of your AI ethics knowledge? (2) What is your prior experience with AI ethics? (3) What are the ethical concerns you have when conducting AI research? (4) How as a research team do you deliberate ongoing ethical issues you face? (5) In your current workplace, what training or learning opportunities are there with AI ethics? (6) What can your supervisor or the institution or university do to support you in understanding and identifying ethical issues in AI research? The interview guide was developed and piloted by members of the research team.

All focus groups (roughly 60-90 minutes each) were recorded and transcribed verbatim. No repeat interviews or focus groups were conducted. Transcripts were cleaned for errors and deidentified to protect participant confidentiality. Recordings and transcripts were stored on the university's secure password-protected server, and only members of the research team had access to the data. Thematic saturation was assessed through iterative review during and after the second focus group. At that stage, no substantively new themes emerged, and only minor variations of existing concepts were observed. We therefore determined that thematic saturation had been sufficiently achieved for the purposes of this pilot, exploratory study, and data collection concluded accordingly. In addition, members of the research team contacted different principal investigators within their respective academic units to enhance diversity in disciplinary backgrounds, institutional affiliations, and research areas. This strategy broadened the participant pool and helped capture a wider range of perspectives on AI ethics while maintaining the feasibility of this exploratory qualitative study.

Data Analysis

We applied conventional content analysis, an inductive qualitative approach used when limited theory or research exists on a topic [28]. Analysis occurs directly from the data without the use of preconceived frameworks or codebooks, allowing researchers to conduct in-depth exploration of raw data [28]. All focus group transcripts (roughly 70 pages of transcript data) were redacted and anonymized, and participants were given a specific numerical code to ensure the accuracy of responses. The 2 focus group transcripts were disseminated to 3 members of the research team (known as the coding team) to review independently from one another. Transcripts were reviewed, and data were coded inductively to form new insights and perspectives on medical AI ethics. Initial coding was conducted first to highlight exact words or phrases to denote emerging concepts. The whole research team met to discuss initial thoughts and impressions from the transcript and to develop a codebook. The transcripts were reviewed a few more times by the three independent coders to organize: (1) codes into categories, (2) categories into clusters, and (3) clusters into emerging themes. The coding team met frequently to finalize the codebook and to reach consensus on emerging themes and patterns from the data. Discrepancies in coding were discussed by the coding team to reach consensus. Codes were iteratively clustered into broader conceptual categories and then synthesized into higher-order themes that reflected shared meanings across participants. Throughout this process, the team also noted and discussed negative or divergent cases to ensure that contrasting perspectives were represented and that the final themes captured the full range of participant experiences. Once the codebook was finalized, the whole research team met to review findings and finalize themes. Rigor and trustworthiness were attained through peer debriefing with other AI and data science experts. The themes were grounded in participant data to capture their perspectives, thoughts, and insights on medical AI ethics.

Ethical Considerations

The study was approved by the Institutional Review Boards (IRBs) of Texas A&M University (approval number: IRB2023-0396D) and the University of Texas at Arlington (approval number: 2023-0234) prior to participant recruitment. Before the scheduled focus groups, participants received a copy of an informed consent form, which they signed and returned electronically. At the beginning of each focus group, the research team reviewed the consent information again orally, provided time for clarifying questions, and reiterated confidentiality limits and group norms. To protect the privacy and confidentiality of the participants, we transcribed the video recordings and removed all identifying information, including names, geographic location, or university affiliations. We conducted the data analysis based on the anonymized transcripts instead of videos. Each participant was paid US $20 in an Amazon gift card as compensation for their time.

XSL•FO
RenderX

Results

Overview

We interviewed a total of 13 participants employed in medical AI research and development. No participants declined or withdrew participation either before or during the focus group. Six (46%) participants were women, and 7 (54%) were men. Six (46%) participants identified as Chinese, and other participants identified as Asian Indian, Middle Eastern, Egyptian, Bangladeshi, Pacific Islander, or Taiwanese; only 1 participant identified as White. Participants held a range of positions, including research faculty (5/13, 38%), graduate students or research assistants (7/13, 54%), and programmer (1/13, 0.07%). Focus group participants represented 5 distinct academic institutes from different regions in the United States. Participants' AI research included (1) disease or surgery outcomes prediction, (2) prediction and optimization of treatment, (3) analysis of electronic medical records or diagnostic imaging, (4) genetic analysis and genotype–phenotype correlation, (5) AI in counseling and behavioral health, and (6) AI in digital health and clinical trial work.

Four themes emerged from the analysis of focus group transcripts: (1) AI ethics knowledge acquisition that demonstrates how and where participants obtain AI ethics information, (2) ethical encounters that identify the main ethical issues that arise from algorithm development and design, (3) perceptions of ethical encounters to understand the implications of unresolved ethical encounters, and (4) recommendations and strategies to facilitate ethical deliberation and debrief in the workplace (Table 1).

XSL•FO
RenderX

Table 1. Themes and examples identified through conventional content analysis of focus groups on medical artificial intelligence (AI) ethics among US-based AI developers and researchers (2024).

Themes (innovation-decision process) and subthemes	Description	Quote
AI knowledge acquisition (knowledge)		
Peer-reviewed publications	Learning about AI ethics from published studies discussing bias, fairness, and responsible AI.	"So I was following up some publications, so I started to see the trend of new publications coming up and talking about like as I mentioned AI bias…"
Reviewers' feedback	Gaining awareness of ethical issues through reviewers' comments during the publication process.	"The first time I realized it was when I submitted my manuscript to Nature-like journals; most reviewers pointed it out, and that's when I started thinking about ethics seriously. "
Social media and AI policy updates	Following experts and organizations online to stay current with national and international AI ethics guidelines.	"I don't know if you've seen the recent news, like DeepMind's phone app for ChatGPT."
Informal mentorship and seminars	Receiving ethics training through informal networks, research supervisors, and academic workshops.	"Some competitions from big tech companies like Microsoft and Meta discussed these topics, and in our school, we also have weekly seminars about them."
Lack of formal training	Having little to no structured ethics education leads to uncertainty about ethical risks in AI research.	"I haven't received much training in AI ethics, so sometimes I don't even know what the problems are. Getting more training would help me recognize the issues and address them better."
Ethical encounters		
Data bias and fairness	Challenges related to underrepresentation in training data and unfair model outputs.	"Sometimes the data we use don't really represent everyone, so the model ends up being unfair to certain groups."
Privacy concerns	Issues with using patient data without proper consent or beyond the original intended use.	"Using patient data can be tricky; we're not always sure if we have full consent or if it's okay to reuse it for other purposes."
Use of generative AI	Concerns about researchers using tools like ChatGPT to fabricate or skip steps.	"Some people just ask ChatGPT to write sections for them, and that really blurs the line between help and fabrication."
Commercialization pressures	Ethical concerns regarding profit-driven deployment by tech companies over academic integrity.	"Once big companies get involved, the focus often shifts from research integrity to making profits."
Focus on accuracy over ethics	Some researchers prioritize performance over ethical considerations.	"Everyone talks about model accuracy, but barely anyone mentions the ethical side of it."
Reflections on ethical implications		
Model generalizability and explainability	Ethical concerns arise when models cannot be applied broadly or are not easily interpretable.	"Sometimes the model works great on one dataset but fails completely on another, and we don't really know why."
Impact on patient care	Researchers worry about AI models causing harm or failing to help diverse patient populations.	"If the model gives the wrong prediction, it could actually harm patients instead of helping them."
Clinician autonomy and displacement	Fears that AI may replace doctors or alter clinician-patient relationships.	"Some doctors worry that AI might start making decisions for them or replace parts of their job."
Technological pace vs evaluation speed	Difficulty in evaluating AI tools quickly enough to match their development speed.	"AI is moving so fast that our evaluation methods can't really keep up."
Ethical burden on researchers	Responsibility to address AI ethics falls heavily on developers without adequate support.	"We're the ones expected to think about ethics, but no one really gives us the tools or training to do it properly."
Strategies to mitigate ethical concerns (implementation and confirmation)		
Guideline communication	Improve access to updated ethical AI guidelines and standards.	"We really need clearer and more accessible guidelines on AI ethics; sometimes it's hard to even find the latest ones."
Ethics checklists and scenarios	Using predefined lists or cases to test and evaluate model ethics and bias.	"Having a checklist or real cases to go through would make it easier to see where our model might go wrong ethically."
Data collaboration and diversity	Partnering with other institutions to diversify datasets and reduce bias.	"If we could share data across more institutions, the models would be less biased and more reliable."
IRB[a] support and governance	Having AI-specific ethics experts within IRBs to guide responsible research.	"It would help a lot if IRBs had someone who actually understands AI to guide us on the ethical parts."

XSL•FO
RenderX

Themes (innovation-decision process) and subthemes	Description	Quote
Inclusion of bioethicists	Adding bioethics experts to AI teams to help identify and resolve ethical issues.	"Having a bioethicist on the team would make us think about these issues more seriously from the start."

[a]IRB: Institutional Review Board.

Medical AI Ethics Knowledge Acquisition

Participants discussed various avenues through which they sought AI ethics information and knowledge, including peer-reviewed publications, journal feedback, social media, and informal institutional learning. Several participants mentioned journal submission guidelines or peer review feedback that relayed information on AI ethics or included statements that mentioned AI use and plagiarism:

> *Although I had studied ethics during my medical school, I never paid attention during machine learning research. The first time I came to know was when I submitted my manuscript…the first thing they [reviewers] pointed out was about this [ethics]…most of the reviewers were concerned about it. So that's when I started thinking about it more seriously. I came to start thinking about ethics because if you publish in good journals, people will point out those things.* [P2]

Others observed a rise in publications on AI ethics and begun to read peer-reviewed articles for information: "I started seeing the trend of new publications coming up, talking about bias, fairness, and then AI ethics" (P5).

Other participants relied on social media to inform AI ethics knowledge:

> *I've been following [on X] anyone that has their hands in AI ethics or AI policy and checking all the guidelines, not just institutional levels, but national and international levels. It's hard to keep up with all the literature that's being pumped out right now. But it's important to at least familiarize yourself with some of the different pieces…and what the relevant concerns are that are transcending that international sphere.* [P6]

Social media was perceived as more current and relevant than peer-reviewed publications, able to keep up with fast-paced developments.

A small number of participants described that AI research ethics knowledge derived from informal discussions with supervisors or participation in university seminars or workshops:

> *I have really lucked out into having good people in my circle and training me. I think that's a huge resource in terms of understanding ethics and AI, and then also intentional engagement with current guidelines that are being put out.* [P6]

Yet, one participant remained silent during this discussion. When prompted, the participant stated having little to no knowledge about AI ethics:

> *I have received not too much training in AI ethics, so that is why I don't even know what the problems are. Even if I'm making some mistakes, I don't know if those are things that I should have been careful about.* [P2]

This participant's experience is important to elucidate, as it shows potential training gaps and impacts on students and researchers.

Ethical Encounters in AI Development

Participants reflected on ethical challenges encountered in their research environments, including concerns about data bias, privacy, commercialization, and the use of generative AI tools. Participants discussed the fabrication of data, in which the reliance on AI-generated tools, like ChatGPT, has enabled colleagues to skip steps through automated written responses. Yet, the primary ethical obstacle in medical AI research, as reported by participants, was bias and underrepresentation within training datasets. The ethical issues were consequences related to predictive modeling and fairness, especially how data omission could disproportionately exclude people of color. As a participant who worked on radiation therapy, observed:

> *When we build the predictive model, if our model is just purely based on the data we collected, it seems like it's not very fair for Black people or Asian people. That's the issue we are currently facing.* [P4]

Another participant agreed, stating:

> *Most of the data are coming from European [and] the therapeutics will be ultimately optimized for a certain group of people [so] it can't be generalized. If you are not careful with what kind of metric you're using to assess and evaluate that model, you're pretty much classifying everything as negative. And institutions like to incorporate these models into their systems. If you put more weight on them [the positive cases] you might be identifying the white skin tone but not the darker skin tone.* [P1]

There were also concerns of data security and the risk of breaching patient data privacy:

> *If you're using patient data without their proper consent or you use data that trains a model that is then used for something else that's not within the previously defined scope, that is not ethical.*

A participant discussed that privacy was also the "need to test it [the model] and then be transparent to the community and [provide] the proper instructions of the model's performance degree" (P10) to ensure that the model is explainable and interpretable to key stakeholders.

Participants deliberated on the dangers of the commercialization of AI technologies and limited regulations:

XSL•FO
RenderX

> *What I am really concerned with is that these big tech companies are pushing very hard to deploy their AI model into the hospital system. It's linked to profit, a very profitable market; if those big tech companies want to push their product, I don't know whether they will do it with the same level of checks and balances because it's profit-driven, and they can promise a lot of money, and we cannot make that promise.* [P9]

The overall ethical concern related to commercialization was the fast-paced development of AI and the time sensitivity of implementing AI into health care spaces; for-profit companies will be selected over evidence-informed AI research programs.

> *The perceived competitiveness between big tech corporate research and also academic research and I feel like they are not playing by the rules because they can skirt and essentially do things that we have to abide by like privacy issues and so forth.* [P8]

In contrast, there were participants who did not perceive these as ethics issues but rather as an accuracy issue: "I'm not really focused on ethics. I tend to focus more on accuracy, something that will make the model better but not actually the ethics" (P3).

Reflections on Ethical Implications

Participants described how the ethical encounters stated above influenced research design and modeling choices and raised broader concerns about patient care, clinician roles, and the future of health care. For example, issues related to fairness and bias influenced generalizability:

> *from a data scientist perspective, it's an issue; you cannot have a very accurate model with very high bias. You can build your model, but we want the model to have higher generalizability; we need to take this issue from a data scientist perspective.* [P4]

Participants noted that limits to explainability of AI impeded solutions to resolve ethical issues:

> *with so much advancement in AI technology, there is still no standard correct ways of evaluating my model because I haven't understood my data or the distribution of the data yet.* [P1]

Other participants considered ways ethical issues in AI development may impact patient care and physician interactions. For example, participants who worked on large language models deliberated on how AI tools can generate clinical notes for the patient and questioned the accuracy of how "clinical notes could be to the specific patient…. we don't know how that benefits the patients" (P7). Questions related to predictive modeling also drew fears of perpetuating patient harms:

> *How do we balance advancing healthcare to truly help patients in this unprecedented way, but also make sure that we're not exploiting them or using models that aren't appropriate for them?* [P6]

Participants reported that physician autonomy and patient-physician relationships were another important area to identify the ethical implications of AI deployment in health care. One participant asked, "whether AI is going to replace certain jobs and tasks and maybe even eventually replace doctors; that's a conversation I have about my research" (P8). The timing to evaluate technology was an added concern that could impact patient care:

> *By the time you come up with a standard metric that you need to satisfy your AI model to be deployed in a healthcare facility…maybe the technology has completely changed by then. I don't know what the solution would be, but clinicians, researchers, lawmakers, you know, everybody needs to be on board because they can no longer take that long to evaluate a technology.* [P1]

The perceived impacts of AI ethics placed added burdens and responsibilities on AI researchers and developers. Heightened attention on AI ethics placed more obligations on AI developers and researchers to resolve these issues, yet without training or learned mitigation strategies.

Strategies to Mitigate Ethical Concerns

Participants described a range of strategies to address AI ethics in research, including individual practices, team-based approaches, and institutional-level interventions. Participants suggested individual and team-based approaches that foster transparent communication and knowledge mobilization. For example, a participant emphasized the need for

> *…good communication about the latest guidelines that are available from different communities. If that becomes available to use as students and even as faculty that will be more helpful to make us more compliant with those regulations.* [P5]

Guidelines, in turn, can assist in the design and development of checklists or critical scenarios to mitigate biased models. One participant stated

> *…maybe we can have a checklist on what we need to see before doing something that's more concrete. I know it's difficult to do that in AI ethics because we don't directly use it for patient outcomes right now. But I think it will be a good starting point to have some kind of checklist on what we should be careful about.* [P2]

Another participant wanted the actual model to counteract biases:

> *The first check should be done on the data and how the data biases have been handled by the AI models. And last, what are the abusive ways this model can be used? We should have some critical scenario by which we can test our model, like some exerted test on the product to see whether this product is stable up to two years…whether it is up to our expectation.* [P1]

Participants advocated for organizational and institutional strategies to support ethical AI development. One approach mentioned was multi-institutional collaboration to improve data quality and diversity and mitigate bias by increasing access to larger, more representative datasets. One participant said: "Where you don't have enough data to support the deep learning [models], we have to collaborate with other institutes to not

XSL•FO
RenderX

only expand the sample size but also introduce diversity into the data" (P4). Participants also called for ethics consortiums to foster ethical awareness and skill development. For instance, a participant said: "I feel like getting more training will help me more to even identify what the problems are and then address them" (P2).

An added strategy was to equip the IRB with AI-specific guidance or AI ethics expertise on regulatory committees to enhance regulation through a uniform approach and facilitate adherence to best practices. IRB involvement could promote a more consistent approach to oversight and improve adherence to best practices. However, several participants expressed frustration that current assistance or guidance sought from the IRB often resulted in confusion rather than clarity:

> *We have those IRB boards and maybe some better governance…to have somebody also on AI ethics and being responsible for sharing that awareness as well as ensuring that we are going through the guidelines and sticking through the regulations.* [P5]

Finally, for some participants, a bioethicist or ethics expert should be involved as a potential interdisciplinary member of the research team:

> *When you don't have a bioethicist at your beck and call or infused in the research in some degree, that makes it really tricky too because you might not have the checks and balances that are appropriate in maybe expanding your research or getting into the right market.* [P6]

Discussion

Overview

The integration of medical AI in preventive care and clinical decision-making means that researchers, data scientists, and those involved in the design and development of AI need to start becoming attuned to its clinical impacts. This study aims to address gaps in scholarship by examining AI researchers' ethics perspectives. Academic institutions, such as universities and research institutes, play a central role in educating the future generation of AI developers on AI ethics and design.

Findings from the study inform how medical AI may be diffused into health care settings and how its use communicated effectively between physicians, staff, and patients. The themes from this study may be adapted into Rogers' diffusion of innovation framework [29]. Rogers maps a 5-stage decision-making process to evaluate and adopt AI innovation into practice. The series of stages is as follows: (1) knowledge (gains understanding), (2) persuasion (reflect on attitudes), (3) decision (activities and experiences that lead to choice), (4) implementation (its actual use in practice), and (5) confirmation (to avoid dissonance and conflict).

Our findings ought to be conceptualized within the diffusion of innovation framework to understand how perspectives of AI researchers and developers offer insight into the steps, attitudes, and barriers that influence the decision-making of medical AI adoption and integration. Findings from this study may inform policy, practice, and education efforts to readily prepare AI researchers and developers to identify and examine ethical encounters in their work and to illustrate how medical AI attitudes, perceptions, and support may influence adoption or rejection [29].

AI Ethics Knowledge Acquisition

In the knowledge stage of individual decision-making, participants in this study received information about AI ethics from a multitude of sources, including social media, peer review journal commentary and publications, and voluntary workshops and seminars. Students' particular focus on social media as an access point for AI knowledge may be an important consideration to assess (1) the type of accuracy of messaging received and (2) the ethical issues being described and disseminated. Knowledge garnered through social media may filter into how students understand and evaluate their own research and ethical encounters, including how early adopters may rely more on social media than on peer-reviewed sources. For example, participants who described issues as rooted in accuracy rather than ethics may benefit from conversations and messaging that deciphers ethical issues from technical issues; the ability to understand how to identify and label issues as ethical (rather than solely technical) could enhance medical AI ethics knowledge and lead to a more nuanced and robust deliberation on how ethics may impact medical AI design, development, and deployment.

Ethical Encounters and Resolution Strategies

Past research experiences of participants impacted their attitudes and perspectives of medical AI. Bias and fairness were central ethical challenges identified by participants, particularly the underrepresentation of people of color in training datasets. Such omissions risk reinforcing structural inequities and limiting the generalizability of medical AI systems. To address these issues, future research should prioritize diversifying medical datasets and integrating fairness auditing across development, supported by multi-institutional collaborations and community engagement to ensure representativeness, transparency, and accountability [30-32]. Together, these efforts can help mitigate the disproportionate exclusion of marginalized groups and promote more equitable AI-driven health outcomes to address negative attitudes toward medical AI innovation.

Extant scholarship echoes the current study's findings by demonstrating that AI developers possess some awareness of key ethical principles, such as fairness, data security, transparency, and reliability [14], that may persuade their decision-making in adoption. Nichol et al [14], for example, conducted semistructured interviews with 40 employees from AI organizations. Participants in the study identified potential impacts of ethical issues, such as violation of patients' privacy, misdirected health care practices, and disrupted health care systems. Other studies have similarly shown that some AI developers are sensitive to broader societal impacts of ethical AI [24,25], beyond technical issues and optimization of their algorithms [33,34].

Although participants in this study were able to identify emerging ethical issues and had thoughtfully evaluated how

XSL•FO
RenderX

these issues would impact patient and clinician experiences, there were limited resolution strategies. These experiences impacted AI researchers' decision-making, including not knowing whether to adopt or reject innovations in their work. For participants, ethical encounters were often left unresolved with no clear direction on how to proceed. The lack of informed decision-making was rooted in a lack of clarity from institutions and left participants feeling that AI researchers and developers held an undue burden in deploying ethical AI tools without further scrutiny or analysis. The added pressure on AI developers and researchers to perform was perceived as a challenge, and our findings suggest that other key stakeholders (including physicians or clinicians and health systems) ought to contribute to ethical decision-making when AI is used in practice. Thus, findings from this study show that decisions of whether to adopt or reject AI ought to include diverse perspectives to allow for more information, to identify problems, and to have support [29].

AI models must continue to be questioned and analyzed by stakeholders even after deployment. With AI technologies changing so rapidly, participants struggled to balance the fast-paced development of AI algorithms with the ethical concerns that arose. This led participants to articulate that medical AI tools ought to be continuously reviewed and evaluated.

The barriers and limited support indicate that implementation and confirmation, the final stages in the innovation-decision process, may be difficult to reach. Participants described that in the development of medical AI innovation, they often evaluated long-term impacts on patients and families and desired further support from mentors, supervisors, and organizational leadership. The perspectives of participants show that there continue to be conflicting messages and dissonance among researchers and developers regarding the adoption and use of medical AI in practice settings. Further organizational practices and policies ought to be considered to assist in decision-making activities to facilitate a more robust and comprehensive adoption process.

This study's findings echo prior work wherein AI developers voiced confusion regarding their own roles and responsibilities in mitigating the potential harm of their tools, compounded by perceived limited authority, external pressures to produce, and the difficulty of balancing productivity and ethical considerations [26,33]. Algorithm development is highly complex and iterative, making it difficult for researchers to predict its ethical impact and apply oversight in the process. As participants in this study noted, transparency and explainability were key ethical issues, and a gap in accessible checklists or guidelines heightened obstacles to elucidating datasets and explaining patterns to clinicians and patients who may rely on these algorithms for diagnosis and treatment. The issue here is that resolving ethical encounters requires additional time and energy from AI developers, which may be an added challenge in a high-stress environment that is at odds with the fast-paced development of commercial AI tools [14]. The capacity to build collaborative environments, hold ethics consortiums, and have a robust network of people and resources to support AI ethics awareness, knowledge, and action is critical to support AI developers and researchers. Relieving some of the burdens on AI developers and researchers with institutional mechanisms can model an environment that supports ethical rigor and deliberation and lead to reinforcement and confirmation of medical AI technologies in practice settings.

As pointed out by Mittelstadt [35], compared with medicine, the field of AI research is much more heterogeneous, without defined common aims, fiduciary duties, or historical professional norms. The constant changes and shifts related to AI policy and procedure create difficulties in outlining consistent guidelines or measures to follow. Additionally, AI developers typically have backgrounds in computer science with limited training in ethics. The relative unfamiliarity with ethical principles and their implications could add further barriers to ethical medical AI development, potentially leading to ethically flawed AI products that could impose unintended harm to patients [34]. Thus, multisite collaborations, interdisciplinary communication, and IRB guidance and best standards may help to reduce the burden on AI developers, create more teachable moments, and establish more thoughtful and intentional mechanisms for deliberating ethical encounters, along with clear resolution pathways to facilitate implementation and confirmation.

Bioethics-Informed Guidance

The inclusion of bioethicists on research teams, as stated by participants, has been suggested in prior theoretical scholarship. For example, McLennan et al [36] proposed the concept of "embedded ethics," a collaborative approach that creates interdisciplinary research teams whereby AI developers and ethicists can anticipate, identify, and address ethical issues as they arise in the development process. Other studies have suggested a practical ethics checklist for AI developers [37] that recognizes ethical and social responsibility within AI development [38] or ethics guidelines and review processes specifically designed for AI developers [39,40] to support research design and analysis. These efforts offer improvements by having a refined focus on the practicality of how to use ethics recommendations and an emphasis on frontline AI developers, who can help mitigate ethical issues prior to AI deployment and use. This study's findings demonstrate that AI developers are interested in gaining knowledge about AI ethics, are already deliberating on the ethical encounters in their design and development, and are thoughtful about the longer-term practical implications of their work in health systems. Future research ought to consider strategies to mitigate ethical encounters and to advocate for heightened ethics knowledge, training, and conversations specifically targeted for AI developers and researchers. Foundational seminars on how to identify and label an issue as ethical (as opposed to technical) are a critical first step in training to ensure all developers and researchers can recognize these encounters in practice. Supervisors and managers must consider ways to encourage ethical dialogue and empower students and faculty to seek ways to mitigate ethical concerns and bridge their work to practice. This may involve bringing in bioethicists or other ethics experts who can speak diligently, thoughtfully, and comprehensively about these topics. AI developers and researchers should not be working in siloes but rather placed in communities with other medical AI stakeholders to heighten ethical dialogue and theorize novel

XSL•FO
RenderX

mitigation strategies. Additionally, to enhance the actionability of these recommendations, institutions could develop sample ethics checklists (eg, addressing data representativeness, model explainability, and patient privacy), workflow templates that map ethical review points within the AI development process, and metrics to evaluate ongoing compliance. Such practical tools can help translate ethical principles into consistent, operational practices for research offices and AI teams. These steps may facilitate the diffusion of innovation processes to allow for an easier and more transparent decision, implementation, and confirmation process that can lead to ethical adoption of medical AI. It is clear that ethical conflict and encounters of ethical dilemmas in the development and deployment of medical AI have stark impacts on the diffusion of innovation and the ability to effectively implement and reinforce the decision to adopt. Future research may seek to understand how this process may infiltrate the decision-making and ethical attitudes and perspectives of other key stakeholders, including physicians, allied health workers, health care administrators, patients, and families.

Limitations

This is one of the first studies in North America to examine AI developers' knowledge, encounters, and recommendations of medical AI ethics. Yet, there are several limitations. The small and relatively homogenous sample limits the diversity and generalizability of the findings. Future research should pursue broader and more inclusive investigations that capture perspectives from a wider range of disciplines, institutions, and demographic backgrounds across the United States. The representation of only 5 academic institutions may narrow the findings and may overlook other ethical concerns that emerge in distinct research areas. This was also a qualitative focus group study, and, thus, participants may have had concerns regarding privacy and confidentiality, reputation, and status in responses, and the emergence of potential power imbalances with student participants. Furthermore, as with most focus group studies, participants may have provided more socially desirable responses due to the group setting or the presence of peers, which could have influenced the depth or candor of some discussions. The focus group facilitator mitigated any ethical concerns by setting group norms, ensuring privacy and confidentiality, and piloting focus group questions and prompts. Future research may consider an anonymous survey to broadly examine ethical encounters in medical AI research.

Conclusions

As an exploratory pilot study, the current findings provide preliminary insights that can guide future empirical and institutional efforts. Findings from this study are important to determine the next steps to facilitate ethical decision-making among medical AI developers and researchers. There ought to be strategies to effectively deliberate about AI ethics across research teams and create opportunities for multisite collaboration, IRB debriefs and guidelines, protocol checklists and testing mechanisms, and the involvement of key stakeholders in deliberation, including bioethicists, clinicians, patients, and hospital leadership or administration with AI research teams. These initial insights lay the groundwork for larger-scale, multi-institutional investigations that can further validate and expand on the patterns identified here. The perspectives of key stakeholders may inform stages in the innovation-decision process and gain insight into barriers, supports, and resources necessary to ethically adopt medical AI into practice.

Funding

This work was supported by the National Institute of Health, United States (grant number: 3U01AG070112-02S2) as well as the Texas A&M Institute of Data Sciences.

Data Availability

The data that support the findings of this study are not publicly available due to privacy and confidentiality agreements.

Conflicts of Interest

None declared.

Multimedia Appendix 1
COREQ Checklist.
[PDF File (Adobe PDF File), 423 KB - jmir_v28i1e79613_app1.pdf]

References

1. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. In: Artificial Intelligence in Healthcare. California: Academic Press; 2020:25-60.
2. Chew HSJ, Achananuparp P. Perceptions and needs of artificial intelligence in health care to increase adoption: scoping review. J Med Internet Res 2022;24(1):e32939 [FREE Full text] [doi: 10.2196/32939] [Medline: 35029538]
3. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]

XSL•FO
RenderX

4. Afshar M, Adelaine S, Resnik F, Mundt MP, Long J, Leaf M. Deployment of real-time natural language processing and deep learning clinical decision support in the electronic health record: pipeline implementation for an opioid misuse screener in hospitalized adults. JMIR Med Inform 2023;11:e44977 [FREE Full text] [doi: 10.2196/44977] [Medline: 37079367]
5. Sapci AH, Sapci HA. Innovative assisted living tools, remote monitoring technologies, artificial intelligence-driven solutions, and robotic systems for aging societies: systematic review. JMIR Aging 2019;2(2):e15429 [FREE Full text] [doi: 10.2196/15429] [Medline: 31782740]
6. Chen Y, Stavropoulou C, Narasinkan R, Baker A, Scarbrough H. Professionals' responses to the introduction of AI innovations in radiology and their implications for future adoption: a qualitative study. BMC Health Serv Res 2021;21(1):813 [FREE Full text] [doi: 10.1186/s12913-021-06861-y] [Medline: 34389014]
7. Wangmo T, Lipps M, Kressig RW, Ienca M. Ethical concerns with the use of intelligent assistive technology: findings from a qualitative study with professional stakeholders. BMC Med Ethics 2019;20(1):98 [FREE Full text] [doi: 10.1186/s12910-019-0437-z] [Medline: 31856798]
8. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021;27(12):2176-2182 [FREE Full text] [doi: 10.1038/s41591-021-01595-0] [Medline: 34893776]
9. Hendrix N, Hauber B, Lee CI, Bansal A, Veenstra DL. Artificial intelligence in breast cancer screening: primary care provider preferences. J Am Med Inform Assoc 2021;28(6):1117-1124 [FREE Full text] [doi: 10.1093/jamia/ocaa292] [Medline: 33367670]
10. Miller DD, Brown EW. Artificial intelligence in medical practice: the question to the answer? Am J Med 2018;131(2):129-133. [doi: 10.1016/j.amjmed.2017.10.035] [Medline: 29126825]
11. Keskinbora KH. Medical ethics considerations on artificial intelligence. J Clin Neurosci 2019;64:277-282. [doi: 10.1016/j.jocn.2019.03.001] [Medline: 30878282]
12. Tang L, Li J, Fantus S. Medical artificial intelligence ethics: a systematic review of empirical studies. Digit Health 2023;9 [FREE Full text] [doi: 10.1177/20552076231186064] [Medline: 37434728]
13. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 2020;3:118 [FREE Full text] [doi: 10.1038/s41746-020-00324-0] [Medline: 32984550]
14. Nichol AA, Sankar PL, Halley MC, Federico CA, Cho MK. Developer perspectives on potential harms of machine learning predictive analytics in health care: qualitative analysis. J Med Internet Res 2023;25:e47609 [FREE Full text] [doi: 10.2196/47609] [Medline: 37971798]
15. Čartolovni A, Malešević A, Poslon L. Critical analysis of the AI impact on the patient-physician relationship: a multi-stakeholder qualitative study. Digit Health 2023;9:20552076231220833 [FREE Full text] [doi: 10.1177/20552076231220833] [Medline: 38130798]
16. Bergquist M, Rolandsson B, Gryska E, Laesser M, Hoefling N, Heckemann R. Trust and stakeholder perspectives on the implementation of AI tools in clinical radiology. Eur Radiol 2024;34(1):338-347 [FREE Full text] [doi: 10.1007/s00330-023-09967-5] [Medline: 37505245]
17. Blease C, Kaptchuk TJ, Bernstein MH, Mandl KD, Halamka JD, DesRoches CM. Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views. J Med Internet Res 2019;21(3):e12802 [FREE Full text] [doi: 10.2196/12802] [Medline: 30892270]
18. Bourla A, Ferreri F, Ogorzelec L, Peretti C, Guinchard C, Mouchabac S. Psychiatrists' attitudes toward disruptive new technologies: mixed-methods study. JMIR Ment Health 2018;5(4):e10240 [FREE Full text] [doi: 10.2196/10240] [Medline: 30552086]
19. Ongena YP, Haan M, Yakar D, Kwee TC. Patients' views on the implementation of artificial intelligence in radiology: development and validation of a standardized questionnaire. Eur Radiol 2020;30(2):1033-1040 [FREE Full text] [doi: 10.1007/s00330-019-06486-0] [Medline: 31705254]
20. Witkowski K, Dougherty RB, Neely SR. Public perceptions of artificial intelligence in healthcare: ethical concerns and opportunities for patient-centered care. BMC Med Ethics 2024;25(1):74 [FREE Full text] [doi: 10.1186/s12910-024-01066-4] [Medline: 38909180]
21. Akinrinola O, Okoye CC, Ofodile OC, Ugochukwu CE. Navigating and reviewing ethical dilemmas in AI development: strategies for transparency, fairness, and accountability. GSC Adv Res Rev 2024;18(3):050-058. [doi: 10.30574/gscarr.2024.18.3.0088]
22. Drabiak K, Kyzer S, Nemov V, El Naqa I. AI and machine learning ethics, law, diversity, and global impact. Br J Radiol 2023;96(1150). [doi: 10.1259/bjr.20220934] [Medline: 37191072]
23. Mathrani A, Susnjak T, Ramaswami G, Barczak A. Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics. Comput Educ Open 2021;2. [doi: 10.1016/j.caeo.2021.100060]
24. Morley J, Kinsey L, Elhalal A, Garcia F, Ziosi M, Floridi L. Operationalising AI ethics: barriers, enablers and next steps. AI Soc 2021;38(1):411-423. [doi: 10.1007/s00146-021-01308-8]
25. Sanderson C, Douglas D, Lu Q, Schleiger E, Whittle J, Lacey J. AI ethics principles in practice: perspectives of designers and developers. IEEE Trans Technol Soc 2023;4(2):171-187. [doi: 10.1109/tts.2023.3257303]

XSL•FO
RenderX

26. Griffin TA, Green BP, Welie JVM. The ethical agency of AI developers. AI Ethics 2023;4(2):179-188. [doi: 10.1007/s43681-022-00256-3]
27. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: implications for clinical decision-making. PLOS Digit Health 2024;3(11). [doi: 10.1371/journal.pdig.0000651] [Medline: 39509461]
28. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
29. Rogers EM. Diffusion of Innovation. 4th Edition. New York: Free Press; 1995.
30. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: a call for open science. Patterns (N Y) 2021;2(10):100347 [FREE Full text] [doi: 10.1016/j.patter.2021.100347] [Medline: 34693373]
31. Drukker K, Chen W, Gichoya J, Gruszauskas N, Kalpathy-Cramer J, Koyejo S. Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. J Med Imaging (Bellingham) 2023;10(6):061104 [FREE Full text] [doi: 10.1117/1.JMI.10.6.061104] [Medline: 37125409]
32. Chinta SV, Wang Z, Palikhe A, Zhang X, Kashif A, Smith MA. AI-driven healthcare: fairness in AI healthcare: a survey. PLOS Digit Health 2025;4(5):e0000864. [doi: 10.1371/journal.pdig.0000864] [Medline: 40392801]
33. Griffin TA, Green BP, Welie JV. The ethical wisdom of AI developers. AI Ethics 2024;5(2):1087-1097. [doi: 10.1007/s43681-024-00458-x]
34. Kim JP, Ryan K, Kasun M, Hogg J, Dunn LB, Roberts LW. Physicians' and machine learning researchers' perspectives on ethical issues in the early development of clinical machine learning tools: qualitative interview study. JMIR AI 2023;2:e47449 [FREE Full text] [doi: 10.2196/47449] [Medline: 38875536]
35. Mittelstadt B. Principles alone cannot guarantee ethical AI. Nat Mach Intell 2019;1(11):501-507. [doi: 10.1038/s42256-019-0114-4]
36. McLennan S, Fiske A, Tigard D, Müller R, Haddadin S, Buyx A. Embedded ethics: a proposal for integrating ethics into the development of medical AI. BMC Med Ethics 2022;23(1):6 [FREE Full text] [doi: 10.1186/s12910-022-00746-3] [Medline: 35081955]
37. Morley J, Cowls J, Taddeo M, Floridi L. Ethical guidelines for COVID-19 tracing apps. Nature 2020;582(7810):29-31. [doi: 10.1038/d41586-020-01578-0] [Medline: 32467596]
38. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1(5):206-215 [FREE Full text] [doi: 10.1038/s42256-019-0048-x] [Medline: 35603010]
39. Ryan M, Stahl BC. Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. J Inf Commun Ethics Soc 2020;19(1):61-86. [doi: 10.1108/jices-12-2019-0138]
40. Srikumar M, Finlay R, Abuhamad G, Ashurst C, Campbell R, Campbell-Ratcliffe E. Advancing ethics review practices in AI research. Nat Mach Intell 2022;4(12):1061-1064. [doi: 10.1038/s42256-022-00585-2]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
IRB: institutional review board

Edited by A Stone; submitted 24.Jun.2025; peer-reviewed by I Adefolaju, X Liang, LP Gorrepati; comments to author 06.Aug.2025; revised version received 15.Dec.2025; accepted 29.Dec.2025; published 28.Jan.2026.

Please cite as:
Fantus S, Li J, Wang T, Tang L
Ethical Knowledge, Challenges, and Institutional Strategies Among Medical AI Developers and Researchers: Focus Group Study
J Med Internet Res 2026;28:e79613
URL: https://www.jmir.org/2026/1/e79613
doi:10.2196/79613
PMID:

©Sophia Fantus, Jinxu Li, Tianci Wang, Lu Tang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Educating Students About Digital Health Research Ethics: Curricula Review and Expert Interview Study

Yier Zhu[1], MPH; Brian McInnis[1], PhD; Tanya Punater[1], BS; Brittany York[1], MPH; Camille Nebeker[1,2,3], EdD, MS

[1]Herbert Wertheim School of Public Health and Human Longevity Science, University of California San Diego, La Jolla, CA, United States

[2]The Design Lab, University of California San Diego, La Jolla, CA, United States

[3]Qualcomm Institute, University of California San Diego, La Jolla, CA, United States

Corresponding Author:
Camille Nebeker, EdD, MS
Herbert Wertheim School of Public Health and Human Longevity Science
University of California San Diego
9500 Gilman Drive, MC 0811
La Jolla, CA, 92093
United States
Phone: 1 619 518 2875
Email: nebeker@health.ucsd.edu

Abstract

Background: The rapid growth of digital health research, involving wearable devices, mobile apps, and sociotechnical health systems, raises complex ethical, legal, and social considerations. While institutional review boards and research ethics frameworks address some concerns, less is known about how students and trainees in digital health are systematically educated to recognize and navigate these challenges. Understanding the scope and content of ethics training is critical to ensuring the responsible development and application of digital health technologies.

Objective: This study investigated how college students are trained to identify and address ethical considerations in digital health research through an analysis of formal curricula and expert perspectives.

Methods: Researchers reviewed 132 syllabi from 76 academic programs across 62 universities and conducted semistructured interviews with 6 leading digital health scholars. All syllabi were coded for instructional content and learning objectives. Researchers conducted open coding and collaboratively applied affinity diagramming to organize the data into hierarchical themes.

Results: All syllabi included instructional content, and most included explicit learning objectives. Analysis identified 7 key themes, which captured both explicit knowledge imparted through formal instruction and tacit knowledge cultivated through laboratory work, mentorship, and applied experiences. Findings highlighted gaps between formal ethics instruction and the realities of research practice.

Conclusions: Ethics education in digital health research develops through the interplay of formal coursework and practice-based training, each fostering complementary skills needed for data-intensive and collaborative environments. Together, these pathways support students in identifying ethical issues, applying principles contextually, anticipating emerging risks, and communicating across disciplines; however, access to experiential learning opportunities remains inconsistent. Strengthening ethics training will require expanding structured early research engagement, cultivating communities of practice, and translating tacit ethical reasoning into accessible resources. Integrating ethical reflection into routine research activities may better prepare future digital health researchers to responsibly design and govern sociotechnical health systems.

(J Med Internet Res 2026;28:e82861) doi:10.2196/82861

KEYWORDS

digital health research; ethics education; practice-based learning; qualitative research; research ethics; research training; sociotechnical health systems

XSL•FO
RenderX

Introduction

As digital technologies become integral to health research and clinical care, students must be prepared to examine related ethical, social, and organizational implications. Digital health tools, from mobile apps and wearables to artificial intelligence (AI) systems, can improve care; yet, their implementation often outpaces understanding of effects on patients, providers, and health systems [1,2]. Our research focuses on digital health research ethics as a sociotechnical concern: technologies both shape and are shaped by institutional contexts, stakeholder relationships, and data infrastructures to coordinate people, groups, and computing resources with a goal of improving patient health and well-being [3,4]. For example, a conversational agent located in a patient's room could be used to take notes during a clinical visit [5], notify nursing teams when the patient is experiencing increased anxiety [6], and converse with the patient to reduce their feelings of social isolation [7]. Digital health tools, like wearable sensors, offer new ways to monitor patients, both in and outside of the clinic, which have important ethical implications [8]. By adopting a sociotechnical lens, we attend to system components (ie, technologies, data practices, and institutional policies) and social interactions between team workflows and patient/participant engagement to examine emerging ethical issues, including unanticipated privacy and data management risks.

Academic education and training programs play a critical role in preparing digital health researchers and clinicians to navigate ethical challenges like algorithmic bias, data privacy, and access inequities, yet little empirical work has examined the availability and scope of such programs across organizations contributing to digital health research [6]. Unlike general research ethics, digital health research includes characteristics that can strain traditional academic ethics models. These differences include streams of continuous, granular individual-level data that may remain identifiable and blurred lines between research and industry data flows [9-12]. Importantly, informed consent communications are challenged by difficulties explaining passive sensing and other pervasive technologies, data management practices, and evolving algorithms [13-15].

Researchers learn to recognize and mitigate ethical challenges through university courses, workshops, and readings (explicit knowledge) as well as through firsthand opportunities to learn by training with expert research practitioners and working with clinicians, caregivers, and digital health technology designers in the field (tacit knowledge). As digital health ethical and social implications relate to the design and deployment of technology within social, cultural, and organizational contexts, it is critical for students to learn through both explicit coursework and experience-based learning (tacit knowledge). However, less is known about the specific components of this education among disciplines involved in the design and testing of digital health tools. Explicit training supplies common concepts and language; experiential learning develops judgment for context-specific decisions as teams design, deploy, and evaluate sociotechnical systems in health settings [10,11,16]. Because these sociotechnical properties scale through computing infrastructures, the stakes and failure modes in digital health differ from those typically addressed in general research ethics training.

To examine how students gain ethical competencies in digital health research, we analyzed university syllabi to characterize explicit knowledge conveyed through coursework and conducted interviews with leading digital health researchers to understand how tacit knowledge is conveyed through the conduct of research, which can include decisions about consent, data protection, stakeholder engagement, and dissemination [17,18]. Our goals for this study were to (1) evaluate how programs that contribute to digital health research socialize students to ethical and responsible research practices through coursework and (2) describe how expert researchers convey tacit ethical practices to trainees through hands-on training. We do not test hypotheses; rather, this exploratory, mixed evidence synthesis (syllabi and interviews) maps explicit and tacit learning opportunities in digital health research ethics.

Methods

Overview

This study used a multimethod qualitative design, integrating 2 complementary components to examine how ethical competencies in digital health research are taught and learned. We analyzed course syllabi to characterize explicit ethics instruction and conducted semistructured interviews with experienced digital health researchers to understand how tacit ethical knowledge is conveyed through research practice. We used the sociotechnical framework to integrate insights from syllabi (explicit knowledge) and interviews (tacit knowledge), examining how instructional content aligned with or diverged from the ethical challenges researchers encounter in sociotechnical research environments. These 2 data sources provide a comprehensive synthesis of explicit and experiential learning.

Ethical Considerations

The research activities involved minimal risk to participants, and the protocol was reviewed and verified as exempt under 45 CFR 46.104(d)2 by the institutional review board (IRB) at the University of California San Diego (exemption verification 210941). All interview participants received recruitment materials and provided verbal informed consent prior to audio recording. Participants did not receive compensation. Reporting followed the Standards for Reporting Qualitative Research (SRQR), identified by the EQUATOR (Enhancing the Quality and Transparency of Health Research) Network.

Audio recordings and transcripts were stored in University of California San Diego's secure Microsoft Teams environment, which provides encrypted, access-restricted cloud storage; only authorized study personnel with institutional credentials could access study files to ensure participant privacy and confidentiality.

Study Design

The study used 2 complementary components to examine how ethical competencies in digital health research are taught and learned. First, a syllabi analysis was conducted to characterize

XSL•FO
RenderX

explicit ethics instruction delivered through university coursework. Second, expert interviews were conducted to examine tacit ethics knowledge conveyed through hands-on research practice. Although designed and analyzed independently, the 2 components together highlight complementary dimensions of ethics education in sociotechnical digital health research. To identify academic programs and researchers relevant to digital health ethics training, this study drew upon a preliminary literature review of sociotechnical digital health research in computing and information science venues. This review informed the selection of academic units for the syllabi search as well as the identification of researchers eligible for the interview component.

Course Syllabi Collection—Explicit Knowledge

Institutional websites and departmental materials were examined to determine whether units associated with the publications included in the systematic literature review (forthcoming) offered formal or informal training opportunities focused on ethical and responsible research practices. These included required ethics coursework, elective offerings, workshops, journal clubs, and other structured learning activities relevant to sociotechnical systems and digital health research.

A total of 132 course syllabi were collected from academic units identified through the search strategy. Of these, 98 syllabi contained learning objectives and were therefore included in the learning objective analysis. The remaining syllabi were excluded because they lacked learning objectives or did not provide sufficient detail to determine ethics-related instructional content. The final analytic set represents 64 academic programs from universities in the United States and internationally.

When course information was not publicly available, research staff contacted instructors or administrative personnel to request syllabi, course descriptions, and information about relevant instructional activities. Syllabi served as the primary data source for examining explicit ethics-related knowledge conveyed through coursework.

Criteria for identifying ethics-related courses were derived from the syllabi analysis protocol, which specified how the research team determined whether a course contained substantial ethics content. Courses were classified as ethics-focused when the title, description, learning outcomes, or listed topics indicated a primary emphasis on ethical issues relevant to sociotechnical systems that contribute to digital health research. Courses in which ethics was addressed only briefly within a broader technical curriculum were not classified as ethics-focused. When syllabi were unavailable, these criteria were applied to publicly accessible catalog descriptions and course topic summaries.

To determine who to contact within each academic program, research staff reviewed departmental web pages, faculty directories, graduate program administrator listings, and instructional staff profiles to identify individuals most likely to have knowledge of course offerings. Depending on the structure of each department, emails were directed to course instructors, department chairs, program coordinators, graduate advisors, or general departmental accounts.

Beginning on August 18, 2023, and concluding on December 8, 2023, research staff sent up to 4 email communications to relevant instructors and administrative staff to solicit course information. The communication provided an overview of the study and requested (1) syllabi for targeted courses, (2) course descriptions or lists of required graduate courses, and (3) information about formal and informal training opportunities, such as workshops, seminars, journal clubs, or other instructional activities. Recipients were also invited to forward the request to colleagues who might have access to course materials or additional information.

Syllabi Analysis

After collecting the course syllabi, the content was analyzed to explore the breadth of learning objectives and course topics related to digital health research ethics. Learning objectives, also known as learning goals or outcomes, are statements that define what students should learn or be able to do after completing a course or activity. Course topics refer to the sequence of subject areas covered during a course, often presented along with a schedule for the course lectures and assignments or summarized in a course description. To analyze the collection of course syllabi, our study applied an approach similar to recent studies of courses in technology ethics [17] and social computing [18]. Specifically, 3 researchers reviewed each syllabus to identify learning objectives and course topics. Initial coding was conducted independently by multiple members of the research team, followed by regular consensus-based meetings to review interpretations, resolve discrepancies, and refine analytic decisions. To identify themes, the research team applied an affinity diagramming method [19] to qualitatively organize the learning objectives and course topics into hierarchical clusters. Members of the research team met twice per week for 6 weeks using FigJam (Figma Inc) [20] to support collaborative synthesis. The resulting themes characterize explicit ethics-related knowledge conveyed through coursework and are later interpreted alongside interview findings.

Interview Recruitment and Procedures—Tacit Knowledge

Semistructured interviews were conducted between August 17 and September 7, 2023. Recruitment involved outreach to a pool of 32 eligible individuals, with up to 3 contact attempts per person. Recruitment emails summarized the study purpose and the general topical domains from which participants could select areas they felt most prepared to discuss. Authors were eligible to participate if they were corresponding or lead authors on included publications and could speak to training, ethical decision-making, and research team practices in digital health research. A total of 6 individuals consented to participate and completed the interview. All interviews were analyzed. Together, the 6 interviewees had 516 publications and 85,252 citations, reflecting their standing in the field.

Eligible researchers received an email invitation describing the study purpose, interview procedures, and plans to audio-record the session. Interviews were conducted via Zoom (Zoom Communications) by a trained graduate student researcher, with a supervising team member present to support consistency in

XSL•FO
RenderX

procedures. Verbal informed consent was obtained at the beginning of each interview. All interviews were audio-recorded with participant consent and transcribed using Zoom's built-in transcription feature.

The semistructured interview guide was informed by our prior literature review and publication-specific analyses of sociotechnical health research practices (forthcoming). Questions were organized around the following five themes identified in our prior analyses: (1) prototyping designs, including how research teams collaboratively prototype systems, protocols, and materials; (2) engaging stakeholders, including how stakeholders are involved throughout the research life cycle; (3) developing informed consent procedures and data protections, including how teams design consent processes and data safeguards; (4) mentoring research, including the role of mentorship in training and guiding research practices; and (5) sharing research, including how research teams disseminate findings and return results to participants and stakeholders. The guide included open-ended prompts within each theme to elicit concrete examples from participants' experiences. Experts were invited to select two to three themes they felt most prepared to discuss.

Interview Transcripts Analysis

An iterative, inductive thematic approach was used to analyze interview transcripts. Although the analysis was inductive, interpretation was informed by a digital health research ethics perspective, which sensitized the team to ethical considerations arising from the design and use of engineering-based digital health systems within research and institutional contexts. Following each interview, researchers drafted analytic memos documenting preliminary impressions, emerging concepts, and notable patterns [21]. These memos and associated transcript excerpts were reviewed by supervising researchers to ensure accuracy, deepen interpretation, and maintain analytic rigor.

Building on this initial memoing process, 1 researcher conducted detailed inductive coding to identify recurring concepts, and a second researcher independently reviewed these summaries and coded transcript segments to refine and validate emerging interpretations. The student interviewers and supervising researchers met regularly to discuss analytic decisions, resolve discrepancies through consensus, and iteratively refine the thematic structure; four researchers contributed to the overall analysis. Because the full set of interviews represented a small sample of senior digital health researchers, the team analyzed all available data and refined themes until shared conceptual coherence was reached through regular consensus meetings. Topics from the interview guide are summarized with example prompts in Textbox 1.

Textbox 1. Interview focus areas and example prompts.

Ethics-related research practices discussed during interviews

- Prototyping technology for health research
- Engaging with community stakeholders
- Developing informed consent procedures and data protection policies
- Sharing research data and findings

Reflective prompts about field development and ethics priorities

- How do you find and connect with other researchers in your field?
- Where do you look for funding for your research?
- How do you see this research field evolving over the next few years?
- What do you see as key research ethics topics to include in training programs?
- Who sets the guidelines and standards for ethics training in your field?

Results

Overview

Across our analysis of 98 course syllabi and 6 expert interviews, we identified seven themes that characterize how ethical competencies in digital health research are taught and learned: (1) broad understanding of ethical principles, (2) applied ethics, (3) research in context, (4) ethics in technology design, (5) anticipating the future, (6) communication practice, and (7) professional responsibility. Some conceptual overlap across themes occurs across communication, team collaboration, and professional responsibility. This is expected given the sociotechnical context of digital health research. Ethical competence is developed and demonstrated across interpersonal interaction, team workflows, and organizational norms, as these are interdependent domains. The sections that follow describe the programs along with each theme and illustrate how explicit and tacit learning opportunities manifest across programs and research environments.

Academic Programs

The academic programs identified through the search reflect a breadth of disciplines. A majority (n=48, 77%) of the universities represented by the syllabus collection are located within the United States, with 14 (22%) located internationally. The 132 courses include graduate-level instruction for doctoral and medical school students as well as students in a range of

XSL•FO
RenderX

master's degree programs (Table 1). Courses were open to undergraduate and graduate students (n=24, 18%), with the majority (n=64, 48%) limited to undergraduate enrollment.

Table 1. Total courses by academic discipline based on the name of the university department.

Academic discipline	Count, n
Bioengineering	2
Biology	2
Cognitive science	2
Communication	2
Computer science	62
Digital media	8
Education	1
Engineering	10
Global health	1
Human-centered computing	3
Industrial design	1
Informatics	18
Information science	11
Innovation	1
Nursing	1
Pediatrics	1
Philosophy	1
Public policy	1
Sociology	1
Total	132

Sample Size

We analyzed 98 syllabi that included learning objectives to identify explicit ethics-related learning opportunities and identified 7 themes, as summarized in Table 2. Within the 98 syllabi analyzed, 5 universities provided instruction for 5 of the 7 themes, whereas 12 universities incorporated only 1 (19%) theme. None of the academic programs offered a course series that encompassed all 7 learning-objective themes. The following sections describe each theme. Additional examples of learning objectives and theme-specific references are provided in Multimedia Appendix 1.

XSL•FO
RenderX

Table 2. Descriptions of themes with the number of learning objectives and courses represented in each theme.

Theme descriptions	Objectives, n	Courses (n=98), n (%)
Theme 1: broad understanding. Gain a broad understanding of ethical principles and guidelines [22-25]		
Learn to apply ethical principles in research, in general	20	18 (18.4)
Explore how ethical principles have been applied in different contexts	33	18 (18.4)
Summary	53	31 (31.6)
Theme 2: applied ethics. Practice applying ethical principles to real-world cases and scenarios [26-29]		
Identify health informatics issues from various perspectives	12	4 (4.1)
Identify ethical challenges related to technology	42	27 (27.6)
Identify ethical considerations related to privacy and security	17	13 (13.3)
Learn about ethical practices in data science	28	19 (19.4)
Summary	99	43 (43.9)
Theme 3: research in context. Build research skills necessary to investigate ethical questions within specific context [30-33]		
Explain uses of AI[a]	19	7 (7.1)
Learn about data science methods	23	13 (13.3)
Describe security and privacy processes	37	13 (13.3)
Develop research and analysis skills	33	16 (16.3)
Summary	112	42 (42.9)
Theme 4: ethics in technology design. Learn how ethical principles can play into decisions about human-computer interaction [34-37]		
Develop foundational skills in the design process	28	16 (16.3)
Gain skills in human-computer interaction design	18	6 (6.1)
Learn about ethical design principles	19	7 (7.1)
Learn about the design of digital health systems	16	5 (5.1)
Summary	81	27 (27.6)
Theme 5: anticipating the future. Learn how to identify and assess ethical issues tied to emerging and futuristic technologies [38-46]		
Explore diverse perspectives in technology	15	11 (11.2)
Explore the societal impact of health technology	14	7 (7.1)
Explore the societal impact of novel and future technology	24	21 (21.4)
Explore ELSIs[b] of technology	36	16 (16.3)
Introduced to an overview of computing research	33	19 (19.4)
Gain basic knowledge in business	1	1 (1)
Summary	123	50 (51)
Theme 6: communication practice. Develop skills to communicate ethical concerns in group discussions, presentations, and written reports [47-51]		
Develop professional communication skills	26	18 (18.4)
Learn about engaging with diverse audiences	22	18 (18.4)
Developing critical thinking and reasoning skills	17	13 (13.3)
Summary	65	42 (42.9)
Theme 7: professional responsibility. Understand the ethical and professional standards expected of researchers conducting research in the public interest [52-56]		
Describe ethical and professional responsibilities in computing	17	14 (14.3)
Learn about professional ethics in workplaces	11	7 (7.1)
Develop teamwork and team management skills	13	10 (10.2)
Summary	41	28 (29)

XSL•FO
RenderX

[a]AI: artificial intelligence.

[b]ELSI: ethical, legal, and social implications.

Theme 1: Broad Understanding. Principles and Guidelines (eg, Bioethics and Human Rights)

This theme captures how courses orient students to ethical traditions (bioethics, research ethics, and rights-based perspectives) and institutional guardrails (eg, IRB training). The emphasis is on mastering shared vocabulary and analytic lenses that later anchor applied judgments in digital health contexts.

Our analysis found that a third of all courses (31/98) include learning objectives that emphasize training in diverse philosophical approaches to ethical thinking (Table 2, Theme 1). As summarized in Table 2 (Theme 1), foundational objectives appear across a substantial subset of courses but vary in depth from survey-style introductions to more rigorous treatments that require analytic writing and IRB certification. Programs rarely sequence these foundations with later, hands-on components, creating discontinuities that reappear in downstream themes. One course articulates this by stating that students will "develop a working understanding of major traditions of philosophy and ethical theory to provide foundational tools of analysis." (C108)

Foundational principles serve as a grounding reference point for decisions about informed consent, risk assessment, and mitigation, and expectations for fairness as digital health systems are implemented in new settings. Differences in depth point to gaps in scaffolding from foundational principles to applied and design contexts. These foundational principles matter because they establish the ethical vocabulary and conceptual grounding students later rely upon when making context-specific decisions in data-intensive health environments. Without this grounding, students may struggle to recognize how abstract principles translate into concrete ethical challenges that emerge in sociotechnical systems.

Theme 2: Applied Ethics. Practice Applying Ethical Principles to Real-World Sociotechnical Contexts

Courses in this theme move beyond abstraction and ask students to apply ethical principles to concrete cases involving algorithmic decision-making, data privacy and security, and responsible AI deployment in health contexts. Applied ethics is positioned as a bridge between frameworks and situated decisions made during design, analysis, and deployment, with related learning objectives appearing in 43 (44%) of 98 courses. These courses emphasized students' ability to analyze ethical challenges (ie, information storage, exchange, security, and privacy) across various contexts, such as health informatics, data science, and computer science (Table 2, Theme 2). For example, "Apply the philosophies and theories covered to computer science problems and scenarios, including risks associated with large language models and algorithmic decision-making" (C55). As shown in Table 2 (Theme 2), applied ethics objectives are prevalent yet unevenly embedded. Some courses integrate iterative casework and debate across the term; others contain a single module within broader technical instruction. Assignments commonly require students to reason through value trade-offs across stakeholders and technical constraints, making ethical salience visible at practical decision points.

Applied ethics training equips students to recognize and navigate tensions where harms materialize, such as in data curation, model evaluation, and deployment pipelines. This applied focus is essential in digital health because ethical tensions often arise at the level of implementation, such as privacy trade-offs in data pipelines or fairness issues in algorithmic modeling. These courses help students identify where harms may materialize and how to reason through competing stakeholder values, trade-offs in data pipelines, or fairness issues in algorithmic modeling.

Theme 3: Research in Context. Methods to Study Sociotechnical Interactions

This theme reflects the pairing of technical content (AI/machine learning, security, and data science) with research design and analysis (literature reviews, proposal writing, and qualitative/quantitative reasoning) so students can empirically study sociotechnical systems in health. The focus is not coding per se, but the methodological capacity to investigate ethical questions in context. Results reveal 43% (42/98) of our course syllabi (Table 2, Theme 3) depict broad coverage of research skills objectives. We found that some syllabi emphasize technical fluency, while others center on social scientific inquiry. Few programs explicitly connect these skill sets to institutional realities (eg, governance and clinical workflows), a linkage that becomes crucial in later themes. As an example, C66 states, "This course covers the fundamental principles of the scientific method, and the core skills of planning, designing, executing, evaluating, and presenting research."

Blending computing and social science methods reveals how risks emerge from interactions across data, tools, institutions, and people; however, explicit integration of where risks and benefits can accrue across clinical and organizational constraints appears inconsistent. Syllabi linked to this theme aim to develop competence by enabling students to investigate how risks, biases, and unintended consequences occur, reflecting the sociotechnical nature of digital health systems. Understanding context is critical for anticipating where ethical safeguards may fail.

Theme 4: Ethics in Technology Design. Learn How Ethical Principles Can Play Into Decisions About Human-Computer Interaction

Courses link values (accessibility, equity, and sustainability) to concrete design choices for digital health systems. Here, ethics is framed as an up-front design requirement, which may include incorporating inclusive methods, accessibility standards, and sensitivity to domain-specific issues such as digital biomarkers or health disparities, rather than a post hoc audit. One effective way to bridge the gap between ethical principles in theory and the practice of system design is through the incorporation of human-centered design (HCD) and human-computer interaction

XSL•FO
RenderX

concepts. This includes examining "how digital health technologies can be designed to achieve health equity [with a] central focus on applying human-centered design concepts to the domain of digital health equity" (C19). Our analysis found that 27% (n=26) of the courses teach students HCD processes (Table 2, Theme 4). Unlike Theme 3, courses in Theme 4 prioritize understanding system design concepts, though systems are also often incorporated into specific research methods.

Introducing criteria into design processes that increase awareness of ethical practices can help teams to operationalize what respect for persons and justice look like in the research design phase. Programs that connect HCD/human-computer interaction to clinical governance appear better positioned to align design intent with downstream demonstration of ethical practices.

Explicit and Tacit Knowledge About Research Ethics Through Curriculum and Hands-on Training

As described in the Methods section, digital health research experts, selected based on their prior citation frequency and prominence within the digital health research community, were interviewed to gather their perspectives on the digital health field and to speculate about the ethics training students will need in the future. Quotes from the interviews are uniquely identified by code, P1-P6, with each code corresponding to a specific expert. Because the interview data aligned closely with Themes 5-7, the following theme descriptions integrate evidence from both course syllabi and experts' perspectives.

Theme 5: Anticipating the Future. Learn How to Identify and Assess Ethical Issues Tied to Emerging and Futuristic Technologies

Courses in Theme 5 ask students to anticipate ethical, legal, and social trade-offs of emerging or speculative technologies before systems are deployed, with attention to governance, unintended consequences, and positionality. As summarized in Table 2 (Theme 5) , this theme appears in 50 (51%) of 98 syllabi, making it the most prevalent in the corpus; yet, programs vary in how anticipation is taught. Some emphasize legal and policy analysis, others use scenario debates and stakeholder role-play to surface value trade-offs. One course, for example, requires students to "propose recommendations...that represent the perspectives of stakeholders to develop a resolution." (C2).

Interview accounts show how this anticipatory work becomes a tacit judgment in practice. Researchers described teaching students to reason through downstream risks under uncertainty , particularly around data practices that are easy to overlook but consequential for participants. For instance, with passive sensing of GPS data, students are asked to grapple with consent memory and continuous collection: "it gets turned on, and it runs...continues to pull those data until the person uninstalls the app.... How do you manage that?" (P2). Teams use these moments to clarify what to do when participants later realize some data feel too sensitive to be shared, and to revisit data sharing procedures so that participants are aware of protections throughout a study.

Senior researchers also coach students to communicate the salience of risk when peers view data as "just data." As one expert noted, "Students might ask, 'This is just data; why are you making it so hard to access?' I bring examples of how location data alone can be very compromising or reveal intimate parts of someone's life.... It can have serious consequences if used beyond consent." (P6). Another expert emphasized that even deidentified datasets can create reidentification risk when clinical features are unique, requiring thoughtful protection while retaining research value (P4). In parallel, experts voiced concern that AI/large language model adoption in health research is moving faster than frameworks and training, urging explicit conversations about safe, responsible use. As one explained, "we should...figure out ways to talk about how to use these technologies in a safe, ethical, and responsible way." (P6).

Anticipatory ethics is particularly important in digital health, where emerging technologies outpace regulation and policy. Learning objectives that aim to prepare students to spot plausible harms and governance gaps before deployment, but curricula differ in modality and depth. Interview evidence shows how mentors transform anticipation into on-the-ground judgment, teaching students to (1) notice subtle data risks (eg, passive GPS), (2) communicate why those risks matter to participant welfare, and (3) revisit data management choices as technologies and contexts evolve.

Theme 6: Communication Practice. Develop Skills to Communicate Ethical Concerns in Group Discussions, Presentations, and Written Reports

Theme 6 centers on the individual's capacity to surface and reason about ethical issues through argumentation, critique, writing, and public speaking. As shown in Table 2 (Theme 6), communication-focused learning objectives appear in 42 (43%) of the 98 syllabi, but their sophistication varies. This spans from basic presentation mechanics to rigorous argumentation that engages counterpositions with evidence and adapts messages to different audiences.

Many courses build these competencies by having students read extensively, write analytically, and lead structured discussions. One syllabus puts it plainly: students should "read a lot, write a fair amount, and become more comfortable participating in oral discussions and giving an oral presentation." (C49). Another emphasizes the standard to which this participation is held: "articulate your ethical beliefs precisely and persuasively." (C116). Together, these objectives position rhetoric as the mechanism by which ethical reasoning becomes public, contestable, and actionable in research settings.

Communication is foundational in ethical practice. Ethical issues in digital health typically surface through dialogue ; these syllabi describe tools to practice talking about ethical challenges. The interview narratives show those tools at work explaining risk, negotiating disagreement, and recording decisions so teams remain accountable to participants and stakeholders.

Theme 7: Professional Responsibility. Understand the Ethical and Professional Standards Expected of Researchers

This theme focuses on organization-level norms, including professional codes, role clarity, escalation paths for ethical

XSL•FO
RenderX

concerns, cross-disciplinary collaboration, and stewardship responsibilities for managing volumes of individual-level sensitive data. As indicated in Table 2 (Theme 7), such objectives appear in 28 (29%) of 98 syllabi, less frequently than communication or anticipation themes. Syllabi in this theme ask students to analyze professional association policies and to consider how responsibilities translate into daily practice in research teams and health organizations; for example, "What special responsibilities do we have as computing professionals? How can we apply the Software Engineering Code of Ethics and ACM Code of Ethics in our daily practice?" (C21).

Tacit practices, described by experts, make the organizational mechanics of responsibility concrete. Senior researchers emphasized modeling norms and setting expectations that are then diffused through laboratories and professional communities. For example, "we have to show that this is the way we do the work." (P5). Experts also highlighted the need to work across disciplines and with stakeholders (eg, clinicians, patients, and community partners) so that design decisions reflect health system realities and community values; collaboration remains essential but complex, requiring time, care, and clear alignment with funder constraints (P4 and P1).

Experts train laboratory members to conduct stakeholder-engaged digital health research, which includes building relationships, listening across perspectives, and giving back to communities. These are competencies that extend professional responsibility beyond the laboratory into the settings where digital health systems are designed and implemented (P4 and P5).

Professional responsibility is enacted through governance structures and norms (codes, roles, and processes) and through collaboration with diverse stakeholders. Courses introduce these expectations; interviews show how senior researchers make governance visible in day-to-day work, translating ethical intent into accountable team practices.

Discussion

Principal Findings

Digital health research operates within complex sociotechnical systems [57], making ethical competence foundational to responsible innovation. This study examined how research ethics education prepares trainees for the sociotechnical realities of digital health research. Findings indicate that ethical competence develops through the interaction of classroom instruction and experiential learning; yet, gaps remain in translating abstract principles into research practice. Addressing these gaps will require instructional strategies that more intentionally integrate ethics training with the collaborative, data-intensive environments in which digital health research occurs.

Research ethics education in digital health relies on both explicit instruction and tacit learning. Classroom teaching introduces guiding ethical principles, analytic tools, and case studies [58], while participation on research teams enables students to identify, discuss, and resolve ethical issues as they arise [59]. Experts in our study described mentoring trainees through decisions affecting participant data privacy, allowing students to observe how choices related to data storage, sharing, and stewardship unfold across a study. These complementary experiences reinforce the reciprocal relationship between teaching and research practice.

Our thematic results illustrate how ethical competence in digital health research emerges through sociotechnical interactions rather than individual reasoning alone. For example, the "applied ethics" and "research in context" themes show how ethical issues arise within data pipelines, modeling practices, and clinical workflows, where technical and organizational factors intersect. Themes such as "communication practice" and "professional responsibility" highlight how interpersonal coordination, role clarity, and governance structures shape ethical action within research teams. Our results confirm that ethics education must prepare trainees to navigate these interconnected sociotechnical dynamics rather than isolated ethical dilemmas.

Ethics Education as a Sociotechnical Process

Our findings reveal a persistent sociotechnical tension in ethics education. Coursework often emphasizes individual reasoning and abstract principles, whereas ethical challenges in digital health research are collective and infrastructural, emerging through team workflows, data pipelines, institutional requirements, and evolving stakeholder expectations. Ethical competence, therefore, extends beyond cognitive skill to a situated practice shaped by organizational routines, technical affordances, and supervisory norms [60]. Senior researchers play an important role by modeling everyday ethical decision-making and making these processes visible, enabling norms to diffuse across laboratories and professional communities [61]. Cross-training activities, such as journal clubs, further support this process by cultivating a shared vocabulary for reasoning about risks, roles, and constraints.

Because digital health research frequently involves health care systems and community partners, stakeholder communication and team science skills are prioritized in both coursework and experiential training [62]. Experts reported connecting trainees with community collaborators to build practical experience working across sectors, while classroom exercises in perspective-taking and public speaking reinforce these competencies. Such preparation is critical for sustaining multipartner research, where navigating disagreement is often central to maintaining trust.

Despite these complementary pathways, reliance on experiential learning raises concerns about consistency in training. Ethical norms are frequently transmitted within research groups, meaning that uneven access to hands-on sociotechnical decision-making could reinforce inequities in students' ability to develop ethical judgment [63]. Our syllabus analysis supports this interpretation in that courses commonly address privacy, fairness, and professional responsibilities; however, fewer examine how these principles are enacted within algorithmic pipelines, wearable sensing systems, or clinical governance processes. This gap helps explain why experts view experiential learning as essential for translating abstract principles into practice while underscoring the need for more intentional instructional supports.

XSL•FO
RenderX

Expanding Experiential Learning Through Instructional Infrastructure

One strategy for expanding access to practice-based learning is to scaffold undergraduate participation through a structured critique modeled after the ethics and society review [64]. Students who have completed research ethics coursework could contribute to proposal review in bounded, time-limited ways. Although ethics and society review relies on expert reviewers and may be difficult to scale, prior work in human-computer interaction demonstrates that scaffolding approaches, such as design critique and historical analysis, help novice learners complete complex evaluative tasks while strengthening ethical reasoning [59,65,66]. Instructors could adopt similar workflows by training students to collaboratively analyze case studies [67]. However, implementation challenges remain, including limited institutional support, uneven student preparedness for complex sociotechnical contexts, and the risk that simplified workflows may obscure ethical tensions. Addressing these constraints will require continued evaluation and the development of shared educational resources supported by professional associations and research sponsors.

Translating Tacit Insights Into Shareable Resources

Beyond scaffolding, our findings suggest the importance of translating practice-based insight into reusable instructional materials. Tacit knowledge develops through experience and is often conveyed as intuition, making it difficult to formalize. Yet ethical concerns that emerge during research offer opportunities for documentation. Existing tools, such as ethics checklists (eg, the Digital Health Checklist [68]) and design artifacts such as card decks that embed ethical principles into computing and AI research [69,70] demonstrate how abstract guidance can be externalized into prompts that support reflection in practice.

Building on this approach, researchers could create brief ethics discussion prompts when salient concerns arise during a study. Drawing on Parker's framework [71], such prompts would include a problem statement, a description of the solution space, and a guiding question, accompanied by reflection on how the issue was addressed. Unlike traditional case studies, which often isolate single decisions, these prompts could capture ethical judgment as it unfolds within research workflows. When shared alongside research outputs, they have the potential to convert tacit experience into concise, transferable instructional resources. Adoption, however, raises important questions about documenting sensitive issues, ownership of materials, and incentives for participation, which create tensions that must be addressed to normalize reflective ethical practice in digital health research.

Theoretical Implications for Digital Health Research

Our findings reconnect research ethics training to sociotechnical theory by illustrating that ethical competence emerges through the interaction of data infrastructures, research workflows, and institutional expectations rather than through formal instruction alone. Experts described guiding trainees through real-time decisions about passive data collection, such as GPS and sensor-based monitoring, where ethical judgment developed in relation to both continuous data flows and governance requirements surrounding consent and stewardship. Viewing ethics education through a sociotechnical lens shifts the focus from individual responsibility toward the systems and environments that shape ethical action. By foregrounding ethics education as a sociotechnical phenomenon, this study extends prior work that frames ethical decision-making as an individual competency. Instead, our findings suggest that ethical capacity is distributed across training environments, research infrastructures, and supervisory practices, underscoring the importance of designing educational systems that support ethical action.

Implications for Research Ethics Training in Digital Health

Responding to calls for more practice-oriented ethics training in digital health, this study makes 3 primary contributions. First, it provides a thematic analysis of explicit ethics instruction across 98 syllabi representing 76 international academic programs, offering one of the most comprehensive mappings of current curricular approaches. Second, it advances understanding of how trainees develop ethical competence by identifying experiential pathways through which tacit knowledge is cultivated in research settings. Third, it proposes actionable strategies for strengthening instructional infrastructure, including closer integration of classroom learning with research practice, scaffolded opportunities for student participation, educator-researcher learning communities, and the use of embedded ethics discussion prompts to support learning reflective practice.

Collectively, our findings position research ethics education not simply as a pedagogical challenge but as an instructional infrastructure issue within digital health research. Ethical competence is shaped by the environments in which researchers are trained, suggesting that strengthening this infrastructure requires closer integration of classroom learning with structured experiential opportunities, expanded scaffolded participation in ethical review, and mechanisms for translating practice-based insight into shared instructional resources. Advancing these efforts could promote more consistent training, reduce inequities in access to experiential learning, and better prepare the next generation of researchers to navigate the sociotechnical complexities of data-intensive health research.

Limitations

This study has several limitations related to our sampling strategy. Digital health remains an emerging sociotechnical field with few formal degree programs and no standardized course titles (eg, "digital health research ethics"), making traditional catalog searches impractical. We therefore identified researchers and affiliated programs through Association for Computing Machinery publications to center the study on active research environments; however, this approach may have excluded relevant curricula at teaching-focused colleges and medical schools that prioritize instruction over research. Our sample also reflects a geographic concentration in the United States and Europe, introducing a potential Western bias. Teaching priorities and approaches to ethics education may differ globally, highlighting the need for more internationally

XSL•FO
RenderX

representative research. Additionally, syllabi provide insight into intended instruction but do not capture how courses are experienced by instructors or students. Although interviews with senior researchers offered perspectives on mentoring and training, we did not include trainees or individuals who left academic pathways, limiting our understanding of how ethics education is enacted and received. Future research should examine classroom practices and team-based research environments to better identify effective approaches to teaching ethical digital health research.

Conclusion

Our study demonstrates that ethics education and learning emerge through the interaction of formal coursework and practice-based training, each cultivating distinct capabilities needed to navigate data-intensive and collaborative research environments. While these complementary pathways help prepare students to recognize ethical issues, apply principles in context, anticipate emerging risks, and communicate across disciplines, access to experiential ethics training remains uneven. Strengthening ethics education will therefore require expanding structured opportunities for early research engagement, fostering communities of practice among educators and researchers, and developing mechanisms that translate tacit ethical reasoning into shareable resources. Designing training environments that integrate ethical reflection into everyday research practice may better equip the next generation of digital health researchers to responsibly shape the sociotechnical health systems they both study and build.

Acknowledgments

The authors wish to thank Sneha Jaiswal for her contributions to the research process and to Daniela G Vital for her administrative program management over the course of the project.

Funding

This material is based upon work supported by the National Science Foundation under grant 2124975. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: CN (lead), BM (equal)
Data curation: YZ, BM, TP
Formal analysis: BM, YZ, TP (supporting)
Funding acquisition: CN (lead)
Investigation: YZ, TP, BY
Methodology: BM
Project administration: BY (lead)
Resources: CN
Supervision: BM, CN, BY
Validation: YZ, TP
Visualization: YZ (lead), TP (supporting)
Writing—original draft: YZ (lead), CN, BY (supporting)
Writing—review & editing: YZ, BM, TP, BY, CN

Conflicts of Interest

None declared.

Multimedia Appendix 1
Readings to support themes identified in the Results section.
[DOCX File , 40 KB - jmir_v28i1e82861_app1.docx]

References

1. Lupton D. Critical perspectives on digital health technologies. Sociol Compass 2014;8(12):1344-1359. [doi: 10.1111/soc4.12226]
2. Topol E. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. USA: Basic Books, Inc; 2019.

XSL•FO
RenderX

3. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017;19(11):e8775 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
4. Sittig DF, Singh H. A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. Qual Saf Health Care 2010;19 Suppl 3(Suppl 3):i68-i74 [FREE Full text] [doi: 10.1136/qshc.2010.042085] [Medline: 20959322]
5. Wilcox L, Brewer R, Diaz F. AI consent futures: a case study on voice data collection with clinicians. Proc ACM Hum-Comput Interact 2023;7(CSCW2):1-30. [doi: 10.1145/3610107]
6. Hernandez-Cruz N, Garcia-Constantino M, Beltran-Marquez J, Cruz-Sandoval D, Lopez-Nava I, Cleland I, et al. Study design of an environmental smart microphone system to detect anxiety in patients with dementia. USA: Association for Computing Machinery; 2019 Presented at: Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare; 2029 May 20; New York, NY, USA p. 383-388. [doi: 10.1145/3329189.3329234]
7. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
8. Dimond JP, Fiesler C, Bruckman AS. Domestic violence and information communication technologies. Interact Comput 2011;23(5):413-421. [doi: 10.1016/j.intcom.2011.04.006]
9. Saleheen N, Ullah MA, Chakraborty S, Ones DS, Srivastava M, Kumar S. WristPrint: characterizing user re-identification risks from wrist-worn accelerometry data. Conf Comput Commun Secur 2021;2021:2807-2823 [FREE Full text] [doi: 10.1145/3460120.3484799] [Medline: 36883116]
10. Bloss C, Nebeker C, Bietz M, Bae D, Bigby B, Devereaux M, et al. Reimagining human research protections for 21st century science. J Med Internet Res 2016;18(12):e329 [FREE Full text] [doi: 10.2196/jmir.6634] [Medline: 28007687]
11. Dunseath S, Weibel N, Bloss CS, Nebeker C. NIH support of mobile, imaging, pervasive sensing, social media and location tracking (MISST) research: laying the foundation to examine research ethics in the digital age. NPJ Digit Med 2018;1(1):20171 [FREE Full text] [doi: 10.1038/s41746-017-0001-5] [Medline: 31304345]
12. Shilton K, Greene D. Linking platforms, practices, and developer ethics: levers for privacy discourse in mobile application development. J Bus Ethics 2017;155(1):131-146. [doi: 10.1007/s10551-017-3504-8]
13. McInnis BJ, Pindus R, Kareem D, Vital DG, Hekler EB, Nebeker C. Factors influencing informed consent preferences in digital health research: survey study of prospective participants. J Med Internet Res 2025;27:e63349 [FREE Full text] [doi: 10.2196/63349] [Medline: 39847412]
14. McInnis BJ, Pindus R, Kareem D, Nebeker C. Considerations for the design of informed consent in digital health research: participant perspectives. J Empir Res Hum Res Ethics 2024;19(4-5):175-185 [FREE Full text] [doi: 10.1177/15562646241290078] [Medline: 39397541]
15. McInnis BJ, Pindus R, Kareem D, Gamboa S, Nebeker C. Exploring the future of informed consent: applying a service design approach. Proc ACM Hum-Comput Interact 2024;8(CSCW1):1-31. [doi: 10.1145/3637330]
16. Nebeker C, Harlow J, Espinoza Giacinto R, Orozco-Linares R, Bloss CS, Weibel N. Ethical and regulatory challenges of research using pervasive sensing and other emerging technologies: IRB perspectives. AJOB Empir Bioeth 2017;8(4):266-276. [doi: 10.1080/23294515.2017.1403980] [Medline: 29125425]
17. Fiesler C, Garrett N, Beard N. What do we teach when we teach tech ethics? A syllabi analysis. USA: Association for Computing Machinery; 2020 Presented at: Proceedings of the 51st ACM Technical Symposium on Computer Science Education; 2020 Feb 26; New York, NY, USA p. 289-295. [doi: 10.1145/3328778.3366825]
18. Delcourt C, Venkatagiri S, Chandrasekharan E. What's in a social computing course: analyzing computer and information science syllabi. USA: Association for Computing Machinery; 2024 Presented at: Proceedings of the 6th Annual Symposium on HCI Education; 2024 Jun 05; New York, NY, USA p. 1-8. [doi: 10.1145/3658619.3658623]
19. Holtzblatt K, Beyer H. The affinity diagram. In: Holtzblatt K, Beyer H, editors. Context Des Second Ed Boston: Morgan Kaufmann. Context Des Second Ed Boston: Morgan Kaufmann; 2017:127-146.
20. FigJam: the online collaborative whiteboard for teams. Figma. 2025. URL: https://www.figma.com/figjam/ [accessed 2025-08-08]
21. Birks M, Chapman Y, Francis K. Memoing in qualitative research: probing data and processes. J Res Nurs 2008;13(1):68-75. [doi: 10.1177/1744987107081254]
22. Integrity in scientific research video series. American Association for the Advancement of Science. URL: https://www.aaas.org/resources/integrated-public-use-microdata-series-ipums [accessed 2026-03-10]
23. Jones CP. Confronting institutionalized racism. Phylon (1960) 2002;50(1/2):7. [doi: 10.2307/4149999]
24. Ess C. Digital media ethics. In: Oxford Research Encyclopedia of Communication. Oxford, UK: Oxford University Press; 2017.
25. Responsible conduct of research training. NIH Office of Intramural Research. URL: https://oir.nih.gov/sourcebook/ethical-conduct/responsible-conduct-research-training [accessed 2026-03-10]
26. Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, et al. The moral machine experiment. Nature 2018;563(7729):59-64. [doi: 10.1038/s41586-018-0637-6] [Medline: 30356211]

XSL•FO
RenderX

27. Kolata G. When doctors use a chatbot to improve their bedside manner. The New York Times. 2023. URL: https://www.nytimes.com/2023/06/12/health/doctors-chatgpt-artificial-intelligence.html [accessed 2026-03-10]
28. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366(6464):447-453 [FREE Full text] [doi: 10.1126/science.aax2342] [Medline: 31649194]
29. Waldman AE. Cognitive biases, dark patterns, and the 'privacy paradox'. Curr Opin Psychol 2020;31:105-109. [doi: 10.1016/j.copsyc.2019.08.025] [Medline: 31590106]
30. Foxwell HJ. Creating Good Data: A Guide to Dataset Structure and Data Representation. New York, NY: Apress; 2020.
31. James G, Witten D, Hastie T, Tibshirani R. Introduction to Statistical Learning: With Applications in R. 2nd ed. New York, NY: Springer; 2021.
32. O'Neil C. The ethical data scientist. Slate. 2016. URL: https://slate.com/technology/2016/02/how-to-bring-better-ethics-to-data-science.html [accessed 2026-03-10]
33. Conklin AW. Principles of Computer Security: CompTIA Security+ and Beyond. 5th ed. New York, NY: McGraw-Hill Education; 2026.
34. Dix A. Human–computer interaction: a stable discipline, a nascent science, and the growth of the long tail. Interacting with Computers 2010;22(1):13-27. [doi: 10.1016/j.intcom.2009.11.007]
35. Garrett JJ. The Elements of User Experience: User-Centered Design for the Web and Beyond. 2nd ed. Berkeley, CA: New Riders; 2011.
36. Horton S, Quesenbery W. A Web for Everyone: Designing Accessible User Experiences. New York, NY: Rosenfeld Media; 2014.
37. Lazar J, Goldstein DF, Taylor A. Ensuring Digital Accessibility through Process and Policy. 1st ed. San Francisco, CA: Morgan Kaufmann Publishers Inc; 2015.
38. Barocas S, Selbst AD. Big data's disparate impact. Calif L Rev 2016. [doi: 10.15779/Z38BG31]
39. Calo R. Artificial intelligence policy: a roadmap. SSRN 2017. [doi: 10.2139/ssrn.3015350]
40. Friedler SA, Scheidegger C, Venkatasubramanian S. The (Im)possibility of fairness. Commun ACM 2021;64(4):136-143. [doi: 10.1145/3433949]
41. Kirkpatrick N. Making a living in the toxic world of discarded electronics. The Washington Post. 2015. URL: https://www.washingtonpost.com/news/in-sight/wp/2015/04/15/the-children-who-make-a-living-in-the-toxic-world-of-discarded-electronics [accessed 2026-03-10]
42. Noble SU. Algorithms of Oppression: How Search Engines Reinforce Racism. New York, NY: New York University Press; 2018.
43. Rotman D. The disparity between the rich and everyone else is larger than ever in the United States and increasing in much of Europe. Why? MIT Technology Review. 2014. URL: https://www.technologyreview.com/2014/10/21/170679/technology-and-inequality/ [accessed 2026-03-10]
44. Shaw J. Confronting pitfalls of machine learning, artificial intelligence. Harvard Magazine. 2018. URL: https://www.harvardmagazine.com/2018/12/artificial-intelligence-limitations [accessed 2026-03-10]
45. Wallach W, Allen C. Moral Machines: Teaching Robots Right from Wrong. Oxford, UK: Oxford University Press; 2009.
46. Winner L. Trust and terror: the vulnerability of complex socio‐technical systems. Sci Cult 2004;13(2):155-172. [doi: 10.1080/0950543042000226594]
47. Dainow DB. SIENNA D6.3: Methods for translating ethical analysis into instruments for the ethical development and deployment of emerging technologies. Zenodo 2021. [doi: 10.5281/ZENODO.5541538]
48. Niki A. How to nail the pitch: 17 storytelling tips for startups. The Founder Institute. URL: https://FI.co/insight/how-to-nail-the-pitch-17-storytelling-tips-for-startups [accessed 2026-03-10]
49. Michaelson HB. Creative aspects of engineering writing. IRE Trans Eng Writing Speech 1961;4(3):77-79. [doi: 10.1109/TEWS.1961.4322737]
50. Tebeaux E, Dragga S. The Essentials of Technical Communication. 5th ed. Oxford, UK: Oxford University Press; 2020.
51. Trimble J. Writing with Style: Conversations on the Art of Writing. 3rd ed. Munich: Pearson; 2010.
52. Gotterbarn D, Brinkman B, Flick C, Kirkpatrick MS, Miller K, Varansky K, et al. ACM code of ethics and professional conduct. Association for Computing Machinery. URL: https://www.acm.org/code-of-ethics [accessed 2026-03-10]
53. Compact between postdoctoral appointees and their mentors. Association of American medical colleges. Association of American Medical Colleges. 2017. URL: https://www.aamc.org/about-us/mission-areas/biomedical-research/post-doc-compact [accessed 2026-03-10]
54. Davis M. Thinking Like an Engineer: Studies in the Ethics of a Profession. 1st ed. New York, NY: Oxford University Press; 1998.
55. Hill LA. Becoming a Manager: How New Managers Master the Challenges of Leadership. Boston: Harvard Business Review Press; 2003.
56. National Academies of Sciences, Engineering, and Medicine, Policy and Global Affairs, Board on Higher Education and Workforce, Committee on Effective Mentoring in STEMM. In: Dahlberg ML, Byars-Winston A, editors. The Science of Effective Mentorship in STEMM. Washington, DC: National Academies Press; 2019.

XSL•FO
RenderX

57. Baxter G, Sommerville I. Socio-technical systems: from design methods to systems engineering. Interact Comput 2011;23(1):4-7 [FREE Full text] [doi: 10.1002/iub.1366] [Medline: 25904163]
58. Todd EM, Torrence BS, Watts LL, Mulhearn TJ, Connelly S, Mumford MD. Effective practices in the delivery of research ethics education: a qualitative review of instructional methods. Account Res 2017;24(5):297-321. [doi: 10.1080/08989621.2017.1301210] [Medline: 28332862]
59. Luther K, Tolentino J, Wu W, Pavel A, Bailey B, Agrawala M, et al. Structuring, aggregating, and evaluating crowdsourced design critique. USA: Association for Computing Machinery; 2015 Presented at: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing; 2025 Feb 28; New York, NY, USA p. 473-485. [doi: 10.1145/2675133.2675283]
60. Eraut M. Non-formal learning and tacit knowledge in professional work. Br J Educ Psychol 2000;70(1):113-136. [doi: 10.1348/000709900158001] [Medline: 10765570]
61. Lave J, Wenger E. Situated Learning: Legitimate Peripheral Participation. United Kingdom: Cambridge University Press; 1991.
62. Hall KL, Vogel AL, Huang GC, Serrano KJ, Rice EL, Tsakraklides SP, et al. The science of team science: a review of the empirical evidence and research gaps on collaboration in science. Am Psychol 2018;73(4):532-548. [doi: 10.1037/amp0000319] [Medline: 29792466]
63. Hafferty FW, Franks R. The hidden curriculum, ethics teaching, and the structure of medical education. Acad Med 1994;69(11):861-871. [doi: 10.1097/00001888-199411000-00001] [Medline: 7945681]
64. Bernstein MS, Levi M, Magnus D, Rajala BA, Satz D, Waeiss Q. Ethics and society review: ethics reflection as a precondition to research funding. Proc Natl Acad Sci USA 2021;118(52):e2117261118 [FREE Full text] [doi: 10.1073/pnas.2117261118] [Medline: 34934006]
65. Luther K, Pavel A, Wu W, Tolentino J, Agrawala M, Hartmann B, et al. CrowdCrit: crowdsourcing aggregating visual design critique. USA: Association for Computing Machinery; 2014 Presented at: Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing; 2014 Feb 15; New York, NY, USA p. 21-24. [doi: 10.1145/2556420.2556788]
66. Wang NC, Hicks D, Luther K. Exploring trade-offs between learning and productivity in crowdsourced history. 2018 Presented at: Proceedings of the ACM on Human-Computer Interaction; 2018 Nov 1; Canada p. 1-24. [doi: 10.1145/3274447]
67. Feldman MQ, McInnis BJ. How we write with crowds. 2021 Presented at: Proceedings of the ACM on Human-Computer Interaction; 2021 Jan 5; New York NY USA p. 1-31. [doi: 10.1145/3432928]
68. Nebeker C, Bartlett Ellis RJ, Torous J. Development of a decision-making checklist tool to support technology selection in digital health research. Transl Behav Med 2020;10(4):1004-1015 [FREE Full text] [doi: 10.1093/tbm/ibz074] [Medline: 31120511]
69. Hsieh G, Halperin BA, Schmitz E, Chew YN, Tseng YC. What is in the cardsxploring uses, patterns, trends in design cards. USA: Association for Computing Machinery; 2023 Presented at: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems; 2023 Apr 23; New York, NY, USA p. 1-18. [doi: 10.1145/3544548.3580712]
70. Sadek M, Constantinides M, Quercia D, Mougenot C. Guidelines for Integrating Value Sensitive Design in Responsible AI Toolkits. USA: Association for Computing Machinery; 2024 Presented at: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems; 2024 Feb 29; New York, NY, USA p. 1-20. [doi: 10.1145/3613904.3642810]
71. Parker WC. Public discourses in schools: Purposes, problems, possibilities. Educ Res 2006;35(8):11-18. [doi: 10.3102/0013189x035008011]

Abbreviations

AI: artificial intelligence
EQUATOR: Enhancing the Quality and Transparency of Health Research
HCD: human-centered design
IRB: institutional review board
SRQR: Standards for Reporting Qualitative Research

Edited by A Stone; submitted 22.Aug.2025; peer-reviewed by H Matisonn, S Arabi; comments to author 29.Oct.2025; accepted 18.Feb.2026; published 26.Mar.2026.

Please cite as:
Zhu Y, McInnis B, Punater T, York B, Nebeker C
Educating Students About Digital Health Research Ethics: Curricula Review and Expert Interview Study
J Med Internet Res 2026;28:e82861
URL: https://www.jmir.org/2026/1/e82861
doi:10.2196/82861
PMID:

XSL•FO
RenderX

©Yier Zhu, Brian McInnis, Tanya Punater, Brittany York, Camille Nebeker. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Strategy for Hepatitis B and C Virus Testing Campaigns Through Web Services and Digital Advertising in Japan: Nationwide Cross-Sectional Study With Correspondence Analysis

Kosuke Sakai[1], MD, PhD; Shoko Nakazawa[1], MD, PhD; Yuko Furuya[1], MD, PhD; Kota Fukai[1], MD, PhD; Kei Sano[2], MD, PhD; Masaaki Korenaga[3], MD, PhD; Masayuki Tatemichi[1], MD, PhD

[1]Department of Preventive Medicine, School of Medicine, Tokai University, 143 Shimokasuya, Isehara, Japan

[2]Department of Ophthalmology, The Jikei University School of Medicine, Minatoku, Japan

[3]Hepatitis Information Center, The Research Center for Hepatitis and Immunology, National Institute of Global Health and Medicine, Japan Institute for Health Security, Ichikawa, Japan

Corresponding Author:
Kota Fukai, MD, PhD
Department of Preventive Medicine, School of Medicine, Tokai University, 143 Shimokasuya, Isehara, Japan

Abstract

Background: Public awareness campaigns and testing promotion must be strengthened to eliminate infections with hepatitis B and C viruses (HBV and HCV, respectively) by 2030. Although public health campaigns using various forms of advertising are widely implemented, the most appropriate channels for viral hepatitis testing remain unclear.

Objective: This study aims to identify web services and digital advertising channels appropriate for promoting HBV and HCV testing, segmented by prior testing history and the desire for hepatitis virus testing.

Methods: A nationwide cross-sectional online survey of Japanese adults aged 20 to 69 years was conducted. The respondents answered questions regarding viral hepatitis testing status, routinely used web services (180 options), and exposure to digital advertising (25 options). Correspondence analysis was used to visualize relationships among testing segments, web services, and digital advertising. For individuals classified as "never having been tested and wishing to be tested," channel-specific alignment was quantified using cosine θ. Sensitivity analyses were conducted by repeating the correspondence analysis after excluding respondents uncertain about their testing history and by fitting modified Poisson regression models with robust variance to estimate prevalence ratios and 95% CIs.

Results: Of the 2000 respondents (1011 male and 989 female), 18% (n=359) reported prior HBV and HCV testing, and 22.1% (n=441) were unsure whether they had ever been tested. Web services characteristically associated with "never having been tested and wishing to be tested" included Lawson (convenience store: cosine θ=0.989) and Cosme (shopping: cosine θ=0.987). The corresponding digital advertising channels included in-store and storefront screens at Welcia (pharmacy chain: θ=0.994) and Lawson (cosine θ=0.937). Segment-specific patterns varied according to age group and sex. Sensitivity analyses excluding the unsure group showed similar patterns. Modified Poisson regression results were also consistent; for example, Lawson web service use was associated with a desire for hepatitis virus testing (prevalence ratio 1.75, 95% CI 1.22 - 2.52).

Conclusions: In Japan, the convenience store chain Lawson was a frequently used touchpoint, both online and offline, among individuals seeking viral hepatitis testing. Future studies are needed to determine whether implementing awareness-raising activities through Lawson can increase the uptake of testing and subsequent treatment.

(J Med Internet Res 2026;28:e89585) doi:10.2196/89585

KEYWORDS

hepatitis B; hepatitis C; digital advertising; web services; correspondence analysis

Introduction

The World Health Organization (WHO) has set the goal of eliminating hepatitis B virus (HBV) and hepatitis C virus (HCV) infections by 2030 [1]. In Japan, over 70% of patients with hepatocellular carcinoma are infected with HBV or HCV [2,3]. In 2015, approximately 2 million individuals were infected with HBV or HCV, of whom approximately 1 million were undiagnosed or untreated [4]. From 2015 to 2022, a decreasing trend in acute-onset HBV and HCV hepatitis was reported, partially attributable to strengthened surveillance systems and, in the case of HCV, to remarkable advances in treatment with direct-acting antivirals, which have dramatically reduced the disease burden [5]. However, the burden of viral hepatitis is

XSL•FO
RenderX

predicted to persist until 2035, and more public health measures are required to achieve this goal [4].

In Japan, viral hepatitis testing is offered through community screening programs, health checkups, public health centers, prenatal care, and blood donations [6]. Among these, annual workplace health checkups, which cover many working-age adults, represent a particularly important opportunity for hepatitis virus testing [7]. Several companies have used this opportunity to implement HBV and HCV testing. Nevertheless, viral hepatitis screening is not part of legally required tests, resulting in only a subset of employees undergoing testing [7]. Furthermore, testing is frequently avoided due to social stigma and workplace discrimination [8]. To effectively address these challenges and enhance the response to viral hepatitis, the Ministry of Health, Labor and Welfare has formally urged employers and relevant organizations nationwide to collaborate to facilitate testing and ensure that employees who test positive receive appropriate medical treatment [9]. Hence, our research team has conducted awareness campaigns targeting top management and staff responsible for workplace health checkups within the human resources and general affairs departments [10]. To strengthen countermeasures against viral hepatitis, it is necessary to promote awareness campaigns among not only key workers but also the general population.

Recently, health awareness campaigns using websites and social media advertisements have become increasingly prevalent [11]. For example, as part of a viral hepatitis awareness campaign, educational columns and videos have been distributed via websites such as “Let’s Learn About Hepatitis” [12]. Information on the hepatitis virus has been widely disseminated to the public. However, effective health campaigns require accurately defining the target population and delivering content that is appropriately tailored to their specific attributes [13]. A scoping review of HCV media campaigns found that their effectiveness varied, emphasizing the need for customized approaches [14]. Accordingly, HCV care has been incorporated into opioid treatment programs to improve access for high-risk groups [15]. Additionally, social media strategies have been explored to identify individuals with HBV [16]. Overall, these observations indicate that designing interventions specifically for different populations is crucial for success.

Public health marketing for viral hepatitis should be tailored to a person’s testing history and results. One effective strategy to encourage more people to get tested is to guide those who have never been tested but are interested toward specific locations where they can access testing. However, there is limited knowledge regarding the most effective ways to share information about hepatitis testing with individuals who seek it.

This study aims to identify which web services and digital advertisements are most effective for promoting hepatitis testing in Japan, using groups defined by their prior testing history and interest in testing. The results of this study will inform the development of digital advertising strategies to more effectively disseminate information and promote hepatitis testing. These findings on public health advertising strategies can be applied to a range of health issues.

Methods

Study Design and Selection of Respondents

This nationwide cross-sectional study was conducted in August 2025 using a Japanese-language online questionnaire. Rakuten Insight Inc, a Tokyo-based web research firm, supported the survey. Panelists were users who registered through the company’s standard process and could be invited to specific surveys based on their profile information. Rakuten Insight has been used in previous peer-reviewed public health research. Respondents were recruited from the company’s registry via email and app push notifications. Invitations were randomly sent to individuals with active accounts in the past year who met the inclusion criteria as follows: aged 20 to 69 years, residing in Japan, and consenting to participate. Participants first answered screening questions; only those eligible completed the full questionnaire. Stratified sampling was carried out across 20 segments, divided by sex (male and female) and 5-year age strata, reflecting population estimates from Japan’s 2020 Population Census [17]. Surveys continued until 2000 valid responses were obtained, with target quotas set for each segment. The sample size was determined by the study budget and the maximum feasible recruitment. This study followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for cross-sectional observational studies (Checklist 1) [18], and web survey reporting adhered to the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) standards (Checklist 2) [19].

As this was a closed survey administered by a commercial online survey company, respondents were required to log in using their unique panel numbers before accessing the questionnaire. Once a respondent completed the survey, the questionnaire was no longer displayed. The IP addresses or cookies were not included in the survey. The survey consisted of 1 question per page, totaling 8 questions. No final review of responses was conducted.

Assessment of HBV and HCV Testing History and Desire for Testing

Information regarding the history of hepatitis virus testing and the desire for testing was obtained using a questionnaire. The respondents were asked to indicate the statement that best described their HBV and HCV testing status: (1) having been tested and knowing the results, (2) having been tested but not knowing the results, (3) never having been tested and wishing to be tested, (4) never having been tested and not wishing to be tested, and (5) unsure if tested. In Japan, public guidance recommends that all individuals undergo hepatitis virus testing at least once in their lifetime; therefore, this study focused on ever-tested status and the current desire for testing [7].

Assessment of Daily Used Web Services

Respondents were asked to complete multiple-choice items on regularly used web services. The survey was conducted annually by the research firm with all panel respondents. Data collected over 1 year, from November 2024 (when the questionnaire was distributed) through the completion of this survey, were analyzed. The items were synthesized from 28 genres: video

XSL•FO
RenderX

streaming, video sharing, music, social networking services, search engines, news, newspapers, magazines, books, calendars, hobbies, learning, maps, points, beauty, finance, general e-commerce, electronics stores, retail, manufacturer e-commerce, online shopping, flea markets, games, cooking recipes, restaurants, food delivery, travel, and weather. Multimedia Appendix 1 lists 180 types of web services, including Google, YouTube, Amazon, Netflix, Wikipedia, Facebook, Instagram, X (formerly known as Twitter), LinkedIn, and Spotify.

Assessment of Daily Shown Digital Advertising

Respondents were asked about their routine exposure to digital advertising by broad genres rather than brand names, such as convenience stores, supermarkets, shopping malls, drugstores, discount stores, transit media including in-car train screens, stations, bus stops, taxis, large outdoor video screens, vending machines, other noncommercial facilities, and other digital advertising (Multimedia Appendix 2).

Other Measurements

To characterize the respondents, information was collected on sex, age, educational background, household income, marital status, and employment status. Sex was self-reported, with respondents indicating whether they were male or female. Age was determined using the age on the date of response, based on the date of birth.

Regarding their educational background, respondents were asked to select the most appropriate option from the following: middle school, high school, vocational school, junior or technical college, university, graduate school, currently enrolled, or other.

The participants were further asked to select the most appropriate option from the following choices regarding pretax household income for 2024: up to JP ¥4 million, JP ¥4.01 - 6 million, JP ¥6.01 - 8 million, JP ¥8.01 - 10 million, JP ¥10.01 - 12 million, JP ¥12.01 - 15 million, and over JP ¥15 million (JP ¥159=US $1).

Marital status was defined either as married or unmarried. Regarding employment status, individuals who were currently working were asked about their main occupations and industries. Occupations were classified into 12 categories based on the Japan Standard Occupational Classification (Fifth Revision, December 2009) [20] as follows: administrative and managerial workers; professional and engineering workers; clerical workers; sales workers; service workers; security workers; agriculture, forestry, and fishery workers; manufacturing process workers; transport and machine operation workers; construction and mining workers; carrying, cleaning, packaging, and related workers; and workers unclassified by occupation [20]. Industries were classified into 20 categories based on the Japan Standard Industrial Classification: agriculture and forestry; fisheries; mining and quarrying of stone and gravel; construction; manufacturing; electricity, gas, heat, and water; information and communications; transport and postal services; wholesale and retail trade; finance and insurance; real estate and goods rental and leasing; scientific research, professional, and technical services; accommodation, food, and drink services; living-related, personal, and amusement services; education and learning support; medical, health care, and welfare; compound services; other services; government; and unclassified industries [21].

Statistical Analysis

The numbers and percentages of respondents were compiled by sex, age, education, income, marital status, and HBV or HCV testing status. To assess selection bias in the survey, we compared the distributions of respondents' educational backgrounds, income, occupations, industries, and residential regions with those from the 2020 Population Census [17]. The respondents' selection status was tabulated for 180 types of web services and 25 types of digital advertisements. As all questionnaire items were set as mandatory, no data were missing.

Correspondence analysis was used to identify the web services and digital advertising appropriate for raising awareness, tailored to the history of viral hepatitis testing and the current desire for testing. Correspondence analysis is frequently used in marketing analysis [22], as it can handle large amounts of categorical data and has been employed to examine relationships between purchasing trends across numerous products and buyer attributes. In public health studies, this method has also been applied to verify the distribution of categorical data [23]. Our previous research targeting key individuals for workplace health checkups also employed this method, successfully identifying prioritized information on viral hepatitis and appropriate media for conducting campaigns [10]. Ideally, a comprehensive representation of the distribution across the 5 HBV or HCV response categories would necessitate 4 dimensions to capture the total variance (100% inertia). However, as 4D models are not practically interpretable for visualization, we selected 3 dimensions as the most suitable option for effective exploratory mapping. Statistical analyses were performed using KH Coder software (version 3.0, SCREEN Advanced System Solutions Co Ltd) [24].

To confirm the validity of the responses, a correspondence analysis was performed to visualize web service usage patterns by respondents' sex and age. A 3D visualization of the correspondence analysis results was generated to characterize web service usage patterns associated with each viral hepatitis testing history and testing desire. To quantitatively evaluate the web service characteristics of groups willing to undergo testing (willing group), we assessed the deviation from the origin of each web service relative to the willing group using cosine θ. The same analysis was conducted for the digital advertisements. Since digital service usage patterns can differ by sex and age, the analysis was stratified by sex (male and female) and age group (20 - 49 y and 50 - 69 y).

Responses regarding HBV or HCV testing history and intention were classified into 5 mutually exclusive categories: tested and known result, tested but unknown result, never tested and willing to be tested, never tested and unwilling to be tested, and unsure if previously tested. The category "unsure if tested" was treated as an independent variable, as uncertainty about prior testing was considered meaningful for public health communication and may reflect distinct factors, such as limited recall and incomplete communication of test results.

XSL•FO
RenderX

During the sensitivity analysis, we excluded respondents who selected “unsure if tested” and subsequently performed correspondence analysis to identify web services and digital advertising channels associated with the willing group. Additionally, we fitted a series of modified Poisson regression models with robust variance, treating each of the web services selected by at least 5% (100/2000) and each of the digital advertisements as binary outcomes. Although the correspondence analysis used all web services, the multivariate regression analysis included only selected web services to adjust for covariates. The main independent variable was the desire for viral hepatitis testing (never tested but want to be tested vs all other responses). All models were adjusted for sex, age, household income, educational background (less than university vs university or higher), and marital status. Models with web services as outcomes were further adjusted for the total number of web services selected (0‐180); models with digital advertisements as outcomes were further adjusted for the total number of advertisements selected (0‐25). Age, sex, household income, educational attainment, and marital status were included as covariates as they likely influence prior testing history and desire for hepatitis virus testing [25]. Moreover, the total number of selected web services or digital advertising channels was included, as respondents demonstrating a higher overall propensity to choose multiple digital options are, by definition, more likely to exhibit positive results for any specific outcome. The detailed variable definitions and coding schemes (web services and digital advertising channels) are presented in Multimedia Appendix 3. For each outcome, we estimated prevalence ratios (PRs) and 95% CIs for the desire for viral hepatitis testing, with a 2-sided significance level of 5%. The statistical software used was R (version 4.4.2; R Foundation for Statistical Computing) [26].

Ethical Considerations

This study was approved by the research ethics committee of the Tokai University School of Medicine (23R039). Informed consent was obtained from all participants before they responded to the questionnaire. The consent screen outlined the study’s purpose, procedures, voluntary participation, and the process for withdrawing consent. Respondents were allowed to withdraw their consent up to 3 months after survey completion. To protect privacy, Rakuten Insight removed all personally identifiable information before data delivery. These measures were designed to minimize the risk of information leakage. Participants earned panel points based on the provider’s standard incentive system: those who completed the survey received 5 points, each equivalent to JP ¥1 (US $0.6), which were redeemable within the provider’s partner network.

Results

Participants

Of the 2,379,232 registered individuals in the survey panel, 357,272 were invited to participate in the study via email and push notifications. Of these, 5621 responded to the screening question (Figure 1). A total of 3278 individuals were excluded due to declining participation (n=2676), incomplete responses, or reaching quota targets (n=602). Among the 2343 individuals who answered all questions, 2000 were selected for analysis using the following process. First, respondents were excluded if their response times were below a threshold defined as the median response time multiplied by a device factor (0.3 for PC and 0.2 for smartphone). Subsequently, a random number was assigned to each valid sample using the RAND function in Microsoft Excel. The samples were ranked in ascending order according to assigned random numbers, and the top 2000 were selected for inclusion in the final dataset. The invitation-to-screening proportion was 1.6% (5621/357,272), while the screening-to-analysis proportion was 85.4% (2000/2343).

XSL•FO
RenderX

Figure 1. Flowchart of participant selection for analysis.

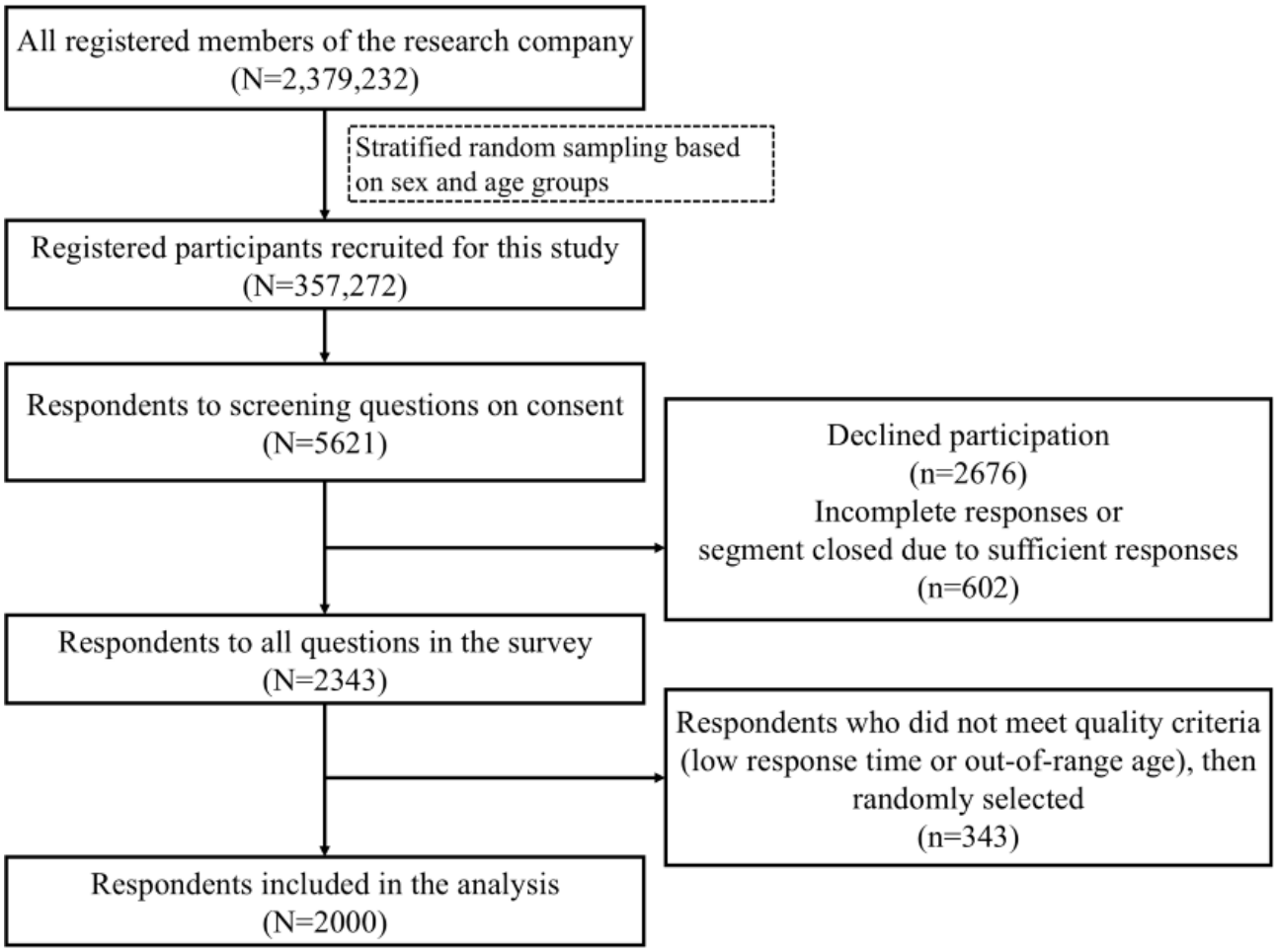

Characteristics of the Respondents

Responses were obtained from 1011 male and 989 female participants (Table 1). Most respondents (484/2000, 24.2%) were aged 50 to 59 years. The educational background was university or graduate school for 56.3% (569/1011) of the male and 37.6% (372/989) of the female participants. A household income of less than JP ¥4 million was the most common, accounting for 31.4% (317/1011) of male and 41.2% (407/989) of female participants. Regarding marital status, 48.4% (489/1011) of male and 52.1% (515/989) of female participants were married. The percentages of respondents who had been tested for HBV or HCV and knew their results were 14.4% (146/1011) among male and 18.7% (185/989) among female participants. The percentage of individuals who were never tested but wanted to be tested was 26% (263/1011) among male and 23.4% (231/989) among female participants.

To assess selection bias, we compared the distributions of educational background (Multimedia Appendix 4), household income (Multimedia Appendix 5), occupation and industry (Multimedia Appendix 6), and residential region (Multimedia Appendix 7) with those from the 2020 Population Census of Japan.

Regarding educational attainment, national statistics indicate that 27.8% of male and 17.9% of female participants hold university degrees, whereas in this study, 50.1% of male and 34.8% of female participants hold university degrees.

Regarding household income, the census showed that 48.9% of households earned less than JP ¥4 million, whereas only 36.2% of respondents in this study were in this income bracket. Among male participants, administrative and managerial workers accounted for 2.1% in the census and 14.5% in this study, sales workers accounted for 12.4% in the census and 5.8% in this study, and manufacturing process workers accounted for 13.4% in the census and 6.3% in this study. The distribution of industries among the respondents in this study was similar to that in the census.

Regarding residential regions, participants were overrepresented in Tokyo (male: +2.41% points; female: +2.7% points), Kanagawa (male: +2.24% points; female: +2.47% points), and Osaka (male: +2.13% points; female: +0.70% points) compared with the national population distribution.

XSL•FO
RenderX

Table . Respondent characteristics.

Characteristics	All (N=2000), n (%)	Male (n=1011), n (%)	Female (n=989), n (%)
Age (y)			
20 - 29	339 (17)	175 (17.3)	164 (16.6)
30 - 39	351 (17.6)	180 (17.8)	171 (17.3)
40 - 49	433 (21.7)	219 (21.7)	214 (21.6)
50 - 59	484 (24.2)	244 (24.1)	240 (24.3)
60 - 69	393 (19.7)	193 (19.1)	200 (20.2)
Educational background			
Middle or high school	534 (26.7)	254 (25.1)	280 (28.3)
Vocational school, junior college, or technical college	434 (21.7)	130 (12.9)	304 (30.7)
University or graduate school	941 (47.1)	569 (56.3)	372 (37.6)
Currently enrolled	73 (3.7)	47 (4.6)	26 (2.6)
Other	18 (0.9)	11 (1.1)	7 (0.7)
Household income (JP ¥[a])			
≤4 million	724 (36.2)	317 (31.4)	407 (41.2)
4.01 - 6 million	414 (20.7)	205 (20.3)	209 (21.1)
6.01 - 8 million	358 (17.9)	192 (19)	166 (16.8)
8.01 - 10 million	236 (11.8)	138 (13.6)	98 (9.9)
10.01 - 12 million	133 (6.7)	79 (7.8)	54 (5.5)
12.01 - 15 million	73 (3.7)	42 (4.2)	31 (3.1)
>15 million	62 (3.1)	38 (3.8)	24 (2.4)
Marital status			
Married	1004 (50.2)	489 (48.4)	515 (52.1)
Unmarried	996 (49.8)	522 (51.6)	474 (47.9)
History of HBV[b] and HCV[c] testing and current desire for testing			
Tested and known result (known)	331 (16.6)	146 (14.4)	185 (18.7)
Tested but unknown result (unknown)	28 (1.4)	19 (1.9)	9 (0.9)
Never tested, want to be (willing)	494 (24.7)	263 (26)	231 (23.4)
Never tested, do not want to be (unwilling)	706 (35.3)	358 (35.4)	348 (35.2)
Unsure if tested (unsure)	441 (22.1)	225 (22.3)	216 (21.8)

[a]JP ¥159=US $1.

[b]HBV: hepatitis B virus.

[c]HCV: hepatitis C virus

Characteristics of Web Services and Digital Advertising

The top 50 selected web services are presented in Multimedia Appendix 8. The top 10 selected services were the general e-commerce company Rakuten Ichiba (n=863), Google (n=740), Amazon (n=690), YouTube (n=669), the search engine Yahoo! Japan (n=658), Yahoo! News (n=624), Google Maps (n=589), Yahoo! Weather & Disaster (n=503), the travel reservation site Rakuten Travel (n=455), and the restaurant review site Tabelog (n=420).

The ranking of the 25 types of digital advertising is presented in Multimedia Appendix 9. The most frequent responses were “Nothing particular” at 53.1% (1062/2000), “FamilyMart in-store and storefront digital advertising” at 18.3% (365/2000), “Digital advertising inside trains” at 17.9% (358/2000), “Digital advertising at stations and inside station facilities” at 17.3% (346/2000), and “Seven-Eleven in-store and storefront digital advertising” at 16.1% (322/2000).

XSL•FO
RenderX

Correspondence Analysis Between Respondent Characteristics and Web Services or Digital Advertisements

The results of the correspondence analysis between sex and web services are presented in Multimedia Appendix 10. The specific web services used by male participants included internet technology services such as IT media and DMM.com, as well as newspaper sites such as Nikkei and Yomiuri. Among female participants, the most common services were cosmetic services and cooking recipe sites.

The results of the correspondence analysis between age groups and web services are shown in Multimedia Appendix 11. In their 20s, Uber Eats (food delivery), TikTok (video sharing), Spotify (music), and gaming services were prominent; in their 30s and 40s, Cosme (cosmetic shopping), Rakuten Fashion (general e-commerce), HOT PEPPER Beauty (beauty salon booking and review site), and Cookpad (recipe sharing and cooking website); in their 50s, Yahoo sites (search engine and relevant services) and Facebook; and in their 60s, Japanet (TV shopping), MSN Japan (online news and portal site), and Belluna (mail-order and online shopping site).

Correspondence Analysis Between HBV and HCV Testing Characteristics and Web Services or Digital Advertising

Figure 2 shows the 3 dimensions: correspondence analysis results for the respondents' HBV and HCV testing statuses, the 60 web services farthest from the origin, and all 25 digital advertisements. Web services specific to people who had not been tested and wanted to be tested were Lawson (convenience store) with a cosine θ of 0.989 and Cosme (cosmetic shopping) with a cosine θ of 0.987. Digital advertising targeted at individuals who had not been tested and wanted to be tested included in-store and storefront advertising at Welcia (pharmacy chain; cosine θ=0.994) and Lawson (convenience store; cosine θ=0.937).

Figure 2. Associations of viral testing history and the desire for testing with (A) web services and (B) digital advertising.

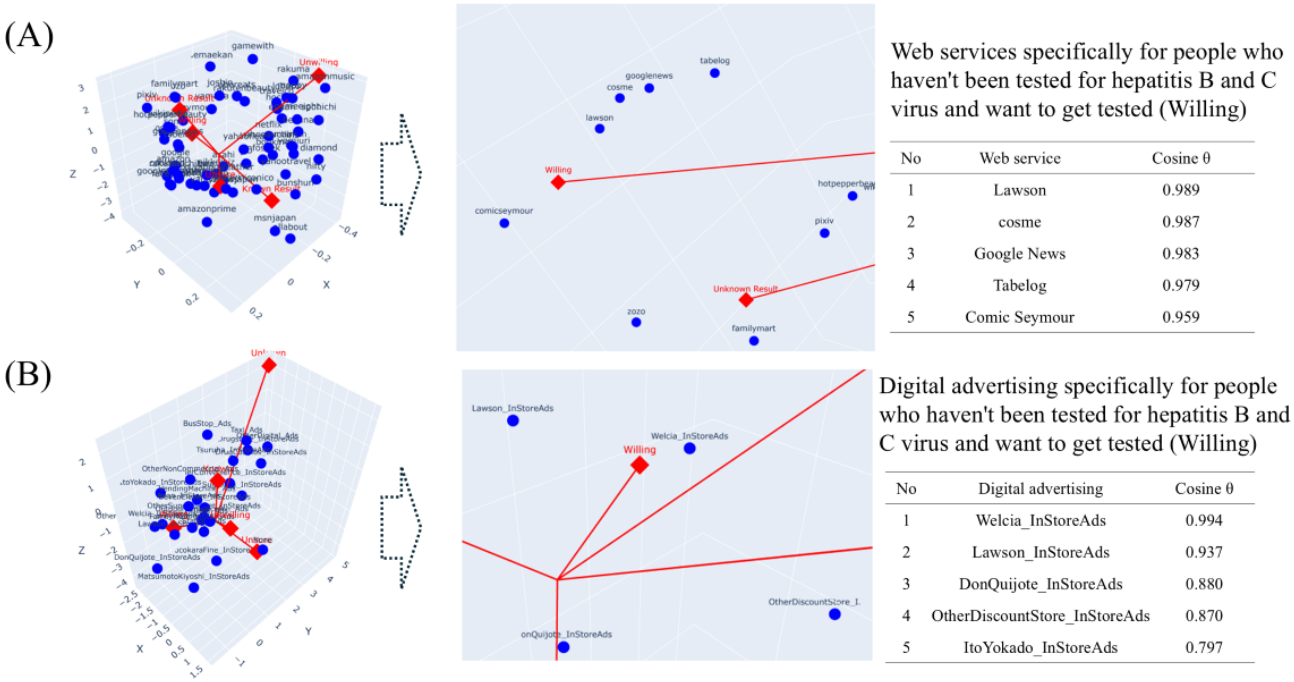

Web services specifically for people who haven't been tested for hepatitis B and C virus and want to get tested (Willing)

No	Web service	Cosine θ
1	Lawson	0.989
2	cosme	0.987
3	Google News	0.983
4	Tabelog	0.979
5	Comic Seymour	0.959

Digital advertising specifically for people who haven't been tested for hepatitis B and C virus and want to get tested (Willing)

No	Digital advertising	Cosine θ
1	Welcia_InStoreAds	0.994
2	Lawson_InStoreAds	0.937
3	DonQuijote_InStoreAds	0.880
4	OtherDiscountStore_InStoreAds	0.870
5	ItoYokado_InStoreAds	0.797

The results of the correspondence analysis between HBV and HCV testing status and web services by sex (male and female) and age group (<50 y or >50 y) are shown in Figures 3 and 4, respectively. Depending on sex and age group, web services specific to individuals who had not been tested and who wanted to be tested were identified. Among individuals who intended to undergo testing, the most frequently used were Nikkei Biz (news media) and Yahoo-related services (search engine) for male participants, Demae-can (food delivery) and e-book platforms (Comic Seymour and Mecha Comic) for female participants, Google-related services for those aged 20 to 49 years, and Zozotown (fashion e-commerce site) and Booking.com (hotel booking site) for those aged 50 to 69 years. In the correspondence analysis, the first 2 dimensions explained most of the variability in the distribution of web services: among male participants, 73.8% of the total inertia (56.3% on axis 1 and 17.5% on axis 2), and among female participants, 71.6% (51% on axis 1 and 20.6% on axis 2). For participants aged 20 to 49 years, the first 2 dimensions accounted for 82.1% of total inertia (70.5% on axis 1 and 11.6% on axis 2); for those aged 50 to 69 years, they accounted for 68% (40.7% on axis 1 and 27.3% on axis 2).

XSL•FO
RenderX

Figure 3. Relationship between hepatitis B virus (HBV) or hepatitis C virus (HCV) screening experience and web services by sex. (A) Results for male participants. (B) Results for female participants.

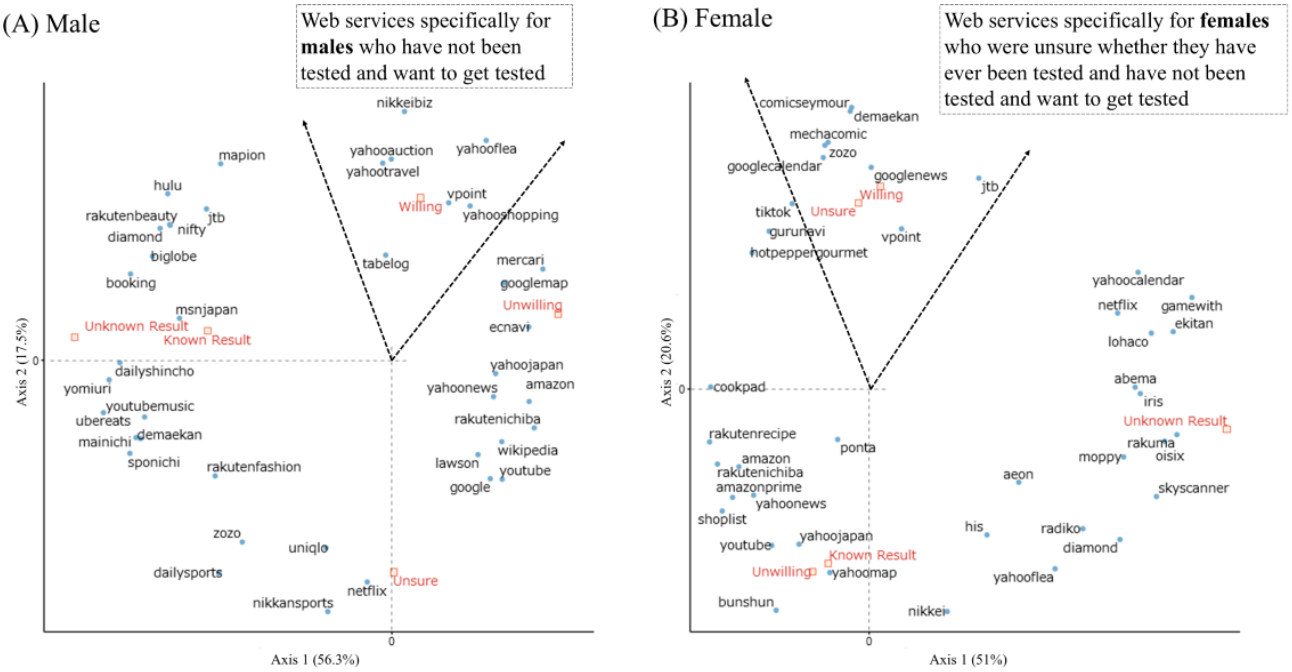

Figure 4. Relationship between hepatitis B virus (HBV) or hepatitis C virus (HCV) screening experience and web services by age groups. (A) Results for respondents aged 20 to 49 years. (B) Results for respondents aged 50 to 69 years.

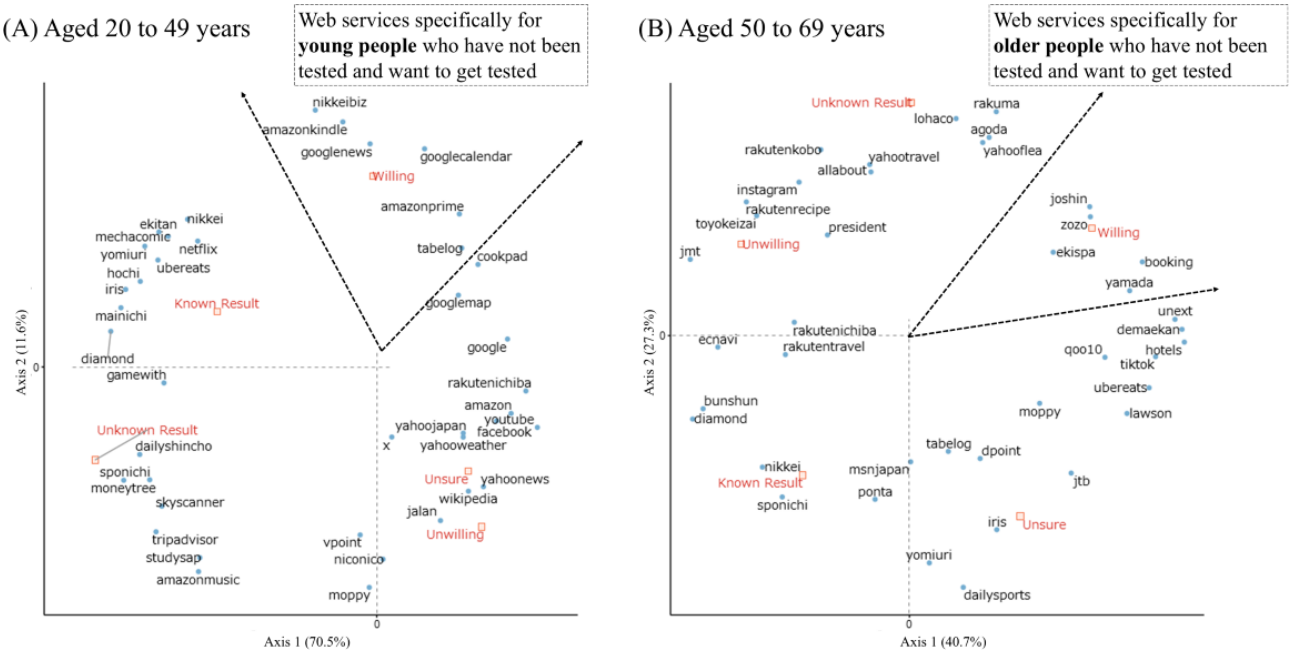

Sensitivity Analysis

The findings from the sensitivity analysis of web services using correspondence analysis are detailed in Multimedia Appendix 12. Web services identified as particularly relevant for individuals who had not been tested but wanted to be included are Google News (cosine θ=0.98), Zozotown—a fashion e-commerce platform—(cosine θ=0.96), Nikkei Biz (news media; cosine θ=0.94), Comic Seymour (e-book platform; cosine θ=0.92), and Lawson (convenience store; cosine θ=0.90).

Similarly, the outcomes of the sensitivity analysis of digital advertising channels indicated that Lawson in-store and storefront advertisements (cosine θ=0.96), as well as Matsumoto Kiyoshi in-store and storefront advertisements (cosine θ=0.92), were specifically associated with individuals who had not been tested but wanted to be (Multimedia Appendix 13).

The results of the sensitivity analysis of the web services using a modified Poisson regression revealed a significant association for Lawson (Multimedia Appendix 14). The highest PRs for the desire for hepatitis testing were observed for Lawson (PR 1.75, 95% CI 1.22 - 2.52), Zozotown (PR 1.67, 95% CI 1.18 - 2.36), Google Calendar (PR 1.63, 95% CI 1.21 - 2.20), and Google News (PR 1.61, 95% CI 1.25 - 2.07).

XSL•FO
RenderX

The sensitivity analysis of digital advertising channels, using modified Poisson regression, demonstrated that the highest PRs were observed for in-store advertisements at other discount stores (PR 2.64, 95% CI 1.51 - 4.59), Ito Yokado (PR 1.72, 95% CI 1.13 - 2.63), DonQuijote (PR 1.72, 95% CI 1.18 - 2.51), Welcia (PR 1.71, 95% CI 1.13 - 2.59), and Lawson (PR 1.49, 95% CI 1.19 - 1.87; Multimedia Appendix 15).

Discussion

Principal Findings

We identified web services and digital advertisements for use in HBV and HCV campaign channels, based on testing history and expressed interest in testing. Individuals who had never been tested for hepatitis but wanted to be tested were regularly exposed to the website and in-store digital advertisements of Lawson, one of the most famous convenience store chains. Segmenting the target population by sex or age group revealed distinct preferences for web services. These observations suggest that effective digital public health campaigns should account for specific attributes of the target population, including their testing history and interests, when determining appropriate communication channels.

Comparison With Prior Work

Lawson might be crucial in conducting viral hepatitis awareness campaigns, as individuals interested in HBV and HCV testing use Lawson's web services and digital advertisements. In 2013, Lawson collaborated with local medical and public health organizations to distribute application forms for free hepatitis testing at selected stores in Saitama Prefecture, illustrating the potential of convenience store chains as access points for hepatitis screening campaigns [27]. Convenience stores are gaining attention as venues for health promotion. For example, initiatives have been undertaken to improve access to healthy foods by stocking them in these stores [28,29]. Convenience stores are key locations within the community, and staff education has been implemented to strengthen the community network for dementia [30]. These initiatives focused on location characteristics, such as the presence of unhealthy food and local staff members. The results of this study enable us to consider intervention settings from different perspectives by designing interventions tailored to the characteristics of visiting customers.

In public health communication, it is preferable to tailor information dissemination to the characteristics and interests of the target audience [31]. As advertising is expensive, cost-effective campaigns can be implemented by targeting those who have never undergone an examination but are willing to be tested. Several interventions have been tailored to specific target groups based on their characteristics [32]. An educational intervention using an app to provide information about HBV screening in their native language to Asian Americans resulted in increased consultations with health care providers and subsequent testing [33]. In the United Kingdom, a combination intervention of sending tailored letters to encourage high-risk and untreated patients with HCV to seek care, training staff, and implementing awareness-raising notices increased HCV testing and treatment, demonstrating good cost-effectiveness [34]. In an intervention in New York, a multidisciplinary team implemented efforts to remove barriers to care for low-income individuals, the homeless, and drug users who needed support, leading many to undergo HCV testing and complete treatment [35]. Consistent with these results, community-based interventions at convenience stores may be a promising approach.

When developing advertising strategies for public health, it is important to consider the sex and age of the target population. The results of this study revealed distinct patterns in web service usage and digital advertising exposure by sex and age. Therefore, stratified by sex and age, the characteristic channels differed according to viral hepatitis testing status. Viral hepatitis testing is recommended at least once in a lifetime, regardless of sex or age [36]. For this awareness campaign, it was preferable to use the analysis results based on the entire target population. When focusing on specific age groups or sex, awareness strategies should be developed based on analysis results adjusted for the characteristics of that group. For example, when implementing HBV awareness campaigns targeting pregnant female individuals, awareness campaigns should be developed by referencing our results in female participants.

Fewer respondents than expected reported having been tested for viral hepatitis. In Japan, it is estimated that more than half of the population has been tested for viral hepatitis [6], far exceeding the 18% testing rate in this study; however, an additional 22.1% were unsure of whether they had been tested. A study investigating Japanese people's knowledge of how viral hepatitis testing is performed reported that 93% of respondents identified blood tests, whereas only 7% mentioned urine tests [37]. In addition, physicians do not inform patients of negative results. A survey targeting ophthalmologists who frequently conducted viral hepatitis tests before surgery revealed that approximately 40% did not explain the results to patients when the test was negative [38]. Even when negative results were verbally communicated to patients, only 42% recalled them 1 year later, leading to the recommendation of using negative cards—written records certifying that the individual tested negative for HBV and HCV—for communicating results [39]. Negative cards are expected to enable information sharing with other health care professionals and to avoid duplicate viral hepatitis testing.

This study neither implemented nor evaluated specific public health awareness campaigns. Instead, we present a methodological case study illustrating how correspondence analysis can inform the development of public health strategies that leverage web services and digital advertising. Medical professionals have used correspondence analysis to address critical issues through dimension reduction, data simplification, clustering, and classification [40]. This study visualized the distributions of 180 web services and 25 digital advertisements. Correspondence analysis has been used to clarify the relationship between age and headache type, to identify associations between personality type and various diagnoses [41], and to clarify the key challenges prioritized by each profession when pursuing innovation in public hospitals [42]. On the basis of these examples, correspondence analysis may be useful as an exploratory tool for examining and planning public health

XSL•FO
RenderX

advertising strategies. A sensitivity analysis using multivariate models adjusted for respondent characteristics yielded patterns similar to those observed in the main correspondence analysis. This consistency indicates that the study's findings are robust to potential confounding from participant attributes.

In this study, respondents tended to have higher educational attainment, higher household income, and a greater proportion of managerial occupations than the working population estimated from the 2020 Population Census. This is consistent with prior research showing that online surveys often include individuals with higher education and income levels than those in national statistics [43,44]. Higher socioeconomic status has been associated with better health literacy, which may explain the interest in hepatitis testing [45]. Moreover, because web service use was assessed among individuals who chose to participate in an online survey, respondents might be more familiar with online services. Thus, the level of web service use in this study might be higher than that in the general population, and our findings reflect patterns among digitally engaged individuals.

Limitations

This study has certain limitations. First, potential selection bias warrants consideration. Compared with national statistics, our study included a higher proportion of individuals with university-level education, lower annual income, and a higher proportion in managerial or professional roles than in production-related occupations. Furthermore, respondents were more likely to reside in urban areas. Given that increased educational attainment, income, and occupational status may correlate with heightened health awareness and interest in viral hepatitis testing, the observed associations may not be generalizable to the broader population. Thus, the findings must be interpreted with caution when extending them beyond digitally engaged adults. Future studies should address this issue by using non–internet-based recruitment methods to enhance external validity.

Second, measurement biases regarding web services and digital advertising exist. To mitigate this effect, we conducted a correspondence analysis of web service usage by sex and age, revealing patterns consistent with real-world behaviors—such as female individuals preferring cosmetic sites and younger individuals favoring TikTok. Subsequent research should seek to collect objective data regarding web service use.

Third, recall bias may have influenced the reporting of HBV or HCV testing history. Notably, 22.1% of respondents indicated uncertainty about previous testing, which may have led to misclassification of past testing status. Additionally, exposure misclassification may have occurred because digital advertising exposure was self-reported and may not accurately reflect media saturation or individual engagement with advertisements.

Finally, we were unable to verify the effectiveness of advertising campaigns in changing behavior. Further investigation is necessary to determine whether interventions that incorporate appropriate web services and digital advertising can facilitate viral hepatitis testing, detection, and treatment.

Conclusions

This study identified distinct channel-specific patterns in the promotion of viral hepatitis testing in Japan. Lawson consistently served as a primary point of contact across web services and digital advertising among respondents who had never been tested but expressed interest in testing. Subgroup analyses revealed differences by sex and age, while sensitivity analyses confirmed comparable overall trends. These results support the development of targeted digital communication strategies and offer a practical framework for designing effective public health campaigns for HBV or HCV testing.

Acknowledgments

The authors acknowledge the support of Rakuten Insight Inc for encouraging recruitment and conducting a web-based survey using online panels. The authors are grateful to Editage for reviewing this manuscript for English language and clarity. Additionally, the authors used ChatGPT (GPT-5.2 Thinking) to assist with English-language editing [46].

Funding

This study was supported by the Ministry of Health, Labour and Welfare, Japan (grant: 23HC1001).

Data Availability

The datasets used and analyzed in this study are not publicly accessible due to confidentiality agreements; however, additional information regarding the study is available from the corresponding author upon reasonable request.

Authors' Contributions

K Sakai and MT conceptualized and designed this study. K Sakai collected the data and performed the statistical analysis. KF, YF, SN, K Sano, and MK contributed to the interpretation of the results. K Sakai wrote the manuscript. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
List of 180 types of web services.
[DOCX File, 44 KB - jmir_v28i1e89585_app1.docx]

Multimedia Appendix 2
List of 25 types of digital advertising.
[DOCX File, 23 KB - jmir_v28i1e89585_app2.docx]

Multimedia Appendix 3
Variables used in modified Poisson regression models for digital exposure and desire for hepatitis virus testing.
[DOCX File, 25 KB - jmir_v28i1e89585_app3.docx]

Multimedia Appendix 4
Distribution of educational background in this study and the Census 2020.
[DOCX File, 22 KB - jmir_v28i1e89585_app4.docx]

Multimedia Appendix 5
Distribution of household income in this study and the Census 2020.
[DOCX File, 34 KB - jmir_v28i1e89585_app5.docx]

Multimedia Appendix 6
Distribution of occupation and industry in this study and the census 2020.
[DOCX File, 27 KB - jmir_v28i1e89585_app6.docx]

Multimedia Appendix 7
Distribution of residential regions in this study and the 2020 Population Census.
[DOCX File, 33 KB - jmir_v28i1e89585_app7.docx]

Multimedia Appendix 8
The top-selected 50 web services.
[DOCX File, 26 KB - jmir_v28i1e89585_app8.docx]

Multimedia Appendix 9
The selection of 25 digital advertising.
[DOCX File, 24 KB - jmir_v28i1e89585_app9.docx]

Multimedia Appendix 10
Correspondence analysis between sex and web services.
[DOCX File, 1144 KB - jmir_v28i1e89585_app10.docx]

Multimedia Appendix 11
Correspondence analysis between age groups and web services.
[DOCX File, 1776 KB - jmir_v28i1e89585_app11.docx]

Multimedia Appendix 12
Sensitivity analysis of web services based on 3D correspondence analysis.
[DOCX File, 308 KB - jmir_v28i1e89585_app12.docx]

Multimedia Appendix 13
Sensitivity analysis of digital advertising channels based on 3D correspondence analysis.
[DOCX File, 284 KB - jmir_v28i1e89585_app13.docx]

Multimedia Appendix 14
Associations between desire for hepatitis virus testing and use of individual web services.
[DOCX File, 46 KB - jmir_v28i1e89585_app14.docx]

XSL•FO
RenderX

Multimedia Appendix 15
Associations between desire for hepatitis virus testing and exposure to individual digital advertising channels.
[DOCX File, 49 KB - jmir_v28i1e89585_app15.docx]

Checklist 1
STROBE checklist.
[DOCX File, 123 KB - jmir_v28i1e89585_app16.docx]

Checklist 2
CHERRIES checklist.
[DOCX File, 23 KB - jmir_v28i1e89585_app17.docx]

References

1. Global health sector strategies on, respectively, HIV, viral hepatitis and sexually transmitted infections for the period 2022-2030. : World Health Organization; 2022 URL: https://iris.who.int/server/api/core/bitstreams/6b18ac34-f56a-466e-9660-b5653cb52ef0/content [accessed 2026-03-13]
2. Ikeda M, Mitsunaga S, Shimizu S, et al. Current status of hepatocellular carcinoma in Japan. Chin Clin Oncol 2013 Dec;2(4):40. [doi: 10.3978/j.issn.2304-3865.2013.09.01] [Medline: 25841919]
3. Nakano M, Yatsuhashi H, Bekki S, et al. Trends in hepatocellular carcinoma incident cases in Japan between 1996 and 2019. Sci Rep 2022 Jan 27;12(1):1517. [doi: 10.1038/s41598-022-05444-z] [Medline: 35087153]
4. Tanaka J, Kurisu A, Ohara M, et al. Burden of chronic hepatitis B and C infections in 2015 and future trends in Japan: a simulation study. Lancet Reg Health West Pac 2022 May;22:100428. [doi: 10.1016/j.lanwpc.2022.100428] [Medline: 35637862]
5. Okushin K, Kanto T, Korenaga M, et al. Real‐world trends in acute viral hepatitis in Japan: a nationwide questionnaire‐based survey. Hepatol Res 2025 Mar;55(3):301-308. [doi: 10.1111/hepr.14123]
6. Korenaga M, Kanto T. Testing, diagnosis of viral hepatitis, and the follow-up policy in Japan. Glob Health Med 2021 Oct 31;3(5):308-313. [doi: 10.35772/ghm.2021.01072] [Medline: 34782874]
7. Tatemichi M, Furuya H, Nagahama S, et al. A nationwide cross-sectional survey on hepatitis B and C screening among workers in Japan. Sci Rep 2020 Jul 10;10(1):11435. [doi: 10.1038/s41598-020-68021-2] [Medline: 32651421]
8. Sasaki N, Wada K, Smith DR, Wang G, Ohta H, Shibuya A. Hepatitis screening in Japanese individuals of working age and prejudice against infected persons in the workplace. J Occup Health 2014;55(5):392-397. [doi: 10.1539/joh.12-0266-fs] [Medline: 23812028]
9. Request for cooperation regarding measures against viral hepatitis in the workplace [Report in Japanese]. : Ministry of Health, Labour and Welfare; 2023 URL: https://www.mhlw.go.jp/content/10900000/001076013.pdf [accessed 2026-03-13]
10. Sakai K, Nakazawa S, Fukai K, Furuya Y, Korenaga M, Tatemichi M. Marketing strategies for promoting workplace hepatitis B and C virus testing: a cross-sectional study using correspondence analysis in Japan. Front Public Health 2025;13:1522850. [doi: 10.3389/fpubh.2025.1522850] [Medline: 40115348]
11. An J, Kwak H, Qureshi HM, Weber I. Precision public health campaign: delivering persuasive messages to relevant segments through targeted advertisements on social media. JMIR Form Res 2021 Sep 24;5(9):e22313. [doi: 10.2196/22313] [Medline: 34559055]
12. Takeuchi Y, Ohara M, Kanto T. Nationwide awareness-raising program for viral hepatitis in Japan: the “Shitte kan-en” project. Glob Health Med 2021 Oct 31;3(5):301-307. [doi: 10.35772/ghm.2021.01063] [Medline: 34782873]
13. Dietrich T. Segmentation in social marketing: five steps to success. In: Dietrich T, Rundle-Thiele S, Kubacki K, editors. Segmentation in Social Marketing: Springer; 2017:77-92.
14. Etoori D, Desai M, Mandal S, Rosenberg W, Sabin CA. A scoping review of media campaign strategies used to reach populations living with or at high risk for hepatitis C in high income countries to inform future national campaigns in the United Kingdom. BMC Infect Dis 2023 Sep 26;23(1):629. [doi: 10.1186/s12879-023-08603-3] [Medline: 37752434]
15. Talal AH, Dharia A, Markatou M, et al. Facilitated telemedicine as a patient-centered, sociotechnical intervention to integrate hepatitis C treatment into opioid treatment programs and overcome the digital divide among underserved populations: qualitative study. JMIR Public Health Surveill 2025 Jul 16;11:e68854. [doi: 10.2196/68854] [Medline: 40669057]
16. Willem T, Zimmermann BM, Matthes N, Rost M, Buyx A. Acceptance of social media recruitment for clinical studies among patients with hepatitis B: mixed methods study. J Med Internet Res 2024 Aug 26;26:e54034. [doi: 10.2196/54034] [Medline: 39186322]
17. Population census [Article in Japanese]. e-Stat. 2020. URL: https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00200521&tstat=000001136464&cycle=0&year=20200&month=24101210&tclass1=000001136466 [accessed 2025-12-14]

XSL•FO
RenderX

18. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med 2007 Oct 16;147(8):573-577. [doi: 10.7326/0003-4819-147-8-200710160-00010] [Medline: 17938396]
19. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
20. Japan Standard Occupational Classification (rev. 5th, December 2009) structure and explanatory notes [Website in Japanese]. Ministry of Internal Affairs and Communications. 2009. URL: https://www.soumu.go.jp/english/dgpp_ss/seido/shokgyou/co09-4.htm [accessed 2026-03-13]
21. Japan Standard Industrial Classification (rev. 13, October 2013) the underlying principles of the classification [Website in Japanese]. Ministry of Internal Affairs and Communications. 2013. URL: https://www.soumu.go.jp/english/dgpp_ss/seido/sangyo/san13-2.htm [accessed 2025-12-14]
22. Beh EJ, Lombardo R. Correspondence Analysis: Theory, Practice and New Strategies: John Wiley & Sons; 2014.
23. Costa PS, Santos NC, Cunha P, Cotter J, Sousa N. The use of multiple correspondence analysis to explore associations between categories of qualitative variables in healthy ageing. J Aging Res 2013;2013:302163. [doi: 10.1155/2013/302163] [Medline: 24222852]
24. Higuchi K. New quantitative text analytical method and KH Coder software. Jpn Sociol Rev 2017;68(3):334-350. [doi: 10.4057/jsr.68.334]
25. Hajek A, Kretzler B, König HH. Determinants of healthcare use based on the Andersen model: a systematic review of longitudinal studies. Healthcare (Basel) 2021 Oct 12;9(10):1354. [doi: 10.3390/healthcare9101354] [Medline: 34683034]
26. Core Team R. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2024. URL: https://www.r-project.org/ [accessed 2026-03-19]
27. News release. Lawson, Inc. URL: https://mldata.lawson.co.jp/company/news/year/077593/ [accessed 2025-12-14]
28. Haboush-Deloye AL, Knight MA, Bungum N, Spendlove S. Healthy foods in convenience stores: benefits, barriers, and best practices. Health Promot Pract 2023 May;24(1_suppl):108S-111S. [doi: 10.1177/15248399221147878] [Medline: 36999507]
29. Jernigan VBB, Williams M, Wetherill M, et al. Using community-based participatory research to develop healthy retail strategies in Native American–owned convenience stores: the THRIVE study. Prev Med Rep 2018 Sep;11:148-153. [doi: 10.1016/j.pmedr.2018.06.012] [Medline: 30003014]
30. Igarashi A, Matsumoto H, Takaoka M, Kugai H, Suzuki M, Yamamoto-Mitani N. Educational program for promoting collaboration between community care professionals and convenience stores. J Appl Gerontol 2020 Jul;39(7):760-769. [doi: 10.1177/0733464819871878] [Medline: 31478434]
31. Schmid KL, Rivers SE, Latimer AE, Salovey P. Targeting or tailoring? Maximizing resources to create effective health communications. Mark Health Serv 2008;28(1):32-37. [Medline: 18389854]
32. Rajkumar V, McCausland K, Lobo R. A rapid review of interventions to increase hepatitis B testing, treatment, and monitoring among migrants living in Australia. Int J Environ Res Public Health 2022 May 13;19(10):5947. [doi: 10.3390/ijerph19105947] [Medline: 35627481]
33. Khalili M, Kim NJ, Tsoh JY, et al. Health Within Reach-a patient-centered intervention to increase hepatitis B screening among Asian Americans: a randomized clinical trial. J Gen Intern Med 2022 Oct;37(13):3242-3250. [doi: 10.1007/s11606-021-07232-3] [Medline: 34993863]
34. Roberts K, Macleod J, Metcalfe C, et al. Cost effectiveness of an intervention to increase uptake of hepatitis C virus testing and treatment (HepCATT): cluster randomised controlled trial in primary care. BMJ 2020 Feb 26;368:m322. [doi: 10.1136/bmj.m322] [Medline: 32102782]
35. Ford MM, Johnson N, Desai P, Rude E, Laraque F. From care to cure: demonstrating a model of clinical patient navigation for hepatitis C care and treatment in high-need patients. Clin Infect Dis 2017 Mar 1;64(5):685-691. [doi: 10.1093/cid/ciw806] [Medline: 27940945]
36. Basic guidelines for promotion of control measures for hepatitis. Ministry of Health, Labour and Welfare. 2011. URL: https://www.mhlw.go.jp/bunya/kenkou/kekkaku-kansenshou09/pdf/hourei-17e.pdf [accessed 2025-12-14]
37. Masuda S, Hidaka I, Nakano S, et al. Knowledge enlightenment about viral hepatitis using a quiz at an awareness-raising event. Kanzo 2025;66(10):443-446. [doi: 10.2957/kanzo.66.443]
38. Ohara M, Ide T, Inoue J, et al. Follow-up survey on hepatitis virus testing among ophthalmologists. Kanzo 2025;66(9):407-410. [doi: 10.2957/kanzo.66.407]
39. Aida M, Ikegami T, Korenaga M. Methods of communicating negative results of hepatitis virus tests. Kanzo 2023;64(6):259-269. [doi: 10.2957/kanzo.64.259]
40. Žlahtič B, Kokol P, Blažun Vošner H, Završnik J. The role of correspondence analysis in medical research. Front Public Health 2024;12:1362699. [doi: 10.3389/fpubh.2024.1362699] [Medline: 38584915]
41. Greenacre M. Correspondence analysis in medical research. Stat Methods Med Res 1992;1(1):97-117. [doi: 10.1177/096228029200100106] [Medline: 1341654]
42. Jończyk JA, Olszewska AM. The use of correspondence analysis in assessing the antecedents of innovativeness in public hospitals. Stud Log Gramm Rhetor 2016 Dec 1;47(1):85-101. [doi: 10.1515/slgr-2016-0048]

XSL•FO
RenderX

43. Tsuboi S, Yoshida H, Ae R, Kojo T, Nakamura Y, Kitamura K. Selection bias of internet panel surveys: a comparison with a paper-based survey and national governmental statistics in Japan. Asia Pac J Public Health 2015 Mar;27(2):NP2390-NP2399. [doi: 10.1177/1010539512450610] [Medline: 22743862]
44. Szolnoki G, Hoffmann D. Online, face-to-face and telephone surveys—comparing different sampling methods in wine consumer research. Wine Econ Policy 2013 Dec;2(2):57-66. [doi: 10.1016/j.wep.2013.10.001]
45. Kagamimori S, Gaina A, Nasermoaddeli A. Socioeconomic status and health in the Japanese population. Soc Sci Med 2009 Jun;68(12):2152-2160. [doi: 10.1016/j.socscimed.2009.03.030] [Medline: 19375838]
46. Introducing GPT-5.2. OpenAI. URL: https://openai.com/index/introducing-gpt-5-2/ [accessed 2025-12-11]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
HBV: hepatitis B virus
HCV: hepatitis C virus
PR: prevalence ratio
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
WHO: World Health Organization

Edited by C Argyropoulos; submitted 15.Dec.2025; peer-reviewed by E Etafo, HS Maan, MZ Hossain; revised version received 23.Feb.2026; accepted 25.Feb.2026; published 02.Apr.2026.

Please cite as:
Sakai K, Nakazawa S, Furuya Y, Fukai K, Sano K, Korenaga M, Tatemichi M
Strategy for Hepatitis B and C Virus Testing Campaigns Through Web Services and Digital Advertising in Japan: Nationwide Cross-Sectional Study With Correspondence Analysis
J Med Internet Res 2026;28:e89585
URL: https://www.jmir.org/2026/1/e89585
doi:10.2196/89585

© Kosuke Sakai, Shoko Nakazawa, Yuko Furuya, Kota Fukai, Kei Sano, Masaaki Korenaga, Masayuki Tatemichi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Factors Influencing Universal Coverage of AI-Assisted Cervical Cancer Screening: Qualitative Study Based on the Macro Model of Health System

Lu Ji[1*], PhD; Xinke Zhou[2*], MD; Lan Yao[2], PhD

[1]School of Nursing, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

[2]School of Medicine and Health Management, Tongji Medical College, Huazhong University of Science and Technology, 13 HangKong Road, Wuhan, China

[*]these authors contributed equally

Corresponding Author:
Lan Yao, PhD
School of Medicine and Health Management, Tongji Medical College, Huazhong University of Science and Technology, 13 HangKong Road, Wuhan, China

Abstract

Background: Improving screening coverage is a central goal of the global strategy to eliminate cervical cancer. In resource-constrained settings, insufficient service accessibility remains a key barrier to expanding coverage. Supported by artificial intelligence (AI)–assisted diagnostic technology, Hubei province has pioneered China's first provincial-level population-wide cervical cancer screening program, serving 12.67 million eligible women. This initiative provides an innovative practice for addressing such challenges.

Objective: This study systematically examines major factors influencing the achievement of universal screening coverage targets through interviews with core managers and implementers of Hubei province's screening program. It aims to provide empirical evidence and strategic recommendations for applying AI technologies in cervical cancer screening and enhancing screening coverage rates.

Methods: The interview guide was developed under the guidance of the macro model of health system. A combination of purposive sampling and multistage stratified sampling was used to capture provincial-level overviews and understand regional implementation variations, respectively. Guided by the macro model of health system, interview outlines were developed. Semistructured interviews were conducted between January and August 2024 with key project personnel (one per institution) from 14 relevant institutions. Interview data were analyzed using thematic analysis, with systematic coding and management facilitated by the NVivo software.

Results: Key informants reported that comprehensive screening has been largely achieved. The analysis identified government stewardship, AI-assisted screening technology, screening funding, and health literacy as the major factors for achieving universal screening coverage. Among these, government leadership and the application of AI-assisted diagnostic technologies provide significant driving factors. Additional factors encompassed structural dimensions, including multisectoral coordination, trained screening technicians, and information systems; process dimensions, such as institutional service delivery capacity, quality control measures, and community mobilization; along with outcome dimensions comprising population coverage, cytology positivity rate, follow-up, and treatment rate.

Conclusions: Achieving large-scale cervical cancer screening requires coordinated efforts across four dimensions: government stewardship, screening technology, screening funding, and health literacy. Government stewardship served as the core driver in advancing population-wide screening coverage. Its mechanisms included coordinated procurement of AI-assisted screening services, secured financial investment, formulation of targeted policies, promotion of multi-sectoral collaboration, and optimization of service delivery models. These efforts systematically improved the accessibility and utilization of screening services, ultimately encouraging and facilitating active participation among residents.

(J Med Internet Res 2026;28:e75372) doi:10.2196/75372

KEYWORDS

cervical cancer screening; screening; qualitative research; health systems; artificial intelligence; AI

XSL•FO
RenderX

Introduction

Cervical cancer remains one of the major health threats for women worldwide. In 2022, there were approximately 660,000 new cases globally, accounting for 3.3% of total cancer cases, with a mortality rate as high as 56% [1]. Data from the China Cancer Registry Annual Report 2022 show a cervical cancer incidence rate of 13.8 per 100,000 women, with approximately 150,000 new cases annually [2]. In recent years, both incidence and mortality rates have shown an upward trend [3]. The World Health Organization launched a global strategy in 2020 aiming to eliminate cervical cancer through human papillomavirus vaccination, cervical cancer screening, and treatment [4]. Cervical cancer screening is a key focus of policy interventions in China. The current screening rate is approximately 36.8%; however, China aims to achieve a 70% screening rate by 2030 [5].

Improving screening coverage depends on multiple factors, particularly in resource-limited settings where accessible and affordable screening technologies are critically needed [6,7]. Significant disparities in screening coverage exist across nations and regions [8-11], influenced by a combination of factors including sociocultural norms, economic status, health care policies, health awareness, educational resources, and technological capabilities [12-17]. In low- and middle-income countries (LMICs), women commonly face multiple barriers, including limited health literacy, inadequate access to services, insufficient funding, and scarce medical resources, which in turn constrain the improvement of screening coverage rates [18-21]. Both the World Health Organization and the National Institutes of Health recommend the use of artificial intelligence (AI) technology in resource-limited settings to effectively improve access to screening services.

AI-assisted diagnostic technology offers advantages in accessibility, cost, and diagnostic efficiency. Leveraging these strengths, China has implemented its first provincial-level cervical cancer screening program with universal coverage. The shortage of pathologists makes it difficult to implement traditional cytological screening effectively in primary health care institutions, limiting its capacity to adequately meet the population's screening needs [22,23]. In Hubei province, the adoption of AI-assisted cervical cancer screening technology achieved a coverage rate of 13.45% within 6 months. Calculated with the same number of pathologists, its efficiency was 87.5 times that of manual slide interpretation [24,25]. The AI-assisted approach showed remarkable diagnostic accuracy and cost-effectiveness, with an average per-person cost of 49 RMB [26-28]. This innovative technology has significantly enhanced screening coverage, particularly in resource-limited settings.

Current research on AI technology predominantly focuses on the diagnostic performance of screening techniques, with a notable scarcity of studies investigating its large-scale implementation and associated influencing factors [29-33]. The macro model of health system (MMHS) applies systems thinking to analyze the interplay and interdependence among various components within a health system, making it well-suited for policy interventions and health system diagnostics. This study developed interview guidelines based on the MMHS theoretical framework and conducted interviews with managers and health care providers involved in Hubei's cervical cancer screening program. The aim was to identify the factors influencing the achievement of universal coverage screening, thereby providing reference suggestions for the wider application of AI-assisted diagnostic technology. Given the core bottleneck of limited cervical cancer screening services due to insufficient pathologist resources in primary health care institutions, this study proposes the following exploratory hypothesis: the introduction of cost-effective and quality-reliable AI-assisted screening technology is expected to substantially overcome the primary barriers to achieving universal screening coverage, thereby accelerating the attainment of coverage targets. Meanwhile, maintaining an open research stance, we developed a theory-driven interview guide to capture the diverse complexities and obstacles that may arise during the implementation of universal coverage programs.

Methods

Overview

This study adopts a qualitative approach following Consolidated Criteria for Reporting Qualitative Research (COREQ), the standardized and widely recognized checklist for qualitative research (Checklist 1), and conducted semistructured interviews with managers and implementers of cervical cancer screening programs from 14 institutions. To systematically explore the factors influencing universal cervical cancer screening, the interview guide was developed based on the MMHS. Thematic analysis was performed following a 6-step analytic framework [34], through which textual data were integrated, categorized, and synthesized to ultimately identify the factors affecting the achievement of universal cervical cancer screening coverage. The researchers' reflections on the predefined statements are presented in detail in the *Discussion* section.

Macro Model of Health System

Framework

The macro model of health system (MMHS) is a theoretical and methodological framework that applies systems thinking to analyze the interactions and interdependencies among the various components within a health system and to reveal its operational patterns [35]. This model integrates systems science, complexity theory, and the theory of social determinants of health to comprehensively explain the combined impact of multiple factors on the health system. The model consists of 2 subsystems: an external submodel and an internal submodel. The external submodel encompasses external factors affecting health, such as politics, economy, culture, population needs, biology, environment, and behavior. The internal submodel represents the internal drivers of health, following the logical framework of structure-process-outcome (Figure 1). Within this framework, administrative, organizational, and resource factors at the structural level influence service delivery processes, which in turn affect system outcomes and ultimately determine health outcomes.

XSL•FO
RenderX

The MMHS is applicable to policy intervention research and health system diagnostics. The comprehensive cervical cancer screening program represents a concrete manifestation of public health policy. The implementation of cervical cancer screening programs is influenced by external factors, such as the policy environment and economic conditions, as well as internal factors including administrative processes, organizational structures, and resource allocation affecting service provision and utilization. These factors ultimately impact screening coverage and population cervical health outcomes. Therefore, this study uses the theoretical framework of the MMHS to design research tools and diagnose the factors influencing the implementation of universal cervical cancer screening.

Figure 1. Conceptual framework of the macro model of health system.

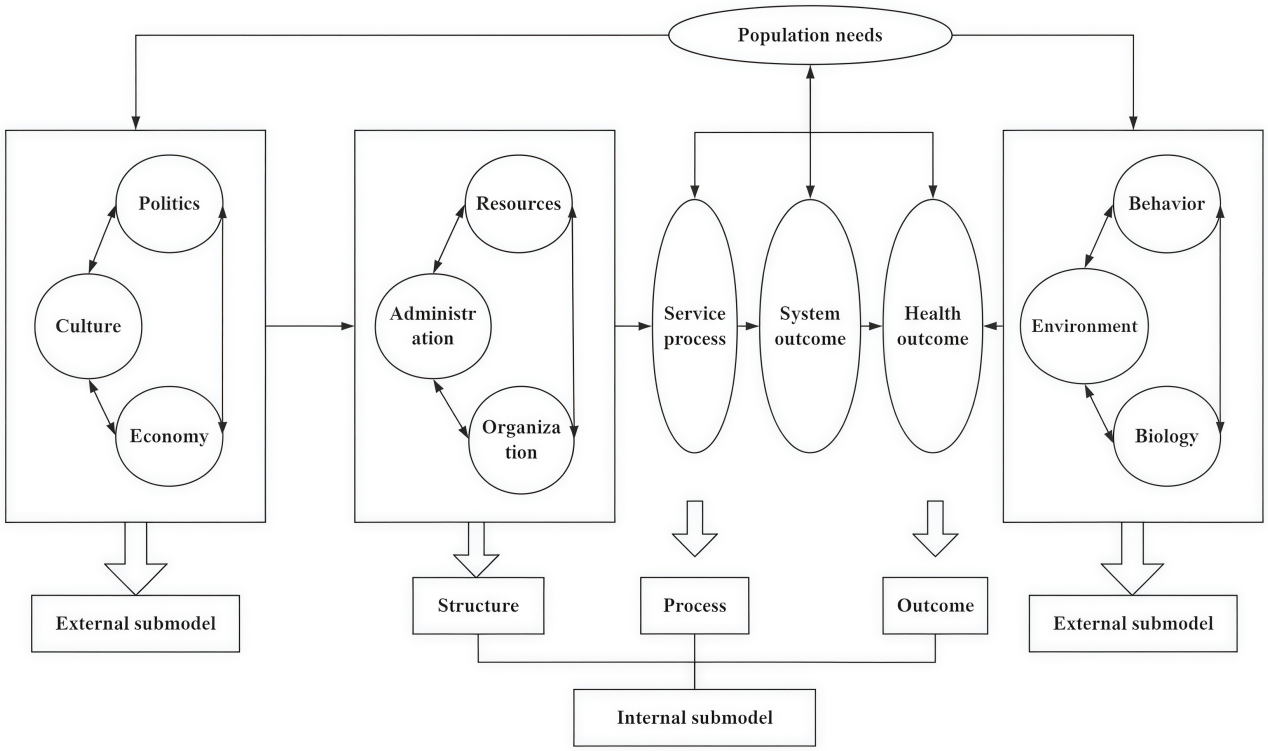

AI-Assisted Cervical Cancer Screening

AI-assisted diagnostic technology has been extensively implemented in China's free cervical cancer screening programs and has facilitated the completion of the nation's first provincial-level universal cervical cancer screening project, which provided screening services to approximately 12.67 million eligible women within 3 years. The most notable advantage of this technology lies in its ability to deliver cervical cancer screening of equivalent quality with high efficiency and low cost in resource-limited settings, thereby significantly advancing health equity.

Within the screening workflow, primary health care institutions are only required to collect cervical exfoliated cell samples from participants. The samples are then transported via logistics to laboratories, where they undergo digitization to generate cellular images uploaded to a cloud platform. Diagnosis is performed through a collaborative effort involving AI algorithms, cytotechnicians, and pathologists. Screening participants can quickly access their test results by scanning a QR code [24].

Study Participants

This study was conducted in Hubei province, China, from January to August 2024. A combination of purposive sampling and multistage stratified sampling was used to obtain an information-rich and representative sample. Hubei province was selected as the study area because it is the first region in China to implement province-wide universal cervical cancer screening.

The sampling procedure was conducted in 2 phases. In the first phase, purposive sampling was applied at the provincial and municipal levels to capture macro-level perspectives on policy-making and program management. Five key administrators were invited from the core management institutions involved in the program. These included 2 officials from the Maternal and Child Health Department of the Provincial/Municipal Health Commission, 2 project managers from the Provincial/Municipal Women's Federation, and 1 project director from the Provincial Maternal and Child Health Hospital.

In the second phase, a multistage stratified sampling method was used to select county-level implementation sites to investigate influencing factors at the grassroots level. The multistage stratified sampling primarily considers factors such as screening rounds, cervical cancer detection rates, and follow-up rates. The detailed sampling flowchart is presented in Figure 2. Based on this framework, 9 county or district-level maternal and child health hospitals were selected as study sites. These hospitals are the sole organizing and implementing institutions for cervical cancer screening within their respective jurisdictions; therefore, their project managers possess

XSL•FO
RenderX

comprehensive knowledge of local screening management and operations. One project manager directly responsible for the program was invited from each hospital to participate in an interview. Thus, this study included 14 key informants as interview participants, whose role distribution ensured multilevel perspective coverage from policy formulation to grassroots implementation.

Figure 2. Flowchart of the multistage stratified sampling process.

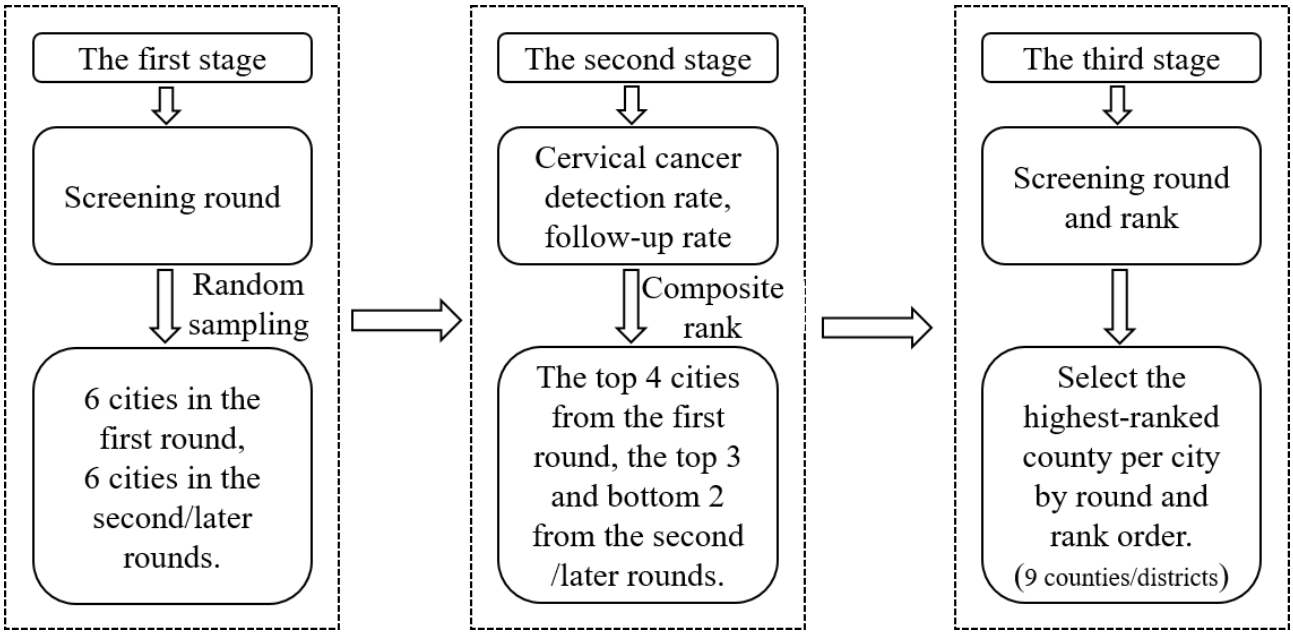

Data Collection

All participants were informed of the researcher's institutional affiliation and the purpose of the qualitative study, which was to understand the implementation of the cervical cancer screening program. Following the provision of verbal informed consent by all participants, the data were collected between January and August 2024. All semistructured interviews were conducted by the same researcher (the first author, holding a PhD in management) to ensure consistency in the interview process and interaction style. The interviewer was female, possessed 7 years of qualitative research experience, and had previously conducted specialized interviews related to cervical cancer screening, ensuring a deep understanding of the research context. Of the 14 participants, 2 had prior professional contact with the interviewer from a 2021 project; to mitigate potential bias toward official stances, the interviewer emphasized academic neutrality and proactively probed for challenges, while the remaining 12 were new contacts. Interviews were carried out either face-to-face or by telephone, in private offices or meeting rooms to protect participant confidentiality. Each interview lasted approximately 60 to 80 minutes and was audio-recorded.

Data Management and Analysis

Within 24 hours post interview, a research team member transcribed the audio recordings verbatim. A second researcher cross-verified transcripts against original recordings to ensure completeness and accuracy. We replaced all personal identifiers with unique codes to anonymize field notes, recordings, and documents. Transcripts were imported into the NVivo 12 (QSR International) software for organization and coding. During the dynamic determination process, we adopted the constant comparative method of "interviewing, coding, and comparing simultaneously." Two professionals independently completed the transcription and coding of interviews; if there were discrepancies in coding, the experienced interview host made the final decision, which effectively avoided subjective biases caused by different researchers. Based on the inductive thematic analysis method under the MMHS framework, we have analyzed and identified the key influencing factors and have also obtained a comprehensive and systematic set of influencing factors.

Thematic saturation was achieved in this study, with no new themes emerging after the interview with the second participant. In the early stage of the study, the applicability of the MMHS framework was verified through a systematic literature review and expert consultations, and a semi-structured interview guide was developed based on this framework. The guide includes sequential questions targeting influencing factors at different levels, supplemented by keyword prompts to minimize potential information gaps, thereby ensuring the systematicity and completeness of the interview content (Multimedia Appendix 1). Coupled with the subsequent adoption of inductive thematic analysis, which enabled the rapid categorization of data into corresponding thematic categories, the study achieved thematic saturation efficiently.

Furthermore, the findings of the thematic analysis were presented and validated for feedback at a cervical cancer screening seminar with 2 provincial-level key informants, who were also participants in this study. Their perspectives were highly consistent with the study's findings, and no revision suggestions were proposed.

Ethical Considerations

Ethics approval was obtained from the Tongji Medical College, Huazhong University of Science and Technology (approval

XSL•FO
RenderX

number [2023] IEC(A235)). Prior to the start of each interview—whether conducted face-to-face or via telephone—verbal informed consent was formally obtained from all participants, who were fully informed of the study purpose, procedures, potential risks, and right to withdraw at any time without any negative consequences before their consent was confirmed. All participant data were anonymized and treated in strict confidence. Each participant was offered a financial compensation of 500 RMB (US $70.60) as a recognition of the time they spent (60 - 80 minutes per interview) and the professional responses they provided. A portion of participants voluntarily chose to participate on an unpaid basis, aiming to promote the smooth implementation of the full-coverage project to serve women's health. The compensation was solely intended to offset participants' time costs and acknowledge their professional inputs, without serving as an incentive to influence the authenticity of their responses.

Results

Participant Characteristics

No participants refused or withdrew during the process. Between January and August 2024, we conducted interviews with 14 administrators from the Health Commission (HC), Women's Federation (WF), and maternal and child health (MCH) hospitals. Participants comprised 9 program managers from MCH hospitals, 2 program directors from provincial HC, and 2 representatives from provincial WF. Among the participants, 92.86% (n=13) reported having more than 10 years of work experience, with 42.86% (n=6) having over 20 years. Most participants held bachelor's degrees, with 3 participants having completed postgraduate education. Detailed participant characteristics are provided in Table 1.

Table . Sociodemographic and occupational characteristics of the participants.

Characteristic	Participants, n (%)
Organization	
The Women's Federation	2 (14.29)
The Maternal and Child Health Hospitals	10 (71.42)
The Health and Wellness Administration	2 (14.29)
Sex	
Male	4 (28.57)
Female	10 (71.42)
Age (y)	
30 - 39	2 (14.29)
40 - 49	9 (64.29)
50 - 59	3 (21.43)
Education	
Associate degree	1 (7.14)
College	10 (71.42)
Master	3 (21.43)
Work experience (y)	
3 - 9	1 (7.14)
10 - 19	7 (50)
20 - 29	6 (42.86)

Thematic Analysis

Overview

Thematic analysis revealed that the core drivers can be summarized into 4 key factors: government stewardship serves as the fundamental guarantee, screening technology provides the foundational support, screening funding supplies sustained momentum, and improving public health literacy is the intrinsic key to achieving broad participation. Further systematic analysis indicates that these factors can be integrated into a comprehensive "structure-process-outcome" framework. Structural factors include government stewardship, multisectoral collaboration, funding, information system, screening technology, and trained screening technicians. Process factors include service delivery capacity, community mobilization, health literacy, and quality control. Outcome measures include coverage, cytology positivity rate, and follow-up rate treatment rate. Detailed results are presented in Table 2.

XSL•FO
RenderX

Table . Themes and Subthemes.

Themes and subthemes	Tertiary theme
Structure	
Administration	• Government stewardship
Organization	• Multisectoral collaboration
Resources	• Funding for screening • Information system • Screening technology • Trained screening technicians
Process	
Service provision and utilization	• Service delivery capacity • Community mobilization • Health literacy
Outcomes	
Screening results	• Coverage • Cytology positivity rate • Follow-up rate • Treatment rate

Government Stewardship

Participants indicated that strong government promotion was a key factor enabling the project to achieve comprehensive coverage. This governmental driving force was primarily manifested by incorporating the screening program into regional government performance metrics, with local government departments overseeing HC or WF and MHC in completing the screening tasks. Participants consistently emphasized:

> *The government's proactive approach demonstrates the decisive role of political will. Sustained political support has been the key driver for program implementation, and policy endorsement of AI technology was indispensable for scaling up this AI-assisted screening initiative.* [MCH01,02,03,04,07,08; HC01]

Another participant noted:

> *The provincial government prioritized free cervical cancer screening, incorporating it into the ' public service delivery' performance metrics. Screening completion rates became key evaluation criteria for local governments. Without such government mobilization, universal screening coverage would be unattainable.* [WF01]

Additionally, the government's leading role is reflected in its support for screening funding and the application of AI-assisted diagnostic technologies.

Screening Technology

Participants indicated that AI-assisted cervical cancer screening technology serves as a critical tool for achieving universal coverage. Previously, primary health care institutions struggled to meet demands of residents due to a shortage of pathologists. By facilitating the decentralization of remote pathological resources, AI-assisted diagnostic technology has enhanced service accessibility and addressed screening bottlenecks in areas with limited health care resources. Furthermore, the screening efficiency of AI-assisted diagnostic technology is remarkably high, enabling significant improvements in screening coverage within a short timeframe.

> *AI-based image analysis reduces physicians' workload and ensures high-quality task completion in primary care settings with insufficient pathology personnel. The technology can automatically handle large volumes of repetitive tasks, such as preliminary screening and marking abnormal areas, thereby alleviating physicians' workload. Under traditional methods, a pathologist could screen a maximum of 100 individuals per day, whereas digitization increases this capacity to nearly 10,000.* [MCH03]

> *The automated system processes cases much faster than manual methods, and the generated reports are highly accurate. Our AI big data cloud diagnosis platform generates up to 50,000 daily diagnoses.* [MCH07]

> *In our region, manual screening previously allowed only 4,000 cases annually, but with AI technology, we can now screen over 73,000 individuals per year.* [MCH10]

> *Some regions face significant resource limitations, particularly in mountainous areas where specialized pathologists are scarce. The introduction of AI-assisted screening technology helps mitigate these disparities. Through cloud-based API calls, primary care institutions can access advanced diagnostic technology, improving the accessibility of cervical cancer screening.* [MCH01]

XSL•FO
RenderX

Screening Funding

Insufficient screening funding, inadequate matching funds, and unclear allocation collectively constitute the primary financial challenges for the screening program. These issues not only hinder smooth program implementation but also strain health care institutions' financial capacity and compromise service quality. Some screening institutions report that large-scale screening has caused more severe funding shortages than before.

> *The per-case subsidy decreased 25.8% (from 66 RMB to 49 RMB) following AI implementation. But we must cover multiple costs including disposable supplies, field visits of personnel and equipment, and follow-up activities, costs that were underestimated or unaccounted for in the budget, making the 49 RMB subsidy insufficient to cover actual costs. We consistently operate at a loss when conducting screenings.* [MCH02,07,10]

> *Healthcare institutions receive less than 40% government funding, relying entirely on self-generated revenue. However, state funding for program operations remains severely inadequate, and trade union funding has been reduced in recent years—we're expected to do more work with significantly less funding.* [MCH04]

Furthermore, insufficient matching funds remain an issue.

> *We face tremendous financial pressure. The government assigned heavy screening targets without allocating sufficient funds, requiring advance payments that overextend local government finances.* [MCH05,08]

Unclear fund allocation exacerbates the funding shortage, remaining a key barrier to achieving universal screening coverage.

> *The core funding issue stems from policy-mandated cost estimates set unrealistically low, resulting in systemic underfunding.* [HC01]

> *Current lump-sum allocations lack itemized funding for dual cancers (cervical or breast), with no specific indicators for individual budget lines…* [MCH07]

> *Lump-sum funding becomes problematic at county-level redistribution, with no clear documentation on allocation plans or expenditure purposes.* [MCH01]

Health Literacy

Participants noted that insufficient health literacy among residents is a significant barrier, particularly in rural areas. Although AI-assisted diagnostic technologies for cervical cancer screening have substantially improved service accessibility, residents' willingness to utilize free screening services remains low. Officials from MCH hospitals emphasized that promoting cervical cancer prevention awareness, disseminating information about free screening, and mobilizing residents to participate require dedicated and proactive efforts, rather than merely providing information about free screening availability.

> *If individuals receive invitations, they may participate; however, without direct contact, they might overlook group notifications.* [MCH08]

> *Many are constrained by cognitive limitations and traditional cultural beliefs, while others decline participation due to factors like travel distance to townships or family responsibilities.* [MCH09]

> *Due to insufficient recognition of its importance, incentives such as transportation assistance and free breast cancer screenings are necessary to achieve screening rates.* [MCH04]

> *Project success fundamentally depends on population health literacy—once literacy improves, residents will proactively seek screening without promotional campaigns.* [MCH07]

> *Some women refuse follow-up after initial negative results, mistakenly viewing normal outcomes as permanent risk exemption—reflecting inadequate screening awareness.* [MCH01]

Multisectoral Collaboration

Participants indicated that a multisectoral collaboration mechanism had been established prior to the implementation of the comprehensive coverage project. During the execution of the project, under government stewardship, the responsibilities of each department were further clarified, detailed, and refined. In the process of implementing the comprehensive coverage plan, the MCH were responsible for providing screening services, and the HC oversaw personnel training, project progress, and quality control, while the WF was tasked with mobilizing residents to actively utilize cervical cancer screening services.

> *This project is spearheaded by the Women's Federation and the Hubei Provincial Health Commission, with joint participation from the Trade Union and the Finance Department. The Women's Federation is responsible for publicity efforts, while the Health Commission provides staff training for maternal and child health hospitals , oversees project progress, and ensures quality control. The Finance Department provides financial support.* [HC02; MCH01]

Information System

Screening programs require the collection of extensive information, and the AI-assisted diagnostic system's information platform streamlines the entire process for residents—from initial screening to receiving results. The implementation of this information platform has enhanced the management efficiency of screening programs. However, several challenges persist, including incomplete system development, data collection gaps, data sharing difficulties, and limited usability at grassroots levels. Participants reported:

> *The system cannot export positive cases, requiring manual entry for each one—extremely inconvenient.* [MCH02]

Another user noted:

XSL•FO
RenderX

> *The system operation is overly complex. A simplified system design would enhance usability.* [MCH06]

Additionally, concerns were raised about data completeness:

> *The system's statistical data isn't entirely complete … trade union and corporate data aren't reported into this system.* [MCH01]

Trained Screening Technicians

Professional personnel required for cervical cancer screening include screening operators, cytopathologists, and diagnostic pathologists, with the shortage of the latter 2 categories being particularly pronounced. Project managers reported that the introduction of AI-assisted diagnostic technology has addressed the shortage of specialized technical personnel in cytological and pathological diagnosis.

> *Current staffing remains suboptimal with inadequate professional qualifications. Nevertheless, AI technology can address challenges in cytology and pathology diagnosis, effectively resolving the difficulties caused by a shortage of specialized personnel.* [MCH01,07; WF02]

> *We conduct annual colposcopy training requiring physicians to pass certification exams. Additionally, the provincial MCH hospital offers specialized three-month training programs to ensure sampling proficiency.* [MCH05,06; HC02]

Service Delivery Capacity

In the comprehensive cervical cancer screening program, MCH's service capacity is primarily reflected in its capabilities for cervical cancer screening and treatment. The screening program encompasses both urban and rural regions, involving provincial or county MCH and township health centers. District or county-level centers exhibit good screening and treatment capabilities. However, township centers face challenges stemming from professional and resource limitations, resulting in restricted screening and treatment capacities. Notably, AI technology has significantly enhanced their screening capabilities.

> *Regarding service capacity, most institutions can adequately manage general gynecological diseases and precancerous lesions.* [MCH02,04,07]

> *Township facilities are experiencing downsizing with insufficient obstetrician and gynecologist, limiting maternal and child health follow-up capacity. However, AI-enabled sample referral to provincial centers now ensures successful rural screening implementation.* [MCH03; HC01]

Quality Control

Quality control is essential for ensuring both progress and quality in achieving universal cervical cancer screening coverage. It involves technical accuracy and reliability at the operational level. In addition to on-site inspections, remote monitoring is conducted through the information system, primarily to track completion rates across all screening sites.

> *We monitor progress through the information system and reported data, including the number of individuals screened, treated, and managed. These metrics ultimately generate performance rankings.* [MCH01,03,05]

Community Mobilization

Community mobilization campaigns, particularly those driven by administrative initiatives and free voucher distribution, have proven effective in enhancing awareness and increasing cervical cancer screening participation rates. Large-scale community mobilization campaigns can improve public understanding of cervical cancer, thereby increasing screening willingness and advancing the goal of universal screening coverage. These strategies have significantly contributed to achieving comprehensive screening coverage.

> *The free cervical cancer screening program will be promoted through flyers, short videos, official social media accounts, television programs, small-scale lectures, and live-streaming sessions.* [MCH04,05,09]

> *Free screening vouchers were printed and distributed to village clinics and women's committee directors, serving both promotional and participatory purposes. Last year, 100,000 vouchers were printed and distributed based on demand.* [MCH02]

Coverage and Cytology Positivity Rate

Nearly all regions reported achieving comprehensive coverage of cervical cancer screening, primarily attributed to the enhanced service accessibility and efficiency facilitated by AI-assisted technology. Compared to previous methods, the technological upgrades in screening have also significantly increased the cytological positive detection rate. This has further heightened health awareness and willingness to proactively utilize screening services. Coverage and positive detection rates not only directly affect early diagnosis and treatment rates in the current screening cycle but also dynamically shape whether the goal of sustained universal coverage can be achieved.

> *The adoption of AI assistance has significantly improved overall detection rates compared to previous methods.* [MCH01]

With advancements in screening technology and updated equipment, the positive detection rate for cervical cancer screening has increased markedly compared to historical levels. Multiple respondents confirmed:

> *Universal coverage has essentially been achieved.* [MCH02-10]

Follow-Up and Treatment Rates

Cervical cancer screening should be conducted at regular intervals to detect lesions at an early stage and enable timely intervention. Participants emphasized that achieving comprehensive screening coverage requires attention not only to the screening rate but, more importantly, to follow-up management and treatment rates. Follow-up and treatment rates are key indicators for the closed-loop management of screening programs, directly influencing their effectiveness and affecting

XSL•FO
RenderX

whether residents will continue to participate in screening regularly in the future. Although universal screening coverage has been achieved, significant challenges remain in the follow-up of positive cases, which may, to some extent, affect future screening rates.

> *Our follow-up reaches approximately 70% of targets, as many patients don't respond to phone calls … with higher loss-to-follow-up rates in rural areas.* [MCH05,10]

Primary challenges include patient noncompliance and disparities between screening, re-screening, and treatment locations. These difficulties complicate treatment rate documentation.

> *Some migrant workers relocate before results are available, while many screen-positive patients seek treatment at top-tier hospitals elsewhere—making data tracking challenging though we maintain documentation efforts.* [WF01,02]

Discussion

Principal Findings

Through interviews, we have identified factors influencing the achievement of universal coverage in cervical cancer screening, which span 3 dimensions: structural, service-process, and outcome-related. Among these, government stewardship is considered the primary determinant, followed by the application of AI-assisted diagnostic technology, sufficient funding, and public health literacy, all of which jointly contribute to the realization of comprehensive screening coverage. During implementation, government-led initiatives provide strong support for the promotion of AI-assisted diagnostic technology, multisectoral collaboration, and funding allocation, while also facilitating public access to screening services. Concurrently, AI-assisted diagnostic technology helps alleviate shortages of trained screening technicians, establishes information management platforms, reduces screening costs, and thereby enhances institutional service capacity. To further advance the goal of universal coverage, it is essential to ensure sustained financial investment, improve public health literacy, and strengthen follow-up and treatment for positive cases.

Government stewardship serves as the critical enabler for implementing large-scale cervical cancer screening programs. In developed nations such as the United States, New Zealand, and the United Kingdom, cervical cancer prevention systems are legally mandated, whereas China adopts national policy-driven approaches [36]. Hubei's universal screening program, integrated into the provincial "Public Service Initiatives," ties screening targets to local government performance metrics, ensuring comprehensive policy support and fiscal allocation [37]. The program established a multisector framework led by the government, coordinated by the Women's Federation, with participation from health care, financial, and labor unions. It is noteworthy that even when funding management and public health literacy failed to exert a significant positive influence, the government effectively achieved the phased objective of "testing all who should be tested" through the impetus of clearly defined assessment targets. In contrast, most LMICs lack both legal frameworks and substantive policy support for cervical cancer prevention [38]. Screening in LMICs remains opportunistic, lacking systematic implementation strategies. Thus, robust governmental commitment and policy infrastructure are prerequisites for population-wide screening.

AI-assisted screening technology represents a pivotal component in achieving large-scale cervical cancer screening. It substantially enhances screening efficiency and coverage while reducing costs, enabling Hubei province to complete population-wide screening within compressed timelines. In 2016, a single Chinese pathologist could manually process approximately 100 cytology screenings daily—a rate that would leave millions of women lifetime-unserved by screening programs [39]. However, AI technology allows a single pathologist to conduct roughly 10,000 screenings per day, significantly cutting down mass screening time. The technology also improves accessibility, allowing rural women to undergo sample collection at local clinics without travel burdens. Economically, AI-assisted screening demonstrates superior cost-effectiveness, reducing per-capita costs from 100-200 RMB (medical service fees) to 49 RMB, enabling broader eligibility for free screening. These advantages make universal screening coverage a feasible public health goal.

Financially, the government has allocated approximately 600 million RMB to advance the universal cervical cancer screening initiative. However, it is noteworthy that screening institutions continue to face challenges in funding security, manifested through 3 dimensions: insufficient total funding, lack of matching funds, and ambiguous allocation mechanisms. Funding shortages are particularly acute in LMICs, where systemic underinvestment persists [40,41]. Despite AI's cost-saving potential, funding gaps endure due to insufficient government reimbursement and overreliance on singular revenue streams [42,43]. Sustainable implementation requires diversified financing mechanisms and evidence-based fiscal policies [44]. Current tiered matching-fund systems fail to alleviate local funding deficits, as reported by stakeholders. This mismatch disproportionately affects economically disadvantaged regions, where heavy reliance on central transfers, limited fiscal autonomy, and low economic capacity exacerbate funding inadequacies [45]. Solutions include needs-based allocation models incorporating regional disparities in fiscal capacity, population size, and geographic coverage. Long-term economic development remains critical for fiscal resilience [46]. China's lump-sum public health funding—which includes screening budgets—lacks transparent suballocation guidelines [47]. Ambiguities persist regarding cost coverage and interagency fund distribution. Unclear allocation risks resource concentration, compromising efficiency and equity [48]. Standardized operational guidelines for fund distribution are thus imperative for program sustainability [49].

Population health literacy has emerged as a critical determinant of overall health status, reflecting growing public health awareness. It directly influences both health service utilization and treatment-seeking behaviors [50]. This study identified multidimensional barriers to screening participation: low

XSL•FO
RenderX

awareness, embarrassment, geographic accessibility issues, and competing household priorities—necessitating incentives, such as gifts or transportation. Knowledge gaps about cervical cancer persist universally across income settings [51]. Despite understanding triennial screening recommendations, women often discontinue after negative results, reflecting insufficient risk perception. Targeted education interventions demonstrably improve screening knowledge and participation rates [52]. Sustainable screening expansion thus requires dual strategies: technological or service optimization and literacy-focused health promotion. Multimodal education can foster proactive health behaviors and sustained screening engagement [53,54].

In the initial phase of our research, we viewed AI technology as the primary driver for achieving universal cervical cancer screening coverage. We believed that addressing the supply bottleneck caused by a shortage of pathologists through technological means would enable rapid attainment of coverage targets. However, through systematic interview analysis, we discovered that while AI technology can effectively enhance service delivery capacity, its effectiveness is highly dependent on systemic support. Government leadership emerged as the most critical factor in achieving screening coverage. By integrating screening into government performance evaluation systems, introducing new technologies (AI-assisted diagnosis), and providing sustained financial support, the government effectively coordinated resources. It mobilized multiple departments—including WFs and MCH hospitals—to collaborate on outreach and organizational efforts, significantly boosting residents' willingness to participate and actual screening rates. This reflection prompted a fundamental shift in our cognitive framework: moving from an initial "technology-centric" singular assumption to recognizing a multilayered, systematic advancement logic characterized by "government leadership, technological support, multistakeholder collaboration, and competency enhancement." This shift lends our research conclusions greater explanatory power and policy relevance.

Strengths and Limitations

Improving cervical cancer screening coverage is a critical public health priority in China and other LMICs. The emergence of AI technology offers a promising solution to enhance the accessibility of these services. Within the context of AI applications in China's basic public health services, this study systematically investigates the challenges faced in its practical implementation for increasing screening coverage. Furthermore, by applying the MMHS, we provide a comprehensive analysis of the determinants of screening, which offers a robust framework for developing the interview guide. However, a limitation of this study is that it only considers the influencing factors from the perspectives of project managers and providers. In actual program implementation, the perspective of the demand side is equally crucial, and its absence may result in factors affecting program execution being identified incompletely.

Conclusions

The achievement of comprehensive cervical cancer screening coverage is mainly influenced by government stewardship, AI-assisted screening, screening funding, and health literacy. Among these, government stewardship serves as the core driving force, specifically reflected in the inclusion of screening in government performance evaluations, which drives policy implementation, technological adoption, funding security, multisectoral collaboration, and activities aimed at improving residents' health literacy and screening willingness. The application of AI-assisted diagnostic technology effectively mitigates constraints, such as shortages of trained screening technicians, limited funding, and insufficient service delivery capacity. To ensure the smooth implementation of the project and achieve preventive goals, coordinated efforts are required in 3 key areas: at the policy level, clear screening plans and supportive measures should be formulated; at the financial level, sufficient funding must be guaranteed, with coordinated multilevel and multidepartmental resources and clearly defined allocation mechanisms; at the service delivery level, diversified health education should be carried out to enhance residents' health literacy, cultivate regular screening habits, and promote the application of AI technology.

Acknowledgments

We extend our gratitude to all the stakeholders in Hubei province who contributed their insights and perspectives on the research topic.

Funding

This work was supported by Health Commission of Hubei Province, China (WJ2023Z008).

Data Availability

All data generated or analyzed during this study are included in this published paper.

Authors' Contributions

LY is the principal investigator of the study and responsible for the overall content as guarantor. LJ designed the study, drafted the study documents, and participated in both data collection and analysis. XZ assisted in drafting the study documents and coding and analyzing the transcripts. LY and LJ provided feedback on the study protocol draft and related documents and reviewed the quality of the data, and XZ assisted data collection. LJ and XZ supported the development of data extraction processes and the

XSL•FO
RenderX

interpretation of the findings. XZ authored the first draft of the manuscript with support from LJ and LY. All other authors contributed feedback on the manuscript and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[DOCX File, 22 KB - jmir_v28i1e75372_app1.docx]

Checklist 1
COREQ checklist.
[DOC File, 73 KB - jmir_v28i1e75372_app2.doc]

References

1. Ferlay J, Ervik M, Lam F, et al. Cancer today. Global Cancer Observatory (GCO). URL: https://gco.iarc.who.int/today [accessed 2024-09-24]
2. Cancer today–China. : Global Cancer Observatory (GCO), International Agency for Research on Cancer (IARC), World Health Organization (WHO); 2024 URL: https://gco.iarc.who.int/media/globocan/factsheets/populations/160-china-fact-sheet.pdf [accessed 2024-09-24]
3. Zheng R, Zhang S, Zeng H, et al. Cancer incidence and mortality in China, 2016. J Natl Cancer Cent 2022 Mar;2(1):1-9. [doi: 10.1016/j.jncc.2022.02.002] [Medline: 39035212]
4. Global strategy to accelerate the elimination of cervical cancer as a public health problem. : World Health Organization (WHO); 2020 URL: https://www.who.int/publications/i/item/9789240014107 [accessed 2024-10-28]
5. Zhang X, Hu S, Zhao F, et al. Cervical cancer screening service policies in China, 2006–2023: a quantitative analysis of government documents. Chinese J Public Health 2024;40(5):521-527. [doi: 10.11847/zgggws1143537]
6. Campos NG, Tsu V, Jeronimo J, Mvundura M, Lee K, Kim JJ. To expand coverage, or increase frequency: quantifying the tradeoffs between equity and efficiency facing cervical cancer screening programs in low-resource settings. Int J Cancer 2017 Mar 15;140(6):1293-1305. [doi: 10.1002/ijc.30551] [Medline: 27925175]
7. Lin S, Gao K, Gu S, et al. Worldwide trends in cervical cancer incidence and mortality, with predictions for the next 15 years. Cancer 2021 Nov 1;127(21):4030-4039. [doi: 10.1002/cncr.33795] [Medline: 34368955]
8. Yagi A, Ueda Y, Kakuda M, et al. Epidemiologic and clinical analysis of cervical cancer using data from the population-based Osaka Cancer Registry. Cancer Res 2019 Mar 15;79(6):1252-1259. [doi: 10.1158/0008-5472.CAN-18-3109] [Medline: 30635276]
9. Ojamaa K, Innos K, Baburin A, Everaus H, Veerus P. Trends in cervical cancer incidence and survival in Estonia from 1995 to 2014. BMC Cancer 2018 Nov 7;18(1):1075. [doi: 10.1186/s12885-018-5006-1] [Medline: 30404606]
10. Bruni L, Serrano B, Roura E, et al. Cervical cancer screening programmes and age-specific coverage estimates for 202 countries and territories worldwide: a review and synthetic analysis. Lancet Glob Health 2022 Aug;10(8):e1115-e1127. [doi: 10.1016/S2214-109X(22)00241-8] [Medline: 35839811]
11. Zhang M, Zhong Y, Wang L, et al. Cervical cancer screening coverage—China, 2018-2019. China CDC Wkly 2022 Dec 2;4(48):1077-1082. [doi: 10.46234/ccdcw2022.217] [Medline: 36751373]
12. Lahole BK, Woldamlak M, Kussia W. Determinants of cervical cancer screening intention among reproductive age women in Ethiopia: a systematic review and meta-analysis. PLoS One 2024;19(10):e0312449. [doi: 10.1371/journal.pone.0312449] [Medline: 39480772]
13. Wang M, Chen H, Wong MCS, Huang J, Jin Y, Zheng ZJ. Effects of screening coverage and screening quality assurance on cervical cancer mortality: implication for integrated framework to monitor global implementation of cervical cancer screening programmes. J Glob Health 2024 Oct 25;14:04189. [doi: 10.7189/jogh.14.04189] [Medline: 39450607]
14. Salehiniya H, Momenimovahed Z, Allahqoli L, Momenimovahed S, Alkatout I. Factors related to cervical cancer screening among Asian women. Eur Rev Med Pharmacol Sci 2021 Oct;25(19):6109-6122. [doi: 10.26355/eurrev_202110_26889] [Medline: 34661271]
15. Sabeena S, Ravishankar N, Kalpana MM. Implementation strategies of cervical cancer screening in South Asia: a systematic review. Int J Gynaecol Obstet 2024 Aug;166(2):483-493. [doi: 10.1002/ijgo.15366] [Medline: 38268412]
16. Surendran S, Foo CD, Tan DHY, et al. Understanding barriers and facilitators of breast and cervical cancer screening among Singapore women: a qualitative approach. Asian Pac J Cancer Prev 2023 Mar 1;24(3):889-895. [doi: 10.31557/APJCP.2023.24.3.889] [Medline: 36974542]
17. Suk R, Hong YR, Rajan SS, Xie Z, Zhu Y, Spencer JC. Assessment of us preventive services task force guideline-concordant cervical cancer screening rates and reasons for underscreening by age, race and ethnicity, sexual orientation, rurality, and

XSL•FO
RenderX

insurance, 2005 to 2019. JAMA Netw Open 2022 Jan 4;5(1):e2143582. [doi: 10.1001/jamanetworkopen.2021.43582] [Medline: 35040970]
18. Shrestha AD, Andersen JG, Gyawali B, et al. Cervical cancer screening utilization, and associated factors, in Nepal: a systematic review and meta-analysis. Public Health 2022 Sep;210:16-25. [doi: 10.1016/j.puhe.2022.06.007] [Medline: 35863158]
19. Mezei AK, Armstrong HL, Pedersen HN, et al. Cost-effectiveness of cervical cancer screening methods in low- and middle-income countries: a systematic review. Int J Cancer 2017 Aug 1;141(3):437-446. [doi: 10.1002/ijc.30695] [Medline: 28297074]
20. Allanson ER, Schmeler KM. Cervical cancer prevention in low- and middle-income countries. Clin Obstet Gynecol 2021 Sep 1;64(3):501-518. [doi: 10.1097/GRF.0000000000000629] [Medline: 34120126]
21. Martinez ME, Schmeler KM, Lajous M, Newman LA. Cancer screening in low- and middle-income countries. Am Soc Clin Oncol Educ Book 2024 Jun;44(3):e431272. [doi: 10.1200/EDBK_431272] [Medline: 38843475]
22. Di J, Rutherford S, Chu C. Review of the cervical cancer burden and population-based cervical cancer screening in China. Asian Pac J Cancer Prev 2015;16(17):7401-7407. [doi: 10.7314/apjcp.2015.16.17.7401] [Medline: 26625735]
23. Ji L, Chen M, Yao L. Strategies to eliminate cervical cancer in China. Front Oncol 2023;13:1105468. [doi: 10.3389/fonc.2023.1105468] [Medline: 37333817]
24. Ji L, Yao Y, Yu D, et al. Performance of a full-coverage cervical cancer screening program using on an artificial intelligence- and cloud-based diagnostic system: observational study of an ultralarge population. J Med Internet Res 2024 Nov 20;26:e51477. [doi: 10.2196/51477] [Medline: 39566061]
25. Bao H, Sun X, Zhang Y, et al. The artificial intelligence-assisted cytology diagnostic system in large-scale cervical cancer screening: a population-based cohort study of 0.7 million women. Cancer Med 2020 Sep;9(18):6896-6906. [doi: 10.1002/cam4.3296] [Medline: 32697872]
26. Bao H, Bi H, Zhang X, et al. Artificial intelligence-assisted cytology for detection of cervical intraepithelial neoplasia or invasive cancer: a multicenter, clinical-based, observational study. Gynecol Oncol 2020 Oct;159(1):171-178. [doi: 10.1016/j.ygyno.2020.07.099] [Medline: 32814641]
27. Chen H, Liu J, Wen QM, et al. CytoBrain: cervical cancer screening system based on deep learning technology. J Comput Sci Technol 2021 Apr;36(2):347-360. [doi: 10.1007/s11390-021-0849-3]
28. Cervical cancer screening can be done at township health centers [Article in Simplified Chinese]. Hubei Provincial Health Commission. 2025. URL: https://wjw.hubei.gov.cn/bmdt/dtyw/202410/t20241021_5377559.shtml [accessed 2025-01-09]
29. Hou X, Shen G, Zhou L, Li Y, Wang T, Ma X. Artificial intelligence in cervical cancer screening and diagnosis. Front Oncol 2022;12:851367. [doi: 10.3389/fonc.2022.851367] [Medline: 35359358]
30. Bi WL, Hosny A, Schabath MB, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. CA Cancer J Clin 2019 Mar;69(2):127-157. [doi: 10.3322/caac.21552] [Medline: 30720861]
31. Adams SJ, Stone E, Baldwin DR, Vliegenthart R, Lee P, Fintelmann FJ. Lung cancer screening. Lancet 2023 Feb 4;401(10374):390-408. [doi: 10.1016/S0140-6736(22)01694-4] [Medline: 36563698]
32. Wu T, Lucas E, Zhao F, Basu P, Qiao Y. Artificial intelligence strengthens cervical cancer screening—present and future. Cancer Biol Med 2024 Sep 19;21(10):864-879. [doi: 10.20892/j.issn.2095-3941.2024.0198] [Medline: 39297572]
33. Gastounioti A, Desai S, Ahluwalia VS, Conant EF, Kontos D. Artificial intelligence in mammographic phenotyping of breast cancer risk: a narrative review. Breast Cancer Res 2022 Feb 20;24(1):14. [doi: 10.1186/s13058-022-01509-z] [Medline: 35184757]
34. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
35. Hao M. Health Policy [Book in Chinese], 2nd edition: People's Medical Publishing House; 2013. URL: https://product.dangdang.com/11956642953.html [accessed 2026-03-23]
36. Bao HL, Fang LW, Wang LH. The strategy for establishment of comprehensive cervical cancer prevention and control in the world [Article in Chinese]. Zhonghua Yu Fang Yi Xue Za Zhi 2017 Jan 6;51(1):96-100. [doi: 10.3760/cma.j.issn.0253-9624.2017.01.019] [Medline: 28056279]
37. Dee EC, Eala MAB, Robredo JPG, et al. Leveraging national and global political determinants of health to promote equity in cancer care. J Natl Cancer Inst 2023 Oct 9;115(10):1157-1163. [doi: 10.1093/jnci/djad123] [Medline: 37402623]
38. Lee J, Ismail-Pratt I, Machalek DA, Kumarasamy S, Garland SM. From barriers to opportunities from COVID-19 pandemic: stakeholder perspectives on cervical cancer screening programs in LMICs of the Asia-Pacific region. PLOS Glob Public Health 2024;4(10):e0003768. [doi: 10.1371/journal.pgph.0003768] [Medline: 39365795]
39. Li Y, Ma L, Yang C, et al. A study on service capacity of primary medical and health institutions for cervical cancer screening in urban and rural areas in China. Chin J Cancer Res 2019 Oct;31(5):838-848. [doi: 10.21147/j.issn.1000-9604.2019.05.13] [Medline: 31814687]
40. Petersen Z, Jaca A, Ginindza TG, et al. Barriers to uptake of cervical cancer screening services in low-and-middle-income countries: a systematic review. BMC Womens Health 2022 Dec 2;22(1):486. [doi: 10.1186/s12905-022-02043-y] [Medline: 36461001]

XSL•FO
RenderX

41. Devarapalli P, Labani S, Nagarjuna N, Panchal P, Asthana S. Barriers affecting uptake of cervical cancer screening in low and middle income countries: a systematic review. Indian J Cancer 2018;55(4):318-326. [doi: 10.4103/ijc.IJC_253_18] [Medline: 30829264]
42. Luo YB, Jia LG, Liu JA, et al. Management on national basic public health service projects and its related issues in Hubei province [Article in Chinese]. Chin J Public Health 2019;35(6):676-679. [doi: 10.11847/zgggws1120669]
43. Zhu X, Yao Q, Dai W, et al. Cervical cancer screening aided by artificial intelligence, China. Bull World Health Organ 2023 Jun 1;101(6):381-390. [doi: 10.2471/BLT.22.289061] [Medline: 37265676]
44. Kutzin J. Health financing for universal coverage and health system performance: concepts and implications for policy. Bull World Health Organ 2013 Aug 1;91(8):602-611. [doi: 10.2471/BLT.12.113985] [Medline: 23940408]
45. Feng G. Local finance from the perspective of government intervention: a literature review [Article in Chinese]. Financial 2016(3):67-79125. [doi: 10.3969/j.issn.1674-7690.2016.03.005]
46. Zhang Y. Reflections on the optimization of China's central finance to local transfer payment system: current situation, question and countermeasure [Article in Chinese]. Finance 2024;14(5):1640-1647. [doi: 10.12677/fin.2024.145168]
47. Chen H, Tao HB, Luo LX, et al. The situation and analysis of the allocation and performance management system of community health service funds at home and abroad. Chin Gen Pract 2011(16):7. [doi: 10.3969/j.issn.1007-9572.2011.16.005]
48. Zhi Q, Meng T. Funding allocation, inequality, and scientific research output: an empirical study based on the life science sector of natural science foundation of China. Scientometrics 2016 Feb;106(2):603-628. [doi: 10.1007/s11192-015-1773-5]
49. Wang LM, Yu J. Analysis of the allocation mechanism of basic public health services in Wuhan. Med Soc (Berkeley) 2013(7):56-59. [doi: 10.3870/YXYSH.2013.07.019]
50. Kim K, Han HR. Potential links between health literacy and cervical cancer screening behaviors: a systematic review. Psychooncology 2016 Feb;25(2):122-130. [doi: 10.1002/pon.3883] [Medline: 26086119]
51. Kung TPH, Gordon JR, Abdullahi A, et al. "My husband says this: if you are alive, you can be someone…": facilitators and barriers to cervical cancer screening among women living with HIV in India. Cancer Causes Control 2019 Apr;30(4):365-374. [doi: 10.1007/s10552-019-01145-7] [Medline: 30809741]
52. Mukama T, Ndejjo R, Musabyimana A, Halage AA, Musoke D. Women's knowledge and attitudes towards cervical cancer prevention: a cross sectional study in Eastern Uganda. BMC Womens Health 2017 Jan 31;17(1):9. [doi: 10.1186/s12905-017-0365-3] [Medline: 28137260]
53. Popalis ML, Ramirez SI, Leach KM, Granzow ME, Stoltzfus KC, Moss JL. Improving cervical cancer screening rates: a scoping review of resources and interventions. Cancer Causes Control 2022 Nov;33(11):1325-1333. [doi: 10.1007/s10552-022-01618-2] [Medline: 35980511]
54. Holme F, Maldonado F, Martinez-Granera OB, et al. HPV-based cervical cancer screening in Nicaragua: from testing to treatment. BMC Public Health 2020 Apr 15;20(1):495. [doi: 10.1186/s12889-020-08601-z] [Medline: 32295562]

Abbreviations

AI: artificial intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research
HC: Health Commission
LMIC: low- and middle-income country
MCH: maternal and child health
MMHS: macro model of health system
WF: Women's Federation

Edited by A Stone; submitted 02.Apr.2025; peer-reviewed by C Long, D Tang, G Raut, M Temmerman, S Tang; accepted 23.Feb.2026; published 01.Apr.2026.

Please cite as:
Ji L, Zhou X, Yao L
Factors Influencing Universal Coverage of AI-Assisted Cervical Cancer Screening: Qualitative Study Based on the Macro Model of Health System
J Med Internet Res 2026;28:e75372
URL: https://www.jmir.org/2026/1/e75372
doi:10.2196/75372

© Lu Ji, Xinke Zhou, Lan Yao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 1.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Opposition to Youth e-Cigarette Prevention Campaigns on Twitter and TikTok: Cross-Platform Observational Mixed Methods Analysis

Chandler C Carter, MA; Simon Page, BA; Mateusz Borowiecki, MSc; Ganna Kostygina, PhD; Sherry L Emery, MBA, PhD; Miao Feng, PhD

Social Data Collaboratory, Public Health Department, NORC at the University of Chicago, 55 E Monroe St 30th Floor, Chicago, IL, United States

Corresponding Author:
Miao Feng, PhD
Social Data Collaboratory, Public Health Department, NORC at the University of Chicago, 55 E Monroe St 30th Floor, Chicago, IL, United States

Abstract

Background: Youth e-cigarette use rose sharply between 2013 and 2024 in the United States, prompting widespread prevention campaigns at national, state, and local levels. However, many campaigns encountered online opposition, sometimes leading to message distortion or campaign withdrawal. While previous studies have examined individual campaigns, little is known about how oppositional dynamics differ across social media platforms with distinct architectures.

Objective: This study aimed to conduct a retrospective, cross-platform surveillance study of oppositional responses to US youth e-cigarette prevention campaigns, comparing tactics, themes, and engagement on Twitter (now X) and TikTok to inform platform-specific public health strategies.

Methods: We collected Twitter (2014 - 2020) and TikTok (2020 - 2023) posts related to major US e-cigarette prevention campaigns using 15 campaign-specific hashtags and 4 verified prevention campaign handles. We included public, English-language posts from geographic regions allowed by the platforms. Machine learning classification and human coding were used to detect oppositional content, characterize narrative frames, and classify user types. Engagement was assessed using post-level metrics, including likes, shares, comments, and retweets. We analyzed message prevalence, engagement patterns, and oppositional themes.

Results: On Twitter, opposition comprised 26.8% (83,074/310,207) of campaign-related posts overall but dominated certain campaigns (eg, *Still Blowing Smoke*: 6052/6113, 99%). A small cluster of advocacy and commercial accounts generated 57.8% of opposition retweets. Dominant narratives included questioning the credibility of health authorities, claims that prevention advertisements backfired, vaping rights, and product promotion. In contrast, TikTok opposition constituted only 3.5% (108/3127, 95% CI 3.1%-3.9%) of posts and was characterized by humor (71/108, 65.7%), mockery (48/108, 44.4%), and ironic portrayals of vaping (30/108, 27.8%). Individual creators comprised 76.1% (153/201) of accounts sharing prevention posts, and opposition videos used the visibility-boosting hashtag #fyp significantly more than prevention posts (51.9% vs 32.2%; $P<.001$). Despite inconsistent hashtag use, prevention posts achieved higher average engagement than oppositional content.

Conclusions: This novel cross-platform, multicampaign analysis of opposition responses to e-cigarette prevention campaigns revealed how opposition reflects distinct platform architectures. Twitter opposition was highly coordinated and amplified by commercial and advocacy accounts, especially during regional campaigns. TikTok opposition was decentralized and humor-driven, aligning with the platform's entertainment-oriented algorithm. The findings strengthen health communication by introducing a framework for evaluating platform-specific vulnerabilities and informing evidence-based campaign design. On Twitter, effective countermeasures may require real-time monitoring of social media discourse to support rapid responses to coordinated opposition. On TikTok, leveraging creator partnerships and remix-friendly content may help public health messages compete with entertainment-dominated discourse. Consistent hashtag use can strengthen engagement by minimizing the fragmentation of content visibility, and credible health sources should increasingly reinforce prevention narratives on both platforms. Greater platform accountability and transparency are needed to ensure that prevention content is not systematically deprioritized by algorithms relative to commercial promotion.

(J Med Internet Res 2026;28:e83791) doi:10.2196/83791

KEYWORDS

health campaigns; social media; tobacco industry; public health; health communication; e-cigarettes

XSL•FO
RenderX

Introduction

Following a long decline in youth tobacco use, the emergence of e-cigarettes in the 2010s prompted a resurgence in tobacco use among adolescents in the United States [1]. By 2023, approximately 10% of US high school students reported current e-cigarette use, making these devices the most popular tobacco product among youth [1]. Adolescents are particularly susceptible to nicotine addiction and may face additional respiratory, cardiovascular, and immunological harms [1,2]. This rise in youth use sparked significant public health concern, including concern about the potential for e-cigarettes to act as a "gateway" to other tobacco products. This was a key driver for the creation of federal and local prevention campaigns analyzed in this study [3-5]. Although interventions such as flavor bans have sought to curb youth appeal [6,7], e-cigarettes remain pervasive [8], underscoring the importance of effective prevention campaigns.

A diverse network of federal, state, and nonprofit stakeholders delivers tobacco prevention messages via television, print, and digital platforms [9-12]. These multichannel campaigns have demonstrated effectiveness in shaping risk perceptions and reducing initiation among youth [9,13-18]. However, the rise of social media presents new challenges for maintaining message integrity and evaluating broader impacts [19,20]. Despite generally being banned across traditional media, both direct and indirect tobacco marketing thrives on social media [21,22], with emerging evidence that such promotion is associated with tobacco use and susceptibility [23,24]. Furthermore, interactions among user behaviors, algorithmic feeds, and peer-to-peer interactions, such as remixing or hashtag hijacking, can amplify or undermine prevention efforts in ways not observed in traditional media environments [25-28]. These dynamics require a framework that accounts not only for content but also for how visibility, engagement, and narrative control are structured.

Simultaneously, counter-narratives thrive on social media [22,29,30]. Tactics such as "Twitter bombing" (ie, mass posting to overwhelm hashtags) or astroturfing (coordinated messaging disguised as organic support) allow commercial or advocacy interests to flood platforms with counter-narratives, creating misleading perceptions of widespread support or dissent [31-35]. Theory-based frameworks suggest that social media may allow commercial or advocacy actors to "set the agenda" in ways once restricted to institutional actors and traditional media. These frameworks also suggest that the structure of social media enables novel forms of gatekeeping that can amplify pro-vaping narratives, potentially increasing youth uptake while subverting prevention efforts [36-39]. Platform affordances shape these strategies: text-based environments such as Twitter (now X) facilitate rapid coordination and volume, while TikTok's short-video format enables the spread of provaping content through easily digestible entertainment [28,40-42].

Drawing on theories of networked gatekeeping and principal-agent dynamics [37,38,43], we examine how digital platform architectures mediate the amplification of health messages versus counter-narratives. Networked gatekeeping theory highlights how control over content visibility is distributed across digital actors rather than centralized within traditional editorial institutions [37,43]. Principal-agent dynamics describe how misaligned incentives between message originators, such as public health agencies, and intermediaries, such as platforms or content creators, can shape how messages are disseminated and transformed [38]. In doing so, we explore how platform affordances shape engagement between public health campaigns and their challengers. Although previous research has examined antitobacco messaging or oppositional discourse on individual platforms, particularly Twitter and YouTube, little is known about how opposition dynamics differ across platforms with distinct sociotechnical architectures. Existing studies typically focus on single-platform case studies or specific campaigns [32], often without directly comparing how algorithmic design, user behaviors, and content norms shape oppositional engagement. This gap limits our understanding of how campaign messaging is received, reframed, or resisted in varied digital environments [31,32]. Our study is the first to provide cross-platform surveillance of opposition to US youth e-cigarette prevention campaigns. Specifically, we compare Twitter's historical network dynamics (ie, retweet-based amplification by advocacy clusters) with TikTok's contemporary algorithmic affordances (ie, individualized For You Page and entertainment-driven content formats). In our study, "opposition" is operationalized as content within specific campaign discursive spaces that diverges from or actively challenges official public health messaging. This approach examines patterns that may reflect platform-specific structural impacts of these divergent narratives, acknowledging the inherent complexity of online discourse without implying a simplistic binary. By analyzing both oppositional structures and engagement characteristics, we aim to illuminate platform-specific vulnerabilities and outline adaptive strategies for strengthening public health communication in increasingly participatory, algorithmically governed media ecosystems. We pursue three core aims: (1) to assess the volume and engagement of prevention campaign–related social media posts, (2) to identify the sources and tactics of oppositional content, and (3) to characterize platform-specific opposition dynamics to inform future strategy development.

Methods

Study Design and Rationale for Mixed Methods Approach

We conducted a US-based cross-platform surveillance study examining opposition to e-cigarette prevention campaigns on Twitter (2014 - 2020) and TikTok (2020 - 2023). We use the name "Twitter" throughout this study as our data collection period for this platform concluded in 2020, before its 2022 acquisition and rebranding as X. We analyzed posts related to campaigns at multiple levels of implementation, including national (eg, the Food and Drug Administration's *The Real Cost*), state (eg, California's *Still Blowing Smoke*), and local (eg, Chicago's *Vaping Truth*). *The Real Cost* campaign initially launched in 2014 to curb cigarette use among youth, but by 2018, the campaign pivoted to address e-cigarette use [44].

XSL•FO
RenderX

This study used a convergent mixed methods design to integrate machine learning classification (quantitative) with thematic analysis (qualitative) to triangulate opposition patterns across platforms. Quantitative and qualitative findings were integrated at the interpretation stage by jointly examining how patterns in volume, engagement, and user type corresponded to emergent opposition themes across platforms. We selected this approach because quantitative classifiers alone cannot adequately capture multimodal content, while qualitative coding alone cannot scale across millions of posts.

Due to nonconcurrent time frames, we focused on identifying structural differences in oppositional engagement rather than conducting comparisons of post volume for cross-platform campaigns. Twitter hashtags were identified from campaign websites and through exploratory searches of social media content. Hashtags identified during Twitter data collection that also appeared on TikTok were used to collect data on TikTok, with TikTok data supplemented by searches for platform-specific hashtags (Table S1 in Multimedia Appendix 1).

This study is reported in accordance with the *American Psychological Association* Journal Article Reporting Standards for Mixed Methods Research (JARS–Mixed Methods) [45].

Analytic Framework

Overview

Twitter and TikTok require distinct analytic approaches due to the text-based nature of the former and the multimodal nature (eg, video, audio, and text) of the latter. Machine learning techniques were only applied to Twitter, as text-based content is well suited for natural language processing techniques. In contrast, TikTok's multimodal format presents challenges for automated classification, necessitating manual content and thematic coding to ensure accurate interpretation.

To systematically evaluate social media discourse related to e-cigarette prevention campaigns, we applied a 4-component framework assessing (1) message volume, (2) message content, (3) message sources, and (4) engagement metrics [40].

Message Volume

The number of tweets (Twitter) and videos (TikTok) was measured over time and by campaign. We used linear regression to assess changes in campaign-related post volume over time on each platform (from January 2020 to December 2023 on TikTok and from February 2016 to October 2020 on Twitter).

Message Content

To capture the advocacy stance (opposition vs prevention or neutral) of tweets, we used a supervised machine learning classifier. Opposition themes were identified using thematic analysis to explore the underlying opposition narratives. For TikTok, content analysis was used to systematically classify the advocacy stance of videos (opposition vs prevention vs neutral) and to identify high-level opposition themes. To further explore nuanced rhetorical strategies and narrative patterns within opposition posts, we conducted a thematic analysis.

Message Sources

For Twitter, user type was determined using a machine learning approach, while TikTok user classification was achieved through content analysis.

Engagement Metrics

For Twitter, engagement was operationalized as the number of retweets each post received. For TikTok, "interactions" comprised the sum of likes, comments, and shares per video, while views were analyzed separately as a measure of reach. TikTok-specific visibility-boosting features were also analyzed, including the proportion of videos using the hashtag #fyp as well as interactive video formats: duets (playing videos side by side from different users) and stitches (embedding clips from existing videos). Given these platform differences, engagement metrics were not compared directly across platforms. The validity of content analyses was assessed using Gwet's AC1 statistic. The performance of the Twitter advocacy stance classification model was evaluated using F_1-scores.

Data Collection and Analysis

Twitter

We collected Twitter data using the now-deprecated Historical PowerTrack API, which provided enterprise-level access to the complete archive of public tweets matching defined rules on a monthly basis from January 2018 through April 2023. Our dataset includes all public English-language tweets (N=908,010) from February 2014 to October 2020 that matched 15 campaign-related hashtags associated with 14 distinct e-cigarette prevention campaigns, along with 4 verified campaign-affiliated accounts (Table S1 in Multimedia Appendix 1). Although the *Fresh Empire* campaign is focused on tobacco prevention, discussions of e-cigarettes appeared in posts using the campaign's hashtag. Additionally, experimental evidence indicates that tobacco prevention messaging can increase perceived effectiveness for discouraging e-cigarette use, suggesting spillover effects across products [46,47].

Opposition content was identified using a supervised machine learning classifier trained on 2282 human-labeled tweets. Coders were instructed to label tweets as "oppositional" only when they expressed support for vaping, rejected the goals of prevention campaigns, or aligned with provaping advocacy. They also considered full textual context, including tone and source, to avoid misclassifying neutral or critical mentions of hashtags. Posts were classified as oppositional if they met any of the following criteria: (1) use of hashtags explicitly rejecting the prevention efforts (eg, #notblowingsmoke); (2) use of hashtags promoting vaping (eg, #vapingsavedmylife and #vapelife); (3) reference to vaping-related advocacy (eg, #wevapewevote); or (4) mention of organizations known for supporting vaping or harm reduction, such as the Consumer Advocates for Smoke-Free Alternatives Association. The classifier achieved a macro-averaged F_1-score of 0.90 using 10-fold cross-validation, with a precision of 0.88 and recall of 0.92. An error analysis indicated that most misclassifications (false positives) occurred in cases where tweets mentioned health-related topics or campaign hashtags without clear oppositional framing.

XSL•FO
RenderX

TikTok

Data from 2020 to 2023 were collected in March 2025 in 2 phases using the TikTok Research API [48], an official platform-provided interface that allows approved researchers to access public content from designated geographic regions for research purposes. Our study focused on public content and restricted analysis to English-language posts, identified through metadata and manual review.

Phase 1 used 18 official hashtags and yielded 2563 posts. Phase 2 focused on 2 prominent national prevention campaign hashtags that were prominent in 2020: #ImmuneUpVapeDown and #ThisisQuitting, selected due to their prevalence and common spelling variations (eg, #thisisquitting was often misspelled as #thisisquiting; and #ImmuneUpVapesDown was also widely disseminated as #ImmuneUpVapeDown, a nonplural version of "Vape"). For *Immune Up*, both hashtag variations appeared in official promotional materials [49]. These 2 hashtags collectively generated over 2.1 million TikTok posts, from which we extracted as much content as application programming interface constraints allowed [50,51]. Hashtag queries were constructed using Boolean operators to include relevant variations and exclude unrelated campaigns.

To ensure our coded sample captured the full spectrum of content, from highly viral to more typical posts, we used quantile-stratified sampling based on total engagement (likes, comments, and shares combined). As engagement distributions on social media are heavily skewed, a simple random sample would risk underrepresenting the small number of high-impact posts that disproportionately shape discourse. We therefore ranked all 15,000 collected posts by engagement and divided them into 3 strata: the top 20% (high engagement), the middle 40% (moderate engagement), and the bottom 40% (low engagement). We then randomly sampled a proportionate number of posts from each stratum to generate our final coded sample of 800 posts. This approach ensured representation across visibility tiers and enhanced thematic diversity. All sampled posts were reviewed to exclude duplicates and nonsubstantive content. Across both phases, we collected 19,208 unique TikTok posts matching relevant prevention campaign hashtags. Of these, 15,000 (78.1%) posts with complete engagement metadata were eligible for ranking and quantile-stratified sampling. From this ranked set, we drew a proportionate stratified sample of 800 (5.3%) posts for manual coding. All posts from phase 1 of the data collection were manually labeled, resulting in a total of 3363 labeled posts. At the time of final analysis, 3127 (93.0%) labeled posts remained viewable and analyzable, with the remainder removed due to deletion and privacy restrictions. These removals accounted for approximately 7% of labeled TikTok posts and were treated as missing completely at random; no imputation was performed because missingness reflected post-level availability rather than content characteristics. A flow diagram summarizing the TikTok data collection and sampling workflow is provided in Figure S1 in Multimedia Appendix 1.

Content and Thematic Coding

Twitter

To develop opposition content themes, we used a thematic analysis approach. Two coders reviewed the top (highest tweet volume) 1000 oppositional messages from the most active users. We focused on the most active users, as their high volume of content is more likely to shape the overall narrative [52]. The coders then collaboratively identified themes, which were refined and grouped together until we arrived at four principal themes: (1) criticism of and mistrust in public health institutions (eg, claims of health authorities spreading inaccurate information and critiques of regulations), (2) claims of a boomerang effect of prevention advertisements (eg, advertisements triggering a desire to use e-cigarettes), (3) vaping rights and advocacy (eg, movements such as #ivapeivote), and (4) product promotion (eg, deals and coupons).

TikTok

For TikTok's short-form, video-first format, 5 trained coders developed and applied a structured codebook through an initial inductive coding phase, in which codes were generated directly from the data, followed by the integration of deductive codes informed by prior literature. The codebook included (1) user-level categories: account type (individual, organization, community [eg, fan accounts or news aggregators], and other), sector (health and government), and advocacy stance (protobacco or anticampaign and antitobacco or procampaign); and (2) video-level features: substance type (e-cigarette, other tobacco product, and other substance), advocacy stance, public service announcement framing (eg, conveying facts about substance harms or providing cessation resources), quit narratives (personal accounts of quitting a substance), brand mentions, product promotions, and entertainment or humor (overall; eg, skits, lip-syncing, jokes, and memes). Intercoder reliability was strong across all dimensions (Gwet AC1>0.81). To uncover more nuanced opposition narratives, we used the same thematic analysis approach applied to Twitter. Two coders jointly reviewed opposition posts (n=108), refining themes through iterative discussion, and identified six principal themes: (1) harm reduction or health claims (eg, mentions of any e-cigarette or tobacco-related health outcomes or comparative risk of tobacco smoking vs vaping), (2) personal testimony about vaping or tobacco benefits, (3) mockery of prevention campaigns (eg, using parody or satire), (4) dissatisfaction with advertisements, (5) ironic or hyperbolic references to vaping (eg, to comment on vaping culture or one's own vaping behaviors, including decisions to ignore prevention information), and (6) claims of a boomerang effect of prevention advertisements. The harm reduction or health claims were each evaluated for accuracy based on information from the Food and Drug Administration's Tobacco Education Resource Library [53]. Thematic saturation was reached when no new themes emerged during the final coding iterations.

Twitter User Classification

Twitter user classification followed a 2-step process. First, 2 coders labeled a random sample of 300 accounts (with up to 10 tweets per account) into 4 mutually exclusive categories that

XSL•FO
RenderX

were determined using a grounded theory approach: health (health-related organizations or individuals, including researchers and health care providers), commercial (organizations or brands explicitly selling or promoting tobacco products), provape advocacy (organizations or individuals advocating for pro–e-cigarette or tobacco policies or agendas), and organic (individuals with identity cues who are not public figures or influencers). Agreement was strong for the first 3 categories (AC1>0.80) and moderate for organic (AC1=0.72). Next, we used GPT-4o, accessed via an OpenAI instance on Microsoft Azure, to classify the same 300 accounts using structured prompts (Multimedia Appendix 2) and compared manual labels with GPT [54]. Accounts unrelated to the abovementioned categories or lacking sufficient information were classified as "other." Compared to human labels, GPT-4o achieved F_1-scores of 0.91 for organic and health accounts, 0.89 for advocacy accounts, and 0.96 for commercial accounts. On the basis of this validation, we applied the same prompting strategy to classify all unique users in the analytic data, using a unique user identifier available in the Twitter data. However, as Twitter display names were used as a feature for GPT's user classification and are nonunique (eg, users can change their display names), we opted to classify unique users in the Twitter data, sorted by highest follower count: this ensured we classified unique users using the most popular observation per unique identifier, where popularity is determined by the follower count. Final classifications were cleaned and merged with post-level data to analyze account type by post stance (opposition vs prevention or neutral). These classifications were used to contextualize quantitative patterns of opposition volume and engagement and to inform the interpretation of qualitative opposition themes across platforms. All codebooks used for data labeling are available upon request.

Ethical Considerations

All research activities for this study were approved by NORC at the University of Chicago's Institutional Review Board (IRB 23061378) and determined to meet the criteria for exemption under Exempt Category 4, as defined in US federal regulations (45 CFR 46.104) for using existing, publicly available social media data. The social media data were collected in accordance with each platform's terms of service and without direct interaction with users; therefore, informed consent was not required by the institutional review board. Privacy and confidentiality were safeguarded by ensuring that all data used in the analysis were publicly available and that all reported data were deidentified. No identifiable information or images of individual users are included in the manuscript or supplementary materials. Finally, no compensation was provided, as the study involved no direct interaction with or intervention upon human subjects.

Results

Campaign-Related Post Volume and Opposition Trends

We collected a total of 908,010 tweets referencing 14 e-cigarette prevention campaigns or mentioning the organization that produced the campaign from 2014 to 2020. After filtering for tweets explicitly referencing campaign names or hashtags, 310,207 (34.2%) tweets were retained for analysis. On TikTok (2020-2023), we analyzed 3127 viewable posts after excluding deleted, private, or non-English or non–US-based content.

Twitter campaign activity peaked during major campaign launches: March 2014 (n=9742) and August 2014 (n=10,867), corresponding with the rollout of *The Real Cost*, and January 2016 (n=14,989), coinciding with the launch of *Vaping Truth* (Figure 1). A peak in June 2017 (n=11,059) aligned with a *Fresh Empire*–sponsored event. From February 2016 to the end of the study period (October 2020), tweet volume declined significantly (β=–122 tweets per month, 95% CI –144.08 to –99.17; P<.001). On TikTok, campaign-related content exhibited smaller peaks, with the largest in October 2020 (n=995) during the launch of *Immune Up Vapes Down*. Post volume declined modestly but significantly from January 2020 to December 2023 (β=–5 posts per month, 95% CI: –7.91 to –1.02; P=.01).

Figure 1. Campaign-related tweet counts over time, overall and by the 5 campaigns with the highest tweet count (2014-2020). This figure illustrates the overall volume of discourse and key peaks corresponding to campaign launches.Table 1

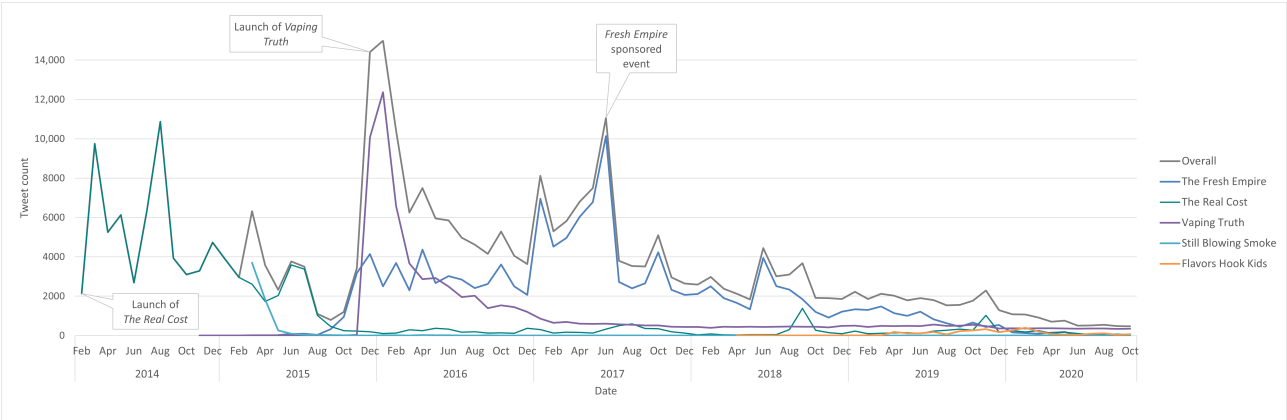

On Twitter, opposition content comprised 26.8% (83,074/310,207) of campaign-related posts, peaking in March 2015 (3766/6314, 59.7%) during the *Still Blowing Smoke* campaign and in January 2016 (11,821/14,989, 78.9%) during the *Vaping Truth* campaign. On TikTok, opposition was lower overall (108/3127, 3.5%) but rose modestly from 1.3% (27/2088) in 2020 to 5.6% (23/408) in 2023, peaking at 9.9% (28/282) of campaign-tagged posts in 2021, during *The Real Cost* campaign.

XSL•FO
RenderX

Opposition intensity varied by campaign and platform. On Twitter, among campaigns with more than 30 posts, state and local campaigns faced the most sustained counter-advocacy. Opposition comprised 99% (6052/6113) of content related to *Still Blowing Smoke*, 97.3% (70,300/72,270) of *Vaping Truth*–related content, and 27.3% (35/128) of messages about *The Vape Talk*. In contrast, national campaigns faced lower opposition: 3.4% (3125/93,246) for *The Real Cost* and 2.3% (3079/134,343) for *Fresh Empire* (Table 1). On TikTok, among campaigns with more than 30 posts, the highest opposition proportions were found in *The Real Cost* (56/246, 22.8%), *Vapes Down* (22/178, 12.4%), and *Ditch Juul* (4/42, 9.5%). Nine campaigns had limited TikTok presence (<30 posts).

Table . Proportion of campaign-related tweets classified as opposition by platform, overall and by campaign hashtag[a].

Platforms and Campaigns	Opposition, n (%)	Total, n (%)
Twitter		
Total	83,074 (26.8)	310,207 (100)
StillBlowingSmoke	6052 (99)	6113 (2)
VapingTruth	70,300 (97.3)	72,270 (23.3)
TheVapeTalk	35 (27.3)	128 (0.0)
VapingEquals	18 (15.5)	116 (0.0)
FlavorsHookKids	383 (12.9)	2974 (1)
TheDirtyTruth	22 (10.5)	209 (0.1)
IJustDidn’tKnow	15 (9.7)	155 (0.0)
VapeFreeNovember	42 (7.1)	592 (0.2)
FactsOverFlavor	2 (6.4)	31 (0.0)
TheRealCost	3125 (3.4)	93,246 (30.1)
FreshEmpire	3079 (2.3)	134,343 (43.3)
TikTok		
Total	108 (3.5)	3127 (100)
TheRealCost	56 (22.8)	246 (7.9)
VapesDown	22 (12.4)	178 (5.7)
DitchJuul	4 (9.5)	42 (1.3)
DotheVapeTalk	2 (6.1)	33 (1.1)
EscapeTheVape	2 (5)	40 (1.3)
DitchVape	7 (3)	230 (7.4)
ImmuneUpVapesDown	11 (0.9)	1290 (41.3)
ThisIsQuitting	5 (0.5)	1011 (32.3)

[a]Total Twitter figures include all tweets collected using 15 prevention campaign-related hashtags associated with 14 distinct campaigns. Total TikTok figures reflect the subset of labeled posts with viewable videos at the time of analysis. Only campaigns with >30 posts are shown individually; others are included in the total.

Engagement Patterns and Visibility-Boosting Features

On Twitter, retweets accounted for 59.7% (185,315/310,207) of campaign posts. Prevention tweets were retweeted 3.6 times more often than opposition posts. The most retweeted campaigns were *Still Blowing Smoke* (4977/6113, 81.4%) and *The Real Cost* (66,060/93,246, 70.8%). Opposition users also posted more frequently than prevention or neutral users (9.6 vs 1.6 posts per user). Engagement was highly concentrated: the top 10% of opposition accounts generated 84.4% (34,080/40,378) of opposition retweets, and the top 1% accounted for 63.9% (25,782/40,378).

On TikTok, opposition posts accounted for a smaller proportion of total posts (108/3127, 3.5%), compared to prevention posts (410/3127, 13.5%). Opposition posts also received lower average engagement (3103 vs 12,008 interactions (likes, comments, and shares). The median engagement was 40.0 (IQR 13.0 - 209.8) for prevention posts and 44.5 (IQR 16.8 - 128.0) for opposition posts. TikTok interactions (likes, comments, and shares) were driven by a small number of high-impact users: 97.9% (328,128/335,074) of opposition interactions came from just 11 accounts (the top 10%).

Opposition content used visibility-boosting platform features more frequently. The hashtag #fyp (“For Your Page”) appeared in 51.9% (56/108) of opposition posts, compared to 32.2% (132/410) of prevention posts. Use of interactive video formats, such as duets and stitches, was similar across conditions

XSL•FO
RenderX

(opposition posts: 6/108, 5.6% vs prevention posts: 20/410, 4.9%).

Opposition Themes

Prominent themes in opposition messaging varied by platform. On Twitter, key opposition themes included (1) criticism of and mistrust in public health institutions, (2) claims of a boomerang effect of prevention advertisements, (3) vaping rights and advocacy, and (4) product promotion. Many high-volume users aligned with advocacy movements such as #ivapeivote or #ABillionLives, framing vaping as a personal liberty or health-related right.

On TikTok, opposition content was primarily entertainment-oriented. The most frequently observed themes included general humor or entertainment (71/108, 65.7%), mockery of e-cigarette prevention campaigns (48/108, 44.4%), and ironic or hyperbolic references to vaping (30/108, 27.8%). One-quarter of opposition posts (27/108, 25%) visually depicted an e-cigarette. Expressions of dissatisfaction with prevention advertisements were present in 15.7% (17/108) of posts. Less frequently coded themes included claims of a boomerang effect of prevention advertisements (n=4, 3.7%), product promotion (n=2, 1.9%), personal testimony about the benefits of tobacco or vaping (n=2, 1.9%), and harm reduction or health claims (n=2, 1.9%). Notably, all health-related claims (n=3, 2.8%) contradicted established scientific consensus (eg, people with asthma are typically found to be healthier when they vape) [53].

Content Creator and Source Analysis

On Twitter, among users with nonmissing profile descriptions, organic users constituted the majority (42,446/49,809, 85.2%), followed by other (6850/49,809, 13.8%), health (2264/49,809, 4.5%), provape advocate (1570/49,809, 3.2%), and commercial accounts (951/49,809, 1.9%). However, among opposition posters, provape advocacy (1376/5465, 25.2%) and commercial (724/5465, 13.2%) accounts were overrepresented, and organic users (2926/5465, 53.5%) were underrepresented compared to overall figures.

Among accounts posting prevention or neutral content, representation by health-related sources was 4.8% (2139/44,344).

On TikTok, individuals made up 91.8% (2023/2204) of labeled users, community accounts made up 3.1% (69/2204), organizations made up 2.8% (61/2204), and other types of accounts made up 2.3% (51/2204). A small proportion (29/2204, 1.3%) of users were categorized as health- or government-affiliated accounts.

Among opposition posters, individuals were the most prevalent (95/103, 92.2%), followed by community accounts (4/103, 3.9%), organizations (2/103, 1.9%), and other types of accounts (2/103, 1.9%). There was no representation by health or government sectors. Provape advocacy represented 6.8% (7/103) of opposition accounts.

Among accounts posting prevention messages, 76.1% (153/201) were individuals, 14.4% (29/201) were organizations, 5% (10/201) were community accounts, and 4.5% (9/201) were some other type of account. Health-related accounts constituted 10% (20/201) of these accounts, while 0.5% (1/201) were government related.

Hashtag Variation and Engagement

Campaign engagement was shaped by both deliberate co-optation strategies and hashtag variation. Most prominently, opposition to *Still Blowing Smoke* on Twitter was marked by extensive use of the counter-hashtag #notblowingsmoke [32], which highlighted the claim that e-cigarette vapor is less dangerous than cigarette smoke and appeared alongside the official campaign hashtag #stillblowingsmoke in 53.7% (3248/6052) of the campaign's opposition tweets.

On TikTok, engagement patterns were affected by the concurrent use of multiple similar hashtags associated with the same campaign. Across the study period, #ImmuneUpVapeDown was tagged in 3552 posts, while #ImmuneUpVapesDown appeared in 1184 posts. Although #ImmuneUpVapesDown appeared less frequently, posts using this hashtag generated higher average engagement, with 11,271 views and 1380 engagements per post, compared to 7831 views and 1105 engagements per post for #ImmuneUpVapeDown. This reflects a 1.4-fold increase in views and a 1.3-fold increase in engagements for #ImmuneUpVapesDown.

A similar pattern was observed for the *This is Quitting* campaign. The official hashtag, #thisisquitting, appeared in 13,114 posts, whereas the common misspelling, #thisisquiting, appeared in 434 posts. Despite its lower volume, posts with the misspelled hashtag generated slightly higher average engagement (1653 vs 1536 interactions per post). However, these posts received lower average viewership compared to posts with the official hashtag (10,877 vs 14,163 views per post).

Discussion

Principal Findings

Our study offers one of the first cross-platform examinations of digital opposition to youth e-cigarette use prevention campaigns, identifying distinct structures, engagement patterns, and thematic frames across Twitter and TikTok. We found that opposition was more prevalent and coordinated on Twitter, driven by commercial and advocacy accounts. In contrast, TikTok opposition was decentralized and humor-driven. Despite these differences, prevention content generally achieved higher engagement than oppositional posts on both platforms. These findings highlight the extent to which oppositional activity is shaped not only by campaign content but also by the unique affordances, user cultures, and algorithmic logics of each platform. Understanding these dynamics is essential. Responsive and resilient prevention strategies must be tailored to the evolving digital media landscape.

Opposition took divergent forms across platforms, reflecting structural and cultural asymmetries. On Twitter, opposition was highly coordinated and driven by commercial and advocacy accounts that leveraged hashtag hijacking and volume-based amplification. These patterns reflect classic principal-agent dynamics and networked gatekeeping. In both cases, vested interests influence which narratives gain visibility and shape

XSL•FO
RenderX

policy discourse at scale. Organized advocacy networks and commercial promotion directly contested health communication and regulatory narratives. Their messages reflected direct challenges to health institutions, calls for deregulation, claims of a boomerang effect, harm reduction narratives, and product promotion.

In contrast, TikTok opposition was decentralized and creator-led, with individual users relying on algorithmically favored formats, such as humor, parody, and remix. Rather than mounting explicit policy critiques, TikTok users often reframed prevention campaigns as objects of satire or trivialization through "duets" that mock campaign messages, lip-syncing to songs about e-cigarette use, or comedic skits. Twitter's text-centric environment and low production barriers facilitate more textual coordination, while TikTok's video-first interface favors performative and affective engagement [55-57].

Prior research [32-35] has documented Twitter "bombing" or astroturfing and branding appropriation; our findings suggest that TikTok's entertainment-driven opposition presents a structurally distinct challenge to campaign coherence. Importantly, public health institutional accounts comprised only 10% of all prevention-aligned TikTok content, indicating that the discourse is overwhelmingly dominated by individual creators. While this does not speak directly to impact or effectiveness, it underscores the relatively modest presence of institutional voices within platform-native discourse. The prevalence of mockery on TikTok, with the opposition's significantly higher use of visibility-boosting hashtags like #fyp, aligns with an algorithmic logic that privileges engagement over informational accuracy. This makes evidence-based public health content less salient unless adapted to the platform's affective and performative norms. The presence of both coordinated and decentralized opposition across platforms suggests that prevention campaigns face 2 qualitatively distinct credibility challenges: organized counter-messaging that seeks to influence policy and diffuse creator-driven content that may erode credibility through parody and cultural reframing.

Engagement dynamics further underscore the differences in the nature of opposition between platforms. On Twitter, opposition users posted or retweeted repeatedly, suggesting coordinated persistence. On TikTok, opposition was more episodic and reliant on the virality of individual posts. Although opposition content more frequently used visibility-boosting features such as #fyp, it did not consistently outperform prevention content in engagement, indicating that tagging alone does not guarantee amplification. These dynamics reflect how each platform's algorithmic logic privileges particular forms of participation. Twitter rewards repeated real-time posting, trending topics, and follower networks, while TikTok prioritizes personalized content that aligns with user engagement patterns and preferred entertainment formats [55,56].

This pattern of engagement is further illustrated by the concentration of opposition interactions among a small number of high-impact TikTok users. Most opposition interactions (likes, comments, and shares) came from just 11 accounts representing the top 10% of opposition users. This aligns with TikTok's influencer-driven model, where a small number of content creators can disproportionately shape discourse and visibility. For prevention campaigns, this suggests a strategic opportunity to increase engagement. This can be achieved by forming partnerships with credible, platform-native, youth-aligned influencers who can compete with oppositional voices in reach rather than relying on broad content distribution.

Notably, opposition visibility did not uniformly suppress engagement with prevention content, particularly on TikTok, suggesting that volume alone does not determine engagement and impact. This underscores the need for platform-specific intervention approaches. On Twitter, effective approaches should focus on detecting coordinated opposition and preemptively framing policy discourse to counter specific antiregulatory narratives. On TikTok, campaigns should invest in creative, platform-native formats, such as duets, remixes, or trending audio strategies, to maximize resonance with youth audiences while expanding public health visibility within the algorithmic ecosystem.

Future research should examine how competing narratives influence youth perceptions of e-cigarette use risk, assess the impact of unsupported and unauthorized health claims on public understanding, and explore regulatory approaches to moderating coordinated opposition. Further work is needed to examine how evolving algorithmic infrastructures govern the diffusion and visibility of public health content, including how oppositional discourse adapts in increasingly sophisticated ways. In addition, campaign-by-campaign thematic analyses could help disentangle the influence of specific prevention messages from broader platform-level dynamics and determine whether certain prevention themes predict opposition. These future efforts would also benefit from deeper integration of theory, particularly frameworks such as networked gatekeeping to better explain the structural vulnerabilities of public health messaging online.

Across both platforms, disruptive hashtag dynamics emerged in two forms: (1) strategic co-optation or hijacking (eg, #notblowingsmoke) and (2) hashtag fragmentation, which included both user-generated "drift" (ie, unintentional spelling variations such as #thisisquiting) and inconsistencies introduced by campaigns themselves (eg, simultaneous promotion of 2 similar but nonidentical hashtags for a single campaign). While hashtag drift can sometimes reflect authentic engagement and meme-like spread, inconsistent use of multiple official hashtags introduces risks. It can confuse the audience, dilute message visibility, and complicate evaluation. To address this, campaigns should promote a single, memorable, and unique hashtag consistently across all materials and partner networks. This helps minimize the hashtag-level fragmentation of campaign content visibility reflected in our findings and strengthens the integrity of message tracking.

Real-time monitoring can help detect and address hijacking and hashtag drift. Campaign staff should set up real-time alerts using social listening tools for sudden spikes in volume or shifts in sentiment. Such monitoring could signal a coordinated hijacking effort like the one seen with some counter-campaigns observed in regional campaigns. Quantitative monitoring should be paired with daily qualitative reviews of the content being posted under the hashtag to understand how the narrative is evolving.

XSL•FO
RenderX

Finally, prevention campaigns need preplanned rapid response protocols to counter oppositional messages. These should extend beyond reactive posts. There should be a preplanned communications protocol that may include a playbook of preapproved messages, coordinated content amplification, and activating credible messengers. Such practices are especially important in environments such as Twitter, where opposition is coordinated and can quickly flood the e-cigarette prevention discourse. It is also crucial in environments where algorithmic feeds (eg, TikTok's For You Page) may amplify small differences in framing, format, or metadata, thereby influencing content visibility and diffusion in unpredictable ways.

Limitations

This study has several limitations. First, platform-level constraints, particularly inconsistencies in the TikTok Research API, limited the volume of analyzable content, especially for high-volume campaigns. In addition, it remains unclear whether promoted or sponsored advertisement content is consistently included via the application programming interface; if excluded, our analysis may underrepresent campaign materials delivered through paid placements. Second, while observed differences in discourse across platforms could be attributed to structural or programmatic features of the platforms, they could also reflect variation in data collection periods or user demographics. As Twitter and TikTok data were collected from nonoverlapping periods (2014 - 2020 and 2020 - 2023, respectively), our goal was not to directly compare levels of opposition or engagement across time. Rather, each dataset reflects the dominant usage period of the platform for prevention campaigns, and our analysis focuses on platform-specific opposition dynamics within those distinct temporal contexts. Third, while engagement metrics provide useful indicators of content visibility and performance, they do not capture user intent, sentiment, or behavioral outcomes. Therefore, high engagement should not be equated with either campaign impact or message suppression. Additionally, Twitter user classification relied on mutable display names rather than persistent identifiers, which constrained our ability to characterize user behavior at scale. Finally, our analyses do not take into account secular trends in platform activity, such as the long-run decline in Twitter activity between 2014 and 2019. As our analysis is primarily concerned with proportions and themes of content, this limitation does not substantively impact our results [58].

Conclusions

This study systematically compared oppositional dynamics across e-cigarette prevention campaigns on 2 major social media platforms using a mixed methods approach. Unlike previous studies that focused on singular platforms or campaigns, this study compared several e-cigarette prevention campaigns on Twitter and TikTok to reveal how distinct sociotechnical architectures shape opposition strategies. Campaign resilience in digital environments depends on more than message design alone. Our findings offer actionable guidance to strengthen digital health communication across platforms. Public health institutions must adapt to platform-specific architectures. On Twitter, strategies to counter opposition content should include real-time surveillance to enable rapid response to coordinated opposition. On TikTok, public health entities may improve engagement by investing in creator partnerships and using visibility-enhancing content formats. Sustained investment in creator collaboration is especially critical to counter entertainment-driven opposition narratives with content forms already familiar and appealing to youth audiences. Across both platforms, campaigns should ensure consistent branding to prevent potential audience fragmentation and message dilution from hashtag variations. Additionally, given the relatively small presence of credible health sources in the prevention discourse, they should be used more frequently to reinforce prevention narratives. On Twitter, credible health voices are essential for combating coordinated advocacy networks, while on TikTok, they should complement influencer partnerships and entertainment-oriented formats.

Meanwhile, regulatory agencies and health campaigns should also advocate platform accountability, including algorithmic transparency. This entails, at minimum, requiring platforms to disclose whether prevention content is systematically downranked compared to commercial promotion and to provide greater elevation of verified public health sources.

Acknowledgments

We would like to acknowledge Alex Kresovich and Haoyu Shi for their contributions to this research.

GPT-4o was used for Twitter user classification, as described in the *Methods* section. All outputs were reviewed and verified by the authors. Grammarly was used for spelling and grammar assistance; no generative artificial intelligence tools were used to draft substantive content or interpret results.

Funding

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health (grant R01CA283038). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The social media data that support the findings of this study are not publicly available due to third-party restrictions and platform terms of service. Data were obtained via the X (formerly Twitter) API and TikTok Research API for approved research use.

XSL•FO
RenderX

Access to these data can be requested from the respective platforms and is subject to their individual data-sharing policies and researcher eligibility criteria.

Authors' Contributions

CCC, SP, GK, SLE, and MF together designed and planned the study. CCC, SP, and MF conducted data analysis. CCC, SP, MB, and MF contributed to data interpretation. CCC, SP, MB, and MF wrote the first draft. CCC, SP, MB, GK, SLE, and MF revised the draft. The final version of the paper has been reviewed and approved by all authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
TikTok data collection and sampling workflow and overview of e-cigarette social media campaigns by hashtag, account handle, region, platform, and post count.
[PDF File, 149 KB - jmir_v28i1e83791_app1.pdf]

Multimedia Appendix 2
ChatGPT prompt for Twitter user classification.
[PDF File, 103 KB - jmir_v28i1e83791_app2.pdf]

References

1. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. E-Cigarette Use Among Youth and Young Adults: A Report of the Surgeon General: Centers for Disease Control and Prevention; 2016. [Medline: 30869850]
2. McGrath-Morrow SA, Gorzkowski J, Groner JA, et al. The effects of nicotine on development. Pediatrics 2020 Mar;145(3):e20191346. [doi: 10.1542/peds.2019-1346] [Medline: 32047098]
3. Visser W, Klerx W, Geraets L, et al. The health risks of using e-cigarettes. : RIVM, Netherlands National Institute for Public Health and the Environment; 2015 URL: https://www.rivm.nl/bibliotheek/rapporten/2015-0144.pdf [accessed 2026-02-26]
4. The real cost e-cigarette prevention campaign. US Food and Drug Administration. 2025. URL: https://www.fda.gov/tobacco-products/real-cost-campaign/real-cost-e-cigarette-prevention-campaign [accessed 2026-02-26]
5. Martinelli T, Candel M, de Vries H, et al. Exploring the gateway hypothesis of e-cigarettes and tobacco: a prospective replication study among adolescents in the Netherlands and Flanders. Tob Control 2023 Mar;32(2):170-178. [doi: 10.1136/tobaccocontrol-2021-056528] [Medline: 34226262]
6. FDA finalizes enforcement policy on unauthorized flavored cartridge-based e-cigarettes that appeal to children, including fruit and mint. US Food & Drug Administration. 2020 Jan. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/enforcement-priorities-electronic-nicotine-delivery-system-ends-and-other-deemed-products-market [accessed 2026-02-26]
7. FDA notifies companies, including puff bar, to remove flavored disposable e-cigarettes and youth-appealing e-liquids from market for not having required authorization. PR Newswire. 2020. URL: https://www.prnewswire.com/news-releases/fda-notifies-companies-including-puff-bar-to-remove-flavored-disposable-e-cigarettes-and-youth-appealing-e-liquids-from-market-for-not-having-required-authorization-301096252.html [accessed 2026-02-26]
8. Ali FRM, Seidenberg AB, Crane E, Seaman E, Tynan MA, Marynak K. E-cigarette unit sales by product and flavor type, and top-selling brands, United States, 2020-2022. MMWR Morb Mortal Wkly Rep 2023 Jun 23;72(25):672-677. [doi: 10.15585/mmwr.mm7225a1] [Medline: 37347717]
9. Farrelly MC, Davis KC, Haviland ML, Messeri P, Healton CG. Evidence of a dose-response relationship between “truth” antismoking ads and youth smoking prevalence. Am J Public Health 2005 Mar;95(3):425-431. [doi: 10.2105/AJPH.2004.049692] [Medline: 15727971]
10. Bal DG, Kizer KW, Felten PG, Mozar HN, Niemeyer D. Reducing tobacco consumption in California. Development of a statewide anti-tobacco use campaign. JAMA 1990 Sep 26;264(12):1570-1574. [doi: 10.1001/jama.1990.03450120082034] [Medline: 2395199]
11. Feng M, Pierce JP, Szczypka G, Vera L, Emery S. Twitter analysis of California’s failed campaign to raise the state’s tobacco tax by popular vote in 2012. Tob Control 2017 Jul;26(4):434-439. [doi: 10.1136/tobaccocontrol-2016-053103] [Medline: 27601452]
12. McAlister A, Morrison TC, Hu S, et al. Media and community campaign effects on adult tobacco use in Texas. J Health Commun 2004;9(2):95-109. [doi: 10.1080/10810730490425231] [Medline: 15204821]

XSL•FO
RenderX

13. Emery S, Kim Y, Choi YK, Szczypka G, Wakefield M, Chaloupka FJ. The effects of smoking-related television advertising on smoking and intentions to quit among adults in the United States: 1999-2007. Am J Public Health 2012 Apr;102(4):751-757. [doi: 10.2105/AJPH.2011.300443] [Medline: 22397350]
14. Farrelly MC, Nonnemaker J, Davis KC, Hussin A. The Influence of the National truth campaign on smoking initiation. Am J Prev Med 2009 May;36(5):379-384. [doi: 10.1016/j.amepre.2009.01.019] [Medline: 19211213]
15. Terry-McElrath YM, Wakefield MA, Emery S, et al. State anti-tobacco advertising and smoking outcomes by gender and race/ethnicity. Ethn Health 2007 Sep;12(4):339-362. [doi: 10.1080/13557850701300723] [Medline: 17701761]
16. Wakefield M, Terry-McElrath Y, Emery S, et al. Effect of televised, tobacco company-funded smoking prevention advertising on youth smoking-related beliefs, intentions, and behavior. Am J Public Health 2006 Dec;96(12):2154-2160. [doi: 10.2105/AJPH.2005.083352] [Medline: 17077405]
17. Durkin SJ, Brennan E, Wakefield MA. Optimising Tobacco Control Campaigns Within a Changing Media Landscape and Among Priority Populations: Tobacco Control BMJ Publishing Group Ltd; 2022. [doi: 10.1136/tobaccocontrol-2021-056558]
18. Durkin S, Brennan E, Wakefield M. Mass media campaigns to promote smoking cessation among adults: an integrative review. Tob Control 2012 Mar;21(2):127-138. [doi: 10.1136/tobaccocontrol-2011-050345] [Medline: 22345235]
19. Diaz MC, Kierstead EC, Edwards D, et al. Online tobacco advertising and current chew, dip, snuff and snus use among youth and young adults, 2018–2019. Int J Environ Res Public Health 2022 Apr 14;19(8):4786. [doi: 10.3390/ijerph19084786] [Medline: 35457651]
20. Huang J, Duan Z, Kwok J, et al. Vaping versus JUULing: how the extraordinary growth and marketing of JUUL transformed the US retail e-cigarette market. Tob Control 2019 Mar;28(2):146-151. [doi: 10.1136/tobaccocontrol-2018-054382]
21. Vassey J, Valente T, Barker J, et al. E-cigarette brands and social media influencers on Instagram: a social network analysis. Tob Control 2023 Aug;32(e2):e184-e191. [doi: 10.1136/tobaccocontrol-2021-057053] [Medline: 35131947]
22. Liang Y, Zheng X, Zeng DD, Zhou X, Leischow SJ, Chung W. Exploring how the tobacco industry presents and promotes itself in social media. J Med Internet Res 2015 Jan 21;17(1):e24. [doi: 10.2196/jmir.3665] [Medline: 25608524]
23. Donaldson SI, Dormanesh A, Perez C, Majmundar A, Allem JP. Association between exposure to tobacco content on social media and tobacco use: a systematic review and meta-analysis. JAMA Pediatr 2022 Sep 1;176(9):878-885. [doi: 10.1001/jamapediatrics.2022.2223] [Medline: 35816331]
24. Kostygina G, Kim Y, Gebhardt Z, et al. Using exogenous social media exposure measures to assess the effects of smokeless tobacco-related social media content on smokeless tobacco sales in the United States. Nicotine Tob Res 2024 Feb 15;26(Supplement_1):S49-S56. [doi: 10.1093/ntr/ntad169] [Medline: 38366341]
25. Manganello J, Bleakley A, Schumacher P. Pandemics and PSAs: rapidly changing information in a new media landscape. Health Commun 2020 Dec;35(14):1711-1714. [doi: 10.1080/10410236.2020.1839192] [Medline: 33106041]
26. Kite J, Chan L, MacKay K, et al. A model of social media effects in public health communication campaigns: systematic review. J Med Internet Res 2023 Jul 14;25(1):e46345. [doi: 10.2196/46345] [Medline: 37450325]
27. Southwell BG, Yzer MC. When (and why) interpersonal talk matters for campaigns. Commun Theory 2009 Feb;19(1):1-8 [FREE Full text] [doi: 10.1111/j.1468-2885.2008.01329.x]
28. Feng M, Carter CC, Page S, Emery SL, Tran H, Kostygina G. Tweeted, trolled, twisted: battling for narrative control in e-cigarette use prevention campaigns (2014-2020). J Health Commun 2025 Mar 28;30(sup1):39-49. [doi: 10.1080/10810730.2025.2462682] [Medline: 39915966]
29. Majmundar A, Le N, Moran MB, Unger JB, Reuter K. Public response to a social media tobacco prevention campaign: content analysis. JMIR Public Health Surveill 2020 Dec 7;6(4):e20649. [doi: 10.2196/20649] [Medline: 33284120]
30. Liang Y, Zheng X, Zeng DD, Zhou X, Leischow SJ, Chung W. Characterizing social interaction in tobacco-oriented social networks: an empirical analysis. Sci Rep 2015 Jun 19;5(1):10060. [doi: 10.1038/srep10060] [Medline: 26091553]
31. Harris JK, Moreland-Russell S, Choucair B, Mansour R, Staub M, Simmons K. Tweeting for and against public health policy: response to the Chicago Department of Public Health's electronic cigarette Twitter campaign. J Med Internet Res 2014 Oct 16;16(10):e238. [doi: 10.2196/jmir.3622] [Medline: 25320863]
32. Allem JP, Escobedo P, Chu KH, Soto DW, Cruz TB, Unger JB. Campaigns and counter campaigns: reactions on Twitter to e-cigarette education. Tob Control 2017 Mar;26(2):226-229. [doi: 10.1136/tobaccocontrol-2015-052757] [Medline: 26956467]
33. Leas EC, Prochaska JJ, Ayers J, Nobles AL, Henriksen L. What to do when tobacco advertisers exploit antitobacco social media campaigns to sell tobacco. Tob Control 2020 Mar;29(2):243-244. [doi: 10.1136/tobaccocontrol-2019-054993] [Medline: 31249102]
34. Keller FB, Schoch D, Stier S, Yang J. Political astroturfing on Twitter: how to coordinate a disinformation campaign. Polit Commun 2020 Mar 3;37(2):256-280. [doi: 10.1080/10584609.2019.1661888]
35. Chan J. Online astroturfing: a problem beyond disinformation. Philos Soc Crit 2024 Mar;50(3):507-528. [doi: 10.1177/01914537221108467]
36. McCombs ME, Shaw DL. The agenda-setting function of mass media. Public Opin Q 1972;36(2):176. [doi: 10.1086/267990]
37. Meraz S, Papacharissi Z. Networked gatekeeping and networked framing on #Egypt. Int J Press Polit 2013 Apr;18(2):138-166. [doi: 10.1177/1940161212474472]

XSL•FO
RenderX

38. Laffont JJ, Martimort D. The Theory of Incentives: The Principal-Agent Model: Princeton University Press; 2009. [doi: 10.2307/j.ctv7h0rwr]
39. Byrne S, Hart PS. The boomerang effect A synthesis of findings and a preliminary theoretical framework. Annals of the International Communication Association 2009 Jan;33(1):3-37. [doi: 10.1080/23808985.2009.11679083]
40. Feng M, Shi H, Page S, et al. Analyzing U.S. e-cigarette prevention campaigns on TikTok (2020-2023). Presented at: World Association for Public Opinion Research's 77th Annual Conference; Jul 28-31, 2024.
41. Jung S, Murthy D, Bateineh BS, Loukas A, Wilkinson AV. The normalization of vaping on TikTok using computer vision, natural language processing, and qualitative thematic analysis: mixed methods study. J Med Internet Res 2024 Sep 11;26(1):e55591. [doi: 10.2196/55591] [Medline: 39259963]
42. Al-Rawi A, Blackwell B, Zemenchik K, Lee K. Twitter misinformation discourses about vaping: systematic content analysis. J Med Internet Res 2023 Nov 10;25(1):e49416. [doi: 10.2196/49416] [Medline: 37948118]
43. West SM. Raging against the machine: network gatekeeping and collective action on social media platforms. MaC 2017 Sep 22;5(3):28-36. [doi: 10.17645/mac.v5i3.989]
44. The real cost campaign. US Food and Drug Administration. 2017. URL: https://www.fda.gov/tobacco-products/public-health-education-campaigns/real-cost-campaign [accessed 2026-02-26]
45. Levitt HM, Bamberg M, Creswell JW, Frost DM, Josselson R, Suárez-Orozco C. Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):26-46. [doi: 10.1037/amp0000151] [Medline: 29345485]
46. Grummon AH, Hall MG, Mitchell CG, et al. Reactions to messages about smoking, vaping and COVID-19: two national experiments. Tob Control 2022 May;31(3):402-410. [doi: 10.1136/tobaccocontrol-2020-055956] [Medline: 33188150]
47. Kowitt SD, Mendel Sheldon J, Vereen RN, et al. The impact of the real cost vaping and smoking ads across tobacco products. Nicotine Tob Res 2023 Feb 9;25(3):430-437. [doi: 10.1093/ntr/ntac206] [Medline: 36006858]
48. TikTok for developers. Research API. URL: https://developers.tiktok.com/products/research-api/ [accessed 2025-08-01]
49. Truth takes to TikTok to talk vaping and the immune system. truth initiative. 2020. URL: https://truthinitiative.org/press/press-release/truth-takes-tiktok-talk-vaping-and-immune-system-during-covid-19 [accessed 2025-07-28]
50. Vanian J. Apple, Google remove TikTok from stores as app halts service in US. CNBC. 2025. URL: https://www.cnbc.com/2025/01/18/apple-google-remove-tiktok-from-stores-as-app-halts-service-in-us.html [accessed 2026-02-26]
51. Pearson GDH, Silver NA, Robinson JY, Azadi M, Schillo BA, Kreslake JM. Beyond the margin of error: a systematic and replicable audit of the TikTok research API. Information, Communication & Society 2025 Feb 17;28(3):452-470. [doi: 10.1080/1369118X.2024.2420032]
52. Guimaraes A, Wang L, Weikum G. Us and them: adversarial politics on Twitter. Presented at: 2017 IEEE International Conference on Data Mining Workshops (ICDMW); Nov 18-21, 2017; New Orleans, LA p. 872-877. [doi: 10.1109/ICDMW.2017.119]
53. Vaping prevention resources tobacco education resource library. US Food and Drug Administration. URL: https://digitalmedia.hhs.gov/tobacco/educator_hub [accessed 2025-07-28]
54. ChatGPT [large language model]. OpenAI. 2025. URL: https://chat.openai.com/chat [accessed 2026-02-26]
55. Tufekci Z. Twitter and Tear Gas: The Power and Fragility of Networked Protest: Yale University Press; 2017.
56. Abidin C. Mapping internet celebrity on TikTok: exploring attention economies and visibility labours. Cultural Science Journal 2020 Jan 19;12(1):77-103. [doi: 10.5334/csci.140]
57. Literat I, Kligler-Vilenchik N. TikTok as a key platform for youth political expression: reflecting on the opportunities and stakes involved. Social Media + Society 2023 Jan;9(1):20563051231157595. [doi: 10.1177/20563051231157595]
58. Visualizing a decade of Twitter's evolution: Jan 2012 – Nov 2022 & Musk's purchase. The GDELT Project. URL: https://blog.gdeltproject.org/visualizing-a-decade-of-twitters-evolution-jan-2012-nov-2022-musks-purchase/ [accessed 2025-11-06]

Edited by S Brini; submitted 08.Sep.2025; peer-reviewed by RR Ouellette, Y Wang; accepted 01.Jan.2026; published 26.Mar.2026.

Please cite as:
Carter CC, Page S, Borowiecki M, Kostygina G, Emery SL, Feng M
Opposition to Youth e-Cigarette Prevention Campaigns on Twitter and TikTok: Cross-Platform Observational Mixed Methods Analysis
J Med Internet Res 2026;28:e83791
URL: https://www.jmir.org/2026/1/e83791
doi:10.2196/83791

©Chandler C Carter, Simon Page, Mateusz Borowiecki, Ganna Kostygina, Sherry L Emery, Miao Feng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted

XSL•FO
RenderX

use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Suicidal Thoughts and Behaviors Among Chinese Adolescents in Relation to Negative Life Events, Internet Addiction, and Sexual Abuse: Cross-Sectional Study

Juanfang Zhu[1*], MPH; Qinqin Jiang[1,2*], MHA; Feng Zhang[1*], MPH; Hao Liu[1], MHA; Shaojie Yu[1], MPH; Jinhai Sun[1], PhD; Zhe Zhao[1], MPH; Lijuan Liu[1], PhD; Lei Yuan[1], MPH

[1]Faculty of Military Health Service, Naval Medical University, No. 800 Xiangyin Road, Shanghai, China

[2]College of Health Management, Southern Medical University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Lei Yuan, MPH
Faculty of Military Health Service, Naval Medical University, No. 800 Xiangyin Road, Shanghai, China

Abstract

Background: Increasing suicidal thoughts and behaviors (STB) among adolescents raise social concerns and have a well-recognized association with sexual abuse (SA). However, research regarding the mechanisms explaining the association between SA and STB remains limited.

Objective: This study aims to examine the chained mediating effects of negative life events (NLE) and internet addiction (IA) between SA and STB among adolescents in China.

Methods: This cross-sectional study used data from the Science Database of the People Mental Health survey conducted between March 2013 and December 2022 by the National Population Health Data Center of the National Research Institute for Family Planning. Through stratified sampling, 20,893 adolescents were recruited from 16 Chinese provinces. After excluding samples with missing relevant variables, 10,664 (55.89%; aged 16-17.9 y; n=5826, 54.63% women) adolescents were included in the final analysis. STB was the outcome variable, with NLE and IA as mediators, all assessed via a questionnaire that was uniformly administered by trained investigators in school settings. The Pearson χ^2 test was used to analyze the association between SA and STB. Using a combination of multiple linear regression and bootstrap testing, the study constructed a chain mediation model to explore how SA influences STB in adolescents through NLE and IA.

Results: The scores for SA, NLE, IA, and STB were 1.330 (SD 1.714), 51.960 (SD 23.822), 34.88 (SD 13.852), and 0.690 (SD 1.396), respectively. Multiple linear regression analysis indicated SA was associated with NLE (β=2.382, 95% CI 2.112 - 2.653; *P*<.001), IA (β=1.094, 95% CI 0.939 - 1.249; *P*<.001), and STB (β=0.135, 95% CI 0.119 - 0.151; *P*<.001). Mediation analysis revealed a significant direct association between SA and STB (74.81%; β=0.101, 95% CI 0.085 - 0.117; *P*<.001). Indirect effects constituted 25.19% of the total association (β=0.034, 95% CI 0.029 - 0.040; *P*<.001), comprised of 3 specific pathways: via NLE (11.85%; β=0.016, 95% CI 0.012 - 0.020; *P*<.001), via IA (10.37%; β=0.014, 95% CI 0.010 - 0.018; *P*<.001), and via the sequential pathway through NLE and then IA (2.96%; β=0.004, 95% CI 0.003 - 0.006; *P*<.001).

Conclusions: Unlike previous research that examined them separately, this study reveals the chain mediation mechanism of NLE and IA in the relationship between SA and STB, providing a new research perspective on the complex evolution from early adversity to extreme outcomes. This research also provided clear direction and scientific evidence for the development of comprehensive suicide prevention strategies. Future efforts should focus on the prevention of SA and trauma intervention while simultaneously strengthening adolescents' cognitive reappraisal of negative events and guiding their online behavior.

(J Med Internet Res 2026;28:e85371) doi:10.2196/85371

KEYWORDS

suicidal thoughts; suicidal behaviors; sexual abuse; SA; internet addiction; IA; chain mediation effect; adolescents

Introduction

Adolescents' susceptibility to suicidal thoughts and behaviors (STB) raises social and academic concerns [1-3]. According to the World Health Organization (WHO) estimates, suicide was the fourth leading cause of mortality for youths aged 15 to 29 years globally in 2019 [4], with significant variations observed across countries, age groups, and genders [5]. Both attempted and completed suicides impose substantial social, economic, and emotional burdens worldwide. Despite efforts to reduce

XSL•FO
RenderX

suicide risk, global rates of youth STB exhibit no improvement [6]. Hence, the identification of risk factors and antecedents predisposing individuals to increased suicide risk is imperative to enable prompt and appropriate monitoring and intervention [7]. Sexual abuse (SA) is a recognized risk factor for suicidal ideation and behavior [8,9]. By disrupting normal neuroendocrine and immune functions [10], SA exerts a direct effect on increasing suicide risk [11]. Accordingly, this study proposes hypothesis 1 as follows: SA is positively associated with STB in adolescents. Although this association is well-established, the mediating mechanisms through which SA elevates suicide risk remain unclear, particularly among Chinese adolescents.

Based on theoretical reasoning, negative life events (NLE) and internet addiction (IA) may serve as potential mediators. According to the integrated motivational-volitional model of STB, NLE are the main stressors driving STB, which play a "trigger" role [12] and are strong predictors thereof [13]. Previous cross-sectional studies have linked NLE such as academic pressure and interpersonal conflict to suicidal ideation and behavior in university students [14,15], an association also supported by longitudinal research [16]. Furthermore, evidence suggests that survivors of SA are more likely to experience other forms of interpersonal harm, including physical assault [17], illicit drug dependence [18], and adolescent dating violence [19]—all constituting NLE. Thus, hypothesis 2 is proposed as follows: NLE mediate the relationship between SA and STB. Recent studies indicate a significant association between adolescent IA and suicidal ideation or behavior [17,20,21], with addicted adolescents exhibiting higher suicide risk. A key risk factor for IA may be childhood trauma, which is strongly predictive of lifelong mental disorders [22,23]. A history of childhood trauma appears equally prevalent among individuals with IA and those with substance dependence [24]. Moreover, a study of Chinese university students found that childhood trauma significantly influences suicidal ideation, an effect mediated by IA [25]. Hence, hypothesis 3 is proposed as follows: IA mediates the relationship between SA and STB. Due to underdeveloped social-cognitive abilities and the inevitability of academic and interpersonal stressors, adolescents often experience NLE. While the causes of IA are still under investigation, NLE are considered a key contributing factor. Research shows a significant positive correlation between NLEs and IA among university students [26-29], a link observed across subtypes such as interpersonal distress, academic pressure, punishment, and health-related problems [30]. Consequently, hypothesis 4 is proposed as follows: NLEs and IA operate as serial mediators in the relationship between SA and STB.

While existing research has separately revealed significant associations between SA and STB, between NLEs and STB, and between IA and STB, how these factors interact within a coherent mechanistic model to collectively explain the pathway from early trauma to ultimate risk remains poorly understood [8-11,14,15,17,21,26]. Building on prior findings, this study aims to use a large representative sample of youth to examine the association between SA and STB, and to explore the potential direct and indirect effects of SA on STB mediated through NLE and IA. This investigation seeks to deepen the understanding of the mechanisms underlying adolescent suicide risk and to provide an evidence-based roadmap for the precise prevention and tiered intervention of suicide risk among Chinese adolescents.

Methods

Data and Sampling

This study used data from the Science Database of the People Mental Health survey conducted between March 2013 and December 2022 by the National Population Health Data Center of the National Research Institute for Family Planning [31]. The data are designed horizontally and vertically, and through dynamic, continuous, and systematic collection of mental health and related risk factor information from children, adolescent, young adult, middle-aged, and older adult populations, as well as from specific populations, occupational populations, and mental health specialties. Thereby, the trajectory of changes in the mental health of individuals, clusters, and populations is presented, and the knowledge mapping of mental health data is developed. The surveys adhere to psychological industry standards, expert consensus, guidelines, and regulations, using widely accepted international and domestically standardized scales adapted for Chinese norms. Data collection involved validity and reliability assessments and was conducted using standardized procedures for both offline and online administration, using surveys and assessments within a mental health context. A multistage stratified random sampling method was used in this study. From March 2013 to December 2022, 20,893 adolescents were recruited from secondary schools across 16 provinces in China. For this analysis, after excluding respondents with missing values on any variable, the final analytical sample comprised 10,664 adolescents. Figure 1 illustrates the data cleaning process. Prior to the survey, informed consent was obtained from both the respondents' parents and the respondents themselves, following local parental consent procedures.

XSL•FO
RenderX

Figure 1. Flowchart of adolescent participants in the cross-sectional study on sexual abuse (SA) and suicidal thoughts and behaviors (STB), China, 2013 - 2022. IA: internet addiction

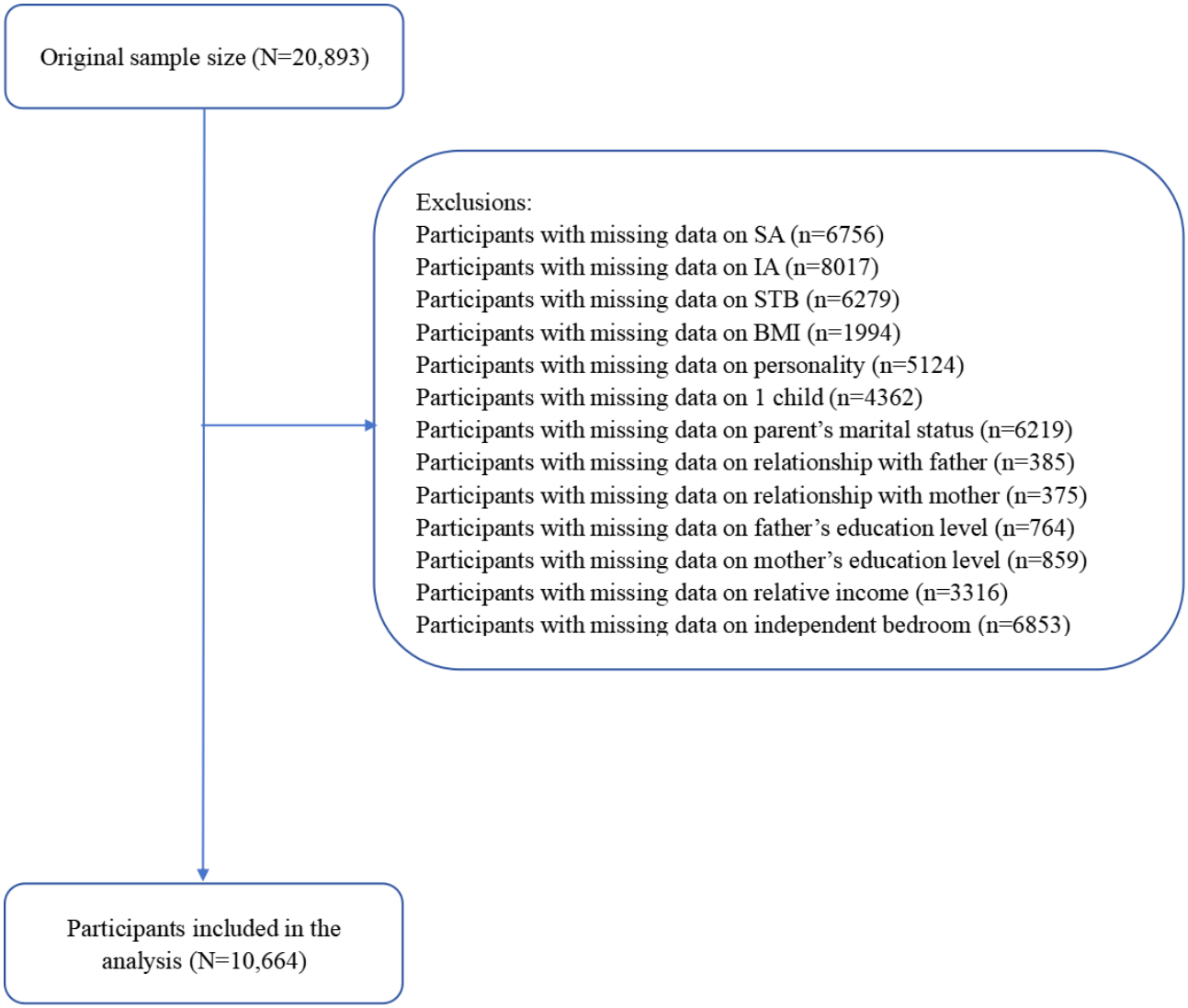

Measures

Sociodemographic Variables

Sociodemographic variables included age, gender, ethnic group, BMI, personality type, 1-child status, parents' marital status, relationship with father, relationship with mother, father's educational level, mother's educational level, region, relative income, and having an independent bedroom. Regions, categorized according to the affiliation of the respondent's province in the *Chinese Health Statistics Yearbook* [32], were divided into central, western, and eastern regions.

Grouping Variable

This study's grouping variable is hukou, which is categorized into urban and rural areas.

Suicidal Thoughts and Behaviors

Suicidal behaviors are measured by using the following 3 items, according to the Global School-Based Student Health Survey, including suicide ideation, plans, and attempts, respectively: "Have you seriously considered suicide in the last 12 months?" "Have you made a specific plan to commit suicide in the last 12 months?" and "Have you attempted suicide in the last 12 months?" All 3 variables were considered dummy variables, with "never," "≤2 days," and ">2 days" coded as 0, 1, and 2, respectively. The total score for STB ranges from 0 to 6, with a higher score indicating more STB. In this study sample, the Cronbach α coefficient for the STB assessment was 0.713, the Kaiser-Meyer-Olkin (KMO) value was 0.667, and the Bartlett test of sphericity yielded a significant result (P<.001).

Sexual Abuse

Respondents were asked the following 9 questions to measure SA: "Has anyone ever kissed you without your consent?" "Has anyone ever exposed their genitals to you?" "Has anyone ever masturbated in front of you?" "Has anyone ever told you dirty jokes, made pornographic gestures, or shown you pornographic pictures?" "Has anyone ever touched your sensitive parts (eg, lower body, breasts, and genitals)?" "Has anyone ever rubbed their genitals on you?" "Has anyone ever made oral contact with your sex organ?" "Has anyone ever tried to have sex with you?" and "Has anyone ever forced sex on you?" All 9 variables were considered dummy variables, with "yes" and "no" coded as 1 and 0, respectively. In this study sample, the Cronbach α coefficient for the SA experience assessment was 0.747, the

XSL•FO
RenderX

KMO value was 0.850, and the Bartlett test of sphericity yielded a significant result (*P*<.001).

Negative Life Events

This study measured NLE using the ASLEC (Adolescent Self-Rating Life Event Checklist; Checklist 2) [33], which is used to assess adverse events within the past 12 months. The ASLEC comprises 27 items in the following 6 dimensions: interpersonal relationships (interpersonal conflict or contradiction), study pressure (stress caused by learning), being punished (punishment by caregivers, teachers, or peers), bereavement (death of beloved ones), health adaptation, and other dimensions (eg, school refusal) [33]. The ASLEC assesses the frequency and intensity of life events experienced by respondents in the past half-year. Respondents rate each NLE's impact on a 5-point Likert scale (1=never to 5=extremely severe). The ASLEC's total score ranges between 27 and 135, with a higher score indicating a greater number of NLE and greater stress experienced over the past half-year. Previous studies have reported that this scale is reliable and valid among Chinese children [27,34]. In this study sample, the Cronbach α coefficient for the Adolescent Life Events Scale was 0.965, the KMO value was 0.971, and the Bartlett test of sphericity yielded a significant result (*P*<.001).

Internet Addiction

The research focused predominantly on IA as measured by the Internet Addiction Test (IAT) developed by Young [35]. The IAT is a 20-item questionnaire that measures characteristics and behaviors associated with compulsive internet use. Each item is weighted using a Likert scale ranging from 1 ("rarely") to 5 ("always"). The IAT scale comprises the following 6 subscales: salience, excessive use, neglect of work, anticipation, lack of control, and neglect of social life; however, precise interval thresholds are not defined for this scale. Therefore, a higher score represents greater addiction. Previous studies have reported that this scale is reliable and valid among Chinese children [36,37]. In this sample, the scale demonstrated excellent internal consistency, with a Cronbach α coefficient of 0.947. KMO measure of sampling adequacy was 0.959, and the Bartlett test of sphericity was significant (*P*<.001).

Figure 2 illustrates the proposed serial mediation model.

Figure 2. Proposed serial mediation model depicting the relationships among sexual abuse, negative life events, internet addiction, and suicidal thoughts and behaviors in Chinese adolescents.

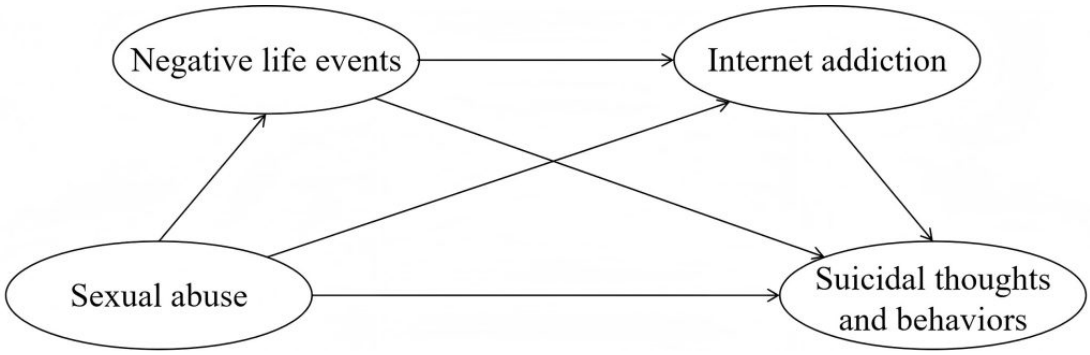

Statistical Analysis

Data analysis was performed using Stata (version 18.0; StataCorp LLC) and SPSS (version 26.0; IBM Corp) software. The study was reported following the Reporting Standards for Studies Using Structural Equation Modeling to ensure compliance with the transparency and completeness requirements stipulated by the JARS (Journal Article Reporting Standards) guidelines (Checklist 1) [38,39]. Descriptive statistics were computed for respondents' sociodemographic characteristics. Frequencies and percentages were reported for categorical variables, whereas means and SDs were reported for continuous variables. Associations between independent and outcome variables were estimated using the Pearson chi-square test (or 2-tailed *t* test for continuous variables). The Pearson correlation analysis measures the strength and direction of a linear relationship between dependent and independent variables. Using a combination of multiple linear regression and bootstrap testing, we constructed a chain mediation model. Multivariate linear regression analysis was used to identify significant independent correlates of STB. To test the hypothesized mediation effects, we conducted path analysis using the bootstrap method, with 5000 bootstrap resamples to generate bias-corrected 95% CI. This approach quantified the indirect effects of the independent variable on the dependent variable through the mediators, as well as the direct effects after controlling for the mediators. The 2-tailed significance level was set at *P*<.05.

Ethical Considerations

This study is a secondary analysis based on the Science Database of the People Mental Health Survey. The original data collection adhered to the principles of the Declaration of Helsinki, and written informed consent was obtained from all participants and their guardians. This research used only processed, publicly available data containing no personally identifiable information and involved no interaction with human participants. No identification of individual participants is presented in any images within this study or supplementary materials. The Science Database of the People Mental Health Survey has specific licensing and citation requirements. The authors obtained permission to use the data and have strictly complied with the relevant data use agreement regarding noncommercial academic research and proper citation. This study was exempted from additional ethics review and approval by the Naval Medical University Ethics Committee.

XSL•FO
RenderX

Results

Sociodemographic Characteristics

Among the participants, 2075 (19.46%), 5960 (55.89%), and 2629 (24.65%) were aged younger than 16 years, 16 to 17.9 years, and 18 years or older, respectively; 4838 (45.37%) and 5826 (54.63%) were male and female adolescents, respectively; 5703 (53.48%) and 4961 (46.52%) were Han Chinese and belonged to an ethnic minority, respectively; 6796 (63.73%), 2511 (23.55%), 998 (9.36%), and 359 (3.37%) had BMIs in the ranges of 18.5 to 23.9 kg/m^2, <18.5 kg/m^2, 24 to 27.9 kg/m^2, and ≥28 kg/m^2, respectively; 1430 (13.41%), 3355 (31.46%), and 5879 (55.13%) were introverts, extroverts, and ambiverts, respectively; 3099 (29.06%) had 1-child status in their families; 421 (3.95%) and 369 (3.46%) had divorced parents and remarried families, respectively; 8833 (82.83%) and 9712 (91.07%) had amicable relationships with their fathers and mothers, respectively; 1012 (9.49%) and 915 (8.58%) respondents' fathers and mothers had an educational level of college and higher, respectively; 4898 (45.93%), 2265 (21.24%), and 3501 (32.83%) were in the central, eastern, and western regions, respectively; 6247 (58.58%), 2539 (23.81%), and 1878 (17.61%) had general, rich, and poor income statuses, respectively; and 8910 (83.55%) had an independent bedroom. The chi-square test and 2-tailed *t* test revealed differences in age, gender, ethnic group, BMI, 1-child status, parents' marital status, father's educational level, mother's educational level, relative income, and independent bedroom of middle school adolescents between urban and rural areas (Table 1). Gender differences exist in the prevalence of SA (57.24%) and STB (24.8%) among adolescent samples in China currently (Table S1 in Multimedia Appendix 1).

XSL•FO
RenderX

Table . Characteristics of Chinese adolescent participants (N=10,664).

Variables	Total (N=10,664)	Rural (n=8028)	Urban (n=2636)	Statistics, chi-square (*df*)	*P* value
Age (y), n (%)				9.3 (2)	.01
<16	2075 (19.46)	1521 (18.95)	554 (21.02)		
16-17.9	5960 (55.89)	4479 (55.79)	1481 (56.18)		
≥18	2629 (24.65)	2028 (25.26)	601 (22.80)		
Gender, n (%)				30.3 (1)	<.001
Woman	5826 (54.63)	4508 (56.15)	1318 (50)		
Man	4838 (45.37)	3520 (43.85)	1318 (50)		
Ethnic group, n (%)				30.3 (1)	<.001
Han Chinese	5703 (53.48)	4171 (51.96)	1532 (58.12)		
Ethnic minority	4961 (46.52)	3857 (48.04)	1104 (41.88)		
BMI (kg/m^2), n (%)				34.0 (3)	<.001
18.5~23.9	6796 (63.73)	5222 (65.05)	1574 (59.71)		
<18.5	2511 (23.55)	1853 (23.08)	658 (24.96)		
24.0~27.9	998 (9.36)	714 (8.89)	284 (10.77)		
≥28.0	359 (3.37)	239 (2.98)	120 (4.55)		
Personality, n (%)				13.3 (2)	<.001
Introvert	1430 (13.41)	1107 (13.79)	323 (12.25)		
Extrovert	3355 (31.46)	2454 (30.57)	901 (34.18)		
Ambivert	5879 (55.13)	4467 (55.64)	1312 (53.57)		
One child, n (%)				1200.0 (1)	<.001
No	7565 (70.94)	6397 (79.68)	1168 (44.31)		
Yes	3099 (29.06)	1631 (20.32)	1468 (55.69)		
Parent's marital status, n (%)				66.7 (3)	<.001
Married	9412 (88.26)	7161 (89.02)	2251 (85.39)		
Divorce	421 (3.95)	275 (3.43)	146 (5.54)		
Remarry	369 (3.46)	227 (2.83)	142 (5.39)		
Other	462 (4.33)	365 (4.55)	97 (3.68)		
Relationship with father, n (%)				7.5 (2)	.02
Well	8833 (82.83)	6671 (83.10)	2162 (82.02)		
General	1449 (13.59)	1092 (13.06)	357 (13.54)		
Other	382 (3.58)	265 (3.30)	117 (4.44)		
Relationship with mother, n (%)				12.2 (2)	.34
Well	9712 (91.07)	7302 (90.96)	2410 (91.43)		
General	768 (7.20)	579 (7.21)	189 (7.17)		
Other	184 (1.73)	147 (1.83)	37 (1.40)		
Father's education level, n (%)				1900.0 (3)	<.001
Primary school and below	2341 (21.95)	2142 (26.68)	199 (7.55)		
Junior high school	4993 (46.82)	4154 (51.74)	839 (31.83)		
High school	2318 (21.74)	1449 (18.05)	869 (32.97)		
College and higher	1012 (9.49)	283 (3.53)	729 (27.66)		

XSL•FO
RenderX

Variables	Total (N=10,664)	Rural (n=8028)	Urban (n=2636)	Statistics, chi-square (*df*)	*P* value
Mother's education level, n (%)				2200.0 (3)	<.001
Primary school and below	3392 (31.8)	3103 (38.65)	289 (10.96)		
Junior high school	4558 (42.74%)	3709 (46.20)	849 (32.21)		
High school	1799 (16.87)	994 (12.38)	805 (30.54)		
College and higher	915 (8.58)	222 (2.77)	693 (26.29)		
Region				111.8 (2)	<.001
Central region	4898(45.93%)	3920(48.83%)	978(37.10%)		
Eastern region	2265(21.24%)	1636(20.38%)	629(23.86%)		
Western region	3501(32.83%)	2472(30.79%)	1029(39.04%)		
Relative income, n (%)				550.0(2)	<.001
General	6247 (58.58)	4843 (60.33)	1404 (53.26)		
Rich	2539 (23.81)	1511 (18.82)	1028 (39)		
Poor	1878 (17.61)	1674 (20.8)	204 (7.74)		
Independent bedroom, n (%)				35.6 (1)	<.001
No	1754 (16.45)	1419 (17.68)	335 (12.71)		
Yes	8910 (83.55)	6609 (82.32)	2301 (87.29)		
Sexual abuse, mean (SD)	—[a]	1.289 (1.672)	1.436 (1.832)	−3.289[b]	.001
Negative life events, mean (SD)	—	52.734 (23.953)	49.596 (23.266)	5.876[c]	<.001
Internet addiction, mean (SD)	—	34.628 (13.777)	35.650 (14.056)	−3.290[c]	.001
Suicidal thoughts and behaviors, mean (SD)	—	0.648 (1.370)	0.805 (1.469)	−5.010[c]	<.001

[a]Not applicable.

[b]Rank-sum test.

[c]2-tailed *t* test.

Correlation Analyses

Table 2 presents the means, SD, and Pearson correlations of all study variables. Notably, SA was significantly positively correlated with NLE (r=0.174; P<.001), IA (r=0.199; P<.001), and STB (r=0.137; P<.001). Moreover, NLE were significantly positively correlated with IA (r=0.281; P<.001) and STB (r=0.173; P<.001). Finally, IA was significantly positively correlated with STB (r=0.197; P<.001). Significant correlations between the variables initially supported our hypotheses.

Table . Means, SDs, and correlation of sex abuse, negative life events, internet addiction, and suicidal thoughts and behaviors (STB; N=10,664).

Variable	Value, mean (SD)	1[a]	2[b]	3[c]	4[d]
Sexual abuse	1.330 (1.714)	1.000			
Negative life events	51.960 (23.822)	0.174[e]	1.000		
Internet addiction	34.88 (13.852)	0.199[e]	0.281[e]	1.000	
STB	0.690 (1.396)	0.137[e]	0.173[e]	0.197[e]	1.000

[a]Correlations between sexual abuse and itself, negative life events, internet addiction, and STB.

[b]Correlations between negative life events and itself, internet addiction, and STB.

[c]Correlations between internet addiction and itself and STB.

[d]Correlation of STB with itself.

[e]P<.001.

XSL•FO
RenderX

Chain Mediation Analysis

To control for the influence of potential confounding factors, covariates were included in all mediation effect test models. We constructed a chain mediation effect model based on different theoretical assumptions (Figure 3). Multicollinearity tests indicated that the variance inflation factor values for all predictors were well below the critical threshold of 5, suggesting no serious multicollinearity issues in the model (Table S2 in Multimedia Appendix 1).

Figure 3. The indirect effects of negative life events and internet addiction in the relationship between sexual abuse and suicidal thoughts and behaviors. a: *P*<.001.

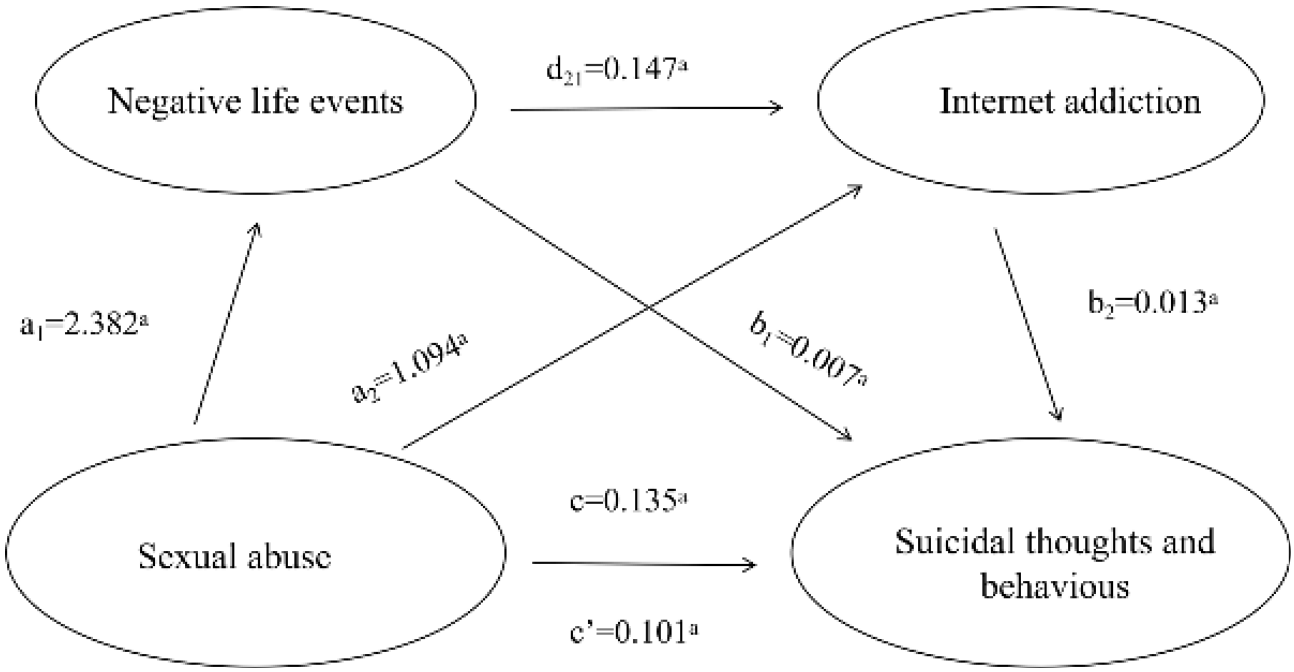

Figure 3 depicts the tested structural model, which statistically demonstrated significant associations among key variables. The model indicated that a higher level of SA was associated with higher levels of NLE (β=2.382, 95% CI 2.112 - 2.653; *P*<.001), IA (β=1.094, 95% CI 0.939 - 1.249; *P*<.001), and STB (β=0.135, 95% CI 0.119 - 0.151; *P*<.001). Higher levels of NLE were further associated with higher levels of IA (β=0.147, 95% CI 0.136 - 0.158; *P*<.001) and STB (β=0.007, 95% CI 0.005 - 0.008; *P*<.001). IA also showed a significant association with higher levels of STB (β=0.013, 95% CI 0.011 - 0.015; *P*<.001). Furthermore, as presented in Table 3, the direct effect of SA on STB was significant (β=0.101, 95% CI 0.085 - 0.117; *P*<.001), accounting for 74.81 % of the total effect. The indirect effect of SA on STB through NLE was significant (β=0.016, 95% CI 0.012 - 0.020; *P*<.001), accounting for 11.85% of the total effect; the indirect effect of SA on STB through IA was significant (β=0.014, 95% CI 0.010 - 0.018; *P*<.001), accounting for 10.37% of the total effect. Moreover, the indirect effect of SA on STB through NLE and IA was significant (β=0.004, 95% CI 0.003 - 0.006; *P*<.001), accounting for 2.96% of the total effect. All regression coefficients are unstandardized coefficients.

XSL•FO
RenderX

Table . Results of the chain mediation model analyzing the pathway from sexual abuse to suicidal thoughts and behaviors via negative life events and internet addiction.

Model pathways	β[a] (95% CI)	*P* value	Ratio (%)
Total effect	0.135 (0.119-0.151)	<.001	100.00
Direct effect	0.101 (0.085-0.117)	<.001	74.81
Total indirect effect	0.034 (0.029-0.040)	<.001	25.19
Sexual abuse →negative life events→ suicidal thoughts and behaviors	0.016 (0.012-0.020)	<.001	11.85
Sexual abuse → Internet addiction→ suicidal thoughts and behaviors	0.014 (0.010-0.018)	<.001	10.37
Sexual abuse →negative life events→ internet addiction →suicidal thoughts and behaviors	0.004 (0.003-0.006)	<.001	2.96

[a]Unstandardized regression coefficient.

Missing Data

Among the 18 variables analyzed in this study, the variables NLE, Age, gender, ethnic group, and region have no missing values. The proportion of missing data is presented in Table 4. The Little test for completely random missingness showed a significant result (χ^2_8=14670.975; *P*<.001).

Table . The proportion of missing data.

Variables	Missing value, n (%)
SA[a]	6756 (32.3)
NLE[b]	0 (0)
IA[c]	8017 (38.4)
STB[d]	6279 (30.1)
Age	0 (0)
Gender	0 (0)
Ethnic group	0 (0)
BMI	1994 (9.5)
Personality	5124 (24.5)
One child	4362 (20.9)
Parent's marital status	6219 (29.8)
Relationship with father	385 (1.8)
Relationship with mother	375 (1.8)
Father's education level	764 (3.7)
Mother's education level	859 (4.1)
Region	0 (0)
Relative income	3316 (15.9)
Independent bedroom	6853 (32.8)

[a]SA: sexual abuse.

[b]NLE: negative life events.

[c]IA: internet addiction.

[d]STB: suicidal thoughts and behaviors.

Following multiple imputation of missing values, the sensitivity analysis yielded results consistent with the primary analysis, demonstrating robustness. The outcomes after imputation are presented in Table S3 in Multimedia Appendix 1.

XSL•FO
RenderX

Discussion

Principal Findings

This study revealed that STB were positively associated with SA among adolescents. Notably, NLE and IA independently mediated the path from SA to STB. Furthermore, NLE and IA played a serial mediating role in the relationship between SA and STB. The study constructs a new theoretical model with adolescents as the research object and provides guidance and recommendations to develop interventions for adolescent suicide.

SA and STB

This study's findings support hypothesis 1 that SA would be positively associated with STB among adolescents, suggesting that adolescents with a history of SA are at a higher risk of developing STB. Our results align with previous cross-sectional and longitudinal studies that noted a consistent association between childhood SA and an increased risk of suicide attempts [11,40,41]. Several biological and psychological theories elucidate how SA increases the risk of STB. According to neuroimaging investigations, childhood abuse is associated with reduced medial prefrontal cortex volume [42], which reduces an individual's cognitive functioning; the resultant cognitive deficits can directly increase suicide risk [43,44]. Regarding psychological theories, Finkelhor and Browne [45] identified 2 key trauma dynamics resulting from early SA, namely, stigmatization and feelings of betrayal—commonly experienced by survivors of early SA, especially when the abuser is a familiar member or caregiver [46]. These trauma dynamics catalyze depression, shame, and alienation from others, further eroding one's sense of self-worth and belonging [47]. These negative effects grow, fostering negative emotions and, after a point, self-harming behaviors and even STB [48].

Notably, gender differences exist in the prevalence of SA and STB among adolescent samples in China currently. Specifically, the probability of STB is higher among girls, while the proportion of experiencing SA is higher among boys [41,49,50]. This pattern likely does not reflect absolute differences in incidence rates but rather profoundly reveals gendered expressions of risk behaviors and sociocultural cognitive biases [51,52]. More concretely, the higher reported rate of suicidal ideation among women may be partly attributable to social norms that more readily permit them to acknowledge and express internalized emotional distress [53]. Conversely, the lower reported rate of SA among male individuals is more likely rooted in the widespread societal neglect of male survivors, the heightened stigma associated with disclosure, and the potential limitations of existing measurement tools in capturing trauma experiences specific to male survivors [50]. It is recommended to focus on the underidentified SA trauma among male adolescents and its potential externalizing behavioral consequences, as well as the clearer pathway from internalizing NLE to suicidal risk among female adolescents.

Mediating Effect of NLE

This study's findings verified that NLE mediated the relationship between SA and STB among adolescents, thus supporting hypothesis 2. To our knowledge, previous literature on the mediating role of NLE in the association between SA and STB is scant. Our study expands this field of research and fills a gap in this area. Notably, SA inflicts profound psychological trauma on survivors, which profoundly disrupts their daily lives and interpersonal relationships [54,55]. Furthermore, survivors may grapple with confusion and negative perceptions of their identities and sexual orientations, accompanied by feelings of shame and guilt [56,57]. Moreover, this trauma may impair survivors' academic, occupational, and daily functioning, as well as precipitate the onset of long-term health issues, including sexually transmitted infections, unintended pregnancies, chronic pain, and physical disabilities [54,58-62].

This study found that NLE were significantly positively correlated with STB, which is consistent with previous research [63-67]. Chronic stress from such events results in depression or despair [68], negative coping strategies, and, ultimately, suicidal thoughts [69]. Rumination may explain how NLE mediate the relationship between SA and STB, suggesting that individuals abused as children, without learning to manage their emotions, may have a reduced sense of control in adulthood [70]. Facing NLE, they tend to repeatedly consider negative outcomes, thus triggering STB.

Mediating Effect of IA

In this study, IA was identified as a mediator in the association between SA and STB among adolescents, thus corroborating hypothesis 3. Consistent with prior research, survivors of childhood abuse were found to be more prone to IA, with a strong link observed between SA and childhood trauma [71-73]. Per the social compensation hypothesis, the internet offers a suitable platform for self-expression for those who have faced prolonged NLE [74]. Furthermore, SA creates a dangerous environment for the survivor, thus resulting in numerous negative consequences [75]. Hence, they may hesitate to disclose their true selves in face-to-face interactions. Given the anonymous and concealed nature of the internet, these individuals are highly likely to turn to it as a platform for self-expression and release, which may, in turn, elevate their susceptibility to IA [76]. Additionally, IA may result in a variety of drastic physical, psychological, and social relation changes that adolescents can neither control nor avoid [77]. This cycle can trigger intense stress, potentially culminating in a mental health crisis, self-harm, and suicidal behaviors among adolescents. Moreover, the very online environments they frequent—through activities like gaming, socializing, or watching videos—are linked to increased suicide risk [78]. Within these spaces, discussions that normalize suicide or even share methods are not uncommon, potentially providing susceptible youths with both the impetus and the means for self-harm [79].

Serial Mediating Effect of NLE and IA

This study supported hypothesis 4, as the association between SA and STB was manifested through a serial mediation pathway involving NLE and IA. Adolescents who have experienced SA are more inclined to adopt negative coping styles when dealing with NLE, such as by frequently using the internet to escape reality, which increases their likelihood of receiving undesirable

XSL•FO
RenderX

information, experiencing negative moods, and exhibiting STB [27].

A significant positive correlation was observed between NLE and IA—consistent with previous research [80]. First, when faced with NLE, adolescents tend to adopt negative coping styles to alleviate negative emotions; adolescents commonly use the internet for emotional venting [80]. Owing to the lack of social support from family, friends, or other people, adolescents choose to use the internet to avoid the unpleasantness and pain of reality and forget their troubles temporarily [81]. Hence, IA may be a compensatory approach to deal with NLE [82]. Adolescents can seek help on the internet and gain a sense of accomplishment through gaming and entertainment. Adolescents who have experienced NLE are more likely to feel lonely, unaccepted, or have low self-confidence and may, thus, seek a sense of belonging and recognition online [67,83]. Such virtual social relationships may increase their feelings of dependence on the internet [84].

Suggestions for Future Research

Based on our findings, potential targets for intervention and multipathway mechanisms have been identified. In practice, we propose the following recommendations. First, in clinical assessments, adolescents with a history of SA should undergo routine screening for negative cognitive biases and problematic internet use patterns to facilitate the early identification of high-risk individuals [85]. Second, efforts to enhance adolescent psychological resilience should focus on helping youth modify their cognitive appraisal of NLE and preventing these appraisals from evolving into maladaptive behavioral patterns [86]. Third, modular content addressing IA should be incorporated into adolescent mental health initiatives. This content should aim to help adolescents understand the emotional needs underlying their internet use and collaboratively establish healthy alternative activities and time management plans [87].

Limitations

This study has several limitations. First, given the cross-sectional design, the mediation model we examined reveals patterns of statistical association among the variables but cannot establish strict causal direction or temporal sequence. Future research should adopt longitudinal or cross-lagged designs to more clearly elucidate the causal timing and dynamic interactions among these variables. Second, there may be reverse causality between the variables or spurious associations due to unmeasured variables. Third, the SA assessment tool used in this study is not a widely validated standardized instrument, and simply summing dichotomous items may not fully capture the complexity of SA experiences. Furthermore, this total score should not be interpreted as a strict interval scale.

Conclusions

SA is significantly correlated with STB among adolescents, and it indirectly exacerbates suicide risk through the serial mediation of NLE and IA.

This study possesses both theoretical and practical value in the prevention of suicide among adolescents. First, unlike previous research that has largely focused on isolated risk factors or single mediators, this study innovatively constructs and validates a chain mediation model, which reveals the chained mediating roles of NLE and IA between SA and STB. Second, this study systematically elucidates the sequential pathway mechanism from early trauma to psychosocial stress, then to behavioral problems, and ultimately to suicide risk, offering a novel perspective on the complex developmental process from early adversity to extreme outcomes. Finally, the findings provide clear direction and scientific evidence for the development of comprehensive suicide prevention strategies. Future efforts should focus on the prevention of SA and trauma intervention while simultaneously strengthening adolescents' cognitive reappraisal of negative events and guiding their online behavior.

Acknowledgments

The authors thank all study participants. The authors declare that no generative artificial intelligence tools were used in the writing of any portion of this manuscript.

Funding

This work was supported by the Talent Plan of Naval Medical University (grant 2025) and National Ministry of Education Collaborative Education Reform Project (grant 2411052361123306). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the study; and decision to submit the study for publication.

Data Availability

The dataset analyzed during this study is publicly available in the Science Database of the People Mental Health survey conducted by the National Population Health Data Archive. The dataset is available online [31].

Authors' Contributions

Conceptualization: QJ, LY
Data curation: JZ, QJ, FZ
Formal analysis: JZ, FZ, SY, LY
Methodology: HL, JS, ZZ, LL

XSL•FO
RenderX

Supervision: ZZ, LL
Validation: JZ, QJ, FZ, HL, SY, JS, LY
Visualization: JZ, QJ, HL, JS
Writing – original draft: JZ, QJ, FZ, LY
Writing – review & editing: ZZ, LL, LY
JZ, QJ, and FZ contributed equally as co-first authors.
ZZ, LL, and LY contributed equally as co-corresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of gender differences in participants, multicollinearity test results, and chained mediation imputation results.
[DOCX File, 19 KB - jmir_v28i1e85371_app1.docx]

Checklist 1
JARS checklist.
[DOCX File, 20 KB - jmir_v28i1e85371_app2.docx]

Checklist 2
ASLEC checklist.
[DOCX File, 14 KB - jmir_v28i1e85371_app3.docx]

References

1. Lipson SK, Zhou S, Abelson S, et al. Trends in college student mental health and help-seeking by race/ethnicity: findings from the National Healthy Minds Study, 2013-2021. J Affect Disord 2022 Jun 1;306:138-147. [doi: 10.1016/j.jad.2022.03.038] [Medline: 35307411]
2. Mathieu S, de Leo D, Koo YW, Leske S, Goodfellow B, Kõlves K. Suicide and suicide attempts in the Pacific Islands: a systematic literature review. Lancet Reg Health West Pac 2021;17:100283. [doi: 10.1016/j.lanwpc.2021.100283] [Medline: 34734201]
3. Pan CH, Lee MB, Wu CY, Liao SC, Chan CT, Chen CY. Suicidal ideation, psychopathology, and help-seeking in 15 to 19-year-old adolescents in Taiwan: a population-based study 2015-2019. J Affect Disord 2021 Mar 1;282:846-851. [doi: 10.1016/j.jad.2020.12.139] [Medline: 33601727]
4. Suicide worldwide in 2019: global health estimates. : World Health Organization; 2021 URL: https://iris.who.int/server/api/core/bitstreams/3bd4ac79-4347-420e-b675-948d36ab3d90/content [accessed 2026-02-24]
5. Glenn CR, Kleiman EM, Kellerman J, et al. Annual research review: a meta-analytic review of worldwide suicide rates in adolescents. J Child Psychol Psychiatry 2020 Mar;61(3):294-308. [doi: 10.1111/jcpp.13106] [Medline: 31373003]
6. Altavini CS, Asciutti APR, Santana GL, et al. Suicide ideation among Brazilian college students: relationship with academic factors, mental health, and sexual abuse. J Affect Disord 2023 May 15;329:324-334. [doi: 10.1016/j.jad.2023.02.112] [Medline: 36849006]
7. Mościcki EK. Epidemiology of completed and attempted suicide: toward a framework for prevention. Clin Neurosci Res 2001 Nov;1(5):310-323. [doi: 10.1016/S1566-2772(01)00032-9]
8. Zhou Y, Liang Y, Cheng J, Zheng H, Liu Z. Child maltreatment in Western China: demographic differences and associations with mental health. Int J Environ Res Public Health 2019 Sep 26;16(19):3619. [doi: 10.3390/ijerph16193619] [Medline: 31561624]
9. Schröder J, Nick S, Richter-Appelt H, Briken P. Psychiatric impact of organized and ritual child sexual abuse: cross-sectional findings from individuals who report being victimized. Int J Environ Res Public Health 2018 Oct 31;15(11):2417. [doi: 10.3390/ijerph15112417] [Medline: 30384461]
10. Berens AE, Jensen SKG, Nelson CA. Biological embedding of childhood adversity: from physiological mechanisms to clinical implications. BMC Med 2017 Jul 20;15(1):135. [doi: 10.1186/s12916-017-0895-4] [Medline: 28724431]
11. Ng QX, Yong BZJ, Ho CYX, Lim DY, Yeo WS. Early life sexual abuse is associated with increased suicide attempts: an update meta-analysis. J Psychiatr Res 2018 Apr;99:129-141. [doi: 10.1016/j.jpsychires.2018.02.001] [Medline: 29454220]
12. O'Connor RC, Kirtley OJ. The integrated motivational-volitional model of suicidal behaviour. Philos Trans R Soc Lond B Biol Sci 2018 Sep 5;373(1754):20170268. [doi: 10.1098/rstb.2017.0268] [Medline: 30012735]
13. Laglaoui Bakhiyi C, Jaussent I, Beziat S, et al. Positive and negative life events and reasons for living modulate suicidal ideation in a sample of patients with history of suicide attempts. J Psychiatr Res 2017 May;88:64-71. [doi: 10.1016/j.jpsychires.2016.12.022] [Medline: 28088052]

XSL•FO
RenderX

14. Wang M, Kou C, Bai W, et al. Prevalence and correlates of suicidal ideation among college students: a mental health survey in Jilin Province, China. J Affect Disord 2019 Mar 1;246:166-173. [doi: 10.1016/j.jad.2018.12.055] [Medline: 30583141]
15. Kang C, Zheng Y, Yang L, et al. Prevalence, risk factors and clinical correlates of suicidal ideation in adolescent patients with depression in a large sample of Chinese. J Affect Disord 2021 Jul 1;290:272-278. [doi: 10.1016/j.jad.2021.04.073] [Medline: 34015621]
16. Zheng H, Liu D, Cheng J, Wang DB, Liu Y, Wu Y. Negative life events increase the risk of suicidal ideation in 6653 Chinese freshmen: from a 1-year longitudinal study. J Affect Disord 2022 Feb 15;299:604-609. [doi: 10.1016/j.jad.2021.12.039] [Medline: 34942231]
17. Xiong A, Liao S, Luo B, Luo S, Tong Y, Li Z. Associations between problematic internet use, life satisfaction, and deliberate self-harm among Chinese adolescents: a multi-centered longitudinal study. Addict Behav 2023 Dec;147:107808. [doi: 10.1016/j.addbeh.2023.107808] [Medline: 37536220]
18. Carr CP, Martins CMS, Stingel AM, Lemgruber VB, Juruena MF. The role of early life stress in adult psychiatric disorders: a systematic review according to childhood trauma subtypes. J Nerv Ment Dis 2013 Dec;201(12):1007-1020. [doi: 10.1097/NMD.0000000000000049] [Medline: 24284634]
19. Hébert M, Daigneault I, Vancamp T. Chapitre 5 agression sexuelle et risque de revictimisation à l'Adolescence: modèles conceptuels et défis liés à la prévention. In: L'agression Sexuelle Envers Les Enfants [Book in French]: Presses de l'Université du Québec; 2012:171-224. [doi: 10.2307/j.ctv18ph433.8]
20. Pan PY, Yeh CB. Internet addiction among adolescents may predict self-harm/suicidal behavior: a prospective study. J Pediatr 2018 Jun;197:262-267. [doi: 10.1016/j.jpeds.2018.01.046] [Medline: 29550226]
21. Kim H, Jhon M, Kim JW, et al. Suicide and non-suicidal self-injury from internet addiction among Korean adolescents. J Korean Acad Child Adolesc Psychiatry 2023 Jul 1;34(3):152-158. [doi: 10.5765/jkacap.230015] [Medline: 37426833]
22. Turner S, Harvey C, Hayes L, et al. Childhood adversity and clinical and psychosocial outcomes in psychosis. Epidemiol Psychiatr Sci 2019 Dec 16;29:e78. [doi: 10.1017/S2045796019000684] [Medline: 31839014]
23. Kessler RC, McLaughlin KA, Green JG, et al. Childhood adversities and adult psychopathology in the WHO World Mental Health Surveys. Br J Psychiatry 2010 Nov;197(5):378-385. [doi: 10.1192/bjp.bp.110.080499] [Medline: 21037215]
24. Schwaninger PV, Mueller SE, Dittmann R, et al. Patients with non-substance-related disorders report a similar profile of childhood trauma experiences compared to heroin-dependent patients. Am J Addict 2017 Apr;26(3):215-220. [doi: 10.1111/ajad.12512] [Medline: 28240800]
25. Lu L, Jian S, Dong M, et al. Childhood trauma and suicidal ideation among Chinese university students: the mediating effect of Internet addiction and school bullying victimisation. Epidemiol Psychiatr Sci 2020 Aug 10;29:e152. [doi: 10.1017/S2045796020000682] [Medline: 32772993]
26. Huang Q, Li J, Wang J, Liu B. Negative life events and internet addiction among college students: role of physical exercise and prosocial behavior. Front Psychol 2025;16:1629818. [doi: 10.3389/fpsyg.2025.1629818] [Medline: 41040106]
27. Yang F, Jiang L, Miao J, et al. The association between non-suicidal self-injury and negative life events in children and adolescents in underdeveloped regions of south-western China. PeerJ 2022;10:e12665. [doi: 10.7717/peerj.12665] [Medline: 35287346]
28. Zeng Z, Hu Y, Peng L, et al. The effect of negative life events on college students' depression: the mediating role of internet addiction and the moderating role of 5-HTT1A gene rs6449693 polymorphism. Curr Psychol 2023 Mar;42(8):6555-6563. [doi: 10.1007/s12144-021-01888-0]
29. Xiao L. Relationship among inclination of internet addiction, life events and self-efficacy in college students [Master's Thesis] [Article in Chinese]. : Hangzhou Normal University; 2012 URL: https://oversea.cnki.net/kcms2/article/[illegible] [accessed 2026-02-24]
30. Li Y, Sun J, Hu QY, Li DS, Lu HQ. Effects of the life events on internet addiction disorder in undergraduate students [Article in Chinese]. Mod Prev Med 2012;39(23):6207-6209 [FREE Full text]
31. Science database of people mental health [Article in Chinese]. Population Health Data Archive. URL: https://www.ncmi.cn/ /phda/dataDetails.do?id=CSTR:17970.11.A001U.202204.281.V1.0 [accessed 2026-02-24]
32. 2023 China health and wellness statistical yearbook. : National Health Commission of the People's Republic of China; 2025 URL: https://www.nhc.gov.cn/mohwsbwstjxxzx/tjtjnj/202501/8193a8edda0f49df80eb5a8ef5e2547c/files/ 1740022743894_10341.pdf [accessed 2026-02-28]
33. Liu XC, Oda S, Peng X, Asai K. Life events and anxiety in Chinese medical students. Soc Psychiatry Psychiatr Epidemiol 1997 Feb;32(2):63-67. [doi: 10.1007/BF00788922] [Medline: 9050346]
34. Yang T, He Y, Wu S, Cui X, Luo X, Liu J. Association between schizoid tendencies and aggressive behaviors: mediating and moderating influences in childhood trauma and life events among Chinese adolescents. Ann Gen Psychiatry 2021 Dec 3;20(1):51. [doi: 10.1186/s12991-021-00371-1] [Medline: 34861879]
35. Young KS. Internet addiction: the emergence of a new clinical disorder. Cyberpsychol Behav 1998 Jan;1(3):237-244. [doi: 10.1089/cpb.1998.1.237]
36. Lai CM, Mak KK, Watanabe H, Ang RP, Pang JS, Ho RCM. Psychometric properties of the internet addiction test in Chinese adolescents. J Pediatr Psychol 2013 Aug;38(7):794-807. [doi: 10.1093/jpepsy/jst022] [Medline: 23671059]

XSL•FO
RenderX

37. Tahir MJ, Malik NI, Ullah I, et al. Internet addiction and sleep quality among medical students during the COVID-19 pandemic: a multinational cross-sectional survey. PLoS ONE 2021;16(11):e0259594. [doi: 10.1371/journal.pone.0259594] [Medline: 34739502]
38. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
39. Journal Article Reporting Standards (JARS). American Psychological Association. URL: https://apastyle.apa.org/jars [accessed 2026-02-09]
40. Ferahkaya H, Türker M, Coşkun F, Uzun N, Akça Ö. Predictors of suicide attempt among subjects who were previously exposed to sexual abuse. Int J Soc Psychiatry 2025 Sep;71(6):1155-1163. [doi: 10.1177/00207640251334268] [Medline: 40249755]
41. Alhassan MA, Alarabi MA, Albalawi WM, et al. Prevalence and risk factors of suicidal ideation and suicide attempts among young people in the MENAT region: a systematic review and meta-analysis. BMC Psychiatry 2025 Nov 14;25(1):1149. [doi: 10.1186/s12888-025-07585-7] [Medline: 41257638]
42. van Harmelen AL, van Tol MJ, van der Wee NJA, et al. Reduced medial prefrontal cortex volume in adults reporting childhood emotional maltreatment. Biol Psychiatry 2010 Nov 1;68(9):832-838. [doi: 10.1016/j.biopsych.2010.06.011] [Medline: 20692648]
43. Szanto K. Cognitive deficits: underappreciated contributors to suicide. Am J Geriatr Psychiatry 2017 Jun;25(6):630-632. [doi: 10.1016/j.jagp.2017.02.012] [Medline: 28302439]
44. Olsson P, Wiktorsson S, Sacuiu S, et al. Cognitive function in older suicide attempters and a population-based comparison group. J Geriatr Psychiatry Neurol 2016 May;29(3):133-141. [doi: 10.1177/0891988715627015] [Medline: 26810564]
45. Finkelhor D, Browne A. The traumatic impact of child sexual abuse: a conceptualization. Am J Orthopsychiatry 1985 Oct;55(4):530-541. [doi: 10.1111/j.1939-0025.1985.tb02703.x] [Medline: 4073225]
46. Cahill C, Llewelyn SP, Pearson C. Long-term effects of sexual abuse which occurred in childhood: a review. Br J Clin Psychol 1991 May;30(2):117-130. [doi: 10.1111/j.2044-8260.1991.tb00927.x] [Medline: 2059744]
47. Hassanpour P, Buchwald S, Mehta AHP, Goldberg SB, Walsh K. Sexual violence and shame: a meta-analysis. Trauma Violence Abuse 2026 Jan;27(1):240-255. [doi: 10.1177/15248380241308828] [Medline: 39819226]
48. Sadath A, Kavalidou K, McMahon E, Malone K, McLoughlin A. Associations between humiliation, shame, self-harm and suicidality among adolescents and young adults: a systematic review. PLoS One 2024;19(2):e0292691. [doi: 10.1371/journal.pone.0292691] [Medline: 38329967]
49. Lai W, Wu H, Yang L, et al. Trends in prevalence of suicidal ideation, suicide attempt and sleep disturbance among Chinese high school students, 2014-2023. Eur Child Adolesc Psychiatry 2025 Nov;34(11):3475-3484. [doi: 10.1007/s00787-025-02762-1] [Medline: 40450187]
50. Abreu L, Hecker T, Goessmann K, Abioye TO, Olorunlambe W, Hoeffler A. Prevalence and correlates of sexual violence against adolescents: quantitative evidence from rural and urban communities in South-West Nigeria. PLOS Glob Public Health 2025;5(2):e0004223. [doi: 10.1371/journal.pgph.0004223] [Medline: 39933016]
51. Chaplin TM, Aldao A. Gender differences in emotion expression in children: a meta-analytic review. Psychol Bull 2013 Jul;139(4):735-765. [doi: 10.1037/a0030737] [Medline: 23231534]
52. Cantor N, Kingsbury M, Warner E, et al. Young adult outcomes associated with adolescent suicidality: a meta-analysis. Pediatrics 2023 Mar 1;151(3):e2022058113. [doi: 10.1542/peds.2022-058113] [Medline: 36810672]
53. Chaplin TM. Gender and emotion expression: a developmental contextual perspective. Emot Rev 2015 Jan;7(1):14-21. [doi: 10.1177/1754073914544408] [Medline: 26089983]
54. Guiney H, Caspi A, Ambler A, et al. Childhood sexual abuse and pervasive problems across multiple life domains: findings from a five-decade study. Dev Psychopathol 2024 Feb;36(1):219-235. [doi: 10.1017/S0954579422001146] [Medline: 36562290]
55. Serrano-Rodríguez E, Luque-Ribelles V, Hervías-Parejo V. Psychosocial consequences of sexual assault on women: a scoping review. Arch Sex Behav 2025 Jan;54(1):231-258. [doi: 10.1007/s10508-024-03013-1] [Medline: 39477846]
56. McElvaney R, Lateef R, Collin-Vézina D, Alaggia R, Simpson M. Bringing shame out of the shadows: identifying shame in child sexual abuse disclosure processes and implications for psychotherapy. J Interpers Violence 2022 Oct;37(19-20):NP18738-NP18760. [doi: 10.1177/08862605211037435] [Medline: 34459692]
57. Guyon R, Fernet M, Couture S, Tardif M, Cousineau MM, Godbout N. “Finding my worth as a sexual being”: a qualitative gender analysis of sexual self-concept and coping in survivors of childhood sexual abuse. Arch Sex Behav 2024 Jan;53(1):341-357. [doi: 10.1007/s10508-023-02693-5] [Medline: 37747584]
58. Ochoa MK, Constantin K. Impacts of child sexual abuse: the mediating role of future orientation on academic outcomes. Child Abuse Negl 2023 Nov;145:106437. [doi: 10.1016/j.chiabu.2023.106437] [Medline: 37659153]
59. Jonsdottir SD, Thordardottir EB, Valdimarsdottir UA, et al. Sexual violence in the workplace and associated health outcomes: a nationwide, cross-sectional analysis of women in Iceland. Lancet Public Health 2024 Jun;9(6):e365-e375. [doi: 10.1016/S2468-2667(24)00075-6] [Medline: 38821683]

XSL•FO
RenderX

60. Skjælaaen K, Nesvold H, Brekke M, et al. Sexually transmitted infections among patients attending a sexual assault centre: a cohort study from Oslo, Norway. BMJ Open 2022 Dec 1;12(12):e064934. [doi: 10.1136/bmjopen-2022-064934] [Medline: 36456026]
61. Ajayi AI, Ezegbe HC. Association between sexual violence and unintended pregnancy among adolescent girls and young women in South Africa. BMC Public Health 2020 Sep 7;20(1):1370. [doi: 10.1186/s12889-020-09488-6] [Medline: 32894130]
62. Bussières A, Hancock MJ, Elklit A, et al. Adverse childhood experience is associated with an increased risk of reporting chronic pain in adulthood: a stystematic review and meta-analysis. Eur J Psychotraumatol 2023;14(2):2284025. [doi: 10.1080/20008066.2023.2284025] [Medline: 38111090]
63. Chan S, Miranda R, Surrence K. Subtypes of rumination in the relationship between negative life events and suicidal ideation. Arch Suicide Res 2009;13(2):123-135. [doi: 10.1080/13811110902835015] [Medline: 19363749]
64. Chang EC, Sanna LJ, Hirsch JK, Jeglic EL. Loneliness and negative life events as predictors of hopelessness and suicidal behaviors in Hispanics: evidence for a diathesis-stress model. J Clin Psychol 2010 Dec;66(12):1242-1253. [doi: 10.1002/jclp.20721] [Medline: 20734320]
65. Lannoy S, Mars B, Heron J, Edwards AC. Suicidal ideation during adolescence: the roles of aggregate genetic liability for suicide attempts and negative life events in the past year. J Child Psychol Psychiatry 2022 Oct;63(10):1164-1173. [doi: 10.1111/jcpp.13653] [Medline: 35766310]
66. Glenn CR, Kleiman EM, Kandlur R, Esposito EC, Liu RT. Thwarted belongingness mediates interpersonal stress and suicidal thoughts: an intensive longitudinal study with high-risk adolescents. J Clin Child Adolesc Psychol 2022;51(3):295-311. [doi: 10.1080/15374416.2021.1969654] [Medline: 34570668]
67. Li Y, Shen X, Wang JL. Association between negative life events and suicidal behavior in adolescents: roles of core self-evaluation, depression and gender. Soc Psychiatry Psychiatr Epidemiol 2025 Mar;60(3):673-684. [doi: 10.1007/s00127-024-02725-y] [Medline: 39001887]
68. Zhao SY, Ren RR, Chen W. A mediated moderation model of negative life events, self-esteem, rumination and parental divorce on depression among Chinese juvenile delinquents. Sci Rep 2023 Jan 31;13(1):1793. [doi: 10.1038/s41598-023-28626-9] [Medline: 36720911]
69. Miranda R, Tsypes A, Gallagher M, Rajappa K. Rumination and hopelessness as mediators of the relation between perceived emotion dysregulation and suicidal ideation. Cogn Ther Res 2013 Aug;37(4):786-795. [doi: 10.1007/s10608-013-9524-5]
70. Nolen-Hoeksema S, Wolfson A, Mumme D, Guskin K. Helplessness in children of depressed and nondepressed mothers. Dev Psychol 1995;31(3):377-387. [doi: 10.1037/0012-1649.31.3.377]
71. Jones L, Hastings RP, Totsika V, Keane L, Rhule N. Child behavior problems and parental well-being in families of children with autism: the mediating role of mindfulness and acceptance. Am J Intellect Dev Disabil 2014 Mar;119(2):171-185. [doi: 10.1352/1944-7558-119.2.171] [Medline: 24679352]
72. Wang H, Luo W, Huang W, et al. How sleep duration mediated childhood trauma and internet addiction of the vocational college population in urban areas of south China. Front Psychiatry 2022;13:1088172. [doi: 10.3389/fpsyt.2022.1088172] [Medline: 36713913]
73. Hsieh YP, Shen ACT, Wei HS, Feng JY, Huang SCY, Hwa HL. Associations between child maltreatment, PTSD, and internet addiction among Taiwanese students. Comput Human Behav 2016 Mar;56:209-214. [doi: 10.1016/j.chb.2015.11.048]
74. McKenna KYA, Green AS, Gleason MEJ. Relationship formation on the internet: what's the big attraction? J Soc Issues 2002 Jan;58(1):9-31. [doi: 10.1111/1540-4560.00246]
75. Laconi S, Kaliszewska-Czeremska K, Gnisci A, et al. Cross-cultural study of problematic internet use in nine European countries. Comput Human Behav 2018 Jul;84:430-440. [doi: 10.1016/j.chb.2018.03.020]
76. Mayer ED, Peterson R, Kim E, et al. What survivors of sexual violence want when disclosing their experiences in person or online: qualitative interview analysis. J Med Internet Res 2025 Sep 5;27:e73497. [doi: 10.2196/73497] [Medline: 40911341]
77. Machulska A, Kleinke K, Eiler TJ, et al. Retraining automatic action tendencies for smoking using mobile phone–based approach-avoidance bias training: a study protocol for a randomized controlled study. Trials 2019 Dec 12;20(1):720. [doi: 10.1186/s13063-019-3835-0] [Medline: 31831080]
78. Shinetsetseg O, Jung YH, Park YS, Park EC, Jang SY. Association between smartphone addiction and suicide. Int J Environ Res Public Health 2022 Sep 15;19(18):11600. [doi: 10.3390/ijerph191811600] [Medline: 36141872]
79. Grant-Allen G, Wang L, Amini J, Dhaliwal S, Sinyor M, Mitchell RH. Self-harm and suicide-related content on TikTok: thematic analysis. J Med Internet Res 2025 Sep 18;27:e77828. [doi: 10.2196/77828] [Medline: 40811047]
80. Wang X, Ding T, Lai X, Yang C, Luo J. Negative life events, negative copying style, and internet addiction in middle school students: a large two-year follow-up study. Int J Ment Health Addict 2023 Mar 28;2023:1-11. [doi: 10.1007/s11469-023-01045-7] [Medline: 37363770]
81. Zhang Y, Qi Y, Ma Y. The associations between social support and problematic mobile phone use among children and adolescents: a three-level meta-analysis. J Youth Adolesc 2025 Jan;54(1):158-175. [doi: 10.1007/s10964-024-02055-x] [Medline: 39033221]

XSL•FO
RenderX

82. Cheng J, Peng C, Rong F, Wang Y, Tan Y, Yu Y. Mobile phone addiction and suicide behaviors among Chinese adolescents: the mediation of poor sleep quality. J Behav Addict 2024 Mar 26;13(1):88-101. [doi: 10.1556/2006.2023.00078] [Medline: 38224348]
83. Hall JA, Pennington N, Merolla AJ. Which mediated social interactions satisfy the need to belong? J Comput Mediat Commun 2022 Nov 4;28(1). [doi: 10.1093/jcmc/zmac026]
84. Li J, Zhan D, Zhou Y, Gao X. Loneliness and problematic mobile phone use among adolescents during the COVID-19 pandemic: the roles of escape motivation and self-control. Addict Behav 2021 Jul;118:106857. [doi: 10.1016/j.addbeh.2021.106857] [Medline: 33676160]
85. Lam KYI. Disclosure and psychological well-being of sexually abused adolescents in Hong Kong. J Child Sex Abus 2015;24(7):731-752. [doi: 10.1080/10538712.2015.1077364] [Medline: 26479960]
86. Zhang J, Li Y, Li J, et al. The effect of life events, resilience, self-esteem, and coping styles on aggressive behavior among left-behind adolescents: structural equation modeling. Front Psychiatry 2023;14:991608. [doi: 10.3389/fpsyt.2023.991608] [Medline: 36793944]
87. He X, Chen S, Yu Q, Yang P, Yang B. Correlations between problematic internet use and suicidal behavior among Chinese adolescents: a systematic review and meta-analysis. Front Psychiatry 2024;15:1484809. [doi: 10.3389/fpsyt.2024.1484809] [Medline: 39588549]

Abbreviations

ASLEC: Adolescent Self-Rating Life Event Checklist
IA: internet addiction
IAT: Internet Addiction Test
JARS: Journal Article Reporting Standards
KMO: Kaiser-Meyer-Olkin
NLE: negative life events
SA: sexual abuse
STB: suicidal thoughts and behaviors
WHO: World Health Organization

Edited by S Brini; submitted 07.Oct.2025; peer-reviewed by A AL-Asadi, L Liu; accepted 06.Feb.2026; published 25.Mar.2026.

Please cite as:
Zhu J, Jiang Q, Zhang F, Liu H, Yu S, Sun J, Zhao Z, Liu L, Yuan L
Suicidal Thoughts and Behaviors Among Chinese Adolescents in Relation to Negative Life Events, Internet Addiction, and Sexual Abuse: Cross-Sectional Study
J Med Internet Res 2026;28:e85371
URL: https://www.jmir.org/2026/1/e85371
doi:10.2196/85371

© Juanfang Zhu, Qinqin Jiang, Feng Zhang, Hao Liu, Shaojie Yu, Jinhai Sun, Zhe Zhao, Lijuan Liu, Lei Yuan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Embracing Co-Design and Interprofessional Teamwork to Build an Innovative Dashboard for a National Social Needs Screening and Referral Clinical Intervention in the Veterans Health Administration: Design and Development Study

Lauren E Russell[1*], MPH, MPP; Sarah M Leder[1*], MSW; Jaime Boris[1], MHA, MS; Alicia J Cohen[2,3,4], MD, MSc; Christopher W Halladay[2], ScM; Kathleen M Mitchell[5], MPH; Sydney C Ruggles[5], MA, MS; Andrea Berkheimer[6], LCSW; Chava Sonnier[1], MSOD; Michelle C Wilcox[7], LCSW, LASUDC; Ernest Moy[1], MD, MPH; Meaghan A Kennedy[5,8], MD, MPH

[1]Office of Health Equity, Veterans Health Administration, United States Department of Veterans Affairs, 810 Vermont Ave NW, Washington, DC, United States

[2]Center of Innovation in Transformative Health Systems Research to Improve Veteran Equity and Independence, Veterans Health Administration, United States Department of Veterans Affairs, Providence, RI, United States

[3]Department of Family Medicine, Alpert Medical School, Brown University, Providence, RI, United States

[4]Department of Health Services, Policy, and Practice, School of Public Health, Brown University, Providence, RI, United States

[5]New England Geriatric Research, Education, and Clinical Center, VA Bedford Healthcare System, United States Department of Veterans Affairs, Bedford, MA, United States

[6]Social Work Services, James E. Van Zandt Veterans Affairs Medical Center, United States Department of Veterans Affairs, Altoona, PA, United States

[7]Mental Health Services, George E. Wahlen Veterans Affairs Medical Center, United States Department of Veterans Affairs, Salt Lake City, UT, United States

[8]Department of Family Medicine, Chobanian and Avedisian School of Medicine, Boston University, Boston, MA, United States

[*]these authors contributed equally

Corresponding Author:
Sarah M Leder, MSW
Office of Health Equity, Veterans Health Administration, United States Department of Veterans Affairs, 810 Vermont Ave NW, Washington, DC, United States

Abstract

Background: Assessing Circumstances and Offering Resources for Needs (ACORN) is a US Department of Veterans Affairs (VA) clinical intervention designed to identify and address social needs to improve health and well-being among all veterans. We co-designed the ACORN Dashboard to facilitate access to real-time social needs and intervention data for VA clinical care teams and leadership.

Objective: This study aimed to (1) describe the iterative development of the ACORN Dashboard, (2) assess end user feedback and Dashboard usage, and (3) discuss the role of social needs dashboards in facilitating continuous quality improvement in health care settings.

Methods: An interprofessional team of subject-matter experts and end user feedback contributed to the design. Phase 1 included more than 7 months of weekly working meetings. We initially constructed a wireframe in Microsoft PowerPoint, then translated it into a prototype in Power BI, a data visualization software. Using Microsoft Power BI, we built data visualizations to communicate population-level sociodemographic and ACORN screening data. Through feedback sessions, staff from 8 VA medical centers (VAMCs) reviewed the prototype and recommended improvements regarding the Dashboard's purpose, content, and usability. Phase 2 involved 6 weeks of weekly working meetings, where we developed and iteratively refined 5 written drafts of clinically relevant variables for potential inclusion in the Patient-Level Data Page. This list informed a Power BI prototype. We also developed the ACORN Implementation Map page in Power BI to display implementation locations and settings. We again used feedback sessions with 8 VAMCs to review and refine the newly added pages and discuss improvements. To assess usage, we obtained metadata from a VA-specific Power BI report and user experience data from an ACORN VAMC survey.

Results: The ACORN Dashboard displays national data that are updated daily, reflecting 83,546 screens administered across 82 VAMCs facilities between July 1, 2021, and April 30, 2025. The Dashboard was viewed 18,192 times by 2251 unique users, and, on average, 263 (SD 91.2) unique users viewed the Dashboard every month between October 1, 2023, and April 30, 2025. Dashboard variables include the number of screens completed, sociodemographic characteristics of veterans screened, prevalence

XSL•FO
RenderX

of social needs, and interventions provided to address needs. Phase 1 semistructured feedback sessions included recommendations for a page with patient-level data to supplement the population-level pages, incorporation of additional filters to select specific data, and development of a user guide. In phase 2, key insights included enhancement of end users' ability to search by veteran or staff name, guidance about screening frequency, changing the display order of variables, and the inclusion of variable definitions.

Conclusions: Using co-design to develop, maintain, and continually refine data dashboards enhances implementation of social screening and interventions in health care settings. In addition to supporting individual-level patient care, population-level dashboard data inform continuous quality improvement, promote health equity, and identify gaps in services to address identified needs.

(J Med Internet Res 2026;28:e81846) doi:10.2196/81846

KEYWORDS

data dashboard; social needs; social determinants; social determinants of health; veterans

Introduction

As electronic health record (EHR) usage has increased over the past two decades, so too has the volume of patient medical and demographic data collected and maintained by health systems [1]. Data-driven technologies in health care settings include tools such as clinical and quality dashboards [2,3] that function through the collection, analysis, and application of individual- and population-level health information [4]. The integration of data-driven technologies into health systems and the use of these technologies as part of routine clinical care can enhance quality and patient safety, improve health outcomes, inform tailored clinical care and the provision of social and medical services, and contribute to public health and health services research [4-6]. However, data-driven tools should be developed and implemented with a thoughtful and intentional approach, with the end users at the center of the design.

Coinciding with the rise in technology use in health care, the value of medical and social care integration has been increasingly recognized among health care professionals, policymakers, and funders [7-11]. This has led health systems to improve their efforts to collect information on health-related social needs (herein referred to as "social needs"), which are social and economic aspects of individuals' lives that affect their health and well-being [12]. Examples of these needs include housing instability, food insecurity, unemployment, lack of access to telehealth, and social isolation. Although many health systems now routinely collect information on social needs, which is an accreditation requirement of the Joint Commission [11], social needs data are typically not readily available to health professionals in a user-friendly way (ie, without requiring a comprehensive chart review) to inform individual-level care and population health management [13,14].

The US Department of Veterans Affairs (VA) has often been at the forefront of integrating social and medical care through its provision of health care and social services supplemented by its financial investment in innovation and quality improvement (QI) efforts. In 2012, the VA implemented universal screening for housing instability [15], followed in 2017 by universal screening for food insecurity [16]. In addition to these efforts, the VA also offers a range of resources and supportive services for veterans with unmet social needs, such as supportive housing services, vocational rehabilitation, assistance for justice-involved veterans, digital device access and digital health literacy support, and peer-support services [17-21]. Given its robust infrastructure of internal supports and collaborations with Veteran Service Organizations, the VA is uniquely positioned to respond to the medical and social needs of the patients it serves. Developed in 2018, Assessing Circumstances and Offering Resources for Needs (ACORN) is a VA clinical intervention aimed at identifying and addressing social needs to improve health and well-being among all veterans [22,23]. Since 2021, ACORN has been implemented in partnership with the VA Office of Health Equity and the VA National Social Work Program, Care Management and Social Work Services. By the end of April 2025, clinical care teams had implemented ACORN in a variety of clinical settings in 82 VA medical centers (VAMCs) across 37 states and Puerto Rico. Adoption has largely been supported by implementation tools and technical support provided through a national ACORN Community of Practice [24]. Since the ACORN pilot in 2019 to 2020 [22], frontline clinical teams using ACORN have consistently shown interest in reviewing their own data. Doing so allows them to identify the most frequently endorsed social needs within their VAMC or clinical setting, as well as the most common resources or referrals provided to veterans who report any social needs. During initial pilot phases, our team responded to data requests on a case-by-case basis, often responding to requests received via email or through secure instant message (ie, Microsoft Teams). As screening expanded to more VAMCs, it became apparent that fulfilling ad hoc, VAMC-specific data requests was an unsustainable method of technical support. Therefore, we focused on developing a data sharing method that would enable VA leadership and VAMCs to access their data directly. To facilitate access to real-time social needs screening and intervention data for VA clinical care teams and leadership, we iteratively developed and refined the ACORN Dashboard, incorporating end user feedback at multiple time points. Similar to the approach used to develop the ACORN screening tool in 2018 [22], we actively involved end users in the ACORN Dashboard design process.

Data dashboards providing clinical care teams with immediate access to patient information can enhance adherence to quality guidelines and improve patient health outcomes [25]. Although existing literature describes the development of clinical dashboards [26-29], little has been published to date describing the development, usage, and effectiveness of social needs dashboards intended for use in health care settings. In this paper,

XSL•FO
RenderX

we aimed to (1) describe the iterative development of the ACORN Dashboard, (2) examine usage and feedback about the ACORN Dashboard, and (3) discuss the role of data dashboards in facilitating continuous QI in health care settings.

Methods

Project Overview

The ACORN Dashboard was developed over two phases: (1) initial development, consisting of co-design and build of population-level data pages; and (2) expansion, which included the addition of a Patient-Level Data Page based on end user feedback (Figure 1). The ACORN Dashboard Workgroup consisted of 4 key members: a public health practitioner involved in co-developing ACORN and co-leading VA-wide implementation (LER); a biostatistician with robust knowledge of VA data, including ACORN-specific data capture elements, and coding expertise (CWH); a data analyst with a background in social work and expertise in the ACORN dataset and Microsoft Power BI (SML); and an exercise physiologist with a background in health administration and experience in building Power BI reports (JB).

Co-design principles were incorporated throughout to ensure that the product was aligned with end user preferences and the needs of interprofessional clinical staff and leaders. Through feedback sessions in both phases, 37 staff (including 4 staff across 3 VAMCs who participated in both phase 1 and phase 2) from a total of 13 VAMCs reviewed the design and provided critical insights regarding the usability of, and content included in, the Dashboard.

Figure 1. Assessing Circumstances and Offering Resources for Needs (ACORN) Dashboard development timeline for phases 1 and 2 (March 2023 to November 2024). This timeline depicts how each phase of development consisted of 3 key steps: in phase 1, build, beta test, and launch; and in phase 2, expand, beta, and release. The ACORN Dashboard data visualization team, which developed the initial wireframes of the Dashboard in PowerPoint, consisted of 8 individuals with diverse backgrounds and expertise, including clinical social work, nursing, public health, clinical informatics and data analysis, and visual design. A smaller team of members from the ACORN Dashboard data visualization team then formed the ACORN Dashboard Workgroup. This workgroup transformed the PowerPoint wireframes into a functional prototype in Power BI, a data visualization software. This same Workgroup supported the beta testing in both phase 1 and phase 2, as well as the expansion of the Dashboard in phase 2. The arrow at the bottom of the figure depicts the incorporation of iterative feedback throughout development. VA: US Department of Veterans Affairs.

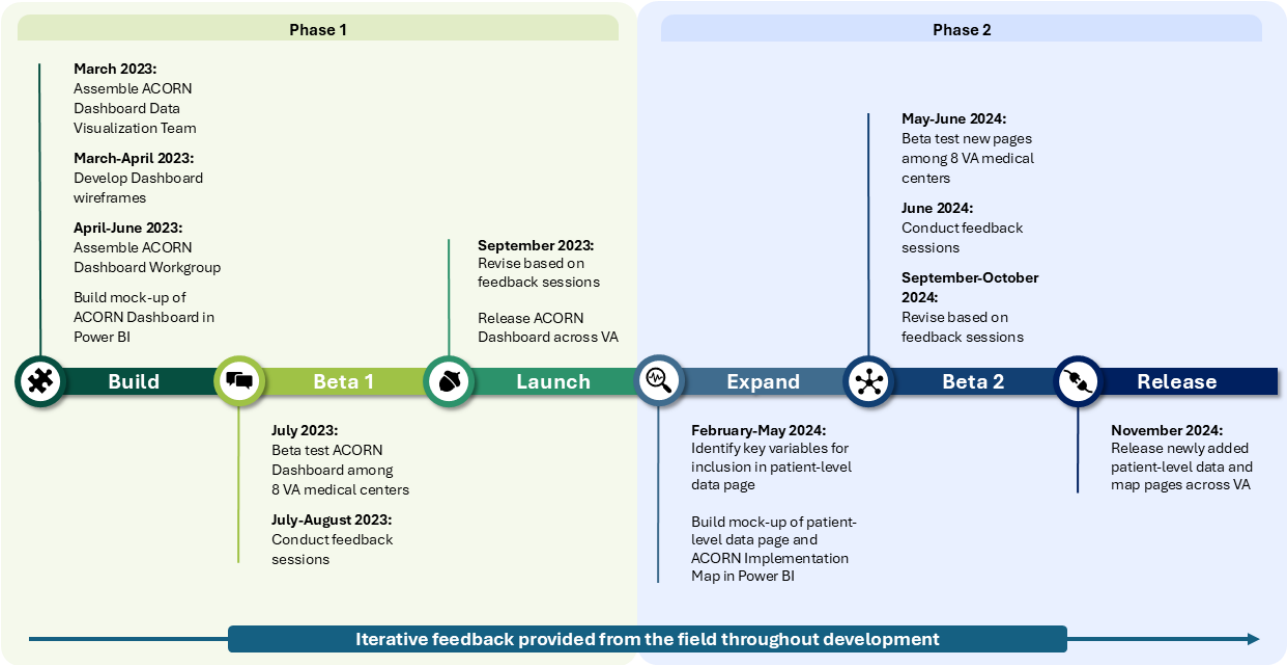

Ethical Considerations

As this work was conducted for the purposes of program evaluation and QI, it was determined by the VA Central Institutional Review Board to be a nonresearch activity and did not require regulatory review. As a nonresearch activity, formal informed consent was not obtained; staff who participated in pilot testing of the ACORN Dashboard and corresponding feedback sessions were informed that their participation was voluntary, that notes would be taken, and that findings would be used to inform the development and enhancement of end user tools. Staff participated within their routine roles and did not receive specific compensation for engaging in pilot testing or feedback sessions.

ACORN Dashboard Development

Phase 1: ACORN Dashboard Creation

Phase 1: Co-Design and Initial Build

We embraced a co-design approach, incorporating interprofessional subject-matter expertise and end user input from the outset. To ensure a functional "end product," we formed the ACORN Dashboard data visualization team in March 2023. We recruited VA staff knowledgeable in the importance of population and individual-level data to ensure the value of the Dashboard to both frontline clinical staff and VA leaders, whom we anticipated would be our end users. Additionally, we included data analysts, specifically, subject-matter experts who possessed a comprehensive understanding of the ACORN

XSL•FO
RenderX

dataset and those who would be able to create a functional end product. This interprofessional team also possessed the knowledge and expertise to determine which data were clinically relevant and feasible to display in a dashboard. The final data visualization team consisted of 8 individuals with expertise spanning clinical social work, nursing, public health, clinical informatics and data analysis, and visual design.

The data visualization team met weekly through the beginning of April 2023. This team brainstormed the purpose, content, and layout of the ACORN Dashboard and created several wireframes (ie, mockups of dashboard pages including tables, charts, and figures). To ensure all members of this team could access and make edits on the wireframes, we used Microsoft PowerPoint slides to create mockups of each page that we would ultimately create in Power BI, the data visualization software used across the Veterans Health Administration. Using Power BI Desktop enabled us to design, build, and iteratively refine the Dashboard, while uploading it to Power BI Service (or Power BI Server) generated a shareable link that could be disseminated to all VA employees. Although we initially planned for the Dashboard to be a single page, it quickly became evident that we would need to expand to multiple pages and determine which data could be grouped together, for example, a quick snapshot of the most frequently requested ACORN data (ie, number and percentage of those endorsing each social need), veterans' sociodemographic information, and the most commonly provided resources and referrals to address needs. Once the data visualization team reached consensus on a complete wireframe, we gathered feedback from staff in the VA Office of Health Equity and members of the ACORN Leadership Team, which consists of Office of Health Equity and National Social Work Program leadership and staff, health services researchers who are also VA primary care physicians, data scientists or analysts, and biostatisticians.

Incorporating this feedback, a subset of members from the data visualization team then formed the smaller ACORN Dashboard Workgroup (hereafter the "workgroup") to code and build the initial Dashboard. The workgroup met weekly between mid-April 2023 and the end of June 2023 to discuss progress made on the Power BI draft of the Dashboard, review outstanding issues, and verify next steps. As the primary goal of the Dashboard was to provide data collected across the entire VA, including all 139 VAMCs, we quickly established a threshold to protect patient privacy. Adhering to the VA Office of Health Equity standard practice for protecting protected health information (PHI), if any cell in any table had fewer than 13 patients, the true count would be hidden.

After approximately 10 weeks, we completed the first draft of the Dashboard and uploaded it to the Office of Health Equity's Power BI online workspace.

Phase 1: Beta Test Dashboard

In early July 2023, we recruited staff from 8 VAMCs to serve as "beta testers." Medical centers were recruited through ACORN Community of Practice calls and direct email outreach. We included medical centers in various stages of ACORN implementation (ie, preimplementation, implementation, and sustainment) and across different clinical settings and geographic locations to ensure a diverse range of perspectives and feedback on the initial dashboard design and content.

Phase 1: Collect End User Feedback Through Semistructured Feedback Sessions

Staff from each of the 8 VAMCs agreed to test the Dashboard over the course of a month (July 2023) and participate in a 1-hour semistructured feedback session with 2 to 3 workgroup members between late July and early August 2023. Prior to conducting semistructured feedback sessions, we sent a list of usability questions to each VAMC (Supplemental File 1 in Multimedia Appendix 1). We then used these questions to guide our feedback sessions. During each session, 1 member of the workgroup took detailed notes in a structured template to capture quotes, barriers to and facilitators of Dashboard usage, and opportunities for improvement, while the other members asked questions.

Phase 1: Analyze End User Feedback From Semistructured Feedback Sessions

After the conclusion of the 9 feedback sessions (1 VAMC requested 2 feedback sessions to accommodate varying staff schedules), findings were summarized in a template based on the feedback session guide, and discrepancies in the interpretation of any of the responses were resolved through consensus discussion between the workgroup members in attendance at the respective feedback session. Additionally, the workgroup members who participated in the feedback sessions reviewed responses for each template domain to identify the emergence of key themes. The full workgroup then reconvened to review the summary template, which included the compilation of "action items" from all the feedback sessions. First, we discussed the key insights identified by workgroup members who participated in the feedback sessions. We reviewed the summary template section by section, both for action item clarity and to resolve any concerns about an action item's categorization (eg, if one team member thought a piece of feedback in the "layout" section should be captured in the "content" section). Second, we wrote a "general recommendation" response for each action item and grouped them into three categories: (1) action items with a clear plan for resolution, (2) action items needing additional research or follow-up from our team to glean additional insight or assess the feasibility, and (3) action items outside of the scope of the Dashboard. Third, we assessed the feasibility of addressing the key action items. Feasibility was assessed based on two criteria: (1) whether our team possessed the skill set and expertise to make the change prior to the planned release on September 30, 2023; and (2) whether we had the ability to make the modification within the constraints of Power BI when applicable for action items. Finally, the action items deemed feasible were prioritized based on anticipated time to complete them, with those taking the least amount of time to resolve being at the top of the list. We addressed all action items on the priority list before the announcement and online release of the Dashboard to all VA staff at the end of September 2023.

XSL•FO
RenderX

Phase 2: Dashboard Expansion

Phase 2: Refinement and Expansion of Dashboard Features

The impetus for phase 2 was 2-fold. First, multiple end users noted the importance of accessing patient-level data during phase 1 feedback sessions (see *Results* section). Furthermore, after the initial ACORN Dashboard launch (phase 1), we received multiple requests from end users to incorporate a Patient-Level Data Page that could be used to inform clinical practice and support care teams in following up with individual patients. The requests often included specific patient-level data elements that end users sought to use to appropriately tailor clinical care. The workgroup used these requests to develop an initial list of variables for inclusion in a Patient-Level Data Page. We also reviewed 2 existing VA dashboards that include patient-level data—the Food Insecurity Summary Screen Dashboard developed by the VA Food Security Office and the Primary Care Equity Dashboard [30] developed by researchers at the VA Center for Health Equity and Research Promotion. To protect PHI on the Patient-Level Data Page, we used a dynamic filtering feature on Power BI Service and VA data tables that limits internal users' data access permissions based on their assigned VAMC (location) and the overall data permissions granted by the VA.

Between the end of February 2024 and the beginning of April 2024, we developed 5 drafts of key variables for inclusion. Once members of the workgroup reached consensus on the complete list of patient-level variables, we reviewed and gathered feedback from staff in the Office of Health Equity and members of the ACORN Leadership Team. In addition to the Patient-Level Data Page, we also agreed to develop an ACORN Implementation Map to display VAMCs and clinical settings in which ACORN had been implemented thus far.

Phase 2: Beta Test Dashboard

After completing a draft version of the Patient-Level Data Page and the ACORN Implementation Map in Power BI, we shared a link to our test version in mid-May 2024 with 8 VAMCs. As phase 2 beta testing included the Patient-Level Data Page, which relies on Veteran PHI, we only included sites that had already implemented ACORN in this testing. This ensured that participants could view real, local patient (specifically PHI) data. Staff at each VAMC could only see patient-level data that they would otherwise have clinical access to view within the VA EHR. Similar to phase 1, we recruited staff at these VAMCs to serve as "beta testers" through the ACORN Community of Practice and direct email outreach based on their involvement in ACORN. The 8 VAMCs agreed to test the Patient-Level Data Page over the course of a month (mid-May to mid-June 2024) and to meet for a 45‐ to 60-minute feedback session with 2 to 3 members of the Workgroup between mid- to late-June 2024.

Phase 2: Collect End User Feedback Through Semistructured Feedback Sessions

Similar to phase 1, we conducted semistructured feedback sessions with each beta testing VAMC. We aimed to obtain insights to refine the Patient-Level Data Page and the ACORN Implementation Map to make sure they were (1) useful to VAMCs, (2) being interpreted as intended, and (3) appropriately tailored to meet the needs of frontline staff. During each session, 1 member of the workgroup served as the lead or facilitator, while the other served as the primary notetaker or recorder. However, both the lead or facilitator and primary notetaker or recorder took detailed notes in a structured template to capture responses about features that should be added or removed and feedback on the content, layout, and usability of the Patient-Level Data Page and any requested changes to the ACORN Implementation Map.

Phase 2: Analyze End User Feedback From Semistructured Feedback Sessions

After completing all 8 feedback sessions, 2 members of the workgroup summarized the findings in a template based on the feedback session guide and resolved discrepancies in the interpretation of any of the findings through consensus discussion. These workgroup members also reviewed responses from each VAMC to identify key themes gleaned from feedback sessions.

The full workgroup reconvened in September 2024 to review the summary template and a list of key action items put together by the workgroup members who facilitated the feedback sessions. We reviewed the key action items one by one and placed them into the following categories: (1) feasible changes to be made to existing content; (2) feasible variables and features to incorporate before release; (3) deletions to existing content; and (4) follow-up and outstanding questions or issues, which included notes about requested changes that were not feasible due to Power BI limitations or the available time window to make modifications. Between September and October 2024, we revised the Patient-Level Page to address the findings in categories 1 to 3 and worked to both answer remaining issues in category 4 and communicate to the "beta testers" which of the requested changes were not feasible.

Dashboard User Data

Usage

We used the VA-developed "Usage Metrics" Power BI report, which allows Power BI workspace administrators to access metadata related to the reports they have created. The Usage Metrics report allowed us to track the total number of Dashboard views and the number of unique viewers over time.

User Experience

As part of an annual survey used to understand how VAMCs are implementing ACORN, staff were asked to complete 2 questions about how frequently they used the ACORN Dashboard and how useful they found it. Staff from all 63 VAMCs that had implemented ACORN between October 2023 and September 2024 were invited to complete the survey via a Qualtrics link distributed via email. Invited staff were asked to provide 1 response per VAMC. When we received more than 1 response for a given VAMC, the most complete response (ie, most questions completed) was used for analyses.

XSL•FO
RenderX

Results

Phase 1 End User Feedback Session Findings

Between late July and early August 2023, 9 feedback sessions were conducted with the 8 VAMCs who served as "beta testers" of the ACORN Dashboard. A total of 27 staff participated, with a median of 3 (IQR 1‑8) staff per VAMC. Participants represented a broad range of professions, including clinical leadership, health informatics, and frontline staff (ie, physicians, social workers, and nurses).

Overall, respondents found the ACORN Dashboard informative, intuitive, easy to navigate, and user-friendly. However, staff suggested several content and layout changes and noted several points of confusion or difficulty (Textbox 1). They reported that a variety of staff roles or disciplines would likely use the ACORN Dashboard, including clinical care team members or frontline staff (eg, social workers and nurses), health informaticists, executive leadership teams (eg, VAMC directors and chiefs of staff), and team members of locally tailored efforts to improve health equity and address health disparities. Key refinements made based on phase 1 end user feedback included modifying the names of certain variables, domains, and categories on the Dashboard; adding additional filters to allow staff to further refine their data search by quarter within a fiscal year and by domain on the Domain Trends page; developing a user guide and frequently asked questions resource; and providing the data source for included variables. This end user feedback also prompted the Workgroup to begin conceptualizing a Patient-Level Data Page (phase 2).

Textbox 1. Key insights and refinements from Assessing Circumstances and Offering Resources for Needs (ACORN) dashboard development phase 1 (July to August 2023).

Include patient-level data
Modify language for ease of use (ie, change "–" to "<13")
Enhance available filters (eg, fiscal year and quarter)
Develop a user guide and frequently asked questions resource
Clarify data source (ie, data derived from the ACORN electronic health record template)

The key insights and refinements presented in Textbox 1 are based on recommendations gleaned from 9 feedback sessions conducted with 27 staff across 8 VAMCs for phase 1 of the ACORN Dashboard development between July and August 2023. Participants represented clinical leadership, health informatics, and care team members or frontline staff (ie, physicians, social workers, and nurses).

Phase 2: End User Feedback Session Findings

In June 2024, we conducted feedback sessions with staff from 8 VAMCs (n=3, 37.5%, of which also participated in phase 1 beta testing) who agreed to serve as "beta testers" of the Patient-Level Data Page and the ACORN Implementation Map. Overall, 13 staff representing a variety of roles or disciplines, including social work, mental health, and nursing, participated in the feedback sessions, with a median of 1.5 (IQR 1‑3) attendees per VAMC.

Feedback regarding the Patient-Level Data Page was overall positive, with one VAMC reporting, "This is perfect. It's exactly what we're looking for," and another sharing, "We really wanted to know which patients were screened and things like that. We've really been hoping for something like this...this will be very helpful." The majority of staff found the page easy to follow and understand. Several VAMCs mentioned their appreciation for the ability to export data from Power BI to Excel or PDF, noting that this functionality made it easier to share data with colleagues and VA leadership. Staff generally agreed that the most critical domains for inclusion were the active suicide flag indicator (an indicator for suicide risk in the VA EHR), positive social needs domains, and the Veterans' Care Assessment Needs score (a VA-validated risk prediction model) [31]. One VAMC noted that they found the Patient-Level Data Page included too many domains, which made it difficult to easily navigate the page and duplicated information readily available in veterans' EHR (eg, sex, race, ethnicity, and sexual orientation). This VAMC also noted that many of the domains were not crucial data points for informing follow-up care. Staff also proposed several enhancements (Textbox 2). Key refinements to the Patient-Level Data Page made based on phase 2 end user feedback included enhancing the end user's ability to search by veteran or staff name; incorporating clarification about screening frequency (ie, date of next recommended screen); modifying the display order of the variables; providing data definitions for variables in an updated version of the user guide and frequently asked questions resource; and improving the overall readability and usability of the Dashboard contents.

Textbox 2. Key insights and refinements from Assessing Circumstances and Offering Resources for Needs Dashboard development phase 2 (July 2024).

Enhance end user's ability to search by veteran or US Department of Veterans Affairs staff name
Incorporate guidance about screening frequency
Modify order of variables
Define included variables
Improve readability and usability (eg, increase font size and add alternative text)

XSL•FO
RenderX

The key insights and refinements presented in Textbox 2 are based on recommendations gathered during the 8 feedback sessions conducted with 13 staff across 8 VAMCs for phase 2 of the ACORN Dashboard development during July 2024. Participants represented clinical leadership and care team members or frontline staff from several disciplines (eg, social work, mental health, and nursing).

For the ACORN Implementation Map, staff were generally appreciative of this added feature, and no proposed changes were suggested during our feedback sessions.

Final ACORN Dashboard Components (Includes Phase 1 and Phase 2)

The original ACORN Dashboard (phase 1) was launched in September 2023, and the Patient-Level Data Page and ACORN Implementation Map (phase 2) were released in November 2024 [24] (Figure 2). The overall aim of the ACORN Dashboard is to increase access to data collected through the ACORN screening tool in the VA EHR. The ACORN Dashboard helps VA care teams and leadership: (1) understand the social needs impacting patients, (2) track the types of resources and referrals provided to patients who screen positive on ACORN, and (3) observe how screening rates and the provision of resources and referrals vary by sociodemographic factors and over time. At the end of the study period, the Dashboard reflected 83,546 screens among 76,357 unique patients across 82 VAMCs since July 1, 2021. Among the 47,683 (62.4%) unique veterans screening positive for one or more social needs on ACORN, 35,897 (75.3%) received a resource and/or referral to a relevant VA or community service.

As previously noted, Dashboard variables, including screened veterans' sociodemographic information, ACORN screening and resource and referral data, as well as clinical characteristics, are pulled from the VA Corporate Data Warehouse and are updated nightly at 11:59 PM Eastern Time (see Supplemental File 2 in Multimedia Appendix 2 for a complete list of variables).

Figure 2. Assessing Circumstances and Offering Resources for Needs (ACORN) Dashboard page examples. (A) The Overview page; (**B**) the ACORN Implementation Map page; (**C**) the Demographics page; (**D**) the Resources and Referrals by Social Needs Domain page. As of May 1, 2025, the date on which the above screenshots were taken and included data gathered through April 30, 2025, the Dashboard reflected 83,546 screens among 76,357 unique patients across 82 Veterans Affairs Medical Centers since July 1, 2021. Among the 47,683 (62.4%) unique veterans screening positive for one or more social needs on ACORN, 35,897 (75.3%) received a resource and/or referral to a relevant VA or community service. Panels A, C, and D were developed as part of ACORN Dashboard development phase 1 (March to September 2023), while panel B was developed as part of ACORN Dashboard development phase 2 (February to November 2024). (The ACORN Dashboard continually updates daily, so the data shown are for the study period.)

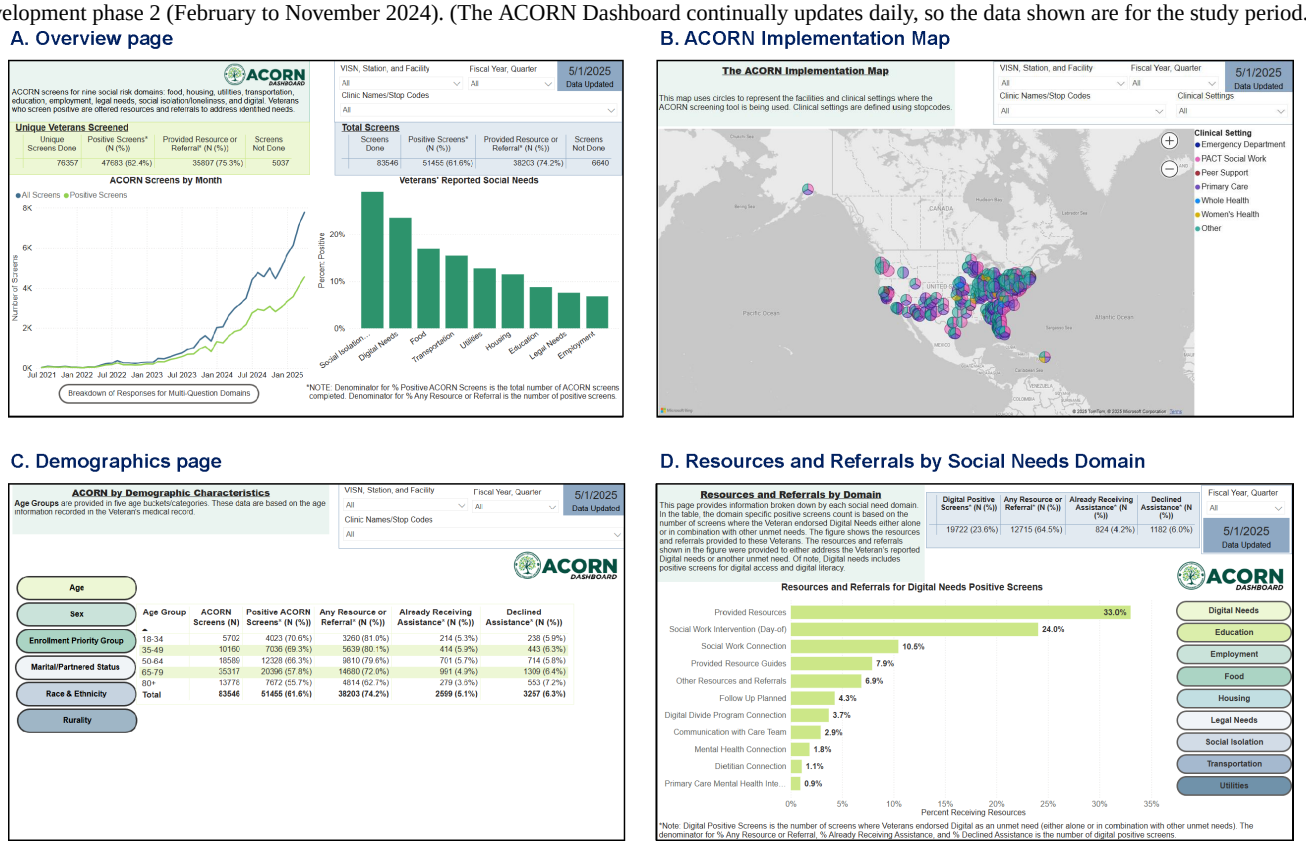

Dashboard Training and Tools

In conjunction with the initial release of the Dashboard (phase 1), members of the workgroup provided a tutorial for the ACORN Community of Practice (October 2023) and on a national VA webinar (January 2024). Additionally, email announcements were sent to internal VA staff and VAMCs involved in the ACORN Community of Practice, and an announcement was included on the ACORN SharePoint VAMC, an internal VA website available to all staff. Concurrent with the release of the Dashboard, the workgroup released the ACORN Dashboard User Guide and Frequently Asked Questions (FAQs; Supplemental File 2 in Multimedia Appendix 2). The User Guide and FAQs was created as an additional

XSL•FO
RenderX

resource to provide information asynchronously and for those who might prefer a written reference. The guide contains a brief overview of the ACORN screening tool, a description of the rationale behind the Dashboard, and in-depth descriptions and screen captures of each Dashboard page. Crucially, the User Guide and FAQs provide explanations of features that could cause trouble for end users who are unfamiliar with Power BI. This includes information about the on-screen filters that allow users to limit their view by categories (eg, VAMC, clinic type, and fiscal year). In alignment with our co-design approach, the User Guide and FAQs were developed based on the review of summary notes and questions gathered from end user feedback sessions. After the Implementation Map page was created and Patient-Level Data Page was updated following the same principles as the Patient-Level Data Page (phase 2), a similar communications approach was taken. An overview of the new pages was provided to the ACORN Community of Practice (November 2024), and a more in-depth tutorial was provided to this group in April 2025. Additionally, descriptions and screenshots of the Patient-Level Data Page and ACORN Implementation Map, along with questions from the phase 2 end user sessions, were added to an updated User Guide and FAQs.

ACORN Dashboard User Data

Usage

Between October 1, 2023, and April 30, 2025, the Dashboard was viewed 18,192 times by 2251 unique users. On average, 263 (SD 91.2) unique users viewed the Dashboard every month, with a range of 107 to 552. Three notable spikes in the number of Dashboard views per month (Figure 3) correspond to key ACORN milestones. The first spike in January 2024 (856 total views and 351 unique views) corresponds to the release of the ACORN national template in the VA EHR, which was announced via an email from the VA Health Information Community of Practice and through the ACORN Community of Practice monthly calls, Microsoft Teams Channel chat, and ACORN SharePoint VAMC. The second spike in April 2024 (1310 total views and 277 unique views) coincides with the ACORN screening tool receiving endorsement from the VA Governance Board, which is composed of VA senior leaders who oversee operational activities to ensure decisions are prioritized and aligned with VA's mission and the strategic direction set forth by the VA Under Secretary for Health. Many regional leaders and members of medical centers' executive leadership teams reported receiving awareness of this change from their regional directors or others in their chain of command. Finally, the most recent spike in November 2024 (1448 total views and 552 unique views) corresponds to the release of the Patient-Level Data Page and ACORN Implementation Map, which was announced through the ACORN Community of Practice monthly calls, Teams Channel chat, and ACORN SharePoint VAMC.

Figure 3. Assessing Circumstances and Offering Resources for Needs (ACORN) Dashboard unique and total views by month. October 2023 marks the first full month after the release of the ACORN Dashboard. Data from September 23, 2023, to September 30, 2023, are not shown. Between October 1, 2023, and April 30, 2025, the Dashboard was viewed 18,192 times by 2251 unique users. On average, 263 (SD 91.2) unique users viewed the Dashboard every month, with a range of 107 to 552.

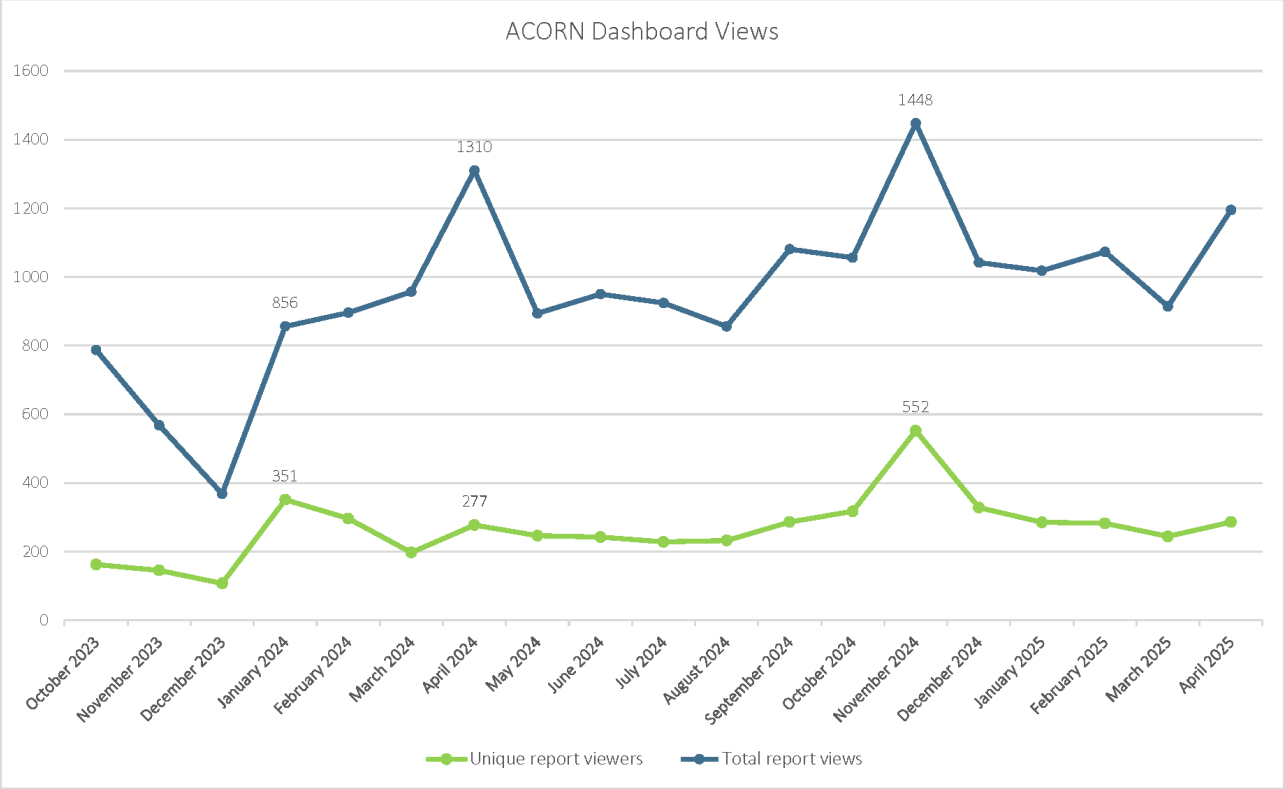

XSL•FO
RenderX

User Experience

Of 63 invited VAMCs, 48 completed the annual VAMC survey in October 2024; of those, 48 completed a question on how frequently they used the Dashboard, and 46 completed a question on how useful they found it. More than half (54.1%) of the VAMCs reported "often" using the Dashboard, 20 (41.7%) reported "sometimes" using the Dashboard, and 2 (4.2%) reported "never" using the Dashboard. Nearly two-thirds (65.2%) of VAMCs reported that the Dashboard was "very useful," 15 (32.6%) reported that the Dashboard was "somewhat useful," and 1 (2.2%) reported that the Dashboard was "not useful at all."

Discussion

Brief Summary of the ACORN Dashboard Co-Design Process and Usage

In this paper, we describe the iterative development, usage, and staff perceptions of a social needs dashboard. The overarching purpose of the ACORN Dashboard is two-fold. First, it is intended to help care team members, including clinicians, nurses, social workers, and other allied health professionals, review individual-level social needs, sociodemographic, and clinical data among veterans screened to guide additional outreach or follow-up and inform local QI efforts to ensure appropriate follow-up for veterans screening positive. Second, it provides population-level social needs data collected through the ACORN EHR template to care team members, local and regional leadership, and national leadership to understand the unmet social needs impacting veterans, identify gaps in the service delivery model, and inform future resource allocation in response to these needs. Using data gathered during this study period, the Dashboard displayed national data gathered from 82 VAMCs facilities during an approximately 4.5-year time span, with 83,546 ACORN screens administered among 76,357 unique veterans. Since the Dashboard's launch, more than 2250 unique users viewed the Dashboard nearly 18,200 times. Phase 1 and phase 2 feedback sessions provided insight into end users' goals and opportunities for Dashboard enhancements. Literature exists on the development of dashboards for patient quality and safety as well as those to address health disparities and promote health equity and to display social determinants (or drivers) of health in relation to certain disease states or clinical conditions [1,29,32-34]. However, to our knowledge, this is the first description of the development and use of a social needs dashboard intended to inform both individual-level clinical care and population-level health management.

Alignment With Existing Literature

Using a co-design process, we leveraged the VA's EHR to create a dashboard for a national VA social needs clinical intervention. Notably, this approach helped us to collaboratively create a data-driven tool that both meets end user needs and provides a range of staff, from clinical providers to senior VAMC and organizational leadership, with access to routinely collected social needs data that are updated daily [4,5,25]. This tool optimizes the use of clinical (individual-level) and population-level health information for end users, which ultimately promotes patient-centered care. Aligned with prior studies, our work demonstrates interest among health professionals in accessing clinical data collected from patients during the provision of routine clinical care [1,35]. The interprofessional co-design approach used to build the ACORN Dashboard also builds upon existing literature highlighting the importance of integrally involving health professionals throughout the development of patient data dashboards to enhance usage [36]. As prior research suggests [29,34], including health professionals, especially from a diverse range of backgrounds and disciplines, as we did in this work, can help ensure that included data are meaningful and clinically relevant to a broad audience.

The Dashboard also aligns with the principles of visual management systems used to monitor and improve care [37]. To maximize the likelihood of engagement and use of this tool, we used a co-design process to build and iteratively revise the ACORN Dashboard. Using this approach meant including relevant partners, the anticipated end users, from the outset. We sought to ensure that VA care team members and leaders relying on this real-time, point-of-care tool could seamlessly identify and use ACORN Dashboard visuals to communicate critical information [38]. As noted in other studies, including end users in the design phase is critical for electronic (or digital) health care tools whose effectiveness largely depends on high levels of care team member usage [39,40]. Additionally, the range of staff roles and disciplines of those involved in the data visualization team, Office of Health Equity, ACORN Leadership Team, and VAMC staff involved in feedback sessions provided valuable insights into diverse perspectives on use contexts and end user needs. This enhanced our understanding of what clinicians, care team members, and leaders perceived as most relevant for tailoring clinical care and informing population health management. Building a strong data visualization team with a diverse skill set played a crucial role in our ability to create an end product that was not only functional but also easy to navigate and aesthetically pleasing. Similar to other studies, engaging end users throughout the Dashboard development and refinement process likely contributed to the acceptability of the prototypes that were beta tested and discussed during phases 1 and 2 [39,41,42]. It may also have contributed to the consistent use of the Dashboard over time, as evidenced by user trends [30,43].

The Dashboard also aligns with the principles of visual management systems used to monitor and improve care [37]. To maximize the likelihood of engagement and use of this tool, we used a co-design process to build and iteratively revise the ACORN Dashboard. Using this approach meant including relevant partners, the anticipated end users, from the outset. We sought to ensure that VA care team members and leaders relying on this real-time, point-of-care tool could seamlessly identify and use ACORN Dashboard visuals to communicate critical information [38]. As noted in other studies, including end users in the design phase is critical for electronic (or digital) health care tools whose effectiveness largely depends on high levels of care team member usage [39,40]. Additionally, the range of staff roles and disciplines of those involved in the data visualization team, Office of Health Equity, ACORN Leadership Team, and VAMC staff involved in feedback sessions provided

XSL•FO
RenderX

valuable insights into diverse perspectives on use contexts and end user needs. This enhanced our understanding of what clinicians, care team members, and leaders perceived as most relevant for tailoring clinical care and informing population health management. Building a strong data visualization team with a diverse skill set played a crucial role in our ability to create an end product that was not only functional but also easy to navigate and aesthetically pleasing. Similar to other studies, engaging end users throughout the Dashboard development and refinement process likely contributed to the acceptability of the prototypes that were beta tested and discussed during phases 1 and 2 [39,41,42]. It may also have contributed to the consistent use of the Dashboard over time, as evidenced by user trends [30,43].

Application of the ACORN Dashboard and Facilitation of Continuous QI Efforts

The practice of creating and using data dashboards aligns with the VA's efforts to not only be a high-reliability organization but also a high-equity-reliability organization [44]. High-equity-reliability organizations consider environmental factors that influence patient health and well-being and recognize that it is imperative to understand patients' feelings and their values [44], which aligns with systematic efforts to identify and address social needs. Many VAMCs have sought to consider environmental factors through the incorporation of social needs screening and referral initiatives. Often, they have conducted this work through QI initiatives.

To inform continuous local and national QI efforts, we collect ongoing end user input to optimize clinical use. In addition, we provide 4 potential avenues for users to find technical assistance and other support [45]. The most accessible is the ACORN Dashboard User Guide and FAQs (described earlier). We also provide support through the ACORN Community of Practice and ACORN "Office Hours" monthly calls. These calls present opportunities for VAMCs to request and receive uniquely tailored technical assistance from members of the ACORN Leadership Team and learn about best practices from other VAMCs involved in the ACORN Community of Practice. Additionally, users can reach out to the Dashboard team using a designated VA email address to troubleshoot errors. To provide teams with the opportunity to access additional VAMC-specific data beyond the scope of what is provided through the Dashboard, our team developed an ACORN Data Pull Request Form in collaboration with colleagues in the Office of Health Equity. Through this process, we also offer consultations to support individual VAMCs in obtaining ACORN screening data for local QI and evaluation efforts. The focus of these requests usually includes in-depth data tailored to local priorities and a desire to track specific outcome measures. These processes keep multiple channels of communication open and have allowed for robust documentation of suggestions for consideration and improvements to the Dashboard.

Using the Dashboard to assess population-level data can support the identification of inequities and provide health systems with data needed to identify and fill gaps in services to address identified social needs. Similar to the VA Primary Care Equity Dashboard [30], we wanted to create a tool through which VA leaders could use the aggregate data to inform VAMC-level QI efforts and resource allocation, and care team members could "drill down" to clinic-level and patient-level data to tailor clinical interventions and enhance patient-centered care offerings. In alignment with existing guidance on incorporating equity into QI efforts [30,46,47], we recognized the importance of allowing staff to view stratified data. To this end, we built the Demographics page, which enables end users to view ACORN data stratified by age, sex, race and ethnicity, and other demographic characteristics.

Insights and Implications for VA and Non-VA Health Settings

It is imperative to acknowledge that the VA is uniquely well positioned to screen for and address social needs, given its emphasis on interprofessional patient-aligned care teams, which include social workers embedded in primary care teams [48], and robust resources and supportive services to address the myriad social needs included in ACORN [17-21]. Although non-VA health systems may not have as many "in-house" comprehensive service offerings as the VA, the co-design process, including the timeline detailing the steps of phases 1 and 2, described in this paper could be used by other health systems to inform a similar process. Even if non-VA health systems are collecting social needs data through different screening instruments, the processes detailed herein, particularly the inclusion of interprofessional health professionals in the design and build and the approach to making iterative refinements made through both phases, could be applied to build a dashboard with individual- and/or population-level data. Additionally, the ACORN Dashboard User Guide & FAQs (Supplemental File 2 in Multimedia Appendix 2) can serve as a template for health systems (and potentially other organizations) looking to develop educational materials and resources for staff.

The link between social needs and health outcomes is well established; thus, collecting social needs data is now viewed by many as a standard of care [7,10,11,22,49-54]. First developed and used in a VAMC in New England [22], the ACORN clinical intervention has now been implemented in approximately 85 VAMCs in all 18 VA regions. In addition to the VA's increased adoption of social needs clinical interventions in recent years, The Joint Commission established a National Patient Safety Goal standard in 2022 that requires hospitals to assess patients' social needs and provide relevant connections to community resources and supportive services to address reported needs [11,54]. To meet this accreditation requirement, both VA and non-VA health systems must screen for, assess, and address social needs using tools such as ACORN [11]. We have received feedback from VAMCs that Dashboard data, including the data visualizations, has supported their efforts to meet this Joint Commission standard. Non-VA health systems may also be able to use a similar strategy if they develop data visualizations to display information gathered through social needs screening instruments used within their clinical settings and programs. Our team is conducting ongoing evaluation efforts to further explore and quantify this impact.

XSL•FO
RenderX

Strengths and Limitations of the ACORN Dashboard

Notable strengths of the ACORN Dashboard include its daily updates, which provide near real-time data to staff and leadership. Using these data, care team members and frontline staff can optimize their provision of clinical care for veterans, track their metrics, and share their innovative work with veterans, other frontline staff, and local, regional, and national leadership. Its user-friendly design makes the ACORN Dashboard well suited to support program evaluation efforts at the local and national levels. The data captured in the Dashboard enable care teams to provide proactive and uniquely tailored services to veterans while optimizing resource use, which can result in more efficient and effective service delivery. Additionally, the development of the Dashboard minimizes delays in data access among care team members and frontline staff by reducing the need for the ACORN Leadership Team to respond to a large volume of VAMC-specific data requests manually. However, the Dashboard does have several limitations. The initial version (phase 1) was restricted to national-level, or aggregate, data; however, we resolved this limitation with the development of the "Patient-Level Data Page" (phase 2). Due to patient privacy concerns, the VAMC-level filters were not initially available for the "Co-Occurring Social Needs by Social Need Domain" and "Resources and Referrals by Social Need Domain" pages due to small cell sizes at the level of granularity for these domain-related data. However, the filters were added in June 2025 after nearly 100,000 screens had been completed across all VAMCs. The Dashboard also only includes data from VAMCs using the ACORN screening tool and the specific clinical settings in which they have implemented ACORN. As a result, data are not representative of all veterans enrolled in the VA, and data within certain social need domains may differ from other VA dashboards or published reports that include similar data (eg, food security).

Limitations of the Approach Used to Build the ACORN Dashboard

Our findings should be interpreted within the context of several limitations. First, one beta testing VAMC in phase 1 was in preimplementation, meaning it had not yet started conducting ACORN screening, and therefore, staff were not able to view their own VAMC-level data during testing. However, their inclusion was intentional as we sought to solicit feedback from VAMCs at all phases of implementation. Those in preimplementation could provide valuable feedback regarding perceived usefulness and perspectives on how to use data from the Dashboard to increase buy-in among local leadership and frontline staff to support ACORN adoption. Additionally, findings may have been susceptible to selection bias because many of the VAMCs involved in our evaluation had a high level of buy-in and engagement. Although the intent of the Dashboard is to provide data to VA leaders at the national, regional, and VAMC levels, leadership was likely underrepresented in beta testing procedures (outside of national leadership from the VA Office of Health Equity and VA National Social Work Program). However, we solicited high-level feedback on the Dashboard components from members of the ACORN Partner Engagement Group, which comprised representatives from offices and programs across the VA and serves as an advisory body to inform the development of objectives and initiatives to support VA-wide implementation of ACORN [24]. Finally, although findings from the end user feedback sessions during phase 1 and phase 2 and the user data support acceptability and use of the ACORN Dashboard, a formal evaluation of ACORN Dashboard usage, satisfaction, and impacts across all VAMCs is needed.

Conclusions

On the basis of continued uptake of the ACORN clinical intervention and use of the ACORN Dashboard, this paper adds critical information about engaging collaborators in a co-design process to build tools that inform patient care and promote the adoption of innovations within health systems. The development, maintenance, and iterative refinement of innovative data dashboards can support the implementation of social needs screening and referral interventions in health care settings. Documenting and tracking patient- and population-level data on social needs informs individual-level care and quality assurance in the provision of necessary follow-up, facilitates population health management, and helps identify health inequities and gaps in available services.

Acknowledgments

The authors would like to express their gratitude to their collaborating VA medical centers (VAMCs), clinical teams, and partners for their willingness to implement Assessing Circumstances and Offering Resources for Needs (ACORN) and contribute to the development of the ACORN Dashboard. The authors appreciate the input provided by other members of the ACORN Leadership Team, including Amy M. Donaldson, LCSW (National Social Work Program, Care Management and Social Work Services); Jennifer A. Koget, MS, LCSW, BCD (National Social Work Program, Care Management and Social Work Services); Jennifer W. Silva, LCSW-S (National Social Work Program, Care Management and Social Work Services); Brittany L. Trabaris, LCSW (Edward Hines Jr Veterans Administration Hospital); and Lisa E. Wootton, LCSW (National Social Work Program, Care Management and Social Work Services). Finally, the authors value the guidance and feedback provided on this manuscript by David A. Frank, MPH (Center for Healthcare Evaluation, Research, and Promotion, Veterans Affairs Pittsburgh Healthcare System) and Leslie R.M. Hausmann, PhD (Center for Healthcare Evaluation, Research, and Promotion, Veterans Affairs Pittsburgh Healthcare System).

The contents do not represent the views of the US Department of Veterans Affairs, the United States Government, or any other organizations.

XSL•FO
RenderX

Funding

This work was funded by the US Department of Veterans Affairs Office of Health Equity and supported by National Social Work Program. It was also supported by resources at VA Bedford Healthcare System and George E. Wahlen Medical Center. AJC was additionally supported by CDA 20-037 from the Department of Veterans Affairs Health Systems Research.

Data Availability

The authors could provide a data dictionary and/or definitions for any included variables for their analyses, although they will not be able to include individual patient data, even if blinded or anonymized, due to privacy concerns.

Authors' Contributions

Conceptualization: SML, LER, AJC, MAK
Data curation: SML, LER, JB, AJC, CWH, MAK, KMM, SCR
Formal analysis: LER, SML, AJC, MAK, SCR
Funding acquisition: AJC, MAK
Investigation: SML, LER, JB, AJC, CWH, MAK, KMM, SCR
Methodology: LER, SML, JB, AJC, CWH, MAK, SCR
Project administration: LER, SML, JB, CWH, KMM, SCR
Resources: SML, LER, AB, JB, AJC, CWH, MAK, KMM, EM, SCR, CS, MCW
Software: SML, JB, CWH, CS
Supervision: EM, MAK
Validation: SML, LER, AJC, MAK, SCR
Visualization: SML, LER, AB, JB, AJC, CWH, MAK, KMM, EM, SCR, CS, MCW
Writing – original draft: LER, SML, JB, AJC, MAK
Writing – review & editing: LER, SML, AB, JB, AJC, CWH, MAK, KMM, EM, SCR, CS, MCW

Conflicts of Interest

None declared.

Multimedia Appendix 1
Assessing Circumstances and Offering Resources for Needs (ACORN) Dashboard development: phase 1 usability questions.
[DOCX File, 15 KB - jmir_v28i1e81846_app1.docx]

Multimedia Appendix 2
Assessing Circumstances and Offering Resources for Needs (ACORN) User Guide and Frequently Asked Questions.
[DOCX File, 4826 KB - jmir_v28i1e81846_app2.docx]

References

1. Janssen A, Donnelly C, Kay J, et al. Developing an intranet-based lymphedema dashboard for breast cancer multidisciplinary teams: design research study. J Med Internet Res 2020 Apr 21;22(4):e13188. [doi: 10.2196/13188] [Medline: 32314968]
2. Callaghan-Koru JA, Newman Chargois P, Tiwari T, Brown CC, Greenfield W, Koru G. Public maternal health dashboards in the United States: descriptive assessment. J Med Internet Res 2024 Sep 17;26:e56804. [doi: 10.2196/56804] [Medline: 39288409]
3. Smith VS. Data dashboard as evaluation and research communication tool. New Drctns Evaluation 2013 Dec;2013(140):21-45. [doi: 10.1002/ev.20072]
4. O'Brien N, et al. Addressing racial and ethnic inequities in data-driven health technologies. : Institute of Global Health Innovation URL: https://spiral.imperial.ac.uk/server/api/core/bitstreams/f1dd3e4d-1a51-462f-9a10-3e919d114057/content [accessed 2025-08-01]
5. Cascini F, Santaroni F, Lanzetti R, Failla G, Gentili A, Ricciardi W. Developing a data-driven approach in order to improve the safety and quality of patient care. Front Public Health 2021;9:667819. [doi: 10.3389/fpubh.2021.667819] [Medline: 34095071]
6. Our data-driven future in healthcare: people and partnerships at the heart of health related technologies. : The Academy of Medical Sciences; 2018 Nov URL: https://acmedsci.ac.uk/file-download/74634438 [accessed 2025-08-01]
7. Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, Institute of Medicine. Capturing Social and Behavioral Domains and Measures in Electronic Health Records: The National Academies Press; 2014. [doi: 10.17226/18951]
8. Alley DE, Asomugha CN, Conway PH, Sanghavi DM. Accountable health communities--addressing social needs through Medicare and Medicaid. N Engl J Med 2016 Jan 7;374(1):8-11. [doi: 10.1056/NEJMp1512532] [Medline: 26731305]

XSL•FO
RenderX

9. Gusoff G, Fichtenberg C, Gottlieb LM. Professional medical association policy statements on social health assessments and interventions. TPJ 2018 Dec;22(4S):22. [doi: 10.7812/TPP/18-092]
10. National Academies of Sciences, Engineering, and Medicine. Integrating Social Care into the Delivery of Health Care: Moving Upstream to Improve the Nation's Health: The National Academies Press. URL: https://www.ncbi.nlm.nih.gov/books/NBK552597/ [accessed 2025-08-01]
11. List JM, Russell LE, Hausmann LRM, et al. Addressing veteran health-related social needs: how Joint Commission Standards accelerated integration and expansion of tools and services in the Veterans Health Administration. Jt Comm J Qual Patient Saf 2024 Jan;50(1):34-40. [doi: 10.1016/j.jcjq.2023.10.002] [Medline: 37923670]
12. Social drivers of health and health-related social needs. Centers for Medicare & Medicaid Services. 2024. URL: https://www.cms.gov/priorities-innovation-key-concepts-social-drivers-health-health-related-social-needs [accessed 2025-08-01]
13. Wang M, Pantell MS, Gottlieb LM, Adler-Milstein J. Documentation and review of social determinants of health data in the EHR: measures and associated insights. J Am Med Inform Assoc 2021 Nov 25;28(12):2608-2616. [doi: 10.1093/jamia/ocab194] [Medline: 34549294]
14. Buitron de la Vega P, Losi S, Sprague Martinez L, et al. Implementing an EHR-based screening and referral system to address social determinants of health in primary care. Med Care 2019;57(Suppl 2):S133-S139. [doi: 10.1097/MLR.0000000000001029] [Medline: 31517789]
15. Montgomery AE, Fargo JD, Byrne TH, Kane VR, Culhane DP. Universal screening for homelessness and risk for homelessness in the Veterans Health Administration. Am J Public Health 2013 Dec;103 Suppl 2(Suppl 2):S210-S211. [doi: 10.2105/AJPH.2013.301398] [Medline: 24148032]
16. Cohen AJ, Rudolph JL, Thomas KS, et al. Food insecurity among veterans: resources to screen and intervene. Fed Pract 2020 Jan;37(1):16-23. [Medline: 32047351]
17. Russell LE, Cornell PY, Halladay CW, et al. Sociodemographic and clinical characteristics associated with veterans' digital needs. JAMA Netw Open 2024 Nov 4;7(11):e2445327. [doi: 10.1001/jamanetworkopen.2024.45327] [Medline: 39546310]
18. VA homeless programs – Permanent housing assistance for veterans. United States Department of Veterans Affairs. 2021 Dec 14. URL: https://www.va.gov/HOMELESS/housing.asp [accessed 2025-08-01]
19. Wewiorski NJ, Gorman JA, Scoglio AAJ, et al. Promising practices in vocational services for the community reintegration of returning veterans: the individual placement and support model and beyond. Psychol Serv 2018 May;15(2):191-199. [doi: 10.1037/ser0000177] [Medline: 29723021]
20. Chinman M, Henze K, Sweeney P. Peer specialist toolkit: implementing peer support services in VHA. : U.S. Department of Veterans Affairs URL: https://www.mirecc.va.gov/visn4/docs/Peer_Specialist_Toolkit_FINAL.pdf [accessed 2025-08-01]
21. Russell LE, Mitchell KM, Kennedy MA, et al. Building tailored resource guides to address social risks and advance health equity in the Veterans Health Administration. Fed Pract 2024 Jan;41(1):22-28. [doi: 10.12788/fp.0446] [Medline: 38835360]
22. Russell LE, Cohen AJ, Chrzas S, et al. Implementing a social needs screening and referral program among veterans: Assessing Circumstances & Offering Resources for Needs (ACORN). J Gen Intern Med 2023 Oct;38(13):2906-2913. [doi: 10.1007/s11606-023-08181-9] [Medline: 37165261]
23. Cohen AJ, et al. The Assessing Circumstances and Offering Resources for Needs (ACORN) initiative. : Veterans Health Administration; 2022 URL: https://www.va.gov/HEALTHEQUITY/docs/ACORN_Screening_Tool.pdf [accessed 2025-08-01]
24. Russell LE, et al. Assessing Circumstances and Offering Resources for Needs (ACORN) – fiscal year 2023 annual report. : Department of Veterans Affairs; 2023 URL: https://www.va.gov/HEALTHEQUITY/docs/ACORN_Annual_Report_FY2023.pdf [accessed 2025-08-01]
25. Dowding D, Randell R, Gardner P, et al. Dashboards for improving patient care: review of the literature. Int J Med Inform 2015 Feb;84(2):87-100. [doi: 10.1016/j.ijmedinf.2014.10.001] [Medline: 25453274]
26. Lau MK, Bounthavong M, Kay CL, Harvey MA, Christopher MLD. Clinical dashboard development and use for academic detailing in the U.S. Department of Veterans Affairs. J Am Pharm Assoc (2003) 2019 Mar;59(2):S96-S103. [doi: 10.1016/j.japh.2018.12.006]
27. Timmerman L, Dutton H, McDannald N, Smitherman EA, Mannion ML. Electronic health record modification and dashboard development to improve clinical care in pediatric rheumatology. Front Pediatr 2024;12:1428792. [doi: 10.3389/fped.2024.1428792] [Medline: 39192884]
28. Chang RC, Thomas K, Penaia CS, Ponce NA. Building the NHPI Data Policy Platform to identify the most pressing SDOH policy domains for the NHPI community. J Health Care Poor Underserved 2024;35(3S):16-26. [Medline: 39069925]
29. Adediran E, Owens R, Gardner E, et al. Development and usability of an EHR-driven hypertension disparities dashboard in primary care: a qualitative study. J Clin Hypertens (Greenwich) 2024 Jul;26(7):797-805. [doi: 10.1111/jch.14834] [Medline: 38850400]
30. Hausmann LRM, Lamorte C, Estock JL. Understanding the context for incorporating equity into quality improvement throughout a national health care system. Health Equity 2023;7(1):312-320. [doi: 10.1089/heq.2023.0009] [Medline: 37284535]
31. Wang L, Porter B, Maynard C, et al. Predicting risk of hospitalization or death among patients receiving primary care in the Veterans Health Administration. Med Care 2013 Apr;51(4):368-373. [doi: 10.1097/MLR.0b013e31827da95a] [Medline: 23269113]

XSL•FO
RenderX

32. German JC, Stirling A, Gorgone P, et al. Interactive visualization tool to understand and monitor health disparities in diabetes care and outcomes. J Clin Transl Sci 2024;8(1):e102. [doi: 10.1017/cts.2024.542] [Medline: 39220819]
33. Jonnalagadda P, Swoboda C, Singh P, et al. Developing dashboards to address children's health disparities in Ohio. Appl Clin Inform 2022 Jan;13(1):100-112. [doi: 10.1055/s-0041-1741482] [Medline: 35081656]
34. Bucalon B, Shaw T, Brown K, Kay J. State-of-the-art dashboards on clinical indicator data to support reflection on practice: scoping review. JMIR Med Inform 2022 Feb 14;10(2):e32695. [doi: 10.2196/32695] [Medline: 35156928]
35. Robinson TE, Janssen A, Harnett P, et al. Embedding continuous quality improvement processes in multidisciplinary teams in cancer care: exploring the boundaries between quality and implementation science. Aust Health Rev 2017 Jul;41(3):291-296. [doi: 10.1071/AH16052] [Medline: 27372543]
36. van de Baan FC, Lambregts S, Bergman E, Most J, Westra D. Involving health professionals in the development of quality and safety dashboards: qualitative study. J Med Internet Res 2023 Jun 12;25:e42649. [doi: 10.2196/42649] [Medline: 37307058]
37. Steenkamp LP, Hagedorn-Hansen D, Oosthuizen GA. Visual management system to manage manufacturing resources. Procedia Manufacturing 2017;8:455-462. [doi: 10.1016/j.promfg.2017.02.058]
38. Taxter A, Johnson L, Tabussi D, et al. Co-design of an electronic dashboard to support the coproduction of care in pediatric rheumatic disease: human-centered design and usability testing. J Particip Med 2022 Apr 22;14(1):e34735. [doi: 10.2196/34735] [Medline: 35133283]
39. Morken V, Perry LM, Coughlin A, et al. Implementation of a co-design strategy to develop a dashboard to support shared decision making in advanced cancer and chronic kidney disease. J Clin Med 2024 Jul 17;13(14):4178. [doi: 10.3390/jcm13144178] [Medline: 39064218]
40. Sligo J, Gauld R, Roberts V, Villa L. A literature review for large-scale health information system project planning, implementation and evaluation. Int J Med Inform 2017 Jan;97:86-97. [doi: 10.1016/j.ijmedinf.2016.09.007] [Medline: 27919399]
41. Slattery P, Saeri AK, Bragge P. Research co-design in health: a rapid overview of reviews. Health Res Policy Syst 2020 Feb 11;18(1):17. [doi: 10.1186/s12961-020-0528-9] [Medline: 32046728]
42. Grindell C, Coates E, Croot L, O'Cathain A. The use of co-production, co-design and co-creation to mobilise knowledge in the management of health conditions: a systematic review. BMC Health Serv Res 2022 Jul 7;22(1):877. [doi: 10.1186/s12913-022-08079-y] [Medline: 35799251]
43. Ergonomics of human-system interaction: part 210: human-centred design for interactive systems, 2 ed. ISO. 2019. URL: https://www.iso.org/standard/77520.html [accessed 2025-08-01]
44. Moy E, Hausmann LRM, Clancy CM. From HRO to HERO: making health equity a core system capability. Am J Med Qual 2022;37(1):81-83. [doi: 10.1097/JMQ.0000000000000020] [Medline: 34506333]
45. Scott VC, Jillani Z, Malpert A, Kolodny-Goetz J, Wandersman A. A scoping review of the evaluation and effectiveness of technical assistance. Implement Sci Commun 2022 Jun 28;3(1):70. [doi: 10.1186/s43058-022-00314-1] [Medline: 35765107]
46. Chin MH, Clarke AR, Nocon RS, et al. A roadmap and best practices for organizations to reduce racial and ethnic disparities in health care. J Gen Intern Med 2012 Aug;27(8):992-1000. [doi: 10.1007/s11606-012-2082-9] [Medline: 22798211]
47. Fichtenberg C, Delva J, Minyard K, Gottlieb LM. Health and human services integration: generating sustained health and equity improvements. Health Aff (Millwood) 2020 Apr;39(4):567-573. [doi: 10.1377/hlthaff.2019.01594] [Medline: 32250685]
48. Cornell PY, Halladay CW, Ader J, et al. Embedding social workers in Veterans Health Administration primary care teams reduces emergency department visits. Health Aff (Millwood) 2020 Apr;39(4):603-612. [doi: 10.1377/hlthaff.2019.01589] [Medline: 32250673]
49. Livingston CJ, Titus TM, Yerokun TA, Patel NA. Screening for health-related social needs: American College of Preventive Medicine's practice statement. Am J Prev Med 2025 May;68(5):1041-1049. [doi: 10.1016/j.amepre.2025.01.005] [Medline: 39793769]
50. Murray GF, Rodriguez HP, Lewis VA. Upstream with a small paddle: how ACOs are working against the current to meet patients' social needs. Health Aff (Millwood) 2020 Feb;39(2):199-206. [doi: 10.1377/hlthaff.2019.01266] [Medline: 32011930]
51. Bretsch JK, Wallace AS, McCoy R. Social needs screening in academic health systems: a landscape assessment. Popul Health Manag 2024 Jun;27(5):312-319. [doi: 10.1089/pop.2024.0111] [Medline: 39069945]
52. Artiga S, Hinton E. Beyond health care: the role of social determinants in promoting health and health equity. KFF. 2018. URL: https://www.kff.org/racial-equity-and-health-policy/beyond-health-care-the-role-of-social-determinants-in-promoting-health-and-health-equity [accessed 2025-08-01]
53. Kreuter MW, Thompson T, McQueen A, Garg R. Addressing social needs in health care settings: evidence, challenges, and opportunities for public health. Annu Rev Public Health 2021 Apr 1;42:329-344. [doi: 10.1146/annurev-publhealth-090419-102204] [Medline: 33326298]
54. National Performance Goals (Effective January 2026 for the Hospital Program). : Joint Commission; 2025 Sep URL: https://digitalassets.jointcommission.org/api/public/content/9ca80055182b4274842a5780a94f2c82?v=149a13a9&_gl=1*klh3z5*_gcl_au*NDQ3MTQ4NDIxLjE3NzI2NDk4MDE.

XSL•FO
RenderX

*_ga*MTkzNDY1Mjk5MS4xNzcyNjQ5ODAy*_ga_K31T0BHP4T*czE3NzM4NDI1MDYkbzQkZzEkdDE3NzM4NDM4ODYkajYwJGwwJGgw [accessed 2026-03-20]

Abbreviations

ACORN: Assessing Circumstances and Offering Resources for Needs
EHR: electronic health record
FAQ: frequently asked question
PHI: protected health information
QI: quality improvement
VA: US Department of Veterans Affairs
VAMC: Veterans Affairs Medical Center

Edited by A Stone; submitted 25.Aug.2025; peer-reviewed by C Turvey, D Fuller; accepted 25.Feb.2026; published 13.Apr.2026.

Please cite as:
Russell LE, Leder SM, Boris J, Cohen AJ, Halladay CW, Mitchell KM, Ruggles SC, Berkheimer A, Sonnier C, Wilcox MC, Moy E, Kennedy MA
Embracing Co-Design and Interprofessional Teamwork to Build an Innovative Dashboard for a National Social Needs Screening and Referral Clinical Intervention in the Veterans Health Administration: Design and Development Study
J Med Internet Res 2026;28:e81846
URL: https://www.jmir.org/2026/1/e81846
doi:10.2196/81846

© Lauren E Russell, Sarah M Leder, Jaime Boris, Alicia J Cohen, Christopher W Halladay, Kathleen M Mitchell, Sydney C Ruggles, Andrea Berkheimer, Chava Sonnier, Michelle C Wilcox, Ernest Moy, Meaghan A Kennedy. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Developing and Validating a Machine Learning Algorithm to Predict the Risk of Incident Opioid Use Disorder Among OneFlorida+ Patients: Prognostic Modeling Study

Jabed Al Faysal[1,2], BSc, MSc; Weihsuan Lo-Ciganic[3,4,5], PhD; Walid F Gellad[3,4], MD; Yonghui Wu[6], PhD; Christopher A Harle[7,8], PhD; Khoa Nguyen[9], PharmD; James L Huang[1], PhD; Gerald Cochran[10], PhD; Debbie L Wilson[1], PhD; Stephanie AS Staras[6], PhD; Siegfried OF Schmidt[11], MD, PhD; Eric I Rosenberg[1,12], MD, MSPH; Danielle Nelson[11], MD; Shunhua Yan[1], MEd; Gary M Reisfield[1], MD; William M Greene[13], MD; Courtney Kuza[4], PhD; Md Mahmudul Hasan[1,14], PhD

[1]Department of Pharmaceutical Outcomes & Policy, University of Florida, 1889 Museum Road, Malachowsky Hall, Suite 6300, Gainesville, FL, United States

[10]Department of Internal Medicine, Division of Epidemiology, University of Utah, Salt Lake City, UT, United States

[11]Department of Community Health and Family Medicine, College of Medicine, University of Florida, Gainesville, FL, United States

[12]Department of Internal Medicine, College of Medicine, University of Florida, Gainesville, FL, United States

[13]Department of Psychiatry, College of Medicine, University of Florida, Gainesville, FL, United States

[14]Department of Information Systems and Operations Management, Warrington College of Business, University of Florida, Gainesville, FL, United States

[2]Computer Science and Engineering Discipline, Khulna University, Khulna, Bangladesh

[3]Division of General Internal Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA, United States

[4]Center for Pharmaceutical Policy and Prescribing, University of Pittsburgh, Pittsburgh, PA, United States

[5]Geriatric Research Education and Clinical Center, North Florida/South Georgia Veterans Health System, Gainesville, FL, United States

[6]Department of Health Outcomes & Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, United States

[7]Department of Health Policy and Management, School of Public Health, Indiana University, Indianapolis, IN, United States

[8]Regenstrief Institute, Indianapolis, IN, United States

[9]Department of Pharmacotherapy and Translational Research, College of Pharmacy, University of Florida, Gainesville, FL, United States

Corresponding Author:
Md Mahmudul Hasan, PhD
Department of Pharmaceutical Outcomes & Policy, University of Florida, 1889 Museum Road, Malachowsky Hall, Suite 6300, Gainesville, FL, United States

Abstract

Background: Opioid use disorder (OUD) remains a critical public health crisis in the United States. Despite widespread policy and clinical interventions, early identification of individuals at risk for developing OUD remains challenging due to limitations in traditional screening approaches and a lack of individualized risk stratification methods. Machine learning (ML) methods offer an opportunity to develop timely, high-performing, and explainable predictive models that can enhance OUD prevention strategies in clinical settings.

Objective: This study aims to develop and validate an ML model using electronic health record (EHR) data to predict the 3-month risk of incident OUD among adults initiating opioid therapy and to stratify patients into clinically actionable risk groups.

Methods: This prognostic modeling study used 2017 - 2022 OneFlorida+ EHR data to develop and validate ML algorithms predicting 3-month incident OUD risk. We included 182,083 adults (≥18 y) without cancer, overdose, or OUD or hospice history who received ≥1 outpatient, noninjectable opioid prescription. Using 183 predictors measured in sequential 3-month intervals, we developed an elastic net, least absolute shrinkage and selection operator, gradient boosting machine (GBM), and random forest models on randomly split training, testing, and validation sets. Model performance was assessed using C-statistics, predictive values, and number needed to evaluate, with patients stratified into risk deciles for clinical applicability. Model explainability was assessed using Shapley additive explanations, and fairness was evaluated using standard metrics. We externally validated the best-performing model using an independent cohort from the 2018 - 2020 UPMC (formerly University of Pittsburgh Medical Center) health system.

Results: In the validation sample (n=60,694), GBM (C-statistics=0.879, 95% CI 0.874 - 0.884) and elastic net (C-statistics=0.872, 95% CI 0.867 - 0.877) outperformed least absolute shrinkage and selection operator (C-statistics=0.846, 95% CI 0.840 - 0.851)

XSL•FO
RenderX

and random forest (C-statistics=0.798, 95% CI 0.792 - 0.804), with GBM model requiring the fewest predictors (n=75) for predicting 3-month incident OUD. Using the GBM algorithm to predict the subsequent 3-month OUD risk, the top decile subgroup had a positive predictive value of 3.26%, a negative predictive value of 99.8%, and a number needed to evaluate of 31. The top decile (n=6696) captured ~68% of patients with OUD. Shapley additive explanations analysis identified age, number of outpatient visits, history of back and other pain conditions, comorbidity burden, and opioid prescribing patterns as the strongest predictors of incident OUD. Fairness assessment showed an acceptable false negative rate parity across race, age, and sex. In external validation on the UPMC cohort, the GBM model maintained good discrimination (C-statistics=0.756, 95% CI 0.750 - 0.762) and effective risk stratification.

Conclusions: An ML algorithm predicting incident OUD derived from OneFlorida+ EHR data performed well in external validation with data using UPMC. The algorithm might be valuable for incident OUD risk prediction and stratification across health systems, with potential to inform early intervention.

(J Med Internet Res 2026;28:e79482) doi:10.2196/79482

KEYWORDS

opioid use disorder; machine learning; OneFlorida+; risk stratification; external validation

Introduction

The United States continues to face a persistent and evolving opioid epidemic. Opioid use disorder (OUD) and overdose affect millions of Americans, leading to increased morbidity, mortality, and health care costs [1,2] In 2022, more than 6 million individuals experienced OUD, imposing significant societal and economic burdens exceeding US $78 billion annually [1,3,4]. Opioid-related deaths increased tenfold from 1999 to 2022 (>82,000 in 2022) [5-8]. Prescription opioids were responsible for approximately 280,000 overdose deaths over the same period [9]. The prevalence of OUD continues to rise despite efforts to curb opioid misuse [10,11]. In response, health systems, payers, and policymakers have implemented various interventions aimed at reducing unsafe prescribing and patient risk [12].

Early identification of OUD can prevent severe addiction, high-risk behaviors, overdose, and death, while enabling timely access to evidence-based treatments such as buprenorphine or referral to recovery services. Existing methods for identifying high-risk opioid users, such as high-dose opioid prescribing thresholds and use of multiple pharmacies, are often based on simple or single criteria [13,14]. These rule-based strategies leave many high-risk individuals undetected and may misclassify others, resulting in unintended consequences like delayed care, unnecessary monitoring, or stigma. For example, a study found that nearly 43% of individuals who developed OUD were missed by such approaches [15]. Furthermore, traditional statistical models emphasize population-level risk factors and assume linear relationships between predictors and outcomes, limiting their ability to capture complex interactions among patient demographics, comorbidities, health care use, and prescriber patterns. Although tools such as the Screener and Opioid Assessment for Patients with Pain-Revised and the Opioid Risk Tool are commonly used in clinical settings, studies report inconsistent estimates of test accuracy and no evidence that they reduce OUD or overdose [16-18]. The 2022 Centers for Disease Control and Prevention Clinical Practice Guideline similarly concluded that current opioid risk screening tools lack proven clinical use. Proprietary platforms such as NarxCare leverage prescription and health record data but lack transparency, public validation, and demonstrated generalizability [19]. These limitations underscore the need for more robust, transparent, and generalizable models to support real-time clinical decision-making across diverse populations. Studies have also highlighted these shortcomings and called for more advanced, data-driven models to improve identification of individuals at risk (or low risk) of OUD [20-24].

Machine learning (ML) offers a promising alternative to traditional statistical methods in improving risk prediction. ML models can analyze large health care data to identify complex nonlinear patterns in patient data. Recent studies have demonstrated the potential of ML to predict opioid-related outcomes, including OUD and overdose, with higher accuracy than that of traditional methods [25-28]. Our previous work showed that ML approaches can improve risk prediction and stratification for incident OUD and subsequent overdose in Medicare and Medicaid beneficiaries. For example, we developed ML models to predict incident OUD among Medicare beneficiaries, achieving high discrimination (C-statistic >0.86) and effectively stratifying patients into clinically meaningful risk subgroups (eg, low, moderate, and high risk for developing OUD) [29]. However, many existing models are developed using Medicare or Medicaid data, which primarily include older adults or low-income populations.

In this study, we extend our work to develop and validate an ML algorithm to predict incident OUD among patients who received care from institutions within the OneFlorida+ network—a large, diverse patient population across Florida, Georgia, Alabama, and Arkansas, including individuals with commercial insurance, Medicaid, or Medicare, and those with no insurance [30]. This study advances our prior Medicare-based work in several important ways. First, although claims-based models can be incorporated into some clinical workflows, electronic health records (EHRs) often lack the full use and medication detail available in claims and may vary in data quality. As such, adapting the model for EHR-based use required methodological refinement to support real-time, point-of-care integration. Second, the broader and more heterogeneous cohort enhances the model's generalizability beyond the older, fee-for-service Medicare population used previously. Third, we externally validated the model using an independent cohort

XSL•FO
RenderX

from the UPMC health system, assessing cross-system generalizability and transportability—an important step toward real-world implementation. Fourth, we incorporated fairness and bias analyses across demographic subgroups to address equity considerations often overlooked in prior OUD risk prediction studies. Finally, we articulate a translational vision for integrating this model into clinical workflows. By organizing predictors to enable risk-based stratification, the model is designed to function as an EHR-embedded decision support tool in primary care settings, alerting providers (eg, physicians) when a patient is flagged as high-risk to prompt timely, preventive actions. We emphasize model interpretability to ensure that our findings can be seamlessly integrated into the clinical environment.

Methods

Study Design and Data Source

This is a prognostic modeling study with a retrospective cohort design. We used the TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + Artificial Intelligence) guidelines for reporting our work (Checklist 1) [31]. This study used data from OneFlorida+, a secure and centralized repository of patient - level data from both health system partners and insurers, and managed by the OneFlorida Clinical Research Consortium [32]. Currently, OneFlorida+ includes EHRs and claims data from 26 million individuals across Florida, Georgia, Alabama, and Arkansas, covering a wide range of populations and health care settings [30,33]. All data sources are mapped to the Patient - Centered Outcomes Research network common data model (version 7.0) to ensure standardization of data elements across sources. Major data elements in the common data model include demographics, enrollment, encounters, diagnoses, procedures, dispensed medications, and deaths. Additionally, while OneFlorida data includes Florida Medicaid recipients with claims data only, our study cohort requires having at least 1 EHR encounter. The study population included adult patients (aged ≥18 y) who had at least 1 opioid prescription filled between 2017 and 2022 (Figure S1 in Multimedia Appendix 1). The index date was defined as the date of a patient's first opioid prescription between October 1, 2017, and September 30, 2022. To construct a cohort appropriate for assessing incident OUD risk, we excluded patients who (1) had cancer diagnoses (Table S1 in Multimedia Appendix 2), (2) received hospice care, (3) had a diagnosis of OUD, opioid overdose, other substance use disorders, drug abuse, or received methadone or buprenorphine for OUD before initiating opioids, or (4) did not have at least a 3-month observation window following the first opioid prescription to allow for predictor measurement. We also excluded patients with a diagnosis of other substance use disorders to minimize confounding, as some physicians may have used this diagnosis for patients with co-occurring OUD and other substance use conditions. Once eligible, patients remained in the cohort until they experienced an outcome of interest (ie, incident of OUD diagnosis) or were censored due to death, regardless of continued opioid use (Figure S2 in Multimedia Appendix 1).

For external validation, we applied the algorithm to 2018 - 2020 EHR data from the UPMC health system in Pennsylvania, a region with demographic and opioid prescribing characteristics distinct from the OneFlorida+ network. Cohort construction and predictor generation followed the same procedures used for the OneFlorida+ dataset.

Outcome Variable

Similar to other claims-based analyses [29,34-36], our primary outcome was incident OUD (Table S2 in Multimedia Appendix 2), defined as the first recorded diagnosis of OUD from all settings in each 3-month window after the index prescription date (Figure S2 in Multimedia Appendix 1). The *ICD-10* (*International Statistical Classification of Diseases and Related Health Problems 10th Revision*) codes for incident OUD diagnosis include F11.1X and F11.2X but exclude F11.11 (opioid-related disorders in remission) and F11.21 (opioid dependence in remission).

Predictor Candidates

We identified 183 candidate predictors of OUD from the prior literature (Table S3 in Multimedia Appendix 2) [37-48]. These predictors included sociodemographic factors, patient health status, and patterns of opioid use and other nonopioid prescriptions measured at baseline (during the 3 month period before the first opioid fill) and in 3-month windows after initiating prescription opioids. This 3-month window was selected for both clinical and operational relevance. Evidence shows that early indicators of problematic opioid use (eg, dose escalation, early refills, or polypharmacy) typically emerge within the first 3 months of therapy [49-52]. Pharmacoepidemiologic studies demonstrate that transitions to sustained or high-risk opioid use most commonly occur within 3 months after treatment initiation [53,54]. This timeframe also aligns with the quarterly monitoring cycles used by many prescription drug monitoring programs and health plans [14,38]. We updated the predictors measured in each 3-month period to account for changes over time for predicting incident OUD risks in each subsequent period (Figure S2 in Multimedia Appendix 1). This time-updating approach mimics active surveillance that a health system might conduct in real time [55] to provide clinicians with timely, actionable opportunities (eg, closer follow-up, patient counseling, or medication adjustments) before the progression to OUD. We examined missingness across candidate predictors and applied prespecified imputation rules before model training. We used simple imputation techniques where missing values for all categorical predictors (including race, ethnicity, and provider sex) were imputed using the modal category. Continuous predictors (eg, use counts, comorbidity scores, and medication counts) with missing values were imputed using the median calculated from the full analytic dataset. Binary and count predictors representing the presence or frequency of diagnoses, procedures, or medications were imputed as zero when missing, consistent with the interpretation that the absence of a code reflects no recorded event during the baseline window.

XSL•FO
RenderX

ML Approaches and Prediction Performance Evaluation

Our machine learning analysis using OneFlorida+ data had two objectives: (1) developing a prediction model to generate individuals' incident OUD risk scores and (2) stratifying individuals into subgroups based on similar OUD risk levels. We developed and tested 4 ML algorithms to predict incident OUD: elastic net (EN), least absolute shrinkage and selection operator (LASSO), gradient boosting machine (GBM), random forest (RF). These algorithms have consistently demonstrated strong predictive performance in prior studies [12,46,56,57]. Details of each algorithm are provided in Multimedia Appendix 3. We randomly divided the cohort into 3 subsets: a training sample for algorithm development, a testing sample for algorithm refinement, and a validation sample for evaluating predictive performance. This split was conducted strictly by patient ID, such that all 3-month episodes from a given patient were assigned to a single dataset (training, testing, or validation), and no patient contributed episodes to more than one set, thereby avoiding information leakage across datasets. Because patients could contribute multiple 3-month episodes, episode-level observations within patients were correlated by design. Models were trained and evaluated at the episode level to reflect the intended real-world use case of repeated, longitudinal risk assessment over time and to maximize learning from rare outcomes. Tree-based ensemble methods such as GBMs are generally robust to correlated observations because they do not rely on independence assumptions and learn predictive structure through recursive partitioning rather than parametric estimation. To evaluate the potential impact of within-patient correlation on model performance estimates, we conducted sensitivity analyses using patient-level random subsets. We used one 3-month period with predictor measurements to forecast risk in the subsequent 3 months within the validation set.

To evaluate discrimination performance, we assessed whether individuals predicted to be at high risk had higher OUD incidence than those predicted to be at low risk. We compared C-statistics across different methods in the validation sample using the DeLong test [58]. A C-statistic between 0.7 and 0.8 indicates good discrimination, while values above 0.8 indicate very good discrimination. Precision-recall curves were also examined [42]. Given that OUD events are rare outcomes and C-statistics do not incorporate information about outcome incidence, we reported other metrics, including sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), number needed to evaluate (NNE) to identify 1 OUD episode, and estimated rate of alerts, to thoroughly assess the algorithms' predictive ability (Figure S3 in Multimedia Appendix 1) [43,59].

There is no universally applicable prediction probability threshold, so we evaluated performance across multiple sensitivity and specificity levels (eg, selecting 90% sensitivity). To determine an optimized threshold that balances sensitivity and specificity, we applied the Youden index in the training sample [60]. In the validation sample, patients were stratified into risk subgroups based on deciles of predicted incident OUD probability. The highest decile was further divided into 3 strata: the top first percentile, the second to fifth percentiles, and the sixth to tenth percentiles. This approach allowed for a more detailed assessment of individuals at the highest risk of developing OUD.

We conducted additional analyses to enhance clinical use. The primary goal of our ML algorithm was to generate risk scores for incident OUD. For comprehensive model explanation and further enhancement of model interpretability, we conducted a Shapley additive explanations (SHAP) analysis [61]. SHAP quantifies the contribution of each predictor to the model's output while accounting for interactions with other features. The SHAP plots indicate whether the mentioned features have a positive correlation (red color) or a negative correlation (blue color) with OUD. To evaluate bias across demographic factors (eg, race, age, and sex), we examined risk score distributions and false positive rates (FPR) and false negative rates (FNR) in the best performing model. We defined significant disparity as a parity ratio between subgroups falling outside the range of 0.80 to 1.25, consistent with the standards used in algorithmic fairness audits [62].

Additionally, we evaluated model calibration to assess the reliability of predicted risk scores. Given the potential for miscalibration in rare-outcome ML models, we applied isotonic regression to recalibrate predicted risk scores [63]. The overall accuracy of probabilistic predictions was assessed using the Brier score metric [64]. To determine clinical use beyond discrimination, we performed decision curve analysis. We calculated the net benefit [65] across a range of clinically relevant risk thresholds, comparing the recalibrated model against default strategies of "treat all" and "treat none". This approach quantifies the trade-off between the benefit of identifying true OUD cases and the potential harm of false-positive alerts.

Statistical Analysis

All analyses were performed using SAS 9.4 (SAS Institute Inc) and Python 3.6 (Python Software Foundation). Continuous variables were summarized using mean (SD) values, while categorical variables were summarized using frequencies and percentages. We compared patients' characteristics in training, testing, and validation datasets using 2-tailed Student *t* tests, chi-square tests, and ANOVA, or corresponding nonparametric tests.

Ethical Considerations

This study was reviewed and approved by the University of Florida (UF) Institutional Review Board (IRB202101897). The research involved secondary analysis of limited electronic health record data from the OneFlorida+ Data Trust. In accordance with federal regulations (45 CFR 46.104[d] [4]) and institutional policies, the requirement for informed consent was waived because the study involved no direct participant contact and posed minimal risk to individuals. All analyses were conducted within the secure UF server environment; no identifiable private information was accessed or transmitted outside the system. The study adhered to institutional data-use agreements and complied with all relevant guidelines for human subjects research and data privacy. The external validation using UPMC data was reviewed by the University of Pittsburgh Human

XSL•FO
RenderX

Research Protection Office and determined not to constitute human subjects research, as only deidentified data were used. Therefore, institutional review board review and informed consent were not required.

Results

Patient Characteristics

OneFlorida+ patients in the training (n=60,694), testing (n=60,695), and validation (n=60,694) samples had similar characteristics and outcome distributions (mean age 53.02, SD 7.60 y; 111,576/182,083, 61.28% female; 125,208/182,083, 68.76% White; 144,316/182,083, 79.26% non-Hispanic; 4328/182,083, 2.38% had incident OUD; Table 1). The external validation cohort from the UPMC included 129,215 patients (mean age 58.91, SD 16.75 y; 78,574/129,215, 60.81% female; 115,907/129,215, 89.7% White; 123,083/129,215, 95.25% non-Hispanic; 2537/129,215, 1.96% had incident OUD). While sex distributions were similar between the 2 health systems, the UPMC cohort was older and less diverse, with more White and non-Hispanic patients.

Table . Selected characteristics of OneFlorida+ and UPMC[a] health system patients.

Characteristics	2017 - 22 OneFlorida+ data			2018 - 20 UPMC[a] data
	Training (n=60,694)	Testing (n=60,695)	Internal validation (n=60,694)	External validation (n=129,215)
Development of incident opioid use disorder, n (%)	1443 (2.38)	1433 (2.36)	1452 (2.39)	2537 (1.96)
Age (years), mean (SD)	53.11 (17.57)	52.98 (17.64)	52.98 (17.57)	58.91 (16.75)
Age group (years), n (%)				
18 - 34	10,200 (16.81)	10,286 (16.95)	10,285 (16.95)	13,852 (10.73)
35 - 50	15,326 (25.25)	15,501 (25.54)	15,477 (25.5)	23,931 (18.52)
51 - 64	19,121 (31.5)	19,067 (31.41)	19,096 (31.46)	44,047 (34.10)
≥65	16,047 (26.44)	15,841 (26.1)	15,836 (26.09)	47,385 (36.65)
Sex, n (%)				
Male	23,534 (38.77)	23,489 (38.70)	23,481 (38.69)	50,641 (39.19)
Female	37,160 (61.23)	37,206 (61.30)	37,210 (61.31)	78,574 (60.81)
Race, n (%)				
White	41,737 (68.77)	41,751 (68.79)	41,720 (68.74)	115,907 (89.70)
Black	13,403 (22.08)	13,458 (22.17)	13,548 (22.32)	10,833 (8.38)
Other or unknown	5554 (9.15)	5486 (9.04)	5426 (8.94)	2475 (1.92)
Ethnicity, n (%)				
Non-Hispanic	47,951 (79)	48,242 (79.48)	48,123 (79.29)	123,083 (95.25)
Hispanic	10,756 (17.72)	10,455 (17.23)	10,540 (17.37)	1013 (0.78)
Other or unknown	1987 (3.27)	1998 (3.29)	2031 (3.35)	5123 (3.96)

[a]UPMC: Formerly known as University of Pittsburgh Medical Center.

Prediction Performance Across ML Methods

Figure 1 summarizes the 4 prediction performance measures of each model in the internal validation sample. GBM (C-statistics=0.879, 95% CI 0.874 - 0.884) and EN (C-statistics=0.872, 95% CI 0.867 - 0.877) models outperformed the LASSO (C-statistics=0.846, 95% CI 0.840 - 0.851) and RF (C-statistics=0.798, 95% CI 0.792 - 0.804) models, with the GBM model requiring the fewest predictors (n=75) for predicting 3-month incident OUD (P<.001; Figure 1A). The GBM and EN models had similar prediction performance, and the GBM model had the best precision-recall performance (Figure 1B). Sensitivity analyses using patient-level data yielded similar results (Figure S4 in Multimedia Appendix 1). To evaluate model robustness across key clinical subgroups, we stratified the GBM model's performance by age, sex, and ethnicity. The model demonstrated consistent discriminative ability across populations, with C-statistics ranging from 0.833 to 0.909 across age groups (Figure S5 in Multimedia Appendix 1), 0.865 to 0.887 between sexes (Figure S6 in Multimedia Appendix 1), and 0.864 to 0.914 by ethnicity (Figure S7 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 1. Performance matrix across machine learning models for predicting incident opioid use disorder in OneFlorida+ patients. This figure shows 4 prediction performance matrices in the validation sample. (A) The areas under receiver operating characteristic curves (or C-statistics). (B) The precision-recall curves (precision=positive predictive value and recall=sensitivity): precision-recall curves that are closer to the upper right corner or are above another method have improved performance. (C) The number needed to evaluate by different cutoffs of sensitivity. (D) Alerts per 100 patients by different cutoffs of sensitivity. AUC: area under the curve; EN: elastic net; GBM: gradient boosting machine; LASSO: least absolute shrinkage and selection operator; PPV: positive predictive value; RF: random forest; ROC: receiver operating characteristics.

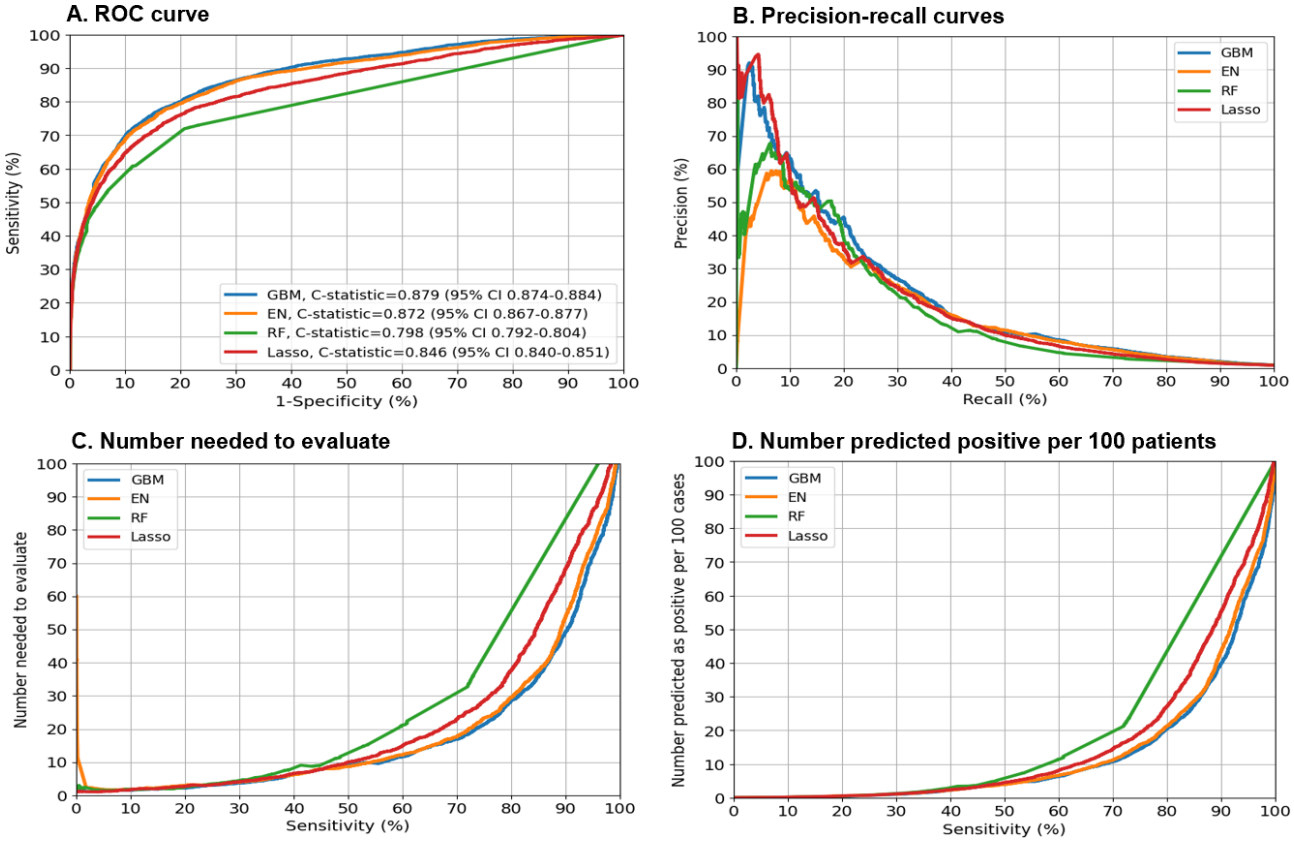

Table S4 in Multimedia Appendix 2 shows the performance measures for predicting incident OUD across different levels (90% - 100%) of sensitivity and specificity for GBM and EN. When set at the optimized sensitivity and specificity as measured by the Youden index, the GBM model had a sensitivity of 76.59%, specificity of 84.75%, PPV of 4.36%, NPV of 99.75%, and NNE of 23; and the EN model had a sensitivity of 77.65%, specificity of 82.74%, PPV of 3.93%, NPV of 99.76%, and NNE of 25 (Figures 1C and 1D; Table S4 in Multimedia Appendix 2).

Risk Stratification, Bias and Fairness Analysis, and Model Explainability

Figure 2 represents the actual OUD rate for individuals in each decile subgroup using GBM. The high-risk subgroup (with risk scores in the top decile; 11% [n=6696] of the validation cohort) had a PPV of 3.26%, a NPV of 99.8%, and NNE of 31. In a hypothetical clinical application involving 1000 patients, applying the top decile threshold would generate alerts for approximately 100 individuals. Based on the observed PPV at this threshold, approximately 4 of the 100 flagged patients would be expected to develop OUD, demonstrating the model's ability to meaningfully concentrate risk among a small subgroup. Conversely, the high NPV (99.8%) indicates that very few incident OUD cases would arise among the 900 unflagged patients, consistent with the model's low false negative rate. Taken together, these estimates illustrate how the model could support prioritization of patients most likely to benefit from closer monitoring or preventive intervention.

XSL•FO
RenderX

Figure 2. Incident opioid use disorder was identified by the gradient boosting machine's decile risk subgroup in the validation sample. Based on the individual's predicted probability of an opioid use disorder event, we classified patients in the internal validation sample into decile risk subgroups, with the highest decile further split into 3 additional strata based on the top first, second to fifth, and sixth to tenth percentiles to allow closer examination of patients at the highest risk of developing opioid use disorder.

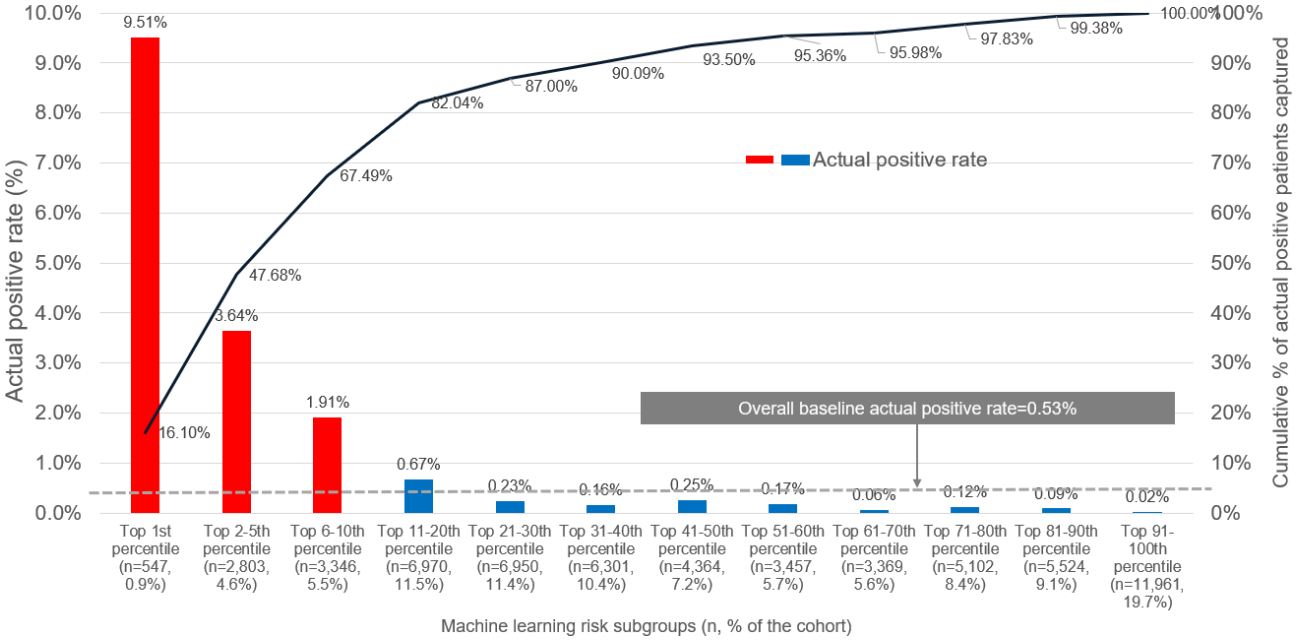

Among all 323 individuals with an incident OUD, 265 (82.04%) occurred in the top two decile subgroups (Decile 1=67.5 % and Decile 2=14.6%). Those in the first decile subgroup had at least a 14-fold higher OUD rate compared to the lower-risk groups (eg, observed OUD rate: Decile 1=9.51%, Decile 2=0.67%, Decile 10=0.02%). The third through tenth decile subgroups had minimal OUD incidence rate (2 to 16 per 10,000).

In our evaluation of racial fairness, we found that among the true positive cases identified in the top decile risk subgroup by the GBM model, 78.44% (171/218) were White individuals and 18.8% (41/218) were Black individuals. In the top first percentile, the distribution was 67.3% (35/52) White individuals and 25% (13/52) Black individuals. The detailed racial distribution of true positives across all risk subgroups is provided in Table S5 in Multimedia Appendix 2. We also compared FNR and FPR by race in the GBM model with and without race as a predictor. FNRs were consistently higher for Black individuals compared to White individuals across most percentiles, while FPR remained similar across racial groups (Figure S8 and Figure S9 in Multimedia Appendix 1). We compared the FNR and FPR between racial groups at the 90th percentile and found that the FNR parity ratio (1.23) and FPR parity ratio (1.03) were within the accepted threshold (0.8 - 1.25) [62], suggesting equitable model performance across racial groups when race was included as a predictor. We also assessed fairness across age and sex. The true positive cases in the top decile were concentrated among the 35 - to 50-year age group (82/218, 37.61%) and females (146/218, 66.97%). The detailed distribution of true positives by age and sex across all risk subgroups is provided in Tables S6 and S7 in Multimedia Appendix 2. FNRs were slightly higher for older individuals (≥65 y) across most percentiles compared to the young adults (18 - 34 y), while FNRs for males and females were closely aligned. FPRs remained consistent across both age groups and sex (Figures S10 and S11 in Multimedia Appendix 1). Parity ratios for sex (FNR=0.91, FPR=1.18) and age-based FPR (0.91; comparing 18 - 34 vs ≥65 y) remained within the acceptable range (0.8 - 1.25), whereas the age-based FNR parity ratio (0.74) was below the threshold.

Figure 3 presents variable importance plots using SHAP values derived from the GBM model. The plots on the right-hand side present the average impact of various features on incident OUD, listed in descending order of their significance. Variables such as age and history of other pain for each 3-month period were negatively correlated with incident OUD, while factors having a positive correlation with incident OUD included the number of outpatient visits, history of back pain, and Elixhauser comorbidity index score.

XSL•FO
RenderX

Figure 3. Variable importance plots using Shapley additive explanations values from the gradient boosting machine model. (A) Shapley additive explanations value computed from individual features' values and their impact (both positive and negative) on incident opioid use disorder. (B) Average Shapley additive explanations values of features showing average impact on and correlation with incident opioid use disorder. SHAP: Shapley additive explanations.

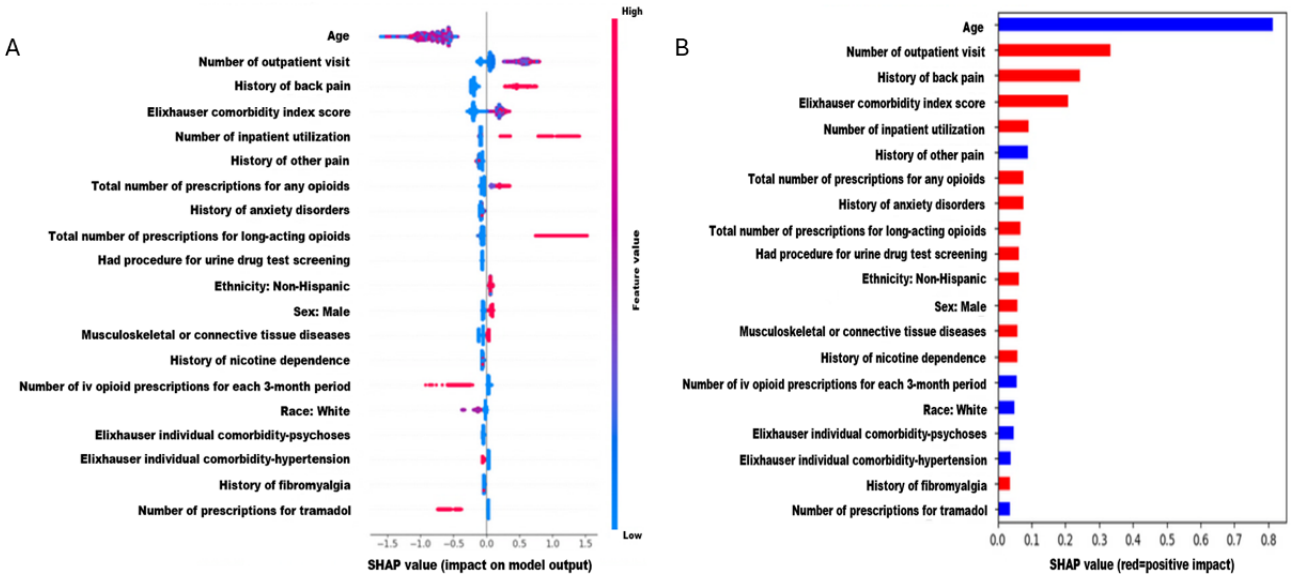

Calibration and Decision Curve Analysis

Calibration assessment showed that the GBM model initially overestimated risk (Brier score=0.09), a known phenomenon in rare-event prediction. After isotonic recalibration, model reliability improved significantly, with a calibration slope of 0.98 (ideal=1), an intercept of –0.03 (ideal=0), and a Brier score of 0.01 (lower is better; Figure S12 in Multimedia Appendix 1). Decision-curve analysis showed that a "treat-all" strategy (intervene on everyone) yields negative net benefit at risk thresholds above 0.5% (Figure S13 in Multimedia Appendix 1), whereas the GBM model maintained positive net benefit across a broader range of clinically relevant thresholds. For example, at the top decile cutoff (probability threshold=0.009), the model achieved a net benefit of 0.0026. This corresponds to identifying approximately 2.6 additional true-positive OUD cases per 1000 patients (after accounting for false-positive alerts) compared with a strategy of no intervention. Relative to the OUD outcome incidence (0.53% in the prediction window), this net benefit reflects ~50% of the maximum achievable use for a perfect classifier in this setting, indicating substantial clinical value despite the low-prevalence context.

External Validation Performance

When applied to the UPMC cohort, the best-performing GBM model achieved a C-statistic of 0.756 (95% CI 0.750 - 0.762; Figure S14 in Multimedia Appendix 1). The high-risk decile subgroup demonstrated a PPV of 2.52%, an NPV of 99.8%, and NNE of 40 (Figure S15 in Multimedia Appendix 1). The top first percentile had an OUD incidence rate 14-fold higher than the overall baseline (8.15% vs 0.58%). Fairness analyses in the UPMC cohort revealed variations in FNR and FPR across age groups (Figure S16 in Multimedia Appendix 1), while error rates remained similar across race (Figure S17 in Multimedia Appendix 1) and between males and females (Figure S18 in Multimedia Appendix 1). For example, using the 90th percentile risk score as the elevated risk cutoff, younger adults (18 - 34 y) had a lower FNR (fewer missed cases) compared to older adults (≥65 y).

Discussion

Principal Findings

This study expanded our previous work using ML approaches to improve the accuracy of predicting incident OUD in the subsequent 3 months of prescription opioid initiation among fee-for-service Medicare beneficiaries and broaden the applicability of these models across a diverse population in the OneFlorida+ network [29]. Our GBM and EN models demonstrated strong predictive performance, achieving a high C-statistic (>0.87) and outperforming LASSO and RF. Our best-performing GBM model offers several advantages, including its ability to eliminate the need for a separate feature selection process and its flexibility in hyperparameter tuning to capture complex interactions between predictors and outcomes. We acknowledge that this flexibility in model tuning can be computationally intensive and time-consuming. The algorithm effectively stratified the population into distinct risk groups based on predicted risk scores. Notably, approximately 80% of the cohort had minimal OUD risk, while the highest-risk decile alone accounted for ~68% (218/323) of all individuals who developed OUD. Given the low incidence of OUD within a 3-month period, the PPV was expectedly low. However, in the context of OUD prevention, a lower PPV may be ethically and clinically acceptable when associated interventions are supportive (eg, closer follow-up, motivational interviewing, review of opioid therapy, or other high-risk concurrent medication use) rather than punitive actions (eg, refusal to prescribe or abrupt discontinuation). For these low-risk, supportive interventions, the potential harm of a false positive (providing additional support to a patient who would not have developed OUD) is minimal compared with the harm of missing a true high-risk case (false negative). It is therefore important to frame the model as augmenting, not replacing, clinician

XSL•FO
RenderX

judgment to ensure that low PPV does not lead to overreaction to false positives or unintended consequences.

A key strength of this study is the successful external validation of our algorithm using an independent UPMC cohort. Our model maintained good performance (C-statistic: 0.76) without retraining, demonstrating transportability across two distinct health care systems. This replication supports the stability of the key predictors across settings and indicates that the model can effectively identify high-risk individuals in diverse clinical environments.

Prior studies, targeting different aspects of OUD risk, varied in their prediction windows and data sources, including 6-month OUD risk based on private insurance claims [66]; 12-month risk of aberrant opioid use behaviors following an initial pain clinic visit [67]; and 12-month OUD risk using private insurance claims [38,68] or pharmacy benefit manager claims data [36]. Other models focused on longer-term risk, such as 2-year problematic opioid use documented in EHR from primary care [69] and 5-year OUD risk using electronic medical record from a medical center [34] or Rhode Island Medicaid data [35]. Despite their contributions, these models had key limitations. Most measured predictors are only at baseline, lacking temporal updates over time. Several relied on case-control designs, which may not generalize well to population-level OUD incidence rates. Additionally, even in non–case-control designs, the highest reported C-statistic was 0.85, indicating room for improvement in predictive performance [35,36,67,69]. Expanding on our prior Medicare claims-based study, this work explored whether predictive performance can be refined using EHR data. Our study addressed the abovementioned limitations by using a population-based cohort and a more immediate prediction window, estimating OUD risk within the next 3 months rather than over a year or longer. This short-term predictive approach, combined with innovative risk stratification and enhanced clinical interpretability, made our method highly applicable for timely intervention.

Limitations

Although the findings of the study are promising, some limitations need to be acknowledged. First, incident OUD was identified solely through *ICD-10* diagnosis codes, which may undercapture true OUD cases due to known under-coding, variability in diagnostic practices, and reluctance to document substance use disorders in clinical settings. As a result, some individuals with OUD may have been misclassified as noncases. Although prior work has similarly relied on diagnosis-based definitions [29,34-36], broader definitions incorporating buprenorphine treatment or combinations of high-risk prescribing behavior indicators may improve case ascertainment. Second, unmeasured predictors, such as socioeconomic determinants and illicit opioid use, could influence risk trajectories. Third, while the study uniquely assessed OUD risk within 3 months of opioid initiation, its reliance on older EHR data may limit real-time applicability due to potential delays in data availability and processing.

Clinical Utility

From a clinical use perspective, the model's ability to stratify patients by risk supports early intervention and more efficient resource allocation in primary care settings. Our analysis confirms that the model's risk estimates are not only accurate in distinguishing higher- from lower-risk patients but also well calibrated after post hoc adjustment. Decision-curve analysis further demonstrated that the model offers meaningful clinical utility across thresholds where default "treat-all" strategies offer little value in the context of a low outcome incidence. At the top decile threshold, the model achieved a positive net benefit equivalent to identifying approximately 2.6 additional true OUD cases per 1000 patients—representing about 50% of the maximum achievable use given the cohort's baseline incidence. These findings align with established decision-analytic frameworks [70] and highlight the model's potential to effectively prioritize individuals most likely to benefit from preventive intervention. Integrating such models into clinical decision support (CDS) tools could enable providers to proactively flag and manage patients at elevated risk for OUD before symptoms escalate. However, ethical considerations must also be addressed, including the potential for unintentionally reinforcing existing health disparities. Although the overall FNR and FPR parity ratio was within acceptable bounds, Black patients exhibited higher false negative rates across percentiles, suggesting that true OUD cases may be underidentified in this group. To address this disparity, we outline potential mitigation strategies for future model refinement. These include group-specific threshold adjustment to harmonize error rates, reweighting or cost-sensitive learning to shift model attention toward underidentified subgroups, and postprocessing approaches such as calibration-by-group or equalized-odds adjustment. Importantly, any fairness intervention should balance reducing disparities with avoiding unintended undertreatment or new forms of inequity. These considerations underscore the need for iterative fairness evaluation as part of a responsible deployment framework.

Clinical Implementation and Future Directions

To support real-world deployment, our translational roadmap includes a structured, multiphase implementation strategy grounded in principles of human-centered design and clinical workflow alignment. In the retrospective phase, we will operationalize the OUD risk model within the UF Health Integrated Data Repository, a comprehensive enterprise clinical data warehouse. This step will enable validation using real-time patient data and ensure backend compatibility with the Epic EHR system. The model will generate individual-level risk scores for patients with recent opioid prescriptions, using secure internal infrastructure and consistent EHR data formatting. In the next phase, we will begin with a feasibility test and pilot integration of the model into the EHR system (Epic) within UF Health, gathering real-world feedback from clinicians before scaling up. Notably, UF Health's Epic EHR infrastructure has already implemented a system-wide approach for an artificial intelligence–driven CDS tool that predicts patients with a high risk of opioid overdose in primary care workflows [71,72].

XSL•FO
RenderX

In the near term, we plan to deploy the OUD risk prediction model through EHR-based CDS tools that seamlessly integrate into routine care. The model's risk score will trigger automated alerts during opioid prescribing encounters, prompting timely interventions for those identified as high risk. For example, in a primary care clinic with approximately 1000 patients initiating opioid therapy annually, applying a top decile (10%) risk threshold would flag approximately 100 individuals as high risk, of whom ~4 would be expected to develop OUD within 3 months. Although the PPV remains low due to the rarity of incident OUD, the alert volume is manageable in routine practice, particularly when recommended interventions are supportive and low-intensity (eg, increased follow-up, medication review, and motivational interviewing). The remaining ~90% low-risk patients would have an extremely high NPV, allowing providers to continue appropriate pain management for those individuals without unnecessary alarm. To further address alert fatigue, we propose specific operational strategies such as (1) embedding alerts within existing opioid or pain management workflows (eg, triggering only when signing a new medication order) to minimize disruption, (2) using tiered alerting triggered by repeated high-risk predictions, and (3) combining risk scores with other clinically meaningful indicators (eg, concurrent benzodiazepine prescribing) to enhance alert specificity. To minimize potential harms such as stigma and inappropriate opioid restriction, alerts should be framed using universal-precaution language that encourages supportive monitoring rather than implying misuse.

Although our findings indicate that the model has characteristics that could support future integration into clinical workflows, real-world feasibility and impact have not yet been evaluated. Additional implementation studies, including usability testing, workflow integration, and prospective evaluation, are needed before broad deployment in routine practice. While our separate pilot trial, Developing and Evaluating a Machine Learning Opioid Prediction & Risk-Stratification E-Platform, is currently assessing the feasibility of integrating an EHR-based overdose alert into clinical workflows within UF Health primary care clinics [71,72], analogous testing would be required for the model developed in this study.

Conclusions

This study demonstrates that an ML algorithm derived from the OneFlorida+ network can accurately predict incident OUD and effectively stratify risk across diverse populations and health systems, as evidenced by successful external validation in UPMC. The next step is to prospectively validate this algorithm within the UF Health clinical workflow to leverage existing decision support infrastructure and facilitate early intervention. Future research should expand the model to incorporate unstructured clinical notes, prescription drug monitoring program data, and social drivers of health (eg, housing instability, employment status, and incarceration history) to further improve prediction accuracy and equity.

Funding

This project is funded by the US National Institute on Drug Abuse (R01DA050676 and R01DA044985). WLC and WFG are named as inventors in one preliminary patent (U1195.70174US00) filing from the University of Florida and University of Pittsburgh for use of a machine-learning algorithm for opioid risk prediction in Medicare. The views presented here are those of the authors alone and do not necessarily represent the views of the Department of Veterans Affairs or the United States Government.

Data Availability

The full modeling pipeline, including code for data preprocessing, training, validation, and fairness analysis, is publicly available in the GitHub repository [73]. The datasets generated or analyzed during this study are not publicly available due to patient privacy regulations, Health Insurance Portability and Accountability Act, but are available from the OneFlorida+ network upon reasonable request and institutional review board approval.

Authors' Contributions

Conceptualization: JAF, WL-C, WFG, YW, CAH, KN, JLH, GC, DLW, SASS, SOFS, EIR, DN, SY, GMR, WMG, CK, MMH
Data curation: JAF, JLH, WLC, and MMH
Formal analysis: JAF, JLH, WLC, and MMH
Funding acquisition: WLC
Investigation: All authors
Methodology: All authors
Project administration: WLC
Resources: JAF, JLH, WLC, and MMH
Software: JAF, WLC, and MMH
Supervision: WLC and MMH
Validation: JAF, JLH, WLC, and MMH
Visualization: JAF, WLC, and MMH
Writing – original draft: JAF and MMH
Writing – review and editing: All authors

XSL•FO
RenderX

Conflicts of Interest

DLW and WLC have received grant funding from Merck, Sharp & Dohme and Bristol Myers Squibb for unrelated research. WLC was compensated for consulting services by Teva Pharmaceuticals.

Multimedia Appendix 1
Study design, model performance (internal and external validation), calibration and decision curves, and risk stratification.
[DOCX File, 2662 KB - jmir_v28i1e79482_app1.docx]

Multimedia Appendix 2
Diagnosis codes, predictor definitions, and model performance.
[DOCX File, 85 KB - jmir_v28i1e79482_app2.docx]

Multimedia Appendix 3
Methods for machine learning model development and evaluation.
[DOCX File, 58 KB - jmir_v28i1e79482_app3.docx]

Checklist 1
TRIPOD+AI checklist.
[DOCX File, 47 KB - jmir_v28i1e79482_app4.docx]

References

1. Florence C, Luo F, Rice K. The economic burden of opioid use disorder and fatal opioid overdose in the United States, 2017. Drug Alcohol Depend 2021 Jan 1;218:108350. [doi: 10.1016/j.drugalcdep.2020.108350] [Medline: 33121867]
2. Hébert AH, Hill AL. Impact of opioid overdoses on US life expectancy and years of life lost, by demographic group and stimulant co-involvement: a mortality data analysis from 2019 to 2022. Lancet Reg Health Am 2024 Aug;36:100813. [doi: 10.1016/j.lana.2024.100813] [Medline: 38978785]
3. Leslie DL, Ba DM, Agbese E, Xing X, Liu G. The economic burden of the opioid epidemic on states: the case of Medicaid. Am J Manag Care 2019 Jul;25(13 Suppl):S243-S249. [Medline: 31361426]
4. Keyes KM, Rutherford C, Hamilton A, et al. What is the prevalence of and trend in opioid use disorder in the United States from 2010 to 2019? Using multiplier approaches to estimate prevalence for an unknown population size. Drug Alcohol Depend Rep 2022 Jun;3:100052. [doi: 10.1016/j.dadr.2022.100052] [Medline: 35783994]
5. About overdose prevention. Centers for Disease Control and Prevention (CDC). URL: https://www.cdc.gov/overdose-prevention/about/index.html [accessed 2024-05-04]
6. Kline D, Hepler SA, Krawczyk N, Rivera-Aguirre A, Waller LA, Cerdá M. A state-level history of opioid overdose deaths in the United States: 1999-2021. PLoS ONE 2024;19(9):e0309938. [doi: 10.1371/journal.pone.0309938] [Medline: 39240938]
7. Chen LH, Hedegaard H, Warner M. Drug-poisoning deaths involving opioid analgesics: United States, 1999-2011. NCHS Data Brief 2014 Sep(166):1-8. [Medline: 25228059]
8. Ruhm CJ. Corrected US opioid‐involved drug poisoning deaths and mortality rates, 1999–2015. Addiction 2018 Jul;113(7):1339-1344. [doi: 10.1111/add.14144] [Medline: 29430760]
9. Overdose prevention: about prescription opioids. Centers for Disease Control and Prevention (CDC). URL: https://www.cdc.gov/overdose-prevention/about/prescription-opioids.html [accessed 2024-10-14]
10. National Institute on Drug Abuse, NIDA. Only 1 in 5 US adults with opioid use disorder received medications to treat it in 2021. URL: https://nida.nih.gov/news-events/news-releases/2023/08/only-1-in-5-us-adults-with-opioid-use-disorder-received-medications-to-treat-it-in-2021 [accessed 2024-10-14]
11. Volkow ND, Blanco C. The changing opioid crisis: development, challenges and opportunities. Mol Psychiatry 2021 Jan;26(1):218-233. [doi: 10.1038/s41380-020-0661-4] [Medline: 32020048]
12. Lo-Ciganic WH, Donohue JM, Yang Q, et al. Developing and validating a machine-learning algorithm to predict opioid overdose in medicaid beneficiaries in two US states: a prognostic modelling study. Lancet Digit Health 2022 Jun;4(6):e455-e465. [doi: 10.1016/S2589-7500(22)00062-0] [Medline: 35623798]
13. Edlund MJ, Martin BC, Russo JE, DeVries A, Braden JB, Sullivan MD. The role of opioid prescription in incident opioid abuse and dependence among individuals with chronic noncancer pain: the role of opioid prescription. Clin J Pain 2014 Jul;30(7):557-564. [doi: 10.1097/AJP.0000000000000021] [Medline: 24281273]
14. Yang Z, Wilsey B, Bohm M, et al. Defining risk of prescription opioid overdose: pharmacy shopping and overlapping prescriptions among long-term opioid users in medicaid. J Pain 2015 May;16(5):445-453. [doi: 10.1016/j.jpain.2015.01.475] [Medline: 25681095]

XSL•FO
RenderX

15. Eguale T, Bastardot F, Song W, et al. A machine learning application to classify patients at differing levels of risk of opioid use disorder: clinician-based validation study. JMIR Med Inform 2024 Jun 4;12:e53625. [doi: 10.2196/53625] [Medline: 38842167]
16. Screener and opioid assessment for patients with pain-revised (SOAPP®-r). : Inflexxion, Inc; 2008 URL: https://ddph-materials.s3.amazonaws.com/HelpIsHere/SOAPP-Tool.pdf [accessed 2026-01-22]
17. Cheatle MD, Compton PA, Dhingra L, Wasser TE, O'Brien CP. Development of the revised opioid risk tool to predict opioid use disorder in patients with chronic nonmalignant pain. J Pain 2019 Jul;20(7):842-851. [doi: 10.1016/j.jpain.2019.01.011] [Medline: 30690168]
18. Chou R, Hartung D, Turner J, et al. Opioid Treatments for Chronic Pain: Agency for Healthcare Research and Quality (US); 2020. URL: https://www.ncbi.nlm.nih.gov/sites/books/NBK556253/ [accessed 2026-02-13]
19. Bamboo health. NarxCare. URL: https://bamboohealth.com/solutions/narxcare/ [accessed 2025-11-03]
20. Wei YJJ, Chen C, Sarayani A, Winterstein AG. Performance of the centers for medicare & medicaid services' opioid overutilization criteria for classifying opioid use disorder or overdose. JAMA 2019 Feb 12;321(6):609-611. [doi: 10.1001/jama.2018.20404] [Medline: 30747958]
21. Canan C, Polinski JM, Alexander GC, Kowal MK, Brennan TA, Shrank WH. Automatable algorithms to identify nonmedical opioid use using electronic data: a systematic review. J Am Med Inform Assoc 2017 Nov 1;24(6):1204-1210. [doi: 10.1093/jamia/ocx066] [Medline: 29016967]
22. Rough K, Huybrechts KF, Hernandez-Diaz S, Desai RJ, Patorno E, Bateman BT. Using prescription claims to detect aberrant behaviors with opioids: comparison and validation of 5 algorithms. Pharmacoepidemiol Drug Saf 2019 Jan;28(1):62-69. [doi: 10.1002/pds.4443] [Medline: 29687539]
23. Goyal H, Singla U, Grimsley EW. Identification of opioid abuse or dependence: no tool is perfect. Am J Med 2017 Mar;130(3):e113. [doi: 10.1016/j.amjmed.2016.09.022] [Medline: 28215952]
24. Wood E, Simel DL, Klimas J. Pain management with opioids in 2019-2020. JAMA 2019 Nov 19;322(19):1912-1913. [doi: 10.1001/jama.2019.15802] [Medline: 31600370]
25. Dong X, Deng J, Rashidian S, et al. Identifying risk of opioid use disorder for patients taking opioid medications with deep learning. J Am Med Inform Assoc 2021 Jul 30;28(8):1683-1693. [doi: 10.1093/jamia/ocab043] [Medline: 33930132]
26. Gao W, Leighton C, Chen Y, Jones J, Mistry P. Predicting opioid use disorder and associated risk factors in a Medicaid managed care population. Am J Manag Care 2021 Apr;27(4):148-154. [doi: 10.37765/ajmc.2021.88617] [Medline: 33877773]
27. Hasan MM, Young GJ, Patel MR, Modestino AS, Sanchez LD, Noor-E-Alam M. A machine learning framework to predict the risk of opioid use disorder. Machine Learning with Applications 2021 Dec;6:100144. [doi: 10.1016/j.mlwa.2021.100144]
28. Segal Z, Radinsky K, Elad G, et al. Development of a machine learning algorithm for early detection of opioid use disorder. Pharmacol Res Perspect 2020 Dec;8(6):e00669. [doi: 10.1002/prp2.669] [Medline: 33200572]
29. Lo-Ciganic WH, Huang JL, Zhang HH, et al. Using machine learning to predict risk of incident opioid use disorder among fee-for-service medicare beneficiaries: a prognostic study. PLoS ONE 2020;15(7):e0235981. [doi: 10.1371/journal.pone.0235981] [Medline: 32678860]
30. Hogan WR, Shenkman EA, Robinson T, et al. The OneFlorida data trust: a centralized, translational research data infrastructure of statewide scope. J Am Med Inform Assoc 2022 Mar 15;29(4):686-693. [doi: 10.1093/jamia/ocab221] [Medline: 34664656]
31. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024 Apr 16;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
32. Smith SM, Winterstein AG, Gurka MJ, et al. Initial antihypertensive regimens in newly treated patients: real world evidence from the OneFlorida+ clinical research network. J Am Heart Assoc 2023 Jan 3;12(1):e026652. [doi: 10.1161/JAHA.122.026652] [Medline: 36565195]
33. OneFlorida+ clinical research network. The OneFlorida+ Data Trust. URL: https://onefl.net/data/ [accessed 2025-01-07]
34. Ellis RJ, Wang Z, Genes N, Ma'ayan A. Predicting opioid dependence from electronic health records with machine learning. BioData Min 2019;12(1):3. [doi: 10.1186/s13040-019-0193-0] [Medline: 30728857]
35. Hastings JS, Howison M, Inman SE. Predicting high-risk opioid prescriptions before they are given. Proc Natl Acad Sci U S A 2020 Jan 28;117(4):1917-1923. [doi: 10.1073/pnas.1905355117] [Medline: 31937665]
36. Ciesielski T, Iyengar R, Bothra A, Tomala D, Cislo G, Gage BF. A tool to assess risk of de novo opioid abuse or dependence. Am J Med 2016 Jul;129(7):699-705. [doi: 10.1016/j.amjmed.2016.02.014] [Medline: 26968469]
37. Hall AJ, Logan JE, Toblin RL, et al. Patterns of abuse among unintentional pharmaceutical overdose fatalities. JAMA 2008 Dec 10;300(22):2613-2620. [doi: 10.1001/jama.2008.802] [Medline: 19066381]
38. White AG, Birnbaum HG, Schiller M, Tang J, Katz NP. Analytic models to identify patients at risk for prescription opioid abuse. Am J Manag Care 2009 Dec;15(12):897-906. [Medline: 20001171]
39. Sullivan MD, Edlund MJ, Fan MY, DeVries A, Braden JB, Martin BC. Risks for possible and probable opioid misuse among recipients of chronic opioid therapy in commercial and medicaid insurance plans: The TROUP Study. Pain 2010 Aug;150(2):332-339. [doi: 10.1016/j.pain.2010.05.020] [Medline: 20554392]

XSL•FO
RenderX

40. Cepeda MS, Fife D, Chow W, Mastrogiovanni G, Henderson SC. Assessing opioid shopping behaviour: a large cohort study from a medication dispensing database in the US. Drug Saf 2012 Apr 1;35(4):325-334. [doi: 10.2165/11596600-000000000-00000] [Medline: 22339505]
41. Cochran G, Gordon AJ, Lo-Ciganic WH, et al. An examination of claims-based predictors of overdose from a large medicaid program. Med Care 2017 Mar;55(3):291-298. [doi: 10.1097/MLR.0000000000000676] [Medline: 27984346]
42. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 2015;10(3):e0118432. [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]
43. Romero-Brufau S, Huddleston JM, Escobar GJ, Liebow M. Why the C-statistic is not informative to evaluate early warning scores and what metrics to use. Crit Care 2015 Aug 13;19(1):285. [doi: 10.1186/s13054-015-0999-1] [Medline: 26268570]
44. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B 2005 Apr 1;67(2):301-320. [doi: 10.1111/j.1467-9868.2005.00503.x]
45. Larochelle MR, Zhang F, Ross-Degnan D, Wharam JF. Rates of opioid dispensing and overdose after introduction of abuse-deterrent extended-release oxycodone and withdrawal of propoxyphene. JAMA Intern Med 2015 Jun;175(6):978-987. [doi: 10.1001/jamainternmed.2015.0914] [Medline: 25895077]
46. Chu A, Ahn H, Halwan B, et al. A decision support system to facilitate management of patients with acute gastrointestinal bleeding. Artif Intell Med 2008 Mar;42(3):247-259. [doi: 10.1016/j.artmed.2007.10.003] [Medline: 18063351]
47. Ives TJ, Chelminski PR, Hammett-Stabler CA, et al. Predictors of opioid misuse in patients with chronic pain: a prospective cohort study. BMC Health Serv Res 2006 Apr 4;6(1):46. [doi: 10.1186/1472-6963-6-46] [Medline: 16595013]
48. Becker WC, Sullivan LE, Tetrault JM, Desai RA, Fiellin DA. Non-medical use, abuse and dependence on prescription opioids among U.S. adults: psychiatric, medical and substance use correlates. Drug Alcohol Depend 2008 Apr 1;94(1-3):38-47. [doi: 10.1016/j.drugalcdep.2007.09.018] [Medline: 18063321]
49. Dowell D, Haegerich TM, Chou R. CDC guideline for prescribing opioids for chronic pain — United States, 2016. MMWR Recomm Rep 2023;65(1):1-49. [doi: 10.15585/mmwr.rr6501e1] [Medline: 26987082]
50. UK NGC, Others. Evidence review: risk factors for dependence. : National Institute for Health and Care Excellence (NICE); 2022 URL: https://www.ncbi.nlm.nih.gov/books/NBK580678/ [accessed 2026-02-13]
51. How opioid use disorder occurs. Mayo Clinic. URL: https://www.mayoclinic.org/diseases-conditions/prescription-drug-abuse/in-depth/how-opioid-addiction-occurs/art-20360372 [accessed 2025-10-31]
52. Guidelines for prescribing controlled substances for pain. Medical Board of California. URL: https://www.mbc.ca.gov/Resources/Medical-Resources/controlled-substance.aspx [accessed 2025-11-03]
53. Ling W, Mooney L, Hillhouse M. Prescription opioid abuse, pain and addiction: clinical issues and implications. Drug Alcohol Rev 2011 May;30(3):300-305. [doi: 10.1111/j.1465-3362.2010.00271.x] [Medline: 21545561]
54. Committee on Pain Management and Regulatory Strategies to Address Prescription Opioid Abuse. Pain Management and the Opioid Epidemic: Balancing Societal and Individual Benefits and Risks of Prescription Opioid Use: National Academies Press; 2017. [doi: 10.17226/24781]
55. Lo-Ciganic WH, Huang JL, Zhang HH, et al. Evaluation of machine-learning algorithms for predicting opioid overdose risk among medicare beneficiaries with opioid prescriptions. JAMA Netw Open 2019 Mar 1;2(3):e190968. [doi: 10.1001/jamanetworkopen.2019.0968] [Medline: 30901048]
56. Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction 2009. [doi: 10.1198/jasa.2004.s339]
57. Al Faysal J, Noor-E-Alam M, Young GJ, et al. An explainable machine learning framework for predicting the risk of buprenorphine treatment discontinuation for opioid use disorder among commercially insured individuals. Comput Biol Med 2024 Jul;177:108493. [doi: 10.1016/j.compbiomed.2024.108493] [Medline: 38833799]
58. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988 Sep;44(3):837-845. [doi: 10.2307/2531595] [Medline: 3203132]
59. Data Mining and Statistics for Decision Making: Wiley; 2011. [doi: 10.1002/9780470979174]
60. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. Biom J 2005 Aug;47(4):458-472. [doi: 10.1002/bimj.200410135] [Medline: 16161804]
61. Nohara Y, Matsumoto K, Soejima H, Nakashima N. Explanation of machine learning models using improved shapley additive explanation. 2019 Sep 4 Presented at: BCB '19: Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics; Sep 7-10, 2019; Niagara Falls NY USA p. 546. [doi: 10.1145/3307339.3343255]
62. Saleiro P, Kuester B, Hinkson L, et al. Aequitas: a bias and fairness audit toolkit. arXiv. Preprint posted online on Nov 14, 2018. [doi: 10.48550/arXiv.1811.05577]
63. Jiang X, Osl M, Kim J, Ohno-Machado L. Smooth isotonic regression: a new method to calibrate predictive models. AMIA Jt Summits Transl Sci Proc 2011:16-20. [Medline: 22211175]
64. Rufibach K. Use of Brier score to assess binary predictions. J Clin Epidemiol 2010 Aug;63(8):938-939. [doi: 10.1016/j.jclinepi.2009.11.009] [Medline: 20189763]
65. Peirce CS. The numerical measure of the success of predictions. Science 1884 Nov 14;4(93):453-454. [doi: 10.1126/science.ns-4.93.453-a] [Medline: 17795531]

XSL•FO
RenderX

66. Dufour R, Mardekian J, Pasquale MK, Schaaf D, Andrews GA, Patel NC. Understanding predictors of opioid abuse: predictive model development and validation. Am J Pharm Benefits 2014;6(5):208-216 [FREE Full text]
67. Webster LR, Webster RM. Predicting aberrant behaviors in opioid-treated patients: preliminary validation of the opioid risk tool. Pain Med 2005;6(6):432-442. [doi: 10.1111/j.1526-4637.2005.00072.x] [Medline: 16336480]
68. Rice JB, White AG, Birnbaum HG, Schiller M, Brown DA, Roland CL. A model to identify patients at risk for prescription opioid abuse, dependence, and misuse. Pain Med 2012 Sep;13(9):1162-1173. [doi: 10.1111/j.1526-4637.2012.01450.x] [Medline: 22845054]
69. Hylan TR, Von Korff M, Saunders K, et al. Automated prediction of risk for problem opioid use in a primary care setting. J Pain 2015 Apr;16(4):380-387. [doi: 10.1016/j.jpain.2015.01.011] [Medline: 25640294]
70. Kerr KF, Brown MD, Zhu K, Janes H. Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use. J Clin Oncol 2016 Jul 20;34(21):2534-2540. [doi: 10.1200/JCO.2015.65.5654] [Medline: 27247223]
71. Nguyen K, Wilson DL, Diiulio J, et al. Design and development of a machine-learning-driven opioid overdose risk prediction tool integrated in electronic health records in primary care settings. Bioelectron Med 2024 Oct 18;10(1):24. [doi: 10.1186/s42234-024-00156-3] [Medline: 39420438]
72. Hong JWJ, Wilson DL, Nguyen K, et al. Protocol for a single-arm pilot clinical trial: developing and evaluating a machine learning opioid prediction & risk-stratification E-platform (DEMONSTRATE). J Clin Med 2025 Dec 1;14(23):8522. [doi: 10.3390/jcm14238522] [Medline: 41375825]
73. Incident-OUD-project. GitHub. URL: https://github.com/faysal22/Incident-OUD-Project [accessed 2026-01-24]

Abbreviations

CDS: clinical decision support
EHR: electronic health record
EN: elastic net
FNR: false negative rate
FPR: false positive rate
GBM: gradient boosting machine
ICD-10: *International Statistical Classification of Diseases and Related Health Problems 10th Revision*
LASSO: least absolute shrinkage and selection operator
ML: machine learning
NNE: number needed to evaluate
NPV: negative predictive value
OUD: opioid use disorder
PPV: positive predictive value
RF: random forest
SHAP: Shapley additive explanations
TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+Artificial Intelligence
UF: University of Florida
UPMC: University of Pittsburgh Medical Center

Edited by J Sarvestan; submitted 22.Jun.2025; peer-reviewed by J Song, MA Roshani; revised version received 13.Dec.2025; accepted 15.Dec.2025; published 05.Mar.2026.

Please cite as:
Faysal JA, Lo-Ciganic W, Gellad WF, Wu Y, Harle CA, Nguyen K, Huang JL, Cochran G, Wilson DL, Staras SAS, Schmidt SOF, Rosenberg EI, Nelson D, Yan S, Reisfield GM, Greene WM, Kuza C, Hasan MM
Developing and Validating a Machine Learning Algorithm to Predict the Risk of Incident Opioid Use Disorder Among OneFlorida+ Patients: Prognostic Modeling Study
J Med Internet Res 2026;28:e79482
URL: https://www.jmir.org/2026/1/e79482
doi:10.2196/79482

© Jabed Al Faysal, Weihsuan Lo-Ciganic, Walid F Gellad, Yonghui Wu, Christopher A Harle, Khoa Nguyen, James L Huang, Gerald Cochran, Debbie L Wilson, Stephanie AS Staras, Siegfried OF Schmidt, Eric I Rosenberg, Danielle Nelson, Shunhua Yan, Gary M Reisfield, William M Greene, Courtney Kuza, Md Mahmudul Hasan. Originally published in the Journal of Medical

XSL•FO
RenderX

Internet Research (https://www.jmir.org), 5.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Breast Cancer Screening Knowledge and Sentiments in Singaporean Women: Mixed Methods Study Using Topic Modeling, Sentiment Analysis, and Structured Questionnaire Data

Peh Joo Ho[1,2,3], PhD; Zi Lin Lim[3], BSc (Hons); Jenny Liu[1,4], MSc; Nur Khaliesah Mohamed Riza[1,5], BSc; Ying Jia Chew[5], BSc (Hons); Yi Ying Lim[1], BSc (Hons); Hui Ling Tan[4], BMgmt (Hons); Su-Ann Goh[2], PhD; Han Boon Oh[4,6], MMed; Chi Hui Chin[7], MMed; Sing Cheer Kwek[8], MMed; Zhi Peng Zhang[9], MMed; Desmond Luan Seng Ong[7], MMed; Swee Tian Quek[10], MBBS; Sujith Wijerathne[5,6], MMed; Philip Tsau Choong Iau[4,5*], MD, PhD; Mikael Hartman[1,2,5*], PhD; Jingmei Li[2,3,11*], PhD

[1]Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore, Singapore

[10]Department of Diagnostic Imaging, National University Hospital and National University Health System, Singapore, Singapore

[11]National Cancer Centre Singapore, Singapore Health Services (SingHealth), Singapore, Singapore

[2]Department of Surgery, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore

[3]Genome Institute of Singapore, Agency for Science, Technology and Research (A*STAR), #02-01, 60 Biopolis Street, Singapore, Singapore

[4]Department of General Surgery, Ng Teng Fong General Hospital, Singapore, Singapore

[5]Department of Surgery, National University Hospital and National University Health System, Singapore, Singapore

[6]Department of General Surgery, Jurong Medical Centre, Singapore, Singapore

[7]Jurong Polyclinic, National University Polyclinics and National University Health System, Singapore, Singapore

[8]Bukit Batok Polyclinic, National University Polyclinics and National University Health System, Singapore, Singapore

[9]Choa Chu Kang Polyclinic, National University Polyclinics and National University Health System, Singapore, Singapore

[*]these authors contributed equally

Corresponding Author:
Jingmei Li, PhD
Department of Surgery, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore

Abstract

Background: Mammography screening uptake in Singapore remains below 40% despite campaigns and subsidies. Natural language processing (NLP) can extract nuanced attitudes from free text that fixed response options miss, revealing latent factors influencing breast cancer (BC) screening behavior.

Objective: This study characterized women's attitudes toward mammography using mixed methods data, examined associations between BC awareness and screening willingness, and identified barriers and facilitators through NLP of free-text responses.

Methods: We conducted a cross-sectional study within the *Breast Screening Tailored for Her* multicenter cohort in Singapore (October 2021-December 2023). In total, 4169 women aged 35 - 59 years (median 48, IQR 43 - 54) were recruited via convenience sampling (3 hospitals and 2 polyclinics). Participants completed online structured questionnaires on demographics and screening history, then a BC education quiz with feedback. Participants answering >80% correctly were classified as "BC-aware." Posteducation, participants reported screening willingness (motivated or neutral) with optional free-text explanations. Logistic regression models (adjusted for study site, age, ethnicity, marital status, housing, and education) examined the associations with willingness. For 3819 English-language respondents, biterm topic modeling identified themes and sentiment analysis quantified emotional tone. Statistical significance: α=.05.

Results: Overall, 79% (3287/4169) were BC-aware, and 94% (3908/4169) reported increased motivation posteducation. BC-aware women had higher screening motivation than BC-unaware women (adjusted odds ratio [aOR] 2.88, 95% CI 2.19 - 3.80; *P*<.001). Motivation was higher among those with larger public housing (OR 1.81, 95% CI 1.30 - 2.50; *P*<.001) and private housing vs 1 - 3 room units (OR 2.69, 95% CI 1.75 - 4.13; *P*<.001), married vs not separated, divorced, or widowed (OR 2.38 [inverse of 0.42], 95% CI 1.75 - 3.13; *P*<.001), and prior screening attendance (OR 3.49, 95% CI 2.71 - 4.50; *P*<.001). Women who disagreed that mammography was expensive had higher motivation (aOR 1.94, 95% CI 1.50 - 2.50; *P*<.001). Among 3819 English respondents, 94% (3579/3819) were motivated and 6% (240/3819) neutral. Free-text responses came from 34% (1220/3579) of motivated and 64% (153/240) of neutral participants. Biterm topic modeling revealed motivated participants emphasized early

XSL•FO
RenderX

detection benefits, health awareness, BC risk, and logistics; neutral participants focused on mammography pain experiences and cost barriers. Mean sentiment was 0.207 (range: −1.00 to 1.65), with motivated participants displaying more positive sentiments than neutral participants (linear regression, $P<.001$). Identical words carried different emotional tones across subgroups: "health" had positive sentiment among motivated participants (mean difference, Welch t tests $P<.05$) but negative sentiment among neutral participants. Word frequency analysis showed motivated participants used positive-sentiment words ("better," "cure," and "prevention"). Neutral participants emphasized negative words ("painful" and "uncomfortable").

Conclusions: Integrating quantitative surveys with NLP revealed that the same screening concepts are emotionally framed differently by motivated vs neutral women, a finding missed by knowledge- or intent-focused approaches alone. In practice, these findings support the need for emotionally tailored BC education and prevention strategies.

(J Med Internet Res 2026;28:e78439) doi:10.2196/78439

KEYWORDS

mammography screening; breast cancer; Asian population; screening attitudes; breast cancer awareness; precision health; proactive health

Introduction

Mammography screening has been widely regarded as the most effective method for achieving early detection of breast cancer (BC), leading to improved survival [1-3]. Despite a decade of annual campaigns and various subsidies, the mammography screening rate remains suboptimal, even among women who know about mammogram services in Singapore [4,5].

A review by Momenimovahed et al [6] found that barriers to mammography screening in Asia include factors such as financial limitations, lack of social support, fatalistic tendencies, fear of pain and embarrassment, and religion. Barriers to mammography screening specific to Singapore have also been widely examined in multiple studies, and the common obstacles reflected are the lack of time, fatalism, misconceptions about mammography screening or BC, reluctance to spend on screening, and religious reasons [7-12]. In a review by Rajendram et al [4], the perceived costs of screening played the biggest role in hindering screening participation among Singaporean women.

According to the principle of attitude consistency, a person's attitudes tend to influence and shape their actions and behavior [13]. Common techniques used to understand people's opinions and attitudes toward screening include quantitative tools, such as the Likert scale, or qualitative tools, such as focus groups or in-depth interviews. However, these tools are limited by a few factors. Self-reported measurements using the Likert scale often present participants with closed-ended scenarios. The finite number of response options can limit respondents' ability to express nuanced opinions, and responses can be prone to central tendency bias [14,15]. On the other hand, while focus groups and in-depth interviews offer valuable qualitative data, providing context and information on nuances of responses, they are limited by factors such as response bias [16], researcher bias, and time-intensive analysis processes [17].

In recent years, natural language processing (NLP) has emerged as a novel tool in health behavior and health services research, offering unique advantages in understanding individuals' attitudes and perceptions [18,19]. By leveraging machine learning algorithms, NLP can sift through vast amounts of textual information to extract nuanced insights [20,21]. This approach not only enables researchers to capture the rich complexity of human language but also serves as an avenue to reveal implicit sentiments and identify emerging themes. Subsequently, the use of NLP can offer a more comprehensive understanding of the factors influencing the uptake of health behaviors, such as screening.

Currently, studies involving the use of NLP tools in understanding health-seeking behavior largely revolve around social media posts [22]. A systematic review by Döbrössy et al [23] demonstrated the value of social media as a platform to spread BC awareness and the value of NLP tools in identifying common areas of concern of the lay public about BC. NLP's ability to analyze free text provides valuable insights into the psychological and social factors influencing health behaviors, potentially contributing to more effective public health interventions and communication strategies. This study leverages structured questionnaire data supplemented by NLP tools to analyze free-text responses, aiming to (1) characterize women's attitudes toward mammography, (2) examine the relationship between BC awareness and screening willingness, and (3) identify barriers and facilitators for the willingness to screen.

Methods

Study Design

We conducted a cross-sectional, observational, mixed methods study combining quantitative surveys and qualitative free-text analysis to examine women's attitudes and willingness to attend routine mammography screening. This study was reported following the American Psychological Association Journal Article Reporting Standards for mixed-methods, quantitative, and qualitative research [24].

Setting

The Breast Screening Tailored for Her (BREATHE) study is a multicenter prospective cohort study in Singapore [25,26]. Participants were recruited between October 2021 and December 2023 from 3 restructured hospitals—Ng Teng Fong General Hospital, National University Hospital, and Alexandra Hospital—2 polyclinics—Bukit Batok Polyclinic and Choa Chu Kang Polyclinic—and Jurong Medical Center [25]. Briefly, potential participants were identified through advertisements

XSL•FO
RenderX

(eg, posters, flyers, and blog recruitment pages inviting women to register interest) and an active approach at the clinical sites (ie, study staff approached women in waiting areas of the participating hospitals and polyclinics). Recruitment was based on convenience (ie, convenience sampling) and recruitment feasibility across participating clinical sites rather than a formal a priori power calculation.

Eligible participants were female, aged between 35 and 59 years, and Singapore citizens or permanent residents at the point of enrollment. Participants were excluded if they had a histologically confirmed diagnosis of any cancer, were cognitively impaired, or were pregnant during recruitment. Consistent with the subsidized Screen for Life BC screening criteria in Singapore, women with benign breast disease were eligible for the study, while those with breast implants were excluded. After providing informed consent, participants completed an online structured first-visit questionnaire, which included factors linked to BC and related conditions, such as demographic, lifestyle, and reproductive characteristics, prior treatments, and other environmental exposures (Multimedia Appendix 1). They then undertook a brief, self-administered education session online covering BC knowledge and the importance of regular breast self-examination and screening (Multimedia Appendix 2). Participants answered the online education questionnaire in either English, Chinese, or Malay.

A flowchart illustrating the selection of participants from those enrolled in the BREATHE study is shown in Figure S1 in Multimedia Appendix 3. Between October 2021 and December 2023, 4592 participants were enrolled in the study (corresponding to a participation rate of 83% among the 5536 women approached or expressing interest). Participants were excluded if they withdrew from the study (n=74), were diagnosed with BC within 6 months of enrollment (n=17), had unknown past mammography (n=38), or had incoherent BC screening and mammography screening attendance based on the lifestyle questionnaire and education session (n=294). The analytical cohort comprised 4169 participants. The free-text analysis included participants who answered the English questionnaire (n=3819).

Demographics and BC Risk Factors

Baseline information on demographics and BC risk factors was obtained from the first-visit questionnaire. The variables included ethnicity, medical history, previous benign lump or gynecological surgery (yes, no, or missing), family history of breast and ovarian cancer (yes, no, or missing), reproductive factors, and previous breast examination and screening habits, among others.

Sociodemographic factors were derived from the same questionnaire, where individual factors were further categorized for ease of analysis (Table S1 in Multimedia Appendix 4). Housing (public: 1 - 3 room flat, public: >3-room flat [4-, 5-, or executive-type], or private), highest qualification achieved (no formal or primary, secondary, postsecondary [nontertiary], professional diploma, or tertiary), and marital status (married, never married, widowed, or separated or divorced) were used as proxies for economic, education, and social support status, respectively.

BC Education

Participants' existing screening habits and views about BC were assessed during an online education session (Multimedia Appendix 2). Various statements were used to assess BC awareness and provide education (questions 7 - 13, 16, and 18). These statements were presented to participants to indicate their agreement (agree or disagree). The correct answer and an accompanying explanation were given after every response. Participants were characterized as aware of BC risk (ie, BC-aware) before study enrollment if they answered >80% of the questions correctly.

Following BC education, participants were asked about their willingness to attend regular screening (question 20: "After knowing the above, would you be more willing to attend regular screening?" yes or no) and were categorized as "motivated" or "neutral," respectively. Participants were also asked to explain their choice (question 21: "Referring to your answer in the last question, please provide a reason." free text, optional).

Statements unrelated to BC or BC screening knowledge were considered as BC perceptions for analysis (ie, perceived importance of BC screening, perceived risk of BC, fatalistic attitudes, and finding mammography embarrassing, expensive, inconvenient, or painful).

Data Analysis

Descriptive analyses were conducted to characterize demographic factors and BC screening attitudes. The associations between willingness to attend more regular screening and the various characteristics were studied using the chi-square and Kruskal-Wallis tests for categorical and continuous variables, respectively.

The study's primary outcome was self-reported willingness to screen, measured immediately after an education intervention. The association between the various factors and participants' willingness to attend regular screening was studied using logistic regression models. In addition to univariate analysis, the models were adjusted for study site and participant characteristics (age category, ethnicity, marital status, housing type, and highest academic status). To select the best predictive model, stepwise forward selection was used (ie, the lowest Bayesian information criterion). Odds ratios (ORs) and 95% CIs are presented.

Free-Text Data Preprocessing

The sentiment of words may not be well captured through translation. We thus retained only the reasons provided in English in the sentiment analysis. Of the 4169 participants, 3819 (92%) completed the English version of the education session. Participants (n=350) who completed the Chinese or Malay version of the education session were excluded from text analysis. Free-text analysis was done to assess participants' sentiments toward regular screening. Each participant's response was tokenized (R package *tidytext;* version 0.4.1, Julia Silge and David Robinson) [27], and spelling errors were corrected (R package *hunspell;* version 3.0.4, Jeroen Ooms). Additional spelling errors and corrections were added to the list generated. The full spelling errors and corrections can be found in Table S2 in Multimedia Appendix 4. The participants were categorized

XSL•FO
RenderX

by BC awareness (BC-aware vs BC-unaware) and their willingness to attend regular screening in the future posteducation session (motivated vs neutral; Figure S1 in Multimedia Appendix 3).

Biterm Topic Modeling

Biterm topic modeling (R package *BTM*; version 0.3.7, Jan Wijffels) was used to identify the top topics mentioned by the participants [28]. Although there are other topic modeling methods, such as latent Dirichlet allocation, latent semantic analysis, nonnegative matrix factorization, and transformer-based approaches such as BERTopic, biterm topic modeling was chosen due to the short length of the text inputs (1-3 sentences), its computational efficiency, and the interpretability of the resulting topics [29,30]. As the choice of the number of topics in the topic model will impact the effect of the model, we compared the results of the models with 2-5 topics (ie, parameter k). Expecting a small number of topics within each participant's response, we set α=.01 and β=.01. The clustering of words into topics depends on the initialization step and may change with different iterations. We attempted 6 iterations for each k, using set.seed() and permutations of 123 (ie, 123, 132, 213, 231, 312, and 321). The choice of the number of topics (k) was based on the stability of the topics produced across iterations, and the topic for each set of words was derived by reviewing the comments for relevant words and discussing them with team members. Biterm topic modeling looks at pairs of words and the probability of either of the words being in a particular topic. Topics emerged from the data rather than being predetermined, with the biterm topic modeling algorithmically identifying word co-occurrence patterns without a priori coding categories. This requires word stemming (R package *hunspell*; version 3.0.4) to prevent words with the same meaning from being tokenized into different tokens (eg, the words "prevented" and "prevents" have the same stem word "prevent," but not "prevention") and removing stop words from the text corpus. The list of stop words was derived from "stop_words" in tidytext and was further edited to suit this study. The full list of stem words and stop words is provided in Tables S3 and S4 in Multimedia Appendix 4, respectively.

Sentiment Analysis

In our study, "sentiment" refers to the emotional tone of participants' free-text responses as classified by NLP algorithms and used strictly as a quantitative analytical construct rather than a synonym for perception, attitude, or feeling. Unlike biterm modeling, which uses all pairs of words, sentiment analysis analyzes sentences or single words. Here, sentiment analysis was done using the R package *sentimentr;* version 2.9.0, Tyler W Rinker), as it accounts for valence shifters to determine the sentiment at the sentence level, while estimating the sentiment score by referencing the Jockers sentiment dictionary. Because these free-text responses were brief (1-3 sentences) and expressed lay perspectives and emotions about mammography screening rather than technical medical terminology, the use of a general-purpose sentiment lexicon with minor modifications was appropriate and ensured consistent capture of participants' attitudes across the cohort. The package scores sentiments on a scale where 0 is considered neutral, negative numbers indicate the presence of negative sentiments, and positive numbers indicate the presence of positive sentiments. Sentiment rating of certain words was further modified to suit the context of the study (Table S5 in Multimedia Appendix 4). Free-text responses to question 21 in Multimedia Appendix 2 were first analyzed by sentence and then averaged across sentences for each participant. Linear regression was used to determine significant differences in sentiment ratings across subgroups.

Word clouds (R package *ggplot2;* version 3.4.4, Hadley Wickham) were used to tabulate the top sentiment words of each subgroup. The larger the word in the visual, the more common the word was in the subgroup. To further explore patterns, common sentiment words, and the average sentiment associated were compared across subgroups using the Welch 2-sample *t* test. Unique sentiment words were also extracted. Example quotes of common and unique words were extracted randomly to provide better context.

Interpretive Limitations

The research team comprised epidemiologists, public health researchers, and data scientists with expertise in cancer screening behavior and computational text analysis. The use of algorithmic biterm topic modeling and sentiment analysis minimizes researcher interpretation bias. Researchers' perspectives influenced only the selection of analytical methods and the labeling of algorithmically identified topic clusters. Rather than sampling to thematic saturation as in traditional qualitative research, we analyzed all available free-text responses. Data collection concluded at the end of the study recruitment period (December 2023) rather than when thematic saturation was achieved. The stability of topics identified across 6 biterm topic modeling iterations with different random seeds provides evidence of thematic robustness within our dataset. The relatively brief nature of responses (1 - 3 sentences) and focus on specific questions means our analysis captures targeted attitudes toward screening rather than comprehensive life narratives, which is appropriate for our research objectives.

Ethical Considerations

This study involved human participants and complied with all relevant institutional and national research ethics guidelines. Ethics approval was obtained from the National Healthcare Group Domain-Specific Review Board, Singapore (reference no 2020/01327; approval date: June 7, 2021). Written informed consent was obtained from participants by trained study coordinators in the participant's preferred language (English, Chinese, or Malay). The informed consent process included permission to use the study data for secondary analyses relevant to BC screening research; therefore, no additional consent was required for the current analysis. Participant privacy and confidentiality were safeguarded throughout the study. All research data were deidentified before analysis, stored on secure servers with restricted access, and handled in accordance with institutional data protection policies to ensure anonymity. No individual participants can be identified in any image in the paper or multimedia appendices.

XSL•FO
RenderX

Results

Participant Characteristics

Table 1 shows the descriptive statistics of participant characteristics. Of the 4169 participants included, 3908 (94%) indicated that they were more motivated to attend regular screening posteducation sessions. The median age of the participants was 48 years (IQR 43-54), and the majority (77%, n=3208) of the participants were of Chinese ethnicity. Overall, 79% (n=3287) of the participants were BC-aware, and 22% (n=938) and 41% (n=1690) attended BC screening once every year or once every 2 years, respectively. Seventy-three percent (n=3056) had done at least 1 mammogram in the past. While 44% (n=1845) perceived themselves as low risk and 47% (n=1944) as average risk, 96% (n=4023) of the participants agreed that BC screening is important. Twelve percent (n=485) of the participants reported a history of benign breast disease.

XSL•FO
RenderX

Table . Characteristics of the BREAst Screening Tailored for HEr (BREATHE) study population, by their willingness to attend screening. Participants were recruited between 2021 and 2023 in Singapore's health care institutes. Two-sided *P* values for categorical variables are based on the chi-square test, and *P* values for continuous variables are based on the Kruskal-Wallis test. Counts and column percentages are presented, unless otherwise stated.

Characteristic	Total, n=4169, n (%)	Neutral, n=261, n (%)	Motivated, n=3908, n (%)	*P* value
BC[a] awareness				<.001
BC-unaware	882 (21)	104 (40)	778 (20)	
BC-aware	3287 (79)	157 (60)	3130 (80)	
Age, median (IQR)	48 (43-53)	49 (44-53)	48 (42-53)	.45
Age category (years)				.33
35-39	531 (13)	35 (13)	496 (13)	
40-49	1846 (44)	104 (40)	1742 (45)	
50-59	1792 (43)	122 (47)	1670 (43)	
Participant characteristic				
Ethnicity				.96
Chinese	3208 (77)	204 (78)	3004 (77)	
Malay	465 (11)	28 (11)	437 (11)	
Indian	279 (7)	17 (7)	262 (7)	
Other	217 (5)	12 (5)	205 (5)	
Marital status				<.001
Married	3186 (76)	164 (63)	3022 (77)	
Separated, divorced, or widowed	349 (8)	25 (10)	324 (8)	
Never married	634 (15)	72 (28)	562 (14)	
Housing				<.001
Public 1-3 room	490 (12)	52 (20)	438 (11)	
Public >3 room	2757 (66)	170 (65)	2587 (66)	
Private	922 (22)	39 (15)	883 (23)	
Highest qualification attained				.06
No formal or primary	235 (6)	23 (9)	212 (5)	
Secondary	787 (19)	46 (18)	741 (19)	
Postsecondary	1240 (30)	66 (25)	1174 (30)	
Tertiary	1907 (46)	126 (48)	1781 (46)	
Self-reported breast cancer history, existing behavior, and perceptions				
Family history of BC				.94
No	3756 (90)	236 (90)	3520 (90)	
Yes	413 (10)	25 (10)	388 (10)	
Benign breast disease				.25
No	3673 (88)	235 (90)	3438 (88)	
Yes	485 (12)	24 (9)	461 (12)	
Unknown	11 (0)	2 (1)	9 (0)	
BCS[b] attendance				<.001
No	1113 (27)	140 (54)	973 (25)	
Yes	3056 (73)	121 (46)	2935 (75)	

XSL•FO
RenderX

Characteristic	Total, n=4169, n (%)	Neutral, n=261, n (%)	Motivated, n=3908, n (%)	*P* value
Mammography attendance				<.001
No	1113 (27)	140 (54)	973 (25)	
Yes	3056 (73)	121 (46)	2935 (75)	
Previous BCS type				<.001
Mammography	3056 (73)	121 (46)	2935 (75)	
Ultrasound only	94 (2)	7 (3)	87 (2)	
Neither	1019 (24)	133 (51)	886 (23)	
Current BCS behavior				<.001
Once a year	938 (22)	11 (4)	927 (24)	
Once every 2 years	1690 (41)	32 (12)	1658 (42)	
Do not intend to attend in the future	99 (2)	43 (16)	56 (1)	
Other	320 (8)	35 (13)	285 (7)	
Not in the above categories and aged 35-39 years	506 (12)	35 (13)	471 (12)	
Unknown	616 (15)	105 (40)	511 (13)	
Perceived importance of BCS				<.001
Agree	4023 (96)	198 (76)	3825 (98)	
Neutral	136 (3)	57 (22)	79 (2)	
Disagree	10 (0)	6 (2)	4 (0)	
Perceived risk of BC				.03
Low	1845 (44)	133 (51)	1712 (44)	
Average	1944 (47)	101 (39)	1843 (47)	
High	380 (9)	27 (10)	353 (9)	
Fatalism toward developing BC				.85
Low	2002 (48)	121 (46)	1881 (48)	
Average	1179 (28)	77 (30)	1102 (28)	
High	988 (24)	63 (24)	925 (24)	
Fatalism toward dying from BC				.22
Low	2456 (59)	153 (59)	2303 (59)	
Average	1023 (25)	73 (28)	950 (24)	
High	690 (17)	35 (13)	655 (17)	
Finds mammography: embarrassing				<.001
Agree	586 (14)	74 (28)	512 (13)	
Disagree	3583 (86)	187 (72)	3396 (87)	
Expensive				<.001
Agree	1663 (40)	144 (55)	1519 (39)	
Disagree	2506 (60)	117 (45)	2389 (61)	
Inconvenient				<.001
Agree	866 (21)	105 (40)	761 (19)	

XSL•FO
RenderX

Characteristic	Total, n=4169, n (%)	Neutral, n=261, n (%)	Motivated, n=3908, n (%)	*P* value
Disagree	2779 (67)	120 (46)	2659 (68)	
Not applicable	524 (13)	36 (14)	488 (12)	
Painful				<.001
Agree	1027 (25)	134 (51)	893 (23)	
Disagree	2618 (63)	91 (35)	2527 (65)	
Not applicable	524 (13)	36 (14)	488 (12)	

[a]BC: breast cancer.

[b]BCS: breast cancer screening.

BC Awareness Is Significantly Associated With Posteducation Screening Motivation

In univariate logistic models, more motivated participants were more likely to be BC aware (OR 2.67, 95% $CI_{aware\ vs\ unaware}$ 2.05-3.46; *P*<.001), residing in >3 rooms HDB ($OR_{>3\ rooms\ HDB\ vs\ 1\text{-}3\ rooms\ HDB}$ 1.81, 95% CI 1.30-2.50; *P*<.001) or private housing ($OR_{private\ vs\ 1\text{-}3\ rooms\ HDB}$ 2.69, 95% CI 1.75-4.13; *P*<.001), currently married ($OR_{separated/divorced/widowed\ vs\ currently\ married}$ 0.42, 95% CI 0.32-0.57; *P*<.001), having attended BC screening before ($OR_{yes\ vs\ no}$ 3.49, 95% CI 2.71-4.50; *P*<.001), being currently regular screeners, and being less likely to find the procedure embarrassing, expensive, inconvenient, or painful (Table 2). The complete results table of univariate analysis and the results adjusting for participants' characteristics is provided in Table S6 in Multimedia Appendix 4.

XSL•FO
RenderX

Table . Associations between patient characteristics and willingness to attend regular screening posteducation. Stepwise forward selection was used to select the best predictive model (ie, the lowest Bayesian information criterion [BIC]).

Characteristic	Univariate		Best prediction model by BIC	
	OR (95% CI)	*P* value	OR (95% CI)	*P* value
BC[a] awareness		<.001		<.001
BC-unaware	1.00 (referent)		1.00 (referent)	
BC-aware	2.67 (2.05-3.46)		1.90 (1.41-2.56)	
Age category (years)				
35-39	1.00 (referent)	—[b]	—	—
40-49	1.18 (0.80-1.76)	.41	—	—
50-59	0.97 (0.65-1.43)	.86	—	—
Ethnicity				
Chinese	1.00 (referent)	—	—	—
Malay	1.06 (0.71-1.59)	.78	—	—
Indian	1.05 (0.63-1.74)	.86	—	—
Other	1.16 (0.64-2.11)	.63	—	—
Marital status				
Married	1.00 (referent)	—	1.00 (referent)	—
Never married	0.70 (0.45-1.09)	.11	0.80 (0.49-1.31)	.38
Separated, divorced, or widowed	0.42 (0.32-0.57)	<.001	0.46 (0.33-0.64)	<.001
Family history of breast cancer				
No	1.00 (referent)	—	—	—
Yes	1.04 (0.68-1.59)	.86	—	—
BCS[c] attendance				
No	1.00 (referent)	—	—	—
Yes	3.49 (2.71-4.50)	<.001	—	—
Previous BCS type				
Mammography	1.00 (referent)	—	—	—
Ultrasound only	0.51 (0.23-1.13)	.10	—	—
Neither	0.27 (0.21-0.36)	<.001	—	—
Current BCS behavior				
Once a year	1.00 (referent)	—	1.00 (referent)	—
Once every 2 years	0.61 (0.31-1.23)	.17	0.65 (0.33-1.30)	.23
Do not intend to attend in the future	0.02 (0.01-0.03)	<.001	0.03 (0.01-0.06)	<.001
Other	0.10 (0.05-0.19)	<.001	0.13 (0.07-0.27)	<.001
Not in above categories and aged 35-39 years	0.16 (0.08-0.32)	<.001	0.17 (0.03-0.87)	.03
Unknown	0.06 (0.03-0.11)	<.001	0.10 (0.05-0.20)	<.001
Perceived importance of BCS				
Agree	1.00 (referent)	—	1.00 (referent)	—
Neutral	0.07 (0.05-0.10)	<.001	0.21 (0.14-0.32)	<.001
Disagree	0.03 (0.01-0.12)	<.001	0.11 (0.02-0.50)	.004
Finds mammography painful				

XSL•FO
RenderX

Characteristic	Univariate		Best prediction model by BIC	
	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Agree	1.00 (referent)	—	1.00 (referent)	—
Disagree	4.17 (3.16-5.50)	<.001	2.08 (1.52-2.85)	<.001
Not applicable	2.03 (1.39-2.99)	<.001	1.95 (0.43-8.80)	.39

[a]BC: breast cancer.

[b]Not applicable.

[c]BCS: breast cancer screening.

After adjustment for study site and participant characteristics (age, ethnicity, marital status, housing type, highest academic status, and family history of BC), BC awareness remained significant ($OR_{aware\ vs\ unaware}$ 2.88, 95% CI 2.19-3.80; *P*<.001) (Table S6 in Multimedia Appendix 4). The best combination of factors (according to the lowest Bayesian information criterion) that predicts screening motivation was BC awareness, marital status, current BC screening behavior, perceived importance of BC screening, and pain during mammography.

Demographic Differences in Participants Using the English and the Non-English Questionnaire

Participants who used the English version of the questionnaire (n=3819) were more likely to be BC aware, younger, and living in larger housing, and had higher educational attainment, as compared to participants who used the non-English questionnaire (n=350) (Table S7 in Multimedia Appendix 4). A total of 350 participants answered the Chinese or Malay versions of the questionnaire, of which 31 (9%) answered the free-text question on reasons for screening.

Differences in Characteristics Between Participants Who Provided and Did Not Provide Free-Text Explanation

Participants who provided and did not provide an explanation were not statistically different in age, ethnicity, housing type, and family history of BC (Table S8 in Multimedia Appendix 4). Participants who provided explanations were more likely to have never been married (*P*=.04) and to have higher educational attainment (*P*<.001). These participants differed from those who did not provide free-text responses regarding their current BC screening behavior (*P*=.01), perceived importance of screening (*P*=.002), perceived risk of BC (*P*=.01), and reporting that mammography was painful (*P*=.004) (Table S8 in Multimedia Appendix 4).

Biterm Topic Modeling Reveals Key Themes by Motivation Level

Of the 3819 participants who answered in English, 94% (n=3579) were motivated, and 6% (n=240) remained neutral about regular screening after the breast education session. Among the motivated participants, 1220 (34%) provided additional comments, compared to 153 (64%) among the neutral participants. Tables S9 and S10 in Multimedia Appendix 4 provide the sets of words derived from biterm topic modeling. The top and bottom 20 words are listed.

In the motivated participants, we selected 4 topics as repeats in top words occurred at 5 topics (Table S9 in Multimedia Appendix 4). “Early” and “health” were consistent across all 6 iterations, with “age” and “appointment” occurring 4 times, “better” twice, and “risk” and “painful” once. It was observed that “age” and “risk” belonged to the same set, while “appointment,” “better,” and “painful” were used commonly together. Using the first iteration (Table 3), topics observed were (1) benefits of early detection (with words “early,” “detection,” “better,” “cure,” and “prevention”), (2) health awareness (“health,” “regular,” “know,” “check,” and “important”), (3) BC risk (“age,” “risk,” “40,” “2,” and “safe”), where 40 refers to the age of 40 years and 2 refers to screening once every 2 years, (4) logistics and mammography experience (“appointment,” “regular,” “routine,” “schedule,” and “painful”).

In the neutral participants, we selected 2 topics as repeats in the top words that occurred at 3 topics (Table S10 in Multimedia Appendix 4). No word occurred in all iterations. “Pain” and “painful” were from the same set of words; “time,” “regular,” “health,” and “examination” were commonly from the same set of words. Using the first iteration (Table 3), unlike in the motivated participants, there was a distinction between (1) mammography experience with a focus on pain during a mammogram and (2) logistics, with a mention of “expensive” that was not observed in the top words from motivated participants. This is in line with our finding that participants who disagreed that mammography was expensive were more likely ($OR_{disagree\ vs\ agree}$ 1.94, 95% CI 1.50-2.50) to be willing to attend the screening posteducation session (Table S6 in Multimedia Appendix 4).

XSL•FO
RenderX

Table . Topics surfaced using biterm topic modeling on the free-text response following the question, "After knowing the above, would you be more willing to attend regular screening?" by willingness to screen: motivated and neutral. As different iterations with the same number of clusters produce variations in results, the first iteration is presented. Topic distribution (θ) provides the proportion of each topic in the entire corpus (dataset of additional comments). "k" refers to the number of clusters and was selected based on the occurrence of repeated topics within each iteration.

Cluster	Value (θ)	Top word	Subsequent 19 words	Topic
Motivated participants (n=1205)				
1	0.550	Early	Detection, better, cure, detect, prevention, chance, treatment, life, save, regular, survival, important, increase, live, recovery, believe, stage, always	Benefits of early detection
2	0.248	Health	Early, regular, good, know, check, important, detection, always, time, body, need, prevent, condition, family, age, risk, better, detect	Health awareness
3	0.103	Age	Risk, regular, 40, 2, safe, know, need, check, reduce, women, test, health, importance, 50, diagnose, healthy, aware, yearly	Breast cancer risk
4	0.099	Appointment	Health, better, regular, routine, follow, schedule, painful, women, easy, important, preventive, experience, hard, purpose, issue, patient, question, yearly	Logistics and mammogram experience
Neutral participants (n=145)				
1	0.569	Pain	Time, 2, once, cost, follow, risk, 1, booking, know, change, comfortable, frequent, abnormality, body, chance, depend, factor, increasing	Mammogram experience
2	0.431	Regular	Health, health care, painful, need, expensive, 40, attend, examination, women, schedule, time, change, decision, subsidize, appointment, ease, good, impact	Logistics

Subgrouping by BC Awareness Yields Similar Results to Motivation-Based Grouping

We repeated the analysis comparing BC-aware vs BC-unaware participants. Eighty-four percent (n=1010) of motivated women and 66% (n=95) of neutral women were BC-aware. When participants were grouped by BC awareness ($n_{BC\text{-}aware}$=1105 and $n_{BC\text{-}unaware}$=243), 2 topics were identified for BC-aware (ie, benefits of early detection and health awareness), similar to the motivated group (Table S11 in Multimedia Appendix 4). In BC-unaware participants, the main topic identified was the benefits of early detection (Table S12 in Multimedia Appendix 4). A second topic was not identified, as the word sets varied across the 6 iterations, likely reflecting the limited thematic diversity of responses. Conclusions about additional potential topics should be interpreted carefully.

More Motivated Participants Displayed More Positive Sentiments Toward Regular Screening

The minimum sentiment rating obtained in the analysis was −1.00, and the maximum rating was 1.65. The median and mean of the sentiment scores were 0.177 (IQR 0.000-0.202) and 0.202 (SD 0.327), respectively, indicating that positive sentiments were more prevalent than negative ones among the participants. The mean sentiment values and 95% CIs of each group are displayed in Figure 1. Motivated participants had significantly more positive sentiments toward regular screening than neutral participants, regardless of existing BC awareness.

XSL•FO
RenderX

Figure 1. Comparison of average sentiments across groups according to breast cancer (BC) awareness and willingness to screen (♦=mean). Median sentiment and IQR values of each subgroup are displayed at the bottom. Differences in sentiments were tested using the Wilcoxon test (***P<.001, **P<.01, and *P<.05). NS: not significant.

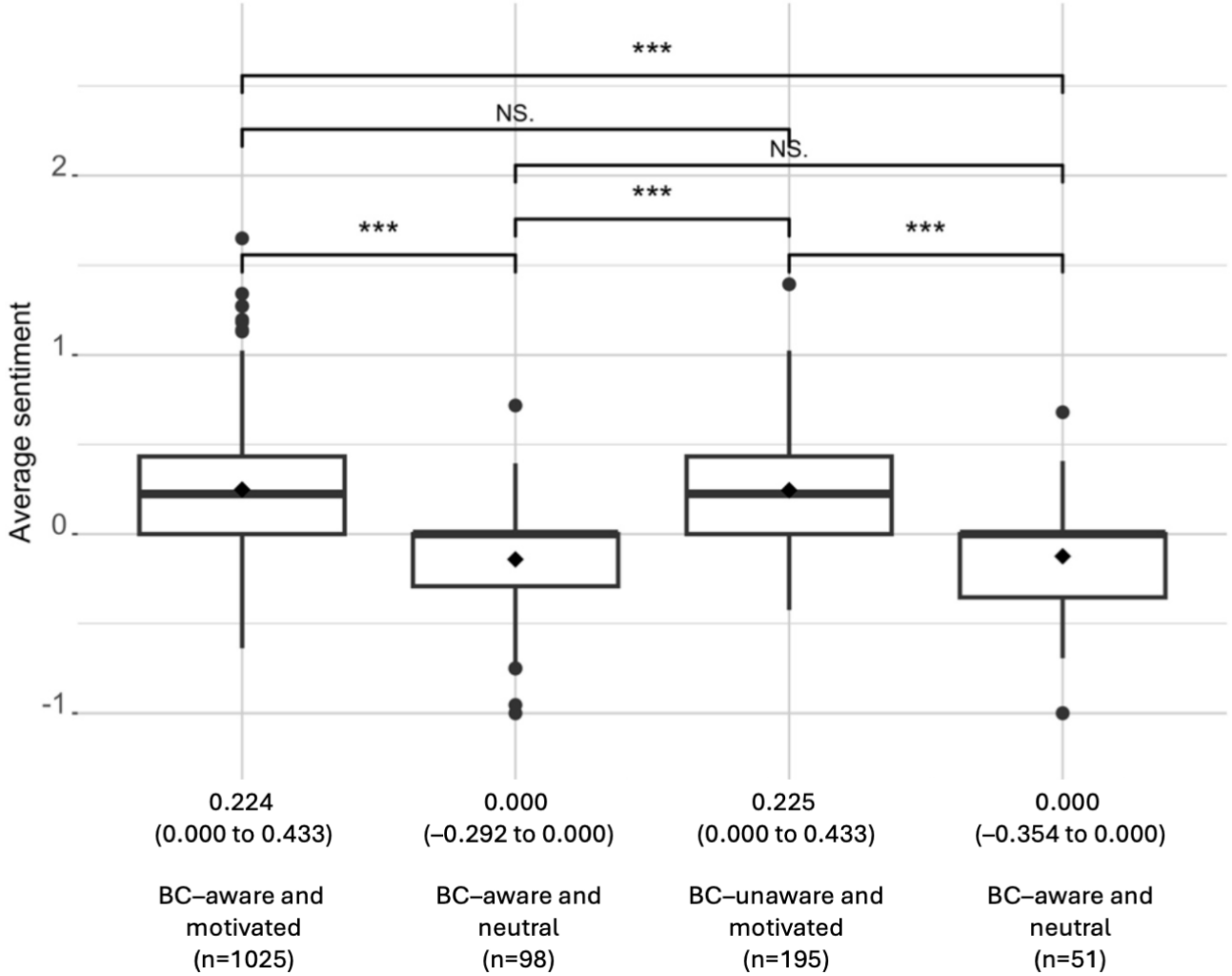

Word Frequencies are Indicative of Subgroup Sentiments

Words, expressed by at least 5% of each subgroup, were summarized and tabulated (Table 4). Motivated participants, both BC-aware and BC-unaware, displayed higher frequencies of words with positive sentiments (eg, “better,” “cure,” and “prevention”). Neutral participants who were BC-unaware displayed higher frequencies of negative sentiment words (eg, “painful,” “pain,” and “uncomfortable”), and only “painful” had over 5% frequency in participants who were BC-aware (Table 4).

XSL•FO
RenderX

Table . Top words mentioned in the free-text responses following the question, "After knowing the above, would you be more willing to attend regular screening?" by breast cancer awareness (BC-aware or BC-unaware) and willingness to attend screening (motivated or neutral), were concluded using biterm topic modeling. Words were mentioned by at least 5% of participants in each respective subgroup.

Subgroups	Positive words	Neutral words	Negative words
BC-aware[a] and motivated (n=1050)	Cure (n=74); better (n=136); important (n=80); detection (n=223); prevention (n=107); detect (n=68)	Early (n=362); health (n=101); regular (n=62)	—[b]
BC-aware and neutral (n=99)	—	Time (n=8); regular (n=8); health (n=6); cost (n=5)	Painful (n=15)
BC-unaware and motivated (n=206)	Better (n=24); important (n=19); detection (n=24); detect (n=22); prevention (n=17); prevent (n=13)	Early (n=48); health (n=20); regular (n=15); know (n=12)	—
BC-unaware and neutral (n=54)	Working (n=2); depends (n=2)	Time (n=6); necessary (n=3); regular (n=3); examination (n=3); cost (n=2); need (n=2); know (n=2)	Painful (n=8); troublesome (n=2); risk (n=2); pain (n=2); uncomfortable (n=2)

[a]BC: breast cancer.

[b]Not available.

Comparison of Words Across Subgroups

To gain deeper insights into the context of words and the subtleties surrounding future screening willingness, comments on the common and unique words from each subgroup were extracted. Here, we considered words that occurred in at least 5% of the subgroup.

Table S13 in Multimedia Appendix 4 provides the common words (n=12), along with the average sentiment of comments involving the words and some example quotes. Four words ("health," "important," "know," and "regular") had different sentence sentiments when used by the different subgroups. "Health" was a common word across 3 subgroups (BC-aware and motivated, BC-aware and neutral, and BC-unaware and motivated). When used by a participant in the BC-aware and motivated group, the sentence had a positive sentiment (eg, "health check is important to take care of one's health," "I want to stay healthy to see my girl grow up"). However, participants in the BC-aware and neutral group used it in sentences that had a negative sentiment ("I find it inconvenient to go for health screening," "the health care system does not allow women younger than 40 years to get subsidized breast screening"). Participants in the BC-unaware and motivated group were not significantly different from the BC-aware and motivated group (eg, "for my own health" and "it's important to keep track of one's health").

"Know" was used by BC-unaware participants, among whom the motivated participants used "know" in positive sentiment sentences (eg, "better to know if there is a sign of BC and get treatment" and "I thought that screening is expensive and did not know that there are subsidies"). This highlighted the utility of the education session in increasing BC awareness. However, participants who remained neutral used "know" to express negative sentiments (eg, "knowing the above does not affect my regular screening" and "better not to know"), which shows the limitations of the education exercise.

"Regular," a common word across all 4 subgroups, was used positively by motivated participants, for example, "important to do regular check-ups" by the BC-aware participants and "regular breast screening is important to detect any abnormality" by BC-unaware participants. "Regular" was used in negative sentences by the neutral participants, for example, "do not like too regular screening" by BC-aware participants and "I will be more open to regular screenings once I am 40" by BC-unaware participants.

Table 5 provides the words (n=4; "cure," "examination," "necessary," and "prevent") unique to the subgroup. While the words "examination" and "necessary" were mentioned by 3 participants (BC-unaware and neutral), the word "cure" was mentioned 74 times by BC-aware and motivated participants. The word "prevent" was mentioned 13 times by BC-unaware and motivated participants. It can be concerning that after the education session, participants used "prevent" regarding the occurrence of cancer (eg, "to prevent myself from any cancer").

XSL•FO
RenderX

Table . Use of unique sentiment words in context according to breast cancer awareness (BC-aware or BC-unaware) and willingness to screen (motivated or neutral).

Word	Behavior group	No of mentions in the group	Sentiment by participants in group, mean (SD; range)	Example quotes (random 3)
Cure	BC-aware[a] and motivated	74	0.803 (0.238; 0.177 to 1.650)	Prevention is better than cure; early detection will mean higher chance of survival; prevention is better than cure; prevention better than cure
Examination	BC-unaware and neutral	3	0.013 (0.022; 0.000 to 0.038)	If I do self-examination regularly it's not necessary; I will do self-breast examination to look out for any early signs; prefer to do self-examination for now
Necessary	BC-unaware and neutral	3	−0.031 (0.055; −0.094 to 0.000)	Still feel not necessary at this moment; depending on if it is necessary for me; if I do self-examination regularly it is not necessary
Prevent	BC-unaware and motivated	13	0.367 (0.345; −0.177 to 1.025)	it is very important need to prevent; no prevent breast cancer; prevention of breast cancer

[a]BC: breast cancer.

Discussion

Principal Findings

This study explored the use of sentiment analysis and biterm topic modeling to examine factors influencing participants' posteducation willingness to attend regular BC screening in a pilot study, using a specific question in a structured questionnaire with free-text questions to obtain participants' responses. Results from the free-text analysis indicate that most participants were BC aware and had positive sentiments toward regular screening. Major topics discovered were related to the benefits of early detection and health awareness in participants who were motivated to attend BC screening. In neutral participants, mammography experience, in particular, pain during the mammogram, was the key topic highlighted.

The focus of this study was to examine the public's views, sentiments, and concerns regarding their willingness to participate in regular BC screening following a BC education session. NLP of free-text responses provided an efficient method to filter the concerns that are prioritized by participants through biterm topic models and word count tabulations and allowed us to explore reasons by accessing the full entry. Key topics of interest to the participants may not surface in structured questionnaires, and the options are restricted by the researchers' choice of questions. In our case, the perceived importance of screening and perceived screening barriers such as inconvenience, pain, embarrassment, and cost were assessed using a limited range of available response options (5-point Likert scale or yes or no).

A review by Hisan et al [31] categorized the use of NLP in health care into clinical applications, such as clinical documentation, medical coding, clinical decision support, and patient engagement, and public health applications, including sentiment analysis, clinical trials, and disease surveillance, demonstrating how NLP enhances health care communication and outcomes. Thackeray et al [32] analyzed over 1.3 million tweets during Breast Cancer Awareness Month in 2012, revealing that most tweets were original posts by individuals, with organizations and celebrities focusing on fundraisers and early detection, while highlighting the need for strategic, ongoing social media engagement to promote preventive behaviors and maximize outreach.

Another study analyzing tweets by Nastasi et al [33] found that nonhealth care users frequently shared unsupported claims and expressed confusion about screening guidelines, highlighting the need for better public education on BC prevention and accurate information dissemination. These studies have demonstrated the immense potential of NLP in identifying current health care barriers and equipping governmental and health care institutions with information to deliver more targeted programs. Applying NLP to this study, we found that more motivated participants were more concerned with screening utility and understood the value of regular screening, while topics brought up by neutral participants were focused on the cost (financial, mental, and time) of screening.

The participation rate for routine mammography screening is poor in Singapore's population (<40%), a country that may be considered to be BC-aware (National Population Health Survey). In this study, we found 79% (n=3287) of our participants to be BC-aware. After completing the BC education questions, 94% (n=3908) reported being more motivated to attend screening. This is in contrast with only 63% (n=2628) who reported regular screening before the education session. BC awareness remained

XSL•FO
RenderX

a significant factor in participants' willingness to attend screening after adjusting for current screening frequency. Biterm topic modeling highlighted that the reinforcement of information on the benefits of early detection motivated the idea of attending screening.

In the area of barriers to screening, the main concern was pain among participants who were still not encouraged to screen. This echoes earlier findings that perceived pain or the experience of pain during mammography was a deterring factor for screening [34-37]. Although pain is a recognized barrier to mammography, our findings demonstrate that it remains a persistent concern. Future research should explore practical strategies to reduce discomfort and support more positive screening experiences. In contrast, other known barriers for screening, such as fatalism and cost, did not surface as concerns among our participants [4]. This may be due to multiple factors. Prior research studies on fatalism in BC were inconsistent with regard to ethnicity. While Straughan and Seow [38] found fatalism to be associated with mammography uptake in a Chinese-majority cross-sectional study, Goh et al [9] found fatalism not to be associated with mammography uptake among the Malay community [39]. In addition, some studies included older populations (aged >60 years in 2016) who were not included in BREATHE (aged 35-59 years) [40]. Furthermore, the higher income found in our population, as compared with previous research that included a larger portion of lower-income participants, may have reduced concerns about cost [37,41].

Contextual differences of the same word emphasize the need for subgroup analysis using multiple methods (examining words and sentences) to capture the information provided by free text. Notably, the word "health" was common across motivated and neutral groups. However, the context in which "health" was used differed. Motivated participants used the term to emphasize the importance of screening for maintaining one's health. This is similar to other studies demonstrating that innate health consciousness and perceived screening importance served as facilitators to screening attendance [42-44]. In contrast, neutral participants used "health" to highlight potential health risks associated with screening and their fatalistic attitudes toward the disease.

In our BC-unaware and neutral group, the words "examination" and "necessary" were used in combination with "regular," reflecting their perception that regular self-examination is sufficient to inform them of their health status and thus that mammography is unnecessary. This presents a gap in the understanding of mammography screening. A randomized controlled trial in Shanghai found that while breast self-examination accompanied by intensive instruction can help detect smaller fibroadenomas, it was not effective in reducing BC mortality [45]. On the other hand, Duffy et al [46] showed that mammography screening can lead to a 41% reduction in BC mortality over 10 years, along with a 25% decrease in the incidence of advanced BCs. Importantly, mammography can detect nonpalpable breast tumors, which would otherwise be missed by self-examination or clinical examination [47]. There is a need to inform the public that although regular breast self-examination is an important habit, it is not a substitute for mammography screening.

Comparison With Previous Studies

Our previous studies in Singapore, including structured surveys, focus group discussions, in-depth interviews, and discrete choice experiments, identified cost and perceived risk (ie, women's belief that they are at low risk of developing BC) as major determinants of participation in both standard and risk-based screening programs [11,12,42,48-50]. In contrast, the current NLP analysis revealed that cost did not emerge as a prominent factor, and perceived risk was not directly represented as a key factor, appearing instead through related concepts such as perceived screening importance. We believe that a strength of the data-driven NLP analytical approach is that it can uncover subtle insights from unstructured text that traditional survey or qualitative methods may overlook.

Considering our findings in light of the low mammography screening uptake rate in Singapore, addressing knowledge gaps and negative sentiments toward screening may increase women's motivation to participate. Evidence suggests that individualized discussions that acknowledge women's specific concerns can increase informed screening uptake, whereas the absence of such discussions or a lack of health care provider recommendation may discourage participation [51]. A personalized risk-based screening approach could provide a practical pathway to improve participation, as it allows for tailored communication and targeted support based on each woman's risk profile. Our experience with the BREATHE implementation study showed that many women underestimated their BC risk, and that the perceived risk of BC can change after receiving results from a personalized predictive risk assessment [52]. However, confidence in the predicted risk result is lower among those who underestimated their risk [52]. Nonetheless, implementing such personalized approaches at scale poses challenges, including resource constraints and the need for trained personnel to deliver individualized feedback.

Challenges of Using Online Education Tools

There were challenges in using a self-administered online questionnaire as an educational tool for BC screening. Participants appeared to be confused about the concept of early detection and cancer prevention, despite the education session emphasizing that mammography detects cancer early but does not prevent its occurrence. This was reflected in the participants' continued use of the terms "prevention" and "prevent" when referring to mammography screening (eg, "prevention of cancer" and "prevention is better than cure"), suggesting persistent misconceptions about the purpose of screening. However, even when the conceptual distinction between cancer prevention and early detection is clarified, broader barriers (eg, health literacy, digital access, emotional concerns, and health care system influences) may continue to influence screening intentions.

Limitations

We acknowledge several limitations of this study. There remains a gap in the validity and suitability of the tools for NLP [53]. In this study, we did not explore the accuracy of each tool in this specific use case. Compared to other studies exploring BC sentiments on social media, the availability of specific questions

XSL•FO
RenderX

in this study shapes participants' responses and provides more relevant details.

However, participants could decide whether they wanted to provide answers, and most did not, resulting in a small dataset for analysis. Additionally, answers provided by participants were short, often 1-2 sentences, plausibly due to the lack of question prompts. Future studies can include more questions with improved prompts to increase the robustness of the topics covered and obtain deeper responses from participants.

We acknowledge that domain-specific sentiment lexicons may provide added nuance for clinically focused text. However, given the lay language and short responses in our dataset, the use of a general-purpose sentiment lexicon was appropriate.

This study only included the opinions of English-speaking participants and, thus, may not capture the diverse perspectives of all women. The cohort, drawn from a risk-based screening study, may also not fully represent the general population of screening-eligible women.

Our findings are based on participants' self-reported willingness to screen following an educational session and may not reflect actual screening behavior. The BC education session likely influenced participants' responses, particularly in shaping positive attitudes and highlighting certain themes such as early detection. As such, the topics and sentiments derived from the free-text responses may reflect immediate posteducation reactions rather than pre-existing beliefs, representing a potential source of bias. Future studies could incorporate baseline measurements or control groups to disentangle the effects of educational interventions on sentiment and topic expression.

Conclusion

We integrated quantitative surveys with NLP to reveal emotional dimensions of screening attitudes. Unlike existing studies focused on knowledge or intent, we showed that identical screening concepts are emotionally framed differently by motivated vs neutral participants. In practice, these findings support the need for emotionally tailored BC education and prevention strategies.

Acknowledgments

We would like to thank our dedicated research and administrative staff, Pooi Yee Wong, Ganga Devi D/O Chandrasegran, Nabilah Binte Supiee, Siti Zulyqha Binte Yazid, Alleza Joeay Balbanero Aquino, Pei Xuan Lim, Jolene Lu Yee Poh, Brenna Jing Jie Quah, Qian Ning Peh, Chun Mei Wang, Cara Wee Ying Wong, Kimiie, Wei Lin Chia, Yi Lin Chen, Jinan May Loewen, Hui Min Lau, Varshaa D/O Saravanan, Vannevia Jedidiah Shi Tong Foo, Nurfilya Binte Hamdil, Hian Ching Ng, Yen Shing Yeoh, Amanda Tse Woon Ong, Jing Jing Hong, and Siew Li Tan, for their contributions to the planning, preparation, and execution of BREAst Screening Tailored for HEr (BREATHE). We would also like to thank Dr Chuan Chien Tan for assisting in the initial setup of the project, the doctors from the Department of Obstetrics and Gynecology at the National University Hospital, Dr Judith Shan Lin Ong and Dr Susan Jane Sinclair Logan, for allowing our team to conduct recruitment at Jade Clinic, and the radiologists from the Department of Diagnostic Imaging, Breast Imaging, at the National University Hospital, Dr Pooja Jagmohan, Dr Premilla Gopinathan Pillay, Dr Yonggeng Goh, Dr Han Yang Ong, and Dr Abrar, for helping with the density reading scores in our BREATHE cohort. Portions of this manuscript were edited for language and clarity with the assistance of ChatGPT (OpenAI; GPT-4). All content, interpretations, and conclusions are solely the work of the authors. Artificial intelligence (AI)-generated content was reviewed and revised by the authors to ensure accuracy, relevance, and scientific integrity.

Funding

This study was funded by the JurongHealth Fund (reference no JHF-20-RE-003), Agency for Science, Technology and Research (A*STAR), and the Precision Health Research Singapore Clinical Implementation Pilot (PRECISE CIP) Fund. MH is supported by the JurongHealth Fund, PRECISE CIP Fund, the Breast Cancer Prevention Programme under the Saw Swee Hock School of Public Health Programme of Research Seed Funding (SSHSPH-Res-Prog-BCPP), the Breast Cancer Screening Prevention Programme under the Yong Loo Lin School of Medicine (NUHSRO/2020/121/BCSPP/LOA), the National University Cancer Institute Singapore (NCIS) Centre Grant Programme (CGAug16M005), and the Asian Breast Cancer Research Fund. The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The data generated by this study are owned by the providing institutions (Ng Teng Fong General Hospital [NTFGH], National University Hospital [NUH], Alexandra Hospital [AH], National University Polyclinics [NUP], and Jurong Medical Center [JMC]). Data may be obtained upon a reasonable request to the principal investigator, MH. The data are not publicly available due to privacy and/or ethical restrictions. Legal agreements will need to be drawn up between data requesters and providers for access to the deidentified data. The proposed studies need to comply with Singapore's laws and regulations regarding human biomedical research and clinical investigation, including the Declaration of Helsinki, International Good Clinical Practice Guidelines, and Good Clinical Practice guidelines by Singapore's Health Science Authority and the Ministry of Health.

XSL•FO
RenderX

Authors' Contributions

PJH and ZLL are joint first authors. J Li and MH conceptualized the study. ZLL and PJH developed the methodology. J Liu, NKMR, YJC, HLT, SAG, HBO, CHC, SCK, ZPZ, DLSO, STQ, SW, and MH curated the data. ZLL and PJH conducted the formal analysis. PJH, ZLL, and J Li prepared the original draft. ZLL, PJH, HBO, CHC, SCK, ZPZ, DLSO, STQ, SW, PTCI, J Liu, NKMR, YJC, HLT, SAG, MH, and J Li reviewed and edited the manuscript. J Li supervised the study. J Liu, NKMR, YJC, and HLT managed the project administration. MH and JL acquired the funding.

Conflicts of Interest

None declared.

Multimedia Appendix 1
BREATHE online structured first visit questionnaire, which includes factors linked to breast cancer and related conditions, such as demographic, lifestyle, and reproductive characteristics, prior treatments, and other environmental exposures.
[PDF File, 1015 KB - jmir_v28i1e78439_app1.pdf]

Multimedia Appendix 2
BREATHE education questionnaire.
[PDF File, 742 KB - jmir_v28i1e78439_app2.pdf]

Multimedia Appendix 3
Flowchart of participant selection from individuals enrolled in BREATHE.
[DOCX File, 18 KB - jmir_v28i1e78439_app3.docx]

Multimedia Appendix 4
Detailed reference classifications for sociodemographic variables (housing and educational attainment; Tables S1) and full natural language processing (NLP) pipeline documentation, including spelling corrections (Table S2), stemming rules (Table S3), stop word exclusions (Table S4), and manually adjudicated sentiment ratings (Table S5), alongside complete logistic regression models examining patient-level predictors of post-education screening willingness (Table S6), attrition analyses comparing English-language respondents to nonrespondents (Table S7) and free-text contributors to noncontributors (Table S8), exhaustive biterm topic model outputs across candidate topic solutions (k = 1–5) stratified by motivational response (more motivated vs. neutral; Tables S9–S10) and education session performance (>80% vs. ≤80%; Tables S11–S12), and a contextualized concordance analysis of high-frequency sentiment terms disaggregated by breast cancer awareness status and screening motivation (Table S13).
[XLSX File, 110 KB - jmir_v28i1e78439_app4.xlsx]

References

1. Ghoncheh M, Pournamdar Z, Salehiniya H. Incidence and mortality and epidemiology of breast cancer in the world. Asian Pac J Cancer Prev 2016;17(S3):43-46. [doi: 10.7314/apjcp.2016.17.s3.43] [Medline: 27165206]
2. Coleman C. Early detection and screening for breast cancer. Semin Oncol Nurs 2017 May;33(2):141-155. [doi: 10.1016/j.soncn.2017.02.009] [Medline: 28365057]
3. Wild CP, Weiderpass E, Stewart BW, editors. World Cancer Report: Cancer Research for Cancer Prevention: International Agency for Research on Cancer (IARC); 2020.
4. Rajendram P, Singh P, Han KT, et al. Barriers to breast cancer screening in Singapore: a literature review. Ann Acad Med Singap 2022 Aug;51(8):493-501. [doi: 10.47102/annals-acadmedsg.2021329] [Medline: 36047524]
5. Epidemiology & Disease Control Division and Policy RSG, Ministry of Health and Health Promotion Board, Singapore: National Population Health Survey; 2021.
6. Momenimovahed Z, Tiznobaik A, Taheri S, Hassanipour S, Salehiniya H. A review of barriers and facilitators to mammography in Asian women. Ecancermedicalscience 2020;14:1146. [doi: 10.3332/ecancer.2020.1146] [Medline: 33343705]
7. Teo CT, Yeo YWS, Lee SC. Screening mammography behavior and barriers in Singaporean Asian women. Am J Health Behav 2013 Sep;37(5):667-682. [doi: 10.5993/AJHB.37.5.11] [Medline: 23985290]
8. Straughan PT, Seow A. Barriers to mammography among Chinese women in Singapore: a focus group approach. Health Educ Res 1995 Dec;10(4):431-441. [doi: 10.1093/her/10.4.431]
9. Goh SA, Lee JK, Seh WY, et al. Multi-level determinants of breast cancer screening among Malay-Muslim women in Singapore: a sequential mixed-methods study. BMC Womens Health 2022 Sep 19;22(1):383. [doi: 10.1186/s12905-022-01972-y] [Medline: 36123600]

XSL•FO
RenderX

10. Surendran S, Foo CD, Tan DHY, et al. Understanding barriers and facilitators of breast and cervical cancer screening among Singapore women: a qualitative approach. Asian Pac J Cancer Prev 2023 Mar 1;24(3):889-895. [doi: 10.31557/APJCP.2023.24.3.889] [Medline: 36974542]
11. Kamila KA, Liow JJK, Giam F, et al. Spousal perceptions and their role in promoting breast cancer screening: a focus group study. BMJ Open 2025 Mar 12;15(3):e086340. [doi: 10.1136/bmjopen-2024-086340] [Medline: 40074274]
12. Liow JJK, Lim ZL, Ho PJ, et al. Attitudes and barriers to mammography screening in Singaporean women through the eyes of their adult children: a focus group study. SSM - Qualitative Research in Health 2022 Dec;2:100168. [doi: 10.1016/j.ssmqr.2022.100168]
13. Glasman LR, Albarracín D. Forming attitudes that predict future behavior: a meta-analysis of the attitude-behavior relation. Psychol Bull 2006 Sep;132(5):778-822. [doi: 10.1037/0033-2909.132.5.778] [Medline: 16910754]
14. Lee JW, Jones PS, Mineyama Y, Zhang XE. Cultural differences in responses to a likert scale. Res Nurs Health 2002 Aug;25(4):295-306. [doi: 10.1002/nur.10041]
15. Westland JC. Information loss and bias in likert survey responses. PLoS ONE 2022;17(7):e0271949. [doi: 10.1371/journal.pone.0271949] [Medline: 35901102]
16. Small ML, Cook JM. Using interviews to understand why: challenges and strategies in the study of motivated action. Sociol Methods Res 2023 Nov;52(4):1591-1631. [doi: 10.1177/0049124121995552]
17. Adams A, Cox AL. Questionnaires, in-depth interviews and focus groups. In: Research Methods for Human Computer Interaction: Cambridge University Press:17-34. [doi: 10.1017/CBO9780511814570.003]
18. Khaleghparast S, Maleki M, Hajianfar G, et al. Development of a patients' satisfaction analysis system using machine learning and lexicon-based methods. BMC Health Serv Res 2023 Mar 23;23(1):280. [doi: 10.1186/s12913-023-09260-7] [Medline: 36959630]
19. Khanbhai M, Anyadi P, Symons J, Flott K, Darzi A, Mayer E. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. BMJ Health Care Inform 2021 Mar;28(1):e100262. [doi: 10.1136/bmjhci-2020-100262]
20. Mehta N, Pandit A. Concurrence of big data analytics and healthcare: a systematic review. Int J Med Inform 2018 Jun;114:57-65. [doi: 10.1016/j.ijmedinf.2018.03.013] [Medline: 29673604]
21. Yim WW, Yetisgen M, Harris WP, Kwan SW. Natural language processing in oncology: a review. JAMA Oncol 2016 Jun 1;2(6):797-804. [doi: 10.1001/jamaoncol.2016.0213] [Medline: 27124593]
22. Clark EM, James T, Jones CA, Alapati A, Ukandu P, Danforth CM, et al. A sentiment analysis of breast cancer treatment experiences and healthcare perceptions across Twitter. arXiv. Preprint posted online on May 25, 2018 URL: https://arxiv.org/abs/1805.09959 [accessed 2026-02-19]
23. Döbrössy B, Girasek E, Susánszky A, Koncz Z, Győrffy Z, Bognár VK. "Clicks, likes, shares and comments" a systematic review of breast cancer screening discourse in social media. PLoS One 2020;15(4):e0231422. [doi: 10.1371/journal.pone.0231422] [Medline: 32294139]
24. Levitt HM, Bamberg M, Creswell JW, Frost DM, Josselson R, Suárez-Orozco C. Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):26-46. [doi: 10.1037/amp0000151] [Medline: 29345485]
25. Liu J, Ho PJ, Tan THL, et al. BREAst screening Tailored for HEr (BREATHE)—a study protocol on personalised risk-based breast cancer screening programme. PLoS ONE 2022;17(3):e0265965. [doi: 10.1371/journal.pone.0265965]
26. Ho PJ, Liu J, Goh SA, et al. Profile of the BREATHE cohort for risk-based breast screening in Singapore. Sci Rep 2025 Oct 10;15(1):35424. [doi: 10.1038/s41598-025-19205-1] [Medline: 41073454]
27. Silge J, Robinson D. tidytext: text mining and analysis using tidy data principles in R. J Open Source Softw 2016;1(3):37 [FREE Full text] [doi: 10.21105/joss.00037]
28. Cheng X, Yan X, Lan Y, Guo J. BTM: topic modeling over short texts. IEEE Trans Knowl Data Eng 2014;26(12):2928-2941. [doi: 10.1109/TKDE.2014.2313872]
29. Yan X, Guo J, Lan Y, Cheng X. A biterm topic model for short texts. 2013 Presented at: WWW '13: Proceedings of the 22nd international conference on World Wide Web; May 13-17, 2013; Rio de Janeiro, Brazil p. 1445-1456. [doi: 10.1145/2488388.2488514]
30. Zhu K, Weng J, Tu B. A comparative study on short-text topic modeling in online collaborative learning platforms: application and performance analysis of BTM and BERTopic. 2024 Presented at: ICAIE '24: Proceedings of the 2024 3rd International Conference on Artificial Intelligence and Education; Nov 22-24, 2024 URL: https://dl.acm.org/doi/proceedings/10.1145/3722237 [accessed 2026-02-19] [doi: 10.1145/3722237.3722309]
31. Hisan UK, Chaaben AB, Adeleye AO, Kaushik A. Impact of natural language processing on healthcare quality: a review of clinical and public health applications. JAAHM 2022 Jan;6(1):19-42 [FREE Full text]
32. Thackeray R, Burton SH, Giraud-Carrier C, Rollins S, Draper CR. Using Twitter for breast cancer prevention: an analysis of breast cancer awareness month. BMC Cancer 2013 Oct 29;13(1):508. [doi: 10.1186/1471-2407-13-508] [Medline: 24168075]

XSL•FO
RenderX

33. Nastasi A, Bryant T, Canner JK, Dredze M, Camp MS, Nagarajan N. Breast cancer screening and social media: a content analysis of evidence use and guideline opinions on Twitter. J Cancer Educ 2018 Jun;33(3):695-702. [doi: 10.1007/s13187-017-1168-9] [Medline: 28097527]
34. Jones T, Wisdom-Chambers K, Freeman K, Edwards K. Barriers to mammography screening among Black women at a community health center in South Florida, USA. Med Res Arch 2023 Apr;11(4). [doi: 10.18103/mra.v11i4.3814] [Medline: 37475892]
35. Gong J, Kampadellis G, Kong Q, Spijker W. Factors determining non-attendance in breast cancer screening among women in the Netherlands: a national study. Health Promot Int 2023 Jun 1;38(3). [doi: 10.1093/heapro/daac009]
36. Williams WM, Fu MR. Understanding breast cancer disparities affecting Black women: obtaining diverse perspectives through focus groups comprised of patients, providers, and others. J Community Health 2023 Oct;48(5):834-839. [doi: 10.1007/s10900-023-01227-3] [Medline: 37131066]
37. Lim SK, Teo XL, Ng JL, Li FX, Tan SM. A survey on Singaporean women's knowledge, perception and practices of mammogram screening. Ann Acad Med Singap 2015 Sep;44(9):317-325. [Medline: 26584660]
38. Straughan PT, Seow A. Fatalism reconceptualized: a concept to predict health screening behavior. Journal of Gender, Culture, and Health 1998 Jun;3(2):85-100. [doi: 10.1023/A:1023278230797]
39. Shaw T, Ishak D, Lie D, et al. The influence of Malay cultural beliefs on breast cancer screening and genetic testing: a focus group study. Psychooncology 2018 Dec;27(12):2855-2861. [doi: 10.1002/pon.4902] [Medline: 30264524]
40. Wee LE, Lim LY, Koh GCH. Two sides of the coin: a qualitative study of patient and provider perspectives on colorectal, breast and cervical cancer screening in a low-income Asian community. Proc Singap Healthc 2016 Jun;25(2):80-91. [doi: 10.1177/2010105815616404]
41. Seetoh T, Siew WF, Koh A, et al. Overcoming barriers to mammography screening: a quasi-randomised pragmatic trial in a community-based primary care setting. Ann Acad Med Singap 2014 Dec;43(12):588-594. [Medline: 25588917]
42. Lim ZL, Ho PJ, Khng AJ, et al. Mammography screening is associated with more favourable breast cancer tumour characteristics and better overall survival: case-only analysis of 3739 Asian breast cancer patients. BMC Med 2022 Aug 4;20(1):239. [doi: 10.1186/s12916-022-02440-y] [Medline: 35922814]
43. Safizade H, Amirzadeh N, Mangolian Shahrbabaki P. Motivational factors for breast cancer screening behaviors in Iranian women: a qualitative study. Asian Pac J Cancer Prev 2020 Oct 1;21(10):3109-3114. [doi: 10.31557/APJCP.2020.21.10.3109] [Medline: 33112574]
44. Jung SM, Jo HS. Intrinsic motivation factors based on the self-determinant theory for regular breast cancer screening. Asian Pac J Cancer Prev 2014;15(23):10101-10106. [doi: 10.7314/apjcp.2014.15.23.10101] [Medline: 25556433]
45. Thomas DB, Gao DL, Ray RM, et al. Randomized trial of breast self-examination in Shanghai: final results. J Natl Cancer Inst 2002 Oct 2;94(19):1445-1457. [doi: 10.1093/jnci/94.19.1445] [Medline: 12359854]
46. Duffy SW, Tabár L, Yen AMF, et al. Mammography screening reduces rates of advanced and fatal breast cancers: results in 549,091 women. Cancer 2020 Jul 1;126(13):2971-2979. [doi: 10.1002/cncr.32859] [Medline: 32390151]
47. Welch HG, Prorok PC, O'Malley AJ, Kramer BS. Breast-cancer tumor size, overdiagnosis, and mammography screening effectiveness. N Engl J Med 2016 Oct 13;375(15):1438-1447. [doi: 10.1056/NEJMoa1600249] [Medline: 27732805]
48. Liow JJK, Lim ZL, Sim TMY, et al. "It Will Lead You to Make Better Decisions about Your Health"-a focus group and survey study on women's attitudes towards risk-based breast cancer screening and personalised risk assessments. Curr Oncol 2022 Nov 25;29(12):9181-9198. [doi: 10.3390/curroncol29120719] [Medline: 36547133]
49. Lim ZL, Giam F, Wong RYX, Liow JJK, McCrickerd K, Li J. Do women with a history of breast cancer recommend risk-based breast cancer screening? An in-depth interview study. Front Psychol 2025;16:1414099. [doi: 10.3389/fpsyg.2025.1414099] [Medline: 40330299]
50. Wang Y, Ho PJ, Mou L, Li J. Women's preferences for testing to predict breast cancer risk – a discrete choice experiment. J Transl Med 2025;23(1). [doi: 10.1186/s12967-025-06119-9]
51. Georga D, Zartaloudi A, Saridi M, et al. Factors determining women's attitudes and knowledge toward breast cancer screening: a systematic review. Healthcare (Basel) 2025 Jul 4;13(13):1605. [doi: 10.3390/healthcare13131605] [Medline: 40648630]
52. Ho PJ, Goh SA, Goh SSN, et al. Impact of personalised risk predictions on breast cancer risk perceptions: insights from the BREATHE study. J Transl Med 2025 May 8;23(1):517. [doi: 10.1186/s12967-025-06515-1] [Medline: 40340901]
53. Gohil S, Vuik S, Darzi A. Sentiment analysis of health care tweets: review of the methods used. JMIR Public Health Surveill 2018 Apr 23;4(2):e43. [doi: 10.2196/publichealth.5789] [Medline: 29685871]

Abbreviations

BC: breast cancer
BREATHE: Breast Screening Tailored for Her
NLP: natural language processing
OR: odds ratio

XSL•FO
RenderX

Edited by S Brini; submitted 02.Jun.2025; peer-reviewed by I Moreira, L Raymond Guo, PC Lin; accepted 25.Dec.2025; published 10.Mar.2026.

Please cite as:
Ho PJ, Lim ZL, Liu J, Mohamed Riza NK, Chew YJ, Lim YY, Tan HL, Goh SA, Oh HB, Chin CH, Kwek SC, Zhang ZP, Ong DLS, Quek ST, Wijerathne S, Iau PTC, Hartman M, Li J
Breast Cancer Screening Knowledge and Sentiments in Singaporean Women: Mixed Methods Study Using Topic Modeling, Sentiment Analysis, and Structured Questionnaire Data
J Med Internet Res 2026;28:e78439
URL: https://www.jmir.org/2026/1/e78439
doi:10.2196/78439

© Peh Joo Ho, Zi Lin Lim, Jenny Liu, Nur Khaliesah Mohamed Riza, Ying Jia Chew, Yi Ying Lim, Hui Ling Tan, Su-Ann Goh, Han Boon Oh, Chi Hui Chin, Sing Cheer Kwek, Zhi Peng Zhang, Desmond Luan Seng Ong, Swee Tian Quek, Sujith Wijerathne, Philip Tsau Choong Iau, Mikael Hartman, Jingmei Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Integration of Digital Therapeutics Into Occupational Rehabilitation in Germany: Multilevel Simulation Study

Christoph Wallner[1], PhD; Sonja Verena Schmidt[1], MD; Felix Reinkemeier[1], MD; Dagmar-Maria Dick[2], MD; Marcus Lehnhardt[1], MD; Marius Drysch[1], MD

[1]Department for Plastic and Hand Surgery, University Hospital Bergmannsheil Bochum, Bürkle de la Camp-Platz 1, Bochum, Germany

[2]BG BAU – German Social Accident Insurance for the Construction Sector, Wuppertal, Germany

Corresponding Author:
Christoph Wallner, PhD
Department for Plastic and Hand Surgery, University Hospital Bergmannsheil Bochum, Bürkle de la Camp-Platz 1, Bochum, Germany

Abstract

Background: Expenditures for physiotherapy and extended outpatient physiotherapy (EAP) are increasing within Germany's statutory accident insurance system (Berufsgenossenschaften), placing growing pressure on rehabilitation capacity and timely access to care. Digital health applications (DiGAs) are reimbursable nationwide and represent a novel component of routine rehabilitation pathways. However, their real-world system-level and economic effects in occupational rehabilitation remain insufficiently understood.

Objective: This study aimed to evaluate how the integration of DiGAs into occupational rehabilitation pathways may influence costs, service capacity, and waiting times within routine care delivered by 5 German statutory accident insurance funds that cover 25.9 million insured individuals.

Methods: Aggregated administrative data from 5 Berufsgenossenschaften (fiscal years 2023 - 2024) were analyzed using a multilevel simulation framework combining (1) probabilistic cost-consequence modeling with Monte Carlo simulation (10,000 iterations), (2) an adherence-based adoption funnel distinguishing long-term engaged users (15%) and short-term users (85%) based on German claims data, and (3) a calibrated M/M/1 queuing model validated through discrete event simulation to estimate the effects on waiting times and system capacity. Primary outcomes included net financial impact, break-even thresholds, and changes in access-related performance metrics.

Results: Combined physiotherapy and EAP expenditures reached €404 million (€1=US $1.18) in 2024, increasing by 10.1% year-over-year. The primary simulation (N=10,000 iterations) indicated mean annual net savings of €18.4 million (median €17.9 million) with a 90.7% probability of cost savings (95% uncertainty range: net cost of €8 million to net savings of €47.7 million). After incorporating adherence dynamics, the projected mean net savings were €16.2 million (95% CI €5-€29.8 million), corresponding to a 100% probability of positive financial impact within the modeled parameter space. Cost neutrality was maintained for DiGA prices up to €617.8 per prescription, nearly 40% above the base-case assumption of €450, indicating substantial economic robustness. Queuing analyses demonstrated that modest reductions in therapeutic demand decreased mean waiting times from 17.3 to 12.8 days (−26%), equivalent to approximately 120,000 cumulative patient waiting days saved annually across 26,705 EAP patients. The validation of discrete event simulation confirmed the magnitude and direction of analytic estimates.

Conclusions: Under conservative assumptions, integrating digital therapeutics into occupational rehabilitation pathways is likely to generate both economic benefits and substantial system-level capacity gains. The break-even threshold of €617.80 per prescription provides a wide margin for pricing policy. Beyond cost effects, DiGAs may function as scalable capacity tools that alleviate systemic bottlenecks and improve timely access to rehabilitation services in capacity-constrained systems.

(J Med Internet Res 2026;28:e93793) doi:10.2196/93793

KEYWORDS

digital therapeutics; digital health applications; simulation modeling; health system capacity; occupational rehabilitation; access to care; cost-consequence analysis

XSL•FO
RenderX

Introduction

Background and Rationale

The German statutory accident insurance system, managed by the Berufsgenossenschaften, represents a unique institutional framework responsible for the prevention, rehabilitation, and compensation of occupational accidents and diseases for over 85 million insured individuals [1]. Against a backdrop of high and rising expenditures—rehabilitation spending across Germany reached €43.6 billion (€1=US $1.18) in 2022 [2]—traditional therapies such as physiotherapy and extended outpatient physiotherapy (EAP) represent a substantial and growing cost driver [3,4]. Given the scale of the Berufsgenossenschaften system and its ring-fenced financing structure, Germany provides an internationally distinctive setting to assess the system-level impact of novel care models.

A pivotal development in Germany was the establishment of a formal regulated pathway for digital health applications (known by their German acronym, DiGAs) introduced by the 2019 Digital Healthcare Act (Digitale-Versorgung-Gesetz), which enables the prescription and reimbursement of digital therapeutics within routine care [5]. While DiGA uptake has accelerated since the first DiGAs were introduced into the German statutory health insurance system in 2020—cumulative reimbursements reached €234 million by the end of 2024—evaluation has largely emphasized clinical efficacy and patient-relevant structural outcomes [6,7]. A critical evidence gap persists around DiGA's economic and operational effects in high-volume, high-cost rehabilitation pathways.

Digital Physiotherapy: International Evidence and the German Context

International studies suggest that digital physiotherapy and telehealth can deliver outcomes comparable to face-to-face care at a lower cost, primarily through reduced resource use and travel [8]; for example, a Swedish analysis reported digital osteoarthritis care at approximately 25% of the cost of in-person care [9]. A large-scale health technology assessment found that virtual musculoskeletal solutions delivered clinically meaningful improvements in pain and function compared with usual care across a range of musculoskeletal disorders [10]. Recent meta-analytic evidence further supports the cost-effectiveness of telerehabilitation relative to traditional in-person rehabilitation [11]. Yet, within Germany, economic evaluations indicate that the DiGA value is sensitive to price and evidentiary strength, with some analyses showing only marginal benefits [12]. To our knowledge, no prior study has quantified the potential cost and capacity effects of DiGA specifically for physiotherapy and EAP within the Berufsgenossenschaften context.

Study Aim

To address this gap, we developed a multistage quantitative modeling framework integrating probabilistic cost-consequence analysis, claims-based adherence stratification, and calibrated queuing theory to provide a robust, multidimensional assessment of whether, and under what assumptions, DiGA can reduce costs while expanding effective care capacity in Germany's occupational rehabilitation system. This study, therefore, examines digital therapeutics not only as clinical interventions but also as system-level tools capable of reshaping rehabilitation capacity and access dynamics.

Methods

Study Design and Modeling Framework

To evaluate the system-level effects of DiGA integration into occupational rehabilitation pathways, we developed a multilevel simulation framework combining economic, behavioral, and operational modeling approaches. Rather than performing a traditional cost-use evaluation, the study was designed as a system-level cost-consequence and implementation simulation assessing how digital therapeutics may influence real-world system performance, including resource use, service capacity, and access to care.

The framework integrates three interconnected components: (1) administrative expenditure data from German statutory accident insurance funds (Berufsgenossenschaften), (2) an adoption and adherence model reflecting real-world patient engagement patterns, and (3) system dynamics modeling using probabilistic simulation and queuing-based capacity analysis. A discrete event simulation (DES) was implemented to validate time-efficiency estimates under stochastic conditions. An overview of the conceptual structure and analytical workflow is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Conceptual framework of the multilevel simulation model evaluating digital health integration into occupational rehabilitation. Administrative data from German statutory accident insurance funds (Berufsgenossenschaften) provide aggregated information on physiotherapy and extended outpatient physiotherapy utilization and expenditures. These data inform an adoption and adherence funnel representing real-world eligibility, uptake, and engagement with digital health applications. Economic and operational effects are estimated using probabilistic simulation modeling substitution and efficiency mechanisms. Changes in therapeutic demand are propagated into a queuing model (M/M/1) to estimate system utilization and waiting-time dynamics, with results validated through discrete event simulation. The integrated framework generates system-level outcomes, including net costs, waiting times, and access-to-care effects. EAP: extended outpatient physiotherapy.

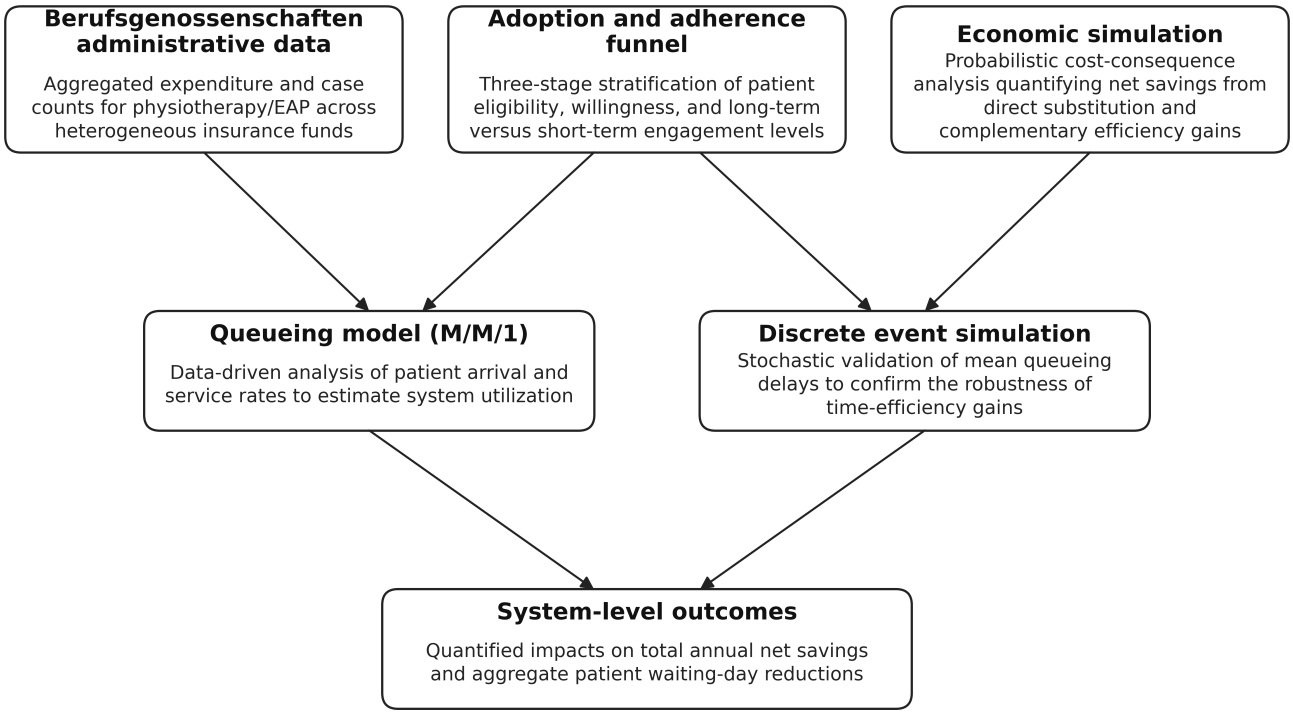

Data Source and Preparation

We analyzed aggregated, nonpublic administrative summaries from a purposive sample of 5 German statutory accident insurance funds (Berufsgenossenschaften) for fiscal years 2023 to 2024, covering 25.9 million insured individuals. The sample spans heterogeneous risk profiles—for example, Berufsgenossenschaften Bau (construction; high-to-severe trauma exposure) and BGW (health and welfare; higher burden of chronic musculoskeletal conditions)—to improve external validity across occupational settings. For each Berufsgenossenschaften, we obtained annual expenditure and case counts for physiotherapy and EAP, as well as the number of insured employees.

Reporting categories were harmonized to common physiotherapy and EAP definitions using fund fee schedules. Quality checks covered nonnegativity and plausible ranges, reconciliation of physiotherapy and EAP subtotals with Berufsgenossenschaften aggregates, and cross-year consistency. Monetary values are nominal euros (no price adjustment). Missing fields were not imputed; the affected strata were excluded from rate calculations but retained in totals when available. Derived indicators included spend per case, cases per 1000 insured, year-over-year growth, and the arrival-rate proxy used in queuing analysis. The dataset contained no individual-level or identifiable information.

Economic Modeling Framework

We developed a cost-consequence model from the perspective of the statutory accident insurance funds (reference year: 2024). Net annual savings from DiGA integration into physiotherapy and EAP are the sum of two mutually exclusive mechanisms: (1) direct substitution of conventional therapy with a DiGA and (2) complementary use of a DiGA as an add-on that increases efficiency in nonsubstituted cases. Physiotherapy and EAP are modeled separately and then aggregated.

Model 1 (Direct substitution): A proportion R_{Sub} of conventional cases is fully replaced by a DiGA; savings per therapy type = avoided conventional cost − DiGA cost. Model 2 (Complementary efficiency): Among the remaining $(1 - R_{Sub})$ cases, a share R_{Comp} receives a DiGA add-on, yielding an efficiency gain E_{Gain} (fractional reduction in standard care cost). Net savings per case equal (E_{Gain} × standard care cost) − DiGA cost. By construction, the 2 mechanisms apply to disjoint case partitions, precluding double counting. Adoption and engagement follow the adherence model described below.

Probabilistic Sensitivity Analysis

To propagate parameter uncertainty, we ran a Monte Carlo probabilistic sensitivity analysis (PSA) with 10,000 iterations. In each iteration, parameters were sampled from their prespecified distributions (Table 1), the model was evaluated for physiotherapy and EAP, and the outputs were aggregated across Berufsgenossenschaften. We report the mean, median, and 95% uncertainty range (2.5th-97.5th percentiles), as well as the probability of any net savings. Heterogeneity was assessed by repeating the PSA per Berufsgenossenschaften to obtain Berufsgenossenschaften-specific means and uncertainty ranges.

XSL•FO
RenderX

Table . Model parameters and evidence-based probability distributions for the probabilistic sensitivity analysis.

Parameter	Symbol	Description	Distribution	Min, Mode, Max	Source
DiGA[a] cost	C_{DiGA}	Cost for one DiGA prescription per case	Triangular	€250[b], €450, €600	Goeldner and Gehder [5]
Physiotherapy substitution rate	R_{Sub}, physiotherapy	% of physiotherapy cases fully replaced by DiGA	Triangular	5%, 10%, 25%	Informed by digital physiotherapy cost literature [9] and conservative assumptions
EAP[c] substitution rate	R_{Sub}, EAP	% of EAP cases fully replaced by DiGA	Triangular	1%, 3%, 8%	Conservative assumption
Efficiency gain	E_{Gain}	Cost reduction per case with DiGA add-on	Triangular	10%, 20%, 30%	Shambushankar et al [11]
Physiotherapy complementary use rate	R_{Comp}, physiotherapy	% of remaining physiotherapy cases using complementary DiGA	Triangular	20%, 35%, 50%	Informed by survey-based evidence on digital health adoption [13] and conservative assumptions
EAP complementary use rate	R_{Comp}, EAP	% of remaining EAP cases using complementary DiGA	Triangular	15%, 25%, 40%	Conservative assumption

[a]DiGA: digital health application.

[b]A currency exchange rate of €1=US $1.18 is applicable.

[c]EAP: extended outpatient physiotherapy.

Model Parameters and Distributions

Key parameters follow the substitution and complementary logic and are specified with triangular distributions (minimum, mode, and maximum), informed by recent evidence on the German DiGA market and international digital physiotherapy literature. The exact values and sources are listed in Table 1. All costs are calculated on a per-case basis. Parameter draws are independent unless stated otherwise.

Patient Adherence Modeling

To enhance the real-world validity of the simulation and reflect that not all patients use DiGA with the same intensity, we integrated an evidence-based patient adherence model (Multimedia Appendix 1). This step is critical, as real-world data from Germany indicate that a significant portion of users discontinue DiGA use after the initial prescription period, and clinical effectiveness is closely linked to the level of patient engagement in a dose-response relationship [14].

The model stratifies the patient population within each simulation run into 2 distinct cohorts based on their adherence level [14]. The first cohort (engaged users) represents patients who exhibit long-term adherence, defined as renewing their DiGA prescription for at least a second cycle. According to real-world claims data from Germany's largest statutory health insurer, this group comprises approximately 15% based on publicly available German market data [15]. We assume that only patients in this cohort can realize the full potential of the intervention. Accordingly, the literature-based parameters for substitution and efficiency gains were exclusively applied to this group.

The second cohort (short-term users) comprises the remaining 85% of patients who use the DiGA only for the initial prescription period and then discontinue [15]. For this group, we conservatively assumed that their limited engagement is insufficient to produce a measurable clinical or economic effect in terms of substitution or efficiency gains, which were set to zero. However, we modeled a partial benefit equivalent to 3 physiotherapy sessions for short-term users, reflecting front-loaded value delivery within the initial prescription cycle. The full cost of the initial DiGA prescription was applied regardless of adherence status. This adherence-based stratification was applied within each of the 10,000 Monte Carlo iterations.

Outcome Measures

The primary outcome was total annual net savings (in euros) across the included Berufsgenossenschaften. The secondary outcomes included Berufsgenossenschaften-specific savings and the probability of any savings. Break-even results are reported separately. Analyses were conducted using Python 3.10 (Pandas, NumPy, and Matplotlib). All simulations were executed with fixed random seeds to ensure reproducibility.

Break-Even Threshold Analysis

With parameters fixed at their mode (base case), we solved two deterministic thresholds: (1) the maximum DiGA price (C_{DiGA}) that yields cost neutrality and (2) the minimum engaged adherence rate required for cost neutrality at the base-case DiGA price. These thresholds define operational and commercial boundary conditions for viability in this setting.

Queuing Model and DES

We quantified the effect of DiGA integration on access and time efficiency by modeling the EAP pathway as an M/M/1 queue (Poisson arrivals, exponential service times, and single server representing total therapeutic capacity). The model is

XSL•FO
RenderX

parameterized by the mean arrival rate (λ) and service rate (μ). λ was derived from administrative counts as annual EAP cases divided by 250 business days; μ was obtained by calibration using an empirical waiting-time sample (N=31 patients) from one author's rehabilitation center. In this framework, the probability of any wait equals system utilization ($\rho=\lambda/\mu$). We defined a clinically relevant delay as greater than 3 days and used the observed share exceeding this threshold (96.8%) as a data-driven estimator of baseline ρ. With λ and ρ known, μ followed algebraically. DiGA scenarios were implemented as proportional reductions in λ according to demand changes predicted by the adherence funnel.

The primary end point was the mean waiting time in queue per patient (Wq), which we extrapolated to the annual EAP population to obtain total patient waiting days saved as a system-level measure of improved accessibility. It is important to note that the calibration sample (N=31) is small and originates from a single center. While this provides a data-driven indication of high system use, the exact magnitude of waiting-time reductions may vary across regions and insurance funds. This limitation is inherent to the first iteration of the model and will be addressed in future multicenter extensions.

To validate the analytic M/M/1 results, we implemented DES with Poisson arrivals, exponentially distributed service times, and first-come first-served discipline. Run length and warm-up were chosen to control Monte Carlo error and remove transients: at least 20,000 post–warm-up patients per replication with a 2000-patient warm-up, repeated over 200 independent replications. Building on this calibrated baseline, we compared the status quo model against several DiGA integration scenarios derived directly from the therapeutic demand reduction predicted by the multistage patient selection and adherence funnel: a realistic case (base-case funnel parameters) and conservative and ambitious cases represented 50% and 150% of this effect, respectively.

Ethical Considerations

This study analyzed aggregated administrative data without individual-level or identifiable information. According to applicable regulations, formal ethical approval and informed consent were not required. All analyses complied with institutional data protection policies and the Declaration of Helsinki.

Results

Descriptive Statistics and Cost Dynamics

The analysis used 2024 baseline data from 5 German statutory accident insurance funds—Berufsgenossenschaften Bau (Construction), BGHM (Wood and Metal), BGHW (Trade and Logistics), VBG (Administration), and BGW (Health and Welfare)—covering 25.9 million insured persons. In 2024, the combined expenditures for physiotherapy and EAP across these Berufsgenossenschaften totaled €404.0 million (nominal). Year-over-year, costs rose by 10.1%, while patient volume increased by 1.2%, indicating that spending growth outpaced volume by approximately 9 percentage points. This widening gap signaled mounting financial pressure in rehabilitation pathways and motivated the subsequent simulation of DiGA integration.

Probabilistic Sensitivity Analysis

The primary simulation, encompassing 10,000 Monte Carlo iterations, quantified the prospective net financial impact of DiGA integration across the German statutory accident insurance landscape. The analysis revealed a mean annual net savings of €18.4 million (median: €17.9 million). The model predicts a 90.7% probability that DiGA adoption will be cost-saving, with a 95% uncertainty range for the annual financial impact spanning from a net cost of €8.0 million to a net saving of €47.7 million. Figure 2 displays the full probability distribution of the 10,000 simulated outcomes, illustrating a significant rightward shift of the probability mass relative to the cost-neutrality threshold.

XSL•FO
RenderX

Figure 2. Probabilistic sensitivity analysis of the net financial impact of digital health application (DiGA) integration. The histogram displays the frequency distribution of potential annual net savings resulting from 10,000 Monte Carlo simulation iterations. The model incorporates evidence-based probability distributions for all key parameters, including DiGA costs, substitution rates, efficiency gains, and patient adherence. Vertical lines indicate the mean annual net savings (€18.4 million), the median (€17.9 million), and the cost-neutrality threshold (€0). The significant shift of the distribution's probability mass into the positive range demonstrates the high likelihood (90.7%) of a favorable financial outcome from DiGA integration in the German occupational rehabilitation setting.

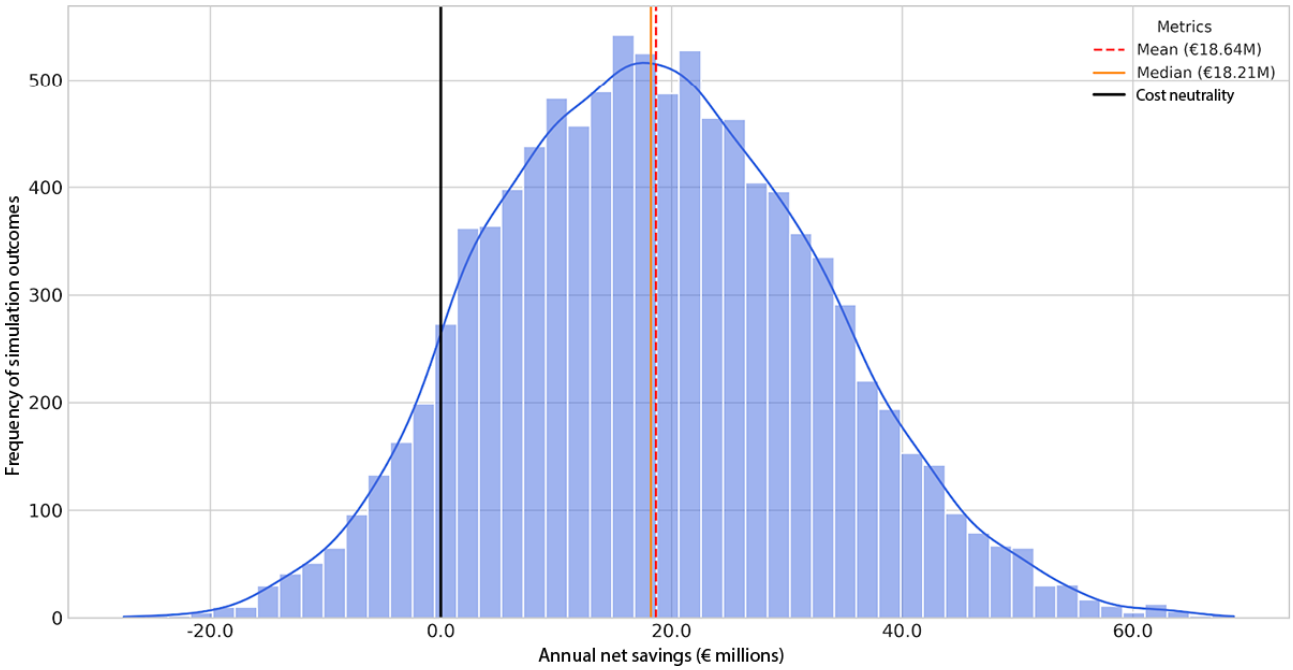

Heterogeneity of Financial Impact Across Insurance Funds

To investigate whether the potential for cost savings varied across different patient populations, the PSA was stratified by each individual Berufsgenossenschaften. This analysis revealed substantial heterogeneity in the estimated financial impact (Figure 3). While all analyzed Berufsgenossenschaften demonstrated a positive mean net saving, the magnitude and certainty of savings differed. Berufsgenossenschaften Bau, which covers the construction sector with a high incidence of severe traumatic injuries, showed the highest mean annual net savings at €4.9 million (95% CI €0.2-€10.3 million). Conversely, Berufsgenossenschaften covering administrative sectors (VBG) and health and welfare services (BGW), which often manage more chronic conditions, showed a smaller, though still positive, mean financial benefit. Notably, the 95% uncertainty ranges for these latter Berufsgenossenschaften overlap with the cost-neutrality threshold, indicating a higher degree of uncertainty regarding the financial outcome in these specific patient populations.

XSL•FO
RenderX

Figure 3. Heterogeneity of financial impact across Berufsgenossenschaften. The forest plot displays the mean annual net savings (points) and the corresponding 95% uncertainty range (horizontal lines) for each of the 5 analyzed German statutory accident insurance funds (Berufsgenossenschaften). Results are derived from a stratified probabilistic sensitivity analysis with 10,000 Monte Carlo iterations per Berufsgenossenschaften. The vertical dashed line at €0 represents the cost-neutrality threshold. The plot illustrates the differential impact of DiGA integration, with the highest point estimate for savings found in the Berufsgenossenschaften serving the construction sector (Berufsgenossenschaften Bau).

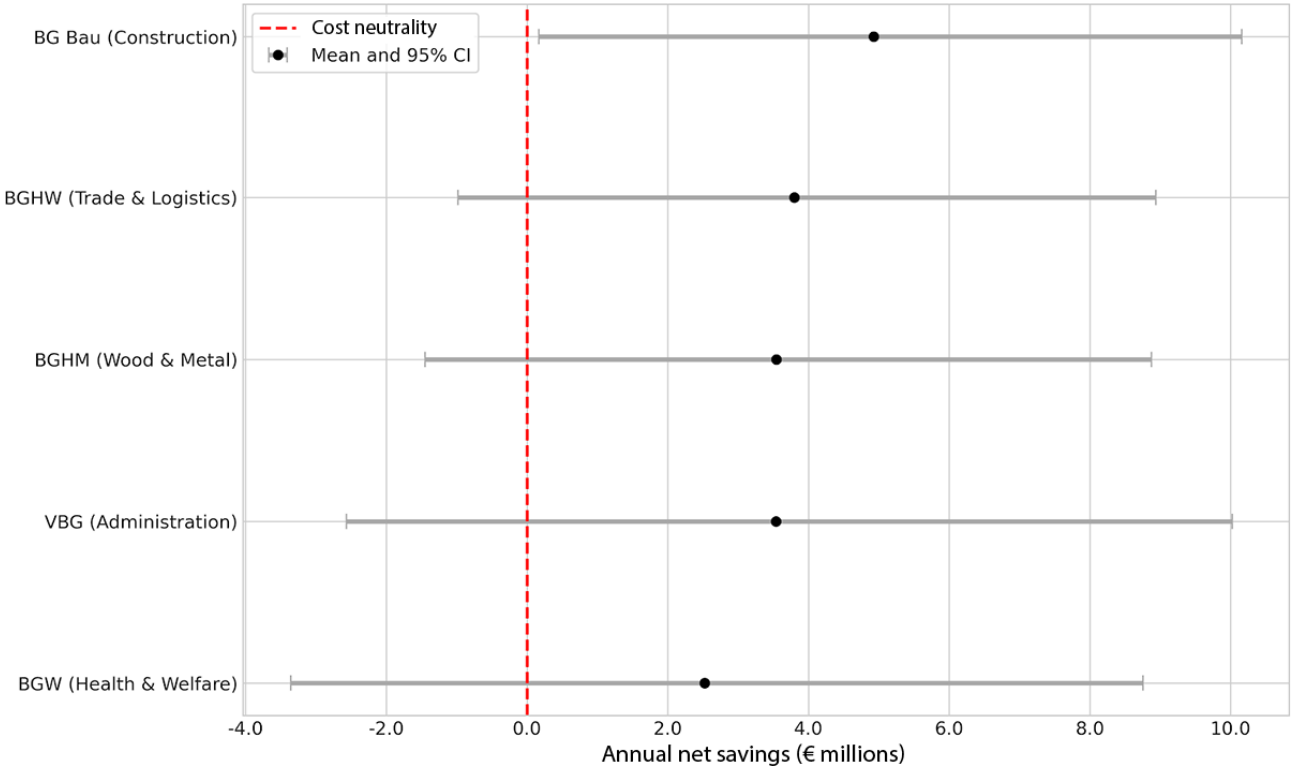

This heterogeneity likely reflects differences in baseline cost structures, injury severity profiles, and the relative share of acute versus chronic rehabilitation pathways across sectors. Construction sector rehabilitation typically involves high-cost posttraumatic pathways where substitution effects generate larger absolute savings, whereas administrative and health care sectors may have lower per-case costs and different rehabilitation patterns. These findings suggest that the economic return on DiGA integration is not uniform but rather dependent on the sectoral composition of the insured population and the cost intensity of prevailing rehabilitation pathways.

Patient Selection and Adherence Funnel

To generate the most realistic forecast of the net financial impact, we applied a 3-stage patient selection and adherence funnel to the simulation (Multimedia Appendix 1). This final model assumes that only 70% of potential candidates are eligible for and accept a DiGA prescription. This cohort is then stratified into long-term engaged users (15%) and short-term users (85%), with the latter achieving a partial benefit equivalent to 3 physiotherapy sessions.

This refined analysis yields a robustly positive financial case for DiGA integration. The simulation forecasts a mean annual net saving of €16.2 million (median: €15.7 million). Critically, the entire 95% uncertainty range (€5.0-€29.8 million) lies in positive territory, resulting in a 100% probability of achieving a net positive financial outcome (Figure 4). This demonstrates that even after applying conservative filters for patient eligibility and accounting for suboptimal real-world adherence patterns, the integration of DiGA remains financially viable.

XSL•FO
RenderX

Figure 4. Net financial impact of digital health application integration under a multistage patient selection and adherence funnel model. The histogram displays the frequency distribution of net savings from 10,000 Monte Carlo iterations. The model incorporates a 70% patient eligibility and acceptance filter, followed by a stratification into long-term (15%) and short-term (85%) users, with the latter achieving a partial benefit. The entire probability mass is located in the positive range, indicating a 100% probability of a cost-saving outcome, with a mean net saving of €16.2 million.

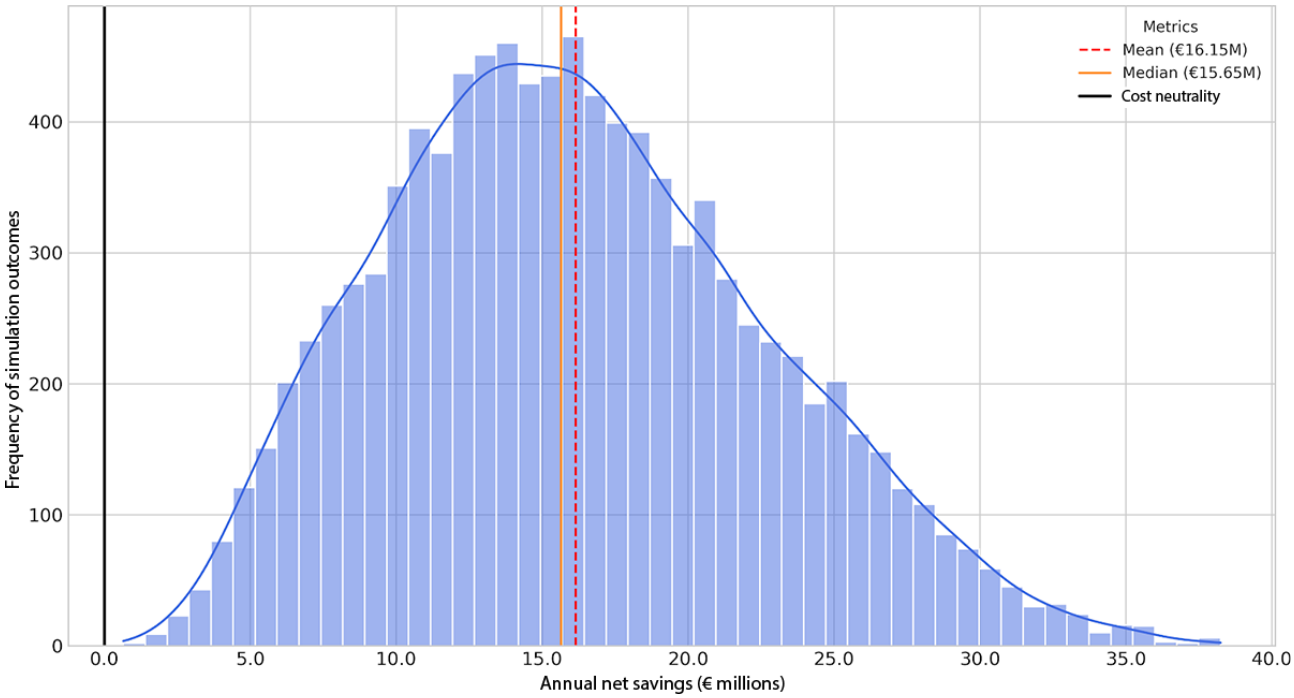

Notably, the structural dependency on partial benefits from short-term users represents a key finding of the adherence analysis. The break-even threshold analysis (see *Break-Even Threshold Analysis* section) confirms that even under maximally optimistic assumptions regarding long-term adherence, the model cannot achieve cost neutrality without capturing value from the initial prescription cycle of short-term users. This underscores that economic sustainability in digital therapeutics is not solely a function of durable adherence but critically depends on early-cycle benefit realization—a finding consistent with evidence on high attrition and the importance of early engagement in digital health interventions [16].

Break-Even Threshold Analysis

To determine the commercial and operational thresholds for implementation feasibility, a deterministic break-even analysis was conducted using the base-case parameters of the final funnel model (Figure 5). The analysis of the price threshold (Figure 5A) shows that DiGA integration remains cost-saving for an average DiGA price of up to a maximum of €617.80. This indicates that even at a price point nearly 40% higher than the base-case assumption of €450, the intervention does not result in a net loss, demonstrating substantial economic robustness. The model further showed that the long-term adherence rate alone cannot drive cost neutrality (Figure 5B): the calculated break-even adherence rate of 164.8% is a mathematical impossibility, confirming that the partial benefit from the 85% of short-term users is structurally indispensable for overall implementation feasibility.

XSL•FO
RenderX

Figure 5. Deterministic break-even analysis for DiGA price and patient adherence. The figure displays 2 one-way sensitivity analyses based on the final funnel model with all other parameters held at their base-case values. (A) The effect of varying the cost per DiGA on annual net savings. The model breaks even at a maximum price of €617.80. (B) The effect of varying the long-term adherence rate on annual net savings. The model fails to break even within the plausible 0% to 100% adherence range, indicating that the partial benefit from short-term users is essential for the overall implementation feasibility. DiGA: digital health application.

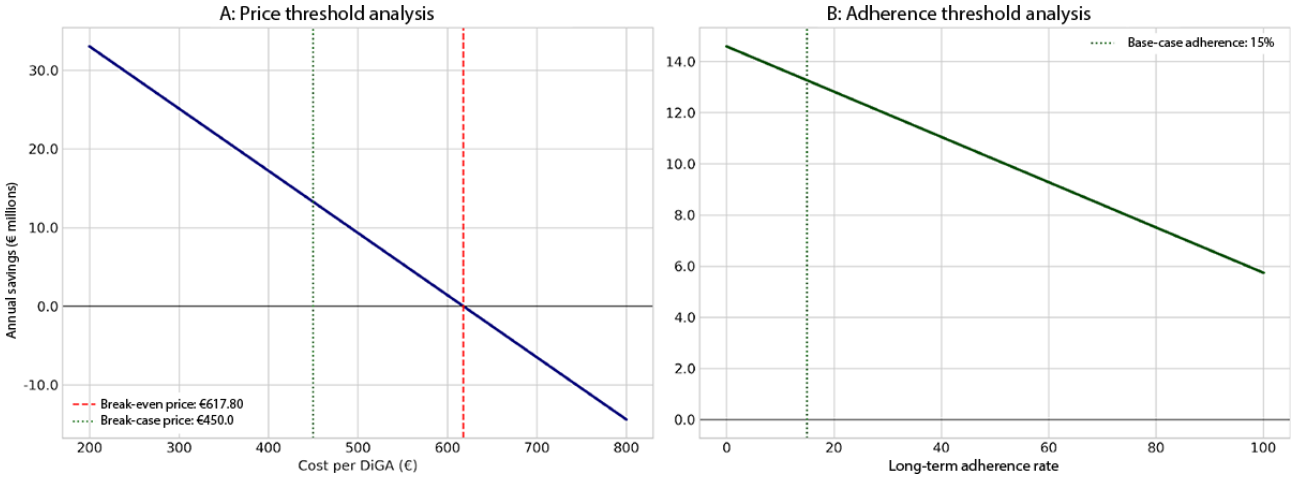

System-Level Impact on Care Accessibility

To quantify the impact of DiGA integration on time efficiencies, we modeled the EAP care pathway as a queuing system. The model was calibrated directly using empirical patient waiting-time data (N=31 patients), as shown in Figure 6A. We defined system use (ρ) as the proportion of patients waiting more than a clinically reasonable period of 3 days for therapy initiation. This data-driven approach revealed a mean system use of 96.8% based on the cohort, which is consistent with operation near capacity in the sampled setting and is therefore highly sensitive to changes in patient demand.

Figure 6. Impact of digital health application (DiGA) integration on extended outpatient physiotherapy (EAP) waiting times, based on a data-driven queuing model. The figure displays the results of the system-level time-efficiency analysis. (A) The frequency distribution of empirical waiting times from a sample of patients (N=31). The proportion of patients waiting more than 3 days was used to calibrate the queuing model to a baseline system utilization of 96.8%, which accurately reproduces the observed mean waiting time of 17.3 days. (B) A dual-axis chart showing the projected impact of different DiGA-induced demand reduction scenarios. The bars (left axis) represent the mean waiting time per patient, which is projected to decrease from 17.3 days to 12.8 days in the realistic scenario. The dashed line (right axis) shows the total annual patient waiting days saved for the entire EAP patient population (26,705), culminating in approximately 120,000 saved days in the realistic scenario. DiGA: digital health application.

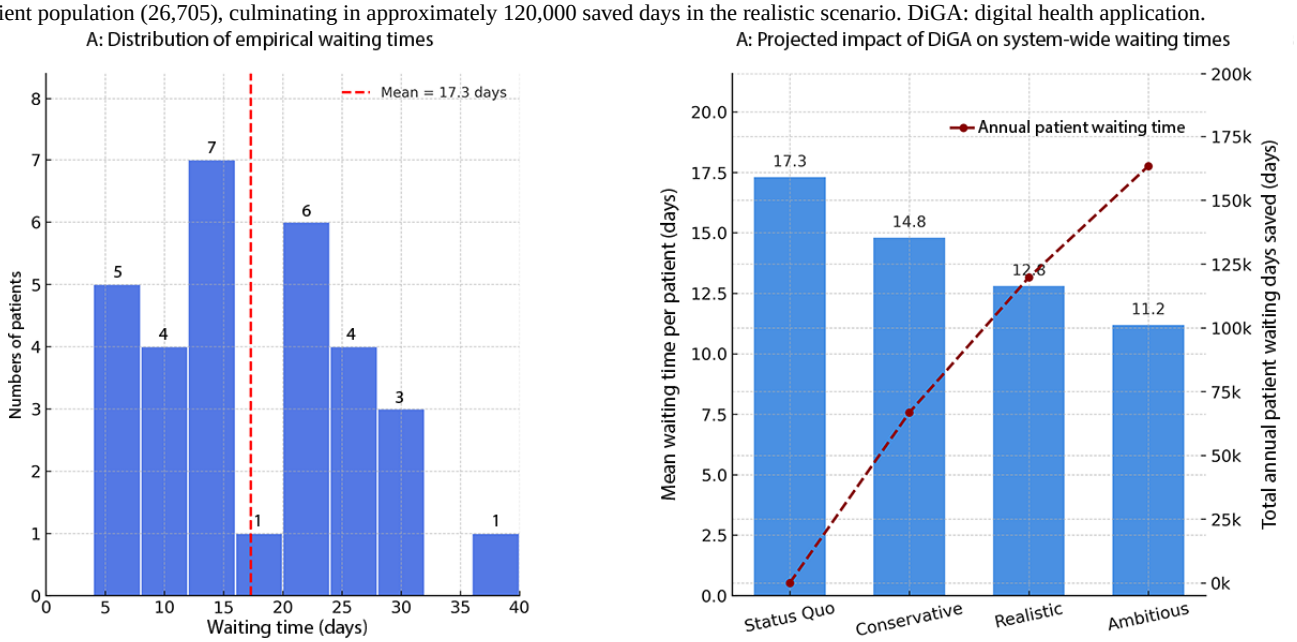

The simulation of DiGA integration under this calibrated model forecasts a substantial reduction in systemic waiting times (Figure 6B). The status quo scenario reproduces the empirically observed mean waiting time of 17.3 days per patient. In a realistic scenario, which assumes a modest reduction in therapeutic demand based on the final funnel model, the mean waiting time per patient decreases by 26% to 12.8 days. When extrapolated to the entire EAP patient population (26,705 patients), this reduction corresponds to a system-wide saving of approximately 120,000 cumulative patient waiting days annually. This finding indicates that in a capacity-constrained environment, DiGA can act as a critical release valve, mitigating systemic bottlenecks and increasing patient access to care.

XSL•FO
RenderX

For validation, results were corroborated in a DES parameterized to the same arrival and service rates. Under the calibrated baseline, the DES yielded a mean waiting time of approximately 17.3 days (95% uncertainty range: 15.5 - 19.6 d). A DiGA scenario with a 15% demand reduction produced a mean waiting time of approximately 10.58 days (95% range: 9.7 - 11.7 d), confirming the magnitude and direction of the analytic M/M/1 estimates (Multimedia Appendix 2).

Estimated Productivity Implications

To provide a preliminary estimate of indirect cost implications, we translated the projected waiting-time reductions into productivity-related outcomes. If the 120,000 cumulative patient waiting days saved annually correspond even partially to accelerated therapy initiation, the downstream implications for return-to-work timelines are substantial. Using publicly available data on average gross daily wages in the German construction sector (approximately €180 per business day) and accounting for Verletztengeld (injury-related wage replacement benefits, typically 68% - 80% of net wages), a conservative extrapolation suggests that the productivity-related value of reduced waiting could reach several million euros annually across the insured population, though precise quantification requires patient-level linkage data not available in this study. This estimate is intended as an order-of-magnitude indication rather than a definitive calculation and underscores the need for prospective studies that capture indirect costs alongside direct rehabilitation expenditures.

Discussion

Principal Findings

This study presents a multilevel simulation demonstrating that the integration of digital therapeutics into Germany's occupational rehabilitation system is likely to generate substantial economic benefits and system-level capacity gains under conservative, real-world assumptions. The primary simulation indicated mean annual net savings of €18.4 million (90.7% probability of cost savings), which increased to €16.2 million with a 100% probability of positive financial impact after incorporating adherence dynamics. Cost neutrality was maintained for DiGA prices up to €617.80 per prescription—nearly 40% above the base-case assumption—indicating substantial economic robustness. A calibrated queuing model projected a 26% reduction in mean EAP waiting times, equivalent to approximately 120,000 patient waiting days saved annually, demonstrating DiGA's potential as a strategic capacity tool in addition to its cost-saving properties.

Germany as a Real-World Laboratory for Digital Therapeutics

As the first country to establish a formal reimbursement pathway for digital therapeutics, Germany offers a unique real-world laboratory for evaluating digital medicine at scale [17]. By anchoring our model in actual cost data from Germany's statutory accident insurance funds and leveraging claims-based adherence rates, this work moves beyond theoretical potential to a more realistic forecast. The finding that the probability of a positive financial return approached 100% after adherence adjustment suggests that the economic case for DiGA integration is not merely marginal but substantial.

The Decisive Role of a Structured Implementation Pathway

Perhaps the most significant methodological contribution of this work is the shift from a simplistic universal-provision model to an evidence-based, multistage patient selection and adherence funnel. This approach underscores that the implementation feasibility of DiGA is not an inherent feature of the technology alone, but rather a function of the care pathway in which it is embedded. The model demonstrates that digital therapeutics cannot be distributed indiscriminately with an expectation of positive returns. Instead, it highlights the necessity of a structured implementation pathway that accounts for patient eligibility, clinical triage, and real-world engagement patterns, highlighting the importance of patient selection and sustained engagement in digital health interventions [18,19].

Heterogeneity of Impact and Precision Reimbursement

The stratified analysis reveals that the financial benefit of DiGA is not uniformly distributed across patient populations. The observation that savings potential is highest in Berufsgenossenschaften Bau (construction sector) implies that DiGA may deliver the greatest value when applied to acute, high-cost, posttraumatic rehabilitation pathways. This heterogeneity challenges current generic approaches to prescribing and reimbursement and suggests that future policy iterations should consider precision reimbursement models with value-based pricing structures that incentivize DiGA application in the highest-need, highest-cost cohorts [20-22]. The differential impact across sectors further highlights that the economic case for DiGA should be evaluated at the level of specific rehabilitation pathways rather than as a blanket system-wide metric.

Economic Boundaries as a Policy Signal

The break-even analysis delineates the boundary conditions for DiGA integration. The model demonstrates substantial robustness to price variation: cost neutrality persists until an average prescription price of €617.80, nearly 40% above the reference case (€450), aligning with recent evaluations of the German DiGA reimbursement framework [5]. In sharp contrast, the analysis shows that partial benefits from short-term users are indispensable for cost-effectiveness. Even under maximally optimistic assumptions, higher long-term adherence alone cannot achieve break-even within our model; without capturing value from short-term engagement, the model remains structurally loss-making. This counterintuitive finding underscores that economic sustainability is not solely a function of durable adherence but critically depends on early-cycle benefit realization, consistent with evidence on digital health attrition and front-loaded value delivery [16]. For policymakers, this implies that price negotiation and reimbursement design may be a more effective lever than adherence-focused interventions alone. For developers, it highlights the strategic imperative of creating interventions that deliver measurable benefits within

XSL•FO
RenderX

the first prescription cycle and remain effective despite predictable attrition in digital health interventions [19].

System-Level Capacity Effects

The queuing analysis reveals a potentially profound impact, demonstrating DiGA's capacity to resolve systemic bottlenecks in care delivery. Calibrating the queuing model against empirical data showed that 96.8% of sampled patients experienced waits exceeding 3 days, consistent with near-capacity operation. Under queuing theory, such conditions predict highly nonlinear effects: the projected 26% reduction in mean waiting time in the realistic scenario—from 17.3 to 12.8 days—and the system-wide saving of approximately 120,000 cumulative patient-days annually underscore this disproportionate improvement resulting from a modest reduction in demand. Consistency with the DES further supports these findings: the DES reproduced the calibrated baseline delay and projected a comparable reduction, demonstrating robustness to stochastic variation. Taken together, analytic and simulated results indicate that modest, targeted demand reductions can produce outsized accessibility gains once the system is nudged below its critical utilization threshold. This suggests that DiGA should be viewed not merely as a cost-saving substitute but as a strategic instrument to generate capacity and enhance patient access.

Implications for Global Digital Health Policy

The findings offer actionable insights beyond Germany. For policymakers in other countries, the model provides a clear economic rationale for investing in digital therapeutic frameworks [23,24], with the critical caveat that these investments must be coupled with strategies to manage patient selection and support engagement. For the global digital health industry, the results highlight that the central challenge is not only demonstrating clinical efficacy but also solving the real-world problem of patient adherence [16,18,19]. Ultimately, the most significant system-level impact of effective digital rehabilitation may extend beyond direct savings. The estimated reduction of approximately 120,000 waiting days annually could translate into substantially earlier return-to-work timelines, mitigating the indirect costs of lost productivity that constitute the largest component of the total economic burden of occupational injuries. The ultimate value of DiGA should therefore be contextualized not only by direct savings for insurers but also by their potential contribution to reducing the macroeconomic burden of workplace accidents [23,24].

Limitations

As a simulation-based implementation study, several limitations should be considered. First, the predictive accuracy of the simulation is contingent upon the validity of its input parameters. While grounded in the best available evidence, key assumptions—particularly the DiGA substitution rate (R_Sub) and the efficiency gain (E_Gain)—are derived from international studies and clinical equivalence trials rather than direct real-world data from the German occupational insurance system. The assumed substitution of complex, multimodal EAP cases, though modeled conservatively, remains especially theoretical. Consequently, the calculated savings should be interpreted as potential estimates under specific evidence-based assumptions, not as a definitive forecast. This reliance on literature-derived parameters is, however, standard practice in simulation modeling for health technologies at early adoption stages, where direct real-world data are not yet available, and the robustness of findings is ensured through a 10,000 iteration PSA.

The adherence model, while innovative in its use of real-world German claims data, simplifies a complex behavioral spectrum into a binary engaged versus short-term cohort. The approximately 15% long-term engagement rate, though derived from a large insurer, may not be perfectly generalizable to the Berufsgenossenschaften population.

The queuing model, while providing strong evidence of nonlinear capacity effects, relies on calibration from a small empirical waiting-time sample (N=31 patients) obtained from a single rehabilitation center. Although this sample provides a data-driven indication of high system use, it limits the precision of absolute waiting-time estimates and may not fully represent regional or fund-specific variation. Future studies should aim to incorporate larger, multiregional datasets and multiserver capacity dynamics. Additionally, while the DES validates the analytic model, this first iteration shares the single-server and exponential-service assumptions; extensions to multiserver capacity, empirical service-time distributions, day-specific capacity calendars, and no-show dynamics are planned to stress-test generalizability.

Finally, the analysis is based on aggregated data, precluding stratification by patient-level factors (diagnoses, comorbidities, age) known to influence outcomes and technology uptake.

Future Directions

While this study models the financial dimension of DiGA integration, a holistic valuation requires further analysis of its impact on care capacity and indirect costs. The next step is a richer DES incorporating multiserver capacity, weekday calendars, and empirical duration and no-show distributions calibrated to larger real-world datasets. To progress from forecasting toward causal inference, findings should be validated with prospective, patient-level data. A randomized controlled trial or rigorously adjusted observational study within the German Berufsgenossenschaften system would be the gold standard. Finally, implementation science studies should focus on clinical decision support for eligibility and triage and on evidence-based engagement strategies that sustain long-term use.

Acknowledgments

The authors thank Ariel Dora Stern and Benedikt Langenberger (Hasso Plattner Institute, University of Potsdam, Potsdam, Germany) for their helpful input and feedback on earlier drafts of this manuscript. Generative artificial intelligence tools were used to support language editing, text refinement, and editorial corrections during manuscript preparation. The authors reviewed,

XSL•FO
RenderX

critically assessed, and approved all artificial intelligence–assisted modifications and take full responsibility for the final content, accuracy, and integrity of the manuscript.

Funding

The authors declared no financial support was received for this work. The participating statutory accident insurance funds provided aggregated administrative data but had no role in study design, analysis, interpretation of results, manuscript preparation, or the decision to submit for publication.

Data Availability

Aggregated administrative data used in this study are not publicly available due to institutional agreements with participating insurance funds. Analytical code and simulation scripts are available from the corresponding author upon reasonable request.

Authors' Contributions

CW conceptualized the study, developed the methodology, performed the analyses, and drafted the manuscript. MD and SVS contributed to data interpretation and methodological refinement. FR supported data preparation and validation. DMD contributed domain expertise and facilitated access to administrative data. ML supervised the project and contributed to critical manuscript revision. All authors reviewed, edited, and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Conceptual flowchart of the multistage patient selection and adherence funnel. This diagram visually outlines the 3-stage model used to simulate the patient journey and its economic consequences under DiGA integration. The process begins with stage 1: patient eligibility and willingness, where a “pool of potential patients” undergoes a clinical and preference-based triage. Only those deemed “eligible & willing” (70%) proceed to DiGA prescription; others remain in the “standard care pathway.” In stage 2: patient adherence profile, the prescribed cohort is stratified based on real-world adherence patterns into “short-term user cohort” (85%) and “engaged user cohort” (15%). Finally, stage 3: outcome assigns differential benefits: “short-term users” achieve a partial economic benefit, while “engaged users” realize the full potential benefit, reflecting a dose-response relationship between engagement and therapeutic outcome.

[PNG File, 145 KB - jmir_v28i1e93793_app1.png]

Multimedia Appendix 2

EAP waiting time (DES): baseline vs DiGA scenario. Bars display the DES estimate of the mean queuing delay per patient E[Wq] in days for the calibrated status quo system and for a DiGA integration scenario that reduces demand by 15% ($\lambda'=0.85\ \lambda$). Using 2025 EAP starts (27 cases; 250 business days), we set $\lambda=0.108$ day^{-1}; the service rate was calibrated to observed waiting times, yielding $\mu\approx0.150$ day^{-1}. The DES reproduces the empirical mean delay at baseline (mean 17.28 days; 95% replication range 15.69-19.42) and projects a shorter delay with DiGA (mean 10.51 days; 95% 9.79-11.67). The implied system-level effect corresponds to a substantial reduction in accumulated waiting time across the annual EAP cohort.

[PNG File, 97 KB - jmir_v28i1e93793_app2.png]

References

1. Versicherte und Unternehmen [Article in German]. Deutsche Gesetzliche Unfallversicherung. 2024. URL: https://www.dguv.de/de/zahlen-fakten/versicherte-unternehmen/index.jsp [accessed 2025-07-18]
2. Ausgaben für rehabilitation und teilhabe [Article in German]. REHADAT-Statistik. 2022. URL: https://www.rehadat-statistik.de/news/Ausgaben-fuer-Rehabilitation-und-Teilhabe-00001/ [accessed 2025-07-18]
3. Trägerübergreifende ausgabenstatistik der BAR [Article in German]. Bundesarbeitsgemeinschaft für Rehabilitation. 2024. URL: https://www.bar-frankfurt.de/service/reha-info-und-newsletter/reha-info-2024/reha-info-012024/traegeruebergreifende-ausgabenstatistik-2024.html [accessed 2025-07-18]
4. Reinhard M. Economic and professional situation of outpatient physiotherapy and occupational therapy in Germany [Report in German]. : ifo Institute – Leibniz Institute for Economic Research at the University of Munich; 2017 URL: https://www.ifo.de/DocDL/ifo_Forschungsberichte_90_2017__Reinhard_Physio_Ergotherapie.pdf [accessed 2025-07-18]
5. Goeldner M, Gehder S. Digital health applications (DiGAs) on a fast track: insights from a data-driven analysis of prescribable digital therapeutics in Germany from 2020 to mid-2024. J Med Internet Res 2024 Aug 29;26:e59013. [doi: 10.2196/59013] [Medline: 39208415]

XSL•FO
RenderX

6. GKV report on the utilization and development of digital health application (DiGA) care in Germany. Med Tech Reimbursement Consulting. URL: https://mtrconsult.com/news/gkv-report-utilization-and-development-digital-health-application-diga-care-germany [accessed 2025-07-18]
7. Mäder M, Timpel P, Schönfelder T, et al. Evidence requirements of permanently listed digital health applications (DiGA) and their implementation in the German DiGA directory: an analysis. BMC Health Serv Res 2023 Apr 17;23(1):369. [doi: 10.1186/s12913-023-09287-w] [Medline: 37069592]
8. Marks D, Kitcher S, Attrazic E, Hing W, Cottrell M. The health economic impact of musculoskeletal physiotherapy delivered by telehealth: a systematic review. Int J Telerehabil 2022;14(2):e6524. [doi: 10.5195/ijt.2023.6524] [Medline: 38026565]
9. Ekman B, Nero H, Lohmander LS, Dahlberg LE. Costing analysis of a digital first-line treatment platform for patients with knee and hip osteoarthritis in Sweden. PLoS One 2020;15(8):e0236342. [doi: 10.1371/journal.pone.0236342] [Medline: 32785226]
10. Virtual Musculoskeletal (MSK) Solutions Assessment Report v1.3. : Peterson Health Technology Institute (PHTI); 2024 URL: https://phti.org/wp-content/uploads/sites/3/2024/10/PHTI-Virtual-MSK-Solutions-Assessment-Report-v1.3.pdf [accessed 2025-10-12]
11. Shambushankar AK, Jose J, Gnanasekaran S, Kaur G. Cost-effectiveness of telerehabilitation compared to traditional in-person rehabilitation: a systematic review and meta-analysis. Cureus 2025 Feb;17(2):e79028. [doi: 10.7759/cureus.79028] [Medline: 40099085]
12. Lewkowicz D, Bottinger E, Siegel M. Economic evaluation of digital therapeutic care apps for unsupervised treatment of low back pain: Monte Carlo simulation. JMIR mHealth uHealth 2023 Jun 29;11:e44585. [doi: 10.2196/44585] [Medline: 37384379]
13. Martinsen L, Østerås N, Moseng T, Tveter AT. Usage, attitudes, facilitators, and barriers toward digital health technologies in musculoskeletal care: survey among primary care physiotherapists in Norway. JMIR Rehabil Assist Technol 2024 Sep 16;11:e54116. [doi: 10.2196/54116] [Medline: 39283661]
14. Zhang ZY, Tian L, He K, et al. Digital rehabilitation programs improve therapeutic exercise adherence for patients with musculoskeletal conditions: a systematic review with meta-analysis. J Orthop Sports Phys Ther 2022 Nov;52(11):726-739. [doi: 10.2519/jospt.2022.11384] [Medline: 35960507]
15. Germany: DiGA facts and figures: German health insurer publishes comprehensive report. CMS Law (CMS Legal Services). URL: https://cms.law/en/int/publication/cms-dtx-legal-navigator/news-on-dtx-legal/germany-diga-facts-and-figures [accessed 2025-07-18]
16. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11. [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
17. Schmidt L, Pawlitzki M, Renard BY, Meuth SG, Masanneck L. The three-year evolution of Germany's digital therapeutics reimbursement program and its path forward. NPJ Digit Med 2024 May 24;7(1):139. [doi: 10.1038/s41746-024-01137-1] [Medline: 38789620]
18. Montalescot L, Baussard L, Charbonnier E. Factors associated with digital intervention engagement and adherence in patients with cancer: systematic review. J Med Internet Res 2024 Dec 11;26:e52542. [doi: 10.2196/52542] [Medline: 39661976]
19. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: a systematic and meta-analytic review. J Consult Clin Psychol 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
20. Gensorowsky D, Witte J, Batram M, Greiner W. Market access and value-based pricing of digital health applications in Germany. Cost Eff Resour Alloc 2022 Jun 13;20(1):25. [doi: 10.1186/s12962-022-00359-y] [Medline: 35698135]
21. Benning L, Teepe GW, Pooth JS, Hans FP. Performance-based reimbursement for digital therapeutics in Germany: a misconceptualized opportunity. Digit Health 2024;10:20552076241281199. [doi: 10.1177/20552076241281199] [Medline: 39347508]
22. Chen Y, Guzauskas GF, Gu C, et al. Precision health economics and outcomes research to support precision medicine: big data meets patient heterogeneity on the road to value. J Pers Med 2016 Nov 2;6(4):20. [doi: 10.3390/jpm6040020] [Medline: 27827859]
23. Gerke S, Stern AD, Minssen T. Germany's digital health reforms in the COVID-19 era: lessons and opportunities for other countries. NPJ Digit Med 2020;3:94. [doi: 10.1038/s41746-020-0306-7] [Medline: 32685700]
24. Yap LK, Ennis E, Mulvenna M, Martinez-Carracedo J. Defining and measuring engagement and adherence in digital mental health interventions: protocol for an umbrella review. JMIR Res Protoc 2025 Jul 28;14:e73438. [doi: 10.2196/73438] [Medline: 40720820]

Abbreviations

DES: discrete event simulation
DiGA: digital health application
EAP: extended outpatient physiotherapy

XSL•FO
RenderX

PSA: probabilistic sensitivity analysis

Edited by J Piera-Jiménez; submitted 19.Feb.2026; peer-reviewed by A Torad, L Herdiman; revised version received 30.Mar.2026; accepted 09.Apr.2026; published 26.May.2026.

Please cite as:
Wallner C, Schmidt SV, Reinkemeier F, Dick DM, Lehnhardt M, Drysch M
Integration of Digital Therapeutics Into Occupational Rehabilitation in Germany: Multilevel Simulation Study
J Med Internet Res 2026;28:e93793
URL: https://www.jmir.org/2026/1/e93793
doi:10.2196/93793

© Christoph Wallner, Sonja Verena Schmidt, Felix Reinkemeier, Dagmar Maria Dick, Marcus Lehnhardt, Marius Drysch. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Rapid Epidemiological Data Collection on Social Media for COVID-19: Comparative Study Between Online Surveys and Conventional Cohorts

Maged Mortaga[1*], MSc; Hendrik Nunner[1*], PhD; Sydney Paltra[2*], MSc; Leonard Stellbrink[1*], MSc; Jens Friedel[3], MSc; Manuela Harries[4], DrPH; Jessica Krepel[4], Dr rer nat; Berit Lange[4], Prof Dr; MuSPAD Study Group[4]; Viola Priesemann[3,5], Prof Dr; André Calero Valdez[1], Prof Dr

[1]Institute of Human-Centered Interactive Systems (IMIS), University of Lübeck, Ratzeburger Allee 160, Lübeck, Germany

[2]Technische Universität Berlin, Berlin, Germany

[3]Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany

[4]Helmholtz Center for Infection Research, Braunschweig, Germany

[5]Institute for the Dynamics of Complex Systems, Faculty of Physics, University of Göttingen, Göttingen, Germany

[*]these authors contributed equally

Corresponding Author:
Maged Mortaga, MSc
Institute of Human-Centered Interactive Systems (IMIS), University of Lübeck, Ratzeburger Allee 160, Lübeck, Germany

Abstract

Background: After COVID-19 was declared a pandemic by the World Health Organization (WHO) in March 2020, global responses relied on nonpharmaceutical interventions such as physical distancing and mask mandates. These measures were guided by mathematical models built on empirical data. Although traditional methods such as surveys and observational studies provide high-quality data, they are often slow and resource-intensive. Social media polls (SMPs) offer a faster, more cost-effective alternative.

Objective: This study aims to evaluate the reliability and biases of SMPs as a rapid supplementary tool for epidemiological data collection and to compare their representativeness and data quality with conventional approaches.

Methods: In this cross-sectional observational study in Germany, we used SMPs to collect data on infections and demographic attributes via Twitter and Mastodon. We collected data directly on the social media platforms as well as through forwarding to an external survey via post. The time frame covered was from 2019 to 2024. Data were analyzed for infection rates, sociodemographic representativeness, and overall data quality.

Results: SMPs demonstrated viability as a rapid data collection tool. Based on a sample of 6127 answers on social media and 867 responses from the external survey, the self-reported frequency of infection aligned well with conventional sources. Across all 4 studies, approximately one-third of respondents reported having never been infected, half reported having had 1 infection, and one-sixth reported having had 2 or more infections. Statistical analyses of differences between data from Twitter, Mastodon, the external survey, and conventional data showed only small effect sizes (Cohen w=0.105 - 0.188). Spearman rank correlation demonstrated strong positive associations between infection dates in the external survey and conventional data (ρ=0.883, P<.001), as well as between the external survey and the Robert Koch Institute (ρ=0.640, P<.001). However, demographic analyses revealed biases in the external survey. By design, SMPs do not provide detailed demographic data, limiting options for subgroup analyses.

Conclusions: We found SMPs to be a practical and cost-effective method for quickly gathering epidemiological insights. In particular, self-reported infection frequency can aid during periods of high availability of self-testing during epidemics. We demonstrate that, even with a nonrepresentative and biased sample, we were able to closely match infection numbers with Multilocal and Serial Prevalence Study of Antibodies Against Respiratory Infectious Diseases in Germany data and produce incidence trends comparable to those in official Robert Koch Institute data. One can argue that SMPs alone are insufficient for public health modeling, as they do not allow real-time monitoring of, for example, population infection rates based on serological data. They are also limited with regard to inherent demographic bias related to recruitment and the inability to collect individual-level covariates. However, they can complement traditional approaches by offering rapid, low-cost insights.

(J Med Internet Res 2026;28:e80311) doi:10.2196/80311

XSL•FO
RenderX

KEYWORDS

cross-sectional studies; pandemics; data collection; communicable diseases; social media; Twitter; Mastodon; COVID-19; Multilocal and Serial Prevalence Study of Antibodies Against Respiratory Infectious Diseases; MuSPAD; COVID-19 Snapshot Monitoring; COSMO; epidemiological models; digital health

Introduction

After the World Health Organization declared COVID-19 a public health emergency of international concern on January 30, 2020 [1] and as a pandemic on March 11, 2020 [2], it triggered unprecedented global responses aimed at mitigating the spread of infections and caused widespread societal, psychological, and economic impacts [3-7]. In the absence of an effective vaccine, various nonpharmaceutical intervention measures (NPIs) to contain the spread of the virus were implemented. NPIs included recommendations for physical distancing, travel restrictions, hygiene and sanitation measures (such as mask mandates), and temporary lockdowns [8-16]. Policy decisions (eg, NPIs, vaccination recommendations) were based on various sources of empirical data, such as case numbers, surveys on public behavior and attitudes, mobility changes, and social contacts [17], which were extensively supported by mathematical models to estimate the efficacy of such measures [18-24]. Such models, however, require high-quality and rapidly collected empirical data to serve as a reliable source of information for public health decisions [25-27].

There are different ways of collecting empirical data, which differ in collection speed and the quality of information. Conventional methods typically produce high-quality information while requiring a long time for data collection. For example, in Germany, seroprevalence studies like the Multilocal and Serial Prevalence Study of Antibodies Against Respiratory Infectious Diseases (MuSPAD) in Germany [28] provided valuable insights. The MuSPAD study uses established protocols to measure the prevalence of antibodies against SARS-CoV-2 in the population at different times to determine when and how many people have been exposed to the virus [29]. It was later adapted to an epidemic panel and supplemented by novel multiplex serological devices able to gather reinfection data for relevant respiratory infections. Another example is the COVID-19 Snapshot Monitoring (COSMO) study [30], which uses repeated cross-sectional surveys to continuously track public perceptions, attitudes, and behaviors regarding the COVID-19 pandemic in Germany to inform public health interventions and improve communication strategies. While these approaches provide high-quality data from representative samples, they are time-consuming and costly, limiting their ability to inform real-time public health decisions. In contrast, novel methods, such as scraping social media data, typically allow faster data collection, however, at the cost of producing lower-quality information [31-33]. Social media data can be less reliable due to several factors, including demographic disparities, selection, self-selection bias, inconsistent user activity levels, and platform bias [34-36]. Previous studies have identified these biases in social media data, particularly in the context of public health crises, such as demographic differences in posting COVID-19-related content [37] and the challenges of ensuring representativeness in Twitter (We will refer to X as Twitter throughout this paper, as that was the official name during data collection.) polling data [38]. In addition, social media was one of the main drivers of spreading misinformation during the COVID-19 pandemic [39-41], with WhatsApp, Facebook, Twitter, and YouTube among the most used platforms [42-44]. Despite these challenges, social media polls (SMPs) offer the advantage of collecting large amounts of data in real-time without labor- and cost-intensive processes. This has been shown, especially in health-related contexts [38,45-47].

In addition to direct data collection, SMPs hold the potential to serve as a rapid and cost-effective recruitment tool for more comprehensive linked online surveys. Such surveys can capture more nuanced and extensive epidemiological information. By leveraging SMPs for recruitment, researchers can therefore balance cost-efficiency and the need for high-quality data.

However, a systematic assessment of the practicability, reliability, and biases of data collected using SMPs and comparison of data quality with conventional epidemiological methods has not been undertaken, leaving significant gaps in understanding the potential of SMPs in epidemiological contexts. We therefore ask: to what extent does COVID-19 data collected through SMPs differ from conventional methods regarding representativeness and reliability?

To test the reliability of SMPs both as data sources and recruitment tools for epidemiological surveys, we hypothesize the following:

- H1: Data related to COVID-19, sourced from cost-effective SMPs and shared surveys, is less representative of the population compared to data derived from more conventional methods, such as the MuSPAD antibody study. However, we assume that the differences show only small effects.

To test the reliability of SMP data in terms of the platform used for data collection, we hypothesize the following:

- H2: The type of social media platform used for gathering data will significantly influence the selectivity of the population represented and, consequently, the relevance and reliability of the data collected.

Methods

Study Design and Objectives

This study employed a quantitative cross-sectional design to evaluate the reliability and biases of SMPs compared to conventional epidemiological methods for collecting health-related data. We used Twitter and Mastodon, a decentralized social media platform, to gather data on COVID-19 infections, integrating this approach with a linked online LimeSurvey questionnaire. Additionally, data from conventional sources, including the MuSPAD and COSMO

XSL•FO
RenderX

studies, as well as official reports from the Robert Koch Institute (RKI) and Federal Statistical Office of Germany (FSO), were used for comparison.

Our study was preregistered on OSF in July 2023 [48]. Our third preregistered hypothesis focused on the derivation of contact network properties (H3: During the COVID-19 pandemic, individuals exhibited more pronounced social selectivity based on a homophily—the tendency to associate with similar others—due to increased awareness of their "second-order contacts" [contacts of contacts]) is tested in a second, forthcoming publication [49]. The recruitment strategy was developed afterward to increase reach and was introduced during the survey development. Bootstrapping was also introduced at a later point to address recruitment bias.

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki. Informed consent was obtained from all participants. It was approved by the ethics committee of the University of Lübeck (approval number 2023-413). Reporting was conducted in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guidelines [50]. Recruiters were not compensated in any way and included study authors, as well as one person affiliated with the study institutions. Sharing the recruitment link via social media merely served as a neutral means of participant recruitment and did not provide the authors with any personal, financial, or competitive advantage influencing the study's outcomes.

Clarification and obtaining consent to share nonpersonal data as part of a social media survey took place in terms of use of the respective social media services. Apart from the publicly accessible survey data, no other, and thus, in particular, no personal data are used.

Recruiters and SMP Distribution

We enlisted 5 German-speaking recruiters, each with an established social media presence, to distribute the SMPs. All recruiters were informed about the study and willingly agreed to participate. The recruiters had follower counts ranging from approximately 750 to 65,000 on Twitter (Table 1). Recruiter 1, who also had a Mastodon account with over 8300 followers, posted identical polls on both platforms, enabling platform comparisons. At the time, Recruiter 1 was the only recruiter with more than 8300 followers on Mastodon, which we deemed necessary for a reliable comparison of results.

Table . Overview of recruiters used for social media poll distribution.[a]

Recruiter	Followers (Twitter), n	Short description	Domain
Recruiter 1	~65,000	Professor	Physics
Recruiter 2	~39,000	Professor	Economics
Recruiter 3	~10,000	Medical researcher	Medicine
Recruiter 4	~2500	Professor	Psychology
Recruiter 5	~750	Professor	Computer science

[a]This table summarizes the accounts leveraged to recruit participants for the comparative study on COVID-19 infections. Included are the number of followers, short description, and domain. The recruitment targeted followers of these accounts in Germany between July 19 and July 26, 2023.

Polls were posted from July 19 to July 26, 2023 (7 days being the maximum due to platform limits) and included 2 multiple-choice questions on COVID-19 infection history. The polls were structured to allow respondents to select their answers or view results without participating.

A custom-built Twitter posting tool automated the posting process for 3 recruiters, ensuring consistent timing across accounts. Recruiters who posted manually followed similar guidelines to minimize timing bias. The polls were posted in a thread, with the last post being a post with a link to an external LimeSurvey questionnaire, allowing respondents to transition seamlessly from poll participation to survey completion. From now on, external survey always refers to the externally hosted LimeSurvey questionnaire that was linked at the end of the Twitter or Mastodon thread. All questions were originally formulated in German and translated for the purpose of this paper (see also Table S1 in Multimedia Appendix 1).

The initial social media post briefly explained the study purpose to ensure transparency and provide context. An anonymized example of the Twitter thread is shown in Figure 1, where identifiable details, including the location of the team and the survey link, have been removed.

XSL•FO
RenderX

Figure 1. Example of a Twitter thread used for the comparative study on COVID-19. The figure illustrates the format and structure of a recruitment thread used to conduct poll-based data collection among any social media users who viewed the post, including followers of 5 recruiter accounts. All identifiable research information has been removed to maintain researcher anonymity. The recruitment took place in Germany between July 19 and July 26, 2023.

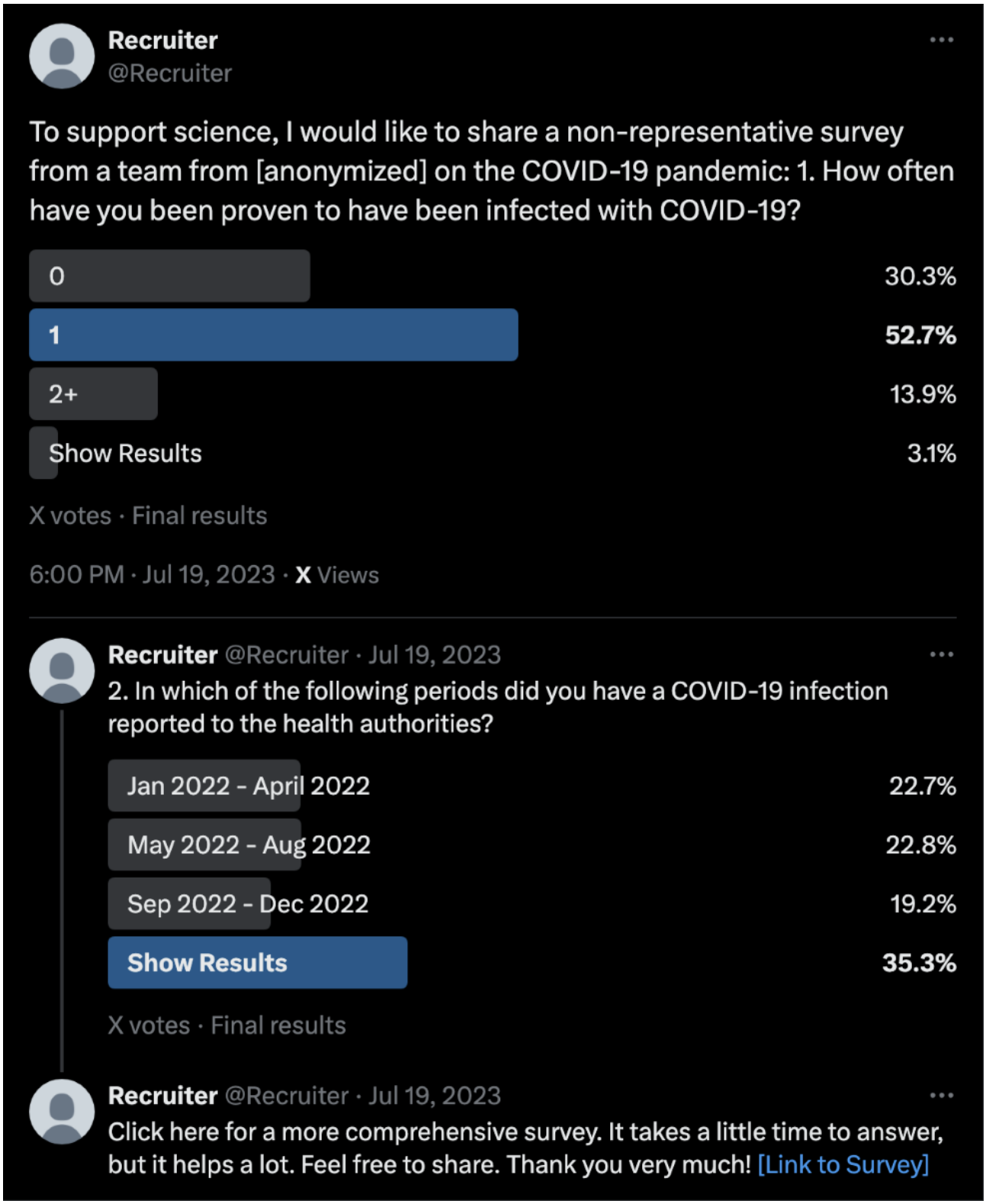

Data Collection

SMPs and External Survey

The external survey was open from July 18 to August 30, 2023. It collected detailed information on sociodemographics, health status (COVID-19 infections, vaccination history, preexisting conditions), and social contacts. The survey formed part of a broader research project and therefore contained several additional sections that are not examined in this paper. Because the overall instrument was comparatively long, we anticipated a substantial dropout rate. For this reason, the study design

XSL•FO
RenderX

explicitly incorporated the use of incomplete questionnaires, provided that respondents had supplied sufficient data for the analyses reported here. Respondents were informed about data collection and processing at the start of the external study, and consent was explicitly requested. The participation requirements were only that participants have access to the survey link and be at least 18 years of age. All variables and the survey are available in the preregistration on OSF [48]. The contact data will be described here but analyzed in depth in a separate paper [49].

A summary of survey responses by medium is described in Table 2.

From the preregistration, we expected 5000 participants in the SMPs, which we exceeded by 1172 participants. For the external survey, we expected 500 responses. The 867 responses received and used for analysis exceeded our target. We also used incomplete responses for our analysis.

Table . Survey response data across social media platforms in the comparative study on COVID-19.[a]

Platform and question/type	Responses, n
Twitter	
Question 1	4370
Question 2	2129
Mastodon	
Question 1	1802
Question 2	738
External survey (excluding speeders[b])	
Started	867
Completed	398
Survey shares	68
Completed shared surveys	12

[a]The table reports response counts from any social media users who answered polls posted via 5 recruiter accounts on Twitter and Mastodon, as well as participants in the external survey. Data are disaggregated by platform and question. Recruitment and data collection occurred in Germany between July 19 and July 26, 2023, for Twitter and Mastodon and between July 18 and August 30, 2023, for the external survey.

[b]A speeder is defined as a participant completing the survey in less than one-third of the median time.

All responses were anonymized during data processing. Personally identifiable information, such as IP addresses and free-text responses, was removed, and speeders were excluded from the analysis as per the definition in the preregistration [48]. It is assumed that participants falling into this category did not provide meaningful data points and therefore were excluded. Data preprocessing and analysis were performed using R version 4.4.1 and the *tidyverse* packages (version 2.0.0) [51,52]. The code to reproduce the analysis is publicly available on GitHub [53].

The average completion time was around 17 minutes. We did not check device compatibility, as our survey did not have any specific device requirements. No attention checks were included in the survey design. All participant answers that were not excluded for qualifying as speeders were used for the analysis. Missing data were not imputed (see supplementary materials in Multimedia Appendix 1 for detailed reporting of missing data by variable).

Multilocal and Serial Prevalence Study of Antibodies Against Respiratory Infectious Diseases in Germany

In the spring of 2022, a subset of 9921 of the invited 33,426 MuSPAD participants took part in the corresponding survey round. These 9921 participants are the source of all demographic data presented in this comparison. The next round of data collection was conducted in the winter of 2022/2023. This is the source of all data on the number of infections, the time of infection, and the number of vaccinations. Here, we only considered the responses of the 9921 participants who had already participated in spring 2022. However, not all participants responded to the winter survey, which reduced the sample size to 5128 participants.

Data Processing

The MuSPAD data were merged from the independent waves using the following procedure: use the “user_id” column to assign results to participants and merge all results. MuSPAD collected the number of infections up to the end of 2022; participants could only report 1 date of infection for the period from April 1, 2023, to August 31, 2023. Therefore, incidence during this period was calculated based on a maximum of 1 reported infection per participant.

We apply weighted bootstrapping to ensure that the age distribution of participants in the external survey matches the age distribution of the German population (see *Demographic Comparison* section). This also allows for comparisons of the 7-day incidence per 100,000 between the external survey, the MuSPAD study, and the official reporting statistics by RKI. To compute the mean 7-day incidence and the empirical 95% CI of this mean for each point in time, we apply bootstrapping. We perform 1000 bootstrap iterations, resampling with replacement

XSL•FO
RenderX

within each age group, with sample sizes weighted according to population age distributions, to generate incidence estimates.

Comparison of Data Sources

We evaluated the representativeness and reliability of the data collected via SMPs and the external survey against the following conventional sources:

1. MuSPAD [28,29]: a sequential seroprevalence study of SARS-CoV-2 infections and vaccinations involving 5128 participants in 2022 - 2023.
2. COSMO [30,54]: a cross-sectional serial online survey capturing public perceptions of COVID-19, with data from 1003 respondents in late 2022.
3. Official reports: 7-day incidence rates and vaccination statistics from RKI [55,56] and demographic data from the FSO collected on a daily basis.

For demographic comparisons, gender [57], age [58], household size [59], education level [60], and occupation [61] data were used, sourced from MuSPAD and national statistics. Age brackets were introduced into the data to facilitate comparison with other data. These consist of the age-in-years brackets: 18 - 39, 40 - 59, 60 - 79, and 80 - 99.

Analysis Framework

Descriptive statistics were used to analyze the distribution of infection rates, vaccination history, and demographic data. Comparisons included differences between Twitter and Mastodon responses, as well as variations among the 5 recruiters. Furthermore, we excluded the “show results” votes on the SMPs from the analysis (“show results” means that on Twitter and Mastodon, users submitted no valid answer themselves but only looked at the votes of others, providing no data points).

To estimate differences in infection distributions, we used contingency tables and reported Cohen *w* as a measure of total sample differences. To compare infection timelines across data sources, we calculated pairwise Pearson and Spearman correlation coefficients and conducted corresponding correlation tests to obtain *P* values for each comparison. Data from the external survey were compared against MuSPAD and the FSO to quantify bias.

We computed 95% CIs for any sample proportion, p^. As we assumed a binomial distribution for the true population proportion and as the binomial distribution is approximately normal for large enough samples, we used *z* scores when computing the CIs.

Results

Summary of Results

We found similar results for all samples regarding infection frequency, timing of infections, number of vaccinations, 7-day incidences, and demographic differences, although some differences could be explained by different measurement time periods or differences in measurement methods. Based on a sample of 6127 responses on Twitter and Mastodon and 867 responses from the external survey, the self-reported frequency of infection aligned closely with conventional sources, such as MuSPAD. Across all 4 studies, 28% to 38% of respondents reported never being infected (Twitter 28%, Mastodon 38%, external survey 34%, MuSPAD 33%), 50% to 56% reported one infection (Twitter 55%, Mastodon 50%, external 56%, MuSPAD 56%), and 8% to 17% reported 2 or more infections (Twitter 17%, Mastodon 12%, external 10%, MuSPAD 12%). The differences compared to MuSPAD were significant but showed only small effect sizes (Cohen *w* for Twitter: 0.152, Mastodon: 0.188, and external survey: 0.105). We also compared the subsample of SMPs by recruiters, indicating that follower size did not strongly impact findings.

COVID-19–Related Comparisons

Table 3 provides an overview of the demographic composition and self-reported COVID-19 infection history across the different samples. While the age distributions of Twitter and Mastodon participants are comparable, both differ from the MuSPAD dataset, which includes a broader age range. The proportion of self-reported infections is also relatively consistent across social media–based samples and MuSPAD, although participants from MuSPAD reported fewer repeat infections. These similarities and differences highlight the role of recruitment methods in shaping sample characteristics.

XSL•FO
RenderX

Table . Demographic characteristics and COVID-19 infection data comparing external survey participants with the Multilocal and Serial Prevalence Study of Antibodies Against Respiratory Infectious Diseases cohort.[a]

Characteristic	Recruited via Twitter (N=565)	Recruited via Mastodon (N=276)	MuSPAD[b] (N=9921)	Overall (N=10,762)
Age (y)				
Mean (SD)	48.9 (11)	48.6 (10.1)	57.1 (16.5)	56.4 (16.2)
Median (minimum-maximum)	49 (18-83)	49 (21-75)	59 (19-101)	58 (18-101)
Missing, n (%)	8 (1.4)	5 (1.8)	111 (1.1)	124 (1.2)
Gender, n (%)				
Female	318 (56.3)	130 (47.1)	5966 (60.1)	6414 (59.6)
Male	242 (42.8)	135 (48.9)	3861 (38.9)	4238 (39.4)
No answer	4 (0.7)	6 (2.2)	79 (0.8)	89 (0.8)
Other	1 (0.2)	5 (1.8)	15 (0.2)	21 (0.2)
Number of COVID-19 infections, n (%)				
0	194 (34.3)	95 (34.4)	5730 (57.8)	6019 (55.9)
1	311 (55)	157 (56.9)	3474 (35)	3942 (36.6)
≥2	60 (10.6)	23 (8.3)	429 (4.3)	512 (4.8)
No answer	0 (0)	1 (0.4)	288 (2.9)	289 (2.7)

[a]This table contrasts participants recruited via Twitter and Mastodon with the MuSPAD dataset in terms of age, gender, and self-reported history of COVID-19 infection. The social media survey sample comprised followers of 5 recruiter accounts in Germany, with data collected between July 19 and July 26, 2023. The MuSPAD dataset serves as a conventional cohort reference to contextualize the representativeness of the social media–based sample.

[b]MuSPAD: Multilocal and Serial Prevalence Study of Antibodies Against Respiratory Infectious Diseases.

The results from Twitter, Mastodon, the external survey, and MuSPAD are largely consistent in terms of reported infection history (Figure 2). For these 4 studies, the proportion of respondents who stated that they had never been infected is comparable, varying between 28% and 38% (Twitter: 1191/4217, 28%, 95% CI 26.9%-29.6%; Mastodon: 666/1765, 38%, 95% CI 35.5%-40%; external survey: 294/866, 34%, 95% CI 30.8%-37.1%; MuSPAD: 1475/4997, 30%, 95% CI 28.3%-30.8%). Similarly, around half of the respondents of each study stated that they had been infected once (Twitter: 2310/4217, 55%, 95% CI 53.3%-56.3%; Mastodon: 882/1765, 50%, 95% CI 47.6%-52.3%; external survey: 483/866, 56%, 95% CI 52.5%-59.1%; MuSPAD: 2920/4997, 58%, 95% CI 57.1%-59.8%). The percentage of respondents who reported experiencing at least 2 infections ranged from 8% to 17%, while the highest share was recorded on Twitter at 17% (716/4217, 95% CI 15.8%-18.1%), followed by Mastodon at 12% (216/1765, 95% CI 10.7%-13.8%), the MuSPAD study at 12% (602/4997, 95% CI 11.1%-12.9%), and the external survey at 10% (89/866, 95% CI 8.3%, 12.3%).

XSL•FO
RenderX

Figure 2. (A) Share of participants who reported having been infected 0, 1, or ≥2 times with COVID-19 since the beginning of the COVID-19 pandemic. Sample sizes reflect that not all participants provided an answer. Error bars represent 95% CIs (see *Analysis Framework* section for details). The visible differences between the COVID-19 Snapshot Monitoring (COSMO) study and the other 4 studies can be traced back to different data collection periods (see panel B). (B) Timeline depicting the different data collection periods. Bold blocks represent the time frame of actual data collection, and the dotted lines represent the time frame during which infections could have occurred. The COSMO study ended in November 2022 and thus does not include infections from 2023 onward. MuSPAD: Multilocal and Serial Prevalence Study of Antibodies Against Respiratory Infectious Diseases.

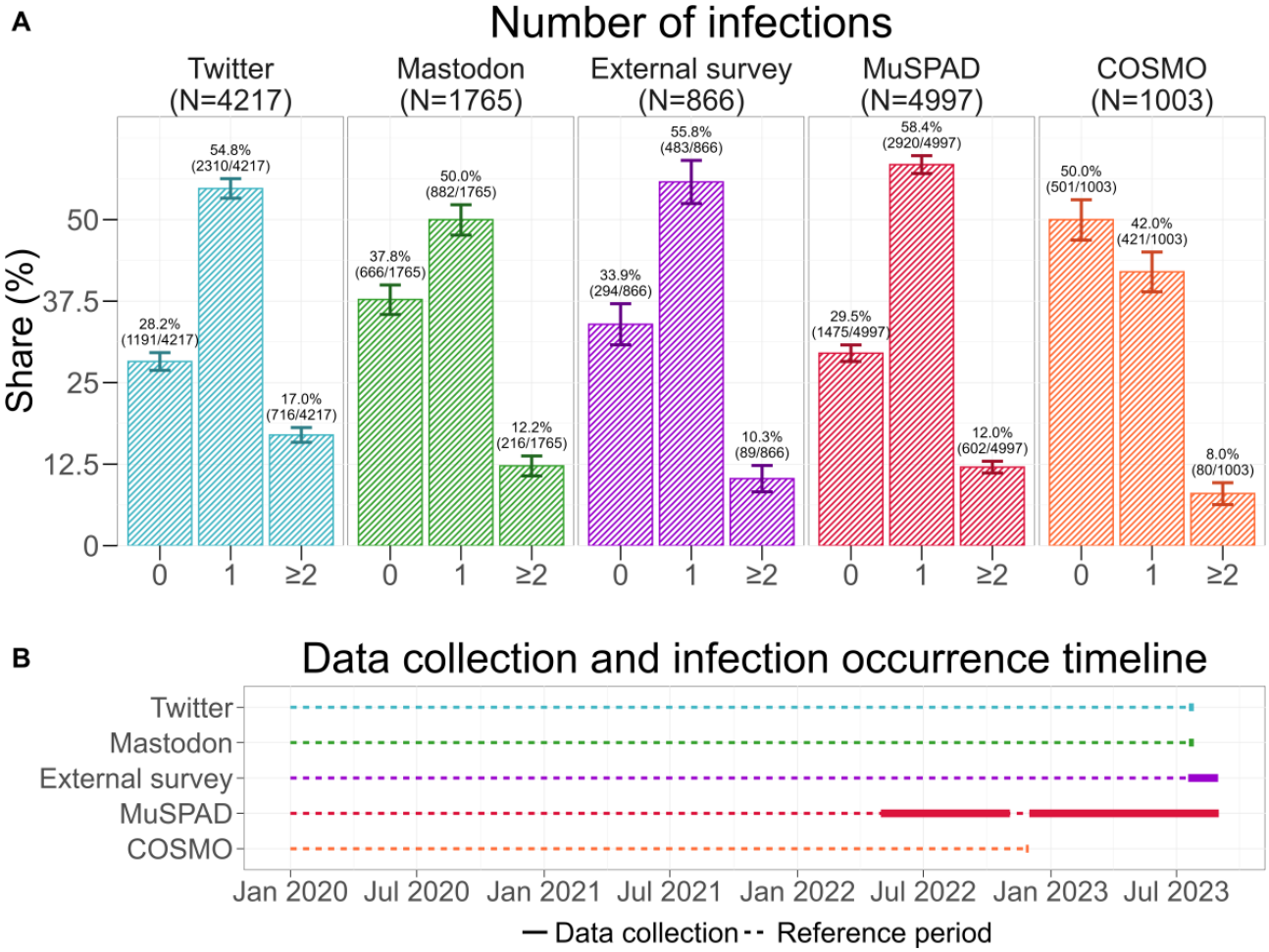

Overall, the number of infections reported on Twitter showed a difference compared to the reference MuSPAD data with Cohen *w* 0.152, Mastodon with Cohen *w* 0.188, and the external survey with Cohen *w* 0.105. We additionally computed a sensitivity analysis, which indicated that, given the smallest sample size (n=866) and assuming α=.01 and power=0.95, the study could detect effects of Cohen *w* of at least 0.154. We opted for a post hoc sensitivity analysis instead of a power analysis, as we are mostly interested in determining the threshold for differences that we could have detected rather than determining the power retrospectively.

In contrast, a comparison with the COSMO study shows visible differences. In the most recent round of the COSMO study, conducted on November 29, 2022, and November 30, 2022, 50% of participants (501/1003) reported that they had never been infected, 42% of participants (421/1003) reported that they had been infected once, and 8% (80/1003) reported that they had been infected at least 2 times. Due to its implementation period, the COSMO study does not account for infections that occurred in 2023, impeding a meaningful comparison between the COSMO study and the other 4 studies. However, higher counts for zero infections seem reasonable at an earlier point in time. Summarizing, we can state that the number of infections was captured similarly between samples.

A comparison of the timing of infections (Figure 3) shows that the 7-day incidence per 100,000 from the external survey, the MuSPAD study, and the officially reported incidence by the RKI follow the same trend from March 2020 until data collection ended in summer 2023 (Twitter/Mastodon questions are excluded from the timing comparison as they do not allow the computation of a 7-day incidence per 100,000 for COVID-19 cases. See the *Data Collection* section for question formulations and Timing of Infection [Twitter and Mastodon] in the Supplementary for discussion of the votes on question 2.). Specifically, waves, local maxima, and local minima occur simultaneously in all 3 data sources. We further note that the local maxima in July 2022 and October 2022 are larger for the external survey and the MuSPAD study, reaching values around 1500, while the 7-day incidence per 100,000, according to the RKI, only reaches values around 500. Overall, patterns of infections are reasonably similar between samples, although measurements were conducted differently.

XSL•FO
RenderX

Figure 3. (A) 7-day incidence per 100,000 from March 2020 until summer 2023. For the external survey and the Multilocal and Serial Prevalence Study of Antibodies Against Respiratory Infectious Diseases (MuSPAD) study, the incidence for individuals aged 18 years and older is depicted, while for the Robert Koch Institute (RKI), the incidence for individuals aged 15 years and older is depicted. As the RKI does not provide case data in appropriate age bins, an exact match of age groups was not possible. The waves, local maxima, and local minima occur simultaneously in all 3 data sources. Ribbons represent 95% CIs (see *Analysis Framework* section for details). Bootstrapped results are presented in Figure S1 in Multimedia Appendix 1. For original German and English translations of corresponding survey items, see Table S3. (B) Timeline depicting the different data collection periods. Bold blocks represent the time frame of actual data acquisition, and the dotted lines represent the reference period, the time frame during which infections could have also occurred. In contrast to the external survey and the MuSPAD study, the RKI continuously collected infection data during the COVID-19 pandemic.

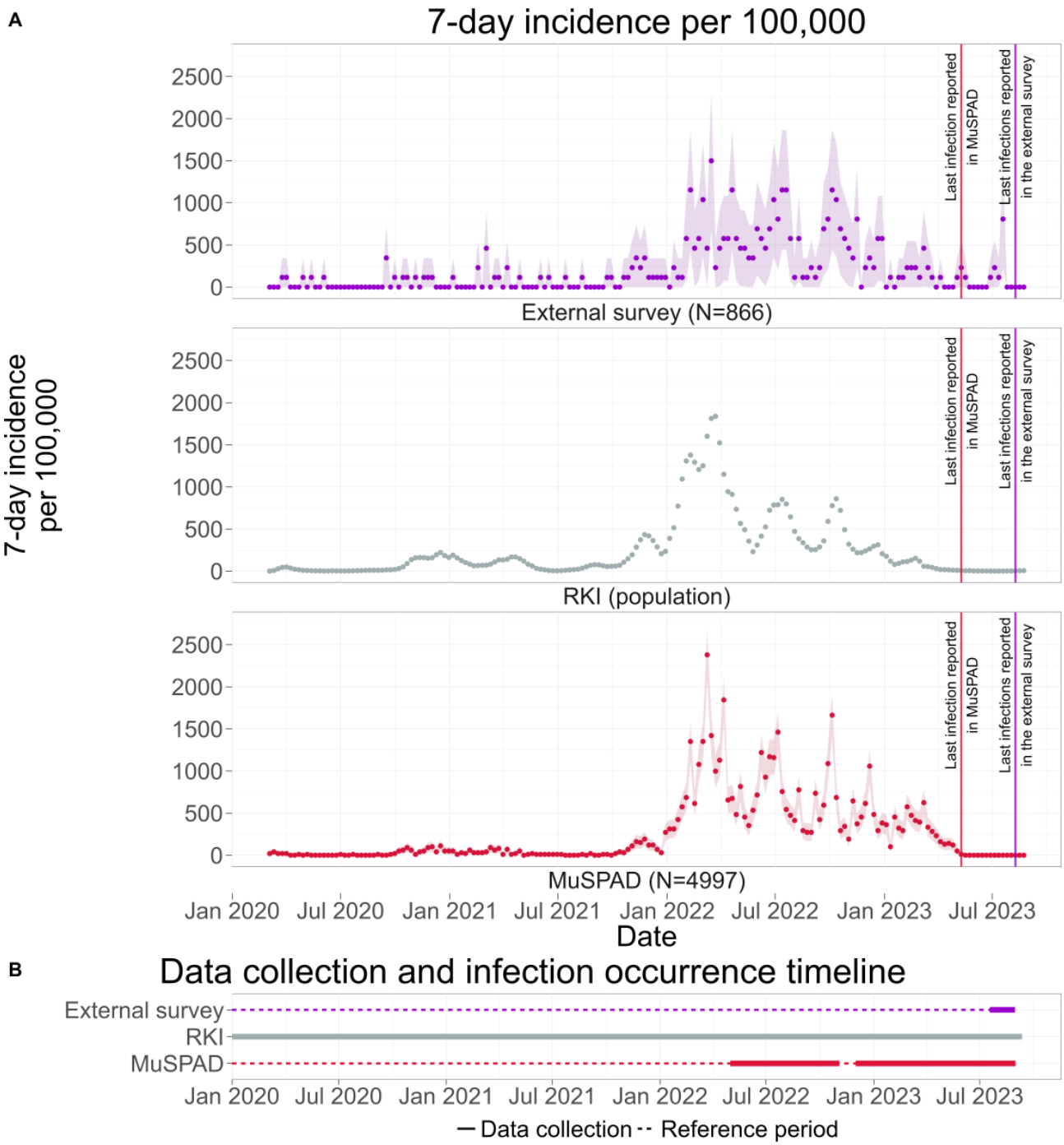

Spearman rank correlation analyses demonstrated strong positive associations between infection dates in the external survey and MuSPAD (ρ=0.884, *P*<.001) and between MuSPAD and RKI counts (ρ=0.831, *P*<.001). The correlation between the external survey and RKI counts was large as well (ρ=0.640, *P*<.001). Corresponding Pearson coefficients (0.839, 0.856, and 0.599, respectively) showed similar patterns of association. All correlations were statistically significant, indicating temporal concordance across the 3 data sources.

Both the external survey and the MuSPAD study undersample 1-person households. Since the external survey overrepresents 40- to 59-year-old people (see *Demographic Comparison* section), we applied bootstrapping to adjust the age distribution for comparability with RKI-reported 7-day incidence per

XSL•FO
RenderX

100,000 (see *Data Processing* section). As the adjustment had a negligible impact on trends, we present the raw data here, with bootstrapped results available in Bootstrapping in Multimedia Appendix 1.

The analysis of the number of vaccination doses received until September 11, 2023, shows comparable results between the external survey and the MuSPAD study (Figure 4). That is, almost all participants received at least 2 doses of a COVID-19 vaccine of any brand, with no discernible differences between the 18- to 39-year-olds, 40- to 59-year-olds, 60- to 79-year-olds, and 80- to 99-year-olds. However, the MuSPAD study shows slightly lower percentages. In both the external survey and the MuSPAD study, we find a small drop between the share of participants who reported receiving at least 2 doses and those who reported receiving 3 doses for the 18 to 39, 40 to 59, and 60 to 79 years age groups, not distinguishing between homologous or heterologous booster shots. A notable difference between the 2 studies emerges in the share of participants who received at least 4 doses of the COVID-19 vaccine, although fourth dose recommendations were intended for older or at-risk populations. Across all age groups, apart from the 80 to 99 years age group, the external survey reports higher vaccination rates than the MuSPAD study.

The comparison of vaccination numbers between the 2 studies and the officially reported numbers by the RKI reveals 2 key differences. First, the RKI splits the adult population only into 2 age groups (18‑59 and 60+ years), thus limiting the comparability. Second, it can be noted that both the external survey and the MuSPAD study struggle to recruit unvaccinated individuals and individuals who decided not to or could not receive a third or fourth vaccine dose. However, the reported variable COVID-19 vaccination status showed a considerable amount of missing data (300/867, 34.6%), which might include unvaccinated individuals.

XSL•FO
RenderX

Figure 4. (A) Shares of individuals who have received at least 1, 2, 3, or 4 doses of any COVID-19 vaccination by September 11, 2023. Both the external survey and the Multilocal and Serial Prevalence Study of Antibodies Against Respiratory Infectious Diseases (MuSPAD) study failed to recruit large numbers of unvaccinated individuals and individuals who decided not to get or could not get a booster shot. Error bars denote 95% CIs (see *Analysis Framework* section for details); for Robert Koch Institute (RKI) data, these are negligible due to the large sample size. (B) This timeline illustrates the data collection periods for 3 different sources: the external survey, the RKI dataset, and the MuSPAD study. The solid bold segments indicate the actual data collection periods, while the dotted lines represent the corresponding reference periods—time frames during which vaccinations could have been administered and retrospectively reported. The alignment of these periods ensures consistency in comparing vaccination data across sources. The external survey and MuSPAD are retrospective data collections, while the RKI uses continuous data collection.

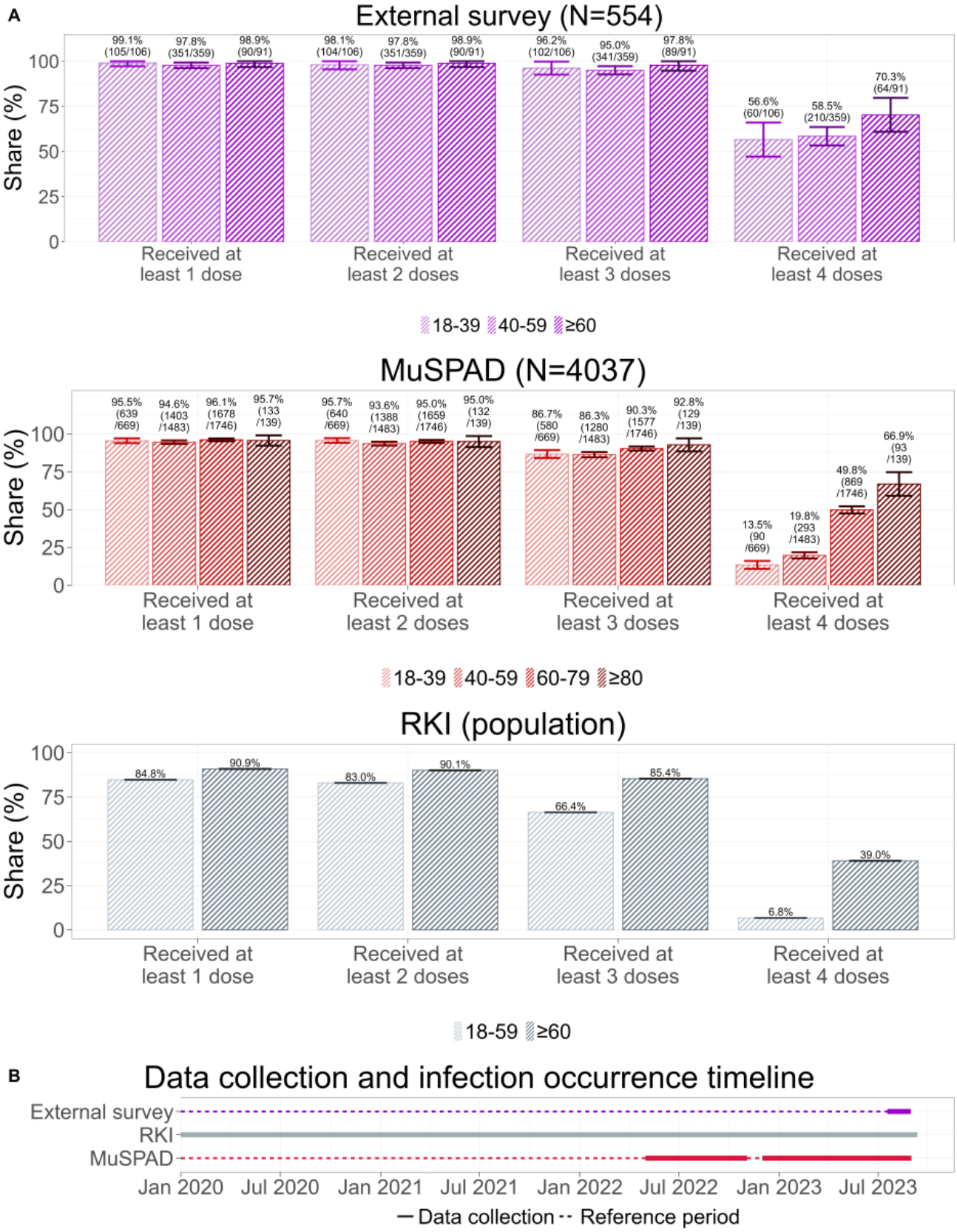

XSL•FO
RenderX

COVID-19–Related Comparison by Recruiter

Analysis of responses to the first Twitter or Mastodon question reveals that most votes stem from Recruiter 1 and Recruiter 2 (5589/6172, ~90%), making them the dominant contributors to the Twitter sample (Table 4).

Breaking down infection history by recruiter reveals variations in reported infection rates. The shares of the participants who reported zero infections on Twitter, for example, fluctuate between 18% (17/97, Recruiter 5) and 31% (642/2054, Recruiter 1 [Twitter]) (see Figure 5). On Mastodon, however, the share of participants who reported zero infections is higher (666/1765, 38%). A similar, though less pronounced, variation is observed among participants who reported one infection: 50% (882/1765) of participants recruited by Recruiter 1 (Mastodon), 54% (863/1605) of participants recruited by Recruiter 2, 54% (1117/2054) of participants recruited by Recruiter 1 (Twitter), 57% (60/106) of participants recruited by Recruiter 4, 59% (57/97) of participants recruited by Recruiter 5, and 60% (211/353) of participants recruited by Recruiter 3 reported one infection. Recruiter 5 recruited, with 23% (22/98), the largest share of participants who reported at least 2 infections. However, the limited number of responses from Recruiter 5's poll had little impact on the overall Twitter sample distribution. In summary, all recruiters drew similar samples regarding total infections.

Figure 5. Raw and estimated proportions of participants reporting 0, 1, or ≥2 COVID-19 infections in the first social media poll, by recruiter account and platform. Estimates were created using a generalized linear model with the recruiter as a random intercept. The figure compares responses from participants recruited via Twitter and Mastodon. Overall patterns were similar across recruiters; however, the largest proportion of participants reporting zero infections came from the Mastodon-recruited group. Error bars were created using parametric bootstrapping (500 simulations) on mixed-effects logistic regression models to generate 95% CIs. Participants who selected "show results" were excluded from this analysis. Data were collected in Germany between July 19 and July 26, 2023.

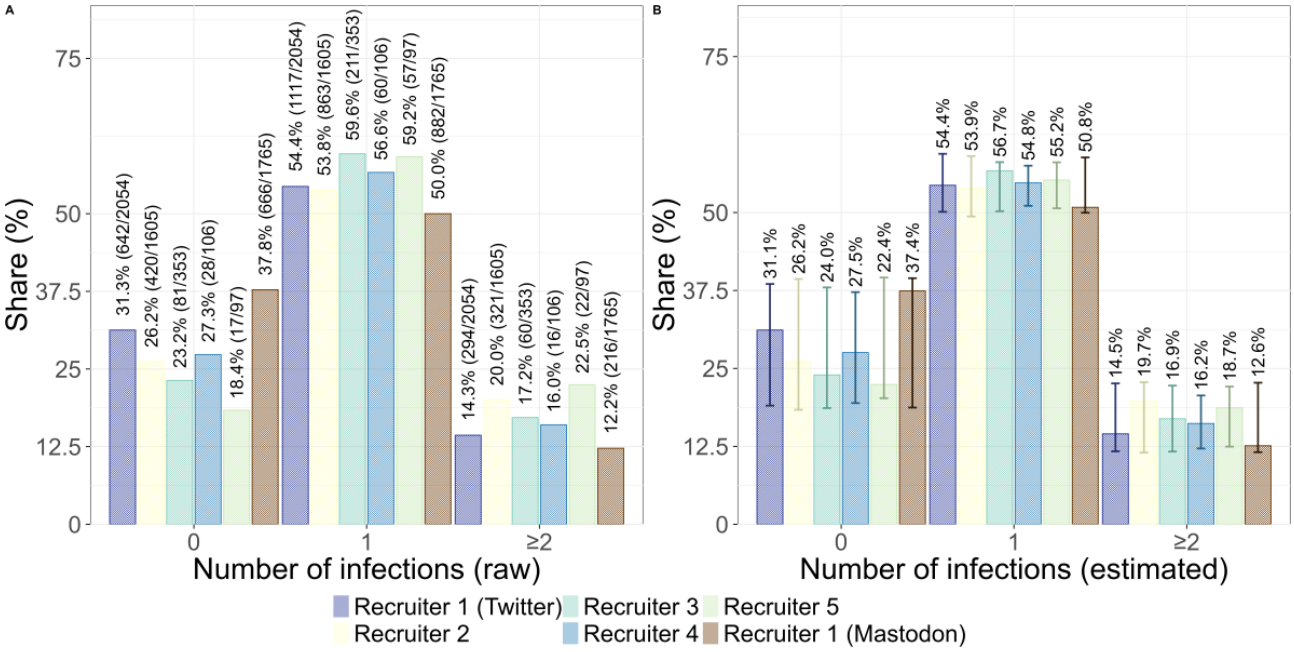

Table . Number of votes on the first social media poll, by recruiter account and platform.[a]

Recruiter	Number of votes (question 1 only)
Recruiter 1 (Twitter)	2120
Recruiter 2	1667
Recruiter 3	371
Recruiter 4	111
Recruiter 5	101
Recruiter 1 (Mastodon)	1802

[a]This table shows the distribution of responses across the 5 recruiter accounts. Only Recruiter 1 shared the poll on both Twitter and Mastodon, while the remaining four recruiters distributed the poll exclusively on Twitter. Data were collected in Germany between July 19 and July 26, 2023.

Demographic Comparison

A demographic comparison of the external survey and the MuSPAD study with official statistics reveals notable deviations on various levels. That is, females are slightly overrepresented in the external survey but less so than in the MuSPAD study (external survey: 464/857, 54%, MuSPAD: 5966/9842, 61%, FSO: 51%; Figure 6A). Consequently, males are slightly underrepresented in both studies (external survey: 386/857, 45%, MuSPAD: 3861/9842, 39%, FSO: 49%), while

XSL•FO
RenderX

respondents who indicated their gender as diverse make up less than 1% of both study's samples (the FSO only distinguishes between "female" and "male"). In addition, participants aged 40 to 59 years are substantially overrepresented in the external survey (external survey: 569/854, 67%, MuSPAD 3486/9814, 36%, FSO: 27%). In contrast, individuals aged 80 to 99 years are underrepresented in both the external survey and the MuSPAD study (external survey: 2/854, <1%, MuSPAD: 631/9814, 6%, FSO: 9%; Figure 6B). Both the survey and the MuSPAD study undersample 1-person households (external survey: 199/857, 23%, MuSPAD: 2134/7768, 28%; Figure 6C), which make up 41% of households in Germany according to the FSO. Consequently, larger household sizes are slightly overrepresented. Furthermore, most external survey and MuSPAD participants (external survey: 421/565, 75%, MuSPAD: 6931/8608, 81%) replied that they had no children under the age of 14 years (Figure 6D). One (external survey: 72/565, 13%, MuSPAD: 864/8608, 10%) and 2 (external survey: 60/565, 11%, MuSPAD: 667/8608, 8%) children under the age of 14 years are similarly likely in both studies, while only 2% (external survey: 12/565, 2%, MuSPAD: 146/8608, 2%) of respondents reported having 3 or more children under the age of 14 years. Participants who have received higher education are overrepresented in the external survey (external survey: 485/565, 86%, MuSPAD: 5085/9679, 52%, FSO: 34%). Thus, only 7% (37/565) reported they had obtained a certification after 10 years (MuSPAD: 2454/9679, 26%, FSO: 30%), and less than 1% (2/565) reported that they had obtained a certificate after 9 years (MuSPAD: 927/9679, 12%, FSO: 29%; Figure 6E). Finally, the external survey oversampled participants who reported their current occupation as "other" (external survey: 407/563, 72%, MuSPAD: 4558/9791, 47%, Federal Employment Agency: 55%), while it undersampled retired participants (external survey: 53/563, 9%, MuSPAD: 3546/9791, 36%, Federal Employment Agency: 30%;).

To conclude, both novel sampling methods (eg, external survey, SMPs) introduce biases regarding demographic data.

XSL•FO
RenderX

Figure 6. Comparison of sociodemographic distributions between the external survey, the Multilocal and Serial Prevalence Study of Antibodies Against Respiratory Infectious Diseases (MuSPAD) study, and the Federal Statistical Office (FSO) and Federal Employment Agency. Panel A shows the gender, panel B the age brackets, panel C the household size, panel D the number of children under 14 in the household, panel E the education level, and panel F the current occupation of participants. Participants who failed to answer the corresponding survey item were excluded from the analysis. Error bars represent 95% CIs (see *Analysis Framework* section for details); for FSO data, these are negligible due to the large sample size. Underage participants were excluded from the analysis, as neither the external survey nor the MuSPAD study recruited them (see panel B). Additionally, the FSO does not provide data on children younger than 14 years (panel D).

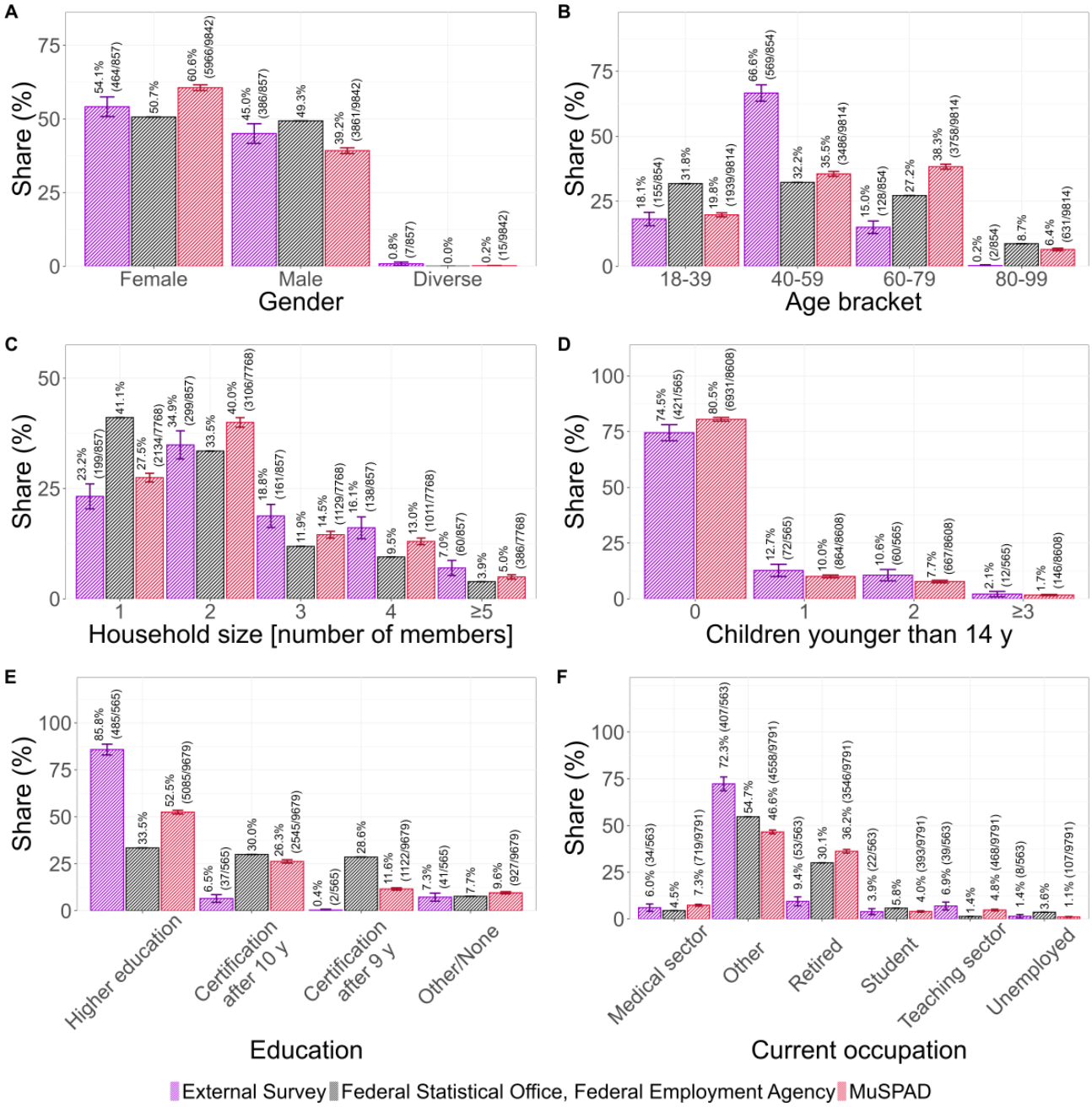

Discussion

Principal Findings

In the time period assessed here—3 years into a coronavirus pandemic and after the establishment of large testing schemes in the population—the data collected via SMPs mirror cumulative self-reported infections in the assessed epidemic panel. In particular, it matches the proportion of people reporting having had infections. This shows the potential of SMPs, which can be an adequate and cost-effective proxy for infection numbers in such situations and can provide rapid and readily available data for dynamic modeling. Epidemic panels are particularly valuable during pandemic situations, as they enable serological estimates of cumulative infection prevalence and real-time monitoring of infection rates, helping to identify potential underreporting of self-reported diagnoses. This applies to both early pandemic phases (before large-scale testing is available) and later intrapandemic periods (after interest in testing has declined), as well as to seasonal infections commonly monitored in epidemic panels, such as influenza, respiratory syncytial virus, or vector-borne diseases, where diagnosed cases often underestimate actual infections, particularly across different age groups.

XSL•FO
RenderX

If the focus is on the cumulative number of infections during mid-intrapandemic periods, near real-time and straightforward data collection methods, such as SMPs and the external survey, can provide rapid, directional insights, showing only small differences compared to larger serological studies like MuSPAD. Incidence gathered through recruitment via social media for an external survey is also strongly correlated with official data from public health authorities. While it is notable that our sample can produce such timely results at minimal cost, claims about reliability and representativeness should be tempered, acknowledging that they are limited to pandemic scenarios with similar disease characteristics such as COVID-19.

On the other hand, demographic comparisons reveal that 40‑to 59-year-olds and individuals with higher education are overrepresented in the data, while 1-person households are underrepresented. The higher education level is expected because all recruiters are part of academia. Due to legal reasons, participants under 18 years could not participate in the SMP, which is a drawback compared to the official testing conducted by the RKI.

While these sample biases and statistically significant differences in infection numbers across different platforms did exist, effect sizes were small, indicating that SMPs could be useful in ascertaining infection trends in the population in a rapid and cost-effective way.

Comparison With Prior Work

Similar to what Schober et al [33] suggest, SMPs might be used as additional, cost-effective interim or on-top data collection tools to enrich official data. We go beyond what Vidal-Alaball et al [47] achieved in their study, focusing on collecting data instead of evaluating public opinion. As Zhao et al [36] point out, we did consider bias in our data and explicitly tested for this, observing inherently higher education levels in the Twitter data compared to other survey types. While social media data have previously been used for opinion mining and behavioral studies in public health, SMPs have not been systematically applied to directly collect self-reported infection or incidence data.

In terms of comparison to serological indicators of cumulative infections at certain time periods, the presented data on self-reported infections are in line with published data from the IMMUNEBRIDGE project, of which the MuSPAD cohort was one part [62]. Comparison with COSMO data revealed that, due to the earlier data collection period of COSMO, infection numbers were lower compared to our data and MuSPAD, possibly indicating that a large portion of first infections occurred in 2023.

Data until 2022 showed that survey-based incidence aligned closely with the officially reported figures from the RKI. However, a discrepancy emerged with the onset of the Omicron BA.5 wave in the summer of 2022. This could be due to individuals confirming infections with rapid antigen tests that were not reported or because many experienced only mild symptoms and did not seek medical attention, thus not entering official statistics. SMPs here allow accounting for this decline in official testing. Additionally, a potential limitation was that in the MuSPAD data, reported infections in the time frame April 1, 2023, to August 31, 2023, were limited to a single date, excluding possible secondary infections. Due to this, infection numbers during that time frame might be underreported.

Recall bias might have affected the reported timing of infections. shows that the incidence still aligns quite well with official reporting, indicating that the effect of recall bias was small overall.

Regarding vaccination doses, our data align well with the MuSPAD data except for the fourth vaccination dose. Here, our limited sample size regarding older age groups (60+) might be problematic because the fourth vaccination is recommended by the RKI only for these older age groups. Overall, compared to the RKI data, our sample and the MuSPAD sample generally have a higher proportion of vaccinated individuals. This might result from sampling bias, as very risk-averse and protective people might have a higher tendency to participate in a survey regarding a potentially dangerous infectious disease. We also report a high number of missing data points for the COVID-19 vaccination status, which might include unvaccinated individuals who do not want to report their status due to social desirability effects. The effect being present in the MuSPAD data also shows the difficulty in mitigating these effects even in high-effort, high-cost studies.

For the successful recruitment of a large enough sample size, a sizable followership on social media is required. Also, demographic comparisons have shown that biases exist when recruiters are not diverse regarding age, education, and background. Public health–related SMPs are more feasible in the later stages of a pandemic, when tests are widely available to the public and sufficient sentinel studies have already been conducted.

Overall, SMPs can complement traditional methods by providing real-time insights but cannot replace them due to data quality and representativeness concerns.

Limitations and Future Work

Finally, it should be noted that there are limitations to SMPs: one does not obtain a list of the users who participated in the survey, but solely the proportion of responses for each option. Thus, this method cannot record demographic attributes, and subanalysis for specific target groups is impossible. Furthermore, only a subset of the population is active on Twitter, limiting the generalizability of these findings.

Another limitation of our study is the very small number of unvaccinated individuals included in the sample (external survey: n=10, 1.76%; MuSPAD: n=489, 3.63%), although the large number of participants who did not provide an answer here might include unvaccinated individuals.

Using 5 academically skewed accounts for recruitment generated 36,416 total Twitter impressions but introduced selection bias through extreme concentration (2 accounts controlled 87% of reach) and network homophily (followers share similar educated, professionally engaged demographics). Response rates of 17% to 19% indicate high engagement; however, the 56% drop in impressions and the 53% drop in votes between Question 1 and

XSL•FO
RenderX

Question 2 reveal limitations in data gathering. Mastodon provided 2540 votes, but no impression data were available due to platform limitations, which prevented the calculation of response rates. This recruitment strategy systematically excludes individuals outside academic networks, thereby limiting the findings to "engaged followers of academic social media accounts" rather than more generalizable populations. Additionally, not all recruiters generated enough responses to detect small effect sizes when comparing them to each other. Although the number of infections shows variation, small sample sizes limit any meaningful analysis here.

Due to the limited data collected via Mastodon, H2 could not be fully answered. The Mastodon data, provided solely by Recruiter 1, included a higher proportion of single-person households compared to the overall sample, which may help explain the greater number of participants reporting no infections.

Future work includes improving the representativeness of SMPs, possibly through targeted outreach via ads on Twitter or Facebook, the inclusion of more platforms for data collection (eg, Threads, Bluesky, Facebook), and generating longitudinal data through weekly sampling via a proposed Twitter poll posting tool. This posting tool automatically posts polls every week or 2 weeks to inquire about infections, offering a highly cost-efficient method and allowing for automatic weekly evaluations. Additionally, to test the reliability of this method, a study with in-person testing could be run in parallel.

Conclusions

Especially for recruitment, SMPs show promising results. We were able to generate a sizable sample, although it must be mentioned that this depends heavily on platform follower numbers. We showed that SMPs are an adequate and cost-effective tool for rapidly collecting health data. While overall demographic representativeness is good, discrepancies in age and education may impact generalizability. Especially for health data—in this case, infection numbers and incidence—data quality is comparable to that of more costly and high-effort panel studies. To respond to pandemic scenarios with similar disease characteristics such as COVID-19, data collected via SMPs can quickly provide accurate enough data to help with modeling efforts.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy [63], the following tasks were delegated to GAI tools under full human supervision: code generation, code optimization, and proofreading and editing.

The GAI tools used were ChatGPT (GPT-4.5), Claude 4.5, Grammarly, and LanguageTool.

Responsibility for the final manuscript lies entirely with the authors.

GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Declaration submitted by: collective responsibility.

Additional note: OpenAI ChatGPT, Claude AI, Grammarly AI, and LanguageTool were used for grammar checks in the main text, and GitHub Copilot served as a coding assistant. The authors assume full responsibility for the final content of the paper.

Funding

The work on the paper was partly funded by the Ministry of Research and Education (BMBF), Germany (grant 031L0300A, 031L0300C, 031L0300D, 031L0302A), by the Max Planck Society, and by TU Berlin. The initial funding for the MuSPAD study was provided by the Initiative and Networking Fund of the Helmholtz Association of German Research Centers under grant number SO-96. The NAKO study is supported by the Federal Ministry of Education and Research (BMBF) (grant 01ER1301A/B/C, 01ER1511D, 01ER1801A/B/C/D, and 01ER2301A/B/C), along with contributions from the federal states of Germany, the Helmholtz Association, participating universities, and institutes of the Leibniz Association. The funder of the study had no role in the design of the study, data collection, analysis, interpretation, the writing of the manuscript, or the decision to submit it for publication. No authors were paid by any pharmaceutical company or other agency to write this article. The authors were not precluded from accessing the data and accept full responsibility for the decision to submit this manuscript for publication.

Data Availability

The datasets generated or analyzed during this study are available in the OSF repository [48], and the analysis code is available on GitHub [53]. Commit 9598b4c was used to produce the figures. We will share anonymized MuSPAD data used in this study with other academic researchers upon request.

Authors' Contributions

Conceptualization: VP (lead), MM (equal), HN (equal), SP (equal), LS (equal), JF (equal), BL (equal), ACV (equal)
Data curation: MM (equal), HN (equal), SP (equal), LS (equal)
Formal analysis: SP

XSL•FO
RenderX

Funding acquisition: ACV (equal), VP (equal)
Investigation: MM (equal), HN (equal), LS (equal)
Methodology: MM (equal), HN (equal), SP (equal), LS (equal), JF (supporting), ACV (supporting), VP (supporting)
Project administration: HN (equal), LS (equal), ACV (equal), VP (equal)
Resources: JF (equal), HN (equal), JK (equal), MH (equal), BL (equal)
Software: MM (equal), HN (equal), SP (equal), LS (equal), JF (equal)
Supervision: ACV (equal), VP (equal)
Validation: SP (equal), LS (equal), MM (equal)
Visualization: SP (lead), LS (supporting)
Writing – original draft: MM (equal), HN (equal), SP (equal), LS (equal)
Writing – review & editing: LS (lead), MM (equal), HN (equal), SP (equal), JF (supporting), JK (supporting), ACV (supporting), VP (supporting), MH (supporting), BL (supporting)
All authors reviewed and approved the final manuscript.
The MuSPAD study group consists of these members: Claudia Denkinger, Lisa Koeppel, Laura-Inés Boehler, Viola Priesemann, Sebastian Contreras, Philipp Dönges, Veronika K. Jaeger, André Karch, Berit Lange, Manuela Harries, Carolina Klett-Tammen, Torben Heinsohn, Isti Rodiah, Olga Hovardovska, Rafael Mikolajczyk, Cornelia Gottschick, Ulrich Reinacher, Felix Guenther, Melanie Schienle, Daniel Wolffram, Johannes Bracher, Alex Dulovic, Patrick Marsall, Daniel Junker, Nicole Schneiderhan-Marra, Wolfgang Bock, Tyll Krüger, Alex Kuhlmann, Rolf Kaiser, Michael Böhm, Nils Bardeck

Conflicts of Interest

VP has received funding from public institutions for this and related research. Additionally, she has received honoraria for talks regarding COVID-19 and scientific outreach. BL is a member of several expert councils regarding vaccination and public health. She has received funding from EU Horizon, EU4Health, Helmholtz, DFG, BMFTR, VW-Foundation, DAAD, GIZ, and state programs. She has been part of committees publishing guidelines and recommendations at state, national, and international levels and is active as a representative of professional societies such as the German Society of Epidemiology. As part of this work, she has also published on the relevance of infectious disease–specific cohorts, changes to the (sentinel) surveillance system to inform vaccine guidelines, and necessary digital transformations. ACV has also received funding from public institutions for this and related research.

Multimedia Appendix 1
Supplementary materials showing further data on bootstrapping, timing of infection, demographic comparisons by Twitter/Mastodon recruiter, and comparisons of vaccine suppliers.
[DOCX File, 2988 KB - jmir_v28i1e80311_app1.docx]

References

1. Jee Y. WHO International Health Regulations Emergency Committee for the COVID-19 outbreak. Epidemiol Health 2020;42:e2020013. [doi: 10.4178/epih.e2020013] [Medline: 32192278]
2. Cucinotta D, Vanelli M. WHO declares COVID-19 a pandemic. Acta Biomed 2020 Mar 19;91(1):157-160. [doi: 10.23750/abm.v91i1.9397] [Medline: 32191675]
3. Clemente-Suárez VJ, Navarro-Jiménez E, Moreno-Luna L, et al. The impact of the COVID-19 pandemic on social, health, and economy. Sustainability 2021;13(11):6314. [doi: 10.3390/su13116314]
4. Delardas O, Kechagias KS, Pontikos PN, Giannos P. Socio-economic impacts and challenges of the coronavirus pandemic (COVID-19): an updated review. Sustainability 2022;14(15):9699. [doi: 10.3390/su14159699]
5. Dubey S, Biswas P, Ghosh R, et al. Psychosocial impact of COVID-19. Diabetes Metab Syndr 2020 Sep;14(5):779-788. [doi: 10.1016/j.dsx.2020.05.035] [Medline: 32526627]
6. Gimma A, Munday JD, Wong KLM, et al. Changes in social contacts in England during the COVID-19 pandemic between March 2020 and March 2021 as measured by the Comix survey: a repeated cross-sectional study. PLoS Med 2022 Mar;19(3):e1003907. [doi: 10.1371/journal.pmed.1003907] [Medline: 35231023]
7. Onyeaka H, Anumudu CK, Al-Sharify ZT, Egele-Godswill E, Mbaegbu P. COVID-19 pandemic: a review of the global lockdown and its far-reaching effects. Sci Prog 2021;104(2):368504211019854. [doi: 10.1177/00368504211019854] [Medline: 34061685]
8. Bielecki M, Patel D, Hinkelbein J, et al. Air travel and COVID-19 prevention in the pandemic and peri-pandemic period: a narrative review. Travel Med Infect Dis 2021;39:101915. [doi: 10.1016/j.tmaid.2020.101915] [Medline: 33186687]
9. Sun KS, Lau TSM, Yeoh EK, et al. Effectiveness of different types and levels of social distancing measures: a scoping review of global evidence from earlier stage of COVID-19 pandemic. BMJ Open 2022 Apr 11;12(4):e053938. [doi: 10.1136/bmjopen-2021-053938] [Medline: 35410924]
10. Islam N, Sharp SJ, Chowell G, et al. Physical distancing interventions and incidence of coronavirus disease 2019: natural experiment in 149 countries. BMJ 2020 Jul 15;370:m2743. [doi: 10.1136/bmj.m2743] [Medline: 32669358]

XSL•FO
RenderX

11. Kharroubi S, Saleh F. Are lockdown measures effective against COVID-19? Front Public Health 2020;8:549692. [doi: 10.3389/fpubh.2020.549692] [Medline: 33194950]
12. Li T, Liu Y, Li M, Qian X, Dai SY. Mask or no mask for COVID-19: a public health and market study. PLoS ONE 2020;15(8):e0237691. [doi: 10.1371/journal.pone.0237691] [Medline: 32797067]
13. Liao M, Liu H, Wang X, et al. A technical review of face mask wearing in preventing respiratory COVID-19 transmission. Curr Opin Colloid Interface Sci 2021 Apr;52:101417. [doi: 10.1016/j.cocis.2021.101417] [Medline: 33642918]
14. Prati G, Mancini AD. The psychological impact of COVID-19 pandemic lockdowns: a review and meta-analysis of longitudinal studies and natural experiments. Psychol Med. p. 201-211. [doi: 10.1017/S0033291721000015] [Medline: 33436130]
15. Przekwas A, Chen Z. Washing hands and the face may reduce COVID-19 infection. Med Hypotheses 2020 Nov;144:110261. [doi: 10.1016/j.mehy.2020.110261] [Medline: 33254560]
16. Teslya A, Rozhnova G, Pham TM, et al. The importance of sustained compliance with physical distancing during COVID-19 vaccination rollout. Commun Med (Lond) 2022 Nov 19;2(1):146. [doi: 10.1038/s43856-022-00207-3] [Medline: 36402924]
17. Perra N. Non-pharmaceutical interventions during the COVID-19 pandemic: a review. Phys Rep 2021 May 23;913(1–52):1-52. [doi: 10.1016/j.physrep.2021.02.001] [Medline: 33612922]
18. Dehning J, Zierenberg J, Spitzner FP, et al. Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. Science 2020 Jul 10;369(6500):eabb9789. [doi: 10.1126/science.abb9789] [Medline: 32414780]
19. Dönges P, Wagner J, Contreras S, et al. Interplay between risk perception, behavior, and COVID-19 spread. Front Phys 2022;10:842180. [doi: 10.3389/fphy.2022.842180]
20. Kretzschmar ME, Rozhnova G, Bootsma MCJ, van Boven M, van de Wijgert JHHM, Bonten MJM. Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. Lancet Public Health 2020 Aug;5(8):e452-e459. [doi: 10.1016/S2468-2667(20)30157-2] [Medline: 32682487]
21. Moore S, Hill EM, Tildesley MJ, Dyson L, Keeling MJ. Vaccination and non-pharmaceutical interventions for COVID-19: a mathematical modelling study. Lancet Infect Dis 2021 Jun;21(6):793-802. [doi: 10.1016/S1473-3099(21)00143-2] [Medline: 33743847]
22. Nunner H, van de Rijt A, Buskens V. Prioritizing high-contact occupations raises effectiveness of vaccination campaigns. Sci Rep 2022 Jan 14;12(1):737. [doi: 10.1038/s41598-021-04428-9] [Medline: 35031651]
23. Teslya A, Pham TM, Godijk NG, Kretzschmar ME, Bootsma MCJ, Rozhnova G. Impact of self-imposed prevention measures and short-term government-imposed social distancing on mitigating and delaying a COVID-19 epidemic: a modelling study. PLoS Med 2020 Jul;17(7):e1003166. [doi: 10.1371/journal.pmed.1003166] [Medline: 32692736]
24. Thompson RN. Epidemiological models are important tools for guiding COVID-19 interventions. BMC Med 2020 May 25;18(1):152. [doi: 10.1186/s12916-020-01628-4] [Medline: 32448247]
25. Alamo T, G Reina D, Millán Gata P, Preciado VM, Giordano G. Data-driven methods for present and future pandemics: monitoring, modelling and managing. Annu Rev Control 2021;52:448-464. [doi: 10.1016/j.arcontrol.2021.05.003] [Medline: 34220287]
26. Kretzschmar ME, Ashby B, Fearon E, et al. Challenges for modelling interventions for future pandemics. Epidemics 2022 Mar;38:100546. [doi: 10.1016/j.epidem.2022.100546] [Medline: 35183834]
27. Shadbolt N, Brett A, Chen M, et al. The challenges of data in future pandemics. Epidemics 2022 Sep;40:100612. [doi: 10.1016/j.epidem.2022.100612] [Medline: 35930904]
28. Kettlitz R, Harries M, Ortmann J, et al. Association of known SARS-CoV-2 serostatus and adherence to personal protection measures and the impact of personal protective measures on seropositivity in a population-based cross-sectional study (MuSPAD) in Germany. BMC Public Health 2023 Nov 17;23(1):2281. [doi: 10.1186/s12889-023-17121-5] [Medline: 37978484]
29. Gornyk D, Harries M, Glöckner S, et al. SARS-CoV-2 seroprevalence in Germany. Dtsch Arztebl Int 2021 Dec 3;118(48):824-831. [doi: 10.3238/arztebl.m2021.0364] [Medline: 35191825]
30. Betsch C, Wieler L, Bosnjak M, et al. Germany COVID-19 Snapshot Monitoring (COSMO Germany): monitoring knowledge, risk perceptions, preventive behaviours, and public trust in the current coronavirus outbreak in Germany. PsychArchives. Preprint posted online on Mar 3, 2020. [doi: 10.23668/psycharchives.2776]
31. Brito KDS, Silva Filho RLC, Adeodato PJL. A systematic review of predicting elections based on social media data: research challenges and future directions. IEEE Trans Comput Soc Syst 2021;8(4):819-843. [doi: 10.1109/TCSS.2021.3063660]
32. Kim A, Murphy J, Richards A, Hansen H, Powell R, Haney C. Can tweets replace polls? A U.S health-care reform case study. In: Social Media, Sociality, and Survey Research: John Wiley & Sons Inc; 2013:61-86. [doi: 10.1002/9781118751534.ch3]
33. Schober MF, Pasek J, Guggenheim L, Lampe C, Conrad FG. Social media analyses for social measurement. Public Opin Q 2016;80(1):180-211. [doi: 10.1093/poq/nfv048] [Medline: 27257310]
34. Morstatter F, Liu H. Discovering, assessing, and mitigating data bias in social media. Online Soc Netw Med 2017 Jun;1:1-13. [doi: 10.1016/j.osnem.2017.01.001]

XSL•FO
RenderX

35. Olteanu A, Castillo C, Diaz F, Kıcıman E. Social data: biases, methodological pitfalls, and ethical boundaries. Front Big Data 2019;2(13):13. [doi: 10.3389/fdata.2019.00013] [Medline: 33693336]
36. Zhao Y, He X, Feng Z, et al. Biases in using social media data for public health surveillance: a scoping review. Int J Med Inform 2022 Aug;164:104804. [doi: 10.1016/j.ijmedinf.2022.104804] [Medline: 35644051]
37. Campos-Castillo C, Laestadius LI. Racial and ethnic digital divides in posting COVID-19 content on social media among US adults: secondary survey analysis. J Med Internet Res 2020 Jul 3;22(7):e20472. [doi: 10.2196/20472] [Medline: 32568726]
38. Eibensteiner F, Ritschl V, Nawaz FA, et al. People's willingness to vaccinate against COVID-19 despite their safety concerns: Twitter poll analysis. J Med Internet Res 2021 Apr 29;23(4):e28973. [doi: 10.2196/28973] [Medline: 33872185]
39. Cinelli M, Quattrociocchi W, Galeazzi A, et al. The COVID-19 social media infodemic. Sci Rep 2020 Oct 6;10(1):16598. [doi: 10.1038/s41598-020-73510-5] [Medline: 33024152]
40. Gabarron E, Oyeyemi SO, Wynn R. COVID-19-related misinformation on social media: a systematic review. Bull World Health Organ 2021 Jun 1;99(6):455-463A. [doi: 10.2471/BLT.20.276782] [Medline: 34108756]
41. Gisondi MA, Barber R, Faust JS, et al. A deadly infodemic: social media and the power of COVID-19 misinformation. J Med Internet Res 2022 Feb 1;24(2):e35552. [doi: 10.2196/35552] [Medline: 35007204]
42. Joseph AM, Fernandez V, Kritzman S, et al. COVID-19 misinformation on social media: a scoping review. Cureus 2022 Apr;14(4):e24601. [doi: 10.7759/cureus.24601] [Medline: 35664409]
43. Malik AA, Bashir F, Mahmood K. Antecedents and consequences of misinformation sharing behavior among adults on social media during COVID-19. Sage Open 2023;13(1):21582440221147022. [doi: 10.1177/21582440221147022] [Medline: 36699545]
44. Srivastava KC, Shrivastava D, Chhabra KC, Naqvi WM, Sahu A. Facade of media and social media during COVID-19: a review. Int J Res Pharm Sci 2020;11(SPL1):142-149. [doi: 10.26452/ijrps.v11iSPL1.2288]
45. Dal Moro F. Online survey on twitter: a urological experience. J Med Internet Res 2013 Oct 25;15(10):e238. [doi: 10.2196/jmir.2719] [Medline: 24164710]
46. Loeb S, Roupret M, Van Oort I, et al. Novel use of Twitter to disseminate and evaluate adherence to clinical guidelines by the European Association of Urology. BJU Int 2017 Jun;119(6):820-822. [doi: 10.1111/bju.13802] [Medline: 28170154]
47. Vidal-Alaball J, Fernandez-Luque L, Marin-Gomez FX, Ahmed W. A new tool for public health opinion to give insight into telemedicine: Twitter poll analysis. JMIR Form Res 2019 May 28;3(2):e13870. [doi: 10.2196/13870] [Medline: 31140442]
48. Twitter survey data generation. Open Science Framework (OSF). 2023. URL: https://osf.io/rtjzu [accessed 2025-07-09]
49. Paltra S, Stellbrink L, Friedel J, et al. Bimodal contact reductions and social homophily during COVID-19. medRxiv. Preprint posted online on Jul 11, 2025. [doi: 10.1101/2025.07.10.25331264]
50. von Elm E, Altman DG, Egger M, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007 Oct 20;335(7624):806-808. [doi: 10.1136/bmj.39335.541782.AD] [Medline: 17947786]
51. R: a language and environment for statistical computing. : R Foundation for Statistical Computing; 2024 URL: https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf [accessed 2026-03-20]
52. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. J Open Source Softw 2019;4(43):1686. [doi: 10.21105/joss.01686]
53. Hciuse/twitter-study. GitHub. 2025. URL: https://github.com/hciuse/twitter-study [accessed 2025-07-09]
54. Ergebnisse aus dem COVID-19 snapshot monitoring COSMO: die psychologische lage [Report in German]. : Universität Erfurt; 2022 URL: https://projekte.uni-erfurt.de/cosmo2020/files/COSMO_W70.pdf [accessed 2026-03-20]
55. 7-tage-inzidenz der COVID-19-fälle in deutschland [Article in German]. Zenodo. 2024. URL: https://zenodo.org/doi/10.5281/zenodo.14021103 [accessed 2026-03-20]
56. COVID-19-impfungen in deutschland [Article in German]. Robert koch-institut. 2024. URL: https://robert-koch-institut.github.io/COVID-19-Impfungen_in_Deutschland/ [accessed 2026-03-20]
57. Bevölkerung nach nationalität und geschlecht [Article in German]. Statistisches Bundesamt (Destatis). 2024. URL: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Bevoelkerungsstand/Tabellen/zensus-geschlecht-staatsangehoerigkeit-2024-basis-2022.html [accessed 2026-03-20]
58. Bevölkerung nach Altersgruppen [Article in German]. Statistisches Bundesamt (Destatis). 2024. URL: https://web.archive.org/web/20251021045952/https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Bevoelkerungsstand/Tabellen/bevoelkerung-altersgruppen-deutschland.html [accessed 2025-10-21]
59. Haushalte und haushaltsmitglieder [Article in German]. Statistisches Bundesamt (Destatis). 2024. URL: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Haushalte-Familien/Tabellen/1-1-privathaushalte-haushaltsmitglieder.html [accessed 2026-03-20]
60. Bevölkerung im alter von 15 jahren und mehr nach allgemeinen und beruflichen bildungsabschlüssen nach jahren [Article in German]. Statistisches Bundesamt (Destatis). 2024. URL: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Bildungsstand/Tabellen/bildungsabschluss.html [accessed 2026-03-20]

XSL•FO
RenderX

61. Beschäftigte nach berufen [Article in German]. Beschäftigte nach Berufen (KldB 2010) (Quartalszahlen). 2010. URL: https://web.archive.org/web/20250708061505/https://statistik.arbeitsagentur.de/Statistikdaten/Detail/202306/iiia6/beschaeftigung-sozbe-bo-heft/bo-heft-d-0-202306-xlsx.xlsx?__blob=publicationFile&v=2 [accessed 2025-07-08]
62. Lange B, Jaeger VK, Harries M, et al. Estimates of protection levels against SARS-CoV-2 infection and severe COVID-19 in Germany before the 2022/2023 winter season: the IMMUNEBRIDGE project. Infection 2024 Feb;52(1):139-153. [doi: 10.1007/s15010-023-02071-2] [Medline: 37530919]
63. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (Generative AI Delegation Taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res 2026 Apr;33(3):1-27. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]

Abbreviations

COSMO: COVID-19 Snapshot Monitoring
FSO: Federal Statistical Office of Germany
MuSPAD: Multilocal and Serial Prevalence Study of Antibodies Against Respiratory Infectious Diseases
NPI: nonpharmaceutical intervention measure
RKI: Robert Koch Institute
SMP: social media poll
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
WHO: World Health Organization

Edited by A Mavragani; submitted 03.Sep.2025; peer-reviewed by H Balakrishna, H Kitiabi, M Ekwueme, O Oluwole; revised version received 27.Feb.2026; accepted 27.Feb.2026; published 27.Apr.2026.

Please cite as:
Mortaga M, Nunner H, Paltra S, Stellbrink L, Friedel J, Harries M, Krepel J, Lange B, MuSPAD Study Group, Priesemann V, Calero Valdez A
Rapid Epidemiological Data Collection on Social Media for COVID-19: Comparative Study Between Online Surveys and Conventional Cohorts
J Med Internet Res 2026;28:e80311
URL: https://www.jmir.org/2026/1/e80311
doi:10.2196/80311

© Maged Mortaga, Hendrik Nunner, Sydney Paltra, Leonard Stellbrink, Jens Friedel, Manuela Harries, Jessica Krepel, Berit Lange, MuSPAD Study Group, Viola Priesemann, André Calero Valdez. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Experiences of Home-Dwelling Older Adults and Their Family Caregivers With Digital Health Services: Qualitative Study

JiaYi Ruan*, MSc; CuiXia Guan*, MSc; QingQing Chen*, MSc; Ting Yao, MSc; Yun Liu, MSc; YuXia Zhang, MSc; LunFang Xie, PhD

School of Nursing, Anhui Medical University, No. 81 Meishan Road, Hefei, AnHui, China

*these authors contributed equally

Corresponding Author:
LunFang Xie, PhD
School of Nursing, Anhui Medical University, No. 81 Meishan Road, Hefei, AnHui, China

Abstract

Background: As the population ages, demand for health services among older adults and caregivers increases. Digital health services help meet this demand. However, acceptance and experiences differ between older adults and caregivers.

Objective: This study aims to explore the specific experiences of home-dwelling older adults and their caregivers with digital health services, identify their willingness to use these services and the factors influencing their adoption, and identify their service needs to inform the subsequent design and optimization of digital health products and services.

Methods: From December 2023 to February 2024, a descriptive qualitative study was conducted. Researchers used purposive and maximum variation sampling to recruit 18 home-dwelling older adults and 17 family caregivers, including 10 matched dyads. Participants came from 7 community health service centers and a tertiary hospital in Hefei, China. Open-ended, semistructured, face-to-face interviews were conducted separately. Data were analyzed using conventional content analysis.

Results: Four main themes emerged. The first were the application characteristics, including usage situations and preferred functions. The second theme described specific service experiences, both positive and negative. The third was the influencing factors. Promoting factors included health needs, user experience, and subjective norms. Obstructive factors differed between groups and included low digital literacy, weak economic foundation, service security concerns, and lack of time. The fourth theme covered suggestions and expectations for improvement. Older adults appreciated the advantages but often had suboptimal experiences due to complex designs and threats to personal dignity. Conversely, caregivers valued efficiency but reported low use, hindered by time constraints and concerns about privacy and safety.

Conclusions: This study reveals a critical gap between the potential and actual use of digital health services. Older adults face barriers with usability and digital literacy. Caregivers struggle to integrate these services and have trust issues. Future optimization requires action at several levels. Government support is needed through policies, authoritative platforms, and subsidies. Service providers should offer age-friendly, personalized design, and continuous support. Social support systems are also essential, such as digital reciprocity from family and peer help. These strategies are critical for enhancing user experience, bridging the digital divide, and enabling both older adults and their caregivers to fully benefit from digital health technologies.

(J Med Internet Res 2026;28:e89496) doi:10.2196/89496

KEYWORDS

older adults; caregiver; digital health service; qualitative study; user experience; health literacy

Introduction

Population aging is a major global challenge that needs immediate action in the 21st century. According to the World Population Prospects 2022 released by the United Nations Department of Economic and Social Affairs (DESA), the proportion of the global population aged 65 years and older is projected to rise significantly from 10% in 2022 to 16% by 2050 [1]. This trend indicates that the global phenomenon of population aging is intensifying continuously. In high-income countries, an aging society has already become a reality. In transitional and low-income countries, the growth of the older adult population is also particularly notable. China, for instance, ranks among the countries at the forefront in terms of both the speed and scale of population aging [2].

With population aging and longer life expectancy, the burden of chronic diseases has grown. As a result, the demand for health management among older adults has increased significantly. Patients with chronic diseases have long-term health management needs. However, due to scarce medical resources, these needs remain unmet. This hinders both health decisions and engagement in self-management [3]. Due to factors such

XSL•FO
RenderX

as multimorbidity, low educational attainment, and poor digital health literacy [4], older adults with multiple chronic conditions face more complex and challenging health management barriers [5] and often need to rely on the assistance of caregivers and geriatric nursing staff.

In Europe, general practitioners or family medicine practitioners are considered the cornerstone of primary care [6]. In contrast, health management for older adults in Chinese communities relies more heavily on long-term family caregivers, such as spouses and children [7]. These caregivers may not be professional health care workers, yet they are required to take on tasks including assisting with daily living activities and accessing medical resources. A cross-sectional study on the burden of caregivers for disabled older adults in rural areas found that 77.6% of caregivers had changed their employment status due to caregiving responsibilities [8]. Behind the caregiver role lie unpaid labor, forgone employment opportunities, and economic costs. Chronic diseases such as cardiovascular and cerebrovascular diseases, diabetes, and osteoporosis have a high prevalence among the older population and require long-term management and care. However, as care needs increase, caregivers may experience a decline in their physical health, along with increased social and psychological stress [9,10].

To address the numerous challenges in health management, a growing number of older adults and their caregivers have turned to digital health services. At present, digital health services are widely used as tools for accessing personal health knowledge and conducting self-health management [11]. In low- and middle-income countries, digital health promotion technologies targeting young people have become increasingly prevalent [12]. However, digital technology is not exclusive to young people; a growing number of older adults are breaking stereotypes and beginning to use digital health services to achieve self-health management and access primary care. In China, the implementation forms of digital health services for older adults mainly include health monitoring, health management, daily living assistance, and internet-based diagnosis and treatment [13-15]. In contrast, European countries focus more on the practical applications of digital health services for older adults, with an emphasis on disabled older adults in communities or nursing homes [16-18]. A longitudinal study on the older population indicated that health interventions based on digital health technologies—such as virtual reality technology, wearable devices, telemedicine, artificial intelligence devices, and online health information seeking—have all demonstrated favorable effects in supporting home-based health management for older adults [19].

However, older adults often have limited digital health skills and low health literacy, and the digital divide continues to widen [20]. A scoping review on older adults' use of digital technologies for health promotion indicated [21] that digital health service technologies will become a primary means for older adults to achieve health goals. However, existing findings have not fully explained the specific application processes of these technologies, and more research is needed to understand the specific experiences of older adults with digital health services [22]. While the unique challenges and pressures faced by family caregivers have attracted considerable attention from researchers, the challenges and experiences they encounter when using digital health services remain poorly understood. Several existing studies suggest that there are specific differences in the perceptions of digital health services between older adults and their caregivers. For example, older adults have a relatively low overall use rate of "Internet+health care," but a high level of satisfaction with its use, with safety as their primary concern [23]. This indicates that older adults may attach greater importance to safety and privacy protection when choosing digital health services. In contrast, caregivers may focus more on how to effectively use these technologies to improve the quality and efficiency of care [24,25].

The accelerating pace of population aging exerts profound impacts on society, the economy, and public health systems. Addressing these challenges requires comprehensive strategies, including a deep understanding of the health needs of older adults and the support needs of their caregivers, to improve the older adults' care system and foster social integration and economic development. Therefore, this study aims to conduct an exploratory qualitative study in China to investigate the experiences with digital health services, willingness to use them, and influencing factors from the dual perspective of home-dwelling older adults and their family caregivers. The findings will identify their specific service needs and provide insights for the future design and optimization of digital health products.

Methods

Research Design

This study used a descriptive, qualitative design using in-person, semistructured interviews. Through this qualitative research, researchers aimed to understand the specific experiences of older adults and their caregivers with digital health services, identify their usage intentions, influencing factors, and service needs, and provide recommendations to advance the development of digital health services. Furthermore, we adhered to the Standards for Reporting Qualitative Research guidelines [26] to present the findings of this research (see Multimedia Appendix 1).

Recruitment

The study population included older adults with experience using digital health services and their caregivers. The researchers adopted a purposive sampling method to select research sites in Hefei City, Anhui Province, China, specifically including (1) seven community health service centers: Yaohai District (eastern Hefei), Shushan District (southwestern Hefei), and Baohe District (southern Hefei), which cover the geographical distribution of both old urban areas and emerging development zones and (2) the Geriatric Inpatient Ward of the First Affiliated Hospital of Anhui Medical University. This hospital is the largest public Grade A tertiary general teaching hospital in Anhui Province. The inclusion and exclusion criteria for older adults and caregivers were shown in Table 1. Eligible older adults and caregivers were sought, and participants were briefed on the study objectives and procedures. If they showed interest, they were invited to participate in the study interview. Data saturation was considered to have been achieved when

XSL•FO
RenderX

semistructured interviews with successive participants failed to generate any new open-ended codes, categories, or thematic insights on 3 consecutive occasions. The research team then collectively confirmed that no additional meaningful information would be derived from further interviews.

Table . Inclusion and exclusion criteria for the study population.

Study population	Inclusion criteria	Exclusion criteria
Older adults	• Aged ≥60 years • Resided in communities in Hefei for ≥6 months • Have experiences in using digital health services	• Comorbid with other severe diseases • Presence of communication disorders
Caregivers	• Aged ≥18 years • Family members of older adults who assumed primary responsibility for home-based care without receiving remuneration • Longest caregiving duration (≥3 months in total) • Have experiences in using digital health services • Clear consciousness, everyday speech, and the ability to complete the interview	• Diagnosis of severe organic diseases or mental illnesses • Hired nannies, nursing attendants, or other paid caregivers

Ethical Considerations

This study has obtained ethical approval from the Biomedical Ethics Committee of Anhui Medical University (Ethical approval number: 83242378). In accordance with the Declaration of Helsinki, the study adheres to the principles of informed consent and privacy protection, and all investigations were conducted with informed consent. All data are strictly confidential and shall be used exclusively for this study. Participants took part in this study voluntarily and did not receive any financial compensation.

Data Collection

Based on the research objectives, a semistructured interview guide was developed through an extensive review of relevant literature and discussions among research team members. Before the formal interview, we conducted a preinterview with 2 older adults and their caregivers and adjusted the interview questions based on the results. The final interview outline was shown in Multimedia Appendix 1. Prior to the formal interviews, we screened the research participants to ensure that they had experience in using digital health services. At the very beginning of the interview, the research participants were provided with descriptions of the 4 types of digital health services, and then they chose the services they had used in their lives, including health monitoring (Mo), health management (Ma), intelligent life assistance (I), and online health care services (O). Detailed descriptions of the digital health services refer to the outline of the interview in Multimedia Appendix 1. Interviews were conducted in Chinese by 2 researchers (JYR and CXG), who had no prior contact with the participants. Interview data were collected between December 2023 and February 2024 via in-person, semistructured, in-depth interviews, supplemented by field notes. Interviews were arranged at locations convenient to participants and in quiet, undisturbed settings to prevent interruptions. Before each interview, participants' permission to record audio was obtained.

Data Analysis

Conventional content analysis guided our analytic approach [27]. Audio recordings were transcribed verbatim immediately after each interview. The textual data were then imported into NVivo (12.0 software; QSR International Pty Ltd). To protect privacy, participants' real names were replaced with letters. Data analysis, coding, and theme extraction were performed using NVivo 12.0 software. Two researchers (co-first authors JYR and CXG), both of whom had training in qualitative methods, repeatedly read the interview materials (including audio transcripts and field notes) to capture the exact meanings participants intended. JYR repeatedly read the first 4 transcripts, marked all meaning units relevant to the phenomenon, and generated an initial list of codes. Subsequently, JYR and CXG collaborated to develop the codebook. JYR drafted names, definitions, and typical excerpt examples for each initial code, forming the preliminary codebook. The 2 researchers then jointly reviewed the first 4 transcripts, checking the fit between each code and its data source line by line, merging similar codes, refining definitions, and resolving any inconsistencies through discussion. Thereafter, the codebook was applied to subsequent transcripts. When new meanings emerged that could not be captured by existing codes, new codes were added and discussed again for revision. This iterative process continued until no new codes emerged and the codebook was considered saturated and stable.

The final determination of themes was completed by all authors (JYR, CZG, QQC, TY, YL, YXZ, and LFX). After all transcripts had been coded, JYR exported the final list of codes with their definitions and preliminarily clustered them into subcategories and candidate themes based on their shared latent meanings. The research team then held multiple discussions, engaging in repeated debate and revision of the naming, definitions, and boundaries of the themes until consensus was reached. Finally, each theme was assigned a formal name and clear definition,

XSL•FO
RenderX

and the most representative participant quotations were extracted as examples.

Fidelity Monitor

Before initiating the study, researchers received specialized training in interview techniques and conducted pilot interviews, refining the interview guide based on their outcomes. Before formal interviews, recording equipment was tested to ensure optimal functionality and audio quality, and written informed consent was obtained from all participants. Interviews with older adults and their caregivers were independently administered by 2 postgraduate nursing students, with no cross-discussion between the interviewers during data collection. All textual data were anonymized and coded; researchers maintained a strictly neutral stance throughout the analytical process to ensure objective interpretation, and study findings were subjected to rigorous review by the research team to uphold methodological objectivity. In cases of discrepancies in extracted themes, the research team engaged in collective deliberation to reach consensus on final thematic classifications, thereby ensuring the reliability and consistency of the study's results.

Results

Characteristics of Study Participants

The study included 18 older adults with a mean age of 70.17 (SD 6.21) years and 17 caregivers with a mean age of 50.65 (SD 15.37) years. Ten of these participants formed matched older adult and caregiver dyads. Additional demographic characteristics are presented in Multimedia Appendix 1.

Thematic Findings

The qualitative analysis began with 630 initial codes. Through an iterative process of comparison and consolidation of similar codes, these were refined into 470 distinct codes. These codes were then organized into a hierarchical structure, resulting in 12 subthemes and 4 main themes: (1) usage and preference of services, (2) specific service experiences, (3) influencing factors, and (4) suggestions and expectations. An example of the coding process and the complete coding framework are provided in Multimedia Appendix 1. The similarities and differences in perceptions of digital health services between older adults and caregivers were summarized in Table 2.

Table . Comparison of perspectives on digital health services between community-dwelling older adults and family caregivers.

Evaluation dimension	Older adults' perspective	Family caregivers' perspective	Similarities	Differences (O[a] and C[b])
Usage and preference of services	Focus on easy-to-operate, low-cost hardware and basic online services	Usage attached to older adults' needs; focus on auxiliary monitoring and comprehensive online medical services	Health monitoring and health management services	O: simple hardware and basic functions C: auxiliary operation and full online medical services
Experience	Positive: usefulness, ease of use, health promotion, psychological comfort Negative: low credibility, complex operation, physical discomfort, dignity concerns	Positive: same as older adults Negative: low credibility, complex operation	Value usefulness, ease of use, health promotion, psychological comfort; Face low credibility and complex operation	O: physical discomfort, dignity concerns
Influencing factors	Promoting: health needs, subjective norms, government support Obstructing: low literacy, weak economy, low acceptance of new things	Promoting: health needs, excellent user experience Obstructing: low literacy, cost concerns, low acceptance of new things, insecurity, limited time	Driven by health needs; blocked by low literacy, low acceptance of new things, cost concerns	O: affected by peers/experts/policy; government support C: excellent user experience, limited by time; focus on privacy
Suggestions and expectations	Age-friendly and personalized design; stricter supervision and review; on-site assistance	Age-friendly and personalized design; on-site assistance; family/peer support; authoritative platforms; more publicity	Suggest personalized and age-friendly design, on-site assistance	O: emphasize stricter supervision and review C: emphasize authoritative platforms, publicity, and family/peer support

[a]O: older adults.

[b]C: caregivers.

Theme 1: Usage and Preference of Services

Online learning enabled older adults to easily acquire health knowledge and skills, thereby enhancing their self-care capacities, while the online shopping function facilitated convenient purchases of health products and services to meet daily needs. Through continuous monitoring via relevant functions, they tracked changes in vital signs like heart rate, blood pressure, and blood sugar, gaining a comprehensive understanding of their health status to detect issues early and prevent complications. The health management feature allowed them to review daily data and obtain personalized advice, reducing hospital visits and enabling effective chronic disease management. Remote diagnosis streamlined health care access, provided tailored services, and allowed them to verify doctors' qualifications online. Community communication offered a

XSL•FO
RenderX

platform for interacting with peers and professionals, sharing experiences, and gaining additional health insights and advice (Textbox 1).

Textbox 1. Theme 1 and subthemes.

Theme 1: Usage and Preference of Services

Subtheme 1: Usage situation

Older adults tended to use health monitoring, health management, and intelligent life assistance services. In contrast, caregivers tended to use health monitoring, health management, and online health care services. A detailed overview of the digital health services used by the older adults and caregivers was presented in Table 3 and Multimedia Appendix 1.

Subtheme 2: Preferred functions

Participants reported five preferred functions of digital health services, as follows:

(1) Online learning or shopping

"I took an online traditional Chinese medicine course and really liked the synchronous learning feature, which let me study and practice alongside the video." [A12, male, 71 years, health management (Ma), online health care services (O)]

"Shopping on the platform is great, and the after-sales service is excellent. We place an order on the platform, and the courier delivers it directly to us." [A17, female, 72 years, health monitoring (Mo), Ma]

(2) Real-time health monitoring

"With a blood pressure monitor, I can check my blood pressure anytime, anywhere. Seeing a doctor requires going to the hospital, but at home, I can measure it 2 or 3 times a day." [A13, female, 64 years, Mo]

(3) Health management

"The blood pressure monitor and glucometer are accurate. If my blood pressure is high, it can also remind me to take my medication." [A5, male, 70 years, Mo]

"This platform also pushes some health information, like how hypertensive patients should control their diet and what they should pay attention to." [A17, female, 72 years, Mo, Ma]

(4) Remote diagnosis/treatment

"The 'window' [to view doctor profiles] is very important; you cannot trust just one doctor's opinion. Although the doctors' introductions on the official account page look similar, you can basically figure it out by looking at a few more... When making an online appointment, you can see the doctor's profile." [A11, female, 61 years, Ma, O]

(5) Community communication

"He [the staff] shared some precautions and instructions for using the features in the large group chat." [A17, female, 72 years, Mo, Ma]

"When I come across an article from this official account, I can sometimes share it with others." [A7, male, 82 years, Ma, intelligent life assistance (I)]

XSL•FO
RenderX

Table . Digital health services used by older adults and caregivers.

Group	Health monitoring (Mo[a])					Health management (Ma[b])					Intelligent life assistance (I[c])		Online health care services (O[d])		
	Electronic BP[e] monitor	Electronic Glucose meter	Smart body fat scale	Wearable activity tracker	Pulse oximeter	Treadmill	Online software and platforms	Health-related short videos	Online purchase of health care products	Professional health websites	Electric wheelchairs	Intelligent cochlear implants	Online appointment and registration	Online health consultations	Online medication purchase
Older adults	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓		
Caregivers	✓	✓	✓	✓			✓	✓	✓	✓			✓	✓	✓

[a]Mo: health monitoring.

[b]Ma: health management.

[c]I: intelligent life assistance.

[d]O: online health care services.

[e]BP: blood pressure.

Theme 2: Specific Service Experiences

Older adults and caregivers shared highly consistent positive experiences, as both groups acknowledged the value of digital health services in improving medical efficiency, simplifying health management, and alleviating health anxiety, findings that confirmed the adaptability of such services to older adults' care scenarios. Caregivers' positive experiences additionally included "reducing care burden," a benefit closely linked to their role as assistants. In contrast, older adults prioritized the psychological comfort and enhanced life autonomy afforded by the services. Their negative experiences exhibited both commonalities and significant group differences. Both groups encountered issues of low credibility of information and cumbersome operation of certain functions. Older adults, however, experienced unique challenges such as physical discomfort and psychological resistance due to physiological aging and dignity demands, while caregivers focused more on privacy, security risks, and potential costs of service use. Older adults' negative experiences were the key inducement for their shift from "unable to use" to "unwilling to use." In contrast, caregivers' negative feelings directly impacted their willingness to assist older adults in service use (Textbox 2).

XSL•FO
RenderX

Textbox 2. Theme 2 and subthemes.

Theme 2: Specific service experiences

Subtheme 1: Positive experience

Participants reported four key positive experiences with digital health services, as follows:

(1) Perceived usefulness

"With online appointments, you do not have to travel back and forth—that is the most significant convenience. I can book in advance with the specialist who knows my condition. It has many benefits." [A15, female, 60 years, health monitoring (Mo), health management (Ma), online health care services (O)]

"It is much more convenient. Since we live far away, it would be very troublesome to travel here and find no available appointment slots." [C10, male, 51 years, Mo]

(2) Perceived ease of use

"It is very simple. You just put your arm in and press a button. The old blood pressure monitors required removing and putting on clothes, which was inconvenient and made our arms sore." [A2, female, 71 years, Mo, O]

"For example, my grandfather can operate the blood pressure monitor on his own. It saves us both worry and effort, so we will definitely continue using it." [C2, female, 48 years, daughter of A3, Mo, O]

(3) Promote health management for the older adults

"I bought a treadmill after retiring, partly out of curiosity. It's convenient and novel—I can walk at home without going outside, saving me the need to run outdoors." [A6, male, 85 years, Ma]

"The health videos are great. Many are from hospital doctors and experts who regularly explain and analyze various diseases. Both the older adults and I can learn a lot from them." [C2, female, 48 years, daughter of A3, Mo, O]

(4) Recover psychological comfort

"When you go for a walk wearing it [the fitness tracker], you can at least see how many steps you have taken and how far you have gone. It also tells the time. I can check my heart rate when I feel palpitations and my blood oxygen if I feel dizzy. It is definitely helpful, at least I feel less panicked." [A15, female, 60 years, Mo, Ma, O]

"When older adults feel unwell, they might not tell you, fearing they are a burden, but they worry privately. These health services or short videos can serve as intermediaries, helping with their psychological adjustment." [C17, female, 33 years, daughter of A11, Mo, Ma]

Subtheme 2: Negative experience

(1) Low credibility

"At our age, we care about health, but some online health information is reliable, and some is not... For example, ginger tea, brown sugar, and red dates warm the stomach in winter, but some public accounts claim they have anti-aging effects. That's just exaggerated." [A15, female, 60 years, Mo, Ma, O]

"I browse on my phone. If something does not seem real, I stop reading and ignore it." [C4, female, 76 years, spouse of A7, Mo]

(2) Complex operation

"For instance, when a video talks about uric acid, it stops at the key point and requires me to follow a public account to learn more. It's too complicated." [A1, male, 70 years, Mo, intelligent life assistance (I)]

"The glucometer is also relatively cumbersome to use. It involves a blood draw, so I always help them with the measurement." [C11, female, 22 years, granddaughter of A4, Ma, O]

(3) Increase physical and mental burden

"We aren't like young people who like wearing earphones. I have ear problems, but when I wear the [cochlear implant], I feel dizzy and my head feels stuffy if I wear both, so I only wear one." [A7, male, 82 years, Ma, I]

"I have opened those little apps on the phone, but the interfaces pop up everywhere, and you need to click different places. I do not understand it; it is very annoying, so I close it." [A16, female, 70 years, Mo]

"When my condition flares up, one leg is immobile. The doctor told me to use an electric wheelchair and have a family member push me. However, I think since I can walk with one leg, I still want to walk by myself. Otherwise, I feel I have no dignity." [A1, male, 70 years, Mo, I]

Theme 3: Influencing Factors

The facilitating factors for the use of digital health services were jointly driven by older adults' internal health needs and external environmental support. Chronic diseases and declining physical function among older adults served as the core internal motivations, while high-quality user experience, subjective norms within social circles, and government policy support were the key external driving forces. Subjective norms exerted a particularly significant impact on older adults, reflecting their susceptibility to guidance from authoritative entities and social circles when accepting new things, whereas caregivers' willingness to use such services was mostly determined by whether the services could reduce their care burden. The core

XSL•FO
RenderX

obstacles differed between the two groups. Older adults faced low digital health literacy, a weak economic foundation, and low acceptance of new things, while caregivers encountered issues such as insufficient time, privacy security concerns, and limited personal digital capabilities. Both groups were affected by inadequate age-friendly service designs; economic costs and service security were common concerns, acting as core external obstacles to the popularization of digital health services, while caregivers' time constraints were an easily overlooked important influencing factor (Textbox 3).

Textbox 3. Theme 3 and subtheme.

Theme 3: Influencing Factors

Subtheme 1: Promoting factors

Participants reported four driving factors affecting usage, as follows:

(1) Health needs

"I have diabetes and hypertension myself, so I need to keep track of my blood sugar and blood pressure levels." [A4, male, 70 years, health monitoring (Mo)]

"My hearing is very poor, which makes communication difficult. He [the older adult] can't hear clearly at all when others speak." [C4, female, 76 years, spouse of A7, Mo]

(2) Excellent user experience

"It is definitely the effectiveness [of the online health platform]. It has practical benefits for the older adult and is helpful for me, too." [C7, female, 52 years, daughter of A17, health management (Ma), intelligent life assistance (I)]

"The advantage is that it's simple and easy to operate. We will certainly continue using it." [C2, female, 48 years, daughter of A3, Mo, online health care services (O)]

(3) Subjective norms

"My social circle shares information about nutrition and health, and there's a sense that we older adults should keep up with the times and learn to use some smart health devices." [A9, male, 65 years, Ma]

"When browsing Douyin or WeChat, I come across health-related videos, such as those from wellness experts or doctors doing science popularization." [A7, male, 82 years, Ma, I]

(4) Government support

"The government is now strongly promoting the 'Big Health' initiative. Following government policy is undoubtedly the right thing to do, and it is also suitable for us older people." [A17, female, 72 years, Mo, Ma]

Subtheme 2: Obstructive factors

(1) Low level of digital health literacy

"We don't have a medical background, so it's hard to know what's true or false. We also don't understand how to use high-tech products." [A12, male, 71 years, Ma, O]

"Smartphone technology is advancing too fast. It is tough for older minds to adapt. Our [referring to caregivers] education and cultural level have not reached that high standard either." [C16, male, 47 years, son of A10, Mo]

(2) Weak economic foundation

"I'm from the countryside. My children have heavy burdens; sometimes they can't even take care of themselves." [A10, female, 70 years, Mo]

"If we were to continue using it, the first thing [we consider] would definitely be the price." [C14, male, 50 years, son of A8, Mo, O]

(3) Lack of awareness of some digital health services

"I personally believe that seeing a doctor still requires going to the hospital. We only feel reassured after hearing a professional doctor's advice." [A6, male, 85 years, Ma]

"Without actually using this thing [digital health device], you don't know where its benefits lie." [C14, male, 50 years, son of A8, Mo, O]

(4) Low service security

"With online consultations, we always feel somewhat distrustful of the platform. We're afraid that leaving personal information could lead to fraud." [C2, female, 48 years, daughter of A3, Mo, O]

(5) Little free time

"I am quite busy with work now and have no time to watch [health videos, etc.]. I spend my evenings accompanying the two older adults, especially my dad, so sometimes I don't have much time to deal with other things [like exploring other digital health services]." [C7, female, 52 years, daughter of A17, Ma, I]

XSL•FO
RenderX

Theme 4: Suggestions and Expectations

Participants advocated for simplified operation procedures, age-adapted interface designs, and personalized service configurations to enhance accessibility and meet older adults' diverse needs, while older adults emphasized the necessity of strengthened supervision and review mechanisms to protect data privacy, property security, and standardized advertising content. To address technical and operational barriers, older adults expected regular on-site assistance and professional in-home guidance, whereas caregivers highlighted the significance of community support and professional services. Caregivers acknowledged the positive role of family members, peers, and social circles in facilitating older adults' acquisition of knowledge and acceptance of digital health services through technical and emotional support. They also proposed establishing official hospital-operated health accounts with physicians as core health information disseminators, whose institutional endorsement could significantly improve information trustworthiness. Additionally, caregivers stressed that insufficient publicity and low public recognition, particularly among older adults, hindered service acceptance, and multichannel promotion was deemed effective to expand coverage, raise awareness, and build public trust (Textbox 4).

Textbox 4. Theme 4 and subthemes.

Theme 4: Suggestions and Expectations

Subtheme 1: Improve the age-friendly and personalized design of the services

"It's worth continuing to promote these services. I suggest adding more options in form to cater to the personalized needs of older adults." [A4, male, 70 years, health monitoring (Mo)]

"Text is definitely better for us because our reactions are slow, and we can ponder the text. Audio finishes playing and is easily forgotten." [A11, female, 61 years, health management (Ma), online health care services (O)]

"The online part should be made as simple as possible, so it's easier for older people like us, or even older ones, to operate. Using audio instead of text would be very convenient for illiterate seniors." [C2, female, 48 years, daughter of A3, Mo, O]

Subtheme 2: Strengthen the intensity of supervision and review

"(Some advertisers) use celebrity endorsements to deceive people. If we believe them, we pay directly by scanning a code on the platform, and end up losing our money. This needs proper regulation." [A18, female, 66 years, Mo]

Subtheme 3: Offer on-site assistance

"It would be great if there were professionals, for example, coming to our home once a month to calibrate the devices." [A18, female, 66 years, Mo]

"It would be better if the community had professionals. If experts (like family doctors) could provide in-home guidance or regularly help us check things, they [the older adults] would certainly find it convenient and wouldn't need to go out." [C8, female, 42 years, daughter of A18, Ma]

Subtheme 4: Encourage digital feedback and peer support

"If an older adult doesn't live with their children, exposure should start within their own social circle. If they live with their children, the children should take the lead in introducing the services. Guidance and information passed on by children are certainly more readily accepted by older adults." [C1, male, 37 years, son of A2, Mo, intelligent life assistance (I)]

Subtheme 5: Establish an authoritative platform

"Now I see many videos are made by doctors' personal private accounts. Could a dedicated hospital account be created, where doctors from various departments regularly explain medical knowledge about different diseases? Using the hospital as the main body would be better—it's more comprehensive and, secondly, more credible. [C2, female, 48 years, daughter of A3, Mo, O]

Subtheme 6: Increase the level of publicity

"Many people don't trust online health product platforms; my mother didn't believe in it at first. Therefore, it's necessary to increase publicity efforts and enhance credibility among older adults." [C7, Female, 52 years, daughter of A17, Ma, I]

"It's definitely worthwhile. This thing [digital, health services] must be popularized." [C1, male, 37 years, son of A2, Mo, I]

Discussion

Principal Findings

This study identified four key findings regarding digital health service experiences among home-dwelling older adults and their family caregivers. First, both groups primarily used health monitoring and basic health management functions but showed distinct preferences for service types. Second, they shared positive experiences of improved convenience and health support, while older adults uniquely reported dignity-related distress and physical discomfort, and caregivers emphasized privacy and safety concerns. Third, usage was driven by health needs and social influences but blocked by low digital literacy, economic constraints, and time shortages. Fourth, both groups prioritized age-friendly design, authoritative platforms, on-site guidance, and family or peer support to optimize future services.

Usage and Preference of Services

In general, the types of services frequently used by older adults include health monitoring, health management, artificial intelligence–powered living assistance, and Internet-based medical services, which echo the categories of customer-system interaction, decision support, and service delivery in the World Health Organization (WHO) classification of digital health interventions [28]. In addition, regarding health monitoring and

XSL•FO
RenderX

management, older adults showed high acceptance and demand. Older adults tend to use functions that support their autonomy in daily life and show a strong preference for low-cost or free services, such as using an electronic blood pressure monitor and an electronic glucose meter, and accessing online health information. Although digital technology has shown great potential for improving health services for older adults [29], the penetration rate of digital technology in transitional and low-income countries is low, older adults' purchasing power is limited, and their health literacy is low. Moreover, older adults in rural and remote areas often face challenges, including inadequate infrastructure and limited access to technology [30]. Therefore, it is reluctant to try services with high technical access requirements, complex operational processes, and high use costs, such as hip protective airbags, smart homes, and body sensors.

Service Experience

Common Experience

Both older adults and their caregivers can benefit from digital health services, but they also have the same negative experiences. First, both recognized the value of digital health services in improving the efficiency of medical treatment, simplifying health management, and providing psychological comfort. At the same time, it faces the dual challenges of low service reliability and high operational complexity. At present, the promotion of digital health services mainly relies on traditional media, online media, and medical institutions [31]. The COVID-19 pandemic fast-tracked the digital transformation of health care [32]. Some digital health service tools can meet people's daily health management needs, improve their understanding of their own health status, thereby relieving the tension and anxiety caused by uncertainty about their bodies, reduce unnecessary medical visits, and reduce the burden on the medical system [19]. However, some older adults who are not familiar with modern technology products often feel strange and uneasy when using smartphones, tablets, and other devices. Complex operations and a large volume of health information that is difficult to distinguish between authenticity and inaccuracy will cause older adults to experience technical anxiety and worry about economic losses from accidental contact [33]. Family caregivers play an important role in the lives of older adults. They not only provide daily care but also need to assist them in collecting health information and using communication technology to ensure the quality of life of older adults [34]. Therefore, caregivers' experience of use is also significantly influenced by older users' experience, and they tend to be more concerned about the risk of privacy leakage among older people. This may be because they expect that digital health services can adapt to the abilities and needs of older adults, save physical and time costs, and reduce their own care burden [35].

Differentiated Experience

In addition, there was a significant difference in the focus of experience between older adult users and caregivers. For many older adults, the Internet breaks down the limitations of physical space, allowing them to participate in social interactions and effectively relieving loneliness and enhancing social belonging [36]. By mastering new technology, older people often gain a sense of control. This increased ability translates directly into a more positive mindset [37]. In contrast, family caregivers affirmed the advantages of digital services in convenience, information access, and remote monitoring. They regarded it as a tool to reduce the burden of care and optimize the efficiency of health management. At the same time, the experience of older people is deeply affected by their physical decline, insufficient digital literacy, and psychological feelings. The impaired vision, hearing, and learning ability make them prone to frustration in the face of pop-up windows and multistep operations, which damages their autonomy and dignity. This is consistent with the study of Moonika et al [38]; when older adults use digital health services, they will feel the loss of control and the increased sense of dependence on others, and this sense of dependence will damage their personal dignity, directly causing some older adults to develop from "cannot use" to "unwilling to use" and actively avoid digital health services. It should be noted that compared with other caregivers, older adults' spousal caregivers have poor health status [39], lack necessary digital health knowledge and digital skills and ability to solve mental disorders of the older adults [40], and may not be able to effectively help the older adults access and use digital health services, thus limiting the effective use of digital health technology by the older adults.

Therefore, future digital health services must adopt a differentiated and collaborative strategy to effectively integrate into the lives of home-dwelling older adults and their caregivers. On the one hand, it is necessary to systematically lower the usage threshold and reduce the trust cost for older adults by simplifying the interface, making the operation process user-friendly, and ensuring the authority of information sources. On the other hand, digital skills support training should be provided to different types of caregivers to varying degrees, leveraging the family collaboration function and enhancing the collective force of family health management through technological convenience.

Influencing Factors

The continuous innovation and widespread adoption of digital health technologies have led both older adults and their caregivers to recognize the advantages of digital health services. Research indicated that the fulfillment of health needs and positive user experiences facilitate the willingness of older adults and caregivers to adopt digital health services, which is consistent with the Technology Acceptance Model [41], in which perceived usefulness and perceived ease of use are key influencing factors. In addition, subjective norms, such as encouragement from family members, peers, and health experts and the sharing of positive usage experiences, serve as important references for older adults. This finding aligns with the Technology Acceptance and Use Unified Theory (UTAUT), which emphasizes social influence as a critical factor in technology adoption [42]. Furthermore, older adults demonstrated positive attitudes toward government-supported digital health services [43], which is consistent with the concept of facilitating conditions in UTAUT, reflecting that policy-level promotion can enhance older adults' awareness and acceptance of digital health services.

XSL•FO
RenderX

Nevertheless, they continue to face multiple barriers in actual use. Older adults generally have low digital health literacy, making it difficult for them to adapt to rapidly evolving internet-based health technologies [22]. The study found that they exhibited avoidance and rejection attitudes toward some digital health services and were more inclined to trust the diagnoses, treatment opinions, and effects of offline medical services. This preference reflects a tendency to adhere to familiar practices even when new alternatives are available. In addition, the financial burden associated with purchasing digital devices or accessing paid digital health services remains a salient challenge for both low-income older adults and caregivers [34].

Caregivers also face barriers related to their roles. Lack of time is a major constraint, particularly for younger adult caregivers, who struggle to balance work, caregiving responsibilities, and personal life, leaving them little time to assist older adults with digital health resources [44]. Although these caregivers recognize the potential benefits of such services, they find it difficult to actively use them, creating a gap between intention and behavior. Older spousal caregivers encounter technical difficulties similar to those faced by older adults. In addition, caregivers expressed heightened concerns regarding safety and privacy, particularly regarding the security of online platforms and the risk of misuse of personal information. They generally perceive older adults as being in a passive position in the digital society and worry about the safety of older adults using digital health services independently, tending to restrict their use [45].

The Technology Acceptance Model and UTAUT models emphasized rational cognitive factors and ignored the influence of additional factors on users' behavioral intentions, such as time, cost, safety, emotions, and whether they possess the necessary skills to use digital services. Therefore, these classic models were insufficient to explain older adults' adoption behaviors of digital health services. Some studies [46,47] had attempted to extend these models by adding new variables or integrating with other theoretical frameworks to make them more applicable to older populations in digital health. Future research should extend these classic models by integrating insights from qualitative findings, constructing a theoretical framework that is more tailored to older adults for digital health services.

Suggestions and Expectations

Based on feedback from older adults and their caregivers, this study proposes strategies to optimize the adoption of digital health services from 3 dimensions, namely government, service providers, and social support. The government should rely on grassroots institutions such as communities to establish an integrated digital health literacy education, training, and promotion system that covers the continuum from home to community and from offline to online [48], enhancing older adults' comprehension of and experience with digital health technologies, thereby facilitating the widespread adoption of digital health services among this population. Additionally, it is recommended that economic incentives be implemented to reduce usage costs—for example, providing subsidies for digital devices and including relevant expenses in medical insurance reimbursement—to alleviate financial barriers for older adults and low-income groups. For instance, in Nanjing, China, the government provides digital devices, such as smart bracelets, to older adults as a welfare benefit [49,50]. Older adults can access services either through automatic feedback from the devices or through active calls, improving service accessibility and acceptability. Simultaneously, the government should strengthen platform supervision, improve mechanisms for digital privacy protection, and establish authoritative information communication channels jointly led by multiple formal medical institutions to enhance public trust in and adoption of such services. Digital health service providers emphasize the need for age-appropriate, personalized design centered on older adults' needs, with a focus on simplifying operational processes, optimizing user interface interactions, and accounting for age-related physical and cognitive characteristics [51]. Design should be grounded in older users' actual usage habits to lower technological barriers. Previous research indicates that interface adaptability and user-centered design can significantly enhance older adults' engagement [52]. For example, older adults have suggested that platforms should offer multiple input and output methods (eg, text and audio) to meet the needs of users with varying educational levels and processing speeds. In terms of after-sales support, it is recommended to establish user profiles, conduct regular follow-ups on usage experiences, and set up after-sales service stations in communities, pharmacies, and other locations to provide continuous support. Additionally, service providers should formulate reasonable pricing in line with government guidance, offer discounts or medical insurance support for specific older adult groups, and balance social responsibility with economic benefits. Social support plays a crucial role in helping older adults improve their digital health literacy. "Digital reciprocity"—where adult child caregivers support older adults in learning and using digital health services—is recognized as the most direct and practical approach. Face-to-face guidance from adult child caregivers can encourage older adults to adopt an open attitude toward digital technology, gradually guiding them to accept and actively learn health knowledge, thereby enabling older adults to receive both technical support and positive emotional incentives [53]. If adult child caregivers are unable to assist due to time constraints or if older adult spousal caregivers themselves face barriers to using digital technology, social forces such as peer groups, community organizations, or young volunteers can be leveraged to establish a "community school" model, forming an extra-familial reciprocity structure [41]. This mechanism can help older adults overcome technology anxiety, better perceive the value of digital health services, and enhance their intention to use and self-efficacy.

Limitations and Future Directions

This study had several limitations. First, regarding the sample selection, the study was confined to Hefei, China, which may limit the generalizability of the findings to regions with different health care and digital infrastructure contexts. Second, to respect participants' willingness and to capture diverse attitudes and experiences, older adults and their caregivers were not fully paired: only 10 dyads were matched, while the remaining participants were recruited independently. This partial pairing reduced the strength of dyadic-level interpretations, although

XSL•FO
RenderX

separate analyses still provided valuable insights into each group's experiences and needs. Third, all participants had prior experience with digital health services, which may have introduced a positive bias toward perceived benefits and facilitators. Future research should adopt large-scale, multicenter sampling strategies, prioritize the recruitment of fully paired older adult and caregiver dyads, and include participants with varying levels of digital health experience to enhance representativeness.

Beyond sample-related improvements, future studies should also investigate the heterogeneity of digital health services themselves. Specifically, within the same population, older adults' experiences, acceptance, and barriers may differ significantly across various types of digital health services. Exploring these intraindividual differences would help identify which service characteristics are most conducive to adoption and sustained use among older adults. Such insights are critical for tailoring service design and implementation strategies to diverse user needs and contexts.

Conclusion

This study analyzed the usage of digital health services from the dual perspectives of older adults and their caregivers. Although all participants could benefit from digital health services, their willingness to use digital health technologies remained low due to various barriers, including low digital health literacy, weak economic foundations, and inadequate service security. Therefore, it is suggested that future promotion and optimization of digital health services be advanced collaboratively across three dimensions: government guidance, service provider optimization, and the construction of social support systems. This integrated approach aims to improve the user experience for older adults and caregivers, facilitate sustained adoption of these services, and ultimately enable older adults to benefit from digital health technologies.

Acknowledgments

The authors thank the study participants for contributing their time and feedback. The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing. The GAI tool used was DeepSeek-V3.2. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. JiaYi Ruan, CuiXia Guan, and QingQing Chen are co-first authors.

Funding

This work was supported by Key Laboratory of Geriatric Long-term Care (Naval Medical University), Ministry of Education (LNYBPY-2023-05), Anhui Medical University School of Nursing 2025 Graduate Student Young Talent Cultivation Project (Hlqm12025025 & Hlqm12025002) and Provincial Innovation and Entrepreneurship Training Program for College Students (S202510366077). The funders had no involvement in the study design, data collection, data analysis, interpretation of results, or the writing of the manuscript.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to institutional data protection policies and participant confidentiality regulations. The data may be made available from the corresponding author upon reasonable request by qualified researchers for academic purposes.

Authors' Contributions

Conceptualization: LFX
Data curation: JYR, CXG
Formal analysis: QQC
Funding acquisition: LFX, JYR, CXG
Investigation: JYR, CXG
Methodology: LFX
Project administration: TY
Resources: YXZ
Software: YL
Supervision: LFX
Validation: QQC
Visualization: TY
Writing – original draft: LFX, JYR, QQC
Writing – review and editing: LFX, JYR, QQC, LFX

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Appendix reporting standards, interview guides, participant demographics, and coding framework.
[DOCX File, 58 KB - jmir_v28i1e89496_app1.docx]

References

1. World population prospects 2022: summary of results population division. United Nations. URL: https://www.un.org/development/desa/pd/content/World-Population-Prospects-2022 [accessed 2026-05-20]
2. Wang MY, Sung HC, Liu JY. Population aging and its impact on human wellbeing in China. Front Public Health 2022;10:883566. [doi: 10.3389/fpubh.2022.883566] [Medline: 35419339]
3. Selakovic V, Ilic S, Radovanovic J, et al. Population aging as a global demographic phenomenon and its implications for public health. Exp Appl Biomed Res 2025;26(3):223-231. [doi: 10.2478/eabr-2025-0014]
4. Wang T, Dong Q, Sun Q, et al. Trajectories and influencing factors of online health information–seeking behaviors among community-dwelling older adults: longitudinal mixed methods study. J Med Internet Res 2025 Nov 5;27:e77549. [doi: 10.2196/77549] [Medline: 41191927]
5. Lanxin W, Yan Z, Yutong T, Lixue M, Li L, Ting Z. Potential profiling of self-management skills in older co-morbid patients. BMC Geriatr 2024 Jun 25;24(1):555. [doi: 10.1186/s12877-024-05137-4] [Medline: 38918703]
6. Cui W. Research progress on treatment burden and intervention measures for elderly chronic disease comorbidity. Theor Nat Sci 2024;58(1):149-154. [doi: 10.54254/2753-8818/58/20241364]
7. Eriksson M, Blomberg K, Arvidsson E, et al. Did the organization of primary care practices during the COVID-19 pandemic influence quality and safety? - An international survey. BMC Health Serv Res 2024 Jun 14;24(1):737. [doi: 10.1186/s12913-024-11173-y] [Medline: 38877434]
8. Waniger A, Gale K, DeNomie M, Nelson D. Community and caregiver perceptions of giving care to seniors. Int J Psychiatry Med 2019 Sep;54(4-5):307-315. [doi: 10.1177/0091217419860726] [Medline: 31269836]
9. López-Martínez C, Orgeta V, Gutiérrez-Sánchez B, Del-Pino-Casado R. Caregiving satisfaction, psychological distress and caregiver burden in family caregivers of dependent older people: a longitudinal study. J Clin Nurs 2025 Oct;34(10):4177-4185. [doi: 10.1111/jocn.17626] [Medline: 39789812]
10. Mao D, Li G, Liang M, Wang S, Ren X. Dietary patterns and multiple chronic diseases in older adults. Nutr Metab (Lond) 2024 Jun 24;21(1):36. [doi: 10.1186/s12986-024-00814-y] [Medline: 38915027]
11. Mitsutake S, Takahashi Y, Otsuki A, et al. Chronic diseases and sociodemographic characteristics associated with online health information seeking and using social networking sites: nationally representative cross-sectional survey in Japan. J Med Internet Res 2022;25:e44741. [doi: 10.2196/44741]
12. Ferretti A, Adjei KK, Ali J, et al. Digital tools for youth health promotion: principles, policies and practices in sub-Saharan Africa. Health Promot Int 2024 Apr 1;39(2):daae030. [doi: 10.1093/heapro/daae030] [Medline: 38558241]
13. Wongiam N, Praditpornsilpa K, Vacharaksa A. Comprehensive geriatric assessment for oral care in older adults: a focus group study. BMC Geriatr 2025 Apr 8;25(1):232. [doi: 10.1186/s12877-025-05854-4] [Medline: 40200164]
14. Tu J, Shen M, Zhong J, Yuan G, Chen M. The perceptions and experiences of mobile health technology by older people in Guangzhou, China: a qualitative study. Front Public Health 2021 Jun 25;9:683712. [doi: 10.3389/fpubh.2021.683712B3] [Medline: 34249848]
15. Chen C, Ding S, Wang J. Digital health for aging populations. Nat Med 2023 Jul;29(7):1623-1630. [doi: 10.1038/s41591-023-02391-8] [Medline: 37464029]
16. Lin CC, Lin YS, Yeh CH, Huang CC, Kuo LC, Su FC. An exergame-integrated IoT-based ergometer system delivers personalized training programs for older adults and enhances physical fitness: a pilot randomized controlled trial. Gerontology 2023;69(6):768-782. [doi: 10.1159/000526951] [Medline: 36617413]
17. Choi J, Lee S, Kim S, Kim D, Kim H. Depressed mood prediction of elderly people with a wearable band. Sensors (Basel) 2022;22(11):4174. [doi: 10.3390/s22114174]
18. Zhang Z, Giordani B, Margulis A, Chen W. Efficacy and acceptability of using wearable activity trackers in older adults living in retirement communities: a mixed method study. BMC Geriatr 2022 Mar 21;22(1):231. [doi: 10.1186/s12877-022-02931-w] [Medline: 35313819]
19. Piccolo R. Digitisation, health literacy and health empowerment of older people. Int Entrep Manag J 2024 Dec;20(4):3001-3034. [doi: 10.1007/s11365-024-00980-8]
20. Aslan A, Mold F, van Marwijk H, Armes J. What are the determinants of older people adopting communicative e-health services: a meta-ethnography. BMC Health Serv Res 2024 Jan 11;24(1):60. [doi: 10.1186/s12913-023-10372-3] [Medline: 38212713]
21. Yang Y, Cao M, Cheng L, Zhai K, Zhao X, De Vos J. Exploring the relationship between the COVID-19 pandemic and changes in travel behaviour: a qualitative study. Transp Res Interdiscip Perspect 2021 Sep;11:100450. [doi: 10.1016/j.trip.2021.100450] [Medline: 34568810]
22. De Santis KK, Mergenthal L, Christianson L, Busskamp A, Vonstein C, Zeeb H. Digital technologies for health promotion and disease prevention in older people: scoping review. J Med Internet Res 2023 Mar 23;25(3):e43542. [doi: 10.2196/43542] [Medline: 36951896]

XSL•FO
RenderX

23. Charenkova J. “Parenting my parents”: perspectives of adult children on assuming and remaining in the caregiver’s role. Front Public Health 2023;11:1059006. [doi: 10.3389/fpubh.2023.1059006] [Medline: 36875393]
24. Klein B, Shandley K, McLaren S, Clinnick L. Older adults’ preferences for Internet-based services: type and content. Australas J Ageing 2023 Dec;42(4):780-785. [doi: 10.1111/ajag.13201] [Medline: 37036842]
25. Farnood A, Johnston B, Evans C. The role of digital health interventions in supporting family caregivers for people with dementia to improve quality of life. Curr Opin Support Palliat Care 2024 Dec 1;18(4):224-229. [doi: 10.1097/SPC.0000000000000734] [Medline: 39373310]
26. O’Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
27. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
28. Classification of digital health interventions v10. World Health Organization. URL: https://www.who.int/publications/i/item/WHO-RHR-18.06 [accessed 2026-05-20]
29. Ma B, Yang J, Wong FKY, et al. Artificial intelligence in elderly healthcare: a scoping review. Ageing Res Rev 2023 Jan;83:101808. [doi: 10.1016/j.arr.2022.101808] [Medline: 36427766]
30. Fletcher CME, Stewart L, Gunn KM. Stressors, barriers and facilitators faced by Australian farmers when transitioning to retirement: a scoping review. Int J Environ Res Public Health 2023 Jan 31;20(3):2588. [doi: 10.3390/ijerph20032588] [Medline: 36767977]
31. Domingos J, Dean J, Fernandes JB, Família C, Fernandes S, Godinho C. Exploring barriers and educational needs in implementing dual-task training for Parkinson’s disease: insights from professionals. Front Med (Lausanne) 2024;11:1325978. [doi: 10.3389/fmed.2024.1325978] [Medline: 38646555]
32. Hvalič-Touzery S, Laznik J, Petrovčič A. “I’m still struggling with it, and it scares me”: a qualitative analysis of older adults’ experiences with digital health portals during and after COVID-19. Digit Health 2024;10:20552076241282247. [doi: 10.1177/20552076241282247] [Medline: 39372805]
33. Kim HN, Freddolino PP, Greenhow C. Older adults’ technology anxiety as a barrier to digital inclusion: a scoping review. Educ Gerontol 2023 Dec 2;49(12):1021-1038. [doi: 10.1080/03601277.2023.2202080]
34. Hwang YS, Jung SM, Dronina Y, Jo HS. Factors associated with caregivers’ preference for digital technologies to support older adults. J Gerontol Nurs 2022 Aug;48(8):33-41. [doi: 10.3928/00989134-20220630-02] [Medline: 35914077]
35. Lumini MJ, França D, Sousa MR, et al. Efficacy of digital technology-based interventions for reducing caregiver burden and stress: a systematic review and meta-analysis. Front Digit Health 2025;7:1636084. [doi: 10.3389/fdgth.2025.1636084] [Medline: 41281293]
36. Ausserhofer D, Piccoliori G, Engl A, et al. Community-dwelling older adults’ readiness for adopting digital health technologies: cross-sectional survey study. JMIR Form Res 2023;8:e54120. [doi: 10.2196/54120]
37. Laukka E, Lakoma S, Harjumaa M, et al. Older adults’ preferences in the utilization of digital health and social services: a qualitative analysis of responses to open-ended questions. BMC Health Serv Res 2024 Oct 4;24(1):1184. [doi: 10.1186/s12913-024-11564-1] [Medline: 39367429]
38. Raja M, Uhrenfeldt L, Galvin KT, Kymre IG. Older adults` sense of dignity in digitally led healthcare. Nurs Ethics 2022 Sep;29(6):1518-1529. [doi: 10.1177/09697330221095140] [Medline: 35724431]
39. Queen TL, Butner J, Berg CA, Smith J. Activity engagement among older adult spousal caregivers. In: Carr D, editor. J Gerontol B Psychol Sci Soc Sci 2019 Sep 15;74(7):1278-1282. [doi: 10.1093/geronb/gbx106] [Medline: 28958018]
40. Ramos MD. Aging immigrant family caregivers health, social engagement, and health literacy. Innov Aging 2021 Dec 17;5(Supplement_1):795-795. [doi: 10.1093/geroni/igab046.2931]
41. Zhu X, Cheng X. Staying connected: smartphone acceptance and use level differences of older adults in China. Univers Access Inf Soc 2022 Oct 13:1-10. [doi: 10.1007/s10209-022-00933-4] [Medline: 36254136]
42. Wilson J, Heinsch M, Betts D, Booth D, Kay-Lambkin F. Barriers and facilitators to the use of e-health by older adults: a scoping review. BMC Public Health 2021 Aug 17;21(1):1556. [doi: 10.1186/s12889-021-11623-w] [Medline: 34399716]
43. Goodarzi F, Barati M, Bashirian S, Ayubi E, Rahbar S, Cheraghi P. The experiences of the elderly regarding the use of rehabilitation assistive technologies: a directed qualitative content analysis. Disabil Rehabil Assist Technol 2024 Nov;19(8):2857-2868. [doi: 10.1080/17483107.2024.2313081] [Medline: 38329221]
44. Hopwood J, Walker N, McDonagh L, et al. Internet-based interventions aimed at supporting family caregivers of people with dementia: systematic review. J Med Internet Res 2018 Jun 12;20(6):e216. [doi: 10.2196/jmir.9548] [Medline: 29895512]
45. Sun E, Ye X. Older and fearing new technologies? The relationship between older adults’ technophobia and subjective age. Aging Ment Health 2024 Apr;28(4):569-576. [doi: 10.1080/13607863.2023.2241017] [Medline: 37553781]
46. Chen Y, Yuan J, Li C, et al. Innovative adoption model for digital health technologies among elderly with chronic diseases: integrating Unified Theory of Acceptance and Use of Technology and Knowledge-Attitude-Practice model in a survey of 1222 patients in Shanghai. BMJ Open 2026 Mar 9;16(3):e105529. [doi: 10.1136/bmjopen-2025-105529] [Medline: 41802793]
47. Anisha SA, Sen A, Ahmad B, Bain C. Exploring acceptance of digital health technologies for managing non-communicable diseases among older adults: a systematic scoping review. J Med Syst 2025 Mar 11;49(1):35. [doi: 10.1007/s10916-025-02166-3] [Medline: 40067482]

XSL•FO
RenderX

48. Dong Q, Liu T, Liu R, Yang H, Liu C. Effectiveness of digital health literacy interventions in older adults: single-arm meta-analysis. J Med Internet Res 2023 Jun 28;25:e48166. [doi: 10.2196/48166] [Medline: 37379077]
49. Bults M, van Leersum CM, Olthuis TJJ, Bekhuis REM, den Ouden MEM. Barriers and drivers regarding the use of mobile health apps among patients with type 2 diabetes mellitus in the Netherlands: explanatory sequential design study. JMIR Diabetes 2022 Jan 27;7(1):e31451. [doi: 10.2196/31451] [Medline: 35084357]
50. More than 10,000 smart bracelets! elderly people living alone in jiangsu get welfare!. Toutiao.com. 2025. URL: https://www.toutiao.com/article/7236382665803350568/ [accessed 2025-12-11]
51. Kebede AS, Ozolins LL, Holst H, Galvin K. Digital engagement of older adults: scoping review. J Med Internet Res 2022 Dec 7;24(12):e40192. [doi: 10.2196/40192] [Medline: 36477006]
52. Stoumpos AI, Kitsios F, Talias MA. Digital transformation in healthcare: technology acceptance and its applications. Int J Environ Res Public Health 2023 Feb 15;20(4):3407. [doi: 10.3390/ijerph20043407] [Medline: 36834105]
53. Portz JD, Fruhauf C, Bull S, et al. “Call a Teenager… That’s What I Do!” - grandchildren help older adults use new technologies: qualitative study. JMIR Aging 2019 Jun 6;2(1):e13713. [doi: 10.2196/13713] [Medline: 31518274]

Abbreviations

DESA: Department of Economic and Social Affairs
UTAUT: Unified Theory of Acceptance and Use of Technology
WHO: World Health Organization

Edited by A Stone; submitted 13.Dec.2025; peer-reviewed by R Singh, T Nakabe; revised version received 03.May.2026; accepted 04.May.2026; published 11.Jun.2026.

Please cite as:
Ruan J, Guan C, Chen Q, Yao T, Liu Y, Zhang Y, Xie L
Experiences of Home-Dwelling Older Adults and Their Family Caregivers With Digital Health Services: Qualitative Study
J Med Internet Res 2026;28:e89496
URL: https://www.jmir.org/2026/1/e89496
doi:10.2196/89496

© JiaYi Ruan, CuiXia Guan, QingQing Chen, Ting Yao, Yun Liu, YuXia Zhang, LunFang Xie. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Harnessing Participatory Surveillance Cohorts and Proxy Indicators to Dynamically Track Epidemic Trends and Undiagnosed COVID-19 Infections in Singapore: Longitudinal Observational Study

Sheng En Alexius Matthias Soh[1], BEng, MSc; Aung Hein Aung[2], MBBS, MPH; Wei Ling Brenda Ong[1], BSc; Kangwei Zeng[3], BSc, MPH; Jean-Marc Chavatte[4], BSc, Maitrise, MSc, PhD; Lin Cui[4], BSc, MSc, PhD; Raymond Valentine Tzer Pin Lin[5], MBBS, FRCPA; Vanessa W Lim[1], PhD; May O Lwin[6], BA, MBA, PhD; I-Cheng Mark Chen[2], MBBS, PhD

[1]NCID Research Office, National Centre for Infectious Diseases, Singapore, Singapore

[2]Department of Epidemiology and Preventive Medicine, Tan Tock Seng Hospital, 11 Jln Tan Tock Seng, Singapore, Singapore

[3]Advanced Methods and Analytics, Communicable Diseases Agency, Singapore, Singapore

[4]National Public Health Laboratory, Communicable Diseases Agency, Singapore, Singapore

[5]Microbiology Division, Department of Laboratory Medicine, National University Hospital, Singapore, Singapore

[6]Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore, Singapore

Corresponding Author:
I-Cheng Mark Chen, MBBS, PhD
Department of Epidemiology and Preventive Medicine, Tan Tock Seng Hospital, 11 Jln Tan Tock Seng, Singapore, Singapore

Abstract

Background: Accurate COVID-19 incidence estimates, including undiagnosed cases, are vital for epidemic management but are often unavailable in real time. Participatory surveillance can capture community illness episodes; however, quantifying undiagnosed infections remains difficult. We assessed a Singaporean cohort to estimate medically unattended COVID-19 infections by combining symptom models with proxy epidemic indicators.

Objective: This study aims to estimate COVID-19 incidence and medically attended fractions using participatory surveillance data and to evaluate the consistency of these estimates against independently derived serological measures of infection in a community cohort in Singapore.

Methods: We analyzed 11 survey waves (September 2021 to November 2022) from the SOCRATEs (Strengthening Our Community's Resilience Against Threats from Emerging Infections) community cohort (n=1899), spanning Delta and Omicron variant waves. Respondents reported recent illness, symptoms, health care use, and COVID-19 diagnoses. Multilevel logistic regression of medically attended episodes estimated the probability of COVID-19 in unattended episodes, incorporating symptoms and external indicators—wastewater viral-load index and health care staff surveillance. The estimates of total infections and medically attended fractions were validated against independent serological survey results.

Results: Among 2284 illness episodes, 756 were diagnosed with COVID-19, of which 62.4% (472/756) were medically attended. Health care–seeking declined from 83.9% (26/31) of COVID-19 episodes early in 2022 to 55.3% (47/85) by late 2022. Regression models demonstrated strong associations between COVID-19 infection and key symptoms and epidemic activity indicators. Estimated total infections were substantially higher than notified cases, reaching 1.0 to 2.9 times the reported incidence across successive variant waves. Model-based estimates of medically attended fractions were broadly consistent with serological benchmarks. Incidence estimates closely matched serological estimates in earlier intervals, with slight overestimation in later periods due to reinfection.

Conclusions: Participatory surveillance, when combined with probabilistic modeling and external indicators of epidemic activity, can generate robust estimates of infection incidence and health care use. Agreement with serological data supports the validity of this integrated framework, although discrepancies persist in later epidemic phases due to reinfection dynamics. This approach provides a scalable and timely complement to traditional and serological surveillance systems.

(J Med Internet Res 2026;28:e85608) doi:10.2196/85608

XSL•FO
RenderX

KEYWORDS

COVID-19; SARS-CoV-2; COVID-19 testing; public health surveillance; wastewater-based epidemiological monitoring; incidence; seroepidemiologic studies; cohort studies; logistic models; multilevel analysis

Introduction

The COVID-19 pandemic disrupted global public health systems in unprecedented and profound ways between 2020 and 2023. With its high transmissibility, overlap in clinical manifestations with endemic respiratory infections, potential for severe health outcomes, and capacity for asymptomatic transmission, COVID-19 rapidly evolved into a worldwide health crisis. Traditional epidemiological surveillance mechanisms, fundamental to policy decisions and epidemic response, encountered significant limitations as the pandemic rapidly progressed [1-4]. These included delays in data collection, analysis, and dissemination. To inform policymakers about preventable mortality and the impact of an infection on health systems, and hence the necessity of containment measures, an appropriate assessment of disease severity is paramount [5]. However, this relies critically on timely and accurate estimates of the proportion of the population that has been infected. Estimating the proportion infected is also critical for predicting an epidemic's trajectory, assessing susceptibility to future waves of infection, and formulating vaccination strategies [6]. However, reliable data were not readily available during the early phases of the pandemic [7] and were largely unavailable in near real time. Serological studies [8,9] eventually provided clarity on the proportions infected, but they arrived too late to influence many critical decisions in earlier pandemic response efforts.

A potential method to address the gap is participatory surveillance [10-12], using digital technology [13-16] to collect near real-time [17,18] community data on illness and unreported cases. The participatory surveillance concept has been widely studied, particularly for influenza, with a 2023 review by Atkins et al [19] describing the advantages of such systems to complement traditional syndromic surveillance by providing almost real-time insights. Although work has compared trends from participatory surveillance [20-23] with data from national-level sources and laboratory-based surveillance [24-29], most of this work has focused on influenza and has not elaborated on how to integrate such data sources to estimate population-level incidence, prevalence, and attack rates for successive epidemic waves.

Interpreting infection prevalence remains challenging due to changing health-seeking behaviors [30] and overlapping symptoms of infections. A potential simplification is to assume episodes without health care visits have the same infection prevalence as outpatient cases, but this may be inaccurate. Our previous participatory surveillance study showed symptoms were associated with increased health care visits and likelihood of infection [31]. Recognizing this, a German study combined participatory and virological surveillance data while stratifying episodes by symptom profile [32]. The approach did yield higher incidence estimates than traditional systems, with a good temporal correlation with wastewater surveillance data. Innovations such as wastewater testing [33-36] and antigen rapid tests (ARTs) are now common [33-36] and can be timely epidemic indicators but cannot estimate infection rates. What is needed is a framework for combining these modalities with participatory cohorts to accurately estimate infection proportions and validate such approaches.

We address the above issues through a participatory surveillance cohort active in Singapore during the COVID-19 pandemic, which conducted repeated, digitally administered surveys to capture illness episodes, associated health care–seeking behaviors, and symptoms. Leveraging this, we aimed to do the following:

1. Evaluate the use of our participatory surveillance cohort for near real-time monitoring of COVID-19 activity in Singapore.
2. Demonstrate an approach that integrates participatory surveillance with other data streams (eg, wastewater testing, health care staff surveillance, percentage positive for COVID-19 at sentinel clinics) to estimate the additional amount of COVID-19 that would not be medically attended and, hence, not notified as cases to the Ministry of Health.
3. Compare our findings against passive case notifications and results from population-based serological surveys to assess the validity of our approach.

In doing so, we articulate a framework to integrate participatory surveillance with other surveillance streams to produce near real-time estimates of incidence rates for diagnosed and undiagnosed COVID-19, which could also be applied to other infections.

Methods

Study Design

Details of the SOCRATEs (Strengthening Our Community's Resilience Against Threats from Emerging Infections) study were previously described in Lim et al [37]. Participants provided consent either in-person or through teleconference platforms. Follow-ups were conducted through online surveys, with each survey open for approximately 2 to 3 weeks after its launch date. The interval between survey launch dates varied depending on the needs at different stages of the pandemic and ranged from 1 to 2 months during the period of interest.

Surveys on COVID-19 Infection Symptoms, Health Care Visits, and COVID-19 Diagnoses

This analysis included 11 survey waves, spanning 4 epidemic peaks caused by the Delta COVID-19 variant, Omicron BA.1/BA.2, BA.5, and XBB subvariants (Figure 1A). The first survey (wave 28) opened on September 22, 2021, while the last survey (wave 38) closed on November 28, 2022. These waves included the period after COVID-19 ART self-test kits became widely available in mid-2021. They coincided with a transition in Singapore's national approach to managing COVID-19 toward "living with COVID-19" [38,39]. This transition marked

XSL•FO
RenderX

the end of comprehensive contact tracing and quarantine, and consequently, the transmission of COVID-19 was no longer confined to localized outbreaks. To support disease surveillance, all surveys during this period asked whether respondents had experienced acute illness episodes in the 28 days before submitting their response. If so, they were asked when the episode occurred, if their illness was accompanied by a core list of symptoms associated with COVID-19 (cough, runny nose, sore throat, breathlessness, loss of sense of smell, and feeling feverish), and whether they had a physical health care visit for their illness. Additional questions assessed whether the health care visit resulted in a diagnosis of COVID-19 ("medically diagnosed") and whether they also self-tested positive by ART. Episodes that self-tested positive without a health care visit were considered "self-diagnosed."

Figure 1. Proportion of episodes with and without COVID-19 diagnosis that were medically attended. (A) Total respondents, illness episodes, and COVID-19 diagnoses, (B) proportion of episodes with COVID-19 diagnoses that were medically attended, and (C) proportion of episodes with no COVID-19 diagnosis that were medically attended.

XSL•FO
RenderX

Data Analysis

Our analysis aimed to estimate the incidence of COVID-19 infections, including those not medically attended and thus not reported to the Ministry of Health. Methodological details and notations used are provided in Table S1-1 in Multimedia Appendix 1. Briefly, the unit of analysis was the participant-response-day, with each participant contributing person-day observations to calculate incidence rates over the 28 days preceding each survey response. Episodes within this period contributed fully, while others were weighted for partial contribution (Tables S2 and S3 in Multimedia Appendix 1). In our analysis, we included all observed seroconversions within each interval and aligned the data by assuming an average lag of approximately 2 weeks [40,41] between infection onset and seroconversion. Syndromic incidence estimates are therefore aggregated over intervals shifted by 2 weeks relative to the corresponding serological sampling windows, defined using the median dates of the initial and terminal samples for each interval. For the estimation of ascertainment fractions, we computed this by including illness episodes reported up to 2 weeks prior to the initial sample of each interval as potentially contributing to observed seroconversions and classified these as medically attended where appropriate.

Time was not modeled using explicit calendar-date terms (eg, linear trends or interval-specific dummy variables). Instead, temporal variation in infection risk was incorporated through time-varying proxy indicators of epidemic activity, evaluated at the calendar dates corresponding to each participant-response-day observation. Depending on model specification, these indicators included the incidence of medically attended COVID-19 within the cohort, incidence among health care staff, wastewater viral-load index, or the proportion positive for COVID-19 at sentinel clinics. This approach allowed the model to capture nonlinear, wave-specific changes in transmission intensity without imposing parametric assumptions on calendar time.

To account for correlations arising from repeated observations, we fitted a multilevel logistic regression with random intercepts at 3 levels: participants with multiple responses, each representing observation time across dates. The models included demographics, symptoms, optional proxy indicators of epidemic activity—medically attended COVID-19 incidence, lab- and self-reported testing among health care staff [42], wastewater viral-load index (WVI) from the National Environment Agency [35], and the proportion positive for COVID-19 at sentinel clinics tested by the National Public Health Laboratory [43]. We did not impose an explicit temporal autocorrelation structure at the day level (eg, autoregressive terms), as survey responses were irregularly spaced in time, and epidemic-level temporal dependence was already represented by the external proxy indicators. All models were therefore interpreted as estimating symptom-infection relationships conditional on contemporaneous epidemic intensity rather than modeling time as an independent covariate.

During the study, the Ministry of Health [44,45] stipulated COVID-19 testing for all adults with respiratory symptoms presenting at health care facilities. We considered COVID-19 status to be accurate for medically attended episodes but less reliable for unattended episodes due to incomplete testing and lower self-testing accuracy. Therefore, only medically attended episodes were used in multivariable models to assess the predictive value of symptoms. These models were then used to predict the likelihood of COVID-19 in unattended episodes based on symptoms and epidemic timing. Since symptoms were crucial, analyses used only episodes with at least 1 of 6 symptoms: cough, runny nose, sore throat, breathlessness, loss of smell, or feeling feverish.

For a given calendar date, the logistic model's predictions were used to estimate the average weighted probability that an illness episode could be COVID-19. This was then applied in 2 methods to estimate the proportion of COVID-19 cases that were medically attended. Both methods involved the sum of medically attended COVID-19 observations as the numerator and 1 part of the denominator. Method 1 included only 1 additional component in the denominator: the sum of observations for unattended episodes weighted by the probability of being COVID-19 over time. This approach ignores reports of self-diagnosed COVID-19 episodes. Conversely, method 2 incorporated into its denominator the total number of observations for both medically attended and self-diagnosed COVID-19 episodes, along with unattended episodes without a COVID-19 diagnosis, weighted by the corresponding probability of COVID-19.

The medically attended fraction of COVID-19 was stratified into 3 intervals guided by the collection dates for serological samples [46] against which the results were compared (Table S4 in Multimedia Appendix 1). These corresponded approximately to periods when the Delta, Omicron BA.1/BA.2, and Omicron BA.5/XBB variants were circulating. We also compared the cumulative incidence rates estimated from seroconversions, which approximate population-level infections when new infections still vastly outnumber reinfections (Multimedia Appendix 1).

The Institutional Review Board of the National Healthcare Group approved the SOCRATEs (reference 2018/01203, approved February 22, 2019) and serological (reference 2020/00805, approved July 23, 2020) studies. All statistical analyses were performed using STATA 18 on Windows 10, and P values <.05 were considered statistically significant. The STROBE checklist (Checklist 1) [47] was used to guide our reporting of results.

Ethical Considerations

The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of the National Healthcare Group (study reference number 2018/01203 and study date of approval February 22, 2019). All participants involved in the study provided informed consent. The data were anonymized and deidentified prior to analysis. This is not a clinical trial.

Results

Of the 2141 total participants in the SOCRATEs cohort (Table 1), 1899 responded to one or more of the survey waves (waves

XSL•FO
RenderX

28 - 38). Of these, 159 (8%) were from door-to-door recruitment, 730 (38%) from other cohorts, and 1010 (53%) from referrals. Respondents for the included survey waves were similar to the overall cohort recruited and distributed evenly across age groups. The majority identified as Chinese (n=1687, 89%), held tertiary education (n=1096, 58%), were female (n=1151, 61%), and resided in 4 - to 5-bedroom publicly owned flats (n=1097, 58%). Minority ethnic groups were underrepresented relative to population data from the Singapore Department of Statistics, as were individuals with fewer years of education (n=260, 14% with "O"/"N" level and below among respondents vs n=1,142,800, 37% in national data).

Table . SOCRATEs (Strengthening Our Community's Resilience Against Threats from Emerging Infections) cohort demographics.

Demographics	SOCRATEs cohort as at December 31, 2022 (N=2141), n (%)	Participants with at least 1 survey after wave 28 (N=1899), n (%)	Population data (year 2023)[a] (%)
Age (y)			
18 - 29	378 (18)	328 (17)	20
30 - 39	457 (21)	390 (21)	18
40 - 49	403 (19)	362 (19)	18
50 - 59	396 (18)	373 (20)	17
>60	507 (24)	446 (23)	28
Gender			
Male	866 (40)	748 (39)	49
Female	1275 (60)	1151 (61)	51
Ethnicity			
Chinese	1862 (87)	1687 (89)	74
Malay	72 (3)	47 (2)	14
Indian	153 (7)	119 (6)	9
Others	54 (3)	46 (2)	3
Education[b]			
"O" or "N" level and below	314 (15)	260 (14)	37
"A" level or polytechnic diploma	626 (29)	543 (29)	27
University or postgraduate	1201 (56)	1096 (58)	36
Housing type			
Publicly owned flat with ≤3 rooms	297 (14)	253 (13)	24
Publicly owned flat with 4 - 5 rooms	1254 (59)	1097 (58)	54
Privately owned property	590 (28)	549 (29)	22
Pre-existing conditions			
Diabetes	98 (5)	84 (4)	—[c]
Hypertension	214 (10)	186 (10)	—

[a]Taken from singstat.gov.sg (figures may not add up to the totals due to rounding).

[b]Population education data restricted to 25 years and older because that is what is available.

[c]Not available.

Each wave included 1300 to 1600 respondents (Figure 1A) and reported 2511 illness episodes between September 22, 2021, and November 28, 2022. After excluding 227 episodes that occurred more than 28 days before the survey date, 2284 episodes were analyzed. Except for 2 brief gaps in December 2021 and July 2022, almost all dates within the study period fell within the survey response windows or were covered by the 28-day lookback period from the start of each window. The proportions reporting illness episodes increased from 7.4% in wave 28 to more than 20% by wave 38. Episodes with COVID-19 diagnoses increased from 1 (0.1%) episode in wave 28 to a peak of 180 (12.5%) in wave 32 and then ranged between 5% and 8% of responses from wave 33 onward. Episodes without a COVID-19 diagnosis ranged from 7% to 9% during waves 28 to 33 and then more than doubled to 15.2% by wave 38. Among episodes with a COVID-19 diagnosis (Figure 1B), 62.4% (472/756) were medically attended. This proportion decreased from about 80% in waves up to early 2022 to less

XSL•FO
RenderX

than 60% by wave 38. In episodes without a COVID-19 diagnosis (Figure 1C), 811 out of 1528 (53.1%) were medically attended, and the proportion decreased from 68.3% in wave 28% to 52.4% by wave 38.

Among episodes with at least 1 of the 6 core symptoms, medically attended and unattended episodes had a median of 3 (IQR 2-4) versus 2 (IQR 1-4) symptoms respectively (P<.001), and all symptoms except runny nose were significantly more common (Table 2). However, most differences were less marked in stratified analysis by COVID-19 status. In COVID-19 episodes, the median number of symptoms was 4 (IQR 3-5) in both attended and unattended (IQR 3-4) episodes (P=.12). Symptom profiles were also similar. For instance, 88.4% versus 83.7% had cough (P=.09), and 91.8% versus 89.0% had sore throat (P=.24) in attended and unattended episodes, respectively. Not unexpectedly, attended episodes had a lower proportion of unreported temperatures (31.7% vs 49.7%; P<.001), but the distribution of temperature readings was not significantly different (P=.86). However, in episodes with no COVID-19 diagnosis, several symptoms were more common in medically attended versus unattended episodes, including cough (54.3% vs 42.4%; P<.001), shortness of breath (14.4% vs 11.7%; P=.07), and feeling feverish (45.9% vs 39.3%; P=.02). When comparing episodes with and without a COVID-19 diagnosis, the median number of symptoms was significantly greater, and all symptoms were significantly more common, both for medically attended and unattended episodes. COVID-19 episodes were also significantly more likely to report temperatures of 38.5 °C or higher. However, temperature readings were unavailable for a large proportion of observations and were therefore not used in multivariable analysis to discriminate COVID-19 from other illness episodes.

XSL•FO
RenderX

Table . Comparison of symptoms by COVID-19 diagnosis and whether the episode is medically attended or unattended.

Characteristics	All episodes with symptoms reported			Had COVID-19 diagnosis			No COVID-19 diagnosis			Comparison by COVID-19 diagnosis	
	A. Medically attended	B. Medically unattended	*P* value, A vs B	C. Medically attended	D. Medically unattended	*P* value, C vs D	E. Medically attended	F. Medically unattended	*P* value, E vs F	*P* value, C vs E	*P* value, D vs F
Number of episodes	1004	791	—[a]	464	209	—	540	582	—	—	—
Symptoms, median (IQR)	3 (2-4)	2 (1-4)	<.001[b]	4 (3-5)	4 (3-4)	.12[b]	2 (1-3)	2 (1-3)	.002[b]	<.001[b]	<.001[b]
Reported symptoms, n (%)											
• Cough	703 (70)	422 (53.4)	<.001	410 (88.4)	175 (83.7)	.09	293 (54.3)	247 (42.4)	<.001	<.001	<.001
• Runny nose	676 (67.3)	516 (65.2)	.35	364 (78.4)	157 (75.1)	.34	312 (57.8)	359 (61.7)	.18	<.001	<.001
• Sore throat	781 (77.8)	544 (68.8)	<.001	426 (91.8)	186 (89)	.24	355 (65.7)	358 (61.5)	.14	<.001	<.001
• Shortness of breath	195 (19.4)	111 (14)	.003	112 (24.1)	43 (20.6)	.31	83 (15.4)	68 (11.7)	.07	<.001	.002
• Loss of smell	119 (11.9)	63 (8)	.007	92 (19.8)	35 (16.7)	.35	27 (5)	28 (4.8)	.88	<.001	<.001
• Felt feverish	579 (57.7)	376 (47.5)	<.001	331 (71.3)	147 (70.3)	.79	248 (45.9)	229 (39.3)	.03	<.001	<.001
Felt feverish but no temperature reported[c], n (%)	225 (38.9)	214 (56.9)	<.001	105 (31.7)	73 (49.7)	<.001	120 (48.4)	141 (61.6)	.004	<.001	.02
Reported temperature (°C) in those felt feverish[d], n (%)			.06			.86			.11	.001	<.001
• <38.0	155 (43.7)	79 (54.1)		84 (37.2)	26 (40)		71 (55)	53 (65.4)			
• 38.0-38.4	81 (22.8)	32 (21.9)		52 (23)	13 (20)		29 (22.5)	19 (23.5)			
• >38.5	119 (33.5)	35 (24)		90 (39.8)	26 (40)		29 (22.5)	9 (11.1)			

[a]Not applicable.

[b]Wilcoxon rank-sum test; all other *P* values by chi-squared tests

[c]Denominator is those who felt feverish.

[d]Denominator is those who felt feverish and reported a valid temperature.

XSL•FO
RenderX

Table 3 shows the regression model results for 1004 medically attended episodes from 705 participants. In univariable analyses, those aged 30 to 39 years had lower odds of COVID-19 than those younger than 30 years, but there were no other significant associations with sociodemographic variables or pre-existing medical conditions. In contrast, significant positive associations (at $P<.01$) were observed for all symptoms and indicators of epidemic activity in univariable analysis. Across all multivariable models, significant associations remained for 4 symptoms and all indicators of epidemic activity in the respective models. For instance, when using WVI as a proxy indicator, the adjusted odds ratios (95% CIs) for cough, sore throat, loss of sense of smell, and feeling feverish were 3.6 (2.2 - 5.7), 3.8 (2.2 - 6.5), 4.3 (2.2 - 8.4), and 2.8 (1.9 - 4.1), respectively, while the log of WVI had an adjusted odds ratio of 2.2 (1.8 - 2.8).

XSL•FO
RenderX

Table . Univariable and multivariable logistic regression showing results for demographics, symptoms, proxy indicators and random intercept terms.

Characteristics	Univariable analyis, OR (95% CI)	Multivariable analysis[a] for different proxy indicator options			
		No proxy indicators, OR (95% CI)	Log of incidence per 1000 person days, OR (95% CI)[b]	Log of incidence per 1000 health care staff days, OR (95% CI)	Log of wastewater viral-load index, OR (95% CI)
Age (y; vs 18 - 29)					
30 - 39	0.58 (0.37 - 0.91)[c]	0.49 (0.25 - 0.94)[c]	0.51 (0.26 - 1)	0.54 (0.28 - 1.04)	0.55 (0.29 - 1.04)
40 - 49	1.03 (0.65 - 1.63)	0.83 (0.43 - 1.6)	0.8 (0.41 - 1.57)	0.83 (0.43 - 1.62)	0.86 (0.46 - 1.64)
50 - 59	0.97 (0.6 - 1.58)	0.91 (0.46 - 1.81)	0.95 (0.47 - 1.92)	0.98 (0.49 - 1.97)	1.01 (0.52 - 1.97)
>60	1.56 (0.93 - 2.63)	1.57 (0.71 - 3.46)	1.43 (0.64 - 3.23)	1.48 (0.67 - 3.29)	1.56 (0.72 - 3.36)
Female gender (vs male)	0.77 (0.58 - 1.03)	0.79 (0.55 - 1.14)	0.73 (0.51 - 1.07)	0.74 (0.51 - 1.07)	0.73 (0.51 - 1.05)
Race (vs Chinese)					
Malay	0.81 (0.39 - 1.69)	0.92 (0.4 - 2.13)	1.07 (0.4 - 2.89)	1.13 (0.43 - 3)	1.06 (0.42 - 2.71)
Indian	0.74 (0.42 - 1.31)	0.65 (0.28 - 1.52)	0.63 (0.26 - 1.49)	0.63 (0.27 - 1.5)	0.64 (0.28 - 1.47)
Others	0.53 (0.2 - 1.38)	0.56 (0.15 - 2.17)	0.49 (0.13 - 1.87)	0.49 (0.13 - 1.84)	0.48 (0.13 - 1.8)
Housing type (vs publicly owned flat with ≤3 rooms)					
Publicly owned flat with 4 - 5 rooms	0.81 (0.51 - 1.28)	0.89 (0.49 - 1.63)	0.87 (0.46 - 1.62)	0.86 (0.47 - 1.59)	0.87 (0.48 - 1.57)
Privately owned property	1.13 (0.67 - 1.91)	1.13 (0.57 - 2.25)	1.14 (0.56 - 2.31)	1.13 (0.57 - 2.26)	1.13 (0.58 - 2.22)
Education (vs ''O" or "N" level & below)					
"A" level or polytechnic diploma	0.69 (0.41 - 1.16)	0.82 (0.4 - 1.7)	0.85 (0.4 - 1.79)	0.89 (0.42 - 1.85)	0.91 (0.44 - 1.89)
University or postgraduate	0.66 (0.41 - 1.06)	0.74 (0.37 - 1.48)	0.81 (0.39 - 1.67)	0.82 (0.4 - 1.66)	0.85 (0.42 - 1.71)
Occupational status (vs employed)					
Student	0.99 (0.55 - 1.78)	0.58 (0.22 - 1.51)	0.63 (0.23 - 1.69)	0.61 (0.23 - 1.61)	0.64 (0.26 - 1.63)
Self-employed	1.15 (0.66 - 2.01)	1.4 (0.76 - 2.6)	1.35 (0.69 - 2.62)	1.43 (0.75 - 2.75)	1.32 (0.71 - 2.47)
Not employed nor studying	1.44 (0.93 - 2.22)	1.24 (0.68 - 2.23)	1.21 (0.66 - 2.22)	1.24 (0.69 - 2.26)	1.19 (0.67 - 2.13)
Pre-existing conditions (vs none)					
Diabetes	1.21 (0.55 - 2.65)	1.21 (0.44 - 3.35)	1.16 (0.44 - 3.05)	1.2 (0.45 - 3.21)	1.24 (0.47 - 3.24)
Hypertension	1.36 (0.84 - 2.21)	0.8 (0.41 - 1.57)	0.76 (0.39 - 1.46)	0.78 (0.41 - 1.51)	0.75 (0.39 - 1.43)
Symptoms					
cough	7.38 (4.88 - 11.16)[e]	4.24 (2.62 - 6.87)[e]	3.56 (2.17 - 5.82)[e]	3.35 (2.06 - 5.44)[e]	3.57 (2.23 - 5.71)[e]
runny nose	3 (2.13 - 4.25)[e]	1.42 (0.95 - 2.12)	1.42 (0.94 - 2.15)	1.4 (0.93 - 2.11)	1.38 (0.93 - 2.05)
sore throat	6.96 (4.35 - 11.15)[e]	4.24 (2.44 - 7.39)[e]	3.69 (2.13 - 6.39)[e]	3.81 (2.21 - 6.57)[e]	3.78 (2.21 - 6.46)[e]
breathlessness	1.95 (1.33 - 2.87)[d]	1.25 (0.78 - 1.99)	1.21 (0.75 - 1.95)	1.17 (0.73 - 1.88)	1.18 (0.74 - 1.88)
loss of smell	6.23 (3.56 - 10.89)[e]	3.89 (2.01 - 7.52)[e]	5.18 (2.49 - 10.77)[e]	5.25 (2.52 - 10.94)[e]	4.31 (2.21 - 8.38)[e]
feeling feverish	3.45 (2.41 - 4.94)[e]	3.24 (2.18 - 4.81)[e]	2.9 (1.96 - 4.3)[e]	2.87 (1.94 - 4.25)[e]	2.83 (1.93 - 4.13)[e]
Proxy indicators, log of					

XSL•FO
RenderX

Characteristics	Univariable analyis, OR (95% CI)	Multivariable analysis[a] for different proxy indicator options			
		No proxy indicators, OR (95% CI)	Log of incidence per 1000 person days, OR (95% CI)[b]	Log of incidence per 1000 health care staff days, OR (95% CI)	Log of wastewater viral-load index, OR (95% CI)
Incidence per 1000 person days	3.37 (2.6 - 4.35)[e]	-[f]	3.11 (2.33 - 4.15)[e]	-[f]	-[f]
Incidence per 1000 health care staff days	2.58 (2.09 - 3.19)[e]	-[f]	-[f]	2.32 (1.84 - 2.94)[e]	-[f]
Wastewater viral-load Index	2.5 (2.07 - 3.02)[e]	-[f]	-[f]	-[f]	2.22 (1.79 - 2.75)[e]
Random intercept terms					
Between participants	-[f]	0.55 (0.13 - 2.33)[e]	0.68 (0.18 - 2.54)[e]	0.69 (0.19 - 2.44)[e]	0.63 (0.17 - 2.32)[e]
Between responses within particpants	-[f]	0 (0 - 0)[e]	0 (0 - 0)[e]	0 (0 - 0)[e]	0 (0 - 0)[e]

[a]Full model includes symptoms, respective options for proxy indicators shown in the table, plus demographics and random intercept terms for variability between participants and between survey responses within participants.

[b]From medically ascertained episodes within SOCRATEs cohort.

[c]$P<.05$.

[d] $p < .01$

[e]$P<.001$.

[f]Not applicable

Key epidemics of COVID-19 were discernible as peaks in the weighted proportions diagnosed with COVID-19 for both medically attended [48] (gray area) and unattended episodes (green bars), but were clearer in the former. When any indicator of epidemic activity was included, the regression models gave similar trends in the estimated proportions infected both for all unattended (Figure 2A) and unattended episodes with no COVID-19 diagnosis (Figure 2B), even when using sentinel clinic percentage positive where data were available only from May 2022 onward. Trends in the model without any proxy indicator were much less discernible. The differential prevalence of key symptoms (Table 2) led to estimated infections for unattended episodes being generally lower than the observed weighted proportions infected for medically attended episodes (Figure 2A). Among unattended episodes, estimated infections were marginally higher than weighted proportions of self-diagnosed COVID-19, particularly during epidemic peaks (Figure 2A). Conversely, estimated infections among unattended episodes with NO COVID-19 diagnosis (Figure 2B) were mostly lower than the observed weighted proportions for self-diagnosed COVID-19.

XSL•FO
RenderX

Figure 2. Model estimates of proportion infected and the incidence rates from estimating infections for unattended episodes with and without COVID-19 diagnosis. (A) Model estimates of proportion infected for unattended episodes, (B) model estimates of proportion infected for unattended episodes with no COVID-19 diagnosis, (C) method 1: incidence rates from estimating infections in unattended episodes (using wastewater viral-load index [WVI] as proxy), and (D) method 2: incidence rates from estimating infections in unattended episodes with no COVID-19 diagnosis (using WVI as proxy).

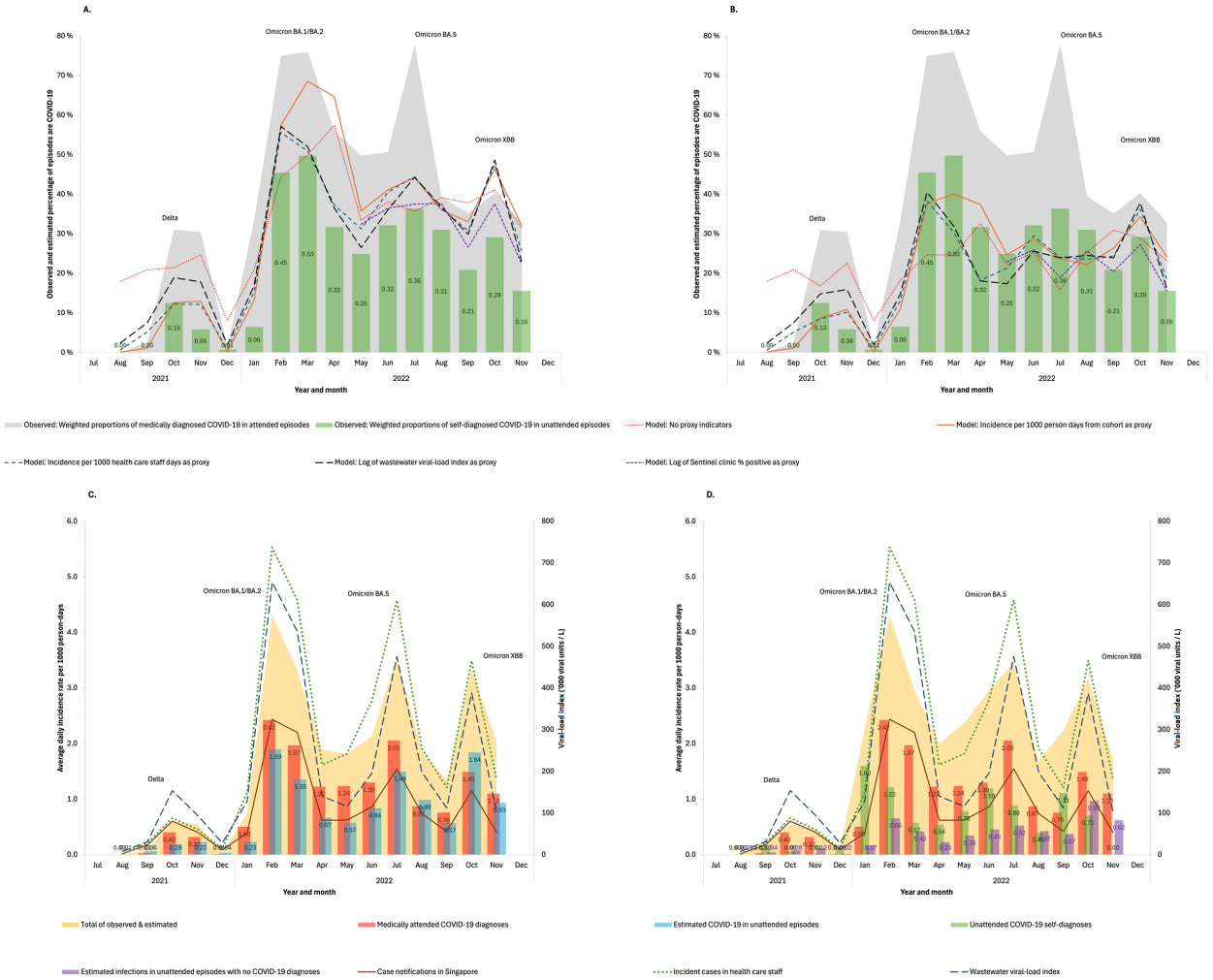

Because WVI time-series data were the most complete, we used them to model COVID-19 incidence for unattended episodes in Figure 2C and D. There was a slight divergence in the latter half of the study period, but nationally notified cases and medically attended COVID-19 infections observed in the cohort were very similar, with incidence rates of 2.4 per 1000 persons per day at the peak of Omicron BA.1/BA.2 activity in February 2022. However, the incidence of infections was likely much higher, particularly at later time points, as a substantial proportion of unattended episodes were likely COVID-19 infections. Adding unnotified infections estimated from method 1 (yellow area), total infections were 1.0, 1.8, 2.3, and 2.9 times the rate of notified cases during the peaks of the Delta, Omicron BA.1/BA.2, Omicron BA.5, and Omicron XBB epidemics, respectively (Figure 2C). Total infections from method 1 were nearly equivalent to those from method 2, which separately estimated undiagnosed COVID-19 in unattended episodes with no COVID-19 diagnosis and summed this with incidence rates of self-diagnosed COVID-19. Although estimated by fitting to WVI data, total infection rates showed strong temporal correlation and were numerically close to those from health care staff.

Figure 3A shows that, across the first, second, and third intervals, respectively, 58.9%, 61.8%, and 42.5% of serologically detected incident COVID-19 infections were medically attended. For the proxy indicators, we omitted the sentinel clinic percentage positive because insufficient data were available for the first 2 intervals. However, models using the other 3 indicators gave fairly similar results (<20% difference), other than for the first interval, where estimates ranged from ~56% to~71%. For the latter 2 intervals, estimates for medically attended fractions deviated from those by serology by <9%. Estimates for the last interval were 3% to 9% higher but had bootstrapped CIs overlapping with the corresponding results from serology. For methods 1 and 2, the total deviance from serology over all 3 intervals was the smallest using WVI and staff surveillance, respectively.

XSL•FO
RenderX

Figure 3. Estimates of medically ascertained fractions and attack rates as compared to those from serology. (A) Estimates of medically ascertained fractions compared with those from serology and (B) estimates of attack rates compared with those from serology.

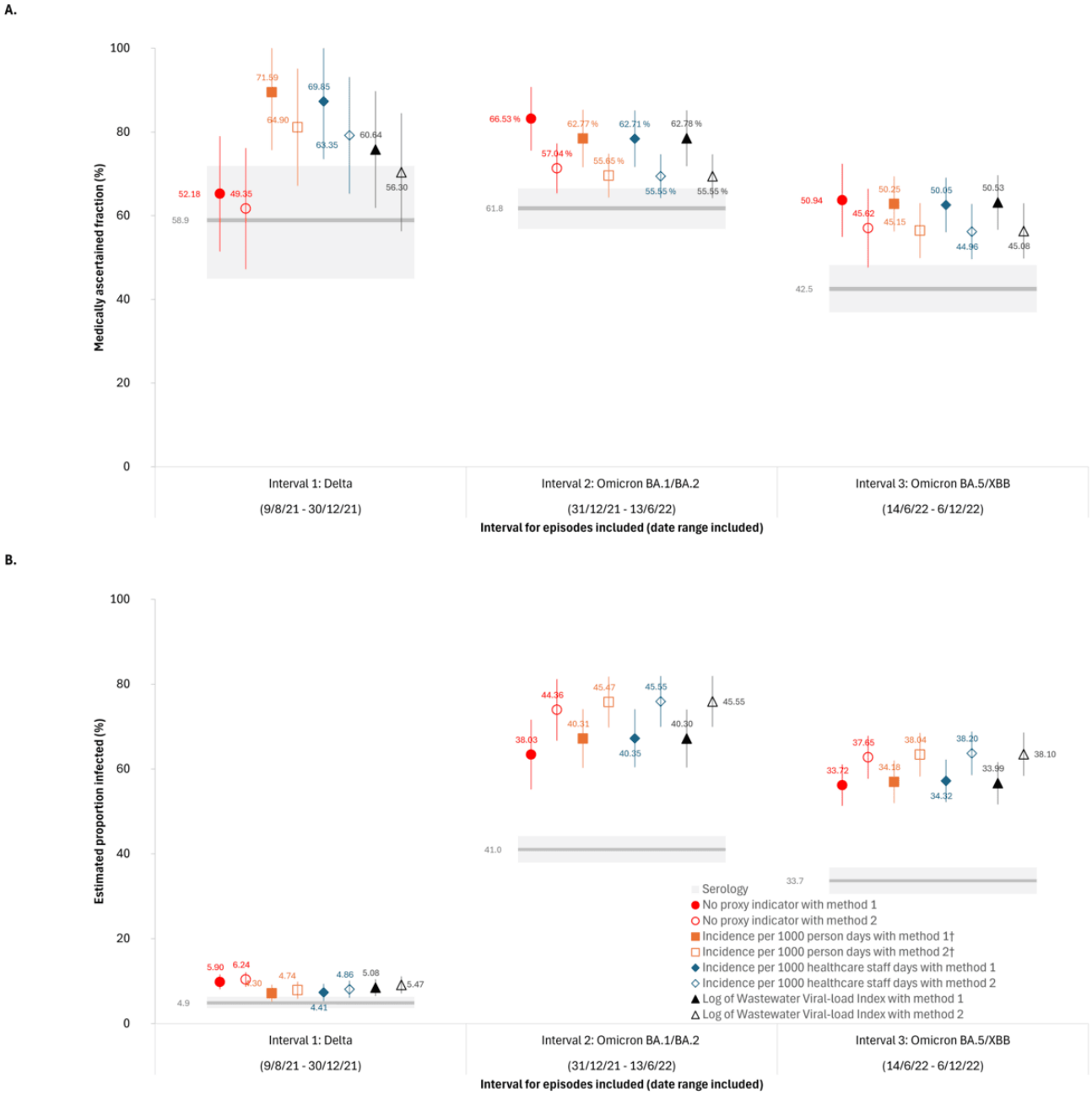

Based on the cumulative incidence of seroconversions divided by the number of valid samples, the serological study estimated that 4.9% were infected in interval 1% and 41% in interval 2 (Figure 3B). The 3 proxy indicators for which data were available gave similar estimates of 4.4% to 5.5% for interval 1% and 40.3% to 45.6% for interval 2. For interval 3, estimates from the 3 indicators ranged between 34% and 38.2%. This was generally higher than the 33.7% from serology, which could not estimate additional infections among those who were already seropositive due to past infections, and the final samples were collected about 1 month after our participatory surveillance data ended.

Discussion

The usefulness of digitally administered participatory surveillance for enhancing situational awareness during public health emergencies has been recognized [10,49-51] for several years. However, applying it to generate quantitative estimates of absolute incidence has remained difficult due to the lack of a known denominator and challenges in determining infection status among digitally reported cases. Our study makes progress in 2 areas. First, by integrating symptom follow-up questionnaires within a pre-enrolled, demographically characterized cohort, we were able to directly calculate person-time denominators rather than estimating them from user activity metrics. Second, by linking cohort data to another indicator of community transmission intensity, we could use regression techniques to estimate the likelihood of undiagnosed

XSL•FO
RenderX

infections, thus overcoming the dependence between infection detection and incidence estimation.

Our analysis shows that a participatory syndromic surveillance cohort—when integrated analytically with an external, closely timed indicator of epidemic activity, such as health care staff incidence rates, WVI, or sentinel clinic test positivity (though we had limited data points for this)—can generate estimates of COVID-19 incidence that are both time-correlated and numerically similar to national notifications over approximately 18 months covering 4 distinct SARS-CoV-2 variant waves. Additionally, the multilevel regression indicated that the likelihood of having COVID-19 among surveyed illness episodes depended systematically on symptom profiles. Symptoms among COVID-19 episodes were strikingly similar between respondents who sought health care and those who did not. Conversely, non-COVID respiratory episodes showed a clear gradient of symptom severity, with medically attended episodes reporting a higher median number of symptoms and greater prevalence of certain symptoms (eg, cough, feeling feverish). These findings support earlier research suggesting that individuals' decision to seek health care is more influenced by how severe they perceive their symptoms to be than by the illness's cause alone [31]. By combining symptom data with proxy indicators of epidemic activity, we could then plausibly estimate undiagnosed infections in episodes that were not medically attended. External validation against a serological cohort on the proportions of COVID-19 cases that were medically attended and the proportion infected further reinforced the reliability of this method.

The increase in self-diagnosed, nonpresenting COVID-19 episodes and the decline in health care–seeking from February 2022 suggest the behavioral impact of widespread ART kits and changing risk perception. As the epidemic evolved, case notifications likely underestimated infections. However, a naive approach applying health care–based test-positivity to all reported episodes of acute respiratory illness in the community would likely overestimate infections, since milder illness is less likely to be COVID-19. This highlights the challenge of designing surveillance systems resilient to changes in testing policy and care-seeking behavior.

Our findings corroborate those of Loenenbach et al [32] on how SARS-CoV-2 incidence rates estimated using participatory surveillance were temporally correlated with wastewater testing indices but were higher than official national data, particularly after free testing ended. Both studies emphasize the limitations of case notifications and the importance of community and environmental surveillance. We additionally propose ways to incorporate alternative indicators of epidemic activity through regression modeling. More crucially, we also demonstrated how our estimates of both medically attended and unattended infections approximate serological data, thereby validating the approach. Doing so builds confidence in an alternative framework to accurately quantify attack rates in near real time. Near real-time point prevalence was previously estimated using household polymerase chain reaction survey platforms, such as England's REACT [52] study, but with higher costs and complexity. A participatory surveillance cohort with limited diagnostic data from a subset of participants, augmented with routinely collected external indicators such as WVI [32], health care staff testing [53], or positivity rates from sentinel surveillance [54], would be more sustainable.

Validation against serological data should also be interpreted in light of evolving population immunity (Multimedia Appendix 1). Agreement between participatory surveillance–based estimates and seroconversion rates was strongest in earlier intervals, when most participants were serologically naïve, and incident infections largely represented primary infections. In the final interval of the serological cohort, a substantial proportion of participants were already seropositive at baseline, where we do not have validated methods for estimating the proportion infected, and it was effectively assumed that infections contributed by these individuals were negligible relative to infections in serologically naïve individuals. This likely explains why the incidence estimated on the basis of seroconversions alone was less than the incidence estimated from syndromic reports.

The strengths of this work include having a clearly defined denominator population, enabling true incidence-rate estimation compared to solely relying on other methods, such as clinic-based sentinels [54] and wastewater testing, both currently used in Singapore; the flexibility to integrate different indicators of community transmission; and triangulation with independent serological end points, lending validity to our approach.

The limitations include that 15% of illness reports fell outside the 28-day recall window, and recall bias may still exist within episodes. Symptom severity can influence testing decisions and health care access, potentially confounding results despite adjustments. Missing data on symptoms, episode dates, or test status were imputed, which might bias findings. Wastewater viral-load data are infrastructure-dependent and can be biased by virological [55] and climatological [56] factors. Due to the limited data points available, our validation of sentinel clinic test positivity as an alternative proxy indicator was less robust. The current method still needs a large, unbiased cohort with a subset willing to be tested and self-report results consistently. To replace unsustainable financial incentives, participatory surveillance could be scaled through citizen science approaches [57]. Alternative approaches could combine cohort symptom data with independent case-control study estimates [58] or sentinel clinic positivity rates stratified to match the symptom profiles of those in the participatory surveillance cohort. Also, the participatory surveillance cohort analyzed in this study was not designed to be population-representative, and participants were over-represented in some demographic areas. Recruitment pathways were also heterogeneous, including door-to-door enrollment, linkage from other cohorts, and participant referrals. Although cohort demographics differ from the national population, the resulting bias cannot be predetermined. Logistic regression (Table 3) showed no significant associations between demographic variables and COVID-19 diagnosis, even after adjustments; thus, demographic weighting was not used. Finally, behavioral responses and policies changed significantly during the study, including shifts in testing access and health care guidance, which may have influenced health care–seeking and testing across subgroups and affected estimates, especially in later epidemic phases. These factors likely contributed to the

XSL•FO
RenderX

decline in medically attended cases across waves (Figure 2A or B), independent of changes in disease severity. These features may be associated with differences in health engagement, health care–seeking behavior, and testing behavior, although the direction and magnitude of any resulting bias cannot be quantified directly from the available data.

This study discusses integrating participatory surveillance into public health monitoring. These are adaptable, not fixed: (1) use wastewater data as a temporal independent signal of epidemic activity to supplement participatory data on illness episodes that estimate attack rates despite changes in incidence of infection, health-seeking, and testing. Public agencies should thus seek to integrate such wastewater reports with routine updates. (2) Focus on maintaining cohort participation and diagnostic consistency rather than merely increasing size. Systems should prioritize the retention of a small subpopulation with linkage to diagnostics and use this to periodically calibrate the parameters used to estimate infection and attack rates.

To conclude, we demonstrated how participatory syndromic surveillance, linked to a reliable epidemic proxy, offers timely, credible estimates of infection burden, complementing traditional and serological systems. By accounting for variability in health care–seeking and testing, it overcomes the limitations of passive surveillance and is timelier than serology. It can provide a scalable model for epidemic monitoring amid rapid behavioral and policy shifts.

Acknowledgments

The authors thank the participants of the SOCRATEs (Strengthening Our Community's Resilience Against Threats from Emerging Infections) study and the HELIOS study team at the Lee Kong Chian School of Medicine, Nanyang Technological University, for supporting the recruitment of SOCRATEs. They also thank Alicia J. Lim and Daniel R. X. Lim for supporting the laboratory work on serological samples. Artificial intelligence was used only to improve readability and language. It was not used to generate scientific insights, analyze or interpret data, or draw conclusions. All outputs were applied with human oversight, and the authors carefully reviewed and edited the results, acknowledging that artificial intelligence can produce incomplete, incorrect, or biased text.

Funding

The estate of the late Irene Tan Liang Kheng sponsored the SOCRATEs (Strengthening Our Community's Resilience Against Threats from Emerging Infections) study. The collection and analysis of serological samples was funded by NMRC COVID-19 Research Fund (Project number: COVID19RF-002). Dr Mark Chen has also received funding by the National Research Council of Singapore under HCSAINV25jul-0006.

Data Availability

The information outlined in this article may be made available upon request from the designated corresponding author. However, we must emphasize that such disclosure is subject to applicable legal and regulatory requirements, as well as the sole discretion of the corresponding author.

Authors' Contributions

Conceptualization: ICMC, SEAMS, WLBO, VWL
Data curation: ICMC, SEAMS
Formal analysis: ICMC, KZ, SEAMS
Funding acquisition: SEAMS, ICMC, WLBO
Investigation: ICMC, SEAMS, WLBO
Methodology: ICMC, SEAMS, WLBO, VWL
Project administration: SEAMS, WLBO
Resources: ICMC, SEAMS, WLBO
Supervision: ICMC
Validation: ICMC
Visualization: ICMC, SEAMS
Writing – original draft: SEAMS
Writing – review & editing: ICMC, SEAMS, AHA, WLBO, KZ, JMC, LC, RVTPL, VWL, MOL
All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Methods used to calculate incidence rates and ascertainment fractions, worked example to calculate weighting rules and bootstrap parameters.
[DOCX File, 153 KB - jmir_v28i1e85608_app1.docx]

Checklist 1
STROBE checklist.
[DOCX File, 35 KB - jmir_v28i1e85608_app2.docx]

References

1. Alvarez E, Bielska IA, Hopkins S, et al. Limitations of COVID-19 testing and case data for evidence-informed health policy and practice. Health Res Policy Syst 2023 Jan 25;21(1):11. [doi: 10.1186/s12961-023-00963-1] [Medline: 36698202]
2. Dimitris MC, Galea S, Marcus JL, Pan A, Sander B, Platt RW. What has the pandemic revealed about the shortcomings of modern epidemiology? What can we fix or do better? Am J Epidemiol 2022 May 20;191(6):980-986. [doi: 10.1093/aje/kwac012] [Medline: 35081616]
3. Zheng P, Li C, Zhang H, et al. Challenges of epidemiological investigation work in the COVID-19 pandemic: a qualitative study of the epidemiology workforce in Guangdong Province, China. BMJ Open 2022 Nov 15;12(11):e056067. [doi: 10.1136/bmjopen-2021-056067] [Medline: 36379656]
4. Xu K, Gao B, Li J, Xiang Y, Cao L, Zhao M. Clinical features, diagnosis, and management of COVID-19 vaccine-associated Vogt-Koyanagi-Harada disease. Hum Vaccin Immunother 2023 Aug 1;19(2):2220630. [doi: 10.1080/21645515.2023.2220630] [Medline: 37282614]
5. Lipsitch M, Riley S, Cauchemez S, Ghani AC, Ferguson NM. Managing and reducing uncertainty in an emerging influenza pandemic. N Engl J Med 2009 Jul 9;361(2):112-115. [doi: 10.1056/NEJMp0904380] [Medline: 19474417]
6. Immunization agenda 2030: a global strategy to leave no one behind. : World Health Organization; 2020 URL: https://cdn.who.int/media/docs/default-source/immunization/strategy/ia2030/ia2030-draft-4-wha_b8850379-1fce-4847-bfd1-5d2c9d9e32f8.pdf?sfvrsn=5389656e_69&download=true [accessed 2026-05-15]
7. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. BMJ 2020 Apr 7;369:m1328. [doi: 10.1136/bmj.m1328] [Medline: 32265220]
8. Kalish H, Klumpp-Thomas C, Hunsberger S, et al. Undiagnosed SARS-CoV-2 seropositivity during the first 6 months of the COVID-19 pandemic in the United States. Sci Transl Med 2021 Jul 7;13(601):eabh3826. [doi: 10.1126/scitranslmed.abh3826] [Medline: 34158410]
9. Winter AK, Martinez ME, Cutts FT, et al. Benefits and challenges in using seroprevalence data to inform models for measles and rubella elimination. J Infect Dis 2018 Jul 2;218(3):355-364. [doi: 10.1093/infdis/jiy137] [Medline: 29562334]
10. Lipsitch M, Santillana M. Enhancing situational awareness to prevent infectious disease outbreaks from becoming catastrophic. In: Global Catastrophic Biological Risks: Springer International Publishing; 2019:59-74. [doi: 10.1007/82_2019_172]
11. Guerrisi C, Turbelin C, Blanchon T, et al. Participatory syndromic surveillance of influenza in Europe. J Infect Dis 2016 Dec 1;214(suppl_4):S386-S392. [doi: 10.1093/infdis/jiw280] [Medline: 28830105]
12. Paolotti D, Carnahan A, Colizza V, et al. Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience. Clin Microbiol Infect 2014 Jan;20(1):17-21. [doi: 10.1111/1469-0691.12477] [Medline: 24350723]
13. Allen WE, Altae-Tran H, Briggs J, et al. Population-scale longitudinal mapping of COVID-19 symptoms, behaviour and testing. Nat Hum Behav 2020 Sep;4(9):972-982. [doi: 10.1038/s41562-020-00944-2] [Medline: 32848231]
14. Drew DA, Nguyen LH, Steves CJ, et al. Rapid implementation of mobile technology for real-time epidemiology of COVID-19. Science 2020 Jun 19;368(6497):1362-1367. [doi: 10.1126/science.abc0473] [Medline: 32371477]
15. Weights and methodology brief for the COVID-19 symptom survey by University of Maryland and Carnegie Mellon University, in partnership with Facebook. : Facebook Research; 2020 URL: https://cmu-delphi.github.io/delphi-epidata/symptom-survey/symptom-survey-weights.pdf?utm_source=chatgpt.com [accessed 2026-05-15]
16. Radin JM, Quer G, Pandit JA, et al. Sensor-based surveillance for digitising real-time COVID-19 tracking in the USA (DETECT): a multivariable, population-based, modelling study. Lancet Digit Health 2022 Nov;4(11):e777-e786. [doi: 10.1016/S2589-7500(22)00156-X] [Medline: 36154810]
17. Menni C, Valdes AM, Freidin MB, et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. Nat Med 2020 Jul;26(7):1037-1040. [doi: 10.1038/s41591-020-0916-2] [Medline: 32393804]
18. Rossman H, Keshet A, Shilo S, et al. A framework for identifying regional outbreak and spread of COVID-19 from one-minute population-wide surveys. Nat Med 2020 May;26(5):634-638. [doi: 10.1038/s41591-020-0857-9] [Medline: 32273611]
19. Atkins N, Harikar M, Duggan K, et al. What are the characteristics of participatory surveillance systems for influenza-like-illness? J Glob Health 2023 Oct 20;13:04130. [doi: 10.7189/jogh.13.04130] [Medline: 37856769]

XSL•FO
RenderX

20. Greenleaf AR, Francis S, Zou J, et al. Influenza-like illness in Lesotho from July 2020 to July 2021: population-based participatory surveillance results. JMIR Public Health Surveill 2024 Oct 8;10:e55208. [doi: 10.2196/55208] [Medline: 39378443]
21. Guerrisi C, Turbelin C, Souty C, et al. The potential value of crowdsourced surveillance systems in supplementing sentinel influenza networks: the case of France. Euro Surveill 2018 Jun;23(25):1700337. [doi: 10.2807/1560-7917.ES.2018.23.25.1700337] [Medline: 29945696]
22. Gertz A, Rader B, Sewalk K, Varrelman TJ, Smolinski M, Brownstein JS. Decreased seasonal influenza rates detected in a crowdsourced influenza-like illness surveillance system during the COVID-19 pandemic: prospective cohort study. JMIR Public Health Surveill 2023 Dec 28;9:e40216. [doi: 10.2196/40216] [Medline: 38153782]
23. Adler AJ, Eames KTD, Funk S, Edmunds WJ. Incidence and risk factors for influenza-like-illness in the UK: online surveillance using Flusurvey. BMC Infect Dis 2014 May 1;14:232. [doi: 10.1186/1471-2334-14-232] [Medline: 24885043]
24. Prieto JT, Jara JH, Alvis JP, et al. Will participatory syndromic surveillance work in Latin America? Piloting a mobile approach to crowdsource influenza-like illness data in Guatemala. JMIR Public Health Surveill 2017 Nov 14;3(4):e87. [doi: 10.2196/publichealth.8610] [Medline: 29138128]
25. Baltrusaitis K, Brownstein JS, Scarpino SV, et al. Comparison of crowd-sourced, electronic health records based, and traditional health-care based influenza-tracking systems at multiple spatial resolutions in the United States of America. BMC Infect Dis 2018 Aug 15;18(1):403. [doi: 10.1186/s12879-018-3322-3] [Medline: 30111305]
26. Kjelsø C, Galle M, Bang H, Ethelberg S, Krause TG. Influmeter—an online tool for self-reporting of influenza-like illness in Denmark. Infect Dis (Lond) 2016 Apr;48(4):322-327. [doi: 10.3109/23744235.2015.1122224] [Medline: 26654752]
27. Carlson SJ, Dalton CB, Butler MT, Fejsa J, Elvidge E, Durrheim DN. Flutracking weekly online community survey of influenza-like illness annual report 2011 and 2012. Commun Dis Intell Q Rep 2013 Dec 31;37(4):E398-E406. [doi: 10.33321/cdi.2013.37.51] [Medline: 24882237]
28. Richard A, Müller L, Wisniak A, et al. Grippenet: A new tool for the monitoring, risk-factor and vaccination coverage analysis of influenza-like illness in Switzerland. Vaccines (Basel) 2020 Jun 27;8(3):343. [doi: 10.3390/vaccines8030343] [Medline: 32605076]
29. Rehn M, Carnahan A, Merk H, et al. Evaluation of an internet-based monitoring system for influenza-like illness in Sweden. PLoS One 2014;9(5):e96740. [doi: 10.1371/journal.pone.0096740] [Medline: 24824806]
30. Bajaj S, Chen S, Creswell R, et al. COVID-19 testing and reporting behaviours in England across different sociodemographic groups: a population-based study using testing data and data from community prevalence surveillance surveys. Lancet Digit Health 2024 Nov;6(11):e778-e790. [doi: 10.1016/S2589-7500(24)00169-9] [Medline: 39455191]
31. Jiang L, Lee VJ, Lim WY, et al. Performance of case definitions for influenza surveillance. Euro Surveill 2015 Jun 4;20(22):21145. [doi: 10.2807/1560-7917.es2015.20.22.21145] [Medline: 26062645]
32. Loenenbach A, Lehfeld AS, Puetz P, et al. Participatory, virologic, and wastewater surveillance data to assess underestimation of COVID-19 incidence, Germany, 2020-2024. Emerg Infect Dis 2024 Sep;30(9):1939-1943. [doi: 10.3201/eid3009.240640] [Medline: 39174033]
33. Ahmed W, Simpson SL, Bertsch PM, et al. Minimizing errors in RT-PCR detection and quantification of SARS-CoV-2 RNA for wastewater surveillance. Sci Total Environ 2022 Jan 20;805:149877. [doi: 10.1016/j.scitotenv.2021.149877] [Medline: 34818780]
34. McClary-Gutierrez JS, Aanderud ZT, Al-faliti M, et al. Standardizing data reporting in the research community to enhance the utility of open data for SARS-CoV-2 wastewater surveillance. Environ Sci: Water Res Technol 2021 Aug 26;7(9):1545-1551. [doi: 10.1039/D1EW00235J]
35. Media factsheet: wastewater-based surveillance for COVID-19 and Zika. : National Environment Agency; 2024 URL: https://www.pub.gov.sg/-/media/PUB/Resources/Press-Releases/2024/04/LKYWP/For-website-upload-NEA-Media-Factsheet_Wastewater-Surveillance-FINAL.pdf [accessed 2025-05-27]
36. Wong JCC, Tan J, Lim YX, et al. Non-intrusive wastewater surveillance for monitoring of a residential building for COVID-19 cases. Sci Total Environ 2021 Sep 10;786:147419. [doi: 10.1016/j.scitotenv.2021.147419] [Medline: 33964781]
37. Lim VW, Lim RL, Tan YR, et al. Government trust, perceptions of COVID-19 and behaviour change: cohort surveys, Singapore. Bull World Health Organ 2021 Feb 1;99(2):92-101. [doi: 10.2471/BLT.20.269142] [Medline: 33551503]
38. Antigen rapid test kits vending machines deployed islandwide and updates to home recovery and travel classifications. Ministry of Health. URL: https://www.moh.gov.sg/newsroom/antigen-rapid-test-kits-vending-machines-deployed-islandwide-and-updates-to-home-recovery-and-travel-classifications_17sep2021/ [accessed 2026-05-15]
39. Resuming our transition towards COVID resilience. Ministry of Health. 2021. URL: https://www.moh.gov.sg/newsroom/resuming-our-transition-towards-covid-resilience/ [accessed 2025-02-27]
40. Muench P, Jochum S, Wenderoth V, et al. Development and validation of the Elecsys Anti-SARS-CoV-2 immunoassay as a highly specific tool for determining past exposure to SARS-CoV-2. J Clin Microbiol 2020 Sep 22;58(10):01694-01620. [doi: 10.1128/JCM.01694-20] [Medline: 32747400]
41. Lou B, Li TD, Zheng SF, et al. Serology characteristics of SARS-CoV-2 infection after exposure and post-symptom onset. Eur Respir J 2020 Aug;56(2):2000763. [doi: 10.1183/13993003.00763-2020] [Medline: 32430429]

XSL•FO
RenderX

42. Sutjipto S, Aung AH, Soon MML, et al. Plastic waste and COVID-19 incidence among hospital staff after deescalation in PPE use. JAMA Netw Open 2025 Apr 1;8(4):e255264. [doi: 10.1001/jamanetworkopen.2025.5264] [Medline: 40232716]
43. Mah T, Chae SR, Wang L, et al. Evaluating acute respiratory infection and influenza-like illness case definitions for community COVID-19 and influenza surveillance [poster]. Presented at: 12th International Conference on Options for the Control of Influenza; Sep 29 to Oct 2, 2024 URL: https://www.optionsxii2024.org.uk/programme [accessed 2026-05-29]
44. Enablers to support safe re-opening. Ministry of Health. 2020. URL: https://www.moh.gov.sg/newsroom/enablers-to-support-safe-re-opening/ [accessed 2025-05-27]
45. Further rationalisation of SMMs, boosting vaccinations, and updates to COVID-19 subsidies. Ministry of Health. 2022. URL: https://www.moh.gov.sg/newsroom/further-rationalisation-of-smms-boosting-vaccinations-and-updates-to-covid-19-subsidies/ [accessed 2025-05-27]
46. Soh AS, Ong BWL, Chen MI. Longitudinal serological assessment of SARS-cov-2 infection and medical ascertainment in a community cohort in singapore between year 2020–2023. Presented at: 19th Singapore Public Health & Occupational Medicine Conference; Oct 23-24, 2025 URL: https://www.optionsxii2024.org.uk/programme [accessed 2026-05-29]
47. Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Int J Surg 2014 Dec;12(12):1500-1524. [doi: 10.1016/j.ijsu.2014.07.014] [Medline: 25046751]
48. WHO COVID-19 dashboard. World Health Organization. 2023. URL: http://data.who.int/dashboards/covid19/summary [accessed 2025-08-15]
49. Lwin MO, Lu J, Sheldenkar A, et al. Effectiveness of a mobile-based influenza-like illness surveillance system (FluMob) among health care workers: longitudinal study. JMIR Mhealth Uhealth 2020 Dec 7;8(12):e19712. [doi: 10.2196/19712] [Medline: 33284126]
50. Lwin MO, Jayasundar K, Sheldenkar A, et al. Lessons from the implementation of Mo-Buzz, a mobile pandemic surveillance system for dengue. JMIR Public Health Surveill 2017 Oct 2;3(4):e65. [doi: 10.2196/publichealth.7376] [Medline: 28970191]
51. Lwin MO, Yung CF, Yap P, et al. FluMob: enabling surveillance of acute respiratory infections in health-care workers via mobile phones. Front Public Health 2017;5:49. [doi: 10.3389/fpubh.2017.00049] [Medline: 28367433]
52. Riley S, Atchison C, Ashby D, et al. REal-time Assessment of Community Transmission (REACT) of SARS-CoV-2 virus: study protocol. Wellcome Open Res 2021;5:200. [doi: 10.12688/wellcomeopenres.16228.2] [Medline: 33997297]
53. Venkatachalam I, Conceicao EP, Aung MK, et al. Healthcare workers as a sentinel surveillance population in the early phase of the COVID-19 pandemic. Singapore Med J 2022 Oct;63(10):577-584. [doi: 10.11622/smedj.2021083] [Medline: 34157807]
54. Tan CY, Zeng K, Cui L, Lin RTP, Chen M. Diagnostic performance of rapid antigen tests (RAT) for COVID-19 and factors associated with RAT-negative results among RT-PCR-positive individuals during Omicron BA.2, BA.5 and XBB.1 predominance. BMC Infect Dis 2024 May 21;24(1):504. [doi: 10.1186/s12879-024-09408-8] [Medline: 38769524]
55. Prasek SM, Pepper IL, Innes GK, et al. Variant-specific SARS-CoV-2 shedding rates in wastewater. Sci Total Environ 2023 Jan 20;857(Pt 1):159165. [doi: 10.1016/j.scitotenv.2022.159165] [Medline: 36195153]
56. Bitter LC, Kibbee R, Garant T, Örmeci B. Impact of wastewater characteristics and weather events on the N2 and N1 gene target ratios during wastewater surveillance of SARS-CoV-2 at five treatment plants and an upper sewershed location. Sci Total Environ 2025 Jun 15;981:179592. [doi: 10.1016/j.scitotenv.2025.179592] [Medline: 40347754]
57. Tan YR, Agrawal A, Matsoso MP, et al. A call for citizen science in pandemic preparedness and response: beyond data collection. BMJ Glob Health 2022 Jun;7(6):e009389. [doi: 10.1136/bmjgh-2022-009389] [Medline: 35760438]
58. Mao Y, Tan YR, Thein TL, et al. Identifying COVID-19 cases in outpatient settings. Epidemiol Infect 2021 Apr 5;149:e92. [doi: 10.1017/S0950268821000704] [Medline: 33814027]

Abbreviations

ART: antigen rapid test
SOCRATEs: Strengthening Our Community's Resilience Against Threats from Emerging Infections
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
WVI: wastewater viral-load index

XSL•FO
RenderX

Edited by A Mavragani; submitted 10.Oct.2025; peer-reviewed by D Yaman, Z Su; revised version received 17.Apr.2026; accepted 06.May.2026; published 10.Jun.2026.

Please cite as:
Soh SEAM, Aung AH, Ong WLB, Zeng K, Chavatte JM, Cui L, Lin RVTP, Lim VW, Lwin MO, Chen ICM
Harnessing Participatory Surveillance Cohorts and Proxy Indicators to Dynamically Track Epidemic Trends and Undiagnosed COVID-19 Infections in Singapore: Longitudinal Observational Study
J Med Internet Res 2026;28:e85608
URL: https://www.jmir.org/2026/1/e85608
doi:10.2196/85608

© Sheng En Alexius Matthias Soh, Aung Hein Aung, Wei Ling Brenda Ong, Kangwei Zeng, Jean-Marc Chavatte, Lin Cui, Raymond Valentine Tzer Pin Lin, Vanessa W Lim, May O Lwin, I-Cheng Mark Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

National Acceptance and Determinants of Immersive Extended Reality in Health Care in China: Cross-Sectional Study

Jiaying Li[1], PhD; Patricia M Davidson[2], PhD; Helen YL Chan[1], PhD; Cho Lee Wong[1], PhD; Xiang Qi[3], PhD; Zengjie Ye[4], PhD; Ankie Tan Cheung[1], PhD; Daniel YT Fong[5], PhD; Yibo Wu[6], PhD; Junxin Li[7], PhD

[1]The Nethersole School of Nursing, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

[2]University of New South Wales Sydney, Sydney, Australia

[3]Rory Meyers College of Nursing, New York University, New York, NY, United States

[4]School of Nursing, Guangzhou Medical University, Guangzhou, China

[5]School of Nursing, Li Ka Shing Faculty of Medicine, The University of Hong Kong, 5/F, HKUMed Academic Building, 3 Sassoon Road, Pokfulam, Hong Kong, China (Hong Kong)

[6]Department of Nursing, The Fourth Affiliated Hospital of School of Medicine, International School of Medicine, International Institutes of Medicine, Zhejiang University, Yiwu, China

[7]School of Nursing, Johns Hopkins University, Baltimore, MD, United States

Corresponding Author:
Daniel YT Fong, PhD
School of Nursing, Li Ka Shing Faculty of Medicine, The University of Hong Kong, 5/F, HKUMed Academic Building, 3 Sassoon Road, Pokfulam, Hong Kong, China (Hong Kong)

Abstract

Background: Immersive extended reality (XR) promises to transform health care, but public acceptance and user-side determinants of acceptance remain largely unknown.

Objective: This study aimed to estimate national immersive XR acceptance and identify sociodemographic, psychosocial, health, and digital determinants among adults in China.

Methods: This nationwide cross-sectional survey was conducted from June to September 2024 among 35,861 Chinese adults aged 18 years or older across 33 provincial-level regions and 800 communities, using multistage sampling. Immersive XR acceptance (0 - 100) and 139 potential predictors across demographic, adversity, personality, literacy, lifestyle, physical, and psychosocial domains were assessed. Poststratification weights were calibrated to the national age-sex distribution. Determinants were identified using survey-weighted hierarchical linear regression with Benjamini-Hochberg false discovery rate correction. Elastic net validation assessed predictor robustness, and classification and regression tree analysis identified profiles of likely nonacceptors.

Results: Mean acceptance was 63.11 (95% CI 62.75-63.46), varying by province (95% CI 47.9 - 72.3). Acceptance was highest in younger adults (women aged 18 - 24 years and men aged 30 - 34 years) and declined with age; sex differences were minimal. Acceptance varied across 15 chronic conditions (lowest: rare diseases, mean 54.24, 95% CI 45.40 - 63.07; highest: hyperlipidemia, mean 62.26, 95% CI 60.19-64.33). Acceptance was strongly associated with socioeconomic factors (higher social status: standardized β=0.17, 95% CI 0.16-0.18; higher youth socioeconomic status: standardized β=0.14, 95% CI 0.13-0.16; better youth economic environment: standardized β=0.06, 95% CI 0.04-0.07), digital capital (prior digital health use: standardized β=0.15, 95% CI 0.14-0.16; eHealth literacy: standardized β=0.09, 95% CI 0.07-0.10), and key traits (self-efficacy: standardized β=0.10, 95% CI 0.09-0.12; personal meaning: standardized β=0.06, 95% CI 0.04-0.07). Other positive predictors included having 2 types of medical insurance (standardized β=0.07, 95% CI 0.04-0.09), stable sleep duration (standardized β=0.07, 95% CI 0.03-0.10), and childhood psychological abuse (standardized β=0.07, 95% CI 0.05 - 0.09). Strong negative predictors included older age (standardized β=−0.08, 95% CI −0.10 to −0.06), couple-only household (standardized β=−0.08, 95% CI −0.11 to −0.04), childhood sexual abuse (standardized β=−0.08, 95% CI −0.10 to −0.06), attention-deficit/hyperactivity disorder (standardized β=−0.07, 95% CI −0.09 to −0.05), more siblings (standardized β=−0.07, 95% CI −0.08 to −0.05), childhood physical abuse (standardized β=−0.05, 95% CI −0.06 to −0.03), and collective violence exposure (standardized β=−0.05, 95% CI −0.07 to −0.03). The 6-node classification tree showed modest discrimination (test area under the curve=0.61; accuracy=0.681), high specificity (specificity=0.903), and low sensitivity (sensitivity=0.242), suggesting better identification of likely nonacceptors than likely acceptors.

Conclusions: Acceptance of immersive XR in health care in China was moderate but uneven. Adoption varied by age, region, socioeconomic resources, digital capital, psychosocial factors, household context, health status, and adversity exposure, suggesting that XR implementation is both a digital health innovation and a health equity challenge. Deployment should include targeted

XSL•FO
RenderX

education, accessible demonstrations, usability support, and trusted guidance for less accepting groups, especially older adults, socioeconomically disadvantaged groups, people with limited digital health experience, and psychosocially vulnerable populations.

(J Med Internet Res 2026;28:e88000) doi:10.2196/88000

KEYWORDS

immersive health care technologies; extended reality; technology acceptance; determinants; digital health

Introduction

Immersive extended reality (XR) technologies—encompassing virtual, augmented, and mixed reality (VR, AR, and MR)—are increasingly used in health care for rehabilitation, symptom management, procedural support, and training [1-3]. Recent evidence syntheses show that XR is now being tested across diverse clinical areas, including pain, neurorehabilitation, cancer-supportive care, mental health, older adult care, and health care education [4-8]. Although benefits have been reported for outcomes such as pain, anxiety, fatigue, and functional impairment, evidence quality and generalizability remain uneven [4-8]. With a market projected to exceed US $5 billion by 2030 (>25% annual growth) [1], and as clinical applications expand, the key implementation question is no longer solely whether immersive XR is clinically effective but also whether it is acceptable to intended users in real-world health care contexts. This issue is particularly important because immersive XR is not a low-friction digital tool. It often requires head-mounted devices, embodied interaction, sensory immersion, user confidence, and clinical or home-based support [9-11]. As a result, even clinically promising XR interventions may fail to scale if users are unwilling to adopt them.

Population-level evidence is needed to guide the real-world scale-up of immersive XR in health care. However, current evidence on XR acceptance remains fragmented and is drawn largely from small or highly specific cohorts, leaving broader population patterns unclear [12-16]. Existing evidence highlights stark contrasts between lay and professional perspectives, as well as across differing cultural and clinical contexts [12-16]. For instance, a gap exists even in high-income settings: general public awareness remains low and only 18.3% of UK adults recognize VR's clinical applications [17], while professional clinical integration also lags, with only 7.1% of 658 US rehabilitation specialists routinely using VR by 2025 [13]. Among patients, receptivity demonstrates similarly deep divisions: while 90% of patients in the United States who experience chronic pain express willingness to join a VR trial [14], Chinese oncology patients remain ambivalent [15], yet obese Chinese individuals show strong intent to try VR-based exergaming [16]. Although these studies provide useful evidence on XR acceptance in specific groups, their findings may not reflect broader public attitudes. In particular, they do not clarify how acceptance varies across geographic regions, age-sex groups, or people living with different noncommunicable diseases. This limits the ability of policymakers, clinicians, and technology developers to identify which populations are most likely to accept or resist immersive XR in health care.

Beyond estimating overall acceptance, understanding the determinants of immersive XR acceptance is essential for designing accessible interventions and guiding equitable implementation. However, current evidence remains fragmented [9,18-24]. Existing evidence suggests that acceptance may be lower among older adults [18,19], higher among women [20], and lower in groups with less education or minority status [21,22]. Psychological factors also appear relevant: openness and technology self-efficacy have been associated with greater acceptance [9,23], whereas technology anxiety may deter use [24]. Yet these studies usually examine a narrow set of determinants and rarely assess broader upstream factors that may shape digital health acceptance before users encounter a specific XR system. While established frameworks such as the Technology Acceptance Model and the Unified Theory of Acceptance and Use of Technology effectively capture immediate, system-specific interactions [25], they are less suited to capturing wider biopsychosocial and equity-related characteristics, such as social resources, prior technology exposure, health literacy, perceived health needs, lifestyle, psychological vulnerability, and adverse life experiences. A broader biopsychosocial lens is therefore needed to understand not only whether the public is willing to use immersive XR in health care but also which population groups may be left behind during implementation.

To address the limitations of prior small, condition-specific, and nonnational studies, we conducted what is, to our knowledge, the largest national survey of health care XR acceptance to date. Our study is distinguished by its national scope; assessment of 139 determinants across socioeconomic, psychosocial, experiential, and health-related domains; equity-oriented analysis; and development of a practical tool to identify likely nonacceptors. Specifically, our main aims were to (1) estimate national levels of immersive XR acceptance and map geographic, age-sex, and noncommunicable disease-specific patterns; (2) evaluate 139 candidate determinants spanning socioeconomics, childhood adversity, lifestyle, health literacy, and psychosocial factors; and (3) develop and validate a streamlined classification tree to flag profiles of immersive XR nonacceptors.

We hypothesized that immersive XR acceptance varies significantly by geography, age, and disease cohort, and those biopsychosocial factors independently drive nonacceptance, converging into distinct, classifiable user profiles.

Methods

Study Design

This study was a nationwide, population-based, and cross-sectional survey conducted in China between June and September 2024 as part of the Psychology and Behavior Investigation of Chinese Residents program.

XSL•FO
RenderX

Study Setting

The study was conducted across 33 provincial regions in China as part of a national psychological and behavioral health program. Data collection took place between June 23 and September 29, 2024, at the selected survey sites.

Participants

Inclusion and Exclusion

Eligible participants were Chinese nationals who were aged 18 years or older, resided in China for at least 11 months per year, were able to complete the online questionnaire, and provided informed consent. We excluded individuals with psychosis, cognitive impairment, or those with prior participation in similar surveys.

Participant Characteristics

The analytic sample had a mean age of 47.50 (SD 16.74) years; 43.0% (15,428/35,861) were male, 56.5% (20,275/35,861) were female, and 0.4% (158/35,861) identified as other sex.

Sampling Procedures

To improve national coverage of Chinese adults, we used a 5-stage geographic sampling framework across 33 provincial-level administrative regions. The sampling process combined multistage area selection with quota-based recruitment at the final stage. Specifically, 150 cities, 202 districts or counties, 390 townships or streets, and 800 communities or villages were selected through multiple geographic stages, after which participant recruitment within selected communities or villages was guided by demographic quota control. Poststratification raking weights were then applied to align the sample with the national age- and sex-specific distribution.

Sample Size, Power, and Precision

Of 38,793 questionnaires distributed, 38,424 (99.0%) were returned; after exclusions, 36,240 valid cases remained. A final 2-investigator audit yielded 35,861 records for analysis. Given this large national sample, the study was expected to provide stable estimates and reasonably narrow CIs for the primary descriptive and regression analyses. Only 4 of 139 predictors had missing data (range 1.54% - 15.30%); the outcome and all other predictors were complete. We therefore used complete-case analysis and did not perform multiple imputation.

Measures and Covariates

Primary Outcome

The primary outcome was self-rated acceptance of immersive health care technologies, measured on a 0 (completely unacceptable) to 100 (completely acceptable) scale. Acceptance was defined as participants' willingness to use immersive XR (VR, AR, or MR) in health care. Participants were provided with standardized definitions of each modality to ensure comprehension.

Exposures, Predictors, and Covariates

We assessed 139 candidate variables as potential exposures, predictors, and covariates across 7 prespecified domains. To improve readability, the main text summarizes each domain only; full instrument names, acronyms, scoring ranges, item details, and reliability metrics are provided in Table S1 in Multimedia Appendix 1.

Specifically, block 1 (demographics and socioeconomics) included 26 variables: age, sex, gender identity, sexual orientation, handedness, race, religion, education, occupation, income, city tier, Hukou registration, perceived social status, relationship status, family structure, living alone, siblings, dwelling area, number of rooms, assets, debt, urban or rural status, years in community, insurance coverage, insurance enrollment site, and difficulty paying medical bills. Block 2 (life adversity and stress) included 30 variables covering 14 past year stressors, 15 adverse childhood experiences, and perceived childhood socioeconomic environment. Block 3 (personality and self-efficacy) included 7 variables: extraversion, agreeableness, conscientiousness, neuroticism, openness, general self-efficacy, and narcissistic admiration or rivalry. Block 4 (health literacy and empowerment) included 15 variables: family-neighbor relationships, presence of a confidant, receipt of tangible help, perceived social support, social isolation, loneliness, social connection, family health, family communication, eHealth literacy, general health literacy, antibiotic knowledge, implicit health beliefs, eating self-regulation, and prior digital health use. Block 5 (lifestyle behaviors) included 14 variables: tobacco use, alcohol consumption, physical activity, chronotype, sleep duration, sleep difficulty, snoring, daytime sleepiness, sleep stability, salt-adding habits, and engagement in music, visual arts, dance, and other arts. Block 6 (physical health and exposures) included 37 variables: BMI; 14 physician-diagnosed chronic diseases; 14 types of injuries in the past 14 months; vaccination status for human papillomavirus, influenza, shingles, hepatitis B virus, and COVID-19; number of COVID-19 infections; no vaccinations; and overall health status. Block 7 (mental and psychosocial health) included 10 variables: depressive symptoms, anxiety, perceived stress, attention-deficit/hyperactivity disorder traits, work burnout, cyberchondria, resilience, maladaptive daydreaming, social media addiction, and meaning in life.

Data Collection and Quality of Measurements

Between June 23 and September 29, 2024, trained surveyors visited selected sites, confirmed eligibility, obtained e-consent, and administered an electronic questionnaire via Wenjuanxing. Responses were uploaded to a secure server and subjected to weekly supervisor review and automated quality control, checking for logic errors, duplicates, >20% missing data, or completion times <300 seconds.

Data Analysis

Analyses comprised three stages: (1) descriptive summaries using poststratification raking weights, (2) multivariable modeling via 7-block hierarchical regression and elastic net validation, and (3) development of a classification and regression tree. Analyses used R (version 4.1.1; R Core Team), with 2-sided $P<.05$.

First, poststratification raking weights aligned the sample to 2020 census age-sex margins. We produced unweighted and

XSL•FO
RenderX

weighted summaries (continuous: mean, SD; categorical: n [%]) and plotted weighted acceptance with 95% CIs by province, age-sex strata, single-year age (locally estimated scatterplot smoothing—smoothed by sex), and chronic condition. Province-level estimates were presented as descriptive summaries of geographic variation and were not intended to model province-level contextual effects.

Second, we fit survey-weighted hierarchical linear regression in 7 blocks—sociodemographics, life adversity and stress, personality, literacy and empowerment, lifestyle, physical health, and mental health—ordering blocks by presumed causal order to quantify independent effects while avoiding overadjustment. This hierarchical approach was used to examine whether each predictor domain explained additional variance in XR acceptance beyond domains entered earlier. For block *k*, we fitted a weighted linear model of acceptance on retained predictors from blocks 1 to *k*, applying Benjamini-Hochberg false discovery rate (FDR) correction within each block to address multiple testing, and carrying forward only predictors that remained significant after FDR correction. Continuous variables entered in original units; categorical variables were dummy-coded. Variance inflation factor >5 flagged multicollinearity. All candidate predictors were then weighted, z-standardized, and entered into a 10-fold cross-validated elastic net (α=.50) with optimal penalty λ_{min} selected by minimum mean squared error, which was used as a complementary data-driven approach because the study included a large number of potentially correlated predictors; this allowed us to assess the robustness of findings from the hierarchical regression and identify a more parsimonious set of predictors. We quantified agreement between elastic net and hierarchical selection via overlap statistics. For reporting, we prespecified the smallest effect size of interest for standardized β coefficients (β_std). Predictors with an absolute value |β_std| ≥0.03 were reported by name in the main text. Other FDR-significant predictors are detailed in main text tables and figures; nonsignificant predictors are tabulated in the supplement.

Third, to build a fast classification tool, we dichotomized acceptance at ≥ 80 (“acceptors”) versus <80 (“nonacceptors”) and randomly split the weighted sample 70/30 (outcome-stratified) into training and test sets. We constructed a Gini-based classification and regression tree on the training data, selecting the complexity parameter by the elbow method. For classification performance, acceptor status (score ≥80) was defined as the positive class; therefore, sensitivity represents the proportion of acceptors correctly classified and specificity represents the proportion of nonacceptors correctly classified. In the test set, we assessed discrimination using the survey-weighted receiver operating characteristic curve and its area under the curve with 1000 bootstrap CIs, and reported accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and balanced accuracy at a 0.50 cutoff.

Ethical Considerations

This study was approved by the Health-Economics and Policy Research (approval number NHC-HEPR202401) and the Shanghai Jiao Tong University Ethics Committee (approval number H20240237I). Written informed consent was obtained from all participants prior to their participation. To protect privacy and confidentiality, all study data were fully anonymized prior to analysis. Participants were not compensated for their participation. Finally, no individual participants or users can be identified in any of the images or figures included in this manuscript or its supplementary materials.

Results

Characteristics of Respondents

Among the 35,861 weighted respondents (Figure 1), mean age was 47.5 (SD 6.7) years; 50.4% (18,084/35,861) were male, 93.1% (33,381/35,861) were of Han ethnicity, 82.0% (29,409/35,861) had at least secondary education, and 72.5% (26,005/35,861) lived in urban areas. Complete sociodemographic, economic, residential, and health characteristics are shown in Table 1.

XSL•FO
RenderX

Figure 1. Flow diagram of data screening and participant selection in this nationwide cross-sectional survey of immersive extended reality acceptance in health care among adults in China (June-September 2024).

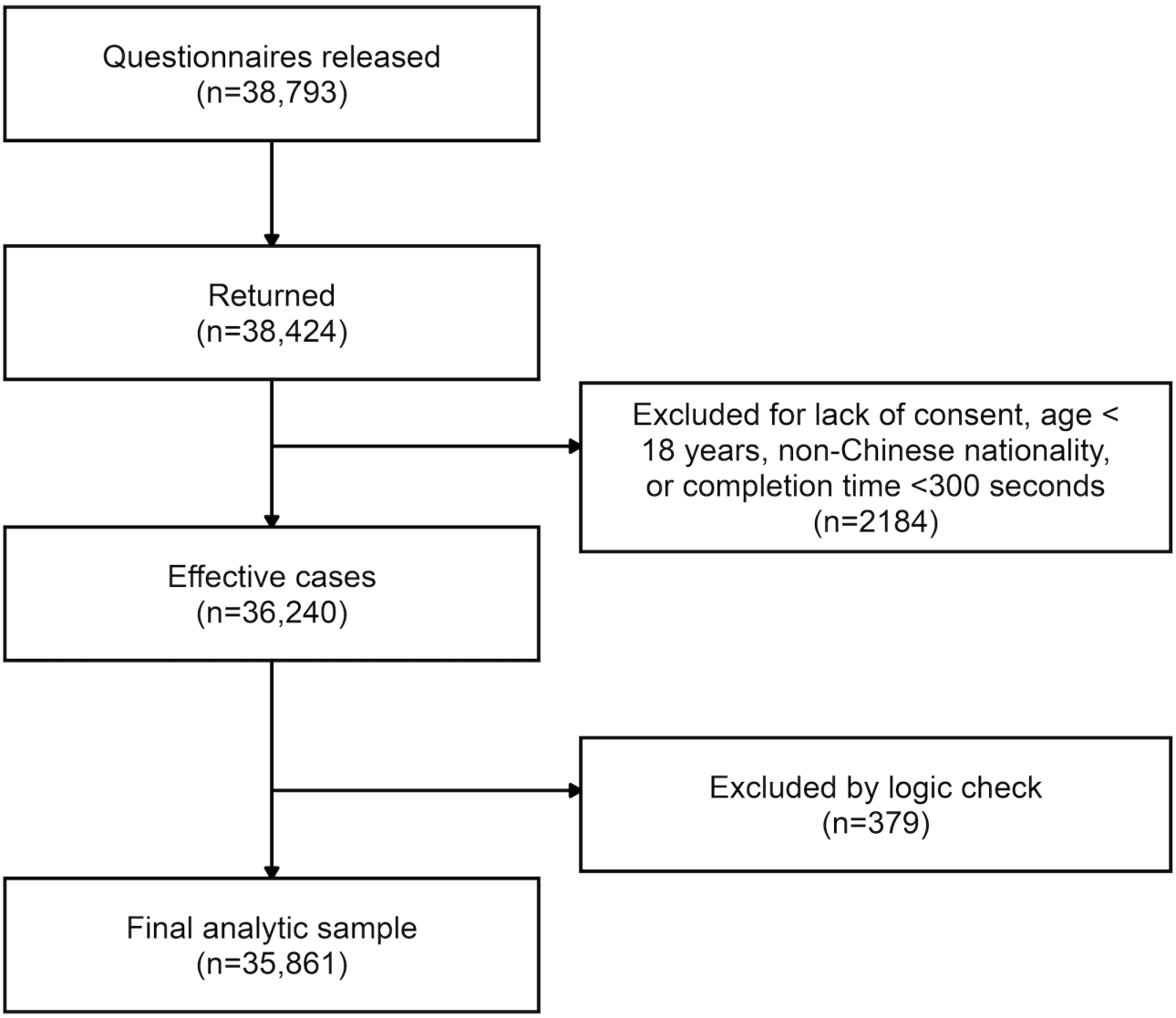

XSL•FO
RenderX

Table . Characteristics of participants included in this nationwide cross-sectional survey of immersive extended reality acceptance in health care in China, presented before and after weighting, June to September 2024 (N=35,861).

Characteristics	Unweighted (N=35,861)	Weighted (N=35,861)
Age (years), mean (SD)	34.88 (16.48)	47.50 (16.74)
Sex, n (%)		
Male	15,428 (43.0)	18,084 (50.4)
Female	20,275 (56.5)	17,639 (49.2)
Other	158 (0.4)	138 (0.4)
Race/ethnicity, n (%)		
Han ethnicity	32,228 (89.9)	33,381 (93.1)
Other	3633 (10.1)	2480 (6.9)
Religion (any), n (%)		
No	32,635 (91.0)	32,134 (89.6)
Yes	3226 (9.0)	3727 (10.4)
Education, n (%)		
Primary or below	3211 (9.0)	6452 (18.0)
Secondary	11,189 (31.2)	13,182 (36.8)
Undergraduate	19,862 (55.4)	14,416 (40.2)
Graduate	1599 (4.5)	1812 (5.1)
Employment status, n (%)		
Employed	10,449 (29.1)	13,889 (38.7)
Retired	2922 (8.1)	6677 (18.6)
Self-employed	5781 (16.1)	8615 (24.0)
Student	14,870 (41.5)	3735 (10.4)
Unemployed	1839 (5.1)	2945 (8.2)
Per capita monthly income (in Chinese yuan; CNY ¥1=US $0.148 as of June 10, 2026), n (%)		
≤1000	2068 (5.8)	1888 (5.3)
1001 - 2000	3053 (8.5)	2946 (8.2)
2001 - 3000	4461 (12.4)	4601 (12.8)
3001 - 4000	5362 (15.0)	5547 (15.5)
4001 - 5000	4789 (13.4)	4974 (13.9)
5001 - 6000	4601 (12.8)	4689 (13.1)
6001 - 9000	4591 (12.8)	4638 (12.9)
9001 - 12,000	2945 (8.2)	2809 (7.8)
12,001 - 15,000	1693 (4.7)	1610 (4.5)
≥15,001	2298 (6.4)	2160 (6.0)
City tier (by Hukou location), n (%)		
First-tier city	2052 (5.7)	2620 (7.3)
New first-tier city	4645 (13.0)	5190 (14.5)
Other city tiers	29,164 (81.3)	28,051 (78.2)
Residence in last 3 months, n (%)		
Urban	27,062 (75.5)	26,005 (72.5)
Rural	8799 (24.5)	9856 (27.5)
Household registration (Hukou), n (%)		

XSL•FO
RenderX

Characteristics	Unweighted (N=35,861)	Weighted (N=35,861)
Urban	16,728 (46.6)	17,883 (49.9)
Rural	19,133 (53.4)	17,978 (50.1)
Medical insurance, n (%)		
None	2344 (6.5)	1661 (4.6)
Government-funded only	2779 (7.7)	2336 (6.5)
Urban resident only	14,742 (41.1)	13,021 (36.3)
Employee only	5528 (15.4)	7604 (21.2)
New rural cooperative only	3870 (10.8)	4874 (13.6)
Commercial only	547 (1.5)	475 (1.3)
2 types	4858 (13.5)	4799 (13.4)
>2 types	1193 (3.3)	1092 (3.0)
Diagnosed hypertension, n (%)		
No	33,045 (92.1)	30,174 (84.1)
Yes	2816 (7.9)	5687 (15.9)
Diagnosed diabetes, n (%)		
No	34,868 (97.2)	33,871 (94.4)
Yes	993 (2.8)	1990 (5.6)
Diagnosed hyperlipidemia, n (%)		
No	35,070 (97.8)	34,352 (95.8)
Yes	791 (2.2)	1509 (4.2)
Diagnosed coronary heart disease, n (%)		
No	35,408 (98.7)	34,923 (97.4)
Yes	453 (1.3)	939 (2.6)
Diagnosed stroke, n (%)		
No	35,751 (99.7)	35,659 (99.4)
Yes	110 (0.3)	202 (0.6)
Diagnosed respiratory disease, n (%)		
No	35,350 (98.6)	35,092 (97.9)
Yes	511 (1.4)	770 (2.1)
Diagnosed urinary disease, n (%)		
No	35,608 (99.3)	35,458 (98.9)
Yes	253 (0.7)	403 (1.1)
Diagnosed digestive disease, n (%)		
No	35,047 (97.7)	34,767 (96.9)
Yes	814 (2.3)	1094 (3.1)
Diagnosed osteoporosis, n (%)		
No	35,063 (97.8)	34,270 (95.6)
Yes	798 (2.2)	1591 (4.4)
Diagnosed arthritis, n (%)		
No	34,801 (97.0)	33,917 (94.6)
Yes	1060 (3.0)	1944 (5.4)
Diagnosed tumor, n (%)		
No	35,665 (99.5)	35,569 (99.2)

XSL•FO
RenderX

Characteristics	Unweighted (N=35,861)	Weighted (N=35,861)
Yes	196 (0.5)	292 (0.8)
Diagnosed rare disease, n (%)		
No	35,792 (99.8)	35,790 (99.8)
Yes	69 (0.2)	71 (0.2)
Diagnosed obesity, n (%)		
No	29,153 (82.6)	27,642 (78.5)
Yes	6156 (17.4)	7585 (21.5)
Diagnosed anxiety, n (%)		
No	29,732 (82.9)	30,702 (85.6)
Yes	6129 (17.1)	5160 (14.4)
Diagnosed depressive, n (%)		
No	27,669 (77.2)	28,970 (80.8)
Yes	8192 (22.8)	6891 (19.2)
Social status score, mean (SD)	4.00 (1.35)	4.11 (1.38)

Weighted Acceptance by Region, Age-Sex, and Chronic Condition

Overall mean weighted acceptance was 63.11 (95% CI 62.75 - 63.46). Provincial means varied widely from 55.96 (95% CI 53.07-58.85) in Guizhou province to 72.34 (95% CI 66.28-78.39) in Jilin province. First-tier cities scored 63.01 (95% CI 60.92-65.10) in Beijing, 66.15 (95% CI 63.95-68.35) in Guangdong province, and 67.38 (95% CI 65.43-69.33) in Shanghai (Figure 2; Table S2 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 2. Province-level weighted mean acceptance of immersive extended reality in health care among adults in China in a nationwide cross-sectional survey conducted between June and September 2024 (N=35,861).

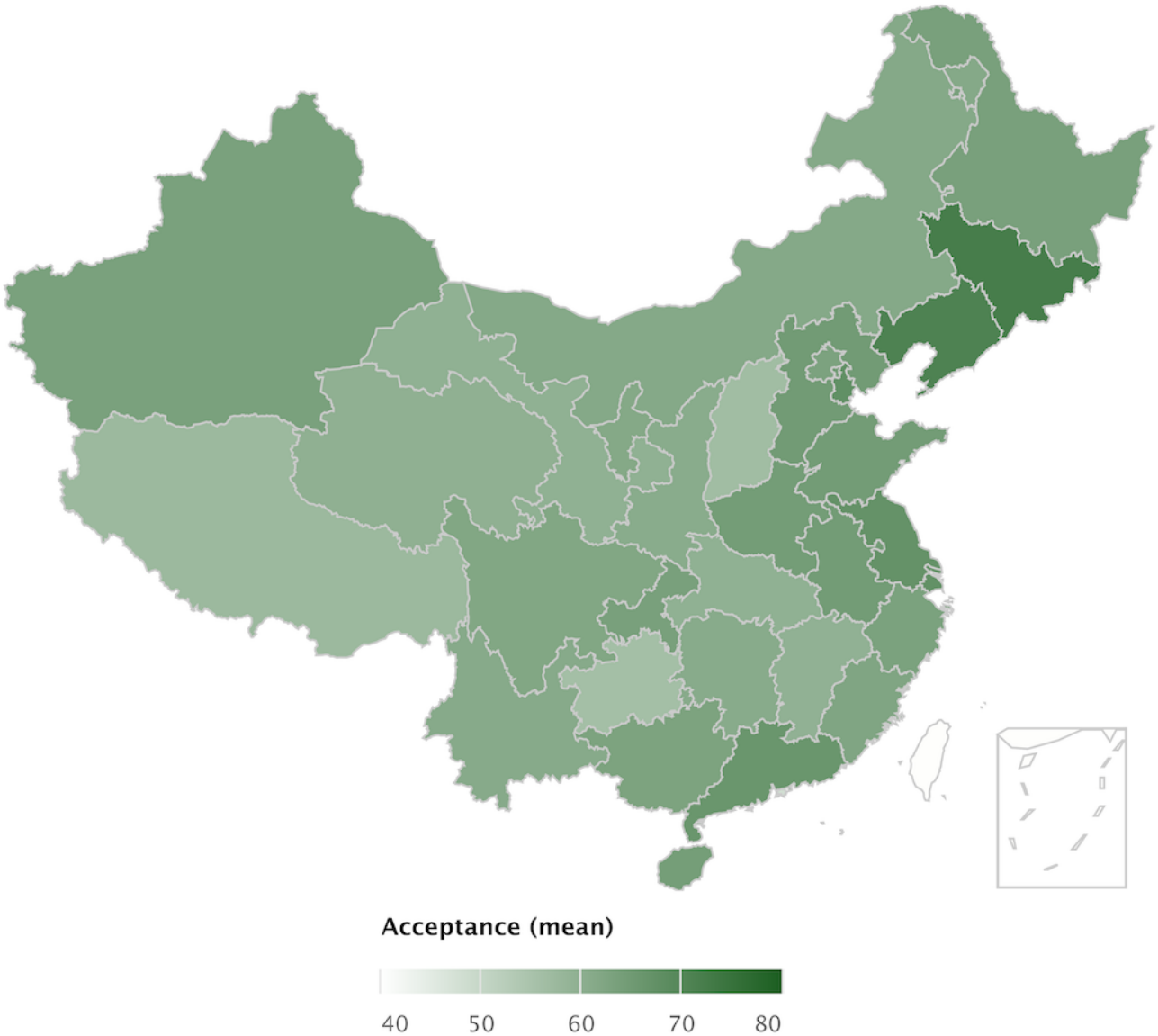

Acceptance varied by age, peaking at 18 - 24 years for women (mean 68.51, 95% CI 68.05 - 68.96) and 30 - 34 years for men (mean 66.44, 95% CI 65.09 - 67.79), with a general decline thereafter. Locally estimated scatterplot smoothing curves confirmed minimal sex differences (<3 points) (Figure 3A and B; Table S3 in Multimedia Appendix 1).

Individuals with any chronic condition had lower acceptance (mean 61.15, 95% CI 60.63 - 61.67) than the overall sample (mean 63.11, 95% CI 62.75 - 63.46). Scores ranged from rare diseases (mean 54.24, 95% CI 45.40 - 63.07) to hyperlipidemia (mean 62.26, 95% CI 60.19 - 64.33). Other common conditions showed relatively similar acceptance: obesity (mean 61.90, 95% CI 61.10 - 62.70), hypertension (mean 60.31, 95% CI 59.23 - 61.39), diabetes (mean 59.59, 95% CI 57.70 - 61.48), anxiety (mean 59.32, 95% CI 58.40 - 60.23), and depression (mean 59.37, 95% CI 58.58 - 60.16) (Figure 3C; Table S4 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 3. Weighted mean acceptance of immersive extended reality in health care among adults in China, shown by (A) sex and age group, (B) locally estimated scatterplot smoothing of single-year age curves stratified by sex, and (C) chronic condition, from a nationwide cross-sectional survey conducted between June and September 2024 (N=35,861). The pink bar in Figure 3C represents the mean and CI for the subgroup with at least 1 chronic condition.

XSL•FO
RenderX

Predictors of Acceptance

Overview

Of 139 candidate predictors, 76 were FDR-significant. The elastic net model showed high concordance, retaining 70 of these 76 (92.1%) significant predictors. The variance inflation factor for 7 regression models ranged from 1.01 to 2.23. Fully adjusted β_std, 95% CIs, *P* values, and adjusted *P* values for FDR-significant predictors are presented in Figure 4 and Table 2. Figure 4A-G refer to significant predictors from (A) demographics and socioeconomics, (B) life adversity and stress, (C) personality, (D) health literacy and empowerment, (E) lifestyle behaviors, (F) physical health and exposures, and (G) mental health and psychosocial factors. We summarize only the strongest predictors (prespecified |β_std|≥0.03) in this section. Model statistics for nonsignificant predictors are presented in Tables S5 - S11 in Multimedia Appendix 1. Although the results are reported according to the 7 prespecified analytic blocks, these blocks were also used to inform the 4 broader domains synthesized in the *Discussion* section: blocks 1 and 2 were grouped as structural-environmental factors, blocks 3 and 7 as trait-resilience factors, block 4 as social-digital capital, and blocks 5 and 6 as behavioral-health factors.

Figure 4. Significant predictors of immersive extended reality acceptance in health care among adults in China across 7 domains: (A) demographics and socioeconomics, (B) life adversity and stress, (C) personality, (D) health literacy and empowerment, (E) lifestyle behaviors, (F) physical health and exposures, and (G) mental health and psychosocial factors from hierarchical weighted linear regression models (N=35,861). ACE: adverse childhood experience; ADHD: attention-deficit/hyperactivity disorder; ASRS-6: 6-item Adult ADHD Self-Report Scale; BSMAS: Bergen Social Media Addiction Scale; CCBI-7: 7-item Chinese version of Copenhagen Burnout Inventory; FCS-SF: Family Communication Scale–Short Form; MET: metabolic equivalent of task; NARQ: Narcissistic Admiration and Rivalry Questionnaire; NGSES: New General Self-Efficacy Scale; PHQ-9: 9-item Patient Health Questionnaire; PSSS: Perceived Social Support Scale; SCS: Short Cyberchondria Scale; SREBQ: Self-Regulation Eating Behavior Questionnaire.

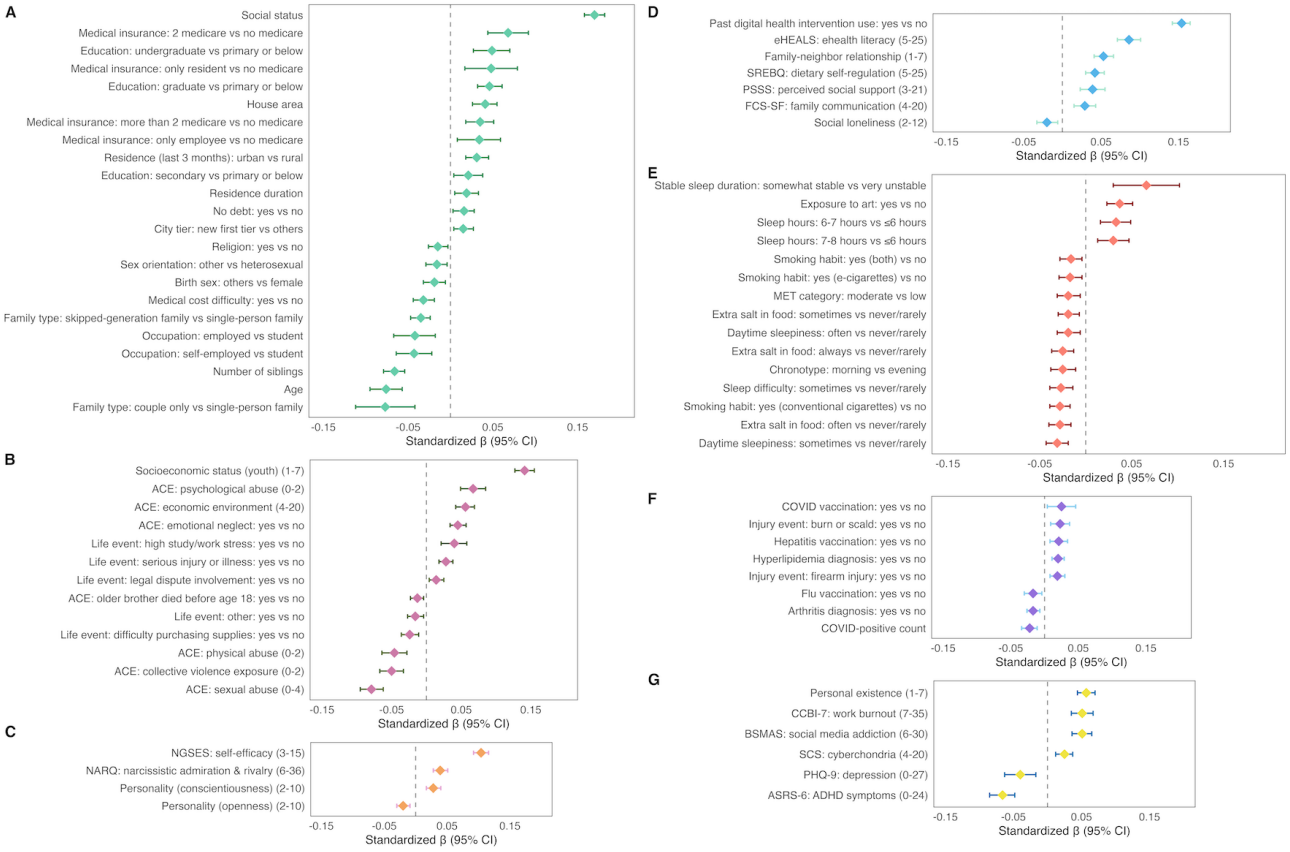

XSL•FO
RenderX

Table . Significant predictors of immersive extended reality acceptance in health care among adults in China across 7 sequential blocks from hierarchical weighted linear regression models in a nationwide cross-sectional survey, June to September 2024 (N=35,861).

Predictors	Standardized β coefficient (95% CI)	*P* value	Adjusted *P* value
Block 1: Demographics and socioeconomics			
Social status	0.17 (0.16 to 0.18)	<.001	<.001
Medical insurance: 2 Medicare vs no Medicare	0.07 (0.04 to 0.09)	<.001	<.001
Education: graduate vs primary or below	0.05 (0.03 to 0.06)	<.001	<.001
Education: undergraduate vs primary or below[a]	0.05 (0.03 to 0.07)	<.001	<.001
Medical insurance: only resident vs no Medicare[a]	0.05 (0.02 to 0.08)	.003	.01
House area	0.04 (0.03 to 0.06)	<.001	<.001
Medical insurance: more than 2 Medicare vs no Medicare	0.04 (0.02 to 0.05)	<.001	<.001
Medical insurance: only employee vs no Medicare	0.03 (0.01 to 0.06)	.01	.03
Residence (last 3 months): urban vs rural	0.03 (0.02 to 0.04)	<.001	<.001
Education: secondary vs primary or below	0.02 (0.00 to 0.04)	.02	.045
No debt: yes vs no[a]	0.02 (0.00 to 0.03)	.02	.04
Residence duration	0.02 (0.00 to 0.03)	.01	.03
City tier: new first tier vs others	0.01 (0.00 to 0.03)	.01	.03
Religion: yes vs no	−0.01 (−0.03 to −0.00)	.01	.03
Birth sex: others vs female[a]	−0.02 (−0.03 to −0.01)	.005	.02
Sex orientation: other vs heterosexual	−0.02 (−0.03 to −0.00)	.009	.03
Medical cost difficulty: yes vs no	−0.03 (−0.04 to −0.02)	<.001	<.001
Family type: skipped-generation family vs single-person family	−0.04 (−0.05 to −0.02)	<.001	<.001
Occupation: employed vs student	−0.04 (−0.07 to −0.02)	.001	.003
Occupation: self-employed vs student	−0.04 (−0.06 to −0.02)	<.001	<.001
Number of siblings	−0.07 (−0.08 to −0.05)	<.001	<.001
Age (years)[a]	−0.08 (−0.10 to −0.06)	<.001	<.001
Family type: couple only vs single-person family	−0.08 (−0.11 to −0.04)	<.001	<.001
Block 2: Life adversity and stress			
Socioeconomic status (youth) (1-7)	0.14 (0.13 to 0.16)	<.001	<.001
ACE[b]: psychological abuse (0 - 2)	0.07 (0.05 to 0.09)	<.001	<.001
ACE: economic environment (4-20)[a]	0.06 (0.04 to 0.07)	<.001	<.001
ACE: emotional neglect: yes vs no	0.04 (0.03 to 0.06)	<.001	<.001

XSL•FO
RenderX

Predictors	Standardized β coefficient (95% CI)	*P* value	Adjusted *P* value
Life event: high study or work stress: yes vs no	0.04 (0.02 to 0.06)	<.001	<.001
Life event: serious injury or illness: yes vs no	0.03 (0.02 to 0.04)	<.001	<.001
Life event: legal dispute involvement: yes vs no	0.01 (0.00 to 0.03)	.009	.02
ACE: older brother died before 18 years of age: yes vs no	−0.01 (−0.02 to −0.00)	.007	.02
Life event: difficulty purchasing supplies: yes vs no	−0.02 (−0.04 to −0.01)	<.001	.001
Life event: other: yes vs no	−0.02 (−0.03 to −0.00)	.007	.02
ACE: collective violence exposure (0 - 2)	−0.05 (-0.07 to −0.03)	<.001	<.001
ACE: physical abuse (0 - 2)	−0.05 (−0.06 to −0.03)	<.001	<.001
ACE: sexual abuse (0 - 4)	−0.08 (−0.10 to −0.06)	<.001	<.001
Block 3: Personality and self-efficacy			
NGSES[c]: self-efficacy (3-15)	0.10 (0.09 to 0.12)	<.001	<.001
NARQ[d]: narcissistic admiration and rivalry (6-36)	0.04 (0.03 to 0.05)	<.001	<.001
Personality (conscientiousness) (2-10)	0.03 (0.02 to 0.04)	<.001	<.001
Personality (openness) (2-10)	−0.02 (−0.03 to −0.01)	<.001	.001
Block 4: Literacy and health empowerment			
Past digital health intervention use: yes vs no	0.15 (0.14 to 0.16)	<.001	<.001
eHEALS[e]: eHealth literacy (5-25)	0.09 (0.07 to 0.10)	<.001	<.001
Family-neighbor relationship (1-7)	0.05 (0.04 to 0.07)	<.001	<.001
PSSS[f]: perceived social support (3-21)	0.04 (0.02 to 0.06)	<.001	<.001
SREBQ[g]: dietary self-regulation (5-25)	0.04 (0.03 to 0.05)	<.001	<.001
FCS-SF[h]: family communication (4-20)	0.03 (0.01 to 0.04)	<.001	<.001
Social loneliness (2-12)	−0.02 (−0.03 to −0.01)	.005	.01
Block 5: Lifestyle behaviors			
Stable sleep duration: somewhat stable vs very unstable	0.07 (0.03 to 0.10)	<.001	.001
Exposure to art: yes vs no	0.04 (0.02 to 0.05)	<.001	<.001
Sleep hours: 6 - 7 hours vs ≤6 hours	0.03 (0.02 to 0.05)	<.001	<.001
Sleep hours: 7 - 8 hours vs ≤6 hours	0.03 (0.01 to 0.05)	.001	.002
Daytime sleepiness: often vs never or rarely	−0.02 (−0.03 to −0.01)	.004	.01
Extra salt in food: sometimes vs never or rarely	−0.02 (−0.03 to −0.01)	.001	.003
MET[i] category: moderate vs low	−0.02 (−0.03 to −0.01)	.004	.01

XSL•FO
RenderX

Predictors	Standardized β coefficient (95% CI)	*P* value	Adjusted *P* value
Smoking habit: yes (both) vs no	−0.02 (−0.03 to −0.00)	.01	.02
Smoking habit: yes (e-cigarettes) vs no	−0.02 (−0.03 to −0.00)	.008	.02
Chronotype: morning vs evening	−0.03 (−0.04 to −0.01)	<.001	.001
Daytime sleepiness: sometimes vs never or rarely	−0.03 (−0.04 to −0.02)	<.001	<.001
Extra salt in food: always vs never or rarely	−0.03 (−0.04 to −0.01)	<.001	<.001
Extra salt in food: often vs never or rarely	−0.03 (−0.04 to −0.02)	<.001	<.001
Sleep difficulty: sometimes vs never or rarely	−0.03 (−0.04 to −0.01)	<.001	<.001
Smoking habit: yes (conventional cigarettes) vs no	−0.03 (−0.04 to −0.02)	<.001	<.001
Block 6: Physical health and exposures			
COVID-19 vaccination: yes vs no	0.03 (0.00 to 0.05)	.02	.047
Hepatitis vaccination: yes vs no	0.02 (0.01 to 0.03)	.001	.004
Hyperlipidemia diagnosis: yes vs no	0.02 (0.01 to 0.03)	<.001	<.001
Injury event: burn or scald: yes vs no	0.02 (0.01 to 0.04)	.001	.004
Injury event: firearm injury: yes vs no	0.02 (0.01 to 0.03)	.001	.003
Arthritis diagnosis: yes vs no	−0.02 (−0.03 to −0.01)	.001	.003
COVID-19–positive count	−0.02 (−0.03 to −0.01)	<.001	.001
Flu vaccination: yes vs no	−0.02 (−0.03 to −0.00)	.01	.03
Block 7: Mental health and psychosocial			
Personal existence (1-7)	0.06 (0.04 to 0.07)	<.001	<.001
BSMAS[j]: social media addiction (6-30)	0.05 (0.04 to 0.06)	<.001	<.001
CCBI-7[k]: work burnout (7-35)	0.05 (0.04 to 0.07)	<.001	<.001
SCS[l]: cyberchondria (4-20)	0.03 (0.01 to 0.04)	<.001	.001
PHQ-9[m]: depression (0 - 27)	−0.04 (−0.06 to −0.02)	.001	.002

XSL•FO
RenderX

Predictors	Standardized β coefficient (95% CI)	*P* value	Adjusted *P* value
ASRS-6[n]: ADHD[o] symptoms (0 - 24)	−0.07 (−0.09 to −0.05)	<.001	<.001

[a]Variables not retained by the elastic net validation.
[b]ACE: adverse childhood experience.
[c]NGSES: New General Self-Efficacy Scale.
[d]NARQ: Narcissistic Admiration and Rivalry Questionnaire.
[e]eHEALS: eHealth Literacy Scale.
[f]PSSS: Perceived Social Support Scale.
[g]SREBQ: Self-Regulation Eating Behavior Questionnaire.
[h]FCS-SF: Family Communication Scale–Short Form.
[i]MET: metabolic equivalent of task.
[j]BSMAS: Bergen Social Media Addiction Scale.
[k]CCBI-7: 7-item Chinese version of Copenhagen Burnout Inventory.
[l]SCS: Short Cyberchondria Scale.
[m]PHQ-9: 9-item Patient Health Questionnaire.
[n]ASRS-6: 6-item Adult ADHD Self-Report Scale.
[o]ADHD: attention-deficit/hyperactivity disorder.

Demographics and Socioeconomics

The largest positive predictor was higher social status (β_std=0.17, 95% CI 0.16-0.18). Other positive factors included having “2 Medicare,” “only resident” Medicare, or “more than 2 Medicare”; and “graduate,” “undergraduate,” and larger “house area” (β_std=0.04-0.07). Strong negative predictors were “age” and “couple-only” family type (both β_std=−0.08). Other negative factors included “number of siblings” (β_std=−0.07, 95% CI −0.08 to −0.05), followed by “number of siblings,” “skipped-generation family” type, “employed,” and “self-employed” (β_std=−0.04 to −0.07).

Life Adversity and Stress

Youth socioeconomic status was the strongest positive predictor (β_std=0.14, 95% CI 0.13-0.16). Other positive predictors were “psychological abuse,” “better economic environment,” “emotional neglect,” “high study or work stress,” and “serious injury or illness” (β_std=0.03-0.07). The strongest negative predictors were “sexual abuse,” “physical abuse,” and “collective violence exposure” (β_std=−0.05 to −0.08).

Personality and Self-Efficacy

Self-efficacy was the most prominent predictor (β_std=0.10, 95% CI 0.09-0.12), followed by “narcissistic admiration and rivalry” (β_std=0.04, 95% CI 0.03-0.05) and “conscientiousness” (β_std=0.03, 95% CI 0.02-0.04).

Literacy and Health Empowerment

“Past digital health intervention use” (β_std=0.15, 95% CI 0.14-0.16) and “eHealth literacy” (β_std=0.09, 95% CI 0.07-0.10) were the strongest positive predictors. Other positive predictors included “family-neighbor relationship,” “perceived social support,” “family communication,” and “dietary self-regulation” (β_std=0.03-0.05).

Lifestyle Behaviors

“Stable sleep duration” (somewhat stable vs very unstable: β_std=0.07, 95% CI 0.03-0.10) was the strongest positive predictor, followed by “exposure to art,” “6 - 7 hours sleep,” and “7 - 8 hours sleep” (β_std=0.03-0.04). Negative predictors included “smoking habit,” “daytime sleepiness,” “extra salt in food,” “morning chronotype,” and “sleep difficulty” (all β_std=−0.03).

Physical Health and Exposures

“COVID-19 vaccination” (β_std=0.03, 95% CI 0.00-0.05) was the only predictor in this domain to meet the reporting threshold. No other predictors within this block were statistically significant.

Mental Health and Psychosocial

Positive predictors included “personal existence,” “work burnout” (β_std=0.05, 95% CI 0.04-0.07), “social media addiction,” and “cyberchondria” (β_std=0.03 - 0.06). Negative predictors were “attention-deficit/hyperactivity disorder symptoms” (β_std=−0.07, 95% CI −0.09 to −0.05) and “depression” (β_std=−0.04, 95% CI −0.06 to −0.02).

Classification-Tree Analysis of Acceptance

A complete-case subsample (N=29,333) was used. The decision tree, trained on 70% (n=20,995) of the sample (Figure 5), identified 3 pathways for “acceptor” (score ≥80) and 4 pathways for “nonacceptor” (score <80). On the test set (n=8338), the tree achieved a modest weighted area under the curve of 0.61 (95% CI 0.60 - 0.62), indicating modest discrimination (Figure S1 in Multimedia Appendix 1). Key performance metrics were accuracy=0.681, sensitivity=0.242 for acceptors, and specificity=0.903 for nonacceptors. Thus, the tree had limited ability to identify acceptors but performed substantially better in identifying likely nonacceptors.

XSL•FO
RenderX

Figure 5. Classification decision tree model of weighted immersive extended reality acceptance in health care among adults in China, fitted on the training set from a nationwide cross-sectional survey conducted between June and September 2024 (N=20,955). MD-SF5: 5-item Maladaptive Daydreaming Short Form; PSSS: Perceived Social Support Scale.

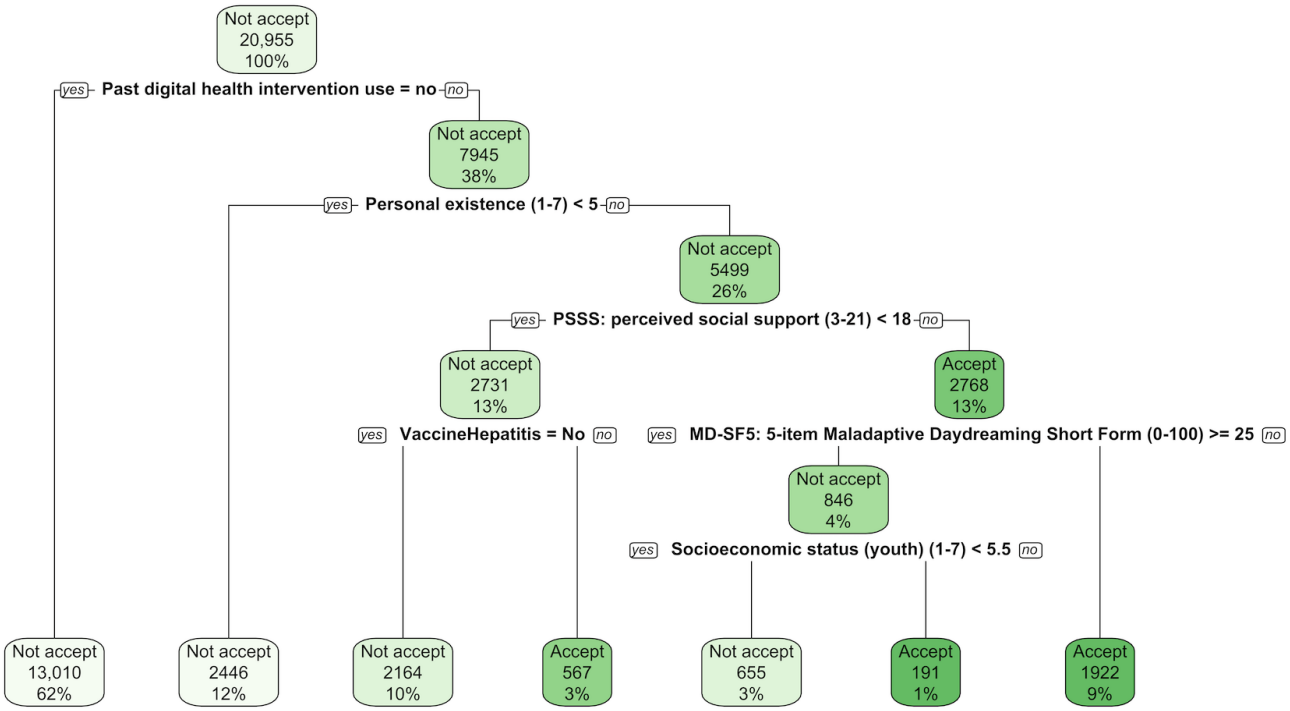

The 3 “acceptor” profiles all required past digital health use and high personal existence (≥5). The largest (1922/20,955, 9%) also had high social support (Perceived Social Support Scale score ≥18) and low maladaptive daydreaming (5-item Maladaptive Daydreaming Short Form [MD-SF5] <25). The second (567/20,955, 3%) had low social support (<18) and hepatitis B vaccination. The smallest (191/20,955, 1%) had high social support, high MDS5 (≥25), and high youth socioeconomic status (≥5.5).

Conversely, 4 “nonacceptor” profiles were identified. The largest (13,010/20,955, 62% of sample) was defined simply by no past digital health use. The other 3 profiles all had past digital use but were defined by low personal existence (<5) (2446/20,955, 12%); or high personal existence, low social support (<18), and no hepatitis B vaccination (2164/20,955, 10%); or high personal existence, high social support, high MDS5 (≥25), and low youth socioeconomic status (<5.5) (655/20,955, 3%).

Discussion

Main Findings

This national survey addressed three aims: (1) estimating population-level acceptance of immersive XR in health care, (2) examining multidomain correlates of acceptance, and (3) developing a parsimonious tool for identifying likely nonacceptors. In relation to these aims, our findings suggest that public acceptance of immersive XR in China is moderate but uneven, with variation across geographic, demographic, and health-related subgroups. Consistent with our hypothesis, acceptance was associated not only with demographic and disease-related characteristics but also with structural, psychosocial, digital, lifestyle, and health-related factors. To facilitate interpretation, we synthesize the seven prespecified analytic blocks into four broader dimensions shaping acceptance: (1) structural-environmental factors, encompassing sociodemographic conditions and life adversity; (2) trait-resilience factors, encompassing personality, self-efficacy, and mental health; (3) social-digital capital, encompassing health literacy, empowerment, social support, and prior digital health experience; and (4) behavioral health profiles, encompassing lifestyle and physical health factors. Finally, the classification tree further suggested that a small set of variables may help identify groups with lower acceptance, although its performance indicates that it should be interpreted as a pragmatic triage aid rather than an individual-level prediction model.

Findings in Context

Overall acceptance was modest, and our findings may provide a useful benchmark for immersive XR pilot programs before wider-scale implementation. We found that acceptance was higher in affluent coastal cities and lower with increasing age. However, even among older adults, acceptance is still influenced by prior exposure, onboarding needs, sensory and cognitive demands, and the availability of caregiver or staff support [26]. This may underscore the need to co-design age-friendly XR systems and highlights the value of future subgroup analyses within older adult populations [27]. For older adults, simplified interfaces, structured orientation, self-paced interaction, and caregiver or staff support may be especially important [26,28]. We also observed heterogeneity by health status, suggesting that acceptance is not uniform across clinical groups. This may imply that immersive XR implementation should be condition-sensitive rather than one-size-fits-all [10].

Structural and environmental factors were important correlates of acceptance. Our findings indicate that higher socioeconomic status, education, perceived social standing, and better insurance

XSL•FO
RenderX

coverage were associated with higher acceptance, whereas older age, lower youth socioeconomic conditions, and some family structure indicators were associated with lower acceptance. These patterns are consistent with the possibility that digital access, cumulative social advantage, and resource constraints may shape receptivity to new health care technologies [29-35]. Notably, the positive association of psychological adverse childhood experiences (abuse and neglect) and recent stressors with higher acceptance might reflect unmeasured confounding rather than a true effect of adversity; these findings should therefore be interpreted cautiously. In contrast, sexual or physical abuse and collective adversity were linked to lower acceptance, plausibly by undermining trust and self-efficacy [36]. From a structural-environmental perspective, equitable implementation of immersive XR requires attention to digital infrastructure, affordability, and shared access models, particularly in underserved settings [27].

Acceptance was also correlated with individual traits. We found that self-efficacy was a strong positive predictor, consistent with the possibility that perceived confidence and capability facilitate receptivity to immersive health tools [25]. Newly identified correlates included conscientiousness, narcissistic admiration, burnout, social media addiction, cyberchondria, attention-deficit/hyperactivity disorder traits, and depressive symptoms. The positive associations of narcissistic admiration, burnout, social media addiction, and cyberchondria are less straightforward and should be interpreted cautiously; they may reflect greater digital immersion or health information seeking rather than inherently greater readiness for sustained XR use [37,38]. Further research is needed to determine whether these factors predict actual XR uptake and sustained engagement, or whether they are markers of broader digital engagement and health information seeking. From an implementation perspective, the strong association with self-efficacy suggests that acceptance may depend partly on users' confidence in trying, navigating, and troubleshooting unfamiliar digital tools. Users with lower confidence or fewer social resources may therefore require more intensive guidance and follow-up [28], such as confidence-building orientation, supervised first use, and simple step-by-step instructions.

Social and digital capital also showed strong associations. Prior digital health use and higher eHealth literacy were among the strongest predictors, consistent with established adoption models, such as the Unified Theory of Acceptance and Use of Technology [39,40]. Stronger family or neighbor ties and perceived social support likely act as motivational and practical resources (eg, encouragement and troubleshooting) [40]. This may be especially true in China's collectivist context, where endorsement from close others functions as a powerful social norm [41]. Finally, dietary self-regulation may proxy a domain-general self-regulatory capacity, which is known to predict adherence to digital health tools [42]. From an implementation perspective, the associations with eHealth literacy, prior digital health use, and social support suggest that some users may need more than access to XR equipment. Implementation may require structured onboarding, hands-on demonstrations, plain-language instructions, troubleshooting support, and optional family or caregiver involvement [9]. For example, health systems could incorporate a brief digital health literacy or digital health readiness screener into intake workflows to identify low-readiness patients for staff-assisted XR orientation, troubleshooting, and follow-up support [43,44].

To our knowledge, few prior XR acceptance studies have examined lifestyle-related factors. Our findings therefore extend previous acceptance research by suggesting that behavioral health profiles may provide additional insight into population-level acceptance of immersive XR. Specifically, we found that more stable and sufficient sleep, arts exposure, evening chronotype, and COVID-19 vaccination were associated with higher acceptance, whereas smoking and salty diet were associated with lower acceptance. These patterns may reflect broader differences in health orientation, novelty receptivity, and familiarity with digital tools [45,46]. For instance, vaccination uptake is often associated with greater institutional trust and health engagement [47]. However, these associations should be interpreted as hypothesis-generating rather than causal, and future studies should examine whether lifestyle-related profiles predict actual XR uptake, adherence, and sustained use. In terms of implementation, strategies should account for users' behavioral health profiles and focus on tailoring interventions to these specific lifestyle factors [48], such as incorporating more motivational framing into VR design for those who smoke.

The classification tree should be interpreted as a pragmatic implementation triage aid, rather than as an individual-level predictive classifier. According to standard interpretation of classification performance metrics [49], its low sensitivity indicates limited ability to identify individuals likely to accept immersive XR. However, its high specificity suggests better performance in identifying participants unlikely to accept immersive XR. Therefore, the tree may be more useful for resource allocation and targeted outreach. In this study, the tree used 6 implementation-relevant variables: past digital health intervention use, perceived meaning in life, perceived social support, hepatitis vaccination status, childhood socioeconomic status, and MDS5. These variables may help implementation teams identify subgroups less likely to accept immersive XR in health care and proactively allocate educational resources and intensive onboarding support to those who require the most assistance [50]. XR onboarding may include prebriefing, written and verbal controller instructions, a supervised first exposure, seated low-intensity familiarization, and real-time language or technical support [51].

Limitations

This study has several limitations. First, its cross-sectional design captures acceptance at a single time point, precluding causal inference or trend analysis. Second, all measures were self-reported and subject to recall and social desirability biases. Third, as noted, the classification-tree model performed poorly on the test set, with very low sensitivity for identifying acceptors. Fourth, the vast regional disparities observed suggest that these individual-level predictors are moderated by powerful cultural, political, or infrastructural factors that were not fully captured in this model. Therefore, our China-based sample may limit generalizability to other cultural or health system contexts.

XSL•FO
RenderX

Fifth, by focusing solely on user-side determinants, this study did not examine interactions with technology-side factors, which are likely to affect uptake and warrant further investigation. Sixth, although the survey used a nationwide multistage framework, participant recruitment at the final stage was guided by quota sampling rather than a fully probability-based design. In addition, poststratification weights were calibrated only to the age-sex distribution, so residual imbalance in other characteristics, such as province, urbanicity, or education, may remain. Moreover, clustering and stratification were not incorporated into variance estimation because the necessary design information was unavailable. Seventh, the geographic analysis was strictly descriptive and did not examine the ecological drivers underlying regional variation. Future research should build on these findings by using multilevel modeling to incorporate province-level covariates to distinguish individual-level determinants from broader contextual effects.

Conclusions

To our knowledge, this is among the first large-scale national studies to examine public acceptance of immersive health care technology using a multidomain framework in China. Unlike prior smaller or population-specific studies, it provides population-level evidence and suggests that receptivity is associated not only with demographic factors but also with structural, psychological, social, digital, behavioral, and health-related characteristics. This study contributes an equity-oriented and implementation-relevant perspective on which population groups may be more or less likely to accept immersive XR in health care. The findings suggest that equitable adoption may require more than technological availability alone; it may also depend on age-friendly design, condition-sensitive implementation, and targeted support for users with lower self-efficacy, lower digital readiness, or fewer social resources. In addition, the 6-item classification tree may offer a pragmatic starting point for identifying likely nonacceptors for outreach and support prioritization, but it requires external validation before practical use. Overall, our findings suggest that successful and equitable implementation of immersive XR is not only a technical issue but also a social, behavioral, and implementation challenge.

Acknowledgments

The authors thank the 48 regional and disciplinary experts for their invaluable feedback on the questionnaire and the surveyors across 33 provinces in China for facilitating data collection. The authors attest that no generative artificial intelligence was used in any portion of the manuscript generation or preparation.

Funding

The authors declared that no financial support was received for this work.

Data Availability

Custom scripts for data cleaning, model development, and figure generation are available from the corresponding author upon reasonable request. The deidentified individual participant data, data dictionary, survey instrument, and analysis code will be available on reasonable request from the corresponding author. Access begins upon publication for 5 years and requires a brief proposal and data use agreement. Public posting is not permitted due to privacy or ethics restrictions.

Authors' Contributions

Conceptualization: JYL (lead), DYTF (equal)
Methodology: JYL (lead)
Data curation: YW
Investigation: YW (lead)
Supervision: JYL (lead)
Writing – original draft: JYL
Writing – review & editing: JXL (equal), PMD (equal), CLW (equal), HYLC (equal), XQ (equal), ZY (equal), ATC (equal)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials including the receiver operating characteristic curve for the classification and regression tree model; predictor definitions and reliability metrics; weighted descriptive results by province, age group, gender, and health condition; and hierarchical weighted linear regression results across 7 predictor blocks.
[DOCX File, 583 KB - jmir_v28i1e88000_app1.docx]

References

XSL•FO
RenderX

1. Kafes M, Ileri YY. Current status of virtual reality researches at healthcare: thematic and bibliometric analysis. Front Virtual Real 2025;6:1411075. [doi: 10.3389/frvir.2025.1411075]
2. Iqbal AI, Aamir A, Hammad A, et al. Immersive technologies in healthcare: an in-depth exploration of virtual reality and augmented reality in enhancing patient care, medical education, and training paradigms. J Prim Care Community Health 2024;15:21501319241293311. [doi: 10.1177/21501319241293311] [Medline: 39439304]
3. Faizan Siddiqui M, Jabeen S, Alwazzan A, et al. Integration of augmented reality, virtual reality, and extended reality in healthcare and medical education: a glimpse into the emerging horizon in LMICs—a systematic review. J Med Educ Curric Dev 2025 May 29;12:23821205251342315. [doi: 10.1177/23821205251342315] [Medline: 40452894]
4. Arthur T, Melendez-Torres GJ, Harris D, Robinson S, Wilson M, Vine S. Extended reality interventions for health and procedural anxiety: panoramic meta-analysis based on overviews of reviews. J Med Internet Res 2025 Jan 8;27:e58086. [doi: 10.2196/58086] [Medline: 39778203]
5. Tang P, Cao Y, Vithran DTAV, et al. The efficacy of virtual reality on the rehabilitation of musculoskeletal diseases: umbrella review. J Med Internet Res 2025 Apr 25;27:e64576. [doi: 10.2196/64576] [Medline: 40279163]
6. Fereidooni M, Toni E, Toni E, Ayatollahi H. Application of virtual reality for supportive care in cancer patients: a systematic review. Support Care Cancer 2024 Aug 5;32(9):570. [doi: 10.1007/s00520-024-08763-1] [Medline: 39103681]
7. Hao J, Crum G, Siu KC. Effects of virtual reality on stroke rehabilitation: an umbrella review of systematic reviews. Health Sci Rep 2024 Sep;7(9):e70082. [doi: 10.1002/hsr2.70082] [Medline: 39328980]
8. Sung H, Kim M, Park J, Shin N, Han Y. Effectiveness of virtual reality in healthcare education: systematic review and meta-analysis. Sustainability 2024;16(19):8520. [doi: 10.3390/su16198520]
9. Kouijzer MMTE, Kip H, Bouman YHA, Kelders SM. Implementation of virtual reality in healthcare: a scoping review on the implementation process of virtual reality in various healthcare settings. Implement Sci Commun 2023 Jun 16;4(1):67. [doi: 10.1186/s43058-023-00442-2] [Medline: 37328858]
10. Elser A, Lange M, Kopkow C, Schäfer AG. Barriers and facilitators to the implementation of virtual reality interventions for people with chronic pain: scoping review. JMIR XR Spat Comput 2024;1:e53129. [doi: 10.2196/53129] [Medline: 41341873]
11. Arthur T, Robinson S, Vine S, Asare L, Melendez-Torres GJ. Equity implications of extended reality technologies for health and procedural anxiety: a systematic review and implementation-focused framework. J Am Med Inform Assoc 2025 May 1;32(5):945-957. [doi: 10.1093/jamia/ocaf047] [Medline: 40112188]
12. Wang EY, Qian D, Zhang L, et al. Acceptance of virtual reality in trainees using a technology acceptance model: survey study. JMIR Med Educ 2024 Dec 23;10:e60767. [doi: 10.2196/60767] [Medline: 39727193]
13. Felsberg DT, McGuirt JT, Ross SE, Raisbeck LD, Howard CK, Rhea CK. Clinician perspectives on virtual reality use in physical therapy practice in the United States. PLoS One 2025;20(4):e0320215. [doi: 10.1371/journal.pone.0320215] [Medline: 40173138]
14. Raghuraman N, Bedford T, Tran N, Haycock NR, Wang Y, Colloca L. The interplay between health disparities and acceptability of virtual reality: a survey study. Cyberpsychol Behav Soc Netw 2024 Dec;27(12):894-902. [doi: 10.1089/cyber.2023.0694] [Medline: 39088356]
15. Zeng F, Li Q, Cai S, et al. Cancer patients' acceptance of virtual reality interventions for self-emotion regulation. Sci Rep 2025 Apr 9;15(1):12185. [doi: 10.1038/s41598-025-95160-1] [Medline: 40204807]
16. Chen Y, Guan B, Zhang Y, et al. Acceptability of and willingness to use virtual reality exergames for weight loss among young adults with overweight or obesity in China: qualitative study. JMIR Serious Games 2025 May 1;13:e66998. [doi: 10.2196/66998] [Medline: 40311124]
17. Sauchelli S, Pickles T, Voinescu A, et al. Public attitudes towards the use of novel technologies in their future healthcare: a UK survey. BMC Med Inform Decis Mak 2023 Feb 22;23(1):38. [doi: 10.1186/s12911-023-02118-2] [Medline: 36814262]
18. Huygelier H, Schraepen B, van Ee R, Vanden Abeele V, Gillebert CR. Acceptance of immersive head-mounted virtual reality in older adults. Sci Rep 2019 Mar 14;9(1):4519. [doi: 10.1038/s41598-019-41200-6] [Medline: 30872760]
19. Moore RC, Hancock JT, Bailenson JN. From 65 to 103, older adults experience virtual reality differently depending on their age: evidence from a large-scale field study in nursing homes and assisted living facilities. Cyberpsychol Behav Soc Netw 2023 Dec;26(12):886-895. [doi: 10.1089/cyber.2023.0188] [Medline: 38011717]
20. Keller MS, Park HJ, Cunningham ME, Fouladian JE, Chen M, Spiegel BMR. Public perceptions regarding use of virtual reality in health care: a social media content analysis using Facebook. J Med Internet Res 2017 Dec 19;19(12):e419. [doi: 10.2196/jmir.7467] [Medline: 29258975]
21. Dy M, Olazo K, Lisker S, et al. Virtual reality for chronic pain management among historically marginalized populations: systematic review of usability studies. J Med Internet Res 2023 Jun 6;25:e40044. [doi: 10.2196/40044] [Medline: 37279039]
22. Siette J, Adam PJ, Harris CB. Acceptability of virtual reality to screen for dementia in older adults. BMC Geriatr 2024 Jun 5;24(1):493. [doi: 10.1186/s12877-024-05115-w] [Medline: 38840041]
23. Wong EYC, Hui RTY, Kong H. Perceived usefulness of, engagement with, and effectiveness of virtual reality environments in learning industrial operations: the moderating role of openness to experience. Virtual Real 2023 Apr 24;27:1-17. [doi: 10.1007/s10055-023-00793-0] [Medline: 37360804]

XSL•FO
RenderX

24. Al Khalifah E, Hammady R, Abdelrahman M, et al. Technology anxiety in virtual reality adoption: examining the impact of age, past experience, and cybersickness. IEEE Access 2025;13:71858-71879. [doi: 10.1109/ACCESS.2025.3562383]
25. Liu Y, Lu X, Zhao G, Li C, Shi J. Adoption of mobile health services using the unified theory of acceptance and use of technology model: self-efficacy and privacy concerns. Front Psychol 2022;13:944976. [doi: 10.3389/fpsyg.2022.944976]
26. Schaumburg M, Imtiaz A, Zhou R, Bernard M, Wolbers T, Segen V. Immersive virtual reality for older adults: challenges and solutions in basic research and clinical applications. Ageing Res Rev 2025 Jul;109:102771. [doi: 10.1016/j.arr.2025.102771] [Medline: 40373984]
27. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119. [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
28. Li Y, Shiyanov I, Muschalla B. Older adults' acceptance of a virtual reality group intervention in nursing homes: pre-post study under naturalistic conditions. JMIR Hum Factors 2024 Aug 6;11:e56278. [doi: 10.2196/56278] [Medline: 39373557]
29. Tuitert I, Marinus JD, Dalenberg JR, van 't Veer JT. Digital health technology use across socioeconomic groups prior to and during the COVID-19 pandemic: panel study. JMIR Public Health Surveill 2024 Sep 13;10:e55384. [doi: 10.2196/55384] [Medline: 39269755]
30. Li Q, Cheng F, Zeng H, Xu J. Health insurance payment for telehealth services: scoping review and narrative synthesis. J Med Internet Res 2024 Dec 9;26:e56699. [doi: 10.2196/56699] [Medline: 39652868]
31. Montagni I, Cariou T, Feuillet T, Langlois E, Tzourio C. Exploring digital health use and opinions of university students: field survey study. JMIR Mhealth Uhealth 2018 Mar 15;6(3):e65. [doi: 10.2196/mhealth.9131] [Medline: 29549071]
32. Bertolazzi A, Quaglia V, Bongelli R. Barriers and facilitators to health technology adoption by older adults with chronic diseases: an integrative systematic review. BMC Public Health 2024 Feb 16;24(1):506. [doi: 10.1186/s12889-024-18036-5] [Medline: 38365698]
33. Zainal H, Xiaohui X, Thumboo J, Seah SJ, Leng LL, Kok Yong F. Exploring caregiver challenges, digital health technologies, and healthcare support: a qualitative study. Front Digit Health 2025;7:1587162. [doi: 10.3389/fdgth.2025.1587162] [Medline: 40547053]
34. Downey DB. Number of siblings and intellectual development. The resource dilution explanation. Am Psychol 2001;56(6-7):497-504. [doi: 10.1037//0003-066x.56.6-7.497] [Medline: 11413873]
35. Silver MP. Socio-economic status over the lifecourse and internet use in older adulthood. Ageing Soc 2014 Jul;34(6):1019-1034. [doi: 10.1017/S0144686X12001420]
36. Diehl AS, Prout MF. Effects of posttraumatic stress disorder and child sexual abuse on self-efficacy development. Am J Orthopsychiatry 2002 Apr;72(2):262-265. [doi: 10.1037/0002-9432.72.2.262] [Medline: 15792065]
37. El-Zayat A, Namnkani SA, Alshareef NA, Mustfa MM, Eminaga NS, Algarni GA. Cyberchondria and its association with smartphone addiction and electronic health literacy among a Saudi population. Saudi J Med Med Sci 2023;11(2):162-168. [doi: 10.4103/sjmms.sjmms_491_22] [Medline: 37252023]
38. Casale S, Banchi V. Narcissism and problematic social media use: a systematic literature review. Addict Behav Rep 2020 Jun;11:100252. [doi: 10.1016/j.abrep.2020.100252] [Medline: 32467841]
39. Rahimi B, Nadri H, Lotfnezhad Afshar H, Timpka T. A systematic review of the technology acceptance model in health informatics. Appl Clin Inform 2018 Jul;9(3):604-634. [doi: 10.1055/s-0038-1668091] [Medline: 30112741]
40. Shiu LS, Huang YS, Liu CY, Cheng YS, Chen YC. eHealth literacy mediating social support and technology acceptance among patients with chronic illnesses: a cross-sectional study. J Adv Nurs 2026 May;82(5):5049-5059. [doi: 10.1111/jan.70207] [Medline: 40922516]
41. Venkatesh V, Zhang X. Unified Theory of Acceptance and Use of Technology: U.S. Vs. China. Journal of Global Information Technology Management 2010 Jan;13(1):5-27. [doi: 10.1080/1097198X.2010.10856507]
42. Plaitano EG, McNeish D, Bartels SM, et al. Adherence to a digital therapeutic mediates the relationship between momentary self-regulation and health risk behaviors. Front Digit Health 2025;7:1467772. [doi: 10.3389/fdgth.2025.1467772] [Medline: 39981105]
43. Nelson LA, Pennings JS, Sommer EC, Popescu F, Barkin SL. A 3-item measure of digital health care literacy: development and validation study. JMIR Form Res 2022 Apr 29;6(4):e36043. [doi: 10.2196/36043] [Medline: 35486413]
44. Rising KL, Guth A, Gentsch AT, et al. Development and preliminary validation of a screener for digital health readiness. JAMA Netw Open 2024 Sep 3;7(9):e2432718. [doi: 10.1001/jamanetworkopen.2024.32718] [Medline: 39254972]
45. Kortesoja L, Vainikainen MP, Hotulainen R, Merikanto I. Late-night digital media use in relation to chronotype, sleep and tiredness on school days in adolescence. J Youth Adolesc 2023 Feb;52(2):419-433. [doi: 10.1007/s10964-022-01703-4] [Medline: 36401709]
46. White TM, Wyka K, Rabin K, El-Mohandes A. Trust in the science behind COVID-19 vaccines as a driver of vaccine acceptance in the United States, 2021-2023. Vaccine X 2024 Dec;21:100576. [doi: 10.1016/j.jvacx.2024.100576] [Medline: 39555242]
47. Krastev S, Krajden O, Vang ZM, et al. Institutional trust is a distinct construct related to vaccine hesitancy and refusal. BMC Public Health 2023 Dec 12;23(1):2481. [doi: 10.1186/s12889-023-17345-5] [Medline: 38082287]
48. Noar SM, Benac CN, Harris MS. Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. Psychol Bull 2007 Jul;133(4):673-693. [doi: 10.1037/0033-2909.133.4.673] [Medline: 17592961]

XSL•FO
RenderX

49. Pewsner D, Battaglia M, Minder C, Marx A, Bucher HC, Egger M. Ruling a diagnosis in or out with “SpPIn” and “SnNOut”: a note of caution. BMJ 2004 Jul 24;329(7459):209-213. [doi: 10.1136/bmj.329.7459.209] [Medline: 15271832]
50. Kappen TH, van Klei WA, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, Moons KGM. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. Diagn Progn Res 2018;2:11. [doi: 10.1186/s41512-018-0033-6] [Medline: 31093561]
51. Leung RYF, Ye MZ, Zhang FY, et al. Feasibility and efficacy of commercial-off-the-shelf virtual reality applications for managing chronic pain and enhancing well-being among older adults in the community: mixed methods pilot study. JMIR Form Res 2025 Aug 18;9:e67765. [doi: 10.2196/67765] [Medline: 40825190]

Abbreviations:

AR: augmented reality
β_std: standardized β coefficient
FDR: false discovery rate
MD-SF5: 5-item Maladaptive Daydreaming Short Form
MR: mixed reality
VR: virtual reality
XR: extended reality

Edited by S Brini; submitted 22.Nov.2025; peer-reviewed by B Birckhead, K Mouloudj, Z Liu; revised version received 18.May.2026; accepted 19.May.2026; published 26.Jun.2026.

Please cite as:
Li J, Davidson PM, Chan HYL, Wong CL, Qi X, Ye Z, Cheung AT, Fong DYT, Wu Y, Li J
National Acceptance and Determinants of Immersive Extended Reality in Health Care in China: Cross-Sectional Study
J Med Internet Res 2026;28:e88000
URL: https://www.jmir.org/2026/1/e88000
doi:10.2196/88000

© Jiaying Li, Patricia M Davidson, Helen YL Chan, Cho Lee Wong, Xiang Qi, Zengjie Ye, Ankie Tan Cheung, Daniel YT Fong, Yibo Wu, Junxin Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Open Syndrome Definition as a Machine-Readable Standard for Public Health: Design and Implementation Study

Ana Paula Gomes Ferreira[1,2], MSc; Aleksandar Anžel[1], PhD; Izabel Marcilio[3], Dr med; Helen Hughes[4], PhD; Alex J Elliot[4], PhD; Jude Dzevela Kong[5], PhD; Madlen Schranz[6], MSc; Alexander Ullrich[6], PhD; Georges Hattab[1,2], PhD

[1]Center for Artificial Intelligence in Public Health Research, Robert Koch Institute, Nordufer 20, Berlin, Germany

[2]Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany

[3]Center for Data and Knowledge Integration for Health, Fundação Oswaldo Cruz, Salvador, Brazil

[4]Real-time Syndromic Surveillance Team, UK Health Security Agency, Birmingham, United Kingdom

[5]Department of Mathematics and Statistics, Africa-Canada Artificial Intelligence and Data Innovation Consortium, York University, Toronto, ON, Canada

[6]Infectious Disease Epidemiology Department, Robert Koch Institute, Berlin, Germany

Corresponding Author:
Georges Hattab, PhD
Center for Artificial Intelligence in Public Health Research, Robert Koch Institute, Nordufer 20, Berlin, Germany

Abstract

Background: Case definitions are essential for effectively communicating public health threats. However, the absence of a standardized, machine-readable format poses significant challenges to interoperability, epidemiological research, data sharing, and the application of computational methods, including artificial intelligence. These barriers complicate collaboration across regions and organizations and hinder technological progress in public health.

Objective: This study aims to propose and release the first open, machine-readable format for representing case and syndrome definitions, together with tools and resources that enable their standardized and scalable use.

Methods: We developed the Open Syndrome Definition, a structured, machine-readable schema for representing case and syndrome definitions. We compiled official public health case definitions from multiple institutions and converted them into standardized, machine-readable representations using open-source tools. These tools, available through GitHub under the Massachusetts Institute of Technology license, automate the translation of narrative definitions into structured data. We also created a platform for browsing, analyzing, and contributing new definitions on our initiative website.

Results: The Open Syndrome Definition format enabled consistent, automated representation of case definitions across different diseases and jurisdictions. The conversion tools achieved high semantic fidelity, as assessed by qualitative expert review, between narrative and structured representations, supporting human verification and automated analysis. The dataset and accompanying tools demonstrated structural and semantic interoperability by standardizing definitions from various health systems into a unified format and integrating existing medical ontologies through JSON for Linked Data. To further illustrate practical applicability and downstream usage, we introduced a data filtering prototype that allows users to upload their own datasets and verify the results against the standardized definitions.

Conclusions: The Open Syndrome Definition establishes a foundation for consistent and machine-readable public health definitions, facilitating reproducible research and interoperability at scale. By enabling systematic data exchange and artificial intelligence–driven analysis, it strengthens public health preparedness and supports more rapid, coordinated responses to emerging health threats.

(J Med Internet Res 2026;28:e86249) doi:10.2196/86249

KEYWORDS

public health surveillance; epidemiological monitoring; health information exchange; public health; AI; data standardization; artificial intelligence

Introduction

Case definitions are essential tools for public health practitioners. They are used to identify, monitor, and respond to diseases or groups of diseases [1,2]. They inform the public, orient public health policies, and guide surveillance indicators, such as syndrome definitions [3]. In the context of public health surveillance, terminology can often be fragmented. While case definitions typically describe strict clinical and laboratory criteria for identifying specific diseases, the criteria used for

XSL•FO
RenderX

early warning systems are referred to in the literature as either syndrome definitions or syndromic indicators. Due to the lack of global consensus [4,5], these 2 terms are frequently used interchangeably in practice; therefore, throughout this paper, we treat them as synonymous.

Developing case definitions requires expert knowledge of the targeted disease(s), consultation of existing definitions, and analysis of available clinical data. Case definitions are usually written as free-text descriptions of the key characteristics and criteria of the target disease or public health threat. The goal of a case definition is to provide a consistent description for public health officials, health workers, policymakers, and the general public so they can understand a given threat. Figure 1 shows two different case definitions for the same disease from different provenance [6,7]. The lack of interoperability and standardization of syndromic indicators and case definitions makes it difficult to compare systems and epidemiological situations across different regions and time periods [8]. Free-text definitions, in particular, introduce a fundamental gap in translation between human and machine interpretation, resulting in inconsistent and error-prone automated processing [9]. Beyond mere structural inconsistencies, free-text descriptions are highly susceptible to linguistic ambiguity and semantic fragmentation [10]. While complex logical relationships and clinical nuances appear straightforward to trained public health professionals, they become deeply problematic for computational systems.

Figure 1. Measles case definition comparison. The definitions from the European Centre for Disease Prevention and Control [6] and India [7] are presented on the left and right, respectively. ECDC: European Centre for Disease Prevention and Control.

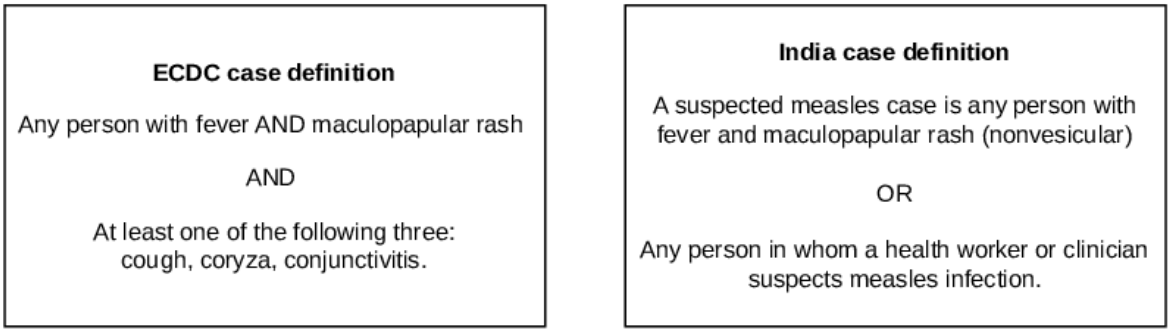

Consequently, effectively translating these definitions for automated surveillance requires more than just a technical data container; it necessitates a robust methodology to resolve semantic ambiguities and ensure true interoperability across diverse health systems [11]. Examples of these relationships are conditions and nested criteria. These inconsistencies can delay coordinated responses during critical time periods when containing emerging threats is essential. For example, inconsistent case definitions had far-reaching consequences for global disease surveillance and control efforts during the COVID-19 pandemic. The consequences of these issues include underreporting and misclassifications of cases [12], compromised accuracy of surveillance data [13-15], compromised resource allocation [16], and problems comparing disease burden and intervention effectiveness across countries [17]. The variability in case definitions and the absence of structured formats to standardize them represent a surprisingly underexplored gap in public health infrastructure. Moreover, the dearth of structured formats imposes substantial constraints on embracing artificial intelligence (AI) and eventually implementing AI for surveillance and outbreak detection. As health systems increasingly rely on computational approaches, a significant challenge to the use of AI as a technology for protecting public health is posed by the gap between narrative definitions and machine-readable formats. When disease indicators are monitored in real-time or near real-time for the purpose of early outbreak detection, it is imperative to use a structured format for case definitions. This is especially true in the context of syndromic surveillance, where automated data acquisition is used [18].

Earlier studies [19,20] verify that using a structured format for clinical standards in case definitions leads to enhanced reporting accuracy compared to narrative descriptions. These studies emphasize that even well-defined case definitions can be interpreted differently by various users, which can lead to misclassifications, decreased sensitivity, and lower positive predictive value. A highlight from the related work is when case definitions were successfully used to automate the generation of a new case definition and translate an existing one from medical conditions [21]. Notwithstanding this evidence, at the time of writing, no comprehensive effort has been made to develop a standardized, machine-readable format for case definitions. Adopting this format would eliminate any inconsistencies in interpretation and lay the groundwork for the infrastructure needed to support next-generation public health surveillance systems.

As a matter of fact, clearly structured definitions can help AI systems identify diseases and track outbreaks more effectively. This enables machine learning (ML) programs to detect cases more accurately, notice subtle changes in disease patterns, and alert us to emerging health threats, even when traditional symptoms are not yet evident.

To achieve these capabilities, the core scientific contribution of this work is the development of a methodology for semantic interoperability in public health surveillance. As the practical implementation of this methodology, we introduce the Open Syndrome Definition (OSD) framework, a name chosen to reflect its capacity to encompass both case and syndrome definitions. Moving beyond basic structural formatting, our approach leverages a machine-readable representation integrated

XSL•FO
RenderX

with semantic web principles (JSON for Linked Data [JSON-LD]) [22,23] and established medical ontologies, successfully bridging the gap between human readability and computational precision. Consequently, OSD offers the flexibility and clarity required to create precise, unambiguous descriptions suitable for AI applications and diverse software tools. Furthermore, by incorporating metadata often missing from traditional narratives, the framework enables the reuse of definitions, cross-jurisdictional comparisons, version control, and downstream applications in ML projects, benchmarks, and publications. Importantly, OSD complements, rather than replaces, traditional narrative structures. By systematically disambiguating epidemiological criteria, it allows computational systems to precisely interpret case definitions, facilitating automated classification and seamless integration into modern digital surveillance pipelines.

In conjunction with the OSD format, this work presents the inaugural dataset of its kind: a collection of machine-readable case definitions for a plethora of diseases from a network of countries spanning 5 continents: the Americas, Europe, Africa, Oceania, and Asia. True interoperability between jurisdictions, reproducible research, and a foundation for more responsive, data-driven approaches to disease surveillance and public health emergency management are all enabled by this common "language."

Case definitions play a critical role in public health and epidemiological research. However, they often suffer from ambiguity and a lack of standardization. This limits their utility for computational processing and interoperability. This study addresses the critical gap in translation between human-readable and machine-processable case definitions. To accomplish this, the study has three main goals: (1) to create the OSD, a standardized, machine-readable format that preserves the logical complexity of case definitions while eliminating ambiguity, (2) to enable the conversion of the machine-readable format to free text and vice versa, which is a necessary operational feature, and (3) to compile the first extensive dataset of structured case definitions that covers various diseases and jurisdictions. This study aims to support future advances in public health surveillance, syndromic surveillance, and broader epidemiological research by pursuing these goals.

Methods

Overview

This section describes the methodology used to develop the OSD format, dataset, and supporting tools. We took an iterative approach to the development process. The format evolved through continuous refinement as we collected case definitions, which became our dataset, and created supporting tools. Although this section is organized into distinct subsections for clarity, these components were developed concurrently and informed each other throughout the research process.

Definitions Dataset

Case definitions have specific guidelines regarding writing style. These guidelines emphasize simplicity and conciseness and encourage the use of a narrative format [1]. Many governmental [24-26] and public health organizations [6,27] publicly share their case definitions to ensure that health workers and the public can access information about monitored diseases and their characteristics.

To identify scientific papers mentioning datasets related to case or syndrome definitions, we searched for the keywords *case definition AND dataset* and *syndrome definition AND dataset* across various scientific sources, as shown in Table 1. The choice of terms reflects the core concepts of interest, case definitions, and their association with publicly available or structured datasets. Our goal was to cast a broad net without overly constraining the results, so we used simple, inclusive phrases and applied no filters based on time or language. However, despite these efforts, we were unable to identify any relevant datasets.

Table 1 summarizes the results of searches conducted in 2 scientific article databases (PubMed and OpenAlex) and 3 dataset platforms: Kaggle, Hugging Face, and Harvard Dataverse. The searches yielded 286 results for "case definition dataset" and 12 results for "syndrome definition dataset," but none of these were actual datasets for case or syndrome definitions.

Table . Search queries for case and syndrome definition datasets accessed on March 24, 2025.

Source	"case definition" AND "dataset"	"syndrome definition" AND "dataset"
PubMed [28]	53 results	0 results
OpenAlex [29]	213 results	12 results
Kaggle [30]	4 results	0 results
HuggingFace [31]	2 results	0 results
Harvard Dataverse [32]	14 results	0 results
Total	286 results	12 results
Confirmed datasets	0	0

Because our initial search yielded no results, we took a targeted approach to compile case and syndrome definitions for our dataset. Our methodology was a multistep heuristic process.

First, we focused exclusively on World Health Organization (WHO) member countries. Second, to ensure balanced representation, we aimed to include countries from each of the following continents: the Americas, Europe, Africa, Oceania,

XSL•FO
RenderX

and Asia. Third, for each country, we began with the search query <country name> AND case definitions; if no relevant results were found, we successively tried <country name> AND syndromic surveillance definitions and <country name> AND syndrome definitions. Fourth, we aimed to include at least one definition from at least 10% of the countries within each continent. Once this target was met, we moved on to the next country and then the next continent. When a country provided multiple definitions on a single webpage or PDF, we selected a disease or group of diseases accordingly. DuckDuckGo was our primary search engine. For every successful find, we exported the web page or PDF file featuring case definitions. Since a page or file may contain multiple definitions, we extracted the text to create a machine-readable version. Among WHO member countries, Japan, Indonesia, Russia, and Cuba had no public definitions.

Ultimately, we collected a total of 40 case definitions. This collection comprises 36 national and regional definitions that collectively represent 60 countries. This coverage is reached primarily because a single regional definition from the Pacific Public Health Surveillance Network accounts for a group of 22 nations. Of the remaining 4 definitions, 3 are from continental organizations (the Pan American Health Organization, the European Centre for Disease Prevention and Control, and the Africa Centers for Disease Control and Prevention) and 1 is from the WHO as a global entity.

As we gathered definitions, common themes surfaced across various diseases and regions, including symptoms, diseases, epidemiological links, laboratory tests, and medical evaluations. We also recognized recurring logical patterns in how definitions combined criteria. Operators like AND, OR, and especially AT LEAST were frequently used. The AT LEAST operator called for a specific number of criteria to be met, as illustrated in definitions such as “fever AND at least two symptoms from: cough, headache, loss of smell, back pain.” Through careful analysis of existing definitions and relevant scientific literature [1,4,13], we pinpointed important metadata elements that, while often found in the broader context of portals and publications, were not consistently included with the definitions themselves. These insights led to the continuous development of our OSD format, which we cover in detail in the Schema subsection.

The dataset is organized into sections for human and machine readability. The human-readable section contains original PDF publications, which may cover multiple diseases, as well as TXT files that contain case definitions converted into our format. To ensure consistency, all web-published definitions were exported to PDF. The machine-readable section includes JSON representations of these definitions that have been validated with the Ajv validator [33]. The dataset is available on the Open Syndrome Initiative (OSI) page at HuggingFace and GitHub.

Schema

The OSD format converts traditional, narrative case definitions into a structured JSON schema with defined properties and types. These properties were derived through an in-depth review of a wide range of existing case definitions to ensure they reflect common structural and semantic elements found in real-world usage. When analyzing case definitions of different public health threats from various countries and continents, we identified consistent patterns in information groups. More details are provided in the Definitions Dataset subsection. We observed that the majority (217/407, 53%) relied on symptoms to describe conditions, followed by diagnosis (34/407, 8%). Additionally, half of the case definitions included criteria that resembled logical operators (such as AND, OR, and AT LEAST) to group-related conditions. Based on these observations, we developed a nested structure using a data-driven approach combined with our established principles. As illustrated in Figure 2, we organized the format into 2 main groups of information: metadata and criteria. We describe both groups of information below.

XSL•FO
RenderX

Figure 2. The human- and machine-readable case definitions of influenza-like illness from the Ministry of Health of Brazil are in the Open Syndrome definition format. The human-readable case definition is shown in the gray box on the right. The machine-readable definition is on the left and uses the JSON format proposed by this work. There are two additional boxes on the sides: the green upper box indicates the metadata fields, and the pink box outlines the inclusion and exclusion criteria.

XSL•FO
RenderX

First, the metadata information group provides essential context about the definition itself. This includes version information, scope (broad or sensitive vs narrow or specific), publication details, the responsible organization, the language used, and other provenance information. Typically, in narrative formats, this metadata is presented implicitly within published documents or websites (eg, European Centre for Disease Prevention and Control case definitions [6], WHO Outbreak Toolkit [34]). We enabled efficient information retrieval, version tracking, and proper attribution by explicitly structuring this information in a machine-readable format. Furthermore, the metadata information group consists of metadata properties that capture essential contextual information about the case definition, ranging from basic identification (title and description) to publication details (published_in, published_at, and authors) and geographical context (location and language). We included properties for tracking the status and version of the case definition itself and the OSD schema. Most metadata elements are derived directly from the original case definitions, though several properties, such as Open Syndrome Version, $schema, and the JSON-LD @context, are specific to our format's technical infrastructure and initiative, which we discuss further in the OSI section. These and other individual metadata properties and their meaning can be seen in Table 2.

Table 2 outlines the first-level properties of the Open Syndrome Definition format, dividing them into metadata and criteria. The metadata category contains information about the format and syndrome, including location, organization, and version. The criteria category contains structured information that defines the syndrome itself.

XSL•FO
RenderX

Table . Breakdown of the format by group of information types.

Key components and their properties	Description
Metadata	
$schema	URI[a] pointing to the JSON[b] schema that validates this document
@context	JSON-LD[b,c] context mapping for semantic interoperability
@type	Semantic class (eg, OSD[d]: case definition or OSD[d]: syndrome definition)
ID	Unique ID from the OSI[e]
Title	Case definition title
Description	A detailed description of the definition
Human-readable definition	A human-readable description of the case definition for user interfaces
Scope	Level of specificity. Options: broad or specific
Created at	Date when the definition was created
Published in	Source or platform where the definition was published
Published at	Date and time in UTC[f] when the definition was published
Published by	List of case definition publishers in the OSI[e]
Authors	List of the case definition authors
Location	Geographical location relevant to the schema's application
Language	Language in which the definition is written (eg, English and Spanish)
Organization	Organization or initiative responsible for the schema maintenance
Status	Current status of the definition. Options: draft, published, and deprecated
Keywords	Keywords related to a definition (eg, COVID-19 and mpox, outbreaks)
Target public health threats	List of public health threats that this definition targets
Notes	Any notes that may be relevant to this definition
Category	Case definition categories [1]. Options: confirmed, probable, suspected
Version	Case definition version to be established by the author
Open Syndrome version	Open Syndrome Definition schema version. Currently: v1 (version 1)
Definition type	Case definition or syndromic indicator
Surveillance system type	Type of surveillance system this definition is
Criterion	
Criterion (meta-type)	Set of properties and operators to select a case (details in Table 3)
Inclusion criteria (criterion)	Criterion used to include a case
Exclusion criteria (criterion)	Criterion used to exclude a case
References	Scientific references that have supported this definition

[a]URI: uniform resource identifier.
[b]JSON: JavaScript object notation.
[c]LD: linked data.
[d]OSD: Open Syndrome definition.
[e]OSI: Open Syndrome initiative.
[f]UTC: Coordinated Universal Time

Second, the inclusion and exclusion criteria specify the conditions that determine whether a case meets the definition or should be excluded. These conditions, referred to as criteria, can be combined using logical operators (AND, OR, and AT_LEAST) to express complex clinical relationships. We organized the data into key and value pairs (properties) within the JSON schema, enabling computational systems to interpret simple and compound conditions precisely. This structured approach eliminated ambiguities often present in narrative text [35] while preserving the clinical intent of the original definitions. The criteria properties form the backbone of the format and are structured to capture the logical relationships

XSL•FO
RenderX

inherent in case definitions. The format distinguishes between inclusion and exclusion criteria, both built upon our criterion meta-type, as described in Table 3. This structure enables the representation of complex clinical reasoning within a machine-readable framework.

Table . Criterion meta-type properties (summary of properties of a criterion and a fundamental component of a definition).

Property	Description	Usage details
Type	Type of criterion. Options: criterion, syndrome, symptom, diagnosis, diagnostic_test, professional_judgment, epidemiological_history, and demographic_criteria	The type property is mandatory, followed by name or values
Name	Criterion label	__[a]
Description	Detailed description of the criterion	—
Ontology ID	The compacted identifier from an ontology (eg, Disease Ontology (DOID))	Examples: hpo:0002045 and mondo:0020674. Facilitates semantic interoperability
Logical operator	Keywords that represent a logical operation on criteria. Options: AND, OR, and AT_LEAST	The logical operator AT_LEAST must be used with the number specified in logical_operator_arguments
Logical operator arguments	List of arguments to be passed to the logical operator	—
Attribute	The referred attribute, for example, body temperature, age, and onset	It is used in composition with operator and value
Value	The reference value for the referred attribute. It could be of any data type because it can represent anything in the real world	Examples: true, 37.6, abnormal but non-specific bowel gas pattern. This property is used in composition with attribute and operator
Operator	Comparison and matching operators. Options: >, >=, <, <=, ==, !=, regex	It is used in composition with value and attribute
Regex pattern	Regular expression for evaluation and pattern matching	It is used in composition with operator
Regex flags	Regular expression flags for extra configuration	It is used in composition with regex_pattern
Code	A system-agnostic diagnosis code object that holds system, code, and display	Useful to represent values from the *ICD*[b], SNOMED CT[c], and others
Values	A list whose types are a criterion	The criteria items should be unique

[a]Not applicable.

[b]*ICD*: *International Statistical Classification of Diseases and Related Health Problems*.

[c]SNOMED-CT: Systematized Nomenclature of Medicine—Clinical Terms.

To ensure true semantic interoperability, the OSD format is designed as a valid JSON-LD document. JSON-LD allows a standard JSON file to act as a graph of linked data simply by defining a context. Each definition includes a @context field that explicitly maps the schema properties to universally accepted medical ontologies and vocabularies, such as Schema.org [36], BioSchemas [37], and the Human Phenotype Ontology (HPO) [38]. This transformation gives explicit semantic meaning to each string field, allowing automated systems to interpret the exact biological or epidemiological concept behind the text [39]. Furthermore, to guarantee long-term machine readability, every generated JSON file includes a $schema property pointing to a persistent, community-governed uniform resource identifier hosted via the W3C permanent identifier infrastructure (eg, [40]). This approach provides a stable reference to the validation schema regardless of future repository migrations.

To capture the logical nature of the collected definitions, we implemented a flexible attribute-operator-value pattern. This pattern can accommodate diverse clinical observations ranging from simple Boolean conditions to complex pattern matching through regular expressions. The format supports numerical comparisons (>, >=, <, <=, ==, and !=) and text pattern matching (eg, regular expressions and through regex), enabling precise representation of quantitative thresholds and textual patterns. Please note that logical operators are available at the criteria level (see the Schema section). To handle semantic mapping, the format provides two specialized properties: ontology_id is used to represent compacted uniform resource identifiers for broad medical concepts (eg, HPO or MONDO Disease Ontology [MONDO] identifiers), while the code object is used to capture specific, system-agnostic clinical billing or diagnostic codes from systems like *ICD* (*International Statistical Classification of Diseases and Related Health Problems*) [41] or SNOMED CT (Systematized Nomenclature of Medicine—Clinical Terms) [42].

The criterion structure can be defined recursively. The values property enables criteria to be nested within other criteria. This

XSL•FO
RenderX

design allows us to represent complex logical relationships, ranging from simple symptom lists to intricate decision trees with multiple levels of criteria. With this property structure, we strike a balance between human readability and machine interpretability. We preserve the clinical logic of narrative case definitions while enabling computational processing and analysis.

OSD Format

Our schema was developed iteratively, resulting in the creation of the OSD format. We designed the format to include both inclusion and exclusion criteria, reflecting common patterns found in various case definitions. We extracted essential metadata components, such as location, publication date, title, authors, and citation details, from the publishing websites and accompanying papers. This framework for metadata ensures proper attribution and provides contextual information for each definition.

We manually adapted the case definition texts into JSON format and used validation tools to verify their structural integrity and compliance. Transforming the texts from human-readable to machine-readable format revealed additional patterns and edge cases, which informed iterative refinements to the format. Throughout this iterative development process, we prioritized flexibility while maintaining structural consistency. This allowed the format to represent definitions across varied health systems, geographical regions, and clinical contexts.

Tooling

As we developed the OSD format and dataset, we realized that supporting tools were necessary to facilitate the conversion of human-readable case definitions into machine-readable formats. We developed a Python-based toolkit to streamline this process, enabling researchers and public health professionals to efficiently translate traditional text-based case definitions into structured JSON representations, and vice versa. The toolkit used Python version 3.11 (Python Software Foundation).

We implemented a conversion utility that leverages large language models (LLMs) to automate the structuring process. To accommodate different user preferences and computational resources, the tool supports local, privacy-preserving deployment via Ollama [43] (version 0.9.6), as well as direct application programming interface (API) integration with major cloud providers, including OpenAI, Anthropic, Google Gemini, Mistral, and DeepSeek. We systematically evaluated multiple local models, including llama-3.2 (Meta) [44], mistral-7b (Mistral AI) [45], and deepseek-r1 (DeepSeek; both 7b and 8b variants) [46], medllama2 (Siraj Raval) [47], and qwen2.5-coder (Alibaba Cloud) [48]. We evaluated the performance of the models based on their ability to comprehend medical terminology accurately and remain faithful to the original clinical meaning. Of the models evaluated, llama-3.2, mistral-7b, and deepseek-r1 demonstrated adequate performance for the conversion task. Mistral-7b was selected as the recommended default due to its optimal balance of accuracy and resource requirements.

Additionally, we developed a reverse conversion function that transformed machine-readable JSON syndrome definitions into human-readable formats that support multiple languages. This bidirectional conversion capability ensures accessibility for diverse user groups and facilitates international collaboration in syndromic surveillance systems.

To evaluate the structural and semantic interoperability of the OSD format, we developed a data filtering prototype designed to process tabular health datasets against our standardized JSON-LD schema [49].

OSI

To promote the adoption and ongoing development of the OSD format, we created the OSI, a collaborative community platform. The OSI serves as the central hub for all OSD-related resources and provides an infrastructure for sharing, collaboration, and knowledge exchange. We cover the methodology behind the website and its functionalities in the following text.

First, to ensure transparency and promote collaboration, we developed the OSI website using open-source tools and libraries. We use Hugo [50] (version v0.145.0) to create the static website, which is built using Markdown [51] files. This allows other community members to easily improve or expand the website content with new blog posts or tutorials.

Second, the website provides a user-friendly contribution workflow for submitting new definitions. This functionality is enabled through pull requests or a simplified web interface, as depicted in Figure 3.

All contributions undergo a community review and are published on our website after receiving approval and technical validation. When a definition is submitted via the form, one of the maintainers will create a GitHub pull request on the user's behalf in our repository. Despite their technical skill level, the user can view the submission publicly available in our repository and provide feedback. This process helps us maintain a clear history of contributions and ensures that all definitions are version-controlled. The submission then follows the same process as any other pull request, including validation and community review. We developed the entire toolkit as open-source software and hosted it on GitHub to enable collaborative development and continuous improvement. To ensure quality and authenticity, we verify the institutional affiliations of first-time contributors.

We implemented the automated validation workflows using GitHub Actions to maintain quality and consistency. These workflows verify JSON schema compliance for each definition and validate new submissions against established formatting requirements. The workflow verifies the conformity of submitted JSON files with a predefined schema. This process ensures that the JSON structure is well-formed, that all required fields are present and correctly named according to the schema specifications, and that the values adhere to the expected data formats (eg, URLs and dates). We open-sourced the entire toolkit and hosted it on GitHub to enable community contributions and continuous improvement through collaborative development. Our workflow is straightforward: we use a GitHub repository as the central hub where all definitions are stored and managed. Anyone can submit a new definition through the submission contact form or directly via a GitHub pull request.

XSL•FO
RenderX

Additionally, individuals can verify their affiliations and become recognized contributors. To do so, they must provide a valid email address and their organization's name through the verification form. Definitions submitted by verified contributors receive a verified tag.

Third, the website features an interactive graph visualization of the definitions dataset that we developed using D3 [52] (version v7.9.0), an open-source JavaScript library for visualizing data, and D3 force-directed graph layout using velocity Verlet integration. This feature allows users to explore the dataset intuitively by panning, zooming, and viewing tooltips that display important information about each criterion. The visualization automatically updates to reflect changes in the underlying dataset, so users always see the most current information.

Figure 3. Open Syndrome Initiative contribution workflow. PR: pull request.

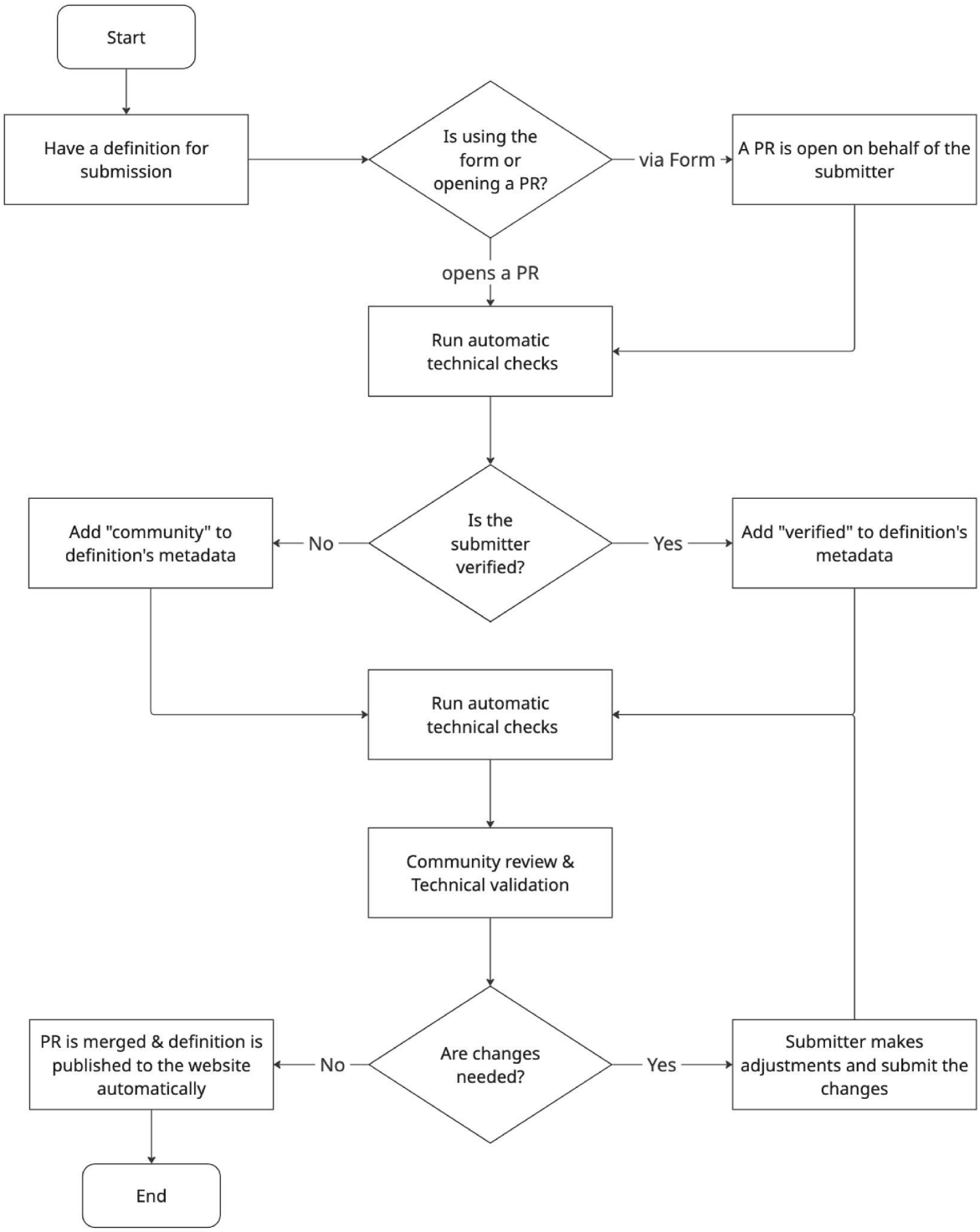

XSL•FO
RenderX

Fourth and last, the web interface includes extra documentation with instructions and suggestions on how to read, work with, and upload case definitions. For those unfamiliar with the format, we provide extensive instructions on how to convert text-based definitions to the OSI format using our tooling.

Ethical Considerations

This study was exempt from institutional ethics review as it exclusively uses publicly available case definitions published by governmental and public health organizations. No patient data, personal information, or human subjects were involved.

Results

Overview

This section examines the structure and characteristics of the case definition format. It introduces the first case definition dataset and its maintenance tools. Finally, it presents the OSI, a collaborative community built around this ecosystem.

The OSD Format

The OSD format was designed to accurately preserve the clinical meaning of case definitions in a machine-readable representation. Examining various case definitions reveals that, despite differences in language and style, they consistently strive to be clear, simple, and concise [1,53]. We have also deliberately incorporated these essential qualities into our format.

The OSD format is a JSON schema, a standard for defining the structure and rules of JSON data. With a JSON schema, one can describe the overall structure of a definition, including its inclusion and exclusion criteria, metadata, properties, and more. The OSD format transforms traditional narrative case definitions into structured, machine-readable representations. The version used for the proposed OSD format is draft 2020 - 12 [54]. For example, a free-text definition for influenza-like illness (right side of Figure 2) becomes a structured JSON object (left side of Figure 2) specifying each criterion with precise logical relationships and the metadata of this definition.

This structured format explicitly defines signs, quantitative attributes, and temporal constraints, which could be ambiguous in a free-text narrative. Additionally, it incorporates metadata, such as provenance and version history, which is often missing from traditional formats. This structured approach enables computational systems to accurately interpret case definitions, thereby facilitating automated case classification, cross-jurisdictional comparison, and application in ML pipelines. Several core principles guided our design decisions.

1. Preservation of clinical meaning: Maintaining the integrity of case definitions while keeping a structured format and preserving essential clinical details.
2. AI Readiness: Designing with computational processing in mind to enable the large-scale analysis and application of AI.
3. Interoperability: Ensuring seamless integration with existing and future systems across different platforms, following the Findability, Accessibility, Interoperability, and Reusability (FAIR) principles [55] and promoting interoperability [56].
4. Openness and Accessibility: Developing a reusable, open format that is free to use and not tied to proprietary platforms.
5. Decentralization: Allowing for independent implementation across websites, scientific publications, and surveillance systems without central control.
6. Versionability: Supporting the evolution of definitions over time with clear tracking of changes.

The format addresses several critical challenges in the current landscape. These challenges include a lack of standardization in case definitions, potential ambiguity in text-based definitions, fragmented information caused by inconsistent metadata, and barriers to large-scale AI applications resulting from format inconsistencies. Even though the technical nature of the format might require some knowledge of JSON, we alleviate this by providing user-friendly tools described in the Tooling section.

This format has several promising applications, including testing definitions against electronic health record data, creating ML models for automated case detection, comparing disease definitions across countries, and enhancing reproducibility in epidemiological research.

While our collection of data effectively documented a wide array of diseases and methods of definition, we recognize the necessity for subsequent improvements. To that end, we included a version field in the format specification. This allows us to maintain backward compatibility and accommodate future improvements as the field evolves. This versioning approach allows the OSD format to adapt to emerging needs while preserving access to historical definitions. The resulting format balances machine readability with a faithful representation of the original clinical intent. This enables automated processing while preserving the essential diagnostic criteria established by public health authorities.

Definitions Dataset

The first comprehensive dataset of machine-readable case definitions was developed, including definitions from 60 countries, 3 continental organizations, and one global organization. The dataset contains 40 case definitions for various diseases and is available in 3 formats: OSD JSON, plain text, and the original PDF publications. The definitions cover a wide range of disease categories: vector-borne illnesses (n=8), viral respiratory diseases (n=7), bacterial diseases, vaccine-preventable diseases (n=6 each), gastrointestinal diseases (n=5), hemorrhagic fevers (n=4), nervous system and other infectious diseases (n=2 each), and climate-related and noninfectious conditions (n=1 each).

To quantify the structural characteristics of the collected data, we assessed the logical complexity of the definitions within our dataset (Figure 4). Regarding logical operator usage, AND was the most frequent operator (n=44), followed by OR (n=33) and AT_LEAST (n=28). Regarding maximum nesting depth, most definitions had a depth of 3 (n=21), followed by depth 2 (n=11), depth 4 or more (n=7), and depth 1 (n=1). The analysis revealed that the majority of the definitions use up to three levels of nesting depth to represent their clinical criteria.

XSL•FO
RenderX

Measles is the most represented disease, with 4 definitions. Cholera, influenza-like illness, and COVID-19 each have two definitions. This likely reflects their status as priority conditions in national and international public health surveillance initiatives, in which consistent case definitions are crucial for monitoring morbidity and responding to outbreaks. The remaining 30 diseases, ranging from common conditions like dengue fever to rare diseases like Lujo hemorrhagic fever, are each represented by one definition.

Figure 4. Distribution of maximum nesting depth and the frequency of logical operators in the machine-readable dataset definitions.

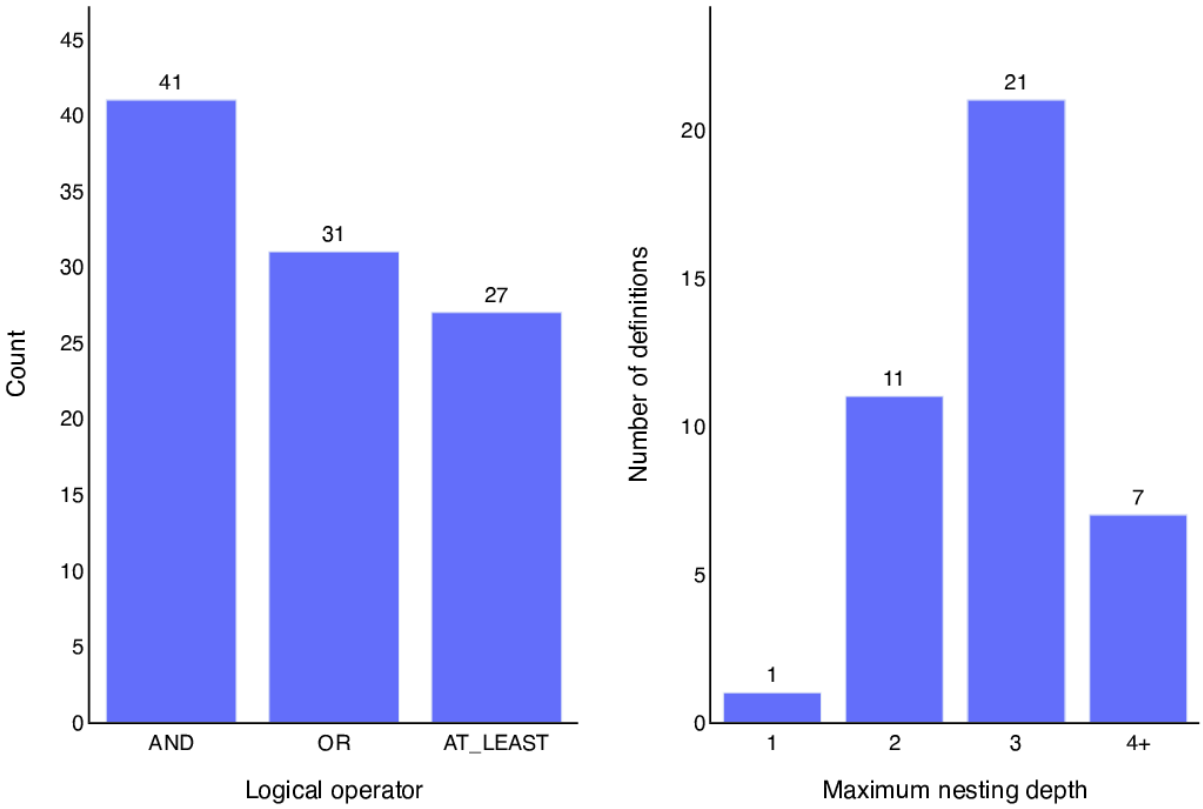

Additionally, we ensured geographic diversity by selecting at least 10% of countries from each continent, based on the number of countries represented in the United Nations, to guarantee diversity in geographic representation. For example, the Pacific Islands are represented collectively through the PPHSN. This collaborative effort involves 22 Pacific Island countries and territories that develop case definitions together. Although using English for our search terms may have introduced some bias, the definitions were available in a range of languages, as depicted in Figure 5.

XSL•FO
RenderX

Figure 5. Distribution of definitions by location and language.

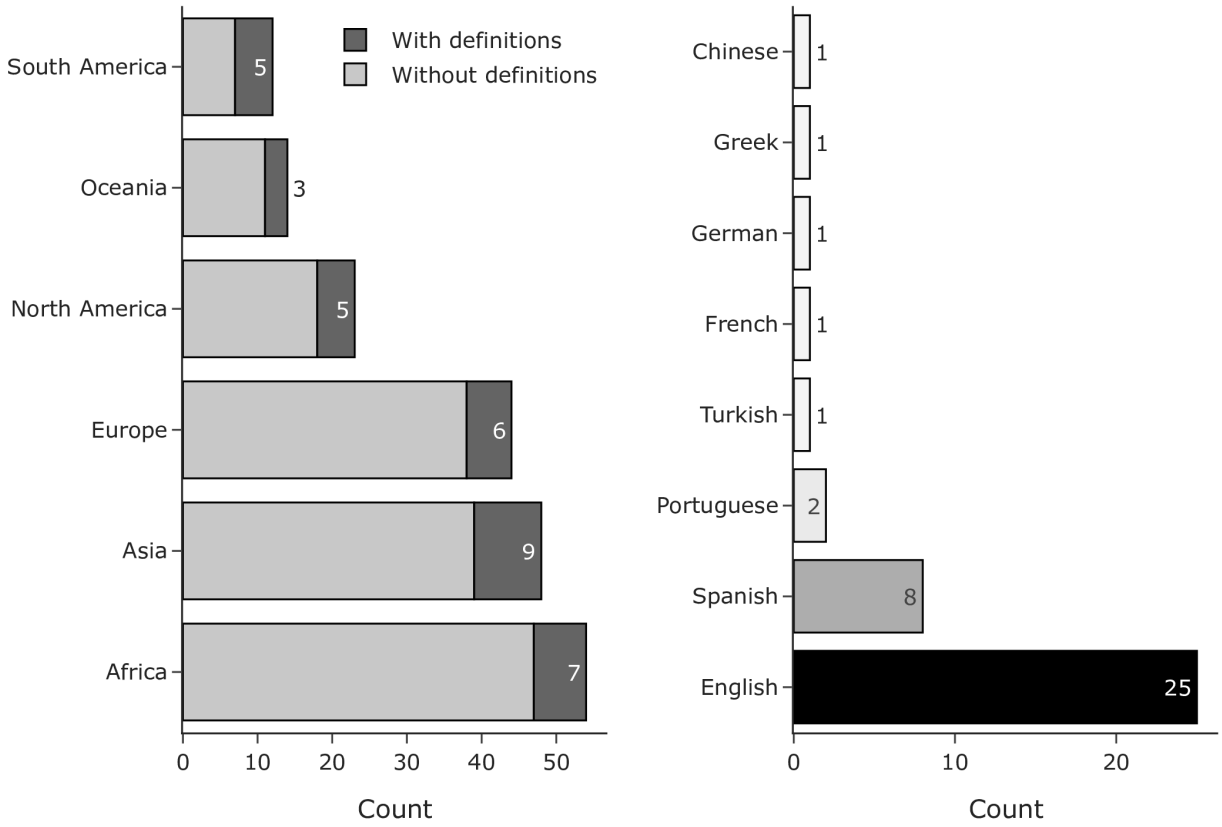

Data Filtering Prototype

To demonstrate the practical utility and downstream applicability, we successfully implemented a data filtering prototype as a proof-of-concept. This web-based tool, publicly accessible at the Open Syndrome Initiative web page [57] (see Figure 6), enables users to upload their own tabular health datasets and apply our standardized case definitions to automatically identify matching patient records.

By leveraging the newly integrated JSON-LD schema, which maps definition criteria to established medical ontologies (eg, HPO and Disease Ontology), the prototype ensures semantic consistency even when processing data from diverse health systems.

This successful implementation validates the structural and semantic interoperability of our format, showcasing its capacity to facilitate automated, cross-platform epidemiological surveillance without ambiguity.

XSL•FO
RenderX

Figure 6. The data filtering prototype allows users to upload a sample of their own data or interact with a toy dataset to see which cases are matched with Open Syndrome definitions available on the website.

XSL•FO
RenderX

Tooling

To overcome barriers to adopting our machine-readable case definition format, we developed a Python command-line tool leveraging LLMs via Ollama to convert between traditional text definitions and our structured format [43]. The library offers two main features: it automatically generates structured JSON representations from case definitions and converts JSON definitions back to human-readable text for verification and sharing.

The tool reliably converts between human- and machine-readable formats using either locally hosted LLMs (via Ollama) or prominent cloud-based models. We do not provide a quantitative evaluation of the tool, as this lies beyond the scope of the current work and reflects its auxiliary role in our study. However, our qualitative analysis indicates that Mistral [45] yields the most reliable results. The tool demonstrates strong fidelity, avoiding hallucinations and accurately translating text into a machine-readable format. This qualitative assessment was conducted by two coauthors (APGF and AA). The library was also designed to be flexible, offering two distinct customization options for the conversion process.

First, users can select their preferred LLM model for conversion. Second, users can specify their preferred output language, enabling multilingual text generation, which is essential for the international adoption of public health systems. By incorporating this flexibility into our library, we ensure local usability and minimize dependencies.

We made the tool available through the Python Package Index under the namespace *opensyndrome* and the project's GitHub repository, enabling integration into existing workflows. We aimed to bridge the gap between technical and nontechnical users by providing extensive documentation on the format, library, and database. This makes the OSD format and its accompanying tools and data accessible to health workers in organizations without dedicated data teams.

OSI

The OSI is a central platform for sharing and accessing case definitions. The initiative's ecosystem facilitates the practical application of standardized case definitions in public health surveillance. It does so by providing a website where users can search, browse, and analyze definitions. The website displays and indexes definitions from the OSI GitHub repository, offering the most current versions and statistical information about the definitions. Users can contribute definitions or suggest modifications via GitHub pull requests or an online form. This dual approach ensures participation from technical and nontechnical users alike. To ensure high-quality dataset definitions, we implemented a two-stage validation process that combines automated static checks with manual semantic verification upon uploading new definitions. The first stage detects missing required fields, formatting inconsistencies, and invalid property types and provides immediate feedback for correction. The second stage involves manually reviewing each definition following the automated checks to semantically validate the uploaded data and ensure its accuracy. The OSI also provides supplementary resources, such as documentation and educational blog posts on best practices.

These resources [58-60], alongside the technical infrastructure, support a community of practice around standardized syndrome definitions and enhance global health surveillance capabilities.

Discussion

Principal Findings

This work introduces the OSD format, a novel approach to representing case definitions in machine-readable form. Alongside this format, we presented the first comprehensive dataset of case definitions and developed supporting tools through the OSI to foster a community of practitioners. These contributions address a significant gap in public health surveillance infrastructure, as standardized digital representations of case definitions have been notably absent.

Our findings demonstrate that the OSD format is flexible enough to accommodate case and syndrome definitions in various clinical contexts. The format's nested structure effectively captures the logical relationships present in traditional case definitions. Based on our curated dataset of structured definitions, we found that 53% (217/407) of the definitions primarily rely on symptom-based criteria. We assessed the logical complexity of each definition by analyzing the depth of nested Boolean expressions and observed that, although logical groupings were common, nesting rarely exceeded three levels. This suggests that complex clinical conditions can be represented within relatively constrained logical structures.

The development process revealed important insights about case definition structures internationally. The predominance of symptom-based criteria across geographical regions and disease categories indicates a universal approach to case definitions, despite variations in health care systems and resources. Additionally, the recurring use of logical operators across definitions from different origins indicates a natural convergence toward structured diagnostic thinking, which our format explicitly codifies.

Despite these promising results, our work has several limitations that warrant consideration. First, while the current dataset is diverse, it is relatively small, containing only 40 case definitions. Although our selective sampling technique covered all continents and major health care systems, the limited sample size may not capture the full range of definitions used globally. A further limitation of the current study lies in the evaluation of semantic fidelity during the conversion process. At this stage, the high semantic fidelity of the machine-readable definitions was assessed qualitatively through manual review by two domain experts. While this human-in-the-loop approach established a highly accurate baseline and ensured that clinical intent was preserved, it is inherently resource-intensive and difficult to scale. Furthermore, the lack of established quantitative benchmarks or baselines in this specific field makes comparative evaluation difficult.

We recognize that this initial, one-of-a-kind dataset will serve as a solid foundation for work ultimately developed by and for

XSL•FO
RenderX

the community. This is why we developed additional tools to ensure a streamlined, user-friendly expansion process accessible to nonexperts. Additionally, the multilingual nature of our source definitions enriches the dataset's diversity but introduces complexity in accurately translating clinical concepts across language barriers. To address this issue, we have included a versioning field in the format of our open-source repositories. This enables the community to propose modifications or corrections for any errors or omissions. Furthermore, the lack of established benchmarks or baselines in this field makes comparative evaluation difficult.

Nevertheless, our work establishes the OSD format at the intersection of significant developments in public health informatics. Its JSON-based structure allows for seamless integration with AI and ML projects, enabling researchers to use these standardized definitions when developing and validating algorithms. As Wang et al [61] noted in their work on automating case identification, structured case definitions can significantly improve the efficiency of surveillance systems. Similar applications have demonstrated that standardized case definitions enhance the assessment of information while alleviating the burden on physicians and clinic managers [62-65].

Our approach's interoperability means that any software capable of parsing JSON can work with these definitions, significantly lowering technical barriers to adoption. This universal accessibility promotes international collaboration and knowledge sharing, enabling public health officials to quickly adopt and implement definitions from other regions. This capability could substantially accelerate the establishment of effective surveillance systems during disease outbreaks, especially in settings with limited resources. While established health care interoperability standards, such as Fast Healthcare Interoperability Resources and Observational Medical Outcomes Partnership Common Data Model, provide comprehensive frameworks for structuring broad medical information, they often require substantial infrastructural investment, specialized technical expertise, and complex data mapping. These requirements can present significant barriers to entry, particularly for resource-constrained public health departments or during rapid, localized outbreak responses. In contrast, the OSD format is purposefully designed to be lightweight and accessible. By using a straightforward JSON-LD schema focused exclusively on epidemiological case definitions, it allows health workers and researchers to standardize their criteria immediately without needing to overhaul their existing IT systems. Table 4 provides a structured comparison across key dimensions. The OSD format is not intended to replace these robust electronic health record standards; rather, it serves as an agile, domain-specific complement. It dramatically lowers the technical barrier to entry for epidemiologists while retaining the structural capacity to be mapped to Fast Healthcare Interoperability Resources (eg, Measure or PlanDefinition) for downstream institutional integration.

By grounding the OSD format in semantic web principles, our work directly connects to and builds upon large-scale existing efforts aimed at making medical diagnoses machine-readable. Initiatives such as the representation of rare diseases in BioSchemas and the MedicalCondition and MedicalCause classes in Schema.org have demonstrated the vast potential of linked data in health care. The OSD format complements these efforts by providing a specialized, deeply nested structural representation specifically tailored for public health surveillance and early warning systems. Future work will explore formally suggesting an epidemiological "MedicalCase" or "Criterion" class extension to the BioSchemas community, thereby contributing our public health modeling back to the broader semantic web ecosystem.

Looking ahead, the OSD format offers several promising avenues for future development. Users will be able to download definitions from various global sources and estimate the number of potential cases they could identify in their local data. This feature will speed up the development and adoption of locally optimized definitions and facilitate the implementation of ML models that leverage this standardized format. As more institutions contribute to and adopt this approach, we anticipate accelerated collaboration across borders and more responsive surveillance capabilities worldwide.

While our present efforts are concentrated on case definitions, there is also considerable potential to broaden the ready implementation of our methodology to syndromic surveillance. The principle of syndromic surveillance is to monitor patients presenting with symptoms, chief complaints, or other nonlaboratory or confirmatory diagnoses. Public health experts coordinating syndromic surveillance map individual clinical codes to more generic syndromic indicators. While these are not case definitions, as they are purely data- and code-based, similar problems exist with the consistency of syndromic indicators internationally.

Using OSD to present the code mappings that underpin syndromic indicators and to standardize these indicators internationally would support the global effort to coordinate syndromic surveillance between countries.

The versioning system in our format makes it easy to adapt to changing needs while also ensuring that older versions remain compatible. This addresses a crucial requirement for sustainable public health infrastructure. As the OSI expands its community of practitioners, we anticipate that the collective expertise will further refine the format and available definitions, thereby creating a robust ecosystem of standardized syndromic surveillance resources accessible to all.

In conclusion, the OSD format represents a significant step toward standardizing and digitizing case definitions for public health surveillance. It establishes a foundation for enhanced interoperability, collaboration, and automation in syndromic surveillance. The format's potential applications extend beyond traditional public health contexts into clinical research, healthcare delivery, and emerging disease response.

XSL•FO
RenderX

Table . Comparison of Open Syndrome definition, CQL[a] or FHIR[b], and OHDSI[c] Phenotype Library across key dimensions. CQL and the OHDSI Phenotype Library are mature, platform-coupled standards optimized for clinical care and retrospective research, respectively. The Open Syndrome definition distinguishes itself through minimal infrastructure requirements, a decoupled definition-execution architecture, and native support for cross-country interoperability and AI pipelines.

Dimension	CQL or FHIR	OHDSI phenotype library	Open Syndrome definition
Required infrastructure	FHIR-compatible EHR[d] with CQL engine	OMOP[e] CDM[f] with ATLAS	JSON[g] parser; optional Python package (opensyndrome)
Format and execution	Domain-specific language; coupled to CQL engine	Platform-generated JSON; executed via HADES[h] (R)	JSON-LD[i]; decoupled from execution pipeline
Ontology support	SNOMED[j], LOINC[k], RxNorm via FHIR bindings	OMOP Vocabulary (Athena); locked to CDM	Agnostic; reuses HPO[l], MONDO[m], *ICD*[n], SNOMED via JSON-LD
Cross-country interoperability	Depends on FHIR adoption	Depends on CDM adoption	Primary design goal; explicit provenance and location metadata
Multilingual support	Partial	No	Yes
ML[o] or AI pipeline integration	Indirect via structured output	Via R with HADES package	Structured for direct ingestion into ML pipelines

[a]CQL: Clinical Quality Language.
[b]FHIR: Fast Healthcare Interoperability Resources.
[c]OHDSI: Observational Health Data Sciences and Informatics.
[d]EHR: electronic health record.
[e]OMOP: observational medical outcomes partnership.
[f]CDM: Common Data Model
[g]JSON: JavaScript object notation.
[h]HADES: Health-Analytics Data to Evidence Suite.
[i]LD: linked data.
[j]SNOMED: Systematized Nomenclature of Medicine – Clinical Terms.
[k]LOINC: Logical Observation Identifiers Names and Codes.
[l]HPO: human phenotype ontology.
[m]MONDO: MONDO disease ontology.
[n]*ICD*: *International Statistical Classification of Diseases and Related Health Problems*.
[o]ML: machine learning.

Future Work

While our current dataset provides a robust baseline for validating the OSD format, we recognize the potential for scaling the repository. A major technical challenge remains the heterogeneity of source PDFs, which makes fully automated bulk extraction prone to errors. To address this, future development will introduce specialized extraction skills within our tooling. These modules will use a 2-step pipeline to first identify multiple conditions within a single document and then structure them, keeping a human in the loop for quality assurance. To enhance the quality assessment process, we intend to adopt automated validation frameworks, such as leveraging advanced LLMs as evaluators (LLM-as-a-judge), to methodically and quantitatively evaluate semantic fidelity in conjunction with our community-driven review process. Furthermore, as an open-source initiative, the platform is now fully equipped to receive community contributions, allowing the dataset to grow organically and incorporate definitions discovered through advanced search methodologies.

To enhance the semantic precision of this extraction process, we plan to implement deterministic mapping to standardized clinical codes. While LLMs excel at generating the initial structural hierarchy, assigning precise ontological codes requires a more rigid approach. Our tooling architecture will be updated to integrate with biomedical ontology repositories, such as the National Center for Biomedical Ontology BioPortal API. Through this integration, terms extracted by the LLM can be automatically queried against standard vocabularies (eg, SNOMED CT or *ICD-10* [*International Statistical Classification of Diseases and Related Health Problems Tenth Revision*]) to retrieve the most accurate concept identifiers, which are then validated by human experts and populated into the format's code fields.

Additionally, to maximize the utility of the OSI platform, we intend to develop an interactive comparative tool. As the repository grows, this feature will allow researchers and public health officials to visually and logically compare case definitions for the same disease across different jurisdictions. This will facilitate the identification of surveillance discrepancies and support the harmonization of global syndromic indicators.

Finally, a significant future development is improving the AI-driven natural language processing techniques that transform

XSL•FO
RenderX

narrative definitions into OSD-format representations. The local LLMs of the current system, which are accessed through Ollama, could be supplemented with BioBERT [66] or ClinicalBERT [67]. Additionally, the Structured Prompt Interrogation and Recursive Extraction of Semantics (SPIRES) framework [68], which combines zero-shot LLM prompting with deterministic ontology grounding, could complement our current extraction pipeline by improving identifier precision for HPO and MONDO terms. These models are specifically trained on multilingual and multicultural clinical datasets, which could improve parsing precision and flexibility. Furthermore, AI anomaly detection techniques (eg, isolation forest and transformer-based systems) could be used to automate the validation and correction of definitions, thereby improving the continuous quality of the dataset. AI-federated approaches would also enable collaborative algorithm development across jurisdictions while ensuring data confidentiality, which is essential for international disease monitoring and surveillance. Public health–federated learning frameworks, such as Flower [69] and PySyft [70], could train and evaluate OSD format–based ML models across regions without sharing sensitive health information. This would improve cross-border analysis, equity, fairness, and public health. OSD format may advance from a defined structure for collaborative model development by progressing in this way.

Conclusions

The OSD provides a flexible, interoperable, and machine-readable representation of case definitions. This innovation enables public health professionals, researchers, and technically proficient individuals to exchange epidemiological information more efficiently and consistently. It also enables the broader application of AI and machine learning techniques to public health data. Along with the format, this work introduces the first dataset of case definitions available in both human- and machine-readable formats. The dataset demonstrates the adaptability of the OSD format across various diseases and geographic settings, paving the way for global comparative analyses of case definition methodologies. To promote collaboration, we have launched the OSI, a platform where users can share definitions and access tools for converting between human- and machine-readable formats. Public health is inherently collaborative, and the OSD format contributes to this shared effort by promoting better data and more effective tools for preparing for and preventing public health threats.

Acknowledgments

AJE is affiliated with the National Institute for Health and Care Research Health Protection Research Unit in Emergency Preparedness and Response at the University of Birmingham, as well as the National Institute for Health and Care Research Health Protection Research Unit in Gastrointestinal Infections at the University of East Anglia. The views expressed in this article are those of the authors and do not necessarily reflect the views of the National Institute for Health and Care Research, UK Health Security Agency, or the Department of Health and Social Care.

Funding

The authors declared no financial support was received for this work.

Data Availability

All resources associated with this work are publicly available. The schema, case definitions, official website [59], and the Python package are hosted under the Open Syndrome Initiative GitHub organization [40]. The dataset of definitions, including JSON, TXT, and PDF files, is available on Hugging Face [60]. The Python package can also be accessed via Python Package Index [58].

Authors' Contributions

APGF: Conceptualization, data curation, formal analysis, investigation, methodology, validation, visualization, writing - original draft.
AA: Conceptualization, visualization, writing - review & editing.
IM: Discussion, review, and editing.
HH: Discussion and review.
AJE: Discussion, review, writing and editing.
JDK: Discussion, review, writing and editing.
MS: Discussion, review and editing.
AU: Discussion, review and editing.
GH: Conceptualization, investigation, supervision, writing - review and editing.

Conflicts of Interest

None declared.

References

1. Bassil K. Case definition. In: Encyclopedia of Epidemiology: SAGE Publications, Inc; 2008, Vol. 2:133-134. [doi: 10.4135/9781412953948]

XSL•FO
RenderX

2. Devlin RK, Andrews MM, von Reyn CF. Recent trends in infective endocarditis: influence of case definitions. Curr Opin Cardiol 2004 Mar;19(2):134-139. [doi: 10.1097/00001573-200403000-00012] [Medline: 15075740]
3. Lazarus R, Kleinman KP, Dashevsky I, DeMaria A, Platt R. Using automated medical records for rapid identification of illness syndromes (syndromic surveillance): the example of lower respiratory infection. BMC Public Health 2001;1(1):9. [doi: 10.1186/1471-2458-1-9] [Medline: 11722798]
4. Chapman WW, Dowling JN, Baer A, et al. Developing syndrome definitions based on consensus and current use. J Am Med Inform Assoc 2010;17(5):595-601. [doi: 10.1136/jamia.2010.003210] [Medline: 20819870]
5. Purdy S, Griffin T, Salisbury C, Sharp D. Ambulatory care sensitive conditions: terminology and disease coding need to be more specific to aid policy makers and clinicians. Public Health (Fairfax) 2009 Feb;123(2):169-173. [doi: 10.1016/j.puhe.2008.11.001] [Medline: 19144363]
6. EU case definitions. European Centre for Disease Prevention and Control. URL: https://www.ecdc.europa.eu/en/all-topics/eu-case-definitions [accessed 2026-05-22]
7. India case definitions - P form Case Definitions 2024. URL: https://web.archive.org/web/20240811013244/https://ihip.mohfw.gov.in/idsp/download?fileid=38 [accessed 2026-06-10]
8. Sanchez Ruiz MA, Marques DF, Lomholt FK, et al. Surveillance of severe acute respiratory infections associated with SARS-CoV-2, influenza virus and RSV using ICD-10 codes: a case definition accuracy study across five European countries, 2021 to 2023. Euro Surveill 2025 Jul;30(27):2400748. [doi: 10.2807/1560-7917.ES.2025.30.27.2400748] [Medline: 40642769]
9. Li I, Pan J, Goldwasser J, et al. Neural natural language processing for unstructured data in electronic health records: a review. Comput Sci Rev 2021:2107. [doi: 10.1016/j.cosrev.2022.100511]
10. Wang N, Leem SG. Modernizing data quality: evolving dimensions for unstructured text data in big data, AI and ethical contexts. In: Ventura S, Luna JM, Martín-Castaño ARM, editors. Data Quality Matters - Best Practices for Integrity and Assurance: IntechOpen. [doi: 10.5772/intechopen.1013232]
11. Gomes Ferreira AP, Anžel A, Ullrich A, Hattab G. Advocating the potential of artificial intelligence for syndrome discovery in syndromic surveillance systems: a scoping review. iScience 2026 Mar 20;29(3):115103. [doi: 10.1016/j.isci.2026.115103] [Medline: 41867626]
12. Koh D, Cunningham AC. Counting coronavirus disease 2019 (COVID-19) cases: case definitions, screened populations and testing techniques matter. Ann Acad Med Singap 2020 Mar;49(3):161-165. [doi: 10.47102/annals-acadmedsg.202038] [Medline: 32301478]
13. Guérin PJ, McLean ARD, Rashan S, et al. Definitions matter: heterogeneity of COVID-19 disease severity criteria and incomplete reporting compromise meta-analysis. PLOS Glob Public Health 2022;2(7):e0000561. [doi: 10.1371/journal.pgph.0000561] [Medline: 36962738]
14. Tsang TK, Wu P, Lin Y, Lau EHY, Leung GM, Cowling BJ. Effect of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China: a modelling study. Lancet Public Health 2020 May;5(5):e289-e296. [doi: 10.1016/S2468-2667(20)30089-X] [Medline: 32330458]
15. Badker R, Miller K, Pardee C, et al. Challenges in reported COVID-19 data: best practices and recommendations for future epidemics. BMJ Glob Health 2021 May;6(5):e005542. [doi: 10.1136/bmjgh-2021-005542] [Medline: 33958393]
16. Raveendran AV, Joshi SR. Evolution of a pandemic and changing concepts in the case definition. J Assoc Physicians India 2024 Aug;72(8):75-77. [doi: 10.59556/japi.72.0606] [Medline: 39163071]
17. Suthar AB, et al. Coronavirus Disease Case Definitions, Diagnostic Testing Criteria, and Surveillance in 25 Countries with Highest Reported Case Counts. 2022 Jan. URL: https://doi.org/10.3201/eid2801.211082 [doi: 10.3201/eid2801.211082]
18. Overview of syndromic surveillance what is syndromic surveillance? Centers for Disease Control and Prevention. URL: https://www.cdc.gov/mmwr/preview/mmwrhtml/su5301a3.htm [accessed 2026-06-10]
19. Krause G, Brodhun B, Altmann D, Claus H, Benzler J. Reliability of case definitions for public health surveillance assessed by Round-Robin test methodology. BMC Public Health 2006 May 10;6:129. [doi: 10.1186/1471-2458-6-129] [Medline: 16686946]
20. Ghodsi Z, Rahimi Movaghar V, Zafarghandi M, et al. The minimum dataset and inclusion criteria for the national trauma registry of Iran: a qualitative study. Arch Trauma Res 2017;6(2):1-7. [doi: 10.5812/atr.39725]
21. Botsis T, Ball R. Automating case definitions using literature-based reasoning. Appl Clin Inform 2013;4(4):515-527. [doi: 10.4338/ACI-2013-04-RA-0028] [Medline: 24454579]
22. Json-ld 1.0: a json-based serialization for linked data. Travesía. 2014. URL: http://hdl.handle.net/10421/7478 [accessed 2026-06-10]
23. Lanthaler M, Gütl C. On using JSON-LD to create evolvable restful services. : ACM; 2012 Apr 17 Presented at: WS-REST '12: Proceedings of the Third International Workshop on RESTful; Apr 17, 2012; Lyon France p. 25-32 URL: https://dl.acm.org/doi/proceedings/10.1145/2307819 [doi: 10.1145/2307819.2307827]
24. Case definitions. Robert Koch Institut. URL: https://www.rki.de/DE/Themen/Infektionskrankheiten/Meldewesen/Falldefinitionen/falldefinitionen-node.html [accessed 2026-05-22]
25. Surveillance case definitions for current and historical conditions. Centers for Disease Control and Prevention. 2025. URL: https://ndc.services.cdc.gov/ [accessed 2026-05-22]

XSL•FO
RenderX

26. Case definitions: nationally notifiable diseases. Government of Canada. 2000. URL: https://diseases.canada.ca/notifiable/diseases-list [accessed 2026-06-10]
27. Collaboration is advancing science and practice of public health. Knowledge Repository. URL: https://knowledgerepository.syndromicsurveillance.org [accessed 2026-05-22]
28. PubMed. US National Library of Medicine. URL: https://pubmed.ncbi.nlm.nih.gov/ [accessed 2026-06-12]
29. OurResearch. Openalex: a fully open catalog of the global research system. 2025 URL: https://openalex.org/ [accessed 2026-05-13]
30. Kaggle: your machine learning and data science community. Kaggle Inc. 2025. URL: https://www.kaggle.com/ [accessed 2026-06-10]
31. Hugging face – the ai community building the future. Hugging Face. 2025. URL: https://huggingface.co/ [accessed 2026-06-10]
32. Harvard Dataverse. Harvard University. 2025. URL: https://data.harvard.edu/dataverse [accessed 2026-06-10]
33. The fastest JSON schema Validator. Supports JSON Schema draft-04/06/07/2019-09/2020-12 and JSON Type Definition (RFC8927). GitHub. URL: https://github.com/ajv-validator/ajv [accessed 2026-06-10]
34. Outbreak toolkit. World Health Organization. URL: https://www.who.int/emergencies/outbreak-toolkit [accessed 2026-05-22]
35. Newman-Griffis D, Divita G, Desmet B, Zirikly A, Rosé CP, Fosler-Lussier E. Ambiguity in medical concept normalization: an analysis of types and coverage in electronic health record datasets. J Am Med Inform Assoc 2021 Mar 1;28(3):516-532. [doi: 10.1093/jamia/ocaa269] [Medline: 33319905]
36. Guha RV, Brickley D, Macbeth S. Schema.org: evolution of structured data on the web. Commun ACM 2016;59:44-51. [doi: 10.1145/2844544]
37. Gray AJ, Goble CA, Jimenez RC, et al. Bioschemas: from potato salad to protein annotation. 2017. URL: https://doi.org/10.7490/f1000research.1114493.1 [accessed 2026-06-10]
38. Köhler S, Gargano M, Matentzoglu N, et al. The Human Phenotype Ontology in 2021. Nucleic Acids Res 2021 Jan 8;49(D1):D1207-D1217. [doi: 10.1093/nar/gkaa1043] [Medline: 33264411]
39. Anjum F, Maqbool F, Razzaq MS, et al. Semantic web ontology for structured knowledge representation and clinical decision support in eye diseases. Sci Rep 2025;15(1):29986. [doi: 10.1038/s41598-025-15885-x]
40. Open Syndrome. GitHub. URL: https://github.com/OpenSyndrome [accessed 2026-05-23]
41. International statistical classification of diseases and related health problems (ICD). World Health Organization. URL: https://www.who.int/standards/classifications/classification-of-diseases [accessed 2026-05-22]
42. Overview of SNOMED CT. National Library of Medicine. URL: https://www.nlm.nih.gov/healthit/snomedct/snomed_overview.html [accessed 2026-05-22]
43. Ollama. 2024. URL: https://ollama.com [accessed 2025-04-21]
44. Touvron H, Thibaut L, Gautier I, et al. LLaMA: open and efficient foundation language models. arXiv. Preprint posted online on Feb 27, 2023. [doi: 10.48550/arXiv.2302.13971]
45. Jiang AQ, et al. Mistral 7B. . Preprint posted online on 2023 URL: https://arxiv.org/abs/2310.06825
46. Guo D, Yang D, Zhang H, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature 2025 Sep;645(8081):633-638. [doi: 10.1038/s41586-025-09422-z] [Medline: 40962978]
47. Medllama2. Ollama. URL: https://ollama.com/library/medllama2 [accessed 2026-06-10]
48. Hui B, et al. Qwen2.5-coder technical report. arXiv. Preprint posted online on Sep 18, 2024. [doi: 10.48550/arXiv.2409.12186]
49. Ferreira APG, Anžel A, Marcilio de IO. The open syndrome definition. GitHub. URL: https://github.com/OpenSyndrome/schema/ [accessed 2026-05-23]
50. The world's fastest framework for building websites. Hugo. URL: https://gohugo.io [accessed 2026-05-22]
51. Markdown. Wikipedia. 2025.
52. The javascript library for bespoke data visualization. D3. URL: https://d3js.org [accessed 2026-05-22]
53. Toolkit for investigation and response to food and waterborne disease outbreaks with an EU dimension. European Centre for Disease Prevention and Control. 2012. URL: https://www.ecdc.europa.eu/en/publications-data/toolkit-investigation-and-response-food-and-waterborne-disease-outbreaks-eu [accessed 2026-05-22]
54. Draft 2020-12. JSON Schema. URL: https://json-schema.org/draft/2020-12 [accessed 2026-05-22]
55. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3(1):160018. [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
56. Hattab G, Irrgang C, Körber N, Kühnert D, Ladewig K. The way forward to embrace artificial intelligence in public health. Am J Public Health 2025 Feb;115(2):123-128. [doi: 10.2105/AJPH.2024.307888] [Medline: 39571129]
57. Live. data browser. Open Syndrome Initiative. URL: https://opensyndrome.org/live/ [accessed 2026-05-23]
58. Open Syndrome 0.3.0. PyPI. URL: https://pypi.org/project/opensyndrome/ [accessed 2026-05-23]
59. Open Syndrome Initiative. URL: https://opensyndrome.org [accessed 2025-05-23]
60. Open Syndrome dataset. Hugging Face. URL: https://huggingface.co/datasets/opensyndrome/case-definitions [accessed 2026-05-23]

XSL•FO
RenderX

61. Wang L, Zipursky AR, Geva A, McMurry AJ, Mandl KD, Miller TA. A computable case definition for patients with SARS-CoV2 testing that occurred outside the hospital. JAMIA Open 2023 Oct;6(3):ooad047. [doi: 10.1093/jamiaopen/ooad047] [Medline: 37425487]
62. Radoï L, Luce D. A review of risk factors for oral cavity cancer: the importance of a standardized case definition. Community Dent Oral Epidemiol 2013 Apr;41(2):97-109. [doi: 10.1111/j.1600-0528.2012.00710.x] [Medline: 22882534]
63. Chan DJ, Donovan B. What's in a word? case definitions in sexual health medicine. Int J STD AIDS 2005 Feb;16(2):91-94. [doi: 10.1258/0956462053057657] [Medline: 15807934]
64. Williamson T, Miyagishima RC, Derochie JD, Drummond N. Manual review of electronic medical records as a reference standard for case definition development: a validation study. CMAJ Open 2017 Dec 11;5(4):E830-E833. [doi: 10.9778/cmajo.20170077] [Medline: 29242256]
65. Collin L, Reisner SL, Tangpricha V, Goodman M. Prevalence of transgender depends on the "Case" definition: a systematic review. J Sex Med 2016 Apr;13(4):613-626. [doi: 10.1016/j.jsxm.2016.02.001] [Medline: 27045261]
66. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
67. Huang K, Altosaar J, Ranganath R. Clinicalbert: modeling clinical notes and predicting hospital readmission. arXiv. Preprint posted online on Apr 10, 2019. [doi: 10.48550/arXiv.1904.05342]
68. Caufield JH, Hegde H, Emonet V, et al. Structured Prompt Interrogation and Recursive Extraction of Semantics (SPIRES): a method for populating knowledge bases using zero-shot learning. Bioinformatics 2024 Mar 4;40(3). [doi: 10.1093/bioinformatics/btae104]
69. Beutel DJ, et al. Flower: a friendly federated learning research framework. arXiv. Preprint posted online on Jul 28, 2020. [doi: 10.48550/arXiv.2007.14390]
70. Ziller A, et al. Pysyft: a library for easy federated learning. In: Federated Learning Systems: Towards next-Generation AI, 111–139: Springer; 2021. [doi: 10.1007/978-3-030-70604-3_5]

Abbreviations

AI: artificial intelligence
API: application programming interface
FAIR: Findability, Accessibility, Interoperability, and Reusability
HPO: Human Phenotype Ontology
ICD: *International Statistical Classification of Diseases and Related Health Problems*
ICD-10: *International Statistical Classification of Diseases and Related Health Problems Tenth Revision*
JSON-LD: JSON for Linked Data
LLM: large language model
ML: machine learning
MONDO: Mondo Disease Ontology
OSD: Open Syndrome Definition
OSI: Open Syndrome Initiative
PPHSN: Pacific Public Health Surveillance Network
SNOMED CT: Systematized Nomenclature of Medicine—Clinical Terms
SPIRES: Structured Prompt Interrogation and Recursive Extraction of Semantics
WHO: World Health Organization

Edited by A Stone; submitted 21.Oct.2025; peer-reviewed by V Emonet, Z Tomasovic; revised version received 29.Apr.2026; accepted 29.Apr.2026; published 18.Jun.2026.

Please cite as:
Ferreira APG, Anžel A, Marcilio I, Hughes H, Elliot AJ, Kong JD, Schranz M, Ullrich A, Hattab G
The Open Syndrome Definition as a Machine-Readable Standard for Public Health: Design and Implementation Study
J Med Internet Res 2026;28:e86249
URL: https://www.jmir.org/2026/1/e86249
doi:10.2196/86249

© Ana Paula Gomes Ferreira, Aleksandar Anžel, Izabel Marcilio, Helen Hughes, Alex J Elliot, Jude Dzevela Kong, Madlen Schranz, Alexander Ullrich, Georges Hattab. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any

XSL•FO
RenderX

medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring Reddit Discourse and Information Needs Surrounding Extreme Heat: Topic, Sentiment, and Engagement Analysis

Melissa MacKay[1], MPH, PhD; Soroush Zamani Alavijeh[2], MSc; Sydney Gosselin[1], MSc; Fattane Zarrinkalam[3], PhD; Jennifer E McWhirter[1], PhD

[1]Department of Population Medicine, University of Guelph, 50 Stone Rd E, Guelph, ON, Canada
[2]Department of Computing and Software, McMaster University, Hamilton, ON, Canada
[3]College of Engineering, University of Guelph, Guelph, ON, Canada

Corresponding Author:
Jennifer E McWhirter, PhD
Department of Population Medicine, University of Guelph, 50 Stone Rd E, Guelph, ON, Canada

Abstract

Background: As Canada's climate changes, extreme heat events have become more frequent, a trend that is expected to continue. Extreme heat can lead to several negative health outcomes, which disproportionately impact vulnerable populations. Evidence-based, equitable interventions are needed to inform and protect the public from the health effects. Effective communication can aid this effort to improve health outcomes by emphasizing the connection between health risks and climate change and empowering people to act. Machine learning has applications in understanding current attitudes, beliefs, experiences, and behaviors within the target audience for public health messaging. Machine learning analyses of social media data have elucidated user perceptions of heat events in the literature; however, research is limited with respect to social media user perceptions, beliefs, and behaviors related to extreme heat, particularly in the Canadian context. Analyzing Canadian social media discourse related to extreme heat will help to address this research gap and inform future research and communications to reduce the risks of extreme heat.

Objective: The purpose of this research is to better understand Canadian discourse and emotions related to extreme heat by examining social media (Reddit). Our objectives include (1) identifying common discussion topics, concerns, and questions related to extreme heat among Canadian Reddit users; (2) analyzing sentiment and emotional responses to extreme heat discussions; and (3) investigating the relationship between topics, sentiment, and engagement for posts.

Methods: We collected data using the Reddit application programming interface (API), retrieving posts from 30 Canada-specific subreddits between February 12, 2023, and February 11, 2024, based on a predefined set of heat- and climate-related keywords. Posts and comments were structured as hierarchical tree models, with text consolidated into documents for analysis. Topic modeling, sentiment analysis, and emotion analysis were conducted; engagement was assessed using net upvote scores to gauge community approval.

Results: The analysis of 607 Reddit posts from 15,366 users revealed that discussions about extreme heat were most frequently centered around the keyword "heat," which appeared in 82.5% (n=501) of the posts and 81.1% (n=25,253) of the comments. Topic analysis identified key themes related to heating and cooling costs, weather records, air conditioning, and health impacts, while sentiment and emotion analyses showed varying levels of positivity and negativity across subreddits.

Conclusions: Our findings present an initial snapshot into Canadian perspectives and information needs about extreme heat in Canada. In our sample, discussions on Reddit about extreme heat in Canada are dominated by concerns over heating and cooling costs, weather patterns, and personal adaptation strategies, reflecting both practical and policy-related challenges. Additionally, sentiment and emotion analyses suggest significant regional differences in public perception, which may be useful for informing health and risk messaging initiatives to better protect Canadians from the adverse health effects of climate change.

(J Med Internet Res 2026;28:e82426) doi:10.2196/82426

KEYWORDS

social media; Reddit; climate change; extreme heat; information ecosystem; health communication

Introduction

Background

Climate change due to human activity has been driving shifts in temperatures and weather patterns [1]. Climate change has many consequences, including fires, flooding, droughts and water scarcity, and intense storms [1]. Because of these impacts, climate change affects health in diverse ways, and some individuals are more vulnerable to these health risks than others [1]. The extent of the impact on individuals is influenced by

XSL•FO
RenderX

their exposure to hazards related to climate change, the degree of impact, and their capacity to cope with the climate hazard [2]. Infants, children, pregnant women, older adults, and persons with physical and mental impairments are more vulnerable to climate-related hazards [2]. Additionally, various social determinants of health, including socioeconomic status, race or racism, health care access, education, and the built environment, interact with climate change to produce patterns of disadvantage that lead to additional vulnerabilities [2].

In Canada, climate change is resulting in extreme weather events, including extreme heat, which has significant direct impacts on human health and the environment [3]. Direct health risks from extreme heat include the exacerbation of existing health conditions, heat exhaustion, heat stroke, and death [4]. Extreme heat can also intersect with the indirect health risks of a changing climate, which are related to the social determinants of health and can impact our food system, culture, housing, infectious disease distribution, water quality and availability, and the economic status of individuals and communities [5]. Indeed, extreme heat is one of the most deadly forms of extreme weather affecting Canadians and is a growing public health risk in Canada [6,7].

Effective communication about climate change–related health risks, including extreme heat, can improve individual and population-level health outcomes [8]. Evidence-based strategies for climate change communication emphasize highlighting health impacts, relying on trusted sources, minimizing barriers to adaptive behaviors, and leveraging social media to disseminate accurate information [8]. To support effective public health messaging, it is essential to understand public risk perceptions as well as their information needs and preferences regarding climate change [8].

Social media provides an important communication channel through which public health can share information and people can share their experiences and opinions [9]. Public health can potentially reach a larger audience, raise awareness and share information about climate-related hazards and issues, and mobilize action. It can also foster public discussion and positively impact attitudes toward climate policy [9,10]. Social media also presents challenges, including the spread of misinformation or disinformation, the creation of echo chambers, and the bias toward certain actors or ideologies because of algorithmic curation [10].

Social media provides a platform for public health professionals to not only reach wider audiences but also to monitor information, assess public awareness, and engage in ongoing discussions. Platforms such as Reddit, which has over 15.55 million users in Canada [11], offer valuable insights into public perceptions and discourse on health-related topics. Canadian users actively participate across subreddits, including r/canada, facilitating country-specific discussions. The large Canadian user base and the diversity of subreddits and topics discussed provide an important social media platform to assess extreme heat and climate-related discussions and trends to understand perceptions, sentiment, and information needs and wants. An analysis of 16 years of Reddit discussion suggests that trends in Reddit discussions are reflective of broader trends across Internet-based channels [12], further supporting the evaluation of Reddit to better understand discourse on climate change.

Although limited research has been conducted to examine climate change discourse on Reddit [13,14], no such research has been conducted with a focus on Canada. This study extends prior research on climate change discourse by offering a Canadian case study and demonstrating the potential of computational methods for examining emerging public concerns about extreme heat. To guide our study, we draw on principles from risk communication theory and the concept of the information ecosystem. Risk communication theory emphasizes the role of trusted channels, perceived risk, and emotional responses in shaping adaptive behaviors during climate hazards [8]. The information ecosystem perspective recognizes social media platforms, such as Reddit, as sites where diverse actors contribute to meaning-making, often shaping how risk information is sought, shared, and acted upon [10]. These perspectives position Reddit not only as a source of descriptive discourse but also as a meaningful part of how Canadians understand, interpret, and respond to extreme heat.

Study Goal and Objectives

The goal of this research is to understand social media (Reddit) discourse and emotions related to extreme heat in Canada. Specific subobjectives of this research are as follows: (1) to identify common discussion topics, questions, and concerns related to extreme heat among Canadian Reddit users; (2) to analyze sentiment and emotional responses to extreme heat discussions; and (3) to investigate the relationship between topics, sentiment, and engagement for posts.

Methods

This study was designed as a descriptive computational analysis, intended to provide an initial overview of Reddit discourse on extreme heat in Canada. As such, the results present descriptive outputs of topic, sentiment, emotion, and engagement analyses without qualitative interpretation.

Ethical Considerations

Research ethics was not required as this study analyzed publicly available data only as per the University of Guelph's Research Ethics Board [15]. Under Canada's Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans, Research Ethics Board review is not required for research that relies exclusively on publicly accessible information where individuals have no reasonable expectation of privacy; as our study analyzes only public Reddit posts without interaction or private identifiers, ethics review is not required [16]. No personally identifiable information or direct identifiers (eg, usernames, links, or IDs) were collected or analyzed as part of this study. No other personally identifiable information was collected or stored, such as user profile information or IP addresses.

Data Source

To collect data for our analysis, we utilized the Reddit API [17] to search for posts within a selection of Canadian subreddits. These searches targeted posts that contained at least one of our

XSL•FO
RenderX

designated keywords in their titles. Specifically, we examined 30 subreddits focused on national interests (eg, r/canada), provincial or territorial interests (eg, r/ontario, r/britishcolumbia), and local interests within provincial or territorial capital cities and most populous cities (if different; eg, r/vancouver, r/halifax, r/saskatoon). The Reddit posts included were made between February 12, 2023, and February 11, 2024. The data collected and analyzed were publicly available Reddit posts; however, our analyses did not use any personally identifiable information or direct identifiers (eg, usernames, links, or IDs).

Data Acquisition

This research sought to understand prevalent topics, sentiment, and engagement information within Reddit posts and comments related to extreme heat events in Canada. Reddit users post content within designated communities, known as subreddits. When a user posts content in a subreddit, the other members of the subreddit can "upvote" (ie, indicate a favorable reaction), "downvote," (ie, indicate an unfavorable reaction), and comment on the post. Subreddits can be focused on any topic or interest, including national, provincial or territorial, or local news and issues.

The research team collected posts and comments from a predetermined list of Canada-specific subreddits using Reddit's API [17]. We used the Reddit API to search the selected subreddits for the following heat- and climate-related search terms: "heat," "extreme heat," "hot weather," "hot summer," "hottest day," "heatwave," "heat dome," "hot day," "hottest year," "climate change AND heat," and "BC heat wave." These terms were determined after searching the subreddits for the keywords used to discuss extreme heat.

To ensure that our keyword set and subreddit sample appropriately captured discussion relevant to extreme heat in Canada, we conducted preliminary pilot searches. We manually browsed Canadian subreddits to identify the actual terms users used when discussing extreme heat, which helped refine our candidate keywords (eg, "heat dome," "hottest day") beyond generic terms, such as "hot weather." For subreddit selection, we began with lists of national, provincial or territorial, and major city subreddits, then excluded those with very low activity (few heat-related posts during pilot months) or that were clearly irrelevant (eg, subreddits focused exclusively on entertainment or sports). This process ensured that the sample remained both manageable and relevant. Similar approaches are used in related social media studies, such as that by Janaswamy and Blackburn [18], who identified climate change–relevant subreddits by starting with seed communities and empirically validating activity levels, and that by Fariello and Jemielniak [12], who validated keyword sets through historical analyses of evolving climate-related terminology.

Text Analytics

The textual data from Reddit posts and comments follow a hierarchical tree structure. Each discussion thread on Reddit begins with a main post, which forms the root node of the tree. Subsequent comments within the thread serve as child nodes. Specifically, a comment that directly responds to the main post is attached as a child to the root node. The comments that are replies to other comments are linked as child nodes to their respective parent comments. This recursive linkage continues, mapping the entire discussion thread into a tree structure where the path from the root to any node represents the flow of conversation from the initial post to a particular comment. Finally, the comments that have no replies are leaves of the tree with no children.

After constructing the tree structures to represent individual Reddit threads, we performed topic analysis on these threads. To facilitate effective topic detection by the language model, it is essential to create more cohesive documents that encapsulate the main context and key information within the discussions. This was achieved by traversing the tree in a level-order fashion. During this traversal, the text of each visited node (comment) was concatenated to form a continuous document. This process continued until the document reached a predefined size suitable for the input constraints of the language model. In this way, the comments that are replies to a specific comment were sequentially concatenated, preserving the flow of discussion within the document. The document construction was terminated either when it reached its maximum allowable size or when there were no further comments to process. If a cutoff occurred due to reaching the maximum size, a new document was initiated following the same methodology, continuing until all comments in the thread were processed. If a single comment or a single post was longer than the maximum size, it was constructed as one document by itself. This process began with the main Reddit post itself, considering it as the parent node from which all subsequent comments in the thread were derived and processed.

To mitigate the sparsity challenges of short comments and to preserve conversational flow, we concatenated comments and replies into larger "documents." Similar strategies have been used in prior research to improve topic modeling and clustering of fragmented user-generated text. For example, Llewellyn et al [19] demonstrated that combining online newspaper comments with their replies produced more coherent topics when applying Latent Dirichlet Allocation . Likewise, Sun et al [20] developed a conversational structure-aware topic model, which leverages reply structures in threaded discussions to yield more interpretable topics. These approaches, like ours, show that aggregating related comments along conversational paths can enhance coherence while retaining contextual meaning that might be lost if each comment were treated entirely independently.

While traversing the 607 tree structures, we excluded comments generated by bots that were deleted or written in languages other than English. This filtering process removed 22,069 comments from the initial pool of 31,118, leaving 9049 valid comments along with the original 607 posts. These were consolidated into 1352 documents for analysis. On average, each document contained 270.56 words, aligning well with our language model's input requirements, which accommodate a maximum token size of 384.

Topic Analysis

For topic modeling, we used BERTopic (bidirectional encoder representations topic modeling), a neural topic modeling

XSL•FO
RenderX

technique, to conduct topic analysis on the documents extracted from Reddit, which were assembled using the document consolidation method outlined in the text analytics section. BERTopic leverages class-based term frequency-inverse document frequency for improved topic representation [21]. BERTopic combines transformer-based embeddings with hierarchical clustering, enabling coherent topic extraction even from short or noisy texts, making it well-suited for our dataset. To assess coherence, we manually inspected representative documents for each identified topic to verify that the extracted clusters reflected consistent underlying themes.

Sentiment Analysis

Sentiment analysis is used to ascertain the prevailing sentiment within textual data, specifically determining whether the sentiment of a given text is positive or negative, and assigning a score to quantify the intensity of the identified sentiment direction. We utilized the consolidated documents, assembled using the method described above, as input for the *distilbert-base-uncased-finetuned-sst-2-english* model [22,23]. This model was specifically designed for text classification tasks and was used to assess the sentiment direction and sentiment score for each document [24]. To provide a check on model performance, we also manually inspected a random sample of 50 documents classified as positive and 50 documents classified as negative. This validation suggested that the sentiment assignments generally aligned with the intended tone of the posts.

Emotion Analysis

Emotion analysis is conducted on textual data to detect and identify various emotional elements within the text. For emotion analysis, we used roberta-base-go_emotions, an off-the-shelf model fine-tuned on the GoEmotions dataset [25]. RoBERTa's optimized pretraining enhances contextual understanding, while fine-tuning on GoEmotions enables precise detection of fine-grained emotions [26,27]. GoEmotions dataset includes 12 positive, 11 negative, 4 ambiguous, and 1 neutral emotion, making it suitable for conversation and differentiating between subtle emotion expression [28]. Using this fine-tuned language model on the consolidated documents previously described, we discerned the level of negative (eg, anger, sadness, or fear) and positive (eg, admiration or optimism) emotions, as well as approval and disapproval emotions within the extracted Reddit documents.

Engagement Analysis

Reddit users often express their agreement or disagreement with a given post or comment through upvotes and downvotes, respectively. The net upvote score of a Reddit post or comment serves as an indicator of engagement and is determined by subtracting the number of downvotes from the number of upvotes [29]. A post or comment with a high net upvote score is generally considered to be approved by the subreddit community, whereas one with a low net upvote score is viewed as disapproved [29].

Results

Data Collection

Based on the search terms and included subreddits, 607 Reddit posts were collected from over 15,000 users. On average, there were approximately 20 posts per subreddit and 52 comments per post. Table 1 provides the total number of posts, users, comments, and distribution of posts across the subreddits.

Table . Total posts, unique users, comments, comments per post, subreddits, and posts per subreddit.

Characteristic	Value
Reddit posts	607
Unique users	15,366
Comment	31,118
Comments per post, SD[a]	81.9
Comments per post, mean[a]	52.8
Comments per post, median (IQR)[a]	24.00 (9-58)
Subreddit	30
Posts per subreddit, SD	17.0
Posts per subreddit, mean	20.2
Posts per subreddit, median (IQR)	18.5 (6-29.5)

[a]Based on posts with at least 1 comment.

Across the 11 keywords used to identify posts in the included subreddits, "heat" was the most common (501 posts, 25,253 comments), followed by "extreme heat" (28 posts, 1320 comments) and "hot weather" (19 posts, 1561 comments; Table 2).

XSL•FO
RenderX

Table . Post and comment distribution across keywords.

Keyword	Post, n (%)	Comment, n (%)
heat	501 (82.5)	25,253 (81.1)
extreme heat	28 (4.6)	1,320 (4.2)
hot weather	19 (3.1)	1,561 (5)
hot summer	17 (2.8)	622 (2)
hottest day	11 (1.8)	344 (1.1)
heatwave	8 (1.3)	348 (1.1)
heat dome	7 (1.2)	868 (2.8)
hot day	6 (1)	225 (0.7)
hottest year	5 (0.8)	363 (1.1)
climate change AND heat	3 (0.5)	212 (0.7)
BC heat wave	2 (0.3)	2 (<0.01%)

Across the included subreddits, r/canada had the most posts and comments (62 and 6010, respectively), followed by r/halifax (53 posts, 1772 comments), and r/britishcolumbia (47 posts; 2927 comments). Table 3 shows the distribution of the posts and comments across all included subreddits.

XSL•FO
RenderX

Table . Post and comment distribution across subreddits.

Subreddit[a]	Post, n (%)	Comment, n (%)
r/canada	62 (10.2)	6010 (19.3)
r/halifax	53 (8.7)	1772 (5.7)
r/britishcolumbia	47 (7.7)	2927 (9.4)
r/vancouver	45 (7.4)	2475 (8.0)
r/ontario	43 (7.1)	2194 (7.1)
r/Winnipeg	34 (5.6)	1379 (4.4)
r/onguardforthee	34 (5.6)	960 (3.1)
r/ottawa	31 (5.1)	1701 (5.5)
r/alberta	25 (4.1)	1949 (6.3)
r/VictoriaBC	25 (4.1)	890 (2.9)
r/NovaScotia	23 (3.8)	648 (2.1)
r/Edmonton	23 (3.8)	656 (2.1)
r/CanadaPolitics	21 (3.5)	545 (1.8)
r/Canada_sub	21 (3.5)	1671 (5.4)
r/Calgary	20 (3.3)	933 (3.0)
r/newfoundland	17 (2.8)	334 (1.1)
r/toronto	14 (2.3)	1575 (5.1)
r/saskatchewan	12 (2.0)	880 (2.8)
r/saskatoon	12 (2.0)	532 (1.7)
r/PEI	12 (2.0)	213 (0.7)
r/montreal	7 (1.2)	371 (1.2)
r/StJohnsNL	6 (1.0)	67 (0.2)
r/Manitoba	6 (1.0)	142 (0.5)
r/newbrunswickcanada	4 (0.7)	145 (0.5)
r/moncton	3 (0.5)	23 (0.1)
r/regina	2 (0.3)	75 (0.2)
r/fredericton	2 (0.3)	9 (<0.01%)
r/Yukon	1 (0.2)	33 (0.1)
r/NWT	1 (0.2)	9 (<0.01%)
r/Whitehorse	1 (0.2)	0 (0.0)

[a]Subreddit names are presented as originally created on Reddit. Capitalization varies across communities and does not necessarily reflect conventional geographic naming standards.

Topic Analysis

We extracted 14 meaningful topics from 1111 of the documents, characterized by their respective scores (Table 4). However, 241 documents were associated with a mixture of topics or contained topics that were not clearly recognizable and, as such, were invalid topics (topic -1 in Table 4). The overall topic distribution across all 1352 documents is presented in Table 4. Heating and cooling mechanics and costs associated with each appeared frequently throughout the topics.

XSL•FO
RenderX

Table . Topic and top word distribution across documents (n=1111 valid topics).

Topic	Top words	Values, n (%)
-1 (invalid topic)	__[a]	241 (17.8)[b]
0	pump, furnace, gas, heat, house	571 (51.4)[c]
1	weather, climate, rain, records, wave	218 (19.6)[c]
2	[air] conditioners, people, health...	51 (4.6)[c]
3	tax, carbon, moe, saskatchewan, money	42 (3.7)[c]
4	walmart, prices, food, grocery, costco	40 (3.6)[c]
5	landlord, rent, tenants, building...	35 (3.2)[c]
6	ac, windows, curtain, fans, sleep	29 (2.6)[c]
7	cat, dog, pets, fan, ice	26 (2.3)[c]
8	wear, hoodie, pants, wool, shoes	23 (2.1)[c]
9	grid, power, nuclear, shortages...	21 (1.9)[c]
10	valve, condo, baseboard, apartment...	19 (1.7)[c]
11	cool, pillow, ice, pools, healthlink	15 (1.4)[c]
12	love, hate, like, humidity, outside	11 (1.0)[c]
13	schools, manitoba, strike, grade...	10 (0.9)[c]

[a]Invalid topic excluded from the top word analysis.

[b]Denominator is all documents, n=1352.

[c]Denominator is all valid documents, n=1111.

Sentiment Analysis

Figure 1 illustrates the distribution of negative and positive sentiment documents across the top 13 subreddits (the largest subreddits with a comparable number of documents). The subreddit r/calgary (positive: n=47, 89% vs negative: n=6, 11%) had the highest positive sentiment, followed by r/Edmonton (n=32, 67% vs n=16, 33%) and r/ontario (n=80, 62% vs n=49, 38%). The subreddits with the highest negative sentiment included r/canada (positive: n=47, 34% vs negative: n=92, 66%), followed by r/saskatchewan (n=11, 41% vs n=16, 59%) and r/alberta (n=30, 41% vs n=43, 59%).

XSL•FO
RenderX

Figure 1. Distribution of negative and positive sentiment documents across top 13 subreddits.

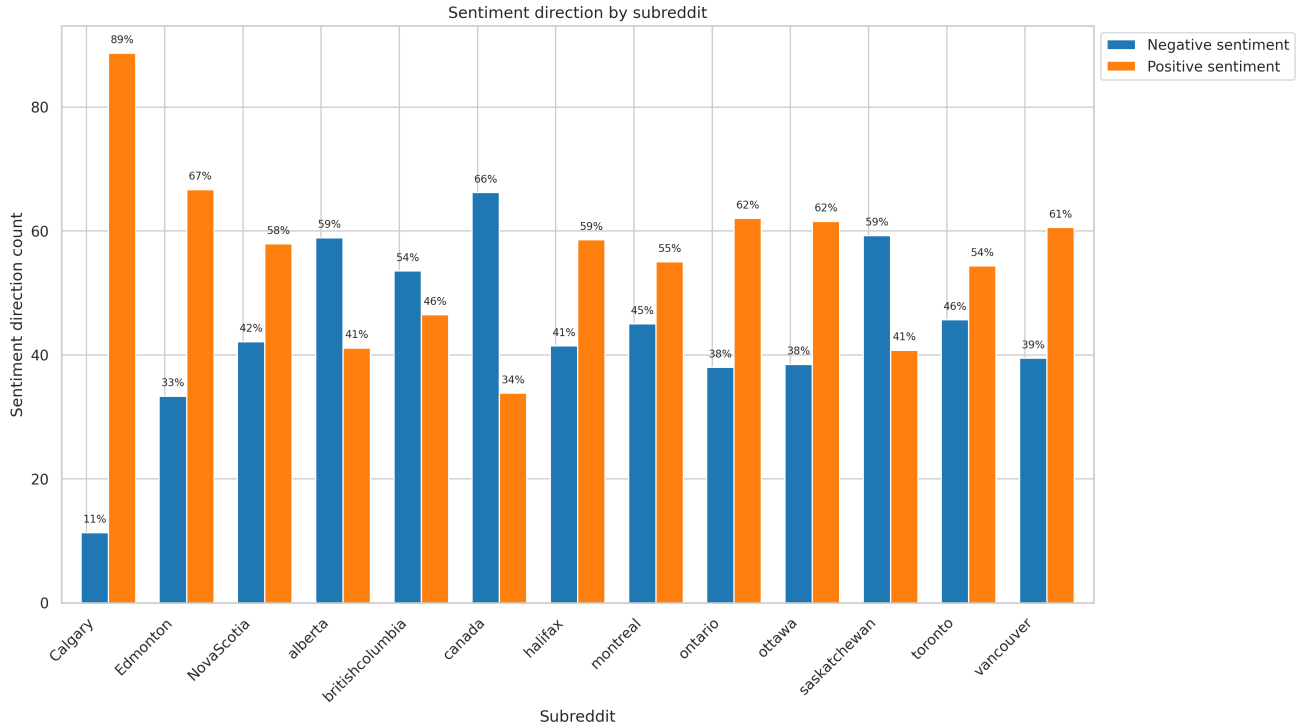

Emotion Analysis

We discerned diverse emotions within the top 13 subreddits (the largest subreddits with a comparable number of documents). Figure 2 displays the average scores for the emotions of approval and disapproval across each subreddit. Within the subreddits, r/Calgary (0.17) had the highest approval emotion score, followed by r/montreal (0.16) and r/alberta (0.15). However, r/montreal (0.11) had the highest disapproval emotion score, followed by r/alberta (0.084), and r/saskatchewan (0.083).

Figure 2. Approval and disapproval emotion distribution by top 13 subreddits.

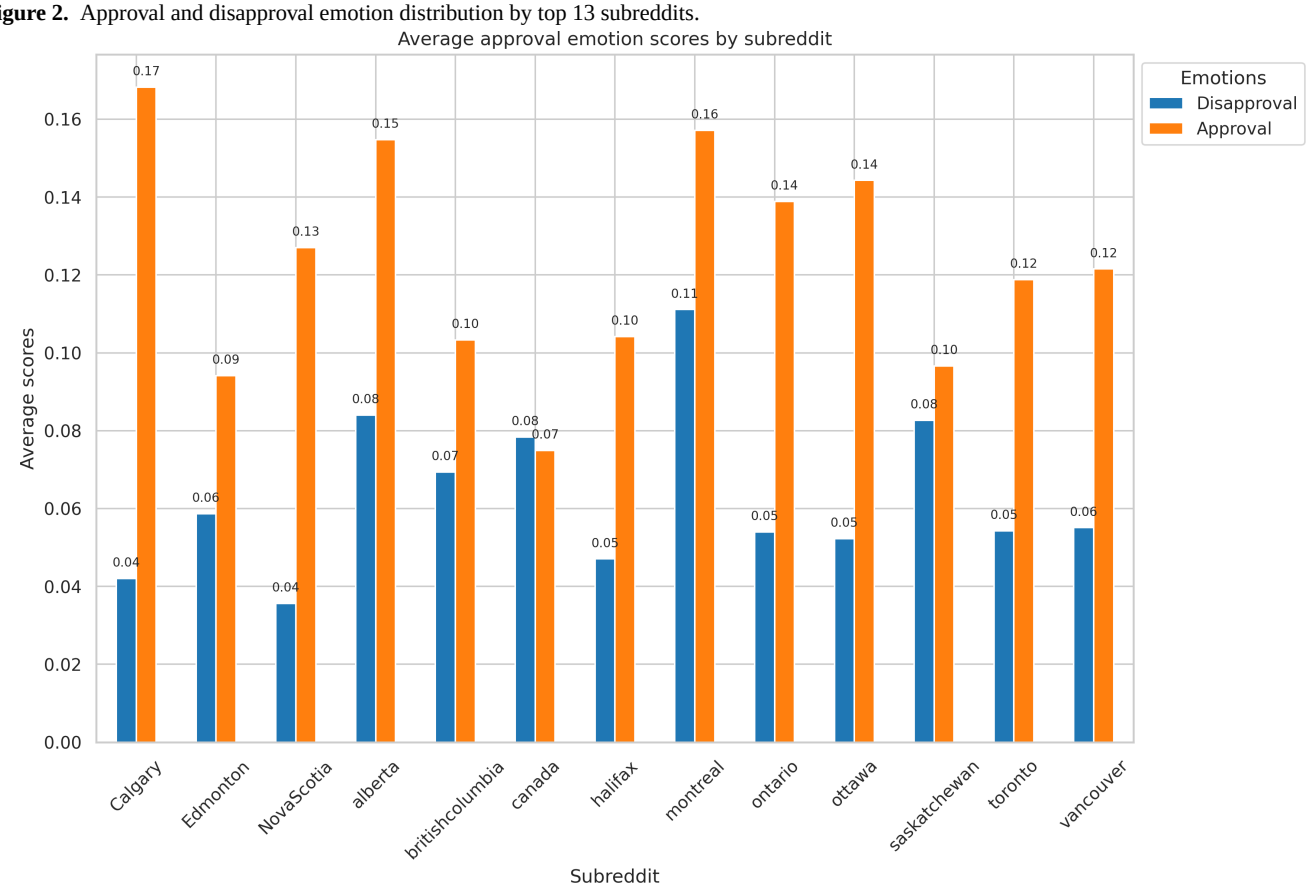

XSL•FO
RenderX

Similarly, Figure 3 presents the average scores for anger, sadness, and fear across the top 13 subreddits. Anger was the highest in r/montreal (0.035), r/canada (0.034), and r/toronto (0.027). Sadness was the highest in r/britishcolumbia (0.043), r/ottawa (0.042), and r/toronto (0.038). Finally, fear was the highest in r/montreal (0.033), r/britishcolumbia (0.018), and r/canada (0.015).

Figure 3. Negative emotion (anger, sadness, and fear) distribution by top 13 subreddits.

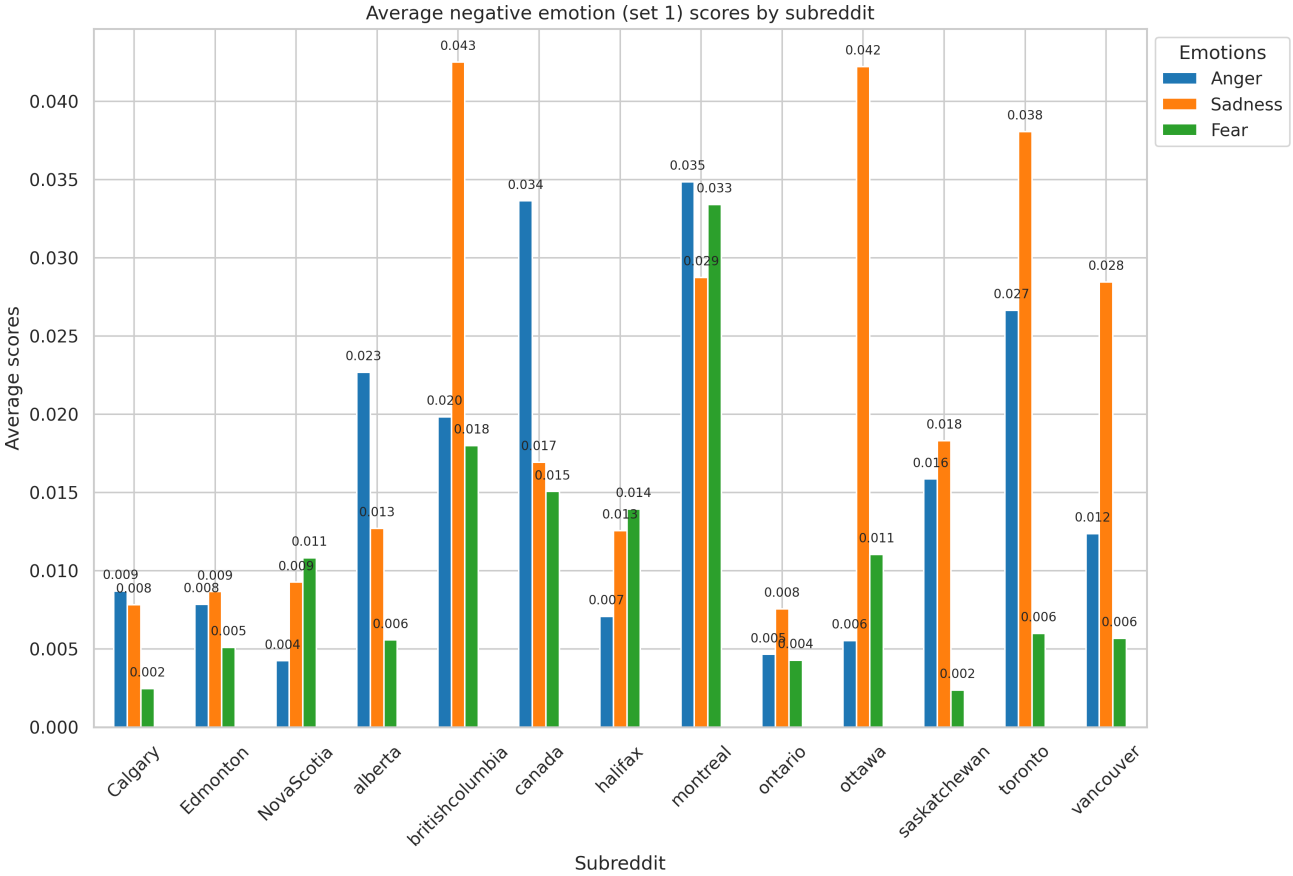

Figure 4 illustrates the average scores for admiration and optimism for the top 13 subreddits. Admiration was found to be the highest in r/saskatchewan (0.094), r/britishcolumbia (0.084), and r/Calgary (0.082). Optimism was the highest in r/saskatchewan (0.096), r/toronto (0.095), and r/Edmonton (0.074).

XSL•FO
RenderX

Figure 4. Positive emotion (admiration and optimism) distribution by top 13 subreddits.

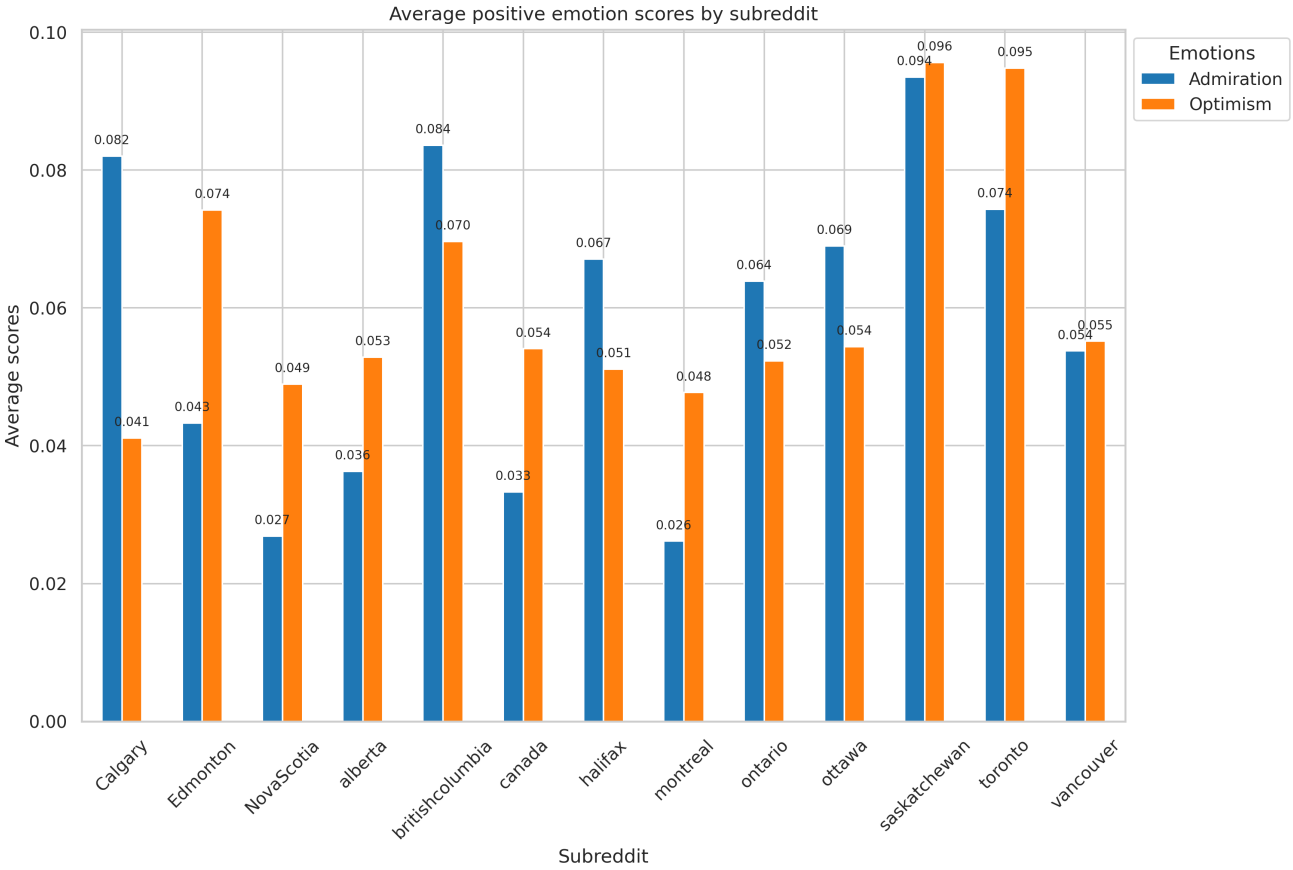

Engagement Analysis

Table 5 illustrates the average upvote score for the posts and comments extracted from the top 13 subreddits, reflecting the levels of community engagement and approval. In our dataset, the subreddit r/toronto (23.68) had the highest level of engagement and approval, followed by r/canada (17.87) and r/montreal (17.14).

Table . Average normalized upvote by top 13 subreddits.

Subreddit	Upvote number
r/toronto	23.686033
r/canada	17.872152
r/montreal	17.13952
r/britishcolumbia	15.612242
r/vancouver	14.76366
r/ottawa	12.568449
r/Calgary	12.30941
r/alberta	11.324873
r/saskatchewan	9.57762
r/ontario	9.345084
r/Edmonton	6.49853
r/Halifax	6.349839
r/NovaScotia	5.056377

XSL•FO
RenderX

Discussion

Principal Findings

This study examined Reddit discussions related to extreme heat in Canada to identify common topics and concerns, explore sentiment and emotional responses, and analyze how these relate to community engagement. The findings indicate that extreme heat and climate change are prominent themes among Canadian Reddit users, with substantial regional variation in both the discussion focus and emotional tone. Engagement patterns coincided with discussions of climate-related issues, suggesting that social media can provide signals of public perception and discourse about extreme weather events. Interpreted through a risk communication lens, these discussions reflect how people perceive health risks, express uncertainty, and negotiate adaptation strategies in an online environment. From an information ecosystem perspective, Reddit acts as a dynamic venue where public health–relevant information is coconstructed, contested, and amplified. While our findings suggest that Reddit can surface emerging concerns and perspectives related to extreme heat, these insights should be understood as complementary to other forms of evidence (eg, surveys, interviews, or other social media analyses) rather than representative of public perceptions in Canada as a whole.

Reddit can provide an exploratory view of public perceptions, particularly among users who are active on the platform, especially with regard to current events that impact health. Prior research has highlighted its utility for tracking perceptions during events such as influenza outbreaks [30]. In this study, Reddit discussions captured public concerns about extreme heat and its growing threats to health, safety, and economic well-being, particularly in cities such as Halifax and Vancouver. Users referenced real-world impacts such as elevated mortality following heatwaves, heightened fire risks, and deteriorating quality of life, which align with national data on climate-related health outcomes [7]. These discussions may reflect broader concerns about climate anxiety among some Canadians, with many linking local events to larger concerns about global climate change [31]. The themes we identified, such as concerns about energy costs, adaptation strategies, and climate anxiety, should be viewed as potential information needs. They may help guide future research and message development, but actionable public health communication strategies will require additional data sources and direct engagement with communities.

Common Discussion Topics and Concerns About Extreme Heat in Canada

Reddit users across Canada expressed widespread interest in heat-related issues, as reflected in the dominance of heat-related content in the analyzed posts. Specifically, 82.5% (n=501) of the posts and over 80% (n=25,253) of the comments centered on the topic of “heat.” In addition to general heat concerns, other common themes included extreme heat, hot weather, and hot summer. These patterns align with national polling data, such as a recent Leger survey showing that 70% of Canadians are concerned about climate change and its role in increasing extreme weather events [31].

The r/Canada subreddit had the highest number of posts, comments, and engagement overall, indicating the breadth of public discourse about extreme heat and climate change across the country. Region-specific subreddits, such as r/halifax and r/britishcolumbia, followed closely, with high volumes of posts and engagement—likely reflecting local experiences with recent extreme heat events in regions historically known for more moderate climates [32]. These observations parallel other Reddit-based studies on climate discourse that similarly found widespread concern about the impacts of climate change, with the consequences of climate change a main topic of discussion on 100,000 Reddit posts [33].

Beyond general climate concerns, the topic modeling analysis revealed several dominant themes related to daily life and broader societal challenges. The largest topic cluster focused on heating and cooling, including practical terms, such as “pump,” “furnace,” and “gas.” These reflect public concern over energy infrastructure and affordability, particularly in the context of extreme temperatures. This is supported by evidence showing that many Canadian households face energy poverty, leading to discomfort, health risks, and financial strain [34]. Reddit users expressed anxiety around increasing energy costs and insufficient cooling infrastructure, especially during heatwaves.

Health impacts emerged as another important discussion topic. Climate change and extreme heat are contributing to health inequities across the determinants of health, leading to impacts on community and individual health in Canada [35]. A popular news article in 2024 focused on extreme heat and outlined how racialized communities, those renting their homes, and those living in energy poverty were found to be at higher risk [36]. These findings correspond with the broader recognition of how climate change intersects with social determinants of health [35] and demonstrate how Reddit discussions mirror evolving media, political, and institutional attention to these equity concerns.

Attitudes and Emotions Expressed in Reddit Discussions on Extreme Heat

The emotional tone of Reddit discussions varied substantially by region, illustrating different levels of concern, belief, and lived experience with climate change across the country. Subreddits such as r/Calgary and r/Edmonton showed more positive sentiment overall, while r/canada and r/saskatchewan leaned toward more negative sentiment. This contrast highlights the nuanced ways that attitudes toward climate-related challenges manifest online.

Regional differences in emotion were also apparent in the emotion analysis. For example, subreddits such as r/saskatchewan and r/calgary expressed relatively high levels of admiration and optimism. This may reflect regional differences in belief systems about climate change. Some research suggests that individuals who downplay the severity of climate change tend to report more positive feelings, which can lead to a phenomenon known as “denial-based hope” [37]. While this outlook may reduce perceived urgency, it can also hinder individual and collective action to address climate risks [38].

XSL•FO
RenderX

In contrast, negative emotions, such as anger, sadness, and fear, were more pronounced in subreddits such as r/montreal, r/britishcolumbia, and r/canada.

These emotions may stem from direct experiences with climate-related disasters, such as the 2021 heat dome in British Columbia and the 2023 wildfires in Nova Scotia. Consistent with previous research, negative emotional responses, particularly fear, were associated with heightened climate awareness and a greater likelihood of supporting or engaging in adaptive action [39]. In our dataset, these emotions appeared in ways that may align with those patterns, though our study did not measure behavior directly. Sustained exposure to such distress can also lead to climate-related mental health challenges. A recent survey found that 78% of young Canadians reported negative mental health impacts from climate change, underscoring the psychological burden of climate anxiety [40].

Engagement Patterns in Heat-Related Reddit Discussions: Connections With Topics and Sentiment

Community engagement, as measured by upvotes, was the highest in subreddits r/toronto and r/canada, indicating high levels of user participation and approval in discussions surrounding extreme heat. Upvotes make a post more visible on Reddit and show others' support or approval of the topic [29].

These findings reflect broader trends in climate change communication on social media, where approximately one-third of users engage with climate-related content, particularly Gen Z adults [41]. Engagement is often higher for posts that include calls to action or connect the problem of climate change with potential solutions [41]. Credible sources, linking the problem to solutions, and including calls to action are key best practices for engaging people in climate change information [42]. In this context, high engagement with Reddit content may indicate a public appetite for action-oriented or relatable discussions, especially when aligned with credible sources and practical implications for everyday life (eg, energy costs, health impacts).

Engagement levels also intersected with topic and sentiment trends. Subreddits with high discussion volumes and strong negative emotional expression also saw high levels of engagement. This pattern indicates that emotionally charged posts may coincide with more active participation, though causality cannot be inferred. While our results suggest that stronger emotional expression often coincided with higher engagement, our study was not designed to explain why some emotionally charged posts gain more traction than others. Understanding these dynamics would require qualitative analysis of discourse and context, or triangulation with other data sources.

Taken together, these findings suggest that Reddit can serve as an exploratory lens into how Canadians discuss and react to extreme heat, surfacing concerns that range from household-level adaptation strategies to broader issues of affordability, energy systems, and equity. While not representative of the population as a whole, these insights highlight potential information needs and public concerns that public health agencies and policymakers may wish to consider alongside other information sources when designing communication strategies around extreme heat and climate risks. More broadly, this study illustrates how computational analyses of social media can complement traditional research approaches by identifying emergent themes and emotions in real time, thereby offering early signals of public discourse that warrant further investigation. This research demonstrates the value of integrating social media monitoring into the broader toolkit for public health communication and climate resilience planning.

Future Research

Future research should examine Reddit posts related to extreme heat and climate change in Canada for misinformation or disinformation. Within this research, it would be beneficial to examine how misinformation or disinformation about climate change and extreme heat impacts the engagement and emotions found within posts. Further exploring the relationship between emotions and action in climate adaptation would also help shape initiatives. Developing communication reflective of the topics and discussions found in the subreddits and evaluating them across communities would also help understand the utility of social media discourse for informing official messaging. Future research should also build on this descriptive foundation by incorporating qualitative coding of excerpts, deeper contextual interpretation, and comparisons with external datasets, such as meteorological records or health outcomes. Finally, discourse found on Reddit should be compared with other social media platforms and Google trends to understand similarities and differences found in online sources.

Limitations

This research was based on the social media platform Reddit and common subreddits in Canada. Generalizability is limited because Reddit users skew toward younger demographics and each subreddit has its own norms, moderation practices, and community perspectives. As Caplan and Purser [43] highlight, subreddit-specific cultures and sociotechnical contexts shape discourse; our inclusion of 30 national, provincial or territorial, and local subreddits was intended to mitigate this, but the findings should still be interpreted as a broad snapshot rather than as representative of each subreddit's culture. Future research could extend this work by analyzing subreddit-specific practices in greater depth and comparing results with other online platforms.

A key technical limitation is Reddit's API restriction, which caps retrieval at 1000 posts per query. In high-traffic subreddits, this constraint can bias the dataset toward more recent content and limit temporal coverage.

Model-related limitations must also be considered. The BERTopic and sentiment or emotion models used were trained on general-purpose corpora that differ from Canadian Reddit discussions of extreme heat. This may lead to systematic misclassification; for example, sarcasm, humor, or colloquial expressions may be interpreted incorrectly. These constraints affect both topic coherence and sentiment assignments, underscoring the need for caution in drawing substantive conclusions.

XSL•FO
RenderX

Finally, despite filtering out bot-generated, deleted, and non-English comments, undetected bots remain a persistent challenge. Sophisticated bots can mimic human discourse and potentially amplify particular narratives, subtly influencing topic distributions and sentiment profiles. This is especially relevant in politically and socially salient discussions.

Conclusions

This study demonstrates the impacts of extreme heat and climate-related challenges across Canadian communities, found within subreddit discussions. Topics around heating and cooling, financial burdens, and broader social implications of adapting to a changing climate were found. Community engagement was particularly high in r/toronto and r/Canada, signifying active discourse and collective concern. The emotional analysis showed regional differences, with positive emotions such as optimism and admiration common in r/saskatchewan and r/Calgary, and negative emotions such as fear and anger more common in r/montreal and r/canada. Terms such as “pump,” “costs,” and “climate” were common and reinforced the need for adaptive strategies that address social, economic, and environmental concerns. Overall, the findings demonstrate how Reddit may serve as a useful exploratory platform for examining discourse around climate change and extreme heat, offering signals that could inform future tailored and community-specific communication efforts.

Acknowledgments

Generative artificial intelligence was not used in the research or manuscript development associated with this study.

Funding

This research was funded by a University of Guelph CARE-AI Seed Fund AI Innovation Award.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request. We have placed the source code, sample negative sentiment documents, sample positive sentiment documents, and representative documents for each topic in [44].

Authors' Contributions

Conceptualization: FZ, JEM, MM, SG, SZ
Formal analysis: FZ, SZ
Funding acquisition: JEM, MM, SG
Investigation: SZ
Methodology: FZ, JEM, MM, SG, SZ
Project administration: JEM, MM
Supervision: FZ, JEM, MM
Writing – original draft: MM, SZ
Writing – review and editing: FZ, JEM, MM, SG, SZ

Conflicts of Interest

None declared.

References

1. What is climate change? United Nations. 2025. URL: https://www.un.org/en/climatechange/what-is-climate-change [accessed 2026-01-31]
2. Who is most impacted by climate change. Government of Canada. 2022. URL: https://www.canada.ca/en/health-canada/services/climate-change-health/populations-risk.html [accessed 2026-02-10]
3. Climate change adaptation in Canada. Natural Resources Canada. 2008. URL: https://natural-resources.canada.ca/climate-change/what-adaptation/10025 [accessed 2026-01-31]
4. Extreme heat events: health risks and who is at risk of extreme heat events. Government of Canada. 2022. URL: https://www.canada.ca/en/health-canada/services/climate-change-health/extreme-heat/who-is-at-risk.html [accessed 2026-01-31]
5. Risks to health from climate change. Government of Canada. 2022. URL: https://www.canada.ca/en/health-canada/services/climate-change-health/risks-to-health.html [accessed 2026-01-31]
6. Extreme heat events: overview. Government of Canada. 2018. URL: https://www.canada.ca/en/health-canada/services/climate-change-health/extreme-heat.html [accessed 2026-01-31]
7. Fact sheet: climate change and heat waves. Canadian Climate Institute. 2024. URL: https://climateinstitute.ca/news/fact-sheet-heat-waves/ [accessed 2026-01-31]
8. Peters E, Boyd P, Cameron LD, et al. Evidence-based recommendations for communicating the impacts of climate change on health. Transl Behav Med 2022 May 25;12(4):543-553. [doi: 10.1093/tbm/ibac029] [Medline: 35613000]

XSL•FO
RenderX

9. León B, Negredo S, Erviti MC. Social engagement with climate change: principles for effective visual representation on social media. Clim Policy 2022;22(8):976-992. [doi: 10.1080/14693062.2022.2077292]
10. Schäfer MS. Social media in climate change communication: state of the field, new developments and the emergence of generative AI. Dialogues Clim Change 2025 Apr;2(1):49-55. [doi: 10.1177/29768659241300666]
11. Reddit users by country. Oberlo. 2024. URL: https://www.oberlo.com/statistics/reddit-users-by-country [accessed 2026-01-31]
12. Fariello G, Jemielniak D. The changing language and sentiment of conversations about climate change in Reddit posts over sixteen years. Commun Earth Environ 2025;6(1):1-9. [doi: 10.1038/s43247-024-01974-8]
13. Treen K, Williams H, O'Neill S, Coan TG. Discussion of climate change on Reddit: polarized discourse or deliberative debate? Environ Commun 2022 Jul 4;16(5):680-698. [doi: 10.1080/17524032.2022.2050776]
14. Gadanidis T. The discourse of climate change on reddit. Toronto Working Papers in Linguistics. 2020. URL: https://twpl.library.utoronto.ca/index.php/twpl/article/view/41010 [accessed 2026-01-31]
15. Research ethics/protection of human participants. University of Guelph. 2025. URL: https://www.uoguelph.ca/research/services-divisions/ethics [accessed 2026-01-31]
16. Interagency Advisory Panel on Research Ethics. Does research using social media platforms require research ethics board review?—Guidance in applying TCPS 2 (2022). Government of Canada. 2022. URL: https://ethics.gc.ca/eng/reb-cer_social-sociaux.html [accessed 2026-01-31]
17. Boe B. PRAW: The Python Reddit API Wrapper 7.7.1 documentation. 2023. URL: https://praw.readthedocs.io/en/stable/ [accessed 2026-01-31]
18. Janaswamy S, Blackburn J. Exploring climate change discourse: measurements and analysis of reddit data. arXiv. Preprint posted online on Dec 2, 2024. [doi: 10.48550/arXiv.2412.01111]
19. Llewellyn C, Grover C, Oberlander J. Improving topic model clustering of newspaper comments for summarisation. Presented at: Proceedings of the ACL 2016 Student Research Workshop; Aug 7-12, 2016. [doi: 10.18653/v1/P16-3007]
20. Sun Y, Loparo K, Kolacinski R. Conversational structure aware and context sensitive topic model for online discussions. Presented at: 2020 IEEE 14th International Conference on Semantic Computing (ICSC); Feb 3-5, 2020. [doi: 10.1109/ICSC.2020.00019]
21. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv. Preprint posted online on Mar 11, 2022. [doi: 10.48550/arXiv.2203.05794]
22. Socher R, Perelygin A, Wu J, et al. Recursive deep models for semantic compositionality over a sentiment treebank. Presented at: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing; Oct 20, 2013; Seattle, Washington, USA p. 1631-1642. [doi: 10.18653/v1/D13-1170]
23. Sanh V, Debut L, Chaumond J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv. Preprint posted online on Oct 2, 2019. [doi: 10.48550/arXiv.1910.01108]
24. Distilbert Base Uncased Finetuned Sst 2 English. dataloop. 2024. URL: https://dataloop.ai/library/model/distilbert_distilbert-base-uncased-finetuned-sst-2-english/ [accessed 2026-01-31]
25. Demszky D, Movshovitz-Attias D, Ko J, Cowen A, Nemade G, Ravi S. GoEmotions: a dataset of fine-grained emotions. Presented at: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Jul 5-10, 2020. [doi: 10.18653/v1/2020.acl-main.372]
26. Liu Y, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. Preprint posted online on Jul 26, 2019. [doi: 10.48550/arXiv.1907.11692]
27. SamLowe/roberta-base-go_emotions. Hugging Face. URL: https://huggingface.co/SamLowe/roberta-base-go_emotions [accessed 2026-01-31]
28. Alon D, Ko J. GoEmotions: a dataset for fine-grained emotion classification. Google Research. 2021. URL: https://research.google/blog/goemotions-a-dataset-for-fine-grained-emotion-classification/ [accessed 2026-01-31]
29. Rechel E. What is Reddit upvotes? Hybrid Traffic. 2021. URL: https://www.hybridtraffic.net/single-post/what-is-reddit-upvotes [accessed 2026-01-31]
30. Park A, Conway M. Tracking health related discussions on Reddit for public health applications. AMIA Annu Symp Proc 2018;2017:1362-1371. [Medline: 29854205]
31. Rabson M. 7 in 10 Canadians worried about climate change, poll suggests. CBC News. 2023. URL: https://www.cbc.ca/news/politics/majority-canadians-worried-about-climate-change-1.6964334 [accessed 2026-01-31]
32. Extreme heat in Canada. Canadian Climate Institute. 2023. URL: https://climateinstitute.ca/reports/extreme-heat-in-canada/ [accessed 2026-01-31]
33. Parsa MS, Shi H, Xu Y, Yim A, Yin Y, Golab L. Analyzing climate change discussions on reddit. Presented at: 2022 International Conference on Computational Science and Computational Intelligence (CSCI); Dec 14-16, 2023. [doi: 10.1109/CSCI58124.2022.00150]
34. Martiskainen M, Das R. Worried about high energy bills, some Canadians risk discomfort, illness and even death. The Conversation. 2022. URL: http://theconversation.com/worried-about-high-energy-bills-some-canadians-risk-discomfort-illness-and-even-death-186371 [accessed 2026-01-31]
35. Health Canada. Health of Canadians in a changing climate: advancing our knowledge for action. : Government of Canada; 2022 URL: https://changingclimate.ca/site/assets/uploads/sites/5/2022/02/CCHA-REPORT-EN.pdf [accessed 2026-01-31]

XSL•FO
RenderX

36. Ram A, Shingler B, Pelley L. Extreme heat is a global killer—and worse for our health than previously thought, new research shows. CBC News. 2024. URL: https://www.cbc.ca/news/health/heat-second-opinion-1.7241714 [accessed 2026-01-31]
37. Bechtoldt MN, Schermelleh-Engel K. Worry's Clout: Concern, not positive affectivity, drives climate activism. J Environ Psychol 2025 Feb;101:102517. [doi: 10.1016/j.jenvp.2024.102517]
38. Why some Americans do not see urgency on climate change. Pew Research Center. 2023. URL: https://www.pewresearch.org/science/2023/08/09/why-some-americans-do-not-see-urgency-on-climate-change/ [accessed 2026-01-31]
39. Lawrance EL, Thompson R, Newberry Le Vay J, Page L, Jennings N. The impact of climate change on mental health and emotional wellbeing: a narrative review of current evidence, and its implications. Int Rev Psychiatry 2022 Aug;34(5):443-498. [doi: 10.1080/09540261.2022.2128725] [Medline: 36165756]
40. Galway LP, Field E. Climate emotions and anxiety among young people in Canada: a national survey and call to action. J Clim Chang Health 2023 Jan;9:100204. [doi: 10.1016/j.joclim.2023.100204]
41. Tyson A, Kennedy B, Funk C. Climate engagement and activism. Pew Research Centre. 2021. URL: https://www.pewresearch.org/science/2021/05/26/climate-engagement-and-activism/ [accessed 2026-01-31]
42. Communicating on climate change. United Nations. URL: https://www.un.org/en/climatechange/communicating-climate-change [accessed 2026-01-31]
43. Caplan MA, Purser G. Qualitative inquiry using social media: a field-tested example. Qual Soc Work 2019 May;18(3):417-435. [doi: 10.1177/1473325017725802]
44. Reddit heat Canada study. GitHub. URL: https://github.com/szamani20/Reddit-Heat-Canada-Study [accessed 2026-02-05]

Abbreviations

API: application programming interface
BERTopic: bidirectional encoder representations topic modeling

Edited by A Stone, A Mavragani; submitted 14.Aug.2025; peer-reviewed by AW Bello, Q Ng; revised version received 28.Oct.2025; accepted 28.Oct.2025; published 25.Feb.2026.

Please cite as:
MacKay M, Zamani Alavijeh S, Gosselin S, Zarrinkalam F, McWhirter JE
Exploring Reddit Discourse and Information Needs Surrounding Extreme Heat: Topic, Sentiment, and Engagement Analysis
J Med Internet Res 2026;28:e82426
URL: https://www.jmir.org/2026/1/e82426
doi:10.2196/82426

© Melissa MacKay, Soroush Zamani Alavijeh, Sydney Gosselin, Fattane Zarrinkalam, Jennifer E McWhirter. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Detection of Antithrombotic-Related Bleeding in Older Inpatients: Multicenter Retrospective Study Using Structured and Unstructured Electronic Health Record Data

Claire Coumau[1,2,3], PhD; Frederic Gaspar[1,2,3], PhD; Mehdi Zayene[4], MSc; Elliott Bertrand[4], MSc; Lorenzo Alberio[5], Prof Dr Med; Christian Lovis[6], Prof Dr Med; Patrick E Beeler[7,8], PD Dr med; Fabio Rinaldi[9,10,11], Prof Dr; Monika Lutters[12], PhD; Marie-Annick Le Pogam[13*], MD, PhD; Chantal Csajka[1,2,3*], Prof Dr; SwissMADE Collaborators[14,17]

[1]Center for Research and Innovation in Clinical Pharmaceutical Sciences, Lausanne University Hospital and University of Lausanne, Rue du Bugnon 19, Lausanne, Switzerland

[10]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

[11]Faculty of Biology and Medicine, Fondazione Bruno Kessler, Trento, Italy

[12]Head of Hospital Pharmacy, Kantonsspital Aarau, Aarau, Switzerland

[13]Department of Epidemiology and Health Systems, Unisanté, University Center for Primary Care and Public Health & University of Lausanne, Lausanne, Switzerland

[14]See Acknowledgments

[17]

[2]School of Pharmaceutical Sciences, University of Geneva, Geneva, Switzerland

[3]Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, Geneva, Switzerland

[4]Artefact Company, Lausanne, Switzerland

[5]Service of Haematology and Central Haematology Laboratory, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

[6]Division of Medical Information Sciences, Geneva University Hospitals and University of Geneva, Geneva, Switzerland

[7]Division of Occupational and Environmental Medicine, Epidemiology, Biostatistics and Prevention Institute, University of Zurich and University Hospital Zurich, Zurich, Switzerland

[8]Center for Clinical Research, University of Lucerne, Lucerne, Switzerland

[9]Dalle Molle Institute for Artificial Intelligence Research, IDSIA USI-SUPSI, Lugano, Switzerland

[*]these authors contributed equally

Corresponding Author:
Chantal Csajka, Prof Dr
Center for Research and Innovation in Clinical Pharmaceutical Sciences, Lausanne University Hospital and University of Lausanne, Rue du Bugnon 19, Lausanne, Switzerland

Abstract

Background: Bleeding complications are a major contributor to adverse drug events among older inpatients, particularly in those treated with antithrombotic agents. Timely and accurate detection of bleeding events is essential for improving drug safety surveillance and clinical risk management.

Objective: The study aimed to develop and validate automated algorithms for detecting major bleeding (MB) and clinically relevant nonmajor bleeding (CRNMB) events from electronic medical records (EMRs) by combining structured data-based rule models and a natural language processing (NLP) approach, and to evaluate their performance and generalizability against a manually reviewed gold standard and an external dataset.

Methods: We conducted a multicenter retrospective study using routinely collected EMR data from 3 Swiss university hospitals. Patients 65 years or older who received at least one antithrombotic agent and were hospitalized between January 2015 and December 2016 were included. To detect MB and CRNMB events, rule-based algorithms were developed using structured data (*International Statistical Classification of Diseases, 10th Revision, German Modification* [*ICD-10-GM*] codes, laboratory values, transfusion records, and antihemorrhagic prescriptions), with variables and cutoff values defined according to adapted International Society on Thrombosis and Haemostasis definitions and expert consensus. In parallel, a supervised NLP model was applied to discharge summaries from one hospital. A manual review of 754 EMRs served as the reference standard for internal validation, and the algorithm performance of the structured data algorithms (SDA), NLP, and their combination (SDA+NLP) was evaluated against this manually reviewed gold standard using standard performance metrics. External validation was performed on an independent dataset from the Lausanne University Hospital to assess model robustness and generalizability.

XSL•FO
RenderX

Results: Among 36,039 inpatient stays, SDA identified 8.26% (n=2979) as MB and 15.04% (n=5419) as CRNMB cases. *ICD-10-GM* codes alone detected 28.5% (n=849) of MB and 31.48% (n=1706) of CRNMB cases, while laboratory data contributed most to event detection (n=1994, 66.94% for MB and n=3663, 67.60% for CRNMB). Integrating SDA with NLP improved detection, identifying 12.2% (920/7513) of MB and 27.4% (2062/7513) of CRNMB cases at 1 hospital. The combined model achieved the best performance (sensitivity 0.84, positive predictive value 0.51, F_1-score 0.64). External validation on Lausanne University Hospital 2021 - 2022 data (n=24,054 stays) confirmed the algorithms' reproducibility; the prevalence of MB decreased while CRNMB increased, reflecting evolving clinical practices and antithrombotic use patterns.

Conclusions: Our integrated approach, combining SDA with NLP, enhances the detection of hemorrhagic events in older hospitalized patients treated with antithrombotic agents, suggesting its potential usefulness for drug safety monitoring and clinical risk management.

International Registered Report Identifier (IRRID): RR2-10.2196/40456

(J Med Internet Res 2026;28:e77809) doi:10.2196/77809

KEYWORDS

adverse drug events; adverse drug reactions; older inpatients; structured data mining; machine learning; natural language processing; electronic medical records; multicenter study; antithrombotic; hemorrhage; artificial intelligence; pharmacovigilance

Introduction

Over 16% of older inpatients experience at least 1 adverse drug event (ADE) during their hospital stay [1], often with more severe consequences than in younger patients [2]. Among the medications most frequently implicated, antithrombotic agents, widely prescribed in older adults for the prevention and treatment of cardiovascular disease, stand out as a major cause of bleeding-related ADEs [1,3]. Hemorrhagic complications represent a substantial share of drug-related harm in this population and are associated with longer hospital stays, higher readmission rates, and increased mortality. Continuous and accurate measurement of these events is therefore essential to inform prevention strategies, strengthen pharmacovigilance, and promote safer antithrombotic use in clinical practice.

Various approaches have been developed to detect ADEs in hospital settings, each with advantages and limitations. Spontaneous reporting systems, though simple to implement, notoriously underestimate the true frequency of ADEs due to underreporting [4]. Systematic chart reviews of electronic medical records (EMRs), often considered the reference standard, provide detailed clinical information but are too resource- and time-intensive for routine surveillance [5]. To overcome these constraints, automated detection methods using routinely collected EMR data have emerged. These approaches leverage both structured data, such as diagnostic codes, medication records, laboratory results, and vital signs, and unstructured clinical narratives, including discharge summaries, progress notes, and consultation reports. Structured data are accessible and standardized, supporting large-scale analyses but may lack contextual nuances needed to capture complex clinical events such as bleeding [6-8]. Conversely, textual data, although unstructured, often contain richer clinical detail but require advanced computational methods for analysis. Recent advances in machine learning (ML) and natural language processing (NLP) have markedly improved the ability to extract this information and are increasingly applied to pharmacovigilance and ADE detection [9]. Integrating both structured and textual data appears particularly promising for identifying bleeding events, potentially enhancing accuracy and completeness [10].

Despite growing interest in automated ADE detection to support drug safety monitoring, important knowledge gaps remain, particularly in the Swiss context. Most existing studies focusing on bleeding events have relied exclusively on either structured or unstructured data [11-15], have prioritized prediction rather than detection [8,16], or have focused on specific bleeding types or patient groups [7,10,17-19]. Furthermore, clear operational definitions distinguishing major bleeding (MB) from clinically relevant nonmajor bleeding (CRNMB) are often lacking, limiting comparability across studies [20]. To date, no study in Switzerland has comprehensively evaluated the combined contribution of structured and textual data for ADE detection in a general inpatient population receiving antithrombotic therapy. To address this gap, we conducted a multicenter study integrating rule-based algorithms and NLP to detect MB and CRNMB events among older inpatients treated with antithrombotics. We hypothesized that combining structured and textual EMR data would improve the accuracy and completeness of bleeding event identification compared with using either data source alone. The study aimed to develop rule-based algorithms for bleeding detection from structured data sources (diagnoses, laboratory results, transfusions, and antihemorrhagic prescriptions) based on international definitions; design and train an NLP model to identify bleeding mentions in discharge summaries; assess and compare the diagnostic performance of structured data algorithms (SDA), NLP, and their combination (SDA+NLP) against a manually reviewed gold standard; and evaluate the generalizability of the best-performing models through external validation on an independent dataset.

Methods

Study Design

We conducted a multicenter cross-sectional study using retrospective data covering the period from January 1, 2015, to December 31, 2016. Data were obtained from 4 large Swiss hospitals: Lausanne University Hospital (CHUV; approximately

XSL•FO
RenderX

1500 beds [21]), Geneva University Hospital (HUG; approximately 2000 beds [22]), both located in the French-speaking region and serving the cantons of Vaud and Geneva, respectively, Zürich University Hospital (USZ; approximately 900 beds [23]) serving the Zurich metropolitan area, and Baden Cantonal Hospital (KSB; approximately 400 beds [24]) serving the canton of Aargau in the German-speaking region. This study was conducted in accordance with the SRTOBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement (Checklist 1).

The 2015 - 2016 dataset was used for algorithm development as it was the most recent period with harmonized, high-quality structured and unstructured EMR data across all hospitals. Later years were excluded due to EMR vendor transitions, database restructuring, and new data-governance restrictions limiting access to deidentified text. A more recent CHUV dataset (2021 - 2022) was used for temporal and external validation to test algorithm robustness under evolving clinical practices and documentation standards.

Study Participants and Hospital Stays

Eligible participants were Swiss residents 65 years or older treated with at least 1 antithrombotic agent during their hospital stay. Antithrombotic agents included vitamin K antagonists, heparins, platelet aggregation inhibitors, direct thrombin inhibitors, direct factor Xa inhibitors, or fondaparinux. Hospitalizations had to last at least 24 hours and to occur between January 2015 and December 2016 (test dataset). For the external validation, an additional dataset from CHUV covering January 2021 to December 2022 was used (validation dataset). Only patients who had provided explicit consent for the reuse of their health data for research purposes, as indicated by the signature of the general consent form, were eligible for inclusion. Hospital stays lasting less than 24 hours were excluded from the analysis.

Data Sources and Preprocessing

Each participating hospital extracted relevant clinical data from its institutional data warehouses for all inpatient stays meeting the inclusion criteria. The extracted datasets included both structured and unstructured data. Structured data comprised administrative information, patient movements within the hospital, key clinical and laboratory parameters, and prescribed medications coded using the anatomical therapeutic chemical classification. Diagnostic codes were drawn from the *International Statistical Classification of Diseases, 10th Revision, German Modification* (*ICD-10-GM*), and procedures were coded according to the Swiss Classification of Surgical Procedures (CHOP). Diagnoses and procedures were obtained from the hospital billing records associated with each inpatient stay. Unstructured data included discharge summaries. Further details on data extraction and handling are available in the published study protocol [25].

Prior to analysis, structured data were cleaned, harmonized, and verified for consistency at each site, then locally deidentified before being transferred to a centralized database hosted at CHUV. Unstructured data were deidentified and, where necessary, converted into machine-readable formats, but were stored locally on secure hospital servers to comply with data governance policies. Due to the extent of missing and inconsistent information, such as discrepancies in data structure, coding systems, variable definitions, and extensive missing values, reliable harmonization of KSB data with the other hospitals was not feasible, and data from KSB were excluded from the analysis. In addition, only unstructured data from CHUV were analyzed, as full deidentification of textual data from the other sites could not be ensured. The same preprocessing workflow was applied to the 2021 - 2022 CHUV dataset used for external validation. An overview of the data processing workflow is provided in Figure 1.

Figure 1. Overview of the data extraction and preprocessing pipeline for structured and unstructured electronic medical record (EMR) data. CHOP: Swiss Classification of Surgical Procedures; CHUV: Lausanne University Hospital; *ICD-10-GM*: *International Statistical Classification of Diseases, 10th Revision, German Modification*.

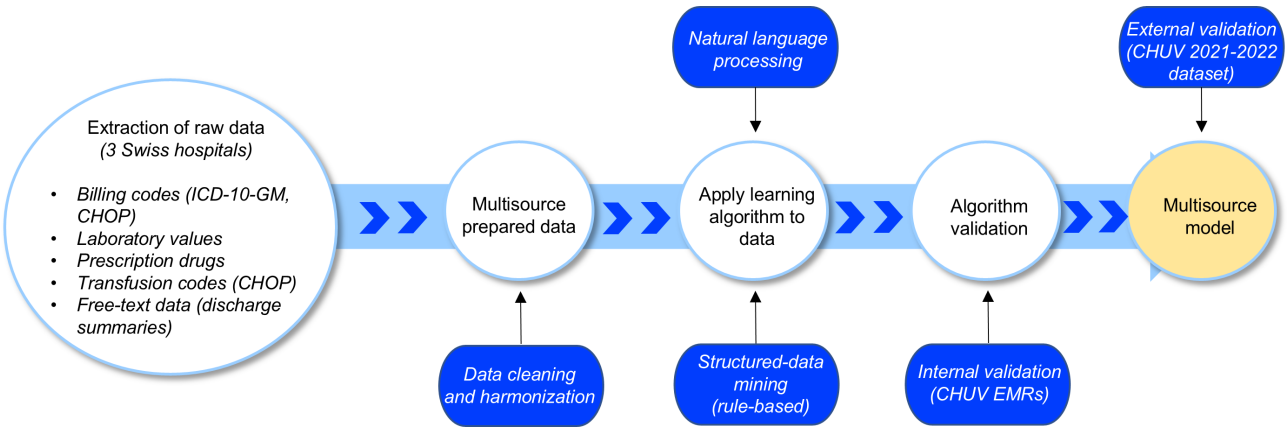

Bleeding Detection Algorithms

We selected variables of interest and cutoff values for our algorithms based on an adaptation of the International Society on Thrombosis and Haemostasis (ISTH) definitions of MB [26] and CRNMB [27], informed by an extensive review of international guidelines (Multimedia Appendix 1). MB was defined as a hemoglobin drop of 4 g/dL or more within 48 hours, a 2 to 4 g/dL drop associated with death within 24 hours, a hemoglobin level less than 7 g/dL, a hemoglobin level between 7 and 9 g/dL associated with death within 24 hours, or a transfusion of more than 5 units of blood or red blood cells. CRNMB was defined as a hemoglobin drop of 2 to 4 g/dL within

XSL•FO
RenderX

48 hours not associated with death or a hemoglobin nadir between 7 and 9 g/dL without subsequent death. The ISTH hemoglobin thresholds were adapted to improve specificity in older inpatients and to reduce the risk of misclassifying nonhemorrhagic anemias. The 5-unit transfusion threshold was pragmatically chosen due to the limited granularity of CHOP procedural codes. Additional structured indicators were also integrated to refine case classification: the prescription of antihemorrhagic agents (idarucizumab, andexanet alfa, prothromplex, octaplex, and beriplex) was considered indicative of MB. MB-related in-hospital mortality was defined as any hospital stay involving at least 1 MB event followed by death during the same admission. We then developed rule-based algorithms using Boolean logic to detect MB and CRNMB cases from structured data (Figure 2).

Figure 2. Algorithmic framework for detection of major bleeding (MB) and nonmajor clinically relevant bleeding (CRNMB) cases using structured data. Antihemorrhagic agent: idarucizumab, andexanet alfa, prothromplex, octaplex, and beriplex. ΔHb: drop in hemoglobin levels within 48 hours; *ICD-10-GM*: *International Statistical Classification of Diseases, 10th Revision, German Modification*; Min Hb: minimum hemoglobin value during the stay.

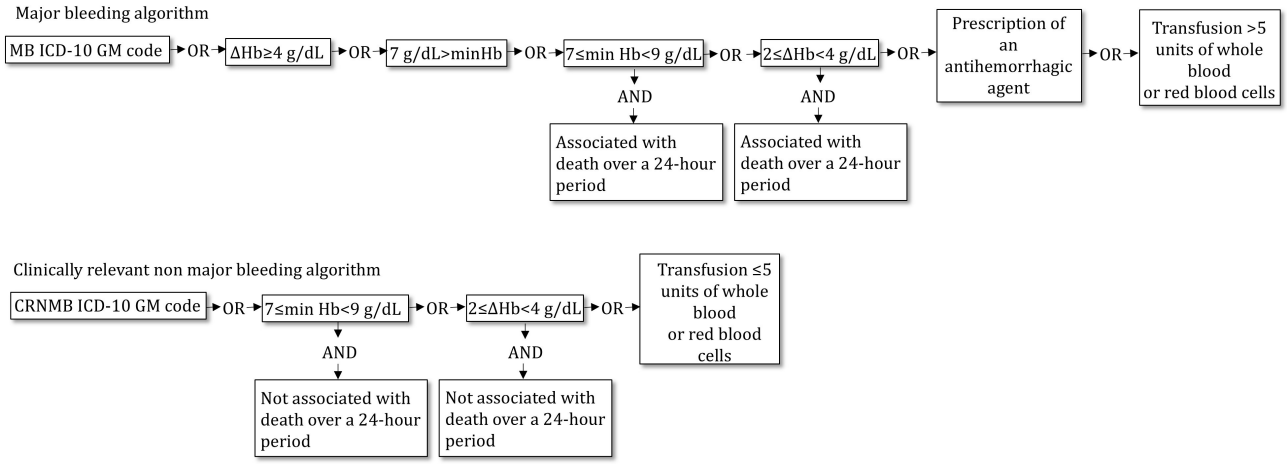

The *ICD-10-GM* list comprised 12 codes for MB and 41 codes for CRNMB (Table 1). We defined an MB in-hospital mortality case as a stay containing an MB occurring during hospitalization followed by death of the patient during the same hospitalization period. We measure the prevalence of bleeding cases, corresponding to inpatient stays with at least 1 MB or CRNMB event, either present on admission or occurring during hospitalization. We quantified the relative and absolute contribution of each structured data source (diagnoses, laboratory, transfusions, and medications), both individually and in combination, in terms of overall detection capacity and proportion of identified bleeding events.

XSL•FO
RenderX

Table . Lists of deficient systems/organs and distribution of *ICD-10-GM*[a] chapters and codes identifying MB[b] and CRNMB[c] cases.

Types of hemorrhagic events	*ICD-10-GM* codes
MB	
Hyphema	H21.0
Hemorrhage and rupture of the choroid	H31.3
Retinal, vitreous, or subarachnoid hemorrhage	H35.6, H43.1, I60_
Hemopericardium not classified elsewhere	I31.2
Intracerebral hemorrhage	I61_
Other nontraumatic intracranial hemorrhages	I62_
Hemoperitoneum	K66.1
Hemarthrosis	M25.0_
Hypovolemic shock	R57.1
Shock during or after a procedure for diagnostic and therapeutic purposes, not classified elsewhere	T81.1
CRNMB	
Conjunctival hemorrhage	H11.3
Otorrhagia	H92.2
Hemorrhagic esophageal varices	I85.0, I98.3
Other specified diseases of the esophagus	K22.8
Non-traumatic hemothorax	J94.2
Gastric, duodenal, or gastrojejunal ulcer with hemorrhage and/or perforation	K25.0, K25.2, K25.6, K26.0, K26.2, K26.4, K26.6, K27.0, K27.2, K27.4, K27.6, K28.0, K28.2, K28.4, K28.6
Acute hemorrhagic gastritis	K29.0
Rectal and anal hemorrhage	K62.5
Hematemesis	K92.0
Melena	K92.1
Unspecified gastrointestinal hemorrhage	K92.2
Prostatic congestion and hemorrhage	N42.1
Hematoma of the broad ligament	N83.7
Hematometra	N85.7
Abnormal bleeding from the uterus and vagina	N93.8, N93.9
Postmenopausal bleeding	N95.0
Epistaxis	R04.0
Throat hemorrhage	R04.1
Hemoptysis	R04.2
Respiratory tract hemorrhage	R04.8, R04.9
Spontaneous ecchymosis	R23.3
Unspecified hematuria	R31
Hemorrhage, not classified elsewhere	R58, T81.0

[a]*ICD-10-GM*: *International Classification of Diseases, 10th Revision, German Modification*.

[b]MB: major bleeding.

[c]CRNMB: clinically relevant nonmajor bleeding.

XSL•FO
RenderX

Natural Language Processing Model

To complement structured data detection, we developed a supervised ML model to identify MB, CRNMB, and past bleeding cases documented in discharge summaries.

A dataset of 400 discharge summaries from CHUV was randomly divided into a training set (n=280) and a test set (n=120), including 100 summaries with MB, 100 with CRNMB, and 200 with no bleeding. Three independent physicians manually annotated the 400 discharge summaries using 4 mutually exclusive labels: (A) 'presence of CRNMB,' (B) 'presence of MB' (as previously defined), (C) 'history of bleeding' (when a discharge summary mentioned bleeding in the EMR before the hospital admission), and (D) 'absence of any bleeding.' Preprocessing steps included tokenization, lemmatization, and sentence segmentation using the French spaCy model (v3.0) [28]. The classification pipeline combined logistic regression and support vector machine models, selected for their interpretability and robustness with limited training data. We deliberately used a classical supervised ML model rather than deep learning architectures to ensure interpretability, reproducibility, and computational efficiency, which are essential for clinical validation and routine pharmacovigilance applications. This approach also better suited the relatively small, annotated corpus, allowing transparent feature weighting and easier auditability across institutions. The model was trained using the scikit-learn library (Python v3.9.1). The classification pipeline proceeded in 3 stages: step 1: binary classifier to identify bleeding-relevant versus irrelevant sentences; step 2: multiclass classifier to distinguish between *irrelevant*, *antecedent bleeding*, and *active bleeding*; step 3: binary classifier to further differentiate between MB and CRNMB within sentences flagged as active bleeding. Sentence-level predictions were aggregated to assign a final label to each document. Rules were prioritized as follows: MB>CRNMB>history of bleeding>no bleeding. This ensured a conservative classification hierarchy, favoring the identification of more severe bleeding cases when multiple labels were present. Further methodological details are available in the study proposal previously published [25], the related article [29], and summarized in Figure 3.

Figure 3. Natural language processing workflow from raw text input to final classification output. BERT: Bidirectional Encoder Representations From Transformers; DS: discharge summary; HBDSCAN: Hierarchical Density-Based Spatial Clustering of Applications With Noise; UMAP: Uniform Manifold Approximation and Projection.

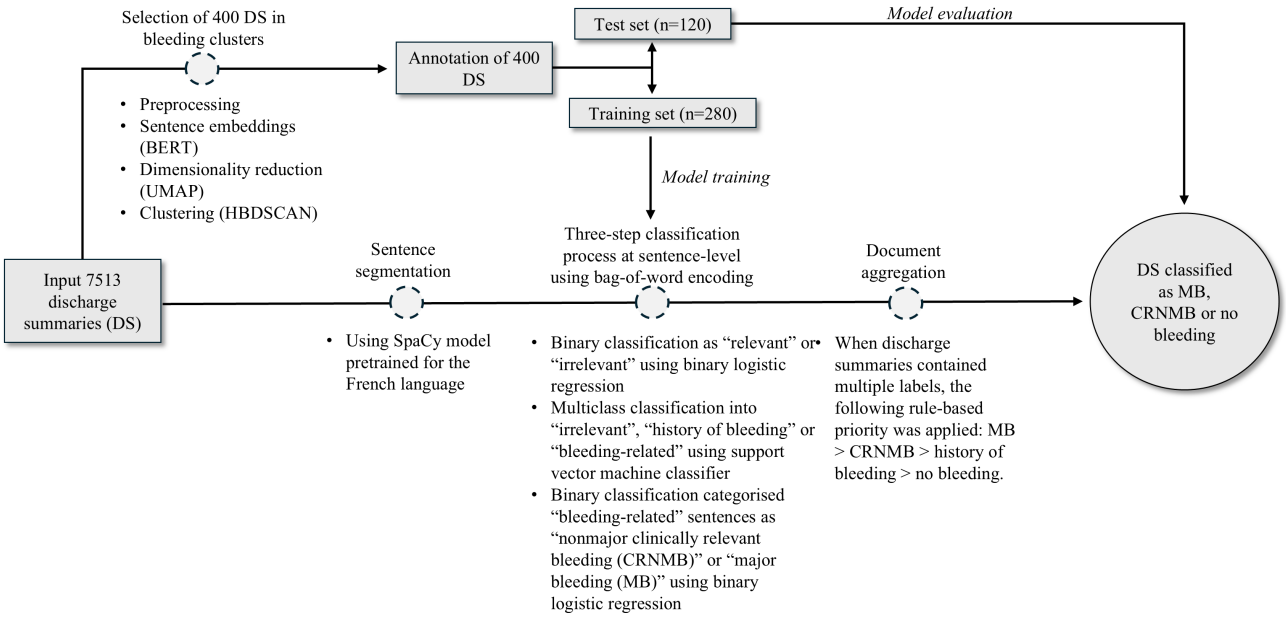

Validation of the Bleeding Detection Algorithms

Internal Validation Using CHUV 2015-2016 Data

To validate the SDA and SDA combined with NLP (SDA+NLP) models, we conducted a manual review of 754 EMRs from CHUV's 2015 - 2016 dataset. The sample size for validation was determined using a test result-based sampling method [30]. Assuming a 7% MB, a 10% CRNMB accuracy, and a sensitivity of 0.7, at least 704 EMRs had to be reviewed, and 754 EMRs were effectively reviewed. Four physicians independently reviewed the records to compare algorithm-detected with clinician-identified MB and CRNMB cases. The review process followed a structured protocol aligned with ISTH definitions [26,27] and adapted for retrospective application to routinely collected hospital data. Reviewers assessed each inpatient stay according to 4 key criteria: (1) evidence of active bleeding, (2) severity of the event (eg, hemodynamic instability), (3) need for therapeutic intervention (eg, transfusion volume, administration of antihemorrhagic agents), and (4) temporal relationship to hospital admission (present on admission versus occurred during stay). A complete list of synonyms used to identify MB and CRNMB cases during manual chart review is provided in Multimedia Appendix 2. Two binary classification scenarios were evaluated: (1) MB versus all other cases (CRNMB or no bleeding), and (2) CRNMB versus no bleeding (excluding MB). Algorithm performance was evaluated at the inpatient-stay level using standard binary classification metrics (sensitivity, specificity, positive predictive value [PPV], negative predictive value, accuracy, and F_1-score), with manual chart review as the gold standard. Comparisons between SDA, NLP, and combined models were descriptive, and sensitivity was prioritized due to the study's patient safety focus. Interrater

XSL•FO
RenderX

reliability among reviewers was evaluated using Fleiss κ on a subset of 40 cases, with agreement levels interpreted according to Landis and Koch [31] (>0.80: *almost perfect*; 0.61 - 0.80: *substantial*; 0.41 - 0.60: *moderate*; 0.21 - 0.40: *fair*; 0.00 - 0.20: *slight*; <0.00: *poor*). A *P* value associated with the Fleiss κ coefficient was also calculated, with a *P* value less than .05 indicating statistically significant agreement. Additional details and results are provided in Multimedia Appendix 3. In a subanalysis, a causal relationship between antithrombotic therapy and each bleeding event was also assessed during the manual review, using a structured tool based on temporal association, biological plausibility, and alternative explanations. Cases were rated as *certain*, *probable*, *possible*, or *unclassified*, in relation to antithrombotic exposure, according to the WHO-Uppsala Monitoring Center scale [32]. The methodology, sample size calculation, and findings of the causality assessment of the subanalysis are presented in Multimedia Appendix 4.

External Validation Using CHUV 2021-2022 Data

An external validation was performed using CHUV data from 24,054 inpatient stays between January 2021 and December 2022. We applied the same detection algorithms (SDA and SDA+NLP) to this independent dataset to evaluate their performance, robustness, and reproducibility. Results were compared to those from the 2015 - 2016 CHUV dataset.

Statistical Analysis

Descriptive statistics were used to summarize population characteristics. Comorbidity was assessed using the Charlson and Elixhauser indexes [33,34], which are validated tools for risk adjustment and mortality prediction based on administrative health data. Comparisons of patient characteristics between hospitals were conducted using a 1-way analysis of variance on ranks (Kruskal-Wallis test) for continuous variables and Pearson χ^2 test for categorical variables. Hyperparameters of the NLP classifier were optimized through 5-fold cross-validation on the training set, and final performance was estimated on an independent test set. All performance metrics were reported with 95% CIs calculated using the Wilson method. Analyses were conducted using StataCorp. 2021. Stata Statistical Software: Release 17. College Station, TX: StataCorp LLC software for structured data and Python (v3.9.1) for NLP development.

Ethical Considerations

Human Subject Ethics Review Approvals or Exemptions

This study was conducted in accordance with the *Declaration of Helsinki* and Swiss federal regulations governing research on human data. Ethical approval was obtained from all relevant cantonal ethics committees, coordinated by the lead committee of the Canton of Vaud (CER-VD No. 2018 - 00272). As the study involved secondary analysis of routinely collected, deidentified hospital data, it qualified for a simplified review under Swiss Human Research Act article 2, paragraph 2(c).

Informed Consent

The study relied exclusively on existing clinical data that were deidentified before analysis. According to Swiss regulations and institutional data governance policies, informed consent was waived for patients who had not explicitly objected to the use of their medical data for research purposes. All participating hospitals operate an institutional opt-out procedure, allowing patients to refuse the secondary use of their data for research.

Privacy and Confidentiality

All data were deidentified at source before analysis. Structured data were transferred through secure institutional channels to a restricted-access research environment hosted at CHUV. Unstructured textual data remained stored locally on hospital servers and were processed within each institution's secure infrastructure to comply with data protection requirements. No directly identifiable information was accessible to the investigators.

Compensation Details

No compensation was provided to patients, as the study involved secondary analysis of preexisting, routinely collected data and did not include direct contact with participants.

Protection of Identifiable Information in Figures and Supplementary Materials

No image, document, or figure contains any identifiable patient information. Consequently, no individual consent for image publication was required.

Ethics Approval

Approved by the Cantonal Ethics Committee of Vaud, Switzerland (CER-VD No. 2018 - 00272); informed consent was waived for patients who did not opt out of research data use.

Results

Study Population Characteristics

A total of 36,039 inpatient stays, involving 24,991 unique patients, were included in the analysis: 7677 stays (5754 patients) at CHUV, 18,015 stays (11,356 patients) at HUG, and 10,347 stays (7881 patients) at USZ. Patient characteristics are detailed in Table 2. The median age at admission was 78 (IQR 65 - 99) years, with a balanced sex distribution (51.40% male). Comorbidity was generally low across the cohort, with a median Charlson index and Elixhauser index of 0.0; USZ patients had the lowest overall comorbidity burden.

XSL•FO
RenderX

Table . Baseline patient characteristics and treatments: overall and by university hospital.

Characteristics	All hospitals (n=36,039)[a]	CHUV[b] (n=7677)	HUG[c] (n=18,015)	USZ[d] (n=10,347)
Admission age (years), median (IQR)	78 (65 - 99)	79 (65 - 99)	80 (65 - 99)	75 (65 - 92)
Sex, n (%)				
Male	18,525 (51.40)	3987 (51.93)	8638 (47.95)	5900 (57.02)
Female	17,514 (48.60)	3690 (48.07)	9377 (52.05)	4447 (42.98)
Length of stay (d), median (IQR)	9 (1-342)	9 (1-293)	12 (1 - 342)	6 (1-145)
Transfer to intensive care, n (%)	1534 (4.26)	467 (6.1)	1067 (5.92)	__[e]
In-hospital mortality, n (%)	1416 (3.93)	345 (4.5)	850 (4.7)	221 (2.1)
Comorbidity, n (%)				
Chronic renal dysfunction	8418 (23.36)	2163 (28.18)	5662 (31.43)	593 (5.7)
Dialysis	622 (1.7)	176 (2.3)	241 (1.3)	205 (2.0)
Acute renal dysfunction	1151 (3.19)	266 (3.5)	623 (3.5)	262 (2.5)
Chronic liver dysfunction	1020 (2.83)	294 (3.8)	497 (2.8)	229 (2.2)
Acute liver dysfunction	498 (1.4)	145 (1.9)	244 (1.4)	109 (1.1)
Hypertension	18,316 (50.82)	3158 (41.14)	9271 (51.46)	5887 (56.90)
Alcohol abuse	1354 (3.76)	388 (5.1)	663 (3.7)	303 (2.9)
Stroke	3001 (8.33)	813 (10.6)	1625 (9.02)	563 (5.4)
Cancer	6776 (18.80)	1572 (20.48)	2905 (16.13)	2299 (22.22)
Platelet coagulation defect	2178 (6.04)	496 (6.5)	1029 (5.71)	653 (6.3)
Anemia	7624 (21.15)	1998 (26.03)	4380 (24.31)	1246 (12.04)
Risk fall	11,376 (31.57)	2932 (38.20)	6021 (33.42)	2423 (23.42)
Diabetes	6638 (18.42)	1314 (17.12)	3575 (19.84)	1749 (16.90)
Recent myocardial infection	1923 (5.34)	609 (7.9)	761 (4.2)	553 (5.3)
Low weight	4059 (11.26)	967 (12.6)	2533 (14.06)	559 (5.4)
Thrombolysis	695 (1.9)	180 (2.3)	512 (2.8)	3 (0.0)
Vascular malformation	955 (2.6)	153 (2.0)	334 (1.9)	468 (4.5)
Charlson comorbidity index, median (IQR)	0.0 (0.0 - 9.0)	0.0 (0.0 - 9.0)	0.0 (0.0 - 7.0)	0.0 (0.0 - 7.0)
Elixhauser comorbidity index, median (IQR)	0.0 (0.0 - 6.0)	0.0 (0.0 - 6.0)	0.0 (0.0 - 6.0)	0.0 (0.0 - 5.0)
Antithrombotic categories, n (%)				
Direct factor Xa inhibitors	3297 (9.15)	599 (7.8)	1478 (8.20)	1220 (11.80)
Vitamin K antagonists	7469 (20.72)	1324 (17.25)	4943 (27.44)	1202 (11.62)
Heparin group	24,784 (6877)	5045 (65.71)	11,918 (66.17)	7821 (75.59)
Direct thrombin inhibitors	255 (0.7)	87 (1.1)	134 (0.7)	34 (0.3)
Platelet aggregation inhibitors	14,220 (39.46)	4354 (56.71)	4700 (26.09)	5166 (49.93)
Thrombolytics	104 (0.3)	15 (0.2)	89 (0.5)	0.0 (0.0)
Other antithrombotic agents: fondaparinux	1365 (3.79)	212 (2.8)	1140 (6.33)	13 (0.1)
Antidotes, n (%)	137 (0.4)	15 (0.2)	122 (0.7)	0.0 (0.0)
Transfusion, n (%)	582 (1.6)	264 (3.4)	318 (1.8)	—

XSL•FO
RenderX

Characteristics	All hospitals (n=36,039)[a]	CHUV[b] (n=7677)	HUG[c] (n=18,015)	USZ[d] (n=10,347)
≤5 UI[f] plasma or red blood cells	225 (0.6)	100 (1.3)	125 (0.7)	—
>5 UI plasma or red blood cells	357 (1.0)	164 (2.1)	193 (1.1)	—
Number of antithrombotic agents received during hospitalization, n (%)				
1	22,397 (62.15)	4257 (55.45)	12,381 (68.73)	5759 (55.66)
2	11,918 (33.07)	2904 (37.83)	4924 (27.33)	4090 (39.53)
3	1641 (4.55)	495 (6.4)	669 (3.7)	477 (4.6)
≥4	83 (0.2)	21 (0.3)	41 (0.2)	21 (0.2)

[a]n: total number of recorded measurements for the respective parameter.

[b]CHUV: Lausanne University Hospital.

[c]HUG: Geneva University Hospital.

[d]USZ: Zürich University Hospital.

[e]Not available (missing or nontransferred data).

[f]UI: unit of blood component.

Distinct prescribing patterns were observed across hospitals: HUG had the highest use of vitamin K antagonists (n=4943, 27.44%), CHUV had the highest prescription rate of antiplatelet agents (n=4354, 56.71%), and USZ reported the highest use of direct factor Xa inhibitors (n=1220, 11.79%) and heparins (n=7821, 75.59%). Hypertension (n=18,316, 50.82%), chronic renal dysfunction (n=8418, 23.36%), anemia (n=7624, 21.15%), and cancer (n=6776, 18.80%) were among the most prevalent comorbidities. Overall, in-hospital mortality was 3.93% (n=1416).

Bleeding Detection Using SDA

SDA detected 8748 (24.27%) overall bleeding cases, of which 2979 (8.26%) were MB cases and 5419 (15.04%) were CRNMB cases (Table 3). Fatal MB occurred in 1.0% (n=350) of all stays. MB prevalence varied across hospitals, with the highest proportion observed at CHUV (n=769, 10.0%), followed by USZ (n=998, 9.6%) and HUG (n=1212, 6.73%). CRNMB prevalence was highest at USZ (n=1682, 16.26%). Missing values for each variable used to identify MB and CRNMB events are presented in Multimedia Appendix 5.

Table . Prevalence of bleeding cases detected by SDA[a], overall and by university hospital[b].

	All hospitals, n (%)	CHUV[c], n (%)	HUG[d], n (%)	USZ[e], n (%)	*P* value[f]
Nonbleeding-related	27,641 (76.70)	5822 (75.84)	14,152 (78.56)	7667 (74.10)	<.001
CRNMB[g]	5419 (15.04)	1086 (14.15)	2651 (14.72)	1682 (16.26)	<.001
MB[h]	2979 (8.26)	769 (10.0)	1212 (6.73)	998 (9.6)	<.001
MB in-hospital mortality	350 (1.0)	119 (1.6)	137 (0.8)	94 (0.9)	<.001
Total	36,039	7677	18,015	10,347	__[i]

[a]SDA: structured data algorithms (ie, rule-based algorithm for structured data).

[b]Bleeding cases: number of stays for patients treated with at least 1 antithrombotic agent during which at least 1 bleeding episode occurred.

[c]CHUV: Lausanne University Hospital.

[d]HUG: Geneva University Hospital.

[e]USZ: Zürich University Hospital.

[f]Using Pearson χ^2 test.

[g]CRNMB: clinically relevant nonmajor bleeding.

[h]MB: major bleeding.

[i]Not applicable.

Relative and Absolute Contribution of Structured Data Sources

Laboratory data were the most influential source for detecting both MB and CRNMB, contributing to two-thirds of identified cases, while *ICD-10-GM* codes contributed to approximately one-third. Prescriptions for antihemorrhagic agents had a minimal added value for MB detection, while transfusion data contributed modestly. Figure 4 illustrates the relative contribution of each data source.

XSL•FO
RenderX

Figure 4. Relative contribution of structured data sources (laboratory data, *ICD-10-GM* codes, prescription of antihemorrhagic agents, and transfusions) to the detection of major bleeding (MB) and clinically relevant nonmajor bleeding (CRNMB). *ICD-10-GM*: *International Statistical Classification of Diseases, 10th Revision, German Modification*.

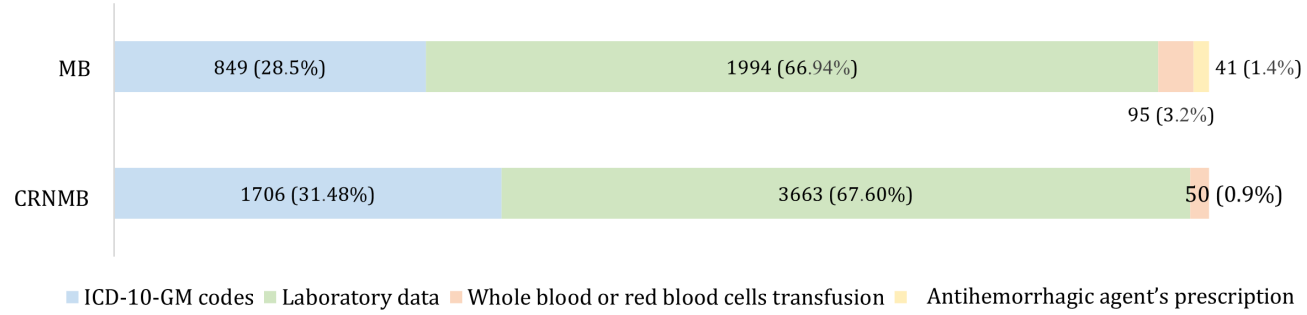

Overlap between data sources was limited. Only 12.1% (n=361) of MB stays and 8.7% (n=458) of CRNMB stays were identified by 2 data sources, while detection by all 4 sources occurred in 0% of MB cases and only 0% (n=12) of CRNMB cases (Figure 5). This limited overlap highlights the complementarity, but also fragmentation, of structured data signals.

Figure 5. Absolute contribution of structured data sources (laboratory data, *ICD-10-GM* codes, prescription of antihemorrhagic agents, and transfusions) to the detection of (A) major bleeding and (B) clinically relevant nonmajor bleeding. *ICD-10-GM: International Statistical Classification of Diseases, 10th Revision, German Modification*.

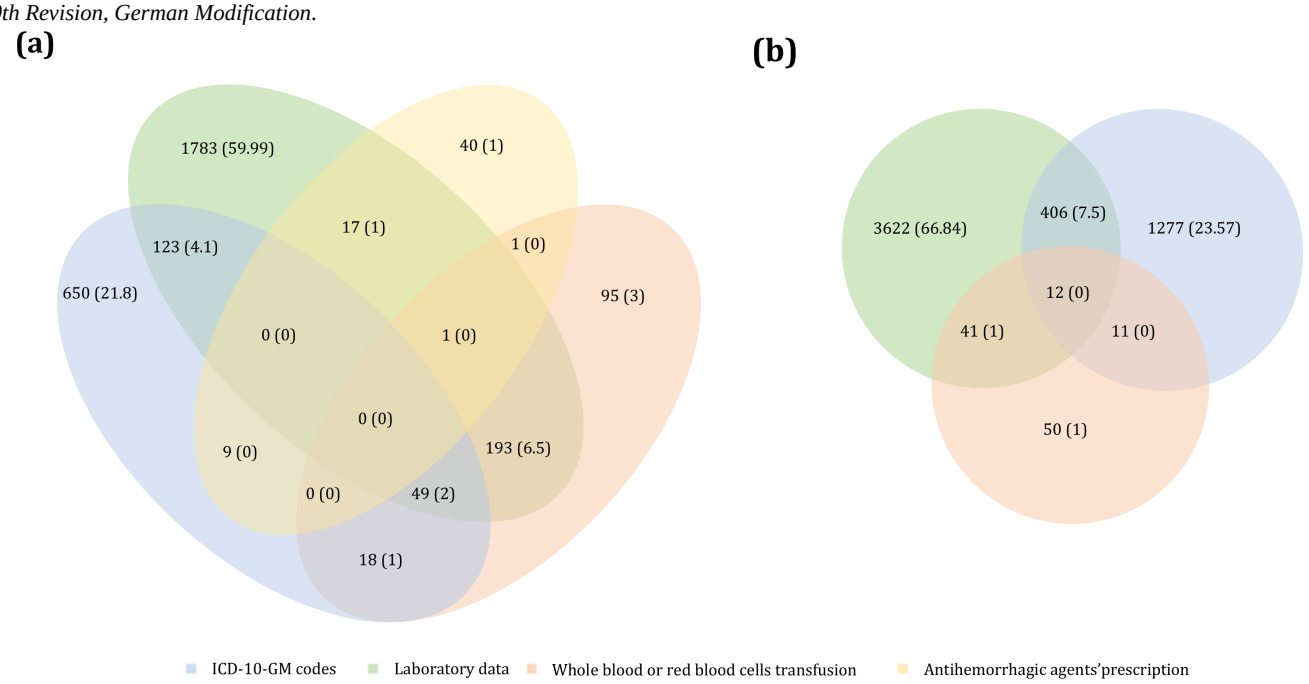

Combined Detection Using SDA and NLP (CHUV Only)

Among 7513 CHUV stays with discharge summaries, combining SDA and NLP increased case detection: In total, 39.69% (n=2982) of hemorrhagic cases were detected: 12.2% (n=920) were identified as MB and 27.45% (n=2062) as CRNMB.

For MB cases, 56.6% (n=521) were detected by SDA alone, 19.8% (n=182) by NLP alone, and 23.6% (n=217) by both. For CRNMB cases, 35.1% (n=724) were detected by SDA alone, 48.2% (n=994) by NLP alone, and 16.7% (n=344) by both.

Classification discrepancies were observed between SDA and NLP: 217 cases identified as MB by SDA were reclassified as CRNMB by NLP, and conversely, 81 CRNMB cases by SDA were reclassified as MB by NLP. NLP also enabled the detection of a history of bleeding in 8.5% (n=642) of cases, improving the temporal resolution of hemorrhage onset.

Internal Validation Using CHUV 2015-2016 Data

The manual review of 754 EMRs identified 276 bleeding cases: 144 MB and 132 CRNMB. Structured laboratory data showed the highest sensitivity (0.58, 95% CI 0.52 - 0.64), while *ICD-10-GM* codes had the highest PPV (0.89, 95% CI 0.83 - 0.98), and F_1-score (0.60). SDA outperformed NLP in sensitivity (0.77 vs 0.61), but NLP had higher PPV (0.70 vs 0.51) and F_1-score (0.65 vs 0.62). The best performance was achieved by the combined *ICD-10-GM*∪NLP algorithm, with a sensitivity of 0.71 (95% CI 0.66 - 0.76), PPV of 0.72 (95% CI 0.66 - 0.87), and F_1-score of 0.72. Algorithms combining SDA and NLP yielded the highest sensitivity (0.84), confirming the benefit of multimodal approaches. However, intersection-based algorithms (eg, SDA∩NLP) demonstrated higher specificity at the cost of reduced sensitivity.

Performance metrics for MB and CRNMB subgroups followed similar trends, with reduced sensitivity but high specificity for

XSL•FO
RenderX

ICD-10-GM–based detection. Table 4 presents a comprehensive summary of all performance metrics, including sensitivity, specificity, PPV, negative predictive value, accuracy, and F_1-score.

XSL•FO
RenderX

Table . Performance metrics of bleeding detection algorithms compared to manual electronic medical records review (gold standard; n=754)[a].

	Sensitivity[b] (95% CI)	Specificity[c] (95% CI)	PPV[d] (95% CI)	NPV[e] (95% CI)	Accuracy[f] (95% CI)	F_1-score[g]
Bleeding all type (MB[h] or CRNMB[i])						
Individual structured data sources						
ICD-10-GM[j]	0.46 (0.40 - 0.51)	0.97 (0.95 - 0.98)	0.89 (0.83 - 0.98)	0.75 (0.72 - 0.79)	0.78 (0.75 - 0.81)	0.60
Laboratory data	0.58 (0.52 - 0.64)	0.60 (0.55 - 0.64)	0.45 (0.40 - 0.64)	0.71 (0.66 - 0.75)	0.59 (0.55 - 0.62)	0.51
Whole blood or red blood cells transfusion	0.18 (0.14 - 0.23)	0.95 (0.93 - 0.97)	0.68 (0.57 - 0.97)	0.67 (0.63 - 0.70)	0.67 (0.63 - 0.70)	0.29
Detection algorithms						
SDA[k]	0.77 (0.72 - 0.82)	0.58 (0.54 - 0.62)	0.51 (0.47 - 0.62)	0.81 (0.77 - 0.85)	0.65 (0.62 - 0.68)	0.62
NLP[l]	0.61 (0.55 - 0.67)	0.85 (0.81 - 0.88)	0.70 (0.64 - 0.88)	0.79 (0.75 - 0.82)	0.76 (0.73 - 0.79)	0.65
Combined data sources and algorithms						
SDA∪NLP	0.84 (0.79 - 0.88)	0.54 (0.49 - 0.58)	0.51 (0.47 - 0.58)	0.85 (0.81 - 0.89)	0.65 (0.61 - 0.68)	0.64
SDA∩NLP	0.47 (0.41 - 0.53)	0.92 (0.89 - 0.94)	0.78 (0.71 - 0.94)	0.75 (0.71 - 0.78)	0.76 (0.72 - 0.79)	0.59
ICD-10-GM∪NLP	0.71 (0.66 - 0.76)	0.84 (0.80 - 0.87)	0.72 (0.66 - 0.87)	0.83 (0.80 - 0.86)	0.79 (0.76 - 0.82)	0.72
ICD-10-GM∩NLP	0.31 (0.26 - 0.37)	0.99 (0.98 - 1.00)	0.95 (0.89 - 1.00)	0.71 (0.68 - 0.74)	0.74 (0.71 - 0.77)	0.47
MB						
Individual structured data sources						
ICD-10-GM	0.34 (0.27 - 0.42)	0.99 (0.97 - 0.99)	0.84 (0.73 - 0.99)	0.86 (0.84 - 0.89)	0.86 (0.84 - 0.88)	0.49
Laboratory data	0.47 (0.39 - 0.55)	0.81 (0.77 - 0.84)	0.36 (0.30 - 0.84)	0.86 (0.83 - 0.89)	0.74 (0.71 - 0.77)	0.41
Whole blood or red blood cells transfusion	0.22 (0.16 - 0.29)	0.97 (0.96 - 0.98)	0.66 (0.52 - 0.98)	0.84 (0.81 - 0.87)	0.83 (0.80 - 0.85)	0.32
Algorithms						
SDA	0.72 (0.64 - 0.78)	0.79 (0.76 - 0.82)	0.45 (0.39 - 0.82)	0.92 (0.90 - 0.94)	0.78 (0.75 - 0.81)	0.55
NLP	0.35 (0.28 - 0.44)	0.95 (0.93 - 0.96)	0.63 (0.51 - 0.96)	0.86 (0.83 - 0.89)	0.84 (0.81 - 0.86)	0.45
Combined data sources and algorithms						
SDA∪NLP	0.76 (0.68 - 0.82)	0.79 (0.75 - 0.81)	0.46 (0.39 - 0.82)	0.93 (0.91 - 0.95)	0.78 (0.75 - 0.81)	0.57
SDA∩NLP	0.30 (0.23 - 0.39)	0.96 (0.94 - 0.97)	0.64 (0.52 - 0.97)	0.85 (0.83 - 0.88)	0.83 (0.81 - 0.86)	0.41
ICD-10-GM∪NLP	0.56 (0.48 - 0.64)	0.94 (0.92 - 0.96)	0.69 (0.60 - 0.96)	0.90 (0.87 - 0.92)	0.87 (0.84 - 0.89)	0.62
ICD-10-GM∩NLP	0.14 (0.09 - 0.21)	1.0 (0.99 - 1.00)	0.91 (0.72 - 1.00)	0.83 (0.80 - 0.86)	0.83 (0.80 - 0.86)	0.25
CRNMB						
Individual structured data sources						
ICD-10-GM	0.30 (0.23 - 0.39)	0.91 (0.88 - 0.93)	0.41 (0.32 - 0.93)	0.86 (0.83 - 0.88)	0.80 (0.77 - 0.83)	0.35
Laboratory data	0.25 (0.18 - 0.33)	0.73 (0.70 - 0.77)	0.17 (0.12 - 0.77)	0.82 (0.79 - 0.85)	0.65 (0.61 - 0.68)	0.20
Whole blood or red blood cells transfusion	0.03 (0.01 - 0.07)	0.96 (0.95 - 0.98)	0.15 (0.06 - 0.98)	0.82 (0.79 - 0.85)	0.80 (0.77 - 0.83)	0.05
Detection algorithms						

XSL•FO
RenderX

	Sensitivity[b] (95% CI)	Specificity[c] (95% CI)	PPV[d] (95% CI)	NPV[e] (95% CI)	Accuracy[f] (95% CI)	F_1-score[g]
SDA	0.65 (0.42 - 0.58)	0.65 (0.62 - 0.69)	0.23 (0.19 - 0.69)	0.86 (0.82 - 0.89)	0.63 (0.59 - 0.66)	0.32
NLP	0.53 (0.45 - 0.62)	0.77 (0.74 - 0.81)	0.34 (0.28 - 0.81)	0.89 (0.86 - 0.91)	0.73 (0.70 - 0.76)	0.41
Combined data sources and algorithms						
SDA∪NLP	0.66 (0.57 - 0.73)	0.56 (0.52 - 0.60)	0.24 (0.20 - 0.60)	0.88 (0.85 - 0.91)	0.58 (0.54 - 0.61)	0.35
SDA∩NLP	0.38 (0.30 - 0.47)	0.88 (0.84 - 0.90)	0.40 (0.32 - 0.90)	0.87 (0.84 - 0.89)	0.79 (0.76 - 0.82)	0.39
ICD-10-GM∪NLP	0.60 (0.52 - 0.68)	0.75 (0.72 - 0.79)	0.35 (0.29 - 0.79)	0.90 (0.87 - 0.92)	0.73 (0.69 - 0.76)	0.44
ICD-10-GM∩NLP	0.24 (0.17 - 0.32)	0.93 (0.91 - 0.95)	0.42 (0.31 - 0.95)	0.85 (0.82 - 0.87)	0.81 (0.78 - 0.83)	0.30

[a]It should be noted that no patient record contained the variable antihemorrhagic agent for the detection of MB. Consequently, the performance for this variable was not included in the table.

[b]Sensitivity: proportion of bleeding cases that have been correctly identified.

[c]Specificity: proportion of nonbleeding-related cases that have been correctly identified.

[d]PPV: positive predictive value; proportion of bleeding cases among all those classified as bleeding cases by the algorithm.

[e]NPV: negative predictive value; proportion of nonbleeding-related cases among all those classified as nonbleeding-related cases by the algorithm.

[f]Accuracy: overall prediction accuracy (ie, the proportion of bleeding and nonbleeding-related cases that the algorithm has correctly identified.

[g]F_1-score: harmonic mean of the precision and recall (ie, F_1-score = 2 × [recall × precision]/[recall + precision]).

[h]MB: major bleeding.

[i]CRNMB: clinically relevant nonmajor bleeding.

[j]*ICD-10-GM*: *International Statistical Classification of Diseases, 10th Revision, German Modification*.

[k]SDA: structured data algorithm.

[l]NLP: natural language processing.

Interrater reliability for manual review of 40 EMRs showed substantial agreement: Fleiss' Kappa was 0.65 for bleeding detection and 0.61 for MB versus CRNMB classification. Of 276 manually reviewed inpatient stays with bleeding events, 17% (n=48) were attributed to antithrombotic agents. The causal relationship was classified as "certain" in 25% of cases (n=12), "probable/likely" in 23% (n=11), and "possible" in 52% (n=25).

External Validation Using CHUV 2021-2022 Data

Application of the SDA and SDA+NLP algorithms to the CHUV validation dataset (24054 stays) demonstrated generalizability. The prevalence of MB cases significantly decreased from 10.0% in the 2015 - 2016 period to 5.55% (n=1336) in the 2021 - 2022 period, while the prevalence of CRNMB cases increased significantly from 14.15% to 16.63% (n=4000). MB in-hospital mortality also rose, from 1.6% to 2.6% (n=616).

Patient characteristics differed significantly between cohorts (Multimedia Appendix 6). Direct oral anticoagulant prescriptions increased from 7.8% to 22.7%, while vitamin K antagonist use decreased from 17.2% to 7.6%. The incidence of elevated INR values >4 declined from 3.3% to 1.8%. Both Charlson and Elixhauser scores increased, reflecting higher comorbidity. Transfusions involving ≤5 units of blood rose from 1.3% to 8.8%. Notably, the proportion of patients receiving ≥3 antithrombotic agents during hospitalization increased fivefold (from 6.7% to 35.0%).

Discussion

Principal Findings

To our knowledge, this is one of the first multicenter studies assessing the feasibility and effectiveness of combining structured and unstructured EMR data to detect bleeding cases in older inpatients treated by one or more antithrombotic agents. Across 3 large university hospitals, our SDA identified 8.26% of MB and 15.4% of CRNMB cases. Laboratory variables contributed most to event detection, while *ICD-10-GM* codes alone captured only about one-third of cases, achieving a sensitivity of 0.84 when both data sources were combined. These findings confirm the feasibility of automated bleeding surveillance in real-world hospital data and demonstrate the added value of leveraging free-text information to complement structured data sources.

Comparison to Prior Work

Our estimated bleeding rates (MB: 8.26% and CRNMB: 15.04%) are consistent with prior hospital-based studies in older adults, which reported MB incidences ranging from 1.8% to 11.3% [35,36] and CRNMB from 3.5% to 13.0% [35,37]. These findings confirm that antithrombotic-related bleeding remains a major cause of ADEs in older populations, associated with increased hospitalization length, morbidity, and mortality [38], highlighting the need for targeted preventive strategies.

The algorithms' performance varied across structured data sources and aligns with prior research. *ICD-10-GM* codes detected only one-third of MB and CRNMB cases, consistent

XSL•FO
RenderX

with previous evidence of underreporting anticoagulant-related bleeding events [39,40]. Yap et al [15] found similarly low sensitivity (16% - 24%) but very high PPV (>0.97), indicating that diagnostic codes are reliable confirmatory markers but poor screening tools. The inclusion of laboratory data markedly improved sensitivity in our SDA model, consistent with findings by Dyas et al [6] and Shung et al [10]. The modest decline in PPV was likely due to false positives generated by hemoglobin thresholds, a limitation noted in earlier work [15].

Detection of CRNMB was more challenging than MB, partly due to broader definitions and lower specificity of *ICD-10-GM* codes and transfusion data, echoing the moderate performance reported by Yap et al [15] (sensitivity 50% - 56%, PPV 43 - 50%).

The NLP model contributed substantially to overall detection, with a sensitivity of 61% and PPV of 70%, in line with earlier NLP-based models for bleeding and ADE detection [10,41]. Importantly, only about 20% of events overlapped with those captured by structured data, demonstrating that text analysis retrieves unique clinical insights often missing from coded data. NLP also enhanced temporal resolution by identifying prior bleeding episodes in 8.5% of cases, information generally unavailable from structured data alone. The combined SDA+NLP model achieved high sensitivity (0.84), thereby minimizing the risk of missed events, with only 16% of cases being false negatives. Although this proportion is relatively low, it still represents missed hemorrhagic events that could impact the accuracy of retrospective surveillance and safety signal detection. However, our detection algorithm provides a notable proportion of false positives (49%), which could contribute to alert fatigue in clinical practice and increase the workload associated with unnecessary chart reviews. For real-world deployment, performance thresholds depend on the intended use: for surveillance or signal detection, a sensitivity above 0.80 with PPV above 0.50 is generally acceptable, as false positives can be secondarily reviewed; for clinical decision support, stricter thresholds (eg, PPV≥0.70) are needed to prevent alert fatigue. Improving true positive detection to 70% would strengthen reliability and clinical applicability, potentially through prioritization or triage of clinically significant cases.

External validation revealed a decline in MB prevalence and a concurrent increase in CRNMB and MB-related mortality in the validation dataset (2021 - 2022), compared to the CHUV 2015 - 2016 dataset. These trends may reflect evolving prescribing patterns, such as increased use of direct oral anticoagulants and reduced use of vitamin K antagonists, and a shift in clinical profiles, with higher comorbidity scores and greater treatment complexity in the more recent cohort. These observations are consistent with the known bleeding risk profiles of antithrombotic agents, direct oral anticoagulants being more frequently associated with gastrointestinal bleeding (CRNMB), and vitamin K antagonists with intracranial bleeding (MB) [42], and underscore the need for dynamic algorithmic models capable of adjusting for changing treatment patterns and patient characteristics [43].

Strengths and Limitations

This study has several notable strengths. It is one of the first multicenter initiatives to integrate structured and unstructured EMR data for ADE detection in older hospitalized patients. The inclusion of 3 university hospitals provided a large, diverse dataset, while the harmonization of over 1 million clinical variables ensured robust data quality. Algorithms were developed using internationally accepted definitions of MB and CRNMB and validated through manual chart review, ensuring clinical credibility. External validation on a temporally distinct dataset further reinforced reproducibility and robustness.

Several limitations should also be considered.

First, the test dataset (2015 - 2016) was relatively dated and spanned only 2 years, reflecting mostly the time-consuming extraction and harmonization process required to merge data from three hospitals before data interoperability infrastructures were implemented. Consequently, it may not entirely capture current clinical practices. Nevertheless, this limitation was mitigated by validating our pipeline on an independent and more recent dataset.

Second, NLP development and validation were performed using CHUV data only and did not take into consideration interinstitutional variations in coding practices, hospital information system architecture and interoperability, clinical documentation standards, or local prescribing patterns, which may limit the generalizability of our findings. To mitigate this, the model was trained on a balanced, manually annotated corpus reviewed by 3 independent physicians. Future studies should externally validate the NLP model on datasets from other French-speaking institutions to confirm its performance and enhance its applicability.

Third, data from 1 hospital (Baden hospital) were excluded due to missing information and harmonization challenges, and CHOP codes could not be extracted from the USZ hospital; this could have led to underestimation of certain bleeding events. Recent efforts have been undertaken to improve data harmonization across sites, which now largely mitigate the harmonization challenges previously encountered.

Fourth, while *ICD-10-GM* code selection was based on international guidelines and expert review, some misclassification may have occurred. This limitation was partly mitigated by manual validation. However, the adoption of a standardized bleeding classification would help overcome this limitation and harmonize bleeding-event categorization across studies.

Fifth, causality assessment between bleeding cases and antithrombotic agents was not formally assessed by our algorithms, as this requires strict criteria and necessitates a comprehensive EMR review. Causality was manually evaluated using the WHO–Uppsala Monitoring Center framework, which provided valuable contextual insights but is resource-intensive. Future work should investigate semiautomated causal-inference tools to scale this process efficiently.

Sixth, structured data were insufficient to capture the timing of bleeding cases prior to admission, as such information is

XSL•FO
RenderX

documented in discharge summaries, underscoring the need for unstructured data in ADE detection.

Finally, all participating hospitals were tertiary academic centers with strong data infrastructures and comprehensive documentation practices. While this ensured data reliability and methodological consistency, it may limit the generalizability of our findings to other contexts, such as secondary or community hospitals, or to health systems with different digital maturity levels. Compared to many international settings, Swiss university hospitals operate within a decentralized but highly standardized health care system, characterized by universal coverage and well-developed inpatient services. Future research should evaluate these algorithms in more diverse hospital types and countries to assess their adaptability and scalability beyond tertiary Swiss institutions.

Future Directions

This study illustrates the value of combining structured and unstructured clinical data to improve the detection of bleeding events in older inpatients exposed to antithrombotic therapy. This integrated approach can enhance pharmacovigilance systems, reduce underreporting, and support timely clinical interventions. Future efforts should expand algorithm coverage to additional unstructured sources (eg, nursing notes and consultation letters), improve clinical documentation practices, and incorporate semiautomated causality assessment tools. Combining these detection models with multivariate risk stratification that integrates patient-specific factors (age, comorbidities, comedications, and clinical service) could enable prioritization of clinically meaningful alerts. Finally, embedding such tools within common data models and privacy-preserving data-sharing infrastructures, such as those promoted by the Swiss Personalized Health Network, could facilitate cross-institutional learning health systems and accelerate artificial intelligence-supported pharmacovigilance in real-world clinical practice.

Acknowledgments

The authors thank all contributors to data extraction and processing across participating sites, with special appreciation to Walid Gharib-Blanc and Alexandre Wetzel. We gratefully acknowledge Tapio Niemi, Marie Bettex, and Patrick Taffé (Unisanté, Department of Epidemiology and Health Systems, University of Lausanne, Switzerland).

The SwissMADE collaborators are Bernard Burnand, Department of Epidemiology and Health Systems, Unisanté and University of Lausanne, Lausanne, Switzerland; Pierre Olivier Lang, Geriatric Medicine and Geriatric Rehabilitation Division, Department of Medicine, University Hospital of Lausanne, Lausanne, Switzerland; Nicola Colic, Dalle Molle Institute for Artificial Intelligence Research, Università della Svizzera Italiana, Lugano, Switzerland; Angela Schulthess-Lisibach, Institute of Primary Health Care, University of Bern, Bern, Switzerland; Christophe Gaudet-Blavignac, Division of Medical Information Sciences, Geneva University Hospitals, Geneva, Switzerland and Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland; Nathalie Casati, Clinical Data Science Group, Lausanne University Hospital, Lausanne, Switzerland; Jean-Philippe Goldman, Faculty of Letters, Department of Linguistics, University of Geneva, Geneva, Switzerland; and Vasiliki Foufi, Division of Medical Information Sciences, University Hospitals of Geneva, Geneva, Switzerland and Faculty of Medicine, University of Geneva, Geneva, Switzerland.

Funding

This study was partially supported by the Swiss National Research Fund (project 167381). The funder had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Data Availability

Requests for access to the anonymized dataset and study materials should be directed to the corresponding author and will be considered on a case-by-case basis in accordance with institutional policies and applicable data protection regulations.

Authors' Contributions

MALP and C Csajak designed the study; MALP, FG, and C Csajak wrote and submitted the proposal. CL, Christophe Gaudet-Blavignac, Bernard Burnand, FR, ML, Pierre Olivier Lang, and PEB contributed to the development of the study proposal and participated in its revision. CL, PEB, Christophe Gaudet-Blavignac, Nicola Colic, ML, and FR contributed to data extraction and curation. C Coumou, MALP, C Csajak, and FG conducted the study and designed the rule-based algorithms. C Coumou, MALP, C Csajak, FG, and MZ performed the formal analysis. LA contributed to the conceptualization of the rule-based algorithms. MZ and EB designed and implemented the natural language processing algorithms. C Coumou has written the original draft of the manuscript. MALP, C Csajak, FG, and LA contributed to the conceptualization, supervision, and review and editing of the manuscript. The SwissMADE collaborators were consulted on methodological aspects throughout the study. All authors have read and approved the final version of the manuscript.

XSL•FO
RenderX

Conflicts of Interest

PEB reports having received fees and funding from AstraZeneca for matters unrelated to the present study. All other authors declare no competing interests.

Multimedia Appendix 1
International definitions of major bleeding and clinically relevant nonmajor bleeding: a comparative overview.
[DOCX File, 68 KB - jmir_v28i1e77809_app1.docx]

Multimedia Appendix 2
List of synonyms for major bleeding and clinically relevant nonmajor bleeding cases.
[DOCX File, 17 KB - jmir_v28i1e77809_app2.docx]

Multimedia Appendix 3
Interrater reliability assessment among manual reviewers.
[DOCX File, 18 KB - jmir_v28i1e77809_app3.docx]

Multimedia Appendix 4
Causality assessment between antithrombotic treatment and bleeding.
[DOCX File, 16 KB - jmir_v28i1e77809_app4.docx]

Multimedia Appendix 5
Overview of missing values per variable used in major bleeding and clinically relevant nonmajor bleeding algorithms.
[DOCX File, 15 KB - jmir_v28i1e77809_app5.docx]

Multimedia Appendix 6
Comparison of clinical characteristics of Lausanne University Hospital 2015-2016 vs 2021-2022 cohorts.
[DOCX File, 23 KB - jmir_v28i1e77809_app6.docx]

Checklist 1
STROBE checklist.
[DOCX File, 24 KB - jmir_v28i1e77809_app7.docx]

References

1. Jennings ELM, Murphy KD, Gallagher P, O'Mahony D. In-hospital adverse drug reactions in older adults; prevalence, presentation and associated drugs-a systematic review and meta-analysis. Age Ageing 2020 Oct 23;49(6):948-958. [doi: 10.1093/ageing/afaa188] [Medline: 33022061]
2. Long SJ, Brown KF, Ames D, Vincent C. What is known about adverse events in older medical hospital inpatients? A systematic review of the literature. Int J Qual Health Care 2013 Oct;25(5):542-554. [doi: 10.1093/intqhc/mzt056] [Medline: 23925507]
3. Andreotti F, Rocca B, Husted S, et al. Antithrombotic therapy in the elderly: expert position paper of the European Society of Cardiology Working Group on Thrombosis. Eur Heart J 2015 Dec 7;36(46):3238-3249. [doi: 10.1093/eurheartj/ehv304] [Medline: 26163482]
4. Hazell L, Shakir SAW. Under-reporting of adverse drug reactions: a systematic review. Drug Saf 2006;29(5):385-396. [doi: 10.2165/00002018-200629050-00003] [Medline: 16689555]
5. Murff HJ, Patel VL, Hripcsak G, Bates DW. Detecting adverse events for patient safety research: a review of current methodologies. J Biomed Inform 2003;36(1-2):131-143. [doi: 10.1016/j.jbi.2003.08.003] [Medline: 14552854]
6. Dyas AR, Zhuang Y, Meguid RA, et al. Development and validation of a model for surveillance of postoperative bleeding complications using structured electronic health records data. Surgery 2022 Dec;172(6):1728-1732. [doi: 10.1016/j.surg.2022.08.021] [Medline: 36150923]
7. Kashkoush J, Gupta M, Meissner MA, Nielsen ME, Kirchner HL, Garg T. Performance characteristics of a rule-based electronic health record algorithm to identify patients with gross and microscopic hematuria. Methods Inf Med 2023 Dec;62(5-06):183-192. [doi: 10.1055/a-2165-5552] [Medline: 37666279]
8. Hung CY, Lin CH, Chang CS, Li JL, Lee CC. Predicting gastrointestinal bleeding events from multimodal in-hospital electronic health records using deep fusion networks. Annu Int Conf IEEE Eng Med Biol Soc 2019 Jul;2019:2447-2450. [doi: 10.1109/EMBC.2019.8857244] [Medline: 31946393]
9. Salas M, Petracek J, Yalamanchili P, et al. The use of artificial intelligence in pharmacovigilance: a systematic review of the literature. Pharmaceut Med 2022 Oct;36(5):295-306. [doi: 10.1007/s40290-022-00441-z] [Medline: 35904529]

XSL•FO
RenderX

10. Shung D, Tsay C, Laine L, et al. Early identification of patients with acute gastrointestinal bleeding using natural language processing and decision rules. J Gastroenterol Hepatol 2021 Jun;36(6):1590-1597. [doi: 10.1111/jgh.15313] [Medline: 33105045]
11. Li R, Hu B, Liu F, et al. Detection of bleeding events in electronic health record notes using convolutional neural network models enhanced with recurrent neural network autoencoders: deep learning approach. JMIR Med Inform 2019 Feb 8;7(1):e10788. [doi: 10.2196/10788] [Medline: 30735140]
12. Mitra A, Rawat BPS, McManus DD, Yu H. Relation classification for bleeding events from electronic health records using deep learning systems: an empirical study. JMIR Med Inform 2021 Jul 2;9(7):e27527. [doi: 10.2196/27527] [Medline: 34255697]
13. Taggart M, Chapman WW, Steinberg BA, et al. Comparison of 2 natural language processing methods for identification of bleeding among critically Ill patients. JAMA Netw Open 2018 Oct 5;1(6):e183451. [doi: 10.1001/jamanetworkopen.2018.3451] [Medline: 30646240]
14. Walker AL, Watson C, Butcher R, Abedin Z, Yandell M, Shah RU. Use of commercially available natural language processing software to identify bleeding from the medical record. medRxiv. Preprint posted online on Nov 19, 2021. [doi: 10.1101/2021.11.19.21266586]
15. Yap AJY, Teo DCH, Ang PS, et al. Validation of a major and clinically relevant nonmajor bleeding phenotyping algorithm on electronic health records. Pharmacoepidemiol Drug Saf 2024 Aug;33(8):e5875. [doi: 10.1002/pds.5875] [Medline: 39090800]
16. Hung LC, Su YY, Sun JM, Huang WT, Sung SF. Clinical narratives as a predictor for prognosticating functional outcomes after intracerebral hemorrhage. J Neurol Sci 2023 Oct 15;453:120807. [doi: 10.1016/j.jns.2023.120807] [Medline: 37717279]
17. Lee HJ, Jiang M, Wu Y, et al. A comparative study of different methods for automatic identification of clopidogrel-induced bleedings in electronic health records. AMIA Jt Summits Transl Sci Proc 2017;2017:185-192. [Medline: 28815128]
18. Deng B, Zhu W, Sun X, et al. Development and validation of an automatic system for intracerebral hemorrhage medical text recognition and treatment plan output. Front Aging Neurosci 2022;14:798132. [doi: 10.3389/fnagi.2022.798132] [Medline: 35462698]
19. Hansen RS, Lynggaard RB, Laursen MS, Lykke FM, Vinholt PJ. Identification of hematuria with a natural language processing model and validation of hematuria diagnosecodes. Thromb Res 2024 Dec;244:109182. [doi: 10.1016/j.thromres.2024.109182] [Medline: 39426095]
20. Mehran R, Rao SV, Bhatt DL, et al. Standardized bleeding definitions for cardiovascular clinical trials: a consensus report from the Bleeding Academic Research Consortium. Circulation 2011 Jun 14;123(23):2736-2747. [doi: 10.1161/CIRCULATIONAHA.110.009449] [Medline: 21670242]
21. Le CHUV en chiffres [Web Page in French]. Centre Hospitalier Universitaire Vaudois. 2023. URL: https://www.chuv.ch/fr/a-propos/le-chuv-en-chiffres [accessed 2026-01-10]
22. Facts, figures and dates. Hôpitaux Universitaires De Genève. URL: https://www.hug.ch/en/facts-figures-and-dates [accessed 2026-01-10]
23. About the USZ 2025. Universitäts Spital Zürich. URL: https://www.usz.ch/en/about-university-hospital-zurich [accessed 2025-10-07]
24. KSB weiterhin auf wachstumskurs [Web Page in German]. Kantonsspital Baden. 2024. URL: https://www.kantonsspitalbaden.ch/news/ksb-weiterhin-auf-wachstumskurs [accessed 2026-01-10]
25. Gaspar F, Lutters M, Beeler PE, et al. Automatic detection of adverse drug events in geriatric care: study proposal. JMIR Res Protoc 2022 Nov 15;11(11):e40456. [doi: 10.2196/40456] [Medline: 36378522]
26. Schulman S, Kearon C, Subcommittee on Control of Anticoagulation of the Scientific and Standardization Committee of the International Society on Thrombosis and Haemostasis. Definition of major bleeding in clinical investigations of antihemostatic medicinal products in non-surgical patients. J Thromb Haemost 2005 Apr;3(4):692-694. [doi: 10.1111/j.1538-7836.2005.01204.x] [Medline: 15842354]
27. Kaatz S, Ahmad D, Spyropoulos AC, Schulman S, Subcommittee on Control of Anticoagulation. Definition of clinically relevant non-major bleeding in studies of anticoagulants in atrial fibrillation and venous thromboembolic disease in non-surgical patients: communication from the SSC of the ISTH. J Thromb Haemost 2015 Nov;13(11):2119-2126. [doi: 10.1111/jth.13140] [Medline: 26764429]
28. Honnibal M, Montani I. spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Sentometrics Res 2017;7(1):411-420 [FREE Full text]
29. Gaspar F, Zayene M, Coumau C, et al. Natural language processing and ICD-10 coding for detecting bleeding events in discharge summaries: comparative cross-sectional study. JMIR Med Inform 2025 Aug 29;13:e67837. [doi: 10.2196/67837] [Medline: 40882207]
30. Taffé P, Halfon P, Ghali WA, Burnand B, International Methodology Consortium for Coded Health Information (IMECCHI). Test result-based sampling: an efficient design for estimating the accuracy of patient safety indicators. Med Decis Making 2012;32(1):E1-12. [doi: 10.1177/0272989X11426176] [Medline: 22065144]
31. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]

XSL•FO
RenderX

32. The use of the WHO-UMC system for standardised case causality assessment. : World Health Organization - Uppsala Monitoring Centre; 2018 URL: https://www.who.int/docs/default-source/medicines/pharmacovigilance/whocausality-assessment.pdf [accessed 2026-01-10]
33. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987;40(5):373-383. [doi: 10.1016/0021-9681(87)90171-8] [Medline: 3558716]
34. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care 1998 Jan;36(1):8-27. [doi: 10.1097/00005650-199801000-00004] [Medline: 9431328]
35. Ferrazzini E, Méan M, Stalder O, Limacher A, Rodondi N, Aujesky D. Incidence and clinical impact of bleeding events in older patients with acute venous thromboembolism. Blood Adv 2023 Jan 24;7(2):205-213. [doi: 10.1182/bloodadvances.2022007263] [Medline: 35381071]
36. Gómez-Cuervo C, Rivas A, Visonà A, et al. Predicting the risk for major bleeding in elderly patients with venous thromboembolism using the Charlson index. Findings from the RIETE. J Thromb Thrombolysis 2021 May;51(4):1017-1025. [doi: 10.1007/s11239-020-02274-6] [Medline: 32945982]
37. Poli D, Antonucci E, Bertù L, et al. Very elderly patients with venous thromboembolism on oral anticoagulation with VKAs or DOACs: results from the prospective multicenter START2-Register Study. Thromb Res 2019 Nov;183:28-32. [doi: 10.1016/j.thromres.2019.08.024] [Medline: 31536872]
38. Prasad N, Lau ECY, Wojt I, Penm J, Dai Z, Tan ECK. Prevalence of and risk factors for drug-related readmissions in older adults: a systematic review and meta-analysis. Drugs Aging 2024 Jan;41(1):1-11. [doi: 10.1007/s40266-023-01076-8] [Medline: 37864770]
39. Shehab N, Ziemba R, Campbell KN, et al. Assessment of ICD-10-CM code assignment validity for case finding of outpatient anticoagulant-related bleeding among Medicare beneficiaries. Pharmacoepidemiol Drug Saf 2019 Jul;28(7):951-964. [doi: 10.1002/pds.4783] [Medline: 31144403]
40. Joos C, Lawrence K, Jones AE, Johnson SA, Witt DM. Accuracy of ICD-10 codes for identifying hospitalizations for acute anticoagulation therapy-related bleeding events. Thromb Res 2019 Sep;181:71-76. [doi: 10.1016/j.thromres.2019.07.021] [Medline: 31357146]
41. Mitra A, Rawat BPS, McManus D, Kapoor A, Yu H. Bleeding entity recognition in electronic health records: a comprehensive analysis of end-to-end systems. AMIA Annu Symp Proc 2020;2020:860-869. [Medline: 33936461]
42. Xu W, Lv M, Wu S, et al. Severe bleeding risk of direct oral anticoagulants versus vitamin K antagonists for stroke prevention and treatment in patients with atrial fibrillation: a systematic review and network meta-analysis. Cardiovasc Drugs Ther 2023 Apr;37(2):363-377. [doi: 10.1007/s10557-021-07232-9] [Medline: 34436708]
43. Mant J. Process versus outcome indicators in the assessment of quality of health care. Int J Qual Health Care 2001 Dec;13(6):475-480. [doi: 10.1093/intqhc/13.6.475] [Medline: 11769750]

Abbreviations

ADE: adverse drug event
CHOP: Swiss Classification of Surgical Procedures
CHUV: Lausanne University Hospital
CRNMB: clinically relevant nonmajor bleeding
EMR: electronic medical record
HUG: Geneva University Hospital
***ICD-10-GM*:** *International Statistical Classification of Diseases, 10th Revision, German Modification*
ISTH: International Society on Thrombosis and Haemostasis
KSB: Baden Cantonal Hospital
MB: major bleeding
ML: machine learning
NLP: natural language processing
PPV: positive predictive value
SDA: structured data algorithms
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
USZ: Zürich University Hospital

XSL•FO
RenderX

Edited by A Schwartz; submitted 20.May.2025; peer-reviewed by M Forgerini, S Sazzad, U Modebelu; revised version received 18.Nov.2025; accepted 27.Nov.2025; published 29.Jan.2026.

Please cite as:
Coumau C, Gaspar F, Zayene M, Bertrand E, Alberio L, Lovis C, Beeler PE, Rinaldi F, Lutters M, Le Pogam MA, Csajka C, SwissMADE Collaborators
Detection of Antithrombotic-Related Bleeding in Older Inpatients: Multicenter Retrospective Study Using Structured and Unstructured Electronic Health Record Data
J Med Internet Res 2026;28:e77809
URL: https://www.jmir.org/2026/1/e77809
doi:10.2196/77809

© Claire Coumau, Frederic Gaspar, Mehdi Zayene, Elliott Bertrand, Lorenzo Alberio, Christian Lovis, Patrick E Beeler, Fabio Rinaldi, Monika Lutters, Marie-Annick Le Pogam, Chantal Csajka, SwissMADE Collaborators. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Behavioral Dynamics of AI Trust and Health Care Delays Among Adults: Integrated Cross-Sectional Survey and Agent-Based Modeling Study

Xueyao Cai[1*], MD; Weidong Li[1*], MD; Wenjun Shi[2], MD; Yuchen Cai[1*], MD; Jianda Zhou[1*], MD

[1]Department of Plastic Surgery, The Third Xiangya Hospital, Central South University, Changsha, China

[2]Department of Plastic and Reconstructive Surgery, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Jianda Zhou, MD
Department of Plastic Surgery
The Third Xiangya Hospital
Central South University
138 Tongzipo Road
Changsha
China
Phone: 86 13508493668
Email: zhoujianda@csu.edu.cn

Abstract

Background: While artificial intelligence (AI) holds significant promise for health care, excessive trust in these tools may unintentionally delay patients from seeking professional care, particularly among patients with chronic illnesses. However, the behavioral dynamics underlying this phenomenon remain poorly understood.

Objective: This study aims to quantify the influence of AI trust on health care delays through integrated survey-based mediation analysis and real-world research, and to simulate intervention efficacy using agent-based modeling (ABM).

Methods: A cross-sectional online survey was conducted in China from December 2024 to May 2025. Participants were recruited via convenience sampling on social media (WeChat and QQ) and hospital portals. The survey included a 21-item questionnaire measuring AI trust (5-point Likert scale), AI usage frequency (6-point scale), chronic disease status (physician-diagnosed, binary), and self-reported health care delay (binary). Responses with completion time <90 seconds, logical inconsistencies, missing values, or duplicates were excluded. Analyses included descriptive statistics, multivariable logistic regression (α=.05), mediation analysis with nonparametric bootstrapping (500 iterations), and moderation testing. Subsequently, an ABM simulated 2460 agents within a small-world network over 14 days to model behavioral feedback and test 3 interventions: broadcast messaging, behavioral reward, and network rewiring.

Results: The final sample included 2460 adults (mean age 34.46, SD 11.62 years; n=1345, 54.7% female). Higher AI trust was associated with increased odds of delays (odds ratio [OR] 1.09, 95% CI 1.00-1.18; *P*=.04), with usage frequency partially mediating this relationship (indirect OR 1.24, 95% CI 1.20-1.29; *P*<.001). Chronic disease status amplified the delay odds (OR 1.42, 95% CI 1.09-1.86; *P*=.01). The ABM demonstrated a bidirectional trust erosion loop, with population delay rates declining from 10.6% to 9.5% as mean AI trust decreased from 1.91 to 1.52. Interventions simulation found broadcast messaging most effective in reducing delay odds (OR 0.94, 95% CI 0.94-0.95; *P*<.001), whereas network rewiring increased odds (OR 1.04, 95% CI 1.04-1.05; *P*<.001), suggesting a "trust polarization" effect.

Conclusions: This study reveals a nuanced relationship between AI trust and delayed health care–seeking. While trust in AI enhances engagement, it can also lead to delayed care, particularly among patients with chronic conditions or frequent AI users. Integrating survey data with ABM highlights how AI trust and delay behaviors can strengthen one another over time. Our findings indicate that AI health tools should prioritize calibrated decision support rather than full automation to balance autonomy, odds, and decision quality in digital health. Unlike previous studies that focus solely on static associations, this research emphasizes the dynamic interactions between AI trust and delay behaviors.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e82170) doi:10.2196/82170

KEYWORDS

agent-based modeling; artificial intelligence; chronic disease; health care delay; real-world research

Introduction

Background

With the rapid development and widespread adoption of artificial intelligence (AI) in the medical field, AI has emerged as a pivotal tool across multiple domains, including health management, disease prediction, and clinical decision support [1,2]. Its potential to enhance health care efficiency, assist in disease diagnosis, optimize treatment strategies, and facilitate personalized health management is substantial [3,4]. However, this promise is tempered by growing concerns about its psychosocial and behavioral effects. The prevailing view of AI as an unequivocal force for good is being challenged by evidence that human-AI interaction can lead to new cognitive and behavioral issues [5,6]. In particular, AI's capacity to provide health recommendations and interventions based on extensive data analysis marks a significant breakthrough [7,8]. However, this ability may alter traditional health decision-making processes. Trust is a key concept for adopting technology, involving aspects such as performance, process, and purpose. A growing body of research indicates that while AI offers convenient access to health-related information, excessive reliance on these tools may lead to detrimental health behaviors, such as overreliance on automated suggestions and disregarding clinical intuition [9-11].

Traditional health decision-making models typically depend on professional medical judgment and individuals' self-awareness of their health status [12]. The advent of AI is reshaping this framework by providing automated diagnostic recommendations and personalized health guidance [13]. While such "digital health advice" can enhance confidence in medical decision-making in the short term, it may foster excessive dependence on AI tools over the long term. Patients with chronic diseases, who require ongoing health management and exhibit relatively stable symptoms, may be particularly susceptible to this shift [14]. This population often practices prolonged self-management, leading to frequent use of digital tools [15]. The relative stability of their conditions can create a sense of security, in which they may consider AI-generated reassuring feedback adequate. This dependence on AI could lead to delays in pursuing health care, as patients with chronic disease tend to give priority to AI-generated symptom evaluations and treatment options rather than traditional medical advice [16]. Consequently, they might overlook complex diagnostic and therapeutic needs that AI cannot fully assess [17]. Moreover, biases in medical AI can persist throughout its life cycle, potentially leading to serious repercussions in clinical decision-making. If these biases are not addressed, they can result in inaccurate medical judgments and exacerbate existing health care disparities [18]. While existing research has shed light on the potential risks of AI reliance in health care, a comprehensive framework that formally analyzes the complex relationship between AI trust and health behavior, particularly in the context of chronic disease, is still lacking. Additionally, few studies have used computational simulation to assess how various public health intervention strategies might mitigate this system-level risk.

This study aims to systematically analyze the interrelationships among AI trust, AI usage frequency, chronic disease status, and delayed health care–seeking behavior. The analysis was conducted through a cross-sectional online survey from December 2024 to May 2025. We propose that AI trust and frequency of use jointly contribute to the odds of delayed medical care through a behavioral feedback mechanism, particularly pronounced among patients with chronic conditions. High levels of AI trust may inadvertently lead individuals to postpone health care–seeking by increasing usage frequency, highlighting a complex interaction between trust and behavior. To investigate these dynamic interdependencies, we use an agent-based modeling (ABM) framework to simulate the trust-behavior feedback over time. By embedding individuals within social networks and allowing trust and delay behaviors to evolve together, the ABM enables us to examine how microlevel decisions accumulate into population-level odds. This includes issues such as collective delays or systemic trust collapse, which are often difficult to capture using conventional cross-sectional data. A schematic overview of the study design is provided in Figure 1.

XSL•FO
RenderX

Figure 1. Schematic overview of the study design and methodology. Based on online questionnaires, the analysis framework includes mediation analysis to explore the relationship between artificial intelligence (AI) trust, usage frequency, and delayed health care–seeking behavior. Additionally, agent-based modeling was used to model the dynamic feedback loops over a 14-day period, incorporating 3 intervention strategies: broadcast, reward, and rewire. These strategies aimed to evaluate how the interventions affected trust and delay behavior within the population. Image created in Biorender by WL [19].

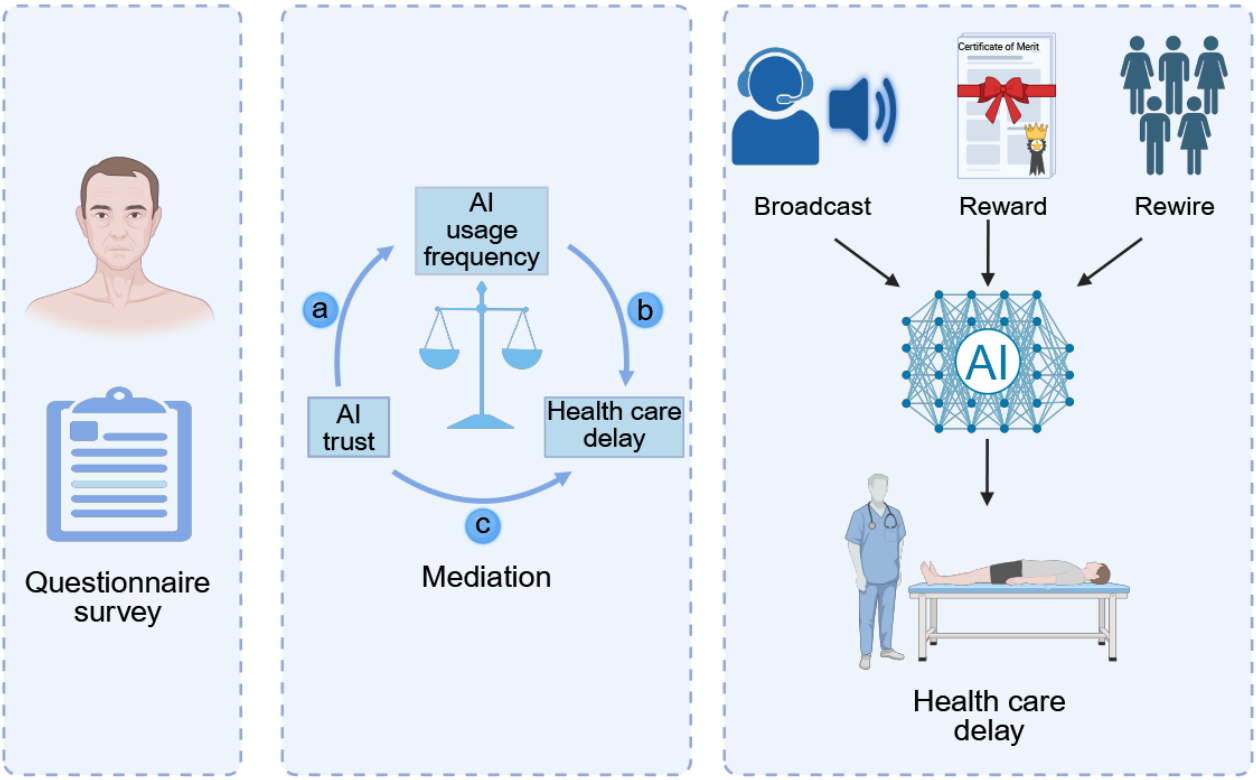

Objective

This study not only addresses a theoretical gap in understanding the impact of AI on delayed health care–seeking but also offers a fresh perspective on optimizing the design of AI-based health tools. A key challenge in future medical applications of AI will be achieving the right balance between fostering user trust and ensuring accuracy and timeliness in health care–seeking. We aim to provide theoretical support for designing AI-driven health interventions. Additionally, we seek to inform more effective public health strategies, ensuring that technological advancements promote rather than hinder appropriate health care decisions.

Methods

Survey Design and Data Collection

This study was conducted in 2 integrated phases: a cross-sectional survey and an ABM simulation. To ensure comprehensive and transparent reporting, the methods are reported following the Journal Article Reporting Standards guidelines under relevant subheadings [20,21]. A cross-sectional online survey was conducted between December 1, 2024, and May 20, 2025. The 21-item questionnaire was conducted in Mandarin Chinese and developed based on a systematic literature review and focus-group discussions. It underwent refinement through 2 rounds of expert review by 4 specialists in public health and AI, and was finalized after a pilot test involving 5 target participants. The full survey questionnaire (English translation) is provided in Multimedia Appendix 1.

Eligible participants were adults aged 18-75 years who had resided in mainland China for the past 12 months and were able to read Chinese. The age range was selected to include the digitally engaged adult population while excluding minors and the very old people, who may exhibit different health-seeking patterns. The 12-month residency criterion ensured consistent exposure to the local health care, minimizing confounding from recent immigration or transient populations. The survey was disseminated via WeChat (Tencent) and QQ (Tencent), and the patient-education portals of collaborating hospitals using convenience sampling. Participants accessed the questionnaire after reading an electronic informed consent form and selecting “Agree.” All items were mandatory, and skip logic was applied to minimize irrelevant questions. Responses were transmitted over HTTPS and encrypted at rest on the institution’s private server, accessible only to authorized investigators.

To ensure the validity and interpretability of the logistic regression and mediation models, which require complete-case data, we performed rigorous data cleaning. Questionnaires completed in under 90 seconds, or those containing logical inconsistencies, missing values, or duplicate entries (only the first complete entry was retained), were excluded. Given the limited amount of missing data and the potential risk of introducing bias through imputation, listwise deletion was considered the most appropriate and conservative approach. To examine the mechanism of missingness, we conducted Little’s

XSL•FO
RenderX

missing completely at random (MCAR) test using the *mice* package in R software (versions 4.2.3; R Foundation for Statistical Computing). Descriptive statistics were used to summarize the data (frequencies, percentages, means, and SDs), and multivariable logistic regression was conducted to examine the association between AI usage frequency and delayed health care–seeking behavior. Likert-type scales ranging from 1 to 5 (or 0 to 5 for AI usage and trust items) were used to assess frequency, trust level, and willingness to recommend.

Chronic disease was defined as any physician-diagnosed condition lasting 6 months or longer, including chronic allergic rhinitis, asthma, hypertension, diabetes, chronic skin diseases, etc. The variables included demographic characteristics (eg, age, sex, and occupation), AI usage behaviors (eg, frequency and exposure), trust perceptions, and health care–seeking outcomes.

Self-reported health care delay was assessed using a single binary item specifically targeting AI-influenced delay behaviors. Participants were asked: "Has advice provided by ChatGPT ever caused you to postpone or cancel seeking medical care?" with response options 0=no and 1=yes. This measure captures intentional delay or avoidance of health care that participants attributed directly to ChatGPT's recommendations, rather than delays caused by logistical or accessibility barriers. This operationalization aligns with behavioral delay frameworks used in prior studies and reflects clinically meaningful patterns of health care–seeking behavior.

Ethical Considerations

The study protocol was approved by the Institutional Review Board of Xiangya Third Hospital, Central South University (approval number 2025255). All procedures complied with the Declaration of Helsinki and the Personal Information Protection Law of China. Electronic informed consent was obtained from all participants. After reading a digital information sheet that outlined the study's purpose, procedures, risks, benefits, and their rights (including voluntary participation and withdrawal), participants indicated their agreement by selecting "Agree" before proceeding to the questionnaire. All collected data were deidentified. No personally identifiable information (eg, name, ID number, and contact details) was stored with the response data. Data were transmitted securely via HTTPS and stored in encrypted form on a private, access-controlled institutional server. Results are reported in aggregate to prevent any possibility of individual identification. Participants did not receive any financial or material compensation for their involvement in this study. The manuscript and its supplementary materials do not contain any images, videos, or textual data that could lead to the identification of an individual participant. Therefore, specific consent for the publication of identifiable information was not applicable.

Key Predictors of Delayed Care: Logistic Regression Analysis

We used logistic regression models to evaluate key predictors of delayed health care–seeking behavior. Initially, univariate logistic analyses were conducted to estimate the association between each candidate variable and the outcome, with results reported as odds ratios (ORs) accompanied by 95% CIs. Variables demonstrating potential significance were subsequently included in a series of hierarchical multivariate models (models 1-4). These models progressively incorporated individual characteristics, intervention exposure, AI usage patterns, and social influence to assess their independent contributions to delay behavior. This approach allowed us to identify the relative importance of each factor in influencing health care–seeking delays.

The data collection period encompassed the public release of DeepSeek, a major large language model in China, which occurred in early 2025 [22]. This event represented a significant shock to public awareness toward AI. To test the robustness of our core findings against this potential confounding effect, we performed a stratified analysis by dividing the sample into pre- and post-DeepSeek release subgroups. The cutoff date was set to February 1, 2025, allowing a sufficient time frame for the model's public impact to materialize within our survey window. We then reran the univariate logistic analyses within each subgroup to assess the consistency of the associations between AI trust, usage frequency, and health care delays.

Mediation Analysis: Indirect Effect of AI Trust via Usage Frequency

To examine whether AI trust influences delay behavior indirectly through AI usage frequency, we conducted mediation analysis using 2 approaches. First, following the traditional Baron and Kenny framework with the Sobel test, we estimated path *a* (the association between AI trust and usage frequency) via linear regression, and paths *b* and *c'* (the associations of usage frequency and AI trust with delay behavior) via multivariable logistic regression. The significance of the indirect effect ($a \times b$) was assessed using the Sobel *z* test. Second, to obtain robust CIs and significance estimates, we performed nonparametric bootstrap resampling (n=500). In each iteration, we reestimated paths *a* and *b*, calculated the product $a \times b$, and derived the empirical distribution of the indirect effect. The 95% bootstrap CI and *P* value were computed accordingly. All models were adjusted for age, gender, and chronic disease status.

Moderation by AI Recommendation Exposure: Stratified and Interaction Models

To further examine whether the level of AI recommendation exposure moderates the relationship between AI trust and delayed health care–seeking behavior, we conducted stratified logistic regression and interaction modeling. Participants were divided into low and high exposure groups based on their reported frequency of receiving AI recommendations (≤2 vs >2). Separate multivariable logistic regression models were fitted for each group, adjusting for age, gender, chronic disease status, and AI usage frequency. ORs and 95% CIs were reported to assess effect heterogeneity across exposure levels.

Additionally, we introduced an interaction term between AI trust and recommendation exposure into the full model. The interaction coefficient was used to calculate the interaction OR and the corresponding 95% CI. Predicted probability curves were plotted to visualize how the relationship between AI trust and delay behavior varies by exposure group, providing insights

XSL•FO
RenderX

into the moderating effect of AI recommendation exposure on health care–seeking behavior.

Moderation by Recommendation Willingness: Stratified Trust-Delay Associations

To evaluate whether individuals' willingness to recommend AI tools moderates the relationship between AI trust and delayed medical care, we conducted stratified logistic regressions and interaction analysis. Participants were grouped based on their scores (1-5) regarding the likelihood of recommending AI: those scoring ≤2 were classified as the "low recommendation intensity" group, and those scoring ≥3 as the "high recommendation intensity" group. Multivariable logistic regression models were fitted within each stratum, adjusting for age, gender, chronic disease status, and AI usage frequency. ORs and 95% CIs were calculated for AI trust in each group.

Subsequently, a full model including an interaction term between AI trust and recommendation intensity was constructed. We estimated the interaction effect (OR, 95% CI, and *P* value) and generated a prediction grid with standardized covariates to visualize the predicted probability of delayed care across AI trust levels in each group. Interaction plots were created to illustrate potential effect modification, enhancing our understanding of how willingness to recommend AI tools influences the relationship between AI trust and health care–seeking behavior.

Scenario-Based Modeling: Joint Effects of AI Behavior Factors on Delay

To theoretically evaluate the potential impact of various intervention strategies on delayed health care–seeking behavior, we conducted a scenario-based simulation analysis using logistic regression. A multivariable logistic model was initially constructed with AI trust, frequency of use, and chronic disease status as predictors. ORs, 95% CIs, and *P* values were reported for each explanatory variable.

Based on this model, the following six intervention scenarios were simulated: (1) baseline (trust=3, frequency=3, no chronic disease); (2) increased AI trust (set to 5); (3) increased frequency of AI use (set to 5); (4) chronic disease status switched to "yes"; (5) combined trust + frequency increase; and (6) combined trust + frequency + chronic disease. For each scenario, predicted probabilities of delay were computed across all individuals and averaged to reflect group-level effects.

The results were visualized using a bar plot displaying the mean predicted probability of delay under each strategy. Annotations highlighted the ORs and significance levels of the 3 key predictors, facilitating an intuitive understanding of their relative contributions to delay behavior. This comprehensive approach allowed us to identify the most effective interventions for reducing delays in health care–seeking.

ABM: Broadcast, Reward, Rewire

To evaluate the impact of different intervention mechanisms on delayed health care–seeking behavior, we used an ABM framework. ABM is a computational simulation approach particularly suited for studying complex systems, where population-level outcomes emerge from the interactions of autonomous, diverse individuals ("agents") operating within a defined environment and set of rules [23-25]. This method is particularly appropriate for our research question for 3 main reasons. First, AI trust is not a static trait; it is a dynamic belief influenced by personal experience and social factors, which ABM is designed to capture. Second, the decision to delay care involves weighing personal trust against the behaviors of peers, a scenario well modeled by embedding agents within a social network where attitudes and actions spread [26,27]. Third, ABM allows us to test the "trust-delay" feedback loop, a causal chain that cannot be directly identified from our cross-sectional survey data but can be explored through simulation [28].

To improve the transparency and reproducibility, the ABM model is described following the overview, design concepts, and details (ODD) protocol [29]. The purpose of the simulation was to examine how AI trust, peer influence, and intervention strategies jointly shape delayed health care–seeking behavior over time. Each agent represented 1 survey respondent and was initialized using the individual's empirical AI trust (1-5), AI usage frequency (1-5), and chronic disease status. Agents were embedded in a Watts-Strogatz small-world social network (n=2460; average degree=4; rewiring probability=0.2), which captures realistic social clustering and intermittent long-distance ties. The simulation progressed in daily cycles for 14 days, during which agents first computed a probability of delay using a logistic model derived from the survey data, then made a probabilistic delay decision, and subsequently updated their trust and usage behaviors according to personal outcomes, peer context, and intervention conditions. A total of 100 repetitions were performed for each scenario.

The model incorporated key design concepts of agent-based systems. Interaction occurred exclusively through local network neighbors; both average neighbor trust and neighbor delay behaviors influenced an individual's own updates. Stochasticity was present in network initialization, delay decisions, and trial-level replications. Agents adapted their trust and usage frequency over time, increasing them when surrounded by high-trust peers or after not delaying care, and decreasing them when neighbors frequently delayed or after personally delaying care. No explicit learning mechanism was included; behavioral dynamics emerged entirely from rule-based adaptation. Model outputs consisted of daily population-level delay rates and comparative effects of intervention strategies relative to baseline.

Detailed implementation followed ODD guidelines [29]. At initialization, agents with an empirically predicted baseline delay probability above 0.20 were assigned a 1-point reduction in trust and usage frequency to represent structural vulnerability. During each daily update, delaying care reduced trust by 0.2 and usage frequency by 0.3, whereas not delaying increased trust by 0.1; all values were bounded between 1 and 5. A total of 3 intervention strategies were embedded into this baseline framework. In the broadcast condition, trust was reduced daily by a small fixed penalty to reflect exposure to trust-eroding messages. The reward condition increased trust and usage frequency for agents who consistently sought timely care, with rewards provided every 2 days. In the rewiring condition, network edges were periodically redirected toward the highest-trust agents to model opinion leader amplification. All

XSL•FO
RenderX

intervention rules were implemented on top of the core behavioral update mechanism. Finally, 1-way sensitivity analyses were conducted by varying initial trust levels (means of 2.5, 3.0, and 3.5), broadcast penalty intensities (0.05-0.20 per day), reward magnitudes (0.03-0.10), and rewiring frequencies (intervals of 2-10 days). Each parameter set was simulated in 100 trials, and the resulting delay trajectories were compared to assess model robustness.

Results

Demographic Characteristics, AI Use, and Health Decision Outcomes

Of 2785 initial submissions, 325 (11.7%) responses were excluded based on prespecified criteria (completion time <90 seconds, logical inconsistencies, duplicate entries, or missing values in key variables). Specifically, 136 (4.9%) exclusions from the initial sample were due to missing values. Little's MCAR test indicated that the data were MCAR (χ^2_{14}=12.87; *P*=.54), supporting the use of complete-case analysis. Consequently, a total of 2460 valid responses were retained, predominantly female (n=1345, 54.7%), with an average age of 34.46 (SD 11.62) years. Occupations included students (n=825, 33.5%), technology workers (n=715, 29.1%), other professions (n=640, 26%), and health care personnel (n=280, 11.4%). A significant majority (n=2215, 90%) reported being aware of generative AI tools, and 62% (n=1525) had previously used AI for health advice. Regarding the frequency of AI-based advice use, 38% (n=935) of participants never used it, while 16.2% (n=398) used it weekly (1-2 times) and 15.7% (n=385) monthly. Trust in AI-generated advice varied, with 38% (n=935) never using it, and the remaining participants distributed across differing trust levels (Table 1).

In terms of health care decision outcomes, 11.6% (285/2460) of participants deferred or canceled health care due to AI advice, while 18.9% (465/2460) changed their health care decisions based on it. Additionally, 16.9% (415/2460) adopted alternative therapies following AI recommendations. The primary source of health information was physicians (1203/2460, 48.9%), followed by search engines (465/2460, 18.9%) and generative AI tools (375/2460, 15.2%). Perceived influence of online discussions varied, with 19.7% (485/2460) considering it very likely to impact their decisions. Lastly, 30.1% (740/2460) of participants reported having a chronic disease (Table 2).

XSL•FO
RenderX

Table 1. Demographic characteristics and artificial intelligence (AI) use (n=2460).

Variable	Values
Sex, n (%)	
Female	1345 (54.7)
Male	1115 (45.3)
Age (years), mean (SD)	34.46 (1.62)
Occupation category, n (%)	
Students	825 (33.5)
Technology workers	715 (29.1)
Other	640 (26)
Health care personnel	280 (11.4)
Awareness of generative-AI tools, n (%)	
Yes	2215 (90)
No	245 (10)
Previous use of AI for health advice, n (%)	
Yes	1525 (62)
No	935 (38)
Frequency of AI-based advice use, n (%)	
Never	935 (38)
Weekly 1-2 times	398 (16.2)
Monthly	385 (15.7)
Weekly 3-4 times	308 (12.5)
Occasional	293 (11.9)
Almost daily	141 (5.7)
Trust in AI-generated advice, n (%)	
1 (“none”)	320 (13)
2	290 (11.8)
3	280 (11.4)
4	345 (14)
5 (“high”)	290 (11.8)
0 (“never used”)	935 (38)

XSL•FO
RenderX

Table 2. Health decision outcomes and contextual factors (n=2460).

Variable	Value, n (%)
Deferred or cancelled health care due to AI[a] advice	
Yes	285 (11.6)
No	2175 (88.4)
Changed health care decision due to AI advice	
Yes	465 (18.9)
No	1995 (81.1)
Adopted alternative therapy due to AI advice	
Yes	415 (16.9)
No	2045 (83.1)
Primary source of health information	
Physician	1203 (48.9)
Search engines	465 (18.9)
Generative AI tools	375 (15.2)
Family or friends	255 (10.3)
Government or traditional media	85 (3.5)
Social networking	77 (3.1)
Perceived influence of online discussions	
Very likely	485 (19.7)
Likely	835 (33.9)
Uncertain	775 (31.5)
Unlikely	265 (10.8)
Very unlikely	100 (4.1)
Presence of chronic disease	
Yes	740 (30.1)
No	1720 (69.9)

[a]AI: artificial intelligence.

AI Trust and Usage Frequency Are Key Predictors of Delay

In the univariate logistic regression, higher AI trust was positively associated with delayed health care–seeking behavior, with an OR of 1.27 (95% CI 1.19-1.36; *P*<.001; Figure 2A). Similarly, the frequency of AI use also showed a significant positive correlation with delay (OR 1.41, 95% CI 1.18-1.69; *P*<.001; Figure 2A). Although not statistically significant, trends were observed in the effects of receiving AI recommendations (OR 0.95, 95% CI 0.81-1.11) and actively recommending AI (OR 0.92, 95% CI 0.781.09; Figure 2A), both suggesting a potential reduction in delay. In the fully adjusted model (model 4), AI trust remained a significant predictor (OR 1.09, 95% CI 1.00-1.18; *P*=.04; Figure 2B). Frequency of AI use exhibited the strongest association with delay (OR 1.39, 95% CI 1.14-1.68; *P*<.001; Figure 2B), indicating that more frequent users had higher odds of postponing medical visits. Additionally, chronic disease status demonstrated a persistent positive association with delay (OR 1.43, 95% CI 1.10-1.88; *P*=.009; Figure 2B). Detailed analyses for univariate and multivariate logistic regression can be found in Tables S1 and S2 in Multimedia Appendix 2.

To address potential confounding from a major AI market event, we stratified the analysis by the DeepSeek release period. The results confirmed the robustness of our primary findings. The association between AI usage frequency and delayed health care–seeking remained nearly identical in both direction and magnitude in the pre- and postrelease periods (prerelease: OR 1.41, 95% CI 1.17-1.70; *P*<.001; postrelease: OR 1.41, 95% CI 1.27-1.56; *P*<.001). For AI trust, while the positive association was slightly attenuated and not statistically significant in the prerelease period (OR 1.05, 95% CI 0.89-1.25; *P*=.56), it was significant in the postrelease period (OR 1.10, 95% CI 1.00-1.21; *P*=.046). This suggests that the fundamental behavioral mechanism linking AI trust and usage to delay is robust. The DeepSeek release may have slightly amplified the measurable effect of AI trust, possibly due to increased public reliance on AI tools. Detailed results of this stratified analysis are available in Table S3 in Multimedia Appendix 2.

XSL•FO
RenderX

Figure 2. Association between artificial intelligence (AI) trust, usage frequency, and delayed health care–seeking behavior. (A) Univariate logistic regression and (B) fully adjusted multivariable model evaluating key predictors of delayed health care–seeking behavior.

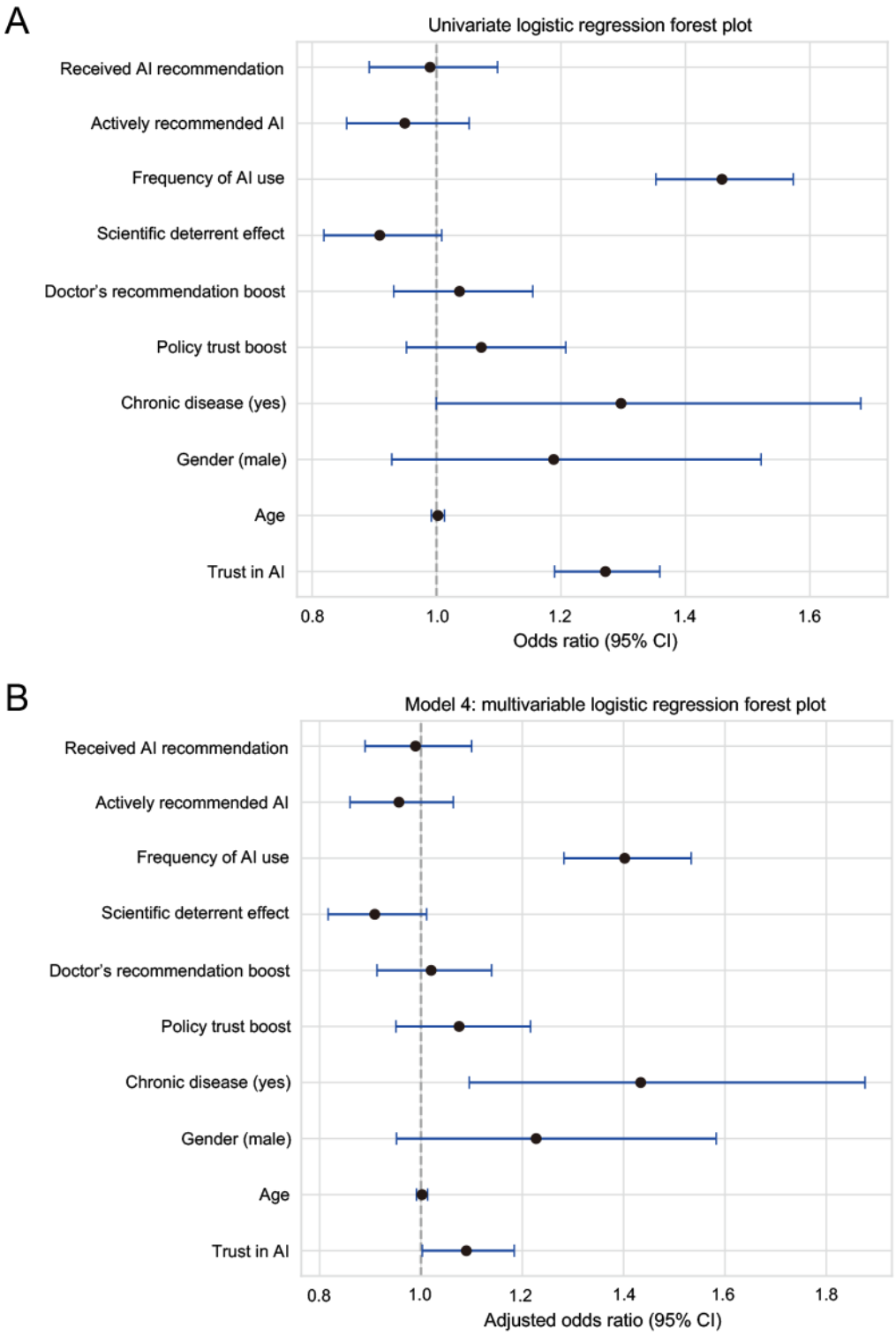

AI Trust Influences Delay via Frequency: Evidence of Mediation

We further evaluated the indirect effect of AI trust on delayed health care–seeking through AI usage frequency, supporting a partial mediation model. In the linear regression analysis, AI trust significantly predicted usage frequency (path *a*: β=.5754; *P*<.001). In the multivariable logistic model, usage frequency was also significantly associated with delay (path *b*: OR 1.40, 95% CI 1.28-1.55; *P*<.001). The product term (indirect effect $a \times b$) was calculated to be 0.1949, and the Sobel test yielded a significant result (*P*<.001), indicating a statistically robust mediation path (Table 3). To further validate these findings, a nonparametric bootstrap analysis with 500 replications was conducted. The mean indirect effect was found to be 0.2152 (OR 1.24, 95% CI 1.20-1.29), with a bootstrap *P* value of <.001.

XSL•FO
RenderX

This confirms that AI trust contributes to delay behavior partly through increased frequency of AI usage. The direct effect of AI trust on delay remained statistically significant after adjusting for the mediator (path *c'*: OR 1.09, 95% CI 1.00-1.18; *P*=.04), supporting the conclusion of a partial mediation model (Table 3). A diagram illustrating this mediation relationship can be found in Figure 3.

Table 3. Mediation effect of artificial intelligence (AI) trust on health care delay behavior.

Path	β (log OR[a])	OR (95% CI)	SE	*P* value	*P* value (Sobel)
a (AI trust-AI frequency)	.5754	N/A[b]	0.014	<.001	N/A
b (AI frequency-delay)	.3387	1.4031 (1.28-1.53)	0.0455	<.001	N/A
a × *b* (indirect)	.1949	N/A	0.0266	N/A	<.001
c (AI trust-delay, total effect)	.2473	1.2806 (1.20-1.37)	N/A	N/A	N/A
c' (AI trust-delay, direct effect)	.085	1.0887 (1.0-1.18)	0.0423	.04	N/A

[a]OR: odds ratio.

[b]N/A: not applicable.

Figure 3. Mediation model demonstrating how artificial intelligence (AI) trust influences delayed health care–seeking through increased frequency of AI usage. Path *a* shows the significant effect of AI trust on usage frequency, while path *b* indicates the association between usage frequency and delay. The indirect effect (*a* × *b*) and direct effect (*c'*) are also represented, highlighting the statistical significance of the mediation pathway. Image created in Biorender by WL [30]. OR: odds ratio.

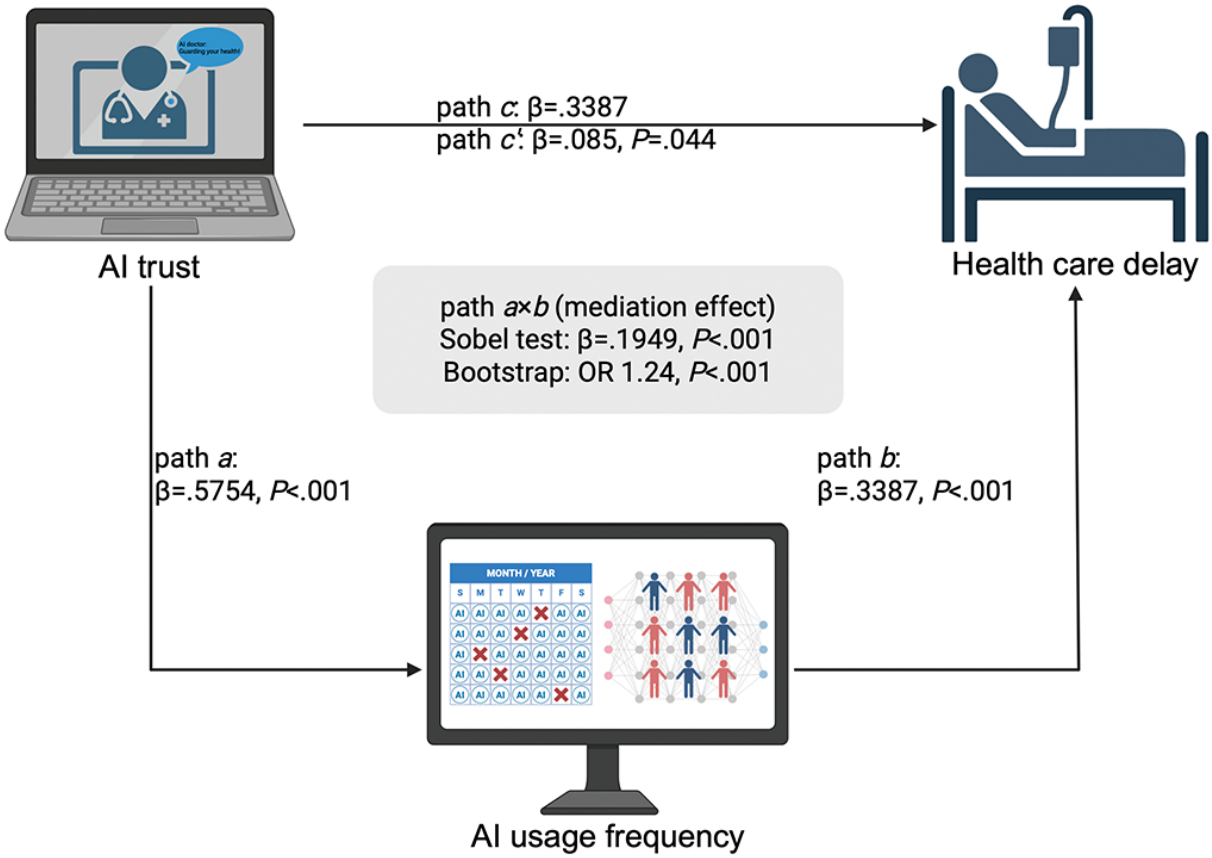

No Significant Moderation by Recommendation Exposure Level

In the stratified analysis by recommendation exposure, AI trust was significantly associated with delayed health care–seeking in the high exposure group (OR 1.11, 95% CI 1.01-1.23; *P*=.03), but not in the low exposure group (OR 1.04, 95% CI 0.88-1.22; *P*=.67). Notably, frequency of AI use was consistently associated with delay across both groups (low: OR 1.47, 95% CI 1.24-1.73; *P*<.001; high: OR 1.38, 95% CI 1.24-1.53; *P*<.001), suggesting a stable effect regardless of exposure level (Table 4). In the full model with interaction terms, the AI trust × recommendation exposure interaction was not statistically significant (interaction OR 0.97, 95% CI 0.83-1.14; *P*=.75; Figure 4A; detailed stratified analyses can be found in Table S4 in Multimedia Appendix 2). As shown in the predicted probability plot, the slopes of AI trust

XSL•FO
RenderX

on delay were similar between groups, with overlapping CIs, supporting the absence of a significant interaction (Figure 4A).

In summary, although AI trust showed a stronger association with health care delay in the high exposure group, the overall model did not indicate a statistically significant moderation effect. This suggests that while recommendation exposure may influence the magnitude of the trust-delay relationship, it does not alter its fundamental nature.

Table 4. Stratified analysis by recommendation exposure for the artificial intelligence (AI) trust-health care delay associations.

Outcome and variable	OR[a] (95% CI)	*P* value
Low recommendation exposure		
Trust in AI	1.04 (0.88-1.22)	<.001
Frequency of AI use	1.47 (1.24-1.73)	.67
Chronic disease (yes)	0.80 (0.47-1.37)	<.001
Age	1.01 (0.99-1.04)	.42
Gender (male)	1.35 (0.84-2.19)	.16
High recommendation exposure		
Trust in AI	1.11 (1.01-1.23)	.21
Frequency of AI use	1.38 (1.24-1.53)	<.001
Chronic disease (yes)	1.78 (1.30-2.44)	.03
Age	1.00 (0.99-1.01)	<.001
Gender (male)	1.17 (0.87-1.58)	<.001

[a]OR: odds ratio.

XSL•FO
RenderX

Figure 4. Stratified logistic regression and interaction modeling between artificial intelligence (AI) trust and recommendation exposure. Moderating effect of (A) AI recommendation exposure and (B) recommendation intensity on health care–seeking behavior. OR: odds ratio.

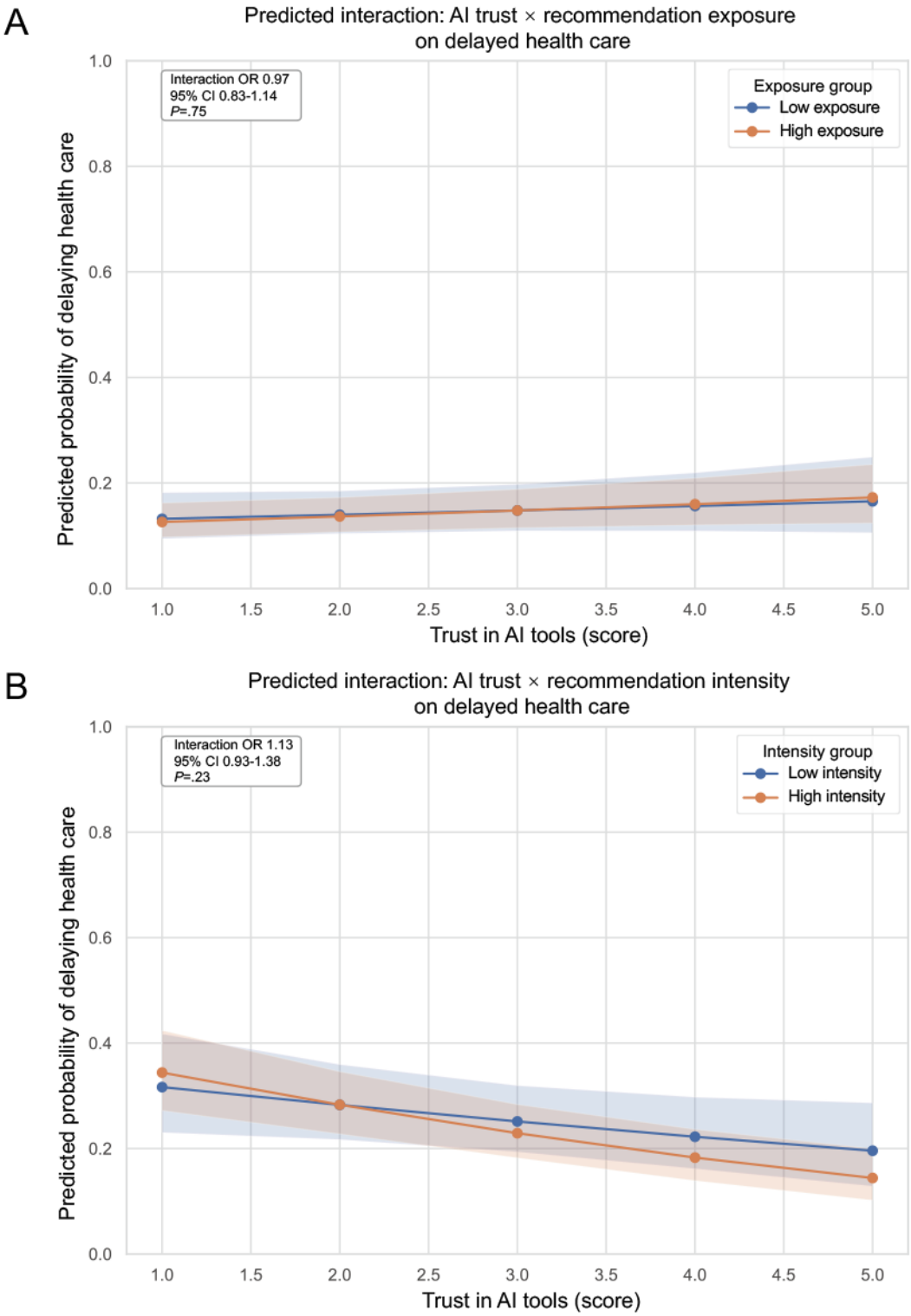

Limited Moderation by Recommendation Willingness Intensity

When stratified by willingness to recommend AI tools, AI trust was significantly associated with higher odds of health care delay in the low-intensity group (OR 1.16, 95% CI 1.00-1.34; *P*=.047), but not in the high-intensity group (OR 1.06, 95% CI 0.96-1.17; *P*=.27). Frequency of AI use remained a significant predictor of delay across both strata (low: OR 1.35, 95% CI 1.15-1.57; *P*<.001; high: OR 1.43, 95% CI 1.28-1.60; *P*<.001; Table 5). In the interaction model, the interaction term between AI trust and recommendation intensity was not statistically significant (interaction OR 1.13, 95% CI 0.93-1.38; *P*=.23; Figure 4B; detailed stratified analyses can be found in Table S5 in Multimedia Appendix 2). The interaction plot showed that the predicted probability of delaying care decreased with

XSL•FO
RenderX

increasing AI trust in both groups, with overlapping CIs, indicating no significant moderation effect (Figure 4B).

In summary, while AI trust appears to be more strongly associated with the odds of health care delay in the low-intensity group, recommendation intensity did not significantly moderate this association statistically. Its influence may reflect subtle variation in effect size rather than a true interaction.

Table 5. Stratified analysis by recommendation intensity for the artificial intelligence (AI) trust-health care delay associations.

Outcome and variable	OR[a] (95% CI)	*P* value
Low recommendation intensity		
Trust in AI	1.16 (1.00-1.34)	.047
Frequency of AI use	1.35 (1.15-1.57)	<.001
Chronic disease (yes)	1.53 (0.94-2.48)	.09
Age	1.00 (0.98-1.02)	.75
Gender (male)	1.07 (0.68-1.68)	.79
High recommendation intensity		
Trust in AI	1.06 (0.95-1.17)	.27
Frequency of AI use	1.43 (1.28-1.60)	<.001
Chronic disease (yes)	1.41 (1.02-1.95)	.04
Age	1.00 (0.99-1.02)	.71
Gender (male)	1.30 (0.96-1.77)	.09

[a]OR: odds ratio.

Scenario Simulations Reveal Combined Risk of AI Trust, Frequency, and Chronic Disease

In the multivariable logistic regression analysis, all 3 key predictors—AI trust, frequency of use, and chronic disease status—were significantly associated with health care delay. Specifically, each unit increase in AI trust was linked to 9% higher odds of delay (OR 1.09, 95% CI 1.00-1.18; *P*=.04). Frequency of AI use demonstrated an even stronger association (OR 1.40, 95% CI 1.28-1.53; *P*<.001), and chronic disease status significantly increased the odds of delay as well (OR 1.42, 95% CI 1.09-1.86; *P*=.01; Table 6).

Table 6. Multivariable logistic regression analysis for the associations between key predictors and health care delay.

Variable	OR[a] (95% CI)	*P* value
Trust in AI[b]	1.09 (1.00-1.18)	.04
Frequency of AI use	1.40 (1.28-1.53)	<.001
Chronic disease (yes)	1.42 (1.08-1.86)	.01

[a]OR: odds ratio.

[b]AI: artificial intelligence.

Scenario simulations indicated that, compared to a baseline of moderate trust and usage without a chronic condition (predicted probability=11.6%), increasing either AI trust (13.97%) or frequency (25.2%) alone raised the likelihood of delayed health care. The effect was most pronounced with frequency increases. When chronic illness was present alone, the predicted delay probability rose to 14.1%. Combining high trust and high frequency led to a delay probability of 30.5%, and when all 3 factors—high trust, high frequency, and chronic illness—were present, the probability escalated to 35.8%. These findings suggest that while increased trust and usage may enhance engagement with AI, they may paradoxically elevate the odds of behavioral delay due to potential overreliance or false reassurance. Therefore, intervention strategies should carefully balance cognitive trust with medical decision accuracy to mitigate potential adverse outcomes.

Agent-Based Simulation Reveals Bidirectional Feedback Between Trust and Delay

Simulation results demonstrated a consistent decline in the overall rate of health care delay, decreasing from 10.6% on day 1 to 9.5% by day 14 (Figure 5A). This trend reflects behavioral feedback, where agents who experienced delays reduced their AI usage frequency, ultimately lowering delay probabilities at the population level. Concurrently, the mean AI trust score declined steadily from approximately 1.95 to 1.49 over the 14-day period (Figure 5B). This indicates a natural erosion of trust in the absence of external reinforcement, suggesting a feedback loop where behavioral delays contribute to progressive trust deterioration. When stratifying agents by their predicted baseline delay probability, both high-risk and low-risk groups exhibited declining trust; however, the high-risk group experienced a steeper decline (from 2.55 to 1.74) compared to

XSL•FO
RenderX

the low-risk group (from 1.84 to 1.53; Figure 5C). This suggests that high-risk individuals may fall into a "high expectation-high disappointment" loop, accelerating trust collapse and reinforcing delay behavior. Detailed analyses for ABM can be found in Table S6 in Multimedia Appendix 2.

Figure 5. Agent-based modeling for the bidirectional feedback between artificial intelligence (AI) trust and delay behavior over a 14-day simulation period. (A) The overall rate of delayed health care. (B) The mean AI trust score. (C) AI trust score for the high-risk and low-risk groups.

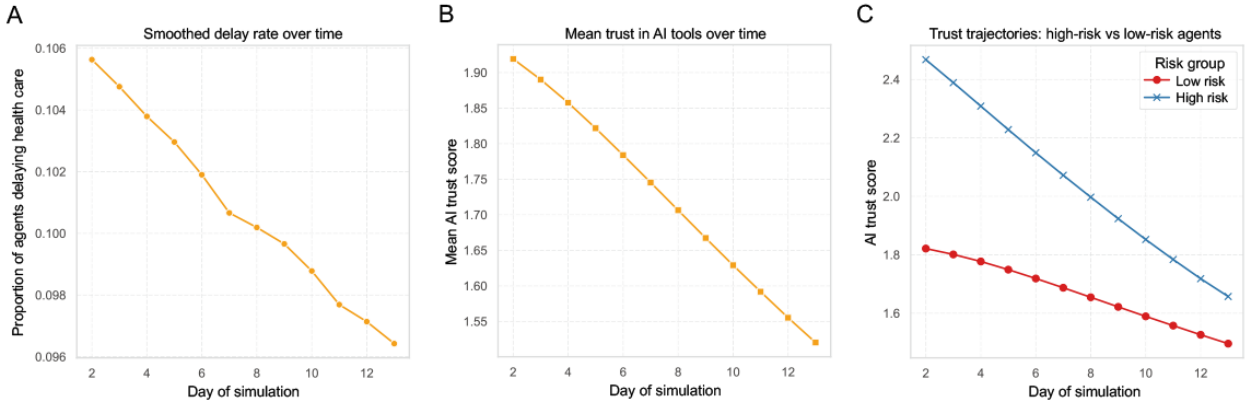

Overall, these simulations demonstrate the reciprocal feedback between trust and behavior, with individual heterogeneity compounding over time. The findings highlight the potential need for risk-stratified interventions or strategies designed to counteract trust erosion to prevent systemic trust breakdown.

Strategy Comparison: Broadcast Is the Most Effective, While Rewire May Backfire

Across 100 simulation trials, we examined the temporal effects of 4 intervention strategies on delay behavior. The broadcast strategy resulted in the most significant reduction in delay rates, reaching approximately 9.7% by day 14 (Figure 6). The reward strategy showed a moderate effect, with a slight downward trend compared to the baseline (Figure 6). In contrast, the rewire strategy led to an upward trend in delay rates after day 7, ultimately surpassing the baseline group, which suggests a potential amplification of trust polarization (Figure 6).

Figure 6. Comparison of 4 intervention strategies on delay behavior using artificial intelligence simulations. Strategies include baseline, broadcast messaging, behavioral reward, and network rewiring.

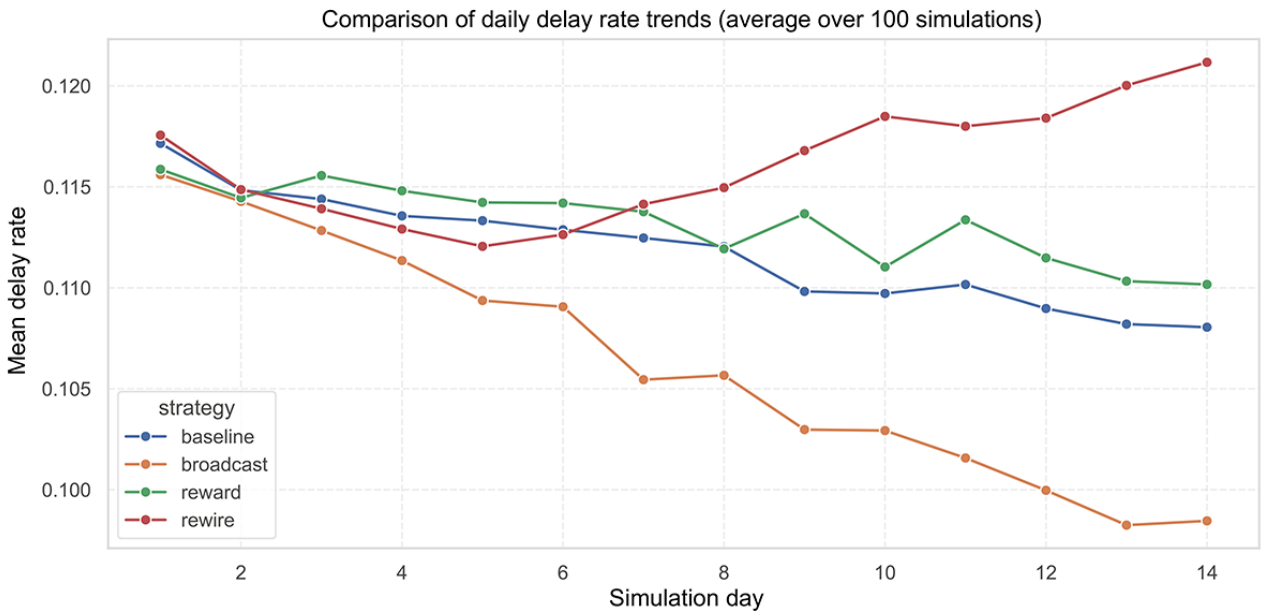

Logistic regression further quantified these effects (Table 7). Compared to the baseline, the broadcast strategy significantly reduced the odds of delay (OR 0.94, 95% CI 0.94-0.95; $P<.001$). The reward strategy indicated a slight increase in the odds of delay (OR 1.01, 95% CI 1.01-1.02; $P<.001$), while the rewire strategy demonstrated the most substantial increase in odds (OR 1.04, 95% CI 1.04-1.05; $P<.001$), indicating that this approach may exacerbate delay behavior under certain conditions. In summary, the broadcast strategy emerged as the most effective in reducing delay behavior within the current simulation framework, highlighting the potential advantages of proactive information signaling and group-level risk awareness alerts.

XSL•FO
RenderX

Table 7. Logistic regression analysis for the intervention strategies on health care delay.

Strategy	OR[a] (95% CI)	*P* value
Broadcast	0.94 (0.94-0.95)	<.001
Reward	1.01 (1.01-1.02)	<.001
Rewire	1.04 (1.04-1.05)	<.001

[a]OR: odds ratio.

To validate the robustness of the simulation across varying intervention configurations, we performed sensitivity analyses on 4 key parameters: initial trust level, broadcast penalty intensity, reward magnitude, and rewiring frequency. When varying the initial trust level (2.5, 3.0, and 3.5), higher values led to significantly increased rates of health care delay over time, suggesting that excessive trust may backfire despite its general benefits (Figure S1A in Multimedia Appendix 3). For broadcast penalty intensity, tested at values of 0.05, 0.1, and 0.2, the delay trajectories remained nearly identical, indicating minimal impact of penalty strength on intervention stability (Figure S1B in Multimedia Appendix 3). Similarly, altering the reward magnitude (0.03, 0.05, and 0.10) produced overlapping delay curves, demonstrating the strategy's robustness and absence of nonlinear effects (Figure S1C in Multimedia Appendix 3). In the case of rewiring frequency (2, 5, and 10 days), more frequent rewiring only slightly reduced early-stage delays, with long-term outcomes remaining consistent across intervals (Figure S1D in Multimedia Appendix 3). Overall, these results confirm that the simulation framework is stable, with behavioral outcomes staying similar across a range of parameters, highlighting the broader relevance and effectiveness of the modeled interventions.

Discussion

Principal Findings

This study is the first to systematically investigate the joint influence of AI trust, AI usage frequency, and chronic disease status on predicting delayed health care–seeking from the perspective of a behavior-trust feedback mechanism in China. Our findings demonstrate that both AI trust and usage frequency are significant predictors of health care delay. Importantly, AI trust not only indirectly elevates the odds of delay through increased usage frequency but also exerts a significant direct association when controlling for usage frequency, indicating a partial mediation pathway. Of note, individuals with chronic conditions inherently exhibit a predisposition toward delayed care. These odds are further exacerbated by the combined effects of high AI trust and frequent AI use. These results expand the current understanding of delayed health care behaviors. They emphasize the complex and dynamic interplay between technological trust and health-related decision-making. Additionally, they highlight that enhanced trust in AI may inadvertently contribute to adverse behavioral outcomes among patients with chronic diseases.

To investigate the psychological mechanisms underlying the "AI trust-frequent AI use-health care delay" pathway, we propose that cognitive bias and overreliance are key driving factors [31,32]. On one hand, patients with chronic conditions often experience stable and slowly progressing symptoms. They are more likely to view AI's reassuring suggestions as indicators of safety. As a result, they may underestimate the seriousness of their own symptoms. This sense of "false reassurance" is particularly pronounced among individuals with high levels of trust in AI [33]. Previous studies have shown that excessive trust in health AI recommendations can result in users neglecting bodily warning signs, ultimately leading to delays in health care–seeking [5]. On the other hand, frequent AI usage does not necessarily reflect higher health literacy. It also does not always indicate better self-management capacity [34]. Instead, it may suggest a psychological tendency to avoid traditional health care services. This tendency is especially common when medical care is expensive, time-consuming, or perceived as untrustworthy [35]. This form of "instrumental dependence" may drive individuals to rely on AI as a substitute source of advice when experiencing discomfort, instead of seeking timely professional care [36]. Therefore, this study emphasizes that the behavioral consequences of AI trust and usage frequency should not be interpreted simply as empowerment. Instead, it is important to explore the underlying psychological and behavioral mechanisms. This is essential for accurately assessing the true impact of technological interventions on health care behavior.

Although we further investigated the moderating effects of AI recommendation exposure and recommendation willingness on the "AI trust-health care delay" pathway, the interaction terms were not statistically significant. This indicates that these factors have a limited influence on the primary effect pathway. Subgroup analyses revealed some trend-level differences; for instance, individuals with higher recommendation willingness showed a slight reduction in delay behavior [37], but these effects did not achieve statistical significance. We speculate that a nonlinear threshold effect may be present. Once recommendation exposure reaches a certain frequency, its impact may plateau. This could result in a failure to further enhance trust or promote behavioral change. Another possibility is that the effectiveness of recommendations is highly context dependent. They may exert influence primarily when users experience high health anxiety, evident symptoms, or a strong sense of urgency [38,39]. Additionally, AI trust may be inherently unstable among individuals, easily influenced by emotional states, prior experiences, or public opinion [40]. These factors may further diminish the practical effectiveness of recommendations. While recommendation behaviors have the potential for positive influence, their underlying mechanisms appear more complex [41]. They may not adequately address the fundamental contradiction between AI trust and health care delay. Therefore, future interventions should not concentrate solely on increasing recommendation frequency or willingness.

XSL•FO
RenderX

Instead, it is crucial to address system-level design issues. This includes enhancing decision transparency and improving users' perceived control over health-related choices. Additionally, developing prompting strategies that better align with users' psychological models is essential. Such improvements are essential for facilitating meaningful and lasting behavioral change.

Scenario-based simulations further confirmed that the combination of high AI trust, frequent AI usage, and chronic disease status results in the greatest odds of delayed health care–seeking. This finding underscores the behavioral risks inherent in the trust feedback mechanism. Using behavioral decision theories, particularly the dual-process model, helps illuminate the psychological underpinnings of this phenomenon [42]. In high-trust situations, individuals tend to rely more on the intuitive system for decision-making, leading to quick, automatic judgments rather than thorough risk evaluations [43]. This "intuitive trust" is especially evident among patients with chronic illnesses, who often feel a sense of safety due to the relatively stable nature of their symptoms. When coupled with high trust in and frequent reliance on AI, these individuals are more likely to accept AI-generated suggestions without critically assessing their health status or seeking professional medical advice [44]. This can create a "false reassurance" effect, significantly heightening the risk of delayed health care. These findings serve as a cautionary note for the future design of AI-based health tools. Simply pursuing user trust is insufficient; it is essential to strike a balance between fostering trust and maintaining risk awareness. Both clinical practitioners and AI developers need to consider how to enhance user experience while mitigating the unintended consequences of overreliance [45]. Future AI systems should be designed to promote rational judgment, especially among high-risk groups with chronic conditions. Incorporating interactive features that provide calibrated risk reminders may be necessary to prevent the unconscious shift from trust to delayed health care behavior.

Our ABM offered a dynamic and systems-level understanding of how AI trust and delay behavior coevolve over time. Unlike traditional regression models that provide static snapshots [23], ABM captures how small differences in initial trust and exposure can compound through individual learning and social influence. Over a 14-day simulation period, we observed a steady decline in both trust and health care–seeking behavior, indicating a reciprocal erosion loop: delayed health care leads to dissatisfaction or unmet expectations, which in turn diminishes trust and further health care delay [46]. Importantly, agents stratified by predicted baseline odds (above vs below 20%) showed divergent trust trajectories. High-risk individuals began with higher trust but experienced a sharper decline—suggesting a "high expectation-high disappointment" loop. This highlights how perceived safety in chronic illness, when coupled with excessive AI reliance, may paradoxically lead to trust collapse and amplified delay behavior. Such patterns would be difficult to detect using empirical data alone, underlining the value of simulation modeling in behavioral health research. Beyond reproducing observed behaviors, ABM also allowed us to simulate system-wide interventions and uncover nonlinear responses to trust regulation strategies. Of the 3 mechanisms tested, broadcast messaging consistently reduced delay by maintaining population-level risk awareness. In contrast, network rewiring unexpectedly increased delay, likely due to the formation of echo chambers among high-trust individuals, a phenomenon we term "trust polarization" [47]. This emergent property demonstrates that even well-intentioned peer-based reinforcement may backfire under certain trust dynamics. Overall, these simulations underscore the importance of viewing AI trust not merely as an individual attribute, but as a collective behavioral variable that evolves across time, context, and networks [48]. While the ABM provides valuable insights into potential system dynamics and the comparative theoretical performance of different strategies, these findings serve as a proof of concept. Their real-world effectiveness and causal impact must be rigorously tested in future randomized controlled trials or natural experiments.

Simulation-based evaluations of 3 intervention strategies indicate that broadcast messaging is the most effective for reducing delayed health care behavior. This suggests that public health interventions delivering wide-reaching, consistent risk reminders may be the most cost-effective approach [49]. Conversely, the reward-based strategy was less effective, likely due to insufficient incentives or limited reach, which hindered meaningful behavior change at the population level [50]. Unexpectedly, the network rewiring strategy not only failed to reduce delays but exacerbated them. We hypothesize that this may result from the formation of information echo chambers among high-trust individuals [51]. In these tightly connected groups, AI trust can become mutually reinforced, amplifying the "AI trust-health care delay" pathway and leading to what we term trust polarization. These findings underscore the need to prioritize intervention strategies that broadly disseminate risk information and continuously enhance risk awareness. Future health interventions should focus on scalable and sustainable communication mechanisms rather than relying on individual "opinion leaders" or short-term incentives. This approach provides valuable insights into integrating AI and health behavior, while also offering empirical evidence to inform public health policy development.

Limitations

Despite proposing a novel "trust-behavior" feedback mechanism and validating various intervention strategies through simulation, this study has several limitations. First, the cross-sectional and observational nature of our survey data precludes definitive causal inference. While we controlled for several demographic and health factors, the potential for omitted variable bias remains. Unmeasured confounders, such as general health literacy and prior negative experiences with the health care system, could independently influence both AI trust and the propensity to delay care. To establish causality, future research should use randomized controlled trials that investigate participants' trust in AI. This approach would enable more conclusive mediation and moderation analyses and would represent a critical continuation of our future research efforts. Second, measurements of AI trust and usage frequency were based on self-reported data. This approach may introduce recall bias, social desirability effects, or subjective judgment. Consequently, these factors could potentially affect the accuracy

XSL•FO
RenderX

of the findings. Although we introduced a group-based behavioral simulation model to address the static nature of cross-sectional data, this model simplifies the complex dynamics of trust evolution in real-world contexts. Future research could incorporate reinforcement learning or dynamic trust modeling approaches to better capture human cognitive and behavioral trajectories [52]. Third, the generalizability of our findings may be limited by the specific sociotechnical context of China. Our recruitment relied on dominant Chinese digital platforms (WeChat and QQ). The observed relationships between AI trust, usage, and health care delays are likely influenced by China's unique health care system, technology ecosystem, and cultural norms. Therefore, caution is warranted when extending these results to other countries with different health care policies, technology adoption patterns, and cultural attitudes toward AI in medicine. Fourth, our recruitment via social media and hospital portals, while enabling broad access, prevents precise quantification of response rates from each channel. Although we estimate that most responses came from WeChat and QQ, with a smaller proportion from hospital portals, the anonymous nature of the survey precluded channel-specific stratification. Additionally, this approach likely oversampled individuals who are more digitally literate or proactively engaged with health care information. This may limit the generalizability of our findings to populations with limited digital access or lower health literacy. Fifth, the questionnaire did not collect data on rural or urban residence or socioeconomic status, limiting our ability to examine these potential sociodemographic influences on the trust-delay pathway. Future studies would benefit from incorporating these variables. Sixth, our sample contained a slightly higher proportion of female participants (1345/2460, 54.7%), which may influence generalizability. However, we adjusted for gender in all analyses and found no evidence of significant effect modification. Finally, we advocate for the development of AI health tools with stratified trust management functions [53]. Such systems should provide tailored risk alerts for high-risk patients with chronic conditions, promoting rational trust for genuine health empowerment rather than fostering passive dependence.

Conclusions

This study identifies a nuanced relationship between AI trust and delayed health care–seeking behavior. While trust in AI tools can enhance user engagement, it may also be associated with delayed health care, particularly among individuals with chronic conditions or higher levels of AI use. By integrating survey data with ABM, our study moves beyond static associations and illustrates how AI trust and delay behaviors may interact and reinforce one another over time. This dynamic perspective highlights how individual-level trust processes can accumulate into system-level patterns, including trust polarization and collective delay. Our findings suggest that the design of AI health tools should prioritize calibrated decision support rather than full automation. Encouraging a balanced interaction between user autonomy and technological assistance may help mitigate delay-related risks and improve decision quality in digital health contexts. Beyond individual-level tool design, these insights may also inform population-level strategies for trust governance and risk communication in AI-driven health care systems.

Acknowledgments

All authors thank the Central South University for providing the research platform. We also thank the BioRender platform [54] for providing the tools to create Figure 1 and Figure 3, with all elements authorized for use.

Generative AI was not used in any portion of the manuscript writing.

Funding

This work was supported by the National Natural Science Foundation of China (82501245), Key Research and Development Projects in the Field of Social Development in Hunan Province (2020SK3030), Project of Science and Technology of Hunan Province (2021JJ40932), China Postdoctoral Science Foundation (2025M771943), and Shanghai Post-doctoral Excellence Program (2024409).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: XC
Methodology: XC, YC
Formal analysis: WL
Investigation: XC, WL
Resources: WL
Writing – original draft: XC
Writing – review & editing: WL, WS, YC
Supervision: YC, JZ
Funding acquisition: YC, JZ

XSL•FO
RenderX

Project administration: JZ

Authors YC (1917@sjtu.edu.cn) and JZ (zhoujianda@csu.edu.cn) are co-corresponding authors for this article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
English-translated survey questionnaire.
[DOCX File , 16 KB - jmir_v28i1e82170_app1.docx]

Multimedia Appendix 2
Supplementary Tables S1-S6.
[DOCX File , 29 KB - jmir_v28i1e82170_app2.docx]

Multimedia Appendix 3
Supplementary Figure S1.
[DOCX File , 315 KB - jmir_v28i1e82170_app3.docx]

References

1. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J 2021;8(2):e188-e194 [FREE Full text] [doi: 10.7861/fhj.2021-0095] [Medline: 34286183]
2. Gulfidan G, Beklen H, Arga K. Artificial intelligence as accelerator for genomic medicine and planetary health. OMICS 2021;25(12):745-749. [doi: 10.1089/omi.2021.0170] [Medline: 34780300]
3. Topol E. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
4. Grossi E. The long journey of artificial intelligence in medicine: an overview. Clin Exp Rheumatol 2025;43(5):815-821 [FREE Full text] [doi: 10.55563/clinexprheumatol/oamfed] [Medline: 40338059]
5. Quinn T, Senadeera M, Jacobs S, Coghlan S, Le V. Trust and medical AI: the challenges we face and the expertise needed to overcome them. J Am Med Inform Assoc 2021;28(4):890-894 [FREE Full text] [doi: 10.1093/jamia/ocaa268] [Medline: 33340404]
6. Aung Y, Wong D, Ting D. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. Br Med Bull 2021;139(1):4-15. [doi: 10.1093/bmb/ldab016] [Medline: 34405854]
7. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
8. Lynch C, Liston C. New machine-learning technologies for computer-aided diagnosis. Nat Med 2018;24(9):1304-1305. [doi: 10.1038/s41591-018-0178-4] [Medline: 30177823]
9. Ploug T, Holm S. The right to refuse diagnostics and treatment planning by artificial intelligence. Med Health Care Philos 2020;23(1):107-114. [doi: 10.1007/s11019-019-09912-8] [Medline: 31359302]
10. Panch T, Mattie H, Celi L. The "inconvenient truth" about AI in healthcare. NPJ Digit Med 2019;2:77 [FREE Full text] [doi: 10.1038/s41746-019-0155-4] [Medline: 31453372]
11. Dalton-Brown S. The ethics of medical ai and the physician-patient relationship. Camb Q Healthc Ethics 2020;29(1):115-121. [doi: 10.1017/S0963180119000847] [Medline: 31858938]
12. Krist A, Tong S, Aycock R, Longo D. Engaging patients in decision-making and behavior change to promote prevention. Stud Health Technol Inform 2017;240:302. [doi: 10.3233/978-1-61499-790-0-284]
13. Johnson K, Wei W, Weeraratne D, Frisse M, Misulis K, Rhee K, et al. Precision medicine, AI, and the future of personalized health care. Clin Transl Sci 2021;14(1):86-93 [FREE Full text] [doi: 10.1111/cts.12884] [Medline: 32961010]
14. ALHosni F, Al Qadire M, Omari O, Al Raqaishi H, Khalaf A. Symptom prevalence, severity, distress and management among patients with chronic diseases. BMC Nurs 2023;22(1):155 [FREE Full text] [doi: 10.1186/s12912-023-01296-8] [Medline: 37149599]
15. Ajayi K, Wachira E, Onyeaka H, Montour T, Olowolaju S, Garney W. The use of digital health tools for health promotion among women with and without chronic diseases: insights from the 2017-2020 health information national trends survey. JMIR Mhealth Uhealth 2022;10(8):e39520 [FREE Full text] [doi: 10.2196/39520] [Medline: 35984680]
16. Yun H, Bickmore T. Online health information-seeking in the era of large language models: cross-sectional web-based survey study. J Med Internet Res 2025;27:e68560 [FREE Full text] [doi: 10.2196/68560] [Medline: 40163112]
17. Al-Antari M. Artificial intelligence for medical diagnostics-existing and future ai technology!. Diagnostics (Basel) 2023;13(4):688 [FREE Full text] [doi: 10.3390/diagnostics13040688] [Medline: 36832175]
18. Cross J, Choma M, Onofrey J. Bias in medical AI: implications for clinical decision-making. PLOS Digit Health 2024;3(11):e0000651. [doi: 10.1371/journal.pdig.0000651] [Medline: 39509461]

XSL•FO
RenderX

19. Figure 1. Schematic overview of the study design and methodology. Based on online questionnaires, the analysis framework includes mediation analysis to explore the relationship between AI trust, usage frequency, and delayed healthcare seeking behavior. Additionally, agent-based modeling (ABM) were employed to model the dynamic feedback loops over a 14-day period, incorporating three intervention strategies: Broadcast, Reward, and Rewire. These strategies aimed to evaluate how the interventions affected trust and delay behavior within the population.. BioRender. URL: https://app.biorender.com/citation/694800462abb8b925ab7f3cb [accessed 2026-01-18]
20. Levitt H, Bamberg M, Creswell J, Frost D, Josselson R, Suárez-Orozco C. Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA publications and communications board task force report. Am Psychol 2018;73(1):26-46 [FREE Full text] [doi: 10.1037/amp0000151] [Medline: 29345485]
21. Journal Article Reporting Standards (JARS). American Psychological Association. URL: https://apastyle.apa.org/jars [accessed 2025-12-31]
22. Conroy G, Mallapaty S. How China created AI model deepSeek and shocked the world. Nature 2025;638(8050):300-301. [doi: 10.1038/d41586-025-00259-0] [Medline: 39885352]
23. Tracy M, Cerdá M, Keyes K. Agent-based modeling in public health: current applications and future directions. Annu Rev Public Health 2018;39:77-94 [FREE Full text] [doi: 10.1146/annurev-publhealth-040617-014317] [Medline: 29328870]
24. Li Y, Lawley M, Siscovick D, Zhang D, Pagán JA. Agent-based modeling of chronic diseases: a narrative review and future research directions. Prev Chronic Dis 2016;13:E69 [FREE Full text] [doi: 10.5888/pcd13.150561] [Medline: 27236380]
25. Yang Y. A narrative review of the use of agent-based modeling in health behavior and behavior intervention. Transl Behav Med 2019;9(6):1065-1075. [doi: 10.1093/tbm/iby132] [Medline: 30649559]
26. Murase Y, Jo H, Török J, Kertész J, Kaski K. Deep learning exploration of agent-based social network model parameters. Front Big Data 2021;4:739081 [FREE Full text] [doi: 10.3389/fdata.2021.739081] [Medline: 34661097]
27. Blok D, Simoski B, van Woudenberg TJ, Buijzen M. The usefulness of web-based communication data for social network health interventions: agent-based modeling study. JMIR Pediatr Parent 2023;6:e44849 [FREE Full text] [doi: 10.2196/44849] [Medline: 37991813]
28. Marshall B, Galea S. Formalizing the role of agent-based modeling in causal inference and epidemiology. Am J Epidemiol 2015;181(2):92-99 [FREE Full text] [doi: 10.1093/aje/kwu274] [Medline: 25480821]
29. Grimm V, Berger U, DeAngelis D, Polhill J, Giske J, Railsback S. The ODD protocol: a review and first update. Ecol Model 2010;221(23):2760-2768 [FREE Full text] [doi: 10.1016/j.ecolmodel.2010.08.019]
30. Figure 3. Mediation model demonstrating how AI trust influences delayed healthcare seeking through increased frequency of AI usage. Path a shows the significant effect of AI trust on usage frequency, while path b indicates the association between usage frequency and delay. The indirect effect (ab) and direct effect (c′) are also represented, highlighting the statistical significance of the mediation pathway.. BioRender. URL: https://app.biorender.com/citation/69480d355b576217dd8c87e3 [accessed 2026-01-18]
31. Vicente L, Matute H. Humans inherit artificial intelligence biases. Sci Rep 2023;13(1):15737 [FREE Full text] [doi: 10.1038/s41598-023-42384-8] [Medline: 37789032]
32. Klingbeil A, Grützner C, Schreck P. Trust and reliance on AI—an experimental study on the extent and costs of overreliance on AI. Comput Hum Behav 2024;160:108352 [FREE Full text] [doi: 10.1016/j.chb.2024.108352]
33. Stein J, Messingschlager T, Gnambs T, Hutmacher F, Appel M. Attitudes towards AI: measurement and associations with personality. Sci Rep 2024;14(1):2909 [FREE Full text] [doi: 10.1038/s41598-024-53335-2] [Medline: 38316898]
34. Dong C, Ji Y, Fu Z, Qi Y, Yi T, Yang Y, et al. Precision management in chronic disease: an AI empowered perspective on medicine-engineering crossover. iScience 2025;28(3):112044 [FREE Full text] [doi: 10.1016/j.isci.2025.112044] [Medline: 40104052]
35. Khanna N, Maindarkar M, Viswanathan V, Fernandes J, Paul S, Bhagawati M, et al. Economics of artificial intelligence in healthcare: diagnosis vs treatment. Healthcare (Basel) 2022;10(12):2493 [FREE Full text] [doi: 10.3390/healthcare10122493] [Medline: 36554017]
36. Huang S, Lai X, Ke L, Li Y, Wang H, Zhao X, et al. AI technology panic-is AI dependence bad for mental health? A cross-lagged panel model and the mediating roles of motivations for AI use among adolescents. Psychol Res Behav Manag 2024;17:1087-1102 [FREE Full text] [doi: 10.2147/PRBM.S440889] [Medline: 38495087]
37. Rung J, Madden G. Experimental reductions of delay discounting and impulsive choice: a systematic review and meta-analysis. J Exp Psychol Gen 2018;147(9):1349-1381 [FREE Full text] [doi: 10.1037/xge0000462] [Medline: 30148386]
38. Asmundson G, Taylor S. How health anxiety influences responses to viral outbreaks like COVID-19: what all decision-makers, health authorities, and health care professionals need to know. J Anxiety Disord 2020;71:102211 [FREE Full text] [doi: 10.1016/j.janxdis.2020.102211] [Medline: 32179380]
39. Thakkar A, Gupta A, De Sousa A. Artificial intelligence in positive mental health: a narrative review. Front Digit Health 2024;6:1280235 [FREE Full text] [doi: 10.3389/fdgth.2024.1280235] [Medline: 38562663]
40. Li Y, Wu B, Huang Y, Luan S. Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-AI trust. Front Psychol 2024;15:1382693 [FREE Full text] [doi: 10.3389/fpsyg.2024.1382693] [Medline: 38694439]

XSL•FO
RenderX

41. Laursen B, Veenstra R. Toward understanding the functions of peer influence: a summary and synthesis of recent empirical research. J Res Adolesc 2021;31(4):889-907 [FREE Full text] [doi: 10.1111/jora.12606] [Medline: 34820944]
42. Barrouillet P. Dual-process theories and cognitive development: advances and challenges. Dev Rev 2011;31(2-3):79-85 [FREE Full text] [doi: 10.1016/j.dr.2011.07.002]
43. Chick C. Cooperative versus competitive influences of emotion and cognition on decision making: a primer for psychiatry research. Psychiatry Res 2019;273:493-500. [doi: 10.1016/j.psychres.2019.01.048] [Medline: 30708200]
44. Shekar S, Pataranutaporn P, Sarabu C, Cecchi GA, Maes P. People overtrust AI-generated medical advice despite low accuracy. NEJM AI 2025;2(6):AIoa2300015 [FREE Full text] [doi: 10.1056/aioa2300015]
45. Siala H, Wang Y. SHIFTing artificial intelligence to be responsible in healthcare: a systematic review. Soc Sci Med 2022;296:114782 [FREE Full text] [doi: 10.1016/j.socscimed.2022.114782] [Medline: 35152047]
46. Nong P, Ji M. Expectations of healthcare AI and the role of trust: understanding patient views on how AI will impact cost, access, and patient-provider relationships. J Am Med Inform Assoc 2025;32(5):795-799. [doi: 10.1093/jamia/ocaf031] [Medline: 40036944]
47. Fränken J, Pilditch T. Cascades across networks are sufficient for the formation of echo chambers: an agent-based model. JASSS 2021;24(3):1 [FREE Full text] [doi: 10.18564/jasss.4566]
48. Nong P, Platt J. Patients' Trust in Health Systems to Use Artificial Intelligence. JAMA Netw Open 2025;8(2):e2460628 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.60628] [Medline: 39951270]
49. Muuraiskangas S, Harjumaa M, Kaipainen K, Ermes M. Process and effects evaluation of a digital mental health intervention targeted at improving occupational well-being: lessons from an intervention study with failed adoption. JMIR Ment Health 2016;3(2):e13 [FREE Full text] [doi: 10.2196/mental.4465] [Medline: 27170553]
50. Michaelsen M, Esch T. Understanding health behavior change by motivation and reward mechanisms: a review of the literature. Front Behav Neurosci 2023;17:1151918 [FREE Full text] [doi: 10.3389/fnbeh.2023.1151918] [Medline: 37405131]
51. Hartmann D, Wang S, Pohlmann L, Berendt B. A systematic review of echo chamber research: comparative analysis of conceptualizations, operationalizations, and varying outcomes. J Comput Soc Sc 2025;8(2):52 [FREE Full text] [doi: 10.1007/s42001-025-00381-z]
52. Das A, Islam M. SecuredTrust: a dynamic trust computation model for secured communication in multiagent systems. IEEE Trans Dependable and Secur Comput 2012;9(2):261-274 [FREE Full text] [doi: 10.1109/tdsc.2011.57]
53. Carroll N, Jones A, Burkard T, Lulias C, Severson K, Posa T. Improving risk stratification using AI and social determinants of health. Am J Manag Care 2022;28(11):582-587 [FREE Full text] [doi: 10.37765/ajmc.2022.89261] [Medline: 36374616]
54. bioRender. URL: https://www.biorender.com/ [accessed 2026-01-08]

Abbreviations

ABM: agent-based modeling
AI: artificial intelligence
MCAR: missing completely at random
ODD: overview, design concepts, and details
OR: odds ratio

Edited by S Brini; submitted 10.Aug.2025; peer-reviewed by MF Wibowo, Z Cui; comments to author 29.Sep.2025; revised version received 12.Dec.2025; accepted 21.Dec.2025; published 03.Feb.2026.

Please cite as:
Cai X, Li W, Shi W, Cai Y, Zhou J
Behavioral Dynamics of AI Trust and Health Care Delays Among Adults: Integrated Cross-Sectional Survey and Agent-Based Modeling Study
J Med Internet Res 2026;28:e82170
URL: https://www.jmir.org/2026/1/e82170
doi:10.2196/82170
PMID:

©Xueyao Cai, Weidong Li, Wenjun Shi, Yuchen Cai, Jianda Zhou. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Screen Time and Chronic Pain Health: Mendelian Randomization Study

Jiahui Jiang[1*], MS; Chunyan Pu[1*], MPH; Jiarui Cai[1*], MA, MPH; Chuan Yu[1], PhD; Zhenmi Liu[1], PhD; Chenghan Xiao[2], PhD

[1]Department of Maternal and Child Health, West China School of Public Health and West China Fourth Hospital, Sichuan University, Chengdu, Sichuan, China

[2]The Medical Laboratory Center, Hainan General Hospital, Hainan Medical University Hainan Hospital, Haikou, Hainan, China

[*]these authors contributed equally

Corresponding Author:
Zhenmi Liu, PhD
Department of Maternal and Child Health
West China School of Public Health and West China Fourth Hospital
Sichuan University
No.18, Section 3, Renmin South Road, Wuhou District
Chengdu, Sichuan, 610061
China
Phone: 86 18116636482
Email: zhenmiliu@scu.edu.cn

Abstract

Background: The rapid proliferation of electronic devices has increased screen time, raising concerns about its potential health effects, including chronic pain. However, existing studies have limitations in scope and causal inference, with inconsistent findings and a lack of exploration of potential biological mechanisms.

Objective: The objective of our study was to investigate the causal associations and potential shared biological mechanisms between different forms of screen time and various chronic pain phenotypes.

Methods: Leveraging genome-wide association study data, we investigated the association and potential shared biological mechanisms between screen time (time spent watching television, time spent using computer, and length of mobile phone use) and chronic pain phenotypes (including multisite chronic pain [MCP], back, knee, neck or shoulder, hip pain, and headaches). Two-sample Mendelian randomization (MR), reverse MR and multivariable Mendelian randomization (MVMR) analysis were performed to examine associations between screen time and chronic pain. Summary data–based Mendelian randomization (SMR), transcriptome-wide association study (TWAS), and colocalization analysis were used to identify the shared genes and potential biological mechanism.

Results: MR analysis revealed that time spent watching television and length of mobile phone use were positively associated with several types of chronic pain, while time spent using computer showed a negative association. Specifically, time spent watching television was positively associated with the risk of MCP (P=1.05×10^{-31}; odds ratio [OR] 1.61, 95% CI 1.49-1.74), back pain (P=2.41×10^{-8}; OR 1.14, 95% CI 1.09-1.19), knee pain (P=7.10×10^{-6}; OR 1.09, 95% CI 1.05-1.13), neck or shoulder pain, and hip pain. Length of mobile phone use was positively associated with the risk of MCP (P=2.15×10^{-5}; OR 1.22, 95% CI 1.11-1.34), headaches, and neck or shoulder pain. However, time spent using computer was negatively associated with the risk of MCP (P<.001; OR 0.83, 95% CI 0.75-0.92), back pain, and knee pain. The reverse MR results showed that MCP was positively associated with time spent watching television (P=4.8×10^{-7}; OR 1.27, 95% CI 1.16-1.4) and length of mobile phone use (P=3.38×10^{-5}; OR 1.29, 95% CI 1.14-1.45), while the association with time spent using computer (P=.61; OR 0.97, 95% CI 0.87-1.09) was not statistically significant. The MVMR results failed to meet the criterion that all conditional F-statistics exceed 10. Integrative 3 analysis methods identified overlapping genes, with *CEP170* emerging as a key gene consistently supported by SMR, TWAS, and colocalization analysis in the relationship between time spent using computer and MCP.

Conclusions: Our findings demonstrate an association between screen time and various aspects of chronic pain. The *CEP170* gene might contribute to the shared biological mechanism between time spent using computer and MCP risk. However, due to the absence of robust MVMR results, the potential influence of confounding factors cannot be ruled out.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e78233) doi:10.2196/78233

KEYWORDS

causality; chronic pain; medical genetics; multisite chronic pain; screen time

Introduction

With the widespread popularity of electronic devices, the screen time spent by people on phones, computers, and televisions has increased significantly every day. Recently, an increasing number of studies have shown that excessive screen time may bring a series of health risks [1,2]. A study on children and adolescents has shown that long-term screen time exposure is closely related to obesity, restricted cognitive and language development, decreased academic performance, and altered sleep patterns [3,4]. Studies on adults have found that excessive screen time is associated with reduced physical activity and increased BMI [5]. Moreover, studies have found that the incidence of headaches increases with increased screen time, with similar trends observed in both adolescents and adults [6-9]. Research has also shown that prolonged computer use can contribute to pain in several areas, such as the neck, shoulders, and back [10].

While some studies have explored the relationship between screen time and chronic pain, the current evidence is still limited. An important limitation is that existing studies have shortcomings in assessment scope and underlying relationship to fully reveal the association between the two. We found that most studies have focused on the use of a single type of screen time. One study only pointed out that prolonged computer use can cause pain in multiple areas [10], but ignored the effects of different types of screen time. Moreover, most studies on screen time and chronic pain tend to focus on specific body regions, such as lower back pain or headaches [11,12], while overlooking the impact of screen time on multisite chronic pain (MCP) and the combined effects of different types of chronic pain. In addition, the results of existing studies are inconsistent, with some studies finding an association between screen time and chronic pain [13], while others fail to observe a significant association [14]. This inconsistency may be caused by confounding bias or a weaker ability to infer causality.

Another important limitation in the current research concerns the mechanisms associated with screen time and chronic pain. Some studies have suggested that the brightness or light wave frequency of screens may trigger migraine attacks, while prolonged exposure to screens may lower the threshold for migraines [15-17], making them more likely to be triggered by other factors. However, this explanation lacks a detailed understanding of the biological mechanisms involved. To date, the underlying biological mechanisms linking screen time and chronic pain have remained largely unexplored, highlighting the need for future research to address this issue and explore whether different devices cause chronic pain through common or different mechanisms.

The genome-wide association studies (GWASs) of MCP [18], the development of genetic epidemiology, and multiomics integration analysis methods provide new opportunities to address these limitations [19,20]. Two-sample Mendelian randomization (MR) analysis can be used to examine the potential relationship between screen time and chronic pain. MR analysis uses genetic variants significantly associated with exposure as instrumental variables (IVs) to infer causality, thereby minimizing confusion and reverse causality [21]. In addition, summary data–based Mendelian randomization (SMR) and transcriptome-wide association study (TWAS) analysis can help identify overlapping genes between the two. In addition, colocalization analysis can assess whether these associations share causal genetic variation, thereby strengthening the evidence for shared biological mechanisms. By using multiple genetic epidemiological methods, we can explore the biological relationship between screen time and chronic pain.

In this study, we aimed to address these limitations. First, we examined whether screen time was associated with MCP and chronic pain in 5 other body parts. Next, we identified overlapping genes through genetic epidemiological methods approach and explored their potentially shared biological mechanism, aiming to provide new insights into the complexity of chronic pain.

Methods

Study Overview

As shown in Figure 1, this study follows a 3-stage approach. In Phase 1, we performed a 2-sample MR analysis and reverse MR to investigate the phenotypic association between screen time and chronic pain. Considering the mutual influence among multiple exposures, we conducted a multivariate Mendelian randomization (MVMR) analysis. In Phase 2, we first performed SMR analysis combined with expression Quantitative Trait Locus (eQTL) summary datasets to explore associations between gene expression and phenotypic traits. We then performed TWAS analysis to identify gene expression associated with screen time and chronic pain, and assessed whether the same genetic variants were shared through colocalization analysis, providing stronger genetic evidence. In Phase 3, we further integrated the results of SMR, TWAS, and colocalization analysis. We aimed to identify genes that were consistently supported across these methods. The overlapping genes provide more robust evidence of shared biological mechanisms in the association between screen time and chronic pain.

XSL•FO
RenderX

Figure 1. Flowchart of the overall study design. eQTL: expression Quantitative Trait Locus; FDR: false discovery rate; GTEx: Genotype-Tissue Expression; GWAS: genome-wide association study; IV: instrumental variable; IVW: inverse variance weighted; LD: linkage disequilibrium; MCP: multisite chronic pain; MR: Mendelian randomization; MVMR: multivariate Mendelian randomization; OR: odds ratio; SMR: summary data–based Mendelian randomization analysis; SNP: single-nucleotide polymorphism; TWAS: transcriptome-wide association study; TV: television.

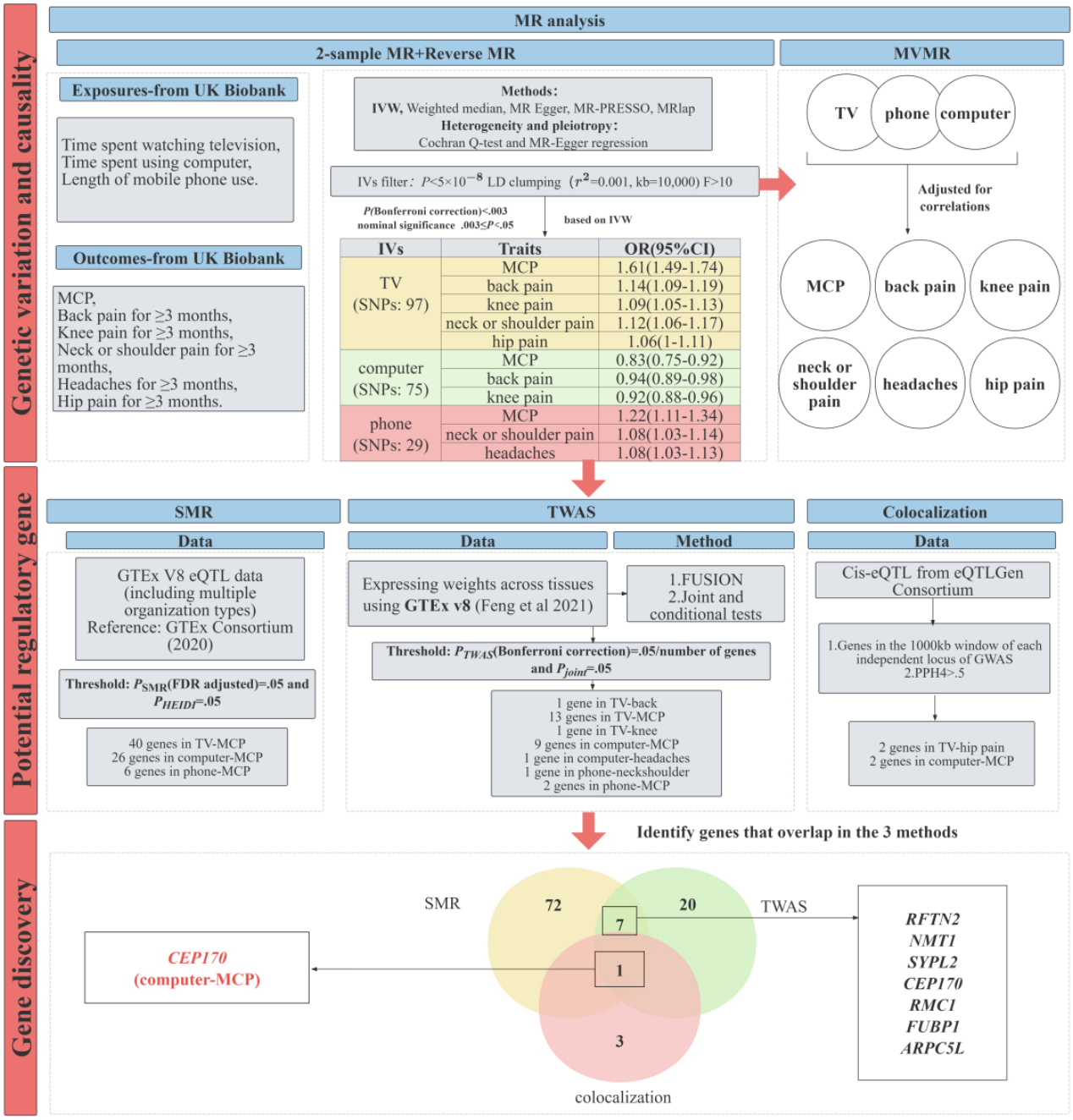

Data Sources for Exposures, Outcomes, and eQTL Summary Data

Exposures

The exposure phenotype in this study was screen time, which included time spent watching television, time spent using computer, and length of mobile phone use. These phenotypes were derived from the UK Biobank touchscreen questionnaire. The question for time spent using computer explicitly stated, "Do not include using a computer at work," whereas the questions for television watching and mobile phone use captured overall daily use. We provided detailed data sources about screen time in Multimedia Appendix 1. To address the issue of weak instrument bias, a genome-wide significance threshold of $P<5\times10^{-8}$ was used as the default criterion for identifying single-nucleotide polymorphisms (SNPs); simultaneously, to optimize the results and eliminate SNPs in strong linkage disequilibrium (LD), an LD clustering method was adopted, with thresholds of r^2=0.001 and kilobase, kb=10,000 to ensure accurate and precise clustering. r^2 is the squared correlation coefficient between 2 SNPs' allele counts, used to quantify LD. In the 2-sample MR analysis, we selected 112, 82, and 32 independently significant SNPs located on the autosomes as IVs of time spent watching television, time spent using

XSL•FO
RenderX

computer, and length of mobile phone use, respectively. Details of these SNPs are presented in Multimedia Appendices 2-4.

Outcomes

The outcomes included 6 chronic pain phenotypes, namely MCP, back pain for ≥3 months, knee pain for ≥3 months, neck or shoulder pain for ≥3 months, headaches for ≥3 months, and hip pain for ≥3 months. Among them, MCP is by far the largest GWAS summary statistic for all chronic pain phenotypes, involving 387,649 people from the UK Biobank [18]. Data for the remaining 5 types of chronic pain came from the Integrative Epidemiology Unit (IEU) OpenGWAS, which covered a total of about 9,851,867 SNPs in both male and female individuals [22]. Multimedia Appendix 1 presents the sources of these outcome data. The participants were all European populations to reduce bias caused by population heterogeneity. In reverse MR analysis, Multimedia Appendix 5 shows that 31 independently significant SNPs located on the autosomes as IVs of MCP.

eQTL Data

The Genotype-Tissue Expression (GTEx) project was established to characterize the genetic influence on transcriptomic variation across human tissues and to link regulatory mechanisms to traits and diseases. GTEx v8 provides 15,201 RNA-sequencing samples from 49 tissues [23], collected from 838 postmortem donors. The dataset comprehensively characterizes cis- and trans-eQTLs, revealing regulatory associations for nearly all genes and offering insights into allele heterogeneity, pleiotropy, and tissue-specific genetic effects. For SMR analysis, we used 8 eQTL summary data from GTEx (v8): Adipose Subcutaneous, Adipose Visceral Omentum, Brain Cerebellum, Brain Cortex, Brain Spinal cord cervical c-1, Muscle Skeletal, Nerve Tibial, and Whole Blood [23]. This is a set of cis-eQTL summary data for 8 human tissues from GTEx v8.

For colocalization analysis, we additionally used eQTL data from the eQTLGen Consortium [24]. eQTLGen provides a larger scale of eQTL data than GTEx, which can improve the confidence of colocalization analysis.

Statistical Analysis

2-Sample MR Analysis

2-sample MR analysis was conducted for exposure and outcome. The strength of IVs was assessed using the F-statistics: $F = R^2 \times (N-2)/(1-R^2)$, where R^2 represents the proportion of variation in the exposure variable explained by IVs [21]. The calculation of R^2 involved multiplying beta (the estimated genetic association of each SNP with the trait) by minor allele frequency (MAF), using the formula $2 \times \text{MAF}\,(1-\text{MAF})\,\beta^2$. An F-statistic >10 indicates a robust instrument [21]. IVs with F-statistics below 10 were considered weak instruments and excluded from analysis. We applied inverse variance weighted (IVW) regression as the primary method to estimate the effects of screen time on chronic pain [25]. In order to assess the robustness of the results, we also performed different sensitivity analyses. First, we complement the IVW MR with the weighted median, MR-Egger methods. The weighted median method can robustly estimate the causal relationship even when less than 50% of the genetic variants are invalid IVs [26]. The MR-Egger method includes an intercept term to account for directional pleiotropy [27]. Pleiotropy means that SNPs not only affect outcomes through exposures but also directly influence outcomes through other independent pathways. However, the slope of the MR-Egger regression provides valid MR estimates in the presence of horizontal pleiotropy [27,28]. It should be noted that when the number of SNPs is small, the statistical ability of MR-Egger is limited, making its results less reliable [27]. Second, we used Mendelian Randomization Pleiotropy Residual Sum and Outlier (MR-PRESSO) to detect the presence of outliers and to reassess the effect after removing the detected outlier SNPs [29]. Finally, we performed a leave-one-out analysis to reevaluate the IVW effect by excluding SNPs one by one [30]. To investigate potential bias due to sample overlap between screen time and chronic pain, we further conducted the MRlap method [31]. We examined the $P_{difference}$ of MRlap results, which is the P value used to test for differences between the observed (uncorrected) and corrected effects. We used the Bonferroni correction to account for multiple tests in the IVW result. An association was considered significant if the P value in the primary analysis was below 2.778×10^{-3} (.05/18) and the direction of effect estimates remained consistent across all methods. Correspondingly, suggestive evidence was considered if the P value for the IVW result was between .003 and .05.

To ensure methodological rigor and robustness, we also conducted a reverse MR, which is similar to the approach of 2-sample MR.

MVMR Analysis

To clarify the independent effects of each type of screen time on chronic pain, we performed MVMR analysis. The MVMR takes into account the combined effects of multiple related exposures, thereby adjusting for potential confounders and providing a more precise estimate of the independent contribution of each exposure. In the MVMR analysis, 3 phenotypes related to screen time exposures (time spent watching television, time spent using computer, and length of mobile phone use) were used as common exposure variables, and SNPs with $P<5\times10^{-8}$ genomic significance were used as IVs. At the same time, SNPs in strong LD with exposure variables were removed through LD clustering (r^2=0.001, kb=10,000). Next, we merged the SNPs of 3 exposures and deleted the duplicate SNPs. To ensure validity, only SNPs present in the result data were retained and any SNPs that did not match were excluded. The IVW method was used for main effect estimation. We evaluated the strength of the IVs using the conditional F-statistics, which measure the strength of each exposure conditional on the others in the model. If the conditional F-statistics for all exposures exceeded the rule-of-thumb value of 10, the IVs were considered adequately strong for the purposes of MVMR [32].

SMR Analysis

We performed SMR analysis to explore whether gene expression is causally linked to both screen time and chronic pain [20], thereby identifying potential functional genes involved in the

XSL•FO
RenderX

relationship. SMR integrates GWAS and eQTL data, enabling the investigation of associations between gene expression and phenotypic traits. In the SMR analysis, exposures and outcomes were analyzed separately with each eQTL summary data to assess whether gene expression is linked to both screen time and chronic pain. During the analysis, SNPs significantly associated with gene expression in cis-eQTL summary data were used as IVs, as they serve as genetic proxies for gene expression levels. These SNPs were selected based on their genome-wide significant association with gene expression in eQTL summary data. False discovery rate (FDR) correction was applied to control the FDR of multiple tests. Additionally, to verify whether the observed associations could be caused by a single causal variable, we evaluated the association results using the heterogeneity in dependent instruments (HEIDI) test implemented in the SMR tool, retaining only probes with P_{HEIDI} values indicating low heterogeneity [33]. Ultimately, the SMR analysis results for screen time and chronic pain were integrated to identify overlapping genes and further explore their potential roles in the association between screen time and chronic pain.

TWAS Analysis

We conducted TWAS to identify genes whose expression may mediate the relationship between screen time and chronic pain. GWAS datasets were obtained from IEU OpenGWAS, and using the FUSION tool developed by Gusev et al [34], we performed single-trait TWAS using the 3 cross-tissue weights for cross-tissue features generated through sparse canonical correlation analysis (sCCA) on GTEx v8 gene expression (including sCCA1, sCCA2, and sCCA3) to identify regulatory genes that may be involved in chronic pain pathways [19]. To control for type I error rates, Bonferroni correction was applied to each exposure and result to interpret multiple tests, setting the significance level for each trait or tissue to P=.05/number of genes (adjusted for the number of genes on each chromosome in the 3 GTEx v8 weights). We then extracted significant GWAS of exposure and outcome, performed joint and conditional tests on loci with multiple related features to assess whether loci contain signals independent of expression. Finally, we sought overlapping genes between exposures and outcomes in joint tests.

Colocalization Analysis

To further verify the reliability of the SMR analysis results and determine whether screen time and pain are driven by the same genetic variation, we used colocalization analysis. First, we determined all independent signals of GWAS across the genome through LD independence analysis. Genes within a 1000-kb window of each independent locus were subjected to colocalization analysis [35]. Within these windows, we combined the corresponding genes with eQTL data, which is cis-eQTL data in SMR format, and then conducted colocalization analysis. We performed colocalization analysis using the default priors of $P_1=1\times10^{-4}$, $P_2=1\times10^{-4}$, and $P_{12}=1\times10^{-5}$ [36]. Colocalization analysis assesses whether 2 traits may be driven by the same causal variant by estimating the joint posterior probability (PP) of GWAS and eQTL signals at the same locus. This method assumes a maximum of one causal variant per trait in a gene region and uses approximate Bayesian factor calculations to derive the PP for 4 mutually exclusive hypotheses (H_0-H_4), representing all possible association configurations between 2 traits: (1) H_0: neither trait has a genetic association in the region; (2) H_1 or H_2: only trait 1 or trait 2 has a genetic association in the region; (3) H_3: both traits are associated, but with different causal variants; and (4) H_4: both traits are associated and share a single causal variant. The PP of each configuration is denoted as PPH_0, PPH_1, PPH_2, PPH_3, and PPH_4, respectively [37]. We used PPH_4 to characterize the possibility of colocalization. The probability value of .5< PPH_b <.8 suggests a moderate support for colocalization, whereas PPH_4≥.8 indicates a strong support for colocalization, indicating that the 2 signals share a causal variant at this locus [38]. Using eQTL data, colocalization analysis was conducted separately for exposure and outcome traits, and genes showing evidence of colocalization were subsequently overlapped at the gene level.

Finally, to ensure the robustness of the analysis results, we further screened and consolidated the results of 3 analysis methods. We compared the colocalized genes with the overlapping genes identified in SMR and TWAS analysis to enhance the understanding of the potential biological mechanisms of these genes in the relationship between screen time and chronic pain.

Ethical Considerations

All data used in this study were deidentified publicly available data; therefore, no ethical approval was required for this study. All original studies received ethical approval from their respective institutional review boards, and all participants provided informed consent. The data used were anonymized to ensure privacy and confidentiality. No compensation was provided to participants. Additionally, this study does not include any identifiable images or figures.

Results

The Putative Association Between Screen Time and Chronic Pain

The F-statistics of IVs are all greater than 10 in Multimedia Appendices 2-4, indicating that the IVs are relatively strong. We used a 2-sample MR method to make causal inference, primarily relying on the IVW method. Figure 2 shows that the 2-sample MR results of screen time and chronic pain based on the IVW. The box in Figure 2 indicates the point estimate of the causal effect, and the error bars represent the 95% CI. Multimedia Appendix 6 showed that time spent watching television and length of mobile phone use were positively associated with chronic pain, while time spent using computer was negatively associated. Our IVW results suggested significant positive associations of time spent watching television with MCP (P=1.05×10^{-31}; odds ratio [OR] 1.61, 95% CI 1.49-1.74), back pain (P=2.41×10^{-8}; OR 1.14, 95% CI 1.09-1.19), knee pain (P=7.10×10^{-6}; OR 1.09, 95% CI 1.05-1.13), and neck or shoulder pain (P=1.18×10^{-5}; OR 1.12, 95% CI 1.06-1.17). The association between time spent watching

XSL•FO
RenderX

television and hip pain (P=.03; OR 1.06, 95% CI 1-1.11) reached nominal significance. It is notable that even though there is horizontal pleiotropy between time spent watching television and back pain ($P_{MR\text{-}Egger\ intercept}$=.02), it is significant after MR-Egger correction ($P_{MR\text{-}Egger\ slope}$=4.20×10^{-4}). Similarly, our IVW results indicated significant positive associations of length of mobile phone use with MCP (P=2.15×10^{-5}; OR 1.22, 95% CI 1.11-1.34), headaches (P=.003; OR 1.08, 95% CI 1.03-1.13), and neck or shoulder pain (P<.001; OR 1.08, 95% CI 1.03-1.14). However, IVW results suggested significant negative associations of time spent using computer with MCP (P<.001; OR 0.83, 95% CI 0.75-0.92), and knee pain (P=3.20×10^{-5}; OR 0.92, 95% CI 0.88-0.96). The association between time spent using computer and back pain (P=.01; OR 0.94, 95% CI 0.89-0.98) reached nominal significance. For the significant associations, concordant estimates were basically suggested by weighted median, MR-Egger, MR-PRESSO, and MRlap.

Figure 2. Mendelian randomization analysis of screen time and chronic pain based on the inverse variance weighted. MCP: multisite chronic pain; OR: odds ratio; TV: television.

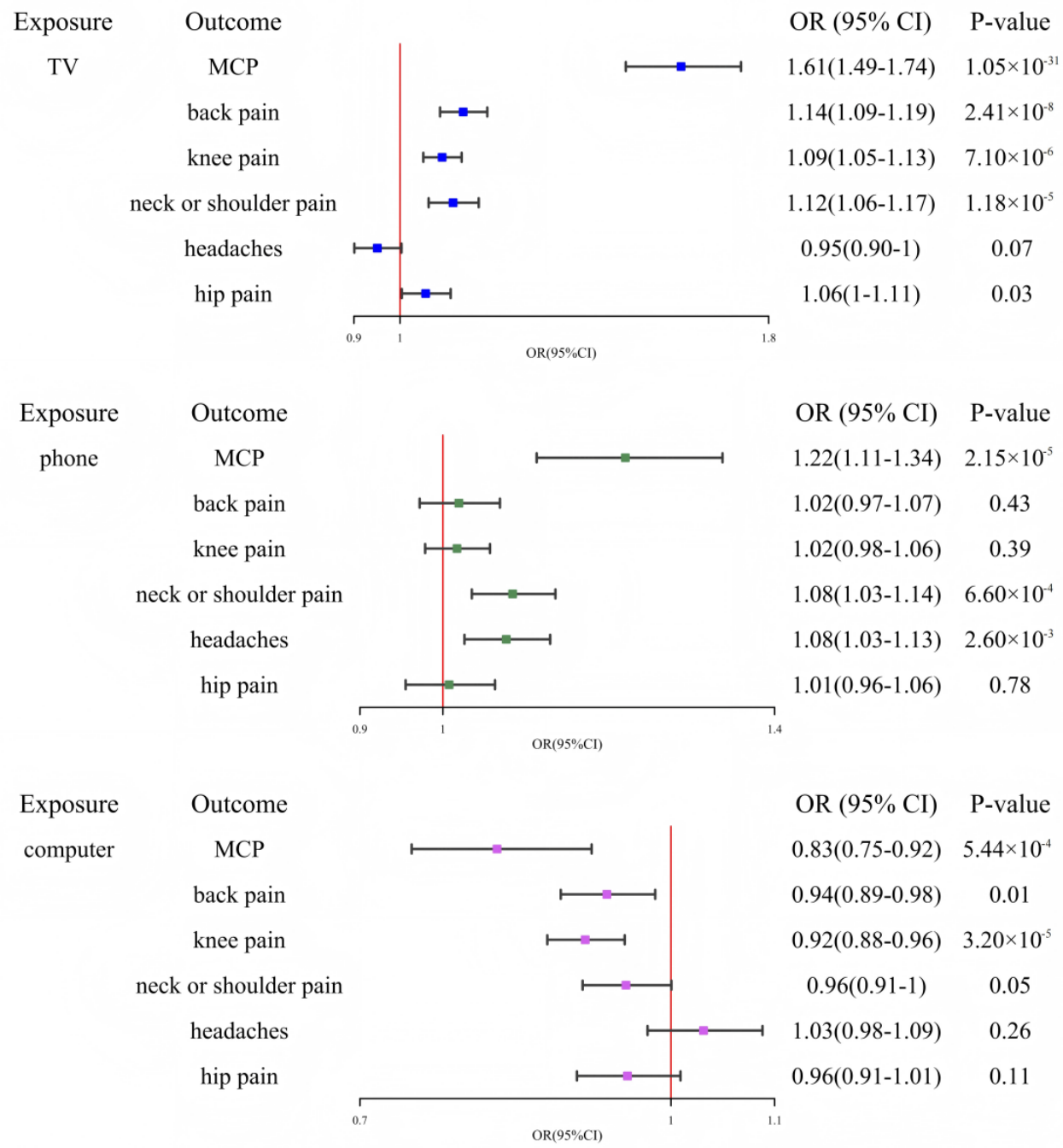

MRlap-corrected results in Multimedia Appendix 6 showed that time spent watching television significantly increased the risk of MCP ($P_{corrected}$=3.57×10^{-30}; $OR_{corrected}$=1.53, 95% CI 1.43-1.65), back pain ($P_{corrected}$=8.46×10^{-6}; $OR_{corrected}$=1.27, 95% CI 1.14-1.41), knee pain ($P_{corrected}$=1.14×10^{-4}; $OR_{corrected}$=1.22, 95% CI 1.11-1.36), and neck or shoulder pain

XSL•FO
RenderX

($P_{corrected}$=6.99×10^{-4}; $OR_{corrected}$=1.19, 95% CI 1.08-1.31). Length of mobile phone use significantly increased the risk of MCP ($P_{corrected}$=1.04×10^{-4}; $OR_{corrected}$=1.39, 95% CI 1.18-1.64), headaches ($P_{corrected}$=.03; $OR_{corrected}$=1.34, 95% CI 1.04-1.74), and neck or shoulder pain ($P_{corrected}$=.01; $OR_{corrected}$=1.55, 95% CI 1.1-2.17). In contrast, time spent using computer was negatively associated with the risk of MCP ($P_{corrected}$=1.65×10^{-3}; $OR_{corrected}$=0.81, 95% CI 0.71-0.92), back pain ($P_{corrected}$=4.28×10^{-3}; $OR_{corrected}$=0.83, 95% CI 0.73-0.94), and knee pain ($P_{corrected}$=3.48×10^{-3}; $OR_{corrected}$=0.83, 95% CI 0.73-0.94). Overall, the MRlap-corrected causal estimates were consistent in direction and statistical significance with the primary IVW results in Multimedia Appendix 6, supporting the robustness of the findings against biases induced by sample overlap. Leave-one-out analyses in Multimedia Appendix 7 showed no outlying SNPs.

In the reverse MR analysis, MCP was used as the exposure, and different types of screen time were treated as outcomes. The F-statistics of IVs of MCP in Multimedia Appendix 5 are all greater than 10, indicating that the IVs are relatively strong. The results in Multimedia Appendix 8 indicated that MCP was positively associated with time spent watching television (P=4.8×10^{-7}; OR 1.27, 95% CI 1.16-1.4) and length of mobile phone use (P=3.38×10^{-5}; OR 1.29, 95% CI 1.14-1.45), while the association with time spent using computer (P=.61; OR 0.97, 95% CI 0.87-1.09) was not statistically significant. For other exposures, the analysis could not be conducted due to the limited number of available IVs or F-statistics below the conventional threshold of 10.

Independent Effect of Screen Time on Chronic Pain

Multimedia Appendix 9 showed that the causal effects of screen time on chronic pain based on IVW MVMR. The conditional F-statistics for both time spent using computer and length of mobile phone use is less than 10 in the MVMR results.

Discovery of Screen Time and Chronic Pain Genes Based on SMR Analysis

As the main analysis result, SMR was corrected by FDR and combined with the HEIDI test to screen for overlapping genes between exposure and outcome in 8 eQTL summary data. Multimedia Appendices 10 and 11 provided the full results, including the exact P_{FDR} and P_{HEIDI} values for all tested genes after FDR correction. A total of 72 overlapping genes related to exposure and outcome were identified in Multimedia Appendix 12, and the results showed that these overlapping genes originated from 3 exposures and MCP: 40 genes for time spent watching television and MCP, 26 genes for time spent using computer and MCP, and 6 genes for length of mobile phone use and MCP. In the SMR analysis, the Adipose Subcutaneous tissue had the most genes, while only one gene was found in the Brain Spinal cord cervical c-1.

Genetic Findings of TWAS-Based Screen Time and Chronic Pain

After Bonferroni correction, we examined genome-wide significant associations in Multimedia Appendices 13 and 14, followed by joint and conditional tests. Multimedia Appendices 15 and 16 provided the results of the jointly significant genes. Then we identified 28 genes with overlapping exposures and outcomes in Table 1. No genes overlapping with the exposure were found in the hip pain for ≥3 months GWAS dataset.

XSL•FO
RenderX

Table 1. Transcriptome-wide significant genes identified by TWAS[a] in screen time and chronic pain.

Trait	Cross-tissue expression	Gene (chromosome)
Time spent watching television – back pain for ≥3 months	sCCA1	• *PMS2P3* (7)
Time spent watching television – MCP[b]	sCCA2	• *FUBP1* (1) • *WDR47* (1) • *PLEKHO1* (1) • *FAM172A* (5)
Time spent watching television – MCP	sCCA3	• *SUSD3* (9) • *RPL35* (9) • *RELA* (11) • *RMC1* (18) • *SYPL2* (1) • *SF3B4* (1) • *ARPC5L* (9) • *PTPDC1* (9) • *FAM53B* (10)
Time spent watching television – knee pain for ≥3 months	sCCA2	• *GATC* (12)
Length of mobile phone use – neck or shoulder pain for ≥3 months	sCCA3	• *RBM42* (19)
Length of mobile phone use – MCP	sCCA3	• *CSTPP1* (11) • *TM9SF4* (20)
Time spent using computer – MCP	sCCA2	• *WDR47* (1) • *RFTN2* (2) • *NMT1* (17) • *RMC1* (18) • *FASTKD5* (20) • *TM9SF4* (20)
Time spent using computer – MCP	sCCA3	• *SYPL2* (1) • *CEP170* (1) • *SCOC-AS1* (4)
Time spent using computer – headaches for ≥3 months	sCCA2	• *SHMT2* (12)

[a]TWAS: transcriptome-wide association study.

[b]MCP: multisite chronic pain.

By combining the results of SMR and TWAS analysis, we found that 7 overlapping genes (*SYPL2*, *RMC1*, *FUBP1*, *ARPC5L*, *RFTN2*, *NMT1*, and *CEP170*) were identified between exposures (time spent watching television and time spent using computer) and outcome (MCP) in Multimedia Appendix 17 and Figure 3. In addition, we found that *SYPL2* gene expression was associated with both time spent watching television and time spent using computer.

XSL•FO
RenderX

Figure 3. Results supported by genetic evidence. SMR: summary data–based Mendelian randomization analysis; TWAS: transcriptome-wide association study.

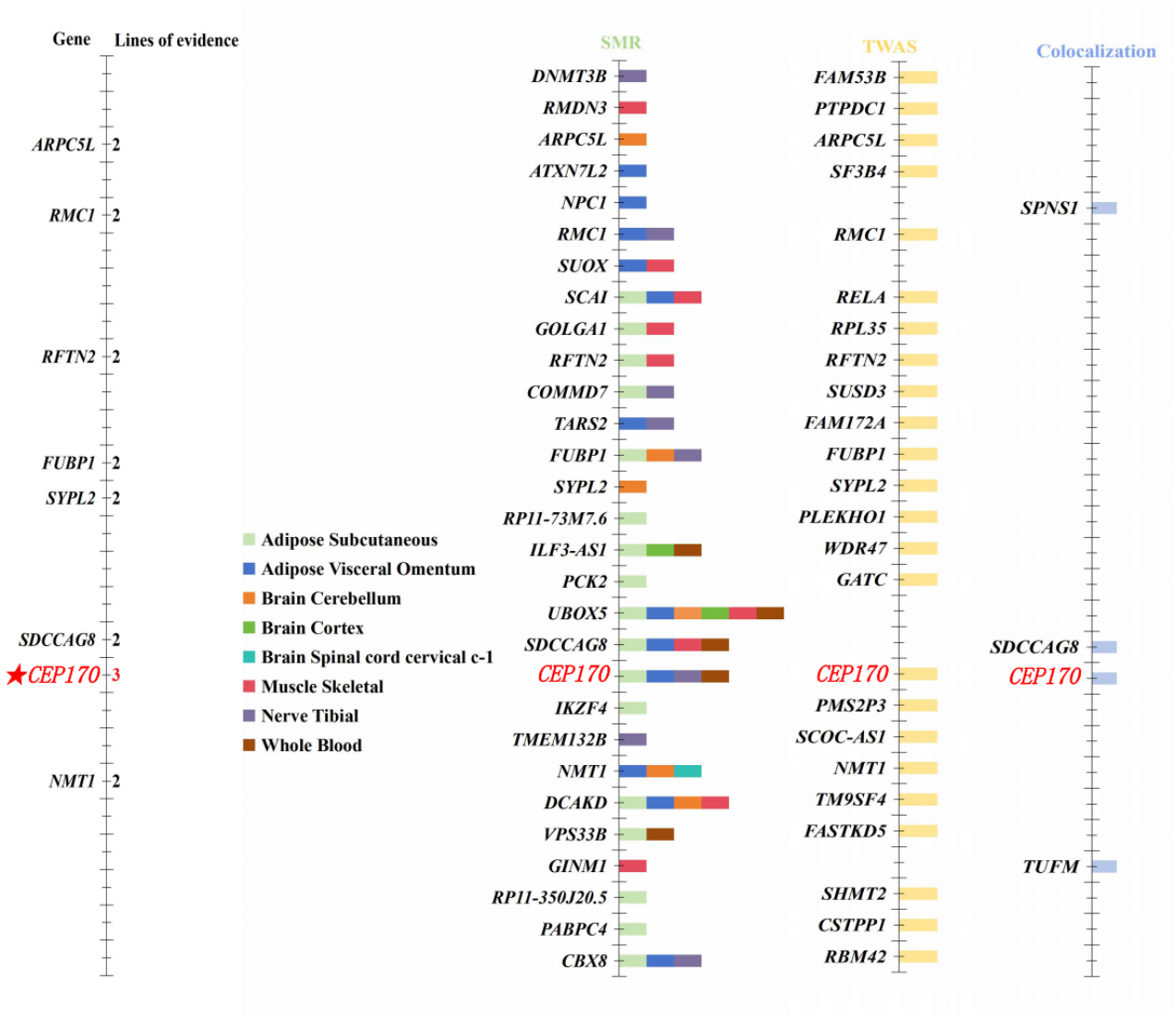

In Figure 3, the panel on the far left shows genes supported by at least 2 lines of evidence. A total of 3 lines of evidence were used to support genes, with each column representing one type of supporting evidence. The SMR evidence shows different colors according to 8 different eQTL summary data. TWAS evidence is shown in light yellow and colocalization evidence is shown in light blue. *CEP170* was supported by all analyses. *ARPC5L*, *RMC1*, *RFTN2*, *FUBP1*, *SYPL2*, *SDCCAG8*, and *NMT1* were supported by 2 lines of evidence.

Colocalization Analysis of Shared Genetic Variation

Colocalization analysis identified multiple genes with moderate or strong support for colocalization signals in the shared genetic variation between screen time and chronic pain in Multimedia Appendix 18. The results showed that 4 genes suggest a moderate support for colocalization (*SPNS1*, *TUFM*, *SDCCAG8*, and *CEP170*). Multimedia Appendix 19 showed that the 4 genes are overlapped between screen time and chronic pain.

Gene Discovery

As shown in Figure 3, we presented the distribution of significant genes in each analysis method, and the results showed that the *CEP170* gene stood out in the association analysis of "Time spent using computer" and MCP, receiving triple support from SMR, TWAS, and colocalization analysis (PPH_4>.75). The signaling distribution of this gene involved 4 eQTL tissues: Adipose Subcutaneous, Adipose Visceral Omentum, Nerve Tibial, and Whole Blood.

Discussion

Principal Findings

This study delves into potential phenotypic associations and underlying genetic connections between screen time and chronic pain. The results show that screen time is associated with chronic pain. In the 2-sample MR results, time spent watching television and length of mobile phone use were significantly and positively associated with chronic pain, including MCP, back pain, knee pain, neck or shoulder pain, hip pain, and headaches. In contrast, time spent using computer inversely was negatively associated with MCP, back pain, and knee pain. It is particularly important to note that we still list the significant IVW results of exposure to the length of mobile phone use. Even though the effect value directions of the IVW and MR-Egger results are not consistent,

XSL•FO
RenderX

it is consistent with the direction of our main analytical method in other methods. The statistical power of MR-Egger decreases for fewer IVs [27]. Therefore, we still take the results of IVW as the standard for the length of mobile phone use. In the reverse MR results, we found that there is a bidirectional positive correlation among time spent watching television, length of mobile phone use, and MCP. In the MVMR analysis, not all the conditional F-statistics were >10, suggesting the presence of weak instruments. Therefore, our interpretations were based on the 2-sample MR results. However, the inverse association between time spent using computer and MCP should be interpreted cautiously, as it stems from the 2-sample MR results. The lack of robust MVMR findings means we cannot rule out the potential influence of confounding factors. In addition, by integrating SMR, TWAS, and colocalization analysis, we identified an overlapping gene *CEP170* associated with time spent using computer and MCP. These findings highlight the important role of screen time in the development of chronic pain.

Differential Associations Between Screen Time and Chronic Pain

Previous studies have shown a positive correlation between mobile phone use and neck pain [39], and a strong correlation between watching television and back pain [40], which is consistent with our research findings. In contrast, our research found that time spent using computer may have a protective effect on certain types of chronic pain, which differs from some studies that concluded, “longer computer time increases the risk of multi-site pain” [8,39]. This difference may be due to the study population, with earlier studies focusing on adolescents, while our analysis was conducted on adults. Due to continued musculoskeletal development, adolescents may be more susceptible to the negative physical effects of sedentary behavior.

Furthermore, 2 potential explanations may underlie the differential associations between time spent using computer and chronic pain, although there is no direct evidence available at present. One possible explanation is that the perception of pain is influenced by the allocation of cognitive resources; that is, when attention is occupied by other highly attractive stimuli, the perception of pain may be reduced [41,42]. According to the characteristics of the UK Biobank population (aged 40-69 years during 2006-2010) in this study [43], the use of mobile phones and televisions was often associated with activities that had relatively low cognitive requirements and less attention input at that time. In contrast, even leisure-time computer use generally involved more cognitively demanding activities among middle-aged adults, such as reading, writing, online communication, or other interactive tasks. Research indicates that pain is less likely to enter conscious awareness when cognitive resources are occupied by goal-directed information maintained in working memory and when sustained attention is devoted to the task [44]. This indicates that computer tasks with high cognitive load may potentially reduce the pain experience through distraction. Moreover, it is also important to consider that individuals who frequently use computers tend to have higher educational attainment, higher income, and better overall health [45]. Given that individuals in nonmanual occupations tend to have a lower likelihood of experiencing chronic pain compared with those in manual labor occupations [46,47], part of the observed lower risk associated with time spent using computer may therefore reflect underlying socioeconomic or occupational factors rather than a sole reflection of the nature of computer-based activities themselves. This may explain why length of mobile phone use and time spent watching television were associated with a higher risk of pain, while time spent using computer showed a protective effect in this study.

Another possible explanation is that different types of screen time may affect pain perception by engaging the brain’s reward system to varying degrees. Dopamine is involved in pleasure, reward, and incentive behavior [48]. Although mobile phone and television use may induce pleasure, the limited interactivity prevents them from triggering dopamine release to the same extent as engaging in more goal-directed and cognitively engaging computer activities. Successfully completing such goals may elicit a sense of accomplishment. The reward system may be activated, promoting dopamine release, which has been proven to have analgesic effects [49,50]. Moreover, bidirectional association analysis suggests that MCP may influence the choice of screen behavior. People with MCP tend to increase passive and low-effort screen behaviors (such as watching television or casually browsing their mobile phones). This behavioral adaptation may contribute to a pattern of bidirectional reinforcement, where an increase in passive activity further restricts physical movement and may exacerbate pain. However, we emphasize that these explanations are speculative, as direct evidence is currently lacking.

The Potential Mechanisms by Which CEP170 Contributes to Screen Time and Chronic Pain

Leveraging multiomics approaches, we identified *CEP170*, whose cis-regulated expression may contribute to the biological mechanisms between time spent using computer and MCP. In SMR results, the *CEP170* gene expression level was positively correlated with MCP but negatively correlated with time spent using computer. This aligns with our MR results, where time spent using computer was associated with a lower risk of MCP. One possible explanation is that *CEP170* may influence behavioral patterns related to screen time, where individuals with higher gene expression tend to spend less time on the computer, which in turn is associated with increased MCP risk. Although its colocalization results suggest a moderate support for colocalization, *CEP170* was detected in all analytical methods.

CEP170 is located at the centrosome and spindle microtubules and participates in microtubule organization and assembly [51,52]. Additionally, *CEP170* plays an important role in supporting ciliary homeostasis [53]. Cilia may potentially play a central role in cell signaling, and recent animal experiments have further revealed that cilia have key regulatory functions in controlling mechanical nociceptive thresholds and inflammatory and neuropathic pain [54]. The SMR results show that *CEP170* exhibited significant signals in multiple eQTL tissues, including Adipose Subcutaneous, Adipose Visceral Omentum, Nerve Tibial, and Whole Blood. This indicates that

XSL•FO
RenderX

CEP170 may be expressed in these tissues and may contribute to the biological mechanisms behind the observed associations. A large UK Biobank study revealed that abdominal adipose tissue (including visceral fat and subcutaneous fat) was associated with chronic musculoskeletal pain. It suggested that excessive and ectopic fat depositions may be involved in the pathogenesis of multisite and widespread chronic musculoskeletal pain [55]. This is likely attributable to the chronic low-grade inflammation driven by accumulated visceral fat, which can sensitize peripheral nerves and promote MCP [56,57]. In line with this, our SMR analysis found that *CEP170* expression in subcutaneous and visceral adipose tissue is associated with MCP. Given the established importance of microtubules in lipid metabolic homeostasis [58,59], we hypothesize that *CEP170* may potentially protect against MCP by enhancing microtubule stability in adipose tissue. This mechanism might potentially constrain adipose-derived inflammatory signaling, thereby possibly alleviating peripheral nerve sensitization and lowering MCP risk. However, the current results only provide preliminary genetic clues. Direct links remain unclear and more functional studies, especially tissue-specific experiments, are needed in the future to clarify the potential contribution of *CEP170* to pain pathways.

Strengths and Limitations

Our research has several advantages. First, we integrated a variety of analytical methods to duplicate our findings and strengthen the robustness of our results. Second, we leveraged large-scale GWAS data from the UK Biobank, ensuring a robust analysis with broad generalizability. Third, by identifying overlapping genes and potential biological pathways, our study provides new genetic insights into the association between screen time and chronic pain, which, despite limited evidence, may inform future research in this field. Finally, our study considered different types of screen time separately, which allows for a more nuanced understanding of their different effects on chronic pain. These findings may serve as a useful reference for subsequent studies on the genetic and epidemiological links between screen time and chronic pain.

Although this study provides new insights, there are still some limitations. First, using data from the same sources in multiple analyses may introduce the bias of the winner's curse, potentially inflating effect sizes or significance levels. Replication in fully independent datasets is necessary to confirm the discovery. Second, our analyses were restricted to European populations, which may limit generalizability to other populations with different screen use habits and pain reporting. Third, the lack of detailed data on screen use posture, activity type, and behavior patterns may limit the depth of causal reasoning. Future studies could incorporate longitudinal and behavioral data to better clarify the mechanisms and causal pathways. Fourth, as both exposures and outcomes are self-reported, recall and measurement bias may lead to misclassification and reduce precision. Self-reported pain may not capture clinical heterogeneity, so future studies using clinically validated phenotypes are needed. Fifth, due to the weak strength of IVs, we were unable to adequately control for crucial potential confounding factors, such as socioeconomic status and educational attainment. Consequently, the observed protective association between "time spent using computer" and chronic pain might be partially attributable to residual confounding by socioeconomic status and related factors. Sixth, reverse MR could not be performed for the outcomes of back pain, knee pain, neck or shoulder pain, headaches, and hip pain due to limited IVs, limiting the assessment of reverse causality. In addition, the smaller sample size for hip pain may have limited statistical power, contributing to borderline significance in the 2-sample MR analysis and precluding reverse MR for this phenotype. Seventh, environmental exposure, such as air pollution, was not taken into account in this analysis. Particulate matter can affect inflammatory and stress-related biomarkers involved in chronic pain [60-62]. Future studies should integrate genetic and environmental data to clarify these relationships. Eighth, the colocalization signal for *CEP170* is not the strongest, and direct evidence is lacking. Therefore, its role remains speculative and warrants further investigation. Finally, this study lacks validation in large surveys or independent cohorts, highlighting the need to combine genetic and epidemiological studies in future research.

Conclusions

This study provides evidence that reveals an association between screen time and chronic pain. We found the *CEP170* gene might contribute to the shared biological mechanism between time spent using computer and MCP risk. Future studies should further validate this association and clarify the functional role of *CEP170* in the development of chronic pain.

Acknowledgments

The authors would like to thank the participants and investigators involved in the GWASs used in this study.

Funding

This work was supported by the Hainan Provincial People's Hospital Scientific Research Start-up Fund (grant number YJRC2025009).

Data Availability

The code used in this study is available from the corresponding authors on reasonable request. The GWAS summary statistics data used in this study are available from IEU OpenGwas and publicly available GWAS. The eQTL summary data for GTEx used in the SMR analysis are available in the public database [63]. Precomputed sCCA weights for GTEx v8 used in this study

XSL•FO
RenderX

are available in the public database [64]. The eQTL data used in the colocalization analysis are available from the eQTLGen Consortium [65].

Authors' Contributions

CX and ZL contributed equally as corresponding authors. JJ, CP, and JC contributed equally to this work and share first authorship. JJ and CX conceived the study. JJ and CP performed the data analysis and interpretation. JJ and CP wrote the manuscript. JC, CP, and JJ critically revised the manuscript. CY, CP, JC, and CX provide additional methodological guidance and technical support. CX and ZL supervised the study and were responsible for overall project administration. All authors reviewed and approved the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data source of the exposures and outcomes.
[XLSX File (Microsoft Excel File), 12 KB - jmir_v28i1e78233_app1.xlsx]

Multimedia Appendix 2
Characteristics of genetic instruments of Time spent watching television and their effect sizes with chronic pain.
[XLSX File (Microsoft Excel File), 45 KB - jmir_v28i1e78233_app2.xlsx]

Multimedia Appendix 3
Characteristics of genetic instruments of Time spent using computer and their effect sizes with chronic pain.
[XLSX File (Microsoft Excel File), 36 KB - jmir_v28i1e78233_app3.xlsx]

Multimedia Appendix 4
Characteristics of genetic instruments of Length of mobile phone use and their effect sizes with chronic pain.
[XLSX File (Microsoft Excel File), 21 KB - jmir_v28i1e78233_app4.xlsx]

Multimedia Appendix 5
Characteristics of genetic instruments of multisite chronic pain.
[XLSX File (Microsoft Excel File), 15 KB - jmir_v28i1e78233_app5.xlsx]

Multimedia Appendix 6
Putative causal effect of screen time on chronic pain from 2-sample Mendelian randomization analysis.
[XLSX File (Microsoft Excel File), 24 KB - jmir_v28i1e78233_app6.xlsx]

Multimedia Appendix 7
The leave-one-out plot of screen time on chronic pain.
[DOCX File , 365 KB - jmir_v28i1e78233_app7.docx]

Multimedia Appendix 8
Reverse Mendelian randomization analysis assessing the causal effect of chronic pain on screen time.
[XLSX File (Microsoft Excel File), 14 KB - jmir_v28i1e78233_app8.xlsx]

Multimedia Appendix 9
Causal effects of the screen time on chronic pain based on inverse variance weighted multivariable Mendelian randomization.
[XLSX File (Microsoft Excel File), 13 KB - jmir_v28i1e78233_app9.xlsx]

Multimedia Appendix 10
The top hit probes identified in summary data–based Mendelian randomization analysis after false discovery rate correction of screen time.
[XLSX File (Microsoft Excel File), 533 KB - jmir_v28i1e78233_app10.xlsx]

Multimedia Appendix 11

XSL•FO
RenderX

The top hit probes identified in summary data–based Mendelian randomization analysis after false discovery rate correction of chronic pain.
[XLSX File (Microsoft Excel File), 108 KB - jmir_v28i1e78233_app11.xlsx]

Multimedia Appendix 12
The genes with overlapping exposure and results were screened by PFDR and PHEIDI in summary data–based Mendelian randomization analysis.
[XLSX File (Microsoft Excel File), 13 KB - jmir_v28i1e78233_app12.xlsx]

Multimedia Appendix 13
Transcriptome-wide association study of screen time in three cross-tissue (sparse canonical correlation analysis) weights using Bonferroni correction and cross-tissue expression Genotype-Tissue Expression v8.
[XLSX File (Microsoft Excel File), 269 KB - jmir_v28i1e78233_app13.xlsx]

Multimedia Appendix 14
Transcriptome-wide association study of chronic pain in three cross-tissue (sparse canonical correlation analysis) weights using Bonferroni correction and cross-tissue expression Genotype-Tissue Expression v8.
[XLSX File (Microsoft Excel File), 74 KB - jmir_v28i1e78233_app14.xlsx]

Multimedia Appendix 15
The results of the jointly significant genes for screen time.
[XLSX File (Microsoft Excel File), 73 KB - jmir_v28i1e78233_app15.xlsx]

Multimedia Appendix 16
The results of the jointly significant genes for chronic pain.
[XLSX File (Microsoft Excel File), 23 KB - jmir_v28i1e78233_app16.xlsx]

Multimedia Appendix 17
Overlapping genes by both transcriptome-wide association study and summary data–based Mendelian randomization methods.
[XLSX File (Microsoft Excel File), 10 KB - jmir_v28i1e78233_app17.xlsx]

Multimedia Appendix 18
Results of the colocalization analysis for genes identified in the exposure and outcome traits.
[XLSX File (Microsoft Excel File), 41 KB - jmir_v28i1e78233_app18.xlsx]

Multimedia Appendix 19
Overlapping genes in colocalization.
[XLSX File (Microsoft Excel File), 10 KB - jmir_v28i1e78233_app19.xlsx]

References

1. Pandya A, Lodha P. Social connectedness, excessive screen time during COVID-19 and mental health: a review of current evidence. Front Hum Dyn 2021;3. [doi: 10.3389/fhumd.2021.684137]
2. Van Dyck D, Cardon G, Deforche B, Owen N, De Cocker K, Wijndaele K, et al. Socio-demographic, psychosocial and home-environmental attributes associated with adults' domestic screen time. BMC Public Health 2011 Aug 25;11:668 [FREE Full text] [doi: 10.1186/1471-2458-11-668] [Medline: 21864412]
3. Goldfield GS, Murray M, Maras D, Wilson AL, Phillips P, Kenny GP, et al. Screen time is associated with depressive symptomatology among obese adolescents: a HEARTY study. Eur J Pediatr 2016;175(7):909-919. [doi: 10.1007/s00431-016-2720-z] [Medline: 27075014]
4. Nightingale CM, Rudnicka AR, Donin AS, Sattar N, Cook DG, Whincup PH, et al. Screen time is associated with adiposity and insulin resistance in children. Arch Dis Child 2017;102(7):612-616 [FREE Full text] [doi: 10.1136/archdischild-2016-312016] [Medline: 28288985]
5. Ahmadi-Abhari S, Sabia S, Shipley MJ, Kivimäki M, Singh-Manoux A, Tabak A, et al. Physical activity, sedentary behavior, and long-term changes in aortic stiffness: the Whitehall II study. J Am Heart Assoc 2017;6(8):e005974 [FREE Full text] [doi: 10.1161/JAHA.117.005974] [Medline: 28784651]
6. Chetty-Mhlanga S, Fuhrimann S, Eeftens M, Basera W, Hartinger S, Dalvie MA, et al. Different aspects of electronic media use, symptoms and neurocognitive outcomes of children and adolescents in the rural Western Cape region of South Africa. Environ Res 2020;184:109315. [doi: 10.1016/j.envres.2020.109315] [Medline: 32172071]

XSL•FO
RenderX

7. Marques A, Calmeiro L, Loureiro N, Frasquilho D, de Matos MG. Health complaints among adolescents: associations with more screen-based behaviours and less physical activity. J Adolesc 2015;44:150-157. [doi: 10.1016/j.adolescence.2015.07.018] [Medline: 26275746]
8. Taehtinen RE, Sigfusdottir ID, Helgason AR, Kristjansson AL. Electronic screen use and selected somatic symptoms in 10-12 year old children. Prev Med 2014;67:128-133. [doi: 10.1016/j.ypmed.2014.07.017] [Medline: 25045838]
9. Torsheim T, Eriksson L, Schnohr CW, Hansen F, Bjarnason T, Välimaa R. Screen-based activities and physical complaints among adolescents from the Nordic countries. BMC Public Health 2010;10:324 [FREE Full text] [doi: 10.1186/1471-2458-10-324] [Medline: 20534116]
10. Hakala PT, Saarni LA, Punamäki RL, Wallenius MA, Nygård CH, Rimpelä AH. Musculoskeletal symptoms and computer use among Finnish adolescents--pain intensity and inconvenience to everyday life: a cross-sectional study. BMC Musculoskelet Disord 2012;13:41 [FREE Full text] [doi: 10.1186/1471-2474-13-41] [Medline: 22439805]
11. Guerra PH, Martelo R, da Silva MN, de Andrade GF, Christofaro DGD, Loch MR. Screen time and low back pain in children and adolescents: a systematic review of Brazilian studies. Rev Paul Pediatr 2023;41:e2021342 [FREE Full text] [doi: 10.1590/1984-0462/2023/41/2021342] [Medline: 37042940]
12. Montagni I, Guichard E, Carpenet C, Tzourio C, Kurth T. Screen time exposure and reporting of headaches in young adults: A cross-sectional study. Cephalalgia 2016 Oct;36(11):1020-1027. [doi: 10.1177/0333102415620286] [Medline: 26634831]
13. Roman-Juan J, Roy R, Jensen MP, Miró J. The explanatory role of sedentary screen time and obesity in the increase of chronic back pain amongst European adolescents: the HBSC study 2002-2014. Eur J Pain 2022;26(8):1781-1789 [FREE Full text] [doi: 10.1002/ejp.2003] [Medline: 35761772]
14. Silva AG, Couto PS, Queirós A, Neto M, Rocha NP. Chronic pain in high school students is associated with physical activity and sleeping hours but not with screen time. Int J Adolesc Med Health 2017;31(3):2017. [doi: 10.1515/ijamh-2017-0014] [Medline: 28628479]
15. Henry P, Auray JP, Gaudin AF, Dartigues JF, Duru G, Lantéri-Minet M, et al. Prevalence and clinical characteristics of migraine in France. Neurology 2002;59(2):232-237. [doi: 10.1212/wnl.59.2.232] [Medline: 12136063]
16. Kowacs P, Piovesan E, Werneck L, Fameli H, Pereira da Silva H. Headache related to a specific screen flickering frequency band. Cephalalgia 2004;24(5):408-410. [doi: 10.1111/j.1468-2982.2004.00686.x] [Medline: 15096230]
17. Robbins L. Precipitating factors in migraine: a retrospective review of 494 patients. Headache 1994;34(4):214-216. [doi: 10.1111/j.1526-4610.1994.hed3404214.x] [Medline: 8014037]
18. Johnston KJA, Adams MJ, Nicholl BI, Ward J, Strawbridge RJ, Ferguson A, et al. Genome-wide association study of multisite chronic pain in UK Biobank. PLoS Genet 2019;15(6):e1008164 [FREE Full text] [doi: 10.1371/journal.pgen.1008164] [Medline: 31194737]
19. Feng H, Mancuso N, Gusev A, Majumdar A, Major M, Pasaniuc B, et al. Leveraging expression from multiple tissues using sparse canonical correlation analysis and aggregate tests improves the power of transcriptome-wide association studies. PLoS Genet 2021;17(4):e1008973 [FREE Full text] [doi: 10.1371/journal.pgen.1008973] [Medline: 33831007]
20. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet 2016;48(5):481-487. [doi: 10.1038/ng.3538] [Medline: 27019110]
21. Burgess S, Thompson SG. Avoiding bias from weak instruments in Mendelian randomization studies. Int J Epidemiol 2011;40(3):755-764. [doi: 10.1093/ije/dyr036] [Medline: 21414999]
22. Rusk N. The UK Biobank. Nat Methods 2018;15(12):1001. [doi: 10.1038/s41592-018-0245-2] [Medline: 30504882]
23. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 2020;369(6509):1318-1330 [FREE Full text] [doi: 10.1126/science.aaz1776] [Medline: 32913098]
24. Võsa U, Claringbould A, Westra HJ, Bonder MJ, Deelen P, Zeng B, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet 2021 Sep;53(9):1300-1310 [FREE Full text] [doi: 10.1038/s41588-021-00913-z] [Medline: 34475573]
25. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. Genet Epidemiol 2013;37(7):658-665 [FREE Full text] [doi: 10.1002/gepi.21758] [Medline: 24114802]
26. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. Genet Epidemiol 2016;40(4):304-314 [FREE Full text] [doi: 10.1002/gepi.21965] [Medline: 27061298]
27. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol 2015;44(2):512-525 [FREE Full text] [doi: 10.1093/ije/dyv080] [Medline: 26050253]
28. Burgess S, Thompson SG. Interpreting findings from Mendelian randomization using the MR-Egger method. Eur J Epidemiol 2017;32(5):377-389 [FREE Full text] [doi: 10.1007/s10654-017-0255-x] [Medline: 28527048]
29. Verbanck M, Chen C, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nat Genet 2018;50(5):693-698 [FREE Full text] [doi: 10.1038/s41588-018-0099-7] [Medline: 29686387]

XSL•FO
RenderX

30. Burgess S, Bowden J, Fall T, Ingelsson E, Thompson SG. Sensitivity analyses for robust causal inference from Mendelian randomization analyses with multiple genetic variants. Epidemiology 2017;28(1):30-42 [FREE Full text] [doi: 10.1097/EDE.0000000000000559] [Medline: 27749700]
31. Mounier N, Kutalik Z. Bias correction for inverse variance weighting Mendelian randomization. Genet Epidemiol 2023;47(4):314-331. [doi: 10.1002/gepi.22522] [Medline: 37036286]
32. Sanderson E, Davey Smith G, Windmeijer F, Bowden J. An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. Int J Epidemiol 2019;48(3):713-727 [FREE Full text] [doi: 10.1093/ije/dyy262] [Medline: 30535378]
33. Mo X, Guo Y, Qian Q, Fu M, Lei S, Zhang Y, et al. Mendelian randomization analysis revealed potential causal factors for systemic lupus erythematosus. Immunology 2020;159(3):279-288 [FREE Full text] [doi: 10.1111/imm.13144] [Medline: 31670388]
34. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet 2016;48(3):245-252 [FREE Full text] [doi: 10.1038/ng.3506] [Medline: 26854917]
35. GTEx Consortium, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Lead analysts:, NIH program management:, Biospecimen collection:, Pathology:, eQTL manuscript working group:, et al. Genetic effects on gene expression across human tissues. Nature 2017;550(7675):204-213 [FREE Full text] [doi: 10.1038/nature24277] [Medline: 29022597]
36. Kia DA, Zhang D, Guelfi S, Manzoni C, Hubbard L, Reynolds RH, et al. Identification of candidate parkinson disease genes by integrating genome-wide association study, expression, and epigenetic data sets. JAMA Neurol 2021;78(4):464-472 [FREE Full text] [doi: 10.1001/jamaneurol.2020.5257] [Medline: 33523105]
37. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet 2014;10(5):e1004383 [FREE Full text] [doi: 10.1371/journal.pgen.1004383] [Medline: 24830394]
38. Yi M, Feng X, Guan Q, Liu Y, Liu Y, Su Z. A multi-omics Mendelian randomization study reveals PAM as a potential therapeutic target for type 2 diabetes. J Transl Med 2025;23(1):1067 [FREE Full text] [doi: 10.1186/s12967-025-07086-x] [Medline: 41063246]
39. Cankurtaran F, Menevşe O, Namlı A, Kızıltoprak, Altay S, Duran M, et al. The impact of digital game addiction on musculoskeletal system of secondary school children. Niger J Clin Pract 2022;25(2):153-159 [FREE Full text] [doi: 10.4103/njcp.njcp_177_20] [Medline: 35170440]
40. Shehab DK, Al-Jarallah KF. Nonspecific low-back pain in Kuwaiti children and adolescents: associated factors. J Adolesc Health 2005;36(1):32-35. [doi: 10.1016/j.jadohealth.2003.12.011] [Medline: 15661594]
41. Buhle J, Wager TD. Performance-dependent inhibition of pain by an executive working memory task. Pain 2010;149(1):19-26 [FREE Full text] [doi: 10.1016/j.pain.2009.10.027] [Medline: 20129735]
42. Buhle JT, Stevens BL, Friedman JJ, Wager TD. Distraction and placebo: two separate routes to pain control. Psychol Sci 2012;23(3):246-253. [doi: 10.1177/0956797611427919] [Medline: 22261568]
43. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 2015;12(3):e1001779 [FREE Full text] [doi: 10.1371/journal.pmed.1001779] [Medline: 25826379]
44. Legrain V, Crombez G, Verhoeven K, Mouraux A. The role of working memory in the attentional control of pain. Pain 2011;152(2):453-459 [FREE Full text] [doi: 10.1016/j.pain.2010.11.024] [Medline: 21238855]
45. Czaja SJ, Charness N, Fisk AD, Hertzog C, Nair SN, Rogers WA, et al. Factors predicting the use of technology: findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE). Psychol Aging 2006 Jun;21(2):333-352 [FREE Full text] [doi: 10.1037/0882-7974.21.2.333] [Medline: 16768579]
46. Saastamoinen P, Leino-Arjas P, Laaksonen M, Lahelma E. Socio-economic differences in the prevalence of acute, chronic and disabling chronic pain among ageing employees. Pain 2005;114(3):364-371. [doi: 10.1016/j.pain.2004.12.033] [Medline: 15777862]
47. Vinstrup J, Sundstrup E, Andersen LL. Psychosocial stress and musculoskeletal pain among senior workers from nine occupational groups: cross-sectional findings from the SeniorWorkingLife study. BMJ Open 2021;11(3):e043520 [FREE Full text] [doi: 10.1136/bmjopen-2020-043520] [Medline: 33782021]
48. Chinta SJ, Andersen JK. Dopaminergic neurons. Int J Biochem Cell Biol 2005;37(5):942-946. [doi: 10.1016/j.biocel.2004.09.009] [Medline: 15743669]
49. Navratilova E, Porreca F. Reward and motivation in pain and pain relief. Nat Neurosci 2014;17(10):1304-1312 [FREE Full text] [doi: 10.1038/nn.3811] [Medline: 25254980]
50. Wood PB. Role of central dopamine in pain and analgesia. Expert Rev Neurother 2008;8(5):781-797. [doi: 10.1586/14737175.8.5.781] [Medline: 18457535]

XSL•FO
RenderX

51. Mazo G, Soplop N, Wang W, Uryu K, Tsou MB. Spatial control of primary ciliogenesis by subdistal appendages alters sensation-associated properties of cilia. Dev Cell 2016;39(4):424-437 [FREE Full text] [doi: 10.1016/j.devcel.2016.10.006] [Medline: 27818179]
52. Welburn JPI, Cheeseman IM. The microtubule-binding protein Cep170 promotes the targeting of the kinesin-13 depolymerase Kif2b to the mitotic spindle. Mol Biol Cell 2012;23(24):4786-4795 [FREE Full text] [doi: 10.1091/mbc.E12-03-0214] [Medline: 23087211]
53. Weijman JF, Vuolo L, Shak C, Pugnetti A, Mukhopadhyay AG, Hodgson LR, et al. Roles for CEP170 in cilia function and dynein-2 assembly. J Cell Sci 2024;137(8):jcs261816 [FREE Full text] [doi: 10.1242/jcs.261816] [Medline: 38533689]
54. Fitzsimons LA, Staurengo-Ferrari L, Khomula EV, Bogen O, Araldi D, Bonet IJM, et al. The nociceptor primary cilium contributes to mechanical nociceptive threshold and inflammatory and neuropathic pain. J Neurosci 2024;44(47):e1265242024. [doi: 10.1523/JNEUROSCI.1265-24.2024] [Medline: 39349056]
55. Kifle ZD, Tian J, Aitken D, Melton PE, Cicuttini F, Jones G, et al. MRI-derived abdominal adipose tissue is associated with multisite and widespread chronic pain. Reg Anesth Pain Med 2025. [doi: 10.1136/rapm-2024-105535] [Medline: 39256036]
56. Neogi T. The epidemiology and impact of pain in osteoarthritis. Osteoarthritis Cartilage 2013;21(9):1145-1153 [FREE Full text] [doi: 10.1016/j.joca.2013.03.018] [Medline: 23973124]
57. Pou KM, Massaro JM, Hoffmann U, Vasan RS, Maurovich-Horvat P, Larson MG, et al. Visceral and subcutaneous adipose tissue volumes are cross-sectionally related to markers of inflammation and oxidative stress: the Framingham Heart Study. Circulation 2007;116(11):1234-1241. [doi: 10.1161/CIRCULATIONAHA.107.710509] [Medline: 17709633]
58. Jin Y, Ren Z, Tan Y, Zhao P, Wu J. Motility plays an important role in the lifetime of mammalian lipid droplets. Int J Mol Sci 2021;22(8):3802 [FREE Full text] [doi: 10.3390/ijms22083802] [Medline: 33916886]
59. Liang Y, Zhu Z, Lu Y, Ma C, Li J, Yu K, et al. Cytoskeleton regulates lipid droplet fusion and lipid storage by controlling lipid droplet movement. Biochim Biophys Acta Mol Cell Biol Lipids 2025;1870(4):159610. [doi: 10.1016/j.bbalip.2025.159610] [Medline: 40189192]
60. Dolcini J, Landi R, Ponzio E, Picchini L, Luciani A, Formenti L, et al. Association between TNF-α, cortisol levels, and exposure to PM10 and PM2.5: a pilot study. Environ Sci Eur 2024;36 [FREE Full text] [doi: 10.1186/s12302-024-00961-2]
61. Duan Y, Chen S, Li Q, Zang Y. Neuroimmune mechanisms underlying neuropathic pain: the potential role of TNF-α-necroptosis pathway. Int J Mol Sci 2022;23(13):7191 [FREE Full text] [doi: 10.3390/ijms23137191] [Medline: 35806192]
62. Knezevic E, Nenic K, Milanovic V, Knezevic NN. The role of cortisol in chronic stress, neurodegenerative diseases, and psychological disorders. Cells 2023;12(23):2726 [FREE Full text] [doi: 10.3390/cells12232726] [Medline: 38067154]
63. Data table. GTEx Consortium. URL: https://yanglab.westlake.edu.cn/data/SMR/GTEx_V8_cis_eqtl_summary.html [accessed 2024-05-25]
64. Feng H. GTEx cross-tissue (sCCA) expression. gusevlab. URL: http://gusevlab.org/projects/fusion/#gtex-cross-tissue-scca-expression [accessed 2024-05-25]
65. cis-eQTLs. eQTLGen Consortium. URL: https://www.eqtlgen.org/cis-eqtls.html [accessed 2024-05-25]

Abbreviations

eQTL: expression Quantitative Trait Locus
FDR: false discovery rate
GTEx: Genotype-Tissue Expression
GWAS: genome-wide association study
HEIDI: heterogeneity in dependent instruments
IEU: Integrative Epidemiology Unit
IV: instrumental variable
IVW: inverse variance weighted
LD: linkage disequilibrium
MAF: minor allele frequency
MCP: multisite chronic pain
MR: Mendelian randomization
MR-PRESSO: Mendelian Randomization Pleiotropy Residual Sum and Outlier
MVMR: multivariable Mendelian randomization
OR: odds ratio
PP: posterior probability
sCCA: sparse canonical correlation analysis
SMR: summary data–based Mendelian randomization
SNP: single-nucleotide polymorphism
TWAS: transcriptome-wide association study

XSL•FO
RenderX

Edited by A Coristine; submitted 29.May.2025; peer-reviewed by J Dolcini, Y Zhu; comments to author 24.Oct.2025; accepted 24.Dec.2025; published 09.Feb.2026.

Please cite as:
Jiang J, Pu C, Cai J, Yu C, Liu Z, Xiao C
Screen Time and Chronic Pain Health: Mendelian Randomization Study
J Med Internet Res 2026;28:e78233
URL: https://www.jmir.org/2026/1/e78233
doi:10.2196/78233
PMID:

©Jiahui Jiang, Chunyan Pu, Jiarui Cai, Chuan Yu, Zhenmi Liu, Chenghan Xiao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Genre-Specific Gaming Addiction and Flourishing in Adolescents: Cross-Sectional Survey Study

Yan Luo[1], PhD; Xiaofei Chen[1]; Yaoyu Liu[1]; Wanjia Hong[1]; Miner Fan[1], LLB; David Thomas[2,3], MS; Sean P Mullen[3,4], PhD

[1]Department of Physical Education, Shanghai Jiao Tong University, Shanghai, Shanghai, China

[2]Informatics Programs, University of Illinois Urbana-Champaign, Urbana, IL, United States

[3]Department of Health and Kinesiology, University of Illinois Urbana-Champaign, Urbana, IL, United States

[4]Beckman Institute for Advanced Science and Technology, University of Illinois Urbana-Champaign, Urbana, IL, United States

Corresponding Author:
Yan Luo, PhD
Department of Physical Education
Shanghai Jiao Tong University
800 Dongchuan Road
Shanghai, Shanghai, 200240
China
Phone: 86 19901790215
Email: yluo2@sjtu.edu.cn

Abstract

Background: Adolescent gaming addiction (GA) has been linked to a range of adverse health outcomes. However, whether the associated health risks differ across game genres remains poorly understood.

Objective: Guided by VanderWeele's multidimensional flourishing framework, this study aims to examine genre-specific associations between GA and flourishing among adolescents.

Methods: This study used a cross-sectional observational design. A total of 2194 middle school students were recruited via convenience sampling from a private tutoring center in a northwestern city in China. Eligibility criteria were (1) enrollment in participating classes at the tutoring center, (2) provision of both student and parental consent, and (3) presence during questionnaire administration. The mean age of participants was 14.53 (SD 0.76) years; 985 (44.90%) were boys and 1174 (53.51%) were girls. During class time, students completed paper-based questionnaires that assessed their demographics, gaming addiction, and flourishing. Participants listed up to 3 video games played in the past month and rated their addiction to each. Games were classified into 8 genres: action and adventure (AA), sandbox and simulation (SS), multiplayer online battle arena (MOBA), shooting, strategy, casual, sports, and role-playing. Flourishing was assessed using the Human Flourishing Index across 5 domains: happiness and life satisfaction, mental and physical health, meaning and purpose, character and virtue, and close social relationships.

Results: Robust linear regression analyses (α=.05) showed that AA addiction was associated with lower overall flourishing (b=–3.11, 95% CI –4.34 to –1.88) and all 5 subdomains (happiness and life satisfaction: b=–0.46, 95% CI –0.75 to –0.17; mental and physical health: b=–0.61, 95% CI –0.88 to –0.34; meaning and purpose: b=–0.55, 95% CI –0.82 to –0.27; character and virtue: b=–0.74, 95% CI –1.06 to –0.43; and close social relationships: b=–0.62, 95% CI –0.92 to –0.32). MOBA addiction was associated with lower overall flourishing (b=–1.33, 95% CI –2.34 to –0.32), character and virtue (b=–0.34, 95% CI –0.59 to –0.08), and meaning and purpose (b=–0.34, 95% CI –0.56 to –0.11). SS addiction was associated with lower overall flourishing (b=–3.42, 95% CI –5.80 to –1.04), close social relationships (b=–0.86, 95% CI –1.46 to –0.27), and mental and physical health (b=–1.09, 95% CI –1.60 to –0.58).

Conclusions: This study provides novel evidence that the association between GA and adolescent flourishing is genre dependent. In contrast to prior research that conceptualizes health narrowly or unidimensionally, a multidimensional perspective provides a more nuanced understanding of the health risks associated with GA. The findings advance the field by showing that addiction to AA, MOBA, and SS games is associated with greater health risks than addiction to other genres. Accordingly, prevention, education, and policy efforts should prioritize higher-risk genres to promote adolescent health.

(J Med Internet Res 2026;28:e89319) doi:10.2196/89319

XSL•FO
RenderX

KEYWORDS

digital game; online game; problematic gaming; gaming disorder; youth

Introduction

Background

Video gaming is one of the most prevalent leisure activities among adolescents worldwide. For many, gaming provides entertainment, opportunities for social interaction, and cognitive benefits [1]. Yet, for a notable proportion of adolescents, gaming becomes addictive, characterized by preoccupation, impaired control, and persistence despite negative consequences. A meta-analysis estimated the prevalence of gaming disorder, a severe and clinical form of gaming addiction (GA), among adolescents to be 8.8%, based on over 400,000 participants from 155 studies across 33 countries [2]. This rate is considerably higher than that observed in other age groups [3]. Therefore, it is critical to investigate the associations between GA and adolescents' health to inform interventions and policies that foster healthier gaming practices.

A substantial body of research has documented that GA is associated with a range of negative health outcomes in adolescents [2,4]. Although these studies have significantly advanced our understanding of GA-related harms, a notable limitation is that most have adopted a relatively narrow conceptualization of health, focusing disproportionately on pathological conditions. According to the World Health Organization, health is defined as "a state of complete physical, mental, and social wellbeing" [5], which highlights the need to investigate the relationship between GA and health from a broader perspective. One construct that captures this view is flourishing, which reflects a state in which "all aspects of a person's life are good" [6]. VanderWeele [6] proposed that flourishing should, at a minimum, encompass 5 domains of human life: (1) happiness and life satisfaction, (2) mental and physical health, (3) meaning and purpose, (4) character and virtue, and (5) close social relationships [6]. Collectively, these domains reflect distinct facets of a well-lived life. Happiness and life satisfaction reflect how good life feels overall. Mental and physical health capture how well individuals are functioning psychologically and biologically. Meaning and purpose speak to whether life feels worthwhile and directed. Character and virtue reflect the capacity for moral self-regulation and for acting well when facing challenges. Close social relationships reflect the quality of one's relational world. Although these domains are often correlated, they are conceptually distinct. Flourishing entails doing well across all 5 domains, and impairment in any single domain indicates a substantive shortfall in overall well-being. Therefore, examining the association between GA and flourishing is crucial, as it provides a more holistic understanding of how GA may constrain adolescents' ability to achieve well-rounded lives and may inform prevention and policy efforts to promote adolescents' health.

Review of Relevant Scholarship

Previous research suggests that GA is associated with multiple flourishing-related domains. In the domain of happiness and life satisfaction, a large cross-national study of over 14,000 adolescents found that GA was linked to lower life satisfaction [7]. A longitudinal study further showed that GA predicted subsequent declines in adolescents' life satisfaction over 6 months, whereas the reverse effect was not observed [8]. These findings suggest that GA may function as an antecedent of reduced life satisfaction. Although no studies to date have directly examined the association between GA and happiness among adolescents, GA has been consistently shown to correlate with several factors strongly connected to happiness [9-13], including poorer academic achievement [2], lower self-esteem [2], reduced health-related quality of life [14], more frequent negative life events [15], and worse family financial conditions [16]. In summary, existing evidence suggests that GA is associated with lower life satisfaction and may also be linked to lower levels of happiness among adolescents.

The relationship between GA and mental health outcomes has been the most extensively investigated. Studies using composite indicators of mental health have shown that GA was associated with more frequent psychological complaints [7], elevated psychopathological symptoms [17-19], and a higher likelihood of receiving at least one mental health diagnosis [20]. Research focusing on specific mental health conditions further revealed that GA was associated with increased depression [2,21], anxiety [22-24], general and academic stress [2,25,26], loneliness [27], emotional distress [2,23,28,29], suicidal ideation [19,30], suicide risk [31], emotional and behavioral problems [24], sleep difficulties [4,32], and executive dysfunction [32]. A few studies have examined the association between GA and physical health. Evidence suggests that GA is linked to unhealthy weight status, including both underweight and overweight [33,34], more somatic complaints [17], musculoskeletal pain [4], and dry eye symptoms [4]. Overall, these findings suggest that GA is associated with a broad range of mental and physical health problems, with the literature to date focusing predominantly on mental health outcomes.

Evidence supporting associations between GA and the remaining flourishing domains is limited in the current literature. Existing findings suggest that adolescents with GA reported both a lower presence of meaning in life and a stronger search for meaning in life [35]. Although no adolescent studies have directly examined GA in relation to sense of purpose, research with young adults indicates a negative association between GA and purpose in life [36,37]. Regarding character and virtue, studies have shown that GA is linked to lower empathy [38], lower moral levels [39], reduced academic perseverance [40], and fewer prosocial behaviors [38]. In the relational domain, GA has been associated with lower interpersonal trust [41], conflictual parent-child relationships [42,43], peer alienation [26], less social support [44], and greater reliance on online-only friendships [14]. Overall, the available evidence suggests that GA may have broad adverse associations across these flourishing domains in adolescents.

Beyond research conducted on individual domains of flourishing, no study to date has investigated GA in relation to

XSL•FO
RenderX

flourishing as an integrated construct. Existing work on other problematic behaviors has found that lower flourishing was related to greater cannabis use [45], higher media addiction [46], higher smartphone addiction [47], and higher Facebook addiction [48,49]. These findings imply that GA, as a form of problematic behavior, may likewise be associated with reduced flourishing. It is important to note that the flourishing framework used in these studies is derived from Diener and colleagues [50], which captures all but the health dimension of flourishing proposed by VanderWeele [6]. As a result, it remains unclear whether VanderWeele's broader conceptualization of flourishing is associated with any form of problematic behavior, including GA. A recent study that employed VanderWeele's flourishing framework examined the relationship between digital game use and flourishing among older adults. No differences were found in overall flourishing or in any of the 5 subdomains between current gamers, former gamers, and nongamers [51]. As this study did not focus on GA and was conducted in an older adult population, its findings provide minimal guidance regarding how GA relates to flourishing among adolescent gamers.

Although prior studies have documented associations between GA and individual elements of flourishing, a critical gap remains regarding how GA relates to overall flourishing and its subdomains among adolescents. As flourishing is a multidimensional construct encompassing various aspects of well-being, investigating these associations is essential for establishing a more comprehensive understanding of how GA may relate to adolescents' well-being across multiple life domains. A further limitation of the existing literature is that most studies have conceptualized gaming as a single, homogeneous activity. It remains unclear whether GA is uniformly associated with flourishing across different game genres. Emerging evidence suggests that the psychological correlates of gaming may vary substantially by genre. For example, Park and colleagues [52] reported that adolescents who primarily played massively multiplayer online role-playing games exhibited the highest levels of social anxiety, those who preferred first-person shooter games showed the lowest anxiety levels, and strategy-game players reported the highest self-esteem [52]. Other studies have demonstrated that certain genres were associated with increased aggression, reduced empathy, and prosocial behaviors [53-56]. As these factors are important components of flourishing, such findings suggest that the association between GA and flourishing may differ depending on the game genres adolescents play. Understanding these genre-specific associations could provide a more nuanced picture of how GA relates to adolescent health, thereby informing the development of targeted intervention strategies tailored to specific game genres.

Hypothesis and Aim

To address these gaps, this study aimed to investigate genre-specific associations between GA and flourishing among adolescents, using VanderWeele's multidimensional flourishing framework. Drawing on prior evidence, we specified a single primary hypothesis: *GA would be negatively associated with overall flourishing and its subdomains for certain game genres, but not uniformly across all genres.* Our secondary hypothesis was that *overall GA across game genres would be negatively associated with overall flourishing and its subdomains.*

Methods

Inclusion and Exclusion

This study was reported in accordance with the Journal Article Reporting Standards for Quantitative Research (JARS-Quant) checklist [57]. A total of 2194 middle school students from a private tutoring center located in a northwestern city in China participated in this study. Inclusion criteria were as follows: (1) enrollment in participating classes at the tutoring center; (2) provision of both student and parental consent; and (3) presence during questionnaire administration. No exclusion criteria were applied.

Participant Characteristics

As shown in Table 1, participants had a mean age of 14.53 (SD 0.76) years. Of the 2194 students, there were 985 (44.90%) boys and 1174 (53.51%) girls. Regarding grade level, 52 (2.37%) were in the seventh grade, 969 (44.17%) were in the eighth grade, and 1092 (49.77%) were in the ninth grade. On average, participants reported a family economic status of 2.23 (SD 0.70) and 18.81 (SD 16.40) hours of social media use per week. In terms of gaming behavior, 610 (27.80%) participants reported not playing any games, 327 (14.90%) reported playing 1 game, 525 (23.93%) reported playing 2 games, and 732 (33.36%) reported playing 3 games. The most frequently played game genre was multiplayer online battle arena (MOBA; n=808, 36.83%), followed by shooting (n=589, 26.85%), action and adventure (AA; n=342, 15.59%), casual (n=302, 13.76%), sports (n=156, 7.11%), sandbox and simulation (SS; n=99, 4.51%), role-playing (n=89, 4.06%), and strategy (n=54, 2.46%) games.

XSL•FO
RenderX

Table 1. Descriptive statistics of demographic and gaming characteristics (N=2194).

Variables	Values
Age (years), mean (SD)	14.53 (0.76)
Sex, n (%)[a]	
Men	985 (44.90)
Women	1174 (53.51)
Grade level, n (%)	
Seventh grade	52 (2.37)
Eighth grade	969 (44.17)
Ninth grade	1092 (49.77)
Family economic status[b], mean (SD)	2.23 (0.70)
Social media time (hours), mean (SD)	18.81 (16.40)
Number of games played, n (%)	
Do not play a game	610 (27.80)
Playing 1 game	327 (14.90)
Playing 2 games	525 (23.93)
Playing 3 games	732 (33.36)
Number of participants per genre[c], n (%)	
Multiplayer online battle arena	808 (36.83)
Shooting	589 (26.85)
Action and adventure	342 (15.59)
Casual	302 (13.76)
Sports	156 (7.11)
Sandbox and simulation	99 (4.51)
Role-playing	89 (4.06)
Strategy	54 (2.46)

[a]35 missing values for this variable.

[b]Used a 5-point Likert scale for evaluation (0=extremely poor to 4=excellent).

[c]The number of participants per genre refers to those who reported playing at least one game within that genre.

Sampling Procedures

Participants were recruited using convenience sampling. A total of 2269 middle school students were approached, of whom 2194 (96.69%) agreed to participate and completed the survey. Self-selection occurred at both the teacher and student levels. At the teacher level, instructors at the tutoring center were invited to assist with data collection, and 10 teachers agreed to participate. At the student level, only students who expressed interest and provided both student assent and parental consent were included. Each participating teacher administered the survey once to each of their classes, all of which were held at the tutoring center. Each participant received ¥10 (US $1.44) in cash as compensation for completing the survey. Study procedures were conducted in compliance with the Declaration of Helsinki and approved by the ethics committee of the first author's institution (approval number H20250747I). Data were collected between September 10 and October 10, 2025.

Sample Size, Power, and Precision

We conducted a priori power analysis using G*Power for a linear multiple regression model (fixed model, R^2 increase) to determine the required sample size [58]. We assumed a small effect size for each genre-specific addiction predictor (partial R^2=0.01), an α level of .05, and a desired statistical power of 0.80. The model specified 1 tested predictor and a total of 12 predictors (8 genre-specific addiction variables and 4 covariates). Under these assumptions, the minimum required sample size was 779 participants. Given the achieved sample size of 2194 students, the study was sufficiently powered to detect genre-specific addiction predictors accounting for as little as 1% of unique variance in the outcome.

Measures and Covariates

Measures and covariates are presented in Textbox 1.

XSL•FO
RenderX

Textbox 1. Measures and covariates.

1. Genre-specific addiction

The extent to which an individual exhibits excessive and compulsive engagement with a specific genre of video games.

2. Flourishing

A multidimensional state of well-being encompassing domains of happiness and life satisfaction, mental and physical health, meaning and purpose, character and virtue, and close social relationships.

3. Sex

A biological classification of individuals, typically men or women.

4. Social media use

The extent to which students engage with social networking platforms.

5. Family economic status

The financial conditions of a household, reflecting the family's access to economic resources and its relative position within the socioeconomic hierarchy.

Data Collection

We contacted a friend of the first author (YL), a physics teacher at the private tutoring center, to inform him about the study's objectives and data collection procedures. After agreeing to assist, he introduced us to other teachers at the center. Ten additional teachers consented to participate in data collection.

We compiled measurement instruments into a questionnaire, which also included demographic items (age, sex, and grade). Each questionnaire was labeled with a unique identifier number at the bottom of every page to distinguish individual responses. Digital versions of the questionnaire and consent form were sent to the teachers via WeChat (Tencent Holdings Limited). Teachers were instructed to print these materials for use in class.

Each teacher administered the survey once to each of their classes, all of which were held at the tutoring center. At the beginning of each class, teachers explained the study's purpose and distributed parental consent forms to students who expressed interest in participating. Students who missed 2 classes were excluded from data collection. During the third class session, teachers distributed the questionnaires. Students were told that the survey was anonymous and that neither their parents nor their teachers would have access to their responses. They were instructed to answer honestly, write legibly, and seek clarification if any survey items were unclear. Completing the survey took approximately 10 minutes. The friend of the first author collected all the questionnaires from the teachers and shipped them to the first author. Each teacher received ¥800 (US $115.01) in cash as compensation. The survey questionnaire is available in Multimedia Appendix 1.

Quality of Measurements

To enhance the quality of measurements, we conducted an online group meeting with all participating teachers before data collection to provide standardized training and explain the data collection protocol in detail. In addition, we trained research assistants in standardized procedures for digitizing paper questionnaires to ensure accurate data entry.

Instrumentation

Video Game Name

The video game name was obtained using the following question: "Please provide the names of all the games you have played during the past month (up to 3). If you have played more than 3, please list the 3 you played most frequently."

Game-Specific Addiction

For each game, participants rated their level of addiction by responding to the following question: "Please indicate your level of addiction for each of the games listed above." Responses were given on a 5-point Likert scale (0=none, 1=mild, 2=moderate, 3=moderately severe, and 4=severe), with higher scores indicating stronger addiction to a particular game. We used a single item to measure GA for 3 reasons. First, widely used instruments such as the Gaming Addiction Scale for Adolescents [59] and the Internet Gaming Disorder Scale [60] assess overall GA and are therefore unsuitable for measuring game-specific addiction. Second, even if these scales were adapted for game-specific assessment, participants who played 2 or 3 games would need to complete the same scale multiple times, increasing response burden and the risk of response bias. Third, multi-item scales would substantially increase completion time, which was impractical given that the survey was administered during class. Similar single-item measures of GA have been used in prior research, although primarily in adult samples [61,62]. In adolescent populations, single self-diagnostic items assessing other behavioral addictions have shown strong correlations with standardized multi-item instruments [63,64], suggesting that adolescents can comprehend and self-evaluate the severity of behavioral addictions using a single question. The process of transforming game-specific addiction into genre-specific addiction is described in the "Data Processing" section.

Flourishing

Flourishing was assessed using the Chinese version of the Human Flourishing Index (HFI) [65]. The scale consists of 10 items covering 5 domains (2 items per domain): (1) happiness and life satisfaction; (2) mental and physical health; (3) meaning and purpose; (4) character and virtue; and (5) close social

XSL•FO
RenderX

relationships. Items were rated on an 11-point Likert scale ranging from 0 to 10, with response labels varying across items. Domain scores were computed by summing the scores of the respective items. Overall flourishing was calculated as the sum of all item scores, with higher values reflecting greater flourishing. If any item required to compute a composite score was missing, the corresponding composite score was coded as missing.

Covariates

Previous research has found associations between GA, sex [66], social media use [67], and socioeconomic status [68]. To account for their confounding effects [69], we measured participants' weekly social media time and perceived family economic status using 2 questions. Participants were asked, "In general, how many hours per day do you spend browsing social media platforms (eg, Douyin, Weibo, WeChat Moments, Xiaohongshu, and Bilibili) on weekdays and weekends? Please report to the nearest 0 or 30 minutes (eg, 2 hours 0 minutes, 4 hours 30 minutes)." Total weekly social media time (in hours) was calculated using the formula: 5 × (weekday hours + weekday minutes/60) + 2 × (weekend hours + weekend minutes/60). The family economic status question was, "How would you describe your family's economic status?" Responses were on a 5-point Likert scale (0=extremely poor to 4=excellent), with higher scores reflecting better family economic status.

Masking, Psychometrics, Conditions and Design, and Data Processing

As this study employed a cross-sectional observational design without experimental manipulation, masking of participants, data collectors, or outcome assessors was not applicable.

The HFI demonstrated good internal consistency in our sample (Cronbach α=0.93 for overall flourishing, and ranging from 0.76 to 0.89 across domains). Game-specific addiction and family economic status were assessed using single-item measures, and sex and social media use were measured using non–Likert-type items. Therefore, internal consistency coefficients were not applicable for these measures.

This study used a cross-sectional observational design with no experimental manipulation.

Details of the data processing are provided in Multimedia Appendix 2.

Data Diagnostics

We screened social media use for implausible values and recoded 44 values exceeding 98 hours per week (>14 hours/day) as missing. Extreme outliers were identified using the IQR method. Fourteen values falling below Q1 – 3 × IQR or above Q3 + 3 × IQR were recoded as missing. All other variables were examined for implausible values, and none were identified. Responses were considered missing when participants left an item unanswered. Missing data were handled using multiple imputation by chained equations before statistical analysis. We examined the distributions of the outcome variables and found that overall flourishing and its 5 subscales were negatively skewed across all datasets.

Analytic Strategy

All statistical analyses were conducted in R (R Foundation) [70]. To reduce the influence of nonnormal outcome distributions on parameter estimation, we employed robust linear regression models [71]. To test the primary hypothesis, for each of the 6 flourishing outcomes, we fitted robust multiple regression models for each dataset, including the 8 genre-specific addiction variables as predictors, and sex, social media time, family economic status, and addiction to other games as covariates. To test the secondary hypothesis, we fitted separate robust regression models for each of the 6 flourishing outcomes with overall GA as a predictor, and sex, social media time, and family economic status as covariates. Parameter estimates were pooled across datasets using Rubin's rules. We controlled the type I error rate across all models for genre-specific addiction predictors (6 pooled estimates × 8 genres=48 P values) and the overall GA predictors (6 pooled estimates) by adjusting the 54 P values using the Benjamini-Hochberg false discovery rate procedure. The adjusted P values were hereafter referred to as q values. Statistical significance for all q values was set at α=.05.

As sensitivity analyses, we refitted all models using complete-case data to assess the robustness of the findings to the handling of missing data. The false discovery rate procedure was also applied in complete-case analyses. We further evaluated the robustness of the observed associations to unmeasured confounding by calculating E values for statistically significant pooled regression coefficients using the "EValue" package in R [72]. E values quantify the minimum strength of association (on the risk ratio scale) that unmeasured confounder(s) would need to have with both the exposure and the outcome, beyond the measured covariates, to fully explain the observed effect [73]. We also calculated E values for the CI limit closest to the null, which quantify the minimum strength of association that unmeasured confounder(s) would need to render the observed effect nonsignificant. E values were calculated separately for each imputed dataset, and the mean and SD across the 20 imputed datasets were reported.

Ethical Considerations

This study was approved by the Ethics Committee of Shanghai Jiao Tong University (approval number H20250747I). After the study procedures were explained, students who expressed interest were given parental consent forms to take home. Written informed consent was obtained from parents through signed consent forms before participation. Only students who returned signed parental consent forms and agreed to participate completed the survey, and participation was voluntary. Data were collected anonymously to protect participants' confidentiality, and only the research team had access to the dataset. Each participant received ¥10 (US $1.44) in cash as compensation for completing the survey. No identifiable images of participants are included in the manuscript or its multimedia appendices.

XSL•FO
RenderX

Results

Participant Flow and Recruitment

Of the 2269 middle school students eligible for participation, 68 were excluded due to absence from 2 consecutive data collection sessions, and 7 declined to participate. Figure 1 shows the flow of participants across each stage of the study. The final sample included 2194 participants.

Participants were recruited between September 10 and September 24, 2025.

Figure 1. Flow of participants through each stage of the study.

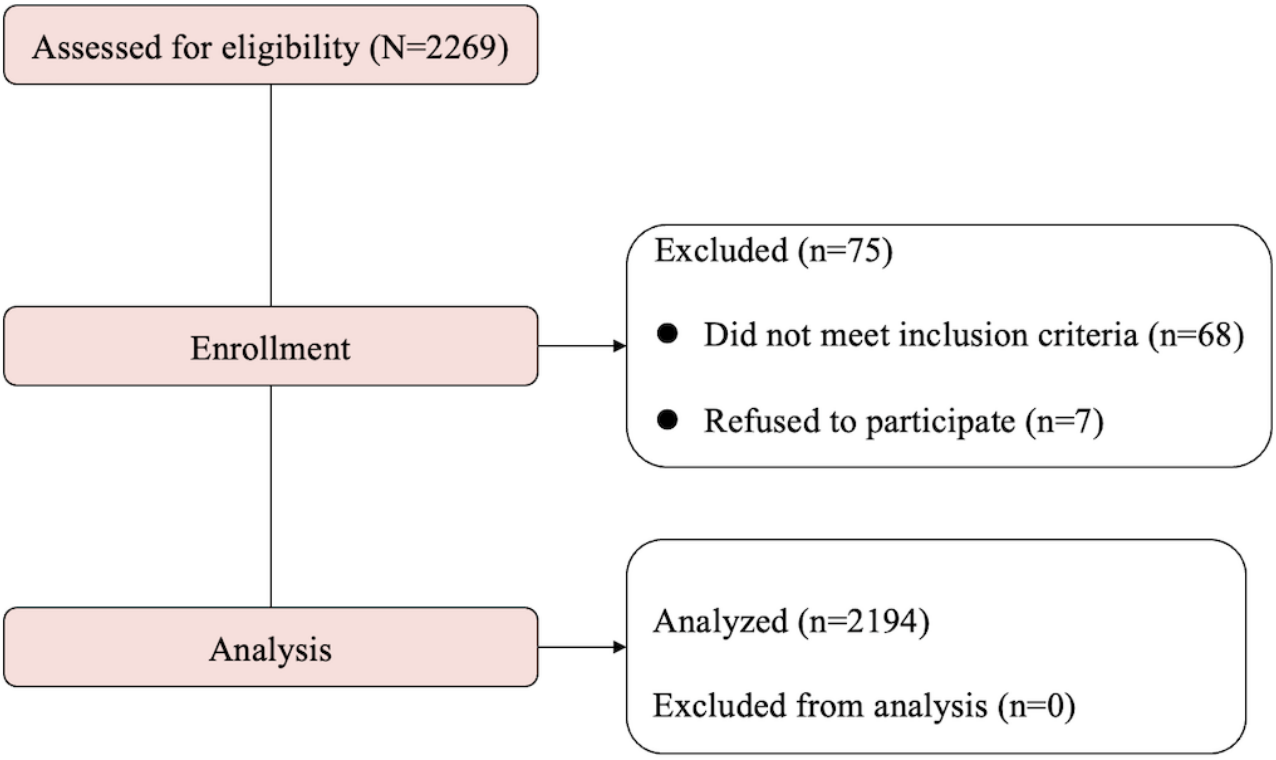

Statistics and Data Analysis

Inspection of missing data revealed that, among the 2194 participating students, data were missing for age in 22 (1%) cases, sex in 35 (1.60%), grade in 81 (3.69%), family economic status in 254 (11.58%), and social media time in 378 (17.23%). For the HFI, missing data were observed for happiness and life satisfaction in 34 (1.55%) participants, mental and physical health in 34 (1.55%), meaning and purpose in 170 (7.75%), character and virtue in 174 (7.93%), close social relationships in 164 (7.47%), and overall flourishing in 226 (10.30%). Across individual HFI items, missing data ranged from 24 (1.09%) to 164 (7.47%). For genre-specific addiction variables, missing data ranged from 3 (0.14%) to 76 (3.46%). In addition, data were missing for the addiction-to-other-games variable in 24 (1.09%) participants, whereas no missing data were observed for the overall addiction variable.

The Little missing completely at random test applied to the dataset was statistically significant, χ^2_{1298}=1495.00 (P<.001), indicating that the missing mechanism was not missing completely at random. As several variables exceeded 5% missingness and complete-case analysis can reduce statistical power and introduce bias [74], we assumed that the missingness mechanism was missing at random and applied multiple imputation by chained equations to impute missing data [75]. The predictor matrix included sex, family economic status, social media time, 10 flourishing items, 8 genre-specific addiction variables, and the addiction-to-other-games variable. The overall addiction variable was excluded due to its high linear dependence on other addiction variables. Sex was imputed using logistic regression, and all other variables were imputed using predictive mean matching. We generated 20 imputed datasets, each with 20 iterations. Convergence was confirmed using trace plots, and the distributions of the imputed values aligned well with those of the observed data. After imputation, we calculated composite scores for each dataset.

As shown in Table 2, on average, participants reported a mean overall GA score of 1.49 (SD 1.23). Among genre-specific addictions, participants reported the highest levels of addiction to strategy games (mean 1.65, SD 1.11), followed by AA (mean 1.62, SD 1.24), shooting (mean 1.58, SD 1.21), MOBA (mean 1.46, SD 1.10), SS (mean 1.38, SD 1.36), role-playing (mean 1.36, SD 1.02), casual (mean 1.11, SD 1.09), and sports (mean 0.96, SD 1.01) games. Regarding the HFI, the mean domain scores were 13.54 (SD 4.82) for happiness and life satisfaction, 14.54 (SD 4.72) for mental and physical health, 13.72 (SD 4.58) for meaning and purpose, 12.50 (SD 4.96) for character and virtue, and 14.24 (SD 4.99) for close social relationships. The overall flourishing score averaged 68.60 (SD 20.99).

XSL•FO
RenderX

Table 2. Descriptive statistics of gaming addiction and flourishing outcomes (N=2194).

Variables	Values, mean (SD)
Overall addiction	1.49 (1.23)
Gaming addiction per genre	
Multiplayer online battle arena	1.46 (1.10)
Shooting	1.58 (1.21)
Action and adventure	1.62 (1.24)
Casual	1.11 (1.09)
Sports	0.96 (1.01)
Sandbox and simulation	1.38 (1.36)
Role-playing	1.36 (1.02)
Strategy	1.65 (1.11)
Human Flourishing Index	
Happiness and life satisfaction	13.54 (4.82)
Mental and physical health	14.54 (4.72)
Meaning and purpose	13.72 (4.58)
Character and virtue	12.50 (4.96)
Close social relationships	14.24 (4.99)
Overall flourishing	68.60 (20.99)

Table 3 and Figure 2 present robust linear regression results examining the associations between GA and flourishing outcomes using multiply imputed data. Regarding the primary hypothesis, AA addiction was the only predictor that was negatively associated with all flourishing outcomes (overall flourishing: b=–3.11, 95% CI –4.34 to –1.88, q<.001; happiness and life satisfaction: b=–0.46, 95% CI –0.75 to –0.17, q=.009; mental and physical health: b=–0.61, 95% CI –0.88 to –0.34, q<.001; meaning and purpose: b=–0.55, 95% CI –0.82 to –0.27, q=.001; character and virtue: b=–0.74, 95% CI –1.06 to –0.43, q<.001; and close social relationships: b=–0.62, 95% CI –0.92 to –0.32, q<.001). MOBA addiction was negatively associated with overall flourishing (b=–1.33, 95% CI –2.34 to –0.32, q=.03), meaning and purpose (b=–0.34, 95% CI –0.56 to –0.11, q=.02), and character and virtue (b=–0.34, 95% CI –0.59 to –0.08, q=.03). SS addiction was negatively associated with overall flourishing (b=–3.42, 95% CI –5.80 to –1.04, q=.02), mental and physical health (b=–1.09, 95% CI –1.60 to –0.58, q<.001), and close social relationships (b=–0.86, 95% CI –1.46 to –0.27, q=.02). Other genre-specific associations were not statistically significant (q≥.05 for all; discussed later). Analyses testing the secondary hypothesis showed that overall GA was negatively associated with all flourishing outcomes (overall flourishing: b=–1.99, 95% CI –2.74 to –1.24, q<.001; happiness and life satisfaction: b=–0.24, 95% CI –0.41 to –0.07, q=.02; mental and physical health: b=–0.38, 95% CI –0.55 to –0.22, q<.001; meaning and purpose: b=–0.45, 95% CI –0.62 to –0.28, q<.001; character and virtue: b=–0.60, 95% CI –0.78 to –0.41, q<.001; and close social relationships: b=–0.35, 95% CI –0.53 to –0.17, q=.001).

XSL•FO
RenderX

Table 3. Robust linear regression results for the associations between gaming addiction and flourishing outcomes based on multiple imputed data (N=2194).

Variables[a]	Multiple imputation analysis[b]	
	b (95% CI)	*q* value[c]
Outcome: overall flourishing		
Predictors (genre-specific model)		
Multiplayer online battle arena	–1.33 (–2.34 to –0.32)	.03
Shooting	0.39 (–0.69 to 1.47)	.57
Casual	–1.54 (–3.24 to 0.15)	.17
Action and adventure	–3.11 (–4.34 to –1.88)	<.001
Sandbox and simulation	–3.42 (–5.80 to –1.04)	.02
Sports	–1.43 (–3.94 to 1.07)	.36
Strategy	–2.17 (–5.05 to 0.71)	.22
Role-playing	–0.39 (–3.09 to 2.31)	.84
Predictor (overall addiction model)		
Overall addiction	–1.99 (–2.74 to –1.24)	<.001
Outcome: happiness and life satisfaction		
Predictors (genre-specific model)		
Multiplayer online battle arena	–0.28 (–0.51 to –0.05)	.05
Shooting	0.22 (–0.03 to 0.47)	.18
Casual	–0.23 (–0.61 to 0.16)	.36
Action and adventure	–0.46 (–0.75 to –0.17)	.009
Sandbox and simulation	–0.61 (–1.16 to –0.05)	.08
Sports	–0.06 (–0.64 to 0.52)	.84
Strategy	–0.40 (–1.07 to 0.27)	.36
Role-playing	–0.07 (–0.69 to 0.55)	.84
Predictor (overall addiction model)		
Overall addiction	–0.24 (–0.41 to –0.07)	.02
Outcome: mental and physical health		
Predictors (genre-specific model)		
Multiplayer online battle arena	–0.18 (–0.40 to 0.04)	.20
Shooting	0.19 (–0.05 to 0.43)	.20
Casual	–0.37 (–0.74 to 0.00)	.12
Action and adventure	–0.61 (–0.88 to –0.34)	<.001
Sandbox and simulation	–1.09 (–1.60 to –0.58)	<.001
Sports	–0.39 (–0.94 to 0.17)	.26
Strategy	–0.38 (–1.05 to 0.30)	.36
Role-playing	–0.46 (–1.06 to 0.15)	.22
Predictor (overall addiction model)		
Overall addiction	–0.38 (–0.55 to –0.22)	<.001
Outcome: meaning and purpose		
Predictors (genre-specific model)		
Multiplayer online battle arena	–0.34 (–0.56 to –0.11)	.02
Shooting	–0.02 (–0.27 to 0.22)	.84

XSL•FO
RenderX

Variables[a]	Multiple imputation analysis[b]	
	b (95% CI)	*q* value[c]
Casual	–0.37 (–0.76 to 0.01)	.13
Action and adventure	–0.55 (–0.82 to –0.27)	.001
Sandbox and simulation	–0.40 (–0.93 to 0.13)	.22
Sports	–0.36 (–0.93 to 0.21)	.32
Strategy	–0.55 (–1.20 to 0.09)	.18
Role-playing	–0.12 (–0.72 to 0.49)	.78
Predictor (overall addiction model)		
Overall addiction	–0.45 (–0.62 to –0.28)	<.001
Outcome: character and virtue		
Predictors (genre-specific model)		
Multiplayer online battle arena	–0.34 (–0.59 to –0.08)	.03
Shooting	–0.09 (–0.36 to 0.19)	.61
Casual	–0.18 (–0.60 to 0.25)	.51
Action and adventure	–0.74 (–1.06 to –0.43)	<.001
Sandbox and simulation	–0.51 (–1.09 to 0.08)	.18
Sports	–0.50 (–1.13 to 0.14)	.22
Strategy	–0.35 (–1.05 to 0.36)	.43
Role-playing	0.07 (–0.59 to 0.73)	.84
Predictor (overall addiction model)		
Overall addiction	–0.60 (–0.78 to –0.41)	<.001
Outcome: close social relationships		
Predictors (genre-specific model)		
Multiplayer online battle arena	–0.29 (–0.54 to –0.04)	.07
Shooting	0.15 (–0.11 to 0.41)	.36
Casual	–0.35 (–0.76 to 0.05)	.18
Action and adventure	–0.62 (–0.92 to –0.32)	<.001
Sandbox and simulation	–0.86 (–1.46 to –0.27)	.02
Sports	–0.15 (–0.75 to 0.46)	.72
Strategy	–0.22 (–0.90 to 0.46)	.61
Role-playing	0.27 (–0.37 to 0.92)	.51
Predictor (overall addiction model)		
Overall addiction	–0.35 (–0.53 to –0.17)	.001

[a]Game genres denote addiction to the corresponding genres.

[b]All models included sex, family economic status, and social media time as covariates; genre-specific models additionally controlled for addiction-to-other-games.

[c]*q* indicates that raw *P* values were adjusted using the Benjamini-Hochberg false discovery rate procedure.

XSL•FO
RenderX

Figure 2. Robust linear regression results showing the associations between gaming addiction and flourishing outcomes based on multiple imputed data (N=2194). HFI: Human Flourishing Index; MOBA: multiplayer online battle arena.

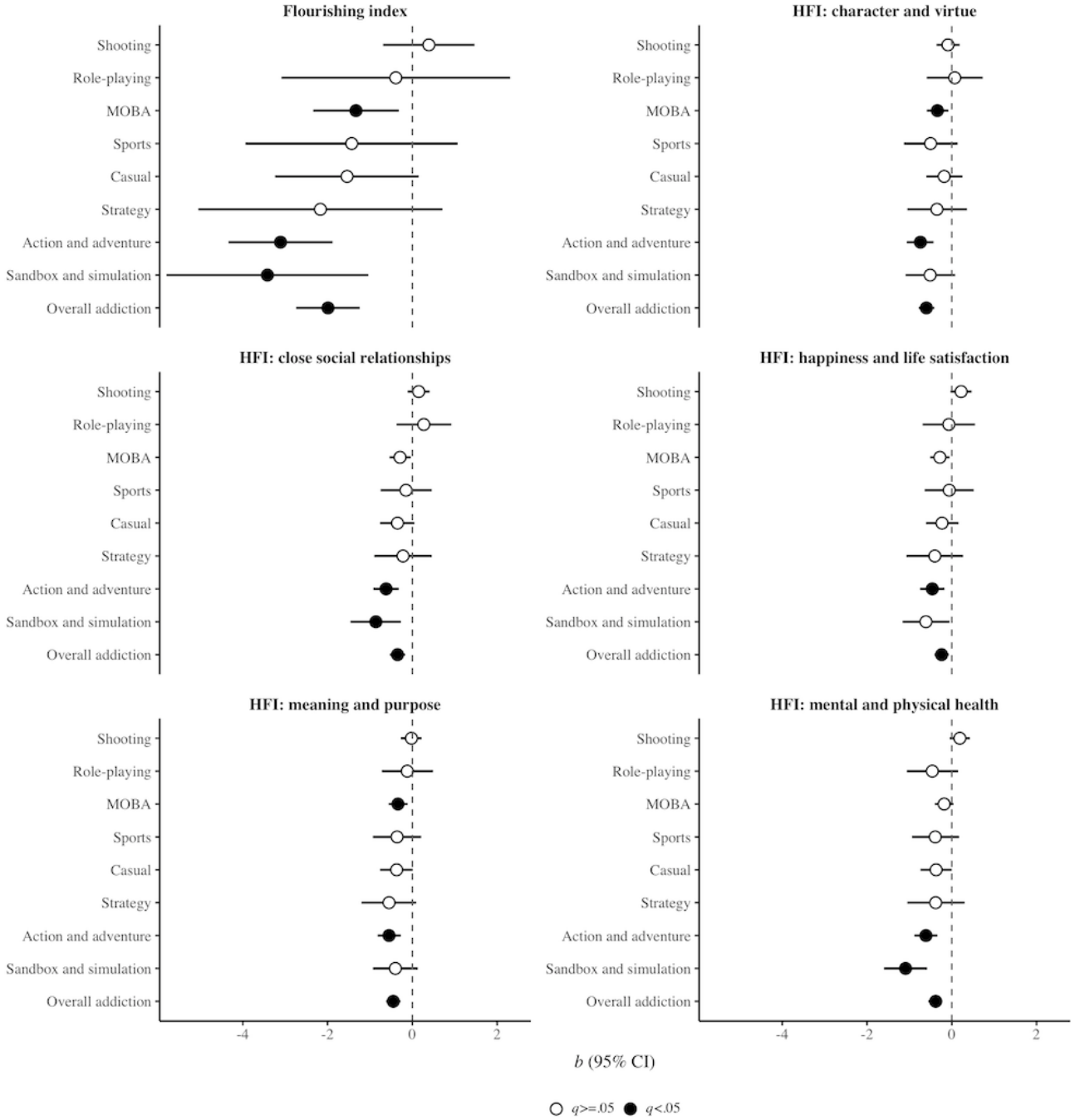

Although overall GA was negatively associated with all flourishing domains, the genre-specific findings indicate substantial heterogeneity in the addiction-flourishing relationship across game genres. The most prominent genre-specific predictor was AA addiction, which was associated with lower scores across all flourishing domains, suggesting that adolescents with AA addiction may constitute a particularly high-risk subgroup with broadly impaired well-being. Although both SS and MOBA addiction were associated with lower overall flourishing, their domain-specific patterns differed. The SS-flourishing association appeared to be driven primarily by poorer mental and physical health and weaker close social relationships, whereas the MOBA-flourishing association was driven mainly by lower meaning and purpose and character and virtue. The absence of significant associations for other genres (overall flourishing: q=.57, .17, .36, .22, and .84; happiness and life satisfaction: q=.18, .36, .84, .36, and .84; mental and physical health: q=.20, .12, .26, .36, and .22; meaning and purpose: q=.84, .13, .32, .18, and .78; character and virtue: q=.61, .51, .22, .43, and .84; and close social relationships: q=.36, .18, .72, .61, and .51, for shooting, casual, sports, strategy, and role-playing, respectively) suggests that addiction to these genres may be comparatively less detrimental to flourishing than addiction to AA, SS, and MOBA.

We conducted model diagnostics to evaluate regression assumptions. Multicollinearity among predictors was assessed using variance inflation factors. The maximal variance inflation factor across the 20 imputed datasets and all outcome variables (including the overall GA models) was 1.22, indicating no

XSL•FO
RenderX

problematic collinearity among predictors [76]. Visual inspection of residuals-versus-fitted plots across all imputed datasets and models indicated no substantial departures from linearity or other patterns suggesting model misspecification.

Multimedia Appendix 3 presents the results of the robust linear regression analyses examining the associations between GA and flourishing outcomes using complete-case data. The pattern of significant predictors was consistent with the analyses based on multiple imputed data for overall flourishing, mental and physical health, meaning and purpose, and character and virtue. Compared with the multiple imputation analyses, the effect of overall GA on happiness and life satisfaction became nonsignificant in the complete-case analysis (b=–0.22, 95% CI –0.41 to –0.03, q=.07), whereas the effect of MOBA on close social relationships became statistically significant in the complete-case analysis (b=–0.39, 95% CI –0.66 to –0.12, q=.02). Despite these discrepancies, the overall pattern of results was generally consistent across the 2 approaches, suggesting that the findings were largely robust to the handling of missing data.

Multimedia Appendix 4 presents the E values for the significant pooled effects from the robust linear regression analyses based on multiple imputed data. Across all flourishing outcomes, the E values for the regression coefficients ranged from 1.31 to 1.55, indicating moderate robustness to unmeasured confounding. This suggests that unmeasured confounder(s) would need to be associated with both GA and flourishing outcomes by at least a 1.3-fold risk ratio, beyond the measured covariates, to fully explain the observed effects. The E values for the CI limits closest to the null ranged from 1.14 to 1.43, with 7 values below 1.20. These findings indicate that relatively weak unmeasured confounder(s) could attenuate the statistical significance of some estimates. Therefore, certain genre-specific associations should be interpreted with caution, given their potential susceptibility to residual confounding.

Discussion

Principal Findings

Using VanderWeele's multidimensional flourishing framework, this study examined genre-specific associations between GA and flourishing among adolescents. Consistent with our primary hypothesis, the associations between GA and flourishing varied across game genres rather than being uniform. Addiction to AA, SS, and MOBA games showed significant negative associations with all or some flourishing outcomes, whereas addiction to other genres was not significantly associated with any flourishing outcomes. Consistent with our secondary hypothesis, overall GA was negatively associated with all flourishing outcomes.

Interpretations and Comparisons to Existing Literature

A substantial body of research has linked adolescents' GA to a range of health outcomes, including reduced life satisfaction [7], poorer health [7,17], lower meaning in life [35], lower empathy and moral levels [38,39], and difficulties in social relationships [14,26]. Consistent with this literature, this study found that overall GA was negatively associated with all 5 flourishing subdomains (ie, happiness and life satisfaction, mental and physical health, meaning and purpose, character and virtue, and close social relationships). In addition, our findings extend prior work by showing that the negative correlates of GA are not confined to isolated facets of well-being but instead span the full spectrum of adolescents' flourishing. This suggests that addictive gaming may constrain adolescents' capacity to thrive across multiple life domains. Moreover, our results indicate that the adverse association between GA and flourishing is not uniform across game genres. While addiction to certain genres was negatively related to flourishing, addiction to others was not associated with any flourishing outcomes.

The genre-specific analyses revealed that AA addiction was the only consistent negative predictor of overall flourishing and all 5 of its domains. Several mechanisms may explain these findings. First, AA games possess design features that may heighten the risk of reduced flourishing. Compared with other genres, AA games often feature open-world exploration and story-driven adventures that lack natural stopping points [77]. Such features can promote prolonged, uninterrupted play, frequently in solitary contexts, which may displace real-life activities essential for flourishing [78,79]. Second, AA games commonly include fast-paced, skill-based activities (eg, defeating enemies and solving puzzles) that produce rapid cycles of in-game achievement [80]. Although such virtual accomplishments may temporarily "scratch the itch" for success [81], prior research suggests that they function more as compensators than as substitutes for real-world competence [82]. Thus, adolescents with addictive gaming patterns may feel competent during gameplay, yet their underlying need for real-world competence remains unmet, potentially leading to reduced well-being [83]. Third, adolescents with preexisting low levels of flourishing may be particularly drawn to AA games. Those with low life satisfaction, weak social relationships, poor emotional well-being, or a lack of life purpose may find AA games especially appealing because they offer immersive narratives, clear missions, and rapid reward cycles that provide a temporary sense of escape, competence, and progression [84]. When real life feels unsatisfying, completing quests, advancing storylines, and gaining in-game achievements can serve as substitutes for the meaning and accomplishments they struggle to experience in daily life, making these games particularly attractive [85].

Higher MOBA addiction was associated with poorer character and virtue. This relationship may be explained by the highly competitive and performance-driven nature of these games, which emphasize winning, dominance, and skill mastery rather than empathy or moral reflection [86]. Prior research has shown that many MOBA players experience in-game hostility and toxic communication [87], which may erode prosocial tendencies and empathy—core elements of character and virtue. Higher MOBA addiction was also associated with a lower sense of meaning and purpose. Similar to AA games, MOBA games provide clear goals (eg, destroying the enemy nexus) and immediate rewards (eg, rank increases after winning) [88], which are often more attainable than real-life achievements. When adolescents repeatedly turn to these games for accomplishment, real-world goals such as academic progress

XSL•FO
RenderX

or personal growth may feel less rewarding or more difficult to pursue [82], leading to fewer meaningful experiences and a weakened sense of purpose [89]. It is also plausible that adolescents who already exhibit lower character and virtue or a reduced sense of meaning and purpose are more vulnerable to MOBA addiction, as these game-design features may temporarily compensate for these deficits [90].

Higher SS addiction was associated with fewer close social relationships and poorer mental and physical health. Several mechanisms may account for these associations. First, SS games are typically structured around solitary play, allowing players to advance independently, in contrast to team-based genres that require interpersonal coordination [91]. Such solitary engagement may displace real-world interaction, which is essential for building and maintaining close social relationships [92]. Second, SS games provide a strong sense of autonomy and control (eg, players can design cities, manage resources, and make strategic decisions about budgets and policies) [93], which may be particularly appealing to adolescents who feel powerless in real life due to limited social support [94,95]. Third, adolescents with poorer mental health resulting from real-world challenges may be prone to playing these games, as they allow them to simulate idealized achievements and life scenarios (eg, building perfect cities, running successful businesses, or creating ideal families) [96]. These experiences may temporarily compensate for dissatisfaction with real-life circumstances [82].

The absence of associations between addiction to shooting, sports, casual, strategy, or role-playing games and flourishing may be explained by several factors. First, these genres are highly heterogeneous in their modes of play [97-100]. For example, shooting games may involve competitive player-versus-player matches, cooperative missions against computer-controlled opponents, or single-player narrative progression [100]. Sports games can be played through team-on-team matches or career modes that simulate long-term athlete development [100]. Casual, strategy, and role-playing games can be played in either solitary or highly social modes [97-99]. This variability means that potential positive and negative associations with flourishing at the game level may offset one another at the genre level. Second, many games in these genres feature short matches or rounds, which may make disengagement easier compared with genres that lack natural stopping points [101]. This, in turn, reduces the likelihood of prolonged immersion that displaces real-life activities essential for flourishing [102,103].

Implications

The findings of this study have several practical implications. First, higher addiction to shooting, sports, casual, strategy, or role-playing games was not associated with poorer flourishing outcomes. This suggests that encouraging adolescents to spend their gaming time within these relatively lower-risk genres may not increase the likelihood of adverse outcomes associated with GA. This does not imply that addiction to such games is harmless. Rather, when gaming is already part of adolescents' lives, guiding them toward lower-risk genres may represent a more realistic harm-reduction strategy. Second, when adolescents are highly engaged in AA, MOBA, or SS games, targeted strategies may be required to reduce addiction-related risks. For parents, setting clear time limits and enforcing mandatory breaks may help prevent prolonged, immersive play. Conversations that unpack why certain games feel particularly absorbing may also increase adolescents' awareness of risky game design features, thereby empowering them to better self-regulate their play. Providing alternative activities that offer similar psychological rewards may also be effective. For example, adolescents drawn to SS games may find comparable satisfaction in robotics clubs, art workshops, or sandbox-style board games, which can fulfill needs for autonomy and creativity. MOBA players may benefit from real-world team sports, which can satisfy desires for competition, collaboration, and skill mastery. Third, for policy makers, rather than imposing one-size-fits-all rules, policies could be more targeted toward high-risk game genres. For instance, China's current "Online Game Anti-Indulgence System" restricts adolescents to a single hour of play on weekend evenings across all games (although many adolescents have learned ways to circumvent this rule). A more nuanced policy could impose stricter limits on high-risk genres while allowing more time for genres associated with fewer negative health outcomes. Such differentiation could better balance adolescents' enjoyment of gaming with the goal of minimizing potential harm to their well-being.

Limitations

Several limitations should be considered when interpreting the findings. First, the results may be influenced by the genre classification approach adopted in this study. Although the categorization was informed by established frameworks, substantial variation may still exist among games within the same genre with respect to design features that could lead to distinct associations with health-related outcomes. Future research would benefit from developing or validating a genre classification system specifically optimized for health research. Such a system could reduce within-genre heterogeneity and facilitate more robust conclusions regarding genre-specific associations with adolescent health. Second, the cross-sectional design of this study limits causal inference between GA and flourishing. The *E* value analysis also suggested that some associations could potentially be explained by unmeasured confounders of small magnitude. To strengthen causal inference, future studies could employ longitudinal and experimental designs to clarify whether genre-specific GA undermines flourishing, whether adolescents with lower flourishing are more likely to develop addiction to particular genres, or whether these processes operate reciprocally. Third, due to practical constraints, GA was measured using single-item self-reports, which are inherently less precise than multi-item scales and may have introduced measurement error. Future studies should use validated multi-item measures when feasible.

Conclusions

This study provides novel evidence that the association between GA and adolescent flourishing varies by game genre. In contrast to prior work that treats health as a narrow or unidimensional construct, the adoption of a multidimensional health framework offers a more nuanced understanding of gaming-related risks. The findings contribute to the field by indicating that addiction

XSL•FO
RenderX

to AA, MOBA, and SS games poses greater risks to adolescent health than addiction to other genres. Accordingly, prevention, education, and policy efforts should prioritize these higher-risk genres to more effectively support adolescent health.

Acknowledgments

We gratefully acknowledge the teachers who contributed to data collection for this study. We thank the research assistants for their support in digitizing the paper questionnaires. We are also grateful to Dr. Dorota W ziak-Białowolska for providing the Chinese version of the Human Flourishing Index. The funder had no role in the study design, data collection, analysis, interpretation, or manuscript preparation. We declare the use of generative artificial intelligence in the research and writing process. According to GAIDeT (Generative AI Delegation Taxonomy) (2025), the following tasks were delegated to generative artificial intelligence tools under full human supervision: proofreading and editing. The generative artificial intelligence tool used was ChatGPT-5.0 (OpenAI). Responsibility for the final manuscript lies entirely with the authors. Generative artificial intelligence tools are not listed as authors and do not bear responsibility for the outcomes.

Funding

This work was supported by the Shanghai Jiao Tong University Young Scholars Program in Humanities (grant 2025QN029) and the Shanghai Jiao Tong University Research Start-up Fund (grant WH220525103).

Data Availability

The datasets generated or analyzed during this study are not publicly available because they will be used for secondary analyses in other ongoing studies, but are available from the corresponding author (YL) upon reasonable request.

Authors' Contributions

Conceptualization: Y Luo
Data curation: Y Luo, XC, Y Liu, WH, MF
Formal analysis: Y Luo
Funding acquisition: Y Luo
Investigation: Y Luo
Methodology: Y Luo
Project administration: Y Luo
Resources: Y Luo
Software: Y Luo
Supervision: Y Luo
Visualization: Y Luo
Writing – original draft: Y Luo
Writing – review & editing: Y Luo, XC, Y Liu, WH, MF, DT, SM

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey questionnaire.
[DOCX File , 37 KB - jmir_v28i1e89319_app1.docx]

Multimedia Appendix 2
Data processing.
[DOCX File , 22 KB - jmir_v28i1e89319_app2.docx]

Multimedia Appendix 3
Results of the robust linear regression analyses examining the associations between gaming addiction and flourishing outcomes using complete-case data.
[DOCX File , 22 KB - jmir_v28i1e89319_app3.docx]

Multimedia Appendix 4
E values for the significant pooled effects from the robust linear regression analyses based on multiple imputed data.

XSL•FO
RenderX

[DOCX File , 17 KB - jmir_v28i1e89319_app4.docx]

References

1. Granic I, Lobel A, Engels RCME. The benefits of playing video games. Am Psychol 2014 Jan;69(1):66-78. [doi: 10.1037/a0034857] [Medline: 24295515]
2. Gao Y, Wang J, Dong G. The prevalence and possible risk factors of internet gaming disorder among adolescents and young adults: systematic reviews and meta-analyses. Journal of Psychiatric Research 2022 Oct;154:35-43. [doi: 10.1016/j.jpsychires.2022.06.049]
3. Kim HS, Son G, Roh E, Ahn W, Kim J, Shin S, et al. Prevalence of gaming disorder: a meta-analysis. Addict Behav 2022 Mar;126:107183. [doi: 10.1016/j.addbeh.2021.107183] [Medline: 34864436]
4. Byeon G, Jo S, Park J, Jeong H, Lee HK, Yim HW. Risk factors and outcomes of internet gaming disorder identified in Korean prospective adolescent cohort study. J Behav Addict 2022 Dec 27;11(4):1035-1043 [FREE Full text] [doi: 10.1556/2006.2022.00071] [Medline: 36194504]
5. World Health Organization (WHO). Constitution of the World Health Organization. Am J Public Health Nations Health 1946 Nov;36(11):1315-1323. [doi: 10.2105/ajph.36.11.1315]
6. VanderWeele TJ. On the promotion of human flourishing. Proc Natl Acad Sci U S A 2017 Aug 01;114(31):8148-8156. [doi: 10.1073/pnas.1702996114] [Medline: 28705870]
7. van der Neut D, Peeters M, Boniel-Nissim M, Klanšček HJ, Oja L, van den Eijnden R. A cross-national comparison of problematic gaming behavior and well-being in adolescents. J Behav Addict 2023 Jun 29;12(2):448-457 [FREE Full text] [doi: 10.1556/2006.2023.00010] [Medline: 37060555]
8. Teng Z, Pontes HM, Nie Q, Xiang G, Griffiths MD, Guo C. Internet gaming disorder and psychosocial well-being: a longitudinal study of older-aged adolescents and emerging adults. Addict Behav 2020 Nov;110:106530. [doi: 10.1016/j.addbeh.2020.106530] [Medline: 32683173]
9. Arslan G, Jarden A, Yıldırım M, Jarden R, Silapurem L. Longitudinal associations between life events, hope, life satisfaction, happiness, and depressive symptoms. Cogn Ther Res 2025 Jul 09:e1 (forthcoming). [doi: 10.1007/s10608-025-10630-0]
10. Freire T, Ferreira G. Health-related quality of life of adolescents: relations with positive and negative psychological dimensions. International Journal of Adolescence and Youth 2016 Dec 23;23(1):11-24. [doi: 10.1080/02673843.2016.1262268]
11. Freire T, Ferreira G. Do I need to be positive to be happy? Considering the role of self-esteem, life satisfaction, and psychological distress in Portuguese adolescents' subjective happiness. Psychol Rep 2020 Aug 29;123(4):1064-1082. [doi: 10.1177/0033294119846064] [Medline: 31033417]
12. López-Pérez B, Fernández-Castilla B. Children's and adolescents' conceptions of happiness at school and its relation with their own happiness and their academic performance. J Happiness Stud 2017 Jun 7;19(6):1811-1830. [doi: 10.1007/s10902-017-9895-5]
13. Shin W, Jyung M, Choi I, Sul S. Perceived financial well-being and its association with frontostriatal functional connectivity, real-life anticipatory experiences, and everyday happiness. Sci Rep 2023 Oct 31;13(1):18739 [FREE Full text] [doi: 10.1038/s41598-023-44001-0] [Medline: 37907524]
14. Wartberg L, Kriston L, Kammerl R. Associations of social support, friends only known through the internet, and health-related quality of life with internet gaming disorder in adolescence. Cyberpsychol Behav Soc Netw 2017 Jul;20(7):436-441. [doi: 10.1089/cyber.2016.0535] [Medline: 28715266]
15. Zhang L, Han J, Liu M, Yang C, Liao Y. The prevalence and possible risk factors of gaming disorder among adolescents in China. BMC Psychiatry 2024 May 21;24(1):381 [FREE Full text] [doi: 10.1186/s12888-024-05826-9] [Medline: 38773555]
16. Liu Y, Gong R, Yu Y, Xu C, Yu X, Chang R, et al. Longitudinal predictors for incidence of internet gaming disorder among adolescents: the roles of time spent on gaming and depressive symptoms. Journal of Adolescence 2021 Oct;92(1):1-9. [doi: 10.1016/j.adolescence.2021.06.008]
17. Canale N, Elgar FJ, Pivetta E, Galeotti T, Marino C, Billieux J, et al. Problem gaming and adolescents' health and well-being: evidence from a large nationally representative sample in Italy. Computers in Human Behavior 2025 Jul;168:108644. [doi: 10.1016/j.chb.2025.108644]
18. Falcione K, Weber R. Psychopathology and gaming disorder in adolescents. JAMA Netw Open 2025 Jul 01;8(7):e2528532 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.28532] [Medline: 40728787]
19. Yu Y, Wu AMS, Fong VWI, Zhang J, Li J, Lau JTF. Association between internet gaming disorder and suicidal ideation mediated by psychosocial resources and psychosocial problems among adolescent internet gamers in China: cross-sectional study. JMIR Serious Games 2024 Sep 19;12:e48439 [FREE Full text] [doi: 10.2196/48439] [Medline: 39298753]
20. Ferreira FDM, Bambini B, Tonsig G, Fonseca L, Picon FA, Pan PM, et al. Predictors of gaming disorder in children and adolescents: a school-based study. Braz J Psychiatry 2021;43(3):289-292 [FREE Full text] [doi: 10.1590/1516-4446-2020-0964] [Medline: 32756806]
21. Coutelle R, Balzer J, Rolling J, Lalanne L. Problematic gaming, psychiatric comorbidities, and adolescence: a systematic review of the literature. Addict Behav 2024 Oct;157:108091. [doi: 10.1016/j.addbeh.2024.108091] [Medline: 38901145]

XSL•FO
RenderX

22. Meng Y, Shi X, Cai D, Ran M, Ye A, Qiu C. Prevalence, predictive factors, and impacts of internet gaming disorder among adolescents: a population-based longitudinal study. J Affect Disord 2024 Oct 01;362:356-362. [doi: 10.1016/j.jad.2024.06.020] [Medline: 38844168]
23. Rahul R, Kareem SA, Gowda KC, Sajna SJ. Impact of internet gaming disorder on mental well-being: prevalence and psychological distress among adolescents and young adults. Psychiatry Res 2025 Jul;349:116512. [doi: 10.1016/j.psychres.2025.116512] [Medline: 40318564]
24. Zhang W, Pu J, He R, Yu M, Xu L, He X, et al. Demographic characteristics, family environment and psychosocial factors affecting internet addiction in Chinese adolescents. J Affect Disord 2022 Oct 15;315:130-138. [doi: 10.1016/j.jad.2022.07.053] [Medline: 35901990]
25. Jeong EJ, Ferguson CJ, Lee SJ. Pathological gaming in young adolescents: a longitudinal study focused on academic stress and self-control in South Korea. J Youth Adolesc 2019 Dec 20;48(12):2333-2342. [doi: 10.1007/s10964-019-01065-4] [Medline: 31327099]
26. Mohamed NF, Ab Manan N, Muhammad Firdaus Chan MF, Rahmatullah B, Abd Wahab R, Baharudin SNA, et al. The prevalence of internet gaming disorders and the associated psychosocial risk factors among adolescents in Malaysian secondary schools. Clin Child Psychol Psychiatry 2023 Oct 17;28(4):1420-1434. [doi: 10.1177/13591045231164870] [Medline: 36927134]
27. Yu Y, Peng L, Mo PK, Yang X, Cai Y, Ma L, et al. Association between relationship adaptation and internet gaming disorder among first-year secondary school students in China: mediation effects via social support and loneliness. Addictive Behaviors 2022 Feb;125:107166. [doi: 10.1016/j.addbeh.2021.107166]
28. Deng X, Hu Y, Liu C, Li Q, Yang N, Zhang Q, et al. Psychological distress and aggression among adolescents with internet gaming disorder symptoms. Psychiatry Research 2024 Jan;331:115624. [doi: 10.1016/j.psychres.2023.115624]
29. Wartberg L, Kriston L, Kramer M, Schwedler A, Lincoln T, Kammerl R. Internet gaming disorder in early adolescence: associations with parental and adolescent mental health. Eur. psychiatr 2020 Mar 23;43:14-18. [doi: 10.1016/j.eurpsy.2016.12.013]
30. Yu Y, Yang X, Wang S, Wang H, Chang R, Tsamlag L, et al. Serial multiple mediation of the association between internet gaming disorder and suicidal ideation by insomnia and depression in adolescents in Shanghai, China. BMC Psychiatry 2020 Sep 23;20(1):460. [doi: 10.1186/s12888-020-02870-z] [Medline: 32967648]
31. Chamarro A, Díaz-Moreno A, Bonilla I, Cladellas R, Griffiths MD, Gómez-Romero MJ, et al. Stress and suicide risk among adolescents: the role of problematic internet use, gaming disorder and emotional regulation. BMC Public Health 2024 Jan 30;24(1):326. [doi: 10.1186/s12889-024-17860-z]
32. Moshel ML, Warburton W, Thomasius R, Paschke K. Sleep quality as a mediator of internet gaming disorder and executive dysfunction in adolescents: cross-sectional questionnaire study. J Med Internet Res 2025 Jul 09;27:e68571 [FREE Full text] [doi: 10.2196/68571] [Medline: 40632805]
33. Che Mokhtar M, McGee R. Impact of internet addiction and gaming disorder on body weight in children and adolescents: a systematic review. J Paediatrics Child Health 2024 Dec 06;61(2):136-147. [doi: 10.1111/jpc.16726]
34. Männikkö N, Ruotsalainen H, Tolvanen A, Kääriäinen M. Problematic gaming is associated with some health-related behaviors among Finnish vocational school students. Int J Ment Health Addiction 2019 May 17;18(4):993-1007. [doi: 10.1007/s11469-019-00100-6]
35. Zhao H, Li X, Zhou J, Nie Q, Zhou J. The relationship between bullying victimization and online game addiction among Chinese early adolescents: the potential role of meaning in life and gender differences. Children and Youth Services Review 2020 Sep;116:105261. [doi: 10.1016/j.childyouth.2020.105261]
36. Wu AMS, Lei LLM, Ku L. Psychological needs, purpose in life, and problem video game playing among Chinese young adults. Int J Psychol 2013 Apr 16;48(4):583-590. [doi: 10.1080/00207594.2012.658057] [Medline: 22506646]
37. Zhang MX, Wang X, Yu SM, Wu AM. Purpose in life, social support, and internet gaming disorder among Chinese university students: a 1-year follow-up study. Addict Behav 2019 Dec;99:106070. [doi: 10.1016/j.addbeh.2019.106070] [Medline: 31430620]
38. Verrastro V, Albanese CA, Ritella G, Gugliandolo MC, Cuzzocrea F. Empathy, social self-efficacy, problematic internet use, and problematic online gaming between early and late adolescence. Cyberpsychol Behav Soc Netw 2021 Dec 01;24(12):806-814. [doi: 10.1089/cyber.2021.0042] [Medline: 34515543]
39. Yildiz Durak H, KIDIMAN DEMİRHAN E, Citil M. Examining various risk factors as the predictors of gifted and non-gifted high school students' online game addiction. Computers & Education 2022 Feb;177:104378. [doi: 10.1016/j.compedu.2021.104378]
40. Li L, Liang J, Chen S. How autonomy support environment influences student online game addiction: the mediating roles of academic motivation and academic perseverance. IJMHP 2024;26(10):777-790. [doi: 10.32604/ijmhp.2024.056988]
41. Wartberg L, Kriston L, Thomasius R. Internet gaming disorder and problematic social media use in a representative sample of German adolescents: Prevalence estimates, comorbid depressive symptoms and related psychosocial aspects. Computers in Human Behavior 2020 Feb;103:31-36. [doi: 10.1016/j.chb.2019.09.014]
42. Rosales-Navarro C, Torres Pérez I. Parental behaviors and video game addiction in adolescents: a systematic review. Int J Ment Health Addiction 2025 Mar 27:e1 (forthcoming). [doi: 10.1007/s11469-025-01473-7]

XSL•FO
RenderX

43. Torres-Rodríguez A, Griffiths MD, Carbonell X, Oberst U. Internet gaming disorder in adolescence: psychological characteristics of a clinical sample. J Behav Addict 2018 Sep 01;7(3):707-718 [FREE Full text] [doi: 10.1556/2006.7.2018.75] [Medline: 30264606]
44. Jeong H, Yim HW, Lee S, Lee HK, Potenza MN, Lee H. Factors associated with severity, incidence or persistence of internet gaming disorder in children and adolescents: a 2‐year longitudinal study. Addiction 2021 Jan 28;116(7):1828-1838. [doi: 10.1111/add.15366]
45. Butler A, Patte KA, Ferro MA, Leatherdale ST. Interrelationships among depression, anxiety, flourishing, and cannabis use in youth. Addict Behav 2019 Feb;89:206-215. [doi: 10.1016/j.addbeh.2018.10.007] [Medline: 30321693]
46. Blachnio A, Przepiórka A, Gorbaniuk O, McNeill M, Bendayan R, Ben-Ezra M, et al. Lonely online: a social model of digital media addiction: a study in 21 countries. The Social Science Journal 2024 Jan 10;62(3):717-728. [doi: 10.1080/03623319.2023.2228439]
47. Okur S, Engin M, Kütük H, Satıcı SA. Cross-lagged relations between smartphone addiction and flourishing in adolescents. Personality and Individual Differences 2025 Apr;236:113008. [doi: 10.1016/j.paid.2024.113008]
48. Satici SA, Uysal R. Well-being and problematic Facebook use. Computers in Human Behavior 2015 Aug;49:185-190. [doi: 10.1016/j.chb.2015.03.005]
49. Uysal R. The predictive roles of social safeness and flourishing on problematic Facebook use. South African Journal of Psychology 2014 Nov 21;45(2):182-193. [doi: 10.1177/0081246314560010]
50. Diener E, Wirtz D, Tov W, Kim-Prieto C, Choi D, Oishi S, et al. New well-being measures: short scales to assess flourishing and positive and negative feelings. Soc Indic Res 2009 May 28;97(2):143-156. [doi: 10.1007/s11205-009-9493-y]
51. Nakagomi A, Ide K, Kondo K, Shiba K. Digital gaming and subsequent health and well-being among older adults: longitudinal outcome-wide analysis. J Med Internet Res 2025 Jan 27;27:e69080 [FREE Full text] [doi: 10.2196/69080] [Medline: 39869904]
52. Park JH, Han DH, Kim B, Cheong JH, Lee Y. Correlations among social anxiety, self-esteem, impulsivity, and game genre in patients with problematic online game playing. Psychiatry Investig 2016 May;13(3):297-304 [FREE Full text] [doi: 10.4306/pi.2016.13.3.297] [Medline: 27247595]
53. Fraser AM, Padilla-Walker LM, Coyne SM, Nelson LJ, Stockdale LA. Associations between violent video gaming, empathic concern, and prosocial behavior toward strangers, friends, and family members. J Youth Adolesc 2012 May 1;41(5):636-649. [doi: 10.1007/s10964-012-9742-2] [Medline: 22302216]
54. Greitemeyer T. The dark and bright side of video game consumption: effects of violent and prosocial video games. Curr Opin Psychol 2022 Aug;46:101326 [FREE Full text] [doi: 10.1016/j.copsyc.2022.101326] [Medline: 35366605]
55. Greitemeyer T, Osswald S. Effects of prosocial video games on prosocial behavior. J Pers Soc Psychol 2010 Feb;98(2):211-221. [doi: 10.1037/a0016997] [Medline: 20085396]
56. Whitaker JL, Bushman BJ. “Remain calm. Be kind.” Effects of relaxing video games on aggressive and prosocial behavior. Social Psychological and Personality Science 2011 May 23;3(1):88-92. [doi: 10.1177/1948550611409760]
57. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: the APA Publications and Communications Board Task Force report. Am Psychol 2018 Jan;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
58. Faul F, Erdfelder E, Buchner A, Lang A. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
59. Lemmens JS, Valkenburg PM, Peter J. Development and validation of a game addiction scale for adolescents. Media Psychology 2009 Mar 05;12(1):77-95. [doi: 10.1080/15213260802669458]
60. Lemmens JS, Valkenburg PM, Gentile DA. The Internet Gaming Disorder Scale. Psychol Assess 2015 Jun;27(2):567-582. [doi: 10.1037/pas0000062] [Medline: 25558970]
61. King DL, Delfabbro PH, Griffiths MD. Trajectories of problem video gaming among adult regular gamers: an 18-month longitudinal study. Cyberpsychol Behav Soc Netw 2013 Jan;16(1):72-76. [doi: 10.1089/cyber.2012.0062] [Medline: 23098213]
62. Najavits L, Lung J, Froias A, Paull N, Bailey G. A study of multiple behavioral addictions in a substance abuse sample. Subst Use Misuse 2014 Mar 04;49(4):479-484. [doi: 10.3109/10826084.2013.858168] [Medline: 24304172]
63. Vogel EA, Prochaska JJ, Rubinstein ML. Measuring e-cigarette addiction among adolescents. Tob Control 2020 May 11;29(3):258-262 [FREE Full text] [doi: 10.1136/tobaccocontrol-2018-054900] [Medline: 31079033]
64. van den Eijnden RJ, Lemmens JS, Valkenburg PM. The Social Media Disorder Scale. Computers in Human Behavior 2016 Aug;61:478-487. [doi: 10.1016/j.chb.2016.03.038]
65. W ziak-Białowolska D, McNeely E, VanderWeele TJ. Human flourishing in cross cultural settings. Evidence from the United States, China, Sri Lanka, Cambodia, and Mexico. Front Psychol 2019 May 29;10:1269 [FREE Full text] [doi: 10.3389/fpsyg.2019.01269] [Medline: 31191421]
66. Hyun GJ, Han DH, Lee YS, Kang KD, Yoo SK, Chung U, et al. Risk factors associated with online game addiction: a hierarchical model. Computers in Human Behavior 2015 Jul;48:706-713. [doi: 10.1016/j.chb.2015.02.008]

XSL•FO
RenderX

67. Moreno M, Riddle K, Jenkins MC, Singh AP, Zhao Q, Eickhoff J. Measuring problematic internet use, internet gaming disorder, and social media addiction in young adults: cross-sectional survey study. JMIR Public Health Surveill 2022 Jan 27;8(1):e27719 [FREE Full text] [doi: 10.2196/27719] [Medline: 34081596]
68. Toker S, Baturay MH. Antecedents and consequences of game addiction. Computers in Human Behavior 2016 Feb;55:668-679. [doi: 10.1016/j.chb.2015.10.002]
69. VanderWeele TJ. Principles of confounder selection. Eur J Epidemiol 2019 Mar 6;34(3):211-219 [FREE Full text] [doi: 10.1007/s10654-019-00494-6] [Medline: 30840181]
70. R Core Team. R: a language and environment for statistical computing. R Foundation. URL: https://www.R-project.org [accessed 2026-01-02]
71. Maechler M, Rousseeuw P, Croux C. Package 'robustbase'. R Foundation. 2021. URL: https://cran.r-project.org/web/packages/robustbase/index.html [accessed 2026-02-09]
72. Mathur MB, Ding P, Riddell CA, VanderWeele TJ. Web site and R package for computing E-values. Epidemiology 2018;29(5):e45-e47. [doi: 10.1097/ede.0000000000000864]
73. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. Ann Intern Med 2017 Aug 15;167(4):268-274. [doi: 10.7326/M16-2607] [Medline: 28693043]
74. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts. BMC Med Res Methodol 2017 Dec 06;17(1):162 [FREE Full text] [doi: 10.1186/s12874-017-0442-1] [Medline: 29207961]
75. Van BS, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. Journal of Statistical Software 2011;45:1-67 [FREE Full text] [doi: 10.18637/jss.v045.i03]
76. James G, Witten D, Hastie T, Tibshirani R, Taylor J. An Introduction to Statistical Learning: With Applications in Python. Cham, Switzerland: Springer; 2023.
77. Hughes N. Understanding specific gaming experiences: the case of open world games. Doctor of Philosophy thesis. University of York. White Rose eThesis Online. 2023. URL: https://etheses.whiterose.ac.uk/id/eprint/33608/1/Hughes_CorrectedThesisClean.pdf [accessed 2025-11-24]
78. Buecker S, Simacek T, Ingwersen B, Terwiel S, Simonsmeier BA. Physical activity and subjective well-being in healthy individuals: a meta-analytic review. Health Psychol Rev 2021 Dec 10;15(4):574-592. [doi: 10.1080/17437199.2020.1760728] [Medline: 32452716]
79. Kroencke L, Harari GM, Back MD, Wagner J. Well-being in social interactions: examining personality-situation dynamics in face-to-face and computer-mediated communication. J Pers Soc Psychol 2023 Feb;124(2):437-460. [doi: 10.1037/pspp0000422] [Medline: 35834202]
80. Imbert D. What makes video games engaging: deep dive into the design, mechanics and psychology of games. Bachelor's thesis. University of Applied Sciences. Theseus. URL: https://www.theseus.fi/bitstream/handle/10024/797212/Daniel_Imbert.pdf?sequence=2&isAllowed=y [accessed 2025-11-24]
81. Rieger D, Wulf T, Kneer J, Frischlich L, Bente G. The winner takes it all: the effect of in-game success and need satisfaction on mood repair and enjoyment. Computers in Human Behavior 2014 Oct;39:281-286. [doi: 10.1016/j.chb.2014.07.037]
82. Chang S, Hsieh GM, Lin SS. The mediation effects of gaming motives between game involvement and problematic internet use: escapism, advancement and socializing. Computers & Education 2018 Jul;122:43-53. [doi: 10.1016/j.compedu.2018.03.007]
83. Allen JJ, Anderson CA. Satisfaction and frustration of basic psychological needs in the real world and in video games predict internet gaming disorder scores and well-being. Computers in Human Behavior 2018 Jul;84:220-229. [doi: 10.1016/j.chb.2018.02.034]
84. Anto A, Basu A, Selim R, Foscht T, Eisingerich AB. Open-world games' affordance of cognitive escapism, relaxation, and mental well-being among postgraduate students: mixed methods study. J Med Internet Res 2024 Dec 17;26:e63760 [FREE Full text] [doi: 10.2196/63760] [Medline: 39689301]
85. Bender PK, Gentile DA. Internet gaming disorder: relations between needs satisfaction in-game and in life in general. Psychology of Popular Media 2020 Apr;9(2):266-278. [doi: 10.1037/ppm0000227]
86. Xia T, Lin X, Mo X, Su Q, Ding S. Players' continuous willingness to play in MOBA game ranking mode: through the lens of self-determination theory and social comparison theory. Humanit Soc Sci Commun 2024 Oct 22;11(1):1398. [doi: 10.1057/s41599-024-03934-1]
87. Kordyaka B, Laato S, Weber S, Niehaves B. What constitutes victims of toxicity - identifying drivers of toxic victimhood in multiplayer online battle arena games. Front Psychol 2023 Jun 16;14:1193172 [FREE Full text] [doi: 10.3389/fpsyg.2023.1193172] [Medline: 37397342]
88. Morris R. MOBA games: a critical analysis into what makes and breaks them, and what influences players to keep returning. BSc Games Programming thesis. University of London. ResearchGate. 2019 May. URL: https://www.researchgate.net/profile/Rees-Morris/publication/344483735_MOBA_Games_A_Critical_Analysis_Into_What_Makes_And_Breaks_Them_And_What_Influences_Players_to_Keep_Returning/links/5f7b9fd5a6fdccfd7b4a6ee9/

XSL•FO
RenderX

MOBA-Games-A-Critical-Analysis-Into-What-Makes-And-Breaks-Them-And-What-Influences-Players-to-Keep-Returning.pdf [accessed 2025-11-24]
89. Bu L, Ding B. A comparative study of online gaming between high-performing and low-performing students-a case from China. Heliyon 2024 Jan 30;10(2):e24291 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e24291] [Medline: 38293356]
90. T'ng ST, Ho KH, Pau K. Need frustration, gaming motives, and internet gaming disorder in mobile multiplayer online battle arena (MOBA) games: through the lens of self-determination theory. Int J Ment Health Addict 2022 Apr 25;21(6):1-21 [FREE Full text] [doi: 10.1007/s11469-022-00825-x] [Medline: 35497075]
91. Hooper J. Freeplay in videogames. CORE. URL: https://files01.core.ac.uk/download/pdf/196607253.pdf [accessed 2025-11-24]
92. Pinker S. The Village Effect: How Face-to-Face Contact Can Make Us Healthier and Happier. Toronto, ON, Canada: Random House Canada; 2015.
93. Zhong Y, Fryer LK, Zheng S, Shum A, Chu SKW. The power of play: integrating competitive sandbox game for experiential learning to foster twenty-first century skills. Int J Educ Technol High Educ 2025 Jun 06;22(1):34. [doi: 10.1186/s41239-025-00528-y]
94. An G, Kim M. Powerlessness, social support, and glycemic control in Korean adults with type 2 diabetes. Contemporary Nurse 2014 Dec 17;42(2):272-279. [doi: 10.5172/conu.2012.42.2.272] [Medline: 23181377]
95. Lammers J, Stoker JI, Rink F, Galinsky AD. To have control over or to be free from others? The desire for power reflects a need for autonomy. Pers Soc Psychol Bull 2016 Apr 16;42(4):498-512. [doi: 10.1177/0146167216634064] [Medline: 26984014]
96. Rahimi S, Walker JT, Lin-Lipsmeyer L, Shin J. Toward defining and assessing creativity in sandbox games. Creativity Research Journal 2023 Jan 03;36(2):194-212. [doi: 10.1080/10400419.2022.2156477]
97. Nielsen JL, Jensen BF, Mahlmann T, Togelius J, Yannakakis GN. AI for general strategy game playing. In: Handbook of Digital Games. Hoboken, NJ: Wiley; 2014:13-46.
98. Deterding S, Zagal J. Role-Playing Game Studies: Transmedia Foundations. New York, NY: Routledge; 2018.
99. Johnson MR. Casual games before casual games: historicizing paper puzzle games in an era of digital play. Games and Culture 2018 Aug 01;14(2):119-138. [doi: 10.1177/1555412018790423]
100. Wolf MJP, Perron B. The Routledge Companion to Video Game Studies. Abingdon, UK: Taylor & Francis; 2023.
101. Alexandrovsky D, Gerling K, Opp MS, Hahn CB, Birk MV, Alsheail M. Disengagement from games: characterizing the experience and process of exiting play sessions. Proc ACM Hum-Comput Interact 2024 Oct 15;8(CHI PLAY):1-27. [doi: 10.1145/3677066]
102. Stieger S, Lewetz D, Willinger D. Face-to-face more important than digital communication for mental health during the pandemic. Sci Rep 2023 May 17;13(1):8022 [FREE Full text] [doi: 10.1038/s41598-023-34957-4] [Medline: 37198196]
103. Lee PSN, Leung L, Lo V, Xiong C, Wu T. Internet communication versus face-to-face interaction in quality of life. Soc Indic Res 2010 Oct 31;100(3):375-389. [doi: 10.1007/s11205-010-9618-3]

Abbreviations

AA: action and adventure
GA: gaming addiction
HFI: Human Flourishing Index
JARS-Quant: Journal Article Reporting Standards for Quantitative Research
MOBA: multiplayer online battle arena
SS: sandbox and simulation

Edited by S Brini; submitted 10.Dec.2025; peer-reviewed by PS Liao; comments to author 13.Jan.2026; revised version received 03.Feb.2026; accepted 04.Feb.2026; published 12.Feb.2026.

Please cite as:
Luo Y, Chen X, Liu Y, Hong W, Fan M, Thomas D, Mullen SP
Genre-Specific Gaming Addiction and Flourishing in Adolescents: Cross-Sectional Survey Study
J Med Internet Res 2026;28:e89319
URL: https://www.jmir.org/2026/1/e89319
doi:10.2196/89319
PMID:

©Yan Luo, Xiaofei Chen, Yaoyu Liu, Wanjia Hong, Miner Fan, David Thomas, Sean P Mullen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Feb.2026. This is an open-access article distributed under the

XSL•FO
RenderX

terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Implementation of Reference Framework for Preventive Care for Older Adults: Cross-sectional Qualitative Study

Claire Chenwen Zhong[1,2]; Mingtao Chen[1]; Chung Yi Lo[1]; Man Kin YIM[1]; Xiaoshu Zhang[1]; William CW Wong[3]; Junjie Huang[1,2]; Martin CS Wong[1,2]

[1]JC School of Public Health and Primary Care, Faculty of Medicine, Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

[2]Centre for Health Education and Health Promotion, Faculty of Medicine, Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

[3]Department of Family Medicine & Primary Care, LKS Faculty of Medicine, University of Hong Kong, Hong Kong, China (Hong Kong)

Corresponding Author:
Martin CS Wong
JC School of Public Health and Primary Care
Faculty of Medicine
Chinese University of Hong Kong
4/F, JC School of Public Health
Prince of Wales Hospital
Hong Kong,
China (Hong Kong)
Phone: 852 22528782
Email: wong_martin@cuhk.edu.hk

Abstract

Background: The Hong Kong Government developed the Reference Framework for Preventive Care for Older Adults (RF) to provide evidence-based recommendations for preventive care in the primary care setting. However, no study has systematically evaluated the adoption and awareness of this framework among primary care physicians (PCPs).

Objective: This study aimed to explore PCPs' perceptions of the RF and identify barriers and facilitators to its adoption in primary care settings.

Methods: This cross-sectional qualitative study involved interviews with 40 PCPs in Hong Kong to assess their views on the RF's adoption and implementation. The Consolidated Framework for Implementation Research (CFIR) was used to guide the analysis, focusing on five domains: intervention characteristics, outer setting, inner setting, individual characteristics, and implementation process.

Results: Relative advantage (65%, n = 26), innovation design (45%, n = 18), and access to knowledge and information (70%, n = 28) were the facilitators that mostly discussed by interviewee. By contrast, interviewees also raised concerns regarding policy and legal (n=29, 73%) and availability of resources (n=29, 73%) in the implementation of the RF. In general, participants acknowledged the RF's evidence-based foundation and comprehensive design, appreciating its potential to improve care for older adults. However, barriers included the complexity and limited personalization of the RF, competing clinical priorities, time constraints, and resource limitations. Government support and policy initiatives facilitated engagement, but clearer integration into existing workflows and stronger promotion were needed. Tailored updates, user feedback, and technological enhancements were seen as essential for improving the RF's usability and ensuring its relevance in clinical practice.

Conclusions: This study highlights that while the RF has potential to improve preventive care in Hong Kong's primary care setting, its adoption is constrained by systemic, organizational, and individual barriers. To ensure its successful integration, flexible implementation, institutional support, tailored incentives, and enhanced clinician and patient engagement are essential. These findings offer practical implications for policymakers and healthcare practitioners to refine and promote the RF in routine consultations, and underscore the need for future research to test theory-informed implementation strategies. Furthermore, this study offered novel contributions to the CFIR literatures in systematically investigate physician perspectives on RF for the older adult population within the distinct socio-cultural and structural context of Hong Kong, COVID-19 pandemic impact to the current healthcare system and actionable implementation strategies for Asian populations. Applying the findings from this study, the CFIR-ERIC Matching Tool could be used to address the specific barriers identified in this study and improving implementation outcomes across various healthcare settings.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e78814) doi:10.2196/78814

KEYWORDS

primary care; guideline adherence; geriatrics; implementation science; Consolidated Framework for Implementation Research

Introduction

The Hong Kong population has encountered a serious aging burden, with a higher old-age dependency ratio and old-age population proportion than other East Asian districts such as Macau and Singapore [1]. It has been estimated that the proportion of residents aged 65 years or older will increase from 1.45 million (20.5%) in 2021 to more than 2.74 million (36%) in 2046 [2,3]. Apart from implications for policy and the provision of social services, this increase could exert a substantial influence on the demand for health care services, including a higher population prevalence of multimorbidity [4]. A systematic review reported that multimorbidity has affected 76.5% of Chinese older adults [5], and its disease burden continues to escalate in aging communities [6]. The challenge posed by these observations is compounded by the increasing prevalence of disabilities, functional decline, cognitive impairment, mood problems, and lack of social support among the older adult population [7], imposing a substantial burden on patients, their caregivers, and the health care system.

A significant body of evidence shows that adherence to guideline-directed preventive interventions can be translated into clinical benefits [8,9]. In addition, evidence-based guidelines can improve public health outcomes by informing physicians, allied health professionals, and policy makers to practice and design tailored health programs in the most effective way.

The Primary Healthcare Office of the Department of Health, the Hong Kong Government, has produced a Reference Framework for Preventive Care for Older Adults (hereafter "RF"), dedicated to providing recommendations for evidence-based preventive care for the older adult population in the primary care setting. A Task Force on Conceptual Model and Preventive Protocols (hereafter referred to as the "Task Force") was established under the Working Group on Primary Care [10]. The Task Force synthesized recommendations based on the latest evidence for adoption in primary care settings in Hong Kong. A Clinical Advisory Group, consisting of academic professors, stakeholders from professional organizations, experts from the primary care sector, and advocacy groups, was assembled to contribute to and support guideline development. One reference framework is dedicated specifically to preventive care for older adults [7]. The Task Force is charged with promulgating, maintaining, and revising the framework content and promoting its adoption by primary care physicians (PCPs). The RF contains a core document accompanied by 6 modules on specific areas, including (1) health assessment; (2) falls in older adults; (3) dental health care; (4) visual impairment; (5) mental health problems; and (6) cognitive impairment. The RF adopts a life-course approach and was developed with reference to 2 existing reference frameworks on diabetes and hypertension [11,12]. The RF also highlights key success factors for implementation, including adaptability to local environments, needs, and structures. It was anticipated that the adoption of the RF would improve patient care through health promotion, disease prevention, and addressing major health risks among older adults [7].

It is crucial to devise and implement strategies that could enhance health promotion, disease prevention, and preservation of functional ability among older adults [10]. Active aging has been recognized as multidimensional, encompassing a low risk of disease, high functional capacity, and active engagement in one's life [13]. With increasing age, lifestyle factors become more important determinants of active aging. Hence, engagement of older adults in disease-preventive activities is of paramount importance.

PCPs are in a privileged position to act as a first point of contact, providing comprehensive, continuous, and community-oriented care in a coordinated manner. They offer health education, risk assessment, and regular care for common chronic diseases [14]. These roles also include support for family members and caregivers, as well as referral to a specialist and community care when applicable. The effectiveness of disease prevention in preserving functional ability among the older adult population has been supported by a substantial body of evidence. For instance, a systematic review of randomized controlled trials summarizing the impact of interventions aimed at preventing frailty progression showed that group exercise programs and cognitive training could successfully reduce or delay frailty [15]. These preventive programs were also found to exert favorable effects on frailty indicators with good value for money, especially among very frail community-dwelling participants.

Several implementation science models are available. The Consolidated Framework for Implementation Research (CFIR) was particularly well-suited for this study because of its comprehensive and flexible nature, which allows for exploration of multiple factors influencing the adoption and implementation of the RF. Unlike other implementation science models that focus on specific domains or stages of implementation—for instance, the Theoretical Domains Framework, which emphasizes individual behavior change, or the Reach, Effectiveness, Adoption, Implementation, and Maintenance (RE-AIM) framework, which focuses on reach and effectiveness [16,17]—CFIR provides a holistic approach by integrating 5 key domains: Intervention Characteristics, Outer Setting, Inner Setting, Characteristics of Individuals, and the Implementation Process [18]. This breadth enables CFIR to capture the dynamic interplay among the RF's design, the health care system's infrastructure, external influences, and individual practitioner behaviors. Its adaptability to varying contexts makes CFIR especially relevant for understanding barriers and facilitators to implementing a complex intervention such as the RF in diverse health care settings [18], where factors including resource constraints, time limitations, and local needs are critical. Thus, CFIR's ability to identify patterns across these

XSL•FO
RenderX

domains provides actionable insights that are essential for improving the RF's adoption and impact.

To our knowledge, no study has systematically assessed this RF or examined the level of awareness and adoption of the RF among PCPs, as well as perceived facilitators and barriers to its implementation. This study aims to explore the implementation of the RF in Hong Kong's primary care settings, using the CFIR to identify key facilitators and barriers.

Methods

Study Design

The study employed a cross-sectional qualitative design involving in-depth individual interviews with 40 PCPs. A sample size of 10-50 participants has been recommended in previous qualitative studies, or until data saturation is reached [19-22], to ensure a comprehensive understanding of the RF. This study was designed, conducted, and reported in accordance with the Standards for Reporting Qualitative Research (SRQR; see Multimedia Appendix 2) [23].

This study was conducted within a constructivist paradigm, which posits that knowledge is socially constructed through interactions between individuals and shaped by context. To explore the implementation of the RF, we employed a 2-stage qualitative analysis. First, we conducted a thematic analysis to inductively identify patterns and emergent themes from the interview data. Subsequently, we applied a framework analysis guided by the CFIR to systematically organize and interpret the data within a structured implementation science framework. This approach allowed for both inductive exploration and deductive mapping of findings to established constructs.

Participant Recruitment

Participants were recruited between February and September 2024 through emails or phone calls. Eligible participants were (1) PCPs who provide preventive care to older adults as part of their routine clinical practice and (2) those who had read the core document and relevant modules of the RF before the interviews. PCPs who did not consult older adult patients in routine practice were excluded. No restrictions were placed on years of clinical experience to ensure a broad range of viewpoints. Eight distinct groups of PCPs were invited to share their experiences with the adoption of the RF. These groups included (1) PCPs employed by health maintenance organizations; (2) PCPs in group practices within the private sector but not affiliated with health maintenance organizations; (3) solo practitioners providing care independently in private clinics; (4) general outpatient clinic physicians working in publicly funded clinics; (5) physicians at Family Medicine Specialist Clinics; (6) Fellows of the Hong Kong College of Family Physicians who are actively involved in teaching and provide preventive care for older adults; (7) PCPs serving as referral partners for District Health Centers; and (8) physicians with formal qualifications in gerontology or geriatric medicine recognized by the Medical Council of Hong Kong. A purposive sampling method was employed to recruit participants from the 8 predefined PCP groups, ensuring comprehensive representation across diverse practice settings.

For groups 1-3, participants were recruited through established professional networks under the Hong Kong Medical Association. For groups 4 and 5, we contacted the Chiefs of Service across all 7 Hospital Authority clusters, with whom we have long-standing research and teaching collaborations, to nominate suitable participants. Participants in group 6 were identified in collaboration with the Fellows of the Hong Kong College of Family Physicians. For groups 7 and 8, eligible physicians were selected based on referral network lists published by the Health Bureau of the Hong Kong Special Administrative Region Government.

Data Collection

All interviews were conducted via Zoom (Zoom Communications). A semistructured interview guide (Multimedia Appendix 1) was used to explore PCPs' awareness of, attitudes toward, and perceived barriers to and enabling factors for use of the RF. Before formal data collection, the guide and the moderator's manual were pilot-tested with a small sample of PCPs from diverse practice settings. Feedback from the pilot informed revisions to improve the clarity and contextual relevance of the questions. During the interviews, the final version of the guide was used, with additional probing questions introduced as new topics emerged. All interviews were conducted by 2 trained research assistants (CYL and XZ)—one serving as the facilitator and the other as an observer—both of whom had prior experience in qualitative data collection.

All interviews were audio-recorded and transcribed verbatim in Chinese. A standardized transcription protocol was established to guide and ensure consistency throughout the transcription process. Transcribers received training and were supervised by an experienced researcher, who also validated the transcripts to ensure accuracy and reliability. Transcripts were translated into English only after completion of data analysis to preserve the original meanings and avoid introducing interpretive bias during coding. Interviewee transcript review was not conducted, as evidence from previous studies indicates that this practice may have a limited impact on transcript accuracy [24]. However, each transcript was independently reviewed by an experienced researcher (CCZ) to minimize potential bias and misinterpretation. The research team further reviewed all transcripts to ensure both linguistic and conceptual accuracy.

Data Analysis

All data were analyzed concurrently and sequentially using NVivo 12 software (QSR International). Data analysis was conducted in 2 phases using thematic analysis based on the original Chinese transcripts. This approach was adopted to avoid potential misinterpretations that could arise from premature translation. Quotations were translated into English only after completion of data analysis to ensure that original meanings were preserved. The accuracy of the translations was subsequently checked and confirmed by a senior researcher (CCZ). In the first phase of analysis, inductive coding was performed to identify key patterns and themes emerging from the data. In the second phase, these preliminary themes were systematically reexamined using the CFIR as a guiding framework. The CFIR comprises 38 constructs across 5

XSL•FO
RenderX

domains—Innovation, Inner Setting, Outer Setting, Individual Characteristics, and Implementation Process—which may act as barriers or facilitators depending on the implementation context [25]. Applying the CFIR enabled a structured, theory-driven analysis of the contextual factors shaping RF implementation.

The coding process was conducted independently by 2 researchers (MC and XZ). Any discrepancies were discussed and resolved through team consensus. When disagreements persisted, a senior researcher with expertise in implementation science was consulted to adjudicate and facilitate consensus. Redundant or overlapping codes were merged to ensure clarity and consistency. To identify the most frequently reported implementation determinants, the number of PCPs who described experiences or views related to each CFIR construct was also calculated. This approach helped highlight the most salient barriers and facilitators across participants.

Quality Assurance

In this study, interviewers received comprehensive training to ensure competency in conducting interviews, adherence to the CFIR framework, and consistency across interviews. The training covered key areas, including understanding the CFIR framework, interviewing skills, ethical considerations, mock interviews, and cultural sensitivity. In addition, quality control measures were adopted to ensure the accuracy and reliability of the data collection process. A standardized procedure was developed to guide the interview process, including preinterview preparation and interview conduct. Furthermore, senior researchers (CCZ and MCSW) periodically observed interviews to ensure adherence to the protocol and to provide feedback to interviewers.

Ethical Considerations

The study was reviewed and approved by the Survey and Behavioral Research Ethics Committee, The Chinese University of Hong Kong (approval number SBRE-22-0193). Oral informed consent was obtained from all participants before the interviews. The original informed consent permitted secondary analysis without additional consent. Consent information included details about the study's purpose, procedures, potential risks, and benefits, as well as participants' right to withdraw at any time without penalty. All data were deidentified to protect participant privacy and confidentiality. For identifiable information collected during the primary study, strict data protection protocols were followed, including secure storage on password-protected systems accessible only to authorized personnel. No identifiable information of individual participants is presented in this manuscript or the multimedia appendices. As a token of appreciation, each participant received an HK $100 (US $12.79) coupon upon completion of the interview.

Results

Characteristics of the Included Participants

A total of 40 PCPs participated in this qualitative study, offering diverse professional backgrounds and experiences. Most participants were male (n=31, 78%), while 9 (23%) were female (Table 1). Nearly half (n=18, 45%) were recruited from public health care settings, and 36 (90%) had more than 16 years of clinical experience; 24 (60%) reported seeing fewer than 40 patients per day. The majority were family physicians (n=31, 78%).

XSL•FO
RenderX

Table 1. Demographic characteristics of participants.[a]

Demographic characteristics	Values, n (%)
Gender	
Male	31 (78)
Female	9 (23)
Age (years)	
35-45	7 (18)
46-55	24 (60)
56-65	8 (20)
≥66	1 (3)
Practice experience (years)	
5-15	4 (10)
16-25	13 (33)
26-35	16 (40)
36-45	7 (18)
Practice type (public)	
General outpatient clinic	15 (38)
Family medicine specialist clinics	2 (5)
Specialist outpatient clinics	1 (3)
Practice type (private)	
Solo	13 (33)
With partners	2 (5)
Health maintenance organizations	4 (10)
University	3 (8)
Specialty	
Family medicine	31 (78)
Internal medicine	2 (5)
Community medicine	3 (8)
Psychiatry	1 (3)
Geriatric medicine	3 (8)
Practice size (number of patients seen daily)	
0-19	15 (38)
20-39	9 (23)
40-59	7 (18)
60-79	7 (18)
>100	1 (3)
Missing	1 (3)

[a]Participants were recruited from 8 distinct groups of service providers from February to September 2024 through emails or phone calls.

Barriers and Facilitators to the Adoption of the RF Using CFIR

Barriers and facilitators to the adoption of the RF in daily consultations were identified using the CFIR framework, with key themes emerging from the interviews. Figure 1 illustrates the CFIR domains reflected in the interviews: Innovation, Outer Setting, Inner Setting, Characteristics of Individuals, and the Implementation Process. Table 2 summarizes the frequencies for each identified theme, while Tables 3 and 4 illustrate quotations for each identified theme. Common facilitators included relative advantage (n=26, 65%), innovation design (n=18, 45%), and access to knowledge and information (n=28, 70%). Major barriers were policy and legal concerns and

XSL•FO
RenderX

availability of resources, each cited by 29 (73%) participants. Key themes are detailed below.

Figure 1. Schematic diagram of the Consolidated Framework for Implementation Research.

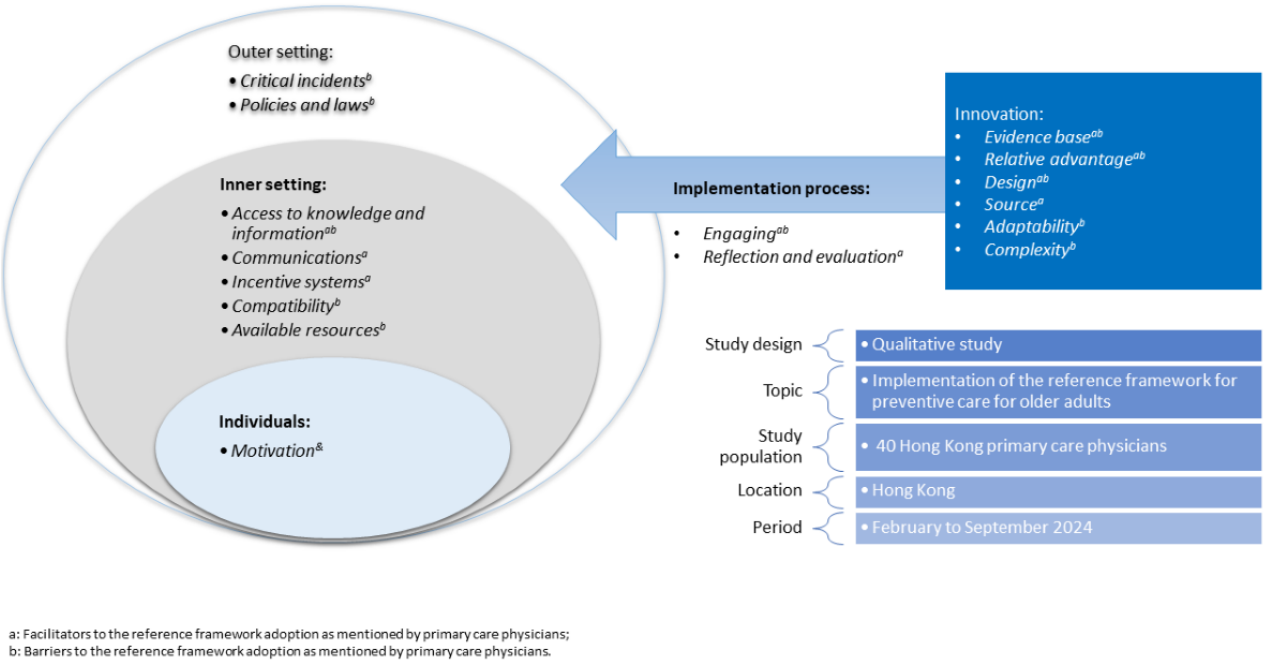

XSL•FO
RenderX

Table 2. The barriers and facilitators for the reference framework.[a]

CFIR[b] domains and constructs	Barriers, n (%)	Facilitators, n (%)
Innovation		
Innovation evidence base	4 (10)	7 (18)
Innovation relative advantage	12 (30)	26 (65)
Innovation design	10 (25)	18 (45)
Innovation source	0 (0)	9 (23)
Innovation adaptability	7 (18)	0 (0)
Innovation complexity	6 (15)	0 (0)
Inner Setting		
Access to knowledge	10 (25)	28 (70)
Communications	0 (0)	9 (23)
Incentive systems	0 (0)	3 (8)
Compatibility	9 (23)	0 (0)
Available resources	29 (73)	0 (0)
Outer Setting		
Critical incidents	3 (8)	0 (0)
Policies and laws	29 (73)	0 (0)
Individuals		
Motivation	3 (8)	0 (0)
Implementation Process		
Engaging	5 (13)	10 (25)
Reflection and evaluation	0 (0)	12 (30)

[a]Participants were recruited from 8 distinct groups of service providers from February to September 2024 through emails or phone calls.
[b]CFIR: Consolidated Framework for Implementation Research.

XSL•FO
RenderX

Table 3. Quotations of facilitators for the reference framework.[a]

CFIR[b] domains, subdomains, and example categories	Quotations of facilitators
Innovation	
Evidence base	
Confidence in evidence-based approach	• "Because it is evidence based, rather than purely based on my own policies for the clinic or purely to meet the needs of patient, it reinforces my beliefs." (023) • "I think that if it will encourage more doctors to, you know, provide evidence-based care, maybe, but it might be worthwhile to sort of know, you know, how much we actually are doing or how successful we are given our current approach" (029) • "That is, a lot of evidence base information can actually remind us of some updated treatment plan or some care plan." (04)
Relative advantage	
Value in holistic geriatric care	• "We often neglect to consider the living environment of the elderly or whether there is anyone at home to take care of them. This is another important reminder for us—not just to focus on their illness or the medications they are using." (004) • "Dementia is something everyone has heard of, but what exactly is it? What are the symptoms? What should we pay attention to? These are things that, within this framework, help us realize that we need to be aware of these issues, especially for the elderly." (021)
RF[c] as a structured clinical decision support tool	• "It acts as a structured checklist that ensures comprehensive patient assessment." (009) • "A summary checklist facilitates workflow efficiency by allowing clinicians to systematically track completed and pending clinical steps." (010)
Design	
Visual enhancements improve RF usability and implementation	• "The color coding is helpful, so you can see things at a glance." (029) • "The hardcopy is of a good size to be put in clinics." (031)
Integration of technology and artificial intelligence to enhance RF implementation	• "If we could develop some apps or use AI technology, for example, we could key in certain patient information and come up with conclusions or suggestions. This would be a good direction." (003) • "If there is an advanced version, that is, the RF steps to form a different path, so that the content is more concise, can help me follow the path to care for patients and it will be more convenient for my daily use." (014)
Source	
Trust in the expert-driven development	• "You're talking about doing it based on evidence. Of course, because there's a large team of experts who spend a lot of time putting it together." (021) • "Having experts provide an official guide for local colleagues is very worthwhile." (022) • "I have confidence in its robustness or feasibility because it is made by the Primary Health Care Office." (026)
Inner Setting	
Access to knowledge and information	
Accessibility and availability of RF information facilitate uptake	• "Regular updates provided through the Primary Care Office website contribute to sustained engagement with RF." (011) • "Email notifications and printed summary materials enhance accessibility and reinforce RF awareness." (004)
Communications	
Communication and peer support	• "Our own doctors' peer groups have discussions, so from these three aspects, we are aware of this framework." (005) • "Sometimes, there may be specific content that we can discuss together...Maybe we can encourage more to do it." (010)
Incentive systems	
Incentives and Continuing Medical Education accreditation promote RF engagement	• "They can also earn some CME credits. This encourages all doctors to reference, learn, and study." (001) • "The government may need to provide more incentives, like funding, to support and motivate doctors to perform these screenings." (017)

XSL•FO
RenderX

CFIR[b] domains, subdomains, and example categories	Quotations of facilitators
Implementation Process	
Engaging	
Patient education and shared decision-making facilitate RF integration	• "Important points should be highlighted, possibly using charts or visuals to help elderly patients." (023) • "I think the version you give to the patient should not be the English version but the Chinese version...it is a simplified version plus some explanations...we know exactly what we are doing, but for them (the patient), they may think that what you are asking has nothing to do with the disease...Firstly, it needs to be simplified, secondly, it needs some description...Then they (the patients) won't find you annoying and know that what you are asking is really related to them." (009)
Reflection and evaluation	
Continuous evaluation and iterative refinement of RF to ensure relevance	• "It would be ideal if there were committees that regularly review this, as that would be very practical." (015) • "I hope the relevant authorities can promptly review the effectiveness of this guideline or framework and propose a post-pandemic version." (026) • "For example...if new decisions arise, do we need to make modifications or release a second edition?...it's worth paying attention to how often updates occur, or when necessary, updates can be made. (022)

[a]Participants were recruited from 8 distinct groups of service providers from February to September 2024 through emails or phone calls.

[b]CFIR: Consolidated Framework for Implementation Research.

[c]RF: Reference Framework for Preventive Care for Older Adults.

XSL•FO
RenderX

Table 4. Quotations of barriers for the reference framework.[a]

CFIR[b] domains, subdomains, and example categories	Quotations of facilitators
Innovation	
Evidence base	
Concerns over outdated or overly theoretical content	• "The framework is too theoretical, vague academic, and complex, especially for beginners." (013) • "Sometimes I think it's so theoretical, if we can execute the essence of it in front line?" (015) • "I also noticed that many of the references in here are from a long time ago, or even 20th century." (032)
Relative advantage	
Limited utility for experienced physicians	• "No, it's not very useful for me. I rely more on what I know than on it...for those who are already trained...it may not be very useful." (006) • "I find it hard to recommend this material. It's quite simple, and everyone should know it." (017)
Perceived lack of clinical relevance	• "The information in the document would not help anyone make significant clinical decisions and is insufficient in determining the next step." (003) • "There might be conditions not currently covered." (033)
Design	
Lack of user-friendliness and accessibility	• "It's not user-friendly because when doctors want to implement certain recommendations, they have to go back to the original documents." (040) • "It almost 70, 80 pages for a module, and not many doctors will read it all...We prefer to read things like lecture notes." (002) • "I'm not sure how much is involved in transitioning to an app version or if there's a plan for it. Nowadays, everyone accesses information via smartphones." (034)
Adaptability	
Limited personalization and outdated design	• "Therefore, I find that what's mentioned in it is quite vague and not personalized enough to apply to my patients in family medicine." (006) • "After COVID-19, I think there have been significant changes in elderly health...Utilizing technology for smart health care to help the elderly participate in their health has certainly changed compared to the initial design." (026)
Complexity	
Challenges due to breadth and length	• "It is also too extensive and should, instead, suggest what choices to consider under different situations." (014) • "It was impractical to follow the guidelines as the RF is too comprehensive with too many pages to follow." (002)
Inner Setting	
Access to knowledge and information	
Lack of access to training and awareness	• "When we are working, there may be some CME activities on primary care related topics, but we do not pay special attention to attend a workshop on framework." (022) • "I didn't even know there were any training programs about this" (025)
Compatibility	
Compatibility with clinical practice	• "Actually, I work in geriatrics. When I talk to patients, the concepts in this overview are quite similar." (001) • "I don't need to go through the overview, because in fact many of them are routines, such as diabetes, hypertension" (014) • "We don't usually follow this RF, although sometimes we do things that overlap with the RF, but I don't actually see it from this RF." (016)
Competing priorities and limited focus on the RF[c]	• "I haven't specifically focused on the overview because there are internal workflows and the consultations., we can provide are quite brief due to limited resources." (014)
Available resources	

XSL•FO
RenderX

CFIR[b] domains, subdomains, and example categories	Quotations of facilitators
Time constraints and resource limitations	• "In government institutions, like in our general outpatient clinics, it's quite challenging because we only have 6 minutes for each consultation." (027) • "I look for opportunities during the consultation; if I have extra time, I can discuss a bit more." (018) • "The practicality of the guidelines really depends on how many staff members are available in your clinic and how well you manage your own time." (007)
Outer Setting	
Critical incidents	
Disruption caused by COVID-19	• "In the past few years, we have been implementing this framework, but many events, like COVID, have occurred." (019) • "I haven't participated in many workshops in the past few years due to the epidemic. Our frontline is mainly responsible for the epidemic." (023)
Policies and laws	
Concerns over compulsory adoption and loss of professional autonomy	• "You might have a template for everyone to follow...if someone is entering a nursing home, I need you to fill out all these things, which feels compulsory and may not give the doctor a good impression." (008) • "If it is a policy that we have to follow, then I will have qualified opinion about it...it will make many frontline doctors worry...no more room for freedom...may offend patients, or some patients who have medical needs will not be able to receive suitable treatment." (004)
Characteristics of Individuals	
Motivation	
Preference for alternative guidelines	• "Some colleagues may follow foreign guidelines instead of the local recommendations." (040) • "Doctors in Hong Kong will refer to a lot of different information...even if I don't follow Hong Kong's guideline, but follow US's guideline, it's still evidence base" (035)
Implementation Process	
Engaging	
Challenges with patient adherence and expectations	• "However, when it comes to private practices, if it involves blood tests, it requires citizens to be willing to pay. If a doctor suggests a blood test and the patient is unwilling, that can pose difficulties." (027) • "Currently, the situation is that people come to see a doctor only when they are sick." (021)

[a]Participants were recruited from 8 distinct groups of service providers from February to September 2024 through emails or phone calls.

[b]CFIR: Consolidated Framework for Implementation Research.

[c]RF: Reference Framework for Preventive Care for Older Adults.

Innovation

Innovation Evidence Base

Confidence in Evidence-Based Approach

The RF's credibility, grounded in an evidence-based approach, was acknowledged by participants. A total of 7 (18%) respondents mentioned confidence in the use of the RF based on sufficient evidence-based research for its implementation.

> *Because it is evidence based, rather than purely based on my own policies for the clinic or purely to meet the needs of patient, it reinforces my beliefs.* [023]
>
> *I think that if it will encourage more doctors to, you know, provide evidence-based care, maybe, but it might be worthwhile to sort of know, you know, how much we actually are doing or how successful we are given our current approach.* [029]
>
> *That is, a lot of evidence base information can actually remind us of some updated treatment plan or some care plan.* [04]

Concerns Over Outdated or Overly Theoretical Content

However, the evidence for RF has been subject to some criticism, particularly in terms of its timeliness and clinical validity. Some participants (n=4) indicated that the current evidence for RF was outdated and too theoretical for frontline use.

> *The framework is too theoretical, vague academic, and complex, especially for beginners.* [013]
>
> *Sometimes I think it's so theoretical, if we can execute the essence of it in front line?* [015]

XSL•FO
RenderX

I also noticed that many of the references in here are from a long time ago, or even 20th century. [032]

Innovation Relative Advantage

Value in Holistic Geriatric Care

Some participants noted its practical utility. For example, 26 (65%) participants highlighted its value in ensuring no critical aspects are overlooked, particularly in older adults' care.

We often neglect to consider the living environment of the older adults or whether there is anyone at home to take care of them. This is another important reminder for us—not just to focus on their illness or the medications they are using. [004]

Dementia is something everyone has heard of, but what exactly is it? What are the symptoms? What should we pay attention to? These are things that, within this framework, help us realize that we need to be aware of these issues, especially for the older adults. [021]

RF As a Structured Clinical Decision-Support Tool

In addition, respondents mentioned that this RF helped remind them of some clinical processes that may have been forgotten.

It acts as a structured checklist that ensures comprehensive patient assessment. [009]

A summary checklist facilitates workflow efficiency by allowing clinicians to systematically track completed and pending clinical steps. [010]

Limited Utility for Experienced Physicians

However, 12 (30%) participants pointed out that this RF was not useful to them in practice. On the one hand, they felt that the content of the RF was too basic for experienced physicians.

No, it's not very useful for me. I rely more on what I know than on it...for those who are already trained...it may not be very useful. [006]

I find it hard to recommend this material. It's quite simple, and everyone should know it. [017]

Perceived Lack of Clinical Relevance

On the other hand, they believed that the consultation in this RF was insufficient to help them make useful clinical decisions.

The information in the document would not help anyone make significant clinical decisions and is insufficient in determining the next step. [003]

There might be conditions not currently covered. [033]

Innovation Design

Visual Enhancements Improve RF Usability and Implementation

Participants (n=18) were optimistic about the design. For example, they claimed that the RF is well-sized and color-coded with visual aids, and that these features improve clarity and accessibility.

The color coding is helpful, so you can see things at a glance. [029]

The hardcopy is of a good size to be put in clinics. [031]

Integration of Technology and Artificial Intelligence to Enhance RF Implementation

In addition, respondents made several suggestions to make the RF more user-friendly. Their suggestions focused on how to make it easier and quicker to access the information they need and use the RF to improve clinical efficiency, emphasizing the integration of emerging technologies, especially artificial intelligence.

If we could develop some apps or use AI technology, for example, we could key in certain patient information and come up with conclusions or suggestions. This would be a good direction. [003]

If there is an advanced version, that is, the RF steps to form a different path, so that the content is more concise, can help me follow the path to care for patients and it will be more convenient for my daily use. [014]

Lack of User-Friendliness and Accessibility

In addition, the design of the RF was criticized by some respondents. Ten participants expressed concerns and dissatisfaction with the current design of the RF, especially in terms of the user-friendliness of interactions and the lack of usability of the currently available version in different situations.

It's not user-friendly because when doctors want to implement certain recommendations, they have to go back to the original documents. [040]

It almost 70, 80 pages for a module, and not many doctors will read it all...We prefer to read things like lecture notes. [002]

I'm not sure how much is involved in transitioning to an app version or if there's a plan for it. Nowadays, everyone accesses information via smartphones. [034]

Innovation Source: Trust in the Expert-Driven Development

Within this construct, 9 (23%) participants expressed trust in the expert panels involved in developing the RF. They generally believed that the expertise of these panels ensured the reliability of the RF and supported the evidence-based approach adopted.

You're talking about doing it based on evidence. Of course, because there's a large team of experts who spend a lot of time putting it together. [021]

Having experts provide an official guide for local colleagues is very worthwhile. [022]

I have confidence in its robustness or feasibility because it is made by the Primary Health Care Office. [026]

XSL•FO
RenderX

Innovation Adaptability: Limited Personalization and Outdated Design

Participants (n=7) raised concerns about the local use of this RF at the frontline. They claimed that the lack of personalization or localization in the current version of the RF hinders its use in their practice.

> *Therefore, I find that what's mentioned in it is quite vague and not personalized enough to apply to my patients in family medicine.* [006]
>
> *After COVID-19, I think there have been significant changes in older adults' health...Utilizing technology for smart health care to help the older adults participate in their health has certainly changed compared to the initial design.* [026]

Innovation Complexity: Challenges Due to Breadth and Length

The comprehensiveness of the RF presented a challenge. Some participants (n=6) pointed out that it contained extensive content and too many pages, which limited their ability to implement and follow it in a real-world setting.

> *It is also too extensive and should, instead, suggest what choices to consider under different situations.* [014]
>
> *It was impractical to follow the guidelines as the RF is too comprehensive with too many pages to follow.* [002]

Inner Setting

Access to Knowledge and Information

Accessibility and Availability of RF Information Facilitate Uptake

Most participants (n=28) indicated that they obtained RF-related information and updates mainly through self-directed behavior or informal pathways, such as checking the websites of government health care departments and emails and hard copies sent by relevant departments, and also considered that this made them more familiar with the RF.

> *Regular updates provided through the Primary Care Office website contribute to sustained engagement with RF.* [011]
>
> *Email notifications and printed summary materials enhance accessibility and reinforce RF awareness.* [004]

Lack of Access to Training and Awareness

However, some participants (n=10) pointed out that information on RF training was limited and that they had not even received information on RF training or workshops.

> *When we are working, there may be some CME activities on primary care related topics, but we do not pay special attention to attend a workshop on framework.* [022]
>
> *I didn't even know there were any training programs about this* [025]

Communications: Communications and Peer Support

Maintaining communication with peers is important for encouraging the use of the RF. Nine participants mentioned that they often update information through peer discussions and thus increase peer collaboration, informal feedback on existing content, and review of its utility in the clinic to ensure more effective implementation of the RF.

> *Our own doctors' peer groups have discussions, so from these three aspects, we are aware of this framework.* [005]
>
> *Sometimes, there may be specific content that we can discuss together...Maybe we can encourage more to do it.* [010]

Incentive Systems: Incentives and Continuing Medical Education Accreditation

Incentives, tangible or intangible, are crucial to the implementation of new measures. Three participants mentioned that additional incentives may be needed to increase the use of the RF, including the inclusion of the RF in Continuing Medical Education (CME) credit courses to encourage more physicians to participate. In addition, providing financial assistance was cited as an effective approach.

> *They can also earn some CME credits. This encourages all doctors to reference, learn, and study.* [001]
>
> *The government may need to provide more incentives, like funding, to support and motivate doctors to perform these screenings.* [017]

Compatibility

Compatibility With Clinical Practice

Some physicians (n=9) indicated that they were reluctant to use or refer to the RF in their day-to-day work, and one of the primary reasons was that they recognized the content of the RF as similar to what they did on a day-to-day basis.

> *Actually, I work in geriatrics. When I talk to patients, the concepts in this overview are quite similar.* [001]
>
> *We don't usually follow this RF, although sometimes we do things that overlap with the RF, but I don't actually see it from this RF.* [016]
>
> *I don't need to go through the overview, because in fact many of them are routines, such as diabetes, hypertension.* [014]

Competing Priorities and Limited Focus on the RF

In addition, the prioritization of internal workflows may hinder their use of the RF, as one participant mentioned.

> *I haven't specifically focused on the overview because there are internal workflows and the consultations.* [014]

Available Resources: Time Constraints and Resource Limitations

When discussing how the RF works in practice, most participants (n=29) complained that real-world resource

XSL•FO
RenderX

constraints limited their use of the RF, especially time constraints. Resource constraints have been a longstanding theme in the health care system, especially in the public health care system, where respondents reported that their consultation time per patient was only a few minutes, making it difficult to implement the RF. In addition, the lack of staff presents an additional challenge.

> *In government institutions, like in our general outpatient clinics, it's quite challenging because we only have 6 minutes for each consultation.* [027]
>
> *I look for opportunities during the consultation; if I have extra time, I can discuss a bit more.* [018]
>
> *The practicality of the guidelines really depends on how many staff members are available in your clinic and how well you manage your own time.* [007]

Outer Setting

Critical Incidents: Disruption Caused by COVID-19

The COVID-19 pandemic posed challenges to the RF's promotion and implementation, with 3 participants mentioning that the pandemic limited its outreach. They noted that the arrival of COVID-19 shifted their focus to COVID-19–related activities, to the extent that they had little exposure to RF-related knowledge or training, despite having started RF implementation several years ago.

> *In the past few years, we have been implementing this framework, but many events, like COVID, have occurred.* [019]
>
> *I haven't participated in many workshops in the past few years due to the epidemic. Our frontline is mainly responsible for the epidemic.* [023]

Policies and Laws: Concerns Over Compulsory Adoption and Loss of Professional Autonomy

When we addressed their views on whether the RF should be made a policy, most participants (n=29) showed clear opposition, believing that if the RF is made mandatory, it may cause physicians to lose their clinical freedom, thereby ignoring the real needs of their patients.

> *If it is a policy that we have to follow, then I will have qualified opinion about it...it will make many frontline doctors worry...no more room for freedom...may offend patients, or some patients who have medical needs will not be able to receive suitable treatment.* [004]
>
> *You might have a template for everyone to follow...if someone is entering a nursing home, I need you to fill out all these things, which feels compulsory and may not give the doctor a good impression.* [008]

Individuals (Motivation: Preference for Alternative Guidelines)

We found that PCPs' beliefs about different health care systems may influence their acceptance and adoption of the RF. Participants (n=3) noted that they may compare practices and guidelines from different regions on their own and place more faith in foreign guidelines than in local recommendations.

> *Some colleagues may follow foreign guidelines instead of the local recommendations.* [040]
>
> *Doctors in Hong Kong will refer to a lot of different information...even if I don't follow Hong Kong's guideline, but follow US's guideline, it's still evidence base.* [035]

Implementation Process

Engaging

Challenges With Patient Adherence and Expectations

We found that some patients' uncooperative behaviors or habits may affect physicians' confidence and willingness to implement the RF. Some participants (n=5) noted that even when the RF is followed to make recommendations for patients, its effectiveness requires patient participation, and that uncooperative attitudes make the RF difficult to implement. In addition, respondents pointed to doubts about the usefulness of the RF, emphasizing that confidence in the RF is low in the current style of patient health service seeking.

> *However, when it comes to private practices, if it involves blood tests, it requires citizens to be willing to pay. If a doctor suggests a blood test and the patient is unwilling, that can pose difficulties.* [027]
>
> *Currently, the situation is that people come to see a doctor only when they are sick...however, the government's future direction hopes to detect diseases earlier, intervene sooner, and prevent illnesses from occurring.* [021]

Patient Education and Shared Decision-Making Facilitate RF Integration

Although there are some limitations to RF adoption, some participants (n=10) noted that the RF may be adopted because of the role it plays, particularly in empowering patients and helping doctors and patients reach consensus. Participants noted that uncooperative behaviors may decrease and physician-patient communication may be enhanced if patients are aware of the RF.

> *I think the version you give to the patient should not be the English version but the Chinese version, and it should be simpler, it is a simplified version plus some explanations...Of course, we know exactly what we are doing, but for them (the patient), they may think that what you are asking has nothing to do with the disease, they are just seeing a doctor...Firstly, it needs to be simplified, secondly, it needs some description...Then they (the patients) won't find you annoying and know that what you are asking is really related to them.* [009]
>
> *Important points should be highlighted, possibly using charts or visuals to help older adults.* [023]

XSL•FO
RenderX

Reflection and Evaluation: Continuous Evaluation and Iterative Refinement of RF to Ensure Relevance

When discussing further promotion of RF adoption, participants (n=12) emphasized the importance of regular review and updating to keep up with the latest trends, especially after experiencing critical incidents such as the COVID-19 pandemic.

> *It would be ideal if there were committees that regularly review this, as that would be very practical.* [015]

> *After COVID-19, I think there have been significant changes in older adults' health...Utilizing technology for smart health care to help the older adults participate in their health has certainly changed compared to the initial design. I hope the relevant authorities can promptly review the effectiveness of this guideline or framework and propose a post-pandemic version.* [026]

> *For example, after five years or a few years, if new decisions arise, do we need to make modifications or release a second edition? I know they would do that, but it's worth paying attention to how often updates occur, or when necessary, updates can be made.* [022]

Discussion

Summary of Findings

This qualitative study examined the views of 40 PCPs on the adoption of the RF in Hong Kong. The findings revealed a complex interplay of systemic, organizational, and individual-level factors influencing RF implementation. At the policy level, many participants expressed concern about making RF adoption mandatory, fearing a loss of clinical autonomy and flexibility in tailoring care to individual patient needs. Participants preferred the RF as a recommendation rather than a compulsory protocol. Within the Inner Setting, peer discussions emerged as a facilitator of RF awareness and informal learning. However, some physicians perceived the RF as redundant, given its conceptual overlap with their routine practice. Competing clinical priorities and limited consultation time also hindered engagement with the RF. Although most participants accessed RF information through government websites or printed materials, many cited a lack of structured training opportunities and expressed interest in clearer dissemination and CME-accredited courses. Resource constraints—especially time limitations and staffing shortages—posed significant barriers to practical implementation. Furthermore, physicians highlighted the importance of financial and educational incentives to support adoption. At the individual level, some PCPs demonstrated a preference for foreign guidelines over local ones, reflecting skepticism toward the RF's added value. Patient factors, including low adherence, financial considerations, and a curative (rather than preventive) care-seeking mindset, were seen as major challenges to RF uptake. Despite these barriers, several participants acknowledged the potential of the RF to enhance physician-patient communication and support shared decision-making, particularly if supported by simplified, culturally appropriate patient materials. Continuous evaluation and iterative updates to the RF were viewed as essential to maintaining relevance, especially in the post–COVID-19 context. To further understand how these factors interact, we mapped them across the CFIR domains to identify patterns and relationships that shape implementation outcomes.

Interpretation Across CFIR Domains

Within the Intervention Characteristics domain, trust in the RF's evidence-based foundation and expert development fostered credibility, yet its perceived complexity and limited adaptability constrained its practical use—especially in time-constrained environments (Inner Setting). External pressures such as the COVID-19 pandemic and resistance to policy mandates (Outer Setting) further shaped adoption, interacting with individual beliefs and preferences (Characteristics of Individuals), including skepticism toward local guidelines. Peer communication and informal learning (Inner Setting) helped mitigate gaps in formal training, while patient engagement and shared decision-making (Process) emerged as critical to overcoming adherence challenges. These cross-domain patterns suggest that effective implementation requires not only credible and well-designed content but also responsive adaptation to clinical workflows, supportive infrastructure, and active stakeholder involvement.

Possible Explanation and Recommendations

This study received both positive and negative feedback from practitioners regarding their impressions of the RF, with some practitioners indicating that the RF was useful, while others criticized that the content was not applicable to daily consultation routines. Based on our results, practitioner noncompliance may be related to personal attitudes toward the RF and limitations of the RF, such as too many pages and differing physiological conditions of patients. The results revealed that improvement of the RF was necessary to enhance practitioner compliance. High complexity and a high number of conditional recommendations in guidelines have been found to be major barriers to physicians adopting guidelines [26]. The current RF contains comprehensive information but lacks a quick checklist for practitioners to apply. A simple checklist for practitioners and patients may help improve RF compliance. Furthermore, comments from physicians reflected a lack of collaboration and communication between expert panels and frontline physicians, resulting in a tool that prioritizes general guidance over actionable, context-specific solutions. Lu and colleagues [27] emphasized and provided key strategies for successful implementation, in which stakeholder engagement was identified as an important process for success. To bridge this gap, co-designing updates to the RF with frontline practitioners is essential. Regular feedback loops and workshops involving diverse stakeholders, including PCPs, specialists, and nurses, can help refine the framework. These updates should focus on creating adaptable modules or case-based scenarios that address specific clinical complexities while retaining the RF's evidence-based rigor. A critical interpretation of our results points to a fundamental implementation challenge: the RF was developed with insufficient input from its end users. This top-down approach likely explains why physicians viewed it as a source of general guidance rather than a practical clinical

XSL•FO
RenderX

tool. This aligns with the implementation science literature, which consistently identifies stakeholder engagement as a cornerstone of success [27]. Therefore, a paradigm shift from consultation to co-creation is warranted. We recommend that updates to the RF be developed through participatory workshops with frontline PCPs, specialists, and nurses. This co-design process would focus on embedding actionable, case-based scenarios directly into the framework, enhancing its relevance and utility in daily practice.

The RF received a positive response regarding the user-friendliness and convenient design of the framework, especially in terms of size and color coding. User-friendliness also emerged as a crucial factor in promoting regular use. Catho et al [28] conducted a qualitative study to examine factors determining guideline adherence and adoption of computerized decision support systems by physicians. The study indicated that user-friendliness was one of the key factors affecting the adoption of new tools in the workplace. The underlying mechanism may be attributed to the fact that an intuitive interface and adequate technical support may help clinicians use the RF easily, even if they are not familiar with the technology [29].

By contrast, users also identified barriers to implementing the RF in daily consultations. Resource constraints, especially insufficient consultation time, were reported by most participants and substantially affected RF adoption. A systematic review covering 18 countries and 28 million consultations across 111 publications from 1946 to 2016 revealed that in half of the cases, consultation time was 5 minutes or less [30]. The main contributor to this phenomenon was heavy practitioner workloads, for example, in China [31]. No guideline recommends an optimal consultation time for outpatient clinics, but consultation times of less than 7 minutes were found to be more likely to result in practitioners missing some psychosocial problems that patients have [32], let alone following the RF within such a short consultation. Solo practitioners must negotiate optimal medication for patients, address concerns from patients and families, manage schedule constraints, and—above all—consider patients', families', or caregivers' capacity to follow the recommended regimen within a limited consultation. Despite participants having positive responses to the robustness, feasibility, benefits for patient care, user-friendliness, and convenience features of the RF, they also reflected that sufficient consultation time would be key to whether the RF was practically applicable. Notably, our findings emphasize physicians' resistance to mandatory use of the RF, possibly due to impairment of job autonomy or negative effects from increased external controls [33].

This study highlights physicians perceived role of patient engagement in the decision-making process of whether to practice the RF and points out that low patient compliance and poor care outcomes may undermine physicians' confidence and willingness to perform the RF, which may be related to physicians perceived meaning of their work. Previous research has indicated that physicians' perceptions of health care outcomes and the meaning of their work may influence or prevent the onset of burnout [34], and that there is a negative relationship between burnout and physicians' perceived value of an innovation or willingness to adopt an innovation [35]. This mechanism reflects the fact that physicians may abandon use of the RF because of perceived lack of patient engagement or ineffective care. Therefore, it may be important to enhance physicians' sense of meaning and willingness to use innovations by improving patient engagement or adherence. In addition, we found that physicians offered many ideas for improving patient engagement, particularly in the areas of patient empowerment and physician-patient consensus. They noted that the RF should be expanded to include patient education, which would help them better implement the content of the RF. This approach may be effective in that, on the one hand, patient empowerment is evident in stimulating patient engagement and optimizing the use of limited resources such as time, as well as reducing the burden on frontline health care professionals [36,37]. On the other hand, improved physician-patient communication may enhance the likelihood of long-term care outcomes, thereby increasing physician confidence in the use of the RF [34].

Continuous improvement is vital for the sustained relevance and effectiveness of the RF. Regularly seeking user feedback and conducting usability testing could help identify areas for enhancement [38]. Implementing long-term engagement strategies, such as personalized reminders, could also help sustain routine use effectively [39-41]. This study revealed insufficient training or workshops concerning the RF, especially during critical incidents such as COVID-19. To support the adoption of the RF, ongoing training and user support are critical. Workshops, clear instructional materials, and accessible technical support [42] could help practitioners become more familiar with the RF as a whole. By implementing these measures, health care providers could navigate the RF more effectively and better utilize the 6 modules in the RF, improving their overall experience and patient health outcomes.

In this study, some constructs, such as structural characteristics and local attitudes, were not coded or mentioned by physicians. This absence appears to reflect the local context. For example, infrastructure was not perceived as a barrier or facilitator, likely because the RF did not require significant changes to physical or technological systems. Similarly, the limited engagement of patients in the Implementation Process may be attributed to the top-down nature of Hong Kong's health care system, where patients typically have less access to information and decision-making power compared with health care professionals.

Implications for Practice and Research

Beyond the specific barriers and facilitators identified, our findings underscore that the broader cultural, organizational, and policy contexts in Hong Kong are not passive settings but active forces shaping RF implementation. Culturally, the longstanding preference for curative care over preventive approaches presents a foundational challenge [43]. This orientation suggests that the success of the RF is not only dependent on its technical integration into clinical workflows but also on a broader societal shift toward preventive health literacy. Without sufficient patient understanding and engagement, even a well-integrated RF may encounter resistance or low adherence in practice. At the organizational level, the structure and pace of the public health care system impose

XSL•FO
RenderX

significant constraints. Physicians often operate under intense time pressure, and many clinical workflows are already well established. In this context, any new intervention must demonstrate clear added value and seamless compatibility with existing practices; otherwise, it risks being perceived as redundant or burdensome. Therefore, implementation strategies must be designed with these operational realities in mind, ensuring that the RF enhances rather than disrupts clinical efficiency. From a policy perspective, top-down mandates may be counterproductive if they conflict with the cultural value placed on clinical autonomy and the practical limitations of consultation time. Rather than imposing rigid requirements, policy should play a facilitative role. This includes providing resources to alleviate time constraints and supporting integration of the RF into existing digital infrastructures, such as the eHealth app [44]. Embedding the RF into such platforms could streamline access to patient-specific recommendations, reduce the cognitive and administrative load on practitioners, and ultimately improve adherence.

In addition to digital integration, the development of a lay version of the RF could empower older adults by offering accessible, user-friendly guidance on lifestyle changes and risk awareness. This would not only support patient engagement but also reinforce the preventive orientation that the RF seeks to promote. By equipping patients with the knowledge and tools to participate actively in their care, the RF could become a more effective and sustainable component of routine practice. While financial incentives have been proposed to increase practitioner adherence, previous studies have shown that such incentives often have limited impact on clinical behavior [45-47]. As such, financial incentives should be considered a last resort rather than a primary strategy. Instead, the focus should be on creating a supportive ecosystem in which the RF is perceived as both useful and feasible to implement.

This study offers several novel contributions to the CFIR literature. First, to our knowledge, it is the first to systematically investigate physician perspectives on the RF for the older adult population within the distinct sociocultural and structural context of Hong Kong. Second, our application of the CFIR uncovered specific barriers, such as how the COVID-19 pandemic introduced unique, large-scale disruptions that directly impeded the RF's promotion and integration. Finally, our findings provide actionable insights for tailoring implementation strategies in similar high-density, public payer–led Asian metropolises, addressing a gap in the current evidence base.

Importantly, future research should build on this foundation by applying the CFIR-Expert Recommendations for Implementing Change (ERIC) Matching Tool, which links CFIR-identified barriers to evidence-based implementation strategies from the ERIC taxonomy [48]. This tool maps 73 strategies to CFIR constructs, supporting the selection of targeted interventions such as preparing champions, conducting educational meetings, and tailoring strategies to local contexts [49]. Empirical studies have demonstrated the utility of this approach in improving implementation outcomes across various health care settings, including primary care and geriatric models [50,51]. Future research should focus on testing CFIR-ERIC–informed strategies to address the specific barriers identified in this study, such as limited leadership engagement, resource constraints, and competing clinical priorities. Comparative studies across different health care systems would help assess the generalizability of these strategies, while mixed methods and longitudinal designs could explore their impact on practitioner compliance, patient outcomes, and system-level sustainability.

Strengths and Limitations

This study is the first to comprehensively assess the adoption of the RF in the primary care setting, offering new insights and potential policy implications for enhancing clinical adoption of the RF. It draws strength from a substantial sample size of 40 PCPs, whose extensive clinical experience enriched the depth and relevance of the findings. Use of the CFIR provided a structured, theory-informed approach to identifying barriers and facilitators, enhancing the analytical rigor of the study. Moreover, the study contributes to the CFIR literature by applying the framework in a unique sociocultural and health care context, highlighting context-specific determinants that may not be captured in differently structured health systems.

However, several limitations should be acknowledged. First, the study relied on self-reported perceptions, which may introduce bias due to social desirability or recall limitations. Second, the findings are context specific and may not be directly transferable to other health care systems. Comparative studies across different health care systems would help assess the generalizability of these findings. Third, the study focused solely on the perspectives of PCPs, without incorporating views from other stakeholders such as patients, administrators, or policy makers, which may have provided a more holistic understanding of RF adoption. Including perspectives from patients, administrators, and policy makers will be essential to developing a more comprehensive understanding of RF implementation and its long-term viability.

Conclusions

This is the first study investigating the perception of RF implementation among PCPs in Hong Kong guided by the CFIR, offering essential insights for future enhancements of RF and aiding the translation of evidence into clinical practice. While the RF holds promise for enhancing preventive care and shared decision-making in Hong Kong's primary care setting, its adoption is currently hindered by systemic, organizational, and individual barriers. To support effective integration into routine practice, flexible implementation, stronger institutional support, tailored incentives, and improved clinician and patient engagement strategies are essential. Policy makers could begin by promoting awareness through targeted advertising and professional workshops, followed by revising and refining RF content to better align with clinical realities and patient needs. Future research should empirically test CFIR-ERIC–informed implementation strategies to address identified barriers, explore patient outcomes to support practitioner engagement, and conduct comparative studies to assess the transferability of RF adoption across diverse health care systems.

XSL•FO
RenderX

Acknowledgments

MCSW is an honorary medical advisor to GenieBiome Ltd., SunRise, and BGI Health; is an advisory committee member of Pfizer; an external expert for GlaxoSmithKline Limited; a member of the advisory board of AstraZeneca; has received consultancy fees for providing research advice; and has received grants from Pfizer, Merck Sharp & Dohme, AstraZeneca, GSK, and Moderna. We acknowledge the contribution of Ms. Stephanie Kwan, research assistant at The Chinese University of Hong Kong, for assisting with the interview study. During the preparation of this work, the authors used ChatGPT (OpenAI) solely to improve language and readability, with caution. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

This study was funded by the Health and Medical Research Fund of the Hong Kong Special Administrative Region (reference number 20210381). The funder had no role in the study design, data collection, analysis, interpretation, or writing of the manuscript.

Data Availability

The data supporting the findings of this study are available from the corresponding author (MCSW) upon reasonable request.

Authors' Contributions

JH, CCZ, WCWW, and MCSW conceptualized and supervised this study. CYL, XZ, MC, and CCZ were responsible for data curation and formal analysis. CCZ, MC, and MKY drafted the manuscript. JH, WCWW, and MCSW reviewed and edited the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[DOCX File , 24 KB - jmir_v28i1e78814_app1.docx]

Multimedia Appendix 2
SRQR Checklist.
[DOCX File , 22 KB - jmir_v28i1e78814_app2.docx]

References

1. Xue D, Bai Q, Bian Y. How working-age population education and health of older people shape the burden of population aging: a comparative study of Macau, Hong Kong, and Singapore. Front Public Health 2022;10:1031229 [FREE Full text] [doi: 10.3389/fpubh.2022.1031229] [Medline: 36408011]
2. Hong Kong population projections for 2022 to 2046. Census and Statistics Department. 2023. URL: https://www.censtatd.gov.hk/en/EIndexbySubject.html?pcode=FA100061&scode=190 [accessed 2026-02-17]
3. Hong Kong population projections for 2017 to 2066. Census and Statistics Department. 2017. URL: https://www.censtatd.gov.hk/hkstat/sub/sp190.jsp?productCode=FA100061 [accessed 2026-02-17]
4. Wetterling T. Pathogenesis of multimorbidity-what is known? Z Gerontol Geriatr 2021 Oct 10;54(6):590-596 [FREE Full text] [doi: 10.1007/s00391-020-01752-z] [Medline: 32651847]
5. Hu X, Huang J, Lv Y, Li G, Peng X. Status of prevalence study on multimorbidity of chronic disease in China: systematic review. Geriatr Gerontol Int 2015 Jan 27;15(1):1-10. [doi: 10.1111/ggi.12340] [Medline: 25163532]
6. Kim K, Lee JH, Kim C. Impaired health-related quality of life in elderly women is associated with multimorbidity: results from the Korean National Health and Nutrition Examination Survey. Gend Med 2012 Oct;9(5):309-318. [doi: 10.1016/j.genm.2012.08.001] [Medline: 22940276]
7. Hong Kong reference framework for preventive care for older adults in primary care settings (revised edition 2021). Health Bureau. 2021. URL: https://www.fhb.gov.hk/pho/rfs/english/reference_framework/pre_care_for_older_adults.html [accessed 2026-02-17]
8. Piccini JP, Xu H, Cox M, Matsouaka RA, Fonarow GC, Butler J, et al. Adherence to guideline-directed stroke prevention therapy for atrial fibrillation is achievable. Circulation 2019 Mar 19;139(12):1497-1506. [doi: 10.1161/circulationaha.118.035909]
9. Wang Z, Norris SL, Bero L. The advantages and limitations of guideline adaptation frameworks. Implement Sci 2018 May 29;13(1):72-13 [FREE Full text] [doi: 10.1186/s13012-018-0763-4] [Medline: 29843737]

XSL•FO
RenderX

10. Sin CK, Fu S, Tsang CS, Tsui WW, Chan FH. Prevention in primary care is better than cure: The Hong Kong Reference Framework for Preventive Care for Older Adults--translating evidence into practice. Hong Kong Med J 2015 Aug 4;21(4):353-359 [FREE Full text] [doi: 10.12809/hkmj144326] [Medline: 26238132]
11. Hong Kong Reference Framework for Diabetes Care. Primary Health Commission. URL: https://www.healthbureau.gov.hk/phcc/rfs/english/reference_framework/diabetes_care.html [accessed 2026-02-17]
12. Hypertension care. Primary Health Commission. URL: https://www.healthbureau.gov.hk/phcc/rfs/english/reference_framework/hypertension_care.html [accessed 2026-02-17]
13. World Health Organization (WHO). Active ageing: a policy framework. WHO. 2021. URL: https://iris.who.int/server/api/core/bitstreams/0418705f-1c82-4fe2-82d6-d4faee89dfa0/content [accessed 2026-02-17]
14. Jimenez G, Matchar D, Koh GCH, Tyagi S, van der Kleij RMJJ, Chavannes NH, et al. Revisiting the four core functions (4Cs) of primary care: operational definitions and complexities. Prim Health Care Res Dev 2021 Nov 10;22(1):e68 [FREE Full text] [doi: 10.1017/S1463423621000669] [Medline: 34753531]
15. Apóstolo J, Cooke R, Bobrowicz-Campos E, Santana S, Marcucci M, Cano A, et al. Effectiveness of interventions to prevent pre-frailty and frailty progression in older adults: a systematic review. JBI Database System Rev Implement Rep 2018 Jan;16(1):140-232 [FREE Full text] [doi: 10.11124/JBISRIR-2017-003382] [Medline: 29324562]
16. Atkins L, Francis J, Islam R, O'Connor D, Patey A, Ivers N, et al. A guide to using the Theoretical Domains Framework of behaviour change to investigate implementation problems. Implement Sci 2017 Jun 21;12(1):77 [FREE Full text] [doi: 10.1186/s13012-017-0605-9] [Medline: 28637486]
17. Holtrop JS, Estabrooks PA, Gaglio B, Harden SM, Kessler RS, King DK, et al. Understanding and applying the RE-AIM framework: clarifications and resources. J Clin Transl Sci 2021;5(1):e126 [FREE Full text] [doi: 10.1017/cts.2021.789] [Medline: 34367671]
18. Consolidated Framework for Implementation Research. CFIR Research Team-Center for Clinical Management Research. URL: https://cfirguide.org/ [accessed 2026-02-18]
19. Creswell J, Creswell J. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Thousand Oaks, CA: Sage Publications; 2017.
20. Barbie E. The Basics of Social Research. Belmont, CA: Thomson/Wadsworth; 2008.
21. Salkind NJ. Encyclopedia of Research Design. Thousand Oaks, CA: SAGE Publications; 2010.
22. Vasileiou K, Barnett J, Thorpe S, Young T. Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. BMC Med Res Methodol 2018 Nov 21;18(1):148-118 [FREE Full text] [doi: 10.1186/s12874-018-0594-7] [Medline: 30463515]
23. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research. Academic Medicine 2014;89(9):1245-1251. [doi: 10.1097/acm.0000000000000388]
24. Hagens V, Dobrow MJ, Chafe R. Interviewee transcript review: assessing the impact on qualitative research. BMC Med Res Methodol 2009 Jul 06;9(1):47 [FREE Full text] [doi: 10.1186/1471-2288-9-47] [Medline: 19580666]
25. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated Consolidated Framework for Implementation Research based on user feedback. Implementation Sci 2022 Oct 29;17(1):75 [FREE Full text] [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]
26. Qumseya B, Goddard A, Qumseya A, Estores D, Draganov PV, Forsmark C. Barriers to clinical practice guideline implementation among physicians: a physician survey. IJGM 2021 Nov;Volume 14:7591-7598. [doi: 10.2147/ijgm.s333501]
27. Lu AD, Kaul B, Reichert J, Kilbourne AM, Sarmiento KF, Whooley MA. Implementation strategies for frontline healthcare professionals: people, process mapping, and problem solving. J Gen Intern Med 2021 Feb 11;36(2):506-510 [FREE Full text] [doi: 10.1007/s11606-020-06169-3] [Medline: 32918200]
28. Catho G, Centemero NS, Catho H, Ranzani A, Balmelli C, Landelle C, on the behalf of the Q-COMPASS study group. Factors determining the adherence to antimicrobial guidelines and the adoption of computerised decision support systems by physicians: a qualitative study in three European hospitals. Int J Med Inform 2020 Sep;141:104233 [FREE Full text] [doi: 10.1016/j.ijmedinf.2020.104233] [Medline: 32736330]
29. Morrow S, Daines L, Wiener-Ogilvie S, Steed L, McKee L, Caress A, et al. Exploring the perspectives of clinical professionals and support staff on implementing supported self-management for asthma in UK general practice: an IMPART qualitative study. NPJ Prim Care Respir Med 2017 Jul 18;27(1):45 [FREE Full text] [doi: 10.1038/s41533-017-0041-y] [Medline: 28720751]
30. Irving G, Neves AL, Dambha-Miller H, Oishi A, Tagashira H, Verho A, et al. International variations in primary care physician consultation time: a systematic review of 67 countries. BMJ Open 2017 Nov 08;7(10):e017902 [FREE Full text] [doi: 10.1136/bmjopen-2017-017902] [Medline: 29118053]
31. Jin G, Zhao Y, Chen C, Wang W, Du J, Lu X. The length and content of general practice consultation in two urban districts of Beijing: a preliminary observation study. PLoS One 2015 Aug 10;10(8):e0135121 [FREE Full text] [doi: 10.1371/journal.pone.0135121] [Medline: 26258911]
32. Caroe TD. Consultation length. Br J Gen Pract 2019 Sep 26;69(687):488.1-48488. [doi: 10.3399/bjgp19x705665]

XSL•FO
RenderX

33. Jacob C, Sanchez-Vazquez A, Ivory C. Social, organizational, and technological factors impacting clinicians' adoption of mobile health tools: systematic literature review. JMIR Mhealth Uhealth 2020 Feb 20;8(2):e15935 [FREE Full text] [doi: 10.2196/15935] [Medline: 32130167]
34. Hiefner AR, Constable P, Ross K, Sepdham D, Ventimiglia JB. Protecting family physicians from burnout: meaningful patient-physician relationships are "more than just medicine". J Am Board Fam Med 2022 Jul 27;35(4):716-723 [FREE Full text] [doi: 10.3122/jabfm.2022.04.210441] [Medline: 35896474]
35. Bian D, Xiao Y, Song K, Dong M, Li L, Millar R, et al. Determinants influencing the adoption of internet health care technology among Chinese health care professionals: extension of the value-based adoption model with burnout theory. J Med Internet Res 2023 Mar 10;25:e37671 [FREE Full text] [doi: 10.2196/37671] [Medline: 36897630]
36. Barenfeld E, Ekman I, Cederberg M, Fors A, Ali L, Gyllensten H. Patient and healthcare professional engagement and time use within a randomised controlled trial: investigating intervention costs associated with remote person-centred care in Sweden. BMJ Open 2025 Oct 09;15(10):e099034 [FREE Full text] [doi: 10.1136/bmjopen-2025-099034] [Medline: 41067773]
37. Salmon P, Hall GM. Patient empowerment and control: a psychological discourse in the service of medicine. Soc Sci Med 2003 Nov;57(10):1969-1980. [doi: 10.1016/s0277-9536(03)00063-7] [Medline: 14499519]
38. Heen AF, Vandvik PO, Brandt L, Achille F, Guyatt GH, Akl EA, et al. Decision aids linked to evidence summaries and clinical practice guidelines: results from user-testing in clinical encounters. BMC Med Inform Decis Mak 2021 Jun 29;21(1):202 [FREE Full text] [doi: 10.1186/s12911-021-01541-7] [Medline: 34187484]
39. Ricciardi L, Mostashari F, Murphy J, Daniel JG, Siminerio EP. A national action plan to support consumer engagement via e-health. Health Aff (Millwood) 2013 Feb;32(2):376-384. [doi: 10.1377/hlthaff.2012.1216] [Medline: 23381531]
40. Jacob C, Bourke S, Heuss S. From testers to cocreators-the value of and approaches to successful patient engagement in the development of eHealth solutions: qualitative expert interview study. JMIR Hum Factors 2022 Oct 06;9(4):e41481 [FREE Full text] [doi: 10.2196/41481] [Medline: 36102548]
41. Virani SS, Ramsey DJ, Westerman D, Kuebeler MK, Chen L, Akeroyd JM, et al. Cluster randomized trial of a personalized clinical decision support intervention to improve statin prescribing in patients with atherosclerotic cardiovascular disease. Circulation 2023 May 02;147(18):1411-1413. [doi: 10.1161/CIRCULATIONAHA.123.064226] [Medline: 36871214]
42. Rödlund A, Toropova A, Lengnick-Hall R, Powell BJ, Schäfer Elinder L, Björklund C, et al. A cluster-randomized controlled trial assessing the effectiveness of a multifaceted versus a discrete implementation strategy on fidelity to an occupational guideline for the prevention of mental health problems at the workplace: a dual perspective from Swedish schools. Implement Sci Commun 2025 Nov 21;6(1):127 [FREE Full text] [doi: 10.1186/s43058-025-00821-x] [Medline: 41272910]
43. Babatunde AO, Shobanke HA, Akinade AA, Michael AJ, Osadare M, Akanbi OK, et al. Enhancing preventive medicine over curative medicine: role of telemedicine. Public Health Pract (Oxf) 2021 Nov;2:100130 [FREE Full text] [doi: 10.1016/j.puhip.2021.100130] [Medline: 36101617]
44. What's eHealth. Electronic Health Record Sharing System (eHealth). URL: https://www.ehealth.gov.hk/en/whats-ehealth/index.html [accessed 2026-02-17]
45. van Dijk CE, Verheij RA, Spreeuwenberg P, van den Berg MJ, Groenewegen PP, Braspenning J, et al. Impact of remuneration on guideline adherence: empirical evidence in general practice. Scand J Prim Health Care 2013 Mar 21;31(1):56-63 [FREE Full text] [doi: 10.3109/02813432.2012.757078] [Medline: 23330604]
46. van Dijk CE, Verheij RA, te Brake H, Spreeuwenberg P, Groenewegen PP, de Bakker DH. Changes in the remuneration system for general practitioners: effects on contact type and consultation length. Eur J Health Econ 2014 Jan 28;15(1):83-91. [doi: 10.1007/s10198-013-0458-3] [Medline: 23446626]
47. Lugtenberg M, Burgers JS, Han D, Westert GP. General practitioners' preferences for interventions to improve guideline adherence. J Eval Clin Pract 2014 Dec 21;20(6):820-826. [doi: 10.1111/jep.12209] [Medline: 24953439]
48. Waltz TJ, Powell BJ, Fernández ME, Abadie B, Damschroder LJ. Choosing implementation strategies to address contextual barriers: diversity in recommendations and future directions. Implement Sci 2019 Apr 29;14(1):42 [FREE Full text] [doi: 10.1186/s13012-019-0892-4] [Medline: 31036028]
49. Waltz TJ, Powell BJ, Matthieu MM, Damschroder LJ, Chinman MJ, Smith JL, et al. Use of concept mapping to characterize relationships among implementation strategies and assess their feasibility and importance: results from the Expert Recommendations for Implementing Change (ERIC) study. Implement Sci 2015 Aug 07;10(1):109 [FREE Full text] [doi: 10.1186/s13012-015-0295-0] [Medline: 26249843]
50. Shin M, Montano A, Adjognon O, Harvey K, Solimeo S, Sullivan J. Identification of implementation strategies using the CFIR-ERIC matching tool to mitigate barriers in a primary care model for older veterans. Gerontologist 2023 Mar 21;63(3):439-450. [doi: 10.1093/geront/gnac157] [Medline: 36239054]
51. Yakovchenko V, Lamorte C, Chinman MJ, Goodrich DE, Gibson S, Park A, et al. Comparing the CFIR-ERIC matching tool recommendations to real-world strategy effectiveness data: a mixed-methods study in the Veterans Health Administration. Implement Sci 2023 Oct 12;18(1):49 [FREE Full text] [doi: 10.1186/s13012-023-01307-x] [Medline: 37828539]

XSL•FO
RenderX

Abbreviations

CFIR: Consolidated Framework for Implementation Research
CME: Continuing Medical Education
ERIC: Expert Recommendations for Implementing Change
PCP: primary care physician
RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance
RF: Reference Framework for Preventive Care for Older Adults
SRQR: Standards for Reporting Qualitative Research

Edited by S Brini; submitted 10.Jun.2025; peer-reviewed by P Vatcharavongvan, K Mouloudj; comments to author 02.Oct.2025; revised version received 23.Oct.2025; accepted 23.Oct.2025; published 04.Mar.2026.

Please cite as:
Zhong CC, Chen M, Lo CY, YIM MK, Zhang X, Wong WCW, Huang J, Wong MCS
Implementation of Reference Framework for Preventive Care for Older Adults: Cross-sectional Qualitative Study
J Med Internet Res 2026;28:e78814
URL: https://www.jmir.org/2026/1/e78814
doi:10.2196/78814
PMID:

©Claire Chenwen Zhong, Mingtao Chen, Chung Yi Lo, Man Kin YIM, Xiaoshu Zhang, William CW Wong, Junjie Huang, Martin CS Wong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Virtual Care and Health Care Access: Pragmatic Evaluation of Implementation, Acceptance, and Use in General Practice and Aged Care Homes

Adeola Bamgboje-Ayodele[1], PhD; Tamasha Jayawardena[2], BSc; Anita Higgins[2], AssDipAppSc; Margaret Watkiss[3]; Fiona Robinson[4], PhD; Susan Kurrle[4], PhD; Pan Teng[5], BSc (Hons); Melissa Baysari[2], PhD; Meredith Makeham[4], PhD

[1]Discipline of Design, School of Architecture, Design and Planning, The University of Sydney, 148 City Road, Darlington, Australia

[2]Digital Health Human Factors Research Group, Sydney School of Nursing, Faculty of Medicine and Health, The University of Sydney, Camperdown, Australia

[3]Twilight Aged Care Homes, New South Wales, Australia

[4]Faculty of Medicine and Health, The University of Sydney, Camperdown, Australia

[5]Sydney North Health Network, New South Wales, Australia

Corresponding Author:
Adeola Bamgboje-Ayodele, PhD
Discipline of Design, School of Architecture, Design and Planning, The University of Sydney, 148 City Road, Darlington, Australia

Abstract

Background: Health care access plays a central role in reducing inequities across populations. Virtual care can mitigate these inequities by facilitating more inclusive and accessible health care delivery. In residential aged care homes (RACHs), virtual care has the potential to enable timely and efficient access to general practitioners (GPs) for residents. However, as context, technologies, and users are complex, the implementation, acceptance, and use of virtual care technologies in RACHs remain challenging.

Objective: This study aimed to evaluate the barriers and facilitators to implementing, accepting, and using virtual care technologies to connect residents with GPs for health care delivery in RACHs, and to identify benefits, unintended consequences, and opportunities for optimization.

Methods: We conducted a pragmatic, cross-sectional qualitative study guided by interpretivist principles. Semistructured interviews were undertaken with 32 participants (11 GPs, 11 RACH nurses, 3 practice managers, 5 residents, and 2 carers). Data were analyzed inductively using reflexive thematic analysis and mapped deductively to the Systems Engineering Initiative for Patient Safety model to examine sociotechnical interactions influencing virtual care delivery.

Results: Our investigation revealed that barriers to implementing, accepting, and using virtual care technologies to deliver care to RACH residents were more pervasive and salient in participants' accounts than enablers. While barriers were found across all Systems Engineering Initiative for Patient Safety domains for GPs, most of the barriers for residents and carers were identified in the "people" and "organizational" domains, and in addition to these, "technology and tools" domain for RACH nurses. Although many barriers are common across the people (eg, resistance to using new technology), technology (eg, inadequate system integration), and organization (eg, logistical challenges) domains for RACH nurses and GPs, our study revealed unique barriers to virtual care delivery for residents and carers (eg, interruptions and potential to exclude residents from conversations) whose views are often absent from existing literature. Our findings also revealed that there is no standardized virtual care consultation process between RACHs and GPs—a key concern strongly associated with the identified work system barriers. While virtual care was seen as beneficial, participants identified some unintended consequences to patients (eg, loss of doctor-patient relationship), clinicians (eg, additional workload), and health care organizations (eg, infection control).

Conclusions: Virtual care can improve access to timely, high-quality general practice services in RACHs, but its potential is constrained by sociotechnical, organizational, and workflow challenges. Addressing system integration, usability, staffing, training, and policy gaps, including funding and billing structures, will be crucial to realizing the benefits of virtual care. This study provides new evidence to inform design, implementation, and policy decisions supporting equitable virtual care delivery aligned with Sustainable Development Goals 3 and 10.

(J Med Internet Res 2026;28:e89019) doi:10.2196/89019

XSL•FO
RenderX

KEYWORDS

virtual care; implementation; acceptance; adoption; aged care; primary care

Introduction

The world is aging [1]. As the needs of an aging population are severe and complex, the demand for health care access will continue to rise [2,3]. Health care access plays a central role in reducing disparities across populations. The 2030 Agenda for Sustainable Development Goals (SDGs), adopted by all United Nations members, identified health and well-being (Goal 3) and reduced inequalities (Goal 10) as key priorities [4]. With only 5 years to reach the SDGs, pervasive global disparities continue to undermine the equitable realization of these goals [5]. Virtual care technologies have transformative potential to reduce inequities in health care by making services more inclusive and accessible. However, without inclusive design, context-aware implementation, and fair scaling, they could unintentionally reinforce existing inequities.

Virtual models of care encompass a spectrum of modalities and technological complexity. These models may involve synchronous modalities (eg, telephone or video consultations) or asynchronous modalities (eg, email or text-based communication) [6], and may be supported by technologies ranging from relatively simple tools to complex, integrated systems such as remote monitoring devices, clinical decision support, and Bluetooth-enabled diagnostic instruments [7,8]. Many benefits of virtual care have been reported in primary care settings. For example, virtual care can improve primary care efficiency by facilitating the remote triage of patients, identifying those who require urgent face-to-face care from those who can be managed virtually [9,10]. In residential aged care homes (RACHs), existing research suggests that virtual care has the potential to enable timely and efficient access to primary care providers for residents [11]. While some evidence suggests that virtual care models can reduce avoidable hospitalizations [12] and subsequent health deterioration in residents due to shorter wait times [13], context, technologies, and users are complex, and there are many reports of challenges in adoption and intended outcomes not being achieved [13]. Additionally, adoption challenges associated with telephone consultations are likely to differ from those associated with video-enabled consultations requiring peripheral devices, clinical system integration, and staff facilitation. In a previous review of 26 articles describing the experiences and perspectives of health care providers involved in virtual care services in RACHs, usability problems emerged as a key barrier to uptake [14]. This has also been identified as an issue for residents using virtual care services [15]. In addition to poor usability, barriers to virtual care uptake include end user factors (eg, hearing difficulties) and implementation factors (eg, using multiple platforms and poor workflow integration) [15,16]. Our more recent systematic review of 13 articles on the barriers, enablers, processes, and outcomes of virtual care delivery in RACH settings echoes these findings and indicates limited evidence on residents' and carers' perspectives [17]. These studies highlight the importance of virtual care aligning with residents' and health care providers' preferences and capacity, and the significance of considering the complex contexts of their use in a technology implementation (ie, the sociotechnical work system) [16].

Despite significant efforts to optimize the implementation, acceptance, and use of virtual care technologies in RACHs, these remain an ongoing challenge. Limited comprehensive evaluations that provide in-depth insights into the various sociotechnical factors that enable successful virtual care implementation in RACH settings have been conducted. While challenges related to usability, workflow integration, and system interoperability have been documented following existing evaluations [17], far less is known about how virtual care is experienced by residents of aged care homes, particularly in relation to communication, involvement in care, and the doctor-patient relationship.

This study aimed to evaluate the implementation, acceptance, and use of virtual care technologies used in RACHs to connect residents with general practitioners (GPs) for health care delivery. In this study, we conceptualize virtual care as encompassing synchronous and asynchronous consultations supported by varying levels of technological complexity. Our investigation focused on three key questions: (1) what are the barriers and enablers to implementing, accepting, and using virtual care technologies to deliver care to RACH residents from the perspective of GPs and their teams, RACH residents (or their carers), and RACH staff? (2) What are the benefits and unintended consequences of virtual care implementation on residents, clinicians, and health service organizations? And (3) How might we optimize the implementation, acceptance, and use of virtual models of care from the perspective of GPs and their teams, RACH residents (or their carers), and RACH staff?

Methods

Study Design

The study was informed by a pragmatic qualitative orientation as we focused on understanding real-world implementation challenges and informing practice improvements. Additionally, our study was guided by interpretivist principles to explore participant experiences. Accordingly, we conducted a cross-sectional qualitative study, which is a design appropriate for providing insights into people's experiences of a complex phenomenon or activity [18]. We conducted semistructured interviews to explore participants' experiences. The COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist was used to report the study (Checklist 1).

The Virtual Care Platform and Setting

The suite of virtual care technologies included a platform with videoconferencing functions, secure document and image sharing, a 6-in-1 remote monitoring device for measuring vital signs (eg, temperature and oxygenation), and other remote technologies that connected to the 6-in-1 device via Bluetooth technology (eg, digital stethoscope, otoscope, throat scope, and urine analyzer). The virtual care platform also provides

XSL•FO
RenderX

integration with some RACH and GP clinical information systems.

The suite of virtual care technologies was implemented at the Sydney North Health Network (SNHN), which spans a large geographic area (approximately 900 km^2 covering 9 local government areas). It included 103 RACHs with around 9000 beds, and 288 general practices with 756 GPs. The virtual care service was implemented to improve the level of service from GPs to RACHs. Across the RACHs that opted to use the service, SNHN provided a suite of hardware, software, and training in 2024. For this study, we included those who had implemented any type of virtual care technology, including the SNHN virtual care platform.

In Australia, GPs providing care to residents of residential aged care homes are predominantly funded through the national universal health insurance scheme, Medicare, via fee-for-service payments under the Medicare Benefits Schedule (MBS). These arrangements cover both face-to-face and telehealth consultations, including services delivered to residents in aged care settings. During the study period (2024 - 2025), telehealth consultations were a permanent feature of general practice funding in Australia.

Participant Recruitment

GP participants and RACHs were invited via the SNHN Newsletter, flyers, and targeted email invitations from SNHN. All RACHs who had engaged with a virtual care technology and all GPs across SNHN who provide services to these homes were invited to participate. General practice participants (GPs and practice managers) were invited via letters delivered by post. The managers at the RACH sites were asked to nominate staff who were involved in, or interested in, supporting a virtual care consult between a resident and GP, and these staff were invited to participate. RACH managers were also asked to identify potential residents who satisfy eligibility criteria for researchers to invite into the study. Eligibility criteria for our participant groups (ie, GPs, RACH staff, practice managers, residents, and their carers) are shown in Table 1.

Table . Eligibility criteria for the 5 participant groups.

Participant group	Eligibility criteria
GP[a]	• Current Australian Health Practitioner Regulation Agency registration as Specialist General Practitioners • Current or previous experience with visiting RACHs[b] to provide care for residents
RACH staff	• Working in a RACH in the SNHN[c] region that implemented a virtual care platform for conducting consultations with GPs
Practice manager	• Working in the general practices of consenting GPs
Resident	• Able to read, understand, and speak English • Able to provide informed consent • Aged 65 years or more • Had no severe cognitive function restrictions (eg, established dementia) • Identified as suitable by their RACH manager and usual nursing staff, based upon their clinical judgment
Carer	• Invited by their RACH resident participant and • Could read, understand, and speak English

[a]GP: general practitioner.

[b]RACH: residential aged care home.

[c]SNHN: Sydney North Health Network.

Data Collection Procedure

Data were collected from September 2024 to April 2025, when thematic saturation occurred. A semistructured interview guide was developed by the research team with expertise in qualitative research, human factors, implementation science, psychology, information technology, medicine, and nursing (refer to Multimedia Appendix 1 for the guide). The interview guide was piloted with clinician-researchers who were members of the broader team, and the guide was subsequently refined for clarity. All interviews were conducted in person or online via videoconference and audioconference, recorded and transcribed verbatim. Interviews were led by an experienced qualitative researcher (AB-A) or trained assistant (TJ), neither of whom had prior ties to the organizations. Data collection continued until thematic saturation was reached across the overall dataset, particularly for the primary stakeholder groups (GPs and RACH staff), recognizing that saturation was not sought or achieved for all participant subgroups individually.

Data Analysis

Demographic data were analyzed descriptively using Microsoft Excel. For the qualitative data, our analytical approach included data familiarization, coding, generating initial themes, reviewing potential themes, defining and naming themes, and producing the manuscript. Specifically, 2 researchers (AB-A and TJ)

XSL•FO
RenderX

familiarized themselves with the transcribed data by reading it multiple times before independent coding. Interviews were inductively analyzed using NVivo software (Lumivero) [19]. Two researchers (AB-A and TJ) independently coded the first interview transcript and met to discuss and reach consensus on codes. After initial code generation, the researchers developed themes by merging codes with a shared meaning. The remaining interviews were coded by one researcher (AB-A) using the coding framework agreed upon. Following this, the researchers reviewed the themes to ensure coherent patterns were formed, which contributed to the overall narrative and interpretation of the data. Researchers met periodically throughout data collection to discuss the interviews (AB-A and TJ).

After inductive coding, themes from interviews were deductively mapped to elements of the Systems Engineering Initiative for Patient Safety (SEIPS) model by AB-A and TJ in a 3-hour workshop. The SEIPS model focuses on how people, tasks, technologies and tools, organizational factors, and environments interact to shape care processes and outcomes [20]. Specifically, we adopted the SEIPS model to examine virtual care because it explicitly conceptualizes technology use and acceptance as outcomes of interactions within a sociotechnical work system. This perspective is particularly well-suited to residential aged care settings, where virtual care delivery is highly mediated by staff roles, workflow coordination, organizational policies, and infrastructure. By using SEIPS, we position acceptance and use of virtual care not as purely individual psychological phenomena, but as emergent properties of work system design and implementation. This enables a more comprehensive examination of barriers, enablers, unintended consequences, and opportunities for optimization across multiple stakeholder groups.

Following this, AB-A presented the codes and themes to other members of the team (MM and MB) for further refinement until consensus was reached in a 2-hour workshop. Results are presented based on the SEIPS work system elements, including people, tasks, tools, organization, and environment, identifying barriers, facilitators, and elements with dual traits (ie, viewed as both a facilitator and a barrier).

In qualitative research, 4 key criteria: credibility, confirmability, dependability, and transferability, are commonly used to strengthen the rigor and trustworthiness of a study. To enhance credibility, most (20/32) interview sessions were led by a senior researcher (AB-A) with substantial experience in digital health, virtual care, and qualitative research. The senior researcher trained, supervised, and/or was present in the remaining interviews conducted by a junior researcher (TJ). To ensure confirmability, we adopted an investigator-triangulation approach. Multiple members of the research team collaboratively reviewed themes in a 2-hour workshop and brought diverse perspectives to the interpretation of the data. Dependability was supported through the development of a detailed study protocol, drafted collaboratively by the research team in consultation with relevant stakeholders. We also ensured coding reliability by recording, transcribing, and carefully reviewing all audio data. To promote transferability, we used purposive sampling to recruit participants who represented the full range of stakeholder groups relevant to the project. We aimed to provide rich contextual information about these groups; however, identifying details were removed from all transcripts and quotations to protect anonymity and ensure that individual stakeholders could not be recognized.

Ethical Considerations

Ethics approval was obtained from the University of Sydney's Human Research Ethics Committee (HE000483). Written informed consent was obtained from all participants, who were informed of their right to withdraw from the study at any time. Participation was voluntary, with gift cards offered as compensation for their time. All identifying characteristics have been removed to protect anonymity.

Results

Demographic Characteristics

Individual semistructured interviews were conducted with 32 participants and lasted 25 minutes on average (total: 801 min). A list of the participants interviewed and their roles is provided in Tables 2 and 3.

Table . Participant demographics for the GP[a], nurse, and practice manager groups.

Participant type	Number	Age (years), mean (SD; range)	Sex, or sex assigned at birth, n (%)	Participants who have used audio-based telehealth (n)	Participants who have used video-based telehealth (n)
GP	11	42 (7.8; 25 - 54)	Female: 7 (63.6) Male: 4 (36.4)	11	2
RACH[b] nurse	11	48 (9.2; 35 - 55)	Female: 11 (100)	8	5
Practice manager[c]	3	43 (5.7, 35 - 54)	Female: 3 (100)	—[d]	—

[a]GP: general practitioner.

[b]RACH: residential aged care home.

[c]One practice staff had a dual role as a practice nurse and a practice manager.

[d]Not reported.

XSL•FO
RenderX

Table . Participant demographics for the resident and carer groups.

Participant type	Number	Age (years), mean (SD; range)	Sex, or sex assigned at birth, n (%)	Resident's LOS[a] at RACH[b], duration (months), mean (SD; range)	Audio-based telehealth, n (%)	Video-based telehealth, n (%)
Resident	5	88 (3.9; 82 - 91)	Female: 3 (60) Male: 2 (40)	11 (8.2; 3 - 36)	3 (60)	2 (40)
Carer	2	79 (7.7; 73 - 84)	Female: 1 (50) Male: 1 (50)	24 (6; 12 - 36)	1 (50)	0 (0)

[a]LOS: length of stay.

[b]RACH: residential aged care home.

All participating GPs were in practices that had a manager and a nurse on site. Most GPs serviced only one RACH (n=6), while some serviced more than one (n=3), with one servicing 20 RACHs. Insights from participants were either specific to the SNHN Virtual Care platform (with 4 RACH nurses trained to use the platform) or generally on virtual care technologies. At the general practices where our GP participants were based, the number of RACH patients serviced ranged from 25 to 70. All residents who took part in the interviews reported that their GPs had previously visited them at their RACH. Only one carer had experienced GP visits and telephone consultations together with their resident (refer to Multimedia Appendix 2 for further demographic details).

Work System Elements

An overview of how the themes align with the SEIPS model is shown in Figure 1, and a list of themes, subthemes, and illustrative participant quotes appears in Multimedia Appendix 3.

Different participant groups identified varying barriers and facilitators, with some SEIPS domains of more concern to some groups than others. A summary of the perceptions of the participant groups concerning the barriers and facilitators to virtual care delivery is shown in Table 4.

Figure 1. An overview of themes mapped to the Systems Engineering Initiative for Patient Safety model. RACH: residential aged care home.

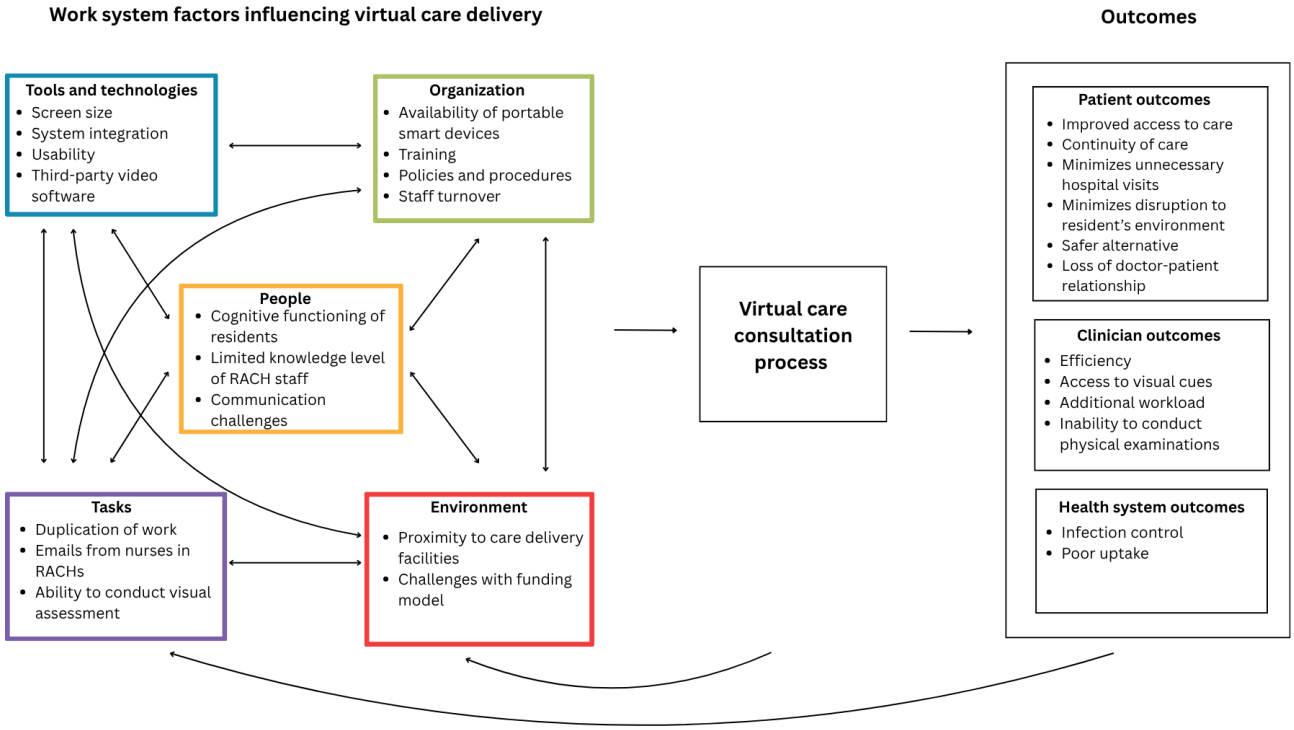

XSL•FO
RenderX

Table . Perceptions of participant groups mapped to the SEIPS[a] domains[b].

SEIPS domain and perceptions (+ enablers, – barriers)	RACH[c] staff	Primary care providers	Residents and carers
People			
Cognitive functioning of residents –	✓		
Limited knowledge level of RACH staff –	✓	✓	
Resistance to using new technology –	✓	✓	
GP[d] preference and willingness –	✓	✓	
Digital literacy ±		✓	
Communication challenges –			✓
Potential to exclude residents from conversations –			✓
Familiarity between the resident, RACH nurse, and GP ±			✓
Task			
Duplication of work –	✓		
Emails from nurses in RACHs pass along the burden of responsibility –		✓	
Inefficiency –		✓	
Ability to conduct visual assessment +		✓	
Ability to collect and share vital signs data +		✓	
Interruptions –			✓
Technology			
Screen size ±	✓	✓	
Lack of system integration –	✓	✓	
Poor usability –	✓	✓	
Poor network connectivity –	✓	✓	
Technology glitches –	✓		
Inability to conduct full assessment –		✓	
Difficulties accessing clinical information systems –		✓	
Third-party video software +		✓	
Remote access to clinical information systems +		✓	
Organization			
Awareness of virtual care options –	✓		
Availability of portable smart devices –	✓		
Training –	✓		
Management of privacy –	✓	✓	✓

XSL•FO
RenderX

SEIPS domain and perceptions (+ enablers, – barriers)	RACH[c] staff	Primary care providers	Residents and carers
Logistical challenges –	✓	✓	
Organizational policies and procedures –	✓	✓	
Availability of registered nurses during the virtual consult +	✓	✓	
Limited working hours –		✓	
Staff turnover –		✓	
Inadequate RACH staff training –			✓
Environment			
Proximity to care delivery facility –		✓	
Challenges with the funding model –		✓	

[a]SEIPS: Systems Engineering Initiative for Patient Safety.

[b]Legend: (–) indicates barrier, (+) indicates enabler, and (±) indicates both barrier and enabler.

[c]RACH: residential aged care home.

[d]GP: general practitioner.

People Domain

The complexity of a resident's care needs were reported by RACH nurses as a key factor influencing health care providers' decisions on whether a virtual or face-to-face consult was appropriate. Simple care needs with clear and well-defined problems were viewed as suitable for virtual care consultations, whereas complex care needs requiring investigations were said to be more attuned to face-to-face consultations. RACH nurses explained that patients with mental health challenges or those requiring end-of-life discussions or advanced care planning would benefit more from in-person consults.

> *...end of life care is something that I think you need, that personal touch is very important...* [P2, registered nurse]

Digital literacy levels of RACH nurses and residents were viewed by GPs as either an enabler (high level) or a barrier (low level) to visual assessments in virtual care consultations. Familiarity between the resident, RACH nurse, and GP was reported by nurses and residents as an enabler of virtual care consultations, which was seen to result in trust and a meaningful relationship being built over time. However, lack of familiarity was said to worsen the virtual care experience. Good cognitive functioning in residents enabled the acceptance of videoconferencing from the perspective of RACH nurses, whereas residents with cognitive decline were said to be hesitant, resistant, and unable to understand the virtual care consultation.

> *...it could be bit challenging, and sometimes residents with dementia here, they may not understand the process, and they are more resistant and hesitated* [P12, registered nurse]

Communication challenges were identified as barriers to virtual care, with residents reporting that these arose from hearing impairments and language differences between the resident, RACH nurse, and GP, often resulting in information being lost in translation. Additionally, both residents and carers reported that RACH residents lose their autonomy and direct access to their GP, and this concern could be further exacerbated by virtual consultations where the nurse oversees the consultation resulting in the potential to exclude residents from conversations pertaining to them.

> *He does everything I want on the phone and not direct with me, just the nurse. I tell the nurse, yeah, and she goes I will talk to him. She comes back and interprets whatever he says. I've never seen him. I don't know if he exists.* [P29, resident]

Participants' accounts suggested that resident exclusion during virtual consultations occurred through interacting factors rather than a single mechanism. For example, the physical positioning and handling of the device by the RACH nurse, combined with small screen sizes and audio limitations, could limit residents' ability to see or hear the GP clearly. These constraints were compounded for residents with hearing impairment or language differences, particularly when GPs and nurses communicated primarily with each other to progress the consultation efficiently.

Furthermore, limited virtual care knowledge levels of RACH staff were flagged by GPs and RACH nurses as another barrier to virtual care consultations, which participants noted could lead to information being passed on incorrectly, resulting in confusion. Also, resistance to new technology was reported by RACH nurses and GP participants as they explained that older GPs were perceived to have a negative perception of virtual care, which was viewed to have stemmed from previous experiences with technologies. Similarly, some RACH nurse participants were of the opinion that some GPs preferred in-person visits to virtual visits, as they believed these GPs lacked the willingness to learn new technologies.

XSL•FO
RenderX

> *...it's so frustrating to see that like us as GP...we're really resistant to change...* [P4, general practitioner]

Task Domain

RACH nurses explained that the tasks that triggered either phone or video consultation requests to GPs included wound and skin care, medication management, palliative care, case conferences, check-ups, and determining treatment plans, particularly for residents with low to medium acuity levels. Once any of these tasks triggered the need for a consultation, it was stated that the nurses then progressed to preconsultation tasks. Appointment booking was said to occur via phone calls, emails, and, on rare occasions, text messages from the RACH to the practice. The approach to booking was perceived to be determined by the urgency of the resident's concern or GP-specific practices, which might involve nurses calling the GP's practice within a specified time. Following appointment booking, it was mentioned that the nurse may be asked to collect some preconsultation observations, for example, blood pressure, urinalysis, and blood oxygen saturation. However, it was reported that this task (ie, collecting vital signs data) may not be performed unless it is specifically requested by the GP prior to the consult. From the GP's perspective, a task sometimes performed prior to consultations was the review of the resident's record. Documentation was reported to be the major postconsultation task that is completed by either or both the nurse and the GP. It was further mentioned that some documentation tasks following a virtual care consult were done via email between the GP and the RACH nurse.

> *...if it's basically weekend, or if it's basically, I am out of Sydney, so I don't have access to my own PC at home to get access to the remote access to the Best Practice. Then I will get nurse to write on my behalf…by the time that I get to the telehealth or to the Sydney I have access either to my remote control at home or at the practice, I'll do my notes as well.* [P18, general practitioner]

Barriers to appointment booking, consultation, and documentation tasks were reported. For appointment booking, GP participants explained that emails from RACH nurses pass along the burden of responsibility to GPs, which can both be challenging and consequential, resulting in delayed patient care. Participants noted that this task was experienced as burdensome in the context of existing organizational workflows and funding arrangements. This led to emails being discouraged in general practices as a communication approach with RACHs, with GP participants suggesting that RACHs should be responsible for ensuring care continuity. However, RACH nurses reported that emails are preferred as it fulfills the requirement for documentation. The main barrier to consultation was reported to be interruptions as this was flagged as a concern from the resident's perspective.

> *...he doesn't ever turn his things off properly. Even when we're in the room, there'll be people you know in the room or you can hear the click of his phone that he's got in front of him… Apparently they [other clinicians] aren't allowed to do some things without his authorisation. So rather than hold them up all the time, they come in and interrupt you as you are talking…And he tended to be combining sometimes, dealing with me and dealing with something coming over the computer...* [P31, resident]

For documentation tasks, inefficiency was highlighted as a concern, as GPs noted that it takes too much time to document and upload their notes into their own clinical information systems following a virtual consult. RACH nurses also added that the burden of documentation stemmed from mandatory reporting requirements, which were further exacerbated by duplication of work (ie, documenting the same information multiple times in multiple places).

> *When you've got to repeat what you've done potentially three or four times, you know, whether you're a PDF document or you cut and paste, it's a real issue...* [P6, registered nurse]

Enablers of virtual care consults were perceived to include the ability to conduct a visual assessment as opposed to a telehealth (audio-only) consult from the perspective of GPs. Also, GP participants highlighted that the ability to collect and share the vital signs data between the RACH and the GP facilitates improved virtual care delivery.

> *I think the biggest thing is always just, it's just being able to see the patient. I think that's the main thing. To be able to see them and talk to them and see what they're doing is a massive deal.* [P3, general practitioner]

Technology and Tool Domain

Within this domain, GPs and RACH nurse participants described a range of technologies and tools that supported virtual care delivery, as well as facilitators and barriers to the use of those technologies. Four types of technologies were described that supported or enabled virtual care, including videoconferencing platforms (eg, Zoom [Zoom Communications, Inc], FaceTime [Apple Inc], Coviu [Coviu Global Pty Ltd], and WhatsApp [Meta Platforms, Inc]), remote monitoring technologies (eg, digital stethoscope and pulse oximeter), clinical information systems (eg, electronic medication record and electronic patient record), and communication technologies (eg, email and text message). RACH nurses reported that paper-based medication charts were the only tool supporting virtual consults.

A large number of barriers and facilitators relating to technologies and tools were mentioned in interviews. Screen size was viewed to either enable (if larger) or impede (if smaller) visual assessments in virtual care consultations from the perspective of both GPs and RACH nurses. The most frequently reported barriers were lack of system integration, poor usability, and poor network connectivity. Lack of system integration was highlighted as the key source of data fragmentation in clinical information systems, with GP participants reporting that this makes documentation and continuity of care difficult to achieve in a seamless manner. Within RACH settings, GP participants further noted that different patient record systems were in place in different homes, and RACHs were using a national medication system that hospitals could not access, making continuity of care challenging. Poor usability was seen to be

XSL•FO
RenderX

another barrier to conducting virtual care consultations, with both GP and RACH nurse participants emphasizing that some virtual care technologies are not user-friendly, not easy to learn, and that infrequent use meant that users were not able to remember how to navigate them, resulting in difficulties in setting up the technology. Poor network connectivity was identified as another problem, with GP participants explaining that it could impact the video and the audio quality of a virtual care consult, resulting in the need to terminate such calls and, on occasion, opting for high-resolution photos shared by text messaging services.

> *...good internet, so enough bandwidth to facilitate a good, clear call, you know, without it being choppy. And so that's number one. And then also, you know, that affects audio. It affects the clarity of the video. You know, if I'm doing a telehealth review of the wound, if I get really grainy, terrible video, it's of no value… So sometimes, in those cases, I will then get the RN to send me a high-resolution photo via text, okay, just to back up that kind of review of the wound.* [P17, general practitioner]

The reported enablers of virtual care relating to technology and tools were the use of third-party video software and remote access to clinical information systems. GP participants explained that third-party video software such as Apple's FaceTime and Meta's WhatsApp, which are readily available on smart devices, was preferred to purpose-built videoconferencing software, as they were easy to use and saved time. Additionally, remote access to clinical information systems enabled and simplified documentation tasks associated with each consultation, as stated by practice managers and RACH nurses. Also, access to these clinical information systems was viewed by GP participants as a facilitator to prescribing medications and a means of having visibility over the patient's trajectory. Although different RACHs have different systems with varying levels of access to clinical information systems for GPs, instances where GPs have remote access to RACH systems were viewed positively.

> *...so I could, wherever I am, actually look at the notes that if they've done something, I can see. Also the prescription, the medication chart was also online as well, so it's easy when I need to prescribe something...* [P4, general practitioner]

Organization Domain

Awareness of virtual care options, availability of portable smart devices, training, and management of privacy were all seen to both impede and facilitate effective virtual care delivery. RACH nurse participants noted that it was imperative to educate residents and their families about the virtual care options available to increase awareness. Conversely, when consumers lacked awareness of the available options, RACH nurses explained that they tended to choose emergency department visits, impeding virtual care delivery. Additionally, RACH nurses said that the availability of portable smart devices enabled virtual care calls due to the ease of moving the camera of such devices (eg, smartphone) during a virtual care consultation when compared to a much bigger device (eg, computer). However, RACH nurses also stated the lack of portable smart devices resulted in them using their own personal devices to facilitate the virtual care call, highlighting the lack of organization-issued portable smart devices to support virtual care consultations.

> *...at the moment, if we do have to have a telehealth, it needs to be done by our own private phone, which is not really ideal. We don't have anything in place at this stage… [we need RACH organisations to] buy a device for the nurses, because we need [it]...* [P16, registered nurse]

Training was perceived as both enabling and hindering virtual care consultations: RACH nurses felt it helped GPs and nurses learn to use the system, while residents viewed it as inadequate. It was further reported by RACH nurses that virtual care consultation training should be extended to care staff in RACHs, as there are instances when nurses are unavailable to support the call, and participants noted that inadequate staff training results in poor resident experience during the virtual care consult. In addition, how a resident's privacy was managed during videoconferencing could either enable the consult when conducted in the resident's own room (if not shared) or impede the consult in cases where residents shared rooms. GPs expressed that maintaining privacy for residents was often not feasible, particularly when cognitive impairments meant nursing staff had to be present during consultations. Although third-party software was popular for virtual care, GPs and RACH nurses noted privacy concerns when using personal devices, as it risked sharing their mobile numbers without consent.

> *...FaceTime would be more effective, but I don't use that because it gives my phone number to the person, which I don't want to do...* [P20, general practitioner]

The most frequently reported organization-level barriers to virtual care consultations were logistical challenges and organizational policies and procedures. Logistical challenges in organizing virtual care calls often impeded consultations. GP and RACH nurse participants highlighted the absence of a clear scheduling structure, with difficulties in aligning the right people at the right time. These issues persisted across both video and audio modalities. Organizational policies also created barriers: some RACHs required staff to transfer residents to the hospital after any incident, regardless of severity, limiting opportunities for virtual care. GP participants further noted that policy development has lagged behind the implementation of virtual care technologies.

> *...an issue is the company policies on various issues that happens or incidents that happen. We have policies that mandate that we have to transfer to people to hospital to be investigated…so that's something to think [about] and each organisation has different policies and procedures around transfer to hospital...* [P2, registered nurse]

The availability of registered nurses during the virtual consult was identified by GP participants as an important organizational enabler to their ability to undertake a virtual care consultation with a resident.

XSL•FO
RenderX

Environment Domain

A total of 3 unique barriers to virtual care delivery were identified in this domain: 2 in the internal environment and 1 in the external environment. Within the internal environment, practice managers and GP participants reported that virtual care was usually not necessary as they were physically located in proximity to the RACH. Noise was seen as another barrier to virtual care consults, as practice manager participants noted that the presence of noise during the call could hamper communication.

> *the voice [was] not clear due to the background noise...* [P10, practice manager]

Within the external environment, a challenge with the funding model was flagged as a deterrent to conducting virtual care in RACHs. GP participants reported that the process of receiving payments was often laborious and frustrating, and that the requirement to always have the client on the call is not realistic. This barrier was said to have resulted in unpaid work or GPs choosing to abandon virtual care as a means of conducting consultations with RACH residents altogether.

> *So there's still a lot of unpaid work being done at my nursing home...which is really like frustrating.*
>
> *[P3, general practitioner]*

Virtual Care Consultation Process

RACH nurse participants reported that there was no standard way of conducting virtual care consultations between RACHs and GPs. They also noted varying approaches to conducting a virtual care consultation with differences occurring based on urgency, capacity of the resident, availability of a registered nurse, technologies available, time of day, and other work system factors. It was also highlighted by GP participants that nurses in RACHs often sought informal consultations from GPs about residents via text messages and emails without booking appointments. This was sometimes viewed by GPs as unpaid work due to the difficulty in billing processes for informal consultations that were not booked appointments directly involving the resident.

> *We don't have a process as such...We don't have that. I have never even looked into if there is a policy of the video call, what exactly you're supposed to do.* [P16, registered nurse]

Outcomes

Within this domain, participants reported positive and negative outcomes of virtual care delivery for residents, clinicians, and health services. For residents, the most significant benefit GP participants identified was improved access to care for residents in RACHs. Other benefits noted by RACH nurses and GPs included reduced unnecessary hospital visits and fewer disruptions to residents' familiar environment. However, the potential loss of the doctor-patient relationship was flagged by GP participants as an unintended consequence of virtual care delivery in RACHs.

> *There's something that happens when you're in a room with someone...it's just a human-to-human interaction. Yeah, there's something that it's just lost ... you know that patient-doctor relationship, there is something about...meeting people, safe in their hand.* [P17, general practitioner]

GP participants stated that virtual care delivery saved clinicians travel time, which enabled faster access to care for more residents. However, they also highlighted some inefficiencies due to the lack of integration of clinical information systems and logistical challenges. Additionally, GPs mentioned access to visual cues as a benefit of virtual care consultations with video. Both GP and RACH nurse participants stated that the unintended consequences of virtual care delivery for clinicians were additional workload and the inability to conduct physical examinations.

> *And there's other things... they've got to do with assessment and maintaining the aged care quality standards ... that keeps them time poor ... because they're trying to do this paperwork. And, I think those reasons are the reason they see this as [extra workload] ...the impact of this on our already busy life, or work life, does put a strain on them…and the workload. The workload is huge.* [P6, registered nurse]

For health services, RACH nurses explained that virtual care consultations minimize the spread of infections (as seen during the COVID-19 pandemic) but simultaneously create the need for additional infection control procedures in RACHs, for example, disinfecting devices like tablets and phones. Poor uptake of virtual care was viewed by GPs as an unintended consequence of poor implementation in RACHs, with many reports of virtual care technologies not being used and resulting in desired outcomes not being achieved. When nurses were asked how often virtual care calls were conducted in their RACHs, many reported that it was very infrequent.

Recommendations

Participants provided some recommendations for improving virtual care delivery between RACHs and general practices, as shown in Table 5.

XSL•FO
RenderX

Table . Recommendations for improving virtual care delivery.

Theme and code	Quote
External environment	
Improve funding model	• "I think that [NDIS funding] model is much better suited because a lot of the Medicare rules actually just don't work for nursing home patients, especially with the patients with dementia. Technically, if I bill for 20 minutes of my time, I have to have spent all that 20 minutes with you, not talking to the nurse or… And actually, Medicare says that all that doesn't count … So it is very not compatible. And secondly, it doesn't allow you to collaborate with other people, where NDIS does that all the time, … I, as a GP, cannot speak to a cardiologist about your care and bill for that… I can't bill the cardiologist for taking my time, and I can't bill you" (P13, GP)
Organization-focused implementation strategies	
Create efficient process for video calls in RACHs[a]	• "So having that ability, that knowing that … you can still do that telehealth conference without having to first call them and say, Hey, I'm about to call you on the program" (P15, GP)
Provide more realistic and ongoing training	• "So I think telehealth is a great intervention. Maybe more training, doctors need to try it out for themselves and to be with the patients and see what does work for them and what doesn't" (P31, resident)
Broaden population of staff who receive training	• "Although it would be good if my nurse got trained…because my nurse could help with some of it too, and she could set up, yeah, she could almost get it ready for me, and so that I'm not wasting my time waiting for them" (P3, GP)
Provide dedicated smart devices in RACHs	• "there needs to be a dedicated phone that the doctor can call and arrange an appointment with it" (P3, GP)
Provide stepwise resident onboarding	• "introduce this stuff to residents little bit at a time, given the reason why it's happening, present it at their meetings. Tell them where we're going to use it, why we're going to use it. Ask them why they think that's not going to be the right thing to do, because it's really just the doctors that are going to be at the other end of the machine" (P7, RN)
Promote virtual care via success stories	• "if you have a nursing home that seems to have a very smooth use of this and an efficient use of this, then often one person showing another person how it's good" (P4, GP)
Use word of mouth approach to promote virtual care	• "I think part of it maybe, is probably with no word of mouth, actually, like one GP telling another GP" (P4, GP)
Technology-focused strategies	
Provide system integration	• "I think virtual care would be amazing if they can start to logistically integrate everything into the one database that people can access" (P5, GP)
Provide an integrated app for virtual care	• "Can we have national thing? Can we have an app we can download… I just think if the Department of Health, they can start an app just to facilitate that part of the consultation, that would be cool" (P1, RN) • "If I had an app that could be on my phone, like, theoretically, they could set up a consult with all the obs like that would be ideal if they said, here's your consult coming in, and then I take it … almost like an Uber Doctor consult" (P3, GP)
Provide remote monitoring devices	• "I think a [digital] stethoscope would be one of the most useful things, because I feel as though with, you know, with guidance from a very senior clinician, you could probably direct a relatively well skilled nurse, to elicit the information that you require, just through placement…I would love to utilise those types of things, because I think that is useful" (P27, RN)

XSL•FO
RenderX

Theme and code	Quote
Ensure technology usability	• "I guess the set-up steps needs to be very specific, like very detailed. It shouldn't be a lot of steps… So it needs to be very simple, like telephone simple, like they only need to click a few buttons to get connected…Because it's kind of like a big difference from Zoom or Teams like, you have to go into a waiting room until you got connected" (P9, RN)
Consider providing real-time language translation	• "Heidi [platform] and all are working on it, the ability to translate in real time would be really good" (P13, GP)
Provide 3D rotation technology	• "some type of 3d rotation thing with a picture, [to] try to take picture of some ulcer, or some lesion" (P11, GP)

[a]RACH: residential aged care home.

Discussion

Principal Findings

In this study, we sought to evaluate the implementation, acceptance, and use of virtual care technologies used in RACHs to connect residents with general practitioners for health care delivery. Our investigation revealed that barriers to implementing, accepting, and using virtual care technologies to deliver care to RACH residents were more pervasive and salient in participants' accounts than enablers. While barriers were found across all SEIPS domains for primary care providers, we mostly identified barriers for residents and carers in the "people" and "organizational" domains, and in the same domains, including "technology and tools" for RACH staff. Although many barriers are common across people (eg, resistance to using new technology), technology (eg, lack of system integration), and organization (eg, logistical challenges) domains for RACH staff and primary care providers, our study revealed unique barriers to virtual care delivery from the perspective of residents and carers (eg, interruptions, potential to exclude residents from conversations). This provides an important perspective, as the views of residents and their carers are largely absent from the existing literature describing RACH and general practice virtual care. Our findings also revealed that there is no standardized virtual care consultation process between RACHs and GPs. This was a key concern strongly associated with the identified work system barriers (eg, logistical challenges). Additionally, we found that virtual care delivery in RACH settings was largely beneficial to patients, but we also identified some unintended negative consequences, including loss of doctor-patient relationships, additional workload for clinicians, and impacts on health care organizations (increased requirements for infection control)—which warrant further investigation.

Our findings shed light on residents' perceptions of virtual care delivery, noting interruptions, potential to exclude residents from conversations, communication challenges, inadequate RACH staff training, and poor management of privacy as key barriers. Our recent systematic review of 13 articles on the barriers, enablers, processes, and outcomes of virtual care delivery in RACHs and primary care identified only one article providing insights on residents' views [17]. The article explored residents' views on ethical challenges associated with the use of virtual care and identified communication barriers and privacy concerns, aligning with our findings [21]. From the residents' perspectives, our study highlights 2 novel findings: interruptions to GPs during virtual consultations and the risk of residents being excluded from virtual care conversations. These novel barriers to virtual care delivery can negatively impact residents' experiences and can contribute to the loss of the doctor-patient relationship, which was identified as an unintended consequence. Viewed through a sociotechnical lens, our findings illustrate how interactions between tools (device form factor and audio-visual quality), people (residents' sensory and cognitive capacities), and tasks (clinically focused communication between staff) can unintentionally marginalize residents within virtual care encounters, despite the technology being introduced to improve access. These findings also raise ethical considerations regarding how residents are positioned within virtual care encounters. Participants' accounts suggest that virtual care may risk marginalizing residents' experiential knowledge and involvement in decision-making. Such interactional shifts help explain participants' perceptions of a loss of the doctor-patient relationship and highlight the importance of designing virtual care models that actively support residents' participation to maintain person-centered care.

Consistent with our study findings, inadequate RACH staff training is a well-documented barrier to virtual care delivery in the literature [17]. Although existing studies have recommended effective and regular training on virtual care technologies in RACH settings [22,23], our findings were aligned with other studies that highlighted the barriers to achieving this due to the unique characteristics of this setting (eg, high workload, high staff turnover rates, and low digital literacy) [11,22,24-27]. These factors make it challenging to rely on training, and there have been calls for virtual care technologies in RACH settings to be easy to learn [28].

This study highlighted that no standardized virtual care consultation process currently exists between RACHs and GPs within the context investigated. This finding aligns with our existing systematic review that found that none of the 13 articles reviewed provided details on their virtual care consultation process [17]. Previous studies have clearly identified how virtual care is performed using process maps in virtual hospitals [29] and clinical pharmacy settings [30], but this level of detail is lacking in RACH and GP virtual care delivery settings. One

XSL•FO
RenderX

reason for this may be that RACH organizations are yet to define their virtual care processes, procedures, or workflows. Although virtual care frameworks and operational guidance exist within Australian health systems, these are not consistently translated into procedural governance at the RACH level, suggesting challenges related to implementation and dissemination rather than an absence of guidance. At the work system level, the absence of a standardized virtual care consultation process reflects an organizational gap in the provision of formal structures, such as clear procedures and workflows. In the absence of these structures, GPs and RACH staff were required to compensate through increased coordination, decision-making, and cognitive effort to organize and conduct consultations. This reliance on informal, ad hoc processes helps explain the additional workload reported by participants and illustrates how organizational gaps translate into process inefficiencies and increased burden at the individual level. Consequently, existing research has called for standardized virtual care processes to be created within RACH and GP settings, and our findings further support this [31,32].

From a SEIPS perspective, the burden associated with asynchronous email communication reflects an interaction between tasks, organizational workflows, and the external funding environment. While email correspondence emerged as a task performed by GPs, it was experienced as burdensome largely because it is not routinely recognized or remunerated within existing Medicare funding structures. As a result, this form of work becomes effectively "invisible," contributing to clinician workload and shaping perceptions of virtual care as inefficient or unsustainable. This finding highlights how task-level burdens cannot be fully understood without considering the organizational and policy contexts in which they occur. Recent policy reforms intended to strengthen continuity of care, such as the introduction of voluntary patient registration models, may partially address fragmentation in the general practice relationship with RACH settings. However, where virtual care continues to rely predominantly on fee-for-service billing for discrete clinical encounters, forms of asynchronous coordination such as email triage and communication between RACH staff and GPs are likely to remain undervalued or unpaid. As such, the "invisible" work identified by participants may persist despite policy reform, reinforcing the need for funding models that better align with the realities of virtual care delivery in RACH settings.

Virtual care funding models remain a topic of concern emerging from this study. A scoping review of 23 articles investigated health insurance payment models for virtual care (telehealth) services and found variability across countries [33]. For example, China uses region-specific insurance quotas and reimbursement rates based on hospital level or service price, resulting in varied coverage across provinces [33]. France offers national health insurance that covers 70% of costs, with full coverage for certain groups during the COVID-19 pandemic [33]. In Australia, the universal health system, funded through Medicare, provides low-cost access to public hospital and primary care services [34]. Primary care is largely delivered by private providers in community settings, with patients paying fees offset by standardized Medicare rebates via the MBS [34]. In 2021, virtual care (telehealth) became a permanent feature of Australian primary care with policy rapidly adapted to include changes to MBS items, practice incentive payments, increased scope for GPs to accept Medicare-funded bulk billing payments for telehealth services, and continuity provisions [34,35]. However, our findings reveal that existing funding structures inadequately address the needs of GPs who deliver care to RACHs due to the billing rules, logistical challenges, and a lack of standardized virtual care consultation process. These result in limited collaboration with specialists, challenges in conducting virtual care consults for patients with certain conditions (eg, dementia), clinician frustration, unpaid work, and the propensity to boycott virtual care delivery in RACHs altogether, limiting health care access to residents. With the promise of virtual care being thwarted by poor funding structures and challenges with policy alignment, our findings shed light on how virtual care implementation within this setting in Australia undermines the equitable realization of the SDGs.

The benefits of virtual care delivery in RACHs are well-documented and are consistent with our findings [17]. However, unintended consequences (eg, additional workload, poor uptake, and additional infection control) continue to limit the realization of intended benefits. In our study, virtual care introduced an important system trade-off: while remote consultations reduced GP travel time and supported continuity of care, they simultaneously created additional tasks for RACH nursing staff, such as the new requirement to be present during virtual GP visits and additional infection control tasks related to cleaning and disinfecting shared devices. This illustrates how gains in efficiency for one part of the work system may redistribute workload elsewhere, shaping how virtual care is experienced and sustained in practice. Beyond these trade-offs, our findings also highlight broader systemic constraints. Although virtual care technologies could be more usable and easier to learn, challenges such as lack of system integration and poor network connectivity played a significant role in negatively impacting virtual care delivery, consistent with existing evidence [17,36]. Lack of system integration contributed to duplication of work, particularly when performing virtual care documentation tasks, resulting in additional workload—an unintended consequence for clinicians. Taken together, these findings stress the need for concerted efforts at the macro (government), meso (primary health networks), and micro (RACHs and general practices) levels to tackle work system barriers to ensure the benefits of virtual care can be fully realized, thereby contributing to the earnest achievement of the SDGs.

Strengths and Limitations

To the best of our knowledge, this is the first study to identify domains of concern to different stakeholder groups, using the SEIPS model to examine work system facilitators and barriers to virtual care delivery between general practices and RACH settings. Our qualitative study was strengthened by the large number of participants (n=32) from multiple RACHs and general practices to provide a complete picture of virtual care delivery between aged care homes and primary care from the perspectives of all relevant stakeholder groups.

XSL•FO
RenderX

Most participants in this study were either RACH staff or primary care providers, whose perspectives were shaped by their roles and responsibilities regarding virtual care delivery. While their insights provided valuable depth regarding work system barriers and enablers, the limited representation from residents and carers may have constrained the breadth of perspectives captured and may further constrain the transferability of findings. It is possible that greater participation from these groups could have yielded additional insights. Future research may benefit from a more balanced representation across stakeholder groups to enrich understanding of virtual care delivery in this setting. To achieve this, future research would benefit from purposive, consumer-led recruitment strategies and methods tailored to the capacities and preferences of older adults and their carers. Also, our findings are based on data collected across one large metropolitan region of New South Wales, Australia, which may limit transferability to other contexts and settings (eg, rural and remote regions). Strengthening consumer representation will be essential for developing virtual care models that are genuinely person-centered and equitable.

Conclusion

This study generates new evidence to guide how GPs deliver virtual care in RACHs, supporting residents' access to general practice services and promoting more timely, high-quality care. Our findings have implications for health care organizations seeking to involve GPs in virtual care within RACHs. They also inform investment in virtual care models and technologies, as well as education and training for GPs and RACH staff, alongside policy development by peak bodies and governments. These efforts contribute to SDG 3 (Good Health and Well-Being) by improving residents' access to timely, high-quality general practice services and to SDG 10 (reduced inequalities) by promoting equitable access to care in aged care settings.

Acknowledgments

We would like to thank the Royal Australian College of General Practice (RACGP) and the Medibank Foundation for funding the study.

During the preparation of this manuscript, Microsoft Copilot was used for minor sentence refinement. The authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Funding

This study was funded by the Royal Australian College of General Practice (RACGP) and Medibank foundation. The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated and analyzed during this study are not publicly available, as all identifying characteristics have been removed to protect participant anonymity.

Authors' Contributions

AB-A, MB, and MM conceptualized the research design. AB-A, MM, MB, SK, FR, PT, and MW contributed to funding acquisition. PT and MW contributed to project resources. AB-A and TJ collected and analyzed the data. AB-A drafted the manuscript. All authors contributed to data interpretation and manuscript writing and read and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[DOCX File, 19 KB - jmir_v28i1e89019_app1.docx]

Multimedia Appendix 2
Participant demographic details.
[DOCX File, 15 KB - jmir_v28i1e89019_app2.docx]

Multimedia Appendix 3
Systems Engineering Initiative for Patient Safety domains, subthemes, and quotes.
[DOCX File, 38 KB - jmir_v28i1e89019_app3.docx]

Checklist 1
COREQ checklist.

XSL•FO
RenderX

[DOCX File, 17 KB - jmir_v28i1e89019_app4.docx]

References

1. World Population Prospects 2022: Summary of Results: United Nations Department of Economic and Social Affairs; 2022.
2. Final report: care, dignity and respect. : Royal Commission into Aged Care Quality and Safety; 2021 URL: https://www.aph.gov.au/About_Parliament/Parliamentary_departments/Parliamentary_Library/Research/Quick_Guides/2020-21/AgedCareQuality#heading_257d83dd2e5441eb98239f8522fe1464 [accessed 2026-06-08]
3. Lind KE, Raban MZ, Brett L, Jorgensen ML, Georgiou A, Westbrook JI. Measuring the prevalence of 60 health conditions in older Australians in residential aged care with electronic health records: a retrospective dynamic cohort study. Popul Health Metr 2020 Oct 8;18(1):25. [doi: 10.1186/s12963-020-00234-z] [Medline: 33032628]
4. Goal 10: reduce inequality within and among countries. United Nations Sustainable Development Goals. 2015. URL: https://www un org/sustainabledevelopment/inequality [accessed 2026-05-29]
5. The Sustainable Development Goals Report 2025. : United Nations Department of Economic and Social Affairs; 2025 URL: https://unstats.un.org/sdgs/report/2025/ [accessed 2026-06-08]
6. Mold F, Cooke D, Ip A, Roy P, Denton S, Armes J. COVID-19 and beyond: virtual consultations in primary care-reflecting on the evidence base for implementation and ensuring reach: commentary article. BMJ Health Care Inform 2021 Jan;28(1):e100256. [doi: 10.1136/bmjhci-2020-100256] [Medline: 33436372]
7. Hutchings OR, Dearing C, Jagers D, et al. Virtual health care for community management of patients with COVID-19 in Australia: observational cohort study. J Med Internet Res 2021 Mar 9;23(3):e21064. [doi: 10.2196/21064] [Medline: 33687341]
8. Fronczek AE. Nursing theory in virtual care. Nurs Sci Q 2019 Jan;32(1):35-38. [doi: 10.1177/0894318418807926] [Medline: 30798757]
9. Monaghesh E, Hajizadeh A. The role of telehealth during COVID-19 outbreak: a systematic review based on current evidence. BMC Public Health 2020 Aug 1;20(1):1193. [doi: 10.1186/s12889-020-09301-4] [Medline: 32738884]
10. Lin CH, Tseng WP, Wu JL, et al. A double triage and telemedicine protocol to optimize infection control in an emergency department in Taiwan during the COVID-19 pandemic: retrospective feasibility study. J Med Internet Res 2020 Jun 23;22(6):e20586. [doi: 10.2196/20586] [Medline: 32544072]
11. Savira F, Gupta A, Gilbert C, et al. Virtual care initiatives for older adults in Australia: scoping review. J Med Internet Res 2023 Jan 18;25:e38081. [doi: 10.2196/38081] [Medline: 36652291]
12. Hofmeyer J, Leider JP, Satorius J, Tanenbaum E, Basel D, Knudson A. Implementation of telemedicine consultation to assess unplanned transfers in rural long-term care facilities, 2012-2015: a pilot study. J Am Med Dir Assoc 2016 Nov 1;17(11):1006-1010. [doi: 10.1016/j.jamda.2016.06.014] [Medline: 27477614]
13. Wade V, Whittaker F, Hamlyn J. An evaluation of the benefits and challenges of video consulting between general practitioners and residential aged care facilities. J Telemed Telecare 2015 Dec;21(8):490-493. [doi: 10.1177/1357633X15611771] [Medline: 26556062]
14. Tan AJ, Rusli KD, McKenna L, Tan LL, Liaw SY. Telemedicine experiences and perspectives of healthcare providers in long-term care: a scoping review. J Telemed Telecare 2024 Feb;30(2):230-249. [doi: 10.1177/1357633X211049206] [Medline: 34666535]
15. Mao A, Tam L, Xu A, et al. Barriers to telemedicine video visits for older adults in independent living facilities: mixed methods cross-sectional needs assessment. JMIR Aging 2022 Apr 19;5(2):e34326. [doi: 10.2196/34326] [Medline: 35438648]
16. Warmoth K, Lynch J, Darlington N, Bunn F, Goodman C. Using video consultation technology between care homes and health and social care professionals: a scoping review and interview study during COVID-19 pandemic. Age Ageing 2022 Feb 2;51(2):b279. [doi: 10.1093/ageing/afab279]
17. Makeham MAB, Jayawardena T, Elkheir S, et al. Virtual care in residential aged care and primary care settings: a systematic literature review using the SEIPS framework. Int J Med Inform 2026 Mar 1;207:106218. [doi: 10.1016/j.ijmedinf.2025.106218] [Medline: 41411901]
18. Creswell JW, Creswell JD. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches: Sage Publications; 2017.
19. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Aug 8;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
20. Holden RJ, Carayon P, Gurses AP, et al. SEIPS 2.0: a human factors framework for studying and improving the work of healthcare professionals and patients. Ergonomics 2013;56(11):1669-1686. [doi: 10.1080/00140139.2013.838643] [Medline: 24088063]
21. Sira N, Groß D, Optimal@NRW Research Group, Wilhelmy S. Ethical acceptability of telemedicine: nursing home resident's perspective on telemedical consultations. Digit Health 2024;10:20552076241288368. [doi: 10.1177/20552076241288368] [Medline: 39484651]
22. Banbury A, Taylor ML, Gray LC, Reid N, Smith AC. Sustaining and expanding telehealth activity: training requirements for Australian residential aged care front-line staff. PEC Innov 2023 Dec;2:100109. [doi: 10.1016/j.pecinn.2022.100109] [Medline: 37214526]

XSL•FO
RenderX

23. Kuek A, Hakkennes S. Healthcare staff digital literacy levels and their attitudes towards information systems. Health Informatics J 2020 Mar;26(1):592-612. [doi: 10.1177/1460458219839613] [Medline: 30983476]
24. Hodroj B, Way KA, Scott TL, Wright AL, Manchha A. Does context count? The association between quality of care and job characteristics in residential aged care and hospital settings: a systematic review and meta-analysis. Gerontol 2023 Jul 18;63(6):1012-1027. [doi: 10.1093/geront/gnac039] [Medline: 35323966]
25. Thwaites C, McKercher JP, Fetherstonhaugh D, et al. Factors impacting retention of aged care workers: a systematic review. Healthcare (Basel) 2023 Nov 21;11(23):3008. [doi: 10.3390/healthcare11233008] [Medline: 38063576]
26. Manchha AV, Walker N, Way KA, Dawson D, Tann K, Thai M. Deeply discrediting: a systematic review examining the conceptualizations and consequences of the stigma of working in aged care. Gerontol 2021 Jun 2;61(4):e129-e146. [doi: 10.1093/geront/gnaa166] [Medline: 33103188]
27. Dyrbye LN, Shanafelt TD, Sinsky CA, et al. Burnout among health care professionals: a call to explore and address this underrecognized threat to safe, high-quality care. NAM Perspectives.: National Academy of Medicine; 2017 Jul 5 URL: http://nam.edu/perspectives/?sf_paged=7 [accessed 2026-06-08]
28. Bamgboje-Ayodele A, Makeham M, Higgins A, et al. Usability evaluation of a virtual care platform for residential aged care homes: a cognitive walkthrough study. SSRN. Preprint posted online on Jul 11, 2025. [doi: 10.2139/ssrn.5346704]
29. Bamgboje-Ayodele A, Boscolo A, Newton N, et al. Detection and management of clinical deterioration in a virtual hospital: a work system analysis. Appl Ergon 2025 Sep;127:104530. [doi: 10.1016/j.apergo.2025.104530] [Medline: 40215709]
30. Marchese M, Heintzman A, Pasetka M, Charbonneau F, DeAngelis C, Peragine C. Development of a process map for the delivery of virtual clinical pharmacy services at Odette Cancer Centre during the COVID-19 pandemic. J Oncol Pharm Pract 2021 Apr;27(3):650-657. [doi: 10.1177/1078155221991202] [Medline: 33554738]
31. Warmoth K, Bennett C, Lynch J, Goodman C. Using online consultations to facilitate health and social care delivery during COVID-19: an interview study of care home staff. J Long Term Care 2023 Jun 9:100-109. [doi: 10.31389/jltc.184]
32. Sunner C, Giles M, Parker V, Kable A, Foureur M. COVID-19 preparedness in aged care: a qualitative study exploring residential aged care facility managers experiences planning for a pandemic. J Clin Nurs 2025 Dec;34(12):5249-5259. [doi: 10.1111/jocn.15941] [Medline: 34254376]
33. Li Q, Cheng F, Zeng H, Xu J. Health insurance payment for telehealth services: scoping review and narrative synthesis. J Med Internet Res 2024 Dec 9;26:e56699. [doi: 10.2196/56699] [Medline: 39652868]
34. Hall Dykgraaf S, Desborough J, de Toca L, et al. “A decade’s worth of work in a matter of days”: the journey to telehealth for the whole population in Australia. Int J Med Inform 2021 Jul;151:104483. [doi: 10.1016/j.ijmedinf.2021.104483] [Medline: 33984625]
35. Fisher K, Davey AR, Magin P. Telehealth for Australian general practice: the present and the future. Aust J Gen Pract 2022 Aug;51(8):626-629. [doi: 10.31128/AJGP-11-21-6229] [Medline: 35908757]
36. Haines SJ, FitzGerald J, Clune S, et al. Virtual nursing in residential aged care: what is known? A rapid review. J Adv Nurs 2026 Apr;82(4):2810-2826. [doi: 10.1111/jan.70071] [Medline: 40726070]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
GP: general practitioner
MBS: Medicare Benefits Schedule
RACH: residential aged care home
SDG: Sustainable Development Goal
SEIPS: Systems Engineering Initiative for Patient Safety
SNHN: Sydney North Health Network

Edited by A Stone; submitted 05.Dec.2025; peer-reviewed by G Davlyatov, J Singh; revised version received 15.Apr.2026; accepted 16.Apr.2026; published 12.Jun.2026.

Please cite as:
Bamgboje-Ayodele A, Jayawardena T, Higgins A, Watkiss M, Robinson F, Kurrle S, Teng P, Baysari M, Makeham M
Virtual Care and Health Care Access: Pragmatic Evaluation of Implementation, Acceptance, and Use in General Practice and Aged Care Homes
J Med Internet Res 2026;28:e89019
URL: https://www.jmir.org/2026/1/e89019
doi:10.2196/89019

XSL•FO
RenderX

© Adeola Bamgboje-Ayodele, Tamasha Jayawardena, Anita Higgins, Margaret Watkiss, Fiona Robinson, Susan Kurrle, Pan Teng, Melissa Baysari, Meredith Makeham. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Institutionalizing Digital Parenting Programs in Low Resource Settings in China: Comparative Case Study of Health Care and Education Sectors Using the RE-AIM Framework

Xinyu Shi[1], BA; Ruochen Ruan[2], MEd; Yi Qie[3,4], MA; Jamie M Lachman[4,5,6], PhD; Na Zhong[7], MA; Zuyi Fang[8,9], PhD

[1]School of Government, Beijing Normal University, Beijing, China
[2]Faculty of Education, Beijing Norma University, Beijing, China
[3]China Development Research Foundation (CDRF), Child Development Research Institute, Beijing, China
[4]Department of Social Policy and Intervention, University of Oxford, London, United Kingdom
[5]Centre for Social Science Research, University of Cape Town, Rondebosch, Cape Town, South Africa
[6]Parenting for Lifelong Health, Barnett House, Oxford, United Kingdom
[7]Chengbei Preschool, Xinyu, China
[8]Institute of Population Research, Peking University, No.5 Yiheyuan Road, Haidian District, Beijing, China
[9]Department of Social Policy and Intervention, University of Oxford, Oxford, United Kingdom

Corresponding Author:
Zuyi Fang, PhD
Institute of Population Research, Peking University, No.5 Yiheyuan Road, Haidian District, Beijing, China

Abstract

Background: Digital parenting programs offer a promising way to disseminate evidence-based parenting knowledge and support early childhood development. They help reduce costs while improving scalability and fidelity. However, their successful implementation is context-dependent, and existing research offers limited guidance on how the implementation of digital parenting interventions unfolds across diverse settings.

Objective: This study aims to identify the shared and unique facilitators and barriers affecting each dimension of implementation, as well as differentiated mechanisms that support the effective implementation and institutionalization of such interventions across diverse settings.

Methods: Using a multiple-case study design, this research compared the implementation of a digital (chatbot-led) parenting program across 2 distinct settings in China: urban educational and rural health care contexts. The intervention content remained consistent, while the contexts and formats of local human-led support differed. Guided by the RE-AIM framework, this study examines the program's reach, adoption, implementation, and maintenance in both settings. Data sources included program documents, field observations, semistructured interviews, and focus group discussions with 83 stakeholders. Thematic analysis was conducted using ATLAS.ti until thematic saturation was reached.

Results: Data were collected from 83 stakeholders, and findings are based on an analysis of 18 interviews and 4 focus groups with caregivers, village doctors, and health officials from the rural health care setting, and 29 interviews and 4 focus groups with caregivers, teachers, social workers, and managers from the urban educational setting. Regarding reach, strong relationships between parents and implementers and the credibility of program developers were shared facilitators in both settings. Parenting conservatism and limited understanding of the program were shared barriers. In rural health care settings, parents' perception of village doctors as lacking parenting expertise posed an additional challenge. For adoption, trust between managers and program developers, program alignment with organizational functions, and organizational empowerment supported implementation are shared facilitators in both settings. At the individual level, task-driven motivation helped, while time constraints hindered adoption in the health care setting. Teachers adopted the program due to its relevance to their roles in the educational setting, unlike village doctors, who did not see it as part of their core duties. For implementation, supportive management and clear guidelines were shared facilitators in both settings, while a lack of purpose and psychological pressure acted as barriers. Rural implementation was aided by scheduling during off-seasons and standardized workflows, whereas flexible workflows were essential in the educational setting. Regarding maintenance, alignment with organizational functions and internal resources facilitated sustainability in both settings, while overreliance on government authorization posed challenges. Educational settings required contextual adaptation, while health care settings needed more content adaptation.

XSL•FO
RenderX

Conclusions: Implementing digital parenting programs is a complex process, influenced by multilevel facilitators and barriers that vary across regions (rural vs urban) and settings (educational vs health care). This study highlights the importance of context-specific implementation strategies and proposes differentiated delivery models tailored to local structures and needs.

(J Med Internet Res 2026;28:e79848) doi:10.2196/79848

KEYWORDS

digital delivery; parenting program; implementation science; Reach, Effectiveness, Adoption, Implementation, and Maintenance; RE-AIM; child development

Introduction

Early childhood is a critical period for long-term educational, emotional, and economic development [1]. Globally, more than 43% of children younger than 5 years of age are at risk of not reaching their full developmental potential, with children in low- and middle-income countries (LMICs) being disproportionately affected [2]. The failure to achieve developmental potential during early childhood has far-reaching consequences, not only for individual well-being but also for societal progress, which leads to losses in human capital and increased risks of social instability [3]. These challenges underscore the critical importance of investing in effective strategies that support ECD.

Among the most influential predictors of positive ECD outcomes is the quality of parenting [4,5]. Improving parenting practices, especially in under-resourced settings, is therefore a key priority in global ECD efforts. Parenting interventions, which are typically grounded in social learning and attachment theories, aim to equip parents with the knowledge and skills necessary to enhance parenting quality and to foster home environments that promote children's health and development [6-8]. Such programs have been shown to improve ECD outcomes, enhance positive parenting practices, reduce the use of harsh or violent discipline, and improve caregivers' mental health [9,10]. Meta-analytic findings indicate that such interventions in LMICs yield greater improvements in children's cognitive, language, and motor development, highlighting the potential for large-scale benefits in resource-limited settings [10].

Traditional parenting interventions, delivered through home visits or small-group sessions, are resource-intensive and face barriers such as time constraints, geographic distance, and limited personnel [3,11]. These limitations make conventional delivery methods difficult to scale and sustain, particularly in low-resource settings [12,13]. Nowadays, the growing accessibility of the internet and the widespread ownership of mobile phones present a promising opportunity to overcome these challenges by leveraging digital technology [14]. Digital parenting programs provide a cost-effective way to deliver evidence-based parenting support, particularly in low-resource settings, by enhancing scalability while maintaining high fidelity [12,15]. They reduce costs associated with time and travel and increase accessibility to high-quality parenting resources [16,17]. Emerging evidence indicates that digital parenting interventions can positively influence multiple outcomes, including child development, parental confidence, parenting stress, and children's behavioral problems [18-20]. These programs are particularly effective when complemented by human-led support, either in person or remotely [21].

The success of social and behavioral interventions hinges not only on the inclusion of key components designed to facilitate behavior change, but also on the fidelity and quality of their implementation in real-world settings [22]. While the body of research demonstrating the effectiveness of digital parenting interventions in promoting child development is growing, limited attention has been given to their implementation processes. This gap in knowledge poses significant barriers to the replication, adaptation, and scale-up of such interventions, particularly in diverse and resource-constrained contexts [23,24].

Implementation research seeks to uncover what works, how it works, and why it works in everyday settings beyond the confines of controlled trials [25]. It emphasizes identifying contextual factors that influence successful implementation, such as characteristics of the delivery environment, community context, interorganizational partnerships, and intervention delivery mechanisms [26,27]. The RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) framework is a widely recognized model that offers a comprehensive approach for evaluating the implementation and public health impact of interventions [28]. It helps researchers and practitioners identify the strengths and barriers in intervention implementation, thereby facilitating the optimization and broader dissemination of public health practices [28-30]. This approach also offers insights into how short-term programs can be transformed into sustainable, institutionalized practices within organizations and service systems [31]. While existing studies applying the RE-AIM framework have examined digital interventions in various health domains, such as diabetes management, psychotherapy, vaccination, and health behavior promotion [32-35], there is a notable gap in research on the implementation of digital parenting programs, especially across different settings. This gap is especially evident in LMICs, where such interventions remain underexplored and underdeveloped.

Guided by the RE-AIM framework, this study addresses these gaps by evaluating the implementation of a digital parenting program, "Keyushiguang", in 2 distinct service delivery systems in China: an urban educational setting and a rural health care setting. In China, where significant early childhood development gaps and resource disparities persist, there is an urgent need for scalable, context-sensitive parenting support models [36,37]. This study, therefore, examines the program's reach, adoption, implementation, and maintenance to identify shared and setting-specific facilitators and barriers. By comparing these 2 delivery models, the study aims to provide new insights into

XSL•FO
RenderX

the differentiated mechanisms required for scaling and sustaining digital parenting support in diverse, low-resource contexts. The program's impact evaluation will be reported separately.

Methods

Study Design and Context

A multiple-case study approach was used to facilitate a deeper understanding of individual cases while also identifying overarching patterns. A case study is an empirical research method used to investigate contemporary phenomena within their real-world context. By analyzing a specific case in depth, researchers gain insights into the background, processes, and influencing factors, which is especially valuable for studying complex social phenomena or contexts [38]. Case studies can be categorized into single-case and multiple-case studies, depending on the number of cases analyzed. The multiple-case study approach was adopted to enhance the validity of findings from different cases, leading to more robust and reliable conclusions [39]. This method allowed us to identify commonalities and differences in the facilitators and barriers to implementation across different contexts.

Theoretical Underpinnings: RE-AIM

This paper focuses on 4 key dimensions of the RE-AIM framework: Reach, Adoption, Implementation, and Maintenance. Effectiveness was assessed using quantitative methods and will be presented separately.

Reach refers to the total number, proportion, and representativeness of individuals who are willing to participate in a given initiative, intervention, or program [28]. It primarily focuses on understanding the factors that influence individuals' decisions to engage—or not engage—with the program. Adoption refers to the absolute number, proportion, and representativeness of settings and intervention agents (people who deliver the program) who are willing to initiate a program, as well as the reasons behind their decisions [28]. Implementation primarily focused on the program's fidelity, which is the extent to which the program is delivered as intended, and examined how this fidelity was achieved throughout the delivery process [40]. Maintenance at the setting level of a program refers to its institutionalization or integration into routine practices and policies [30].

This paper evaluated each of these dimensions across the 2 distinct settings, analyzing the factors that were perceived to facilitate or hinder participation in the programs, as well as the differences in these factors between the 2 settings.

We also reported a descriptive overview of reach, adoption, and implementation. Reach was calculated by the ratio of participants who consented to participate relative to the total number of eligible individuals in both settings. The recruitment criteria for both programs were similar, with the primary requirement being that the caregivers of children be the primary caregivers. Adoption at the staff level is calculated by dividing the number of staff who agreed to participate in program implementation by the total number of staff eligible to participate in the program within the organization. In addition, we assessed the implementation based on the fidelity checklists of implementers.

Case Selection

The selection of cases in this research was guided by the study objectives. This paper examines the implementation of the digital (chatbot-led) parenting intervention in 2 distinct settings in China: an urban preschool (educational) setting and 2 rural health centers (health care) settings. Intervention delivery in both settings was led by local implementation partners, with support from the research team that developed the digital intervention. Consequently, the intervention's digital delivery method and content remained highly consistent across the 2 settings, with only minor adaptations made for the rural health care setting, such as colloquial expressions used, without altering the core parenting principles.

Case boundaries were defined as follows. The urban educational case referred to the implementation within one preschool in Xinyu city, Jiangxi Province, while the rural health care case referred to the implementation within the village-level public health system, supported by the county-level Maternal and Child Health Department and 2 township-level Health Centers in Huining County, Gansu Province. The analysis covered the full pilot implementation period: the preschool intervention that ran from March 2024 to June 2024, and the health care intervention that ran from October 2024 to January 2025. Stakeholders included in the case analysis were those directly engaged in program implementation—caregivers, teachers, social workers, village doctors, program managers, and local government officers. Broader community members who were not directly involved in program delivery were excluded.

The digital parenting program "Keyushiguang" was culturally adapted from the Parenting for Lifelong Health's ParentText chatbot for parents of children aged 2 to 9 years by a local research team [41]. Delivered via a rule-based chatbot on WeChat, the program covers 8 key parenting topics, each consisting of 3 to 6 modules depending on the child's age. Each module includes an introduction, quizzes, core parenting tips (in video or text format), and home exercises. The chatbot sends one parenting module every 23.5 hours, allowing parents to gradually develop their parenting skills and apply them in daily situations. Designed to integrate smoothly into parents' daily routines, the chatbot enables flexible engagement at convenient times, such as after work, during breaks, or before bedtime. This gradual approach encourages parents to apply the learned skills in real-life scenarios, such as managing child behavior or establishing daily routines. By offering relevant, practical, and easily accessible support, the chatbot helps parents stay engaged consistently, enhancing the program's long-term impact (Detailed program description can be seen in Multimedia Appendix 1).

The main distinctions between the 2 implementations lay in the format of additional human-led parenting support and were driven by the unique implementation contexts and available resources in each setting. In the urban educational setting, chatbot-led digital delivery was supplemented by message-based WeChat group interactions via web, held once or twice per week, and facilitated by headteachers and trained social workers.

XSL•FO
RenderX

In contrast, the rural health care setting incorporated biweekly 40-minute individual home visits conducted by village doctors to reinforce the content delivered online and foster engagement.

The choice between group-based and individual-based human-led parenting support models was based on a comprehensive assessment of the target population and implementation environment [42]. In China, rural areas typically face greater challenges to child development than urban areas [43]. Therefore, the program was designed as a universal intervention (available to all families regardless of child development status) in the urban setting, while in rural areas, it was implemented as a selective intervention (targeted at high-risk families). Additionally, resource feasibility and cultural acceptance also informed program design. For example, most rural families have access to village doctors, whereas urban preschools face teacher shortages. In addition, rural parents are more receptive to home visits [44]. Therefore, we adopted a group-based WeChat interaction model for the urban preschool setting and a combined individual and group home visit model for the rural health care setting to better suit local needs and contexts. The comparison of the 2 pilot implementations is provided in Multimedia Appendix 2.

The rationale for selecting these 2 cases was that by controlling for similarities in program content, the study could minimize the impact of content-related differences and more precisely analyze the result variations arising from differing implementation contexts (ie, urban education vs rural health care settings).

Data Collection

We used data triangulation to enhance the comprehensiveness and validity of our findings. Multiple sources of data were used, including program documentation (such as program manuals and materials), field observation notes (such as the behaviors of stakeholders in the process of program pre-preparation, implementation, and evaluation), semistructured individual interviews, and focus group discussions (FGDs). Our triangulation involved 2 levels. First, we conducted within-case triangulation by integrating these diverse data sources to validate findings within each of the 2 settings. For example, we compared implementers' subjective perceptions of their implementation fidelity with parents' feedback (as reflected in the fidelity checklist) and triangulated these with our field observation records. This approach allowed us to more accurately interpret and evaluate the perspectives expressed by stakeholders. Following this, we conducted between-case triangulation (ie, a cross-case analysis). This second level systematically compared the validated themes from the urban educational setting against those from the rural health care setting to identify the overarching facilitators, barriers, and divergent patterns presented in our results.

Program-related documents retained since the program's inception were collected and analyzed. These included program proposals, curriculum designs, daily implementation records, and training materials, which served as foundational resources for understanding the intervention. All documents were reviewed systematically, and key implementation details were extracted for analysis. In addition, as members of the research team, the authors were actively involved in and directly observed the implementation processes in both settings, providing them with in-depth insights into the contextual dynamics and operational realities of the program delivery.

Field observations related to stakeholders' behaviors, program structure, and implementation processes. We observed the behavior and actions of stakeholders throughout the program and recorded those, which allowed for a more accurate understanding of the viewpoints they expressed.

We adopted a purposive sampling strategy, which is consistent with qualitative research practices that emphasize differentiation (Figure 1). Participants were deliberately selected to reflect diversity in roles, experiences, and contextual backgrounds. This approach allowed us to explore the phenomenon under investigation in a more nuanced and comprehensive manner. Specifically, participants were categorized into 5 key stakeholder groups: program officers, local leadership, local program managers, program implementers, and program participants. The final sample included 1 program officer from the health care setting; 7 local program leaderships (ie, 3 preschool leaderships and 4 government officers from the health care setting); 3 local program managers (ie, 2 Women's and children's health workers and 1 local program staff from the health care setting); 36 program implementers (ie, 9 preschool teachers and 7 social workers from the preschool setting, and 20 village doctors from the health care setting); and 36 program participants (ie, 26 caregivers from the preschool setting; and 10 caregivers from the health care setting).

The figure 1 illustrates the overall structure of data collection. As detailed in the Methods section, the implementation and data collection were conducted sequentially (urban case first, followed by the rural case). FGD: focus group discussion; GO: government officers from the healthcare setting; LPC: local program coordinator; PM: preschool managers; PO: program officer; PP: program participants; PT: preschool teachers; SI: semistructured interview; SW: social worker; VD: village doctor; WACHW: Women's and Children's Health Worker.The interview and FGD guides were developed based on the RE-AIM framework, incorporating a structured set of open-ended questions. Data related to the "Reach" dimension were primarily obtained from program participants, while "Implementation," "Adoption," and "Maintenance" were explored through interviews and focus groups with program donors, local leadership, local program managers, and program implementers. Each FGD lasted approximately one hour, and individual semistructured interviews averaged 40 minutes in duration. Prior to data collection, informed consent was obtained from all participants. With participants' permission, all interviews and FGDs were audio-recorded. Recordings were subsequently transcribed, reviewed for accuracy, and anonymized to ensure confidentiality.

XSL•FO
RenderX

Figure 1. Stakeholder sampling and data collection design for the 2 cases.

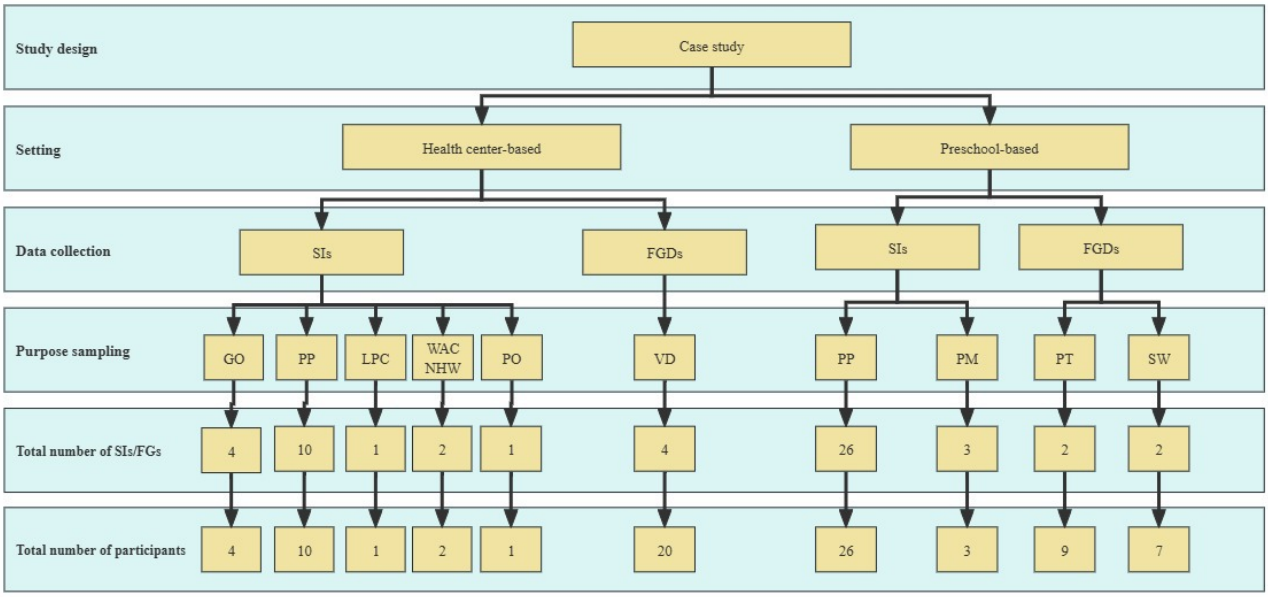

Data Analysis

Data were coded using ATLAS.ti (version 8; Lumivero). A thematic analysis was conducted using a hybrid inductive-deductive approach to identify patterns across the dataset [45]. As data collection was conducted at different times, we first collected and analyzed data from the urban educational setting following a stepwise thematic analysis procedure. After completing data collection in the rural health care setting, we applied the same analytical process. To ensure consistency and mitigate the risk of bias from this sequential design, the same coders applied the same predefined RE-AIM framework to both datasets independently. Only after both independent analyses were complete were the results from the 2 settings formally compared to identify similarities and differences. The coding process was conducted by 2 researchers (XS and RR) and followed the steps outlined by standard thematic analysis [45,46]. Thematic analysis was conducted through a structured 5-step process—data familiarization, initial coding, theme development, collaborative review, and final theme definition—guided by the RE-AIM framework to identify facilitators and barriers to program implementation (Figure 2). During the coding process, we continuously compared new data with existing codes and themes. We considered thematic saturation to be achieved when no new codes or themes emerged from subsequent interviews and FGDs. Given the diversity of stakeholders included—parents, teachers, village doctors, local managers, and government officers—the data provided a comprehensive reflection of perspectives across both settings.

Finally, as the capstone of our cross-case analysis, we synthesized the shared and setting-specific facilitators and barriers identified through the RE-AIM framework. This synthesis allowed us to construct the differentiated implementation models presented as the final component of our Results.

XSL•FO
RenderX

Figure 2. The process of data analysis.

XSL•FO
RenderX

Ethical Considerations

Ethics approvals were obtained from Beijing Normal University (approval numbers: SSDPP-HSC-2024003 and BNU202410100062) and the University of Oxford (SPI_DREC_24_006). We provided participants with a detailed introduction to the program and obtained their informed consent voluntarily. All data is stored in a secure database and is used only for research purposes by the researchers.

Positionality Statement

We acknowledge that our dual roles as program designers and researchers may have influenced data interpretation. To mitigate this, we applied both personal and methodological reflexivity throughout the study.

Our research team has diverse academic backgrounds, including evidence-based social intervention, child development and protection, public administration, education, and social work. These perspectives shaped how we approached the research: team members with social intervention and public administration backgrounds tended to focus on macro-level structures and policies, while those with child development, education, and social work backgrounds were more attuned to participants' lived experiences. To minimize potential biases from these disciplinary differences, the team engaged in ongoing reflexive discussions, considering multiple perspectives at all stages of analysis.

Methodologically, 2 researchers (XS and RR) independently coded different portions of the dataset and cross-checked each other's work. A third researcher (ZF) reviewed the coding, provided feedback, and helped resolve discrepancies through iterative discussion until consensus was reached. This collaborative process ensured rigor and consistency in the analysis.

In addition, practical measures were taken to reduce researcher influence during data collection. Data collectors were trained to use semistructured interview guides with open-ended, nonleading questions, fostering a neutral environment that encouraged participants to share their authentic experiences.

Results

Overview

To ensure methodological rigor and transparency, the reporting of this study followed the iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations (Checklist 1) [47]. Descriptions of interview participants can be seen in Multimedia Appendix 3.

Case Context and Stakeholder Profiles

The Urban Educational Setting: A Preschool in Xinyu City

This pilot was implemented in a preschool located in Xinyu City, an urban, medium-income area in eastern China. The preschool provides early childhood education before primary school, with parenting integrated into daily activities. The main participants were parents, most of whom hold stable jobs; their participation time was influenced by their work schedules. Key implementers included teachers who hold educational credibility and trained social workers who offered professional parenting support and assisted teachers in carrying out the program. The preschool leadership's primary motivation was to enhance the school's educational quality and reputation.

The Rural Health Care Setting: 2 Townships in Gansu Province

This pilot took place in health centers across 2 townships in Huining County, Gansu Province, a low-income rural region in western China. The centers are responsible for public health services, and promoting early childhood development was one of their key goals. The main participants were parents recruited from rural communities, most of whom engage in agricultural work. Their participation was shaped by the agricultural calendar and seasonal migration for work. The key implementers were village doctors who were trusted in their communities with strong local knowledge but generally older and less experienced in parenting. The health care leaderships were primarily motivated by fulfilling national public health directives. We labeled the interview data based on the source and role of the interviewees. The initial letters in each label represent the interviewee's role, as illustrated in Figure 1. HB and SB denote data from the hospital-based program and preschool-based program, respectively. The last 3 digits of each label indicate the document number, and each number corresponds to a unique interviewee or focus group.

Reach—Facilitators and Barriers

Overview

In total, 541 (81.4%) caregivers (402 [74.3%] were mothers, 135 [25%] were fathers, and 2 [0.3%] were grandparents) of the 665 eligible families consented to participate in both the program and the research.

In contrast, the township central health center–based program aimed to recruit 120 parents of children aged 24 - 59 months from 2 pilot townships, out of approximately 795 eligible families. A total of 83 out of the targeted amount of 120 were able to participate (69.2%).

Shared Facilitators to Reach

The main motivations for participation in the program include its strong emphasis on child development, strong relationships between program implementers and caregivers, the perceived credibility of the program developers, and the convenience of learning through a chatbot.

A primary motivation in both settings was a recognized need for practical parenting skills. In the rural health care settings, caregivers expressed that they faced ongoing challenges in daily parenting and felt that their current knowledge was insufficient. They recognized the need to learn practical parenting skills to better support their children. In the urban educational setting, parents expressed similar concerns.

> *We don't know much about how to support our children's growth or deal with their emotional problems. This program gives us a chance to learn parenting ways.* [PP, HBD05]

XSL•FO
RenderX

Another shared facilitator was the influence of implementers, driven by social relationships. Some caregivers in the health care setting reported that their agreement to participate was influenced by their trust and social pressures from program implementers, such as face-saving norms, social obligations, or a sense of personal favor. Similarly, in the educational setting, the motivation to participate also stemmed from the influence of teachers.

> *The village doctor is our family doctor. He has helped us a lot and has a good relationship with us. He wants us to join the program, so we are willing to support him.* [PP, HBD11]

> *At first, the teacher encourages us to participate in this parenting program. I think if other children are participating, we shouldn't seem uncooperative with the teacher's efforts.* [PP, SBD11]

Parents in the health care setting also considered the background of the program developers. Knowing that the program was developed by a research team from a prestigious university in China enhanced its credibility and increased parents' perceived benefits. Similarly, parents in the educational setting shared the same view, trusting the program's credibility due to its association with a reputable institution, which further reinforced their belief in the program's potential benefits.

> *There is too much parenting information online, and it is often hard to tell what is true or false. But this online course is developed by a professional team, so we feel the content is reliable and will be helpful to me.* [PP, HBD10]

The convenience of learning was also an essential factor for program reach. Parents in both educational and health care settings appreciated that the chatbot-based delivery allowed them to access the content anytime and anywhere. The use of diverse formats, such as videos and text, accommodated different learning environments and individual learning preferences.

> *The timing is fine because it is flexible. You can choose when to learn, so it is convenient when you have free time.* [PP, SBD15]

> *The course offers both video and text options, which is suitable for different scenarios. When I have free time, I choose to read the text. When I am busy, I play the video and listen to the content.* [PP, HBD11]

Shared Barriers to Reach

In both settings, shared factors hindering parents' participation include parenting conservatism, a lack of understanding of the program, and limited time for learning.

A shared barrier in both settings was 'parenting conservatism,' where parents preferred to stick to traditional parenting methods. During the program promotion process, some parents in the educational setting with low engagement exhibited a form of 'parenting conservatism'. They preferred to maintain their existing parenting practices and felt it was unnecessary to learn about other approaches, believing that new practices might not be suitable for their own situations. Similarly, in the health care setting, some caregivers also displayed resistance to adopting new parenting knowledge, preferring to rely on their established practices.

> *I feel that every family has its own parenting style. My other children are raised according to my own parenting methods, and there is no need to change it.* [PP, SBD13]

Furthermore, a limited understanding of the program served as a barrier in both contexts. Some caregivers in the educational setting indicated that they did not fully understand the program, such as its content, learning format, and flexibility. Some caregivers claimed they had no spare time to engage with the digital program and chose not to participate, unaware that the learning sessions were brief and could be easily completed anywhere in just a few minutes.

> *I think the course will take up a lot of time. Since I have to work and take care of my family, I don't have much spare time for studying.* [PP, SBD13]

While some parents in the health care setting remained unaware of what they could gain from the program, this ultimately reduced their enthusiasm for participating.

> *At first, they don't tell me the specific details of the program. I don't know its benefits, so I initially see it as just a favor to village doctors.* [PP, HBD13]

Distinct Facilitators and Barriers to Reach

The urban educational setting presented several unique facilitators. These included urban parents' recognition of the parenting values conveyed by the program and their acknowledgment of the implementers' credibility and rationale in delivering parenting knowledge. These factors played a significant role in enhancing parental engagement in the educational setting. However, such facilitators were not observed in the health center-based program.

Some parents in the urban areas reported that the parenting concepts and attitudes conveyed through the program resonated with their own parenting beliefs. In contrast, parents in rural areas rarely expressed this kind of value-level recognition.

> *I agree with the parenting concepts in the course, and I am willing to learn it.* [PP, SBD08]

The perceived professional authority of the program implementers influenced caregivers' decisions to participate. Parents viewed preschools as appropriate platforms for parenting support, and headteachers were seen as credible and persuasive figures in promoting parenting knowledge.

> *This preschool and its teachers are the best in the area, and I trust the quality of both the preschool and the teachers. I believe their actions will be beneficial to me.* [PP, SBD18]

In contrast, the rural setting faced a unique barrier related to the implementers' professional role. In rural areas where village doctors delivered the parenting program for the first time, some doctors noted that because their routine work is not typically associated with parenting, parents were skeptical about their true intentions behind offering such services.

XSL•FO
RenderX

> *The parents doubt why I invite them to join the parenting program. They ask if I have received any benefits from the program organizers.* [VD, HBD25]

The summary of barriers and facilitators to reach can be seen in Multimedia Appendix 4.

Adoption—Facilitators and Barriers

Overview

At the setting level, since the program was conducted as a pilot, 1 urban preschool and 2 rural health centers agreed to participate upon invitation. At the staff level, in the preschool-based program, the primary implementers included headteachers and recruited social workers, all of whom joined the program in response to directives from their supervisors. In the health center-based program, the main implementers were 2 women's and children's health workers and 30 village doctors. Eventually, both maternal and child health workers participated, while 23 village doctors (76.7%) remained engaged, and 7 withdrew from the program. Overall, the adoption level of the program was relatively high.

Shared Facilitators for Adoption

At the setting level, relevant information was gathered through interviews with local leadership. The main shared factors that promoted adoption included trust in the program developer, alignment with organizational functions, and empowerment of the organization and staff.

A primary factor in facilitating successful cooperation and adoption in both settings was trust in the program developers. Leadership in the educational setting acknowledged that the developers' affiliation with a prestigious university in China enhanced their perception of the program's credibility and value. In the health care setting, management also demonstrated trust in the collaborating research team.

> *You are a research team from a top university in China, and the quality of what you developed is guaranteed. We also have confidence that parents will accept the program, and we believe it will have a positive impact.* [PM, SBD33]

Another shared facilitator was the importance of the alignment between the program and the implementing organization's core functions. In the educational setting, leaders reported that before adopting the program, they would assess whether their organization possessed the capacity and resources to support implementing the program and whether the initiative aligned with their institutional mandate. For instance, preschools regarded parenting services as a fundamental aspect of their work, viewing it as consistent with their educational responsibilities.

> *This is something I want to do. In 2010, our preschool established an 'Early Education Center,' and parenting is a part of our work.* [PM, SBD32]

Likewise, in the rural setting, health centers considered parenting support closely related to child health care and broader public health objectives.

> *This online-based parenting program focuses on promoting health education, especially concerning children's growth and development. This is also connected to our daily work in public health.* [GO, HBD32]

Local leadership in both settings also highlighted its potential to empower staff and foster organizational growth. In the educational setting, managers noted that the program could enhance headteachers' professional skills in parenting, positioning it as a valuable resource for supporting parents.

> *Our teachers gain many benefits during the implementation of the program, such as learning about family education and strengthening their connection with parents.* [PM, SBD30]

In the health care setting, leaderships emphasized that the program could strengthen the capabilities of village doctors as well as maternal and women's and children's health workers in delivering comprehensive children's healthcare services.

> *Our county pays great attention to the implementation of early childhood development work in rural areas. Involving village doctors and other staff in this parenting program allows them to gain experience in carrying out such work.* [GO, HBD27]

At the staff level, both task-driven (motivation that arises from external obligations to complete tasks assigned by others, such as supervisors or program coordinators) and intrinsic motivation (motivation from personal interest or intrinsic desire) were key shared factors promoting adoption. For example, in the health care settings, implementers reported that the program was perceived as a mandated task from higher-level leadership, who required them to carry out the program and follow the specified timeline and procedures.

> *Our leaders attach great importance to this program and repeatedly emphasize the need to cooperate with the program developers to complete it. We feel that this is a very important task for us.* [VD, HBD24]

Intrinsic motivation also played an important role. Some implementers in the health care setting expressed that they recognized the significance of this pilot and the potential benefits it could bring to caregivers. This sense of mission motivated them to actively adopt and support the program. In the educational setting, teachers also felt that the program could support the development of both parents and children, contributing to better educational outcomes.

> *Another reason I participate in this program is that I feel what we are doing is meaningful. In our rural area, such resources are very precious, and many parents also have a demand for parenting knowledge. By completing this program, we can provide help to children and their families.* [VD, HBD31]

Shared Barriers to Adoption

A major shared factor hindering adoption in both settings was the perceived difficulty of implementation due to limited time. Both implementers and local leadership expressed concerns that implementers were already occupied with their regular

XSL•FO
RenderX

responsibilities, which consumed most of their time. The entire program's especially tight schedule further compounded the issue, as they were expected to complete tasks within a short timeframe. This increased the pressure on implementers, making them feel even more stressed.

> *Our village doctors have a lot of work and are very busy. We have to both prescribe medical treatment and complete public health tasks such as chronic disease management and vaccination. The program requires us to complete the tasks within a short time, and it feels very stressful.* [VD, HBD25]

> *We also have a lot of teaching tasks at school to complete, and the time is not very sufficient.* [PT, SBD34]

Distinct Facilitators and Barriers to Adoption

At the staff and setting level, prior successful collaboration experiences served as a unique facilitator for the health centers' adoption of the program. However, health center-based implementers often lacked familiarity with parenting knowledge, and some faced difficulties in recruiting appropriate participants. The perceived difficulty in recruitment hindered program adoption among village doctors. Moreover, remuneration emerged as a significant motivating factor for implementers in the health care setting. In the educational setting, the teachers' professional roles in parenting acted as a key facilitator.

At the institution level, factors promoting adoption differed between the 2 contexts. In the health center-based setting, health care leaders and program officers highlighted that prior successful collaboration experience played a key role. The local government, supported by the donor, had previously implemented similar ECD parenting programs in other townships within the county, which helped establish mutual trust and efficient working mechanisms.

> *They agree because we have previously implemented an early childhood development program here. We are familiar with the local leaders and institutions, and have built a strong trust relationship.* [PD, HBD35]

At the staff level, the relevance of the task to the implementers' professional roles directly influenced their motivation to adopt the program. In the urban preschool setting, headteachers reported that parenting support was closely tied to their core duties, contributing to child development and the enhancement of educational quality. This strong connection fostered greater intrinsic motivation among them.

> *We also provide some parenting education in daily work. I review this program and feel it is quite good. If parents attend, it will benefit them.* [PT, SBD34]

In contrast, some village doctors and maternal and child health workers in the health care setting felt their routine work was less directly related to parenting. Consequently, they viewed the program more as a task assigned by higher-level leadership, with task-driven motivation playing a prominent role.

> *We have not done similar work before. I believe it is not our duty, as we are mainly responsible for public health. The main reason for implementing this program is that it is assigned by our leadership.* [VD, HBD24]

Some village doctors also reported difficulty in identifying parents who met the program's participation criteria and were able to join. This challenge led to frustration and, over time, caused some of them to gradually withdraw from participating in the program altogether.

> *The main challenge for me is that I really cannot find suitable parents, because many children in the village are migrant or left-behind children.* [VD, HBD30]

Additionally, village doctors noted that the remuneration provided by the program officers was an important incentive for them to complete their responsibilities.

> *We carry out the task as required and receive some compensation. If it is voluntary, I will not be willing to complete the task.* [VD, HBD24]

The summary of barriers and facilitators to adoption can be seen in Multimedia Appendix 5.

Implementation—Facilitators and Barriers

Overview

The investigation of implementation primarily focused on the program's fidelity, which is the extent to which the program is delivered as intended, and examined how this fidelity was achieved throughout the delivery process [40].

The program was delivered with high fidelity in both settings, as demonstrated by fidelity checklists. In the preschool setting, headteachers recruited parents and reminded them to complete the online modules, while 10 social workers facilitated group discussions and completed all assigned tasks. In the health care setting, village doctors conducted home visits, supported by maternal and child health workers. Follow-up phone calls with 74 parents confirmed the content of the visits. The village doctors achieved an average fidelity score of 7.81 out of 8, indicating strong adherence to the program protocol.

Shared Facilitators for Implementation

Several key shared factors were identified as promoting the fidelity of program implementation: adequate onboarding training, a supportive management system, timely external support, and clear work guidelines.

Adequate onboarding training was a shared facilitator in both settings. Implementers in the health care setting believed that the training helped them understand the parenting principles, tasks, potential challenges, and proposed solutions. They also noted that the time allocated for discussion created opportunities to express concerns, ask questions, and receive prompt responses. Similarly, in the educational setting, teachers and social workers affirmed that the training enabled them to understand their respective tasks and effectively contribute to the program's implementation.

> *The training before the program starts is essential. It helps us understand what we need to do, and the*

XSL•FO
RenderX

> *subsequent discussion sessions also address some issues.* [VD, HBD24]

A supportive management system involved a balanced combination of appropriate incentives and supervision to facilitate the successful implementation of the intervention. In both health care and educational settings, local leadership perceived that, to enhance fidelity, implementers required suitable incentives that align with their expectations, including both material and intrinsic (spiritual) rewards.

> *After all, the main work of village doctors is not parenting. They have limited time and energy, so they need to be compensated.* [GO, HBD27]

> *Our teachers need to have a sense of achievement when doing this program. Even giving them a certificate is an effective encouragement for them.* [PM, SBD32]

Meanwhile, while supervision was crucial in both contexts, its focus differed. Local program managers in the health care setting emphasized that supervision was essential to achieve a balance between quality control and task implementation efficiency. They believed that supervision should cause necessary work pressure while allowing sufficient flexibility in how tasks are carried out. In the educational setting, the focus of supervision was more on ensuring timely task completion, with managers highlighting the importance of regularly reminding teachers to meet deadlines.

> *Village doctors work hard. Sometimes, tasks cannot be completed due to external reasons. When supervising their work, we have to both consider the quality and avoid discouraging their motivation.* [WACNHW, HBD26]

Providing timely external support to implementers was perceived to be a critical shared factor for successful implementation, which involves access to expert advice, consultations, and problem-solving assistance from external consultants or organizations when issues arise during program implementation. Implementers in the health care setting reported that timely and adequate support—whether material, psychological, or resource-based—from organizations and program developers helped alleviate their stress, overcome challenges, and enabled them to successfully complete their tasks. Likewise, in the educational setting, implementers affirmed that support from program developers and managers was crucial for the smooth execution of their work.

> *I feel that the support is quite sufficient. You are always there whether we or the village doctors have any questions. I feel touched.* [WACNHW, HBD26]

Finally, having clear work guidelines was a key shared facilitating factor. Implementers in both educational and preschool settings noted that well-defined procedures contributed to smoother implementation by clearly outlining what needed to be done, when, and how. This clarity made it easier for them to carry out their tasks efficiently.

> *The clear work guidelines let me know what I need to do, making it easier to carry out my work.* [VD, HBD31]

> *This program provides clear instructions, and the guidance in all aspects is very detailed. So, we can implement it well according to the guidance.* [SW, SBD27]

Shared Barriers to Implementation

A lack of sense of purpose and psychological pressure was perceived as key shared challenges faced by some implementers in both settings. Psychological pressures refer to stressors such as heavy workloads, tight deadlines, and the emotional burden of ensuring the success of the program. These pressures can negatively impact the quality of program delivery and the well-being of implementers. While leadership demonstrated a clear understanding of the program's goals, some implementers lacked this clarity. During the implementation process, it was observed that certain implementers were unsure about the program's objectives and the value it could provide to caregivers. This uncertainty at times led to self-doubt, which had a negative impact on their motivation. Further communication with implementers confirmed this observation—several social workers from the educational setting and village doctors from the health care setting expressed confusion about the purpose of their roles and the potential benefits their efforts could bring to families.

> *When conducting home visits, we do not know our purpose. What exactly are we trying to do?* [VD, HBD24]

> *You do not know your role in this process, but you must do it. Moreover, at this point, you can feel quite confused.* [SW, SBD28]

Implementers in both settings shared the challenge of generating demand and encouraging active participation in the program, which caused psychological pressure among implementers.

> *Because not everyone is enthusiastic about participating, I think the appeal of the content to parents is crucial. Is the content we are providing useful to parents? Is it something they want?* [SW, SBD27]

Distinct Facilitators and Barriers to Implementation

A unique facilitator in the health care setting was the timing of implementation, which aligned well with the villagers' work schedules. However, in China, village doctors typically only conduct home visits for specific health care tasks. Also, in the health care setting, parenting home visits are irregular, and the long distances they needed to travel posed a challenge during program delivery. Furthermore, while clarity in work procedures was a shared facilitating factor across both settings, there was a notable difference in the need for implementation flexibility across the 2 settings.

First, in the rural setting, village doctors shared that the program was implemented during the off-season for farming, just before the Chinese New Year, a period when there was less agricultural work and many parents who normally worked away from home had returned. This created free time for caregivers to participate in the program.

XSL•FO
RenderX

> *Parents are generally at home in winter. As the weather is cold, there is no farming to do. Besides, they do not go out for labor during this season, so most of them stay home.* [VD, SBD24]

Second, village doctors also reported that the physical distance between households and their base locations created challenges in conducting home visits on time.

> *It is still too far, the road is inconvenient, and it is not easy to return.* [VD, D30]

Third, standardized workflow served as a double-edged sword, with its impact varying across settings. In the educational setting, especially among headteachers, implementers preferred greater flexibility. They expressed a desire to customize token economy systems to encourage daily participation in their own class and requested more adaptable web-based group interaction structures to better address the diverse needs of caregivers.

> *I feel that our online group discussions are a bit stiff. The content shared is fixed and does not capture parents' interest. If we integrate additional parenting support with our daily activities, it will be better.* [PT, SBD35]

In contrast, implementers in the rural health care setting, particularly village doctors who were typically older and had limited experience with parenting support, preferred a highly standardized process. They valued having clear guidance on what tasks to perform, when to perform them, and how to carry them out.

> *The workflow you provided is very clear. It outlines each step, and we just need to follow it to complete the task.* [WACNHW, HBD26]

The summary of barriers and facilitators to implementation can be seen in Multimedia Appendix 6.

Maintenance at the Setting Level—Facilitators and Barriers

Overview

This article primarily explores the factors that may facilitate or hinder the integration of the program into the established service system. In both settings, the leadership actively worked towards promoting the program's integration into the existing service infrastructure.

Shared Facilitators to Maintenance at a Setting Level

The shared factors were the integration of the program into an organization's daily operations, alignment between the program content and the organization's core functions, the availability of sufficient and appropriate internal human resources, and the low cost associated with digital delivery.

First, alignment with the organization's core mission was a key shared facilitator. Local leadership in the health care setting emphasized that the program's alignment with their existing responsibilities was essential for successful institutionalization beyond initial adoption. By supporting the organization's core mission, the program minimized additional costs and streamlined integration into routine work. Likewise, in the educational setting, managers emphasized that the parenting program was a natural fit for their institution's existing work and goals.

> *Normalizing it means integrating it with our public health services, which does not create any work pressure for us or add any burden. It should also benefit our daily work.* [GO, HBD27]

Second, leadership in both settings believed they had access to a well-suited workforce for program delivery. Village doctors in the rural setting and headteachers in the urban setting were described as stable, professionally trained in child development, and capable of ongoing learning. Their regular contact with families and their established trust and authority within the community made them particularly suitable for engaging caregivers and carrying out the program.

> *The village doctor team can handle this program. On one hand, they have good relationships with the villagers. On the other hand, they possess a lot of medical knowledge and have the ability to learn.* [GO, HBD32]

> *Our teachers are well-suited to take on this program. They already handle educational work and can easily use it as a tool in their daily teaching.* [PM, SBD33]

Third, local leadership in the educational setting emphasized that the low cost of digital delivery was a key factor facilitating program integration. By delivering core content via web through a WeChat-based chatbot, the program minimized the need for intensive offline operations. This digital approach not only helped maintain content quality but also significantly reduced the workload and operational costs for implementing organizations, making the program more sustainable and scalable within existing structures. Similarly, in the health care setting, managers highlighted that controllable costs were crucial for sustaining the program in rural areas.

> *I think the advantage of the online format is that, firstly, it can push information in real-time and is more flexible. At the same time, since we are in the information technology era, parents also use their phones often.* [PM, SBD27]

Shared Barriers to Maintenance at the Setting Level

Shared barriers to integrating the program into the organization's daily operations included institutional dependence on higher government authorization, challenges in sustaining staff motivation, and difficulties in generating parental demand for parenting support.

A primary barrier in both settings was the institution's reliance on higher-level government approval. In the rural context, health care leadership noted that integration into routine services would require authorization from senior government officials overseeing health affairs.

> *To integrate the program into our daily work, we need approval from the senior leaders in charge of this work at the higher level.* [GO, HBD33]

Similarly, in the urban context, preschool managers indicated that continued implementation in preschools would depend on approval and support from the education department.

XSL•FO
RenderX

> *To integrate the program into the preschool's daily work, we need support from the local education bureau leaders. It would be best to also gain official support from your university.* [PM, SBD32]

In addition, leaders in both settings expressed concerns about sustaining staff motivation over the long term. While staff had fulfilled their responsibilities during the pilot phase, there was uncertainty about how to provide sufficient and appropriate incentives to support ongoing, routine implementation once the program became part of daily work. This concern was echoed in the educational setting, where managers also stressed the need for incentives to maintain teachers' continued engagement.

> *If the work is unpaid and voluntary labor, as mentioned by the village doctor, it is a challenge for the village doctor.* [GO, HBD27]

Another significant barrier was the lack of perceived demand from parents. Government officers and organizational managers from both settings expressed concern that many parents did not recognize the value or necessity of parenting support. Without sufficient parental buy-in, they feared that efforts to integrate the program into routine services would result in low participation rates and limited impact.

> *After all, we are in the countryside. Some parents have a higher level of awareness and think that interacting with their children is meaningful. However, some parents feel it is unnecessary; if the children can play alone and do not cause trouble, they have done their job.* [WACNHW, HBD26]

> *Many parents, when it comes to parenting, understand but not to a great extent, which causes a disconnect. They feel they somewhat understand but not fully, and accepting some new concepts is difficult.* [PM, SBD30]

Distinct Facilitators and Barriers to Maintenance at the Setting Level

A unique facilitating factor in the health center-based delivery was the close relationship between village doctors and local families, which enhanced trust and communication. However, implementation in this setting also faced a unique challenge in securing resources for sustainable material incentives. Both the health care and educational settings shared the need for further localized contextual adaptation of the program. Nonetheless, the focus of these recommended adaptations differed between the 2 settings.

In the rural health care setting, leaderships perceived that parenting programs facilitated by village doctors were feasible for sustained implementation in rural health care settings. They believed that village doctors, who often had strong, long-standing relationships with families—sometimes dating back to prenatal care—were uniquely positioned to gain family acceptance of the program and successfully conduct home visits. In contrast, the preschool program relied on teachers leading group-based remote support and did not require the same depth of ongoing personal relationships with students' families as village doctors had.

> *In our rural areas, village doctors are like family doctors, closely connected to the families. People know each other well, and home visits are easy. In the city, living in apartment buildings makes it less convenient, and people are more guarded.* [GO, HBD32]

In addition, the approach to sustaining staff motivation differed between the 2 settings. In the educational setting, preschool managers noted that parenting work is closely aligned with headteachers' core educational duties. As a result, integrating the program into daily operations was more natural and did not significantly increase their workload. Therefore, spiritual incentives, such as recognition and a sense of professional fulfillment, were seen as important for promoting long-term motivation.

> *Parenting is already part of our teachers' work, so it does not add extra burden. However, some motivation is still needed, such as certificates or awards.* [PM, SBD33]

In contrast, in the health care setting, leadership emphasized that village doctors primarily focused on public health responsibilities. Parenting-related tasks were viewed as additional duties, which increased their overall workload. In this context, material incentives were considered more effective, as they better compensated for the extra effort required for sustained engagement. By comparison, for preschool teachers, parenting was closely integrated with their daily work. The program did not add to their workload but rather complemented it, serving as an educational tool to support their parenting-related tasks.

> *Our village doctors' main work is in public health. This task adds extra duties for them, so compensation is needed.* [GO, HBD28]

> *I don't think this is a burden for the teachers, as they are already engaged in education. Parenting is also familiar to them. They can adapt the program content into teaching tools based on their needs.* [PM, SBD30]

Furthermore, adaptations to each specific implementation context were perceived as important. In the preschool, leaderships suggested that for the program to be effectively integrated into daily work, the digital learning format, duration, and additional parenting support needed to be further adapted to better align with teachers' schedules and classroom activities. This type of adaptation primarily involved organizational alignment, ensuring the program fit seamlessly within existing school routines.

> *To become part of our daily work, the program needs to align with the kindergarten's teaching needs and integrate with our activities. Teachers have to be the leaders.* [PM, SBD30]

In the health care setting, health care managers believed that for long-term implementation, the program needed to be adjusted to include special groups, such as left-behind children and children cared for by grandparents, who make up the majority in the local area, and also need parenting support. However,

XSL•FO
RenderX

this time, the adaptation focused more on tailoring the program to parents who stayed at home to care for their children, as a result of careful consideration of available implementation resources and the exploratory nature of the pilot as an initial test of digital interventions in rural areas.

> *Like in our area, most young parents go out to work, leaving grandparents behind. Therefore, these families need more services.* [GO, HBD33]

The summary of barriers and facilitators to maintenance at the setting level can be seen in Multimedia Appendix 7.

Differentiated Models of Digital Parenting Program Implementation in Two Sectors

Based on the results, we summarize two models for implementing and institutionalizing the digital parenting program in rural health care and urban educational settings (Figure 3). The basic sequence starts with promoting organizational adoption of the program, followed by encouraging staff adoption, then expanding reach. After that, the focus shifts to implementation and finally to maintenance. In impoverished rural areas, digital interventions need to be complemented by individualized human-led offline support to enhance the impact. Accordingly, we propose a model that involves collaboration with the health center and leverages village doctors to deliver the digital program alongside home visits. For urban areas, we suggest a model where digital components are combined with online parent groups led by teachers.

During the organizational adoption stage, building trust with managers, emphasizing the alignment between the program and the organization, and highlighting the program's benefits are crucial. Trust can be fostered by demonstrating the evidence base and credibility of the program, which helps managers recognize the program's potential. Emphasizing how the program aligns with the organization's goals encourages adoption of the program and fosters positive expectations regarding its completion.

Figure 3. Differentiated models of delivery in urban preschool and rural health care settings.

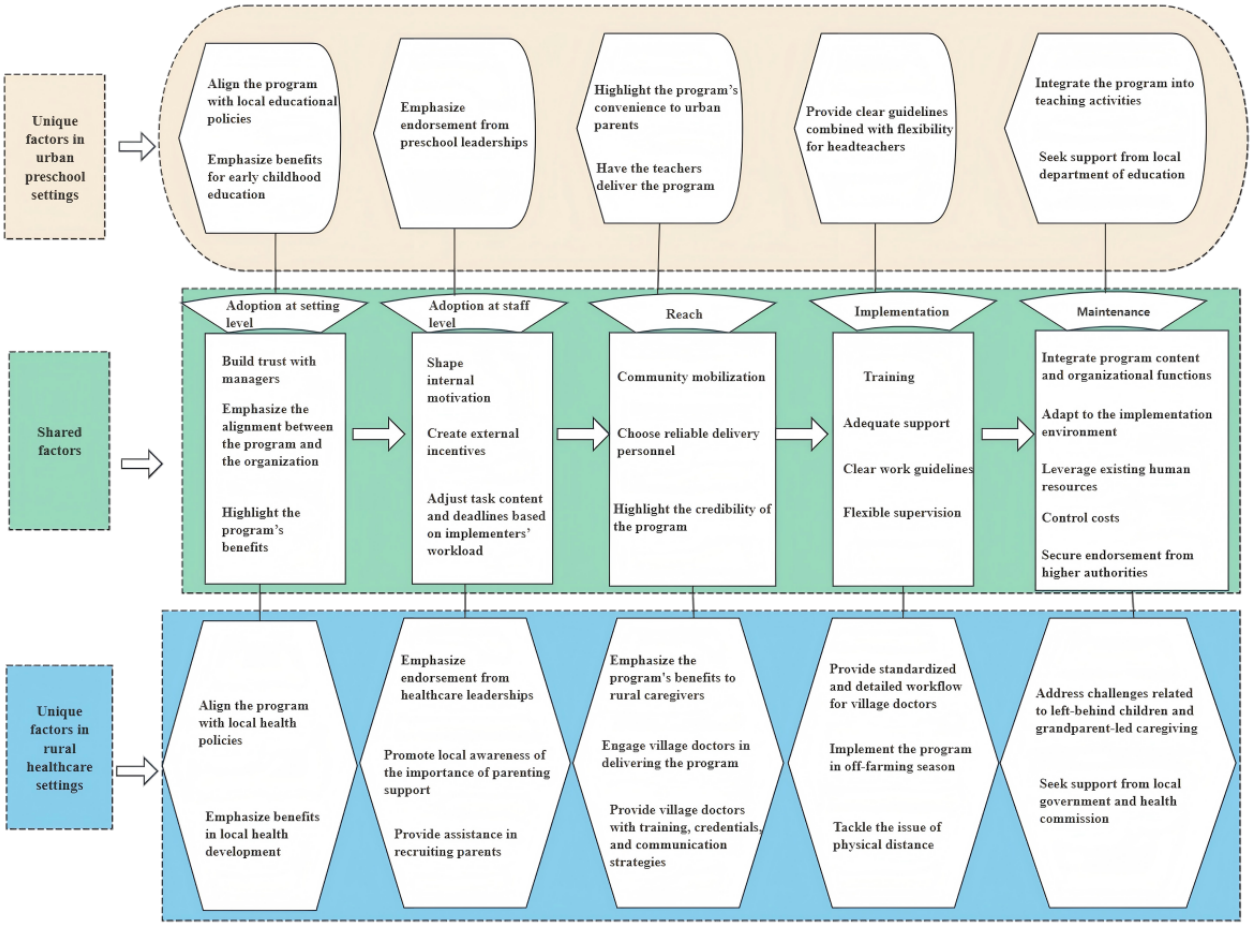

In rural health care settings, aligning the program with local health policies can strengthen its acceptance, particularly by emphasizing benefits such as improving early childhood development and addressing population health challenges. In contrast, urban educational settings can benefit from closer ties to educational policies, where the program's role in enhancing teaching quality, promoting parent-teacher collaboration, and reducing behavioral problems in preschool environments is more prominent. Thus, the focus in rural areas is more health-centered, while urban settings prioritize educational outcomes.

At the staff-level adoption stage, it is essential to shape internal motivation, create external incentives, and adjust task content and deadlines based on implementers' workload. First,

XSL•FO
RenderX

effectively communicating the program's significance helps foster internal motivation and enables implementers to recognize its benefits. Preschool teachers typically have stronger internal motivation because their daily work is directly related to parenting, allowing them to see immediate benefits. In contrast, village doctors in health centers, whose routine duties are less connected to parenting, require communication and training to understand how parenting relates to their role and to enhance their perception of the program's importance.

Second, when internal motivation alone is insufficient, forming external incentives becomes essential. Depending on the program's context, it is necessary to strike a balance between material and intrinsic incentives to fully motivate implementers. For example, in the rural pilot, providing village doctors with reasonable compensation effectively promoted their active participation in program implementation.

Regarding recruitment, urban preschools generally experience less difficulty, as parents tend to cooperate more readily with teachers. However, in rural areas, recruiting parents is more challenging due to complex family structures and village doctors' limited experience with parenting work. This necessitates additional support from implementers and institutional leaders.

In the reach stage, community mobilization, choosing reliable delivery personnel, and highlighting the credibility of the program content are essential. Through community mobilization, parents can understand the program's goals, content, and requirements, and most importantly, recognize the benefits of the program and that the learning process is convenient and low in time cost. In addition, it is worthwhile to highlight the program's convenience to urban parents and emphasize its benefits to rural parents. During the promotion, it is also essential to highlight the program's being systematic, distinguishing it from fragmented parenting information online. In urban preschools, teachers can serve as program implementers due to their close relationships with parents. In rural health care settings, village doctors can take on this role, as they are generally trusted by families. However, since parenting is not a routine part of village doctors' responsibilities, providing them with adequate training, relevant credentials, and communication strategies is necessary to reduce parental skepticism about the program.

In the implementation stage, training, adequate support, clear work guidelines, and flexible supervision are essential for ensuring fidelity. However, the specificity of work guidelines should be tailored to the implementers' roles. In rural health care settings, village doctors, who are generally unfamiliar with parenting work, require standardized and highly detailed workflows to guide their delivery. Conversely, urban preschool teachers, already experienced with parenting-related tasks, benefit from clear guidelines combined with flexibility to facilitate group discussions during parent meetings or school activities, supported by suggested discussion topics. Additionally, in rural areas, particular consideration must be given to the timing and logistics of home visits, as busy farming seasons often limit the availability of both village doctors and families for such visits.

In the maintenance stage, several factors are critical to ensuring the program's sustainability: integrating program content with organizational functions, adapting to the implementation environment, leveraging existing human resources, controlling costs, and securing endorsement from higher authorities. Controlling costs is paramount, as high expenses threaten the long-term sustainability of the program.

Various strategies can be used to reduce implementation costs. Similar to adoption, alignment between organizational functions and program content is important, but in the maintenance stage, this alignment emphasizes better integration of the program with the organization's daily work content and clarifying the position of the program task in the organization.

Using existing human resources by providing adequate onboarding and ongoing training can make the best use of the organization's current staff rather than hiring new personnel. Furthermore, it is necessary to obtain endorsement from higher authorities, with preschools needing support from the education department and health centers from public health authorities.

Adaptations must reflect the distinct realities of urban educational and rural health care settings. In urban preschools, the program needs to be integrated into teaching activities, becoming a practical tool within educators' daily routines. Conversely, rural areas must also address challenges related to left-behind children and grandparent-led caregiving to ensure these families are included in the services.

Discussion

Principal Findings

This study aimed to identify multilevel factors influencing the implementation of a digital parenting program across 2 distinct settings. The findings addressed this aim by revealing both shared and setting-specific facilitators and barriers across the RE-AIM dimensions, providing a comprehensive understanding of how digital parenting programs may be effectively implemented and sustained in different institutional contexts.

While well-known digital parenting programs such as Triple P Online and Incredible Years Online have shown positive outcomes in several countries [48-50], this evidence is predominantly from high-income countries, with limited exploration of their adaptability and sustainability across diverse implementation contexts. Our study directly addresses this gap by examining the implementation of a digital parenting program in 2 distinct settings in China—an urban educational setting and a rural health care setting—thereby contributing crucial evidence from a large low- and middle-income country.

Parents in both urban and rural areas expressed a desire for parenting knowledge, which was motivated by their wish to support their children in growing up healthy. Although the internet provided access to information, much of it was fragmented and lacked a systematic approach. As a result, parents often struggled to build a coherent knowledge system or apply these concepts in practice. This finding is consistent with previous research [51]. In contrast, the digital parenting

XSL•FO
RenderX

program offers culturally adapted, evidence-based content that is systematic, practical, and easy to access.

Trust is indispensable to a program's success, as previous study shows [52]. It must be built among organization managers, implementers, and the beneficiaries. First, building trust with managers and implementers to get their support is important [53]. It can be fostered by showcasing the program's evidence base, emphasizing its benefits, and selecting those with prior cooperation experience. Second, building trust with beneficiaries requires the credibility of the delivery channel and the implementer. A trusted implementer can boost parents' confidence in the program and provide extra motivations, such as social consideration (eg, face-saving or relationship building). In rural areas, village doctors, trusted as family doctors, are well-suited for this role [54], while in urban areas, preschool teachers hold similar credibility and respect.

Similar to previous research [55], we found that community mobilization for both parents and implementers before implementation is important for generating demand, thereby increasing the adoption and program reach. Parents should be introduced to the importance of evidence-based parenting, the benefits of the program, and how it works. At the same time, implementers need to understand the connection between the program and their routine work, its significance for them, and the necessity of the human-led components. Additionally, fostering a shared vision with all stakeholders is essential.

Providing support for implementers is also crucial, which aligns with previous research [56]. It is important to provide adequate training to equip them with implementation skills and clearly define their responsibilities. Establishing communication channels related to parenting support within the organization, such as regular meetings to discuss issues and digital systems to monitor progress. This will facilitate efficient vertical communication, supervision, and problem-solving mechanisms between implementers and organization managers.

Appropriate incentives for implementers are essential. Previous research emphasizes their importance, but it often lacks guidance on how to provide them [57]. We found that material and psychological incentives should be based on the workload and the fit between the implementer's tasks and capabilities. When human-led responsibilities exceed usual duties, greater incentives may be required to ensure motivation and sustained engagement.

Building on these findings, several practical and theoretical implications emerge for strengthening digital parenting interventions across diverse contexts. Program adaptation is necessary for maintenance, which aligns with previous research [53]. Adaptation occurs at 2 levels: content and context [58]. Content adaptation involves adding information relevant to the organization's functions and the local population's needs, such as child nutrition and disease prevention in rural health care, and early education in urban preschools. Context adaptation involves adjusting delivery methods to fit the organization's characteristics, ensuring that the program becomes integrated into regular work [56]. In rural areas, additional focus on left-behind children and grandparent caregivers is necessary because of their population structure [59]. While urban settings should incorporate the program into daily school activities, such as parent meetings and lectures.

Implications for Policy, Practice, and Research

To better support families in both urban and rural areas, governments should invest in providing trustworthy and accessible parenting resources. Trusted institutions, such as preschools and health centers, can serve as effective channels for delivering this information. With government support and institutional credibility, families may be more likely to engage in the programs.

When implementing digital parenting programs, it is essential to tailor strategies to local contexts, considering organizational mandates, local culture, demographic profiles, and economic conditions. Integrating the program into existing institutional workflows can reduce costs and promote sustainability. Establishing trust with organizational leaders, frontline implementers, and the beneficiaries is also key and should be supported by early and continuous community mobilization and engagement. Importantly, implementation must strike a balance between fidelity to core content and flexibility in delivery. Aligning program delivery with the institutional capabilities and work patterns, while providing clear and structured workflows, can help maintain quality and adapt to implementers' needs on the ground.

Future research should explore a broader range of settings and more diverse implementation strategies. Additionally, this study primarily used the RE-AIM framework to evaluate implementation at the mezzo- and macro-levels. However, parents, as direct beneficiaries, also play a crucial role. Their acceptance and participation significantly influence program success. Therefore, future research should adopt a micro-perspective focusing on parents' perceptions, attitudes, acceptance, and engagement. Such insights would inform future implementation of digital parenting programs.

Conclusions

This study's findings go beyond demonstrating the feasibility of digital parenting programs in low-resource settings to highlight a key lesson for their institutionalization: successful scaling is not about a single "one-size-fits-all" digital solution. Effective implementation requires hybrid models that strategically combine low-cost technology with trusted local human infrastructure, such as teachers in urban schools or village doctors in rural clinics.

Our comparison of urban and rural settings shows that the human-led component must be carefully tailored. Urban environments can benefit from flexible online group support, whereas rural contexts often require the structure and accountability of in-person visits. These insights provide a roadmap for policymakers and practitioners to move beyond standardized rollouts and develop a flexible "implementation playbook." By prioritizing adaptation to local social and organizational contexts, evidence-based digital parenting interventions can bridge the gap from efficacy to sustainable, equitable, and real-world impact.

XSL•FO
RenderX

Limitations and Strengths

First, this study did not report on the program's effectiveness. This may raise concerns about its impact. We plan to address this in a forthcoming publication specifically focusing on impact evaluation. Second, this study did not delve into micro-level aspects, especially parents' adoption of and attitudes toward the program. While this is an important area, it could not be adequately covered due to space limitations and will be examined in future research. Third, this study was limited to 2 pilot sites. Including more case studies in future research will help enrich the findings and enhance external validity.

The strength of this study lies in its focus on the implementation process of digital parenting programs, providing valuable insights for future program implementation. We also explored different implementation models across urban and rural education and health care sectors. This can support potential scale-up in diverse contexts. Additionally, we discussed the use of human-led models in different practical settings. This can provide guidance for choosing appropriate approaches in future hybrid (digital and human-led) parenting programs.

Acknowledgments

We extend our sincere gratitude to Xuechen Zhang, Ruinan Zhou, Taoran Li, Xinran Liu, Qingyang Zhang, Laurie Markle, Chiara Facciolà, Wenhao Ma, Zhen Liu, Yicong Guo, Edmud Moss, Ian Stride, Lily Clements, Dongping Qiao, and Xiying Wang for their valuable contributions to the program adaptation. We are especially grateful to Yicong Guo for her assistance with data collection. We wish to thank Xinnan Zhao, Bing Shen, and Jiawen Zhu for providing implementation assistance. We are deeply grateful to the county and township governments, local program staff, maternal and child health workers, village doctors, Jiangxi Xinyu Chengbei Preschool, headteachers, social workers, and participating families. Their engagement and support were vital to the success of this research. Language in this manuscript was improved using generative artificial intelligence tools. No content was generated by artificial intelligence.

Funding

China Postdoctoral Science Foundation (2022M720467), LEGO Foundation, CICC, and the CICC Charity Foundation. The funders were not involved in the design and conduct of the study, the collection, management, analysis, and interpretation of the data.

Data Availability

Deidentified participant data will be made available from the corresponding author upon reasonable request, following publication of the article.

Authors' Contributions

ZF and XS conceptualized the study. YQ and NZ supported the local implementation. RR and XS curated and analyzed the data. XS, ZF, RR, and NZ interpreted the results. XS drafted the manuscript with support from ZF. All authors, including XS, RR, YQ, JML, NZ, and ZF, reviewed and revised the manuscript.

Conflicts of Interest

YQ is an employee of the China Development Research Foundation. JML is the CEO of Parenting for Lifelong Health (PLH), a charitable organization based in the United Kingdom that developed the adapted program. PLH programs are open access and licensed under a Creative Commons 4.0 Attribution Share-Alike license. ZF has worked as a consultant for PLH in the past. JML has (and is participating) in a number of research studies involving the program, as an investigator, and the University of Oxford and University of Cape Town receive research funding for these.

Multimedia Appendix 1
Program description.
[DOCX File, 302 KB - jmir_v28i1e79848_app1.docx]

Multimedia Appendix 2
Comparison of the two programs.
[DOCX File, 21 KB - jmir_v28i1e79848_app2.docx]

Multimedia Appendix 3
Description of interview participants.
[DOCX File, 21 KB - jmir_v28i1e79848_app3.docx]

XSL•FO
RenderX

Multimedia Appendix 4
Summary of barriers and facilitators to reach.
[DOCX File, 15 KB - jmir_v28i1e79848_app4.docx]

Multimedia Appendix 5
Summary of barriers and facilitators to adoption.
[DOCX File, 16 KB - jmir_v28i1e79848_app5.docx]

Multimedia Appendix 6
Summary of barriers and facilitators to implementation of the digital parenting program.
[DOCX File, 15 KB - jmir_v28i1e79848_app6.docx]

Multimedia Appendix 7
Summary of barriers and facilitators to maintenance at the setting level of the digital parenting program.
[DOCX File, 16 KB - jmir_v28i1e79848_app7.docx]

Checklist 1
Checklist of iCHECK-DH guidelines. iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations.
[DOCX File, 23 KB - jmir_v28i1e79848_app8.docx]

References

1. Britto PR, Lye SJ, Proulx K, et al. Nurturing care: promoting early childhood development. Lancet 2017 Jan 7;389(10064):91-102. [doi: 10.1016/S0140-6736(16)31390-3] [Medline: 27717615]
2. Black MM, Walker SP, Fernald LCH, et al. Early childhood development coming of age: science through the life course. Lancet 2017 Jan 7;389(10064):77-90. [doi: 10.1016/S0140-6736(16)31389-7] [Medline: 27717614]
3. Aboud FE, Yousafzai AK. Global health and development in early childhood. Annu Rev Psychol 2015 Jan 3;66(433–457):433-457. [doi: 10.1146/annurev-psych-010814-015128] [Medline: 25196276]
4. Lugo-Gil J, Tamis-LeMonda CS. Family resources and parenting quality: links to children's cognitive development across the first 3 years. Child Dev 2008;79(4):1065-1085. [doi: 10.1111/j.1467-8624.2008.01176.x] [Medline: 18717907]
5. Mistry KB, Minkovitz CS, Riley AW, et al. A new framework for childhood health promotion: the role of policies and programs in building capacity and foundations of early childhood health. Am J Public Health 2012 Sep;102(9):1688-1696. [doi: 10.2105/AJPH.2012.300687] [Medline: 22813416]
6. Eyberg SM, Nelson MM, Boggs SR. Evidence-based psychosocial treatments for children and adolescents with disruptive behavior. J Clin Child Adolesc Psychol 2008 Jan;37(1):215-237. [doi: 10.1080/15374410701820117] [Medline: 18444059]
7. Kaminski JW, Claussen AH. Evidence base update for psychosocial treatments for disruptive behaviors in children. J Clin Child Adolesc Psychol 2017;46(4):477-499. [doi: 10.1080/15374416.2017.1310044] [Medline: 28459280]
8. WHO Guidelines on Parenting Interventions to Prevent Maltreatment and Enhance Parent–Child Relationships with Children Aged 0–17 Years: World Health Organization; 2022.
9. Aboud FE, Prado EL. Measuring the implementation of early childhood development programs. Ann N Y Acad Sci 2018 May;1419(1):249-263. [doi: 10.1111/nyas.13642] [Medline: 29791725]
10. Jeong J, Franchett EE, Ramos de Oliveira CV, Rehmani K, Yousafzai AK. Parenting interventions to promote early child development in the first three years of life: a global systematic review and meta-analysis. PLoS Med 2021;18(5):e1003602. [doi: 10.1371/journal.pmed.1003602]
11. Luoto JE, Lopez Garcia I, Aboud FE, et al. Group-based parenting interventions to promote child development in rural Kenya: a multi-arm, cluster-randomised community effectiveness trial. Lancet Glob Health 2021 Mar;9(3):e309-e319. [doi: 10.1016/S2214-109X(20)30469-1] [Medline: 33341153]
12. Corralejo SM, Domenech Rodríguez MM. Technology in parenting programs: a systematic review of existing interventions. J Child Fam Stud 2018 Sep;27(9):2717-2731. [doi: 10.1007/s10826-018-1117-1]
13. Solís-Cordero K, Duarte LS, Fujimori E. Effectiveness of remotely delivered parenting programs on caregiver-child interaction and child development: a systematic review. J Child Fam Stud 2022;31(11):3026-3036. [doi: 10.1007/s10826-022-02328-8] [Medline: 35615461]
14. 55th statistical report on china's internet development. : China Internet Network Information Center (CNNIC); 2024 URL: https://www.cnnic.com.cn/IDR/ReportDownloads/202411/P020241101318428715781.pdf
15. Hall CM, Bierman KL. Technology-assisted interventions for parents of young children: emerging practices, current research, and future directions. Early Child Res Q 2015;33:21-32. [doi: 10.1016/j.ecresq.2015.05.003]
16. Fang Z, Martin M, Copeland L, Evans R, Shenderovich Y. Parenting interventions during the COVID-19 pandemic: a systematic review of the rationales, process, feasibility, acceptability, and impacts of adaptation. Trauma Violence Abuse 2024 Dec;25(5):3887-3902. [doi: 10.1177/15248380241266183] [Medline: 39082191]

XSL•FO
RenderX

17. Jäggi L, Hartinger SM, Fink G, et al. Parenting in the digital age: a scoping review of digital early childhood parenting interventions in low- and middle-income countries (LMIC). Public Health Rev 2025 Jan 21;45:1607651. [doi: 10.3389/phrs.2024.1607651]
18. Baumel A, Pawar A, Kane JM, Correll CU. Digital parent training for children with disruptive behaviors: systematic review and meta-analysis of randomized trials. J Child Adolesc Psychopharmacol 2016 Oct;26(8):740-749. [doi: 10.1089/cap.2016.0048]
19. Breitenstein SM, Gross D, Christophersen R. Digital delivery methods of parenting training interventions: a systematic review. Worldviews Ev Based Nurs 2014 Jun;11(3):168-176. [doi: 10.1111/wvn.12040]
20. Xie EB, Jung JW, Kaur J, Benzies KM, Tomfohr-Madsen L, Keys E. Digital parenting interventions for fathers of infants from conception to the age of 12 months: systematic review of mixed methods studies. J Med Internet Res 2023 Jul 26;25:e43219. [doi: 10.2196/43219] [Medline: 37494086]
21. Leijten P, Rienks K, Groenman AP, et al. Online parenting support: meta-analyses of non-inferiority and additional value to in-person support. Child Youth Serv Rev 2024 Apr;159:107497. [doi: 10.1016/j.childyouth.2024.107497]
22. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
23. Ahun MN, Ali NB, Hentschel E, Jeong J, Franchett E, Yousafzai AK. A meta-analytic review of the implementation characteristics in parenting interventions to promote early child development. Ann N Y Acad Sci 2024 Mar;1533(1):99-144. [doi: 10.1111/nyas.15110] [Medline: 38354095]
24. Richter LM, Daelmans B, Lombardi J, et al. Investing in the foundation of sustainable development: pathways to scale up for early childhood development. Lancet 2017 Jan 7;389(10064):103-118. [doi: 10.1016/S0140-6736(16)31698-1] [Medline: 27717610]
25. Peters DH, Adam T, Alonge O, Agyepong IA, Tran N. Republished research: implementation research: what it is and how to do it. Br J Sports Med 2014 Apr;48(8):731-736. [doi: 10.1136/bmj.f6753]
26. Britto PR, Singh M, Dua T, Kaur R, Yousafzai AK. What implementation evidence matters: scaling-up nurturing interventions that promote early childhood development. Ann N Y Acad Sci 2018 May;1419(1):5-16. [doi: 10.1111/nyas.13720] [Medline: 29791739]
27. Lachman JM, Kelly J, Cluver L, Ward CL, Hutchings J, Gardner F. Process evaluation of a parenting program for low-income families in South Africa. Res Soc Work Pract 2018 Feb;28(2):188-202. [doi: 10.1177/1049731516645665]
28. Glasgow RE, Harden SM, Gaglio B, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. Front Public Health 2019 Mar 29;7. [doi: 10.3389/fpubh.2019.00064]
29. Gaglio B, Shoup JA, Glasgow RE. The RE-AIM framework: a systematic review of use over time. Am J Public Health 2013 Jun;103(6):e38-e46. [doi: 10.2105/AJPH.2013.301299]
30. Glasgow RE, Askew S, Purcell P, et al. Use of RE-AIM to address health inequities: application in a low-income community health center-based weight loss and hypertension self-management program. Behav Med Pract Policy Res 2013 Jun;3(2):200-210. [doi: 10.1007/s13142-013-0201-8]
31. Bakken S, Ruland CM. Translating clinical informatics interventions into routine clinical care: how can the RE-AIM framework help? J Am Med Inform Assoc 2009 Nov 1;16(6):889-897. [doi: 10.1197/jamia.M3085]
32. MacDonald B, Gibson AM, Janssen X, Kirk A. A mixed methods evaluation of a digital intervention to improve sedentary behaviour across multiple workplace settings. Int J Environ Res Public Health 2020 Jun 24;17(12):4538. [doi: 10.3390/ijerph17124538] [Medline: 32599730]
33. Stephens AB, Wynn CS, Stockwell MS. Understanding the use of digital technology to promote human papillomavirus vaccination – a RE-AIM framework approach. Hum Vaccin Immunother 2019 Aug 3;15(7-8):1549-1561. [doi: 10.1080/21645515.2019.1611158]
34. Taylor ME, Liu M, Abelson S, Eisenberg D, Lipson SK, Schueller SM. The reach, effectiveness, adoption, implementation, and maintenance of digital mental health interventions for college students: a systematic review. Curr Psychiatry Rep 2024 Dec;26(12):683-693. [doi: 10.1007/s11920-024-01545-w]
35. Yoshida Y, Patil SJ, Brownson RC, et al. Using the RE-AIM framework to evaluate internal and external validity of mobile phone–based interventions in diabetes self-management education and support. J Am Med Inform Assoc 2020 Jun 1;27(6):946-956. [doi: 10.1093/jamia/ocaa041]
36. Wang J, Hedley D, Bury SM, Barbaro J. A systematic review of screening tools for the detection of autism spectrum disorder in mainland China and surrounding regions. Autism 2020 Feb;24(2):285-296. [doi: 10.1177/1362361319871174]
37. Xiong N, Yang L, Yu Y, et al. Investigation of raising burden of children with autism, physical disability and mental disability in China. Res Dev Disabil 2011 Jan;32(1):306-311. [doi: 10.1016/j.ridd.2010.10.003]
38. Eisenhardt KM. Building theories from case study research. The Academy of Management Review 1989 Oct;14(4):532. [doi: 10.2307/258557]
39. Meredith J. Building operations management theory through case and field research. J of Ops Management 1998 Jul;16(4):441-454. [doi: 10.1016/S0272-6963(98)00023-0]

XSL•FO
RenderX

40. Bellg AJ, Borrelli B, Resnick B, et al. Enhancing treatment fidelity in health behavior change studies: best practices and recommendations from the NIH behavior change consortium. Health Psychol ;23(5):443-451. [doi: 10.1037/0278-6133.23.5.443]
41. Ward CL, Wessels IM, Lachman JM, et al. Parenting for lifelong health for young children: a randomized controlled trial of a parenting program in South Africa to prevent harsh parenting and child conduct problems. J Child Psychol Psychiatry 2020 Apr;61(4):503-512. [doi: 10.1111/jcpp.13129] [Medline: 31535371]
42. Designing, Implementing, Evaluating, and Scaling Up Parenting Interventions: A Handbook for Decision-Makers and Implementers: World Health Organization; 2024. URL: https://www.who.int/publications/i/item/9789240095595
43. Emmers D, Jiang Q, Xue H, et al. Early childhood development and parental training interventions in rural China: a systematic review and meta-analysis. BMJ Glob Health 2021 Aug;6(8):e005578. [doi: 10.1136/bmjgh-2021-005578] [Medline: 34417271]
44. Zhou J, Heckman J, Wang F, Liu B. Early childhood learning patterns for a home visiting program in rural China. J Community Psychol 2023 Mar;51(2):584-604. [doi: 10.1002/jcop.22872] [Medline: 35567396]
45. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Meth Thousand 2017 Oct 2;16(1). [doi: 10.1177/1609406917733847]
46. Kiger ME, Varpio L. Thematic analysis of qualitative data: AMEE guide no. 131. Med Teach 2020 Aug;42(8):846-854. [doi: 10.1080/0142159X.2020.1755030] [Medline: 32356468]
47. Perrin Franck C, Babington-Ashaye A, Dietrich D, et al. iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res 2023 May 10;25(1):e46694. [doi: 10.2196/46694]
48. Entenberg GA, Mizrahi S, Walker H, et al. AI-based chatbot micro-intervention for parents: meaningful engagement, learning, and efficacy. Front Psychiatry 2023 Jan 20;14. [doi: 10.3389/fpsyt.2023.1080770]
49. Day JJ, Sanders MR. Do parents benefit from help when completing a self-guided parenting program online? a randomized controlled trial comparing triple P online with and without telephone support. Behav Ther 2018 Nov;49(6):1020-1038. [doi: 10.1016/j.beth.2018.03.002] [Medline: 30316482]
50. Villalta L, Elias M, Llorens M, Vall-Roqué H, Romero-González M, Valencia-Agudo F. Implementation of the incredible years-ASLD® program in autism and preterm children with communication and/or socialization difficulties in Spain (FIRST STEPS): a feasibility study. Child Youth Serv Rev 2025 Oct;177:108422. [doi: 10.1016/j.childyouth.2025.108422]
51. Liu TL, Hsiao RC, Chou WJ, Yen CF. Parenting stress, anxiety, and sources of acquiring knowledge in Taiwanese caregivers of children with attention-deficit/hyperactivity disorder. BMC Public Health 2024 Jun 24;24(1):1675. [doi: 10.1186/s12889-024-18761-x] [Medline: 38914984]
52. Burke SM, Shapiro S, Petrella RJ, et al. Using the RE-AIM framework to evaluate a community-based summer camp for children with obesity: a prospective feasibility study. BMC Obes 2015;2(21):21. [doi: 10.1186/s40608-015-0050-8] [Medline: 26217536]
53. Wozniak LA, Soprovich AL, Johnson JA, Eurich DT. Adopting and implementing an innovative model to organize diabetes care within First Nations communities: a qualitative assessment. BMC Health Serv Res 2021 May 3;21(1):415. [doi: 10.1186/s12913-021-06424-1] [Medline: 33941176]
54. Nielsen JV, Skovgaard T, Bredahl TVG, Bugge A, Wedderkopp N, Klakk H. Using the RE-AIM framework to evaluate a school-based municipal programme tripling time spent on PE. Eval Program Plann 2018 Oct;70:1-11. [doi: 10.1016/j.evalprogplan.2018.05.005] [Medline: 29890448]
55. Cassar S, Salmon J, Timperio A, et al. Optimizing intervention dissemination at scale: a qualitative study of multi-sector partner organization experiences. Transl Behav Med 2024 Oct 6;14(10):621-633. [doi: 10.1093/tbm/ibae042] [Medline: 39216008]
56. Wozniak L, Soprovich A, Rees S, Al Sayah F, Majumdar SR, Johnson JA. Contextualizing the effectiveness of a collaborative care model for primary care patients with diabetes and depression (teamcare): a qualitative assessment using RE-AIM. Can J Diabetes 2015 Oct;39:S83-S91. [doi: 10.1016/j.jcjd.2015.05.004]
57. Cohen DJ, Crabtree BF, Etz RS, et al. Fidelity versus flexibility: translating evidence-based research into practice. Am J Prev Med 2008 Nov;35(5 Suppl):S381-S389. [doi: 10.1016/j.amepre.2008.08.005] [Medline: 18929985]
58. Miller CJ, Barnett ML, Baumann AA, Gutner CA, Wiltsey-Stirman S. The FRAME-IS: a framework for documenting modifications to implementation strategies in healthcare. Implementation Sci 2021 Dec;16(1):36. [doi: 10.1186/s13012-021-01105-3]
59. Lyu L, Mei Z, Yan F, Wang X, Duan C. The status of rural children left-behind in China: 2010–2020. China popul dev stud 2024 Jun;8(2):97-111. [doi: 10.1007/s42379-024-00159-2]

Abbreviations

FGD: focus group discussion
HBD: hospital-based program
LMIC: low- and middle-income country
PP: program participants

XSL•FO
RenderX

RE-AIM : Reach, Effectiveness, Adoption, Implementation, and Maintenance
SBD: preschool-based program

Edited by N Cahill; submitted 03.Jul.2025; peer-reviewed by M Galani, X Liang; accepted 22.Oct.2025; published 06.Jan.2026.

Please cite as:
Shi X, Ruan R, Qie Y, Lachman JM, Zhong N, Fang Z
Institutionalizing Digital Parenting Programs in Low Resource Settings in China: Comparative Case Study of Health Care and Education Sectors Using the RE-AIM Framework
J Med Internet Res 2026;28:e79848
URL: https://www.jmir.org/2026/1/e79848
doi:10.2196/79848

©Xinyu Shi, Ruochen Ruan, Yi Qie, Jamie M Lachman, Na Zhong, Zuyi Fang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Complex Digital Health Intervention to Support People With HIV: Organizational Readiness Survey Study and Preimplementation Planning for a Hybrid Effectiveness-Implementation Study

Jacqueline Hodges[1], MD, MPH; Wendy Cohn[2], PhD; Amanda Castel[3], MD, MPH; Tabor Flickinger[4], MD, MPH; Ava Lena Waldman[5], CHES, CCRP, MHS; Michelle Hilgart[6], MEd, PhD; Olivia Kirby[3], MPH; Sylvia Caldwell[5], MPH, DHSc; Karen Ingersoll[6], ABPP, PhD

[1]Division of Infectious Diseases, Department of Medicine, Duke University, 315 Trent Drive, Durham, NC, United States

[2]Department of Public Health Sciences, University of Virginia, Charlottesville, VA, United States

[3]Milken Institute School of Public Health, George Washington University, Washington, DC, United States

[4]General, Geriatric & Palliative Medicine, Department of Medicine, University of Virginia, Charlottesville, VA, United States

[5]Division of Infectious Diseases and International Health, Department of Medicine, University of Virginia, Charlottesville, VA, United States

[6]Department of Psychiatry and Neurobehavioral Sciences, University of Virginia, Charlottesville, VA, United States

Corresponding Author:
Jacqueline Hodges, MD, MPH
Division of Infectious Diseases, Department of Medicine, Duke University, 315 Trent Drive, Durham, NC, United States

Abstract

Background: Evaluating implementation of digital health interventions (DHIs) in practice settings is complex, involving diverse users and multistep processes. Proactive planning can ensure implementation determinants and outcomes are captured for hybrid studies, but operational guidance for designing or planning hybrid DHI studies is limited.

Objective: This study aimed to proactively define, prioritize, and operationalize measurement of implementation outcomes and determinants for a DHI hybrid effectiveness-implementation trial. We describe unique advantages and limitations of planning the trial implementation evaluation among a large-scale cohort study population and share results of a pretrial organizational readiness assessment.

Methods: We planned a cluster-randomized, type II hybrid effectiveness-implementation trial testing *PositiveLinks*, a smartphone app for HIV care, compared to usual care (n=6 sites per arm), among HIV outpatient sites in the DC Cohort Longitudinal HIV Study in Washington, DC. We (1) defined components of the DHI and associated implementation strategy; (2) selected implementation science frameworks to accomplish evaluation aims; (3) mapped framework dimensions, domains, and constructs to implementation strategy steps; (4) modified or created instruments to collect data for implementation outcome measures and determinants; and (5) developed a compatible implementation science data collection and management plan. Provider baseline surveys administered at intervention sites probed usage of digital tools and assessed provider readiness for implementation with the Organizational Readiness to Implement Change tool.

Results: We specified DHI and implementation strategy toward planning measurement of DHI and broader program reach and adoption. Mapping of implementation strategy steps to the Reach Effectiveness Adoption Implementation Maintenance framework prompted considerations for how to capture understudied aspects of each dimension: denominators and demographic representativeness within reach or adoption, and provider or organization-level adaptations, dose, and fidelity within the implementation dimension. Our process also prompted the creation of tools to obtain detailed determinants across domains and constructs of the Consolidated Framework for Implementation Research within a large sample at multiple time points. Some aspects of real-world *PositiveLinks* implementation were not reflected within the planned hybrid trial (eg, research assistants selected as de facto site implementation leads) or were modified to preserve internal validity of effectiveness measurement (eg, “Community of Practice”). Providers and research assistants (n=17) at intervention sites self-reported high baseline use of digital tools to communicate with patients. Readiness assessment revealed high median (48, IQR 45‐54) total Organizational Readiness to Implement Change scores, with research assistants scoring higher than physicians (52.5, IQR 44-55 vs 48.0, IQR 46-49).

Conclusions: Key takeaways, challenges, and opportunities arose in planning the implementation evaluation within a hybrid DHI trial among a cohort population. Prospective trial planning must balance generalizability of implementation processes to

XSL•FO
RenderX

"real world" conditions with rigorous procedures to measure intervention effectiveness. Rapid, scalable tools require further study to enable evaluations within large multisite hybrid studies.

Trial Registration: ClinicalTrials.gov NCT04998019; https://clinicaltrials.gov/study/NCT04998019

International Registered Report Identifier (IRRID): RR2-10.2196/37748

(J Med Internet Res 2026;28:e76327) doi:10.2196/76327

KEYWORDS

mobile health; digital health; HIV; RE-AIM; CFIR; hybrid effectiveness-implementation study; implementation science; Reach Effectiveness Adoption Implementation Maintenance; Consolidated Framework for Implementation Research

Introduction

Digital health interventions (DHIs) including web-based and mobile health (mHealth) interventions can improve clinical outcomes for chronic health conditions. DHIs apply behavior change theories with variable mechanisms of action, such as enhanced motivation, self-management, and peer support. Munoz et al [1] and Hermes et al [2] describe the spectrum of behavioral intervention technologies ranging from adjunctive tools embedded within clinic-based care to support provider tasks, to fully-automated direct-to-consumer technologies for patient self-management. Implementation outcomes or the 'who,' 'how,' 'how much,' and 'how well' needed to implement DHIs [3] can vary widely as a result. Clinic-embedded DHIs may require significant engagement by multiple providers and staff. Additionally, DHIs with multiple features can generate large amounts of backend paradata related to usage that are important for quantifying implementation reach and adoption. Capturing DHI implementation outcomes and determinants can thus be complicated. Curran et al [4] describe 3 designs of hybrid effectiveness-implementation studies. However, the literature provides little operational guidance or shared procedural knowledge on planning hybrid studies to capture implementation outcomes and identify salient implementation determinants for complex multifeature multiuser DHIs.

We reviewed several comprehensive efforts to recharacterize or adapt broader health service implementation science frameworks for the study of DHIs. Recent examples related to evaluation frameworks, which evaluate implementation outcomes, include (1) a workshop conducted by the Dissemination and Implementation Core of the Center for Technology and Behavioral Health at Dartmouth College [5], (2) Hermes et al's [2] recharacterization of Proctor's outcomes for implementation research for technology-based behavioral interventions, and (3) De la Vega et al's [6] post hoc app of this recategorized framework against Glasgow's Reach Effectiveness Adoption Implementation Maintenance (RE-AIM) framework [7,8]. Determinant frameworks answer the question: 'Why was the intervention/practice/innovation implemented or not implemented?' Several frameworks applied to DHI implementation research include the Consolidated Framework for Implementation Research (CFIR) [9-12], the Theoretical Domains Framework [13], Promoting Action on Research Implementation in Health Services [14], and others. Application of these frameworks for DHI trials is often done post hoc rather than proactively, and practical guidance on incorporating frameworks into DHI hybrid trial planning is limited, despite their importance in revealing why implementation of DHIs tested within real-world settings did or did not meet expectations.

The *PositiveLinks* platform is a clinic-embedded multifeature smartphone app with patient and provider-facing components. It was developed and refined following a rigorous, iterative process of user-centered design to support people with HIV receiving outpatient care [15]. Program implementation among a cohort in Central Virginia where the intervention was developed and refined has demonstrated long-term usage and significant improvement in clinical outcomes at 1 [16], 2 [17], and 3 years [18]. The platform has been adapted for other chronic conditions, end users, and contexts [19-24]. To date, *PositiveLinks* has been implemented as part of routine clinical care at 9 clinics in Virginia, and 8 sites in other states, and is considered an evidence-based intervention for HIV care by several national consensus guidelines [25-28]. Clinical effectiveness of PositiveLinks is currently being tested against usual care in a hybrid effectiveness-implementation trial, the *PositiveLinks* in DC Cohort Study, using a cluster randomized controlled trial design (ClinicalTrials.gov NCT04998019) [29]. The trial is being conducted among sites in the DC Cohort Longitudinal HIV Study (DC Cohort Study) following over 12,800 people with HIV at 14 outpatient HIV practice settings, including Federally Qualified Health Centers and academic medical centers [30]. The DC Cohort context for this trial, including cohort site characteristics, as well as study design, site selection, randomization, recruitment, data collection, and statistical analysis procedures, is outlined in the study protocol paper [29].

For complex DHIs like *PositiveLinks* engaging multiple end users collectively within an 'implementation climate,' it is important to establish readiness for the coordinated actions needed to implement the intervention across the organization. The theory-based measure, Organizational Readiness for Change (ORIC), was designed and validated by Weiner et al [31] to measure collective readiness for implementation of health care innovations. We opted to measure organizational readiness at each intervention site, including the survey items within provider baseline surveys, in order to understand (1) which relatively lower-scoring sites may require additional support or attention during implementation and (2) to understand at the back end postimplementation if baseline readiness was an influencing factor in provider adoption of *PositiveLinks*. To this end, we also assessed baseline provider usage of mHealth or digital health tools to understand how this experience might shape

XSL•FO
RenderX

provider adoption of a new tool within each 'implementation climate.'

We share our process to proactively define, prioritize, and operationalize evaluation of relevant implementation outcomes and determinants for this type II hybrid effectiveness-implementation trial [4] testing a complex DHI among six DC Cohort sites randomized to the intervention over a 12-month study period. We highlight the unique opportunities and challenges that emerged for planning of a hybrid DHI trial among a large-scale epidemiologic cohort study population. This manuscript shares a practical set of takeaways and considerations that stood out to us as novel or distinct from the available literature we reviewed as we prepared for our DHI trial, that is, what we 'wish we knew' before embarking on our extensive planning stage. Several lessons are considerations for teams that would ideally be incorporated as early as the study conception and design stage. Finally, we share the results of an assessment of provider baseline technology usage and pretrial readiness to implement the intervention across participating sites and discuss how results may inform posttrial analyses of implementation outcomes and determinants.

Methods

Study Team and Process Refinement

The hybrid trial planning phase spanned an 18-month period preceding onboarding of the first site in December 2022, conducted by an interdisciplinary research team. Research team members hold an established record of clinical research experience, including conducting formative evaluations and observational studies testing clinical efficacy of *PositiveLinks*. Investigators' primary expertise includes clinical psychology, program evaluation, qualitative methods, instructional design, software development, and implementation research. Program managers contributed empirical observation of *PositiveLinks* implementation processes over a decade that assisted with conceptualization of the intervention, implementation contexts, and components of the implementation strategy. The trial planning team also included DC Cohort Study investigators with expertise in epidemiologic and intervention studies at the cohort sites. The methodological approach to proactively integrate implementation evaluation activities within the hybrid trial required iterative steps conducted through multiple cycles of team feedback, consensus, and refinement (Checklist 1).

Specify Components of the DHI and Associated Implementation Strategy

Given the multifeature, multiuser nature of the intervention, we created a specified list of components of the DHI implementation process and discrete steps of the implementation strategy. This process was informed by team experience with implementation of *PositiveLinks* in other contexts as part of usual care, including a prior rigorous qualitative study summarizing key in-clinic processes necessary for *PositiveLinks* implementation [32], and a formative preimplementation study engaging stakeholders within the DC Cohort Study context to tailor the app and implementation strategy [33]. Implementation strategy steps were further specified in terms of actors, corresponding actions, and action targets mapped specifically to the DC Cohort Study context [34,35].

Select Appropriate Implementation Science Frameworks to Accomplish Evaluation Aims

We identified aims for the implementation evaluation arm of the hybrid trial: (1) define and measure implementation outcomes of interest and (2) elucidate determinants of implementation in a rapid, scaled fashion across participating sites. We first used narrative reviews of theories, models, frameworks, and strategy categorization to assess the most widely used determinant and evaluation frameworks [36,37]. We then reviewed technology-specific compendia and original research studies reconceptualizing broader health service frameworks toward DHIs, including hybrid study designs [2,5,6,8,10]. We subsequently selected the RE-AIM evaluation framework and CFIR determinant framework for our first and second implementation evaluation aims, respectively.

Map Framework Dimensions, Domains, and Constructs to Steps of the Implementation Strategy

We mapped each dimension of the evaluation framework, RE-AIM, to specified components of the intervention and implementation strategy steps, informed by prior efforts in the literature [2,6]. We prioritized measurement of specific steps based on impact of findings for informing future *PositiveLinks* implementations in this context and others, and the feasibility of measurement within the hybrid trial study. For implementation determinants, we selected salient domains/constructs identified from our prior detailed qualitative CFIR-guided assessment of several sites implementing *PositiveLinks* [32]. We also identified salient CFIR interview guide questions for conversion into survey items, which could be rapidly deployed among a larger sample of up to 50 or more cohort providers employed across trial intervention sites at multiple predefined time points (baseline, 6 months, and 12 months).

Modify or Create Instruments to Support Data Collection for Implementation Outcome Measures and Determinants

Existing data collection instruments developed for the DC Cohort Study or for prior *PositiveLinks* real-world implementation were modified with additional items or created *de novo* within Research Electronic Data Capture to completely assess each RE-AIM implementation outcome measure and salient implementation determinants. Provider survey items were generated using close-ended questions (eg, dichotomous or Likert scale responses), as well as optional free-text responses, and planned for distribution at 6- and 12-months into implementation. Semistructured interviews and focus groups were planned to elicit more detailed feedback post-implementation among a smaller subset of both provider and patient participants, respectively, with guides designed using CFIR.

XSL•FO
RenderX

Develop a Compatible Data Collection and Management Plan for Implementation Evaluation

Finally, we generated an overall plan for participant data collection and management that ensured compatibility between the clinical effectiveness arm of the trial and the implementation evaluation. Plans for implementation outcome or determinant data collection and abstraction were incorporated into patient approaches for follow-up, study monitoring, and data abstraction already planned for the trial's effectiveness arm. Provider-related activities specific to the implementation evaluation were incorporated into study onboarding processes.

Measurement of Baseline Provider Technology Usage and Implementation Readiness

Providers completed an electronic baseline assessment upon enrollment in the study with items related to familiarity, knowledge, and usage of available patient and provider-facing mHealth tools, apps, and portals in routine HIV care, probing specific experience with mHealth tools sharing functionality with components of the *PositiveLinks* intervention. The ORIC 12-item tool [31,38,39] assesses organizational members' shared resolve to implement a change (change commitment) and shared belief in their collective capability to do so (change efficacy). The ORIC was distributed as part of the provider baseline survey to provider participants at intervention sites.

Analysis

Descriptive statistics were used to analyze responses to provider survey items related to baseline technology usage and implementation readiness (close-ended response options or free-text responses). Each ORIC item is rated on a 5-point ordinal scale (1="disagree" to 5="agree"). Responses were analyzed for frequency, median value, or free-text content as appropriate. ORIC scores were characterized using medians and IQR for total scores (sum of all items) and individual subscores related to change commitment (n=5 items) and change efficacy (n=7) [38]. Statistical analysis was conducted using R version 4.1.2 (R Foundation for Statistical Computing).

Ethical Considerations

Ethical approval was obtained upon human subject research ethics review by the George Washington University Institutional Review Board (IRB) (Protocol NCR202829; ClinicalTrials.gov NCT04998019) and site-specific IRBs as required. Written study-specific informed consent for all procedures of this trial was collected in addition to the consent obtained for inclusion of clinical data among patient participants upon enrollment in the DC Cohort Study. Written informed consent was obtained for all trial provider participants using procedures outlined in the IRB-approved study protocol. Data collection, storage, and management followed all outlined procedures (eg, deidentification of baseline survey data and use of an assigned study ID with a separate link long). No monetary compensation was provided to participants for study activities described in this paper.

Results

Key Process Steps and Takeaways

Our hybrid DHI trial planning process is summarized in Figure 1, including key process steps and takeaways for research teams aiming to perform similar prospective hybrid trials in real-world settings, where complex DHI implementation and associated study procedures can be planned in advance. We highlight takeaways that emerged from our experience and lacked more applied experience or specific guidance within the literature.

Figure 1. This figure summarizes the 5-step process to plan an implementation evaluation within a hybrid trial for prospective testing of a digital health intervention (DHI). Process steps are outlined, along with respective generalizable takeaways for research teams planning comprehensive implementation evaluations for trials testing DHIs.

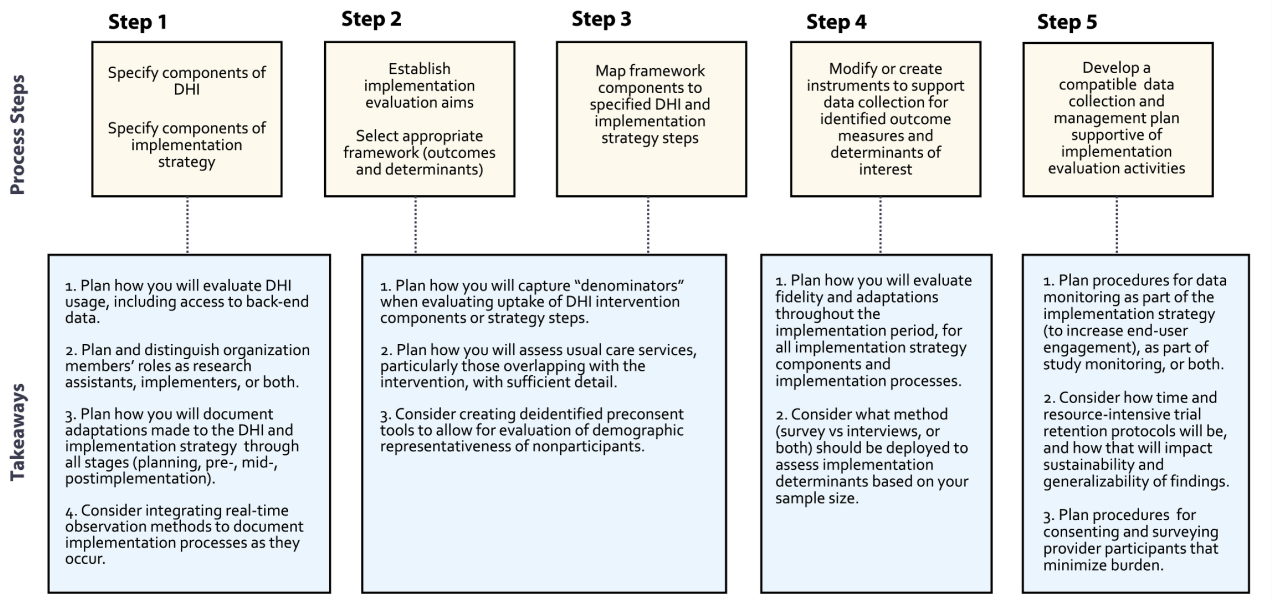

XSL•FO
RenderX

Define Components of Intervention and Implementation Strategy

Step 1 of the planning phase included specification of *PositiveLinks* intervention components (both patient and provider-facing elements) as well as the respective paradata generated by usage of each feature. The *PositiveLinks* platform contains a large amount of backend data (often described as metadata or paradata) that can be analyzed individually for every participant by feature and is accessible to the study team for analysis. Specific paradata were defined by end user group and data format (individual activity reports over selected time intervals with frequency of use, or free-text post content) in order to plan metrics of uptake for related implementation outcomes defined in steps 2 and 3. Takeaway 1: Research teams should consider during the planning stage what backend data is directly accessible for tools they are studying early in the planning process. Other pragmatic DHI trials may test commercially available tools developed by external companies or vendors, necessitating data-sharing agreements with vendors or suppliers of these tools, proactive design of trial assessments that allow participants to self-report usage at regular (eg, weekly) intervals, or integration of other commercially available tools to track DHI usage into study budgeting, design, consenting, and data collection processes.

Steps of the planned implementation strategy specified for the DC Cohort context are shown in Figure 2. “Actors” include the *PositiveLinks* core team’s program managers who assist remotely with implementation activities and troubleshooting at partner sites. The study team includes researchers involved with coordinating the specific intervention study being conducted to test *PositiveLinks* in the DC Cohort. Study team members were planned to have a larger role in site onboarding processes (determine site needs and resources, conduct research assistant and provider training) within the hybrid trial when compared to real-world implementation. Research assistants employed at DC Cohort Study sites are familiar with recruiting and consenting patients, uploading relevant patient data, and conducting assessments throughout the study period. Site research assistants were thus selected as actors to conduct several steps of the implementation strategy in this context, whereas in real-world implementation, the program is administered by individuals with a range of roles, typically identified by leadership or self-identified as the site lead for implementation. This was a decision made across sites given the extensive role and protected time research assistants have to assist with studies. Takeaway 2: Research teams designing or planning hybrid trials must consider “actors” both in terms of their research roles and implementation roles. If implementation activities in a trial are primarily assigned to a research team member, organization members already responsible for existing usual care activities overlapping with the intervention (eg, for *PositiveLinks*, retention-in-care activities) should also be engaged as early as possible. Shifting more implementation responsibilities to a research team member reduces the pragmatic nature of the trial and challenges intervention sustainability, as teams must decide who will execute those responsibilities when the trial is over.

XSL•FO
RenderX

Figure 2. Discrete implementation strategy steps expected for program deployment at DC Cohort sites are demonstrated in the figure and described, specified by step, actions, and actor in the table below it. The arrows show the sequential nature of the activities, but also note the gray dotted arrows indicate the iterative nature of using evaluation and feedback to inform various process steps as implementation proceeds. HIPAA: Health Insurance Portability and Accountability Act; PL: PositiveLinks

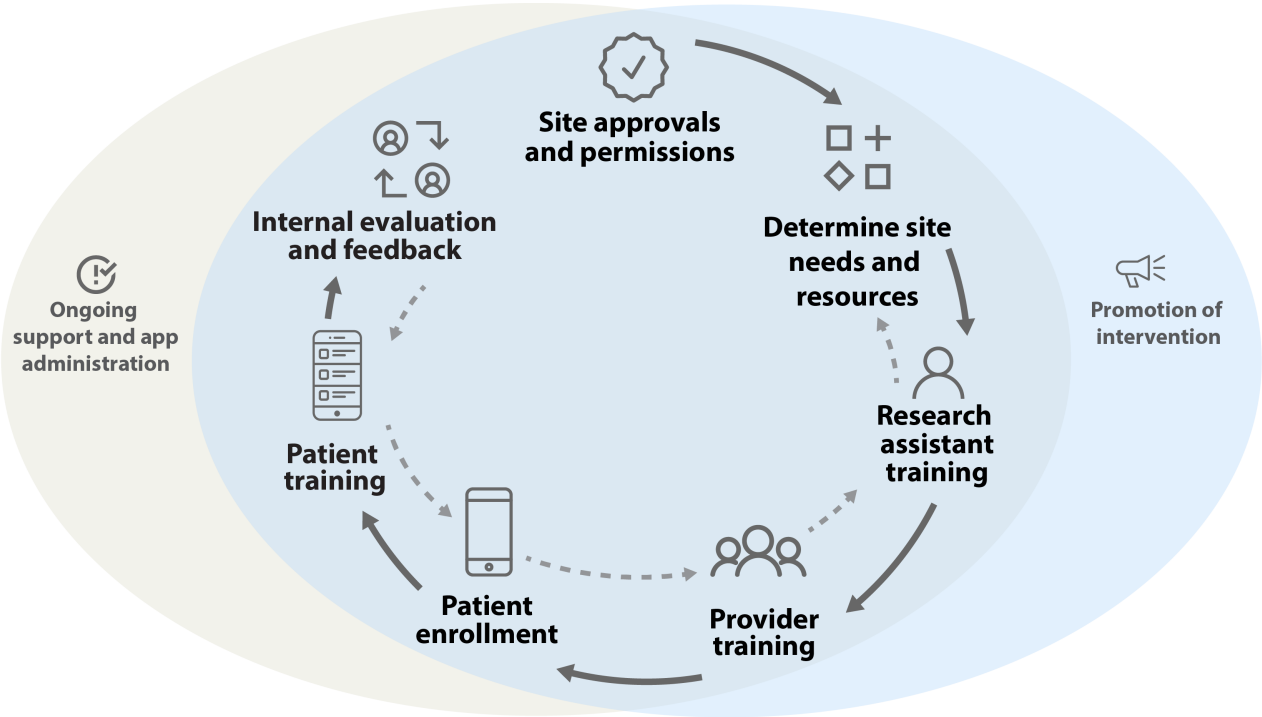

Step in implementation strategy	Actions	Actors	Action target
Site approvals and permissions	• Promote intervention to clinic and higher level leadership (study team) • Secure organizational approval of site use of intervention (study team, site principal investigator) • Identify and obtain required permissions related to patient IT security or privacy or HIPAA (study team, site principal investigator)	• Study team • Site principal investigator	• Organizational leadership
Determine site needs and resources	• Clarify roles of providers in implementation, establish staffing needs • Establish how providers will communicate with each other about program • Identify clinic values or clinical goals and how they may or may not align with PL • Determine how clinic providers would like to use portal (content shown, functionality) • Identify patient subgroups likely to benefit, needs that can be addressed with PL and issues that may arise for them related to using PL	• Study team • Site research assistant • Site Providers	• Site providers
Promotion of intervention	• Disseminate information about PL or site-specific benefits of PL to rest of site and larger organization	• Study team • Site principal investigator • Site research assistant	• Site providers • Patients • Organizational leadership
Research assistant training	• Inform site research assistant of app features so they understand how to use the app themselves and can explain benefits to providers and clinic processes, and benefits to their patients	• PL core team • Study team	• Site research assistant
Provider training	• Inform providers of app features so they understand how to use the app themselves • 'Sell' providers on the app's potential benefits to their patients, efficacy • 'Sell' app functionalities (mechanics of program) to site providers and how it will benefit them or their ability to provide care • Ensure newer providers are aware of procedures if there is turnover or new hiring	• Site research assistant	• Site providers
Patient enrollment	• Offer enrollment in the app as part of the research study and obtain informed consent • Provide preview of benefits or ease of use to patient • Sign patient up on the spot (requires WiFi access, time during appointment) • Provide smartphone or dataplan if needed	• Site research assistant	• Patients
Patient training	• Inform patient of how to navigate app interface or features and explain each in real time • Demonstrate app security or anonymity to patient • Demonstrate the customizable way patient can use the app	• Site research assistant	• Patients
Ongoing Support and app administration	• Communicate with PL core team about logistical or technical issues (site research assistant) • Troubleshoot patients' technical issues in real time or in-house when core team not available (site research assistant) • Assist research assistant with software or advanced technical issues nonurgently (PL core team)	• Site research assistant • PL core team	• PL core team • Site research assistant
Internal evaluation and feedback	• Perform surveillance or tracking of program progress, present to rest of site providers (study team and site research assistant) • Develop and carry out action plans based on interim site-specific progress or feedback (study team, site research assistant and site providers) • Ongoing champion activity to support maintenance of intervention (site providers and site research assistant)	• Study team • Site research assistant • Site providers	• Site providers • Patients

Distinctions between real-world implementation processes and those expected for the trial emerged at this planning stage. The learning management system (LMS) is a series of interactive online training modules developed for providers and staff, aimed

XSL•FO
RenderX

primarily at increasing knowledge and familiarity with the use of platform features (followed by a posttraining feedback survey). Prior to site onboarding for the study, several cohort site principal investigators expressed they would expect significantly lower enrollment of providers if LMS completion was mandatory, primarily due to significant provider clinical burden or competing demands. As a result, participating providers could opt out of LMS completion but were expected to undergo group onboarding or one-on-one training with the site research assistant prior to study participation. Takeaway 3: Research teams should consider incorporating a process to regularly document throughout the pre-, mid-, and postimplementation period, how their specific instance of DHI implementation for the trial diverges from other implementations, adaptations, or real-world practice, representing fidelity of implementation through every stage (not just during the intervention implementation period).

Real-world *PositiveLinks* implementation offers collaboration between sites in a "Community of Practice," to share implementation challenges, solutions, and unique adaptations to improve patient experiences and engagement. In contrast, we sought to maintain internal validity and avoid cross-site contamination for the trial, so intervention sites were not included in the "Community of Practice." Additionally, steps like administration of as-needed technical app support/troubleshooting and promotion of the intervention are usually more informally evaluated through site discussions and internal meetings during real-world implementation. During planning, we determined they would be challenging to evaluate with structured tools and opted to evaluate them through qualitative interviews at completion of the study. Takeaway 4: Clinic-embedded DHIs implemented in real-world practice can involve complex interactions between actors both across sites and within sites, and over time. Reserving evaluation of these processes to the late stages of implementation can result in missed opportunities to capture important real-world factors influencing implementation. Within the constraints of budget and time, research teams should consider inclusion of study methods to obtain real-time observations of these implementation processes as they occur. These could include direct observations (with field logs or notes) by implementers (providers, RAs), and small "check-ins" or "debriefs" conducted by a study team member throughout the study period.

Select Appropriate Frameworks and Map Components to Implementation Strategy Steps

Output for steps 2 and 3 are summarized in Table 1. Specifically, RE-AIM evaluation framework dimensions were mapped to each prioritized step within the implementation strategy. During this step, we identified discrete datapoints that required collection to assess each RE-AIM dimension. Other DHI hybrid trials with detailed reporting of framework app describe similar approaches, aligning DHI usage and other aspects of program uptake by patients and providers with reach and adoption dimensions, respectively [6,32]. Several unique takeaways emerged in our effort.

XSL•FO
RenderX

Table . RE-AIM[a] framework mapped to intervention/implementation strategy, corresponding outcome measures, data collection methods[b].

RE-AIM dimension	Outcome measures	Data source or data collection instrument
Reach: *How do I reach the targeted population with the intervention?*	Number of patients offered enrollment/number seen at site	DC Cohort Database, DC Cohort Study Patient Consent Logs
	Number of patients completing enrollment/number offered	PL[c] Platform Paradata, DC Cohort Study Patient Consent Logs
	Number of patients accessing PL app/number enrolled	PL Platform Paradata
	Sociodemographics of patients enrolled (representativeness)	DC Cohort Database
	Sociodemographics of patients declining enrollment (representativeness)	DC Cohort Study Patient Consent Logs
Effectiveness: *How do I know my intervention is effective?*	HIV viral load suppression (<200 copies/mL)	DC Cohort Database
	Visit constancy	DC Cohort Database
	Retention in care	DC Cohort Database
	Demographics for patients meeting efficacy end points versus those not meeting	DC Cohort Database
	PL usage patterns for patients meeting efficacy end points versus those not meeting	PL Platform Paradata
Adoption: *How do I develop organizational support to deliver my intervention?*	Number of providers who completed onboarding/number employed	Provider Baseline Survey, DC Cohort Site Assessment Survey (Multimedia Appendix 1)
	Number of providers who completed LMS[d]/number enrolled	PL Posttraining feedback survey
	Number of providers accessing PL app/number enrolled	PL Platform Paradata
	Demographics for providers	Provider Baseline Survey (Multimedia Appendix 2)
	Cohort site characteristics	DC Cohort Site Assessment Survey
Implementation: *How do I ensure the intervention is delivered properly?*	Dose, fidelity, and adaptations: patient training	Provider Follow-up Survey (Multimedia Appendix 3), Patient Interview Guides
	Dose, fidelity, and adaptations: provider training	PL Posttraining feedback survey, Provider Interview Guides
	Dose, fidelity, and adaptations: provider PL usage	PL Platform Paradata, Provider Follow-up Survey, Provider Interview Guides
Maintenance: *How do I incorporate the intervention so that it is delivered over the long term?*	Patient intent/interest to continue using PL following trial completion	Patient Interview Guides
	Provider intent/interest to continue using PL following trial completion	Provider Follow-Up Survey, Provider Interview Guides

[a]RE-AIM: Reach Effectiveness Adoption Implementation Maintenance.

[b]Data collection methods include abstraction from one or more items from existing data sources, modification or addition of items to existing data collection instruments, or creation of new instruments. Visit constancy is the proportion of 4-month intervals a visit is completed in 12 months. Retention in care per HRSA-1 definition: 2 appointments attended at least 90 days apart within 12 months.

[c]PL: PositiveLinks.

[d]LMS: learning management system.

As we reviewed the reach and adoption dimensions, we conceptualized how we could measure not only absolute numbers as measures of uptake (eg, number of patients/providers who download and use the app), but also estimate proportions: how many people could have taken up the DHI when it was offered at each site? The DC Cohort Site Assessment Survey is periodically deployed to cohort sites for updated assessment of available service delivery. We added several items to this Site Assessment Survey to capture the denominator of people employed for each type of provider role (eg, attending physician, clinic nurse, case manager, social worker, etc) as well as site-level baseline mHealth or technology use (Multimedia Appendix 1) to more adequately assess the adoption dimension for *PositiveLinks* implementation. Takeaway 1: Research teams should consider in advance how they will capture ‘denominators’ for both patients (reach) and providers/staff

XSL•FO
RenderX

serving as implementers (adoption), including a need for pretrial site-level data collection that can provide denominators, and with what frequency they need to be deployed surrounding implementation.

Several existing Site Assessment items were identified for abstraction for evaluation of the adoption dimension of RE-AIM (eg, on-site clinical services and support services, updated staffing of clinical and nonclinical providers, specialty training). Takeaway 2: Research teams testing DHIs across multiple care settings must consider how they will sufficiently characterize existing "usual care" services available, which can vary widely across care settings and change over time. This step is particularly important for DHIs that do not study direct-to-consumer tools, but instead require DHIs to become integrated into, or penetrate, an implementation climate and existing practices. Understanding to what extent the tool is serving as an adjunct or supplementary tool layered on top of an already robust organizational usual practice, versus creating a service where none existed, is crucial for interpreting and contextualizing effectiveness findings from a hybrid trial, in addition to other implementation outcomes.

A distinct opportunity also arose from conducting the hybrid trial among the DC Cohort Study. DC Cohort patients who choose not to participate in the *PositiveLinks* program at intervention sites have already consented to inclusion of their data in the cohort database, so these nonparticipants could be examined to understand sociodemographic representativeness, a component of patient "reach" within RE-AIM that appears infrequently in published applications of this framework. Patient consent logs standardized for the DC Cohort Study include patients' reasons for declining participation, with up to 3 approaches. These logs were modified to query a selected number of demographics for cohort participants declining to participate in the study (age, sex, race, ethnicity, insurance status, last CD4 count, and last HIV viral load). Takeaway 3: Teams designing DHI trials outside of existing cohorts with available data should consider use of preconsent tools that collect deidentified demographics of interest among clinic patients who decline the intervention, in order to evaluate this understudied component of patient reach.

Modify or Create Instruments to Support Data Collection for Implementation Outcome Measures and Determinants

Our process yielded several strategies to support data collection of all identified implementation outcomes: (1) *creation* of new instruments for prospective data collection specific to the implementation evaluation, (2) *modification* of standardized tools used for DC Cohort intervention studies or *PositiveLinks* evaluations or (3) *abstraction* from existing *PositiveLinks* or DC Cohort sources (eg, DC Cohort Database storing patient encounter, laboratory, and sociodemographic data). Modification and abstraction planned based on steps 2 and 3 from existing data sources are described previously.

Instruments created specifically for the implementation arm of the hybrid trial included surveys directed at providers not usually targeted by data collection for intervention studies at cohort sites. The provider baseline survey (Multimedia Appendix 2) and follow-up survey (Multimedia Appendix 3), annotated with respective framework components, were designed to capture otherwise unincorporated data points for measurement for the implementation dimension of RE-AIM (eg, fidelity, dose, and any adaptations made to steps of the implementation strategy by site providers or research assistants). Many DHIs require a complex set of steps to ensure both patient and provider engagement. For *PositiveLinks*, this included providers themselves understanding the app features, how they work and their impact, then remembering and being motivated to bring up the app and promote it during a routine clinical encounter, be able to describe its features and benefits, and refer the patient to a staff member (eg, research assistant). The research assistant must then effectively assist the patient to download the app, register for the account, and train them on its usage (Figure 2). The provider follow-up surveys were thus designed to capture the extent to which providers performed each of these tasks over multiple time points (Multimedia Appendix 3), in addition to questions probing self-rated fidelity and adaptations made to the use of the tool and its features over multiple time points. Takeaway 1: Comprehensive evaluations of DHIs should consider fidelity and adaptations of aspects of implementation beyond direct end user engagement with the digital tool, and data collection instruments may consequently need to be created to assess implementation outcomes related to these specific steps over multiple time points.

Salient implementation determinants adapted for inclusion from our prior CFIR-guided rapid evaluation study of *PositiveLinks* real-world implementation were: inner setting (compatibility), outer setting (patient needs and resources, external policy, and incentives), characteristics of individuals (knowledge and beliefs), innovation characteristics (adaptability and complexity), and implementation process (planning and engagement of key stakeholders). Postimplementation patient focus group and provider in-depth interview guides were also adapted from prior *PositiveLinks* implementations using salient CFIR 1.0 domains or constructs. Takeaway 2: Surveys designed using CFIR or other determinant frameworks offer an opportunity to more rapidly probe a wide array of domains and constructs among a larger sample, but should be planned in conjunction with richer data collection methods (qualitative). The surveys we designed were limited to previously identified salient constructs during our rapid evaluation study in other contexts (cite), and if used alone would miss important contextual factors for this trial.

Develop a Compatible Data Collection and Management Plan for Implementation Evaluation

Finally, we developed a plan to specify timing and frequency of data abstraction (eg, from the DC Cohort Database) and collection in relation to planned activities for the effectiveness arm of the trial (eg, patient consent or enrollment, administration of baseline, 6 mo, and 12 mo assessments). For patient data, plans were designed to minimize separate approaches as well as 'data pulls' from existing sources anticipated to support evaluation of clinical effectiveness outcomes. Further, monitoring of feature usage from platform paradata for patients and providers is a routine part of real-world *PositiveLinks* implementation, to guide efforts to engage and re-engage staff

XSL•FO
RenderX

and enrolled patients, troubleshoot concerns in real-time, and ensure sustainability. Frequency of paradata abstraction for monitoring was thus predetermined for specific features at timepoints throughout implementation. Takeaway 1: Research teams should consider data monitoring both as a study activity and as a part of the implementation strategy. If regular data monitoring is expected to increase engagement or generate actionable data for implementers, this activity itself becomes a component of the implementation strategy and should both be integrated into program planning and measured for fidelity to plans along with other strategy components. Takeaway 2: Similarly, trial retention protocols whereby participants showing low engagement are contacted by study team members through a series of time and resource-intensive activities ultimately impact intervention uptake (reach and adoption). The design of these protocols should be considered in the interpretation of generalizability/sustainability of observed implementation outcomes.

We required informed consent for all provider activities in the evaluation, including collection of provider survey responses and participation in postimplementation in-depth interviews. Selection of the timing and frequency of provider survey administration required consideration of provider turnover, particularly in participating intervention sites with higher expected turnover (eg, trainees like infectious disease fellows rotating within academic centers). Takeaway 3: Implementation evaluations engage providers as participants, which is distinct from most intervention efficacy trials, and onboarding processes should consider integrating consent procedures and baseline survey administration to reduce the burden and frequency of study procedures on participating providers.

Baseline mHealth/technology Use and PositiveLinks Implementation Readiness

A total of 17 providers and RAs have completed provider baseline surveys to date. Self-reported mHealth/technology use for various aspects of patient care at baseline is summarized in Table 2. Among the 17 respondents, 9 reported access to a patient messaging feature via their electronic medical record system (EMR). Usage of additional non-EMR messaging tools was reported by 6 providers. Overall satisfaction with non-EMR methods of messaging was high for those reporting usage (n=6), and reported frequency of use was higher for non-EMR tools compared to EMR-based messaging. Fewer respondents completed optional survey items related to non-EMR tools for sharing lab results or exchanging documents with patients, with variable frequency and satisfaction with described tools.

We incorporated the ORIC tool into the provider baseline survey to assess collective baseline readiness at each site [31]. The median total ORIC score for all providers was 48 (IQR 45 - 54, possible score range 12 - 60), with RAs scoring slightly higher than physicians (52.5 (44-55) vs 48.0 (46-49)). Median total scores for change commitment and change efficacy were 20 (IQR: 19 - 22, possible score range 5 - 25) and 28 (27 - 30, possible score range 7 - 35). Total change efficacy scores were slightly higher for RAs compared to providers (30.5 vs 28.0).

XSL•FO
RenderX

Table . Provider and research assistant baseline usage of mobile health (mHealth) tools by functionality.

Baseline survey questions	Response rate, n (%)
Messaging	
Does your EMR[a] system have a patient portal that allows you to directly message with your patients? (n=17)	
Yes	9 (53)
No	6 (35)
Not sure	2 (12)
If yes, how frequently do you use it to message with your patients? (n=9)	
Never	2 (22)
Rarely	2 (22)
Occasionally	3 (33)
Frequently	2 (22)
Very frequently	0 (0)
Please specify name of apps or tools or websites used to message patients (n=6)	
Ring central	2 (33)
Halo	1 (17)
EMR mobile or web-based platform	1(17)
Text messaging	2 (33)
In the past 3 months, how many times have you used this app/tool/site to message patients? (n=6)	
Never	0 (0)
Rarely	0 (0)
Occasionally	1 (17)
Frequently	4 (67)
Very frequently	1 (17)
To what extent are you satisfied with these telemedicine services used for messaging? (n=6)	
Very unsatisfied	0 (0)
Unsatisfied	0 (0)
Neutral	0 (0)
Satisfied	4 (67)
Very satisfied	2 (33)
Laboratory results	
Please specify name(s) of app/tool/website(s) to share lab results. (n=3)	
EMR autosend letter	1 (33)
EMR mobile or web-based platform	1 (33)
Clinic phone	1 (33)
In the past 3 months, how many times have you used this app/tool/site to share laboratory results? (n=3)	
Never	0 (0)
Rarely	0 (0)
Occasionally	1 (33)
Frequently	1 (33)
Very frequently	1 (33)
To what extent are you satisfied with this app or tool or website for sharing laboratory results? (n=3)	
Very unsatisfied	0 (0)

XSL•FO
RenderX

Baseline survey questions	Response rate, n (%)
Unsatisfied	0 (0)
Neutral	0 (0)
Satisfied	2 (67)
Very satisfied	1 (33)
Documents	
Please specify the name of app or tool or website to share or receive documents (n=2)	
Encrypted email	1 (50)
EMR mobile/web-based platform	1 (50)
In the past 3 months, how many times have you used this app/tool/site to share or receive documents? (n=2)	
Never	0 (0)
Rarely	0 (0)
Occasionally	1 (50)
Frequently	1 (50)
Very frequently	0 (0)
To what extent are you satisfied with this app or tool or website for sharing or receiving documents? (n=2)	
Very unsatisfied	0 (0)
Unsatisfied	1 (50)
Neutral	0 (0)
Satisfied	1 (50)
Very satisfied	0 (0)

[a]EMR: electronic medical record.

Individual items corresponding to change efficacy (CE1-CE7) were scored on average between 4 and 4.2, with median scores of 4 (IQR 3-4) for all items (“Somewhat agree”), and no items scoring lower than 3 (“Neither agree nor disagree”; Figure 3). All change efficacy scores ranged from 3 to 5 except for CE7 (‘People who work here feel confident that they can manage the politics of implementing this change’), which along with CE3 (”People who work here feel confident that the organization can support people as they adjust to this change”) had a distribution with greater spread due to relatively higher frequency of neutral or “3” scores.

ORIC items corresponding to change commitment (CC1-CC5) similarly scored with a median of 4 (IQR 3-5) for all items (Figure 4), means ranging from 3.88 to 4.2, and scores also ranged from 3 to 5 for all items. Change commitment score distributions were slightly more negatively skewed for items CC2 (“People who work here will do whatever it takes to implement this change”) and CC4 (“People who work here are determined to implement this change”). When analyzed by provider type, mean scores were higher across all individual ORIC items for RAs compared to physicians. Median scores within subgroups were 4 (with IQR values ranging from 3.5 to 5, Figure 4) for a majority of items except for two change commitment items with higher median scores for RAs: CC3 (“People who work here want to implement this change”) and CC5 (“People who work here are motivated to implement this change”).

XSL•FO
RenderX

Figure 3. Baseline ORIC scores summarized for all respondents (n=17) at 6 sites randomized to implement the intervention (A) as well as by respondent type such as research assistants and attending physicians (B). One social worker completed the survey but ultimately withdrew from the program. ORIC: Organizational Readiness to Implement Change.

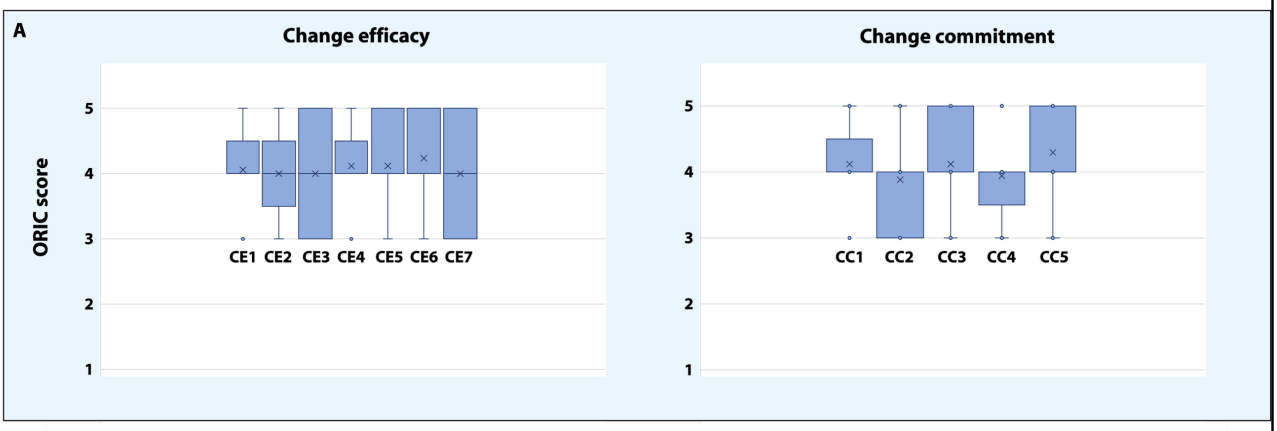

B

ORIC Item		All (n=17)	Research assistants (n=9)	Attending physicians (n=7)
Change efficacy				
CE1: People who work here feel confident that the organization can get people invested in implementing this change.	Median (IQR) Mean (SD, Range)	4.0 (4,4) 4.1 (0.7, 3-5)	4.0 (4,5) 4.3 (0.5, 4-5)	4.0 (3.5,4) 3.9 (0.7, 3-5)
CE2: People who work here feel confident that they can keep track of progress in implementing this change.		4.0 (4,4) 4.0 (0.7, 3-5)	4.0 (4,5) 4.2 (0.8, 3-5)	4.0 (4,4) 3.9 (0.4, 3-4)
CE3: People who work here feel confident that the organization can support people as they adjust to this change.		4.0 (3,5) 4.0 (0.8, 3-5)	4.0 (4,5) 4.2 (0.8, 3-5)	4.0 (3,4) 3.6 (0.5, 3-4)
CE4: People who work here feel confident that they can keep the momentum going in implementing this change.		4.0 (4,4) 4.1 (0.6, 3-5)	4.0 (4,5) 4.3 (0.7, 3-5)	4.0 (3,4) 4.0 (0.0, 4-4)
CE5: People who work here feel confident that they can handle the challenges that might arise in implementing this change.		4.0 (4,5) 4.1 (0.7, 3-5)	4.0 (4,5) 4.1 (0.8, 3-5)	4.0 (4,4) 4.0 (0.6, 3-5)
CE6: People who work here feel confident that they can coordinate tasks so that implementation goes smoothly.		4.0 (4,5) 4.2 (0.7, 3-5)	4.0 (4,5) 4.2 (0.8, 3-5)	4.0 (4,4) 4.1 (0.4, 4-5)
CE7: People who work here feel confident that they can manage the politics of implementing this change.		4.0 (3,4) 4.0 (0.8, 3-4)	4.0 (4,5) 4.2 (0.8, 3-5)	4.0 (3.5,4) 3.9 (0.7, 3-5)
Change commitment				
CC1: People who work here are committed to implementing this change.	Median (IQR) Mean (SD, Range)	4.0 (4,4) 4.1 (0.6, 3-5)	4.0 (4,5) 4.4 (0.5, 4-5)	4.0 (4,4) 3.9 (0.4, 3-4)
CC2: People who work here will do whatever it takes to implement this change.		4.0 (3,4) 3.9 (0.7, 3-5)	4.0 (4,5) 4.1(0.8, 3-5)	4.0 (3.5,4) 3.7 (0.5, 3-4)
CC3: People who work here want to implement this change.		4.0 (4,5) 4.1 (0.7, 3-5)	5.0 (4,5) 4.3 (0.9, 3-5)	4.0 (4,4) 4.0 (0.0, 4-4)
CC4: People who work here are determined to implement this change.		4.0 (4,4) 3.9 (0.7, 3-5)	4.0 (4,5) 4.1 (0.8, 3-5)	4.0 (4,4) 3.9 (0.4, 3-4)
CC5: People who work here are motivated to implement this change.		4.0 (4,5) 4.3 (0.7, 3-5)	5.0 (4,5) 4.4 (0.7, 3-5)	4.0 (4,4.5) 4.3 (0.5, 4-5)

Discussion

Principal Results

We describe a study team's experience planning and proactively integrating implementation science methodology into a hybrid DHI study conducted within an existing cohort study. This experience included a process to proactively define, prioritize, and operationalize measurement of implementation outcomes salient to a prospective DHI hybrid trial, as well as development of data collection instruments compatible with rapid assessment of implementation determinants and outcomes at scale for a large cohort and over multiple time points throughout the trial. Our experience highlights the limitations of prospective "pragmatic" trial designs to reflect real-world DHI implementation, and the unique opportunities and challenges of DHI trial planning within a large-scale cohort study population. We assessed baseline provider mHealth or technology usage and organizational readiness at baseline, which will be applied as a site-level characteristic or covariate during exploratory analysis of differential site and provider-level adoption of the intervention, including specific features, throughout the implementation period. We predict findings may align with site-specific implementation determinants revealed by CFIR application (eg, innovation relative advantage) during qualitative analysis, as well as account for variation in implementation outcome measures (eg, adoption by platform feature) across sites where providers and research assistants, for example, were already frequently using existing digital tools overlapping with *PositiveLinks* functionality that were fully integrated into their clinic's EMR with a high degree of satisfaction and had high ORIC scores. Readiness assessment results varied across sites but were not sufficiently different to prompt differential approaches to implementation support across sites. Results suggest that among the providers and RAs queried at the 6 intervention sites preimplementation, belief in their sites' collective capability to implement the program was high overall, particularly in relation to internal abilities to coordinate tasks and handle challenges arising within the sites. However,

XSL•FO
RenderX

respondents' confidence was lower in external actors' influence on implementation, or the organizational politics and support necessary to execute program components. To increase collective change efficacy of providers and research assistants, concerted preimplementation evaluation efforts may be needed in future *PositiveLinks* implementations to build site members' confidence that higher organizational levels of support for the program are present.

Comparison With Prior Work

To apply relevant implementation science frameworks to DHI implementation evaluation, we reviewed how investigators compared Proctor's Outcomes for Implementation Research recategorized for DHIs against RE-AIM [2,6]. These and other original research studies had limitations including a lack of measurable eligible "denominators" available for certain outcomes (eg, provider adoption dimension of RE-AIM). Prior studies also focused on conceptualizing outcomes for a primarily patient-facing intervention itself (eg, different ways of leveraging back-end usage capture data and tracking referrals of patients by clinic to the study), rather than including additional strategy steps to implement a dual-facing (patient and provider) intervention (mapped in Figure 2), which require dedicated measurement of both patient and provider inputs.

While RE-AIM offered adequate operational guidance to evaluate this DHI and implementation strategy as with some prior studies [2,6], several takeaways, opportunities, and challenges arose throughout our planning process as detailed throughout this article and summarized in Figure 1. Since the completion of our planning process and preparation of this manuscript, a framework for designing DHI hybrid trials has been published [32], which shares similarities with our identified planning process steps (specify components of the digital intervention as well as support services and implementation strategy, delineate domains being studied, ie, actors and action targets). Our takeaways delve deeper into considerations researchers should make based on specific experience with: (1) designing a randomized trial where the digital intervention is tested against a usual care condition, contains multiple features and engages both patients and providers or organization members; (2) integrating understudied components of RE-AIM and related evaluation frameworks into DHI trial design (eg, representativeness, fidelity, dose, and adaptation); and (3) conducting a trial among an epidemiologic cohort.

There is a well-recognized tension between maintaining the rigor and validity of randomized trial designs and establishing pragmatic conditions more relevant to real-world implementation [40]. By performing a preimplementation planning phase for a prospective DHI trial that incorporated frameworks proactively, we ensured collection of a robust set of quantitative and qualitative data. In contrast, post hoc evaluations conducted in other studies available in the literature are frequently reported among "real world" conditions or 'usual care' expected of implementation research as most strictly defined. When not planned in advance for inclusion, there are significant limits on the extent to which additional implementation outcomes like fidelity, dose, program adaptations, and uptake of specific components or steps of implementation have been evaluated with sufficient granularity and coverage of participants over the course of implementation [41,42] in these studies. Simultaneously, however, tighter control of prospective hybrid trial conditions and administration of the program through protocolized research activities inherently limits how well 'real-world' or pragmatic conditions are reflected within a hybrid trial.

Limitations

Several limitations emerged within our process. This more comprehensive planning process for the implementation evaluation was undertaken following receipt of funding for the award, which is common for hybrid trials. We found that ideally, this process should be undertaken as early as possible. Several takeaways from our planning process require attention this early, impacting study design and planned procedures, and consequently even impacting study budgeting and scope of work for research staff.

We noted several key limitations regarding the intended pragmatic nature of our prospective hybrid DHI trial. Real-world *PositiveLinks* implementation often relies on outreach by partner sites, by individuals who serve as champions of the intervention with an active role in obtaining site approvals (related to data security and patient privacy), and who continue to promote the intervention. Research assistants were assigned as de facto program managers at sites participating in the hybrid trial; however, this is a major distinction from real-world implementation. This decision, while needed to rapidly plan and conduct a multisite trial within budgetary and time-related constraints, represents a tradeoff in terms of generalizability of this planned prospective *PositiveLinks* implementation research when the intervention requires a distinct, multilayered 'implementation climate' and 'champions' within that climate. These 'climates' typically require gradual building of multiple, interacting implementers' self-efficacy, motivation, and longitudinal intervention promotion efforts to ensure site readiness and penetration.

Additional considerations for hybrid trials implementing DHIs may represent challenges to generalizability in terms of real-world maintenance and sustainability, including providing technology to patients (eg, smartphones, data plans), incentives for usage of the intervention or specific features, and participant retention protocols common for clinical efficacy trials. Real-world PositiveLinks implementation variably includes provision of phones and data plans, depending on specific partner site funding availability and patient need, and re-engagement protocols are also a routine part of implementation at several sites. For this trial, we planned to rely on existing site-specific or federal subsidy programs available to participants in the context (eg, Federal Communications Commission Lifeline program) before providing smartphones and data plans, and a retention protocol was used by site RAs to periodically re-engage patients. No additional incentives were planned for higher levels of app usage, however, and patients who do receive smartphones can keep them for the duration of the study regardless of app usage.

Examining implementation in parallel with a cluster randomized effectiveness trial among an epidemiological cohort presented

XSL•FO
RenderX

another set of unique challenges and opportunities. This hybrid approach allows for a scaled implementation evaluation to occur across multiple sites simultaneously, leveraging existing research infrastructure, including site staffing with research assistants and existing data collection instruments. The scaled, simultaneous multisite approach, however, engages a larger number of patients and providers within time and funding constraints of a single hybrid trial and necessitates consideration of more rapid, cost-effective approaches to implementation evaluation (acknowledged by de la Vega et al [6]). Provider surveys, for example, were designed to capture implementation outcome measures and determinants for the larger expected sample of respondents at 6 intervention sites than are typically engaged with more in-depth qualitative processes applying these frameworks, in particular CFIR. There is no consensus in the implementation science field about how these frameworks reflecting complex psychosocial/behavioral constructs should be applied, including whether to attempt to dichotomize or categorize items for broader, rapid distribution. Surveys could, for example, introduce study team bias during creation and selection of specific items to probe and limit more systematic application of the framework [37]. Combining surveys and qualitative approaches can offer opportunities to validate the latter methods, but more extensive psychometric validation of survey tools is needed to ensure generalizability and validity. Finally, readiness assessment findings are based on self-reported survey items, subject to scoring biased by the individuals' level of involvement with pretrial planning and other procedures (eg, research assistants vs physicians).

Conclusions

Implementation research for complex DHIs can expand understanding of how these interventions will behave in "real world" conditions. Prospective hybrid effectiveness-implementation trials can facilitate more in-depth implementation evaluations at scale if appropriately planned. Our experience highlights the ways in which evaluations must attempt to balance rigor, proximity to "real world" implementation climates, and incorporate multiple key implementation outcomes and determinants within the time and resource constraints of a prospective DHI hybrid trial. Based on our experience, planning processes for hybrid DHI trials should include:

- Specification of discrete DHI and associated implementation strategy components, considering how end users will engage with each, and what study procedures should be planned and budgeted to adequately measure that engagement (eg, backend paradata).
- Strategies to observe and document adaptations and real-time implementation processes throughout the planning, pre-, mid-, and postimplementation periods.
- Plans ahead of time to capture denominators of uptake outcomes (reach for patients and adoption for providers), demographic representativeness within those reached by the DHI versus not, and a method to capture usual care services that overlap with DHI functionalities across trial sites.
- Plans to evaluate fidelity and adaptations to the DHI and implementation strategy steps carried out among the implementing organization site that go beyond the use of the tool itself.
- Considerations for how to obtain detailed descriptive data related to implementation determinants in a larger sample size of participants (eg, design survey tools, interviews, or both using determinant frameworks).
- Plans for study procedures that minimize provider or implementer burden but enable consenting, data monitoring, and surveying of those providers, and consider in protocol design how retention protocols and use of research staff over "usual care" staff challenges generalizability or sustainability.

Acknowledgments

The authors would like to acknowledge all the collaborators of the participating DC Cohort clinics as well as input from members of the PositiveLinks team. No generative artificial intelligence tools were used for manuscript preparation.

Data in this manuscript were collected by the DC Cohort Study Group with investigators and research staff located at: Children's National Hospital Pediatric clinic (Natella Rakhmanina); the Senior Deputy Director of the DC Department of Health HAHSTA (Clover Barnes); Family and Medical Counseling Service (Rita Aidoo); Georgetown University (Princy Kumar); The George Washington University Biostatistics Center (Tsedenia Bezabeh, Vinay Bhandaru, Asare Buahin, Nisha Grover, Lisa Mele, Susan Reamer, Alla Sapozhnikova, Greg Strylewicz, and Marinella Temprosa); The George Washington University Department of Epidemiology (Shannon Barth, Morgan Byrne, Amanda Castel, Alan Greenberg, Shannon Hammerlund, Paige Kulie, Anne Monroe, Lauren O'Connor, James Peterson, and Mark Storey) and Department of Biostatistics and Bioinformatics; The George Washington University Medical Faculty Associates (Jose Lucar); Howard University Adult Infectious Disease Clinic (Jhansi L. Gajjala) and Pediatric Clinic (Sohail Rana); Kaiser Permanente Mid-Atlantic States (Michael Horberg); La Clinica Del Pueblo (Ricardo Fernandez); MetroHealth (Duane Taylor); Washington Health Institute, formerly Providence Hospital (Jose Bordon); Unity Health Care (Gebeyehu Teferi); Veterans Affairs Medical Center (Debra Benator and Rachel Denyer); Washington Hospital Center (Adam Klein); and Whitman-Walker Institute (Stephen Abbott).

Funding

AC and KI are multiple principal investigators on this project which is funded by the National Institute of Mental Health (R01MH122375). The DC Cohort Longitudinal HIV Study is funded by the National Institute of Allergy and Infectious Diseases

XSL•FO
RenderX

(UM1AI069503 and 1R24AI152598-01). JH received additional support from the National Institutes of Health (T32AI007046 and 5P30AI064518-20). The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets used and analyzed during the current study are available from the DC Cohort Longitudinal HIV Study following approval of request(s) from the study team and DC Cohort Executive Committee. Access is restricted in this way to protect the privacy of cohort participants, who have consented to inclusion of their data according to this restricted access policy. Requests for access can be submitted to the study authors and will be followed by a screening process first by the study team, then by the DC Cohort Executive Committee.

The datasets used and analyzed during the current study are available from the DC Cohort Longitudinal HIV Study following approval of request(s) from the study team and DC Cohort Executive Committee.

Conflicts of Interest

ALW and KI have active consulting agreements with Warm Health Technology, Inc., a wholly owned subsidiary of the University of Virginia Licensing & Ventures Group, which distributes the *Positive Links* program as *PL Cares*. JH is the founder and CEO of emPath technologies, LLC.

Multimedia Appendix 1
DC cohort site assessment survey with modifications.
[DOCX File, 18 KB - jmir_v28i1e76327_app1.docx]

Multimedia Appendix 2
Provider baseline survey.
[DOCX File, 18 KB - jmir_v28i1e76327_app2.docx]

Multimedia Appendix 3
Provider follow-up survey.
[DOCX File, 25 KB - jmir_v28i1e76327_app3.docx]

Checklist 1
CONSORT-EHEALTH (V 1.6.1) checklist.
[PDF File, 1238 KB - jmir_v28i1e76327_app4.pdf]

References

1. Muñoz RF. The efficiency model of support and the creation of digital apothecaries. Clin Psychol Sci Pract 2017;24(1):46-49. [doi: 10.1111/cpsp.12174]
2. Hermes EDA, Lyon AR, Schueller SM, Glass JE. Measuring the implementation of behavioral intervention technologies: recharacterization of established outcomes. J Med Internet Res 2019;21(1):e11752. [doi: 10.2196/11752]
3. Curran GM. Implementation science made too simple: a teaching tool. Implement Sci Commun 2020;1(1):27. [doi: 10.1186/s43058-020-00001-z] [Medline: 32885186]
4. Curran GM, Bauer M, Mittman B, Pyne JM, Stetler C. Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. Med Care 2012 Mar;50(3):217-226. [doi: 10.1097/MLR.0b013e3182408812] [Medline: 22310560]
5. Lord SE, Campbell ANC, Brunette MF, et al. Workshop on implementation science and digital therapeutics for behavioral health. JMIR Ment Health 2021 Jan 28;8(1):e17662. [doi: 10.2196/17662] [Medline: 33507151]
6. de la Vega R, Ritterband L, Palermo TM. Assessing digital health implementation for a pediatric chronic pain intervention: comparing the RE-AIM and BIT frameworks against real-world trial data and recommendations for future studies. J Med Internet Res 2020 Sep 1;22(9):e19898. [doi: 10.2196/19898] [Medline: 32870158]
7. Glasgow RE, Harden SM, Gaglio B, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. Front Public Health 2019;7:64. [doi: 10.3389/fpubh.2019.00064] [Medline: 30984733]
8. Glass JE, Dorsey CN, Beatty T, et al. Study protocol for a factorial-randomized controlled trial evaluating the implementation, costs, effectiveness, and sustainment of digital therapeutics for substance use disorder in primary care (DIGITS Trial). Implement Sci 2023 Feb 1;18(1):3. [doi: 10.1186/s13012-022-01258-9] [Medline: 36726127]
9. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implementation Sci 2009 Dec;4(1):50. [doi: 10.1186/1748-5908-4-50]

XSL•FO
RenderX

10. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146. [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
11. Rangachari P, Mushiana SS, Herbert K. A scoping review of applications of the Consolidated Framework for Implementation Research (CFIR) to telehealth service implementation initiatives. BMC Health Serv Res 2022 Nov 30;22(1):1450. [doi: 10.1186/s12913-022-08871-w] [Medline: 36447279]
12. Heinsch M, Wyllie J, Carlson J, Wells H, Tickner C, Kay-Lambkin F. Theories informing eHealth implementation: systematic review and typology classification. J Med Internet Res 2021 May 31;23(5):e18500. [doi: 10.2196/18500] [Medline: 34057427]
13. Jennings HM, Morrison J, Akter K, et al. Developing a theory-driven contextually relevant mHealth intervention. Glob Health Action 2019;12(1):1550736. [doi: 10.1080/16549716.2018.1550736] [Medline: 31154988]
14. Connolly SL, Hogan TP, Shimada SL, Miller CJ. Leveraging implementation science to understand factors influencing sustained use of mental health apps: a narrative review. J Technol Behav Sci 2021;6(2):184-196. [doi: 10.1007/s41347-020-00165-4] [Medline: 32923580]
15. Laurence C, Wispelwey E, Flickinger TE, et al. Development of PositiveLinks: a mobile phone app to promote linkage and retention in care for people with HIV. JMIR Form Res 2019 Mar 20;3(1):e11578. [doi: 10.2196/11578] [Medline: 30892269]
16. Dillingham R, Ingersoll K, Flickinger TE, et al. PositiveLinks: a mobile health intervention for retention in HIV care and clinical outcomes with 12-month follow-up. AIDS Patient Care STDS 2018 Jun;32(6):241-250. [doi: 10.1089/apc.2017.0303] [Medline: 29851504]
17. Canan CE, Waselewski ME, Waldman ALD, et al. Long term impact of PositiveLinks: clinic-deployed mobile technology to improve engagement with HIV care. PLoS ONE 2020;15(1):e0226870. [doi: 10.1371/journal.pone.0226870] [Medline: 31905209]
18. Bielick C, Canan C, Ingersoll K, Waldman AL, Schwendinger J, Dillingham R. Three-year follow-up of PositiveLinks: higher use of mHealth platform associated with sustained HIV suppression. AIDS Behav 2024 Aug;28(8):2708-2718. [doi: 10.1007/s10461-024-04405-z] [Medline: 38869759]
19. Hodges J, Zhdanova S, Koshkina O, et al. Implementation of a mobile health strategy to improve linkage to and engagement with HIV care for people living with HIV, Tuberculosis, and substance use in Irkutsk, Siberia. AIDS Patient Care STDS 2021 Mar;35(3):84-91. [doi: 10.1089/apc.2020.0233] [Medline: 33538649]
20. Hodges J, Waselewski M, Harrington W, et al. Six-month outcomes of the HOPE smartphone application designed to support treatment with medications for opioid use disorder and piloted during an early statewide COVID-19 lockdown. Addict Sci Clin Pract 2022 Mar 7;17(1):16. [doi: 10.1186/s13722-022-00296-4] [Medline: 35255965]
21. Clement ME, Lovett A, Caldwell S, et al. Development of an mHealth app to support the prevention of sexually transmitted infections among Black men who have sex with men engaged in pre-exposure prophylaxis care in New Orleans, Louisiana: qualitative user-centered design study. JMIR Form Res 2023 Feb 27;7:e43019. [doi: 10.2196/43019] [Medline: 36848209]
22. Mugabirwe B, Flickinger T, Cox L, Ariho P, Dillingham R, Okello S. Acceptability and feasibility of a mobile health application for blood pressure monitoring in rural Uganda. JAMIA Open 2021 Jul;4(3):ooaa068. [doi: 10.1093/jamiaopen/ooaa068] [Medline: 34514350]
23. Hodges J, Waldman AL, Koshkina O, et al. Process evaluation for the adaptation, testing and dissemination of a mobile health platform to support people with HIV and tuberculosis in Irkutsk, Siberia. BMJ Open 2022 Mar 29;12(3):e054867. [doi: 10.1136/bmjopen-2021-054867] [Medline: 35351714]
24. Flickinger TE, Sherbuk JE, Petros de Guex K, et al. Adapting an m-Health intervention for Spanish-speaking Latinx people living with HIV in the Nonurban Southern United States. Telemed Rep 2021 Feb;2(1):46-55. [doi: 10.1089/tmr.2020.0018] [Medline: 33817694]
25. Interventions: positivelinks. : Center for innovation and engagement / National Alliance of State and Territorial AIDS Directors (NASTAD); 2021 URL: https://ciehealth.org/wp-content/uploads/2022/01/PositiveLinks_12-16-21_compressed.pdf [accessed 2025-12-16]
26. Compendium of evidence-based interventions and best practices for HIV prevention: background, methods, and criteria. : Centers for Disease Control and Prevention; 2024 URL: https://stacks.cdc.gov/view/cdc/149681 [accessed 2025-12-16]
27. Thompson MA, Horberg MA, Agwu AL, et al. Primary care guidance for persons with human immunodeficiency virus: 2020 update by the HIV medicine association of the infectious diseases society of America. Clin Infect Dis 2021 Dec 6;73(11):e3572-e3605. [doi: 10.1093/cid/ciaa1391] [Medline: 33225349]
28. Ryan white HIV/AIDS program (RWHAP) best practices compilation. Health Resources and Services Administration (HRSA). 2022. URL: https://targethiv.org/library/cie-positivelinks [accessed 2025-12-16]
29. Hodges J, Caldwell S, Cohn W, et al. Evaluation of the implementation and effectiveness of a mobile health intervention to improve outcomes for people with HIV in the Washington, DC Cohort: study protocol for a cluster randomized controlled trial. JMIR Res Protoc 2022 Apr 22;11(4):e37748. [doi: 10.2196/37748] [Medline: 35349466]
30. Greenberg AE, Hays H, Castel AD, et al. Development of a large urban longitudinal HIV clinical cohort using a web-based platform to merge electronically and manually abstracted data from disparate medical record systems: technical challenges

XSL•FO
RenderX

and innovative solutions. J Am Med Inform Assoc 2016 May;23(3):635-643. [doi: 10.1093/jamia/ocv176] [Medline: 26721732]
31. Shea CM, Jacobs SR, Esserman DA, Bruce K, Weiner BJ. Organizational readiness for implementing change: a psychometric assessment of a new measure. Implement Sci 2014 Jan 10;9(1):7. [doi: 10.1186/1748-5908-9-7] [Medline: 24410955]
32. Matson TE, Hermes EDA, Lyon AR, et al. A framework for designing hybrid effectiveness-implementation trials for digital health interventions. Ann Epidemiol 2025 Apr;104:35-47. [doi: 10.1016/j.annepidem.2025.02.007] [Medline: 40015542]
33. Caldwell S, Flickinger T, Hodges J, et al. An mHealth platform for people with HIV receiving care in Washington, district of Columbia: qualitative analysis of stakeholder feedback. JMIR Form Res 2023 Sep 19;7:e48739. [doi: 10.2196/48739] [Medline: 37725419]
34. Proctor EK, Powell BJ, McMillen JC. Implementation strategies: recommendations for specifying and reporting. Implement Sci 2013 Dec 1;8(1):139. [doi: 10.1186/1748-5908-8-139] [Medline: 24289295]
35. Fernandez ME, Ten Hoor GA, van Lieshout S, et al. Implementation mapping: using intervention mapping to develop implementation strategies. Front Public Health 2019;7:158. [doi: 10.3389/fpubh.2019.00158] [Medline: 31275915]
36. Powell BJ, McMillen JC, Proctor EK, et al. A compilation of strategies for implementing clinical innovations in health and mental health. Med Care Res Rev 2012 Apr;69(2):123-157. [doi: 10.1177/1077558711430690] [Medline: 22203646]
37. Nilsen P. Making sense of implementation theories, models and frameworks. Implement Sci 2015 Apr 21;10(1):53. [doi: 10.1186/s13012-015-0242-0] [Medline: 25895742]
38. Weiner BJ, Lewis CC, Stanick C, et al. Psychometric assessment of three newly developed implementation outcome measures. Implement Sci 2017 Aug 29;12(1):108. [doi: 10.1186/s13012-017-0635-3] [Medline: 28851459]
39. Weiner BJ. A theory of organizational readiness for change. Implement Sci 2009 Oct 19;4:67. [doi: 10.1186/1748-5908-4-67] [Medline: 19840381]
40. Geng EH, Peiris D, Kruk ME. Implementation science: relevance in the real world without sacrificing rigor. PLoS Med 2017 Apr;14(4):e1002288. [doi: 10.1371/journal.pmed.1002288] [Medline: 28441435]
41. Proctor E, Silmere H, Raghavan R, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011 Mar;38(2):65-76. [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
42. Murray E, Hekler EB, Andersson G, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016 Nov;51(5):843-851. [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]

Abbreviations

CFIR: Consolidated Framework for Implementation Research
DHI: digital health intervention
EMR: electronic medical record
IRB: institutional review board
LMS: learning management system
mHealth: mobile health
ORIC: Organizational Readiness to Implement Change
RE-AIM: Reach Effectiveness Adoption Implementation Maintenance

Edited by A Mavragani, T Leung; submitted 21.Apr.2025; peer-reviewed by K Mouloudj, M Al-Mujtaba; revised version received 20.Oct.2025; accepted 30.Nov.2025; published 21.Jan.2026.

Please cite as:
Hodges J, Cohn W, Castel A, Flickinger T, Waldman AL, Hilgart M, Kirby O, Caldwell S, Ingersoll K
A Complex Digital Health Intervention to Support People With HIV: Organizational Readiness Survey Study and Preimplementation Planning for a Hybrid Effectiveness-Implementation Study
J Med Internet Res 2026;28:e76327
URL: https://www.jmir.org/2026/1/e76327
doi:10.2196/76327

© Jacqueline Hodges, Wendy Cohn, Amanda Castel, Tabor Flickinger, Ava Lena Waldman, Michelle Hilgart, Olivia Kirby, Sylvia Caldwell, Karen Ingersoll. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Using the NASSS–Complexity Assessment Tool to Evaluate the Implementation of “Cadê O Kauê?”: Chat-Story Intervention for Youth Participation in Mental Health Promotion in Brazil

Sakshi Setia[1*], MBBS, MSc; Gabriela Pavarini[2*], MPhil, DPhil; Mona Koshkouei[1], MPharm, DPhil; Sheila Giardini Murta[3], MSc, PhD

[1]Nuffield Department of Primary Care Health Sciences, University of Oxford, Radcliffe Primary Care Building, Radcliffe Observatory Quarter, Woodstock Rd, Oxford, United Kingdom

[2]Department of Social Policy and Intervention, University of Oxford, 32 Wellington Square, Oxford, United Kingdom

[3]Department of Clinical Pyschology, Universidade de Brasília, Brasilia, Brazil

[*]these authors contributed equally

Corresponding Author:
Gabriela Pavarini, MPhil, DPhil
Department of Social Policy and Intervention, University of Oxford, 32 Wellington Square, Oxford, United Kingdom

Abstract

Background: The global public health challenge of young people’s mental health is particularly evident in Brazil, which is considered the most anxious and the fifth most depressed country globally. In response to this, a digital intervention, that takes the form of a storytelling chatbot named “Cadê o Kauê?” (translation: “Where is Kauê?”), was coproduced by an interdisciplinary team of researchers, young people, and narrative designers as part of a project titled Engajadamente. The intervention was designed to strengthen Brazilian adolescents’ skills in promoting their peers’ mental health through direct support and collective action for mental health. The intervention was rolled out in select schools in Brazil and online via social media, but it encountered challenges during the initial implementation phase in school settings.

Objective: This project’s primary objective is to evaluate the implementation of “Cadê o Kauê” in school settings in Brazil, assess current and potential complexities, and formulate recommendations to reduce or manage these complexities. The secondary objective of this project is to understand the current and potential sociotechnical change resulting from the implementation of the “Cadê o Kauê” in a real-world setting.

Methods: The Non-adoption, Abandonment, Scale-up, Spread, Sustainability (NASSS) framework was selected as the theoretical framework to identify different areas of complexity across domains and further thinking of ways to reduce or manage these complexities. Data were gathered by conducting semistructured online interviews with the Engajadamente team members based in the United Kingdom and Brazil. The NASSS–Complexity Assessment Tool was used as a sensemaking device to facilitate a collective understanding of the implementation, and narratives were mapped onto specific NASSS domains. Key themes were identified across overarching domains using thematic analysis to meet the objectives of this evaluation.

Results: The results mapped onto the NASSS domains generated a narrative about the initial implementation of “Cadê o Kauê?” in schools. Existing and potential complexities across technology, organization, and wider system domains were characterized by interdependencies, unintended consequences, and uncertainties. The implementation of “Cadê o Kauê?” highlighted challenges related to internet dependency, infrastructure limitations, and varying levels of organizational readiness, teacher preparedness, and school culture. Addressing these through offline solutions, additional teacher training, whole-school engagement strategies, and a value-informed complexity approach can enhance accessibility, alignment with school priorities, and adaptability within Brazil’s dynamic sociopolitical context.

Conclusions: Using the NASSS framework, the evaluation captured both current and emerging complexities in implementing “Cadê o Kauê?” in schools. By engaging with theory-informed approaches to technology implementation projects and deep-diving into discursive dynamic interactions between technology and its “context” of implementation, we have explored ways to manage these complexities and developed suitable recommendations to guide the Engadajamente team and support similar projects in the future.

(J Med Internet Res 2026;28:e79106) doi:10.2196/79106

KEYWORDS

digital mental health; implementation science; mental health; knowledge translation; youth participation

XSL•FO
RenderX

Introduction

Mental ill health represents the leading cause of disability among individuals aged 10 - 24 years, accounting for 45% of the global burden of disease [1,2], where half of mental health disorders begin by the age of 14 years [3]. The global public health challenge of young people's mental health is particularly evident in Brazil, which is considered the most anxious and fifth most depressed country globally [4]. The COVID-19 pandemic severely impacted Brazil; around 1 in 3 young people exhibited symptoms of depression or anxiety [5].

In low- and middle-income countries (LMICs), sociocultural factors, including religion, cultural values, economic inequality, and political instability, disproportionately impact mental health outcomes [2,6,7]. For instance, Brazilian adolescents from lower socioeconomic backgrounds face greater mental health challenges and poor access to care [8]. Determinants such as stigma surrounding mental health also affect care-seeking and mental health service usage across the country [9]. Novel solutions are required to address these sociocultural determinants for promoting young people's mental health in Brazil and other LMICs [10].

Since the United Nations Conventions on Rights of the Child (1989), researchers, practitioners, and policymakers have increasingly acknowledged young people's potential to be active agents of social change capable of promoting health in their communities through participation in advisory groups, representative councils, or activism [11]. Grounded in the Ottawa Charter's definition of health promotion (1986), mental health promotion reflects a view of health as a positive concept encompassing social and personal resources, as well as physical capacities. Accordingly, youth participation in promoting mental health within the community involves engaging in social interactions and collective action to address social determinants and to build community capacities and competencies for well-being, rather than focusing solely on illness and its risk factors [11].

Since adolescents spend a significant amount of their time in school, schools have become natural environments for promoting positive mental health outcomes [12-14]. School initiatives focusing on mental health promotion have strengthened students' preparedness to offer peer support and a sense of belonging within the school community [15].

Over the past decade, digital interventions, such as mobile apps, have been recognized as essential tools for addressing young people's mental health needs by overcoming traditional barriers and expanding access to care, particularly in underserved communities [16]. In parallel, co-designing these interventions with young people provides scalable and culturally relevant solutions to support their mental health [17]. However, most co-designed digital mental health interventions have been developed and evaluated in high-income regions, including Australasia, North America, and Europe [18]. This highlights an important gap, as there is a pressing need for contextually relevant digital mental health resources in LMICs, where health care access remains limited, and stigma surrounding mental health persists [19,20].

"Cadê o Kauê?," is one such digital mental health intervention, which takes the form of a storytelling chatbot, developed to enhance young people's participation in promoting mental health and well-being within their peer communities in Brazil [21]. This approach reflects a potentially scalable intervention model that builds on peer networks, strengthens community connections, and encourages youth engagement in collective action and social advocacy [22]. It was co-designed by *Engajadamente's* [22] interdisciplinary team of researchers, adolescents, and narrative designers to strengthen Brazilian adolescents' skills in promoting their peers' mental health through direct support and collective action for mental health [23]. The development also involved input from teachers, psychologists, and other high school professionals through advisory committees [24]. The detailed auto-ethnography of the pluralistic coproduction model [21,23] is summarized in another paper.

In this digital narrative, players interact with fictional, automated characters (bots) via SMS text messaging and learn that their friend is struggling with his mental health. As the narrative progresses, they "unlock" skills (eg, signposting and establishing partnerships) that enable them to support him and others [24]. The chat-story was initially envisioned for remote, stand-alone use among adolescents aged 15 - 18 years. However, following the mapping stage of the co-design process, the tool was tailored for use within schools to create a more comprehensive pathway for change, with teachers responsible for facilitating the introduction of "Cadê o Kauê?" in classrooms with students [21]. A feasibility trial accompanied by a qualitative acceptability study of "Cadê o Kauê?" was conducted in 7 schools [21]. Several unanticipated challenges emerged during the initial implementation in school settings, indicating the necessity of further real-world evaluation of "Cadê o Kauê?"

Implementing large-scale technology-supported projects does not simply involve introducing new technology or disrupting old practices but also navigating an intricate web of interconnected factors within those systems [25]. This requires a mindset grounded in complexity thinking, recognizing these systems' dynamic and unpredictable nature [26]. Complexity, as defined by Cohn et al [27], is a "dynamic and constantly emerging set of processes that not only interact with each other but come to be determined by those interactions." Effective implementation, therefore, necessitates viewing technological interventions as part of a larger social system, where success depends on embracing the system's inherent complexity [26]. Complex adaptive systems have fuzzy boundaries that allow for dynamic interactions between components of the system, leading to emergent behaviors. Undertaking sociotechnical evaluations of complex technology projects offers an opportunity to understand the changing and evolving interrelationships between technology and human or organizational (or socio-) factors [28]. A sociotechnical perspective assumes that "organizational and human (socio) factors and technology's (technical) materiality are interrelated parts of one system, recursively (not simultaneously) shaping each other" [29,30].

Previous research has evaluated the feasibility and acceptability of "Cadê o Kauê?" through user surveys as part of the co-design process [21]. Survey findings indicated that the intervention

XSL•FO
RenderX

enhanced adolescents' motivation and self-efficacy to engage in peer support and collective action related to mental health, as well as increasing their openness in discussing mental health issues [21].

However, to understand "how" "Cadê o Kauê?" functions within the broader social and institutional context of implementation, this qualitative study uses a complexity science approach for evaluation of this intervention. The primary aim of this study is to evaluate the implementation of "Cadê o Kauê?" to identify existing and potential complexities and develop recommendations to address or mitigate these challenges. The secondary aim is to explore potential sociotechnical changes associated with implementing "Cadê o Kauê?" in a real-world context.

Methods

Theoretical Framework

The Non-adoption, Abandonment, and Challenges to Scale-up, Spread, and Sustainability (NASSS) framework is selected as the theoretical framework to gain insights into the multifaceted and interdependent components of the initial implementation of "Cadê o Kauê?." This framework was developed against the backdrop of complex adaptive systems, synthesizing various theories to map possible areas of complexity when planning, evaluating, and synthesizing a technology initiative in health and social care [25]. NASSS framework (Figure 1) consists of 7 domains—the condition or illness, the technology, the value proposition, the adopter system (intended users), the organizations, and the wider system (mainly regulatory, legal, and policy issues); the seventh, cross-cutting, domain considers how all these interact and emerge over time [24].

Figure 1. The NASSS framework (reproduced from Greenhalgh et al [31]) published under the Creative Commons Attribution 4.0 International License [32]. NASSS: Non-adoption, Abandonment, Scale-up, Spread, Sustainability.

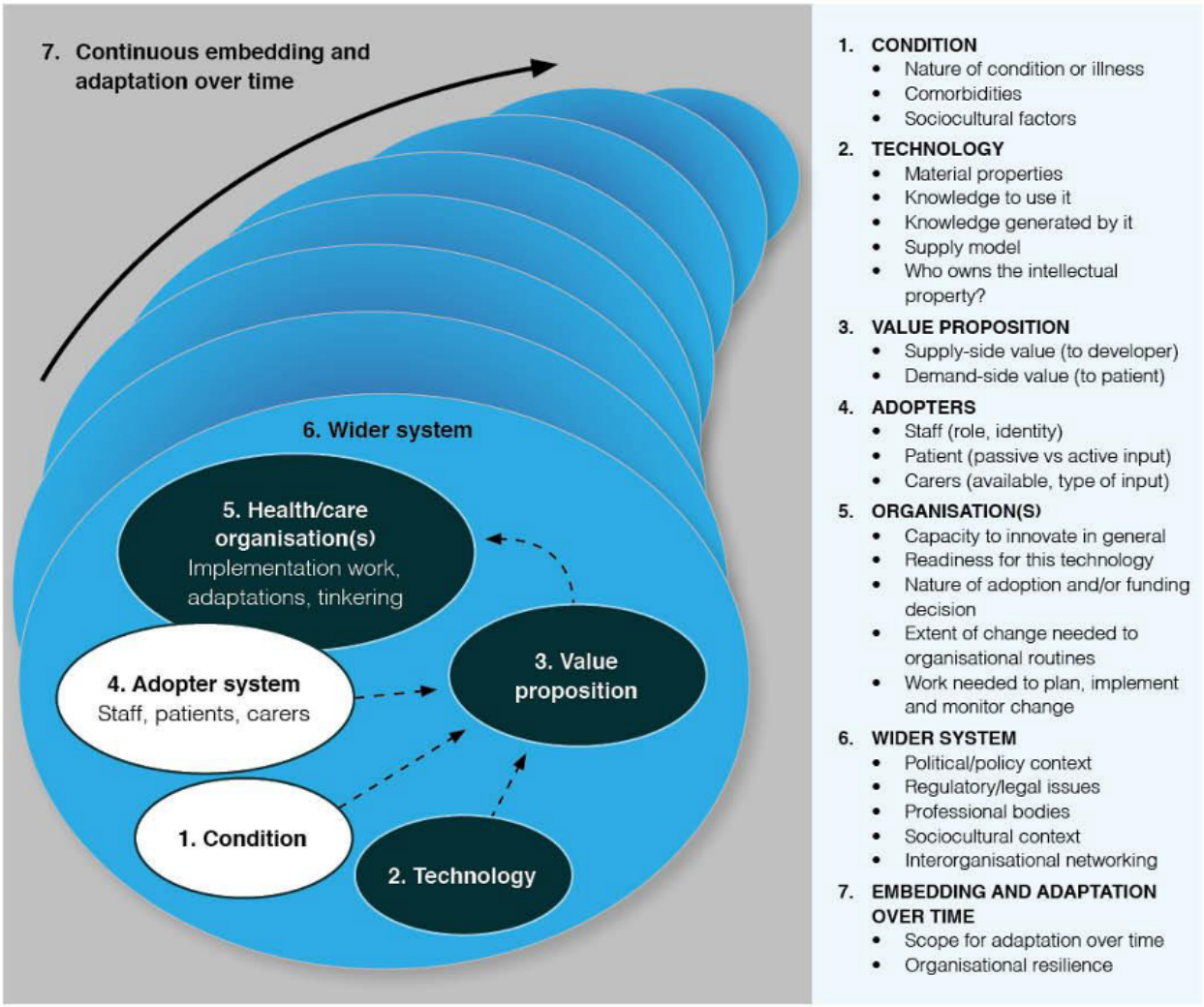

NASSS is used for this study on the account of the following: first, "Cadê o Kauê?" was introduced into a complex landscape with overlapping system boundaries involving young people, teachers, school administration, and wider policy structures. NASSS is well-suited to address cultural, infrastructural, and policy-specific complexities by addressing interactions between the technology and the implementing organization and the wider social system [25]. Second, the framework supports the use of retrospective exploration of "how" and "why" "Cadê o Kauê?" functioned as it did during early implementation, rather than

XSL•FO
RenderX

focusing solely on whether the intervention was effective. This is particularly important for identifying emergent interactions, contextual influences, and unintended consequences that were not fully visible at the outset of implementation. Third, it has been previously used to evaluate digital mental health interventions, including those in LMICs [33].

Participants and Procedure

With the help of a senior researcher (GP), we identified a purposive sample of stakeholders with varied roles and involvement in the co-design and implementation of the intervention to capture their diverse perspectives. This sample included core co-design team members, including researchers, youth collaborators, technology and policy experts, and implementation facilitators, including 2 teachers from schools that tested out the intervention (Table 1). Stakeholders from the United Kingdom and Brazil (Table 1) were invited via personalized emails, after which they reviewed a participant information sheet and gave informed consent through Microsoft Forms. Participant Information Sheets and consent forms were provided in English and Portuguese, depending on participants' preferences.

Table . Summary of study participants' characteristics.

Participant ID	Location of the participant	Role	Mode of data collection	Language	Role in Engajadamente project
Participant ID 1	Brazil	Youth collaborator	Synchronous	English	Member of core co-design team
Participant ID 2	United Kingdom	Researcher	Synchronous	English	Member of core co-design team
Participant ID 3	United Kingdom	Researcher	Synchronous	English	Member of core co-design team
Participant ID 4	Brazil	Researcher	Synchronous	English	Member of core co-design team
Participant ID 5	Brazil	Researcher	Synchronous	English	Member of core co-design team
Participant ID 6	Brazil	Policy expert	Synchronous	Portuguese	Member of advisory committee advising on policy aspects
Participant ID 7	Brazil	Tech expert	Synchronous	Portuguese	Member of core co-design team
Participant ID 8	Brazil	Teacher	Asynchronous	Portuguese	Testing in schools
Participant ID 9	Brazil	Teacher	Asynchronous	Portuguese	Testing in schools
Participant ID 10	Brazil	Youth collaborator	Asynchronous	Portuguese	Member of core team for co-design process

Data Collection

Data collection took place in 2024 virtually with support from Microsoft Teams. Interviews were conducted virtually due to the geographical dispersion of participants across different countries and time zones, as well as their varied professional commitments. Online interviews provided a flexible and feasible approach that enabled participation without the need for travel, while still allowing for in-depth discussion. Virtual platforms also supported secure audio recording and note-taking, and did not compromise the quality of interaction. In total, 7 interviews took place synchronously and lasted for an average of 56 minutes (43 - 60 min). Furthermore, 3 interviews (2 teachers and 1 youth collaborator) were conducted asynchronously, with the questions answered through audio messages. This alternate strategy was used to facilitate participation, as teachers had limited time to participate in the study and were, therefore, difficult to recruit.

A semistructured interview guide (Multimedia Appendix 1) was designed for online interviews using the prompts from the theoretical NASSS framework, covering 7 domains—condition, technology, value proposition, adopters, organization, wider system, and emergence over time. For online interviews, questions were divided into three sections, that are (1) participant's background, Brazilian context, and technology; (2) questions about the implementation process in schools, including the success factors and challenges faced; and (3) reflections and future recommendations. Questions were adapted according to participants' background and their role in the implementation. For example, the interview guide was adapted to cover the technical and material components for the tech-expert, aligning with the prompts from the technology domain of the NASSS framework. For teachers, questionnaires were adapted to fit the relevant domains (organization and adopter system) of the framework. Similarly, for the policy expert to understand the broader context, the guide included detailed follow-up questions on educational and social policies on youth mental health in Brazil. To mitigate recall bias, as interviews were conducted retrospectively, participants were prompted to anchor their reflections to specific implementation milestones, and accounts were interpreted alongside other participant perspectives and available project documentation.

XSL•FO
RenderX

Out of 7 synchronous interviews, 5 were conducted in English, and 2 of them were conducted in Portuguese. All 3 asynchronous interviews were conducted in Portuguese. Interviews were conducted by SS, and GP helped facilitate the conversations for non-English speaking interviewees. The interviews were transcribed using Microsoft Teams transcription software. The transcriptions in Portuguese were translated into English thereafter and checked for accuracy with the help of a native Portuguese-speaking researcher.

To support participant well-being during interviews, questions focused on their experiential learnings of the implementation of the tool, which posed minimal risk. Interviews were conducted by SS, who had previous experience in facilitating qualitative interviews requiring sensitive communication. Participants were reminded that they may pause, decline to answer any question, or stop the interview at any time. All data were handled confidentially, with identifying details removed during transcription and reporting. These measures ensured that risks remain minimal while respecting the sensitive context of youth mental health–related work.

Ethical Considerations

The Oxford Tropical Research Ethics Committee reviewed the study's protocol. It was determined that the project did not require further ethical approval in the United Kingdom as it qualifies as a service evaluation project. In Brazil, local ethics approval was sought and granted by the University of Brasilia Humanities and Social Sciences Ethics Committee (reference 6.980.535).

This study focuses on professionals' and team members' experiences of implementation rather than their personal mental health; however, the authors acknowledge the broader sensitivity of youth mental health work and co-design contexts. Several safeguards were therefore put together in place. Informed consent was obtained before participation via a Microsoft Form, supported by a detailed participant information sheet (Multimedia Appendix 2). This clearly explained the study purpose, the voluntary nature of involvement and participation, the right to withdraw at any time, and how data will be stored, anonymized, and used. Contact details for the researcher and the University's ethics or information compliance teams were also provided.

Data Availability

The University of Oxford is the data controller with respect to participants' personal data and determined how those data were used. The University processed personal data for the purpose of the evaluation outlined above. Interview recordings were stored securely on the University drives, with access restricted to the study team, and were deleted after data analysis was complete. Consent records were also stored securely on the University drives with restricted access and will be retained for at least 3 years following final publication of the work. Following transcription, any identifying information relating to participants or their organizations was removed, and transcripts were labeled using random IDs. These transcripts may be uploaded to data repositories or shared with other researchers via OneDrive, where appropriate, and were retained for at least 3 years. A separate file linking participant IDs to names and contact details was maintained in case recontact was required and was deleted after data analysis was completed.

Data Analysis

Data analysis was led by the first author (SS), at the time of the study, an MSc researcher in Translational Health Sciences with experience in qualitative research. Qualitative data from interview transcripts were analyzed using the approach to thematic analysis by Braun and Clarke [34]. Data analysis was conducted simultaneously with data collection, allowing for themes identified during data analysis to inform data collection. In line with this approach, analysis was recursive rather than strictly linear, with ongoing engagement with emerging patterns throughout the period of data collection. Early analytic insights informed areas of attention in subsequent interviews [34]. To understand "the depth and breadth of data" [35], SS familiarized herself with the data by reading and rereading the transcripts multiple times. As a nonnative Portuguese speaker, SS remained attentive to potential linguistic and cultural nuances during interpretation. This was addressed through consultation with bilingual members of the research team to ensure conceptual clarity and contextual accuracy in the analysis. Her year-long engagement with the *Engajadamente* [22] team also enriched her awareness of the pertinent sociopolitical factors surrounding the project. Furthermore, initial interpretations of the data were discussed in meetings with GP and MK, both senior researchers. Personal biases and assumptions that may have been held while analyzing the content were recognized and discussed during those meetings. These sense-checking discussions supported contextual validity while maintaining analytic independence. After the first 5 interviews, the transcripts were imported into NVivo software (Lumivero) [36] to support the data analysis.

Following data familiarization, a coding framework was developed using NASSS as a sensitizing device while keeping the initial evaluation objectives in mind. However, this alone did not define the scope of the findings. Subthemes were identified using NASSS domains as overarching main themes. The coding and subtheme generation process evolved throughout the data analysis process; this process was refined iteratively, with modifications made to the coding framework to capture newly arising concepts and subthemes. After 8 interviews, no substantially new insights were generated across NASSS domains and 2 additional interviews supported the development of a sufficiently rich and coherent set of themes. Subthemes were generated not solely based on the number of codes they contained but also on their significance and how well they related to the evaluation objectives. Earlier transcripts were checked again after data collection was complete to ascertain if new codes could be applied to these transcripts. Any codes that did not fit the overall analysis were checked to see if they could be merged with another code. In the second cycle of analysis, codes were organized into subthemes under the 7 broader domains of the NASSS framework (refer to Results section).

XSL•FO
RenderX

Results

Mapping Complexity Onto NASSS Domains

Thematic analysis of interview transcripts was mapped onto the NASSS domains (Table 2). Within each domain, subthemes were identified and illustrated with participant quotes (referenced by participant ID). These subthemes were further interpreted to generate insights, which then informed a detailed examination of complexity within the relevant domains, as presented in the subsequent section.

XSL•FO
RenderX

Table . Summary of NASSS[a] domains with corresponding subthemes and supporting participant quotes.

NASSS domain and subtheme	Example data	Insights generated
Condition		
Capturing diverse perspectives of young people's mental health across Brazil	• *"The tool needed to be culturally relevant and sensitive to the diverse realities of Brazilian adolescents. A poor adolescent may have a different set of responsibilities, a Black adolescent faces others, and an adolescent from a middle-class background will have different ones. However, they are all united by the shared adolescent experience."* [Participant ID 10]	Brazil's youth mental health landscape is influenced by several sociocultural determinants however the Engajadamente team managed to incorporate plurality of experiences into the design of "Cadê o Kauê?"
Challenges in incorporating the voices of Indigenous populations	• *"We wanted to include the experiences of Indigenous people in Brazil into the design of "Cadê o Kauê?", but were mindful of the need to have appropriate resources and support in places; especially those in remote or rural areas."* [Participant ID 3]	Additional resources are required to incorporate the perspectives of Indigenous populations
Technology		
Technical integration and platform dependencies	• *"A significant challenge was our reliance on platform updates from services like Messenger and Telegram, as any changes to their code affected our development."* [Participant ID 7]	Technology has complex technical integrations and is dependent on multiple platforms for functioning
Technology's functionality is dependent on the internet	• *"We provided 3G routers to make up for inconsistent internet access, but the signal needed to be improved, especially when large groups of students were using it simultaneously."* [Participant ID 2]	Technology's smooth functioning is dependent on stable internet connectivity
Knowledge generated after introduction of technology	• *"There was also some resistance from teachers who were concerned about students accessing social media."* [Participant ID 5]	Some teachers expressed resistance especially those who held negative beliefs about the use of social media in schools
Knowledge generated after introduction of technology	• *"We observed that young people place significant value on internet access, especially when students do not have it but want to check social media on their phones."* [Participant ID 2]	This dependency highlighted the importance of internet access and the growing use of social media among young people
Innovative nature of technology	• *"We decided to use the Hero's Journey framework, which is commonly used in novel writing, for structuring the story. Since our target audience is young people, we incorporated gamification elements to make the story engaging and relevant."* [Participant ID 3]	The chatbot incorporated an innovative and interactive design, using storytelling elements grounded in principles of positive psychology
Value proposition		
Added value to the school system	• *"The primary objective is to spark dialog and empower students to identify and address the mental health needs within their school."* [Participant ID 4]	"Cadê o Kauê?" can act as a catalyst in initiating meaningful dialogue surrounding mental health and well-being within the school community

XSL•FO
RenderX

NASSS domain and subtheme	Example data	Insights generated
Added value to the school system	• *"The tool facilitated important conversations about mental health in schools and among adolescents and teachers, breaking down stigmas and enabling a more open dialog. The project encouraged male students to talk about their feelings and interpersonal relationships and provide support to classmates who cried during classes.*" [Participant ID 9]	By fostering conversations about mental health, the intervention engaged adolescents, teachers, and parents, contributing to stigma reduction and promoting values, such as empathy and social responsibility
Added value to the school system	• *"The administration, coordinators, and other teachers viewed this project as a valuable addition given the limited support for students experiencing symptoms of anxiety and depression.*" [Participant ID 8]	Within the school system, the intervention was perceived as adding value, especially after pandemic, when schools struggled to respond adequately to the behavioral and mental health needs of students
Improving access to universal mental health services	• "*The toolkit informs students about free, state-provided resources beyond the support available in schools or from parents, including psychologists and social workers who should be present in Brazilian schools. The intervention encourages students to reach out to these professionals*." [Participant ID 1]	The intervention provides young people with a toolkit highlighting available mental health support resources, mainly through Brazil's universal health system, that offers mental health services for youth in most states
High costs of maintaining the tool	• *"Maintaining a presence on Instagram is very expensive because you have to pay for Meta. If the tool were to be standalone, the maintenance cost is low. However, if we want to integrate it into popular platforms like Instagram and Facebook where young people are active, the cost increases significantly."* [Participant ID 4]	The sustainment of "Cadê o Kauê?" depends on external funding, which is contingent on demonstrating value on the demand side by delivering benefits to the direct users (young people and their peers) and the broader system, including schools, policymakers, and the healthcare system
Adopter system		
High acceptability	• *"Comments like 'It felt close to my life,' 'It represented my experiences,' and 'I love that character,' 'that character is me' were particularly touching.*" [Participant ID 2]	The representation in the tool was particularly appreciated, students felt a personal connection to the tool, stating how its characters and scenarios reflected their own experiences
Co-design enhanced user engagement	• *"The pluralistic co-production model went beyond mere participation, as we shared power and responsibilities among the senior members of the team, this immersive and iterative approach helped understand and adapt "Cadê o Kauê?" to the sociocultural specificities of the population,*" [Participant ID 1]	The co-design process was attributed as a success factor for the high acceptability of "Cadê o Kauê?" among young people .
Organization		

XSL•FO
RenderX

NASSS domain and subtheme	Example data	Insights generated
Assessing organization-innovation fit	• *"Expanded coordination with school managers is very important as the leadership guides the sustainment of these opportunities. The administrators must see value in prioritizing these projects; otherwise, they will not put all their efforts into making that project happen."* [Participant ID 6]	Involvement of school administrators in the co-production process is required to deliver a streamlined intervention in schools
Teachers' preparedness and need for additional training	• *"Teachers expressed concerns about managing difficult conversations when students raised mental health issues."* [Participant ID 3] • *"In a few schools, students felt that teachers needed more time to be ready to handle the discussion circles."* [Participant ID 5]	Teachers highlighted the need for additional training to facilitate postintervention discussion circles and address student mental health concerns
Shifting organizational routines	• *"Ideally, incorporating the tool and conversation circle into the curriculum should mean no extra work for teachers, as the lesson would already be prepared, however during implementation, the whole process took longer than anticipated."* [Participant ID 2]	The intervention was designed to be used in a classroom setting in high school and expected to last an hour and a half, which corresponded to approximately 2 class periods
Wider system		
Uncertainty in sociopolitical environment	• *"Brazil faced significant challenges over the last few years, including economic decline, political instability, and a major shift towards extreme right-wing policies. The political climate created uncertainty around government acceptance of educational interventions, especially concerning adolescent mental health."* [Participant ID 5]	The project's values, which focused on ensuring equitable access to quality health care through community and school involvement, conflicted with the existing government's policies
Emergence over time		
Need for ongoing adaptation	• *"There is uncertainty regarding the technology's adaptation over time, as what is considered 'relevant' today may require 'reinvention' tomorrow."* [Participant ID 6]	"Cadê o Kauê?' needs to adapt to shifting technology policies, educational practices and funding priorities

[a]NASSS: Non-adoption, Abandonment, Scale-up, Spread, and Sustainability.

Complexity Assessment

As illustrated in Table 2, the identified subthemes were mapped onto the NASSS domains. The findings were subsequently applied to the Complexity Assessment Tool, which revealed significant areas of complexity within the domains of technology, organization, and wider system. These were characterized by uncertainties, interdependencies, and unintended consequences (Table 3). Instead of categorizing NASSS domains as simple, complicated, or complex, the complexity assessment focused on examining how different forms of complexity unfolded during the implementation. This approach also acknowledges the dynamic nature of complexity, recognizing that a domain that is considered simple or complicated at one stage may evolve into a complex one over time, and vice versa.

XSL•FO
RenderX

Table . Table of complexities.

NASSS[a] domain	Interdependencies	Unintended consequences	Uncertainties
Technology	Complex technical integrations require compatibility across multiple platforms. Smooth functioning is dependent on stable internet access and adequate infrastructure.	Provision of 3G routers during initial implementation enabled access, but students shared passwords for social media use, causing system crashes.	Teachers' knowledge, beliefs, and attitudes towards technology still remain uncertain.
Organization	Adoption is contingent on organizational culture, readiness to innovate, and supportive leadership.	Implementation disrupted established routines, intervention delivery took longer than anticipated and teachers required additional training to both use the intervention and lead discussion circles effectively.	Uncertainty remains around the conditions required for routinization, as well as *who* should be responsible for introducing and mainstreaming the technology.
Wider System	Sustainability depends on political support, policy alignment, and system-level backing for spread and scale-up.	Polarized political climate and introduction of a new high-school system created disorganization in the initial implementation.	Fragmented technology policy landscape raises uncertainty regarding long-term sociopolitical support for sustaining the intervention.

[a]NASSS: Non-adoption, Abandonment, Scale-up, Spread, and Sustainability.

Complexity in Technology

The initial implementation of "Cadê o Kauê?" revealed complexity in its current state of materiality, as its smooth functioning depends on internet connectivity. While the selection criteria for the pilot implementation included schools with access to computers, some schools either lacked internet connectivity or faced frequent disruptions.

Using "Cadê o Kauê?" through the website also caused stability issues due to its dependency on the internet, particularly when the internet signal was interrupted, resulting in the loss of interaction history.

> *If you are using the intervention through the website and the signal is cut, you would lose the history of your interaction, so you would have to start all over again. This made it difficult for students to play the whole game without losing interest.* [Participant ID 3]

A teacher from a rural school noted,

> *Due to the school's rural location, the internet signal was weak, and the students preferred to access the tool on their mobile phones.* [Participant ID 8]

This dependence also resulted in unintended consequences, such as in one school, where students shared router passwords, causing network crashes, thereby disrupting the effectiveness of the intervention.

> *At one school, we provided a router, and the students ended up sharing the password with other students to check social media on their phones. Soon, the entire school started using the router, which was meant for 40 and not 700, leading to a system crash.* [Participant ID 1]

Participants observed that these dynamics not only exposed structural challenges related to technology but also generated some knowledge about the digital behaviors of young people.

> *We discovered the value of the internet among young people, especially when students do not have it and want to check their social media apps on their phones.* [Participant ID 2]

In some schools, some teachers were resistant to the use of technology via social media platforms. Their concerns primarily stemmed from negative perceptions about social media and worries that it would distract students from their learning. However, a participant noted that collaboration and the process of coproduction can help overcome these fears.

> *There was an initial hesitation among teachers to join the project; however, through the process of co-production and forming a teacher advisory committee, some teachers eventually became strong advocates and ran campaigns in their schools to promote the intervention.* [Participant ID 6]

There was a shared recognition among participants that intervening offline would reduce this dependency on the internet.

> *From this experience, we concluded that an offline intervention would be more reliable given the current infrastructure. An offline version, similar to older software games, would work better than relying on an online connection.* [Participant ID 2]

Complexity in Organization

In this context, the organization refers to schools, and their readiness to adopt the technology was influenced by several factors. Key factors that emerged during the initial implementation included the schools' capacity for innovation, ability to absorb new knowledge, overall implementation climate, and the strength of leadership.

"Cadê o Kauê?" was piloted in 7 state schools in Brasilia, and typically, the school committees recommended a teacher from the *Life Project* course to introduce the intervention within classrooms. The *Life Project* is a course proposed in the new high school curriculum in Brazil, designed to promote students' self-awareness and autonomy, as well as a purposeful approach to life. This course aligned well with the objectives of the Engadajamente project. "Cadê o Kauê?" was designed to fit within regular class hours without adding extra work for

XSL•FO
RenderX

teachers; however, in practice, the whole implementation process took 3-4 hours.

> *Ideally, incorporating the tool and discussion circle into curriculum… should mean no extra work for teachers… but the whole process took longer than anticipated and this disrupted the normal class routines.* [Participant ID 5]

Apart from technical difficulties, teachers attributed adoption uncertainties to transitional and communication issues associated with the recently introduced high-school system. They observed that the intervention might be more effective after the school's curriculum was fully restructured.

> *I believe the project could be carried out again now that the school curriculum is more organised and we have adapted to the new high school system, even though further changes are expected. Things are less fragmented within the school now for better adoption of the intervention.* [Participant ID 9]

Teachers also emphasized the need for broader engagement to address the structural challenges, especially from the school's leadership.

> *Expanded coordination with school managers is very important…the administrators must see value in prioritizing these projects; otherwise, they will not put all their efforts into this project's implementation.* [Participant ID 6]

The intervention was also tested in a military school setting, which posed unique challenges due to its authoritarian structure and rigid leadership, contrasting sharply with the participatory environment typically required for youth mental health initiatives. However, the teacher representing this school challenged these constraints by implementing "Cadê o Kauê?" in her classroom.

Participants also highlighted teachers' requests for additional training and the schools' need for mental health resources. Teachers acknowledged that they often encounter mental health concerns among students, but their training typically lacks mental health literacy on handling such issues.

> *This request for further training is not a minor adjustment; it suggests a need for more comprehensive training.* [Participant ID 2]

> *It is also important to provide more training and didactic support to teachers, as they can take valuable lessons from this project on how to handle difficult situations in classrooms concerning student mental health.* [Participant ID 10]

Complexity in External or Wider System

The co-design and implementation phase of "Cadê o Kauê?" were affected by the challenging sociopolitical climate, the introduction of a new high-school system, and fragmented social policies in Brazil. The project's values, which focused on community participatory approaches in the school involvement, conflicted with the government's policies. The ruling government did not prioritize the rights of vulnerable populations, including the mental health care needs of young people.

> *Brazil faced significant challenges over the last few years, including economic decline, political instability, and a major shift towards extreme right-wing policies. The political climate created uncertainty around acceptance of educational interventions concerning adolescent mental health.* [Participant ID 5]

During the initial implementation of "Cadê o Kauê?," Brazil's educational system also underwent significant restructuring with the introduction of a new high-school curriculum. This reform coincided with rising school dropout rates, sparking criticism that the reform was contributing to the decline in quality of Brazil's education system.

Despite these challenges, the Engadajamente team navigated this complexity in the sociopolitical landscape. They were cautious about the design of the chat-story as a youth collaborator noted,

> *We did not omit any mental health topics, but we did have to cut certain content due to political concerns. I suggested to tone down discussions on LGBTQ+ issues. A previous incident influenced this decision, in 2013, when an infographic addressing homophobia caused significant controversy and was not accepted in schools. I recommended taking a vigilant approach to avoid risking our access to schools* [Participant ID 1]

Discussion

Principal Findings

Using the NASSS framework as a sensemaking device, this study captured diverse perspectives of stakeholders to generate a narrative of the initial implementation of "Cadê o Kauê?" in school settings in Brazil. By mapping the complexities across the NASSS domains and conducting a complexity assessment, we shed light on the dynamic relationships between technology and the broader social system around it. Building on these insights, we propose recommendations for addressing current and emerging complexities that may evolve over time.

Managing Complexity in Technology

Access to the internet and technological infrastructure are prerequisites for most eHealth innovations [37], and these requirements could limit technology's adaptability in resource-constrained settings [38]. Using a sociotechnical lens, 2 challenges are identified that may emerge over time due to this reliance on the internet. First, building on the views of Meyer et al [38], that while end users may have an overall positive view of "Cadê o Kauê?" the trust in its overall use may fade over time if users experience more than occasional failures and frustrations. Second, it could hinder the scale-up of intervention in schools with poor internet access and/or appropriate infrastructure. In turn, this dependency may widen existing inequalities, as schools without adequate infrastructure may be excluded from the benefits of this intervention. This

XSL•FO
RenderX

digital divide is a concern among bioethicists, as digital mental health interventions may exacerbate existing inequities [35].

The first step to address complexities in this domain is to move away from viewing technology as a "plug-and-play" solution. To reduce the reliance of technology on the internet, one potential solution identified is to develop an offline version of the intervention; this approach could help maintain user engagement while also addressing the issues of equity. However, dependency on computers remains a concern if some are nonfunctioning. This factor largely depends on the organization's capacity and willingness to support the innovation.

Knowledge, perceptions, and attitudes about the technology also shape its acceptance [37], so it is essential to thoroughly understand these dynamics early and integrate strategies to address them in project planning. In some schools, teachers were initially resistant to using the technology, particularly because it was delivered through social media platforms. Their concerns largely stemmed from negative perceptions of social media and fears that it could distract students from learning. However, these reservations were gradually addressed through collaborative approaches. Participation in coproduction processes and teacher advisory committees helped build trust. Over time, some teachers became strong advocates who actively promoted the intervention within their schools.

Maximizing Value

The Engadajamente team members envision a future for "Cadê o Kauê?" where the intervention can be scaled nationally and embedded in the school curricula. However, lack of sustainable funding is a potential barrier to scale-up. Making a plausible business case for "Cadê o Kauê?" to generate long-term and alternative funding streams is challenging, owing to uncertainties in the redistribution costs over time [39].

The value of technology in a health policy context lies in the nature and scale of the benefits it generates [36]. To navigate the complex external environment, value-generation strategies are proposed to reframe the intervention's benefits in line with policy priorities, enhancing its appeal to alternative investors and government stakeholders.

First, the intervention can be repositioned as a strategy to improve adolescents' access to Brazil's public mental health services. Although understanding care-seeking pathways to mental health care services is complex, nonetheless, there is a substantial unmet need in Brazil, with only 20% of children and adolescents using public mental health services [40]. Adolescents consulting primary care centers reported solving mental health problems generally on their own or with peers [41]. One component of "Cadê o Kauê?" directly addresses this gap by providing an information toolkit that maps free mental health services available through Brazil's Unified Health System (Sistema Único de Saúde) across different states. This approach raises adolescents' awareness of their right to access these services and encourages them to seek professional care for their mental health issues. This rights-based approach aligns with the principles of Brazil's community mental health network (RAPS) [42]. A youth collaborator in the Engadajamente project with lived experience of mental health issues reflects how, as a teenager, they were unaware of such services and how having that information would have made a significant difference for them. Additionally, using schools as a platform to recognize early signs of mental distress, facilitate early detection, and refer cases to mental health care has been highlighted as a promising approach to improving care-seeking pathways [43]. Integrating "Cadê o Kauê?" into the school environment positions it as a bridge linking adolescents to existing public health infrastructure. This process aligns the intervention with national policies' planned efforts to strengthen adolescent mental health care within Sistema Único de Saúde.

Recognizing the Organization's Readiness

Organizational readiness for change is considered a key factor in determining whether evidence-supported interventions are successfully adopted in a system [44]. Contextual factors, such as leadership, teamwork, trust, implementation climate, and resource availability, also shape implementation outcomes [44]. Drawing upon some of these factors, we explain how these link to the implementation of "Cadê o Kauê?" in schools. The school system is considered an inherently complex setting [45] where multiple components are interconnected and interact dynamically, and any successful intervention must account for and adapt to such complexity [46]. A "whole school approach" [47] that promotes collaborative efforts among administrative staff, teachers, and students, offers a promising strategy for implementation in schools as it addresses school-system complexities [47] and mobilizes collective action [48] across stakeholder groups.

It is, therefore, also essential to clarify role-specific responsibilities, particularly regarding *who* is responsible for introducing the innovation in the school and *who* drives it forward. Teachers are the primary facilitators of "Cadê o Kauê?" within classrooms and are responsible for conducting discussion circles afterwards. The complexity in pilot implementation in schools emerged from the need for additional training to support facilitation, as some teachers felt unprepared to address complex concerns around mental health that surfaced during the discussion circles.

Unfamiliarity and skepticism about innovation can also create initial hurdles in its uptake and sustainability [45]. Although a teacher advisory committee was constituted and consulted with as a part of the coproduction process, teachers must also be consulted about their willingness, knowledge, and beliefs about the technology and broader meaning of mental health through dedicated research. Engaging in collective sensemaking [31] with teachers from the outset, coupled with the provision of additional training, can help mitigate initial resistance toward the technology. Additionally, aligning the intervention with the school's values and fostering mutual understanding through involvement of school administrators in the coproduction process can facilitate the transition and promote a supportive culture for innovation.

XSL•FO
RenderX

Running With "Sociopolitical" and "Emerging" Complexity

During the COVID-19 pandemic, Brazil's sociopolitical climate faced significant challenges [49], which interfered with the planning, design, and implementation of "Cadê o Kauê?" These tensions were faced due to political polarizations complicating the acceptance of adolescent mental health interventions. Similarly, conservative religious influences created resistance, especially concerning LGBTQ+ rights, leading to the cautious design of characters in the story "Cadê o Kauê?"

As we conducted the interviews with the Engadajamente team, participants shared coherent views and acknowledged inherent complexities in Brazil's technology and educational policy landscape, especially with the introduction of technology regulatory policies and a new high school system. Until now, the team has navigated this uncertainty by "running" with it despite several roadblocks. We drew upon the values-informed approach to complexity by Greenhalgh et al [50], where engaging with such uncertainty requires human qualities like courage, humility, and flexibility. Researchers and youth collaborators from the Engadajamente team possess these qualities and have contributed to these human actions in navigating the complexity of the external environment. For instance, a youth collaborator showed "flexibility" when they recommended toning down the "language" and "certain characters" used in the "Cadê o Kauê?" story design to fit the complexities in the sociopolitical landscape. The policy expert embodied "humility" when she acknowledged that uncertainty in Brazil's unripe technology policy landscape may be too early to sustain intervention like "Cadê o Kauê?" in its current version. The teacher in a military school exhibited "courage" by resisting the conflicting authoritarian values of the school and introduced "Cadê o Kauê?" within classrooms in the initial testing study.

Strengths and Limitations

This study represents the first service evaluation project conducted by the Engajadamente team to assess the implementation of "Cadê o Kauê?" in schools. The evaluation design is strengthened by considering the broader social context of implementation as opposed to previous user testing studies conducted in isolation. The knowledge produced from this narrative will provide valuable insights to the Engajadamente team in identifying their priorities for further implementation. This evaluation is timely as the Engajadamente project has recently secured funding from the UK Economic and Social Research Council for further development and evaluation of "Cadê o Kauê?" which will enable the team to address some of the complexities identified in the study.

This study is not without limitations. Data analysis was conducted by SS, a non-Brazilian researcher, using translated transcripts. To minimize potential misunderstandings, she dedicated significant time to understanding the sociocultural context of Brazil, the transcripts were thoroughly checked for accuracy, and GP and SGM provided substantial input to support both data analysis and interpretation. Another limitation is the retrospective nature of participants' reflections on the implementation. Future research conducted concurrently with the implementation phase could offer additional insights. Finally, the study could have been strengthened by diversifying the participant sample. Including more teachers and students involved in intervention testing in schools would have enabled a more comprehensive understanding of their diverse experiences.

Translation to Policy and Practice

While this study was situated within Brazilian public school contexts, several findings may be transferable to other LMICs and school settings where digital mental health interventions are introduced into complex and resource-constrained systems. In particular, challenges related to organizational readiness, infrastructure variability, and alignment with existing institutional priorities are likely to resonate across similar educational and policy environments.

The application of NASSS–Complexity Assessment Tool in this study also demonstrates its utility for examining sociotechnical evaluation of technology implementation in LMIC contexts. Given that the framework has been used only in a small number of LMIC studies to date, our findings add to the emerging evidence based on its applicability for examining the sociotechnical complexity of technology implementation.

Conclusion

Using the NASSS framework, this evaluation project has generated a rich narrative that captures both (1) the retrospective complexities encountered during the initial implementation of "Cadê o Kauê?" in schools and (2) the emerging complexities that may unfold with time, affecting its future implementation. By engaging with theory-informed approaches to technology implementation projects and deep-diving into discursive dynamic interactions between technology and its "context" of implementation, we have explored ways to manage these complexities and developed suitable recommendations to guide the Engadajamente team. The implementation of "Cadê o Kauê?" highlighted challenges related to internet dependency, infrastructure limitations, and varying levels of organizational readiness, teacher preparedness, and school culture factors that influence its successful integration and scalability. Addressing these through offline solutions, additional teacher training, whole-school engagement strategies, and a value-informed complexity approach can enhance accessibility, alignment with school priorities, and adaptability within Brazil's dynamic socio-political context.

Acknowledgments

The authors would like to thank the entire Engadajamente team in Brazil and the UK for their valuable insights and participation, and the teachers in Brasilia for taking the time to participate in this project.

XSL•FO
RenderX

Funding

This study was funded by an NDPH Intermediate Fellowship (Oxford Population Health, University of Oxford) to GP and an UK Economic and Social Research Council Research Grant (ES/Z503745/1) to GP and SM.

Authors' Contributions

SS, GP, and MK contributed to the conceptualization of the study and development of the initial ideas. Data curation was undertaken by SS, who also carried out the data analysis under the supervision of GP, MK, and SGM. SS and GP were responsible for the investigation, and, together with MK, contributed to the development of the study methodology. SS, GP, and SGM undertook the project administration. SS led the preparation of the original draft, and GP, MK, and SGM contributed to the review and editing of the manuscript.

Conflicts of Interest

None declared

Multimedia Appendix 1
Semistructured Interview Guide.
[DOCX File, 18 KB - jmir_v28i1e79106_app1.docx]

Multimedia Appendix 2
Participation Information Sheet.
[PDF File, 134 KB - jmir_v28i1e79106_app2.pdf]

Checklist 1
COREQ checklist.
[PDF File, 401 KB - jmir_v28i1e79106_app3.pdf]

References

1. Guidance on Community Mental Health Services: Promoting Person-Centred and Rights-Based Approaches, 1st edition: World Health Organization; 2021.
2. McGorry PD, Mei C, Dalal N, et al. The Lancet Psychiatry Commission on youth mental health. Lancet Psychiatry 2024 Sep;11(9):731-774. [doi: 10.1016/S2215-0366(24)00163-9] [Medline: 39147461]
3. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity survey replication. Arch Gen Psychiatry 2005 Jun;62(6):593-602. [doi: 10.1001/archpsyc.62.6.593] [Medline: 15939837]
4. Feter N, Caputo EL, Doring IR, et al. Sharp increase in depression and anxiety among Brazilian adults during the COVID-19 pandemic: findings from the PAMPA cohort. Public Health (Fairfax) 2021 Jan;190:101-107. [doi: 10.1016/j.puhe.2020.11.013]
5. Barros MDA, Lima MG, Malta DC, et al. Mental health of Brazilian adolescents during the COVID-19 pandemic. Psychiatry Res Commun 2022 Mar;2(1):100015. [doi: 10.1016/j.psycom.2021.100015] [Medline: 34977912]
6. Roberts T, Miguel Esponda G, Torre C, Pillai P, Cohen A, Burgess RA. Reconceptualising the treatment gap for common mental disorders: a fork in the road for global mental health? Br J Psychiatry 2022 Sep;221(3):553-557. [doi: 10.1192/bjp.2021.221] [Medline: 35137680]
7. Clarke AM, Kuosmanen T, Barry MM. A systematic review of online youth mental health promotion and prevention interventions. J Youth Adolesc 2015 Jan;44(1):90-113. [doi: 10.1007/s10964-014-0165-0] [Medline: 25115460]
8. Ribeiro IDS, Correa MM, Oliveira G, Cade NV. Common mental disorders and socioeconomic status in adolescents of ERICA. Rev Saude Publica 2020;54:04. [doi: 10.11606/S1518-8787.2020054001197] [Medline: 31939577]
9. Mental health promotion. Intellectual Disability Mental Health Connect. URL: https://idmhconnect.health/i-am-professional/mental-health-promotion [accessed 2026-02-10]
10. Rose-Clarke K, Bitta M, Evans-Lacko S, et al. Centring youth mental health discourse on low-income and middle-income countries. Lancet Psychiatry 2024 Sep;11(9):671-672. [doi: 10.1016/S2215-0366(24)00211-6] [Medline: 39147454]
11. The Ottawa Charter (The WHO, 1986) marked a global turning point in health policies. URL: https://www.regione.lombardia.it/wps/wcm/connect/938b9ac0-e2fd-4376-8586-486cabfcc05e/Mental+Health+Glossary.pdf?MOD=AJPERES&CACHEID=938b9ac0-e2fd-4376-8586-486cabfcc05e [accessed 2026-02-11]
12. Santre S. Mental health promotion in adolescents. Journal of Indian Association for Child and Adolescent Mental Health 2022 Apr;18(2):122-127. [doi: 10.1177/09731342221120709]
13. Vijayaraghavan J, Vidyarthi A, Livesey A, et al. Strengthening adolescent agency for optimal health outcomes. BMJ 2022 Oct 27;379:e069484. [doi: 10.1136/bmj-2021-069484] [Medline: 36302546]

XSL•FO
RenderX

14. Bower M, Donohoe-Bales A, Nguyen AQH, et al. Mental health and wellbeing outcomes of youth participation: a scoping review protocol. PLoS ONE 2023;18(10):e0293006. [doi: 10.1371/journal.pone.0293006] [Medline: 37847717]
15. Gatera G, Pavarini G. The voices of children in the global health debate. Lancet 2020 Feb 22;395(10224):541-542. [doi: 10.1016/S0140-6736(20)30364-0] [Medline: 32085820]
16. Baka E, Tan YR, Wong BLH, Xing Z, Yap P. A scoping review of digital interventions for the promotion of mental health and prevention of mental health conditions for young people. Oxf Open Digit Health 2025;3:oqaf005. [doi: 10.1093/oodh/oqaf005] [Medline: 40230873]
17. Gonsalves PP, Ansari S, Berry C, et al. Co-designing digital mental health interventions with young people: 10 recommendations from lessons learned in low-and-middle-income countries. Health Educ J 2025 Jun;84(4):373-384. [doi: 10.1177/00178969241297821]
18. Bevan Jones R, Stallard P, Agha SS, et al. Practitioner review: co-design of digital mental health technologies with children and young people. J Child Psychol Psychiatry 2020 Aug;61(8):928-940. [doi: 10.1111/jcpp.13258] [Medline: 32572961]
19. Carter H, Araya R, Anjur K, Deng D, Naslund JA. The emergence of digital mental health in low-income and middle-income countries: a review of recent advances and implications for the treatment and prevention of mental disorders. J Psychiatr Res 2021 Jan;133:223-246. [doi: 10.1016/j.jpsychires.2020.12.016] [Medline: 33360867]
20. Naslund JA, Deng D. Addressing mental health stigma in low-income and middle-income countries: a new frontier for digital mental health. Ethics Med Public Health 2021 Dec;19:100719. [doi: 10.1016/j.jemep.2021.100719] [Medline: 35083375]
21. Pavarini G, Murta SG, Mendes JDA, et al. Cadê o Kauê? co-design and acceptability testing of a chat-story aimed at enhancing youth participation in the promotion of mental health in Brazil. J Child Psychol Psychiatry 2025 May;66(5):697-715. [doi: 10.1111/jcpp.14078] [Medline: 39703174]
22. General 2. Engajadamente. URL: https://www.engajadamente.org/o-projeto [accessed 2026-02-11]
23. Siston FR, Murta SG, Mendes JDA, et al. A collective autoethnography of coproduction in mental health research by academic researchers and young people in Brazil. BMJ Glob Health 2023 Nov;8(11):e012443. [doi: 10.1136/bmjgh-2023-012443] [Medline: 37949502]
24. Pavarini G, Murta SG, Mendes J, et al. Your best friend is missing and only you can find him: a chat-story to promote peer support and collective action for mental health. 2023 Presented at: Proceedings of the 22nd Annual ACM Interaction Design and Children Conference; Jun 19-22, 2023; Chicago, IL p. 732-735. [doi: 10.1145/3585088.3594496]
25. Greenhalgh T, Abimbola S. The NASSS framework - a synthesis of multiple theories of technology implementation. Stud Health Technol Inform 2019 Jul 30;263:193-204. [doi: 10.3233/SHTI190123] [Medline: 31411163]
26. Greenhalgh T, Papoutsi C. Studying complexity in health services research: desperately seeking an overdue paradigm shift. BMC Med 2018 Jun 20;16(1):95. [doi: 10.1186/s12916-018-1089-4] [Medline: 29921272]
27. Cohn S, Clinch M, Bunn C, Stronge P. Entangled complexity: why complex interventions are just not complicated enough. J Health Serv Res Policy 2013 Jan;18(1):40-43. [doi: 10.1258/jhsrp.2012.012036] [Medline: 23393041]
28. Cresswell KM, Sheikh A. Undertaking sociotechnical evaluations of health information technologies. Inform Prim Care 2014;21(2):78-83. [doi: 10.14236/jhi.v21i2.54] [Medline: 24841408]
29. Papoutsi C, Wherton J, Shaw S, Morrison C, Greenhalgh T. Putting the social back into sociotechnical: case studies of co-design in digital health. J Am Med Inform Assoc 2021 Feb 15;28(2):284-293. [doi: 10.1093/jamia/ocaa197] [Medline: 33043359]
30. Leonardi PM. Materiality, sociomateriality, and socio-technical systems: what do these terms mean? how are they related? do we need them? Social Science Research Network. Preprint posted online on Aug 15, 2012. [doi: 10.2139/ssrn.2129878]
31. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
32. Attribution 4.0 international (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/#ref-appropriate-credit [accessed 2026-05-15]
33. Shin HD, Hamovitch E, Gatov E, et al. The NASSS (non-adoption, abandonment, scale-up, spread and sustainability) framework use over time: a scoping review. PLOS Digit Health 2025 Mar;4(3):e0000418. [doi: 10.1371/journal.pdig.0000418] [Medline: 40096260]
34. Braun V, Clarke V. Thematic Analysis: A Practical Guide: Sage Publications Ltd; 2021.
35. Adams J. Examining ethical and social implications of digital mental health technologies through expert interviews and sociotechnical systems theory. DISO 2024 Aug;3(2):24. [doi: 10.1007/s44206-024-00110-5]
36. Lehoux P, Miller FA, Daudelin G, Denis JL. Providing value to new health technology: the early contribution of entrepreneurs, investors, and regulatory agencies. Int J Health Policy Manag 2017 Sep 1;6(9):509-518. [doi: 10.15171/ijhpm.2017.11] [Medline: 28949463]
37. Shuvo T, Islam R, Hossain S, et al. eHealth innovations in LMICs of Africa and Asia: a literature review exploring factors affecting implementation, scale-up, and sustainability. IEH 2015;20152:95. [doi: 10.2147/IEH.S88809]
38. Meyer AJ, Armstrong-Hough M, Babirye D, et al. Implementing mHealth interventions in a resource-constrained setting: case study from Uganda. JMIR Mhealth Uhealth 2020 Jul 13;8(7):e19552. [doi: 10.2196/19552] [Medline: 32673262]

XSL•FO
RenderX

39. Greenhalgh T, Robert G, Macfarlane F, Bate P, Kyriakidou O. Diffusion of innovations in service organizations: systematic review and recommendations. Milbank Q 2004;82(4):581-629. [doi: 10.1111/j.0887-378X.2004.00325.x] [Medline: 15595944]
40. Fatori D, Salum GA, Rohde LA, et al. Use of mental health services by children with mental disorders in two major cities in Brazil. Psychiatr Serv 2019 Apr 1;70(4):337-341. [doi: 10.1176/appi.ps.201800389] [Medline: 30651056]
41. Amaral CE, Onocko-Campos R, de Oliveira PRS, et al. Systematic review of pathways to mental health care in Brazil: narrative synthesis of quantitative and qualitative studies. Int J Ment Health Syst 2018;12:65. [doi: 10.1186/s13033-018-0237-8] [Medline: 30450125]
42. Guidance on community mental health services: promoting person-centred and rights-based approaches. World Health Organization. URL: https://www.who.int/publications/i/item/9789240025707 [accessed 2026-02-09]
43. Vieira MA, Gadelha AA, Moriyama TS, Bressan RA, Bordin IA. Evaluating the effectiveness of a training program that builds teachers' capability to identify and appropriately refer middle and high school students with mental health problems in Brazil: an exploratory study. BMC Public Health 2014 Feb 28;14(1):210. [doi: 10.1186/1471-2458-14-210] [Medline: 24580750]
44. Caci L, Nyantakyi E, Blum K, et al. Organizational readiness for change: a systematic review of the healthcare literature. Implement Res Pract 2025;6:26334895251334536. [doi: 10.1177/26334895251334536] [Medline: 40385227]
45. Bradshaw M, Gericke H, Coetzee BJ, Stallard P, Human S, Loades M. Universal school-based mental health programmes in low- and middle-income countries: a systematic review and narrative synthesis. Prev Med 2021 Feb;143:106317. [doi: 10.1016/j.ypmed.2020.106317] [Medline: 33159922]
46. Hoare E, Bott D, Robinson J. Learn it, live it, teach it, embed it: implementing a whole school approach to foster positive mental health and wellbeing through positive education. Intnl J Wellbeing ;7(3):56-71 [FREE Full text] [doi: 10.5502/ijw.v7i3.645]
47. Clarke AM, Barry MM. Supporting a whole-school approach to mental health promotion and wellbeing in post-primary schools in ireland. In: Kutcher S, Wei Y, Weist MD, editors. School Mental Health. URL: https://www.cambridge.org/core/books/abs/school-mental-health/supporting-a-wholeschool-approach-to-mental-health-promotion-and-wellbeing-in-postprimary-schools-in-ireland/3AFFE8C936243D7949AF1B25BE03BBC7 [Medline: 20011679]
48. May CR, Johnson M, Finch T. Implementation, context and complexity. Implement Sci 2016 Oct 19;11(1):141. [doi: 10.1186/s13012-016-0506-3] [Medline: 27756414]
49. Avritzer L, Rennó L. The pandemic and the crisis of democracy in Brazil. Journal of Politics in Latin America 2021 Dec;13(3):442-457. [doi: 10.1177/1866802X211022362]
50. Greenhalgh T, Engebretsen E, Bal R, Kjellström S. Toward a values-informed approach to complexity in health care: hermeneutic review. Milbank Q 2023 Sep;101(3):646-674. [doi: 10.1111/1468-0009.12656] [Medline: 37219239]

Abbreviations

LMIC: low- and middle- income country
NASSS: Non-adoption, Abandonment, Scale-up, Spread, Sustainability

Edited by A Stone; submitted 06.Sep.2025; peer-reviewed by A Ionescu, Y Li; revised version received 10.Apr.2026; accepted 13.Apr.2026; published 27.May.2026.

Please cite as:
Setia S, Pavarini G, Koshkouei M, Giardini Murta S
Using the NASSS–Complexity Assessment Tool to Evaluate the Implementation of "Cadê O Kauê?": Chat-Story Intervention for Youth Participation in Mental Health Promotion in Brazil
J Med Internet Res 2026;28:e79106
URL: https://www.jmir.org/2026/1/e79106
doi:10.2196/79106

© Sakshi Setia, Gabriela Pavarini, Mona Koshkouei, Sheila Giardini Murta. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Explaining the Use Behavior of Digital Technologies in Pediatric Rehabilitation: Structural Equation Modeling Analysis of a Cross-Sectional European Survey

Johanne Mensah Gourmel[1,2,3,4], MD, PhD; Sylvain Brochard[1,2,3,4], MD, PhD; Saranda Bekteshi[5], PhD; Elegast Monbaliu[5], PhD; Gwenaël Cornec[1,2,4], MSc, MD; Anca Irina Grigoriu[4,6], MD, PhD; Christopher J Newman[7,8], MD, PhD; Marco Konings[5], PhD; Javier De La Cruz[9*], PhD; Christelle Pons[1,2,3,4*], MD, PhD

[1]UMR1101, Université de Bretagne Occidentale, Brest, France
[2]Physical Medicine and Rehabilitation Department, Centre Hospitalier Universitaire de Brest, 2 avenue Foch, Brest, France
[3]Pediatric Rehabilitation Department, Fondation Ildys, Brest, France
[4]INSERM - UMR1101, Laboratoire de Traitement de l'Information Médicale, Brest, France
[5]Department of Rehabilitation Sciences, Neurorehabilitation Technology Lab, KU Leuven, Bruges, Belgium
[6]Multifactorial Gait Analysis Laboratory, National Center of Neurorehabilitation for Children "Robanescu-Padure", Bucharest, Romania
[7]Paediatric Neurology and Neurorehabilitation Unit, University Hospital of Lausanne, Lausanne, Switzerland
[8]Clinical Sciences Section, Faculty of Biology and Medicine, Lausanne University, Lausanne, Switzerland
[9]UiCEC+12, University Hospital 12 de Octubre Research Institute (imas12), Madrid, Spain
[*]these authors contributed equally

Corresponding Author:
Johanne Mensah Gourmel, MD, PhD
UMR1101, Université de Bretagne Occidentale, Brest, France

Abstract

Background: Digital technologies for rehabilitation (DT4R), such as robotics and treadmill systems (RobTS), virtual reality and active video gaming (VR-AVG), and telehealth and apps (T&Apps), are promising tools for pediatric motor rehabilitation. Identifying acceptance factors is essential for effective clinical adoption.

Objective: This study aimed to analyze the use of 3 different technologies for rehabilitation—RobTS, VR-AVG, and T&Apps—through a causal model based on the Unified Theory of Acceptance and Use of Technology (UTAUT).

Methods: This study was part of RehaTech4child, a cross-sectional survey (2022) supported by the European Academy of Childhood-onset Disability, aimed at professionals working in pediatric motor rehabilitation across Europe. It assessed DT4R use, intention to use, and UTAUT concepts (performance expectancy, effort expectancy, social influence, and barriers). Structural equation modeling was performed to analyze the data and understand relationships between observed and latent variables.

Results: A total of 1397 responses were received, and 635 fulfilled the eligibility criteria. The fitness indices suggested a satisfactory fit between the data and the model. The model explained 67% of the variance in the use of RobTS, 62% in VR-AVG, and 57% in T&Apps. Among all studied determinants, access had the strongest impact on use for all 3 categories of DT4R (RobTS: β=0.78, VR-AVG: β=0.73, and T&Apps: β=0.70; *P*<.001). Intention to use significantly impacted use behavior for all technologies; it was the second determinant after access for VR-AVG (β=0.18, *P*<.001) and T&Apps (β=0.21, *P*<.001), with a lower weight for RobTS (β=0.06, P=.007; *P*<.001). In the subgroup analysis of respondents reporting easy access, intention to use was the strongest determinant of use. The model explained 61% of the variance in intention to use for RobTS, 67% for VR-AVG, and 68% for T&Apps. Performance expectancy had the strongest effect on intention to use for the 3 technologies (RobTS: β=0.81, VR-AVG: β=0.84, and T&Apps: β=0.90; *P*<.001). For this concept, the items with the highest weights were significantly related to the effectiveness of DT4R on rehabilitation. Social influence and effort expectancy had a slight impact on intention to use.

Conclusions: These results underscore the need to ensure easy access as a prerequisite for assessing relevant determinants of acceptance. Developing the evidence base for DT4R effectiveness and ensuring the availability of existing evidence may facilitate DT4R implementation. In our study, within the framework of the UTAUT model, no acceptance barrier was linked to the use of DT4R with children. Gathering families' views may be useful for the implementation of RobTS. T&Apps may be useful for involving parents in their child's rehabilitation. Further studies should focus on children's and families' points of view.

Trial Registration: ClinicalTrials.gov NCT05176522; https://clinicaltrials.gov/search?cond=NCT05176522

(J Med Internet Res 2026;28:e80355) doi:10.2196/80355

XSL•FO
RenderX

KEYWORDS

digital technologies; rehabilitation; pediatrics; Unified Theory of Acceptance and Use of Technology model; UTAUT; structural equation modeling

Introduction

Motor impairment hinders a child's development and participation [1,2]. Specialized care and continuous management with individualized rehabilitation programs reduce the burden of disabilities. Keeping rehabilitation fun and engaging to optimize its effect is a real challenge to maintain during growth; including recent technological advances that enable new approaches could help address this issue. Digital technologies (DT), including robotics and treadmill systems (RobTS), virtual reality and active video gaming (VR-AVG), and telehealth and apps (T&Apps), are creating new opportunities in pediatric rehabilitation. Their use shows promise in areas such as upper limb rehabilitation [3], improving walking and balance [4,5], and facilitating the activities and participation of children with disabilities [6,7]. In addition, these technologies can be used to motivate children and encourage their participation [8,9]. Preliminary evidence showed that digital technologies for rehabilitation (DT4R) are both acceptable [10] and enjoyable for children [9]. These 3 groups of technologies represent growing opportunities in pediatric motor care; however, there is still room for optimization regarding their integration into clinical practice. Children who require motor rehabilitation are a heterogeneous population. Furthermore, they grow and their developmental abilities change constantly, which creates specific challenges for the implementation of technologies [11] and supports the need to conduct specific studies. Further research is required to understand the behavioral factors influencing acceptance to improve implementation. In this context, RehaTech4child, a cross-sectional survey (2022) supported by the European Academy of Childhood-onset Disability (EACD), was disseminated to professionals working in pediatric motor rehabilitation across Europe. The first aim was to assess access to and use of DT for pediatric motor rehabilitation in Europe as a function of individual and environmental factors, as well as the potential barriers to their use. The initial results highlighted that DT4R was regularly used in clinical practice by around two-thirds of the European respondents.

The gap between the provision of equipment in a clinical facility and its actual use has been highlighted in the literature [12], and this gap is particularly pronounced for technologies. In the late 1980s [13], theories were developed to explain an individual's behavior toward a new technology and to understand the levers that may help providers move from purchase to effective practical implementation within their departments. To increase usage, the core concepts highlight the need to improve technology acceptance [12]. In 2003, Venkatesh et al [14] identified 8 models that explain how and why individuals adopt new information technologies and synthesized them into a unique model. Finally, they retained the core concepts that explained the largest proportion of variance in the Unified Theory of Acceptance and Use of Technology (UTAUT) model: approximately 50% of the variance in use and 70% in intention to use. In this model, intention to use and facilitating conditions are the 2 direct determinants of use. Three main determinants are directly related to intention to use: performance expectancy (PE), effort expectancy (EE), and social influence (SI). PE can be defined as the degree to which an individual believes that using the system will help them attain gains and enhance their job performance, EE is defined as the degree of ease associated with the use of the system, and SI is defined as the degree to which an individual perceives that important others believe they should use the system. Facilitating conditions pertain to the existing organizational and technical infrastructure supporting the system use. Applying this conceptual framework to a given technology improves understanding of its acceptance. This can help individuals and teams who purchase and implement technologies within clinical facilities to identify actions aimed at facilitating successful implementation [14].

The UTAUT model has been widely used to explain health care professionals' behavior toward clinical information systems (eg, the picture archiving and communication system and health information system) and to identify levers and actions to be taken to facilitate their implementation and adoption [15-17]. More recently, the model has also been applied to specific rehabilitation technologies, such as the use of 3D printing systems by occupational therapists [18] and the use of an exergame-based telerehabilitation system by older adults and their health care professionals [19]. Previous UTAUT studies have generally shown that intention to use is an important determinant of use, confirming the need to address behavioral aspects of technology adoption. Regarding the determinants of intention to use, PE is generally a key factor [20], and this effect appears to hold across various professions [21,22]. However, to our knowledge, these studies have remained focused on isolated technologies and contexts, without drawing generalizable insights into the acceptance of rehabilitation technologies. Recent studies [23] highlighted that there are still challenges in the implementation of DT4R, including using these devices to their full potential. The question of the relevance of DT4R for children was also raised in terms of device size and developmental issues [24], as well as cost-effectiveness and patient preferences [25]; however, research in pediatric populations has been focused on technical rather than behavioral aspects. To our knowledge, no study has evaluated the acceptance of digital technologies in pediatric rehabilitation. Children requiring motor rehabilitation have specific needs that must be understood. Furthermore, they are widely dispersed geographically. Therefore, we chose a macroscopic perspective to identify levers and actions to facilitate DT4R implementation in children. The objective was to gather the opinions of a large and diverse European sample across 3 groups of digital technologies to obtain generalizable and actionable findings. Considering technologies with varied levels of technicity and usability by nonprofessionals may help deepen the understanding of acceptance across the different categories.

This study aimed to assess the behavioral components of use and intention to use for RobTS, VR-AVG, and T&Apps in

XSL•FO
RenderX

pediatric motor rehabilitation through a causal model based on the UTAUT, and thus to identify key acceptance factors for improving implementation in clinical practice, complementing previous descriptions of access and use in relation to individual and environmental determinants [26].

Based on responses from health care professionals participating in the European RehaTech4Child survey, we hypothesized that (1) the UTAUT model would explain use behavior for the 3 DT4R technologies; (2) intention to use would have the strongest impact on use for each technology; (3) PE would significantly impact intention to use, regardless of profession; and (4) the impact of determinants on use would differ across technologies.

Methods

Study Design

This study was part of RehaTech4child, an online European cross-sectional survey (2022) sponsored by the EACD, targeting professionals working in pediatric motor rehabilitation across Europe. This report follows the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) and CHERRIES (Checklist for Reporting Results of Internet E-Surveys) guidelines. The study was registered on ClinicalTrials.gov (NCT05176522).

Ethical Considerations

The study was conducted in compliance with current French legislation (Loi Jardé 2012‑300) [27-29]. The Brest CHU (Centre Hospitalo-Universitaire/University Hospital Center) Institutional Review Board considered that ethical committee approval was not necessary.

Survey Description

The first part of the survey focused on individual determinants (respondents' sociodemographic characteristics) and environmental determinants, including the type of professional practice and the characteristics of respondents' patients. The second part assessed DT4R use, intention to use, and UTAUT concepts for each of the 3 categories of DT4R: RobTS, VR-AVG, and T&Apps [8,26]. The UTAUT model is recommended in the health care context but must be adapted to a given technology and the specific context of each study [12,14]. On the basis of previous research and discussions with the steering committee during the elaboration of the survey, we oriented the questions to focus on barriers rather than facilitating conditions to align with previous research on professional practice conditions in rehabilitation, where organizational and technical infrastructures appeared as barriers, as well as feedback from the pilot testers [8,30,31]. Furthermore, in rehabilitation, access has been stated as a key priority to facilitate DT4R adoption [32]. However, access is not always considered a barrier per se in the literature and is not often included as such in studies about barriers and enablers [33]. A previous study showed the major impact of access on use [26]. Access was therefore considered separately from other barriers. Use (and frequency of use), intention to use, PE (8 questions), EE and SI (2 questions), access (and ease of access), and barriers were evaluated on a 7-point Likert scale. The survey is available online [34]. A detailed description of the survey protocol has been published previously [8].

Participants

The survey was translated into 20 different languages and disseminated by the EACD national coordinators through European and national professional and personal networks using convenience and snowball sampling. Responses were excluded if respondents did not practice motor rehabilitation with children, were from a non-EACD member country, or did not provide responses about DT4R [8,26].

Data Analysis

Structural equation modeling was used to analyze the data and elucidate the relationships between observed and latent variables, based on the UTAUT model as defined by Venkatesh et al [14] (Figure 1). R software (Lavaan library) was used. Model fit indices were calculated (comparative fit index [CFI], Tucker-Lewis index [TLI], and root mean square error of approximation [RMSEA]) to evaluate the quality of the model, as recommended by Stone [35]. R^2 coefficients were determined for the explained variance of use and intention to use for the 3 DT4R categories in the main analysis. Relationships between latent and observed variables were evaluated using the standardized regression coefficient (β) and its associated P value. Statistical significance was set at $P<.05$.

XSL•FO
RenderX

Figure 1. Relationship between observed (rectangular boxes) and latent (oval boxes) variables as specified in the structural equation modeling implemented using R software. The dashed line includes all items in between. SI: social influence; PE: performance expectancy.

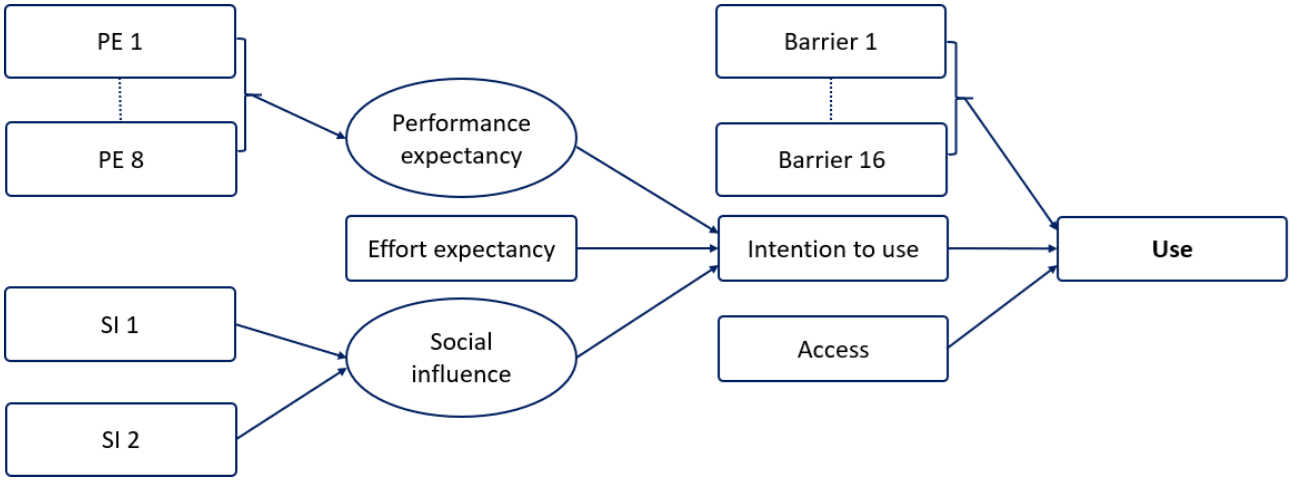

Subgroup Analysis

As access and ease of access were previously identified as major determinants of use and frequency of use [26], exploratory analyses were performed for subgroups based on ease of access to explore if the impact of various determinants differed for respondents reporting easier access (defined as ≥4 points on the Likert scale).

The literature suggests that EE may be underestimated by individuals who have not experienced the studied technology at least once [12,18]. Therefore, to determine whether this could be a leverage point for improvement to facilitate the implementation of DT4R, we conducted an exploratory analysis of the data for the subgroup of actual users.

Other authors highlighted differences in the impact of SI when studying 2 countries with different cultures [36]. We therefore performed a subgroup analysis on areas of Europe to explore the impact of SI across geographical subregions. The countries listed by Pons et al [8] in the protocol were classified according to the United Nations Geoschemes [37].

Evidence in the literature shows that PE influences the intention to use, and this effect appears to hold across various professions. Therefore, to evaluate this in the specific context of pediatric rehabilitation, we performed a subgroup analysis on professions.

Results

Participants' Characteristics

A total of 1395 responses were received between January 1, 2022, and November 1, 2022, of which 635 were included in the analysis; 599 (94%) responses were complete for RobTS, 563 (89%) for VR-AVG, and 553 (87%) for T&Apps. Of the respondents, 83% (528/635) were female participants. There were 53% (334/635) physiotherapists, 26% (166/635) physicians, and 15% (95/635) occupational therapists. It has been reported that 37% (219/599), 45.5% (256/563), and 50.8% (281/553) of respondents used RobTS, VR-AVG, and T&Apps, respectively. When asked about their intention to use DT4R if they had access to it, 76.6% (459/599), 74.2% (418/563), and 69% (382/553) of all respondents at least somewhat agreed that they would use RobTS, VR-AVG, and T&Apps, respectively. A detailed description of participant characteristics has been published previously [26].

Structural Equation Modeling: Model Fit Indices

Standard model fit indices showed that the model was acceptable (CFI=0.839/0.871/0.876, TLI=0.818/0.855/0.86, and RMSEA=0.081/.073/0.07 for RobTS, VR-AVG, and T&Apps, respectively).

UTAUT Determinants of Use

The model explained 67% of the variance in use for RobTS, 62% for VR-AVG, and 56% for T&Apps. Among all the variables, access had the strongest impact on use for all 3 categories of DT4R (RobTS: β=0.78, VR-AVG: β=0.73, and T&Apps: β=0.70; P<.001) (Figure 2A, B, and C and Table 1).

XSL•FO
RenderX

Figure 2. Determinants of use and intention to use for robotics and treadmill systems (A), virtual reality and active video gaming (B), and telehealth and apps (C) according to the Unified Theory of Acceptance and Use of Technology model. Dotted arrows represent β coefficients with *P*>.05. β coefficients are specified when *P*<.05. Dashed box contours indicate negative coefficients. Barriers are represented by light gray boxes. β coefficients were estimated using structural equation modeling. Comparative fit index=0.839/0.871/0.876, Tucker-Lewis index=0.818/0.855/0.86, and root mean square error of approximation=0.081/0.073/0.07 for robotics and treadmill systems, virtual reality and active video gaming, and telehealth and apps, respectively.

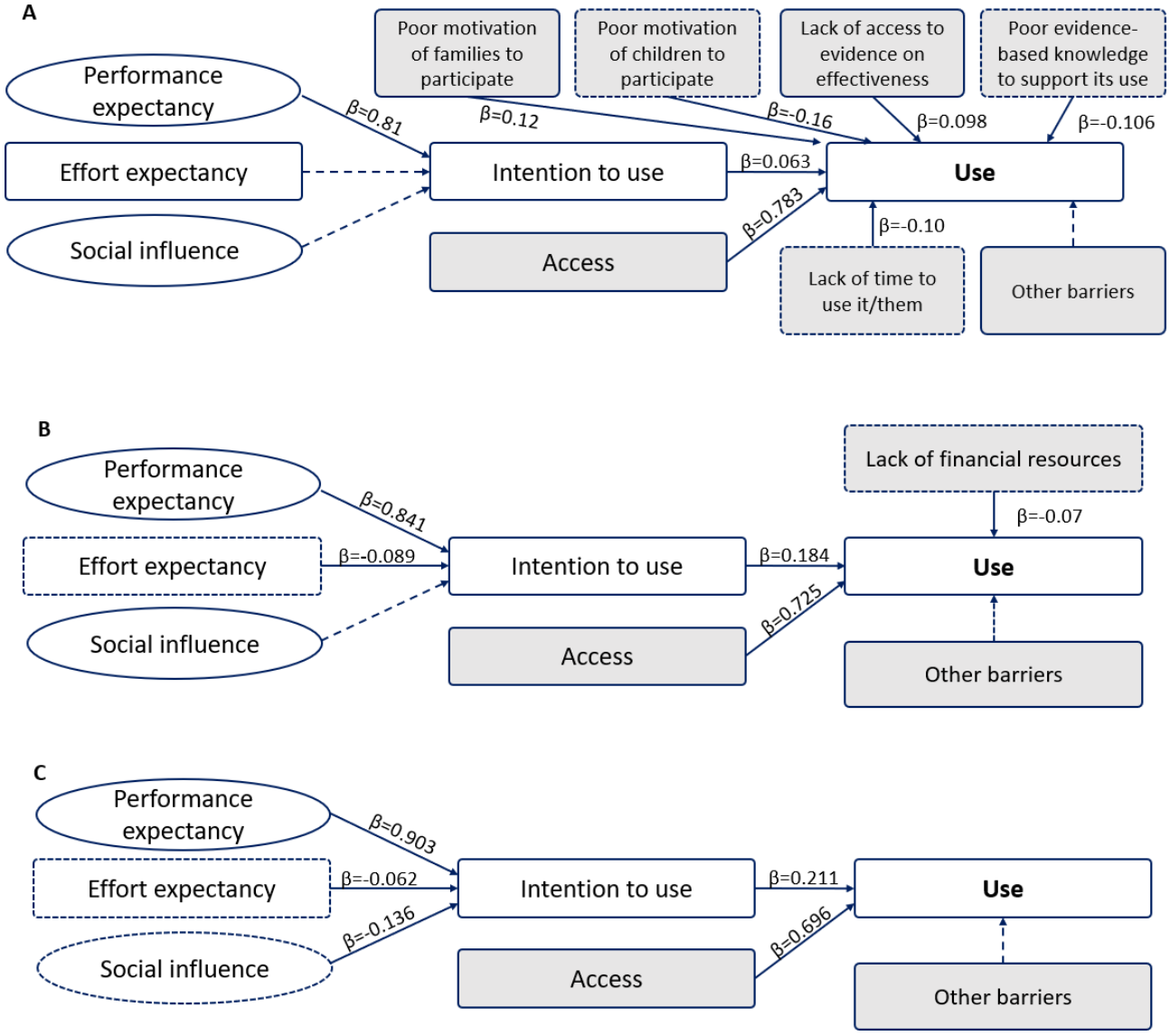

XSL•FO
RenderX

Table . Results of the structural equation modeling for use of robotics and treadmill systems (RobTS), virtual reality and active video gaming (VR-AVG), and telehealth and apps (T&Apps) with the standardized regression coefficients (β) and their associated *P* value for each of the latent and observed variables.

Use	RobTS		VR-AVG		T&Apps	
	β	*P* value	β	*P* value	β	*P* value
Intention to use	0.06[a]	.01	0.18[a]	<.001	0.21[a]	<.001
Access	0.78[a]	<.001	0.73[a]	<.001	0.70[a]	<.001
Barrier 1: Lack of financial resources	0.02	.39	−0.07[a]	.02	−0.04	.21
Barrier 2: Lack of confidence (technophobia)	0.01	.84	0.01	.79	−0.02	.63
Barrier 3: Lack of time to learn how to use it/them	0.05	.22	0.01	.91	−0.02	.76
Barrier 4: Lack of time to use it/them	−0.010[a]	.03	−0.07	.17	0.04	.46
Barrier 5: Lack of time to set up and clean up	−0.002	.10	0.04	.42	0.05	.27
Barrier 6: Lack of robustness (in situations like heavy usage, pulling, and drooling)	0.05	.11	0.03	.49	−0.06	.13
Barrier 7: Lack of access to evidence on effectiveness	0.10[a]	.02	−0.002	.97	0.03	.54
Barrier 8: Poor evidence-based knowledge to support its effectiveness and use	−0.12[a]	.01	0.01	.89	−0.03	.63
Barrier 9: Lack of accessible assistance (specialized instruction or specific person)	−0.005	.88	0.003	.94	−0.02	.75
Barrier 10: Lack of initial training	−0.05	.23	−0.08	.15	0.02	.77
Barrier 11: Lack of educational opportunities	0.01	.30	0.04	.58	−0.01	.90
Barrier 12: Lack of training opportunities	−0.10	.05	−0.02	.77	−0.09	.24
Barrier 13: Space-related issues (eg, insufficient or inappropriate space)	0.03	.35	0.001	.97	0.06	.12
Barrier 14: DT[b] that are not adapted to clients	0.05	.08	−0.01	.82	0.02	.66
Barrier 15: Poor motivation of children to participate	−0.16[a]	<.001	−0.001	.99	−0.07	.19

XSL•FO
RenderX

Use	RobTS		VR-AVG		T&Apps	
	β	*P* value	β	*P* value	β	*P* value
Barrier 16: Poor motivation of families to participate	0.12[a]	.01	0.05	.26	0.07	.24

[a]*P*<.05.

[b]DT: digital technologies.

Intention to use significantly impacted use behavior for all technologies; it was the second determinant after access for VR-AVG (β=0.18) and T&Apps (β=0.21), with a very low weight for RobTS (β=0.06; *P*<.001).

In the subgroup analysis of respondents reporting easier access, among all studied variables, the impact of intention to use was the strongest (RobTS: β=0.41, *P*<.001; VR-AVG: β=0.53, *P*<.001; and T&Apps: β=0.35, *P*<.001). The impact of access was then nonsignificant and slight.

Regarding the other determinants (Table 1), some were statistically significant but had slight weights. For RobTS, barrier 7: "*Lack of access to evidence on effectiveness*" (β=0.098) and barrier 16: "*Poor motivation of families to participate*" (β=0.12) had a statistically significant impact on use, whereas barrier 4: "*Lack of time to use it/them*" (β=–0.10), barrier 8: "*Poor evidence-based knowledge to support its effectiveness and use*" (β=−0.106), and barrier 15: "*Poor motivation of children to participate*" (β=−0.16) had a statistically negative impact on use. For VR-AVG, barrier 1: "*Lack of financial resources*" (β=−0.07) had a statistically negative impact on use.

UTAUT Determinants of Intention to Use

The model explained 61% of the variance in intention to use for RobTS, 67% for VR-AVG, and 68% for T&Apps. PE had the strongest effect on intention to use for all 3 DT4R categories (β=0.81 for RobTS, β=0.84 for VR-AVG, and β=0.90 for T&Apps; *P*<.001) (Figure 2A, B, and C and Table 2). This effect persisted in the subgroup analysis of professions.

XSL•FO
RenderX

Table . Results of the structural equation modeling for intention to use of robotics and treadmill systems (RobTS), virtual reality and active video gaming (VR-AVG), and telehealth and apps (T&Apps) with the standardized regression coefficients (β) and their associated *P* value for each of the latent and observed variables.

Structural equation modeling	RobTS		VR-AVG		T&Apps	
	β	*P* value	β	*P* value	β	*P* value
Performance expectancy						
PE[a] 1: I believe that rehabilitation with DT[b] complements or enhances the therapists' abilities	0.87[c]	<.001	0.88[c]	<.001	0.87[c]	<.001
PE 2: I think that the use of rehabilitation DT has the potential to improve motor outcomes	0.89[c]	<.001	0.90[c]	<.001	0.84[c]	<.001
PE 3: I think that the use of rehabilitation DT systems encourages active participation of the patient	0.88[c]	<.001	0.84[c]	<.001	0.83[c]	<.001
PE 4: I think that the use of rehabilitation DT adds value to what a conventional approach offers	0.87[c]	<.001	0.90[c]	<.001	0.88[c]	<.001
PE 5: I think that DT4R[d] give the opportunity for family members to participate in the rehabilitation process	0.61[c]	<.001	0.64[c]	<.001	0.75[c]	<.001
PE 6: I think that the use of rehabilitation DT is appropriate within my clinical practice	0.72[c]	<.001	0.78[c]	<.001	0.79[c]	<.001
PE 7: I think that the use of rehabilitation DT is a strategy to increase the number of potential clients	0.55[c]	<.001	0.53[c]	<.001	0.58[c]	<.001
PE 8: I think that rehabilitation DT devices make my work more interesting	0.74[c]	<.001	0.78[c]	<.001	0.77[c]	<.001
Social influence						
SI[e] 1: People whose opinions I value think that I should use rehabilitation DT	0.98[c]	<.001	0.96[c]	<.001	0.99[c]	<.001

XSL•FO
RenderX

Structural equation modeling	RobTS		VR-AVG		T&Apps	
	β	*P* value	β	*P* value	β	*P* value
SI 2: My supervisor thinks that I should use rehabilitation DT	0.76[c]	<.001	0.78[c]	<.001	0.75[c]	<.001
Intention to use						
PE	0.81[c]	<.001	0.84[c]	<.001	0.90[c]	<.001
EE[f]	−0.05	.06	−0.09[c]	.001	−0.06[c]	.02
SI	−0.046	.20	−0.04	.25	−0.14[c]	<.001

[a]PE: performance expectancy.

[b]DT: digital technologies.

[c]*P*<.05.

[d]DT4R: digital technology for rehabilitation.

[e]SI: social influence.

[f]EE: effort expectancy.

Regarding PE determinants (Table 2), items 1 to 4 evaluated the effectiveness of DT4R on rehabilitation. They had the strongest impact on PE, and their weights were similar. PE 6 evaluated the relevance of DT4R in pediatric rehabilitation, and its impact was slightly lower than that of PE 1 to PE 4. The item PE 5: *"I think that DT4R give the opportunity for family members to participate in the rehabilitation process"* had a moderate impact, apart from T&Apps, for which the impact was higher. PE 7: *"I think that the use of rehabilitation digital technologies is a strategy to increase the number of potential clients"* had the lowest weight for the 3 categories of DT4R.

SI and EE (Table 2) had a slight impact on use in comparison with PE. This effect persisted in the subgroup analysis of European subregions. The impact of the opinions of people whose opinions were valued by the respondents was slightly higher than the impact of supervisors' opinions.

The exploratory analysis of the results for users found only a slight impact of EE on the intention to use compared with PE.

Discussion

Principal Findings

In this study, acceptance was modeled within the framework of the UTAUT model using structural equation modeling for 3 types of DT4R: RobTS, VR-AVG, and T&Apps. The model included data from more than 500 different professionals across Europe working in pediatric motor rehabilitation, with or without experience in pediatric rehabilitation technologies. The model explained 55% to 70% of the variance in intention to use and use, as expected by Venkatesh et al [14]. It also contributed to establishing general principles for rehabilitation technologies while highlighting specific characteristics across the different categories of DT4R. Generalizable insights were first highlighted. Within the UTAUT framework, access had the strongest impact on use among all determinants for all 3 DT4R categories. Furthermore, when access was easier, the impact of intention to use became the strongest. PE was the main determinant of intention to use for all 3 DT4R categories, whereas the impacts of SI and EE were slight. For PE, the items regarding the effectiveness of DT4R on rehabilitation had the highest impact. Specific characteristics were then underlined. For instance, for RobTS, the item on children's poor motivation was negatively correlated with use, whereas the item on families' poor motivation was positively correlated. For T&Apps, the item concerning family involvement had a higher weighting than for the other DT4R categories. These analyses helped identify actions needed to promote behavioral change and support the adoption of DT4R.

Lack of access was highlighted as a barrier in the first step of this study [26] and is consistently reported as such in the literature [38]. This barrier is particularly relevant to technologies such as RobTS and VR-AVG, which require space and are relatively expensive. At first sight, the impact of access on use seemed to overwhelm the other determinants of the UTAUT model. However, when access was not only provided but also made easier, intention to use had the strongest impact on use. In summary, ease of access can be understood as a prerequisite for use. Therefore, we recommend conducting a proper evaluation of access conditions before addressing the issue of acceptance for the implementation of DT4R.

PE was the main determinant of intention to use for all 3 DT4R categories. Furthermore, its impact was far greater than that of the other UTAUT determinants. The importance of PE has already been highlighted in the literature on the acceptability of digital technologies in health care, both in the general context of acceptance [12,39,40] and in the specific context of the UTAUT model [41,42] and is now confirmed in the specific context of pediatric rehabilitation. This effect globally persisted in the subgroup analysis of professions. If professionals are convinced that DT4R will improve their practice, they are more likely to use it. This aligns with the items that had higher weights for PE in our study, namely those concerning the effectiveness of DT4R on rehabilitation. To increase intention to use,

XSL•FO
RenderX

professionals must first be confident in the performance of DT4R.

To this aim, some of our results could provide an initial lead; however, they carry a slight weight compared to the main determinants. Notably, 2 barriers that had a statistically significant influence on intention to use for RobTS were related to scientific evidence: negatively for the proposition *"poor evidence-based knowledge to support its use"* and positively for *"lack of access to evidence on effectiveness."* In previous studies, *"lack of evidence for the use of RobTS"* has been highlighted as a limitation to DT4R use by professionals [23], despite the fact that evidence suggests promising results for RobTS in training specific functions, at least in some indications [4]. Our findings suggested that the professionals who responded may already be convinced of the utility of RobTS but were hindered in its use by limited access to evidence. This interpretation should be taken with caution, given the slight weights; however, this information may pave the way for further actions to be taken by rehabilitation department leaders to facilitate the implementation of DT4R in clinical practice: highlighting the evidence of effectiveness (by conducting research in their departments, providing precise recommendations, and facilitating access to current evidence) of DT4R to rehabilitation professionals. Furthermore, digital technology developers should be made aware of this information and should consider developing evidence in accordance with health care guidelines.

The impact of EE on intention to use was lower than that of PE for all 3 DT4R categories. Slegers et al [18] suggested that initial testing of the system reduced perceived ease of use, and Holden and Karsh [12] suggested that EE might be underestimated by people with no or little experience of the technology. To explore this point, we conducted an exploratory analysis of the data for the subgroup of actual users. Although the results of this subgroup analysis should be interpreted with caution, given the smaller sample size, the impact of EE did not change. However, as suggested in previous studies [12], a double impact of experience may occur: having no actual experience of the DT may lead to an underestimation of EE, and, on the contrary, with time and practice, the EE of the system becomes no longer important to acceptance and use. Thus, if the initial training is not sufficient, the EE may increase. However, with sufficient training, it will no longer be a determining factor. In further studies, as done by Nicora et al [23], specifying the level of experience, both in terms of years of practice and practice frequency, may facilitate the analysis of the impact of EE to confirm these hypotheses.

Technical and logistical issues should be considered when implementing DT4R for children, such as screen exposure duration [24], the size of devices, etc. In our acceptance model, the weight of items specifically related to the relevance of DT4R for children, such as *"digital technologies are not adapted to the clients (for professionals working with children)*," was rather slight. Furthermore, the barrier "*poor motivation of children to participate*" was significantly negatively linked to use for RobTS; however, it carried only slight weight. These findings suggest that there was no specific acceptance barrier linked to the use of DT4R in children. Regarding the involvement of families, opinions varied across technologies. Indeed, the barrier "*Poor motivation of families to participate*" was a significant determinant of RobTS use, but with a slight weight. In contrast, evidence in the literature suggests that parents have positive expectations and are generally satisfied with the use of robotic devices by their children [43]. It would be interesting to deepen this specific point in further studies and explore the risk of preconceived ideas of professionals on the motivation of families. On the contrary, some DT4R characteristics could even be assets in the context of pediatric rehabilitation. Thus, in our model, the possibility of involving families more strongly determined PE for T&Apps than for other DT4R categories. For pediatric rehabilitation, T&Apps offer tremendous new possibilities to involve families in their child's rehabilitation [44]. This specific objective may be considered and valued when developing or purchasing a T&Apps device.

Clinicians directly mediate their patients' access to DT. Therefore, the factors influencing therapists' intentions to use the technology are crucial to the implementation process and a condition for access to DT for children and families. To address this, we first questioned health care professionals. However, the overall picture will not be complete without the perspectives of children and families, who are the final users. To ensure efficient use, both the development and evaluation of new digital technologies should involve users. Evaluating acceptance from children's and families' perspectives is a priority in future research to avoid a "one-size-fits-all" approach and to ensure personalized and individualized proposals [45]. Interestingly, recent evidence showed that for the deployment of new interventions, including numeric content for children with emotional disorders, parents' intention to use was influenced by perceived usefulness and efficacy [46]. It would be interesting to determine whether this also applies to DT4R.

SI had a slight impact on the respondents' intention to use DT4R, in line with previous findings [23], for example, in a study of the use of an exergame-based telerehabilitation system by patients and health care professionals [19]. It has been suggested that SI has a higher impact in mandatory contexts, when the use of a device is imposed [14]. Some studies also found that SI increased in specific contexts, such as the use of diabetic footwear by patients [15]. Moreover, we found studies that reported a high impact of SI; those studies defined SI as the "degree of support for one's use of technology from individuals or groups that one considers important to oneself" [42], rather than the opinion of significant others on the technology itself.

To our knowledge, few studies based on the UTAUT model have compared the influence of SI through the lens of culture. In another continent, Akiogbe et al [36] showed that SI has a significantly stronger impact on Chinese students than on Japanese students regarding the use of mobile health (mHealth). They noted that in collectivist societies, SI plays a stronger role [36,47] as individuals rely on group norms and institutional endorsements. Metallo et al [48] showed that cultural values may hinder the use of health care technologies due to social norms. The exploratory subgroup analysis found that the impact of SI was not modified by European subregions. However, the subgroup sample was relatively small; therefore, the results

XSL•FO
RenderX

should be considered with caution. Further studies could aim to deepen the understanding of the impact of culture on SI.

Furthermore, European countries differ in their institutions, laws, and ethical principles [49], and these differing institutional contexts may affect acceptance.

Limitations and Future Work

Our model explained approximately 60% to 70% of the variance in intention to use and 55% to 67% of the variance in use, which corresponds to the expected performance of the UTAUT model at its development [12,14] and its current use thereafter. To the best of our knowledge, models that find higher proportions of explained variance are often modified from the original UTAUT model by adding parameters [17,50]. For example, Ngusie et al [17] explained 84.5% of the variance in intention to use with their extended UTAUT3 model, in which they introduced additional constructs such as attitude as a mediator and technology anxiety and digital technology self-efficacy as exogenous variables. This raises the question of the risk of overfitting. Adding parameters may increase the explained variance but limit the generalizability of the results. It also raises the question of the difficulty in respecting the consistency of the UTAUT model, which should be adapted to a given context without being denatured. The UTAUT model was not developed with the aim of explaining all the variance in a use behavior; rather, it was built to be robust. It offers an observational lens, which may be enriched to ensure a comprehensive understanding of professionals' behavior toward DT4R. That is why we chose to use the UTAUT model without adding additional parameters.

Some fit indices (CFI and TLI) were slightly below commonly cited thresholds. In accordance with recommendations from the methodological literature, these indices were interpreted as guidelines rather than strict decision thresholds. The results were consistent with the literature in terms of the weights of major determinants.

Because of the convenience and snowball sampling methods used to obtain the largest possible sample of European respondents, the sample may not be representative. However, to obtain a representative sample, data on the epidemiological characteristics of rehabilitation professionals would have been necessary; to our knowledge, such data are not available at the European level. We aimed to limit this impact by translating the survey into numerous European languages and disseminating it through national coordinators, national and local professionals, and personal networks. This strategy resulted in responses from 30 out of 43 countries. Nonetheless, the respondents may have had a greater interest in DT4R than the general population of rehabilitation professionals, which could affect the behavioral components of acceptance, especially the intention to use. By disseminating the survey in 20 languages through personal and professional networks, we aimed to reach as varied a sample as possible. Responses were obtained from people with different levels of exposure to and experience with a given technology, but the years of experience with the technology were not gathered in our survey. Further studies could include more precise characterization of exposure levels and experience and propose appropriate strategies according to levels of training and experience to facilitate implementation, as done by Nicora et al [23]. In that Italian study, respondents were stratified according to their years of experience in robot-assisted rehabilitation.

Some studies that combined the perspectives of professionals and patients on a given technology have shown that their views may differ. Seinsche et al [19] showed that health care professionals underestimated the ability of patients to properly set up and use the studied exergame-based telerehabilitation system. Therefore, further studies aiming to evaluate the opinions of children and their relatives would be enlightening, especially given that the 3 DT4R categories, like those studied by Seinsche et al [19], and in contrast with digital technologies such as imaging or shared medical record software, are likely to be directly used by patients and families, especially in the context of remote rehabilitation and telehealth.

Some of the standardized coefficients, although statistically significant, were small. The results should be explored further by additional studies.

We conducted subgroup analyses to improve understanding of the impact of effort expectancy, ease of access, European subregions, and professions. However, the subgroup samples were small; therefore, the results should be interpreted with caution.

Conclusions

This study modeled acceptance based on the UTAUT framework for 3 categories of DT4R—RobTS, VR-AVG, and T&Apps—in pediatric motor rehabilitation. Data were gathered from more than 500 respondents across Europe, and responses were compared across the 3 categories. We then proposed actions to improve acceptance and, consequently, the implementation of DT4R. Within the framework of the UTAUT model, access and the conditions required for easy access emerged as prerequisites for evaluating DT4R acceptance, regardless of the type of DT4R. It is known that offering access to a technology is not sufficient to ensure its implementation within a clinical facility. However, rehabilitation centers must provide and assess access conditions before assessing behavioral barriers and facilitators. Acceptance of DT4R was positively influenced by PE, the main determinant of intention to use for all 3 DT4R categories, outweighing the other determinants. These findings suggest that, regardless of the type of DT4R being implemented in a clinical setting, improving acceptance and bridging the gap between access and effective use require efforts to demonstrate and communicate the effectiveness of the device. To achieve this aim, the elaboration of evidence-based clinical guidelines explicating indications and protocols for use (including, for instance, session length and frequency) is warranted. DT development would benefit from being grounded in the scientific evidence of what works and what does not. The weights of SI and EE on intention to use were slight. In our study, within the framework of the UTAUT model, no acceptance barrier was linked to the use of DT4R in children. Gathering families' views may be useful for the implementation of RobTS. T&Apps may be assets for involving families in their child's rehabilitation. Limitations included self-reported measures and the use of convenience and snowball sampling methods. A large and diverse European sample was chosen to identify actionable and generalizable

XSL•FO
RenderX

levers; however, this should be subsequently adapted to local contexts. Further studies should focus on children's and families' points of view.

Acknowledgments

The authors are grateful to all the people who contributed to the development and refinement of the survey, helped in distributing it, and to all participants who took the time to respond. They thank the EACD and its national coordinators for their support in designing and disseminating the survey. The authors also thank Johanna Robertson, PT, PhD, for her assistance with language editing.

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the Generative AI Delegation Taxonomy (GAIDeT, 2025), the following tasks were delegated to GAI tools under full human supervision: code optimization, translation, and reformatting. The GAI tool used was ChatGPT 4.1 and 4.5. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research or authorship of this paper. Financial support for the publication will be provided by the BEaChild research group.

Data Availability

Data are available upon request to the corresponding author.

Authors' Contributions

Conceptualization: JMG (lead), JDLC (equal), CP (equal), S Brochard (equal)
Data curation: JMG (lead), JDLC (equal), CP (equal), S Brochard (supporting)
Formal analysis: JMG (lead), JDC (equal), CP (equal), S Brochard (supporting)
Investigation: CP (lead), JMG (supporting), S Bekteshi (supporting), JDLC (supporting), S Brochard (supporting), EM (supporting), AIG (supporting), GC (supporting), CJN (supporting), MK (supporting)
Methodology: JMG (lead), JDLC (equal), CP (equal), S Brochard (supporting)
Software: JMG (lead), JDLC (equal)
Supervision: CP (lead), JDLC (equal)
Visualization: JMG (lead), JDLC (equal), CP (equal)
Writing – original draft: JMG (lead), JDLC (equal), CP (equal), S Brochard (supporting)
Writing – review & editing: JMG (lead), S Bekteshi (supporting), JDLC (supporting), CP (supporting), S Brochard (supporting), EM (supporting), AIG (supporting), GC (supporting), CJN (supporting), MK (supporting)

Conflicts of Interest

None declared.

References

1. Sacaze E, Garlantezec R, Rémy-néris O, et al. A survey of medical and paramedical involvement in children with cerebral palsy in Britanny: preliminary results. Ann Phys Rehabil Med 2013 May;56(4):253-267. [doi: 10.1016/j.rehab.2012.11.003] [Medline: 23312436]
2. Mihaylov SI, Jarvis SN, Colver AF, Beresford B. Identification and description of environmental factors that influence participation of children with cerebral palsy. Dev Med Child Neurol 2004 May;46(5):299-304. [doi: 10.1017/s0012162204000490] [Medline: 15132259]
3. Burin-Chu S, Baillet H, Leconte P, Lejeune L, Thouvarecq R, Benguigui N. Effectiveness of virtual reality interventions of the upper limb in children and young adults with cerebral palsy: a systematic review with meta-analysis. Clin Rehabil 2024 Jan;38(1):15-33. [doi: 10.1177/02692155231187858] [Medline: 37499213]
4. Castelli E, Beretta E, De Tanti A, et al. Robot-assisted rehabilitation for children with neurological disabilities: results of the Italian consensus conference CICERONE. NeuroRehabilitation 2022;51(4):665-679. [doi: 10.3233/NRE-220036] [Medline: 36530098]
5. Ghai S, Ghai I. Virtual reality enhances gait in cerebral palsy: a training dose-response meta-analysis. Front Neurol 2019;10:236. [doi: 10.3389/fneur.2019.00236] [Medline: 30984095]
6. Liu C, Wang X, Chen R, Zhang J. The effects of virtual reality training on balance, gross motor function, and daily living ability in children with cerebral palsy: systematic review and meta-analysis. JMIR Serious Games 2022 Nov 9;10(4):e38972. [doi: 10.2196/38972] [Medline: 36350683]

XSL•FO
RenderX

7. Wang N, Liu N, Liu S, Gao Y. Effects of nonimmersive virtual reality intervention on children with spastic cerebral palsy: a meta-analysis and systematic review. Am J Phys Med Rehabil 2023 Dec 1;102(12):1130-1138. [doi: 10.1097/PHM.0000000000002321] [Medline: 37535642]
8. Pons C, Brochard S, Grigoriu A, et al. Digital technologies for motor rehabilitation in children: protocol for a cross-sectional European survey. BMJ Open 2023 Apr 25;13(4):e069034. [doi: 10.1136/bmjopen-2022-069034] [Medline: 37185191]
9. Phelan I, Carrion-Plaza A, Furness PJ, Dimitri P. Home-based immersive virtual reality physical rehabilitation in paediatric patients for upper limb motor impairment: a feasibility study. Virtual Real 2023;27(4):3505-3520. [doi: 10.1007/s10055-023-00747-6] [Medline: 36686613]
10. Alonazi A. Effectiveness and acceptability of telerehabilitation in physical therapy during COVID-19 in children: findings of a systematic review. Children (Basel) 2021 Nov 29;8(12):1101. [doi: 10.3390/children8121101] [Medline: 34943295]
11. Sulzer J, Karfeld-Sulzer LS. Our child's TBI: a rehabilitation engineer's personal experience, technological approach, and lessons learned. J NeuroEngineering Rehabil 2021 Dec;18(1):59. [doi: 10.1186/s12984-021-00862-y]
12. Holden RJ, Karsh BT. The technology acceptance model: its past and its future in health care. J Biomed Inform 2010 Feb;43(1):159-172. [doi: 10.1016/j.jbi.2009.07.002] [Medline: 19615467]
13. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. Manage Sci 1989 Aug;35(8):982-1003. [doi: 10.1287/mnsc.35.8.982]
14. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]
15. Hemler SL, Sommerich CM, Correia JC, Pataky Z. User perceptions of intelligent offloading diabetic footwear. Front Endocrinol (Lausanne) 2024;15:1380525. [doi: 10.3389/fendo.2024.1380525] [Medline: 39170738]
16. Alameraw TA, Asemahagn MA, Gashu KD, et al. Intention to use telemonitoring for chronic illness management and its associated factors among nurses and physicians at public hospitals in Bahir Dar, Northwest Ethiopia: using a modified UTAUT2 model. Front Health Serv 2025;5:1460077. [doi: 10.3389/frhs.2025.1460077] [Medline: 40115328]
17. Ngusie HS, Kassie SY, Zemariam AB, et al. Understanding the predictors of health professionals' intention to use electronic health record system: extend and apply UTAUT3 model. BMC Health Serv Res 2024 Aug 3;24(1):889. [doi: 10.1186/s12913-024-11378-1] [Medline: 39097725]
18. Slegers K, Krieg AM, Lexis MAS. Acceptance of 3D printing by occupational therapists: an exploratory survey study. In: Hilton C, editor. Occup Ther Int 2022;2022:4241907. [doi: 10.1155/2022/4241907] [Medline: 36636037]
19. Seinsche J, de Bruin ED, Saibene E, et al. A newly developed exergame-based telerehabilitation system for older adults: usability and technology acceptance study. JMIR Hum Factors 2023 Dec 7;10:e48845. [doi: 10.2196/48845] [Medline: 38060283]
20. Thanthrige A, Lu B, Sako Z, Wickramasinghe N. Determinants of health care technology adoption using an integrated unified theory of acceptance and use of technology and task technology fit model: systematic review and meta-analysis. J Med Internet Res 2025 Dec 30;27:e64524. [doi: 10.2196/64524] [Medline: 41468595]
21. Yousef CC, Salgado TM, Farooq A, et al. Health care providers' acceptance of a personal health record: cross-sectional study. J Med Internet Res 2021 Oct 21;23(10):e31582. [doi: 10.2196/31582] [Medline: 34569943]
22. Sahan F, Gutermuth A, Müller JA, Muth T, Panitz K, Apolinário-Hagen J. Acceptance and use of digital health technologies among physiotherapists in Germany: a web-based cross-sectional survey. BMC Health Serv Res 2026 Jan 10;26(1):147. [doi: 10.1186/s12913-026-14005-3] [Medline: 41519786]
23. Nicora G, Parimbelli E, Mauro MC, et al. Healthcare practitioners and robotic-assisted rehabilitation: understanding needs and barriers. J NeuroEngineering Rehabil 2025;22(1). [doi: 10.1186/s12984-025-01593-0]
24. Kaimara P, Oikonomou A, Deliyannis I. Could virtual reality applications pose real risks to children and adolescents? A systematic review of ethical issues and concerns. Virtual Real 2022;26(2):697-735. [doi: 10.1007/s10055-021-00563-w] [Medline: 34366688]
25. Newman CJ. Do children with cerebral palsy dream of electric legs? The effects of robot-assisted gait training. Dev Med Child Neurol 2024 Aug;66(8):972-973. [doi: 10.1111/dmcn.15863] [Medline: 38303160]
26. Mensah-Gourmel J, Bekteshi S, Brochard S, et al. Digital technologies for pediatric rehabilitation: current access and use in the European RehaTech4child survey. Eur J Phys Rehabil Med 2024 Dec;60(6):970-979. [doi: 10.23736/S1973-9087.24.08559-9] [Medline: 39387853]
27. LOI n° 2012-300 du 5 mars 2012 relative aux recherches impliquant la personne humaine (1) [Article in French]. Légifrance (République Française). URL: https://www.legifrance.gouv.fr/loda/id/JORFTEXT000025441587 [accessed 2026-05-29]
28. Ordonnance n° 2016-800 du 16 juin 2016 relative aux recherches impliquant la personne humaine [Article in French]. Légifrance (République Française). URL: https://www.legifrance.gouv.fr/loda/id/JORFTEXT000032719520 [accessed 2026-05-29]
29. Décret n° 2016-1537 du 16 novembre 2016 relatif aux recherches impliquant la personne humaine [Article in French]. Légifrance (République Française). URL: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033394083 [accessed 2026-05-29]

XSL•FO
RenderX

30. Glegg SMN, Holsti L, Velikonja D, Ansley B, Brum C, Sartor D. Factors influencing therapists' adoption of virtual reality for brain injury rehabilitation. Cyberpsychol Behav Soc Netw 2013 May;16(5):385-401. [doi: 10.1089/cyber.2013.1506] [Medline: 23713844]
31. Levac D, Glegg S, Colquhoun H, Miller P, Noubary F. Virtual reality and active videogame-based practice, learning needs, and preferences: a cross-Canada survey of physical therapists and occupational therapists. Games Health J 2017 Aug;6(4):217-228. [doi: 10.1089/g4h.2016.0089] [Medline: 28816511]
32. Kerr A, Smith M, Reid L, Baillie L. Adoption of stroke rehabilitation technologies by the user community: qualitative study. JMIR Rehabil Assist Technol 2018 Aug 17;5(2):e15. [doi: 10.2196/rehab.9219] [Medline: 30120086]
33. Li L, Tyson S, Weightman A. Professionals' views and experiences of using rehabilitation robotics with stroke survivors: a mixed methods survey. Front Med Technol 2021;3:780090. [doi: 10.3389/fmedt.2021.780090] [Medline: 35047969]
34. A cross-European survey «Rehatech4child». SurveyMonkey. URL: http://www.surveymonkey.com/r/Rehatech4childTEST [accessed 2026-05-15]
35. Stone BM. The ethical use of fit indices in structural equation modeling: recommendations for psychologists. Front Psychol 2021;12:783226. [doi: 10.3389/fpsyg.2021.783226] [Medline: 34887821]
36. Akiogbe O, Feng H, Kurata K, et al. Cross-analysis of mHealth social acceptance among youth: a comparative study between Japan and China. Behav Sci (Basel) 2025 Feb 14;15(2):213. [doi: 10.3390/bs15020213] [Medline: 40001844]
37. Standard country or area codes for statistical use (M49). United Nations Statistics Division (UNSD). URL: https://unstats.un.org/unsd/methodology/m49/ [accessed 2026-05-15]
38. Aranda JM, Moreno-Fergusson ME, Guerrero WJ, et al. Technology acceptance and use among nursing staff in Latin American hospitals: a mixed methods study. Int J Nurs Stud Adv 2025 Jun;8:100290. [doi: 10.1016/j.ijnsa.2024.100290] [Medline: 39872435]
39. Azam M, Bin Naeem S, Kamel Boulos MN, Faiola A. Modelling the predictors of mobile health (mHealth) adoption among healthcare professionals in low-resource environments. Int J Environ Res Public Health 2023 Nov 26;20(23):7112. [doi: 10.3390/ijerph20237112] [Medline: 38063542]
40. Ahmed A, Ho CW, Grant Y, Archer S, Carrington EV. Acceptability of digital health interventions in perioperative care: a systematic review and narrative synthesis of clinician perspectives. BMJ Open 2025 Mar 15;15(3):e086412. [doi: 10.1136/bmjopen-2024-086412] [Medline: 40090692]
41. Sultan H, Najjar S, Awartani F, Abu-Odah H. The unmet needs of adopting clinical decision support system among physicians in Palestinian hospitals. PLOS One 2024;19(12):e0310765. [doi: 10.1371/journal.pone.0310765] [Medline: 39666711]
42. Mengying L, Dongquan F, Lin L, et al. Pathway analysis of clinical nurse educator's intention to use virtual reality technology based on the UTAUT model. Front Public Health 2024;12:1437699. [doi: 10.3389/fpubh.2024.1437699] [Medline: 39618955]
43. Gilardi F, De Falco F, Casasanta D, et al. Robotic technology in pediatric neurorehabilitation. a pilot study of human factors in an Italian pediatric hospital. Int J Environ Res Public Health 2020 May 17;17(10):3503. [doi: 10.3390/ijerph17103503] [Medline: 32429562]
44. Ryan AK, Worthy P, Miller L, Johnston LM. I can decide MyGOALS! Co-designing an online tool for child-led goal setting and evaluation. Disabil Rehabil Assist Technol 2026 Feb;21(2):446-460. [doi: 10.1080/17483107.2025.2522782] [Medline: 40560899]
45. Te Boekhorst KE, Kuipers SJ, Ribbers GM, Cramm JM. Exploring rehabilitation patients' perspectives on what matters for the adoption of home-based rehabilitation technology: Q-methodology study. JMIR Rehabil Assist Technol 2025 Jul 9;12:e71515. [doi: 10.2196/71515] [Medline: 40632833]
46. Moreira H, Góis AC, Pereira AM, et al. Parents' acceptability of blended psychological interventions for children with emotional disorders. Internet Interv 2023 Dec;34:100687. [doi: 10.1016/j.invent.2023.100687] [Medline: 38023966]
47. Muthukrishna M, Schaller M. Are collectivistic cultures more prone to rapid transformation? Computational models of cross-cultural differences, social network structure, dynamic social influence, and cultural change. Pers Soc Psychol Rev 2020 May;24(2):103-120. [doi: 10.1177/1088868319855783] [Medline: 31253070]
48. Metallo C, Agrifoglio R, Lepore L, Landriani L. Explaing users' technology acceptance through national cultural values in the hospital context. BMC Health Serv Res 2022 Jan 17;22(1):84. [doi: 10.1186/s12913-022-07488-3] [Medline: 35039014]
49. Bente BE, Van Dongen A, Verdaasdonk R, van Gemert-Pijnen L. eHealth implementation in Europe: a scoping review on legal, ethical, financial, and technological aspects. Front Digit Health 2024;6:1332707. [doi: 10.3389/fdgth.2024.1332707] [Medline: 38524249]
50. Farooq MS, Salam M, Jaafar N, et al. Acceptance and use of lecture capture system (LCS) in executive business studies: extending UTAUT2. Interact Technol Smart Educ 2017 Nov 20;14(4):329-348. [doi: 10.1108/ITSE-06-2016-0015]

Abbreviations

CFI: comparative fit index
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
DT4R: digital technologies for rehabilitation

XSL•FO
RenderX

EACD: European Academy of Childhood-onset Disability
EE: effort expectancy
mHealth: mobile health
PE: performance expectancy
RMSEA: root mean square error of approximation
RobTS: robotics and treadmill systems
SI: social influence
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
T&Apps: telehealth and apps
TLI: Tucker-Lewis index
UTAUT: Unified Theory of Acceptance and Use of Technology
VR-AVG: virtual reality and active video gaming

Edited by M Balcarras; submitted 24.Jul.2025; peer-reviewed by FA Etindele Sosso, K Mouloudj; revised version received 24.Mar.2026; accepted 10.Apr.2026; published 22.Jun.2026.

Please cite as:
Mensah Gourmel J, Brochard S, Bekteshi S, Monbaliu E, Cornec G, Grigoriu AI, Newman CJ, Konings M, De La Cruz J, Pons C Explaining the Use Behavior of Digital Technologies in Pediatric Rehabilitation: Structural Equation Modeling Analysis of a Cross-Sectional European Survey
J Med Internet Res 2026;28:e80355
URL: https://www.jmir.org/2026/1/e80355
doi:10.2196/80355

© Johanne Mensah Gourmel, Sylvain Brochard, Saranda Bekteshi, Elegast Monbaliu, Gwenaël Cornec, Anca Irina Grigoriu, Christopher J Newman, Marco Konings, Javier De La Cruz, Christelle Pons. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Health Communication Campaign Performance During the HEALing Communities Study: Cross-Sectional Examination of Digital Advertising Methods

Nicky Lewis[1], PhD; Jennifer Reynolds[2], MPH; Diane Krause[2], MS, MPH; Philip M Reeves[2], PhD; Jamie Luster[3], MPH; Michael D Stein[4], MD; Sharon L Walsh[5], PhD; Amy Farmer[3], MA; Michelle R Lofwall[5], MD; Monica F Roberts[5], PharmD; Hilary L Surratt[5], PhD; Brooke N Crockett[3], MPH; Kara Stephens[2], MPH, MSA; Kelli Bursey[2], MPH; Kristin Mattson[2], MPH, MCHES; Michael D Slater[3], PhD

[1]Indiana University Bloomington, 601 E Kirkwood Avenue, Bloomington, IN, United States

[2]Health and Resilience Strategies, Oak Ridge Associated Universities, Oak Ridge, TN, United States

[3]The Ohio State University College of Medicine, Columbus, OH, United States

[4]Boston University School of Public Health, Boston, MA, United States

[5]University of Kentucky, Lexington, KY, United States

Corresponding Author:
Nicky Lewis, PhD
Indiana University Bloomington, 601 E Kirkwood Avenue, Bloomington, IN, United States

Abstract

Background: Research on the effectiveness of digital health campaign strategies is lacking. Understanding performance outcomes is essential for the successful implementation of campaigns. Two studies examined platforms, tactics, and content of digital health campaigns using paid media performance data.

Objective: This analysis compared 2 digital advertising methods (social media and banner or display) using click-through rate (CTR) and cost-per-click (CPC) as performance measures. Performance differences by state, community type, message approach, format, and image type were assessed. CTR and CPC served as measures in determining performance differences between social media and banner or display.

Methods: This cross-sectional secondary analysis examined campaign performance for the HEALing (Helping to End Addiction Long-Term) Communities Study, which served 85,875,105 impressions. Data were collected from media buy reports, entered into templates that included method (display or banner and social media) and key performance indicators (impressions, clicks, and media spend), and CTR and CPC were calculated. Study 1 assessed differences in CTR and CPC for social media and banner or display by state (KY, NY, MA, and OH) and community type (urban and rural). Study 2 assessed differences in CTR for social media and banner or display by state (KY, NY, MA, and OH), community type (urban and rural), message approach (testimonial and information-based), format (motion graphic or graphics interchange format, video, and static image), and image type (local and stock). Separate analyses were conducted for each advertising method.

Results: Study 1 found significant differences between advertising methods, where social media had higher CTR compared to banner or display. Social media had a significant main effect for state, where OH had the highest CTR. There was a statistically significant difference in CPC based on advertising method, where social media had a lower CPC compared to banner or display. Social media had a significant main effect for state, where OH had the lowest CPC. Banner or display had a significant main effect for state and community type, where OH and urban communities had the highest CPC. Study 2 found significant differences between advertising methods, where social media had higher CTR than banner or display. For social media, urban communities, static format, and local spokespersons had the highest CTR. There were significant differences between all pairs of states, where OH had the highest CTR. For display or banner, static format and local spokespersons had the highest CTR.

Conclusions: This analysis provides guidance for digital health campaigns. It examined the performance of opioid use disorder campaigns using CTR and CPC measures, demonstrating utility in future campaign evaluations. Social media was more related to stimulating responses to campaign messages compared to banner or display. State-to-state variations emphasized the importance of message pilot testing. Using local spokespersons versus stock spokespersons is recommended.

Trial Registration: ClinicalTrials.gov NCT04111939; https://clinicaltrials.gov/study/NCT04111939

(J Med Internet Res 2026;28:e75458) doi:10.2196/75458

XSL•FO
RenderX

KEYWORDS

communications; opiate overdose; social stigma; naloxone; methadone; buprenorphine evidence-based practice

Introduction

Overview

Studies that examine the platforms (ie, banner or display and social media), tactics (ie, campaign strategy), and content (ie, campaign materials) that influence health communication campaign effectiveness are limited [1,2]. There are few published empirical reports on the comparative effectiveness of online advertising platforms in regard to online public health messaging content [3]. The current research effort examines the online advertising tactics, platforms, and messaging across the communication campaigns of the HEALing (Helping to End Addiction Long-Term) Communities Study (HCS) using paid media performance data.

Background and Rationale

Examinations of the effectiveness of health communication campaigns are minimal and essentially absent in the domain of opioid use disorder (OUD); see Rath et al [4] for an exception. Most research on the effectiveness of health communication campaigns is centered on tobacco cessation efforts, including the Tips from Former Smokers campaign [5]. Durkin et al [6] highlight the potential of digital and social media in enhancing campaign effectiveness. Digital and social media may be particularly important when producing opioid-related health campaigns because of the stigmatized nature of opioid use. This is because digital and social media can often be consumed in private or semiprivate locations and on one's individual media device (eg, tablet and smartphone), reducing the likelihood of others observing the content.

Most opioid-related campaigns focus on prevention or stigma reduction; some target the overdose reversal medication naloxone, but few focus on treatment with medications for opioid use disorder (MOUD) [7]. These campaigns frequently communicate to the general public rather than targeting messaging to priority audiences who could most benefit, such as individuals and loved ones affected by OUD and sectors that may influence local programs and policies regarding substance use, including criminal or legal services, emergency medical services, local government officials, faith-based leaders, and media [8]. Geographical, regional, and cultural differences can also contribute to varying outcomes associated with campaigns as well as the way campaigns are managed and the materials used [9]. Moreover, many opioid-related health campaigns fail to conduct formative research to inform audience segmentation and campaign messaging tactics and content or to use evidence-based practices for campaign development, implementation, and evaluation [7].

Although data regarding widespread opioid-related campaign effectiveness are limited, some evidence can be gleaned from smaller-scale message testing studies. Visual campaigns about stigmatizing language and MOUD treatment, combined with narrative vignettes from either people with lived experience experiencing OUD or of health care administrators, resulted in a significant reduction in stigmatizing attitudes among a large survey of health care providers [10]. Focus groups of young adults who use illicit drugs in Illinois provided evidence that effective campaigns should highlight the risk of fentanyl exposure, fill in gaps about fentanyl knowledge, form a personal connection in messaging, and use realistic details and images [11]. Among a sample of college students without personal OUD experience, a blog-style testimonial about an individual in recovery who used MOUD followed by abstinence-based treatment resulted in strong perceptions of the character's fluidity and predicted stigma reduction [12]. These findings suggest the importance of personalized messaging with a testimonial-driven element, consistent with the literature on engagement with digital advertisements in general. Both personal relevance, defined as the process through which viewers determine if an advertisement or message relates to their personal goals or values, and social relevance, a situational element involving likes, comments, or other engagement with advertisements or messages by one's social network, can lead to increased engagement [13].

HCS Campaigns

The HCS was a multisite, parallel-group, community-level cluster randomized wait-listed trial supported by the National Institutes of Health (NIH) and the Substance Abuse and Mental Health Services Administration (SAMHSA). The HCS sought to test the effectiveness of an intervention for reducing opioid overdose deaths in 67 rural (n=29) and urban (n=38) communities heavily impacted by the opioid crisis across 4 states (KY, MA, NY, and OH) [8]. The HCS tested the Communities That HEAL (CTH) intervention, which deployed a community-engaged, data-driven process and tailored communication campaigns to assist communities in adopting and implementing evidence-based practices (EBPs) to reduce opioid fatalities [8]. Communities were randomized to either receive the CTH first (wave 1; n=34) or serve as the wait-list control group (wave 2; n=33) to receive the CTH after the HCS comparison period was completed [8]. Randomization only occurred at the community level in the main study, where communities were selected as part of the intervention group or wait-list control group. There was no randomization of the observational data reported in this paper. Complicated by the COVID-19 pandemic, an extreme rise of fentanyl in the illicit drug supply and the limited measurement period, the intervention did not significantly reduce the opioid overdose death rate in intervention communities [14]. However, an estimated total of 483 deaths were averted in the intervention arm (wave 1) compared to the control arm (wave 2).

The HCS communication campaigns aimed to support the adoption of EBPs (1) by improving awareness of and access to naloxone (an overdose reversal medication) and MOUD, (2) by driving demand for these EBPs, and (3) by reducing stigma, a key barrier to treatment initiation and recovery for individuals with substance use disorders [8]. A workgroup of communication experts from the research sites (Boston Medical Center, Columbia University, University of Kentucky, and The

XSL•FO
RenderX

Ohio State University), the HCS Data Coordinating Center (Research Triangle Institute International), the National Institute on Drug Abuse, SAMHSA, and Oak Ridge Associated Universities (ORAU), an outside consultant on the study, conducted formative research and developed a standardized strategy (ie, topics, objectives, and priority audiences) and approach for the campaigns (ie, activities that shared messages with priority audiences). Priority audiences across all 4 sites included local community leaders and influencers, people at risk for an overdose or with OUD, and family and friends of those at risk. There were options at each site for crafting community-tailored messages (ie, including local photos), but all messages contained similar calls to action (eg, Carry naloxone. Save a life).

Objectives

The development, dissemination, and monitoring of HCS campaigns created an opportunity for a more robust analysis of the performance of health communication campaigns, particularly those focused on OUD. Campaign development also involved in-depth formative research conducted by communications experts within multiple sectors, which is considered a best practice but is sometimes limited due to time, financial, and other resource constraints [15]. Thus, the current research effort adds to the existing literature on the effectiveness of health communication campaigns, expanding beyond smaller-scale OUD campaign message testing studies [11] and accounting for the exposure and engagement outcomes associated with the implementation of multigeographical site and multicommunication platform OUD campaigns [10]. Considering the above factors, we planned our assessment around 2 key campaign flight components: the exposure and engagement outcomes associated with specific tactics used in standard or template-based message dissemination (study 1) and the exposure and engagement outcomes associated with specific features of customizable, community-tailored campaign messages (study 2).

Considering this, the following research questions (RQs) were posed for the HCS tactics dataset (study 1):

- RQ1: How did click-through rate (CTR) differ based on the type of advertising method (social media and banner or display)?
- RQ2: How did CTR for social media advertisements differ across states (KY, MA, NY, and OH) and community types (urban and rural)?
- RQ3: How did CTR for banner or display advertisements differ across states (KY, MA, NY, and OH) and community types (urban and rural)?
- RQ4: How did cost-per-click (CPC) differ based on the type of advertising method (social media and banner or display)?
- RQ5: How did CPC for social media advertisements differ across states (KY, NY, and OH) and community types (urban and rural)?
- RQ6: How did CPC for banner or display advertisements differ across states (KY, NY, and OH) and community types (urban and rural)?

The following RQs were posed for the HCS community-tailored messages dataset (study 2):

- RQ1: How did CTR differ based on the type of advertising method (social media and banner or display)?
- RQ2: How did CTR for social media advertisements differ based on (1) approach (testimonial and information-based), (2) format (motion graphic or graphics interchange format [GIF], video, and static), (3) image (local and stock), (4) state (KY, MA, NY, and OH), and (5) community type (urban and rural)?
- RQ3: How did CTR for banner or display advertisements differ based on (1) approach (testimonial and information-based), (2) format (motion graphic or GIF, video, and static), (3) image (local and stock), (4) state (KY, MA, and OH), and (5) community type (urban and rural)?

Methods

Overview

The study start date was October 23, 2019, and the estimated study completion date was September 30, 2025.

The STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) cross-sectional checklist was used in the writing of this manuscript [16]. This cross-sectional secondary analysis examines the relationships between different paid digital advertising strategies from a standard or template-based set of advertisements (study 1) and message content factors from a customizable, community-tailored set of advertisements (study 2) on engagement key performance indicators (KPIs) from communication campaigns developed for the HCS. KPIs are reported from each community and site as a whole, not as part of the original randomization of the communities into their respective intervention group or wait-list control group as part of the main study.

The outcomes presented here build upon a conceptual framework of KPIs that uses CTR and CPC as performance outcomes [1]. We compared 2 digital advertising methods—social media and banner or display—measuring cost and engagement KPIs. In study 1: HCS tactics, we examined the differences in CTR and CPC for each advertising method—social media or banner or display—across the states of KY, NY, and OH (Note: cost data from MA were not used because of inconsistencies, where multiple communities were grouped together for their advertising spend), and community type—urban and rural—separately. In Study 2: HCS community-tailored messages, no cost data were available, so only differences in CTR were examined within each advertising method (social media and banner or display). The dataset included detailed information about the characteristics of messages, so each analysis included one of the following independent variables: approach (testimonial and information-based), format (motion graphic or GIF, video, and static), image type (local and stock), state (KY, MA, OH, NY; Note: NY was excluded from banner or display analyses due to small sample size), and community type (urban and rural).

XSL•FO
RenderX

Setting

The research site and ORAU staff collaborated closely with community coalitions and "champions" in each HCS community across KY, MA, NY, and OH to assess local needs and communication preferences, review local OUD data, and tailor campaign audience segments, materials, and dissemination strategies. Champions were coalition members who facilitated communication and activities between the coalitions, the HCS research teams, and partner organizations. HCS staff also provided technical assistance and training throughout the study to support coalitions in developing, implementing, and evaluating their campaigns. In total, the HCS designed, disseminated, and evaluated 264 local, community-tailored, and data-informed health communication messages about 5 different OUD topics from April 2020 to December 2023 (Table 1).

Table . Characteristics of HEALing[a] Communities Study health communication campaign messages disseminated in wave 1 and wave 2 communities across KY, MA, NY, and OH.

Wave	Communities, n	Total campaigns, n	Topics	Objectives	Date range
1	33	165	• Naloxone • Stigma • MOUD[b] treatment • Retention in MOUD treatment • Community choice[c]	• Increase demand for naloxone • Increase access to and availability of naloxone • Increase demand for, access to, and availability of MOUD • Increase MOUD prescribing • Improve acceptance and support for those in MOUD treatment • Increase demand for, access to, and availability of MOUD • Increase MOUD prescribing • Increase retention in MOUD treatment for OUD[d] • Improve support for those in MOUD treatment	April 2020 to February 2022
2	33	99	• Naloxone and fentanyl • MOUD treatment and stigma • Community choice[c]	• Increase demand for naloxone • Increase access to and availability of naloxone • Increase knowledge about the prevalence and risk of fentanyl • Increase demand for, access to, and availability of MOUD • Increase MOUD prescribing • Improve acceptance and support for those in MOUD treatment	November 2022 to December 2023

[a]HEAL: Helping to End Addiction Long-Term.

[b]MOUD: medication for opioid use disorder.

[c]"Community choice" indicates that communities had the option to repeat one of the prior campaigns based on their specific needs and preferences.

[d]OUD: opioid use disorder.

XSL•FO
RenderX

Study 1: HCS Tactics Design

Coalitions used a study-provided template to identify paid and unpaid dissemination tactics they wished to implement for each HCS campaign. These plans included specifics about budget, types of advertising, and targeting parameters. Given the study campaigns aimed to drive traffic to community web pages containing OUD education and treatment information, all campaigns were optimized to generate "web page clicks" (engagement) over impressions (exposure). The study required all advertising to be, as much as possible, limited to the geographic boundaries or zip codes of each community. This requirement, and the relatively small budgets allocated for paid media, prevented communities from using "mass media" advertising, such as cable television. Research sites provided funding to communities for campaign dissemination. Most communities chose to purchase a mix of traditional (eg, print, local radio, local television, and billboards) and digital (eg, social media, digital display, and streaming television) advertising, although some communities did not purchase advertising for every campaign.

Most communities elected to target advertisements to similar audiences, including people with lived experience and community leaders. Communities used common digital media platforms and diverse audience targeting tactics to reach these audiences with advertisements; however, most communities elected to target campaign advertising using the demographics (eg, age and gender) associated with higher overdose deaths in each county and selected the media channels most used by those demographic groups. The second-most common targeting was behavioral, involving sharing advertisements with those who searched for or showed interest in drugs, alcohol, addiction, and treatment information.

To facilitate the purchasing and placement of paid advertisements, ORAU, on behalf of each research site, subcontracted with 4 media buyers, each with experience and expertise in 1 of the 4 state markets (KY, MA, NY, and OH). As part of their contracts, media buyers were required to provide communities and study staff with detailed paid advertising performance reports after each campaign, which formed the basis for the study data. The duration of communities' paid advertising flights ranged from 5 to 133 days, with an average length of 59 days. The paid media costs per community per campaign ranged from US $75 to US $44,002, with an average cost of US $4630. Across all campaigns, the total impression volume was 85.9 million, and the total number of clicks to the HCS web pages was 369,679. Researchers could not estimate the number of specific audience members in each community (eg, number of people who use drugs in a specific community and number of community leaders).

Study 2: HCS Community-Tailored Messages Design

HCS provided communities with a set of customizable social media and print materials (assets) and a website linking community members to additional information and local resources related to each campaign theme. The campaign assets were designed with "core" (unchangeable) messages and calls to action. Campaign assets could be locally tailored with options including imagery (local faces and places), content (medications, language, and personal testimonials or quotes), product (type of material, format, size, and design specifications), and destination (local provider, website, and phone number).

Data Sources or Measurement

The method for collecting and analyzing paid media data across research sites involved a collaborative and iterative approach. Health communication experts from each site worked closely with media buyers to ensure strategic media placements and accurate data collection. Upon conclusion of the study, investigators sought to compare digital media performance across states and communities. To facilitate this process, a data manager developed a Microsoft Excel template that stored media buy method information (ie, display or banner and social media) and KPIs reported across all campaigns. KPIs included impressions, clicks, and media spend, from which CTR and CPC could be calculated (if not already provided in media buy reports). The template was refined through iterative feedback from staff at each site, ensuring that it was comprehensive and adaptable to the varied types of data collected across sites. Some KPIs, such as video completions, video completion rate, and reach, were not included in the final template, given that they were not consistently used or reported across states. This template, complemented by a codebook for data standardization, was shared across all sites. Study staff from each state entered media buy data from their community reports into this standard template, such that data could easily be combined and compared. The data manager reviewed data for errors, cleaned to address inconsistencies, and confirmed with site staff for accuracy.

Variables

All KPIs used in these studies are reported based on the dates the campaigns were implemented in the HCS communities. As indicated earlier, the data were compiled, and the following variables were examined during the analysis.

In the HCS tactics dataset (study 1), CPC and CTR were used as dependent variables. CPC was calculated by dividing the total spend on the paid media tactic or platform within specific dates by the total number of clicks. CTR was calculated by dividing the total number of clicks by the total number of impressions. The independent variables were community type (urban and rural) and state (KY, NY, and OH; Note: MA was excluded from these analyses due to inconsistencies in their cost data, where multiple communities were grouped together for their advertising spend). Separate analyses were conducted for each advertising method (banner or display and social media). Banner or display advertising included cross-platform displays, device ID targeting, and native advertisements for seamless integration, along with search retargeting and in-app videos to engage users based on behavior and context. Strategies also leveraged standard display, preroll behavioral, and preroll contextual advertisements for targeted visibility. Social media advertising used Facebook for its broad reach and targeting capabilities, including boosted posts, and Instagram for visual engagement. Snapchat and TikTok targeted younger audiences with dynamic content, and YouTube provided a platform for video content.

XSL•FO
RenderX

In the HCS community-tailored messages dataset (study 2), CTR was the only dependent variable. No cost data were available, and as a result, CPC could not be calculated. Community type (urban and rural) and state (KY, MA, and OH; Note: NY was excluded from banner or display analyses due to small sample size) were used as independent variables. Additional message characteristics were also used as independent variables in this study. These characteristics included the approach for conveying the message (testimonial or information-based), the message format (motion graphic or GIF, video, or static image; Note: GIFs were excluded from social media analysis due to small sample size), and the image type (image of a local spokesperson or image of a stock person). Separate analyses were conducted for each advertising method (banner or display and social media).

Statistical Methods

When analyzing the data, only cases with data for every variable of interest were included, resulting in different sample sizes for each analysis. Two datasets, HCS tactics (study 1) and HCS community-tailored messages (study 2), were analyzed. Data were not normalized by state or community for any of the analyses in either study. For Study 1, there were no missing data for any of the independent variables. When values for the dependent variables were unavailable, the corresponding media buy was removed from the analysis. Missing data for CTR represented 3.92% (18/459) of cases that met the inclusion criteria. Missing data for CPC represented 6.10% (28/459) of cases that met the inclusion criteria. For Study 2, there were no missing data on any of the independent variables. When values for the dependent variable were unavailable, the corresponding media buy was removed from the analysis. Missing data for CTR represented 0.99% (59/5913) of cases that met the inclusion criteria.

In the HCS tactics dataset (study 1), a Mann-Whitney *U* test was conducted to test differences in median percentage CTR (or median multiplied by 100) based on advertising method (RQ1). A Mann-Whitney *U* test was selected due to violations of normality, homogeneity, and linearity assumptions of parametric tests. Because differences were found, subsequent tests looked at each advertising method individually. Due to violations of normality and homogeneity assumptions in the data structure, Scheirer-Ray-Hare tests were conducted that looked at each advertising method individually (ie, one equation for social media [RQ2] and one for banner or display [RQ3]). Each test examined differences based on 2 independent variables, community type (rural vs urban) and state (KY, NY, OH). Dunn test with Bonferroni corrections was calculated to examine differences between pairs of states. The analytical process was duplicated with CPC as the dependent variable (RQ4, 5, and 6).

In the HCS community-tailored messages dataset (study 2), 3 statistical tests were conducted. CTR was the dependent variable for all tests. First, a Mann-Whitney *U* test was performed with the advertising method (social media vs banner or display) as the independent variable (RQ1). Because differences were found based on advertising method, 2 quantile regression equations were calculated that looked at each advertising method individually (ie, one equation for social media [RQ2] and one equation for banner or display [RQ3]). Quantile regression equations were used due to the nonnormal nature of the data. This test predicts the median percentage CTR (or median multiplied by 100), and when there were multiple levels of an independent variable, we used the level with the largest sample size as the reference group. In the social media regression equation (RQ2), the independent variables were state (KY, NY, OH, and MA), community type (rural vs urban), approach (testimonial or information-based), format (video or static image), and image type (image of a local spokesperson or image of a stock person). In the banner or display regression equation (RQ3), the independent variables were state (KY, OH, and MA), community type (rural vs urban), approach (testimonial or information-based), format (GIF, video, or static image), and image type (image of a local spokesperson or image of a stock person). No interaction terms were specified in either equation.

Ethical Considerations

The HCS study protocol was approved by Advarra Inc (Pro00038088), a single institutional review board. The data analyzed here were observations of public behaviors and not primary analyses of the original study. Advarra Inc allowed for the secondary analyses conducted here to be conducted without additional consent. The data in this study were received from each state's media buyer as aggregated, paid advertisement performance reports of KPIs not linked to any single individual. No individual participants were recruited, participated in, or compensated for this analysis, and as such, no identification of individual participants was possible (ie, data are deidentified).

Results

Study 1

HCS Tactics CTR

An overview of the state-level campaign tactics, audiences, and impressions can be seen in Table 2, and an overview of state-level campaign metric descriptives is reported in Table 3.

XSL•FO
RenderX

Table . Overview of state-level campaign impressions, target audiences, display or banner tactics, and social media tactics for the digital health communication campaigns of the HEALing[a] Communities Study from April 2020 to December 2023.

State	Impressions, n	Audiences	Display or banner[b]	Social media
KY	31,122,896	Community leaders, people with lived experience[c]	Cross-platform display, device ID targeting, display, in-app video, native, preroll, search retargeting	Facebook, Instagram, Snapchat, YouTube
MA	18,555,039	Community leaders, people with lived experience[c], health care providers	Cross-platform display, display, preroll	Facebook, Instagram
NY	27,535,463	Community leaders, people with lived experience[c]	Cross-platform display, preroll contextual	Facebook, Instagram, Snapchat, YouTube
OH	8,661,707	Community leaders, people with lived experience[c], health care providers	Display, preroll behavioral	Facebook, Snapchat, TikTok, YouTube

[a]HEAL: Helping to End Addiction Long-Term.

[b]Tactics definitions: cross-platform display: advertisements displayed across multiple devices and platforms. Device ID targeting: advertisements targeted based on unique device identifiers. Display: banner or visual advertisements shown on websites or apps. In-app video: video advertisements embedded within apps. Native: advertisements that blend with the platform's content and design. Preroll: video advertisements played before the main video content. Preroll behavioral: prevideo advertisements targeted based on user behavior patterns. Preroll contextual: prevideo advertisements tailored to the context of the content being viewed. Search retargeting: advertisements targeted to users based on their online search behavior.

[c]Including people who use drugs, people with opioid use disorder, and their loved ones.

Table . State-level campaign key performance indicators for the digital health communication campaigns of the HEALing[a] Communities Study from April 2020 to December 2023.

State	Total impressions, n	Total clicks, n	Total cost (US $)	Click-through rate (%)	Cost-per-click (US $)
KY	31,122,896	90,843	223,330.30	0.291	2.46
MA	18,555,039	84,455	N/A[b]	0.455	N/A
NY	27,535,463	109,619	265,481.20	0.398	2.42
OH	8,661,707	84,762	92,598.90	0.978	1.09

[a]HEAL: Helping to End Addiction Long-Term.

[b]N/A: not applicable.

The Mann-Whitney *U* test resulted in a statistically significant difference in CTR based on the advertising method (z=9.27; P<.001; r=0.44, 95% CI 0.22%-0.44%; RQ1). Social media (mean 0.84%, SD 1.1%; median 0.54%, IQR 0.15%-1.03%) had a higher CTR when compared to banner or display (mean 0.21%, SD 0.27%; median 0.13%, IQR 0.08%-0.21%). Because the CTR was different for each advertising method, additional tests were conducted that examined each method separately based on state and community type. For social media (RQ2), the Scheirer-Ray-Hare test demonstrated that there was a statistically significant main effect for the state (H(2, 240)=27.16; P<.001). However, there was not a statistically significant main effect for community type (H(1, 240)=0.183; P=.67), and there was not a statistically significant interaction (H(2, 240)=0.163; P=.92). Dunn test with Bonferroni corrections found statistically significant pairwise differences between KY and OH (z=4.60; P<.001) and NY and OH (z=5.08; P<.001; Table 4). OH (mean 1.83%, SD 1.39%; median 1.35%) had a higher median CTR than both KY (mean 0.70%, SD 0.83%; median 0.52%) and NY (mean 0.67%, SD 0.95%; median 0.35%). For banner or display (RQ3), the Scheirer-Ray-Hare test demonstrated that there was not a statistically significant main effect for the state (H(2, 189)=5.78; P=.06) or community type (H(1, 189)=8.53; P=.003), nor was there a statistically significant interaction (H(2, 189)=4.16; P=.12). Dunn test with Bonferroni corrections found no statistically significant pairwise differences between states (Table 4).

XSL•FO
RenderX

Table . Click-through rate descriptives for social media and banner or display by state and community type for the digital health communication campaigns of the HEALing[a] Communities Study from April 2020 to December 2023.

State and community	Social media			Banner or display		
	Advertisements, n	Mean (SD) (%)	Median (95% Mdn CI[b]) (%)	Advertisements, n	Mean (SD) (%)	Median (95% Mdn CI) (%)
KY						
Rural	40	0.70 (0.77)	0.52 (0.23-0.61)	66	0.28 (0.39)	0.14 (0.12-0.17)
Urban	48	0.70 (0.88)	0.53 (0.21-0.66)	71	0.17 (0.16)	0.12 (0.11-0.15)
Total	88	0.70 (0.83)	0.52 (0.28-0.62)	137	0.22 (0.30)	0.14 (0.12-0.15)
NY						
Rural	56	0.61 (0.53)	0.61 (0.24-0.69)	18	0.26 (0.24)	0.22 (0.17-0.27)
Urban	69	0.72 (1.18)	0.29 (0.23-0.50)	23	0.16 (0.14)	0.10 (0.08-0.13)
Total	125	0.67 (0.95)	0.35 (0.26-0.61)	41	0.20 (0.19)	0.15 (0.10-0.21)
OH						
Rural	19	2.08 (1.61)	2.00 (0.56-2.48)	11	0.13 (0.06)	0.11 (0.09-0.12)
Urban	14	1.49 (0.99)	1.30 (0.85-2.26)	6	0.07 (0.04)	0.05 (0.05-0.10)
Total	33	1.83 (1.39)	1.35 (0.75-2.45)	17	0.11 (0.06)	0.10 (0.05-0.12)
Total						
Rural	115	0.88 (1.02)	0.60 (0.46-0.65)	95	0.26 (0.34)	0.14 (0.13-0.17)
Urban	131	0.79 (1.08)	0.47 (0.29-0.63)	100	0.16 (0.15)	0.12 (0.10-0.14)
Total	246	0.84 (1.05)	0.54 (0.40-0.63)	195	0.21 (0.27)	0.13 (0.12-0.14)

[a]HEAL: Helping to End Addiction Long-Term.

[b]Mdn CI: median CI.

HCS Tactics CPC

The Mann-Whitney *U* test resulted in a statistically significant difference in CPC based on the advertising method ($z=-7.47$; $P<.001$; $r=0.36$; 95% CI US $2.47-$3.89; RQ4). Social media (mean US $5.59, SD US $8.95; median US $1.61, IQR US $0.98-$4.90) had a lower CPC when compared to banner or display (mean US $7.91, SD US $7.84%; median US $6.09, IQR US $3.79-$8.97). Because the CPC was different for each advertising method, additional tests were conducted that examined each method separately based on state and community type. For social media (RQ5), the Scheirer-Ray-Hare test demonstrated that there was a significant main effect for the state ($H(2, 240)=12.05$; $P=.002$), but no effect for community type ($H(1, 240)=1.28$; $P=.26$) nor an interaction ($H(2, 240)=3.04$; $P=.22$). Dunn test with Bonferroni corrections found significant pairwise differences between KY and OH ($z=2.64$; $P=.03$) and NY and OH ($z=3.47$; $P=.002$; Table 5). OH (mean US $2.50, SD US $5.68; median US $1.08) had a lower median CPC than both KY (mean US $4.03, SD US $5.19; median US $2.31) and NY (mean US $7.51, SD US $11.10; median US $2.04). For banner or display (RQ6), the Scheirer-Ray-Hare test demonstrated that there was a significant main effect for state ($H(2, 189)=17.07$; $P<.001$) and community type ($H(1, 189)=6.99$; $P=.008$), but not for the interaction ($H(2, 189)=2.26$; $P=.32$). Dunn test with Bonferroni corrections found significant pairwise differences between KY and OH ($z=3.98$; $P<.001$) and NY and OH ($z=3.82$; $P<.001$; Table 5). OH (mean US $15.50, SD US $12.50; median US $11.30) had a higher median CPC than both KY (mean US $7.51, SD US $7.46; median US $5.82) and NY (mean US $6.00, SD US $3.93; median US $5.48). Urban communities (mean US $8.83, SD US $8.26; median US $6.94) had a higher median CPC when compared to rural communities (mean US $6.94, SD US $7.29; median US $6.94). See Table 5 for a summary of CPC descriptives.

XSL•FO
RenderX

Table . Cost-per-click descriptives for social media and banner or display by state and community type for the digital health communication campaigns of the HEALing[a] Communities Study from April 2020 to December 2023.

State and community	Social media			Banner or display		
	Advertisements, n	Mean (SD) (US $)	Median (95% Mdn CI[b]) (US $)	Advertisements, n	Mean (SD) (US $)	Median (95% Mdn CI) (US $)
KY						
Rural	40	4.24 (5.52)	2.56 (1.31 - 3.89)	61	7.02 (8.18)	5.69 (4.02 - 6.26)
Urban	48	3.86 (4.95)	2.31 (1.26 - 3.13)	66	7.97 (6.77)	6.16 (5.14 - 7.72)
Total	88	4.03 (5.19)	2.31 (1.35 - 3.44)	127	7.51 (7.46)	5.82 (5.01 - 6.56)
NY						
Rural	56	6.76 (10.80)	1.32 (1.14 - 2.37)	18	4.33 (2.01)	4.19 (3.21 - 5.17)
Urban	69	8.13 (11.30)	2.58 (1.74 - 3.88)	23	7.31 (4.57)	7.49 (5.47 - 8.91)
Total	125	7.51 (11.10)	2.04 (1.46 - 2.77)	41	6.00 (3.93)	5.48 (4.04 - 6.55)
OH						
Rural	19	2.93 (7.15)	1.15 (0.73 - 1.16)	11	10.80 (6.01)	10.20 (5.31 - 11.30)
Urban	14	1.90 (2.85)	0.82 (0.76 - 1.58)	6	24.10 (17.10)	21.10 (9.85 - 40.00)
Total	33	2.50 (5.68)	1.08 (0.81 - 1.25)	17	15.50 (12.50)	11.30 (6.94 - 13.90)
Total						
Rural	115	5.25 (8.80)	1.32 (1.20 - 1.68)	90	6.94 (7.29)	5.52 (4.41 - 6.23)
Urban	131	5.90 (9.10)	2.33 (1.46 - 2.58)	95	8.83 (8.26)	6.94 (5.82 - 7.88)
Total	246	5.59 (8.95)	1.61 (1.38 - 2.33)	185	7.91 (7.84)	6.09 (5.48 - 6.57)

[a]HEAL: Helping to End Addiction Long-Term.

[b]Mdn CI: median CI.

Study 2

For the community-tailored messages dataset, the Mann-Whitney *U* test also resulted in a significant difference in CTR based on the advertising method (z=31.9; P<.001; r=0.42; 95% CI 0.18%-0.22%; RQ1). Social media (mean 0.84%, SD 1.40%; median 0.3%, IQR 0.12%-1.92%) had a higher CTR when compared to banner or display (mean 0.20%, SD 0.56%; median 0.08%, IQR 0.00%-0.19%). Because the CTR was different for each advertising method, additional tests were conducted that examined each method separately. The quantile regression equation that predicted CTR for social media advertisements (RQ2) found statistically significant predictors within each independent variable except for the approach (testimonial vs information-based). Urban communities, static format, and having a local spokesperson all produced a higher median CTR than their respective reference group (Tables 6 and 7). Dunn test with Bonferroni corrections found statistically significant pairwise differences between all pairs of states: KY and MA (z=5.97; P<.001), KY and NY (z=3.90; P<.001), OH and KY (z=5.54; P<.001), OH and MA (z=7.85; P<.001), OH and NY (z=6.75; P<.001), and NY and MA (z=2.10; P=.04; Tables 6 and 7).

XSL•FO
RenderX

Table . Click-through rate descriptives for social media by state, image type, format, community type, and approach for the digital health communication campaigns of the HEALing[a] Communities Study from April 2020 to December 2023.

	Advertisements, n	Mean (SD) (%)	Median (95% Mdn CI[b]) (%)
State			
KY	2178	0.81 (1.26)	0.31 (0.27-0.34)
MA	140	0.42 (0.84)	0.12 (0.11-0.14)
NY	142	0.30 (0.23)	0.24 (0.21-0.29)
OH	137	2.31 (3.16)	1.26 (0.73-2.00)
Image			
Stock person	288	0.47 (0.91)	0.16 (0.12-0.24)
Local spokesperson	2309	0.88 (1.44)	0.31 (0.29-0.35)
Format			
Static	394	0.60 (1.02)	0.36 (0.24-0.45)
Video	2203	0.88 (1.45)	0.29 (0.27-0.32)
Community			
Urban	1468	0.94 (1.61)	0.33 (0.30-0.39)
Rural	1129	0.70 (1.06)	0.25 (0.22-0.30)
Approach			
Testimonial	2428	0.85 (1.43)	0.30 (0.27-0.33)
Information	169	0.60 (0.80)	0.37 (0.16-0.53)
Total	2597	0.84 (1.40)	0.30 (0.27-0.33)

[a]HEAL: Helping to End Addiction Long-Term.
[b]Mdn CI: median CI.

Table . Quantile regression coefficients for predicting click-through rate on social media using state, image type, format, community type, and approach for the digital health communication campaigns of the HEALing[a] Communities Study from April 2020 to December 2023.

	Coefficient (SE) (%)	*t* test (*df*)	*P* value	95% CI (%)
Intercept	0.34 (0.03)	10.76 (8, 2589)	N/A[b]	0.24 to 0.40
State (reference=KY)				
MA	−0.24 (0.03)	−9.03 (3, 2589)	<.001	−0.31 to −0.16
NY	−0.06 (0.03)	−2.24 (3, 2589)	.02	−0.10 to −0.04
OH	1.00 (0.35)	2.89 (3, 2589)	.003	0.50 to 1.59
Image (reference=stock person)				
Local spokesperson	0.28 (0.10)	2.73 (1, 2589)	.01	0.21 to 0.43
Community (reference=urban)				
Rural	−0.08 (0.03)	−2.76 (1, 2589)	.01	−0.11 to −0.04
Format (reference=static)				
Video	−0.29 (0.10)	−2.88 (1, 2589)	.004	−0.39 to −0.25
Approach (reference=testimonial)				
Information	0.04 (0.04)	1.03 (1, 2589)	.31	−0.01 to 0.09

[a]HEAL: Helping to End Addiction Long-Term.
[b]N/A: not applicable.

The quantile regression equation predicting the median CTR for display or banner advertisements found (RQ3) statistically significant predictors for image (stock person vs local spokesperson) and format (static vs motion graphic or video). Static advertisements and advertisements with an image of a local spokesperson had a higher median CTR (Tables 8 and 9).

XSL•FO
RenderX

Dunn test with Bonferroni corrections did not find any statistically significant pairwise differences between any pairs of states, and the only statistically significant pairwise difference based on format was between static advertisements and video advertisements (z=2.38; P=.02).

Table . Click-through rate descriptives for display by state, image type, format, community type, and approach for the digital health communication campaigns of the HEALing[a] Communities Study from April 2020 to December 2023.

	Advertisements, n	Mean (SD) (%)	Median (95% Mdn CI[b]) (%)
State			
KY	2558	0.21 (0.60)	0.08 (0.07-0.09)
MA	639	0.17 (0.35)	0.09 (0.08-0.09)
OH	119	0.26 (0.59)	0.11 (0.09-0.12)
Image			
Stock person	2230	0.19 (0.59)	0.07 (0.07-0.08)
Local spokesperson	1086	0.23 (0.48)	0.12 (0.10-0.13)
Format			
Static	1931	0.21 (0.63)	0.09 (0.08-0.09)
Motion graphic or GIF[c]	942	0.15 (0.18)	0.08 (0.06-0.11)
Video	443	0.29 (0.71)	0.06 (0.03-0.07)
Community			
Urban	1899	0.20 (0.50)	0.08 (0.08-0.09)
Rural	1417	0.21 (0.63)	0.08 (0.07-0.09)
Approach			
Testimonial	2771	0.21 (0.60)	0.08 (0.07-0.08)
Information	545	0.17 (0.25)	0.10 (0.09-0.11)
Total	3316	0.20 (0.56)	0.08 (0.08-0.09)

[a]HEAL: Helping to End Addiction Long-Term.

[b]Mdn CI: median CI.

[c]GIF: graphics interchange format.

XSL•FO
RenderX

Table . Quantile regression coefficients for predicting click-through rate on banner or display using state, image type, format, community type, and approach for the digital health communication campaigns of the HEALing[a] Communities Study from April 2020 to December 2023.

	Coefficient (SE) (%)	*t* test (*df*)	*P* value	95% CI (%)
Intercept	0.08 (0.00)	21.66 (8, 3308)	<.001	0.07 to 0.08
State (reference=KY)				
MA	0.00 (0.01)	−0.18 (2, 3308)	.86	−0.01 to 0.02
OH	0.01 (0.02)	0.61 (2, 3308)	.54	0.02 to 0.06
Image (reference=stock person)				
Local spokesperson	0.09 (0.01)	11.27 (1, 3308)	<.001	0.07 to 0.11
Community (reference=urban)				
Rural	−0.01 (0.01)	−1.69 (1, 3308)	.09	−0.02 to 0.00
Format (reference=static)				
Motion graphic or GIF[b]	−0.04 (0.01)	−4.99 (2, 3308)	<.001	−0.05 to −0.02
Video	−0.12 (0.01)	−8.59 (2, 3308)	<.001	−0.14 to −0.09
Approach (reference=testimonial)				
Information	0.01 (0.01)	0.63 (1, 3308)	.53	−0.01 to 0.02

[a]HEAL: Helping to End Addiction Long-Term.

[b]GIF: graphics interchange format.

Discussion

Key Results

This paper assessed 2 key health communication campaign flight components: the exposure and engagement outcomes associated with specific tactics used in standard or template-based message dissemination (study 1) and the exposure and engagement outcomes associated with specific features of customizable, community-tailored campaign messages (study 2). In Study 1 (standard or template-based messages), social media had a higher CTR compared to banner or display (RQ1). Social media had a statistically significant main effect for state, but not for community type (RQ2), where OH had the highest CTR. There were no significant differences based on state or community type for display or banner (RQ3). Social media had a lower CPC compared to banner or display (RQ4). Social media also had a statistically significant main effect for state, but not community type, where OH had the lowest CPC (RQ5). Banner or display had a significant main effect for state and community type, where OH and urban communities had the highest CPC (RQ6). In Study 2 (community-tailored messages), social media also had higher CTR compared to banner or display (RQ1). For social media, advertisements in static format, advertisements featuring local spokespersons, advertisements in OH, and advertisements in urban communities had the highest CTR. There were no significant differences based on approach, that is, testimonial or information-based (RQ2). For display or banner, static advertisements and advertisements featuring local spokespersons had the highest CTR. Approach, state, and community type differences were not significant (RQ3).

Interpretation

The results suggest that social media advertisements outperformed banner or display advertisements on CTR and CPC KPIs in the context of health communication campaigns intended to encourage information-seeking via click-throughs to website landing pages. However, these rates and costs varied as a function of the state where the campaign took place, the type of image used in the advertisement, the community type, and the format of the advertisement, suggesting that these contextual factors play a key role in the performance of digital campaigns.

Although means can be highly variable, median CPC was consistently lower for social media advertisements compared to banner or display advertisements; the extent varied by state, with OH having a much lower median CPC for social media compared to KY and NY. CTR differences by state were not simply attributable to the total dollar amount spent on media buys at each site but reflected differences in actual CTR. Social media CTR was similar for KY and NY but was higher in OH, which also had the best-performing CPC. There were no significant pairwise differences in the banner or display CTR between states, although the overall main effect for the state was significant and, descriptively, KY and NY CTR were similar while OH CTR was lower.

The differences in CTR and CPC performance by state provide a cautionary tale for the health communication practitioner. One state (OH) performed very well in response to social media advertisement placements but not in banner or display advertisements. Another state (NY) performed better with banner or display advertisements. The design and content of these banner or display advertisements were similar across sites. Considering the resources often invested in digital health communication campaigns, pilot testing across regions is a

XSL•FO
RenderX

recommended best practice to direct planners toward the most effective advertising methods [17].

There are also some important caveats regarding the better overall CTR and CPC performance for social media advertisements compared to banner or display advertisements. It is possible that the placement of the social media advertisements in these campaigns was more appropriate than the placement of the banner or display advertisements or that the message content and strategy were better suited to social media advertising [18]. Furthermore, the data here demonstrate that CTR and CPC are key criterion variables when the desired outcome is to drive targeted audiences to act by visiting a website landing page. However, if the targeted outcome is different, such as increased attention to or enhanced recall of specific health information, banner or display advertising may prove to be more efficient and effective. Past research demonstrates that advertising message complexity is a key factor in attention and attitudinal responses, where fewer message features result in more recognition, comprehension, recall, and positive attitudes [19,20].

The results associated with CTR as a function of urban versus rural community types varied across advertising methods. In Study 1, CTR for banner or display advertisements was higher in rural communities than in urban communities. In Study 2, CTR for social media advertisements was higher in urban communities than in rural communities. In general, banner or display advertisements reach a broader audience across the internet, while social media advertisements are targeted to specific social media networks (eg, Facebook or Instagram). Social media use is generally higher in urban areas, as internet access is more reliable [21]. It is possible that banner or display advertisements are more appropriate for reaching rural individuals because of their more expansive reach. This is because banner or display advertisements appear on a wider range of websites (eg, news and shopping) and are not specifically related to one's browsing activity. Alternatively, social media advertisements are only displayed within the specific platform that one actively uses.

There were 2 additional message factors that emerged as significant in study 2: the type of individual featured in the advertisement (stock photo person and local spokesperson) and the format of the advertisement (static, motion graphic or GIF, and video). Advertisements featuring local spokespeople and static advertising formats produced higher CTR than their respective reference groups. Here, using photographs of real people from the HCS communities initiated more engagement with campaign messages than stock photography. The stock images may have been perceived as less authentic or more unnatural by audiences than the photographs featuring real community members [22].

Kite et al [23] caution against creating media campaigns that use inappropriate or low-quality creative messaging; using photographs of real people may be one way to increase the quality and authenticity of messages. Regarding the performance of static formatted messaging in these studies, it seems counterintuitive that a less attention-getting format would be more successful than other types. However, when the intended outcome is to communicate clear and simple messaging, static may be a preferred format. Fan et al [24] also found that static and video messaging can elicit similar emotional responses. Static content is also more affordable to place than motion graphics or GIFs and video formats, which benefits those managing campaigns with budget constraints.

One final message factor was examined in this research effort: message approach (testimonial vs information-based). In general, narrative content garners more attention than purely informational content. In the present studies and counter to prior research findings [25], testimonial approaches were not more effective than information-based approaches [24]. However, considering that the target audience for these campaigns was communities highly affected by the opioid crisis, issue fatigue may have influenced engagement (or lack thereof). Issue fatigue occurs when a topic is consistently and regularly repeated and can result in issue avoidance and the dismissal of information [26]. Capturing and understanding community attitudes at the start of health campaigns via surveys appears to be a key component of formative research before launch. Furthermore, assessing community perceptions of campaign message repetition mid-flight could lend insight into issue fatigue. Both practices are recommended for future studies.

Limitations

There are several limitations to report with this study. First, inconsistencies in the cost data from MA prevented us from including it in the analyses for study 1. Although we could account for cost variables, there were different campaign budgets and media buyers for each site, which increased variability in the data as a whole. Second, although each RQ presented here assessed different categorical types of advertisements, the overall number of RQs increases the possibility of type I error. Third, CTR and CPC are key criterion variables in evaluating campaign performance, but there are many others not assessed here (eg, time spent on page) that could also contribute to our understanding of campaign engagement. To this point, postclick behavioral outcomes (ie, landing and subsequent action on the HCS website) were not reported here because the landing website was maintained and reviewed by a third-party data coordinating center that was unable to verify the data. Algorithmic changes across Google and Meta regarding the placement of the advertisements were also unaccounted for. Finally, community coalitions and champions steered the types of paid advertising selected for each campaign. While media buyers and consultants provided guidance on effective strategies, communities varied widely in the types of advertising used and the specific audience segments and targeting parameters they chose to receive those advertisements. This high degree of variability did not allow for comparisons by audience or targeting parameters (eg, specific geographic locations and behavioral interests), which, along with the advertisements themselves, are known to influence campaign performance. Further, many communities experimented with more novel forms of digital media advertising (eg, streaming television advertising and mobile app advertising), but these efforts were not large enough to allow for meaningful analysis or comparison.

XSL•FO
RenderX

Generalizability

We hope the data presented here provide some guidance for health communication scholars and practitioners in their implementation of digital media campaigns, especially those associated with OUD. This paper is one of the first to report on the performance of OUD campaigns using CTR and CPC measures, demonstrating the utility of their inclusion in future campaign evaluations. To date, many local, state, federal, and nonprofit agencies are implementing media campaigns with little guidance for best practices.

Conclusions

In all, we can conclude that social media advertisements generally outperform banner or display advertisements in health communication campaigns, but only when considering key contextual factors, such as geographical location and type, imaging used, and format of the advertisement. Furthermore, using median CPC as a comparison metric instead of mean CPC is an effective metric for explaining campaign performance. Health communication practitioners should use pilot testing before launching digital campaigns to guide advertising methods. Desired campaign outcomes (eg, visits to a website landing page vs increased awareness and/or enhanced recall of information) should also be considered when implementing campaign strategies, tactics, and messages.

Acknowledgments

Generative artificial intelligence was not used in any portion of this manuscript writing.

Funding

This research was supported by the National Institutes of Health (NIH) and the Substance Abuse and Mental Health Services Administration (SAMHSA) through the NIH HEAL (Helping to End Addiction Long-Term) Initiative (awards UM1DA049394, UM1DA049406, UM1DA049412, UM1DA049415, and UM1DA049417). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, SAMHSA, or the NIH HEAL Initiative. The funders of this study, NIH and SAMHSA through the NIH HEAL, had no involvement in the study design, data collection, analysis, interpretation, or writing of this manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request. Please email nhlewis@iu.edu directly for requests.

Conflicts of Interest

SLW has served as a scientific advisor for the following entities, which either have or are developing medications for the treatment of opioid use disorder: AstraZeneca, Braeburn Pharmaceuticals, Cerevel Therapeutics, Indivior, Kinnoxis, and ReacX. In the last 3 years, MRL has received research consulting fees from Braeburn Pharmaceuticals, Journey Colab, and Berkshire Biomedical and an honorarium from Camurus for developing and giving a scientific talk on long-acting injectable buprenorphine.

References

1. Chan L, O'Hara B, Phongsavan P, Bauman A, Freeman B. Review of evaluation metrics used in digital and traditional tobacco control campaigns. J Med Internet Res 2020 Aug 11;22(8):e17432. [doi: 10.2196/17432] [Medline: 32348272]
2. Noar SM. A 10-year retrospective of research in health mass media campaigns: where do we go from here? J Health Commun 2006;11(1):21-42. [doi: 10.1080/10810730500461059] [Medline: 16546917]
3. Kite J, Chan L, MacKay K, et al. A model of social media effects in public health communication campaigns: systematic review. J Med Internet Res 2023 Jul 14;25(3):e46345. [doi: 10.2196/46345] [Medline: 37450325]
4. Rath JM, Perks SN, Vallone DM, et al. Educating young adults about opioid misuse: evidence from a mass media intervention. Int J Environ Res Public Health 2021 Dec 21;19(1):22. [doi: 10.3390/ijerph19010022] [Medline: 35010279]
5. Davis KC, Duke J, Shafer P, Patel D, Rodes R, Beistle D. Perceived effectiveness of antismoking ads and association with quit attempts among smokers: evidence from the tips from former smokers campaign. Health Commun 2017 Aug;32(8):931-938. [doi: 10.1080/10410236.2016.1196413] [Medline: 27435919]
6. Durkin SJ, Brennan E, Wakefield MA. Optimising tobacco control campaigns within a changing media landscape and among priority populations. Tob Control 2022 Mar;31(2):284-290. [doi: 10.1136/tobaccocontrol-2021-056558] [Medline: 35241601]
7. Frkovich J, Hedrick H, Anakaraonye AR, Bornkessel A, Lefebvre RC. Opioid-related public health communication campaigns: an environmental scan. Am J Health Promot 2022 Jul;36(6):913-919. [doi: 10.1177/08901171221082471] [Medline: 35365055]

XSL•FO
RenderX

8. Lefebvre RC, Chandler RK, Helme DW, et al. Health communication campaigns to drive demand for evidence-based practices and reduce stigma in the HEALing communities study. Drug Alcohol Depend 2020 Dec 1;217:108338. [doi: 10.1016/j.drugalcdep.2020.108338] [Medline: 33152673]
9. Flora JA, Schooler C, Pierson RM. Effective Health Promotion Among Communities of Color: The Potential of Social Marketing: Psychology Press eBooks; 2018:353-373. [doi: 10.4324/9781315805795-20]
10. Kennedy-Hendricks A, McGinty EE, Summers A, Krenn S, Fingerhood MI, Barry CL. Effect of exposure to visual campaigns and narrative vignettes on addiction stigma among health care professionals: a randomized clinical trial. JAMA Netw Open 2022 Feb 1;5(2):e2146971. [doi: 10.1001/jamanetworkopen.2021.46971] [Medline: 35119460]
11. Fernandez P, Azucar D, Zambole K. A dose of truth: a qualitative assessment of reactions to messages about fentanyl for people who use drugs. Subst Use Misuse 2023;58(4):520-527. [doi: 10.1080/10826084.2023.2177112] [Medline: 36762461]
12. Smith RA, Osoro RA, Sterner G. Understanding the effects of digital storytelling on OUD stigma reduction: exploring character and the recovery plot. Health Commun 2025 Nov;40(12):2467-2477. [doi: 10.1080/10410236.2025.2463535] [Medline: 39950341]
13. Geng S, Yang P, Gao Y, Tan Y, Yang C. The effects of ad social and personal relevance on consumer ad engagement on social media: the moderating role of platform trust. Comput Human Behav 2021 Sep;122:106834. [doi: 10.1016/j.chb.2021.106834]
14. Walters ST, Drainoni ML, Oga EA, Byard J, Chandler RK. Solving the “last mile” problem in overdose prevention: lessons from the HEALing Communities Study. Drug Alcohol Depend 2024 Nov 1;264:112453. [doi: 10.1016/j.drugalcdep.2024.112453] [Medline: 39379270]
15. Atkin CK, Freimuth V. Guidelines for formative evaluation research in campaign design. In: Anonymous Public Communication Campaigns: SAGE; 2012:52-68. [doi: 10.4135/9781544308449.n4]
16. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. The STROBE statement: guidelines for reporting observational studies. PLoS Med 2007;4(10). [doi: 10.1371/journal.pmed.0040297] [Medline: 17941715]
17. Gaysynsky A, Heley K, Chou WYS. An overview of innovative approaches to support timely and agile health communication research and practice. Int J Environ Res Public Health 2022 Nov 16;19(22):15073. [doi: 10.3390/ijerph192215073] [Medline: 36429796]
18. Namin A, Hamilton ML, Rohm AJ. Impact of message design on banner advertising involvement and effectiveness: an empirical investigation. J Mark Commun 2020 Feb 17;26(2):115-129. [doi: 10.1080/13527266.2017.1393767]
19. Pieters R, Wedel M, Batra R. The stopping power of advertising: measures and effects of visual complexity. J Mark 2010 Sep;74(5):48-60. [doi: 10.1509/jmkg.74.5.048]
20. Yeun Chun K, Hee Song J, Hollenbeck CR, Lee JH. Are contextual advertisements effective? Int J Advert 2014 Jan;33(2):351-371. [doi: 10.2501/IJA-33-2-351-371]
21. Vogels EA. Some digital divides persist between rural, urban and suburban America. Pew Research Center. 2021. URL: https://www.pewresearch.org/short-reads/2021/08/19/some-digital-divides-persist-between-rural-urban-and-suburban-america/ [accessed 2026-03-02]
22. Ward CG. Stock images, filler content and the ambiguous corporate message. M/C J 2007;10(5). [doi: 10.5204/mcj.2706]
23. Kite J, Thomas M, Bellew B, Bauman A, Grunseit A. Mass media campaigns and the “file drawer problem”: a mixed methods study of how to avoid campaign failure. PLoS ONE 2024;19(4):e0294372. [doi: 10.1371/journal.pone.0294372] [Medline: 38625844]
24. Fan S, Shen Z, Koenig BL, Ng TT, Kankanhalli MS. When and why static images are more effective than videos. IEEE Trans Affective Comput 2023;14(1):308-320. [doi: 10.1109/TAFFC.2020.3040399]
25. Adebayo AL, Davidson Mhonde R, DeNicola N, Maibach E. The effectiveness of narrative versus didactic information formats on pregnant women’s knowledge, risk perception, self-efficacy, and information seeking related to climate change health risks. Int J Environ Res Public Health 2020 Sep 23;17(19):6969. [doi: 10.3390/ijerph17196969] [Medline: 32977683]
26. Schumann C, Arlt D. When citizens get fed up. Causes and consequences of issue fatigue—results of a two-wave panel study during the coronavirus crisis. Communications 2023 Mar 6;48(1):130-153. [doi: 10.1515/commun-2021-0014]

Abbreviations

CPC: cost-per-click
CTH: Communities That HEAL
CTR: click-through rate
EBP: evidence-based practice
GIF: graphics interchange format
HCS: HEALing Communities Study
HEAL: Helping to End Addiction Long-Term
KPI: key performance indicator
MOUD: medication for opioid use disorder
NIH: National Institutes of Health

XSL•FO
RenderX

ORAU: Oak Ridge Associated Universities
OUD: opioid use disorder
RQ: research question
SAMHSA: Substance Abuse and Mental Health Services Administration
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by S Brini; submitted 03.Apr.2025; peer-reviewed by J Wang, N Hu, P Shafer; revised version received 29.Jan.2026; accepted 30.Jan.2026; published 01.Apr.2026.

Please cite as:
Lewis N, Reynolds J, Krause D, Reeves PM, Luster J, Stein MD, Walsh SL, Farmer A, Lofwall MR, Roberts MF, Surratt HL, Crockett BN, Stephens K, Bursey K, Mattson K, Slater MD
Health Communication Campaign Performance During the HEALing Communities Study: Cross-Sectional Examination of Digital Advertising Methods
J Med Internet Res 2026;28:e75458
URL: https://www.jmir.org/2026/1/e75458
doi:10.2196/75458

© Nicky Lewis, Jennifer Reynolds, Diane Krause, Phillip M Reeves, Jamie Luster, Michael D Stein, Sharon L Walsh, Amy Farmer, Michelle R Lofwall, Monica F Roberts, Hilary L Surratt, Brooke N Crockett, Kara Stephens, Kelli Bursey, Kristin Mattson, Michael D Slater. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 1.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Engagement With Meditation Apps: Cross-Sectional Survey of Use and Associations

Julia Adams[1], BSc, BA, BSc(Hons); Jonathan Davies[1], PhD; Prai Wattanatakulchat[1], BA; Julieta Galante[1], MD, PhD; Felicity Miller[1], MPsych(Clin); Simon D'Alfonso[2], PhD; Nicholas T Van Dam[1], PhD

[1]Contemplative Studies Centre, Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Melbourne, Australia

[2]School of Computing and Information Systems, University of Melbourne, Melbourne, Australia

Corresponding Author:
Julia Adams, BSc, BA, BSc(Hons)
Contemplative Studies Centre
Faculty of Medicine, Dentistry and Health Sciences
University of Melbourne
700 Swanson St
Melbourne, 3010
Australia
Phone: 61 2904309289
Email: adamsjuliaalice@gmail.com

Abstract

Background: Meditation apps are increasingly popular, yet there is limited understanding of how much users actually engage with them. While meditation apps show promise for supporting mental health, engagement in real-world settings appears to be notably low. The patterns of app use and the factors that influence usage remain relatively unclear.

Objective: This study aims to examine the extent of meditation app use and the factors associated with user engagement.

Methods: We conducted a cross-sectional survey of 536 recent meditation app users across 5 English-speaking countries. Engagement data were collected via self-report and app-verified screenshots. Assessed factors included user characteristics (age, education, income, sex, country, personality, self-efficacy, readiness and expectations for change, self-compassion, and quality of life), mental health (distress, well-being, life satisfaction, anxiety, depression, support, and stress), and app-related elements (therapeutic alliance, appeal, functionality, aesthetics, information, quality, and perceived impact). The 4 outcome variables representing engagement were app-verified minutes, self-reported minutes, app-verified minutes per year (adjusted for app download date), and self-reported minutes per year (adjusted for app download date). Associations between app use and variables of interest were examined using correlations. Factors with significant associations were then included in multivariable regression models to identify those most strongly associated with engagement.

Results: Age (ρ=0.13-0.15, PP^{FDR}, where FDR is false discovery rate), expectations for sleep (ρ=0.12-0.33, P^{FDR}<.05), and expectations for thriving (ρ=0.12-0.18, P^{FDR}<.05) were associated with all outcome measures except adjusted objective minutes. Readiness to change was associated with all outcome measures (ρ=0.24-0.33, P^{FDR}<.05). Among app factors, appeal (ρ=0.18-0.23, P^{FDR}<.05) and perceived impact (ρ=0.23-0.32, P^{FDR}<.05) were associated with all outcome measures except adjusted self-report minutes, while perceived quality (r=0.28-0.51, P^{FDR}<.05) was associated with all outcome measures. Robust linear regressions showed that greater readiness to change (β=0.005-0.026, *P*=.006-.02), higher education level (β=0.029-0.540, *P*<.001), and higher openness (β=0.004-0.010, *P*=.008-.03) were associated with increased engagement. Additionally, greater expectations for sleep (β=0.004-0.009, *P*=.02-.04), greater expectation match (β=0.023, *P*=.03), and higher perceived app quality (β=0.008-0.042, *P*=.001-.01) were uniquely associated with increased engagement.

Conclusions: Most individuals who download meditation apps engage minimally. Our findings suggest that users who are more educated, open to new experiences, and hold strong beliefs in the effectiveness of meditation apps are more likely to use them regularly. Longitudinal studies are needed to examine patterns of use and strengthen causal inferences.

(J Med Internet Res 2026;28:e71960) doi:10.2196/71960

XSL•FO
RenderX

KEYWORDS

meditation; mindfulness; engagement; app; meditation apps; digital mental health intervention; behavior change

Introduction

Background

Around 1 billion people globally live with a mental health disorder [1], creating a demand that exceeds available resources. As the rise of technology has coincided with increasing strain on mental health systems, digital mental health interventions have gained popularity due to their accessibility [2,3]. Fully automated versions of these interventions may reduce reliance on limited human resources. However, engagement remains a challenge, with fewer than 20% of users continuing beyond 7 days [4].

Meditation apps are among the most common digital mental health tools [4,5]. Meditation encompasses a wide range of techniques across different traditions and religions and typically involves emotional and attentional regulation [6]. Mindfulness meditation, for example, emphasizes nonjudgmental awareness of the present moment [7]. Among regular meditators, most have used a meditation app [8]. While global rates of meditation remain unknown, 70 million people had downloaded Headspace by 2022 [9]. Thus, it is likely that a very large proportion of the population has tried meditation through an app. Given the high accessibility of meditation apps among those facing barriers to mental and physical health care, it is important to examine the practical limitations of such programs, with engagement being a key shortfall in app-based behavior change [10,11].

Most current information on meditation app use comes from clinical trials, which are not representative of real-world use. While randomized controlled trials of meditation apps show small- to medium-sized effects [12], real-world estimates suggest exceedingly high discontinuation rates [4]. As behavioral interventions are only effective if used, this presents a major challenge for apps [4]. For digital offerings to be truly useful, a better understanding of the factors associated with sustained use is essential. This study examines engagement rates and identifies who is most likely to engage with meditation apps.

Mindfulness-Based Programs and Meditation Apps

While a limited number of meditation apps have shown some efficacy for mental health outcomes [12], they should be considered separately from the established evidence base for mindfulness-based programs (MBPs) [13]. MBPs are arguably the most popular form of meditation training in clinical and academic settings, likely due to their strong evidence base [14,15]. By contrast, most popular meditation apps depart from the guided, intensive structure of MBPs [16]. Only 4% of popular apps provide evidence of their benefits [17]. Even where apps show potential efficacy, recent reviews highlight engagement as a major limitation to intervention effectiveness [12,18-20].

The Digital Transition

Digital health interventions generally face engagement issues, which likely reduce their benefits [4]. In nonpharmacological interventions, adherence is linked to outcomes [21], yet it tends to be worse in digital formats than in face-to-face interventions [22]. For behavior change apps (and apps more broadly), discontinuation occurs in 40%-60% of users [11]. In naturalistic settings, 21%-88% of users engage with an app at least once, but only 0.5%-28% sustain engagement (eg, completing all assigned modules or continuing use beyond 6 weeks [23]). Engagement decreases when digital interventions lack interactive human or human-like support [24], posing challenges for fully automated meditation apps. Clarifying who engages meaningfully with meditation apps is therefore important, given the link between adherence and outcomes.

Attrition, Adherence, and Engagement

Engagement refers to the extent of intervention use, including the amount, frequency, duration, and depth of use [25]. Attrition and adherence are related terms that describe levels of (dis)engagement in research studies. Attrition refers to discontinuation or dropout from the intervention program or from research data provision during a study [11]. For MBPs, attrition is around 19% [26], whereas app-based interventions show an average attrition of 42% in studies lasting 10 days to 12 weeks [12]. Real-world estimates for meditation apps indicate disengagement rates as high as 94% within the first 2 weeks [4].

Adherence refers to the extent to which an individual follows a prescribed treatment or intervention [21]. As no clear guideline exists for the amount of practice required to achieve an effect in mindfulness or meditation [16,27], adherence can only be considered in relation to recommended practice amounts (see example in [28]). In meditation training, prescribed engagement time ranges from as little as 35 minutes [28] to 3 hours per week in the widely used Mindfulness-Based Stress Reduction program [7]. By contrast, many apps recommend as little as 5 minutes per day or provide no clear guidance regarding minimum practice length, session duration, or overall time commitment needed to establish a practice [29]. Given the limited knowledge about dose-response relationships in meditation [16,29] and the tendency for most people to discontinue practice relatively early [4], engagement serves as a useful proxy for understanding who practices, what type of practice they follow, and why.

Why Do People (Dis)Engage With Meditation and Apps?

Overview

Understanding engagement in meditation apps requires consideration of various behavior change and persuasive systems design frameworks. Behavior change frameworks—including habit formation theory, social cognitive theory, the theory of planned behavior, and the transtheoretical model of change—suggest that user characteristics such as expectations, motivation, readiness to change, consistency of use, and self-efficacy influence engagement with behavioral interventions [30-33]. The persuasive systems design framework highlights

XSL•FO
RenderX

app design features that shape engagement and therapeutic alliance [34].

Habit formation theory emphasizes reward and associative cues as central to establishing habits, with positive outcomes reinforcing continued engagement [33]. Context can shape how rewards are perceived, influencing whether a habit is formed [35]. The theory of planned behavior further posits that perspective and context guide behavior [30,36]. Broad factors such as sociodemographics, mental health, and personality also influence engagement [30]. Expectations are shaped by attitudes and norms, with positive expectations and attitudes predicting greater meditation app engagement [36]. The transtheoretical model outlines stages of change, with later stages—more closely aligned with commitment to action—linked to more sustained behavior change [32]. Readiness to change reflects an individual's stage of change and is associated with successful maintenance of behavior change [37]. The Sussex Meditation Model identifies preintention, preparation, action, and maintenance stages as relevant to establishing a meditation practice [38,39]. Persuasive systems design, which examines how digital interventions can be structured to influence user behavior, highlights app features that enhance engagement, such as reminders and personalization [34]. Drawing on these frameworks, factors relevant to engagement in behavior change, meditation, or app use were categorized into user-related factors (sociodemographics, personal/user characteristics, and mental health factors) and app-related factors.

Sociodemographic Factors

Sociodemographic factors associated with disengagement from meditation include lower levels of education [40]; however, men, people with less education, and those with poorer health are less likely to begin meditating [41]. Meditators are also more likely to be wealthier than nonmeditators [41]. In online and app-based meditation, older age, positive expectations, and intrinsic motivation are associated with greater engagement [42,43].

Personal/User Factors

Personality factors have also been shown to influence engagement with meditation apps. Conscientiousness has been associated with meditation in general [44]. Openness predicts meditation practice outside formal program training, reflecting the "in-the-wild" context of app use [45].

Behavior change factors may also influence engagement with meditation apps. Self-efficacy and readiness to change have been linked to successful habit formation [46]. A higher intention to practice is associated with greater engagement [46]. Intrinsic motivation moderates behavior change success across demographic groups [47] and is crucial for making initial behavior change choices. Self-compassion and self-efficacy have also been found to be positively associated with engagement in behavior change [47,48].

Expectations for program efficacy can also influence behavior change. Positive experiences that meet expectations can facilitate ongoing engagement. Conversely, engagement may decline when a program or behavior does not deliver the anticipated positive outcomes [49]. Experiences of progress enhance engagement in both behavior change apps and meditation apps [50,51]. Positive expectations also predict higher engagement with digital meditation resources [42].

Mental Health Factors

Health characteristics are also important for engagement. People may be motivated by physical or mental health issues, but these same issues can also act as barriers [43]. This paradox can be explained by the desire to address a problem that simultaneously hinders the ability to engage in practice. Additionally, limited perceived gains may lead to early discontinuation. Barriers to mental or physical health care, which can impact quality of life, may further motivate meditation app use to address unmet health needs [11,41]. Although meditation use among individuals with mental health problems is common, depression is associated with low adherence to behavior modification recommendations in clinical populations [52,53]. The very symptoms people seek to address—such as amotivation, distressing thoughts, and irritability—can also complicate their efforts. Meditation apps may be moderately effective for depression, anxiety, and stress [12,19,50], potentially fostering an experience of progress. However, a minimal level of engagement is necessary to achieve efficacy [4,27]. Consequently, failure to achieve expected outcomes may lead to decreased engagement.

App Factors

The user's relationship with the app is also relevant. Therapeutic alliance—the collaborative relationship between the user and the app—predicts engagement with mental health apps [53]. Ease of use, the ability to personalize settings, reminders, progress tracking, and positive perceptions of the app also predict higher engagement with mental health apps, though these factors have not been extensively examined in meditation apps [42,53,54]. Usability (ie, the app's functionality) was identified as a key factor related to engagement in a systematic review of mental health apps [55].

This Study

Previous literature highlights factors that may be associated with meditation app use. In a cross-sectional survey capturing demographics, retrospective reports of app use, mental health factors, and perspectives on apps, we aimed to examine engagement rates and identify factors significantly associated with engagement.

This study focused on several preregistered questions:

1. To what extent are *user-related factors*—including sociodemographic characteristics, spirituality, personality, self-efficacy, self-regulation, motivation, expectations, self-compassion, mental health care status, and psychological distress—associated with mindfulness *app engagement*?
2. To what extent are *user-app relationship factors*—including therapeutic alliance, agreement on tasks and goals, and perceived app empathy and expertise—associated with mindfulness app *engagement*?
3. To what extent are *app-related factors*—including appeal, functionality, aesthetics, information quality, quantity, and

XSL•FO
RenderX

credibility, customization, accessibility, and usability—associated with mindfulness app *engagement*?

Additional questions included in the preregistration are not addressed in this paper.

Methods

Deviations From Preregistration

For clarification, we have changed the term "mindfulness apps" to "meditation apps" to capture a broader range of relevant practices. Mindfulness can, but does not necessarily, entail meditation and is variably represented as a capacity, skill, or technique. Meditation, by contrast, encompasses a broad array of spiritual and secular practices that use techniques such as focusing on an object, experience, image, or idea [56].

Deviations from the preregistration included the following: (1) focusing on 4 definitions of minutes as the primary outcome and omitting the second preregistered outcome variable—regular practice hours—for simplicity. Regular practice hours were not included because their calculation combined multiple variables and, therefore, could be subject to estimation error. The 4 variations of the outcome variable were included to capture the complexity of user behavior. (2) We did not report 95% odds ratios, as continuous outcomes were used. (3) The final sample size was substantially reduced to 536 from the target of 1000 due to a smaller-than-anticipated eligible pool. This reduction decreased statistical power, although it still allowed adequate power to detect small effects. The reduced pool also led to a fourth deviation. (4) Recruitment was extended to Australia, Canada, the United Kingdom, and New Zealand. An additional deviation involved (5) not analyzing motivation for use, as this information was captured in open-text responses and could not be used in this quantitative analysis.

Ethical Considerations

This study was conducted in accordance with ethical guidelines and was approved by the Office of Research Ethics and Integrity at the University of Melbourne (approval number 2025-23969-62994-8).

Participants provided informed consent to participate in the study via the Qualtrics survey (Qualtrics International Inc). Consent was obtained within the survey, which also included a downloadable copy of the plain language statement. The plain language statement is available in Section S1 in Multimedia Appendix 1. Provision of informed consent included acknowledgment of the right to withdraw at any time without providing an explanation. Participants also consented to secondary analyses. Survey questions were coded so that participants could not proceed without providing consent. All included responses were double-checked to ensure consent had been given.

Participants were compensated Aus $0.30-0.50 (US $0.20-0.33) for completing the screening survey (mean duration 1 minute 49 seconds) and Aus $6-8 (US $3.96-5.29) for completing the follow-up survey (mean duration 22 minutes 57 seconds), averaging Aus $20.59 (US $13.60) per hour. Survey compensation varied slightly based on median completion time; compensation was occasionally increased to better approximate the proposed hourly rate if the median completion time indicated the study took longer than expected.

Privacy and Confidentiality

Where possible, identifying information was removed from the dataset. Any copies of datasets containing identifying information were stored securely in accordance with relevant privacy guidelines and encrypted using Transport Layer Security (also known as HTTPS).

Study Design

Overview

This was a cross-sectional analysis of data collected from participants.

Procedure

Participants were recruited via Prolific (Prolific Academic Ltd) to complete a survey hosted on Qualtrics. The survey was accessible to potential participants in the United States, the United Kingdom, Canada, Australia, and New Zealand between August 1 and October 6, 2023. Participants were invited to complete a prescreening survey, and eligible individuals were sent the full survey within 1-2 days. Surveys were completed online using a laptop or mobile device. Participants were asked to upload a screenshot of their app use statistics, which provided information such as minutes, days, sessions, streaks, and the original date of download, depending on the app.

App Selection

We collected engagement information for popular meditation apps listed on the iOS (Apple Inc) and Android (Google LLC/Alphabet Inc) app stores (see Section S2 in Multimedia Appendix 1). Participants using apps in which meditation—including mindfulness meditation—was the primary intended function were included, based on app descriptions, marketing, and in-app features.

Participants using any app could complete the prescreening survey. Two (JA and JD) researchers assessed whether the app (1) prominently promoted itself as a mindfulness meditation tool and (2) provided techniques to practice mindfulness or another form of meditation. Meditation or mindfulness could not be a secondary component. We did not evaluate app content in relation to any specific definition. We adopted this approach because meditation apps do not offer a single type of meditation, nor do mindfulness apps (eg, Headspace) necessarily adhere to the MBP definition of meditation. Apps were excluded if they focused exclusively on fitness/exercise, employee well-being, cognitive behavioral therapy, or other mental health interventions. All included apps were fully automated (ie, without human support).

Participants

Inclusion criteria required participants to have used an eligible meditation app within the past 180 days; be fluent in English; and reside in Australia, Canada, New Zealand, the United Kingdom, or the United States. Exclusion criteria included failure to provide evidence of app use or use of an app in which meditation was not the primary focus.

XSL•FO
RenderX

A total of 6137 prescreening surveys were completed. We excluded 5307 responses: 4343 (70.77%) were unable to demonstrate access to an app, 316 (5.15%) had not used the app in the past 180 days, 319 (5.20%) had downloaded an ineligible app, and 329 (5.36%) self-reported zero use. Of the remaining surveys, 800 (13.04%) met the inclusion criteria.

Of the 800 participants invited to the survey, 677 (84.6%) completed it. Among the 675 survey responses received, 18 (2.3%) were identified as likely bots or fraudulent responses based on fraudulent screenshots or failed reCAPTCHA (reverse Completely Automated Public Turing Test to Tell Computers and Humans Apart), and 13 (1.6%) exhibited suspiciously high average session lengths (>3× the IQR, 70.35 minutes). An additional 22 participants (2.8%) timed out before completing the survey, 25 (3.1%) failed attention checks, 59 (7.4%) failed screenshot checks, 86 (10.75%) responded twice, and 21 (2.6%) declined to complete the survey. These categories were not mutually exclusive. The resulting sample consisted of 563 (70.4%) participants who consented and completed the full survey. Finally, 27 (4.8%) multivariate outliers were excluded according to the preregistration, yielding a final sample of 536 participants.

Measures

Engagement

To verify the reliability of self-reported information, we collected both subjective self-reports and objective, app-verified data (screenshots provided by participants), which included minutes, days, streaks (consecutive days of use), number of sessions, average session length, and duration of app ownership (Table 1). These metrics were collected using recent app duration, defined as the number of days between first and last app use. App-verified duration was recorded if a participant validated the download period via screenshot.

The primary engagement variable was minutes of app use. Four variations were analyzed: (1) objective unadjusted minutes, representing total app-verified minutes; (2) self-reported or "subjective" unadjusted minutes, representing total unverified self-reported minutes; (3) objective adjusted minutes, calculated as app-verified minutes adjusted for app-verified duration of use, expressed as minutes per year; and (4) self-reported or "subjective" adjusted minutes, calculated as self-reported minutes adjusted for self-reported duration of use, expressed as minutes per year. Adjusted variables accounted for the duration of access to the app (from the download date to the last use). As only a limited number of apps reported objective start dates, and app-verified duration correlated highly with self-reported duration, the adjusted variables were calculated using the time between the first and last reported use.

Table 1. Descriptive statistics for engagement outcomes among meditation app users.

Statistics	n	Mean (SD)	5th percentile	25th percentile	50th percentile (median)	75th percentile	95th percentile
Subjective							
Total minutes	483	3562.78 (8616.39)	4.10	76.00	420.00	2474.00	21735.10
Total sessions	452	108.43 (197.00)	0.58	9.40	37.33	130.30	407.25
Duration (days)	477	894.16 (854.20)	16.80	158.00	621.00	1342.00	2689.80
Minutes per session	454	16.60 (29.74)	3.00	5.81	10.51	18.09	35.41
Estimated minutes per month[a]	477	148.77 (304.56)	0.48	8.41	40.25	137.20	789.44
Estimated sessions per month[a]	437	9.35 (16.61)	0.13	0.93	3.29	11.05	34.79
Objective							
Total minutes	483	3358.69 (8607.36)	12.00	96.50	465.00	2410.00	21735.10
Total sessions	151	61.89 (194.90)	0.61	3.36	12.22	39.95	167.23
Minutes per session	151	49.21 (89.20)	1.98	7.26	25.92	55.90	154.08
Estimated minutes per month[a]	151	73.58 (223.73)	0.61	2.39	10.61	35.09	313.31
Estimated sessions per month[a]	148	9.44 (47.63)	0.02	0.10	0.66	2.32	17.29

[a]Estimated minutes per month and sessions per month were calculated by total engagement in minutes divided by duration of app use in years divided by 12.

Self-Reported Measures

Sociodemographic Information and Meditation History

Sociodemographic information included household income, education level, religion, app name, and approximate start and stop dates of use. Most data were self-reported via the survey. Prolific provided additional information, including age, sex, language, student and employment status, country of birth, and current residence. Regular practice information included minutes per session, sessions per day, and days per week. Meditation

XSL•FO
RenderX

history was assessed by asking participants to report their previous meditation experience in hours, ranging from 0-100 hours to 1000+ hours. See Table 1 for regular practice information, and Sections S3-S5 in Multimedia Appendix 1 for sociodemographic statistics, meditation app frequencies, and a detailed survey flow.

Attention Checks

Three attention checks were included in the survey to assess participant engagement. Participants failing 2 or more attention checks were excluded. The attention checks were designed to mimic the scale items within which they appeared; for example, "In general, select dissatisfied to show that you are paying attention."

The EuroQoL Health and Wellbeing Assessment—Short Form

The 9-item EuroQoL Health and Wellbeing (EQ-HWB-9) is a newly developed quality-of-life measure by Brazier and colleagues [57]. It assesses quality of life with a focus on health and well-being. The scale consists of 9 items, each rated from 1 (no difficulty, none of the time, and no physical pain) to 5 (unable, most or all of the time, and very severe physical pain). In this study, the 9 items of the EQ-HWB-9 demonstrated very good internal consistency (Cronbach α=0.873), and McDonald hierarchical omega was relatively high (ω=0.732). The EQ-HWB-9 was used with permission from the EuroQol Group.

The Kessler Psychological Distress Scale

The Kessler Psychological Distress Scale (K10) assesses psychological distress over the past 30 days [58]. This 10-item questionnaire, measuring anxiety and depressive symptoms, uses a 5-point scale ranging from 1 (none of the time) to 5 (all of the time). In this study, the scale demonstrated excellent internal consistency (Cronbach α=0.930; McDonald ω=0.799).

The Warwick-Edinburgh Mental Wellbeing Scale

The Short Warwick-Edinburgh Mental Wellbeing Scale (WEMWBS) was used to assess positive aspects of mental health [59]. This 7-item scale uses a 5-point response format, ranging from 1 (none of the time) to 5 (all of the time). In this study, the scale demonstrated high internal consistency (Cronbach α=0.879; McDonald ω=0.790).

The Satisfaction with Life Survey (Single Item)

The Satisfaction with Life Survey Single Item is an abbreviated version of the established Satisfaction with Life Survey (SWLS) [60]. The scale demonstrates reasonable criterion validity with the full SWLS (zero-order r=0.62-0.64). This single-item measure asks participants to rate their life satisfaction on a scale from 1 (extremely dissatisfied) to 7 (extremely satisfied).

7-Item Generalized Anxiety Disorder Scale

The 7-Item Generalized Anxiety Disorder Scale is used to screen for general anxiety symptoms [61]. Each item is rated from 0 (not at all) to 3 (nearly every day). In this study, the scale demonstrated high internal consistency (Cronbach α=0.899; McDonald ω=0.863).

Depression (8-Item Patient Health Questionnaire)

The 8-item Patient Health Questionnaire is used to assess depressive symptoms [62]. Each item is rated from 0 (not at all) to 3 (nearly every day). In this study, the scale demonstrated high internal consistency (Cronbach α=0.881; McDonald ω=0.781).

Self-Compassion Scale

The Self-Compassion Scale consists of 12 items assessing participants' ability to be compassionate toward themselves [63]. Each item is rated from 1 (almost never) to 5 (almost always). In this study, the scale demonstrated high internal consistency (Cronbach α=0.881; McDonald ω=0.656).

Self-Efficacy (6-Item Generalized Self-Efficacy)

The 6-item Generalized Self-Efficacy assesses self-efficacy, or an individual's perceived ability to achieve goals [64]. Each item is rated from 1 (not at all true) to 3 (exactly true). The scale demonstrated high internal consistency (Cronbach α=0.821; McDonald ω=0.788).

Readiness to Change 1-Item

The Readiness to Change 1-item assessment is a 10-point scale that measures an individual's preparedness to enact a behavioral change [37]. The scale ranges from 0 (not prepared to change) to 10 (already changing). It has been validated to reflect actual readiness in clinical contexts [37,65,66].

6-Item Digital Working Alliance Inventory

The 6-item Digital Working Alliance Inventory (DWAI-6) [5] is a rating scale that assesses the therapeutic alliance between an individual and their health care provider, adapted for smartphone interventions (ie, referring to "the app" rather than "the therapist"). Items are rated on a 7-point scale from 1 (strongly disagree) to 7 (strongly agree), with subscales evaluating goal alliance (agreement on goals), task alliance (agreement on tasks), and bond (connection between app and user). The overall scale demonstrated good internal consistency (Cronbach α=0.850; McDonald ω=0.830). The Goal subscale showed high consistency (Cronbach α=0.766), the Bond subscale demonstrated acceptable consistency (Cronbach α=0.676), and the Task subscale showed poor consistency (Cronbach α=0.402).

Common Factors Domains (Modum Process Outcome Questionnaire)

We used a subset of items from the Common Therapeutic Relationship Factors Questionnaire (Modum Process Outcome Questionnaire) to assess the therapeutic relationship beyond the DWAI-6 [67]. The original questionnaire, which focuses on the clinician-patient relationship, was adapted to refer to the app-user relationship (eg, "I am able to be open and honest when interacting with the app"). Three items from the 12-item scale were included, each rated from 1 (strongly disagree) to 7 (strongly agree), with a "not applicable" option. The items demonstrated low internal consistency (Cronbach α=0.612; McDonald ω<0.001), likely due to being an unintended subset. Consequently, results will be reported for each item individually rather than as a total score.

XSL•FO
RenderX

The Big Five Inventory Short Form 2 (BFI-S-2)

The Big Five Inventory Short Form 2 (BFI-S-2) is a 30-item questionnaire assessing 5 personality domains: Extraversion, Agreeableness, Conscientiousness, Negative Emotionality/Neuroticism, and Open-Mindedness [68]. Each subscale consists of 6 items. Internal consistency, assessed using Cronbach α and McDonald ω, ranged from acceptable (Extraversion, α=0.766; ω=0.636/Agreeableness, α=0.744; ω=0.549/Openness, α=0.785; ω=0.664) to high (Conscientiousness, α=0.812; ω=0.738/Negative Emotionality, α=0.884; ω=0.841).

User Mobile Application Rating Scale

The user Mobile Application Rating Scale (uMARS) is a 27-item instrument for assessing the quality of mobile apps [69]. The scale includes subscales evaluating engagement (referred to as "appeal" in this study for clarity: "Is the app fun/entertaining to use?"), functionality ("How accurately and quickly do the app features and components work?"), aesthetics ("How good does the app look?"), information ("Is the app content correct, well written, and relevant to the goal/topic of the app?"), perceived/subjective quality ("Would you pay for this app?"), and perceived impact ("This app has increased my knowledge/understanding of meditating"). The scale asks participants to rate app elements on a 5-point scale ranging from 1 (poor) to 5 (excellent), with specific descriptions for each item. The overall scale demonstrated high internal consistency (Cronbach α=0.866) but only moderate reliability (McDonald ω=0.606). The additional subscale for perceived impact was also highly consistent (Cronbach α=0.863; McDonald ω=0.761). The Functionality subscale demonstrated high internal consistency and reliability (Cronbach α=0.803; McDonald ω=0.739), whereas the Engagement (Cronbach α=0.734; McDonald ω=0.628), Aesthetics (Cronbach α=0.761; McDonald ω=0.658), and Information (Cronbach α=0.762; McDonald ω=0.688) subscales showed acceptable consistency and reliability. The Subjective Quality subscale was consistent (Cronbach α=0.629; McDonald ω=0.667). The Aesthetics subscale showed variable internal consistency (Cronbach α=0.762; McDonald ω=0.065), indicating unequal item contributions.

Analysis Plan

Sample Size Determination

Given an unknown effect size due to the absence of robust data, and guided by prior estimates of effect sizes for meditation apps [12], we aimed to calculate statistical power based on the smallest effect size we could reasonably detect. The target sample size was 1000 participants, providing 90% power to detect an effect of r=0.102, corresponding to a small effect. Because of recruitment challenges, the final sample size before analyzing the main engagement variable was 536, representing nearly 54% of our target population (n=1000). Although this reduction decreased statistical power (80% power to detect r=0.122), given that the a priori effect size was unknown, the study remained adequately powered to detect relatively small effects, albeit slightly larger than initially anticipated.

Planned Statistics

As outlined in the preregistration, we explored associations between user-related factors (Q1, H1), user-app relationship factors (Q2, H2), app-related factors (Q3, H3), mental health factors (Q5, H5), and predefined engagement variables. All engagement variables were heavily skewed and nonnormal (see Figure 1 and Sections S11-S14 in Multimedia Appendix 1); therefore, Spearman rho correlations were estimated. Variables that were significantly associated with any outcome variable were subsequently used to build regression models for each of the 4 outcomes. As transformations did not normalize the variables, untransformed variables were analyzed using robust regression with the "robustbase" package in R (R Foundation), employing an MM-type regression estimator with a bisquare redescending score function [70,71]. This method applies case weighting to account for nonnormality. All analyses were conducted in RStudio (R Foundation).

XSL•FO
RenderX

Figure 1. Histogram of self-reported practice hours (x axis) shown as the percentage of the sample (y axis), with values converted from minutes for visualization purposes.

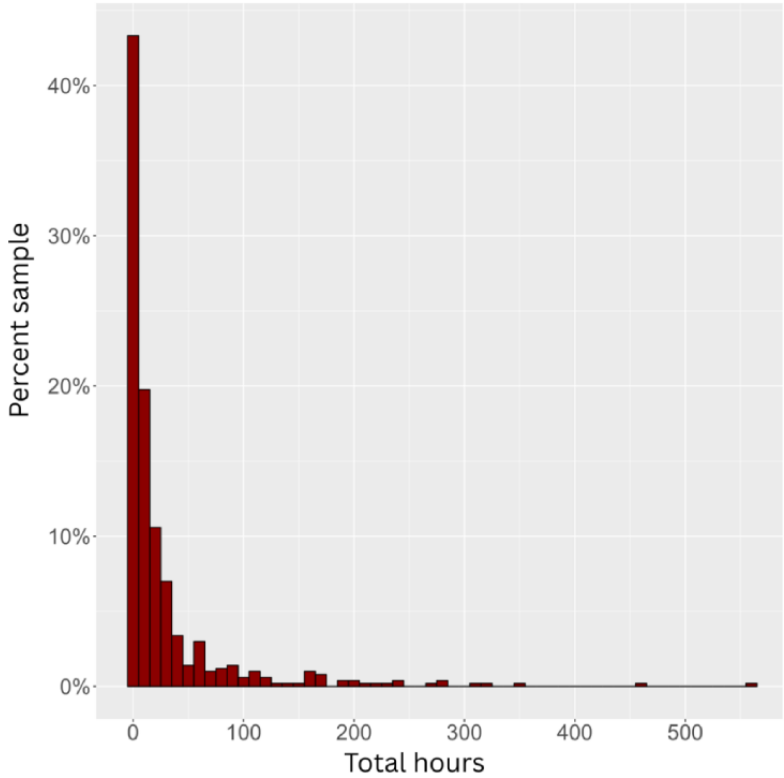

Regression Models

Robust linear regression was used to investigate which factors accounted for significant variance in engagement. For each of the 4 outcome variables—adjusted objective minutes, adjusted self-report minutes, objective minutes, and self-report minutes—a separate regression model was created using the respective measure as the outcome variable. No stepwise regression was implemented. Instead, independent variables were selected from user, mental health, and app factors that were significantly associated with at least one outcome measure in the correlation analyses, following correction for multiple comparisons.

Multiple Comparisons Correction

We explored correlations between the 4 engagement outcomes and related factors, applying a false discovery rate (FDR) correction to account for multiple comparisons [72].

Data Cleaning

Duplicate and invalid responses were removed. The data demonstrated weak correlations, positive skew, and extreme values, indicating potentially high variability or inconsistency (see Sections S6-S10 in Multimedia Appendix 1). After adjusting for app duration, we implemented several data-cleaning procedures. Intraindividual response validity calculation, LongString Identification [73], and inconsistency of responses on the BFI-S-2 [68] were each used to identify and exclude extreme cases; however, none of these approaches resulted in major changes to the results (see Sections S10-S12 in Multimedia Appendix 1). As specified in the preregistration, multivariate outliers were removed, identified as cases with a Mahalanobis distance greater than the 95th percentile on the BFI-S-2 (see Section 13 in Multimedia Appendix 1).

Results

Overall Engagement

Overall, most users completed only a few minutes across limited sessions. Despite generally low engagement, 134 out of 536 (25%) users reported more than 11 sessions per month (approximately 1 session every 3 days), while the top 5% (27/536) reported around 35 sessions per month (more than 1 session per day). These patterns align with prior findings, including a median of 90% of users dropping off within the first week of real-world use and an average 42% drop in participation in meditation app randomized controlled trials spanning 1-2 months [4,12]. Notably, 402 (75%) participants reported more than 9 sessions, which contrasts with prior findings suggesting that most users disengage completely within a week of download. However, only the top 5% (25/536, 4.7%) engaged at levels consistent with clinically meaningful change [16], while the top 25% (134/536) engaged at levels comparable to the dose of mindfulness-based interventions [27].

Participant Characteristics: Descriptives

Sociodemographic Features

Participants (N=536) ranged from 18 to 70 years of age (mean 36.56 years, SD 10.68 years) and were predominantly female (n=366, 68.3%). See Figure 2 for participant flow. Most resided in the United States 253 (47.20%) or the United Kingdom

XSL•FO
RenderX

(n=226, 42.2%), with smaller proportions from Australia (n=27, 5%), Canada (n=21, 3.9%), and New Zealand (n=6, 1.1%). The majority identified as White (n=422, 78.73%) and were highly educated, with 387 (72.20%) holding at least a bachelor's degree. Participants were also relatively wealthy, with nearly one-third reporting a combined income of $100,000 or more (n=145, 27.05%). Note that income brackets were not adjusted across countries. Nearly half of the participants reported no religious affiliation (n=264, 49.25%). The most frequently used apps were Headspace (n=191, 35.6%) and Calm (n=123, 22.9%), which together accounted for 58.6% of the sample (n=314). Full details are provided in Sections S3 and S4 in Multimedia Appendix 1.

Figure 2. Participant flow.

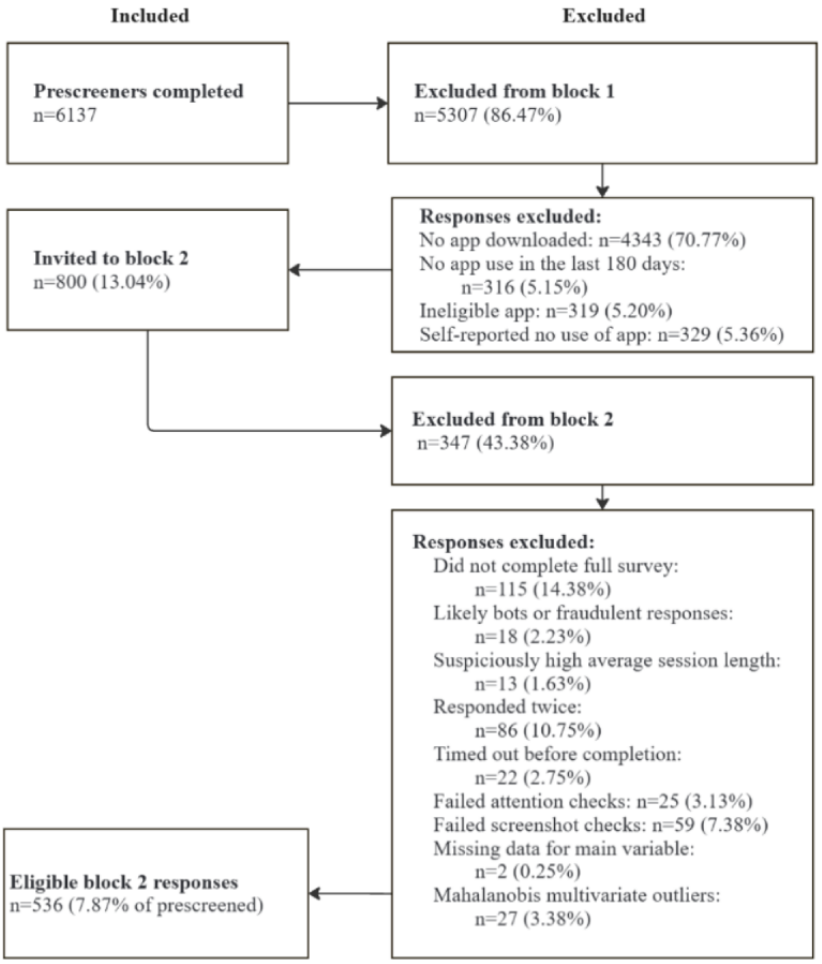

Meditation Experience

Most users (n=330, 61.6%) reported between 0 and 100 hours of overall meditation experience. Meditation experience varied across meditation apps (χ^2_4=34.18, *P*<.001); users with 0-100 hours were most likely to use Headspace (124/377, 32.9%), followed by Calm (75/377, 19.9%) and Insight Timer (25/377, 6.6%; see Table 2).

Table 2. Frequency and relative percentage of Calm, Headspace, and Insight Timer users by self-reported meditation experience level (n=377).

Duration (hours)	Calm, n (%)	Headspace, n (%)	Insight Timer, n (%)	Total, n (%)
0-100	75 (19.9)	124 (32.9)	25 (6.6)	224 (59.4)
101-1000	35 (9.3)	46 (12.2)	46 (12.2)	127 (33.7)
1001+	6 (1.6)	11 (2.9)	9 (2.4)	26 (6.9)
Total	116 (30.8)	181 (48.0)	80 (21.2)	377 (100)

Engagement

After excluding invalid responses and duplicates, engagement levels remained low, with a positive skew for both hours and sessions (see Figures 2 and 3). Adjusting for app duration showed low engagement regardless of how long the app had been available to users (see Figures 4 and 5). Estimated minutes per month and sessions per month were adjusted within each user to provide clearer engagement metrics. The median number

XSL•FO
RenderX

of sessions per month was 3.29. With a median of 40.2 minutes per month, this equated to roughly three 12-minute sessions.

Figure 3. Histogram showing the total number of self-reported sessions (x axis) by the percentage of participants (y axis).

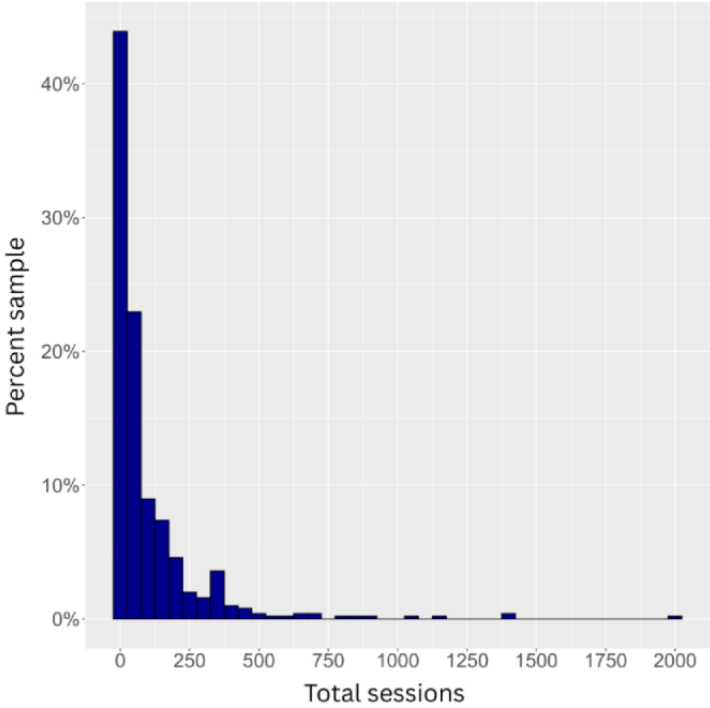

Figure 4. Total hours of practice by categorical groups based on time since app download. The y axis is truncated at 50 for visualization purposes, capturing 75% of the data. Horizontal bars indicate the median value for each category.

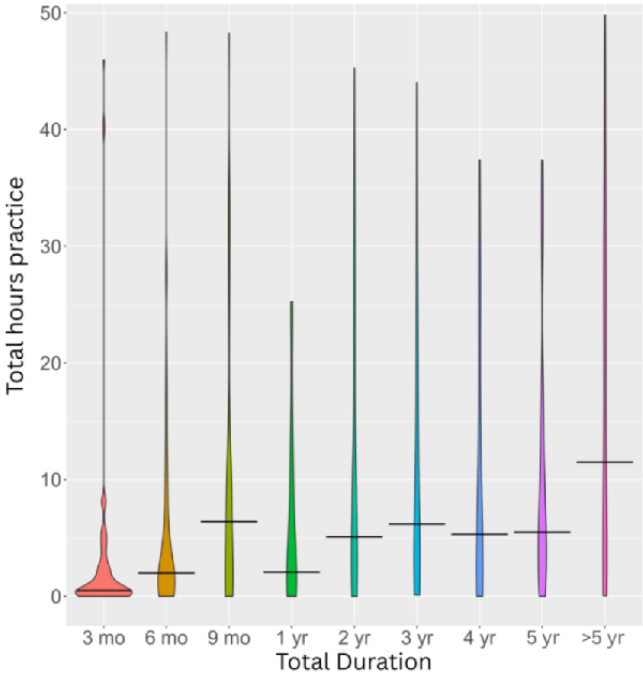

XSL•FO
RenderX

Figure 5. Total sessions by categorical groups based on time since app download. The y axis has been truncated for visualization purposes, capturing 75% of the data. Horizontal bars indicate median values for each category.

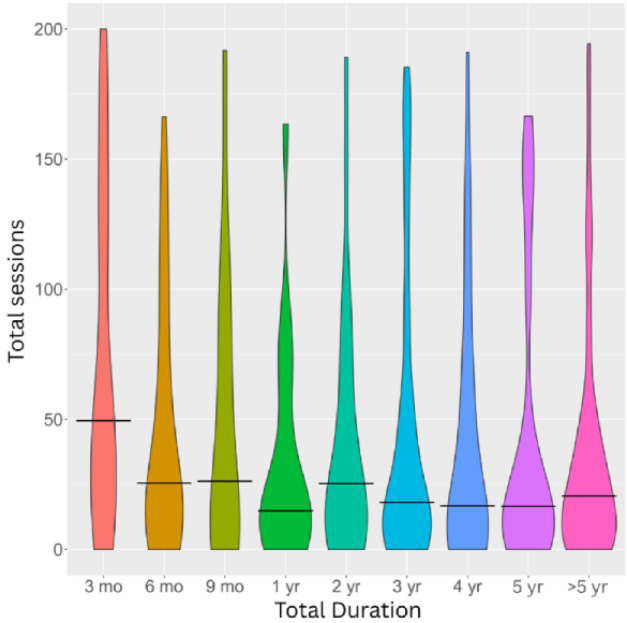

Engagement Statistics

Engagement variables were highly intercorrelated (r=0.495-0.999; see Section S14 in Multimedia Appendix 1). Objective adjusted minutes were the most reliable variable, but the sample was small and limited to 2 apps (n=156; Headspace and Waking Up). By contrast, subjective minutes with a subjective start date had a sample size 3 times larger (n=536). Given the strong association between objective and subjective start dates (r=0.783, P<.001), adjusted subjective minutes were calculated using subjective duration as the denominator to maximize sample size.

Categorical Demographic Factors and Engagement

Categorical demographic associations with engagement are reported in Table 3. Being female was associated with lower engagement on 1 outcome, while residing in the United Kingdom was associated with higher engagement across 3 of the 4 outcome variables.

Table 3. Point-biserial correlations for the association between noncontinuous variables and engagement.

Variables	Objective minutes	Subjective minutes	Adjusted objective	Adjusted subjective
Sex[b]	–0.019	0.080	–0.296[a]	–0.016
Residence				
Australia	–0.027	–0.016	–0.004	–0.040
United States	–0.071	–0.069	–0.065	–0.068
United Kingdom	0.102[a]	0.110[a]	0.075	0.106[b]
Canada	–0.032	–0.062	–0.029	–0.068
New Zealand	–0.021	–0.031	–0.007	0.038

[a]P<.05 without multiple comparisons correction.

[b]P<.05 with multiple comparisons correction. For the biserial sex correlation, 1=female, 0=male.

Engagement by App

Three 1-way analyses of variance indicated a significant effect of app type on engagement for subjective minutes ($F_{2,362}$=9.03, P=.002, η^2=.05) and adjusted subjective minutes ($F_{2,358}$=7.63, P=.001; see Sections S15 and S16 in Multimedia Appendix 1). No significant differences were found for objective minutes ($F_{2,362}$=0.972, P=.38).

User Factors

We defined robust associations as those present across 3 or more of the 4 engagement outcomes. Among user factors, only 9 of 20 variables met this criterion (see Figure 6): age, openness (BFI-S-2), readiness to change, expectation match, expectations for sleep, expectations for anxiety, expectations for happiness, expectations for thriving, and expectations for performance enhancement. After FDR correction, only 4 of 20 remained: age, readiness to change, expectations for sleep, and expectations for thriving (see Section 17 in Multimedia Appendix 1 for CIs).

XSL•FO
RenderX

Figure 6. Heatmap of correlation results for user factors and adjusted self-report, adjusted objective, total self-report, and total objective minutes of meditation app use. The heatmap illustrates the direction and magnitude of Spearman correlations, as shown in the legend. Numbers indicate correlation coefficients. Asterisks denote correlations significant at *P*<.05 after correction for multiple comparisons. Nonsignificant correlations are omitted. Outcomes include self-reported total minutes of meditation app use, self-reported minutes adjusted for app duration, objectively verified total minutes of use (via app screenshot), and objective minutes adjusted for app duration.

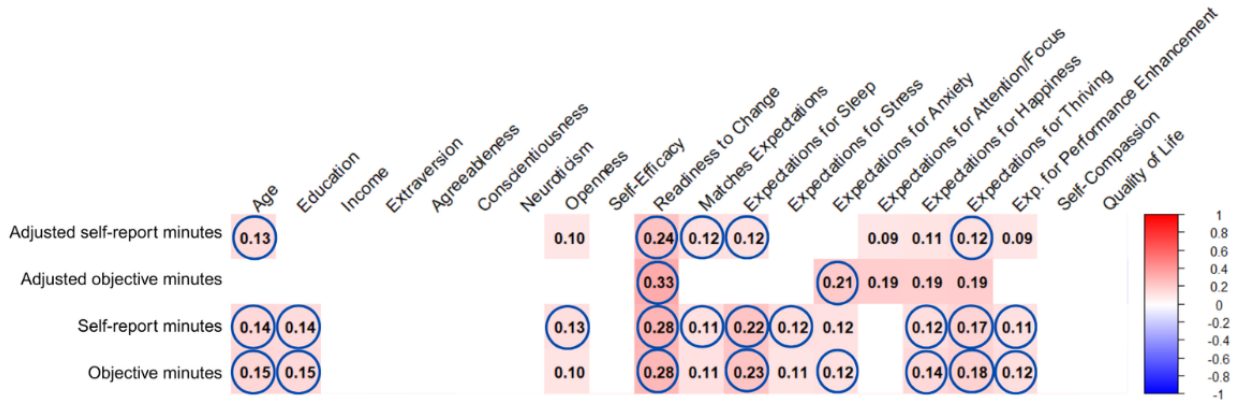

Mental Health Factors

Self-reported stress, depression, and psychological distress were negatively associated with app use. Specifically, distress was negatively associated with adjusted self-reported minutes, depression with unadjusted self-reported minutes, and current stress with both unadjusted self-reported minutes and objective minutes. However, no mental health factors remained significantly associated with any engagement outcome after correction for multiple comparisons (see Figure 7).

Figure 7. Heatmap of correlations between mental health factors and adjusted self-report, adjusted objective, total self-report, and total objective minutes of meditation app use. The heatmap shows the direction and magnitude of Spearman correlations, as indicated by the legend. Numbers represent correlation coefficients. Circles denote significant correlations at *P*<.05 after correction for multiple comparisons. Nonsignificant correlations are omitted. Outcomes include self-reported total minutes of meditation app use, self-reported minutes adjusted for app duration, objectively verified total minutes of use (via app screenshot), and objective minutes adjusted for app duration.

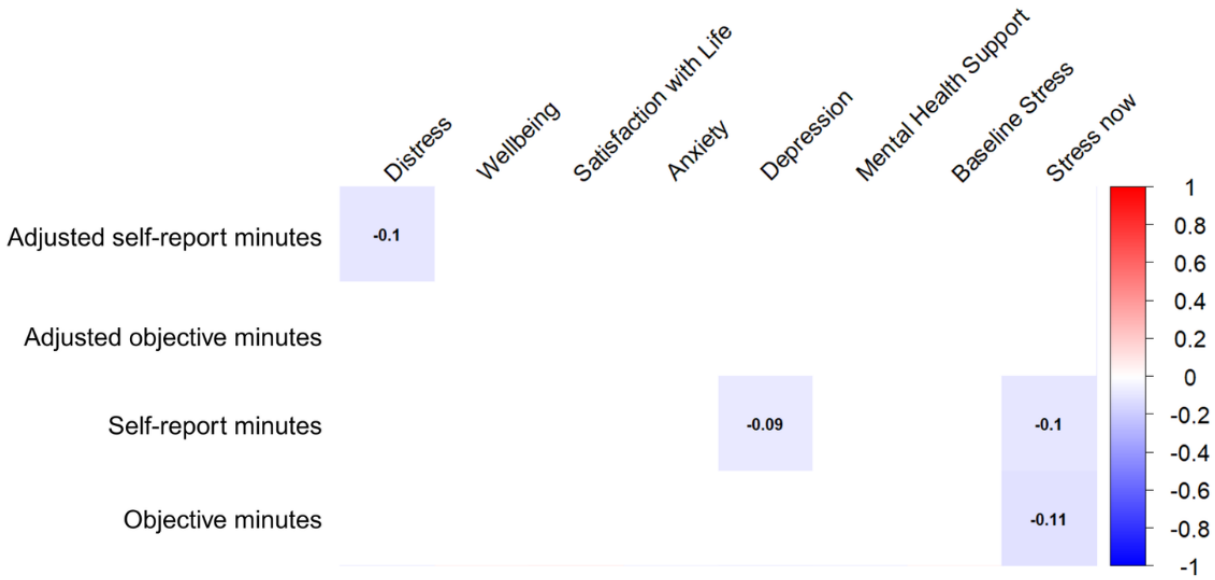

App Factors

DWAI-6 Total, uMARS Appeal, uMARS Perceived Quality, and Perceived Impact were associated with 3 of the 4 outcome variables. Of the app factors investigated, 7 were associated with at least one engagement outcome after FDR correction (see Figure 8): DWAI-6 Total, as well as Goal, Bond, and Task subscales, and uMARS Appeal, Perceived Quality, and Perceived Impact.

XSL•FO
RenderX

Figure 8. Heatmap of correlations between app factors and meditation app use outcomes. Displayed are Spearman correlation coefficients (numbers) showing the direction and magnitude of associations as indicated in the legend. Circles denote correlations significant at *P*<.05 after multiple comparison correction, and nonsignificant correlations are omitted. Outcomes include self-reported total minutes of app use, self-reported minutes adjusted for app duration, objective total minutes of use (verified via app screenshot), and objective minutes adjusted for app duration. DWAI: Digital Working Alliance Inventory; uMARS: user Mobile Application Rating Scale.

Outcome Regression Models

All models demonstrated a reasonable fit, explaining 12%-16% of the variance (see Table 4 and Section 18 in Multimedia Appendix 1). Significant (*P*<.05) predictors in 1 or more regression models included education, readiness to change, expectations for sleep, expectation match, the Perceived Quality subscale of the uMARS, and the Openness subscale of the BFI-S-2.

XSL•FO
RenderX

Table 4. Regression model coefficients for adjusted objective minutes, total objective minutes, adjusted self-report minutes, and total self-report minutes.

Standardized β coefficients for predictors in models 1–4	Adjusted objective minutes[a], standardized β coefficients	Objective minutes[b], standardized β coefficients	Adjusted self-report minutes[c], standardized β coefficients	Subjective minutes[d], standardized β coefficients
Intercept	–0.260[e]	–0.228[e]	–0.355[e]	–0.371[e]
User factors				
Sex	0.001	<0.001	–0.008	0.002
Country of residence (United Kingdom)	<.0001	0.002	0.008	0.004
Age	0.003	<–.001	0.004	–0.001
Education	0.033	0.153[e]	0.532[e]	0.237[e]
Big Five Inventory Short form 2 openness	0.010[f]	0.003	0.005	0.006
Readiness to change	<0.001	0.005[f]	0.027[g]	0.008[f]
Expectations (match)	<0.001	–0.003	0.023[f]	–0.004
Expectations for sleep	–0.002	0.004	–0.008	0.009[f]
Expectations for stress	<0.001	<–.001	0.012	<.001
Expectations for anxiety	0.001	–0.001	–0.020	–0.003
Expectations for happiness	–0.005	0.002	0.009	<.001
Expectations for thriving	0.008	0.003	–0.006	0.006
Expectations for performance enhancement	–0.004	–0.003	–0.015	–0.003
App factors				
DWAI-6[h] (Goal)	–0.009	0.004	–0.020	0.008
DWAI-6 (Bond)	–0.004	–0.003	–0.013	–0.002
DWAI-6 (Task)	0.010	0.002	0.024	0.004
uMARS[i] (Appeal)	–0.012	<.001	–0.016	0.005
uMARS (Perceived Quality)	0.025[g]	0.003[g]	0.041[e]	0.010[f]
uMARS (Perceived Impact)	–0.006	<–.001	–0.003	0.002
Adjusted R^2	0.158	0.150	0.126	0.137

[a]App-verified minutes of use per year adjusted for total duration of use in years.
[b]Total app-verified minutes of use.
[c]Self-report minutes of use per year adjusted for total duration of use in years.
[d]Self-report total minutes of use.
[e]$P<.001$.
[f]$P<.05$.
[g]$P<.01$.
[h]DWAI-6: 6-item Digital Working Alliance Inventory.
[i]uMARS: user Mobile Application Rating Scale.

Discussion

Principal Findings

We examined factors associated with engagement in popular meditation apps among 536 participants. Consistent with prior findings, most participants engaged minimally. Although apps were available for an average of 894 days (about 2.5 years), participants reported an average of 108 sessions, while app-verified data from about one-third of participants indicated 62 sessions on average. Half of the sample engaged in 3 or fewer sessions per month. Notably, engagement did not increase with longer app availability, suggesting a pattern of persistently low overall engagement.

Few significant correlations between individual user factors and engagement were observed, most of which were small in magnitude (r=0.09-0.30), with a few reaching the moderate range (r=0.30-0.50). After correction for multiple comparisons, positive associations with engagement remained for male sex, older age, higher education level, readiness to change, and expectations of the app for stress reduction, sleep improvement,

XSL•FO
RenderX

anxiety reduction, happiness, thriving, and performance enhancement. These results suggest that older, more educated users with greater readiness to change and higher expectations of the app are more likely to engage regularly. App factors—including perceived appeal, quality, and impact—were consistently associated with higher engagement, as were the 3 digital working alliance subscales (Goal, Task, and Bond). This indicates that both perceptions of the app and the perceived relationship with it may be important determinants of engagement.

User Factors Related to Use

Sociodemographics

Education was the variable most consistently associated with higher engagement. Meditation is more common among individuals who are White, middle-aged, wealthier, and better educated [41]. Lower levels of education have been linked to earlier disengagement or a failure to engage in meditation at all [41]. Higher education is also associated with greater engagement in mindfulness practices in US nationally representative surveys [11]. Lower levels of education have also been linked to poorer health outcomes [74,75] and are related to health literacy, which partially mediates health-promoting behaviors [76,77]. While lower education may contribute to poorer health outcomes and lower health literacy, other social factors may also reduce engagement. For example, individuals with lower education often have more fragmented leisure time, leaving less opportunity for regular, recurrent activities [78]. Male sex was associated with 1 engagement measure, which contrasts with prior research showing that females are generally more likely to engage in meditation practice, even after controlling for other demographic factors [12,42,79]. Previous studies have also found that males may demonstrate greater persistence in meditation [80]. As this correlation was observed only for adjusted objective use (available for Headspace and Waking Up), it may reflect patterns specific to users of these apps rather than meditation app use more broadly [40]. Notably, sex did not emerge as a significant predictor in the regression models.

Personality

Of the personality factors, only openness was related to engagement. Openness reflects general curiosity and a willingness to explore novel perspectives of one's subjective experience [81]. Individuals higher in openness are more likely to try meditation initially and persist despite encountering difficulties. Openness has also been associated with meditation practice outside of group meditation class settings [45]. In contrast to prior research, we found no associations between engagement and conscientiousness, extraversion, agreeableness, or neuroticism [55,82,83]. While conscientiousness was not related to engagement in this study, it has previously been linked to positive attitudes toward practice [43]. Similarly, neuroticism showed no association with engagement here, although prior work has linked it to perceiving more barriers to practice [84,85].

Mental Health

None of the 8 mental health factors were significantly associated with engagement. Previous research has found meditation apps to be modestly effective for depression and anxiety [12,19], potentially serving as a form of self-managed treatment for individuals facing barriers to mental health care [41]. However, no such associations with mental health factors were observed in this study. In a previous study, motivation for mental health was negatively associated with app use [40]. Meditation can negatively impact mental health [86]. While these meditation-related adverse events do not always result in impairment, about half of meditators report experiencing an adverse effect, and 9.1% report functional impairment as a result [86]. Individuals who do not experience benefits or who encounter adverse effects may disengage shortly after download. Furthermore, meditating for mental health reasons has been negatively associated with the total amount of meditation practice completed over the long term [27,87]. Individuals with higher lifetime meditation practice often shift toward spiritual motivations as their practice progresses [29]. However, the retrospective design of our study limits causal inferences.

App Factors

uMARS

Five of the 6 uMARS subscales were associated with engagement. Previous research suggests that aesthetics and appeal relate to meditation app engagement [53], although in our study, aesthetics were not robustly associated after FDR correction. The Perceived Quality subscale showed the strongest association (r=0.51), indicating that user perceptions may drive both usage and beliefs in the app's effectiveness. Perceived impact was also robustly associated with engagement. Given the retrospective design, survivorship bias should be considered: users who continued using the apps likely enjoyed them, while those who did not may have stopped. It is also possible that users who experienced benefits from their chosen app developed increasingly positive app appraisals over time.

Digital Working Alliance

The DWAI-6 Goal, Bond, and Task subscales, as well as the overall score, were associated with engagement, consistent with prior findings [88]. All subscales correlated with adjusted objective minutes—the most reliable outcome measure, computed using app-verified minutes and download date—but this could only be calculated for apps that provide download dates (Headspace and Waking Up; n=151). Therapeutic alliance and engagement may promote each other [88]. While therapeutic alliance is considered important in digital mental health [89], current measures are adaptations of traditional, human-centered alliance scales. Incorporating human-computer interaction perspectives may provide greater nuance, particularly for anthropomorphic scale items [90]. Despite this limitation, therapeutic alliance with apps remains relevant to engagement, as alignment between a user's goals and perceived app support may encourage continued use.

One consideration for both the uMARS and DWAI-6 is that several subscales demonstrated relatively poor internal consistency. The reliability of the uMARS Engagement and

XSL•FO
RenderX

Subjective Quality subscales, as well as the DWAI-6 Bond and Task subscales, ranged from acceptable to poor, which reduces confidence in the constructs being measured.

Expectations for Efficacy

Higher expectations of efficacy across 6 of the 7 domains assessed (sleep, stress, anxiety, happiness, thriving, and performance enhancement) were generally associated with higher engagement, with the exception of expectations for attention/focus. Only expectations for sleep were significant in the regression model. These findings align with our predictions. Experimental and prospective studies have shown that failing to meet expectations is more predictive of behavior than matched expectations [91,92]. Unmet or low expectations negatively influence engagement and perceived usefulness, whereas met or exceeded expectations positively affect behavior and perceptions [92]. Expectations are closely linked to app ratings, as features such as goal setting and feedback enhance beliefs in an app's effectiveness [34]. These features can also foster positive experiences of progress, creating a feedback loop that promotes further engagement [33,48]. In the absence of human interaction, the relationship between a user and an app is shaped by the "user journey"—the path a user follows through the app's design. Persuasive design can help establish and meet user expectations.

A general rating of whether expectations were met was weakly associated with engagement. This result aligns with literature suggesting that matched expectations have a positive influence on behavior and mismatched expectations have a negative effect [93]. It is worth noting that we asked, "To what extent did your experience match your initial expectations?" without specifying which expectations participants should consider. Consequently, this approach may have captured only an overall impression of expectation match.

Readiness to Change

Readiness to change showed robust, moderate associations and accounted for a significant proportion of variance in the regression model. The readiness-to-change ruler used in this study is actively employed in behavior change interventions and is based on the Transtheoretical Model of Change, which conceptualizes behavior change in stages [32,94,95]. Readiness to change shows promise as one of the most predictive factors of actual behavior change, as it is conceptually closely linked to both motivation and behavior. These findings align with broader evidence connecting readiness ratings to actual behavior change, particularly in health-related contexts [63,95,96]. This relationship could inform app design, allowing offerings to be tailored to users' readiness levels. The same single-item measure used in our study could be implemented immediately after app download to tailor the length, complexity, and type of practice to users' readiness levels. For example, users with lower readiness could be offered shorter, simpler meditations or psychoeducational content about meditation to reduce perceived barriers and enhance understanding of the practice.

Self-Efficacy and App Ratings in Building Habits

Contrary to our expectations, self-efficacy was not related to engagement. Previous research on habit formation suggests that self-efficacy may support the maintenance of a target behavior before a habit is established. There is limited evidence that self-efficacy promotes habit-building [33,48,97] and increases with ongoing meditation practice [98]; however, results are mixed [99,100]. One likely reason self-efficacy did not predict engagement is that expectations, perceptions, and habit formation played larger roles. A person may believe they can achieve a goal, but if they are not committed or do not perceive long-term utility, they may lack motivation to engage. This may explain why readiness to change was associated with engagement, whereas self-efficacy was not.

Limitations

One key limitation of this study is that its retrospective design precludes causal inferences, although research on meditation app engagement is generally scarce. Additionally, we cannot confirm detailed usage patterns, such as extended gaps or cessation points; however, our estimates of sessions per month provide a rough indication of practice regularity. This study included cross-app comparisons, which few prior studies have conducted. Such comparisons are valuable, given that all therapeutic alliance subscales and half of the uMARS subscales were associated with engagement after correction for multiple comparisons. However, by not focusing on a specific app, the sample was disproportionately composed of users of the most popular apps.

Another limitation was that our most reliable outcome variable—objective minutes adjusted for verified app duration—was restricted to apps that displayed the month or year of joining. As a result, the sample for objective-adjusted minutes comprised only about one-third of the self-reported sample. Nevertheless, objective minutes were highly correlated with self-reported minutes, which may mitigate some concerns, although it is possible that individuals who can view their app-recorded minutes rely on these records when self-reporting.

Our data quality may have been influenced by self-selection, socioeconomic skew, and the compensation structure in our Prolific sample. Nevertheless, research indicates that among popular online survey platforms, Prolific consistently provides high-quality data across a wide range of measures [101]. Additionally, our data may have been skewed by the overrepresentation of the most popular meditation apps, limiting the generalizability of the findings to less popular apps or those with a narrower focus.

A final significant issue concerns what the outcome measures captured. While meditation was the central focus of the included apps, many also offer alternative exercises that contribute to the measured minutes, including—but not limited to—breathwork, sleep stories, and podcasts. This is an issue because sleep stories may continue running for hours after an individual falls asleep and be recorded as meditation. Future studies that can distinguish between different activities will provide more accurate statistics on meditation engagement.

Future Directions

A group-level comparison of engagers and disengagers could reveal cluster effects, where active users share similar characteristics. Baumel et al [4] observed a drop-off trend in a

XSL•FO
RenderX

large sample but noted that understanding precisely why people engage or disengage during this period would be of interest. While there is a high use rate among those who continue engaging beyond the first week, this represents only a small portion of users. By including all users over the past 180 days, we obtain a general picture of app use across the population; however, this approach results in a large sample of disengagers and only a small sample of active users, limiting our power to detect effect sizes within the subsample of engagers.

Longitudinal analysis could directly examine temporal and causal aspects of engagement, account for changes in contributing factors over time, and provide a clearer understanding of baseline predictors. For example, longitudinal data could track whether changes in mental health outcomes influence engagement. This study did not find any significant associations for mental health outcomes that survived multiple comparisons. However, we relied on participants' reports of mental health status following meditation app use. Apps have been shown to reduce outcomes such as stress, depression, and anxiety [12], and such changes could positively or negatively reinforce app use. Moods, circumstances, and lifestyles can fluctuate widely over extended periods. User ratings of apps using scales such as the uMARS may better explain engagement when app rating and user engagement occur close together. Longitudinal analysis also allows for baseline measurement of variables, such as expectations, which can then be compared with actual experiences at follow-up. The low proportion of variability accounted for suggests that factors outside the model have a significant impact on engagement.

Conclusions

This study aimed to explore factors influencing engagement with popular meditation apps, highlighting a substantial early drop-off. Although the models accounted for only a small proportion of overall variance, the findings emphasize the importance of user characteristics and app quality in sustaining engagement. This exploratory study aimed to examine a wide range of factors potentially relevant to meditation app engagement. The results indicated that older, more educated users, as well as those with higher expectations of apps and greater readiness to change, were more likely to engage with the apps regularly.

Acknowledgments

We acknowledge the coding contributions of Alex Burger and Karen Trapani, as well as the administrative support of Cathleen Benevento. This study would not have been possible without the support of the Contemplative Studies Centre and the Melbourne School of Psychological Sciences community at the University of Melbourne. Funding was provided to establish the Contemplative Studies Centre via a philanthropic gift from the Three Springs Foundation, Pty, Ltd. No original content in the manuscript was generated by artificial intelligence. The authors occasionally used the free version of Grammarly (Grammarly, Inc) for content flow, such as identifying extraneous words in a sentence.

Data Availability

Data and analytic code are uploaded to the Open Science Framework registry [102].

Authors' Contributions

Conceptualization: JA, JD, NTVD, JG
Data curation: JA, JD, PW, NTVD
Formal analysis: JA, JD, NTVD
Funding acquisition: NTVD
Methodology: JA, JD, FM, JG, NTVD
Resources: JA, JD, JG, NTVD
Visualization: JA, NTVD
Writing – original draft: JA, JG, NTVD
Writing – review & editing: JA, JG, SDA, NTVD

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional analysis.
[DOCX File , 727 KB - jmir_v28i1e71960_app1.docx]

References

1. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022 Feb;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]

XSL•FO
RenderX

2. Lattie EG, Stiles-Shields C, Graham AK. An overview of and recommendations for more accessible digital mental health services. Nat Rev Psychol 2022 Feb;1(2):87-100 [FREE Full text] [doi: 10.1038/s44159-021-00003-1] [Medline: 38515434]
3. Schueller S, Hunter J, Figueroa C, Aguilera A. Use of digital mental health for marginalized and underserved populations. Curr Treat Options Psych 2019 Jul 5;6(3):243-255 [FREE Full text] [doi: 10.1007/s40501-019-00181-z] [Medline: 38515434]
4. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567-e14255 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
5. Goldberg SB, Baldwin SA, Riordan KM, Torous J, Dahl CJ, Davidson RJ, et al. Alliance with an unguided smartphone app: validation of the digital working alliance inventory. Assessment 2022 Sep 25;29(6):1331-1345 [FREE Full text] [doi: 10.1177/10731911211015310] [Medline: 34000843]
6. Lutz A, Slagter HA, Dunne JD, Davidson RJ. Attention regulation and monitoring in meditation. Trends Cogn Sci 2008 Apr 16;12(4):163-169 [FREE Full text] [doi: 10.1016/j.tics.2008.01.005] [Medline: 18329323]
7. Kabat-Zinn J. Mindfulness-based interventions in context: past, present, and future. Clinical Psychology: Science and Practice 2003;10(2):144-156 [FREE Full text] [doi: 10.1093/clipsy.bpg016]
8. Lam S, Riordan K, Simonsson O, Davidson R, Goldberg S. Who sticks with meditation? Rates and redictors of persistence in a population-based sample in the USA. Mindfulness (N Y) 2023 Jan 06;14(1):66-78 [FREE Full text] [doi: 10.1007/s12671-022-02061-9] [Medline: 36777474]
9. Lee RL. Review of headspace: meditation and sleep. Fam Syst Health 2023 Mar 06;41(1):114-116. [doi: 10.1037/fsh0000737] [Medline: 36951700]
10. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11-116 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
11. Olano HA, Kachan D, Tannenbaum SL, Mehta A, Annane D, Lee DJ. Engagement in mindfulness practices by U.S. adults: sociodemographic barriers. J Altern Complement Med 2015 Feb;21(2):100-102 [FREE Full text] [doi: 10.1089/acm.2014.0269] [Medline: 25685958]
12. Gál É, tefan S, Cristea IA. The efficacy of mindfulness meditation apps in enhancing users' well-being and mental health related outcomes: a meta-analysis of randomized controlled trials. J Affect Disord 2021 Jan 15;279(1):131-142 [FREE Full text] [doi: 10.1016/j.jad.2020.09.134] [Medline: 33049431]
13. Crane RS, Brewer J, Feldman C, Kabat-Zinn J, Santorelli S, Williams JMG, et al. What defines mindfulness-based programs? The warp and the weft. Psychol Med 2017 Apr;47(6):990-999. [doi: 10.1017/S0033291716003317] [Medline: 28031068]
14. Goldberg SB, Tucker RP, Greene PA, Davidson RJ, Wampold BE, Kearney DJ, et al. Mindfulness-based interventions for psychiatric disorders: a systematic review and meta-analysis. Clin Psychol Rev 2018 Feb;59(6):52-60 [FREE Full text] [doi: 10.1016/j.cpr.2017.10.011] [Medline: 29126747]
15. Galante J, Friedrich C, Collaboration of Mindfulness Trials (CoMinT) 3, Dalgleish T, Jones PB, White IR, Collaboration of Mindfulness Trials (CoMinT). Individual participant data systematic review and meta-analysis of randomised controlled trials assessing adult mindfulness-based programmes for mental health promotion in non-clinical settings. Nat Ment Health 2023 Jul 10;1(7):462-476 [FREE Full text] [doi: 10.1038/s44220-023-00081-5] [Medline: 37867573]
16. Bowles NI, Davies JN, Van Dam NT. Dose-response relationship of reported lifetime meditation practice with mental health and wellbeing: a cross-sectional study. Mindfulness (N Y) 2022 Feb;13(10):2529-2546. [doi: 10.1007/s12671-022-01977-6] [Medline: 36193220]
17. Schultchen D, Terhorst Y, Holderied T, Stach M, Messner EM, Baumeister H, et al. Stay present with your phone: a systematic review and standardized rating of mindfulness apps in European app stores. Int J Behav Med 2021 Oct 10;28(5):552-560. [doi: 10.1007/s12529-020-09944-y] [Medline: 33215348]
18. Spijkerman MPJ, Pots WTM, Bohlmeijer ET. Effectiveness of online mindfulness-based interventions in improving mental health: a review and meta-analysis of randomised controlled trials. Clin Psychol Rev 2016 Apr;45(4):102-114 [FREE Full text] [doi: 10.1016/j.cpr.2016.03.009] [Medline: 27111302]
19. Sommers-Spijkerman M, Austin J, Bohlmeijer ET, Pots W. New evidence in the booming field of online mindfulness: an updated meta-analysis of randomized controlled trials. JMIR Ment Health 2021 Jul 19;8(7):e28168-e28114 [FREE Full text] [doi: 10.2196/28168] [Medline: 34279240]
20. Jiang A, Rosario M, Stahl S, Gill JM, Rusch HL. The effect of virtual mindfulness-based interventions on sleep quality: a systematic review of randomized controlled trials. Curr Psychiatry Rep 2021 Jul 23;23(9):62 [FREE Full text] [doi: 10.1007/s11920-021-01272-6] [Medline: 34297230]
21. DiMatteo MR, Giordani PJ, Lepper HS, Croghan TW. Patient adherence and medical treatment outcomes: a meta-analysis. Med Care 2002 Sep;40(9):794-811. [doi: 10.1097/00005650-200209000-00009] [Medline: 12218770]
22. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on Internet-based mental health interventions — a systematic review. Internet Interventions 2014 Oct;1(4):205-215. [doi: 10.1016/j.invent.2014.08.003]
23. Fleming T, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018 Jun 06;20(6):e199-e215 [FREE Full text] [doi: 10.2196/jmir.9275] [Medline: 29875089]

XSL•FO
RenderX

24. Szinay D, Jones A, Chadborn T, Brown J, Naughton F. Influences on the uptake of and engagement with health and well-being smartphone apps: systematic review. J Med Internet Res 2020 May 29;22(5):e17572 [FREE Full text] [doi: 10.2196/17572] [Medline: 32348255]
25. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017 Jun 29;7(2):254-267 [FREE Full text] [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
26. Lam SU, Kirvin-Quamme A, Goldberg SB. Overall and Differential Attrition in Mindfulness-Based Interventions: A Meta-Analysis. Mindfulness (N Y) 2022 Nov;13(11):2676-2690 [FREE Full text] [doi: 10.1007/s12671-022-01970-z] [Medline: 36506616]
27. Strohmaier S. The relationship between doses of mindfulness-based programs and depression, anxiety, stress, and mindfulness: a dose-response meta-regression of randomised controlled trials. Mindfulness 2020 Mar 02;11(6):1315-1335 [FREE Full text] [doi: 10.1007/s12671-020-01319-4] [Medline: 27966189]
28. Yik LL, Ling LM, Ai LM, Ting AB, Capelle DP, Zainuddin SI, et al. The effect of 5-minute mindfulness of peace on suffering and spiritual well-being among palliative care patients: a randomized controlled study. Am J Hosp Palliat Care 2021 Sep 28;38(9):1083-1090. [doi: 10.1177/1049909120965944] [Medline: 33078627]
29. Bowles NI, Van Dam NT. Dose-response effects of reported meditation practice on mental-health and wellbeing: a prospective longitudinal study. Appl Psychol Health Well Being 2025 Aug 20;17(4):e70063-e71090. [doi: 10.1111/aphw.70063] [Medline: 40785526]
30. Bandura A. Social Foundations of Thought and Action: A Social Cognitive Theory. Hoboken, NJ: Prentice-Hall; 1986.
31. Ajzen I. The theory of planned behavior. Organizational Behavior and Human Decision Processes 1991 Dec 1;50(2):179-211.
32. DiClemente CC, Prochaska O. Toward a comprehensive, transtheoretical model of change: stages of change and addictive behaviors. In: Treating Addictive Behaviors (2nd Edition). Berlin/Heidelberg, Germany: Springer; 1998:3-27.
33. Gardner B, Lally P, Wardle J. Making health habitual: the psychology of 'habit-formation' and general practice. Br J Gen Pract 2012 Dec;62(605):664-666 [FREE Full text] [doi: 10.3399/bjgp12X659466] [Medline: 23211256]
34. Oinas-Kukkonen H, Harjumaa M. Persuasive systems design: key issues, process model, and system features. CAIS 2009;24(605):664-666 [FREE Full text] [doi: 10.17705/1cais.02428]
35. Lally P, Gardner B. Promoting habit formation. Health Psychology Review 2013 May;7(sup1):S137-S158. [doi: 10.1080/17437199.2011.603640]
36. Crandall A, Cheung A, Young A, Hooper AP. Theory-based predictors of mindfulness meditation mobile app usage: a survey and cohort study. JMIR Mhealth Uhealth 2019 Mar 22;7(3):e10794-e1S158 [FREE Full text] [doi: 10.2196/10794] [Medline: 30900992]
37. Hesse M. The Readiness Ruler as a measure of readiness to change poly-drug use in drug abusers. Harm Reduct J 2006 Jan 25;3(3):3 [FREE Full text] [doi: 10.1186/1477-7517-3-3] [Medline: 16436208]
38. Bowen M, Beam M. Mapping mindfulness: assessing the stages of meditation habit formation in the USA using the Sussex Mindfulness Meditation (SuMMed) Model. Mindfulness 2025 Apr 07;16(5):1340-1351. [doi: 10.1007/s12671-025-02555-2]
39. Miles E, Matcham F, Strauss C, Cavanagh K. Making mindfulness meditation a healthy habit. Mindfulness 2023 Nov 28;14(12):2988-3005. [doi: 10.1007/s12671-023-02258-6]
40. Lam SU, Xie Q, Goldberg SB. Situating meditation apps within the ecosystem of meditation practice: population-based survey study. JMIR Ment Health 2023 Apr 28;10:e43565 [FREE Full text] [doi: 10.2196/43565] [Medline: 37115618]
41. Davies JN, Faschinger A, Galante J, Van Dam NT. Prevalence and 20-year trends in meditation, yoga, guided imagery and progressive relaxation use among US adults from 2002 to 2022. Sci Rep 2024 Jul 01;14(1):14987 [FREE Full text] [doi: 10.1038/s41598-024-64562-y] [Medline: 38951149]
42. Jakob R, Harperink S, Rudolf AM, Fleisch E, Haug S, Mair JL, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371 [FREE Full text] [doi: 10.2196/35371] [Medline: 35612886]
43. Osin EN, Turilina II. Mindfulness meditation experiences of novice practitioners in an online intervention: trajectories, predictors, and challenges. Appl Psychol Health Well Being 2022 Feb 15;14(1):101-121. [doi: 10.1111/aphw.12293] [Medline: 34268871]
44. Kim S, Park JY, Chung K. The relationship between the big five personality traits and the theory of planned behavior in using mindfulness mobile apps: cross-sectional survey. J Med Internet Res 2022 Nov 30;24(11):e39501 [FREE Full text] [doi: 10.2196/39501] [Medline: 36449344]
45. Canby NK, Eichel K, Peters SI, Rahrig H, Britton WB. Predictors of out-of-class mindfulness practice adherence during and after a mindfulness-based intervention. Psychosom Med 2021 Oct 8;83(6):655-664 [FREE Full text] [doi: 10.1097/PSY.0000000000000873] [Medline: 33038188]
46. Stojanovic M, Fries S, Grund A. Self-efficacy in habit building: how general and habit-specific self-efficacy influence behavioral automatization and motivational interference. Front Psychol 2021 Aug;12(8):643753-643899. [doi: 10.3389/fpsyg.2021.643753] [Medline: 34025512]

XSL•FO
RenderX

47. Schiwal AT, Fauth EB, Wengreen H, Norton M. The gray matters app targeting health behaviors associated with Alzheimer's risk: improvements in intrinsic motivation and impact on diet quality and physical activity. J Nutr Health Aging 2020 Dec 22;24(8):893-899. [doi: 10.1007/s12603-020-1421-5] [Medline: 33009542]
48. Phillips W, Hine D. Self-compassion, physical health, and health behaviour: a meta-analysis. Health Psychol Rev 2021 Mar 22;15(1):113-139. [doi: 10.1080/17437199.2019.1705872] [Medline: 31842689]
49. Jones F, Harris P, Waller H, Coggins A. Adherence to an exercise prescription scheme: the role of expectations, self-efficacy, stage of change and psychological well-being. Br J Health Psychol 2005 Sep;10(Pt 3):359-378. [doi: 10.1348/135910704X24798] [Medline: 16238853]
50. Laurie J, Blandford A. Making time for mindfulness. Int J Med Inform 2016 Dec 01;96(4):38-50. [doi: 10.1016/j.ijmedinf.2016.02.010] [Medline: 26965526]
51. Banerjee A, Banerji R, Berry J. From proof of concept to scalable policies: challenges and solutions, with an application. Journal of Economic Perspectives 2017;31(4):73-102.
52. Prince M, Patel V, Saxena S, Maj M, Maselko J, Phillips MR, et al. No health without mental health. Lancet 2007 Sep 08;370(9590):859-877. [doi: 10.1016/S0140-6736(07)61238-0] [Medline: 17804063]
53. Baumel A, Kane JM. Examining predictors of real-world user engagement with self-guided eHealth interventions: analysis of mobile apps and websites using a novel dataset. J Med Internet Res 2018 Dec 14;20(12):e11491 [FREE Full text] [doi: 10.2196/11491] [Medline: 30552077]
54. Alqahtani F, Al Khalifah G, Oyebode O, Orji R. Apps for mental health: an evaluation of behavior change strategies and recommendations for future development. Front Artif Intell 2019 Dec 17;2:30 [FREE Full text] [doi: 10.3389/frai.2019.00030] [Medline: 33733119]
55. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387. [doi: 10.2196/24387] [Medline: 33759801]
56. Van Dam NT, van Vugt MK, Vago DR. . Mind the Hype: A Critical Evaluation and Prescriptive Agenda for Research on Mindfulness and Meditation. Perspectives on Psychological Science. 2018. URL: https://doi.org/10.1177/1745691617709589 [accessed 2025-10-24]
57. Brazier J, Peasgood T, Mukuria C, Marten O, Kreimeier S, Luo N, et al. The EQ-HWB: Overview of the Development of a Measure of Health and Wellbeing and Key Results. Value Health 2022 Apr;25(4):482-491 [FREE Full text] [doi: 10.1016/j.jval.2022.01.009] [Medline: 35277337]
58. Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SLT, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002 Aug;32(6):959-976 [FREE Full text] [doi: 10.1017/s0033291702006074] [Medline: 12214795]
59. Tennant R, Hiller L, Fishwick R, Platt S, Joseph S, Weich S, et al. The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. Health Qual Life Outcomes 2007 Nov 27;5(6):63-76 [FREE Full text] [doi: 10.1186/1477-7525-5-63] [Medline: 18042300]
60. Cheung F, Lucas RE. Assessing the validity of single-item life satisfaction measures: results from three large samples. Qual Life Res 2014 Dec 27;23(10):2809-2818 [FREE Full text] [doi: 10.1007/s11136-014-0726-4] [Medline: 24890827]
61. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
62. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr 22;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
63. Raes F, Pommier E, Neff KD, Van Gucht D. Construction and factorial validation of a short form of the Self-Compassion Scale. Clin Psychol Psychother 2011 Apr 08;18(3):250-255. [doi: 10.1002/cpp.702] [Medline: 21584907]
64. Romppel M, Herrmann-Lingen C, Wachter R, Edelmann F, Düngen HD, Pieske B, et al. A short form of the General Self-Efficacy Scale (GSE-6): Development, psychometric properties and validity in an intercultural non-clinical sample and a sample of patients at risk for heart failure. Psychosoc Med 2013;10(3):Doc01-Doc05 [FREE Full text] [doi: 10.3205/psm000091] [Medline: 23429426]
65. Dixon JB, Laurie CP, Anderson ML, Hayden MJ, Dixon ME, O'Brien PE. Motivation, readiness to change, and weight loss following adjustable gastric band surgery. Obesity (Silver Spring) 2009 Apr;17(4):698-705 [FREE Full text] [doi: 10.1038/oby.2008.609] [Medline: 19148126]
66. Eshah NF. Readiness for Behavior Change in Patients Living With Ischemic Heart Disease. J Nurs Res 2019 Dec;27(6):e57-705 [FREE Full text] [doi: 10.1097/jnr.0000000000000336] [Medline: 31283634]
67. Finsrud I, Nissen-Lie HA, Vrabel K, Høstmælingen A, Wampold BE, Ulvenes PG. It's the therapist and the treatment: The structure of common therapeutic relationship factors. Psychother Res 2022 Feb;32(2):139-150 [FREE Full text] [doi: 10.1080/10503307.2021.1916640] [Medline: 33938407]
68. Soto CJ, John OP. The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. J Pers Soc Psychol 2017 Jul 02;113(1):117-143 [FREE Full text] [doi: 10.1037/pspp0000096] [Medline: 27055049]

XSL•FO
RenderX

69. Stoyanov SR, Hides L, Kavanagh DJ, Wilson H. Development and Validation of the User Version of the Mobile Application Rating Scale (uMARS). JMIR Mhealth Uhealth 2016 Jun 10;4(2):e72-143 [FREE Full text] [doi: 10.2196/mhealth.5849] [Medline: 27287964]
70. Yohai V. High Breakdown-Point and High Efficiency Robust Estimates for Regression. Ann. Statist 1987 Jun 1;15(2):e72 [FREE Full text] [doi: 10.1214/aos/1176350366] [Medline: 27287964]
71. Koller M, Stahel W. Sharpening Wald-type inference in robust regression for small samples. Computational Statistics & Data Analysis 2011 Aug 1;55(8):2504-2515. [doi: 10.1016/j.csda.2011.02.014]
72. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) 1995 Jan 01;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]
73. Yentes R, Wilhelm F. careless: Procedures for Computing Indices of Careless Responding. Package ‘careless’ 2021 [FREE Full text]
74. Kunst AE, Bos V, Lahelma E, Bartley M, Lissau I, Regidor E, et al. Trends in socioeconomic inequalities in self-assessed health in 10 European countries. Int J Epidemiol 2005 Apr;34(2):295-305 [FREE Full text] [doi: 10.1093/ije/dyh342] [Medline: 15563586]
75. Mackenbach JP, Stirbu I, Roskam AJR, Schaap MM, Menvielle G, Leinsalu M, European Union Working Group on Socioeconomic Inequalities in Health. Socioeconomic inequalities in health in 22 European countries. N Engl J Med 2008 Jun 05;358(23):2468-2481 [FREE Full text] [doi: 10.1056/NEJMsa0707519] [Medline: 18525043]
76. Barber MN, Staples M, Osborne RH, Clerehan R, Elder C, Buchbinder R. Up to a quarter of the Australian population may have suboptimal health literacy depending upon the measurement tool: results from a population-based survey. Health Promot Int 2009 Sep 05;24(3):252-261. [doi: 10.1093/heapro/dap022] [Medline: 19531559]
77. van der Heide I, Wang J, Droomers M, Spreeuwenberg P, Rademakers J, Uiters E. The relationship between health, education, and health literacy: results from the Dutch Adult Literacy and Life Skills Survey. J Health Commun 2013 Sep 13;18 Suppl 1(Suppl 1):172-184 [FREE Full text] [doi: 10.1080/10810730.2013.825668] [Medline: 24093354]
78. Sevilla A, Gimenez-Nadal JI, Gershuny J. Leisure inequality in the United States: 1965-2003. Demography 2012 Aug;49(3):939-964 [FREE Full text] [doi: 10.1007/s13524-012-0100-5] [Medline: 22589003]
79. Winter N, Russell L, Ugalde A, White V, Livingston P. Engagement Strategies to Improve Adherence and Retention in Web-Based Mindfulness Programs: Systematic Review. J Med Internet Res 2022 Jan 12;24(1):e30026-e30064 [FREE Full text] [doi: 10.2196/30026] [Medline: 35019851]
80. Burke A, Lam CN, Stussman B, Yang H. Prevalence and patterns of use of mantra, mindfulness and spiritual meditation among adults in the United States. BMC Complement Altern Med 2017 Jun 15;17(1):316 [FREE Full text] [doi: 10.1186/s12906-017-1827-8] [Medline: 28619092]
81. Goldberg LR. The structure of phenotypic personality traits. Am Psychol 1993 Jan 15;48(1):26-34. [doi: 10.1037//0003-066x.48.1.26] [Medline: 8427480]
82. Khwaja M, Pieritz S, Faisal AA, Matic A. Personality and Engagement with Digital Mental Health Interventions. In: Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization. 1993 Jan Presented at: UMAP Confefence on User Modelling, Adaptation and Personalization; June 21-25, 2021; Utrecht, Netherlands p. 235-239. [doi: https://doi.org/10.1145/3450613.3456823]
83. Alqahtani F, Meier S, Orji R. Personality-based approach for tailoring persuasive mental health applications. User Model User-Adap Inter 2021 Jul 16;32(3):253-295 [FREE Full text] [doi: 10.1007/s11257-021-09289-5]
84. Whitford S, Warren K. Perceived Barriers to Meditation Among College Students. Building Healthy Academic Communities Journal 2019;32(3):23-33. [doi: 10.18061/bhac.v3i1.6678]
85. Delmonte MM. Personality correlates of meditation practice frequency and dropout in an outpatient population. J Behav Med 1988 Dec 29;11(6):593-597. [doi: 10.1007/BF00844908] [Medline: 3075240]
86. Van Dam N, Targett J, Davies J, Burger A, Galante J. Incidence and Predictors of Meditation-Related Unusual Experiences and Adverse Effects in a Representative Sample of Meditators in the United States. Clinical Psychological Science 2025 Jan 06;13(3):632-648. [doi: 10.1177/21677026241298269] [Medline: 3075240]
87. Jiwani Z, Lam SU, Richard JD, Goldberg SB. Motivation for Meditation and its Association with Meditation Practice in a National Sample of Internet Users. Mindfulness (N Y) 2022 Oct 06;13(10):2641-2651 [FREE Full text] [doi: 10.1007/s12671-022-01985-6] [Medline: 36506892]
88. Clarke J, Proudfoot J, Whitton A, Birch MR, Boyd M, Parker G, et al. Therapeutic Alliance With a Fully Automated Mobile Phone and Web-Based Intervention: Secondary Analysis of a Randomized Controlled Trial. JMIR Ment Health 2016 Feb 25;3(1):e10-2651 [FREE Full text] [doi: 10.2196/mental.4656] [Medline: 26917096]
89. Tremain H, McEnery C, Fletcher K, Murray G. The Therapeutic Alliance in Digital Mental Health Interventions for Serious Mental Illnesses: Narrative Review. JMIR Ment Health 2020 Aug 07;7(8):e17204 [FREE Full text] [doi: 10.2196/17204] [Medline: 32763881]
90. D'Alfonso S, Lederman R, Bucci S, Berry K. The Digital Therapeutic Alliance and Human-Computer Interaction. JMIR Ment Health 2020 Dec 29;7(12):e21895 [FREE Full text] [doi: 10.2196/21895] [Medline: 33372897]

XSL•FO
RenderX

91. Armitage CJ, Norman P, Alganem S, Conner M. Expectations are more predictive of behavior than behavioral intentions: evidence from two prospective studies. Ann Behav Med 2015 Apr 29;49(2):239-246 [FREE Full text] [doi: 10.1007/s12160-014-9653-4] [Medline: 25623893]
92. Armitage CJ, Norman P, Alganem S, Conner M. Expectations are more predictive of behavior than behavioral intentions: evidence from two prospective studies. Ann Behav Med 2015 Apr 26;49(2):239-246 [FREE Full text] [doi: 10.1007/s12160-014-9653-4] [Medline: 25623893]
93. Bhattacherjee A. Understanding Information Systems Continuance: An Expectation-Confirmation Model. MIS Quarterly 2001 Sep 26;25(3):351-617. [doi: 10.2307/3250921]
94. Zimmerman GL, Olsen CG, Bosworth MF. A 'stages of change' approach to helping patients change behavior. Am Fam Physician 2000 Mar 01;61(5):1409-1416 [FREE Full text] [Medline: 10735346]
95. Zimmerman GL, Olsen CG, Bosworth MF. A 'stages of change' approach to helping patients change behavior. Am Fam Physician 2000 Mar 01;61(5):1409-1416. [Medline: 10735346]
96. Moyers TB, Martin T, Houck JM, Christopher PJ, Tonigan JS. From in-session behaviors to drinking outcomes: a causal chain for motivational interviewing. J Consult Clin Psychol 2009 Dec;77(6):1113-1124 [FREE Full text] [doi: 10.1037/a0017189] [Medline: 19968387]
97. Singh A. Self-efficacy and well-being among students: role of goal meditation. International Journal of Indian Psychology 2019;7(2):405-414. [doi: 10.25215/0702.049] [Medline: 19968387]
98. Goldstein L, Nidich SI, Goodman R, Goodman D. The effect of transcendental meditation on self-efficacy, perceived stress, and quality of life in mothers in Uganda. Health Care Women Int 2018 Jul;39(7):734-754 [FREE Full text] [doi: 10.1080/07399332.2018.1445254] [Medline: 29494787]
99. Goldstein L, Nidich SI, Goodman R, Goodman D. The effect of transcendental meditation on self-efficacy, perceived stress, and quality of life in mothers in Uganda. Health Care Women Int 2018 Jul;39(7):734-754 [FREE Full text] [doi: 10.1080/07399332.2018.1445254] [Medline: 29494787]
100. Wells RE, Burch R, Paulsen RH, Wayne PM, Houle TT, Loder E. Meditation for migraines: a pilot randomized controlled trial. Headache 2014 Oct;54(9):1484-1495. [doi: 10.1111/head.12420] [Medline: 25041058]
101. Peer E, Rothschild D, Gordon A, Evernden Z, Damer E. Data quality of platforms and panels for online behavioral research. Behav Res Methods 2022 Aug;54(4):1643-1662 [FREE Full text] [doi: 10.3758/s13428-021-01694-3] [Medline: 34590289]
102. Adams J. User and app related factors associated with engagement with mindfulness apps and health and wellbeing: a cross-sectional survey of US mindfulness app users. Open Science Framework (OSF). 2025 Mar 21. URL: https://osf.io/jcv5n/overview [accessed 2025-11-07]

Abbreviations

BFI-S-2: Big Five Inventory Short Form 2
DWAI-6: 6-item Digital Working Alliance Inventory
EQ-HWB-9: 9-item EuroQoL Health and Wellbeing
FDR: false discovery rate
K10: Kessler Psychological Distress Scale
MBP: mindfulness-based program
reCAPTCHA: reverse Completely Automated Public Turing Test to Tell Computers and Humans Apart
SWLS: Satisfaction with Life Survey
uMARS: user Mobile Application Rating Scale
WEMWBS: Warwick-Edinburgh Mental Wellbeing Scale

Edited by A Mavragani, T de Azevedo Cardoso; submitted 30.Jan.2025; peer-reviewed by SU Lam, G Cain; comments to author 03.Mar.2025; revised version received 29.Mar.2025; accepted 20.Aug.2025; published 02.Feb.2026.

Please cite as:
Adams J, Davies J, Wattanatakulchat P, Galante J, Miller F, D'Alfonso S, Van Dam NT
Engagement With Meditation Apps: Cross-Sectional Survey of Use and Associations
J Med Internet Res 2026;28:e71960
URL: https://www.jmir.org/2026/1/e71960
doi:10.2196/71960
PMID:41627883

©Julia Adams, Jonathan Davies, Prai Wattanatakulchat, Julieta Galante, Felicity Miller, Simon D'Alfonso, Nicholas T Van Dam. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.Feb.2026. This is an open-access

XSL•FO
RenderX

article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Stepped Care Approach to Promoting Engagement in a Digital Health Intervention: Secondary Analysis of a Randomized Controlled Trial

Esha Dwibedi[1], PhD; Wen You[2], PhD; Lee M Ritterband[3], PhD; Donna-Jean P Brock[1], MS; Annie L Reid[1], MPH, RDN; Christina Frederick[3], BS; Jamie M Zoellner[1], RDN, PhD

[1]Department of Public Health Sciences, School of Medicine, University of Virginia, Christiansburg, VA, United States

[2]Department of Public Health Sciences, School of Medicine, University of Virginia, Charlottesville, VA, United States

[3]Department of Psychiatry and Neurobehavioral Sciences, School of Medicine, University of Virginia, Charlottesville, VA, United States

Corresponding Author:
Jamie M Zoellner, RDN, PhD
Department of Public Health Sciences
School of Medicine
University of Virginia
16 East Main Street
Christiansburg, VA, 24073
United States
Phone: 1 4349624488
Email: Jz9q@virginia.edu

Abstract

Background: Digital health interventions (DHIs) often struggle with participant engagement. A stepped care approach, starting with low-resource intensity strategies and escalating as needed, can optimize resource use. Yet its application and cost implications remain underexplored.

Objective: This study uses data from the iSIPsmarter experimental arm of a 2-group randomized controlled trial targeting sugar-sweetened beverage consumption in rural Appalachia. This study examines the demand and implementation costs associated with iSIPsmarter's stepped care engagement approach and simulates how variations in monitoring efficiency, demand, and stepped care intensity influence resource use and implementation costs to inform future implementation.

Methods: iSIPsmarter's stepped care process combined automated and human-supported components to enhance engagement across 6 web-based modules ("Cores") over 9 weeks. Participants who did not complete a Core received an automated email, followed by stepped care if still incomplete: a text (step 1, low-resource intensity) after 7 days and up to 3 telephone attempts (step 2, high-resource intensity) after another 7 days. Staff time was tracked to estimate implementation costs: monitoring averaged 3 minutes (US $1.68), texts 2.83 minutes (US $1.58), and calls 5.1 minutes (US $2.85). Simulations explored 18 scenarios varying monitoring efficiency (20%, 50%, and 80% of trial-observed monitoring time and costs), stepped care demand (20%, 50%, and 80% of participants needing stepped care), and intervention intensity (low vs high).

Results: Among 126 participants, the mean stepped care contact was 1.2 (SD 1.3): 52 (41%) required none, 42 (33%) required 1 Core contact, 26 (21%) required 2, and 7 (6%) required 3. On average, participants completed 5.2 (SD 1.6) of 6 Cores. The mean stepped care implementation time per participant was 26.46 (SD 11.02) minutes, with a corresponding mean cost of US $14.80 (SD 6.16). Monitoring accounted for 78% of total cost (mean cost US $11.61, SD 2.37), with initial monitoring contributing 58% of total cost (mean cost US $8.51, SD 2.35). Simulations showed variation in time and cost based on monitoring efficiency. In low-demand, low-intensity scenarios, efficient monitoring required mean of 7.47 (95% CI 7.36-7.57) minutes and mean cost of US $4.18 (95% CI 4.12-4.24), while inefficient monitoring required a mean of 19.58 (95% CI 19.21-19.95) minutes and mean cost of US $10.95 (95% CI 10.74-11.16). In high-demand, high-intensity scenarios, efficient monitoring required a mean of 101.80 (95% CI 101.65-101.96) minutes and mean cost of US $56.92 (95% CI 56.84-57.01), while inefficient monitoring increased time to a mean of 146.32 (95% CI 145.92-146.71) minutes and mean cost of US $81.82 (95% CI 81.60-82.04).

Conclusions: A stepped care approach can efficiently sustain engagement in DHIs by targeting support to higher-need participants. These findings offer actionable guidance for designing scalable, cost-effective interventions for real-world settings, as resource-efficient engagement strategies remain a persistent challenge for DHIs.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e80921) doi:10.2196/80921

KEYWORDS

digital health interventions; engagement; stepped care; implementation cost; simulations

Introduction

Digital health interventions (DHIs) offer a promising approach to support patient self-management and promote health behavior change [1,2] while addressing barriers related to availability, accessibility, geography, and time constraints [3-5]. Despite this potential, systematic reviews report mixed evidence on DHI effectiveness, with limited user engagement identified as a key barrier [6,7]. Because engagement is closely linked to intervention effectiveness [8-10], low and inconsistent engagement poses a major barrier to the scalability and sustainability of DHIs, especially in underserved rural populations where resources are constrained and few evidence-based strategies exist [11,12].

Human-supported strategies, such as personalized support, tailored feedback, and reminders, are consistently linked to better DHI engagement [12,13]. Yet, these strategies are often resource-intensive and costly, posing major challenges for scalability and implementation [12]. Consequently, there is a critical need to identify and evaluate scalable, resource-efficient strategies that can strengthen user engagement in DHIs.

A stepped care framework may offer a potential solution to this challenge. Stepped care models are designed to maximize resource efficiency by initiating treatment with low-intensity interventions and escalating to high-intensity options only as necessary [14-17]. Traditionally used in behavioral and mental health contexts [18,19], stepped care has been shown to improve access [20], reduce treatment duration [21,22], and lower costs [23,24]. In dietary interventions, stepped care has been applied in contexts such as eating disorders [25] and weight management [26]. However, its application specifically as an engagement strategy within DHIs remains unexplored.

To evaluate the feasibility and potential scalability of a stepped care engagement approach within DHIs, it is essential to understand its demand, impact on engagement, and implementation costs. However, current research offers limited evidence on these Core indicators, leaving a critical gap in how engagement strategies are operationalized within digital platforms. In particular, assessing stepped care implementation costs from the provider's perspective is particularly important when the long-term goal is to sustain implementation of DHIs within health care systems. Simulation methods enable systematic examination of how realistic variations in health care delivery (eg, shifts in demand for stepped care, monitoring efficiency, and stepped care intensity) influence total implementation costs and resource allocation [27]. By linking the engagement process with economic and operational dynamics, this framework provides a scalable, data-driven approach for optimizing DHIs and guiding implementation strategies in real-world settings.

This study presents secondary aims and analyses of iSIPsmarter, the experimental arm of a previously conducted randomized controlled trial (RCT). iSIPsmarter integrated a stepped care engagement framework, which is the focus of the present analysis. iSIPsmarter is a DHI aimed at reducing sugar-sweetened beverage (SSB) consumption among rural adults in Appalachia [28]. In an RCT, iSIPsmarter demonstrated effectiveness in reducing SSB intake following a 9-week intervention period (effect size=0.37; P=.005) and 6 months later (effect size=0.35; P=.009), as well as in improving percent weight loss at the 6-month assessment (effect size=0.23; P=.046), compared to a static patient education control group [29]. Engagement with the iSIPsmarter program was high, with most participants completing the majority of the educational modules ("Cores") [29]. This high engagement, in part, might be attributed to the human-supported stepped care component designed to reengage disengaged participants. However, the demand and implementation cost of iSIPsmarter's stepped care approach have not yet been examined and are the focus of this study. Furthermore, because these estimates are specific to the research context of the RCT, simulating how variations in demand and monitoring efficiency might affect implementation costs in real-world health care systems represents an essential step toward advancing iSIPsmarter along the translation pipeline.

Thus, the objectives of this study are 2-fold. First, this process evaluation addresses the trial's a priori secondary aims by examining demand and implementation costs associated with a stepped care engagement approach [29,30]. Second, simulation analyses were conducted to explore additional research questions that were not an explicit focus of the original study. Specifically, the study simulates how variations in monitoring costs, demand, and stepped care intensity influence overall resource use and implementation costs to inform future implementation efforts.

Methods

Study Design

This study uses process data from a 2-group RCT where participants were randomized into either the iSIPsmarter experimental condition or the Patient Education website control, with both groups receiving content focused on reducing SSB [28]. Details on the effectiveness of the RCT and CONSORT (Consolidated Standards of Reporting Trials) diagrams can be found in previously published work [29]. The control group did not receive stepped care. Therefore, this study focuses solely on the stepped care data from the iSIPsmarter experimental condition, assessing both the demand for stepped care and the associated implementation cost. All costs reported are in US dollars (US $).

Study Sample

The broader 2-group RCT recruited participants from Southwest Virginia and neighboring Appalachian counties. Eligibility criteria included being 18 years or older, English-speaking, and reporting the consumption of more than 200 calories per day from SSBs. A total of 249 participants consented, enrolled, and

XSL•FO
RenderX

were randomized with 127 assigned to the iSIPsmarter intervention.

iSIPsmarter Intervention Description

The iSIPsmarter intervention is based on the theory of planned behavior, incorporating behavior change strategies and health literacy techniques [28]. The development process of the iSIPsmarter intervention was further guided by the Internet Intervention Model [31], which emphasizes the role of user characteristics, environment, intervention content, level of support, and targeted outcomes in shaping behavior change [32], as well as practices from human-centered design and instructional design [32]. Additional details on the development process can be found in the iSIPsmarter development manuscript [32].

The iSIPsmarter intervention includes 6 interactive Cores delivering behavioral change and educational content through integrated text, audio, graphics, animation, and video. Each Core is enhanced by vignettes depicting relatable storylines and situations focused on goal setting, behavior change, overcoming barriers, and managing relapses. iSIPsmarter also included behavioral diary tracking and personalized action planning [28]. The only nondigital component is a human-supported stepped care strategy to encourage Core completion. As further described below, all iSIPsmarter participants have access to stepped care, although not all require it.

Stepped Care Theoretical Rationale and Description

iSIPsmarter's stepped care approach was informed by the Internet Intervention Model [31], specifically the domain addressing level of support, and by the Supportive Accountability Model, which posits that individuals are more likely to adhere to a DHI when they feel accountable to a supportive, trustworthy person who provides oversight and encouragement [12,13,33]. It was designed to promote reengagement among disengaged participants and to enhance behavioral engagement [34,35], specifically Core (module) completion. Unlike traditional stepped care approaches, which typically escalate intervention intensity through personalized coaching targeting behavioral outcomes [18-22,25] (eg, reductions in SSB intake), the iSIPsmarter stepped care approach focused on providing supportive accountability to encourage Core completion and address DHI disengagement [12,13]. This novel design may offer efficiency and scalability if future implementation efforts, particularly in settings where nutrition and weight management professionals are limited, such as low-resourced health care and community-based settings.

iSIPsmarter's stepped care process is illustrated in Figure 1. During the 9-week intervention period, each Core was unlocked 7 days after the prior Core was completed. Participants who did not complete a Core within 3 days of its availability received an automated email reminder. One week after the Core became available, a study coordinator assessed completion status, triggering an initial monitoring cost. If the Core remained incomplete, the research coordinator evaluated the need for human-supported stepped care and initiated step 1, which involved sending a text reminder. This step generated both a text cost and a monitoring cost to track engagement. If the Core was still incomplete after an additional 7 days, step 2 was initiated: a phone call was made (up to 3 attempts to reach the participant with voicemail if necessary), incurring a phone call cost. These calls provided encouragement, technical assistance, and strategies to support task completion. Subsequent follow-up generated additional monitoring and phone costs.

Participants who did not complete the Core within 7 days after the last phone call attempt were classified as nonadherent. Nonadherent participants included those reached by phone but who did not complete the Core as well as those who were never reached (despite leaving voicemails). Participants who spoke with study coordinators once did not receive additional calls for the same Core. These individuals continued receiving automated emails every 2 weeks but no further human-supported stepped care. However, if they completed the Core at any point during the 9-week intervention period, they regained adherence status, and the next Core was unlocked after 7 days. Participants who completed a Core after the 9-week intervention period were still able to access subsequent Cores, but their engagement data were no longer tracked. Throughout the intervention, nonadherent participants were monitored every 2 weeks, contributing to nonadherence monitoring costs. As participants were required to complete each Core before advancing to the next, some were unable to complete all Cores within the 9 weeks. The monitoring and engagement steps are detailed in Figure 1 and the time spent by study coordinators and costs incurred are defined in the subsection below.

XSL•FO
RenderX

Figure 1. Stepped care approach to promote Core (module) completion in the digital iSIPsmarter intervention for Appalachian adults.

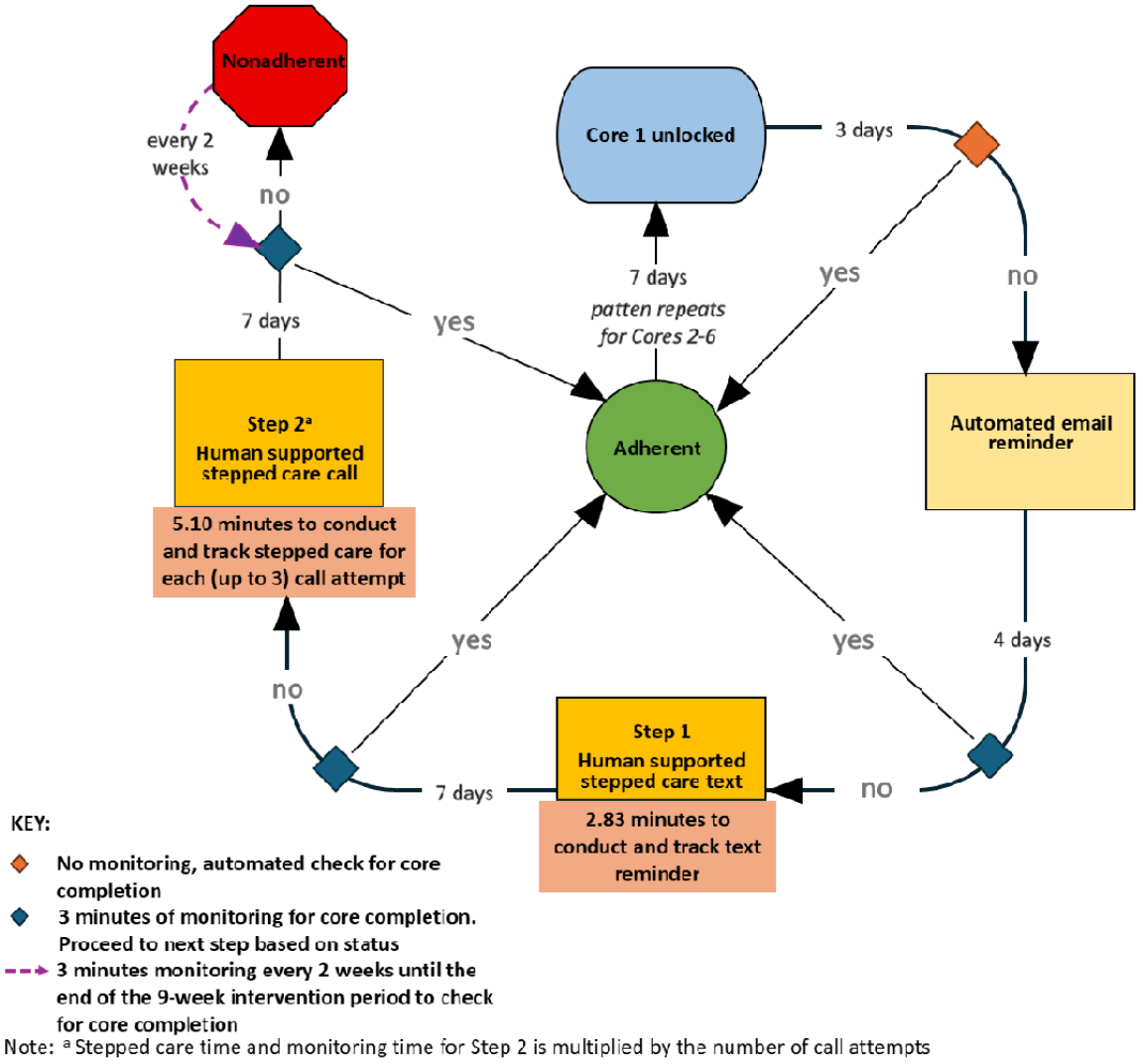

Definition of Time Spent and Costs Incurred in the Stepped Care Engagement Process

Each level the of stepped care engagement process incurs specific costs, primarily reflecting staff time required for monitoring and outreach. These costs and corresponding times are defined as follows:

1. Initial monitoring cost (CIM) or time (TIM): The cost or time of assessing Core completion 1 week after being available, following the automated email reminder.
2. Text cost ( ) or time ( ): The cost or time of sending a text reminder (step 1) to noncompleters.
3. Monitoring text cost ( ) or time ( ): The cost or time of monitoring participant engagement after the text reminder.
4. Phone call cost ( ) or time ( ): The cost or time of making up to 3 phone call attempts (step 2), including voicemails and participant support.
5. Monitoring phone cost ( ) or time ( ): The cost or time of monitoring engagement following phone call attempts to determine the need for additional follow-up.
6. Nonadherence monitoring cost (CNAM) or time (TNAM): The cost or time of biweekly monitoring participants classified as nonadherent until the end of the intervention.

Data Analysis

Stepped Care Engagement Demand

Engagement was measured by whether participants completed a Core, including following 1 of the 2 stepped care engagement strategies: the text reminder (step 1), or phone calls (step 2). Demand for stepped care was evaluated based on the extent to which the stepped care process successfully reengaged participants in the iSIPsmarter experimental condition. This approach provided a way to evaluate the success of each stepped care component in sustaining participant involvement and promoting Core completion. In addition to examining overall Core completion and aggregate stepped care demand, participant-level stepped care demand was tracked longitudinally to explore individual variation in support needs.

Stepped Care Implementation Time and Costs

Recent work on cost assessment for behavioral interventions highlights the importance of tracking staff time as a key driver of implementation costs [36-39]. Staff time, a major and variable

XSL•FO
RenderX

cost, particularly for training and support, is often harder to estimate than material costs. Time-driven activity-based costing (TDABC) offers a practical assessment method [39-41]. In this study, implementation costs were assessed as the cost of delivering the stepped care engagement strategies in the iSIPsmarter experimental condition, using TDABC's uniform estimate of time for tasks self-report approach [42].

For each participant, the average time and cost of implementing the stepped care process for each participant were calculated based on study coordinator-reported time data. Each of the 5 study coordinators (DJB, TM, AR, BK, and HW) provided time estimates for each step of the process across 3 separate instances. The average of these self-reported time estimates was included in the analysis with the monitoring time (defined as checking for Core completion) at 3 minutes, text time at 2.83 minutes, and phone call time at 5.1 minutes. The labor costs of stepped care were valued at the market wage rate for the associated occupation by taking the 2023 national mean wage for a "Health Education Specialist" reported as US $33.55/hour in the Occupational Employment and Wage Statistics (OEWS) data from the US Bureau of Labor Statistics [43]. The time estimates in minutes were converted to hours and multiplied with the mean wage to get a cost estimate of the time spent on each step of the stepped care process. For each of the 6 Cores, the per-participant cost of implementing the stepped care process was estimated based on the defined cost components. Initial monitoring (US $1.68) covered checking the Core completion after the automated email. Text (US $1.58) covered typing and sending the text and monitoring text (US $1.68) accounted for tracking completion after the text reminder (step 1). Phone (US $2.85) covered up to 3 call attempts, while monitoring phone (US $1.68) tracked completion after each attempt. Nonadherence monitoring (US $1.68) reflected biweekly checks for participants who remained nonadherent.

Simulation Models

The iSIPsmarter trial was conducted with a relatively small and homogeneous sample, and the observed demand and engagement patterns are specific to a controlled research context. To explore how these dynamics might vary in applied settings, we used deterministic scenario simulations [44] to assess how anticipated variations in demand, monitoring costs, and stepped care intensity influence overall resource use and implementation costs.

These simulations are not intended to test the robustness of a single empirical estimate, as would be the case in a traditional sensitivity analysis. Instead, they operationalize a scenario-driven framework that links engagement dynamics with implementation costs to anticipate how a stepped care engagement strategy would perform under diverse real-world conditions, which is the information that cannot be obtained from the trial alone. Specifically, the simulation parameters reflect resource allocation frameworks, highlighting the need to balance intervention effectiveness with cost efficiency [44]. A total of 18 distinct scenarios were modeled to explore these variations. Because the 6 Cores were unlocked sequentially based on prior completion, the simulations use longitudinal nested data draws to maintain the logical flow of Core access throughout the intervention.

Monitoring costs were systematically varied by reducing them to 20%, 50%, and 80% of the trial-observed per-participant stepped care implementation costs related to monitoring, reflecting decreasing monitoring efficiency encompassing all monitoring-related components: initial monitoring, text reminders, phone outreach, and nonadherence monitoring. These levels illustrate a plausible range of monitoring intensity reductions likely in real-world implementation.

Additionally, demand for stepped care was also varied at 3 levels (ie, 20%, 50%, and 80% of participants) to represent different degrees of need for human-supported engagement during the intervention. These scenarios were selected to provide a wide range of plausible conditions, informed by the observed heterogeneity in engagement with iSIPsmarter [45] as well as prior literature documenting variability in participant adherence and engagement with DHIs [46,47].

The intensity of stepped care was categorized into two levels: (1) low intensity: participants required only the initial monitoring step of automated email and text reminders to reengage (step 1); and (2) high intensity: participants required the full stepped care engagement process, including up to 3 phone calls (step 2) to reengage them. This reflected the study design and illustrated cost differences between lower- and higher-intensity support.

The total simulated cost $C_{total,i}$ for each scenario was modeled as a function of the monitoring costs, the demand for stepped care, and the intensity of the stepped care process, as shown in the equations below.

where i represent individual participant; $\chi \in \{0.2,0.5,0.8\}$ is the monitoring efficiency multiplier, applied to all monitoring-related costs; is the cost of initial monitoring; $D_i \in \{0,1\}$ indicates whether participant i receives stepped care (1 if required, 0 otherwise), with stepped care requirements governed by the demand scenario parameter $\delta \in \{0.2,0.5,0.8\}$, which represent the proportion of participants receiving stepped care in each simulation scenario; $I_{low} \in \{0,1\}$ is an indicator equal to 1 if the participant receives only step 1 (ie, text-based engagement); $I_{high} \in \{0,1\}$is an indicator equal to 1 if the participant receives step 1 and step 2 (ie, phone-based engagement); C_{low} is the cost of participants receiving only step 1; C_{high} is the cost of participants receiving step 1 and step 2; is the cost of tracking and sending text reminders; is the cost of conducting phone call attempts and/or voicemail; is the cost of monitoring response to text reminders; is the cost of monitoring response to phone calls; and is the cost of monitoring for nonadherence after step 2. The indicators $I_{low,i}$

XSL•FO
RenderX

and $I_{high,i}$ are mutually exclusive (ie, $I_{low,i} + I_{high,i} = 1$) for participants who receive stepped care, reflecting that only one intensity level applies per participant.

By varying the demand for care and the intensity of the intervention, the model simulates the cost impact under different conditions of monitoring cost efficiency, participant engagement, and resource use. This simulation approach provides insights into how different components of the stepped care process contribute to the overall costs and resource needs, guiding optimization strategies for the intervention.

Anticipated Simulation Patterns

We anticipated the following patterns to emerge from the simulations:

1. Among scenarios with the same percentage of stepped care demand, those with higher intensity engagement (eg, requiring phone calls) would result in higher overall implementation costs compared to those with lower intensity engagement (eg, text reminders only).
2. Among scenarios with the same stepped care demand and intensity level, those with more efficient monitoring (ie, lower monitoring cost) would result in lower overall implementation costs than those with less efficient monitoring (higher monitoring cost).

Ethical Considerations

All study procedures were approved by the University of Virginia Institutional Review Board (#22130) and were registered prospectively at ClincialTrials.gov as NCT05030753 on August 26, 2021. There were no deviations from the trial registration. The study is described in accordance with the Template for Intervention Description and Replication (TIDieR) checklist (Multimedia Appendix 1) [48] to enhance transparency and reproducibility of reporting. We also follow the CONSORT-EHEALTH reporting guidelines for RCTs of eHealth interventions (Multimedia Appendix 2) [49]. Prior to enrollment in the RCT [28], participants provided written informed consent to participate in the study, including consent for the data presented in this paper. The paper and all supplementary materials do not include any photographs, screenshots, or other images that could permit identification of individual participants or users. Participants were allowed to withdraw from the study at any time. All data used in this analysis were deidentified to ensure participant privacy and confidentiality. Participants received compensation in the form of a cellular-enabled scale at baseline and gift cards (US $50) for completing each of the assessments.

Results

Demographics

As shown in the CONSORT diagram (Figure 2), 778 individuals were assessed for eligibility, of which 502 (65%) were eligible and 249 (50%) were enrolled and randomized. Among the 127 participants in the iSIPsmarter intervention, 1 participant withdrew from the study midway and was consequently dropped from subsequent analysis. Analyzed participants (N=126) were predominantly women (n=102, 81%) and White (non-Hispanic) (n=112, 89%), with an average age of 41.9 (SD 12.01) years. Most participants (n=112, 89%) had at least some college education. Half of the participants (n=62, 50%) reported an annual income of US $55,000 or more. The mean health literacy score was 11.57 out of 12 (SD 0.99). The characteristics of the participants are given in Table S1 in Multimedia Appendix 3.

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) diagram for a 2-arm randomized controlled trial of Appalachian adults, with analyses focused on stepped care within the iSIPsmarter arm. SSB: sugar-sweetened beverage; SWVA: southwest Virginia.

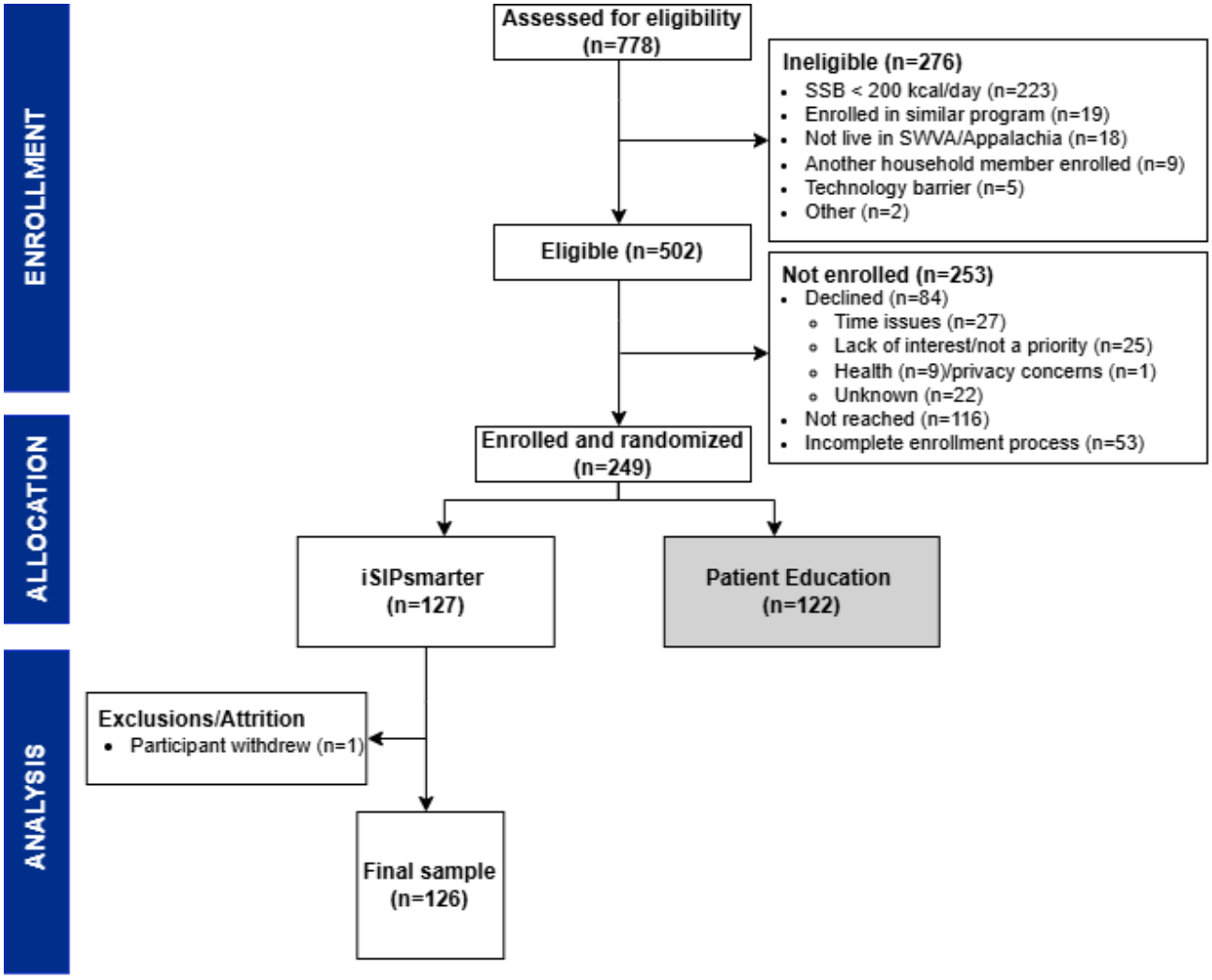

Stepped Care Engagement Demand

Aggregated stepped care contact information showed that 76 (60%) participants required at least 1 stepped care contact over the course of the intervention, with participants averaging 1.2 stepped care contacts (Table S2 in Multimedia Appendix 3). On average, iSIPsmarter participants completed 5.2 (SD 1.6) of 6 total Cores, with 74 (59%) participants completing all 6 Cores and an additional 20 (16%) participants completing 5 Cores.

Table 1 summarizes participant demand for components of the stepped care process for each Core and the subsequent Core completion rates after each step of the stepped care process. As shown, the overall demand for text reminders (step 1) was approximately twice that of phone calls (step 2), with totals of 148 (range across Cores: 14-34) and 73 (range across Cores: 7-20), respectively. Among those who required stepped care, low-intensity text reminders (step 1) were generally effective: in 75 (51%) instances, participants successfully completed the Core after receiving the text reminders (Table 1). In contrast, high-intensity phone outreach (step 2) had a lower success rate, with 12 (16%) instances of recipients completing the Core after receiving the first phone call or voicemail, 6 (10%) instances after the second and 3 (5%) instances after the third (Table 1), leading to an overall success rate of 29% for step 2.

Figure 3 illustrates the distribution of engagement with the stepped care levels for each Core. Across the 6 Cores, the majority of participants completed each Core without requiring stepped care, indicating that a substantial proportion of participants remained engaged with the iSIPsmarter intervention without requiring additional support. Step 1 and Step 2 showed varied success across the Cores, with Core completion after Step 1 and Step 2 peaking at Core 2. Core nonadherence after stepped care was low. Because participants could not proceed to the next Cores without completing the previous one, some did not finish all 6 Cores within the 9-week intervention period. By the end of the 9-week intervention period, 36 participants had timed out, including 20 who were still on Core 6.

XSL•FO
RenderX

Table 1. Demand and Core (module) completion success for each level of stepped care within the digital iSIPsmarter[a] intervention among Appalachian adults.

Core	Needed step 1 (text)	Completed Core after step 1 (text)	Needed Step 2 (phone call)	Completed Core after phone call 1	Completed Core after phone call 2	Completed Core after phone call 3
Core 1	14	7	7	4	0	0
Core 2	34	20	14	6	0	2
Core 3	30	17	13	0	6	1
Core 4	24	14	10	0	0	0
Core 5	22	13	9	2	0	0
Core 6	24	4	20	0	0	0
Total[b]	148	75	73	12	6	3
Impact[c]	—[d]	51	—	16	10	5

[a]iSIPsmarter is a digital health intervention aimed at reducing sugar-sweetened beverage consumption; data represent the initial 9-week intervention period.

[b]Raw count of participants who needed each level of stepped care or completed Cores after each type of stepped care contacts, across all Cores.

[c]Impact of stepped care contacts given by the percentage of participants who completed Cores after receiving stepped care contacts (for text and phone calls).

[d]Not applicable.

Figure 3. Aggregate participant completion rates for each Core (module) by stepped care levels in the digital iSIPsmarter intervention among Appalachian adults (data represent the initial 9-week intervention period).

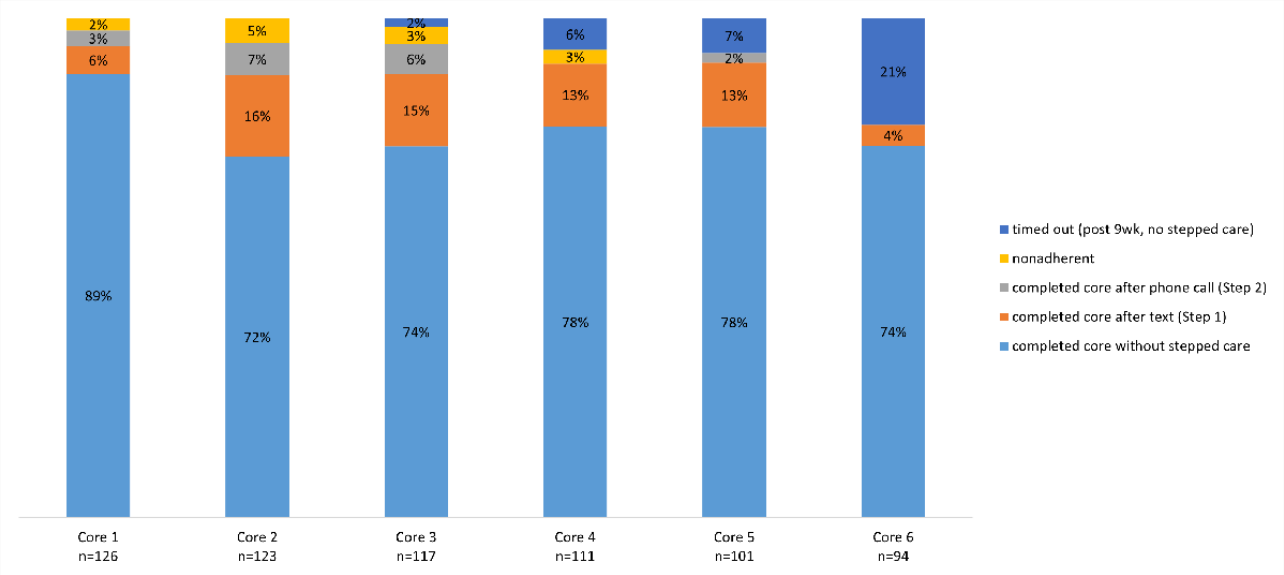

Longitudinal tracking of participant-level stepped care demand revealed that support was used by many participants across Cores, rather than concentrated among a few. This pattern is illustrated in Figure S1 in Multimedia Appendix 3.

Stepped Care Implementation Time and Costs

Table 2 presents descriptive statistics for the time and cost associated with each component of the stepped care process. On average, implementing the full stepped care process required 26.46 (SD 11.02) minutes per participant and cost US $14.80 (SD 6.16; Table 2). The initial monitoring step was the most resource-intensive, averaging US $8.51 (SD 2.35) and accounting for 58% of the total implementation cost. Altogether, monitoring-related costs, including initial, text, phone, and nonadherence monitoring, comprised US $11.61, or 78% of the total average cost (Table 2).

Implementation costs varied across Cores. Core 2 had the highest average cost (US $3.52, SD 4.2) and time commitment (6.3, SD 7.51 minutes), while Core 6 had the lowest average cost (US $1.17, SD 1.17) and time (2.09, SD 2.1 minutes; Table 2).

Figure 4 illustrates the distribution of stepped care component costs by Core. Initial monitoring was the highest cost across all Cores, peaking at Core 1 and gradually declining. Costs for step 1 (text + text monitoring) and step 2 (phone + phone monitoring) peaked at Core 2, then declined in subsequent Cores. In contrast, nonadherence monitoring costs rose over time, peaking at Core

XSL•FO
RenderX

6. The total Core-specific cost, the sum of all components, was highest at Core 2 and declined thereafter.

Table 2. Descriptive statistics (time and cost) of stepped care implementation for the digital iSIPsmarter[a] intervention among Appalachian adults.

Variable[b]	Time (min), mean (SD; range)	Cost (US $), mean (SD; range)
Total	26.46 (11.02; 15.00-58.16)	14.80 (6.16; 8.39-32.52)
Initial monitoring	15.21 (4.19; 3-18)	8.51 (2.35; 1.68-10.07)
Text (step 1)	2.58 (2.58; 0-8.49)	1.44 (1.44; 0-4.75)
Monitoring text (step 1)	2.57 (2.57; 0-9)	1.44 (1.43; 0-5.03)
Phone (step 2)	3.12 (5.88; 0-25.5)	1.74 (3.29; 0-14.26)
Monitoring phone (step 2)	1.60 (3.21; 0-12)	0.89 (1.80; 0-6.71)
Nonadherent monitoring	1.38 (3.79; 0-15)	0.77 (2.12; 0-8.39)
By Core		
Core 1	4.07 (4.31; 3.00-33.13)	2.28 (2.41; 1.68-18.53)
Core 2	6.30 (7.51; 0-33.13)	3.52 (4.20; 0-18.53)
Core 3	5.72 (7.27; 0-33.13)	3.20 (4.06; 0-18.53)
Core 4	3.84 (5.14; 0-33.13)	2.15 (2.88; 0-18.53)
Core 5	3.06 (2.99; 0-16.93)	1.71 (1.67; 0-9.47)
Core 6	2.09 (2.10; 0-8.83)	1.17 (1.17; 0-4.94)

[a]iSIPsmarter is a digital health intervention aimed at reducing sugar-sweetened beverage (SSB) consumption; data represent the initial 9-week intervention period

[b]The sample size for the stepped care intervention is 126.

Figure 4. Distribution of stepped care engagement cost (in US $) across the 6 Cores (modules) of the digital iSIPsmarter intervention among Appalachian adults (data represents the initial 9-week intervention period).

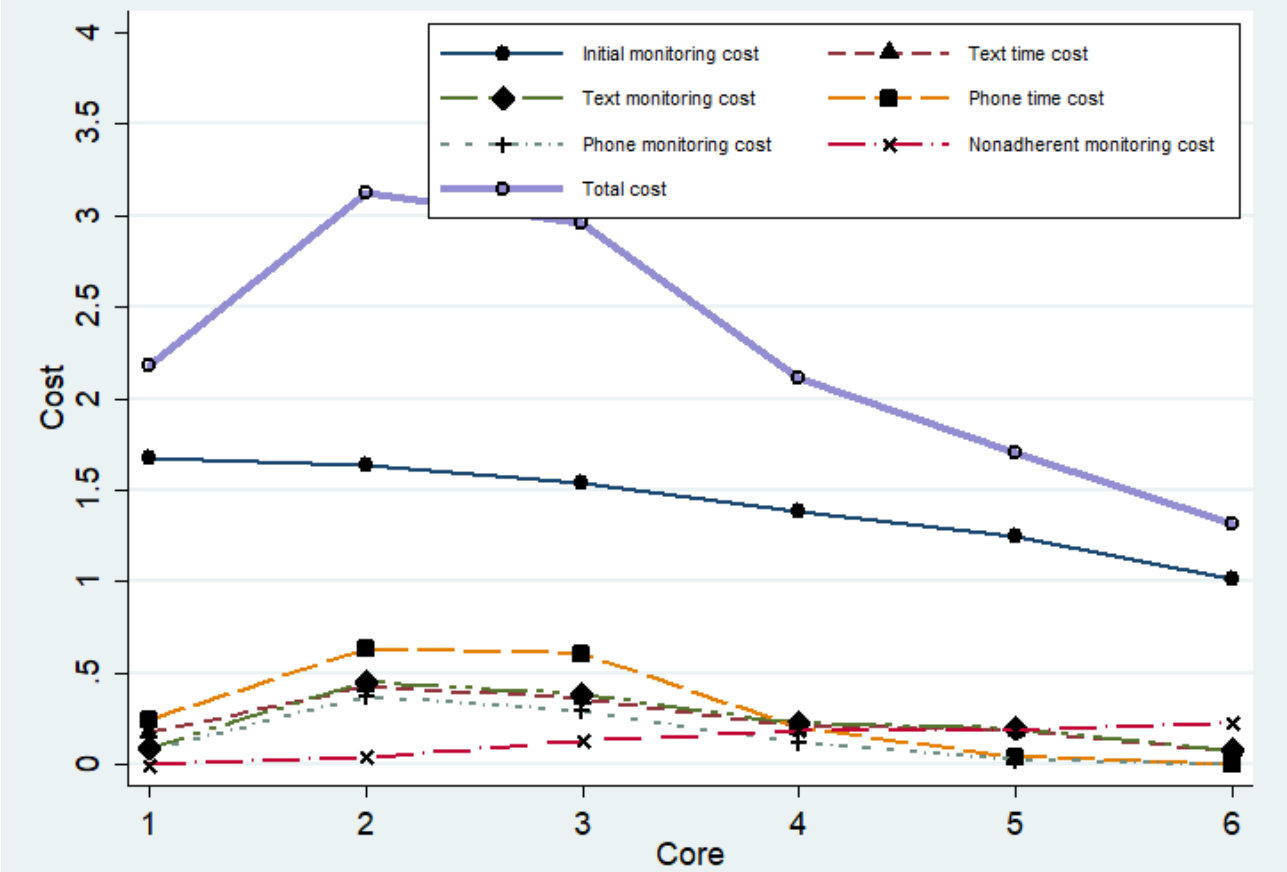

XSL•FO
RenderX

Simulations

Table 3 presents the aggregate cost estimates from the deterministic scenario simulation modeling, varying stepped care demand, monitoring efficiency, and stepped care intensity, with corresponding time estimates provided in Table S4 in Multimedia Appendix 3. Total simulated costs varied substantially across stepped care demand, monitoring cost efficiency, and stepped care intensity, ranging from mean cost of US $4.18 (95% CI 4.12-4.24) in the low demand, high monitoring efficiency, and low intensity scenario to mean cost of US $81.82 (95% CI 81.60-82.04) in the high demand, low monitoring efficiency, and high intensity scenario (Table 3). Similarly, time estimates varied across the different simulation scenarios with total time spent on low demand, high monitoring efficiency, and low intensity scenario estimated at mean time of 7.47 (95% CI 7.36-7.57) minutes while total time spent on high demand, low monitoring efficiency, and high intensity scenario was estimated at mean time of 146.32 (95% CI 145.92-146.71) minutes (Table S4 in Multimedia Appendix 3).

Consistent with the anticipated simulation (see “pattern 1” in the section “Anticipated Simulation Pattern”), high-intensity stepped care resulted in substantially greater implementation costs than low-intensity care across all demand and monitoring levels. For example, under a high-demand scenario (101/126, 80%) of participants requiring stepped care), shifting from high- to low-intensity care decreased total costs by US $60.37 and time by 107.96 minutes, a 74% decrease, under inefficient monitoring, and by US $45.87 and 82.03 minutes, an 81% decrease, under efficient monitoring. This cost and time pattern was consistent across moderate (63/126, 50%) and low (25/126, 20%) demand scenarios. Likewise, consistent with anticipated simulation (see “pattern 2” in the section “Anticipated Simulation Pattern”), reductions in monitoring costs led to substantial savings. Under low-intensity care and low demand, reducing monitoring costs from 80% to 20% of baseline levels resulted in a US $6.77 (62%) reduction in total costs. Under high-intensity care and high demand, this reduction yielded a US $24.9 (30%) decrease in total costs.

XSL•FO
RenderX

Table 3. Aggregate participant cost of implementation of stepped care engagement of the digital iSIPsmarter[a] intervention among Appalachian adults: results from simulating varying levels of demand for stepped care, monitoring, and intensity of stepped care (values are means of 1000 simulations; in US $; mean, 95% CI)[b].

Monitoring level[c]	Demand for stepped care[d]	Intensity[e]	C_{total}	C^{IM}	[illegible]	[illegible]	[illegible]	[illegible]	C^{NAM}
Low	Low	Low	4.18; 4.12-4.24	1.70; 1.62-1.78	1.92; 1.90-1.94	0.40; 0.40-0.41	—[f]	—	0.15; 0.08-0.23
Low	Low	High	15.60; 15.51-15.70	1.70; 1.62-1.78	1.92; 1.90-1.94	0.40; 0.40-0.41	10.22; 10.15-10.29	1.21; 1.20-1.22	0.15; 0.08-0.23
Low	Medium	Low	7.61; 7.55-7.67	1.70; 1.62-1.78	4.75; 4.72-4.78	1.00; 0.99-1.01	—	—	0.15; 0.08-0.23
Low	Medium	High	36.28; 36.16-36.39	1.70; 1.62-1.78	4.75; 4.72-4.78	1.00; 0.99-1.01	25.64; 25.55-25.73	3.02; 3.01-3.03	0.15; 0.08-0.23
Low	High	Low	11.05; 10.99-11.11	1.70; 1.62-1.78	7.58; 7.56-7.61	1.61; 1.60-1.61	—	—	0.15; 0.08-0.23
Low	High	High	56.92; 56.84-57.01	1.70; 1.62-1.78	7.58; 7.56-7.61	1.61; 1.60-1.61	41.04; 40.98-41.11	4.83; 4.82-4.84	0.15; 0.08-0.23
Medium	Low	Low	7.56; 7.43-7.69	4.25; 4.05-4.46	1.92; 1.90-1.94	1.00; 0.99-1.01	—	—	0.39; 0.20-0.57
Medium	Low	High	20.80; 20.65-20.96	4.25; 4.05-4.46	1.92; 1.90-1.94	1.00; 0.99-1.01	10.22; 10.15-10.29	3.02; 3.01-3.04	0.39; 0.20-0.57
Medium	Medium	Low	11.90; 11.77-12.03	4.25; 4.05-4.46	4.75; 4.72-4.78	2.51; 2.50-2.52	—	—	0.39; 0.20-0.57
Medium	Medium	High	45.10; 44.93-45.27	4.25; 4.05-4.46	4.75; 4.72-4.78	2.51; 2.50-2.52	25.64; 25.55-25.73	7.56; 7.53-7.58	0.39; 0.20-0.57
Medium	High	Low	16.24; 16.12-16.38	4.25; 4.05-4.46	7.58; 7.56-7.61	4.02; 4.01-4.03	—	—	0.39; 0.20-0.57
Medium	High	High	69.37; 69.22-69.52	4.25; 4.05-4.46	7.58; 7.56-7.61	4.02; 4.01-4.03	41.04; 40.98-41.12	12.08; 12.06-12.10	0.39; 0.20-0.57
High	Low	Low	10.95; 10.74-11.16	6.81; 6.48-7.14	1.92; 1.90-1.94	1.61; 1.59-1.62	—	—	0.62; 0.32-0.92
High	Low	High	26.00; 25.78-26.23	6.81; 6.48-7.14	1.92; 1.90-1.94	1.61; 1.59-1.62	10.22; 10.15-10.29	4.84; 4.81-4.86	0.62; 0.32-0.92
High	Medium	Low	16.19; 15.99-16.40	6.81; 6.48-7.14	4.75; 4.72-4.78	4.02; 4.00-4.04	—	—	0.62; 0.32-0.92
High	Medium	High	53.92; 53.69-54.16	6.81; 6.48-7.14	4.75; 4.72-4.78	4.02; 4.00-4.04	25.64; 25.55-25.73	12.09; 12.05-12.13	0.62; 0.32-0.92
High	High	Low	21.44; 21.24-21.65	6.81; 6.48-7.14	7.58; 7.56-7.61	6.43; 6.42-6.45	—	—	0.62; 0.32-0.92
High	High	High	81.82; 81.60-82.04	6.81; 6.48-7.14	7.58; 7.56-7.61	6.43; 6.42-6.45	41.04; 40.98-41.11	19.33; 19.30-19.36	0.62; 0.32-0.92

[a]iSIPsmarter is a digital health intervention aimed at reducing sugar-sweetened beverage consumption; data represent the initial 9-week intervention period

[b]Refer to "Definition of Time Spent and Costs Incurred in the Stepped Care Engagement Process" for the definition of the cost variables.

[c]Monitoring levels: low (20% monitoring cost); medium (50% monitoring cost); high (80% monitoring cost).

[d]Demand for stepped care levels: low (20% participants needing stepped care); medium (50% participants needing stepped care); high (80% participants needing stepped care).

[e]Intensity levels: low (stepped care up to step 1); high (stepped care up to step 2).

[f]Not applicable.

Simulated cost estimates by each Core are presented in Table S3 in Multimedia Appendix 3 with the corresponding time estimates in Table S5 in Multimedia Appendix 3. For low demand, high monitoring efficiency, and low intensity scenario, Core 1 is estimated to have mean cost of US $0.72 (95% CI 0.72-0.73) and require mean of 1.29 (95% CI 1.29-1.30) minutes while Core 6 is estimated to mean cost of US $0.59 (95% CI 0.56-0.62) and require mean of 1.05 (95% CI 1.00-1.11) minutes (Tables S3 and S5 in Multimedia Appendix 3). Similar to aggregated results, time and cost estimates increased with

XSL•FO
RenderX

high-resource intensity strategies, with higher demand for stepped care, and with low monitoring efficiency across all Cores.

Figure 5 visually compares the implementation costs across all simulation scenarios. Here, D denotes the demand for stepped care. I denotes intensity (low I - up to Step 1 [text reminders], high I - up to Step 2 [phone calls]). Each panel represents 1 of the 3 monitoring efficiency levels (low, medium, and high). Within each panel, cost components are stacked to illustrate their contribution under varying demand and intensity conditions. This figure highlights how total costs shift depending on simulation inputs and illustrates the dominant role of monitoring and intensity in driving cost differences across scenarios. For example, comparing the first panel (low monitoring) of high monitoring efficiency with the third panel (high monitoring) of low monitoring efficiency, the monitoring components (initial monitoring + text monitoring + phone monitoring + nonadherent monitoring) are noticeably larger in the third panel as compared to the first panel, illustrating the higher implementation cost of scenarios with inefficient monitoring. Similarly, comparing low- to high-intensity care, for higher intensity (high I) scenarios, costs of Step 2 components (phone time + phone monitoring) are added to the implementation costs as compared to lower intensity (low I) scenarios, under similar levels of monitoring and demand for stepped care. As also shown in Table 3, nonadherence monitoring contributes only a negligible share of total costs, which is why it appears nearly invisible in Figure 5.

Figure 5. Distribution of the cost (in US $) of stepped care from simulating varying levels of demand for stepped care, monitoring, and intensity of stepped care for the digital iSIPsmarter intervention among Appalachian adults (data represent the initial 9-week intervention period).

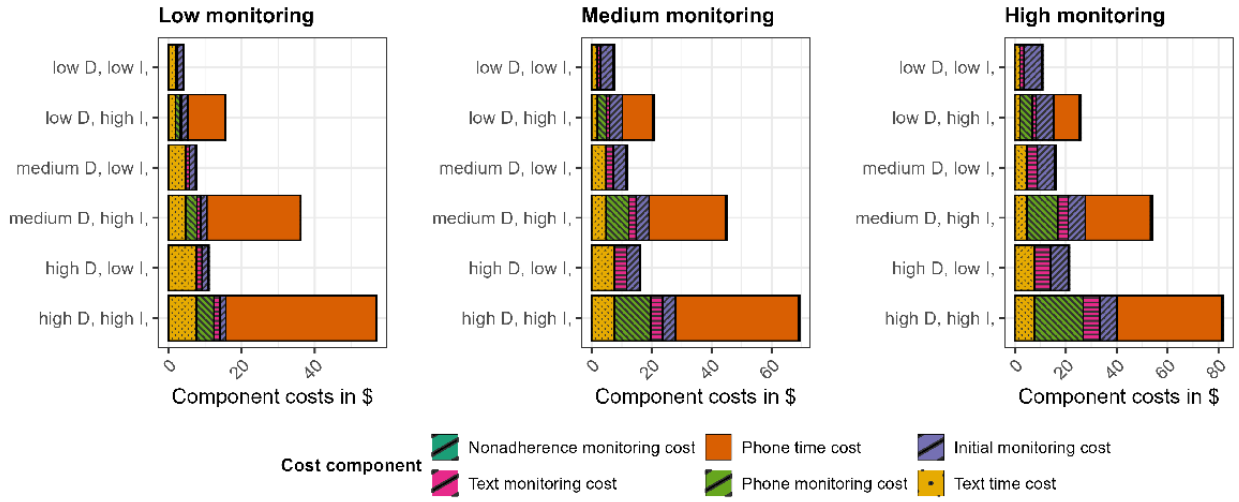

Discussion

Principal Results

This study evaluated the demand and implementation costs of a stepped care engagement approach within the iSIPsmarter intervention and explored how variations in monitoring, demand, and intensity influence overall resource use. Our findings indicate that the intervention can engage most participants effectively without additional human support, while a subset of participants benefits from stepped care for specific or multiple intervention components. Monitoring was the largest driver of implementation costs, and simulation analyses demonstrated that both demand and intensity strongly influence total resource use. These results highlight the potential for adaptive, tiered engagement strategies to sustain participant involvement while efficiently allocating resources and providing actionable insights for optimizing future implementation of DHIs.

Comparison With Prior Work

Our study introduces a novel application of stepped care to address one of the persistent challenges in DHIs, namely, sustaining user engagement [6,7]. Engagement in DHIs is a multifaceted construct, encompassing behavioral (ie, observable user actions within a DHI, such as Core completion), affective (ie, user's emotional response, attitudes and feelings during a DHI), and cognitive (ie, mental effort, thought, and attention users invest during a DHI interaction) components [34,35]. Guided by the Supportive Accountability Model [33], which suggests that accountability to a supportive, trustworthy person enhances adherence, this study focused on a behavioral engagement strategy delivered by study personnel. While human-supported strategies, such as e-coaching (a technology-mediated process that provides guidance, feedback, and support to help individuals achieve behavior change), can enhance engagement [46], offering them universally is resource-intensive and often unsustainable. In contrast, a stepped care approach allocates resources efficiently, escalating human support only when disengagement occurs [14-17]. This is consistent with frameworks such as the Internet Intervention Model that emphasize the interplay of user needs, intervention features, and contextual factors in shaping engagement [12,13,31].

In our stepped care engagement approach, human-supported text messages and phone calls were used to reengage participants, troubleshoot barriers, and encourage completion of digital modules (Cores). Qualitative findings from a summative evaluation of iSIPsmarter indicated that participants valued these personal contacts, expressing appreciation for their

XSL•FO
RenderX

interactions with the research team while complementary quantitative results showed that more than half of the participants believed that phone contact with an expert would have helped them be more successful [50]. Prior research on adding human facilitation into engagement strategies for DHIs has shown mixed results [51,52]. Although reviews suggest that incorporating human or social support can improve user experience and perceived engagement, effects on objective use outcomes are generally small and inconsistent across studies [52]. Remote facilitation delivered through telephone, email, or text-based coaching has shown particular promise in helping participants implement interventions and sustain adherence in a scalable and cost-efficient manner [12,13,53]. Consistent with this, findings from mobile health research suggest that hybrid interventions combining automated digital features with elements of human support tend to achieve higher adherence than unguided approaches [13].

Our findings suggest that while the iSIPsmarter intervention is relatively successful at engaging participants without additional human support, a subset of participants benefit from stepped care across specific or multiple Cores. This underscores the importance of flexible systems that can allocate intensive resources to participants with varying needs. Notably, support demand was distributed across many participants rather than concentrated among a few, suggesting that disengagement reflects a broad, recurring need rather than isolated cases. Although our analysis did not assess whether early high-intensity support predicted disengagement in later Cores, future research could build on these insights to identify patterns of sustained or escalating support needs to guide resource allocation. Such insights would help determine whether high resource intensity stepped care remains cost-efficient across all Cores or is most effective when limited to initial Cores. Future interventions could refine stepped care triggers by identifying participant characteristics linked to greater support needs, allowing for more personalized and targeted engagement strategies. This approach would be especially important for supporting participants who may face greater structural or social barriers to engagement. By tailoring resource allocation to those with higher needs, a stepped care engagement approach can help reduce disparities in DHI outcomes. Future research is needed to better understand how participant characteristics may also influence the demand for stepped care. However, our previous findings suggest that Core completion rates were not significantly affected by rurality, race, gender, income, or education [45]. Notably, older age was a significant predictor of higher Core completion. This suggests that while demand for stepped care may be relatively consistent across most demographic groups, younger participants may require more stepped care support to complete the Core modules. Our stepped care framework provides a generalizable framework for optimizing engagement in DHIs where sustained participation is essential. By optimizing resources, this novel design could support scalable implementation in health care and community settings with few providers. However, direct comparisons with other interventions are limited due to the unique design and context of this approach, emphasizing the need for further research to establish benchmarks and assess generalizability.

From a cost perspective, we used a TDABC approach to estimate the implementation cost of the stepped care engagement components [42]. Monitoring accounted for the majority of implementation expenses, highlighting it as a primary driver of resource use in our model of stepped care. Although the iSIPsmarter intervention was not designed to test the effectiveness of stepped care, it offers valuable insights into cost optimization strategies for stepped care engagement approaches. Simulation analysis demonstrated that demand and intensity strongly influence costs, with high-intensity, high-demand scenarios producing the greatest burden. Importantly, the simulated efficiency gains, reflecting potential automation of monitoring processes, indicate the potential for substantial cost savings across all demand levels. These findings point to a key opportunity for future DHI design: integrating scalable automation tools such as tracking systems, workflow platforms, or learning management systems, to reduce monitoring burden while preserving the human-supported elements critical to engagement. Prior literature on the application of simulation-driven decision support systems in health care highlights the value of using simulation to generate quantitative insights that inform effective resource management [54,55]. Such approaches have been applied to analyze capacity constraints, reduce costs, and identify optimal process configuration [56,57]. Thus, these simulation findings help define key parameters for progressing from the current RCT efficacy trial [29,32] to a hybrid effectiveness-implementation study in partnership with resource-constrained health care systems that may ultimately adopt stepped care engagement strategies within DHIs.

Limitations

We acknowledge several limitations of this investigation. First, the estimates of demand and costs for the stepped care components are specific to the study population and intervention context, and as such, they cannot be readily generalized to other settings without further evidence. Second, the time data used in this study are self-reported and estimated by project staff, which may introduce response biases and affect the accuracy of the cost and time estimates. Third, the absence of a control group limits our ability to evaluate the effectiveness of stepped care in reducing SSB consumption or other behavioral outcomes. Fourth, we assume no change in Core completion rates in simulation scenarios with lower monitoring costs.

Future Work

Future research should focus on exploring the effectiveness and cost-effectiveness of stepped care on both engagement and health outcomes, perhaps through an optimization trial whereby stepped care is experimentally varied (eg, turned on or off, or delivered at higher or lower intensity) [58]. Similarly, the integration of additional automated technologies should be explored to further enhance participant engagement and improve health outcomes, while containing implementation costs. Future research should also aim to optimize engagement by determining which types of support are most effective for reengaging participants and by identifying a priori which individuals are most likely to benefit from each approach. Finally, examining the success of stepped care engagement strategies across various

XSL•FO
RenderX

digital health contexts and populations will help expand the generalizability of these findings.

Conclusions

In conclusion, this study demonstrates an innovative application of stepped care to support engagement in DHIs, addressing a common challenge in sustaining user participation. Unlike most prior approaches that provide either universal human support [12,13] or fully automated interventions [59], this tiered, adaptive strategy selectively allocates support to participants who need it most. By showing how stepped care can maintain engagement efficiently while optimizing resource use, the findings contribute actionable insights for designing scalable and cost-effective interventions. This is especially timely and relevant to the field, particularly as DHIs rapidly expand while sustainable, resource-efficient engagement strategies remain a persistent challenge. In practice, these results suggest that integrating stepped care with automated technologies could enhance adherence, reduce disparities, and support sustainable implementation of digital health programs across diverse populations and resource-limited real-world settings.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: code optimization, proofreading, and editing. The GAI tool used was ChatGPT-4.5. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by: ED and JZ.

Funding

This study was funded by the National Institutes of Health, National Institute on Minority Health and Health Disparities (R01MD015033). The National Institutes of Health was not involved in the design of this study or writing of this manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JMZ, WY, and LMR were responsible for the acquisition of funding. ED, WY, and JMZ prepared the first draft of the manuscript. All authors provided critical review, editing, and approval of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
TIDieR checklist.
[PDF File (Adobe PDF File), 235 KB - jmir_v28i1e80921_app1.pdf]

Multimedia Appendix 2
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1117 KB - jmir_v28i1e80921_app2.pdf]

Multimedia Appendix 3
Tabular and visual summaries of all the analyses with the digital intervention.
[DOCX File , 241 KB - jmir_v28i1e80921_app3.docx]

References

1. Forman EM, Evans BC, Flack D, Goldstein SP. Could technology help us tackle the obesity crisis? Future Sci OA 2016;2(4):FSO151 [FREE Full text] [doi: 10.4155/fsoa-2016-0061] [Medline: 28116133]
2. Moller AC, Merchant G, Conroy DE, West R, Hekler E, Kugler KC, et al. Applying and advancing behavior change theories and techniques in the context of a digital health revolution: proposals for more effectively realizing untapped potential. J Behav Med 2017;40(1):85-98 [FREE Full text] [doi: 10.1007/s10865-016-9818-7] [Medline: 28058516]
3. Rudd BN, Beidas RS. Digital mental health: the answer to the global mental health crisis? JMIR Ment Health 2020;7(6):e18472 [FREE Full text] [doi: 10.2196/18472] [Medline: 32484445]

XSL•FO
RenderX

4. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019;21(7):e12869. [doi: 10.2196/12869] [Medline: 31333198]
5. Patel S, Akhtar A, Malins S, Wright N, Rowley E, Young E, et al. The acceptability and usability of digital health interventions for adults with depression, anxiety, and somatoform disorders: qualitative systematic review and meta-synthesis. J Med Internet Res 2020;22(7):e16228 [FREE Full text] [doi: 10.2196/16228] [Medline: 32628116]
6. Cole-Lewis H, Ezeanochie N, Turgiss J. Understanding health behavior technology engagement: pathway to measuring digital behavior change interventions. JMIR Form Res 2019;3(4):e14052 [FREE Full text] [doi: 10.2196/14052] [Medline: 31603427]
7. Grady A, Yoong S, Sutherland R, Lee H, Nathan N, Wolfenden L. Improving the public health impact of eHealth and mHealth interventions. Aust N Z J Public Health 2018;42(2):118-119 [FREE Full text] [doi: 10.1111/1753-6405.12771] [Medline: 29384248]
8. Donkin L, Christensen H, Naismith SL, Neal B, Hickie IB, Glozier N. A systematic review of the impact of adherence on the effectiveness of e-therapies. J Med Internet Res 2011;13(3):e52 [FREE Full text] [doi: 10.2196/jmir.1772] [Medline: 21821503]
9. Alexander GL, McClure JB, Calvi JH, Divine GW, Stopponi MA, Rolnick SJ, MENU Choices Team. A randomized clinical trial evaluating online interventions to improve fruit and vegetable consumption. Am J Public Health 2010;100(2):319-326 [FREE Full text] [doi: 10.2105/AJPH.2008.154468] [Medline: 20019315]
10. Tate DF, Wing RR, Winett RA. Using internet technology to deliver a behavioral weight loss program. JAMA 2001;285(9):1172-1177. [doi: 10.1001/jama.285.9.1172] [Medline: 11231746]
11. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017;7(2):254-267 [FREE Full text] [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
12. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
13. Jakob R, Harperink S, Rudolf AM, Fleisch E, Haug S, Mair JL, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022;24(5):e35371 [FREE Full text] [doi: 10.2196/35371] [Medline: 35612886]
14. Davison GC. Stepped care: doing more with less? J Consult Clin Psychol 2000 Aug;68(4):580-585. [Medline: 10965633]
15. Kazdin AE. Addressing the treatment gap: a key challenge for extending evidence-based psychosocial interventions. Behav Res Ther 2017;88:7-18. [doi: 10.1016/j.brat.2016.06.004] [Medline: 28110678]
16. Kendall PC, Makover H, Swan A, Carper MM, Mercado R, Kagan E, et al. What steps to take? How to approach concerning anxiety in youth. Clin Psychol Sci Pract 2016 Sep;23(3):211-229. [doi: 10.1111/cpsp.12156]
17. Salloum A. Minimal therapist-assisted cognitive–behavioral therapy interventions in stepped care for childhood anxiety. Prof Psychol Res Pr 2010;41(1):41-47. [doi: 10.1037/a0018330]
18. Scogin FR, Hanson A, Welsh D. Self-administered treatment in stepped-care models of depression treatment. J Clin Psychol 2003;59(3):341-349. [doi: 10.1002/jclp.10133] [Medline: 12579549]
19. Espie C. "Stepped care": a health technology solution for delivering cognitive behavioral therapy as a first line insomnia treatment. Sleep 2009;32(12):1549-1558. [doi: 10.1093/sleep/32.12.1549] [Medline: 20041590]
20. McKellar J, Austin J, Moos R. Building the first step: a review of low-intensity interventions for stepped care. Addict Sci Clin Pract 2012;7(1):26 [FREE Full text] [doi: 10.1186/1940-0640-7-26] [Medline: 23227807]
21. Pettit JW, Rey Y, Bechor M, Melendez R, Vaclavik D, Buitron V, et al. Can less be more? Open trial of a stepped care approach for child and adolescent anxiety disorders. J Anxiety Disord 2017;51:7-13 [FREE Full text] [doi: 10.1016/j.janxdis.2017.08.004] [Medline: 28843575]
22. Rapee RM, Lyneham HJ, Wuthrich V, Chatterton ML, Hudson JL, Kangas M, et al. Comparison of stepped care delivery against a single, empirically validated cognitive-behavioral therapy program for youth with anxiety: a randomized clinical trial. J Am Acad Child Adolesc Psychiatry 2017;56(10):841-848. [doi: 10.1016/j.jaac.2017.08.001] [Medline: 28942806]
23. Jobes DA, Gregorian MJ, Colborn VA. A stepped care approach to clinical suicide prevention. Psychol Serv 2018;15(3):243-250. [doi: 10.1037/ser0000229] [Medline: 30080081]
24. Yeguez CE, Page TF, Rey Y, Silverman WK, Pettit JW. A cost analysis of a stepped care treatment approach for anxiety disorders in youth. J Clin Child Adolesc Psychol 2020;49(4):549-555 [FREE Full text] [doi: 10.1080/15374416.2018.1539913] [Medline: 30644757]
25. Wood M, Smith J, Healey H, Görges M, Lokker C. Enhanced recovery support for people with eating disorders during the COVID-19 pandemic: quality improvement using a web-based, stepped-care programme in Canada. BMJ Open Qual 2023;12(4):e002366 [FREE Full text] [doi: 10.1136/bmjoq-2023-002366] [Medline: 37935516]
26. Carels RA, Selensky JC, Rossi J, Solar C, Hlavka R. A novel stepped-care approach to weight loss: the role of self-monitoring and health literacy in treatment outcomes. Eat Behav 2017;26:76-82. [doi: 10.1016/j.eatbeh.2017.01.009] [Medline: 28189945]

XSL•FO
RenderX

27. Banks J, editor. Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice. New York, NY: John Wiley & Sons; 1998.
28. Zoellner JM, Porter KJ, You W, Reid AL, Frederick C, Hilgart M, et al. Study protocol for iSIPsmarter: a randomized-controlled trial to evaluate the efficacy, reach, and engagement of a technology-based behavioral intervention to reduce sugary beverages among rural appalachian adults. Contemp Clin Trials 2021;110:106566 [FREE Full text] [doi: 10.1016/j.cct.2021.106566] [Medline: 34492306]
29. Zoellner JM, You W, Porter K, Reid AL, Brock DJP, Markwalter T, et al. A digital behavioral intervention to reduce sugar-sweetened beverage consumption: a randomized, controlled trial. Am J Clin Nutr 2025;122(2):544-555. [doi: 10.1016/j.ajcnut.2025.06.010] [Medline: 40513952]
30. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions: medical research council guidance. BMJ 2015;350:h1258 [FREE Full text] [doi: 10.1136/bmj.h1258] [Medline: 25791983]
31. Ritterband LM, Thorndike FP, Cox DJ, Kovatchev BP, Gonder-Frederick LA. A behavior change model for internet interventions. Ann Behav Med 2009;38(1):18-27 [FREE Full text] [doi: 10.1007/s12160-009-9133-4] [Medline: 19802647]
32. Zoellner J, Reid A, Porter K, Frederick C, Hilgart M, Ritterband L. Development of a digital behavioral intervention to reduce the consumption of sugar-sweetened beverages among rural appalachian adults: multiphased, human-centered design approach. JMIR Hum Factors 2023;10:e41262 [FREE Full text] [doi: 10.2196/41262] [Medline: 36724036]
33. Mohr DC, Cuijpers P, Lehman K. Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions. J Med Internet Res 2011;13(1):e30 [FREE Full text] [doi: 10.2196/jmir.1602] [Medline: 21393123]
34. Nahum-Shani I, Shaw SD, Carpenter SM, Murphy SA, Yoon C. Engagement in digital interventions. Am Psychol 2022;77(7):836-852 [FREE Full text] [doi: 10.1037/amp0000983] [Medline: 35298199]
35. Milne-Ives M, Homer S, Andrade J, Meinert E. The conceptualisation and measurement of engagement in digital health. Internet Interv 2024;36:100735 [FREE Full text] [doi: 10.1016/j.invent.2024.100735] [Medline: 38558760]
36. Roberts SLE, Healey A, Sevdalis N. Use of health economic evaluation in the implementation and improvement science fields-a systematic literature review. Implement Sci 2019;14(1):72 [FREE Full text] [doi: 10.1186/s13012-019-0901-7] [Medline: 31307489]
37. Smith JD, Hasan M. Quantitative approaches for the evaluation of implementation research studies. Psychiatry Res 2020;283:112521 [FREE Full text] [doi: 10.1016/j.psychres.2019.112521] [Medline: 31473029]
38. Eisman AB, Kilbourne AM, Dopp AR, Saldana L, Eisenberg D. Economic evaluation in implementation science: making the business case for implementation strategies. Psychiatry Res 2020;283:112433 [FREE Full text] [doi: 10.1016/j.psychres.2019.06.008] [Medline: 31202612]
39. Cidav Z, Mandell D, Pyne J, Beidas R, Curran G, Marcus S. A pragmatic method for costing implementation strategies using time-driven activity-based costing. Implement Sci 2020;15(1):28 [FREE Full text] [doi: 10.1186/s13012-020-00993-1] [Medline: 32370752]
40. Kaplan RS, Witkowski M, Abbott M, Guzman AB, Higgins LD, Meara JG, et al. Using time-driven activity-based costing to identify value improvement opportunities in healthcare. J Healthc Manag 2014;59(6):399-412. [doi: 10.1097/00115514-201411000-00005]
41. Keel G, Savage C, Rafiq M, Mazzocato P. Time-driven activity-based costing in health care: a systematic review of the literature. Health Policy 2017;121(7):755-763 [FREE Full text] [doi: 10.1016/j.healthpol.2017.04.013] [Medline: 28535996]
42. Huebschmann AG, Trinkley KE, Gritz M, Glasgow RE. Pragmatic considerations and approaches for measuring staff time as an implementation cost in health systems and clinics: key issues and applied examples. Implement Sci Commun 2022;3(1):44 [FREE Full text] [doi: 10.1186/s43058-022-00292-4] [Medline: 35428326]
43. Occupational employment and wage statistics. US Bureau of Labor Statistics. URL: https://www.bls.gov/oes/ [accessed 2026-03-24]
44. Laxminarayan R, Chow J, Shahid-Salles S. Intervention cost-effectiveness: overview of main messages. In: Jamison DT, Breman JG, Measham AR, editors. Disease Control Priorities in Developing Countries 2nd ed. Washington (DC): The International Bank for Reconstruction and Development/The World Bank; 2006.
45. Brock DP, Ritterband LM, You W, Reid AL, Porter KJ, Markwalter T, et al. An exploratory study to understand how rurality status and demographic characteristics are associated with enrollment, engagement, and retention in a digital health intervention targeting the appalachian region. J Rural Health 2025;41(2):e70042. [doi: 10.1111/jrh.70042] [Medline: 40515483]
46. Saleem M, Kühne L, De Santis KK, Christianson L, Brand T, Busse H. Understanding engagement strategies in digital interventions for mental health promotion: scoping review. JMIR Ment Health 2021;8(12):e30000 [FREE Full text] [doi: 10.2196/30000] [Medline: 34931995]
47. Mclaughlin M, Delaney T, Hall A, Byaruhanga J, Mackie P, Grady A, et al. Associations between digital health intervention engagement, physical activity, and sedentary behavior: systematic review and meta-analysis. J Med Internet Res 2021;23(2):e23180 [FREE Full text] [doi: 10.2196/23180] [Medline: 33605897]
48. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: Template for Intervention Description and Replication (TIDieR) checklist and guide. BMJ 2014;348:g1687 [FREE Full text] [doi: 10.1136/bmj.g1687] [Medline: 24609605]

XSL•FO
RenderX

49. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
50. Porter KJ, Brock DJP, Reid AL, Markwalter T, Jones C, Ritterband LM, et al. Participants' experiences with and perceptions of a digital health intervention to reduce sugar-sweetened beverage intake: a mixed methods summative evaluation of iSIPsmarter. Internet Interv 2026 Mar;43:100921. [doi: 10.1016/j.invent.2026.100921]
51. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on Internet-based mental health interventions — a systematic review. Internet Interv 2014;1(4):205-215. [doi: 10.1016/j.invent.2014.08.003]
52. Grady A, Pearson N, Lamont H, Leigh L, Wolfenden L, Barnes C, et al. The effectiveness of strategies to improve user engagement with digital health interventions targeting nutrition, physical activity, and overweight and obesity: systematic review and meta-analysis. J Med Internet Res 2023;25:e47987 [FREE Full text] [doi: 10.2196/47987] [Medline: 38113062]
53. Spring B, Duncan JM, Janke EA, Kozak AT, McFadden HG, DeMott A, et al. Integrating technology into standard weight loss treatment: a randomized controlled trial. JAMA Intern Med 2013;173(2):105-111 [FREE Full text] [doi: 10.1001/jamainternmed.2013.1221] [Medline: 23229890]
54. Power DJ, Sharda R. Model-driven decision support systems: concepts and research directions. Decis Support Syst 2007;43(3):1044-1061. [doi: 10.1016/j.dss.2005.05.030]
55. Ruiz M, Orta E, Sánchez J. A simulation-based approach for decision-support in healthcare processes. Simul Model Pract Theory 2024;136:102983. [doi: 10.1016/j.simpat.2024.102983]
56. Roy SN, Shah BJ, Gajjar H. Application of simulation in healthcare service operations: a review and research agenda. ACM Trans Model Comput Simul 2020;31(1):1-23. [doi: 10.1145/3427753]
57. Kar E, Eldabi T, Fakhimi M. Hybrid simulation in healthcare: a review of the literature. : IEEE; 2022 Presented at: Winter Simulation Conference (WSC); 2022 December 11-14; Singapore p. 1211-1222. [doi: 10.1109/wsc57314.2022.10015418]
58. Collins LM. Optimization of Behavioral, Biobehavioral, and Biomedical Interventions: The Multiphase Optimization Strategy (MOST). Cham, Switzerland: Springer; 2018.
59. Woo Baidal JA, Nichols K, Charles N, Chernick L, Duong N, Finkel MA, et al. Text messages to curb sugar-sweetened beverage consumption among pregnant women and mothers: a mobile health randomized controlled trial. Nutrients 2021;13(12):4367 [FREE Full text] [doi: 10.3390/nu13124367] [Medline: 34959919]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
DHI: digital health intervention
OEWS: Occupational Employment and Wage Statistics
RCT: randomized controlled trial
SSB: sugar sweetened beverages
TDABC: time-driven activity-based costing
TIDieR: Template for Intervention Description and Replication

Edited by S Brini; submitted 18.Jul.2025; peer-reviewed by C Gissel, B Wolfe-Piggott; comments to author 22.Sep.2025; accepted 17.Mar.2026; published 04.May.2026.

Please cite as:
Dwibedi E, You W, Ritterband LM, Brock DJP, Reid AL, Frederick C, Zoellner JM
A Stepped Care Approach to Promoting Engagement in a Digital Health Intervention: Secondary Analysis of a Randomized Controlled Trial
J Med Internet Res 2026;28:e80921
URL: https://www.jmir.org/2026/1/e80921
doi:10.2196/80921
PMID:

©Esha Dwibedi, Wen You, Lee M Ritterband, Donna-Jean P Brock, Annie L Reid, Christina Frederick, Jamie M Zoellner. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development of Virtual Reality Health Literacy: Delphi Expert Consensus Study

Junghee Yoon[1,2], PhD; Mangyeong Lee[3,4], PhD; Dokyoon Kim[3,5], MSc; Joungwon Park[1,6], BSc; Su jin Kim[3,4], MSc; Jiyoon Han[1,4], MSc; Juhee Cho[1,2,4,7], PhD

[1]Department of Clinical Research Design and Evaluation, Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University, 115 Irwon-ro, Gangnam-gu, Seoul, Republic of Korea

[2]Institute for Quality of Life in Cancer, Samsung Medical Center, Seoul, Republic of Korea

[3]Department of Digital Health, Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University, Seoul, Republic of Korea

[4]Center for Clinical Epidemiology, Samsung Medical Center, Seoul, Republic of Korea

[5]AI Research Center, Research Institute for Future Medicine, Samsung Medical Center, Seoul, Republic of Korea

[6]Division of Social Work, Samsung Medical Center, Seoul, Republic of Korea

[7]Cancer Education Center, Samsung Medical Center, Seoul, Republic of Korea

Corresponding Author:
Juhee Cho, PhD
Department of Clinical Research Design and Evaluation, Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University, 115 Irwon-ro, Gangnam-gu, Seoul, Republic of Korea

Abstract

Background: Virtual reality (VR) is a promising tool in health care, offering immersive and interactive environments that can enhance patient education, rehabilitation, and mental health interventions. However, effective patient engagement with head-mounted display (HMD)–based immersive VR depends on a combination of functional competencies and readiness-related determinants that have not yet been systematically defined.

Objective: This study aimed to conceptualize an initial framework of VR health literacy, focused on HMD-based immersive VR in clinical settings, and to achieve expert consensus on its definition, domains, and subdomains.

Methods: A 3-phase modified Delphi study was conducted between January and April 2024, including a literature review in MEDLINE (via PubMed) and Embase (2017-2023) informed by scoping review methodology, a multidisciplinary expert panel formation, and 2 online survey rounds, in which panelists rated each subdomain on a 4-point necessity scale and provided open-ended feedback, with 15 experts from the health care, VR, and health literacy fields. Consensus was defined using IQRs and agreement thresholds; items with moderate consensus were further evaluated through structured internal deliberation.

Results: A total of 15 experts participated in Round 1, and 13 continued to Round 2 (retention rate, 87%). An initial structure of 7 candidate domains with 23 subdomains was iteratively refined across the two rounds based on consensus levels, expert panel feedback, and internal deliberation, with consensus thresholds applied as guiding criteria rather than automatic exclusion rules; subdomains with moderate consensus were further evaluated through structured internal deliberation to determine theoretical necessity within the framework. The final framework comprised 5 domains and 14 subdomains: performance expectancy (perceived usefulness of VR for health management; expectations of future VR benefits); effort expectancy (perceived immersion or embodiment; perceived interactivity and responsiveness; understanding of VR-related terms); facilitating conditions (access to VR devices and platforms; digital knowledge and confidence; technical proficiency with VR devices; digital self-efficacy); attitudes toward VR (awareness of VR in health contexts; interest in VR technology; problem-solving ability using VR content); and behavioral intention (intention to use VR technology or services; willingness to engage with VR for health).

Conclusions: This study presents an initial consensus-based framework of VR health literacy for HMD-based immersive VR in clinical settings. Developed through a multidisciplinary Delphi process, the framework combines operational competencies with engagement-related determinants to provide both theoretical clarity and practical use, offering guidance for clinicians, educators, and policymakers to design and implement VR interventions that are accessible, equitable, and effective in health care contexts.

(J Med Internet Res 2026;28:e85842) doi:10.2196/85842

KEYWORDS

virtual reality; VR; health literacy; Delphi study; domains

XSL•FO
RenderX

Introduction

Virtual reality (VR) is a technology that generates an artificial, 3-dimensional environment using computer systems, allowing users to experience simulations that resemble the real world through displays, ranging from isolated screens to wearable options like head-mounted displays (HMDs) [1,2]. By engaging multiple senses, VR offers immersive experiences with realistic 360-degree visuals, real-time interaction, and artificial intelligence–driven personalization through intelligent avatars, enabling active user participation and creating a dynamic virtual environment unlike traditional media [3]. In this study, the term "VR" specifically refers to immersive VR systems delivered through HMD devices. While extended reality (XR) encompasses a broader spectrum, including augmented reality and mixed reality, the present framework focused on immersive HMD-based VR, as this modality currently represents the most widely implemented form of VR in clinical settings and presents distinct cognitive demands and safety considerations for users [4]. VR technology is a groundbreaking tool in health care interventions, offering immersive and interactive environments that enable patients to actively engage in their health management [5,6]. By simulating real-life scenarios, enhancing experiential learning, and providing tailored interventions such as multisensory high-end systems and game-based systems, VR has demonstrated promise in areas such as rehabilitation, mental health, and chronic disease management [7-9]. The capability of VR to deliver highly personalized and immersive experiences has been recognized as a pivotal advancement in the domain of patient-centered interventions [10].

The adoption of VR in health care faced several challenges, including the lack of a comprehensive framework to support patient engagement. The unique skills required for effective VR use, such as navigation and content interpretation, were often underdeveloped, limiting its accessibility and increasing inequalities in its use. A previous study highlighted that VR interaction capabilities differ significantly based on literacy and technological familiarity, with nonliterate users facing challenges in navigation, visual comprehension, and complex task execution [11]. The clinical adoption of VR has been reported to face obstacles such as limited VR knowledge, unfavorable attitudes, reluctance to use the technology, and usability difficulties, largely stemming from insufficient VR-related literacy and experience [5,12,13]. These challenges highlighted the urgent need to define and develop VR health literacy as a critical framework.

To the best of our knowledge, VR health literacy remains an unfamiliar concept within the existing literature. Therefore, developing this concept is essential to ensure that individuals can benefit from VR technologies regardless of their prior experience or technological proficiency [14,15]. Previous studies predominantly focused on usability or feasibility evaluations of VR use for patients or health-related applications among individuals with limited literacy skills [8,16,17].

To address this gap, we aimed to conceptualize VR health literacy as a framework to enhance individuals' capacity to engage with VR-based health care interventions equitably and effectively. Therefore, we conducted a modified Delphi study to systematically establish the scope and components of VR health literacy, achieving consensus on the definition and identifying the essential subdomains.

Methods

Study Design

Overview of the Delphi Process

We conducted a 3-phase modified Delphi study to develop and achieve consensus on a conceptual framework for VR health literacy. This design is suitable for identifying VR-specific factors within the framework of digital health literacy to advance the effective adoption and use of VR in clinical practice [18]. It enables expert consensus through iterative surveys and controlled feedback and offers a systematic framework for addressing complex challenges (Figure 1). The Delphi process involved 2 rounds of online surveys: the first survey was conducted between January 27 and February 8, 2024, and the second between April 18 and 25, 2024. Surveys were distributed via email, and the study adhered to the principles outlined in the Guidance on Conducting and Reporting Delphi Studies [19].

XSL•FO
RenderX

Figure 1. Overview of the 3-phase modified Delphi process used to develop the virtual reality (VR) health literacy framework. Phase 1 involved a scoping review to identify facilitators, barriers, and literacy-related determinants of VR use in health care. Phase 2 comprised internal expert discussions to develop the working definition of VR health literacy, generate preliminary domains, and form the multidisciplinary expert panel. Phase 3 consisted of 2 Delphi survey rounds. In Round 1, experts provided open-ended definitions of VR health literacy and rated preliminary domains for necessity using a 4-point Likert scale. In Round 2, experts reviewed and confirmed revised domains and subdomains and rated their necessity in relation to VR health literacy using a 4-point Likert scale.

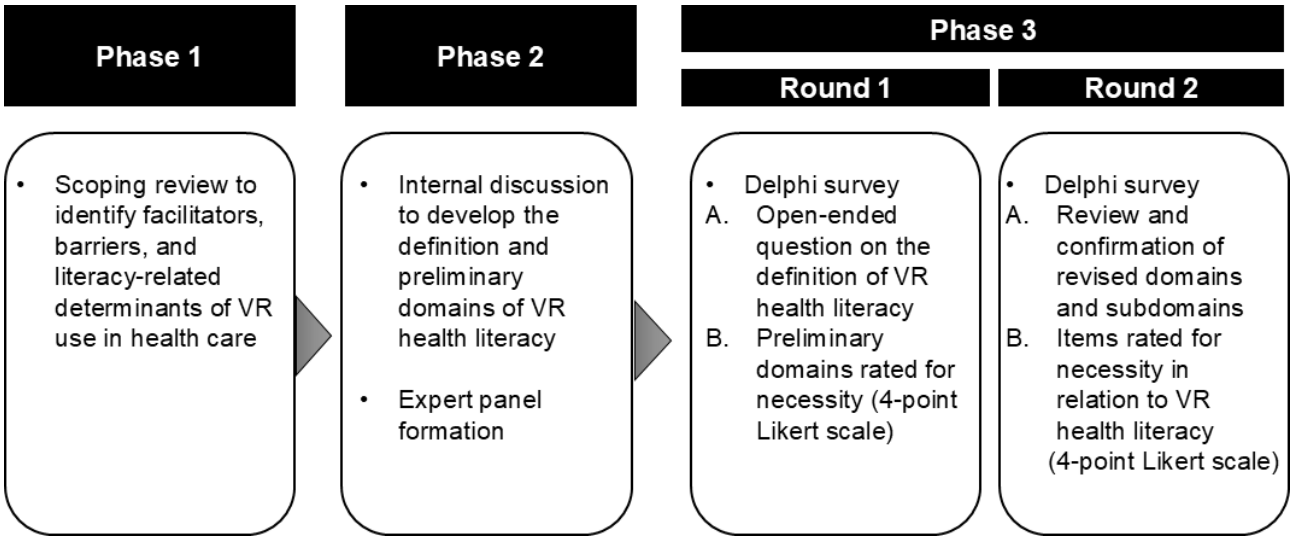

Phase 1. Literature Review Informed by Scoping Review Methodology for Item Generation

The initial phase consisted of a literature review informed by scoping review methodology to identify facilitators and barriers to VR use in health care and to explore literacy components associated with its application. An initial pilot search combining "virtual reality" and "health literacy" yielded a limited number of studies, most of which examined VR as a health literacy intervention rather than exploring literacy-related competencies required for VR engagement. Therefore, the search scope was expanded to include studies addressing usability, implementation barriers, and facilitators of VR use. PubMed and Embase were searched for studies published from January 2017 to December 2023, combining VR-related terms (eg, "virtual reality," "VR") with engagement-related terms (eg, literacy, usability, barriers, facilitators, adoption, implementation) and health-related terms (eg, health, health care, clinical); the full Boolean search strings are provided in Multimedia Appendix 1. The combined search yielded approximately 210 records after deduplication. Studies examining user engagement, implementation, or experiential aspects of health care VR in adult patients were included; records were excluded if they addressed non-VR digital technologies, pediatric or adolescent populations, provider-training contexts, or lacked a clear clinical application context. Following title and abstract screening against these criteria, approximately 30 studies were retained for closer reading, and approximately 15 of these were examined in depth as primary sources for factor extraction. Records were reviewed collaboratively by 3 research team members through iterative discussion until consensus was reached on each study's conceptual relevance, and the extracted factors informed the initial domain and subdomain generation for the Delphi rounds.

This literature review revealed that expectancy-related constructs consistently emerged as determinants of VR adoption and sustained use in health care contexts. Traditional health literacy frameworks predominantly focus on cognitive and functional skills required to access, understand, and apply health information [20]. However, engagement with technology-mediated health interventions involves an additional layer of evaluation: users must appraise not only the health content itself but also the technological medium through which this content is delivered, including its usability and perceived benefit. In immersive VR contexts, these evaluations directly influence whether users initiate and sustain engagement with the system, thereby conditioning the enactment of literacy-related competencies in practice. This gap between traditional health literacy conceptualizations and the requirements of VR-based interventions informed the incorporation of expectancy-related constructs as engagement-enabling components within our framework.

Phase 2. Expert Panel Formation and Item Development

An expert panel was convened with professionals representing health care, VR technology, and health literacy. Eligibility criteria required at least 1 prior experience in VR-related research or educational activities and included (1) health care professionals (eg, physicians, nurses, researchers), (2) experts in health care or health communication, and (3) specialists in digital health literacy research or interventions. The 1-year threshold for VR-related experience was determined after extensive internal discussion, balancing the need for practical exposure to immersive VR implementation against the limited availability of professionals with combined expertise in VR technology and health literacy in this emerging field. The second phase focused on developing the initial survey instrument for the Delphi study. Based on findings from the scoping review, an internal expert panel consisting of 2 digital health experts, 1 VR specialist, and 2 patient education researchers was convened.

Furthermore, to ensure integration of health literacy theory with VR-specific technological expertise, the research team was

XSL•FO
RenderX

deliberately composed to intentionally include members who would bring complementary disciplinary perspectives. The core team comprised investigators with extensive backgrounds in VR technology development and clinical implementation (including 1 member with over 20 years of experience in immersive technology design for health care and education), as well as investigators with established expertise in health literacy research, including the development of nationally validated health literacy assessment instruments and empirical investigations of digital health literacy across diverse populations. This multidisciplinary composition functioned as a theoretical and methodological safeguard, informing all stages of the study methodology. By using a consensus-based process, this panel identified essential factors, proposed preliminary domains, and subsequently generated, prioritized, and refined the corresponding survey items. Each item was carefully reviewed to ensure its relevance, clarity, and alignment with the core aspects of VR health literacy before proceeding to the first Delphi round. Item generation was conducted collaboratively, with each preliminary item reviewed by both health literacy scholars and VR technology experts to balance theoretical grounding with medium-specific competencies. Throughout the Delphi process, panel feedback and emerging consensus were continuously evaluated by team members from both disciplinary backgrounds to mitigate potential disciplinary bias and maintain alignment between health literacy theory and VR implementation considerations.

Phase 3. Delphi Rounds

Two rounds of Delphi surveys were conducted to refine the definition and domains of VR health literacy and to assess the relevance of its constituent domains and subdomains.

In Round 1 (R1), 2 components were included to capture initial expert input and prioritize the necessity of proposed domains. First, participants provided open-ended definitions of VR health literacy, ensuring diverse perspectives for conceptualization. Items related to expert-identified needs for integrating VR in health care management were rated using an agreement scale ranging from “strongly disagree” to “strongly agree.” In contrast, items assessing VR health literacy domains and subdomains were evaluated using a necessity scale ranging from “not necessary” to “very necessary.” Second, participants were presented with a list of detailed domains related to VR health literacy and asked to rate their necessity on a 4-point Likert scale ranging from not necessary to very necessary.

In Round 2 (R2), the survey focused on achieving consensus on the refined framework. Participants reviewed the revised domains and subdomains from R1 and evaluated the necessity of each item using a 4-point Likert scale (“not necessary at all” to “very necessary”). Additionally, participants were invited to provide qualitative feedback to suggest further refinement. In this round, participants were informed that the evaluation would focus specifically on HMD-based immersive VR systems. This clarification was provided because several R1 responses referenced broader XR modalities. Narrowing the scope ensured conceptual consistency and alignment with the modality most commonly implemented in clinical health care settings.

Statistical Analysis

Descriptive statistics, including mean (SD), median (IQR), and agreement levels, were used to analyze the quantitative data from each Delphi round. Agreement was defined as the proportion of respondents who rated a statement as 3 or 4 on a 4-point Likert scale [18]. For the qualitative data, thematic analysis was applied to the open-ended responses to extract common themes related to the definition and components of VR health literacy.

The criteria for consensus were defined as follows: “high consensus” required an IQR ≤1 and agreement percentage ≥80%; “moderate consensus” required an IQR ≤1 and agreement percentage between 50% and 79%; and “low consensus” was identified by an IQR >1 and agreement percentage <50%. The dataset was analyzed using R software (version 4.2.3; R Foundation for Statistical Computing). Consensus thresholds were applied as guiding criteria rather than automatic exclusion rules. Subdomains with moderate consensus were further evaluated through structured internal deliberation to determine theoretical necessity within the framework.

Ethical Considerations

This study was approved by the institutional review board of Samsung Medical Center (approval number SMC 2023-12-070-011). All participants provided informed consent prior to their involvement in the study. Participant privacy and confidentiality were strictly maintained throughout the study. All collected data were deidentified and stored securely, and only the research team had access to the study data. The study was conducted in accordance with applicable ethical requirements and relevant regulations concerning the protection of personal information, privacy, and human rights. Participants received compensation of KRW 50,000 (approximately US $37) per round for their participation in the 2-round Delphi process (total KRW 100,000, approximately US $73).

Results

A total of 15 experts participated in R1, with 13 continuing to R2, as summarized in Table 1. The panel included various health care professionals, including hospital researchers and professors (n=5), IT developers (n=5), and researchers from industry promotion foundations, public institutions, or government organizations (n=5). While some experts had relatively limited VR-related experience, averaging 1 year, the panel also included senior participants with up to 12 years of VR expertise and over 25 years of experience in health care research. This diversity provided a balanced perspective; however, we acknowledge the limited pool of dual-expertise professionals as a study limitation. Of the 15 participants in R1, 13 (87%) continued to R2, reflecting a high retention rate for Delphi studies.

XSL•FO
RenderX

Table . Descriptive characteristics of expert panel participants (N=15).

Number	Career	VR[a]-related experience (y)	Health care (y)	Education and research (y)	Health literacy experience (y)
1	Researcher	1	0	4	0
2	IT developer	1	0	0	0
3	Industry promotion foundation researcher	10	0	0	0
4	IT developer[b]	12	9	14	5
5	Hospital researcher	2	6	12	0
6	Hospital researcher	2	20	3	0
7	Industry promotion foundation researcher	3	0	12	0
8	IT developer	1	0	4	0
9	Hospital researcher	2	7	7	2
10	IT developer	1	0	0	0
11	Professor (doctor)[b]	5	25	24	5
12	Industry promotion foundation researcher	4	10	10	4
13	Research director	1	0	10	0
14	Hospital researcher	3	1	15	10
15	IT developer	1	2	0	0

[a]VR: virtual reality.

[b]Participated only in Round 1.

In R1, experts evaluated the need for VR applications in health care (Table 2). The highest consensus included VR-based content development research (mean 3.9, IQR 0, Q1-Q3 4-4, agreement 100%), simulations for patient and caregiver education (mean 3.9, IQR 0, Q1-Q3 4-4, agreement 100%), psychological therapy applications (mean 3.8, IQR 0, Q1-Q3 4-4, agreement 100%), and student or health care professional education (mean 3.8, IQR 0, Q1-Q3 4-4, agreement 100%). In contrast, areas related to direct medical care and treatment received lower scores, including diagnostic assistance (mean 2.7, IQR 1, Q1-Q3 2-3, agreement 60%), surgical support (mean 2.6, IQR 1, Q1-Q3 2-3, agreement 53.3%), prognosis monitoring and survivorship care (mean 2.8, IQR 1, Q1-Q3 2-3, agreement 64.3%), and symptom management, such as pain reduction (mean 2.9, IQR 2, Q1-Q3 2-4, agreement 66.7%).

XSL•FO
RenderX

Table . Expert-identified needs for integrating virtual reality in health care management and clinical practice[a].

Domains and subdomains	Mean (SD)	Median	IQR (Q1-Q3)	Agreement (%)
Medical care				
Diagnostic assistance	2.7 (0.9)	3	1 (2-3)	60
Treatment				
Surgical support	2.6 (1.0)	3	1 (2-3)	53.3
Rehabilitation therapy	3.5 (0.5)	4	1 (3-4)	100
Psychological therapy (eg, trauma)	3.8 (0.4)	4	0 (4-4)	100
Symptom management (eg, pain reduction)	2.9 (1.1)	3	2 (2-4)	66.7
Cognitive training (eg, dementia prevention)	3.6 (0.5)	4	1 (3-4)	100
Prognosis monitoring and survivorship care (eg, cancer survivors)	2.8 (0.9)	3	1 (2-3)	64.3
Research				
VR-based[b] content development research	3.9 (0.3)	4	0 (4-4)	100
Intervention efficacy studies	3.6 (0.5)	4	1 (3-4)	100
Patient or caregiver education				
Simulation (eg, emergency response)	3.9 (0.3)	4	0 (4-4)	100
Spatial orientation and navigation guidance	3.3 (0.7)	3	1 (3-4)	86.7
General public education				
Health promotion and disease prevention	3.1 (0.8)	3	1.5 (2.5-4)	73.3
Behavioral simulation (eg, exercise methods, healthy eating)	3.5 (0.7)	4	1 (3-4)	92.9
Student or health care professional education				
Clinical simulation (eg, surgery, CPR[c])	3.8 (0.4)	4	0 (4-4)	100
Staff education				
Safety training simulation (eg, fire response)	3.7 (0.5)	4	1 (3-4)	100

[a]Ratings based on a 4-point agreement scale (1=strongly disagree; 4=strongly agree).

[b]VR: virtual reality.

[c]CPR: cardiopulmonary resuscitation.

As part of Delphi R1, experts were asked to provide open-ended definitions of VR health literacy. A diverse range of conceptual interpretations reflecting technical, cognitive, and practical dimensions of VR use in health contexts is summarized in Table S1 in Multimedia Appendix 2. Following iterative review and conceptual refinement of expert input, we defined VR health literacy as "the competency to acquire, understand, use, and manage health information and data through VR devices and services, enabling individuals to effectively engage with health-related content, promote health behaviors, and support health management." This definition reflected a deliberate conceptual positioning. While VR-specific technological characteristics were considered during refinement, the framework ultimately positioned VR health literacy as health literacy applied within immersive VR contexts, with VR-specific competencies operationalized through the identified domains. Further empirical work may inform refinement of VR-specific competencies in future iterations of the framework.

Table 3 summarizes the mean (SD), median (IQR), and agreement percentages across 2 Delphi rounds, highlighting the

XSL•FO
RenderX

iterative refinement of subdomains. Through iterative refinement, the initial 7-domain candidate structure entering R1 was refined through 2 rounds of consensus into a final 5-domain framework: (1) performance expectancy, (2) effort expectancy, (3) social influence, (4) facilitating conditions, (5) attitudes toward VR, (6) VR information literacy, and (7) behavioral intention, with 23 subdomains initially proposed during R1. Following R1, 9 subdomains were removed. Subdomains with moderate consensus or lower were excluded after internal deliberation, ensuring that the final set of subdomains was relevant and robust. Despite achieving high consensus (mean 3.4, IQR 0.6, agreement 93.3%) in R1, the "anticipated participation in VR-based health services" subdomain was excluded after internal deliberation. Experts noted that the absence of an established VR-based health service infrastructure and the limited availability of such programs undermined the current feasibility of this subdomain. In contrast, 3 new subdomains, including "expectations of future VR benefits," "intention to use VR technology/services," and "willingness to engage with VR for health," were added, resulting in 17 subdomains being evaluated in R2. The 3 newly added subdomains were developed based on feedback from the expert panel, internal discussions, and additional literature reviews. However, the "understanding of VR-related terms" subdomain (mean 3.8, IQR 0.5, agreement 73.3%), despite achieving only moderate consensus, was retained based on its perceived importance as determined during expert panel discussions.

In R2, experts demonstrated high levels of consensus across multiple subdomains, particularly those related to immersive experiences, such as "perceived immersion/embodiment," user-system interaction, such as "perceived interactivity and responsiveness," and "digital self-efficacy," each reaching 100% agreement (Table 3). The inclusion of 2 behavioral intention subdomains, "intention to use VR" and "willingness to engage with VR technology/services," was also validated with strong support, reinforcing the importance of motivational factors. Finally, we included these 3 moderated consensus subdomains, including "expectations of future VR benefits," "understanding of VR-related terms," and "digital knowledge and confidence," in the final framework based on internal deliberations regarding their importance and relevance to VR health literacy.

XSL•FO
RenderX

Table . Descriptive statistics and agreement levels for each subdomain in Delphi Rounds 1 and 2[a].

Domains	Round 1				Decision after Round 1	Round 2			
	Mean (SD)	Median	IQR[b] (Q1-Q3)	Agreement (%)		Mean (SD)	Median	IQR (Q1-Q3)	Agreement (%)
Performance expectancy									
Perceived usefulness of VR[c] for health management	3.3 (0.6)	3	1 (3-4)	93.3	No change	3.5 (0.9)	4	1 (3-4)	92.3
Anticipated participation in VR-based health services	3.4 (0.6)	3	1 (3-4)	93.3	Deleted	—[d]	—	—	—
Expectations of future VR benefits	—	—	—	—	Added	2.9 (0.8)	3	1 (2-3)	69.2
Effort expectancy									
Perceived immersion or embodiment	3.1 (0.8)	3	1 (3-4)	86.7	No change	3.7 (0.5)	4	1 (3-4)	100
Perceived interactivity and responsiveness	3 (0.8)	3	0.5 (3-3.5)	80	No change	3.5 (0.5)	4	1 (3-4)	100
Understanding of VR-related terms	3.8 (0.6)	3	0.5 (2.5-3)	73.3	No change	2.5 (1)	3	1 (2-3)	53.8
Social influence									
Support from peers or caregivers	2.6 (0.7)	3	1 (2-3)	60	Deleted	—	—	—	—
Recommendations or encouragement from others	2.5 (0.6)	2	1 (2-3)	40	Deleted	—	—	—	—
Health care professional endorsement	3.1 (0.8)	3	1.5 (2.5-4)	73.3	Deleted	—	—	—	—
Facilitating conditions									
Access to VR devices and platforms	3.1 (0.7)	3	1 (3-4)	80	No change	3.2 (0.4)	3	0.2 (3–3.2)	100
Digital knowledge and confidence	3.1 (0.7)	3	0.5 (3-3.5)	80	No change	2.9 (0.6)	3	0 (3-3)	76.9

XSL•FO
RenderX

Domains	Round 1				Decision after Round 1	Round 2			
	Mean (SD)	Median	IQR[b] (Q1-Q3)	Agreement (%)		Mean (SD)	Median	IQR (Q1-Q3)	Agreement (%)
Technical proficiency with VR devices	3 (0.7)	3	0 (3-3)	80	No change	3.3 (0.8)	3	1 (3-4)	84.6
Prior experience with VR (direct/indirect)	2.5 (0.7)	2	1 (2-3)	40	Deleted	—	—	—	—
Digital self-efficacy	3.1 (0.6)	3	0.5 (3-3.5)	86.7	No change	3.5 (0.5)	4	1 (3-4)	100
Attitudes toward VR									
Awareness of VR in health contexts	3.2 (0.6)	3	0.5 (3-3.5)	93.3	No change	3.4 (0.7)	3	1 (3-4)	92.3
Interest in VR technology	3.2 (0.7)	3	1 (3-4)	86.7	No change	3.1 (0.5)	3	0 (3-3)	92.3
VR information literacy									
Ability to acquire and process VR-based health info	2.5 (0.5)	2	1 (2-3)	46.7	Deleted	—	—	—	—
Evaluation of reliability, relevance, and technology	2.7 (0.7)	3	1 (2-3)	53.3	Deleted	—	—	—	—
Critical thinking about VR content and tools	2.7 (0.7)	3	1 (2-3)	53.3	Deleted	—	—	—	—
Problem-solving ability using VR content	3.1 (0.7)	3	0.5 (3-3.5)	80	No change	3.2 (0.9)	3	1 (3-4)	84.6
Behavioral intention									
Intention to use VR technology or services	—	—	—	—	Added	3.5 (0.9)	4	1 (3-4)	92.3
Willingness to engage with VR for health	—	—	—	—	Added	3.8 (0.4)	4	0 (4-4)	100

XSL•FO
RenderX

Domains	Round 1				Decision after Round 1	Round 2			
	Mean (SD)	Median	IQR[b] (Q1-Q3)	Agreement (%)		Mean (SD)	Median	IQR (Q1-Q3)	Agreement (%)
Sharing VR experiences with others	2.8 (0.8)	3	1 (2-3)	60	Deleted	—	—	—	—

[a]Seven candidate domains entered Round 1; the final framework comprises 5 domains and 14 subdomains. “Social influence” and “VR information literacy” did not reach consensus and were dissolved. “Problem-solving ability using VR content” was reassigned to “Attitudes toward VR.”

[b]The IQR is calculated by subtracting the first quartile (Q1) from the third quartile (Q3). "High consensus" required an IQR ≤1, agreement percentage ≥80%; "moderate consensus" required an IQR ≤1, agreement percentage between 50% and 79%; and "low consensus" was identified by an IQR >1, agreement percentage. Ratings based on a 4-point agreement scale (1=not necessary; 4=very necessary).

[c]VR: virtual reality.

[d]Not applicable.

In the refinement process, several subdomains were deleted, some newly added, and others reclassified (Table 4). Notably, the subdomain “problem-solving ability using VR content,” initially categorized under VR information literacy, was integrated into the domain “attitudes toward VR.” Experts noted that problem-solving reflects cognitive engagement and motivational attitudes toward VR use rather than technical literacy alone. Based on this iterative process, the final VR health literacy framework comprised 5 overarching domains and 14 subdomains, each representing essential competencies for engaging with VR in health contexts (Table 3). “Performance expectancy” included “perceived usefulness of VR for health management” and “expectations of future benefits” from its integration into health care. “Effort expectancy” encompassed perceptions of immersion, interactivity, and the ability to understand VR-related terminology. “Facilitating conditions” captured the practical enablers of VR use, such as access to devices and platforms, digital knowledge and confidence, technical proficiency, and digital self-efficacy. “Attitudes toward VR” reflected individuals’ awareness of VR applications in health and interest in VR technology, including their problem-solving capabilities. Lastly, “behavioral intention” was represented by individuals’ intention to use and willingness to engage in VR-based health interventions.

XSL•FO
RenderX

Table . Definitions of finalized subdomains of virtual reality health literacy.

Domains and subdomains	Definitions
Performance expectancy	
Perceived usefulness of VR[a] for health management	The individual's belief that VR technologies and services can generate meaningful value and improved outcomes in the context of health management
Expectations of future VR benefits	The degree to which individuals anticipate positive future outcomes or health-related advantages from the continued advancement and integration of VR technologies
Effort expectancy	
Perceived immersion or embodiment	The subjective sense of being physically and psychologically present in the virtual environment while using VR devices for health-related purposes
Perceived interactivity and responsiveness	The extent to which users perceive VR systems and content as interactive, responsive, and capable of enabling real-time engagement in a health context
Understanding of VR-related terms	The ability to comprehend basic terminology, concepts, and language necessary to navigate and use VR platforms and services effectively
Facilitating conditions	
Access to VR devices and platforms	The individual's physical and digital accessibility to VR equipment, applications, and platforms applicable to health monitoring, education, or intervention
Digital knowledge and confidence	The foundational knowledge of digital technologies and the confidence to navigate digital tools, including those specific to VR, in health-related settings
Technical proficiency with VR devices	The practical ability to operate essential functions of VR devices and applications required for effective participation in digital health experiences
Digital self-efficacy	An individual's belief in their own capacity to successfully engage with and perform tasks related to digital health, including VR-mediated interventions
Attitudes toward VR	
Awareness of VR in health contexts	Recognition of the presence, relevance, and potential applications of VR technology within medical, therapeutic, or public health settings
Interest in VR technology	The cognitive and emotional inclination to explore, learn about, and engage with VR technologies, especially as they pertain to health and well-being
Problem-solving ability using VR content	The capability to address or manage health-related challenges through the strategic use of VR content, tools, or simulations
Behavioral intention	
Intention to use VR technology or services	The expressed willingness and planned behavior to engage with VR-based tools and services as part of ongoing or future health practices
Willingness to engage with VR for health	The readiness to participate in VR-assisted health interventions or programs, including openness to adoption and behavioral engagement

[a]VR: virtual reality.

Discussion

We conducted a modified Delphi study to establish the scope and components of VR health literacy, with the aim of developing a comprehensive framework and reaching consensus on its definition and essential subdomains. In this study, we defined VR health literacy as the competency to acquire, understand, use, and manage health information and data through VR devices and services, enabling individuals to engage with health-related content, promote health behaviors, and support health management. This definition builds on the integrated health literacy framework while extending it with VR-specific competencies such as immersion, interactivity, and technical proficiency unique to virtual environments. This study identified 5 core domains of VR health literacy, including (1) performance expectancy, (2) effort expectancy, (3) facilitating conditions, (4) attitudes toward VR, and (5) behavioral intention, as well as 14 subdomains.

Aligned with prior research, our expert panel recognized education, training, and mental health interventions as pivotal areas for VR use, underscoring the technology's transformative potential in these domains [21,22]. However, in practice, the implementation of VR in health care is largely limited to basic

XSL•FO
RenderX

applications, such as scenic distractions or the adaptation of pre-existing interventions [23,24]. While the adoption of VR in health care is steadily expanding, current applications are limited to basic functions, such as visualization and passive content delivery. Consequently, the broader potential of VR to support preventive health and behavioral improvement has not been fully realized in practice. This gap between technological potential and applied use highlights the need for a systematic framework to assess user competencies. For health care professionals who design and implement VR interventions, such insights are critical to ensuring that strategies are tailored to users' cognitive, emotional, and contextual capacities, thereby maximizing both clinical effectiveness and educational impact.

In this study, we defined VR health literacy as an extension of digital health literacy that incorporates competencies unique to virtual environments, including immersion, interactivity, and technical proficiency. Early discussions of VR health literacy emphasized that VR is not only a technology but also a medium with its own emerging language and symbolic system [25]. Within this framework, VR health literacy includes the capacity to interpret and apply interaction cues such as gestures, avatars, and spatial representations in health care contexts. It is shaped by both users and content developers, underscoring its relevance for patients, clinicians, and educators. As the language of VR remains at an early stage, VR health literacy should be understood as a dynamic construct evolving alongside the expansion of immersive technologies in health care.

The Delphi results showed variation in consensus across domains. Subdomains related to performance and effort expectancy, particularly perceived usefulness, reached consistently high agreement. This finding is consistent with prior studies indicating that users value VR mainly for its immersive features and outcome-oriented benefits in health contexts [26]. Our Delphi findings underscore the fact that immersion and interactivity are central elements of VR health literacy, aligning with qualitative evidence [21] showing that these features enhance learner motivation and engagement. Notably, domains related to social influence were excluded. Although prior studies highlight the importance of peer and clinician recommendations for technology adoption [21], our Delphi panel gave these items lower priority. This may reflect the early stage of VR adoption in health care, where individual technical competencies are viewed as more urgent prerequisites than external influences. Future studies with broader public exposure to VR may yield different results. In addition, this contrast suggests that while assessment tools may appropriately focus on individual competencies, successful implementation in clinical practice will also depend on social and contextual drivers. Our results also underscore behavioral intention as a key determinant within VR health literacy. Consistent with prior research [27] in telerehabilitation, behavioral intention emerged as a key factor influencing adoption. These findings suggest that the successful integration of VR in health care relies not only on technical feasibility but also, importantly, on willingness to engage with and sustain the use of the technology. In contrast, despite acknowledging factors such as prior experience with VR, evaluation of reliability and relevance, critical thinking about VR content and tools, and sharing VR experiences as potentially relevant to VR health literacy, these subdomains were ultimately excluded due to low consensus within our Delphi panel. Although VR technologies are advancing rapidly, widespread engagement among the general public is still limited [28], with most exposure occurring through gaming rather than health-related applications. The exclusion of these domains from our Delphi consensus likely reflects this contextual backdrop of restricted exposure rather than conceptual irrelevance, indicating that their relevance should be revisited in future iterations of the framework as VR adoption expands beyond entertainment into health-related contexts.

In this study, expectancy- and intention-related constructs were included within the framework of VR health literacy to reflect the engagement-dependent nature of immersive VR environments. Contemporary developments in digital and eHealth literacy increasingly recognize that the enactment of literacy skills in technology-mediated contexts depends not only on operational ability but also on perceived capability, confidence, and readiness to engage [20,29]. HMD-based immersive VR requires active participation, spatial navigation, and sustained interaction for users to access health information. In such environments, literacy extends beyond the possession of technical skills to include the capacity to mobilize those skills within embodied and action-oriented contexts. Accordingly, performance expectancy, effort expectancy, and behavioral intention were conceptualized as engagement-enabling components that influence whether literacy-related competencies can be meaningfully enacted, rather than as direct substitutes for core cognitive literacy skills. Further field-based empirical validation is necessary to assess construct validity and clarify the functional interplay between motivational readiness and operational competencies in applied VR environments.

The 5 domains identified in this study constitute an initial operational framework of core competencies for the effective use of VR in health care. The framework was developed through a structured process integrating insights from health care professionals, IT developers, and public sector representatives, thereby reflecting clinical, technical, and policy perspectives. This multidisciplinary composition allowed diverse priorities to be reconciled, resulting in a model that combines theoretical clarity with practical applicability. The value of this framework extends beyond its theoretical structure. Developed through a multidisciplinary consensus process, it provides a practical tool for clinicians, educators, and policymakers to support the integration of VR into health care. The identified competencies may inform patient education curricula, guide assessment of readiness for VR-based interventions, and shape policy standards for safe and equitable implementation. VR health literacy was conceptualized not as proficiency confined within immersive environments, but as the capacity to engage with, interpret, and apply health information delivered through VR technologies across contexts. Although the framework does not include a standalone "transfer of knowledge" domain, the translation of VR-mediated information into real-world health understanding, decision-making, and communication is treated as a core functional outcome of literacy enactment. In health care contexts, immersive understanding is only meaningful as long as it informs subsequent health decisions, discussions with

XSL•FO
RenderX

providers, and behavioral implementation beyond the virtual setting. This conceptualization positions VR health literacy as an applied construct oriented toward real-world health action, rather than as a measure of in-system proficiency alone. As such, the framework offers a foundation for translating VR health literacy from concept to practice [27,30].

In the present Delphi process, subdomains related to critical appraisal and information evaluation were carefully considered during item generation and expert review. Although these competencies are central to established health literacy frameworks, they did not reach consensus for inclusion in the final framework. This outcome was interpreted not as a theoretical devaluation of appraisal, but as a reflection of the current clinical context of immersive VR use, in which health content is typically curated and implemented within supervised health care settings. In such environments, functional readiness, embodied interaction, and engagement capacities may operate as prerequisites for subsequent evaluative competencies. Accordingly, the framework adopts a staged conceptualization of VR health literacy, emphasizing access and enactment capacities in its initial iteration while recognizing that appraisal-related competencies may become increasingly salient as VR platforms evolve toward more open and user-driven information ecosystems.

Furthermore, the present framework focuses specifically on HMD-based immersive VR systems rather than the broader spectrum of XR modalities. This staged conceptualization also informs our treatment of immersion-related competencies. Although immersion is fundamentally a property of immersive media, effective engagement in VR-based health interventions requires users to cognitively and behaviorally adapt to embodied and spatially dynamic environments. In this framework, perceived immersion or embodiment does not refer to the intensity or quality of subjective experience, but rather to the individual's capacity to maintain spatial orientation, regulate embodied interaction, and integrate multisensory stimuli while processing health information. This distinction explicitly separates user experience quality from competency-related adaptation within immersive contexts, addressing the conceptual boundary between media properties and literacy-related capacities.

Regarding safety-related competencies, including recognition of cybersickness and maintenance of spatial awareness, we acknowledge their importance in immersive VR contexts. Although safety did not emerge as a distinct domain in the Delphi consensus, these elements were conceptually recognized during framework refinement and intentionally integrated across relevant domains given their cross-cutting nature. In interpreting these findings, we considered that many current health care VR applications are implemented in structured clinical environments with guided use and curated content, where safety-related awareness is often embedded within broader operational competencies. Accordingly, safety was positioned as an integral component of technical proficiency and facilitating conditions rather than defined as an independent construct in this initial framework. This positioning reflects a deliberate conceptual focus rather than an omission and may be revisited as VR applications expand into less supervised or nonclinical settings.

A key strength of this study is its innovative approach to establishing a foundational framework for the emerging field of VR health literacy. Nonetheless, this study has some limitations. First, the composition of the Delphi panel included variation in VR-specific experience levels among panel members, ranging from early-career professionals to senior experts, as well as a higher proportion of VR technology specialists relative to health literacy researchers. While this composition was intentionally designed to capture medium-specific competencies unique to immersive VR environments, differences in experience and disciplinary perspective may have influenced the prioritization of operational and engagement-related constructs. Future iterations of this framework would benefit from broader and more balanced panel representation, including patient advocates, health education researchers, and clinical practitioners, to further enhance conceptual balance and generalizability. In addition, as this study was conducted with experts based in Korea, cultural and systemic factors may have shaped the consensus. Future Delphi studies with multinational panels will be critical to validate and generalize the VR health literacy framework globally. Lastly, as VR technology continues to advance rapidly, the specific components of this framework may require updating over time. As a foundational study in this nascent field, the framework will naturally require further validation and refinement as the concept of VR health literacy evolves and its application in health care settings expands.

In conclusion, this study establishes an initial, consensus-based framework for VR health literacy. As VR technology approaches widespread clinical integration, this framework addresses the urgent need to enhance patient-centered competencies and align technological capabilities with clinical needs. It serves as both a theoretical model and a practical roadmap for researchers, clinicians, and developers to ensure that VR is implemented equitably and effectively. As immersive VR continues to evolve from supervised clinical implementation toward more autonomous and home-based use, domains that did not reach consensus in the present Delphi process—most notably critical appraisal of VR content and safety-related competencies such as recognition of cybersickness and maintenance of spatial awareness—are likely to gain increasing salience and warrant dedicated empirical investigation in future iterations of the framework. Ultimately, fostering VR health literacy is pivotal in unlocking the full potential of VR to transform health care, and this study provides a foundational step in that critical journey.

Acknowledgments

Generative artificial intelligence tools (ChatGPT [GPT-4o; OpenAI]) were used solely for English language editing and grammar refinement. No artificial intelligence tools were used for data analysis, interpretation, or generation of scientific content.

XSL•FO
RenderX

Funding

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health and Welfare, Republic of Korea (grant RS-2023-KH135171).

Data Availability

The data generated in this study are available from the corresponding author upon reasonable request. These data are not publicly available because of privacy restrictions.

Authors' Contributions

Conceptualization: JY, ML, JC
Data curation: JY
Formal analysis: JY, ML
Funding acquisition: JC
Investigation: JC
Methodology: ML, JC, DK, JP, SJK, JH
Software: JY, ML
Visualization: JY
Writing – original draft: JY
Writing – review & editing: ML, JC, DK, JP, SJK, JH
All authors approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full Boolean search strategies.
[DOCX File, 15 KB - jmir_v28i1e85842_app1.docx]

Multimedia Appendix 2
Expert perspectives on definition of virtual reality (VR) health literacy.
[DOCX File, 18 KB - jmir_v28i1e85842_app2.docx]

References

1. Hamad A, Jia B. How virtual reality technology has changed our lives: an overview of the current and potential applications and limitations. Int J Environ Res Public Health 2022 Sep 8;19(18):11278. [doi: 10.3390/ijerph191811278] [Medline: 36141551]
2. Bohil C, Owen CB, Jeong EJ, Alicea B, Biocca FA. Virtual reality and presence. In: Eadie WF, editor. 21st Century Communication: A Reference Handbook: SAGE Publications; 2009:534-544. [doi: 10.4135/9781412964005.n59]
3. Slater M, Sanchez-Vives MV. Enhancing our lives with immersive virtual reality. Front Robot AI 2016 Dec;3. [doi: 10.3389/frobt.2016.00074]
4. Spiegel BMR, Rizzo A, Persky S, et al. What is medical extended reality? A taxonomy defining the current breadth and depth of an evolving field. J Med Ext Real 2024 Jan 1;1(1):4-12. [doi: 10.1089/jmxr.2023.0012] [Medline: 38505474]
5. Kouijzer MMTE, Kip H, Bouman YHA, Kelders SM. Implementation of virtual reality in healthcare: a scoping review on the implementation process of virtual reality in various healthcare settings. Implement Sci Commun 2023 Jun 16;4(1):67. [doi: 10.1186/s43058-023-00442-2] [Medline: 37328858]
6. Abbas JR, O'Connor A, Ganapathy E, et al. What is virtual reality? A healthcare-focused systematic review of definitions. Health Policy Technol 2023 Jun;12(2):100741. [doi: 10.1016/j.hlpt.2023.100741]
7. Tokgöz P, Stampa S, Wähnert D, Vordemvenne T, Dockweiler C. Virtual reality in the rehabilitation of patients with injuries and diseases of upper extremities. Healthcare (Basel) 2022 Jun 16;10(6):1124. [doi: 10.3390/healthcare10061124] [Medline: 35742176]
8. Skurla MD, Rahman AT, Salcone S, et al. Virtual reality and mental health in older adults: a systematic review. Int Psychogeriatr 2022 Feb;34(2):143-155. [doi: 10.1017/S104161022100017X] [Medline: 33757619]
9. Hajesmaeel Gohari S, Gozali E, Niakan Kalhori SR. Virtual reality applications for chronic conditions management: a review. Med J Islam Repub Iran 2019;33:67. [doi: 10.34171/mjiri.33.67] [Medline: 31456991]
10. Pandita S, Stevenson Won A. Chapter 7 - Clinical applications of virtual reality in patient-centered care. In: Kim J, Song H, editors. Technology and Health: Academic Press; 2020:129-148. [doi: 10.1016/B978-0-12-816958-2.00007-1]

XSL•FO
RenderX

11. Gul MI, Khan IA, Shah S, El-Affendi M. Can nonliterates interact as easily as literates with a virtual reality system? A usability evaluation of VR interaction modalities. Systems 2023;11(2):101. [doi: 10.3390/systems11020101]
12. Glegg SMN, Levac DE. Barriers, facilitators and interventions to support virtual reality implementation in rehabilitation: a scoping review. PM R 2018 Nov;10(11):1237-1251.e1. [doi: 10.1016/j.pmrj.2018.07.004] [Medline: 30503231]
13. Zhao W, Baker S, Waycott J. Challenges of deploying VR in aged care: a two-phase exploration study. Presented at: 32nd Australian Conference on Human-Computer Interaction (OzCHI '20); Dec 2-4, 2020; Sydney, NSW, Australia p. 87-98. [doi: 10.1145/3441000.3441018]
14. Zhang T, Booth R, Jean-Louis R, et al. A primer on usability assessment approaches for health-related applications of virtual reality. JMIR Serious Games 2020 Oct 28;8(4):e18153. [doi: 10.2196/18153] [Medline: 33112243]
15. van der Kruk SR, Zielinski R, MacDougall H, Hughes-Barton D, Gunn KM. Virtual reality as a patient education tool in healthcare: a scoping review. Patient Educ Couns 2022 Jul;105(7):1928-1942. [doi: 10.1016/j.pec.2022.02.005] [Medline: 35168856]
16. Mosadeghi S, Reid MW, Martinez B, Rosen BT, Spiegel BMR. Feasibility of an immersive virtual reality intervention for hospitalized patients: an observational cohort study. JMIR Ment Health 2016 Jun 27;3(2):e28. [doi: 10.2196/mental.5801] [Medline: 27349654]
17. Finkelstein J, Parvanova I, Huo X. Feasibility of a virtual reality app to promote pulmonary rehabilitation. Stud Health Technol Inform 2023 May 18;302:458-462. [doi: 10.3233/SHTI230172] [Medline: 37203716]
18. Nasa P, Jain R, Juneja D. Delphi methodology in healthcare research: how to decide its appropriateness. World J Methodol 2021 Jul 20;11(4):116-129. [doi: 10.5662/wjm.v11.i4.116] [Medline: 34322364]
19. Jünger S, Payne SA, Brine J, Radbruch L, Brearley SG. Guidance on Conducting and REporting Delphi Studies (CREDES) in palliative care: recommendations based on a methodological systematic review. Palliat Med 2017 Sep;31(8):684-706. [doi: 10.1177/0269216317690685] [Medline: 28190381]
20. van der Vaart R, Drossaert C. Development of the digital health literacy instrument: measuring a broad spectrum of Health 1.0 and Health 2.0 skills. J Med Internet Res 2017 Jan 24;19(1):e27. [doi: 10.2196/jmir.6709] [Medline: 28119275]
21. Skidmore N, Ryan CG, Mankelow J, Martin D. Acceptability and feasibility of virtual reality to promote health literacy in primary care from the health professional's view: a qualitative study. Patient Educ Couns 2024 Jun;123:108179. [doi: 10.1016/j.pec.2024.108179] [Medline: 38367303]
22. Shiner CT, Croker G, McGhee J, Faux SG. Perspectives on the use of virtual reality within a public hospital setting: surveying knowledge, attitudes, and perceived utility among health care professionals. BMC Digit Health 2024;2(1):18. [doi: 10.1186/s44247-024-00076-x]
23. Skidmore N, Ryan C, Mankelow J, Bradford C, Graham A, Martin D. Exploring the potential of virtual reality for the self-management of chronic pain: a scoping review of its use to address health literacy. Musculoskelet Sci Pract 2024 Aug;72:102962. [doi: 10.1016/j.msksp.2024.102962] [Medline: 38703701]
24. Dhar E, Upadhyay U, Huang Y, et al. A scoping review to assess the effects of virtual reality in medical education and clinical care. Digit Health 2023;9:20552076231158022. [doi: 10.1177/20552076231158022] [Medline: 36865772]
25. Sherman WR, Craig AB. Literacy in virtual reality. SIGGRAPH Comput Graph 1995 Nov;29(4):37-42. [doi: 10.1145/216876.216887]
26. Jacobs C, Foote G, Williams M. Evaluating user experience with immersive technology in simulation-based education: a modified Delphi study with qualitative analysis. PLoS One 2023;18(8):e0275766. [doi: 10.1371/journal.pone.0275766] [Medline: 37531361]
27. Chan YK, Tang YM, Teng L. A comparative analysis of digital health usage intentions towards the adoption of virtual reality in telerehabilitation. Int J Med Inform 2023 Jun;174:105042. [doi: 10.1016/j.ijmedinf.2023.105042] [Medline: 36965405]
28. Na JY. A study on factors affecting VR game usage trends and perceived satisfaction. J Digit Contents Soc 2024 Sep 30;25(9):2689-2698. [doi: 10.9728/dcs.2024.25.9.2689]
29. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9. [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
30. Wang EY, Qian D, Zhang L, et al. Acceptance of virtual reality in trainees using a technology acceptance model: survey study. JMIR Med Educ 2024 Dec 23;10:e60767. [doi: 10.2196/60767] [Medline: 39727193]

Abbreviations

HMD: head-mounted display
R1: Round 1
R2: Round 2
VR: virtual reality
XR: extended reality

XSL•FO
RenderX

Edited by A Stone; submitted 14.Oct.2025; peer-reviewed by JF Costa, T Basmaji; revised version received 30.Apr.2026; accepted 30.Apr.2026; published 24.Jun.2026.

Please cite as:
Yoon J, Lee M, Kim D, Park J, Kim SJ, Han J, Cho J
Development of Virtual Reality Health Literacy: Delphi Expert Consensus Study
J Med Internet Res 2026;28:e85842
URL: https://www.jmir.org/2026/1/e85842
doi:10.2196/85842

© Junghee Yoon, Mangyeong Lee, Dokyoon Kim, Joungwon Park, Su jin Kim, Jiyoon Han, Juhee Cho. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Virtual Exposure With Response Prevention for Obsessive-Compulsive Disorder: Randomized Controlled Trial

Lara Rolvien[1], PhD; Lena Jelinek[1], Prof Dr; Luzie Lohse[1], MSc; Steffen Moritz[1], Prof Dr; Lisa Borgmann[1], MSc; Jürgen Gallinat[1], Prof Dr; Simone Kühn[1,2], Prof Dr; Franziska Miegel[1], PD, PhD

[1]Department of Psychiatry and Psychotherapy, University Medical Center Hamburg-Eppendorf, Martinistr 52, Hamburg, Germany

[2]Center for Environmental Neuroscience, Max Planck Institute for Human Development, Berlin, Germany

Corresponding Author:
Franziska Miegel, PD, PhD
Department of Psychiatry and Psychotherapy, University Medical Center Hamburg-Eppendorf, Martinistr 52, Hamburg, Germany

Abstract

Background: Despite being highly effective and guideline-recommended for obsessive-compulsive disorder, exposure with response prevention is often underused due to barriers from both therapists and patients. Virtual reality may help address these barriers.

Objective: This study aimed to assess the efficacy and acceptance of virtual exposure with response prevention (VERP) in patients with contamination obsessive-compulsive disorder (conOCD) and checking obsessive-compulsive disorder (checkOCD) and to identify factors influencing treatment success.

Methods: Within a randomized controlled trial, 80 participants (40 with conOCD and 40 with checkOCD) were allocated to an intervention group, which received outpatient care as usual (CAU) supplemented by 6 weekly VERP sessions (VERP group), or to a control group receiving CAU alone (CAU group). Assessments occurred at baseline (t0), postintervention (t1), and at 3-month follow-up (t2). The reduction in obsessive-compulsive symptoms, as measured with the Yale-Brown Obsessive-Compulsive Scale, served as the primary outcome.

Results: Overall, obsessive-compulsive symptoms showed no greater improvement in the VERP group compared to the CAU group from t0 to t1 or t2 (η_p^2=0.003-0.014). In exploratory subsample analyses, we observed a trend toward an interaction effect, showing greater reduction in obsessions among participants with checkOCD from t0 to t1 ($F_{1,52}$=3.91; P=.05; η_p^2=0.070) and greater anxiety reduction among participants with conOCD from t0 to t2 ($F_{1,45}$=4.11; P=.05; η_p^2=0.084). Exploratory moderation analysis suggested that adding VERP to CAU was more effective for those in the VERP group who had not prematurely discontinued prior psychological treatment and those with checking behavior. Sense of presence was only moderate, unrelated to outcome, and patient satisfaction with VERP was positive overall.

Conclusions: While overall symptom reduction did not significantly differ between groups, exploratory analyses suggested trend-level differences in treatment response across obsessive-compulsive disorder subtypes and patient characteristics. These findings should be interpreted cautiously and are best viewed as hypothesis-generating.

Trial Registration: German Clinical Trials Register DRKS00016929; https://drks.de/search/en/trial/DRKS00016929/details

(J Med Internet Res 2026;28:e79326) doi:10.2196/79326

KEYWORDS

virtual reality; exposure with response prevention; obsessive-compulsive disorder; contamination; checking

Introduction

Background

Obsessive-compulsive disorder (OCD) is a common chronic and disabling condition characterized by recurrent intrusive, distress-provoking thoughts, images, ideas, and impulses (obsessions) as well as repetitive overt or covert rituals (compulsions) aimed at reducing distress and unpleasant emotions such as anxiety or disgust. Obsessions and compulsions can burden and disrupt daily life, which highlights the negative impact on patients' quality of life [1]. With lifetime prevalence estimates of OCD ranging from 1.3% to 2.5% in recent meta-analyses [2,3] and OCD's ranking as one of the 10 leading causes of disability worldwide due to its chronic nature [4], the need for appropriate treatment is evident.

Cognitive behavioral therapy (CBT) that includes exposure with response prevention (ERP) is the recommended first-line treatment for OCD [5-7]. In ERP, patients are encouraged to face their fears stemming from their obsessions (exposure) and

XSL•FO
RenderX

are taught to resist the urge to engage in compulsive behaviors (response prevention). Classic ERP can take various forms, with the most common being in vivo exposure in real-life scenarios. ERP in imagined scenarios is called in sensu, with the latter yielding lower long-term success rates [8]. The idea behind ERP is to help patients realize that the feared stimuli are not as threatening as perceived and that their distress, anxiety, or other unpleasant emotions naturally decrease (or can be managed) without the need for compulsive behaviors, ultimately leading to changes in maladaptive cognitive processes [5,9,10].

Despite its well-established effectiveness in treating OCD [11], ERP is not used to its full potential. Several patient- and therapist-related factors may be responsible for this [12]. For the therapist, ERP can be time-intensive and may pose practical challenges, including the need for longer or more frequent sessions and the coordination of in vivo exposures [13,14]. Moreover, therapists have reported that the exposure is stressful for them, they do not feel safe conducting it, they are afraid of patients' reactions, and they fear harming the patient [15]. According to Maltby and Tolin [16], even among therapists who identify themselves as adhering to a cognitive behavioral approach, the implementation of evidence-based therapies such as ERP remains insufficient, with use rates remaining below 50%. As for the patients, they often fear exposure, lack motivation, prefer self-help exposure, and/or perceive logistical challenges [14]. It has been estimated that around 25% of patients with OCD refuse ERP [16]. Together, the barriers on the patients' and the therapists' sides may result in what Moritz et al [14] refer to as "phobie à deux," hindering successful treatment and symptom improvement in the patient.

Exposure Therapy in Virtual Reality

Exposure therapy in virtual reality (VRET) is a promising new treatment strategy with the potential to overcome some of the described treatment barriers, particularly logistical challenges and time difficulties for therapists as well as patient fears. Virtual reality (VR) is a computer-generated simulation, enabling real-time interaction with the environment. It offers a nurturing and safe environment for patients, fostering increased self-efficacy [17]. Consequently, VRET is often considered a promising or advantageous alternative compared to in vivo or in sensu exposure in certain contexts, particularly due to its potential to increase patient acceptance and reduce refusal rates, rather than being universally superior [18,19]. Leveraging VR within psychotherapy offers significant benefits, including the logistical advantages of tailoring environments and scenarios, controlling the exposure process, and conducting therapy sessions in the therapist's room [20]. By using VR, therapists can tailor the virtual environment to match the specific fears of each patient, subsequently offering guidance and support while monitoring their progress within the virtual environment [21]. Moreover, the interactivity of VR allows for standardized, controlled, and replicable exposure environments, which can improve the therapeutic process, by ensuring consistent delivery of exposure tasks, reducing reliance on patient imagination or therapist improvisation, and facilitating the safe practice of challenging situations. These features may lower barriers to treatment access and enhance efficiency [17,18].

VRET for anxiety disorders has been well studied and has shown efficacy for specific phobias, social anxiety disorder, posttraumatic stress disorder, and panic disorder, as a meta-analysis of 30 studies on VRET revealed [22]. However, the success of VRET in this meta-analysis was similar to in vivo exposure. In more severe anxiety disorders and posttraumatic stress disorder, according to another meta-analysis [23], VRET's superiority to CBT with ERP was small and not significant.

Virtual Exposure With Response Prevention for OCD

The inherent challenges posed by the nature of OCD make the use of VR in patients with OCD more difficult; it is less researched and less disseminated in clinical practice despite its possible advantages for treatment. So far, it is unclear whether the treatment effects for anxiety disorders generalize to OCD, which is characterized by high idiosyncrasy and cognitive avoidance, especially during in sensu exposure [24]. Moreover, the stimuli in OCD are more heterogeneous than in most anxiety disorders, making it difficult to cover all comprehensively in one virtual environment. For patients, particularly those afflicted with contamination obsessive-compulsive disorder (conOCD), the use of a shared device, such as VR glasses, may pose considerable challenges due to concerns regarding potential contamination by previous users [25].

However, preliminary research has shown that virtual exposure with response prevention (VERP) can be a valuable tool in the treatment of OCD. Studies have found that VERP can effectively provoke anxiety, disgust, and checking in patients with OCD [26-28], and patients have reported a good sense of "presence" (ie, the feeling of being there in a virtual environment) and increased anxiety (standardized mean difference=2.92; 95% CI 1.89 - 3.94; $P<.001$; $I^2=95\%$) and disgust (standardized mean difference=2.52; 95% CI 1.36 - 3.68; $P<.001$; $I^2=95\%$) levels in response to VERP scenarios compared to healthy controls in a meta-analysis of 11 primary studies [27]. Furthermore, a trend correlation between anxiety level and emotional engagement ($r=0.87$; $P\geq.05$) as well as anxiety and sense of presence ($r=0.88$; $P\geq.05$) during VERP has been found in a primary study including 4 women with OCD [29]. Laforest et al [24] investigated VERP in 3 patients over the course of 12 weeks and found a significant reduction in obsessions and compulsions in all patients, as assessed by visually inspecting the graphs of the daily assessments and an autoregressive moving average analysis, although the effects were not maintained in 2 of the 3 patients after 12 months. In addition, VERP has been found to be effective in managing obsessive-compulsive (OC) symptoms, particularly when traditional ERP has been unsuccessful, as reported in a single case report [30]. A recent case series that included 8 patients with contamination-related OCD revealed that VERP was able to induce distress (subjectively measured, $d=1.11$) and arousal (assessed with Galvanic Skin Response: $d=4.84$ and heart rate: $d=2.38$) associated with disgust at large effect sizes. VERP also evoked a moderate sense of presence but found a low rate of symptom reduction as measured with the Yale-Brown Obsessive-Compulsive Scale (Y-BOCS) total scale ($t_7=2.30$; $P=.06$; $d=1.82$ [25]). In another study, 21 nonclinical participants with contamination fears either received

XSL•FO
RenderX

VERP or were assigned to a control group, resulting in reduced anxiety ($F_{1,19}$=11.75; P<.01; η_p^2=0.38), disgust ($F_{1,19}$=18.53; P<.001; η_p^2=0.49), and urge to wash hands ($F_{1,19}$=17.03; P<.01; η_p^2=0.47) in the experimental group [31]. In line with these findings, Javaherinenani et al [21] found that CBT with VERP significantly reduced the severity of OC symptoms in the intervention group undergoing a 12-week treatment compared to a conventional CBT treatment including ERP components (Y-BOCS total score: F=60.97; P<.001; η_p^2=0.82).

Although there are initial indications that VERP can be effectively used in the treatment of OCD, larger-scale research is needed to determine its full potential, especially when compared to care as usual (CAU; defined to allow all outpatient treatments to be used, such as CBT with ERP). A recent systematic review evaluating the evidence for the use of VR in both the diagnosis and treatment of mental disorders found that the smallest number of studies focused on OCD [32]. This review is not the only one to conclude that the research gap in VERP for OCD is marked and that further studies are urgently needed. Beyond establishing efficacy, it is also important to evaluate patients' satisfaction with the treatment [25] as well as their sense of presence—a factor known to elicit anxiety in VRET and thus one that may influence treatment effectiveness [29,33].

In this study, we focused specifically on the symptom dimensions of conOCD and checking obsessive-compulsive disorder (checkOCD). These 2 domains are among the most prevalent and clinically relevant manifestations of OCD, consistently identified in factor-analytic studies and frequently associated with substantial distress and impairment [34-36]. They also represent prototypical forms of compulsive behavior that are well-suited to examine the feasibility and potential effects of novel interventions such as VERP. Importantly, although many individuals with OCD experience symptoms across multiple domains, classification into patients with conOCD or patients with checkOCD in our trial was based on the predominant symptom dimension, as determined through clinical interview and supported by self-report. This approach reflects the conceptual and empirical relevance of symptom dimensions while acknowledging the dimensional nature of OCD presentations.

Study Objectives

The primary objective of this study was to examine the incremental benefit of adding VERP to CAU in patients with OCD, compared to CAU alone. Specifically, we aimed to test whether VERP plus CAU would lead to greater reductions in overall OCD severity (measured by the Y-BOCS) than CAU alone. In addition, we investigated potential differences between conOCD and checkOCD, hypothesizing that patients with conOCD would benefit more strongly from VERP than patients with checkOCD due to the higher ecological validity of contamination-related VR scenarios. Finally, we explored patient satisfaction with and sense of presence in VERP and examined the potential relationship between these factors and efficacy.

Methods

Design

In the framework of a randomized controlled trial (see Checklist 1 for the CONSORT list[Consolidated Standards of Reporting Trials]), patients with conOCD and patients with checkOCD were randomly assigned to either the VERP or the CAU group after baseline assessment (t0). After 6 weeks, all patients were invited to a second assessment (post assessment, t1). After 3 more months, an online survey and a telephone interview were conducted (follow-up, t2). The principal investigators (LR and FM) carried out the randomization, which was based on a randomization plan (1:1 allocation rule) that was conducted by an online randomization program. Additionally, randomization was stratified by type of compulsions (contamination or checking), with half of the patients with conOCD being allocated to the intervention group and half to the CAU group, and the same procedure applied to patients with checkOCD. Patients were informed about the group assignment by a sequentially numbered envelope they received after t0. Before leaving the room after the t0 assessment, the raters, who were blind to the patients' group assignments, reminded the patients not to reveal their group assignment at the following assessments. Additionally, patients in the intervention group as well as the CAU group were permitted to continue their routine medical care (such as CBT or psychiatric medication), which was carefully recorded.

Ethical Considerations

The study was conducted in accordance with the Declaration of Helsinki, received ethics approval (local ethics committee of the clinic: LPEK-0020), and was preregistered with the German Clinical Trials Register (DRKS00016929). All participants provided written informed consent prior to participation. Informed consent procedures also included permission for the use of anonymized study data for analysis and publication. No additional consent was required for the present analyses. All data were pseudonymized at the point of collection and stored on secure servers in accordance with data protection regulations. No personal identifiers were retained, and confidentiality of participants' information was ensured throughout the study. No financial compensation was offered for study participation, but, as an incentive, after the last assessment, patients in the CAU group received a self-help manual for OCD—My Metacognitive Training [37]. No identifying images or other personal details of study participants are included in this paper.

Participants and Procedure

The recruitment strategies included a Google AdWords campaign, promotion in related forums and posts on websites, posters, describing the study to the anxiety outpatient ward of the clinic and local therapists, as well as getting in touch with previous study participants who had given permission to be contacted again for future studies. A telephone interview conducted by the principal investigators (LR, FM, and LL) screened interested individuals for eligibility. In this screening, interested individuals were asked whether they predominantly experienced washing and/or checking compulsions. This

XSL•FO
RenderX

inclusion criterion was reassessed during the baseline interview (t0) using the Y-BOCS symptom checklist to ensure accurate identification of each participant's predominant symptom dimension. Participants who reported both types of compulsions were further asked to identify which behavior was more dominant, distressing, and restrictive. Based on this self-assessment, participants were classified as either having conOCD or checkOCD. It is important to note that overlap of compulsive behaviors is common, and the classification reflects the symptom domain most prominently impacting the individual, while other OCD symptoms may also be present to a lesser extent. Inclusion and exclusion criteria are presented in Textbox 1. If eligible, patients were invited to an in-person t0 assessment and a t1 assessment 6 weeks later that were conducted by trained interviewers, and they received a link to an online assessment containing the questionnaires (secondary outcome measures). After the t0 assessment, patients in the intervention group participated in VERP sessions for 6 weeks, with weekly sessions lasting 60 to 90 minutes each. The t2 assessment 3 months after t1 was carried out online and via telephone.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Predominant presence of contamination obsessive-compulsive disorder or checking obsessive-compulsive disorder
- Provision of written informed consent
- Age between 18 and 75 years
- Sufficient command of the German language
- Willingness to participate in virtual exposure with response prevention for 6 weeks

Exclusion criteria

- Lifetime symptoms of psychosis or mania
- Severe neurological disorder
- Acute suicidality

Treatment

All 6 treatment sessions (1 session per week) were highly structured and used a treatment manual (based on a CBT manual for OCD by Oelkers and Hautzinger [38]). The first 2 sessions prepared the patient for the VERP sessions (symptom anamnesis, psychoeducation on ERP, presentation of a video of the VR environment, and rating of the OC-triggering stimuli to design the individual ERP). In the third session, the patient was familiarized with the VR equipment through testing the VR devices in a neutral VR setting (ie, a neutral, unfurnished virtual room). After that, the therapist guided the patient through the actual VERP session (eg, by encouraging them to look at and engage with the stimuli). In sessions 4 to 6, further VERPs were conducted, starting at medium intensity in session 3 and increasing the intensity throughout the following sessions. At the beginning of sessions 3 to 6, expectations on what would happen during VERP were identified beforehand and discussed afterward (eg, by noting that the feared consequence, such as fainting, had not occurred, representing an expectancy violation [39]). To minimize the occurrence of mental checking during VERP sessions, patients were explicitly informed about the concept of "mental checking" and instructed to refrain from engaging in it. Therapists provided detailed guidance on how to behave following each exposure session, including instructions not to wash hands, use disinfectants, or change clothes immediately. These measures were implemented to ensure response prevention and to reduce the likelihood that participants classified as having checkOCD would engage in covert checking or that patients with conOCD would perform their compulsive rituals immediately after the session. Therapists were psychotherapists in CBT training or licensed psychotherapists supervised by an experienced CBT psychotherapist (LJ).

An HTC Vive Pro headset was used to conduct the VERP. The VR environments were conducted by Unity (version 2018.2.9f1). In the virtual environment, patients moved by using controllers that teleported (StreamVR Teleport) them to the intended location. During VERP, therapists were able to guide patients because they could see on a computer screen what the patient was seeing. Two different virtual environments were available: 1 public restroom addressing conOCD (Figure 1A) and 1 apartment with various technical and other checking-related devices addressing checkOCD (Figure 1B). The public restroom included, among other things, a dirty sink, urine, feces, blood, and a wet floor. In this environment, patients were also instructed to interact directly with the stimuli, for example, by touching the "dirty" floor or sitting on a "dirty toilet" (represented by a regular chair in the real room). Participants were instructed not to wash or disinfect their hands, body, or clothes afterward, ensuring response prevention. In the apartment, participants were confronted with various electrical devices, candles, and water taps. To ensure response prevention in this environment, participants here were explicitly instructed to switch on relevant devices (eg, stove or microwave and water tap) at the beginning of the exercise, then switch them off only once, and immediately leave the room and subsequently the apartment. The front door could be closed with a key, and repeated checking was not permitted.

XSL•FO
RenderX

Figure 1. (A) Excerpt of the VR environment for patients with washing compulsions. (B). Excerpt of the VR environment for patients with checking compulsions. VR: virtual reality.

A

B

Control Condition

In this study, a CAU control group was included. CAU was defined so that all outpatient treatments could be used, including CBT with ERP. It was assessed whether participants had received outpatient psychotherapy in general during the intervention period, and additionally, whether they had received guided ERP in vivo or unguided ERP in vivo.

Measures

At baseline (t0), sociodemographic and clinical variables were assessed through a structured interview. The following information was collected: gender (women or men), age (in years), and qualification for university study. In addition, treatment-related variables were recorded, including current medication (antidepressant, combination, other, and none), psychotherapy experience (current, in the past, current and in the past, or discontinued prematurely), illness duration (in years), and age at onset (in years). No monitoring, management, or reassessment of medication occurred during the study, as all participants were outpatients. All measures used in the study are summarized in Table 1.

XSL•FO
RenderX

Table . Measures used in the study.

Title	Abbreviation	Content and psychometrics	Assessment times
Primary outcome			
Yale-Brown Obsessive-Compulsive Scale	Y-BOCS	The Y-BOCS [40,41] is a half-structured interview assessing OC[a] symptoms by a symptom checklist and a severity rating (including 19 items). Psychometric properties of the German version have been shown to be good to very good [42].	t0, t1, and t2
Secondary outcomes			
Obsessive-Compulsive Inventory—Revised	OCI-R	The OCI-R [43,44] is a self-rated questionnaire capturing OC symptoms. The German version of the OCI-R showed good psychometric properties [44].	t0, t1, and t2
Beck Depression Inventory-II	BDI-II	The BDI-II [45] is a widely used 21-item self-report questionnaire measuring the severity of depression symptomatology in adolescents and adults. The German version of the BDI-II has been shown to have good reliability and validity in both clinical and nonclinical populations [46].	t0, t1, and t2
Generalized Anxiety Disorder Scale-7	GAD-7	The GAD-7 [47] is a 7-item questionnaire that aims to detect the severity of clinical symptoms of generalized anxiety disorder. The scale has shown good psychometric properties in a German sample and can therefore be considered a reliable scale [48].	t0, t1, and t2
Obsessive Beliefs Questionnaire-44	OBQ-44	The OBQ-44 [49,50] evaluates belief domains pertinent to OCD[b], including control of thoughts, importance of thoughts, responsibility, intolerance of uncertainty, overestimation of threat, and perfectionism. The questionnaire has shown very good psychometric properties [51].	t0, t1, and t2
World Health Organization Quality of Life—BREF	WHOQOL-BREF	Quality of life was assessed with the global item of the WHOQOL-BREF, which served as a brief indicator of participants' overall quality of life while minimizing assessment burden [52].	t0, t1, and t2
Subjective Appraisal Rating	__[c]	The questionnaire capturing the subjective appraisal of the VERP[d] (only by patients in the VERP group) was based on ones used by Jelinek et al [53,54] and was adapted to fit the assessment of VERP. One open question asks patients for their appraisal of the VERP and leaves space for a personal comment.	t1

XSL•FO
RenderX

Title	Abbreviation	Content and psychometrics	Assessment times
Igroup Presence Questionnaire	IPQ	The IPQ [55] records the degree patients feel being (or feeling present) in the virtual environment. The scale consists of three subscales assessing (1) spatial presence, (2) involvement, and (3) experienced realism. One additional item assesses the general sense of presence. Psychometric properties of the IPQ have been reported to be good [55].	t1
Mini-International Neuropsychiatric Interview	MINI	Psychiatric diagnoses were assessed using the MINI [56], a structured diagnostic interview designed to identify major axis I psychiatric disorders according to DSM-IV[e] and ICD-10[f] criteria. The MINI was administered by a trained interviewer to confirm eligibility and to document comorbid psychiatric conditions.	t0

[a]OC: obsessive-compulsive.

[b]OCD: obsessive-compulsive disorder.

[c]Not available.

[d]VERP: virtual exposure with response prevention.

[e]DSM-IV: Diagnostic and Statistical Manual of Mental Disorders IV.

[f]ICD-10: International Classification of Diseases, 10th Revision.

Power Analysis

As Carl et al [22] reported medium to large effect sizes in their meta-analysis on VRET in anxiety disorders, we expected similar effect sizes for this study. Thus, to detect an effect size of d=0.80 (α=.05) with a power of 0.80, a sample size of 70 participants was needed according to G*Power [57]. ERP often has a high dropout rate [58], but we expected it to be lower in this study due to the less frightening nature of the operationalization by VR; thus, we expected a dropout rate of 15%. This led us to recruit a total sample of 80 participants.

Strategy of Data Analysis

A complete case (CC) analysis (patients who provided data at all assessments) and an intention-to-treat (ITT) analysis (patients who provided only t0 data) were conducted. Analyses of covariance (ANCOVAs) were conducted with group (VERP group vs CAU group) as the between-subject factor, the t0 scores of the respective outcome as the covariates, and change scores as the dependent variables. Change scores were defined as the difference between baseline and posttreatment (t0–t1) for the Y-BOCS, Obsessive-Compulsive Inventory—Revised (OCI-R), Generalized Anxiety Disorder Scale-7 (GAD-7), Beck Depression Inventory-II (BDI-II), Obsessive Beliefs Questionnaire-44 (OBQ-44), and World Health Organization Quality of Life—BREF (WHOQOL-BREF), and as the difference between baseline and follow-up (t0–t2) for the same measures. Exploratory analyses included the Y-BOCS subscales (obsessions and compulsions). These analyses are reported as exploratory due to the high interrelatedness of the outcomes and the risk of type I error; results should therefore be interpreted cautiously. The imputation of missing values for the ITT analysis was performed by multiple imputation. Imputations were performed using fully conditional specification with predictive mean matching (20 nearest neighbors), generating 100 imputed datasets. The imputation model included all primary and secondary outcome change scores (Y-BOCS total and subscales, Obsessive Compulsive Inventory, Beck Depression Inventory, GAD-7, Obsessive Beliefs Questionnaire, and quality of life), as well as relevant baseline covariates (age, gender, marital status, duration of illness, age at onset, and number of previous outpatient treatments). Pooled estimates were obtained across imputations according to Rubin's rules. Within-group differences on the Y-BOCS (and for exploratory purposes on its subscales), OCI-R, GAD-7, BDI-II, OBQ-44, and WHOQOL-BREF were analyzed with paired-sample 2-tailed t tests. Pearson correlations were analyzed for the relationship between sense of presence (measured with the Igroup Presence Questionnaire [IPQ]) and reduction in the primary outcome (Y-BOCS). For exploratory subsample analyses addressing the 2 types of OCD (conOCD and checkOCD), a 2×2 ANCOVA that included type of OCD as an additional fixed factor was conducted for both t0-t1 and t0-t2 for the Y-BOCS (and for exploratory purposes for its subscales), OCI-R, GAD-7, BDI-II, OBQ-44, and WHOQOL-BREF. Here, only trend or significant effects are reported for clarity and conciseness. The assumption of homogeneity of variances was tested with the Levene test. An exploratory moderation analysis was conducted using the SPSS macro PROCESS [59] for the CC sample for hypothesis-generating purposes. Here, the difference in the primary outcome (Y-BOCS total score) from t0-t1 and t0-t2 served as the dependent variable and group as the independent variable. Baseline sociodemographic variables as well as total and subscales of all primary and secondary

XSL•FO
RenderX

measures assessed at t0 were included as possible moderators (only significant or trend interaction effects are reported). Patient satisfaction was analyzed descriptively by displaying positively (combined completely agree and agree) and negatively formulated (combined disagree and completely disagree) separately.

Results

Overview

A total of 123 patients were screened for eligibility (see Figure 2 for details). The final sample consisted of 80 patients with OCD (40 patients with conOCD and 40 with checkOCD). At baseline, 1 participant per group did not complete the secondary outcome measures, resulting in a baseline sample size of 39 per group for these variables, while the full ITT sample (n=40 per group) was retained for all analyses. In total, 59 (73.8%) participants were women, and the mean age of the total sample was 36.85 (SD 11.44) years. The mean OCD duration was 18.04 (SD 14.22) years, and on average, the participants were aged 19.14 (SD 10.67) years at the onset of their illness.

Figure 2. Study flowchart. OCD: obsessive-compulsive disorder; VR: virtual reality.

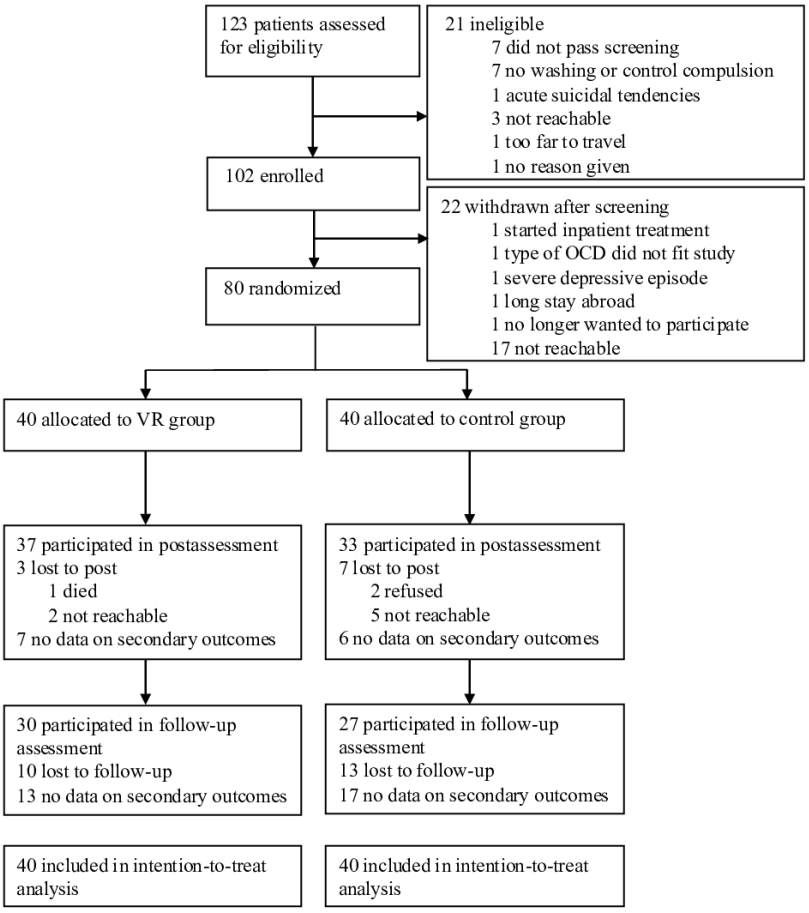

The VERP and CAU groups did not differ in sociodemographics, treatment variables, or psychopathology at t0 (Table 2). The sample showed severe OC symptoms (Y-BOCS) and moderate depressive symptoms (BDI-II) and anxiety symptoms (GAD-7). According to the Mini-International Neuropsychiatric Interview [56], main comorbidities were depression (53/80, 66.2%, with 18/80, 22.5% having a current depressive episode and 34/80, 42.5% a past depressive episode) and anxiety (current or lifetime panic disorder: 12/80, 15%; generalized anxiety disorder: 10/80, 12.5%; agoraphobia: 5/80, 6.3%; social phobia: 3/80, 3.8%). A majority of the sample had previously undergone psychotherapy (69/80, 86.2%), of which 14 of 80 (17.5%) had done so only currently, 35 of 80 (43.8%) only in the past, and 20 of 80 (25%) both currently and in the past. In total, 26 of 80 (32.5%) patients had prematurely discontinued a treatment in the past.

XSL•FO
RenderX

Table . Baseline characteristics with psychopathology (intention-to-treat sample) and treatment use during the intervention period (reported at post)[a].

	VERP[b] group (n=40)	CAU[c] group (n=40)	Statistics	*P* value
Demographic characteristics				
Gender, n (%)			χ^2_1=0.6	.45
Women	28 (70%)	31 (77.5%)		
Men	12 (30%)	9 (22.5%)		
Age (years), mean (SD)	38.18 (11.61)	35.53 (11.26)	t_{78}=1.04	.15
Qualified for university study	32 (80%)	29 (72.5%)	χ^2_2=0.7	.71
Treatment variables				
Medication, n (%)			χ^2_4=5.3	.26
Antidepressant	15 (37.5%)	14 (35%)		
Combination	4 (10%)	5 (12.5%)		
Other	6 (15%)	1 (2.5%)		
None	15 (37.5%)	20 (50%)		
Psychotherapy experience, n (%)			χ^2_3=2.8	.43
Current	6 (15%)	8 (20%)		
In the past	17 (42.5%)	18 (45%)		
Current and in the past	9 (22.5%)	11 (27.5%)		
None	8 (20%)	3 (7.5%)		
Discontinued a previous treatment prematurely, n (%)	15 (37.5%)	11 (27.5%)	χ^2_1=1.1	.30
Illness duration (years), mean (SD)	18.23 (13.57)	17.86 (15.02)	t_{78}=0.11	.46
Age at onset (years), mean (SD)	18.60 (11.58)	19.68 (9.80)	t_{78}=–0.45	.33
Psychopathology, mean (SD)				
Y-BOCS[d] total	24.40 (6.33)	24.25 (5.53)	t_{78}=0.11	.46
Y-BOCS obsessions	11.78 (3.33)	11.60 (3.10)	t_{78}=0.24	.40
Y-BOCS compulsions	12.63 (3.29)	12.65 (3.05)	t_{78}=–0.04	.49
OCI-R[e] total	31.77 (12.07)	31.18 (10.37)	t_{76}=0.23	.41
OCI-R washing	7.16 (4.30)	6.69 (4.24)	t_{76}=–0.48	.32
OCI-R obsessions	6.05 (3.21)	5.74 (3.08)	t_{76}=–0.43	.33
OCI-R hoarding	3.54 (3.33)	3.49 (2.59)	t_{76}=–0.08	.47
OCI-R ordering	4.54 (3.47)	4.72 (3.14)	t_{76}=0.24	.41
OCI-R checking	7.05 (3.64)	7.44 (3.07)	t_{76}=0.50	.31
OCI-R neutralizing	3.44 (3.72)	3.10 (3.39)	t_{76}=–0.41	.34
BDI-II[f]	19.69 (11.83)	21.64 (11.74)	t_{76}=–0.73	.23
GAD-7[g]	11.31 (5.07)	11.46 (4.47)	t_{76}=–0.14	.44
OBQ-44[h]	188.54 (45.96)	191.44 (53.68)	t_{76}=–0.26	.40
QoL[i]	2.90 (0.88)	2.82 (1.00)	t_{76}=0.36	.36
Treatment use during the intervention period[j]				

XSL•FO
RenderX

	VERP[b] group (n=40)	CAU[c] group (n=40)	Statistics	*P* value
Outpatient psychotherapy	17 (63%)	13 (56.5%)	χ^2_1=0.2	.64
Guided ERP[k] in vivo	8 (29.6%)	4 (17.4%)	χ^2_1=1.0	.31
Unguided ERP in vivo	12 (44.4%)	7 (30.4%)	χ^2_1=1.0	.31

[a]For psychopathological questionnaires that were collected online (OCI-R, BDI-II, GAD-7, OBQ-44, and QoL), data are only available for 39 per group due to missing baseline responses from 1 participant in each group. All participants were retained in the intention-to-treat sample.

[b]VERP: virtual exposure with response prevention.

[c]CAU: care as usual.

[d]Y-BOCS: Yale-Brown Obsessive-Compulsive Scale.

[e]OCI-R: Obsessive-Compulsive Inventory—Revised.

[f]BDI-II: Beck Depression Inventory-II.

[g]GAD-7: Generalized Anxiety Disorder Scale-7.

[h]OBQ-44: Obsessive Beliefs Questionnaire-44.

[i]Global item of the World Health Organization Quality of Life—BREF (quality of life).

[j]VERP group: n=27; CAU group: n=23.

[k]ERP: exposure with response prevention.

At t1, 17 (63%) participants of the VERP group reported having received outpatient psychotherapy during the intervention period, with 8 (29.6%) participants having received therapist-guided ERP in vivo, and 12 (44.4%) having received unguided ERP in vivo (Table 2). In contrast, 13 (56.5%) participants in the CAU group reported having received outpatient psychotherapy during the intervention period, with 4 (17.4%) participants having received therapist-guided ERP in vivo, and 7 (30.4%) having received unguided ERP in vivo. Participation in outpatient therapy, guided or unguided ERP, did not significantly differ between the 2 groups (all *P*>.05). Regarding the differentiation according to OCD subtype (conOCD and checkOCD), 13 (56.5%) patients with conOCD reported having received outpatient psychotherapy, with 5 (21.7%) participants having received therapist-guided ERP in vivo and 10 (43.5%) unguided ERP in vivo. In contrast, 17 (63%) patients with checkOCD reported having received outpatient psychotherapy, with 7 (25.9%) participants having received therapist-guided ERP in vivo and 9 (33.3%) unguided ERP in vivo. Again, these differences did not reach statistical significance (all *P*>.05).

Completion

The ITT sample had 80 participants, of whom 70 (87.5%) participants completed the t1 interview, and 57 (71.3%) participants completed the t2 interview (Figure 2). Regarding the secondary outcomes that were collected online, some missing data must also be reported. In total, 13 (16.3%) participants did not complete the t1 online questionnaires, and 30 (37.5%) participants did not complete the t2 online questionnaires. Therefore, the CC sample included 57 participants for the primary outcome and 50 participants for the secondary outcomes. Completion did not differ between the 2 groups for the primary (t_{78}=–0.73; *P*=.23) and secondary (t_{78}=–0.92; *P*=.18) outcomes. Of the participants in the VERP group, 4 participants did not take part in any therapy sessions. The remaining 36 participants attended an average of 5 (mean 5.31, SD 1.45) of the 6 sessions.

CC Analyses

The results of the CC analyses are summarized in Table 3 for the primary outcome and in Table 4 for the secondary outcomes. For the primary outcome, there were no significant group differences (for the total scale and the 2 subscales of the Y-BOCS; all *P*>.05; η_p^2=0.003-0.014). This applies both to the change from t0 to t1 and to the change from t0 to t2.

XSL•FO
RenderX

Table . Primary outcomes at each assessment time for complete cases (CC) sample[a].

	VERP[b] group			CAU[c] group			ANCOVA[j] with respective baseline values as covariates							
	t0	t1	t2	t0	t1	t2	CC t0-t1			ITT[e] t0-t1	CC t0-t2			ITT t0-t2
							F test (*df*=1,54)	*P* value	η_p^2	*P* value	*F* test (*df*=1,54)	*P* value	η_p^2	*P* value
Y-BOCS[f] total	24.33 (6.03)	20.67 (7.01) [g]	20.10 (7.22) [g]	23.56 (5.50)	21.04 (6.07) [h]	20.64 (6.11) [h]	0.71	.40	0.013	.47	0.59	.45	0.011	.37
Y-BOCS obsessions	11.77 (3.16)	10.23 (3.81) [h]	9.27 (3.78) [g]	11.26 (2.92)	10.33 (3.16)	9.72 (3.31) [i]	0.32	.57	0.006	.72	0.76	.39	0.014	.36
Y-BOCS compulsions	12.57 (3.22)	10.43 (3.40) [g]	10.83 (3.82) [g]	12.30 (3.12)	10.70 (3.41) [g]	10.92 (3.32) [h]	0.69	.41	0.013	.36	0.16	.70	0.003	.53

[a]ANCOVA with respective baseline values as covariates for CCs (n=57) and ITT (n=80) samples. ITT analyses were conducted using multiple imputation in SPSS. For these models, SPSS provides pooled significance tests (*P* values) but does not output pooled *F* statistics, degrees of freedom, or effect size estimates (eg, partial eta squared). Accordingly, only *P* values are reported for the ITT analyses. CC analyses are presented with full test statistics.

[b]VERP: virtual exposure with response prevention.

[c]CAU: care as usual.

[d]ANCOVA: analysis of covariance.

[e]ITT: intention-to-treat.

[f]Y-BOCS: Yale-Brown Obsessive-Compulsive Scale.

[g]$P \leq .001$.

[h]$P < .01$.

[i]$P < .05$.

XSL•FO
RenderX

Table . Secondary outcomes at each assessment time for complete cases (CC) sample (n=50)[a].

	VERP[b] group			CAU[c] group			ANCOVA[d] with respective baseline values as covariates							
	t0	t1	t2	t0	t1	t2	CC t0-t1			ITT[e] t0-t1	CC t0-t2			ITT t0-t2
							F test (*df*=1,47)	*P* value	η_p^2	*P* value	*F* test (*df*=1,47)	*P* value	η_p^2	*P* value
OCI-R[f]	33.74 (12.26)	30.41 (11.94)[g]	31.52 (12.03)	29.44 (11.67)	27.91 (13.01)	30.57 (15.45)	0.34	.56	0.007	.49	1.18	.28	0.024	.29
BDI-II[h]	19.93 (11.47)	18.07 (11.39)	20.15 (12.72)	20.83 (11.79)	19.78 (13.63)	19.04 (14.24)	0.20	.66	0.004	.86	0.47	.49	0.010	.55
GAD-7[i]	11.30 (5.14)	11.41 (5.73)	11.07 (4.84)	10.74 (4.63)	10.35 (6.01)	9.61 (5.66)	0.24	.62	0.005	.70	0.85	.36	0.018	.36
OBQ-44[j]	189.82 (49.38)	189.67 (50.44)	171.78 (51.62)[g]	187.44 (60.06)	180.09 (55.56)	167.52 (58.43)[k]	0.71	.40	0.015	.32	0.06	.81	0.001	.69
QoL[l]	2.81 (0.79)	2.93 (0.73)	2.85 (0.95)	2.87 (0.87)	3.17 (0.94)[g]	3.43 (0.99)[g]	1.24	.27	0.026	.27	4.64	.04	0.090	.03

[a]ANCOVA with respective baseline values as covariates for CC (n=50) and ITT (n=80) samples. ITT analyses were conducted using multiple imputation in SPSS. For these models, SPSS provides pooled significance tests (*P* values) but does not output pooled *F* statistics, degrees of freedom, or effect size estimates (eg, partial eta square). Accordingly, only *P* values are reported for the ITT analyses. CC analyses are presented with full test statistics.

[b]VERP: virtual exposure with response prevention.

[c]CAU: care as usual.

[d]ANCOVA: analysis of covariance.

[e]ITT: intention-to-treat.

[f]OCI-R: Obsessive-Compulsive Inventory—Revised.

[g]*P*<.05.

[h]BDI-II: Beck Depression Inventory-II.

[i]GAD-7: Generalized Anxiety Disorder Scale-7.

[j]OBQ-44: Obsessive Beliefs Questionnaire-44.

[k]*P*<.001.

[l]Global item of the World Health Organization Quality of Life—BREF (quality of life).

For the within-group differences, both groups showed a significant improvement in the primary outcome from t0 to t1 (Y-BOCS total, VERP group: t_{29}=3.95; *P*<.001; CAU group: t_{26}=3.27; *P*=.002) and from t0 to t2 (Y-BOCS total, VERP group: t_{29}=3.95; *P*<.001; CAU group: t_{26}=2.95; *P*=.003).

With regard to the secondary outcomes, there was only 1 significant group difference in the change in quality of life (global item of the World Health Organization Quality of Life—BREF) from t0 to t2, favoring the CAU group ($F_{1,47}$=4.64; *P*=.04; η_p^2=0.090). For all other secondary outcomes, there were no group differences from t0 to t1 or from t0 to t2 (all *P*>.05).

Significant within-group differences for secondary outcomes were only found for the change from t0 to t1 in the OCI-R (t_{26}=2.14; *P*=.02) and from t0 to t1 in the OBQ-44 (t_{26}=2.23; *P*=.02) in the VERP group. For the CAU group, within-group differences were found for the OBQ-44 between t0 to t2 (t_{22}=3.60; *P*<.001) and for quality of life between t0 to t1 as well as between t0 and t2 (t_{22}=−2.73; *P*=.006).

ITT Analyses

The results of the ITT analyses were largely comparable to those of the CC analyses (Tables 3 and 4). Again, only the improvement in quality of life showed a significant group difference from t0 to t2, favoring the CAU group (*P*=.03).

Remission and Response

In total, 10 (17.5%) participants of the CC sample showed remission (Y-BOCS score ≤14 [60]) at t1, 6 (20%) patients of the VERP group, and 4 (14.8%) patients of the CAU group. At t2, 7 (23.3%) patients of the VERP group and 5 (18.5%) patients of the CAU group showed remission. In the VERP group, 5 of 30 (16.7%) participants can be regarded as responders (ie, reduction of >35% in primary outcome Y-BOCS) at t1 and 7 of 30 (23.3%) participants at t2. In the CAU group, 2 of 27 (7.4%) participants can be regarded as responders at t1, and 4 of 27 (14.8%) participants at t2.

Sense of Presence

The mean value for the perceived sense of presence at post was 37.5 (SD 14.35) for the overall scale of IPQ, with the following mean values for the subscales: spatial presence 16.12 (SD 6.00), involvement 9.81 (SD 4.14), and experience realism 8.15 (SD 5.01). No significant correlations were present for any subscale

XSL•FO
RenderX

of the IPQ nor for the total scale with the difference on the primary outcome and its subscales at each assessment time ($r \leq 0.379$; $P \geq .06$). Sense of presence (total scale and subscales) did not differ between the 2 symptom subtypes ($P \geq .27$).

Exploratory Analyses Including OCD Subtype

When including OCD subtype (conOCD and checkOCD) as a fixed factor within a 2×2 factorial ANCOVA, we observed a trend toward an interaction effect between group and type of OCD for the change in obsessions (Y-BOCS subscale obsessions) from t0 to t1 ($F_{1,52}$=3.91; P=.05, η_p^2=0.070) favoring the combination VERP group and patients with checkOCD (Figure 3). To further examine the trend-level interaction between OCD subtype and treatment group, we conducted simple-effects analyses. For patients with checkOCD, there was a statistical trend toward greater improvement in the VERP group compared to CAU alone ($F_{1,52}$=3.16; P=.08). In contrast, among patients with conOCD, differences between VERP and CAU were not significant ($F_{1,52}$=1.09; P=.30). Thus, the interaction pattern appears to be primarily driven by relatively greater symptom reduction among patients with checkOCD receiving CAU plus VERP, rather than by reduced improvement among patients with conOCD in CAU plus VERP. Additionally, we found a main effect for type of OCD for change in anxiety (GAD-7) from t0 to t2 ($F_{1,45}$=4.11; P=.048; η_p^2=0.084), showing that patients with conOCD had a significantly greater decrease in anxiety compared to patients with checkOCD (mean difference for patients with conOCD 0.74, SD 3.84; mean difference for patients with checkOCD –0.41, SD 4.72).

XSL•FO
RenderX

Figure 3. Significant interactions of exploratory 2-factor (group and type of obsessive-compulsive disorder) ANCOVA. Error bars: 95% CI. (A) Estimated marginal means of difference on subscale obsession of the Y-BOCS from baseline to post for the complete cases sample. Covariates appearing in the model were evaluated at Y-BOCS subscale obsession baseline=11.60. (B) Estimated marginal means of difference on the Y-BOCS total scale from baseline to post. Covariates appearing in the model were evaluated at Y-BOCS total scale baseline =24.36. (C) Estimated marginal means of difference on subscale obsession of the Y-BOCS from baseline to post for all available lost-to-follow-up data. Covariates appearing in the model were evaluated at Y-BOCS subscale baseline=11.60. ANCOVA: analysis of covariance; CAU: care as usual group; VERP: virtual exposure with response prevention; Y-BOCS: Yale-Brown Obsessive-Compulsive Scale.

XSL•FO
RenderX

To carry out the analysis with the largest possible sample, the same analysis was calculated with all available lost-to-t2 data for exploratory purposes; we additionally found a significant interaction effect between group and type of OCD for the change in OC symptoms (Y-BOCS total scale) from t0 to t1 ($F_{1,65}$=4.64; P=.04; η_p^2=0.067) and a significant interaction effect between group and type of OCD for the change in obsessions (Y-BOCS subscale obsessions) from t0 to t1 ($F_{1,65}$=6.66; P=.01; η_p^2=0.093), both favoring the combination VERP group and patients with checkOCD (Figure 3).

Descriptive data (means and SDs of Y-BOCS total and subscale scores at all 3 time points) for both subtypes are presented in Multimedia Appendices 1 and 2.

Exploratory Moderation Analysis

Results of the exploratory moderation analysis are summarized in Table 5. Individuals in the VERP group who did not prematurely discontinue a previous treatment showed a significantly greater reduction on the Y-BOCS at t1 (P=.04) and t2 (P=.008) relative to the CAU group. In addition, individuals in the VERP group who scored high on compulsive checking (OCI-R subscale checking) showed a trend toward greater reduction on the primary outcome at post relative to the CAU group at t1 (OCI-R subscale checking: P=.05).

Table . Moderators for obsessive-compulsive symptom improvement (Y-BOCS[a] total difference scores, means are centered), results of the complete cases sample.

	B (SE)	*t* test	*P* value	LLCI[c]	ULCI[d]	*P* for –1SD	*P* for 0	*P* for +1SD
Y-BOCS difference baseline to post								
Discontinuation of prior treatment	6.08 (2.83)	2.15	.04	0.40	11.75	.16	—[e]	.07
OCI-R[b] subscale checking	0.75 (0.37)	2.00	.05	–0.002	1.50	.35	.31	.04
Y-BOCS difference baseline to follow-up								
Discontinuation of prior treatment	9.12 (3.31)	2.76	.008	2.49	15.76	.07	—	.03

[a]Y-BOCS: Yale-Brown Obsessive-Compulsive Scale.

[b]OCI-R: Obsessive-Compulsive Inventory—Revised.

[c]LLCI: lower level confidence interval.

[d]ULCI: upper level confidence interval.

[e]Not available.

Subjective Appraisal

Results of the subjective appraisal are summarized in Figure 4. The majority evaluated VERP as useful and meaningful (25/33, 75.8%) and that they would recommend it to others (24/32, 75%). The purpose of the intervention was clear to all participants (33/33, 100%), and all participants felt comfortable with the therapist (33/33, 100%). Half of the participants reported that their compulsions (16/32, 50%) and obsessions (16/33, 48.5%) decreased due to VERP. Only 1 (1/33, 3%) participant stated that VERP had a negative effect on other interventions, that they had to force themselves to attend the VERP sessions, and that they developed new obsessions due to VERP. Qualitative feedback on the intervention was positive. Most of the participants expressed their appreciation of and gratitude to the study therapists. One participant stated:

> *The threshold for entering exposure therapy has been significantly lowered through the use of virtual reality. In addition, you can playfully test and explore your own fears and compulsions in a safe environment and approach fearful objects with a safety net.*

XSL•FO
RenderX

Figure 4. Subjective appraisal of VERP (n=33; n=32 for 2 items: My compulsions decreased due to VERP and I would recommend VERP to others). (A) Positive items and (B) reverse items. VERP: virtual exposure with response prevention.

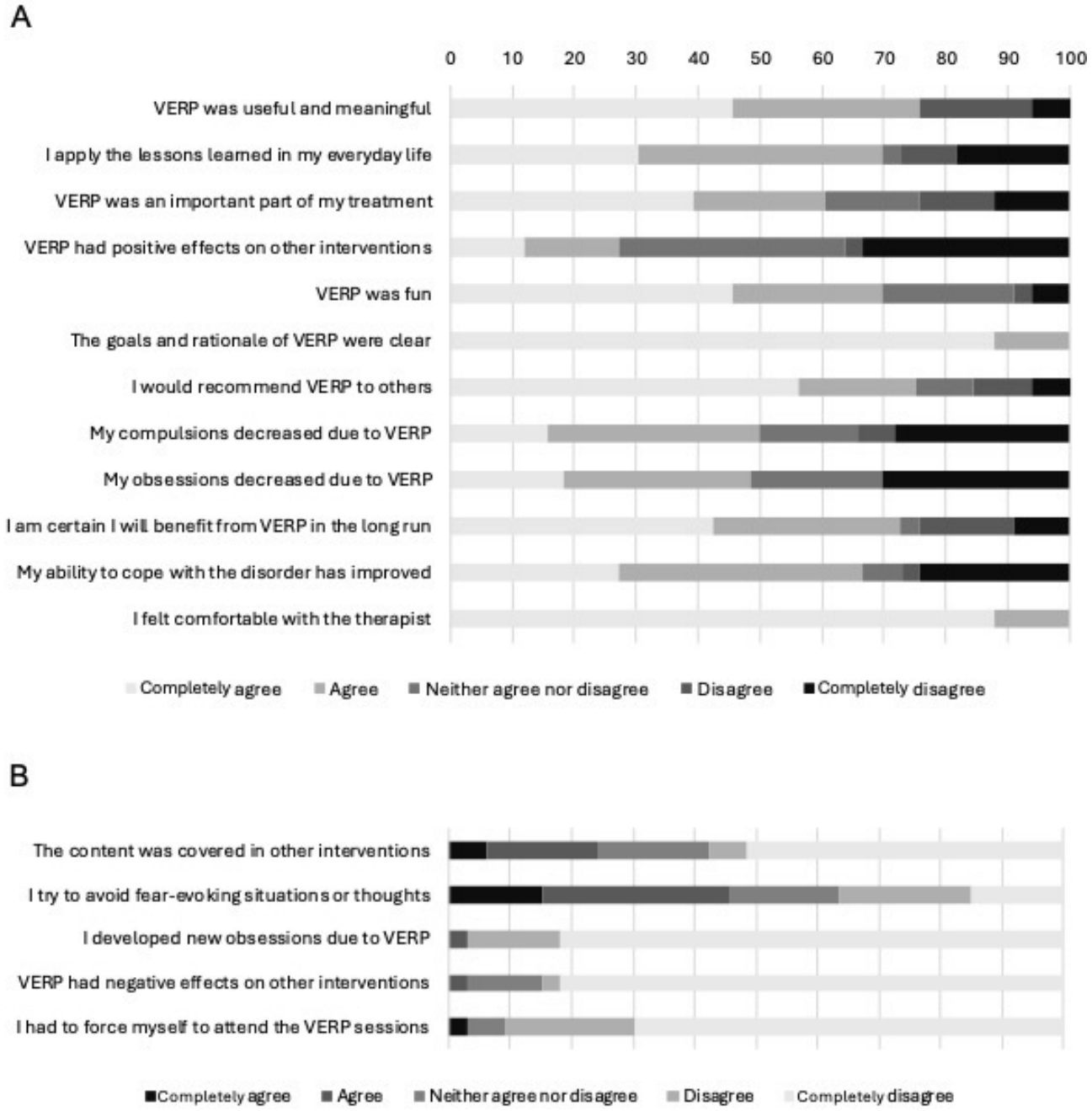

Discussion

The aim of the study was to investigate the efficacy and acceptance of VERP as an add-on to CAU in a sample of patients with OCD and contamination-related (patients with conOCD) or checking-related (patients with checkOCD) compulsions and to identify possible moderators of treatment success.

Main Findings

Within this randomized controlled trial, we did not find additional efficacy of VERP in the total sample of patients with OCD (patients with conOCD and patients with checkOCD) when measured after the 6-week intervention period (post) and 3 months after post. Both the VERP group and the CAU group improved in the primary outcome (OC symptom reduction). This result contrasts with findings from prior studies that have shown the efficacy of VERP in reducing OC symptoms [21,31]. One reason for this could be that the intervention period in the study by Javaherirenani et al [21] lasted for 12 weeks (ie, twice as long as in our study), which could have played a role in the more favorable outcomes observed in their study. In addition, the unusually high baseline Y-BOCS scores in their sample may have facilitated greater symptom reduction compared to our study. Furthermore, with only 1 session per week in our study, there is a lot of time between sessions, so old behavior patterns can quickly reappear. In line with this, there are highly effective concentrated therapeutic approaches in the treatment of OCD that aim for block-based treatment [61,62]. It is also conceivable that in VERP, the transfer from digital to everyday life is more

XSL•FO
RenderX

difficult compared to ERP in vivo, meaning that generalization is less likely to take place. It is further possible that there was too little variance in our scenario (always the same therapy room with some repeated OC-relevant objects), as Craske et al [63] postulate that repeated practice in different contexts is important for the generalization of the knowledge acquired during ERP. In addition, compulsions are highly idiosyncratic, so our selection of objects may not have been truly relevant to every patient (some patients are more afraid of radiation or asbestos, eg, which we did not address in our study). Furthermore, in our study, a large proportion of the CAU group received outpatient psychotherapy during the intervention period, some with guided or unguided ERP in vivo, which may explain the improvement in the CAU group. Several technical and design-related factors may have also contributed to the absence of significant between-group effects. Locomotion in the virtual environments was implemented via teleportation, which—while reducing simulator sickness—may have disrupted spatial presence and attenuated emotional engagement compared to more naturalistic movement. In addition, the lack of haptic feedback (eg, touching a clean chair while viewing a contaminated stimulus) may have reduced perceptual realism and contributed to the moderate presence ratings observed.

Ecological validity may also have been limited in the checking scenario. Participants were instructed to switch devices off once and immediately leave the apartment, which may not have fully captured postdeparture doubt and urges to recheck—core features of checking compulsions. Together, these factors may have constrained anxiety induction and thereby reduced the observable impact of VERP on global OCD severity.

Importantly, the absence of even a trend-level difference between groups deserves further consideration. More fundamental explanations may include that the VERP component was not sufficiently potent to add therapeutic benefit beyond CAU, and that the heterogeneity of concurrent treatments (including therapist-guided ERP in both groups, albeit more frequently in the VERP group) likely diluted potential differences attributable to the experimental intervention. In addition, the limited sample size may have restricted our ability to detect subtle effects. Taken together, these factors suggest that the lack of group differences cannot be attributed solely to technical aspects of the intervention but may also reflect broader methodological and conceptual limitations.

Additionally, a low level of sense of presence could have played a role. However, only 1 of the 2 studies mentioned earlier assessed sense of presence [31], and its authors did not provide an interpretation of the scores assessed, so we cannot come to any conclusion regarding whether the observed effects might be due to a higher sense of presence. In our study, the reported level of sense of presence was only moderate (mean 37.5, SD 14.35; range 14-94). Although correlations between sense of presence and treatment outcome in our study did not reach statistical significance ($r \leq 0.379$; $P \geq .06$), others found a higher sense of presence associated with higher experienced anxiety in VRET for anxiety disorders [33], which could also apply to OCD. The induction of distress or an unpleasant emotion such as anxiety or disgust is an essential component of ERP [5], and it is therefore possible that more positive results could have been achieved in our study if the experienced sense of presence had been higher (eg, if patients, especially patients with conOCD, could see their own hands in the VR environment [25,64]). Compared to in vivo ERP, the absence of genuine haptic contamination cues (eg, real dirt or moisture) may have limited the intensity of obsessive thoughts and emotional responses during VERP. This limitation, together with the only moderate levels of sense of presence reported in our study, might partly explain why VERP did not yield additional effects, in contrast to studies such as Javaherirenani et al [21], where patients physically touched dirty objects and repeated this as homework.

Another aspect might be that on an even more general level, the construct of sense of presence may not fully capture disorder-specific conditions of exposure in OCD, such as the threat of having actually contaminated oneself or overlooked something with catastrophic consequences. Since these fears often arise from patients' own actions or omissions rather than visual input alone, VR-based exposure may elicit weaker emotional responses in OCD than in anxiety disorders.

Surprisingly, the only significant group difference found was an increase in quality of life from t0 to t2 in the CAU group compared to the VERP group, which was contrary to our expectations. A possible reason for this could be that participants who knew they were taking part in a trial may have experienced an improvement in their quality of life regardless of the actual treatment because they paid more attention to their behavior and symptoms and took better care of themselves overall. Another explanation could be that those participants of the CAU group who were not receiving treatment experienced a feeling of relief because they temporarily did not have to actively deal with their OCD.

Subsample Analyses

If we take a closer look at the subsample analyses by content of OCD (patients with conOCD vs patients with checkOCD), we found a trend-level interaction between group and subtype for the change in obsessions from t0 to t1. This may indicate that the 2 subtypes responded differently to the intervention, with a possible tendency toward greater reduction in obsessions for patients with checkOCD. However, these findings should be interpreted with caution. Given the absence of an overall effect in the full sample, as well as the exploratory nature of these analyses, the results should be regarded as only preliminary indications that warrant replication in future studies with a priori hypotheses and larger samples. Moreover, the factorial validity of the Y-BOCS subscales themselves has been debated [65,66], further limiting the interpretability of subscore-based findings. From a clinical perspective, it is conceivable that differences in the predominant emotions associated with each subtype (fear in checkOCD and disgust in conOCD) may play a role in how patients engage with VR-based exposure. In our opinion, it would have seemed more plausible that patients with conOCD would show greater improvement, given that disgust can typically be elicited more directly by visual stimuli in VR, whereas anxiety in checkOCD is often tied to imagined future consequences (eg, forgetting to turn off

XSL•FO
RenderX

the stove), which may be less readily activated in a virtual environment. At the same time, patients with conOCD might have engaged in more cognitive avoidance during exposure, which could have reduced the efficacy of VERP for this group—although this remains speculative, as our measures did not allow us to test this assumption. It is also possible that the abovementioned reasons for the general lack of effects apply even more to patients with conOCD (high idiosyncrasy, too little variance in the scenario, unsuitable stimuli, etc).

We also found a main effect for subtype, with patients with conOCD showing a greater decrease in anxiety from t0 to t2. A possible explanation could be that washing compulsions are often more directly linked to immediate anxiety reduction, whereas checking compulsions are often tied to future-oriented threats. This might make patients with conOCD more responsive to interventions aimed at anxiety reduction.

Exploratory Moderation Analysis

The findings of the exploratory moderation analysis provide deeper insight into the factors that may influence the effectiveness of adding VERP to CAU. Individuals in the VERP group who scored high on compulsive checking showed a trend toward greater reductions in the primary outcome than those in the CAU group at t1. This suggests that CAU plus VERP may be particularly effective for individuals with compulsive checking behavior, indicating that this subtype of OCD may benefit more from VERP, which is also implied by the subgroup analyses. Comparable studies examining the efficacy of VERP have so far only included patients with washing compulsions [21,31], so we cannot draw a comparison here.

Individuals in the VERP group who had not previously discontinued treatment prematurely showed a significantly greater reduction in OC symptoms at both t1 and t2 compared to those in the CAU group, indicating that patients with a history of consistent treatment may respond better to CAU plus VERP or other forms of treatment, highlighting the importance of adherence to treatment protocols for therapeutic success [67]. Consistency in treatment reflects a level of engagement and commitment to the therapeutic process. It indicates that the patient has built a degree of resilience and is willing to work through the discomfort that is often part of exposure-based therapies such as VERP. This consistency allows for greater opportunities for learning and adaptation, which are crucial factors in reducing OC symptoms.

Subjective Acceptance

The subjective evaluation of VERP was positive overall, particularly with regard to the usefulness, purpose, fun, and applicability of the intervention as well as the therapeutic relationship, but it was rather moderate with regard to the reduction of OC symptoms. However, the results of the subjective evaluation indicate that VERP may be better accepted than ERP in vivo [14], although no direct comparison is possible based on the present study design.

Strengths and Limitations

Strengths of the study were the inclusion of an active control group, the verification of diagnoses using standardized clinical interviews by trained and blinded raters, and the assessment of self-rating psychopathology questionnaires as well as the collection of follow-up data. At the same time, several limitations should be noted.

First, the intervention period was rather short, consisting of only 6 weekly sessions, of which the first 2 were primarily preparatory and only approximately 4 sessions involved active exposure exercises. This represents a substantially lower treatment “dose” than standard ERP protocols, which typically include a higher number of exposure sessions over a longer time frame. Given the chronic nature of OCD in this sample (mean illness duration of 18.04, SD 14.22 years), it is questionable whether this limited exposure intensity was theoretically sufficient to induce measurable changes in global OCD severity as assessed by the Y-BOCS. Consequently, the relatively brief intervention duration and limited number of active exposure sessions may have been a primary contributor to the small and nonsignificant effect sizes observed, particularly with respect to overall symptom reduction.

Second, although the study used a randomized controlled design, the control condition was treatment as CAU in outpatient care and not an active comparator (eg, ERP in vivo); therefore, conclusions about the relative efficacy of VERP compared to standard ERP in vivo cannot be drawn. In addition, a substantial proportion of participants in the CAU group received concurrent outpatient psychotherapy, and more than half were exposed to guided or unguided ERP as part of their usual care. This overlap with established evidence-based interventions likely reduced the contrast between groups and may have diluted potential add-on effects of VERP, thereby contributing to the absence of significant between-group differences.

Third, the sample was restricted to 2 symptom subtypes (conOCD and checkOCD). The findings may thus not generalize to the broader spectrum of OC presentations. Participants were assigned to subtypes based on their predominant symptom dimension, although OCD symptoms frequently co-occur across dimensions. This binary classification may therefore oversimplify the heterogeneous nature of OCD and may have resulted in misclassification, potentially attenuating or obscuring differential intervention effects between subgroups.

In addition, although diagnoses were verified using the Mini-International Neuropsychiatric Interview and OCD symptom severity was assessed with the gold-standard semistructured Y-BOCS interview, several secondary outcomes (eg, OCI-R and BDI-II) relied on self-report and were assessed without blinding. These measures may thus have been susceptible to response biases, including demand characteristics in the intervention group or resentful demoralization in the control group, which may have influenced symptom ratings independently of true treatment effects. In addition, sample size within subgroups was modest, which limited the power to detect smaller effects and to explore interaction effects across different clinical characteristics.

Fourth, dropout and prior treatment history influenced treatment response, as indicated by the moderation analysis. This points to selection effects and highlights the need to replicate the findings in a larger, more heterogeneous sample. Although

XSL•FO
RenderX

missing data were handled using multiple imputation under a missing-at-random assumption, no formal sensitivity analyses under nonmissing-at-random assumptions were conducted. If participants who discontinued follow-up assessments experienced poorer outcomes, treatment effects may have been overestimated.

Fifth, longer-term effects beyond the 3-month follow-up were not assessed. Therefore, the durability of the effects, particularly the differential effects for conOCD and checkOCD, remains to be determined. A further limitation concerns the sample size estimation. Because empirical data on VERP in OCD were scarce at the time of study planning, our power analysis was based on effect sizes reported for VRET in anxiety disorders. Given that previous VERP trials in OCD have not shown effects as large as those in anxiety disorders, this assumption may have been overly optimistic. As a result, our study might have been underpowered to detect more modest effects of VERP in OCD.

Moreover, the relatively small sample size may limit the reliability of the trend-level interaction effects observed in subgroup analyses, which should therefore be interpreted with caution.

Another limitation concerns the exploratory nature of the analyses of Y-BOCS subscales. These measures are highly interrelated, and multiple testing increases the risk of type I error. Therefore, the results on subscales need to be interpreted cautiously, and we note that the study was primarily powered to detect differences in the Y-BOCS total score. Exploratory moderation analyses based on single-item measures should also be interpreted with caution and warrant replication using validated instruments and prespecified hypotheses.

Furthermore, medication use was assessed only at baseline, and no monitoring or management of psychotropic medication occurred during the study. Consequently, changes in medication during the intervention period could not be controlled and may have confounded symptom change, potentially masking or mimicking intervention effects.

Additionally, recruitment partly involved contacting participants from previous studies, which may have introduced a selection bias toward individuals with more chronic or treatment-refractory OCD, potentially contributing to the high chronicity of the sample and the limited treatment response observed.

Finally, the add-on nature of the trial limits the interpretation of our findings. Because both groups received CAU and only the intervention group received additional VERP sessions, the results demonstrate the incremental benefit of VERP but do not allow conclusions about the efficacy of VERP as a stand-alone treatment or in direct comparison with established ERP. We deliberately chose this design because VERP is still in an early stage of evaluation, and testing it as an adjunct to CAU offered a pragmatic and ethically appropriate first step that mirrors real-world clinical practice. Future studies using alternative designs (eg, noninferiority trials or head-to-head comparisons) will be needed to clarify the relative effectiveness of VERP within the broader spectrum of OCD interventions.

Recommendations

Further research should examine whether VERP offers greater efficacy than standard treatments with extended intervention periods and advanced virtual environments, as well as in comparison to in vivo ERP. Studies should also investigate how the sense of presence relates to treatment outcomes and how it can be increased (possibly by using mixed reality approaches) and assess VERP's effectiveness for other OCD types.

Conclusions

The study revealed that CAU plus VERP did not show a significant effect in reducing OC symptoms in the whole sample of patients with conOCD and patients with checkOCD after a 6-week intervention and a 3-month follow-up. However, exploratory subsample analyses provided preliminary indications that adding VERP to CAU might be associated with greater reductions in obsessions among patients with checkOCD and with greater reductions in anxiety among patients with conOCD. Similarly, patients in the VERP group who reported checking compulsions and who had not previously discontinued treatment prematurely showed signals of greater improvement in OC symptoms at posttreatment and follow-up compared to those in the CAU group. Given the exploratory nature and methodological limitations of these analyses, these findings should be interpreted with caution and considered hypothesis-generating. While VERP was generally well accepted by participants, several factors could play a role in the current nonsignificant treatment effects.

Acknowledgments

The authors thank the study's therapists, Brigit Hottenrott, Ruth Veckenstedt, Jakob Scheunemann, Marcel Riele, Mira Berthold, and Swantje Borsutzky, as well as the developers of the virtual reality scenarios, Jannik Blömer and Karsten Grzella. The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: summarizing text. The GAI tool ChatGPT-5 was used to summarize and optimize self-generated text. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

XSL•FO
RenderX

Funding

The study was funded by Freie und Hansestadt Hamburg; Behörde für Arbeit, Gesundheit, Soziales, Familie und Integration (Free and Hanseatic City of Hamburg; German Ministry of Labor, Health, Social Affairs, Family and Integration). The funders had no involvement in any step of the study or the publication process.

Data Availability

Data can be made available upon request to the corresponding author.

Authors' Contributions

Conceptualization: LR, LJ, SM, SK, FM
Funding acquisition: LR, FM
Study design: LR, FM
Project administration: LR, LL, FM
Resources: LJ, SM, JG, SK
Data collection: LR, LL, FM
Data analysis: LR
Data interpretation: LR, LJ, FM
Literature search: LB
Figures: LR
Writing—original draft: LR, LB, FM
Writing—review and editing: LJ, LL, SM, LB, JG, SK, FM

Conflicts of Interest

None declared.

Multimedia Appendix 1
Yale-Brown Obsessive-Compulsive Scale interview.
[PDF File, 543 KB - jmir_v28i1e79326_app1.pdf]

Multimedia Appendix 2
Descriptives (means and SDs) of the Yale-Brown Obsessive-Compulsive Scale total score and its subscales (obsession and compulsions) at all 3 assessment times for the 2 obsessive-compulsive disorder subtypes.
[DOCX File, 18 KB - jmir_v28i1e79326_app2.docx]

Checklist 1
CONSORT checklist.
[PDF File, 128 KB - jmir_v28i1e79326_app3.pdf]

References

1. Stein DJ, Costa DLC, Lochner C, et al. Obsessive-compulsive disorder. Nat Rev Dis Primers 2019 Aug 1;5(1):52. [doi: 10.1038/s41572-019-0102-3] [Medline: 31371720]
2. Fawcett EJ, Power H, Fawcett JM. Women are at greater risk of OCD than men: a meta-analytic review of OCD prevalence worldwide. J Clin Psychiatry 2020 Jun 23;81(4):19r13085. [doi: 10.4088/JCP.19r13085] [Medline: 32603559]
3. Ferentinos P, Preti A, Veroniki AA, et al. Comorbidity of obsessive-compulsive disorder in bipolar spectrum disorders: systematic review and meta-analysis of its prevalence. J Affect Disord 2020 Feb 15;263:193-208. [doi: 10.1016/j.jad.2019.11.136] [Medline: 31818777]
4. Koran LM. Quality of life in obsessive-compulsive disorder. Psychiatr Clin North Am 2000 Sep;23(3):509-517. [doi: 10.1016/s0193-953x(05)70177-5] [Medline: 10986724]
5. Law C, Boisseau CL. Exposure and response prevention in the treatment of obsessive-compulsive disorder: current perspectives. Psychol Res Behav Manag 2019;12:1167-1174. [doi: 10.2147/PRBM.S211117] [Medline: 31920413]
6. Öst LG, Havnen A, Hansen B, Kvale G. Cognitive behavioral treatments of obsessive-compulsive disorder. A systematic review and meta-analysis of studies published 1993-2014. Clin Psychol Rev 2015 Aug;40:156-169. [doi: 10.1016/j.cpr.2015.06.003] [Medline: 26117062]
7. Voderholzer U. Zwangsstörungen und zwangsspektrumsstörungen [Obsessive-compulsive disorder and obsessive-compulsive spectrum disorders]. Nervenarzt 2022 Jul;93(7):659-660. [doi: 10.1007/s00115-022-01335-w]
8. Foa EB, Steketee G, Grayson JB. Imaginal and in vivo exposure: a comparison with obsessive-compulsive checkers. Behav Ther 1985 Jun;16(3):292-302. [doi: 10.1016/S0005-7894(85)80017-4]

XSL•FO
RenderX

9. Abramowitz JS. The psychological treatment of obsessive-compulsive disorder. Can J Psychiatry 2006 Jun;51(7):407-416. [doi: 10.1177/070674370605100702] [Medline: 16838822]
10. Abramowitz JS, Arch JJ. Strategies for improving long-term outcomes in cognitive behavioral therapy for obsessive-compulsive disorder: insights from learning theory. Cogn Behav Pract 2014 Feb;21(1):20-31. [doi: 10.1016/j.cbpra.2013.06.004]
11. Olatunji BO, Davis ML, Powers MB, Smits JAJ. Cognitive-behavioral therapy for obsessive-compulsive disorder: a meta-analysis of treatment outcome and moderators. J Psychiatr Res 2013 Jan;47(1):33-41. [doi: 10.1016/j.jpsychires.2012.08.020] [Medline: 22999486]
12. Bey K, Willems S, Dueren AL, Philipsen A, Wagner M. Help-seeking behavior, treatment barriers and facilitators, attitudes and access to first-line treatment in German adults with obsessive-compulsive disorder. BMC Psychiatry 2025 Mar 11;25(1):235. [doi: 10.1186/s12888-025-06655-0] [Medline: 40069700]
13. Gillihan SJ, Williams MT, Malcoun E, Yadin E, Foa EB. Common pitfalls in exposure and response prevention (EX/RP) for OCD. J Obsessive Compuls Relat Disord 2012 Oct 1;1(4):251-257. [doi: 10.1016/j.jocrd.2012.05.002] [Medline: 22924159]
14. Moritz S, Külz A, Voderholzer U, Hillebrand T, McKay D, Jelinek L. “Phobie à deux” and other reasons why clinicians do not apply exposure with response prevention in patients with obsessive-compulsive disorder. Cogn Behav Ther 2019 Mar;48(2):162-176. [doi: 10.1080/16506073.2018.1494750] [Medline: 30064295]
15. Pittig A, Kotter R, Hoyer J. The struggle of behavioral therapists with exposure: self-reported practicability, negative beliefs, and therapist distress about exposure-based interventions. Behav Ther 2019 Mar;50(2):353-366. [doi: 10.1016/j.beth.2018.07.003] [Medline: 30824251]
16. Maltby N, Tolin DF. A brief motivational intervention for treatment-refusing OCD patients. Cogn Behav Ther 2005;34(3):176-184. [doi: 10.1080/16506070510043741] [Medline: 16195056]
17. Bush J. Viability of virtual reality exposure therapy as a treatment alternative. Comput Human Behav 2008 May;24(3):1032-1040. [doi: 10.1016/j.chb.2007.03.006]
18. Botella C, Fernández-Álvarez J, Guillén V, García-Palacios A, Baños R. Recent progress in virtual reality exposure therapy for phobias: a systematic review. Curr Psychiatry Rep 2017 Jul;19(7):42. [doi: 10.1007/s11920-017-0788-4] [Medline: 28540594]
19. Garcia-Palacios A, Botella C, Hoffman H, Fabregat S. Comparing acceptance and refusal rates of virtual reality exposure vs. in vivo exposure by patients with specific phobias. Cyberpsychol Behav 2007 Oct;10(5):722-724. [doi: 10.1089/cpb.2007.9962] [Medline: 17927544]
20. Rothbaum BO, Rizzo AS, McDaniel DD, Zanov MV. Virtual reality exposure therapy. In: Encyclopedia of Mental Health: Elsevier; 2016:370-374. [doi: 10.1016/B978-0-12-397045-9.00196-8]
21. Javaherirenani R, Mortazavi SS, Shalbafan M, Ashouri A, Farani AR. Virtual reality exposure and response prevention in the treatment of obsessive-compulsive disorder in patients with contamination subtype in comparison with in vivo exposure therapy: a randomized clinical controlled trial. BMC Psychiatry 2022 Nov 28;22(1):740. [doi: 10.1186/s12888-022-04402-3] [Medline: 36443695]
22. Carl E, Stein AT, Levihn-Coon A, et al. Virtual reality exposure therapy for anxiety and related disorders: a meta-analysis of randomized controlled trials. J Anxiety Disord 2019 Jan;61:27-36. [doi: 10.1016/j.janxdis.2018.08.003] [Medline: 30287083]
23. van Loenen I, Scholten W, Muntingh A, Smit J, Batelaan N. The effectiveness of virtual reality exposure-based cognitive behavioral therapy for severe anxiety disorders, obsessive-compulsive disorder, and posttraumatic stress disorder: meta-analysis. J Med Internet Res 2022 Feb 10;24(2):e26736. [doi: 10.2196/26736] [Medline: 35142632]
24. Laforest M, Bouchard S, Bossé J, Mesly O. Effectiveness of in virtuo exposure and response prevention treatment using cognitive-behavioral therapy for obsessive-compulsive disorder: a study based on a single-case study protocol. Front Psychiatry 2016;7:99. [doi: 10.3389/fpsyt.2016.00099] [Medline: 27378951]
25. Miegel F, Bücker L, Kühn S, et al. Exposure and response prevention in virtual reality for patients with contamination-related obsessive-compulsive disorder: a case series. Psychiatr Q 2022 Sep;93(3):861-882. [doi: 10.1007/s11126-022-09992-5] [Medline: 35779165]
26. Cullen AJ, Dowling NL, Segrave R, Carter A, Yücel M. Exposure therapy in a virtual environment: validation in obsessive compulsive disorder. J Anxiety Disord 2021 May;80:102404. [doi: 10.1016/j.janxdis.2021.102404] [Medline: 33894550]
27. Dehghan B, Saeidimehr S, Sayyah M, Rahim F. The effect of virtual reality on emotional response and symptoms provocation in patients with OCD: a systematic review and meta-analysis. Front Psychiatry 2021;12:733584. [doi: 10.3389/fpsyt.2021.733584] [Medline: 35177996]
28. Ferraioli F, Culicetto L, Cecchetti L, et al. Virtual reality exposure therapy for treating fear of contamination disorders: a systematic review of healthy and clinical populations. Brain Sci 2024 May 17;14(5):510. [doi: 10.3390/brainsci14050510] [Medline: 38790488]
29. Belloch A, Cabedo E, Carrió C, Lozano-Quilis JA, Gil-Gómez JA, Gil-Gómez H. Virtual reality exposure for OCD: is it feasible. Rev Psicopatol Psicol Clin 2014;19:37-44. [doi: 10.5944/rppc.vol.19.num.1.2014.12981]

XSL•FO
RenderX

30. Dua D, Jagota G, Grover S. Management of obsessive-compulsive disorder with virtual reality-based exposure. Ind Psychiatry J 2021;30(1):179-181. [doi: 10.4103/ipj.ipj_33_19] [Medline: 34483545]
31. Inozu M, Celikcan U, Akin B, Mustafaoğlu Cicek N. The use of virtual reality (VR) exposure for reducing contamination fear and disgust: can VR be an effective alternative exposure technique to in vivo? J Obsessive Compuls Relat Disord 2020 Apr;25:100518. [doi: 10.1016/j.jocrd.2020.100518]
32. Wiebe A, Kannen K, Selaskowski B, et al. Virtual reality in the diagnostic and therapy for mental disorders: a systematic review. Clin Psychol Rev 2022 Dec;98:102213. [doi: 10.1016/j.cpr.2022.102213] [Medline: 36356351]
33. Ling Y, Nefs HT, Morina N, Heynderickx I, Brinkman WP. A meta-analysis on the relationship between self-reported presence and anxiety in virtual reality exposure therapy for anxiety disorders. PLoS ONE 2014;9(5):e96144. [doi: 10.1371/journal.pone.0096144] [Medline: 24801324]
34. Grover S, Sarkar S, Gupta G, et al. Factor analysis of symptom profile in early onset and late onset OCD. Psychiatry Res 2018 Apr;262:631-635. [doi: 10.1016/j.psychres.2017.10.006] [Medline: 29033063]
35. du Mortier JAM, Visser HAD, van Balkom A, et al. Examining the factor structure of the self-report Yale-Brown Obsessive Compulsive Scale Symptom Checklist. Psychiatry Res 2019 Jan;271:299-305. [doi: 10.1016/j.psychres.2018.11.042] [Medline: 30521999]
36. Bloch MH, Landeros-Weisenberger A, Rosario MC, Pittenger C, Leckman JF. Meta-analysis of the symptom structure of obsessive-compulsive disorder. AJP 2008 Dec;165(12):1532-1542. [doi: 10.1176/appi.ajp.2008.08020320]
37. Moritz S, Jelinek L, Hauschildt M, Naber D. How to treat the untreated: effectiveness of a self-help metacognitive training program (myMCT) for obsessive-compulsive disorder. Dialogues Clin Neurosci 2010;12(2):209-220. [doi: 10.31887/DCNS.2010.12.2/smoritz] [Medline: 20623925]
38. Oelkers C, Hautzinger M. Zwangsstörungen—Kognitiv-Verhaltenstherapeutischen Behandlungsmanual, 2nd edition: BELTZ; 2013.
39. Craske MG, Treanor M, Conway CC, Zbozinek T, Vervliet B. Maximizing exposure therapy: an inhibitory learning approach. Behav Res Ther 2014 Jul;58:10-23. [doi: 10.1016/j.brat.2014.04.006] [Medline: 24864005]
40. Goodman WK, Price LH, Rasmussen SA, et al. The Yale-Brown Obsessive Compulsive Scale. II. Validity. Arch Gen Psychiatry 1989 Nov;46(11):1012-1016. [doi: 10.1001/archpsyc.1989.01810110054008] [Medline: 2510699]
41. Hand I, Büttner-Westphal H. Die Yale-Brown Obsessive Compulsive Scale (Y-BOCS): ein halbstrukturiertes interview zur beurteilung des schweregrades von denk- und handlungszwängen [The Yale-Brown Obsessive Compulsive Scale (Y-BOCS): a semi-structured interview for assessing the severity of obsessive thoughts and compulsive behaviors]. Verhaltenstherapie 1991;1(3):223-225. [doi: 10.1159/000257972]
42. Jacobsen D, Kloss M, Fricke S, Hand I, Moritz S. Reliabilität der deutschen version der Yale-Brown Obsessive Compulsive Scale [Reliability of the German version of the Yale-Brown Obsessive Compulsive Scale]. Verhaltenstherapie 2003;13(2):111-113. [doi: 10.1159/000072184]
43. Foa EB, Huppert JD, Leiberg S, et al. The Obsessive-Compulsive Inventory: development and validation of a short version. Psychol Assess 2002 Dec;14(4):485-496. [Medline: 12501574]
44. Gönner S, Leonhart R, Ecker W. Das zwangsinventar OCI-R—die deutsche version des Obsessive-Compulsive Inventory-Revised—ein kurzes selbstbeurteilungsinstrument zur mehrdimensionalen messung von zwangssymptomen [The OCI-R Compulsion Inventory—the German version of the Obsessive-Compulsive Inventory-Revised—a brief self-report instrument for the multidimensional measurement of compulsive symptoms]. Psychother Psychosom Med Psychol 2007:395-404. [doi: 10.1055/s-2007-970894] [Medline: 17590836]
45. Beck AT, Steer RA, Brown GK. Beck Depression Inventory, 2nd edition: The Psychological Corporation; 1996.
46. Richter P, Werner J, Heerlein A, Kraus A, Sauer H. On the validity of the Beck Depression Inventory. A review. Psychopathology 1998;31(3):160-168. [doi: 10.1159/000066239] [Medline: 9636945]
47. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
48. Hinz A, Klein AM, Brähler E, et al. Psychometric evaluation of the Generalized Anxiety Disorder Screener GAD-7, based on a large German general population sample. J Affect Disord 2017 Mar 1;210:338-344. [doi: 10.1016/j.jad.2016.12.012] [Medline: 28088111]
49. Obsessive Compulsive Cognitions Working Group. Development and initial validation of the Obsessive Beliefs Questionnaire and the interpretation of intrusions inventory. Behav Res Ther 2001 Aug;39(8):987-1006. [doi: 10.1016/S0005-7967(00)00085-1]
50. Ertle A, Wahl K, Bohne A, Moritz S, Kordon A, Schulte D. Dimensionen zwangsspezifischer einstellungen [Dimensions of compulsion-specific attitudes]. Z Klin Psychol Psychother 2008 Oct;37(4):263-271. [doi: 10.1026/1616-3443.37.4.263]
51. Obsessive Compulsive Cognitions Working Group. Psychometric validation of the obsessive belief questionnaire and interpretation of intrusions inventory—Part 2: Factor analyses and testing of a brief version. Behav Res Ther 2005 Nov;43(11):1527-1542. [doi: 10.1016/j.brat.2004.07.010]
52. The Whoqol Group. Development of the World Health Organization WHOQOL-BREF Quality of Life Assessment. Psychol Med 1998 May;28(3):551-558. [doi: 10.1017/S0033291798006667]

XSL•FO
RenderX

53. Jelinek L, Moritz S, Hauschildt M. Patients' perspectives on treatment with Metacognitive Training for Depression (D-MCT): results on acceptability. J Affect Disord 2017 Oct 15;221:17-24. [doi: 10.1016/j.jad.2017.06.003] [Medline: 28628763]
54. Jelinek L, Zielke L, Hottenrott B, et al. Patients' perspectives on treatment with metacognitive training for OCD. Z Neuropsychol 2018 Mar 1;29(1):20-28. [doi: 10.1024/1016-264X/a000211]
55. Schubert T, Friedmann F, Regenbrecht H. The experience of presence: factor analytic insights. Presence: Teleop Virt Environ 2001 Jun;10(3):266-281. [doi: 10.1162/105474601300343603]
56. Sheehan DV, Lecrubier Y, Sheehan KH, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 1998;59 Suppl 20:22-33. [Medline: 9881538]
57. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007 May;39(2):175-191. [doi: 10.3758/bf03193146] [Medline: 17695343]
58. Ong CW, Clyde JW, Bluett EJ, Levin ME, Twohig MP. Dropout rates in exposure with response prevention for obsessive-compulsive disorder: what do the data really say? J Anxiety Disord 2016 May;40:8-17. [doi: 10.1016/j.janxdis.2016.03.006] [Medline: 27061971]
59. Hayes AF. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach: Guilford Press; 2013.
60. Farris SG, McLean CP, Van Meter PE, Simpson HB, Foa EB. Treatment response, symptom remission, and wellness in obsessive-compulsive disorder. J Clin Psychiatry 2013 Jul;74(7):685-690. [doi: 10.4088/JCP.12m07789] [Medline: 23945445]
61. Havnen A, Hansen B, Öst LG, Kvale G. Concentrated ERP delivered in a group setting: an effectiveness study. J Obsessive Compuls Relat Disord 2014 Oct;3(4):319-324. [doi: 10.1016/j.jocrd.2014.08.002]
62. Hansen B, Kvale G, Hagen K, Havnen A, Öst LG. The Bergen 4-day treatment for OCD: four years follow-up of concentrated ERP in a clinical mental health setting. Cogn Behav Ther 2019 Mar;48(2):89-105. [doi: 10.1080/16506073.2018.1478447] [Medline: 30088441]
63. Craske MG, Kircanski K, Zelikowsky M, Mystkowski J, Chowdhury N, Baker A. Optimizing inhibitory learning during exposure therapy. Behav Res Ther 2008 Jan;46(1):5-27. [doi: 10.1016/j.brat.2007.10.003] [Medline: 18005936]
64. Miegel F, Jelinek L, Moritz S, Lohse L, Yassari AH, Bücker L. Mixed results for exposure and response prevention therapy in mixed reality for patients with contamination-related obsessive-compulsive disorder: a randomized controlled pilot study. J Clin Psychol 2023 Oct;79(10):2317-2336. [doi: 10.1002/jclp.23550] [Medline: 37317567]
65. Schulze D, Reuter B, Elsner B, Kathmann N. Measurement invariance widely holds for the Yale-Brown Obsessive Compulsive Scale. Eur J Psychol Assess 2025 Mar;41(2):126-136. [doi: 10.1027/1015-5759/a000788]
66. Fatori D, Costa DL, Asbahr FR, et al. Is it time to change the gold standard of obsessive-compulsive disorder severity assessment? Factor structure of the Yale-Brown Obsessive-Compulsive Scale. Aust N Z J Psychiatry 2020 Jul;54(7):732-742. [doi: 10.1177/0004867420924113]
67. Ojalehto HJ, Abramowitz JS, Hellberg SN, Buchholz JL, Twohig MP. Adherence to exposure and response prevention as a predictor of improvement in obsessive-compulsive symptom dimensions. J Anxiety Disord 2020 May;72:102210. [doi: 10.1016/j.janxdis.2020.102210] [Medline: 32208231]

Abbreviations

ANCOVA: analysis of covariance
BDI-II: Beck Depression Inventory-II
CAU: care as usual
CBT: cognitive behavioral therapy
CC: complete case
checkOCD: checking obsessive-compulsive disorder
conOCD: contamination obsessive-compulsive disorder
CONSORT: Consolidated Standards of Reporting Trials
ERP: exposure with response prevention
GAD-7: Generalized Anxiety Disorder Scale-7
IPQ: Igroup Presence Questionnaire
ITT: intention-to-treat
OBQ-44: Obsessive Beliefs Questionnaire-44
OC: obsessive-compulsive
OCD: obsessive-compulsive disorder
OCI-R: Obsessive-Compulsive Inventory—Revised
VERP: virtual exposure with response prevention
VR: virtual reality
VRET: exposure therapy in virtual reality
Y-BOCS: Yale-Brown Obsessive-Compulsive Scale

XSL•FO
RenderX

Edited by A Coristine; submitted 19.Jun.2025; peer-reviewed by B Reuter, M Nunez; revised version received 06.Jan.2026; accepted 20.Jan.2026; published 27.May.2026.

Please cite as:
Rolvien L, Jelinek L, Lohse L, Moritz S, Borgmann L, Gallinat J, Kühn S, Miegel F
Virtual Exposure With Response Prevention for Obsessive-Compulsive Disorder: Randomized Controlled Trial
J Med Internet Res 2026;28:e79326
URL: https://www.jmir.org/2026/1/e79326
doi:10.2196/79326

© Lara Rolvien, Lena Jelinek, Luzie Lohse, Steffen Moritz, Lisa Borgmann, Jürgen Gallinat, Simone Kühn, Franziska Miegel. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Automated Inference of Social Anxiety From Behavior in Social Virtual Reality: Cross-Sectional Observational Study

Gayoung Son[1], MSc; Marius Rubo[1], PhD

Cognitive Psychology, Perception and Research Methods, Institute of Psychology, University of Bern, Bern, Switzerland

Corresponding Author:
Gayoung Son, MSc
Cognitive Psychology, Perception and Research Methods
Institute of Psychology
University of Bern
Fabrikstrasse 8
Bern, 3012
Switzerland
Phone: 41 31 684 40 42
Email: gayoung.son@unibe.ch

Abstract

Background: Social anxiety often manifests through behaviors such as reduced gaze to the eyes and lower speech volume. While these markers have been examined in face-to-face interactions, large-scale assessments remain challenging. Social virtual reality (VR) offers a promising alternative by enabling naturalistic interactions in which behavior can be captured at scale, but it remains unclear if people show naturalistic behavior in such artificial environments.

Objective: We examined whether behavioral and physiological markers associated with social anxiety in real-life interactions similarly emerge in dyadic social VR. We additionally examined whether these patterns overlap with patterns linked to the broader constructs of psychopathology and verticality.

Methods: In this cross-sectional study, 128 participants (105 females, 22 males, 1 diverse; age 18-51 years; mean age 22.60, SD 3.57 years), recruited from a university student population, engaged in a 30-minute avatar-mediated dyadic conversation in social VR while physically located in separate rooms. We assessed gaze toward the partner's eyes, smiling, and speaking behavior by using the VR headsets' built-in eye trackers, face trackers, and microphones, and assessed high-frequency heart rate variability (HF-HRV) by using an electrocardiogram.

Results: Relationships between traits and behavioral and physiological measures were analyzed using linear mixed-effects models (α=.05). Social anxiety was linked to reduced gaze toward the partner's eyes while speaking (β=–.20, 95% CI –0.35 to –0.04; t_{126}=–2.51; P=.01), quieter speech (β=–.18, 95% CI –0.35 to –0.01; t_{126}=–2.12; P=.04), and reduced HF-HRV (β=–.23, 95% CI –0.39 to –0.08; t_{119}=–3.00; P=.003). These findings were not entirely specific to social anxiety, as Pearson correlations revealed similar patterns for social anxiety and psychopathology (r=0.94, 95% CI 0.75-0.99; t_7=7.61; P<.001), whereas verticality was linked to an overall opposite pattern (social anxiety: r=–0.92, 95% CI –0.98 to –0.65; t_7=–6.14; P<.001; psychopathology: r=–0.83, 95% CI –0.96 to –0.38; t_7=–3.98; P=.005).

Conclusions: In dyadic interactions in social VR, social anxiety was associated with behavioral and physiological modulations similar to those observed in face-to-face interactions, indicating heightened social stress and submissiveness even in avatar-mediated communication. Patterns were similar for heightened psychopathology and reversed for verticality, indicating that these traits may lie on a shared social-behavioral spectrum. Extending previous research focused on face-to-face interactions or reactions toward artificial agents displayed in VR, this study is the first to provide a comprehensive account of the behavioral and physiological modulations associated with social anxiety in avatar-based human-human interactions. Since social VR setups allow researchers to assess a rich set of behavioral data as a byproduct of the setup's core functionality, the technique opens novel possibilities to detect social stress, track therapeutic progress, or tailor interventions to individual behavior when interactions take place in VR.

(J Med Internet Res 2026;28:e79147) doi:10.2196/79147

XSL•FO
RenderX

KEYWORDS

eye gaze; interaction behavior; psychopathology; social anxiety; social virtual reality; speaking behavior; verticality

Introduction

Overview

Social anxiety is a psychological trait deeply rooted in social functioning, shaping how individuals perceive and behave in social situations, often manifesting in distinct behavioral and physiological patterns during social interactions [1]. It is widely distributed across the population, with higher levels linked to increased emotional distress and impaired social functioning. Individuals with severe expressions of social anxiety can meet the diagnostic criteria for social anxiety disorder, with a lifetime prevalence of 13% and a 12-month prevalence of 8% among adults [2]. Even moderate, subclinical levels are common and linked to elevated comorbidity and reduced quality of life [3]. Capturing how social anxiety is behaviorally expressed in social interactions is essential for understanding its impact and identifying individuals at risk.

Research has identified a set of robust behavioral and physiological markers associated with social anxiety [4], often conceptualized as safety behaviors—coping mechanisms aimed at preventing negative social outcomes [1,5]. These include gaze avoidance [6], which has been shown to increase during speaking phases in individuals with higher social anxiety [7,8]. However, not all studies have observed this phase-specific pattern [9,10], suggesting that gaze behavior may vary depending on situational factors and may reflect context-sensitive manifestations of underlying traits.

Changes in facial expression have also been observed. In face-to-face interactions, individuals with higher social anxiety have been found to smile more [11], in some cases specifically mimicking the partner's polite smiles [12], which has been suggested to reflect appeasement strategies aimed at avoiding negative evaluation. Overall, speaking behavior is also affected: social anxiety has been associated with the tendency to speak less [13], interrupt less often [14], speak more quietly, and pause longer between speaking turns [15], though these associations may be modulated by gender or context [16]. Together, these patterns suggest a tendency to limit social visibility and reduce opportunities for self-expression, possibly as a means of avoiding evaluation. In parallel, physiological studies consistently report elevated heart rate and reduced high-frequency heart rate variability (HF-HRV) in socially anxious individuals [17,18]—markers of reduced parasympathetic activity and emotional self-regulation [19,20].

While social anxiety has frequently been linked to specific behavioral markers, it is increasingly recognized that these expressions are not exclusive to the trait. Psychological traits often overlap, and patterns attributed to social anxiety may reflect broader constructs. As an internalizing condition, social anxiety shows behavioral similarities with other anxiety and mood syndromes (eg, increased gaze avoidance in depression [21]), but also with externalizing disorders (eg, reduced HF-HRV in substance use disorders [22]). These overlaps align with the notion of a general factor of psychopathology (*p*-factor), which reflects shared variance across mental disorders and has been proposed as a transdiagnostic dimension of psychopathology severity [23,24].

Similarly, verticality—a trait reflecting perceived dominance and social rank—has been linked to reduced social anxiety [25]. Individuals high in social anxiety have been shown to display more low-verticality behaviors such as gaze aversion and reduced speech [4,15,26,27], which are interpreted as submissive strategies for reducing ostracism [28-30]. In contrast, higher verticality is associated with dominant behaviors such as increased gaze, louder speech, and more frequent interruptions [31,32], suggesting an inverse relationship between these constructs.

Notably, verticality itself shows conceptual and empirical overlap with broader psychopathology, as internalizing profiles are associated with perceived inferiority [33], while externalizing traits align with dominance-seeking behavior [25]. To this end, this study assessed a range of psychological traits—including social anxiety, general psychopathology, and verticality—to examine both their shared behavioral expressions and trait-specific associations.

While these behavioral and physiological markers have been extensively researched, capturing them reliably in naturalistic settings remains a significant methodological challenge [34]. Traditional laboratory paradigms—such as eye-tracking studies using static images or videotaped face-to-face interactions—offer experimental control but have typically required restricted participant movement and relied on time-intensive manual coding [35]. To overcome these limitations, researchers have increasingly turned to digital phenotyping, which leverages passive data from smartphones to infer psychological traits [36]. For example, mobility and communication patterns have been shown to predict social anxiety symptom severity [37], while a recent systematic review found that smartphone sensor data can detect behavioral patterns linked to stress, anxiety, and mild depression in student populations [38].

While such findings demonstrate the growing use of mobile sensing, these approaches primarily rely on indirect proxies—such as phone use, physical activity, or ambient audio—that are often captured outside interactive contexts. Moreover, technical challenges such as inaccurate readings when phones are stowed, or poor audio quality in noisy environments, limit the reliability of such data [38].

Critically, smartphone sensing lacks access to the social interactive behavioral cues—such as gaze coordination, conversational timing, or facial expressivity—that are central to social anxiety in interpersonal contexts. For instance, reliably determining whether someone is making eye contact while speaking requires precise tracking of eye movements and speaking behavior of both partners—capabilities that remain difficult to scale during face-to-face interactions.

XSL•FO
RenderX

Social virtual reality (VR) directly addresses these limitations by combining immersive, ecologically valid interaction with high experimental control [39]. Participants engage through avatars that, in the most modern installments, mirror their gaze, facial expressions, gestures, and speech in real time, enabling naturalistic social behavior to be studied under replicable conditions [40,41]. Our previous publication based on the same dataset [41] documented natural interaction behavior at the group level (eg, more gaze to the eyes while listening compared to speaking; more gaze to the eyes at the end of a speaking turn compared to the beginning of a speaking turn; more smiling and interrupting in small-talk compared to more serious conversations; and preserved interaction patterns in anonymity). Building on that work, this study maps the links between interindividual differences in behavior, physiology, and psychological traits, providing detailed insights into expressions of clinically relevant personal characteristics.

Although recent studies have increasingly leveraged VR to examine ecologically valid expressions of behavior and physiological responses [42,43], the technology was traditionally used in contexts where participants interacted alone or with virtual agents [44]. Early work already demonstrated that interpersonal distance toward virtual agents with differential gaze behavior mirrored real-life proxemic behavior [45]. More recent research has begun to focus on human-human interaction within virtual environments [46], with studies investigating gaze avoidance [47], physiological synchrony [48], and facial expressions [49] in dyadic social VR settings, as well as speaking behavior and social coordination in larger group interactions [50]. Crucially, social VR enables the automated collection of multimodal behavioral data in ways comparable to established single-user techniques [51,52], highlighting its promise as a scalable platform for studying social interaction behavior. Our study contributes to this growing line of research by integrating behavioral and physiological measures within a single analytical framework to provide a more comprehensive account of real-time interpersonal dynamics in social VR.

Objective

We assessed whether behavioral alterations associated with social anxiety (but also related constructs such as general psychopathology and verticality) can be observed in interactions in social VR, which allows for naturalistic social interactions while enabling automated and scalable measurement of interaction behavior. To the best of our knowledge, our study is the first to simultaneously assess multiple social behaviors and physiological signals during real-time dyadic interactions in VR. Participants engaged in a dyadic conversation while embodying virtual avatars that displayed their eye movements, facial expressions, and gestures in real time. We specifically assessed gaze, speaking, and smiling behavior, which are processed automatically as part of the social VR environment, but also heart rate and HF-HRV, a robust physiological indicator of psychological stress and anxiety. Specifically, the hypotheses are as follows:

- Hypothesis 1: We expected different self-report measures of social anxiety, as well as different measures of psychopathology and verticality, to load on a common underlying factor.
- Hypothesis 2: We expected higher social anxiety (and similarly, higher psychopathology but lower verticality) to be associated with behavioral alterations:
  - Hypothesis 2.1: less gaze at the partner's eyes while speaking and listening
  - Hypothesis 2.2: heightened overall smiling behavior
  - Hypothesis 2.3: reduced overall speaking time, fewer interruptions during the partner's speech, longer pauses (ie, gaps before starting to speak after the partner finished their turn), and loudness of speech
- Hypothesis 3: We expected higher social anxiety (and similarly, higher psychopathology but lower verticality) to be associated with alterations in physiological functioning, specifically, heightened heart rate and reduced HF-HRV.

Methods

Participants

A total of 128 participants (105 female, 22 male, 1 diverse; aged 18-51 years; mean age 22.60, SD 3.57 years) participated in a dyadic interaction in the social VR environment. This predetermined sample size allows for an 85% chance of detecting a correlation of r=0.26 with α=.05 in a repeated-measures correlation involving 3 measures. Most participants (n=122, 95.31%) were undergraduate psychology students at the University of Bern, 118 with Swiss nationality, 8 with German nationality, and 14 from other European or non-European countries (multiple nationalities are possible). Exclusion criteria included neurological conditions, use of central nervous system–acting substances, and severe visual impairments requiring glasses that are incompatible with VR equipment. Within this sample, none of the individuals had more than a brief acquaintance with one another (n=5, 3.91% participants). To preserve conceptual clarity and avoid overlap between constructs, we kept measures distinct across factors. We included additional behavioral variables (ie, smiling and heart rate) and examined eye gaze separately during speaking and listening phases to offer a broader understanding of participant behavior. Data and analyses are accessible at [53].

Ethical Considerations

This study was approved by the institutional review board of the University of Bern (reference 2023-09-01) and was preregistered [54]. Prior to giving consent, all participants were verbally informed of the research aims, procedure, the voluntary nature of participation, and the right to withdraw at any point in the study freely and without consequences. They were also informed that they might experience symptoms of VR sickness and were instructed to remove the headset and inform the experimenter if this occurred; however, no such instances were reported. All collected data were anonymized and managed securely. Participants received 2 course credits as compensation. All figures containing persons depict example scenes performed by the study authors and do not include actual participants.

XSL•FO
RenderX

Procedure

This cross-sectional study followed the reporting guidelines for the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) Statement for observational studies [55]. Participants were recruited through the university's participant recruitment system between November 2023 and February 2024. Throughout the entire study (lasting approximately 1.5 hours), participants were seated in separate rooms and interacted solely through the social VR software (Figures 1 and 2), which was custom-developed in Unity 3D [56] for this study. Participants additionally arrived and departed separately in order to ensure that they did not see each other at any point. The virtual environment was displayed using Meta Quest Pro headsets [57] and ran on Windows 11 desktop laptops (Intel Core i9-13900HX CPU, 32 GB RAM, and NVIDIA GeForce RTX 4070 GPUs). Validation of study measures is reported in Son and Rubo [41]. A detailed description of the software architecture is provided in Rubo [58]. Subsequently, participants were equipped with surface electrodes to obtain an electrocardiogram throughout the entire study.

Figure 1. The experimental setup seen from a third-person perspective. Participants meet as avatars in virtual reality (top image) while physically located in separate rooms (bottom images). Avatars convey participants' movements, eye gaze, and facial expressions in real time, and voice is transmitted via voice chat. The virtual environment's position and orientation are calibrated so that the position of a virtual table matches the position of a physical table for both participants. All dyads conversed on predefined topics for a total of 30 minutes. People displayed in the images are the study authors demonstrating the experimental setup and procedure.

XSL•FO
RenderX

Figure 2. First-person perspective of a participant during the virtual reality interaction. Each participant views the scene from the perspective of their avatar's eyes, allowing for technology-mediated direct eye contact between participants.

Before the interaction, all participants completed an initial set of questionnaires, which included the self-reported measures of interest. Once the questionnaires were completed, participants were introduced to the virtual environment. This involved headset calibration, a 9-point eye-tracking calibration, and adjustments to the interpupillary distance. Participants also aligned their seating height and orientation to ensure that the virtual table was positioned to correspond precisely with the physical table. Next, participants observed their avatar in a virtual mirror, allowing them to see their avatar's full range of movements. Each avatar displayed the head and eye movement, blinking, mouth movement, facial expressions, and gestures of the participants in real time. Avatars were generated using Character Creator 4 along with the Headshot extension [59] and were configured to render only the head, upper torso, and hands. Each avatar featured extensive facial details and was individually adjusted by a research assistant for eye color, facial shape, hair, and body size (for more details on avatar creation, see Multimedia Appendix 1 [41,59-66]).

In addition, half of the dyads interacted while embodying self-similar avatars generated from participants' portrait photos, and the other half interacted while embodying gender-matching avatars created using generic images of nonexisting individuals from a random pool of 6 male and 6 female avatars. As detailed in a separate report [41], this condition investigated the effects of avatar appearance on social interactions. Since no behavioral differences were identified, and participants furthermore reported no differences in satisfaction with their avatars or in the attractiveness of their avatars, conditions were aggregated in this study.

The experimenter then refined the avatar's size, head orientation, and eye-tracking calibration based on participant feedback. Once the setup was complete, participants were instructed that the study would include three 10-minute conversations on predefined topics (small talk, personal talk, and opinion talk; Multimedia Appendix 1) presented in randomized order. Participants were asked to remain on topic for the duration of the task. The experimenter then exited the room, and participants were connected to their partners within the VR environment. During the interaction, participants were guided by prerecorded computerized voices, and each discussion prompt and the remaining time were displayed on a virtual tablet. After the interaction, they completed control questions on the quality of the interaction (eg, VR sickness symptoms or satisfaction with the avatar), which are described in a separate report [41]. Analyses reported here refer to interactions as a whole.

Self-Report Measures

Social Interaction Anxiety

Social interaction anxiety symptoms were examined using the Social Interaction Anxiety Scale (SIAS [67]; German version [68]). The SIAS assesses anxiety in social interactions in 20 items, rated on a 5-point Likert scale (0=not at all, 4=extremely). Internal consistency measured with Cronbach α was 0.91.

Fear of Positive and Negative Evaluation

Fear of positive and negative evaluation is a core symptom of social anxiety, consisting of fears of both negative and positive judgment from others [69]. Fear of positive evaluation was assessed using the German translation of the Fear of Positive Evaluation Scale [70,71], consisting of 10 items asking to what degree participants experienced positive evaluations as anxiety-inducing (α=.81). Items are rated on a 10-point Likert scale (0=not at all true, 9=very true). Fear of negative evaluation was assessed using the revised Brief Fear of Negative Evaluation Scale [72] (German version: [73]). Participants rate 12 items on a 5-point Likert scale (1=not at all characteristic of me, 5=extremely characteristic of me). Cronbach α was 0.94.

XSL•FO
RenderX

Social Comparison

Social Comparison was evaluated with the Social Comparison Scale [74] (German translation [75]). This scale measures upward social comparison evaluations across 11 different comparison dimensions. Participants respond to bipolar items on a scale from 1 to 10 (eg, In relationship to others I feel:" 1=incompetent, 10=more competent). Higher scores reflect a more favorable self-assessment relative to others, while lower scores reflect a less favorable self-assessment (α=.87).

Trait Anxiety

Trait Anxiety was assessed using the Trait subscale of the State-Trait-Anxiety Scale [76,77]. Participants rate 10 items ranging from 1=not at all to 4=extremely (α=.81).

Stress

Stress was measured using the Perceived Stress Scale-10 [78]. The 10 items assess the degree of stress experienced in the past month (1=never, 5=very often). Cronbach α was 0.81.

Depression Level

Depression level was assessed using the Beck Depression Inventory-Fast Screen [79] (German version [80]). Participants respond to 7 items assessing various depressive symptoms (sadness, pessimism, past failure, loss of pleasure, self-dislike, self-criticalness, and suicidal ideation). The scores vary between 0 and 21, where higher scores signify more severe depressive symptoms (α=.75).

Impulsivity

Impulsivity was assessed using the shortened Barratt Impulsiveness Scale-15 [81] (German version [82]) consisting of 15 items (α=.76). This scale measures 3 subtraits (motor, attentional, and nonplanning) on a 5-point Likert scale (1=never, 4=always).

Empathizing

Empathizing was measured with the German translation of the Empathy Quotient-10 [83]. This questionnaire is associated with autistic traits and also exhibits a broad range within the general population [83]. The 10 items are scored on a 5-point Likert scale (1=strongly disagree, 4=strongly agree). Cronbach α was 0.70.

Neuroticism and Extraversion

Neuroticism and extraversion were assessed using the respective subscales of the Big Five Inventory [84] (German version [85]). Neuroticism and extraversion are highly predictive of depressive and anxiety disorders, as previous meta-analytic research has shown that individuals in various diagnostic groups exhibit high levels of neuroticism and low levels of extraversion [86,87]. Both subscales consist of 4 items scored by a 5-point Likert scale (1=strongly disagree, 5=strongly agree). Cronbach α for neuroticism was 0.74, and for extraversion, 0.82.

Self-Esteem

Self-esteem was assessed with the Rosenberg Self-Esteem Scale [88] (German version [89]). Participants rate 10 items on a 6-point Likert scale (1=not at all, 6=very much). Cronbach α was 0.90.

Subjective Social Status

Subjective social status was measured using the MacArthur Subjective Social Status Scale [90] (German version [91]). Participants rate their socioeconomic position (income, education, and occupation) relative to their community on a scale of 1 to 10, visualized as a ladder with 10 rungs.

Power

Power refers to one's belief in having influence over others. It was assessed using the German translation of the Personal Sense of Power Scale [92,93], consisting of 6 items on a 7-point Likert scale (1=strongly disagree, 7=strongly agree). Cronbach α was 0.76.

Dominance and Prestige

Dominance and Prestige are 2 facets of the social rank attainment model, which posits that social hierarchy is achieved through force (dominance) or competence (prestige) [94]. These were assessed using the Dominance and Prestige Scale [95], which is a previously validated scale consisting of 2 subscales with 8 items measuring dominance (α=.77) and 9 items measuring prestige (α=.69), ranging from 1=not at all to 7=very much.

Agency

Agency was evaluated using the German adaptation of the Inventory of Interpersonal Problems-32 [96]. This tool comprises 32 items addressing 8 dimensions of interpersonal issues (α=.82), categorized into 2 overarching dimensions of agency and communion [97]. Participants indicate the extent to which each interpersonal challenge distresses them on a 5-point Likert scale (0=not at all, 4=extremely). For this study, the agency subscale score was used ([96]). This score specifically addresses interpersonal difficulties related to dominance, agency, and control [97].

Table 1 presents summaries of self-report measures. Responses in anxiety and depression measures showed that 27 (21.09%) participants exceeded the suggested diagnostic threshold of 34 for probable social anxiety [98] and 26 (20.31%) participants met the proposed cutoff score of 4 for mild depression [99].

XSL•FO
RenderX

Table 1. Descriptive statistics for self-report measures of participants completed before interaction, divided into measures for social anxiety, general psychopathology, and verticality (N=128).

Measure	Mean (SD; range)	Possible range[a]
Social anxiety		
Social interaction anxiety	25.20 (11.58; 3 to 53)	0 to 80
Fear of negative evaluation	35.45 (10.16; 15 to 58)	12 to 60
Fear of positive evaluation	22.69 (12.21; 0 to 54)	0 to 90
Social comparison	65.79 (11.45; 40 to 97)	40 to 100
General psychopathology		
Trait anxiety	20.32 (4.36; 11 to 33.5)	10 to 40
Stress	27.24 (4.99; 18 to 42)	10 to 50
Depression levels	2.20 (2.32; 0 to 9.4)	0 to 21
Impulsivity	2.19 (0.38; 1.33 to 3.2)	1 to 4
Empathizing	12.21 (3.13; 3 to 19)	0 to 20
Neuroticism	3.20 (0.81; 1.75 to 5)	1 to 5
Extraversion	3.50 (0.84; 1.5 to 5)	1 to 5
Self-esteem	4.67 (0.77; 2.28 to 5.9)	1 to 6
Verticality		
Social status	5.88 (1.22; 3 to 10)	1 to 10
Power	29.52 (4.27; 17 to 38)	6 to 42
Dominance	2.59 (0.87; 1 to 5)	1 to 7
Prestige	4.91 (0.65; 3 to 6.33)	1 to 7
Agency	–0.79 (0.50; –1.93 to 0.66)	–2.90 to 2.90

[a]Minimum and maximum scores that can be reached on each measure.

Behavioral Measures

We examined several behavioral and physiological markers during the interaction. For eye gaze, we followed conventional thresholds [100,101] and excluded eye-tracking data recorded during blinks as well as 100 ms before and after each blink to avoid blink-related artifacts (average data loss 3.85%, SD 2.81% across participants). Missing data were interpolated using the Piecewise Cubic Hermite Interpolating Polynomial and smoothed with a second-order Butterworth filter (cutoff frequency=10 Hz). Gaze to the partner's eye region was defined as visual attention directed toward the partner's eye region, operationalized as within 2 degrees of visual angle around the center of a partner's eye [102,103]. To account for common minor eyesight irregularities (eg, strabismus), gaze was coded to be directed toward a partner's eyes if at least one of the participant's eyes or a ray combining the 2 eyes was directed toward one of the partner's eyes. Following previous research on differential gaze behavior during conversational roles [104], gaze was analyzed separately for speaking and listening phases.

Speaking behavior was defined as continuous time periods during which the sound amplitude recorded by the head-mounted display's microphone exceeded a threshold of 2. Speech segments shorter than 1 second were classified as backchanneling vocalizations (eg, "ok" or "hmm") and not counted as speaking turns, consistent with prior research on nonverbal behavior [105,106]. Speaking turns of one person could include brief pauses of up to 2 seconds before being classified as 2 consecutive turns from the same person. The threshold was established based on Heldner and Edlund [107], indicating that typical within-speaker pauses last around 1 second, and was specifically selected to avoid classifying brief hesitations as new turns or fusing separate turns. Furthermore, moments when 1 participant remained silent while their partner spoke were categorized as listening [108]. Instances of overlapping speech were classified as interruptions, attributed to the participant who began speaking while previously listening [109].

Smiling was measured using the head-mounted display's built-in face tracking, which detects facial muscle activity based on action units defined by the Facial Action Coding System [110]. The intensity of smiling behavior was calculated by averaging the activation values from both sides of the mouth, yielding a score between 0 (no smile) and 100 (maximum smile expression).

Electrocardiograms were recorded using BItalino devices [111] at a sampling rate of 1000 Hz. QRS complexes (ventricular complex) were first detected automatically using an algorithm described by Elgendi et al [112]. Data were visually inspected individually and corrected manually if needed. Equidistantly sampled RR interval (duration between successive heartbeats)

XSL•FO
RenderX

series were constructed using cubic spline interpolation. Heart rate (in beats per minute) was defined as 60 divided by the mean of the RR interval series (in seconds). To obtain HF-HRV, RR interval time series were first resampled at 4 Hz using bandlimited interpolation. Power spectral density was estimated using the Welch method, with 30-second Hann-windowed segments and 50% overlap. Each segment was processed using the Fast Fourier Transform, and only positive frequencies were retained. HF-HRV was quantified by integrating the power spectral density within the high-frequency band (0.15-0.4 Hz) and log-transforming the resulting power values.

As reported in Son and Rubo [41], participants spent an average of 23.05% (SD 11.96%) of their speaking time and 49.45% (SD 16.96%) of their listening time gazing toward their partner's eyes. The mean proportion of speaking time was 50.73% (SD 9.90%). The mean proportion of time spent interrupting was 4.82% (SD 3.79%), and the average length of pauses before responding to the partner was 1.36 (SD 0.70) seconds. Average speech loudness was 2.70 (SD 0.77), measured relative to the maximum microphone recording capacity (100). The average smiling magnitude was 13.40 (SD 7.99) over the entire conversation period. The mean heart rate was 79.61 (SD 11.95) beats per minute, and the average HF-HRV was 6.36 (SD 0.96).

Data Analysis

Analyses were performed using R software (version 4.3.2; R Foundation for Statistical Computing) [113]. In the first step, principal component analysis was applied to self-report measures related to social anxiety, general psychopathology, and verticality (dominance and power), respectively. Principal component analysis is a widely used method to derive a common component from variables, which maximizes the total variance [60]. Focusing on the dominant components captures the most significant sources of variability in the data, providing a parsimonious representation of the underlying constructs and helping to avoid overly specific interpretations [23].

To further validate these components, two-sided Pearson correlations were conducted between the extracted components and the individual self-report measures that contributed to each factor. For these analyses, measures were aggregated across the entire conversation period. As expected, measures of social anxiety, psychopathology, and verticality were each correlated with their respective common factors (Multimedia Appendix 1). Consequently, subsequent analyses were carried out using these first dominant factors. The effect of each factor on behavioral and physiological measures was examined using mixed-effects models, with random intercepts for participant ID as well as random slopes for topics nested within ID to account for within-subject variability. For all variables, values beyond 3 SD from the mean were set to 3 SD from the mean and *z*-standardized prior to analyses. The percentage of corrected outliers ranged from 0.78% to 3.13% (Multimedia Appendix 1). The significance of results was determined with α set to 5%. *P* values were adjusted for multiple comparisons using the Benjamini-Hochberg false discovery rate (FDR) procedure.

To reduce potential bias, analyses were preregistered, data acquisition was standardized, and preprocessing algorithms were held identical for all participants. Furthermore, mixed-effects models were used to account for within-subject and sampling variability.

Results

Effect of Psychological Traits on Behavior

Principal component analyses were conducted separately for self-report measures related to social anxiety, general psychopathology, and verticality, which reflects dominance and rank. Each analysis yielded a dominant first component (Multimedia Appendix 1). All factors showed theoretically consistent and large associations with their corresponding measures, confirming their suitability as parsimonious representations of the underlying traits (Figure 3). Detailed results are provided in Multimedia Appendix 1.

The effect of social anxiety, general psychopathology, and verticality on the assessed behavioral measures was examined in linear mixed models. Figure 4 shows the β estimates and CIs for the relationships between factors and behaviors. A detailed summary of results is shown in Multimedia Appendix 1.

Figure 3. Pearson correlation coefficients between each first component of the principal component analysis (PCA) for social anxiety, general psychopathology, and verticality with self-report measures on individual scales. As expected, results show overall strong convergence between self-report measures relating to each of the 3 overarching trait constructs.

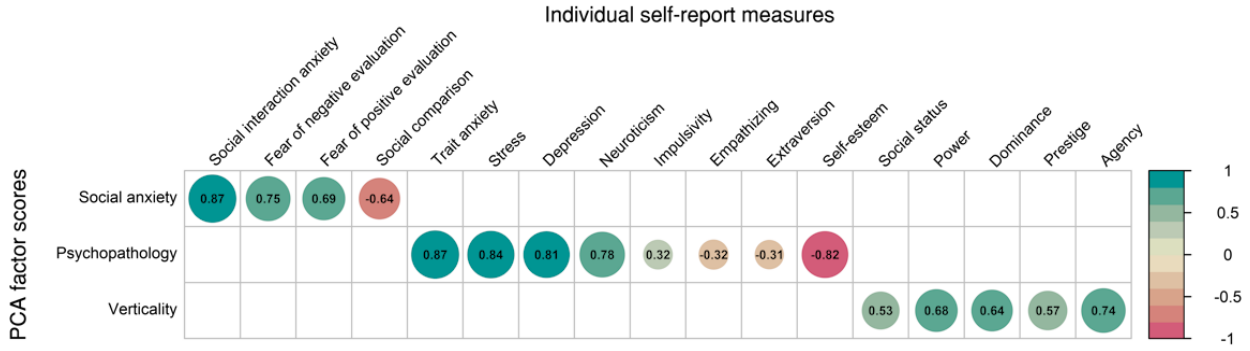

XSL•FO
RenderX

Figure 4. Associations between social anxiety, general psychopathology, and verticality with behavioral and physiological measures collected during the social virtual reality interaction. Values are standardized β estimates, indicating lower or higher measurements on behavioral or physiological metrics with higher trait expression. For example, higher social anxiety predicts reduced gaze to the partner's eyes while speaking. Lines represent 95% CIs. HF-HRV: high-frequency heart rate variability.

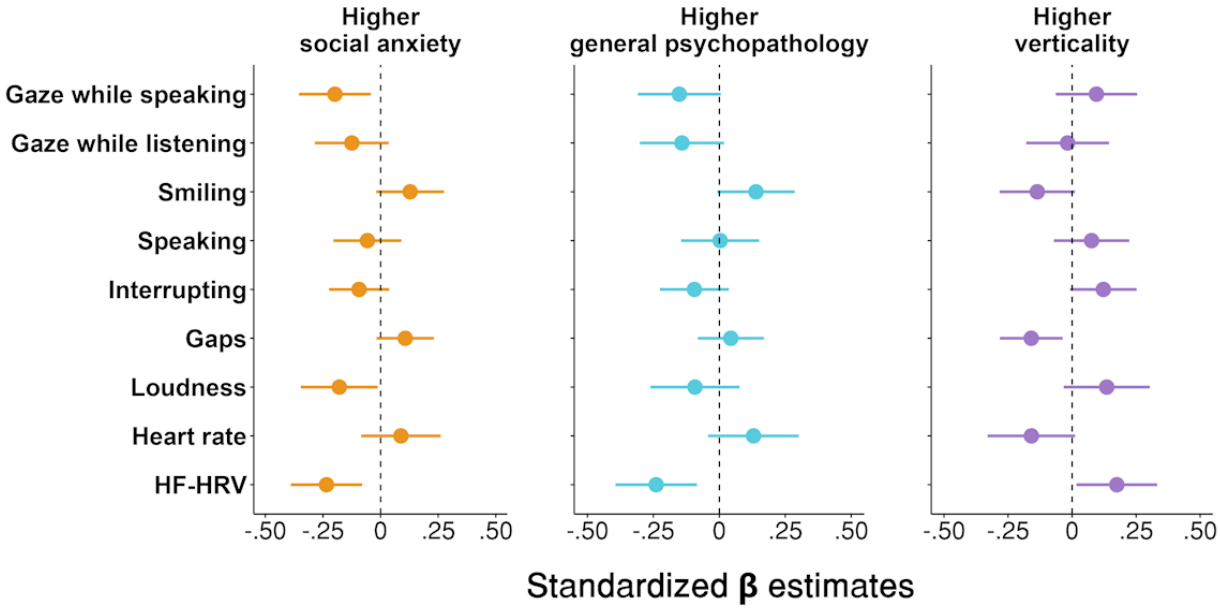

Higher social anxiety levels were associated with reduced gaze toward the partner's eyes during one's own speaking turn (β=–.20, 95% CI –0.35 to –0.04; t_{126}=–2.51; *P*=.01). The effect of underlying psychopathology on the amount of gaze toward the partner's eyes while speaking (β=–.15, 95% CI –0.31 to 0.01; t_{126}=–1.90; *P*=.06) and while listening was marginally significant (β=–.14, 95% CI –0.30 to 0.02; t_{126}=–1.75; *P*=.08). The association between social anxiety and psychopathology with smiling behavior was marginally significant, such that higher social anxiety (β=.13, 95% CI –0.02 to 0.28; t_{126}=1.72; *P*=.09) and higher psychopathology (β=.14, 95% CI –0.01 to 0.29; t_{126}=1.87; *P*=.06) was positively associated with smiling. In contrast, verticality was marginally significantly associated with less smiling behavior (β=–.14, 95% CI –0.28 to 0.01; t_{126}=–1.83; *P*=.07).

Social anxiety was furthermore associated with reduced speaking volume (β=–.18, 95% CI –0.35 to –0.01; t_{126}=–2.12; *P*=.04). The effect of social anxiety on gaps was marginally significant, such that higher social anxiety was associated with longer gaps between one's speaking turns (β=.11, 95% CI –0.02 to 0.23; t_{126}=1.70; *P*=.09). Verticality influenced gap duration, such that higher verticality was associated with shorter gaps before responding (β=–.16, 95% CI –0.28 to –0.04; t_{126}=–2.57; *P*=.01). Verticality was additionally marginally significantly associated with interrupting behavior, such that higher verticality was associated with more interrupting behavior (β=.12, 95% CI –0.03 to 0.25; t_{126}=1.86; *P*=.07). The effect of verticality on heart rate was marginally significant, such that individuals high in verticality displayed lower heart rate (β=–.16, 95% CI –0.33 to 0.01; t_{119}=–1.84; *P*=.07).

Social anxiety and psychopathology were associated with reduced HF-HRV (social anxiety: β=–.23, 95% CI –0.39 to –0.08; t_{119}=–3.00; *P*=.003; psychopathology: β=–.24, 95% CI –0.39 to –0.08; t_{119}=–3.07; *P*=.003), whereas higher verticality was associated with increased HF-HRV (β=.17, 95% CI 0.02 to 0.33; t_{119}=2.20; *P*=.03). After adjusting for multiple comparisons, the observed significant effects only held for social anxiety and psychopathology on HF-HRV.

Behavioral Similarities Between Social Anxiety, Psychopathology, and Verticality

We additionally compared similarities and differences in how social anxiety, psychopathology, and verticality related to behavior. Pearson correlations on the obtained β estimates were performed to examine the pattern of behavior between the 3 factors. The analysis revealed very strong correlations between the factors, such that the pattern of behavior in social anxiety was strongly positively associated with that in psychopathology (*r*=0.94, 95% CI 0.75 to 0.99; t_7=7.61; *P*<.001), and the pattern of behavior in verticality was highly negatively associated to that in social anxiety (*r*=–0.92, 95% CI –0.98 to –0.65; t_7=–6.14; *P*<.001) and psychopathology (*r*=–0.83, 95% CI –0.96 to –0.38; t_7=–3.98; *P*=.005).

Discussion

Principal Results

We examined whether previously reported associations between social anxiety and behavioral and physiological markers could be traced during real-time, dyadic interaction within social VR. This platform enables the automated assessment of fine-grained behavioral data during lifelike dyadic interactions, which has remained relatively difficult to assess in previous paradigms. We found that behavioral and physiological patterns previously associated with social anxiety also emerged in interactions in social VR, replicating patterns observed in face-to-face interactions. These associations were not exclusive to social

XSL•FO
RenderX

anxiety but were also found in relation to broader psychopathology and lowered verticality.

Findings suggested both shared and distinct behavioral patterns across psychological traits. Gaze avoidance was associated with social anxiety, replicating previous findings from face-to-face interaction studies [4,6] and from social VR studies involving agent-based interactions [114]. This effect was specific to speaking phases, consistent with prior work showing heightened gaze avoidance during speech, but not during listening [7,8]. While some studies have reported stronger associations with social anxiety during listening or found no clear phase-specific effects [9,10], these results support the view that gaze behavior in social anxiety may be particularly sensitive to moments of self-expression, when perceived social evaluation is likely heightened. In addition, reduced loudness was relatively specific to social anxiety, aligning with earlier research linking reduced loudness of speech to heightened nervousness in patients with social anxiety [15].

Although associations between social anxiety and behavior more consistently reached statistical significance compared to general psychopathology, the overall pattern of results was highly similar. Both gaze behavior during speaking and listening, as well as smiling, showed trends in the expected direction with higher general psychopathology. These findings suggest that certain interaction behaviors may also reflect broader psychological vulnerability in interpersonal settings. Furthermore, these findings may also reflect the heterogeneity of general psychopathology, which spans both internalizing and externalizing symptoms and may therefore give rise to less distinct patterns of social behavior. Future experimental studies that systematically manipulate the social-evaluative context could help determine whether gaze avoidance is a specific behavioral marker for social anxiety, or whether it generalizes to other dimensions of psychopathology [23].

Conversely, verticality, which encompasses constructs such as dominance and status, showed a partly complementary behavioral profile, marked by increased gaze during speaking (but not listening) and shorter turn-taking gaps—patterns previously linked to assertive or high-dominance behavior [31,32]. These results align with social rank theory [29], which posits that perceived social status regulates behavior in interpersonal contexts. Individuals higher in verticality may perceive less social threat and engage more confidently, reducing the need for avoidant or submissive strategies.

While these associations were theoretically coherent, most were small and did not remain significant after adjusting for multiple comparisons. Nonetheless, the pattern of results was consistent across measures and aligned with theoretical expectations. For instance, 6 of 36 bivariate associations were at least significant at $P<.05$ (expected by chance: 1.8), and 14 were at least marginally significant at $P<.10$ (expected by chance: 3.6). Importantly, all of these associations were in the expected direction, suggesting that the overall pattern aligns with hypothesized relationships.

Regarding physiological data, HF-HRV was predictive for all 3 traits, decreasing with higher social anxiety and general psychopathology, and increasing with verticality. As a physiological index of autonomic regulation and stress, reduced HF-HRV is broadly linked to impaired emotion regulation and social withdrawal [17,19,20,115]. In contrast to behavioral markers, this pattern suggests that HF-HRV may capture a general dimension of stress sensitivity, consistent with its proposed role as a transdiagnostic marker of stress [116,117].

Taken together, behavioral modulations in social anxiety and psychopathology were highly parallel, with verticality showing a largely inverted pattern. Both social anxiety and broader psychopathology were overall associated with an increase in interpersonal safety behaviors, particularly those expressed through submissive strategies such as reduced assertiveness and appeasing social cues [4,118]. These patterns are consistent with prior work linking these traits to diminished perceptions of social power [25,27]. Submissive behavior, in turn, has been consistently linked to reduced psychological well-being [119], reinforcing its clinical relevance as an important marker for heightened psychological vulnerability [27]. Conversely, verticality was associated with reduced submissive behavior, suggesting it may represent the opposing pole of a social-behavioral continuum relative to the submissive tendencies observed in social anxiety and broader psychopathology.

In addition, the strong behavioral convergence between social anxiety and general psychopathology lends support to the *p*-factor framework as a useful model for understanding shared variance in social behavior across psychological traits [23,24]. The behavioral overlap between social anxiety and general psychopathology indicates that the specificity of commonly used markers warrants further examination: rather than uniquely reflecting any single construct, interaction behavior may commonly indicate broader impairments in social functioning and well-being [120,121].

Implications

Our results highlight important implications. Compared to traditional eye-tracking paradigms [35] or smartphone phenotyping [38], our study demonstrates that social VR enables the synchronous collection of behavior during real-time interactions, enabling the study of social dynamics and mutual feedback that characterize human interactions. In social VR, participants can express and observe natural interaction behavior as it unfolds in a shared environment, creating conditions for more ecologically valid expressions of gaze, speech, and physiological responses [36]. This highlights this medium's potential as a powerful research tool for studying social processes and understanding psychological traits in socially embedded contexts.

Social VR presents promising opportunities for clinical research and novel forms of mental health care. Recent findings show that it is not only a growing digital medium, especially among younger users, but also a psychologically meaningful space for interactions. Supportive interactions in these environments have been linked to greater social presence, enhanced self-efficacy, and reduced loneliness [122]. As such, the technique may offer a valuable context for large-scale, ecologically valid monitoring of social behavior among multiple users.

XSL•FO
RenderX

Social VR also offers distinct advantages as a medium for delivering remote mental health care. Unlike conventional videoconferencing, it preserves key nonverbal cues such as eye contact and conversational timing, helping to overcome limitations such as reduced social presence and Zoom fatigue [123]. Moreover, it may enable novel therapeutic formats that are difficult to implement in traditional settings, such as fully anonymous group therapy sessions, which may lower barriers for individuals with social anxiety or those who fear stigmatization.

In addition, its capacity for automated data capture allows researchers to conduct passive, high-frequency assessments, which may be extended toward repeated interactions—such as multiple psychotherapy sessions or ongoing team meetings in virtual environments—facilitating the continuous monitoring of changes in well-being. Similar to passive smartphone phenotyping, which has demonstrated that passive sensor data can predict symptom changes in mental disorders [124], social VR offers new opportunities to assess social functioning in ecologically valid, interactive contexts. For example, fine-grained behavioral data collected in VR may aid in early risk detection, provide objective measures of therapeutic progress, or support the tailoring of interventions to individual behavioral patterns. Such data may be used to inform clinical decisions or even provide feedback within the VR setting itself [125].

At the same time, the technique's capability for unobtrusive, fine-grained data collection may also introduce significant security and ethical challenges. The passive and automated nature of data collection in these environments, which can be carried out by a server as well as any other computer in the network simulation, increases the risk of sensitive behavioral data being collected without the user's knowledge. While data collection in this study was conducted using an in-house software tool with no reliance on external services or internet connection, large-scale implementations of social VR in research contexts and beyond will be required to carefully consider data security principles and ethical implications of their use [39,126].

Limitations

We mapped a range of behavioral and physiological markers to interindividual differences in self-reported psychological traits in a naturalistic social VR interaction in 128 participants. Despite their strengths, 6 factors limit the generalizability of the findings.

First, the sample was composed of students from the German-speaking region of Switzerland, which may have introduced selection bias and limited the representativeness of the findings. Previous research indicates that social behaviors can vary across the lifespan; for example, avoidant behaviors such as gaze aversion tend to increase during adolescence [127]. Cultural factors may likewise influence nonverbal behavior and have been linked to differences in satisfaction and even health outcomes in cross-cultural clinical contexts [128]. As many aspects of nonverbal communication are culturally learned, the generalizability of our findings may decrease with greater cultural distance from the studied population [129]. At the same time, it was observed that interindividual differences in traits such as social anxiety often modulate modes of experience and behavior across population subgroups [130].

Second, although the sample showed substantial variability across behavioral and trait measures—with approximately 1 in 5 participants meeting thresholds for social anxiety disorder—the study lacked sufficient statistical power to examine gender differences in behavior, as the sample consisted predominantly of female participants. Exploratory analyses from a separate report [41] indicated that women smiled more frequently than men, consistent with prior findings in nonverbal behavior research [131]. However, the unequal group sizes in this study limit the interpretability of gender effects. Future research should therefore aim to systematically examine how demographic and cultural factors moderate behavioral dynamics in VR interactions to achieve more generalizability.

Third, the prevalence of mental disorders among university students is comparable to that in the general population [132], and psychological traits examined here—such as general psychopathology—are increasingly understood as dimensional rather than categorical [133]. Although some participants exceeded clinical thresholds for anxiety or depression, a clinically diagnosed sample was not included. It remains uncertain whether the behavioral modulations observed in this nonclinical sample would extend to full-blown mental disorders. In particular, stronger expressions of psychopathology may be associated with more specific modulations of interaction behavior than those observed here [134]. However, prior research has also shown that subclinical social anxiety typically exhibits behavioral patterns similar to those in social anxiety disorder, such as heightened anxiety in facial expressions [135]. Likewise, physiological indices such as heart rate have been shown to distinguish subclinical social anxiety during speech tasks in young adults [136]. These findings support the notion of social anxiety as a continuum, suggesting that the patterns observed here may represent expressions along the same underlying spectrum.

Fourth, while we included a broad range of self-report measures covering internalizing symptoms (eg, anxiety or depression) and select externalizing symptoms (eg, impulsivity or empathizing), other common externalizing symptoms such as substance abuse and aggressive behaviors were not assessed, as these may be less commonly observed in a student sample. This omission may have limited the scope of the general psychopathology construct captured. However, internalizing and externalizing psychopathology have been shown to be highly correlated across various age groups [137,138], and prior research has successfully modeled general psychopathology without including externalizing dimensions [139]. Nonetheless, future research should incorporate a broader range of symptom domains to more comprehensively assess the influence of psychopathology on social behavior.

Fifth, the study adopted a correlational design without a comparison condition in which participants interacted outside of VR. It therefore remains uncertain whether absolute behavioral levels in VR fully align with those observed in real-world interactions. A direct in vivo-in virtuo comparison would provide a stronger test of behavioral equivalence.

XSL•FO
RenderX

However, our previous work using the same dataset [41] compared key interaction metrics to established face-to-face benchmarks and found highly similar patterns, consistent with other recent findings comparing VR to real-life interactions [40], which suggests that fundamental aspects of social behavior are preserved in this social VR setting.

Lastly, although the presence of cohesion between stable traits and observed behavior implies some level of stability in behavior, our design did not allow for directly testing within-person consistency across interactions with varying partners. In order to better characterize how interpersonal processes contribute to individual variation in behavior [140], future studies should assess behavior in individuals across various interactions.

Conclusion

We investigated behavioral and physiological correlates of social anxiety as well as broader psychopathology and verticality in dyadic interactions in social VR. Extending previous work that examined interpersonal dynamics in face-to-face settings or responses to virtual agents, this study provides a comprehensive account of behavioral and physiological modulations associated with social anxiety in real-time avatar-based human-human interactions. In an integrated analytical framework, we assessed gaze, speaking, smiling expression, as well as heart rate and HF-HRV, and their associations with individual traits. As observed in natural face-to-face interactions, social anxiety levels were associated with behavioral modulations such as quieter speaking and reduced eye gaze to a partner while speaking and with reduced HF-HRV, indicating increased social stress and submissiveness. Behavioral and physiological patterns were not exclusive to social anxiety but appeared to reflect broader psychological characteristics, including higher general psychopathology and lower verticality. These results indicate that social interaction patterns may reflect general psychological vulnerability rather than effects specific to social anxiety levels. Behavioral traces that can be obtained automatically in social VR carry sensitive psychological information and, problematically, may be used for unintended profiling or psychological inference. At the same time, these data also provide valuable opportunities to assess psychologically relevant behavioral patterns in social interactions at scale in various settings, ultimately helping to understand pathological mechanisms in social interaction behavior and improving psychotherapy.

Acknowledgments

We thank Luca Panico, Susanne Felder, Michelle Bürki, Ana Rita Calçada da Silva, and Michael Mozhacherry for their help in data collection. Generative artificial intelligence was used in proofreading this manuscript.

Funding

This research was supported by grants from the Swiss National Science Foundation (grant PZ00P1_208909). The funding agencies were not involved in decisions concerning our study design, data collection, analysis, interpretation of data, writing of the report, or the submission of the article for publication.

Data Availability

This study's sample size, measures, and analysis approach were preregistered [54]. The study conformed to the principles expressed in the Declaration of Helsinki and was approved by the local ethics committee at the University of Bern (reference number 2023-09-01). Data that support the findings of this study are available at [53].

Authors' Contributions

GS contributed to conceptualization, project administration, formal analysis, visualization, and writing of the original draft, as well as review and editing. MR contributed to conceptualization, funding acquisition, project administration, formal analysis, software, supervision, and writing, review, and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Results with detailed explanations.
[DOCX File , 267 KB - jmir_v28i1e79147_app1.docx]

References

1. Heimberg RG, Brozovich FA, Rapee RM. A cognitive-behavioral model of social anxiety disorder. In: Social Anxiety. Amsterdam: Elsevier; 2014:705-728.

XSL•FO
RenderX

2. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005;62(6):593-602. [doi: 10.1001/archpsyc.62.6.593] [Medline: 15939837]
3. Fehm L, Beesdo K, Jacobi F, Fiedler A. Social anxiety disorder above and below the diagnostic threshold: prevalence, comorbidity and impairment in the general population. Soc Psychiatry Psychiatr Epidemiol 2008;43(4):257-265. [doi: 10.1007/s00127-007-0299-4] [Medline: 18084686]
4. Piccirillo ML, Taylor Dryman M, Heimberg RG. Safety behaviors in adults with social anxiety: review and future directions. Behav Ther 2016;47(5):675-687. [doi: 10.1016/j.beth.2015.11.005] [Medline: 27816080]
5. Clark D, Wells A. In: Heineberg RG, Liebowitz MR, Hope DA, Schneier FR, editors. A Cognitive Model of Social Phobia in Social Phobia: Diagnosis, Assessment, and Treatment. New York: Guilford Press; 1995.
6. Konovalova I, Antolin JV, Bolderston H, Gregory NJ. Adults with higher social anxiety show avoidant gaze behaviour in a real-world social setting: a mobile eye tracking study. PLoS One 2021;16(10):e0259007 [FREE Full text] [doi: 10.1371/journal.pone.0259007] [Medline: 34695140]
7. Baker SR, Edelmann RJ. Is social phobia related to lack of social skills? Duration of skill-related behaviours and ratings of behavioural adequacy. Br J Clin Psychol 2002;41(Pt 3):243-257. [doi: 10.1348/014466502760379118] [Medline: 12396253]
8. Chen J, van den Bos E, Westenberg PM. Does gaze anxiety predict actual gaze avoidance and is it more informative than social anxiety? J Behav Ther Exp Psychiatry 2024;82:101896. [doi: 10.1016/j.jbtep.2023.101896] [Medline: 37741178]
9. Chen J, van den Bos E, Karch JD, Westenberg PM. Social anxiety is related to reduced face gaze during a naturalistic social interaction. Anxiety Stress Coping 2023;36(4):460-474 [FREE Full text] [doi: 10.1080/10615806.2022.2125961] [Medline: 36153759]
10. Tönsing D, Schiller B, Vehlen A, Spenthof I, Domes G, Heinrichs M. No evidence that gaze anxiety predicts gaze avoidance behavior during face-to-face social interaction. Sci Rep 2022;12(1):21332 [FREE Full text] [doi: 10.1038/s41598-022-25189-z] [Medline: 36494411]
11. Shechter T, Asher M, Aderka IM. Man vs. machine: a comparison of human and computer assessment of nonverbal behavior in social anxiety disorder. J Anxiety Disord 2022;89:102587. [doi: 10.1016/j.janxdis.2022.102587] [Medline: 35661576]
12. Dijk C, Fischer AH, Morina N, van Eeuwijk C, van Kleef GA. Effects of social anxiety on emotional mimicry and contagion: feeling negative, but smiling politely. J Nonverbal Behav 2018;42(1):81-99 [FREE Full text] [doi: 10.1007/s10919-017-0266-z] [Medline: 29497222]
13. Gorlin EI, Teachman BA. Inhibitory control as a moderator of threat-related interference biases in social anxiety. Cogn Emot 2015;29(4):723-735 [FREE Full text] [doi: 10.1080/02699931.2014.931275] [Medline: 24967719]
14. Natale M, Entin E, Jaffe J. Vocal interruptions in dyadic communication as a function of speech and social anxiety. J Pers Soc Psychol 1979;37(6):865-878. [doi: 10.1037//0022-3514.37.6.865] [Medline: 490308]
15. Laukka P, Linnman C, Åhs F, Pissiota A, Frans Ö, Faria V, et al. In a nervous voice: acoustic analysis and perception of anxiety in social phobics' speech. J Nonverbal Behav 2008;32(4):195-214. [doi: 10.1007/s10919-008-0055-9]
16. Galili L, Amir O, Gilboa-Schechtman E. Acoustic properties of dominance and request utterances in social anxiety. J Soc Clin Psychol 2013;32(6):651-673. [doi: 10.1521/jscp.2013.32.6.651]
17. Cheng Y, Su M, Liu C, Huang Y, Huang W. Heart rate variability in patients with anxiety disorders: a systematic review and meta-analysis. Psychiatry Clin Neurosci 2022;76(7):292-302 [FREE Full text] [doi: 10.1111/pcn.13356] [Medline: 35340102]
18. Rösler L, Göhring S, Strunz M, Gamer M. Social anxiety is associated with heart rate but not gaze behavior in a real social interaction. J Behav Ther Exp Psychiatry 2021;70:101600 [FREE Full text] [doi: 10.1016/j.jbtep.2020.101600] [Medline: 32882674]
19. Kemp AH, Quintana DS, Gray MA, Felmingham KL, Brown K, Gatt JM. Impact of depression and antidepressant treatment on heart rate variability: a review and meta-analysis. Biol Psychiatry 2010;67(11):1067-1074. [doi: 10.1016/j.biopsych.2009.12.012] [Medline: 20138254]
20. Alvares GA, Quintana DS, Kemp AH, Van Zwieten A, Balleine BW, Hickie IB, et al. Reduced heart rate variability in social anxiety disorder: associations with gender and symptom severity. PLoS One 2013;8(7):e70468 [FREE Full text] [doi: 10.1371/journal.pone.0070468] [Medline: 23936207]
21. Suslow T, Hoepfel D, Kersting A, Bodenschatz CM. Depressive symptoms and visual attention to others' eyes in healthy individuals. BMC Psychiatry 2024;24(1):184. [doi: 10.1186/s12888-024-05633-2]
22. Quintana DS, McGregor IS, Guastella AJ, Malhi GS, Kemp AH. A meta-analysis on the impact of alcohol dependence on short-term resting-state heart rate variability: implications for cardiovascular risk. Alcohol Clin Exp Res 2013;37(Suppl 1):E23-E29. [doi: 10.1111/j.1530-0277.2012.01913.x] [Medline: 22834996]
23. Caspi A, Houts RM, Belsky DW, Goldman-Mellor SJ, Harrington H, Israel S, et al. The p factor: one general psychopathology factor in the structure of psychiatric disorders? Clin Psychol Sci 2014;2(2):119-137 [FREE Full text] [doi: 10.1177/2167702613497473] [Medline: 25360393]

XSL•FO
RenderX

24. Kotov R, Krueger RF, Watson D, Cicero DC, Conway CC, DeYoung CG, et al. The hierarchical taxonomy of psychopathology (HiTOP): a quantitative nosology based on consensus of evidence. Annu Rev Clin Psychol 2021;17(1):83-108. [doi: 10.1146/annurev-clinpsy-081219-093304] [Medline: 33577350]
25. Tharp JA, Johnson SL, Dev A. Transdiagnostic approach to the dominance behavioral system. Pers Individ Dif 2021;176:110778 [FREE Full text] [doi: 10.1016/j.paid.2021.110778] [Medline: 33746322]
26. Chen J, van den Bos E, Westenberg PM. A systematic review of visual avoidance of faces in socially anxious individuals: influence of severity, type of social situation, and development. J Anxiety Disord 2020;70:102193. [doi: 10.1016/j.janxdis.2020.102193] [Medline: 32058889]
27. Patel TA, Cougle JR. Submissive behavior as a transdiagnostic construct: associations with internalizing symptoms and distress intolerance. J Clin Psychol 2023;79(5):1314-1327. [doi: 10.1002/jclp.23478] [Medline: 36603203]
28. Gilboa-Schechtman E, Shachar I, Helpman L. Evolutionary perspective on social anxiety. In: Social Anxiety. Amsterdam: Elsevier; 2014:599-622.
29. Gilbert P. Evolution and social anxiety. The role of attraction, social competition, and social hierarchies. Psychiatr Clin North Am 2001;24(4):723-751. [doi: 10.1016/s0193-953x(05)70260-4] [Medline: 11723630]
30. Swee MB, Hudson CC, Heimberg RG. Examining the relationship between shame and social anxiety disorder: a systematic review. Clin Psychol Rev 2021;90:102088. [doi: 10.1016/j.cpr.2021.102088] [Medline: 34598054]
31. Carney DR. The nonverbal expression of power, status, and dominance. Curr Opin Psychol 2020;33:256-264. [doi: 10.1016/j.copsyc.2019.12.004] [Medline: 32361679]
32. Hall JA, Coats EJ, LeBeau LS. Nonverbal behavior and the vertical dimension of social relations: a meta-analysis. Psychol Bull 2005;131(6):898-924. [doi: 10.1037/0033-2909.131.6.898] [Medline: 16351328]
33. Johnson SL, Leedom LJ, Muhtadie L. The dominance behavioral system and psychopathology: evidence from self-report, observational, and biological studies. Psychol Bull 2012;138(4):692-743 [FREE Full text] [doi: 10.1037/a0027503] [Medline: 22506751]
34. Hadley LV, Naylor G, Hamilton AFDC. A review of theories and methods in the science of face-to-face social interaction. Nat Rev Psychol 2022;1(1):42-54. [doi: 10.1038/s44159-021-00008-w]
35. Risko EF, Laidlaw K, Freeth M, Foulsham T, Kingstone A. Social attention with real versus reel stimuli: toward an empirical approach to concerns about ecological validity. Front Hum Neurosci 2012;6:143 [FREE Full text] [doi: 10.3389/fnhum.2012.00143] [Medline: 22654747]
36. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021;20(3):318-335 [FREE Full text] [doi: 10.1002/wps.20883] [Medline: 34505369]
37. Jacobson NC, Summers B, Wilhelm S. Digital biomarkers of social anxiety severity: digital phenotyping using passive smartphone sensors. J Med Internet Res 2020;22(5):e16875 [FREE Full text] [doi: 10.2196/16875] [Medline: 32348284]
38. Choi A, Ooi A, Lottridge D. Digital phenotyping for stress, anxiety, and mild depression: systematic literature review. JMIR Mhealth Uhealth 2024;12:e40689 [FREE Full text] [doi: 10.2196/40689] [Medline: 38780995]
39. Parsons TD. Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. Front Hum Neurosci 2015;9:660 [FREE Full text] [doi: 10.3389/fnhum.2015.00660] [Medline: 26696869]
40. Abdullah A, Kolkmeier J, Lo V, Neff M. Videoconference and embodied VR: communication patterns across task and medium. Proc ACM Hum-Comput Interact 2021;5:1-29. [doi: 10.1145/3479597]
41. Son G, Rubo M. Social virtual reality elicits natural interaction behavior with self-similar and generic avatars. Int J Hum Comput Stud 2025;199:103488. [doi: 10.1016/j.ijhcs.2025.103488]
42. Girondini M, Frigione I, Marra M, Stefanova M, Pillan M, Maravita A, et al. Decoupling the role of verbal and non-verbal audience behavior on public speaking anxiety in virtual reality using behavioral and psychological measures. Front Virtual Real 2024;5. [doi: 10.3389/frvir.2024.1347102]
43. Torres SC, Gracia Laso DI, Minissi ME, Maddalon L, Chicchi Giglioli IA, Alcañiz M. Social signal processing in affective virtual reality: human-shaped agents increase electrodermal activity in an elicited negative environment. Cyberpsychol Behav Soc Netw 2024;27(4):268-274 [FREE Full text] [doi: 10.1089/cyber.2023.0273] [Medline: 38394167]
44. Pan X, Hamilton AFDC. Why and how to use virtual reality to study human social interaction: the challenges of exploring a new research landscape. Br J Psychol 2018;109(3):395-417 [FREE Full text] [doi: 10.1111/bjop.12290] [Medline: 29504117]
45. Bailenson JN, Blascovich J, Beall AC, Loomis JM. Interpersonal distance in immersive virtual environments. Pers Soc Psychol Bull 2003;29(7):819-833. [doi: 10.1177/0146167203029007002] [Medline: 15018671]
46. Belina SP, Czaja SJ, Difede J, Pain KJ, Reid MC, Won AS. In sickness and in health-a scoping review on social virtual reality clinical applications. Digit Health 2024;10:20552076241306460 [FREE Full text] [doi: 10.1177/20552076241306460] [Medline: 39741985]
47. Li C, Dai Y, Chen G, Liu J, Li P, Ip HH. Avatar-mediated communication in collaborative virtual environments: a study on users' attention allocation and perception of social interactions. Comput Hum Behav 2025;167:108598. [doi: 10.1016/j.chb.2025.108598]

XSL•FO
RenderX

48. Moharana B, Keighrey C, Murray N. Physiological synchrony in a collaborative virtual reality task. : IEEE; 2023 Presented at: 15th International Conference on Quality of Multimedia Experience (QoMEX); 2023 June 20; Ghent, Belgium p. 288-293. [doi: 10.1109/qomex58391.2023.10178661]
49. Kimmel S, Jung F, Matviienko A, Heuten W, Boll S. Let's face it: influence of facial expressions on social presence in collaborative virtual reality. 2023 Presented at: Conference on Human Factors in Computing Systems - Proceedings; 2016 Apr 19; New York, NY p. 1-16. [doi: 10.1145/3544548.3580707]
50. Léchappé A, Johnstone R, Milliat A, Williamson JH, Chollet M, Williamson JR. Understanding social interactions in reality versus virtuality. : ACM; 2025 Presented at: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. CHI '25; 2025 Apr 25; New York, NY p. 1-18. [doi: 10.1145/3706598.3713628]
51. Xenakis I, Gavalas D, Kasapakis V, Dzardanova E, Vosinakis S. Nonverbal communication in immersive virtual reality through the lens of presence: a critical review. Presence 2022;31:147-187. [doi: 10.1162/pres_a_00387]
52. Yaremych HE, Persky S. Tracing physical behavior in virtual reality: a narrative review of applications to social psychology. J Exp Soc Psychol 2019;85:103845 [FREE Full text] [doi: 10.1016/j.jesp.2019.103845] [Medline: 32831397]
53. Son G, Rubo M. Data and analyses. OSF. URL: https://osf.io/9yr8p [accessed 2026-02-23]
54. Son G, Rubo M. Preregistration. OSF. URL: https://osf.io/kawtn [accessed 2026-02-23]
55. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007;370(9596):1453-1457 [FREE Full text] [doi: 10.1016/S0140-6736(07)61602-X] [Medline: 18064739]
56. Unity Engine. Unity. URL: https://unity.com/ [accessed 2026-02-23]
57. Meta Quest Pro headsets. Meta. URL: https://www.meta.com/ [accessed 2026-02-23]
58. Rubo M. Creating a social virtual reality application for psychological research: A tutorial. Behav Res Methods 2025 Jun 04;57(7):192 [FREE Full text] [doi: 10.3758/s13428-025-02693-4] [Medline: 40468093]
59. Reallusion. URL: https://www.reallusion.com/ [accessed 2026-02-23]
60. Schreiber JB. Issues and recommendations for exploratory factor analysis and principal component analysis. Res Social Adm Pharm 2021;17(5):1004-1011. [doi: 10.1016/j.sapharm.2020.07.027] [Medline: 33162380]
61. Little RJA. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 1988;83(404):1198-1202. [doi: 10.1080/01621459.1988.10478722]
62. Aron A, Melinat E, Aron EN, Vallone RD, Bator RJ. The experimental generation of interpersonal closeness: a procedure and some preliminary findings. Pers Soc Psychol Bull 1997;23(4):363-377. [doi: 10.1177/0146167297234003]
63. Matz DC, Wood W. Cognitive dissonance in groups: the consequences of disagreement. J Pers Soc Psychol 2005;88(1):22-37. [doi: 10.1037/0022-3514.88.1.22] [Medline: 15631572]
64. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. IEEE Trans Pattern Anal Mach Intell 2021;43(12):4217-4228. [doi: 10.1109/TPAMI.2020.2970919] [Medline: 32012000]
65. Rubo M. AvatarTransferPro: a tool for automatic transfer of realistic personalized avatars in networked social virtual reality. SoftwareX 2024;27:101873. [doi: 10.1016/j.softx.2024.101873]
66. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 1995;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]
67. Mattick RP, Clarke JC. Development and validation of measures of social phobia scrutiny fear and social interaction anxiety. Behav Res Ther 1998;36(4):455-470. [doi: 10.1016/s0005-7967(97)10031-6] [Medline: 9670605]
68. Stangier U, Heidenreich T, Berardi A, Golbs U, Hoyer J. Die erfassung sozialer phobie durch die Social Interaction Anxiety Scale (SIAS) und die Social Phobia Scale (SPS). Z Klin Psychol 1999;28(1):28-36 [FREE Full text] [doi: 10.1026//0084-5345.28.1.28]
69. Weeks JW, Howell AN. The bivalent fear of evaluation model of social anxiety: further integrating findings on fears of positive and negative evaluation. Cogn Behav Ther 2012;41(2):83-95. [doi: 10.1080/16506073.2012.661452] [Medline: 22489647]
70. Weeks JW, Heimberg RG, Rodebaugh TL. The Fear of Positive Evaluation Scale: assessing a proposed cognitive component of social anxiety. J Anxiety Disord 2008;22(1):44-55. [doi: 10.1016/j.janxdis.2007.08.002] [Medline: 17884328]
71. Schwarz M, Reichenberger J, Schwitalla M, König D, Wilhelm FH, Voderholzer U, et al. Furcht vor positiver sozialer bewertung: ubersetzung und validierung einer deutschsprachigen version der FPES. Diagnostica 2016;62(3):157-168. [doi: 10.1026/0012-1924/a000147]
72. Carleton RN, McCreary DR, Norton PJ, Asmundson GJG. Brief Fear of Negative Evaluation Scale-revised. Depress Anxiety 2006;23(5):297-303. [doi: 10.1002/da.20142] [Medline: 16688736]
73. Reichenberger J, Schwarz M, König D, Wilhelm FH, Voderholzer U, Hillert A, et al. Angst vor negativer sozialer bewertung: ubersetzung und validierung der furcht vor negativer evaluation–Kurzskala (FNE-K). Diagnostica 2016;62(3):169-181. [doi: 10.1026/0012-1924/a000148]
74. Allan S, Gilbert P. A social comparison scale: psychometric properties and relationship to psychopathology. Pers Individ Differ 1995;19(3):293-299. [doi: 10.1016/0191-8869(95)00086-l]
75. Ascone L, Jaya ES, Lincoln TM. The effect of unfavourable and favourable social comparisons on paranoid ideation: an experimental study. J Behav Ther Exp Psychiatry 2017;56:97-105. [doi: 10.1016/j.jbtep.2016.08.002] [Medline: 27567982]

XSL•FO
RenderX

76. Spielberger CD. State-Trait Anxiety Inventory for Adults. PsycTESTS Dataset 1983. [doi: 10.1037/t06496-000]
77. Grimm J. State-Trait-Anxiety Inventory nach Spielberger. Deutsche lang-und kurzversion. Methodenforum der Universität Wien 2009;2009(02) [FREE Full text]
78. Cohen S, Williamson GM. Perceived stress in a probability sample of the United States. In: Spacapan S, Oskamp S, editors. The Social Psychology of Health. California: Sage Publications; 1988:31-67.
79. Beck AT, Brown GK. BDI-II Fast Screen for Medical Patients Manual. London: The Psychological Corporation; 2000.
80. Kliem S, Mößle T, Zenger M, Brähler E. Reliability and validity of the Beck Depression Inventory-Fast Screen for Medical Patients in the general German population. J Affect Disord 2014;156:236-239. [doi: 10.1016/j.jad.2013.11.024] [Medline: 24480380]
81. Meule A, Vögele C, Kübler A. Psychometrische evaluation der deutschen Barratt Impulsiveness Scale – Kurzversion (BIS-15). Diagnostica 2011;57(3):126-133 [FREE Full text] [doi: 10.1026/0012-1924/a000042]
82. Spinella M. Normative data and a short form of the Barratt impulsiveness scale. Int J Neurosci 2007;117(3):359-368. [doi: 10.1080/00207450600588881] [Medline: 17365120]
83. Greenberg DM, Warrier V, Allison C, Baron-Cohen S. Testing the empathizing-systemizing theory of sex differences and the extreme male brain theory of autism in half a million people. Proc Natl Acad Sci USA 2018;115(48):12152-12157 [FREE Full text] [doi: 10.1073/pnas.1811032115] [Medline: 30420503]
84. John OP, Donahue EM, Kentle RL. Big Five Inventory. APA PsycTests 1991 [FREE Full text] [doi: 10.1037/t07550-000]
85. Rammstedt B, John OP. Kurzversion des Big Five Inventory (BFI-K). Diagnostica 2005;51(4):195-206. [doi: 10.1026/0012-1924.51.4.195]
86. Kotov R, Gamez W, Schmidt F, Watson D. Linking "big" personality traits to anxiety, depressive, and substance use disorders: a meta-analysis. Psychol Bull 2010;136(5):768-821. [doi: 10.1037/a0020327] [Medline: 20804236]
87. Hakulinen C, Elovainio M, Pulkki-Råback L, Virtanen M, Kivimäki M, Jokela M. Personality and depressive symptoms: individual participant meta-analysis of 10 cohort studies. Depress Anxiety 2015;32(7):461-470 [FREE Full text] [doi: 10.1002/da.22376] [Medline: 26014798]
88. Rosenberg M. Society and the Adolescent Self-Image. USA: Princeton University Press; 1965:326.
89. von Collani G, Herzberg PY. Eine revidierte fassung der deutschsprachigen Skala zum selbstwertgefühl von Rosenberg. Z Differ Diagn Psychol 2006;24(1):3-7. [doi: 10.1024//0170-1789.24.1.3]
90. Adler NE, Epel ES, Castellazzo G, Ickovics JR. Relationship of subjective and objective social status with psychological and physiological functioning: preliminary data in healthy, white women. Health Psychol 2000;19(6):586-592. [doi: 10.1037/0278-6133.19.6.586]
91. Hoebel J, Müters S, Kuntz B, Lange C, Lampert T. [Measuring subjective social status in health research with a German version of the MacArthur Scale]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2015;58(7):749-757. [doi: 10.1007/s00103-015-2166-x] [Medline: 25986532]
92. Anderson C, John OP, Keltner D. The personal sense of power. J Pers 2012;80(2):313-344. [doi: 10.1111/j.1467-6494.2011.00734.x] [Medline: 21446947]
93. Körner R, Heydasch T, Schütz A. It's all about power: validation of trait and state versions of the German personal sense of power scale. Eur J Psychol Assess 2022;38(1):36-48. [doi: 10.1027/1015-5759/a000642]
94. Cheng JT, Tracy JL, Foulsham T, Kingstone A, Henrich J. Two ways to the top: evidence that dominance and prestige are distinct yet viable avenues to social rank and influence. J Pers Soc Psychol 2013;104(1):103-125. [doi: 10.1037/a0030398] [Medline: 23163747]
95. Cheng JT, Tracy JL, Henrich J. Pride, personality, and the evolutionary foundations of human social status. Evol Hum Behav 2010;31(5):334-347. [doi: 10.1016/j.evolhumbehav.2010.02.004]
96. Thomas A, Brähler E, Strauß B. IIP-32: entwicklung, validierung und normierung einer kurzform des inventars zur erfassung interpersonaler probleme. Diagnostica 2011;57(2):68-83 [FREE Full text] [doi: 10.1026/0012-1924/a000034]
97. Barkham M, Hardy GE, Startup M. The IIP-32: a short version of the inventory of interpersonal problems. Br J Clin Psychol 1996;35(1):21-35. [doi: 10.1111/j.2044-8260.1996.tb01159.x] [Medline: 8673033]
98. Brown EJ, Turovsky J, Heimberg RG, Juster HR, Brown TA, Barlow DH. Validation of the Social Interaction Anxiety Scale and the Social Phobia Scale across the anxiety disorders. Psychol Assess 1997;9(1):21-27. [doi: 10.1037/1040-3590.9.1.21]
99. Alsaleh M, Videloup L, Lobbedez T, Lebreuilly J, Morello R, Thuillier Lecouf A. Improved detection and evaluation of depression in patients with chronic kidney disease: validity and reliability of screening (PHQ-2) and diagnostic (BDI-FS-Fr) tests of depression in chronic kidney disease. Kidney Dis (Basel) 2019;5(4):228-238 [FREE Full text] [doi: 10.1159/000497352] [Medline: 31768380]
100. Gee de JW, Knapen T, Donner TH. Decision-related pupil dilation reflects upcoming choice and individual bias. Proc Natl Acad Sci USA 2014;111(5):E618-E625 [FREE Full text] [doi: 10.1073/pnas.1317557111] [Medline: 24449874]
101. Geller J, Winn MB, Mahr T, Mirman D. GazeR: a package for processing gaze position and pupil size data. Behav Res Methods 2020;52(5):2232-2255 [FREE Full text] [doi: 10.3758/s13428-020-01374-8] [Medline: 32291732]

XSL•FO
RenderX

102. Aziz S, Lohr D, Friedman L, Komogortsev O. Evaluation of eye tracking signal quality for virtual reality applications: a case study in the Meta Quest Pro. : ACM; 2024 Presented at: Proceedings of the 2024 Symposium on Eye Tracking Research and Applications. ETRA '24; 2024 June 04; United States p. 1-8. [doi: 10.1145/3649902.3653347]
103. Wei S, Bloemers D, Rovira A. A preliminary study of the eye tracker in the Meta Quest Pro. : ACM; 2023 Presented at: Proceedings of the 2023 ACM International Conference on Interactive Media Experiences. IMX '23; 2023 June 12; USA p. 216-221. [doi: 10.1145/3573381.3596467]
104. Ho S, Foulsham T, Kingstone A. Speaking and listening with the eyes: gaze signaling during dyadic interactions. PLoS One 2015;10(8):e0136905 [FREE Full text] [doi: 10.1371/journal.pone.0136905] [Medline: 26309216]
105. Frauendorfer D, Schmid Mast M, Nguyen L, Gatica-Perez D. Nonverbal social sensing in action: unobtrusive recording and extracting of nonverbal behavior in social interactions illustrated with a research example. J Nonverbal Behav 2014;38(2):231-245. [doi: 10.1007/s10919-014-0173-5]
106. Lee JJ, Breazeal C, DeSteno D. Role of speaker cues in attention inference. Front Robot AI 2017;4. [doi: 10.3389/frobt.2017.00047]
107. Heldner M, Edlund J. Pauses, gaps and overlaps in conversations. J Phonet 2010;38(4):555-568. [doi: 10.1016/j.wocn.2010.08.002]
108. Skantze G. Turn-taking in conversational systems and human-robot interaction: a review. Comput Speech Lang 2021;67:101178. [doi: 10.1016/j.csl.2020.101178]
109. Holler J, Kendrick K, Casillas M, Levinson S. Turn-taking in human communicative interaction. Frontiers Media SA 2016. [doi: 10.3389/978-2-88919-825-2]
110. Face tracking for movement SDK for Unity. Meta. URL: https://developers.meta.com/horizon/documentation/unity/move-face-tracking/ [accessed 2026-02-23]
111. Plux Biosignals BITalino. Biosignals Plux. URL: https://www.pluxbiosignals.com/ [accessed 2026-02-23]
112. Elgendi M, Jonkman M, DeBoer F. Frequency bands effects on QRS detection. : SciTePress - Science, and Technology Publications; 2010 Presented at: Proceedings of the Third International Conference on Bio-Inspired Systems and Signal Processing; 2010 June 20; Valencia, Spain p. 428-431. [doi: 10.5220/0002742704280431]
113. R Core Team. R: A Language and Environment for Statistical Computing. Denmark: R Foundation for Statistical Computing; 2023.
114. Rubo M, Munsch S. Social stress in an interaction with artificial agents in virtual reality: effects of ostracism and underlying psychopathology. Comput Hum Behav 2024;153:107915. [doi: 10.1016/j.chb.2023.107915]
115. Chalmers JA, Quintana DS, Abbott MJ, Kemp AH. Anxiety disorders are associated with reduced heart rate variability: a meta-analysis. Front Psychiatry 2014;5:80 [FREE Full text] [doi: 10.3389/fpsyt.2014.00080] [Medline: 25071612]
116. Gidron Y, Deschepper R, De Couck M, Thayer JF, Velkeniers B. The vagus nerve can predict and possibly modulate non-communicable chronic diseases: introducing a neuroimmunological paradigm to public health. J Clin Med 2018;7(10):371 [FREE Full text] [doi: 10.3390/jcm7100371] [Medline: 30347734]
117. Tomasi J, Zai CC, Pouget JG, Tiwari AK, Kennedy JL. Heart rate variability: evaluating a potential biomarker of anxiety disorders. Psychophysiology 2024;61(2):e14481. [doi: 10.1111/psyp.14481] [Medline: 37990619]
118. Russell JJ, Moskowitz DS, Zuroff DC, Bleau P, Pinard G, Young SN. Anxiety, emotional security and the interpersonal behavior of individuals with social anxiety disorder. Psychol Med 2010;41(3):545-554. [doi: 10.1017/s0033291710000863]
119. Allan S, Gilbert P. Submissive behaviour and psychopathology. Br J Clin Psychol 1997;36(4):467-488. [doi: 10.1111/j.2044-8260.1997.tb01255.x] [Medline: 9403141]
120. Achterhof R, Kirtley OJ, Schneider M, Hagemann N, Hermans KS, Hiekkaranta AP, et al. General psychopathology and its social correlates in the daily lives of youth. J Affect Disord 2022;309:428-436. [doi: 10.1016/j.jad.2022.04.147] [Medline: 35500686]
121. Carver CS, Johnson SL, Timpano KR. Toward a functional view of the P factor in psychopathology. Clin Psychol Sci 2017;5(5):880-889. [doi: 10.1177/2167702617710037] [Medline: 29057170]
122. Oh HJ, Kim J, Chang JJ, Park N, Lee S. Social benefits of living in the metaverse: the relationships among social presence, supportive interaction, social self-efficacy, and feelings of loneliness. Comput Human Behav 2023;139:107498. [doi: 10.1016/j.chb.2022.107498]
123. Bailenson JN. Nonverbal overload: a theoretical argument for the causes of Zoom fatigue. Technol Mind Behav 2021;2(1):1-6. [doi: 10.1037/tmb0000030]
124. Bufano P, Laurino M, Said S, Tognetti A, Menicucci D. Digital phenotyping for monitoring mental disorders: systematic review. J Med Internet Res 2023;25:e46778 [FREE Full text] [doi: 10.2196/46778] [Medline: 38090800]
125. Bell IH, Nicholas J, Alvarez-Jimenez M, Thompson A, Valmaggia L. Virtual reality as a clinical tool in mental health research and practice . Dialogues Clin Neurosci 2020;22(2):169-177 [FREE Full text] [doi: 10.31887/DCNS.2020.22.2/lvalmaggia] [Medline: 32699517]
126. Giaretta A. Security and privacy in virtual reality: a literature survey. Virtual Real 2024;29(1):10. [doi: 10.1007/s10055-024-01079-9]
127. Halberstadt AG, Parker AE, Castro VL. Nonverbal Communication: Developmental Perspectives. Berlin, Germany: Nonverbal Communication. DE GRUYTER; 2013:93-128.

XSL•FO
RenderX

128. Levine CS, Ambady N. The role of non-verbal behaviour in racial disparities in health care: implications and solutions. Med Educ 2013;47(9):867-876. [doi: 10.1111/medu.12216] [Medline: 23931536]
129. Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? Behav Brain Sci 2010;33(2-3):61-83. [doi: 10.1017/s0140525x0999152x]
130. Hofmann SG, Anu Asnaani MA, Hinton DE. Cultural aspects in social anxiety and social anxiety disorder. Depress Anxiety 2010;27(12):1117-1127 [FREE Full text] [doi: 10.1002/da.20759] [Medline: 21132847]
131. Lafrance M, Hecht M. Gender and smiling: a meta-analysis. In: Gender and Emotion. United Kingdom: Cambridge University Press; 2000:118-142.
132. Hunt J, Eisenberg D. Mental health problems and help-seeking behavior among college students. J Adolesc Health 2010;46(1):3-10. [doi: 10.1016/j.jadohealth.2009.08.008] [Medline: 20123251]
133. Caspi A, Houts RM, Fisher HL, Danese A, Moffitt TE. The general factor of psychopathology (p): choosing among competing models and interpreting p. Clin Psychol Sci 2024;12(1):53-82 [FREE Full text] [doi: 10.1177/21677026221147872] [Medline: 38236494]
134. McGorry P, Van Os J. Redeeming diagnosis in psychiatry: timing versus specificity. Lancet 2013;381(9863):343-345. [doi: 10.1016/s0140-6736(12)61268-9]
135. Crişan LG, Vulturar R, Miclea M, Miu AC. Reactivity to social stress in subclinical social anxiety: emotional experience, cognitive appraisals, behavior, and physiology. Front Psychiatry 2016;7:5 [FREE Full text] [doi: 10.3389/fpsyt.2016.00005] [Medline: 26858658]
136. Shaukat-Jali R, van Zalk N, Boyle DE. Detecting subclinical social anxiety using physiological data from a wrist-worn wearable: small-scale feasibility study. JMIR Form Res 2021;5(10):e32656 [FREE Full text] [doi: 10.2196/32656] [Medline: 34617905]
137. Cosgrove VE, Rhee SH, Gelhorn HL, Boeldt D, Corley RC, Ehringer MA, et al. Structure and etiology of co-occurring internalizing and externalizing disorders in adolescents. J Abnorm Child Psychol 2011;39(1):109-123 [FREE Full text] [doi: 10.1007/s10802-010-9444-8] [Medline: 20683651]
138. Lahey BB, Krueger RF, Rathouz PJ, Waldman ID, Zald DH. A hierarchical causal taxonomy of psychopathology across the life span. Psychol Bull 2017;143(2):142-186 [FREE Full text] [doi: 10.1037/bul0000069] [Medline: 28004947]
139. van der Linden D, Dunkel CS, Prinzie P, Yamanaka-Altenstein M, von Wyl A, Hengartner MP. Overlap between general factors of psychopathology and personality: they share associations with daily life functioning and communication style. Curr Psychol 2021;42(13):10964-10976. [doi: 10.1007/s12144-021-02354-7]
140. Kenny DA, Mohr CD, Levesque MJ. A social relations variance partitioning of dyadic behavior. Psychol Bull 2001;127(1):128-141. [doi: 10.1037//0033-2909.127.1.128]

Abbreviations

HF-HRV: high-frequency heart rate variability
SIAS: Social Interaction Anxiety Scale
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
VR: virtual reality

Edited by S Brini; submitted 16.Jun.2025; peer-reviewed by K Wen, M Chollet, C Corciova; comments to author 03.Oct.2025; accepted 22.Dec.2025; published 11.Mar.2026.

Please cite as:
Son G, Rubo M
Automated Inference of Social Anxiety From Behavior in Social Virtual Reality: Cross-Sectional Observational Study
J Med Internet Res 2026;28:e79147
URL: https://www.jmir.org/2026/1/e79147
doi:10.2196/79147
PMID:

©Gayoung Son, Marius Rubo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Automated Identification of Nursing Diagnoses and Interventions From Nursing Records Using a Retrieval-Augmented Large Language Model Approach: Quantitative Study

Minye Li[1*], MAST, RN; Dianjie Chen[2*], RN; Qun Xiao[3], RN; Zhengji He[3], RN; Yanyan Zhang[4], RN; Jinlan Zhong[1], RN; Yiwei Luo[1], PhD; Hui Ma[1], PhD

[1]Department of Nursing, Chinese People's Liberation Army General Hospital, 28 Fuxing Road, Haidian District, Beijing, Beijing, China

[2]Department of Infectious Diseases, 5th Medical Center, Chinese People's Liberation Army General Hospital, Beijing, Beijing, China

[3]School of Nursing, Southern Medical University, Guangzhou, Guangdong, China

[4]Department of Respiratory and Critical Care Medicine, 1st Medical Center, Chinese People's Liberation Army General Hospital, Beijing, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Hui Ma, PhD
Department of Nursing, Chinese People's Liberation Army General Hospital, 28 Fuxing Road, Haidian District, Beijing, Beijing, China

Abstract

Background: Electronic health records (EHRs) have been widely adopted, but most nursing records remain in unstructured free-text format, which limits the secondary use of nursing data. Standardized terminologies improve semantic interoperability; however, manual annotation is labor intensive and yields inconsistent results. Advances in large language models (LLMs) and retrieval-augmented generation (RAG) have created new possibilities for automating the mapping of nursing records to standardized terminologies, thereby enhancing the utility of nursing data.

Objective: This study aimed to develop and evaluate Clinical Care Classification nursing terminology with retrieval-augmented mapping (CNTRAM), a 2-stage RAG framework incorporating an LLM, for the automated mapping of nursing diagnoses and interventions from free-text intensive care unit nursing records to standardized Clinical Care Classification (CCC) terms.

Methods: CNTRAM is a 2-stage retrieval-augmented framework that integrates dense embedding retrieval, retrieval-enhanced prompting, and few-shot LLM guidance to map free-text nursing records to standardized CCC terminology. Free-text records and their segments were embedded as subqueries to retrieve the most relevant CCC reference entries and annotated examples, which were merged to construct context windows. Each subquery was combined with its retrieved context using a predefined RAG prompt template that enforces CCC coding rules and a structured JSON schema and was then processed by an LLM to generate CCC outputs. A gold standard dataset of 100 intensive care unit nursing records was annotated by 3 senior nurses and finalized via consensus, with interrater reliability quantified using the Fleiss κ. Model performance was compared with traditional baselines (term frequency–inverse document frequency, Bidirectional Encoder Representations from Transformer, and fine-tuned Bidirectional Encoder Representations from Transformers model) and 4 LLMs (Mistral-7B, Qwen3-14B, Llama3.3-70B, and DeepSeek-R1) across no-RAG, zero-shot, and few-shot settings, using precision, recall, F_1-score, and intersection over union (IoU) as metrics.

Results: Interrater agreement was substantial, with Fleiss κ=0.6449 for diagnoses and κ=0.6180 for interventions. CNTRAM achieved substantial performance gains over all baseline approaches. For nursing diagnoses, DeepSeek-R1 with RAG+few-shot prompting achieved the best performance, with a precision of 0.7909, a recall of 0.7901, an F_1-score of 0.7836, and an IoU of 0.7614. These results were significantly higher than those of traditional baselines (F_1-score 0.0268 - 0.2027), no-RAG LLMs (F_1-score 0.0299 - 0.0588), and RAG+zero-shot LLMs (F_1-score 0.0716 - 0.2160). For nursing interventions, the same configuration achieved a precision of 0.8453, a recall of 0.8504, an F_1-score of 0.8413, and an IoU of 0.8097, outperforming traditional baselines (F_1-score 0.1200 - 0.2323), no-RAG LLMs (F_1-score 0.0077 - 0.0189), and RAG+zero-shot LLMs (F_1-score 0.2744 - 0.4461).

Conclusions: This study developed CNTRAM, an LLM-based 2-stage RAG framework that combines dense embedding retrieval and few-shot prompting for CCC terminology mapping. Using DeepSeek-R1, CNTRAM outperformed baseline models, improved mapping accuracy, and provided a feasible solution for standardizing unstructured nursing data.

(J Med Internet Res 2026;28:e89850) doi:10.2196/89850

XSL•FO
RenderX

KEYWORDS

large language model; LLM; retrieval-augmented generation; RAG; nursing terminology; Clinical Care Classification; nursing records; nursing informatics

Introduction

The widespread adoption of electronic health records (EHRs) has significantly enhanced the collection and use of clinical data [1], providing a solid foundation for data-driven research and continuous improvement of health care practices [2]. Within EHR systems, nursing records constitute a core data source, documenting nurses' clinical observations, patient assessments, symptom responses, and nursing interventions delivered during routine care [3]. These records support clinical decision-making and care planning and also constitute an important primary data source for nursing research [4]. Despite the transition of nursing documentation from paper-based to electronic systems, most nursing records are still recorded in unstructured free-text form. This lack of structure substantially constrains their secondary use, semantic interoperability, and integration into clinical decision support systems [5].

To address these challenges, standardized nursing terminologies (SNTs) have been developed to enhance the consistency, comparability, and interoperability of nursing records [6,7]. Among them, the Clinical Care Classification (CCC) system is a comprehensive, coded nursing terminology designed specifically for the digital documentation of nursing care [8]. The CCC system comprises 176 nursing diagnoses and 201 core nursing interventions, organized into 21 care components covering physiological, functional, psychological, and health behavioral domains. By integrating nursing diagnoses, interventions, and outcomes within a unified structure, CCC provides a systematic framework for representing nursing practice [9]. Compared to the more complex and granular North American Nursing Diagnosis Association International (NANDA-I), Nursing Interventions Classification (NIC), or International Classification for Nursing Practice systems, the CCC system has a more compact semantic space, reducing conceptual ambiguity during the automatic mapping process [10]. A Chinese translation of the CCC system was released in 2018 to support its application in Chinese-language clinical settings, enabling standardized nursing documentation in Chinese hospitals and improving the interoperability of nursing data [11].

Mapping free-text nursing records to CCC terms enables the structured representation of nursing care processes, supports care quality assessment, and facilitates large-scale nursing research [12]. However, CCC-based annotation still relies heavily on manual coding, which is labor intensive, time consuming, and subject to interannotator variability, highlighting the urgent need for automated solutions that can reliably link unstructured nursing text to standardized terminology [13,14].

Recent advances in natural language processing and machine learning have substantially improved the automated structuring of unstructured clinical text [15-17]. Early rule-based and dictionary matching approaches were limited by heavy manual maintenance and poor generalizability to complex nursing language [18,19]. The advent of pretrained contextual models, such as Bidirectional Encoder Representations from Transformers (BERT) [20], enabled deeper semantic understanding and improved performance in tasks such as entity recognition, relation extraction, and text classification. Recently, large language models (LLMs) have further revolutionized the field, exhibiting strong capabilities in long-text comprehension, zero- and few-shot learning, and structured text generation [21]. Techniques such as retrieval-augmented generation (RAG) further enhance factual accuracy and interpretability by integrating external knowledge retrieval [22]. For example, a nursing-focused study showed that integrating a curated breast cancer nursing knowledge base with GPT-4 (OpenAI) improved the accuracy and clinical relevance of nursing responses while preserving empathetic communication [23].

Despite these technological advances, applications within nursing informatics remain limited. Most existing methods are rule driven, limited in scalability, or focused on general clinical text processing, while studies specifically targeting the automated identification of nursing diagnoses and interventions remain scarce. To address this gap, this study proposes an LLM-based framework for the automated extraction of nursing diagnoses and interventions aligned with the CCC system, aiming to reduce reliance on manual coding by nurses, enhance interoperability, and support data-driven decision-making in nursing practice.

Methods

Ethical Considerations

This study was approved by the ethics committee of the Chinese People's Liberation Army General Hospital (S2024-767-01). Given the retrospective nature of this study, the requirement for written informed consent was waived by the ethics committee. No compensation was provided to participants. All data were fully anonymized before being included in the study. The study adhered to relevant ethical guidelines and protected the privacy and confidentiality of all patient records throughout the research process.

Dataset

This study included 197,337 nursing records from the intensive care unit (ICU) of Chinese People's Liberation Army General Hospital, covering the period from January 2023 to May 2025. These records consisted of free-text descriptions of time-stamped nursing assessments, interventions, and patient responses. All records were deidentified using a validated deidentification pipeline.

Gold Standard Data Annotation

To establish a gold standard dataset for model evaluation, 100 ICU nursing records were randomly selected for expert annotation. Three senior registered nurses, each with >5 years of ICU experience and prior familiarity with SNTs, independently annotated the nursing diagnoses and interventions

XSL•FO
RenderX

for all samples. Before annotation, the annotators received a structured briefing on the CCC terminology, including term definitions, hierarchical structure, and mapping principles. Annotation was conducted in accordance with the official CCC terminology definitions and classification guidelines. Interrater reliability among the 3 annotators was assessed using the Fleiss κ, which indicated substantial agreement. The gold standard dataset was established through discussions among the 3 nurses to reach consensus on differing annotations. A representative annotated example is shown in Figure 1.

Figure 1. Annotation example. This figure shows how a segment of unstructured intensive care unit nursing text from a local hospital was manually annotated using Clinical Care Classification nursing diagnoses and interventions. Text segments with the same background color indicate the correspondence between the narrative content and the assigned nursing diagnosis or intervention.

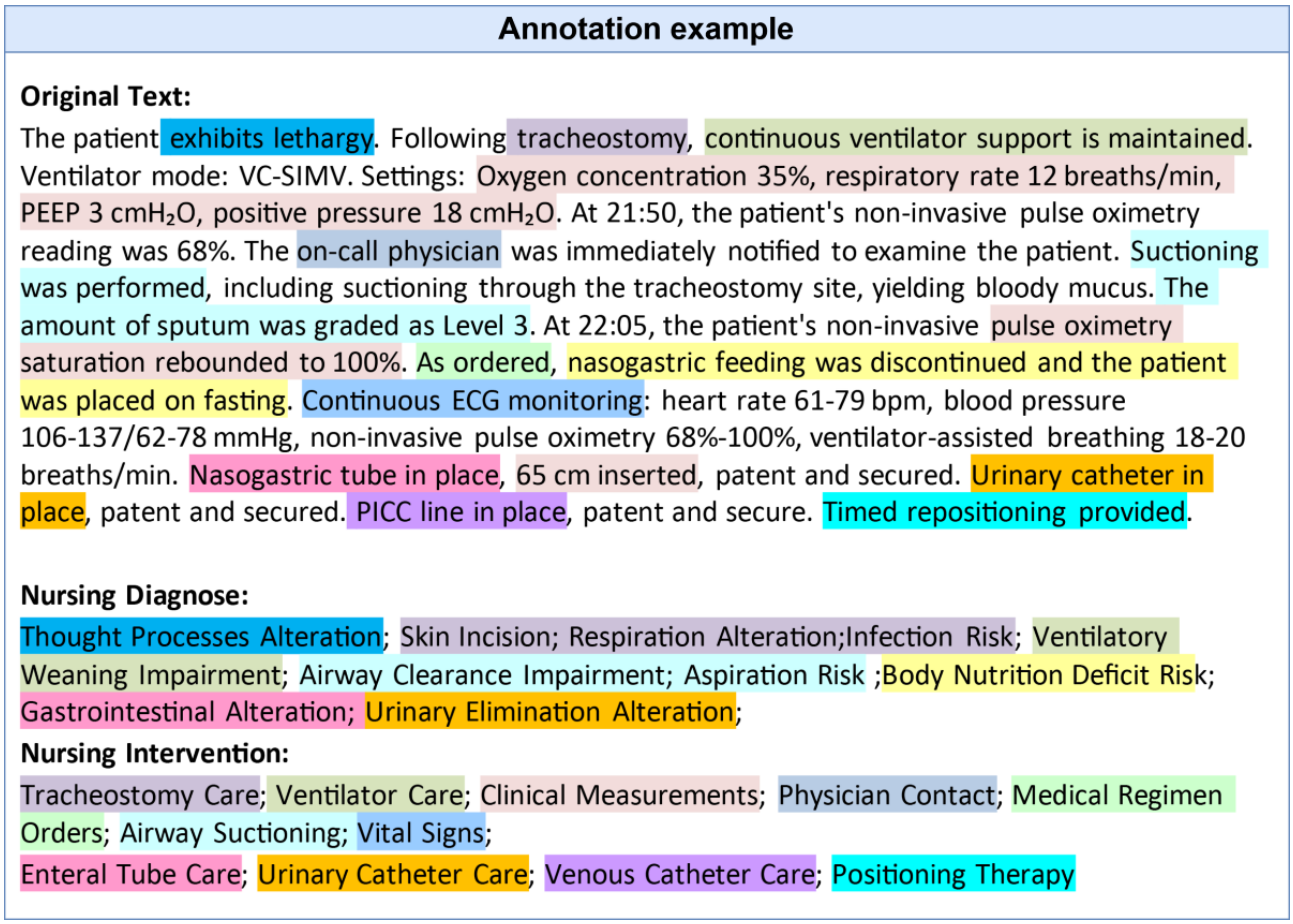

Annotated Corpus

In total, 1500 nursing text samples were randomly selected from the full dataset to construct a manually annotated corpus. Two nursing postgraduate students, each with at least 3 months of ICU training, independently performed annotation after receiving standardized training based on the CCC terminology guidelines. The task involved identifying nursing diagnoses and interventions mentioned in each text and mapping them to the most appropriate CCC terms. When exact matches were unavailable, corresponding NANDA-I and NIC terms were applied as supplementary standards (Multimedia Appendix 1). In cases of disagreement, a senior ICU nurse adjudicated the final labels. To ensure data separation, the 100 expert-annotated evaluation records and the 1500 manually annotated corpus samples were sourced from distinct patients.

CCC Nursing Terminology With Retrieval-Augmented Mapping

We developed CCC nursing terminology with retrieval-augmented mapping (CNTRAM), a 2-stage RAG framework, to map free-text nursing records to standardized CCC diagnoses and interventions (Figure 2). CNTRAM integrates dense embedding retrieval, retrieval-enhanced prompting, and few-shot LLM guidance to improve mapping accuracy and coverage.

XSL•FO
RenderX

Figure 2. Clinical Care Classification (CCC) nursing terminology with retrieval-augmented mapping (CNTRAM) framework. This figure illustrates the 2-stage CNTRAM process for mapping free-text nursing records to standardized CCC diagnoses and interventions. Free-text records are segmented into text chunks, while the original records are retained as subqueries. All subqueries, together with CCC reference entries and annotated corpora, are embedded into 512-dimensional vectors using the BAAI (Beijing Academy of Artificial Intelligence) General Embedding–small-zh model. For each subquery, the top 5 most relevant entries are separately retrieved from CCC references and annotated corpora, then merged and deduplicated to construct context windows. Each subquery is combined with its context window to generate prompts under a predefined retrieval-augmented generation template that adheres to CCC coding guidelines and structured JSON output requirements. Prompts are fed into a locally deployed large language model, and the generated JSON outputs are parsed, cross-referenced against the CCC database for metadata validation, and formatted into standardized nursing coding results.

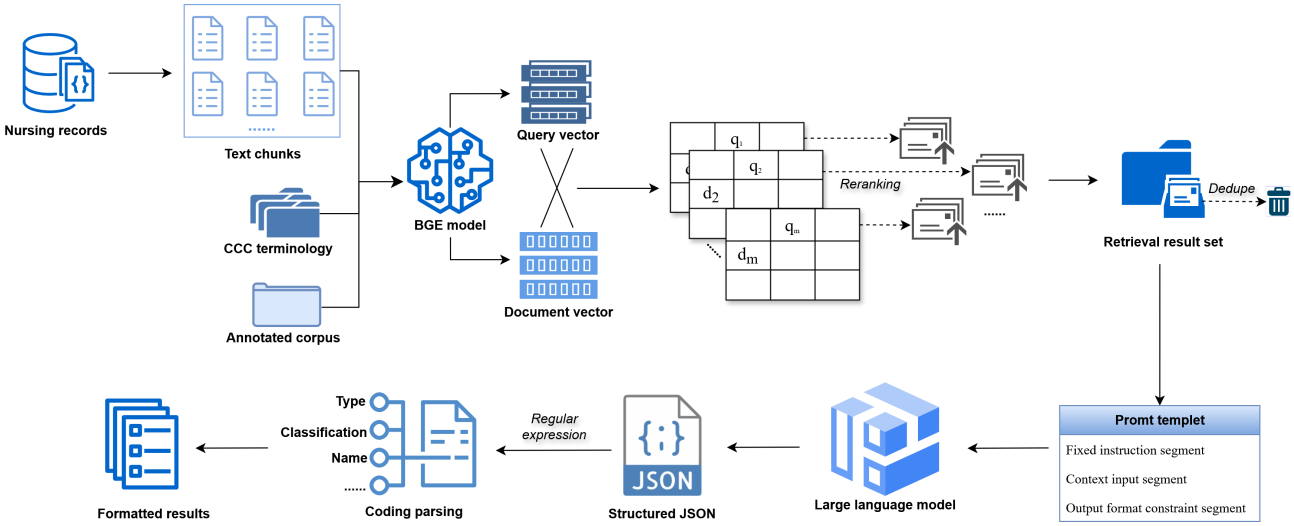

In the first stage, free-text records were segmented into text chunks using Chinese and English punctuation, while the original records were retained as subqueries. Given the small scale of the CCC terminology set and its abundant clinically equivalent paraphrases, we prioritized a fine-grained semantic matching strategy. Specifically, we used the BAAI (Beijing Academy of Artificial Intelligence) General Embedding–small-zh model to embed each subquery, all CCC reference entries, and the full annotated corpus into 512-dimensional dense vectors. Semantic similarity between subqueries, reference entries, and annotated corpora was computed using matrix-based cosine scoring. In detail, the embedding vectors of subqueries and CCC reference entries were L2 normalized to unit norm. Subsequently, matrix multiplication was performed to compute the dot product of these normalized vectors, generating a similarity matrix of dimensions (number of subqueries×number of reference entries), where each element quantifies the semantic similarity between an individual subquery and a reference entry. Retrieval results from all subqueries were aggregated, and the unique entries remaining after duplicate removal were sorted by their similarity scores. To determine the retrieval window, we evaluated retrieval coverage via Recall@k (k=1, 3, 5, and 10). Recall@5 provided the optimal balance between retrieval sensitivity and computational cost and was adopted as the default setting (Multimedia Appendix 2). For each subquery, the top 5 most relevant entries were retrieved separately from both the CCC reference entries and the annotated corpus, merged, and deduplicated to construct context windows.

In the second stage, prompts were generated by combining each subquery with its retrieved context following a predefined RAG template. This template enforces CCC coding guidelines, which incorporate the supplementary NANDA-I and NIC terms, and defines a structured JSON output schema (Figure 3). These prompts were then processed by a locally deployed LLM, which generated structured JSON outputs [24]. The resulting JSON was parsed with regular expressions, cross-referenced against the term database for metadata validation, and formatted into a standardized structure suitable for downstream analysis.

XSL•FO
RenderX

Figure 3. Clinical Care Classification (CCC) nursing terminology with retrieval-augmented mapping (CNTRAM) prompt. This figure shows the standardized prompt format used in CNTRAM and the expected JSON-based output. The prompt specifies that the model should return CCC nursing diagnosis and intervention codes as a JSON array, without explanatory text or duplication; an empty array is returned if no valid codes are identified.

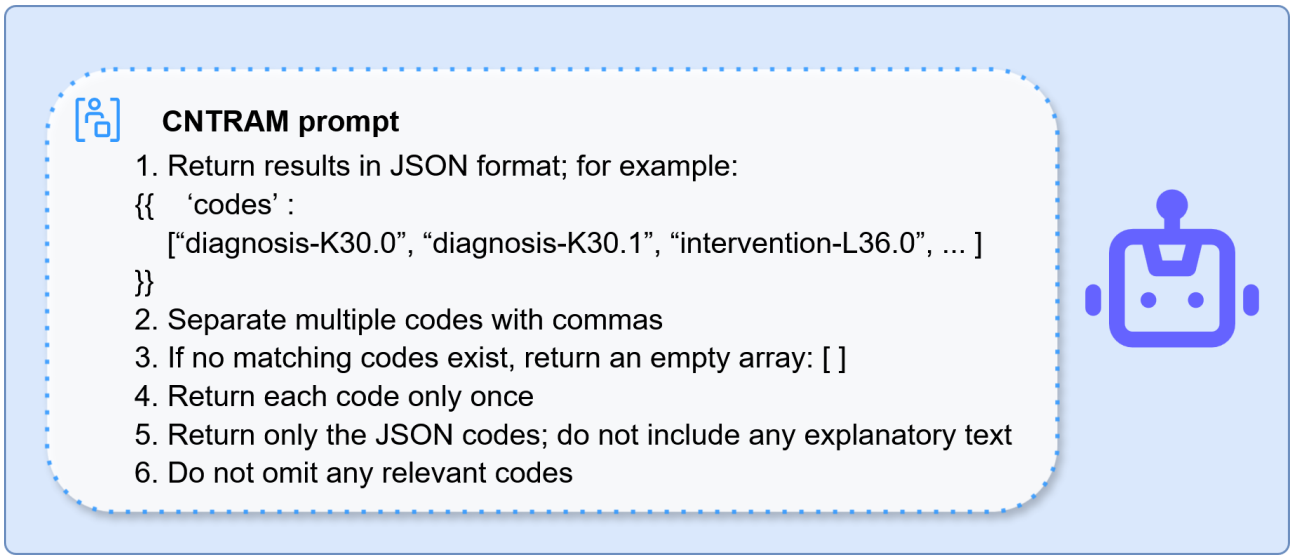

Experiments

To establish comparative baselines and evaluate CNTRAM's relative performance, multiple models were implemented, encompassing traditional statistical methods, pretrained language models, and LLMs.

Free-text inputs were transformed into term frequency–inverse document frequency (TF-IDF) vectors [25], and candidate CCC terms were ranked by cosine similarity. This served as a lexical-matching baseline reflecting rule-driven coding.

Contextual embeddings were generated using BERT [20]. Semantic similarity between text segments and CCC entries was calculated via cosine distance. This represented a rule-driven baseline with contextual semantic awareness. Additionally, BERT was fine-tuned on the annotated corpus of 1500 records to provide a supervised baseline for comparison.

Llama3.3-70B (Meta Platforms Inc) [26], Qwen3-14B (Alibaba Group) [27], Mistral-7B (Mistral AI) [28], and DeepSeek-R1 (DeepSeek) [29] were evaluated under zero-shot and few-shot prompting to test their intrinsic ability to align free-text nursing notes with CCC terms based solely on coding rules. Additionally, no-RAG experiments were conducted for each LLM to serve as a comparison.

Prompting Strategies

To examine the influence of contextual retrieval and example-based guidance, 2 prompting strategies were compared under the CNTRAM framework.

First, zero-shot prompting was used; for each generated subquery derived from both the segmented text and the original records, the top 5 most semantically relevant CCC coding rules were retrieved and incorporated into the prompt [30]. No example demonstrations were included.

For few-shot prompting, few-shot examples were selected by retrieving the top 5 most semantically relevant examples for each subquery generated from both the segmented text and the original records [31]. These examples were combined with retrieved CCC coding rules, which were generated following the same approach as zero-shot prompting, to form the final prompt. We further assessed the contribution of few-shot examples without retrieval using 20 static examples in the prompt.

Evaluation Metrics

Model performance was evaluated using precision, recall, F_1-score, and intersection over union (IoU), with each nursing record treated as a set of terms. CCC terms and supplementary NANDA-I and NIC terms were included in the evaluation. As a single record may contain multiple clinically related diagnoses and interventions, IoU was additionally used to more comprehensively assess the model's ability to identify all nursing terms, capturing their combinatorial characteristics. IoU is defined as the ratio of the intersection to the union of the predicted and gold standard term sets.

For a given record i, let G_i denote the gold standard term set and P_i denote the predicted term set. In this notation, "|·|" denotes set cardinality, "\" denotes set difference, and "∩" denotes intersection. The evaluation components were defined as follows:

TPi/Gi∩Pi/ ,FPi=/Pi Gi/ ,FNi=/Gi Pi/

Metrics for each sample were calculated as follows:

Precisioni=TPiTPi+FPi,Recalli=TPiTPi+FNi,F1i=2PrecisioniRecalliPrecisioni+Recalli

and

IoUi=TPiTPi+FPi+FNi

where TP_i is the number of correctly predicted terms in record i, FP_i is the number of predicted terms not in the gold standard set, and FN_i is the number of gold standard terms missed by the prediction. Overall performance was calculated as the macroaverage of each metric across all records.

XSL•FO
RenderX

Results

Characteristics of the Dataset

This dataset contained 197,337 records from 880 patients, representing an older cohort (median age 67, IQR 57-74 years) with a wide range of ICU lengths of stay (median 1 day, IQR 0-6 days), as shown in Table 1.

Table . Dataset characteristics (N=880).

Characteristics	Values
Nursing note entries, n	197,337
Age (years), median (IQR)	67 (57-74)
Female sex, n (%)	287 (32.6)
Length of stay in the intensive care unit (days), median (IQR)	1 (0 - 6)

Annotator Agreement and Coverage of CCC Terminology

Interrater reliability was evaluated using the Fleiss κ, which measures the degree of agreement among multiple annotators beyond chance. The results showed substantial consistency for both categories, with κ=0.65 for nursing diagnoses and κ=0.62 for nursing interventions (Table 2). These results indicate that the 3 nurses demonstrated a consistent application of CCC terminology during independent annotation, providing a solid foundation for constructing the final consensus-based reference standard used for model evaluation. In terms of terminology coverage, a total of 65 unique nursing diagnoses were observed in the annotated 100-record test set, corresponding to 36.9% of the 176 CCC diagnosis codes. For nursing interventions, 106 unique codes were present, representing 52.7% of the 201 CCC intervention codes.

Table . Fleiss κ for interrater agreement among 3 senior registered nurses.

Items	Observed agreement	Expected agreement	1 minus expected agreement	Fleiss κ
Nursing diagnoses	0.9886	0.9679	0.0321	0.6449
Nursing interventions	0.9830	0.9555	0.0445	0.6180

Model Performance

The CNTRAM framework demonstrated comprehensive superiority over TF-IDF, BERT, fine-tuned BERT, and zero-shot LLMs in mapping nursing records to standardized CCC terminology by integrating retrieval-enhanced and few-shot prompting. For nursing diagnoses (Table 3), the best-performing model was DeepSeek-R1. It achieved a precision of 0.7909, a recall of 0.7901, an F_1-score of 0.7836, and an IoU of 0.7614. All other approaches showed lower results across all metrics. Among traditional methods and language models, the F_1-score ranged from 0.0268 (BERT) to 0.5253 (Qwen3-14B with RAG+few-shot), and the IoU values ranged from 0.0142 (BERT) to 0.4987 (Qwen3-14B with RAG+few-shot).

XSL•FO
RenderX

Table . Model performance for mapping nursing diagnoses.

Models and prompting strategies	Precision	Recall	F_1-score	Intersection over union
Term frequency–inverse document frequency	0.0738	0.2733	0.0856	0.0685
BERT[a]	0.0155	0.2460	0.0268	0.0142
Fine-tuned BERT	0.2223	0.2233	0.2027	0.1798
No RAG[b]				
Mistral-7B	0.0396	0.0316	0.0299	0.0252
Qwen3-14B	0.0482	0.0406	0.0412	0.0394
Llama3.3-70B	0.0570	0.0511	0.0515	0.0511
DeepSeek-R1	0.0601	0.0582	0.0588	0.0580
RAG+zero-shot				
Mistral-7B	0.0727	0.0706	0.0716	0.0704
Qwen3-14B	0.0992	0.0804	0.0849	0.0800
Llama3.3-70B	0.1044	0.0906	0.0937	0.0903
DeepSeek-R1	0.2310	0.2241	0.2160	0.2003
RAG+few-shot				
Mistral-7B	0.4657	0.4374	0.4298	0.3914
Qwen3-14B	0.5614	0.5213	0.5253	0.4987
Llama3.3-70B	0.4819	0.4351	0.4455	0.4245
DeepSeek-R1	*0.7909*	*0.7901*	*0.7836*	*0.7614*

[a]BERT: Bidirectional Encoder Representations from Transformers.

[b]RAG: retrieval-augmented generation.

For nursing interventions (Table 4), DeepSeek-R1 again showed the best overall performance, with a precision of 0.8453, a recall of 0.8504, an F_1-score of 0.8413, and an IoU of 0.8097. Other traditional methods and language model setups exhibited lower performance, with F_1-score varying from 0.0077 (Mistral-7B with no RAG) to 0.5557 (Qwen3-14B with RAG+few-shot) and IoU values ranging from 0.0071 (Mistral-7B with no RAG) to 0.5072 (Qwen3-14B with RAG+few-shot).

XSL•FO
RenderX

Table . Model performance for mapping nursing interventions.

Model and prompting strategies	Precision	Recall	F_1-score	Intersection over union
Term frequency–inverse document frequency	0.2112	0.5508	0.2323	0.1597
BERT[a]	0.0818	0.3854	0.1200	0.0681
Fine-tuned BERT	0.2297	0.1686	0.1723	0.1290
No RAG[b]				
Mistral-7B	0.0079	0.0506	0.0077	0.0071
Qwen3-14B	0.0129	0.0148	0.0132	0.0129
Llama3.3-70B	0.0142	0.0131	0.0134	0.0130
DeepSeek-R1	0.0203	0.0182	0.0189	0.0166
RAG+zero-shot				
Mistral-7B	0.3885	0.2411	0.2744	0.1907
Qwen3-14B	0.4874	0.4147	0.4298	0.3174
Llama3.3-70B	0.4074	0.3932	0.3849	0.2782
DeepSeek-R1	0.4973	0.4903	0.4461	0.3563
RAG+few-shot				
Mistral-7B	0.5218	0.4049	0.4299	0.3561
Qwen3-14B	0.6018	0.5458	0.5557	0.5072
Llama3.3-70B	0.5757	0.4844	0.5052	0.4368
DeepSeek-R1	*0.8453*	*0.8504*	*0.8413*	*0.8097*

[a]BERT: Bidirectional Encoder Representations from Transformers.

[b]RAG: retrieval-augmented generation.

Prompting Strategies

We evaluated the impact of prompting strategies by comparing zero-shot and few-shot setups across all models. Few-shot prompting consistently improved performance for both nursing diagnoses and interventions. For nursing diagnoses (Figure 4), F_1-score across the 4 models ranged from 0.0716 to 0.216 under RAG+zero-shot and increased to a range of 0.4298 to 0.7836 under RAG+few-shot. Corresponding IoU values ranged from 0.0704 to 0.2003 with RAG+zero-shot and from 0.3914 to 0.7614 with RAG+few-shot. For nursing interventions (Figure 5), F_1-score ranged from 0.2744 to 0.4461 with RAG+zero-shot and improved to 0.4299 - 0.8413 with RAG+few-shot. IoU values ranged from 0.1907 to 0.3563 with RAG+zero-shot and from 0.3561 to 0.8097 with RAG+few-shot. In both tasks, DeepSeek-R1 with RAG+few-shot achieved the best performance. Ablation results for the no-RAG+few-shot configuration are provided in Multimedia Appendix 3.

XSL•FO
RenderX

Figure 4. Model performance for nursing diagnosis recognition under different prompting strategies: (left) retrieval-augmented generation (RAG)+zero-shot prompting and (right) RAG+few-shot prompting. This figure shows the performance of different large language models (Mistral-7B, Qwen3-14B, Llama3.3-70B, and DeepSeek-R1) under zero-shot and few-shot prompting conditions in recognizing Clinical Care Classification nursing diagnoses from free-text intensive care unit nursing records. Metrics include precision, recall, F_1-score, and intersection over union (IoU).

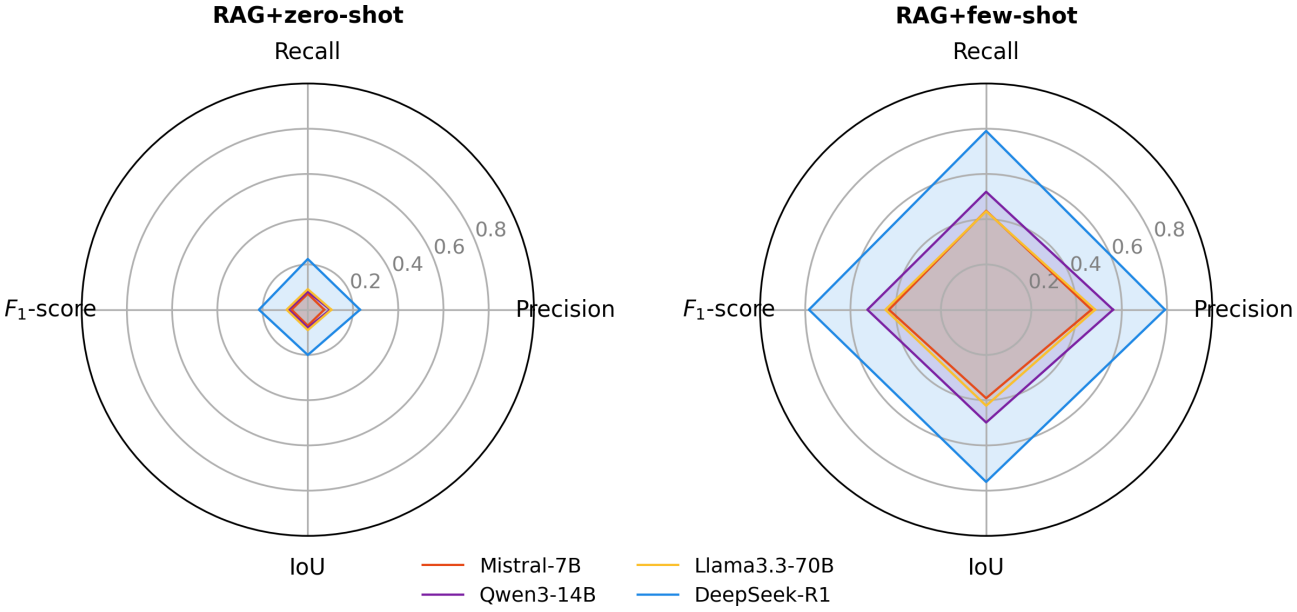

Figure 5. Model performance for nursing intervention recognition under different prompting strategies: (left) retrieval-augmented generation (RAG)+zero-shot prompting and (right) RAG+few-shot prompting. This figure shows the performance of different large language models (Mistral-7B, Qwen3-14B, Llama3.3-70B, and DeepSeek-R1) under zero-shot and few-shot prompting conditions in recognizing Clinical Care Classification nursing diagnoses from free-text intensive care unit nursing records. Metrics include precision, recall, F_1-score, and intersection over union (IoU).

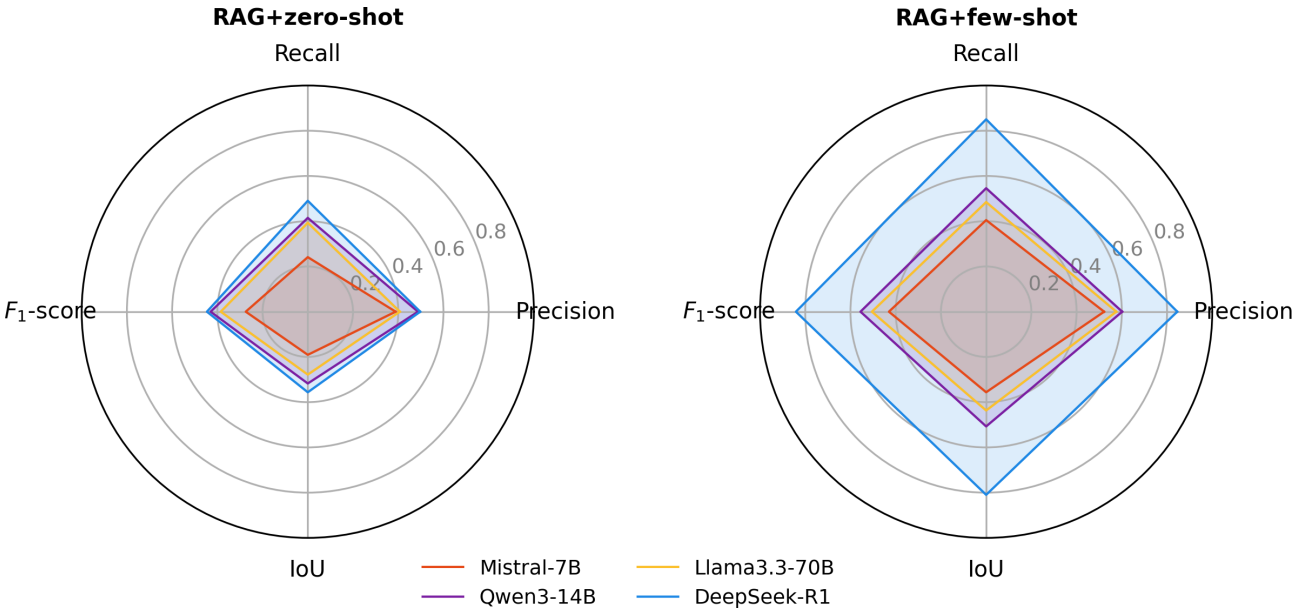

Discussion

Principal Findings

In this study, we developed CNTRAM, a retrieval-augmented terminology mapping framework integrating dense embedding retrieval, retrieval-enhanced prompting, and few-shot LLM guidance to automatically identify nursing diagnoses and interventions from free-text nursing records based on standardized CCC terminology. Our findings demonstrate that combining RAG with carefully designed few-shot examples markedly improves mapping accuracy compared with TF-IDF, BERT, fine-tuned BERT, and zero-shot LLMs. Specifically, our CNTRAM framework, based on DeepSeek-R1, achieved substantial performance improvements in both nursing diagnosis and intervention mapping. These results align with growing evidence that retrieval grounding improves the factual consistency of clinical text generation models [32-34].

The core challenge addressed by CNTRAM stems from the semantic characteristics inherent in nursing terminology. Nursing diagnoses are typically abstract conceptual constructs requiring inference, demanding that nurses synthesize multiple clues and arrive at diagnoses through reasoning [35,36]. Many diagnoses represent higher-order concepts that do not explicitly appear in nursing narratives. In contrast, nursing interventions

XSL•FO
RenderX

are more concrete, operationalized, and linguistically salient, typically expressed through explicit action verbs such as "monitor," "turn," or "assess." Previous analyses of CCC semantic structure have also noted that intervention concepts are more operationalized and lexically consistent, while diagnostic concepts may overlap semantically and contain broader conceptual boundaries, complicating automated classification [37,38]. These semantic differences likely explain the higher precision and F_1-scores observed in mapping nursing interventions and diagnoses in this study. This also underscores the need for technical approaches that support abstraction and reasoning capabilities during nursing terminology standardization.

On the basis of the professional attributes of nursing terminology, the performance advantages of the CNTRAM constructed in this study are primarily demonstrated through its 3 components. First, the RAG module retrieves contextually relevant CCC entries for each subquery, providing semantic grounding that reduces ambiguity and stabilizes concept alignment [39]. This design is consistent with growing evidence that retrieval augmentation improves reliability in medical question answering and EHR summarization [40]. Second, the 2-stage generation structure enforces strict adherence to CCC coding rules while producing stable JSON outputs that support downstream analysis. Finally, few-shot examples offer essential syntactic and semantic anchors for interpreting abstract or compositionally complex nursing expressions such as "airway clearance impairment," "aspiration risk," "infection risk," "venous catheter care," and "positioning therapy." Evidence from biomedical natural language processing similarly shows that combining retrieval with example-guided prompting markedly improves terminology normalization when annotated data are scarce [41].

Beyond task-level performance, CNTRAM maintains methodological continuity with broader concept normalization research while advancing terminology standardization in nursing. Within medical concept normalization, hybrid approaches that integrate rule-based matching, multilevel semantic similarity, machine learning, or multistage LLM workflows have become increasingly prominent for Unified Medical Language System or Systematized Nomenclature of Medicine–Clinical Terms normalization [42-44]. Meanwhile, nursing informatics literature has long emphasized that SNTs enhance interoperability and enable meaningful analysis of nursing patterns [45]. Previous cross-mapping studies demonstrated the feasibility of aligning narrative nursing documentation with vocabularies such as NANDA-I and NIC, although manual mapping workflows proved time consuming and inconsistent across institutions [46,47]. By demonstrating that retrieval-enhanced LLMs can adapt effectively to nursing-specific terminology systems and documentation practices, CNTRAM extends these research directions and offers a practical pathway for automated nursing terminology mapping in real-world environments. Additionally, in this study, the use of LLMs with varying parameter scales showed differing recognition performance. Among them, DeepSeek-R1 outperformed the other models, likely due to its stronger reasoning ability, instruction-following capacity, and better adaptation to Chinese clinical text. This reasoning advantage allows it to handle ambiguous or complex medical and nursing expressions more effectively [29]. Coupled with reliable instruction-following capacity, this ensures accurate mapping within the constrained output schema [48]. Moreover, its specialization in Chinese clinical data further enhances its ability to accurately map nursing terminology, which is essential given the unique linguistic and contextual features of Chinese clinical language [49].

From a nursing practice perspective, CNTRAM offers a scalable method for converting unstructured nursing records into structured, analyzable data. Identifying nursing practice features, such as nursing diagnoses and interventions, within practitioners' observation records helps validate the value of SNT for enhancing interoperability and the applicability of nursing data [45,50,51]. By identifying patients' nursing problems and clarifying corresponding nursing actions, CNTRAM can provide a clinically interpretable basis for risk stratification and outcome assessment. Recent observational evidence further supports this direction. Studies from Italian pediatric [52,53] and adult oncology settings [54] have shown that CCC-based nursing data can capture nursing care complexity and are associated with intrahospital and ICU transfers, diagnosis related group–based medical complexity, and length of stay. These findings reinforce the potential of structured nursing data to reflect patient risk and care demand. Accordingly, the CCC-coded information produced by CNTRAM offers a structured foundation that can be incorporated into future analyses of complexity, outcomes, and resource use. In addition, the high IoU score achieved in this study indicates the model's ability to effectively represent comprehensive sets of nursing diagnoses and interventions, reflecting the combinatorial nature of the nursing process. Notably, nursing diagnoses such as risk of infection and risk of aspiration require comprehensive judgment based on multiple clinical conditions and are prone to certain recognition errors [55]. Generalized nursing interventions, such as clinical measurements and physical examination, may have mapping biases due to unclear classification granularity [56]. Misclassification of nursing diagnoses can affect nurses' risk stratification assessment of patients, while misclassification of nursing interventions may disturb the calculation of nursing workload and the correlation between intervention frequency and patient risk [57]. In practical application, CNTRAM outputs can serve as an aid for clinical decision support and documentation, but they still require judgment and review by clinical nurses.

Limitations and Future Work

However, this study has several limitations. First, the analysis was based on data from a single ICU with a relatively limited dataset, and the test set covered only a subset of commonly used CCC codes. This partial coverage constrains our ability to generalize the findings to the full CCC terminology system. Second, although few-shot prompting improved model performance, it introduced a dependence on manually curated examples. Such reliance may lead to sampling bias. Third, as highlighted in recent health care RAG literature, robust evaluation frameworks are essential to mitigate issues related

XSL•FO
RenderX

to bias, trust, and privacy [58]. Future research could further validate the generalization capability of CNTRAM through multicenter and multilingual data verification. Additionally, the flexibility of CNTRAM to adapt to different nursing terminology systems, such as NANDA-I, NIC, and International Classification for Nursing Practice, could be examined. This would involve integrating diverse domain knowledge resources into the RAG framework to enhance semantic granularity and expand the coverage of terminology mapping [59]. Furthermore, embedding CNTRAM into EHR systems for real-time semantic standardization could provide actionable support for nursing practice and clinical decision-making.

Conclusions

In this study, we developed CNTRAM, an LLM-based 2-stage RAG framework designed to map free-text nursing records to the CCC terminology. CNTRAM combines dense embedding retrieval and few-shot prompting. Using DeepSeek-R1 as its backbone LLM, this framework outperformed traditional methods, no-RAG models, and zero-shot prompting approaches. The results demonstrate that the integration of RAG, few-shot learning, and LLMs improves mapping accuracy and offers a feasible solution for mapping unstructured nursing data to standardized terminologies.

Acknowledgments

The authors thank Zhang Lei for proofreading the manuscript. During the preparation of this manuscript, the authors used DeepSeek-R1 (DeepSeek) to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published paper.

Funding

The authors declare that no financial support was received for this study.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to institutional and ethical restrictions but are available from the corresponding author on reasonable request, subject to approval by the relevant institutional ethics committee. The code used in this study is publicly available [60].

Authors' Contributions

ML and DC, as co–first authors, contributed equally to this study and were jointly responsible for conceptualization and drafting the original manuscript. QX and ZH contributed to data interpretation and results review. YZ and JZ prepared the figures and tables. YL contributed to writing, review, and editing. HM, as the corresponding author, led study supervision and project administration and critically revised the manuscript. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary terms.
[DOCX File, 18 KB - jmir_v28i1e89850_app1.docx]

Multimedia Appendix 2
Recall@k Values for Different Retrieval Window k
[DOCX File, 16 KB - jmir_v28i1e89850_app2.docx]

Multimedia Appendix 3
Performance of large language models under “No retrieval-augmented generation” and “Few-shot prompting” for nursing diagnosis and intervention tasks.
[DOCX File, 19 KB - jmir_v28i1e89850_app3.docx]

References

1. Opmeer BC. Electronic health records as sources of research data. JAMA 2016 Jan 12;315(2):201-202. [doi: 10.1001/jama.2015.15419] [Medline: 26757473]
2. Anderson C, Kaul M, Gullapalli N, Pitani S. Electronic health records and clinical documentation in medical residency programs: preparing residents to become master clinicians. J Am Med Inform Assoc 2023 Nov 17;30(12):1965-1972. [doi: 10.1093/jamia/ocad158] [Medline: 37573135]

XSL•FO
RenderX

3. Salanterä S. Advanced use of electronic health records: the depth of nursing notes. Nurs Res 2015;64(6):411-412. [doi: 10.1097/NNR.0000000000000129] [Medline: 26505153]
4. von Gerich H, Moen H, Peltonen LM. Identifying nursing sensitive indicators from electronic health records in acute cardiac care-towards intelligent automated assessment of care quality. J Nurs Manag 2022 Nov;30(8):3726-3735. [doi: 10.1111/jonm.13802] [Medline: 36124426]
5. Macieira TGR, Chianca TCM, Smith MB, et al. Secondary use of standardized nursing care data for advancing nursing science and practice: a systematic review. J Am Med Inform Assoc 2019 Nov 1;26(11):1401-1411. [doi: 10.1093/jamia/ocz086] [Medline: 31188439]
6. Monsen KA, Heermann L, Dunn-Lopez K. FHIR-up! Advancing knowledge from clinical data through application of standardized nursing terminologies within HL7® FHIR®. J Am Med Inform Assoc 2023 Oct 19;30(11):1858-1864. [doi: 10.1093/jamia/ocad131] [Medline: 37428893]
7. Zhang T, Wu X, Peng G, et al. Effectiveness of standardized nursing terminologies for nursing practice and healthcare outcomes: a systematic review. Int J Nurs Knowl 2021 Oct;32(4):220-228. [doi: 10.1111/2047-3095.12315] [Medline: 33580632]
8. Saba VK, Taylor SL. Moving past theory: use of a standardized, coded nursing terminology to enhance nursing visibility. Comput Inform Nurs 2007;25(6):324-331. [doi: 10.1097/01.NCN.0000299654.13777.9f] [Medline: 18000428]
9. Feng RC, Tseng KJ, Yan HF, Huang HY, Chang P. A preliminary study on the use of clinical care classification in nursing documentation data sets. Comput Methods Programs Biomed 2013 Dec;112(3):713-719. [doi: 10.1016/j.cmpb.2013.07.019] [Medline: 24079965]
10. Kang MJ, Dykes PC, Korach TZ, et al. Identifying nurses' concern concepts about patient deterioration using a standard nursing terminology. Int J Med Inform 2020 Jan;133:104016. [doi: 10.1016/j.ijmedinf.2019.104016] [Medline: 31707264]
11. Clinical nursing terminology index. Zhongwei Nursing Information Research Institute. URL: http://zwini.org.cn/nurseinfo-web/page/nurseterms/index [accessed 2026-04-15]
12. Feng RC, Tseng KJ, Yan HF, Huang HY, Chang P. Capability of using clinical care classification system to represent nursing practice in acute setting in Taiwan. NI 2012 (2012) 2012;2012:113. [Medline: 24199066]
13. Choi M, Kim Y, Kim Y. Development of a mapping table for nursing notes based on nurses' concerns in ICU patients. Stud Health Technol Inform 2024 Jul 24;315:725-726. [doi: 10.3233/SHTI240299] [Medline: 39049400]
14. Torres FBG, Gomes DC, Hino AAF, Moro C, Cubas MR. Comparison of the results of manual and automated processes of cross-mapping between nursing terms: quantitative study. JMIR Nurs 2020;3(1):e18501. [doi: 10.2196/18501] [Medline: 34345784]
15. Elvas LB, Almeida A, Ferreira JC. Natural language processing in medical text processing: a scoping literature review. Int J Med Inform 2025 Dec;204:106049. [doi: 10.1016/j.ijmedinf.2025.106049] [Medline: 40706199]
16. Hossain E, Rana R, Higgins N, et al. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. Comput Biol Med 2023 Mar;155:106649. [doi: 10.1016/j.compbiomed.2023.106649] [Medline: 36805219]
17. Korach ZT, Cato KD, Collins SA, et al. Unsupervised machine learning of topics documented by nurses about hospitalized patients prior to a rapid-response event. Appl Clin Inform 2019 Oct;10(5):952-963. [doi: 10.1055/s-0039-3401814] [Medline: 31853936]
18. Redjdal A, Novikava N, Kempf E, Bouaud J, Seroussi B. Leveraging rule-based NLP to translate textual reports as structured inputs automatically processed by a clinical decision support system. Stud Health Technol Inform 2024 Aug 22;316:1861-1865. [doi: 10.3233/SHTI240794] [Medline: 39176854]
19. Lou Y, Zhu X, Tan K. Dictionary-based matching graph network for biomedical named entity recognition. Sci Rep 2023 Dec 8;13(1):21667. [doi: 10.1038/s41598-023-48564-w] [Medline: 38066007]
20. Nishigaki D, Suzuki Y, Wataya T, et al. BERT-based transfer learning in sentence-level anatomic classification of free-text radiology reports. Radiol Artif Intell 2023;5(2):e220097. [doi: 10.1148/ryai.220097] [Medline: 37035437]
21. Xiao H, Zhou F, Liu X, et al. A comprehensive survey of large language models and multimodal large language models in medicine. Inf Fusion 2025 May;117(C):102888. [doi: 10.1016/j.inffus.2024.102888]
22. Ke YH, Jin L, Elangovan K, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. NPJ Digit Med 2025 Apr 5;8(1):187. [doi: 10.1038/s41746-025-01519-z] [Medline: 40185842]
23. Xu R, Hong Y, Zhang F, Xu H. Evaluation of the integration of retrieval-augmented generation in large language model for breast cancer nursing care responses. Sci Rep 2024 Dec 28;14(1):30794. [doi: 10.1038/s41598-024-81052-3] [Medline: 39730573]
24. JSON Output. DeepSeek API Docs. URL: https://api-docs.deepseek.com/guides/json_mode [accessed 2026-04-15]
25. Bafna P, Pramod D, Vaidya A. Document clustering: TF-IDF approach. Presented at: 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT); Mar 3-5, 2016. [doi: 10.1109/ICEEOT.2016.7754750]
26. Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, et al. The Llama 3 herd of models. arXiv. Preprint posted online on Jul 31, 2024. [doi: 10.48550/arXiv.2407.21783]
27. Yang A, Li A, Yang B, Zhang B, Hui B, Zheng B, et al. Qwen3 technical report. arXiv. Preprint posted online on May 14, 2025. [doi: 10.48550/arXiv.2505.09388]

XSL•FO
RenderX

28. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS. Mistral 7B. arXiv. Preprint posted online on Oct 10, 2023. [doi: 10.48550/arXiv.2310.06825]
29. Guo D, Yang D, Zhang H, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature 2025 Sep;645(8081):633-638. [doi: 10.1038/s41586-025-09422-z] [Medline: 40962978]
30. Levy O, Seo M, Choi E, Zettlemoyer L. Zero-shot relation extraction via reading comprehension. Presented at: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017); Aug 3-4, 2017. [doi: 10.18653/v1/K17-1034]
31. Parnami A, Lee M. Learning from few examples: a summary of approaches to few-shot learning. ArXiv. Preprint posted online on Mar 7, 2022. [doi: 10.48550/arXiv.2203.04291]
32. Li M, Zhan Z, Yang H, et al. Benchmarking retrieval-augmented large language models in biomedical NLP: application, robustness, and self-awareness. Sci Adv 2025 Nov 21;11(47):eadr1443. [doi: 10.1126/sciadv.adr1443] [Medline: 41270181]
33. Koga S, Ono D, Obstfeld A. Retrieval-augmented generation versus document-grounded generation: a key distinction in large language models. J Pathol Clin Res 2025 Jan;11(1):e70014. [doi: 10.1002/2056-4538.70014] [Medline: 39817433]
34. Alkhalaf M, Yu P, Yin M, Deng C. Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records. J Biomed Inform 2024 Aug;156:104662. [doi: 10.1016/j.jbi.2024.104662] [Medline: 38880236]
35. Pirret AM, Neville SJ, La Grow SJ. Nurse practitioners versus doctors diagnostic reasoning in a complex case presentation to an acute tertiary hospital: a comparative study. Int J Nurs Stud 2015 Mar;52(3):716-726. [doi: 10.1016/j.ijnurstu.2014.08.009] [Medline: 25234268]
36. Smith SK, Benbenek MM, Bakker CJ, Bockwoldt D. Scoping review: diagnostic reasoning as a component of clinical reasoning in the U.S. primary care nurse practitioner education. J Adv Nurs 2022 Dec;78(12):3869-3896. [doi: 10.1111/jan.15414] [Medline: 35986584]
37. Fine S, Chaudhri A, Englebright J, Dan Roberts W. Nursing process, derived from the clinical care classification system components, as an earlier indicator of nursing care during a pandemic. Int J Med Inform 2023 May;173:104954. [doi: 10.1016/j.ijmedinf.2022.104954] [Medline: 36842361]
38. Jansen K, Kim TY, Coenen A, Saba V, Hardiker N. Harmonising nursing terminologies using a conceptual framework. Stud Health Technol Inform 2016;225:471-475. [Medline: 27332245]
39. Zhong M, Wu Z, Honda N. Deep learning based dense retrieval: a comparative study. ArXiv. Preprint posted online on Oct 27, 2024. [doi: 10.48550/arXiv.2410.20315]
40. Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. J Am Med Inform Assoc 2025 Apr 1;32(4):605-615. [doi: 10.1093/jamia/ocaf008] [Medline: 39812777]
41. Kresevic S, Giuffrè M, Ajcevic M, Accardo A, Crocè LS, Shung DL. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. NPJ Digit Med 2024 Apr 23;7(1):102. [doi: 10.1038/s41746-024-01091-y] [Medline: 38654102]
42. Chen L, Fu W, Gu Y, et al. Clinical concept normalization with a hybrid natural language processing system combining multilevel matching and machine learning ranking. J Am Med Inform Assoc 2020 Oct 1;27(10):1576-1584. [doi: 10.1093/jamia/ocaa155] [Medline: 33029642]
43. Yoon D, Han C, Kim DW, et al. Redefining health care data interoperability: empirical exploration of large language models in information exchange. J Med Internet Res 2024 May 31;26:e56614. [doi: 10.2196/56614] [Medline: 38819879]
44. Afshar M, Gao Y, Gupta D, Croxford E, Demner-Fushman D. On the role of the UMLS in supporting diagnosis generation proposed by large language models. J Biomed Inform 2024 Sep;157:104707. [doi: 10.1016/j.jbi.2024.104707] [Medline: 39142598]
45. Dunn Lopez K, Heermann Langford L, Kennedy R, et al. Future advancement of health care through standardized nursing terminologies: reflections from a Friends of the National Library of Medicine workshop honoring Virginia K. Saba. J Am Med Inform Assoc 2023 Oct 19;30(11):1878-1884. [doi: 10.1093/jamia/ocad156] [Medline: 37553233]
46. D'Agostino F, Zeffiro V, Vellone E, et al. Cross-mapping of nursing care terms recorded in Italian hospitals into the standardized NNN terminology. Int J Nurs Knowl 2020 Jan;31(1):4-13. [doi: 10.1111/2047-3095.12200] [Medline: 29328554]
47. Juvé Udina ME, Gonzalez Samartino M, Matud Calvo C. Mapping the diagnosis axis of an interface terminology to the NANDA international taxonomy. ISRN Nurs 2012;2012:676905. [doi: 10.5402/2012/676905] [Medline: 22830046]
48. Chen X, Liao B, Qi J, Eustratiadis P, Monz C, Bisazza A, et al. The SIFo benchmark: investigating the sequential instruction following ability of large language models. Presented at: Findings of the Association for Computational Linguistics: EMNLP 2024; Nov 12-16, 2024. [doi: 10.18653/v1/2024.findings-emnlp.92]
49. Yao Z, Duan L, Xu S, Chi L, Sheng D. Performance of large language models in the non-English context: qualitative study of models trained on different languages in Chinese medical examinations. JMIR Med Inform 2025 Jun 27;13:e69485. [doi: 10.2196/69485] [Medline: 40577654]
50. Jedwab RM, Holzhauser K, Raghunathan K, et al. Applicability and benefits of standardised nursing terminology in Australia: a scoping review. Collegian 2024 Dec;31(6):404-420. [doi: 10.1016/j.colegn.2024.10.001]

XSL•FO
RenderX

51. Ranegger R, Baumberger D, Grgic P, Jagfeld G. Semantic interoperability of nursing data - mapping an interface terminology to SNOMED CT. Stud Health Technol Inform 2024 Aug 22;316:1297-1301. [doi: 10.3233/SHTI240650] [Medline: 39176619]
52. Cesare M, D'Agostino F, Sebastiani E, Nursing And Public Health Group, Damiani G, Cocchieri A. Deciphering the link between diagnosis-related group weight and nursing care complexity in hospitalized children: an observational study. Children (Basel) 2025 Jan 17;12(1):103. [doi: 10.3390/children12010103] [Medline: 39857934]
53. Cesare M, Cocchieri A. Can an increase in nursing care complexity raise the risk of intra-hospital and intensive care unit transfers in children? A retrospective observational study. J Pediatr Nurs 2025;80:91-99. [doi: 10.1016/j.pedn.2024.11.015] [Medline: 39602875]
54. Cesare M, Magliozzi E, D'Agostino F, Zeffiro V, Cocchieri A. Prevalence and accuracy of nursing diagnoses in patients with malignant bronchial and lung cancer: a retrospective observational study. Eur J Oncol Nurs 2025 Aug;77:102931. [doi: 10.1016/j.ejon.2025.102931] [Medline: 40712222]
55. Hasegawa T, Ogasawara C, Katz EC. Measuring diagnostic competency and the analysis of factors influencing competency using written case studies. Int J Nurs Terminol Classif 2007;18(3):93-102. [doi: 10.1111/j.1744-618X.2007.00057.x] [Medline: 17714237]
56. Coenen A, Ryan P, Sutton J. Mapping nursing interventions from a hospital information system to the Nursing Interventions Classification (NIC). Nurs Diagn 1997;8(4):145-151. [doi: 10.1111/j.1744-618x.1997.tb00469.x] [Medline: 9624992]
57. Müller-Staub M, Lavin MA, Needham I, van Achterberg T. Nursing diagnoses, interventions and outcomes - application and impact on nursing practice: systematic review. J Adv Nurs 2006 Dec;56(5):514-531. [doi: 10.1111/j.1365-2648.2006.04012.x] [Medline: 17078827]
58. Amugongo LM, Mascheroni P, Brooks S, Doering S, Seidel J. Retrieval augmented generation for large language models in healthcare: a systematic review. PLOS Digit Health 2025 Jun;4(6):e0000877. [doi: 10.1371/journal.pdig.0000877] [Medline: 40498738]
59. Zhao X, Liu S, Yang SY, Miao C. MedRAG: enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. Presented at: WWW '25: Proceedings of the ACM on Web Conference 2025; Apr 28 to May 2, 2025. [doi: 10.1145/3696410.3714782]
60. GitHub. URL: https://github.com/miye-ha/CNTRAM [accessed 2026-04-17]

Abbreviations

BAAI: Beijing Academy of Artificial Intelligence
BERT: Bidirectional Encoder Representations from Transformers
CCC: Clinical Care Classification
CNTRAM: Clinical Care Classification nursing terminology with retrieval-augmented mapping
EHR: electronic health record
ICU: intensive care unit
IoU: intersection over union
LLM: large language model
NANDA-I: North American Nursing Diagnosis Association International
NIC: Nursing Interventions Classification
RAG: retrieval-augmented generation
SNT: standardized nursing terminology
TF-IDF: term frequency–inverse document frequency

Edited by A Coristine; submitted 18.Dec.2025; peer-reviewed by J Yang, M Cesare; revised version received 29.Mar.2026; accepted 29.Mar.2026; published 29.Apr.2026.

Please cite as:
Li M, Chen D, Xiao Q, He Z, Zhang Y, Zhong J, Luo Y, Ma H
Automated Identification of Nursing Diagnoses and Interventions From Nursing Records Using a Retrieval-Augmented Large Language Model Approach: Quantitative Study
J Med Internet Res 2026;28:e89850
URL: https://www.jmir.org/2026/1/e89850
doi:10.2196/89850

© Minye Li, Dianjie Chen, Qun Xiao, Zhengji He, Yanyan Zhang, Jinlan Zhong, Yiwei Luo, Hui Ma. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Apr.2026. This is an open-access article distributed under

XSL•FO
RenderX

the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Temporal Annotation of German Clinical Language in Real and Synthetic Clinical Documents: Corpus Development and Baseline Tagger Validation Study

Luise Modersohn[1,2], MSc; Udo Hahn[2,3], Prof Dr

[1]Chair of Medical Informatics, Institute for AI and Informatics in Medicine, Technical University of Munich (TUM) university hospital, Munich, Bavaria, Germany

[2]JULIELab, Friedrich Schiller University Jena, Jena, Germany

[3]Institute for Medical Informatics, Statistics and Epidemiology (IMISE), Leipzig University, Leipzig, Germany

Corresponding Author:
Luise Modersohn, MSc
Chair of Medical Informatics
Institute for AI and Informatics in Medicine
Technical University of Munich (TUM) university hospital
Ismaninger Str. 22
Munich, Bavaria, 81675
Germany
Phone: 49 8941404320
Email: luise.modersohn@tum.de

Abstract

Background: Temporal information about patients constitutes a precious source for clinical decision-making and medical treatment. The automatic extraction of such data from unstructured clinical narratives requires time-annotated clinical reports and notes from which time-informed taggers can be learned. Unfortunately, the non-English clinical language community, the German one as a typical example, with only a few exceptions, generally lacks such time-annotated resources to train and evaluate temporal taggers.

Objective: To overcome this metadata bottleneck, we developed a TimeML-conformant annotation schema for both temporal entities and temporal relations adapted to the needs of German medical language. Based on the annotations derived therefrom, we trained state-of-the-art baseline taggers to recognize temporal expressions in clinical documents.

Methods: Starting from temporal annotation guidelines for English clinical documents, we developed preliminary annotation guidelines for temporal named entities and temporal relations for the German language. These guidelines were subsequently refined and adapted to German clinical jargon, incorporating the work experience of 5 clinically trained annotators (students of medicine). For this task, we used randomly selected smaller subsets of 2 German clinical corpora—a real-world one ($3000PA_J$) and a synthetic one (GraSCCo). Both corpora were annotated ($3000PA_J$ partially, GraSCCo completely), randomly selecting 10% of the documents as an agreement part on $3000PA_J$. To measure interannotator agreement (IAA), we computed pairwise F_1-scores. We used that metadata to develop BERT (Bidirectional Encoder Representations from Transformers)-based language models for the creation of time-sensitive baseline taggers. All annotations are based on TimeML, the international de facto standard for time information markup.

Results: We created $3000PA_J$-temp, a time-annotated corpus of real clinical documents (which cannot be distributed because of the rigid privacy legislation enforced for German clinical data), and GraSCCo-temp, a synthetic one (which is publicly available without any restrictions). Based on the final guidelines, we achieved an IAA F_1-score of 0.9 on both corpora for the temporal named entity recognition task. For the temporal relation extraction task, the IAA on GraSCCo plummeted to an F_1-score of 0.57 and 0.41 on $3000PA_J$, respectively. Still, those results are comparable with English clinical datasets. Our baseline tagger for named entities reached F_1-scores between 0.64 and 0.85. For automatic relation extraction, we achieved F_1-scores ranging between 0.60 and 0.64.

Conclusions: We here introduce the first TimeML-compliant annotation scheme for time expressions occurring in German clinical language and apply it to 2 clinical corpora, one with nondistributable real clinical data, the other with distributable synthetic

XSL•FO
RenderX

ones. The latter constitutes the first publicly accessible, temporally annotated clinical corpus for the German language. The time tagger trained on these datasets is the first of its kind, fully compliant with the TimeML markup language. The amounts of temporal metadata in our corpora are among the largest datasets ever produced for the clinical domain, both compared with English and German predecessors.

(J Med Internet Res 2026;28:e71458) doi:10.2196/71458

KEYWORDS

natural language processing; clinical NLP; German clinical language; corpus development; temporal named entity; temporal relations; annotation guidelines; annotation process

Introduction

Overview

Information extraction is a field of natural language processing (NLP) that is concerned with automatically identifying mentions of entities of interest (so-called named entities) and semantic relations holding among them [1-3]. In the medical domain, commonly researched named entity types are, for example, Symptom, Finding, Diagnosis, Disease, or Drug, that are conceptually linked via medically relevant semantic relation types, such as Drug – hasDosage – Dosage, Drug – hasEffectOn – Disease (examples are shown in Figures 1A and 1B). Whereas named entity types address the terminological level of medical knowledge, including its taxonomic structure, semantic relations help represent single pieces of factual knowledge (medical assertions or facts).

Sets of medical entities and facts identified by information extraction tools form a bag of information units lacking further structure. One particularly important dimension to add structure to otherwise unrelated information pieces focuses on their temporal ordering on the time axis, a class of information we here refer to as temporal knowledge [4-7]. Information about the temporal order of (clinical) events is crucial in a hospital setting, because, for example, a drug reported on admission has a medically different status compared with one prescribed at discharge. Ultimately, temporal knowledge can be used to visualize timelines of patients' histories, which might prove helpful for doctors in medical decision making and for nurses in their daily care routines. This type of knowledge can either be encoded in terms of the named entity types “Event” and “Temporal Expressions,” such as Date, Duration, and Time, and “temporal relations,” which impose a (temporal) ordering on named entities or semantic relations. For instance, we may state a temporal relation between named entities, such as AFTER (examples are shown in Figure 1C), or between semantic relations, such as BEFORE (examples are shown in Figure 1D).

Prior to the automatic extraction of named entities, semantic and temporal relations, specialized classifiers (taggers) have to be pretrained or fine-tuned, typically using manually supplied annotations to guide the machine learning process. We here face a serious disparity between corpora for English and non-English languages (such as German, the focus of our work [8]). Based on an extensive survey of temporal datasets and tagging tools in the Related Work section, we encounter a large number of temporally annotated resources for the English language, both for clinical and nonclinical domains, that have no comparable counterpart for the German (clinical and nonclinical) language. For German, such resources are scarce and, if available, rather small in the nonclinical domain, while in the clinical domain, they are almost nonexistent. Moreover, annotations obeying the full TimeML annotation language standard are entirely missing for German.

Accordingly, in the Methods section, the unique contributions of this article are described in depth: (1) the provision of 2 clinical German-language corpora, a nondistributable real clinical corpus ($3000PA_J$) and a distributable synthetic corpus (GraSCCo), both of which have been richly annotated, adhering to TimeML, the current meta-language standard for temporal annotations. $3000PA_J$ and GraSCCo are the first corpora annotated with TimeML markup for the German language ever. (2) Exploiting these annotations, we trained the first transformer-based baseline taggers that automatically extract temporal information encoded in German clinical reports and notes. Up until now, transformer-based temporal taggers are not even available for the general German language (eg, covering news or social media content).

XSL•FO
RenderX

Figure 1. Examples for named entities, semantic, and temporal relations.

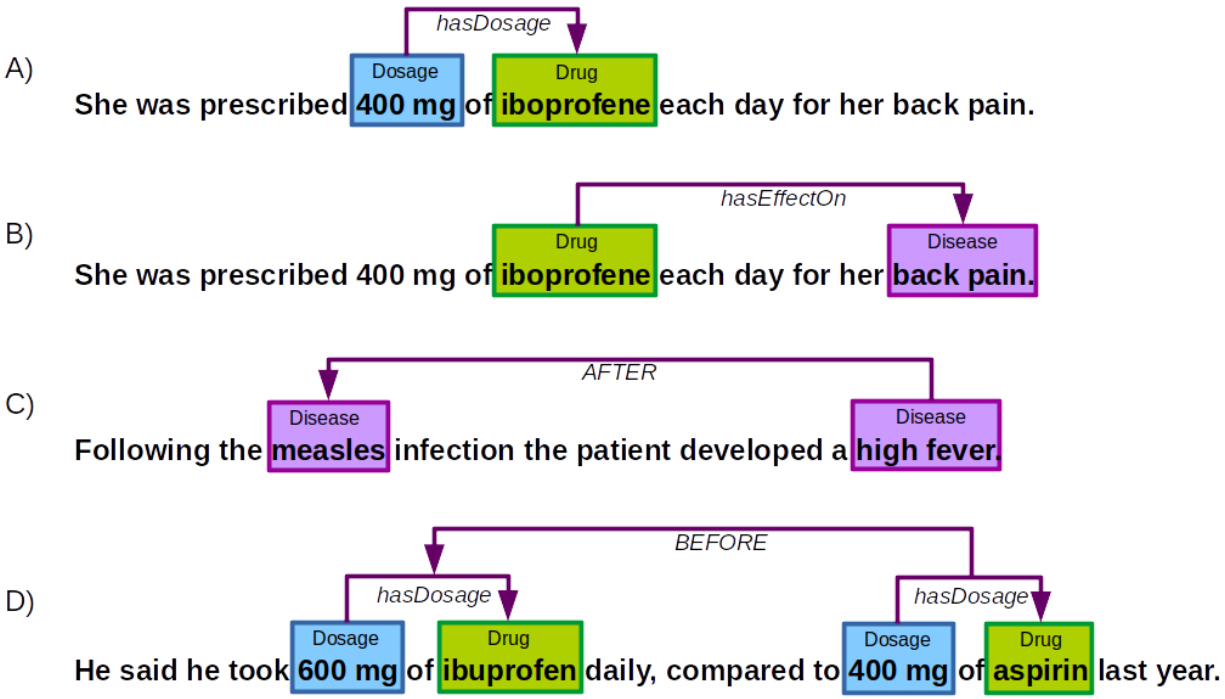

Related Work

Temporal information about patients constitutes a precious source for clinical decision making and medical treatment. It forms cohesive ties between single factual statements by linking them via temporal relations (see the examples for AFTER and BEFORE in Figure 1), which are the basis for temporal reasoning (eg, AFTER (x,y) and AFTER (y,z) transitively imply AFTER (x,z)). The automatic extraction of such temporal data is a prerequisite for the overarching goal to automatically construct consistent, maximally complete, and chronologically coherent timelines of the patients' medical history from unstructured clinical documents and structured electronic medical records to help support medical diagnosis, treatment decisions, and prognosis [9-20].

The automatic extraction of such temporal information from unstructured medical records (clinical reports, notes, or free text fields in the electronic medical record) requires language models trained on raw or metadata-enriched, that is, annotated, textual data. However, annotated temporal data resources are still quite rare and often seriously limited in size. Shortage of corpora with temporal metadata, however, directly hinders supervised learning of time-sensitive language models.

The starting point of time-centered knowledge representation research dates back to Allen's time algebra [21], which was then adopted in the design of TimeML [22], a specification and annotation language for temporal natural language utterances that combines temporal entities (Dates, Durations, Events, etc) and relations (eg, BEFORE, AFTER, OVERLAP); temporal relations impose temporal orderings on temporal entities. After years of experimentation, TimeML evolved into an ISO (International Organization for Standardization) standard as ISO-TimeML [23]. Part of TimeML, and 2 of its historical predecessors, are TIDES TIMEX2 [24] and TIMEX3 [23], which are used to mark up explicit temporal expressions, such as time expressions, dates, durations, etc. Table 1 illustrates the dominant role of TimeML as the representational foundation (markup or annotation language) for time-annotated corpora.

XSL•FO
RenderX

Table 1. Time-focused (mostly English only) nonclinical corpus resources.

Corpus/(challenge) task (year)	Functional characterization of (challenge) task	Annotation language	Text genre	Size in tokens	Type/number of metadata
TimeBank 1.1 [25] (2003)	N/A[a] (corpus)	TimeML 1.1	186 news articles	68.5k	13 temporal relation types: 5.1k temporal relation annotations; 11.2k annotation items, in total
TimeBank 1.2 [26] (2007); for error correction see the work of Ocal et al [27]	N/A (corpus)	TimeML 1.2	186/183 news articles	68.5k	1.4k TIMEX3 annotations; 7.9k Event annotations; 13 temporal relation types (6.4k temporal relation annotations and 27.6k annotation items, in total)
TimeBank-Dense [28,29] (2014)	N/A (corpus)	Subset of TimeML 1.2	36 news articles (subset of TimeBank 1.2)	13.3k (estimated)	0.3k TIMEX3 annotations; 1.7k Event annotations; 6 temporal relation types (12.7k [10.8k] temporal relation annotations; 14.7k annotation items, in total)
TempEval-1 [30] (2009)	Temporal relation (TLink) identification	Subset of TimeML 1.2 (TIMEX3, Event, TLink)	182 news articles (subset of TimeBank 1.2)	61.4k [31]	6 temporal relation types: 5.8k temporal relation annotations; 13.9k annotation items, in total [28]
TempEval-2 [32] (2010)	TIMEX3 identification; Event identification; temporal relation (TLink) identification	Subset of TimeML 1.2 (TIMEX3, Event, TLink)	Same as TempEval-1	63k	6 temporal relation types: 4.9k temporal relation annotations; 12.7k annotation items, in total (English data only) [28]
TempEval-3 [31] (2013)	TIMEX3 extraction and normalization; Event extraction and classification; end-to-end temporal relation (TLink) identification	Subset of TimeML 1.2 (TIMEX3, Event, TLink)	182 news reports (TimeBank) + 73 news reports(tiny subset of≈1M)(AQUAINT [33]);255 news articles, in total	Gold: 61.4 k (TimeBank) + 6.4 k +34.0 k (tiny subset of AQUAINT [33]) = 101.8k, in total; Silver: 666k (Gigaword [32])	13 temporal relation types (full TimeML set): 11.1k temporal relation annotations; 24.2k annotation items, in total (English data only) [28]
SemEval 2015 - TimeLine Task 4 [34] (2015)	Cross-document analysis of news articles to construct timelines for target events, that is, detecting, anchoring in time, and ordering events for a target entity	Subset (and minor adaptation) of TimeML 1.2	120 WikiNews articles, with 4 main topics: Apple, Airbus & Boeing, General Motors, Chrysler & Ford, stock market	10k per main topic, 40k, in total	0.9k Event annotations;0.7k Event chain annotations; 38 target entities; 37 timelines
RED [35] (2016)	N/A (corpus)	Subset of TimeML	95 news articles and casual discussion forum chats	54.3k tokens	1.1k TIMEX3 annotations; 8.7k Event annotations; 5 TLink types (Before, Overlap, Begins_on, Ends_on, Simultaneous): 1.4k temporal relation annotations (TLinks)
CaTeRS [36] (2016)	N/A (corpus)	Event definition based on the TRIPS ontology; subset of TimeML TLINKs	320 five-sentence short stories	1600 sentences	2.7k Event annotations;4 TLink types (Before, Overlaps, Contains, Identity): approximately 2.2k temporal relation annotations (TLinks) out of 2.7k causal and temporal relations
Event StoryLine (ESC) v0.9 [37] (2017)	N/A (corpus)	Compatible with TimeML	258 news articles about calamityevents, that is, natural disasters, shootings, killings, accidents, trials, etc.	N/A	1.3k TIMEX3 annotations; 7.3k Event annotations; 4 TLink types (Before, After, Contains, Overlap): 6.9k temporal relation annotations (TLinks)

XSL•FO
RenderX

Corpus/(challenge) task (year)	Functional characterization of (challenge) task	Annotation language	Text genre	Size in tokens	Type/number of metadata
Universal Decompositional Semantics Time (UDS-T) [29] (2019)	Categorical temporal relations are represented as real-valued relative timelines (on top of Universal Dependency parse trees)	Mapping of events to their likely durations and event pairs as real-valued relative timelines	Mixed Web documents (blog posts, newsgroup threads, emails, product reviews, and answers from QA websites) [38]	254.8k tokens [38]	32.3k Event annotations; 70.4k temporal relation annotations
TDG [39] (2020)	N/A (corpus)	Subset of TimeML + Temporal Dependency Graphs	500 WikiNews articles	Not reported	2.5k TIMEX3 annotations; 15.0k Event annotations; 4 TLink types (Before, After, Overlap, Includes): 28.4k temporal relation annotations (TLinks)
Maven-Ere [40] (2022)	N/A (corpus)	Subset of TimeML: TIMEX3, TLink	4480 Wikipedia articles	Not reported	25.8k TIMEX3 annotations; 6 TLink types (Before, Contains, Overlap, Begins_on, Ends_on, Simultaneous): 1216k temporal relation annotations (TLinks)
TimeBankNT [41] (2024)	Full timeline annotation (all possible TLinks) of the sparse TLink TimeBankDense [28] annotations	Subset of TimeML: TIMEX3, Event, TLink	36 news articles(subset of TimeBank 1.2)	13.3k (estimated)	0.3k TIMEX3 annotations; 1.7k Event annotations; 102.3k (104.8k) temporal relation annotations (TLinks)

[a]N/A: not applicable.

The early TimeBank corpus [25,26,28,29] witnessed the consolidation of TimeML over a decade. As an experimental vehicle, it is characterized by small document and token numbers as well as low amounts of annotated temporal data.

A simplified version of expressionally rich TimeML became the metadata backbone for TempEval [30-32,42,43], a series of time-focused shared task challenges run by NLP researchers within the framework of SemEval, a workshop series focusing on a broad range of semantic interpretation topics organized under the auspices of the Association for Computational Linguistics.

Shared tasks, such as TempEval or SemEval, have become a major driver of progress for NLP [44,45]. They are based on a common gold standard dataset on which several competing research group teams run their own systems against predefined tasks; their submitted solutions are, finally, compared against the gold standard, which yields a ranking of the teams involved based on their performance for specific tasks. Top-ranked teams are then given the opportunity at a closing workshop and invited to journal publications to feature the most successful methodologies—thus, continuously advancing the methodological repertoire of NLP on the basis of empirical evidence from benchmarking experiments.

The TempEval-3 corpus [31] was the largest in this series and contains 255 news articles, with roughly 95k tokens, selected from the TimeBank 1.2 corpus [25,26] (183/186 articles) and the Aquaint corpus (73 news reports) [33] (numbers slightly differ dependent on the cited literature); TimeBank-Dense [28,29], its relationally richer successor, comes with a total of 12,715 temporal relations (more than 2 times the number of the original TimeBank corpus). Interestingly, the second edition of TempEval [32] contained data not only for English, but also for Chinese, French, Italian, Korean, and Spanish, whereas the third edition [31] dealt with English and Spanish data only. A variant of TempEval was considered in the Timeline Task 4 of SemEval 2015, which featured cross-document analysis of news articles to construct timelines for events [34] described in a series of documents. For all 3 challenges in TempEval, 9 subtasks had to be tackled that fell into 3 categories: time expression identification (spans and TIMEX3 classes, such as Date, Time, Duration, etc), event expression identification (spans and Modality, Degree, Polarity, etc), and temporal relation identification (temporal relations, such as BEFORE, OVERLAP, AFTER, and a narrative container relation typed as CONTAINS).

In an intermediary phase, new corpora were generated featuring different text genres (eg, user-generated contents [35,36,46], Wikipedia articles [40]) and playing with different sets of temporal relations, yet with only insignificant growth of document sets or metadata volumes. One of the rare departures from strict adherence to TimeML can be found in the UDS-T corpus [46], where TimeML's categorical temporal relations are represented as real-valued relative timelines (on top of Universal Dependency parse trees). UDS-T also stands for a trend towards increasing quantities of documents/tokens (>250k tokens) and metadata (>100k annotation items). Truly impressive breakthroughs in both of these dimensions have been achieved only recently by getting large numbers of crowdworkers involved, such as for Maven-Ere [40], which excels in, for example, more than a million temporal relation annotations. In addition, TimeBankNT [41] increases the number of metadata by an order of magnitude (>100k temporal relation annotations) relative to TimeBank-Dense. Almost all of the corpora in Table 1 incorporate newswire/newspaper documents only (with some exceptions mentioned above). In contradistinction, Table 2 features corpora with clinical contents.

XSL•FO
RenderX

Table 2. Time-focused (English only) clinical corpus resources.

Corpus/(challenge) task (Year)	Functional characterization of (challenge) task	Annotation language	Text genre	Size in tokens	Type/number of metadata
CLEF [47] (2009)	N/A[a] (corpus)	TIMEX3 & self-defined set of 8 temporal relations (Before, After, Overlap, Is_included, Ended_by, Begun_ by)	50 clinical reports; 50 histopathology reports;50 imaging reports	Not reported	Clinical named entity types (Drug-or-Device, Intervention, Condition, Investigation, Locus, etc); 3.8k named entity annotations; clinical relation types (has_Target, has_Finding, has_Location, Modifies, etc); 2.4k clinical relation annotations; 0.5k temporal relation annotations (for 10 documents, only)
i2b2 2012 [48] (2013)	Spans and types/attributes of temporal expressions (TIMEX3), referring to Dates, Times, Durations (incl. Normalization to ISO[b]8601 standard); Spans and types/attributes of Events, including. Both clinical concepts, such as Problems, Tests, Treatments, and Events relevant to the patient's clinical timeline, such as Admissions, Transfers between departments; Temporal relations between the clinical Events and temporal expressions (Before, After, Overlap)	Simplified version of TimeML adapted to clinical narratives and purposes (eg, the inclusion of clinical concepts, such as Problems, Tests, and Treatments, into Events) [49]	310 discharge summaries	About 178k tokens	2.4k (train) + 1.8k (test) annotation units for temporal expressions (TIMEX3); 15.6k (train) + 13.6k (test) annotation units for temporal Events; 33.5k (train) + 27.7k (test) annotation units for temporal relations; 94.6k annotation units, in total
THYME [50] (2014)	N/A (corpus)	ISO-TimeML (THYME-TimeML) [23]; TIMEX3, Event, 5 temporal relations (Before, Overlap, Begins_on, Ends_on, Contains)	1254 clinical notes and pathology reports (oncology: brain cancer, colon cancer)	Not reported	Subset of the THYME colon cancer corpus (107 documents): 1.4k annotation units for temporal expressions (TIMEX3); 15.8k annotation units for temporal Events; 7.9k annotation units for temporal relations; 25.1k annotation units, in total
Clinical TempEval-1, subbranch within TempEval as part of SemEval [51] (2015) and Clinical TempEval-2 [52] (2016)	Spans and types/attributes of temporal expressions (TIMEX3), referring to Dates, Times, Durations, etc; Spans and types/attributes of Events, including both clinical concepts, such as Problems, Tests, Treatments; temporal relations between the clinical events and the document creation time (Before, After, Overlap, Before-Overlap) and narrative container relations	Extension of TimeML for temporal expressions, medical events, and temporal relations between events and times (Styler et al [50] shows the details of extensions of TimeML)	591 clinical notes and pathology reports (colon cancer)	Not reported	7.9k annotation units for TIMEX3 time expressions; 78.9k annotation units for Events; 23.2k annotation units for temporal relations; 110.0k annotation units, in total

XSL•FO
RenderX

Corpus/(challenge) task (Year)	Functional characterization of (challenge) task	Annotation language	Text genre	Size in tokens	Type/number of metadata
Clinical TempEval-3 [53] (2017)	Same tasks as in Clinical TempEval-1 and Clinical TempEval-2	Same annotation language as in Clinical TempEval-1 and Clinical TempEval-2	Same colon cancer dataset as in Clinical TempEval-1 and Clinical TempEval-2 + a second dataset (595 documents) with focus on brain cancer	Not reported	(Virtually the same number of annotation items for colon cancer as in Clinical TempEval-1 and Clinical TempEval-2) 6.6k annotation units for TIMEX3 time expressions; 48.9k annotation units for Events; 7.3k annotation units temporal relations; 62.8k annotation units for colon cancer, in total; 172.8k annotation units; in total.
n2c2 [54], follow-up event of i2b2 2012 (2023)	Identification of medication change (Event): determine the Action, Negation, Temporality, Certainty, and Actor for any change events (Context)	Categories for temporality (Past, Present, Future, Unknown) are considerably coarser in granularity than those from the i2b2 2012 Shared Task and are not grounded in TimeML	500 clinical notes (discharge summaries, correspondence notes)	About 308k tokens (estimated from the study by Kumar et al [55])	About 1.7k annotation units
CALEX[56] (2019)	N/A (corpus)	TimeML-based TIMEX3 (Date, Time, Duration, Frequency) plus an extension (CALEX) for timeline construction based on calendar expressions	180 discharge summaries, 53 psychiatry reports, 70 pediatrics reports, 75 emergency room reports, 378 clinical reports, in total	55,433 (discharge summaries); 67,569 (psychiatry); 36,675 (pediatrics); 52,041 (emergency), 211.7k clinical reports, in total	2378 TIMEX3 annotations

[a]N/A: not applicable.

[b]ISO: International Organization for Standardization.

The CLEF corpus [47] was the first clinical time-focused dataset ever, with limited size though (150 documents, <7k annotated items from which only a small fraction cover temporal information), restricted to TIMEX expressions, yet with a rich repertoire of named entities, (self-defined) semantic and temporal relations. It became the forerunner of the seminal THYME corpus [50], a flagship enterprise that, unlike CLEF, fully adhered to TimeML representation conventions and came with roughly 1.3k clinical reports and slightly more than 25k temporal annotations on a 107-document subset thereof.

Fortunately, the idea of challenge competitions also attracted clinical NLP researchers. Inspired by TempEval, temporal aspects were targeted within the i2b2 2012 Shared Task on Temporal Relations in Clinical Text [48], here with focus on time information contained in clinical reports. Participating systems were required to extract virtually the same types of temporal knowledge as those targeted by TempEval:

1. Spans and types/attributes of temporal expressions, referring to Dates, Times, Durations, or frequency phrases in the clinical text (subsequently, the values of the extracted TIMEX temporal expressions had to be normalized to the ISO8601 specification standard).
2. Spans and types/attributes of clinically significant events, including both clinical concepts (ie, named entities) such as Problems, Tests, Treatments, and ClinicalDepartments, and events relevant to the patient's clinical timeline, such as Admissions, Transfers between departments.
3. Temporal relations between the clinical events and temporal expressions.

For these challenge tasks, the i2b2 2012 Shared Task organizers added a temporal annotation layer to the dataset, 310 discharge summaries (about 178k tokens), that had already been used in previous i2b2 rounds. For annotation purposes, a simplified version of TimeML was taken as a starting point, but had to be adapted to better suit clinical narratives and purposes (eg, the inclusion of clinical concepts, such as Problems, Tests, and Treatments, into Events) [49]. The i2b2 2012 Shared Task corpus, finally, is accessible on the basis of signing a data use agreement and comes (for training) with 2366 annotation units for clinically relevant temporal expressions (TIMEX3 data), 15,567 annotation units for clinically relevant temporal events, and 33,543 annotation units for clinically relevant temporal relations; train and test data sum up to almost 95k annotation units. This landmark corpus exceeded previous annotation efforts in this field (most notably CLEF [47]) both in terms of the number of annotated documents and annotation units by a large margin and became a benchmark for future research.

Recently, a time-wise much narrower shared task devoted to medication change events documented in clinical notes was run within the context of n2c2, the follow-up event of i2b2 [54].

XSL•FO
RenderX

Participants had to identify medication change (Event) and determine the Action, Negation, Temporality, Certainty, and Actor for any change events (Context). The categories for temporality (Past, Present, Future, Unknown; about 1.7k annotation units) are considerably coarser in granularity than those from the previous i2b2 2012 Shared Task and are not grounded in TimeML. An annotation effort that extends TIMEX3 expressions for timeline construction based on calendar expressions is reported by Viani [56].

Complementary to these activities, within the TempEval branch of SemEval, a clinical TempEval subbranch was established. In the first 2 runs of this medically focused TempEval challenge [51-53], roughly 600 deidentified clinical notes and pathology reports from colon cancer patients at Mayo Clinic were manually annotated with an extension of TimeML for temporal expressions, medical events, and temporal relations between events and times (details of the annotation process and extensions of TimeML as part of the THYME corpus are described by Styler et al [50]). In the third run of TempEval [53], this dataset was complemented by a second one with a focus on brain cancer, as the organizers wanted to explore the potential for domain adaptation, that is, how do systems trained on colon cancer data perform when tested on brain cancer data?

The TimeML-annotated corpus from TempEval is available on a data use agreement basis as well, and comes (for colon cancer) with roughly 83k annotation units in its training (50%), and development (25%) splits, hence contains around 110k annotation units together with the held-out test set (25% of the entire dataset). For brain cancer, 48k annotation units are available for training (50%) and development (25%), and 15k in the test split, summing up to 63k, in total. Thus, the corpus from TempEval-3 exceeds the i2b2 2012 Shared Task corpus by slightly less than a factor of 2 in terms of annotation units and figures as the second milestone for temporal clinical corpora for the English language.

All of the above-mentioned corpus building activities almost only cover the English (clinical) language. Since our work deals with German language data, Table 3 gives an overview of the few efforts to deal with nonclinical German from a temporal perspective.

Table 3. German time-focused nonclinical resources (Corpora and Taggers).

Resource type	Name of resource (Year)	Functional characterization	Annotation language	Text genre	Size in tokens	Type/number of metadata
Corpus	EuroParl [57] (2008)	Time annotation mappings in parallel corpora: English → German	Subset of TimeML: TIMEX3, Events, no temporal relations	Transcribed speeches in the European Parliament	960k bisentences (train) and 0.3k bisentences (dev/ test)	Not reported
Corpus	WikiWarsDE [58] (2011)	N/A[a]	TIMEX2 [22], a subset of TimeML restricted to dates, time expressions, and durations)	22 German Wikipedia articles about famous wars in history corresponding to the English selection in WikiWars [59]	95.6k tokens	2.2k TIMEX2 annotation units
Corpus	KRAUTS [60] (2018)	N/A	TIMEX3	192 newspaper articles	75.7k tokens	1.1k TIMEX3 annotation units
Time tagger	HeidelTime [61] (2013)	Rule-based tagger using gazetteers and regular expressions	TIMEX2, TIMEX3 subset of TimeML plus normalizations into ISO[b] 8601 standard	22 German Wikipedia articles about famous wars in history corresponding to the English selection in WikiWars [59]	100.7k tokens	2.2k TIMEX2 annotation units
Time tagger	HeidelTimeext [62] (2022)	Extended rule-based tagger using gazetteers and regular expressions	TIMEX3	Mix of 10 parliament protocols, 10 books, 766 newspaper and 10 scholarly articles, 10k Wikipedia sentences	5.121k tokens	56.8k/53.8k TIMEX3 annotation units (automatically generated)

[a]N/A: not applicable.

[b]ISO: International Organization for Standardization.

Three German-language corpora have been built with a focus on time information. EuroParl [57] collects transcribed speeches (roughly 1m bisentences) from the European Parliament. TimeML-conformant temporal annotations are automatically mapped from English source texts to German targets. WikiWarsDE [58] assembles 22 Wikipedia articles on famous wars; its time representations are restricted to TIMEX expressions. So is the annotation in KRAUTS [60] based on 192 newspaper articles. For all 3 corpora, token counts do not exceed 100k, and at most, roughly 2k manual gold standard annotations are provided (HeidelTimeext [62] offers a substantially larger yet automatically generated silver standard). Overall, these are, in comparison with the numbers in Table 1, tiny corpora with low annotation density and limited representational expressiveness (with the exception of Spreyer and Frank's study [57]).

XSL•FO
RenderX

Two versions of the same time tagger also deserve mention. HeidelTime [61] was originally built on English data and then augmented for tagging German data. HeidelTime is a classical rule-based system using regular expressions and additional gazetteers. The step from HeidelTime to HeidelTimeext [62] comes with an augmented and updated rule set, an increase of the document size and token number (more than 5m tokens), and the number of temporal annotations (more than 50k items). However, in both cases, the representational expressiveness is limited to TIMEX expressions.

The state of German-language clinical corpora with temporal metadata is basically a short story summarized in Table 4. The MACCS corpus [63] is made of slightly more than 1.7k discharge summaries and clinical notes from the nephrology domain whose temporal annotations are based on a corpus-specific, that is, non–TimeML-conformant annotation language. Our work, instead, is fully TimeML-compliant and features 2 different corpora. $3000PA_J$ comes with 1100 real clinical documents, 1.7M tokens, and 120k temporal annotations, whereas GraSCCo is a much smaller corpus product featuring 63 synthetic documents, 44k tokens, and almost 12k temporal annotation units. Due to rigid data privacy legislation in Germany, $3000PA_J$ cannot be made publicly available, whereas GraSCCo can (it consists of fictitious clinical data). With roughly 131k annotation items, however, both corpora not only exceed all German-language corpora (whether clinical or nonclinical) in terms of temporal metadata but also all clinical English time-focused corpora, with one exception—only the million-scale nonclinical Maven-Ere corpus [40] offers a (much) larger number of temporal annotations.

How do $3000PA_J$ and GraSCCo compare with other non-English clinical corpora annotated with temporal metadata? Table 5 lists such textual resources whose size exceeds 50 documents and 1k annotation units. Perhaps, the closest in kinship is the French MERLOT corpus [64] that also adheres to TimeML specifications but falls behind by orders of magnitude both in terms of tokens and the amount of temporal metadata. Three other French corpora [65-67] are much smaller than MERLOT and also contain fewer temporal annotations; only Bannour et al [66] adheres to TimeML, including temporal relations. From 4 Chinese corpora, only the work of Liu et al [68] contains TimeML specifications for TIMEX3 expressions and temporal relations, yet on a much smaller scale than $3000PA_J$, both document, token, and metadata-wise. Pan et al [69] and Liu et al [70] focus on TIMEX3 expressions only (Liu et al [70] offering the largest corpus in terms of the number of documents, yet not of the number of annotation items), whereas Hu et al [71] use a self-supplied time ontology for annotation. Swedish [72], Portuguese [73], Spanish [17], and Italian [74] resources are tiny in terms of the number of documents and annotation items they contain (in the range of less than some hundred documents) and the number of temporal relations (less than 5k, if at all). Finally, Cheng et al [75] contain 2.6k THYME-conformant temporal relation annotations from a Japanese corpus based on slightly more than 1k documents. Another Japanese corpus, MedTxt-CR-JA [76], comes with 27.4k temporal relation annotations that are inferred from 2.6k explicitly annotated start-point annotations. TIMEX3-only annotations are provided by Hamon et al [65], Jeyafreeda et al [67], Pan et al [69], and Velupillai [72]. With the exception of the multilingual E3C corpus [77] (composed of French, Spanish, Basque, and Italian clinical case reports which differ significantly in style and format from real clinical reports), none of these corpora is publicly available.

Table 4. German time-focused clinical corpus resources with our own contribution highlighted.

Resource type	Name of resource (Year)	Annotation language	Text genre	Size in tokens	Type/number of metadata
Corpus	MACSS [63] (2016)	Corpus-specific, that is, non-TimeML-conformant temporal annotations: time points, dates, temporal courses	118 discharge summaries;1607 clinical notes from the nephrology domain (kidney transplantations)	89.7k tokens; 68.5k tokens	Not reported
Corpus	$3000PA_J$ (our work)[a]	TimeML:TIMEX3, Events, Six temporal relations (Before/After, Contains, Overlap, Begins_on, Ends_on)	1100 real clinical documents, mostly discharge summaries	1.7M tokens[a]	13.2k annotation units for TIMEX3 time expressions; 46.9k annotation units for Events; 59.1k annotation units for temporal relations; 119.2k[a] annotation units, in total
Corpus	GraSCCo (our work)[a]	TimeML:TIMEX3, Events, Six temporal relations (Before/After, Contains, Overlap, Begins_on, Ends_on)	63 synthetic clinical documents	44k tokens[a]	1.2k annotation units for TIMEX3 time expressions; 4.5k annotation units for Events; 5.8k annotation units for temporal relations; 11.6k[a] annotation units, in total

[a]Our own contribution.

XSL•FO
RenderX

Table 5. Non-English and non-German time-focused clinical corpus resources.

Natural language (ISO[a] 639 codes)	Name of resource (Year)	Annotation language	Text genre	Size in tokens	Type/number of metadata
FR	No name [65] (2014)	TIMEX3 expressions (plus normalizations)	182 clinical records (train) and 120 records (test, 25 with reference annotations), 302 records, in total	Not reported	Not reported
SV	Stockholm EPR [72] (2014)	TIMEX3 expressions	112 clinical notes from an ICU[b] unit	≈50k (estimate)	<2k TIMEX annotations
ZH	CMedTEX [70] (2016)	Simplified TimeML TIMEX3 expressions: Date, Time, Duration, and Frequency (plus ISO8601 normalizations)	1778 clinical notes	Not reported	46.6k TIMEX3 annotations
PT	No name [73] (2018)	TimeML TIMEX3: Date, Time, Duration, Set (plus ISO8601 normalizations)	130 discharge summaries	Not reported	Not reported
FR	MERLOT [64] (2018)	Simplified TimeML: Date, Time, Duration, and Frequency; 6 TimeML-compliant temporal relations: Before, Begins_on, Ends_on, During, Overlap, Simultaneous	500 discharge summaries, procedure reports (eg, radiology reports), physician letters, and prescriptions	148.5k	3.9k temporal entity annotations; 4.6k temporal relation annotations
ZH	No name [68] (2019)	Three TimeML-compatible temporal relations: Before, After, Simultaneous	563 clinical notes	Not reported	4.0k temporal expressions (ako TIMEX3 expression); 4.0k time intervals (relations)
IT	No name [74] (2019)	THYME-conformant clinical events (Problems, Tests, Treatments, Occurrences) and temporal relations (Before, Before/Overlap, Overlap, After)	75 clinical reports	57,263	4.4k temporal relation annotations
ES	No name [17] (2020)	TIMEX3 expressions (plus expression normalization) and TimeML-noncompliant temporal relations	50 clinical notes and 50 clinical reports for TIMEX3 identification & 100 clinical notes and 100 clinical reports for temporal relation identification, 300 documents, in total	Not reported	TIMEX3 annotations not reported; 200 temporal relations
ZH	TNorm [69] (2020)	TimeML-compliant TIMEX3 (plus expression normalization) only, no temporal relations	900 discharge summaries	Not reported	12.1k TIMEX3 annotations
ZH	No name [71] (2022)	Self-designed Clinical Time Ontology as a basis for an OWL-based temporal logic, with 11 temporal classes (various types of Instants, Intervals, Instant and Interval Collections, and Modifiers) and 16 temporal relations (extension of Allen's 13 relations set)	300 case reports containing 3000 clinical statements (presumably sentences)	Not reported	3.5k temporal entity annotations; 3.0k temporal relation annotations; 6.5k annotation units, in total
JA	No name [75] (2022)	THYME-compliant, simplified annotation schema: TIMEX3 + 5 temporal relations: On, Before, After, Start, Finish (complemented by 5 medical relations: Change, Compare, Feature, Region, Value)	1000 radiography interpretation reports of lung cancer + 156 medical history reports of idiopathic pulmonary fibrosis	Not reported	2.6k temporal relation annotations

XSL•FO
RenderX

Natural language (ISO[a] 639 codes)	Name of resource (Year)	Annotation language	Text genre	Size in tokens	Type/number of metadata
FR, ES, IT, EU	E3C [77] (2023)	THYME-conformant annotation: TIMEX3 expressions, Events, temporal relations	81 French and Spanish, 86 Italian, and 90 Basque clinical case reports + 615 clinical case reports with silver standard (automatic) annotations	French: 25,196; Spanish: 24,681; Italian: 24,319; Basque: 22,505 + >200k (silver standard)	French: 0.3k TIMEX3, 4.3k Event, 4.0k temporal relation annotations; Spanish: 0.4k TIMEX3, 4.8k Event, 4.7k temporal relation annotations; Italian: 0.3k TIMEX3, 3.4k Event, 1.2k temporal relation annotations; Basque: 0.6k TIMEX3, 7.9k Event, 8.0k temporal relation annotations
FR	No name [66] (2023)	Subset of TimeML Document; Creation Time and 5 temporal relations: Before, Before_Overlap, Overlap, and After + TemporalReference (noncompliant with TimeML)	220 (train) + 57 (test) clinical reports	Not reported	2.9k temporal relation annotations
JA	MedTxt-CR-JA [76] (2024)	Start-point ordering of events from which 3 temporal relations can be inferred: Before, Equal, and After	62 (from 148) case reports	Not reported	27.4k temporal relation annotations (based on 2.6k explicitly annotated start-point annotations)
FR	No name [67] (2024)	Corpus-specific, that is, non-TimeML-conformant temporal annotations: Date, Time, Frequency, Duration, and Age	150 (train) + 50 (test) clinical notes	Not reported	0.4k temporal entity annotations

[a]ISO: International Organization for Standardization.

[b]ICU: intensive care unit.

Goals of This Work

As stated above, temporal knowledge, its annotation, and automatic extraction in the medical domain have almost only been researched in depth in the Anglo-American language community. The non-English clinical language community is basically a low-/nonresource area that, with only a few exceptions, generally lacks temporal metadata. In this article, we propose a 2-fold solution, specifically responding to the strict privacy legislation constraints in the European part of the world.

The unique contributions of this article are the following: We provide temporal annotations for German clinical documents, including quality checks via interannotator agreement (IAA) metrics. More specifically, we supply temporal annotations for a nondistributable corpus composed of real clinical documents and a publicly distributable corpus composed of synthetic clinical documents. Furthermore, we train transformer-based baseline taggers to automatically extract temporal information encoded in German clinical reports and notes.

Methods

Overview

In this section, we will give a detailed overview of our 2 clinical corpora, the annotation schema for temporal named entities and relations, the annotation process and its evaluation, as well as the presentation of our baseline time tagging model for clinical documents. Figure 2 summarizes our experimental setting and illustrates the key steps in our workflow.

XSL•FO
RenderX

Figure 2. Overview of the experimental setup and workflow. BOHB: Bayesian Optimization and Hyperband; GBERT: German BERT model.

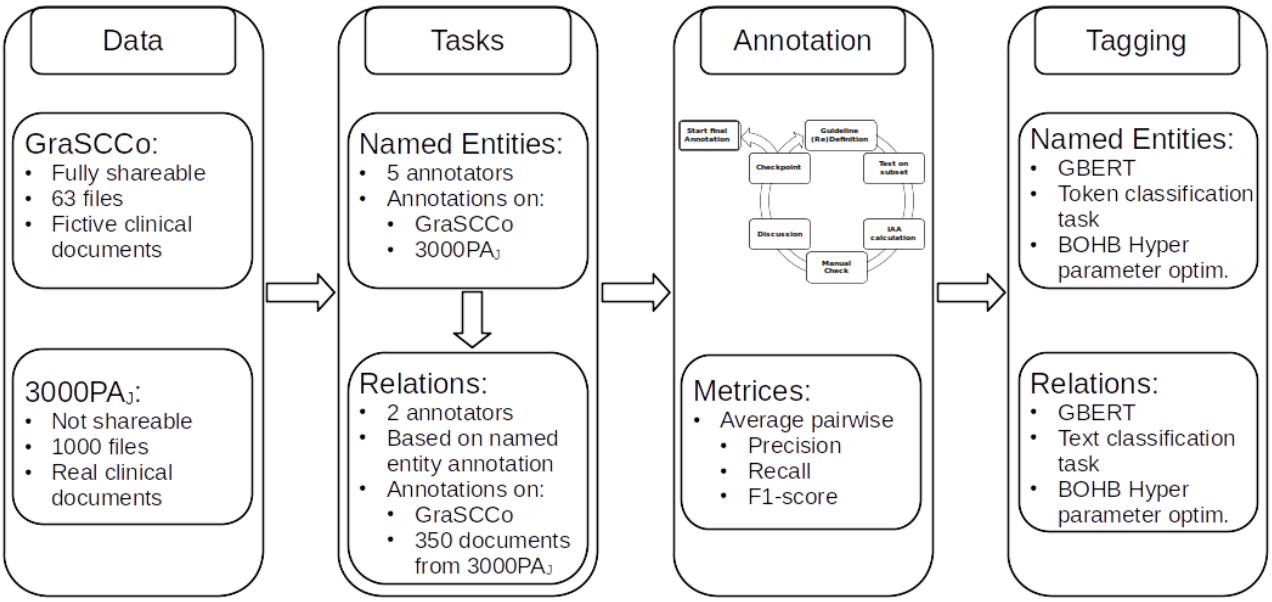

Ethical Considerations

The usage of 3000PA$_J$ is based on the approval by Jena University Hospital ethics committee (4639-12/15) and its data protection officer. The data was collected during the routine clinical care of adults at the intensive care unit or internal medicine ward. All persons were deceased at the time of data collection, thus no informed consent was needed for local data analysis according to the German law at that time. All data were hosted, annotated, and processed at the premises on secure servers with access restrictions and regular data protection training of the annotators. For GraSCCo, no ethics approval is needed as this dataset contains only fictional data written in the style of real clinical documents.

Clinical Data Resources

For temporal annotation, we used the 2 corpora already introduced in Table 4, namely 3000PA$_J$ and GraSCCo. 3000PA$_J$, a subset of the 3000PA corpus [78], consists of about 1100 real, that is, authentic, clinical documents, mostly discharge summaries, with roughly 145k sentences and 1.7m tokens hosted by the Jena University Hospital under strict access restrictions. In effect, this means that according to German data protection legislation, this dataset cannot be made available for public use outside the walls of the Jena University Hospital. 3000PA$_J$ has already been annotated at multiple semantic layers, including section headings [79], Personally Identifiable Information [80], medications [78], and major nontemporal clinical named entities, namely diagnoses, findings, and symptoms [81].

As an alternative to siloing much-wanted clinical data in the barred 3000PA$_J$ dataset, we developed GraSCCo [82], a synthetic twin corpus which is open for use to the entire clinical NLP community without any access restrictions. Albeit being a tiny document collection with only 63 German multiply-alienated and noise-infused real documents with about 5.4k sentences and 44k tokens, it is available via the Creative Commons license on Zenodo [83]. As a noteworthy remark, the authors have already gathered preliminary evidence that GraSCCo may act as a reasonable substitute for 3000PA in terms of clinical language use and genre patterns [82]. This observation is fully consistent with other experimental results for various languages, which show that synthetic corpora can indeed replace real clinical corpora with only marginal quality loss regardless of the application tasks and clinical text genres chosen (see, eg, the evidence found for English [84-86], French [87], and Norwegian [88]).

Temporal Annotation

Overview

Manually generating temporal annotations is well-known as a notoriously complex and time-consuming task [49,50]. Hence, to avoid cognitive overload of the human annotators, we split our annotation campaign into 2 separate phases—an entity annotation and a relation annotation phase. Both followed the same schema (Figure 3 depicts one phase equals a complete cycle) and were performed by at least 3 different annotators. This organizational scheme follows the widely acknowledged annotation conventions established by James Pustejovsky and Amber Stubbs [89]. Generally speaking, at least 3 independent annotators are necessary to run a proper statistical analysis. All of our annotators were qualified medical students who had already passed their first medical exam after 2 years of study (Physikum, in German).

Each annotation cycle was divided into a training-only and a production annotation phase. During the training phase, all annotators received the same partition of the entire dataset for annotation, usually about ten documents. Once this batch was completed, their annotations were checked both automatically

XSL•FO
RenderX

via an IAA metric based on the average of the pairwise F_1-scores, and manually to identify and resolve individual/systematic errors and misunderstandings. The results of the latter phase were then discussed within the annotation team and led to changes in the annotation guidelines, if needed (typical shortcomings arising from premature annotation guidelines are a lack of clarity or ambiguity of instructions, instruction gaps, etc). After that, a new subset was drawn and, again, annotated by all annotators. These training iterations continued until the F_1-score, primarily the pairwise F_1-score as well as the averaged F_1-score, stabilized and reached a reasonable, that is, community-wide accepted, plateau (in the range of 0.9 on a scale of 0 to 1.0 agreement) [90]. After this preparatory phase, the final guidelines were frozen and applied to the entire dataset, again, with a randomly chosen partition as the agreement portion, usually 5% to 10% of all documents. These documents will be called agreement documents, as they are the documents on which the IAA is calculated. The final F_1-score was determined on these agreement documents and reported as the final agreement value.

Once the entity annotation task was completed, we unified all annotations of all agreement documents. Thus, label disagreements and missing annotations were detected. These merged entity annotations with all their identified disagreements were then used for the second phase, the relation annotation, and manually corrected by the human annotators. The annotators were also encouraged to modify/correct entity annotations if they disagreed with the results of the first entity-focused round.

Figure 3. Schema of the iterative workflow for annotation guideline development. The optimization cycle is repeated until a stopping criterion is reached on a document subset, mostly a predefined IAA threshold or the number of iterations. The final guideline is subsequently used to annotate the entire corpus in the production phase. IAA: interannotator agreement.

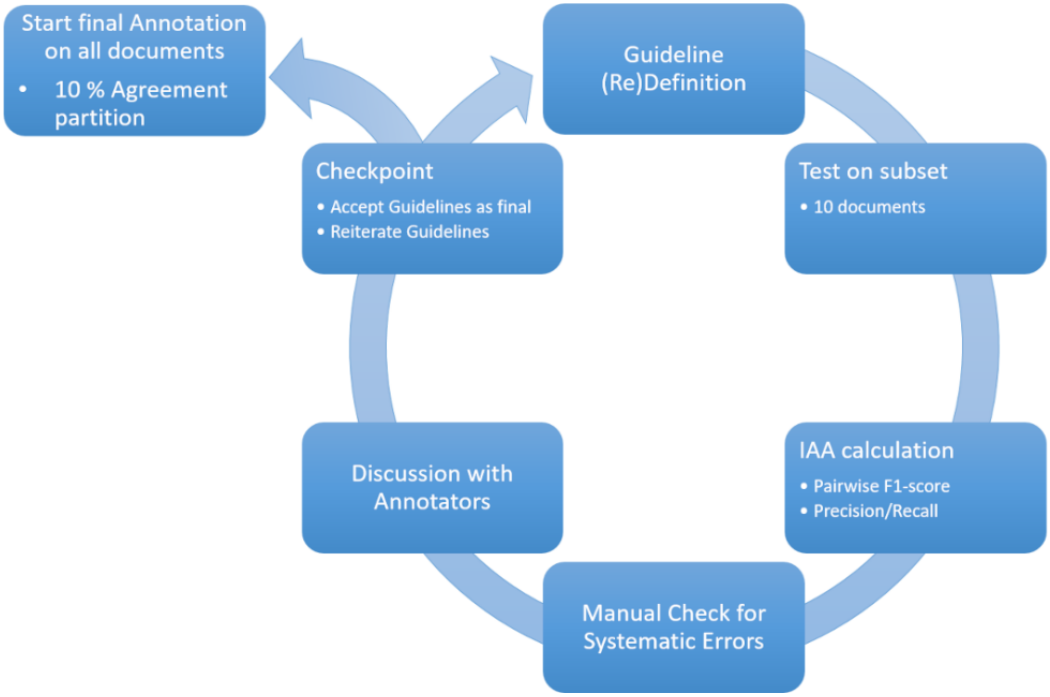

Temporal Entity Definition

Our definitions of time-related entities are identical to THYME [50]. Accordingly, those entities are defined as follows:

- EVENT: Any token or phrase that is related to things that occurred or happened.
- TIMEX3: Any token or short phrase that denotes a time expression of the following subtypes:
  - Date: Expressions or phrases that describe a time with accuracy up to a day (06/2020, next Wednesday),
  - Time: A specific time of the day (12 am, midnight),
  - Duration: A time frame (until yesterday, for 10 minutes),
  - Set: A recurring event (twice a day, 1-0-0),
  - Quantifier: The number of times something has happened or occurred, needs to be paired with an EVENT (second cycle of chemotherapy),
  - PrePost: Expressions like postoperative or premenarche.

Additionally, entities annotated as an EVENT needed to be marked with one of the following tags according to their point in time with respect to the reported hospital stay: BEFORE (if something happened prior to admission), BEFORE-OVERLAP (if something occurred before but is still present at the time the patient was admitted to the hospital), OVERLAP (if something happened during the hospital stay) and AFTER (if occurrences will happen in the future after the discharge). An example of said annotated entities can be found in Figure 4.

XSL•FO
RenderX

Figure 4. Sentences to illustrate the annotation of entities annotated with the BRAT annotation tool. Yellow indicates EVENTS, blue stands for TIMEX3 temporal expressions. The labels are written over the marked words or phrases.

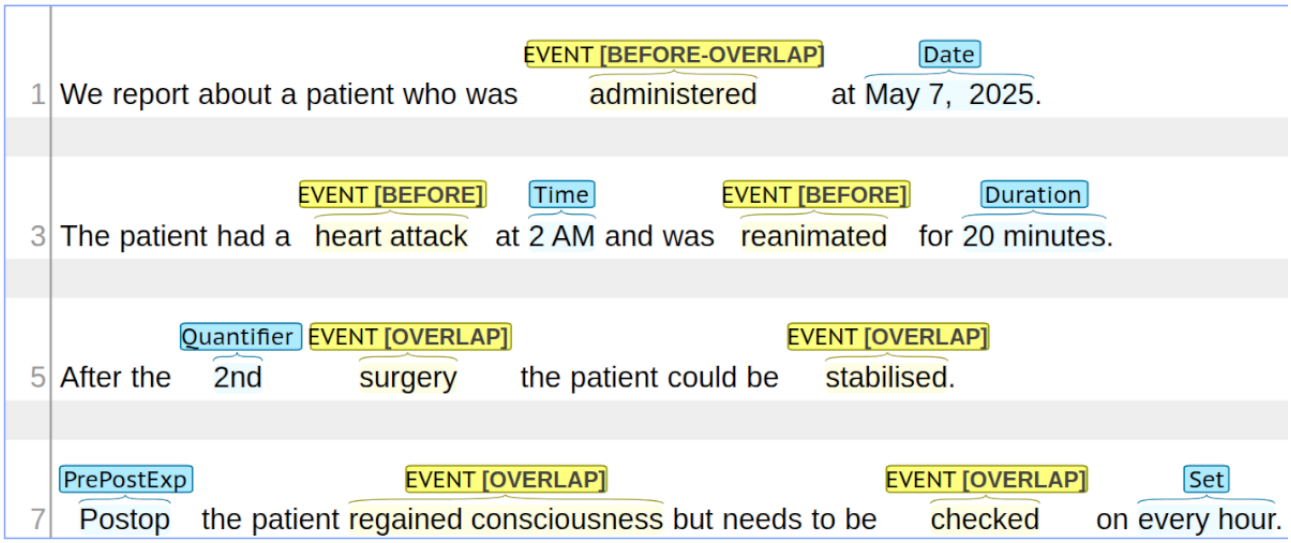

Temporal Relation Definition

Temporal entities themselves do not contain sufficient information for the adequate construction of timelines. Without context or ordering information, most of the time-relevant information is missing. Accordingly, temporal relations between those entities have to be determined. All of the relation types we chose are a subset of Allen's interval algebra [21]. We followed the definitions in THYME [50] and used the following relation types:

- BEFORE (X,Y) X happened entirely before Y
- (AFTER (Y,X) is treated as the inverse of BEFORE (X,Y))
- CONTAINS (X,Y) Y happened entirely during X
- OVERLAP (X,Y) X and Y share a common partition of time
- BEGINS-ON (X,Y) X started with Y
- ENDS-ON (X,Y) X finished with Y

In contrast to THYME and more in line with the i2b2 2012 guidelines, we have instructed our annotators to link each entity to at least one other entity in such a way that each entity is connected to every other entity via a path. This way, we ended up with a connected temporal graph where the entities are the vertices and the relations are represented as edges. This condition is crucial, however, to create a timeline of the entire hospital stay in later processing stages. In contrast to i2b2 2012, we did not use direct connections with the document time or hospital stay (time). This was already covered in the named entity phase by encoding the respective Event tag. Figure 5 demonstrates the application of our annotation conventions to the sentences with already annotated named entities from Figure 4.

XSL•FO
RenderX

Figure 5. Extension of the already annotated named entities from Figure 4 with temporal relations using the BRAT annotation tool. The arrows indicate the relation together with its relation type.

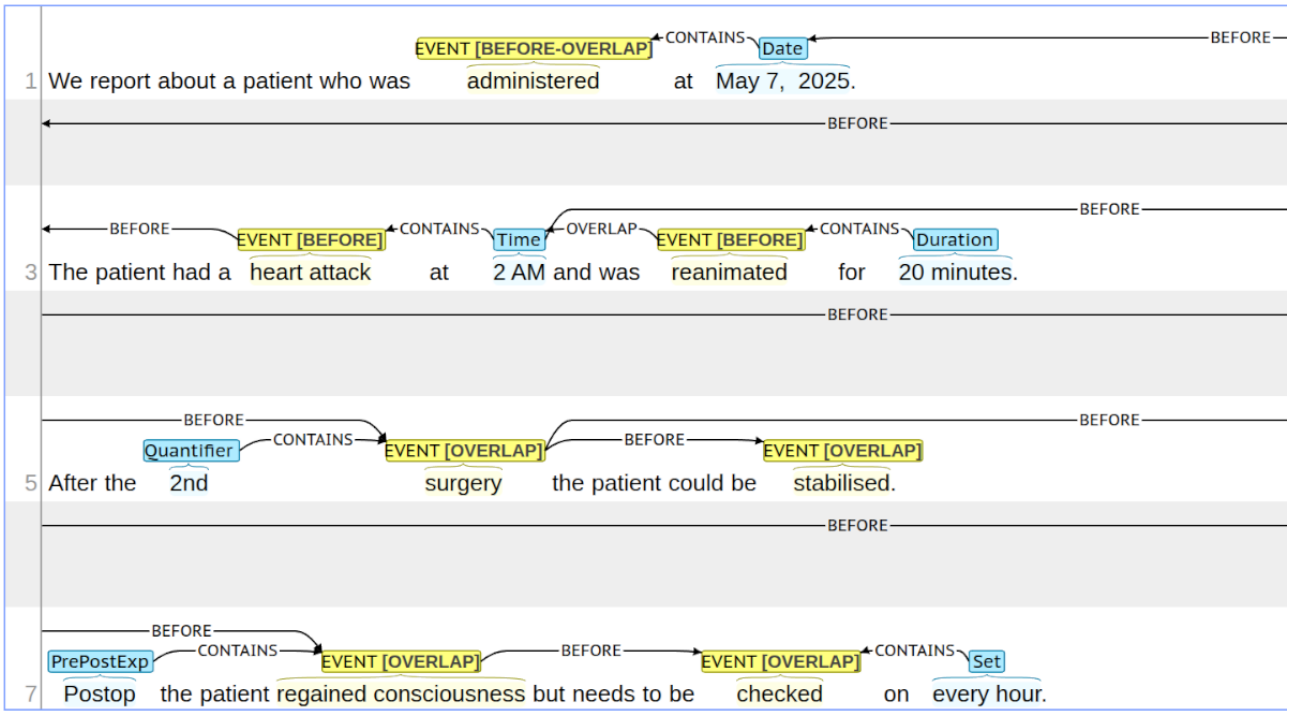

Evaluation Methods

To determine the IAA scores in our temporal named entity annotation campaign, we calculated the mean of the pairwise precision, recall, and F_1-score for the annotated entities. The implementation we used can be found on GitHub [91].

In contrast to temporal named entities, temporal relations need a different treatment for their evaluation. Every entity has temporal connections to every other entity. However, annotating the entire fully connected temporal graph is not feasible, since the number of temporal relations grows quadratically with the number of event mentions (by anchoring events in time, however, annotation of temporal relations indeed scales linearly with the number of events [92]). Thus, each annotator was encouraged to annotate the least possible number of relational edges, resulting in something akin to a minimum spanning tree. To guide the annotators towards this goal, they were instructed not only to connect named entities but to think in terms of units of meaning. These units consist of smaller groups of events, such as the timeline of a surgery, the family history, or the events that happened before admission. These smaller units should then be linked to each other, creating a larger group. This should be repeated until all larger units of meaning are linked into one large, comprehensive timeline. If possible, the anchors for these temporal units should be TIMEX3 entities, preferably Dates. A fictitious example for these linkage conventions can be found in Multimedia Appendix 1.

These conventions result in a number of different graphs that, in the worst case, might share no edges at all. This would lead to a much worse agreement score than justified. The solution we propose here is to use inference and temporal closure [93]. We infer the missing edges for each annotator and are thus able to calculate a pairwise agreement score (namely precision, recall, and F_1-score; see below). Based on the overlap of temporal relation definitions, we used a slightly adjusted version of the THYME evaluation script. The changes included adjustments to our annotation format (BRAT) [94], extending the evaluation, and the visualization of annotations.

Training and Evaluation of a Baseline Temporal Tagger

To demonstrate the usefulness of the outcome of our annotation campaign and to determine the first baseline for the automatic temporal tagging of German clinical documents, we used state-of-the-art transformer-based classification technology, that is, BERT (Bidirectional Encoder Representations from Transformers)-ish language models. For German, the most recent and best-performing language model is the German BERT model (GBERT) [95]. For our experiments, we used the base model for each of the downstream tasks we specified. To automatically optimize hyperparameters for each task, we used Bayesian Optimization and Hyperband [96] (Multimedia Appendix 2). Our experiments incorporate the huggingface library [97], as well as the Bayesian Optimization and Hyperband implementation of the Microsoft NNI framework [98]. This configuration was applied to and evaluated on the nonsharable 3000PA (ie, $3000PA_J$) and the publicly sharable GraSCCo corpus. The ground truth data used for training, testing, and evaluation were taken from the annotated documents without further correction. If the document was part of the agreement partition, and thus had been annotated by all annotators independently, we randomly selected one annotation variant from one annotator.

Evaluation Metrics

For the statistical evaluation, we apply the following metrics.

Precision (P), also known as positive predictive value, is the ratio between the gold-standard matching annotations of all

XSL•FO
RenderX

named entities or temporal relations for each label (true positives) and all annotated named entities or temporal relations for each label (true positives + false positives).

Recall (R), also called sensitivity, is the ratio between the gold-standard matching annotations of all named entities or temporal relations for each label (true positives) and all annotated named entities or temporal relations for each label (true positives and false negatives).

F_1-score is defined as the harmonic mean of precision and recall.

Accuracy (Acc) is defined as the ratio between the gold-standard matching annotations (true positives and true negatives) and all annotations, regardless of their annotations being true or false (sum of annotated entities).

All of these measurements were calculated pairwise between all possible annotator pairs. The final agreement score for the labels or the entire corpus reports the averaged pairwise agreement scores.

Results

Overview

We present the results of our experiments, distinguishing between manual temporal named entity and temporal relation annotations, as well as the evaluation of the GBERT baseline tagging model that automatically generates temporal tags for clinical documents.

Annotation of Temporal Named Entities

In the first cycle of our 2 annotation campaigns, we annotated temporal named entities on both $3000PA_J$ and GraSCCo. The results presented here were calculated on the corrected named entity annotations after the relation annotation campaign with its named entity correction subtask.

As can be seen in Table 6, we achieved pairwise F_1-scores above 0.8 for nearly all named entity types. Moreover, our overall agreement even exceeds an average F_1-score above 0.9. When looking at the token-level evaluation (which is less strict), we achieved an overall agreement average F_1-score for $3000PA_J$ of 0.912 and 0.910 for GraSCCo.

On both datasets, the entity type Duration yielded slightly worse scores. There is only one major outlier in terms of IAA in the $3000PA_J$ results, namely Quantifier, both in terms of F_1-score and SD. Interestingly, this is not observed in GraSCCo. When looking at the confusion matrix of the named entities of both datasets in Figure 6, we are able to explain this deviant behavior.

As can be seen in the confusion matrices for entity types of both corpora in Figure 6, the main reason for mismatching labels was mostly missing annotations, labeled “None” in the confusion matrix. The Event type, in particular, was often only annotated by one annotator. This behavior can be observed for both corpora. Interestingly, in $3000PA_J$, the type Quantifier was also often missed by one of the annotators. The most often confused label pair was Date and Duration. Some translated examples for confused label pairs are:

- {since 2017}DURATION vs since {2017}DATE
- {the entire night}DURATION vs the entire {night}DATE
- {after about 30 minutes}DURATION vs after about {30 minutes}TIME
- {day of self-harm}DATE vs day of {self-harm}EVENT

Some systematic errors also appear in the distinction between Duration and Date. Most of the mismatches were grounded in some form of trigger word, such as “since” or “after”. However, it needs to be stressed that these kinds of errors did not occur too often relative to all cases, as can be seen in the confusion matrices in Figure 6. Mismatches between labels other than the ones described above are very rare (per milli range).

XSL•FO
RenderX

Table 6. Absolute number and relative number in percentage of named entities, mean pairwise F1-score with SD for each entity type of the temporal named entity annotation campaign for 3000PAJ and GraSCCo, respectively, as well as the overall agreement for each dataset, which also includes the nonannotated tokens.

Named entity	$3000PA_J$				GraSCCo			
	Absolute number	Percentage (%)	F_1-score	SD	Absolute number	Percentage (%)	F_1-score	SD
Date	9099	15.13	0.903	0.055	663	11.51	0.947	0.020
DocTime	886	1.47	0.988	0.007	92	1.60	0.944	0.031
Time	593	0.99	0.893	0.058	3	0.75	0.852	0.075
Duration	824	1.38	0.776	0.083	138	2.40	0.747	0.091
Set	1107	1.84	0.822	0.081	158	2.74	0.802	0.107
Quantifier	563	0.94	0.550	0.209	118	2.05	0.938	0.023
PrePost	146	0.24	0.831	0.041	24	0.42	0.932	0.041
EVENT	46,909	78.02	0.910	0.017	4523	78.54	0.906	0.053
Total	60,127	100.0	0.834	0.068	5759	100.0	0.883	0.055
Overall Agreement	__[a]	—	0.904	0.024	—	—	0.906	0.047

[a]Not applicable.

Figure 6. Confusion matrices for temporal named entity types after correction, with 3000PAJ (left) and GraSCCo (right). The color and intensity indicate the relative number of entity pairs (with 1.0 being all and 0.0 no existing pairs). Agreeing pairs (both annotators chose the same label) can be found on the main diagonal.

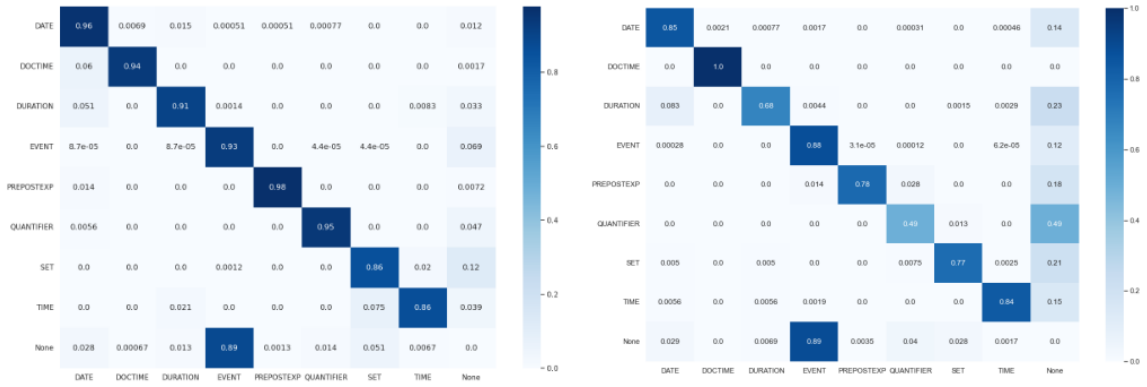

Annotation of Temporal Relations

Because of their complexity and the human efforts required for this task, temporal relations were annotated only on a randomly chosen subset of 350 documents of the original $3000PA_J$ dataset, whereas all 63 documents of the GraSCCo dataset were dealt with. Both datasets were annotated by the same 4 annotators. The resulting IAA scores incorporating the aforementioned inference rules are depicted in Table 7. Because of the nature of points in time and their relation to each other (they form a fully connected temporal graph), relations between 2 named entity pairs always have a relation type, even though it might be automatically inferred rather than manually annotated.

As can be seen in the total count of relation types, in GraSCCo all entities seem to be connected with each other, as the number of relation types is slightly larger (40 relations more than necessary) than the total number of entity types. But in $3000PA_J$ we seem to miss a total of about 1k relations, which means approximately 3 relations per document. However, given the total number of about 60k annotated relations, this is just 1.76% of the anticipated number of relations.

The confusion matrices for the 5 temporal relation types of both datasets are depicted in Figure 7.

The most prominent observation in both confusion matrices is the dominance of the CONTAINS relation type. Especially, OVERLAP was often confused with CONTAINS in both datasets. This is not so surprising given that CONTAINS is a subset of OVERLAP. However, it seems that the stricter CONTAINS relation is used more often compared with the less restrictive OVERLAP relation. We assume that the reason for this phenomenon might be the fact that we hired medical students rather than linguists or computer scientists (an issue often observed in other annotation campaigns, as well). Thus, our annotators have attributed the relation types a more literal meaning, “something contains something”, even partially, and

XSL•FO
RenderX

did not share our set theory-based understanding of CONTAINS and OVERLAP. Unfortunately, in the 3000PA$_J$ corpus, all of the other relation types are often confused with the CONTAINS relation type. Discussions with the annotators revealed that they used the CONTAINS relation type as a residual category, which resulted in generally coding CONTAINS if they were unsure about other alternatives. However, mismatches between other labels occur very rarely (per mille range).

Table 7. Absolute numbers and relative numbers in percentage of temporal relations, mean pairwise F1-score with SD for each annotated type of the temporal relation annotation campaign on 3000PAJ and GraSCCo using temporal closure.

Relation type	3000PA$_J$				GraSCCo			
	Absolute number	Percentage (%)	F_1-score	SD	Absolute number	Percentage (%)	F_1-score	SD
BEFORE	5948	10.07	0.2	0.07	547	9.43	0.45	0.28
BEGINS-ON	1030	1.74	0.48	0.07	146	2.52	0.64	0.19
ENDS-ON	724	1.23	0.69	0.06	74	1.28	0.73	0.14
OVERLAP	5046	8.54	0.09	0.13	679	11.71	0.33	0.36
CONTAINS	46,318	78.42	0.63	0.07	4352	75.06	0.68	0.16
Overall agreement	59,066	100.0	0.41	0.08	5798	100.0	0.57	0.27

Figure 7. Confusion matrices for temporal relation types, with 3000PAJ (left) and GraSCCo (right). The color and intensity indicate the relative number of relation label pairs (with 1.0 being all and 0.0 no existing pairs). Agreeing pairs can be found on the main diagonal.

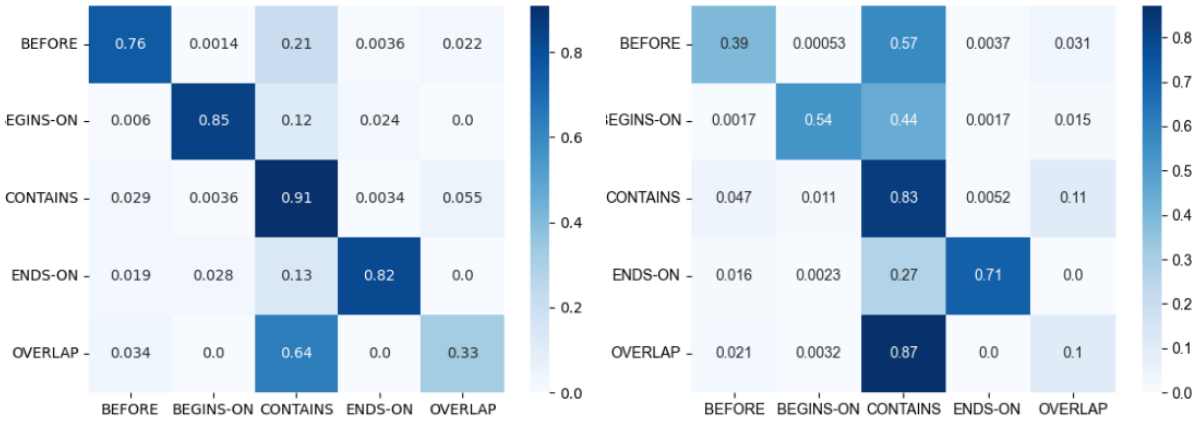

Baseline Model for Automatic Temporal Tagging

For the evaluation of the performance of the automatic temporal taggers trained on our annotated datasets, we fine-tuned and optimized a GBERT on both 3000PA$_J$ and GraSCCo. A summary of the results of our experiments can be found in Table 8.

All in all, GBERT performs well for both the named entity recognition (NER) and the relation extraction task, respectively, on the 3000PA$_J$ dataset. Training GBERT on 3000PA$_J$ and testing on GraSCCo results in an expected, yet acceptable, drop in performance. Interestingly, this drop mostly affects the recall score. The union of both datasets results in a similar performance rate compared with the model trained on 3000PA$_J$ only, yet with a slight decrease in recall and a slight increase in precision, which results in a similar F_1-score for both the named entity recognition and the relation extraction task.

XSL•FO
RenderX

Table 8. Mean pairwise accuracy (Acc), F1-score, precision, and recall for automatic temporal named entity recognition (NER) and relation extraction (REX) on 3000PAJ and GraSCCo, respectively, and a unified set of both (union).

Model	Train set	Test set	Acc	F_1-score	Precision	Recall
NER						
GBERT[a]	$3000PA_J$	$3000PA_J$	0.95	0.78	0.73	0.84
GBERT	GraSCCo	GraSCCo	0.94	0.85	0.84	0.87
GBERT	$3000PA_J$	GraSCCo	0.90	0.64	0.82	0.56
GBERT	union	union	0.95	0.77	0.75	0.80
REX						
GBERT	$3000PA_J$	$3000PA_J$	0.78	0.64	0.63	0.63
GBERT	GraSCCo	GraSCCo	0.81	0.60	0.63	0.58
GBERT	$3000PA_J$	GraSCCo	0.90	0.64	0.82	0.56
GBERT	union	union	0.81	0.63	0.70	0.59

[a]GBERT: German BERT model.

Discussion

Principal Findings

Temporal information is crucial for comprehensive medical reasoning; yet, only a few data resources are available for NLP tools to automatically extract such information from clinical documents. Publicly accessible clinical document collections have primarily emerged from challenge tasks, such as i2b2, n2c2, or Clinical TempEval, but cover the English clinical language only (Table 2). For a few non-English clinical language communities—French, Chinese (Mandarin), Spanish, Italian, Portuguese, German, Swedish, and Japanese—clinical corpora have been annotated with temporal metadata, but (except for E3C, a multilingual case report corpus) none of them is publicly available due to prohibitive data privacy regulations (Tables 4 and 8). We here introduced 2 corpora for the German clinical language annotated with TimeML-compliant metadata—a nondistributable real-world clinical corpus ($3000PA_J$), and a distributable synthetic one (GraSCCo). The first stands out with unprecedented quantitative scales in terms of the number of documents, tokens, and temporal metadata, and its adherence to qualitative standards, that is, TimeML specifications, lacking for many non-English alternatives (Table 8). The latter, though small-sized and synthetic, is one of the rare (if not the only one) non-English clinical text resources marked up for temporal information that are accessible without distribution restrictions. With reasonable agreement scores for manual annotation, we trained and tested temporal BERT-based taggers under varying experimental conditions and achieved decent performance values using both, 3000PAJ and GraSCCo.

Principal Results

In this article, we introduced 2 corpora equipped with temporal metadata for the German clinical language. The first of these corpora, $3000PA_J$, contains real clinical documents, but cannot be distributed in the public domain due to the restrictive privacy legislation codified in Germany and (more generally) Europe. To bypass this severe accessibility hurdle, we enriched a second corpus, GraSCCo, with temporal metadata, which can be distributed without any restrictions since it contains synthetic data, that is, fictitious and noise-adding content, only. The metadata we created for temporal named entities and relations, all of them fully compliant with TimeML, have a high level of credibility, given IAA scores ranging in the high 80s and low 90s.

As can also be seen from our evaluation results, we are on par performance-wise with comparable English corpora, such as THYME. For named entities and the CONTAINS relation, we even achieved better results on both corpora, $3000PA_J$ and GraSCCo. In addition, all documents used in the agreement evaluation were annotated by 4 individual annotators per document and yielded acceptable IAA scores.

In the end, we come up with about 66k temporal entities and 65k temporal relations, which sum up to 131k temporal annotation units for both corpora. These numbers not only exceed all clinical and nonclinical German-language corpora in terms of temporal metadata created up until now, but also outdo all clinical English time-focused corpora. Only the million-scale nonclinical Maven-Ere corpus [40] offers a (much) larger number of temporal annotations.

Limitations

We still have to increase the size of GraSCCo (this is ongoing work in our research group). Also, the baseline tagger needs to be complemented by more sophisticated and larger language models. We plan to experiment with temporal (knowledge graph) embeddings, which encode temporal information directly in time-enhanced embeddings [99-101] and document-level event extraction models [102-104], which will better accommodate long-range cross-sentence and cross-paragraph dependencies typical of temporal relations in long documents, such as clinical reports.

As discussed in the results section, the decision to use only medical students resulted in misconceptions regarding the definitions of the temporal relations. Potentially, mixed teams with medical students and annotators with a background in linguistics or computer science might have resulted in better

XSL•FO
RenderX

results (as small preliminary experiments for i2b2 2012 and CLEF suggest [49,105]). However, as shown by Raghavan et al [106], the level of clinical expertise has only a small impact on the inter-annotator reliability for temporal relations. Furthermore, medical students are more likely to recognize relations between clinical events [105], as well as relations based on medical causal dependence [105]. Additionally, we used the involuntary break and access restrictions during the COVID pandemic for extensive training iterations for our annotator team. Thus, if we compare our work with THYME, which was annotated by mixed teams as well, they reported similar difficulties and F_1-scores to those discussed above.

Comparison With Prior Work

To the best of our knowledge, $3000PA_J$ and GraSCCo are the first German clinical corpora to be annotated with temporal named entities and relations fully compliant with the TimeML de facto markup standard. Surprisingly, for nonclinical documents, no comprehensive TimeML-compliant corpus currently exists for German. Furthermore, providing unconstrained access to the synthetic GraSCCo corpus constitutes a major step in metadata availability for the German clinical NLP community. Moreover, the construction principles underlying GraSCCo [82] outline a way to extend the generation (not only) of temporal metadata beyond the German language. For the first time ever, a TimeML-compliant baseline temporal tagger for German has been trained and properly validated.

Conclusions

Our work can and will be the basis for algorithms and models to automatically detect and combine temporal entities and their relations in order to create clinical timelines [107,108]. These timelines will be visualized and are thus able to provide a comprehensive summary of the patients' clinical history, a potentially valuable source of abstraction for all kinds of clinical decision makers and caretakers.

Acknowledgments

We thank Prof Dr André Scherag, Dr Danny Ammon, and all members of the Data Integration Center of the Jena University Hospital for their continuous support in managing access to $3000PA_J$.

No artificial intelligence assistance has been used in the writing process of this paper; all text, figures, and tables have been created manually by the authors.

Funding

This work was supported by the German Ministry of Education and Research (BMBF) within the SMITH and DIFUTURE clusters of the Medical Informatics Initiative (MI-I) under the grants 01ZZ1803G and 01ZZ2009A, respectively.

Authors' Contributions

LM contributed to the conceptualization of the study, data curation, formal analysis, conduct of experiments, methodology design, provision of resources, software development, validation, visualization, and writing of both the original draft and the review and editing of the manuscript. UH was responsible for funding acquisition, provision of resources, project administration, and supervision, and also contributed to writing the original draft and reviewing and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example for relation annotation schema.
[PDF File (Adobe PDF File), 464 KB - jmir_v28i1e71458_app1.pdf]

Multimedia Appendix 2
Hyperparameter for transformer training.
[DOCX File , 9 KB - jmir_v28i1e71458_app2.docx]

References

1. Wang Y, Wang L, Rastegar-Mojarad MA, Moon SR, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018;77:34-49 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
2. Bose P, Srinivasan S, Sleeman WC, Palta J, Kapoor R, Ghosh P. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. Appl Sci 2021;11(18):8319. [doi: 10.3390/app11188319]
3. Nasar Z, Jaffry SW, Malik MK. Named entity recognition and relation extraction. ACM Comput Surv 2021;54(1):1-39. [doi: 10.1145/3445965]

XSL•FO
RenderX

4. Madkour M, Benhaddou D, Tao C. Temporal data representation, normalization, extraction, and reasoning: a review from clinical domain. Comput Methods Programs Biomed 2016;128:52-68 [FREE Full text] [doi: 10.1016/j.cmpb.2016.02.007] [Medline: 27040831]
5. Leeuwenberg A, Moens MF. A survey on temporal reasoning for temporal information extraction from text. J Artif Intell Res 2019;66:341-380. [doi: 10.1613/jair.1.11727]
6. Alfattni G, Peek N, Nenadic G. Extraction of temporal relations from clinical free text: a systematic review of current approaches. J Biomed Inform 2020;108:103488 [FREE Full text] [doi: 10.1016/j.jbi.2020.103488] [Medline: 32673788]
7. Gumiel YB, Silva e Oliveira LE, Claveau V, Grabar N, Paraiso EC, Moro C, et al. Temporal relation extraction in clinical texts. ACM Comput Surv 2021;54(7):1-36. [doi: 10.1145/3462475]
8. Hahn U. Clinical document corpora-real ones, translated and synthetic substitutes, and assorted domain proxies: a survey of diversity in corpus design, with focus on German text data. JAMIA Open 2025;8(3):ooaf024. [doi: 10.1093/jamiaopen/ooaf024] [Medline: 40371384]
9. Sun W, Rumshisky A, Uzuner Ö. Temporal reasoning over clinical text: the state of the art. J Am Med Inform Assoc 2013;20(5):814-819 [FREE Full text] [doi: 10.1136/amiajnl-2013-001760] [Medline: 23676245]
10. Kamsu-Foguem B, Tchuenté-Foguem G, Foguem C. Verifying a medical protocol with temporal graphs: the case of a nosocomial disease. J Crit Care 2014;29(4):690.e1-690.e9. [doi: 10.1016/j.jcrc.2014.02.006] [Medline: 24636931]
11. Wu STI, Juhn YJ, Sohn S, Liu H. Patient-level temporal aggregation for text-based asthma status ascertainment. J Am Med Inform Assoc 2014;21(5):876-884 [FREE Full text] [doi: 10.1136/amiajnl-2013-002463] [Medline: 24833775]
12. Caballero Barajas KL, Akella R. Dynamically modeling patient's health state from electronic medical records: a time series approach. : Association for Computing Machinery (ACM); 2015 Presented at: KDD - 15 - Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2015 Aug 10; Sydney, New South Wales, Australia p. 69-78. [doi: 10.1145/2783258.2783289]
13. Jo Y, Loghmanpour N, Penstein-Rosé C. Time series analysis of nursing notes for mortality prediction via a state transition topic model. : Association for Computing Machinery (ACM); 2015 Presented at: CIKM - 15 - Proceedings of the 24th ACM International Conference on Information and Knowledge Management; 2015 Oct 19; Melbourne, Victoria, Australia p. 1171-1180. [doi: 10.1145/2806416.2806541]
14. Qiao Z, Zhao S, Xiao C, Li X, Qin Y, Wang F. Pairwise-ranking based collaborative recurrent neural networks for clinical event prediction. 2018 Presented at: IJCAI-ECAI '18 - Proceedings of the 27th International Joint Conference on Artificial Intelligence & 23rd European Conference on Artificial Intelligence; 2018 July 13; Stockholm, Sweden p. 3520-3526. [doi: 10.24963/ijcai.2018/489]
15. Khadanga S, Aggarwal K, Joty S, Srivastava J. Using clinical notes with time series data for ICU management. : Association for Computational Linguistics (ACL); 2019 Presented at: EMNLP-IJCNLP - Proceedings of the Conference on Empirical Methods in Natural Language Processing & 9th International Joint Conference on Natural Language Processing; November ; Hong Kong, China. Association for Computational Linguistics (ACL); 2019 Nov 03; Hong Kong, China p. 6432-6437 URL: https://aclanthology.org/D19-1678/ [doi: 10.18653/v1/d19-1678]
16. Najafabadipour M, Zanin M, Rodríguez-González A, Torrente M, Nuñez García B, Cruz Bermudez JL, et al. Reconstructing the patient's natural history from electronic health records. Artif Intell Med 2020;105:101860. [doi: 10.1016/j.artmed.2020.101860] [Medline: 32505419]
17. Deznabi I, Iyyer M, Fiterau M. Predicting in-hospital mortality by combining clinical notes with time-series data. : Association for Computational Linguistics (ACL); 2021 Presented at: Findings of the Association for Computational Linguistics - ACL-IJCNLP 2021; 2021 Aug 01; USA p. 4026-4031 URL: https://aclanthology.org/2021.findings-acl.352/ [doi: 10.18653/v1/2021.findings-acl.352]
18. Olex AL, McInnes BT. Review of temporal reasoning in the clinical domain for timeline extraction: where we are and where we need to be. J Biomed Inform 2021;118:103784 [FREE Full text] [doi: 10.1016/j.jbi.2021.103784] [Medline: 33862232]
19. Xie F, Yuan H, Ning Y, Ong MEH, Feng M, Hsu W, et al. Deep learning for temporal data representation in electronic health records: a systematic review of challenges and methodologies. J Biomed Inform 2022;126:103980 [FREE Full text] [doi: 10.1016/j.jbi.2021.103980] [Medline: 34974189]
20. Zhao D, Shi Y, Cheng L, Li H, Zhang L, Guo H. Time interval uncertainty-aware and text-enhanced based disease prediction. J Biomed Inform 2023;139:104239 [FREE Full text] [doi: 10.1016/j.jbi.2022.104239] [Medline: 36356933]
21. Allen JF. Towards a general theory of action and time. Artif Intell 1984;23(2):123-154. [doi: 10.1016/0004-3702(84)90008-0]
22. Pustejovsky JD, Castaño J, Ingria RJP, Saurí R, Gaizauskas RJ, Setzer A. TIMEML: robust specification of event and temporal expressions in text. In: New Directions in Question Answering. Washington, D.C: AAAI Press; 2003:28-34.
23. Pustejovsky JD, Lee K, Bunt HC, Romary L. ISO-TIMEML : an international standard for semantic annotation. : European Language Resources Association (ELRA); 2010 Presented at: LREC 2010 - Proceedings of the 7th International Conference on Language Resources and Evaluation; 2010 May 17; La Valletta, Malta p. 394-397 URL: https://aclanthology.org/L10-1027/

XSL•FO
RenderX

24. Ferro L, Gerber L, Mani I, Sundheim BM, Wilson G. TIDES 2005 standard for the annotation of temporal expressions. Linguistic Data Consortium. 2005. URL: https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-timex2-guidelines-v0.1.pdf [accessed 2025-08-14]
25. Pustejovsky JD, Hanks P, Saurí R, See A, Gaizauskas RJ, Setzer A. The TIMEBANK corpus. 2003 Presented at: Proceedings of the Corpus Linguistics 2003 Conference; 2003 March 28; Lancaster, UK p. 647-656.
26. Boguraev BK, Pustejovsky JD, Ando RK, Verhagen M. TimeBank evolution as a community resource for TimeML parsing. Lang Resour Evaluation 2007;41(1):91-115. [doi: 10.1007/s10579-007-9018-8]
27. Ocal M, Radas A, Hummer J, Megerdoomian K, Finlayson MA. A comprehensive evaluation and correction of the Time Bank corpus. : European Language Resources Association (ELRA); 2022 Presented at: Proceedings of the 13th International Conference on Language Resources and Evaluation; 2022 June 20; Marseille, France p. 20-25 URL: https://aclanthology.org/2022.lrec-1.313/
28. Cassidy T, McDowell B, Chambers N, Bethard SJ. An annotation framework for dense event ordering. : Short Papers; 2014 Presented at: ACL 2014 - Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics; 2014 June 22; Baltimore, Maryland, USA p. 501-506. [doi: 10.3115/v1/p14-2082]
29. Chambers N, Cassidy T, McDowell B, Bethard SJ. Dense event ordering with a multi-pass architecture. Trans Assoc Comput Linguist 2014;2:273-284. [doi: 10.1162/tacl_a_00182]
30. Verhagen M, Gaizauskas RJ, Schilder F, Hepple M, Moszkowicz JL, Pustejovsky JD. The TempEval challenge: identifying temporal relations in text. Lang Resour Evaluation 2009;43(2):161-179. [doi: 10.1007/s10579-009-9086-z]
31. UzZaman N, Llorens H, Derczynski LRA, Verhagen M, Allen JF, Pustejovsky JD. SemEval-2013 Task 1: TempEval-3 - evaluating time expressions, events, and temporal relations. : Association for Computational Linguistics (ACL); 2013 Presented at: SEM 2013 - Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics; 2013 June 14; Atlanta, Georgia, USA p. 14-15 URL: https://aclanthology.org/S13-2001/
32. Verhagen M, Saurí R, Caselli T, Pustejovsky JD. SemEval-2010 Task 13: TempEval-2 - Evaluating events, time expressions, and temporal relations. : Association for Computational Linguistics (ACL); 2010 Presented at: SemEval 2010 - Proceedings of the 5th International Workshop on Semantic Evaluation @ ACL 2010; 2010 July 15; Uppsala, Sweden p. 57-62 URL: https://dl.acm.org/doi/10.5555/1859664.1859674
33. Graff D. The AQUAINT corpus of English news text. International Standard Language Resource Number. 2002. URL: https://www.islrn.org/resources/153-002-267-999-9/ [accessed 2026-01-06]
34. Minard AL, Speranza M, Agirre E, Aldabe I, van Erp M, Magnini B. Semeval-2015 Task 4: TimeLine - cross-document event ordering. : Association for Computational Linguistics (ACL); 2015 Presented at: SemEval 2015 - Proceedings of the 9th International Workshop on Semantic Evaluation @ NAACL-HLT 2015; 2015 June 04; Denver, Colorado, USA p. 778-786. [doi: 10.18653/v1/s15-2132]
35. O'Gorman T, Wright-Bettner K, Palmer M. Richer event description: integrating event coreference with temporal, causal and bridging annotation. : Curran Associates, Inc; 2016 Presented at: CNewsStory 2016 - Proceedings of the 2nd Workshop on Computing News Storylines @ EMNLP 2016; 2016 Nov 05; Austin, Texas, USA p. 47-56. [doi: 10.18653/v1/w16-5706]
36. Mostafazadeh N, Grealish A, Chambers N, Allen J, Vanderwende L. CaTeRS : causal and temporal relation scheme for semantic annotation of event structures. : Association for Computational Linguistics (ACL); 2016 Presented at: EVENTS - Proceedings of the 4th Workshop on Events: Definition, Detection, Coreference, and Representation @ NAACL-HLT 2016; 2016 June 17; San Diego, California p. 51-61. [doi: 10.18653/v1/w16-1007]
37. Caselli T, Vossen P. The Event StoryLine corpus: a new benchmark for causal and temporal relation extraction. : Association for Computational Linguistics (ACL); 2017 Presented at: EventStory 2017 - Proceedings of the [1st] Workshop on Events and Stories in the News @ ACL 2017; 2017 Aug 04; Vancouver, British Columbia, Canada p. 77-86. [doi: 10.18653/v1/w17-2711]
38. Silveira N, Doza T, de Marneffe MC, Bowman S, Connor M, Bauer J, et al. A gold standard dependency corpus for English. : European Language Resources Association (ELRA); 2014 Presented at: LREC 2014 - Proceedings of the 9th International Conference on Language Resources and Evaluation; 2014 May 26; Reykjavik, Iceland p. 2897-2904 URL: https://aclanthology.org/L14-1067/
39. Yao J, Qiu H, Min B, Xue N. Annotating temporal dependency graphs via crowdsourcing. : Association for Computational Linguistics (ACL); 2020 Presented at: EMNLP 2020 - Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing; 2020 Nov 16; USA p. 5368-5380. [doi: 10.18653/v1/2020.emnlp-main.432]
40. Wang X, Chen Y, Ding N, Peng H, Wang Z, Lin Y, et al. Maven-Ere: a unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. : Association for Computational Linguistics (ACL); 2022 Presented at: EMNLP 2022 - Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; 2022 Dec 07; Abu Dhabi, United Arab Emirates p. 926-941. [doi: 10.18653/v1/2022.emnlp-main.60]
41. Rogers A, Karpinska M, Gupta A, Lialin V, Smelkov G, Rumshisky A. Narrative Time : dense temporal annotation on a timeline. : European Language Resources Association (ELRA); 2024 Presented at: LREC-COLING 2024 - Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation; 2024 May 20; Torino, Italia p. 12053-12073 URL: https://aclanthology.org/2024.lrec-main.1054/

XSL•FO
RenderX

42. Verhagen M, Gaizauskas RJ, Schilder F, Hepple M, Katz G, Pustejovsky J. SemEval-2007 Task 15: TempEval - Temporal relation identification. : Association for Computational Linguistics (ACL); 2007 Presented at: SemEval 2007 - Proceedings of the 4th International Workshop on Semantic Evaluations @ ACL 2007; 2007 June 23; Prague, Czech Republic p. 75-80 URL: https://aclanthology.org/S07-1014/ [doi: 10.3115/1621474.1621488]
43. Llorens H, Chambers N, UzZaman N, Mostafazadeh N, Allen J, Pustejovsky J. SemEval-2015 task 5: QA TEMPEVAL - evaluating temporal information understanding with question answering. : Curran Associates, Inc; 2015 Presented at: SemEval 2015 - Proceedings of the 9th International Workshop on Semantic Evaluation @ NAACL-HLT 2015; 2015 June 04; Denver, Colorado, USA p. 792-800. [doi: 10.18653/v1/s15-2134]
44. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner Ö. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. J Am Med Inform Assoc 2011;18(5):540-543 [FREE Full text] [doi: 10.1136/amiajnl-2011-000465] [Medline: 21846785]
45. Filannino M, Uzuner Ö. Advancing the state of the art in clinical natural language processing through shared tasks. Yearb Med Inform 2018;27(1):184-192 [FREE Full text] [doi: 10.1055/s-0038-1667079] [Medline: 30157522]
46. Vashishtha S, Van Durme B, White AS. Fine-grained temporal relation extraction. : Association for Computational Linguistics (ACL); 2019 Presented at: ACL 2019 - Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; - August 2; Florence, Italy. Association for Computational Linguistics (ACL). 2019; 2019 July 28; Florence, Italy p. 2906-2919. [doi: 10.18653/v1/p19-1280]
47. Roberts A, Gaizauskas RJ, Hepple M, Demetriou G, Guo Y, Roberts I, et al. Building a semantically annotated corpus of clinical texts. J Biomed Inform 2009;42(5):950-966 [FREE Full text] [doi: 10.1016/j.jbi.2008.12.013] [Medline: 19535011]
48. Sun W, Rumshisky A, Uzuner Ö. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. J Am Med Inform Assoc 2013;20(5):806-813 [FREE Full text] [doi: 10.1136/amiajnl-2013-001628] [Medline: 23564629]
49. Sun W, Rumshisky A, Uzuner Ö. Annotating temporal information in clinical narratives. J Biomed Inform 2013;46:S5-S12 [FREE Full text] [doi: 10.1016/j.jbi.2013.07.004] [Medline: 23872518]
50. Styler WF, Bethard SJ, Finan S, Palmer M, Pradhan S, de Groen PC, et al. Temporal annotation in the clinical domain. Trans Assoc Comput Linguist 2014;2:143-154 [FREE Full text] [Medline: 29082229]
51. Bethard S, Derczynski L, Savova G, Pustejovsky J, Verhagen M. SemEval-2015 task 6: clinical TempEval. : Curran Associates; 2015 Presented at: SemEval 2015 - Proceedings of the 9th International Workshop on Semantic Evaluation @ NAACL-HLT 2015; 2015 June 04; Denver, Colorado, USA p. 806-814. [doi: 10.18653/v1/s15-2136]
52. Bethard SJ, Savova G, Chen WT, Derczynski L, Pustejovsky J, Verhagen M. SemEval-2016 task 12: clinical TempEval. : Association for Computational Linguistics (ACL); 2016 Presented at: SemEval 2016 - Proceedings of the 10th International Workshop on Semantic Evaluation @ NAACL-HLT 2016; 2016 June 16; San Diego, California, USA p. 1052-1062. [doi: 10.18653/v1/s16-1165]
53. Bethard S, Savova G, Palmer M, Pustejovsky J. SemEval-2017 task 12: clinical TempEval. : Association for Computational Linguistics (ACL); 2017 Presented at: SemEval 2017 - Proceedings of the 11th International Workshop on Semantic Evaluation @ ACL 2017; 2017 Aug 03; Vancouver, British Columbia, Canada p. 565-572. [doi: 10.18653/v1/s17-2093]
54. Mahajan D, Liang JJ, Tsou CH, Uzuner Ö. Overview of the 2022 n2c2 shared task on contextualized medication event extraction in clinical notes. J Biomed Inform 2023;144:104432 [FREE Full text] [doi: 10.1016/j.jbi.2023.104432] [Medline: 37356640]
55. Kumar V, Stubbs A, Shaw SY, Uzuner Ö. Creation of a new longitudinal corpus of clinical narratives. J Biomed Inform 2015;58(Suppl):S6-S10 [FREE Full text] [doi: 10.1016/j.jbi.2015.09.018] [Medline: 26433122]
56. Viani N, Tissot H, Bernardino A, Velupillai S. Annotating temporal information in clinical notes for timeline reconstruction: towards the definition of calendar expressions. 2019 Presented at: BioNLP 2019 - Proceedings of the 18th SIGBioMed Workshop on Biomedical Natural Language Processing and Shared Task @ ACL 2019; 2019 Aug 01; Florence, Italy p. 201-210. [doi: 10.18653/v1/w19-5021]
57. Spreyer K, Frank A. Projection-based acquisition of a temporal labeller. : Asian Federation of Natural Language Processing (AFNLP); 2008 Presented at: IJCNLP 2008 - Proceedings of the 3rd International Joint Conference on Natural Language Processing; 2008 Jan 07; Hyderabad, India p. 489-496 URL: https://aclanthology.org/I08-1064/
58. Strötgen J, Gertz M. WikiWarsDE: a German corpus of narratives annotated with temporal expressions. In: Multilingual Resources and Multilingual Applications.: Hamburger Zentrum für Sprachkorpora; 2011 Presented at: GSCL 2011 - Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology; 2011 Sep 28; Hamburg, Germany p. 129-134 URL: http://dbs.ifi.uni-heidelberg.de/fileadmin/Team/jannik/publications/StroetgenGertz_WikiWarsDE_GSCL2011.pdf
59. Mazur P, Dale R. WIKIWARS : a new corpus for research on temporal expressions. : Association for Computational Linguistics (ACL); 2010 Presented at: EMNLP 2010 - Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing; 2010 Oct 09; Boston, Massachusetts, USA p. 913-922 URL: https://aclanthology.org/D10-1089/
60. Strötgen J, Minard A, Lange L, Speranza M, Magnini B. KRAUTS: a German temporally annotated news corpus. : European Language Resources Association (ELRA); 2018 Presented at: LREC 2018 - Proceedings of the 11th International Conference on Language Resources and Evaluation; 2018 May 07; Miyazaki, Japan p. 536-540 URL: https://aclanthology.org/L18-1085/

XSL•FO
RenderX

61. Strötgen J, Gertz M. Multilingual and cross-domain temporal tagging. Lang Resour Evaluation 2012;47(2):269-298. [doi: 10.1007/s10579-012-9179-y]
62. Lücking A, Stoeckel M, Abrami G, Mehler A. I still have time(s): extending HeidelTime for German texts. : European Language Resources Association (ELRA); 2022 Presented at: LREC 2022 - Proceedings of the 13th International Conference on Language Resources and Evaluation; 2022 June 20; Marseille, France p. 4723-4728 URL: https://aclanthology.org/2022.lrec-1.505/
63. Roller R, Uszkoreit H, Xu F, Seiffe L, Mikhailov M, Staeck O. A fine-grained corpus annotation schema of German nephrology records. : COLING 2016 Organizing Committee; 2016 Presented at: ClinicalNLP 2016 - Proceedings of the 1st Workshop on Clinical Natural Language Processing @ COLING 2016; 2016 December 11; Osaka, Japan p. 69-77 URL: https://aclanthology.org/W16-4210/
64. Campillos L, Deléger L, Grouin C, Hamon T, Ligozat A, Névéol A. A French clinical corpus with comprehensive semantic annotations: development of the medical entity and relation LIMSI annOtated Text corpus (MERLOT). Lang Resour Evaluation 2017;52(2):571-601. [doi: 10.1007/s10579-017-9382-y]
65. Hamon T, Grabar N. Tuning HeidelTime for identifying time expressions in clinical texts in English and French. 2014 Presented at: LOUHI 2014 - Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis @ EACL 2014; 2014 April 27; Gothenburg, Sweden p. 101-105. [doi: 10.3115/v1/w14-1116]
66. Bannour N, Rance B, Tannier X, Névéol A. Event-independent temporal positioning: application to French clinical text. 2023 Presented at: BioNLP 2023 - Proceedings of the 22nd Workshop on Biomedical Language Processing (BioNLP) & BioNLP Shared Tasks (BioNLP-ST) @ ACL 2023; 2023 July 13; Ontario, Canada p. 191-205. [doi: 10.18653/v1/2023.bionlp-1.16]
67. Jeyafreeda AJ, Vincent M, Burgun A, Garcelon N. Evaluating LLMs for temporal entity extraction from pediatric clinical text in rare diseases context. 2024 Presented at: CL4Health 2024 - Proceedings of the 1st Workshop on Patient-Oriented Language Processing @ LREC-COLING 2024; 2024 May 20; Torino, Italia p. 145-152 URL: https://aclanthology.org/2024.cl4health-1.18/
68. Liu Z, Wang X, Chen Q, Tang B, Xu H. Temporal indexing of medical entity in Chinese clinical notes. BMC Med Inform Decis Mak 2019;19(Suppl 1):17 [FREE Full text] [doi: 10.1186/s12911-019-0735-x] [Medline: 30700331]
69. Pan X, Chen B, Weng H, Gong Y, Qu Y. Temporal expression classification and normalization from Chinese narrative clinical texts: pattern learning approach. JMIR Med Inform 2020;8(7):e17652 [FREE Full text] [doi: 10.2196/17652] [Medline: 32716307]
70. Liu Z, Tang B, Wang X, Chen Q, Li H, Bu J, et al. CMedTEX: a rule-based temporal expression extraction and normalization system for Chinese clinical notes. AMIA Annu Symp Proc 2016;2016:818-826 [FREE Full text] [Medline: 28269878]
71. Hu D, Wang M, Gao F, Xu F, Gu J. Knowledge representation and reasoning for complex time expression in clinical text. Data Intell 2022;4(3):573-598. [doi: 10.1162/dint_a_00152]
72. Velupillai S. Temporal expressions in Swedish medical text: a pilot study. 2014 Presented at: BioNLP 2014 - Proceedings of the Workshop on Biomedical Natural Language Processing @ ACL 2014; 2024 June 27; Baltimore, Maryland, USA p. 88-92. [doi: 10.3115/v1/w14-3413]
73. de Azevedo RF, Santos Rodrigues JP, da Silva Reis MR, Cabral Moro CM, Cabrera Paraiso E. Temporal tagging of noisy clinical texts in Brazilian Portuguese. In: Computational Processing of the Portuguese Language. 2018 Presented at: PROPOR 2018 - Proceedings of the 13th International Conference on Computational Processing of the Portuguese Language; 2018 Sep 24; Canela, Brazil p. 231-241. [doi: 10.1007/978-3-319-99722-3_24]
74. Viani N, Miller TA, Napolitano C, Priori SG, Savova GK, Bellazzi R, et al. Supervised methods to extract clinical events from cardiology reports in Italian. J Biomed Inform 2019;95:103219 [FREE Full text] [doi: 10.1016/j.jbi.2019.103219] [Medline: 31150777]
75. Cheng F, Yada S, Tanaka R, Aramaki E, Kurohashi S. JaMIE: a pipeline Japanese medical information extraction system with novel relation annotation. 2022 Presented at: LREC 2022 - Proceedings of the 13th International Conference on Language Resources and Evaluation; 2022 June 20; Marseille, France p. 3724-3731 URL: https://aclanthology.org/2022.lrec-1.397/
76. Shimizu S, Kanashiro PL, Yada S, Aramaki E. QA-based event start-points ordering for clinical temporal relation annotation. 2024 Presented at: LREC-COLING 2024 - Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation; 2024 May 20; Torino, Italia p. 13371-13381 URL: https://aclanthology.org/2024.lrec-main.1171/
77. Magnini B, Altuna B, Lavelli A, Minard A, Speranza M, Zanoli R. European clinical case corpus. In: Rehm G, editor. European Language Grid: a Language Technology Platform for Multilingual Europe. New York: Springer; 2023:283-288.
78. Hahn U, Matthies F, Lohr C, Löffler M. 3000PA-towards a national reference corpus of German clinical language. Stud Health Technol Inform 2018;247:26-30. [Medline: 29677916]
79. Lohr C, Luther S, Matthies F, Modersohn L, Ammon D, Saleh K, et al. CDA-compliant section annotation of German-language discharge summaries: guideline development, annotation campaign, section classification. AMIA Annu Symp Proc 2018;2018:770-779 [FREE Full text] [Medline: 30815119]

XSL•FO
RenderX

80. Kolditz T, Lohr C, Hellrich J, Modersohn L, Betz B, Kiehntopf M. Annotating German clinical documents for de-identification. Stud Health Technol Inform 2019;264:203-207. [doi: 10.3233/shti190212]
81. Lohr C, Modersohn L, Hellrich J, Kolditz T, Hahn U. An evolutionary approach to the annotation of discharge summaries. Stud Health Technol Inform 2020;270:28-32. [doi: 10.3233/shti200116]
82. Modersohn L, Schulz S, Lohr C, Hahn U. GraSCCo: the first publicly shareable, multiply-alienated German clinical text corpus. Stud Health Technol Inform 2022;270:66-72. [doi: 10.3233/shti220805]
83. Schulz S. GraSCCo. Zenodo (CERN European Organization for Nuclear Research). 2022. URL: https://doi.org/10.5281/zenodo.6539131 [accessed 2025-08-14]
84. Kweon S, Kim J, Kim J, Im S, Cho E, Bae S, et al. Publicly shareable clinical large language model built on synthetic clinical notes. : Association for Computational Linguistics (ACL); 2024 Presented at: Findings of the Association for Computational Linguistics - ACL 2024; 2024 Aug 11; Bangkok, Thailand p. 5148-5168. [doi: 10.18653/v1/2024.findings-acl.305]
85. Li J, Zhou Y, Jiang X, Natarajan K, Pakhomov SVS, Liu H, et al. Are synthetic clinical notes useful for real natural language processing tasks: a case study on clinical entity recognition. J Am Med Inform Assoc 2021;28(10):2193-2201 [FREE Full text] [doi: 10.1093/jamia/ocab112] [Medline: 34272955]
86. Ive J, Viani N, Kam J, Yin L, Verma S, Puntis S, et al. Generation and evaluation of artificial mental health records for natural language processing. NPJ Digit Med 2020;3(69):1-9 [FREE Full text] [doi: 10.1038/s41746-020-0267-x] [Medline: 32435697]
87. Hiebel N, Ferret O, Fort K, Névéol A. Can synthetic text help clinical named entity recognition? a study of electronic health records in French. : Association for Computational Linguistics (ACL); 2023 Presented at: EACL 2023 - Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics; 2023 May 06; Dubrovnik, Croatia p. 2320-2338. [doi: 10.18653/v1/2023.eacl-main.170]
88. Brekke PH, Rama T, Pilán I, Nytrø Ø, Øvrelid L. Synthetic data for annotation and extraction of family history information from clinical text. J Biomed Semantics 2021;12(1):11 [FREE Full text] [doi: 10.1186/s13326-021-00244-2] [Medline: 34261535]
89. Pustejovsky J, Stubbs A. Natural Language Annotation for Machine Learning - A Guide to corpus-building for Applications. California: O'Reilly Media, Inc; 2012.
90. Artstein R, Poesio M. Inter-coder agreement for computational linguistics. Comput Linguist 2008;34(4):555-596. [doi: 10.1162/coli.07-034-R2]
91. Kolditz T. Bratiaa: inter-annotator agreement for brat annotation projects. GitHub. 2019. URL: https://github.com/kldtz/bratiaa [accessed 2025-08-14]
92. Reimers N, Dehghani N, Gurevych I. Temporal anchoring of events for the TIMEBANK corpus. : Association for Computational Linguistics (ACL); 2016 Presented at: ACL 2016 - Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics; 2016 Aug 07; Berlin, Germany p. 2195-2204. [doi: 10.18653/v1/p16-1207]
93. Verhagen M. Temporal closure in an annotation environment. Language Res Eval 2006;39(2-3):211-241. [doi: 10.1007/s10579-005-7884-5]
94. Stenetorp P, Pyysalo S, Topi G, Ohta T, Ananiadou S, Tsujii J. BRAT: a Web-based tool for NLP-assisted text annotation. : Association for Computational Linguistics (ACL); 2012 Presented at: EACL 2012 - Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations; 2012 Apr 25; Avignon, France p. 25-26 URL: https://aclanthology.org/E12-2021/
95. Chan B, Schweter S, Möller T. German's next language model. : International Committee on Computational Linguistics (ICCL); 2020 Presented at: COLING 2020 - Proceedings of the 28th International Conference on Computational Linguistics; 2020 Dec 08; Barcelona, Spain p. 6788-6796. [doi: 10.18653/v1/2020.coling-main.598]
96. Falkner S, Klein A, Hutter F. BOHB: robust and efficient hyperparameter optimization at scale. : Proceedings of Machine Learning Research; 2018 Presented at: ICML 2018 - Proceedings of the 35th International Conference on Machine Learning; 2018 July 10; Stockholm, Sweden p. 1437-1446 URL: https://proceedings.mlr.press/v80/falkner18a.html
97. Huggingface: Huggingface/Transformers. GitHub. 2020. URL: https://github.com/huggingface/transformers [accessed 2025-08-14]
98. Microsoft: neural network intelligence. GitHub. 2021. URL: https://github.com/microsoft/nni [accessed 2025-08-14]
99. Goyal T, Durrett G. Embedding time expressions for deep temporal ordering models. : Association for Computational Linguistics (ACL); 2019 Presented at: ACL 2019 - Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Aug 02; Florence, Italy p. 4400-4406. [doi: 10.18653/v1/p19-1433]
100. Han Z, Zhang G, Ma Y, Tresp V. Time-dependent entity embedding is not all you need: a re-evaluation of temporal knowledge graph completion models under a unified framework. : Association for Computational Linguistics (ACL); 2021 Presented at: EMNLP 2021 - Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; 2021 Nov 07; Punta Cana, Dominican Republic p. 8104-8118. [doi: 10.18653/v1/2021.emnlp-main.639]
101. Gao Y, He Y, Kan Z, Han Y, Qiao L, Li D. Learning joint structural and temporal contextualized knowledge embeddings for temporal knowledge graph completion. : Association for Computational Linguistics (ACL); 2023 Presented at: Findings

XSL•FO
RenderX

of the Association for Computational Linguistics - ACL 2023; 2023 July 09; Toronto, Ontario, Canada p. 417-430. [doi: 10.18653/v1/2023.findings-acl.28]
102. Mathur P, Jain R, Dernoncourt F, Morariu V, Tran QH, Manocha D. TIMERS: document-level temporal relation extraction. : Association for Computational Linguistics (ACL); 2021 Presented at: ACL-IJCNLP 2021 - Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics & 11th International Joint Conference on Natural Language Processing; 2021 Aug 01; Stroudsburg, PA p. 524-533. [doi: 10.18653/v1/2021.acl-short.67]
103. Sun Q, Zhang K, Huang K, Xu T, Li X, Liu Y. Document-level relation extraction with two-stage dynamic graph attention networks. Knowl Based Syst 2023;267:110428. [doi: 10.1016/j.knosys.2023.110428]
104. Li Y, Wehbe RM, Ahmad FS, Wang H, Luo Y. A comparative study of pretrained language models for long clinical text. J Am Med Inform Assoc 2023;30(2):340-347 [FREE Full text] [doi: 10.1093/jamia/ocac225] [Medline: 36451266]
105. Roberts A, Gaizauskas R, Hepple M, Davis N, Demetriou G, Guo Y, et al. The CLEF corpus: semantic annotation of clinical text. AMIA Annu Symp Proc 2007;2007:625-629 [FREE Full text] [Medline: 18693911]
106. Raghavan P, Fosler-Lussier E, Lai AM. Inter-annotator reliability of medical events, coreferences and temporal relations in clinical narratives by annotators with varying levels of clinical expertise. AMIA Annu Symp Proc 2012;2012:1366-1374 [FREE Full text] [Medline: 23304416]
107. Bastardo R, Castro M, Pavão JA, Ramos L. A timeline model for clinical events: empowering data. Procedia Comput Sci 2022;196:1053-1058. [doi: 10.1016/j.procs.2021.12.109]
108. Singh A, Hummer J, Ocal M, Finlayson MA. PYTLEX : A PYTHON library for TimeLine extraction. : Association for Computational Linguistics (ACL); 2024 Presented at: EACL 2024 - Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations; 2024 March 17; St. Julian's, Malta p. 27-34. [doi: 10.18653/v1/2024.eacl-demo.4]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
GBERT: German BERT model
IAA: interannotator agreement
ISO: International Organization for Standardization
NLP: natural language processing

Edited by J Sarvestan; submitted 18.Jan.2025; peer-reviewed by C Li, G Recski, K Arzideh; comments to author 15.Feb.2025; revised version received 30.Dec.2025; accepted 31.Dec.2025; published 25.Feb.2026.

Please cite as:
Modersohn L, Hahn U
Temporal Annotation of German Clinical Language in Real and Synthetic Clinical Documents: Corpus Development and Baseline Tagger Validation Study
J Med Internet Res 2026;28:e71458
URL: https://www.jmir.org/2026/1/e71458
doi:10.2196/71458
PMID:

©Luise Modersohn, Udo Hahn. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Extracting and Classifying Drug Discontinuations From Estonian Electronic Health Records: Development and Validation Study

Hendrik Šuvalov[1], MSc; Nikita Umov[2], MD; Maria Malk[1], MSc; Markus Haug[1], MSc; Sven Laur[1], PhD; Marek Oja[1], PhD; Sirli Tamm[1], MSc; Sulev Reisberg[1,3], PhD; Jaak Vilo[1,3], PhD; Raivo Kolde[1], PhD

[1]Institute of Computer Science, University of Tartu, Tartu, Estonia

[2]Institute of Family Medicine and Public Health, University of Tartu, Tartu, Estonia

[3]Software Technology and Applications Competence Center, Tartu, Estonia

Corresponding Author:
Hendrik Šuvalov, MSc
Institute of Computer Science
University of Tartu
Narva mnt 18
Tartu, 51009
Estonia
Phone: 372 7375100
Email: hendrik.suvalov@ut.ee

Abstract

Background: Drug adherence is crucial for chronic disease management, yet treatment discontinuation remains common due to factors such as side effects, inefficacy, or cost. These reasons are often recorded only in free-text clinical notes, making large-scale analysis difficult. While large language models (LLMs) can interpret such unstructured data more effectively than traditional natural language processing methods, few studies have systematically categorized reasons for discontinuation or identified whether the decision was initiated by the patient or the clinician, especially in low-resource languages such as Estonian.

Objective: This study aimed to assess the ability of LLMs to extract and classify reasons for drug discontinuation and identify who initiated it using Estonian electronic health records and characterize the observed discontinuation patterns and initiators for statins and antidiabetic medications.

Methods: We combined prescription data with free-text anamneses from a 10% sample of the Estonian population (2012-2019). LLMs (Llama 3.1-70B and GPT-4o) were applied to extract discontinuation phrases and reasons, classify them into a clinician-developed taxonomy, and identify who discontinued the treatment. Performance was evaluated on 100 randomly chosen cases per drug group.

Results: Extraction yielded 625 antidiabetic drug and 233 statin discontinuation cases. Validation confirmed a precision of 0.93 to 0.98 for extracting phrases and 0.95 to 0.96 for extracting reasons. Classification of discontinuation reasons achieved weighted F_1-scores of 0.81 to 0.84, whereas classification of who initiated discontinuation achieved weighted F_1-scores of 0.64 to 0.78. Adverse reactions were the most frequent reason overall, accounting for 70% (163/233) of statin discontinuations and 44.8% (280/625) of antidiabetic drug discontinuations. Regarding antidiabetic drugs, treatment inefficacy and contraindications were more common. Patients more often stopped due to adverse reactions or nonmedical reasons, whereas physicians more often initiated discontinuation for contraindications.

Conclusions: LLMs can accurately extract and classify medication discontinuation reasons and show variable performance in identifying discontinuation initiators in Estonian clinical narratives. Both local and proprietary models showed promising results, enabling scalable analyses that complement structured health records. This demonstrates the potential of LLMs to unlock information from clinical notes, turning this underused electronic health record component into a valuable resource for monitoring treatment patterns and detecting adverse event signals.

(J Med Internet Res 2026;28:e86183) doi:10.2196/86183

KEYWORDS

natural language processing; NLP; large language model; LLM; Estonian; health care data; drug adherence; drug discontinuations; clinical decision support; data mining

XSL•FO
RenderX

Introduction

Drug adherence and persistence are critical for the management of chronic diseases as they directly affect both patient outcomes and health care costs [1]. Nevertheless, discontinuation of prescribed medications is common [2], driven by factors such as adverse events, lack of perceived efficacy, patient preference, and drug cost [3,4]. While understanding these reasons is essential for improving adherence and informing clinical decision-making, they are rarely standardized or systematically coded in electronic health records (EHRs). When such information is recorded at all, it is typically documented in free-text clinical notes, which are difficult to analyze at scale without manual review [5].

Traditional natural language processing (NLP) approaches, including rule-based methods and classic machine learning, can capture common patterns but often fail to detect the nuanced or less frequent circumstances underlying discontinuation decisions. Large language models (LLMs) offer a promising alternative as they are able to handle semantic complexity and variability in clinical text, making it feasible to classify reasons for discontinuation across thousands of records. Clinical narratives such as patient anamneses (case history), though unstructured and heterogeneous, contain valuable information about patients' experiences and the clinical rationale behind stopping medications and, thus, represent an important but underused data source.

Prior research on extracting and classifying drug discontinuation reasons has primarily relied on classic NLP methods such as rule-based approaches or machine learning models based on bidirectional encoder representations from transformers architectures [5,6]. More recently, LLMs have shown strong performance on these tasks. For example, LLMs have been applied to study discontinuations in web-based forums [7], which provide valuable patient-reported insights but often lack clinical context. EHRs have been analyzed to identify discontinued medications using LLMs [8], but reasons for discontinuation were not categorized. Related work has also applied NLP and LLM-based approaches to adverse drug event detection in clinical narratives and EHR data, highlighting the potential of these methods for pharmacovigilance and medication safety research [9,10]. More broadly, recent studies have demonstrated the utility of LLMs for extracting, structuring, and classifying information from unstructured clinical narratives and EHR data, with a strong performance [11-14]. Research in lower-resource languages is very scarce, and studies addressing whether discontinuation was initiated by the patient or the clinician are also lacking.

In this study, we used LLMs to examine the discontinuation of antidiabetics and statins among Estonian patients. Specifically, we aimed to (1) identify patients who were prescribed a drug from either category and subsequently stopped purchasing it, (2) extract text segments from anamneses that mention discontinuation or reasons for stopping, (3) determine whether the decision was made by the patient or the physician, and (4) categorize the identified reasons using a clinically developed ontology. By linking structured prescription records with unstructured anamneses, we ensured that our cohort truly reflected patients who discontinued therapy and that negative mentions corresponded to actual discontinuation of treatment. This approach provides systematic insights into the drivers of medication discontinuation in chronic disease management in Estonia and demonstrates the utility of LLMs for pharmacoepidemiological studies in a low-resource language.

Methods

Dataset Curation

The dataset used for this study was RITA MAITT, a 10% random sample of the Estonian population (n=150,824) from 2012 to 2019 consisting of claims, prescription, and EHR data [15]. For this study, only unique and nonempty anamneses were used.

From the prescription records, we identified individuals who had purchased drugs classified under the Anatomical Therapeutic Chemical (ATC) groups A10 (antidiabetics) or C10 (lipid-modifying agents). To detect discontinuation, we focused on purchases made in 2012 to 2018 and required a subsequent 1-year observation window without further purchases of the same drug, ensuring that the absence of refills reflected treatment discontinuation. This 1-year window was chosen to align with clinical practice, where patients with chronic conditions are typically reviewed by a physician at least annually, and to ensure the reliability of discontinuation reasons, some of which were self-reported and may be less accurate over longer recall periods. Importantly, patients were considered to have discontinued a medication if they stopped purchasing one specific ATC code drug even if they later initiated another drug within the same class (eg, discontinuation of C10AA05 followed by initiation of C10AA07). This ensured that discontinuation was defined at the level of individual drugs rather than drug classes.

After establishing the discontinuation cohort based on these predefined rules, we extracted free-text clinical notes from the EHRs, specifically anamnesis sections documented within 1 year after the last recorded purchase. Anamneses are particularly informative in this context as they often capture both patient-reported experiences and clinician assessments. Unlike structured fields, anamneses provide the narrative detail necessary to understand the circumstances surrounding treatment discontinuation, making them a valuable source for identifying and classifying underlying reasons.

We obtained 24,040 unique anamneses from 4928 patients for A10 and 27,290 unique anamneses from 7363 patients for C10. The data flow diagram is shown in Figure 1 (cohort creation section).

XSL•FO
RenderX

Figure 1. Data flow diagram consisting of two parts: (1) cohort creation—rule-based filtering of anamneses relevant for discontinuation mentions and (2) large language model (LLM) extraction—extracting documents containing discontinuations linked to antidiabetics (A10) or statins (C10). a: anamneses; n: number of patients.

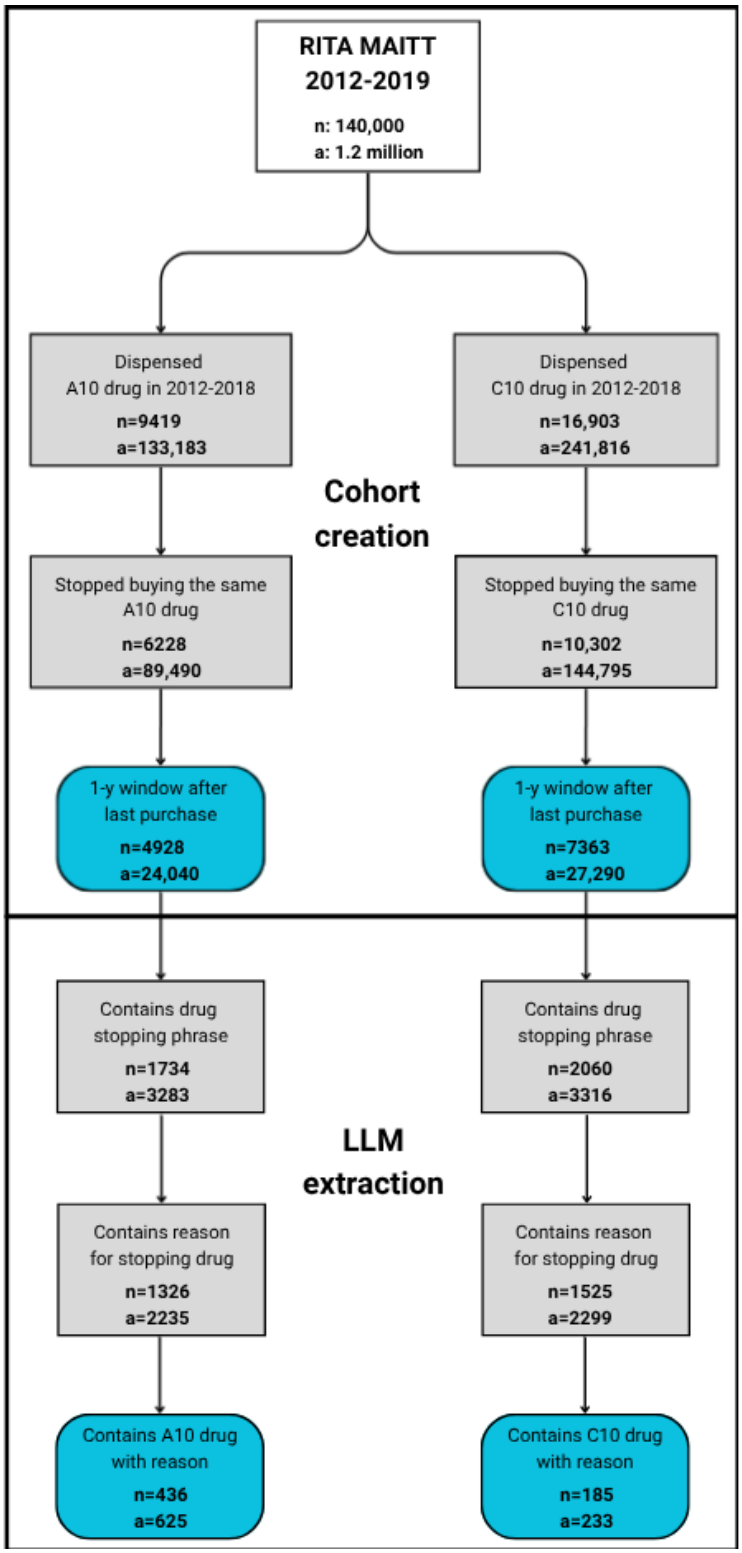

Extracting Discontinuation Information From Clinical Narratives

On the basis of expert review of the data, many anamneses did not explicitly state discontinuation but, instead, described adverse effects or lack of efficacy. As structured prescription data confirmed that drug dispensing had stopped, such negative experiences (eg, “had no effect” and “caused diarrhea”) were classified as discontinuation reasons even without an explicit mention of stopping.

To systematically identify and structure the information on drug discontinuation from free-text anamneses, we applied an LLM. Specifically, we used Meta-Llama-3.1-70B-Instruct-quantized.w4a16 [16], with a

XSL•FO
RenderX

temperature of 0 to ensure consistent outputs across runs. The model was hosted locally in a secure environment to ensure that no sensitive patient data were transmitted or shared externally. Each anamnesis associated with A10 or C10 drug discontinuation in the previous step was provided to the model together with a structured instruction prompt. In a few-shot fashion, each prompt included 10 representative examples curated by a clinician. Although the source data were in Estonian, prompts were written in English as LLMs are predominantly trained on English-language corpora [16,17] and prior work has shown that few-shot prompting with English-language instructions yields the best performance on Estonian EHR data [18]. The prompt template is provided in Multimedia Appendix 1.

To enforce a reproducible and machine-readable format, we constrained the model output using a predefined schema defined using Pydantic, a Python package for validating JSON objects [19]. The schema required the output to include the fields shown in Table 1.

Table 1. Extracted fields and their description.

Extracted field	Description
drug_stop_phrase	The full discontinuation-related sentence or clause from the anamnesis (eg, “patient stopped taking metformin because of headaches”), explicitly including the action of stopping, the drug name, and the reported reason
reason_for_stopping	A shorter, normalized span capturing only the reason for discontinuation (eg, “headaches”)
drug_name	The medication name as it appeared in the text

The extracted drug names were subsequently reviewed by a clinician. Any variants, abbreviations, or nonstandard mentions were manually harmonized and mapped to their corresponding ATC codes (A10 or C10). This step ensured alignment between free-text mentions and structured prescription data.

This approach allowed us to capture discontinuation events at the level of individual drugs with both the narrative context (via the *drug_stop_phrase*) and a shorter representation of discontinuation reasons suitable for downstream quantitative analysis. This pipeline is illustrated in Figure 1 (LLM extraction section).

Classifying Discontinuations

The World Health Organization’s 5-dimension model attributes poor medication adherence to patient, therapy, condition, system, and socioeconomic factors [20]. However, this framework is not specific to discontinuation and is too broad to be applied as practical labeling categories. Subsequent studies have proposed varying taxonomies for discontinuation reasons, often tailored to the type of medication or clinical context [5,7,21]. In this study, we developed a taxonomy of discontinuation reasons based on the framework proposed by Trevena et al [7] but adapted it through a qualitative exploration of our data. Two considerations guided these modifications: first, that the source material reflects physician-entered records rather than patient-reported notes and, second, that the drugs of interest fall within the A10 and C10 classes. The full set of classes and definitions is provided in Table 2.

Table 2. Discontinuation reasons and definitions.

Class	Definition
Adverse reactions	Discontinuation due to adverse side effects, allergic reactions, or negative interactions with other medications
Treatment success	Discontinuation due to successful treatment completion or sufficient improvement in health
Treatment inefficacy	Perceived ineffectiveness of the treatment or loss of belief in its efficacy
Contraindication	Discontinuation due to the emergence or discovery of a medical condition or risk factor that makes the continued use of the treatment unsafe or inappropriate
Nonmedical reasons	Discontinuation due to factors unrelated to the patient’s health or the treatment’s medical effects, such as financial constraints, access issues, personal choice, cultural beliefs, or social circumstances
Other	Other medical reasons not explicitly covered by the other categories
Indeterminate	Unclear or unspecified reasons for discontinuation

To assign each discontinuation instance to one of these categories, we again applied an LLM-based approach. All analyses were first performed locally using Meta-Llama-3.1-70B-Instruct-quantized.w4a16. The model was provided with the extracted *reason_for_stopping* text along with the list of possible categories (Table 2) in the system prompt. Each query was framed in a few-shot format using 10 clinician-curated examples to illustrate the mapping process. To ensure structured outputs, we enforced a predefined Pydantic schema, which required a valid JSON response containing the predicted category.

For external validation, the same task was subsequently repeated using GPT-4o (OpenAI) [17] as the extraction of narrowly defined phrases excluded potential privacy concerns. All the extracted reasons were manually reviewed beforehand to confirm that no sensitive information was present. This allowed us to compare outputs between a local deployment and a proprietary model while maintaining patient privacy.

XSL•FO
RenderX

To explore potential differences in discontinuation reason patterns, we stratified texts into switches—texts from patients who switched from one drug to another within the same class—and nonswitches—texts from patients who discontinued without switching—and analyzed the differences. We focused on results from the local model to establish a reproducible workflow for future studies.

In addition, we used the same LLM framework to classify who initiated the discontinuation. The model was prompted (Multimedia Appendix 1) to assign each case to 1 of 3 categories:

- Doctor—discontinuation recommended by the doctor
- Patient—the patient stopped independently
- Unspecified—insufficient information to determine the initiator

Evaluation Metrics

A random sample of 100 discontinuation cases was evaluated for both statins and antidiabetics. Each case was independently reviewed by 2 medical experts.

These sampled cases were used to evaluate both extraction precision and classification performance. A medical expert assessed whether the extracted discontinuation phrases and reasons were valid representations of the underlying clinical events. Precision was reported as the proportion of outputs judged valid, with 95% CIs calculated using the Wilson method for binomial proportions. For the classification tasks (discontinuation reason and initiator of discontinuation), model predictions were compared against the expert ground truth. Interannotator agreement between the 2 medical experts was assessed using the Cohen κ [22]. We report accuracy (equivalent to microaveraged F_1-scores); weighted F_1-score to account for class imbalance; and the Cohen κ, which adjusts for agreement expected by chance. CIs for classification metrics were estimated using nonparametric bootstrapping with 5000 resamples.

The recall evaluation was concentrated on documents in which no reason for stopping the drug was initially detected. Sets of 100 documents were randomly sampled from this group for both drug classes, and a medical expert reviewed these texts to identify both the reason for stopping the drug and whether the drug of interest was mentioned. The observed number of missed cases was used to estimate the false negative rate for reason detection, and exact binomial CIs were calculated for this rate. The estimated rate and its CIs were then extrapolated to the full set of documents without a detected drug-stopping reason to obtain estimates of the number of false negatives and corresponding CIs. Recall was subsequently calculated, with CIs derived by propagating the false-negative CI bounds through the same formula.

Ethical Considerations

The medical texts were accessed according to the approval of the Estonian Committee on Bioethics and Human Research (1.1-12/1279), which waived the requirement for informed consent for the use of the data. All data were pseudonymized prior to analysis to prevent the identification of individual patients. All data processing and model inference were performed in a secure local environment to ensure patient confidentiality. The only textual information shared with third-party services consisted of extracted discontinuation reasons that were manually verified to contain no sensitive or personally identifiable information. All other analyses were conducted locally in compliance with applicable data protection regulations and ethical standards governing the use of medical data.

Results

Cohort Overview

We focused on clinical texts that simultaneously mentioned a drug from the target group in the extracted phrase and provided an associated reason for discontinuation. Of 24,040 antidiabetic cases, 625 (2.6%) contained both a reason and a drug from the target group, and of 27,290 statin cases, 233 (0.9%) contained both a reason and a drug. For antidiabetic discontinuation, 315 patients had 1 case, 76 had 2 cases, and 45 had 3 or more cases. For statin discontinuation, 152 patients had 1 case, 24 had 2 cases, and 9 had 3 or more cases.

Of the 625 antidiabetic cases, 530 (84.8%) were switches—texts from patients who stopped one drug but started another within the same drug class—and 95 (15.2%) were nonswitches—texts from patients who discontinued without switching to another drug in the same class. Similarly, of the 233 statin cases, 102 (43.8%) were switches and 131 (56.2%) were nonswitches.

The age and sex distributions of patients in the extracted cases were broadly similar to those of all patients who discontinued the respective drugs (Multimedia Appendix 1), suggesting that the extracted subset was reasonably representative of the overall population. However, the relatively high average age of statin and antidiabetic users implies a higher burden of comorbidities, which may introduce additional noise and variability into the extracted discontinuation reasons due to concurrent conditions and treatments. The distribution of time between the last prescription purchase and the epicrisis date containing the analyzed discontinuation case is provided in Multimedia Appendix 1.

Each identified case was further classified by discontinuation reason and by who initiated the decision (patient or physician).

Evaluation

Extraction Performance

Validation of 100 sampled cases per drug group demonstrated a high precision of the extraction process. For antidiabetics, a medical expert reviewed the local LLM-extracted discontinuation phrases and confirmed a precision of 0.93 (93/100, 93% of the cases; 95% CI 0.86-0.97) for discontinuation phrases and 0.95 (95/100, 95% of the cases; 95% CI 0.89-0.98) for extracted reasons. For statins, precision was 0.98 (98/100, 98% of the cases; 95% CI 0.93-0.99) for discontinuation phrases and 0.96 (96/100, 96% of the cases; 95% CI 0.90-0.98) for extracted reasons.

To understand how many cases were missed by the LLM, a negative validation sample of 100 randomly selected

XSL•FO
RenderX

discontinuation documents per drug group in which no reason was detected by the model was reviewed. This analysis revealed, for both drug classes, 0 cases in which both a discontinuation reason and the correct drug were present. Consequently, an accurate estimate of the false negative rate and, thus, recall could not be derived for this set of documents.

Using a relaxed criterion requiring only the presence of a discontinuation reason irrespective of drug attribution, this review identified one missed reason among antidiabetic documents and 2 missed reasons among statin documents. On the basis of these observed false negative rates, the estimated recall for reason extraction was 0.91 (95% CI 0.65-1.00) for antidiabetics and 0.82 (95% CI 0.57-0.97) for statins. The calculation of both precision and recall, including their corresponding CIs, is provided in Multimedia Appendix 2.

To highlight the variability in how discontinuation reasons are expressed in the free text, Table 3 summarizes the distribution of unique *reason_for_stopping* patterns by frequency and their overall coverage.

Table 3. Distribution of unique discontinuation reason text patterns as they appeared in the texts by frequency and their coverage of the total textsa.

Frequency range	Antidiabetic (A10) patterns and coverage (n=625), n (%)	Statin (C10) patterns and coverage (n=233), n (%)
Exactly once	454 (72.6)	166 (71.2)
2-4 times	58 (22.2)	22 (22.7)
5-9 times	3 (3.2)	2 (6)
≥10 times	1 (1.9)	0 (0)

[a]Coverage represents the percentage of all extracted reason texts accounted for by the occurrence counts of patterns in each frequency range. Patterns were counted after converting the extracted reason text strings to lowercase to avoid treating capitalization variants as distinct patterns.

Classification Performance

For discontinuation reasons and initiator (who stopped), predictions from both the local LLM and GPT-4o were compared against annotations from the first medical expert. Table 4 summarizes the performance for both drug groups. One case in antidiabetic reason classification by GPT-4o in the validation set resulted in an error, likely due to a profanity filter, and was considered misclassified. Confusion matrices illustrating true vs predicted labels for discontinuation reason (both models vs the first medical expert) and for discontinuation initiator (local LLM vs the first medical expert) are provided in Multimedia Appendix 1.

Table 4. Performance of the local large language model (Llama 3.1-70B; Meta AI) and GPT-4o (OpenAI) in classifying discontinuation reasons and initiators based on validation sets (n=100). Accuracy is equivalent to the micro–F1-score.

Drug group, task, and model	Accuracy (95% CI)	Weighted F_1-score (95% CI)	Cohen κ (95% CI)
Antidiabetics (A10)			
Reason			
Llama 3.1-70B	0.81 (0.72-0.88)	0.82 (0.74-0.89)	0.70 (0.58-0.81)
GPT-4o	0.80 (0.71-0.87)	0.84 (0.76-0.90)	0.70 (0.58-0.80)
Initiator			
Llama 3.1-70B	0.80 (0.71-0.87)	0.78 (0.68-0.86)	0.69 (0.58-0.80)
GPT-4o	—[a]	—	—
Statins (C10)			
Reason			
Llama 3.1-70B	0.81 (0.72-0.88)	0.82 (0.74-0.89)	0.62 (0.48-0.75)
GPT-4o	0.78 (0.69-0.85)	0.81 (0.73-0.88)	0.59 (0.46-0.72)
Initiator			
Llama 3.1-70B	0.67 (0.57-0.75)	0.64 (0.54-0.74)	0.45 (0.31-0.59)
GPT-4o	—	—	—

[a]GPT-4o was not evaluated for this task as this would have required submitting longer extracted discontinuation phrases to an external model, which was not performed for privacy reasons.

To contextualize model performance, interannotator agreement between 2 medical experts was assessed for the classification of discontinuation reason and initiator. Agreement was substantial for antidiabetics (Cohen κ=0.71 for reason and 0.69 for initiator) and statins (Cohen κ=0.64 for reason and 0.65 for initiator). These values indicate moderate to substantial consistency between human annotators and provide an upper bound for achievable performance. Model agreement with both

XSL•FO
RenderX

medical experts was comparable to interannotator agreement across most tasks, suggesting that model predictions approach human-level reliability for these classification categories.

Corresponding validation results using annotations from the second medical expert are reported in Multimedia Appendix 1.

Classifying Discontinuations

Discontinuation Reasons

Both GPT-4o and the local Llama model were used to classify the extracted reasons for treatment discontinuation (Figure 2). Across most categories, the 2 models produced broadly similar distributions. The main differences were that GPT-4o assigned more cases to *contraindication* and fewer to *other* in the statin group.

Figure 2. Results of classifying extracted reasons using GPT-4o (OpenAI) and the local Llama 3.1-70B model (Meta AI). For antidiabetics (A10), 12 GPT-4o outputs resulted in an error and were excluded from the figure to reflect unassisted model performance (effective n=613).

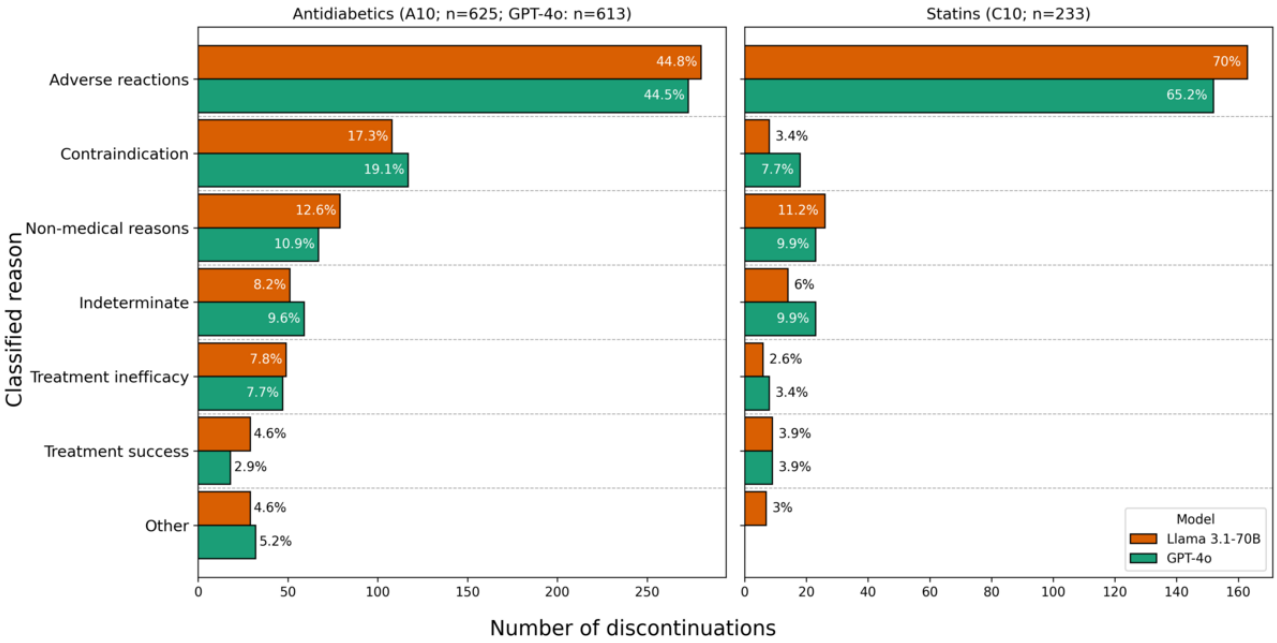

When comparing discontinuations of antidiabetic drugs and statins, *adverse reactions* were by far the most common reason in both groups. For statins, using the results from the local LLM, *adverse reactions* accounted for 70% (163/233) of all discontinuations, whereas for antidiabetics, they represented 44.8% (280/625). The lower proportion for antidiabetics can largely be explained by higher frequencies of *contraindication* and *treatment inefficacy* compared with statins.

Instances in which the local LLM classified discontinuation reasons as *adverse reactions* were examined across the full extracted dataset. The corresponding reasons were then grouped into clinically meaningful categories by a medical expert based on domain knowledge without use of a predefined coding schema or standardized taxonomy.

Among antidiabetic drug discontinuations attributed to adverse reactions (280/625, 44.8%), gastrointestinal symptoms were most frequent, accounting for 40.7% (114/280), followed by general malaise at 15% (42/280) and hypoglycemia at 12.9% (36/280). In contrast, among statin-related adverse reactions (163/233, 70%), the leading causes were muscle-related concerns at 32.5% (53/163), general malaise at 20.9% (34/163), and abnormal liver function at 6.7% (11/163).

Using results from the local Llama model, we stratified discontinuation texts based on whether they came from patients who switched to another drug within the same class or discontinued without switching. The results can be observed in Figure 3. Corresponding analyses using shorter gap definitions are provided in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 3. Distribution of discontinuation reasons in patient texts stratified by whether the text came from a patient who switched to another drug within the same class or discontinued entirely using results from the local Llama model.

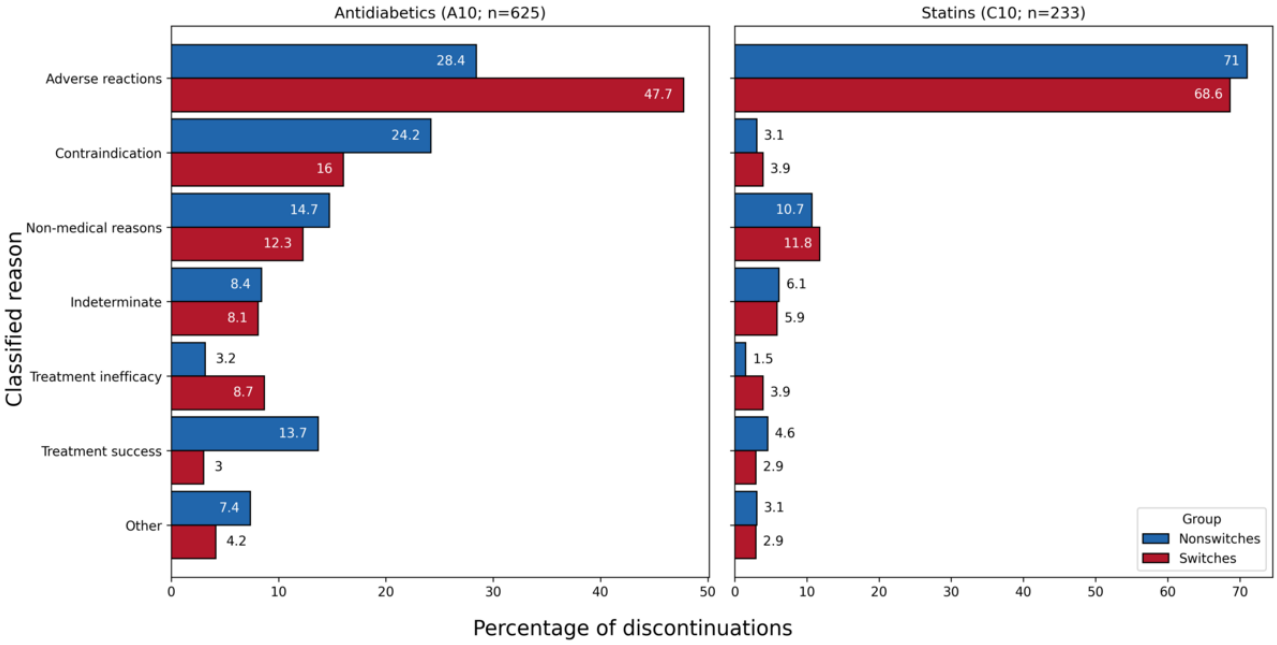

For statins, the distribution of reasons was largely similar between the 2 groups. However, for antidiabetics, differences were apparent: *adverse reactions* and *treatment inefficacy* were more common in texts from switchers, whereas *contraindication* and *treatment success* were more frequent in texts from nonswitchers.

Discontinuation Initiator

The local Llama model was used to classify whether the patient or the physician initiated discontinuation. For this task, extracted discontinuation phrases (rather than reasons) were analyzed as they contained the necessary contextual cues. Although both reasons and phrases were derived from the same clinical notes, they differed substantially in length and content. Discontinuation reasons were short (mean 26.8, SD 18.7 characters) and typically consisted of brief, abstract clinical labels (eg, adverse effect and lack of efficacy), whereas discontinuation phrases were considerably longer (mean 77.7, SD 44.4 characters) and frequently included narrative context, references to individuals (eg, patient or clinician actions), and additional clinical details. The results are shown in Figure 4.

Figure 4. Results of classifying who initiated the discontinuation based on the extracted phrases using the local Llama model.

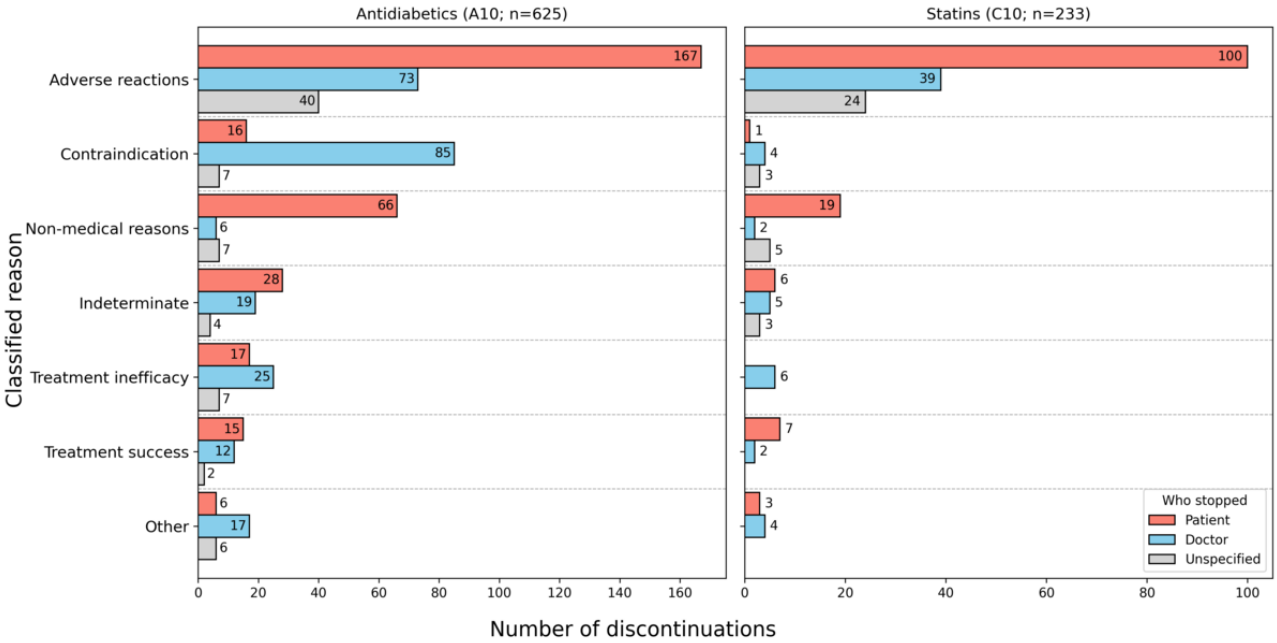

Across both drug groups, *adverse reactions* were most often associated with patient-initiated discontinuation. In contrast, *contraindications* were primarily identified as physician initiated. For *nonmedical reasons*, the decision was again more commonly attributed to patients. An interesting difference emerged in the category of *treatment inefficacy*: for

XSL•FO
RenderX

antidiabetics, patient-initiated discontinuations were noticeably more frequent than for statins.

Discussion

Principal Results

In this study, we showed that LLMs can be used to systematically extract and classify reasons for drug discontinuation from Estonian unstructured clinical narratives. By combining structured prescription data with free-text anamneses, we ensured that discontinuations were real events and that narrative mentions reflected meaningful reasons rather than speculation. This step also allowed us to substantially reduce the number of documents to annotate, focusing on only those we knew were relevant (51,330/1,200,000, 4.3% of the documents), thereby increasing efficiency and minimizing unnecessary review.

Our results show that the local model (Llama 3.1-70B; Meta AI) achieved high precision in extracting the discontinuations. Although recall could not be directly estimated for the final drug-specific extraction due to the absence of observed false negatives in the validation sample, the secondary analysis at the level of reason extraction suggests that the model captures most discontinuation reasons, with only a small number of missed cases observed. For classifying the reason for discontinuation, both the local model and GPT-4o achieved strong results. In the case of classifying the initiator of the discontinuation, the local model performed slightly worse than for reason classification. A possible explanation is that, because the documents are written from the perspective of the physician, it can be challenging to identify the subject of the statement from the context. Moreover, a likely contributing factor is the linguistic and stylistic characteristics of Estonian clinical documentation, which frequently uses implicit or agentless constructions that do not explicitly state who performed the action. For example, expressions such as "*statiinid - lõpetatud*" ("statins - stopped") describe the discontinuation event without specifying whether the decision was made by the physician or the patient. Error analysis showed that these agentless cases accounted for a disproportionate share of misclassifications in the initiator classification task.

The most frequent discontinuation reason was *adverse reactions*, which accounted for a much larger share among statin users. In contrast, for antidiabetic drugs, other drivers such as *treatment inefficacy* and *contraindications* played a more prominent role, likely reflecting challenges in glycemic control and the need to manage comorbidities. Expert review of reported *adverse reactions* indicated that the most common reactions for statins were muscle related, whereas for antidiabetic drugs, they were primarily gastrointestinal symptoms, consistent with previously published literature [23,24].

The observed patterns of switching vs nonswitching provide insights into patient behavior. For antidiabetic drugs, patients were more likely to switch when experiencing *adverse reactions* or perceived *treatment inefficacy*, reflecting the direct and immediate effects on daily glycemic control. For statins, switches were less common, and reasons for switching closely mirrored complete discontinuations. Among antidiabetic nonswitchers, *treatment success* and *contraindications* were more frequent, suggesting that patients are more likely to stop entirely when they feel that the therapy is effective or not suitable, whereas they continue adjusting treatment when efficacy or tolerability is an issue.

Classification of who initiated discontinuation revealed mostly expected patterns, with patients more often stopping due to *adverse reactions* or *nonmedical reasons* and physicians more often initiating discontinuation for *contraindications*. Notably, for antidiabetics, *treatment inefficacy* was much more patient driven than for statins. Analysis of the texts showed that, in these instances, patients measured their blood sugar levels, perceived no effect from the medication, and decided to stop taking it.

Limitations

First, recall could only be estimated on a limited sample of negative cases. While extracted discontinuation reasons showed high precision, recall was assessed by manually reviewing 100 randomly selected anamneses per drug group in which no reason was detected. This negative validation step identified few missed reasons and provided approximate recall estimates, but it cannot exclude systematic errors such as consistently missing specific categories of discontinuation reasons. Larger negative validation samples would improve the robustness of recall estimation. The results primarily characterize precision and classification fidelity rather than the complete capture of all discontinuations.

Second, causal linkage between symptoms and discontinuation cannot be definitively established from free-text anamneses. Explicit statements such as "the patient discontinued drug X because drug Y caused side effect Z" were uncommon. To mitigate this, discontinuation was confirmed using prescription purchase data, and analyses were restricted to anamneses recorded within 12 months of the last purchase. Nevertheless, some extracted negative mentions may reflect coincidental symptoms rather than true causes of discontinuation.

Third, the number of extracted cases with documented reasons was relatively small, reflecting the rarity of explicit discontinuation reasoning in routine clinical text. This resulted in wide CIs and limited the ability to analyze rare adverse reactions or individual drugs.

Fourth, the validation of discontinuation reason classification for statins showed variability in human annotations, with GPT-4o achieving a substantially lower agreement with the first medical expert (Cohen κ=0.59) than with the second expert (Cohen κ=0.85). This indicates that statin discontinuation reason labels may be subject to interannotator variability, reflecting the inherent ambiguity of clinical text interpretation, which may affect estimates of model performance depending on the reference standard used.

Fifth, the subcategorization of adverse drug reactions into clinically meaningful groups was performed by a single medical expert based on domain knowledge rather than using a predefined standardized taxonomy. This may introduce a degree of subjectivity in the grouping of adverse reactions.

XSL•FO
RenderX

Sixth, although the overall pipeline was designed for scalable analysis, the harmonization of extracted drug names required manual review by a clinician to resolve nonstandard mentions and map them to standardized ATC codes. This step introduces a human bottleneck and currently limits full automation, which may constrain scalability for large-scale deployment without further development of automated normalization methods.

Future Directions

Future work could enhance this approach in several ways. Performance may improve through evaluation of alternative prompting strategies and newer language models trained on clinical text. The classification scheme could be refined to capture more detailed reasons for medication discontinuation, whereas linking extracted data with longitudinal clinical or genetic information may enable insights into individual susceptibility to adverse reactions. Additionally, applying the pipeline to larger, more diverse EHR datasets could extend its utility for population-level pharmacovigilance and hypothesis generation. Incorporating patient-reported data or other non-EHR sources may further capture discontinuation reasons that are inconsistently documented, such as cost concerns or personal preferences.

Conclusions

This study demonstrated that LLMs can be applied to Estonian EHRs to identify discontinuation phrases, extract underlying reasons, and classify causes of discontinuation, with lower performance in identifying the initiator of treatment cessation. Validation of 100 sampled cases per drug group showed that, for antidiabetics, precision was 0.93 for extracted discontinuation phrases and 0.95 for extracted reasons, whereas for statins, precision was 0.98 for phrases and 0.96 for reasons. Classification of discontinuation reasons achieved a weighted F_1-score above 0.80 for both antidiabetics and statins, whereas classification of the initiator achieved weighted F_1-scores of 0.78 for antidiabetics and 0.64 for statins. Our findings showed that *adverse reactions* dominated among statin users, whereas *contraindications* and *treatment inefficacy* were more frequent for antidiabetic drugs, with patients and physicians contributing to discontinuation in distinct and clinically consistent ways.

Together, these results highlight the feasibility of transforming this previously underused data source into a valuable asset for pharmacovigilance and routine adverse event signal screening. This methodology is primarily intended for retrospective, population-level surveillance using large real-world datasets that include substantial unstructured clinical text. It enables identification of patterns and potential drivers of medication discontinuation that cannot be detected from structured data. In addition, the approach can support hypothesis generation for adverse event signals or adherence studies. While not designed for real-time clinical decision support, the pipeline demonstrates how LLMs can extract valuable information from narrative low-resource–language EHR data for medication safety surveillance beyond traditionally accessible data sources.

Acknowledgments

Generative large language models (Llama 3.1-70B [Meta AI] and GPT-4o [OpenAI]) were used as part of the study methodology for data extraction and classification, as described in the Methods section. Separately, ChatGPT (OpenAI [25]) was used for language editing and grammatical revision of the manuscript text. ChatGPT was not used for creating scientific content or analysis beyond the described study methodology.

Data Availability

The medical texts analyzed during this study were used according to the Estonian Committee on Bioethics and Human Research approval (1.1-12/1279). These data are not publicly available because they may contain identifiable information about the participants. Access to the medical texts cannot be granted by the authors and is subject to approval by the relevant ethics committee and applicable data protection requirements. The code for the pipeline is publicly available in GitHub [26].

Funding

This research was carried out with the support of the Estonian Research Council (grant PRG1844) and the Optimal Treatment for Patients With Solid Tumours in Europe Through Artificial Intelligence program, which is funded through the Innovative Medicines Initiative 2 Joint Undertaking and is listed under grant agreement 101034347. The Innovative Medicines Initiative 2 receives support from the European Union's Horizon 2020 research and innovation program and the European Federation of Pharmaceutical Industries and Associations. This work was also supported by the Estonian Centre of Excellence in Artificial Intelligence funded by the European Union and the Estonian Ministry of Education and Research via project TEM-TA72 and by Nordic Innovation (grant 407-7003-P22026). This research was also cofunded by the European Union through the European Regional Development Fund (project 2021-2027.1.01.24-0444; TeamPerMed). This work was also funded by the European Union under the Horizon Europe Action Grant Programme agreement 101057639. However, the views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authorities can be held responsible for them.

Authors' Contributions

HŠ conceptualized the study, curated the data, performed formal analysis and validation, conducted the investigation, developed the methodology and software, administered the project, managed resources, created visualizations, prepared the original draft

XSL•FO
RenderX

of the manuscript, and contributed to manuscript review and editing. NU contributed to the conceptualization, formal analysis, investigation, and methodology of the study; validated the results; and participated in manuscript review and editing. MH performed formal analysis and validation, conducted the investigation, and participated in manuscript review and editing. SL conducted the investigation and supervised the project. MO curated the data, developed the methodology, and supervised the project. ST performed formal analysis, conducted the investigation, and supervised the project. SR curated the data, acquired funding, and administered the project. JV curated the data, acquired funding, administered the project, managed resources, and contributed to manuscript review and editing. RK curated the data, acquired funding, conducted the investigation, administered and supervised the project, and contributed to manuscript review and editing. All authors reviewed and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prompt templates, cohort comparisons, temporal analyses, classification confusion matrices, secondary validation results, and sensitivity analyses.
[DOCX File , 536 KB - jmir_v28i1e86183_app1.docx]

Multimedia Appendix 2
Calculations of precision and recall.
[DOCX File , 27 KB - jmir_v28i1e86183_app2.docx]

References

1. Achterbosch M, Aksoy N, Obeng GD, Ameyaw D, Ágh T, van Boven JF. Clinical and economic consequences of medication nonadherence: a review of systematic reviews. Front Pharmacol 2025 Jun 25;16:1570359 [FREE Full text] [doi: 10.3389/fphar.2025.1570359] [Medline: 40635744]
2. Mooses K, Oja M, Malk M, Loorents H, Pajusalu M, Umov N, et al. Systematic evaluation of medication adherence determinants across 137 ingredients on population-level real-world health data. medRxiv. Preprint posted online on February 12, 2025 [FREE Full text] [doi: 10.1101/2025.02.10.25319884]
3. Kardas P, Lewek P, Matyjaszczyk M. Determinants of patient adherence: a review of systematic reviews. Front Pharmacol 2013 Jul 25;4:91 [FREE Full text] [doi: 10.3389/fphar.2013.00091] [Medline: 23898295]
4. Walsh CA, Cahir C, Tecklenborg S, Byrne C, Culbertson MA, Bennett KE. The association between medication non-adherence and adverse health outcomes in ageing populations: a systematic review and meta-analysis. Br J Clin Pharmacol 2019 Nov;85(11):2464-2478 [FREE Full text] [doi: 10.1111/bcp.14075] [Medline: 31486099]
5. Liu F, Pradhan R, Druhl E, Freund E, Liu W, Sauer BC, et al. Learning to detect and understand drug discontinuation events from clinical narratives. J Am Med Inform Assoc 2019 Oct 01;26(10):943-951 [FREE Full text] [doi: 10.1093/jamia/ocz048] [Medline: 31034028]
6. Sarraju A, Zammit A, Ngo S, Witting C, Hernandez-Boussard T, Rodriguez F. Identifying reasons for statin nonuse in patients with diabetes using deep learning of electronic health records. J Am Heart Assoc 2023 Apr 04;12(7):e028120 [FREE Full text] [doi: 10.1161/JAHA.122.028120] [Medline: 36974740]
7. Trevena W, Zhong X, Alvarado M, Semenov A, Oktay A, Devlin D, et al. Using large language models to detect and understand drug discontinuation events in web-based forums: development and validation study. J Med Internet Res 2025 Jan 30;27:e54601 [FREE Full text] [doi: 10.2196/54601] [Medline: 39883487]
8. Shao C, Snyder D, Li C, Gu B, Ngan K, Yang CT, et al. Scalable medication extraction and discontinuation identification from electronic health records using large language models. J Clin Epidemiol 2026 Jan;189:112049 [FREE Full text] [doi: 10.1016/j.jclinepi.2025.112049] [Medline: 41232578]
9. Kopacheva E, Henriksson A, Dalianis H, Hammar T, Lincke A. Identifying adverse drug events in clinical text using fine-tuned clinical language models: machine learning study. JMIR Form Res 2025 Sep 11;9:e71949 [FREE Full text] [doi: 10.2196/71949] [Medline: 40934508]
10. Golder S, Xu D, O'Connor K, Wang Y, Batra M, Hernandez GG. Leveraging natural language processing and machine learning methods for adverse drug event detection in electronic health/medical records: a scoping review. Drug Saf 2025 Apr;48(4):321-337 [FREE Full text] [doi: 10.1007/s40264-024-01505-6] [Medline: 39786481]
11. Ntinopoulos V, Rodriguez Cetina Biefer H, Tudorache I, Papadopoulos N, Odavic D, Risteski P, et al. Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation. BMJ Health Care Inform 2025 Jan 19;32(1):e101139 [FREE Full text] [doi: 10.1136/bmjhci-2024-101139] [Medline: 39832824]
12. Carlisle MN, Pace WA, Liu AW, Krumm R, Cowan JE, Carroll PR, et al. Development and validation of a generative artificial intelligence-based pipeline for automated clinical data extraction from electronic health records: technical

XSL•FO
RenderX

implementation study. JMIR Bioinform Biotechnol 2026 Jan 06;7:e70708 [FREE Full text] [doi: 10.2196/70708] [Medline: 41494167]
13. Schipper A, Belgers P, O'Connor RD, van de Wouw L, Builtjes L, Bosma JS, et al. Large language model automated extraction of clinical signs and symptoms from emergency department reports for machine learning prediction models: development and validation study. JMIR Med Inform 2026 Apr 30;14:e81500 [FREE Full text] [doi: 10.2196/81500] [Medline: 42076868]
14. Sakai H, Lam SS. Large language models for health care text classification: systematic review. JMIR AI 2026 Feb 11;5:e79202 [FREE Full text] [doi: 10.2196/79202] [Medline: 41672471]
15. Oja M, Tamm S, Mooses K, Pajusalu M, Talvik HA, Ott A, et al. Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model: lessons learned. JAMIA Open 2023 Dec 5;6(4):ooad100 [FREE Full text] [doi: 10.1093/jamiaopen/ooad100] [Medline: 38058679]
16. Llama Team. The Llama 3 herd of models. arXiv. Preprint posted online on July 31, 2024 . [doi: 10.48550/arXiv.2407.21783]
17. OpenAI. GPT-4o System Card. arXiv. Preprint posted online on October 25, 2024 . [doi: 10.48550/arXiv.2410.21276]
18. Šuvalov H, Lepson M, Kukk V, Malk M, Ilves N, Kuulmets HA, et al. Using synthetic health care data to leverage large language models for named entity recognition: development and validation study. J Med Internet Res 2025 Mar 18;27:e66279 [FREE Full text] [doi: 10.2196/66279] [Medline: 40101227]
19. Pydantic validation: data validation using Python type hints. GitHub. URL: https://github.com/pydantic/pydantic [accessed 2026-05-27]
20. Adherence to long-term therapies: evidence for action. World Health Organization. 2003. URL: https://iris.who.int/items/bf8058c0-03b2-4b47-838f-5534849927fb [accessed 2026-05-27]
21. Golder S, Weissenbacher D, O'Connor K, Hennessy S, Gross R, Hernandez GG. Patient-reported reasons for switching or discontinuing statin therapy: a mixed methods study using social media. Drug Saf 2022 Sep;45(9):971-981 [FREE Full text] [doi: 10.1007/s40264-022-01212-0] [Medline: 35933649]
22. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr;20(1):37-46 [FREE Full text] [doi: 10.1177/001316446002000104]
23. Ruscica M, Ferri N, Banach M, Sirtori CR, Corsini A. Side effects of statins: from pathophysiology and epidemiology to diagnostic and therapeutic implications. Cardiovasc Res 2023 Jan 18;118(17):3288-3304 [FREE Full text] [doi: 10.1093/cvr/cvac020] [Medline: 35238338]
24. Shurrab NT, Arafa ES. Metformin: a review of its therapeutic efficacy and adverse effects. Obes Med 2020 Mar;17:100186 [FREE Full text] [doi: 10.1016/j.obmed.2020.100186]
25. ChatGPT. URL: https://chatgpt.com/ [accessed 2026-06-06]
26. Discontinuation extraction and classification. GitHub. URL: https://github.com/hendriksuvalov/discontinuation-extraction-and-classification [accessed 2026-05-27]

Abbreviations

ATC: Anatomical Therapeutic Chemical
EHR: electronic health record
LLM: large language model
NLP: natural language processing

Edited by A Coristine; submitted 20.Oct.2025; peer-reviewed by H Tibble, D Chrimes, S Matsuda, H Xu; comments to author 15.Jan.2026; revised version received 18.May.2026; accepted 18.May.2026; published 17.Jun.2026.

Please cite as:
Šuvalov H, Umov N, Malk M, Haug M, Laur S, Oja M, Tamm S, Reisberg S, Vilo J, Kolde R
Extracting and Classifying Drug Discontinuations From Estonian Electronic Health Records: Development and Validation Study
J Med Internet Res 2026;28:e86183
URL: https://www.jmir.org/2026/1/e86183
doi:10.2196/86183
PMID:

©Hendrik Šuvalov, Nikita Umov, Maria Malk, Markus Haug, Sven Laur, Marek Oja, Sirli Tamm, Sulev Reisberg, Jaak Vilo, Raivo Kolde. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Investigating the Effect of Hospital Infection Control Informatization on Optimizing Microbiological Specimen Submission Before Antibiotic Therapy: Failure Mode and Effects Analysis

Jianxiong Wu[1], PhD; Weijiang Zhan[2], BEng; Jun Hua[2], BEng; Qinling Ge[1], BS; Yuexian Zhu[1], BS; Huixian Yu[3], BS; Min Zhao[4], BPharm; Xiaoyan Zhan[5], BS; Bingwei Zhu[6], BS; Tongqi Xiang[7], BS; Longxi Lu[8*], PhD; Tieying Dai[1*], PhD

[1]Department of Public Health, The First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), 54 Youdian Road, Xihu District, Hangzhou, Zhejiang, China

[2]Department of Computer Center, the First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), Hangzhou, Zhejiang, China

[3]Department of Nursing, the First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), Hangzhou, Zhejiang, China

[4]Department of Pharmacy, the First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), Hangzhou, Zhejiang, China

[5]Discipline Inspection Commission Office, the First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), Hangzhou, Zhejiang, China

[6]Microbiology Laboratory, the First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), Hangzhou, Zhejiang, China

[7]Logistics Support Department, the First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), Hangzhou, Zhejiang, China

[8]Department of Environmental Health, Zhejiang Provincial Center for Disease Control and Prevention, Hangzhou, Zhejiang, China

[*]these authors contributed equally

Corresponding Author:
Tieying Dai, PhD
Department of Public Health, The First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), 54 Youdian Road, Xihu District, Hangzhou, Zhejiang, China

Abstract

Background: Antimicrobial resistance (AMR) poses a critical global health threat, with inappropriate antibiotic use being a major driver. Timely microbiological specimen submission before initiating antibiotic therapy is a cornerstone of antimicrobial stewardship (AMS), enabling pathogen-directed therapy and reducing unnecessary broad-spectrum exposure. However, suboptimal compliance remains common due to workflow interruptions, technological barriers, and behavioral factors. Failure Mode and Effects Analysis (FMEA), a proactive risk-assessment method widely used in health care quality improvement, provides a systematic framework to identify process vulnerabilities and prioritize corrective actions. Despite its increasing application, few studies have integrated FMEA with hospital informatization to optimize microbiological specimen submission workflows in routine AMS practice.

Objective: This study aimed to systematically identify workflow risks affecting preantibiotic microbiological specimen submission and to design, implement, and evaluate informatization-enabled interventions using an FMEA-based framework.

Methods: FMEA was conducted at a tertiary hospital in China. A multidisciplinary team identified potential failure modes across 4 domains: health information systems, personnel, administration, and external support. Risk Priority Numbers (RPNs) and Action Priority (AP) indices were calculated for each failure mode. Targeted interventions were implemented, including dual-verification barcode scanning, artificial intelligence-driven clinical decision support alerts, EHR-integrated training modules, and automated compliance dashboards. Pre- and postintervention specimen submission rates (January 2024-December 2024) were analyzed using the Mann-Kendall trend test.

Results: The top 5 failure modes included PDA barcode scanning failures (RPN=175), inadequate clinical decision support (RPN=140), insufficient clinician awareness (RPN=56), suboptimal oversight mechanisms, and patient-related barriers. Postintervention, significant upward trends were observed in overall specimen submission rates ($P<.001$), with similar improvements for restricted-use ($P<.001$) and special-use antibiotics ($P<.001$).

XSL•FO
RenderX

Conclusions: FMEA-based risk management combined with hospital informatization effectively optimized specimen submission workflows. Real-time decision support, process standardization, and interdisciplinary collaboration significantly enhanced compliance. Future research should evaluate long-term impacts on AMR reduction and diagnostic integration. Integrating FMEA with hospital informatization effectively strengthened the microbiological specimen submission process. Digital decision support, standardized workflows, and real-time monitoring substantially improved compliance with preantibiotic specimen submission. This approach provides an actionable model for data-driven AMS enhancement. Future studies should assess scalability, cost-effectiveness, and impacts on downstream AMR outcomes.

(J Med Internet Res 2026;28:e78118) doi:10.2196/78118

KEYWORDS

microbiological specimen submission; failure mode and effects analysis; antibiotic stewardship; hospital informatization; risk management

Introduction

Antimicrobial resistance (AMR) poses a critical threat to global public health and health care systems [1,2]. It was estimated that AMR was associated with 4.71 million deaths worldwide in 2021 [3]. Inappropriate use of antibiotics is a primary driver of AMR [4]. Accurate microbiological pathogen detection is essential for guiding rational antibiotic therapy [5]. On one hand, identifying the causative pathogens and their antimicrobial susceptibility profiles can reduce unnecessary exposure to broad-spectrum antibiotics. On the other hand, microbiological test results support clinicians in refining therapeutic strategies, thereby shortening the infection control cycle. However, multiple studies have reported low rates of microbiological specimen submission prior to initiating antibiotic therapy. A multicenter study in Europe found that only 50% of patients had specimens submitted before antibiotic administration [6]. Similarly, a study conducted in Beijing, China, showed that among patients receiving carbapenem therapy, the preantibiotic microbiological submission rate was only 64.76% [7].

The accuracy and timeliness of pathogen detection rely heavily on standardized and efficient specimen submission workflows. Many countries have established tailored guidelines and policies [8]. Specimens should be obtained prior to initiating antibiotic therapy to ensure timely microbiological culture [9]. For example, China has included the microbiological specimen submission rate for high-priority antibiotics as a core quality indicator in *Healthcare-Associated Infection Management Quality Control Indicators (2024 Edition*) [10]. In clinical practice, multiple challenges hinder preantibiotic specimen submission, including insufficient awareness among clinicians, delays in specimen collection by nursing staff, and transport inefficiencies [11]. As the specimen collection and submission process involves multiple departments and procedural steps, its complexity has hindered improvement of submission rates [12]. Currently, there is a limited understanding of risk factors across the entire microbiological specimen submission pathway.

Failure mode and effects analysis (FMEA), as a risk management tool, has been extensively used in research on optimizing and refining hospital management processes [13,14]. The core principle of FMEA involves assembling an interdisciplinary team to systematically identify potential failure modes within a workflow, analyze their causes and consequences, and prioritize high-risk steps by calculating the Risk Priority Number (RPN), thereby guiding the development of targeted improvement interventions. For example, Yang et al [15] applied FMEA to hospital infection prevention in a stomatology department by mapping the clinical workflow, identifying potential failure modes through expert consensus, and ranking high-risk steps using RPN to inform targeted control strategies. Chiereghin et al [16] used FMEA to improve the quality of an organized colorectal cancer screening program. These studies highlight FMEA's value in systematically dissecting complex clinical processes and guiding resource allocation toward high-impact improvements. However, its application to microbiological specimen submission within antimicrobial stewardship programs remains limited, particularly when integrated with hospital digital infrastructure.

Digital health and information systems can support this process by embedding real-time alerts into electronic prescribing workflows, prompting clinicians to submit microbiological specimens before initiating antibiotic therapy. Integrated informatics tools, such as electronic health records (EHRs), computerized physician order entry, barcode-based specimen tracking systems, and clinical decision support systems (CDSS), enable end-to-end, closed-loop management of specimen collection, transport, and laboratory processing. However, digital health and information systems may also impose additional workload on health care professionals [17]. Therefore, developing feasible implementation strategies and continuously evaluating the real-world performance of these technologies are essential to ensure sustainable optimization.

Therefore, this study aimed to identify the failure modes contributing to low microbiological specimen submission rates in a tertiary hospital of Chinese Medicine using FMEA and to develop and evaluate targeted interventions to improve specimen submission practices.

Methods

Study Setting and Design

This study was conducted at Zhejiang Provincial Hospital of Chinese Medicine, a general tertiary hospital located in Hangzhou, Zhejiang Province, China. Applying FMEA to assess risks in preantibiotic therapy microbiological specimen submission, formulate targeted interventions, and evaluate their effectiveness. A complete checklist is available in Checklist 1. The implementation steps are as follows (Figure 1):

XSL•FO
RenderX

Figure 1. Schematic figure of Failure Mode and Effects Analysis for microbiological specimen submission. IPC: infection prevention and control.

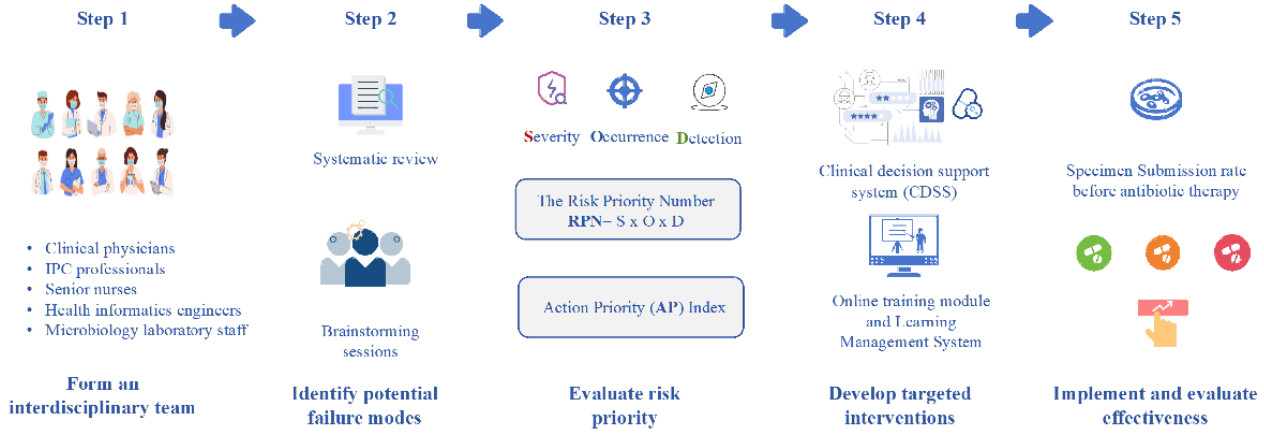

Step 1: Form an Interdisciplinary Team

A multidisciplinary expert team was established to ensure comprehensive identification of workflow risks across clinical, laboratory, managerial, and informatics domains. The team consisted of 9 members, including clinical physicians, Disease Prevention and Control Experts, infection prevention and control (IPC) professionals, microbiology laboratory staff, senior nurses, and health informatics engineers. Team members were selected based on predefined criteria: (1) at least 5 years of professional experience in their respective fields; (2) direct involvement in antimicrobial stewardship, specimen management, or infection surveillance activities; (3) familiarity with routine workflows and known bottlenecks within the hospital; and (4) willingness to participate in structured meetings, independent FMEA scoring, and consensus discussions. This interdisciplinary composition ensured that all key steps of the specimen submission process, from clinical decision-making to specimen collection, laboratory processing, and information system operation, were adequately represented, providing a balanced and comprehensive perspective for risk analysis.

Step 2: Identify Potential Failure Modes

We first conducted a systematic review of the literature to identify common influencing factors in the microbiological specimen submission workflow (Multimedia Appendix 1). Commonly reported factors affecting microbiological specimen submission were extracted and categorized into four predefined domains: (1) health information systems factors, (2) personnel-related factors, (3) administrative and management factors, and (4) external support factors. This framework was used to guide subsequent discussions. Building on this framework, the multidisciplinary expert panel participated in 2 rounds of structured brainstorming sessions. During each session, experts were instructed to identify specific, concrete scenarios within each domain that could plausibly lead to specimen submission failure or delay. To enhance consistency, failure modes were defined as observable deviations from the expected workflow (eg, missed test ordering, delayed collection, incomplete data capture), rather than abstract causes or outcomes. All proposed failure modes were collated, merged, and refined by the research team. Redundant items were consolidated, and ambiguous descriptions were clarified through iterative discussion with panel members. Items were retained only if consensus was reached that the scenario was (1) operationally distinct, (2) locally relevant to routine practice in the study hospital, and (3) theoretically supported by either literature evidence or expert experience. A final set of 30 locally relevant failure modes was generated. These items formed the basis for subsequent FMEA scoring and prioritization.

Step 3: Evaluate Risk Priority

Before formal scoring, all team members received standardized training on the FMEA scoring criteria provided by the research team. Each member independently evaluated the Severity (S), Occurrence (O), and Detection (D) of each potential failure mode. After all 9 members completed their independent assessments, a consensus meeting was convened to discuss discrepancies and clarify scoring decisions. Final S, O, and D scores for each failure mode were determined through group consensus. The RPN was calculated using the formula: RPN=S×O×D.

The scoring range for S, O, and D was 1 - 10. Conventionally, corrective actions are mandated when the RPN exceeds 125. To address limitations of the RPN method, we introduced the Action Priority (AP) index. The determination of AP followed the standard methodology outlined in the Automotive Industry Action Group–Verband der Automobilindustrie FMEA guidelines. Unlike the RPN, the AP level is not calculated through a mathematical formula; instead, it is assigned by consulting the guideline's AP lookup table, which specifies the corresponding action level based on different combinations of Severity (S), Occurrence (O), and Detection (D). As shown in Multimedia Appendix 2, AP levels are categorized as high, medium, or low. High AP indicates that corrective actions should be taken with priority, medium AP suggests that action is recommended but can be scheduled, and low AP indicates that existing controls are acceptable and require only routine monitoring. High-priority failure modes identified by the FMEA (based on both RPN and AP indices).

Step 4: Develop Targeted Interventions

For each mode, the multidisciplinary team collaboratively designed corresponding digital health solutions. All digital interventions underwent iterative refinement through pilot testing in selected departments. Feedback from clinicians, nurses, infection control practitioners, and IT personnel was

XSL•FO
RenderX

collected to optimize alert thresholds, interface design, workflow integration points, and data visualization formats.

The final intervention package consisted of three interrelated components:

1. An artificial intelligence-driven (DeepSeek-R1) CDSS was developed and integrated into the hospital's EHR and computerized physician order entry system to promote microbiological specimen submission before antibiotic therapy. When a clinician initiates an order for therapeutic antibiotics, particularly restricted-use or special-use agents, the system automatically evaluates whether a corresponding microbiological specimen has been ordered or collected within a predefined time window.
2. Online training module and learning management system (LMS): The module was jointly created by the Infection Prevention and Control Department, the Microbiology Laboratory, and the IT Department. The training curriculum covered standardized procedures for microbiological specimen collection, requirements for specimen transport, electronic order entry workflows, and regulations related to preantibiotic specimen submission.
3. Supplementary digital process-enhancement measures: Additional supportive measures included enhanced barcode-based specimen tracking and monitoring dashboards with feedback mechanisms. Barcode-based tracking strengthened specimen traceability during collection and transport, reducing data loss caused by scanning failures or workflow deviations. Dashboards aggregated compliance data by department and antibiotic category, generating periodic feedback reports for clinical units and infection control teams to facilitate continuous quality improvement.

Step 5: Implement and Evaluate Effectiveness

After developing the intervention measures, we proceeded with the implementation phase following the standard quality improvement process. The preparatory stages of FMEA, including workflow mapping, identification and scoring of failure modes, prioritization, and intervention design, were completed between January 1 and March 31, 2024. Beginning on April 1, 2024, the FMEA-informed interventions were formally implemented. Physicians, nurses, and support staff received training on the updated specimen submission workflow and information system functions through the online learning platform. The effectiveness evaluation covered the period from January to December 2024. Improvements were assessed by comparing pre- and postintervention microbiological specimen submission rates.

Three core outcome indices were evaluated:

1. Specimen submission rate before antibiotic therapy of inpatients (%)=number of inpatients with pathogen submission before antibiotic therapy/number of inpatients with therapeutic application of antibiotics during the same period *100%.
2. Specimen submission rate before restricted-use antibiotic therapy of inpatients (%)=number of inpatients with pathogen submission before restricted-use antibiotic therapy/number of inpatients with therapeutic application of restricted-use antibiotics during the same period *100%.
3. Specimen submission rate before special-use antibiotic therapy of inpatients (%)=number of inpatients with pathogen submission before special-use antibiotic therapy/number of inpatients with therapeutic application of special-use antibiotics during the same period *100%.

Data Collection

Specimen submission data from January 1, 2024, to December 31, 2024, was extracted from the Hospital Infection Surveillance System (Xinglin System). The dataset included antibiotic order time, drug category (restricted-use, special-use), and timing of administration relative to specimen submission. To ensure data quality and consistency, only complete records with clearly documented timestamps for both specimen submission and antibiotic prescribing were included in the analysis. Duplicate entries, canceled orders, and specimens rejected by the laboratory due to quality issues were excluded.

Data Analysis

Statistical analyses were performed using SPSS (version 27.0; IBM Corp). Categorical variables were reported as counts (percentages). The Mann-Kendall trend test was used to examine the trend of submission rate over time. When the z score was positive, it meant that the composition ratio of each indicator showed an upward trend. When the z score was negative, it meant that the composition ratio of each indicator presented a downward trend. Statistical significance was set at 2-tailed $P<.05$.

Ethical Considerations

This study was approved by the Ethics Committee of the First Affiliated Hospital of Zhejiang Chinese Medical University (2025-KLS-038-01). As this study used retrospective data, the ethics committee waived the requirement for participants to provide written informed consent. The research data does not involve individual patient information. No compensation was provided to participants as this study was a secondary analysis of preexisting data with no new patient recruitment or intervention. The research was conducted in accordance with the Declaration of Helsinki.

Results

Hazard Analysis of Specimen Submission

The top five failure modes limiting preantibiotic specimen submission were (1) PDA barcode scanning failures, (2) inadequate clinical decision support alerts, (3) insufficient clinician awareness of testing protocols, (4) suboptimal oversight mechanisms, and (5) patient-related health constraints (Table 1).

XSL•FO
RenderX

Table . Failure modes summary for Failure Mode and Effects Analysis.

Potential failure modes	RPN[a]	AP[b]
Health information system		
Interoperability gaps exist between the inpatient electronic medical records system and the health care–associated infection surveillance system, resulting in inaccurate and delayed data sharing and synchronization between the 2 platforms.	146	Moderate
Failure of PDA barcode scanning or over-rapid scanning operations led to unrecorded specimen collection data in the system, compromising preanalytical traceability.	168	High
Incomplete capture or delayed updating of pathogen detection-related data compromised the integrity of microbiological specimen submission records.	158	High
The absence of proactive alerts for pathogen detection failed to prompt clinicians to initiate microbiological testing before antimicrobial therapy, resulting in prescription restrictions due to noncompliance with stewardship protocols.	70	Low
The hospital information system failed to accurately distinguish the purpose of antimicrobial use (therapeutic vs prophylactic) during medication documentation, undermining the precision of stewardship analytics.	35	Low
The calculation methodology for pathogen detection-related metrics in the information system was not aligned with national standards, resulting in inaccurate statistical outcomes.	84	Low
Hospital personnel		
Delayed communication and coordination among IPC[c] professionals due to ineffective communication channels or the absence of established mechanisms hindered timely interventions.	48	Low
Incomplete data verification by IPC professionals was attributed to the overwhelming volume of data and the inherent limitations of manual review processes, leading to frequent oversights.	32	Low
Misinterpretation of infection protection and control standards by IPC professionals occurred due to frequent updates and inadequate, timely training, resulting in noncompliant practices.	48	Low
Health care providers demonstrated insufficient awareness of pathogen detection protocols, with inadequate understanding of its clinical significance, leading to missed testing opportunities in cases where microbiological analysis was indicated.	126	High
Delays in specimen submission by health care providers were attributed to heavy workloads leading to oversight or missed submissions, as well as suboptimal time window settings that conflicted with clinical workflows.	75	Moderate
Noncompliant specimen collection practices by health care providers, stemming from insufficient familiarity with collection protocols and deviations from aseptic techniques, resulted in specimen contamination or invalid samples.	50	Low

XSL•FO
RenderX

Potential failure modes	RPN[a]	AP[b]
Noncompliant specimen transport practices by health care providers, due to insufficient knowledge of transport protocols and inadequate protective measures during transit, compromised sample integrity.	80	Low
Health care providers mistakenly prescribed therapeutic antibiotics instead of prophylactic regimens, leading to inappropriate antimicrobial use.	80	Low
Ineffective communication between nurses and physicians, due to the absence of established collaboration mechanisms, resulted in suboptimal specimen collection practices, as nurses lacked clear guidance on proper techniques and timing, leading to noncompliant samples.	96	Low
IT personnel failed to facilitate tripartite coordination, lacking effective communication and collaboration mechanisms.	48	Low
IT personnel lacked sufficient knowledge of infection surveillance data extraction guidelines: the complexity of guidelines for building infection surveillance information systems made them difficult to understand.	48	Low
Insufficient expertise among hospital microbiology laboratory personnel and inadequate detection technology levels.	40	Low
Health care administration		
Inadequate regulatory mechanisms for pathogen detection specimen submission: issues include unclear regulatory responsibilities, limited regulatory methods, and insufficient regulatory enforcement.	120	Low
Inadequate performance evaluation mechanisms for pathogen detection specimen submission: evaluation metrics are not linked to health care providers' performance or career advancement, lacking incentives.	48	Low
Inadequate feedback mechanisms for pathogen detection specimen submission: test results are not timely or accurately communicated to health care providers.	48	Low
Inadequate communication and collaboration mechanisms among hospital departments: information flow and interdepartmental coordination are suboptimal.	96	Low
Inadequate training content on pathogen detection specimen submission: issues include outdated content, lack of relevance, or limited training methods.	84	Low
Lack of standardized specimen collection and submission processes: issues include unclear procedures, inconsistent operational standards, or poor process execution.	96	Low
External support		
Some patients may not understand or be willing to cooperate with pathogen detection specimen submission, leading to difficulties in sample collection.	49	Low

XSL•FO
RenderX

Potential failure modes	RPN[a]	AP[b]
Patients' poor health status or physical condition may hinder cooperation, affecting the smooth progress of sample collection (eg, difficulty in expectorating sputum leading to challenges in sputum sample collection).	49	Low
Patients' health status or physical condition may affect the smooth progress of sample collection.	192	Moderate
Limited space in the microbiology laboratory may reduce the efficiency of sample processing. During peak hours, sample backlogs may occur, prolonging testing times.	80	Low
Convenience of specimen collection: health care providers may face space and equipment limitations when collecting samples in patient rooms. For example, the lack of appropriate sputum induction devices may compromise sample quality, reducing motivation for submission.	64	Low
Inadequate cleaning and disinfection of patient rooms: a high pathogen load in the room environment may lead to cross-contamination, interfering with the diagnosis of the patient's own infectious pathogens.	128	Moderate

[a]RPN: Risk Priority Number.

[b]AP: action priority.

[c]IPC: infection prevention and control.

Develop Targeted Interventions

For the potentially high-risk failure modes, recommended interventions were identified with a view to reducing their occurrence or improving their detection (Table 2). An LMS-based online training was implemented. All personnel involved in specimen collection or antibiotic order entry were required to complete the module and pass a competency assessment. A score of at least 80% was considered a passing result. The LMS automatically recorded training completion and assessment outcomes; individuals who did not complete the module or did not achieve a passing score were required to undergo retraining before being allowed to independently perform the related tasks. A real-time CDSS was embedded into the antibiotic prescribing process.

XSL•FO
RenderX

Table . Interventions recommended to improve the specimen submission rate.

High-risk failure modes	Targeted interventions
PDA barcode scanning failures	• Implement dual-verification protocols combining manual entry with barcode scanning, accompanied by real-time error alerts in the mobile application.
Inadequate clinical decision support alerts	• Deploy AI[a]-driven decision support systems to automatically flag high-risk cases requiring preantimicrobial testing, with escalating alert tiers.
Insufficient clinician awareness of testing protocols	• Develop role-specific e-Learning modules with embedded competency assessments. • Integrate protocol reminders into electronic health record order entry workflows.
Suboptimal oversight mechanisms	• Establish automated dashboards tracking testing compliance, with weekly feedback reports to department heads. • Introduce “process champions” in each unit to monitor adherence and facilitate peer-to-peer learning.
Patient-related health constraints	• Create adaptive sampling protocols. • Train nursing staff in motivational interviewing techniques to improve patient engagement.

[a]AI: artificial intelligence.

When clinicians initiated orders for therapeutic antibiotics, particularly restricted-use or special-use agents, the CDSS automatically evaluated whether an appropriate microbiological specimen had been ordered. If noncompliance was detected, the system generated real-time alerts prompting clinicians to initiate specimen submission. This workflow enabled prospective intervention at the point of care, reducing missed submissions caused by time pressure or oversight. Supportive process-enhancement measures were integrated to reinforce execution and oversight. Enhanced PDA-based barcode scanning strengthened specimen traceability during collection and transport, reducing data loss related to scanning failures or rapid operations. In parallel, automated monitoring dashboards aggregated specimen submission compliance data by department and antibiotic category. Regular feedback reports were provided to clinical units and infection control teams, supporting targeted supervision and continuous workflow optimization.

Effect of FMEA Management on Specimen Submission

Following the implementation of FMEA management in April 2024, significant improvements were observed in the overall specimen submission rate (Figure 2), the submission rate for restricted antibiotics (Figure 3), and the submission rate for special-use antibiotics (Figure 4). The Mann-Kendall trend test revealed that all 3 categories of specimen submission rates exhibited a statistically significant upward trend ($P<.001$).

XSL•FO
RenderX

Figure 2. Changes in the specimen submission rate before antibiotic therapy among in-patients, 2024.

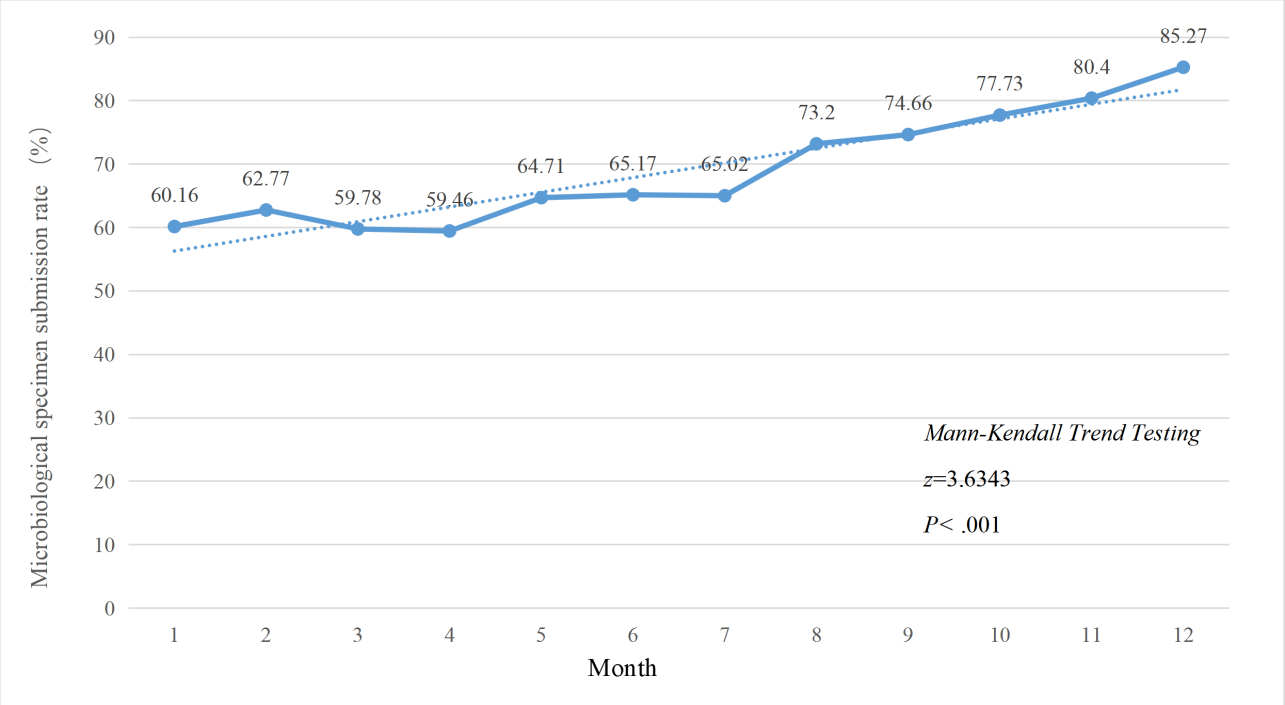

Figure 3. Changes in the specimen submission rate before restricted-use antibiotic therapy among in-patients, 2024.

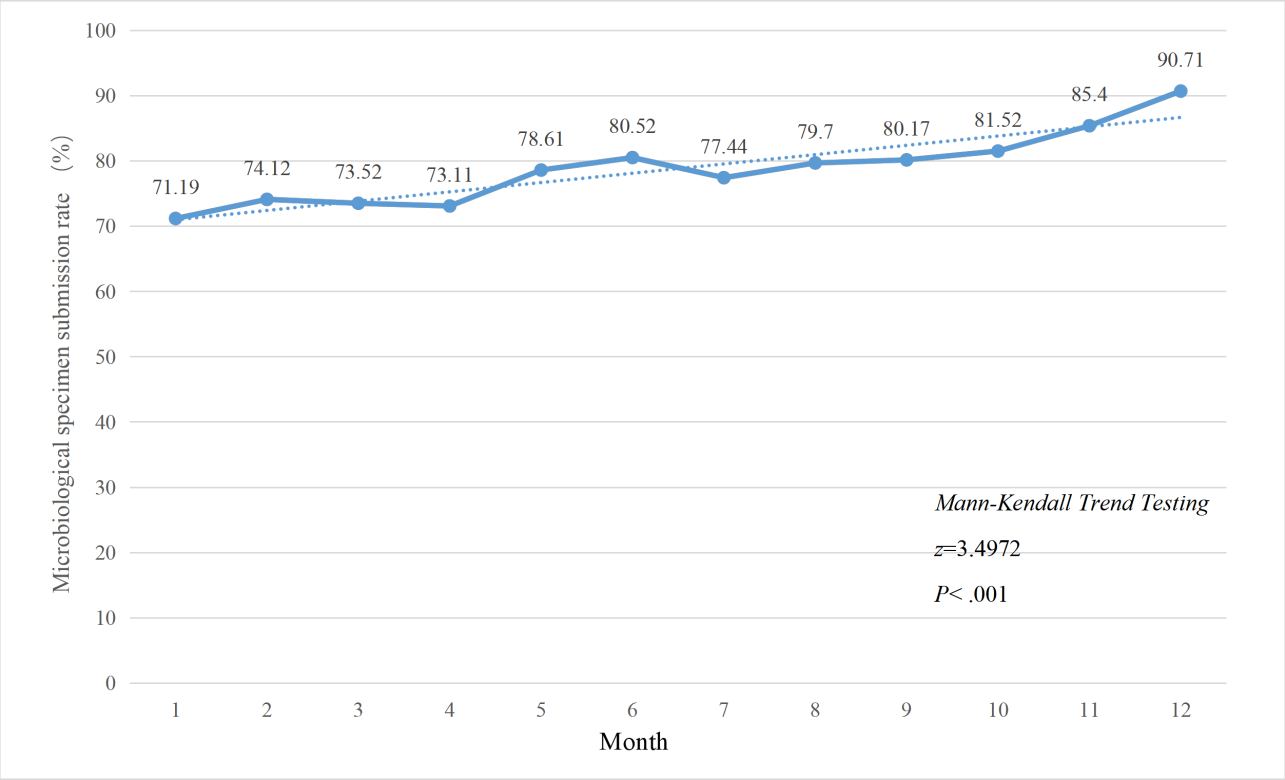

XSL•FO
RenderX

Figure 4. Changes in the specimen submission rate before special-use antibiotic therapy among in-patients, 2024.

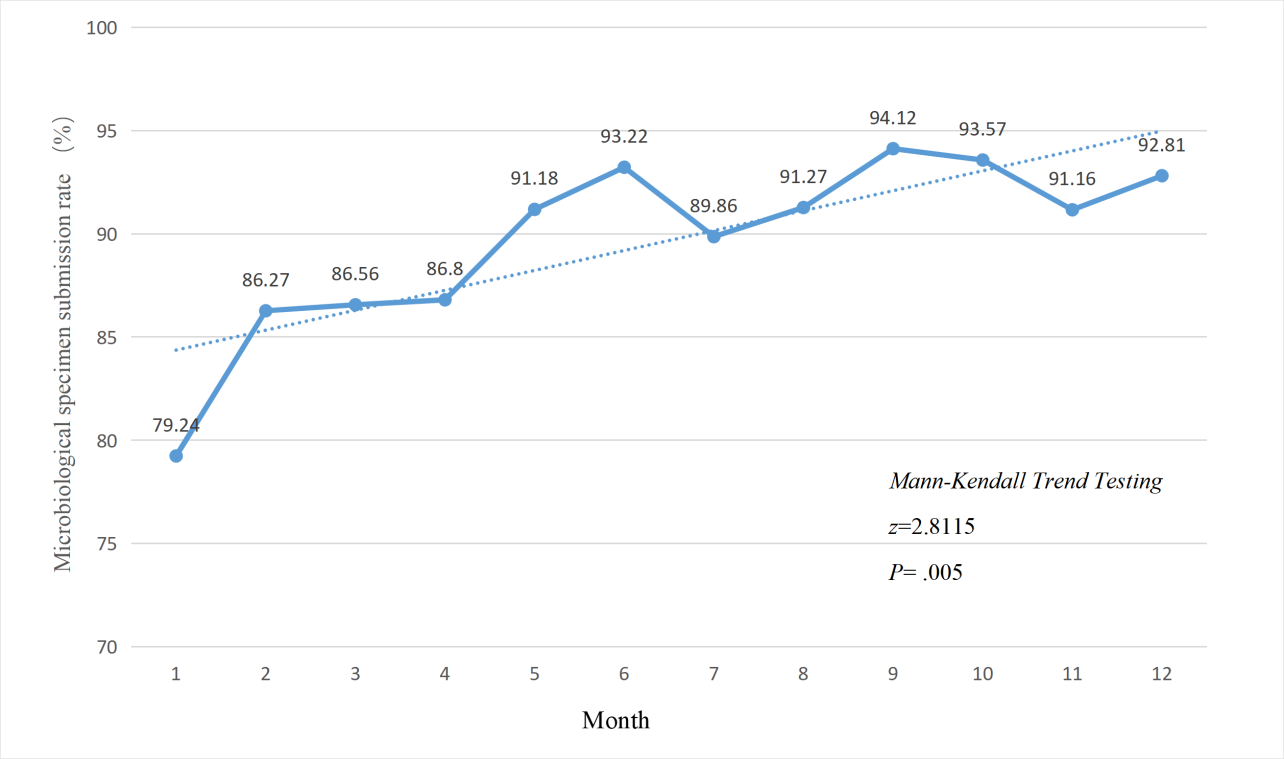

Discussion

Principal Findings

Microbiological specimen submission plays a pivotal role in hospital infection control and promoting rational antibiotic use, representing a complex systemic endeavor [18,19]. This study implemented the internationally recognized risk assessment tool—FMEA, to conduct a multidimensional evaluation of the entire etiological specimen submission workflow, spanning information systems, health care personnel, hospital administration, and external support. Through expert panel discussions, the research identified multiple failure modes and prioritized targeted intervention strategies, providing scientifically grounded decision-making support for improving the quality of specimen submission practices. This study provides a replicable pathway and methodological framework for improving microbiological specimen submission workflows, and may indirectly support hospital infection prevention efforts by enhancing antimicrobial stewardship and strengthening infection surveillance.

The advancement of intelligent health care information systems has emerged as a critical driver for improving specimen submission rates [20]. FMEA-based investigation identified PDA barcode scanning failures and insufficient CDSS alerts as high-priority intervention targets. Mechanistically, real-time alert mechanisms within information systems can trigger specimen collection reminders before antimicrobial administration, thereby minimizing workflow delays caused by human oversight and ensuring precise temporal control of submissions [21]. Concurrently, integrated multisource data platforms, by consolidating heterogeneous data from EHRs, laboratory results, and other sources, enable dynamic decision-making models that assist clinicians in accurately evaluating the necessity of testing indications and the specificity of required assays [22]. These findings reveal a dual-driven mechanism of information system optimization: process monitoring reduces operational deviations, while data integration enhances diagnostic decision-making efficacy.

This study also revealed that inadequate clinician awareness of specimen submission constitutes a critical bottleneck in improving submission quality. To address this, the research team implemented a dual-track capacity-building framework, “theoretical training+practical supervision,” alongside intelligent system upgrades: standardized operational training was used to reinforce protocol adherence, while interdepartmental supervision mechanisms established closed-loop quality improvement cycles. Postintervention evaluation demonstrated statistically significant improvements in specimen submission rates (P<.001). Although the observational study design precludes causal inferences regarding direct relationships between interventions and outcome improvements, the synergistic decline across multidimensional surveillance metrics indirectly suggests potential systemic spillover effects of FMEA-based management in hospital infection prevention and antimicrobial stewardship.

Despite achieving improvements in nosocomial infection specimen submission processes and promoting rational antimicrobial use through hospital-wide informatization, several areas for development remain. Externally, budget constraints limit access to advanced yet costly testing modalities, hindering comprehensive testing. Variations in patient case mix create department-specific testing demands that cannot be uniformly addressed. Additionally, diagnosis-related group reimbursement constraints introduce cost-benefit trade-offs that indirectly influence testing decisions. Therefore, hospital management

XSL•FO
RenderX

must continue refining collaborative mechanisms across clinical, technical, and administrative departments. Strategic initiatives should include optimizing test menu configurations through cost-effectiveness analysis, developing disease-specific testing guidelines, and leveraging informatization for continuous monitoring to enhance staff efficiency. These measures will ultimately strengthen AMR control, promote rational antibiotic use, improve surveillance capabilities for emerging pathogens, and drive health care quality toward higher precision and patient safety.

This study has several limitations that warrant consideration. First, some of the interventions relied on a relatively well-developed hospital information infrastructure, such as artificial intelligence-driven CDSS alerts. Institutions with limited digital capacity may need to adopt manual checklists or simplified reminder mechanisms as alternatives. Second, the study was conducted in a tertiary TCM hospital. Although the hospital includes Western medicine departments such as the intensive care unit, respiratory medicine, and emergency medicine, with antimicrobial use patterns comparable to general hospitals, the patient population and service model of TCM hospitals may differ from other health care settings, which may limit generalizability. Nevertheless, the methodological framework of FMEA is highly transferable, and individual institutions may adapt specific interventions according to their local resources and workflows. Additionally, this was a single-center, before and after study without a control group, which imposes inherent limitations on causal inference. Improvements in specimen submission rates may not be attributable solely to the FMEA-based interventions; they may also reflect the hospital's broader emphasis on antimicrobial stewardship, underlying secular trends in workflow optimization, or concurrent quality improvement initiatives or policy changes. Future research should validate the effectiveness of these interventions using multicenter designs with appropriate control groups.

Conclusions

This study validated FMEA as a risk management tool for optimizing preantimicrobial specimen submission workflows. By addressing critical failures like barcode errors and inadequate decision support, FMEA-based interventions significantly improved specimen submission rates for restricted/special-use antibiotics, aligning with national quality standards. Challenges, including cost constraints and patient barriers, necessitate policy-level solutions. This model offers a scalable framework for antimicrobial stewardship, with future research needed to evaluate long-term outcomes and advanced diagnostic integration.

Funding

This research was supported by grants from the University-level Scientific Research Project of Zhejiang Chinese Medical University (2025GJYY41) and Zhejiang Provincial Basic Public Welfare Research Program (LTGY23H260002).

Data Availability

The data used in this research are available upon individual request from the corresponding author.

Authors' Contributions

TYD and LXL contributed to the conception and design of the study. JXW contributed to writing the data analysis and drafted the original draft. WJZ, JH, and BWZ contributed to methodology and data curation. QLG, MZ, XYZ, and YXZ critically revised the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Systematic review of the submission rate of microbiological specimen before antibiotic therapy.
[DOCX File, 142 KB - jmir_v28i1e78118_app1.docx]

Multimedia Appendix 2
Action Priority lookup table.
[DOCX File, 26 KB - jmir_v28i1e78118_app2.docx]

Checklist 1
iCHECK-DH (Guidelines and Checklist for the Reporting on Digital Health Implementations) guidelines and checklist for the reporting on digital health implementations.
[DOCX File, 24 KB - jmir_v28i1e78118_app3.docx]

References

XSL•FO
RenderX

1. The Lancet. Antimicrobial resistance: an agenda for all. The Lancet 2024 Jun;403(10442):2349. [doi: 10.1016/S0140-6736(24)01076-6]
2. Painter C, Limmathurotsakul D, Roberts T, et al. Sustainable antimicrobial resistance surveillance: time for a global funding mechanism. Lancet Infect Dis 2025 Feb;25(2):e99-e103. [doi: 10.1016/S1473-3099(24)00649-2] [Medline: 39706207]
3. Naghavi M, Vollset SE, Ikuta KS, et al. Global burden of bacterial antimicrobial resistance 1990–2021: a systematic analysis with forecasts to 2050. The Lancet 2024 Sep;404(10459):1199-1226. [doi: 10.1016/S0140-6736(24)01867-1]
4. Lin L, Sun R, Yao T, Zhou X, Harbarth S. Factors influencing inappropriate use of antibiotics in outpatient and community settings in China: a mixed-methods systematic review. BMJ Glob Health 2020 Nov;5(11):e003599. [doi: 10.1136/bmjgh-2020-003599] [Medline: 33184066]
5. Lim C, Ashley EA, Hamers RL, et al. Surveillance strategies using routine microbiology for antimicrobial resistance in low- and middle-income countries. Clin Microbiol Infect 2021 Oct;27(10):1391-1399. [doi: 10.1016/j.cmi.2021.05.037] [Medline: 34111583]
6. Schuts EC, Hulscher M, Mouton JW, et al. Current evidence on hospital antimicrobial stewardship objectives: a systematic review and meta-analysis. Lancet Infect Dis 2016 Jul;16(7):847-856. [doi: 10.1016/S1473-3099(16)00065-7] [Medline: 26947617]
7. Zou Q, Zhang Y, Ai X, et al. Impact of multidisciplinary collaboration on microbiological specimen submission and appropriate use of carbapenems: a pre-post cohort study. Infect Drug Resist 2025;18:4301-4309. [doi: 10.2147/IDR.S532011] [Medline: 40895614]
8. Centers for Disease Control and Prevention. Core elements of hospital antibiotic stewardship programs[r]. : US Department of Health and Human Services, CDC; 2019 URL: https://www.cdc.gov/antibiotic-use/hcp/core-elements/hospital.html [accessed 2026-02-17]
9. Sánchez-Romero MI, García-Lechuz Moya JM, González López JJ, Orta Mira N. Collection, transport and general processing of clinical specimens in microbiology laboratory. Enferm Infecc Microbiol Clin (Engl Ed) 2019 Feb;37(2):127-134. [doi: 10.1016/j.eimc.2017.12.002] [Medline: 29426791]
10. Circular on the issuance of medical quality control indicators for emergency medicine and five other specialties (2024 edition) by the General Office of the National Health Commission [EB/OL]. China's National Health Commission. URL: https://www.nhc.gov.cn/yzygj/c100068/202405/7d1d82090e8b44d2af3339d7ee8c7cca.shtml [accessed 2026-02-17]
11. Mokrani D, Chommeloux J, Pineton de Chambrun M, Hékimian G, Luyt CE. Antibiotic stewardship in the ICU: time to shift into overdrive. Ann Intensive Care 2023 May 6;13(1):39. [doi: 10.1186/s13613-023-01134-9] [Medline: 37148398]
12. Wang Q, Zhang X, Zheng F, Wang L, Yu T. Clinicians' intention to submit microbiological pathogenic test before antibiotics use and its influencing factors: new evidence from the perspective of hospital management. IDR 2022;Volume 15:3013-3023. [doi: 10.2147/IDR.S366892]
13. Prieto-Molina A, Aranda-Gallardo M, Moya-Suárez AB, Rivas-Ruiz F, Peláez-Cherino J, Canca-Sánchez JC. Patient transfers from emergency departments to other in-hospital areas: a failure mode and effects analysis. Emergencias 2023 Dec;35(6):456-462. [doi: 10.55633/s3me/E010.2023] [Medline: 38116970]
14. Weber L, Schulze I, Jaehde U. Using failure mode and effects analysis to increase patient safety in cancer chemotherapy. Res Social Adm Pharm 2022 Aug;18(8):3386-3393. [doi: 10.1016/j.sapharm.2021.11.009] [Medline: 34838476]
15. Yang S, Yang L, Chuang YC, Asihaer G, Lin X. Identification of key risk areas and failure modes using the FMEA method for the prevention and control of major infectious diseases in a stomatology department. BMC Oral Health 2025 Jan 8;25(1):45. [doi: 10.1186/s12903-024-05321-3] [Medline: 39780161]
16. Chiereghin A, Squillace L, Pizzi L, Bazzani C, Roti L, Mezzetti F. Applying the healthcare failure mode and effects analysis approach to improve the quality of an organised colorectal cancer screening programme. J Med Screen 2024 Jun;31(2):70-77. [doi: 10.1177/09691413231197300] [Medline: 37674432]
17. Wosny M, Strasser LM, Hastings J. The paradoxes of digital tools in hospitals: qualitative interview study. J Med Internet Res 2024 Jul 15;26:e56095. [doi: 10.2196/56095] [Medline: 39008341]
18. Skodvin B, Wathne JS, Lindemann PC, et al. Use of microbiology tests in the era of increasing AMR rates– a multicentre hospital cohort study. Antimicrob Resist Infect Control 2019;8(1):28. [doi: 10.1186/s13756-019-0480-z] [Medline: 30740221]
19. Lao Y, Tang Q, Zeng J, Gong W, Shen Y. Effect analysis of multi-department cooperation in improving etiological submission rates before antibiotic treatment. Int J Qual Health Care 2023 Jun 6;35(2):mzad018. [doi: 10.1093/intqhc/mzad018] [Medline: 37040384]
20. Zheng F, Wang K, Wang Q, Yu T, Zhang X. The pre-analytical process management status and influencing factors of laboratory test before prescribing antimicrobial in developing country. BMC Health Serv Res 2023 Mar 25;23(1):283. [doi: 10.1186/s12913-023-09243-8] [Medline: 36966281]
21. Catho G, Sauser J, Coray V, et al. Impact of interactive computerised decision support for hospital antibiotic use (COMPASS): an open-label, cluster-randomised trial in three Swiss hospitals. Lancet Infect Dis 2022 Oct;22(10):1493-1502. [doi: 10.1016/S1473-3099(22)00308-5] [Medline: 35870478]
22. Dutey-Magni PF, Gill MJ, McNulty D, et al. Feasibility study of hospital antimicrobial stewardship analytics using electronic health records. JAC Antimicrob Resist 2021 Mar;3(1):dlab018. [doi: 10.1093/jacamr/dlab018] [Medline: 34223095]

XSL•FO
RenderX

Abbreviations

AMR: antimicrobial resistance
AP: Action Priority
CDSS: clinical decision support systems
EHR: electronic health record
FMEA: Failure Mode and Effects Analysis
LMS: learning management system
RPN: Risk Priority Number
TCM: traditional Chinese medicine

Edited by A Schwartz; submitted 28.May.2025; peer-reviewed by O Ajileye, SLL Holgate; accepted 16.Jan.2026; published 10.Mar.2026.

Please cite as:
Wu J, Zhan W, Hua J, Ge Q, Zhu Y, Yu H, Zhao M, Zhan X, Zhu B, Xiang T, Lu L, Dai T
Investigating the Effect of Hospital Infection Control Informatization on Optimizing Microbiological Specimen Submission Before Antibiotic Therapy: Failure Mode and Effects Analysis
J Med Internet Res 2026;28:e78118
URL: https://www.jmir.org/2026/1/e78118
doi:10.2196/78118

© Jianxiong Wu, Weijiang Zhan, Jun Hua, Qinling Ge, Yuexian Zhu, Huixian Yu, Min Zhao, Xiaoyan Zhan, Bingwei Zhu, Tongqi Xiang, Longxi Lu, Tieying Dai. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Immersive Technologies for Cognitive Rehabilitation in Dementia and Mild Cognitive Impairment: Systematic Review

Jaromir Konecny[1], PhD; Giuseppe Lanza[2,3], PhD; Serafino Buono[3,4], PhD; Raffaele Ferri[3], MD; Antonina Luca[4], PhD; Radek Martinek[1], PhD; Sabrina Musso[3], PsyD; Aurora Palmigiano[2,3], PsyD; Barbora Prauzkova[5], MSc; Angelica Quercia[3], PhD; Francesco Rundo[6], EngD; Ramachandran Avala Subramanian[1], MSc; Alessandro Serretti[3,4], PhD; Michal Prauzek[1], PhD

[1]Department of Cybernetics and Biomedical Engineering, VSB – Technical University of Ostrava, 17. listopadu 2172/15, Ostrava - Poruba, Czech Republic

[2]Department of Surgery and Medical-Surgical Specialties, University of Catania, Catania, Italy

[3]Department of Clinical Neurosciences, Oasi Research Institute-IRCCS, Troina, Italy

[4]Department of Medicine and Surgery, “Kore” University of Enna, Enna, Italy

[5]Department of Management, VSB – Technical University of Ostrava, Ostrava - Poruba, Czech Republic

[6]Unit of Bioinformatics and Statistics, Oasi Research Institute-IRCCS, Troina, Italy

Corresponding Author:
Michal Prauzek, PhD
Department of Cybernetics and Biomedical Engineering, VSB – Technical University of Ostrava, 17. listopadu 2172/15, Ostrava - Poruba, Czech Republic

Abstract

Background: Cognitive decline across the mild cognitive impairment (MCI)–dementia continuum is a major driver of loss of independence and growing health- and social-care burden. Immersive technologies, such as virtual reality (VR), augmented reality (AR), and Cave Automatic Virtual Environment (CAVE) systems, are increasingly explored as tools to enhance engagement, personalization, and ecological validity in cognitive rehabilitation.

Objective: This systematic review synthesizes current evidence on the usability, therapeutic effects, and implementation challenges of immersive technologies for cognitive rehabilitation in MCI and dementia.

Methods: A systematic search of Scopus and Web of Science was conducted for peer-reviewed journal articles published between 2021 and 2026. Eligible studies investigated VR, AR, or CAVE interventions targeting cognitive rehabilitation outcomes in MCI and/or dementia and reported measures related to usability or acceptability, or cognitive, functional, or behavioral outcomes. Due to heterogeneity in technologies, intervention content, and outcome measures, findings were synthesized narratively with comparisons across modalities and study designs.

Results: In total, 119 studies met the inclusion criteria. Across immersive VR interventions, signals of benefit were most consistently reported for memory, attention, and executive functioning, with several studies also targeting outcomes with higher ecological relevance (eg, everyday task performance and functional skills). AR approaches primarily support context-aware cueing and task guidance in real-world settings, aiming to strengthen daily functioning and independence. CAVE-based systems were frequently used for spatial navigation and embodied interaction, offering advantages for supervised clinical deployment. Key barriers included cybersickness and comfort issues, interface complexity, and onboarding demands in cognitively impaired users, limited accessibility and standardization of outcome measures, small samples and short follow-up periods, and practical constraints related to cost, space, staffing, and caregiver involvement.

Conclusions: Immersive VR, AR, and CAVE systems are feasible and often engaging for cognitive rehabilitation in MCI and dementia, with promising therapeutic signals but substantial uncertainty driven by methodological and implementation heterogeneity. Future work should prioritize standardized reporting (intervention components, dose, and adverse events), clinically meaningful outcomes (including functional end points), adequately powered comparative trials, and explicit evaluation of scalability and real-world deployment pathways.

(J Med Internet Res 2026;28:e84349) doi:10.2196/84349

KEYWORDS

cognitive rehabilitation; dementia; CAVE system; virtual reality; augmented reality; mild cognitive impairment; mixed reality; PRISMA; Cave Automatic Virtual Environment system

XSL•FO
RenderX

Introduction

Dementia represents a rapidly growing global health challenge. In 2019, approximately 55 million people were estimated to be living with dementia worldwide, and this number is projected to rise to 139 million by 2050, underscoring the urgency of scalable, evidence-based rehabilitation and support strategies [1]. Cognitive decline across the mild cognitive impairment (MCI)–dementia continuum is a leading driver of loss of independence in later life and a growing burden for health and social care systems. MCI is commonly conceptualized as an intermediate clinical state between typical cognitive aging and dementia, but its course is heterogeneous—some individuals remain stable or improve, whereas others progress to dementia—which makes early, targeted intervention both attractive and complex [2-4]. Dementia syndromes, including Alzheimer disease (AD), involve progressive impairments across multiple cognitive domains with downstream effects on everyday functioning and quality of life, reinforcing the need for interventions that are not only clinically meaningful but also feasible in real-world settings [5,6].

Nonpharmacological approaches remain central to cognitive rehabilitation and support, particularly because disease-modifying pharmacotherapies are limited in scope and may not address day-to-day functional challenges across disease stages [5,7]. Conventional cognitive training and rehabilitation can improve specific skills, yet long-term engagement, personalization to heterogeneous deficits, and transfer to everyday functioning are persistent challenges [5,7,8]. In this context, outcomes related to activities of daily living (ADL) and instrumental activities of daily living (IADL) are especially important because they capture functionally meaningful abilities that shape independence and care needs [9]. Digital and computer-based interventions have therefore been explored as scalable tools for cognitive support, but questions remain regarding ecological validity, long-term adherence, and implementation barriers outside controlled research contexts [7,8].

Immersive technologies are increasingly proposed as a route to address these limitations by enabling interactive, multisensory, context-rich tasks that can approximate real-life demands. Conceptually, these systems can be situated within the broader extended reality (XR) landscape and the reality-virtuality continuum, which clarifies the relationships among virtual reality (VR), augmented reality (AR), and mixed reality [10-12]. In this review, we use VR to denote predominantly computer-generated environments experienced with varying levels of immersion (eg, head-mounted displays or projection-based systems), AR to denote digital overlays and cues integrated into the real world (often via mobile or wearable devices), and CAVE (Cave Automatic Virtual Environment) systems as a projection-based, room-scale configuration of VR that supports embodied interaction and navigation without headgear [10,11,13]. Distinguishing these modalities is practically relevant because they differ in interaction demands, supervision requirements, space and cost constraints, and potential adverse effects, which can influence both feasibility and outcomes in older adults with cognitive impairment [14-16].

Across cognitive rehabilitation, immersive tasks have been used to target memory, attention, executive functions, and spatial or navigation abilities, and they can embed cognitively demanding activities within ADL or IADL-like scenarios to enhance ecological validity and functional relevance [13,17-19]. However, translation beyond controlled settings remains constrained by substantial heterogeneity in platforms and immersion levels, interaction modalities, task designs, intervention dose (session duration and frequency), comparator conditions, and outcome selection, which complicates synthesis and limits the strength and generalizability of inferences [4,20,21]. These challenges are amplified by the fact that XR-based outcomes cannot be interpreted independently of feasibility; usability, acceptability, and accessibility directly shape adherence, delivered exposure (effective “dose”), safety, and the interpretability of cognitive and functional measures in cognitively impaired older adults [14,15].

To address this, this review adopts an explicit usability lens. We use usability in the human-computer interaction sense as a multidimensional construct encompassing effectiveness, efficiency, and satisfaction in a specified context of use. In this literature, usability evidence may be drawn from standardized instruments, task performance indicators (eg, errors and assistance needs), completion and adherence metrics, qualitative user-experience reports, and stakeholder perspectives (patients, caregivers, and clinicians) [18,22]. This framing is particularly important in dementia and MCI because age- and disease-related sensory, motor, and cognitive constraints can make interface complexity, onboarding demands, fatigue, and tolerability decisive determinants of whether an intervention is deliverable, interpretable, and scalable [8,14,15].

Several systematic reviews have examined technology-supported cognitive rehabilitation, but the evidence base remains fragmented with respect to target population (MCI vs dementia or AD), XR modality (VR-only vs broader XR), and the depth and standardization of usability and implementation synthesis [3,4,20,23-27]. In parallel, broader assistive-technology reviews in dementia synthesize important adjacent evidence (including immersive and wearable technologies) but do not provide a modality-explicit, rehabilitation-focused synthesis that treats usability and translation constraints as analytically coded outcomes [28]. Likewise, AR-focused reviews for cognitively impaired populations emphasize action assistance and learning, but their scope is not centered on cognitive rehabilitation outcomes across the MCI-dementia continuum and does not integrate projection-based immersive configurations, such as CAVE [29]. Taken together, these gaps motivate a consolidated systematic review that (1) spans the MCI-dementia continuum, (2) compares immersive technology types with explicit attention to VR (including CAVE as projection-based VR) and AR, and (3) treats usability and implementation constraints as first-class evidence streams alongside cognitive and functional outcomes [14-16].

Accordingly, this review is guided by the following research question: What is the current evidence on usability, feasibility, and reported intervention outcomes of VR, AR, and CAVE systems for cognitive rehabilitation in individuals with MCI and dementia? In addressing this question, we additionally

XSL•FO
RenderX

examine how intervention characteristics (technology type, immersion level, interaction demands, and dose), study design choices, and outcome selection contribute to heterogeneity in reported findings and shape the feasibility of translation beyond controlled settings [20,21,30].

Methods

Review Contributions and Structure

First, it synthesizes current evidence on the usability, therapeutic efficacy, and limitations of immersive systems (VR, AR, and CAVE) in cognitive rehabilitation for individuals with MCI and dementia.

Second, it compares approaches across technologies, study designs, and outcome measures to identify key trends and limitations in the field.

Finally, it outlines challenges and proposes priorities for future research, including methodological standardization and scalability of interventions.

The overall structure of this systematic review is summarized in Figure 1. After the Introduction section, the Methods section details the search strategy, eligibility criteria, study selection, data extraction, and synthesis approach. The Results section then reports the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)-based selection process and presents the included evidence across VR, AR, and CAVE modalities. Next, the Discussion section interprets the findings with an emphasis on usability, feasibility, and translational considerations and highlights implications for research and practice. Finally, the Conclusions section summarizes the key takeaways and priorities for future work.

Figure 1. Graphical outline of the study. ADL: activities of daily living; AR: augmented reality; CAVE: Cave Automatic Virtual Environment; IADL: instrumental activities of daily living; MCI: mild cognitive impairment; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; UX: user experience; VR: virtual reality.

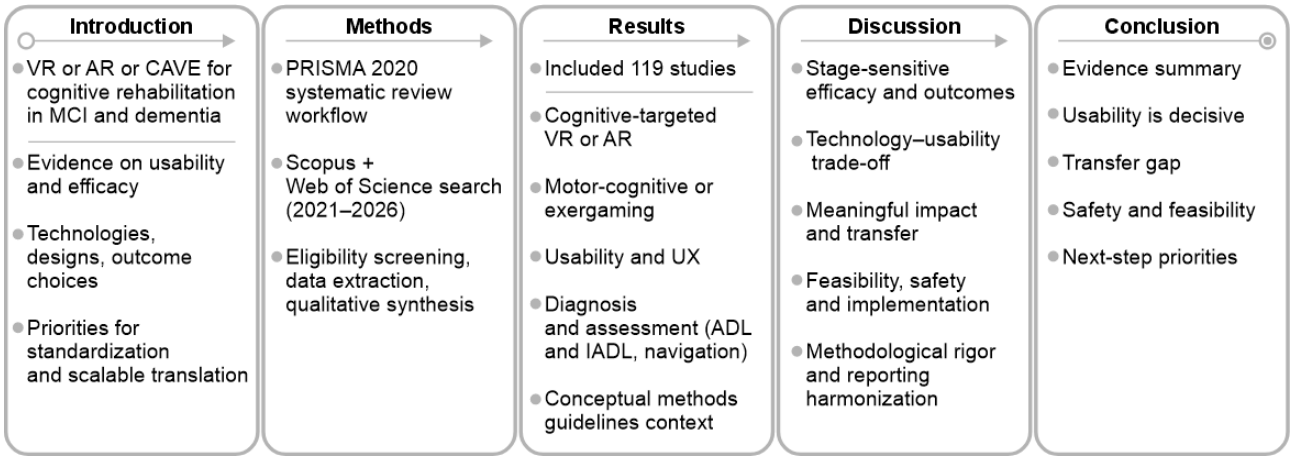

Protocol and Registration

This systematic review was conducted in accordance with the PRISMA 2020 guidelines. The protocol was prospectively developed to ensure methodological rigor and transparency. Trial registration was not applicable, as this work did not involve primary clinical trials.

Eligibility Criteria

Studies were considered eligible if they addressed the use of immersive technologies—specifically VR, AR, or CAVE systems—in the context of cognitive rehabilitation for individuals with MCI and/or dementia. The literature search was limited to publications from 2021 to 2026 and used the following query, "(virtual reality OR augmented reality OR CAVE) AND cognitive AND (dementia OR MCI) AND rehabilitation." Only peer-reviewed journal articles written in English were included. Records without a DOI, missing author information or source title, or those that could not be accessed in full-text form were excluded. Conference papers, abstracts, editorials, and other nonarticle document types were removed. Furthermore, records deemed unrelated to cognitive rehabilitation, MCI, or dementia upon title and abstract screening were excluded. The final set of eligible studies was subsequently divided into 2 categories: that review articles and regular (original research) articles.

Information Sources and Search Strategy

The literature search was conducted using 2 major scientific databases—Web of Science (WoS) and Scopus. These databases were selected due to their broad coverage of peer-reviewed journals in the fields of health care, rehabilitation, neuroscience, and immersive technologies. A structured search strategy was applied in both databases using a consistent combination of keywords and Boolean operators to identify relevant studies. The search focused on titles, abstracts, and author keywords and was designed to capture publications addressing the application of VR, AR, or CAVE systems in cognitive rehabilitation for individuals with MCI and/or dementia. The same search query and inclusion limits were applied across both databases to ensure comparability and reproducibility of the results.

Selection Process

A total of 2 reviewers (RAS and JK) independently screened titles, abstracts, and full texts against predefined criteria. Disagreements were resolved through consensus or consultation with a third reviewer (MP). The selection process was documented using the PRISMA 2020 flow diagram.

XSL•FO
RenderX

Data Extraction

Data were extracted using standardized forms capturing: study design, authorship, publication year, country, participant characteristics, intervention specifics (device type, duration, frequency, and context), cognitive domains targeted, and outcome measures (cognitive tests, usability scales, and emotional or behavioral outcomes). Extraction was performed in duplicate by independent reviewers to minimize errors.

Risk of Bias and Quality Considerations

We performed a structured qualitative appraisal of methodological quality and potential sources of bias during data extraction and narrative synthesis. For each included publication, we recorded study design features and reporting characteristics that most strongly influence interpretability and transferability in immersive-technology research, including (1) study design and the presence or absence of comparator conditions, (2) sample size and participant characterization, (3) intervention or system description and dose (duration, frequency, and setting), (4) completeness and transparency of outcome reporting (including usability or acceptability metrics and adverse events such as cybersickness), (5) follow-up duration, and (6) practical implementation factors (equipment constraints, supervision requirements, staffing, and caregiver involvement). These quality considerations were used to contextualize findings, identify recurring limitations across studies, and guide cautious interpretation of reported outcomes in the narrative synthesis.

Data Synthesis

Given heterogeneity in technologies, intervention content, and outcome measures, a meta-analysis was not feasible. We therefore conducted a narrative synthesis structured by technology type (VR, AR, and CAVE) and outcome domain (usability or acceptability, cognitive performance, daily functioning, and emotional well-being). Within each category, we summarized intervention and system characteristics (eg, hardware setup, immersion level, interaction modalities, delivery setting, and dose where reported) and synthesized reported outcomes qualitatively. For studies that included comparator conditions, we described the direction and consistency of reported between-group findings without statistical pooling. We also synthesized implementation-relevant factors, including onboarding demands, supervision requirements, and adverse events (eg, cybersickness) when reported.

Results

Literature Selection

This section was conducted in accordance with the PRISMA guidelines. Studies were included if they met clearly defined eligibility criteria, that are (1) publication period from 2021 to 2026, (2) English language, (3) peer-reviewed journal articles only, and (4) a clear focus on cognitive rehabilitation in individuals with MCI and/or dementia. Searches were conducted on November 2, 2025. Eligible interventions involved immersive or semi-immersive technologies, specifically VR, AR, or CAVE systems. Records without a DOI, missing author information or source title, or without accessible full-text availability were excluded, as were studies unrelated to cognitive rehabilitation, MCI, or dementia.

Comprehensive electronic searches were conducted in 2 major bibliographic databases, that is, Scopus and WoS. The search strategy used Boolean operators and the following search string, (virtual reality OR augmented reality OR CAVE) AND cognitive AND (dementia OR MCI) AND rehabilitation. The same search strategy and filters were applied consistently across both databases. Retrieved records were exported to reference management software, where duplicates were removed before screening.

Titles and abstracts were screened to exclude clearly irrelevant records, followed by full-text assessment of the remaining articles against the predefined inclusion and exclusion criteria. Only articles that could be downloaded in full text were retained. The final set of included studies was subsequently divided into 2 categories—review articles and regular (original research) articles, to enable structured comparison across study types.

Due to heterogeneity in study designs, intervention characteristics, and outcome measures, a meta-analysis was not feasible. Instead, a narrative synthesis was performed, focusing on the characteristics of the VR, AR, or CAVE interventions, targeted cognitive domains, reported outcomes related to rehabilitation and usability, and identified methodological limitations across studies.

The PRISMA flow diagram (Figure 2) summarizes the study selection process. A total of 236 records were identified from databases (Scopus and WoS), with no records from registers (n=0). Before screening, 57 duplicate records and 9 records marked as ineligible by automation tools were removed, leaving 170 records for screening [31]. No records were excluded at the screening stage (n=0); therefore, 170 reports were sought for retrieval, of which 28 reports could not be retrieved. The remaining 142 reports were assessed for eligibility, and 23 reports were excluded as not related. Ultimately, 119 studies [2-8,13-21,23-27,30-127] were included in the review (119 reports of included studies).

XSL•FO
RenderX

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram illustrating the article selection process.

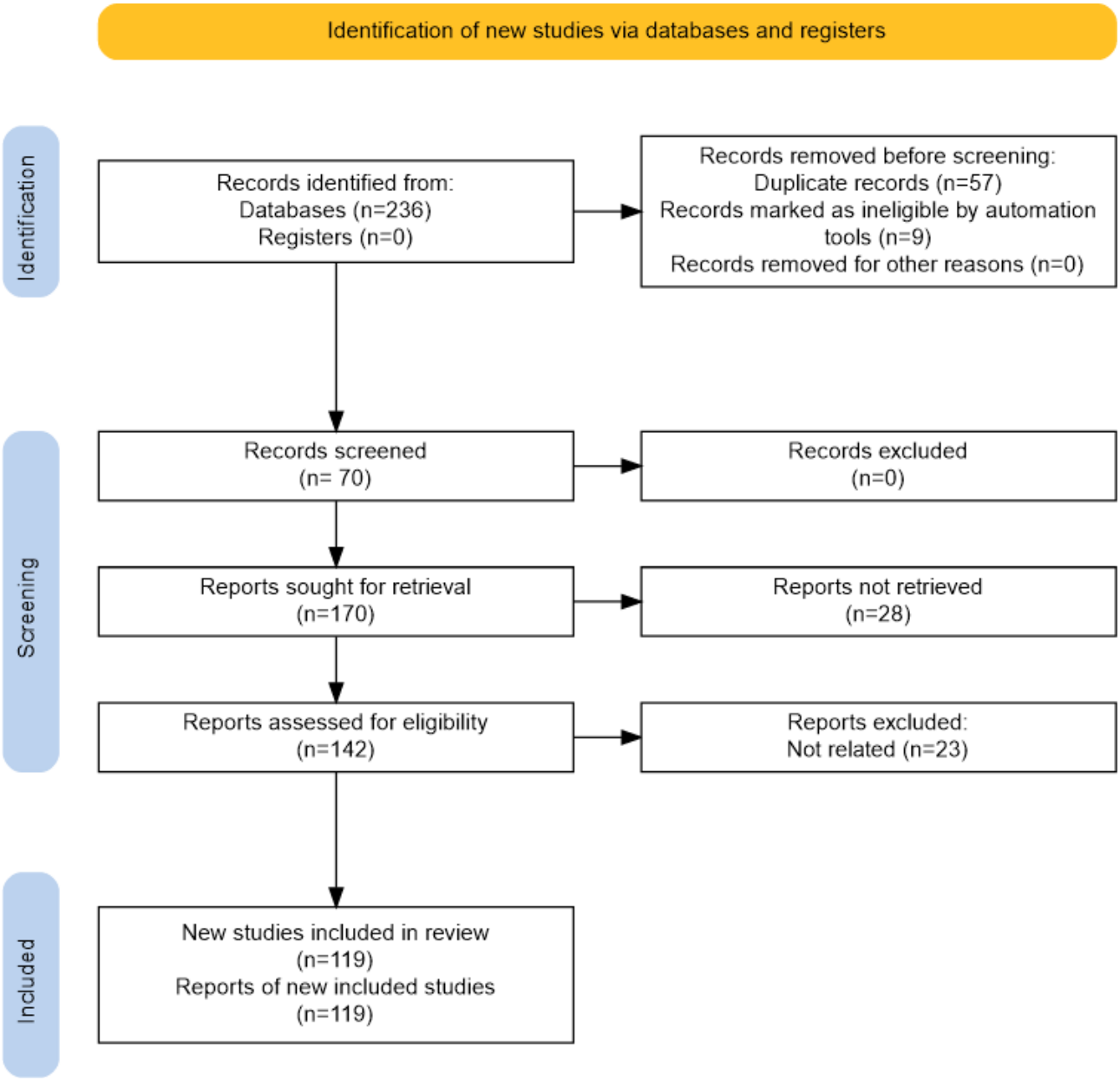

Positioning of This Review Relative to Existing Systematic Reviews

A growing number of systematic reviews have already examined XR-supported cognitive rehabilitation, but their scopes remain fragmented with respect to target population, XR modality, immersion level, and the depth of usability and implementation synthesis. Several focus primarily on MCI and VR-based cognitive rehabilitation, often emphasizing cognitive efficacy and protocol characteristics while providing only limited, nonstandardized coverage of usability, feasibility, and real-world deployment constraints [3,4,20,23-27]. Other reviews address dementia, AD, or functional outcomes (ADL or IADL) and ecological validity, yet typically remain centered on VR (frequently non- or semi-immersive) and treat usability, acceptability, adverse events, and service integration as secondary considerations rather than as analytically coded dimensions [5,6,17]. In parallel, more specialized reviews concentrate on narrower constructs—such as technology-assisted dual-task rehabilitation or spatial navigation paradigms—providing valuable conceptual framing but not offering an implementation-oriented synthesis of immersive therapeutic XR across the MCI-dementia continuum [13,19]. Finally, several systematic reviews adopt broader clinical populations or technology umbrellas (eg, survivors of stroke, Parkinson disease, acquired cognitive disorders, or general digital-technology interventions in older adults), which supports methodological insight but limits direct transferability to dementia-focused cognitive rehabilitation pathways and to cross-modality comparisons including AR and CAVE [7,32-39,128]. A related strand reviews fully immersive VR-based cognitive remediation in adult psychiatric or psychosocial rehabilitation (including some MCI and mild dementia samples), highlighting intervention-component reporting gaps that constrain real-world implementation [40]. In addition, reviews outside the XR rehabilitation core (eg, bibliometric mapping in pediatrics or topic-distant clinical domains) further illustrate how rapidly expanding evidence can

XSL•FO
RenderX

become fragmented and difficult to translate to dementia-specific implementation decisions [41-43].

Beyond cognitive-rehabilitation–centered syntheses, the XR literature also contains systematic reviews that cluster around 2 adjacent themes which inform, but do not substitute for, our focus on immersive XR (VR, AR, or CAVE) cognitive rehabilitation in MCI and dementia.

Reviews in the category "Psychosocial, Emotional, and Behavioral Interventions Using Immersive Technologies" predominantly frame immersive and digital technologies as instruments for psychosocial support, emotional well-being, or broader behavioral outcomes rather than as structured cognitive rehabilitation protocols targeted to the MCI-dementia continuum. For example, one line of work synthesizes XR use in older adults at a general level (including engagement and experience-oriented outcomes) without restricting scope to cognitive rehabilitation in MCI or dementia or to immersive XR modality comparisons [8]. Other reviews adopt an even broader "aging and well-being" perspective, discussing strategies to improve quality of life across normal versus pathological aging, which is informative for contextual motivation but not specific to immersive XR cognitive rehabilitation efficacy or implementation in dementia care pathways [44]. Several reviews examine psychosocially oriented digital interventions and symptom-focused cognitive supports (rather than immersive XR rehabilitation protocols), which limits transferability to our modality-specific and deployment-oriented synthesis in dementia or MCI [45]. A related strand focuses on reminiscence-oriented applications using VR-based approaches, which aligns with person-centered care and engagement, yet differs from our emphasis on cognitive rehabilitation outcomes and cross-technology synthesis across VR, AR, or CAVE [46]. Some reviews concentrate on specific training paradigms or outcome contexts—such as attention rehabilitation via serious games, or changes in cognitive screening performance across training regimes—where dementia-focused immersive XR rehabilitation and implementation or scalability considerations are not the primary analytic axis [47]. Importantly, additional reviews in this group address adjacent rehabilitation or stimulation concepts (eg, multisensory stimulation in dementia rehabilitation) that overlap conceptually but do not provide the XR-modality–differentiated usability and implementation synthesis that our review targets [48]. Finally, reviews focused on digital cognitive interventions in nondementia neurological populations (eg, traumatic brain injury) are methodologically relevant but population-mismatched for the MCI or dementia rehabilitation pathway that defines our eligibility boundaries [49]. Taken together, these help justify the broader psychosocial value and acceptability of immersive or digital interventions, but our review deliberately differentiates itself by restricting the population to MCI and dementia, expanding the technology scope to immersive XR (VR, AR, or CAVE), and systematically integrating usability and real-world implementation or scalability evidence alongside cognitive rehabilitation outcomes.

Reviews in the category "Design, Usability, Safety, and Implementation of Immersive Systems for Older Adults" are closer to our aims in that they foreground design, usability, safety, feasibility, and translational barriers; however, they commonly span heterogeneous populations, modalities, and use cases that dilute direct inference for dementia or MCI cognitive rehabilitation with immersive XR. For instance, a dedicated synthesis of cybersickness and fully immersive VR provides essential safety framing; yet, it is not specifically anchored in dementia or MCI cognitive rehabilitation workflows or in VR, AR, or CAVE cross-modality comparison [50]. Other work reviews XR technologies or interface classes relevant to implementation (eg, smart glasses for cognitive impairment), but it typically emphasizes device feasibility and adoption rather than mapping therapeutic cognitive rehabilitation effects across immersive XR modalities [51]. Several reviews focus on assessment- and function-oriented constructs (eg, IADL assessment) that are highly relevant to outcome selection, while still differing from our rehabilitation-centric scope and our modality-explicit synthesis across immersive XR systems [52]. Additional reviews address intervention families that are broader than immersive XR (eg, digital game-based or computer-based cognitive interventions in cognitively impaired community-dwelling populations), which informs feasibility and design considerations but does not resolve the immersion- and modality-specific trade-offs central to our review [53]. Similarly, comparative or umbrella-type syntheses of nonpharmacologic interventions in older adults provide important taxonomic and implementation context, yet they do not isolate immersive XR (VR, AR, or CAVE) cognitive rehabilitation for MCI or dementia as the analytic core [54]. Bibliometric and mapping-type reviews contribute useful evidence on research trends and gaps, but they do not offer the structured, outcome- and usability-centered comparative synthesis required to guide scalable deployment of immersive XR cognitive rehabilitation [55]. Finally, several reviews in this category focus on other clinical indications or technology ecosystems (eg, poststroke cognitive impairment, Parkinson disease, and sensor or interaction technologies), which provide transferable implementation lessons but remain population- or aim-mismatched to dementia or MCI cognitive rehabilitation [56]. Therefore, while these provide critical inputs for human-centered design, safety, and adoption, our review distinguishes itself by (1) focusing specifically on MCI and dementia, (2) treating immersive XR modality (VR, AR, and CAVE) and immersion level as explicit comparative dimensions, and (3) linking usability or feasibility and implementation or scalability evidence directly to cognitive rehabilitation outcomes and methodological standardization needs in real-world care settings.

Therapeutic VR or AR Interventions Primarily Targeting Cognitive Functions

Table 1 summarizes therapeutic interventions in which the primary target is cognitive functioning, spanning neurodegenerative conditions (AD or dementia and Parkinson disease), MCI, poststroke cognitive impairment, and acquired brain injury, as well as prevention-oriented programs in older adults [57-65]. Across this body of work, interventions aim to improve global cognition and/or specific domains such as executive functions, memory (including working and prospective memory), attention, and visuospatial abilities through structured,

XSL•FO
RenderX

task-based training embedded in virtual or augmented environments [66-70].

XSL•FO
RenderX

Table . Therapeutic VR[a] or AR[b] interventions primarily targeting cognitive functions.

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
Lanzoni et al (2022) [69]	Adults with retrograde amnesia after brain stroke	Usability or design N=NR[c]	VR—Nonimmersive (screen) customizable serious games built from 3D scans of familiar environments for memory rehabilitation	• Dose: Single or few sessions for usability testing • Domains: Spatial memory, episodic memory, orientation • Outcome: Demonstrated feasibility and usability; cognitive outcomes exploratory only
Oliveira et al (2021) [72]	Older adults with mild-to-moderate Alzheimer dementia	Pilot RCT[d] N=17	VR—Immersive (HMD[e] or CAVE[f]) VR cognitive stimulation reproducing IADL[g]	• Dose: 2 months, 10 sessions (2 sessions/week) • Domains: Global cognition, memory, attention, and executive functions • Outcome: Improved global cognition with large effect size in the VR group compared with control
Ip et al (2025) [71]	Community-dwelling older adults with MCI[h]	Pilot RCT N=18	VR exergame—Nonimmersive (large screen) full-immersive VR game-based cognitive-motor training compared with group-based Baduanjin exercise	• Dose: 8 weeks, 16 sessions (2 sessions/week) • Domains: Executive functions, global cognition, and dual-task performance • Outcome: Greater improvements in cognitive-motor performance, mobility, balance, and gait speed in the VR group; no clear benefit in executive function
Huang and Yang (2022) [58]	Older adults with dementia in long-term care	Observational N=20 (7 followed at 3-6 months)	VR—Immersive (HMD) VR reminiscence therapy using immersive scenes related to participants' life history	• Dose: 3 months, 2 sessions/week • Domains: Global cognition (secondary), autobiographical memory, and mood • Outcome: No significant cognitive changes immediately post intervention, but depressive symptoms decreased; cognition declined again at 3-6 months

XSL•FO
RenderX

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
De Simone et al (2023) [60]	Patients with Parkinson disease and MCI; healthy older adults	Protocol N=60 (30 PD-MCI and 30 healthy)	VR+telemedicine—Immersive (HMD) home-based VR cognitive training targeting executive functions to improve prospective memory	• Dose: Planned multiweek intervention (pre- and post–2-month follow-up) • Domains: Prospective memory and executive functions (planning, shifting, and updating) • Outcome: Only baseline and feasibility results available; cognitive efficacy not yet reported
Han et al (2021) [67]	Older adults at risk of cognitive decline (community dwelling)	Prototype and expert evaluation; exploratory effectiveness N=NR	AR—Mobile AR gamified AR cognitive training system inspired by Trail Making Test, performed in real space	• Dose: Short-term pilot (single or few sessions) • Domains: Working memory, attention, and visual perception • Outcome: Experts judged that the system could improve working memory and attention; effectiveness in older adults not yet fully tested
De Luca et al (2024) [62]	Adults with severe acquired brain injury (traumatic and nontraumatic)	RCT N=40 (20 VR-based ROT[i], 20 control)	VR—Nonimmersive (screen) VR-based ROT compared with standard reality orientation therapy	• Dose: Inpatient rehabilitation cycle (several weeks) • Domains: Global cognition (MMSE[j]), orientation, and attention • Outcome: Both groups improved, but VR-ROT showed greater MMSE gains and reduced depressive symptoms
Buele et al (2024) [66]	Older adults with MCI	Single-blind RCT N=34 (17 VR, 17 control)	VR—Semi or full immersive group motor training followed by VR-based cognitive training vs motor training plus traditional cognitive training	• Dose: 6 weeks, 12 sessions (40 minutes each) • Domains: Global cognition, executive functions, and memory • Outcome: Both groups improved in cognition and depression; between-group differences were not significant, but VR group reached higher task difficulty
Downey et al (2023) [64]	Middle-aged and older adults with and without hearing loss	Single-blind RCT N=NR	CCT[k]+VR assessment—Immersive (assessment) At-home computerized executive-function training vs wait-list control, with dual-task gait	

XSL•FO
RenderX

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
				• Dose: 12 weeks of at-home training • Domains: Executive functions, attention, and dual-task performance • Outcome: Planned to test whether executive function training improves cognitive-motor dual-task performance; primary outcomes not fully reported yet
Liu et al (2023) [61]	Older adults with poststroke MCI	Pilot RCT N=30 (15 VR, 15 control)	VR—Immersive (HMD) immersive VR puzzle games added to routine rehabilitation vs traditional cognitive training	• Dose: 6 weeks of treatment • Domains: Executive function, visuospatial attention, and processing speed • Outcome: Significant improvement in DSST in VR group, suggesting gains in executive and visuospatial functions; other measures improved but not sign…
Oliveira et al (2021) [72]	Institutionalized nonrobust older adults	Single-blind RCT N=NR	VR exergame—Nonimmersive (exergame) 2D VR exercise training vs equivalent exercises without VR	• Dose: 2 months, 16 sessions (2 sessions/week) • Domains: Spatial navigation and executive function • Outcome: VR group showed significantly greater improvement in spatial navigation compared with active control
Jeong et al (2023) [68]	Adults with acquired brain injury	Randomized crossover trial N=NR	VR (music)—Immersive (HMD) VR-based music attention training (drumming tasks) vs conventional computerized cognitive training	• Dose: Two 4-week phases (VR vs control) in crossover • Domains: Selective, alternating, and divided attention; executive function • Outcome: VR-MAT improved attention and some executive measures more than conventional training in at least 1 phase
Maeng et al (2021) [70]	Older adults with and without MCI	RCT N=NR	VR—Nonimmersive VR-based cognitive training tasks (eg, ADL[1] scenarios and puzzles) compared with conventional or no training	

XSL•FO
RenderX

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
				• Dose: Multiweek program with repeated sessions • Domains: Global cognition, memory, executive functions, and attention • Outcome: Improvement or stabilization of cognition in older adult participants, particularly those with MCI
Baldimtsi et al (2023) [63]	Older adults with MCI	Randomized intervention trial N=NR	VR (physical+cognitive)—Nonimmersive (likely) combined physical exercise and VR-based cognitive tasks vs conventional or usual care	• Dose: Several weeks with multiple sessions per week • Domains: Executive functions, memory, and attention • Outcome: VR physical+cognitive training associated with greater cognitive gains than comparison conditions
Fiorenzato et al (2025) [59]	Patients with Parkinson disease and healthy older adults	RCT N=NR	VR—Nonimmersive VR-based cognitive training tasks targeting prospective memory and executive functions	• Dose: Multisession intervention over several weeks • Domains: Executive functions and prospective memory • Outcome: Improvements in prospective memory and executive functions in both PD and healthy aging groups
Zhu et al (2022) [65]	Adult patients undergoing laparoscopic surgery	Protocol N=NR	VR—Immersive (HMD) biophilic VR exposure pre- and/or postoperatively compared with sham VR	• Dose: Short perioperative VR sessions • Domains: Postoperative cognitive function (attention, working memory, and global cognition) • Outcome: No results yet; hypothesis that biophilic VR will mitigate postoperative cognitive decline
Początek et al (2023) [73]	Patients with Alzheimer disease	Narrative or systematic review N=NR	Mixed (including VR or AR)—not applicable, overview of cognitive training approaches, including emerging VR-based methods	• Dose: Various across included trials • Domains: Memory, executive functions, language, and attention • Outcome: Cognitive training yields small-to-moderate benefits; VR or AR highlighted as promising but underresearched
Cavallo and Lasaponara (2025) [74]		Narrative review N=NR		

XSL•FO
RenderX

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
	Patients with various neurodegenerative diseases (eg, Alzheimer disease and Parkinson disease)		Mixed (including VR or AR)—not applicable, overview of cognitive deficits and cognitive training modalities in neurodegenerative diseases	• Dose: Various across discussed studies • Domains: Multiple domains (memory, executive functions, attention, and visuospatial abilities) • Outcome: Supports use of structured cognitive training; notes need for more high-quality RCTs and the potential of VR or AR

[a]VR: virtual reality.
[b]AR: augmented reality.
[c]NR: not reported.
[d]RCT: randomized controlled trial.
[e]HMD: head-mounted display.
[f]CAVE: Cave Automatic Virtual Environment.
[g]IADL: instrumental activities of daily living.
[h]MCI: mild cognitive impairment.
[i]ROT: reality orientation therapy.
[j]MMSE: Mini-Mental State Examination.
[k]CCT: computerized cognitive training.
[l]ADL: activities of daily living.

Most studies prioritize executive functioning and attentional control as core therapeutic targets, reflecting their relevance for everyday functioning and their frequent impairment in clinical aging and neurological populations [59,60,62,68]. Memory-focused approaches include working memory training (notably via mobile AR) and prospective memory training in Parkinson disease and healthy aging, with several programs attempting to increase ecological validity by using realistic or meaningful task contexts [57,59,67].

In poststroke and acquired brain injury contexts, VR is used both for domain-specific cognitive rehabilitation (eg, visuospatial attention and executive processes) and for broader orientation and behavioral support, often as an adjunct to standard rehabilitation [61,62].

The subgroup includes immersive head-mounted display (HMD)–based VR (eg, reminiscence and cognitive training), nonimmersive or screen-based VR platforms, and mobile AR serious games [58,61,67,69,70]. A notable direction is the integration of immersive VR with telemedicine or home-based delivery concepts intended to improve accessibility and scalability for older adults and patients with mobility constraints [60]. In contrast, several studies remain clinic- or laboratory-based and emphasize feasibility or usability, platform design, and therapist-friendly monitoring rather than definitive efficacy [67,69,70]. Home-oriented delivery is also reflected in at-home computerized executive-function training designs that use immersive VR environments as ecologically valid contexts for dual-task assessment of cognition and mobility, extending evaluation beyond standard clinic-based testing [64].

Methodological rigor varies substantially, ranging from pilot and single-blind randomized controlled trials to crossover designs and observational follow-up studies [57,58,66,68]. Control conditions include traditional cognitive training, standard rehabilitation or orientation therapy, and non-VR exercise programs, which help isolate VR or AR-specific contributions but also introduce heterogeneity in active ingredients and “dose” [61,62,66,71]. Outcome batteries typically combine global cognitive screening tools with domain-specific neuropsychological tests (executive function, attention, and memory), and—less consistently—functional end points (ADL or IADL), mood, or user experience (UX) measures [57,58,62,66]. Of the total, 2 included articles provide a broader conceptual and methodological context for cognitive training in neurodegenerative disease, highlighting persistent challenges around outcome heterogeneity, transfer to daily functioning, and evidence quality [73,74].

Across populations, many studies report favorable feasibility indicators (acceptability, adherence, and tolerability) and within-group improvements on at least some cognitive measures, particularly when interventions are engaging, adaptive, and anchored in meaningful tasks [61,66,68,70]. However, between-group superiority is not uniform, and several studies are explicitly positioned as pilots or preliminary reports, limiting statistical power and confidence in effect estimates [57,58,60,61]. Reminiscence-oriented immersive VR appears especially relevant for affective and engagement-related outcomes in dementia, with cognitive effects that may be modest and time-limited without maintained intervention [58]. For MCI, combined or multicomponent programs (eg, motor plus

XSL•FO
RenderX

VR-based cognitive training) reflect a practical clinical strategy, although gains in daily functioning are not consistently observed [66,71].

Overall, the subgroup supports the feasibility of delivering cognitively oriented interventions using VR or AR across diverse older and neurological populations, but comparability is constrained by heterogeneity in immersion level, task design, intervention dose, and outcome selection [57,59,67,70]. To strengthen future quantitative synthesis, priorities include (1) harmonized core outcome sets (global cognition+domain-specific measures+functional end points), (2) standardized reporting of intervention dose and fidelity, (3) longer follow-up to assess durability, and (4) clearer mapping between task mechanics and hypothesized transfer targets [73,74]. Finally, protocol-driven work targeting cognitive vulnerability in perioperative settings illustrates ongoing diversification of VR applications, but such studies should be clearly separated from dementia or MCI rehabilitation pathways when interpreting population-specific clinical relevance and implementation needs [65].

Motor-Cognitive and/or Exergaming VR Interventions in Neurorehabilitation

Table 2 summarizes studies in which VR-based approaches combine explicit cognitive demands with concurrent motor engagement (eg, stepping, balance training, cycling, or exergaming) and are applied in neurorehabilitation-relevant contexts. The included evidence spans community-dwelling older adults with cognitive frailty [75], individuals with MCI [76,77], mild AD (protocol) [78], and poststroke cognitive disorders [79,80]. Overall, delivery most often relies on nonimmersive or semi-immersive setups, typically screen-based interaction combined with motion tracking or instrumented exercise equipment [81].

XSL•FO
RenderX

Table . Motor-cognitive and/or exergaming VR[a] interventions in neurorehabilitation.

Study (Year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
Kwan et al (2024) [75]	Community-dwelling older adults with cognitive frailty (physical frailty+MCI[b]), no dementia	Multicenter, assessor-blinded parallel-group RCT[c] (VRMCT[d] vs usual care); N=293 (≈146 VRMCT, 147 control)	Nonimmersive VR motor-cognitive exergame platform with motion sensing; nonimmersive group-based VR motor-cognitive exercises; 16 one-hour sessions led by interventionists	16×60 min; 2 times per week; 8 weeks Domains: Global cognition, executive function, memory, interference control VRMCT significantly improved global cognitive function vs usual care; some benefits in executive measures
Fu and Wang (2025) [76]	Patients with MCI treated in a neurology department	Randomized controlled clinical observation (VR+acupuncture vs VR only); N=48 (24 VR+acupuncture; 24 VR only)	VR rehabilitation training system with virtual scenes plus acupuncture in treatment arm; Non-immersive; Control: VR training only; Treatment: same VR program plus scalp/eye acupuncture during each session	VR training protocol described qualitatively; acupuncture needles retained ~30 min each session Domains: Global cognition (MMSE[e], MoCA[f]), daily living skills Treatment group showed larger gains in MMSE and MoCA and higher overall efficacy vs VR-only control
De Luca et al (2022) [82]	Single patient with Nasu–Hakola disease (TREM2-related dementia) and motor-cognitive deficits	Single-case crossover (conventional rehabilitation phase vs conventional + VRRS[g] phase); N=1	VRRS with cognitive and physical modules; Nonimmersive Phase 1: standard physical + cognitive training; Phase 2: 30 speech + 30 PT sessions plus 60 VRRS sessions (cognitive + motor)	VRRS phase: 3 sessions/week, 60 min; total 60 VRRS sessions after 12 weeks of conventional rehab. Domains: Attention, memory, executive function; affective symptoms Improved cognitive tests and marked reduction in anxiety, apathy, indifference, and depressive symptoms after VRRS phase
Huber et al (2024) [80]	Adults in chronic phase after stroke with residual motor or cognitive deficits	Single-blind parallel-group RCT protocol (exergame + usual care vs usual care); Target N=38 chronic survivors of stroke	Concept-guided personalized motor-cognitive exergame training using video game–based tasks; Nonimmersive Intervention group performs exergame training in addition to usual care; control continues usual care only	2 times per week; 30-40 min; 12 weeks (24 sessions) Domains: Global cognition (MoCA) and domain-specific cognitive functions Results pending; expected superior improvement in cognition and gait in exergame group
Faria et al (2023) [83]	Adults with acquired brain injury, dementia, or age-related cognitive decline (conceptual target)	Methods or architecture paper (no clinical trial); Not applicable	VR-based activities-of-daily-living simulations integrated with AI-driven user profiling and adaptation; Non- or semi-immersive VR ADL environments Framework where VR ADL tasks are adapted in real time using AI models to personalize neurorehabilitation	Not fixed; example scenarios only (conceptual) Domains: Executive function, attention, memory, functional cognition No empirical outcomes; framework is intended to support individualized improvement in functional cognition

XSL•FO
RenderX

Study (Year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
Manenti et al (2024) [77]	Adults with MCI from multiple Italian centers	Multicenter randomized active-controlled trial with five treatment arms; N=109 participants with MCI allocated to 5 groups	Clinic- and home-based VRRS cognitive training combined with anodal or sham tDCS[h], or TAU; Nonimmersive VR / telerehabilitation All active arms: 12 × 60-min face-to-face VRRS sessions; some arms also receive anodal tDCS and/or 12 weeks of home VRRS telerehabilitation or unstructured activity	Clinic VRRS: 12 × 60 min over 4 weeks; plus arm-specific telerehabilitation or unstructured… Domains: Episodic memory (primary), plus global cognition and other… VRRS + anodal tDCS followed by VRRS telerehabilitation produced the largest and most durable gains in episodic memory; VRRS-based protocols outperformed TAU
Maggio et al (2025) [84]	Informal caregivers of people with MCI	Nonrandomized single-arm pilot pre-post study; N=10 caregivers	Semi-immersive VR-based physical training program (K-HERO); Semi-immersive Individual low-intensity VR exercise sessions performed while accompanying relatives to rehabilitation	6 sessions of 30–40 min each during caregivers' clinic visits Domains: No direct cognitive training; outcomes focus on psychological well-… Not applicable; reported reductions in avoidance and social-support–based coping and good usability
Hwang et al (2021) [81]	Community-dwelling older adults with age-related cognitive and gait concerns	Randomized controlled trial (VRCT+locomotor vs tabletop cognitive training); N=18 (9 VRCT[i]+locomotor, 9 control)	Semi-immersive VR cognitive training system combined with locomotor tasks; Semi-immersive experimental group performs VR cognitive tasks while stepping or locomoting; control receives tabletop cognitive activities with the same schedule	30 min/day, 3 times per week for 6 weeks (18 sessions) Domains: Processing speed, attention, working memory, and executive function VR group showed greater improvements in executive and attention measures (eg, TMT-A, Digit Span backward) vs control
Contrada et al (2025) [79]	Adults in the chronic phase after ischemic stroke with cognitive impairment	Within-subject longitudinal study (single group; pre, post, 6-month follow-up); N=18	Home-based multidomain cognitive training via VRRS Tele-NeuroRehabilitation platform; Nonimmersive VR or telerehabilitation Patients perform therapist-prescribed VRRS cognitive exercises at home with remote monitoring and feedback	5 sessions/week, 60 min each, for 4 weeks (20 sessions) Domains: Logical reasoning, attention, executive function, memory, language Significant improvements in working memory and language abilities and reduction in depressive symptoms, maintained at 6 months
Bourrelier et al (2021) [85]	Frail older adults with cognitive impairment (MMSE 15–28) and slow gait (<0.65 m/s)	Randomized controlled trial (VR balance training vs Nintendo Wii); N=22 analyzed (from 37 randomized)	VR balance-training system using game-like upper-limb and postural tasks; Likely semi-immersive (screen-based) Ten VR balance-training sessions vs ten Wii-bowling sessions over 3 weeks; APAs assessed using EMG	10 sessions in 3 weeks (~3-4 sessions/week); session length not clearly reported Domains: Not primarily cognitive; cognitive engagement through interactive… Main effects on postural adjustments; VR training improved APA timing and functional reach more than Wii

XSL•FO
RenderX

Study (Year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
Gambella et al (2022) [78]	Community-dwelling adults with mild Alzheimer disease living at home with caregiver	Randomized controlled trial protocol (VR cycling + CCT[j] vs standard bike + CCT); Planned N=78 (39 VR+CCT; 39 control)	jDome VR cycling system combined with computerized cognitive training; Semi-immersive VR (projected dome environment) Both groups receive computerized cognitive training; experimental group cycles in jDome VR, control uses standard exercise bike	16 × 60-min sessions; 2×/week over 8 weeks with 3-month follow-up Domains: Global cognition (MMSE, ADAS-Cog) plus quality of life, mood, behavior Outcomes not yet reported; aim is stabilization or improvement of global cognition in mild AD
Tang et al (2025) [86]	Adults from the general public with no dementia or caregiving history (empathy training target)	Controlled pre-post study (VR dementia tour vs control); N=80 adults (visual display terminal vs control)	Fully immersive VR dementia tour using interactive narrative persona; Fully immersive Single VR session simulating sensory and cognitive impairments of dementia; compared with non-VR control condition	Single VR session (duration not clearly specified) Domains: Not rehabilitation of patients; focuses on dementia knowledge and… VDT participants showed greater improvements in dementia knowledge and empathy compared with control

[a]VR: virtual reality.
[b]MCI: mild cognitive impairment.
[c]RCT: randomized controlled trial.
[d]VRMCT: virtual reality motor-cognitive training.
[e]MMSE: Mini-Mental State Examination.
[f]MoCA: Montreal Cognitive Assessment.
[g]VRRS: Virtual Reality Rehabilitation System.
[h]tDCS: transcranial direct current stimulation.
[i]VRCT: virtual reality cognitive training.
[j]CCT: computerized cognitive training.

Across intervention studies, the motor-cognitive rationale is commonly framed as training attention, working memory, and executive control under dual-task or enriched sensorimotor conditions. In a large multicenter randomized controlled trial among older adults with cognitive frailty, structured VR motor-cognitive training yielded improvements in global cognition compared with usual care [75]. Smaller trials in community-dwelling older adults further suggest that coupling cognitive tasks with locomotor activity can produce concurrent gains in cognitive performance and mobility-related measures (eg, walking speed), supporting the premise that integrated training may facilitate transfer to functional capacities [81]. Balance-oriented VR interventions target motor control processes that are sensitive to cognitive status; for example, VR-based training improved anticipatory postural adjustment timing relative to an active exergaming comparator (Nintendo Wii), indicating potential benefits for neuromotor preparation in frail older adults with cognitive and motor deficits [85].

A second theme concerns scalability, personalization, and continuity of care. Home-based telerehabilitation delivered via VR rehabilitation platforms with remote monitoring enables high-frequency practice and may support maintenance of gains in chronic poststroke populations [79]. Several contributions in this group are protocols rather than completed efficacy trials, reflecting ongoing consolidation of feasibility and methodological standards. The Physical Exercise and Mobile Cognitive Stimulation protocol, for instance, formalizes a concept-guided personalized motor-cognitive exergame program for chronic stroke, with global cognition (Montreal Cognitive Assessment [MoCA]) as a primary end point and gait or dual-task metrics as complementary outcomes [80]. In AD, the jDome protocol integrates VR-supported cycling with concurrent cognitive training to test whether adding a VR exercise context enhances outcomes beyond conventional exercise plus cognitive training [78]. Complementing these clinical studies, an AI-driven methodology paper proposes adaptive profiling and real-time task adjustment within VR-based ADLs scenarios to address heterogeneity in impairment severity and learning rates [83].

Some studies combine VR training with adjunct modalities, which may be clinically pragmatic but complicate causal attribution. In MCI, a multicenter randomized active-controlled study reported long-lasting improvements in episodic memory when VR-based cognitive treatment was combined with transcranial direct current stimulation and followed by telerehabilitation, suggesting that technology-enabled continuation at home may contribute to long-term effects [77]. Another clinical observation compared VR training alone with VR combined with acupuncture and reported larger improvements on screening measures (Mini-Mental State Examination [MMSE] or MoCA) in the combined arm [76]. In a rare neurodegenerative disease, a case report described a multimodal program incorporating VR rehabilitation and reported improvements across cognitive and neuropsychiatric

XSL•FO
RenderX

measures, although the single-case design and concurrent therapies limit generalizability [82].

Several included publications provide implementation-relevant context rather than direct efficacy evidence. A cross-sectional analysis highlights the co-occurrence of cognitive impairment and psychological morbidity among survivors of stroke in rehabilitation, underscoring the need for scalable cognitive-motor rehabilitation approaches [87]. An opinion article summarizes recent VR developments for poststroke cognitive impairment and discusses practical considerations and limitations of the current evidence [88]. A national clinician survey similarly characterizes adoption, perceived value, and barriers to technology-enabled rehabilitation (including VR and telerehabilitation), informing feasibility and scalability considerations [89]. Beyond patient rehabilitation, an immersive "dementia tour" design study illustrates broader VR content development around dementia and highlights usability and engagement considerations that may be transferable to therapeutic design [86]. In the wider care ecosystem, a pilot study delivering low-intensity VR exercise to caregivers emphasized acceptability and potential psychosocial benefits, although it did not directly address patient cognitive outcomes [84].

In summary, motor-cognitive and exergaming VR interventions show promising signals across cognition- and mobility-relevant outcomes, but the evidence remains heterogeneous in platforms, immersion, intervention dose, and outcome selection [81,85]. The strongest inferences arise from larger randomized designs [75,77], whereas smaller trials, single-group telerehabilitation studies, and protocols primarily support feasibility and hypothesis generation [78-80]. Future work would benefit from harmonized cognitive end points beyond screening tools, explicit reporting of motor-cognitive task composition and progression, and longer follow-up to test durability and transfer to real-world functioning.

Usability, Acceptability, and UX of VR or AR Systems

Across the included studies, usability and acceptability (Table 3) were treated as primary feasibility outcomes rather than secondary implementation considerations, reflecting the field's current emphasis on early-stage validation of immersive and semi-immersive interventions. Evidence spans community-dwelling older adults and mixed-ability samples [90-92], clinical groups with MCI [18,93-98], AD [99], Parkinson disease [100], and dementia care contexts incorporating carers and professionals [101,102]. Collectively, these studies show that VR-based approaches are typically perceived as engaging and feasible when sessions are structured, onboarding is provided, and interaction demands are calibrated to users' capabilities [18,90,91,93,94].

XSL•FO
RenderX

Table . Usability, acceptability, and UX[h] of VR[b] or AR[c] systems.

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose or key cognitive outcomes
Chau et al (2021) [90]	Community-dwelling older adults and people with various disabilities; mixed physical and cognitive abilities	Single-arm pretest–posttest feasibility study; N=NR[d]	Non-immersive motion-tracking VR games targeting upper or lower limb movements, cognitive tasks, community-living skills, and relaxing scenes (exergaming)	Approximately 30-min sessions, 3 times per week for 6 weeks in community or service centers Physical and cognitive function, mood, quality of life, usage statistics, adverse events Preliminary improvements in some physical or cognitive outcomes; high acceptability and good adherence; few adverse events → VR training feasible for heterogeneous older and disabled
Chuang et al (2025) [103]	Community-dwelling older adults, mainly without diagnosed dementia but with age-related cognitive concerns	Single-arm pretest–posttest pilot study; N=NR	Fully immersive HMD[e]-based VR leisure and cognitive games; standing and seated interaction with light exergaming elements	Multiple VR sessions over several weeks in laboratory or clinic setting Cognitive performance (attention, memory, and executive function), self-reported enjoyment and usability, feasible and safe; participants enjoyed the intervention; some mild cybersickness or discomfort; preliminary cognitive improvements but no definitive efficacy conclusions
Kim et al (2025) [91]	Community-dwelling older adults, mostly cognitively intact or with mild decline	Single-arm feasibility and acceptability study; N=NR	Immersive HMD VR puzzle tasks requiring visuospatial problem-solving and manual interaction; low-to-moderate motor demands	One or several short VR sessions in a controlled environment Usability ratings, enjoyment, motion sickness, basic cognitive/engagement measures VR puzzles acceptable and engaging; older adults could use the system with guidance; main issues were headset comfort and controller complexity
Hassandra et al (2021) [93]	Older adults with MCI[f]	Mixed methods feasibility study (quantitative plus qualitative); N=NR	Semi-immersive VR exergame combining cycling or stepping with cognitive tasks (dual-task motor-cognitive training)	Repeated training sessions over several weeks in outpatient or rehabilitation setting Physical performance, cognitive outcomes, usability metrics, qualitative UX motor-cognitive exergaming feasible and acceptable for MCI; participants reported enjoyment and perceived benefit; some technical or usability issues;

XSL•FO
RenderX

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose or key cognitive outcomes
Kua et al (2025) [94]	Older adult participants with MCI	Feasibility randomized pilot trial (VR vs control or alternative training); N=NR	Non- or semi-immersive VR tasks simulating everyday activities with cognitive focus and limited motor interaction	Multisession program over several weeks in a clinical setting; neuropsychological tests (memory, attention, and executive function) and feasibility indicators VR group showed at least comparable and, in some domains, better cognitive outcomes than control; acceptable adherence and dropout; mild fatigue or discomfort in some participants
Perra et al (2024) [96]	Older adults with MCI	Feasibility randomized controlled trial – study protocol; N=NR	VR-based cognitive remediation platform with everyday-like tasks and motor interaction (navigation and object manipulation)	Planned multiweek VR program with repeated sessions in clinical or outpatient setting Planned: cognitive function, daily functioning, quality of life, usability, adherence Protocol outlines rationale, design, and feasibility end points; emphasizes user-centered design, safety monitoring, and adherence for older adults with MCI
Latella et al (2024) [18]	Patients diagnosed with MCI	Single-arm feasibility and usability study; N=NR	Nonimmersive VR platform simulating everyday environments (eg, shopping and navigation) to train ecological cognitive skills; limited motor components	Series of VR sessions over several weeks in clinic or rehabilitation center Ecological cognitive performance, standardized cognitive tests, usability, and satisfaction Most participants with MCI could use VESPA[g] 2; tasks perceived as ecologically valid; some technical difficulties and need for therapist guidance; early signals of improvement in ecolog…
Maggio et al (2025) [100]	Patients with mild-to-moderate Parkinson disease, with cognitive involvement	Pilot randomized controlled trial (VR telerehabilitation vs control or standard care); N=NR	Home-based VR telerehabilitation platform for cognitive tasks with simple motor interaction	Scheduled remote VR sessions at home over several weeks Cognitive outcomes, Parkinson-specific measures, quality of life, feasibility, and caregiver-related measures Telerehabilitation via VR was feasible and well-accepted; adherence adequate; cognitive outcomes showed promising trends; remote delivery reduced travel burden

XSL•FO
RenderX

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose or key cognitive outcomes
Jeong et al (2025) [99]	Patients with mild-to-moderate Alzheimer disease	Single-arm preliminary feasibility study; N=NR	Nonimmersive VR cognitive training tasks (eg, navigation and memory) with simple motor interaction	Short multisession VR program in clinical or day-care setting Cognitive test scores, behavioral responses, usability indicators, adverse events VR training was feasible for selected patients with Alzheimer disease with adequate support; some had difficulties with navigation and task understanding; no serious adverse events;
Mondellini et al (2022) [95]	Older adults with MCI from Estonia and Italy	Cross-sectional comparative study between countries; N=NR	Immersive HMD VR cognitive task with mild motor demands (head or hand movements)	Single VR session in laboratory or clinic setting UX, presence, comfort, motion sickness, basic task performance Both groups completed the task; cultural and individual differences in UX and motion sickness; overall feasibility acceptable with supervision
Munoz et al (2025) [92]	Older adults with varying cognitive status using VR games	Observational data-driven study using classifier models; N=NR	VR exergame or cognitive game platform with motion-tracked head movements; focus on kinematic data rather than training efficacy	Single or repeated gaming sessions in controlled environment Head kinematics, subjective game UX, cognitive test scores Classifiers could link kinematic patterns to cognitive status and UX; supports feasibility of VR sensor data as digital biomarkers;
Stasolla and Di Gioia (2023) [97]	Individuals with mild neurocognitive impairment and their caregivers	Studying Usability and acceptability (conceptual and small sample); N=NR	Early-stage VR system where reinforcement learning adapts cognitive training parameters; some motor interaction	Limited number of sessions or system walkthroughs in clinic or laboratory Usability ratings, satisfaction, perceived usefulness, caregiver feedback Patients and caregivers generally accepted the RL-VR concept; main issues were technical or interface-related
Matsangidou et al (2023) [102]	People living with dementia in care settings	Participatory design study with pilot evaluation; N=NR	VR system for physical rehabilitation (movement tasks or games) with cognitive engagement; non- or semi-immersive	Iterative co-design workshops and pilot VR sessions in care facilities Qualitative feedback from patients, staff, caregivers; observations of engagement and physical activity Co-designed VR rehab matched abilities and preferences; participants could engage in guided VR tasks; staff noted motivational benefits and practical challenges

XSL•FO
RenderX

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose or key cognitive outcomes
Gokani et al (2025) [101]	Informal carers and health or social care professionals working with people with dementia	Qualitative study using interviews or focus groups; N=NR	Various VR scenarios (often immersive or 360°) used for stimulation, reminiscence, and engagement	Discussion plus limited VR demonstrations: no fixed dose Perceived usefulness, risks, ethical issues, feasibility in routine care VR seen as promising for engagement and reminiscence; concerns about cost, motion sickness, staff training, and integration into routine care;
Chen et al (2025) [104]	Community-dwelling and/or institutionalized older adults, mixed cognitive status	Qualitative or mixed methods user-experience study; N=NR	Immersive VR experiences (sightseeing, nature, and simple interactive content) with low physical intensity	One or a few VR sessions in laboratory, community, or residential settings UX, presence, enjoyment, discomfort, perceived value Most older adults were curious and engaged; VR is generally acceptable, though some anxiety, disorientation, or cybersickness; stresses gradual introduction and ergonomic adjustments
Siette et al (2024) [105]	Adults and/or older adults using the LEAF CAF VR module	Think-aloud qualitative usability study; N=NR	VR training or assessment module with cognitive and functional tasks; interaction via controllers or gestures	Single-session interaction with the module in a laboratory Usability problems, interaction patterns, perceived workload, user feedback Identified specific usability barriers (interface complexity, confusing feedback, and navigation issues) and facilitators
Pedroli et al (2022) [106]	Adults or older adults with cognitive impairment using home-based training	Usability and acceptability pilot study; N=NR	Nonimmersive mobile or tablet application with structured cognitive exercises, possibly using enriched media but no full VR HMD	Home-based use over several weeks; self-administered with minimal or remote supervision Task completion, app usage, usability questionnaires, satisfaction Home-based app generally usable; adherence depended on support and motivation; participants appreciated flexibility; technical literacy and device issues were barriers
Tseng and Giau (2022) [98]	Older adults evaluating potential living environments	Feasibility and usability study; N=NR	Non- or semi-immersive VR walkthroughs of housing or environmental designs; focus on navigation and environmental fit	One or a few VR walkthrough sessions in laboratory or facility Perceived safety and usability, environmental preferences, motion sickness, orientation VR pre-occupancy evaluation was feasible; older adults could meaningfully comment on design features; some required support with navigation and controls

XSL•FO
RenderX

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose or key cognitive outcomes
Ghosh et al (2022) [107]	Older adults with MCI participating in paired or social activities	Iterative user-centered design study with small pilots; N=NR	Robot-mediated interaction platform (non-VR) supporting cognitive and social activities with light motor components	Multiple design–test cycles with sessions in laboratory or care setting Usability, acceptability, engagement, perceived social and cognitive benefits Older adults with MCI could engage with robot-mediated paired tasks; emphasizes gradual complexity, clear feedback, and social aspects
Haley et al (2021) [108]	Adults with acquired apraxia of speech	Feasibility and therapeutic-effect pre-post study (sometimes with small comparison group); N=NR	Non-VR tablet or computer-based speech therapy program with autonomy-supportive features; motor speech rehabilitation	Multiple therapy sessions over several weeks in clinical setting Speech production measures, communication outcomes, adherence, and user satisfaction Feasible with positive therapeutic effects on speech measures; autonomy-supportive design enhanced engagement and motivation

[a]UX: user experience.

[b]VR: virtual reality.

[c]AR: augmented reality.

[d]NR: not reported.

[e]HMD: head-mounted display.

[f]MCI: mild cognitive impairment.

[g]VESPA: Virtual Environment for Spatial Processing Assessment.

Quantitative usability or acceptance assessments commonly relied on standardized questionnaires (most frequently the System Usability Scale and related acceptance models), with generally favorable perceptions reported by older end users in both community and clinical settings [91,93,94,103]. Importantly, perspectives sometimes diverged between stakeholders. For example, older adults tended to report high enjoyment and ease of use in fully immersive, leisure-based cognitive training, while professional raters judged overall usability more conservatively [103]. In technology-driven cognitive rehabilitation platforms, feasibility was often demonstrated through completion or adherence alongside positive usability ratings, supporting the practicality of repeated-session protocols in outpatient or community environments [18,91,93,94]. At the same time, measurement approaches were not always comprehensive; the Virtual Environment for Spatial Processing Assessment 2.0 feasibility work explicitly cautioned that relying on a single usability scale is insufficient to capture the multidimensional nature of feasibility, usability, and UX in complex VR rehabilitation systems [18]. This aligns with the broader need to report usability, tolerability, workload, and adverse events in parallel, particularly for populations at risk of fatigue, dizziness, or anxiety during immersion [93,99,100].

Qualitative and mixed methods evidence further clarifies what drives (or undermines) positive UX. Think-aloud evaluation of a VR module identified concrete interaction and interface-level issues that can influence comprehension, confidence, and task flow, underscoring the value of formative testing before efficacy trials [105]. Broader experiential accounts emphasized that perceived realism, meaningfulness of activities, and alignment with older adults' routines shape engagement and perceived usefulness [92,103,104]. Cross-cultural comparison during an immersive VR cognitive task highlighted that UX is not solely determined by hardware or task difficulty, but also by contextual factors that may influence comfort and interpretation of the virtual environment [95]. Complementing self-report, head-kinematics–based modeling demonstrated that behavioral signals during VR gameplay can be leveraged to characterize UX and its associations with cognitive factors, suggesting a pathway toward more objective, continuous UX monitoring in future systems [98].

Several studies explicitly adopted human-centered or participatory design strategies, reinforcing that usability in older and cognitively impaired populations benefits from iterative refinement and stakeholder inclusion. Participatory design in dementia-related VR physical rehabilitation foregrounded feasibility in real-world care contexts and supported tailoring to users' preferences and constraints [102]. In parallel, carers' and professionals' perspectives emphasized practical barriers and facilitators to uptake (eg, training requirements, perceived appropriateness, and care workflows), indicating that acceptance is not only an end user property but also an organizational and implementation outcome [101]. Iterative user-centered

XSL•FO
RenderX

development of robot-mediated paired activities likewise illustrates how repeated testing cycles can be used to refine interaction patterns for older adults with MCI, particularly when social and motivational dynamics are central to the intervention experience [107].

Finally, the included set suggests an emerging trend toward adaptive and home-oriented solutions, while also revealing unresolved UX challenges. Combining reinforcement learning with VR for mild neurocognitive impairment reflects a move toward personalization and dynamic difficulty regulation, with usability assessment extending to both patients and caregivers [97]. Home-based or remotely oriented cognitive rehabilitation tools (including app-based platforms) further broaden access but amplify the importance of intuitive interaction, autonomy support, and long-term engagement over time [96,106,108]. Overall, the evidence indicates that VR or AR systems are frequently acceptable and usable in controlled pilots, but comparability is limited by heterogeneous UX constructs and reporting practices. Future trials should therefore adopt multidimensional usability frameworks, report stakeholder-specific acceptance (patients, carers, and clinicians), and integrate objective interaction markers alongside standardized questionnaires to better explain adoption and adherence [18,98,100,101].

VR or AR for Diagnosis and Assessment (ADL or IADL, Spatial Navigation, and Functional Skills)

The studies in this subgroup leverage VR or AR (and closely related interactive 3D systems) primarily as measurement tools to assess clinically relevant cognitive-functional constructs with greater ecological validity than traditional clinic-based tests (Table 4). A clear concentration is observed in (1) spatial navigation and orientation as sensitive markers of age-related decline and MCI-related impairment, and (2) functional skills or ADL-IADL performance captured through structured, everyday-like scenarios. Across the subgroup, VR systems are used either as standalone diagnostic probes or as performance-based end points embedded within feasibility, validation, or intervention protocols, with heterogeneous outcomes and reporting practices limiting direct cross-study comparability.

XSL•FO
RenderX

Table . VR[a] or AR[b] for diagnosis and assessment (ADL[c] or IADL[d], spatial navigation, and functional skills).

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
Mccracken et al (2025) [109]	Healthy younger adults vs healthy older adults; no diagnosed cognitive impairment	Cross-sectional experimental study; within-subject comparison of real vs VR homing; N≈40 (younger+older adults)	VR navigation task (HMD[e]) plus matched real-world corridor task—Fully immersive HMD VR Participants performed homing trials in both real and virtual replicas of an indoor environment; accuracy and path metrics compared across age groups and modal	Single laboratory session with multiple trials in each condition Domains: spatial navigation; path integration; visuospatial memory Outcome: older adults showed reduced homing accuracy in both environments; VR reproduced age-related navigation deficits observed in the real world.
Tuena et al (2024) [110]	Older adults with amnestic MCI[f] and cognitively healthy controls	Experimental study with between-group comparison of navigation aids; N≈30-40 (MCI+controls)	Desktop VR maze or navigation task—nonimmersive desktop VR Participants learned routes in a virtual environment with different navigation aids (bodily vs visual-cognitive cues), followed by spatial recall tests	Single experimental session with several learning–recall blocks Domains: spatial memory; spatial navigation; executive functions Outcome: Bodily and visual-cognitive aids improved spatial recall, particularly in participants with MCI.
Silva et al (2023) [111]	Community-dwelling older adults without dementia	Feasibility and validation study; N≈40-50 older adults	Immersive VR outdoor-navigation scenario (SOIVET system)—fully immersive HMD VR; older adults explored a virtual neighborhood and performed egocentric orientation tasks (pointing or walking to remembered locations); usability and cybersickness	Single VR assessment session Domains: Egocentric spatial orientation; visuospatial memory; attention Outcome: Immersive VR assessment was feasible, well tolerated, and showed meaningful variability in egocentric orientation.
Tuena et al (2024) [112]	Older adults with MCI	Proof-of-concept interventional trial with VR training vs control; N≈30 MCI participants (training vs control)	Desktop VR navigation game with embodied interaction—nonimmersive or low-immersion VR Participants with MCI completed repeated embodied VR navigation tasks over several weeks; pre-post cognitive and navigation measures were collected	Multisession VR training over several weeks Domains: Spatial navigation; episodic memory; executive functions Outcome: VR-based embodied navigation training improved spatial navigation and some cognitive scores relative to controls.
Panerai et al (2023) [113]	Adults and older adults with major neurocognitive disorder (degenerative and nondegenerative etiologies)	Nonrandomized interventional study (VR training vs. control); N=67 (40 experimental; 27 control)	Tablet or PC-based nonimmersive VR applications simulating ADL or IADL tasks—nonimmersive VR Experimental group trained four functional living skills in virtual scenarios, with in vivo performance of the same tasks assessed at multiple time points	Structured multisession VR training with three assessment time points Domains: Functional living skills (ADL or IADL); executive functions; attention; memory Outcome: VR-trained participants improved both in virtual tasks and in real-world execution of the trained functional skills.

XSL•FO
RenderX

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
Stasolla et al (2024) [114]	People with Alzheimer disease and cognitively healthy controls (where applicable)	Comparative study of traditional vs VR-based or digital storytelling approaches; N=NR[g] (to be extracted)	Nonimmersive VR or computer-based digital storytelling platform—nonimmersive VR or screen-based Patients engaged in digital storytelling activities delivered via VR-inspired interactive environments; outcomes compared with more standard assessment or rehabilitation	Multisession intervention over several weeks (details to be specified from full text) Domains: Episodic memory; autobiographical memory; executive functions; language Outcome: VR or digital-storytelling condition showed comparable or better gains in some cognitive and functional measures relative to standard approaches.
Mirmiran and Moussavi (2026) [115]	Institutionalized older adults with advanced dementia	Single-arm pilot intervention study; N=10 (9 completing ≥20 sessions)	Semi-immersive driving simulator with large display and steering controls—semi-immersive VR Participants repeatedly used a VR driving simulator presenting pleasant outdoor environments; quality of life and behavioral symptoms were monitored	Several sessions per week over approximately four months Domains: Well-being and engagement (primary); visuomotor coordination; secondary behavioral outcome Outcome: Feasible in advanced dementia; associated with improved quality-of-life measures and reduced behavioral disturbance in many residents.
Varela-Aldás et al (2022) [116]	Healthy adults (middle-aged and older)	Feasibility and validation study; N≈30 healthy participants	VR kitchen or cupboard scenario using an HMD—fully immersive HMD VR Participants memorized the location of everyday objects in virtual cupboards and later recalled or placed them; task metrics were related to standard neuropsychology	Single VR assessment session (~30-45 minutes) Domains: Everyday episodic memory; prospective memory; executive functions Outcome: Cupboard task was feasible, well accepted, and produced memory indices correlated with traditional tests, supporting ecological validity.
Bhargava et al (2025) [117]	Game developers and neuroscience/cognition researchers (no patient sample)	Qualitative mixed methods study (interviews, workshops); N=approximately 20-30 professionals	Digital serious games for cognition (including VR or 3D concepts)—mainly nonimmersive; conceptual discussion of multiple levels of immersion, explored design processes, constraints, and opportunities when cocreating games aimed at assessing cognition	Not applicable (design or consultation study) Domains: Multiple domains: attention, memory, executive functions, processing speed Outcome: Identified key design principles and tensions between playability and psychometric rigor for cognition-assessing games.

XSL•FO
RenderX

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
Mahncke (2025) [118]	Conceptual paper; focuses on clinical and research populations reporting cognitive complaints	Commentary or methodological article; N=not applicable	Not specific to one technology; implications for digital and VR-based assessments—not applicable, provides guidance on interpretation of self-reported cognitive complaints and their relation to objective performance measures	Not applicable Domains: Metacognition; subjective cognitive complaints; self-awareness of decline Outcome: Argues that self-report alone is insufficient as a diagnostic tool and should be combined with objective (including VR) assessments and longitudinal follow-up.

[a]VR: virtual reality.

[b]AR: augmented reality.

[c]ADL: activities of daily living.

[d]IADL: instrumental activities of daily living.

[e]HMD: head-mounted display.

[f]MCI: mild cognitive impairment.

[g]NR: not reported.

Several studies operationalize navigation using route learning, homing, and egocentric orientation tasks. A real-world versus virtual comparison of homing performance indicates that VR can reproduce age-related decrements in navigation or path integration, supporting the construct validity of VR-derived navigation metrics in older adults [109]. Immersive assessment of egocentric orientation in older adults further demonstrates feasibility of HMD-based protocols to quantify orientation performance while capturing usability- and tolerability-relevant factors inherent to immersive testing [111]. In MCI-focused work, desktop-based paradigms examine how navigation task design (eg, bodily and visual-cognitive aids) shapes spatial recall and diagnostic sensitivity, suggesting that the embedded cue structure may meaningfully influence performance and interpretability [110]. Related protocols extend to embodied spatial navigation paradigms where repeated task performance can be systematically tracked, offering a potential route to longitudinal assessment signatures (eg, learning curves, variability, and strategy shifts) rather than single-session snapshots [112].

A complementary line of research targets functional capacity via simulated everyday tasks (eg, shopping, medication management, and packing), thereby enabling standardized measurement of instrumental activities that are directly relevant to clinical staging and care needs [113]. Immersive everyday-memory assessment similarly embeds memory demands in familiar household contexts, providing object-location and prospective-memory–like indices intended to reflect real-world failures more directly than decontextualized neuropsychological tasks [116]. Collectively, these approaches illustrate how VR-based assessment can bridge the gap between laboratory cognition measures and functional outcomes, although the extent of transfer to real-world functioning remains variably demonstrated and often depends on task selection and scoring definitions.

In more advanced dementia, formal cognitive testing can be constrained by disease severity; therefore, studies sometimes prioritize pragmatic assessment outcomes, such as engagement, behavioral symptoms, and quality-of-life proxies within repeatable, structured experiences [115]. Other work uses VR-supported or VR-inspired digital storytelling frameworks to compare outcomes against more standard approaches, using the virtual environment as a scaffold for eliciting autobiographical or episodic content and monitoring response patterns [114]. In addition, immersive VR cognitive stimulation trials in AD incorporate performance-based and clinical outcome assessments as end points (even when the primary intent is therapeutic), providing data relevant to feasibility and evaluability of VR protocols in mild-to-moderate dementia [72].

In total, 2 contributions provide interpretive context for assessment-focused VR. A qualitative examination of collaboration between developers and neuroscience researchers highlights tensions between engagement-driven game mechanics and psychometric requirements, underscoring the importance of validity-by-design, telemetry planning, and transparency of scoring logic in cognition-assessing systems [117]. A methodological commentary on interpreting self-report cognition measures reinforces that subjective reports should not be used as standalone diagnostic indicators and motivates combining them with objective performance-based measures—including VR tasks—preferably within longitudinal frameworks [118]. Taken together, the subgroup supports VR or AR as a promising modality for assessing navigation, everyday memory, and functional capacity, but it also emphasizes the need for improved standardization (task definitions, scoring, adverse-event reporting, and benchmark validation) to strengthen synthesis and clinical interpretability across studies.

Conceptual, Methodological, and Guideline Literature: Broader Context

Table 5 summarizes conceptual, methodological, and guideline-oriented publications that provide the broader context

XSL•FO
RenderX

for VR- or AR-based approaches in cognitive rehabilitation and related functional outcomes in older adults. These papers do not primarily aim to estimate intervention efficacy; instead, they shape the field by (1) articulating design and usability guidance for older users, (2) proposing conceptual frameworks and taxonomies for intervention content and mechanisms, (3) outlining methodological pathways for development and evaluation, and (4) mapping research trends to identify gaps and priorities for scalable, reproducible work [2,14-16,21,30,119-126].

XSL•FO
RenderX

Table . Conceptual, methodological, and guideline literature: broader context.

Study (Year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
Abeele et al (2021) [14]	Older adults (aged 55-95 y), diverse cognitive abilities or care settings; first IVR[a] experience	Guideline paper: literature-derived guidelines+empirical qualitative study (interviews; laddering); N=37	Immersive VR[b] (HMD[c]-based IVR); guidelines spanning content/hardware/context immersive (HMD IVR) compiled 67 design guidelines from literature; exposed 37 older adults to first IVR experience and interviewed them to empirically ground or refine guidelines	Design or accessibility or user experience focus; cognitive screening used in sample characterization, empirically grounded design guidance for accessible or engaging IVR for older adults; links IVR attributes to functional or psychosocial benefits
Seifert and Schlomann (2021) [15]	Older adults (focus on adults aged older than 75 y in Europe; heterogeneous abilities and technology experience)	Opinion or short communication (narrative overview)	VR and AR[d] (broad use cases: health gaming, social contact, rehabilitation, and daily-life support) varies (not limited to a specific device class), discusses opportunities and barriers for older adults' VR or AR use; proposes developer or research recommendations (participatory design, and training or support)	General cognition or ADL[e] support context; highlights risks for users with cognitive limitations and cybersickness considerations, synthesizes key barriers (digital divide, skills or access, cybersickness, and environment constraints) and recommends participatory design and user support to improve adoption
Moon and Han (2022) [16]	People with neurological disorders (including cognitive impairment); accessibility for patients underexplored	Perspective paper	Immersive VR (stereo HMD) in neurology (rehabilitation+potential for diagnosis or assessment; metaverse context), immersive (explicitly defines VR as immersive stereo HMD), reviews current state of VR in neurology and proposes future directions, emphasizing VR-based cognitive assessment/diagnosis and accessibility or safety	Working memory, executive function, spatial navigation or spatial memory, cognitive planning (as examples for VR-based daily-life tasks), argues VR is underused for diagnosis or evaluation; proposes VR tasks simulating daily life with richer performance parameters and calls for standardization and diverse normative datasets
He et al (2022) [30]	Publications on VR-based cognitive rehabilitation (2001-2021); diseases include stroke, dementia, MCI[f], etc	Bibliometric analysis (Web of Science; visualization or clustering tools); N=1259 papers	VR for cognitive rehabilitation (field-level mapping) varies (depends on included studies), retrieved WoS[g] records (2001-2021) and analyzed authors, countries, keywords, and trends	Field-level: cognition or cognitive rehabilitation broadly maps research hotspots and trends; highlights telerehabilitation growth and notes lack of consensus on treatment methods and safety concerns (per authors)
Manser and de Bruin (2021) [120]	Older adults with mild neurocognitive disorder; project "Brain-IT" (design/development focus)	Methodological paper+case study applying Motivation–Interaction–Design–Evaluation framework	Exergame-based motor-cognitive training system (Dividat Senso or Senso Flex); gamification and multisensory feedback primarily nonimmersive exergaming (no HMD emphasized), describes step-by-step design or development process guided by MIDE (contextual research, game design or development, evaluation) to create exergame training aimed at halting or reducing	Motor-cognitive training (multiple domains; tailored to user capabilities and therapeutic needs) provides replicable methodological workflow and design requirements to increase usability, acceptance, and adherence of exergame interventions in clinical practice

XSL•FO
RenderX

Study (Year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
Guzmán et al (2024) [21]	Older adults (cognitive rehabilitation context; framework-level; expert review included clinicians and therapists)	Conceptual framework development (Jabareen methodology)+expert review; N=5 experts (framework review)	Serious games for cognitive rehabilitation; emphasizes VR; includes AR as proposed extension primarily immersive VR context (mentions HMD; also discusses AR integration), develops a conceptual framework linking VR, physical activity, social interaction, difficulty adaptation, and content customization	Multiple domains (framework emphasizes multidomain cognitive rehab+real-world scenarios; examples include executive functions and others), Framework intended to guide design and implementation of serious games for older adults' cognitive rehabilitation; highlights iterative, multidisciplinary, participatory development and personalization
Cheng et al (2025) [119]	Older adults (community-dwelling); cognitive and independent living skills training in a home environment simulation	Feasibility study (qualitative interviews after gameplay); N=30	VR serious game ("Brainland") in home-like virtual environment with mini-games Immersive VR (HMD implied) designed a VR game with 6 mini-games targeting learning or cognitive and independent living skills; evaluated feasibility via semistructured interviews after play	Single trial session ~20 minutes (per author), cognitive skills+independent living skills (life skills or ADLs); mentions attention, working memory, processing speed as longer-term targets: Positive perceived usefulness of use and reported competency or cognitive developments; identifies engagement benefits and negative experiences (cybersickness or anxiety) informing best practices

[a]IVR: immersive virtual reality.

[b]VR: virtual reality.

[c]HMD: head-mounted display.

[d]AR: augmented reality.

[e]ADL: activities of daily living.

[f]MCI: mild cognitive impairment.

[g]WoS: Web of Science.

Empirically grounded guidance underscores that acceptability and UX are not peripheral outcomes but prerequisites for any clinically meaningful deployment. Design recommendations for immersive VR tailored to older adults emphasize onboarding, interaction simplicity, comfort, and accessibility as core requirements for both clinical and home use [14]. Broader discussions of VR or AR adoption in later life further highlight structural and individual barriers (eg, technology access, digital skills, perceived relevance, and tolerability), motivating usability-by-design and inclusive implementation strategies [15]. Additional works in this group complement these contributions by further operationalizing usability, safety, and user-centered evaluation considerations relevant to older populations [124,126].

Conceptual papers provide shared vocabularies for describing intervention components (eg, task structure, feedback, engagement mechanisms, and functional relevance), which is essential for comparing heterogeneous VR or AR studies within systematic reviews. A conceptual framework for serious games for cognitive rehabilitation synthesizes design and content considerations intended to support both therapeutic intent and long-term engagement [21]. Related conceptual and synthesis contributions in this set further refine how VR or AR interventions can be specified and categorized, supporting clearer reporting and improved cross-study comparability [122,123].

Methodological publications stress iterative, user-centered development and transparent reporting of technology characteristics (hardware, interaction modality, and immersion level) and intervention parameters (dose, progression, and feedback). A structured methodological approach to exergame development illustrates how therapeutic goals can be aligned with game mechanics and feasibility constraints in older adults with neurocognitive disorders [120]. Complementary methodological contributions in this group further strengthen evaluation pipelines (eg, feasibility-to-trial progression, outcome selection, and implementation-relevant metrics) to reduce variability and improve reproducibility [2,125].

Perspective and bibliometric studies contextualize why certain paradigms, outcomes, and technologies dominate the literature and where underexplored opportunities remain. A neurological disorder perspective highlights both the promise of immersive VR for ecologically valid task simulation and measurement and the practical considerations affecting translation [16]. Bibliometric mapping of VR in cognitive rehabilitation

XSL•FO
RenderX

characterizes publication growth and thematic concentrations, informing priorities, such as methodological standardization, reporting quality, and scalability [30,121]. Additional mapping work (including domain-wide bibliometric analyses) and interaction-technology–focused syntheses (eg, hand-tracking interfaces used in VR-based assessment or intervention) further contextualize research growth and methodological gaps relevant to scalable deployment and measurement fidelity [41,127].

Finally, feasibility-oriented design work demonstrates how the above principles can translate into ecologically meaningful content and UX evidence in older adults, providing practical signals on tolerability and perceived usefulness that can guide refinement and future trial design [119].

Overall, this body of context literature supports three interpretive lenses for the present evidence synthesis, that are (1) usability and acceptability should be treated as core outcomes rather than secondary considerations [14,15,124,126], (2) heterogeneity in study designs and outcomes can be partly mitigated through clearer specification of intervention components and technology characteristics [2,21,122,123,125], and (3) future research should prioritize methodological rigor, standardized reporting, and scalable deployment pathways to move beyond proof-of-concept demonstrations [16,30,119,121].

Discussion

Principal Findings

This systematic review synthesizes evidence on immersive technologies for cognitive rehabilitation across the MCI-dementia continuum, with an explicit focus on usability, therapeutic efficacy, and limitations spanning VR, AR, and CAVE-like embodied systems. Overall, the literature indicates that immersive interventions are feasible and frequently engaging, with signals of benefit across multiple cognitive domains (eg, attention, memory, executive functions, and spatial navigation), but the strength of conclusions remains constrained by substantial heterogeneity in technology, protocols, and outcomes, as well as the predominance of small-sample and early-phase designs [2,4,20,23-26].

Figure 3 provides a schematic overview of the core challenges shaping the interpretation, synthesis, and real-world applicability of immersive XR-based cognitive rehabilitation across the MCI-dementia continuum. The diagram positions the XR intervention at the center and highlights 5 interrelated domains that structure the subsequent discussion. First, therapeutic goals and outcome selection must be stage-sensitive, reflecting fundamental differences between restorative aims in MCI and maintenance-oriented or quality-of-life outcomes in dementia. Second, the figure emphasizes the trade-off between technological immersion and usability, illustrating how interaction demands and interface complexity can confound cognitive outcomes. Third, it underscores the persistent gap between improvements on cognitive test scores and meaningful functional transfer to ADL or IADL. Fourth, methodological rigor and harmonized reporting are identified as prerequisites for cumulative evidence synthesis in a field characterized by heterogeneous protocols and outcome measures. Finally, feasibility, safety, and scalability are highlighted as cross-cutting determinants of effective dose, adherence, and implementation beyond controlled laboratory settings. Together, these interconnected dimensions frame the interpretation of the current evidence base and guide priorities for future research and clinical translation.

Figure 3. Key challenges for immersive extended reality cognitive rehabilitation (VR, AR, or CAVE) across the mild cognitive impairment–dementia continuum: stage-sensitive outcomes, immersion–usability trade-offs, functional transfer (ADL or IADL), methodological or reporting harmonization, and feasibility, safety, and implementation at scale. ADL: activities of daily living; AR: augmented reality; CAVE: Cave Automatic Virtual Environment; IADL: instrumental activities of daily living; MCI: mild cognitive impairment; VR: virtual reality; XR: extended reality.

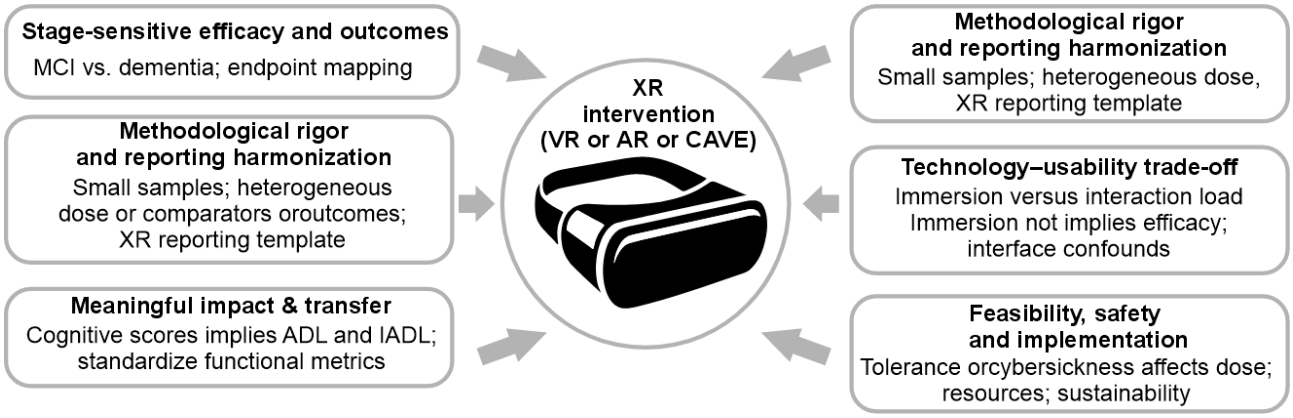

Interpretation of the Evidence Base Across MCI and Dementia

Across MCI, immersive training is commonly positioned as a restorative or capacity-enhancing approach, aiming to improve domain-specific performance (eg, executive control, working memory, and visuospatial abilities) and to strengthen skills relevant to everyday functioning [4,13,20]. In dementia, intervention goals often shift toward maintaining function, supporting compensatory strategies, and enhancing quality of life rather than achieving large gains on conventional neuropsychological end points; this distinction has important implications for outcome selection, intervention dose, and expectations of change [5,6]. Reviews spanning dementia or AD-oriented rehabilitation emphasize the value of ecologically meaningful tasks and the importance of aligning training scenarios with the functional challenges faced by patients and caregivers [5,6,17]. Taken together, the field would benefit from

XSL•FO
RenderX

more explicit stage-sensitive theories of change—what should improve (or be preserved), for whom, and through which mechanisms in MCI versus dementia [16,21].

Technology Matters: Immersion, Interaction Demands, and Ecological Validity

A core promise of immersive systems is ecological validity, the ability to engage cognitive processes within realistic, context-rich activities (eg, navigation, shopping-like scenarios, and multistep functional routines) that may better approximate everyday demands than traditional desktop-based training [13,17,112]. However, greater immersion is not synonymous with better outcomes. Higher immersion can increase interaction complexity and sensory load, which may introduce usability barriers for older adults with cognitive, motor, or sensory constraints and can confound interpretation (eg, performance differences driven by interface demands rather than the targeted cognitive construct) [8,14]. Accordingly, the usability literature emphasizes that interaction design, onboarding, and accessibility features are prerequisites for meaningful interpretation of outcome data, particularly in populations susceptible to fatigue, confusion, or anxiety during first exposure to novel technology [14,15].

AR remains less represented than VR in the cognitive rehabilitation literature, but it is conceptually attractive for supporting in situ cueing, errorless learning, and context-aware scaffolding of everyday activities [8,15,67]. A key research priority is to clarify where AR may offer a comparative advantage (eg, functional cueing and transfer) versus where VR may be preferable (eg, controlled manipulation of stimuli and graded exposure), while using comparable outcome frameworks that allow cross-modality synthesis [21,122].

From Cognitive Test Scores to Meaningful Functional Outcomes

Many studies and syntheses emphasize cognitive test performance; yet, fewer consistently demonstrate robust transfer to ADL or IADL, participation, or caregiver-relevant outcomes [4,17,20]. This gap is particularly consequential for dementia, where functional preservation and symptom management may be primary goals [5,6]. XR platforms are well-suited to embed functional tasks and to collect rich, process-level behavioral data (eg, error patterns, sequencing, and wayfinding efficiency), but these measures are rarely standardized across studies, limiting comparability and interpretability [13,124,126]. Future trials should prespecify clinically meaningful primary end points that map to the target population and intervention rationale and should report functional measures with sufficient detail to enable aggregation and replication [21,122,123].

Usability, Tolerability, and Safety as Determinants of Feasibility and Interpretability

Usability and acceptability are not secondary considerations in this domain; they are prerequisites for feasibility and for interpretable outcome measurement. Design and evaluation guidance for older users emphasizes minimizing interaction complexity, providing clear instructions and feedback, and ensuring comfort and accessibility in immersive experiences [8,14]. Cybersickness and related adverse effects can further influence adherence and exposure (eg, selective withdrawal, shortened sessions, and reduced “dose”), complicating cross-study comparisons and the interpretation of reported changes. Evidence focused on fully immersive VR in older adults indicates that cybersickness is a material barrier that warrants systematic monitoring and reporting, alongside mitigation strategies, such as gradual acclimation, careful locomotion design, and appropriate session duration [50]. More consistent reporting of adverse events, tolerability, and reasons for withdrawal is therefore essential for evaluating real-world feasibility and for understanding which systems and protocols are most likely to translate beyond controlled settings [14,15].

Methodological Limitations and Sources of Heterogeneity

The current evidence base is constrained by several recurring methodological issues, such as (1) small sample sizes and early-phase designs, (2) variability in comparator conditions and control interventions, (3) wide variation in dose and supervision, and (4) heterogeneous outcome measures and reporting practices [2,4,20,23]. These limitations reduce statistical power and complicate meta-analytic synthesis, while also increasing the risk that positive findings reflect context-specific effects rather than generalizable efficacy. Methodological and reporting frameworks emphasize the importance of clear intervention specification, stage-appropriate evaluation (feasibility vs efficacy), and transparent reporting of implementation conditions (training, supervision, equipment, and setting) [119,120,122,123]. The field would benefit from harmonized reporting templates tailored to XR rehabilitation (technology characteristics, immersion level, interaction modality, content design, progression rules, and support requirements), enabling cumulative synthesis across platforms and sites [16,21,122].

Scalability and Implementation in Real-World Care Pathways

Beyond efficacy, the practical deployment of XR interventions depends on scalability constraints, such as costs, maintenance, space requirements (particularly for room-scale systems), clinician time, and the need for caregiver involvement. Field-level syntheses and trend analyses consistently identify implementation, standardization, and scalability as major barriers to broader clinical uptake [16,30,121]. For home or community delivery, studies should evaluate feasibility under realistic constraints (digital literacy, connectivity, supervision, and safety) and should explicitly report the resources required for long-term use (setup time, troubleshooting, and staff training) [15,119]. In parallel, implementation of science constructs (eg, acceptability, feasibility, fidelity, and sustainability) should be incorporated prospectively rather than inferred post hoc [14,120].

Emerging Opportunities: Personalization, Multimodal Sensing, and Mechanistic Evaluation

Several contemporary directions may increase the clinical relevance of immersive rehabilitation. First, adaptive personalization (rule-based or data-driven) may better accommodate heterogeneity in baseline impairment, fatigue,

XSL•FO
RenderX

and learning profiles across MCI and dementia, but it requires transparent specification and validation to avoid "black-box" intervention effects [21,122]. Second, integration with multimodal sensing (eg, physiological measures and behavioral telemetry) may enable more sensitive monitoring of cognitive load, engagement, and progression but raises methodological questions about reliability, privacy, and clinical interpretability [3,125]. Finally, stronger mechanistic evaluation—linking task demands to hypothesized cognitive processes and to functionally meaningful outcomes—is needed to move from promising prototypes to reproducible treatment programs [13,16,21].

Strengths and Limitations of This Review

A key strength of this review is the integrated synthesis of usability, therapeutic efficacy, and limitations across immersive modalities, rather than treating usability as a secondary concern or focusing narrowly on 1 technology class [8,14]. Nevertheless, our conclusions are bounded by the limitations of the underlying literature, including heterogeneity in intervention content and outcomes, and the frequent reliance on pilot studies. As in many rapidly evolving technology domains, publication and selective-reporting biases remain plausible, and inconsistent reporting of implementation factors limits external validity judgments [4,16,122].

Implications for Clinical Practice and Future Research

From a practice perspective, the evidence supports cautious optimism—immersive interventions appear feasible and engaging for many individuals with MCI and, under appropriate support conditions, for some people living with dementia [5,6,14]. However, adoption should be guided by usability screening, conservative session dosing, and systematic monitoring of adverse effects, particularly in fully immersive VR [14,50].

For research, priorities include (1) standardized reporting of technology and intervention components, (2) stage-sensitive outcome frameworks linking cognition to ADL or IADL and quality of life, (3) adequately powered comparative trials with transparent comparator selection, (4) routine usability and safety reporting, and (5) implementation-focused evaluation to support scalability [16,21,30,119-123]. Addressing these challenges will improve the cumulative value of future studies and help transform immersive cognitive rehabilitation from a promising research area into a dependable component of care across the MCI-dementia continuum.

Conclusion

This systematic review synthesized current evidence on immersive technologies (VR, AR, and CAVE) for cognitive rehabilitation in MCI and dementia, with an emphasis on therapeutic outcomes, usability, and translational limitations. Overall, the literature suggests that immersive interventions are feasible and often engaging, with promising signals of benefit across multiple cognitive domains. However, the strength and generalizability of conclusions remain limited by substantial heterogeneity in intervention content, technology and immersion levels, interaction modalities, dosage, and outcome selection, alongside the frequent use of small samples and early-phase study designs.

A key implication is that outcome interpretation cannot be separated from usability and tolerability. In cognitively impaired older adults, interface complexity, onboarding requirements, comfort, and adverse effects can influence adherence and delivered exposure (ie, the effective "dose"), thereby shaping observed performance and functional outcomes. Future studies should therefore treat usability, safety, and acceptability as core outcomes, report them transparently (including adverse events and reasons for withdrawal), and analyze them alongside cognitive and functional measures.

For the next generation of research, priorities include (1) clearer specification and reporting of intervention components (tasks, feedback, progression, and supervision), (2) better alignment of outcomes with clinically meaningful goals, including functional end points (ADL or IADL) and quality-of-life measures, (3) adequately powered comparative trials with well-justified control conditions, and (4) explicit evaluation of implementation and scalability, particularly for home and community deployment. Advancing methodological standardization and reporting consistency will be essential to enable cumulative synthesis and to support translation of immersive cognitive rehabilitation into dependable, real-world care pathways across the MCI-dementia continuum.

Acknowledgments

This work was supported by the project SP2025/019, "Development of algorithms and systems for control, measurement and safety applications XI" of the Student Grant System, VSB-TU Ostrava. During the preparation of this work, the authors used ChatGPT (version 5.2; OpenAI) in order to improve the clarity and grammar of the manuscript draft. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

This article has been produced with the financial support of the European Union under the REFRESH – Research Excellence For REgion Sustainability and High-tech Industries project CZ.10.03.01/00/22_003/0000048 via the Operational Programme Just Transition.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during this study.

XSL•FO
RenderX

Authors' Contributions

JK led the conceptualization and coordination of the extensive review, contributed to methodological design (eligibility criteria and synthesis plan), oversaw screening and extraction consistency, and led drafting and integration of the revised manuscript.
GL contributed clinical domain expertise in dementia/MCI and cognitive rehabilitation, supported refinement of eligibility interpretation across interventions, participated in screening/extraction, and critically revised the manuscript for clinical accuracy.
SB contributed to screening and data extraction, supported interpretation of intervention feasibility and outcomes, and participated in manuscript revision.
RF provided senior clinical-scientific input on outcomes and interpretation and critically revised the manuscript.
AL contributed to interpretation of clinical applicability and feasibility considerations and participated in manuscript revision.
RM contributed to technical reporting rigor and consistency checks (eg, structured synthesis, tabulation coherence) and revised the manuscript.
SM participated in screening/data extraction and supported synthesis of usability/acceptability and translational limitations.
AP contributed to screening/extraction and interpretation of clinical relevance and tolerability findings.
BP supported coordination and critical revision focused on clarity, structure, and reporting completeness.
AQ contributed to screening/data extraction and manuscript revision.
FR provided methodological/statistical consultation for synthesis framing and critically revised the manuscript.
RAS contributed to screening/data extraction and consistency/proofreading checks.
AS provided senior supervision, contributed to clinical interpretation, and critically revised the manuscript.
MP provided overall supervision and project administration, ensured reporting completeness and compliance, coordinated final approval, and led critical revision. All authors reviewed and approved the final manuscript and agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Checklist 1
PRISMA checklist.
[PDF File, 231 KB - jmir_v28i1e84349_app1.pdf]

References

1. World Alzheimer Report 2024: Global changes in attitudes to dementia. Alzheimer's Disease International. 2024. URL: https://www.alzint.org/resource/world-alzheimer-report-2024/ [accessed 2026-02-22]
2. Sasaninezhad M, Moradi AR, Farahimanesh S, Choobin MH, Almasi-Dooghaee M. Enhancing cognitive flexibility and working memory in individuals with mild cognitive impairment: exploring the impact of virtual reality on daily life activities. Geriatr Nurs (Lond) 2024 Mar;56:32-39. [doi: 10.1016/j.gerinurse.2023.12.008]
3. Gkintoni E, Vassilopoulos SP, Nikolaou G, Vantarakis AC. Neurotechnological approaches to cognitive rehabilitation in mild cognitive impairment: a systematic review of neuromodulation, EEG, virtual reality, and emerging AI applications. Brain Sci 2025 May 28;15(6):582. [doi: 10.3390/brainsci15060582] [Medline: 40563754]
4. Tortora C, Di Crosta A, La Malva P, et al. Virtual reality and cognitive rehabilitation for older adults with mild cognitive impairment: a systematic review. AGEING Res Rev 2024 Jan;93:102146. [doi: 10.1016/j.arr.2023.102146] [Medline: 38036103]
5. Son C, Park JH. Ecological effects of VR-based cognitive training on ADL and IADL in MCI and AD patients: a systematic review and meta-analysis. IJERPH 2022 Nov 29;19(23):15875. [doi: 10.3390/ijerph192315875] [Medline: 36497946]
6. Vásquez-Carrasco E, Hernandez-Martinez J, Sepúlveda-Ramírez M, et al. Effectiveness of virtual reality interventions on quality of life, cognitive function and physical function in older people with Alzheimer's disease: a systematic review. Ageing Res Rev 2025 Jul;109:102785. [doi: 10.1016/j.arr.2025.102785] [Medline: 40436354]
7. Zuschnegg J, Schoberer D, Häussl A, et al. Effectiveness of computer-based interventions for community-dwelling people with cognitive decline: a systematic review with meta-analyses. BMC Geriatr 2023 Apr 12;23(1):229. [doi: 10.1186/s12877-023-03941-y] [Medline: 37041494]
8. Margrett JAA, Ouverson KMM, Gilbert SBB, Phillips LA, Charness N. Older adults' use of extended reality: a systematic review. Front VIRTUAL Real 2022;2. [doi: 10.3389/frvir.2021.760064]
9. Roley SS, DeLany JV, Barrows CJ, et al. Occupational therapy practice framework: domain & practice, 2nd edition. Am J Occup Ther 2008;62(6):625-683. [doi: 10.5014/ajot.62.6.625] [Medline: 19024744]
10. Milgram P, Kishino F. A taxonomy of mixed reality visual displays. IEICE Trans Inf & Syst 1994 Dec;E77-D(12):1321-1329 [FREE Full text]
11. Rauschnabel PA, Felix R, Hinsch C, Shahab H, Alt F. What is XR? Towards a framework for augmented and virtual reality. Comput Human Behav 2022 Aug;133:107289. [doi: 10.1016/j.chb.2022.107289]

XSL•FO
RenderX

12. Flavián C, Ibáñez-Sánchez S, Orús C. The impact of virtual, augmented and mixed reality technologies on the customer experience. J Bus Res 2019 Jul;100:547-560. [doi: 10.1016/j.jbusres.2018.10.050]
13. Tuena C, Borghesi F, Bruni F, et al. Technology-assisted cognitive motor dual-task rehabilitation in chronic age-related conditions: systematic review. J Med Internet Res 2023 May 22;25:e44484. [doi: 10.2196/44484] [Medline: 37213200]
14. Abeele VV, Schraepen B, Huygelier H, Gillebert C, Gerling K, van Ee R. Immersive virtual reality for older adults: empirically grounded design guidelines. ACM Trans Access Comput 2021 Sep;14(3):1-30. [doi: 10.1145/3470743]
15. Seifert A, Schlomann A. The use of virtual and augmented reality by older adults: potentials and challenges. Front VIRTUAL Real 2021;2. [doi: 10.3389/frvir.2021.639718]
16. Moon HJ, Han S. Perspective: present and future of virtual reality for neurological disorders. Brain Sci 2022 Dec 9;12(12):1692. [doi: 10.3390/brainsci12121692] [Medline: 36552152]
17. Buele J, Varela-Aldás JL, Palacios-Navarro G. Virtual reality applications based on instrumental activities of daily living (iADLs) for cognitive intervention in older adults: a systematic review. J NeuroEngineering Rehabil 2023;20(1). [doi: 10.1186/s12984-023-01292-8]
18. Latella D, Formica C, Ielo A, et al. A feasibility and usability study of a virtual reality tool (VESPA 2.0) for cognitive rehabilitation in patients with mild cognitive impairment: an ecological approach. Front Psychol 2024;15:1402894. [doi: 10.3389/fpsyg.2024.1402894] [Medline: 39492810]
19. Tuena C, Mancuso V, Stramba-Badiale C, et al. Egocentric and allocentric spatial memory in mild cognitive impairment with real-world and virtual navigation tasks: a systematic review. JAD 2021;79(1):95-116. [doi: 10.3233/JAD-201017] [Medline: 33216034]
20. Yang Q, Zhang L, Chang F, Yang H, Chen B, Liu Z. Virtual reality interventions for older adults with mild cognitive impairment: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2025 Jan 10;27:e59195. [doi: 10.2196/59195] [Medline: 39793970]
21. Guzmán DE, Rengifo CF, García-Cena CE. Serious games for cognitive rehabilitation in older adults: a conceptual framework. MTI 2024;8(8):64. [doi: 10.3390/mti8080064]
22. Alonso-Ríos D, Vázquez-García A, Mosqueira-Rey E, Moret-Bonillo V. Usability: a critical analysis and a taxonomy. Int J Hum Comput Interact 2009 Dec 28;26(1):53-74. [doi: 10.1080/10447310903025552]
23. Maggio MG, Maione R, Cotelli M, et al. Cognitive rehabilitation using virtual reality in subjective cognitive decline and mild cognitive impairment: a systematic review. Front Psychol 2025;16:1641693. [doi: 10.3389/fpsyg.2025.1641693] [Medline: 40994868]
24. Yu J, Song J, Shen Q. Effects of fully immersive virtual reality training on cognitive function in patients with mild cognitive impairment: a systematic review and meta-analysis. Front Hum Neurosci 2024;18:1467697. [doi: 10.3389/fnhum.2024.1467697] [Medline: 39713172]
25. Yuan P, Chen J, Peng D, Yang Q, Liu B, Lu C. Effectiveness of VR-based cognitive training and games on cognitive rehabilitation in patients with MCI: a systematic review and meta-analysis. Front Neurol 2025;16:1691344. [doi: 10.3389/fneur.2025.1691344] [Medline: 41281571]
26. Zhu S, Sui Y, Shen Y, et al. Effects of virtual reality intervention on cognition and motor function in older adults with mild cognitive impairment or dementia: a systematic review and meta-analysis. Front AGING Neurosci 2021;13:586999. [doi: 10.3389/fnagi.2021.586999] [Medline: 34025384]
27. Mehrinejad Khotbehsara M, Soar J, Lokuge S, Mehrinejad Khotbehsara E, Ip WK. The potential of virtual reality-based multisensory interventions in enhancing cognitive function in mild cognitive impairment: a systematic review. J Clin Med 2025 Aug 4;14(15):5475. [doi: 10.3390/jcm14155475] [Medline: 40807098]
28. Ienca M, Fabrice J, Elger B, et al. Intelligent assistive technology for Alzheimer's disease and other dementias: a systematic review. Journal of Alzheimer's Disease 2017 Feb 20;56(4):1301-1340. [doi: 10.3233/JAD-161037]
29. Blattgerste J, Renner P, Pfeiffer T. Augmented reality action assistance and learning for cognitively impaired people. 2019 Jun 5 Presented at: PETRA '19: The 12th Pervasive Technologies Related to Assistive Environments Conference; Jun 5-7, 2019; Rhodes Greece p. 270-279 URL: https://dl.acm.org/doi/proceedings/10.1145/3316782 [accessed 2026-02-22] [doi: 10.1145/3316782.3316789]
30. He D, Cao S, Le Y, Wang M, Chen Y, Qian B. Virtual reality technology in cognitive rehabilitation application: bibliometric analysis. JMIR SERIOUS GAMES 2022 Oct 19;10(4):e38315. [doi: 10.2196/38315] [Medline: 36260388]
31. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. Campbell Syst Rev 2022 Jun;18(2):e1230. [doi: 10.1002/cl2.1230] [Medline: 36911350]
32. Gattoni MF, Gobbo S, Feroldi S, et al. Identification of cognitive training for individuals with Parkinson's disease: a systematic review. Brain Sci 2025 Jan 11;15(1):61. [doi: 10.3390/brainsci15010061] [Medline: 39851429]
33. Catania V, Rundo F, Panerai S, Ferri R. Virtual reality for the rehabilitation of acquired cognitive disorders: a narrative review. BIOENGINEERING (BASEL) 2024;11(1):35. [doi: 10.3390/bioengineering11010035]
34. Chen C, Huang N, Hu B, Zhang M, Yuan J, Guo J. The effectiveness of digital technology interventions for cognitive function in older adults: a systematic review and meta-analysis of randomized controlled trials. GEROSCIENCE 2025 Feb;47(1):653-683. [doi: 10.1007/s11357-024-01446-z] [Medline: 39688787]

XSL•FO
RenderX

35. Gao Y, Ma L, Lin C, et al. Effects of virtual reality-based intervention on cognition, motor function, mood, and activities of daily living in patients with chronic stroke: a systematic review and meta-analysis of randomized controlled trials. Front AGING Neurosci 2021;13:766525. [doi: 10.3389/fnagi.2021.766525] [Medline: 34966267]
36. Parisi A, Bellinzona F, Di Lernia D, et al. Efficacy of multisensory technology in post-stroke cognitive rehabilitation: a systematic review. J Clin Med 2022 Oct 26;11(21):21. [doi: 10.3390/jcm11216324] [Medline: 36362551]
37. Tobar A, Jaramillo AP, Costa SC, Costa KT, Garcia SS. A physical rehabilitation approach for Parkinson's disease: a systematic literature review. CUREUS 2023 Sep;15(9):e44739. [doi: 10.7759/cureus.44739] [Medline: 37809151]
38. Campo-Prieto P, Cancela JM, Rodríguez-Fuentes G. Immersive virtual reality as physical therapy in older adults: present or future (systematic review). VIRTUAL Real 2021 Sep;25(3):801-817. [doi: 10.1007/s10055-020-00495-x]
39. Moulaei K, Sharifi H, Bahaadinbeigy K, Dinari F. Efficacy of virtual reality-based training programs and games on the improvement of cognitive disorders in patients: a systematic review and meta-analysis. BMC PSYCHIATRY 2024 Feb 12;24(1):116. [doi: 10.1186/s12888-024-05563-z] [Medline: 38342912]
40. Perra A, Riccardo CL, De Lorenzo V, et al. Fully immersive virtual reality-based cognitive remediation for adults with psychosocial disabilities: a systematic scoping review of methods intervention gaps and meta-analysis of published effectiveness studies. Int J Environ Res Public Health 2023 Jan 14;20(2):1527. [doi: 10.3390/ijerph20021527] [Medline: 36674283]
41. Kacmaz KS, Kaçmaz C. Bibliometric analysis of research in pediatrics related to virtual and augmented reality: a systematic review. Curr Pediatr Rev 2024;20(2):178-187. [doi: 10.2174/1573396319666230214103103] [Medline: 36786143]
42. Azhari HF. Advancing stroke diagnosis and management through nuclear medicine: a systematic review of clinical trials. Front Med (Lausanne) 2024;11:1425965. [doi: 10.3389/fmed.2024.1425965] [Medline: 39224610]
43. Lo YT, Su HC, Chuenchomnoy C, et al. Umbrella review of nonpharmacological interventions for intrinsic capacity in older adults. Ageing Res Rev 2025 Jun;108:102742. [doi: 10.1016/j.arr.2025.102742] [Medline: 40194665]
44. Bacanoiu MV, Danoiu M. New strategies to improve the quality of life for normal aging versus pathological aging. J Clin Med 2022 Jul 20;11(14):4207. [doi: 10.3390/jcm11144207] [Medline: 35887969]
45. Cabreira VA, Wilkinson T, Frostholm L, Stone J, Carson AJ. Systematic review and meta-analysis of standalone digital interventions for cognitive symptoms in people without dementia. npj Digit Med 2024;7(1). [doi: 10.1038/s41746-024-01280-9]
46. Mao Q, Zhao Z, Yu L, Zhao Y, Wang H. The effects of virtual reality-based reminiscence therapies for older adults with cognitive impairment: systematic review. J Med INTERNET Res 2024 Nov 12;26:e53348. [doi: 10.2196/53348] [Medline: 39531267]
47. Shahmoradi L, Mohammadian F, Rahmani Katigari M. A systematic review on serious games in attention rehabilitation and their effects. Behav Neurol 2022;2022:2017975. [doi: 10.1155/2022/2017975] [Medline: 35256889]
48. Calderone A, Marra A, De Luca R, et al. Multisensory stimulation in rehabilitation of dementia: a systematic review. Biomedicines 2025 Jan 9;13(1):149. [doi: 10.3390/biomedicines13010149] [Medline: 39857732]
49. Chi K, Chen J, Zhou S, Han Z. The effectiveness of digital cognitive intervention in patients with traumatic brain injury: systematic review and meta-analysis. Front Neurol 2025;16:1651443. [doi: 10.3389/fneur.2025.1651443] [Medline: 41127285]
50. Drazich BF, McPherson R, Gorman EF, et al. In too deep? A systematic literature review of fully‐immersive virtual reality and cybersickness among older adults. J AMERICAN GERIATRICS SOCIETY 2023 Dec;71(12):3906-3915. [doi: 10.1111/jgs.18553]
51. Burch BF, Gu J, Betz G, et al. Smart glasses for older adults with cognitive impairment: a scoping review. J Am Med Dir Assoc 2025 Nov;26(11):105831. [doi: 10.1016/j.jamda.2025.105831] [Medline: 40882953]
52. Romero-Ayuso D, Castillero-Perea Á, González P, et al. Assessment of cognitive instrumental activities of daily living: a systematic review. Disabil Rehabil 2021 May;43(10):1342-1358. [doi: 10.1080/09638288.2019.1665720] [Medline: 31549907]
53. Tan SB, Tan J, Raczkowska MN, et al. Digital game-based interventions for cognitive training in healthy adults and adults with cognitive impairment: protocol for a two-part systematic review and meta-analysis. BMJ Open 2023 May 4;13(5):e071059. [doi: 10.1136/bmjopen-2022-071059] [Medline: 37142320]
54. Kooijmans ECM, Hoogendijk EO, Drapała N, et al. Defining and categorizing nonpharmacologic interventions in the older population: a systematic review. J Am Med Dir Assoc 2025 Jan;26(1):105306. [doi: 10.1016/j.jamda.2024.105306] [Medline: 39424279]
55. Jun P, Chengye H, Hui W. Bibliometric analysis of rehabilitation in Alzheimer's disease (2000-2023): trends, hotspots and prospects. Front Aging Neurosci 2024;16:1457982. [doi: 10.3389/fnagi.2024.1457982] [Medline: 39691162]
56. Li Z, Yang L, Qiu H, Wang X, Zhang C, Zhang Y. Comparative efficacy of 5 non-pharmacological therapies for adults with post-stroke cognitive impairment: a Bayesian network analysis based on 55 randomized controlled trials. Front Neurol 2022;13. [doi: 10.3389/fneur.2022.977518]
57. Oliveira J, Gamito P, Souto T, et al. Virtual reality-based cognitive stimulation on people with mild to moderate dementia due to Alzheimer's disease: a pilot randomized controlled trial. Int J Environ Res Public Health 2021 May 16;18(10):5290. [doi: 10.3390/ijerph18105290] [Medline: 34065698]

XSL•FO
RenderX

58. Huang LC, Yang YH. The long-term effects of immersive virtual reality reminiscence in people with dementia: longitudinal observational study. JMIR SERIOUS GAMES 2022 Jul 25;10(3):e36720. [doi: 10.2196/36720] [Medline: 35877169]
59. Fiorenzato E, Zabberoni S, De Simone MS, et al. Effects of virtual reality cognitive training on executive function and prospective memory in Parkinson's disease and healthy aging. J Neurol Sci 2025 Jun 15;473:123507. [doi: 10.1016/j.jns.2025.123507] [Medline: 40250092]
60. De Simone MS, Costa A, Tieri G, et al. The effectiveness of an immersive virtual reality and telemedicine-based cognitive intervention on prospective memory in Parkinson's disease patients with mild cognitive impairment and healthy aged individuals: design and preliminary baseline results of a placebo-controlled study. Front Psychol 2023;14:1268337. [doi: 10.3389/fpsyg.2023.1268337] [Medline: 37928597]
61. Liu Z, He Z, Yuan J, et al. Application of immersive virtual-reality-based puzzle games in elderly patients with post-stroke cognitive impairment: a pilot study. BRAIN Sci 2023;13(1):79. [doi: 10.3390/brainsci13010079]
62. De Luca R, Calderone A, Gangemi A, et al. Is virtual reality orientation therapy useful to optimize cognitive and behavioral functioning following severe acquired brain injury? An exploratory study. Brain Sci 2024 Apr 23;14(5):410. [doi: 10.3390/brainsci14050410] [Medline: 38790389]
63. Baldimtsi E, Mouzakidis C, Karathanasi EM, et al. Effects of virtual reality physical and cognitive training intervention on cognitive abilities of elders with mild cognitive impairment. Journal of Alzheimer's Disease Reports 2023 Nov 1;7(1):1475-1490. [doi: 10.3233/ADR-230099]
64. Downey R, Gagné N, Mohanathas N, et al. At-home computerized executive-function training to improve cognition and mobility in normal-hearing adults and older hearing aid users: a multi-centre, single-blinded randomized controlled trial. BMC Neurol 2023 Oct 20;23(1):378. [doi: 10.1186/s12883-023-03405-1] [Medline: 37864139]
65. Zhu B, Zhou Y, Weng Q, et al. Effects of biophilic virtual reality on cognitive function of patients undergoing laparoscopic surgery: study protocol for a sham randomised controlled trial. BMJ Open 2022 Jul 8;12(7):e052769. [doi: 10.1136/bmjopen-2021-052769] [Medline: 35803619]
66. Buele J, Avilés-Castillo F, Del-Valle-Soto C, Varela-Aldás J, Palacios-Navarro G. Effects of a dual intervention (motor and virtual reality-based cognitive) on cognition in patients with mild cognitive impairment: a single-blind, randomized controlled trial. J Neuroeng Rehabil 2024 Aug 1;21(1):130. [doi: 10.1186/s12984-024-01422-w] [Medline: 39090664]
67. Han K, Park K, Choi KH, Lee J. Mobile augmented reality serious game for improving old adults' working memory. Appl Sci (BASEL) 2021;11(17):7843. [doi: 10.3390/app11177843]
68. Jeong E, Ham Y, Lee SJ, Shin JH. Virtual reality-based music attention training for acquired brain injury: A randomized crossover study. Ann N Y Acad Sci 2024 Nov;1541(1):151-162. [doi: 10.1111/nyas.15249] [Medline: 39476208]
69. Lanzoni D, Vitali A, Regazzoni D, Rizzi C. Design of customized virtual reality serious games for the cognitive rehabilitation of retrograde amnesia after brain stroke. J Comput Inf Sci Eng 2022 Jun 1;22(3). [doi: 10.1115/1.4053149]
70. Maeng S, Hong JP, Kim WH, et al. Effects of virtual reality-based cognitive training in the elderly with and without mild cognitive impairment. PSYCHIATRY Investig 2021 Jul;18(7):619-627. [doi: 10.30773/pi.2020.0446] [Medline: 34265200]
71. Ip WK, Soar J, James C, Wang SY, Fong KNK. Virtual reality game-based training for preventing falls among community-dwelling older adults with mild cognitive impairment: a pilot randomized control trial study. Virtual Real 2025;29(1). [doi: 10.1007/s10055-024-01084-y]
72. Oliveira LM, Evangelista E Souza EH, Alves MR, et al. 2D virtual reality-based exercise improves spatial navigation in institutionalized non-robust older persons: a preliminary data report of a single-blind, randomized, and controlled study. Front Neurol 2020;11:609988. [doi: 10.3389/fneur.2020.609988] [Medline: 33584510]
73. Początek G, Wyszomirska J, Segiet N. Cognitive training for Alzheimer's disease. Aktualn Neurol 2023;23(4):138-144 [FREE Full text] [doi: 10.15557/AN.2023.0021]
74. Cavallo M, Lasaponara S. Cognition and cognitive training in patients with neurodegenerative diseases. Brain Sci 2025 Feb 12;15(2):183. [doi: 10.3390/brainsci15020183] [Medline: 40002516]
75. Kwan RYC, Liu J, Sin OSK, et al. Effects of virtual reality motor-cognitive training for older people with cognitive frailty: multicentered randomized controlled trial. J Med Internet Res 2024 Sep 11;26:e57809. [doi: 10.2196/57809] [Medline: 39259959]
76. Fu Y, Wang H. Clinical observation of VR virtual reality rehabilitation training combined with acupuncture in the treatment of mild cognitive impairment. SLAS Technol 2025 Apr;31:100250. [doi: 10.1016/j.slast.2025.100250] [Medline: 39863045]
77. Manenti R, Baglio F, Pagnoni I, et al. Long-lasting improvements in episodic memory among subjects with mild cognitive impairment who received transcranial direct current stimulation combined with cognitive treatment and telerehabilitation: a multicentre, randomized, active-controlled study. Front Aging Neurosci 2024;16:1414593. [doi: 10.3389/fnagi.2024.1414593] [Medline: 38966802]
78. Gambella E, Margaritini A, Benadduci M, et al. An integrated intervention of computerized cognitive training and physical exercise in virtual reality for people with Alzheimer's disease: The jDome study protocol. Front Neurol 2022;13:964454. [doi: 10.3389/fneur.2022.964454] [Medline: 36034306]
79. Contrada M, Pignolo L, Vatrano M, et al. Application of multidomain cognitive training in a tele-neurorehabilitation setting for treatment of post-stroke cognitive disorders. BRAIN Sci 2025;15(1):11. [doi: 10.3390/brainsci15010011]

XSL•FO
RenderX

80. Huber SK, Knols RH, Held JPO, Betschart M, de Bruin ED. PEMOCS: Evaluating the effects of a concept-guided, personalised, motor-cognitive exergame training on cognitive functions and gait in chronic stroke-study protocol for a randomised controlled trial. Trials 2024 Jul 4;25(1):451. [doi: 10.1186/s13063-024-08283-7] [Medline: 38965612]
81. Hwang NK, Choi JB, Choi DK, et al. Effects of semi-immersive virtual reality-based cognitive training combined with locomotor activity on cognitive function and gait ability in community-dwelling older adults. HEALTHCARE (Basel) 2021 Jun 28;9(7):814. [doi: 10.3390/healthcare9070814] [Medline: 34203233]
82. De Luca R, Marra A, Pollicino P, et al. Advances in neurorehabilitation of TREM2-related dementia: A case report on a novel multimodal approach using virtual reality. Medicine (Baltimore) 2022 May 27;101(21):e29470. [doi: 10.1097/MD.0000000000029470] [Medline: 35623081]
83. Faria AL, Almeida Y, Branco D, et al. NeuroAIreh@b: an artificial intelligence-based methodology for personalized and adaptive neurorehabilitation. Front Neurol 2023;14:1258323. [doi: 10.3389/fneur.2023.1258323] [Medline: 38322797]
84. Maggio MG, Maione R, Migale S, et al. Low-intensity virtual reality exercise for caregivers of people with mild cognitive impairment: a pilot study. J Funct Morphol Kinesiol 2025 Sep 16;10(3):353. [doi: 10.3390/jfmk10030353] [Medline: 40981052]
85. Bourrelier J, Fautrelle L, Haratyk E, et al. Enhancement of anticipatory postural adjustments by virtual reality in older adults with cognitive and motor deficits: a randomised trial. GERIATRICS (Basel) 2021 Jul 22;6(3):72. [doi: 10.3390/geriatrics6030072] [Medline: 34449605]
86. Tang YM, Tse SY, Chan HS, Yip HT, Cheung HT, Geda MW. An innovative interactive narrative persona (INP) approach for virtual reality-based dementia tour design (VDT) in rehabilitation contexts. Bioengineering (Basel) 2025 Sep 17;12(9):987. [doi: 10.3390/bioengineering12090987] [Medline: 41007233]
87. Bumbea AM, Glavan DG, Vasile RC, et al. Cognitive impairment and psychological morbidity among stroke survivors in rehabilitation: a cross-sectional analysis. J Clin Med 2025 Oct 31;14(21):21. [doi: 10.3390/jcm14217735] [Medline: 41227130]
88. Zhang T, Liu W, Bai Q, Gao S. Virtual reality technology in the rehabilitation of post-stroke cognitive impairment: an opinion article on recent findings. Front Psychol 2023;14:1271458. [doi: 10.3389/fpsyg.2023.1271458] [Medline: 37849482]
89. Capecci M, Gandolfi M, Straudi S, et al. Advancing public health through technological rehabilitation: insights from a national clinician survey. BMC Health Serv Res 2024 Dec 19;24(1):1626. [doi: 10.1186/s12913-024-11991-0] [Medline: 39702315]
90. Chau PH, Kwok YYJ, Chan MKM, et al. Feasibility, acceptability, and efficacy of virtual reality training for older adults and people with disabilities: single-arm pre-post study. J Med INTERNET Res 2021 May 4;23(5):e27640. [doi: 10.2196/27640] [Medline: 33944795]
91. Kim Y, Kim J, Kim J, Lee C, Ory MG. Feasibility and acceptability of an immersive puzzle-based virtual reality activity for community-dwelling older adults. Front VIRTUAL Real 2025;6. [doi: 10.3389/frvir.2025.1655981]
92. Munoz JE, Ali F, Basharat A, et al. Development of classifiers to determine factors associated with older adult's cognitive functions and game user experience in VR using head kinematics. IEEE Trans GAMES 2025;17(3):604-612. [doi: 10.1109/TG.2023.3317825]
93. Hassandra M, Galanis E, Hatzigeorgiadis A, et al. A virtual reality app for physical and cognitive training of older people with mild cognitive impairment: mixed methods feasibility study. JMIR Serious Games 2021 Mar 24;9(1):e24170. [doi: 10.2196/24170] [Medline: 33759797]
94. Kua Z, Ong RHS, Chen NYC, et al. Virtual reality-based cognitive training for MCI in the elderly: a feasibility randomised pilot study. Ann Acad Med Singap 2025 Jun 24;54(7):445-447. [doi: 10.47102/annals-acadmedsg.2024162] [Medline: 40799097]
95. Mondellini M, Arlati S, Gapeyeva H, et al. User experience during an immersive virtual reality-based cognitive task: a comparison between Estonian and Italian older adults with MCI. Sensors (Basel) 2022 Oct 27;22(21):21. [doi: 10.3390/s22218249] [Medline: 36365947]
96. Perra A, Primavera D, Leonetti V, et al. Virtual reality cognitive remediation tool for individuals with mild cognitive impairment: study protocol for a feasibility randomized clinical trial. Front PUBLIC HEALTH 2024;12:1477279. [doi: 10.3389/fpubh.2024.1477279] [Medline: 39651467]
97. Stasolla F, Di Gioia M. Combining reinforcement learning and virtual reality in mild neurocognitive impairment: a new usability assessment on patients and caregivers. Front Aging Neurosci 2023;15:1189498. [doi: 10.3389/fnagi.2023.1189498] [Medline: 37293666]
98. Tseng KC, Giau DTN. A feasibility study of using virtual reality as a pre-occupancy evaluation tool for the elderly. Automation in Construction 2022 Feb;134:104037. [doi: 10.1016/j.autcon.2021.104037]
99. Jeong H, Kang D, Kim JE, Lim J, Lee HW. Preliminary study on the feasibility of virtual reality-based cognitive training on patients with mild to moderate Alzheimer's disease. J Alzheimers Dis Rep 2025 Jan;9:25424823251385901. [doi: 10.1177/25424823251385901] [Medline: 41048557]
100. Maggio MG, Rizzo A, Benenati A, et al. Efficacy and feasibility of virtual reality-based cognitive tele-rehabilitation in Parkinson's disease: a pilot randomized controlled trial on patients and caregivers. Digit HEALTH 2025 May;11:20552076251376534. [doi: 10.1177/20552076251376534] [Medline: 41018508]

XSL•FO
RenderX

101. Sommerlad A, Jawharieh H, Ang CS, Huntley J, Gokani HA. Carers and professionals' views on using virtual reality in dementia care: a qualitative study. DEMENTIA (London) 2025 Jan;24(1):71-90. [doi: 10.1177/14713012241272786] [Medline: 39121239]
102. Matsangidou M, Frangoudes F, Schiza E, et al. Participatory design and evaluation of virtual reality physical rehabilitation for people living with dementia. VIRTUAL Real 2023 Mar;27(1):421-438. [doi: 10.1007/s10055-022-00639-1]
103. Chuang IC, Huang XT, Chen IC, Wu CY, Wu CY. Feasibility, usability, and effects of leisure-based cognitive training using a fully immersive virtual reality system in older adults: single-arm pretest-posttest pilot study. JMIR SERIOUS GAMES 2025 Oct 9;13:e66673. [doi: 10.2196/66673] [Medline: 41066691]
104. Chen CY, Chen F, Tsai CE. Exploring older people's user experiences with VR: a case study of Taiwanese seniors. Entertain Comput 2025 Sep;55:100974. [doi: 10.1016/j.entcom.2025.100974]
105. Siette J, Campbell C, Adam PJ, Harris CB. Exploring the usability of the virtual reality module LEAF CAFÉ: a qualitative think-aloud study. BMC Geriatr 2024 Feb 16;24(1):162. [doi: 10.1186/s12877-024-04767-y] [Medline: 38365613]
106. Pedroli E, Mancuso V, Stramba-Badiale C, et al. Brain m-app's structure and usability: a new application for cognitive rehabilitation at home. Front Hum Neurosci 2022;16:898633. [doi: 10.3389/fnhum.2022.898633] [Medline: 35782042]
107. Ghosh R, Khan N, Migovich M, et al. Iterative user centered design of robot-mediated paired activities for older adults with mild cognitive impairment (MCI). In: Lecture Notes in Computer Science 13331 LNCS: Springer; 2022:14-28. [doi: 10.1007/978-3-031-05654-3_2]
108. Haley KL, Cunningham KT, Kim I, Shafer JS. Autonomy-supportive treatment for acquired apraxia of speech: feasibility and therapeutic effect. Aphasiology 2021 Apr 3;35(4):539-559. [doi: 10.1080/02687038.2019.1705662]
109. McCracken MK, Shayman CS, Fino PC, Stefanucci JK, Creem-Regehr SH. A comparison of the effects of older age on homing performance in real and virtual environments. IEEE Trans Visual Comput GRAPHICS 2025;31(5):3213-3222. [doi: 10.1109/TVCG.2025.3549901]
110. Tuena C, Serino S, Goulene KM, Pedroli E, Stramba-Badiale M, Riva G. Bodily and visual-cognitive navigation aids to enhance spatial recall in mild cognitive impairment. Journal of Alzheimer's Disease 2024 May 28;99(3):899-910. [doi: 10.3233/JAD-240122]
111. Silva JD, Santos MD, Costa RD, et al. Applicability of an immersive virtual reality system to assess egocentric orientation of older adults. Arq Neuropsiquiatr 2023 Jan;81(1):019-026. [doi: 10.1055/s-0042-1759762]
112. Tuena C, Serino S, Pedroli E, et al. Embodied spatial navigation training in mild cognitive impairment: a proof-of-concept trial. Journal of Alzheimer's Disease 2024 Jul 30;100(3):923-934. [doi: 10.3233/JAD-240200]
113. Panerai S, Catania V, Rundo F, et al. Functional living skills in patients with major neurocognitive disorder due to degenerative or non-degenerative conditions: effectiveness of a non-immersive virtual reality training. Sensors (Basel) 2023 Feb 8;23(4):1896. [doi: 10.3390/s23041896] [Medline: 36850494]
114. Stasolla F, Di Gioia M, Messina I, et al. Assessing and recovering Alzheimer's disease: a comparative analysis of standard neuropsychological approaches and virtual reality interventions with the use of digital storytelling. Front Psychol 2024;15:1406167. [doi: 10.3389/fpsyg.2024.1406167] [Medline: 39114597]
115. Mirmiran S, Moussavi Z. Improving quality of life for institutionalized individuals with advanced dementia: a pilot study on efficacy of a semi-immersive virtual reality driving simulator for individuals with advanced dementia. DISPLAYS 2026 Jan;91:103182. [doi: 10.1016/j.displa.2025.103182]
116. Varela-Aldás J, Buele J, Amariglio R, García-Magariño I, Palacios-Navarro G. The cupboard task: an immersive virtual reality-based system for everyday memory assessment. Int J Hum Comput Stud 2022 Nov;167:102885. [doi: 10.1016/j.ijhcs.2022.102885]
117. Bhargava Y, Sumanth SV, Deshmukh A, Baths V. A peek into the minds of game developers and neuroscience researchers collaborating on cognition assessing games. IEEE ACCESS 2025;13:62674-62690. [doi: 10.1109/ACCESS.2025.3557454]
118. Mahncke HW. Interpreting self-report measures about cognition: what matters and what to do with the data: a commentary on "Awareness of baseline functioning and sensitivity to improvement in older people with and without mild cognitive impairment receiving a computerized functional skills training program". Int Psychogeriatr 2025 Jun;37(3):100082. [doi: 10.1016/j.inpsyc.2025.100082] [Medline: 40360334]
119. Cheng L, Pang WYJ, Kong A. A virtual reality game for older adults' immersive learning. Sci Rep 2025;15(1). [doi: 10.1038/s41598-025-07354-2]
120. Manser P, de Bruin ED. Making the best out of IT: design and development of exergames for older adults with mild neurocognitive disorder - a methodological paper. Front AGING Neurosci 2021;13:734012. [doi: 10.3389/fnagi.2021.734012] [Medline: 34955806]
121. He X, Liu C, Li Z, Cai X. Leisure activity interventions on cognition in mild cognitive impairment patients: a meta-analysis. Neuropsychiatr Dis Treat 2025;21:1671-1687. [doi: 10.2147/NDT.S536627] [Medline: 40827136]
122. Bhargava Y, Kumar Kattupalli K, Baths V. CogTrack: a proof of concept for cognition tracker. IEEE Access 2024;12:194609-194629. [doi: 10.1109/ACCESS.2024.3519220]
123. Dombrowski W, Mims A, Kremer I, et al. Dementia ideal care: ecosystem map of best practices and care pathways enhanced by technology and community. J Alzheimers Dis 2024;100(1):87-117. [doi: 10.3233/JAD-231491] [Medline: 38848182]

XSL•FO
RenderX

124. Kaiser W, de Bruin ED, Manser P. Domain-specific evaluation of exergame metrics among older adults with mild neurocognitive disorder: secondary analysis of 2 randomized controlled trials. JMIR SERIOUS GAMES 2025 May 21;13:e65878. [doi: 10.2196/65878] [Medline: 40397948]
125. Peng Y, Song P, Karako T, Asakawa T. Blocking progression from intervenable mild cognitive impairment to irreversible dementia, what can we do? Biosci Trends 2024 Nov 15;18(5):409-412. [doi: 10.5582/bst.2024.01304] [Medline: 39401896]
126. Singh SM, Kathiresan P, Gaur N. Clinical practice guidelines on addressing cognitive impairment in vascular dementias. Indian J Psychiatry 2025 Jan;67(1):54-64. [doi: 10.4103/indianjpsychiatry.indianjpsychiatry_713_24] [Medline: 40046475]
127. Colombini G, Duradoni M, Carpi F, Vagnoli L, Guazzini A. LEAP Motion technology and psychology: a mini-review on hand movements sensing for neurodevelopmental and neurocognitive disorders. Int J Environ Res Public Health 2021 Apr 11;18(8):4006. [doi: 10.3390/ijerph18084006] [Medline: 33920362]
128. Zhou L, Huang X, Wang J, Wang F, Liu J, Liu N. The influence of eight cognitive training regimes upon cognitive screening tool performance in post-stroke survivors: a network meta-analysis. Front Aging Neurosci 2024;16:1374546. [doi: 10.3389/fnagi.2024.1374546] [Medline: 39100748]

Abbreviations

AD: Alzheimer disease
ADL: activity of daily living
AR: augmented reality
CAVE: Cave Automatic Virtual Environment
HMD: head-mounted display
IADL: instrumental activity of daily living
MCI: mild cognitive impairment
MMSE: Mini-Mental State Examination
MoCA: Montreal Cognitive Assessment
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
UX: user experience
VR: virtual reality
WoS: Web of Science
XR: extended reality

Edited by A Schwartz, M Balcarras; submitted 18.Sep.2025; peer-reviewed by G Spalla, T Llana; revised version received 18.Dec.2025; accepted 18.Dec.2025; published 18.Mar.2026.

Please cite as:
Konecny J, Lanza G, Buono S, Ferri R, Luca A, Martinek R, Musso S, Palmigiano A, Prauzkova B, Quercia A, Rundo F, Subramanian RA, Serretti A, Prauzek M
Immersive Technologies for Cognitive Rehabilitation in Dementia and Mild Cognitive Impairment: Systematic Review
J Med Internet Res 2026;28:e84349
URL: https://www.jmir.org/2026/1/e84349
doi:10.2196/84349

© Jaromir Konecny, Giuseppe Lanza, Serafino Buono, Raffaele Ferri, Antonina Luca, Radek Martinek, Sabrina Musso, Aurora Palmigiano, Barbora Prauzkova, Angelica Quercia, Francesco Rundo, Ramachandran Avala Subramanian, Alessandro Serretti, Michal Prauzek. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating Encoder and Decoder Models for Extended Clinical Concept Recognition in Japanese Clinical Texts: Comparative Study With Weighted Soft Matching

Yuya Tsukiji[1], MMSc; Satoshi Kataoka[1], MS; Masafumi Itokazu[1], MD; Ryozo Nagai[2], MD, PhD; Takeshi Imai[1], PhD

[1]Center for Disease Biology and Integrative Medicine, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Clinical Research Center A646, The University of Tokyo Hospital, Tokyo, Japan

[2]Jichi Medical University, Tochigi, Japan

Corresponding Author:
Takeshi Imai, PhD
Center for Disease Biology and Integrative Medicine, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Clinical Research Center A646, The University of Tokyo Hospital, Tokyo, Japan

Abstract

Background: Extracting medical knowledge for secondary purposes, such as diagnostic support, continues to pose a substantial challenge. Conventional named entity recognition has focused on short terms (eg, genes, diseases, and chemicals), whereas extraction and assessment of longer, complex expressions remain underexplored. Clinically vital concepts, such as diseases, pathologies, symptoms, and findings, often appear as long phrases, and accurate extraction is crucial for applications such as constructing causal knowledge from case reports. Consequently, a framework addressing both short terms and clinically meaningful long phrases—termed extended Clinical Concept Recognition (E-CCR)—is essential.

Objective: This study, the first comprehensive investigation of E-CCR model selection, aimed to identify optimal strategies by comparing encoder versus decoder models and general-purpose versus domain-specific pretraining. We analyzed variation in effectiveness by target length and proposed a novel E-CCR evaluation metric.

Methods: We evaluated 17 encoder and decoder models using J-CaseMap, a database of approximately 20,000 Japanese case reports annotated with clinical concepts. Performance was primarily assessed using the weighted soft matching score, which penalizes fragmentation of long extraction targets and weights scores by target length to account for the greater difficulty of extracting longer expressions.

Results: On J-CaseMap, JMedDeBERTa(s)—an encoder model pretrained on domain-specific medical text—achieved the highest mean performance (F1-score=0.758, SD 0.002), with similarly strong results from JMedDeBERTa(c), suggesting comparable performance among the top encoder models. As the fragmentation penalty increased, performance generally declined; however, no consistently severe degradation was observed. On the Medical Report Named Entity Recognition for positive disease dataset, the general-domain DeBERTaV2-base yielded the highest mean F1 score, and differences among the medical-domain JMedDeBERTa(s) and JMedDeBERTa(c) variants were small, suggesting limited benefit of domain-specific pretraining. Overall, under our experimental settings (low-rank adaptation fine-tuning for decoders and full fine-tuning for encoders), encoder models outperformed decoder models, and token classification outperformed our instruction tuning setup.

Conclusions: Under our experimental setting, encoder-based token classification achieved the highest mean performance on our internal dataset. Differences among the top encoder models were small and should be interpreted as comparable within the uncertainty implied by our annotation review, whereas decoder-based approaches did not surpass encoder-based models in this setup, suggesting that encoder models can deliver high accuracy with fewer parameters and may offer practical advantages in resource-constrained environments. Token classification outperformed instruction tuning for extracting long expressions, whereas instruction tuning was better suited to short terms. Using the weighted soft matching score, we found that performance did not substantially deteriorate as the fragmentation penalty increased, indicating that extracted spans were rarely fragmented. Similar trends in external validation datasets suggest that findings under our setup may generalize to information extraction tasks on Japanese medical text. Further investigation is needed to determine whether these findings hold across other languages and medical document types.

(J Med Internet Res 2026;28:e78681) doi:10.2196/78681

XSL•FO
RenderX

KEYWORDS

natural language processing; NLP; named entity recognition; instruction tuning; token classification; large language model; LLM; transformer model

Introduction

With advancements in digitization within the medical field, various medical documents, including electronic medical records, case reports, and discharge summaries, are increasingly digitized. There is a growing demand for the secondary use of these digitized documents, such as extracting medical knowledge for diagnostic support, and research in this area is advancing. For example, Japan has developed a diagnostic support system called "Difficult Diagnosis Case Search: J-CaseMap" [1]. J-CaseMap uses a manually constructed knowledge graph created by approximately 150 internists, extracting the causal relationships between diseases and the associated pathophysiological processes and symptoms from approximately 20,000 case reports collected from local chapters of the Japan Society of Internal Medicine. A system has been developed that, through search and inference on this large-scale knowledge graph integrating all causal chains from these cases, presents both a "list of differential diagnoses suggested by a given combination of input symptoms and findings" and "the causal chain leading to those diagnoses." This system has been implemented as a service for members of the Japan Society of Internal Medicine.

Such diagnostic support systems can assist in searching similar cases, suggesting diseases, and supporting diagnostic reasoning, making them highly valuable for standardizing the quality of medical care. However, constructing and updating the underlying structured knowledge database requires substantial time and effort, posing substantial challenges to efficiency and automation. Consequently, demand for automated methods leveraging large language models (LLMs), which have seen rapid advances in recent years, has increased to structure knowledge from medical documents using computers.

When constructing a knowledge database of the causal relationships among diseases, pathologies, symptoms, and findings that are considered essential for differential diagnosis from clinical texts (such as case reports), the first crucial step is to extract the terms and phrases representing these key clinical concepts. Conventionally, named entity recognition (NER) has been widely used to extract patient-related information from medical texts, including clinical records, medical literature, and electronic health records [2,3]. The expressions extracted can then be used for aggregating clinical data, substantiating biomedical findings, supporting clinical decision-making, improving patient care, and enhancing health care management and resource allocation [4]. However, much of the prior research has concentrated solely on extracting terms or short expressions (eg, disease names, symptoms, drug names, and treatment methods) [5-10].

The clinical concepts critical for differential diagnosis—diseases, pathologies, symptoms, and findings—are not necessarily confined to single terms or short expressions. For instance, the expression "Anti-GAD antibody level was elevated at 3790.0 U/mL" in its entirety denotes the concept "anti-GAD-positive." If this concept is considered as a node in a causal knowledge graph, then extracting the corresponding long expression from the text becomes essential. Thus, when addressing the task of "extracting medical concepts that reflect the clinically meaningful units considered by physicians during diagnosis," a comprehensive extraction framework is needed that targets not only terms (short expressions) but also longer expressions (long expressions) that may span multiple sentences. In this study, this task is referred to as Extended Clinical Concept Recognition (E-CCR). This distinguishes our task from conventional NER, which typically targets shorter named entities such as disease names, drug names, or laboratory items.

Several studies [11-13] have investigated Japanese medical text processing, including work on deidentification, document classification, and extraction of relatively short medical entities. However, to our knowledge, no prior study has systematically addressed the extraction of "clinically important long expressions" that form the nodes of a clinical knowledge graph.

Previous studies have primarily used evaluation metrics such as ROUGE-1, ROUGE-2, ROUGE-L, recall, precision, and F_1-score. ROUGE calculates scores based on string similarity [14]. F_1-scores have been computed by segmenting annotation schemes (ie, the extraction targets) into tokens [2] or characters [15]. Scores are typically computed on a token or character basis rather than per extracted unit, as the number of characters in each extraction target may vary. However, when comparing multiple models, token-based metrics should be avoided because each LLM uses its tokenizer and varying tokenizers yield different segmentation granularities. Conversely, even when ROUGE or F_1-scores are computed on a character basis, their validity is questionable for long extraction targets. This is because a fragmented extraction of a long expression, although undesirable, may still attain a high score on a character basis, simply because the majority of characters are captured. Therefore, when comparing multiple models that include long expressions as extraction targets, it becomes imperative to develop a new evaluation metric that penalizes fragmented outputs and is unaffected by differences in tokenizers.

Expression extraction can be approached as either a sequence labeling task or a generation task. Conventionally, many researchers have addressed expression extraction as a sequence labeling task. Techniques such as conditional random fields [16] and long short-term memory (LSTM) [17] have been proposed, with transformer-based encoder models, particularly those based on Bidirectional Encoder Representations from Transformers (BERT) [18], gaining substantial popularity [19]. More recently, with the emergence of LLMs, interest in performing expression extraction as a generation task has increased, and various information extraction techniques such as cross-domain learning, zero-shot prompting, in-context learning, supervised fine-tuning, and data augmentation have emerged [20]. However, as generation tasks require

XSL•FO
RenderX

simultaneous text generation and expression extraction, they are inherently more complex than sequence labeling approaches. Therefore, extracting expressions via generation tasks necessitates careful prompt design and fine-tuning [21-25].

However, whether encoder or decoder models are more suitable for expression extraction remains unclear. For example, even when techniques such as designing prompt formats that facilitate the proper noun extraction [22], tailoring prompts to specific tasks [23], and customizing datasets [24] are used, in zero-shot or few-shot settings, decoder models (eg, GPT-3 [25], GPT-3.5, and GPT-4 [26]) have not outperformed encoder models (eg, BERT [23], BioClinicalBERT [24], and RoBERTa-Large [27]). In contrast, BioNER-LLAMA, a decoder model specialized for the medical domain, has reportedly surpassed the encoder model BioBERT by using instruction tuning on a dataset for proper noun extraction in biomedical contexts and by using carefully designed prompts [28]. In another study [29], when encoder and decoder models were both used for token classification in expression extraction, LLaMA2 outperformed RoBERTa.

Additionally, insights into the effectiveness of domain-specific texts for expression extraction in the medical domain remain limited. Previous studies have compared models pretrained on domain-specific medical texts or further adapted to them with models pretrained on general text. For example, BERT_mimic, pretrained on medical texts, demonstrated superior performance compared to a BERT model pretrained on general English texts [30]. Similarly, BioBERT, initially pretrained on general English texts and subsequently adapted to medical texts, outperformed BERT [31]. In contrast, BioELECTRA, pretrained on a relatively small corpus (0.8 GB) of German medical texts, performed worse than DBMDZ ELECTRA, which was pretrained on general texts. Lentzen et al [32] reported that BioGottBERT—first pretrained on general texts and then further pretrained on similar medical texts—achieved the best performance in some cases. Subies et al [33] confirmed the effectiveness of encoder models in Spanish clinical tasks but noted that the best-performing model was not domain adapted; rather, domain-adapted models performed worse than general or multilingual models. Thus, the current insights regarding the effectiveness of domain-adapted models remain inadequate. Moreover, prior studies have primarily focused on extracting short expressions such as drug names, disease names, symptoms, and treatment methods, and evidence on long expression extraction remains limited.

Therefore, this study aimed to elucidate effective model selection strategies for the E-CCR task by comparing encoder and decoder models, as well as general-purpose models with domain-specific models. We propose a novel evaluation metric tailored for E-CCR that not only incorporates conventional measures but also accounts for penalties when long extraction targets are fragmented. We analyzed how model effectiveness varies with the length of the extraction target.

Methods

Overview

Figure 1 illustrates the flowchart of our research design. We compared and evaluated sequence labeling tasks (token classification) with generation tasks (instruction tuning) using a corpus of clinically essential concepts extracted from case reports. Performance was assessed using a weighted soft matching score, and 17 measures—including F_1-score, recall, and precision—were used for both encoder and decoder models.

Figure 1. Overall study design for comparing encoder and decoder language models on the extended Clinical Concept Recognition task. Japanese clinical case reports from the J-CaseMap database were used as input texts. Seventeen language models (encoder and decoder architectures and instruction tuning vs token classification settings) were fine-tuned in a 5-fold cross-validation framework to extract clinically important concepts from these reports. Model outputs were evaluated using the weighted soft matching score with varying fragmentation penalties (p=1, 1.5, 2, 100), along with conventional F_1, recall, and precision, followed by qualitative review of output examples.

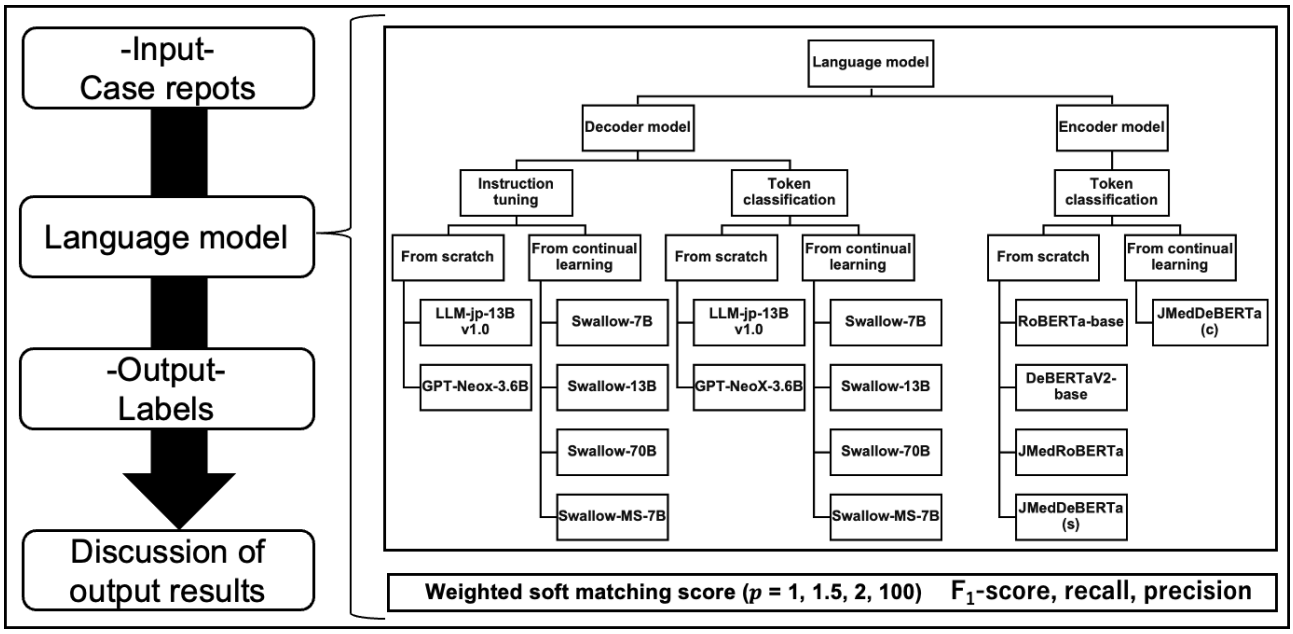

XSL•FO
RenderX

Materials

We used data from the J-CaseMap database, which comprises abstracts from approximately 20,000 case reports presented at local chapters of the Japanese Society of Internal Medicine. Each abstract contains approximately 600 characters. The database was manually constructed by approximately 150 internists, who extracted causal chains of diseases, resulting pathophysiological processes, and associated clinical signs and symptoms from these case reports. The knowledge graph comprises "nodes," representing term expressions that are normalized by specific rules, and "edges," which represent the causal relationships between them. Although such a knowledge database supports decision systems in differential diagnosis, its construction requires substantial manual effort, thereby necessitating the development of automated methods.

Using J-CaseMap as an example, the task of automatically constructing a causal chain knowledge base from case reports can be divided into three steps:

1. Extraction: identify and extract terms or phrases from the case reports that are considered clinically essential for differential diagnosis (eg, diseases, pathophysiological processes, signs, and symptoms)
2. Normalization: map the extracted expressions to standardized terminology
3. Causal inference: infer causal relationships among the normalized expressions

As shown in Figure 2, the first critical step in developing an automated approach is extracting clinically essential concepts that form the nodes of the knowledge graph. When these essential clinical concepts appear in the text, they often manifest as long expressions. For instance, the sentence "Glucagon負荷試験で内因性インスリン分泌反応の低下を認めた" might be represented as "Glucagon負荷試験 = 内因性インスリン分泌/低下" (ie, *"The glucagon stimulation test indicated a reduced endogenous insulin secretory response"* → *"glucagon stimulation test=endogenous insulin secretion/reduced"*), constituting a long expression.

Figure 2. Correspondence between clinically essential expressions in a Japanese case report and the J-CaseMap knowledge graph. The left panel shows an example Japanese case report (with English translation) describing type 1 diabetes complicated by Graves disease. Phrases highlighted in light blue indicate text spans selected by physicians as clinical concepts (eg, diagnoses, symptoms, and test findings) that are important for differential diagnosis. The right panel illustrates how these expressions are normalized as nodes and linked by causal relationships (edges) within the J-CaseMap knowledge graph. These node-defining spans constitute the gold-standard corpus for the extended Clinical Concept Recognition task. GAD: glutamic acid decarboxylase; HbA_{1c}: hemoglobin A_{1c}; ICA: islet cell antibody; SPIDDM: slowly progressive insulin-dependent diabetes mellitus; TBII: thyrotropin binding inhibitory immunoglobulin.

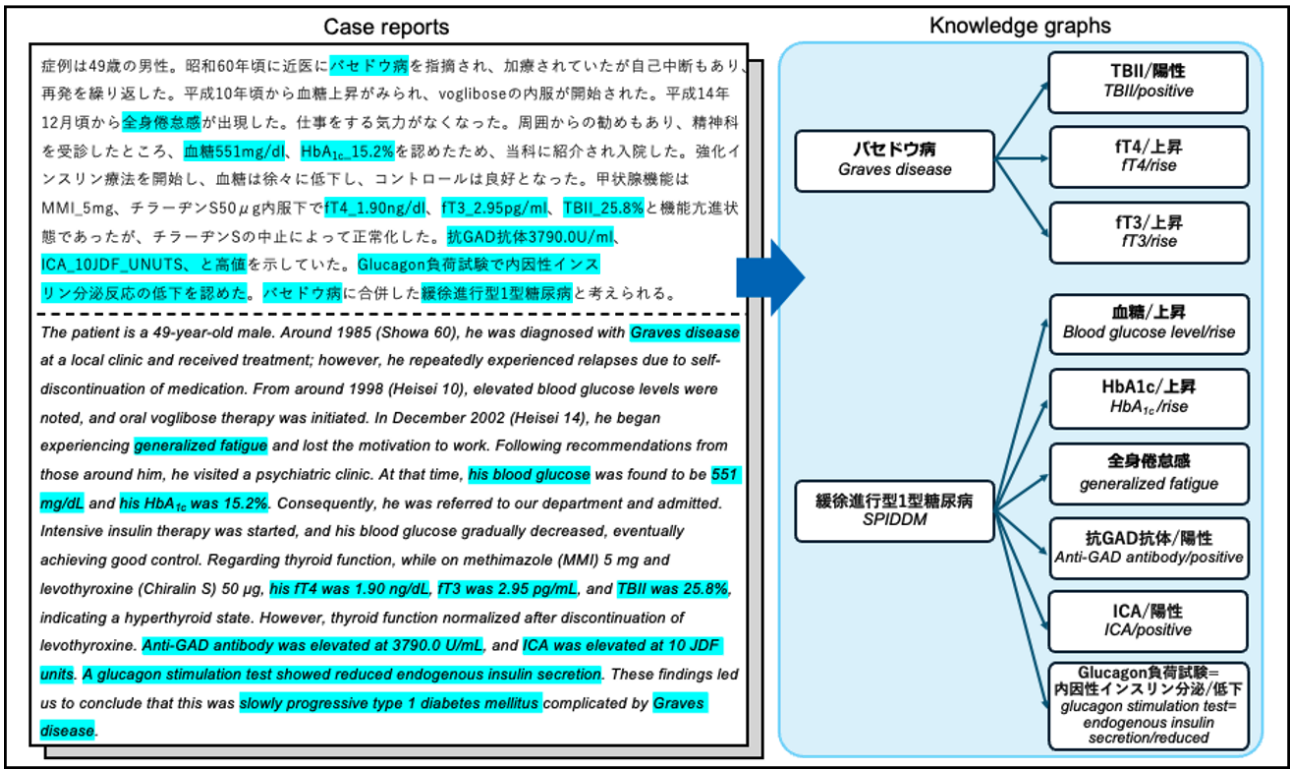

Using the data from J-CaseMap, we created a corpus of approximately 150,000 pairs of sentences. Six nurses identified segments corresponding to the normalized "term expressions" manually from the original case report sentences. This corpus was used to define clinically essential expressions in sentences as descriptions and definitions of nodes in the knowledge graph of J-CaseMap. This study aimed to extract these clinically meaningful expressions in sentences. Converting expressions from extracted sentences into normalized terminology was beyond the scope of this study.

The medical text corpus was used to fine-tune the encoder and decoder models. We used 0.56 GB of medical domain texts, which combined real-world clinical texts, such as discharge summaries (0.53 GB), and textbook-style medical texts, such as medical textbooks (0.03 GB), to construct a unique DeBERTa

XSL•FO
RenderX

[34] pretraining model and an additional pretraining model with complex medical knowledge.

Annotation Procedure

To minimize the subjectivity inherent in the definition of "clinically important expressions," we designed our annotation scheme on the basis of the J-CaseMap knowledge graph. For each case report, we first regarded as "descriptions important for clinical reasoning" those concepts that the physicians editing J-CaseMap had adopted as component nodes of the knowledge graph for that case (this node selection process itself involved multiple rounds of review by several physicians). We then constructed the corpus by marking, in the original case report text, the descriptions that corresponded to these node expressions. The annotation task was performed by 6 nurses, with 2 nurses independently annotating each case report. When their results did not agree, the 2 annotators discussed the discrepancies and revised the annotations until full consensus was reached. However, because we did not retain the preadjudication annotations, we could not compute the initial interannotator agreement (IAA). Accordingly, the resulting labels should be interpreted as an adjudicated consensus reference standard rather than an objective ground truth. Thus, our evaluation primarily quantifies agreement with this consensus, and some apparent "errors" may reflect clinically acceptable alternative span selections rather than model failures.

Creating Correct Answer Label Data

Figure 3 illustrates the process for generating correct label data and handling input and output for the model.

XSL•FO
RenderX

Figure 3. Creation of gold-standard labels and model inputs for instruction tuning versus token classification. The example sentence "感染性心内膜炎 (IE）はその経過" ("The course of infective endocarditis (IE)") is taken from a Japanese clinical case report. In the instruction tuning setting (left), clinically important concepts within the sentence are enclosed by special markers (@@ and ##) in the output text. In the token classification setting (right), the same sentence is split into characters, and binary labels are assigned to each character (1 for characters within clinically essential expressions and 0 otherwise), which are then mapped to tokens based on the tokenization. Character-level labels are used to calculate the weighted soft matching score, enabling fair comparison across models with different tokenizers.

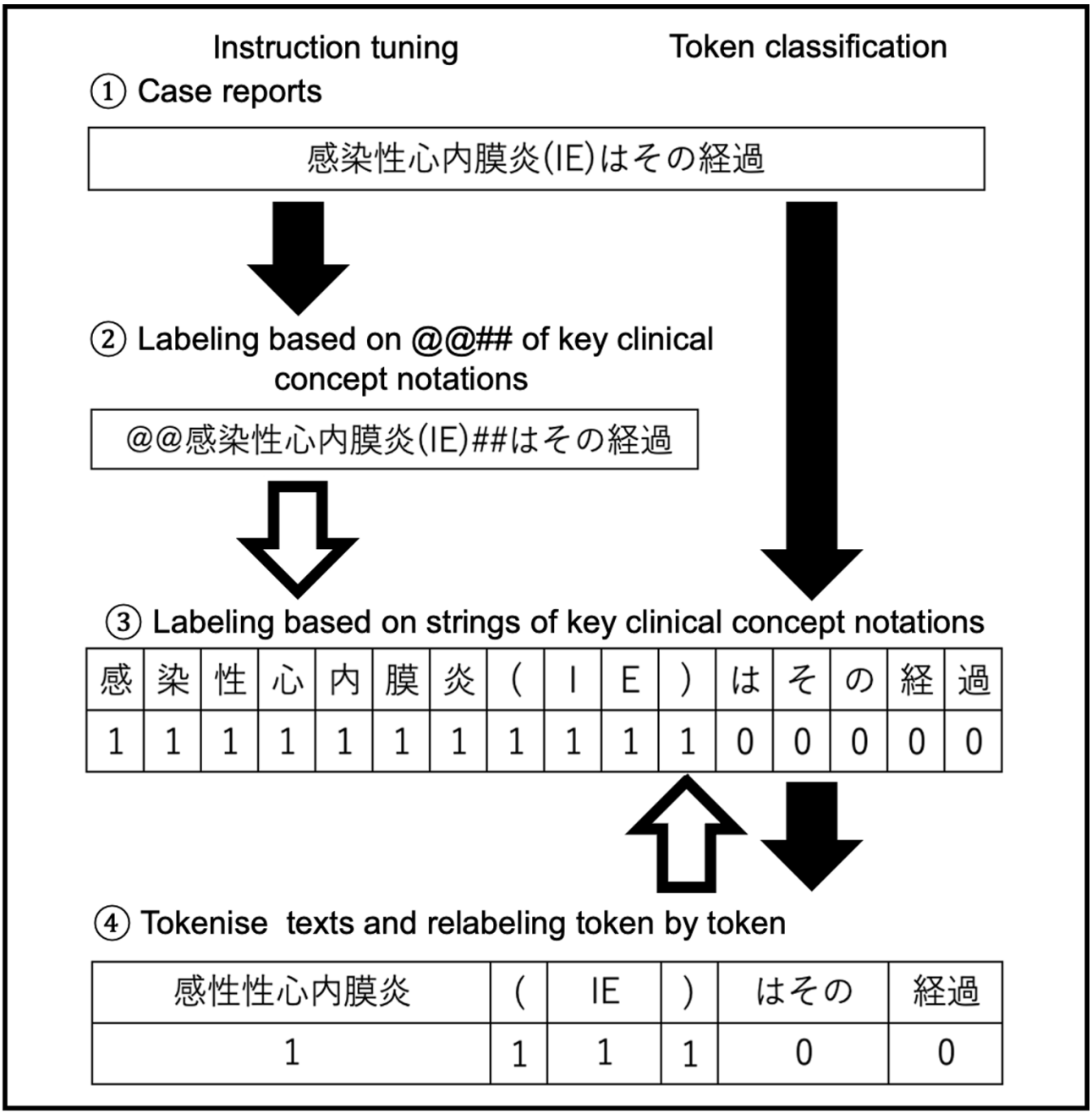

In instruction tuning, following Wang et al [23] and Keloth et al [28], ① case report sentences are provided as input to the models, and ② important clinical concepts are labeled by enclosing them with @@ and ## markers. These marked expressions are then extracted, and performance evaluation is conducted after splitting the text into characters (③). We enabled byte fallback to ensure robust handling of out-of-vocabulary characters.

In token classification, ① case report sentences are split into characters (③); a label of "1" is assigned to character strings identified as critical clinical concepts, and "0" is assigned to all other character strings. The case report text is then tokenized and labeled with "1" for tokens representing important clinical concepts and "0" for all other tokens. At this point, if the byte fallback function of each model is enabled, expressions that are not in the vocabulary set of the tokenizer are subdivided, making it difficult to label them (eg, "蛔 (*Ascaris*)" → "'<0xE8>', '<0x9B>', '<0x94>', '虫'"). Notably, the encoder models used in this study originally had byte fallback disabled by default. Accordingly, for consistency across models and to avoid label fragmentation, the byte fallback function was disabled in all encoder and decoder models used for token classification, and unknown tokens were allowed. Next, the case report sentences are input to multiple models, and the labels after step ④

XSL•FO
RenderX

represent the model outputs. Performance evaluation is conducted on a character-by-character basis.

Performance was evaluated on the labels generated after the aforementioned steps for both tasks. The performance evaluation used a weighted soft matching score inspired by the Hölder mean.

∑i=1mailp1p

For each extracted entity, l is the length of the entity string, m is the number of markers assigned to that entity, and a_i is the string length of the i-th marker. Finally, p represents the penalty strength.

This evaluation metric imposes a penalty on the extraction process when segmented parts are present. Each extraction target was scored using this weighted soft matching score; furthermore, considering that extracting longer targets is more challenging, the scores were aggregated using a length-weighted average based on the target length. This assigns relatively higher scores to more challenging extractions. This evaluation metric is designed such that when p=1, it corresponds to a character-level evaluation. As p increases, there is a penalty for the extracted part if there is a split in the extracted marker, and when $p=\infty$, it converges to the ratio of the longest extracted part in the extracted target.

Performance was also assessed on a marker-by-marker basis (marker matching score)—a simple average that does not account for target length—as well as on token-split and character-split bases.

Results are presented in Tables S1-S3 in Multimedia Appendix 1. In this study, the number of extraction targets differed for each case; therefore, we evaluated all target extraction markers contained in multiple instances in the test set. Case-specific performance evaluations are presented in Tables S4-S7 in Multimedia Appendix 1.

Using the aforementioned evaluation metrics, we computed true positives, false positives, and false negatives, focusing on the target extraction spans, and then calculated the F_1-score from these values. We did not count true negatives for strings outside the extraction targets; therefore, even if instruction tuning produces outputs that differ from the source text in nonextraction spans, such differences do not affect the evaluation.

Training Settings

Token classification and instruction tuning were conducted on the corpus of important clinical concepts using the following models:

- General-purpose models pretrained on Japanese text include:
  - RoBERTa [35] (rinna/japanese-roberta-base; hereinafter RoBERTa-base)
  - DeBERTa (izumi-lab/deberta-v2-base-japanese; hereinafter DeBERTaV2-base)
  - GPT2 [36] (llm-jp/llm-jp-13b-v1.0, llm-jp/llm-jp-13b-instruct-full dolly_en-dolly_ja-ichikara_003_001-oasst_en-oasst_ja-v1.1; hereinafter LLM-jp-13B v1.0)
  - GPT-NeoX [37] (rinna/japanese-gpt-neox-3.6b, rinna/japanese-gpt-neox-3.6b-instruction-sft-v2; hereinafter GPT-NeoX-3.6B)
- Domain-specific models pretrained on medical texts include:
  - RoBERTa (alabnii/jmedroberta-base-sentencepiecevocab50000; hereinafter JMedRoBERTa)
  - DeBERTa pretrained on original medical texts (JMedDeBERTa(s))
- Models additionally pretrained on general-purpose Japanese text include:
  - LLaMA2 [38]: (tokyotech-llm/Swallow-7bhf, tokyotech-llm/Swallow-7b-instruct-v0.1, tokyotech-llm/Swallow-13b-hf, tokyotech-llm/Swallow-13b-instruct-v0.1, tokyotech-llm/Swallow-70b-hf, tokyotech-llm/Swallow-70b-instruct-v0.1, hereinafter Swallow-7B, Swallow-13B, and Swallow-70B)
  - Mistral [39]: (tokyotech-llm/Swallow-MS-7b-v0.1, tokyotech-llm/Swallow-MS-7b-instruct-v0.1, hereinafter Swallow-MS-7B)
- Models further pretrained on medical texts include:
  - DeBERTa (izumi-lab/deberta-v2-base-japanese) with additional pretraining on medical texts (hereinafter JMedDeBERTa(c))

In the list mentioned earlier, the descriptions in parentheses represent the model identifiers on Hugging Face, the largest data-sharing platform for LLMs.

Prior research has reported that even slight modifications in prompt design can markedly alter model performance, indicating that developing effective prompts requires domain-specific expertise [2]. On the basis of the studies by Wang et al [23] and Keloth et al [28], we designed the prompt for instruction tuning as follows: “以下は, タスクを説明する指示と, 文脈のある入力の組み合わせです。”“要求を適切に満たす応答を書きなさい。\n\n” “### 指示:\n与えられた文に対して重要臨床概念を@@と##で強調表示し, 抽出してください。” “重要臨床概念が存在しない場合は同じ文を出力してください。\n\n### 入力:{input}\n\n### 応答:” (*“The following is a combination of instructions describing a task and contextual input.” “Write a response that appropriately fulfills the request.” “###Instruction:\n For the given sentence, highlight the key clinical concepts using @@ and ##, and extract them.” “If there are no key clinical concepts, output the same sentence as it is. \n\n### Input: {input}\n\n### Response:”*) After fine-tuning, we compared these 17 models.

Constructing the Pretraining and Additional Pretraining Models Using Medical Texts

As shown in Figure 4, 0.56 GB of medical domain text was used to construct JMedDeBERTa(s) and JMedDeBERTa(c). When constructing JMedDeBERTa(s), SentencePiece [40] was used as the tokenizer.

XSL•FO
RenderX

Figure 4. Construction of JMedDeBERTa(s) and JMedDeBERTa(c) using Japanese medical texts. JMedDeBERTa(s) is an encoder model with the DeBERTaV2-base architecture pretrained from scratch on 0.56 GB of Japanese medical text (0.53 GB clinical documents such as discharge summaries and 0.03 GB medical textbooks), then fine-tuned on the extended Clinical Concept Recognition (E-CCR) data set derived from J-CaseMap. JMedDeBERTa(c) starts from a general-domain DeBERTaV2-base model pretrained on 362 GB of mixed-domain Japanese text, continues pretraining on the same 0.56 GB medical corpus, and is then fine-tuned on the same E-CCR data set.

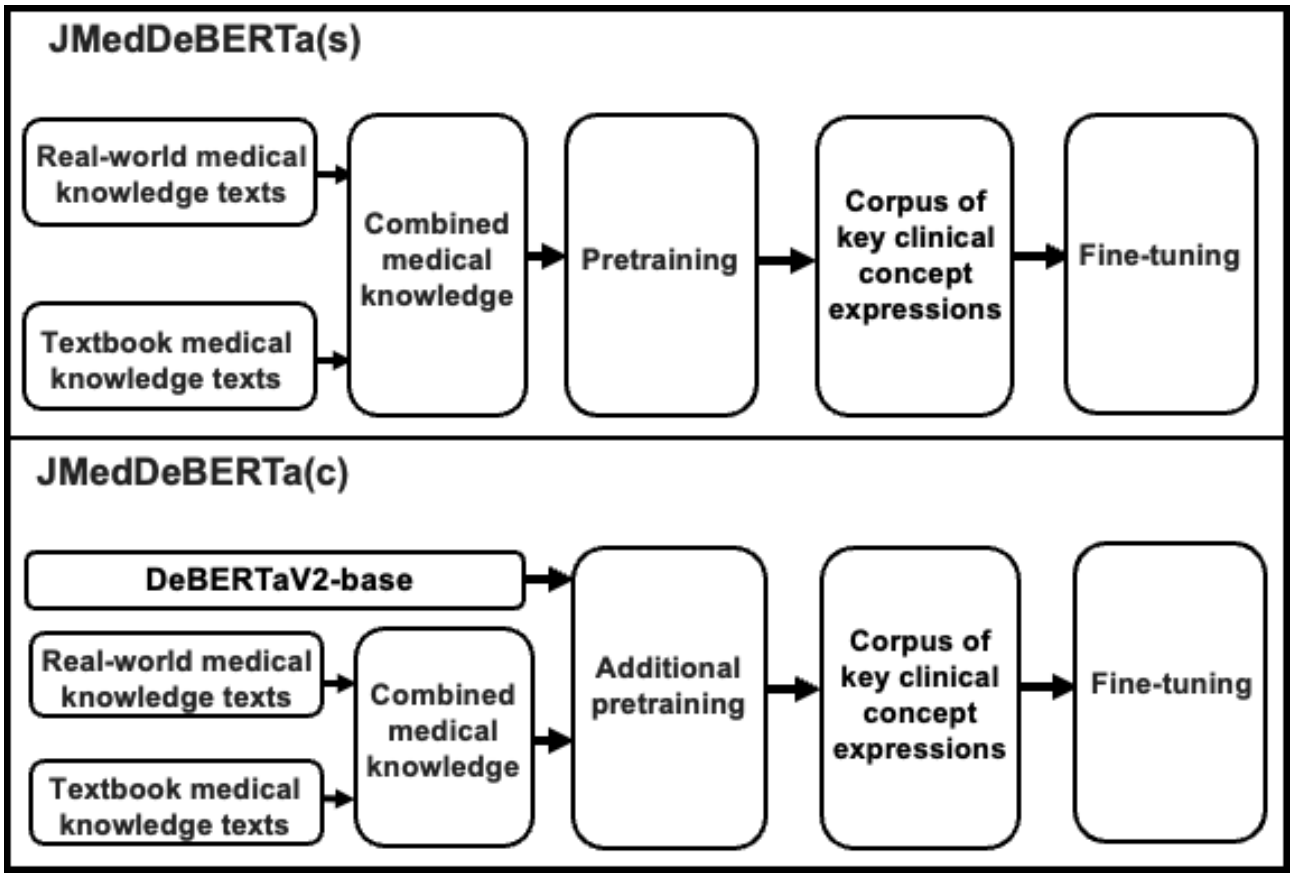

For JMedDeBERTa(s) and JMedDeBERTa(c), pretraining and additional pretraining were conducted on 8 V100 GPUs (16 GB each). The medical corpus was split into training and validation sets at a ratio of 0.8:0.2, and the models were trained using the hyperparameters listed in Table S8 in Multimedia Appendix 1. The effective global batch size was 2048 sequences (32 per GPU×8 GPUs, with gradient accumulation of 8). Under this setting, the maximum of 60 training epochs corresponded to 81,960 optimizer update steps (1366 updates per epoch). Therefore, the learning curves in Figures S1 and S2 in Multimedia Appendix 1 represent the full 60-epoch training run, plotted up to approximately 80,000 update steps. In general, substantially larger datasets are required to learn robust language representations. As our domain corpus is relatively small, we additionally monitored the masked language modeling (MLM) loss for both JMedDeBERTa(s) and JMedDeBERTa(c) to verify that training proceeded appropriately.

Fine-Tuning

Overview

Figure 5 lists the 17 comparison models used for fine-tuning, and Table 1 provides an overview of the various models. For fine-tuning, the corpus of clinically significant concepts was divided into training, validation, and test sets at a ratio of 0.6:0.2:0.2, and 5-fold cross-validation was performed for 10 epochs (with early stopping). The 70B decoder model could be trained on an NVIDIA H100 (80 GB) GPU, while all other models were trainable on a single NVIDIA A6000 (48 GB) GPU. The hyperparameters for the encoder and decoder models were selected based on settings used in prior studies on NER tasks [28,29] and are detailed in Table S9 in Multimedia Appendix 1. Fine-tuning performance metrics included F1 score, recall, and precision, evaluated using the weighted soft matching score. Due to computational resource constraints, we applied low-rank adaptation (LoRA) to the decoder models and full fine-tuning to the encoder models. This choice was motivated by prior work showing that LoRA can achieve performance comparable to full fine-tuning [41].

XSL•FO
RenderX

Figure 5. Overview of the 17 encoder and decoder language models compared in this study. All models were fine-tuned to extract clinically important concepts from Japanese case reports in the J-CaseMap database and evaluated on the E-CCR task. Decoder models (LLM-jp-13B v1.0, GPT-NeoX-3.6B, Swallow-7B/13B/70B, Swallow-MS-7B) were fine-tuned by instruction tuning or token classification, while encoder models (RoBERTa-base, DeBERTaV2-base, JMedRoBERTa, JMedDeBERTa(s), JMedDeBERTa(c)) were fine-tuned by token classification only. All models used identical data splits and evaluation metrics.

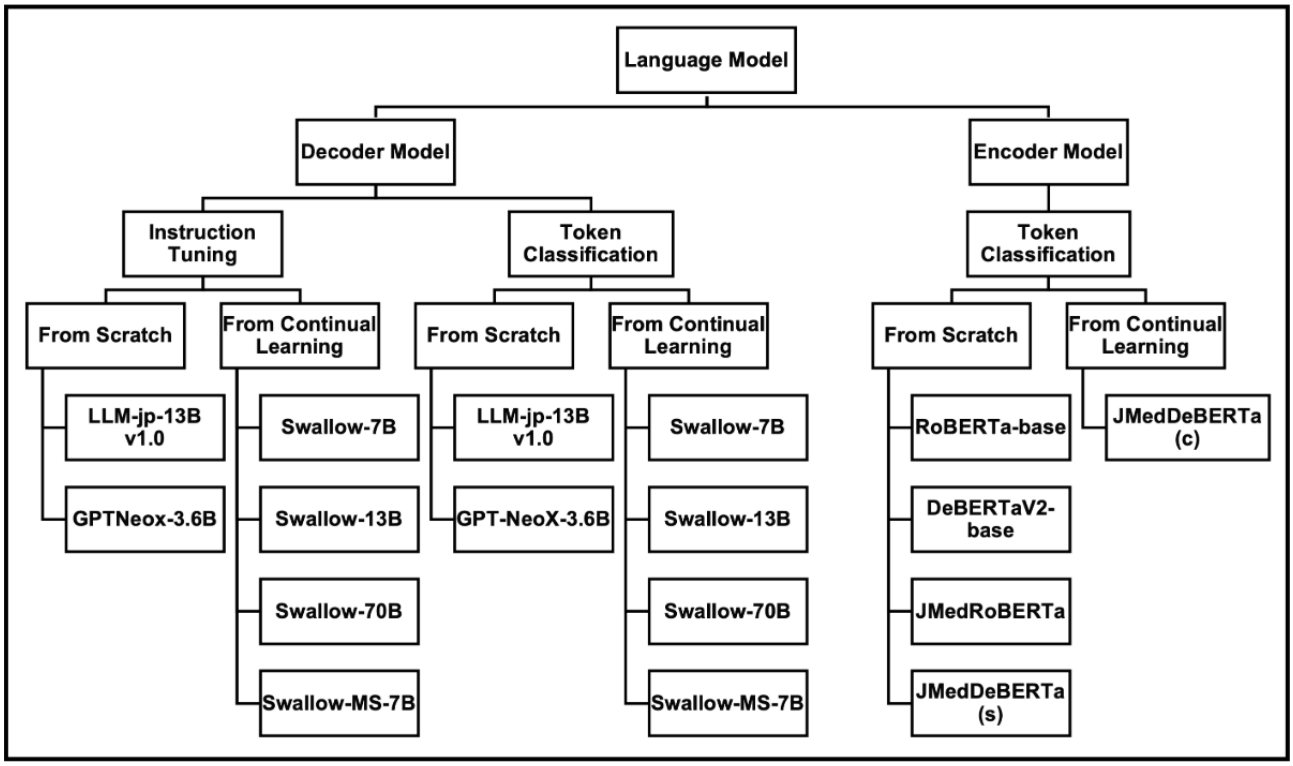

XSL•FO
RenderX

Table . Characteristics of the 17 encoder and decoder language models evaluated on the extended clinical concept recognition task in Japanese clinical case reports.

Models[a]	Architecture	Parameters	Pretraining methods	Pretraining corpus type	Pretraining corpus size	Vocab size
Decoder models						
GPT-NeoX-3.6B	GPT-NeoX-3.6B	3.6B	From scratch	General (Japanese)	312.5B tokens	32,000
LLM-jp-13B v1.0	GPT2-13B	13B	From scratch	General (Japanese)	300B tokens	50,570
Swallow-7B	LLaMA2-7B	7B	Continual learning	English+general (Japanese)	+100B tokens	43,176
Swallow-13B	LLaMA2-13B	13B	Continual learning	English+general (Japanese)	+100B tokens	43,176
Swallow-70B	LLaMA2-70B	70B	Continual learning	English+general (Japanese)	+100B tokens	43,176
Swallow-MS-7B	Mistral-7B	7B	Continual learning	English+general (Japanese)	+105B tokens	42,800
Encoder models						
DeBERTaV2-base	DeBERTa V2-base	110M	Continual learning	General+finance	357 GB+5.2 GB	32,000
RoBERTa-base	RoBERTa-base	110M	From scratch	General (Japanese)	75 GB	32,000
JMedRoBERTa	RoBERTa-base	124M	From scratch	Medical (Japanese)	1.8 GB	50,000
JMedDeBERTa(s)	DeBERTa V2-base	125M	From scratch	Medical (Japanese, this study)	0.56 GB	32,000
JMedDeBERTa(c)	DeBERTa V2- base	110M	Continual learning	General+medical (this study)	362 GB (general)+0.56 GB (medical)	32,000

[a] Models are grouped by architecture (decoder or encoder) and described by parameter count, pretraining methods, pretraining corpus type and size, and vocabulary size. All models were subsequently fine-tuned on the data set constructed from the J-CaseMap database.

For decoder-only models in the token classification setting, we retained the default causal (left-to-right) self-attention behavior used during pretraining; thus, the hidden state at position t could attend only to tokens at positions ≤t. In implementation, we used only the standard padding attention mask to ignore padded tokens and did not modify the causal mask or the attention pattern (eg, no full unmasking, no layer-wise causal mask removal, and no other bidirectionality-enabling modification) [42,43]. We attached a token-level classification head to the final-layer hidden states and computed label logits for all token positions in parallel from a single forward pass under this causal mask.

Pretraining Methods

"From scratch" indicates that the base model was pretrained from random initialization on the corpus listed under "pretraining corpus type/size." "Continual learning" indicates that the base model was initialized from an existing pretrained checkpoint and further pretrained on additional corpora (eg, Swallow models continued training on approximately 100 billion tokens of English plus general Japanese text following LLaMA2 or Mistral; JMedDeBERTa(c) continued training on 0.56 GB of medical text after DeBERTaV2-base).

Pretraining Corpus Type

"General" indicates general-domain Japanese text; "medical" indicates the 0.56 GB of Japanese medical documents used in this study; "general+finance" indicates a mixture of general and financial text; and "general+medical" indicates the combination of the general corpus used for DeBERTaV2-base and the 0.56 GB medical corpus used for JMedDeBERTa(c).

External Evaluation on a Public Clinical NER Dataset

To assess the reproducibility and generalizability of our findings using a public benchmark, we conducted an additional evaluation on Medical Report Named Entity Recognition for positive disease (MRNER disease), a Japanese clinical NER dataset used in JmedBench [44]. MRNER disease comprises 100 clinical documents (50 case reports and 50 radiology reports) and 810 annotated entities. The task focuses on extracting relatively short disease names actually observed in patients, corresponding to a conventional clinical NER setting rather than our E-CCR task, which targets longer, diagnostically important phrases.

Although the number of annotated entities is relatively small for training large models, we considered the dataset sufficient for benchmark evaluation. We applied the same preprocessing, fine-tuning procedures, and evaluation framework as in the main analysis. Specifically, all 17 models were fine-tuned on MRNER disease and evaluated using the weighted soft matching score with fragmentation penalty parameter p=1, 1.5, 2, 100. Detailed scores for each model (F_1-score, recall, and precision at each p) are reported in Tables S10-S17 in Multimedia Appendix 1.

XSL•FO
RenderX

Ethical Considerations

In this study, to protect patient privacy, all personally identifiable information was removed from the datasets used. The case reports of the Japanese Society of Internal Medicine used to construct the J-CaseMap database are abstracts of case reports presented at regional meetings of the Society that do not contain personal information, and they have already been made publicly available on the web; therefore, no additional ethical review was required. In contrast, for the development of the language model using clinical record texts, to protect patient privacy, we used data from which all traceable personally identifiable information had been removed from the analysis dataset, and the study was conducted in accordance with the Declaration of Helsinki with the approval of the Ethics Committee of the Faculty of Medicine and Graduate School of Medicine, The University of Tokyo (approval 2018-NI). As this study analyzed only anonymized case report texts (secondary-use data) and involved no direct intervention with or interaction toward human subjects, the ethics committee determined that neither individual informed consent from patients nor financial compensation was required. Neither this article nor the supplementary materials contains any images or other materials that could be used to identify individual patients or clinicians.

Results

Overview

Table 2 summarizes the average results from 5-fold cross-validation of the 17 comparison models. Tables S1-S3 in Multimedia Appendix 1 display the average results for markers, characters, and tokens across all cases. Tables S4-S7 in Multimedia Appendix 1 show average results for weighted soft matching scores, marker matching scores, characters, and tokens for each case. Model performance declined as the penalty for splits in extracted parts increased. When comparing the results for $p=1$ and p=100, the decoder model GPT-NeoX-3.6B was impacted, with an F_1-score reduction of −0.047.

For other decoder models using token classification, the F_1-score ranged from −0.042 to −0.046. Encoder model F_1-scores ranged from −0.032 to −0.036. When using instruction tuning, the F_1-score for all decoder models ranged from −0.022 to −0.030.

Regardless of the penalty size, the encoder model using token classification, that is, JMedDeBERTa(s), achieved the highest mean scores, while the top encoder models showed very similar performance. The decoder model using instruction tuning, that is, GPT-NeoX-3.6B, showed the worst results. Among the decoder models, Swallow-70B, which exhibited the best results for instruction tuning and token classification, could not produce better results than the encoder models.

Next, we compared and verified the results for p=1, which performed the best. Among all the models, the encoder model using token classification, that is, JMedDeBERTa(s), achieved the highest performance: *mean* F_1-score 0.758 (SD 0.002), mean recall 0.768 (SD 0.011), and mean precision 0.749 (SD 0.009). In contrast, the decoder model GPT-NeoX-3.6B, under instruction tuning, recorded the lowest performance: mean F_1-score 0.621 (SD 0.023), mean recall 0.582 (SD 0.044), and mean precision 0.669 (SD 0.006). The encoder model JMedDeBERTa(s), with 125 million parameters, is approximately 1/70th the size of Swallow-70B with 70 billion parameters. The decoder model Swallow-70B, under instruction tuning, achieved mean F_1-score 0.713 (SD 0.016), mean recall 0.680 (SD 0.035), and mean precision 0.751 (SD 0.020). The decoder model using token classification, that is, Swallow-70B, achieved mean F_1-score 0.739 (SD 0.005), mean recall 0.744 (SD 0.013), and precision mean 0.734 (SD 0.003). The metric scores of JMedDeBERTa(s) were better than those of Swallow-70B by margins of F_1-score=+0.045 and recall=+0.088 for instruction tuning and F_1-score=+0.020 and recall=+0.024 for token classification.

XSL•FO
RenderX

Table . Performance of 17 encoder and decoder models in terms of weighted soft matching score on the J-CaseMap test set.[a]

Models[b]	p=1, mean (SD)			p=1.5, mean (SD)			p=2, mean (SD)			p=100, mean (SD)		
	F_1-score	Recall	Precision	F_1-score	Recall	Precision	F_1-score	Recall	Precision	F_1-score	Recall	Precision
Decoder models												
Instruction tuning												
GPT-Neox-3.6B	0.621 (0.023)	0.582 (0.044)	0.669 (0.006)	0.605 (0.022)	0.566 (0.042)	0.651 (0.005)	0.599 (0.022)	0.560 (0.042)	0.645 (0.005)	0.591 (0.022)	0.553 (0.041)	0.637 (0.005)
LLM-jp-13B v1.0	0.673 (0.011)	0.602 (0.022)	*0.763 (0.007)* [c]	0.660 (0.011)	0.588 (0.021)	*0.753 (0.008)* [c]	0.656 (0.011)	0.584 (0.021)	*0.750 (0.008)* [c]	0.651 (0.011)	0.577 (0.021)	*0.746 (0.008)* [c]
Swallow-7B	0.697 (0.009)	0.667 (0.037)	0.732 (0.027)	0.681 (0.007)	0.652 (0.037)	0.717 (0.032)	0.676 (0.006)	0.647 (0.037)	0.712 (0.034)	0.670 (0.005)	0.640 (0.037)	0.706 (0.036)
Swallow-13B	0.706 (0.010)	0.668 (0.024)	0.749 (0.016)	0.692 (0.011)	0.652 (0.024)	0.739 (0.019)	0.688 (0.011)	0.647 (0.023)	0.735 (0.020)	0.682 (0.011)	0.640 (0.023)	0.731 (0.021)
Swallow-70B	0.713 (0.016)	0.680 (0.035)	0.751 (0.020)	0.699 (0.015)	0.665 (0.034)	0.738 (0.023)	0.694 (0.015)	0.660 (0.034)	0.734 (0.024)	0.688 (0.015)	0.653 (0.034)	0.730 (0.025)
Swallow-MS-7B	0.693 (0.018)	0.648 (0.046)	0.747 (0.022)	0.679 (0.017)	0.633 (0.044)	0.735 (0.024)	0.675 (0.017)	0.628 (0.043)	0.732 (0.025)	0.669 (0.017)	0.621 (0.042)	0.727 (0.026)
Token classification												
GPT-Neox-3.6B	0.725 (0.004)	0.731 (0.014)	0.719 (0.007)	0.698 (0.005)	0.683 (0.014)	0.714 (0.007)	0.689 (0.005)	0.668 (0.014)	0.713 (0.008)	0.678 (0.005)	0.649 (0.014)	0.710 (0.008)
LLM-jp-13B v1.0	0.732 (0.003)	0.742 (0.017)	0.722 (0.010)	0.707 (0.004)	0.696 (0.017)	0.718 (0.010)	0.698 (0.004)	0.681 (0.016)	0.716 (0.010)	0.688 (0.004)	0.663 (0.016)	0.714 (0.010)
Swallow-7B	0.732 (0.003)	0.733 (0.010)	0.731 (0.008)	0.706 (0.004)	0.686 (0.012)	0.727 (0.009)	0.697 (0.004)	0.671 (0.013)	0.725 (0.009)	0.686 (0.005)	0.653 (0.013)	0.723 (0.009)
Swallow-13B	0.735 (0.002)	0.739 (0.013)	0.730 (0.008)	0.710 (0.003)	0.696 (0.012)	0.725 (0.008)	0.702 (0.003)	0.683 (0.012)	0.724 (0.008)	0.692 (0.003)	0.666 (0.012)	0.721 (0.009)
Swallow-70B	0.739 (0.005)	0.744 (0.013)	0.734 (0.003)	0.714 (0.006)	0.698 (0.014)	0.730 (0.003)	0.706 (0.007)	0.684 (0.015)	0.729 (0.003)	0.695 (0.007)	0.666 (0.015)	0.727 (0.003)
Swallow-MS-7B	0.736 (0.002)	0.755 (0.010)	0.719 (0.008)	0.711 (0.002)	0.709 (0.009)	0.714 (0.008)	0.703 (0.002)	0.694 (0.009)	0.712 (0.008)	0.692 (0.002)	0.676 (0.009)	0.709 (0.008)
Encoder models												
Token classification												
DeBERTaV2-base	0.755 (0.006)	0.761 (0.015)	0.750 (0.006)	0.738 (0.007)	0.728 (0.016)	0.747 (0.006)	0.732 (0.007)	0.718 (0.016)	0.746 (0.006)	0.724 (0.008)	0.704 (0.017)	0.744 (0.006)
RoBERTa-base	0.748 (0.003)	0.748 (0.013)	0.749 (0.009)	0.729 (0.003)	0.713 (0.012)	0.746 (0.009)	0.722 (0.003)	0.701 (0.012)	0.745 (0.009)	0.714 (0.003)	0.687 (0.011)	0.743 (0.009)

XSL•FO
RenderX

Models[b]	p=1, mean (SD)			p=1.5, mean (SD)			p=2, mean (SD)			p=100, mean (SD)		
	F_1-score	Recall	Precision	F_1-score	Recall	Precision	F_1-score	Recall	Precision	F_1-score	Recall	Precision
JMedRoBERTa	0.756 (0.003)	*0.769 (0.015)* [c]	0.743 (0.009)	0.735 (0.003)	0.733 (0.015)	0.737 (0.010)	0.728 (0.003)	0.722 (0.016)	0.735 (0.010)	0.719 (0.004)	0.707 (0.016)	0.732 (0.011)
JMedDeBERTa(s)	*0.758 (0.002)* [d], [c]	0.768 (0.011)	0.749 (0.009)	*0.740 (0.002)* [c]	*0.734 (0.012)* [c]	0.746 (0.010)	*0.734 (0.002)* [c]	0.724 (0.012)[c]	0.744 (0.010)	*0.726 (0.002)* [c]	*0.710 (0.012)* [c]	0.743 (0.010)
JMedDeBERTa(c)	0.757 (0.004)	0.764 (0.014)	0.751 (0.010)	0.739 (0.004)	0.730 (0.014)	0.748 (0.010)	0.732 (0.004)	0.719 (0.014)	0.747 (0.010)	0.724 (0.004)	0.705 (0.014)	0.746 (0.010)[c]

[a]F_1-score, recall, and precision are shown for 4 conditions of the Hölder-type parameter p (1, 1.5, 2, 100), which controls the strength of the fragmentation penalty (p=1 corresponds to pure character-level evaluation; larger p penalizes split predictions of long target spans more heavily).

[b]Model performance is calculated using the weighted soft matching score, evaluating all extraction targets (clinical concepts) in the test set collectively.

[c]Highest score among the comparative models.

[d]Highest performance.

Next, we compared the performance of the decoder models. In the instruction tuning task, Swallow-70B achieved the highest mean F_1-score 0.713 (SD 0.016) and mean recall 0.680 (SD 0.035), whereas GPT-NeoX-3.6B achieved the lowest mean F_1-score 0.621 (SD 0.023) and mean recall 0.582 (SD 0.044). In the token classification task, Swallow-70B achieved the highest mean F_1-score 0.739 (SD 0.005), and Swallow-MS-7B achieved the highest mean recall 0.755 (SD 0.010), whereas GPT-NeoX-3.6B achieved the lowest mean F_1-score 0.725 (SD 0.004) and mean recall 0.731(SD 0.014). In the LLAMA2 decoder model, increasing parameters from 7 billion to 70 billion enhanced performance: F_1-score +0.016 and recall +0.013 for the instruction tuning task, and F_1-score +0.007 and recall +0.010 for the token classification task.

Finally, we compared the various encoder models. JMedDeBERTa(s) achieved the numerically highest mean F_1-score (0.758, SD 0.002), and JMedRoBERTa achieved the highest mean recall (0.769, SD 0.015), whereas RoBERTa-base showed the lowest performance (mean F_1-score 0.748, SD 0.003; and mean recall 0.748, SD 0.013). Notably, among encoder models, the DeBERTa family models (JMedDeBERTa(s), JMedDeBERTa(c), and DeBERTaV2-base) achieved the highest mean scores. However, differences within this top tier were marginal ($\Delta F_1 \leq 0.003$ at p=1) and should be interpreted as effectively comparable, given the estimated annotation uncertainty (4% case-level corrections in our sampled review). DeBERTaV2-base achieved a higher F_1-score (+0.007) than RoBERTa-base, and JMedDeBERTa(s) achieved a higher F_1-score (+0.003) than JMedRoBERTa. JMedRoBERTa achieved a higher F_1-score (+0.007) than RoBERTa-base.

Performance on a Public Clinical NER Benchmark

Consistent with the E-CCR results on J-CaseMap, encoder models outperformed decoder models on the MRNER disease benchmark (Table S10 in Multimedia Appendix 1). The best-performing encoder model was DeBERTaV2-base, which achieved an mean F_1-score of 0.794 (SD 0.018) at p=1, closely followed by JMedDeBERTa(c) (mean F_1-score 0.788, SD 0.016) and JMedDeBERTa(s) (mean F_1-score 0.788, SD 0.027). Among decoder models, LLM-jp-13B v1.0 in the token classification setting showed the highest performance (mean F_1-score 0.694, SD 0.029), whereas instruction-tuned GPT-NeoX-3.6B performed worst (mean F_1-score 0.373, SD 0.067). These results reinforce the main conclusion that, for Japanese clinical concept extraction, encoder-based token classification is consistently stronger than decoder-based approaches—whether instruction-tuned or token classification models—under the same span-level evaluation.

Discussion

Principal Findings

A key finding of this study is that, under the experimental conditions, encoder models consistently outperformed decoder models; no decoder model outperformed any encoder model.

On the J-CaseMap dataset, JMedDeBERTa(s)—an encoder model pretrained on domain-specific medical text—achieved the highest performance (F_1-score=0.758, SD 0.002), whereas the instruction-tuned GPT-NeoX-3.6B produced the worst. While performance generally declined as the fragmentation penalty increased, no consistently severe degradation was observed. On the MRNER disease dataset, the general-domain DeBERTaV2-base yielded the highest mean F_1-score. The differences among the medical-domain JMedDeBERTa(s) and JMedDeBERTa(c) variants were minimal, suggesting that the benefit of domain-specific pretraining was limited in this context. Overall, under the evaluated instruction tuning setup (a single prompt template requiring inline @@/## boundary tags), token classification encoders were more effective than instruction-tuned decoders in our experiments. However, because the generative setting additionally required faithful text reproduction plus precise character-level boundary-tag placement, the observed performance gap may partly reflect the difficulty of this particular output representation rather than model architecture alone. We did not evaluate alternative prompting strategies (eg, structured span lists or JSON outputs) or reasoning-augmented prompting (eg, chain-of-thought),

XSL•FO
RenderX

which could affect generative performance. Importantly, encoder-based token classification is nonautoregressive and can assign labels to all tokens in parallel (ie, in a single forward pass), whereas decoder-based extraction via instruction tuning typically relies on autoregressive generation, whose runtime scales with the number of generated tokens. This architectural difference is expected to yield lower and more predictable inference latency and higher throughput for encoder-based token classification, which is particularly relevant for time-sensitive clinical workflows. Although we did not benchmark wall-clock inference latency in this study, we highlight this implication as a practical advantage of encoder-based approaches. Moreover, the smaller memory footprint of encoder models can make on-premise deployment more feasible when handling sensitive clinical text.

If, due to hallucination, the decoder extracts a string that differs from the original only in trivial ways—such as the insertion or deletion of a very small number of characters—then counting such negligible discrepancies as correct could, in principle, yield higher practical accuracy. However, as shown in Table S18 in Multimedia Appendix 1, after removing @@/## from both the gold-set inputs and the model outputs, the 5-fold mean exact match rate was 98.40% (SD 0.42%). Moreover, inspection of the mismatched cases indicated that most differences were minor: considering mismatches only, the mean SequenceMatcher similarity was 0.993 (SD 0.002), the mean Levenshtein distance was 7.54 (SD 1.99), and the mean length difference was −2.30 (SD 1.82). These findings suggest that even under a lenient assumption in which all such cases are treated as correct, the conclusion stated earlier would not change.

In recent years, the development of biomedical language models prioritized increasingly complex architectures and a rapid, often exponential growth in parameter count, in line with the consensus in LLM research [45,46]. However, some studies have reported that even a substantial increase in the parameter count of the decoder models does not necessarily improve model performance beyond that of encoder models [23,24], which is consistent with the findings of this study.

One possible explanation is that token classification constitutes a more constrained formulation of the prediction problem than free-form text generation via instruction tuning. Specifically, instead of having to localize the target span while generating fluent text, the model only needs to assign a binary label to each token. As the output space is thus more restricted, encoder models with far fewer parameters may plausibly compete with much larger decoder models. At the very least, a similar tendency was observed in both our main experiments and the additional experiments on the MRNER disease dataset. In our study, we demonstrated that even for tasks beyond conventional NER, such as E-CCR that targets longer phrases, encoder models with far fewer parameters are far more effective than decoder models. These findings offer valuable insight into model selection under practical constraints for general research institutions with limited computational resources, representing one of the substantial contributions of our work. Prior research has reported that decoder models require substantial computational resources and incur high costs, making them unsuitable when such resources are constrained [2].

Our goal was to compare off-the-shelf encoder versus decoder backbones under commonly used fine-tuning recipes, without altering the pretrained attention mechanism. An additional factor contributing to the encoder-decoder gap is attention directionality. Encoder models compute token representations bidirectionally, whereas decoder-only models under the default causal mask cannot incorporate future tokens when predicting token-level labels. This limitation may particularly affect boundary decisions in span labeling. Prior work indicates that partially or fully relaxing the causal mask (eg, layer-wise causal mask removal) can markedly improve sequence labeling with decoder-only LLM [42,43]. Under the standard fine-tuning recipes considered in this study, encoder backbones outperform decoder-only backbones on token classification; however, relaxing the causal mask (eg, via layer-wise mask removal) may improve decoder-only models and help narrow this gap, which we leave for future work.

Second, a comparison of the evaluated decoder models showed that instruction-tuned variants consistently underperformed their token classification counterparts. This contrasts with prior work [28], which suggested that decoder models can match or even surpass encoder models when supported by carefully designed prompts. Our results indicate that, even within decoder architectures, token classification is more effective than instruction tuning for extraction tasks. Although we did not exhaustively explore all prompt designs, the E-CCR task requires extracting entire phrases corresponding to single clinical concepts. This is substantially more complex than conventional NER, making effective prompt design for instruction tuning considerably more challenging. Overall, these findings suggest that token classification provides a more robust approach for specialized extraction scenarios across both encoder and decoder models.

Among the various token classification–based encoder models investigated in this study, JMedDeBERTa(s) achieved the highest mean performance, whereas RoBERTa-base performed the worst. However, the differences among the top encoder models were very small and should be interpreted as comparable within the uncertainty implied by our annotation review. This is consistent with previous research [34] and indicates that the DeBERTa architecture is more effective than the RoBERTa architecture. In particular, DeBERTa uses the disentangled attention mechanism that allows for a more precise interpretation of the importance of word positions and the contextual relationships among words, as well as a masked decoder that predicts not only the masked tokens but also their positions [47,48]. These mechanisms likely enhance the understanding of linguistic structure and context, resulting in high performance even on E-CCR tasks.

With respect to the effect of domain-specific text versus general text, no significant difference was detected; however, the highest accuracy was achieved by the model pretrained on domain-specific documents, followed by the model that underwent additional pretraining on a generally pretrained model, and finally by the model pretrained solely on general text. For example, among the DeBERTa models, JMedDeBERTa(s)—pretrained on medical documents—outperformed both DeBERTaV2-base (pretrained

XSL•FO
RenderX

on general Japanese texts) and JMedDeBERTa(c), which was additionally pretrained on medical records. Although JMedDeBERTa(c) performed better than DeBERTaV2-base, it did not reach the performance level of JMedDeBERTa(s). Similarly, among the RoBERTa models, MedRoBERTa, which was pretrained on medical documents, outperformed RoBERTa-base, which was pretrained on general Japanese texts. However, the differences were minimal, indicating that the distinctions among these pretraining approaches are negligible.

Comparison to Prior Work

This observation aligns with previous studies. Kim et al [49] reported that a model pretrained on domain-specific Korean medical documents outperformed a general Korean language model. Conversely, Subies et al [33] found that a model pretrained on domain-specific Spanish medical records did not surpass the performance of a model pretrained on general Spanish texts or that of a multilingual model. As with our findings, the differences in model comparison results in these studies were minimal.

Similar to the minimal performance differences reported by Kim et al [49] and Subies et al [33], the slight differences in performance observed in our study may be attributed to the quantity and content of the texts used for both pretraining and additional pretraining. Specifically, only 0.56 GB of medical text was used to train the domain-specific models JMedDeBERTa(c) and JMedDeBERTa(s), compared to the 362 GB of text used to train the general-corpus DeBERTaV2-base model.

Although the pretraining corpus is small compared with typical general-language pretraining datasets, the MLM loss curves for JMedDeBERTa(s) and JMedDeBERTa(c) (Figures S1 and S2 in Multimedia Appendix 1) indicate stable optimization and convergence under our setting. We note, however, that convergence of the MLM loss does not by itself guarantee broad linguistic coverage or robustness across institutions and document types. A possible reason why effective learning may be achievable with a relatively small corpus is the qualitative nature of discharge summaries: compared with fragmented note types (eg, brief progress notes), discharge summaries are often more narrative and syntactically coherent, potentially enabling MLM pretraining to capture clinically salient vocabulary and patterns efficiently. Our findings indicate that, at least for the E-CCR task examined in this study, even a model pretrained on a very small domain-specific corpus can achieve performance comparable to, or slightly exceeding, that of a model trained on a large general corpus. This highlights the high data efficiency of domain-specific pretraining in this setting: even a small amount of domain-specific text may be sufficient to obtain performance on par with a model trained on a much larger general corpus. In contrast, our results also indicate that, even when only a limited amount of domain-specific text is available, a model trained solely on general text may still achieve satisfactory performance on the E-CCR task. This is presumably because, in the E-CCR task, the syntactic characteristics of the text play a crucial role even when specialized terminology has not been thoroughly acquired. For example, even if the model has not explicitly learned each specialized term, when processing a sentence such as “SAA was abnormally elevated to 1940 mg/mL, and FMF was suspected,” it may still be able to extract spans such as “SAA was abnormally elevated to 1940 mg/mL” and “FMF” by using syntactic information as clues. Therefore, training strategies should be flexibly adjusted according to the amount and type of data available.

External Benchmark and Implications for Domain-Specific Versus General Pretraining

Additional experiments using the MRNER disease dataset further support these conclusions. Even in this conventional clinical NER task, encoder models consistently outperformed decoder models; however, the general-domain DeBERTaV2-base slightly outperformed the domain-specific JMedDeBERTa(s) (eg, F_1=0.794, SD 0.018 vs 0.788, SD 0.027 at p=1). However, this difference was marginal and likely reflects multiple factors.

We interpret this result not as a lack of generalization ability in JMedDeBERTa(s) but primarily as a reflection of task mismatch and distribution shift. First, MRNER disease is a conventional clinical NER task targeting short disease name entities, whereas E-CCR, our primary evaluation target, focuses on longer, more complex, and clinically meaningful spans that often exceed typical NER boundaries. Therefore, the relative advantage of domain-specific versus general-domain pretraining can vary depending on specific task requirements. Additionally, distribution shifts arising from differences in document types, medical institutions, and annotation criteria may have also influenced these results.

The substantial disparity in the size and composition of pretraining data (362 GB vs 0.56 GB) further complicates this comparison. Although JMedDeBERTa(s) achieved the highest average performance on the E-CCR task, the differences among the top-tier encoder models were generally comparable within the margin of uncertainty suggested by our annotation review. Taken together, these findings suggest that while pretraining on a massive general-domain corpus provides a solid linguistic foundation, pretraining on a much smaller domain-specific corpus can achieve competitive performance—comparable to or slightly exceeding general-domain models—depending on the task constraints and linguistic characteristics of the target documents.

However, as the external benchmark consists of only 100 documents, these slight performance differences are not necessarily robust and should be interpreted with caution. Moreover, previous research has reported substantially low zero-shot and few-shot performance of pretrained LLMs on MRNER disease (entity-level F_1 ≤0.3). In contrast, the significantly higher scores achieved in this study highlight the critical importance of supervised fine-tuning in Japanese clinical NER, demonstrating that merely prompting non–fine-tuned LLMs is insufficient for achieving adequate performance in this context.

XSL•FO
RenderX

Differences in Characteristics Between Instruction Tuning and Token Classification

The differences in the properties of instruction tuning and token classification were analyzed from the perspective of the length of the extracted segments. Table S1 in Multimedia Appendix 1 presents evaluation results using the marker matching score, which does not account for the length of the extraction targets. In contrast to Table 2, where longer extracted segments yield higher scores, Table S1 in Multimedia Appendix 1 simply counts each extraction as one regardless of its length. Consequently, frequently occurring short terms (eg, "lung cancer") that are repeatedly extracted may appear to achieve a high score.

Comparing Table 2 and Table S1 in Multimedia Appendix 1 reveals that, for instruction tuning, all models except GPT-NeoX-3.6B attained higher F_1-scores in Table S1 in Multimedia Appendix 1.

In addition, we stratified the gold entities into 5 character-length bins (1 - 2, 3 - 6, 7 - 10, 11 - 20, and ≥21) and computed precision, recall, and F_1-score within each bin (Tables S19-S21 in Multimedia Appendix 1). Under the weighted soft matching score with p=1 (equivalent to character-level evaluation), the decoder model (Swallow-70B) achieved the highest F_1-score for the shortest spans (1 - 2 characters; F_1-score=0.814, SD 0.017), but its performance dropped rapidly as the span length increased (≥21 characters; F_1-score=0.677, SD 0.015). In contrast, the encoder model (JMedDeBERTa(s)) showed relatively low F_1-score for the extremely short spans (1 - 2 characters), which account for only 9% of the dataset, while consistently outperforming the decoder model for spans of 3 characters or longer and maintaining robust performance even for long spans (≥10 characters). A similar trend was also observed in the stratified analysis based on the marker matching score, which does not weight span length.

This suggests that instruction tuning is effective for extracting short terms but less so for long expressions. When using instruction tuning, both sides of the extraction segments must be enclosed with markers. For E-CCR, which includes long phrases, the attention mechanism must retain these extended, marker-enclosed segments. This requirement may account for the discrepancy with a previous study [28] that reported superior performance for instruction tuning compared with encoder models when extracting terms or short phrases (eg, disease names or gene identifiers). The fact that encoder models outperformed decoder models for spans of 3 characters or longer—which constitute the majority (91%) of entities in E-CCR—is also clinically meaningful.

A comparison of decoder models using token classification reveals higher F_1-scores in Table 2 than in Table S1 in Multimedia Appendix 1. This suggests that token classification remains effective for extracting long phrases. When using token classification, the model assigns a 0 or 1 to each token, and this token-level information can be used in subsequent outputs, potentially making it easier to handle long markers than instruction tuning.

When long expressions are extracted, unintended portions of the text may also be included, potentially reducing precision. However, in applications such as automated knowledge graph construction, such noise is not critical, as it can be filtered out during subsequent processing (eg, normalization). The key point is that the relatively long expressions that need to be extracted are reliably captured. Token classification achieved a higher recall than instruction tuning and is therefore well suited to applications where E-CCR is followed by a series of postprocessing steps, such as normalization of the extracted segments and causal inference among terms, as in automatic knowledge graph construction.

Seventeen models were compared using the weighted soft matching score, which accounts for penalties when the extracted segments are fragmented. In particular, when p=1, the evaluation is equivalent to that based on character-level segmentation. Across all p values, token classification consistently outperformed instruction tuning, and encoder models outperformed decoder models. The scores did not decrease substantially at higher p, suggesting that cases in which the marker positions become fragmented during extraction are rare.

Figure 6 shows an example of the output results (for the test set) of JMedDeBERTa(s), which exhibited the best performance. The output results include (1) cases in which the extraction of critical clinical concepts was performed correctly and (2) cases in which the extraction of critical clinical concepts failed. In particular, in (2), as shown in Figure 6, cases exist where the model does not extract the part that should be extracted; however, cases also exist where the output results are not necessarily wrong, as shown in Figures S3-S5 in Multimedia Appendix 1, and cases exist where the F_1-score decreases because of labeling-related problems. Therefore, the actual model performance may be underestimated and warrants further investigation in future research.

XSL•FO
RenderX

Figure 6. Example of perfect extraction by JMedDeBERTa(s) in the test set. The upper row shows the original Japanese case report text, and the lower row shows the corresponding English translation. Blue highlights indicate gold-standard clinically essential expressions, and red highlights indicate spans predicted by JMedDeBERTa(s) fine-tuned with token classification. In this example, the predicted spans exactly match the gold standard, resulting in a weighted soft matching F1 score of 1.0.

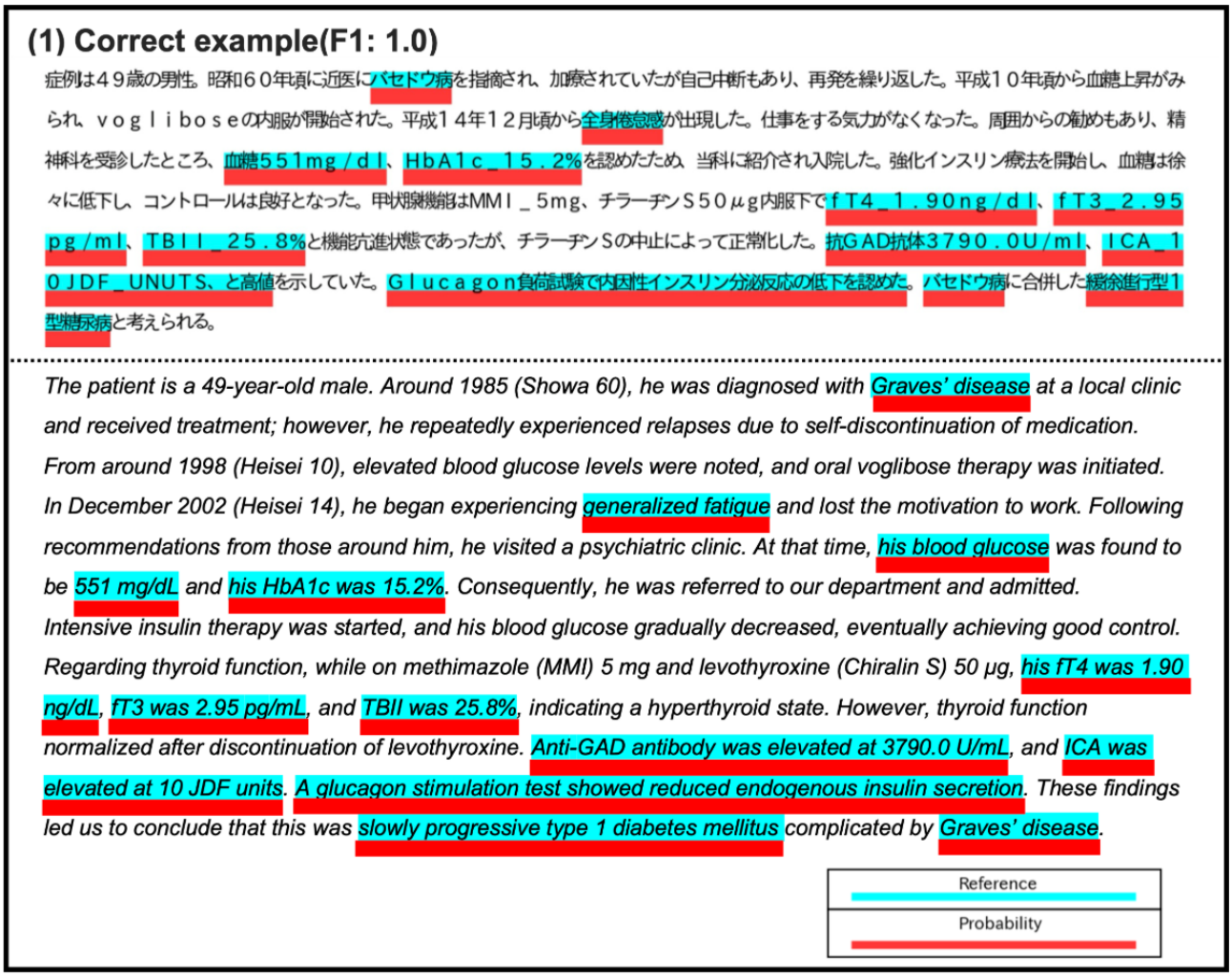

To gauge the potential impact of such labeling issues on performance evaluation, we randomly sampled 200 cases from the test set and reassessed the correspondence between the gold-standard labels and the original case report texts. As a result, we found that in 8 of the 200 cases (4%), at least one gold-standard label required correction; in most of these cases, the problem was attributable not to completely inappropriate concepts, but rather to omissions of important expressions that should have been included or to deviations in span boundaries. Thus, although the overall quality of the gold standard appears to be generally high, the actual model performance may be slightly underestimated, and a more systematic investigation of this issue will be needed in future work. These findings suggest that the measured F_1-score may be partially bounded by residual label noise and annotation subjectivity, potentially constituting a performance ceiling in agreement with the reference standard.

Novelty of This Study

This study is the first to compare instruction tuning and token classification and decoder and encoder models, in the context of the comprehensive E-CCR task.

In clinical concept extraction, capturing both long and conventional short expressions is essential. This study extended conventional NER, which typically targets words or short phrases, by proposing an E-CCR task that includes long expressions as extraction targets together with a weighted soft matching score. This extension represents a significant contribution of this work. To the best of our knowledge, few studies have focused on extracting long expressions to the extent achieved in this study. The study by Chen et al [15] addressed long expression extraction by summarizing discharge summaries through segment identification from the texts using AlphaBERT. Although Chen et al focused on extracting expressions representing single clinical concepts with AlphaBERT, their evaluation was limited to character-level analysis and did not incorporate metrics penalizing fragmented extraction. In contrast, the E-CCR framework explicitly includes long expressions as extraction targets. Chen et al.'s study was limited to comparative evaluations among AlphaBERT, BERT, BioBERT, and LSTM models, without incorporating analyses using the rapidly advancing decoder models. The inclusion of comparative evaluations using various decoder models alongside encoder models constitutes a significant advancement.

XSL•FO
RenderX

Future Directions

E-CCR can be applied not only to knowledge extraction for diagnostic support but also to other extraction tasks involving relatively long expressions. For example, consider a case in which information on daily living functions is extracted from a text and coded using the International Classification of Functioning, Disability and Health. When making a judgment of “嚥下機能低下(中等度)”<b5105.2> (“swallowing function decline (moderate)”<b5105.2>) in response to the expression “嚥下造影検査:反射惹起遅延(+)。液体で少量の誤嚥あり，ペースト食で少量の残留あり。咳反射あるが微弱” (“swallowing contrast study: reflex-induced delay (+). Small amount of aspiration with liquid, small amount of residual with paste food. Cough reflex present but weak”), the part corresponding to b5105.2 must be extracted, which is not a single word but a single description that spans multiple sentences. Additionally, numerous other applications can be envisioned, such as International Statistical Classification of Diseases and Related Health Problems coding of patient condition descriptions in clinical text, thereby providing valuable insights for extraction tasks across a wide range of medical texts.

Limitations

First, all experiments were conducted on Japanese clinical case reports. As Japanese differs substantially from English and other languages that dominate clinical natural language processing research—most notably in the lack of whitespace and the use of mixed scripts—our findings may not directly generalize to other languages or document types.

Second, for instruction-tuned decoder-based models, we evaluated only a single prompt template and a single instruction or output format, following prior work [23,28], in which extraction spans are generated as inline boundary tags using the @@ and ## markers. We did not perform a systematic prompt engineering study (eg, ablations over instruction wording, formatting, or response schemas). As LLM extraction performance is sensitive to both prompt design and output representation, alternative prompting strategies and constrained output formats—such as returning a structured list of extracted spans, structured JSON with character offsets, or other boundary representations—may lead to different results. We also did not explore reasoning-augmented prompting (eg, rationale or chain-of-thought prompting), including variants where reasoning is elicited but the final answer is constrained to a structured schema, which could affect generative extraction behavior. Thus, the observed performance gap between encoder and decoder models may partly reflect the difficulty of generating precise character-level boundary tags under this specific template or format. Future work should systematically compare multiple prompting or output strategies to disentangle the effects of model architecture from output format constraints.

Third, to maintain consistent label-token alignment across models, we disabled byte fallback. Consequently, out-of-vocabulary characters were normalized to <unk>, potentially discarding surface cues for rare kanji and specialized medical symbols. This may have disadvantaged models with smaller vocabularies, including GPT-NeoX-3.6B, DeBERTaV2-base, RoBERTa-base, and the JMedRoBERTa/JMedDeBERTa variants. Notably, however, all models used for token classification—both decoder only and encoder based—were evaluated under this same unfavorable setting (ie, with byte fallback disabled), yet the token classification still outperformed instruction tuning overall.

Fourth, owing to computational constraints, we applied parameter-efficient fine-tuning with LoRA to decoder models, whereas encoder models were fully fine-tuned. Although LoRA has been reported to achieve performance comparable to full fine-tuning [41], its effectiveness can vary depending on the task and hyperparameter settings; therefore, caution is warranted when interpreting these comparative results.

Fifth, for decoder-only models in token classification, we retained the default causal attention mask; therefore, token-level predictions could not leverage right-side context. As bidirectional information is often helpful for sequence labeling, this design may disadvantage causal-masked decoders relative to encoders. Approaches that enable bidirectional information in decoder-only LLM (eg, layer-wise causal mask removal) have been reported to improve sequence labeling performance [42,43]. Therefore, our conclusion that encoder-based token classification outperformed decoder-based approaches should be interpreted as holding for decoder-only models used with the default causal mask under our fine-tuning constraints and may not generalize to decoder variants explicitly adapted to incorporate right-context information.

Sixth, comparisons between general-domain and medical-domain pretraining are confounded by substantial differences in corpus size (362 GB vs 0.56 GB) and data composition. The strong performance of general-domain models—despite limited exposure to specialized medical terminology—may reflect not only syntactic knowledge but also other advantages conferred by massive training data. Rigorous ablation studies that control for architecture, corpus size, and training conditions are therefore needed. Further research is also required to determine whether the observed encoder-decoder performance gap, the relative benefits of domain-specific versus general-domain pretraining, and the effects of prompting strategies generalize to other languages, clinical settings, and medical document types.

Seventh, the weighted soft matching score is an automated evaluation metric whose clinical validity has not yet been validated, including its alignment with clinician judgments, and its usefulness for downstream tasks. Moreover, as the score is computed based on the span boundaries of the reference annotations, it may penalize predictions even when boundary differences are clinically acceptable. Future work should quantify correlations with clinician ratings and downstream task performance to establish the metric’s clinical validity.

Finally, our reference annotations represent an adjudicated consensus rather than an objective ground truth. As preadjudication annotations were not retained, we cannot report IAA to quantify the inherent subjectivity of the task. In a post hoc audit, 8 (4%) of 200 randomly sampled test cases contained at least one span requiring correction, mostly due to omissions or boundary deviations, indicating residual label noise and

XSL•FO
RenderX

subjective boundary decisions. Moreover, as illustrated by examples where model predictions are clinically plausible despite differing from the reference, span-level F_1-score should be interpreted as agreement with this consensus reference, and may be bounded by a ceiling on reference-based performance rather than model capability. This limitation primarily affects the absolute magnitude of the scores; however, as all models were evaluated against the same reference standard under the same metric, relative comparisons across models remain informative. Future work will retain independent preadjudication labels to compute IAA and will complement reference-based metrics with clinician-centered acceptability assessments and/or more tolerant scoring schemes.

Conclusions

This study is the first to investigate effective model selection for the E-CCR task—an extraction task targeting terms and extended expressions (eg, diseases, conditions, and clinical findings considered critical for differential diagnosis) that are indispensable for the automatic construction of causal relationship knowledge from case reports—by comparing encoder versus decoder models, as well as general-purpose versus domain-specific models. Additionally, we introduced a novel evaluation metric tailored to the E-CCR task, the weighted soft matching score. Furthermore, we analyzed model effectiveness in relation to the length of the extraction targets.

As a result, encoder-based token classification achieved the best performance in our evaluation. Under our evaluation setting—where decoder models were constrained to output entity spans using inline boundary tags (eg, @@ and ##)—token classification generally outperformed instruction tuning. Moreover, among the 17 models compared, we observed no case where a decoder model outperformed its corresponding encoder model. These findings suggest that, for the E-CCR task, token classification–based extraction with encoder models can achieve high accuracy with relatively fewer parameters, making it advantageous in resource-constrained environments. This efficiency is also beneficial for real-world deployment in clinical practice, where inference latency and throughput are critical. That said, although encoder token classification outperformed decoder approaches under our standard fine-tuning recipe, decoder performance may improve with alternative prompting strategies, the use of reasoning-oriented models (ie, increased inference-time computation), or architectural or training modifications such as layer-wise relaxation of the causal mask for instruction tokens.

In contrast, while JMedDeBERTa(s), pretrained on medical text, obtained the highest average score on the E-CCR task, DeBERTaV2-base achieved the highest average score on the MRNER disease benchmark. These results indicate that there is no single universally optimal model; the relative advantage of domain-specific versus general-domain pretraining likely depends on the characteristics of the target task and the dataset.

Our results examining the effect of domain-specific texts relative to general texts align with previous findings [33], indicating only marginal performance differences between models trained with and without domain-specific texts.

Furthermore, our findings indicate that models trained on relatively small amounts of domain-specific text can match or exceed the performance of models trained on large general corpora. Conversely, in situations where domain-specific text is scarce, models trained solely on general corpora may still deliver high performance. Therefore, model selection should be guided by the quantity and characteristics of the available training texts.

An analysis of the differences between instruction tuning and token classification with respect to the length of the extracted segments revealed that instruction tuning is well suited for extracting short terms but is less effective for extracting long expressions. In contrast, token classification proved effective for extracting extended phrases. Furthermore, evaluation using the weighted soft matching score showed that increasing the fragmentation penalty did not substantially degrade model performance. This suggests that marker positions split during extraction are infrequent, even when extracting long expressions.

A detailed analysis of the outputs from JMedDeBERTa(s) revealed inaccuracies in some manually assigned correct labels. This observation implies that actual model accuracy may surpass the reported figures.

Extracting clinical concepts necessitates targeting both conventional short expressions and extended phrases. To our knowledge, few prior studies have comprehensively addressed extended expression extraction in clinical texts. In this study, we performed model comparisons on a more comprehensive E-CCR task using the weighted matching score, obtaining insights into model selection, effectiveness with respect to expression length, and the impact of domain-specific texts. These findings constitute a substantial contribution to the advancement of clinical information extraction methodologies.

Funding

This work was partially supported by the Cross-ministerial Strategic Innovation Promotion Program on “Integrated Health Care System” (grant JPJ012425).

Data Availability

The datasets generated or analyzed in this study are not publicly available due to restrictions imposed by a data use agreement with the Japanese Society of Internal Medicine. However, subject to approval by the Japanese Society of Internal Medicine and the relevant affiliated institutions and/or regulatory authorities, the data may be made available from the corresponding author upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

YT contributed to reviewing and editing, writing the original draft, visualization, validation, software, methodology, and investigation. SK contributed to reviewing and editing and methodology. MI contributed to reviewing and editing and investigation. RN contributed to data curation, resources, funding acquisition, and supervision. TI contributed to conceptualization, data curation, resources, reviewing and editing, writing the original draft, methodology, investigation, funding acquisition, and supervision.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Extended experimental details and evaluation results that support the main findings. Contents include (1) overall and case-level performance of the 17 encoder or decoder models on the J-CaseMap and MRNER disease datasets under 4 scoring metrics (weighted soft matching, marker matching, character segment, and token segment scores), (2) pretraining and fine-tuning hyperparameters for JMedDeBERTa(s) or JMedDeBERTa(c) and all comparison models, (3) training and validation loss curves, (4) input-output agreement analysis for the instruction-tuned decoder, (5) gold entity length distribution and length-stratified performance, and (6) representative examples illustrating discrepancies between model predictions and reference annotations. [DOCX File, 10144 KB - jmir_v28i1e78681_app1.docx]

References

1. About “J-CaseMap: case search for difficult-to-diagnose cases” [Article in Japanese]. The Japanese Society of Internal Medicine. 2024. URL: https://www.naika.or.jp/j-casemap [accessed 2024-12-01]
2. Keraghel I, Morbieu S, Nadif M. Recent advances in named entity recognition: a comprehensive survey and comparative study. arXiv. Preprint posted online on Jan 19, 2024. [doi: 10.48550/arXiv.2401.10825]
3. Madan S, Lentzen M, Brandt J, Rueckert D, Hofmann-Apitius M, Fröhlich H. Transformer models in biomedicine. BMC Med Inform Decis Mak 2024 Jul 29;24(1):214. [doi: 10.1186/s12911-024-02600-5] [Medline: 39075407]
4. Cho HN, Jun TJ, Kim YH, et al. Task-specific transformer-based language models in health care: scoping review. JMIR Med Inform 2024 Nov 18;12(12):e49724. [doi: 10.2196/49724] [Medline: 39556827]
5. Luo X, Deng Z, Yang B, Luo MY. Pre-trained language models in medicine: a survey. Artif Intell Med 2024 Aug;154(154):102904. [doi: 10.1016/j.artmed.2024.102904]
6. Hu Z, Hou W, Liu X. Deep learning for named entity recognition: a survey. Neural Comput & Applic 2024 Jun;36(16):8995-9022. [doi: 10.1007/s00521-024-09646-6]
7. Yada S, Nakamura Y, Wakamiya S, et al. REAL-MedNL: overview of REAL document-based medical natural language processing task. 2022 Presented at: Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies; Jun 14-17, 2022. [doi: 10.20736/0002002295]
8. Kajiyama K, Horiguchi H, Okumura T, Morita M, Kano Y. De-identifying free text of Japanese electronic health records. J Biomed Semantics 2020 Sep 21;11(1):11. [doi: 10.1186/s13326-020-00227-9] [Medline: 32958039]
9. Nath N, Lee SH, Lee I. NEAR: named entity and attribute recognition of clinical concepts. J Biomed Inform 2022 Jun;130(130):104092. [doi: 10.1016/j.jbi.2022.104092] [Medline: 35533990]
10. Yada S, Joh A, Tanaka R, et al. Towards a versatile medical-annotation guideline feasible without heavy medical knowledge: starting from critical lung diseases. 2020 Presented at: Proceedings of the 12th Language Resources and Evaluation Conference; May 11-16, 2020 URL: https://aclanthology.org/2020.lrec-1.561/ [accessed 2026-04-28]
11. Ohno Y, Kato R, Ishikawa H, et al. Using the natural language processing system medical named entity recognition-Japanese to analyze pharmaceutical care records: natural language processing analysis. JMIR Form Res 2024 Jun 4;8(8):e55798. [doi: 10.2196/55798] [Medline: 38833694]
12. Kawazoe Y, Shibata D, Shinohara E, Aramaki E, Ohe K. A clinical specific BERT developed using a huge Japanese clinical text corpus. PLoS One 2021;16(11):e0259763. [doi: 10.1371/journal.pone.0259763] [Medline: 34752490]
13. Nishiyama T, Yamaguchi A, Han P, et al. Automated system to capture patient symptoms from multitype Japanese clinical texts: retrospective study. JMIR Med Inform 2024 Sep 24;12:e58977. [doi: 10.2196/58977] [Medline: 39316418]
14. Lin C. ROUGE: a package for automatic evaluation of summaries. : Association for Computational Linguistics; 2004 Presented at: Proceedings of the Text Summarization Branches Out Workshop; Jul 25-26, 2004; Barcelona, Spain p. 74-81 URL: https://aclanthology.org/W04-1013/ [accessed 2026-04-28]
15. Chen YP, Chen YY, Lin JJ, Huang CH, Lai F. Modified bidirectional encoder representations from transformers extractive summarization model for hospital information systems based on character-level tokens (AlphaBERT): development and performance evaluation. JMIR Med Inform 2020 Apr 29;8(4):e17787. [doi: 10.2196/17787] [Medline: 32347806]
16. Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. 2001. URL: https://repository.upenn.edu/handle/20.500.14332/6188 [accessed 2026-04-14]
17. Graves A. Supervised sequence labelling with recurrent neural networks. In: Studies in Computational Intelligence: Springer; 2012. URL: https://link.springer.com/10.1007/978-3-642-24797-2 [accessed 2026-04-14] [doi: 10.1007/978-3-642-24797-2]

XSL•FO
RenderX

18. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: Proceedings of the 2019 Conference of the North; Jun 2-7, 2019. [doi: 10.18653/v1/N19-1423]
19. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. 2017 Presented at: Proceedings of the 31st Conference on Neural Information Processing Systems; Dec 4-9, 2017 URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html [accessed 2026-04-28]
20. Xu D, Chen W, Peng W, et al. Large language models for generative information extraction: a survey. Front Comput Sci 2024 Dec;18(6):186357. [doi: 10.1007/s11704-024-40555-y]
21. Wang Z, Zhou Q, Junfeng Z, Wang Y, Ding H, Song J. A knowledge-enhanced medical named entity recognition method that integrates pre-trained language models. 2023 Presented at: 2023 IEEE International Conference on Medical Artificial Intelligence (MedAI); Apr 14-19, 2024. [doi: 10.1109/MedAI59581.2023.00046]
22. Ashok D, Lipton ZC. PromptNER: prompting for named entity recognition. arXiv. Preprint posted online on May 24, 2023. [doi: 10.48550/arXiv.2305.15444]
23. Wang S, Sun X, Li X, et al. GPT-NER: named entity recognition via large language models. Presented at: Findings of the Association for Computational Linguistics; Apr 29 to May 4, 2025. [doi: 10.18653/v1/2025.findings-naacl.239]
24. Hu Y, Chen Q, Du J, et al. Improving large language models for clinical named entity recognition via prompt engineering. J Am Med Inform Assoc 2024 Sep 1;31(9):1812-1820. [doi: 10.1093/jamia/ocad259] [Medline: 38281112]
25. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. 2020 Presented at: Proceedings of the 34th International Conference on Neural Information Processing Systems; Dec 6-12, 2020; Vancouver, BC, Canada (virtual p. 1877-1901 URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html [accessed 2026-04-28]
26. Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. arXiv. Preprint posted online on Mar 15, 2023. [doi: 10.48550/arXiv.2303.08774]
27. Sun X, Dong L, Li X, et al. Pushing the limits of chatgpt on NLP tasks. arXiv. Preprint posted online on Jun 16, 2023. [doi: 10.48550/arXiv.2306.09719]
28. Keloth VK, Hu Y, Xie Q, et al. Advancing entity recognition in biomedicine via instruction tuning of large language models. Bioinformatics 2024 Mar 29;40(4). [doi: 10.1093/bioinformatics/btae163]
29. Li Z, Li X, Liu Y, et al. Label supervised llama finetuning. arXiv. Preprint posted online on Oct 2, 2023. [doi: 10.48550/arXiv.2310.01208]
30. Yu Z, Yang X, Sweeting GL, et al. Identify diabetic retinopathy-related clinical concepts and their attributes using transformer-based natural language processing methods. BMC Med Inform Decis Mak 2022 Sep 27;22(Suppl 3):255. [doi: 10.1186/s12911-022-01996-2] [Medline: 36167551]
31. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682]
32. Lentzen M, Madan S, Lage-Rupprecht V, et al. Critical assessment of transformer-based AI models for German clinical notes. JAMIA Open 2022 Dec;5(4):ooac087. [doi: 10.1093/jamiaopen/ooac087] [Medline: 36380848]
33. García Subies G, Barbero Jiménez Á, Martínez Fernández P. A comparative analysis of Spanish Clinical encoder-based models on NER and classification tasks. J Am Med Inform Assoc 2024 Sep 1;31(9):2137-2146. [doi: 10.1093/jamia/ocae054] [Medline: 38489543]
34. He P, Liu X, Gao J, et al. DeBERTa: decoding-enhanced BERT with disentangled attention. 2021 Presented at: Proceedings of the International Conference on Learning Representations; May 3-7, 2021 URL: https://openreview.net/forum?id=XPZIaotutsD [accessed 2026-04-28]
35. Liu Y, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. Preprint posted online on Jul 26, 2019. [doi: 10.48550/arXiv.1907.11692]
36. Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. : OpenAI Blog; 2019 URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf [accessed 2026-04-28]
37. Black S, Biderman S, Hallahan E, et al. GPT-NeoX-20B: an open-source autoregressive language model. 2022 Presented at: Proceedings of BigScience Episode #5 -- Workshop on Challenges & Perspectives in Creating Large Language Models; May 22-27, 2022. [doi: 10.18653/v1/2022.bigscience-1.9]
38. Touvron H, Martin L, Stone K, et al. Llama 2: open foundation and fine-tuned chat models. arXiv. Preprint posted online on Jul 18, 2023. [doi: 10.48550/arXiv.2307.09288]
39. Jiang AQ, Sablayrolles A, Mensch A, et al. Mistral 7B. arXiv. Preprint posted online on Oct 10, 2023. [doi: 10.48550/arXiv.2310.06825]
40. Kudo T, Richardson J. SentencePiece: a simple and language independent subword tokenizer and detokenizer for neural text processing. 2018 Presented at: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; Oct 31 to Nov 4, 2018; Brussels, Belgium p. 66-71. [doi: 10.18653/v1/D18-2012]
41. Hu EJ, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models. arXiv. Preprint posted online on Jun 17, 2021. [doi: 10.48550/arXiv.2106.09685]

XSL•FO
RenderX

42. Dukić D, Snajder J. Looking right is sometimes right: investigating the capabilities of decoder-only LLMs for sequence labeling. Presented at: Findings of the Association for Computational Linguistics ACL 2024; Aug 11-16, 2024. [doi: 10.18653/v1/2024.findings-acl.843]
43. Kukić ML, Čuljak M, Dukić D, Tutek M, Šnajder J. Sequence repetition enhances token embeddings and improves sequence labeling with decoder-only language models. 2026 Presented at: Findings of the Association for Computational Linguistics; Mar 24-29, 2026. [doi: 10.18653/v1/2026.findings-eacl.339]
44. Jiang J, Huang J, Aizawa A. JMedBench: a benchmark for evaluating Japanese biomedical large language models. 2025 Presented at: Proceedings of the 31st International Conference on Computational Linguistics; Jan 19-24, 2025 URL: https://aclanthology.org/2025.coling-main.395/ [accessed 2026-04-28]
45. Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. arXiv. Preprint posted online on Jan 23, 2020. [doi: 10.48550/arXiv.2001.08361]
46. Nazi ZA, Peng W. Large language models in healthcare and medical domain: a review. Informatics (MDPI) 2024;11(3):57. [doi: 10.3390/informatics11030057]
47. Jeong Y, Kim E. SciDeBERTa: learning DeBERTa for science technology documents and fine-tuning information extraction tasks. IEEE Access 2022;10(10):60805-60813. [doi: 10.1109/ACCESS.2022.3180830]
48. Luo Z, Xi Y, Ma J, et al. DecBERT: enhancing the language understanding of BERT with causal attention masks. : Association for Computational Linguistics; 2022 Presented at: Findings of the Association for Computational Linguistics; Jul 10-15, 2022; Seattle, United States p. 1185-1197. [doi: 10.18653/v1/2022.findings-naacl.89]
49. Kim Y, Kim JH, Lee JM, et al. A pre-trained BERT for Korean medical natural language processing. Sci Rep 2022;12(1). [doi: 10.1038/s41598-022-17806-8]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
E-CCR: extended clinical concept recognition
IAA: interannotator agreement
LLM: large language model
LoRA: low-rank adaptation
LSTM: long short-term memory
MLM: masked language modeling
MRNER disease: Medical Report Named Entity Recognition for positive disease
NER: Named Entity Recognition

Edited by A Coristine; submitted 13.Jun.2025; peer-reviewed by C Ma, M Torii; revised version received 01.Mar.2026; accepted 17.Mar.2026; published 14.May.2026.

Please cite as:
Tsukiji Y, Kataoka S, Itokazu M, Nagai R, Imai T
Evaluating Encoder and Decoder Models for Extended Clinical Concept Recognition in Japanese Clinical Texts: Comparative Study With Weighted Soft Matching
J Med Internet Res 2026;28:e78681
URL: https://www.jmir.org/2026/1/e78681
doi:10.2196/78681

© Yuya Tsukiji, Satoshi Kataoka, Masafumi Itokazu, Ryozo Nagai, Takeshi Imai. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Human-AI Interaction in Kidney Transplant Decision Support Systems: Qualitative Study of Patient and Support Person Expectations

Zeineb Sassi[1,2], MSc; Sascha Eickmann[3], PhD; Roland Roller[4], PhD; Bilgin Osmanodja[5], MD; Jakob Joachim Spencker[6], MA, MD; Ömer Ege Ömeroğlu[5], MD; Aljoscha Burchardt[4], PhD; Michael Hahn[7], MTh; Peter Dabrock[7], Prof Dr; Sebastian Möller[4,8], Prof Dr; Klemens Budde[5], Prof Dr; Anne Herrmann[1,9,10], Prof Dr

[1]Department of Epidemiology and Preventive Medicine, Medical Sociology, Faculty of Medicine, University of Regensburg, Dr. Gessler-Straße 17, Regensburg, Germany

[10]Bavarian Centre for Cancer Research (BZKF), Regensburg, Germany

[2]Central Clinical School, University of Sydney, Sydney, NSW, Australia

[3]Institute of General Practice, University Hospital Regensburg, Regensburg, Bavaria, Germany

[4]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany

[5]Department of Nephrology and Medical Intensive Care, Charité-Universitätsmedizin Berlin, Corporate Member of Free University of Berlin, Berlin Institute of Health, Berlin, Germany

[6]Department of Nephrology and Medical Intensive Care, Charité-Universitätsmedizin Berlin, Corporate Member of Free University of Berlin, Berlin Institute of Health, Humboldt-University of Berlin, Berlin, Germany

[7]Chair for Systematic Theology II (Ethics), Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Nuremberg, Germany

[8]Quality and Usability Lab, Technical University of Berlin, Berlin, Germany

[9]Department of Internal Medicine III, University Hospital Regensburg, Regensburg, Germany

Corresponding Author:
Zeineb Sassi, MSc
Department of Epidemiology and Preventive Medicine, Medical Sociology, Faculty of Medicine, University of Regensburg, Dr. Gessler-Straße 17, Regensburg, Germany

Abstract

Background: Artificial intelligence (AI) is increasingly applied in medicine, including clinical decision-making. AI-based decision support systems (DSS) can enhance early risk detection and treatment optimization. However, the perspectives of patients and their support persons on AI-assisted DSS in clinical care, particularly regarding shared decision-making (SDM), remain underexplored.

Objective: This study investigates the expectations, informational needs, and perceptions of patients who underwent kidney transplantation and their support persons regarding AI-assisted DSS and its influence on SDM in posttransplant care.

Methods: In a longitudinal qualitative study, 36 semistructured interviews were conducted with patients who underwent kidney transplantation and their support persons at a German kidney transplant center. Participants were asked about their views on AI's role in follow-up care, its impact on communication, trust, and decision-making, as well as their informational needs regarding AI-assisted DSS. Interviews were transcribed, pseudonymized, and analyzed using framework analysis.

Results: Participants recognized AI's potential to support clinicians by identifying risks of transplant loss, rejection, and infection, and by providing data-driven treatment recommendations. However, they emphasized that final decisions should remain with physicians. A majority of participants (n=28, 78%) expressed concern that AI might depersonalize care and diminish physician-patient communication due to a lack of "human touch." Participants demonstrated limited understanding of AI-based DSS functionality and highlighted the need for simple, accessible educational materials (eg, leaflets) explaining AI operations. While most doubted AI could replicate human empathy, some acknowledged that AI might be perceived as more attentive than time-pressured physicians, offering consistent monitoring and support. Participants consistently stressed that AI should augment, not replace, clinical decision-making.

Conclusions: Patients who underwent kidney transplantation and support persons endorse the integration of AI in follow-up care when it enhances clinical decision-making without supplanting the physician's role. Acceptance and trust depend on transparency, accountability, and preserving the "human touch" in care. The development of educational tools to communicate AI functions and limitations is crucial to empower patients and support persons in SDM processes and to ensure AI complements, rather than undermines, patient-centered care.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e83195) doi:10.2196/83195

KEYWORDS

artificial intelligence; AI; decision-making; medical informatics; decision support system; DSS; health care; human-computer interaction

Introduction

Impact of AI on Medical Decision Support in Nephrology

Artificial intelligence (AI) has the potential to transform health care by improving diagnostic accuracy, supporting medical decision-making, and enhancing patient care [1]. In various medical fields such as radiology, oncology, and cardiology [2], the successful use of AI has supported early detection, personalized treatment planning, and risk prediction, providing a comparative context for its emerging role in nephrology, where AI-based decision support systems (DSS) are growing as valuable tools to assist health care professionals. In the context of posttransplant care, particularly for patients who have undergone kidney transplantation, these systems may predict outcomes such as graft survival, detect patients at risk for rejection, and optimize medication regimens. They can improve patient management by identifying trends and patterns in patient data that may not be easily recognized by human physicians [3]. Posttransplant care involves complex, long-term decision-making processes that require close monitoring and individualized treatment adjustments. For example, managing immunosuppressive therapy demands high levels of patient-physician interaction, making kidney transplant recipients particularly sensitive to changes introduced by AI-based DSS [1-4]. By analyzing extensive patient data, AI systems provide insights—such as predictive modeling for graft survival—that complement traditional clinical assessments [5]. While AI offers significant potential to enhance posttransplant care, integrating these technologies into clinical practice presents several challenges. Effective AI systems depend on high-quality, standardized data. However, health care data often originates from diverse sources with varying formats and standards, leading to interoperability and integration challenges [3,6,7]. Such a lack of standardization impedes AI model training and deployment, potentially compromising their accuracy and reliability [8]. Moy et al's [9] scoping review highlights that patients' needs and expectations are often overlooked in the development and application of AI technologies, underscoring the critical need to incorporate patient and end-user perspectives in health care AI design [10]. Patients are rarely involved in the design of AI systems and often have a limited understanding of how these technologies work, including their underlying data and limitations. This lack of knowledge may foster uncertainty and unaddressed concerns, which in turn can undermine trust [9,10]. To counteract this, it is essential to identify and integrate the expectations and needs of patients and their support persons, as their preferences directly shape the acceptance and effectiveness of AI technologies in health care.

Shared Decision-Making: The Importance of Understanding the Preferences and Concerns of Patients in Medical Decisions

Shared decision-making (SDM) refers to a collaborative process in which patients and physicians jointly make treatment decisions, with patient values and preferences serving as a central guiding principle [11]. In the context of AI, SDM gains particular importance, since patients often hold complex and ambivalent views on the role of AI in their care [12]. In post–kidney transplant follow-up care, patient engagement in SDM is vital, as treatment decisions often involve balancing complex risks, such as graft rejection [13,14]. Patients fear that AI may reduce the "human touch" of health care, leading to depersonalized treatment [15,16]. They also express worries about data privacy and security [17,18], potential diagnostic errors or misjudgments in rapidly changing health conditions [18-20], and a lack of transparency in algorithmic decision-making [20,21]. Such concerns highlight the need to understand patient perspectives and integrate them into SDM, as acceptance of AI-based decision support depends on patients' trust and willingness to engage with these technologies.

While prior literature has addressed general patient attitudes toward AI in health care, empirical evidence on the expectations of patients and support persons regarding AI-based SDM in nephrology remains lacking. Most studies focus on physician perspectives, leaving the voices of patients and support persons under-represented [9]. In Germany, there is no qualitative evidence on experiences with AI-based care in high-stakes contexts such as kidney transplantation. This study addresses this gap by providing first insights into expectations and needs related to AI-based DSS within a German transplant center.

Support Persons' Role in AI-Assisted SDM Matters

Support persons, such as family members, are key sources of patients' health information and play vital roles in diagnosis, treatment, and recovery [22]. Previous research indicates that surrogate decision-makers, such as family members, make approximately 75% of the decisions for hospitalized patients with life-threatening illnesses [23]. Despite the importance of support persons in patient care, their perspectives—particularly in posttransplant contexts—remain largely underexplored [24,25]. Support persons also face unique concerns when participating in decision-making throughout a patient's health care journey, with these concerns being intensified by the introduction of AI. Navigating complex medical AI systems adds both emotional and practical challenges, as support persons must assist patients while also understanding and interpreting AI-generated data and recommendations [4]. One prominent concern for support persons is information overload, where frequent AI alerts and large volumes of data can create cognitive stress rather than provide support. Support persons also worry that AI may weaken the emotional connection between patients

XSL•FO
RenderX

and their treating physician, potentially leading to feelings of isolation [26,27]. Additionally, support persons fear that over-reliance on AI may reduce personalized care in medical decision-making [28], and they may feel uncertain about endorsing treatment decisions if they do not understand how AI functions or how it influences patient care [29].

In all, these emotional and practical concerns illustrate that support persons are not only critical to patient care but also pivotal stakeholders in AI-based SDM. Understanding and integrating their perspectives and expectations are, therefore, essential to ensure that AI systems support both patients and their support persons effectively.

Patient-Centered Care Through Studying the Perspectives of Patients and Support Persons on AI-Based Decision-Making

AI-based DSS are increasingly applied in high-stakes medical contexts, including posttransplant care, where they can support complex clinical decisions. The effectiveness of these systems, however, depends not only on technical performance but also on their alignment with patient-centered care principles, particularly SDM, which integrates patient preferences, values, and the perspectives of support persons [11,23,30].

Patients and support persons bring their own expectations and concerns to medical decision-making, which strongly influence the acceptance of AI. Previous research indicates that mistrust in medical AI is common and often stems from limited knowledge of its functioning, personal experiences, and concerns about data privacy, potential errors, and the depersonalization of care [31]. Support persons, who frequently act as surrogate decision-makers and key sources of health information for patients [32], face additional challenges, including the cognitive burden of AI-generated information, uncertainty about AI's impact on patient outcomes, and anxiety that AI may reduce personalized care [4,24,26-29,32]. Emotions such as anxiety play a decisive role in shaping support persons' engagement with AI-based technologies, influencing whether they accept or reject these tools [32].

Despite these insights, qualitative research examining how AI-based DSS affects SDM in nephrology remains scarce. Existing studies often focus on general patient attitudes or physician perspectives, leaving the voices of patients and support persons under-represented. This study addresses this gap by exploring the expectations and needs of patients and support persons regarding AI-based DSS in posttransplant care. By capturing these perspectives, the research aims to inform the development and implementation of AI technologies that support SDM, enhance patient autonomy, foster trust, and align with patient-centered care principles, such as transparency and accountability.

Methods

Study Design

This study is part of a 2-year longitudinal qualitative interview study conducted as part of an interventional study in a German kidney transplant center (KTC). The protocols for both the interventional study as well as the nested interview study were published before data collection, and the study has been registered (NCT06056518) [13,14]. Semistructured interviews with patients who underwent kidney transplantation and their support persons were conducted at baseline and are reported here. The AI-based DSS applied in this study—a predictive machine learning model—assesses the risk of graft loss within the upcoming 360 days for patients who underwent kidney transplantation [28,33]. The system was rigorously developed and pretested before deployment and is monitored by a team of technicians supporting the technical subproject of this study [28,34].

Sample and Recruitment

Eligible patients were identified by study physicians after applying all inclusion and exclusion criteria to the entire transplant cohort of the KTC. Ongoing analysis revealed thematic redundancy in later interviews, indicating sufficient thematic depth and suggesting that data saturation had been reached.

Inclusion and Exclusion Criteria

Patients were eligible if they were 18 years or older, had undergone kidney transplantation, and had a functioning kidney graft with an estimated glomerular filtration rate of less than 30 mL/min/1.73 m^2. They were scheduled for routine follow-up at the participating KTC and provided written informed consent. Patients and their support persons were able to communicate in German. Support persons were eligible if they were 18 years or older and could provide informed consent.

Data Collection

Consenting patients and support persons were contacted by a researcher to arrange an interview appointment. Participants could choose the interview mode according to their preferences. All participants indicated that they preferred to be interviewed via telephone. Interviews were scheduled according to the availability of the patients and support persons to reduce research-related burden on participants. Through semistructured interviews, participants were encouraged to express their views on how AI-based DSS may impact physician-patient–support person communication and the decision-making process in the way they preferred. A narrative approach was used to elicit the variety and interplay of potential factors related to physician-patient communication in this area. This open approach naturally led to a relatively generic interpretation and assessment of AI. At the end of the interview, participants were given the opportunity to provide additional comments or ask questions. The researchers received intensive training from an interdisciplinary expert team in conducting these interviews to reduce bias in question framing, administration, and interpretation. Additionally, standardized interview protocols were used to document each interview.

Interviews

The interview guide was developed based on a literature review [35,36] and discussions among the interdisciplinary research team, which included experts in medicine, communication and behavioral science, health services research, ethics, and

XSL•FO
RenderX

computer science. Participants were asked about their expectations regarding the use of AI-based DSS, particularly in the context of SDM. Topics included patients' communication preferences, perceptions of trust, transparency, and responsibility in relation to AI-based DSS, as well as anticipated benefits, risks, and barriers to implementing this tool in routine care [14]. The interviews were conducted by a researcher with a background in psychology and health services research, along with prior academic engagement with topics related to AI in health care, which informed a sensitized but critically reflective stance toward AI throughout data collection and analysis. This disciplinary perspective may have influenced the framing of questions and participants' reflections on digital tools. Regular reflexive discussions within the interdisciplinary research team were used to critically examine potential assumptions during data collection.

Data Analysis

Interviews were transcribed verbatim. Two trained researchers reviewed each transcript for accuracy by comparing it with the original audio recordings, correcting any transcription errors, and ensuring completeness.

The data were then analyzed using framework analysis, which allows for both inductive and deductive coding [37]. Initially, 2 researchers conducted independent open coding, identifying concepts and patterns emerging directly from the data (inductive coding). Based on these initial insights and the study's guiding research questions, a framework of analytical categories was developed and applied to the data (deductive coding) [37]. The deductive coding framework was informed by established SDM [11,30,38,39] and human-computer interaction models [40], which guided the development of analytical categories related to trust, transparency, responsibility, and collaborative deliberation. In addition, principles from human-centered AI [39] and the European Commission's Ethics Guidelines for Trustworthy AI [41] informed categories concerning explicability, accountability, and stakeholder engagement.

To enhance the credibility and consistency of the coding process, 2 researchers discussed discrepancies and reached a consensus through iterative refinement of the coding framework. The qualitative data management software, ATLAS.ti (ATLAS.ti GmbH), was used to organize, code, and retrieve relevant data segments efficiently. The results of the analysis were discussed in close collaboration with the entire interdisciplinary research team.

Ethical Considerations

All patients were provided with written informed consent for the study intervention as well as the interview study. The support person accompanying the patient to the appointment was also provided with written informed consent [14]. The ethics committee of Charité - Universitätsmedizin Berlin approved this study (EA1/177/23; date of approval: September 19, 2023). Data were pseudonymized, and the confidentiality and privacy of participants were strictly maintained in accordance with institutional and national ethical standards and the Declaration of Helsinki. Each participant received financial compensation of €50 (approximately US $58.60).

Results

Overview

A total of 36 participants were included in the final analysis. Of these, 22 (61%) were males and 14 (39%) were females. Interviews were conducted together with a support person in 21 (58%) cases; among those, 9 (43%) support persons were identified as the participant's spouse. Fifteen (42%) patients stated that they preferred not to involve a support person in the call. Participants' ages ranged from 37 to 74 years (mean 57, SD 11 y). Sixteen participants resided in Berlin, whereas 20 lived elsewhere in Germany (Table 1). The duration of the interviews varied, with an average length of 34 (SD 14) minutes. The median interview length was 32 (IQR 28-42; range 20-90) minutes. Main themes derived from the Interview data are summarized in Table 2. Patients' sociodemographic and transplantation-related characteristics are presented in Table 1 below.

XSL•FO
RenderX

Table . Patients' sociodemographic and transplantation-related characteristics (N=36).

Characteristics	Values
Age (y), mean (SD; range)	57 (11; 37 - 74)
Gender, n (%)	
Male	22 (61)
Female	14 (39)
Involvement of support person, n (%)	
Yes	21 (58)
Spouse	9 (43)
Other	12 (57)
No	15 (42)
Time since kidney transplantation (y), n (%)	
<1	0 (0)
1 - 5	4 (11)
5 - 10	6 (17)
>10	26 (72)
Place of residence, n (%)	
Berlin	16 (44)
Germany, not Berlin	20 (56)

Patients' Perceived Benefits and Expectations Regarding the Use of AI-Based DSS

Patients expressed mixed feelings when asked what they generally associate with AI in health care. Some (15/36, 42%) were open and optimistic, viewing AI as a helpful tool that could support them and their physicians by providing additional insights for individualized treatment and improved SDM. A commonly mentioned benefit was AI's ability to process vast amounts of data quickly and potentially identify treatment options or risks that a human physician might overlook. As 1 participant explained:

> *It can be a help, like if it reminds the doctor that a certain value is off and that medication B shouldn't be given. That's a help. But in the end, if it still gets prescribed, well, then it's a human error.* [Male, 57 years]

Patients also recognized the potential of AI to enhance communication between health care providers by transferring patient data, which could enable more coordinated and informed care.

Another participant emphasized AI's potential to broaden clinical perspectives, stating that it may "show people possibilities that might not have come to their minds at that moment" (male, 55 years).

Importantly, participants emphasized that they expect AI not to make decisions independently.

"Yes, together with the doctor, yes—on its own, no" (male, 71 years), summarized a patient when asked about AI making medical decisions regarding his treatment. Another participant commented, "Definitely to be part of the decision-making—that's exactly the trend that's developing right now" (male, 63 years), underscoring the growing desire for collaborative approaches in which AI supports, but, according to most participants, does not replace human judgment and patient involvement.

While many acknowledge that AI can provide rapid and comprehensive analysis, potentially identifying patterns in laboratory results or drawing insights from large datasets, participants emphasized that AI should only serve as a supportive tool in the decision-making process. As 1 patient with her support person put it:

> *Ultimately, the doctor should still be the one in charge of my treatment decision. The AI should just support them. That's my view.* [Female, 50 years]

Another participant praised AI's analytical capabilities but highlighted the importance of human oversight:

> *The AI has much faster access to a lot more data than my treating physician. [...] It might raise a red flag when something is off, but then the doctor looks at it and says, "Okay, what's going on here?" [...] And then we go back to the doctor-patient conversation, trust, and competence. That's [the] key.* [Male, 65 years]

These expectations closely align with established models of SDM: trust and transparency are core prerequisites for meaningful patient participation in SDM [11], particularly in complex and high-stakes contexts such as kidney transplantation. Participants' insistence that AI should support rather than replace physicians reflects a preference for AI as an enabling tool within SDM, rather than as an autonomous decision-maker.

XSL•FO
RenderX

Interpreting these findings through an SDM lens highlights that the value of AI-based DSS lies in strengthening, rather than disrupting, relational and communicative aspects of care.

Patients' and Support Persons' Fears and Concerns Regarding AI-Based DSS

Most patients (26/36, 78%) expressed concerns about AI's inability to replicate the human qualities of empathy and individualized care—care that meets their unique needs, preferences, and values to understand their decision preferences. They feared that AI-based care, including DSS, might create distance between them and their treating physicians, who, according to the patients, are better equipped to understand their concerns, unique circumstances, medical history, and preferences, thereby avoiding a one-size-fits-all approach. They emphasized that while AI could be helpful, physicians should not rely on "machines" and thus should not leave the decision solely to the AI-based DSS. The majority of patients and their support persons (30/36, 83%) emphasized that their trust in final decisions regarding their care should always rest with their treating physicians. Physicians should never rely too heavily on automated systems, as 1 patient noted:

> *[...]It should never end up in a situation where the doctor no longer thinks and it becomes an automatic process, and if we were to be very dystopian, that eventually no doctor would be there at all, but rather we would feed data to the AI, which then says, "Do this and that," and we would only execute it [...] So, it [the DSS] should never replace a doctor!* [Male, 57 years]

In addition to that, patients and their support persons also identified several weaknesses and potential fears in relation to AI-based DSS. The most common concerns were related to the lack of "human touch" (29/36, 80%), referring to the absence of empathy, emotional understanding, and personalized communication that physicians offer while navigating complex treatment decisions:

> *Because human interaction still needs to be there [...] People don't want to only talk to a computer. They also want to have the human element.* [Male, 48 years]

Participants also expressed concerns about the potential for AI to make errors or miss important contextual information about a patient's condition. Many patients feared that over-reliance on AI could lead to a loss of personalized care (31/36, 78%) and treatment decisions, where their unique needs and circumstances might be overlooked.

Impact of Patients' and Support Persons' Knowledge Gaps on AI-Based DSS

Notably, many patients and support persons expressed a limited understanding of how AI-based DSS functions within medical settings (31/36, 86%), which often led to skepticism and uncertainty. Limited AI literacy emerged as one of the core findings in this study. A poor understanding of how AI-based DSS functions directly shaped participants' levels of trust, acceptance, and their willingness to engage in AI-assisted SDM. In addition to that, uncertainty about data sources, decision logic, and error handling amplified concerns about over-reliance on AI and loss of control. These findings demonstrate that improving AI literacy is not merely an educational "nice-to-have," but a prerequisite for meaningful and equitable AI-assisted SDM.

Another often-recurring theme was the concern about data privacy and the potential misuse of sensitive health information. Several participants voiced fears about who has access to their medical data and how it is processed, particularly if AI systems are used without sufficient transparency or oversight.

One participant reflected on this concern, saying, "If data about your body, your health, your care is stored and evaluated, it's fine—as long as it's not misused, and that's my fear when I say this" (female, 72 years).

Another patient noted:

> *I hope [AI] turns out to be positive. But it could also go the other way. After everything we've experienced in recent years, I'm not sure whether I could really put my trust in it. [...] Ultimately, the final decision should always be ours [Patient, SP and physician].*
>
> *[Female, 38 years]*

This quote captures the cautious optimism expressed by many—acknowledging potential benefits while remaining wary of unintended consequences or over-reliance on technology.

Moreover, patients consistently emphasize the importance of transparent and accessible information about how AI is used in their posttransplant care. They expressed a desire to better understand the technology's role, benefits, and risks to feel more secure and involved in decision-making processes. Transparency, for many, was a key factor in building trust in AI-assisted SDM.

Overall, while patients recognized AI's potential to contribute valuable insights, many stressed that only a physician can contextualize these insights and make appropriate, individualized decisions through direct conversation and shared understanding. According to most of the participants, trust in AI, therefore, depends not just on its accuracy but also on how clearly its role is communicated and how responsibly it is integrated into the decision-making process.

Perceived Impact of AI-Based DSS on SDM

Across interviews, participants expressed a strong desire to actively participate in the decision-making process between physicians and the AI-based DSS. Rather than seeing AI as a substitute for the physician or the patient's perspective, many envisioned a triadic model of decision-making involving the patient and the support person as one unit, the physician, and the AI-based DSS—not as a third independent decision-maker, but as a supportive participant. This approach was perceived as enriching the process, increasing trust, and reinforcing the role of patients in their care.

One patient articulated this ideal of SDM powerfully:

> *[...] But if the results are reviewed together and I, as the patient with my partner, can also help decide—then there are 3 things that can be brought*

XSL•FO
RenderX

> *into it: the technology, the doctor, and myself as the patient.* [Male, 57 years]

Across multiple interviews, patients and their support persons consistently emphasized the importance of transparency in AI-based decision-making. They expressed a desire to understand not only how the AI system itself generates recommendations, but also how these recommendations are integrated into their physician's decision-making process. For instance, one support person described an ideal scenario in which AI would become part of a joint consultation, allowing both the patient and the physician to review and discuss the AI's input together:

> *We [the patient and I] would find it good if we could look at the AI together with the doctor, for example, to see what data it has about him. If I had the chance to look at the computer, too. I would like that—that both the doctor and I see the same thing. So really a kind of teamwork with the AI, the doctor, and us.* [Male, 63 years]

At the same time, not all participants felt sufficiently included in current decision-making processes. Some reported a lack of involvement and insufficient dialogue about treatment choices—particularly in relation to potential AI-based recommendations:

> *If I were included, I would want to be asked: Did you understand this? Is it okay for you? But I don't think I will be asked that.* [Female, 38 years]

In summary, AI was seen as having the potential to enhance SDM, especially when its use facilitates dialogue, offers additional perspectives, or helps physicians explain and justify their decisions. However, participants made it clear that realizing this potential depends on the physician's communication style and the extent to which the patient has been introduced to and informed about the AI and whether the patient is treated as an equal partner in the decision-making process.

It is important to note that participants' perceptions varied depending on the type of AI system envisioned. References to predictive models were often associated with background data processing and risk estimation, whereas conversational or interactive AI was more frequently linked to concerns about authority, communication, and the involvement of patients.

The main themes and subthemes derived from the interview analysis are summarized in Table 2

XSL•FO
RenderX

Table . Main themes derived from interview data (N=36).

Theme	Values, n (%)
Previous experience with AI[a]-based medical devices	
Yes	21 (58)
Male	19 (90)
Female	2 (10)
No	15 (42)
General attitude toward the inclusion of AI in health care	
Rather optimistic	15 (42)
Rather pessimistic	5 (14)
Mixed	16 (44)
Trust in AI-based DSS[b] regarding final treatment decisions	
More trust in AI-based DSS	0 (0)
More trust in treating physician	30 (83)
Mixed	6 (17)
Concerns expressed with the integration of AI in the treatment	
Perceived personal lack of knowledge about the AI-assisted DSS	31 (86)
Perceived lack of "human touch" (eg, empathy) of treatment	29 (80)
Fear of less personalized treatment	28 (78)
Other	16 (44)
Perceived impact of AI on SDM[c]	
Desire to include AI-assisted DSS in SDM	22 (62)
Desire for more transparency on how the AI-based DSS generates treatment recommendations	31 (86)
No perceived impact	0

[a]AI: artificial intelligence.

[b]DSS: decision support system.

[c]SDM: shared decision-making.

Discussion

Principal Findings

This qualitative study provides insights into how patients who have undergone kidney transplants and their support persons perceive AI-based DSS in their posttransplant care, especially in the context of SDM. While participants largely welcomed AI as a supportive tool, they consistently emphasized the irreplaceable role of human judgment and the fear of the loss of the "human touch," meaning emotional connection and individualized care. These concerns appeared to have different origins. Some were grounded in participants' prior experiences with AI technologies, such as impersonal digital systems or fragmented communication, which informed fears of further depersonalization. Other concerns reflected more general uncertainty toward AI as an emerging and poorly understood technology, particularly in relation to accountability and error potential. Distinguishing between experiential and abstract sources of concern helps to explain the breadth of apprehension expressed and highlights different points of intervention. The modality-specific perceptions were reflected in the empirical material, where participants associated predictive systems with passive support, while interactive AI was more often perceived as actively shaping communication and decision-making processes.

These findings point to a critical challenge for future implementation: how to meaningfully integrate AI into the clinical encounter without eroding the essential elements of SDM: Charles et al [42] define SDM as an interactive process requiring at least 2 active participants, typically the physician and the patient, each contributing to the decision with their expertise and values. As previously mentioned, in our findings, patients envisioned a triadic model involving the patient, the support person, and the treating physicians, in which AI serves not as a third independent decision-maker but as a supportive participant, providing additional information without replacing human actors. This aligns with Elwyn's [11] model of SDM, which underscores the importance of team talk, option talk, and decision talk, emphasizing collaborative deliberation based on trust and clarity. However, our participants made it clear that this process depends heavily on whether AI can be transparently

XSL•FO
RenderX

explained and meaningfully integrated into conversations, without diminishing their own agency or emotional safety.

Patient- and Human-Centered AI: Empowering Patients Through Co-Design

Many patients lack an understanding of how AI-based DSS work, revealing a clear need for patient-focused approaches to AI development. Patient-centered and human-centered AI aim to ensure that these technologies are designed not merely for efficiency, but to align with patients' values, expectations, and experiences. Co-design is crucial in this context: involving patients as partners in the development process helps ensure that AI systems are understandable, transparent, and sensitive to patients' concerns [39]. International efforts, such as the European Commission's Ethics Guidelines for Trustworthy AI, stress that AI must be explicable, human-centered, and designed with stakeholder engagement, foremost including patients. In addition, there is a growing recognition that explainable AI must be designed not only for regulatory transparency but also for patient comprehension [43]. While initiatives such as the SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials—Artificial Intelligence) guidelines [44] or the CONSORT-AI (Consolidated Standards of Reporting Trials—Artificial Intelligence) [45] extensions have focused on professional audiences, future efforts must expand to systematically include patients in shaping how AI fits into care processes and SDM. Without active efforts to co-design and involve patients in development and to train clinicians in communicating about AI, there is a risk of eroding trust or creating asymmetries in SDM. Thus, the findings reinforce that effective integration of AI into SDM is not only a technical challenge but also a collaborative and communicative one. For AI to genuinely promote SDM in transplant care, it must be introduced in a way that strengthens, rather than displaces, the patient–support person–physician relationship.

Although participants referred to "AI" in AI-based DSS as a general concept, it is important to note that contemporary medical AI applications encompass a wide range of modalities with distinct implications for trust, transparency, and decision-making. These may include predictive machine learning used for risk stratification or outcome prediction, as well as conversational agents or generative AI applications that directly interact with end users. Such modalities differ substantially in their visibility, interaction style, explainability, and user design. While "background" predictive AI models may be experienced as less intrusive, interactive AI systems may be perceived as more present in the medical encounter and thus more directly affect the patient-physician relationship [46]. Interpreting participants' concerns and expectations through this lens helps to contextualize why issues such as the preservation of the "human touch," accountability, and transparency were particularly noticeable in our findings.

The "Human Touch" in the Age of AI: Empathy, Emotion, and Patient Perspectives

Interestingly, while many participants doubted that AI could replicate the so-called "human touch," this skepticism merits deeper examination. The term "human touch" was used by participants to denote a complex interplay of empathy, emotional understanding, nonverbal cues, and situational sensitivity—aspects that are perceived as central to patient-centered care. This translates into the ability to recognize and respond to a patient's emotional state, convey reassurance, and build rapport—capacities typically associated with human caregivers. According to most of the participants (26/36, 73%), the concept of empathy is presumed to be uniquely human. However, emerging studies suggest that AI-based chatbots or support tools may sometimes be perceived as more empathetic or nonjudgmental than rushed or fatigued physicians, especially in structured communication contexts (eg, symptom checks, emotional support) [8,47-49]. While our participants viewed AI as inherently lacking empathy, it remains unclear whether this is an ontological limitation or simply a function of current design and communication strategies. Across interviews, a recurring pattern suggested differences between patients and support persons in how AI modalities were evaluated. While not uniformly expressed by all participants, support persons more frequently emphasized the practical benefits of predictive, background AI systems that support physicians in risk monitoring and data processing. In contrast, patients more often articulated ambivalence when AI was described as directly interacting with them or shaping physician-patient communication. While some support persons appeared more receptive to background predictive AI systems, and several patients expressed ambivalence toward more interactive or conversational AI, these patterns were not uniformly observed across all interviews. Rather than representing a consistent role-based distinction, these differences should be understood as interpretive tendencies within the data, shaped by individual experiences, expectations, and levels of engagement with AI. These patterns suggest that AI tools may need to address the distinct informational and emotional needs of patients and their support persons. However, this area warrants further investigation as the perceived emotional sufficiency of AI may shape both trust and acceptance of end users.

AI Literacy and the "AI Package Leaflet": Building Understanding and Trust Through Transparent Information

The findings highlight that knowledge gaps are a significant barrier to the acceptance of AI in SDM. Patients and support persons often expressed uncertainty about how AI arrives at its recommendations, who controls the data, and what exactly the system "knows" about them. To address this, structured informational tools, such as package leaflets for AI systems, may help bridge this gap [28,34,50]. Such a tool might include: The purpose of the AI-DSS, what type of data it uses, how it generates risk estimates, what its limitations are, how well it performs for specific subgroups, and a brief disclaimer of who is accountable for final decisions. This concept aligns with emerging international efforts to enhance algorithmic transparency and patient-centered communication [44,45]. In this light, an "AI package leaflet" could serve as a translational interface between technical system design and patient-centered AI-based care and decision-making. It would also support the ethical imperative of informed consent, which in the context of AI not only requires understanding treatment options but also the nature and role of digital tools involved [51]. Ideally, such

XSL•FO
RenderX

information should also be co-designed with patients, evaluated for health literacy, and tested in different clinical settings to ensure it meets the diverse informational needs of users [51]. As an illustrative example (not derived from participant data), an "AI package leaflet" could briefly outline what the system does, what type of data it uses, how recommendations are generated, and what its limitations are. Transparent, accessible information like this is key to supporting informed consent, fostering trust, and empowering patients to participate actively in AI-assisted SDM. AI literacy not only shapes general attitudes but also affects how patients and support persons participate in SDM during consultations. If patients do not understand how an AI-based risk estimate is generated, they may feel unable to question or discuss it meaningfully. In contrast, basic understanding of the system's purpose and limitations enables more active engagement, supporting genuinely collaborative decision-making. AI literacy emerged as a key factor shaping how patients engaged with AI-DSS, particularly within SDM processes. In practical terms, limited understanding of AI systems may constrain patients' ability to question recommendations, interpret uncertainty, or actively participate in deliberation. Within clinical workflows, this suggests that AI outputs must not only be explained but also contextualized within the decision-making process, enabling patients to understand how recommendations relate to their individual situation, preferences, and alternatives. Without such integration, AI risks reinforcing passive acceptance rather than supporting meaningful SDM.

Limitations

Limitations of this study may include potential social desirability and interviewer effects. Given the sensitive nature of discussing personal health decisions and emerging technologies such as AI, participants may have been inclined to express favorable views or align their responses with what they perceived to be the interviewer's expectations. In addition, participants' baseline familiarity with AI or digital health tools was not systematically assessed. As a result, differences in technological literacy may have influenced how participants understood and reflected on the role of AI in SDM processes. The study is subject to recall bias as participants reflected retrospectively on clinical appointments. The sample was recruited purposefully at a single German transplant center, which limits the generalizability of the findings to broader transplant populations.

Conclusions

Our findings support previous research suggesting that the needs and expectations of patients regarding AI technologies have often not been explicitly studied [9,10]. In the context of nephrology and SDM, our study shows that patients and their support persons particularly emphasize concerns about a potential loss of the "human touch" in treatment due to increased reliance on AI, as well as a lack of trust in AI-based DSS stemming from limited understanding of these technologies [31,52].

However, participants welcomed AI-based DSS as a supportive tool and simultaneously emphasized its role as a supplement in SDM—not a replacement for human judgment. Furthermore, the successful integration of AI-based DSS into posttransplant care may depend less on technical precision and more on its alignment with the interpersonal and ethical foundations of SDM, including the preservation of the "human touch."

Uncertainty about how AI works, its potential benefits, and its challenges for medical encounters revealed patients' knowledge gaps, highlighting the need for further patient involvement in the design of these tools as well as improved patient information strategies, such as AI "package leaflets." Such strategies may strengthen patient trust through improved patient-physician communication on AI and help patients and support persons become involved in decisions related to their care. Future research should explicitly differentiate between AI modalities when examining patients' perceptions, acceptance, and experiences with AI-based DSS. Comparative studies are needed to assess how predictive background AI-based systems versus interactive AI applications are perceived and integrated into decision-making processes by patients and their support persons.

Acknowledgments

The authors thank all the participants in the study for their valuable insights.

Funding

This study is supported by the German Federal Ministry of Education and Research (grant 01GP2202A).

Data Availability

The datasets generated and analyzed during the study are available from the corresponding author upon reasonable request.

Authors' Contributions

AH devised the project, together with PD, KB, and SM. ZS wrote the manuscript with input from all authors. AB, RR, BO, JJS, ÖEÖ, MH, and SE critically reviewed the article and added remarks based on their areas of expertise. ZS accepts full responsibility for the completed work and controlled the decision to publish.

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J 2021 Jul;8(2):e188-e194. [doi: 10.7861/fhj.2021-0095] [Medline: 34286183]
2. Gala D, Behl H, Shah M, Makaryus AN. The role of artificial intelligence in improving patient outcomes and future of healthcare delivery in cardiology: a narrative review of the literature. Healthcare (Basel) 2024 Feb 16;12(4):481. [doi: 10.3390/healthcare12040481] [Medline: 38391856]
3. Seyahi N, Ozcan SG. Artificial intelligence and kidney transplantation. World J Transplant 2021 Jul 18;11(7):277-289. [doi: 10.5500/wjt.v11.i7.277] [Medline: 34316452]
4. Desai P, Dobes A, Shah A, Abdulhay L, Peterson C, Ancker J, et al. Patient and caregiver perspectives on generative artificial intelligence in patient-centered clinical decision support. : Agency for Healthcare Research and Quality; 2024 URL: https://www.ncbi.nlm.nih.gov/books/NBK617040/ [accessed 2026-05-12]
5. Badrouchi S, Bacha MM, Ahmed A, Ben Abdallah T, Abderrahim E. Predicting long-term outcomes of kidney transplantation in the era of artificial intelligence. Sci Rep 2023 Dec 2;13(1):21273. [doi: 10.1038/s41598-023-48645-w] [Medline: 38042904]
6. Reddy S, Fox J, Purohit MP. Artificial intelligence–enabled healthcare delivery. J R Soc Med 2019 Jan;112(1):22-28. [doi: 10.1177/0141076818815510] [Medline: 30507284]
7. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec;2(4):230-243. [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
8. Li YH, Li YL, Wei MY, Li GY. Innovation and challenges of artificial intelligence technology in personalized healthcare. Sci Rep 2024 Aug 16;14(1):18994. [doi: 10.1038/s41598-024-70073-7] [Medline: 39152194]
9. Moy S, Irannejad M, Manning SJ, et al. Patient perspectives on the use of artificial intelligence in health care: a scoping review. J Patient Cent Res Rev 2024;11(1):51-62. [doi: 10.17294/2330-0698.2029] [Medline: 38596349]
10. Adams SJ, Tang R, Babyn P. Patient perspectives and priorities regarding artificial intelligence in radiology: opportunities for patient-centered radiology. J Am Coll Radiol 2020 Aug;17(8):1034-1036. [doi: 10.1016/j.jacr.2020.01.007] [Medline: 32068006]
11. Elwyn G. Shared decision making: what is the work? Patient Educ Couns 2021 Jul;104(7):1591-1595. [doi: 10.1016/j.pec.2020.11.032] [Medline: 33353840]
12. Al Kuwaiti A, Nazer K, Al-Reedy A, et al. A review of the role of artificial intelligence in healthcare. J Pers Med 2023 Jun 5;13(6):951. [doi: 10.3390/jpm13060951] [Medline: 37373940]
13. Osmanodja B, Sassi Z, Eickmann S, et al. Investigating the impact of AI on shared decision-making in post-kidney transplant care (PRIMA-AI): protocol for a randomized controlled trial. JMIR Res Protoc 2024 Apr 1;13:e54857. [doi: 10.2196/54857] [Medline: 38557315]
14. Sassi Z, Eickmann S, Roller R, et al. Prospectively investigating the impact of AI onshared decision-making in post kidney transplant care (PRIMA-AI): protocol for a longitudinal qualitative study among patients, their support persons and treating physicians at a tertiary care centre. BMJ Open 2024 Oct 1;14(10):e081318. [doi: 10.1136/bmjopen-2023-081318] [Medline: 39353696]
15. Katirai A, Yamamoto BA, Kogetsu A, Kato K. Perspectives on artificial intelligence in healthcare from a Patient and Public Involvement Panel in Japan: an exploratory study. Front Digit Health 2023;5:1229308. [doi: 10.3389/fdgth.2023.1229308] [Medline: 37781456]
16. Esmaeilzadeh P, Mirzaei T, Dharanikota S. Patients' perceptions toward human-artificial intelligence interaction in health care: experimental study. J Med Internet Res 2021 Nov 25;23(11):e25856. [doi: 10.2196/25856] [Medline: 34842535]
17. Wang C, Zhang J, Lassi N, Zhang X. Privacy protection in using artificial intelligence for healthcare: Chinese regulation in comparative perspective. Healthcare (Basel) 2022 Sep 27;10(10):1878. [doi: 10.3390/healthcare10101878] [Medline: 36292325]
18. Mondal H, Mondal S. Ethical and social issues related to AI in healthcare. In: Srivastava A, Mishra V, editors. Artificial Intelligence in Microbiology: Scope and Challenges Volume 1: Elsevier; 2024:247-281. [doi: 10.1016/bs.mim.2024.05.009]
19. Evans H, Snead D. Understanding the errors made by artificial intelligence algorithms in histopathology in terms of patient impact. NPJ Digit Med 2024 Apr 10;7(1):89. [doi: 10.1038/s41746-024-01093-w] [Medline: 38600151]
20. Kotsifa E, Mavroeidis VK. Present and future applications of artificial intelligence in kidney transplantation. J Clin Med 2024 Oct 5;13(19):5939. [doi: 10.3390/jcm13195939] [Medline: 39407999]
21. Khosravi M, Zare Z, Mojtabaeian SM, Izadi R. Artificial intelligence and decision-making in healthcare: a thematic analysis of a systematic review of reviews. Health Serv Res Manag Epidemiol 2024;11:23333928241234863. [doi: 10.1177/23333928241234863] [Medline: 38449840]
22. DeMartino ES, Dudzinski DM, Doyle CK, et al. Who decides when a patient can't? Statutes on alternate decision makers. N Engl J Med 2017 Apr 13;376(15):1478-1482. [doi: 10.1056/NEJMms1611497] [Medline: 28402767]
23. Kim SYH, Karlawish JHT, Caine ED. Current state of research on decision-making competence of cognitively impaired elderly persons. Am J Geriatr Psychiatry 2002;10(2):151-165. [Medline: 11925276]
24. Mauthner O, De Luca E, Poole J, et al. Preparation and support of patients through the transplant process: understanding the recipients' perspectives. Nurs Res Pract 2012;2012:547312. [doi: 10.1155/2012/547312] [Medline: 23125923]

XSL•FO
RenderX

25. Nour El Hadi S, Zanotti R, Danielis M. Lived experiences of persons with heart transplantation: a systematic literature review and meta-synthesis. Heart Lung 2025;69:174-184. [doi: 10.1016/j.hrtlng.2024.10.013] [Medline: 39481147]
26. Witkowski K, Dougherty RB, Neely SR. Public perceptions of artificial intelligence in healthcare: ethical concerns and opportunities for patient-centered care. BMC Med Ethics 2024 Jun 22;25(1):74. [doi: 10.1186/s12910-024-01066-4] [Medline: 38909180]
27. Čartolovni A, Malešević A, Poslon L. Critical analysis of the AI impact on the patient-physician relationship: a multi-stakeholder qualitative study. Digit Health 2023;9:20552076231220833. [doi: 10.1177/20552076231220833] [Medline: 38130798]
28. Roller R, Hahn M, Ravichandran AM, et al. One size fits none: rethinking fairness in medical AI. Presented at: Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP); Aug 1, 2025. [doi: 10.18653/v1/2025.gebnlp-1.25]
29. Robertson C, Woods A, Bergstrand K, Findley J, Balser C, Slepian MJ. Diverse patients' attitudes towards Artificial Intelligence (AI) in diagnosis. PLOS Digit Health 2023 May 19;2(5):e0000237. [doi: 10.1371/journal.pdig.0000237] [Medline: 37205713]
30. Elwyn G, Lloyd A, May C, et al. Collaborative deliberation: a model for patient care. Patient Educ Couns 2014 Nov;97(2):158-164. [doi: 10.1016/j.pec.2014.07.027] [Medline: 25175366]
31. Payerchin R. Survey shows patient mistrust of AI use in health care. Urology Times. 2023. URL: https://www.urologytimes.com/view/survey-shows-patient-mistrust-of-ai-use-in-health-care [accessed 2025-08-16]
32. Park EH, Werder K, Cao L, Ramesh B. Why do family members reject AI in health care? Competing effects of emotions. J Manag Inf Syst 2022;39(3):765-792. [doi: 10.1080/07421222.2022.2096550]
33. Roller R, Mayrdorfer M, Duettmann W, et al. Evaluation of a clinical decision support system for detection of patients at risk after kidney transplantation. Front Public Health 2022;10:979448. [doi: 10.3389/fpubh.2022.979448] [Medline: 36388342]
34. Samhammer D, Beck S, Budde K, et al. Klinische Entscheidungsfindung Mit Künstlicher Intelligenz Ein Interdisziplinärer Governance-Ansatz [Book in German]: Springer; 2023.
35. Agency. Stanford Encyclopedia of Philosophy. URL: https://plato.stanford.edu/archives/win2019/entries/agency/ [accessed 2026-05-16]
36. Butow PN, Maclean M, Dunn SM, Tattersall MHN, Boyer MJ. The dynamics of change: cancer patients' preferences for information, involvement and support. Ann Oncol 1997 Sep;8(9):857-863. [doi: 10.1023/a:1008284006045] [Medline: 9358935]
37. Draucker CB, Martsolf DS, Ross R, Rusk TB. Theoretical sampling and category development in grounded theory. Qual Health Res 2007 Oct;17(8):1137-1148. [doi: 10.1177/1049732307308450] [Medline: 17928484]
38. Elwyn G, Frosch D, Thomson R, et al. Shared decision making: a model for clinical practice. J Gen Intern Med 2012 Oct;27(10):1361-1367. [doi: 10.1007/s11606-012-2077-6] [Medline: 22618581]
39. Sassi Z, Eickmann S, Roller R, et al. Human-centered AI in healthcare: empowering patients and support persons in clinical decision-making. BMC Med Inform Decis Mak 2025 Nov 26;25(1):431. [doi: 10.1186/s12911-025-03298-9] [Medline: 41299433]
40. Panda S, Roy ST. Reflections on emerging HCI–AI research. AI & Soc 2024 Feb;39(1):407-409. [doi: 10.1007/s00146-022-01409-y]
41. Ethics and governance of artificial intelligence for health: WHO guidance. : World Health Organization; 2021 URL: https://iris.who.int/server/api/core/bitstreams/f780d926-4ae3-42ce-a6d6-e898a5562621/content [accessed 2026-05-12]
42. Charles C, Gafni A, Whelan T. Shared decision-making in the medical encounter: what does it mean? (or it takes at least two to tango). Soc Sci Med 1997 Mar;44(5):681-692. [doi: 10.1016/s0277-9536(96)00221-3] [Medline: 9032835]
43. Sadeghi Z, Alizadehsani R, Cifci MA, et al. A review of explainable artificial intelligence in healthcare. Comput Electr Eng 2024 Aug;118:109370. [doi: 10.1016/j.compeleceng.2024.109370]
44. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 2020 Sep;26(9):1351-1363. [doi: 10.1038/s41591-020-1037-7] [Medline: 32908284]
45. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 2020 Sep;26(9):1364-1374. [doi: 10.1038/s41591-020-1034-x] [Medline: 32908283]
46. Hahn M, Samhammer D, Tretter M, Dabrock P. Aktanten im Gesundheitswesen? KI-Systeme aus der Perspektive technischer Vermittlung [Article in German]. Z med Ethik 2025;71(4):446-468. [doi: 10.30965/29498570-20250141]
47. Chen D, Chauhan K, Parsa R, et al. Patient perceptions of empathy in physician and artificial intelligence chatbot responses to patient questions about cancer. NPJ Digit Med 2025 May 13;8(1):275. [doi: 10.1038/s41746-025-01671-6] [Medline: 40360673]
48. Howcroft A, Bennett-Weston A, Khan A, Griffiths J, Gay S, Howick J. AI chatbots versus human healthcare professionals: a systematic review and meta-analysis of empathy in patient care. Br Med Bull 2025 Sep 22;156(1):ldaf017. [doi: 10.1093/bmb/ldaf017] [Medline: 41115171]

XSL•FO
RenderX

49. Chin H, Song H, Baek G, et al. The potential of chatbots for emotional support and promoting mental well-being in different cultures: mixed methods study. J Med Internet Res 2023 Oct 20;25:e51712. [doi: 10.2196/51712] [Medline: 37862063]
50. Ott T, Dabrock P. Transparent human – (non-) transparent technology? The Janus-faced call for transparency in AI-based health care technologies. Front Genet 2022;13:902960. [doi: 10.3389/fgene.2022.902960] [Medline: 36072654]
51. Chau M, Rahman MG, Debnath T. From black box to clarity: strategies for effective AI informed consent in healthcare. Artif Intell Med 2025 Sep;167:103169. [doi: 10.1016/j.artmed.2025.103169] [Medline: 40446592]
52. Longoni C, Bonezzi A, Morewedge CK. Resistance to medical artificial intelligence. J Consum Res 2019 Dec;46(4):629-650. [doi: 10.1093/jcr/ucz013]

Abbreviations

AI: artificial intelligence
CONSORT-AI: Consolidated Standards of Reporting Trials—Artificial Intelligence
DSS: decision support system
KTC: kidney transplant center
SDM: shared decision-making
SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials—Artificial Intelligence

Edited by A Stone; submitted 29.Aug.2025; peer-reviewed by KH Lin, X Liang; revised version received 01.Apr.2026; accepted 02.Apr.2026; published 21.May.2026.

Please cite as:
Sassi Z, Eickmann S, Roller R, Osmanodja B, Spencker JJ, Ömeroğlu ÖE, Burchardt A, Hahn M, Dabrock P, Möller S, Budde K, Herrmann A
Human-AI Interaction in Kidney Transplant Decision Support Systems: Qualitative Study of Patient and Support Person Expectations
J Med Internet Res 2026;28:e83195
URL: https://www.jmir.org/2026/1/e83195
doi:10.2196/83195

© Zeineb Sassi, Sascha Eickmann, Roland Roller, Bilgin Osmanodja, Jakob Joachim Spencker, Ömer Ege Ömeroğlu, Aljoscha Burchardt, Michael Hahn, Peter Dabrock, Sebastian Möller, Klemens Budde, Anne Herrmann. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Detecting Uncoded Self-Harm in Veterans' Electronic Health Records Using Positive and Unlabeled Learning: Retrospective Cohort Study

Praveen Kumar[1], PhD; Alexandria D Viszolay[1], BS; Rajesh Upadhayaya[1], BS; Fariha Moomtaheen[1], MS; Donald R Greer[2], BBA; Cristian G Bologa[1,3], PhD; Kristan A Schneider[1], PhD; Sharon E Davis[4,5], PhD; Michael E Matheny[4,5], MD, MS, MPH; David van der Goes[3,6], PhD; Gerardo Villarreal[3,7], MD; Yiliang Zhu[1], PhD; Mauricio Tohen[7], MBA, MD, DrPH; Scott A Malec[1,3], PhD; Jeremy J Yang[1,3], PhD; Elliot M Fielstein[4,8], PhD; Christophe Gerard Lambert[1,3], PhD

[1]Department of Internal Medicine, School of Medicine, University of New Mexico Health Sciences Center, 1 University of New Mexico, MSC10 5550, Albuquerque, NM, United States

[2]Greer Black Company, Bozeman, MT, United States

[3]Raymond G. Murphy VA Medical Center, Albuquerque, NM, United States

[4]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

[5]VA Tennessee Valley Healthcare System, Nashville, TN, United States

[6]Department of Economics, University of New Mexico, Albuquerque, NM, United States

[7]Department of Psychiatry and Behavioral Sciences, School of Medicine, University of New Mexico Health Sciences Center, Albuquerque, NM, United States

[8]Office of Mental Health, United States Department of Veterans Affairs, Washington, DC, United States

Corresponding Author:
Christophe Gerard Lambert, PhD
Department of Internal Medicine, School of Medicine, University of New Mexico Health Sciences Center, 1 University of New Mexico, MSC10 5550, Albuquerque, NM, United States

Abstract

Background: Underdiagnosis and undercoding are common across mental health conditions, particularly suicide and self-harm. This leaves health care datasets lacking reliable negative examples needed for predictive modeling, phenotype prevalence estimation, and identification of individuals at elevated risk. We use positive and unlabeled (PU) learning to address this challenge.

Objective: This study aims to identify US Veterans whose self-harm events were not explicitly captured through diagnostic codes in electronic health records (EHRs) and estimate the underlying prevalence using a novel PU learning algorithm.

Methods: We performed a retrospective cohort study using Veterans Health Administration EHRs (from October 1, 1999, to August 31, 2019), selecting a random 25% sample of 1,329,120 Veterans out of 5,316,480 (1,193,563 males and 135,557 females) with at least 2 years of observation. The study cohort comprised 24,625 Veterans with coded self-harm and 1,304,495 uncoded, with the mean ages of 38.39 (SD 12.17) and 48.76 (SD 15.04) years, respectively. We applied our PULSNAR (positive unlabeled learning selected not at random) algorithm to estimate the proportion of individuals with uncoded self-harm. Covariates included age, medical conditions, procedures, and clinical observations. Four experts (raters) independently reviewed charts of 97 uncoded Veterans, each selected from 1% intervals of calibrated PULSNAR probabilities from 0.01 to 0.97. Agreement was assessed among raters, PULSNAR classifications, and consensus review decisions. Post hoc calibration was used to refine prevalence estimates.

Results: Of the 159,049 covariates in the dataset, PULSNAR's Extreme Gradient Boosting (XGBoost) model identified 1302 (0.82%) as informative for classification. Only 1.85% (24,625/1,329,120) of Veterans had diagnostic codes indicating self-harm events, while PULSNAR estimated an overall prevalence of 10.46% (139,026/1,329,120) by identifying an additional α=8.77% (114,404/1,304,495) of self-harm cases among the uncoded population. Of the 97 chart-reviewed patients, 39 had documented but uncoded self-harm. PULSNAR probabilities were post hoc calibrated such that their sum over the 97 cases equaled 39, which adjusted the combined coded and imputed prevalence downward from 10.46% to 7.91% (105,133/1,329,120). By applying this calibration to shift the probabilities of all uncoded Veterans, with bootstrapping for confidence intervals, PULSNAR estimates that coded self-harm represents only 23.4% (95% CI 17.76% to 31.51%) of all documented (coded+notes) self-harm.

Conclusions: Under the "selected not at random" assumption, PULSNAR provides an innovative and scalable framework for estimating the clinically documented prevalence of mental health conditions and identifying the uncoded individuals with calibrated

XSL•FO
RenderX

prediction, without requiring confirmed negative labels. This method offers an alternative to time-consuming chart reviews for detecting likely cases missing structured coding capture. By addressing diagnostic undercoding of mental health conditions in EHRs, this approach has the potential to enhance the estimation of mental health prevalence and support screening, activation of automated clinical decision support, targeted intervention, better resource allocation, and research to improve outcomes in real-world settings.

(J Med Internet Res 2026;28:e89071) doi:10.2196/89071

KEYWORDS

self-injurious behavior; machine learning; Veterans' health; PULSNAR; electronic health record; positive unlabeled learning selected not at random

Introduction

Suicide and self-harm remain significant public health concerns in the United States, consistently ranking among the leading causes of death. In 2023, suicide was among the top 8 causes of death for individuals aged 10‑64 years and the second leading cause for those aged 10-34 years [1,2]. According to the Centers for Disease Control and Prevention, suicide accounted for 49,316 deaths in 2023 [1], 1.6% of all deaths. The prevalence of suicidal thoughts and attempts is even higher than suicide deaths. In 2023, an estimated 12.8 million American adults aged 18 years or older experienced serious thoughts of suicide, 3.7 million made a suicide plan, and 1.5 million attempted suicide [1,3]. These data suggest that for each suicide death, there are hundreds of individuals experiencing suicidal ideation or engaging in self-harm.

Veterans are at a disproportionately higher risk of experiencing trauma and self-destructive behavior compared to the general population, accounting for nearly 14% of adult suicide deaths despite representing only 7.6% of the population [4]. The overall unadjusted suicide rate among Veterans was roughly double that of non-Veteran adults (34.7 per 100,000 vs 17.1 per 100,000) [5]. Both male (37.3 vs 28.7 per 100,000) and female Veterans (13.5 vs 7.2 per 100,000) had markedly higher suicide rates than their non-Veteran counterparts. In addition, younger Veterans (ages 18‑34 years) faced the highest risk (47.6 per 100,000)—far exceeding that of any other age group [5]. Co-occurring conditions, such as posttraumatic stress disorder, depression, bipolar disorder, substance use disorders, traumatic brain injury, and prior self-injurious behavior, significantly contribute to this elevated risk [4,6].

Suicide rarely occurs in isolation; it is often preceded by identifiable risk factors, with trauma, past self-harm, and suicidal ideation being among the strongest predictors of future suicide risk [7,8]. Early identification of individuals exhibiting these behaviors is critical, as timely intervention can significantly reduce the risk of suicide. Studying self-harm, therefore, is essential for comprehensive suicide prevention efforts. Within the Veterans Health Administration (VHA), policy and quality metrics primarily focus on suicidal self-directed violence within the past 12 months, rather than self-harm broadly defined. Although our study centers on ever self-harm as a phenotype, it is important to distinguish this analytic focus from operational VHA surveillance, which prioritizes identification and follow-up of suicidal behaviors in the near-term versus capturing all self-harm behaviors with and without intent.

The widespread adoption of electronic health records (EHRs) in the United States has generated large repositories of patient health care data, comprising both structured data (eg, demographics, diagnoses, procedures, and prescriptions) and unstructured data (eg, clinical notes, imaging, and pathology reports) [9-11]. While these data are invaluable for observational studies and health care analytics, they also reveal inconsistencies in data quality across different clinical settings and heterogeneous data structures that span structured codes, free-text notes, and temporal measurements [11]. Additionally, these data are affected by missingness due to multiple mechanisms, including the incomplete capture of out-of-network care, underreporting of sensitive conditions, and not documenting or undercoding of clinical diagnoses. Undercoding refers to instances in which events or conditions, such as self-harm behaviors and mental illnesses, are not recorded in structured diagnosis codes in the EHR, leading to physicians overlooking past diagnoses, inaccurate prevalence estimates, and hindering effective intervention strategies. Importantly, undercoding of mental health conditions, including suicidality and self-harm, is common in EHRs, limiting research that relies on accurate reporting of these conditions [12-15]. VHA's suicide risk surveillance relies heavily on policy-driven, standardized national note templates such as the Suicide Behavior and Overdose Report (SBOR), which capture suicidal and other self-directed violence events in structured clinical documentation. These templates function in parallel to diagnostic coding and are a primary mechanism by which VHA monitors suicide behaviors. Undercoding of *ICD* (*International Classification of Diseases*) self-harm diagnoses reflects partially the extent of VHA suicide risk monitoring efforts.

The increased availability of EHRs and advancements in machine learning (ML) methodologies have led to increased application of ML techniques to identify and predict instances of self-harm and suicidal ideation using EHRs and insurance claims data [12,13,15-22]. Kumar et al [12] and Nestsiarovich et al [13] used an Extreme Gradient Boosting (XGBoost) [23] model on visit-level data to estimate the uncoded self-harm events among individuals with major mental illness. Simon et al [16] developed random forest–based models aimed at predicting fatal or nonfatal self-harm events within 90 days following a sampled encounter. Simon et al [17] used LASSO (least absolute shrinkage and selection operator)–based logistic regression models to predict suicide attempts and suicide deaths post-outpatient visits. Rozova et al [18] used natural language processing (NLP) supervised learning techniques on free-text triage notes from emergency department (ED) visits to detect

XSL•FO
RenderX

self-harm and suicidal ideation among ED patients. Walsh et al [19] used random forest and nonregularized logistic regression models on longitudinal clinical data to detect the risk of nonfatal suicide attempts in adolescents. Tsui et al [20] leveraged NLP models on clinical notes along with 4 ML techniques, including Naïve Bayes, LASSO regression, random forest, and an ensemble of extreme gradient boosting, for the prediction of first-time suicide attempts. Su et al [21] used LASSO logistic regression to predict suicidal behavior among children and adolescents based on their longitudinal clinical records, identifying both short- and long-term risk factors. Barak-Corren et al [22] developed Bayesian models using a retrospective cohort approach to predict future documented suicidal behavior.

Previous studies used traditional positive-negative ML classifiers to identify/predict instances of self-harm in health records. However, due to undercoding common in mental health data, unlabeled instances contain both positive (diagnosed and undiagnosed) and negative (unaffected) cases, leading to biased classification and predictions. To address this, we applied a novel positive and unlabeled (PU) learning algorithm, PULSNAR (positive unlabeled learning selected not at random) [24], to Veterans' EHR data to estimate the proportion of Veterans with ever self-harm. Notably, we are not predicting future self-harm, but rather classifying whether patients had experienced self-harm at any point during the study period. PU learning is a semisupervised approach that uses labeled (coded) positive examples and unlabeled (uncoded) examples containing an unknown mixture of positives and negatives [24].

To our knowledge, this is the first study to use PU learning algorithms to estimate the proportion of Veterans ever with self-harm imputing uncoded individuals in Veterans' EHR data. By applying PULSNAR, we aim to improve the detection and estimation of self-harm prevalence among US Veterans, demonstrating a framework applicable to detecting other undetected mental health diagnoses. This approach supports earlier screening as well as novel intervention strategies to reduce self-harm and suicide rates in this high-risk population. Our findings are likely to improve awareness of risk factors of self-destructive behaviors among Veterans and highlight the broader use of PU learning in mental health informatics.

Methods

PU Learning Background

In many real-world applications, annotating all records can be challenging, expensive, or even impossible due to the volume of data. Often, only positive instances are labeled, leaving a considerable portion of the data unlabeled [25]. Notably, an unlabeled instance does not necessarily indicate a negative case because the absence of a diagnosis code does not confirm the absence of a condition. Given that only a fraction of records are labeled, learning from PU data has emerged as an active area of research [24-32]. The majority of current PU learning methods are based on the "selected completely at random" (SCAR) assumption [25-30], which posits that labeled positives are independent and identically distributed (i.i.d.) random samples from the positive distribution, meaning that the probability of an instance being labeled as positive is independent of its attributes [26]. However, in real-world applications, this assumption often does not hold due to labeling bias; for example, patients with more severe or specific symptoms are more likely to be labeled with a disease than those with milder or unspecific symptoms.

Despite the prevalence of real-world non-SCAR PU data, only a few studies have focused on PU learning under the non-SCAR assumption [24,31,32]. In this study, we have applied PULSNAR, a PU algorithm based on the SNAR (selected not at random) assumption. PULSNAR uses a divide-and-conquer approach to cluster SNAR positives into subtypes and estimate the proportion of each subtype among the unlabeled using the PULSCAR (positive unlabeled learning selected completely at random) algorithm. Under the SNAR assumption, the probability that a positive example is labeled depends on its attributes [24,32], which is more appropriate for health care data for which labeling bias is prevalent.

PU Learning Algorithms to Estimate the Proportion of Self-Harm Among Uncoded Individuals

We found that existing state-of-the-art PU learning methods suffer from scalability issues and fail to execute on large datasets [24]. As a result, in earlier work, we developed 2 novel PU learning algorithms aimed at estimating the proportion (α) of positive instances among unlabeled examples and subsequently imputing these instances: PULSCAR for when the SCAR assumption holds, and PULSNAR for when it does not. What sets our methods apart from other PU methods is that they not only estimate α but also calculate calibrated probabilities using the estimated α, leading to markedly improved classification performance. We focus our analysis on the application of the PULSNAR framework to universally available structured EHR data; extensive methodological comparisons demonstrating its superiority over standard supervised baselines have been detailed in our prior work [24,33].

EHR data may contain various types of self-harm events, such as hanging, poisoning, cutting, etc, and the likelihood of coding these different types of self-harm may differ due to the severity of the condition and the source and nature of the underlying trauma. Therefore, it becomes evident that the SCAR assumption may not hold for self-harm data, and SCAR-based PU methods may not be suitable for such data. Recognizing this limitation, we applied the PULSNAR algorithm to estimate the proportion of self-harm-uncoded patients in an EHR dataset of US Veterans.

When dealing with highly class-imbalanced training datasets, ML algorithms often exhibit a bias toward the majority class, struggling to effectively generalize patterns from the minority class, which yields biased predictions [34]. Since only a tiny fraction of records were coded for self-harm in EHR data, we adopted a random undersampling approach [35], where we transformed the imbalanced dataset into k balanced datasets (as illustrated in Figure 1), where $k=floor(|\text{unlabeled}|/|\text{labeled}|)$. Each balanced dataset comprised all labeled self-harm records along with a similar number of unlabeled records. Subsequently, we applied PULSCAR and PULSNAR algorithms (with XGBoost as the classifier) to each balanced dataset to estimate the proportion (α) of self-harm among the unlabeled records.

XSL•FO
RenderX

The final α was determined by computing the mean of the *k* estimates of α. We performed hyperparameter tuning to determine optimal values for the XGBoost parameters. To better explore the high-dimensional covariate space, we used aggressive feature subsampling combined with a low learning rate and a large ensemble of deeper trees. This configuration limits early dominance by a small set of predictors and promotes broader exploration of the covariate space. With this sampling scheme, deeper trees provide more opportunities for additional variables to enter each tree and enable the model to capture complex interactions among covariates. Although deeper trees may increase the risk of overfitting, the large sample size and cross-validation procedures mitigate this risk. The final XGBoost parameter values were *max_depth*=12, *n_estimators*=400, *learning_rate*=0.05, *min_child_weight*=1, *colsample_bytree*=0.3, *colsample_bylevel*=0.6, *n_jobs*=32, *random_state*=0, and *objective='binary:logistic'*. Other parameters were kept at their default values.

Figure 1. Steps for α estimation when the dataset has high class imbalance. The mix of blue and red represents the unlabeled set (U), and blue represents the positive set (P). To address class imbalance, *k* balanced datasets are generated, where *k=floor*(|unlabeled|/|labeled|). Each balanced dataset includes all labeled records and a randomly selected subset of unlabeled records of comparable size. The PU algorithm is applied to each of the *k* balanced datasets to estimate the proportion (α) of positives within the unlabeled set. The final α estimate is obtained by averaging the *k* individual α estimates. PULSCAR: positive unlabeled learning selected completely at random; PULSNAR: positive unlabeled learning selected not at random.

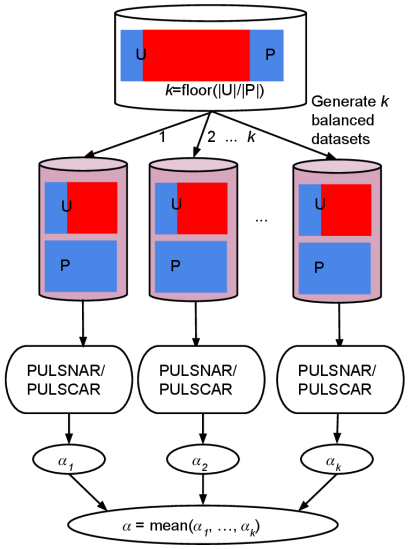

Data Source

For this study, we used the Veterans Health Administration (VHA) EHR data (from October 1, 1999, to August 31, 2019) available in the OMOP CDMv5 (Observational Medical Outcomes Partnership Common Data Model) [36] format, selecting a random 25% sample of 1,329,120 Veterans (out of 5,316,480). The only inclusion criterion was that Veterans had at least 2 years of enrollment. Ethical approvals and data use agreements were obtained from the appropriate institutional review boards to ensure compliance with privacy and confidentiality regulations.

Phenotyping and Covariate Selection

A self-harm phenotype was defined by the presence of one or more *ICD-10-CM* (*International Classification of Diseases, 10th Revision, Clinical Modification*) or *ICD-9-CM* codes (Table S1 in Multimedia Appendix 1). These codes encompass all instances of intentional self-harm or suicide attempts by any means, including a history of self-harm. Patients with any of these codes were labeled as positive cases, while all others remained unlabeled. Accurate identification of self-harm events in structured data is complicated by self-directed violence nomenclature and coding rules that require documentation of clinical intent. Injury and poisoning codes must distinguish accidental events, intentional self-harm, assault, or undetermined causes, and in the absence of clear documentation, intent is often coded as accidental. These challenges are not unique to VHA; they affect coding practices in non-VHA systems and research cohorts as well.

Covariates included patient age at enrollment and the presence/absence of medical conditions, procedures, and clinical observations over the duration of patient observation. *ICD-9-CM* and *ICD-10-CM* diagnosis codes were mapped to their Systematized Nomenclature of Medicine (SNOMED) equivalents (and all ancestors thereof) using the OMOP vocabulary [37]. Procedure codes from *ICD-9-CM Volume 3* (*ICD-9-CM V3*), the *ICD-10 Procedure Coding System* (*ICD-10-PCS*), and *Current Procedural Terminology, Fourth Edition*, were mapped to *ICD-10-PCS* concepts (and all ancestors thereof). Overall, from a dataset of 1,329,120 Veterans, we selected 159,049 covariates for use in the PU learning algorithm. Codes used to define the self-harm phenotype (Table S1 in Multimedia Appendix 1) were excluded from the covariate list to prevent data leakage. Since each patient had an average of only 1203 nonzero features, a compressed sparse row (CSR) matrix with 1,329,120 rows (patients) and 159,049 columns (covariates) was created as input for the PU models. Covariates were encoded as binary values (0/1) in the CSR matrix; if a covariate was not present in an individual's data, it was set to 0, and if it was present, it was set to 1. Therefore, no covariate included in the modeling framework contained missing values. As a result, tests for missingness (eg, missing completely at random testing) and multiple imputation

XSL•FO
RenderX

procedures were not applicable for our study. Figure 2 illustrates the complete schema of our study.

Figure 2. Study schema. First, *k* balanced datasets were generated, and the PU learning method was applied to each of them. The parameter α_j was estimated for each balanced dataset, and the final α for the complete dataset was computed by taking the mean of the *k* estimated values $\alpha_1,...,\alpha_k$. Using the estimated α, the PU method calculated the calibrated probability of being labeled as positive for each uncoded individual. These calibrated classifications were then used to determine probable positive (PP) and probable negative (PN) individuals among the uncoded individuals. CP: coded positive; OMOP: Observational Medical Outcomes Partnership; PULSNAR: positive unlabeled learning selected not at random; VHA: Veterans Health Administration.

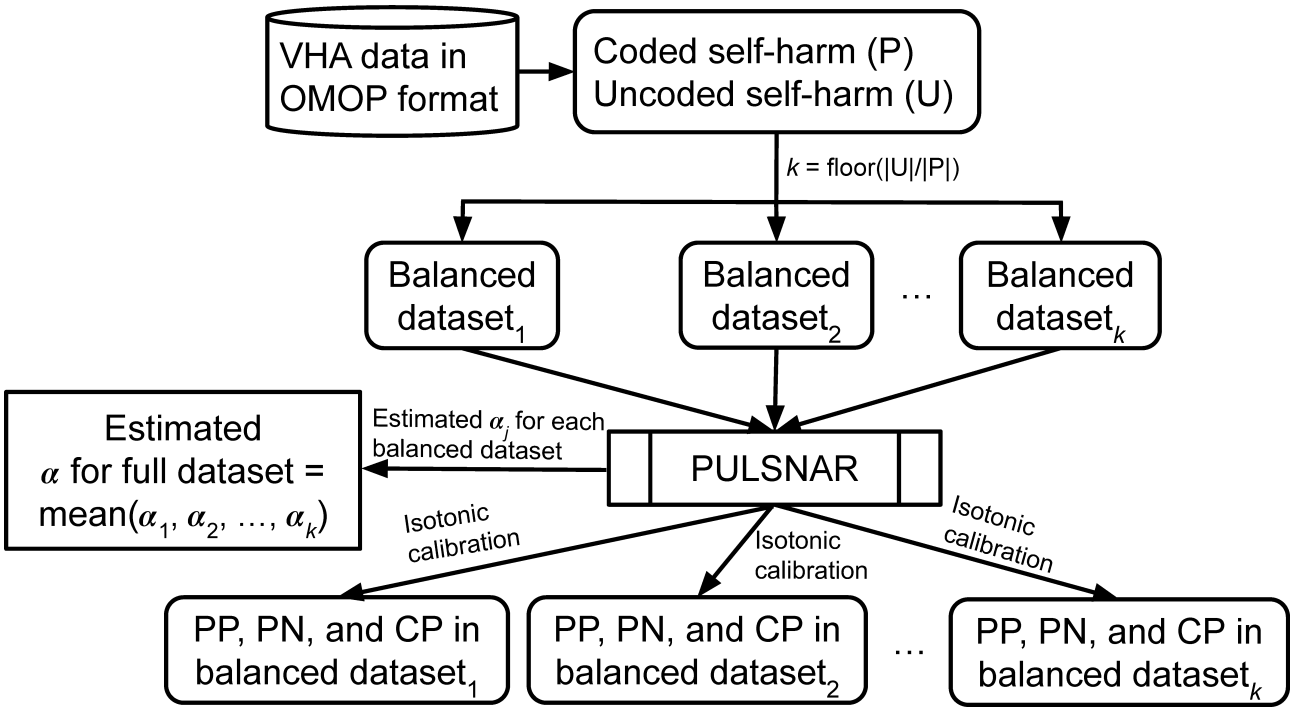

Chart Review Process

To validate our model's classifications, we conducted a chart review of a random sample of unlabeled individuals whose calibrated probabilities fell into each 1% probability bin. Due to the absence of unlabeled individuals in the calibrated probability bins in the first and last two percentiles, the selection was limited to 97 individuals. To ensure rigorous clinical validity, a comprehensive set of search keywords was established a priori in direct collaboration with our clinical psychiatrist coauthor (GV). Four informaticists (coauthors CGB, DvdG, JJY, and CGL) used these clinically vetted keywords and search utilities to identify and extract potential evidence of suicidal or nonsuicidal self-harm from combined charts, which often exceeded 500,000 lines per patient. Interrater reliability was assessed at this stage to ensure consistency among the informatics reviewers. Crucially, the extracted clinical evidence for each of the 97 cases was subsequently evaluated and discussed in detail with the psychiatrist. Rather than limiting clinical oversight to disagreement resolution among informaticists, this joint case-by-case review ensured that definitional and semantic complexities (eg, distinguishing when chronic substance abuse strictly constitutes intentional self-harm) were uniformly adjudicated by a mental health clinician. Through this expert-guided process, individuals were manually classified as positive (class 1) or negative (class 0) for ever self-harm. Finally, we compared the sum of the PULSNAR-calibrated probabilities of the 97 individuals with the sum of their consensus labels identified through expert chart review.

Calibrated Self-Harm Prevalence Estimates for Veterans

To calibrate the PULSNAR-estimated fraction of patients with self-harm against consensus chart review results, we applied a bias-only logit shift—a logistic transformation that adds a uniform constant offset (*c*) to the logits of the original probabilities without altering their ordering. Mathematically, for each probability p_i, the logit is computed as li=log (pi1−pi); *c* is solved numerically (via root-finding) such that ∑i=0|U|11+exp (−(li+c)) is equal to the number of positives confirmed by chart review. The adjusted probabilities are p^i=11+exp (−(li+c^)). This method preserves relative ordering and can shift probabilities lower when correcting overestimation, providing exact sum alignment to the gold standard. We quantified calibration uncertainty using a 100,000-sample bootstrap, sampling chart-reviewed observations with replacement, solving for *c* each time to produce an empirical distribution. The 2.5th and 97.5th percentiles formed the 95% CI for *c*^, which we then applied to the unlabeled population to bound the total estimated positives.

PULSNAR Classification vs Expert Chart Review

The probability of a self-harm diagnosis was stratified into low (probability < 25%), intermediate (25% ≤ probability ≤ 75%), or high (probability > 75%) categories. These stratified probability categories (which also serve as a proxy measure for

XSL•FO
RenderX

the underlying behavioral risk or detection and identification of such) were compared with the outcome of the chart reviews stratified into unanimous decision against self-harm (ie, all reviewers agreed that no evidence of self-harm was found in the patient chart), dissenting evidence for self-harm (ie, some but not all reviewers agreed on self-harm), and unanimous decision for self-harm (ie, all reviewers agreed on self-harm).

Ethical Considerations

Ethical approval for this study was obtained from the institutional review board and research and development committees of the VA New Mexico Healthcare System (H2947), University of New Mexico Health Sciences Center (20-477), and VA Tennessee Valley Healthcare System (#1576576). The requirement for informed consent was waived because the study involved secondary analysis of existing electronic health record data and posed minimal risk to participants. All data were deidentified prior to analysis, and access to the data was restricted to authorized study personnel and maintained on secure, password-protected VA servers to ensure privacy and confidentiality. No participants were contacted for this study, and no compensation was provided. No identifiable individual-level information or images are included in this manuscript or supplementary materials.

Results

Patient Characteristics

Applying our inclusion and exclusion criteria, we identified 1,329,120 individuals (1,193,563 males and 135,557 females) for the study. Table 1 summarizes the key demographic and clinical characteristics of the study population.

XSL•FO
RenderX

Table . Demographic and clinical characteristics of patients with and without coded self-harm selected from Veterans Health Administration electronic health record data (October 1, 1999, to August 31, 2019). The listed comorbidities represent key covariates identified as important features by the XGBoost[a] model within the PULSNAR[b] algorithm.

Patient characteristics (n=1,329,120)	Coded for self-harm (n=24,625, 1.85%)		Uncoded for self-harm (n=1,304,495, 98.15%)		OR[c] (95% CI)
	Patients, n (%)	95% CI	Patients, n (%)	95% CI	
Male	20,723 (84.15)	83.69 - 84.61	1,172,840 (89.91)	89.86 - 89.96	0.60 (0.58 - 0.62)
Female	3902 (15.85)	15.39 - 16.31	131,655 (10.09)	10.04 - 10.14	1.68 (1.62 - 1.74)
Age (years), mean (SD)	38.39 (12.17)	—[d]	48.76 (15.04)	—	—
Age (years)					
0 - 19	1301 (5.28)	5.01 - 5.57	32,128 (2.46)	2.44 - 2.49	2.21 (2.09 - 2.34)
20 - 29	6126 (24.88)	24.34 - 25.42	168,060 (12.88)	12.83 - 12.94	2.24 (2.17 - 2.31)
30 - 39	5424 (22.03)	21.51 - 22.55	164,091 (12.58)	12.52 - 12.64	1.96 (1.90 - 2.02)
40 - 49	7261 (29.49)	28.92 - 30.06	257,077 (19.71)	19.64 - 19.78	1.70 (1.66 - 1.75)
50 - 59	3712 (15.07)	14.63 - 15.53	365,610 (28.03)	27.95 - 28.10	0.46 (0.44 - 0.47)
≥60	801 (3.25)	3.03 - 3.48	317,529 (24.34)	24.27 - 24.41	0.10 (0.10 - 0.11)
Comorbidities					
Personality disorder	2536 (10.30)	9.92 - 10.68	5627 (0.43)	0.42 - 0.44	26.48 (25.22 - 27.82)
Bipolar disorder	2374 (9.64)	9.27 - 10.02	5887 (0.45)	0.44 - 0.46	23.53 (22.38 - 24.72)
Schizophrenia	944 (3.83)	3.60 - 4.08	2277 (0.17)	0.17 - 0.18	22.79 (21.09 - 24.64)
Major depressive disorder	4412 (17.92)	17.44 - 18.40	13,572 (1.04)	1.02 - 1.06	20.76 (20.00 - 21.53)
Psychoactive substance use disorder	3783 (15.36)	14.91 - 15.82	10,933 (0.84)	0.82 - 0.85	21.47 (20.63 - 22.34)
Posttraumatic stress disorder	3563 (14.47)	14.03 - 14.91	10,369 (0.79)	0.78 - 0.81	21.10 (20.27 - 22.00)
Attention-deficit/hyperactivity disorder	509 (2.07)	1.89 - 2.25	1266 (0.10)	0.09 - 0.10	21.72 (19.56 - 24.11)
Chronic pain	2538 (10.31)	9.93 - 10.69	6977 (0.53)	0.52 - 0.55	21.37 (20.37 - 22.41)
Adjustment disorder	2748 (11.16)	10.77 - 11.56	7763 (0.60)	0.58 - 0.61	20.98 (20.03 - 21.97)
Sleep disorder	3570 (14.50)	14.06 - 14.94	10,744 (0.82)	0.81 - 0.84	20.41 (19.59 - 21.25)
Traumatic brain injury	56 (0.23)	0.17 - 0.30	101 (0.01)	0.01 - 0.01	29.45 (20.84 - 41.21)
Anxiety disorder	4424 (17.97)	17.49 - 18.45	13,855 (1.06)	1.04 - 1.08	20.40 (19.66 - 21.16)
Induced psychotic disorder	21 (0.09)	0.05 - 0.13	34 (0.00)	0.00 - 0.00	32.75 (18.05 - 58.10)

[a]XGBoost: Extreme Gradient Boosting.

[b]PULSNAR: positive unlabeled learning selected not at random.

[c]OR: odds ratio.

[d]Not applicable.

Performance of PU Models

Table 2 presents the estimated proportion of self-harm cases by the PULSCAR and PULSNAR algorithms. The proportion of individuals coded for self-harm was only 1.85% (24,625/1,329,120). As expected, PULSCAR provided a lower α estimate (21,524/1,304,495, 1.65%) compared to PULSNAR (114,404/1,304,495, 8.77%), because the EHR self-harm data of Veterans did not satisfy the SCAR assumption. PULSCAR estimated 3.47% (46,120/1,329,120) of the individuals with coded and imputed self-harm, while PULSNAR provided a higher estimate of 10.46% (139,026/1,329,120).

XSL•FO
RenderX

Table . Performance of the PULSCAR[a] and PULSNAR[b] algorithms in imputing uncoded self-harm cases from Veterans Health Administration (VHA) electronic health record data. The parameter α denotes the estimated proportion of uncoded self-harm cases identified by each method.

	VHA (ever self-harm)
Records, n	1,329,120
Coded self-harm	24,625 (1.85%, 95% CI 1.83% - 1.88%)
Covariates, n	159,049
Covariate types	3 (condition, procedure, and observation)
Important covariates, n	1302
α using PULSCAR	1.65%
α using PULSNAR	8.77%
Coded+imputed self-harm using PULSCAR	3.47%
Coded+imputed self-harm using PULSNAR	10.46% (14 clusters)

[a]PULSCAR: positive unlabeled learning selected completely at random.

[b]PULSNAR: positive unlabeled learning selected not at random.

Important Covariates Used by the XGBoost Model

Our covariate selection approach identified 159,049 covariates for the XGBoost model used in the PULSCAR and PULSNAR methods, but XGBoost identified only 1302 (0.82%) as informative for classification. Feature (covariate) importance was measured by the average gain (mean reduction in the loss function) contributed by each covariate. The top 15 covariates with the highest gain scores are shown in Table 3.

XSL•FO
RenderX

Table . Top 15 covariates identified by the XGBoost[a] model, along with their corresponding gain scores. Of the 159,049 covariates included in the positive and unlabeled models based on XGBoost, only 1302 (0.82%) were identified by the model as contributing to training in our cohort. Feature (covariate) importance was quantified using average gain, defined as the mean reduction in the model's loss function attributable to splits on a given covariate across all trees; thus, higher gain values indicate a greater contribution to improving the model's predictive performance.

OMOP[b] concept ID	Concept name	Gain score
438028	Poisoning by a drug AND/OR a medicinal substance	21.61
4009713	Thoughts of self-harm	14.62
440270	Poisoning by antidepressants	11.18
2840354	Medical and Surgical @ Anatomical Regions, Upper Extremities @ Supplement @ Axilla, Left @ Open @ Autologous Tissue Substitute	10.68
44782421	Acute deep venous thrombosis of the upper extremity	10.50
4133169	Operative procedure on the pelvis	10.44
434626	Borderline personality disorder	9.87
4306645	Finding of thought content	9.73
2887059	Medical and Surgical @ Gastrointestinal System @ Bypass @ Descending Colon @ Open	9.59
4181019	Cluster B personality disorder	9.49
2832323	Medical and Surgical @ Anatomical Regions, General @ Repair @ Abdominal Wall @ Percutaneous @ No Device	9.47
2841968	Imaging @ Lower Arteries @ Fluoroscopy @ Abdominal Aorta @ Low Osmolar	9.30
2762368	Supplement Thoracic Vertebra with Nonautologous Tissue Substitute, Percutaneous Approach	9.27
437456	Poisoning by an anticonvulsant	8.90
2895851	Medical and Surgical @ Upper Bones @ Excision @ Clavicle, Left @ Open	8.87

[a]XGBoost: Extreme Gradient Boosting.

[b]OMOP: Observational Medical Outcomes Partnership.

Chart Review

Table 4 presents the interreviewer agreement among 4 reviewers across 97 uncoded self-harm cases, as well as the agreement between individual reviewers and PULSNAR, and between reviewers and the overall reviewer consensus. Table S2 in Multimedia Appendix 1 reports the pairwise Cohen κ coefficients for these agreements. Interreviewer reliability among 4 reviewers across 97 uncoded self-harm cases was substantial (Fleiss κ=0.668, z=16.1, P<.001). Notably, there were 11 patients of 39 (28.2%) where the consensus review was positive, but at least one reviewer missed the relevant evidence. Of 4×39=156 evaluations of positive charts, 17 (10.9%) were false negatives.

XSL•FO
RenderX

Table . Agreement between PULSNAR[a] and each reviewer, between individual reviewers, between each reviewer and their consensus, and between the reviewers' consensus and PULSNAR for 97 uncoded self-harm cases. Agreement is calculated as the percentage agreeing out of the 97 charts. Overall interreviewer reliability was substantial (Fleiss κ=0.668, z=16.1, P<.001). Among 39 consensus-positive cases, 11 (28.2%) had at least one reviewer miss relevant evidence.

	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	PULSNAR	Reviewer consensus
Reviewer 1	100.0%	89.7%	94.9%	82.5%	58.8%	88.7%
Reviewer 2	89.7%	100.0%	88.7%	76.3%	54.6%	82.5%
Reviewer 3	94.9%	88.7%	100.0%	79.4%	59.8%	87.6%
Reviewer 4	82.5%	76.3%	79.4%	100.0%	63.9%	85.6%
PULSNAR	58.8%	54.6%	59.8%	63.9%	100.0%	59.8%
Reviewer consensus	88.7%	82.5%	87.6%	85.6%	59.8%	100.0%

[a]PULSNAR: positive unlabeled learning selected not at random.

For completeness, we also calculated standard classification metrics on the 97 chart-reviewed cases using the PULSNAR probabilities versus expert consensus as the ground truth. The model demonstrated an area under the curve (AUC) of 0.6813, an F_1-score of 0.5517, a precision of 0.5, and a recall of 0.6154.

Post Hoc Calibration

After applying the bias-only logit shift on the PULSNAR probabilities for the 97 reviewed charts, using their consensus chart-reviewed self-harm labels, the sum of transformed PULSNAR probabilities equaled 39 (the expert count), with a logit shift parameter of c=−0.54 (bootstrap 95% CI −1.1420 to −0.0252). Applying this transformation to the 1,304,495 unlabeled patients yielded a sum of adjusted probabilities of 80,574.7, corresponding to 6.18% positives among the uncoded patients, which is lower than the PULSNAR-estimated α of 8.77%. This suggests that, if all notes for all Veterans without coded self-harm were chart-reviewed, 6.18% (95% CI 4.1%-8.74%) would reveal a documented history of self-harm. Thus, coded self-harm represents approximately 24,625/(80,574.7+24,625)=23.4% (95% CI 17.76%-31.51%) of all documented (coded+notes) self-harm.

Efficiency/Scalability Quantification

To quantify the operational utility and scalability of the PULSNAR approach, we evaluated both the manual effort required for expert chart review and the computational resources consumed by our model inference across the full cohort. Expert chart review time ranged from approximately 25 minutes to 2 hours per patient chart, depending on the complexity of the record and the volume of clinical notes. Across the 97 charts reviewed in this study, this corresponds to an estimated total effort of approximately 40 - 200 person-hours per reviewer. In contrast, running the PULSNAR model for inference on the entire cohort of 1.3 million records was computationally tractable, requiring approximately 63 hours of wall-clock time on a machine with 16 vCPUs, 128 GB of RAM, and 128 GB of disk space.

PULSNAR Classification vs Expert Chart Review

Table 5 shows the comparison of PULSNAR-classified self-harm risk categories with chart review outcomes. Among those Veterans for whom expert reviewers could not find evidence of self-harm in some patient charts, PULSNAR classified the probability of self-harm as low in 35.42% of the cases and found it to be intermediate or high in the remaining 64.58% of the cases. Classifying patients as intermediate or high risk despite the absence of self-harm coding is a desirable feature, as the patient chart might also not document self-harm. Among those Veterans for whom some, but not all, expert reviewers agreed on recorded self-harm behavior in the patient chart, PULSNAR would classify the probability of self-harm to be low only in 14.81% of the cases, and intermediate to high in 85.19% of the cases. Among those Veterans for whom the expert reviewers unanimously agreed on evidence of self-harm behavior in the patient chart, PULSNAR estimated the probability of self-harming behavior as low only in 13.64% of the cases, and intermediate to high in 86.36% of the cases. Stratified as such, there was an association between expert chart reviewers and PULSNAR (Fisher exact test: P=.02).

Table . Comparison of PULSNAR[a]-classified self-harm risk categories (low, intermediate, and high probability) with expert chart review outcomes (unanimous against self-harm, dissenting evidence for self-harm, and unanimous for self-harm) among 97 chart-reviewed uncoded cases. There was a statistically significant association between PULSNAR risk categories and chart review outcomes (Fisher exact test, P=.02).

Expert review	PULSNAR classification		
	Low	Intermediate	High
Unanimous no self-harm	35.42	50.00	14.58
Dissenting self-harm	14.81	40.74	44.44
Unanimous self-harm	13.64	68.18	18.18

[a]PULSNAR: positive unlabeled learning selected not at random.

XSL•FO
RenderX

Discussion

Principal Results

In this study, we successfully applied the novel PULSNAR algorithm to a large representative cohort of US Veterans' electronic health records to estimate the burden of self-harm beyond what is captured in structured diagnostic codes. Our primary finding is that relying solely on structured diagnostic codes dramatically underestimates the clinically documented prevalence of self-harm. By using only structured data within a PU learning framework and calibrating predictions against expert chart review of clinical notes, our approach provides a more comprehensive estimate of self-harm prevalence and highlights the extent of undercoding in VA electronic health records. These results support our central hypothesis that PU learning can recover hidden disease burden in routinely collected health care data. Critically, our results suggest that only about 1 in 4 patients with clinical notes documenting self-harm, or a history thereof, have the condition captured in structured diagnostic codes. While the discrimination metrics (AUC=0.68) appear modest, this is primarily because the model relies solely on structured data for prediction, while the validation ground truth is derived from a comprehensive review of free-text clinical notes. Furthermore, the nonrandom, uncoded test set, which lacks confirmed negatives, is optimized not for discrimination assessment, but rather prevalence calibration, which remains the study's primary, successful objective.

Self-harm, like many sensitive mental health conditions, is frequently underreported in EHRs due to confidentiality concerns, stigma, limited help-seeking, barriers to care, and inconsistent screening and documentation practices across health care settings [38]. Additionally, self-harm is likely underdocumented in *ICD* codes in VHA records because VHA facilities receive federal funding on a per-enrolled-patient basis rather than through per-service billing, reducing incentives for exhaustive diagnostic coding [39]. It should be noted that VHA suicide risk surveillance does not rely solely on *ICD* codes. Policy and standardized templates such as SBOR are used to systematically document suicidal behaviors and overdoses [40], meaning that *ICD* undercoding does not present a complete picture of VHA's operational monitoring of suicide risk. This underreporting hinders accurate prevalence estimation, risk identification, resource allocation, risk modeling, and intervention design [12,13,41,42]. To address these gaps, we applied PULSNAR to estimate the proportion of Veterans with both coded and uncoded self-harm and to identify likely cases at scale, serving as a case study for broader phenotyping of undercoded mental illnesses.

Veterans' health records often contain hundreds of thousands of lines of clinical notes, making it challenging and time-consuming for clinicians to manually identify individuals at risk of self-harm, suicidality, or other mental health phenotypes through chart review [43,44]. More broadly, the self-directed violence nomenclature itself is often confusing to frontline clinicians, and the expectation of uniform, intent-based coding is difficult to meet in routine care [45]. Because injury and poisoning codes must encode intent, incomplete or ambiguous documentation pushes events toward "accidental" categories, hindering differentiation between nonsuicidal self-harm and suicidal self-directed violence [46]. This is a fundamental limitation of current surveillance approaches across health systems, and it constrains the accuracy of both administrative data and research phenotypes. Further, while one might hope that a patient's problem list would contain a history of self-harm if it was ever noted, we found that only 22.6% (5556/24,625) of patients with coded self-harm ever had self-harm or self-harm history recorded in their VHA problem list. Given the high workload and time constraints faced by health care providers, thoroughly reviewing lengthy patient records can significantly slow the assessment process, potentially delaying risk assessment and timely intervention [47]. Another structural barrier is the absence of required, standardized training for VHA or non-VHA clinicians on self-directed violence nomenclature and the use of structured diagnostic codes for suicide risk surveillance. In practice, documentation and coding are shaped by local culture, individual comfort, and time constraints rather than uniform training, which contributes to inconsistent recognition, labeling, and coding of self-harm and suicidal behaviors across settings [48,49]. The PULSNAR method addresses this limitation by efficiently estimating the probability of self-harm through analysis of patient data, thereby assisting clinicians in prioritizing high-risk individuals for further evaluation.

Our expert chart review of 97 unlabeled Veterans yielded 3 key insights relevant to validating algorithmic phenotypes. First, probability calibration is essential and can be achieved with limited chart review. Applying post hoc calibration to the full cohort resulted in a more conservative estimate of self-harm prevalence at 7.91%, in contrast to PULSNAR's estimate of 10.46%. Both estimates, however, are consistent with ranges reported in prior studies of Veteran populations, supporting the plausibility of our findings [50-52]. Still, here we privilege the calibrated estimate tied to observed human review.

Second, the main human-human discrepancies were both definitional and false negatives from information overload. The largest contributor to divergence was the treatment of "punching objects" (eg, walls/doors) without stated self-harm intent. One reviewer initially counted many such episodes as nonsuicidal self-injury (NSSI) (preconsensus count 45), whereas other reviewers generally did not—applying an intent or foreseeability standard. Borderline cases (eg, punching a window with tendon laceration) were debated under a reasonable foreseeability framework, absent self-harm intent. Similarly, statements about long-horizon self-destruction (eg, "drinking myself to death") were generally classified as substance use disorder unless the episode reflected acute, explicit self-harm intent (eg, deliberate overdose). Under an inclusive NSSI sensitivity definition that includes "punching objects" [53,54], the consensus chart review identified 39 positives. This is 9 fewer than PULSNAR's estimate of 48, which is expected and desirable, as PULSNAR is designed to provide an upper bound on the proportion of positives among unlabeled cases. Notably, 28.2% of positives had at least one reviewer missing chart evidence—a real concern, as physicians have limited time to comb through notes,

XSL•FO
RenderX

unlike our reviewers who used systematic search tools and spent more time per chart than a typical visit duration.

Third, information asymmetries between data modalities, for example, notes versus codes, explain much of the remaining gap [55]. For example, manual reviewers who examined unstructured clinical notes identified several cases of self-harm history (typically past suicide attempts prior to Veteran enrollment) that were documented only in narrative notes but not captured in structured data fields. In contrast, PULSNAR relied exclusively on structured data and would have missed these cases. Conversely, some PULSNAR "high-risk" cases lacked explicit self-harm documentation in notes but exhibited risk constellations in coded data (injuries/poisonings, major mental illness, and substance use disorders). This pattern is expected given the information asymmetry between notes and codes, reinforcing that neither source is complete [55,56].

Contribution

We emphasize two contributions. (1) Epidemiologic impact: After calibration, PULSNAR yields a population-scale estimate of ever self-harm that materially exceeds code-based prevalence. Notably, prior Veteran studies report ever NSSI between ~6% and 16%, including samples with rates of 14% - 16% [50-52], providing an external range against which our calibrated estimate can be interpreted. (2) Operational utility: PULSNAR enables triage of >1.3 million records, focusing human effort (and/or targeted NLP) where the marginal value of chart review is highest. This offers a blueprint for identifying other undercoded mental health diagnoses. In our experience, expert review of 97 charts required substantial person-hours, whereas model inference over the full cohort is computationally tractable; thus, even moderate accuracy can generate large efficiency gains by shrinking the manual search space.

Our findings support a hybrid workflow generalizable to other undetected conditions: (1) run PULSNAR on structured data to estimate the probability of being a positive case of self-harm (or other target phenotype) for uncoded individuals, (2) select a limited number of uncoded individuals from each probability bin of PULSNAR-estimated probabilities, (3) perform chart review for those selected uncoded individuals, and (4) apply post hoc calibration using chart-reviewed labels and PULSNAR-estimated probabilities of uncoded individuals to obtain calibrated cohort estimates with uncertainty. These steps, collectively, convert a complex validation into actionable epidemiology and a scalable detection pathway for underdocumented self-harm or other mental illnesses.

Limitations

First, this study used a single data source, VHA EHRs, whose patient population, coding practices, and care patterns differ systematically from those of other US and international health systems. Accordingly, the generalizability of our findings to Veterans treated exclusively outside the VHA or to non-Veteran populations needs validation. Second, since the true α is generally unidentifiable [57], PULSNAR estimates an upper bound on α across different positive subtypes. Thus, the corresponding predicted probabilities may be overestimated. Third, the prevalence estimate relies on the assumption that the post hoc calibration factor derived from a small, stratified sample of 97 chart-reviewed cases accurately transfers to the entire unlabeled population of 1.3 million Veterans, a potential source of error we partially addressed with bootstrap resampling to quantify uncertainty. Finally, while we validated this method using chart review for self-harm, which is often documented in notes but uncoded, this approach cannot confirm cases absent from both notes and codes. Furthermore, other conditions may follow different recording mechanisms; for instance, posttraumatic stress disorder in the VHA is typically coded if documented due to disability benefit incentives. Consequently, extending this framework to such diagnoses may require validating against the incidence of future coded diagnoses rather than concurrent notes. Because EHR data do not contain reliable true negative labels for uncoded self-harm, conventional supervised classification metrics such as AUC-ROC, precision-recall, and F_1-scores cannot be unbiasedly estimated [58]. This limitation is common in PU learning settings where only confirmed positives are available [24]. The chart-reviewed subset comprised only previously unlabeled cases and was not selected as a representative validation cohort with confirmed positives and negatives. Accordingly, model evaluation focused on prevalence estimation and agreement with expert review rather than traditional supervised classification metrics.

Finally, our findings should be interpreted in the context of broader limitations of suicide risk surveillance. VHA policy and clinical operations prioritize suicidal self-directed violence within the past 12 months and rely on policy-driven documentation tools such as the SBOR [40,59], whereas our phenotyping targets ever self-harm based primarily on *ICD* coding. Self-directed violence nomenclature and intent-based coding rules are complex [60], and there is no required, standardized training for VHA or non-VHA clinicians on their application [61]; as a result, documentation and coding of self-harm and suicidal behavior are inconsistent across settings and clinical providers. These surveillance and training constraints likely contribute to undercoding and misclassification and should be considered when interpreting our prevalence estimates and model performance.

Future Directions

Future research could validate our PU learning algorithms in other populations and health care settings and across a broader range of mental health diagnoses. Incorporating unstructured data, such as clinical notes, through NLP techniques, may further enhance the detection of uncoded self-harm instances and develop valid and reliable scales to measure self-harm. Additionally, integrating our approach into clinical workflows could facilitate real-time identification of at-risk individuals, enabling timely intervention.

Conclusions

Our study demonstrates the effectiveness of PU learning algorithms under the SNAR assumption in identifying uncoded instances of self-harm among US Veterans. PULSNAR can support both population-level prevalence estimation and individual-level risk stratification using structured health data, although differences between structured billing codes and clinical notes may affect concordance with manual chart

XSL•FO
RenderX

reviews. Our findings reveal a significantly higher prevalence of self-harm than what is captured in diagnostic codes, emphasizing the urgent need for more accurate detection and imputation methods. Our approach offers a scalable and efficient adjunct to manual chart reviews for detecting undetected mental illness diagnoses, with the potential to enhance clinical practice, inform policy decisions, support comparative effectiveness studies with imputed phenotypes [13], improve predictive modeling of self-harm and other conditions, and ultimately contribute to reducing suicide rates among Veterans and improving mental health. We encourage the adoption of similar methods in other health care systems to address undercoding challenges, improve patient outcomes, and advance the application of ML in health care analytics.

Acknowledgments

We gratefully acknowledge Edgar J Villarreal, PhD, and his team from the VA Office of Suicide Prevention for their invaluable guidance on the nuances of Veterans Health Administration policy and operational practices, which substantially strengthened the contextual clarity of this work. Generative artificial intelligence tools were not used at any stage of this study, including study design, methodology development, literature review, code generation, or manuscript writing and editing. However, Grammarly was used as a spell and grammar checker to improve the clarity of the manuscript.

Funding

This research was supported by funding from the US National Institutes of Health, specifically, the National Institute of Mental Health grant R01MH129764, the National Library of Medicine grant R00LM013367, and infrastructure support from the National Center for Advancing Translational Sciences grants UL1TR001449 and UM1TR005466. The views expressed in this paper are those of the authors and do not necessarily reflect those of the National Institutes of Health. Access to VA data and computational environments provided by the Department of Veterans Affairs (VA) Informatics and Computing Infrastructure (VINCI), funded under the research priority to Put VA Data to Work for Veterans (VA ORD 22-D4V).

Conflicts of Interest

None declared.

Multimedia Appendix 1
ICD-CM codes and Cohen κ coefficients. *ICD-CM*: *International Classification of Diseases, Clinical Modification*.
[DOCX File, 1753 KB - jmir_v28i1e89071_app1.docx]

References

1. Facts about suicide. Centers for Disease Control and Prevention. 2025 Mar 26. URL: https://www.cdc.gov/suicide/facts/index.html [accessed 2026-05-09]
2. 10 leading causes of death, United States, 2024. Centers for Disease Control and Prevention. URL: https://wisqars.cdc.gov/lcd/ [accessed 2026-05-09]
3. 2022 National Survey on Drug Use and Health (NSDUH) Annual National Report. : Substance Abuse and Mental Health Services Administration URL: https://www.samhsa.gov/data/report/2023-nsduh-annual-national-report [accessed 2026-05-09]
4. Ruiz F, Burgo-Black L, Hunt SC, Miller M, Spelman JF. A practical review of suicide among veterans: preventive and proactive measures for health care institutions and providers. Public Health Rep 2023;138(2):223-231. [doi: 10.1177/00333549221085240] [Medline: 35403486]
5. 2024 National Veteran Suicide Prevention Annual Report. : U.S. Department of Veterans Affairs; 2024 Dec URL: https://www.mentalhealth.va.gov/docs/data-sheets/2024/2024-Annual-Report-Part-2-of-2_508.pdf [accessed 2026-05-09]
6. Schafer KM, Duffy M, Kennedy G, et al. Suicidal ideation, suicide attempts, and suicide death among veterans and service members: a comprehensive meta-analysis of risk factors. Mil Psychol 2022;34(2):129-146. [doi: 10.1080/08995605.2021.1976544] [Medline: 38536290]
7. Ribeiro JD, Franklin JC, Fox KR, et al. Self-injurious thoughts and behaviors as risk factors for future suicide ideation, attempts, and death: a meta-analysis of longitudinal studies. Psychol Med 2016 Jan;46(2):225-236. [doi: 10.1017/S0033291715001804] [Medline: 26370729]
8. Predescu E, Sipos R. Self-harm behaviors, suicide attempts, and suicidal ideation in a clinical sample of children and adolescents with psychiatric disorders. Children (Basel) 2023 Apr 14;10(4):725. [doi: 10.3390/children10040725] [Medline: 37189974]
9. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in electronic phenotyping: from rule-based definitions to machine learning models. Annu Rev Biomed Data Sci 2018 Jul;1:53-68. [doi: 10.1146/annurev-biodatasci-080917-013315] [Medline: 31218278]
10. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J Am Med Inform Assoc 2013 Dec;20(e2):e206-e211. [doi: 10.1136/amiajnl-2013-002428] [Medline: 24302669]

XSL•FO
RenderX

11. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc 2018 Oct 1;25(10):1419-1428. [doi: 10.1093/jamia/ocy068] [Medline: 29893864]
12. Kumar P, Nestsiarovich A, Nelson SJ, Kerner B, Perkins DJ, Lambert CG. Imputation and characterization of uncoded self-harm in major mental illness using machine learning. J Am Med Inform Assoc 2020 Jan 1;27(1):136-146. [doi: 10.1093/jamia/ocz173] [Medline: 31651956]
13. Nestsiarovich A, Kumar P, Lauve NR, et al. Using machine learning imputed outcomes to assess drug-dependent risk of self-harm in patients with bipolar disorder: a comparative effectiveness study. JMIR Ment Health 2021 Apr 21;8(4):e24522. [doi: 10.2196/24522] [Medline: 33688834]
14. Hedegaard H, Schoenbaum M, Claassen C, Crosby A, Holland K, Proescholdbell S. Issues in developing a surveillance case definition for nonfatal suicide attempt and intentional self-harm using International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) coded data. Natl Health Stat Report 2018 Feb(108):1-19. [Medline: 29616901]
15. Edgcomb JB, Thiruvalluru R, Pathak J, Brooks JO. Machine learning to differentiate risk of suicide attempt and self-harm after general medical hospitalization of women with mental illness. Med Care 2021 Feb 1;59:S58-S64. [doi: 10.1097/MLR.0000000000001467] [Medline: 33438884]
16. Simon GE, Shortreed SM, Johnson E, et al. Predicting risk of suicidal behavior from insurance claims data vs. linked data from insurance claims and electronic health records. Pharmacoepidemiol Drug Saf 2024 Jan;33(1):e5734. [doi: 10.1002/pds.5734] [Medline: 38112287]
17. Simon GE, Johnson E, Lawrence JM, et al. Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. Am J Psychiatry 2018 Oct 1;175(10):951-960. [doi: 10.1176/appi.ajp.2018.17101167] [Medline: 29792051]
18. Rozova V, Witt K, Robinson J, Li Y, Verspoor K. Detection of self-harm and suicidal ideation in emergency department triage notes. J Am Med Inform Assoc 2022 Jan 29;29(3):472-480. [doi: 10.1093/jamia/ocab261] [Medline: 34897466]
19. Walsh CG, Ribeiro JD, Franklin JC. Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. J Child Psychol Psychiatry 2018 Dec;59(12):1261-1270. [doi: 10.1111/jcpp.12916] [Medline: 29709069]
20. Tsui FR, Shi L, Ruiz V, et al. Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. JAMIA Open 2021 Jan;4(1):ooab011. [doi: 10.1093/jamiaopen/ooab011] [Medline: 33758800]
21. Su C, Aseltine R, Doshi R, Chen K, Rogers SC, Wang F. Machine learning for suicide risk prediction in children and adolescents with electronic health records. Transl Psychiatry 2020 Nov 26;10(1):413. [doi: 10.1038/s41398-020-01100-0] [Medline: 33243979]
22. Barak-Corren Y, Castro VM, Javitt S, et al. Predicting suicidal behavior from longitudinal electronic health records. Am J Psychiatry 2017 Feb 1;174(2):154-162. [doi: 10.1176/appi.ajp.2016.16010077] [Medline: 27609239]
23. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: Association for Computing Machinery; 2016:785-794. [doi: 10.1145/2939672.2939785]
24. Kumar P, Lambert CG. Positive unlabeled learning selected not at random (PULSNAR): class proportion estimation without the selected completely at random assumption. PeerJ Comput Sci 2024;10:e2451. [doi: 10.7717/peerj-cs.2451] [Medline: 39650456]
25. Jaskie K, Elkan C, Spanias A. A modified logistic regression for positive and unlabeled learning. In: 2019 53rd Asilomar Conference on Signals, Systems, and Computers: IEEE; 2019. [doi: 10.1109/IEEECONF44664.2019.9048765]
26. Elkan C, Noto K. Learning classifiers from only positive and unlabeled data. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: Association for Computing Machinery; 2008:213-220. [doi: 10.1145/1401890.1401920]
27. Du Plessis MC, Sugiyama M. Class prior estimation from positive and unlabeled data. IEICE Trans Inf Syst 2014;E97.D(5):1358-1362. [doi: 10.1587/transinf.E97.D.1358]
28. Ramaswamy H, Scott C, Tewari A. Mixture proportion estimation via kernel embeddings of distributions. In: Proceedings of the 33rd International Conference on Machine Learning: JMLR; 2016:2052-2060. [doi: 10.5555/3045390.3045607]
29. Bekker J, Davis J. Estimating the class prior in positive and unlabeled data through decision tree induction. Proc AAAI Conf Artif Intell 2018;32(1). [doi: 10.1609/aaai.v32i1.11715]
30. Ivanov D. DEDPUL: difference-of-estimated-densities-based positive-unlabeled learning. In: 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA): IEEE; 2020:782-790. [doi: 10.1109/ICMLA51294.2020.00128]
31. Gerych W, Hartvigsen T, Buquicchio L, Agu E, Rundensteiner E. Recovering the propensity score from biased positive unlabeled data. Proc AAAI Conf Artifi Intell 2022;36(6):6694-6702. [doi: 10.1609/aaai.v36i6.20624]
32. Bekker J, Robberechts P, Davis J. Beyond the selected completely at random assumption for learning from positive and unlabeled data. In: Machine Learning and Knowledge Discovery in Databases: Springer; 2020:71-85. [doi: 10.1007/978-3-030-46147-8_5]
33. Kumar P, Moomtaheen F, Malec SA, et al. Detecting opioid use disorder in health claims data with positive unlabeled learning. IEEE J Biomed Health Inform 2025 Feb;29(2):750-757. [doi: 10.1109/JBHI.2024.3515805] [Medline: 40030473]

XSL•FO
RenderX

34. Pozzolo AD, Caelen O, Johnson RA, Bontempi G. Calibrating probability with undersampling for unbalanced classification. In: 2015 IEEE Symposium Series on Computational Intelligence (SSCI): IEEE; 2015:159-166. [doi: 10.1109/SSCI.2015.33]
35. Philip K, Chan SJ. Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection. In: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining: AAAI Press; 1998:164-168. [doi: 10.5555/3000292.3000320]
36. Voss EA, Makadia R, Matcho A, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc 2015 May;22(3):553-564. [doi: 10.1093/jamia/ocu023] [Medline: 25670757]
37. Lynch KE, Deppen SA, DuVall SL, et al. Incrementally transforming electronic medical records into the observational medical outcomes partnership common data model: a multidimensional quality assurance approach. Appl Clin Inform 2019 Oct;10(5):794-803. [doi: 10.1055/s-0039-1697598] [Medline: 31645076]
38. Waller G, Newbury-Birch D, Simpson D, et al. The barriers and facilitators to the reporting and recording of self-harm in young people aged 18 and under: a systematic review. BMC Public Health 2023 Jan 24;23(1):158. [doi: 10.1186/s12889-023-15046-7] [Medline: 36694149]
39. Dizon MP, Chow A, Ong MK, et al. Lower comorbidity scores and severity levels in Veterans Health Administration hospitals: a cross-sectional study. BMC Health Serv Res 2024 May 8;24(1):601. [doi: 10.1186/s12913-024-11063-3] [Medline: 38714970]
40. Gujral K, Bahraini N, Brenner LA, et al. VA's implementation of universal screening and evaluation for the suicide risk identification program in November 2020 -Implications for Veterans with prior mental health needs. PLoS ONE 2023;18(4):e0283633. [doi: 10.1371/journal.pone.0283633] [Medline: 37040367]
41. Meda N, Angelozzi L, Poletto M, et al. How many people die by suicide each year? Not 727,000: a systematic review and meta-analysis of suicide underreporting across 71 countries over 122 years. Front Psychiatry 2025;16:1609580. [doi: 10.3389/fpsyt.2025.1609580] [Medline: 40873674]
42. Getzen E, Ungar L, Mowery D, Jiang X, Long Q. Mining for equitable health: assessing the impact of missing data in electronic health records. J Biomed Inform 2023 Mar;139:104269. [doi: 10.1016/j.jbi.2022.104269] [Medline: 36621750]
43. McKenzie J, Rajapakshe R, Shen H, Rajapakshe S, Lin A. A semiautomated chart review for assessing the development of radiation pneumonitis using natural language processing: diagnostic accuracy and feasibility study. JMIR Med Inform 2021 Nov 12;9(11):e29241. [doi: 10.2196/29241] [Medline: 34766919]
44. Xie F, Ling Grant DS, Chang J, Amundsen BI, Hechter RC. Identifying suicidal ideation and attempt from clinical notes within a large integrated health care system. Perm J 2022 Apr 5;26(1):85-93. [doi: 10.7812/TPP/21.102] [Medline: 35609162]
45. Teismann T, Eimen JJ, Cwik JC. Misclassification of self-directed violence. Crisis 2023 Nov;44(6):525-528. [doi: 10.1027/0227-5910/a000897] [Medline: 36636794]
46. Simon GE, Shortreed SM, Boggs JM, et al. Accuracy of ICD-10-CM encounter diagnoses from health records for identifying self-harm events. J Am Med Inform Assoc 2022 Nov 14;29(12):2023-2031. [doi: 10.1093/jamia/ocac144] [Medline: 36018725]
47. Arndt BG, Micek MA, Rule A, Shafer CM, Baltus JJ, Sinsky CA. More tethered to the EHR: EHR workload trends among academic primary care physicians, 2019-2023. Ann Fam Med 2024;22(1):12-18. [doi: 10.1370/afm.3047] [Medline: 38253499]
48. Davis M, Siegel J, Becker-Haimes EM, et al. Identifying common and unique barriers and facilitators to implementing evidence-based practices for suicide prevention across primary care and specialty mental health settings. Arch Suicide Res 2023;27(2):192-214. [doi: 10.1080/13811118.2021.1982094] [Medline: 34651544]
49. LeCloux M, Aguinaldo LD, Lanzillo EC, Horowitz LM. PCP opinions of universal suicide risk screening in rural primary care: current challenges and strategies for successful implementation. J Rural Health 2021 Jun;37(3):554-564. [doi: 10.1111/jrh.12508] [Medline: 32845543]
50. Halverson TF, Mann AJD, Zelkowitz RL, et al. Nonsuicidal self-injury in veterans: prevalence, clinical characteristics, and gender differences from a national cohort. Psychiatry Res 2022 Sep;315:114708. [doi: 10.1016/j.psychres.2022.114708] [Medline: 35868073]
51. Gromatsky M, Halverson TF, Dillon KH, et al. The prevalence of nonsuicidal self-injury in military personnel: a systematic review and meta-analysis. Trauma Violence Abuse 2023 Dec;24(5):2936-2952. [doi: 10.1177/15248380221119513] [Medline: 36062896]
52. Bryan C, Bryan A. Nonsuicidal self-injury among a sample of United States military personnel and veterans enrolled in college classes. J Clin Psychol 2014 Sep;70(9):874-885. [doi: 10.1002/jclp.22075] [Medline: 24619940]
53. Kimbrel NA, Thomas SP, Hicks TA, et al. Wall/object punching: an important but under-recognized form of nonsuicidal self-injury. Suicide Life Threat Behav 2018 Oct;48(5):501-511. [doi: 10.1111/sltb.12371] [Medline: 28925016]
54. Patel TA, Dillon KH, Cassiello-Robbins C, Calhoun PS, Beckham JC, Kimbrel NA. Anger, impulsivity and wall/object punching in a sample of U.S. veterans with psychiatric disorders. J Psychiatr Res 2022 Mar;147:269-273. [doi: 10.1016/j.jpsychires.2022.01.036] [Medline: 35074743]

XSL•FO
RenderX

55. Nguyen CT, Richman JS, Chiles JW, Singh JA, McDonald MLN. Osteoarthritis case identification in the Million Veteran Program cohort: comparison of diagnostic codes versus clinical notes. Osteoarthr Cartil Open 2026 Jun;8(2):100758. [doi: 10.1016/j.ocarto.2026.100758] [Medline: 41799924]
56. Moldwin A, Demner-Fushman D, Goodwin TR. Empirical findings on the role of structured data, unstructured data, and their combination for automatic clinical phenotyping. AMIA Jt Summits Transl Sci Proc 2021;2021:445-454. [Medline: 34457160]
57. Blanchard G, Lee G, Scott C. Semi-supervised novelty detection. J Mach Learn Res 2010 Dec 1;11:2973-3009. [doi: 10.5555/1756006.1953028]
58. Zou R, D Williamson B, M Shortreed S, Coley RY. Validation of a risk‐prediction model in the presence of outcome misclassification. Stat Med 2026 Apr;45(8-9):e70377. [doi: 10.1002/sim.70377]
59. Felleman BI, Doran NM, Asamsama OH, Oliva EM, Han BH. New methodology to improve tracking of veteran overdose deaths and characterization of a population of veteran overdose decedents in San Diego County. Drug Alcohol Depend Rep 2025 Dec;17:100392. [doi: 10.1016/j.dadr.2025.100392] [Medline: 41322677]
60. Brenner LA, Breshears RE, Betthauser LM, et al. Implementation of a suicide nomenclature within two VA healthcare settings. J Clin Psychol Med Settings 2011 Jun;18(2):116-128. [doi: 10.1007/s10880-011-9240-9] [Medline: 21626353]
61. Inadequate staff training and lack of oversight contribute to the Veterans Health Administration's suicide risk screening and evaluation deficiencies. : United States Department of Veterans Affairs; 2024 Dec 18 URL: https://www.vaoig.gov/sites/default/files/reports/2024-12/vaoig-23-02939-13.pdf [accessed 2026-05-11]

Abbreviations

AUC: area under the curve
CSR: compressed sparse row
ED: emergency department
EHR: electronic health record
ICD: *International Classification of Diseases*
LASSO: least absolute shrinkage and selection operator
ML: machine learning
NLP: natural language processing
NSSI: nonsuicidal self-injury
OMOP: Observational Medical Outcomes Partnership
PU: positive and unlabeled
PULSCAR: positive unlabeled learning selected completely at random
PULSNAR: positive unlabeled learning selected not at random
SBOR: Suicide Behavior and Overdose Report
SCAR: selected completely at random
SNAR: selected not at random
SNOMED: Systematized Nomenclature of Medicine
VHA: Veterans Health Administration
XGBoost: Extreme Gradient Boosting

Edited by S Brini; submitted 05.Dec.2025; peer-reviewed by A Tai, S Nelson; revised version received 17.Apr.2026; accepted 19.Apr.2026; published 04.Jun.2026.

Please cite as:
Kumar P, Viszolay AD, Upadhayaya R, Moomtaheen F, Greer DR, Bologa CG, Schneider KA, Davis SE, Matheny ME, van der Goes D, Villarreal G, Zhu Y, Tohen M, Malec SA, Yang JJ, Fielstein EM, Lambert CG
Detecting Uncoded Self-Harm in Veterans' Electronic Health Records Using Positive and Unlabeled Learning: Retrospective Cohort Study
J Med Internet Res 2026;28:e89071
URL: https://www.jmir.org/2026/1/e89071
doi:10.2196/89071

© Praveen Kumar, Alexandria D Viszolay, Rajesh Upadhayaya, Fariha Moomtaheen, Donald R Greer, Cristian G Bologa, Kristan A Schneider, Sharon E Davis, Michael E Matheny, David van der Goes, Gerardo Villarreal, Yiliang Zhu, Mauricio Tohen, Scott A Malec, Jeremy J Yang, Elliot M Fielstein, Christophe Gerard Lambert. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 4.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons

XSL•FO
RenderX

Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Addressing Data Quality Challenges in Lung Cancer Data Within the Observational Medical Outcomes Partnership Common Data Model: Observational Study

Jens Declerck[1,2], MSc; Mieke Deschepper[3], PhD; Kirsten Colpaert[3], Prof Dr; Dipak Kalra[1,2], Dr med; Pascal Coorevits[1], Prof Dr

[1]Department of Public Health and Primary Care, Ghent University, Unit of Medical Informatics and Statistics, Corneel Heymanslaan 10, Ghent, Belgium
[2]The European Institute for Innovation through Health Data, Ghent, Belgium
[3]Ghent University Hospital, Data Science Institute, Ghent, Belgium

Corresponding Author:
Jens Declerck, MSc
Department of Public Health and Primary Care, Ghent University, Unit of Medical Informatics and Statistics, Corneel Heymanslaan 10, Ghent, Belgium

Abstract

Background: The secondary use of health data is essential for advancing medical research and improving clinical practice. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) enables large-scale, multicenter studies but faces challenges related to consistency, completeness, and transparency during data mapping from original data sources.

Objective: This study aimed to evaluate the quality of the mapping process for lung cancer data within the Federated Health Innovation Network project, with a focus on consistency, completeness, and challenges encountered throughout the process.

Methods: Clinical data from Ghent University Hospital were mapped to the OMOP CDM using a reference data dictionary. Consistency was assessed using Cohen kappa (κ) scores, while completeness was evaluated by comparing patient and record counts before and after mapping. Challenges, including unstructured data and an evolving reference standard, were documented and analyzed.

Results: High consistency was observed for structured variables, while some unstructured variables, such as "Smoking status," were excluded due to their free-text format and the lack of suitable OMOP concepts. The completeness analysis showed minimal data loss for most structured variables but highlighted substantial challenges associated with unstructured data. Persistent issues included evolving data dictionary versions and mismatches in diagnostic code granularity between institutions, underscoring structural challenges in standardization.

Conclusions: The transformation of lung cancer data to the OMOP CDM highlighted both technical and systemic challenges, including the handling of unstructured data and the resolution of granularity discrepancies. A multidisciplinary approach involving clinical and technical expertise is crucial for ensuring reliable, high-quality datasets for multicenter research.

(J Med Internet Res 2026;28:e90246) doi:10.2196/90246

KEYWORDS

health data quality; Observational Medical Outcomes Partnership Common Data Model; OMOP CDM; primary use; secondary use; extract, transform, and load; ETL

Introduction

The secondary use of health data—leveraging existing health information for purposes beyond direct patient care—has become a cornerstone for advancing medical research [1,2], developing health care policies [3,4], and improving clinical practices [5]. By integrating health data from diverse clinical settings, researchers can uncover valuable insights into disease patterns [6], treatment outcomes [7], and health care processes [8]. This approach is particularly crucial for studying rare diseases or uncommon clinical events, in which data from a single source are often insufficient [9]. The power of large-scale, multicenter datasets lies in their ability to address complex research questions, but this potential can only be fully realized if data quality is ensured throughout the entire data lifecycle—from primary data capture to transformation and integration into a standardized framework for secondary use [10,11].

Data quality remains one of the most substantial challenges in the effective secondary use of health data [11,12]. Poor-quality data can lead to incorrect research findings [13], poor clinical decision-making [14], and misguided health care policies [15].

XSL•FO
RenderX

Data quality is influenced by multiple factors, including the reliability of the primary data sources, the transformation processes used to standardize data, and the quality of the resulting secondary datasets [11]. The extract, transform, and load (ETL) process plays a critical role, as it involves consolidating, standardizing, and integrating data from multiple sources. Each stage of the ETL process presents unique challenges and risks to data quality. Errors at any stage can compromise the usability and reliability of the final dataset, leading to potential misinterpretations in downstream analyses [16,17].

Although frameworks addressing data quality in primary and secondary datasets are well established [11,18-22], mapping clinical data from different sources into a standardized model such as the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) remains prone to challenges [4]. The OMOP CDM is a standardized framework for structuring and analyzing health care data from diverse sources, such as electronic health records (EHRs). By adopting uniform data structures and standardized terminologies, such as Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT), Logical Observation Identifiers Names and Codes (LOINC), and *International Classification of Diseases, 10th Revision* (*ICD-10*), the OMOP CDM facilitates interoperability, collaborative research, and large-scale data analysis. OMOP provides a robust framework for integrating diverse datasets by standardizing both data structure and terminologies [3], thereby enabling multicenter research on treatment outcomes [23] and health care delivery [24]. However, variability in data extraction and mapping practices can introduce inconsistencies, thereby affecting the reliability and reproducibility of research findings [25].

Despite the increasing adoption of the OMOP CDM and the growing usability of secondary datasets [26], there is limited guidance on how to systematically evaluate mapping quality or address discrepancies in granularity and completeness [27]. This lack of structured approaches for assessing the transformation process creates barriers to new implementations, particularly in multicenter settings.

This study sought to address these gaps by focusing on the quality of the mapping process during the implementation of the OMOP CDM for lung cancer data. This effort was part of the Federated Health Innovation Network (FHIN) project, an open-source collaboration among Belgian hospitals to develop a fully automated, federated platform aimed at addressing research questions in the field of lung cancer [28]. As part of this project, a data dictionary was provided, detailing the mapping of raw data to OMOP CDM concepts. This dictionary, which outlined one-to-one relationships between raw data elements and OMOP CDM concept IDs, served as a reference standard.

Specifically, this study aimed to explore strategies for preserving data quality during the process of mapping data to the OMOP CDM. The primary objective was to evaluate the quality of the mapping process by examining the completeness and consistency during the mapping process. The secondary objective was to identify the challenges and complexities encountered during the implementation of the OMOP CDM and to develop a practical framework to guide future OMOP implementations.

Methods

Study Design and Setting

This study was independently conducted by the Data Science Institute (DSI) of Ghent University Hospital and the European Institute for Innovation through Health Data to ensure transparency and traceability of the mapping process of clinical data into the OMOP CDM. This study evaluated the mapping of clinical data to the OMOP CDM within the context of lung cancer data integration, with a focus on reproducibility and data quality assessment. Although not part of the FHIN project, this study aligns with its goals by ensuring rigorous documentation and standardization of the mapping process for lung cancer data. The study design emphasizes reproducibility and adaptability to similar multicenter initiatives.

Ethical Considerations

This study did not undergo a formal institutional review board or research ethics committee assessment because it was based on fully anonymized data and did not involve direct interaction with human participants. No identifiable personal data were accessed, and all data were handled in compliance with applicable data protection regulations.

Reference Standard Provided by the FHIN Project

As part of the FHIN project, a data dictionary was provided that defined key data items relevant to lung cancer and their mapping to specific OMOP concept IDs. This dictionary served as the reference standard for evaluating the consistency of our mapping process. Examples of some of the variables included in the data dictionary are provided in Table 1. However, the dictionary lacked critical details, including the original data sources for the variables, the extraction logic (eg, identification of the relevant tables and fields for each data element and transformation of field values to the standard terminology relevant in OMOP), and the rationale behind assigning specific concept IDs. Additionally, the data dictionary evolved throughout the project, reflecting adjustments made as part of the data quality process. These factors complicated efforts to fully replicate the mapping process, potentially introducing variability and bias. For transparency, we based our evaluation on the version of the data dictionary available as of December 1, 2024.

XSL•FO
RenderX

Table . Example of the data dictionary.

Concept ID	Concept name	Vocabulary ID	Concept code	Observational Medical Outcomes Partnership table
44790293	Radiotherapy delivery	SNOMED CT[a]	231711000000108	PROCEDURE
40483776	Total radiation dose delivered	SNOMED CT	445461008	MEASUREMENT
4155148	Delivered radiation dose	SNOMED CT	371892002	MEASUREMENT

[a]SNOMED CT: Systematized Nomenclature of Medicine–Clinical Terms.

Data Quality Assurance

Data quality assurance was performed to evaluate completeness and consistency. Completeness was assessed by comparing the total number of patients and records extracted from the raw data sources with those successfully transformed into OMOP CDM tables. Consistency assessment was performed to evaluate the agreement between the OMOP concept IDs assigned during the ETL process and those specified in the reference data dictionary. For each data category, which typically included multiple variables, Cohen κ scores [29] were calculated at the variable level. Agreement was defined as an exact match between the concept ID assigned during mapping and the expected concept ID in the dictionary. Variables that were unmapped, mismatched, or lacked a valid concept ID were considered disagreements. To report a single consistency score per category, the final κ score was calculated as the unweighted average of the individual κ scores of all variables within that category, as a descriptive summary measure across heterogeneous variables [30]. To capture within-category variability, the SDs of the κ scores and the number of variables per category were additionally calculated and reported. This approach allowed a balanced assessment across categories, independent of variable count or complexity, and helped identify specific areas of misalignment in the mapping process.

Data Sources and Extraction

On the basis of the variables defined in the data dictionary, all relevant data items were extracted from the data sources at Ghent University Hospital. The extracted data included records from patients between January 1, 2016, and December 31, 2023. The extraction process involved developing and executing SQL queries to retrieve the specified variables from various hospital databases. These data sources included the DSI–Data Warehouse (DWH); Multidisciplinary Oncology Consultation (MOC) application; *Minimale Ziekenhuisgegevens* (MZG); General Laboratory Information Management System; and admission, discharge, and transfer systems, each containing data relevant to the mapping process. The DSI-DWH is a curated database where vital parameters, such as weight and height, are stored. Data with the same meaning (eg, weight) were extracted from different fields within the EHR, cleaned, and standardized. MOC application [31] provides essential histology and pathology information, particularly related to cancer cases. MZG [32] stores diagnostic codes in *ICD-10* format, which are key for mapping clinical diagnoses. The General Laboratory Information Management System contains analysis codes and laboratory values classified using LOINC. Finally, the admission, discharge, and transfer system contains administrative variables and additional patient characteristics.

The data were stored in a staging area with tables reflecting the structure of the source systems. This staging area enabled uniform querying across databases, ensuring that data were harmonized before the ETL process. By implementing a structured extraction workflow, errors were minimized, and traceability from source to target was ensured.

ETL Process

The ETL process was implemented to harmonize extracted data into OMOP CDM version 5.4. During extraction, raw data were stored in a centralized Microsoft SQL database for processing. Transformation involved automated and manual mappings to OMOP standards. Automated mappings were conducted with SQL scripts using the OMOP vocabularies. An example of the mapping from *ICD-10* to standard OMOP codes can be found in Multimedia Appendix 1. This process aligned source data with terminologies such as SNOMED CT, LOINC, and *ICD-10*. Manual mappings were facilitated by Keun [33], particularly for complex variables such as genetic mutations; tumor, node, and metastasis (TNM) staging; and World Health Organization functional scores. SQL scripts were developed to transform raw data into OMOP-compliant tables. These scripts ensured that variables were assigned to appropriate domains and that data transformations adhered to OMOP guidelines. Special attention was given to the handling of unstructured data, such as free-text variables, which posed challenges during mapping. The final load process was executed using the FHIN tool Rabbit-in-a-Blender [34], an ETL pipeline used to transform raw data into the OMOP CDM. Mapping approaches varied by data category, with both automated and manual strategies applied. Manual mapping was used for categories such as “Clinical TNM staging,” “Pathological TNM staging,” and “Genetic mutations.” Automated mapping was applied to standardized classifications, including “Histology,” “Laboratory tests,” and “Diagnosis.”

Results

Consistency

The mapping process involved 12 data categories necessary for transforming lung cancer data into the OMOP CDM. These categories included essential clinical, genetic, and demographic variables such as diagnostic codes, TNM staging, and genetic mutations (eg, Kirsten rat sarcoma and v-raf murine sarcoma viral oncogene homolog B). Mapping methods varied by category, using either manual processes requiring domain

XSL•FO
RenderX

expertise or automated methods for which established standards enabled straightforward mapping. Manual mapping was used for categories such as "Clinical TNM staging," "Pathological TNM staging," and "Genetic mutations," where specialized knowledge was critical to ensure accuracy. Automated mapping was applied to standardized classifications, including "Histology" (based on the International Classification of Diseases for Oncology 3rd edition classification), "Laboratory tests" (based on the LOINC classification), and "Diagnosis" (based on the *ICD-10* classification). Additional details can be found in Multimedia Appendix 1.

To evaluate mapping consistency, κ scores were calculated at the variable level within each data category by comparing assigned OMOP concept IDs with those defined in the reference data dictionary. For each category, the reported κ represents the unweighted mean of the variable-level κ values. In addition, the number of variables per category and the SD were calculated to reflect variability within categories. A mean score of 1 indicates perfect alignment across all variables in that category, while lower scores and higher SD values highlight categories in which mapping challenges or heterogeneity were more pronounced. The mean κ score, SDs, and number of variables per category are presented in (Table 2).

Table . Mean kappa (κ) scores, SDs, and number of variables per category according to the mapping process.

Mapping processes and categories	Variables, n (%)	Cohen κ score, mean (SD)
Automated mapping		
Diagnosis	19 (5.3)	0.842 (0.375)
Histology	26 (7.3)	1.000 (0)
Laboratory tests	185 (52)	0.968 (0.178)
Manual mapping		
Clinical TNM[a] staging	38 (10.7)	0.921 (0.273)
Gender	2 (0.6)	0 (0)
Genetic mutations	4 (1.1)	0.500 (0.577)
Pathological TNM staging	27 (7.6)	0.926 (0.267)
Smoking status	1 (0.3)	0 (N/A[b])
Therapy procedures	3 (0.8)	0.333 (0.577)
Unit	43 (12.1)	0.372 (0.489)
Value	2 (0.6)	1.000 (0)
World Health Organization score	6 (1.7)	0.833 (0.408)

[a]TNM: tumor, node, and metastasis.

[b]N/A: not available.

High consistency was observed in most categories. Categories such as "Value" and "Histology" showed perfect agreement (mean κ 1.000, SD 0), indicating fully consistent mappings across all variables within these categories. "Laboratory tests" and "Diagnosis" demonstrated strong agreement (mean κ 0.968, SD 0.178 and mean κ 0.842, SD 0.375, respectively), with variability observed at the variable level.

Both "Clinical TNM staging" and "Pathological TNM staging" showed high mean κ values (0.921 and 0.926, respectively) but with notable SDs (0.273 and 0.67, respectively), indicating variability across individual variables. Categories such as "Unit" (mean κ 0.372, SD 0.489) and "Therapy procedures" (mean κ 0.333, SD 0.577) exhibited lower consistency and substantial variability, driven by differences in unit recording practices and the absence of specific OMOP concept IDs for certain therapies (eg, immunotherapy).

"Gender" and "Smoking status" showed no agreement (κ=0). For "Smoking status," no variability measure could be calculated (n=1), reflecting complete mapping failure due to unstructured source data. For "Gender," the κ value of 0 was due to the absence of this variable in the reference standard. Consequently, no predefined mapping specification was available, leading to a mismatch between the implemented mapping and the expected reference standard. These findings indicate variability in the data dictionary and the lack of a structured representation of specific variables within the source systems.

Completeness

Completeness was assessed by comparing the number of records and patients before and after mapping. The initial data quality test revealed that only approximately half of the patients were successfully mapped. This low completeness rate was associated with discrepancies in *ICD-10* code mappings and inconsistencies across source systems. For example, some patients diagnosed with lung cancer appeared in the MZG data source but were missing from the MOC application data, or vice versa. These mismatches were accompanied by the temporary exclusion of affected records from the mapped dataset. An iterative data quality improvement process was applied. With each iteration, additional patients were reincluded. By the final iteration, only a single patient remained excluded due to an unresolved classification issue. This patient had relevant clinical data in the laboratory system but was categorized in the source data as

XSL•FO
RenderX

an outpatient consultation. As the ETL pipeline was configured to extract only hospitalized patients, this record could not be incorporated. Consequently, all but one patient were successfully included in the final dataset.

For most variables, patient and record counts remained complete in the final dataset, indicating a successful transformation. However, a few exceptions persisted. The variable "Smoking status" exhibited complete data loss because it was stored in an unstructured free-text field that combined information on smoking, drug abuse, and alcohol use. This format made it impossible to extract smoking-specific content for standardized mapping. Additionally, minor data loss was observed in the "Unit" and "Diagnosis" categories, which stemmed from inconsistencies in data representation or mapping complexity.

Challenges Encountered During the Mapping Process

Implementing the OMOP CDM revealed several structural and semantic challenges that complicated the mapping process. One considerable issue was the difference in data granularity between our hospital and the reference standard. For instance, in the condition table, the *ICD-10* code C34.1, which represents lung cancer of the upper lobe, was inconsistently mapped to different OMOP codes based on whether a 4-digit (C34.1) or 5-digit (C34.10) variation of the code was used. Although there is no clinical difference between codes C34.1 and C34.10, this inconsistency arose because the data dictionary only accounted for 4-digit codes, whereas the system at our hospital used a more granular collection process. Furthermore, SNOMED CT, the standard OMOP vocabulary for conditions, introduced additional complexity by mapping back to multiple codes for a single condition. This duality made maintaining consistency and alignment with the OMOP CDM challenging. This is presented in Figure 1.

Figure 1. Different Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT) codes derived from code C34.1. *ICD-10*: *International Classification of Diseases, 10th Revision*.

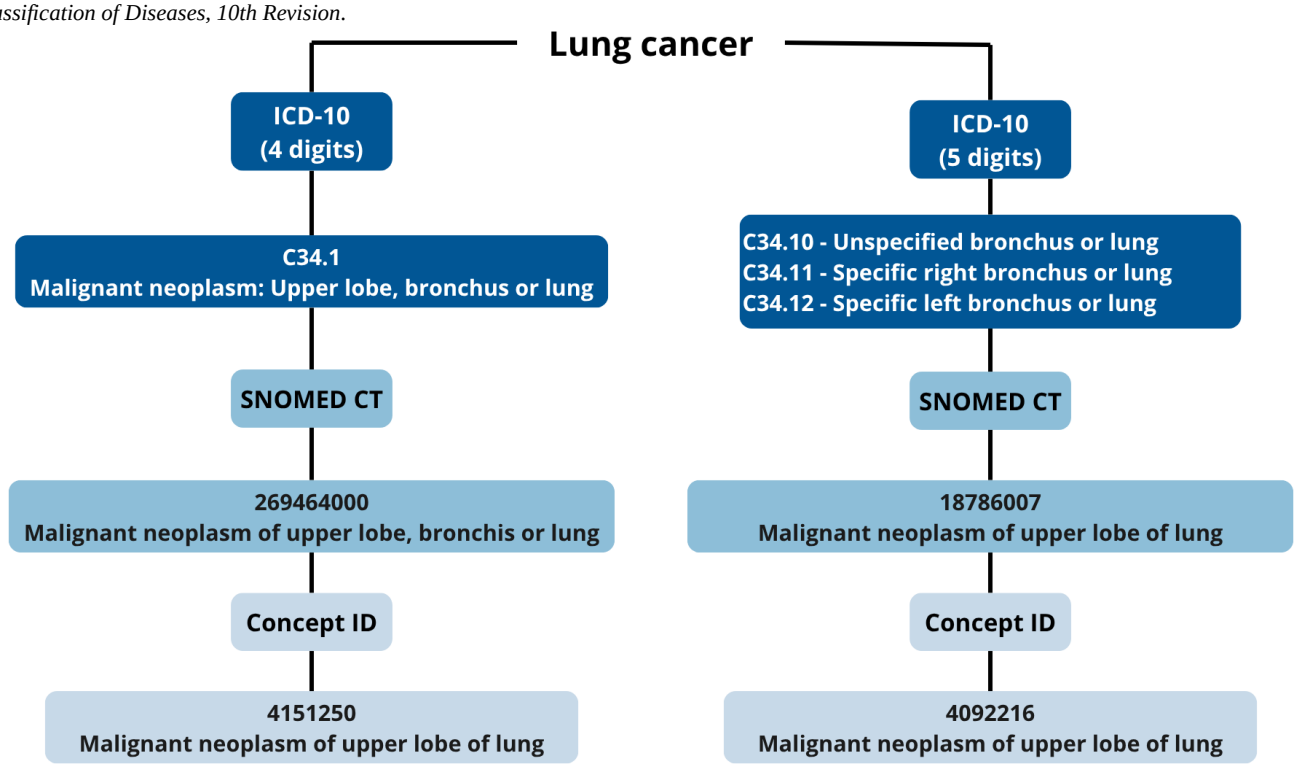

Another major complication was the reliance on free-text fields in the source data. For instance, the "Smoking status" variable was captured in an unstructured field in the EHR that combined drug abuse, smoking, and alcohol abuse into a single text box. Consequently, it was impossible to reliably determine whether the recorded information referred specifically to smoking, alcohol use, or drug abuse. This unstructured format prevented mapping to standardized OMOP concepts and led to the exclusion of these data during transformation. Notably, no neuro-linguistic programming techniques were used in this research, further limiting the ability to computationally extract and interpret such information. These issues were further exacerbated by the design of our hospital's EHR system, which is a home-grown platform historically optimized for clinical documentation rather than structured data capture. The current configuration of the EHR, with limited use of standardized fields, made the extraction and transformation process more challenging. Height and weight were initially derived from unstructured data fields in the EHR. However, a curation and parsing workflow was established, making these variables usable within the project (as stored in the DSI-DWH). A transition toward more structured data entry has recently been initiated, which is expected to facilitate future data standardization.

A more subtle but impactful challenge was the evolving state of the data dictionary during the project. As no finalized version was agreed upon at the project's start, the data dictionary continued to evolve, often introducing inconsistencies. For this study, we used the version of the dictionary established on December 1, 2024. However, several updates (such as the

XSL•FO
RenderX

change in the source of chemotherapy data from procedures to medication records) required periodic reassessment of our mappings. Although these changes were ultimately controlled for, they contributed to mapping delays and highlighted the need for stable definitions early in such projects.

Beyond data structure and documentation, the specialized knowledge required for OMOP CDM mapping proved to be a limiting factor. Mapping variables, resolving inconsistencies, and applying the correct logic demanded not only a solid understanding of the OMOP CDM but also clinical insight into the source data. The lack of domain expertise within the team sometimes caused delays, especially when clinical interpretation was needed to resolve ambiguous cases.

Finally, inconsistencies across the OMOP projects themselves presented challenges. Interactions with other OMOP initiatives revealed differences in data dictionaries and coding approaches. These differences included variations in variable definitions, mapping choices, and levels of coding granularity for similar clinical concepts. Consequently, alignment across projects required additional reconciliation efforts, increasing the complexity of the mapping process.

Discussion

Principal Findings

This study evaluated the quality of the mapping process of lung cancer data to the OMOP CDM, with a focus on completeness and consistency, and identified key challenges encountered during implementation. Overall, high consistency was achieved for structured variables, while unstructured data and variability in coding practices posed challenges. Completeness improved substantially through iterative data quality refinement, highlighting the importance of continuous validation during the ETL process. These findings align with the study objectives of assessing mapping quality and identifying barriers to effective OMOP implementation.

Data Quality Assessment

The findings demonstrated variability in consistency and completeness across categories. Variables with structured data and robust reference standards, such as “Histology,” “Value,” and “Laboratory tests,” achieved high κ scores with low variability, indicating stable and reproducible mappings across variables. In contrast, categories that required greater clinical interpretation or manual mapping, including “Clinical TNM staging,” “Pathological TNM staging,” and “Genetic mutations,” showed high κ values but substantial SDs, reflecting heterogeneous agreement at the variable level. This variability indicates that although overall mapping performance was strong, individual variables within these categories posed specific challenges.

A key contributor to this variability was differences in coding granularity. Variations in the level of detail captured in source systems, such as the use of 4-digit vs 5-digit *ICD-10* codes, introduced inconsistencies during mapping despite representing clinically equivalent concepts. This reflects a broader challenge in OMOP ETL processes, in which differences in coding specificity across institutions can affect semantic alignment and the consistency of standardized data.

Categories such as “Unit,” “Therapy procedures,” “Gender,” and “Smoking status” encountered challenges, reflecting the difficulties associated with ambiguous, incomplete, or unstructured data. These findings were consistent with previous studies, in which effective mappings are established for well-structured and standardized variables [35]. Unstructured variables, such as “Smoking status,” posed a particular challenge. Captured as free text in the EHR, this field combined multiple categories, making it impossible to extract smoking-specific information for mapping to OMOP concepts. These findings highlight a challenge in OMOP implementations related to the handling of unstructured data, which requires additional preprocessing before integration into the standardized model. Previous research has shown that free-text data often leads to data exclusion during OMOP transformation, limiting the accuracy of analyses reliant on such variables [3,4,36].

The completeness analysis revealed that structured data generally retained patient and record counts after mapping. For instance, variables such as “Histology” and “Radiotherapy” achieved nearly complete preservation of records. However, the absence of structured standards for certain categories led to minor data loss. For example, inconsistencies in the recording of units and granularity differences in diagnostic codes resulted in missing data during transformation. Although these losses were minimal, they highlight the need for enhanced preprocessing and harmonization workflows to mitigate discrepancies across source systems. These findings align with previous research that has identified similar challenges [4,35].

Challenges Encountered

The transformation process of mapping lung cancer data to the OMOP CDM highlights several challenges, encompassing both technical data issues and broader systemic and knowledge-related barriers. Although the technical aspects of the data, such as unstructured text and inconsistent coding practices, are well-recognized sources of data quality issues [37], this study demonstrates that the challenges extend beyond these technical constraints.

The major challenge encountered was related to the data dictionary. This data dictionary, provided by the reference hospital, offered a 1-to-1 mapping between raw data elements and OMOP concept IDs, serving as a useful starting point. However, 2 challenges emerged related to the data dictionary. First, frequent updates by the reference hospital invalidated previously consistent mappings, forcing manual remapping efforts. This not only increased the workload but also introduced a higher risk of mapping errors. These disruptions highlight the critical need to finalize a stable and harmonized data dictionary before initiating the project. Establishing such a standardized data dictionary would minimize unnecessary adjustments and reduce inconsistencies during the mapping process. Second, although the 1-to-1 mapping approach provided by the reference site was initially helpful, the transformation process revealed its limitations. A more enriched data dictionary is essential to support consistent and complete mappings. This enriched version should include detailed mapping rules, clinical context,

XSL•FO
RenderX

and clear rationales for assigning raw data elements to specific OMOP IDs to address these gaps.

Another challenge lies in the variability of how data are collected, structured, and recorded across hospitals. Unstructured data, such as free-text entries in EHRs, often result in missing or unusable data during mapping [38]. Similarly, coding discrepancies, such as differences in granularity between source data, can lead to inconsistencies and missing values [39]. These technical issues not only reduce the completeness of the mapped data but also hinder its clinical applicability and analytical utility. Differences in granularity, terminology, and data format between hospitals further exacerbate these challenges, introducing biases during the extraction and mapping process.

Beyond technical challenges, the success of the transformation process depends heavily on the knowledge and expertise of the individuals involved. Mapping requires a deep understanding of the OMOP CDM framework, including its vocabularies, as well as comprehensive knowledge of the clinical and technical aspects of the source data and source systems. Insufficient expertise can result in mapping errors or inconsistencies, particularly for complex variables requiring nuanced interpretation. Differences in how data are collected and structured between the 2 hospitals introduced inconsistencies in the mapping process. Differences in source systems between institutions can introduce biases, as variations in data collection may not always be fully accounted for in the mapping strategy.

Implications for Practice and Research

The findings of this study have several implications for future OMOP implementations. First, they highlight the importance of structured data capture at the source, as unstructured data limits downstream usability. Second, there is a need for stable and enriched data dictionaries that include detailed mapping logic and clinical context. Finally, the results demonstrate that iterative data quality assessment is essential to achieve high completeness and consistency.

Mapping raw health data to the OMOP CDM is a complex process requiring in-depth planning, a structured approach, and multidisciplinary collaboration to ensure high-quality outcomes. On the basis of the insights from this study, the following recommendations are proposed to address gaps identified during the transformation process:

1. Develop an enriched and stable data dictionary: move beyond one-to-one mappings by creating a data dictionary that includes detailed mapping logic and explanatory rationale. This enriched data dictionary should capture the clinical context of variables (eg, the underlying reason and circumstances under which a variable is captured), the structure of the source data, and any transformations applied to align with OMOP conventions. Finalizing a stable and harmonized data dictionary before initiating the project will prevent data quality issues from occurring during the transformation process.
2. Leverage data profiling to address variability across hospitals: systematically profile source data to understand their structure, coding practices, completeness, and variability. This includes identifying differences in data formats, coding systems (eg, *ICD-10* use), value distributions, and missing data patterns prior to mapping.
3. Strengthen expertise through training and collaboration: equip mapping teams with training programs that provide an in-depth understanding of the OMOP CDM framework, including its tables, relationships, and vocabularies [40]. Encourage collaboration between medical and technical experts to ensure that mapping strategies capture both clinical accuracy and technical precision.
4. Automate the mapping process where possible: automate the mapping process where feasible to improve consistency and reduce manual effort, particularly for standardized variables.
5. Conduct data quality assessments: perform data quality assessments before and after mapping to identify inconsistencies and ensure completeness of the transformed dataset.

These recommendations provide a structured framework to improve mapping quality and enhance the reliability of standardized datasets for multicenter research.

Limitations

This study has some limitations. First, no formal data quality assessment was performed after extracting raw data. Consequently, potential inaccuracies or inconsistencies in the source data may have influenced the mapping outcomes, affecting the quality of the final OMOP dataset. Second, the role and timing of medical expert involvement remain unclear. Although their expertise is crucial for interpreting raw data and ensuring consistent mappings, their absence during the creation of the data dictionary may have limited its clinical relevance. Finally, uncertainty about when to involve experts in the workflow—whether during data interpretation, technical mapping, or both—may have affected the consistency of the process. Addressing these limitations through postextraction quality checks and clearer integration of medical expertise can improve the trustworthiness and reliability of future mapping efforts.

Conclusions

This study highlights the challenges of mapping lung cancer data to the OMOP CDM, particularly in managing unstructured data, addressing granularity discrepancies, and adapting to evolving reference standards. Although high consistency was achieved for structured variables, limitations in handling free-text data and incomplete mapping logic documentation revealed areas for improvement. The interplay between technical challenges and nontechnical factors, such as human expertise and system variability, highlights the need for a multidisciplinary approach to OMOP CDM implementation. Collaborations between clinical experts, data scientists, and data engineers are essential to bridge gaps in knowledge and address the complexities of transforming diverse health care data into a standardized format. Furthermore, fostering a shared understanding of the source systems across sites and aligning on best practices for mapping logic can improve the reliability of the mapped data. These insights lay the groundwork for creating harmonized datasets to support robust multicenter research and clinical analytics.

XSL•FO
RenderX

Acknowledgments

The authors thank the Data Science Institute of Ghent University Hospital for its valuable guidance and for providing the opportunity to conduct this study. This work would not have been possible without their support.

ChatGPT (OpenAI) was used as a writing assistant to improve the language and clarity of the manuscript. The authors maintained full control over the content, verified the accuracy of all artificial intelligence–generated suggestions, and take full responsibility for the integrity of the final manuscript.

Funding

The authors declared that no financial support was received for this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
SQL scripts using the vocabularies and relationships defined in the Observational Medical Outcomes Partnership Common Data Model version 5.4.
[DOCX File, 18 KB - jmir_v28i1e90246_app1.docx]

References

1. Jungkunz M, Köngeter A, Mehlis K, Winkler EC, Schickhardt C. Secondary use of clinical data in data-gathering, non-interventional research or learning activities: definition, types, and a framework for risk assessment. J Med Internet Res 2021 Jun 8;23(6):e26631. [doi: 10.2196/26631] [Medline: 34100760]
2. Coorevits P, Sundgren M, Klein GO, et al. Electronic health records: new opportunities for clinical research. J Intern Med 2013 Dec;274(6):547-560. [doi: 10.1111/joim.12119] [Medline: 23952476]
3. Fruchart M, Quindroit P, Jacquemont C, Beuscart JB, Calafiore M, Lamer A. Transforming primary care data into the observational medical outcomes partnership common data model: development and usability study. JMIR Med Inform 2024 Aug 13;12:e49542. [doi: 10.2196/49542] [Medline: 39140273]
4. Oja M, Tamm S, Mooses K, et al. Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) common data model: lessons learned. JAMIA Open 2023 Dec;6(4):ooad100. [doi: 10.1093/jamiaopen/ooad100] [Medline: 38058679]
5. Raman SR, Curtis LH, Temple R, et al. Leveraging electronic health records for clinical research. Am Heart J 2018 Aug;202:13-19. [doi: 10.1016/j.ahj.2018.04.015] [Medline: 29802975]
6. von Lucadou M, Ganslandt T, Prokosch HU, Toddenroth D. Feasibility analysis of conducting observational studies with the electronic health record. BMC Med Inform Decis Mak 2019 Oct 28;19(1):202. [doi: 10.1186/s12911-019-0939-0] [Medline: 31660955]
7. Puttkammer N, Baseman JG, Devine EB, et al. An assessment of data quality in a multi-site electronic medical record system in Haiti. Int J Med Inform 2016 Feb;86:104-116. [doi: 10.1016/j.ijmedinf.2015.11.003] [Medline: 26620698]
8. Hribar MR, Read-Brown S, Goldstein IH, et al. Secondary use of electronic health record data for clinical workflow analysis. J Am Med Inform Assoc 2018 Jan 1;25(1):40-46. [doi: 10.1093/jamia/ocx098] [Medline: 29036581]
9. Bernardi FA, Mello de Oliveira B, Bettiol Yamada D, et al. The minimum data set for rare diseases: systematic review. J Med Internet Res 2023 Jul 27;25:e44641. [doi: 10.2196/44641] [Medline: 37498666]
10. Bernardo BM, Mamede HS, Barroso JM, dos Santos VM. Data governance & quality management—innovation and breakthroughs across different fields. J Innov Knowl 2024 Oct;9(4):100598. [doi: 10.1016/j.jik.2024.100598]
11. Declerck J, Kalra D, Vander Stichele R, Coorevits P. Frameworks, dimensions, definitions of aspects, and assessment methods for the appraisal of quality of health data for secondary use: comprehensive overview of reviews. JMIR Med Inform 2024 Mar 6;12:e51560. [doi: 10.2196/51560] [Medline: 38446534]
12. Bernardi FA, Alves D, Crepaldi N, Yamada DB, Lima VC, Rijo R. Data quality in health research: integrative literature review. J Med Internet Res 2023 Oct 31;25:e41446. [doi: 10.2196/41446] [Medline: 37906223]
13. Goldberg SI, Niemierko A, Turchin A. Analysis of data errors in clinical research databases. AMIA Annu Symp Proc 2008 Nov 6;2008:242-246 [FREE Full text] [Medline: 18998889]
14. Adeniran IA, Efunniyi CP, Osundare OS, Abhulimen AO. Data-driven decision-making in healthcare: improving patient outcomes through predictive modeling. Int J Scholarly Res Multidiscip Studies 2024;5(1):059-067. [doi: 10.56781/ijsrms.2024.5.1.0040]

XSL•FO
RenderX

15. Wiebe N, Xu Y, Shaheen AA, Eastwood C, Boussat B, Quan H. Indicators of missing Electronic Medical Record (EMR) discharge summaries: a retrospective study on Canadian data. Int J Popul Data Sci 2020 Dec 11;5(1):1352. [doi: 10.23889/ijpds.v5i3.1352] [Medline: 34007880]
16. Madigan D, Ryan PB, Schuemie M, et al. Evaluating the impact of database heterogeneity on observational study results. Am J Epidemiol 2013 Aug 15;178(4):645-651. [doi: 10.1093/aje/kwt010] [Medline: 23648805]
17. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19(1):54-60. [doi: 10.1136/amiajnl-2011-000376] [Medline: 22037893]
18. Liaw ST, Guo JG, Ansari S, et al. Quality assessment of real-world data repositories across the data life cycle: a literature review. J Am Med Inform Assoc 2021 Jul 14;28(7):1591-1599. [doi: 10.1093/jamia/ocaa340] [Medline: 33496785]
19. Kahn MG, Callahan TJ, Barnard J, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC) 2016;4(1):1244. [doi: 10.13063/2327-9214.1244] [Medline: 27713905]
20. Liaw ST, Rahimi A, Ray P, et al. Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature. Int J Med Inform 2013 Jan;82(1):10-24. [doi: 10.1016/j.ijmedinf.2012.10.001] [Medline: 23122633]
21. Ozonze O, Scott PJ, Hopgood AA. Automating electronic health record data quality assessment. J Med Syst 2023 Feb 13;47(1):23. [doi: 10.1007/s10916-022-01892-2] [Medline: 36781551]
22. Aerts H, Kalra D, Saez C, et al. Is the quality of hospital EHR data sufficient to evidence its ICHOM outcomes performance in heart failure? A pilot evaluation. Health Informatics 2021;9(8):e27842. [doi: 10.2196/27842] [Medline: 34346902]
23. Wang Z, Penning M, Zozus M. Analysis of anesthesia screens for rule-based data quality assessment opportunities. Stud Health Technol Inform 2019;257:473-478 [FREE Full text] [Medline: 30741242]
24. Rudin RS, Fischer SH, Damberg CL, et al. Optimizing health IT to improve health system performance: a work in progress. Healthc (Amst) 2020 Dec;8(4):100483. [doi: 10.1016/j.hjdsi.2020.100483] [Medline: 33068915]
25. Schmidt L, Olorisade BK, McGuinness LA, Thomas J, Higgins JPT. Data extraction methods for systematic review (semi)automation: a living systematic review. F1000Res 2021;10:401. [doi: 10.12688/f1000research.51117.1]
26. Cascini F, Pantovic A, Al-Ajlouni YA, Puleo V, De Maio L, Ricciardi W. Health data sharing attitudes towards primary and secondary use of data: a systematic review. EClinicalMedicine 2024 May;71:102551. [doi: 10.1016/j.eclinm.2024.102551] [Medline: 38533128]
27. Kumar G, Basri S, Imam AA, Khowaja SA, Capretz LF, Balogun AO. Data harmonization for heterogeneous datasets: a systematic literature review. Appl Sci (Basel) 2021;11(17):8275. [doi: 10.3390/app11178275]
28. OHDSI. The Book of OHDSI: Observational Health Data Sciences and Informatics: OHDSI; 2019. URL: https://books.google.co.uk/books?id=JxpnzQEACAAJ [accessed 2025-05-25]
29. McHugh ML. Interrater reliability: the Kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
30. De Vries H, Elliott MN, Kanouse DE, Teleki SS. Using pooled Kappa to summarize interrater agreement across many items. Field methods 2008 Aug;20(3):272-282. [doi: 10.1177/1525822X08317166]
31. Belgian Cancer Registry. 2024. URL: https://kankerregister.org/nl [accessed 2026-05-27]
32. Koninklijk besluit betreffende de vaststelling en de vereffening van het budget van financiële middelen van de ziekenhuizen. Belgian Federal Government Services. 2002. URL: https://www.ejustice.just.fgov.be/eli/besluit/2002/04/25/2002022335/justel [accessed 2025-05-25]
33. RADar-azdelta/keun. GitHub. URL: https://github.com/RADar-AZDelta/Keun [accessed 2025-05-25]
34. RADar-azdelta/rabbit-in-a-blender. GitHub. URL: https://github.com/RADar-AZDelta/Rabbit-in-a-Blender [accessed 2025-05-25]
35. Biedermann P, Ong R, Davydov A, et al. Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases. BMC Med Res Methodol 2021 Nov 2;21(1):238. [doi: 10.1186/s12874-021-01434-3] [Medline: 34727871]
36. Ehsani-Moghaddam B, Martin K, Queenan JA. Data quality in healthcare: a report of practical experience with the Canadian Primary Care Sentinel Surveillance Network data. Health Inf Manag 2021;50(1-2):88-92. [doi: 10.1177/1833358319887743] [Medline: 31805788]
37. Kent S, Burn E, Dawoud D, et al. Common problems, common data model solutions: evidence generation for health technology assessment. Pharmacoeconomics 2021 Mar;39(3):275-285. [doi: 10.1007/s40273-020-00981-9] [Medline: 33336320]
38. Sedlakova J, Daniore P, Horn Wintsch A, et al. Challenges and best practices for digital unstructured data enrichment in health research: a systematic narrative review. PLOS Digit Health 2023 Oct;2(10):e0000347. [doi: 10.1371/journal.pdig.0000347] [Medline: 37819910]
39. Syed R, Eden R, Makasi T, et al. Digital health data quality issues: systematic review. J Med Internet Res 2023 Mar 31;25:e42615. [doi: 10.2196/42615] [Medline: 37000497]
40. Observational Health Data Sciences and Informatics. The Book of OHDSI: Observational Health Data Sciences and Informatics; 2021. URL: https://ohdsi.github.io/TheBookOfOhdsi/ [accessed 2026-05-28]

XSL•FO
RenderX

Abbreviations

CDM: Common Data Model
DSI: Data Science Institute
DWH: Data Warehouse
EHR: electronic health record
ETL: extract, transform, and load
FHIN: Federated Health Innovation Network
***ICD-10*:** *International Classification of Diseases, 10th Revision*
LOINC: Logical Observation Identifiers Names and Codes
MOC: Multidisciplinary Oncology Consultation
MZG: Minimale Ziekenhuisgegevens
OMOP: Observational Medical Outcomes Partnership
SNOMED CT: Systematized Nomenclature of Medicine–Clinical Terms
TNM: tumor, node, and metastasis

Edited by M Balcarras; submitted 23.Dec.2025; peer-reviewed by A Jouned, M Spotnitz, R Ismayilov; revised version received 04.May.2026; accepted 05.May.2026; published 08.Jun.2026.

Please cite as:
Declerck J, Deschepper M, Colpaert K, Kalra D, Coorevits P
Addressing Data Quality Challenges in Lung Cancer Data Within the Observational Medical Outcomes Partnership Common Data Model: Observational Study
J Med Internet Res 2026;28:e90246
URL: https://www.jmir.org/2026/1/e90246
doi:10.2196/90246

© Jens Declerck, Mieke Deschepper, Kirsten Colpaert, Dipak Kalra, Pascal Coorevits. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Quality of Cancer-Related Clinical Coding in Primary Care in North Central London: Mixed Methods Quality Improvement Project

Afsana Bhuiya[1], BSc, MBBS; Graham Roberts[1], BSc; Katie Tucker[1], BSc, MSc; Stefanie Bonfield[2], BSc, MSc; Georgia Black[2], MSc, PhD

[1]North Central London Cancer Alliance, London, United Kingdom

[2]Wolfson Institute of Population Health, Queen Mary University of London, London, United Kingdom

Corresponding Author:
Afsana Bhuiya, BSc, MBBS
North Central London Cancer Alliance
UCLH, 47 Wimpole Street
London, W1G 8SE
United Kingdom
Phone: 44 7877136123
Email: afsana.bhuiya1@nhs.net

Abstract

Background: The North Central London (NCL) Cancer Alliance carried out a quality improvement (QI) project to fill a distinct knowledge gap regarding the quality of clinical coded data in a primary care electronic health care record system across the whole cancer pathway.

Objective: This study aims to establish the quality of cancer-related clinical coding in NCL primary care, encompassing both quantitative measures (eg, coding completeness and diversity) and qualitative dimensions such as clinical relevance and workflow alignment.

Methods: This was a mixed methods QI project in which we combined an observational dataset review and qualitative data from stakeholder interviews, workshops, and discussions. In the dataset review, we evaluated completeness, diversity, validation, and granularity in cancer clinical coding along the patient cancer pathway, which was split into three domains: (1) patient characteristics and risk factors, (2) cancer screening attendance, and (3) living with cancer. It was conducted in NCL primary care electronic health record systems, covering a population of over 1.4 million adults across 5 boroughs.

Results: Cancer-related clinical coding in NCL primary care revealed significant gaps despite high completeness for ethnicity (912,679/1,055,083, 86.5%) and language (898,023/1,307,601, 68.7%). Employment status (29,848/1,229,644, 2.4%) and family history of cancer (183,424/1,236,580, 14.8%) were underrecorded, with wide variation in coding practices. Screening data showed good alignment with national datasets for cervical and bowel screening but fragmented and inconsistent breast screening data due to a lack of standardized codes. Cancer diagnosis coding was incomplete (4604/5260, 87.5% recorded), and treatment and staging data were almost entirely absent, limiting proactive management of long-term consequences. Stakeholder input highlighted inconsistent template use, limited data updates, and insufficient incentives as key barriers to better coding.

Conclusions: The QI project has provided a detailed insight into the many dimensions of cancer coding and sheds light on many factors that underpin variation and coding preference. We offer a number of recommendations. The prioritized ones include the need for a cancer clinical coding data framework for primary care supported by appropriate funding and incentivization; improvements in the breast screening pathway and its interface with primary care; improvements in the quality of secondary care information that is sent to primary care; and dissemination of the importance of coding of cancer activity in primary care.

(J Med Internet Res 2026;28:e73205) doi:10.2196/73205

KEYWORDS

cancer diagnosis; cancer pathway; cancer risk factors; cancer treatment; cancer; clinical coding; coding completeness; coding data; coding diversity; coding processes; coding quality; coding validation; coding variation; inequalities data; primary care coding; quality improvement; SNOMED CT; Systematized Nomenclature of Medicine – Clinical Terms

XSL•FO
RenderX

Introduction

Overview

Clinical coding of cancer-related data in primary care supports accurate and timely data collection, analysis, and reporting of cancer diagnoses and treatments, which in turn facilitates high-quality patient care [1,2]. Consequently, incomplete or inaccurate clinical coding of cancer-related data has significant implications across the cancer pathway. For example, cancer prevention efforts, including information provision, vaccination, and screening, may be restricted if it is not possible to identify eligible individuals based on data available in primary care records (eg, age, sex, health behaviors, previous medical history) and follow up with those who have not engaged (eg, those who have not responded to previous cancer screening invitations) [3]. Missing data for cancer risk-factors, such as family history or previous medical history, may undermine appropriate referral of symptomatic patients for cancer investigation [4]. Similarly, missing data on precancerous conditions such as Barrett's esophagus or bowel polyps may prevent health care professionals from providing information and support to patients about managing their condition and personal risk and result in patients being excluded from relevant safety-netting efforts and surveillance pathways [5]. Meanwhile, the underreporting of cancer cases can lead to an underestimation of the true cancer burden within a population and limit the ability of primary care to support patients during cancer diagnosis and treatment [5,6]. Beyond clinical impact, poor coding may also contribute to financial losses for health care providers and hinder effective service planning at both practice and system levels [7].

Primary care cancer coding data are often variable and suboptimal, with a poor evidence base for improvement [8-10]. A systematic review by Thiru et al [11], concluded on the lack of standardized measures for data quality, which is supported by previous studies showing the heterogeneity of quality assessment methods in primary care coding [11,12]. Thiru et al [11], uncovered studies that assessed primary electronic patient record data and studies that reviewed survey and questionnaire data. They also found that data quality (reliability) was usually measured with rate comparisons and data validity was expressed under a range of terms (completeness, correctness, accuracy, consistency, and appropriateness), which were rarely defined [11].

Previous studies have identified the need for better communication about patients who have been diagnosed with cancer between primary and secondary care [13], and a UK study reported that 1 in 5 patients with cancer were not recorded to have a cancer diagnosis in primary care records [6]. In England, relatively robust audit systems and regulatory oversight exist to regulate coding for cancer diagnosis, cancer interventions, and procedures in secondary care, underpinned by financial incentives [14,15]. In contrast, equivalent governance mechanisms do not exist in primary care. Clinical entries in primary care rely on SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms), but there are no national standards or frameworks specifically guiding cancer coding across the whole pathway. Instead, coding behavior is shaped primarily by the Quality Outcomes Framework (QOF) [16], which provides financial incentives for documenting selected conditions and activities. For cancer, QOF incentivizes the coding of a cancer diagnosis and the completion of cancer care reviews (CCRs) following diagnosis—although CCRs were removed from general practitioner (GP) QOF contracts in April 2025 [16,17]. Importantly, QOF does not cover the full cancer pathway. Moreover, primary care coding systems often encourage diversity rather than consistency, such as having multiple codes to describe identical clinical events [10,18] (eg, "smoker," "cigarette smoker," "moderate cigarette smoker"). Lack of regulation and research evidence about cancer coding consistency and variation restricts the development of targeted quality improvement (QI) measures. Consequently, there is a need to understand how the quality of clinical coding data in primary care varies to influence outcomes across the cancer pathway. This holistic approach can inform suitable interventions for optimizing the quality of primary care coded data to deliver valuable improvements in cancer prevention, referral, diagnosis, and treatment outcomes [13].

Cancer Coding and Data Curation

The process of how patient information is recorded, updated, and monitored in electronic health records is sometimes called "data curation." Primary and secondary care data curation is influenced by multiple drivers, which in turn affect data quality. In the United Kingdom, the National Cancer Registration and Analysis Service (NCRAS) [19] provides guidance and support for clinical coding of cancer diagnoses and treatment in acute care. NCRAS has developed a set of coding standards that are incorporated in nationally commissioned datasets such as the Cancer Outcomes and Services Dataset and Systemic Anti-Cancer Therapy dataset, which provide guidance on the coding of cancer diagnoses and treatments. The National Health Service (NHS) primary care coding system is called SNOMED CT [20]. The SNOMED CT does not have standard clinical coding specifications for cancer data.

Acute and mental health trusts have standard procedures for regular quality inspections of their coded clinical data for inpatient and day-case episodes by approved clinical coding auditors, who aim to demonstrate compliance with national clinical coding standards [21]. The regulatory component is important, as it supports high-quality data collection that supports secondary uses of the data, such as collaborations with academic and research departments (eg, the "Getting It Right First Time" program to reduce unwarranted variation) [22]. There is no equivalent process for auditing the assignment of the terminology SNOMED CT, which is used in primary care.

Reasons for primary care coding incompleteness and inconsistencies are well documented and include time pressures, finding the right code, and motivation to code [12,23]. As part of the development work that led to this study, NCL Cancer Alliance conducted an online survey of GP respondents in London to understand barriers to good-quality clinical coding in primary care. The findings included lack of standardized coding practices, scarcity of dedicated staff time to code, and inadequate training around coding (for the full list of barriers see Table S2 in Multimedia Appendix 1 [24]).

XSL•FO
RenderX

Context for the Quality Improvement Project

The aim of this QI project was to assess the completeness and variation of clinical cancer coding in NCL primary care data and to understand the reasons for this. The output of this QI project was to develop a robust "case for change" [25] for relevant improvement solutions. The long-term goal of the project was to improve service planning, pathway delivery, and redress health inequalities. The QI project was carried out by the NCL Cancer Alliance team with support from researchers at the Queen Mary University of London.

Our objectives were to: first, examine the quality of clinical coding of cancer-relevant risk factors, processes, and outcomes, encompassing both quantitative measures (eg, coding completeness and diversity) and qualitative stakeholder perspectives (eg, barriers, enablers, clinical relevance, and workflow alignment); second, use data from the first objective to develop recommendations to support data improvement in primary care.

Methods

Overview

This was a mixed methods QI project combining an observational dataset review and stakeholder interviews, workshops, and discussions. The observational dataset and analysis were based on data searches designed in EMIS (Egton Medical Information Systems) Web by Enfield GP Federation. These were set up to take a snapshot of information held in GP records as it stood on October 31, 2023, so that results are consistent and comparable across practices. All participating GP federation teams then ran the searches between November 2023 and January 2024. The results were shared with the NCL Cancer Alliance, which completed the data analysis in April 2024. Preliminary insights from the early stages of qualitative data collection (eg, advice from group interviews) informed the quantitative analysis (eg, process barriers in breast screening to ethnicity codes recorded in 2 different parts of EMIS Web) and the development of the workshop themes and questions. The formal qualitative analysis was conducted after all qualitative data had been collected; this section of analysis was carried from October to December 2024 (Figure S1 in Multimedia Appendix 1 depicts the project timeline).

Setting and Participants

We gathered data on clinical coding from NCL primary care GP systems covering an adult population of >1.4 million. The qualitative data included email conversations, 2 semistructured interviews, and 2 workshops with key primary care stakeholders. The qualitative methods and analyses are reported according to the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist (Table S3 in Multimedia Appendix 1) [24].

Data Collection

We extracted GP data from the electronic health care system, EMIS Web [26], the sole GP electronic health care record provider for all GP practices in NCL. Data were obtained using built-in "searches" within EMIS Web. These are configurable protocols designed to retrieve coded patient information based on predefined clinical or demographic criteria, referred to as data domains (eg, ethnicity, smoking status, cancer diagnosis, or treatment history). Each search identifies patients meeting the selected criteria based on structured clinical codes. The output of these searches is presented in the form of "reports," which summarize the number of patients meeting each criterion and can be exported for further analysis. Figure S2 in Multimedia Appendix 1 provides an example of a search. We used the SNOMED CT [20,27] and EMIS Web clinical codes [28], which coexist in patients' records [29]. The NCL GP federations performed the searches; a GP federation is a group of general practices working collaboratively as an organizational entity to improve patient care, share resources, and enhance service provision within the local health economy [30].

Table 1 illustrates the number of GP practices across each GP federation or primary care network (PCN) in NCL and the completeness of report returns across the 26 searches that were built and run. A PCN in Islington that was not part of the Islington GP Federation (Islington North 2) did not participate in this project. A partial return is where some GP practices have generated results in a search and some have not due to there either being no patients who meet the criteria or technical constraints that prohibit the search from running for particular practices.

Table 1. Shows the number of general practitioner (GP) practices within each GP federation or primary care network (PCN) and the completeness of the reports that were requested.

Category	GP entity						
	Barnet[a], n	Camden[a], n	Camden Health Evolution[a], n	Enfield[a], n	Haringey[a], n	Islington[a], n	Islington North 2[b], n
Number of GP practices	48	23	9	30	34	23	8
Report status							
Complete	7	21	21	13	21	21	0
Partial	17	5	5	13	4	4	0
Did not return	2	0	0	0	0	0	26
Search generated null results	0	0	0	0	1	1	0

[a]Represents GP federations.

[b]Represents a PCN.

XSL•FO
RenderX

Three personalized cancer care metrics—cancer care plan given, end-of-treatment summary, and holistic needs assessment—were unintentionally omitted from the original data request. This gap was identified by the NCL Cancer Alliance after data had already been submitted by all boroughs. To partially address this, Enfield Federation, which had local access, conducted the relevant searches for its own borough (covering the 12 months up to October 2023). Due to timing and resource constraints, it was not feasible to repeat this process across the other boroughs. Enfield's data were compared with HealtheIntent cancer registries [31], which span the entire NCL adult population; while full coding quality could not be assessed, we were able to examine the relative frequency of these codes across populations.

Data Quality Assessment Method

We followed the approach taken by Pineda-Moncusi et al [32], who examined ethnicity data at a large scale and defined data quality across completeness, coverage, and granularity (most prevalent clinical codes used).

We drew on existing cancer data frameworks [33,34] to conceptualize different cancer pathway stages and identify relevant cancer codes. Within cancer alliances, there are established programs [35] spanning the entire pathway—from awareness and prevention through to living with and beyond cancer. Our aim in this study was to create a comprehensive and holistic dataset to review, thereby informing future recommendations. To do this, we built on this existing knowledge, incorporated earlier work from our team [36], and collaborated with our wider NCL Cancer Alliance team to scope data items across each pathway stage. Although we recognized that some of these items were unlikely to be routinely coded in primary care, there was no empirical evidence to confirm this; therefore, part of the purpose of this QI project was to assess the current state of coding. End-of-life care was deemed out of scope for this work. The Enfield GP Federation digital team built EMIS Web searches to cover each element of the pathway and shared these with other NCL GP federations' IT teams to run for each borough. Enfield GP Federation's IT team oversaw the communication, search development, search dissemination, and data submissions. Raw data were transferred to the NCL Cancer Alliance.

The data protection officer for NCL primary care assured that all data sharing complied with UK General Data Protection Regulation.

Qualitative Data Collection

A convenience sample of key stakeholders with expertise in clinical coding were invited (over email and through a GP bulletin) to attend an online group workshop through existing contacts (including GPs, project and program managers, IT staff, and academic researchers). They were told that they were being invited to discuss the quantitative findings relating to clinical coding. We conducted 2 semistructured group interviews and 2 workshops with 11 primary care stakeholders. We also reviewed email correspondence from the Enfield GP Federation team, which captured responses to queries arising from the initial round of quantitative data analysis. These communications provided a systematic method for clarifying data gaps and process-related issues and directly informed the development of key questions explored during the subsequent stakeholder workshops.

All interviews and workshops were held remotely over Microsoft Teams between April and August 2024 (Table S4 in Multimedia Appendix 1) and conducted by AB (female, GP clinical lead for Innovation and Integration at NCL Cancer Alliance). Sessions were also facilitated by 2 other members of the research team, GR (male, head of Data and Analytics, NCL Cancer Alliance) and KT (female, senior innovation consultant, NCL Cancer Alliance). Semistructured interview and workshop topic guides were developed by AB, KT, and GR. We presented key findings for discussion. All remote sessions were video recorded and transcribed using the Microsoft Teams record and transcription functions. Transcripts were not shared with stakeholders.

Analysis

Quantitative data were analyzed descriptively across the different boroughs to characterize coding patterns. In total, 26 searches were built and run. We analyzed these data domains: patient ethnicity, main language spoken, weight or BMI, alcohol consumption, smoking status, family history of cancer, employment status, environmental pollutants exposure, carer, cancer screening attendance, cancer fast-track referrals, presence of malignant neoplastic disease, treatment regimen, and attendance at CCR. To ensure transparency and facilitate replication as far as possible, we have included these extracted search terms in Textbox S1 in Multimedia Appendix 1.

Codes were descriptively analyzed for the following features (1) coding completeness, (2) coding diversity, (3) data validation, and (4) granularity of coding (Table 2).

XSL•FO
RenderX

Table 2. Definitions and methods used to assess clinical coding quality and completeness.

Definitions	Methods
Coding completeness[a]	Percentage of eligible patients with a relevant SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) or EMIS (Egton Medical Information Systems) clinical code captured. This reflects the presence or absence of a code rather than the true or expected prevalence of the underlying condition.
Coding diversity[b]	Number of unique SNOMED CT or EMIS codes used to describe the same information (eg, patient's weight). This serves as a practical proxy for coding breadth, acknowledging limitations such as inaccessible child codes and variation in code list size.
Data validation	Comparison of coding completeness in EMIS Web with national or local reference datasets to assess missed or uncoded data. Calculated as EMIS Web completeness divided by completeness reported in the comparator dataset for the same coded item. Ratios may exceed 100% if EMIS Web demonstrates higher apparent completeness. Comparator datasets included HealtheIntent, Cancer Waiting Times, the National Cancer Registry, South East London Cancer Alliance data[c], and the NHS Futures Screening Dashboard.
Granularity of coding	Review of the most frequently used codes to assess specificity, for example the use of general codes such as "current smoker."

[a]Coding completeness describes the proportion of eligible patients in the denominator population who have a relevant code captured in their primary care record. This definition is intended to reflect the presence or absence of a code rather than to imply that the observed percentages represent the true or expected prevalence of the underlying clinical characteristic or diagnosis. For example, a coding completeness of 0.3% for malignant neoplastic disease reflects the proportion of the registered population with a relevant cancer code, not an assessment of whether this proportion is "correct" or "incorrect." To distinguish between the technical measurement of code capture and the interpretive question of whether coding levels are as expected, we also applied the concept of coding validation. Coding validation involves comparing observed coding completeness against external standards or datasets (eg, cancer wait times) to assess whether the recorded levels are appropriate and consistent with known population rates. In this way, completeness provides a descriptive measure of the presence of codes in primary care data, while validation enables assessment of their adequacy and alignment with clinical or epidemiological expectations.

[b]Coding diversity was defined as the number of unique codes captured in EMIS Web for each search. We recognize this is a practical proxy rather than a full measure, as EMIS Web does not easily expose all child codes, and many codes (eg, for rare conditions or languages) will naturally yield zero results. Some code lists are also inherently larger than others. A distinct count of unique codes therefore provides useful context on the breadth of coding options observed while acknowledging these limitations.

[c]South East London Integrated Care Board (ICB) and Cancer Alliance has its own population health dashboard that overlaps with the data and definitions in this study, meaning it is valid for comparison where we have no other published source. This dashboard is not publicly available.

Data completeness and validation analysis was carried out against all submitting GP practices' adult populations as of January 2024 (Table S5 in Multimedia Appendix 1). Practice populations are relatively stable month to month, so comparing October 2023 search results with the adult population in January 2024 is considered valid. Practices that did not submit data for a profile were excluded from the analysis to maximize data integrity (Table S6 in Multimedia Appendix 1).

Table 3 lays out the report names across the time frames for which the data was searched for, the denominator population, and the validation database used for comparison. Additionally: (1) data on body weight and BMI were assessed through a combined height and weight search); (2) smoking status was assessed through three separate searches; (3) family history of cancer was assessed based on any recorded code, rather than limiting analysis to the preceding 24 months (as the GP federations' IT team were aware this would be captured at one point in records); and (4) breast screening data were retrieved through four distinct searches: screening attendance, normal results, abnormal results, and cancer detected.

XSL•FO
RenderX

Table 3. Each report name shown against time period covered for each report, the denominator population, and the validation database used for comparison, including the validator time period.

Report name	Time period	Denominator population characteristics (age in years)	Validator database	Validator metric and time period
Ethnic origin	Recorded ever	>18	HealtheIntent	Ethnicity coding (April 2024)
Main language spoken	Recorded ever	>18	HealtheIntent	Main language spoken (April 2024)
Employment status	October 2021-October 2023	>18	No comparator	__[a]
Weight or BMI recorded	October 2021-October 2023	>18	South East London Cancer Alliance data completeness	Weight or BMI recorded (April 2024)
Current smoker (recorded in last 24 months)	October 2021-October 2023	>18	HealtheIntent	Current smoker status (April 2024)
Any smoking status (recorded in last 24 months)	October 2021-October 2023	>18	South East London Cancer Alliance data completeness	Any smoking status (April 2024)
Alcohol consumption record	October 2021-October 2023	>18	South East London Cancer Alliance data completeness	Alcohol consumption record (April 2024)
Family history of neoplasm	Recorded ever	>18	No comparator	—
Environmental pollutants	October 2021-October 2023	>18	No comparator	—
Fast track referral coding	New episode added in last 12 months	>18	Cancer Wait Times	Urgent suspected cancer referrals (2023)
Cancer–bowel: did not return screening kit	October 2023-2 years 6 months	60-74	No comparator	—
Cancer–bowel: screening uptake	October 2023	60-74	NHS Futures Screening Dashboard	Bowel screening uptake August 2023 (50-70 years)
Cancer–bowel: abnormal result	October 2023-2 years 6 months	60-74	No comparator	—
Cancer–breast: screened	October 2023-3 years 6 months	50-70 (female)	NHS Futures Screening Dashboard	Breast screening uptake August 2023 (50-70 years)
Cancer–breast: abnormal result	October 2023-3 years 6 months	50-70 (female)	No comparator	—
Cancer–breast: normal result	October 2023-3 years 6 months	50-70 (female)	No comparator	—
Cancer–breast cancer detected	October 2023-3 years 6 months	After screened for breast cancer (female)	National Cancer Registry	Breast cancer diagnosis via screening route February 2020 to July 2023 (3.5 years)
Cancer–cervical: adequate smear	October 2023-3 years 6 months	25-49 (female)	NHS Futures Screening Dashboard	Cervical screening uptake December 2023 (25-49 years)
Cancer–cervical: adequate smear	October 2023-5 years 6 months	50-64 (female)	NHS Futures Screening Dashboard	Cervical screening uptake December 2023 (50-64 years)
Malignant neoplastic disease	New episode added in last 12 months (as of October 2023)	>18	National Cancer Registry	Rapid Cancer Registration – New Diagnosis (2023)
Malignancy stage	New episode added in last 12 months (as of October 2023)	>18	National Cancer Registry	Rapid Cancer Registration – New Diagnosis (2023)
Treatment regimen	New episode added in last 12 months (as of October 2023)	>18	Cancer Waiting Times	Treatment starts (2023)
Cancer care review	New episode added in last 12 months (as of October 2023)	Cancer-diagnosed patients who had a care review	No comparator	—
Has a carer	Recorded ever	>18	No comparator	—

XSL•FO
RenderX

[a]Not applicable.

Missing Data

There were challenges with data collection in participating practices. Running searches at scale across several practices was technically difficult at times. These searches would often time out and fail, resulting in incomplete data collection. Despite repeated running of these searches to attempt to get a complete dataset, failures persisted (Table S3 in Multimedia Appendix 1 and its linked notes provide more detail on this process).

Barnet had the highest level of missing data across the 26 reports, although data gaps were present in all boroughs. The ethnic diversity of Barnet is broadly comparable to the rest of NCL, and given that data from Barnet were available for other nonethnicity reports, we do not believe these gaps materially skew the overall findings. Table S7 in Multimedia Appendix 1 illustrates the proportions of practices that submitted data across each of the 26 searches.

The searches did not examine comparable data across each screening pathway. Direct comparisons between breast, colorectal, and cervical screening were not feasible, as primary care is not uniformly responsible for screening data across the 3 screening pathways. As a result, a process mapping exercise was conducted to trace the pathways of each screening program into primary care, identifying recordable actions within primary care settings as shown in Figure 1.

Figure 1. Diagram showing how codes are generated for all cancer screening programs in primary care.

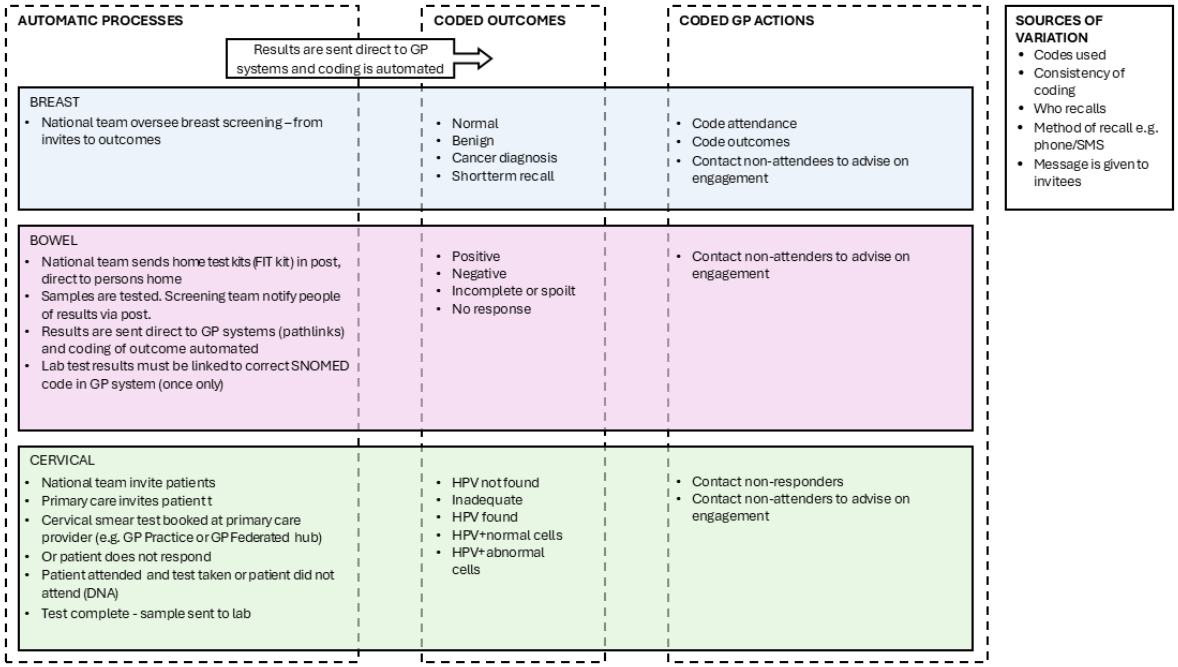

The qualitative data were analyzed by SB, a female behavioral science PhD student, and GB, a female applied health researcher. Both had previous experience conducting qualitative research. The analysis began in October 2024, after qualitative data collection had finished. SB listened to all video recordings and checked the transcripts for accuracy to become familiar with the study context and dataset before coding the data.

Qualitative data collection aimed to contextualize and validate quantitative findings rather than to achieve theoretical saturation. As this study adopted a QI focus, data collection was concluded once sufficient breadth of perspectives had been obtained and no new issues emerged that were relevant to the study objectives.

The analysis was conducted in Google Sheets using framework analysis [37]. Key excerpts from transcripts, email conversations, and comments posted in the chat during the workshops were copied into Google Sheets. All qualitative data were analyzed together. Raw data were arranged into columns. Each column represented a single code, and each row included raw data (eg, a verbatim quote labeled by data source and speaker) pertaining to that code.

An initial framework of 3 themes was developed deductively based on the findings of the quantitative analysis. SB coded the data into each theme before arranging the data into subthemes inductively. SB and GB met after 10%, 50%, and 100% of the data had been coded to discuss and revise the coding framework. A reflexive diary was kept and referred to throughout the analysis, which included field notes and impressions recorded by the researcher who conducted the interviews and workshops (AB) and the researchers who conducted the analysis (SB and GB). Regular meetings were held with the research team to clarify contextual details and discuss key interpretations. A reflexivity statement is available in Textbox S2 in Multimedia Appendix 1.

XSL•FO
RenderX

Ethical Considerations

This study was designed as a QI and health service enhancement initiative and therefore ethics approval was not applied for.

All data analyzed were fully deidentified at source, aggregated, and subject to small-number suppression (<5) in accordance with local information governance standards to protect patient confidentiality. The dataset was generated through local GP federations, with each federation running standardized searches and submitting aggregated data to Enfield Federation. Enfield Federation then securely transferred the data to the NCL Cancer Alliance for analysis.

The NCL primary care data protection officer reviewed and assured all data-sharing processes, confirming they aligned with the UK General Data Protection Regulation and the Caldicott Principles. A Data Protection Impact Assessment was not required, as the data used contained no identifiable information and presented no privacy risk.

Results

Principal Findings

We present our findings combining our descriptive analyses of clinical codes (quantitative findings) and the factors influencing coding (qualitative findings), in three themes: theme 1, precancer pathway, which includes codes relating to demographic characteristics, physical characteristics, risk factors (eg, family history), and cancer referrals; theme 2, cancer screening, which includes codes relating to screening invitations and uptake; and theme 3, postcancer diagnosis, which includes codes relating to staging, treatment, primary care surveillance, and follow-up.

In Table S8 in Multimedia Appendix 1, we include further details on data items included for each theme. Key findings are summarized for each theme before the results of the quantitative and qualitative analyses are presented.

Theme 1: Precancer Pathway

Overview

Completeness, code diversity, and validation of precancer codes are presented in Table 4 (Table S9 in Multimedia Appendix 1 provides additional detail on granularity).

XSL•FO
RenderX

Table 4. Descriptive analyses for data domains in theme 1 (precancer), showing completeness, code diversity, and validation of codes that were searched for in this theme.

Report name	Time period	Completeness and code diversity		Code validation	
		Completeness (% eligible coded), n/N (%)	Number of unique SNOMED CT[a] or EMIS[b] codes, n	Comparator source	EMIS completeness vs comparator, n/N (%)
Ethnic origin (coding)	Recorded ever	912,679/1,055,083 (86.5)	375	HealtheIntent	86.5/90.7 (95.4)
Main language spoken	Recorded ever	898,023/1,307,601 (68.7)	338	HealtheIntent	68.7/62.8 (109.4)
Employment status	24 months up to October 2023	29,848/1,229,644 (2.4)	87	No comparator for coding prevalence of this data item sourced	No comparator for coding prevalence of this data item sourced
Weight or BMI recorded	24 months up to October 2023	485,660/1,159,241 (41.9)	34	SELCA[c]	41.9/45.8 (91.5)
Current smoker (recorded in last 24 months)	24 months up to October 2023	90,029/1,236,580 (7.3)	47	HealtheIntent	7.3/14.9 (49)
Any smoking status (recorded in last 24 months)	24 months up to October 2023	530,720/1,185,812 (44.8)	117	SELCA	44.8/48.6 (92.3)
Alcohol consumption record	24 months up to October 2023	222,753/1,236,580 (18)	39	SELCA	18/29.2 (61.6)
Family history of neoplasm	Recorded ever	183,424/1,236,580 (14.8)	479	No comparator for coding prevalence of this data item sourced	No comparator for coding prevalence of this data item sourced
Environmental pollutants	24 months up to October 2023	494/1,113,575 (0.04)	45	No comparator for coding prevalence of this data item sourced	No comparator for coding prevalence of this data item sourced
Fast track referral coding	New episode added last 12 months up to October 2023	61,562/1,506,746 (4.1)	34	Cancer Waiting Times	61,562/78,989 (78)

[a]SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms.

[b]EMIS: Egton Medical Information Systems.

[c]SELCA: South East London Cancer Alliance.

The completeness of ethnicity coding was high, with 86.5% (912,679/1,055,083) of records containing an ethnicity code. A total of 375 distinct SNOMED CT codes were identified, with a 15.8% variation in coding completeness across boroughs. “Other White background - ethnic category 2001 census” made up 18.6% (170,190/912,679) of total codes captured in NCL. Language coding completeness was 68.7% (898,023/1,307,601) across the eligible population, which compares favorably with the comparator database coverage of 62.8% (929,987/1,482,024). A total of 338 unique codes were identified, with borough-level variation of 15.2% (Camden=75.3% and Enfield=60.1%). The top 2 most prevalent codes for language were “Main spoken language English” (527,227/898,023, 58.7%) and “Main spoken language NOS” (48,311/898,023, 5.4%). Employment status coding was minimal, with completeness at only 2.4% (29,848/1,229,644). The most frequently recorded codes related to unemployment (11,801/29,848, 39.5%) and work-related stress (5063/29,848, 17%), with a total of 87 distinct codes identified.

Coding completeness for BMI was 41.9% (485,660/1,159,241), with 34 individual codes used to describe weight and BMI. The SNOMED CT code for BMI accounted for 88.7% (430,736/485,660) of recorded entries. Overall, 44.8% (530,720/1,185,812) of the population had a recorded smoking status, including classifications such as current smoker and ex-smoker. Approximately 170 variations of smoking-related codes were identified. Data validation demonstrated near completion. Alcohol consumption was recorded in 18% (222,753/1,236,580) of patient records, with borough-level variation ranging from 13.9% (39,198/281,807) to 21% (64,255/281,807). A total of 39 different codes were identified. The “AUDIT-C” screening tool, which assesses excess alcohol consumption, was the prevalent code at 43.4% (96,619/222,753).

Family history of cancer across NCL showed that 14.8% (183,424/1,236,580) of records contained relevant codes, with 479 unique codes identified. There was no available comparator dataset for this parameter. The 2 most prevalent codes were “FH-Cancer” and “FH-Neoplasm” at around 17% each. Coding of environmental exposure was extremely limited, with a completeness rate of 0.04% (494/1,113,575), rendering the data unsuitable for analysis.

XSL•FO
RenderX

The incidence of suspected cancer referrals among eligible patients was 4.1% (61,562/1,506,746), with a 1% variation between the highest and lowest referring boroughs. The comparator database indicated a referral rate of 5.2% (78,989/1,506,746) for NCL residents. The analysis excluded deceased, temporary, and deregistered patients. The discrepancy between the study findings and comparator data was largely attributable to deceased patients.

Factors Influencing Clinical Coding of Precancer Data in Primary Care

Stakeholders reported 3 key factors that contribute to the quality and completeness of clinical coding of precancer data in primary care: opportunities to collect or update precancer data, motivations and capacity to collect and code precancer data, and the nature of the systems used to code precancer data.

Opportunities to Collect or Update Precancer Data

Many stakeholders reported that a key opportunity to capture precancer data is through registration forms and health check appointments. Some explained that registration templates and the commissioning of health check appointments vary by practice, leading to inconsistencies. There was agreement that standardized templates could improve this, although only for those registering subsequently, and it was noted that patients may not provide information if questions are not mandatory and the purpose of data collection is not transparent. While a few GPs suggested that mandating questions could improve data completeness, others were concerned that this could introduce barriers to registration. Family history is not routinely coded but may be documented in free text within clinical notes or referral forms. Environmental exposure data is not routinely collected or standardized within primary care records.

Some stakeholders advised that a lack of data completeness for information that changes over time, such as main language spoken or smoking status, is due to limited opportunities to update patient data after registration unless patients schedule appointments with primary care (eg, long-term conditions review, new medication appointments, e-consultations). It was suggested this could be improved by offering annual health check appointments, inviting patients to update their information, and the sharing of information collected in secondary care.

Motivations and Capacity to Collect and Code Precancer Data

Many GPs indicated that local and national financial incentive schemes (such as the QOF [38]) influence whether they collect and code precancer data at registration and during patient consultations. Some reported that variation in coding completeness between boroughs could be attributed to differences in locally commissioned services. There was agreement that clinical coding is demanding of staff time and capacity, and that improvements to the quality and completeness of clinical coding are unlikely to continue beyond the period of incentivization.

There was a general consensus that, in the absence of mandates and incentives, GPs are motivated to code precancer data that are relevant to the clinical workflow, such as arranging appointments (eg, need for interpreter, has a carer), assessing eligibility for local-level services (eg, smoking cessation, vaccination, information provision), or management of a patient's symptoms or a long-term condition. Some GPs admitted that they prefer to document precancer data in free text responses. This prevented interrupting the flow of conversation with patients, and, if needed, they could provide detail on more complex factors (eg, family history and exposure to environmental pollutants).

Nature of Systems Used to Code Precancer Data

GPs advised that coding will vary depending on whether patients register at GP practices using a paper or online form, whether registration data are coded into the system manually by staff or automatically, and which additional registration processing software practices have access to. It was also raised that there are multiple places within the system for data to be recorded. Furthermore, it was reported that some urgent suspected cancer referral forms may be available in the system but not trigger a SNOMED CT code. There was consensus that automated registration and data capture would improve the consistency and completeness of precancer data in primary care systems.

Several GPs suggested that the consistency of coding is made challenging by the array of codes available for specific types of precancer data such as smoking status, BMI, and family history, where there are different levels and layers, codes with similar or ambiguous meanings, and historic codes and prompts that cannot be removed (Figure 2). GPs admitted that codes higher up the list or those labelled with QOF prompts are most likely to be selected, and a program manager advised that coding prompts should be reserved for data that are most important to capture.

XSL•FO
RenderX

Figure 2. Screenshot of SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) code options for term “smok” when typed into EMIS (Egton Medical Information Systems) Web, showing a drop-down menu of multiple code options that start with “smok.”.

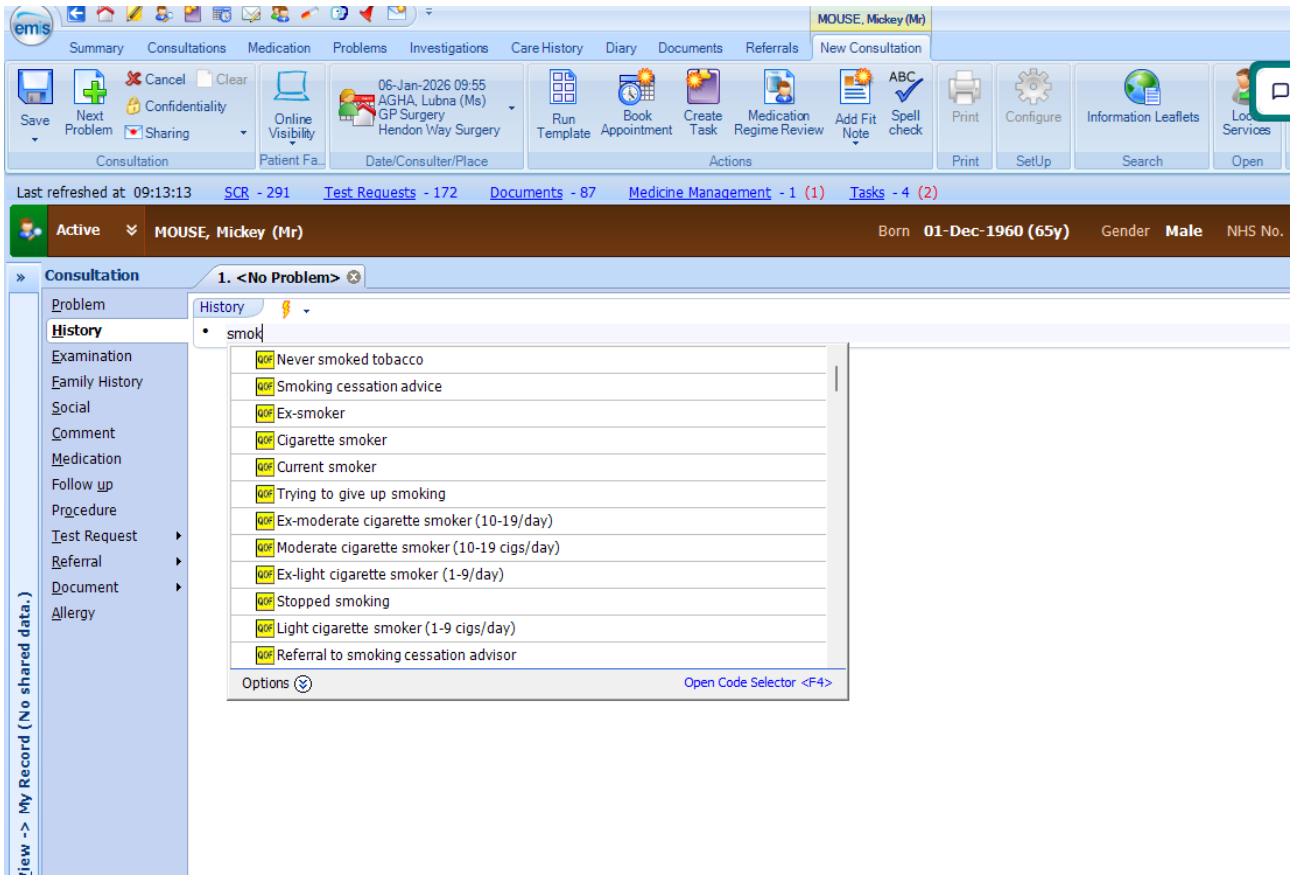

Theme 2: Screening Pathway (Breast, Bowel, and Cervical Screening)

Completeness, code diversity, and validation of cancer screening data is presented in Table 5 (Table S10 in Multimedia Appendix 1 provides additional detail on granularity).

XSL•FO
RenderX

Table 5. Descriptive analyses for data domains in Theme 2 (screening), showing completeness, code diversity, and validation of codes that were searched for in this theme.

Report name	Time period	Completeness and code diversity		Code validation	
		Completeness (% of eligible patients with code captured), n/N (%)	Number of unique SNOMED CT[a] or EMIS[b] codes in search, n	Data validation comparator source	EMIS data completeness as a proportion of comparator, n/N (%)
Cancer–bowel (aged 60-74 years; did not return screening kit)	Run date (October 2023)-2 years 6 months	65,762/188,939 (34.6)	4	No comparator for coding prevalence for kit DNRs[c], only uptake and coverage	No comparator for coding prevalence for kit DNRs, only uptake and coverage
Cancer–bowel (aged 60-74 years; screened)	Run date (October 2023)-2 years 6 months	112,939/184,323 (61.3)	10	NHS[e] Futures Screening Dashboard	61.3/62.2 (98.6)
Cancer–bowel abnormal result	Run date (October 2023)-2 years 6 months	2308/184,323 (1.3)	5	No comparator for coding prevalence for abnormal screening results, only uptake and coverage	No comparator for coding prevalence for abnormal screening results, only uptake and coverage
Cancer–breast (aged 50-70 years; screened)	Run date (October 2023)-3 years 6 months	79,321/175,986 (45.1)	30	NHS Futures Screening Dashboard	45.1/57.2 (78.8)
Cancer–breast (aged 50-70 years; abnormal result)	Run date (October 2023)-3 years 6 months	1646/114,606 (1.4)	7	No comparator for coding prevalence for abnormal screening results, only uptake and coverage	No comparator for coding prevalence for abnormal screening results, only uptake and coverage
Cancer–breast (aged 50-70 years; normal result)	Run date (October 2023)-3 years 6 months	71,043/165,126 (43)	3	No comparator for coding prevalence for normal screening results, only uptake and coverage	No comparator for coding prevalence for normal screening results, only uptake and coverage
Cancer–breast cancer detected	Run date (October 2023)-3 years 6 months	913/167,315 (0.5)	35	National Cancer Registry	913/731 (124.9)
Cancer–cervical (aged 25-49 years; adequate smear)	Run date (October 2023)-3 years 6 months	205,730/356,955 57.6%	103	NHS Futures Screening Dashboard	57.6/57.7 (99.8)
Cancer–cervical (aged 50-64 years; adequate smear)	Run date (October 2023)-5 years 6 months	97,233/139,233 (69.8)	98	NHS Futures Screening Dashboard	69.8/71 (98.3)

[a]SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms.

[b]EMIS: Egton Medical Information Systems.

[c]DNR: did not return.

[d]NHS: National Health Service.

Breast Screening

Coding for breast cancer screening uptake was recorded in 45.1% (79,321/175,986) of screening-eligible patients, with a 12% deficit compared with the comparator database. A total of 30 unique codes were used to document breast screening activity. The number of breast cancer diagnoses following abnormal results exceeded those in the comparative dataset. The most frequently used code was “Mammography normal,” accounting for 39.2% (31,109/79,321) of coded entries. The qualitative data (semistructured interviews and workshops) verified much of the persistent quality issues in the breast screening service, particularly its interface with primary care. This data also highlighted problems such as significant delays in breast screening attendance and nonattendance notifications; delivery of notifications by letters that contain multiple patients on a single sheet and therefore require manual separation; and counterintuitive patient reminders. For instance, proactive reminders for women upon turning age 50 years can cause confusion, because screening invitations may not be issued until age 53 years, with no flexibility for earlier appointments.

XSL•FO
RenderX

Bowel Cancer Screening

Codes for bowel cancer screening suggested an uptake of 61.3% (112,939/184,323). The proportion of screening-eligible patients who did not return test kits was 34.6% (65,762/188,939). Coding of abnormal results was documented in 1.3% (2308/184,323) of cases. The fecal immunochemical test (FIT), which underpins bowel screening, was consistently coded as "Bowel cancer screening program faecal occult blood test" within EMIS Web. The most frequently used codes were for normal and abnormal fecal occult blood (FOB) test results. The qualitative data (semistructured interviews) showed us that the bowel screening processes and coding practices were clearly defined for patients who undergo screening (aligned to QOF-funded activity). However, variability exists in recall, reminder, and engagement activities for nonresponders. Those who did not participate in screening were often coded using a non-SNOMED CT term, "No response to BCSP invitation."

Cervical Screening

Cervical screening data in the eligible population, coded as having had a smear, were 57.6% (205,730/356,955; in the 25 to 49 years age group) and 69.8% (97,233/139,233; in the 50 to 64 years age group). Both numbers aligned with the comparator database at over 99% (57.6%/57.7%) and 98% (69.8%/71%). A total of 98-103 individual codes were used to describe cervical screening coverage among the 2 eligible age cohorts. The qualitative data (semistructured interviews with GP federation's IT team) revealed that cervical screening processes were also clearly defined and aligned with the funding route (QOF). Variability existed in the recall and engagement activities for nonresponders. Reports for cervical nonengagement and recall were not developed because it was understood that data would not exist.

Factors Influencing Clinical Coding of Cancer Screening Data in Primary Care

Stakeholders explained that the coding of cancer screening data in primary care is influenced by primary care staff motivations and the ease of coding screening results in the system.

Relevance of Coding Cancer Screening Data to Clinical Workflow

Primary care staff described varied practices in whether cancer screening attendance was recorded. While some mentioned manual efforts or automated systems for reminding patients about upcoming screening appointments or contacting patients when they were notified of nonattendance, many admitted that they did not code screening attendance and follow up for those who did not attend. In discussing reasons for this, GPs and program managers indicated that coding was motivated by mandates and incentives (QOF) that are often only short term.

Many GPs agreed that recording screening data or contacting those who did not attend was not clinically relevant to primary care workflow and believed it was under the remit of national teams that run the programs. A few also noted that they did not have the most up-to-date information to monitor and facilitate screening attendance as they had received incorrect system prompts around screening attendance and were not aware of changes to screening eligibility. While some GPs and program managers noted local-level efforts to support and improve cancer screening attendance, there was general consensus that this is dependent on practice capacity to follow up those patients and is challenging due to competing priorities.

Ease of Coding Screening Results in Primary Care

GPs and program managers raised that coding breast cancer screening data is time consuming and demanding because paper results are sent to primary care with 2 patients' results per page, meaning they must be cut in 2 before being filed. Some recounted making requests for results to be sent electronically to streamline this process but had accepted that the system cannot be changed. GPs and program managers reported that the multitude of coding options for breast and cervical cancer screening results (eg, cervical screening, smear, cervical smear) contribute to coding inconsistencies. There was agreement that standardized coding could improve this. In contrast, stakeholders reflected that processing bowel cancer screening results is straightforward because the codes on the screening results letters are easy to match to those on the system (SNOMED CT). However, some highlighted that FOB codes (which relate to guaiac fecal occult blood testing [gFOBt] that is no longer used in the bowel screening program) [39] are still being used to code FIT screening results. Some reported that there is unified understanding in primary care that these legacy codes relate to FIT results and that incentive schemes for screening data still acknowledge these codes. However, one GP raised that legacy FOB codes may not be acknowledged in data searches for symptomatic FIT results.

Theme 3: Postcancer Diagnosis

Overview

Completeness, code diversity, and validation for postcancer codes are presented in Table 6 and in Table S12 in Multimedia Appendix 1 (Table S11 in Multimedia Appendix 1 provides additional detail on granularity).

XSL•FO
RenderX

Table 6. Descriptive analyses for data domains in Theme 3 (postcancer diagnosis), showing completeness, code diversity, and validation of codes that were searched for in this theme.

Report name	Time period	Completeness and code diversity		Code validation	
		Completeness (% of eligible patients with code captured), n/N (%)	Number of unique SNOMED CT[a] or EMIS[b] codes in search	Data validation comparator source	EMIS data completeness as a proportion of comparator, n/N (%)
Malignant neoplastic disease	New episode added last 12 months up to October 2023	6,044/1,506,746 (0.4)	677	National Cancer Registry	4604/6319 (73)
Malignancy stage	New episode added last 12 months up to October 2023	6/64,977 (0.01)	4	National Cancer Registry	6/3536 (0.2)
Treatment regime (if coded)	New episode added last 12 months up to October 2023	315/942,260 (0.03)	35	Cancer Waiting Times	315/10,574 (3)
Cancer care review	New episode added last 12 months up to October 2023	3953/1,236,580 (0.3)	1	No comparator for coding prevalence of this data item sourced	No comparator for coding prevalence of this data item sourced
Has a carer	Recorded ever	14,513/1,229,644 (1.2)	17	No comparator for coding prevalence of this data item sourced	No comparator for coding prevalence of this data item sourced

[a]SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms.

[b]EMIS: Egton Medical Information Systems.

The proportion of newly diagnosed cancers coded in EMIS was 73% (4604/6319) of the expected figure. The EMIS searches excluded deceased patients; further data comparison suggests that deceased individuals could account for 11.8%-14% of new cancer diagnoses in the Rapid Cancer Registration Database, reducing the initial coding gap from 27% to approximately 13%.

Cancer staging data showed a total completeness of 0.2% (6/3536) when compared with the validation database. One borough recorded no staging codes. Cancer treatment coding showed a total completeness of 3% (315/10,574) when compared with the validator.

CCRs were coded for 66% (3953/5982) of eligible patients, aligning with the 73% (4604/6319) of new cancer diagnoses recorded in primary care. A single SNOMED CT code is used to document CCRs. A total of 1.2% (14,513/1,229,644) of patients aged 18 years and older had a recorded carer status. The code “Has a carer” accounted for 80.1% (11,631/14,513) of these entries. Beyond CCRs, other personalized cancer care quality indicator data are shown separately in Table S12 in Multimedia Appendix 1. Cancer care plans, end-of-treatment summaries, and holistic needs assessments were recorded in 1%-2% of patients with cancer across NCL.

Factors Influencing Clinical Coding of Postcancer Diagnosis Data in Primary Care

Stakeholders suggested that coding of patients’ cancer diagnoses and treatment data is influenced by the relevance of information to the clinical workflow in primary care, the quality of information sharing from secondary care, and the complexity and consistency of the systems used for coding.

Relevance of Coding Postcancer Diagnosis Data to Clinical Workflow

A few GPs mentioned that coding of cancer diagnoses will improve when incentivized through the QOF. However, some highlighted that coding information about treatment plans organized by secondary care is not perceived as relevant or a priority in daily practice. One GP also reflected that they sometimes felt reluctant to request information from patients during CCR appointments given that patients have already had to discuss their diagnoses in secondary care.

Quality of Postcancer Diagnosis Information Sharing

GPs suggested that cancer diagnosis information is not always shared by secondary care or that it may be sent with some delay. They also reported that information may be missing for patients who are diagnosed and treated privately or those who are diagnosed at an advanced stage whereby primary care is only notified of a cancer diagnosis through the receipt of postmortem information. When cancer diagnosis information is received, there is consensus among GPs that letters from secondary care are long and complex, meaning staff have to scrutinize the whole letter to find and extract the key information. Many agreed that diagnosis information and SNOMED CT codes that require coding should be placed at the top of these letters for easy translation into primary care records. Some GPs also lack trust

XSL•FO
RenderX

in the accuracy of information received from secondary care due to finding previous errors in patient records. For this reason, a few expressed concerns about linked data between primary and secondary care if errors could not be redacted within primary care.

Consistency and Complexity of Systems Used for Coding Postcancer Diagnosis Data

GPs reported that there are multiple places within primary care–based systems (EMIS) where cancer information can be coded or entered as free text, as well as several different coding options that prevent consistent coding of cancer diagnoses. There was agreement that standardized templates were needed. GPs also reported that *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) codes used in secondary care letters are not aligned to SNOMED CT codes used in primary care, which causes ambiguity. Additionally, some explained that practices vary in their capacity to ensure the quality of coding, such as whether formalized coding teams are used and given adequate resources, time, and training to correctly code information received from secondary care.

Discussion

Summary

This QI project provides a detailed assessment of cancer-related clinical coding in NCL primary care across a population of 1.4 million adults. Our findings show that although some demographic data such as ethnicity and language are well captured (coding completeness), many other important codes across the cancer pathway—especially those relating to social determinants of health, cancer treatments, and postdiagnosis care—remain inconsistently coded or significantly absent from primary care records.

Coding quality was strongly influenced by the presence of national or local incentives (such as QOF), which drove completeness for certain indicators like ethnicity, cancer diagnosis, and CCRs. In contrast, areas not linked to performance payments or formalized data capture processes, such as cancer treatment, staging, and screening follow-up, showed substantial gaps. These differences signal that current coding behavior in primary care is shaped by system design, contractual levers, and administrative capacity.

Further qualitative insights helped contextualize these patterns, showing that GPs often prioritize coding activities relevant to their daily clinical workflow or incentivized tasks. Key barriers to good coding included complex and inconsistent coding systems, limited opportunities to update data, and limited structured information sharing from secondary care.

Overall, the findings demonstrate that improving cancer coding quality in primary care requires more than local process changes; it will require a coordinated national approach that includes clearer coding standards, automation, and alignment of incentives.

Interpretation

Our findings align with existing literature relating to cancer data coding in primary care. For example, our observation that ethnicity coding was high at 86.5% (912,679/1,055,083), is consistent with other studies using patient electronic records. This is similar to 78.2% ethnicity coding reported in NHS primary care records in 2022 [40]. These high levels may be related to incentive schemes (QOF) to improve completeness in England [41]. Similarly, 68.7% (898,023/1,307,601) completeness for language coding aligns with findings that emphasize the need for comprehensive demographic information to support better patient health outcomes [42].

The minimal recording of employment status (29,848/1,229,644, 2.4%) and the moderate completeness of BMI data (485,660/1,159,241, 41.9%) underscore challenges in capturing socioeconomic and health metrics. These gaps are consistent with literature indicating that certain health data, such as employment status, are often underreported in primary care settings [43]. Our findings on smoking status (530,720/1,185,812, 44.8%) and alcohol consumption (222,753/1,236,580, 18%) coding completeness, which show significant borough variation, also reflect widespread underrecording of these factors in patient records [44]. This may reflect discomfort in approaching conversations about lifestyle factors as well as lack of time or resources, particularly in more deprived areas [12,45].

The variation in cancer screening data across pathways, particularly the fragmented coding for breast screening, highlights the complexities in capturing screening information. Rafi et al [46] also found that a wide range of codes were used to document a family history of breast cancer in primary care, leading to inconsistencies in data quality and the potential for misclassification of risk. This finding may relate to primary care's role and responsibilities in screening, where their involvement varies significantly across different cancer types [47]. While primary care plays a structured role in bowel and cervical screening (underpinned by funding and agreed processes nationally), breast cancer screening activity is not delivered or incentivized in primary care.

New cancer diagnoses were also undercoded, at 87% (4604/5260) completeness. This aligns with findings from the Netherlands, where only 60.6% of cancer cases were coded according to the national registry [9]. Undercoding is linked to reliance on unstructured secondary care letters and a lack of coding incentives [8,9]. Cancer treatment and staging data were also undercoded, and there was not any existing evidence to compare against.

The recording of CCRs (3953/5982, 66% of eligible patients) closely mirrors the proportion of coded diagnoses, suggesting a correlation between diagnosis coding and care review documentation. However, concerns exist that high completeness rates reflect a "tick-box" approach driven by financial incentives for coding "cancer care review" [48]. Documentation of broader personalized cancer care indicators, such as care plans and holistic needs assessments, was minimal (1%-2%), highlighting gaps in comprehensive cancer care documentation. This aligns with studies calling for broader, patient-centered metrics to sustain confidence in cancer registries [49]. National CCR templates embedded in GP electronic records [50] support additional quality indicator completion alongside CCRs,

XSL•FO
RenderX

promoting standardization [51]. The lack of broader care indicators being coded suggests that these templates are not embedded in daily practice.

While this project was conducted in NCL, the findings have broader relevance across other similar primary care settings. Although population demographics vary between regions, the core structures, processes, and incentives that shape GP coding behaviors are largely consistent across the NHS and similar international health care systems. Primary care operates within a standardized framework of national policies, contractual obligations, and clinical systems, meaning that the challenges and opportunities identified in this study are likely to be mirrored elsewhere. As such, our findings provide valuable insights for informing coding improvement initiatives beyond our study region, with potential applicability across primary care systems nationally.

Limitations

Two boroughs were excluded due to nonengagement; however, the dataset remained robust with high completeness in other areas, minimizing bias. The findings should be interpreted cautiously for nonsubmitting practices. Additionally, our quantitative data searches were completed by GP Federation's IT team, but their searches could not fully align with our specifications. This left gaps in key cancer care metrics, such as patients with cancer on stratified follow-up pathways or active surveillance; alternative data sources such HealtheIntent (the ICBs' provider for linked datasets across the NHS in NCL until September 2025), helped fill in some gaps.

Conclusions

The QI project has provided a unique and detailed insight into the many dimensions of cancer coding across the whole pathway in primary care and sheds light on many factors that underpin variation and coding preference.

We have developed recommendations based on our findings aimed at primary care providers, commissioners, ICBs' digital teams [52], cancer screening teams and the National Cancer team [53].

Implications for Practice

Textbox 1 outlines practical recommendations to enhance primary care data management, integration with secondary care, and overall service quality. The focus is on standardizing coding practices, improving information flow, and leveraging data for informed decision-making and patient care improvements.

XSL•FO
RenderX

Textbox 1. Implications for practice.

Strengthen data infrastructure in primary care

- Develop a structured data framework using SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) with a minimum dataset and distinct code sets, aligning with secondary care processes. This will enable better use of primary care data for analysis, epidemiology, and health care improvement.

Enhance breast screening integration with primary care

- Transition to electronic breast screening outcome reports for general practitioners (GPs), standardize national breast screening activity codes within SNOMED CT, and define and resource primary care's role in patient follow-up and nonattendance intervention. These changes will improve the accuracy and efficiency of breast screening data and ensure clear responsibility for patient engagement.

Improve secondary care information for primary care

- Standardize hospital discharge letters to clearly indicate new diagnoses, multidisciplinary team outcomes, and care plans, ensure clinical coding aligns with SNOMED CT; prioritize consistent implementation of end-of-treatment summaries; and link secondary care data directly to primary care [27-29]. This recommendation aligns with the national strategy [54,55]. This will reduce duplication, enhance continuity of care, and improve data quality for patient management.

Standardize primary care coding practices

- Employ coding professionals in primary care, align screening activity codes, update bowel screening codes to reflect the transition from fecal occult blood to fecal immunochemical tests, and educate the workforce on the value of high-quality coding. These steps will improve consistency, accuracy, and the usefulness of coded data for patient care and service planning.

Optimize patient registration processes

- Automate the integration of coded demographic, behavioral, and risk factor data from National Health Service Digital's updated registration forms (Patient Registration Form 1) [56] into primary care systems and allocate resources for implementation. This will ensure more complete and accurate patient information from the outset.

Improve Quality Outcomes Framework rules and transparency

- Make Quality Outcomes Framework rule changes trackable over time and provide clearer navigation and updates [57,58]. This will help those involved in service improvement and research to understand and respond to coding changes more effectively.

Implement primary care coding audits

- Introduce National Health Service England–funded coding audits, assess data quality and completeness, flag nonrecommended codes, and cross-reference with national datasets. This will improve coding accuracy, highlight inconsistencies, and enhance data reliability.

Develop an analytics dashboard

- Create a live dashboard to track trends, profile data, and support quality improvement, leveraging the London Health Data Service (launched in June 2025) [59]. This will provide real-time insights into primary care data, supporting better decision-making and service planning.

Facilitate knowledge sharing

- Identify regions with superior data completeness and share successful quality improvement initiatives across London. This will promote best practices and drive improvements in data quality and patient care across the system.

Funding

This study did not receive external funding from any public, commercial, or not-for-profit sources. The Enfield Federation was commissioned to build the data searches, run these across NCL GP practices, and extract the data. SB's time was funded by the NCL Cancer Alliance to support the qualitative analysis. GB acknowledges funding from Barts Charity. No funding body was involved in study design, data collection, analysis, interpretation, or manuscript preparation.

Acknowledgments

The authors thank the Enfield general practitioner (GP) Federation data team, General Practice Provider Alliance, North Central London (NCL) GP practices, all workshop participants, patient partners, the internal NCL Cancer Alliance team working group, the NCL Integrated Care Board digital team, and all stakeholders who reviewed and engaged in this study.

We used ChatGPT (free version; OpenAI) [60] to support the manuscript writing process. It was used to optimize sentence structure, readability, and reduce text burden. All content, analyses, and interpretations were developed, verified, and approved by the authors. ChatGPT was not used for data generation, analysis, or creation of original scientific content. The artificial

XSL•FO
RenderX

intelligence (AI) use for this manuscript falls under Domain 1 of the "Recommendations for a Classification of AI Use in Academic Manuscript Preparation" document.

The project was delivered as part of service improvement and evaluation led by the NCL Cancer Alliance.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

AB conceived and led the project, collated a significant portion of the mixed methods data, supported analysis, and drafted the initial manuscript. GR conducted data collation and analysis, identified comparative data sources, and contributed substantially to manuscript development. KT provided project support and made significant contributions to the manuscript. SB analyzed the qualitative data and drafted the corresponding sections. GB provided overall guidance and methodological support, particularly in qualitative data collection and analysis. She advised on this study's structure and contributed significantly to manuscript development. All authors contributed to the analysis, interpretation of findings, and manuscript revisions. All authors reviewed and approved the final version for submission. AB is the guarantor of this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary (clean) 30.10.25.
[DOCX File , 11440 KB - jmir_v28i1e73205_app1.docx]

References

1. High quality patient records. NHS England. URL: https://www.england.nhs.uk/long-read/high-quality-patient-records/ [accessed 2025-01-06]
2. Taylor K, Henderson S. 112 Challenges of clinical coding: an audit of the accuracy of primary diagnosis coding in a specialist palliative care unit. BMJ Supportive & Palliative Care 2019;9(Suppl 1):A49-A50. [doi: 10.1136/bmjspcare-2019-asp.135]
3. Harry ML, Truitt AR, Saman DM, Henzler-Buckingham HA, Allen CI, Walton KM, et al. Barriers and facilitators to implementing cancer prevention clinical decision support in primary care: a qualitative study. BMC Health Serv Res 2019;19(1):534 [FREE Full text] [doi: 10.1186/s12913-019-4326-4] [Medline: 31366355]
4. Greenwood-Lee J, Jewett L, Woodhouse L, Marshall DA. A categorisation of problems and solutions to improve patient referrals from primary to specialty care. BMC Health Serv Res 2018;18(1):986 [FREE Full text] [doi: 10.1186/s12913-018-3745-y] [Medline: 30572898]
5. Morrison Z, Fernando B, Kalra D, Cresswell K, Sheikh A. National evaluation of the benefits and risks of greater structuring and coding of the electronic health record: exploratory qualitative investigation. J Am Med Inform Assoc 2014;21(3):492-500 [FREE Full text] [doi: 10.1136/amiajnl-2013-001666] [Medline: 24186957]
6. Pascoe SW, Neal RD, Heywood PL, Allgar VL, Miles JN, Stefoski-Mikeljevic J. Identifying patients with a cancer diagnosis using general practice medical records and cancer registry data. Fam Pract 2008;25(4):215-220. [doi: 10.1093/fampra/cmn023]
7. Strategies for success: tackling common clinical documentation integrity challenges head-on. Healthcare Financial Management Association. 2024. URL: https://www.hfma.org/revenue-cycle/strategies-for-success-tackling-common-clinical-documentation-integrity-challenges-head-on/ [accessed 2025-09-01]
8. Sollie A, Sijmons RH, Helsper C, Numans ME. Reusability of coded data in the primary care electronic medical record: a dynamic cohort study concerning cancer diagnoses. Int J Med Inform 2017;99:45-52. [doi: 10.1016/j.ijmedinf.2016.08.004] [Medline: 28118921]
9. Sollie A, Roskam J, Sijmons RH, Numans ME, Helsper CW. Do GPs know their patients with cancer? Assessing the quality of cancer registration in Dutch primary care: a cross-sectional validation study. BMJ Open 2016;6(9):e012669 [FREE Full text] [doi: 10.1136/bmjopen-2016-012669] [Medline: 27633642]
10. Zghebi SS, Reeves D, Grigoroglou C, McMillan B, Ashcroft DM, Parisi R, et al. Clinical code usage in UK general practice: a cohort study exploring 18 conditions over 14 years. BMJ Open 2022;12(7):e051456 [FREE Full text] [doi: 10.1136/bmjopen-2021-051456] [Medline: 35879012]
11. Thiru K, Hassey A, Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. BMJ 2003;326(7398):1070 [FREE Full text] [doi: 10.1136/bmj.326.7398.1070] [Medline: 12750210]
12. Davies A, Ahmed H, Thomas-Wood T, Wood F. Primary healthcare professionals' approach to clinical coding: a qualitative interview study in Wales. Br J Gen Pract 2024;75(750):e43-e49. [doi: 10.3399/bjgp.2024.0036]

XSL•FO
RenderX

13. Collaço N, Lippiett KA, Wright D, Brodie H, Winter J, Richardson A, et al. Barriers and facilitators to integrated cancer care between primary and secondary care: a scoping review. Support Care Cancer 2024;32(2):120 [FREE Full text] [doi: 10.1007/s00520-023-08278-1] [Medline: 38252169]
14. National Cancer Registration and Analysis Service (NCRAS). GOV.UK. URL: https://www.gov.uk/guidance/national-cancer-registration-and-analysis-service-ncras [accessed 2025-09-03]
15. Datasets in support of statutory duties. URL: https://digital.nhs.uk/services/data-services-for-commissioners/datasets-in-support-of-statutory-duties [accessed 2025-09-08]
16. Quality and Outcomes Framework (QOF). URL: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/general-practice-data-hub/quality-outcomes-framework-qof [accessed 2024-09-30]
17. Navigating the New QOF Landscape: Your Guide to the 2025/26 Changes for GP Practices. URL: https://www.mypracticemanager.co.uk/advice/qof-2025-26-changes [accessed 2025-09-11]
18. Tai TW, Anandarajah S, Dhoul N, de Lusignan S. Variation in clinical coding lists in UK general practice: a barrier to consistent data entry? Inform Prim Care 2007;15(3):143-150 [FREE Full text]
19. Cancer: find out more about the National Cancer Registration and Analysis Service (NCRAS) - one of the 2 disease registration services within the National Disease Registration Service (NDRS). NDRS. URL: https://digital.nhs.uk/ndrs/about/ncras [accessed 2025-01-11]
20. NHS Digital. The NHS Digital SNOMED CT Browser. URL: https://termbrowser.nhs.uk/ [accessed 2025-01-06]
21. NHS England. The Organisation has a Framework in Place to Support Lawfulness, Fairness and Transparency (1.1). URL: https://tinyurl.com/yc4pnfdt [accessed 2024-02-22]
22. NHS England. Getting It Right First Time - GIRFT. 2020. URL: https://gettingitrightfirsttime.co.uk/ [accessed 2025-01-06]
23. Doktorchik C, Lu M, Quan H, Ringham C, Eastwood C. A qualitative evaluation of clinically coded data quality from health information manager perspectives. Health Inf Manag 2020;49(1):19-27. [doi: 10.1177/1833358319855031] [Medline: 31284769]
24. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
25. How to make a case for change. StaffNet. URL: https://www.staffnet.manchester.ac.uk/od/learning-pathways/professional-and-technical-development/change-management/change-toolkit/how-to/how-to-make-a-case-for-change/ [accessed 2025-11-29]
26. Optum. EMIS Web: Comprehensive, flexible and powerful. URL: https://www.emishealth.com/products/emis-web [accessed 2025-01-23]
27. Arora A, Alderman JE, Palmer J, Ganapathi S, Laws E, McCradden MD, et al. The value of standards for health datasets in artificial intelligence-based applications. Nat Med 2023;29(11):2929-2938 [FREE Full text] [doi: 10.1038/s41591-023-02608-w] [Medline: 37884627]
28. Optum. SNOMED CT and EMIS codes. URL: https://www.emisnow.com/csm?id=kb_article_view&table=kb_knowledge&sys_kb_id=816eb8833b0b5a907290373a85e45a41&searchTerm=emis%20web%20codes [accessed 2025-02-10]
29. NHS England. Read Codes: Retirement of Read Version 2 and Clinical Terms Version 3. URL: https://tinyurl.com/yc7ckkzx [accessed 2016-04-01]
30. What is a GP Federation? GP Health Connect. URL: https://www.gphealthconnect.co.uk/what-is-a-gp-federation/ [accessed 2025-01-06]
31. HealtheIntent. URL: https://nclhealthandcare.org.uk/digital/digital-information-for-health-and-care-professionals/healtheintent/ [accessed 2025-09-03]
32. Pineda-Moncusí M, Allery F, Delmestri A, Bolton T, Nolan J, Thygesen JH, CVD-COVID-UK/COVID-IMPACT Consortium. Ethnicity data resource in population-wide health records: completeness, coverage and granularity of diversity. Sci Data 2024;11(1):221 [FREE Full text] [doi: 10.1038/s41597-024-02958-1] [Medline: 38388690]
33. NHS England. DAPB1521: Cancer Outcomes and Services Data Set. URL: https://tinyurl.com/ypmss6wt [accessed 2023-09-11]
34. Systemic Anti-Cancer Therapy (SACT) Data Set. URL: https://tinyurl.com/j6km6srp [accessed 2025-11-03]
35. Prevention, awareness and screening. North Central London Cancer Alliance. 2025. URL: https://www.nclcanceralliance.nhs.uk/our-work/prevention-awareness-and-screening/ [accessed 2025-09-11]
36. Bhuiya A, Cavanagh S, Nestor C, Fomina M, Ahmed I, Von Wagner C, et al. Development of a cancer pathway support guide for patients and carers: a codesign project. European Journal of Cancer Care 2024;2024:3623136. [doi: 10.1155/2024/3623136]
37. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013;13(1):117. [doi: 10.1186/1471-2288-13-117]
38. NHS England. Quality and Outcomes Framework (QOF) business rules. URL: https://www.google.com/url?q=https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/quality-and-outcomes-framework-qof/business-rules&sa=D&source=docs&ust=1739198517910415&usg=AOvVaw0gPCPGkD2artCnS-Cu87aB [accessed 2025-04-08]

XSL•FO
RenderX

39. National Institute for Health and Care Excellence. The NHS Bowel Screening Programme. URL: https://cks.nice.org.uk/topics/bowel-screening/background-information/the-nhs-bowel-screening-programme/ [accessed 2025-02-10]
40. Andrews CD, Mathur R, Massey J, Park R, Curtis HJ, Hopcroft L, et al. Consistency, completeness and external validity of ethnicity recording in NHS primary care records: a cohort study in 25 million patients' records at source using OpenSAFELY. BMC Med 2024;22(1):288 [FREE Full text] [doi: 10.1186/s12916-024-03499-5] [Medline: 38987774]
41. NHS England. Information Pack for GPs. URL: https://www.england.nhs.uk/london/wp-content/uploads/sites/8/2023/02/GP-Information-Pack-collection-of-patient-ethnicity-to-inform-resource-allocation-November-2022-Public.pdf [accessed 2025-11-29]
42. Cano-Ibáñez N, Zolfaghari Y, Amezcua-Prieto C, Khan KS. Physician-patient language discordance and poor health outcomes: a systematic scoping review. Front Public Health 2021;9:629041 [FREE Full text] [doi: 10.3389/fpubh.2021.629041] [Medline: 33816420]
43. Campbell P, Lewis M, Chen Y, Lacey RJ, Rowlands G, Protheroe J. Can patients with low health literacy be identified from routine primary care health records? A cross-sectional and prospective analysis. BMC Fam Pract 2019;20(1):101 [FREE Full text] [doi: 10.1186/s12875-019-0994-8] [Medline: 31319792]
44. Al Kazzi E, Lau B, Li T, Schneider E, Makary M, Hutfless S. Critical problems of coding data in health care: obesity, smoking, and alcohol use by method of measurement. Value in Health 2015;18(3):A2. [doi: 10.1016/j.jval.2015.03.007]
45. Thelen R, Bhatti S, Rayner J, Grudniewicz A. Collecting sociodemographic data in primary care: qualitative interviews in community health centres. BJGP Open 2025;9(1):BJGPO.2024.0095 [FREE Full text] [doi: 10.3399/BJGPO.2024.0095] [Medline: 39528270]
46. Rafi I, Chowdhury S, Chan T, Jubber I, Tahir M, de Lusignan S. Improving the management of people with a family history of breast cancer in primary care: before and after study of audit-based education. BMC Fam Pract 2013;14(1):105 [FREE Full text] [doi: 10.1186/1471-2296-14-105] [Medline: 23879178]
47. Green T, Atkin K, Macleod U. Cancer detection in primary care: insights from general practitioners. Br J Cancer 2015;112(Suppl 1):S41-S49 [FREE Full text] [doi: 10.1038/bjc.2015.41] [Medline: 25734388]
48. Gopal DP, de Rooij BH, Ezendam NP, Taylor SJ. Delivering long-term cancer care in primary care. Br J Gen Pract 2020;70(694):226-227. [doi: 10.3399/bjgp20x709481]
49. Shokrizadeharani L, Batooli Z, Heydarian S. Evaluation of completeness, comparability, validity, and timeliness in cancer registries: a scoping review. Studies in health technology and informatics 2023:305. [doi: 10.3233/shti230451]
50. Macmillan Cancer Support. Cancer Care Review. URL: https://tinyurl.com/y8b5p8xw [accessed 2025-09-16]
51. Macmillan Cancer Support. Practical Implementation Guide for Cancer Care Reviews. URL: https://www.macmillan.org.uk/healthcare-professionals/cancer-pathways/prevention-and-diagnosis/cancer-care-review [accessed 2025-02-11]
52. NHS England. What are Integrated Care Systems?. URL: https://www.england.nhs.uk/integratedcare/what-is-integrated-care/ [accessed 2025-02-11]
53. NHS England. Cancer. URL: https://www.england.nhs.uk/cancer/ [accessed 2025-02-11]
54. NHS England. Clinical Coding - SNOMED CT. URL: https://www.england.nhs.uk/long-read/clinical-coding-snomed-ct/ [accessed 2025-01-16]
55. Macmillan Cancer Support. Personalised Care for People Living With Cancer. URL: https://www.macmillan.org.uk/healthcare-professionals/innovation-in-cancer-care/personalised-care [accessed 2025-01-16]
56. Londonwide LMCs. Contract Changes for 2024/25. URL: https://www.lmc.org.uk/news/contract-changes-2024-25-and-online-registration-requirement/ [accessed 2024-10-23]
57. NHS England. Questions Patients are Asked in the Online form. URL: https://digital.nhs.uk/services/register-with-a-gp-surgery-service/get-help-using-the-service/questions-asked [accessed 2025-04-14]
58. NHS England. Quality and Outcomes Framework (QOF) business rules v48.0 2023-2024. URL: https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/quality-and-outcomes-framework-qof/quality-and-outcome-framework-qof-business-rules/quality-and-outcomes-framework-qof-business-rules-v48.0-2023-2024 [accessed 2024-01-16]
59. One London. London Secure Data Environment. 2025. URL: https://www.onelondon.online/london-secure-data-environment/ [accessed 2025-09-16]
60. Chat GPT. URL: https://chatgpt.com/ [accessed 2025-11-29]

Abbreviations

CCR: cancer care review
COREQ: Consolidated Criteria for Reporting Qualitative Research
EMIS: Egton Medical Information Systems
FIT: fecal immunochemical test
FOB: fecal occult blood
gFOBt: guaiac fecal occult blood testing
GP: general practitioner

XSL•FO
RenderX

ICB: integrated care board
ICD-10: International Statistical Classification of Diseases, Tenth Revision
NCL: North Central London
NCRAS: National Cancer Registration and Analysis Service
NHS: National Health Service
PCN: primary care network
QI: quality improvement
QOF: Quality Outcomes Framework
SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms

Edited by T de Azevedo Cardoso, A Stone; submitted 20.Mar.2025; peer-reviewed by W Gray, N Calanzani, E Whitfield; comments to author 25.Jul.2025; accepted 10.Nov.2025; published 07.Jan.2026.

Please cite as:
Bhuiya A, Roberts G, Tucker K, Bonfield S, Black G
Quality of Cancer-Related Clinical Coding in Primary Care in North Central London: Mixed Methods Quality Improvement Project
J Med Internet Res 2026;28:e73205
URL: https://www.jmir.org/2026/1/e73205
doi:10.2196/73205
PMID:

©Afsana Bhuiya, Graham Roberts, Katie Tucker, Stefanie Bonfield, Georgia Black. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Multimodal Large Language Models for Cystoscopic Image Interpretation and Bladder Lesion Classification: Comparative Study

Yung-Chi Shih[1], MD; Cheng-Yang Wu[2], MD; Shi-Wei Huang[2,3*], MD, PhD; Chung-You Tsai[1,3,4*], MD, PhD

[1]Division of Urology, Department of Surgery, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[2]Department of Urology, National Taiwan University Hospital, Yunlin Branch, Yunlin, Taiwan
[3]Department of Urology, College of Medicine, National Taiwan University, Taipei, Taiwan
[4]Department of Electrical Engineering, Yuan Ze University, Taoyuan, Taiwan
[*]these authors contributed equally

Corresponding Author:
Chung-You Tsai, MD, PhD
Division of Urology
Department of Surgery
Far Eastern Memorial Hospital
No 21, Sec 2, Nanya S Rd., Banciao Dist.
New Taipei City, 220
Taiwan
Phone: 886 2 8966 7000
Fax: 886 2 8966 5567
Email: pgtsai@gmail.com

Abstract

Background: Cystoscopy remains the gold standard for diagnosing bladder lesions; however, its diagnostic accuracy is operator dependent and prone to missing subtle abnormalities such as carcinoma in situ or misinterpreting mimic lesions (tumor, inflammation, or normal variants). Artificial intelligence–based image-analysis systems are emerging, yet conventional models remain limited to single tasks and cannot produce explanatory reports or articulate diagnostic reasoning. Multimodal large language models (MM-LLMs) integrate visual recognition, contextual reasoning, and language generation, offering interpretive capabilities beyond conventional artificial intelligence.

Objective: This study aims to rigorously evaluate state-of-the-art MM-LLMs for cystoscopic image interpretation and lesion classification using clinician-defined stress-test datasets enriched with rare, diverse, and challenging lesions, focusing on diagnostic accuracy, reasoning quality, and clinical relevance.

Methods: Four MM-LLMs (OpenAI-o3 and ChatGPT-4o [OpenAI]; Gemini 2.5 Pro and MedGemma-27B [Google]) were evaluated under blinded, randomized procedures across two tasks: (1) free-text image interpretation for anatomic site, findings, lesion reasoning, and final diagnosis (n=401) and (2) seven-class tumor-like lesion classification (n=113) within a multiple-choice framework (cystitis, polyps, papilloma, papillary urothelial carcinoma, carcinoma in situ, non-urothelial carcinoma, and none of the above). Three raters independently scored outputs using a 5-point Likert scale, and classification metrics (accuracy, sensitivity, specificity, Youden J index (Youden J), and Matthews correlation coefficient [MCC]) were calculated for lesion detection, biopsy indication, and malignancy endpoints. For optimization, model performance was compared between zero-shot and text-based in-context learning prompts that were prefixed with brief descriptions of tumor features.

Results: The 401-image test set spanned 40 subcategories, with 322 (80.3%) containing abnormal findings in the image interpretation task. OpenAI-o3 demonstrated strong reasoning, with high satisfaction for anatomy (339/401, 84.5%) and findings (305/401, 76%), but lower satisfaction for lesion reasoning (211/401, 52.5%) and final diagnosis (193/401, 48.2%), indicating increasing difficulty with higher-order synthesis. Mean Likert score differences (OpenAI-o3 minus Gemini 2.5 Pro) were +0.27 for findings (adjusted *P* value: q=0.002), +0.24 for lesion reasoning (q=0.047), and +0.19 for final diagnosis. For clinically relevant endpoints in the full set, OpenAI-o3 achieved the most balanced performance, with lesion detection accuracy of 88.3%, sensitivity of 92%, specificity of 73.1%, Youden J of 0.650, and MCC of 0.635. In 7-class tumor-like lesion classification, OpenAI-o3 achieved accuracies of 73.5% for biopsy indication and 62.8% for malignancy, with a balanced sensitivity-specificity

XSL•FO
RenderX

trade-off, outperforming other models. Notably, OpenAI-o3 performed best on prevalent malignant lesions. ChatGPT-4o and Gemini 2.5 Pro showed high sensitivity but low specificity, whereas MedGemma-27B underperformed. In-context learning improved OpenAI-o3 microaverage accuracy (40.7%→46.0%; MCC 0.311→0.370) but yielded only slight specificity gains and minimal accuracy change in other models, likely constrained by the absence of paired image-text context.

Conclusions: MM-LLMs demonstrate meaningful assistive potential in generating interpretable cystoscopy free-text rationales and supporting biopsy triage and training. However, performance in difficult differential diagnoses remains modest and requires further optimization before safe clinical integration.

(J Med Internet Res 2026;28:e87193) doi:10.2196/87193

KEYWORDS

multimodal; large language model; AI; cystoscopy; diagnostic reasoning; finding description; biopsy indication; bladder tumor; artificial intelligence

Introduction

Cystoscopy is one of the most frequently performed procedures in urology [1]. Its effectiveness heavily depends on the urologist's experience, attention to detail, and interpretive skill, making it both technically and diagnostically challenging [2]. Interobserver variability is common, and lesion characterization (tumor vs inflammation vs normal variant) is not always straightforward, often requiring clinical correlation. Bladder cancer, the ninth most common cancer globally [3], relies heavily on cystoscopy as the cornerstone for diagnosis, treatment, and surveillance. However, studies report false-negative rates ranging from 10%-40%, with white-light cystoscopy missing up to one-third of carcinoma in situ (CIS) cases and frequently overlooking small tumors [4]. Accordingly, cystoscopic interpretation is a nuanced clinical process.

Artificial intelligence (AI)-assisted cystoscopic diagnosis and decision-making can be decomposed into distinct tasks: lesion detection (present vs absent), lesion classification, margin segmentation, descriptive reporting, biopsy triage, final diagnosis, and ultimately full report generation. Each task places different demands on algorithms, ranging from visual localization to semantic reasoning and clinical judgment. Previous work in cystoscopy has predominantly framed the problem as image classification or segmentation [5-9], often using specialized vision pipelines that localize or outline lesions but provide limited clinical context and have uncertain generalizability across morphology-diverse appearances.

Evidence from other endoscopic domains provides a useful benchmark. Task-tuned computer-aided detection systems in colonoscopy, for example, improve clinically meaningful endpoints such as polyp or adenoma detection in randomized and real-world settings; however, these gains are achieved by narrowly optimized, single-purpose models rather than by systems capable of broader interpretive reasoning [10-14].

Against this background, multimodal large language models (MM-LLMs) hold substantial potential [15]. By jointly processing images and text, MM-LLMs can, in principle, "see and say": integrate visual features with medical knowledge, generate free-text rationales, and condition decisions on clinical context [16]. Early reports suggest encouraging aggregate performance, but also reveal marked variability across lesions and tasks, indicating a role as assistive rather than autonomous readers at present [17].

Key gaps remain. First, it is unclear how state-of-the-art (SOTA) MM-LLMs perform on morphology-diverse, clinically difficult cystoscopic images curated as a stress test by domain experts. Second, the alignment between their free-text reasoning and expert judgment has not been systematically examined. Third, the practical utility of in-context learning (ICL) in cystoscopy—without task-specific fine-tuning—remains uncertain [18].

To address these gaps, our goal was to characterize the current capabilities and limitations of MM-LLMs in cystoscopic interpretation and to outline directions for model strengthening and additional adaptations required for safe clinical adoption.

Methods

Overview

Building on this objective, we (1) constructed a clinician-defined stress test that reflects real-world interpretive difficulty and spans benign and malignant lesions; (2) implemented a rater-blinded, model-anonymized evaluation across 2 complementary tasks—free-text image interpretation (4 open-ended questions plus a binary lesion detection query) and structured 7-class lesion classification; (3) mapped model outputs to clinically actionable binary endpoints (biopsy indication and malignancy); and (4) quantified the incremental benefit of ICL over zero-shot prompting. The overall study workflow is provided in Figure 1.

XSL•FO
RenderX

Figure 1. Study flow diagram. Evaluation pipeline for cystoscopic image interpretation and tumor-like lesion classification using 4 multimodal large language models (MM-LLMs). ICL: in-context learning; MCC: Matthews correlation coefficient; NPV: negative predictive value; PPV: positive predictive value.

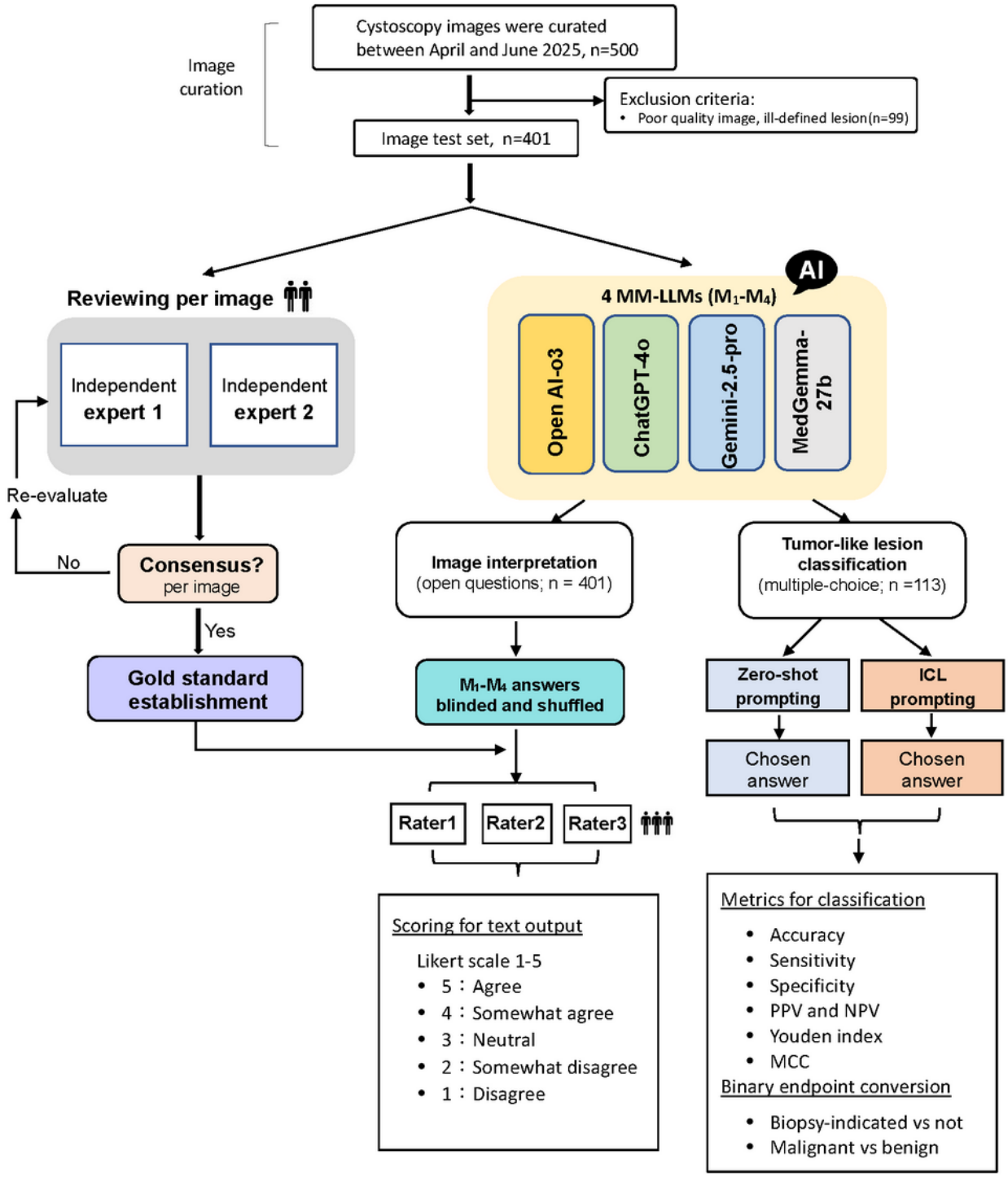

Curation of a Diverse Image Test Set

Inclusion Criteria and Diverse Lesion Coverage

We evaluated vision-enabled LLMs for cystoscopic image interpretation and tumor-like lesion classification. To stress-test model generalization, the test set was curated to maximize the diversity of morphological patterns rather than mirror clinical prevalence. Images were included only if they mapped to a prespecified schema of lower urinary tract presentations—normal anatomy, intraluminal nonmucosal conditions, and focal mucosal lesions—with finer-grained sublabels (eg, verumontanum, trabeculation, and papillary urothelial carcinoma [pUC]). To extend beyond routine cases, we deliberately sampled uncommon entities encountered in

XSL•FO
RenderX

practice, including endometriosis, miscellaneous deposits, fistulas, and erosion-related changes. This approach yielded a corpus with broad lesion coverage suitable for rigorous stress-testing cystoscopic interpretation.

Sources and Image Preprocessing

Images were curated between April and June 2025 from five sources: (1) an industry archive of intra-operative images captured on Karl-Storz endoscopes, (2) reference atlases (eg, Springer's *Diagnostic Cystoscopy* [19] and other urologic textbooks), (3) deidentified cystoscopy images obtained from websites, (4) open-access repositories accompanying PubMed-indexed papers and public datasets, and (5) Creative Commons–licensed surgical or teaching videos hosted on YouTube (Google). A total of 500 images were curated, and 99 were excluded due to poor image quality or an ill-defined lesion. The distribution of image sources and the memorization-test results, which were used to evaluate potential data-leakage risk from overlap with the 4 models' pretraining corpora, are summarized in Table S1 in Multimedia Appendix 1. Raw images were center-cropped to a square aspect ratio, resized to 800×800 pixels, and saved as JPEG files. This standardized pipeline harmonized the field of view and resolution across heterogeneous sources and ensured uniform inputs for all downstream model evaluations.

Multimodal LLMs

The evaluated MM-LLMs comprised 3 general-purpose models and 1 open-weight, medical-specific baseline. Two reasoning-optimized models—OpenAI-o3 [20] and Google Gemini 2.5 Pro [21]—were selected for their native image processing capabilities and emphasis on multistep reasoning. These represent the SOTA reasoning MM-LLMs available before July 2025. ChatGPT-4o (OpenAI) was included as a general-purpose, nonreasoning MM-LLM optimized for everyday assistance and among the earliest widely deployed models capable of accepting image input. MedGemma-27B (Google) [22] was intentionally included as a local and open-weight baseline—an open-source medical model (~27B parameters) suited for on-premises deployment and potential clinical fine-tuning. Because its parameter count and training budget are substantially smaller than the proprietary models (undisclosed), MedGemma-27B serves as a baseline rather than a capacity-matched comparator. Collectively, these systems span closed-source production platforms and an open-weight medical baseline, enabling a balanced, transparent comparison.

Study Design

Establishment of the Gold Standard

The reference standard diagnoses were determined through a multiphase, consensus-based process. Two urological experts, each with more than 25 years of clinical experience, independently reviewed all cystoscopic images, blinded to each other's assessments. The initial interexpert agreement was satisfactory (Cohen κ=0.81). In cases of disagreement, a consensus meeting was convened to establish a single unified diagnosis, integrating both normal anatomical features and pathological findings.

Image Interpretation and Lesion Classification Tasks

For a comprehensive evaluation, we designed 2 complementary tasks that reflect distinct components of diagnostic reasoning. The first, the image interpretation task, evaluated each model's capacity for domain-specific interpretation, logical reasoning, descriptive accuracy, and clinical judgment. The second, the lesion classification task, assessed the model's discriminative performance in differential diagnosis. Together, these 2 tasks provided a systematic assessment of MM-LLMs in both free-form interpretation and constrained classification settings, thereby capturing complementary dimensions of clinical decision-making.

Image Interpretation Task

The task used a structured, stepwise, open-ended question format. Each model was primed with a role-based instruction ("Suppose you are a urologist") and prompted with 4 sequential open-ended questions addressing anatomical site (Q1), findings (Q2), lesion reasoning (Q4), and final diagnosis (Q5). Q3 was not an open-ended interpretation item; it was a binary lesion-detection query (present or absent) embedded in the Q1→Q5 chain-of-thought to assess abnormal-versus-normal detection and was automatically graded against the gold standard. Free-text outputs for Q1, Q2, Q4, and Q5 were independently assessed by 3 raters (urology residents with 2-5 years of cystoscopy experience) using a 5-point Likert scale (1=disagree, 2=somewhat disagree, 3=neutral, 4=somewhat agree, and 5=agree).

Blinded and Randomized Evaluation Procedures

A dedicated evaluation software was developed to ensure complete rater blinding and randomization of both image presentation and model output order (Figure S1 in Multimedia Appendix 1).

- Image-level randomization: the display order of images was randomized once and shared across all raters. Each evaluation screen presented only 1 cystoscopic image at a time.
- Structured display: the upper panel displayed the image and its gold-standard answers to 5 reference questions (4 open-ended and 1 binary detection). The lower panel simultaneously presented anonymized text responses from the 4 MM-LLMs.
- Model-level randomization and anonymization: for each image, the order of model outputs was independently shuffled to minimize position bias. Model identities were fully anonymized to raters.
- Scoring process: raters independently scored the free-text responses using the 5-point Likert scale. The binary lesion-detection item (Q3) was automatically graded against the reference standard and was not rated by humans.

Lesion Classification Task

The lesion classification task was conducted to evaluate the models' discriminative capacity. It simulates a clinical scenario in which a urologist has already identified a tumor-like lesion and requires a differential diagnosis. A subset of tumor-like lesion images was used for this analysis. The task involved a 7-class multiple-choice framework comprising cystitis, polyps,

XSL•FO
RenderX

papilloma, pUC, CIS, nonurothelial carcinoma (non-U Ca), and none of the above (NOTA). Models were tested under 2 prompting strategies: zero-shot prompting and ICL. In the ICL condition, a text-based description of tumor-related features was incorporated into the prompt. This task aimed to assess each model's discriminative performance, adaptability to structured clinical classification, and robustness across prompting paradigms.

A subset of tumor-like lesion images was used for this task. The 7-class classification included cystitis, polyps, papilloma, pUC, CIS, non-U Ca, and NOTA. Models were tested under 2 settings—zero-shot prompting and ICL with added tumor-feature descriptions—to assess discriminative performance.

Clinically Relevant Binary Endpoint Conversion

To mirror real-world cystoscopic decision-making when tumor-like lesions are encountered, the 7-class classification task was collapsed into 2 clinically oriented binary endpoints. The first, the biopsy-indication endpoint, represented immediate clinical decision-making: pUC, CIS, non-U Ca, papilloma, and polyps were labeled as "biopsy indicated," whereas cystitis and NOTA were labeled as "biopsy not indicated." The second, the malignancy endpoint, classifies pUC, CIS, and non-U Ca as malignant, and cystitis, polyps, papilloma, and NOTA as nonmalignant. This mapping preserved the full 7-class framework for granular analysis while providing pragmatic outcomes aligned with bedside triage. Notably, papilloma—though histologically benign—was categorized as "biopsy indicated" to reflect the routine need for histologic confirmation.

Prompt Design

The complete and exact prompt designs are detailed in the Multimedia Appendix 1.

Prompt Design With Open-Ended Questions for Image Interpretation

We used a role-based, zero-shot prompt tailored to cystoscopy. The prompt primed domain reasoning ("Suppose you are a urologist") and briefly contextualized the procedure, followed by stepwise instructions to encourage explicit intermediate reasoning. The query comprised five domains: (1) anatomic site (free text), (2) endoscopic findings (free text), (3) presence or absence of a pathological lesion (binary), (4) lesion diagnostic reasoning and justification if present (free text), and (5) final diagnosis (free text).

Prompt Design for Tumor-Like Lesion Classification Task With Multiple-Choice Diagnostic Framework

We compared 2 prompting strategies for cystoscopic diagnosis of tumor-like lesions: zero-shot and ICL. Both adopted a role-based instruction ("Suppose you are a urologist"). The zero-shot prompt presented a single forced-choice 7-class label set. In contrast, the ICL prompt prefixed brief text-based descriptions of the 7 lesion classes before the same multiple-choice query. Models were instructed to provide the best diagnosis from the given options and include a concise rationale grounded in endoscopic morphology.

Outcome Measures and Statistical Analysis

For each image-question-answer instance, the 3 raters' Likert-scale ratings were averaged to obtain a single consensus score. The distribution of these scores across the test set was summarized using the mean and SD to describe the central tendency and variability of model performance. To compare performance among models, pairwise differences in scores were analyzed using paired *t* tests. Results were reported as mean differences with 95% CIs. Given the ordinal nature of Likert-scale data, Wilcoxon signed-rank tests were conducted as a sensitivity analysis. To account for the multiplicity of pairwise comparisons across the top 3 performing models, the Benjamini-Hochberg procedure was applied to control the false discovery rate and mitigate type I error. Consequently, statistical significance for all intermodel comparisons was defined as a false discovery rate–adjusted *P* value (q value)<.05. Subgroup analyses of final diagnosis were conducted according to cystoscopic finding categories and anatomic sites, following the same statistical procedures.

For interpretability, the mean Likert-scale score for each item was further converted into a binary satisfaction outcome: satisfactory if the mean score was >3 and unsatisfactory if ≤3. The satisfaction rate (percentage of satisfactory responses) was reported and used as a binary outcome in subsequent analyses.

The performance metrics for the classification tasks—including binary domains (lesion detection: present vs absent, biopsy indication: yes or no, and malignancy: yes or no) and the 7-class lesion classification—were derived from confusion matrices. Reported metrics included accuracy, sensitivity, specificity, positive predictive value, negative predictive value, Youden J, and the Matthews correlation coefficient (MCC) [23]. Youden J represents the overall diagnostic effectiveness of a test, defined as (sensitivity + specificity – 1), and reflects the balance between true-positive and true-negative rates. The MCC quantifies the overall agreement between predicted and actual classifications by incorporating all 4 components of the confusion matrix (true or false positives and negatives). Metric comparisons were conducted using the chi-square test.

For the 7-class task (n=113), models were instructed to select exactly 1 forced-choice label from the prespecified options. Outputs failing to provide a single permissible choice (eg, refusals such as "I could not answer this question") were coded as invalid. To ensure a consistent head-to-head comparison and minimize selection bias, the primary (strict) analysis used an intent-to-treat approach: invalid outputs were retained in the denominator and treated as incorrect predictions. However, because an invalid output does not necessarily reflect an incorrect diagnosis and may instead represent abstention—potentially safer than guessing in a human-in-the-loop workflow—we conducted a secondary sensitivity analysis, recalculating performance metrics conditional on valid responses only. All statistical analyses were conducted using SAS software (version 9.4; SAS Institute Inc).

Ethical Considerations

The Research Ethics Committee A of National Taiwan University Hospital determined that this study was exempt from

XSL•FO
RenderX

human participant research (NTUH-REC 202507210W). Informed consent was waived because this study involved a secondary analysis of deidentified cystoscopic images with no patient contact or intervention; for publicly available or published images, consent for the original collection followed the source publication, and the exemption permitted secondary analysis without additional consent. All images were deidentified, stored on access-controlled institutional systems, and reported only in aggregate. No participants were recruited, and no compensation was provided; all figures were reviewed to ensure no individual is identifiable.

Results

Distribution of the Whole Test Set and the Tumor-like Lesion Subset

Among 401 cystoscopic images, most originated from the bladder (n=329), followed by the prostate (n=41) and urethra (n=31). Abnormal findings were present in 322 (80.3%) images. The most common categories were tumor or neoplasm (n=126), structural or outlet abnormalities (n=76), inflammatory or reactive changes (n=69), deposits or foreign bodies (n=43), and vascular lesions (n=8); 79 (19.7%) images showed normal anatomy (Table 1). Table 1 provides the detailed distribution of finding subcategories, reflecting diagnostic diversity and difficulty.

XSL•FO
RenderX

Table 1. Detailed distribution of cystoscopic finding subcategories in the whole test dataset (N=401). This table provides a comprehensive breakdown of all observed cystoscopic findings across 3 hierarchical levels (normality, categories, and subcategories) and anatomic sites (bladder, prostate, and urethra). Values are presented as n (intracategory %); percentages represent the proportion of each subcategory within its respective parent category. The inclusion of both benign and malignant findings illustrates the heterogeneity of endoscopic presentations and underscores the diagnostic complexity represented in the dataset.

Anatomic site	Bladder	Prostate	Urethra	Total
Finding normality, categories, and subcategories, n (intracategory %)				
Abnormal	263	35	24	322
Tumor or neoplasm	114 (100)	—	12 (100)	126
Bladder polyp	7 (6.1)	—	—	7
Suspected bladder CIS[a]	17 (14.9)	—	—	17
Suspected nephrogenic adenoma	2 (1.8)	—	—	2
Papilloma	12 (10.5)	—	—	12
Papillary urothelial carcinoma	52 (45.6)	—	—	52
Nonurothelial carcinoma	18 (15.8)	—	—	18
Endometriosis	5 (4.4)	—	—	5
Teratoma	1 (0.9)	—	—	1
Urethral polyp	—	—	4 (33.3)	4
Urethral tumor	—	—	8 (66.7)	8
Inflammation or reaction	67 (100)	1 (100)	1 (100)	69
Bladder amyloidosis	4 (6.0)	—	—	4
Bladder keratinizing	6 (9.0)	—	—	6
Bladder malakoplakia squamous metaplasia	5 (7.5)	—	—	5
Bladder mucosal break	1 (1.5)	—	—	1
Cystitis	26 (38.8)	—	—	26
Hemorrhagic cystitis	5 (7.5)	—	—	5
Suspected IC[b]	10 (14.9)	—	—	10
Suspected radiation cystitis	7 (10.4)	—	—	7
Suspected Schistosomiasis	3 (4.5)	—	—	3
Urethritis	—	1 (100)	1 (100)	2
Deposits or foreign bodies	39 (100)	2 (100)	2 (100)	43
Bladder encrustation	5 (12.8)	—	—	5
Blood clot	12 (30.8)	1 (50)	1 (50)	14
Foreign body	7 (17.9)	—	1 (50)	8
Stone	15 (38.5)	1 (50)	—	16
Structure or outlet	35 (100)	32 (100)	9 (100)	76
Bladder diverticulum	4 (11.4)	—	—	4
Bladder neck contracture	4 (11.4)	—	—	4
Bladder scar	12 (34.3)	—	—	12
Bladder trabeculation	2 (5.7)	—	—	2
Vesicoureteral reflux	4 (11.4)	—	—	4
Ureterocele	6 (17.1)	—	—	6
Suspected fistula	3 (8.6)	—	2 (22.2)	5
Prostate enlargement	—	31 (96.9)	—	31
Prostatic cyst	—	1 (3.1)	—	1

XSL•FO
RenderX

Anatomic site	Bladder	Prostate	Urethra	Total
Urethra stricture	—	—	4 (44.4)	4
Urethral cyst	—	—	1 (11.1)	1
Urethral trauma	—	—	2 (22.2)	2
Vascularity	8 (100)	—	—	8
Bladder hemangioma	1 (12.5)	—	—	1
Bladder telangiectasia	4 (50.0)	—	—	4
Bladder varices	3 (37.5)	—	—	3
Normal	66	6	7	79

[a]CIS: carcinoma in situ.

[b]IC: interstitial cystitis.

The tumor-like lesion subset included 113 visually and pathologically similar images spanning both benign and malignant lesions: cystitis (n=18), polyps (n=7), papilloma (n=12), pUC (n=20), CIS (n=17), non-U Ca (n=17), and NOTA (n=22) (Table 2).

Table 2. Distribution of the tumor-like lesion subset (n=113) used for the 7-class lesion classification task, representing a focused subset of the whole test dataset. The tumor-like lesion subset comprised 18 cystitis (15.9%), 7 polyps (6.2%), 12 papilloma (10.6%), 20 papillary urothelial carcinoma (pUC; 17.7%), 17 carcinoma in situ (CIS; 15%), 17 non-urothelial carcinoma (non-U Ca; 15%), and 22 none of the above (NOTA; 19.5%).

Lesion type	Value, n (%)
Cystitis	18 (15.9)
Polyps	7 (6.2)
Papilloma	12 (10.6)
pUC[a]	20 (17.7)
CIS[b]	17 (15)
Non-U Ca[c]	17 (15)
NOTA[d]	22 (19.5)
Total	113 (100)

[a]pUC: papillary urothelial carcinoma.

[b]CIS: carcinoma in situ.

[c]Non-U Ca: non-urothelial carcinoma.

[d]NOTA: none of the above.

Comparative Mean Scores of LLMs in Image Interpretation

Among the whole test set (n=401), MM-LLMs demonstrated progressively lower performance as task complexity increased (Table 3 and Figure 2). Mean Likert-scale scores declined from anatomic site recognition (≈ 4.1) to findings (≈ 3.4-3.7), lesion reasoning (≈ 2.7-2.9), and final diagnosis (≈ 2.6-2.8). OpenAI-o3, ChatGPT-4o, and Gemini 2.5 Pro achieved comparable accuracy in anatomical localization, while OpenAI-o3 showed the highest overall consistency and clarity in lesion description. Statistically significant differences emerged in the findings, lesion reasoning, and final diagnosis domains, in which OpenAI-o3 outperformed Gemini 2.5 Pro and the medical-specific MedGemma-27B. Notably, MedGemma-27B lagged substantially behind the general-purpose MM-LLMs across all categories, suggesting that its limited training scope constrained both descriptive precision and diagnostic reasoning. These results indicate that reasoning-optimized general-purpose MM-LLMs currently outperform open-source, domain-specific models in free-text cystoscopic interpretation tasks.

XSL•FO
RenderX

Table 3. Performance of 4 multimodal large language models (MM-LLMs) in cystoscopic image interpretation, presented as mean Likert scores and SDs across open-ended questions and final-diagnosis subgroups.

Question and subgroup	Value	OpenAI-o3, mean (SD)	ChatGPT-4o, mean (SD)	Gemini 2.5 Pro, mean (SD)	MedGemma-27B, mean (SD)
Whole test set (n=401)					
Questions					
Q1: anatomic site	401	4.13 (1.24)	4.06 (1.20)	4.10 (1.21)	2.45 (1.37)
Q2: findings	401	3.69 (1.23)	3.54 (1.25)	3.42 (1.30)	1.80 (1.03)
Q4: lesion reasoning	401	2.94 (1.57)	2.89 (1.49)	2.70 (1.48)	1.62 (1.11)
Q5: final diagnosis	401	2.79 (1.59)	2.75 (1.51)	2.61 (1.51)	1.48 (0.93)
Q5. final diagnosis					
Subgrouping by findings					
Tumor or neoplasm	126	3.12 (1.44)	3.32 (1.36)	3.07 (1.51)	2.06 (1.18)
Inflammation or reaction	69	2.34 (1.21)	2.82 (1.30)	2.87 (1.29)	1.19 (0.45)
Deposits or foreign bodies	43	2.77 (1.59)	2.97 (1.52)	2.64 (1.64)	1.07 (0.26)
Structure or outlet	76	1.71 (1.13)	1.45 (0.69)	2.45 (1.67)	1.03 (0.10)
Vascularity	8	2.42 (1.05)	2.75 (0.71)	2.75 (0.87)	1.29 (0.70)
Normal	79	3.79 (1.76)	2.92 (1.81)	1.76 (1.13)	1.49 (1.04)
Subgrouping by anatomic site					
Bladder	329	3.04 (1.56)	3.06 (1.46)	2.64 (1.47)	1.56 (1.00)
Prostate	41	1.45 (1.02)	1.24 (0.62)	2.95 (1.79)	1.09 (0.31)
Urethra	31	2 (1.34)	1.48 (0.91)	1.80 (1.27)	1.12 (0.38)

Figure 2. Comparative mean scores of multimodal large language models (MM-LLMs) for cystoscopic image interpretation across 4 question domains in the whole test set. Asterisks denote statistically significant pairwise differences among the top 3 models. *q<0.05, where q is the false discovery rate (FDR)-adjusted *P* value.

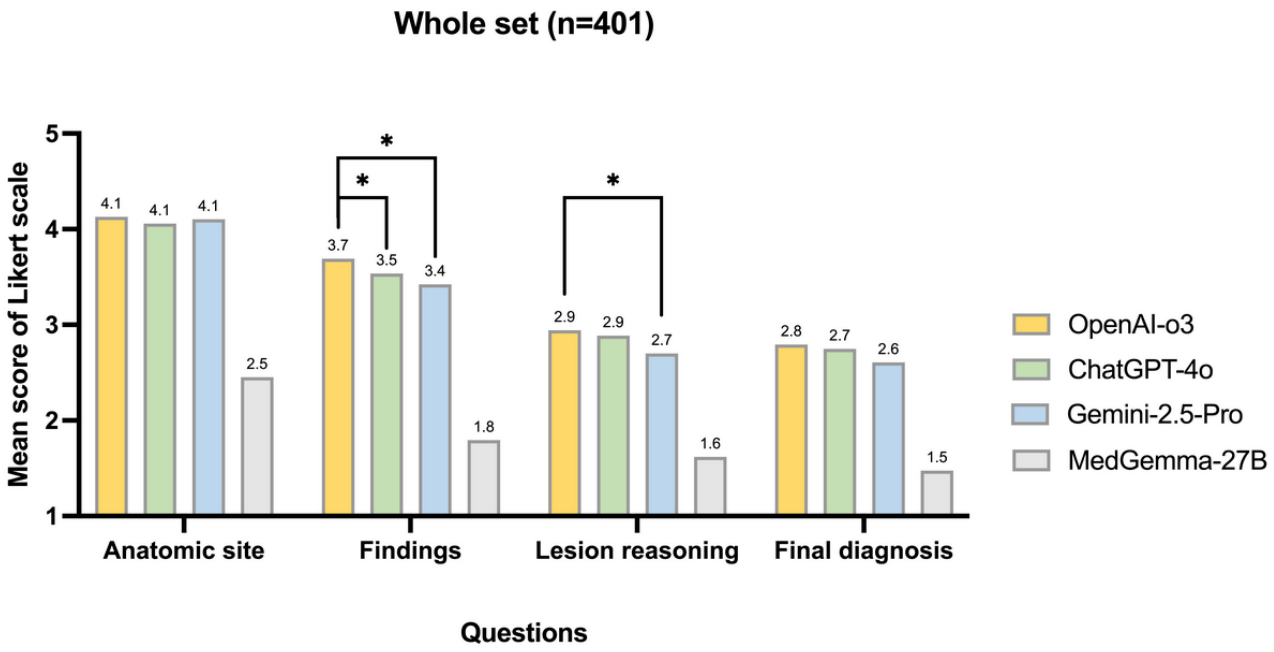

Mean-score pairwise comparisons among models (mean-score deltas) are provided in Table S2 in Multimedia Appendix 1. A 0.2-point difference on the 5-point Likert score corresponds approximately to a 5-point difference on a 100-point scale. The matrix of column-row differences confirmed OpenAI-o3's edge across open-question domains. Versus Gemini 2.5 Pro, OpenAI-o3 scored higher by +0.27 on findings (q=0.002), +0.24 on lesion reasoning (q=0.047), +0.03 on anatomic site, and +0.19 on final diagnosis (not significant). Against ChatGPT-4o, OpenAI-o3 held a small but consistent advantage on findings

XSL•FO
RenderX

(+0.15, q=0.004) with near-parity on anatomic site, lesion reasoning, and final diagnosis (+0.07, +0.06, +0.04; not significant). All general-purpose models substantially outperformed the medical-specific MedGemma-27B; OpenAI-o3's margins were +1.68 (anatomic site), +1.90 (findings), +1.32 (lesion reasoning), and +1.32 (final diagnosis), all q<0.001. Taken together, these deltas indicate that OpenAI-o3 is the most reliable free-text interpreter, with the largest, statistically robust gains in content-heavy domains (Findings → Reasoning → Diagnosis). The significance pattern of the Wilcoxon signed-rank tests was consistent with that of the paired *t* tests.

Intraclass correlation coefficients demonstrated excellent interrater reliability across both model and question domains. Of the 16 intraclass correlation coefficient values, 14 ranged from 0.82 to 0.94, indicating high consistency among raters (Table S3 in Multimedia Appendix 1).

Model Satisfaction Rates

Satisfaction rates for each question across models closely paralleled mean scores of the Likert scale (Figure 3). Overall, satisfaction ranked anatomic site > findings > lesion reasoning ≈ final diagnosis, consistent with mean-score trends. Anatomic site showed uniformly high satisfaction for the top 3 models (339/401, ≈85%), while MedGemma-27B was much lower (184/401, 46%). For findings, OpenAI-o3 (305/401, 76%) and ChatGPT-4o (297/401, 74%) outperformed Gemini 2.5 Pro (277/401, 69%) and MedGemma-27B (92/401, 23%; all q<0.01). In lesion reasoning, OpenAI-o3 (211/401, 53%) and ChatGPT-4o (201/401, 53%) outperformed Gemini 2.5 Pro (184/401, 46%; q=0.003 and q=0.002 vs Gemini 2.5 Pro, respectively), while MedGemma-27B again had the lowest performance (72/401, 18%). For final diagnosis, satisfaction was lowest overall but remained higher for OpenAI-o3 and ChatGPT-4o (192/401, ≈48%) than for Gemini 2.5 Pro (168/401, 42%) or MedGemma-27B (60/401, 15%).

Figure 3. Comparative satisfaction rates (% of cases with mean score >3) of multimodal large language models (MM-LLMs) for cystoscopic image interpretation across 4 question domains. Asterisks denote significance for pairwise comparisons between the top 3 models. *q<0.05, where q is the false discovery rate (FDR)-adjusted *P* value.

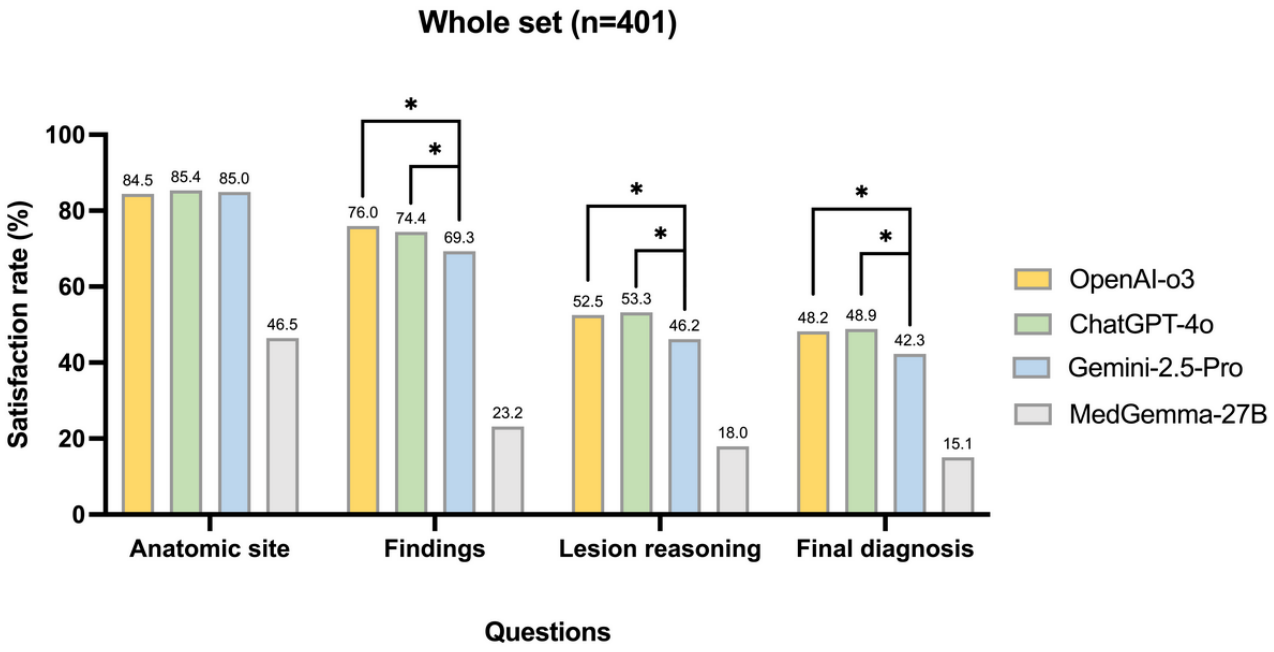

Subgroup Analysis of Final Diagnosis

When mean scores for final diagnosis were stratified by finding category (Figure 4 and Table 3), the largest intermodel differences occurred in the normal and structure or outlet groups. OpenAI-o3 achieved the highest score for normal findings (3.79), significantly outperforming ChatGPT-4o (2.92) and Gemini 2.5 Pro (1.76; q<0.001). Conversely, Gemini 2.5 Pro scored best for structure or outlet findings (2.45), exceeding OpenAI-o3 (1.71) and ChatGPT-4o (1.45; q<0.001). For tumor or neoplasm, ChatGPT-4o slightly surpassed OpenAI-o3 (3.32 vs 3.12; q=0.02), with Gemini 2.5 Pro showing comparable performance (3.07). Inflammation or reaction and vascularity categories both favored ChatGPT-4o and Gemini 2.5 Pro over OpenAI-o3, whereas deposits or foreign bodies showed minimal differences among models.

XSL•FO
RenderX

Figure 4. Subgroup analysis of mean final diagnosis scores across 6 cystoscopic finding categories (tumor or neoplasm, inflammation, deposits, structure, vascularity, and normal) for 4 multimodal large language models (MM-LLMs). Asterisks denote significance for pairwise comparisons between the top 3 models. *q<0.05, where q is the false discovery rate (FDR)-adjusted *P* value.

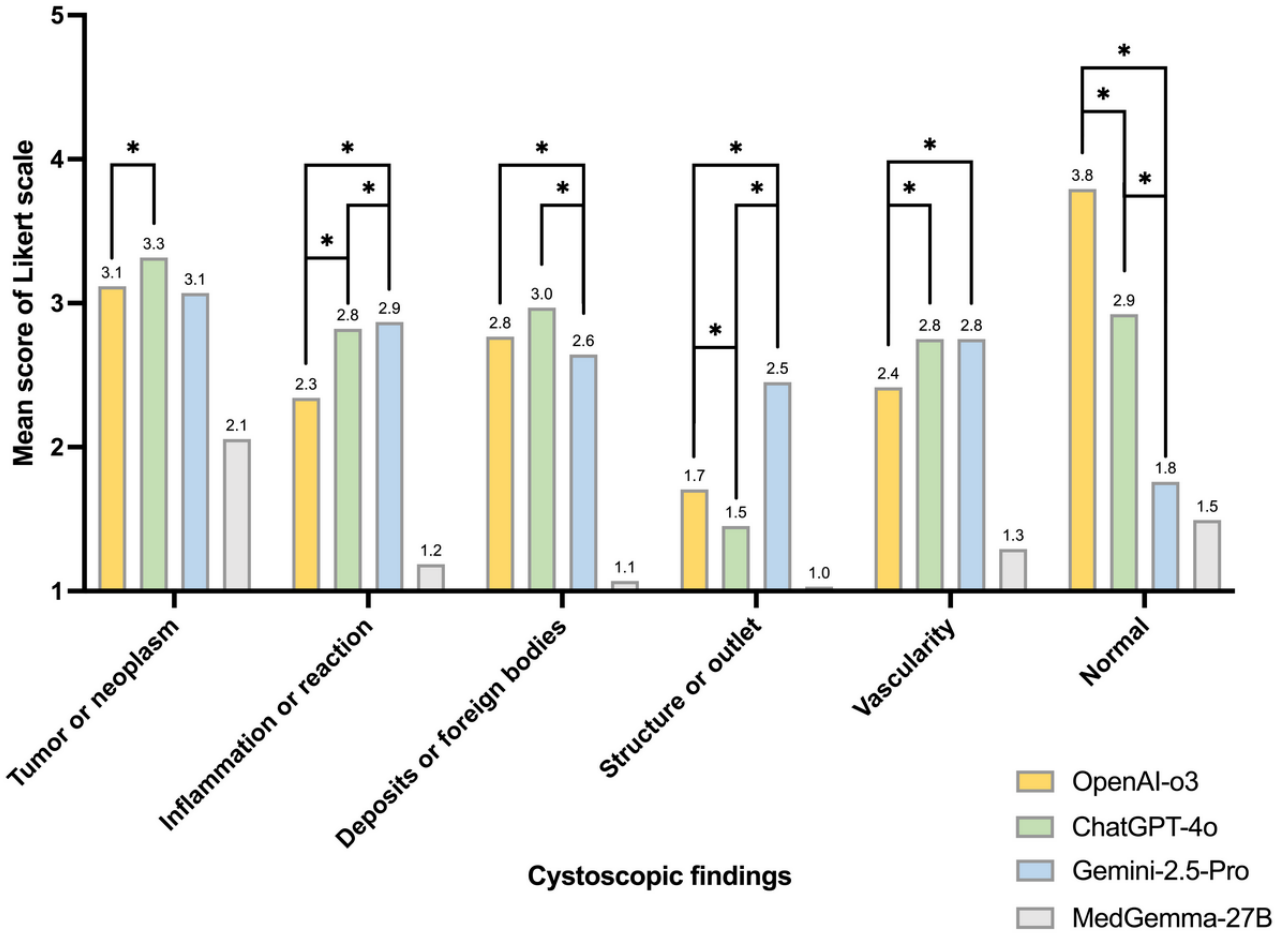

Performance also varied by anatomic site (Figure 5 and Table 3). Gemini 2.5 Pro performed best in the prostate (2.95), significantly exceeding OpenAI-o3 (1.45) and ChatGPT-4o (1.24; q<0.001). In the bladder, OpenAI-o3 (3.04) and ChatGPT-4o (3.06) outperformed Gemini 2.5 Pro (2.64; q<0.001). For the urethra, OpenAI-o3 (2) exceeded ChatGPT-4o (1.48; q=0.02), with Gemini 2.5 Pro intermediate (1.80). These site-specific trends suggest complementary strengths: Gemini 2.5 Pro performs relatively better in structure-dominated prostate views, whereas OpenAI-o3 and ChatGPT-4o perform best in bladder-focused interpretation.

XSL•FO
RenderX

Figure 5. Subgroup analysis of mean final diagnosis scores across 3 anatomic sites (bladder, prostate, and urethra) for 4 multimodal large language models (MM-LLMs). Asterisks denote significance for pairwise comparisons between the top 3 models. *q<0.05, where q is the false discovery rate (FDR)-adjusted *P* value.

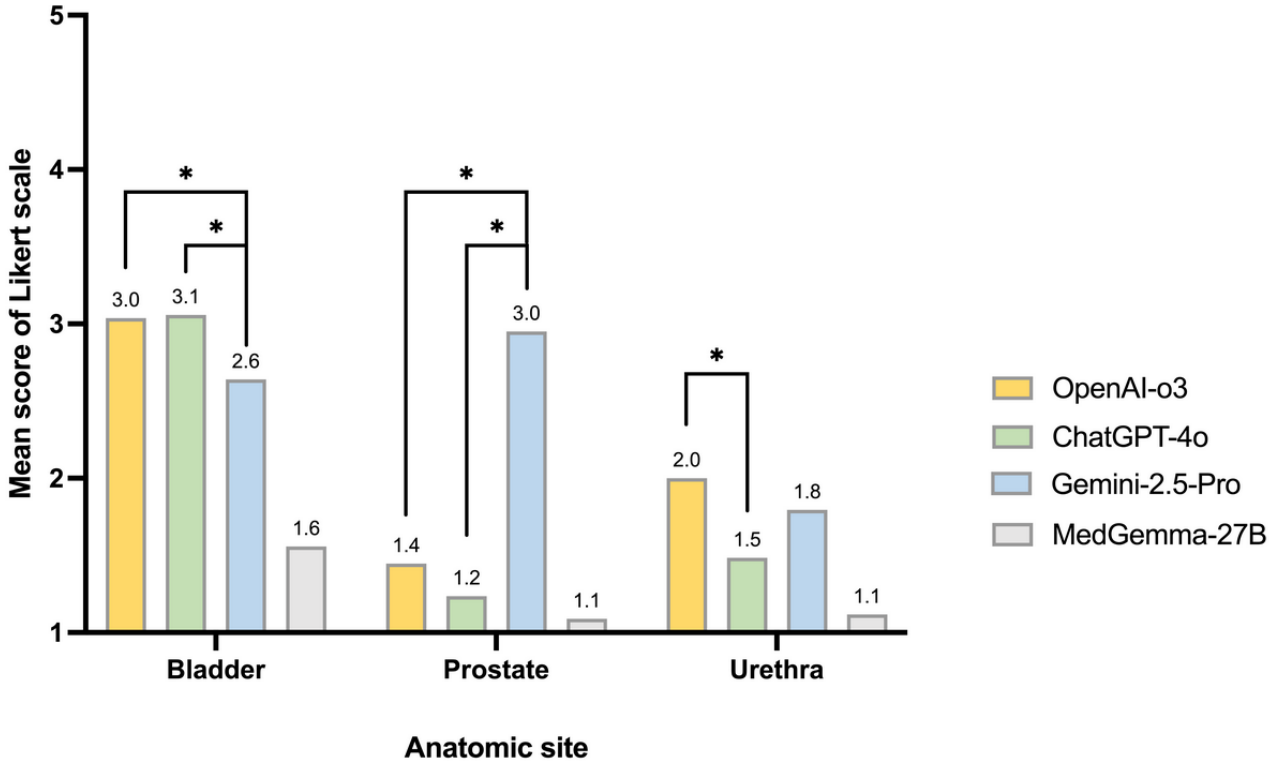

Clinically Relevant Binary Endpoints

For the lesion-detection task in the whole test set, OpenAI-o3 achieved the highest overall performance, with an accuracy of 88.3%, a Youden J of 0.650, and an MCC of 0.635, followed by ChatGPT-4o and Gemini 2.5 Pro, while MedGemma-27B performed the lowest (Table 4). OpenAI-o3 demonstrated the most balanced profile (sensitivity 92% and specificity 73.1%), whereas ChatGPT-4o showed higher sensitivity but lower specificity (94.4% vs 44.2%). Gemini 2.5 Pro exhibited an extreme trade-off—maximal sensitivity (99.7%) but very low specificity (10.3%). MedGemma-27B produced the weakest results overall (accuracy 45.6%, Youden J –0.103, and MCC –0.081).

XSL•FO
RenderX

Table 4. Binary classification performance of 4 multimodal large language models (MM-LLMs) in clinically relevant cystoscopic endpoints (strict analysis).

Task and MM-LLM[a]	VRR[b] (%)	Acc[c] (%)	Sen[d] (%)	Spec[e] (%)	PPV[f] (%)	NPV[g] (%)	Youden J[h]	MCC[i]
Whole test set (n=401)								
Lesion detection (present vs absent)								
OpenAI-o3	100	88.3	92.0	73.1	93.4	68.7	0.650	0.635
ChatGPT-4o	100	84.7	94.4	44.2	87.6	65.4	0.386	0.452
Gemini 2.5 Pro	100	82.3	99.7	10.3	82.1	88.9	0.100	0.266
MedGemma-27B	100	45.6	46.1	43.6	77.2	16.4	-0.103	-0.081
Tumor-like lesion subset (n=113)								
Biopsy indication (yes or no): zero-shot prompting								
OpenAI-o3	99.1	73.5	82.2	57.5	77.9	63.9	0.397	0.407
ChatGPT-4o	92.9	69.0	91.8	27.5	69.8	64.7	0.193	0.258
Gemini 2.5 Pro	100	70.8	86.3	42.5	73.3	63.0	0.288	0.323
MedGemma-27B	100	65.5	97.3	7.5	65.7	60.0	0.048	0.111
Biopsy indication (yes or no): in-context learning								
OpenAI-o3	100	76.1	80.8	67.5	81.9	65.6	0.483	0.481
ChatGPT-4o	98.2	69.0	89.0	32.5	70.7	61.9	0.215	0.265
Gemini 2.5 Pro	100	69.0	75.3	57.5	76.4	56.1	0.328	0.327
MedGemma-27B	100	60.2	93.2	0.0	63.0	0.0	-0.069	-0.159
Presence of malignancy (yes or no): zero-shot prompting								
OpenAI-o3	99.1	62.8	79.6	47.5	58.1	71.8	0.271	0.285
ChatGPT-4o	92.9	55.8	81.5	32.2	52.4	65.5	0.137	0.157
Gemini 2.5 Pro	100	61.1	87.0	37.3	56.0	75.9	0.243	0.278
MedGemma-27B	100	57.5	72.2	44.1	54.2	63.4	0.163	0.169
Presence of malignancy (yes or no): in-context learning								
OpenAI-o3	100	63.7	70.4	57.6	60.3	68.0	0.280	0.282
ChatGPT-4o	98.2	59.3	61.1	57.6	56.9	61.8	0.187	0.187
Gemini 2.5 Pro	100	62.0	66.7	57.6	59.0	65.4	0.243	0.244
MedGemma-27B	100	52.2	87.0	20.3	50.0	63.2	0.074	0.099

[a]MM-LLM: multimodal large language model.

[b]VRR: valid response rate. Valid response rate = (total - invalid) / total. Invalid denotes outputs failing to provide a single permissible choice.

[c]Acc: accuracy.

[d]Sen: sensitivity.

[e]Spe: specificity.

[f]PPV: positive predictive value.

[g]NPV: negative predictive value.

[h]Youden J: Youden J Index.

[i]MCC: Matthews correlation coefficient.

In the tumor-like lesion subset (n=113), OpenAI-o3 again achieved the highest Youden J and MCC performance for both biopsy-indication and malignancy endpoints, followed by Gemini 2.5 Pro and ChatGPT-4o, with MedGemma-27B lowest. For biopsy indication, OpenAI-o3 reached 73.5% accuracy (Youden J=0.397 and MCC=0.407), demonstrating the best specificity-sensitivity balance. ICL modestly improved specificity and accuracy (Table 4). For malignancy detection, OpenAI-o3 similarly performed best (accuracy 62.8%), followed by Gemini 2.5 Pro and ChatGPT-4o, whereas MedGemma-27B underperformed (Tables S4-S5 in Multimedia Appendix 1).

Overall, OpenAI-o3 demonstrated the most balanced diagnostic performance across all 3 binary endpoints, consistently achieving the highest Youden J, largely attributable to its superior specificity (Figure 6). For lesion detection, specificity reached 73.1%, significantly outperforming ChatGPT-4o

XSL•FO
RenderX

(44.2%; q<0.001) and Gemini 2.5 Pro (10.3%; q<0.001; Figure 6A). In biopsy indication, specificity was 57.5%, again higher than ChatGPT-4o (27.5%; q<0.001) and Gemini 2.5 Pro (42.5%; q=0.047; Figure 6B). In malignancy prediction, OpenAI-o3 maintained the highest specificity (47.5%) compared to ChatGPT-4o (32.2%; q=0.042) and Gemini 2.5 Pro (37.3%; q=0.21; Figure 6C). In contrast, MedGemma-27B demonstrated limited generalizability despite its medical-domain optimization.

Figure 6. Radar charts illustrating diagnostic performance across 3 cystoscopic endpoints in classification tasks. Panels represent (A) lesion detection (presence vs absence), (B) biopsy indication (yes vs no), and (C) presence of malignancy (yes vs no). Five key metrics are visualized: accuracy, sensitivity, specificity, Youden J index, and Matthews correlation coefficient (MCC).

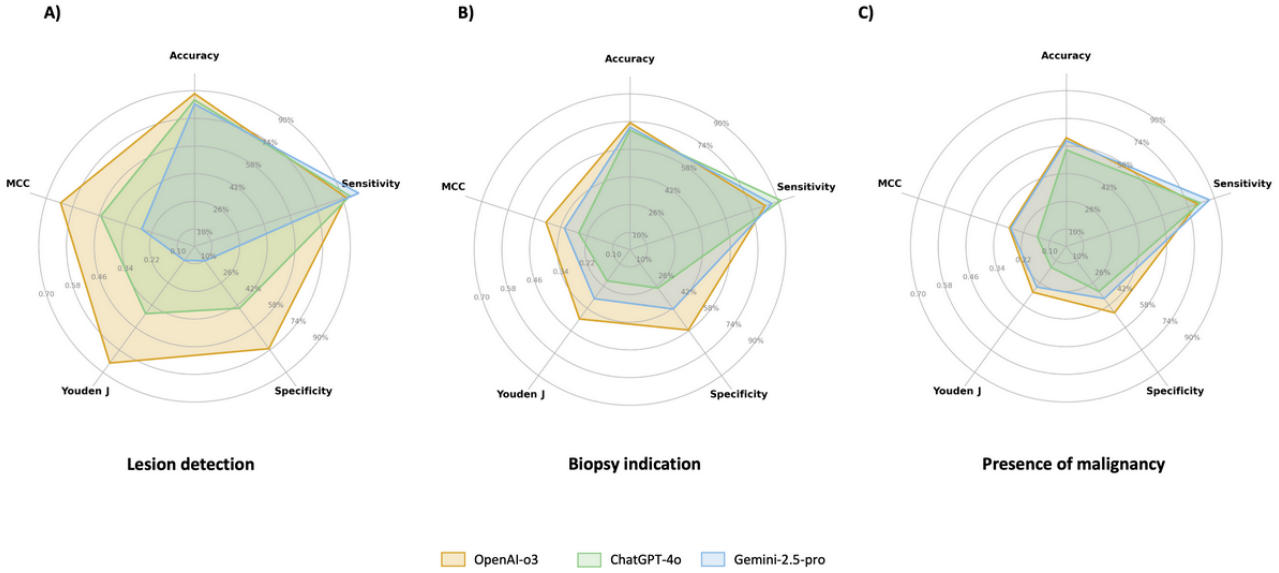

Seven-Class Tumor-Like Lesion Classification

For the 7-class classification (n=113), OpenAI-o3 remained the top-performing model, although overall accuracy was modest, improving only slightly with ICL (microaverage accuracy from 40.7% to 46% and MCC from 0.311 to 0.370) (Table 5). Class-level performance was heterogeneous. In the zero-shot setting, malignant categories (pUC and CIS) achieved relatively balanced sensitivity and specificity, whereas benign lesions (cystitis, polyps, and papilloma) showed high specificity but low sensitivity. Notably, non-U Ca were not recognized. ICL mainly adjusted the sensitivity-specificity balance—enhancing detection of polyps, papilloma, and NOTA while slightly reducing sensitivity for cystitis and CIS. However, non-U Ca remained unrecognized.

XSL•FO
RenderX

Table 5. Confusion matrix outlining the performance of OpenAI-o3 in 7-class tumor-like lesion classification under zero-shot and in-context learning prompting (strict analysis).

	Actual							Total
	Cystitis	Polyps	Papilloma	pUC[a]	CIS[b]	Non-U Ca[c]	NOTA[d]	
Predicted								
Zero-shot prompting								
Cystitis	6	1	1	0	2	1	3	14
Polyps	1	1	0	0	0	1	0	3
Papilloma	0	0	0	0	0	0	0	0
pUC	2	4	11	20	0	14	7	58
CIS	5	0	0	0	9	0	2	16
Non-U Ca	0	0	0	0	0	0	0	0
NOTA	4	1	0	0	5	1	10	21
Invalid[e]	0	0	0	0	1	0	0	1
Total	18	7	12	20	17	17	22	113
ICL[f] prompting								
Cystitis	6	0	1	1	6	0	4	18
Polyps	0	3	0	0	0	3	0	6
Papilloma	0	0	3	0	0	0	0	3
pUC	2	4	8	19	0	12	4	49
CIS	7	0	0	0	7	0	0	14
Non-U Ca	0	0	0	0	0	0	0	0
NOTA	3	0	0	0	4	2	14	23
Total	18	7	12	20	17	17	22	113
Classification metrics								
Zero-shot prompting								
Accuracy	82.3	92.9	89.4	66.4	86.7	85.0	79.6	40.7
Sensitivity	33.3	14.3	0	100	52.9	0	45.5	40.7
Specificity	91.6	98.1	100	59.1	92.7	100	87.9	90.3
PPV[h]	42.9	33.3	0	34.5	56.3	0	47.6	41.1
NPV[i]	87.9	94.5	89.4	100	91.8	85.0	87.0	90.1
Youden J[j]	0.249	0.124	0	0.591	0.456	0	0.334	0.310
MCC[k]	0.277	0.186	0	0.452	0.468	0	0.340	0.311
ICL prompting								
Accuracy	78.8	93.8	92.0	72.6	85.0	85.0	85.0	46.0
Sensitivity	33.3	42.9	25.0	95.0	41.2	0	63.6	46.0
Specificity	87.4	97.2	100	67.7	92.7	100	90.1	91.0
PPV	33.3	50.0	100	38.8	50	0.0	60.9	46.0
NPV	87.4	96.3	91.8	98.4	89.9	85.0	91.1	91.0
Youden J	0.207	0.400	0.250	0.627	0.339	0	0.538	0.370
MCC	0.207	0.430	0.479	0.483	0.368	0	0.529	0.370

[a]pUC: papillary urothelial carcinoma.

[b]CIS: carcinoma in situ.

XSL•FO
RenderX

[c]Non-U Ca: non-urothelial carcinoma.

[d]NOTA: none of the above.

[e]Invalid: model outputs failing to provide a single permissible choice.

[f]ICL: in-context learning.

[g]AVG: microaverage.

[h]PPV: positive predictive value.

[i]NPV: negative predictive value.

[j]Youden J: Youden J Index.

[k]MCC: Matthews correlation coefficient.

The other 3 models performed suboptimally and showed limited responsiveness to ICL (Tables S6-S8 in Multimedia Appendix 1). In the zero-shot setting, microaveraged accuracy ranked as follows: 36.3% (Gemini 2.5 Pro); 31.9% (ChatGPT-4o), 28.3% (MedGemma-27B), with Youden J and MCC both 0.164-0.257. Under ICL, accuracy was similar or slightly lower—34.5% (Gemini 2.5 Pro), 31% (ChatGPT-4o), and 27.4% (MedGemma-27B)—with minimal shifts in Youden J and MCC (0.153-0.236). Class-wise patterns were consistent: malignant categories (pUC, CIS, and non-U Ca) showed the most balanced sensitivity-specificity trade-offs, whereas benign entities (cystitis, polyps, and papilloma) had low sensitivity but high specificity.

Analysis of the NOTA Category

The NOTA category represents a unique "negative exclusion" challenge. Previous research indicates that LLMs often exhibit a bias toward positive selection, struggling to confidently select "None of the Above" even when accurate [24,25]. Our results show that this bias is pervasive, affecting models across different architectures (Tables S6-S8). Despite being a reasoning-optimized model, Gemini 2.5 Pro aligned with the general-purpose ChatGPT-4o and the medical-specific MedGemma-27B in exhibiting a "high specificity, low sensitivity" pattern. Specifically, Gemini 2.5 Pro and ChatGPT-4o achieved high specificity (>97%) but low sensitivity (9.5%-27.3%) across prompting strategies. The open-weight MedGemma-27B exhibited the most severe manifestation of this bias: while its specificity remained high (98.9%), its sensitivity was only 13.6% in the zero-shot setting and collapsed to 0% under ICL prompting. This indicates that for models unable to effectively leverage negative logic, added textual context may inadvertently reinforce positive selection bias.

A distinct divergence was observed between the 2 reasoning-optimized models. In contrast to Gemini 2.5 Pro, OpenAI-o3 demonstrated superior handling of exclusion (Table 5). It achieved a significantly higher baseline sensitivity of 45.5% in the zero-shot setting. Moreover, while ICL yielded negligible or detrimental effects for the other 3 models, OpenAI-o3's sensitivity surged to 63.6% under ICL prompting. This suggests that OpenAI-o3's specific implementation of chain-of-thought reasoning is critical for overcoming the standard positive selection bias, allowing for robust diagnosis through exclusion where other reasoning and general models failed.

Sensitivity Analysis: Conditional on Valid Responses

In the tumor-like lesion classification task, invalid (refusal) outputs were uncommon: valid-response rates were ≈100% for OpenAI-o3, Gemini 2.5 Pro, and MedGemma-27B, whereas ChatGPT-4o had the highest invalid rate (7.1%) in zero-shot prompting. While excluding invalid responses can inflate performance (introducing optimistic bias) relative to the strict analysis, invalid outputs can be interpreted clinically as abstention, which may be safer than guessing in a human-in-the-loop workflow because it prompts clinician confirmation. Accordingly, we report conditional-on-valid performance to better reflect accuracy when the model provides a valid output (Table S9 in Multimedia Appendix 1).

ChatGPT-4o showed the largest strict vs conditional-on-valid differences, consistent with its lower valid-response rate (Table 4 vs Table S9 in Multimedia Appendix 1). Accuracy and Youden J increased from 69% and 0.193 to 74.3% and 0.267 for biopsy indication and from 55.8% and 0.137 to 60% and 0.205 for malignancy. Seven-class changes were small, most notably higher sensitivity for NOTA (from 13.6% to 15%) and cystitis (from 11.1% to 14.3%), with microaverage sensitivity rising from 31.9% to 34.3%.

ICL-focused takeaway across models was that text-only ICL chiefly reweighted sensitivity-specificity rather than boosting overall accuracy; modest gains in benign or NOTA recognition were offset by reduced sensitivity in key malignant classes.

Discussion

Principal Findings

This study is the first to benchmark SOTA MM-LLMs for cystoscopic interpretation under a clinician-defined stress test with rare and diagnostically difficult lesions. The rigorous, blinded design enabled objective assessment of interpretive reasoning and classification. Outputs were also mapped to actionable binary endpoints and used to quantify the incremental effect of text-based ICL over zero-shot prompting, thereby revealing both strengths and current limitations of MM-LLMs in real-world clinical tasks.

Overall, OpenAI-o3 demonstrated superior performance, followed by ChatGPT-4o and Gemini 2.5 Pro, with MedGemma-27B showing the most limited capabilities. The results revealed a progressive decline in model performance as diagnostic complexity increased—from anatomical recognition to higher-order diagnostic synthesis. While models showed meaningful strength in visual recognition and descriptive reporting, performance in challenging differential diagnosis

XSL•FO
RenderX

remained modest, suggesting that current MM-LLMs function best as assistive rather than autonomous diagnostic tools at present.

Image Interpretation and Lesion Classification Tasks

The free-text interpretation task assessed 4 domains of increasing complexity—anatomic site, findings, lesion reasoning, and final diagnosis—simulating real-world diagnostic synthesis that integrates visual recognition with clinical reasoning. Task satisfaction declined with increasing complexity, from anatomic localization (~85%) to definitive diagnosis (~45%). OpenAI-o3 and ChatGPT-4o consistently outperformed Gemini 2.5 Pro and MedGemma-27B, though with distinct profiles: OpenAI-o3 produced concise, accurate descriptions with coherent diagnostic impressions and high specificity for normal anatomy, while ChatGPT-4o showed greater sensitivity for inflammatory and vascular findings. In contrast, Gemini 2.5 Pro often overcalled minor irregularities but performed better on prostate lesions, likely reflecting prostate-predominant pretraining. These discrepancies indicate that MM-LLM behavior depends not only on recognition accuracy but also on underlying reasoning logic, diagnostic thresholds, and domain-specific pretraining.

Regarding clinical decision endpoints, the goal of cystoscopy is to identify abnormal lesions—particularly malignancies—so that biopsies are performed when necessary while avoiding unnecessary procedures that increase cost and risk. Thus, lesion detection, biopsy indication, and malignancy presence were defined as key clinical endpoints. OpenAI-o3 achieved the most balanced performance across sensitivity, specificity, and Youden J, outperforming ChatGPT-4o and Gemini 2.5 Pro, especially in specificity, by accurately distinguishing normal from malignant cases—supporting appropriate biopsy decision-making. These findings highlight the importance of calibrating operating points to clinical priorities and suggest that MM-LLMs, particularly OpenAI-o3, can aid cystoscopic decision-making when optimized for an appropriate specificity-sensitivity balance.

The 7-class lesion classification task evaluated each model's ability to distinguish visually and pathologically similar tumor-like lesions. Models performed best on prevalent malignant lesions (pUC) but struggled with benign mimickers (polyps and papilloma) and rare entities (non-U Ca), often misclassified as pUC—reflecting limited pretraining exposure to uncommon classes. As shown in Table 5, 14 of 17 non-U Ca cases (82.4%) were predicted as pUC, a much more prevalent bladder tumor. One plausible explanation is that LLMs exhibit a tendency to choose majority or high-frequency labels in multiple-choice settings. When pretraining class distributions are imbalanced, the token sequences corresponding to common options (eg, "pUC") can carry higher previous probabilities, biasing the model toward these answers irrespective of correctness. This phenomenon is often described as majority-label bias or common-token bias [26].

On the other hand, the frequent misclassification of papilloma as pUC highlights the inherent challenge of distinguishing these entities based solely on cystoscopic appearance—a difficulty shared by human experts. Rather than indicating model failure, these confusion patterns reflect the substantial macroscopic overlap between papilloma and low-grade pUC. While visual distinction remains experimental, our study addressed this clinical reality by grouping both entities under the "Biopsy Indicated" category in the binary endpoint analysis. In that context, the models successfully flagged these lesions for histologic confirmation, aligning with standard safety protocols despite the specific classification ambiguity. CIS, a flat, high-grade, non-invasive UC subtype, was handled well by OpenAI-o3, achieving strong results (accuracy 86.7, sensitivity 52.9, specificity 92.7, and Youden J 0.46) despite diagnostic difficulty. Overall, OpenAI-o3 showed the most balanced performance, particularly excelling in benign and NOTA classifications, achieving higher specificity than ChatGPT-4o and Gemini 2.5 Pro.

Performance Disparity and the Role of the Open-Weight Baseline

Although MedGemma-27B is a medical-specific model, its performance trailed behind the general-purpose proprietary models (OpenAI-o3, ChatGPT-4o, and Gemini 2.5 Pro). This gap can be attributed to 2 primary factors: domain-specific data misalignment and model scale. First, contrary to the expectation that a medical model should inherently outperform general models, MedGemma's training distribution did not encompass the specific modality of cystoscopy. While its multimodal components (SigLIP encoder) were rigorously pretrained on diverse medical datasets—including chest X-rays, dermatology images, ophthalmology images, and histopathology slides—endoscopic imagery was notably absent from its pretraining corpus. Consequently, the model faced a "zero-shot" challenge in a domain it had not explicitly learned, whereas the massive general-purpose models likely benefited from broader exposure to endoscopic images present in their web-scale training data.

Second, as a local and open-weight baseline, MedGemma-27B (≈27B parameters) operates with significantly constrained capacity compared to the proprietary SOTA architectures. It lacks the extensive parameter count, training budget, and chain-of-thought optimization that allow models such as OpenAI-o3 to generalize across unseen tasks. Therefore, our intent is not to claim parity with these massive systems, but to establish a transparent performance floor for open-weight deployment. Despite the lower accuracy in this zero-shot setting, MedGemma-27B remains a critical benchmark for institutions requiring on-premise, privacy-preserving solutions. Its performance represents the current starting point for future adaptation, such as LoRA fine-tuning or retrieval-augmented prompting, rather than a direct competitor in raw reasoning capability.

Comparing With Previous Studies

Guo et al. [17] reported comparable findings when evaluating ChatGPT-4V (OpenAI) and Claude-3.5 (Anthropic) on 603 cystoscopic images, achieving accuracies of 82.8% and 79.8% but with marked variability across conditions. Both models performed well for cystitis and bladder tumors but poorly for BPH and normal structures, indicating that general-purpose LLMs detect major lesions with high sensitivity but struggle

XSL•FO
RenderX

with subtle findings. Similar variability has been observed in gastrointestinal endoscopy, where ChatGPT-4V showed mixed accuracy across lesion types and underperformed relative to tuned CNN models [27]. Recent work suggests that general multimodal models such as Gemini 2.5 Pro may even surpass specialized AI in certain "edge cases" [28]. Unlike task-specific systems, these models can simultaneously classify images and generate descriptive reasoning and management suggestions, offering value for clinical interpretation and education.

Enhancing MM-LLM Performance in Medical-Specific Domain

Several in-domain strategies can improve general-purpose MM-LLMs without training from scratch, including ICL, contrastive pretraining, and retrieval augmentation.

In our study, text-only ICL with brief cystoscopic tumor descriptions had minimal impact, except for OpenAI-o3. Specificity rose slightly, and sensitivity improved for rare benign lesions, but this was offset by reduced sensitivity for malignant classes (pUC and CIS)—a trade-off in which stricter thresholds reduce false positives and unnecessary biopsies, but risk missed cancers. These findings indicate that text-only ICL confers minimal benefit for image-dominant tasks. Notably, OpenAI-o3 showed a modest gain in micro-average accuracy (0.41-0.46), driven mainly by NOTA, likely reflecting its reasoning-oriented architecture and more flexible use of ICL as contextual support.

A likely reason for the limited effect is insufficient visual grounding: text-only cues do not anchor the model's attention to class-defining morphology. In fine-grained visual discrimination (eg, cystoscopic lesion typing), semantic hints (eg, "papillary fronds" and "flat erythematous base") may not map reliably to visual features unless those associations were learned during pretraining; without image exemplars, the model's visual reasoning remains underconstrained.

Beyond text-only prompts, few-shot image-text exemplars (multimodal ICL) can strengthen grounding and improve accuracy. Presenting paired lesion images with diagnoses exposes prototypical visual features and tightens the link between morphology and class semantics. Across histopathology imaging, image-text ICL has enabled ChatGPT-4V to approach or surpass task-specific classifiers with only 5-10 examples per class, markedly narrowing the gap between zero-shot and fully supervised models [29].

In parallel, contrastive-learned encoders (eg, MedCLIP) [30] and multimodal retrieval augmentation [31] can enrich representations and factual grounding, mitigating data scarcity and hallucination. In summary, improving MM-LLM image classification in medical domains is multifaceted, and combining hybrid ICL (image + text), contrastive pretraining, and retrieval augmentation offers a practical path to greater accuracy, robustness, and interpretability in cystoscopic diagnosis.

Clinical Implications

MM-LLMs offer greater flexibility than task-specific endoscopy AI by combining visual recognition with contextual reasoning and narrative explanation. They can interpret morphology-diverse findings and integrate relevant clinical text, supporting a more context-aware understanding. Our results suggest potential applications in education and workflow support, including serving as virtual tutors for trainees and automating report generation to reduce workload and standardize documentation. However, their moderate diagnostic accuracy, particularly for rare or subtle lesions, limits their current use as autonomous diagnostic tools. Future efforts should focus on vision-conditioned ICL, multimodal retrieval-augmented training, and video-based modeling to enhance interpretive stability and diagnostic confidence [30,31]. Integration with patient-level data, cystoscopy-specific benchmark datasets, and human-in-the-loop oversight will be critical to ensure clinical safety and responsible implementation.

Limitations

This study has several methodological strengths, including an unbiased and rigorous evaluation framework. The dual-task design enabled simultaneous assessment of reasoning transparency, adaptability, and accuracy—helping distinguish superficial pattern recognition from genuine clinical understanding. However, several limitations should be acknowledged. First, our ICL implementation was text-based only: models received brief written descriptions of tumor features without any paired visual exemplars. As a result, this study evaluated "text-based ICL" rather than full multimodal ICL, and the absence of a few-shot image or image-text examples likely constrained the models' multimodal capabilities. The modest gains observed with ICL in our experiments may therefore underestimate the potential benefit of visual or hybrid (image + text) ICL. Future work should directly compare text-based, visual, and hybrid ICL strategies and explore complementary approaches such as contrastive-learned encoders and multimodal retrieval augmentation for cystoscopic diagnosis. Second, because the raw test images were drawn from heterogeneous sources, residual confounding from source-related differences and image-quality artifacts cannot be fully excluded. Although we applied a strict 3-layer quality control pipeline—image exclusion criteria, standardized preprocessing, and human verification of diagnostic utility—to mitigate these effects, some bias related to image source heterogeneity may remain. In addition, our evaluation relied on static images; incorporating temporal cues from cystoscopy videos may improve recognition of subtle or evolving lesions and reduce misclassification. Third, our dataset has an enriched abnormality prevalence (80.3%), substantially higher than that of typical clinical populations (20%-30%). Consequently, the reported positive predictive value and negative predictive value are inflated and not generalizable; they should be interpreted as dataset-specific rather than as estimates for real-world screening or hematuria-clinic settings.

Conclusions

Using a clinically challenging, stress-test image set and a rigorous blinded evaluation framework, this study comprehensively assessed MM-LLMs for cystoscopic interpretation and lesion classification. Among the evaluated models, OpenAI-o3 demonstrated the most balanced and clinically coherent performance, followed by ChatGPT-4o and Gemini 2.5 Pro. These findings highlight the meaningful

XSL•FO
RenderX

assistive potential of MM-LLMs in generating interpretable free-text rationales, supporting biopsy triage, and facilitating training. However, their performance in truly difficult differential diagnoses remains modest and requires further optimization before safe clinical integration.

Acknowledgments

We would like to extend our heartfelt gratitude to our colleagues for their invaluable contributions to this work. ChatGPT (OpenAI) was used only for language editing and grammar correction; it was not used for content generation or any data-related tasks.

Data Availability

The datasets generated and/or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

YCS and CYW contributed equally as first authors. SWH and CYT contributed equally as corresponding authors. Conceptualization was performed by CYT and SWH. Data curation was carried out by YCS and CYW. Formal analysis was conducted by CYW, YCS, SWH, and CYT. Visualization was undertaken by CYT and CYW. The original draft was written by CYW and YCS, and SWH and CYT completed manuscript review and editing. Supervision was provided by CYT and SWH.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional methods, tables, and figures.
[PDF File (Adobe PDF File), 809 KB - jmir_v28i1e87193_app1.pdf]

References

1. David SA, Patil D, Alemozaffar M, Issa MM, Master VA, Filson CP. Urologist use of cystoscopy for patients presenting with hematuria in the United States. Urology 2017;100:20-26. [doi: 10.1016/j.urology.2016.09.018] [Medline: 27645524]
2. Guldhammer CS, Vásquez JL, Kristensen VM, Norus T, Nadler N, Jensen JB, et al. Cystoscopy accuracy in detecting bladder tumors: a prospective video-confirmed study. Cancers (Basel) 2023;16(1):160 [FREE Full text] [doi: 10.3390/cancers16010160] [Medline: 38201586]
3. GLOBOCAN 2022: bladder cancer 9th most common worldwide. WBCP Coalition. URL: https://tinyurl.com/y2p57rbx [accessed 2025-11-02]
4. Devlies W, de Jong JJ, Hofmann F, Bruins HM, Zuiverloon TC, Smith EJ, et al. The diagnostic accuracy of cystoscopy for detecting bladder cancer in adults presenting with haematuria: a systematic review from the European Association of Urology Guidelines Office. Eur Urol Focus 2024;10(1):115-122 [FREE Full text] [doi: 10.1016/j.euf.2023.08.002] [Medline: 37633791]
5. Ye Z, Li Y, Sun Y, He C, He G, Ji Z. Leveraging deep learning in real-time intelligent bladder tumor detection during cystoscopy: a diagnostic study. Ann Surg Oncol 2025;32(5):3220-3226. [doi: 10.1245/s10434-025-17015-3] [Medline: 40050483]
6. Guo Y, Li C, Zhang S, Zhu G, Sun L, Jin T, et al. U-net-based assistive identification of bladder cancer: a promising approach for improved diagnosis. Urol Int 2024;108(2):100-107. [doi: 10.1159/000535652] [Medline: 38081150]
7. Hwang WK, Jo SB, Han DE, Ahn ST, Oh MM, Park HS, et al. Artificial intelligence-based classification and segmentation of bladder cancer in cystoscope images. Cancers (Basel) 2024;17(1):57 [FREE Full text] [doi: 10.3390/cancers17010057] [Medline: 39796686]
8. Jia X, Shkolyar E, Laurie MA, Eminaga O, Liao JC, Xing L. Tumor detection under cystoscopy with transformer-augmented deep learning algorithm. Phys Med Biol 2023;68(16). [doi: 10.1088/1361-6560/ace499] [Medline: 37548023]
9. Huang HH, Cheng P, Tsai CY. Exploring artificial intelligence in functional urology: a comprehensive review. Urol Sci 2025;36:2-10. [doi: 10.1097/us9.0000000000000057]
10. Bretthauer M, Ahmed J, Antonelli G. Use of computer assisted detection (CADe) colonoscopy in colorectal cancer screening and surveillance: European Society of Gastrointestinal Endoscopy (ESGE) position statement. Endoscopy 2025;57:667-673. [doi: 10.1055/s-0045-1805161]
11. Carlini L, Massimi D, Mori Y, Antonelli G, Rizkala T, Spadaccini M, et al. Large language models for detecting colorectal polyps in endoscopic images. Gut 2025 May 24. [doi: 10.1136/gutjnl-2025-335091] [Medline: 40360230]
12. Repici A, Badalamenti M, Maselli R, Correale L, Radaelli F, Rondonotti E, et al. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology 2020;159(2):512-520.e7. [doi: 10.1053/j.gastro.2020.04.062] [Medline: 32371116]

XSL•FO
RenderX

13. Shaukat A, Lichtenstein DR, Somers SC, Chung DC, Perdue DG, Gopal M, SKOUT™ Registration Study Team. Computer-aided detection improves adenomas per colonoscopy for screening and surveillance colonoscopy: a randomized trial. Gastroenterology 2022;163(3):732-741 [FREE Full text] [doi: 10.1053/j.gastro.2022.05.028] [Medline: 35643173]
14. Wang P, Liu X, Berzin TM, Glissen Brown JR, Liu P, Zhou C, et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. Lancet Gastroenterol Hepatol 2020;5(4):343-351. [doi: 10.1016/S2468-1253(19)30411-X] [Medline: 31981517]
15. AlSaad R, Abd-Alrazaq A, Boughorbel S, Ahmed A, Renault M, Damseh R, et al. Multimodal large language models in health care: applications, challenges, and future outlook. J Med Internet Res 2024;26:e59505 [FREE Full text] [doi: 10.2196/59505] [Medline: 39321458]
16. Nam Y, Kim DY, Kyung S, Seo J, Song JM, Kwon J, et al. Multimodal large language models in medical imaging: current state and future directions. Korean J Radiol 2025;26(10):900-923 [FREE Full text] [doi: 10.3348/kjr.2025.0599] [Medline: 41015856]
17. Guo L, Zuo Y, Yisha Z, Liu J, Gu A, Yushan R, et al. Diagnostic performance of advanced large language models in cystoscopy: evidence from a retrospective study and clinical cases. BMC Urol 2025;25(1):64 [FREE Full text] [doi: 10.1186/s12894-025-01740-8] [Medline: 40158093]
18. Ferber D, Wölflein G, Wiest IC, Ligero M, Sainath S, Ghaffari Laleh N, et al. In-context learning enables multimodal large language models to classify cancer pathology images. Nat Commun 2024;15(1):10104 [FREE Full text] [doi: 10.1038/s41467-024-51465-9] [Medline: 39572531]
19. Tenny BC, O'Neill M. Diagnostic Cystoscopy. Cham: Springer; 2022.
20. o3. OpenAI Platform. 2025. URL: https://platform.openai.com/docs/models/o3 [accessed 2025-06-15]
21. Gemini 2.5 Pro. Google Cloud. URL: https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro?utm_source=chatgpt.com&hl=zh-tw [accessed 2025-06-15]
22. MedGemma 1.5 model card. Health AI Developer Foundations. 2025. URL: https://developers.google.com/health-ai-developer-foundations/medgemma/model-card [accessed 2025-07-20]
23. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 2020;21(1):6 [FREE Full text] [doi: 10.1186/s12864-019-6413-7] [Medline: 31898477]
24. Tam ZR, Wu CK, Lin CY, Chen YN. None of the above, less of the right: parallel patterns between humans and llms on multi-choice questions answering. arXiv Preprint posted online on Mar 3, 2025 [FREE Full text]
25. Ren J, Zhao Y, Vu T, Liu PJ, Lakshminarayanan B. Self-evaluation improves selective generation in large language models. arXiv Preprint posted online on Dec 14, 2023 [FREE Full text]
26. Zhao Z, Wallace E, Feng S, Klein D, Singh S. Calibrate before use: improving few-shot performance of language models. arXiv Preprint posted online on Feb 19, 2021. [doi: 10.48550/arXiv.2102.09690]
27. Khalafi MA, Safavi-Naini SAA, Salehi A, Naderi N, Alijanzadeh D, Moghadam PK, et al. Vision language models versus machine learning models performance on polyp detection and classification in colonoscopy images. Sci Rep 2025;15(1):45484 [FREE Full text] [doi: 10.1038/s41598-025-29566-2] [Medline: 41309981]
28. Zhang Y, Chen Q, Zhou T. Can general-purpose omnimodels compete with specialists? A case study in medical image segmentation. arXiv Preprint posted online on Aug 31, 2025 [FREE Full text]
29. Ono D, Dickson DW, Koga S. Evaluating the efficacy of few-shot learning for GPT-4Vision in neurodegenerative disease histopathology: a comparative analysis with convolutional neural network model. Research Square Preprint posted online on May 28, 2024. [doi: 10.21203/rs.3.rs-4462333/v1]
30. Wang Z, Wu Z, Agarwal D, Sun J. MedCLIP: contrastive learning from unpaired medical images and text. 2022 Presented at: Proc Conf Empir Methods Nat Lang Process. 2022 Dec; November 4-9, 2025; Suzhou, China. [doi: 10.18653/v1/2022.emnlp-main.256]
31. Xia P, Zhu K, Li H, Wang T, Shi S, Wang S, et al. Mmed-rag: versatile multimodal rag system for medical vision language models. arXiv Preprint posted online on Oct 16, 2024 [FREE Full text]

Abbreviations

AI: artificial intelligence
CIS: carcinoma in situ
ICL: in-context learning
MCC: Matthews correlation coefficient
MM-LLM: multimodal large language model
Non-U Ca: non-urothelial carcinoma
NOTA: none of the above
pUC: papillary urothelial carcinoma
SOTA: state-of-the-art
Youden J: Youden J index

XSL•FO
RenderX

Edited by A Coristine; submitted 05.Nov.2025; peer-reviewed by HC Yang, CY Chang; comments to author 04.Dec.2025; revised version received 07.Jan.2026; accepted 07.Jan.2026; published 28.Jan.2026.

Please cite as:
Shih YC, Wu CY, Huang SW, Tsai CY
Multimodal Large Language Models for Cystoscopic Image Interpretation and Bladder Lesion Classification: Comparative Study
J Med Internet Res 2026;28:e87193
URL: https://www.jmir.org/2026/1/e87193
doi:10.2196/87193
PMID:

©Yung-Chi Shih, Cheng-Yang Wu, Shi-Wei Huang, Chung-You Tsai. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluation of Large Language Models for Structured Data Extraction From Interstitial Lung Disease Clinical Notes: Comparative Study

Stephanie Ji Chen[1], MD; Manoj Venkat Maddali[1,2,3], MD, MS; Curtis Langlotz[4], MD, PhD; Christian Bluethgen[4], MD, MSc; Jonathan Chen[2], MD, PhD; Rishi Raj[1], MD

[1]Division of Pulmonary, Allergy, and Critical Care Medicine, Stanford Medicine, Stanford, CA, United States

[2]Department of Biomedical Data Science, Stanford Medicine, Stanford, CA, United States

[3]Division of Pulmonary and Critical Care Medicine, Piedmont HealthCare, Atlanta, GA, United States

[4]Department of Radiology and Center for Artificial Intelligence in Medicine and Imaging, Stanford Medicine, Stanford, CA, United States

Corresponding Author:
Stephanie Ji Chen, MD
Division of Pulmonary, Allergy, and Critical Care Medicine
Stanford Medicine
300 Pasteur Dr.
Stanford, CA
United States
Phone: 1 2489713831
Email: sjichen@stanford.edu

Abstract

Background: Most clinically relevant data are in unstructured clinical notes, which are verbose and imprecise, making structured data extraction a costly bottleneck for screening patients for studies or maintaining health care registries. This challenge is particularly pronounced in interstitial lung disease (ILD) and requires significant human effort to interpret notes and determine classification to create an ILD registry. Large language models (LLMs) have the potential to significantly reduce this cost and effort.

Objective: We aim to compare the performance of various LLMs for structured data extraction from unstructured ILD clinic notes. Our primary aim was to evaluate LLM extraction of binary structured data (yes/no answers) from clinical notes regarding key ILD clinical questions. A secondary analysis evaluated select LLMs for the extraction of multiclass data to determine ILD classification.

Methods: We used 12 different LLMs to extract binary answers to 10 ILD clinical questions from the most recent clinic notes of 100 ILD clinic patients. We additionally used 2 LLMs (gpt-oss-20b and gpt-oss-120b) to extract multiclass data regarding ILD classification. Prompts were created with the assistance of ChatGPT (OpenAI) and refined with an iterative approach by testing on a prompt engineering cohort of 10 ILD clinic patient notes. Ground truth was established by consensus among 3 ILD physicians. LLM performance was evaluated using accuracy, precision, recall, and F_1-scores.

Results: LLMs processed each interface call of a clinical note-prompt combination in 1-2 seconds, with estimated costs ranging from less than US $0.001 to US $0.11 (or approximately US $0.05 to US $10.50 per clinical note accounting for 10 runs and 10 binary prompts) depending on the model. Out of the 12 LLMs assessed, 7 models (Claude 3.5 Sonnet [Anthropic], GPT-4o, gpt-oss-20b, gpt-oss-120b, o1, o1-mini, and o3-mini [OpenAI]) performed at human-level accuracy, similar to that of the 3 ILD clinicians (96.2%). A total of 5 LLMs performed significantly worse than humans (Holm-adjusted $P \leq .003$ for all). gpt-oss-120b, o1, and o3-mini models achieved the highest F_1-scores of all the evaluated LLMs. There was no significant difference in model accuracy among the top tier models (Claude 3.5 Sonnet, gpt-oss-20b, gpt-oss-120b, o1, o1-mini, and o3-mini), though GPT-4o achieved significantly lower accuracy than o1 (Bonferroni-adjusted $P=.04$). Multiclass data extraction using gpt-oss-120b and gpt-oss-20b demonstrated lower accuracy when compared to its corresponding binary data extraction (91.1% and 88.0%, respectively). There was no significant difference in accuracy between gpt-oss-120b and gpt-oss-20b for multiclass extraction.

XSL•FO
RenderX

Conclusions: Multiple LLMs consistently achieved human-level accuracy in extracting structured binary data from ILD clinical notes, while being orders of magnitude faster and cheaper. Multiclass data extraction was possible but associated with a lower accuracy. LLMs are promising tools that can be used for clinical data extraction to improve clinical research efficiency.

(J Med Internet Res 2026;28:e90547) doi:10.2196/90547

KEYWORDS

electronic health records; information extraction; interstitial lung diseases; large language models; natural language processing

Introduction

Most clinically relevant data are contained in unstructured text such as clinical notes [1]. Consequently, clinical research often entails significant human effort and cost to review and extract binary data from unstructured text [2,3]. This challenge is particularly pronounced in interstitial lung disease (ILD), where clinical documentation is prone to verbosity and imprecision. Reliable and faithful extraction of structured data by large language models (LLMs) can significantly reduce the cost and effort currently required to screen patients for studies or create and maintain ILD registries or databases.

LLMs are machine learning models designed to predict the next word or token in a sequence based on their training data [4]. Introduction of the transformer architecture in 2017 enabled LLMs to analyze whole sentences or paragraphs simultaneously for relationships, rather than process a single word at a time. This significantly increased the size of LLMs and the data they could be trained on [5]. The extensive data and parameters leveraged by the transformer architecture have enabled LLMs to create rich, multilayered mathematical representations of language, which contribute to the LLM's appearance of intelligence [6]. When given a prompt, these transformer models generate text by predicting the most probable next word based on training data. This process continues iteratively until coherent sentences and paragraphs are produced in response to the prompt.

The most common way to access available LLMs is through a web or app-based interface. The user submits a prompt or query through the LLM web page or application on a device and receives a response, which is displayed in the browser or application. This approach, though convenient, is limited by its inability to adjust model parameters or process more than one document at a time. An alternative for interacting with LLMs involves using an application programming interface (API). This approach involves writing and executing code on a local machine that takes user inputs such as prompts, records, and desired model parameters, and sends them to the LLM via the API for processing. The LLM then generates a response, which is returned to the user and can be further processed to the desired format (Figure 1). By accessing LLMs via an API, the user can process large sets of queries and data sequentially without requiring individual input as in the web-based approach. An additional advantage of this method is the potential to modify various model parameters that are otherwise hidden or fixed in a web or app-based interface, such as the temperature, which determines the stochasticity ("deterministic" vs "creative") of the model responses [7].

XSL•FO
RenderX

Figure 1. Schematic representation of accessing large language models (LLMs). (A) The user programs the application programming interface (API), which can encode multiple records/data, prompts/queries, and model parameters. This is then sent to the LLM either via remote server or locally on a graphics processing unit (GPU) machine, which generates responses that are processed by the API into a unified response and sent back to the user. (B) The user provides a single query to a web browser or application, which accesses the LLM via a remote server. The LLM then generates a single response and sends it back to the user.

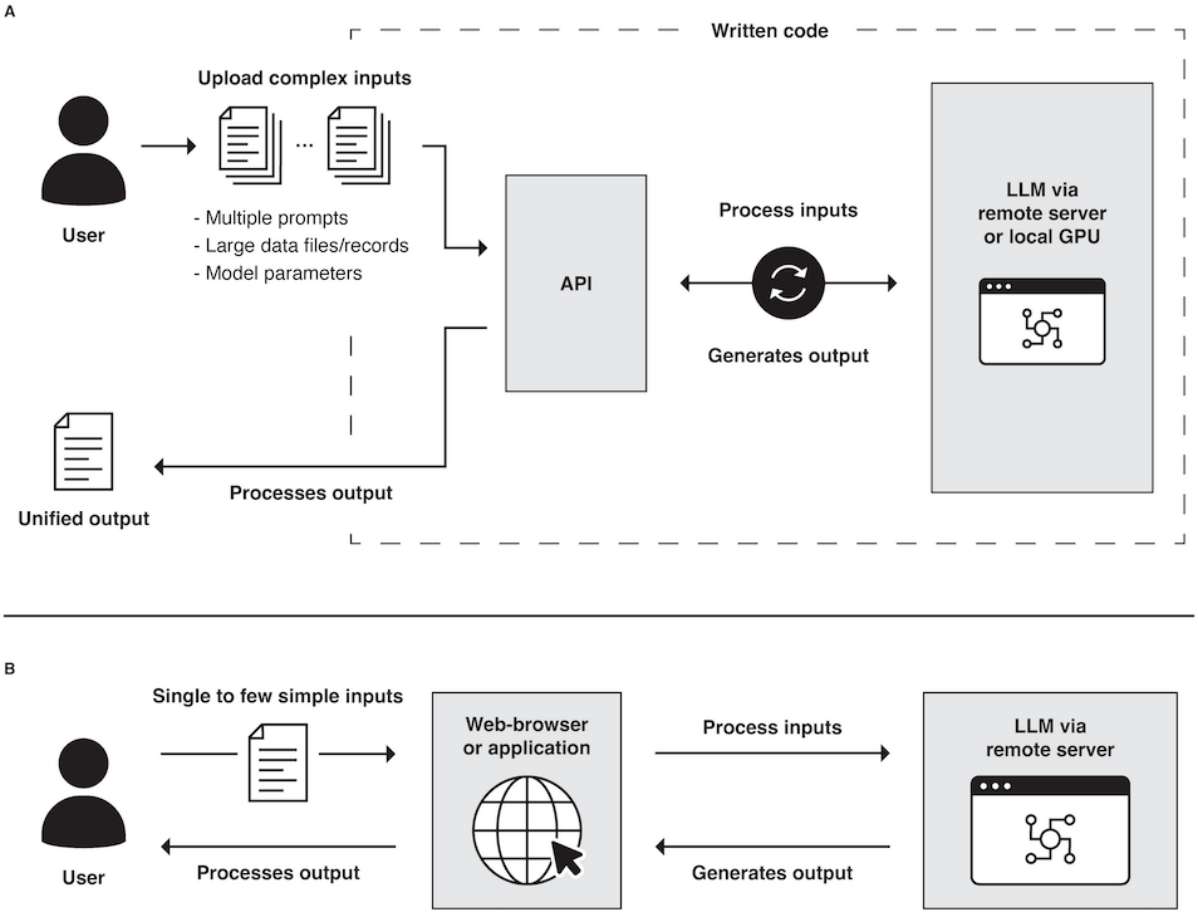

To use LLMs, well-crafted prompts are essential. Prompt engineering is the process of crafting effective and precise instructions or questions as inputs for LLMs to produce the desired output or answer. For example, a simple prompt such as “Does the patient have ILD?” is likely to generate nonspecific answers due to a lack of context and specificity. In such cases, providing additional context and clarification in the prompts is essential to achieve accurate responses.

This study was designed as a proof of concept to assess the performance characteristics of current LLMs in extracting structured binary data (ie, yes or no answers) from clinical notes for patients with ILD regarding key ILD clinical questions. The focus of our study was not to address nuances in ILD classification or to develop all the necessary prompts for a registry, but rather to determine whether LLMs can reliably and faithfully extract binary data from verbose and ambiguous documentation when given a specific set of instructions. Additionally, we also performed a secondary exploratory analysis to assess the feasibility of multiclass extraction using 2 open-weight models of differing sizes. Some of the results of these studies have been previously reported in the form of an abstract [8].

Methods

Cohort

We identified patients who were initially seen in the Stanford Interstitial Lung Disease Clinic between 2018 and 2022 and had at least 3 ILD clinic visits; this was to ensure adequate follow-up visits to reach a diagnosis. From this cohort, 10 patients and 100 patients were randomly selected for the prompt engineering cohort and LLM evaluation validation cohort, respectively (Figure 2). These cohorts were kept separate. The most recent ILD clinic note available for each patient was used for this study.

XSL•FO
RenderX

Figure 2. Flowchart of prompt engineering and large language model evaluation cohorts. ILD: interstitial lung disease.

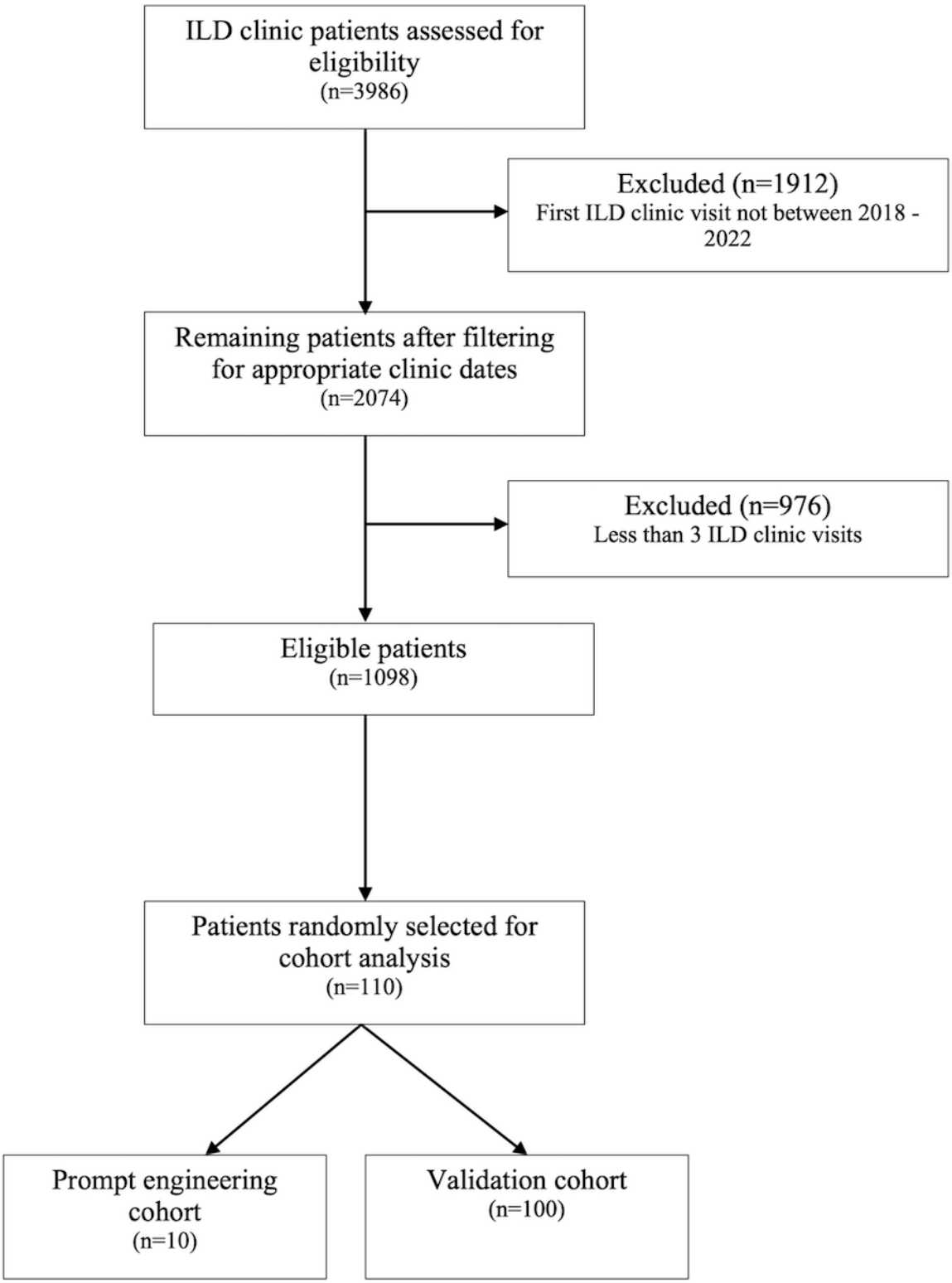

Ethical Considerations

This study was approved by the Institutional Review Board of Stanford University (IRB-83821). A waiver of informed consent was obtained, given that this retrospective chart review study involved no more than minimal risk and does not adversely affect the rights and welfare of the participants. All data used in the study were stored securely by investigators, without sharing of Health Insurance Portability and Accountability Act (HIPAA) information to outside investigators or sources. No identifying patient information has or will be published for the purposes of this study. No compensation was provided for this study. No interventions, clinical recruitment, or patient-level interactions were conducted.

Prompt Engineering Process

An iterative approach was used to create prompts to extract binary (yes or no) responses from clinic notes for the 10 patients

XSL•FO
RenderX

in the prompt engineering cohort. We identified 10 clinical questions related to ILD classification, treatment, and progression. Prompts were initially created in a simple question format and tested using GPT-4o (OpenAI) on the prompt engineering cohort. Given poor performance with simple prompts, chain-of-thought and heuristic prompts were developed for each clinical question following prior published methods on prompt engineering [9,10]. We used ChatGPT (OpenAI) to help create the chain-of-thought and heuristic prompts based on the simple format prompts. Chain-of-thought prompts instructed models to reason step-by-step through relevant clinical criteria before producing a final yes/no output (eg, "Please create a chain-of-thought prompt to determine if a patient has ILD, reasoning through the relevant clinical features before giving a final yes or no answer."). Heuristic prompts detailed specific instructions and constrained the model to output only a binary answer (eg, "Please create a heuristic prompt to determine if a patient has ILD, answering only yes or no."). Heuristic prompts demonstrated the highest accuracy when tested on the prompt engineering cohort with GPT-4o and selected as the final prompt format. Based on a review of incorrect responses in the prompt engineering cohort, an ILD physician iteratively refined the prompts by adjusting the rules accordingly (eg, adding missed definitions, incorporating negative constraints to prevent over-classification, or clarifying ambiguous terminology from notes) and retesting prompts until 100% accuracy was achieved in the prompt engineering cohort. This process took an estimated 3-5 iterations, depending on the prompt. The final heuristic prompts were applied uniformly to all 12 LLMs without model-specific modifications for validation, to allow for direct cross-model comparison under identical prompting conditions. Sample representations of prompts from initial creation to final iteration are shown in Textbox 1.

XSL•FO
RenderX

Textbox 1. Sample interstitial lung disease clinical note binary extraction prompts from initial creation to final version.

Initial prompts

- Does the patient have interstitial lung disease (ILD) as explicitly stated in the note? Answer only "yes" or "no."
- Has the patient had any respiratory-related hospitalizations based on the note? Respiratory-related hospitalizations are defined as any hospitalization due to increase in pulmonary symptoms. Answer only "yes" or "no."

Final prompts

- Does the patient have interstitial lung disease (ILD) as explicitly stated in the note? Answer only "yes" or "no." Keep the following rules in mind:
- Answer Yes if any of the following are explicitly mentioned in the note:

i. Diagnosis of interstitial lung disease (ILD), even if the type is unclassified or pending further work-up (eg, "ILD of uncertain etiology" is considered an ILD)

ii. Diagnosis of idiopathic pulmonary fibrosis (IPF)

iii. Diagnosis of sarcoidosis or sarcoid

iv. Diagnosis of pulmonary fibrosis or fibrotic lung disease

v. Presence of specific CT scan findings such as reticulation, honeycombing, traction bronchiectasis, fibrosis, usual interstitial pneumonia (UIP), nonspecific interstitial pneumonia (NSIP), or hypersensitivity pneumonitis (HP)

vi. Pathology showing ILD, fibrosis, UIP, NSIP, or HP

vii. Patient is on nintedanib (Ofev) or pirfenidone (Esbriet)

- Answer No if:

i. None of the above diagnoses, findings, or medications are mentioned in the note

ii. The diagnosis is explicitly ruled out (eg, No evidence of ILD)

iii. Patient has bronchiectasis without evidence of associated ILD

iv. ILD is not suggested or ruled out, and no related findings are present in the note

- If there is any discordance between different sections of the note (eg, a tentative diagnosis like possible sarcoidosis in the HPI but definitive diagnosis in the assessment and plan section), always use the diagnosis stated in the assessment and plan section to guide your answer
- Has the patient had a respiratory related hospitalization based on the note? Answer only "Yes" or "No." Keep the following rules in mind:
- Answer Yes if any hospitalization or admission is mentioned in the note and is due to any of the following criteria:

i. Respiratory symptoms (eg, cough, shortness of breath/dyspnea, and increased or new oxygen requirement)

ii. Exacerbation of underlying lung disease

iii. Respiratory infection or illness

iv. Pleural effusion

v. Pneumothorax

vi. Volume overload

vii. Left or right sided heart failure (CHF, CHF exacerbation, HFrEF or HFpEF exacerbation)

- Answer No if:

i. There are no hospitalizations mentioned in the note (Note: emergency room or urgent care visits do not count as a hospitalization)

ii. Any hospitalization mentioned does not meet any of the criteria listed above

An additional multiclass response prompt for ILD classification was derived directly from the binary heuristic prompts regarding ILD diagnosis, and a separate iterative process was not undertaken. A complete list of final binary and multiclass prompts used for validation is shown in section S1 in Multimedia Appendix 1.

Protected Health Information Compliance

HIPAA-compliant business associate agreements (BAAs) were signed with OpenAI, Anthropic, and Google to ensure appropriate handling and safeguarding of data prior to transmission, permitting the use of protected health information (PHI) in accordance with HIPAA Privacy and Security rules. Clinical notes were transmitted via API under these BAAs without complete deidentification prior. However, additional preprocessing using regular expressions to remove patient

XSL•FO
RenderX

names, date of birth, and medical record numbers was performed as an added precautionary measure.

LLM Information

We used a HIPAA-compliant pipeline to access the following LLMs: GPT-3.5 Turbo, GPT-4, GPT-4o, GPT-4o mini, o1, o1-mini, o3-mini (OpenAI), Claude 3.5 Sonnet, Claude 3.7 Sonnet (Anthropic), and Gemini 2.0 Pro (Google) [11-19]. We also ran open-weight reasoning models gpt-oss-20b and gpt-oss-120b (OpenAI) models locally on a HIPAA-compliant local graphics processing unit (GPU) machine [20].

Commercial models were accessed through 2 Stanford Health Care institutional gateways: the API Management gateway [21] for GPT-3.5 Turbo, GPT-4, GPT-4o, GPT-4o mini, o1-mini, and Claude 3.5 Sonnet; and the AI Hub gateway [22] for o1, o3-mini, and Claude 3.7 Sonnet. These gateways proxy requests to Azure OpenAI (GPT and o-series) and Amazon Web Services Bedrock (Claude) backends, respectively.

Gemini 2.0 Pro was accessed directly via the Google Vertex AI software development kit (project som-nero-phi-rishiraj-ild-rr, region us-central1). The specific model version used was the experimental build gemini-2.0-pro-exp-02-05, as this was the only Gemini 2.0 Pro variant available on Vertex AI at the time of data extraction. For Anthropic Claude models, the Bedrock model Amazon Resource Names include version-pinning dates (eg, claude-3-5-sonnet-20241022-v2:0 or claude-3-7-sonnet-20250219-v1:0). For OpenAI models accessed through the institutional gateways, the gateway routes requests by deployment alias (eg, gpt-4o or o1); the underlying model snapshot version is managed by the institution and is not returned in the API response, which is an inherent limitation of managed Azure OpenAI deployments. The exact API deployment identifiers, full end point URLs, and access protocols for each model evaluated in this study are summarized in Table S1 in Multimedia Appendix 2.

Both gpt-oss-120b and gpt-oss-20b were obtained from their official HuggingFace repositories (openai/gpt-oss-120b, commit b5c939de8f754692c1647ca79fbf85e8c1e70f8a; openai/gpt-oss-20b, commit 6cee5e81ee83917806bbde320786a8fb61efebee) with no fine-tuning or modifications applied to the released weights. Inference was performed using vLLM 0.10.1+gptoss on NVIDIA H100 NVL GPUs. gpt-oss-120b was served as a single instance with tensor parallelism (TP) across 2 GPUs (TP=2), while gpt-oss-20b was served as 2 independent instances (TP=1, 1 per GPU, ports 8000 and 8001). Both models used fp16 KV cache, max-model-len of 65,536, max-num-seqs of 32, and max-num-batched-tokens of 8192, with chunked prefill, prefix caching, and async scheduling enabled. GPU memory use was set to 0.90. The complete inference configuration is summarized in Table S2 in Multimedia Appendix 2.

Clinical Data Extraction Through the LLM Pipeline

Each note-binary prompt combination in the validation cohort (which was kept strictly separate from the prompt engineering cohort) was processed through the models 10 times to assess consistency of the LLMs, with each run analyzed separately. Therefore, each model had 10,000 inference calls for binary clinical data extraction based on 100 notes, 10 binary prompts, and 10 runs. In addition, each note-multiclass prompt combination was processed through the 2 open-weight models, gpt-oss-20b and gpt-oss-120b 10 times, thus an additional 1000 inference calls for these 2 models. These models were selected specifically for multiclass clinical data extraction because they share the same underlying architecture, differing only in parameter count, allowing for an assessment of how model scale influences performance for more complex extraction tasks. The temperature parameter for GPT-3.5 Turbo, GPT-4, GPT-4o, GPT-4o mini, Claude 3.5 Sonnet, Claude 3.7 Sonnet, and Gemini 2.0 Pro was set at 0.0 to maximize reproducibility [23]. Claude 3.7 Sonnet was evaluated with extended thinking disabled.

For the OpenAI reasoning models accessed via the commercial API (o1, o1-mini, and o3-mini), the temperature parameter was fixed at 1.0 and could not be modified by the user. The reasoning_effort parameter, which controls the number of internal reasoning tokens generated before producing a response, was not explicitly set for these models; per OpenAI's API documentation, this defaulted to "medium." The o1-mini model does not support the reasoning_effort parameter. Similarly, the open-weight reasoning models gpt-oss-20b and gpt-oss-120b temperature parameter was also fixed at 1.0 and not modified by the user.

Commercial and cloud-hosted models (GPT-3.5 Turbo, GPT-4, GPT-4o, GPT-4o mini, o1, o1-mini, o3-mini, Claude 3.5 Sonnet, Claude 3.7 Sonnet, and Gemini 2.0 Pro) were evaluated between February and March 2025. The 2 open-weight models (gpt-oss-20b and gpt-oss-120b) were evaluated after their public release in August 2025.

Reference Standard for Ground Truth

A total of 3 ILD clinicians independently reviewed the same clinical notes and provided responses based on the prompts. They then convened to reach a consensus answer for each note-prompt combination, which served as the ground truth against which to score the LLMs. The scorers were blinded to each other's and LLM responses during individual scoring, and to LLM responses during the consensus meeting.

Statistical Analysis

Accuracy for each LLM model and for each of the ILD clinicians was determined by comparing their responses to the ground truth and reported as the mean across the 10 runs per patient. Precision, recall, and F_1-score were calculated for each LLM. For the multiclass analysis, model outputs were mapped to ILD categories using the following rule: any response containing "unclassifiable" was assigned to the unclassifiable category (reflecting clinical practice in which the unclassifiable designation takes precedence even when specific subtypes are also under consideration); otherwise, the first-listed class was taken as the primary diagnosis. Macro- and weighted-averaged precision, recall, and F_1-score were computed across the ILD categories. Consistency between runs for a given model was assessed by calculating the percent agreement of the majority prediction for each note-prompt combination. Interrater reliability among the 3 ILD physician responses was measured

XSL•FO
RenderX

using the Fleiss κ statistic [24]. For McNemar pairwise comparisons, the majority-vote consensus prediction across 10 runs was taken as each model's response. Differences in model accuracy were analyzed using the Cochran *Q* test followed by post hoc pairwise comparisons conducted using the McNemar test with Bonferroni correction for multiple comparisons (78 pairs; corrected α=.00064). Differences in model accuracy to human accuracy were conducted using the McNemar test with the Holm-Bonferroni step-down procedure (family of 12 comparisons; family-wise α=.05) [25,26]. Differences between LLM models in the multiclass task were assessed using the McNemar test on majority-vote consensus predictions and an independent-samples *t* test on per-run macro-F_1. Statistical analysis was performed in Python. In the absence of established sample size guidelines for LLM evaluation, a cohort of 100 patients was deemed to be an appropriate size for this proof-of-concept study.

Results

Of the 3986 ILD clinic patients assessed for eligibility, 2888 were excluded for not meeting the inclusion criteria. Of the eligible 1098 patients, 110 patients were randomly selected, 10 for prompt engineering and 100 for LLM validation. A flow diagram of cohort selection is shown in Figure 2.

The median word count for the ILD clinical notes was 1788 (IQR 1214-2344) words. Each LLM took approximately 1-2 seconds to process a clinical note-prompt combination. Models were accessed via API with rate limits of up to 30,000,000 tokens/minute (eg, GPT-4o), allowing for efficient parallel processing of clinical notes (1 token is approximately equal to 1 word). Pricing varied across models, with estimated API costs ranging from less than US $0.001 to US $0.11 per clinical note-prompt combination. Accounting for 10 binary prompts per note and 10 repeated runs per binary prompt-note combination, estimated total processing costs ranged from approximately US $0.05 to US $10.50 per clinical note depending on the model used (see section S2 in Multimedia Appendix 1 for detailed cost breakdown for LLMs),

OpenAI's gpt-oss-120b and o1 achieved the highest accuracy at 97.5% (95% CI 97.3%-97.7%) and 97.3% (95% CI 97.1%-97.4%), respectively. The mean accuracy of the 3 ILD physicians was 96.2% (95% CI 92.3%-100.0%). Cochran *Q* test demonstrated statistically significant overall differences in accuracy across models (Q=1437.45, df=12; P<.001). There was no statistically significant difference in accuracy between the ILD physicians and 7 of the models: Claude 3.5 Sonnet, GPT-4o, o1, o1-mini, o3-mini, gpt-oss-20b, and gpt-oss-120b.

Among these top-performing models, there was no significant difference in accuracy from each other except between o1 and GPT-4o. The other 5 models (GPT-4, GPT-4o-mini, GPT-3.5 Turbo, Claude 3.7 Sonnet, and Gemini 2.0 Pro) had significantly lower accuracy compared to ILD physicians. GPT-3.5 Turbo was the worst-performing model, with significantly lower accuracy than every other model. Mean accuracy for each model, along with the ILD physicians, is shown in Figure S1 in Multimedia Appendix 3. Pairwise McNemar test results comparing model accuracy and pairwise McNemar test results comparing model accuracy to human accuracy are shown in Tables S1 and S2, respectively, in Multimedia Appendix 4.

Prompt 8, asking whether a patient has sarcoidosis, performed the best of all the prompts with nearly 100% accuracy for all models except for GPT-3.5 Turbo (accuracy 45.2%). Prompt 9, asking about progression of respiratory symptoms, in general, had the worst accuracy of all the prompts, ranging from 80%-90% accuracy. A detailed breakdown of each model's mean accuracy over 10 runs by prompt is shown in Table S3 in Multimedia Appendix 4.

The precision, recall, and F_1-score for each model averaged over their 10 runs are shown in Table 1. OpenAI's gpt-oss-120b, o1, and o3-mini models achieved the highest F_1-scores of all the evaluated LLMs.

Table 1. Average accuracy, precision, recall, and F1-scores over 10 runs for each large language model for binary clinical data extraction.

Model	Accuracy, % (95% CI)	F_1-score	Precision	Recall
gpt-oss-120b	97.5 (97.3-97.7)	0.962	0.955	0.970
o1	97.3 (97.1-97.4)	0.958	0.969	0.948
o3-mini	97.2 (97.0-97.3)	0.957	0.962	0.953
gpt-oss-20b	97.0 (96.7-97.2)	0.955	0.944	0.965
o1-mini	96.6 (96.4-96.8)	0.948	0.960	0.937
Claude 3.5 Sonnet	95.5 (95.3-95.6)	0.931	0.950	0.913
GPT-4o	95.5 (95.3-95.6)	0.930	0.957	0.903
GPT-4	94.4 (94.4-94.5)	0.931	0.915	0.948
GPT-4o-mini	94.2 (94.1-94.3)	0.909	0.946	0.875
Claude 3.7 Sonnet	92.0 (91.8-92.2)	0.870	0.959	0.796
Gemini 2.0 Pro	90.6 (90.4-90.7)	0.850	0.895	0.811
GPT-3.5 Turbo	70.2 (70.0-70.3)	0.681	0.535	0.938

XSL•FO
RenderX

All models with temperature 0.0 achieved close to 100% agreement over 10 runs. The reasoning models with temperature 1.0 had percent agreement ranging from 93%-100%. A detailed breakdown of each model's mean percent agreement by prompt is shown in Table S4 in Multimedia Appendix 4. The Fleiss κ for the 3 ILD physicians was 0.88, indicating a high level of agreement.

For the multiclass prompt evaluation, precision, recall, and F_1-scores as macro and weighted averages (mean and SD) across the 7 ILD categories are reported in Table 2. Multiclass prompts had lower accuracy compared with binary prompts at 91.1% (vs 97.5%) and 88.0% (vs 97.0%) for gpt-oss-120b and gpt-oss-20b, respectively. gpt-oss-120b model demonstrated significantly higher macro-F_1 across runs (independent-samples t test, t_{18}=2.93; P=.009) compared to gpt-oss-20b and approximately half the run-to-run variability of the 20b model. Patient-level comparison using the McNemar test revealed no significant difference in accuracy between the 2 models (0.91 vs 0.88; McNemar P>.99), with 89 of 100 patients classified concordantly correct. Per-class F_1-score was ≥0.80 in all 7 ILD categories for both models. Of note, both models achieved perfect per-class F_1-scores (1.00) for idiopathic pulmonary fibrosis (n=17) and sarcoidosis (n=4), with small class size a possible contributor when interpreting the results. A detailed breakdown of per-class F_1-score, precision, and recall is shown in Table S5 in Multimedia Appendix 4.

Table 2. Precision, recall, and F1-scores as macro and weighted averages for multiclass response prompt. Values are mean and SD across 10 runs (n=100 patients per run).

Metric	gpt-oss-120b, mean (SD)	gpt-oss-20b, mean (SD)
Accuracy	0.911 (0.014)	0.880 (0.028)
Macro precision	0.926 (0.009)	0.907 (0.022)
Macro recall	0.924 (0.015)	0.897 (0.023)
Macro-F1	0.923 (0.012)	0.901 (0.020)
Weighted F1	0.913 (0.013)	0.889 (0.024)

Discussion

Principal Findings

In this proof-of-concept study, we demonstrate that multiple transformer-based LLMs can reliably and efficiently extract unstructured data from ILD clinical notes into binary data. With appropriate prompt engineering, 7 of the 12 tested models consistently achieved high accuracy over various ILD prompts, similar to that of ILD clinicians (96.2%), while being orders of magnitude faster and cheaper. Additionally, these models all demonstrated excellent performance with an F_1-score of >0.9. High precision, recall, and F_1-scores are critical for appropriate LLM use in clinical research when using it to screen patients for studies or to create and maintain ILD registries or databases. High precision reduces false positives, preventing the inclusion of ineligible patients that can introduce bias and skew research results. High recall reduces false negatives, ensuring that all relevant patients are included in the study, thus preserving the statistical power of the study. A high F_1-score ensures a strong balance between precision and recall.

Interestingly, we noted that there was degradation in performance in 2 of the newer LLMs, Claude 3.7 Sonnet and Gemini 2.0 Pro (both released in February 2025), compared to the older models Claude 3.5 Sonnet (released June 2024) and GPT-4o (released May 2024). This finding could reflect a tension between the model training or alignment and the strict prompt output constraints. The newer LLMs undergo increased layers of training and reinforcement learning from human feedback [18]. While this allows models to perform more complex reasoning tasks and provide elaborate responses, this may paradoxically reduce their ability to follow constrained instructions, such as "answer only yes or no," as specified in our binary prompts. Prior work has shown that LLMs are vulnerable to negative or distractor requirements [27,28]. Our findings may reflect a related but distinct phenomenon where LLMs become vulnerable following instruction requirements that do not align with their pretraining. However, we do note that Claude 3.7 Sonnet was evaluated with extended thinking disabled, and thus the degradation observed in that model cannot be attributed to its extended reasoning mechanism. Instead, it appears to stem from changes in the underlying base model's training and alignment. The role of extended reasoning in constrained-instruction performance remains an open question that this study was not designed to address. Regardless, it is likely important to tailor model selection to the clinical task and recognize that newer, more advanced models are not necessarily better for all tasks.

A qualitative analysis at the prompt level was performed to better understand differences in prompt accuracy across models and human assessors. Prompt 8, asking whether a patient has sarcoidosis, had the best accuracy for all models and among human assessors. One possible reason is that sarcoidosis is generally diagnosed by biopsy, which, if positive, is usually unambiguous and likely to be clearly documented in clinical notes. In contrast, other ILDs may carry lower diagnostic confidence and are often documented with multiple differential diagnoses. In contrast, prompt 9, assessing progression of respiratory symptoms, had the worst accuracy. This is likely due to the vague and often contradictory documentation of symptoms. For example, multiple clinical notes would document increased cough in the *history* section, but in the *assessment and plan* section, indicate "clinically stable" or "no changes in respiratory symptoms." "Carrying forward" text from previous notes contributed to many of these inconsistencies, which led to confusion on how to answer the prompt and greater

XSL•FO
RenderX

disagreements among the human assessors. The difference in prompt accuracies highlights how the quality of clinical note documentation impacts clinical data extraction that is not solely attributable to LLM deficiencies. Therefore, while using and assessing LLMs, it is also important to consider the underlying source from which they are extracting information.

LLMs with temperature 0.0 had close to 100% agreement for each clinical note-prompt combination, demonstrating high consistency in their outputs with repeated runs. In contrast, the reasoning models with temperature 1.0 had a wider range in agreement. Therefore, in certain clinical data extraction tasks where reproducibility is paramount, choosing a model with an adjustable temperature parameter to 0.0 may be preferable without sacrificing accuracy. Furthermore, given the near-deterministic behavior observed with a temperature of 0.0, a single API call per note would likely be sufficient for clinical deployment.

As a secondary exploratory analysis, multiclass data extraction was assessed using the locally deployed gpt-oss-20b and gpt-oss-120b models. Multiclass extraction resulted in lower accuracy and macro-F_1 for ILD classification compared with binary extraction using the same model (gpt-oss-20b: accuracy 88.0% vs 97.0%; macro-F_1 0.901 vs 0.955; gpt-oss-120b: accuracy 91.1% vs 97.5%; macro-F_1 0.923 vs 0.962). Although there was no significant difference in accuracy between the 2 models for multiclass prompts (McNemar P>.99), gpt-oss-120b demonstrated significantly higher macro-F_1 across runs (P=.009) and approximately half the run-to-run variability compared to gpt-oss-20b. This suggests that the increased model scale primarily confers higher consistency rather than peak accuracy in the multiclass task. Similarly, binary extraction also demonstrated minimal differences in accuracy and computational efficiency across model capacities (excluding GPT-3.5 Turbo). Furthermore, the top-performing standard instruction-tuned models (Claude 3.5 Sonnet and GPT-4o) demonstrated no significant difference in accuracy compared to the reasoning models (o1, o1-mini, o3-mini, gpt-oss-20b, and gpt-oss-120b), despite the latter using extended chain-of-thought inference at substantially higher computational cost. This suggests that for structured clinical extraction tasks, whether binary or more complicated multiclass tasks, more advanced or larger LLM architectures may not confer substantial benefit. Rather, the prompts are more important for success in clinical extraction tasks. Accordingly, for structured binary data extraction from clinical notes, smaller and older models are likely to offer a more favorable cost-to-benefit ratio without sacrificing performance.

In addition to lower accuracy, multiclass extraction produced greater variability in model responses, requiring additional postprocessing to standardize outputs for analysis. This variability primarily had 2 forms: inconsistent use of abbreviations and class names (eg, “HP” vs “hypersensitivity pneumonitis”) and production of ranked differential diagnoses rather than single-class labels (eg, output: “unclassifiable, idiopathic pulmonary fibrosis, HP”). The latter raises a methodological question of how such outputs should be graded and requires careful consideration of labeling rules prior to analysis. In our study, we treated any response containing “unclassifiable” as an unclassifiable prediction, reflecting clinical practice in which this designation takes precedence over specific subtype considerations during multidisciplinary discussion. These findings highlight the increased variability of multiclass extraction, placing a greater burden on clinical researchers to correctly interpret and normalize data extraction results, which can reduce the efficiency gains of multiclass extraction compared with binary extraction.

Given that LLMs are designed to predict the next word based on their training data rather than retrieve information from a database, hallucinations are inevitable [29]. Hallucinations occur when LLMs generate plausible outputs that are factually incorrect. However, they can be mitigated through model training and thoughtful prompt engineering. Simple prompts, which asked a direct question without additional detail, performed poorly across all models. In contrast, heuristic prompts, which provided explicit rules, enabled multiple LLMs to accurately extract binary data regarding ILD classification, treatment, and progression. These findings are consistent with prior studies on prompt engineering [9,10]. Thoughtful prompt engineering was crucial to obtain desired clinical data extraction results from LLMs and mitigate some of the limitations stemming from vague or contradictory clinical note documentation. However, as noted above, we suspect that prompt constraints may have contributed to performance degradation in newer models such as Gemini 2.0 Pro. Thus, prompt engineering should be tailored to both the specific LLM and the clinical task at hand to obtain the best performance, rather than a generic prompt that may not optimize performance across all models. While our prompts were developed and validated on ILD clinical notes from a single center, limiting direct generalizability of our specific prompts to other clinical specialties and institutions, the underlying prompt engineering methodology of developing heuristic prompts with explicit clinical rules is broadly applicable.

Our study showed that human assessment also failed to achieve 100% accuracy; in fact, there was a wider range in accuracy among the 3 ILD physicians compared to most of the evaluated LLMs. The human errors were the result of failing to follow the prompt instructions or failing to pick up information hidden within the verbose note. With increased volume of patients and/or length of clinical notes to review, the potential for human error is often likely to become more pronounced and can greatly depend on the individual chosen for evaluation. In contrast, LLM performance and efficiency are not affected by volume.

We believe that our proof-of-concept study has meaningful implications for improving clinical workflows, especially in clinical research. Manual structured clinical data extraction from clinical notes represents a major bottleneck for clinical research and creating patient data registries; it is time-consuming and labor-intensive. Furthermore, it is prone to human error. Use of LLMs would enable automated data extraction for disease-specific patient registry creation (eg, ILD registries) and cohort selection for clinical and quality improvement research (eg, identifying patients with idiopathic pulmonary fibrosis for a study), at a fraction of the time and cost required with traditional chart review. It would also free up clinicians

XSL•FO
RenderX

and researchers to focus their efforts on developing higher-order clinical and scientific questions for the advancement of knowledge, rather than on the tedious process of chart review and data abstraction. The availability and high accuracy of open-weight models such as gpt-oss-20b and gpt-oss-120b further expand implementation possibilities, as these models can be locally housed within an institution's infrastructure without the need to transmit PHI data to an external API, greatly decreasing risks of privacy concerns.

While we demonstrate that multiple LLMs can be effectively leveraged for clinical data extraction, translating this success for use in clinical research practice requires key operational steps. First, a robust and secure infrastructure pipeline is required to ensure PHI compliance when using LLMs. User inputs and prompts sent to LLMs on remote servers (eg, GPT series models from OpenAI, Claude series models from Anthropic, and Gemini series models from Google) are stored and often used to further train these models; however, entities owning and running these LLMs do not provide specific details regarding this process. These LLMs, therefore, cannot be used without appropriate safeguards to process PHI-containing data. While we used regular expressions during the preprocessing of input notes to remove patient names, date of birth, and medical record numbers as an added precautionary measure, this is not sufficient for PHI compliance, given the brittle nature of regular expressions when applied to unstructured narrative clinical notes. Options to process PHI data using these LLMs include either running smaller models locally on a computer with adequate GPU memory (eg, the Llama series models by Meta) or signing a HIPAA-compliant BAA with providers such as OpenAI or Anthropic. The HIPAA-compliant BAA allows for "zero retention," where the data sent to the LLMs are handled appropriately with safeguarding of PHI, not used to train models, and are HIPAA compliant in other respects as well [30,31]. We ensured that all LLMs used in our study were HIPAA compliant, with BAAs in place. Second, careful and thoughtful prompt engineering is required. This involves testing and refining prompts on a representative set of clinical notes to determine that LLMs are providing the desired output. Third, performance characteristics of the final prompts need to be validated on the data they will be used on before incorporating them into the research pipeline.

Comparison to Prior Work

This study builds on prior work by evaluating multiple LLMs and comparing their performance. Prior studies largely compared limited number of models for clinical data extraction, while we systematically compared 12 models spanning different developers and model generations including standard instruction-tuned models (eg, GPT-4 and Claude 3.5 Sonnet), reasoning models (o1, o1-mini, and o3-mini), and open-weight reasoning models that can be deployed locally (gpt-oss-20b and gpt-oss-120b) [32]. Furthermore, we compared LLM performance to a human benchmark (in our case, ILD specialists), which is the current gold standard for clinical data extraction. Additionally, by using ILD clinic notes, we extend upon previous studies that largely focused on structured or semistructured reports such as radiology or pathology reports [33]. ILD notes are often complex and narrative, containing multiple or conflicting diagnoses. However, we were able to demonstrate that multiple LLMs could still faithfully extract structured clinical information with clear instructions. Notably, this was achieved without fine-tuning of models. Liu et al [34] demonstrated that fine-tuning of open-source models was required to achieve consistent human-level performance for pathology extraction tasks; however, fine-tuning requires machine learning expertise that many clinicians are not well-versed in. In contrast, our study demonstrates that multiple commercial and open-weight LLMs can achieve human-level accuracy across ILD clinical extraction tasks through prompt engineering alone without the need for fine-tuning, thus reducing the technical barrier for clinician research adoption. Finally, our comparison of binary vs multiclass extraction offers further methodological insight for clinical data extraction. We demonstrate that multiclass extraction is feasible, although it is associated with a loss of accuracy compared with binary extraction. While binary extraction achieved higher accuracy, the reduction in accuracy with multiclass extraction may be acceptable given gains in extraction practicality and reduction of required queries, depending on the intended clinical or research use of the extracted data.

Limitations and Further Work

Our study had several limitations due to runtime considerations and effort cost. First, our prompt engineering cohort was small and randomly selected, limiting our ability to capture edge cases and further refine prompts. This led to some inaccurate outputs that could have easily been adjusted in the prompts. We suspect that a larger, more variable prompt engineering cohort can further enhance the performance of LLMs in binary and multiclass extraction. Second, our prompts were iteratively refined using GPT-4o and uniformly applied to all models. We recognize that this inherently leads to bias toward GPT models. Furthermore, our prompts were zero-shot prompts and restricted models to "answer only yes or no." As a result, standard instruction-tuned models are at a disadvantage compared to reasoning models, given their inability to internally reason and perform chain-of-thought processes. Prompts optimized for one model family may not elicit the same quality of responses from other model families, and the relative performance gaps should be interpreted with caution, with the understanding that prompts may not have been optimized for the specific model. Additionally, our multiclass prompt used for secondary analysis did not undergo the same iterative refinement as our binary prompts and may contribute to the lower performance scores in multiclass extraction compared to binary extraction. For clinical implementation, model-optimized prompts should be used for the best results. Third, our validation cohort was small and limited to a single clinic and institution. Multi-institutional studies would help to determine whether a standardized prompt library could be developed for ILD clinical data extraction more broadly. Further research should assess LLM performance with well-crafted prompts across a broader range of clinic notes and institutions to better evaluate generalizability across different specialties and note types. However, given that multiple models were noninferior to humans in accuracy despite using complex, narrative ILD notes, we hypothesize that with appropriate prompt engineering, high accuracy comparable to human-level

XSL•FO
RenderX

accuracy would be maintained across other clinical settings. Fourth, the same ILD physicians who determined the consensus reference standard were graded on their individual responses to determine the accuracy of the human assessments. This leads to artificial inflation of human performance. While this represents a methodological limitation, it does not undermine our findings in this proof-of-concept study that multiple LLMs are able to achieve human-level accuracy. Rather, if anything, the limitation is conservative with respect to our primary findings. Assessing human accuracy with an independent adjudicator would likely narrow the performance gap between humans and LLMs further and may show additional models that also achieve human-level accuracy. Fifth, while interrater agreement among the 3 ILD physicians to establish ground truth was high (Fleiss κ=0.88), this still reflects disagreement among physicians for 10%-12% of cases. Given that disagreements were resolved by internal consensus rather than an independent adjudicator, this introduces the risk of convened consensus bias, where the final label may reflect social or hierarchical dynamics of the 3 physicians, rather than a purely objective clinical standard. Future work using an independent fourth reader as a tiebreaker or formal Delphi-style anonymous resolution can help reduce the influence of group dynamics on the reference standard against which both humans and LLMs are evaluated. Sixth, we were unable to adjust the temperature parameter for the reasoning models, thus o1, o1-mini, and o3-mini had a fixed temperature of 1.0 compared to the other models which we had adjusted temperature to 0.0. This may lead to some variations in comparison of model performance although prior studies have shown that changes in temperature from 0.0 to 1.0 did not have significant effect on problem solving performance of LLMs [23]. Seventh, we did not adjust for other model parameters such as top_p or top_k. It is possible that adjustment of these parameters or a different combination of parameters, may improve clinical data extraction, especially for the older models. Eighth, the multiclass prompts did not undergo a rigorous iterative refinement process like that of the binary prompts, as it was derived directly from the binary classification prompts and meant to be a preliminary feasibility assessment. Thus, multiclass accuracy may improve with further prompt refinement in future studies. Finally, our study was designed as a proof-of-concept for structured binary and limited multiclass extraction from clinical notes. We did not evaluate more complex clinical information extraction tasks, such as temporal reasoning across notes, relation extraction, or patient-level aggregation of information from multiple encounters. Given the rapid pace of LLMs' evolution, these represent critical areas for future exploration.

Conclusions

Our findings demonstrate that multiple LLMs can consistently achieve human-level accuracy in extracting structured binary data from unstructured ILD clinical data, while being orders of magnitude faster and cheaper. This was able to be accomplished relying only on prompt engineering and without fine-tuning of models prior to usage. Multiclass extraction is also feasible, although it is associated with a reduction in accuracy.

Acknowledgments

The authors are thankful to Stephanie M Youngquist, who helped design the graphics for Figure 1. We used ChatGPT to assist with improving the grammar and sentence clarity of the final manuscript. These tools were not used for content generation or interpretation of results.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to patient privacy protections and institutional restrictions under the Health Insurance Portability and Accountability Act.

Funding

This study was funded by Boehringer Ingelheim.

Authors' Contributions

SJC and RR conceptualized and designed the study. SJC, MVM, and RR acquired the data and performed statistical analysis. SJC and RR drafted the manuscript. All authors critically revised and finally approved the manuscript.

Conflicts of Interest

MVM received funding to support this study under grant T15LM007033. Authors JC and RR have received research funding from Boehringer Ingelheim. CL has received research support from National Institutes of Health (NIH), National Institute for Biomedical Imaging and Bioengineering (NIBIB), National Heart, Lung and Blood Institute (NHLBI), National Cancer Institute (NCI), Agency for Health Research and Quality (AHRQ); Advanced Research Projects Agency for Health (ARPA-H), and the Gordon and Betty Moore Foundation. He has received consulting fees from Sixth Street, Everlight Radiology, and Gilmartin Capital and honoraria for lectures from the Singapore Ministry of Health, Philips Medical, Canon Medical, and McKinsey & Company. He has received travel support from the Singapore Ministry of Health, Philips Medical, and McKinsey & Company. He has patents pending in collaboration with GE Healthcare. He serves on the Board of Directors of Bunkerhill Health and Sirona Medical. He owns equity in Bunkerhill Health, Sirona Medical, Whiterabbit.ai, Galileo CDS, ADRA.ai, Radiology Partners, TurboRadiology, Abxtract, and Harrison.ai. His institution has received licensing fees from Harrison.ai and grants or gifts from

XSL•FO
RenderX

Amazon Web Services, Bayer, BunkerHill Health, Carestream, CARPL.ai, Clairity, GE HealthCare, Google Cloud, IBM, Kheiron, Lambda, Lunit, Microsoft, Nightingale Open Science, Philips, Siemens Healthineers, Stability.ai, Subtle Medical, VinBrain, Visiana, and Whiterabbit.ai. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Final prompts selected for large language model (LLM) evaluation and LLM processing cost estimation.
[DOCX File , 26 KB - jmir_v28i1e90547_app1.docx]

Multimedia Appendix 2
Deployment identifiers and inference configuration parameters for locally deployed open-weight models (gpt-oss-20b and gpt-oss-120b), including application programming interface end points, HuggingFace repository commits, graphics processing unit hardware specifications, and vLLM serving settings.
[XLSX File (Microsoft Excel File), 20 KB - jmir_v28i1e90547_app2.xlsx]

Multimedia Appendix 3
Mean accuracy of large language models (LLMs) and interstitial lung disease physician assessors with 95% CI. Dark gray fill represents no statistically significant difference in accuracy between humans and LLMs.
[PNG File , 48 KB - jmir_v28i1e90547_app3.png]

Multimedia Appendix 4
Tables for mean accuracy, pairwise McNemar test results comparing model accuracy, pairwise McNemar test results comparing large language model's accuracy to human accuracy, and percent agreement breakdown by model and prompt.
[XLSX File (Microsoft Excel File), 20 KB - jmir_v28i1e90547_app4.xlsx]

References

1. Saczynski JS, McManus DD, Goldberg RJ. Commonly used data-collection approaches in clinical research. Am J Med 2013;126(11):946-950 [FREE Full text] [doi: 10.1016/j.amjmed.2013.04.016] [Medline: 24050485]
2. Penberthy LT, Dahman BA, Petkov VI, DeShazo JP. Effort required in eligibility screening for clinical trials. J Oncol Pract 2012;8(6):365-370 [FREE Full text] [doi: 10.1200/JOP.2012.000646] [Medline: 23598846]
3. Idnay B, Dreisbach C, Weng C, Schnall R. A systematic review on natural language processing systems for eligibility prescreening in clinical research. J Am Med Inform Assoc 2021;29(1):197-206 [FREE Full text] [doi: 10.1093/jamia/ocab228] [Medline: 34725689]
4. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. arXiv. Preprint posted online on March 31, 2023 . [doi: 10.48550/arXiv.2303.18223]
5. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv. Preprint posted online on June 12, 2017 . [doi: 10.48550/arXiv.1706.03762]
6. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent abilities of large language models. arXiv. Preprint posted online on June 15, 2022 . [doi: 10.48550/arXiv.2206.07682]
7. LLM settings? Prompt Engineering Guide. URL: https://www.promptingguide.ai/introduction/settings [accessed 2026-04-04]
8. Chen S, Maddali M, Bluethgen C, Langlotz C, Raj R. Leveraging generative pre-trained transformer (GPT) large language models (LLMs) for interstitial lung diseases (ILD) clinical research. Am J Respir Crit Care Med 2025;211(Supplement_1):A2086. [doi: 10.1164/ajrccm.2025.211.abstracts.a2086]
9. Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. JMIR Med Inform 2024;12:e55318 [FREE Full text] [doi: 10.2196/55318] [Medline: 38587879]
10. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023;25:e50638 [FREE Full text] [doi: 10.2196/50638] [Medline: 37792434]
11. GPT-3.5 Turbo institution. OpenAI. URL: https://openai.com/index/chatgpt/ [accessed 2026-04-04]
12. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 technical report. arXiv. Preprint posted online on March 15, 2023 . [doi: 10.48550/arXiv.2303.08774]
13. OpenAI, Hurst A, Lerer A, Goucher AP, Perelman A, Ramesh A, et al. GPT-4o system card. arXiv. Preprint posted online on October 25, 2024 . [doi: 10.48550/arXiv.2410.21276]
14. GPT-4o mini. OpenAI. URL: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/ [accessed 2026-04-04]
15. OpenAI, Jaech A, Kalai A, Lerer A, Richardson A, El-Kishky A, et al. OpenAI o1 system card. arXiv. Preprint posted online on December 21, 2024 . [doi: 10.48550/arXiv.2412.16720]
16. OpenAI o3-mini system card. OpenAI. 2025. URL: https://openai.com/index/o3-mini-system-card/ [accessed 2026-04-06]

XSL•FO
RenderX

17. Claude 3.5 Sonnet model card addendum. Anthropic. 2024. URL: https://assets.anthropic.com/m/1cd9d098ac3e6467/original/Claude-3-Model-Card-October-Addendum.pdf [accessed 2026-04-04]
18. Claude 3.7 Sonnet system card. Anthropic. 2025. URL: https://www-cdn.anthropic.com/9ff93dfa8f445c932415d335c88852ef47f1201e/claude-3-7-sonnet-system-card.pdf [accessed 2026-06-16]
19. Gemini Team Google, Anil R, Borgeaud S, Alayrac JB, Yu J, Soricut R, et al. Gemini: a family of highly capable multimodal models. arXiv. Preprint posted online on December 19, 2023 . [doi: 10.48550/arXiv.2312.11805]
20. OpenAI, Agarwal S, Ahmad L, Ai J, Altman S, Applebaum A, et al. arXiv. Preprint posted online on August 8, 2025 . [doi: 10.48550/arXiv.2508.10925]
21. API management. Stanford Health Care. URL: http://apim.stanfordhealthcare.org/ [accessed 2026-06-17]
22. AI Hub. Stanford Health Care. URL: http://aihubapi.stanfordhealthcare.org/ [accessed 2026-06-17]
23. Renze M, Guven E. The effect of sampling temperature on problem solving in large language models. 2024 Presented at: Findings of the Association for Computational Linguistics: EMNLP 2024; 2026 June 14; Miami, Florida, USA p. 7356-7356. [doi: 10.18653/v1/2024.findings-emnlp.432]
24. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76(5):378-382. [doi: 10.1037/h0031619]
25. Cochran WG. The comparison of percentages in matched samples. Biometrika 1950;37(3-4):256-266. [doi: 10.1093/biomet/37.3-4.256]
26. Mc N. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12(2):153-157. [doi: 10.1007/BF02295996] [Medline: 20254758]
27. Hwang Y, Kim Y, Koo J, Kang T, Bae H, Jung K. LLMs can be easily confused by instructional distractions. arXiv. Preprint posted online on February 5, 2025 . [doi: 10.48550/arXiv.2502.04362]
28. Jaroslawicz D, Whiting B, Shah P, Maamari K. How many instructions can LLMs follow at once? arXiv. Preprint posted online on July 15, 2025 [FREE Full text]
29. Xu Z, Jain S, Kankanhalli M. Hallucination is inevitable: an innate limitation of large language models. arXiv. Preprint posted online on January 22, 2024 . [doi: 10.48550/arXiv.2401.11817]
30. How can I get a business associate agreement (BAA) with OpenAI for the API services? OpenAI Help Center. URL: https://help.openai.com/en/articles/8660679-how-can-i-get-a-business-associate-agreement-baa-with-openai-for-the-api-services [accessed 2025-05-02]
31. Business associate agreements (BAA) for commercial customers. Anthropic Privacy Center. URL: https://privacy.anthropic.com/en/articles/8114513-business-associate-agreements-baa-for-commercial-customers [accessed 2025-05-02]
32. Chen D, Alnassar SA, Avison KE, Huang RS, Raman S. Large language model applications for health information extraction in oncology: scoping review. JMIR Cancer 2025;11:e65984 [FREE Full text] [doi: 10.2196/65984] [Medline: 40153782]
33. Reichenpfader D, Müller H, Denecke K. A scoping review of large language model based approaches for information extraction from radiology reports. NPJ Digit Med 2024;7(1):222 [FREE Full text] [doi: 10.1038/s41746-024-01219-0] [Medline: 39182008]
34. Liu L, Lian L, Hao Y, Pace A, Kim E, Homsi N, et al. Human level information extraction from clinical reports with finetuned language models. Sci Rep 2025;15(1):45239 [FREE Full text] [doi: 10.1038/s41598-025-28767-z] [Medline: 41286063]

Abbreviations

API: application programming interface
BAA: business associate agreement
GPU: graphics processing unit
HIPAA: Health Insurance Portability and Accountability Act
ILD: interstitial lung disease
LLM: large language model
PHI: protected health information
TP: tensor parallelism

XSL•FO
RenderX

Edited by A Coristine; submitted 30.Dec.2025; peer-reviewed by M Torii, HJ Dai; comments to author 25.Mar.2026; accepted 09.Jun.2026; published 26.Jun.2026.

Please cite as:
Chen SJ, Maddali MV, Langlotz C, Bluethgen C, Chen J, Raj R
Evaluation of Large Language Models for Structured Data Extraction From Interstitial Lung Disease Clinical Notes: Comparative Study
J Med Internet Res 2026;28:e90547
URL: https://www.jmir.org/2026/1/e90547
doi:10.2196/90547
PMID:

©Stephanie Ji Chen, Manoj Venkat Maddali, Curtis Langlotz, Christian Bluethgen, Jonathan Chen, Rishi Raj. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital and Navigational Health Literacy in Swiss Cancer Survivors Compared With the General Population: Cross-Sectional Questionnaire Study

Julia Seinsche[1], PhD; Elena Alder[2], MSc; Tami Wehrmann[1], MSc; Jörg Haslbeck[1], Prof Dr; Karin Ribi[1], Prof Dr; Saskia Maria De Gani[1,2,3], PhD

[1]Careum School of Health, Kalaidos University of Applied Sciences, Gloriastrasse 18a, Zurich, Switzerland

[2]Careum Center for Health Literacy, Careum Foundation, Pestalozzistrasse 3, Zürich, Switzerland

[3]Institute of Primary Care, University of Zurich, University Hospital Zurich, Zurich, Switzerland

Corresponding Author:
Saskia Maria De Gani, PhD
Careum School of Health, Kalaidos University of Applied Sciences, Gloriastrasse 18a, Zurich, Switzerland

Abstract

Background: The number of cancer survivors in Switzerland is increasing. Many individuals face challenges in dealing with health-related information and navigating the health care system. Digitalization offers new care and support opportunities, but its effective use requires digital and navigational health literacy (HL) and self-management skills.

Objective: This study aimed to assess the competencies of cancer survivors in Switzerland, compare them with those of the general population, and identify vulnerable subgroups.

Methods: A cross-sectional online survey was conducted using validated scales from the Health Literacy Survey 2019‐2021 (HLS_{19}), measuring digital HL, interaction with digital devices, frequency of use of digital resources, and navigational HL. Self-management skills were assessed with the Health Education Impact Questionnaire (heiQ). HLS_{19} scores were compared to data from the general Swiss population and a subpopulation with chronic diseases provided by the Federal Office of Public Health, using Mann-Whitney *U* tests, chi-square tests, and independent *t* tests. Regression analyses identified associations of sociodemographic and clinical factors with digital and navigational HL and self-management skills.

Results: A total of 131 cancer survivors (74.8% female, 41% with breast cancer) completed the survey. Only 30% reported high digital HL, while self-management skills were generally high, particularly in "health service navigation" (median 3.6, IQR 0.8 on a 4-point Likert scale) and "constructive attitudes" (median 3.6, IQR 1). Compared to the general population, cancer survivors had significantly higher scores in digital interaction and digital resource use ($P \leq .002$), while no differences were observed in digital or navigational HL. Digital and navigational HL, as well as self-management, were consistently associated with education level, social support, and financial deprivation.

Conclusions: Cancer survivors in Switzerland report low digital and navigational HL, comparable to the general population, while self-management skills were strong overall. Support strategies should simultaneously target individual and systemic factors to improve cancer survivors' ability to access and use digital health resources. Future research should include a more representative sample to validate these findings.

(J Med Internet Res 2026;28:e84228) doi:10.2196/84228

KEYWORDS

digital health literacy; navigational health literacy; self-management skills; cancer survivors; health literacy survey; Switzerland

Introduction

The number of cancer survivors is increasing worldwide due to a combination of factors, including population growth, aging, and improvements in cancer survival [1,2]. In Switzerland, there were approximately 210,350 cancer survivors in 2017 [3], and the Swiss Cancer League estimated that this number will increase to approximately 450,000 in 2022 [4]. The journey of cancer survivors can be characterized by different phases, including the acute phase involving diagnosis and treatment, the transitional phase following successful treatment completion, the extended phase, in which cancer survivors live with cancer as a chronic disease or undergo ongoing treatment, and the cancer- and treatment-free phase, in which late and long-term health and psychological effects may occur [5,6].

Cancer survivors face several challenges throughout these phases, including dealing with a wide range of health-related information and navigating an increasingly complex health and

XSL•FO
RenderX

social care system. They are expected to understand complex information about their diagnosis, learn new medical terminology, take an active role in decision-making regarding treatment and health service options, adhere to treatment plans, assess and manage symptoms, and adopt health-promoting behaviors [7,8]. These challenges are reflected in the priorities reported by cancer survivors, who expressed a desire to better understand their illness's chronicity, learn how to live with cancer, receive information and resources for care and self-management, and be empowered and engaged in their care [9]. The guidelines and models of survivorship care emphasize the importance of supporting cancer survivors' self-management skills to address the medical and emotional consequences of disease and treatment and optimize health and recovery across the cancer journey [6,9-11]. Moreover, self-management is recognized as an essential aspect of general health literacy (HL) in cancer care [9]. HL entails people's knowledge, motivation, and competencies to access, understand, appraise, and apply health information to make judgments and take decisions in everyday life concerning health care, disease prevention, and health promotion to maintain or improve quality of life [12].

With the rapid digital transformation of health care, an increasing number of technology-enabled solutions are being implemented in cancer care [13], providing new opportunities to support cancer survivors and their self-management [14-17]. As digital solutions become more integrated into cancer care, digital HL—the ability to find, understand, evaluate, and use health information from digital sources [18,19]—is becoming increasingly important for cancer survivors' HL. Digital HL enables them to benefit fully from these innovations and manage their care effectively. However, an international report indicates that the general population in many European countries has only moderate levels of digital HL. Notably, the Swiss participants reported particularly low competencies. The corresponding survey conducted in Switzerland found that 72% (1801/2502) of respondents reported difficulties in dealing with digital health-related information to a degree that indicates a widespread low level of digital HL [20]. Furthermore, even lower digital HL scores were observed in those enduring one or more chronic diseases compared with individuals without chronic illnesses [20].

In addition, despite the advances in the digitalization of health care, the Swiss health system remains fragmented and is characterized by insufficient coordination and communication among the different care providers [21,22]. The ability to navigate a health system is described as navigational HL. Specifically, it encompasses the knowledge, motivation, and skills to deal with information for navigating health care systems and services to obtain the most appropriate health care [20,23]. Navigational HL is notably low in most European countries, especially compared to other domains of HL. On average, 45% of the respondents rated navigation tasks as "difficult" or "very difficult." In Switzerland, 74% of the population reported low navigational HL. Individuals living with one or more chronic conditions found it even more difficult to use information from their doctor to manage their condition and reported lower navigational HL than those without chronic conditions [20].

Although the importance of self-management support in survivorship care has been increasingly emphasized [10], digital and navigational HL have not yet been explicitly addressed in survivorship care. Furthermore, there is a paucity of data on digital and navigational HL and self-management skills among cancer survivors in Switzerland.

To address this gap, the primary objective of this study was to assess digital and navigational HL as well as self-management skills among adult cancer survivors in Switzerland, and to compare their levels of digital and navigational HL with those of the general Swiss population.

The secondary objective was to examine the associations between sociodemographic and clinical factors and digital and navigational HL as well as self-management skills.

Methods

Study Design and Procedures

A cross-sectional online survey was conducted to assess digital and navigational HL and self-management skills in adult cancer survivors in Switzerland using nonprobabilistic sampling. To reach cancer survivors in different phases of their disease, a self-declared cancer diagnosis was the only criterion to participate in the survey. Recruitment included advertising the survey at several time points with flyers at health and cancer care institutions, as well as through corporate social media channels (eg, LinkedIn [LinkedIn Corp] and Instagram [Meta Platforms]), especially to minimize the effort required for recruitment by health and social care staff. No preselection of cancer survivors occurred to avoid the screening burden at the supporting institutions. To reach potential participants who preferred analog measures, printed questionnaires were also provided to the recruiting institutions.

Participants completed the survey online (or by paper and pencil, if preferred) using a web-based survey tool (UNIPARK; Tivian XI GmbH) between June and December 2024. The survey was available in German and French, as these are major languages spoken in Switzerland.

Ethical Considerations

The study was submitted to the ethics committee of the Canton of Zurich, which confirmed that the study does not fall under the Human Research Act and therefore granted a waiver (Req-2024 - 00298). The study complies with the principles of the Declaration of Helsinki.

The advertisement for participation included initial information about the study, while more detailed information on the study aim, expected survey completion time, data protection, and anonymity of collected data were provided before starting the survey. After reading this information, participants were asked to confirm their consent to participate.

Measures

The online survey consisted of four parts, including (1) sociodemographic information, (2) questions on digital HL, (3) navigational HL, and (4) self-management skills.

XSL•FO
RenderX

Sociodemographic Information

Sociodemographic characteristics (age, gender, living situation, education based on International Standard Classification of Education [ISCED] levels, experience as a health care professional, place of birth, and mother tongue) were collected alongside clinical characteristics (cancer type, date of diagnosis, and other chronic diseases) and the use of cancer care services.

Additionally, 2 indices were calculated [24]. First, a financial deprivation index, which is based on the following questions answered on a 5-point Likert scale: how easy or difficult it is to (1) afford medication, if needed, (2) afford medical examinations and treatments, if needed, and (3) pay bills at the end of the month. The score was defined as the percentage of items answered with "very difficult" or "difficult" and operationalized into four categories (1) none=0%, (2) some=33.33%, (3) considerable=66.67%, and (4) severe=100%.

Second, a social support index was created based on the Oslo Social Support Scale (OSSS-3) [25]. This scale comprises three items: (1) the number of close and reliable persons (1="no person"-4="6 or more"), (2) interest and concern of other people regarding the participant's life (1="none"-5="a lot"), and (3) difficulty to get practical help from neighbors (1="very difficult"-5="very easy"). A sum score was calculated, ranging between 3 and 14, with a higher score corresponding to a high amount of social support. Following Kocalevent et al [25], the score was operationalized into three categories (1) 3 - 8=poor social support, (2) 9 - 11=moderate social support, and (3) 12 - 14=strong social support.

Digital and Navigational HLs

Overview

To assess digital and navigational HL, the HLS_{19}-DIGI (Health Literacy Survey 2019-2021–Digital Health Literacy) [19,20] and the HLS_{19}-NAV (Health Literacy Survey 2019-2021–Navigational Health Literacy) [26,27] were used as part of the European Health Literacy Survey 2019 - 2021 (HLS_{19}) of the World Health Organization (WHO) Action Network Measuring Population and Organizational Health Literacy (M-POHL). The HLS_{19}-DIGI and the HLS_{19}-NAV indices were constructed following the recommendations of the HLS_{19} Consortium of the WHO Action Network M-POHL [24]. This consortium validated the survey used in each country, including the translations used in the $HLS_{19\text{-}21}$-CH study and this study. The instruments were validated by Cronbach α for reliability, confirmatory factor analysis, and unidimensionality by Rasch analyses. In Switzerland, both instruments showed high reliability (HLS_{19}-DIGI Cronbach α=0.85; HLS_{19}-NAV Cronbach α=0.88) [27].

Digital HL

The HLS_{19}-DIGI consists of three parts:

1. The 8-item HLS_{19}-Digital Health Information (HLS_{19}-DIGI-HI) scale, which assesses the ability to deal with digital health information,
2. The 2-item HLS_{19}-Digital Interaction (HLS_{19}-DIGI-INT) scale, focusing on the interaction with digital resources for health, and
3. The 6-item HLS_{19}-Digital Device (HLS_{19}-DIGI-DD) scale measures the frequency of digital resource usage.

Items of the first 2 parts (HLS_{19}-DIGI-HI and HLS_{19}-DIGI-INT) were operationalized by measuring the perceived difficulty of tasks related to digital health information and resources using a 4-point Likert scale, ranging from "very easy" to "very difficult" [24]. Only participants with >80% (HLS_{19}-DIGI-HI) or 100% (HLS_{19}-DIGI-INT) valid responses (invalid="I don't know" or missing values) were included in all analyses. The respective scores were calculated as the mean of the items' numeric values, scaled from 0 to 100, following the rules proposed by M-POHL. Higher scores indicate a higher level of digital HL [24].

To gain deeper insights into the participants' responses, an index [28] was calculated based on the HLS_{19}-DIGI-HI results. This score is defined as the percentage of items with valid responses that were answered with "very easy" or "easy." Additionally, for easier comprehensibility, the following categories were created.

1. Category excellent: percentage of "very easy" items ≥50%" AND "(very) difficult"<8.334%;
2. Category sufficient: percentage of "(very) easy" items >83.33%;
3. Category inadequate: percentage of "very easy" items <8.334% AND "(very) difficult" ≥50%;
4. Category problematic: all cases which do not fall into one of the other 3 categories.

Participants assigned to the categories "excellent" and "sufficient" were classified as having high digital HL, while those in the "inadequate" and "problematic" categories were classified as having low digital HL [24].

The HLS_{19}-DIGI-DD (frequency of usage) response options range from 1 "not relevant for me or less than once a week" to 5 "more than once per day" in a typical week [28]. As in the HLS_{19}-DIGI-HI scale, only participants with >80% valid responses were included in the analyses. A mean score was calculated as the relative measure for the frequency of use of health-related digital resources. This mean value has no direct interpretation but was used as a potential determinant of digital HL.

Navigational HL

The 12-item HLS_{19}-NAV [26,27] is a self-report questionnaire that evaluates the ability to access, understand, appraise, and apply information on navigational issues, using a 4-point Likert scale ranging from "very easy" to "very difficult." The same response options and score calculations as for the HLS_{19}-DIGI-HI apply, including the analysis of an index score and navigational HL categories. Participants with fewer than 80% valid items were excluded from the analysis. Higher scores indicate a higher level of navigational HL.

XSL•FO
RenderX

Self-Management Skills

Self-management skills were assessed using the Health Education Impact Questionnaire (heiQ). The original self-report questionnaire includes 42 items grouped into 8 dimensions to measure the benefits of patient education programs across chronic conditions [29]. Of these 8 heiQ dimensions, five have been identified as critical dimensions to empower patients with cancer: (1) emotional distress (heiQ3), (2) constructive attitudes and approaches (heiQ5), (3) skill and technique acquisition (heiQ6), (4) social integration and support (heiQ7), and (5) health service navigation [30]. Therefore, for this study, a cancer-specific 25-item version including only these 5 dimensions was used [31,32]. Each question is answered on a 4-point Likert scale ranging from 1 ("strongly disagree") to 4 ("strongly agree"), except for the emotional distress scale, which is scored in reverse. A mean score for each scale was calculated, with higher scores indicating better self-management skills. The heiQ has shown good psychometric properties [31,32].

Statistical Analyses

Sample Size

The sample size calculation was based on the expected proportion of people reporting low digital and navigational HL. For cross-sectional surveys to represent a proportion, the formula for a qualitative variable applies [33]. Different scenarios for the calculation were used, setting the proportion of people with low digital HL to 72% according to the reference of the Swiss population [20] and to 39% or 66%, respectively, according to 2 studies in patients with cancer [34,35], with a 95% CI (z=1.96), and the absolute error set to 5%. This led to a desired sample size between 310 and 366 participants.

Data Analysis

Data were analyzed with SPSS (version 27; IBM Corp). Descriptive statistics were calculated for all sociodemographic variables, HL, and self-management measures. To investigate whether digital HL was associated with the frequency of use of digital devices and resources, a Spearman correlation analysis was performed using the scaled mean values HLS_{19}-DIGI-HI and HLS_{19}-DIGI-INT and the HLS_{19}-DIGI-DD mean value. Differences between the study sample and the Swiss general population (SGP), as well as a subgroup with chronic conditions, were analyzed using Mann-Whitney U tests, chi-square tests, and independent t tests. Concerning the latter, the Welch test was interpreted in case of homogeneity of variances was violated. As gender distribution differed between this study's population and the SGP, comparisons were conducted for the whole sample as well as stratified by gender. The association of sociodemographic and clinical factors with digital and navigational HL and self-management skills was investigated by means of linear regressions with the scaled mean scores of HLS_{19}-DIGI-HI and HLS_{19}-NAV, as well as mean values of all heiQ dimensions, as dependent variables and age, gender, education, health care profession, chronic diseases, migration, financial deprivation, and social support as predictors. For all categorical variables (age categories, education, and chronic diseases), dummy variables were created and included in the regression analyses. As this led to a high number of (potentially irrelevant) predictors, single regression analyses were performed, and only those sociodemographic and clinical factors that showed significant associations with the respective independent variables were then included in the joint regression analyses. In all analyses, a P value of <.05 was considered significant.

Results

Sociodemographic and Clinical Characteristics

A total of 131 participants completed the survey. Table 1 provides an overview of the sociodemographic and clinical characteristics of the participants.

XSL•FO
RenderX

Table . Sociodemographic and clinical characteristics of included participants (N=131).

Variables	Frequency, n (%)
Gender	
Women	98 (74.8)
Men	32 (24.4)
Missing	1 (0.8)
Age group (years)	
30 - 39	9 (6.9)
40 - 49	23 (17.6)
50 - 59	46 (35.1)
60 - 69	38 (29.0)
≥70	14 (10.7)
Missing	1 (0.8)
Highest educational level	
Primary education (ISCED[a] level 1)	2 (1.5)
Lower secondary education (ISCED level 2)	38 (29.0)
Upper secondary education (ISCED level 3)	9 (6.9)
Post-secondary non-tertiary education (ISCED level 4)	34 (26.0)
Short-cycle tertiary education (ISCED level 5)	0 (0.0)
Bachelor (ISCED level 6)	10 (7.6)
Master or Licentiate (ISCED level 7)	25 (19.1)
Doctorate (ISCED level 8)	10 (7.6)
Unknown	3 (2.3)
Living situation	
With a partner, without children	51 (38.9)
With a partner and children	27 (20.6)
Single-parent family	4 (3.1)
Shared flat	3 (2.3)
Single-person household	43 (32.8)
Other type of household (eg, multifamily household)	3 (2.3)
(Previous) work as a health care professional	
Yes	36 (27.5)
No	91 (69.5)
Don't know	4 (3.1)
Born in Switzerland	
Yes	102 (77.9)
No	29 (22.1)
Mother tongue other than language spoken at the place of residence	
Yes	30 (22.9)
No	99 (75.6)

XSL•FO
RenderX

Variables	Frequency, n (%)
Don't know or missing	2 (1.5)
Cancer type	
Pancreatic cancer	3 (2.3)
Bladder cancer	1 (0.8)
Breast cancer	53 (40.5)
Colon cancer	10 (7.6)
Ovarian or cervical cancer	9 (6.9)
Skin cancer	7 (5.3)
Brain tumor or metastases	11 (8.4)
Testicular cancer	1 (0.8)
Lymphoma	13 (9.9)
Liver cancer	3 (2.3)
Leukemia	4 (3.1)
Lung cancer	16 (12.2)
Stomach cancer	1 (0.8)
Kidney cancer	2 (1.5)
Prostate cancer	5 (3.8)
Thyroid cancer	3 (2.3)
Other	21 (16)
Other chronic diseases	
None	71 (54.2)
One	32 (24.4)
More than one	25 (19.1)
Don't know	3 (2.3)
Financial deprivation	
None	103 (79.8)
Some	9 (7.0)
Considerable	7 (5.4)
Severe	10 (7.8)
Social support	
Poor	21 (17.0)
Moderate	48 (39.0)
Strong	55 (44.0)

[a]ISCED: International Standard Classification of Education (8 Levels).

Most participants (53/131, 40.5%) were diagnosed with breast cancer, followed by lung cancer (16/131, 12.2%), lymphoma (13/131, 9.9%), and brain tumors or metastases (11/131, 8.4%). The average time since diagnosis was 6 (SD 5.2) years. Most participants (71/131, 54.2%) did not report any other chronic disease, while 24.4% (32/131) had one and 19.1% (25/131) had multiple chronic diseases. Table 2 displays the number of participants using one or more cancer care services, such as support groups. Complementary medical advice was by far the most frequently sought service (53/131, 40.5%).

XSL•FO
RenderX

Table . Participation in cancer care services (N=131).

Variable	Frequency, n (%)
Nursing care consultation	39 (29.8)
Support groups	25 (19.1)
Complementary medical advice	53 (40.5)
Services by patient organizations	25 (19.1)
None of these services	41 (31.1)
Don't know/missing	2 (1.5)

Digital HL

The mean digital HL index score, thus the average percentage of "very easy" and "easy" items in the HLS_{19}-DIGI-HI, was 57.8 (SD 35.4), with 6.6% (8/122) of the participants having excellent, 30.3% (37/122) sufficient, 36.9% (45/122) inadequate, and 32.8% (40/122) problematic digital HL. Based on these categories, 30.3% (37/122) of participants were classified as having high digital HL and 69.7% (85/122) having low digital HL. The scaled mean score of all participants was 52.88 (SD 20.21), while those with high digital HL had a mean score of 74.77 (SD 10.93).

Looking at the individual items of the HLS_{19}-DIGI-HI (Figure 1), it becomes apparent that the majority struggled with judging whether information is reliable (70/122, 57.40%) or offered with commercial interest (67/122, 54.90%) and to use the information to help solve a health problem (63/117, 53.80%). On the other hand, most participants found it easy to use proper search terms (91/121, 75.20%), understand health-related online information (84/121, 69.40%), and compare the information provided by different websites (84/122, 68.80%). It should be noted that those who responded "I don't know" are not included in the reported percentages, as these responses are considered invalid.

XSL•FO
RenderX

Figure 1. Ability to deal with digital health information (HLS_{19}-DIGI-HI). Note that the response option "I don't know" is not displayed since it is considered an invalid response.

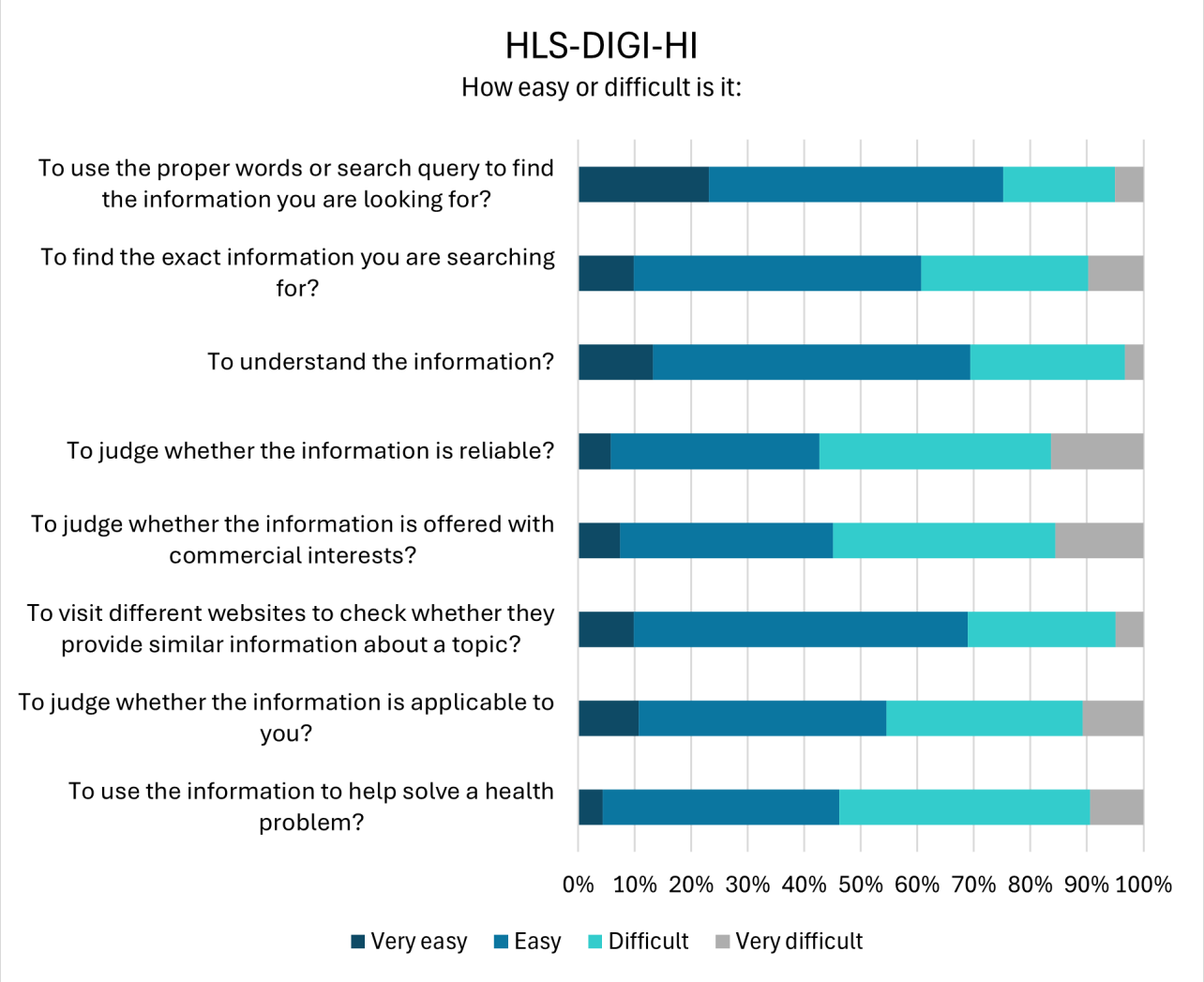

Regarding the interaction with digital resources, the HLS_{19}-DIGI-INT scaled mean score for all participants in this study was 62.86 (SD 27.30). Most participants found it easy to express their opinion in social media forums (65/92, 70.6%) and to formulate a written question for a health care provider (62/92, 67.4%; Multimedia Appendix 1). Notably, 39 out of 131 (29%) participants responded with "I don't know" to either one or both items and were, thus, considered invalid.

In the HLS_{19}-DIGI-DD, participants' mean score was 1.64 (SD 0.67), meaning that on average, participants used digital resources 1 - 3 times per week. Digital devices such as pedometers or smartwatches and health apps were used most frequently, with 25.4% (32/126) or 25.2% (32/127), respectively, using them once or more times a day. Contrary to this, social media and digital interactions with the health system, for instance, for online appointments, were used less than once per week or were considered irrelevant by 77.0% (97/126) and 85.7% (108/126), respectively (Multimedia Appendix 2).

The correlation analyses did not reveal any significant correlations of the mean frequency of use of digital devices and resources (HLS_{19}-DIGI-DD) with either the scaled mean scores of HLS_{19}-DIGI-HI (Spearman ρ=0.06; *P*=.54) and HLS_{19}-DIGI-INT (Spearman ρ=0.09; *P*=.38) or with any of the subitems (Spearman ρ=0.012; *P*=.89; Spearman ρ=0.11; *P*=.32).

Navigational HL

In the HLS_{19}-NAV, participants had a mean navigational HL index score of 53.78 (SD 30.89), with 4.8% (6/124) of the participants categorized as having excellent, 24.2% (30/124) sufficient, 54.8% (68/124) inadequate, and 21% (26/124) problematic navigational HL. Based on these categories, 24.2% (30/124) of participants were classified as having high navigational HL, and overall, participants had a scaled mean score of 51.3 (SD 18.08) points.

Most participants struggled with judging whether a particular health service will meet their expectations (77/122, 63.1%), understanding information on ongoing health care reforms that might affect their health (74/123, 60.2%), and finding out about their rights as a patient (71/122, 58.2%; Figure 2). On the other hand, the majority found it easy to judge which type of health service they need (85/124, 68.5%) and to understand how to get an appointment (103/123, 83.7%).

XSL•FO
RenderX

Figure 2. Level of difficulty of different navigational health literacy (HLS_{19}-NAV) actions. The response option "I don't know" is not displayed since it is considered an invalid response.

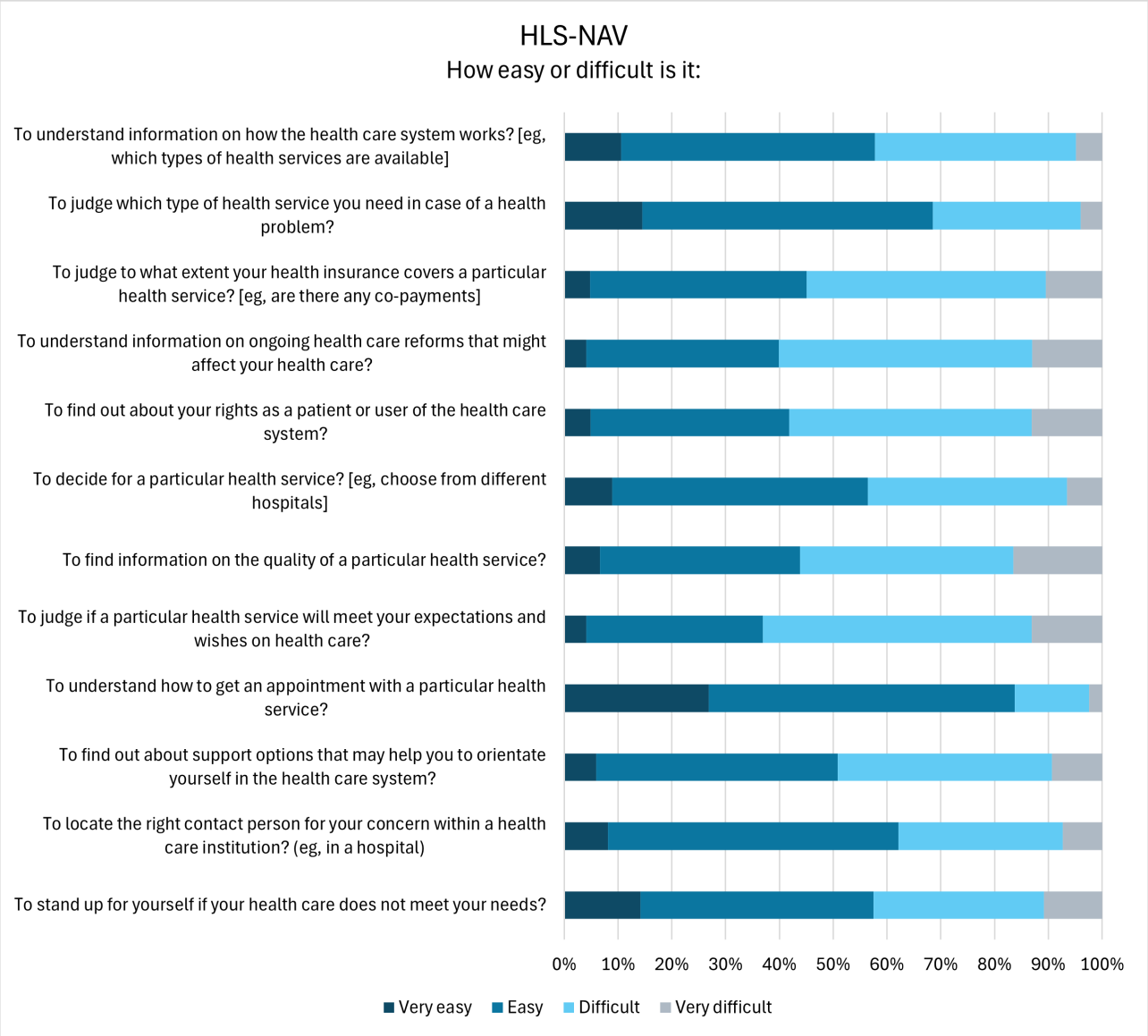

Comparison With the General Swiss Population

Comparing these results with raw data from the $HLS_{19\text{-}21}$ study investigating the general adult population living in Switzerland (SGP), the present study's population was significantly older (U=128055.00; P<.001) and differed in gender distribution (χ^2_1=28.70; P<.001; ϕ=.10), with a higher proportion of female participants compared to the SGP. No significant differences were found in the ability to deal with digital health information (HLS_{19}-DIGI-HI; P=.64). In contrast, both men and women in this study reported a significantly higher ability to interact with digital resources (HLS_{19}-DIGI-INT, P<.001 to P=.02) and a more frequent use of digital devices and resources (HLS_{19}-DIGI-DD, P<.001). Navigational HL (HLS_{19}-NAV) did not differ significantly between groups (P=.87). (Table 3 and Multimedia Appendix 3).

XSL•FO
RenderX

Table . Differences in digital and navigational health literacy between cancer survivors and the Swiss general population. The subpopulation comprises individuals with one or more chronic conditions.

Dependent variable and SGP[a] group	SGP (n)	CS[b] (n)	Test	*t* test (df)	*P* value	MD[c] (95% CI)	Cohen *d*
Digital HL[d] (HLS^{e}_{19}-DI-GI-HI[f] scaled mean value)							
SGP whole group							
Gender							
All	2359	122	*t* test	–0.46 (2479)	.64	−0.85 (–4.46 to 2.75)	0.04
Men	1165	30	*t* test	–1.03 (1193)	.30	–3.76 (–10.91 to 3.38)	0.19
Women	1191	91	*t* test	–0.09 (1280)	.93	–0.18 (–4.43 to 4.06)	0.01
SGP subpopulation							
Gender							
All	852	122	*t* test	–1.96 (972)	.05	–3.84 (–7.69 to 0.00)	0.19
Men	409	30	*t* test	–1.97 (437)	.05	–7.60 (–15.17 to –0.03)	0.37
Women	442	91	*t* test	–1.00 (531)	.32	–2.31 (–6.87 to 2.24)	0.12
Digital HL (HLS_{19}-DI-GI-INT[g] scaled mean value)							
SGP whole group							
Gender							
All	2156	92	*t* test	–3.10 (2246)	<.01	–8.30 (–13.56 to –3.04)	0.33
Men	1082	20	Welch test	–4.08 (20.4)	<.001	−15.95 (–24.10 to –7.80)	0.66
Women	1071	71	*t* test	–2.32 (1140)	.02	−7.44 (–13.73 to –1.14)	0.28
SGP subpopulation							
Gender							
All	752	92	*t* test	–4.15 (842)	<.001	–12.24 (–18.03 to –6.45)	0.46
Men	365	20	Welch test	–5.02 (24.2)	<.001	–20.48 (–28.90 to –12.06)	0.79
Women	386	71	*t* test	–3.02 (455)	<.01	–10.63 (–17.54 to –3.71)	0.39
Navigational HL (HLS_{19}-NAV)[h]							
SGP whole group							
Gender							
All	2478	124	*t* test	–0.17 (2600)	.87	–0.26 (–3.37 to 2.84)	0.02
Men	1214	31	*t* test	–1.56 (1243)	.12	–4.86 (–10.95 to 1.24)	0.28

XSL•FO
RenderX

Dependent variable and SGP[a] group	SGP (n)	CS[b] (n)	Test	*t* test (df)	*P* value	MD[c] (95% CI)	Cohen *d*
Women	1261	92	*t* test	0.38 (1351)	.70	0.71 (–2.94 to 4.36)	0.04
SGP subpopulation							
Gender							
All	910	124	*t* test	–1.04 (1032)	.30	–1.67 (–4.83 to 1.48)	0.10
Men	436	31	*t* test	–2.26 (465)	.02	–6.84 (–12.79 to –0.89)	0.42
Women	473	92	*t* test	–0.04 (563)	.97	–0.09 (–3.94 to 3.76)	0.01

[a]SGP: Swiss general population (data provided by the Federal Office of Public Health).

[b]CS: cancer survivor.

[c]MD: mean difference.

[d]HL: health literacy.

[e]HLS: Health Literacy Survey.

[f]HLS_{19}-DIGI-HI: Health Literacy Survey 2019-2021 - Digital Health Information subscale.

[g]HLS_{19}-DIGI-INT: Health Literacy Survey 2019-2021–Digital Interaction subscale.

[h]HLS_{19}-NAV: Health Literacy Survey 2019-2021–Navigational Health Literacy.

When compared to the SGP subpopulation of participants with a chronic condition, participants were similar in age (U=60785.50; P=.92) but differed in gender distribution (χ^2_1=24.85; P<.001; ϕ=.15). No significant differences were found in digital HL (HLS_{19}-DIGI-HI; P=.05). However, both male and female participants of this study had higher digital interaction abilities (HLS_{19}-DIGI-INT; P<.01) and a higher frequency of digital device use (P<.001). Concerning navigational HL, a significant difference was found only among men (P<.02).

Overall, male and female participants of this study showed a similar pattern when compared with the SGP. However, across all variables, effect sizes were consistently larger among men (Table 3 and Multimedia Appendix 3).

Self-Management Skills

Figure 3 shows that in all subscales, most participants indicated high self-management, with the highest scores in “health service navigation” (median 3.6, IQR 0.8) and “constructive attitudes” (median 3.6, IQR 1) and the lowest in “emotional distress” (median 3.2, IQR 1).

XSL•FO
RenderX

Figure 3. Histograms of Health Education Impact Questionnaire (heiQ) scales. The higher scores indicate better self-management skills; thus, the emotional distress scale was reverse scored so that higher values indicate lower emotional distress.

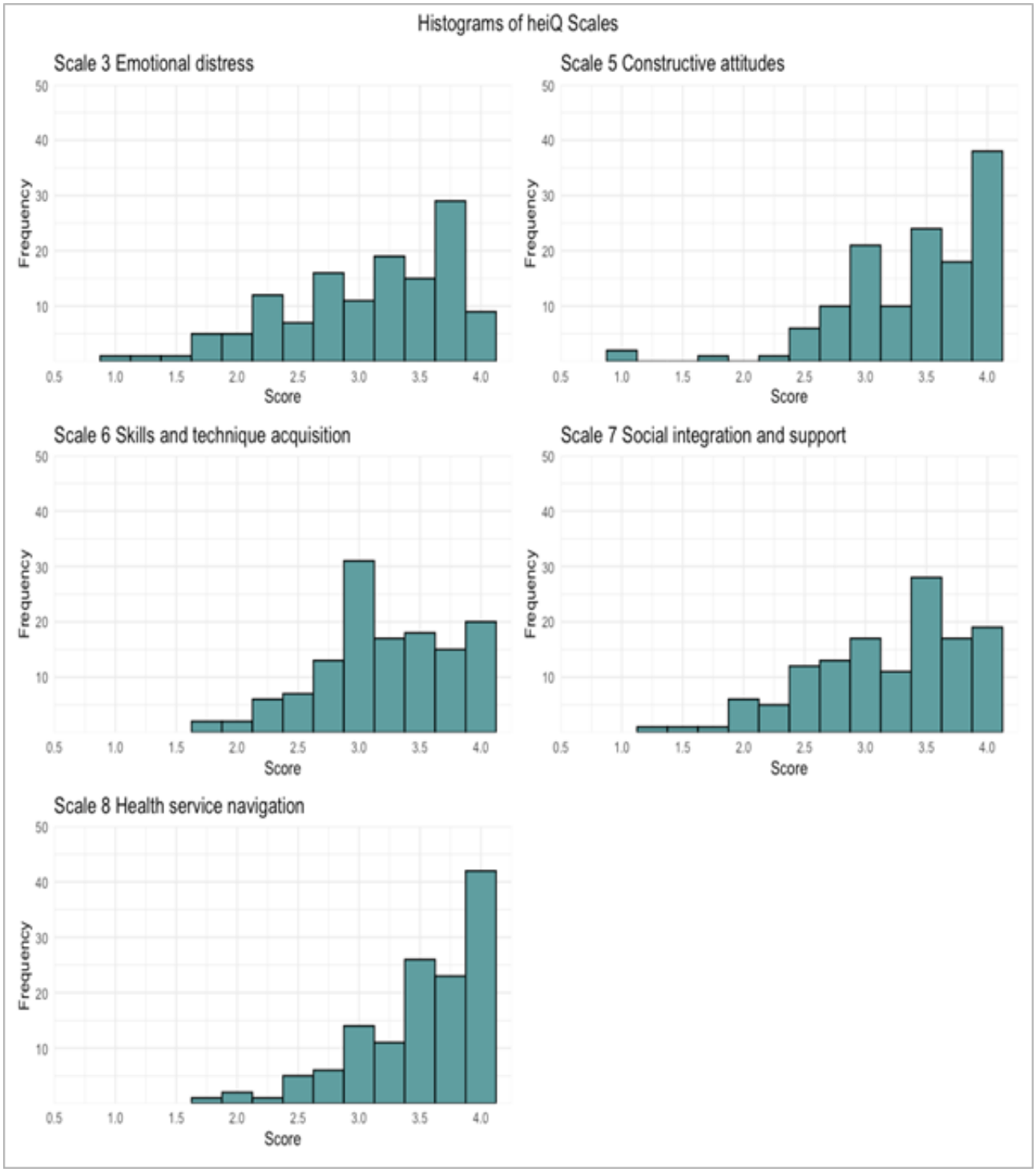

Associations of Sociodemographic and Clinical Factors With Digital and Navigational HL and Self-Management Skills

Digital HL

The single regression analyses revealed significant associations between digital HL (HLS_{19}-DIGI-HI) and financial deprivation (B=−.154; *P*=.01), being a health care professional (B=−9.37; *P*=.02), having the highest level of education (ISCED level 8; B=20.74; *P*=.002), and social support (B=2.30; *P*=.004). Gender, age, migration background, time since diagnosis, and having a chronic illness other than cancer did not show significant associations with digital HL in the single regression analyses. The joint regression model (Regression model 1, Table 4) was statistically significant (*P*<.001) with an adjusted R^2 value of .16, indicative of a moderate/medium goodness-of-fit according to Cohen (1988). In this model, ISCED level 8 (B=19.36; *P*=.003) as well as the level of social support (B=2.11; *P*=.02) had significantly positive effects on digital HL. The standardized coefficients show that education was the strongest predictor in the model (β=0.28).

XSL•FO
RenderX

Table . Regression model 1. The dependent variable in this regression model is the scaled mean score of Health Literacy Survey 2019-2021–Digital Health Information (HLS_{19}-DIGI-HI).

Predictor	B[a] (95% CI)	SE	Standardized β coefficient	*t* test (df)	*P* value
Constant	34.25 (7.04 to 61.47)	13.73	—[b]	2.50 (106)	.01
Financial deprivation	–.06 (–0.19 to 0.07)	0.07	–0.09	–0.92 (106)	.36
Health care professional	3.14 (–5.05 to 11.34)	4.13	0.07	0.76 (106)	.45
ISCED 8[c]	19.36 (6.65 to 32.07)	6.41	0.28	3.02 (106)	.003
Social support	2.11 (0.37 to 3.85)	0.88	0.24	2.41 (106)	.02

[a]B: regression coefficient.

[b]Not applicable.

[c]ISCED 8: International Standard Classification of Education, 8 levels.

Navigational HL

In the single regression analyses, navigational HL was significantly associated with ISCED levels 1 (B=−33.93; *P*=.01) and 8 (B=14.67; *P*=.01), financial deprivation (B=−.24; *P*<.001), being born in Switzerland (B=−11.25; *P*=.004), having no other chronic disease (B=10.08; *P*=.002), and social support (B=2.60; *P*<.001). However, no significant associations with gender, age, time since diagnosis, and having worked as a health care professional were found. The joint regression model (Regression model 2, Table 5) was statistically significant (*P*<.001) and explained a high amount of variance (R^2=.29). In this model, only financial deprivation (B=−.133; *P*=.03) and being born in Switzerland (B=−11.47; *P*=.003) were significantly associated with navigational HL. The standardized coefficients show that being born in Switzerland was the strongest predictor in the model (β=0.248).

Table . Regression model 2.Note. Dependent variable: scaled mean score of Health Literacy Survey 2019-2021–Navigational Health Literacy (HLS_{19}-NAV).

Predictor	B[a] (95% CI)	SE	Standardized β coefficient	*t* test (df)	*P*-value
Constant	42.69 (24.13 to 61.47)	9.36	—[b]	4.56 (106)	<.001
ISCED[c] 1	–22.43 (–45.70 to 0.07)	11.73	–0.16	–1.91 (106)	.06
ISCED 8[c]	9.18 (–1.55 to 11.34)	5.41	0.14	1.70 (106)	.09
Financial deprivation	–.13 (–0.25 to –0.01)	0.06	–0.21	–2.21 (106)	.03
Born in Switzerland	–11.47 (–18.95 to –4.0)	3.77	–0.25	–3.04 (106)	.003
No chronic diseases	5.77 (–0.47 to 12.01)	3.15	0.15	1.83 (106)	.07
Social support	1.45 (–0.04 to 2.95)	0.75	0.18	1.93 (106)	.06

[a]B: regression coefficient

[b]Not applicable.

[c]ISCED: International Standard Classification of Education (8 levels; level 1=primary education–level 8=doctorate).

Self-Management Skills

The regression analyses of the heiQ subscales revealed that three variables significantly influenced the different dimensions of self-management skills: (1) financial deprivation (*P*<.001-.03); (2) the lowest level of education (ISCED 1; *P*=.002-.04); and (3) social support (*P*≤.001-.005) (Multimedia Appendix 4). Only in heiQ6, the time since diagnosis was another significant predictor (*P*=.004). All regression models were statistically significant with *P*<.001 and adjusted R^2 values ranged from R^2=.25 (heiQ3) to R^2=.46 (heiQ7); thus, the models explain a medium to high amount of variation.

Discussion

Principal Findings

Sufficient digital HL, navigational HL, and self-management skills are essential for cancer survivors to successfully manage the many challenges they encounter throughout their cancer journey. The present study delivers first-time quantitative data about digital and navigational HL of cancer survivors in Switzerland in comparison to the general Swiss population and describes their self-management skills.

The results indicate that a substantial proportion of cancer survivors in Switzerland reported difficulties in managing digital health information and services (69.7%) and in successfully

XSL•FO
RenderX

navigating the health care system (75.8%), reflecting generally low levels of digital and navigational HL. Previous studies have yielded inconsistent results regarding digital HL among patients with cancer. For instance, a Canadian study found that 66.3% of patients with lung cancer perceived themselves as having low digital HL [35], whereas other studies reported lower proportions (41.5% [36] and 39.4% [34], respectively) among patients with various cancer types. These discrepancies may partly be explained by the use of different instruments to assess digital HL. In addition, cross-country variations in health care systems and the support provided to cancer survivors by health care professionals and community networks may further contribute to such differences.

In comparison with the general adult population living in Switzerland [20], a slightly lower proportion of cancer survivors in this study reported low digital HL levels, while the proportion of those with low navigational HL was slightly higher. However, statistically significant differences were only found for the ability to interact with digital resources, with cancer survivors reporting higher abilities. The latter finding is unexpected given the increased vulnerability of cancer survivors, their significantly higher age, and the negative correlation between digital HL and age, which previous studies observed in both the general population and cancer survivors [20,37,38]. One possible explanation for why cancer survivors still show similar or even higher levels of digital and navigational HL is that they need to interact with the health care system more frequently than the general population, and in recent years, these interactions have increasingly taken place through digital channels. In line with this assumption, cancer survivors reported a significantly higher use of digital health resources than the general population. Although no significant association between digital resource use and digital HL was found, the combination of frequent engagement with health care services and digital tools may nonetheless contribute to the development of certain digital and navigational HL skills, irrespective of age. In addition, a high proportion of participants reported using complementary medical advice as well as specialist or advanced practice nurse consultations. Previous research suggests that support from health care professionals [39]—such as specialized breast or lung cancer care nurses available to patients with cancer in Switzerland [40-42]—as well as informational and emotional support from health communities [43] may enhance digital HL of cancer survivors by helping them find, understand, appraise, and apply relevant digital health information. Despite these potential facilitating factors, the large proportion of cancer survivors with low digital and navigational HL observed in this study indicates that access to such services may be limited in Switzerland or that existing services may still be insufficiently tailored to cancer survivors' digital HL needs. This interpretation is consistent with previous research showing that difficulties in accessing coordinated, personalized services and care are common rather than exceptional [44].

Beyond the overall HLS_{19} mean scores, an item-level analysis of individual questionnaire items further highlights critical challenges and the importance of support services and improving digital and navigational HL. Cancer survivors reported most difficulties in judging whether the digital information is reliable or offered with commercial interests, and more than half of respondents also indicated difficulties in applying digital information to solve a health problem. These challenges may partly reflect the complexity and individual nature of cancer, as well as the continuous opportunities for new treatment options provided by clinical research. However, similar trends were also observed in the Swiss population study, with 66% of Swiss residents reporting difficulties in assessing the reliability of health information, 63% struggling with determining whether commercial interests were influencing the information, and 53% finding it challenging to apply the information to address a health-related problem [20]. These findings highlight that such difficulties appear to affect people in general, regardless of any specific health condition, and that they may stem from the vast amount of health information available online. This so-called "infodemic"—an overabundance of both accurate health information and mis- or disinformation [45,46]—makes it highly challenging to evaluate the quality, that is, the reliability and the credibility of the information. The spread of mis- and disinformation especially occurs on social media. A study conducted during the COVID-19 pandemic reported that individuals residing in Switzerland did not regard social media as a trustworthy source of health information [47]. This perception may partly explain why most participants in the present study reported using social media for health-related information fewer than once a week.

In addition to improving digital and navigational HL, self-management promotion is recognized as an important strategy to support cancer survivors in dealing with their disease in their everyday life and maintaining and enhancing their health and well-being [10,11]. In this study, most participants reported good self-management skills, consistent with findings from a previous Swiss study of breast cancer survivors [48]. In contrast, the Health Literacy Survey 2019-2021–Switzerland ($HLS_{19\text{-}21}$-CH) study [20] found that 66% (613/928) of participants with one or more chronic diseases had low self-management skills. These differences may partly be explained by the instruments used. While the $HLS_{19\text{-}21}$-CH study assessed self-management skills with only 4 questions and with a focus on health care–related topics, the heiQ covers topics going beyond, capturing a broader range of self-management dimensions. The overall high level of self-management skills observed in this sample may be explained by the participants' relatively high level of social support and low levels of financial deprivation. These factors were identified as significant predictors of self-management skills in this analysis. Another explanation may be the fact that self-management focuses on specific behaviors, and informal and practice-based self-management support, such as counseling by health care professionals, nursing consultations, and support groups, addressing the dimensions assessed in our study, may already be relatively well integrated into comprehensive cancer care. However, while various self-management interventions and programs exist internationally, which are expected to foster self-management skills in cancer survivors [49], structured and theory-based self-management interventions and programs have so far been only partially integrated into routine clinical care in Switzerland, with little data on their effectiveness. Available

XSL•FO
RenderX

interventions include a peer-led self-management program for breast cancer survivors [48], nurse-led basic symptom self-management support for patients receiving first-line systemic outpatient anticancer treatment [50], and a psycho-educational cancer pain self-management intervention [51]. Although these interventions were generally perceived as beneficial, their broader implementation remains limited due to challenges related to delivery, feasibility, and individual preferences regarding timing and format [48,50].

In contrast, measures that specifically promote digital and navigational HL are still rarely used in practice [52]. This imbalance may explain why self-management skills appear relatively strong in this sample, while digital and navigational HL levels remain low. Nevertheless, the high heiQ scores, particularly on the "health service navigation" scale, appear somewhat surprising given the low scores in digital and navigational HL. This discrepancy can be explained by differences in constructs. Although self-management is recognized as an essential aspect of general HL [9], the instruments used to assess self-management skills and navigational HL are based on different constructs. The "health service navigation" items from the heiQ focus more broadly on relationships and communication with health care professionals, while HLS_{19}-NAV items revolve around specific navigational tasks within the entire health system. The lowest self-management skills were observed in relation to emotional distress, which aligns with previous research indicating that emotional needs are among the most frequently unmet needs in patients with breast cancer [53]. This finding underscores the importance of cancer care services addressing not only the physical but also the emotional challenges faced by cancer survivors.

Overall, the findings suggest that efforts to strengthen digital and navigational HL among cancer survivors in Switzerland are needed and that particular attention should be paid to those who experience financial deprivation, have a migration background, lower educational levels, and limited social support. These findings align with previous research that has shown that economic status and HL contribute to inequalities in cancer care experiences in Switzerland [54], suggesting that this may also be true for digital and navigational HL. Moreover, findings are supported by recently published consensus recommendations for comprehensive survivorship care, emphasizing patient empowerment, including the promotion of HL and self-management, as well as psychosocial and socioeconomic support and the involvement of interprofessional experts such as social workers. As a core element for the integration of different types of support, a survivorship plan is recommended [55]. In addition, previous research indicates that interventions, programs, and strategies aimed at strengthening digital and navigational HL of cancer survivors should not only focus on individual competencies. Instead, a more holistic approach is needed, including professional and organizational HL [56]. In addition to individual factors, addressing social, situational, and systemic factors can not only strengthen HL of chronically ill individuals but also reduce burdens and demands placed on them [20,23].

Limitations

This study has several limitations. First, the target sample size could not be reached. With the smaller than expected sample size, the proportion of cancer survivors with low digital and navigational HL may not be adequately reflected. Differences in methodology may limit the comparability of our sample with the Swiss population. Although the modes of administration were quite similar (primarily online), the recruitment approach differed substantially. In the $HLS_{19\text{-}21}$-CH study, participants were recruited via the Swiss Federal Statistical Office's sampling frame and contacted at their postal address. This sampling method could not be used for the present study, as no information about specific chronic conditions was available. Although opting for a broad recruitment approach, it allowed little control over who would be reached by these strategies. Following the outreach to partner organizations and the launch of social media advertisements, the research team noted an increase in survey engagement. Nevertheless, cancer survivors without ties to these institutions or lacking social media access may have missed the chance to participate. Though time-efficient for professionals, this impersonal recruitment strategy may have limited the number of participants compared to more personalized methods involving direct health care provider contact. Second, sampling bias cannot be ruled out, as the participants were predominantly female and breast cancer survivors. The latter, however, reflects the prevalence of cancer in Switzerland, where breast cancer is the most common cancer diagnosis among women [3] and the most often studied cancer type in the field of HL [56]. Furthermore, a considerable number had professional experience in health care, and the survey was conducted predominantly online. Thus, it was completed primarily by people who have sufficient digital skills and/or feel at least comfortable enough to engage with an online survey. In addition, data from the survey software suggest that the length of the survey or the lack of incentives may have demotivated some respondents from completing the survey. Overall, engagement with the landing page of the survey (around 800 clicks) was significantly higher than the number of people who started the survey (n=250), of which more than half ultimately finished it. Cancer survivors might be already burdened by fatigue or other health-related issues affecting their willingness to participate in or finish such a survey. Taken together, these factors limit the representativeness of the study population for Swiss cancer survivors, limiting the generalizability of the results as well as the comparisons with the HL study in the Swiss population that used a representative sample for the adult residents in Switzerland. Third, the instruments used rely on self-reported ability to perform certain tasks related to digital and navigational HL as well as self-management skills. Although self-report instruments have the advantage that they are easy to complete, often answered more honestly, and well accepted by respondents, they are limited by the fact that participants might misestimate their level of HL and self-management skills [57]. However, objective HL measurements are scarce and untenable in population-based studies [58]. Additionally, all instruments applied in this study were standardized and validated.

XSL•FO
RenderX

Conclusions

Cancer survivors in Switzerland generally reported low levels of digital and navigational HL, comparable to those of the general Swiss population and subgroups with chronic conditions. These findings emphasize the need to support cancer survivors in Switzerland in dealing with digital health information and navigating the health care system. Particular attention should be given to tasks perceived as the most difficult, including judging the reliability and relevance of health-related information and services, using this information appropriately, and gaining knowledge and understanding of health care reforms and patient rights. Interventions should be better integrated in existing support, targeting digital and navigational HL as well as self-management skills, while considering individual, social, and situational factors, which showed significant associations with these competencies in this study. Given the limitations of this study, future research should examine digital and navigational HL in a larger sample that more accurately reflects the gender distribution and digital affinity of the population. Nevertheless, this study provides valuable insights into digital and navigational HL among cancer survivors in Switzerland. The findings presented here will be complemented with qualitative interview data from cancer survivors and insights from workshops involving cancer survivors and cancer care professionals. These contributions will be used to develop more explicit intervention recommendations.

Acknowledgments

We would like to thank all the participants who completed our survey. We would also like to thank the Swiss Cancer League, the cantonal and regional cancer leagues, and all the other institutions that supported us with recruitment. Data on digital and navigational health literacy from the general Swiss population was provided by the Swiss Federal Office of Public Health.

Language editing support was provided by ChatGPT-4 and ChatGPT-5 (OpenAI). The tool was used exclusively to refine grammar and phrasing; all scientific content, data interpretation, and conclusions remain the authors' original work, and every artificial intelligence–suggested change was critically reviewed and approved by all coauthors.

Funding

Funding was provided by the Swiss Cancer League (Project Nr. KLS-5941-08-2023).

Data Availability

The datasets used and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: KR, JH, SMDG
Data curation: TW, EA
Formal analysis: JS
Writing – original draft: JS (lead), KR, EA
Writing – review & editing: JH, SMDG, TW

Conflicts of Interest

None declared.

Multimedia Appendix 1
Level of difficulty in interacting with digital resources Health Literacy Survey 2019-2021–Digital Interaction (HLS19-DIGI-INT).
[DOCX File, 40 KB - jmir_v28i1e84228_app1.docx]

Multimedia Appendix 2
Frequency of use of digital health devices and resources Health Literacy Survey 2019-2021–Digital Device (HLS19-DIGI-DD).
[DOCX File, 36 KB - jmir_v28i1e84228_app2.docx]

Multimedia Appendix 3
Differences in frequency of digital device use between cancer survivors and the general Swiss population (Mann-Whitney *U* Test).
[DOCX File, 17 KB - jmir_v28i1e84228_app3.docx]

Multimedia Appendix 4
Regression analyses with all significant sociodemographic variables as independent and Health Education Impact Questionnaire (heiQ) subscales as dependent variables.

XSL•FO
RenderX

[DOCX File, 25 KB - jmir_v28i1e84228_app4.docx]

References

1. Miller KD, Nogueira L, Devasia T, et al. Cancer treatment and survivorship statistics, 2022. CA Cancer J Clin 2022 Sep;72(5):409-436. [doi: 10.3322/caac.21731] [Medline: 35736631]
2. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA A Cancer J Clinicians 2022 Jan;72(1):7-33. [doi: 10.3322/caac.21708]
3. Cirillo P, Feller A, Hošek M, et al. Schweizerischer Krebsbericht 2021: Stand Und Entwicklungen: Federal Statistical Office; 2021.
4. Schweiz K. Krebs in der Schweiz: wichtige Zahlen. Krebsliga Schweiz Ligue Suisse Contre le Cancer Liga Svizzera Contro il Cancro. 2024. URL: https://www.krebsliga.ch/ueber-krebs/zahlen-fakten/-dl-/fileadmin/downloads/sheets/zahlen-krebs-in-der-schweiz.pdf [accessed 2025-07-15]
5. Miller K, Merry B, Miller J. Seasons of survivorship revisited. Cancer J 2008;14(6):369-374. [doi: 10.1097/PPO.0b013e31818edf60] [Medline: 19060601]
6. Vaz-Luis I, Masiero M, Cavaletti G, et al. ESMO expert consensus statements on cancer survivorship: promoting high-quality survivorship care and research in Europe. Ann Oncol 2022 Nov;33(11):1119-1133. [doi: 10.1016/j.annonc.2022.07.1941] [Medline: 35963481]
7. Holden CE, Wheelwright S, Harle A, Wagland R. The role of health literacy in cancer care: a mixed studies systematic review. PLoS ONE 2021;16(11):e0259815. [doi: 10.1371/journal.pone.0259815] [Medline: 34767562]
8. Chan RJ, Mayer DK, Koczwara B, et al. Building capacity in cancer nurses to deliver self-management support: a call for action paper. Cancer Nurs 2020;43(5):341-342. [doi: 10.1097/NCC.0000000000000855] [Medline: 32810041]
9. Mehnert-Theuerkauf A, Esser P. Survivorship-care-programme für Krebspatienten: die Bedeutung von Risikostratifizierung, Selbstmanagement- und Gesundheitskompetenzen im Zeitalter digitaler Versorgung. Bundesgesundheitsbl 2022 Apr;65(4):412-419. [doi: 10.1007/s00103-022-03514-1]
10. Howell D, Mayer DK, Fielding R, et al. Management of cancer and health after the clinic visit: a call to action for self-management in cancer care. J Natl Cancer Inst 2021 May 4;113(5):523-531. [doi: 10.1093/jnci/djaa083] [Medline: 32525530]
11. McCorkle R, Ercolano E, Lazenby M, et al. Self-management: enabling and empowering patients living with cancer as a chronic illness. CA Cancer J Clin 2011;61(1):50-62. [doi: 10.3322/caac.20093] [Medline: 21205833]
12. Sørensen K, Van den Broucke S, Fullam J, et al. Health literacy and public health: a systematic review and integration of definitions and models. BMC Public Health 2012 Jan 25;12(80):80. [doi: 10.1186/1471-2458-12-80] [Medline: 22276600]
13. Shaffer KM, Turner KL, Siwik C, et al. Digital health and telehealth in cancer care: a scoping review of reviews. Lancet Digit Health 2023 May;5(5):e316-e327. [doi: 10.1016/S2589-7500(23)00049-3] [Medline: 37100545]
14. Ahmed S, LePage K, Benc R, et al. Lessons learned from the implementation of a person-centred digital health platform in cancer care. Curr Oncol 2022 Sep 29;29(10):7171-7180. [doi: 10.3390/curroncol29100564] [Medline: 36290841]
15. Chan RJ, Crichton M, Crawford-Williams F, et al. The efficacy, challenges, and facilitators of telemedicine in post-treatment cancer survivorship care: an overview of systematic reviews. Ann Oncol 2021 Dec;32(12):1552-1570. [doi: 10.1016/j.annonc.2021.09.001] [Medline: 34509615]
16. Moradian S, Maguire R, Liu G, et al. Promoting self-management and patient activation through eHealth: protocol for a systematic literature review and meta-analysis. JMIR Res Protoc 2023 Mar 2;12:e38758. [doi: 10.2196/38758] [Medline: 36862481]
17. van den Hurk CJG, Mols F, Eicher M, et al. A narrative review on the collection and use of electronic patient-reported outcomes in cancer survivorship care with emphasis on symptom monitoring. Curr Oncol 2022 Jun 17;29(6):4370-4385. [doi: 10.3390/curroncol29060349] [Medline: 35735458]
18. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9. [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
19. Levin-Zamir D, Van den Broucke S, Bíró É, et al. Measuring digital health literacy and its associations with determinants and health outcomes in 13 countries. Front Public Health 2025;13:1472706. [doi: 10.3389/fpubh.2025.1472706] [Medline: 40182520]
20. Jaks R, De Gani SM, Bieri U, Kocher JP. Health Literacy Survey Schweiz 2019-2021: im Auftrag des Bundesamtes für Gesundheit. Schlussbericht mit überarbeiteten Empfehlungen. : Careum Zentrum für Gesundheitskompetenz; 2023 URL: https://www.bag.admin.ch/dam/de/sd-web/zTvqdtp8UAbm/schlussbericht-health-literacy-survey-careum.pdf [accessed 2026-05-14]
21. Kaltenborn Z, Paul K, Kirsch JD, et al. Super fragmented: a nationally representative cross-sectional study exploring the fragmentation of inpatient care among super-utilizers. BMC Health Serv Res 2021 Apr 14;21(1):338. [doi: 10.1186/s12913-021-06323-5] [Medline: 33853590]
22. Lomazzi M, Borisch B. Game changer in cancer treatment in Switzerland. J Cancer Policy 2022 Sep;33:100343. [doi: 10.1016/j.jcpo.2022.100343] [Medline: 35724957]

XSL•FO
RenderX

23. Griese L, Schaeffer D, Berens EM. Navigational health literacy among people with chronic illness. Chronic Illn 2023 Mar;19(1):172-183. [doi: 10.1177/17423953211073368] [Medline: 35014911]
24. The HLS19 Consortium of the WHO Action Network M-POHL. International report on the methodology, results, and recommendations of the European health literacy population survey 2019-2021 (HLS19) of M-POHL. : Austrian National Public Health Institute; 2021 URL: https://m-pohl.net/Int_Report_methdology_results_recommendations [accessed 2026-05-14]
25. Kocalevent RD, Berg L, Beutel ME, et al. Social support in the general population: standardization of the Oslo social support scale (OSSS-3). BMC Psychol 2018 Jul 17;6(1):31. [doi: 10.1186/s40359-018-0249-9] [Medline: 30016997]
26. Griese L, Finbråten HS, Francisco R, et al. HLS19-NAV-validation of a new instrument measuring navigational health literacy in eight European countries. Int J Environ Res Public Health 2022 Oct 25;19(21):13863. [doi: 10.3390/ijerph192113863] [Medline: 36360755]
27. The HLS19 Consortium of the WHO Action Network M-POHL. The HLS19–NAV Instrument to Measure Navigational Health Literacy: Austrian National Public Health Institute; 2022. URL: https://m-pohl.net/sites/m-pohl.net/files/inline-files/Factsheet%20HLS19-NAV.pdf [accessed 2026-05-14]
28. The HLS19 Consortium of the WHO Action Network M-POHL. The HLS19–DIGI Instrument to Measure Digital Health Literacy: Austrian National Public Health Institute; 2022. URL: https://m-pohl.net/sites/m-pohl.net/files/inline-files/Factsheet%20HLS19-DIGI.pdf [accessed 2026-05-14]
29. Osborne RH, Elsworth GR, Whitfield K. The health education impact questionnaire (heiQ): an outcomes and evaluation measure for patient education and self-management interventions for people with chronic conditions. Patient Educ Couns 2007 May;66(2):192-201. [doi: 10.1016/j.pec.2006.12.002] [Medline: 17320338]
30. Maunsell E, Lauzier S, Brunet J, Pelletier S, Osborne RH, Campbell HS. Health-related empowerment in cancer: validity of scales from the health education impact questionnaire. Cancer 2014 Oct 15;120(20):3228-3236. [doi: 10.1002/cncr.28847] [Medline: 24988944]
31. Bélanger A, Hudon C, Fortin M, Amirall J, Bouhali T, Chouinard MC. Validation of a French-language version of the health education impact Questionnaire (heiQ) among chronic disease patients seen in primary care: a cross-sectional study. Health Qual Life Outcomes 2015 May 24;13(1):64. [doi: 10.1186/s12955-015-0254-0] [Medline: 26306793]
32. Schuler M, Musekamp G, Faller H, et al. Assessment of proximal outcomes of self-management programs: translation and psychometric evaluation of a German version of the health education impact Questionnaire (heiQ). Qual Life Res 2013 Aug;22(6):1391-1403. [doi: 10.1007/s11136-012-0268-6] [Medline: 22987145]
33. Charan J, Biswas T. How to calculate sample size for different study designs in medical research? Indian J Psychol Med 2013 Apr;35(2):121-126. [doi: 10.4103/0253-7176.116232] [Medline: 24049221]
34. Melhem SJ, Nabhani-Gebara S, Kayyali R. Digital trends, digital literacy, and e-Health engagement predictors of breast and colorectal cancer survivors: a population-based cross-sectional survey. Int J Environ Res Public Health 2023 Jan 13;20(2):1472. [doi: 10.3390/ijerph20021472] [Medline: 36674237]
35. Milne RA, Puts MTE, Papadakos J, et al. Predictors of high eHealth literacy in primary lung cancer survivors. J Cancer Educ 2015 Dec;30(4):685-692. [doi: 10.1007/s13187-014-0744-5] [Medline: 25355524]
36. Heiman H, Keinki C, Huebner J, Working Group Prevention and Integrative Oncology of the German Cancer Society. EHealth literacy in patients with cancer and their usage of web-based information. J Cancer Res Clin Oncol 2018 Sep;144(9):1843-1850. [doi: 10.1007/s00432-018-2703-8] [Medline: 30006768]
37. Schaeffer D, Gille S, Berens EM, et al. Digitale gesundheitskompetenz der bevölkerung in Deutschland: ergebnisse des HLS-GER 2. Gesundheitswesen 2023 Apr;85(4):323-331. [doi: 10.1055/a-1670-7636]
38. Hoogland AI, Mansfield J, Lafranchise EA, Bulls HW, Johnstone PA, Jim HSL. eHealth literacy in older adults with cancer. J Geriatr Oncol 2020 Jul;11(6):1020-1022. [doi: 10.1016/j.jgo.2019.12.015] [Medline: 31917114]
39. Hildreth K, Henderson N, Bourne G, et al. Recommendations for improving digital health literacy in patients with cancer receiving care in the deep south. JCO Oncology Practice 2023 Nov;19(11_suppl):202-202. [doi: 10.1200/OP.2023.19.11_suppl.202]
40. Fachgruppe breast care nurse (BCN-OPS). Onkologiepflege Schweiz. 2025. URL: https://www.onkologiepflege.ch/verband/fachgruppen/breast-care-nurse-bcn-ops [accessed 2025-06-18]
41. Eicher M, Kadmon I, Claassen S, et al. Training breast care nurses throughout Europe: the EONS postbasic curriculum for breast cancer nursing. Eur J Cancer 2012 Jun;48(9):1257-1262. [doi: 10.1016/j.ejca.2011.07.011] [Medline: 21865027]
42. Serena A, Dwyer AA, Peters S, Eicher M. Acceptance of the advanced practice nurse in lung cancer role by healthcare professionals and patients: a qualitative exploration. J Nurs Scholarsh 2018 Sep;50(5):540-548. [doi: 10.1111/jnu.12411] [Medline: 30033614]
43. Zhou J, Wang C. Improving cancer survivors' e-health literacy via online health communities (OHCs): a social support perspective. J Cancer Surviv 2020 Apr;14(2):244-252. [doi: 10.1007/s11764-019-00833-2] [Medline: 31820215]
44. Sperisen N, Arditi C, Schaffar R, Dietrich PY, Rapiti E. Experiences and needs of Swiss cancer survivors in the domains of health-related information and the healthcare system. Cancers (Basel) 2024 Dec 15;16(24):4177. [doi: 10.3390/cancers16244177] [Medline: 39766075]

XSL•FO
RenderX

45. Rubinelli S, Purnat TD, Wilhelm E, et al. WHO competency framework for health authorities and institutions to manage infodemics: its development and features. Hum Resour Health 2022 May 7;20(1):35. [doi: 10.1186/s12960-022-00733-0] [Medline: 35525924]
46. Zarocostas J. How to fight an infodemic. The Lancet 2020 Feb;395(10225):676. [doi: 10.1016/S0140-6736(20)30461-X]
47. De Gani SM, Berger FMP, Guggiari E, Jaks R. Relation of corona-specific health literacy to use of and trust in information sources during the COVID-19 pandemic. BMC Public Health 2022 Jan 6;22(1):42. [doi: 10.1186/s12889-021-12271-w] [Medline: 34991525]
48. Ribi K, Schmidt F, Haslbeck J, et al. Pilot testing a peer-led self-management program for women after breast cancer. Semin Oncol Nurs 2025 Apr;41(2):151814. [doi: 10.1016/j.soncn.2025.151814] [Medline: 39915174]
49. Cuthbert CA, Farragher JF, Hemmelgarn BR, Ding Q, McKinnon GP, Cheung WY. Self-management interventions for cancer survivors: a systematic review and evaluation of intervention content and theories. Psychooncology 2019 Nov;28(11):2119-2140. [doi: 10.1002/pon.5215] [Medline: 31475766]
50. Bana M, Ribi K, Kropf-Staub S, et al. Implementation of the Symptom Navi © Programme for cancer patients in the Swiss outpatient setting: a study protocol for a cluster randomised pilot study (Symptom Navi© Pilot Study). BMJ Open 2019 Jul 9;9(7):e027942. [doi: 10.1136/bmjopen-2018-027942] [Medline: 31289075]
51. Valenta S, Miaskowski C, Spirig R, et al. Randomized clinical trial to evaluate a cancer pain self-management intervention for outpatients. Asia Pac J Oncol Nurs 2022 Jan;9(1):39-47. [doi: 10.1016/j.apjon.2021.12.003] [Medline: 35528799]
52. De Gani SM, Berger FMP. Bedeutung und Stärkung von digitaler Gesundheitskompetenz in der Schweiz. In: Langkafel P, Matusiewicz D, editors. Digitale Gesundheitskompetenz: Brauchen Wir Den Digitalen Führerschein [Book in German]: Medhochzwei; 2021:43-54.
53. Capelan M, Battisti NML, McLoughlin A, et al. The prevalence of unmet needs in 625 women living beyond a diagnosis of early breast cancer. Br J Cancer 2017 Oct 10;117(8):1113-1120. [doi: 10.1038/bjc.2017.283] [Medline: 28859057]
54. Jolidon V, Eicher M, Peytremann-Bridevaux I, Arditi C. Inequalities in patients' experiences with cancer care: the role of economic and health literacy determinants. BMC Health Serv Res 2024 Jun 14;24(1):733. [doi: 10.1186/s12913-024-11174-x] [Medline: 38877526]
55. Consensus recommendations for cancer survivorship care in switzerland. Aerzteverlag medinfo AG. 2025. URL: https://www.medinfo-verlag.ch/infooncosuisse/konsensusempfehlungen-fuer-cancer-survivorship-care-in-der-schweiz [accessed 2026-05-04]
56. Jeitani C, Van den Broucke S, Leemans C. Interventions addressing health literacy in cancer care: a systematic review of reviews. Int J Environ Res Public Health 2025 Feb 2;22(2):212. [doi: 10.3390/ijerph22020212] [Medline: 40003438]
57. Schulz PJ, Lindahl B, Hartung U, Naslund U, Norberg M, Nordin S. The right pick: does a self-assessment measurement tool correctly identify health care consumers with inadequate health literacy? Patient Educ Couns 2022 Apr;105(4):926-932. [doi: 10.1016/j.pec.2021.07.045] [Medline: 34366227]
58. Pleasant A, Maish C, O'Leary C, Carmona R. Measuring health literacy in adults: an overview and discussion of current tools. In: Okan O, Bauer U, Levin-Zamir D, Pinheiro P, Sørensen K, editors. International Handbook of Health Literacy: Policy Press; 2019:67-82. [doi: 10.51952/9781447344520.ch005]

Abbreviations

heiQ: Health Education Impact Questionnaire
HL: health literacy
HLS_{19}: Health Literacy Survey 2019‐2021
$HLS_{19\text{-}21}$-CH: Health Literacy Survey 2019-2021–Switzerland
HLS_{19}-DIGI: Health Literacy Survey 2019-2021–Digital Health Literacy
HLS_{19}-DIGI-DD: Health Literacy Survey 2019-2021–Digital Device
HLS_{19}-DIGI-HI: Health Literacy Survey 2019-2021–Digital Health Information
HLS_{19}-DIGI-INT: Health Literacy Survey 2019-2021–Digital Interaction
HLS_{19}-NAV: Health Literacy Survey 2019-2021–Navigational Health Literacy
ISCED: International Standard Classification of Education
M-POHL: Measuring Population and Organizational Health Literacy
OSSS-3: Oslo Social Support Scale
SGP: Swiss general population
WHO: World Health Organization

XSL•FO
RenderX

Edited by A Schwartz, M Balcarras; submitted 16.Sep.2025; peer-reviewed by N Diviani, S Greene; revised version received 22.Jan.2026; accepted 03.Feb.2026; published 25.May.2026.

Please cite as:
Seinsche J, Alder E, Wehrmann T, Haslbeck J, Ribi K, De Gani SM
Digital and Navigational Health Literacy in Swiss Cancer Survivors Compared With the General Population: Cross-Sectional Questionnaire Study
J Med Internet Res 2026;28:e84228
URL: https://www.jmir.org/2026/1/e84228
doi:10.2196/84228

© Julia Seinsche, Elena Alder, Tami Wehrmann, Jörg Haslbeck, Karin Ribi, Saskia Maria De Gani. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Validation of the Updated Digital Health Literacy Instrument and Development of a Short Form: Online Survey Study of the General Population

Rosalie van der Vaart[1], PhD; Inge Tuitert[2], PhD; Job van 't Veer[2], PhD; Constance Drossaert[3], PhD

[1]Centre of Expertise Health Innovation, Technology for Health Care Research Group, The Hague University of Applied Sciences, Johanna Westerdijkplein 75, Den Haag, The Netherlands

[2]Department of Health and Social Studies, NHL Stenden University of Applied Sciences, Leeuwarden, The Netherlands

[3]Department of Psychology, Health and Technology, Faculty of Behavioural, Management and Social Sciences, University of Twente, Enschede, The Netherlands

Corresponding Author:
Rosalie van der Vaart, PhD
Centre of Expertise Health Innovation, Technology for Health Care Research Group, The Hague University of Applied Sciences, Johanna Westerdijkplein 75, Den Haag, The Netherlands

Abstract

Background: The digital health literacy instrument (DHLI) was developed in 2017 to measure individuals' ability to access, understand, evaluate, and apply online health information. Since that time, digital health has shifted from desktop-based internet use to mobile devices, and there has been a rapidly expanding range of health apps. Additionally, heightened privacy and data security requirements have increased the complexity of user competencies needed to engage with digital health tools. These developments underscore the need to update the original DHLI.

Objective: This study aimed to create an updated version of the DHLI (DHLI 2.0) that reflects current digital health practices and to examine its reliability and validity by exploring associations with user characteristics. Additionally, we aimed to develop a short-form version to facilitate broader use in research and practice.

Methods: The instrument was iteratively updated and pilot-tested to retain the original theoretical framework while reflecting current digital health practices, devices, and emerging challenges such as mobile use and data security. Several items were reworded and a new 2-item subscale on digital safety was added. The full DHLI 2.0 comprises 24 items across 8 skill domains. A 16-item short form was developed by iteratively removing 1 or 2 items per subscale based on the "α if item deleted" criterion, while retaining the same subscale structure as the full form. Data to validate the new version of the instrument were collected in June 2024 through an online survey among members of a representative citizen panel in Friesland, a province in the Netherlands (N=2728). Sociodemographics, internet and health-related internet use, general health literacy (measured with the Single Item Literacy Screener), self-reported health, and health care use were assessed. Internal consistency was evaluated using Cronbach α, and construct validity was assessed via Spearman ρ correlations with related constructs.

Results: Internal consistency was high for both the full (α=0.94) and short-form (α=0.90) scales. Most subscales showed satisfactory to excellent reliability (α=0.71–0.93), while "Securing privacy" and "Using security measures" demonstrated moderate reliability (α=0.65-0.66). The DHLI 2.0 total scores were approximately normally distributed (skewness –0.5; kurtosis 0.4). As expected, digital health literacy was negatively correlated with age (ρ=−0.39, *P*<.001) and positively correlated with education (ρ=0.22, *P*<.001), income (ρ=0.27, *P*<.001), time spent online (ρ=0.32, *P*<.001), and general health literacy (ρ=−0.42, *P*<.001).

Conclusions: The DHLI 2.0 provides an updated, reliable, and valid measure of digital health literacy covering 8 key domains, including data security. The 16-item short form offers a concise alternative suitable for research and possibly practical applications in health and eHealth contexts.

(J Med Internet Res 2026;28:e86879) doi:10.2196/86879

KEYWORDS

digital health literacy skills; eHealth; measurement; validation; instrument; health information

Introduction

Over the past decades, the landscape of health care has transformed considerably due to the rise of the internet and digital health technology. Citizens and patients now have access to a broad spectrum of digital health resources, allowing them to search online for health-related information, communicate with care professionals through patient portals, connect with

XSL•FO
RenderX

peers via social media platforms, and apply digital tools and applications to monitor or manage their health.

Although these digital health innovations hold great potential to improve accessibility and quality of care and to improve patient involvement in care, not everyone benefits equally from them. For example, data from the Dutch National Monitor Digital Care show stark differences in use of digital solutions among care users, where almost all care users with a theoretical educational background report using a computer independently or with help, compared to about half of those with a practice-oriented educational background [1,2]. These disparities also extend to the use of digital health tools, where these numbers are 68% and 22%, respectively, for using websites to seek health information. In addition, age-related differences are also evident: while 75% of those aged 15 to 39 years reported having used health websites, this is only 36% in the age group of ≥65 years [1,2]. Similar trends have frequently been observed in international research contexts [3-5]. Effectively engaging with digital health resources requires a set of complex skills that are not evenly distributed across the population. People may struggle to access, navigate, understand, evaluate, or apply online health information, or to use digital health tools efficiently.

These findings highlight the importance of digital health literacy, defined as the ability to seek, find, understand, appraise, and apply online health information and services [6]. As health information and care services are increasingly delivered through digital channels, gaining insight into individuals' digital health literacy is essential. It allows researchers to assess the level of digital health literacy at a population level and provides insight into differences in digital health literacy skills among subgroups. In addition, it supports researchers and developers of health applications to assess whether innovations fit their target group. For health care professionals, knowledge about their patients' digital health literacy could even be useful to assess whether support is needed, for whom, and in which form. To assess digital health literacy, the digital health literacy instrument (DHLI) was developed in 2017 [7]. The DHLI is grounded in theory and empirical observational research and encompasses 7 subscales. Each subscale captures a distinct component of digital health literacy, and together they represent the practical skills needed to access, understand, evaluate, and apply digital health information: operational skills (the skills to operate a device, such as typing, clicking, and swiping), navigation skills, information searching, evaluating reliability, determining relevance, creating self-generated content (being able to express oneself clearly in written language), and protecting privacy while sharing information with others.

Since its introduction, the DHLI has been widely adopted and validated across national and international contexts, including studies among diverse patient populations [8,9], adolescents [10,11], older adults [12], university students [13-15], and health care professionals [16], demonstrating its value as a comprehensive instrument for assessing digital health competencies. Yet, the digital health landscape has evolved considerably since the publication of the DHLI. First, there has been a shift from desktop computers and laptops to mobile devices, such as smartphones and tablets, which require different operational skills (eg, tapping and swiping rather than clicking or typing). Second, the field has seen a rapid increase in digital health tools (eg, apps for self-management, self-monitoring, and communication), making some of the original DHLI items, which primarily focus on internet use, less suitable. Third, growing concerns around privacy and data safety have led to more complex authentication procedures, such as 2-factor identification and the need to manage secure passwords, which demand additional user competencies [17].

In response to these changes, we felt the need for a revision of the DHLI. Therefore, the first aim of this study was to validate an updated version of the DHLI, the DHLI 2.0, including its internal consistency and construct validity, by examining the associations with a range of user characteristics. Additionally, we aimed to examine the feasibility and validity of a short-form version of the instrument and to facilitate broader applications in research and practice.

Methods

Instrument Revision

The revised version of the DHLI, the DHLI 2.0, includes 8 skill domains, each measured with 2 to 4 items, resulting in 24 items in total. The goal of the revision was to keep the original structure and theoretical framework of the DHLI intact, while updating the instrument to current digital health practices and currently used devices. The revision was an iterative process, with some experts, in which we reworded the items to make them more in line with the vocabulary of the current digital landscape. Subsequently, a pilot study was conducted with a preliminary version. The prefinal version of the new instrument was tested in a qualitative pilot study, with 8 survey experts of Planbureau Fryslân, which informed the final adjustments. As main changes, we added the use of tablets and smartphones in the introduction of the items and included words as tapping and swiping (in addition to clicking). A new subscale, including 2 items, was added to address data security challenges based on problems encountered in recent projects. The items include skills to use 2-step authentication and skills in using strong passwords.

Similar to the original version, all items are scored on a 4-point Likert scale, with response options ranging from "very easy" to "very difficult" or from "never" to "often." For the current validation study, we added an option "not applicable/do not know," since this was common practice in the surveys of Planbureau Fryslân. For analysis purposes, scores are reverse-coded so that higher values reflect higher levels of digital health literacy. A total digital health literacy score was computed by averaging responses across all items, provided that at least 16 items have been completed in the full scale and 8 items in the short form (SF). Subscale scores were calculated using the mean of the items within each skill domain, provided that all items of the subscale were completed. Multimedia Appendices 1-3 provide (1) the items and scoring instructions for the revised DHLI, full form (FF DHLI 2.0); (2) the SF DHLI 2.0; and (3) an overview of the changes made in DHLI 2.0 compared to the original DHLI.

XSL•FO
RenderX

Study Design and Participants

For this study, the citizen panel from Planbureau Fryslân was used. Planbureau Fryslân is the independent regional institute for research and monitoring, aiming to inform and enhance local government policy in the province of Friesland, the Netherlands. Their panel consists of a sample of approximately 8000 residents in Friesland (a Northern province of the Netherlands). To ensure representativeness, Planbureau Fryslân regularly draws a random sample of all residents aged 18 years and older in the province, in close coordination with the municipalities. Potential panel members are personally invited by letter to participate in the sample. Self-enrollment in the panel is not possible. Panel members are periodically asked to participate in surveys on topics relevant to regional policy, including health. The data collection for Panel Fryslân follows a standardized procedure as described in their research accountability document [18]. For this study, in June 2024, all active panel members were invited to complete a survey on digital health care use and digital health literacy. The questionnaire included the DHLI 2.0 and 5 other digital health-related questions (see Measures). Panel members could respond to the survey during a 3-week period and received 2 reminders within these 3 weeks. Three months earlier, in January 2024, the panel had filled out a survey on general health, including questions on self-reported general health, health care use, and chronic health conditions (see *Measures*). These variables were merged with the data from June 2024 for the current analyses. In total, 2728 respondents filled out both the survey from January and June 2024. The data of these respondents are used in the analyses.

Procedure

Data collection by Planbureau Fryslân was carried out via an online Dutch-language survey (Quest Software, Inc), which was distributed via email to panel members. Panel members were given 1 month to respond to the survey and received 2 reminder emails during that period to enhance the response rate.

Measures

In addition to the DHLI 2.0, the survey items addressed the following characteristics of the participants: (1) sociodemographics (gender, age, educational attainment, and income); (2) internet use (average number of hours spent online per day); and (3) health-related internet use (the use of 11 different online health activities in the past 12 months) (see Table 1 for a complete overview of all activities); with these 11 items, a sum score was calculated (ranging from 0 to 11), where using 1 specific type of online health activities added 1 point to the sum score; (4) general health literacy (assessed using the Single Item Literacy Screener [SILS]) [19]; (5) self-reported general health on a scale from worst possible health (1) to the best possible health (10); (6) health care use (assessed by number of visits to a general practitioner in the past 12 months, visits to a medical specialist in the past 12 months, and visits to a mental health professional in the past 12 months); and (7) presence of chronic conditions.

XSL•FO
RenderX

Table . Internet use and health-related internet use (N=2728).

Variable	Value
Hours spent online per day, n (%)	
<1	371 (13.6)
1 - 3	1346 (49.3)
3 - 5	551 (20.2)
5 - 8	237 (8.7)
>8	148 (5.4)
Not applicable/don't know	35 (1.3)
Missing	40 (1.5)
Self-rated internet skills, n (%)	
Very good	624 (22.9)
Good	1186 (43.5)
Average	536 (19.6)
Less than average	272 (10)
Poor	50 (1.8)
Not applicable/don't know	20 (0.7)
Missing	40 (1.5)
SILS[a], n (%)	
Never	1692 (62)
Rarely	739 (27.1)
Sometimes	218 (8)
Often	29 (1.1)
Always	3 (0.1)
Not applicable/don't know	7 (0.3)
Missing	40 (1.5)
Number of respondents who have used the internet for health-related reasons	
Online information, n (%)	2353 (86.2)
Accessing a patient portal[b], n (%)	1729 (63.4)
Online prescriptions, n (%)	1220 (44.7)
Accessing a personal health portal[c], n (%)	1175 (43.1)
Wearable, n (%)	1027 (37.6)
Online appointment scheduling, n (%)	855 (31.3)
Information on a health app, n (%)	824 (30.2)
Video consultation, n (%)	428 (15.7)
Information on social media, n (%)	423 (15.5)
Healthy lifestyle app, n (%)	367 (13.5)
Communicating via a patient forum, n (%)	170 (6.2)
Health-related internet use, mean (SD)	3.93 (2.24)

[a]SILS: Single Item Literacy Screener.

[b]Data from a care provider shared via a secured portal.

[c]Health data from care providers combined on a personal secured portal.

XSL•FO
RenderX

Ethical Considerations

All participants provided informed consent upon joining the panel and agreed to the use and sharing of their anonymized data for research purposes. The panel adheres to the code of conduct established by the Dutch Association for Statistics and Research (Vereniging voor Statistiek en Onderzoek, VSO) [20]. This study was reviewed by the Ethical Board of NHL Stenden University of Applied Sciences, who determined that the study is indeed in line with this conduct (reference number Tuitert 202405). Prior to being shared with the authors of this paper, the data were fully anonymized. The authors had no direct contact with the participants, and no financial compensation was provided for participation in the panel.

Data Analyses

Data were analyzed using IBM SPSS version 29.0 for Windows (IBM Corp). Descriptive statistics were computed for all relevant variables, including frequencies, percentages, means, and SDs. Missing values and the option "don't know/don't want to tell" are in most cases reported in the descriptive statistics but were treated as missing values in further analyses.

Cronbach α served as a measure of internal consistency, reflecting the (weighted) average correlation of items within the scale [21]. In general, a Cronbach α of 0.7 to 0.8 is regarded as satisfactory for scales to be used as research tools [22]. The distributional characteristics of the DHLI and its subscales were examined to assess normality and detect floor and ceiling effects. Skewness and kurtosis values were used to evaluate score distributions, with values between +1 and –1 interpreted as indicating no or slight nonnormality [23]. Floor or ceiling effects were considered present when more than 15% of participants achieved either the lowest or highest possible score on a subscale [24]. Evidence for construct validity was determined by studying Spearman ρ correlations between total scores on the DHLI 2.0 and sociodemographics, (health-related) internet use, self-reported general health, health care use, and the SILS [19]. We expected the correlations with the other variables to be in line with the correlations of the original DHLI [7]. Additionally, we explored the reliability and validity of an SF DHLI 2.0. By using the "α if item is deleted" option, for each subscale 1 or 2 items were deleted. These short-form subscales were also taken along in the validity analyses.

Results

Participants

Table 2 shows the characteristics of the sample. Men (1582/2728, 58%) and respondents with a theory-based educational background (1445/2728, 53%) were overrepresented, that is, people with tertiary education. The mean age of the sample was 60.1 (SD 16.8) years, with a range between 19 and 95 years.

Table 3 presents the sample's health-related characteristics. A considerable number of participants (1034/2728, 37.9%) reported having at least 1 chronic condition, and about half (1286/2728, 47.1%) had had contact with a medical specialist in the past 12 months. Still, participants scored their own health as rather good on average (7.5 on a scale from 1 to 10).

XSL•FO
RenderX

Table . Sociodemographics (N=2728).

Characteristic	Value
Gender, n (%)	
Male	1582 (58)
Female	1135 (41.6)
Other	11 (0.4)
Age (y), mean (SD)	60.1 (16.8)
Educational attainment[a], n (%)	
Low	496 (18.2)
Middle (practice-based/vocational)	718 (26.3)
High (theory-based)	1445 (53)
Missing	69 (2.5)
Family income yearly (gross)[b]	
Less than €15,000	63 (2.3)
€15,000– €31,000	308 (11.3)
€31,000–€38,500	282 (10.3)
€38,500–€46,000	363 (13.3)
€46,000–€77,000	643 (23.6)
€77,000–€92,000	235 (8.6)
€92,000 or more	256 (9.4)
Not applicable/don't know	514 (18.8)
Missing	64 (2.3)
Personal income monthly (net)	
€1000 or less – €2000	685 (25.2)
€2000–€3000	826 (30.3)
€3000–€4000	547 (20.1)
€4000–€5000	161 (5.9)
€5000 or higher	136 (4)
Not applicable/don't know	353 (12.9)
Missing	20 (0.7)

[a]Low: less than primary, primary, and lower secondary education (European qualifications framework 1 (EQF1) and part of EQF2); medium: upper secondary and postsecondary nontertiary education (part of EQF2, EQF3, EQF4, EQF4+); high: tertiary education (EQF5, EQF6, EQF7, EQF8) [25].
[b]€1=US $1.15.

Table . Health and health care use.

Variable	Value
Self-reported general health[a] (N=2625), mean (SD)	7.5 (1.3)
Health care use (N=2728), n yes (%)	
General practitioner	1876 (68.8)
Medical specialist	1286 (47.1)
Mental health care	198 (7.3)
Chronic condition (N=2728), n yes (%)	1034 (37.9)

[a]On a scale from 1 to 10.

XSL•FO
RenderX

Table 1 provides an overview of participants' general and health-related internet use. Nearly half of the respondents reported spending approximately 1 to 3 hours online per day. Most participants (1810/2728, 66.4%) rated their own internet skills as (very) good. A large majority had used the internet to search for health-related information (2352/2728, 86.2%), and the use of online portals and digital prescriptions was also relatively high. In contrast, only a small proportion of participants reported using the internet to have a video consultation with their health care provider (428/2728, 15.7%) or to communicate online with fellow patients through a forum (170/2728, 6.2%).

Reliability and Distributional Properties of the Revised DHLI

Table 4 shows the scores and internal consistency of the DHLI 2.0 items and subscales. Both the FF (24 items) and the SF (16 items) showed strong reliability, with Cronbach α coefficients of 0.94 and 0.90, respectively. Only the protecting privacy full subscale and the using security measures subscale had moderate Cronbach α scores (0.65 and 0.66, respectively).

Table . Descriptive statistics of digital health literacy instrument (DHLI) 2.0 scales and subscales.

Scale score	N[a]	Value	Cronbach α
Total scale score, mean (SD)			
FF DHLI 2.0[b] (24 items)	2543	3.17 (0.44)	0.94[c]
SF DHLI 2.0[d] (16 items)	2728	3.11 (0.48)	0.90[e]
Operational skills, mean[f] (SD)			
FF 4 items	2728	3.56 (0.51)	0.86
SF 2 items	2728	3.46 (0.64)	0.83
Navigation skills, mean (SD)			
FF 3 items	2436	3.24 (0.54)	0.77
SF 2 items	2436	3.34 (0.58)	0.71
Information searching, mean (SD)			
FF 3 items	2641	3.15 (0.57)	0.89
SF 2 items	2641	3.08 (0.61)	0.86
Evaluating reliability, mean (SD)			
FF 3 items	2520	2.80 (0.66)	0.84
SF 2 items	2520	2.66 (0.73)	0.83
Determining relevance, mean (SD)			
FF 3 items	2453	2.93 (0.61)	0.89
SF 2 items	2453	2.95 (0.63)	0.86
Generating content, mean (SD)			
FF 3 items	1908	3.07 (0.64)	0.93
SF 2 items	1908	3.11 (0.65)	0.91
Protecting privacy, mean (SD)			
FF 3 items	1387	3.52 (0.50)	0.65
SF 2 items	1387	3.68 (0.51)	0.75
Using security measures, mean (SD)			
FF 2 items	2626	3.06 (0.69)	0.66

[a]Number of respondents per (sub)scale, after excluding missing and "not applicable/do not know" responses.

[b]FF: full form.

[c]n=1055; number of respondents that completed at least 16 items, required to compute a total score for the FF.

[d]SF: short form.

[e]n=1030; number of respondents that completed at least 8 items, required to compute a total scale score for the SF.

[f]Possible range between 1 and 4.

XSL•FO
RenderX

Respondents had total mean scores of 3.17 (SD 0.44) and 3.11 (SD 0.48) for the FF and the SF, respectively. Both the FF (skewness = –0.5; kurtosis =0.4) and the SF (skewness = –0.5; kurtosis =0.50) show low skewness and kurtosis values, indicating that their distributions are approximately normal.

The highest scores on the subscales were reported for operational skills (mean 3.6, SD 0.5 and SF mean 3.5, SD 0.6), navigation skills (mean 3.2, SD 0.5 and SF mean 3.3, SD 0.6), and protecting privacy (mean 3.52, SD 0.50 and SF mean 3.68, SD 0.51). This last subscale is skewed for both the full subscale (−1.5) and the short subscale (−2.1), with a ceiling effect where 32.1% (445/1387) scored 4 in the full subscale and 61.9% (858/1387) in the short version, and showed kurtosis in the full subscale (3.8) and in the SF (5.7). Only in the short subscale, operational skills and navigation skills exhibited a skewness of –1, with a ceiling effect, where for operational skills, 49.3% (1345/2728) scored 4 and for navigation skills 26.9% (655/2436). Participants indicated having more difficulties with evaluating reliability (mean 2.8, SD 0.7 and SF mean 2.7, SD 0.7) and with determining relevance of information (mean 2.9, SD 0.6 and SF mean 3.0, SD 0.6). The scores on the new subscale, using security measures, were relatively high for the question regarding 2-factor authentication (mean 3.28, SD 0.78) and low for the question regarding creating and remembering strong passwords (mean 2.83, SD 0.82).

Construct Validity of the Revised DHLI

Table 5 presents the Spearman ρ correlations between the DHLI 2.0 scale scores, the total scores of both the full and short versions of the DHLI 2.0, and other assessed variables. Looking at sociodemographics, age was moderately negatively correlated with scores on the DHLI 2.0, indicating that older individuals tend to have lower digital health literacy levels. This association was strongest for operational skills and navigation skills. Educational attainment showed a weak positive correlation with the DHLI 2.0, with the highest scores found for generating content, suggesting that educational attainment is somewhat related to digital health literacy skills. Regarding income, family gross income showed a weak positive correlation with the DHLI 2.0, suggesting that a higher income is slightly related to higher digital health literacy skills. Concerning internet use, daily time spent online showed a moderate positive correlation with the DHLI 2.0, suggesting that individuals who spend more time online tend to have higher digital health literacy. Again, the strongest correlations were observed for operational skills. Health-related internet use only weakly correlated with the DHLI 2.0. The DHLI 2.0 showed a moderate positive correlation with SILS, indicating conceptual overlap between the constructs measured by these instruments.

XSL•FO
RenderX

Table . Spearman ρ correlations between the revised digital health literacy instrument, sociodemographics, (health-related) internet use, health-related variables, and the Single Item Literacy Screener (SILS).

(Sub)scale	Age	Education	Family income	Personal income	Hours spent online	Health-related internet use	General health	Use general practitioner	Use medical specialist	Use mental health care	SILS
Full form (FF) and short form (SF)											
FF 24 items ρ	−0.39	0.22	0.27	0.09	0.32	0.07	0.15	0.09	0.06	−0.02	−0.42
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.002	.27	<.001
SF 16 items ρ	−0.39	0.22	0.26	0.10	0.32	0.12	0.14	0.08	0.05	−0.02	−0.40
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.006	.23	<.001
Operational skills											
FF 4 items ρ	−0.46	0.26	0.28	0.09	0.36	0.14	0.12	0.06	0.07	−0.04	−0.35
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.004	<.001	.02	<.001
SF 2 items ρ	−0.47	0.24	0.28	0.09	0.37	0.16	0.11	0.06	0.06	−0.05	−0.34
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.001	.004	.006	<.001
Navigation skills											
FF 3 items ρ	−0.35	0.15	0.19	0.05	0.26	0.007	0.1	0.1	0.07	−0.04	−0.34
P value	<.001	<.001	<.001	<.001	<.001	.72	<.001	<.001	<.001	.03	<.001
SF 2 items ρ	−0.35	0.15	0.20	0.06	0.26	0.02	0.10	0.09	0.07	−0.03	−0.35
P value	<.001	<.001	<.001	.008	<.001	.40	<.001	<.001	<.001	.09	<.001
Information searching											
FF 3 items ρ	−0.26	0.14	0.21	0.09	0.23	0.08	0.13	0.08	0.04	−0.01	−0.31
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.03	.78	<.001
SF 2 items ρ	−0.24	0.12	0.19	0.10	0.22	0.07	0.13	0.07	0.04	−0.00	−0.28
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.001	.03	.87	<.001
Evaluating reliability											
FF 3 items ρ	−0.31	0.15	0.17	0.06	0.25	0.06	0.10	0.06	0.05	−0.03	−0.33
P value	<.001	<.001	<.001	.005	<.001	.003	<.001	.001	.008	.19	<.001
SF 2 items ρ	−0.26	0.13	0.16	0.07	0.21	0.05	0.10	0.05	0.05	−0.02	−0.30
P value	<.001	<.001	<.001	.002	<.001	.02	<.001	.006	.02	.42	<.001
Determining relevance											
FF 3 items ρ	−0.26	0.15	0.17	0.06	0.22	0.08	0.12	0.06	0.01	−0.01	−0.28
P value	<.001	<.001	<.001	.005	<.001	<.001	<.001	.002	.56	.72	<.001
SF 2 items ρ	−0.27	0.14	0.17	0.06	0.22	0.07	0.11	0.06	0.01	−0.00	−0.27
P value	<.001	<.001	<.001	.006	<.001	.001	<.001	.006	.53	.86	<.001
Generating content											

XSL•FO
RenderX

(Sub)scale	Age	Education	Family income	Personal income	Hours spent online	Health-related internet use	General health	Use general practitioner	Use medical specialist	Use mental health care	SILS
FF 3 items ρ	−0.13	0.28	0.20	0.09	0.12	0.05	0.13	0.01	−0.01	0.01	−0.34
P value	<.001	<.001	<.001	<.001	<.001	.02	<.001	.81	.64	.59	<.001
SF 2 items ρ	−0.15	0.28	0.20	0.09	0.13	0.07	0.12	−0.01	−0.02	−0.01	−0.34
P value	<.001	<.001	<.001	.001	<.001	.008	<.001	.81	.30	.74	<.001
Protecting privacy											
FF 3 items ρ	−0.22	0.05	0.05	−0.006	0.13	−0.11	0.05	0.08	0.09	−0.02	−0.28
P value	<.001	.09	.09	.83	<.001	<.001	.07	.005	<.001	.41	<.001
SF 2 items ρ	−0.12	0.04	0.06	−0.003	0.06	−0.09	0.03	0.03	0.09	−0.02	−0.20
P value	<.001	.11	.03	.91	.03	.001	.30	.20	.002	.52	<.001
Using security measures											
FF and SF 2 items ρ	−0.26	0.19	0.21	0.12	0.24	0.12	0.11	0.06	0.04	−0.01	−0.33
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.003	.07	.55	<.001

Discussion

This paper describes the validation of a revised version of the DHLI, the DHLI 2.0. The original instrument was published in 2017, and 8 years later, it has been updated to better reflect the rapidly evolving digital landscape. Several items were reworded to capture the widespread use of smartphones and tablets. Also, a new 2-item subscale was added to address digital safety, specifically the challenges of choosing and using secure passwords and applying two-step authentication when accessing personalized systems. In our view, the DHLI 2.0 offers a more accurate representation of the competencies required for digital health literacy today.

DHLI 2.0 demonstrated strong reliability, with Cronbach α values that were compared to, and in some cases slightly higher than, those of the original instrument. Construct validity was also adequate, with correlations to sociodemographics, internet use, and general health literacy as expected and consistent with those observed for the original DHLI [7] and in other recent systematic reviews focusing on these relationships [26,27]. A recent review by Wang et al [28] identified DHLIs developed over the past 2 decades, each addressing different aspects of digital health literacy, including the DHLI. Compared to the other included instruments, the DHLI and DHLI 2.0 offer a concise yet theoretically well-grounded instrument [7]. It captures a broad spectrum of skills, ranging from basic operational abilities to higher-order information-processing skills, while also addressing the increasingly important domain of privacy and data security. Moreover, to minimize overestimation bias, common for self-reported skills, items are phrased to assess how easy or difficult respondents find certain tasks or how frequently they encounter difficulties. This validation study also included an SF DHLI 2.0, consisting of 16 items instead of 24. The SF demonstrated good reliability and internal consistency. It facilitates easier administration and allows for integration into larger sets of survey instruments.

In revising the DHLI, we deliberately chose to retain only self-report items and exclude the performance-based tasks that were a unique feature of the original version [7]. Although innovative, those tasks turned out to be too complex to administer and provided only limited added value, mostly in terms of face validity. Furthermore, their applicability was highly dependent on the device, browser, and brand used by respondents (eg, desktop vs mobile), making it nearly impossible to develop a universally valid instrument for use in anonymous research settings. Similar challenges are reported by Crocker et al [29] in their systematic scoping review of performance-based eHealth literacy measures, who also highlight the time-consuming nature of these assessments for both participants and researchers, as well as their equipment requirements. Nevertheless, self-report measures have well-known limitations, and future research could explore the integration of performance-based items into the revised DHLI, also for screening in practice settings.

Some limitations of this study need to be acknowledged. First, validation was conducted through an online questionnaire via a citizen panel. Although our sample was large and broadly representative, the online format likely excluded individuals with the lowest digital skills and contributed to the overrepresentation of higher-educated respondents. The response via the citizen panel from Planbureau Fryslân might have contributed to the overrepresentation of older adults. This may limit the generalizability of the findings, particularly for less

XSL•FO
RenderX

educated or younger and middle-aged populations. Validation of the instrument among these currently underrepresented populations is needed in future research to ensure its applicability and reliability across a broad range of people in the population. At present, a study in a patient sample is ongoing, which will provide further insight into the validity of the DHLI 2.0 in a more vulnerable population. Second, the reliability of the scale "Using security measures" did not reach the conventional reliability threshold of α=0.7, which may be due to its coverage of 2 distinct skills (using strong passwords and using 2-step authentication). These items should therefore be considered separately, rather than combined in a sum score, and further research is warranted to determine whether this domain would benefit from the inclusion of additional items. Third, the correlation between DHLI 2.0 scores and health-related internet use was relatively low. This may partly reflect the relatively high mean age of our sample; although many older adults do use the internet regularly for health purposes, they may not possess the full range of skills needed to do so effectively, efficiently, and safely, which is also reported in previous studies [30,31]. Furthermore, motivational factors could influence the relationship, for example, when individuals do not perceive the internet as a good or reliable source of health information and therefore use it less frequently. Finally, health status itself may play a role, as relatively healthy individuals may have limited need for health-related digital tools (eg, patient portals or online communities), despite having adequate or even high levels of digital health literacy [32]. As a final limitation, the total scores of the subscales demonstrated approximately normal distributions, while some subscales exhibited pronounced ceiling effects. The subscale "Protecting privacy" showed the most substantial ceiling effect, potentially influenced by the phrasing of its introductory item: "When you post a message on a public forum or social media…" This condition led to reduced response rates, as only respondents who engage in such online activity reported an answer. Consequently, the responses may reflect a self-selected group with high privacy awareness within the cohort. The ceiling effects in operational and navigation skills might be affected by the online nature of the questionnaire.

The online world and available technologies are changing rapidly. Since the development of the DHLI 2.0, the integration of artificial intelligence into health applications and internet browsers has accelerated significantly. This further stresses the importance of being able to judge the reliability of information, including being aware of its source and the applicability to oneself [33,34]. In this validation study, we specifically see that the skill to evaluate reliability of online information is what many people find challenging. These findings highlight the importance of supporting individuals in developing critical evaluation skills, particularly as digital health environments become more complex and information sources less transparent. This also underscores the need for further research on how to capture the skills required in an increasingly artificial intelligence–driven digital health environment. Future research should also examine the applicability and utility of the DHLI 2.0 among groups at higher risk of limited digital health literacy. Another interesting avenue for future work is to explore the potential of this instrument in practice; this could be within health care, but also within more community- or welfare-oriented organizations. Although not yet validated for that purpose, the DHLI 2.0 (especially the SF) could support decision-making in health consultations, for example, in determining the suitability of specific eHealth tools for patients.

In conclusion, the DHLI 2.0 provides an up-to-date, reliable, valid, and concise measure of digital health literacy, grounded in a strong theoretical framework and covering 8 key domains of digital health literacy. The SF DHLI proves to be a valuable alternative to measure the broad spectrum of digital health literacy skills using 16 items.

Acknowledgments

We thank Jesse David Marinus and Jolijn Hutjes from Planbureau Fryslân for granting access to the panel, for conducting the qualitative pilot study to make final adjustments to the instrument, for their valuable input on the data collection design, and for programming the questionnaire in their survey software. We thank Bart Koemans for his work on the preliminary validation of the DHLI 2.0 in his bachelor thesis. We thank all the respondents for their participation in the study. Artificial intelligence tools were used during the preparation of this manuscript for refinement, correction, and editing of the manuscript to improve the clarity of language, not for any other purposes.

Funding

The authors appreciate the financial support of the Faith Research Consortium.

Data Availability

The data used in this research are available upon request, provided that the request aligns with the established conditions, which can be found on the Planbureau Fryslân website [35]. Researchers interested in using these data should contact the Planbureau Fryslân Office at info@planbureaufryslan.nl.

Authors' Contributions

RvdV and CD were responsible for the conceptualization of DHLI 2.0, and RvdV, CD, and IT conceptualized the validation study. IT initiated the collaboration with Panel Fryslân. All authors contributed to the methodology. IT conducted the formal data

XSL•FO
RenderX

analyses, which were verified by RvdV and CD. RvdV and IT drafted the original manuscript, and CD and JvV critically reviewed and edited the manuscript. All authors read and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Digital health literacy instrument 2.0 full form.
[PDF File, 101 KB - jmir_v28i1e86879_app1.pdf]

Multimedia Appendix 2
Digital health literacy instrument 2.0 short form.
[PDF File, 113 KB - jmir_v28i1e86879_app2.pdf]

Multimedia Appendix 3
Overview of alterations in DHLI 2.0.
[PDF File, 135 KB - jmir_v28i1e86879_app3.pdf]

References

1. van Tuyl L, Keuper J, Standaar L, Alblas E, Keij B. Toegankelijkheid van digitale zorg: signalen uit de e-healthmonitor 2023 [Article in Dutch]. TSG Tijdschr Gezondheidswet 2025;103(S1):3-7. [doi: 10.1007/s12508-025-00451-7]
2. Keij B, Versluis A, Alblas EE, Keuper JJ, van Tuyl LHD, van der Vaart R. 2023 eHealth monitor: state of digital care [Report in Dutch]. : Rijksinstituut voor Volksgezondheid en Milieu (RIVM); 2024 URL: https://www.rivm.nl/bibliotheek/rapporten/2024-0008.pdf [accessed 2026-03-18]
3. Mahajan S, Lu Y, Spatz ES, Nasir K, Krumholz HM. Trends and predictors of use of digital health technology in the United States. Am J Med 2021 Jan;134(1):129-134. [doi: 10.1016/j.amjmed.2020.06.033] [Medline: 32717188]
4. De Santis KK, Jahnel T, Sina E, Wienert J, Zeeb H. Digitization and health in Germany: cross-sectional nationwide survey. JMIR Public Health Surveill 2021 Nov 22;7(11):e32951. [doi: 10.2196/32951] [Medline: 34813493]
5. Yao R, Zhang W, Evans R, Cao G, Rui T, Shen L. Inequities in health care services caused by the adoption of digital health technologies: scoping review. J Med Internet Res 2022 Mar 21;24(3):e34144. [doi: 10.2196/34144] [Medline: 35311682]
6. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9. [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
7. van der Vaart R, Drossaert C. Development of the digital health literacy instrument: measuring a broad spectrum of health 1.0 and health 2.0 skills. J Med Internet Res 2017 Jan 24;19(1):e27. [doi: 10.2196/jmir.6709] [Medline: 28119275]
8. Duimel SLL, Linn AJ, Smets EMA, Smit ES, van Weert JCM. Profiling cancer patients based on their motives for seeking informational and emotional support online. Health Commun 2023 Dec;38(14):3223-3237. [doi: 10.1080/10410236.2022.2144287] [Medline: 36415021]
9. Neter E, Brainin E. Association between health literacy, eHealth literacy, and health outcomes among patients with long-term conditions. Eur Psychol 2019 Jan;24(1):68-81. [doi: 10.1027/1016-9040/a000350]
10. Barbosa MCF, Baldiotti ALP, Braga NS, et al. Cross-cultural adaptation of the digital health literacy instrument (DHLI) for use on Brazilian adolescents. Braz Dent J 2023;34(5):104-114. [doi: 10.1590/0103-6440202305346] [Medline: 38133465]
11. Park E, Kwon M. Testing the Digital Health Literacy Instrument for adolescents: cognitive interviews. J Med Internet Res 2021 Mar 15;23(3):e17856. [doi: 10.2196/17856] [Medline: 33720031]
12. Xie L, Hu H, Lin J, Mo PKH. Psychometric validation of the Chinese digital health literacy instrument among Chinese older adults who have internet use experience. Int J Older People Nurs 2024 Jan;19(1):e12568. [doi: 10.1111/opn.12568] [Medline: 37831059]
13. Bouclaous C, Kamand AA, Daher R, Alrazim A, Kaedbey HD. Digital health literacy and online information-seeking behavior of Lebanese university students in the time of the COVID-19 pandemic and infodemic. Nord J Digit Lit 2023 Apr 11;18(1):60-77. [doi: 10.18261/njdl.18.1.6]
14. Chen SC, Hong Nguyen NT, Lin CY, et al. Digital health literacy and well-being among university students: mediating roles of fear of COVID-19, information satisfaction, and internet information search. Digit Health 2023;9:20552076231165970. [doi: 10.1177/20552076231165970] [Medline: 37009305]
15. Dadaczynski K, Okan O, Messer M, et al. Digital health literacy and web-based information-seeking behaviors of university students in Germany during the COVID-19 pandemic: cross-sectional survey study. J Med Internet Res 2021 Jan 15;23(1):e24097. [doi: 10.2196/24097] [Medline: 33395396]
16. Shudayfat T, Bani Hani S, Al Qadire M. Assessing digital health literacy level among nurses in Jordanian hospitals. Electron J Gen Med 2023;20(5):em525. [doi: 10.29333/ejgm/13466]

XSL•FO
RenderX

17. Alhammad N, Alajlani M, Abd-Alrazaq A, Epiphaniou G, Arvanitis T. Patients' perspectives on the data confidentiality, privacy, and security of mHealth apps: systematic review. J Med Internet Res 2024 May 31;26:e50715. [doi: 10.2196/50715] [Medline: 38820572]
18. Panel Fryslân. Research methodology: sampling and recruitment of a representative internet panel [Report in Dutch]. : Planbureau Fryslân; 2023 URL: https://planbureau.frl/wp-content/uploads/2024/01/Verantwoording-herwerving-panel-2023.pdf [accessed 2026-02-09]
19. Morris NS, MacLean CD, Chew LD, Littenberg B. The single item literacy screener: evaluation of a brief instrument to identify limited reading ability. BMC Fam Pract 2006 Mar 24;7:21. [doi: 10.1186/1471-2296-7-21] [Medline: 16563164]
20. Vereniging voor statistiek en onderzoek handbook GDPR & statistical research. Vsonet. URL: https://vsonet.nl [accessed 2026-03-31]
21. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951 Sep;16(3):297-334. [doi: 10.1007/BF02310555]
22. Streiner DL, Norman GR, Cairney J. Health measurement scales: a practical guide to their development and use: Oxford University Press; 2014. URL: https://academic.oup.com/book/56061 [accessed 2026-03-18]
23. George D, Mallery P. SPSS for Windows Step by Step: A Simple Guide and Reference, 170 Update: Allyn & Bacon; 2010. URL: https://books.google.com.my/books?id=KS1DPgAACAAJ [accessed 2026-03-28]
24. Terwee CB, Bot SDM, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007 Jan;60(1):34-42. [doi: 10.1016/j.jclinepi.2006.03.012] [Medline: 17161752]
25. Description of the eight EQF levels. Europass. 2026. URL: https://europass.europa.eu/en/description-eight-eqf-levels [accessed 2026-02-10]
26. Jiang X, Wang L, Leng Y, et al. The level of electronic health literacy among older adults: a systematic review and meta-analysis. Arch Public Health 2024 Nov 7;82(1):204. [doi: 10.1186/s13690-024-01428-9] [Medline: 39511667]
27. Yuen E, Winter N, Savira F, et al. Digital health literacy and its association with sociodemographic characteristics, health resource use, and health outcomes: rapid review. Interact J Med Res 2024 Jul 26;13:e46888. [doi: 10.2196/46888] [Medline: 39059006]
28. Wang C, Chang L, Chen X, Kong J, Qi H. eHealth literacy assessment instruments: scoping review. J Med Internet Res 2025 Aug 20;27:e66965. [doi: 10.2196/66965] [Medline: 40835422]
29. Crocker B, Feng O, Duncan LR. Performance-based measurement of eHealth literacy: systematic scoping review. J Med Internet Res 2023 Jun 2;25:e44602. [doi: 10.2196/44602] [Medline: 37266975]
30. Waterworth S, Honey M. On-line health seeking activity of older adults: an integrative review of the literature. Geriatr Nurs 2018;39(3):310-317. [doi: 10.1016/j.gerinurse.2017.10.016] [Medline: 29198622]
31. Bachofner Y, Seifert A, Sepahniya S, Fabian C. Online health information-seeking among older adults and predictors of use, motivations, and barriers in the context of healthy aging: cross-sectional study. Online J Public Health Inform 2026 Jan 6;18:e77557. [doi: 10.2196/77557] [Medline: 41493565]
32. Qiu CS, Lunova T, Greenfield G, et al. Determinants of digital health literacy: international cross-sectional study. J Med Internet Res 2025 Jun 30;27:e66631. [doi: 10.2196/66631] [Medline: 40587845]
33. Coşkun AB, Elmaoğlu E, Buran C, Yüzer Alsaç S. Integration of ChatGPT and E-Health literacy: opportunities, challenges, and a look towards the future. J Health Rep Technol 2024;10(1):e139748. [doi: 10.5812/jhrt-139748]
34. Nutbeam D, Okan O, EUPHA-HL,-HP,-DH,-ETH,-FS,-GH, IUHPE Working Group on Health Literacy, Technical University of Munich (Germany) and Chair persons. 5.B. Round Table: improving digital health literacy in the era of generative artificial intelligence. Eur J Public Health 2024 Nov 1;34(Supplement_3). [doi: 10.1093/eurpub/ckae144.276]
35. Planbureau Fryslân. URL: https://planbureau.frl/ [accessed 2026-03-24]

Abbreviations

DHLI: digital health literacy instrument
DHLI 2.0: revised version of the DHLI
FF: full form
SF: short form
SILS: Single Item Literacy Screener

XSL•FO
RenderX

Edited by A Schwartz; submitted 31.Oct.2025; peer-reviewed by H Qi, JL Guo; revised version received 12.Feb.2026; accepted 23.Feb.2026; published 01.Apr.2026.

Please cite as:
van der Vaart R, Tuitert I, van 't Veer J, Drossaert C
Validation of the Updated Digital Health Literacy Instrument and Development of a Short Form: Online Survey Study of the General Population
J Med Internet Res 2026;28:e86879
URL: https://www.jmir.org/2026/1/e86879
doi:10.2196/86879

© Rosalie van der Vaart, Inge Tuitert, Job van 't Veer, Constance Drossaert. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 1.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Relationship Between Online Health Information Acquisition and Shared Decision-Making Among Patients With Diabetes: Cross-Sectional Survey Study

Yue Ming[1], PhD; Miriam L Matteson[2], PhD; Yao Zhang[3], PhD

[1]Rudolph Matas Library of the Health Sciences, Tulane University, New Orleans, LA, United States

[2]College of Communication and Information, Kent State University, Kent, OH, United States

[3]Department of Information Resource Management, School of Information and Communication, Nankai University, 38 Tongyan Road, Jinnan District, Tianjin, China

Corresponding Author:
Yao Zhang, PhD
Department of Information Resource Management, School of Information and Communication, Nankai University, 38 Tongyan Road, Jinnan District, Tianjin, China

Abstract

Background: Shared decision-making allows patients and clinicians to make decisions together to help determine the most appropriate option. Patients need comprehensive health information to participate and evaluate different options during the shared decision-making process. Patients with diabetes need to constantly monitor their health status. They experience an array of health information needs during their ongoing health management. Online health information acquisition is a common behavior among patients with diabetes, and online information can impact the interaction between patients with diabetes and health care providers.

Objective: This study explored the relationship between 2 types of online health information acquisition behavior (active online health information seeking and incidental online health information acquisition) and shared decision-making. It also investigated the mediating role of eHealth literacy during the information acquisition process among US patients with diabetes aged 18 to 44 years.

Methods: Participants were patients with diabetes aged 18 to 44 years in the United States and were recruited by a survey company, Centiment. The sampling process matched the national distribution of gender and age in the United States. An online survey questionnaire was distributed through Qualtrics. A total of 558 valid responses were collected. The average age of the sample was 35.91 (SD 6.04) years. Among the sample, 260 participants were men, 291 participants were women, and 7 participants identified their gender as other. Bivariate analyses and partial least squares structural equation modeling were used for data analysis. All data analyses were performed in R.

Results: The prevalences of active online health information seeking (mean 3.97, SD 0.78) and incidental online health information acquisition (mean 4.27, SD 0.78) were high among participants. Education was a key factor related to eHealth literacy ($P<.001$) and shared decision-making ($P<.001$). Model testing indicated that active online health information seeking was related to eHealth literacy (β=.192, 95% CI .067-.320) and shared decision-making (β=.234, 95% CI .123-.346). Incidental online health information acquisition was related to eHealth literacy (β=.335, 95% CI .205-.461). eHealth literacy was related to shared decision-making (β=.441, 95% CI .334-.536). Therefore, eHealth literacy partially mediated the relationship between active online health information seeking and shared decision-making, while it fully mediated the relationship between incidental online health information acquisition and shared decision-making.

Conclusions: This study contributes to the ongoing development of health communication strategies and the modification of health information training programs for patients with diabetes. The need for the information industry to deliver accurate and easy-to-understand health information to the public to support their decision-making process and encourage positive health behaviors is urgent.

(J Med Internet Res 2026;28:e86137) doi:10.2196/86137

KEYWORDS

online health information acquisition; active online health information seeking; incidental online health information acquisition; eHealth literacy; shared decision-making; health communication; diabetes mellitus

XSL•FO
RenderX

Introduction

Background

Engaging patients in the medical decision-making process has become a common practice to increase patients' self-care adherence and improve their wellness [1,2]. Shared decision-making is a decision-making process involving both physicians and patients [3]. The process allows patients and clinicians to make decisions together to help determine what is the most appropriate option for each patient. The successful implementation of such a process requires health practitioners to share knowledge and patients to consent to the process. More importantly, shared decision-making encourages patients to feel confident in making decisions based on their own circumstances [4]. To take full advantage of the shared decision-making process, patients need comprehensive information to participate in the process and to evaluate different options [5,6]. For example, depending on different health conditions, patients may need time to process new information (eg, new treatments, drug effects) and then take into consideration their personal preferences (eg, affordability, additional underlying health conditions) [7,8]. A few studies have investigated the relationship between information behavior factors and shared decision-making from various perspectives. In a meta review aimed at identifying the influences on information exchange in shared decision-making, researchers argued that health literacy mediates the relationship between information exchange and the shared decision-making process [9]. Other scholars have suggested that during the shared decision-making process, patients may access health information through both health providers' digital channels and external sources [10,11].

With the rapid development of information technology, the internet serves as an important source of health information [12,13]. Researchers have illustrated that young adults tend to use online technologies more actively than the older generation [14]. However, the availability of online health information might not necessarily contribute to better understanding and use of health information [15,16]. Key information practices such as accumulating knowledge, staying well informed, and using information to make decisions are the foundations of eHealth literacy and are critical for effective information seeking and use. Based on the unique characteristics of the online environment, eHealth literacy has become a focal point for researchers studying online information-seeking behavior. eHealth literacy extends various aspects of traditional information literacy to focus on an individual's ability to seek, process, and interpret health information in the online environment [17]. Researchers have suggested that eHealth literacy is associated with online health information-seeking behaviors within different health contexts [18,19]. Moreover, eHealth literacy has also been identified as a key factor related to various health behaviors and health outcomes [20-22].

Information seeking refers to the process of people looking for information to fulfill a purpose or solve a problem [23]. The phrase "information acquisition" provides a broader umbrella to include purposeful information seeking and passive information encountering [24,25]. The ecological model of information use suggests that both active information seeking and incidental information acquisition can occur during our everyday lives [26]. More specifically, active information seeking refers to people purposefully seeking and obtaining information to address problems, while incidental information acquisition refers to people's passive consumption of information from exposure to information sources. Furthermore, when users intentionally seek information that they know they need, they may also unintentionally collect other information they encounter during the information-seeking process, though they do not realize that they may need it. Thus, both types of information acquisition behavior can occur simultaneously, especially within the current digital age, due to the ease of obtaining information online.

The framework of "searching as a learning process" demonstrates the intertwined nature of searching and learning [27]. It suggests that people acquire knowledge critically and receptively through iterative and reflective search behaviors. During this process, comprehensive information skills, such as analyzing, evaluating, and appraising information, can also be learned and improved [28]. Since the model of the information search process [29] demonstrated that information seeking commonly involves knowledge construction and learning, it is important to note that similarly, people can also learn and improve their comprehensive information skills during the process of online health information seeking. Furthermore, online health information acquisition can not only contribute to the knowledge acquisition process but also increase the level of exposure to technology and the internet for health information. Previous studies have demonstrated that exposure, such as the frequency of use of internet and web searches regarding health information, is positively related to eHealth literacy [30,31]. Researchers have also identified the mediating role of eHealth literacy during the process of online information seeking and doctor-patient interaction in China [32]. Specifically, eHealth literacy was one of the mediators that fully and positively mediated the association between online health information-seeking behavior and doctor-patient interactions.

Objectives

Chronic diseases require long-term self-care in both medical and everyday life. People living with chronic diseases need to constantly collect useful information to adjust their behaviors. Diabetes is one of the most common chronic diseases and is a serious public health problem around the world. Data suggest that about 589 million adults worldwide had diabetes in 2024, with a projection that the number would rise to about 853 million by 2050 [33]. Scholars have also illustrated that the rate of diabetes is increasing significantly among young adults [34]. Diabetes progresses through various stages that involve various types of treatments. In addition to medical treatment, patients with diabetes must follow recommended lifestyle practices to slow the decline in their health [35]. Patients with diabetes can decide which treatments or approaches are the best for them. The process of making such decisions may not relate to medical conditions or clinical recommendations but may also involve other factors, such as costs, side effects, and other specific individual contexts [36-38]. Patients with diabetes need to constantly monitor their health status to keep the disease under

XSL•FO
RenderX

control and prevent it from progressing to more complicated stages. This constant monitoring suggests that patients with diabetes experience an array of health information needs during their ongoing health management. By obtaining such information, people living with diabetes can exchange opinions with health care providers and adjust their behaviors accordingly. However, when further examining the impacts of health information acquisition on users' health behaviors in the context of diabetes, researchers [39-42] tend to focus on either active information seeking or incidental information acquisition rather than investigating both behaviors at the same time. It is also important to note that the information sources for both health information acquisition behaviors can vary significantly. Therefore, the cognitive process following active information seeking or incidental information acquisition might differ due to the distinct levels of individuals' perception of various information sources. As a result, effects from both online health information acquisition behaviors might be reflected distinctively in health behaviors through eHealth literacy. Moreover, there is a lack of research that examines shared decision-making through the perspective of information acquisition or observation of patients' information acquisition behaviors. It is necessary to extend the understanding of the links between different patterns of health information acquisition behaviors and shared decision-making to promote effective health communication strategies and online health information consumption skills. Additionally, since only a few studies have investigated the role of eHealth literacy during the shared decision-making process, little is known about the possible roles of eHealth literacy in the link between information acquisition and shared decision-making. Data from the Pew Research Center suggest that young adults tend to use online technologies more actively than older generations [14]. Furthermore, the prevalence of multiple chronic conditions is higher among elderly adults [43]. These additional health conditions could also potentially affect patients' information acquisition behaviors. To minimize the potential impacts of additional chronic conditions and unfamiliarity with internet usage, the study focuses on patients with diabetes who are 18 to 44 years old. Therefore, this study aims to investigate the model in the context of diabetes among young adults aged 18 to 44 years in the United States to understand the potentially influential roles of 2 types of online information acquisition behaviors (active seeking and incidental acquisition) and eHealth literacy on shared decision-making.

Hypotheses

We propose a model (Figure 1) to test the following hypotheses:

- H1: Active online health information seeking is related to eHealth literacy.
- H2: Active online health information seeking is related to shared decision-making.
- H3: Incidental online health information acquisition is related to eHealth literacy.
- H4: Incidental online health information acquisition is related to shared decision-making.
- H5: eHealth literacy is positively related to shared decision-making.

Figure 1. Hypothesis model.

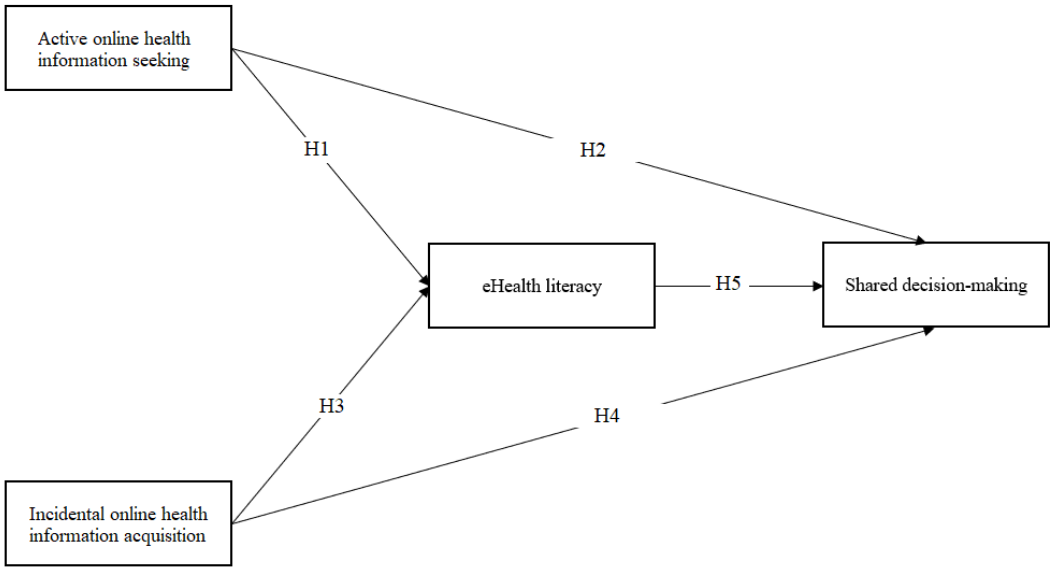

Methods

The CHERRIES (Checklist for Reporting Results of Internet E-Surveys) guideline [44] was followed to report the study.

Ethical Considerations

An institutional review board application for this study was reviewed by the Kent State Institutional Review Board (application number: 627). The application received an exempt approval. Informed consent was obtained from participants. An online consent form was provided on the first page of the online survey. Participants had the option to opt out on the consent

XSL•FO
RenderX

form page. This study collected survey data from participants who consented. No personally identifying information was collected from participants. The data were downloaded from a password-protected Qualtrics account in an anonymous data format. The data were stored in password-protected cloud storage solutions provided by the institution. Participants received compensation from the research panel company, Centiment. They were compensated via PayPal, with an option to donate survey earnings to a nonprofit organization of their choice.

Participants

This study is a part of a larger research project aimed at studying online health information acquisition among patients with diabetes in the United States. To achieve a statistical power of 80% at a .05 significance level to detect a medium effect (0.3) for the statistical analysis of the overall project, a power analysis was conducted, which determined that the recommended sample size is 489. After receiving institutional review board approval, we recruited participants through a research panel company, Centiment, who self-reported that they had been diagnosed with diabetes, were aged 18 to 44 years, and were living in the United States. Centiment mainly recruits survey participants through various social media sites. Its sampling process matches the national distribution of gender and age. An online questionnaire, which took about 12 minutes to complete, was distributed to participants through Qualtrics. The survey was distributed to eligible participants by Centiment from March 15 to March 28, 2023. As each item in the survey was required to be answered, there are no missing data. A total of 558 valid responses were collected.

Measures

The shared decision-making instrument was based on a 9-item shared decision-making questionnaire [45]. A 5-point Likert scale (ranging from strongly disagree to strongly agree) was used to determine participants' agreement with each item. An average score was calculated to reflect each participant's overall shared decision-making involvement. eHealth literacy was measured based on Norman and Skinner's eHealth Literacy Scale [17]. A 5-point Likert scale (ranging from strongly disagree to strongly agree) was used to determine participants' agreement with each item. An average score was calculated to reflect every participant's overall eHealth literacy level. Active online health information seeking was measured based on the active information-seeking instrument from Kelly et al [46]. A 5-point Likert scale (from strongly disagree to strongly agree) was used to determine participants' agreement with each item. An average score was calculated to reflect each participant's overall active online health information-seeking level about diabetes. Incidental online health information acquisition was measured based on the passive information-seeking instrument from Kelly et al [46]. A 5-point Likert scale (ranging from strongly disagree to strongly agree) was used to determine participants' agreement with each item. An average score was calculated to reflect each participant's overall incidental online health information acquisition level about diabetes. For detailed survey items, see Multimedia Appendix 1.

Data Analysis

First, descriptive analyses were conducted to understand the general characteristics of the study sample. Second, bivariate analyses including Pearson's correlation, 2-tailed *t* test, and Kruskal-Wallis test were used to explore bivariate relationships. Then, a partial least squares structural equation model (PLS SEM) was performed to test the proposed hypotheses. Bootstrapping was also conducted afterward. All analyses were performed in R. The R package SEMinR [47] was used for PLS SEM analysis and bootstrapping.

Results

The valid sample had an average age of 35.91 (SD 6.04) years. Among participants, 46.59% (n=260) were men, 52.15% (n=291) were women, and 1.25% (n=7) of the participants reported their gender as other. In terms of race, 57.53% (n=321) of the participants were White or Caucasian, 15.59% (n=87) were Black or African American, 11.65% (n=65) were Hispanic and/or Latinx, and 15.23% (n=85) were categorized as other. Regarding education level, about 31% (n=173) of the participants reported having no college experience, while 69% (n=385) reported having college experience. Regarding marital status, about 45.52% (n=254) of the participants were married, and 32.26% (n=180) of them are single, never married, followed by living as married or living with a romantic partner (n=76, 13.62%), divorced (n=39, 6.99%), and widowed (n=9, 1.61%). The average general health score of the sample is 2.93, with an SD of 1.07 and a range from 1 to 5. Bivariate analyses including Pearson correlation, 2-tailed *t* test, and Kruskal-Wallis test were conducted to identify possible relationships between demographic variables and eHealth literacy and shared decision-making, respectively. Notably, education level was significantly related to both eHealth literacy and shared decision-making. Specifically, compared to those who had no college experience, those with college experience demonstrated a higher level of eHealth literacy (group mean for no college experience=3.96, group mean for "with college experience"=4.18; $P<.001$) and shared decision-making (group mean for "no college experience"=3.84, group mean for "with college experience"=4.01; $P=0.004$). Moreover, the level of general health of the sample was negatively associated with eHealth literacy ($r=-0.18$; $P<.001$) and shared decision-making ($r=-0.22$; $P<.001$). That is, those in the sample who had poorer general health in the sample tended to maintain a higher level of eHealth literacy and participate in shared decision-making at a higher level. Detailed information about the sample is shown in Table 1.

XSL•FO
RenderX

Table . Demographic information of the sample (N=558).

Demographic characteristics	Values	eHealth literacy, *P* value	Shared decision-making, *P* value
Age (y), mean (SD)	35.91 (6.04)	.10	.57
Gender, n (%)		.12	.11
Men	260 (46.59)		
Women	291 (52.15)		
Other	7 (1.26)		
Race, n (%)		.06	.08
Black or African American	87 (15.59)		
Hispanic and/or Latinx	65 (11.65)		
White or Caucasian	321 (57.53)		
Other	85 (15.23)		
Education, n (%)		<.001	.004
No college experience	173 (31)		
With college experience	385 (69)		
Marital status, n (%)		.17	.28
Married	254 (45.52)		
Living as married or living with a romantic partner	76 (13.62)		
Divorced	39 (6.99)		
Widowed	9 (1.61)		
Single, never been married	180 (32.26)		
General health, mean (SD)	2.93 (1.07)	<.001	<.001

Pearson correlation was used to identify relationships between major independent variables and outcome variables. Mean, SD, and Cronbach α for each instrument are also reported. The sample reported an average score of 4.00 for shared decision-making (SD 0.84, α=0.91), 3.97 for active online health information seeking (SD 0.86, α=0.78), 4.28 for incidental online health information acquisition (SD 0.80, α=0.85), 4.11 for eHealth literacy (SD 0.69, α=0.89). Cronbach α for each instrument is over 0.70, which indicates a good fit. Table 2 discusses detailed results.

XSL•FO
RenderX

Table . Correlation between continuous variables.

Variables	1	2	3	4
1. Active online health information seeking				
r value	1			
P value	__[a]			
2. Incidental online health information acquisition				
r value	0.65	1		
P value	<.001	—		
3. eHealth literacy				
r value	0.41	0.45	1	
P value	<.001	<.001	—	
4. Shared decision-making				
r value	0.41	0.37	0.53	1
P value	<.001	<.001	<.001	—
Mean (SD; range)	3.97 (0.78; 1 - 5)	4.27 (0.78; 1 - 5)	4.11 (0.69; 1 - 5)	4.00 (0.84; 1 - 5)
Cronbach α	0.79	0.83	0.89	0.91

[a]Not applicable.

In order to test the model, we performed the Henze-Zirkler multivariate normality test on the data. The results ($P<.05$) indicated that there is not enough evidence to suggest the variables follow a multivariate normal distribution. We then decided to use PLS SEM to test the model because it does not assume data normality and it is more suitable for analyzing the data [48].

To perform the PLS SEM analysis, we first assessed the reflective measures of the study through item loadings, internal consistency reliability, average variance extracted (AVE), and discriminant validity. The general criteria for indicating validity for each index are as follows: (1) all item loadings need to exceed 0.70, (2) composite reliability values need to be between 0.70 and 0.95, (3) AVE values need to exceed 0.50, and (4) discriminant validity values need to be lower than 0.85. According to the results, only 2 items (eHealth literacy item 7 and shared decision-making item 1) had loadings lower than 0.70, while all other item loadings exceeded the threshold. These two items were kept in the analysis for two reasons: (1) all other indices (internal consistency reliability, AVE, and discriminant validity) indicated that the construct is acceptable [49], and (2) for items with a loading between 0.4 and 0.7, it is recommended to remove them only when their removal would result in composite reliability and AVE increasing to meet the thresholds [50]. All other indices for all constructs met the criteria mentioned above. Table 3 shows factor loadings and reliability indices, and Table 4 shows discriminant validity.

XSL•FO
RenderX

Table . Factor loadings for the partial least squares structural equation modeling (PLS SEM) analysis.

Variables	Loading	Alpha	Composite reliability	AVE[a]
Active online health information seeking		0.791	0.856	0.544
1	0.716			
2	0.744			
3	0.738			
4	0.721			
5	0.768			
Incidental online health information acquisition		0.832	0.899	0.748
1	0.835			
2	0.889			
3	0.870			
eHealth literacy		0.891	0.913	0.568
1	0.757			
2	0.761			
3	0.778			
4	0.804			
5	0.792			
6	0.718			
7	0.694			
8	0.715			
Shared decision-making		0.910	0.926	0.586
1	0.576			
2	0.708			
3	0.730			
4	0.802			
5	0.812			
6	0.770			
7	0.821			
8	0.842			
9	0.793			

[a]AVE: average variance extracted.

Table . Discriminant validity.

Variables	1	2	3	4
1. Active online health information seeking				
2. Incidental online health information acquisition	0.810			
3. eHealth literacy	0.489	0.526		
4. Shared decision-making	0.484	0.397	0.586	

Next, we examined the hypotheses using PLS SEM to estimate the coefficients of each path in the model and performed bootstrapping to estimate a CI to determine the statistical significance of the coefficients. The results for each path are shown in Table 5 and Figure 2.

XSL•FO
RenderX

Table . Estimations for each path of the partial least squares structural equation modeling (PLS SEM) analysis (after bootstrapping).

Variables	eHealth literacy[a] β (95% CI)	Shared decision-making[b] β (95% CI)
Active online health information seeking	.192 (.067 to .320)	.234 (.123 to .346)
Incidental online health information acquisition	.335 (.205 to .461)	−.007 (−.132 to .113)
eHealth literacy	__[c]	.441 (.334 to .536)

[a]Adjusted R^2: 0.230.
[b]Adjusted R^2: 0.323.
[c]Not applicable.

Figure 2. Model results of partial least squares structural equation modeling (PLS SEM).

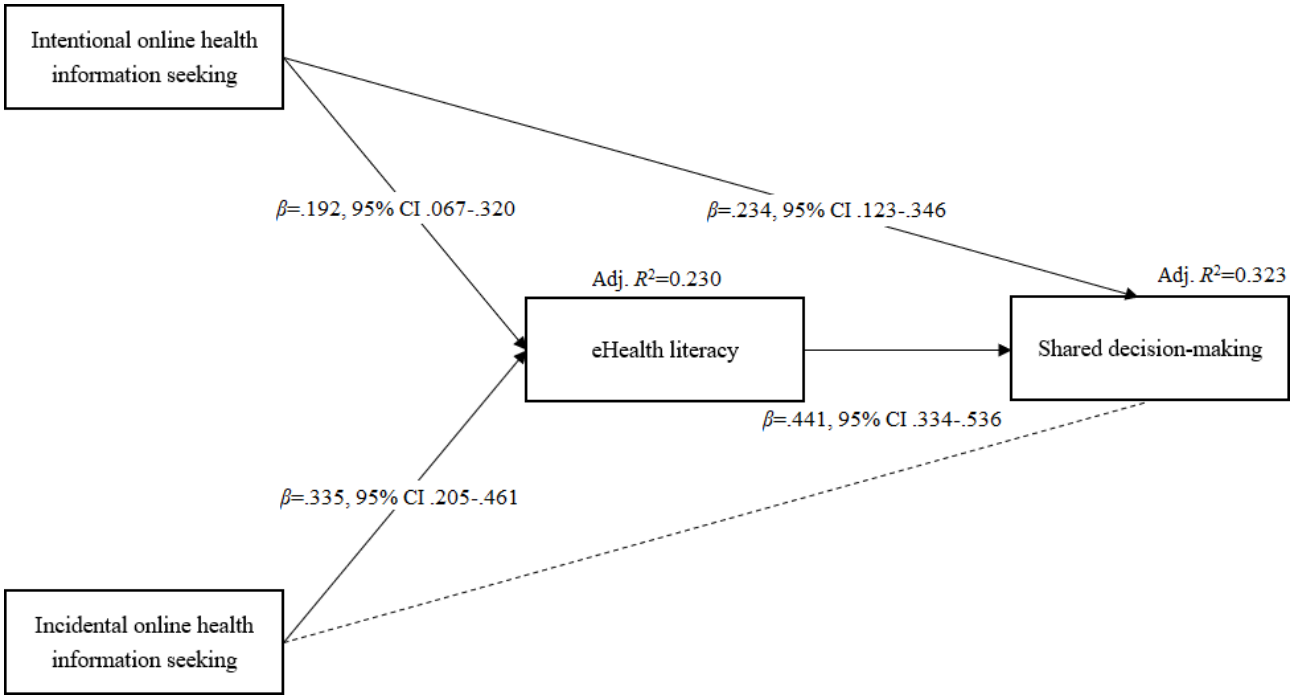

Specifically, both active online health information seeking (β=.192, 95% CI .067-.320) and incidental online health information acquisition (β=.335, 95% CI .205-.461) were positively associated with eHealth literacy; thus, H1 and H3 were supported. Active online health information seeking (β=.234, 95% CI .123-.346) was positively associated with shared decision-making; thus, H2 was supported. eHealth literacy was positively associated with shared decision-making (β=.441, 95% CI .334-.536); thus, H5 was supported. There is not enough evidence to suggest that incidental online health information acquisition was directly associated with shared decision-making; thus, H4 was not supported. In other words, while both active online health information seeking and incidental online health information acquisition were positively associated with eHealth literacy, mediation effects from eHealth literacy on shared decision-making were different for these 2 types of online health information acquisition behaviors. The data show that eHealth literacy partially mediated the relationship between active online health information seeking and shared decision-making and fully mediated the relationship between incidental online health information acquisition and shared decision-making.

Additionally, the value of the adjusted R^2 for eHealth literacy (Adj. R^2=0.230) indicated that active online health information seeking and incidental online health information acquisition contributed to about 23% of the variance in eHealth literacy. Meanwhile, the value of the adjusted R^2 for shared decision-making (Adj. R^2=0.323) suggested that active online health information seeking, incidental online health information acquisition, and eHealth literacy contributed to about 32.3% of the variance in shared decision-making. These results further illustrate that the model explains a significant amount of the variance in eHealth literacy and shared decision-making.

Discussion

Main Findings

This study explored the association between 2 types of online health information acquisition behavior and shared decision-making and investigated the role of eHealth literacy within the process among US patients with diabetes aged 18 to 44 years. Unlike previous studies, this study investigated both active information seeking and incidental information acquisition rather than only concentrating on active information seeking alone or merging the 2 types of information acquisition

XSL•FO
RenderX

together. Multiple hypotheses were proposed within the model and examined through a PLS SEM approach. Descriptive analysis results suggest that the prevalence of incidental online health information acquisition is high among the target population. This aligns with other scholars' arguments that younger people tend to incidentally acquire information through encountering it rather than actually searching for it [51]. This suggests that it is necessary and urgent to raise awareness among information and health professionals of active information-seeking behavior in young people with diabetes and develop appropriate strategies to distribute accurate and timely health information and messages to them. For instance, information professionals and public health agencies could collaborate to increase the promotion of structured and accurate health information through popular online platforms and enhance the readability and interactivity of such information.

The results from bivariate analyses indicated that education is a key factor related to eHealth literacy and shared decision-making. Those who have college experience tend to have a higher level of eHealth literacy and shared decision-making. This finding echoes numerous previous studies and supports the conclusion that it is crucial to increase education levels among the population to help improve population health and that education can be viewed as a form of health policy to enhance quality of life [52-54]. Additionally, negative relationships between general health level and eHealth literacy, as well as general health level and shared decision-making are worth noting. Compared to people with better general health, those with poorer general health have a greater need for health information. As cognitive need is a factor that motivates users to perform information behavior [55], it follows that a greater need for health information could lead to more health information acquisition online. With more frequent online health information acquisition, those with poorer health could get more exposure to health information via the internet. As previously illustrated, frequent exposure to online health information can be related to higher levels of eHealth literacy [30,31]. Furthermore, people with poor health might have a greater need to improve their health status, and shared decision-making is an effective step for them to take to improve their health situation [56], while people with good or excellent health may have less need to participate in shared decision-making. Therefore, these 2 negative relationships observed within our sample are important to highlight as opportunities for health professionals to develop targeted communication strategies.

The results from our model testing using PLS SEM indicated that eHealth literacy mediated the relationship between both types of online health information acquisition behavior and shared decision-making differently. Specifically, active online health information seeking was associated with both eHealth literacy and shared decision-making; therefore, eHealth literacy partially mediated the relationship between active online health information seeking and shared decision-making. Incidental online health information acquisition was only associated with eHealth literacy and shared decision-making; therefore, eHealth literacy fully mediated the relationship between incidental online health information acquisition and shared decision-making. The results show that eHealth literacy is a crucial mechanism that connects online health information acquisition to shared decision-making. This suggests 2 interesting paths that can increase shared decision-making: increasing one's level of active online health information seeking or increasing individual eHealth literacy skills. These results partially align with previous research, which found that active online information seeking is a crucial factor related to shared decision-making [10], and its role can be enhanced through eHealth literacy [32]. Although there are previous studies that indicate that incidental online information acquisition may be related to other types of decision-making [57,58], no other study has examined the relationship between incidental online information acquisition and shared decision-making. This study addresses that gap by exploring the role that incidental online health information acquisition plays in shared decision-making. However, it is worth noting that because this study measured shared decision-making in a general way rather than focusing on diabetes-related shared decision-making, the specific relationship between diabetes online information acquisition and shared decision-making engagement on diabetes still needs further investigation.

The results also suggest that in addition to the active health online information seeking, incidental online health information acquisition could also help an individual search for, evaluate, and use online health information (eHealth literacy). Through an increased level of eHealth literacy, those who have a higher level of incidental online health information acquisition may increase their engagement in shared decision-making. These 2 paths also highlight differences between active online health information seeking and incidental online health information acquisition, while still demonstrating the benefits of incidental online health information acquisition. With the broad use of social media and other internet technologies, incidental exposure to different types of online information is high [58,59]. Thus, it is possible that there are benefits to incidental online information acquisition more generally, beyond the health context.

Implications

Practically, the results of this study can contribute to the continuing development of health communication strategies and the modification of health information training programs for patients with diabetes. First, this study highlights the importance for the government and related agencies to continue enhancing health education policies and increasing health education levels among the population [60,61]. Second, in order to promote shared decision-making among patients with diabetes, it is necessary to improve information seekers and users' eHealth literacy and the credibility of online information that information seekers may incidentally pay attention to. Proper training programs or workshops from information institutions and health agencies could be provided to patients with diabetes to help them identify high-quality health information resources and gain useful online information-seeking skills amid an ever-changing online environment. The need for this training is becoming increasingly urgent, as recent research points to a deterioration in the effectiveness of search engines in returning relevant results [62].

XSL•FO
RenderX

Further, since the shared decision-making process involves a 2-way communication process, it is also necessary to raise awareness of how to improve the communication skills of health professionals. Specific technologies, such as clinical decision support tools, may be harnessed to support medical professional or patient communication [63,64]. To improve the equity of health care services, strategies to help reduce complicit bias, improve communication effectiveness, and enhance health information presenting skills are also highly paramount [65,66]. Health agencies and medical schools can help prepare both patients and health providers with the information and skills to be able to maximize the benefits of the shared decision-making process.

Moreover, as various researchers have illustrated, similar concepts such as artificial serendipity can be refined and shaped within digital environments [67,68]. It is also crucial for health agencies to use emerging technologies to improve the quality of health information that would be recommended or exposed to patients with diabetes incidentally. According to Hassoun et al [51], younger generations tend to start their information journey by scanning and encountering online and then interpret and consume information with other online peers together as a social group. It is essential to keep an eye on social media platforms to monitor the quality of health information. As many researchers have pointed out, social media algorithms manipulate the content that users read and encounter online [69,70]. Thus, it is crucial for the information industry to make an effort to deliver accurate health information to the public in support of their decision-making process and encourage positive health behaviors.

As high-speed internet access remains unreliable in parts of the United States, it is also important to close the gap in access to and use of the internet, especially among patients with diabetes. Community health programs and local institutions can play an important role in this effort. For example, providing free access to computers within local public libraries and organizing tutorial programs to help those not familiar with the online environment could bridge the information gap for patients with diabetes with limited internet access. Moreover, since the number of patients with diabetes is large and patients are not all the same, for instance, older adults tend to be less familiar with digital technologies, it is necessary to provide customized training programs to address issues for different groups of patients with diabetes.

Limitations and Future Directions

There are several limitations within this study. First, it used cross-sectional data; therefore, the results can only contribute to association instead of causation effects. Second, the instrument of each variable is self-reported. There might be potential bias in the responses. For instance, one can self-report with great confidence on finding health information online but actually have false perceptions about the accuracy of online health information sources. Third, the study examined only online information acquisition, while other ways of information acquisition, such as interpersonal information acquisition, were not investigated. Fourth, this study did not collect information from participants regarding the type of diabetes they were diagnosed with. Patients with different types of diabetes (type 1 diabetes, type 2 diabetes, or gestational diabetes) may focus on different types of information from various sources. Fifth, the study investigated only adult patients with diabetes who are under 45 years. Since younger adults tend to use online technologies more frequently compared to older adults, it is important to note that the generalizability of the study findings is limited. Moreover, the sampling process only matched national distribution on gender and age. Matching more key demographic characteristics, such as race, ethnicity, and region, could further enhance the quality of the collected data. Additionally, other important social determinants were not included in this study, such as food insecurity, safety, homelessness, drug use, and comorbidities. These factors may also be related to the information behaviors of patients with diabetes.

Future studies could collect longitudinal data to further examine causation effects. The model could also be modified and extended by including more tailored measurements to other health contexts or populations within other age groups. It would also be useful for future studies to separately measure patients with type 1, type 2, and gestational diabetes to examine whether there are differences in their information acquisition behaviors and related impacts. Furthermore, it would be more appropriate for future studies to contextualize general measurements within the study context to collect context-specific data.

Acknowledgments

The authors did not use generative artificial intelligence in any portion of the manuscript generation.

Funding

This study was funded by the Research and Creative Activity Fund (RACAF) from Kent State University.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: MLM, YM, YZ
Data curation: YM

XSL•FO
RenderX

Formal analysis: YM
Investigation: YM
Methodology: MLM, YM
Project administration: MLM, YM
Writing – original draft: MLM, YM, YZ
Writing – review & editing: MLM, YM, YZ

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey items for key variables.
[DOCX File, 18 KB - jmir_v28i1e86137_app1.docx]

References

1. Riegel B, Dunbar SB, Fitzsimons D, et al. Self-care research: where are we now? Where are we going? Int J Nurs Stud 2021 Apr;116:103402. [doi: 10.1016/j.ijnurstu.2019.103402] [Medline: 31630807]
2. Sweeney JC, Danaher TS, McColl-Kennedy JR. Customer effort in value cocreation activities: improving quality of life and behavioral intentions of health care customers. J Serv Res 2015;18(3):318-335. [doi: 10.1177/1094670515572128]
3. Hoffmann TC, Montori VM, Del Mar C. The connection between evidence-based medicine and shared decision making. JAMA 2014 Oct 1;312(13):1295-1296. [doi: 10.1001/jama.2014.10186] [Medline: 25268434]
4. Alkhayyat SS, Hudairi H, Alqahtani RM, et al. Assessing patient confidence and satisfaction about the shared decision-making meetings for planning cancer chemotherapy. Cureus 2019 Dec 22;11(12):e6445. [doi: 10.7759/cureus.6445] [Medline: 31893191]
5. Bomhof-Roordink H, Gärtner FR, Stiggelbout AM, Pieterse AH. Key components of shared decision making models: a systematic review. BMJ Open 2019 Dec 17;9(12):e031763. [doi: 10.1136/bmjopen-2019-031763] [Medline: 31852700]
6. Montori VM, Ruissen MM, Hargraves IG, Brito JP, Kunneman M. Shared decision-making as a method of care. BMJ Evid Based Med 2023 Aug;28(4):213-217. [doi: 10.1136/bmjebm-2022-112068] [Medline: 36460328]
7. Adams JR, Drake RE, Wolford GL. Shared decision-making preferences of people with severe mental illness. Psychiatr Serv 2007 Sep;58(9):1219-1221. [doi: 10.1176/ps.2007.58.9.1219] [Medline: 17766569]
8. Matthews EB, Savoy M, Paranjape A, et al. shared decision making in primary care based depression treatment: communication and decision-making preferences among an underserved patient population. Front Psychiatry 2021;12:681165. [doi: 10.3389/fpsyt.2021.681165] [Medline: 34322040]
9. Edwards M, Davies M, Edwards A. What are the external influences on information exchange and shared decision-making in healthcare consultations: a meta-synthesis of the literature. Patient Educ Couns 2009 Apr;75(1):37-52. [doi: 10.1016/j.pec.2008.09.025] [Medline: 19036550]
10. Dahl AJ, Milne GR, Peltier JW. Digital health information seeking in an omni-channel environment: a shared decision-making and service-dominant logic perspective. J Bus Res 2021 Mar;125:840-850. [doi: 10.1016/j.jbusres.2019.02.025]
11. Durand MA, Carpenter L, Dolan H, et al. Do interventions designed to support shared decision-making reduce health inequalities? A systematic review and meta-analysis. PLoS ONE 2014;9(4):e94670. [doi: 10.1371/journal.pone.0094670] [Medline: 24736389]
12. Delemere E, Maguire R. Caregivers of children feel confident about using the internet for health information. Health Info Libraries J 2023 Mar;40(1):54-69. [doi: 10.1111/hir.12430] [Medline: 35396806]
13. Wang X, Shi J, Kong H. Online health information seeking: a review and meta-analysis. Health Commun 2021 Sep;36(10):1163-1175. [doi: 10.1080/10410236.2020.1748829] [Medline: 32290679]
14. Faverio M. Share of those 65 and older who are tech users has grown in the past decade. Pew Research Center. 2022. URL: https://www.pewresearch.org/short-reads/2022/01/13/share-of-those-65-and-older-who-are-tech-users-has-grown-in-the-past-decade [accessed 2026-04-15]
15. Cuan-Baltazar JY, Muñoz-Perez MJ, Robledo-Vega C, Pérez-Zepeda MF, Soto-Vega E. Misinformation of COVID-19 on the internet: infodemiology study. JMIR Public Health Surveill 2020 Apr 9;6(2):e18444. [doi: 10.2196/18444] [Medline: 32250960]
16. Wang Z, Fan Y, Lv H, et al. The gap between self-rated health information literacy and internet health information-seeking ability for patients with chronic diseases in rural communities: cross-sectional study. J Med Internet Res 2022 Jan 31;24(1):e26308. [doi: 10.2196/26308] [Medline: 35099401]
17. Norman CD, Skinner HA. eHEALS: the eHealth literacy scale. J Med Internet Res 2006 Nov 14;8(4):e27. [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]

XSL•FO
RenderX

18. Gulec H, Kvardova N, Smahel D. Adolescents' disease- and fitness-related online health information seeking behaviors: the roles of perceived trust in online health information, eHealth literacy, and parental factors. Comput Human Behav 2022 Sep;134:107318. [doi: 10.1016/j.chb.2022.107318]
19. Zhu X, Zheng T, Ding L, Zhang X. Exploring associations between eHealth literacy, cyberchondria, online health information seeking and sleep quality among university students: a cross-section study. Heliyon 2023 Jun;9(6):e17521. [doi: 10.1016/j.heliyon.2023.e17521] [Medline: 37408886]
20. Kim K, Shin S, Kim S, Lee E. The relation between eHealth literacy and health-related behaviors: systematic review and meta-analysis. J Med Internet Res 2023 Jan 30;25:e40778. [doi: 10.2196/40778] [Medline: 36716080]
21. Tsukahara S, Yamaguchi S, Igarashi F, et al. Association of eHealth literacy with lifestyle behaviors in university students: questionnaire-based cross-sectional study. J Med Internet Res 2020 Jun 24;22(6):e18155. [doi: 10.2196/18155] [Medline: 32579126]
22. Xu RH, Zhou LM, Wong ELY, Wang D. The association between patients' eHealth literacy and satisfaction with shared decision-making and well-being: multicenter cross-sectional study. J Med Internet Res 2021 Sep 24;23(9):e26721. [doi: 10.2196/26721] [Medline: 34559062]
23. D. Wilson T. Human information behavior. Informing Sci J 2000;3:049-056. [doi: 10.28945/576]
24. Bettman JR. Consumer information acquisition and search strategies. In: Mitchell AA, editor. The Effects of Information on Consumer and Market Behaviour: American Marketing Association; 1978.
25. Wilson TD. Information behaviour: an interdisciplinary perspective. Inf Process Manag 1997 Jul;33(4):551-572. [doi: 10.1016/S0306-4573(97)00028-9]
26. Williamson K. Discovered by chance: the role of incidental information acquisition in an ecological model of information use. Libr Inf Sci Res 1998 Jan;20(1):23-40. [doi: 10.1016/S0740-8188(98)90004-4]
27. Rieh SY, Collins-Thompson K, Hansen P, Lee HJ. Towards searching as a learning process: a review of current perspectives and future directions. J Inf Sci 2016 Feb;42(1):19-34. [doi: 10.1177/0165551515615841]
28. Smith CL, Matteson ML. Information literacy in the age of machines that learn: desiderata for machines that teach. Libri 2018;68(2):71-84. [doi: 10.1515/libri-2017-0025]
29. Kuhlthau CC. Inside the search process: information seeking from the user's perspective. J Am Soc Inf Sci 1991 Jun;42(5):361-371. [doi: 10.1002/(SICI)1097-4571(199106)42:5<361::AID-ASI6>3.0.CO;2-%23]
30. Guendelman S, Broderick A, Mlo H, Gemmill A, Lindeman D. Listening to communities: mixed-method study of the engagement of disadvantaged mothers and pregnant women with digital health technologies. J Med Internet Res 2017 Jul 5;19(7):e240. [doi: 10.2196/jmir.7736] [Medline: 28679489]
31. Hua Z, Yuqing S, Qianwen L, Hong C. Factors influencing ehealth literacy worldwide: systematic review and meta-analysis. J Med Internet Res 2025 Mar 10;27:e50313. [doi: 10.2196/50313] [Medline: 40063939]
32. Luo A, Yu Z, Liu F, Xie W. The chain mediating effect of the public's online health information-seeking behavior on doctor-patient interaction. Front Public Health 2022;10:874495. [doi: 10.3389/fpubh.2022.874495] [Medline: 35719674]
33. IDF diabetes atlas 11th edition. : International Diabetes Federation; 2025 URL: https://diabetesatlas.org [accessed 2026-04-14]
34. Wang L, Li X, Wang Z, et al. Trends in prevalence of diabetes and control of risk factors in diabetes among US adults, 1999-2018. JAMA 2021 Jun 25;326(8):1-13. [doi: 10.1001/jama.2021.9883] [Medline: 34170288]
35. Healthy living with diabetes. National Institute of Diabetes and Digestive and Kidney Diseases. 2023. URL: https://www.niddk.nih.gov/health-information/diabetes/overview/healthy-living-with-diabetes?dkrd=/health-information/diabetes/overview/diet-eating-physical-activity [accessed 2026-04-14]
36. Peek ME, Odoms-Young A, Quinn MT, Gorawara-Bhat R, Wilson SC, Chin MH. Race and shared decision-making: perspectives of African-Americans with diabetes. Soc Sci Med 2010 Jul;71(1):1-9. [doi: 10.1016/j.socscimed.2010.03.014] [Medline: 20409625]
37. Serrano V, Rodriguez-Gutierrez R, Hargraves I, Gionfriddo MR, Tamhane S, Montori VM. Shared decision-making in the care of individuals with diabetes. Diabet Med 2016 Jun;33(6):742-751. [doi: 10.1111/dme.13143] [Medline: 27105298]
38. Shortus T, Kemp L, McKenzie S, Harris M. "Managing patient involvement": provider perspectives on diabetes decision-making. Health Expect 2013 Jun;16(2):189-198. [doi: 10.1111/j.1369-7625.2011.00700.x] [Medline: 21645187]
39. Bussey LG, Sillence E. The role of internet resources in health decision-making: a qualitative study. Digit Health 2019;5:2055207619888073. [doi: 10.1177/2055207619888073] [Medline: 31741741]
40. Chen YY, Li CM, Liang JC, Tsai CC. Health information obtained from the internet and changes in medical decision making: questionnaire development and cross-sectional survey. J Med Internet Res 2018 Feb 12;20(2):e47. [doi: 10.2196/jmir.9370] [Medline: 29434017]
41. Dean CA, Geneus CJ, Rice S, et al. Assessing the significance of health information seeking in chronic condition management. Patient Educ Couns 2017 Aug;100(8):1519-1526. [doi: 10.1016/j.pec.2017.03.012] [Medline: 28320559]
42. Eke R, Jannat KT, Yang XT, Parton JM. Disparity and factors associated with internet health information seeking among us adults living with diabetes mellitus: cross-sectional study. J Med Internet Res 2022 May 30;24(5):e32723. [doi: 10.2196/32723] [Medline: 35635741]

XSL•FO
RenderX

43. Watson KB, Wiltz JL, Nhim K, Kaufmann RB, Thomas CW, Greenlund KJ. Trends in multiple chronic conditions among US adults, by life stage, Behavioral Risk Factor Surveillance System, 2013-2023. Prev Chronic Dis 2025 Apr 17;22:E15. [doi: 10.5888/pcd22.240539] [Medline: 40245168]
44. Eysenbach G. Improving the quality of web surveys: the checklist for reporting results of internet e-surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
45. Kriston L, Scholl I, Hölzel L, Simon D, Loh A, Härter M. The 9-item Shared Decision Making Questionnaire (SDM-Q-9). development and psychometric properties in a primary care sample. Patient Educ Couns 2010 Jul;80(1):94-99. [doi: 10.1016/j.pec.2009.09.034] [Medline: 19879711]
46. Kelly S, Eldredge SA, Dalton ED, Miller LE. Health-information behavior: an initial validity portfolio for active and passive measures. Commun Res Rep 2014 Apr;31(2):171-182. [doi: 10.1080/08824096.2014.907145]
47. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2021. URL: https://www.R-project.org/ [accessed 2026-04-15]
48. Hair J, Alamer A. Partial Least Squares Structural Equation Modeling (PLS-SEM) in second language and education research: guidelines using an applied example. Res Methods Appl Linguist 2022 Dec;1(3):100027. [doi: 10.1016/j.rmal.2022.100027]
49. Chin WW. The partial least squares approach to structural equation modeling. In: Marcoulides GA, editor. Modern Methods for Business Research: Psychology Press; 1998:295-336 URL: https://www.taylorfrancis.com/chapters/edit/10.4324/9781410604385-10/partial-least-squares-approach-structural-equation-modeling-wynne-chin [accessed 2026-04-15]
50. Hair Jr JF, Hult GTM, Ringle CM, Sarstedt M. A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM), 2nd edition: Sage Publications Inc; 2017. URL: https://eli.johogo.com/Class/CCU/SEM/_A%20Primer%20on%20Partial%20Least%20Squares%20Structural%20Equation%20Modeling_Hair.pdf [accessed 2026-04-15]
51. Hassoun A, Beacock I, Consolvo S, Goldberg B, Kelley PG, Russell DM. Practicing information sensibility: how gen z engages with online information. Presented at: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems; Apr 23-28, 2023. [doi: 10.1145/3544548.3581328]
52. Luy M, Zannella M, Wegner-Siegmundt C, Minagawa Y, Lutz W, Caselli G. The impact of increasing education levels on rising life expectancy: a decomposition analysis for Italy, Denmark, and the USA. Genus 2019 Dec;75(1):11. [doi: 10.1186/s41118-019-0055-0]
53. Mahdavi H, Esmaily H. Impact of educational intervention by community pharmacists on asthma clinical outcomes, quality of life and medication adherence: a systematic review and meta-analysis. J Clin Pharm Ther 2021 Oct;46(5):1254-1262. [doi: 10.1111/jcpt.13419] [Medline: 33817821]
54. Raghupathi V, Raghupathi W. The influence of education on health: an empirical assessment of OECD countries for the period 1995-2015. Arch Public Health 2020;78(1):20. [doi: 10.1186/s13690-020-00402-5] [Medline: 32280462]
55. Wilson TD. Models in information behaviour research. J Doc 1999 Aug 1;55(3):249-270. [doi: 10.1108/EUM0000000007145]
56. Hughes TM, Merath K, Chen Q, et al. Association of shared decision-making on patient-reported health outcomes and healthcare utilization. Am J Surg 2018 Jul;216(1):7-12. [doi: 10.1016/j.amjsurg.2018.01.011] [Medline: 29395026]
57. Hornik R, Parvanta S, Mello S, Freres D, Kelly B, Schwartz JS. Effects of scanning (routine health information exposure) on cancer screening and prevention behaviors in the general population. J Health Commun 2013;18(12):1422-1435. [doi: 10.1080/10810730.2013.798381] [Medline: 24083417]
58. Zhang L, Ye JF, Zhao X. “I saw it incidentally but frequently”: exploring the effects of online health information scanning on lung cancer screening behaviors among Chinese smokers. Health Commun 2025 Feb;40(2):345-356. [doi: 10.1080/10410236.2024.2345948] [Medline: 38683113]
59. Shahin S, Saldaña M, Gil de Zúñiga H. Peripheral elaboration model: the impact of incidental news exposure on political participation. J Inf Technol Polit 2021 Apr 3;18(2):148-163. [doi: 10.1080/19331681.2020.1832012]
60. Education access and quality. Healthy People 2030. URL: https://odphp.health.gov/healthypeople/objectives-and-data/browse-objectives/education-access-and-quality [accessed 2026-04-15]
61. Health literacy in healthy people 2030. Healthy People 2030. URL: https://odphp.health.gov/healthypeople/priority-areas/health-literacy-healthy-people-2030 [accessed 2026-04-15]
62. Bevendorff J, Wiegmann M, Potthast M, Stein B. Is google getting worse? a longitudinal investigation of SEO spam in search engines. Presented at: 46th European Conference on Information Retrieval, ECIR 2024; Mar 24-28, 2024. [doi: 10.1007/978-3-031-56063-7_4]
63. Liu Y, Kornfield R, Yang EF, Burnside E, Keevil J, Shah DV. Patient-provider communication while using a clinical decision support tool: explaining satisfaction with shared decision making for mammography screening. BMC Med Inform Decis Mak 2022 Dec 7;22(1):323. [doi: 10.1186/s12911-022-02058-3] [Medline: 36476612]
64. Pel-Littel RE, Snaterse M, Teppich NM, et al. Barriers and facilitators for shared decision making in older patients with multiple chronic conditions: a systematic review. BMC Geriatr 2021 Feb 6;21(1):112. [doi: 10.1186/s12877-021-02050-y] [Medline: 33549059]
65. Gopal DP, Chetty U, O’Donnell P, Gajria C, Blackadder-Weinstein J. Implicit bias in healthcare: clinical practice, research and decision making. Future Healthc J 2021 Mar;8(1):40-48. [doi: 10.7861/fhj.2020-0233] [Medline: 33791459]

XSL•FO
RenderX

66. Paredes AZ, Idrees JJ, Beal EW, et al. Influence of English proficiency on patient-provider communication and shared decision-making. Surgery 2018 Jun;163(6):1220-1225. [doi: 10.1016/j.surg.2018.01.012] [Medline: 29482884]
67. Chen X, Lin A, Webber S. Designing artificial serendipity. Presented at: 26th International Conference on Human-Computer Interaction (HCII); Jun 29 to Jul 4, 2024. [doi: 10.1007/978-3-031-60405-8_3]
68. Race TM, Makri S. Introducing serendipity. In: Accidental Information Discovery: Cultivating Serendipity in the Digital Age: Elsevier; 2016:1-13. [doi: 10.1016/B978-1-84334-750-7.00001-7]
69. Saurwein F, Spencer-Smith C. Automated trouble: the role of algorithmic selection in harms on social media platforms. Media Commun 2021;9(4):222-233. [doi: 10.17645/mac.v9i4.4062]
70. Willson M. Algorithms (and the) everyday. In: The Social Power of Algorithms: Routledge; 2019:137-150. [doi: 10.4324/9781351200677-9]

Abbreviations

AVE: average variance extracted
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
PLS SEM: partial least squares structural equation modeling

Edited by A Stone; submitted 19.Oct.2025; peer-reviewed by D Bastola, S Song; revised version received 25.Mar.2026; accepted 27.Mar.2026; published 30.Apr.2026.

Please cite as:
Ming Y, Matteson ML, Zhang Y
Relationship Between Online Health Information Acquisition and Shared Decision-Making Among Patients With Diabetes: Cross-Sectional Survey Study
J Med Internet Res 2026;28:e86137
URL: https://www.jmir.org/2026/1/e86137
doi:10.2196/86137

© Yue Ming, Miriam L Matteson, Yao Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Association of Electronic Health Literacy With Self-Care and Health Outcomes Among Patients With Type 2 Diabetes Mellitus: Cross-Sectional Study

Phoenix Kit-Han Mo[1], PhD; Alice PS Kong[2], MD; Luyao Xie[1], PhD; Virginia WY Chan[1,3], PhD; Joseph TF Lau[4,5], PhD

[1]Centre of Health Behaviours Research, Jockey Club School of Public Health and Primary Care, Prince of Wales Hospital, Chinese University of Hong Kong, 202D, Shatin, N.T., Hong Kong, China

[2]Department of Medicine and Therapeutics, Chinese University of Hong Kong, Hong Kong, China

[3]Community Mental Health Association, Hong Kong, China

[4]Public Mental Health Center, School of Mental Health, Wenzhou Medical University, Wenzhou, China

[5]Zhejiang Provincial Clinical Research Center for Mental Disorders, The Affiliated Wenzhou Kangning Hospital, Wenzhou Medical University, Wenzhou, China

Corresponding Author:
Phoenix Kit-Han Mo, PhD
Centre of Health Behaviours Research, Jockey Club School of Public Health and Primary Care, Prince of Wales Hospital, Chinese University of Hong Kong, 202D, Shatin, N.T., Hong Kong, China

Abstract

Background: Diabetes mellitus (DM) continues to be a critical public health issue in Hong Kong. Although self-care behaviors help promote health among patients with DM, adherence remains suboptimal. More attention should be paid to eHealth literacy with the development of modern technologies.

Objective: This study aims to assess the level of eHealth literacy among patients with DM and examine its association with self-care and health outcomes.

Methods: A cross-sectional study was conducted among patients with type 2 DM from the DM clinic of a public hospital in Hong Kong. Data on eHealth literacy, self-care, self-care self-efficacy, diabetes distress, glycated hemoglobin (HbA_{1c}) control, and sociodemographic information were collected. Multivariable regression analyses were performed, adjusting for relevant sociodemographic and medical variables.

Results: Among the 427 patients with DM recruited, around two-thirds (65.1%) were classified as having a high level of eHealth literacy. Compared to those with lower eHealth literacy, participants with higher eHealth literacy demonstrated significantly higher levels of self-care ($P<.001$) and self-care self-efficacy ($P<.001$) and lower levels of diabetes distress ($P<.001$). Higher eHealth literacy was also associated with greater odds of achieving ideal HbA_{1c} control (<7%) in unadjusted analyses (odds ratio 1.90, 95% CI 1.15-2.81); however, this association was not statistically significant after adjustment for sociodemographic and medical covariates (adjusted odds ratio 1.57, 95% CI 0.99-2.52; $P=.07$).

Conclusions: This study evaluated eHealth literacy levels among patients with DM and examined the associations between eHealth literacy and health outcomes (eg, self-care, self-care self-efficacy, diabetes distress, and HbA_{1c} control). Assessing eHealth literacy in patients with DM could be useful in identifying those who are vulnerable to poorer health outcomes. Promoting eHealth literacy among patients with DM may be important.

(J Med Internet Res 2026;28:e77856) doi:10.2196/77856

KEYWORDS

eHealth literacy; diabetes mellitus; patients; health outcomes; self-care

Introduction

Globally, there were 537 million people living with diabetes mellitus (DM) in 2021 [1]. In Hong Kong, the prevalence of DM has increased sharply by 35.1% since 1995 [2]. By 2021, an estimated 7.8% of adults in Hong Kong were affected by DM [1]. It is projected that the prevalence of DM will rise to 9.4% in 2030. DM poses a significant burden to the health care system due to its chronic nature.

There is no known cure for DM, but it is highly controllable. Self-care disease management, including having a healthy diet, exercising, medication adherence, and regular glucose

XSL•FO
RenderX

monitoring, is an essential element of diabetes care [3]. Self-care behaviors are strongly associated with positive health outcomes, including lowering blood glucose level, reducing insulin resistance, and reducing the risk of cardiovascular disease [3]; however, adherence to self-care remains suboptimal. A study in Hong Kong among older adults with DM reported that 41.4% had unsatisfactory metabolic control (ie, glycated hemoglobin [HbA_{1c}] exceeding the threshold of 7%) [4]. It has also been suggested that patients with DM require support and health information to help with disease management, but these needs have not always been met [5-7].

Online platforms are, and will continue to be, an important resource for obtaining health information, as patients with DM are increasingly playing a larger role in managing their own health [8-10]. The widespread use of digital technologies has facilitated access to information previously available primarily through health professionals [11]. Digital health resources can help individuals with chronic diseases share their experiences and gather information and may help them overcome traditional barriers to self-care [12,13]. Moreover, easy access to health resources provided by digital approaches can increase individuals' knowledge and awareness regarding medical and health issues, thus empowering them to actively participate in their own health care by making informed and well-considered decisions [14].

Health literacy is defined as "the degree to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions" [15]. With the advent of digital technologies, the concept of eHealth literacy (also known as digital health literacy) was introduced to extend health literacy in the digital context [16]. The importance of eHealth literacy for patients to properly use online health information and to participate in informed health care decision-making has been increasingly emphasized [17]. Due to the complexity and amount of information available online, it can be challenging for individuals with little experience in searching for online health information to locate the relevant information. In most cases, information retrieved from the internet is not properly understood by all groups of people [18]. There is also evidence that people have difficulty discerning between trustworthy and untrustworthy health information online [19,20]. Poor-quality, incorrect, and confusing health information online could also be harmful to the patients [21]. It is therefore highly important that individuals are able to effectively process and understand medical information available on the internet.

Increasing attention to eHealth literacy has been observed in recent years. For example, studies in various populations, such as older people [22,23], underserved populations [18], or patients with chronic diseases [17,24,25], have suggested that there is, in general, a low level of eHealth literacy among the study samples. Various sociodemographic factors have been shown to affect the prevalence and extent of online health information seeking, including age, gender, education, socioeconomic status, health status, and internet use [26-31]. Research has further shown that higher eHealth literacy is associated with better health outcomes across different populations, such as better subjective health status [32], health-related quality of life [33-35], lower risk of chronic disease [36], healthy behaviors [37,38], and lower psychological distress [39,40].

Several theoretical models, such as the Transactional Model of eHealth Literacy [41] and the Integrative Model of eHealth Use [42], provide frameworks for understanding how eHealth literacy may influence health outcomes. These models suggest that eHealth literacy can directly or indirectly affect health knowledge, patient–provider communication, as well as health outcomes such as proactive health behaviors and health-related quality of life [41,42]. Among patients with diabetes, higher eHealth literacy has been related to better self-care and other health outcomes through multiple psychosocial and behavioral pathways. For example, a study of 453 adults with diabetes in South Korea found that eHealth literacy was positively linked to self-management via self-efficacy and social support [43]. In a Chinese study, eHealth literacy influenced self-care behaviors indirectly through diabetes knowledge and self-efficacy [44]. Additional studies in South Korea and Taiwan also reported that eHealth literacy was associated with health-promoting behaviors and directly influenced self-care behaviors [45,46]. While previous studies have shown that eHealth literacy is associated with health-related behaviors in diabetes, evidence from patients in Hong Kong remains limited.

Given the substantial health needs of individuals with DM, examining their ability to effectively use high-quality digital health resources and make informed decisions to improve their health would be particularly useful. The purposes of this study were to assess the level of eHealth literacy among patients with DM and to examine the association between eHealth literacy and behavioral (self-care), psychosocial (self-care self-efficacy and diabetes distress), and clinical (HbA_{1c} control) outcomes among patients with type 2 DM in Hong Kong.

Methods

Study Design and Participants

A cross-sectional study was conducted among patients with DM. Target participants were patients with type 2 DM in Hong Kong who fulfilled the following criteria: (1) diagnosed with type 2 DM; (2) age between 35 and 65 years; (3) ability to speak and understand Cantonese (the primary language used in Hong Kong). Participants who were receiving psychiatric care at the time of the study were excluded.

Sample Size Calculation

Based on a previous study by Kim et al [45] conducted among patients with type 2 diabetes, which reported a correlation of r=0.15 between eHealth literacy and health-promoting behaviors. The sample size was estimated using G*Power version 3.1, assuming a correlation between eHealth literacy and self-care (r=0.15), with a 2-sided significance level of α=.05, and 80% statistical power, the minimum required sample size was 346 participants.

Data Collection

Eligible patients with type 2 DM attending a diabetes management clinic at a public hospital in Hong Kong were

XSL•FO
RenderX

invited to participate in this cross-sectional survey using a clinic-based convenience sampling approach. For recruitment, informational materials (eg, posters and leaflets) were distributed to patients visiting the clinic. Trained research assistants approached eligible individuals who were waiting to receive routine follow-up consultations. Prospective participants received a detailed explanation of the study objectives and procedures. Those who expressed interest in joining the study received an information sheet, which contained detailed information about the purposes and logistics of the study. Written informed consent and contact information were obtained from participants who agreed to participate, and interview appointments were arranged on site. Telephone surveys were administered by trained research assistants within 1 week of the participants' clinic visit. Participants received a gift coupon valued at HKD 50 (~US $6.5) as a token of appreciation for the time they spent on the study.

Ethical Considerations

Ethical approval was obtained from the Joint Chinese University of Hong Kong–New Territories East Cluster Clinical Research Ethics Committee (reference number T6080016). Written informed consent was obtained from all participants prior to data collection. Participant information was deidentified to ensure privacy and confidentiality, and all data were handled in accordance with institutional ethical guidelines. No compensation was provided to participants.

Measures

eHealth Literacy

eHealth literacy was measured using the 8-item eHealth Literacy Scale (eHEALS), which was developed to measure individuals' perceived skills of internet use for health information [47]. Items are rated on a 5-point Likert scale ranging from "1=Strongly disagree" to "5=Strongly agree." Total scores of the eHEALS range from 8 to 40, with higher scores representing higher levels of eHealth literacy. The scale has been formally translated and validated in Chinese populations [48] and in the Hong Kong context [49-51]. Using the cutoff proposed by Milne et al [52], participants who scored 4 or above on at least 5 out of the 8 eHEALS items were classified as having high level of eHealth literacy. Internal consistency of the scale was satisfactory in the present study (Cronbach α=0.98), which is comparable to that reported in previous studies conducted among general adults (Cronbach α=0.95) [50] and older adults (Cronbach α=0.939) in Hong Kong [51].

Level of Self-Care

Self-care was measured by the 15-item Summary of Diabetes Self-Care Activity scale [53], a self-report questionnaire assessing diabetes self-management behaviors across domains including general diet, specific diet, foot care, self-monitoring of blood glucose, smoking, and exercise. Participants were asked to indicate the frequency (days per week) they engaged in each of these activities, with higher scores indicating higher levels of self-care. A validated Chinese version was used [54]. Internal consistency of the scale was satisfactory in the present study (Cronbach α=0.70).

Self-Care Self-Efficacy

Self-care self-efficacy was measured using the 20-item Chinese Diabetes Empowerment Scale [55]. Items were adapted and translated from the original 37-item Diabetes Empowerment Scale [56]. The Chinese Diabetes Empowerment Scale measures the degree of psychosocial self-efficacy in carrying out diabetes care in 5 domains: achieving goals, overcoming barriers, coping, suitable methods, and obtaining support. Participants were asked to rate their level of agreement on their perceived ability to handle diabetes-related psychosocial problems on a 5-point Likert scale, from "1=Strongly disagree" to "5=Strongly agree," with higher scores indicating a greater level of self-care self-efficacy. The internal consistency of the scale was satisfactory in the present study (Cronbach α=0.94).

Diabetes Distress

Diabetes-related distress was assessed using the 15-item Chinese Diabetes Distress Scale [57], which measures emotional burden distress, regimen- and social support-related distress, and physician-related distress. Participants were asked to rate the degree to which each item may have distressed or bothered them in the past month on a 6-point Likert Scale from "1=Not a problem" to "6=A very serious problem," with higher scores indicating a higher level of diabetes distress. The internal consistency of the scale was satisfactory in the present study (Cronbach α=0.93).

Sociodemographic Characteristics

Sociodemographic characteristics, including gender, age, education level, relationship status, employment status, monthly income, were obtained. Participants were also asked to report the average time spent on the internet per week. Medical information, such as duration of diagnosis, number of complications, types of treatment, and HbA_{1c} control, was retrieved from participants' medical records.

Data Analysis

Descriptive statistics were used to summarize participants' sociodemographic and medical characteristics and levels of eHealth literacy. Correlations between participants' sociodemographic and background characteristics, eHealth literacy, and outcome variables were also examined. To evaluate the association between eHealth literacy and outcome variables, participants were classified into high and low eHealth literacy groups based on the cutoff proposed by Milne et al [52]. Independent-sample *t* tests were used to examine the differences in self-care, self-care self-efficacy, and diabetes distress between groups. One-way analysis of covariance (ANCOVA) was also conducted with covariates being adjusted for in each analysis. Sociodemographic and medical variables that were significantly correlated with the respective outcome variables would be included as covariates. To evaluate the association between eHealth literacy and HbA_{1c} control, participants with HbA_{1c} less than 7% were first defined as having ideal HbA_{1c} control. Univariate logistic regression was conducted for obtaining odds ratio (OR) of eHealth literacy on health-related outcomes. Adjusted logistic regression, adjusting for significant sociodemographic and medical variables, was also conducted; adjusted OR for eHealth literacy was then derived. All analyses

XSL•FO
RenderX

were performed using SPSS version 21 (IBM Corp). Statistical significance was defined as $P<.05$.

Results

Participant Characteristics

A total of 427 patients with type 2 DM completed the survey. The achieved sample size was deemed sufficient for the planned analyses. Slightly more than half of the participants (223/427, 52.1%) were male, and their mean age was 56.53 (SD 7.05) years. About half of them (218/427, 51%) had received a senior secondary level of education or above. The majority of participants (364/427, 85.2%) were married, and more than half of them (240/427, 56.2%) were unemployed. The mean duration of DM diagnosis was 10.69 (SD 8.14) years. Slightly less than two-thirds (261/427, 61.2%) of them had at least one DM-related complication. One hundred ninety out of 427 (44.5%) and 185 out of 427 (43.3%) participants were on oral treatment and on both oral treatment and injections, respectively. Around two-thirds (278/427, 65.1%) of the participants were classified as having high levels of eHealth literacy (Table 1).

XSL•FO
RenderX

Table . Sociodemographic and medical characteristics of participants (N=427).

Characteristic	Sample
Sociodemographic characteristics	
Gender, n (%)	
Male	221 (52.1)
Female	203 (47.9)
Age (y), mean (SD)	56.53 (7.05)
Education level, n (%)	
Primary or below	113 (26.5)
Junior secondary (Year 7 to 9)	96 (22.5)
Senior secondary (Year 10 to 13)	166 (39)
College or university	51 (12)
Relationship status, n (%)	
Single	32 (7.5)
Married	364 (85.2)
Separated/divorced/widowed	29 (6.8)
Others	2 (0.5)
Employment, n (%)	
Employed	240 (56.2)
Housewife	66 (15.5)
Unemployed	15 (3.5)
Retired	102 (23.9)
Others	4 (0.9)
Monthly income (HKD)[a], n (%)	
≤9999	68 (16.5)
10,000-19,999	159 (38.7)
20,000-29,999	117 (28.5)
≥30,000	67 (16.3)
Medical characteristics	
Duration of diagnosis (y), mean (SD)	10.69 (8.14)
Number of complications, n (%)	
None	165 (38.8)
1	162 (38.1)
2	78 (18.4)
≥3	20 (4.7)
Types of treatment, n (%)	
No prescribed medication	21 (4.9)
Oral only	190 (44.5)
Injection only	27 (6.3)
Both oral and injection	185 (43.3)
Others	4 (0.9)
Internet-related variables	
Time spent on the internet per week, n (%)	
0 h	64 (15)

XSL•FO
RenderX

Characteristic	Sample
<1 h	178 (41.6)
Between 1 and 2 h	121 (28.3)
>2 h	64 (14.9)
eHEALS[b] score, mean (SD)	23.68 (11.61)
Level of eHealth literacy[c], n (%)	
Low	149 (34.9)
High	278 (65.1)

[a]1 HKD=0.13 USD.

[b]eHEALS: the 8-item eHealth literacy scale.

[c]A high level of eHealth literacy was defined as those who scored 4 or above with at least 5 eHEALS items.

Correlation Between Variables

The correlation between variables is presented in Table 2. Among sociodemographic and medical variables, male gender (r=−0.26, P<.01), younger age (r=−0.43, P<.01), higher educational level (r=0.50, P<.01), being married (r=−0.11, P<.01), being employed (r=−0.24, P<.01), higher monthly income (r=0.30, P<.01), and shorter duration of diagnosis (r=−0.14, P<.01) had significant correlation with higher level of eHealth literacy. Higher level of eHealth literacy was also significantly correlated with higher level of self-care (r=0.15, P<.01) and higher level of self-care self-efficacy (r=.32, P<.01), as well as lower level of diabetes distress (r=−0.11, P<.05) and lower level of HbA_{1c} (r=−0.11, P<.05).

XSL•FO
RenderX

Table . Correlation between variables.

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Gender[a]	—													
Age	0.12[b]	—												
Education level	−0.30[c]	−0.43[c]	—											
Relationship status[d]	0.04	0.18[c]	−0.13[c]	—										
Employment status[e]	0.17[c]	0.39[c]	−0.20[c]	0.06	—									
Monthly income	−0.10[b]	−0.29[c]	0.40[c]	−0.09	−0.27[c]	—								
Duration of diagnosis	0.13[b]	0.24[c]	−0.21[c]	0.02	0.16[c]	−0.06	—							
Number of complications	−0.10[b]	0.19[c]	−0.13[c]	0.07	0.05	−0.10[b]	0.24[c]	—						
Type of treatment	0.04	0.06	−0.06	−0.01	0.10[b]	−0.04	0.34[c]	0.12[b]	—					
eHealth literacy	−0.26[c]	−0.43[c]	0.50[c]	−0.11[b]	−0.24[c]	0.30[c]	−0.14[c]	−0.09	−0.06	—				
Self-care	0.08	0.01	0.05	−0.10[b]	0.13[c]	0.04	0.10[b]	−0.05	0.09	0.15[c]	—			
Self-care self-efficacy	−0.17[c]	−0.16[c]	0.25[c]	−0.04	−0.05	0.21[c]	−0.05	−0.05	−0.05	0.32[c]	0.30[c]	—		
Diabetes distress	0.07	−0.03	−0.05	0.01	−0.02	-0.20[c]	0.10[b]	0.03	0.02	−0.11[b]	−0.23[c]	−0.46[c]	—	
HbA_{1c}[f]	0.08	0.01	−0.11[b]	0.06	0.06	−0.07	0.34[c]	0.18[c]	0.21[c]	−0.11[b]	0.09	−0.10[b]	0.08	—

[a]Dummy variable was created as 1: male; 2: female.

[b]$P<.05$.

[c]$P<.01$.

[d]Dummy variable was created as 1: married; 2: single/separate/divorced/widowed.

[e]Dummy variable was created as 1: employed; 2: housewife/unemployed/retired.

[f]HbA_{1c}: glycated hemoglobin.

Group Difference in Self-Care, Self-Care Self-Efficacy, and Diabetes Distress

Results of independent-samples *t* tests showed that participants classified as having a high level of eHealth literacy had significantly higher levels of self-care ($P<.001$) and self-care self-efficacy ($P<.001$), as well as significantly lower levels of diabetes distress ($P<.001$).

Results of the ANCOVA showed that such differences remained statistically significant after controlling for significant sociodemographic and medical variables. The adjusted mean differences between the high and low eHealth literacy groups

XSL•FO
RenderX

were 0.442 for self-care, 3.875 for self-care self-efficacy, and −0.237 for diabetes distress (all *P*<.001) (Table 3).

Table . Independent sample *t* test and analysis of covariance analyses of group difference in self-care, self-care self-efficacy, and diabetes distress.

Variable	Low eHealth literacy (n=149), mean (SD)	High eHealth literacy (n=278), mean (SD)	Low eHealth literacy (n=149), adjusted mean (95% CI)	High eHealth literacy (n=278), adjusted mean (95% CI)	Mean difference
Self-care	1.78 (1.03)	2.07 (1.16)[a]	1.71 (1.56–1.87)[b]	2.16 (2.02–2.23)[b]	0.442[c]
Self-care self-efficacy	72.46 (7.87)	77.08 (4.51)[c]	73.37 (72.37 - 74.37)[d]	77.25 (76.45 - 78.05)[d]	3.875[c]
Diabetes distress	2.74 (0.56)	2.44 (0.47)[c]	2.68 (2.60 - 2.76)[e]	2.44 (2.38 - 2.50)[e]	−0.237[c]

[a]*P*<.01.

[b]Adjusted for relationship status, employment, duration of diagnosis, and types of treatment.

[c]*P*<.001.

[d]Adjusted for sex, age, education level, and monthly family income.

[e]Adjusted for monthly family income and duration of diagnosis.

Logistic Regression on Ideal HbA_{1c} Control

There were 38.7% (96/248) and 26.0% (38/146) ideal HbA_{1c} in participants with low level and high level of eHealth literacy, respectively. Results from the univariate logistic regression analysis showed that participants with a high level of eHealth literacy were more likely to report ideal HbA_{1c} control (unadjusted OR 1.90, 95% CI 1.15-2.81; *P*<.01). After adjusting for significant sociodemographic and medical information, the association was not statistically significant (adjusted OR 1.57, 95% CI 0.99-2.52, *P*=0.070).

Discussion

Main Findings and Implications

DM is a chronic disease that requires good knowledge and self-management skills to optimize glycemic control and health outcomes. As digital health technologies are now widely used to support patients, eHealth literacy is of great importance because seeking online health information reflects a complex issue involving information awareness, consciousness, reasoning, and knowledge. It is important to assess patients with DM eHealth literacy, as health care professionals should have a good understanding of the patients' ability to use electronic resources before recommending the use of online health information [47].

Few studies in Hong Kong have reported the eHealth literacy levels of the population, especially in patients. Several local studies reported mean eHEALS scores of 26.10 (SD 7.70) in 1501 general adults [50] and 27.34 (SD 6.44) among 306 adults aged 60 years or older [51]. This study expands the evidence base among patients in Hong Kong and shows that the mean eHEALS score among 427 patients with type 2 DM was 23.68 (SD 11.61). In addition, using the cutoff proposed by Milne et al [52], approximately two-thirds (278/427, 65.1%) of the patients with DM were considered as having a high level of eHealth literacy. Compared with other studies using the same cutoff, the proportion of patients having a high level of eHealth literacy in the present study was significantly higher than that reported among lung cancer survivors (33.7%) in Canada [52]. This difference may be partly attributable to age, as participants in the present study were younger (mean age 56.5, SD 7.05 y) than those in Milne et al's study (median age 71 y). This interpretation is further supported by the negative association between age and eHealth literacy observed in the present study and reported consistently in the literature [11,31]. The patients with DM in the present study were likely confident in using the internet for health information and advice; they also demonstrated reasonable skills and self-efficacy required to achieve maximum benefit from the use.

In addition to age, the present study also revealed that among the sociodemographic and medical variables, male gender, higher level of education, currently married, employed, higher monthly income, and shorter duration of diagnosis were associated with higher eHealth literacy. These findings are consistent with the literature showing that eHealth literacy is generally lower among individuals with lower socioeconomic status [14,58-60]. Because the eHEALS was designed to measure an individual's confidence in their ability to locate and access online health information, it is conceivable that individuals with better socioeconomic status would possess higher ability in understanding and discerning the quality of the health information available on the internet. Individuals who were married and employed may also be more likely to receive a higher level of social support in searching for health information, including those from the online sources.

Understanding the association between eHealth literacy and health-related outcomes is essential for developing effective health communication strategies in the future. It has been found that the use of online health information and eHealth literacy might affect health behaviors [37,45,61,62]. In the present study, higher eHealth literacy was significantly associated with higher levels of self-care and self-care self-efficacy and with lower levels of diabetes distress after controlling for significant sociodemographic and medical variables. These findings are consistent with recent evidence showing that eHealth literacy significantly influences diabetes self-care behaviors among patients with type 2 DM in Saudi Arabia [63]. Although this study also revealed eHealth literacy was significantly associated with better glycemic outcomes in their population [63], the present study did not find a statistically significant association between eHealth literacy and HbA_{1c} among patients with

XSL•FO
RenderX

diabetes in Hong Kong after adjustment for covariates. Patients with higher eHealth literacy are not only more inclined to use the internet for health-related information but also are able to understand, evaluate the information obtained, and use quality information to make informed decisions about health. They may therefore be more motivated to adopt healthier lifestyles based on the information that they obtain online [38], more directly leading to better behavioral and psychosocial outcomes. The integrated Model of eHealth Use model also supports the motivational pathways through which eHealth literacy may influence health behavior [42]. Individuals with a higher ability to understand online health information may also be more likely to access health care and interact with health care professionals, therefore more likely to benefit from their health care utilization. In the conceptual model proposed by Paasche-Orlov and Wolf [64], self-efficacy is considered an important mediator linking health literacy and health outcomes. In digital health contexts, self-efficacy has also been found to be a significant mediator in the relationship between eHealth literacy and health outcomes in patients with DM [43,44]. Individuals with higher eHealth literacy are likely to have greater confidence in their ability to effectively utilize online health resources and engage in self-care behaviors, which in turn, is associated with overall behavioral and mental health in patients with DM. However, HbA_{1c} is a more distal biomedical indicator and reflects average glycemic control over the preceding two to three months, which may be less directly associated with eHealth literacy. Future studies are warranted to further elucidate the pathways linking eHealth literacy to glycemic outcomes in this population. Findings of this study suggest that eHealth literacy is an important concept associated with the health of patients with DM. Improving eHealth literacy of patients with DM would potentially be an effective means in improving the health status of the population.

Implications for Practice

The eHEALS is currently one of the most common tools designed to measure an individual's confidence in their ability to locate and access online health information [65,66]. It demonstrated high internal consistency and predictive validity in Chinese chronic patients [67], and is associated with a range of health-related outcomes among various populations [22,30,67,68]. Findings of the present study further confirmed its utility in Chinese patients with DM and suggested that it could be an effective screening tool for screening patients' level of eHealth literacy in clinical settings.

The higher levels of self-care and self-care self-efficacy, and lower levels of diabetes distress documented among participants with higher eHealth literacy also provided important insights that promoting eHealth literacy can have the potential to support patients with DM with self-management behaviors and may result in better DM outcomes. Recent studies have shown that patients with DM who used DM apps had higher eHealth literacy than nonusers and perceived the app to have a greater effectiveness on their health behaviors [68-70]. Health care providers should therefore empower patients with DM with the skills in utilizing online resources. Improving eHealth literacy in patients with type 2 diabetes has important practical implications for self-care and health outcomes. Evidence suggests that improving diabetes knowledge and attitudes toward online health information are likely to be effective strategies for enhancing eHealth literacy in this population [71]. However, it is also noted that participant engagement in eHealth technologies might wane over time. Given the chronic nature of DM, interventions to increase eHealth literacy should make use of content and features that can be personalized and adapted to the individual over time to achieve a sustainable positive effect [72]. Continuous support, training, and monitoring by health care providers are essential to help patients acquire and apply accurate health information effectively, ultimately improving self-care and health outcomes.

Limitations

There are several limitations that should be noted. First, the present study was cross-sectional in nature; therefore, no causality between the variables could be assumed. Second, self-perceived eHealth literacy was measured in the present study as there was a lack of objective, validated measure in eHealth literacy in the current literature. The current literature seems to have revealed mixed findings on the association between subjective and objective measures of eHealth literacy on health-related internet use [25,36]. However, it is important to note that the eHEALS scale is one of the most used validated scales to measure eHealth literacy. This study was also limited by self-selection bias, that those who agreed to take part in the study might have more favorable views towards internet use and thus a higher level of eHealth literacy. Finally, as the present sample of patients with DM was relatively young and recruited from only one public DM clinic, cautions should be taken when generalizing the results to the general DM population in Hong Kong.

Conclusion

The internet plays an increasingly important part in disease management among patients with DM. Findings of the present study showed that higher level of eHealth literacy was associated with their better self-care, self-care self-efficacy, lower level of diabetes distress, and having ideal HbA_{1c} control. Assessing the level of eHealth literacy of patients with DM could be useful for identifying patients who are more vulnerable to poorer health. Promoting the eHealth literacy of patients with DM could potentially be associated with better health outcomes.

Funding

This work was supported by Health and Medical Research Fund of Hong Kong Food and Health Bureau (12130751).

Data Availability

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to their containing information that could compromise the privacy of research participants.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: PKHM, APSK, VWYC, JTFL
Data curation: PKHM, APSK, VWYC
Formal analysis: PKHM
Methodology: PKHM, APSK, VWYC, JTFL
Project administration: PKHM
Supervision: PKHM
Writing – original draft: PKHM, LX

Conflicts of Interest

None declared.

References

1. IDF Diabetes Atlas, 10th edition: International Diabetes Federation; 2021. URL: https://diabetes.org.br/wp-content/uploads/2022/02/IDF_Atlas_10th_Edition_2021-.pdf [accessed 2026-02-03]
2. Cheung B, Man Y, Lam K, Wat N, Lo J, Chau D, et al. Prevalence of diabetes mellitus (DM) in the Hong Kong Cardiovascular Risk Factor Prevalence Study cohort. Hong Kong Med J 2003;9(Supp 1) [FREE Full text]
3. American Diabetes Association. Standards of medical care in diabetes--2011. Diabetes Care 2011 Jan;34 Suppl 1(Suppl 1):S11-S61. [doi: 10.2337/dc11-S011] [Medline: 21193625]
4. Cheng LY, Leung SY, Leung MKW. The association of glycemic control and fall risk in diabetic elderly: a cross-sectional study in Hong Kong. BMC Prim Care 2022 Aug 1;23(1):192. [doi: 10.1186/s12875-022-01807-7] [Medline: 35915395]
5. Carolan M, Holman J, Ferrari M. Experiences of diabetes self-management: a focus group study among Australians with type 2 diabetes. J Clin Nurs 2015 Apr;24(7-8):1011-1023. [doi: 10.1111/jocn.12724] [Medline: 25363710]
6. Song Y, Song HJ, Han HR, Park SY, Nam S, Kim MT. Unmet needs for social support and effects on diabetes self-care activities in Korean Americans with type 2 diabetes. Diabetes Educ 2012;38(1):77-85. [doi: 10.1177/0145721711432456] [Medline: 22222514]
7. Timmers T, Janssen L, Kool RB, Kremer JA. Educating patients by providing timely information using smartphone and tablet apps: systematic review. J Med Internet Res 2020 Apr 13;22(4):e17342. [doi: 10.2196/17342] [Medline: 32281936]
8. Desveaux L, Shaw J, Saragosa M, et al. A mobile app to improve self-management of individuals with type 2 diabetes: qualitative realist evaluation. J Med Internet Res 2018 Mar 16;20(3):e81. [doi: 10.2196/jmir.8712] [Medline: 29549070]
9. Kruse CS, Beane A. Health information technology continues to show positive effect on medical outcomes: systematic review. J Med Internet Res 2018 Feb 5;20(2):e41. [doi: 10.2196/jmir.8793] [Medline: 29402759]
10. Shahbazi M, Bagherian H, Sattari M, Saghaeiannejad-Isfahani S. The opportunities and challenges of using mobile health in elderly self-care. J Educ Health Promot 2021;10:80. [doi: 10.4103/jehp.jehp_871_20] [Medline: 34084827]
11. Tennant B, Stellefson M, Dodd V, et al. eHealth literacy and Web 2.0 health information seeking behaviors among baby boomers and older adults. J Med Internet Res 2015 Mar 17;17(3):e70. [doi: 10.2196/jmir.3992] [Medline: 25783036]
12. Perakslis E, Ginsburg GS. Digital health—the need to assess benefits, risks, and value. JAMA 2021 Jan 12;325(2):127-128. [doi: 10.1001/jama.2020.22919] [Medline: 33369627]
13. McBride D. Web forums allow patients with colorectal cancer to share their real-world experiences. ONS Connect 2013 Dec;28(4):46. [Medline: 24400584]
14. Yoon H, Jang Y, Vaughan PW, Garcia M. Older adults' internet use for health information: digital divide by race/ethnicity and socioeconomic status. J Appl Gerontol 2020 Jan;39(1):105-110. [doi: 10.1177/0733464818770772] [Medline: 29661052]
15. Parker RM, Ratzan SC, Lurie N. Health literacy: a policy challenge for advancing high-quality health care. Health Aff (Millwood) 2003;22(4):147-153. [doi: 10.1377/hlthaff.22.4.147] [Medline: 12889762]
16. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9. [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
17. Wu Y, Wen J, Wang X, et al. Associations between e-health literacy and chronic disease self-management in older Chinese patients with chronic non-communicable diseases: a mediation analysis. BMC Public Health 2022 Nov 29;22(1):2226. [doi: 10.1186/s12889-022-14695-4] [Medline: 36447176]
18. Chesser A, Burke A, Reyes J, Rohrberg T. Navigating the digital divide: a systematic review of eHealth literacy in underserved populations in the United States. Inform Health Soc Care 2016;41(1):1-19. [doi: 10.3109/17538157.2014.948171] [Medline: 25710808]
19. Nguyen A, Mosadeghi S, Almario CV. Persistent digital divide in access to and use of the Internet as a resource for health information: results from a California population-based study. Int J Med Inform 2017 Jul;103:49-54. [doi: 10.1016/j.ijmedinf.2017.04.008] [Medline: 28551001]
20. Tonsaker T, Bartlett G, Trpkov C. Health information on the Internet: gold mine or minefield? Can Fam Physician 2014 May;60(5):407-408. [Medline: 24828994]

XSL•FO
RenderX

21. Valizadeh-Haghi S, Rahmatizadeh S. Evaluation of the quality and accessibility of available websites on kidney transplantation. Urol J 2018 Sep 26;15(5):261-265. [doi: 10.22037/uj.v0i0.4252] [Medline: 30058062]
22. Xie L, Zhang S, Xin M, Zhu M, Lu W, Mo PKH. Electronic health literacy and health-related outcomes among older adults: a systematic review. Prev Med 2022 Apr;157:106997. [doi: 10.1016/j.ypmed.2022.106997] [Medline: 35189203]
23. Wang Y, Song Y, Zhu Y, Ji H, Wang A. Association of eHealth literacy with health promotion behaviors of community-dwelling older people: the chain mediating role of self-efficacy and self-care ability. Int J Environ Res Public Health 2022 May 17;19(10):6092. [doi: 10.3390/ijerph19106092] [Medline: 35627627]
24. Robinson C, Graham J. Perceived Internet health literacy of HIV-positive people through the provision of a computer and Internet health education intervention. Health Info Libr J 2010 Dec;27(4):295-303. [doi: 10.1111/j.1471-1842.2010.00898.x] [Medline: 21050372]
25. van der Vaart R, Drossaert CHC, de Heus M, Taal E, van de Laar MAFJ. Measuring actual eHealth literacy among patients with rheumatic diseases: a qualitative analysis of problems encountered using Health 1.0 and Health 2.0 applications. J Med Internet Res 2013 Feb 11;15(2):e27. [doi: 10.2196/jmir.2428] [Medline: 23399720]
26. Oh YS, Cho Y. Examining the relationships between resources and online health information seeking among patients with chronic diseases and healthy people. Soc Work Health Care 2015;54(2):83-100. [doi: 10.1080/00981389.2014.987940] [Medline: 25674723]
27. Li J, Theng YL, Foo S. Predictors of online health information seeking behavior: changes between 2002 and 2012. Health Informatics J 2016 Dec;22(4):804-814. [doi: 10.1177/1460458215595851] [Medline: 26261218]
28. Moreland J, French TL, Cumming GP. The prevalence of online health information seeking among patients in Scotland: a cross-sectional exploratory study. JMIR Res Protoc 2015 Jul 15;4(3):e85. [doi: 10.2196/resprot.4010] [Medline: 26177562]
29. Li N, Orrange S, Kravitz RL, Bell RA. Reasons for and predictors of patients' online health information seeking following a medical appointment. Fam Pract 2014 Oct;31(5):550-556. [doi: 10.1093/fampra/cmu034] [Medline: 24963151]
30. Bailey CE, Kohler WJ, Makary C, Davis K, Sweet N, Carr M. eHealth literacy in otolaryngology patients. Ann Otol Rhinol Laryngol 2019 Nov;128(11):1013-1018. [doi: 10.1177/0003489419856377] [Medline: 31195809]
31. Halwas N, Griebel L, Huebner J. eHealth literacy, Internet and eHealth service usage: a survey among cancer patients and their relatives. J Cancer Res Clin Oncol 2017 Nov;143(11):2291-2299. [doi: 10.1007/s00432-017-2475-6] [Medline: 28699035]
32. Kim H, Yang E, Ryu H, Kim HJ, Jang SJ, Chang SJ. Psychometric comparisons of measures of eHealth literacy using a sample of Korean older adults. Int J Older People Nurs 2021 May;16(3):e12369. [doi: 10.1111/opn.12369] [Medline: 33527701]
33. Li S, Cui G, Yin Y, Wang S, Liu X, Chen L. Health-promoting behaviors mediate the relationship between eHealth literacy and health-related quality of life among Chinese older adults: a cross-sectional study. Qual Life Res 2021 Aug;30(8):2235-2243. [doi: 10.1007/s11136-021-02797-2] [Medline: 33661455]
34. Brørs G, Wentzel-Larsen T, Dalen H, et al. Psychometric properties of the Norwegian version of the electronic Health Literacy Scale (eHEALS) among patients after percutaneous coronary intervention: cross-sectional validation study. J Med Internet Res 2020 Jul 28;22(7):e17312. [doi: 10.2196/17312] [Medline: 32720900]
35. Xie L, Mo PKH. A 3-wave longitudinal study of eHealth literacy and older people's health-related quality of life in China: the mediating role of general self-efficacy. J Am Med Dir Assoc 2024 Dec;25(12):105310. [doi: 10.1016/j.jamda.2024.105310] [Medline: 39426796]
36. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. J Med Internet Res 2012 Jan 27;14(1):e19. [doi: 10.2196/jmir.1619] [Medline: 22357448]
37. Kim K, Shin S, Kim S, Lee E. The relation between eHealth literacy and health-related behaviors: systematic review and meta-analysis. J Med Internet Res 2023 Jan 30;25:e40778. [doi: 10.2196/40778] [Medline: 36716080]
38. Xie L, Mo PKH. Prospective relationship between electronic health literacy and health-promoting lifestyle among Chinese older adults: a three-wave longitudinal study. Soc Sci Med 2025 Jan;364:117166. [doi: 10.1016/j.socscimed.2024.117166] [Medline: 39586136]
39. Lin CY, Broström A, Griffiths MD, Pakpour AH. Psychometric evaluation of the Persian eHealth Literacy Scale (eHEALS) among elder Iranians with heart failure. Eval Health Prof 2020 Dec;43(4):222-229. [doi: 10.1177/0163278719827997] [Medline: 30744419]
40. Leung AYM, Parial LL, Tolabing MC, et al. Sense of coherence mediates the relationship between digital health literacy and anxiety about the future in aging population during the COVID-19 pandemic: a path analysis. Aging Ment Health 2022 Mar;26(3):544-553. [doi: 10.1080/13607863.2020.1870206] [Medline: 33438448]
41. Paige SR, Stellefson M, Krieger JL, Anderson-Lewis C, Cheong J, Stopka C. Proposing a transactional model of eHealth literacy: concept analysis. J Med Internet Res 2018 Oct 2;20(10):e10175. [doi: 10.2196/10175] [Medline: 30279155]
42. Bodie GD, Dutta MJ. Understanding health literacy for strategic health marketing: eHealth literacy, health disparities, and the digital divide. Health Mark Q 2008;25(1-2):175-203. [doi: 10.1080/07359680802126301] [Medline: 18935884]
43. Lee EH, Lee YW, Kang EH, Kang HJ. Relationship between electronic health literacy and self-management in people with type 2 diabetes using a structural equation modeling approach. J Nurs Res 2024 Jan 1;32(1):e315. [doi: 10.1097/jnr.0000000000000588] [Medline: 38128065]

XSL•FO
RenderX

44. Ji X, Chi J. Exploring the relationship between eHealth literacy and diabetes knowledge, self-efficacy, and self-care behaviors in Chinese diabetic patients: a cross-sectional study. J Nurs Res 2024 Dec 1;32(6):e359. [doi: 10.1097/jnr.0000000000000642] [Medline: 39593226]
45. Kim KA, Kim YJ, Choi M. Association of electronic health literacy with health-promoting behaviors in patients with type 2 diabetes: a cross-sectional study. Comput Inform Nurs 2018 Sep;36(9):438-447. [doi: 10.1097/CIN.0000000000000438] [Medline: 29742548]
46. Guo SHM, Hsing HC, Lin JL, Lee CC. Relationships between mobile eHealth literacy, diabetes self-care, and glycemic outcomes in Taiwanese patients with type 2 diabetes: cross-sectional study. JMIR mHealth uHealth 2021 Feb 5;9(2):e18404. [doi: 10.2196/18404] [Medline: 33544088]
47. Norman CD, Skinner HA. eHEALS: the eHealth literacy scale. J Med Internet Res 2006 Nov 14;8(4):e27. [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
48. Xu RH, Zhou L, Lu SY, Wong EL, Chang J, Wang D. Psychometric validation and cultural adaptation of the simplified Chinese eHealth literacy scale: cross-sectional study. J Med Internet Res 2020 Dec 7;22(12):e18613. [doi: 10.2196/18613] [Medline: 33284123]
49. Wong DKK, Cheung MK. Online health information seeking and eHealth literacy among patients attending a primary care clinic in Hong Kong: a cross-sectional survey. J Med Internet Res 2019 Mar 27;21(3):e10831. [doi: 10.2196/10831] [Medline: 30916666]
50. Guo Z, Zhao SZ, Guo N, et al. Socioeconomic disparities in eHealth literacy and preventive behaviors during the COVID-19 pandemic in Hong Kong: cross-sectional study. J Med Internet Res 2021 Apr 14;23(4):e24577. [doi: 10.2196/24577] [Medline: 33784240]
51. Kim S, Chow BC, Park S, Liu H. The usage of digital health technology among older adults in Hong Kong and the role of technology readiness and eHealth literacy: path analysis. J Med Internet Res 2023 Apr 12;25(1):e41915. [doi: 10.2196/41915] [Medline: 37043274]
52. Milne RA, Puts MTE, Papadakos J, et al. Predictors of high eHealth literacy in primary lung cancer survivors. J Cancer Educ 2015 Dec;30(4):685-692. [doi: 10.1007/s13187-014-0744-5] [Medline: 25355524]
53. Toobert DJ, Hampson SE, Glasgow RE. The summary of diabetes self-care activities measure: results from 7 studies and a revised scale. Diabetes Care 2000 Jul;23(7):943-950. [doi: 10.2337/diacare.23.7.943] [Medline: 10895844]
54. Tang YH, Pang SMC, Chan MF, Yeung GSP, Yeung VTF. Health literacy, complication awareness, and diabetic control in patients with type 2 diabetes mellitus. J Adv Nurs 2008 Apr;62(1):74-83. [doi: 10.1111/j.1365-2648.2007.04526.x] [Medline: 18352966]
55. Shiu ATY, Wong RYM, Thompson DR. Development of a reliable and valid Chinese version of the diabetes empowerment scale. Diabetes Care 2003 Oct;26(10):2817-2821. [doi: 10.2337/diacare.26.10.2817] [Medline: 14514585]
56. Anderson RM, Funnell MM, Fitzgerald JT, Marrero DG. The diabetes empowerment scale: a measure of psychosocial self-efficacy. Diabetes Care 2000 Jun;23(6):739-743. [doi: 10.2337/diacare.23.6.739] [Medline: 10840988]
57. Ting RZW, Nan H, Yu MWM, et al. Diabetes-related distress and physical and psychological health in chinese type 2 diabetic patients. Diabetes Care 2011 May;34(5):1094-1096. [doi: 10.2337/dc10-1612] [Medline: 21398526]
58. Verma R, Saldanha C, Ellis U, Sattar S, Haase KR. eHealth literacy among older adults living with cancer and their caregivers: a scoping review. J Geriatr Oncol 2022 Jun;13(5):555-562. [doi: 10.1016/j.jgo.2021.11.008] [Medline: 34810146]
59. Choi NG, Dinitto DM. The digital divide among low-income homebound older adults: internet use patterns, eHealth literacy, and attitudes toward computer/internet use. J Med Internet Res 2013 May 2;15(5):e93. [doi: 10.2196/jmir.2645] [Medline: 23639979]
60. Hung LY, Lyons JG, Wu CH. Health information technology use among older adults in the United States, 2009-2018. Curr Med Res Opin 2020 May;36(5):789-797. [doi: 10.1080/03007995.2020.1734782] [Medline: 32096650]
61. Li SJ, Yin YT, Cui GH, Xu HL. The associations among health-promoting lifestyle, eHealth literacy, and cognitive health in older Chinese adults: a cross-sectional study. Int J Environ Res Public Health 2020 Mar 27;17(7):2263. [doi: 10.3390/ijerph17072263] [Medline: 32230935]
62. Kim SH, Son YJ. Relationships between eHealth literacy and health behaviors in Korean adults. Comput Inform Nurs 2017 Feb;35(2):84-90. [doi: 10.1097/CIN.0000000000000255] [Medline: 27258808]
63. AlShehri WS, Khojah YY. Impact of e-health literacy on diabetes self-care activities among people with type 2 diabetes attending primary healthcare centers in Makkah City, Saudi Arabia: a cross-sectional study, 2025. Cureus 2025 Jul;17(7):e87968. [doi: 10.7759/cureus.87968] [Medline: 40666264]
64. Paasche-Orlow MK, Wolf MS. The causal pathways linking health literacy to health outcomes. Am J Health Behav 2007;31 Suppl 1:S19-S26. [doi: 10.5555/ajhb.2007.31.supp.S19] [Medline: 17931132]
65. Griebel L, Enwald H, Gilstad H, Pohl AL, Moreland J, Sedlmayr M. eHealth literacy research-quo vadis? Inform Health Soc Care 2018 Dec;43(4):427-442. [doi: 10.1080/17538157.2017.1364247] [Medline: 29045164]
66. Lee J, Lee EH, Chae D. eHealth literacy instruments: systematic review of measurement properties. J Med Internet Res 2021 Nov 15;23(11):e30644. [doi: 10.2196/30644] [Medline: 34779781]

XSL•FO
RenderX

67. Chang A, Schulz PJ. The measurements and an elaborated understanding of Chinese eHealth literacy (C-eHEALS) in chronic patients in China. Int J Environ Res Public Health 2018 Jul 23;15(7):1553. [doi: 10.3390/ijerph15071553] [Medline: 30041420]
68. Ernsting C, Stühmann LM, Dombrowski SU, Voigt-Antons JN, Kuhlmey A, Gellert P. Associations of health app use and perceived effectiveness in people with cardiovascular diseases and diabetes: population-based survey. JMIR mHealth uHealth 2019 Mar 28;7(3):e12179. [doi: 10.2196/12179] [Medline: 30920383]
69. Rollo ME, Aguiar EJ, Williams RL, et al. eHealth technologies to support nutrition and physical activity behaviors in diabetes self-management. Diabetes Metab Syndr Obes 2016;9:381-390. [doi: 10.2147/DMSO.S95247] [Medline: 27853384]
70. Kebede MM, Pischke CR. Popular diabetes apps and the impact of diabetes app use on self-care behaviour: a survey among the digital community of persons with diabetes on social media. Front Endocrinol (Lausanne) 2019;10:135. [doi: 10.3389/fendo.2019.00135] [Medline: 30881349]
71. Lee M, Shim J. Effects of diabetes knowledge and attitudes toward internet health information on e-health literacy in middle-aged patients with diabetes. Healthcare (Basel) 2025 Feb 27;13(5):512. [doi: 10.3390/healthcare13050512] [Medline: 40077074]
72. Pereira K, Phillips B, Johnson C, Vorderstrasse A. Internet delivered diabetes self-management education: a review. Diabetes Technol Ther 2015 Jan;17(1):55-63. [doi: 10.1089/dia.2014.0155] [Medline: 25238257]

Abbreviations

ANCOVA: analysis of covariance
DM: diabetes mellitus
eHEALS: the 8-item eHealth Literacy Scale
HbA_{1c}: glycated hemoglobin
OR: odds ratio

Edited by A Schwartz, M Balcarras; submitted 02.Jun.2025; peer-reviewed by CL Teng, CW Chan, MA Bujang; revised version received 30.Dec.2025; accepted 02.Jan.2026; published 09.Mar.2026.

Please cite as:
Mo PKH, Kong APS, Xie L, Chan VWY, Lau JTF
Association of Electronic Health Literacy With Self-Care and Health Outcomes Among Patients With Type 2 Diabetes Mellitus: Cross-Sectional Study
J Med Internet Res 2026;28:e77856
URL: https://www.jmir.org/2026/1/e77856
doi:10.2196/77856

© Phoenix Kit-Han Mo, Alice PS Kong, Luyao Xie, Virginia WY Chan, Joseph TF Lau. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Relationship Between Electronic Health Literacy and Health-Related Quality of Life Among Chinese Older Adults: Cross-Sectional Study

Yongqiang Wang[1*], MSc; Baozhen Dai[1,2*], PhD; Jiazhen Yao[1], MSc

[1]Health Policy and Healthcare Security Research Center, Southeast University, No. 87 Dingjiaqiao, Nanjing, Jiangsu, China

[2]Ageing-Responsive Civilization Think Tank, Nanjing, Jiangsu, China

[*]these authors contributed equally

Corresponding Author:
Baozhen Dai, PhD
Health Policy and Healthcare Security Research Center, Southeast University, No. 87 Dingjiaqiao, Nanjing, Jiangsu, China

Abstract

Background: The rapid digitalization of health care has reshaped access to medical services. However, older adults often remain disadvantaged due to the digital divide. Electronic health literacy (EHL) is increasingly recognized as a determinant of health-related quality of life (HRQoL); however, its mechanisms and subgroup differences in China remain underexplored.

Objective: This study aimed to examine the association between EHL and multidimensional HRQoL among Chinese older adults, with a focus on the mediating roles of attitudes toward own aging (ATOA) and self-efficacy (SE), and heterogeneity by age, residence, and lifestyle.

Methods: A cross-sectional survey (July-November 2024) included 8364 adults aged ≥55 years from 4 provinces using stratified multistage sampling. HRQoL was measured by physical health (PH), mental health (MH), and life satisfaction (LS). EHL was assessed with the eHealth Literacy Scale (eHEALS), ATOA with the Philadelphia Geriatric Center Morale Scale subscale, and SE with the General Self-Efficacy Scale. Analyses used seemingly unrelated regressions, PROCESS (Andrew F. Hayes) macro mediation with 5000 bootstraps, and subgroup regressions.

Results: EHL was positively associated with PH (β=0.273; *P*<.001), MH (β=0.190; *P*<.001), and LS (β=0.082; *P*<.001). ATOA and SE significantly mediated these associations (all *P*<.001). For PH, the association was partially mediated by ATOA and SE. The total effect was 0.273 (*P*<.001), with indirect effects accounting for 52.7% of the total. For MH, inconsistent mediation was observed. The direct effect was negative (β=–0.069; *P*<.01), but indirect effects were positive, yielding a positive total effect. For LS, the effect was fully mediated by ATOA and SE. Subgroup analyses showed stronger effects in the younger-old adults (PH: β=0.288, MH: β=0.188, LS: β=0.089; all *P*<.001), urban residents (PH: β=0.237, MH: β=0.214, LS: β=0.083; all *P*<.001), and nonexercisers (PH: β=0.322; *P*<.001).

Conclusions: EHL is strongly associated with HRQoL among Chinese older adults. Its effects on MH and LS operate primarily through psychosocial pathways, while PH benefits directly. The findings highlight EHL's compensatory role, particularly for the younger-old group, rural residents, and nonexercisers, underscoring its importance in digital inclusion and healthy aging policies.

(J Med Internet Res 2026;28:e84700) doi:10.2196/84700

KEYWORDS

electronic health literacy; health-related quality of life; older adult; attitudes toward own aging; self-efficacy

Introduction

Background

The accelerating pace of global population aging has made the pursuit of healthy aging a pressing challenge for both public health and social development [1]. Beard et al [2] argue that healthy aging is not merely about extending life expectancy but about maintaining functional ability, psychological well-being, and social participation across the entire lifespan, thereby improving the holistic welfare of older adults. Under this overall objective, health-related quality of life (HRQoL) has become a key comprehensive indicator for measuring its achievement. HRQoL refers to an individual's subjective assessment of their well-being and ability to perform physical, psychological, and social functions [3]. Considering the landscape of healthy aging, China is undergoing the world's largest and fastest demographic transition, where high prevalence of chronic diseases, functional decline, and social isolation intertwine, severely constraining the progress of healthy aging [4]. Relevant surveys also indicate that the HRQoL among China's older adults remains suboptimal [5,6]. Therefore, enhancing HRQoL is not only paramount to

XSL•FO
RenderX

achieving healthy aging but also an urgent imperative for addressing the challenges of China's aging society.

Concurrently, the digital transformation of health and medical service systems is reshaping the way care is delivered [7]. Electronic health services—such as online prescription, telemedicine, electronic health record, and health information platform—have become vital channels for older adults to obtain health resources. These services transcend traditional barriers of time and space, improving accessibility and the efficiency of health management [8]. Yet not all older adults benefit equally. Owing to the persistent digital divide, many face barriers in accessing, evaluating, and using health information, which may exacerbate health inequalities [9]. In this context, electronic health literacy (EHL) has increasingly been recognized as a critical determinant of HRQoL among older adults [10].

EHL refers to the ability to locate, understand, evaluate, and apply health information from electronic media, as well as the capacity to use acquired health information to address health issues [11]. Existing evidence has shown consistent associations between EHL and a range of health outcomes [12-14]. Older adults with higher levels of EHL are more likely to seek health information online, and in doing so, report better self-rated health, improved chronic disease management, and healthier behaviors [15]. EHL has also been associated with improved mental health (MH) and subjective well-being [14]. Nevertheless, important gaps remain, as existing research has focused on self-assessed health as a single measurement tool [16,17], failing to adopt a comprehensive perspective encompassing physical health (PH), MH, and subjective well-being. Furthermore, the mediating mechanisms underlying these associations and the heterogeneity among population subgroups have not been sufficiently explored [16,18].

In China, the "Internet Plus Healthcare" initiative has expanded digital health services nationwide [15]. However, the general level of EHL among older adults remains relatively low, with marked regional and subgroup disparities [19]. Although older adults should be the main beneficiaries of digital health innovations, they often remain a disadvantaged group in practice due to the digital divide [20]. Consequently, empirical investigation into the role of EHL in shaping HRQoL in China remains scarce, highlighting the need to investigate its mechanisms and contextual variations.

Theoretical Hypotheses

Since the World Health Organization declared 2021-2030 the "Decade of Healthy Aging," enhancing HRQoL has become the key objective for achieving healthy aging [21]. HRQoL encompasses not only the maintenance of physical functioning but also MH, life satisfaction (LS), and cognitive well-being—positive indicators that extend beyond morbidity or disease prevalence [22]. Within the digital era, the impact of internet use on the health of older adults requires positive mediation through human capital. As a critical form of human capital in the digital health field, EHL directly empowers older adults' health management practices by enhancing their capacity to discern health information and optimizing pathways to health services, thereby advancing healthy aging [23]. Similarly, EHL is recognized as both a prerequisite for inclusive health innovation and a super determinant of health equity [24]. Systematic reviews have also confirmed a significant positive correlation between EHL and older adults' health management efficacy [25]. Thus, EHL is not only an essential core health competency for older adults in the digital age but also, due to its enabling role in health management practices and its positive predictive power for health outcomes, emerges as a form of health capital closely linked to HRQoL.

The theory of successful aging emphasizes that older adults' subjective attitudes toward aging play a critical role in maintaining functional capacity [26,27]. EHL, as a resource for information empowerment and cognitive development, can effectively challenge and mitigate negative attitudes toward aging among older adults. This enables them to view aging as a stage where vitality, learning, and growth remain possible [28]. In turn, positive attitudes toward own aging (ATOA) contribute to the enhancement of older adults' HRQoL across multiple dimensions. Psychologically, positive ATOA can alleviate aging-related stress responses, strengthen psychological resilience, and reduce the risk of depression [29]. Behaviorally, older adults with positive attitudes are more likely to adopt and maintain health-promoting behaviors to proactively preserve their capabilities [30]. Physiologically, existing research has confirmed that positive self-perceptions of aging exert direct positive physiological effects on cardiovascular and immune functions by alleviating chronic stress [31]. Based on these findings, our study identifies ATOA as a mediating variable to explore the underlying mechanisms.

The conceptual model proposed by Paasche-Orlow and Wolf emphasizes that self-efficacy (SE) serves as a pivotal mediating factor linking health literacy to health outcomes [32]. SE refers to an individual's subjective judgment and intrinsic confidence in their own capabilities and personal worth within specific contexts [33]. In digital health settings, higher EHL empowers older adults to more effectively access, understand, evaluate, and apply digital health information and services. This process directly enhances their perceived capacity to address health issues. According to social cognitive theory, experiences of successfully mastering and applying relevant skills are key sources of SE. Thus, EHL strengthens older adults' confidence in managing their health by providing experiences of successfully using digital tools for disease prevention, health management, and self-care [34]. SE theory posits that an individual's belief in their ability to perform specific tasks is a core determinant of health behaviors [35,36]. Enhanced SE promotes a range of adaptive health behaviors and psychological processes. Older adults with high SE are more likely to adopt proactive health behaviors, adhere to treatment plans, better cope with disease-related stress, and maintain greater psychological resilience. These factors collectively contribute to improvements in both physical and MH, ultimately manifesting as enhanced HRQoL [37]. Thus, our study examines key mechanisms within the relationship by identifying SE as a mediating variable.

Furthermore, the relationship between EHL and HRQoL varies across contexts and individual characteristics. According to health inequality theory, health outcomes are shaped not only

XSL•FO
RenderX

by individual factors but also by social determinants [38]. Age reflects significant life course characteristics. Older adults often experience cumulative health decline, cognitive impairment, and increased barriers to digital access and use, making their EHL more likely constrained by physiological and cognitive foundations [39]. Younger-old adults typically possess relatively better physical and mental functioning, along with stronger learning abilities. They may more effectively translate EHL into concrete actions for health management and promotion, thereby demonstrating greater health benefits. Regarding residential environments, urban older adults generally benefit from more comprehensive health care systems, denser social support networks, wider internet and smart device coverage, and higher overall educational attainment and digital literacy [40]. These conditions collectively create an enabling environment conducive to accumulating and converting EHL. In contrast, rural older adults generally face multiple constraints, including low accessibility to medical resources, weak digital infrastructure, limited health information channels, and deep-rooted traditional views on aging, exacerbating health inequalities in the digital age [41]. From the perspective of individual adaptive strategies, lifestyle serves as another crucial moderating factor. According to the Selective Compensation Optimization model, individuals cope with resource changes during aging by selecting goals, optimizing resources, and compensating for losses [42]. For older adults with insufficient exercise, EHL may play a crucial compensatory role by promoting other health behaviors to partially offset health risks stemming from physical inactivity [43,44]. Conversely, for those with established exercise habits—whose health behavior systems are relatively well-developed—EHL's impact may manifest more as an optimization support. It can enhance the scientific rigor of exercise routines or integrate diverse health resources, yielding relatively stable incremental benefits.

Overall, existing research has yet to establish an integrated research framework. Most studies focus on single mechanisms such as health-promoting behaviors [45], insomnia and psychological distress [46], sense of coherence [47], and SE of health practices [48]. Meanwhile, existing studies lack exploration of the role of social structural factors and individual adaptive strategies as grouping variables [45,49]. Therefore, our study developed an integrated research framework through theoretical synthesis and conceptual framework construction. Therefore, our study integrates theories and constructs a framework encompassing digital empowerment, psychological-behavioral transformation, comprehensive health outcomes, and multilevel contextual moderation. This framework aims to systematically elucidate the mechanisms underlying the functioning of EHL and identify the key entry points for the delivery of targeted interventions. Based on this framework, the study proposes the following hypotheses:

- Hypothesis 1: EHL is positively associated with HRQoL.
- Hypothesis 2a: ATOA mediates the relationship between EHL and HRQoL.
- Hypothesis 2b: SE mediates the relationship between EHL and HRQoL.
- Hypothesis 3a: The association between EHL and HRQoL differs by age.
- Hypothesis 3b: The association between EHL and HRQoL differs by place of residence.
- Hypothesis 3c: The association between EHL and HRQoL differs by exercise behavior.

Methods

Study Design and Participants

In 2024, our research team independently conducted a special survey titled Digital Inclusion and Healthy Aging System Development. The questionnaire covered demographic information, digital health literacy, and multidimensional indicators of healthy aging. The inclusion criteria were (1) must be at least 55 years of age at the time of the survey, (2) must have resided continuously within the survey area for no less than 6 months, (3) must possess adequate cognitive and communication abilities, and (4) must voluntarily participate in this study and sign an informed consent form before the survey commencement. Exclusion criteria were (1) severe cognitive or psychiatric impairment preventing effective communication, (2) significant PH issues rendering completion of the survey impossible, (3) temporary residents, and (4) individuals who explicitly refused to sign the informed consent form or withdrew from the study midway.

To ensure high data quality and representativeness, a stratified multistage cluster sampling strategy was adopted, combining purposive and random sampling. In the first stage, based on the digital development index, internet penetration rate, and pilot programs of smart health care for older adults published by the National Bureau of Statistics and the Ministry of Industry and Information Technology, along with regional characteristics, population structure, economic development, and resource distribution, 4 representative provinces or autonomous regions were selected, namely, Jiangsu, Shandong, Hubei, and Guangxi. In the second stage, one city from each province was chosen according to the same stratification criteria, namely, Nanjing, Weifang, Enshi, and Nanning. In the third stage, all districts (or counties) were used as the sampling frame, and random sampling identified Gulou District, Xuanwu District, Qixia District, Jiangning District, Qingzhou County, Jianshi County, and Xixiangtang District. In the fourth stage, taking into account factors such as community population size, urban-rural type, and digital environment variation, a total of 57 communities (or villages) were randomly selected from these districts or counties as the final clusters. Eligible older adults within these units were surveyed. The structural characteristics of the final sample demonstrate that the sampling design achieved balance in economic level, urban-rural type, and digital access status.

Data collection was completed through surveyor-led face-to-face structured interviews. Fieldwork was conducted by a team of trained surveyors who had undergone a standardized training program covering survey objectives, questionnaire content, interview techniques, and ethical considerations.

Questionnaire Development and Pretesting

The questionnaire was initially developed in Chinese. To ensure content validity and comprehensibility for the target population, a pretest was conducted among a convenience sample of 200

XSL•FO
RenderX

older adults from nonsampled communities. Based on feedback regarding item clarity, response options, and interview duration, revisions were made to the wording and formatting.

Translation

As this study aims for potential international dissemination, the questionnaire was professionally translated into English by bilingual public health researchers following a forward-backward translation procedure to ensure conceptual equivalence. The English version was used for reporting purposes only.

Quality Control Procedures

Multiple quality control measures were implemented throughout the data collection process. (1) All surveyors were required to pass a mock interview assessment before fieldwork. (2) On-site supervisors randomly observed approximately 10% of the interviews and provided immediate feedback. (3) Surveyors used tablet computers equipped with a survey application to minimize data entry errors. (4) Supervisors reviewed completed questionnaires daily for completeness and consistency. (5) A random subset of participants was selected for telephone callback to verify key information.

Response Rate and Data Cleaning

A total of 8731 questionnaires were distributed to eligible individuals, resulting in a final valid response rate of 95.8%. Figure 1 illustrates the screening process for the study population.

Figure 1. Flowchart showing the selection of participants. EHL: electronic health literacy; HRQoL: health-related quality of life; LS: life satisfaction; MH: mental health; PH: physical health.

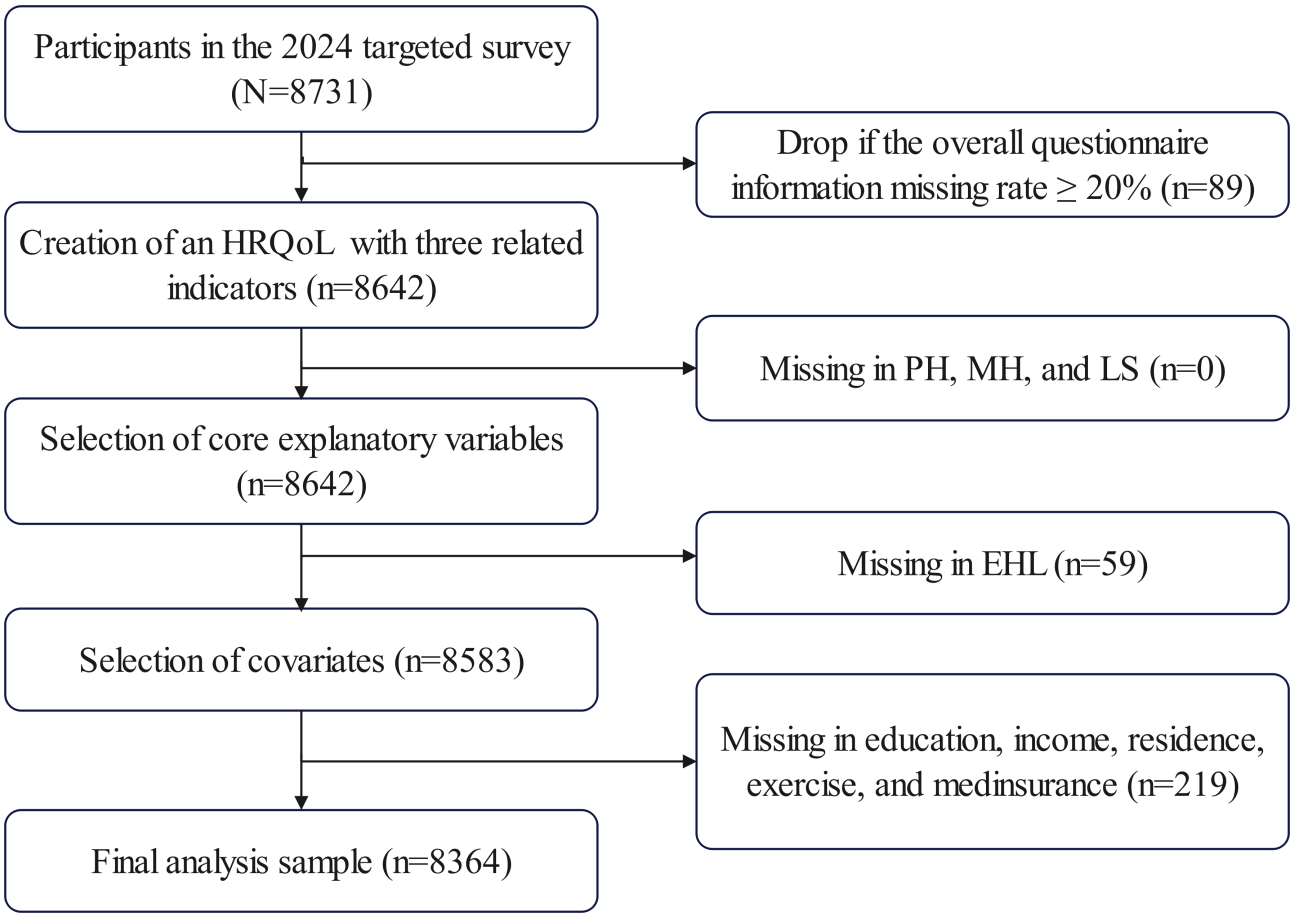

Measures

Assessment of HRQoL

The primary dependent variable was HRQoL. Quality of life is valued as a complement to conventional notions of health and functional status. An optimal health assessment should, therefore, encompass physical, social, and psychological functioning alongside quality of life [50]. Accordingly, this study assessed HRQoL using 3 indicators, namely, PH, MH, and LS.

PH was measured using the EuroQol Visual Analogue Scale (EQ-VAS), which is part of the EQ-5D instrument that evaluates individuals' overall perception of their health status. Respondents rated their current health status on a visual scoring ruler ranging from 0 ("worst conceivable health state") to 100 ("best conceivable health state") [51]. The EQ-VAS is widely recognized as a straightforward, subjective instrument designed to capture general health status in clinical and population-based health studies [52].

MH was evaluated using the World Health Organization-Five Well-Being Index (WHO-5), which assesses subjective well-being and risk of depression. The WHO-5 questionnaire includes 5 specific measurement items, each rated on a 6-point Likert scale ranging from 0 ("at no time") to 5 ("all of the

XSL•FO
RenderX

time”). The raw total score, which spans from 0 to 25, is conventionally converted to a percentage scale by multiplying by 4, resulting in a range of 0 to 100. Higher values on this scale reflect a more favorable MH status [53,54]. In this study, the WHO-5 was found to have high internal consistency (Cronbach α=0.949; Kaiser-Meyer-Olkin [KMO]=.886).

LS is an individual’s overall assessment of their current life situation. It was measured with the Satisfaction With Life Scale (SWLS), developed by Diener et al [55]. As a classic and widely validated tool across populations and cultural contexts [56], the SWLS includes 5 items, each rated on a 7-point agreement scale. The summative score ranges from 5 to 35, and higher values mean better LS. The SWLS demonstrated high internal consistency in our sample (Cronbach α=0.961; KMO=.914).

Assessment of EHL

The core independent variable was EHL. It was measured using the 8-item eHealth Literacy Scale (eHEALS). Items are scored on a 5-point agreement scale (1=“strongly disagree,” 5=“strongly agree”), yielding a total from 8 to 40. Elevated scores correspond to a more positive self-assessment of one’s ability to obtain, judge, and use eHealth information [57]. In our study, the instrument showed excellent internal consistency (Cronbach α=0.986; KMO=.943). The scale has been widely applied in health behavior research across diverse populations and cultural settings [58].

Assessment of Mediating Variables

Two mediators included in this study were ATOA and SE. ATOA refers to an individual’s expectations, cognitive representations, and perceptions regarding their own aging process. It was measured by means of the Attitudes Toward One’s Own Aging subscale of the Philadelphia Geriatric Center Morale Scale. This subscale includes 5 items rated on a 5-point Likert scale (1=“strongly disagree,” 5=“strongly agree”), with 2 negative items reverse-coded. Average scores were calculated, with higher scores indicating more positive ATOA [59]. The internal consistency in this study was acceptable (Cronbach α=0.745; KMO=.811).

SE serves as a key internal motivator that drives individual behavior, influences emotional experiences, and ultimately shapes outcomes. SE was assessed using the General Self-Efficacy Scale (GSES), developed by Schwarzer and Jerusalem in 1995. The scale comprises 10 items rated on a 4-point Likert scale (1=“not at all true,” 4=“exactly true”). Average scores were computed, with higher scores reflecting stronger SE [60]. Reliability was excellent in our study (Cronbach α=0.965; KMO=.968).

Ascertainment of Covariates

To minimize potential confounding, this study incorporated a range of covariates based on prior literature [45,61]. These included demographic characteristics (age, gender, ethnicity, region, residence, education, and spouse status), socioeconomic factors (income, living arrangements, employment status, and number of children), lifestyle behaviors (caregiving for grandchildren, smoking, drinking, and exercise), social security status (participation in medical and pension insurance), and health-related variables (number of chronic diseases and use of health services). Detailed coding of variables is shown in Table 1.

XSL•FO
RenderX

Table . Variables encoding.

Variables	Definition	Type
Age	Actual age in years	Continuous
Sex	0=Female; 1=Male	Categorical
Ethnicity	0=Minority; 1=Han	Categorical
Region	1=Eastern; 2=Central; 3=Western	Categorical
Residence	0=Rural; 1=Urban	Categorical
Education	1=Below primary school; 2=Primary school; 3=Junior school; 4=Senior or vocational school; 5=Bachelor and above	Categorical
Spouse	0=Without; 1=With	Categorical
Income[a]	0=None; 1 =≤1000; 2=1001 - 2000; 3=2001 - 5000; 4 =≥5001	Categorical
Living	0=With family; 1=Empty nest	Categorical
Employment	0=Not; 1=Currently	Categorical
Children	0=None; 1=Have	Categorical
Grandcare	0=No; 1=Yes	Categorical
Smoking	0=No; 1=Yes	Categorical
Drinking	0=No; 1=Yes	Categorical
Exercise	0=No; 1=Yes	Categorical
Medical insurance	0=No; 1=Yes	Categorical
Pension insurance	0=No; 1=Yes	Categorical
Chronic	Actual number	Continuous
Outpatient	Actual frequency	Continuous
Hospitalization	Actual frequency	Continuous

[a]Income is denominated in Chinese yuan.

Statistical Analysis

All statistical analyses were conducted using SPSS 25.0 (IBM Corp) and R 4.4.3 (R Foundation for Statistical Computing). Descriptive statistics, univariate analysis, and correlation analysis were performed. Continuous variables are reported as means and SDs, while categorical variables are reported as frequencies and percentages. To scrutinize the potential for common method variance, Harman single-factor test was performed. Given the correlations among the 3 healthy aging indicators, this study used seemingly unrelated regressions to simultaneously estimate the 3 equations, thereby improving estimation efficiency [62]. To control the increased familywise error rate due to multiple testing, the significance of the associations between EHL and the 3 HRQoL outcomes was assessed using a Bonferroni correction (adjusted α=.05/3). The mediation model, which specified ATOA and SE as parallel mediators, was tested with the PROCESS macro for SPSS (Model 4) [63]. To investigate the significance of the indirect effects, a bootstrapping procedure was applied based on 5000 resamples. Within each HRQoL dimension (PH, MH, and LS), the significance of indirect effects was evaluated using the Holm-Bonferroni sequential correction procedure. Heterogeneity analysis was conducted through subgroup regression. For these exploratory subgroup analyses, the false discovery rate (FDR) across all tests concerning the association between EHL and outcomes was controlled using the Benjamini-Hochberg procedure. All corrections pertained specifically to the hypothesis tests involving EHL. Statistical significance was tested using a 2-tailed test with an α level of .05, and the corrected significance level was fully considered.

Ethical Considerations

This study was approved by the Ethics Committee of Zhongda Hospital, Southeast University (2024ZDSYLL294-Y01). All participants provided informed consent before their inclusion in the study. Anonymous data collection was used to ensure confidentiality, and no personal identifiers, such as names, were included in our study materials. No financial or material compensation was provided for participation in the study. The research strictly adhered to the principles of the Declaration of Helsinki (1964), the Council for International Organizations of Medical Sciences (CIOMS) International Ethical Guidelines, and the World Health Organization's standards and procedures for research involving human participants.

XSL•FO
RenderX

Results

Descriptive Statistics

A total of 8364 participants were included in this study (3546 men and 4818 women), with a mean age of 70.83 (SD 8.39) years. In terms of residence, rural residents slightly outnumbered urban residents (56.5% vs 43.5%). Regarding education level, 31.3% (2614/8364) had not completed primary school, while only 3.9% (329/8364) held a bachelor's degree or higher. In terms of lifestyle, the vast majority did not smoke (7036/8364, 84.1%) or drink alcohol (6743/8364, 80.6%). Regarding chronic disease prevalence, participants had an average of 1.26 types of chronic diseases (SD 1.14). The average EHL score was 20.13 (SD 10.58), indicating a relatively low level of EHL. The mean scores for PH, MH, and LS were 72.01 (SD 18.28), 71.18 (SD 20.47), and 26.71 (SD 5.52), respectively. Detailed participant characteristics are presented in Table 2.

XSL•FO
RenderX

Table . Descriptive statistics (N=8364).

Variables	Values
Age (years), mean (SD)	70.83 (8.39)
Sex, n (%)	
Female	4818 (57.6)
Male	3546 (42.4)
Ethnicity, n (%)	
Minority	1019 (12.2)
Han	7345 (87.8)
Region, n (%)	
Eastern	5259 (62.9)
Central	1670 (20.0)
Western	1435 (17.2)
Residence, n (%)	
Rural	4726 (56.5)
Urban	3638 (43.5)
Education, n (%)	
Below primary school	2614 (31.3)
Primary school	1422 (17.0)
Junior school	1979 (23.7)
Senior or vocational school	2020 (24.2)
Bachelor and above	329 (3.9)
Spouse, n (%)	
Without	1785 (21.3)
With	6579 (78.7)
Income[a], n (%)	
None	1253 (15.0)
≤1000	2310 (27.6)
1001 - 2000	1064 (12.7)
2001 - 5000	2706 (32.4)
≥5001	1031 (12.3)
Living, n (%)	
With family	7207 (86.2)
Empty nest	1157 (13.8)
Employment, n (%)	
Not	6596 (78.9)
Currently	1768 (21.1)
Children, n (%)	
None	352 (4.2)
Have	8012 (95.8)
Grandcare, n (%)	
No	7089 (84.8)
Yes	1275 (15.2)
Smoking, n (%)	

XSL•FO
RenderX

Variables	Values
No	7036 (84.1)
Yes	1328 (15.9)
Drinking, n (%)	
No	6743 (80.6)
Yes	1621 (19.4)
Exercise, n (%)	
No	3127 (37.4)
Yes	5237 (62.6)
Medical insurance, n (%)	
No	156 (1.9)
Yes	8208 (98.1)
Pension insurance, n (%)	
No	532 (6.4)
Yes	7832 (93.6)
Chronic, mean (SD)	1.26 (1.14)
Outpatient, mean (SD)	2.55 (4.82)
Hospitalization, mean (SD)	0.23 (0.67)

[a]Income is denominated in Chinese yuan.

Common Method Bias Test

To assess the potential for common method bias, Harman single-factor test was performed on all measurement items. Six factors with eigenvalues greater than 1 were identified, with the first factor accounting for 38.76% of the variance—below the commonly accepted 50% threshold [64]. Consequently, these findings provide evidence that common method bias did not pose a significant threat to the validity of the study's results.

Univariate Analysis

Results of the univariate analysis are presented in Table 3. Significant differences in PH, MH, and LS were observed across region, residence, education, income, spouse, grandcare, and exercise (all $P<.001$). Overall, older adults living in eastern regions, with higher education and income levels, having a spouse, exercising regularly, or providing grandchild care exhibited better outcomes in healthy aging, whereas the effects of other factors were relatively weaker.

XSL•FO
RenderX

Table . Univariate analysis results.

Variables	PH[a]			MH[b]			LS[c]		
	t test (*df*)	*F* test (*df*)	*P* value	*t* test (*df*)	*F* test (*df*)	*P* value	*t* test (*df*)	*F* test (*df*)	*P* value
Sex	−0.345 (8362)	—[d]	.73	0.791 (8362)	—	.43	2.736 (8362)	—	.006
Ethnicity	−0.543 (8362)	—	.59	−6.845 (8362)	—	<.001	−6.392 (8362)	—	<.001
Region	—	152.358 (2, 8361)	<.001	—	242.853 (2, 8361)	<.001	—	265.219 (2, 8361)	<.001
Residence	6.003 (8362)	—	<.001	−6.136 (8362)	—	<.001	3.661 (8362)	—	<.001
Education	—	18.259 (4, 8359)	<.001	—	42.114 (4, 8359)	<.001	—	21.243 (4, 8359)	<.001
Spouse	−4.252 (8362)	—	<.001	−5.838 (8362)	—	<.001	−6.546 (8362)	—	<.001
Income	—	17.465 (4, 8359)	<.001	—	39.202 (4, 8359)	<.001	—	48.025 (4, 8359)	<.001
Living	1.248 (8362)	—	.21	2.326 (8362)	—	.02	2.588 (8362)	—	.01
Employment	−1.261 (8362)	—	.21	3.691 (8362)	—	<.001	0.854 (8362)	—	.39
Children	6.070 (8362)	—	<.001	−0.484 (8362)	—	.63	−1.143 (8362)	—	.25
Grandcare	−8.881 (8362)	—	<.001	−6.695 (8362)	—	<.001	−5.350 (8362)	—	<.001
Smoking	−0.561 (8362)	—	.58	3.715 (8362)	—	<.001	3.752 (8362)	—	<.001
Drinking	−1.350 (8362)	—	.18	−0.764 (8362)	—	.45	4.253 (8362)	—	<.001
Exercise	−8.346 (8362)	—	<.001	−15.050 (8362)	—	<.001	−11.874 (8362)	—	<.001
Medical insurance	−0.797 (8362)	—	.43	−2.545 (8362)	—	.01	−4.622 (8362)	—	<.001
Pension insurance	1.026 (8362)	—	.31	−1.204 (8362)	—	.23	−0.751 (8362)	—	.45

[a]PH: physical health.

[b]MH: mental health.

[c]LS: life satisfaction.

[d]Not applicable.

Correlation Analysis

Given that our study includes multiple variable types and some continuous variables deviate from the normal distribution, Spearman correlation analysis was used to assess the associations among age, chronic, outpatient, hospitalization, EHL, ATOA, SE, PH, MH, and LS. Results are shown in Figure 2. EHL was positively correlated with ATOA, SE, and HRQoL (ρ=0.102 - 0.370; *P*<.05). In contrast, aside from age and hospitalizations, the number of chronic diseases and outpatient visits was negatively correlated with EHL, ATOA, SE, and HRQoL (ρ= –0.331 to 0.030; *P*<.05).

XSL•FO
RenderX

Figure 2. Correlation heatmap. *P<.05; **P<.01; ***P<.001. ATOA: attitudes toward own aging; EHL: electronic health literacy; LS: life satisfaction; MH: mental health; PH: physical health; SE: self-efficacy.

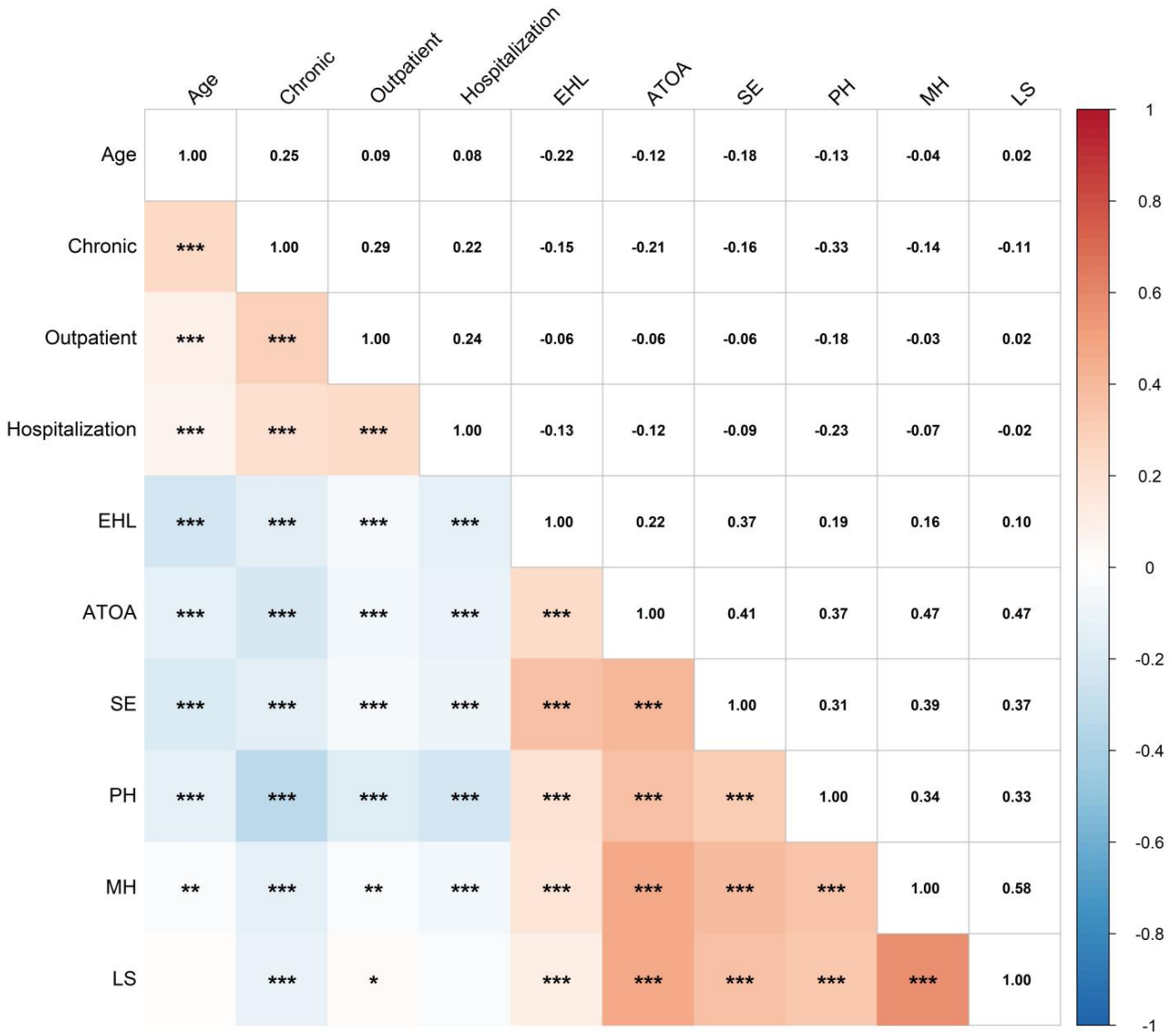

Basic Regression

Based on the univariate analysis results, seemingly unrelated regressions were used to simultaneously estimate PH, MH, and LS (Table 4). In Model 1, EHL was positively associated with PH, MH, and LS (all P<.001). After controlling for covariates in Model 2, the positive associations between EHL and the 3 outcomes remained significant (all P<.001). Notably, these associations retained their statistical significance even after applying a Bonferroni correction for the 3 comparisons (adjusted α=0.0167). These findings confirm that EHL exerts a robust positive influence on multiple dimensions of HRQoL, thereby supporting hypothesis 1.

XSL•FO
RenderX

Table . Baseline regression results.[a]

Variables	Model 1[b]			Model 2[c]		
	PH[d]	MH[e]	LS[f]	PH	MH	LS
EHL[g]	0.235[h] (0.018)	0.273[h] (0.021)	0.064[h] (0.006)	0.273[h] (0.189)	0.190[h] (0.026)	0.082[h] (0.007)
Age	—[i]	—	—	0.147 (0.025)	0.085[j] (0.030)	0.053[h] (0.008)
Ethnicity	—	—	—	−6.734[h] (0.630)	−0.756 (0.811)	−0.902[h] (0.197)
Region						
Central				−7.180[h] (0.501)	−7.655[h] (0.642)	−2.741[h] (0.167)
Western				−15.127[h] (0.644)	−10.843[h] (0.754)	−3.639[h] (0.199)
Residence				−6.761[h] (0.656)	−4.701[h] (0.750)	−3.639[h] (0.202)
Education						
Primary school				−0.399 (0.519)	−1.018 (0.442)	−0.212 (0.170)
Junior school				1.306[k] (0.550)	0.442 (0.675)	0.042 (0.179)
Senior or vocational school				4.399[h] (0.686)	3.041[h] (0.788)	0.922[h] (0.206)
Bachelor and above				4.296[h] (1.098)	4.097[j] (1.321)	1.203[j] (0.348)
Spouse				0.261 (0.624)	2.293[j] (0.762)	0.874[h] (0.192)
Income						
≤1000				0.613 (0.517)	1.929[j] (0.710)	2.074[h] (0.179)
1001–2000				1.888[j] (0.629)	2.372[j] (0.785)	1.450[h] (0.207)
2001–5000				1.723[k] (0.756)	6.774[h] (0.917)	2.034[h] (0.235)
≥5001				2.396[j] (0.832)	5.383[h] (1.085)	2.609[h] (0.285)
Living				0.515 (0.693)	0.406 (0.871)	0.035 (0.226)
Grandcare				1.468[j] (0.446)	2.225[h] (0.573)	0.268 (0.155)
Smoking				−0.192 (0.475)	−1.611[j] (0.626)	−0.586[h] (0.158)
Exercise				2.078[h] (0.397)	4.939[h] (0.483)	1.064[h] (0.125)
Medical insurance				1.198 (1.332)	4.118[k] (1.688)	2.059[h] (0.485)
Chronic				−2.903[h] (0.174)	−2.027[h] (0.207)	−0.569[h] (0.056)
Outpatient				−0.290[h] (0.036)	−0.078 (0.042)	−0.012 (0.012)
Hospitalization				−4.188[h] (0.312)	−1.259[h] (0.344)	−0.256[j] (0.097)
Observations	8364	8364	8364	8364	8364	8364
R^2	0.019	0.02	0.015	0.19	0.116	0.155

[a]All categorical variables are coded with the first group as the reference category. Robust standard errors are shown in parentheses. The significance of the association between EHL and HRQoL dimensions was corrected using the Bonferroni method (corrected α=0.0167), with all P_{Bonf} <.05.

[b]Model 1: Unadjusted.

[c]Model 2: Adjusted for age, ethnicity, region, residence, education, spouse, income, living, grandcare, smoking, exercise, medical insurance, chronic, outpatient, and hospitalization.

[d]PH: physical health.

[e]MH: mental health.

[f]LS: life satisfaction.

[g]EHL: electronic health literacy.

XSL•FO
RenderX

[h]*P*<.001.
[i]Not applicable.
[j]*P*<.01.
[k]*P*<.05.

Mediation Analysis

Mediation analysis was undertaken using PROCESS Model 4 in SPSS, with ATOA and SE specified as parallel mediators. After controlling for covariates, the model included EHL as the predictor, ATOA and SE as mediators, and PH, MH, and LS as outcomes (Table S1 in Multimedia Appendix 1). Results showed that EHL positively predicted both ATOA (β=0.015; *P*<.001) and SE (β=0.019; *P*<.001), indicating that higher EHL was associated with more positive aging attitudes and stronger self-efficacy.

For PH, EHL maintained a significant positive association (β=0.128; *P*<.001), and both ATOA (β=3.270; *P*<.001) and SE (β=5.129; *P*<.001) strongly predicted PH. Thus, EHL directly promoted PH and also indirectly contributed via ATOA and SE. For MH, the direct association between EHL and MH became negative after including the mediators (β=–.069; *P*<.01). However, both ATOA (β=9.417; *P*<.001) and SE (β=6.929; *P*<.001) exerted strong positive effects on MH, suggesting that the positive relationship between EHL and MH was almost entirely mediated by ATOA and SE, with evidence of a suppression effect. For LS, the direct effect of EHL was nonsignificant, while both ATOA (β=2.583; *P*<.001) and SE (β=2.250; *P*<.001) positively predicted LS. This indicates that the association between EHL and LS was fully mediated by ATOA and SE.

Following established recommendations [65], bootstrapping with 5000 resamples was used to assess mediation significance via 95% CI (Table 5; Figure 3). For PH, the direct effect of EHL was significant (95% CI 0.086-0.170), and the indirect effects of ATOA (95% CI 0.038-0.059) and SE (95% CI 0.081-0.112) were also significant, with total indirect effects accounting for 52.7% of the total effect. For MH, the direct effect of EHL was negative and significant (95% CI –0.115 to –0.024), but the indirect effects through ATOA (95% CI 0.119-0.158) and SE (95% CI 0.112-0.149) were positive, yielding a positive total effect (95% CI 0.151-0.248)—indicating inconsistent mediation. For LS, the direct effect of EHL was nonsignificant (95% CI –0.012 to 0.012), while the indirect effects of ATOA (95% CI 0.033-0.043) and SE (95% CI 0.036-0.047) were significant, with total indirect effects accounting for 100% of the total effect. Furthermore, after using Holm-Bonferroni sequential correction, it was found that the bootstrap CI for all indirect effects did not include zero, indicating that all mediating paths remained statistically significant. These results support hypotheses 2a and 2b.

XSL•FO
RenderX

Table . Significance test for mediating effect[a].

Variables	Effect	Paths	β (SE)	95% CI	Proportion
PH[b]	Direct	EHL[c] →PH	.128 (0.022)	[0.086-0.170]	45.8%
	Indirect	EHL→ATOA[d] →PH	.048 (0.005)	[0.038-0.059]	17.6%
		EHL→SE[e]→PH	.096 (0.008)	[0.081-0.112]	35.2%
		Total Indirect	.144 (0.009)	[0.126-0.163]	52.7%
	Total		.273 (0.021)	[0.231-0.314]	__[f]
MH[g]	Direct	EHL→MH	–.069 (0.023)	[–0.115 to –0.024]	—
	Indirect	EHL→ATOA→MH	.138 (0.010)	[0.119-0.158]	—
		EHL→SE→MH	.130 (0.010)	[0.112-0.149]	—
		Total Indirect	.268 (0.014)	[0.240-0.296]	—
	Total		.199 (0.025)	[0.151-0.248]	—
LS[h]	Direct	EHL→LS	.000 (0.006)	[–0.012 to 0.012]	0.0%
	Indirect	EHL→ATOA[d]→LS	.038 (0.003)	[0.033-0.043]	47.5%
		EHL→SE→LS	.042 (0.003)	[0.036-0.047]	52.5%
		Total Indirect	.080 (0.004)	[0.071-0.088]	100%
	Total		0.080 (0.007)	[0.067-0.093]	—

[a]Proportion is the effect size ratio; for mental health, direct and total effects have opposite signs (inconsistent mediation), so proportion is not reported. After applying the Holm-Bonferroni sequential correction procedure, the bootstrapCI for all indirect effects did not include zero.

[b]PH: physical health.

[c]EHL: electronic health literacy.

[d]ATOA: attitudes toward own aging.

[e]SE: self-efficacy.

[f]Not applicable.

[g]MH: mental health.

[h]LS: life satisfaction.

Figure 3. Mediation model. ***P*<.01; ****P*<.001. ATOA: attitudes toward own aging; EHL: electronic health literacy; LS: life satisfaction; MH: mental health; PH: physical health; SE: self-efficacy.

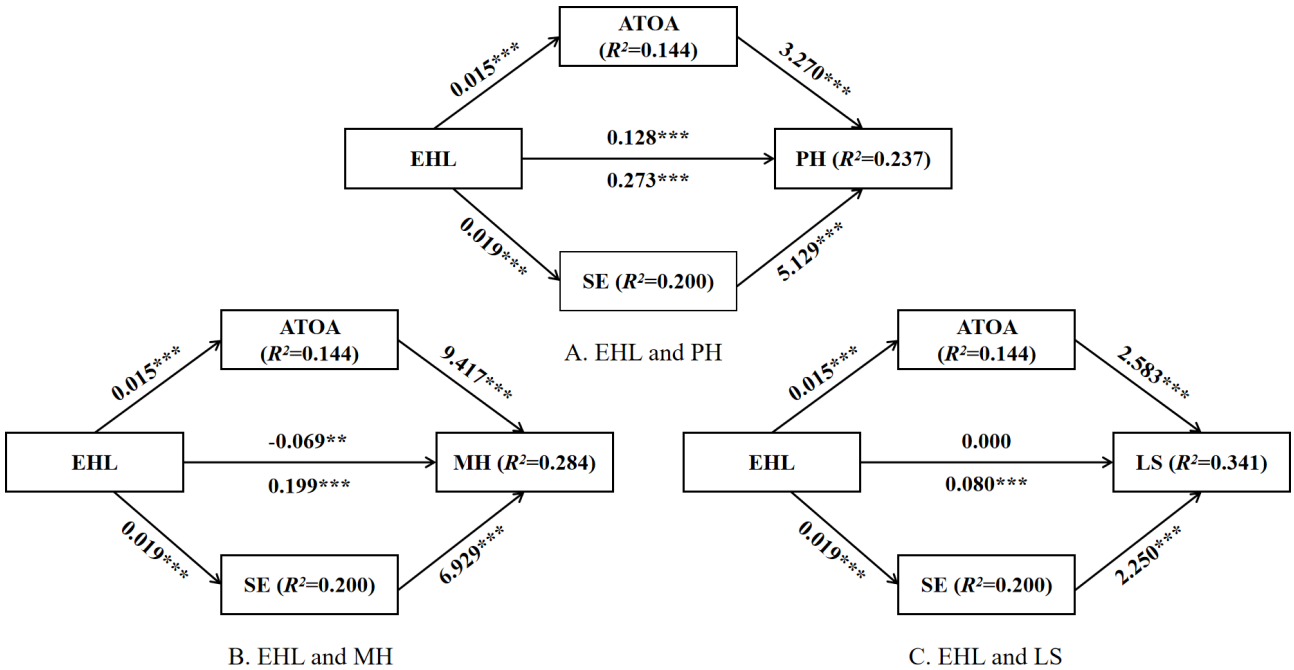

XSL•FO
RenderX

Heterogeneity Test

Our study further conducted subgroup analyses stratified by age, residence, and physical activity behavior, and the results are presented in Tables S2-S4 in Multimedia Appendix 1. Given that subgroup analyses involve multiple comparisons, we performed FDR correction (Benjamini-Hochberg) for all association tests between EHL and different HRQoL dimensions across all subgroups. All significant associations remained significant after correction (all $P_{BH\text{-}FDR}$<.05), indicating that the subgroup analysis results are robust.

The role of EHL showed notable age differences. Among the younger-old adults (≤74 years), EHL had the strongest association with PH (β=0.288; *P*<.001), and also demonstrated significant positive effects on MH (β=0.188; *P*<.001) and LS (β=0.089; *P*<.001). This indicates that, within this group, EHL could more comprehensively enhance their HRQoL. In contrast, among the older-old adults (>74 years), EHL still had significant positive predictive effects on all 3 dimensions, but its strength was generally weaker. The coefficient for PH decreased to 0.222 (*P*<.001), and the promoting effects on MH (β=0.175; *P*<.001) and LS (β=0.060; *P*<.001) were also lower than those in the younger group. These results support hypothesis 3a.

Significant urban-rural differences were observed. Among rural older adults, EHL was significantly and positively associated with PH (β=0.215; *P*<.001), but its association with MH was nonsignificant, and its effect on LS was positive but relatively weak (β=.019; *P*=.032). In contrast, among urban older adults, EHL was more strongly associated with PH (β=0.237; *P*<.001), and its associations with MH and LS were also considerably stronger than in rural areas. Overall, EHL had a broader impact on HRQoL in urban older adults, while in rural populations, its health value was concentrated primarily in PH. These findings support hypothesis 3b.

Associations also show differences with varying exercise habits. Among nonexercisers, EHL had its strongest positive association with PH (β=0.322; *P*<.001), suggesting that EHL effectively compensated for health deficits due to lack of exercise and served as a core support for improving PH. EHL also positively predicted MH (β=0.196; *P*<.001) and LS (β=0.067;*P*<.001) in this group. In contrast, among exercisers, the health effects of EHL, though still significant, were weaker. The association with PH dropped to β=0.193 (*P*<.001), while its effect on MH (β=0.147; *P*<.001) was also lower than in nonexercisers, and the effect on LS showed little difference. These findings suggest that nonexercisers' HRQoL was more strongly dependent on EHL, whereas exercisers already benefited from the baseline health improvements of physical activity, limiting the additional value of EHL. These results support hypothesis 3c.

Discussion

Principal Findings

This study found that EHL is significantly associated with HRQoL among older adults in China. The enhancement of EHL contributes to improving HRQoL in older adults. Previous studies have established a positive association between general health literacy and HRQoL [66], while relevant systematic reviews have clarified the intrinsic relationship between EHL and multidimensional health outcomes [25]. For Chinese older adults, a prior study conducted in Jinan City initially explored the relationship between EHL and HRQoL [45]. These findings are consistent with our conclusions. Our study further expands the representativeness of the sample and the generalizability of conclusions, providing empirical evidence for extending research in this field to broader populations of Chinese older adults.

The mechanism of EHL varies across different dimensions of HRQoL. EHL exerted a direct effect on PH, whereas its associations with MH and LS were almost entirely mediated by ATOA and SE. Older adults with higher EHL can more efficiently retrieve, comprehend, and apply online health information, enabling more scientifically grounded disease prevention, symptom management, medication adherence, and health care decision-making [67]. This process involves a direct conversion of cognitive and behavioral skills, yielding immediate, instrumental positive effects on PH. This direct effect fundamentally reflects the immediacy of translating information literacy into health practices, highlighting EHL's core value as a capacity for health information use [68].

An inconsistent mediation was observed in the MH model. The direct effect of EHL on MH was negative, yet the indirect effects via ATOA and SE were positive, resulting in a positive total effect. Possessing high EHL may also expose individuals to an overwhelming volume of health information, potentially triggering cognitive overload and uncertainty [69]. This fosters negative emotions such as health anxiety and cyberhypochondria. Catastrophizing minor symptoms further exacerbates psychological burden [70]. Older adults may intensify perceptions of health limitations or experience frustration from digital operational barriers, amplifying direct negative impacts. However, EHL can exert protective effects by fostering positive psychological mechanisms, as ATOA reduces age-related health fears, while SE enhances confidence in tackling health challenges [71,72]. Both collectively boost psychological resilience and emotional regulation capacity. This indicates that benefits to older adults' MH stem not from the information itself, but from the psychological coping mechanisms constructed through EHL. Without such transformative processes, high EHL may instead pose risks to the psychological well-being of older adults.

For LS, the association between EHL and outcomes was fully mediated by ATOA and SE. As a cognitive evaluation of overall life quality, LS is more closely associated with universal psychological resources and meaning frameworks [73]. EHL, as a tool-based skill, does not directly contribute to overall life evaluations. Instead, EHL is more likely to enhance LS by strengthening positive aging attitudes (enabling older adults to embrace the life course) and leveraging SE to implement health-promoting behaviors. In other words, EHL functions as an enabler: the knowledge and sense of control afforded by EHL are filtered and sublimated through an individual's meaning system and core beliefs before being translated into more holistic LS.

XSL•FO
RenderX

The analysis also revealed heterogeneity. The health benefits gained by the younger-old from EHL are the most comprehensive and pronounced, indicating that their physiological and cognitive reserves can more effectively translate health literacy into well-being. The intensity of benefits for the older-old adults generally diminishes, potentially stemming from the fact that the marginal benefits of EHL may decrease with age due to cumulative health deficits and the dominance of aging processes [74]. Urban older adults benefited more comprehensively from EHL, likely due to stronger resource conditions that facilitated the translation of literacy into gains in MH and LS, while among rural older adults, the benefits of EHL were concentrated on PH [75]. Similarly, nonexercisers derived greater benefits from EHL, suggesting that it compensates for health deficits due to physical inactivity [76]. Among exercisers, the marginal effects of EHL were weaker, as they already benefited from baseline advantages of physical activity.

Contributions and Implications

This study makes 3 key contributions. First, it extends the literature by showing that EHL is linked not only to PH but also to MH and LS. Second, it identifies ATOA and SE as psychosocial mechanisms through which EHL influences HRQoL. Third, it reveals heterogeneity across subgroups, highlighting the compensatory role of EHL for vulnerable populations such as rural residents and nonexercisers.

These findings confirm the necessity of integrating EHL into healthy aging policies. Policymakers should treat EHL as a social determinant of health and incorporate it into digital inclusion and aging strategies, with particular attention to disadvantaged groups. Practical approaches include (1) recruiting family doctors and volunteers to form a silver-haired lecturer team, implementing EHL training programs at the community level, and establishing personalized advancement pathways for literacy development. The mature models of Shanghai's "Senior Digital Life Workshops" and "Senior Health Cloud Classroom" can serve as references. (2) Rooted in China's unique filial piety and integrity culture, advocating intergenerational interaction models featuring digital reciprocity, mutual learning, and generational integration. (3) With reference to the practical experience of aging-friendly digital infrastructure construction in provinces such as Jiangsu, Zhejiang, and Guangdong, nationwide standardized upgrades and renovations of age-friendly digital infrastructure should be promoted. (4) Expanding digital health services and digital life services that are compatible with the literacy level of the older adults, such as disease prevention and management, health promotion, MH care, and medical electronic commerce, to effectively meet the needs of the older adults. Overall, the comprehensive application of strategies such as community training, intergenerational teaching, infrastructure upgrading projects, and inclusive service expansion can enhance older adults' confidence in using digital health resources, foster a more positive attitude toward aging, strengthen SE, and ultimately improve HRQoL. In addition, EHL should be incorporated into the policy evaluation framework. For example, in the policy evaluation of advancing the informatization construction of compact county medical communities, indicators such as the improvement rate of EHL among the older adults in the county and the use rate of digital health tools among rural older adults should be included in the evaluation system. This measure can align digital health initiatives with the public health goals of health equity and digital inclusion, ensuring that the development of digital health will not widen existing health gaps but instead serve as a catalyst for promoting universal healthy aging in the digital era.

Limitations

Inevitably, our study has several limitations. First, the cross-sectional design precludes causal inference; longitudinal follow-ups are needed. Second, although the sampling ensured regional and population representativeness, generalizability to other cultural contexts remains limited. Third, while common method bias was tested and found to be minimal, the influence of unmeasured confounding cannot be completely excluded. For instance, our study did not account for confounding factors among older adults. Finally, our study treats ATOA and SE as parallel mediating variables. Future research should explore the potential dynamic interdependence and interactions between them to deepen our understanding of the mediating mechanism.

Conclusions

In conclusion, EHL is strongly associated with HRQoL among older adults. Its association with PH appears relatively direct, whereas its associations with MH and LS are largely mediated by ATOA and SE. The observed heterogeneity by age, residence, and exercise behavior suggests that targeted interventions are needed for specific subgroups. Overall, this study unveils the potential of EHL as a key lever for promoting healthy aging in the digital era, emphasizing the urgency of bridging the digital divide and cultivating EHL to support the well-being of older adults.

Acknowledgments

The authors are thankful to the editors and reviewers for their comments and feedback.

Funding

This study was supported by the Major Project of the National Social Science Fund of China (grant number 23&ZD188).

Data Availability

The datasets generated or analyzed during this study are not publicly available due to ethical restrictions as the data contain potentially identifying information but are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: YW, JY
Data curation: YW, JY
Formal analysis: YW, JY
Investigation: YW, JY
Methodology: YW, JY
Project administration: YW
Software: JY, BD
Resources: BD
Supervision: BD
Validation: BD
Writing – original draft: YW
Writing – review & editing: BD

Conflicts of Interest

None declared.

Multimedia Appendix 1
Mediation analysis and heterogeneity analysis results.
[DOCX File, 54 KB - jmir_v28i1e84700_app1.docx]

References

1. Christensen K, Doblhammer G, Rau R, Vaupel JW. Ageing populations: the challenges ahead. Lancet 2009 Oct 3;374(9696):1196-1208. [doi: 10.1016/S0140-6736(09)61460-4] [Medline: 19801098]
2. Beard JR, Officer A, de Carvalho IA, et al. The world report on ageing and health: a policy framework for healthy ageing. Lancet 2016 May;387(10033):2145-2154. [doi: 10.1016/S0140-6736(15)00516-4]
3. Wang HM, Beyer M, Gensichen J, Gerlach FM. Health-related quality of life among general practice patients with differing chronic diseases in Germany: cross sectional survey. BMC Public Health 2008 Jul 21;8:246. [doi: 10.1186/1471-2458-8-246] [Medline: 18638419]
4. Chen X, Giles J, Yao Y, et al. The path to healthy ageing in China: a Peking University-Lancet Commission. Lancet 2022 Dec 3;400(10367):1967-2006. [doi: 10.1016/S0140-6736(22)01546-X] [Medline: 36423650]
5. Chen C, Liu GG, Shi QL, et al. Health-related quality of life and associated factors among oldest-old in China. J Nutr Health Aging 2020;24(3):330-338. [doi: 10.1007/s12603-020-1327-2] [Medline: 32115616]
6. Li H, Tao S, Sun S, Xiao Y, Liu Y. The relationship between health literacy and health-related quality of life in Chinese older adults: a cross-sectional study. Front Public Health 2024;12:1288906. [doi: 10.3389/fpubh.2024.1288906]
7. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85. [doi: 10.2196/jmir.1933] [Medline: 23615206]
8. Hunsaker A, Hargittai E. A review of Internet use among older adults. New Media Soc 2018 Oct;20(10):3937-3954. [doi: 10.1177/1461444818787348]
9. Mitchell UA, Chebli PG, Ruggiero L, Muramatsu N. The digital divide in health-related technology use: the significance of race/ethnicity. Gerontologist 2019 Jan 9;59(1):6-14. [doi: 10.1093/geront/gny138] [Medline: 30452660]
10. Kruse C, Fohn J, Wilson N, Nunez Patlan E, Zipp S, Mileski M. Utilization barriers and medical outcomes commensurate with the use of telehealth among older adults: systematic review. JMIR Med Inform 2020 Aug 12;8(8):e20359. [doi: 10.2196/20359] [Medline: 32784177]
11. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9. [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
12. Çöme O, Gökdemir Ö, Bilik Sezer B, Kasapoğlu SS, Kjær NK, Güldal D. The interplay between cognitive function and digital health literacy among older adults: implications for e-health equity and accessibility. Int J Med Inform 2025 Sep;201:105934. [doi: 10.1016/j.ijmedinf.2025.105934] [Medline: 40273597]
13. Wu Y, Wen J, Wang X, et al. Associations between e-health literacy and chronic disease self-management in older Chinese patients with chronic non-communicable diseases: a mediation analysis. BMC Public Health 2022 Nov 29;22(1):2226. [doi: 10.1186/s12889-022-14695-4] [Medline: 36447176]
14. Ivan L, Marston HR, Prabhu VG, et al. Successful aging across middle versus high-income countries: an analysis of the role of eHealth literacy associated with loneliness and well-being. Gerontologist 2024 Dec 13;65(1):gnae170. [doi: 10.1093/geront/gnae170] [Medline: 39673788]

XSL•FO
RenderX

15. Chen A, Wang X, Xu A, Liu Q, Tong B. Analysis of the impact of e-health literacy on health behavior among the elderly in Jiangsu: structural equation model. Public Health Nurs 2025;42(4):1461-1468. [doi: 10.1111/phn.13565] [Medline: 40266126]
16. Xie L, Mo PKH. A 3-wave longitudinal study of eHealth literacy and older people's health-related quality of life in China: the mediating role of general self-efficacy. J Am Med Dir Assoc 2024 Dec;25(12):105310. [doi: 10.1016/j.jamda.2024.105310] [Medline: 39426796]
17. Stellefson M, Paige SR, Alber JM, et al. Association between health literacy, electronic health literacy, disease-specific knowledge, and health-related quality of life among adults with chronic obstructive pulmonary disease: cross-sectional study. J Med Internet Res 2019 Jun 6;21(6):e12165. [doi: 10.2196/12165] [Medline: 31172962]
18. Zhang M, Tao S, Ge X, et al. Health self-management behaviors as a bridge between electronic health literacy and health-related quality of life: cross-sectional study from China. J Med Internet Res 2025 Aug 13;27:e74056. [doi: 10.2196/74056] [Medline: 40802975]
19. Lai Y, Chen S, Li M, Ung COL, Hu H. Policy interventions, development trends, and service innovations of internet hospitals in China: documentary analysis and qualitative interview study. J Med Internet Res 2021 Jul 20;23(7):e22330. [doi: 10.2196/22330] [Medline: 34283025]
20. Jokisch MR, Schmidt LI, Doh M. Acceptance of digital health services among older adults: findings on perceived usefulness, self-efficacy, privacy concerns, ICT knowledge, and support seeking. Front Public Health 2022;10:1073756. [doi: 10.3389/fpubh.2022.1073756] [Medline: 36582385]
21. Decade of healthy ageing: baseline report. : World Health Organization; 2020 URL: https://apps.who.int/iris/handle/10665/338677 [accessed 2026-03-13]
22. Singh S, Goodwin S, Zhong S, et al. Inequalities in health-related quality of life and functional health of an aging population: a Canadian community perspective. PLoS One 2024;19(7):e0304457. [doi: 10.1371/journal.pone.0304457] [Medline: 38968188]
23. Tan G, Chao J, Jin S, et al. The association between digital health literacy and health inequalities among Chinese older adults: a multicenter cross-sectional study. Digit Health 2025;11:20552076251349901. [doi: 10.1177/20552076251349901] [Medline: 40547434]
24. van Kessel R, Wong BLH, Clemens T, Brand H. Digital health literacy as a super determinant of health: more than simply the sum of its parts. Internet Interv 2022 Mar;27:100500. [doi: 10.1016/j.invent.2022.100500] [Medline: 35242586]
25. Xie L, Zhang S, Xin M, Zhu M, Lu W, Mo PKH. Electronic health literacy and health-related outcomes among older adults: a systematic review. Prev Med 2022 Apr;157:106997. [doi: 10.1016/j.ypmed.2022.106997] [Medline: 35189203]
26. Rowe JW, Kahn RL. Successful aging. Gerontologist 1997 Aug;37(4):433-440. [doi: 10.1093/geront/37.4.433] [Medline: 9279031]
27. Levy BR, Slade MD, Kasl SV. Longitudinal benefit of positive self-perceptions of aging on functional health. J Gerontol B Psychol Sci Soc Sci 2002 Sep;57(5):409-417. [doi: 10.1093/geronb/57.5.p409] [Medline: 12198099]
28. Di X, Wang L. Impact of digital literacy on aging attitudes of the elderly from the perspective of social cognitive theory. BMC Public Health 2025 Jun 6;25(1):2113. [doi: 10.1186/s12889-025-23377-w] [Medline: 40481505]
29. Bellingtier JA, Neupert SD. Negative aging attitudes predict greater reactivity to daily stressors in older adults. J Gerontol B Psychol Sci Soc Sci 2018;73(7):gbw086. [doi: 10.1093/geronb/gbw086]
30. Levy BR, Myers LM. Preventive health behaviors influenced by self-perceptions of aging. Prev Med 2004 Sep;39(3):625-629. [doi: 10.1016/j.ypmed.2004.02.029] [Medline: 15313104]
31. Levy BR, Hausdorff JM, Hencke R, Wei JY. Reducing cardiovascular stress with positive self-stereotypes of aging. J Gerontol B Psychol Sci Soc Sci 2000 Jul;55(4):205-213. [doi: 10.1093/geronb/55.4.p205] [Medline: 11584877]
32. Paasche-Orlow MK, Wolf MS. The causal pathways linking health literacy to health outcomes. Am J Health Behav 2007 Jul 1;31(1):19-26. [doi: 10.5993/AJHB.31.s1.4]
33. Bandura A. Self-Efficacy: The Exercise of Control: W.H.Freeman; 1997.
34. Park C. Electronic health literacy as a source of self-efficacy among community-dwelling older adults. Clin Gerontol 2025 Mar 15;48(2):289-298. [doi: 10.1080/07317115.2024.2373894]
35. Bandura A. Self-efficacy: toward a unifying theory of behavioral change. Psychol Rev 1977 Mar;84(2):191-215. [doi: 10.1037//0033-295x.84.2.191] [Medline: 847061]
36. Bandura A. Health promotion by social cognitive means. Health Educ Behav 2004 Apr;31(2):143-164. [doi: 10.1177/1090198104263660] [Medline: 15090118]
37. Du S, Tian L, Tian Y, Feng Z, Wang Y. The role of self-efficacy and self-care agency as mediating factors in the link between health literacy and health-promoting lifestyle among older adults post covid 19 era: a multiple mediator model. Geriatr Nurs 2023;54:252-257. [doi: 10.1016/j.gerinurse.2023.10.013] [Medline: 37897930]
38. Arcaya MC, Arcaya AL, Subramanian SV. Inequalities in health: definitions, concepts, and theories. Glob Health Action 2015;8:27106. [doi: 10.3402/gha.v8.27106] [Medline: 26112142]
39. Jiang X, Wang L, Leng Y, et al. The level of electronic health literacy among older adults: a systematic review and meta-analysis. Arch Public Health 2024 Nov 7;82(1):204. [doi: 10.1186/s13690-024-01428-9] [Medline: 39511667]

XSL•FO
RenderX

40. Deng SY, Zhao IY, Ho MH, Saravanakumar P, Molassiotis A, Montayre J. Rural-urban disparities in healthy ageing: evidence from a national study in China. Collegian 2022 Dec;29(6):843-849. [doi: 10.1016/j.colegn.2022.06.008]
41. Pailaha AD. Public health nursing: Challenges and innovations for health literacy in rural area. Public Health Nurs 2023;40(5):769-772. [doi: 10.1111/phn.13223] [Medline: 37392096]
42. Baltes PB, Baltes MM. Psychological perspectives on successful aging: the model of selective optimization with compensation. In: Baltes PB, Baltes MM, editors. Successful Aging: Perspectives from the Behavioral Sciences European Network on Longitudinal Studies on Individual Development: Cambridge University Press; 1990:1-34. [doi: 10.1017/CBO9780511665684.003]
43. Zheng X, Xue Y, Dong F, et al. The association between health-promoting-lifestyles, and socioeconomic, family relationships, social support, health-related quality of life among older adults in China: a cross sectional study. Health Qual Life Outcomes 2022 Apr 20;20(1):64. [doi: 10.1186/s12955-022-01968-0] [Medline: 35443689]
44. Cui K, Zou W, Ji X, Zhang X. Does digital technology make people healthier: the impact of digital use on the lifestyle of Chinese older adults. BMC Geriatr 2024 Jan 22;24(1):85. [doi: 10.1186/s12877-023-04651-1] [Medline: 38254001]
45. Li S, Cui G, Yin Y, Wang S, Liu X, Chen L. Health-promoting behaviors mediate the relationship between eHealth literacy and health-related quality of life among Chinese older adults: a cross-sectional study. Qual Life Res 2021 Aug;30(8):2235-2243. [doi: 10.1007/s11136-021-02797-2] [Medline: 33661455]
46. Lin CY, Ganji M, Griffiths MD, Bravell ME, Broström A, Pakpour AH. Mediated effects of insomnia, psychological distress and medication adherence in the association of eHealth literacy and cardiac events among Iranian older patients with heart failure: a longitudinal study. Eur J Cardiovasc Nurs 2020 Feb;19(2):155-164. [doi: 10.1177/1474515119873648] [Medline: 31516036]
47. Leung AYM, Parial LL, Tolabing MC, et al. Sense of coherence mediates the relationship between digital health literacy and anxiety about the future in aging population during the COVID-19 pandemic: a path analysis. Aging Ment Health 2022 Mar;26(3):544-553. [doi: 10.1080/13607863.2020.1870206] [Medline: 33438448]
48. Choi M. Association of eHealth use, literacy, informational social support, and health-promoting behaviors: mediation of health self-efficacy. Int J Environ Res Public Health 2020 Oct 28;17(21):7890. [doi: 10.3390/ijerph17217890] [Medline: 33126469]
49. Neter E, Brainin E. Association between health literacy, eHealth literacy, and health outcomes among patients with long-term conditions. Eur Psychol 2019 Jan;24(1):68-81. [doi: 10.1027/1016-9040/a000350]
50. The Whoqol Group. The World Health Organization quality of life assessment (WHOQOL): development and general psychometric properties. Soc Sci Med 1998 Jun;46(12):1569-1585. [doi: 10.1016/S0277-9536(98)00009-4]
51. Cheng LJ, Tan RLY, Luo N. Measurement properties of the EQ VAS around the globe: a systematic review and meta-regression analysis. Value Health 2021 Aug;24(8):1223-1233. [doi: 10.1016/j.jval.2021.02.003] [Medline: 34372988]
52. Hu W, Zhou L, Chu J, et al. Estimating population norms for the health-related quality of life of adults in southern Jiangsu Province, China. Sci Rep 2022 Jun 14;12(1):9906. [doi: 10.1038/s41598-022-13910-x] [Medline: 35701516]
53. The World Health Organization-five well-being index (WHO-5). World Health Organization. 2024. URL: https://www.who.int/zh/publications/m/item/WHO-UCN-MSD-MHE-2024.01 [accessed 2026-03-13]
54. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 Well-Being Index: a systematic review of the literature. Psychother Psychosom 2015;84(3):167-176. [doi: 10.1159/000376585] [Medline: 25831962]
55. Diener E, Emmons RA, Larsen RJ, Griffin S. The Satisfaction With Life Scale. J Pers Assess 1985 Feb;49(1):71-75. [doi: 10.1207/s15327752jpa4901_13] [Medline: 16367493]
56. Whisman MA, Judd CM. A cross-national analysis of measurement invariance of the Satisfaction With Life Scale. Psychol Assess 2016 Feb;28(2):239-244. [doi: 10.1037/pas0000181] [Medline: 26168309]
57. Norman CD, Skinner HA. eHEALS: the eHealth Literacy Scale. J Med Internet Res 2006 Nov 14;8(4):e27. [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
58. Oh SS, Kim KA, Kim M, Oh J, Chu SH, Choi J. Measurement of digital literacy among older adults: systematic review. J Med Internet Res 2021 Feb 3;23(2):e26145. [doi: 10.2196/26145] [Medline: 33533727]
59. Neupert SD, Bellingtier JA. Aging attitudes and daily awareness of age-related change interact to predict negative affect. Gerontologist 2017 Aug 1;57(suppl_2):S187-S192. [doi: 10.1093/geront/gnx055] [Medline: 28854606]
60. Scholz U, Doña BG, Sud S, et al. Is general self-efficacy a universal construct? Psychometric findings from 25 countries. Eur J Psychol Assess 2002;18(3):242-251. [doi: 10.1027/1015-5759.18.3.242]
61. Liu S, Lu Y, Wang D, et al. Impact of digital health literacy on health-related quality of life in Chinese community-dwelling older adults: the mediating effect of health-promoting lifestyle. Front Public Health 2023;11:1200722. [doi: 10.3389/fpubh.2023.1200722]
62. Chai Y, Xian G, Wang M, Guo L, Luo S. Aging wisely: the impact of Internet use on older adults' mental health. J Affect Disord 2024 Nov 1;364:139-145. [doi: 10.1016/j.jad.2024.08.076] [Medline: 39147146]
63. Preacher KJ, Hayes AF. SPSS and SAS procedures for estimating indirect effects in simple mediation models. Behav Res Methods Instrum Comput 2004 Nov;36(4):717-731. [doi: 10.3758/bf03206553] [Medline: 15641418]

XSL•FO
RenderX

64. Podsakoff PM, MacKenzie SB, Lee JY, Podsakoff NP. Common method biases in behavioral research: a critical review of the literature and recommended remedies. J Appl Psychol 2003 Oct;88(5):879-903. [doi: 10.1037/0021-9010.88.5.879] [Medline: 14516251]
65. Hayes AF. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach: The Guilford Press; 2013.
66. Panagioti M, Skevington SM, Hann M, et al. Effect of health literacy on the quality of life of older patients with long-term conditions: a large cohort study in UK general practice. Qual Life Res 2018 May;27(5):1257-1268. [doi: 10.1007/s11136-017-1775-2] [Medline: 29322478]
67. Lee EH, Lee YW, Kang EH, Kang HJ. Relationship between electronic health literacy and self-management in people with type 2 diabetes using a structural equation modeling approach. J Nurs Res 2024 Jan 1;32(1):e315. [doi: 10.1097/jnr.0000000000000588] [Medline: 38128065]
68. Huhta AM, Hirvonen N, Huotari ML. Health literacy in web-based health information environments: systematic review of concepts, definitions, and operationalization for measurement. J Med Internet Res 2018 Dec 19;20(12):e10273. [doi: 10.2196/10273] [Medline: 30567690]
69. McMullan RD, Berle D, Arnáez S, Starcevic V. The relationships between health anxiety, online health information seeking, and cyberchondria: systematic review and meta-analysis. J Affect Disord 2019 Feb 15;245:270-278. [doi: 10.1016/j.jad.2018.11.037] [Medline: 30419526]
70. Du X, Witthöft M, Zhang T, Shi C, Ren Z. Interpretation bias in health anxiety: a systematic review and meta-analysis. Psychol Med 2023 Jan;53(1):34-45. [doi: 10.1017/S0033291722003427] [Medline: 36349699]
71. Westerhof GJ, Miche M, Brothers AF, et al. The influence of subjective aging on health and longevity: a meta-analysis of longitudinal data. Psychol Aging 2014 Dec;29(4):793-802. [doi: 10.1037/a0038016] [Medline: 25365689]
72. Seeman TE, Unger JB, McAvay G, Mendes de Leon CF. Self-efficacy beliefs and perceived declines in functional ability: MacArthur studies of successful aging. J Gerontol B Psychol Sci Soc Sci 1999 Jul;54(4):214-222. [doi: 10.1093/geronb/54b.4.p214] [Medline: 12382590]
73. Li JB, Dou K, Liang Y. The relationship between presence of meaning, search for meaning, and subjective well-being: a three-level meta-analysis based on the meaning in life questionnaire. J Happiness Stud 2021 Jan;22(1):467-489. [doi: 10.1007/s10902-020-00230-y]
74. Chen X, Wang N. How does digital literacy affect the health status of senior citizens? Micro-level evidence from the CFPS data. BMC Health Serv Res 2025 Jan 27;25(1):153. [doi: 10.1186/s12913-025-12314-7] [Medline: 39871247]
75. Jing R, Jin G, Guo Y, Zhang Y, Li L. The association between constant and new Internet use and depressive symptoms among older adults in China: the role of structural social capital. Comput Human Behav 2023 Jan;138:107480. [doi: 10.1016/j.chb.2022.107480]
76. Chatterjee A, Prinz A, Gerdes M, Martinez S. Digital interventions on healthy lifestyle management: systematic review. J Med Internet Res 2021 Nov 17;23(11):e26931. [doi: 10.2196/26931] [Medline: 34787575]

Abbreviations

ATOA: attitudes toward own aging
CIOMS: Council for International Organizations of Medical Sciences
eHEALS: eHealth Literacy Scale
EHL: electronic health literacy
EQ-VAS: EuroQol Visual Analogue Scale
FDR: false discovery rate
GSES: General Self-Efficacy Scale
HRQoL: health-related quality of life
KMO: Kaiser-Meyer-Olkin
LS: life satisfaction
MH: mental health
PH: physical health
SE: self-efficacy
SWLS: Satisfaction With Life Scale
WHO-5: World Health Organization-Five Well-Being Index

XSL•FO
RenderX

Edited by A Schwartz, M Balcarras; submitted 14.Oct.2025; peer-reviewed by V Dolničar, Y Qian; accepted 05.Feb.2026; published 24.Mar.2026.

Please cite as:
Wang Y, Dai B, Yao J
The Relationship Between Electronic Health Literacy and Health-Related Quality of Life Among Chinese Older Adults: Cross-Sectional Study
J Med Internet Res 2026;28:e84700
URL: https://www.jmir.org/2026/1/e84700
doi:10.2196/84700

© Yongqiang Wang, Baozhen Dai, Jiazhen Yao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Engagement and Cognitive Function Among Older Adults in China: Cross-Sectional Questionnaire Study and Moderated Mediation Model Analysis

Yongqi Du[1*], MPA; Qing Niu[1*], PhD; Gangrui Tan[1], PhD; Jianqian Chao[1], PhD, Prof Dr; Shengxuan Jin[1], PhD; Leixia Wang[1], PhD

Health Management Research Center, School of Public Health, Southeast University, Nanjing, Jiangsu, China

[*]these authors contributed equally

Corresponding Author:
Jianqian Chao, PhD, Prof Dr
Health Management Research Center
School of Public Health
Southeast University
No. 87, Dingjiaqiao
Gulou District
Nanjing, Jiangsu, 210009
China
Phone: 86 18736361189
Email: jqchao1230220@163.com

Abstract

Background: Given the global demographic shifts and rapid digitalization, digital engagement has emerged as a critical determinant of healthy aging. While previous research has linked digital engagement to cognitive outcomes, the underlying mechanisms remain underexplored among Chinese older adults.

Objective: This study aimed to analyze the relationships between digital engagement and cognitive function among older adults in China through a moderated mediation model guided by the technological reserve hypothesis, with digital health literacy (DHL) and social support as mediators and living arrangements as a moderator.

Methods: We conducted a cross-sectional questionnaire survey using stratified multistage sampling from June to November 2024, including 8123 participants aged 55 years and older. Digital engagement, defined as older adults' use of contemporary digital technologies to support routine daily activities, autonomy, independence, and social inclusion, was assessed using a multidimensional questionnaire. The Chinese eHealth Literacy Scale, the 3-item short version of the Perceived Social Support Scale, and the Mini-Cog test were used to assess DHL, social support, and cognitive function. Guided by a directed acyclic graph based on the technological reserve hypothesis, mediation and moderated mediation analyses were performed using the PROCESS macro in SPSS (IBM Corp) with 5000 bootstrap resamples.

Results: Digital engagement was positively associated with cognitive function among older adults (β=0.241, 95% CI 0.216-0.265). This association was partially mediated by DHL (β=0.059, 95% CI 0.049-0.069) and social support (β=0.012, 95% CI 0.008-0.016), with the combined indirect effects accounting for 29.5% of the total effect (β=0.071, 95% CI 0.061-0.082). Additionally, living arrangements significantly moderated the associations between digital engagement and cognitive function (β=0.109, 95% CI 0.052-0.166), digital engagement and DHL (β=0.063, 95% CI 0.014-0.112), and digital engagement and social support (β=0.151, 95% CI 0.089-0.212). These effects were stronger among older adults living alone.

Conclusions: This study contributes to the understanding of cognitive aging in the digital environment from the perspective of the technological reserve hypothesis and digital engagement. Digital engagement influenced cognitive function via DHL and social support, and these associations of digital engagement with cognitive function, DHL, and social support were stronger among older adults living alone. Digital health interventions and public health policies should target both DHL and social support among older populations and prioritize older adults living alone.

(J Med Internet Res 2026;28:e83955) doi:10.2196/83955

XSL•FO
RenderX

KEYWORDS

digital engagement; cognitive function; digital health literacy; social support; living arrangements; moderated mediation model; technological reserve hypothesis

Introduction

Background

Cognitive function is a critical determinant of dementia, functional independence, quality of life, and health care burden [1]. With an increasingly aged population, cognitive impairment and dementia have become major health and social issues worldwide [2]. Research has indicated that there are about 15.07 million patients with dementia in the population aged 60 and older in China, while the prevalence of mild cognitive impairment is 15.54%, and the number of patients is 38.77 million [3]. The disease burden of dementia and cognitive impairment is huge. The estimated total annual costs of dementia in China will reach 114.2 billion US dollars in 2030 [4]. Since there is currently no effective pharmaceutical treatment for dementia and mild cognitive impairment, it is important to identify modifiable intermediate risk factors that could prevent cognitive decline [5].

With rapid digitalization and the widespread integration of technology into daily life, digital technologies have emerged as a potential determinant of healthy aging. In China, as of 2025, the number of internet users aged 60 and older has reached 161 million, accounting for nearly 14.4% of all internet users [6]. Many studies have indicated that older people are competent and skilled users of digital technologies [7-9]. Consequently, the concept of digital engagement has been introduced to emphasize the breadth and extent of digital technology use among older people [10]. Every day, digital engagement provides new opportunities for older adults to address age-related cognitive decline. Engagement in cognitively challenging activities, such as learning new digital skills or knowledge, plays a protective role against age-related cognitive decline [11,12]. Meanwhile, access to communication technology and social media facilitates interpersonal interactions and enhances social support [13], which helps maintain cognitive health in older adulthood.

Against this background, exploring the association between digital technology use and cognitive health in later life has become an important research focus. However, the cognitive impact of digital technology use in China has not been sufficiently studied and understood [7,14]. First, research has focused on how access to the internet relates to cognitive function and the associations between use frequency in specific domains and cognitive outcomes [8,15-17]. Many studies in China have investigated the effect of internet or social media use on cognitive function [18,19]. But limited studies give attention to the concept of digital engagement [20] and comprehensively measure the dimensions and frequency of digital technology use. As digital technologies have become increasingly integrated into older adults' daily lives, it is important to shift research focus from use to meaningful digital engagement to better understand the cognitive effects of digital technology use [7].

Second, although evidence has established the efficacy of digital health interventions for cognitive decline and cognitive impairment, including dementia [21-23], little is known about how they lead to an improvement in symptoms or behavior. The identification of these mediating mechanisms would be useful for tailoring interventions that specifically target these pathways, improving intervention effectiveness. Some studies in China have estimated the mediating roles of physical activity [14] and social support [20,24]. However, few studies have simultaneously examined the roles of multiple mediators. Including multiple mediators can better reflect real-world mechanisms, help understand the relative importance of different intervention pathways, and reduce bias [25].

Third, while digital technologies become increasingly integrated into older adults' everyday life, growing urbanization and economic reforms in China have transformed intergenerational living arrangements patterns [26]. However, limited studies in China have examined how the association between digital technology use and cognitive function may vary by living arrangements. As the number of older adults living alone in China increases, examining the moderating role of living arrangements in this association is meaningful for developing targeted interventions.

The technological reserve hypothesis provides a theoretical framework for addressing these gaps. This hypothesis, developed by Benge and Scullin, focuses on how digital technology use can counteract cognitive decline and reduce disease burden [27-29]. Technological reserve is defined as "the development of a culture and environment of technology use in older adults that can buffer against the impact of cognitive decline on day-to-day activities" [27]. Further study developed the technological reserve concept and summarized 3 central pathways through which digital technology may prevent cognitive decline [28,30]. First, technology can generate cognitive complexity by engaging older adults in cognitively demanding activities that strengthen cognitive reserve [12,31,32]. By enabling access to diverse information sources (eg, online health information), promoting mentally stimulating activities, and requiring continual learning and adaptation, digital technologies help sustain and challenge cognitive capacities [33]. Second, technology fosters social connection and engagement, which are well-established protective factors against cognitive decline [34]. Through platforms such as social media, messaging apps, and video calls, older adults can maintain social ties, reduce loneliness, and access emotional and instrumental support. Finally, technologies can function as cognitive prosthetics by directly compensating for lapses in memory and executive function, particularly those involved in completing activities of daily living. For example, smartphone apps can deliver reminders for medication adherence [35].

Guided by the technological reserve hypothesis, this study aimed to examine the mediating effect of digital health literacy (DHL) and social support on the relationship between digital

XSL•FO
RenderX

engagement and cognitive function, as well as the moderating effect of living arrangements on the relationships among digital engagement, DHL, social support, and cognitive function.

Theoretical Framework

Digital Engagement and Cognitive Function

Within the technological reserve framework, digital technology use as a modifiable lifestyle behavior is a critical factor that can promote better cognitive outcomes than would be expected based on age, brain injury, or disease stage [30]. In this study, we adopted the term "digital engagement" to define digital technology use among older adults. Digital engagement among older adults refers to their use of contemporary digital technologies and devices to carry out routine and enjoyable everyday activities that support autonomy, independence, and social inclusion [8]. This concept emphasizes how older adults integrate information and communication technologies into daily activities and information-seeking behaviors rather than focusing on limitations [36]. Research has investigated the potential association between digital engagement and cognitive function among older adults. Although some studies suggest potential risks such as sleep disruption or social isolation [37-39], the prevailing evidence supports that digital engagement is positively linked to cognitive function [26,40,41]. Empirical findings generally suggest that regular use of digital technologies (such as social media and online social networking) is positively associated with better cognitive outcomes [42-44]. These benefits are often attributed to increased cognitive stimulation, enhanced social connectivity, and greater engagement in mentally active tasks afforded by digital technology [45,46]. Longitudinal studies further suggest that consistent internet use is associated with slower cognitive decline and a lower subsequent risk of dementia compared with nonuse [47,48]. Meta-analyses of randomized controlled trials also support the effectiveness of digital interventions in improving specific cognitive domains [49,50]. Despite growing evidence, the mechanisms through which digital engagement benefits cognition remain insufficiently understood.

The Mediating Role of DHL

DHL is the ability to seek, understand, evaluate, and apply health information from digital sources to support health-related decision-making [51]. Within the technological reserve framework, DHL can strengthen cognition by engaging older adults in cognitively demanding processes such as evaluating online resources, learning new digital skills, and applying health information in daily life. These processes involve active learning, adaptive reasoning, and problem-solving, which are consistent with mechanisms that sustain cognitive reserve [30]. Moreover, empirical studies support this pathway. Higher DHL is associated with greater adoption of preventive health behaviors, better management of chronic conditions, improved adherence to treatment, and more informed health decisions [52-55]. Such behaviors not only enhance health outcomes but also contribute to maintaining and preserving cognitive function in later life. Thus, DHL may mediate the association between digital engagement and cognitive function.

The Mediating Role of Social Support

Within the technological reserve framework, another plausible pathway operates through social connectivity. Social connectivity refers to the structural and functional aspects of individuals' social relationships, and in later life, is often reflected through social support received from their networks [56-58]. Socioemotional selectivity theory points out that social participation requires a certain cost investment, and members who engage in social participation are bound to consider cost-benefit issues [59]. For older adults, declining physical and cognitive abilities raise the cost of offline social participation, leading to a gradual reduction in face-to-face interactions [60]. Digital technologies offer alternative and more accessible avenues for maintaining social support [61]. Some empirical studies have shown that digital engagement is positively associated with increased social support in later life [61-63]. Social support, in turn, is a well-established protective factor for cognitive function: Older adults with stronger support networks tend to perform better cognitively and face a lower risk of cognitive decline or dementia [64-66]. Thus, social support may serve as a mediator linking digital engagement to cognitive outcomes.

The Moderating Role of Living Arrangements

Economic reforms and urbanization in China since the 1980s have profoundly reshaped family structures, particularly impacting older adults. This shift aligns with modernization theory, predicting smaller families and fewer older adults co-residing with children [67]. Consequently, more older adults live only with a spouse or alone [68]. Given the central role of family in Chinese culture, the study shifted to investigate the moderating role of living arrangements. Within the framework of the technological reserve hypothesis, the cognitive benefits of digital engagement are expected to vary across social contexts that shape baseline access to cognitive and social resources. Living arrangements represent a contextual factor in later life, as co-residence with others may provide routine cognitive stimulation and social interaction, whereas living alone is often associated with reduced offline engagement. Consequently, digital engagement may play a more pronounced compensatory role for older adults living alone by supplementing limited in-person cognitive and social resources. This theoretical perspective provides a rationale for examining living arrangements as a moderator in the association between digital engagement and cognitive function. Additionally, living arrangements may shape both the opportunities and the need for engaging with digital technology [69]. Older adults living alone often rely on digital technologies to maintain social ties, bridge social gaps, and manage independent living [70,71]. In contrast, those in multigenerational households may experience "proxy internet use" (eg, reliance on family members for online tasks), reducing direct engagement and the attendant cognitive stimulation [72]. Digital engagement may therefore be especially protective for those living alone.

Hypotheses

Guided by the technological reserve hypothesis, this study tested a moderated mediation model to examine whether digital engagement is associated with cognitive function through DHL

XSL•FO
RenderX

and social support, and whether these pathways are moderated by living arrangements. Based on the theoretical framework and prior empirical evidence, we propose the following hypotheses:

1. Hypothesis 1: higher digital engagement is correlated with greater cognitive function among older adults.
2. Hypothesis 2: higher digital engagement is associated with higher DHL among older adults.
3. Hypothesis 3: higher DHL is correlated with greater cognitive function among older adults.
4. Hypothesis 4: DHL is a mediator between digital engagement and cognitive function among older adults.
5. Hypothesis 5: higher digital engagement is associated with greater social support among older adults.
6. Hypothesis 6: greater social support is correlated with greater cognitive function among older adults.
7. Hypothesis 7: social support is a mediator between digital engagement and cognitive function among older adults in China.
8. Hypothesis 8: living arrangements moderate the associations of digital engagement with cognitive function, DHL, and social support.

To guide the analyses, we specified a directed acyclic graph (DAG) illustrating the hypothesized relationships among digital engagement, cognitive function, DHL, social support, and living arrangements (Figure 1).

Figure 1. The hypothesized moderated mediation model.

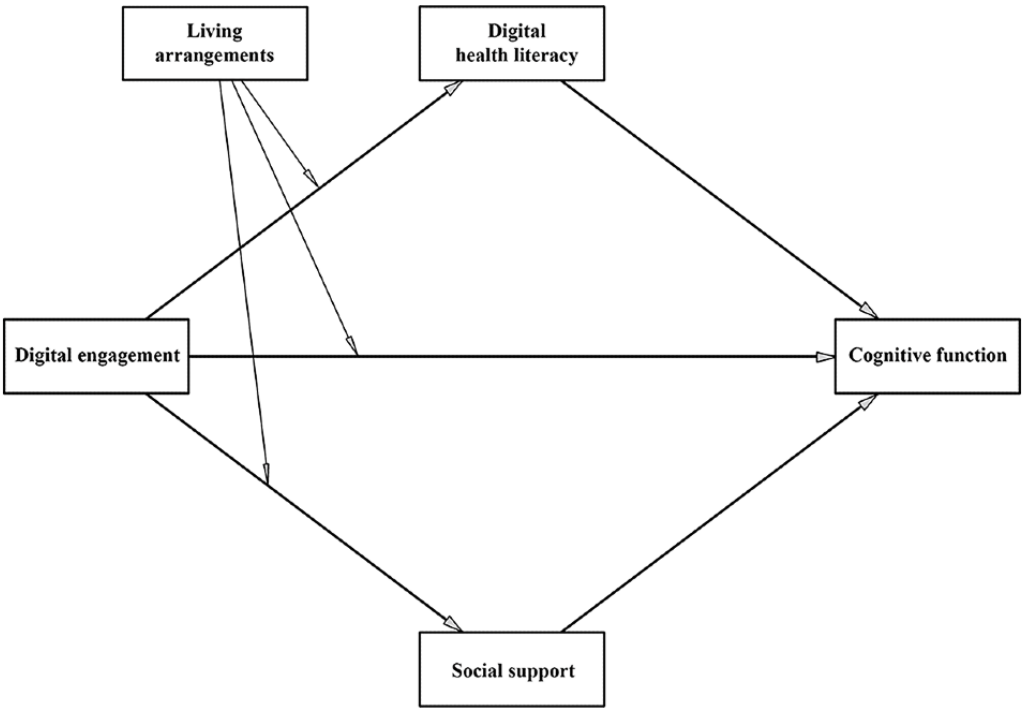

Methods

Study Design and Sampling Procedures

This study used data collected through a large-scale, cross-sectional survey conducted concurrently by 5 academic teams affiliated with 4 major universities in China. To ensure methodological uniformity, all participating sites adhered to a unified research protocol during the implementation phase.

A stratified, multistage sampling framework was used to enhance representativeness across regions with different levels of socioeconomic development. China was first stratified into eastern, central, and western regions, which reflect well-documented gradients in economic development, urbanization, and digital infrastructure. One to 2 provinces were randomly selected from each region. The final sample included Hubei (central China), Shandong and Jiangsu (eastern China), and Guangxi (western China), thereby capturing substantial regional heterogeneity in demographic structure and digital development. Within each selected province, 1 to 2 urban or county-level administrative units were further sampled based on local economic conditions, followed by cluster sampling of communities or villages.

Sample size estimation followed the standard formula for proportion-based calculations: , where u_α represents the critical value for a 95% CI (u_α=1.96), p is the estimated proportion of older internet users based on the China Internet Network Information Center's 51st Statistical Report [73], *q* is the complementary proportion (*q*=1–p), and *d* denotes the allowable error (1.2%). Based on these parameters, the minimum required sample size was calculated as 6616. To account for possible nonresponses and invalid questionnaires, a 20% oversampling rate was applied, resulting in a target sample of approximately 7940 individuals.

Data Collection and Quality Control

Fieldwork was conducted from June to November 2024 by trained surveyors in collaboration with local village committees or community service offices. Face-to-face interviews were

XSL•FO
RenderX

administered at participants' homes using standardized paper questionnaires. The interviews collected information on sociodemographic characteristics, digital technology use, digital literacy, cognitive function, and quality of life. All surveyors received centralized training to ensure consistent questionnaire administration and interpretation. Upon completion of each interview, field supervisors performed a thorough review of the questionnaires to check for completeness, internal consistency, and data accuracy before submission for entry.

Participants

Eligible participants were older adults who met the following inclusion criteria: (1) aged 55 years and older, (2) had resided in the sampled community or village for at least 6 months, and (3) were able to communicate effectively with investigators. Exclusion criteria included: (1) individuals temporarily absent from their households during the survey period, (2) those diagnosed with terminal illnesses, and (3) those who declined to participate. After excluding incomplete responses, duplicate entries, and respondents younger than 55 years, 8302 valid questionnaires remained. Among these, 179 participants (2.2%) had missing values on at least 1 analytic variable and were excluded from the main analyses. The final analytic sample consisted of 8123 participants (Figure 2).

XSL•FO
RenderX

Figure 2. Flow diagram of participant recruitment and data exclusion, resulting in a final analytic sample of N=8123.

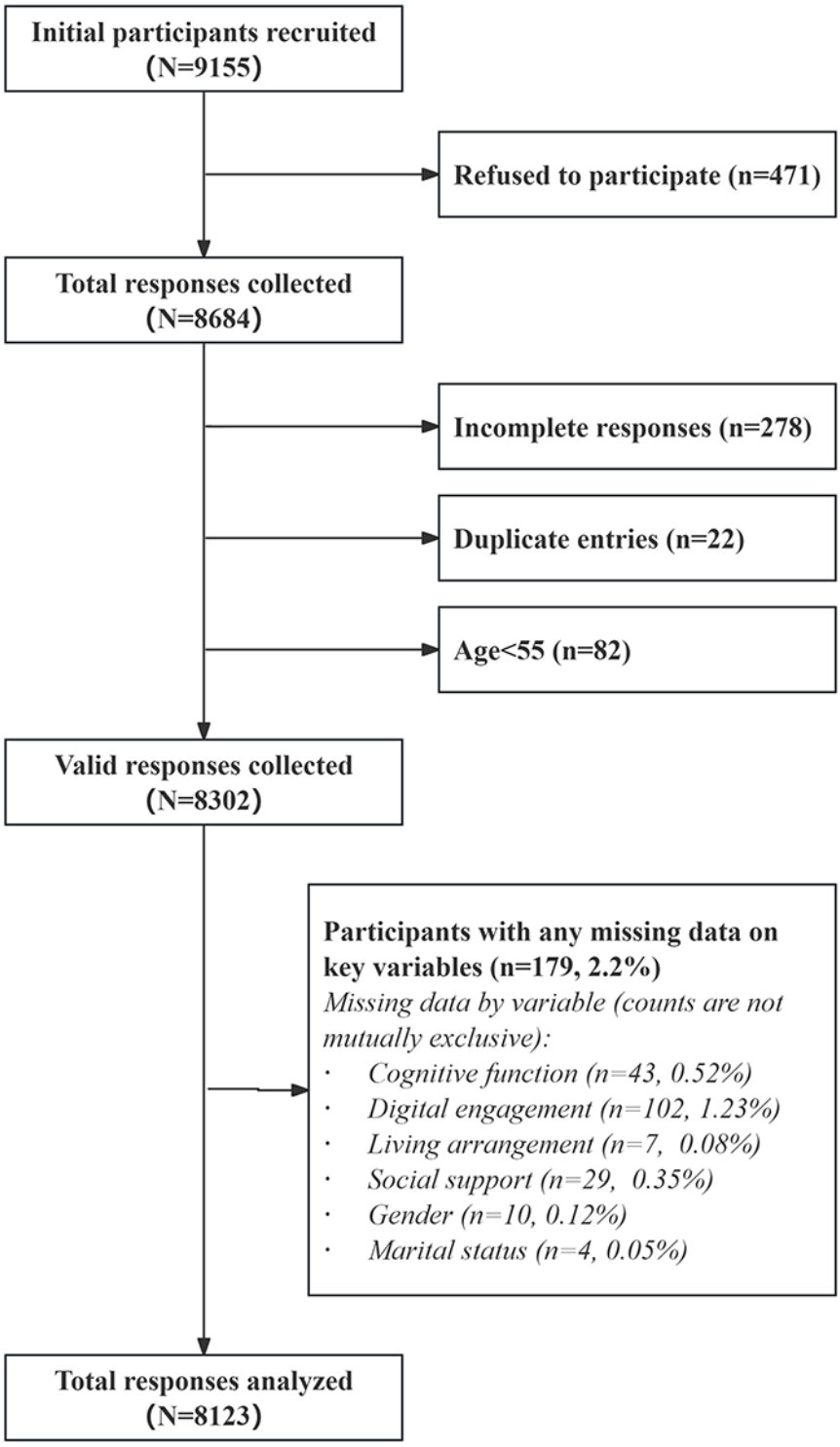

Measurements

Cognitive Function

Cognitive function was assessed using the Mini-Cog test. The Mini-Cog test is a rapid, valid, and reliable screening tool for cognitive impairment [74]. The Mini-Cog Test includes a 3-word recall task (scored 0-3) and the clock drawing test (scored 0-2). The total score ranges from 0 to 5. The Mini-Cog test has demonstrated good screening performance in community-dwelling older adults in China [75]. In this study, the Mini-Cog total score was used as a continuous measure of cognitive function, with higher scores indicating better cognitive performance.

Digital Engagement

Digital engagement was measured using a self-reported scale developed to capture older adults' frequency of participation in various digital activities. The scale included eight items of

XSL•FO
RenderX

digital behaviors: (1) social communication (eg, using WeChat [Tencent] voice or video calls), (2) experience sharing (eg, posting on WeChat Moments, QQ Zone [Tencent], or Weibo [Sina Corporation]), (3) leisure and entertainment (eg, playing online games, listening to music, or watching videos), (4) online transactions (eg, transferring money, making payments, booking services, or trading stocks), (5) information seeking (eg, searching for travel information or reading news), (6) online learning or training, (7) online civic participation (eg, participating in online polls, petitions, or rights protection), and (8) political engagement (eg, online voting or leaving messages on government websites). Participants rated the frequency of each activity on a 5-point Likert scale ranging from 1 (never) to 5 (always). Higher scores indicated higher digital engagement. The scale demonstrated good internal consistency in this sample (Cronbach α=0.876). Although the scale covers multiple items of digital activities, this study conceptualized digital engagement as an overall behavioral tendency reflecting the breadth and extent of digital technology use in daily life. This approach is consistent with the technological reserve hypothesis and the concept of digital engagement, which emphasize cumulative and sustained engagement. Therefore, a composite digital engagement score was used in the analyses.

DHL

DHL was assessed using the eHealth Literacy Scale (eHEALS), a widely validated instrument developed by Norman and Skinner to measure individuals' self-perceived skills in locating, evaluating, and applying electronic health information to health-related problems [51]. The eHEALS consists of 8 items rated on a 5-point Likert scale (1=strongly disagree to 5=strongly agree), reflecting domains such as awareness of available online health resources, confidence in using the internet for health decision-making, and the ability to discern high-quality digital health content. Given the linguistic and cultural differences between the original instrument and the target population of older adults in mainland China, we used the simplified Chinese version (C-eHEALS) translated and validated by Ma and Wu [76]. The C-eHEALS has been confirmed to have good psychometric properties and can therefore be used to evaluate eHealth literacy in Chinese older populations [77]. In this study, the C-eHEALS demonstrated excellent internal consistency, with a Cronbach α coefficient of 0.986, indicating high reliability for use among Chinese older adults.

Social Support

Social support was assessed using the 3-item short version of the Perceived Social Support Scale (PSSS-3), which was developed and validated by Wu et al [78] specifically for use among the Chinese general population. This abbreviated scale was derived from the original 12-item Chinese version of the Multidimensional Scale of Perceived Social Support (MSPSS), originally adapted by Jiang [79] from the version developed by Zimet et al [80]. The PSSS-3 includes 1 item from each of the 3 core dimensions, family support, friend support, and significant others, selected based on the highest factor loadings in a large-scale national sample. The Cronbach α of PSSS-3 in this study was 0.868, demonstrating good internal consistency.

Living Arrangements

Living arrangements were measured as a binary variable indicating whether the older adult lived alone, and were assessed using the following question: "What are your current living arrangements?" Those who reported living alone were coded as 1, and those who reported living with others were coded as 0.

Control Variables

The prior study indicates that demographic and health factors have close links with cognitive function and suggests that these factors should be included in pertinent research [81]. In this study, gender, age, current place of residence, marital status, education, and number of chronic diseases were controlled as covariates.

Statistical Analysis

All analyses were performed using SPSS (version 27; IBM Corp). We first examined the extent and pattern of missing data for all analytic variables. The proportion of missing values for each variable ranged from 0.08% to 1.2% and the overall proportion of missing data was 2.2% (Figure 2). Little's Missing Completely at Random test was conducted using the missing value analysis procedure in SPSS 27. The test indicated that the missing values were independent of the observed or unobserved values (χ^2=14.893; P=.06). Given the low proportion and completely random patterns of missingness, we performed complete-case analyses based on listwise deletion.

Descriptive statistics summarized sample characteristics, with continuous variables reported as mean (SD) and categorical variables as frequencies and percentages. Pearson correlations examined associations among digital engagement, cognitive function, DHL, social support, and living arrangements. Multicollinearity was assessed by the variance inflation factor (VIF), with VIF>5 indicating collinearity. The relationships among variables were specified according to a DAG based on the technological reserve hypothesis (Figure 1) and analyzed using PROCESS models 4 and 8 with 5000 bootstrapped resamples. Effects were considered significant if the 95% bias–corrected CI did not include 0. All models controlled for age, gender, current place of residence, marital status, education, and number of chronic diseases. Continuous variables were standardized prior to analysis. Statistical tests were 2-tailed with α=.05.

Ethical Considerations

The study was reviewed and approved by the Medical Ethics Committee of Zhongda Hospital, Southeast University (approval number 2024ZDSYLL294-Y01). Written informed consent was obtained from all participants before they participated in the study, and they were provided with the opportunity to withdraw at any time during and after the survey. To protect privacy and confidentiality, electronic data were de-identified and stored on password-protected devices accessible only to the research team. No images or supplementary materials in this manuscript contain information that could identify individual participants. There was no compensation for the participants in our study survey.

XSL•FO
RenderX

Results

Demographic Characteristics of the Participants

The final sample consisted of 8123 older adults, with an average age of 71.03 (SD 8.39) years. Men accounted for 42.46% (3449/8123) of the participants. Among the participants, 3990 (49.12%) lived in urban areas, and 4133 (50.88%) lived in rural areas. Most respondents were married (6338/8123, 78.03%), and 51.73% (4202/8123) of the participants had attained a middle school education or above. Overall, 75.28% (6115/8123) of the participants reported at least 1 chronic condition. Summary statistics for main variables, including digital engagement, DHL, social support, living arrangements, and cognitive function, are presented in Table 1.

Table 1. Participant characteristics and descriptive statistics for study variables (N=8123).

Variables	Values
Sex, n (%)	
Male	3449 (42.46)
Female	4674 (57.54)
Age (years), mean (SD)	71.03 (8.39)
Current place of residence, n (%)	
Urban	3990 (49.12)
Rural	4133 (50.88)
Marital status, n (%)	
Married	6338 (78.03)
Unmarried	1785 (21.97)
Education, n (%)	
Primary school or under	3921 (48.27)
Middle or high school	3462 (42.62)
College or above	740 (9.11)
Chronic diseases, n (%)	
0	2008 (24.72)
1	3477 (42.80)
≥2	2638 (32.48)
Digital engagement, mean (SD)	17.37 (8.03)
DHL[a], mean (SD)	20.18 (10.59)
Social support, mean (SD)	16.71 (3.22)
Living arrangements, n (%)	
Living alone	1166 (14.35)
Living with others	6957 (85.65)
Cognitive function, mean (SD)	3.57 (1.49)

[a]DHL: digital health literacy.

Preliminary Correlation Analysis

Table 2 presents the Pearson correlation coefficients among the main variables. Digital engagement was significantly and positively correlated with cognitive function (r=0.365, P<.001), social support (r=0.081, P<.001), and DHL (r=0.575, P<.001), but negatively associated with living arrangements (r=–0.101, P<.001). Cognitive function was also positively correlated with social support (r=0.131, P<.001) and DHL (r=0.347, P<.001), and negatively correlated with living arrangements (r=–0.069, P<.001). VIFs indicated no multicollinearity among digital engagement (VIF=1.732), DHL (VIF=1.666), social support (VIF=1.027), living arrangements (VIF=1.823), and cognitive function.

XSL•FO
RenderX

Table 2. Pearson correlation matrix of digital engagement, digital health literacy, social support, living arrangements, and cognitive function (N=8123).

Variables	Digital engagement	DHL[a]	Social support	Living arrangements	Cognitive function
Digital engagement					
r	1	0.575	0.081	–0.101	0.365
P value	__[b]	<.001	<.001	<.001	<.001
DHL					
r	0.575	1	0.127	–0.073	0.347
P value	<.001	—	<.001	<.001	<.001
Social support					
r	0.081	0.127	1	–0.045	0.131
P value	<.001	<.001	—	<.001	<.001
Living arrangements					
r	–0.101	–0.073	–0.045	1	–0.069
P value	<.001	<.001	<.001	—	<.001
Cognitive function					
r	0.365	0.347	0.131	–0.069	1
P value	<.001	<.001	<.001	<.001	—

[a]DHL: digital health literacy.

[b]Not applicable.

Mediation Analysis

Guided by the DAG-specified conditional process model, we first used model 4 of the PROCESS macro for SPSS [82] to test Hypotheses 1 to 7. The total effect of digital engagement on cognitive function was 0.241 (95% CI 0.216-0.265), of which the direct effect accounted for 70.5% (β=0.170, 95% CI 0.143-0.196) and the combined indirect effects by DHL and social support accounted for 29.5% (β=0.071, 95% CI 0.061-0.082). The findings indicated a moderate but statistically meaningful mediation by DHL and social support.

Table 3 and Figure 3 present the results of the mediation analysis. As shown in model 3, the direct effect of digital engagement on cognitive function was significant (Model 3: β=0.170, 95% CI 0.143-0.196; *P*<.001), thus supporting Hypothesis 1. Digital engagement also demonstrated a significant and positive association with DHL (Model 1: β=0.400, 95% CI 0.379-0.420; *P*<.001), supporting Hypothesis 2. Additionally, DHL was significantly and positively related to cognitive function (Model 3: β=0.148, 95% CI 0.123-0.174; *P*<.001), supporting Hypothesis 3. Furthermore, digital engagement was positively and significantly correlated with social support (Model 2: β=0.129, 95% CI 0.103-0.155; *P*<.001), supporting Hypothesis 5. Social support, in turn, showed a significant and positive correlation with cognitive function (Model 3: β=0.091, 95% CI 0.071-0.112; *P*<.001), supporting Hypothesis 6.

XSL•FO
RenderX

Table 3. Mediation analysis of the association between digital engagement and cognitive function through digital health literacy and social support, adjusted for covariates (PROCESS Model 4; N=8123).

Model 1[a] (DHL[b])			Model 2[c] (social support)		Model 3[d] (cognitive function)	
	β[e] (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value
Explanatory variable						
Digital engagement	0.400 (0.379 to 0.420)	<.001	0.129 (0.103 to 0.155)	<.001	0.170 (0.143 to 0.196)	<.001
Mediator variables						
DHL	__[f]	—	—	—	0.148 (0.123 to 0.174)	<.001
Social support	—	—	—	—	0.091 (0.071 to 0.112)	<.001
Control variables						
Gender	–0.035 (–0.070 to –0.001)	.047	0.169 (0.125 to 0.213)	<.001	–0.044 (–0.085 to –0.003)	.03
Age	–0.041 (–0.060 to –0.021)	<.001	0.065 (0.041 to 0.089)	<.001	–0.095 (–0.117 to –0.072)	<.001
Current place of residence	0.321 (0.280 to 0.362)	<.001	–0.465 (–0.518 to –0.413)	<.001	0.048 (–0.003 to 0.098)	.06
Marital status	0.039 (–0.004 to 0.082)	.076	0.159 (0.104 to 0.213)	<.001	0.134 (0.084 to 0.184)	<.001
Education	0.294 (0.261 to 0.327)	<.001	0.172 (0.131 to 0.213)	<.001	0.179 (0.140 to 0.218)	<.001
Chronic diseases	–0.124 (–0.147 to –0.101)	<.001	–0.078 (–0.107 to –0.048)	<.001	0.034 (0.007 to 0.062)	.01
Constant	–0.471 (–0.563 to –0.380)	<.001	–0.353 (–0.469 to –0.237)	<.001	–0.383 (–0.491 to –0.275)	<.001

[a]$F_{7, 8115}$=811.099; R^2=0.412.

[b]DHL: digital health literacy.

[c]$F_{7, 8115}$=65.515; R^2=0.054.

[d]$F_{9, 8113}$=214.765; R^2=0.192.

[e]β: standardized regression coefficient.

[f]Not applicable.

XSL•FO
RenderX

Figure 3. Mediation model of the association between digital engagement and cognitive function through digital health literacy and social support; values are standardized coefficients adjusted for covariates (PROCESS Model 4; N=8123; ***P<.001).

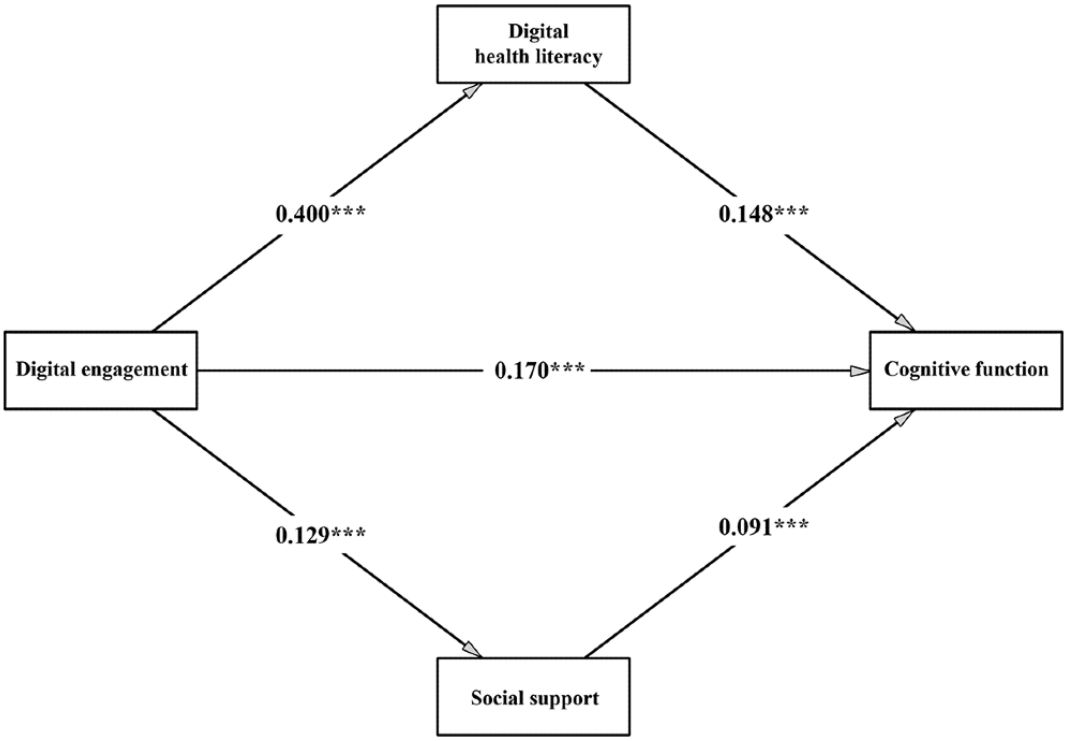

These findings suggest that DHL and social support play a partial mediating role in the relationship between digital engagement and cognitive function, with indirect effects of 0.059 (95% CI 0.049-0.069) and 0.012 (95% CI 0.008-0.016), respectively, supporting Hypotheses 4 and 7. Combining 2 mediation effects, the total indirect effect was 0.071 (95% CI 0.061-0.082). The bootstrap test results for indirect effects are reported in Table 4.

Table 4. Bootstrap estimates of indirect effects of digital engagement on cognitive function through digital health literacy and social support (PROCESS Model 4; N=8123).

Indirect effects path	Effects	SE	95% CI	Proportion of effects (%)
Total indirect effect	0.071	0.005	0.061-0.082	100
DE[a]→DHL[b]→CF[c]	0.059	0.005	0.049-0.069	83.10
DE→SS[d]→CF	0.012	0.002	0.008-0.016	16.90

[a]DE: digital engagement.
[b]DHL: digital health literacy.
[c]CF: cognitive function.
[d]SS: social support.

Moderated Mediation Analysis

To examine the moderated mediation effects involving living arrangements, we used Model 8 of the PROCESS macro for SPSS [82], using 5000 bootstrap resamples and a 95% bias-corrected CI. The results are reported in Table 5. The analysis revealed a significant and positive interaction effect between digital engagement and living arrangements on cognitive function (Model 6: β=0.109, 95% CI 0.052-0.166; P<.001), suggesting a moderating role of living arrangements. As illustrated in Figure 4, the beneficial association between digital engagement and cognitive performance was stronger among older adults who lived alone, relative to those who lived with others.

XSL•FO
RenderX

Table 5. Moderated mediation analysis testing moderation by living arrangements in the associations between digital engagement and digital health literacy, digital engagement and social support, digital engagement and cognitive function, adjusted for covariates (PROCESS Model 8; N=8123).

Variables	Model 4[a] (DHL[b])		Model 5[c] (social support)		Model 6[d] (cognitive function)	
	β[e] (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value
Explanatory variable						
Digital engagement	0.391 (0.370 to 0.413)	<.001	0.109 (0.082 to 0.137)	<.001	0.156 (0.129 to 0.183)	<.001
Mediator variables						
DHL	__[f]	—	—	—	0.147 (0.121 to 0.172)	<.001
Social support	—	—	—	—	0.089 (0.069 to 0.110)	<.001
Moderating variables						
Living arrangements	0.054 (–0.011 to 0.119)	.10	0.058 (–0.024 to 0.141)	.17	0.154 (0.078 to 0.230)	<.001
DE×LA[g]	0.063 (0.014 to 0.112)	.01	0.151 (0.089 to 0.212)	<.001	0.109 (0.052 to 0.166)	<.001
Control variables						
Gender	–0.036 (–0.070 to –0.001)	.044	0.167 (0.123 to 0.211)	<.001	–0.044 (–0.085 to –0.003)	.04
Age	–0.040 (–0.060 to –0.021)	<.001	0.065 (0.041 to 0.089)	<.001	–0.094 (–0.117 to –0.072)	<.001
Current place of residence	0.321 (0.280 to 0.362)	<.001	–0.465 (–0.517 to –0.413)	<.001	0.047 (–0.003 to 0.097)	.07
Marital status	0.063 (0.007 to 0.119)	.03	0.175 (0.103 to 0.247)	<.001	0.210 (0.144 to 0.276)	<.001
Education	0.292 (0.259 to 0.324)	<.001	0.167 (0.126 to 0.208)	<.001	0.176 (0.137 to 0.215)	<.001
Chronic diseases	–0.123 (–0.146 to –0.100)	<.001	–0.077 (–0.106 to –0.047)	<.001	0.034 (0.007 to 0.061)	.01
Constant	–0.492 (–0.591 to –0.393)	<.001	–0.360 (0.485 to –0.234)	<.001	–0.456 (–0.572 to –0.339)	<.001

[a]$F_{9, 8113}$=632.208; R^2=0.412.

[b]DHL: digital health literacy.

[c]$F_{9, 8113}$=53.680; R^2=0.056.

[d]$F_{11, 8111}$=178.558; R^2=0.195.

[e]β: standardized regression coefficient.

[f]Not applicable.

[g]DE×LA: the interaction term between digital engagement and living arrangements.

XSL•FO
RenderX

Figure 4. The moderating effect of living arrangements on the association between digital engagement and cognitive function. The difference in simple slopes indicated that the association between digital engagement and cognitive function was significantly stronger for individuals living alone than for those living with others (PROCESS Model 8, N=8123).

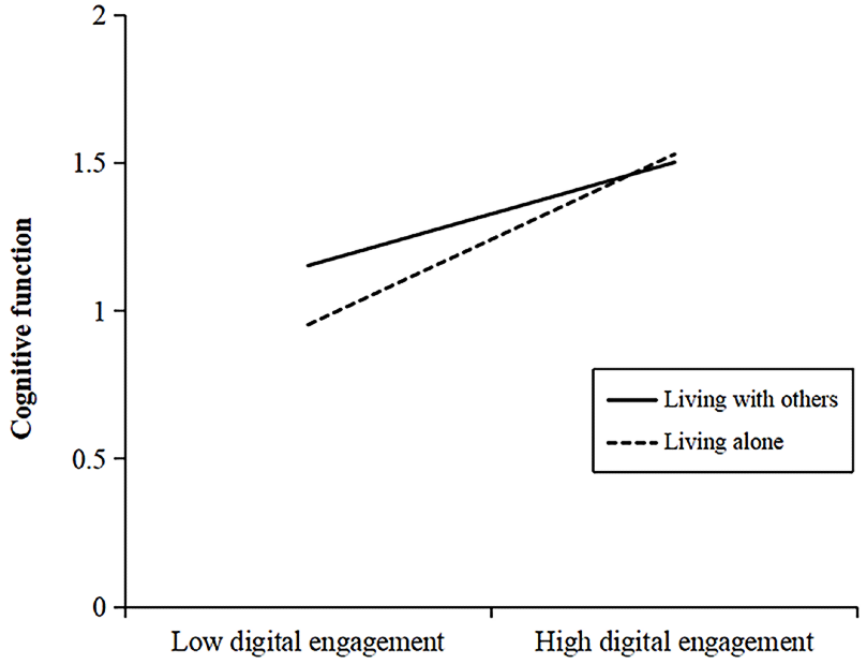

In addition to the interaction effect on the direct path, the results also revealed significant moderating effects of living arrangements on the first stages of both mediation pathways. Specifically, the interaction term between digital engagement and living arrangements significantly predicted DHL (Model 4: β=0.063, 95% CI 0.014-0.112; *P*=.01 and social support (Model 5: β=0.151, 95% CI 0.089-0.212; *P*<.001). As illustrated in Figures 5 and 6, the beneficial association of digital engagement with both DHL and social support was stronger among older adults who lived alone, relative to those who lived with others. These findings indicate that living arrangements moderate the associations of digital engagement with cognitive function, DHL, and social support, supporting Hypothesis 8.

Figure 5. The moderating effect of living arrangements on the association between digital engagement and digital health literacy. The difference in simple slopes indicated that the association between digital engagement and digital health literacy was significantly stronger for individuals living alone than for those living with others (PROCESS Model 8, N=8123).

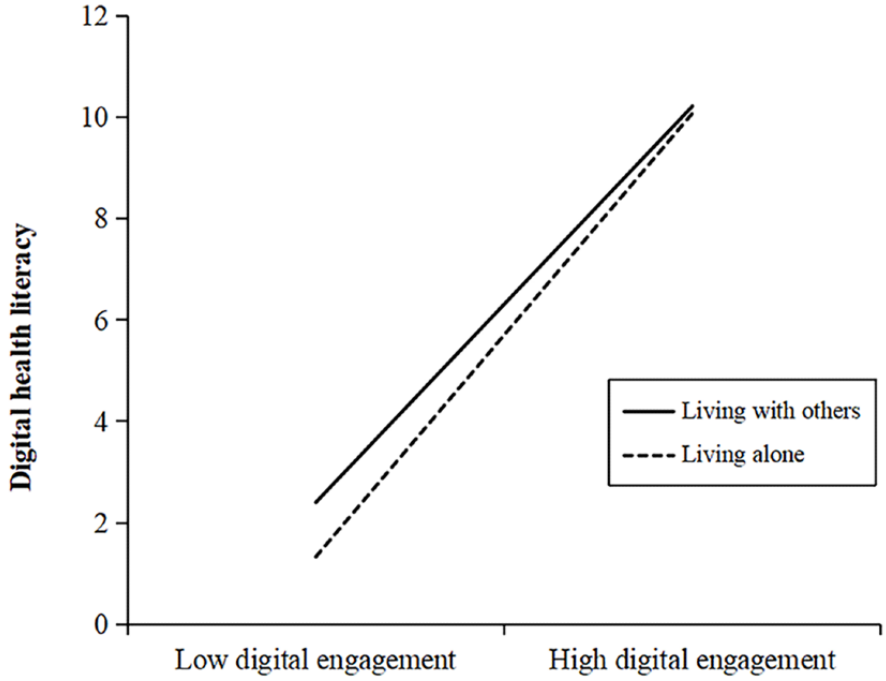

XSL•FO
RenderX

Figure 6. The moderating effect of living arrangements on the association between digital engagement and social support. The difference in simple slopes indicated that the association between digital engagement and social support was significantly stronger for individuals living alone than for those living with others (PROCESS Model 8, N=8123).

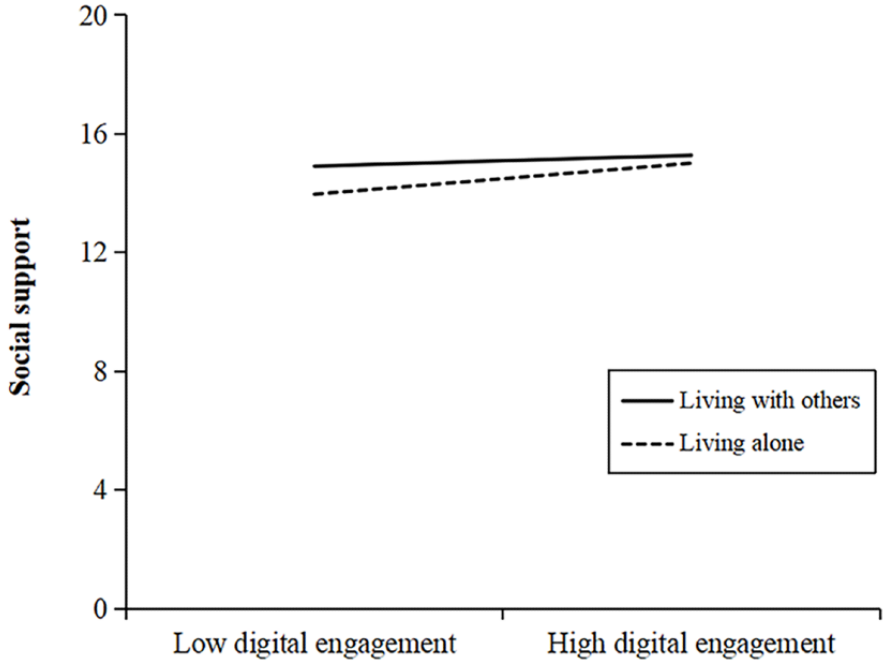

Table 6 reports the conditional indirect effects of digital engagement on cognitive function by each path. For the DHL mediator, the indirect effect of digital engagement on cognitive function was 0.057 (95% CI 0.048-0.068) among participants living with others and 0.067 (95% CI 0.054-0.080) among those living alone. The corresponding index of moderated mediation was 0.009 (95% CI 0.003-0.016), indicating a significantly stronger indirect effect by DHL for participants living alone. Similarly, for the social support mediator, the indirect effect was 0.010 (95% CI 0.007-0.013) for those living with others and 0.023 (95% CI 0.016-0.032) for those living alone. The index of moderated mediation for the social support pathway was 0.013 (95% CI 0.007-0.021). In all cases, the 95% CI values excluded zero, indicating that both indirect effects were significantly stronger among older adults living alone.

Table 6. Indices of moderated mediation for two conditional indirect effects of digital engagement on cognitive function by living arrangements (PROCESS Model 8; N=8123).

Conditional indirect effects path	Effects	SE	95% CI
DE[a]→DHL[b]→CF[c] (living with others)	0.057	0.006	0.048-0.068
DE→DHL→CF (living alone)	0.067	0.007	0.054-0.080
Index of the moderated mediation	0.009	0.003	0.003-0.016
DE→SS[d]→CF (living with others)	0.010	0.002	0.007-0.013
DE→SS→CF (living alone)	0.023	0.004	0.016-0.032
Index of the moderated mediation	0.013	0.003	0.007-0.021

[a]DE: digital engagement.
[b]DHL: digital health literacy.
[c]CF: cognitive function.
[d]SS: social support.

Discussion

Principal Findings

Guided by the technological reserve hypothesis and using a large, community-based sample of older adults in China, this study investigated the mechanisms underlying the association between digital engagement and cognitive function among older Chinese adults. We found that higher digital engagement was associated with better cognitive performance. DHL and social support partially mediated this association, and the combined indirect effects accounted for 29.5% of this association. Living arrangements moderated both the direct and indirect pathways, with stronger benefits among older adults living alone. These findings extend prior work on technology use and cognition in later life and broaden the application of the technological reserve hypothesis in the Chinese context.

XSL•FO
RenderX

The Association Between Digital Engagement and Cognitive Function

Digital engagement was significantly and positively associated with cognitive function. This supports the technological reserve hypothesis that digital technology use is a modifiable behavioral factor that can promote better cognitive outcomes [30]. It also aligns with prior work linking internet or computer use to cognition [83-88] and with recent studies in China demonstrating that internet use and digital activities enhance cognitive function [14,47]. By adopting the construct of digital engagement rather than a simple use-versus-nonuse dichotomy, our study advances the field by situating technology use within the everyday life context of older adults, emphasizing how they integrate information and communication technologies into their ongoing activities, social interaction, and information seeking [36]. This finding implies that encouraging sustained and meaningful digital engagement is a promising strategy for public health and aging policies aiming to strengthen cognitive function among older adults.

The Mediating Role of DHL and Social Support

Consistent with the cognitive-stimulation pathway posited by the technological reserve hypothesis, higher digital engagement was associated with higher DHL, which in turn related to better cognitive performance [28,30]. Notably, the DHL pathway accounted for the majority of the total indirect effect, indicating that health-related digital competencies may be a primary mechanism linking engagement to cognition. Specifically, engaging with digital technology improves DHL because digital skills are among the core skills of DHL [89,90]. In turn, higher DHL denotes a stronger capacity to seek, understand, appraise, and apply health information [91]. These processes involve engaging with cognitively complex information [30], helping older adults build cognitive reserve. This finding advances the eHealth Literacy Model, which posits that DHL is underpinned by cognition [92], and indicates that DHL also serves as a tool that shapes cognition through ongoing, cognitively complex digital activities. Accordingly, interventions should combine user-friendly interfaces with structured, progressive training in cognitively complex digital tasks, ensuring that everyday digital engagement serves as sustained cognitive stimulation. Practical examples include stepwise smartphone or tablet training delivered in community settings (eg, locating health information from reliable sources, evaluating credibility and misinformation, and applying information to everyday self-management tasks).

Social support mediated the association between digital engagement and cognition, consistent with the social-connectivity pathway [28,30]. This finding aligns with previous research showing that digital engagement has the potential to enhance cognitive function among older individuals by addressing feelings of loneliness and improving the social support they receive from relatives and friends [20]. Specifically, digital engagement enables cheap and easy communication between older adults in distant communities, increasing social connections, overcoming social and spatial barriers, and providing a convenient way to stay in touch with families, friends, and the outside world [62]. In turn, better social support is associated with better cognitive outcomes in older adults [65,93-96]. This finding underscores that interventions should help older adults form and maintain digital social ties so that online interactions translate into perceived social support and, ultimately, better cognitive outcomes. For example, programs could incorporate facilitated online peer groups and a “Digital Buddy system” to help older adults translate online interactions into perceived support [97].

The Moderating Role of Living Arrangements

Our study further revealed that living arrangements played a significant moderating role in the associations of digital engagement with cognitive function, DHL, and social support. Compared with older adults who live with others, those living alone experienced a significantly stronger positive effect of digital technology engagement on cognitive function, consistent with the previous studies [26,98]. This moderating effect was significantly present in both mediating pathways: older adults living alone gained greater benefits in terms of DHL and social support from digital engagement than those living with others. Specifically, older adults living alone, due to a lack of effective offline social interactions, are more reliant on virtual social networks facilitated by digital technologies [60]. This reliance partially compensates for the reduced social support associated with solitary living, thereby mitigating its negative impact on cognitive function [70]. Additionally, older adults who live alone are less likely to engage in proxy internet use [72] and thus rely more on themselves to use digital devices (eg, searching for health information online). Furthermore, since older adults living alone are less often burdened with caregiving responsibilities for grandchildren, they have more freedom and time to engage with digital technologies [26]. Our findings suggest that digital engagement serves as a more efficacious strategy for mitigating cognitive decline among older adults living alone compared to those living with others.

Our finding is broadly consistent with other international evidence. A cohort study in America reported that transitioning into Internet use was associated with better cognitive function and slower cognitive decline, and that these benefits were more pronounced among older adults living alone than among those living with others [84]. Additionally, a 2-country longitudinal study in Sweden and the Netherlands observed less decline in global cognition among baseline internet users after adjustment for living situation [86]. Beyond cognitive outcomes, findings based on the Survey of Health, Ageing and Retirement in Europe further indicate that internet use can attenuate the association between living alone and loneliness across different European welfare regimes, implying that digital engagement may buffer psychosocial vulnerabilities of solitary living [99]. Taken together, although the prevalence and social meaning of living alone differ across cultures, converging evidence supports that digital engagement may serve as a compensatory resource for older adults with constrained offline or household-based support.

This moderation finding has practical implications for intervention design. Digital inclusion initiatives to help older people adapt to digital technologies should prioritize this vulnerable group. An integrated community-based approach may be especially useful: individual digital coaching (eg, guided practice in health information seeking) coupled with structured

XSL•FO
RenderX

social support (eg, online group chats). Including a "living alone" priority within such initiatives may help maximize equity and potential cognitive benefits, while also addressing social isolation risks that have been recognized as a public health and policy concern.

Limitations and Future Research

Despite these contributions, several limitations should be acknowledged. First, the cross-sectional design precludes causal inferences. Although our models adjusted for a set of covariates, endogeneity, reverse causation, and unmeasured confounding cannot be fully ruled out. Thus, our results only show associations, not causality. Future studies should use longitudinal designs with extended follow-up periods to elucidate temporal dynamics and disentangle potential reverse causation. Second, reliance on self-reported measures introduces the risk of recall bias, especially among participants with cognitive impairments, despite our use of validated instruments to attenuate this issue. Third, digital engagement was operationalized as a composite measure. Thus, this study could not disentangle potentially differential effects of specific types of digital activities on cognitive function. Finally, due to data constraints, living arrangements were operationalized as solitary versus nonsolitary living, precluding differentiation among various household compositions (eg, living with a spouse, children, or extended family). Given the important role of family structures in the well-being of older adults in China, future research should refine classifications of living arrangements to better explore their moderating effects.

Conclusions

This study contributes to the understanding of cognitive aging in the digital environment from the perspective of the technological reserve hypothesis and digital engagement. First, it offers an innovative framework based on the technological reserve hypothesis for understanding the moderating and mediating mechanisms of DHL, social support, and living arrangements. Second, it advances previous assessment methods of digital technology application by using a comprehensive measure. Our results increase understanding of the mechanisms underlying the cognitive effects of digital technology use and provide insights for designing digital health interventions and public health policies.

Acknowledgments

The authors would like to express their sincere gratitude to all participants in this study, as well as to the staff of the community and village committees who facilitated the fieldwork. This study was supported by the Major Project of the National Social Science Fund of China (grant number 23&ZD188). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to restrictions under existing Data Use Agreements, but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Wang Q, Ni J, Guan Y, Liu X, Li M, Xue H, et al. Associations of depression and symptomatic knee osteoarthritis with cognitive function among middle-aged and older adults: evidence from CHARLS in China. J Gerontol Soc Work Routledge 2025 Jul 28:1-17. [doi: 10.1080/01634372.2025.2539741] [Medline: 40719179]
2. World Alzheimer report 2024: global changes in attitudes to dementia. Alzheimer's Disease International. 2024. URL: https://www.alzint.org/resource/world-alzheimer-report-2024/ [accessed 2025-07-25]
3. Jia L, Du Y, Chu L, Zhang Z, Li F, Lyu D, COAST Group. Prevalence, risk factors, and management of dementia and mild cognitive impairment in adults aged 60 years or older in China: a cross-sectional study. Lancet Public Health 2020 Dec;5(12):e661-e671 [FREE Full text] [doi: 10.1016/S2468-2667(20)30185-7] [Medline: 33271079]
4. Xu J, Wang J, Wimo A, Fratiglioni L, Qiu C. The economic burden of dementia in China, 1990-2030: implications for health policy. Bull World Health Organ 2017;95(1):18-26 [FREE Full text] [doi: 10.2471/BLT.15.167726] [Medline: 28053361]
5. Shah H, Albanese E, Duggan C, Rudan I, Langa KM, Carrillo MC, et al. Research priorities to reduce the global burden of dementia by 2025. Lancet Neurol 2016 Nov;15(12):1285-1294. [doi: 10.1016/S1474-4422(16)30235-6] [Medline: 27751558]

XSL•FO
RenderX

6. The 56th statistical report on China's internet development [Web page in Chinese]. China Internet Network Information Center. URL: https://www3.cnnic.cn/n4/2025/0721/c88-11328.html [accessed 2025-07-31]
7. Kebede A, Ozolins L, Holst H, Galvin K. Digital engagement of older adults: scoping review. J Med Internet Res 2022;24(12):e40192 [FREE Full text] [doi: 10.2196/40192] [Medline: 36477006]
8. Aleti T, Figueiredo B, Reid M, Martin D, Sheahan J, Hjorth L. Older adults' digital competency, digital risk perceptions and frequency of everyday digital engagement. Inf Technol People 2025;38(8):118. [doi: 10.1108/itp-05-2024-0624]
9. Xu L, Ng D, Lee C, Peng P, Chu S. A systematic review of digital literacy in lifelong learning for older adults: challenges, strategies, and learning outcomes. Education Tech Research Dev 2025;73(6):3627-3674. [doi: 10.1007/s11423-025-10530-w]
10. Iqbal S, Fischl C, Asai R. Older persons' social participation, health and well-Being through digital engagement. Act Adapt Aging Routledge 2025;49(4):534-563. [doi: 10.1080/01924788.2025.2512299]
11. Chan M, Haber S, Drew L, Park D. Training older adults to use tablet computers: does it enhance cognitive function? Gerontologist 2016;56(3):475-484 [FREE Full text] [doi: 10.1093/geront/gnu057] [Medline: 24928557]
12. Almeida-Meza P, Steptoe A, Cadar D. Is engagement in intellectual and social leisure activities protective against dementia risk? Evidence from the English Longitudinal Study of Ageing. J Alzheimers Dis 2021;80(2):555-565 [FREE Full text] [doi: 10.3233/JAD-200952] [Medline: 33554903]
13. Czaja S, Boot W, Charness N, Rogers W, Sharit J. Improving social support for older adults through technology: findings from the PRISM randomized controlled trial. Gerontologist 2018 May 08;58(3):467-477 [FREE Full text] [doi: 10.1093/geront/gnw249] [Medline: 28201730]
14. Wang J, Zhang N, Huang C, Wu Q, Tong J. Internet use, physical activity, and cognitive function in Chinese older adults: a cross-lagged panel analysis. Front Aging Neurosci 2025;17:1579874 [FREE Full text] [doi: 10.3389/fnagi.2025.1579874] [Medline: 40405917]
15. Hua Z, Wang F. Association between WeChat use and memory performance among older adults in China: the mediating role of depression. Behav Sci (Basel) 2022;12(9):323 [FREE Full text] [doi: 10.3390/bs12090323] [Medline: 36135127]
16. Zhou Y, Abuduxukuer K, Wang C, Wei J, Shi W, Li Y, et al. WeChat usage and preservation of cognitive functions in middle-aged and older Chinese adults: indications from a nationally representative survey, 2018-2020. BMC Public Health 2024;24(1):1783 [FREE Full text] [doi: 10.1186/s12889-024-19210-5] [Medline: 38965535]
17. Chen B, Yang C, Ren S, Li P, Zhao J. Relationship between internet use and cognitive function among middle-aged and older Chinese adults: 5-year longitudinal study. J Med Internet Res 2024;26:e57301 [FREE Full text] [doi: 10.2196/57301] [Medline: 39539034]
18. Yu D, Fiebig D. Internet use and cognition among middle-aged and older adults in China: a cross-lagged panel analysis. J Econ Ageing 2020;17:100262. [doi: 10.1016/j.jeoa.2020.100262]
19. Ding L, Lu J, Ma X. WeChat use, cognitive function, and depressive symptoms: examining longitudinal relationships among older Chinese adults from a national survey. Mass Commun Soc 2025:1-21. [doi: 10.1080/15205436.2025.2546449]
20. Liu Z, Li Z. Relationships between digital engagement and the mental health of older adults: evidence from China. PLoS One 2024;19(8):e0308071 [FREE Full text] [doi: 10.1371/journal.pone.0308071] [Medline: 39106268]
21. Li A, Qiang W, Li J, Geng Y, Qiang Y, Zhao J. Effectiveness of an exergame-based training program on physical and cognitive function in older adults with cognitive impairment: a randomized controlled trial in rural China. BMC Geriatr 2025;25(1):892 [FREE Full text] [doi: 10.1186/s12877-025-06341-6] [Medline: 41219720]
22. Wen X, Song S, Tian H, Cui H, Zhang L, Sun Y, et al. Intervention of computer-assisted cognitive training combined with occupational therapy in people with mild cognitive impairment: a randomized controlled trial. Front Aging Neurosci 2024;16:1384318 [FREE Full text] [doi: 10.3389/fnagi.2024.1384318] [Medline: 38832072]
23. Cheng Z, Zhou M, Sabran K. Mobile app-based interventions to improve the well-being of people with dementia: a systematic literature review. Assist Technol 2024;36(1):64-74. [doi: 10.1080/10400435.2023.2206439] [Medline: 37115814]
24. Yu X, Ang S, Zhang Y. Exploring rural-urban differences in the association between internet use and cognitive functioning among older adults in China. J Gerontol B Psychol Sci Soc Sci 2024;79(4):gbad195 [FREE Full text] [doi: 10.1093/geronb/gbad195] [Medline: 38147307]
25. Preacher KJ, Hayes AF. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behav Res Methods 2008;40(3):879-891 [FREE Full text] [doi: 10.3758/brm.40.3.879] [Medline: 18697684]
26. Li Y, Han W, Hu M. Does internet access make a difference for older adults' cognition in urban China? The moderating role of living arrangements. Health Soc Care Community 2022;30(4):e909-e920. [doi: 10.1111/hsc.13493] [Medline: 34245201]
27. Benge JF, Scullin MK. Implications for technological reserve development in advancing age, cognitive impairment, and dementia. Behav Brain Sci 2020;43:e157. [doi: 10.1017/S0140525X20000126] [Medline: 32772985]
28. Wolff JL, Benge JF, Cassel CK, Monin JK, Reuben DB. Emerging topics in dementia care and services. J Am Geriatr Soc 2021;69(7):1763-1773 [FREE Full text] [doi: 10.1111/jgs.17341] [Medline: 34245585]
29. Benge JF, Kiselica AM, Aguirre A, Hilsabeck RC, Douglas M, Paydarfar D, et al. Technology use and subjective cognitive concerns in older adults. Arch Gerontol Geriatr 2023;106:104877 [FREE Full text] [doi: 10.1016/j.archger.2022.104877] [Medline: 36459914]

XSL•FO
RenderX

30. Benge JF, Scullin MK. A meta-analysis of technology use and cognitive aging. Nat Hum Behav 2025;9(7):1405-1419 [FREE Full text] [doi: 10.1038/s41562-025-02159-9] [Medline: 40229575]
31. Stern Y, Arenaza-Urquijo EM, Bartrés-Faz D, Belleville S, Cantilon M, Chetelat G, et al. Whitepaper: defining and investigating cognitive reserve, brain reserve, and brain maintenance. Alzheimers Dement 2020;16(9):1305-1311 [FREE Full text] [doi: 10.1016/j.jalz.2018.07.219] [Medline: 30222945]
32. Stern Y, Albert M, Barnes CA, Cabeza R, Pascual-Leone A, Rapp PR. A framework for concepts of reserve and resilience in aging. Neurobiol Aging 2023;124:100-103 [FREE Full text] [doi: 10.1016/j.neurobiolaging.2022.10.015] [Medline: 36653245]
33. Krell-Roesch J, Vemuri P, Pink A, Roberts RO, Stokin GB, Mielke MM, et al. Association between mentally stimulating activities in late life and the outcome of incident mild cognitive impairment, with an analysis of the APOE e4 genotype. JAMA Neurol 2017;74(3):332-338 [FREE Full text] [doi: 10.1001/jamaneurol.2016.3822] [Medline: 28135351]
34. Penninkilampi R, Casey A, Singh MF, Brodaty H. The association between social engagement, loneliness, and risk of dementia: a systematic review and meta-analysis. J Alzheimers Dis 2018;66(4):1619-1633. [doi: 10.3233/JAD-180439] [Medline: 30452410]
35. Scullin MK, Jones WE, Phenis R, Beevers S, Rosen S, Dinh K, et al. Using smartphone technology to improve prospective memory functioning: a randomized controlled trial. J Am Geriatr Soc 2022;70(2):459-469 [FREE Full text] [doi: 10.1111/jgs.17551] [Medline: 34786698]
36. Quan-Haase A, Martin K, Schreurs K. Interviews with digital seniors: ICT use in the context of everyday life. Inf Commun Soc 2016;19(5):691-707. [doi: 10.1080/1369118x.2016.1140217]
37. Ackermann K, Awaworyi Churchill S, Smyth R. Broadband internet and cognitive functioning. Econ Rec Hoboken 2023;99(327):536-563 [FREE Full text] [doi: 10.1111/1475-4932.12757]
38. Schacter DL. Media, technology, and the sins of memory. Mem Mind Media 2022;1:e1 [FREE Full text] [doi: 10.1017/mem.2021.3] [Medline: 34423305]
39. Small GW, Lee J, Kaufman A, Jalil J, Siddarth P, Gaddipati H, et al. Brain health consequences of digital technology use . Dialogues Clin Neurosci 2020;22(2):179-187 [FREE Full text] [doi: 10.31887/DCNS.2020.22.2/gsmall] [Medline: 32699518]
40. Medlock S, Eslami S, Askari M, Arts DL, Sent D, de Rooij SE, et al. Health information-seeking behavior of seniors who use the internet: a survey. J Med Internet Res 2015;17(1):e10 [FREE Full text] [doi: 10.2196/jmir.3749] [Medline: 25574815]
41. Bielak AAM. How can we not 'lose it' if we still don't understand how to 'use it'? Unanswered questions about the influence of activity participation on cognitive performance in older age--a mini-review. Gerontology 2010;56(5):507-519. [doi: 10.1159/000264918] [Medline: 19996570]
42. Yildirim E, Ogel-Balaban H. Cognitive functions among healthy older adults using online social networking. Appl Neuropsychol Adult 2023;30(4):401-408. [doi: 10.1080/23279095.2021.1951269] [Medline: 34310244]
43. Cotten SR, Schuster AM, Seifert A. Social media use and well-being among older adults. Curr Opin Psychol 2022;45:101293. [doi: 10.1016/j.copsyc.2021.12.005] [Medline: 35065352]
44. Han M, Tan XY, Lee R, Lee JK, Mahendran R. Impact of social media on health-related outcomes among older adults in Singapore: qualitative study. JMIR Aging 2021;4(1):e23826 [FREE Full text] [doi: 10.2196/23826] [Medline: 33595437]
45. Ihle A, Bavelier D, Maurer J, Oris M, Kliegel M. Internet use in old age predicts smaller cognitive decline only in men. Sci Rep 2020;10(1):8969 [FREE Full text] [doi: 10.1038/s41598-020-65846-9] [Medline: 32488153]
46. Firth J, Torous J, Stubbs B, Firth JA, Steiner GZ, Smith L, et al. The "online brain": how the internet may be changing our cognition. World Psychiatry 2019;18(2):119-129 [FREE Full text] [doi: 10.1002/wps.20617] [Medline: 31059635]
47. Deng C, Shen N, Li G, Zhang K, Yang S. Digital isolation and dementia risk in older adults: longitudinal cohort study. J Med Internet Res 2025;27:e65379 [FREE Full text] [doi: 10.2196/65379] [Medline: 39969956]
48. Choi NG, Dinitto DM. Internet use among older adults: association with health needs, psychological capital, and social capital. J Med Internet Res 2013;15(5):e97 [FREE Full text] [doi: 10.2196/jmir.2333] [Medline: 23681083]
49. Chen C, Huang N, Hu B, Zhang M, Yuan J, Guo J. The effectiveness of digital technology interventions for cognitive function in older adults: a systematic review and meta-analysis of randomized controlled trials. Geroscience 2025;47(1):653-683. [doi: 10.1007/s11357-024-01446-z] [Medline: 39688787]
50. Gao Y, Liu N. Effects of digital technology-based serious games interventions for older adults with mild cognitive impairment: a meta-analysis of randomised controlled trials. Age Ageing 2025;54(4):afaf080. [doi: 10.1093/ageing/afaf080] [Medline: 40192627]
51. Norman CD, Skinner HA. eHEALS: the eHealth Literacy Scale. J Med Internet Res 2006;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
52. Xin Y, Weina H, Yan D. Digital literacy impacts quality of life among older adults through hierarchical mediating mechanisms. Sci Rep 2025;15(1):19288 [FREE Full text] [doi: 10.1038/s41598-025-04472-9] [Medline: 40456807]
53. Liu S, Lu Y, Wang D, He X, Ren W, Kong D, et al. Impact of digital health literacy on health-related quality of life in Chinese community-dwelling older adults: the mediating effect of health-promoting lifestyle. Front Public Health 2023;11:1200722 [FREE Full text] [doi: 10.3389/fpubh.2023.1200722] [Medline: 37415711]

XSL•FO
RenderX

54. Arias López MDP, Ong BA, Borrat Frigola X, Fernández AL, Hicklent RS, Obeles AJT, et al. Digital literacy as a new determinant of health: a scoping review. PLOS Digit Health 2023;2(10):e0000279 [FREE Full text] [doi: 10.1371/journal.pdig.0000279] [Medline: 37824584]
55. Xie L, Zhang S, Xin M, Zhu M, Lu W, Mo PK. Electronic health literacy and health-related outcomes among older adults: a systematic review. Prev Med 2022;157:106997. [doi: 10.1016/j.ypmed.2022.106997] [Medline: 35189203]
56. Santini ZI, Jose PE, York Cornwell E, Koyanagi A, Nielsen L, Hinrichsen C, et al. Social disconnectedness, perceived isolation, and symptoms of depression and anxiety among older Americans (NSHAP): a longitudinal mediation analysis. Lancet Public Health 2020;5(1):e62-e70 [FREE Full text] [doi: 10.1016/S2468-2667(19)30230-0] [Medline: 31910981]
57. Domènech-Abella J, Mundó J, Haro JM, Rubio-Valera M. Anxiety, depression, loneliness and social network in the elderly: longitudinal associations from The Irish Longitudinal Study on Ageing (TILDA). J Affect Disord 2019;246:82-88. [doi: 10.1016/j.jad.2018.12.043] [Medline: 30578950]
58. Cohen S, Wills TA. Stress, social support, and the buffering hypothesis. Psychological Bulletin 1985;98(2):310-357. [doi: 10.1037/0033-2909.98.2.310] [Medline: 3901065]
59. Carstensen LL. Social and emotional patterns in adulthood: support for socioemotional selectivity theory. Psychol Aging 1992;7(3):331-338. [doi: 10.1037//0882-7974.7.3.331] [Medline: 1388852]
60. Cheng M, Su W, Li H, Li L, Xu M, Zhao X, et al. Factors influencing the social participation ability of rural older adults in China: a cross-sectional study. Front Public Health 2022;10:1001948 [FREE Full text] [doi: 10.3389/fpubh.2022.1001948] [Medline: 36684961]
61. Sun K, Zhou J. Understanding the impacts of internet use on senior citizens' social participation in China: evidence from longitudinal panel data. Telemat Inform 2021;59:101566. [doi: 10.1016/j.tele.2021.101566]
62. Du X, Liao J, Ye Q, Wu H. Multidimensional internet use, social participation, and depression among middle-aged and elderly Chinese individuals: nationwide cross-sectional study. J Med Internet Res 2023;25:e44514 [FREE Full text] [doi: 10.2196/44514] [Medline: 37647119]
63. He T, Huang C, Li M, Zhou Y, Li S. Social participation of the elderly in China: the roles of conventional media, digital access and social media engagement. Telemat Inform 2020;48:101347. [doi: 10.1016/j.tele.2020.101347]
64. Long C, Yang W, Glaser K. Social support, cognition, and mental health among older people in China: a longitudinal life course study. Soc Sci Med 2025;381:118279 [FREE Full text] [doi: 10.1016/j.socscimed.2025.118279] [Medline: 40479799]
65. Ma T, Liao J, Ye Y, Li J. Social support and cognitive activity and their associations with incident cognitive impairment in cognitively normal older adults. BMC Geriatr 2024;24(1):38 [FREE Full text] [doi: 10.1186/s12877-024-04655-5] [Medline: 38191348]
66. Joyce J, Ryan J, Owen A, Hu J, McHugh Power J, Shah R, et al. Social isolation, social support, and loneliness and their relationship with cognitive health and dementia. Int J Geriatr Psychiatry 2022;37(1):00 [FREE Full text] [doi: 10.1002/gps.5644] [Medline: 34741340]
67. Ragab E, Ghannam A. Modernisation in Arab societies: the theoretical and analytical view. Int J Sociol Soc Policy 2001;21:99-131. [doi: 10.1108/01443330110789727]
68. Zeng Y, Wang Z. Dynamics and policy implications of family households and elderly living arrangements in China. American Enterprise Institute. 2019. URL: http://www.jstor.org/stable/resrep24663.5 [accessed 2025-07-30]
69. Cáceres RB, Chaparro AC. Age for learning, age for teaching: the role of inter-generational, intra-household learning in internet use by older adults in Latin America. Inf, Commun Soc 2017:1-17 [FREE Full text] [doi: 10.1080/1369118x.2017.1371785]
70. Silva P, Matos AD, Martinez-Pecino R. Can the internet reduce the loneliness of 50+ living alone? Inf, Commun Soc 2022;25(1):17-33 [FREE Full text] [doi: 10.1080/1369118X.2020.1760917]
71. Berner J, Aartsen M, Deeg D. Predictors in starting and stopping internet use between 2002 and 2012 by Dutch adults 65 years and older. Health Informatics J 2019;25(3):715-730 [FREE Full text] [doi: 10.1177/1460458217720398] [Medline: 28747085]
72. Grošelj D, Reisdorf BC, Petrovčič A. Obtaining indirect internet access: an examination how reasons for internet non-use relate to proxy internet use. Telecommunications Policy 2019;43(3):213-224. [doi: 10.1016/j.telpol.2018.07.004]
73. The 51st statistical report on China's Internet development [Web page in Chinese]. China Internet Network Information Center. 2023. URL: https://www3.cnnic.cn/n4/2023/0302/c199-10755.html [accessed 2025-09-04]
74. Adhikari SP, Dev R, Borson S. Modifying the Mini-Cog to screen for cognitive impairment in nonliterate individuals. Int J Alzheimers Dis 2021;2021:5510093 [FREE Full text] [doi: 10.1155/2021/5510093] [Medline: 34447592]
75. Yang L, Yan J, Jin X, Jin Y, Yu W, Xu S, et al. Screening for dementia in older adults: comparison of Mini-Mental State Examination, Mini-Cog, Clock Drawing Test and AD8. PLoS One 2016;11(12):e0168949 [FREE Full text] [doi: 10.1371/journal.pone.0168949] [Medline: 28006822]
76. Ma Z, Wu M. The psychometric properties of the Chinese eHealth Literacy Scale (C-eHEALS) in a Chinese rural population: cross-sectional validation study. J Med Internet Res 2019;21(10):e15720 [FREE Full text] [doi: 10.2196/15720] [Medline: 31642811]

XSL•FO
RenderX

77. Xu RH, Zhou L, Lu SY, Wong EL, Chang J, Wang D. Psychometric validation and cultural adaptation of the simplified Chinese eHealth Literacy Scale: cross-sectional study. J Med Internet Res 2020;22(12):e18613 [FREE Full text] [doi: 10.2196/18613] [Medline: 33284123]
78. Wu Y, Tang J, Du Z, Chen K, Wang F, Sun X, et al. Development of a short version of the perceived social support scale: based on classical test theory and ant colony optimization. BMC Public Health 2025;25(1):232 [FREE Full text] [doi: 10.1186/s12889-025-21399-y] [Medline: 39833852]
79. Jiang Q. Perceived Social Support Scale. Chin J Behav Med Sci 2001;10(10):41-43 [FREE Full text]
80. Zimet G, Dahlem N, Zimet S, Farley G. The Multidimensional Scale of Perceived Social Support. J Pers Assess Routledge 1988 Mar;52(1):30-41. [doi: 10.1207/s15327752jpa5201_2]
81. Wu X, Tang Y, He Y, Wang Q, Wang Y, Qin X. Prevalence of cognitive impairment and its related factors among Chinese older adults: an analysis based on the 2018 CHARLS data. Front Public Health 2024;12:1500172 [FREE Full text] [doi: 10.3389/fpubh.2024.1500172] [Medline: 39776486]
82. Hayes AF. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach. New York, London: The Guilford Press; Jan 24, 2022.
83. Cho G, Betensky RA, Chang VW. Internet usage and the prospective risk of dementia: a population-based cohort study. J Am Geriatr Soc 2023;71(8):2419-2429. [doi: 10.1111/jgs.18394] [Medline: 37132331]
84. Kim YK, Han SH. Internet use and cognitive functioning in later life: focus on asymmetric effects and contextual factors. Gerontologist 2022;62(3):425-435 [FREE Full text] [doi: 10.1093/geront/gnab149] [Medline: 34614179]
85. Kamin ST, Lang F. Internet use and cognitive functioning in late adulthood: longitudinal findings from the Survey of Health, Ageing and Retirement in Europe (SHARE). J Gerontol B Psychol Sci Soc Sci 2020;75(3):534-539. [doi: 10.1093/geronb/gby123] [Medline: 30346591]
86. Berner J, Comijs H, Elmståhl S, Welmer AK, Sanmartin Berglund J, Anderberg P, et al. Maintaining cognitive function with internet use: a two-country, six-year longitudinal study. Int Psychogeriatr 2019;31(07):929-936 [FREE Full text] [doi: 10.1017/s1041610219000668]
87. d'Orsi E, Xavier AJ, Rafnsson SB, Steptoe A, Hogervorst E, Orrell M. Is use of the internet in midlife associated with lower dementia incidence? Results from the English Longitudinal Study of Ageing. Aging Ment Health 2018;22(11):1525-1533 [FREE Full text] [doi: 10.1080/13607863.2017.1360840] [Medline: 28795579]
88. Klimova B. Use of the internet as a prevention tool against cognitive decline in normal aging. Clin Interv Aging Dove Medical Press Ltd 2016;11:1231-1237 [FREE Full text] [doi: 10.2147/CIA.S113758] [Medline: 27672317]
89. Li S, Cui G, Yin Y, Xu H. Associations between health literacy, digital skill, and eHealth literacy among older Chinese adults: a cross-sectional study. Digit Health 2023;9:20552076231178431 [FREE Full text] [doi: 10.1177/20552076231178431] [Medline: 37256010]
90. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
91. Jung SO, Son YH, Choi E. E-health literacy in older adults: an evolutionary concept analysis. BMC Med Inform Decis Mak 2022;22(1):28 [FREE Full text] [doi: 10.1186/s12911-022-01761-5] [Medline: 35101005]
92. El Benny ME, Kabakian-Khasholian T, El-Jardali F, Bardus M. Application of the eHealth literacy model in digital health interventions: scoping review. J Med Internet Res 2021;23(6):e23473 [FREE Full text] [doi: 10.2196/23473] [Medline: 34081023]
93. Zhang C, Tang D, Wang Y, Jiang S, Liu X. Community support and promoting cognitive function for the elderly. Front Psychol 2022;13:942474 [FREE Full text] [doi: 10.3389/fpsyg.2022.942474] [Medline: 36148108]
94. Li B, Guo Y, Deng Y, Zhao S, Li C, Yang J, et al. Association of social support with cognition among older adults in China: a cross-sectional study. Front Public Health 2022;10:947225 [FREE Full text] [doi: 10.3389/fpubh.2022.947225] [Medline: 36225770]
95. Wang Y, Li J, Fu P, Jing Z, Zhao D, Zhou C. Social support and subsequent cognitive frailty during a 1-year follow-up of older people: the mediating role of psychological distress. BMC Geriatr 2022;22(1):162 [FREE Full text] [doi: 10.1186/s12877-022-02839-5] [Medline: 35227216]
96. Tariq A, Beihai T, Abbas N, Ali S, Yao W, Imran M. Role of perceived social support on the association between physical disability and symptoms of depression in senior citizens of Pakistan. Int J Environ Res Public Health 2020;17(5):1485 [FREE Full text] [doi: 10.3390/ijerph17051485] [Medline: 32106585]
97. Kwan RYC, Ng F, Lai M, Wong D, Chan S. The effects of Digital Buddy programme on older adults' mental well-being: study protocol for a multi-centre, cluster randomized controlled trial. Trials 2023;24(1):95 [FREE Full text] [doi: 10.1186/s13063-023-07130-5] [Medline: 36750879]
98. Zhong R, Ning W. Impact of living arrangements and internet use on the mental health of Chinese older adults. Front Public Health 2024;12:1395181 [FREE Full text] [doi: 10.3389/fpubh.2024.1395181] [Medline: 39712316]
99. Silva P, Delerue Matos A, Martinez-Pecino R. The contribution of the Internet to reducing social isolation in individuals aged 50 years and older: quantitative study of data from the Survey of Health, Ageing and Retirement in Europe. J Med Internet Res 2022;24(1):e20466 [FREE Full text] [doi: 10.2196/20466] [Medline: 34982040]

XSL•FO
RenderX

Abbreviations

C-eHEALS: Chinese version of the eHealth Literacy Scale
DAG: directed acyclic graph
DHL: digital health literacy
eHEALS: eHealth Literacy Scale
MSPSS: Multidimensional Scale of Perceived Social Support
PSSS-3: 3-item short version of the Perceived Social Support Scale
VIF: variance inflation factor

Edited by S Brini; submitted 11.Sep.2025; peer-reviewed by E Calatayud, S Dong; comments to author 17.Nov.2025; accepted 24.Dec.2025; published 20.Jan.2026.

Please cite as:
Du Y, Niu Q, Tan G, Chao J, Jin S, Wang L
Digital Engagement and Cognitive Function Among Older Adults in China: Cross-Sectional Questionnaire Study and Moderated Mediation Model Analysis
J Med Internet Res 2026;28:e83955
URL: https://www.jmir.org/2026/1/e83955
doi:10.2196/83955
PMID:

©Yongqi Du, Qing Niu, Gangrui Tan, Jianqian Chao, Shengxuan Jin, Leixia Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI-Based Triage Decision Support: Multisite Economic Evaluation in the United States

Scott Levin[1], PhD; Ben Steinhart[1], MS; Rohit Sangal[2], MD, MBA; Lesley Meng[3], PhD; K Davina Frick[4], PhD; Cabrina McGinn[5], BS; Jeremiah Hinson[1,6], MD, PhD; Andrew Taylor[7], MD

[1]Danaher Diagnostics, Danaher (United States), 2200 Pennsylvania Avenue NW, Washington, DC, United States
[2]Department of Emergency Medicine, School of Medicine, Yale University, New Haven, CT, United States
[3]Department of Operations, School of Management, Yale University, New Haven, CT, United States
[4]Carey Business School, Johns Hopkins, Baltimore, MD, United States
[5]Radiometer, Copenhagen, Denmark
[6]Department of Emergency Medicine, School of Medicine, Johns Hopkins University, Baltimore, MD, United States
[7]Department of Emergency Medicine, School of Medicine, University of Virginia, Charlottesville, VA, United States

Corresponding Author:
Scott Levin, PhD
Danaher Diagnostics, Danaher (United States), 2200 Pennsylvania Avenue NW, Washington, DC, United States

Abstract

Background: Emergency department (ED) visits have risen in the United States, with demand for emergency care exceeding supply. Resultant ED crowding harms patients, causes staff burnout, and places financial strain on hospitals, payers, and patients alike. Digital tools, including those leveraging artificial intelligence (AI), offer promise for driving efficiency and mitigating the harms of crowding. However, economic frameworks for evaluating these tools remain underdeveloped, limiting adoption.

Objective: This study aimed to develop and apply a generalizable economic model to assess the financial impact of AI-driven efficiency (improved patient throughput) in emergency care. The model contrasts hospital management versus public policy cost-modeling approaches.

Methods: We applied an economic model to operational and financial data collected from 3 EDs (170,723 visits across 180-day preintervention and postintervention periods) following the implementation of an AI triage clinical decision support system. Revenue was directly measured, and costs and operating margin were estimated using 2 cost-modeling frameworks: hospital management (accounting for fixed, variable, and modifiable costs) and public policy (cost-per-visit approach). Sensitivity analysis was used to generalize the model to changes in patient throughput spanning −10% to +10% for 2 ED scenarios: capacity-constrained EDs with volume increases and volume-stable EDs. Break-even analyses determined the maximum sustainable AI tool cost per visit under each scenario that would preserve operating margin for each scenario.

Results: Postintervention revenue increased US $15.4 million (US $138.3 million to US $153.7 million) after AI triage clinical decision support implementation and was accompanied by a 9.6% (7797/81,466) increase in ED visit volume. The financial impact on operating margin from this revenue increase differed substantially by cost-modeling framework. Under the hospital management perspective, costs increased US $2.9 million (US $130.3 million to US $133.2 million), attributing most incremental revenue to an operating margin gain (US $12.6 million). Under the public policy perspective, costs rose proportionally with revenue (US $14.5 million; US $130.3 million to US $144.8 million), generating only US $0.9 million in incremental operating margin. This yielded remarkably different break-even thresholds: US $66.02 per visit (hospital management) versus US $4.69 per visit (public policy) for a 5% (16/311) gain in efficiency (patient throughput) in capacity-constrained ED settings.

Conclusions: The financial impact of tools that promote increased ED efficiency varied greatly based on cost-modeling framework. Traditional policy-level approaches substantially underestimate value from a hospital management perspective. The economic model provides a pragmatic and generalized framework for evaluating AI-driven efficiency in the ED, addressing a critical gap that can limit adoption. The economic model is publicly available and adaptable to local contexts.

(J Med Internet Res 2026;28:e95213) doi:10.2196/95213

KEYWORDS

emergency medicine; triage; economic model; artificial intelligence; AI; clinical decision support

XSL•FO
RenderX

Introduction

Annual emergency department (ED) visits have risen to more than 145 million in the United States, with demand for emergency care far outstripping supply [1,2]. Resultant ED overcrowding harms patients, causes staff burnout, and financially burdens hospitals, payers, and patients [2-5]. While patient care impacts have been well studied, ED crowding's financial consequences have garnered less attention. This is despite ED spending growing faster than other health care sectors [6,7]. Digital tools that leverage artificial intelligence (AI) hold promise for increasing operational efficiency and curbing ED crowding [8-11]. However, a significant gap remains in the economic evaluation of such technologies. Without a practical framework to guide such analyses, health care decision-makers lack the ability to assess cost-effectiveness and financial impact in their local context.

Existing economic evaluations typically apply public policy-level cost perspectives (eg, cost-to-charge ratios) that do not account for the fixed versus variable cost structure central to hospital operations [12-14]. From a hospital management perspective, most ED costs are fixed to maintain 24/7 standby capacity regardless of patient volume [12-14]. Fixed costs include standby labor and facility, equipment, and overhead. The remaining costs are patient throughput–dependent; these include variable costs (eg, supplies and pharmaceuticals) scaling directly with patient volume and modifiable labor costs (eg, contract labor and overtime) that fluctuate with busyness, a product of patient volume and length of stay (LOS) [15-18]. Contract nursing labor has become a particularly significant expense, with disproportionately high use in EDs [17,18]. Under this cost structure, efficiency gains that increase patient throughput and/or reduce care hours delivered can substantially improve operating margins (revenue minus costs) [19]. In contrast, policy-level analyses apply a consistent cost per visit based on hospital charges and cost-to-charge ratios [20,21], which fundamentally differ from hospitals' true cost structures. This limits applicability to decisions regarding AI-driven ED efficiency tools.

The objective of this study was to develop and apply an economic model to assess the financial impact of AI-driven efficiency (gains in ED patient throughput), contrasting hospital management versus public policy cost-modeling approaches. We applied the economic model to data collected from a multisite AI triage clinical decision support (AI triage CDS) intervention [8]. The economic model is publicly available [22] and easily adaptable to local contexts. The study introduces a novel dual-framework for evaluating AI-based ED interventions and demonstrates that choice of cost-modeling approach can substantially alter estimated financial value, with implications for adoption decisions.

Methods

Study Design and Participants

This secondary economic analysis used data from a previously reported preintervention and postintervention study of the AI triage CDS across 3 EDs within a single health system [8]. Implementation occurred in staggered fashion between May 19, 2021, and April 4, 2023. The analysis included 170,723 ED visits during 180-day preintervention and postintervention periods at each site, excluding 2.2% (3925/174,648) of visits in which patients left prior to disposition. Cohort demographics and patient outcomes have been previously reported [8].

Ethical Considerations

The study received approval from the Yale University Institutional Review Board (2000036009) and is reported in accordance with the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement (Checklist 1) [23].

The expected operating margin rate of 5.8% was adjusted from a national report by Wilson and Cutler [19] to account for payer type and admission rate across the study sites.

AI Triage CDS Intervention

The AI triage CDS [8,9,24-26] applies site-specific machine learning algorithms to ED triage data routinely collected on patient arrival to predict risk of 3 patient outcomes: critical care (defined as in-hospital mortality or intensive care unit admission), emergency surgery (any surgery performed in a dedicated operating room suite within 12 hours of ED disposition), and hospital admission. Predictors include demographics (age and sex), arrival mode, vital signs (heart rate, respiratory rate, oxygen saturation, temperature, and systolic blood pressure), chief complaint, and active medical problems.

Predicted probabilities are translated to recommended acuity levels (1‑5, with lower values indicating higher acuity). Probabilities for critical care and emergency surgery meeting specific thresholds trigger recommendations for highest-risk levels (1 or 2), while probability of hospitalization guides recommendations for lower risk levels (3-5) [8]. The AI triage CDS output appears within seconds in the ED nurse workflow, displaying the recommended triage level with individualized explanations of the recommendation generated using Shapley Additive Explanations values transformed to natural language for each patient. Nurses retain triage decision-making autonomy and may assign acuity levels aligned with or divergent from the AI triage CDS recommendations.

Triage acuity levels dictate early course of care for patients and shape ED operational efficiency [27-29]. High-acuity patients are prioritized for immediate evaluation, while low-acuity patients are commonly diverted to high-efficiency workstreams, such as fast-track or vertical care areas (not requiring an ED bed), where they can be treated quickly. Midacuity patients typically follow traditional ED workflows with longer wait times [27].

ED Efficiency Outcomes

Postintervention, low-acuity visit assignments (levels 4‑5) increased by a relative 48.2% (from 19,934/83,404, 23.9% to 32,300/91,244, 35.4%), enabling more patients to benefit from streamlined fast-track pathways. Conversely, midacuity (level 3) visits decreased 18.7% (from 40,701/83,404, 48.8% to 36,223/91,244, 39.7%), reducing the proportion of patients in the ambiguous category that typically experiences prolonged

XSL•FO
RenderX

wait times [8]. Median ED LOS for all patients decreased 6.1% (19/311.0 min; from 311 to 292 min), with 5% (15.6/311.0 minutes) attributed to the intervention after adjustment for confounders [8]. These efficiency gains created capacity for an additional 9.6% (7795/81,466) of visits [8].

Economic Model

The economic model measured total ED revenue (before the intervention and after the intervention) and provided 2 approaches for estimating costs and corresponding operating margin. Revenue was defined as the payment received by the hospital for health care services; this excludes fees for physician-specific services. Costs were defined as expenses incurred by the hospital to provide health care services [20]. Operating margin was defined as revenue minus cost, with the corresponding operating margin rate shown in Textbox 1 [30]. Operating margin indicates profitability of core patient care operations [30,31]. All economic model inputs and assumptions may be adjusted to reflect local hospital characteristics and projections. A spreadsheet (Microsoft Excel) for static calculations and a script (Python version 3.8; Python Software Foundation) to enable additional sensitivity analyses is publicly available for download [22]. Table S1 in Multimedia Appendix 1 provides a guide to support custom model inputs based on published data that have been inflation-adjusted to current value (2025) using the Bureau of Labor Statistics Consumer Price Index for Medical Care [32]. A formal economics analysis plan was not developed prior to this secondary evaluation; the economic model structure, assumptions, and analytic approach are described in this section, and the full model is publicly available.

Textbox 1. Economic model inputs.

Revenue

- US $1698 per emergency department (ED) visit (average) based on data collected during the preintervention period

Hospital costs: hospital management level

- US $1600 per ED visit (average) based on an expected 5.8% operating margin rate on revenue data collected during the preintervention period
- The expected operating margin rate of 5.8% was adjusted from a national report by Wilson and Cutler [19] to account for payer type and admission rate across the study sites.
  - Fixed costs (baseline), 70%
    - Labor, 60%
    - Facility, 5%
    - Equipment (including digital), 3%
    - Overhead (administration), 5%
  - Variable costs, 20%
    - Supplies, 12%
    - Pharmaceuticals, 8%
  - Modifiable costs, 10%
    - Modifiable labor (eg, contract labor and overtime), 10%

Hospital costs: public policy level

- Costs dependent on ED visit volume. Does not account for care hours provided or associated labor.

Operating margin, US $, %

- US $98 per ED visit (average), represented as the difference between revenue and cost at baseline for an expected margin rate of 5.8%: (revenue − costs)/revenue

Revenue Data

Hospital revenue data for the cohort were collected from billing tables housed in the electronic health record system. Revenue for ED discharged patients was attributed wholly to the ED, while revenue for hospital admitted patients was calculated as a fraction of the revenue realized from their entire hospital stay [19]. This fraction was calculated as 24 hours divided by total hospital LOS (hours) for hospital stays exceeding 24 hours, and 12 hours divided by total LOS for stays under 24 hours. To mitigate the influence of outliers common in revenue data, revenue per ED visit was capped at the 95th percentile (US $7241). This approach purposefully yields conservative estimates of total revenue and operating margin by excluding potential high-revenue outliers, which follow high-acuity emergency pathways independent of triage methods. For patients with missing revenue estimates (5447/170,723, 3.19%), values were imputed using the mean total revenue within strata defined

XSL•FO
RenderX

by admission status (discharge or admit) and primary payer type (Medicare, Medicaid, private insurance, and self-pay).

Cost and Operating Margin Estimates

Baseline preintervention cost was estimated by presuming a 5.8% operating margin rate on observed revenue. This rate was derived for our study cohort based on payer-specific margin rates reported by Wilson and Cutler [19], weighted by the distribution of revenue across payer and admission groups in our preintervention data (Textbox 1). Thus, the total preintervention cost was calculated as total revenue multiplied by (1–0.058). The resulting preintervention total costs and average cost per ED visit served as the reference for postintervention comparisons. Postintervention operating margin was calculated as revenue minus cost.

Two distinct cost frameworks were applied to postintervention data. The hospital management perspective accounts for time and resources, allocating 70% of total costs as fixed for standby ED care, 20% as variable costs dependent on patient volume, and 10% as modifiable clinician labor that fluctuates with care hours (Textbox 1). This allocation is consistent with prior ED cost analyses [12-14] and reflects that most ED costs—including standby capacity labor, facility, equipment, and overhead—accrue independent of patient throughput, while supplies and pharmaceuticals vary with volume, and a smaller portion of labor can be flexed through mechanisms such as contract labor [18], overtime, or flexible staffing. Though exact ratios may vary by institution, the model allows adjustment of these assumptions. Alternatively, the public policy perspective applies a consistent cost-per-visit approach based on hospital charges and cost-to-charge ratios, holding average cost per ED visit constant regardless of time and resources consumed (ie, care hours).

Sensitivity Analyses

The economic model was applied to preintervention and postintervention data using measured revenue and both cost frameworks (hospital and public policy) to estimate costs and operating margin changes relative to baseline (before intervention). Sensitivity analyses were performed on these results to both conceptualize and generalize how changes in ED efficiency (ED LOS and patient throughput) translate to financial metrics under 2 common scenarios: (1) *scenario 1, capacity-constrained EDs*, where efficiency gains create new capacity for additional patient volume, applicable where demand for ED services exceeds supply; and (2) *scenario 2, volume-stable EDs*, where efficiency gains do not increase patient volume, applicable where demand for ED services is being met.

For both scenarios, ED LOS changes were modeled from the baseline observed median of 311 minutes to −10% (decreased efficiency, LOS increase to 342 min) to +10% (increased efficiency, LOS decrease to 280 min). Preintervention data (volume, LOS, and revenue) served as baseline comparators. Revenue, costs, and operating margin were estimated across this parameter space.

Break-even analyses were conducted to determine maximum AI tool cost per ED visit to achieve break-even, the point at which tool costs would exactly offset operating margin gains. Break-even estimates per visit were calculated as the difference in margin from the preintervention period to the postintervention period divided by total postintervention visits. This represents the per visit cost an efficiency-enhancing intervention could incur without negating the observed operating margin improvement.

Results

Financial Impact of the AI Triage CDS

Overview

Postintervention revenue increased US $15.4 million (from US $138.3 million to US $153.7 million), driven by a 9.6% (7795/81,466) increase in ED visit volume from 81,464 to 89,259 visits (Table 1). The additional volume was accommodated without expansion of physical infrastructure or staffing, consistent with efficiency gains previously attributed to the intervention after adjustment for confounders. Median ED LOS decreased 19 minutes (15.6 min after adjusting for confounders) [8], during which the EDs accommodated 7795 additional visits.

XSL•FO
RenderX

Table . Preintervention and postintervention financial metrics.

Metric	Preintervention	Postintervention	Difference
ED[a] operations			
Patient volume, n[b]	81,464	89,259	7795
ED LOS[c] (min), median (IQR)[b]	311 (182-554)	292 (173-496)	19
Hospital management level			
Revenue (US $[b])	138,309,296	153,747,123	15,437,827
Costs (US $[d])	130,287,357	133,155,246	2,867,889
Operating margin (US $), %	8,021,939 (5.8)	20,591,878 (13.4)	12,569,938
Public policy level			
Revenue (US $[b])	138,309,296	153,747,123	15,437,827
Costs (US $[d])	130,287,357	144,829,790	14,542,433
Operating margin (US $), %[d]	8,021,939 (5.8)	8,917,333 (5.8)	895,394

[a]ED: emergency department.

[b]ED operations metrics and revenue measures were observed from the artificial intelligence triage clinical decision support intervention study.

[c]LOS: length of stay.

[d]Cost and operating margin were estimated by the economic model.

Average ED revenue per visit was US $1698 (SD US $1834) before intervention and US $1722 (SD US $1843) after intervention (Textbox 1), comparable to published reports (Table S1 in Multimedia Appendix 1). The baseline 5.8% operating margin rate, derived from payer-specific margins weighted by our cohort's revenue distribution, yielded preintervention costs of US $130.3 million (Table 1) and an average cost per visit of US $1600 (Textbox 1). From this baseline, the financial impact of efficiency gains during the postintervention period differed widely based on cost-modeling framework.

Hospital Management Perspective

Costs increased modestly by US $2.9 million (US $130.3 million to US $133.2 million), reflecting the high fixed costs (70%) and modifiable costs (10%) that were unchanged after intervention. This cost structure attributed most incremental revenue (US $12.6 million of the US $15.4 million) to operating margin gain. The operating margin rate improved from 5.8% ($8,021,939/$138,309,396) to 13.4% ($20,591,878/$153,747,123; Table 1).

Public Policy Perspective

Costs increased proportionally with revenue by US $14.5 million (US $130.3 million to US $144.8 million) as the constant cost-per-visit approach holds average costs fixed regardless of efficiency. This yielded a minimal US $0.9 million gain in operating margin (from US $8.0 million to US $8.9 million), with the operating margin rate unchanged ($8,021,939/$138,309,396, 5.8%).

Economic Model Sensitivity Analyses

Overview

Figure 1 depicts sensitivity analysis results for capacity-constrained and volume-stable ED scenarios.

In Figure 1, changes in ED LOS computed from −10% reduction in efficiency (increase in LOS to 342 min) through a 10% increase in efficiency (decrease in LOS to 280 min). Scenario 1 simulates changes in ED LOS with a corresponding (equal and inverse) change in ED visit volume. Scenario 2 simulates changes in ED LOS with no resulting change in ED visit volume. Hospital management is abbreviated as Hosp. Mgmt.

XSL•FO
RenderX

Figure 1. Economic model sensitivity analyses. ED: emergency department; HM: hospital management.

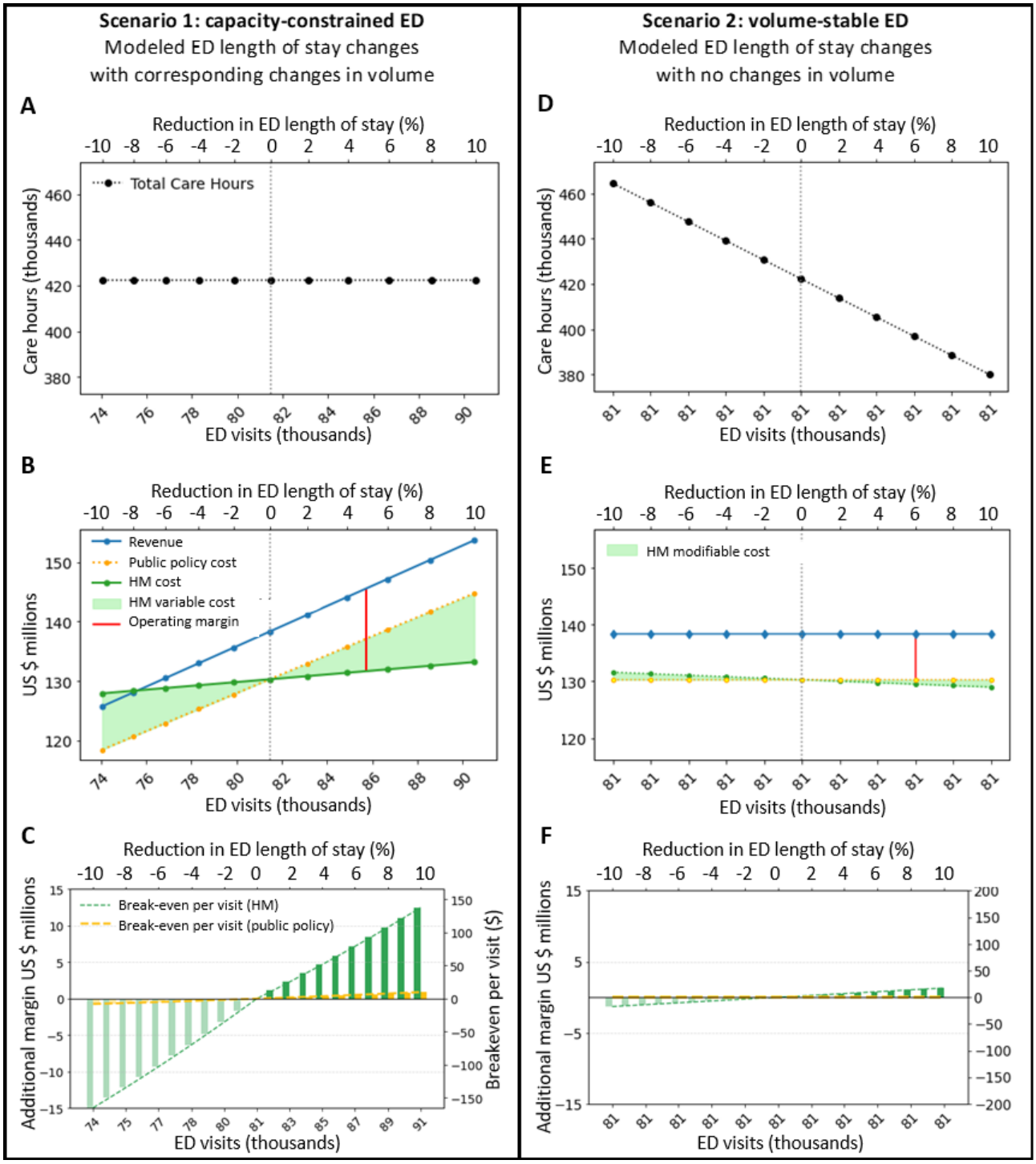

Scenario 1: Capacity-Constrained EDs

In capacity-constrained settings where demand exceeds supply, changes in ED efficiency (LOS) inversely correspond to patient volume changes, holding total care hours (product of patient volume and LOS) constant (Figure 1A). Figure 1B displays the impact on operating margin under the hospital management versus public policy level cost structures. Under the hospital management perspective, operating margin changes substantially with efficiency. Alternatively, the public policy perspective demonstrates costs moving proportionally with volume, minimally impacting operating margin. Contrasting each perspective, an AI tool that increases efficiency by 5% breaks even at US $66.02 per ED visit from a hospital perspective but may only cost US $4.69 to break-even adopting the policy-level perspective on cost (Figure 1C).

Scenario 2: Volume-Stable EDs

In this setting, efficiency is not associated with patient volume but does affect total care hours provided (Figure 1D). There is no impact on revenue (Figure 1E) at baseline. However, changes in total care hours due to LOS often require adjustments in modifiable costs for a hospital. Increased efficiency (decreases in ED LOS) presents an opportunity to reduce modifiable labor

XSL•FO
RenderX

costs (eg, contract labor and overtime), resulting in an increased operating margin. Under this scenario, an AI tool that increases efficiency by 5% breaks even at US $8.00 per ED visit when total modifiable costs may be reduced by a commensurate 5% (US $13.0 million to US $12.3 million). Applying the public policy cost-estimation approach, revenue and cost do not change, resulting in a decrease in total operating margin with the introduction of any new cost (ie, cost of the AI tool).

Discussion

Principal Results

The study demonstrates a financial impact assessment of an AI intervention designed to drive ED efficiency. The assessment illustrates how financial impact varied substantially based on cost-modeling framework. From a hospital management perspective, accounting for fixed and variable cost structures, the AI triage CDS implementation was associated with significant gains in operating margin, amounting to US $12.6 million (158% increase) in our multisite implementation. Conversely, public policy-level cost approaches substantially underestimated economic value, showing only US $0.9 million (11% increase) in margin gain despite identical revenue increases. The large difference in break-even thresholds between perspectives (US $66.02 vs US $4.69 per visit) has direct implications for implementation decisions and highlights the critical need for appropriate economic frameworks when evaluating ED efficiency interventions for a hospital or the public.

The increased financial value under the hospital management perspective reflects ED cost structure: approximately 70% of costs remained fixed (standby capacity), while only 20% varied with volume (supplies and pharmaceuticals) and 10% represented modifiable labor. This meant US $12.6 million of the US $15.4 million revenue increase translated directly to operating margin improvement. The cost structure assumptions (70% fixed, 20% variable, and 10% modifiable) align with prior ED cost analyses [12-14] and reflect the fundamental economics of emergency care delivery. Modifiable labor costs have become increasingly relevant given that contract nursing labor now represents 40% (2022) [18] of hospital-wide nursing expenses, with even higher use in EDs [17]. Our sensitivity analyses demonstrate that even modest reductions in modifiable costs through improved ED efficiency can generate substantial financial value in volume-stable ED settings.

Comparison With Prior Work

While previous economic evaluations of ED interventions have documented efficiency gains [6,7,12,19,21], they typically do not explicitly distinguish between hospital management and policy-level cost perspectives. Our economic model highlights how cost-modeling approach (perspective) fundamentally shapes economic value assessment. Recent studies have highlighted the promise of AI tools for improving ED triage accuracy and patient flow [8,9,24], but comprehensive economic analyses comparing cost perspectives have been limited.

Implications for Practice

Hospital decision-makers should evaluate AI efficiency tools using cost frameworks that reflect operational economics, not policy-level per-visit approaches. In capacity-constrained EDs, even modest efficiency gains can yield substantial operating margin improvements when additional patient volume can be accommodated (Figure 1). In volume-stable ED settings, efficiency gains provide opportunities to reduce modifiable labor costs (eg, contract labor and overtime) where applicable. Although the patient perspective was not explicitly measured in these analyses, the operational efficiency gains (reduced LOS) underlying the modeled financial value are patient-relevant outcomes previously associated with reduced morbidity and improved satisfaction [3,8,9,27]. Efficiency gains that improve operating margin can also reduce patient wait times and crowding-related harms. Policymakers should recognize that traditional cost-per-visit approaches substantially underestimate value to hospitals and may inadvertently discourage adoption of beneficial technologies; reimbursement and incentive policies should account for hospitals' actual cost structures and local market dynamics. Furthermore, developers of AI tools addressing ED crowding should adopt setting-specific value propositions that reflect differential impact based on ED capacity constraints and revenue-generating opportunity from new patient volume.

Economic Model Generalizability and Adaptability

The economic model developed in this study is designed for broad applicability beyond the specific AI triage CDS implementation studied. We provide the model publicly [22] with guidance for adapting inputs to local contexts (Table S1 and user guide in Multimedia Appendix 1). Our baseline financial metrics align with other large-scale reports of ED financials (Table S1 in Multimedia Appendix 1), with observed differences attributable to cohort selection, payer mix, admission rates (acuity), and accounting methods for ED ancillary and physician-based services. While our cost allocation assumptions (70% fixed, 20% variable, and 10% modifiable labor) are a synthesis of published analyses, institutions can adjust these parameters to reflect local cost structures. Similarly, baseline operating margins, revenue per visit, and efficiency projections can be customized using institutional data or published benchmarks adjusted for inflation.

The 2-scenario approach (capacity-constrained vs volume-stable) allows decision-makers to model their specific operational context. Most US EDs operate in capacity-constrained environments where demand exceeds supply [1,2], making scenario 1 widely applicable. However, EDs in lower-demand settings or those with excess capacity may better align with the volume-stable scenario 2, where value derives primarily from modifiable cost reductions rather than volume increases. The model accommodates both scenarios and the full spectrum between them.

Overall, to support practical application of the economic model, Table S1 in Multimedia Appendix 1 provides published ED financial benchmarks with a user guide for selecting inputs appropriate to local cohort, payer mix, and ED type, and Table S2 in Multimedia Appendix 1 reports the underlying numerical

XSL•FO
RenderX

values across the modeled efficiency range (−10% to +10% LOS change), allowing users to read off expected operating-margin and break-even outcomes at specific efficiency assumptions without rerunning the model. Furthermore, EDs assessing their position on the capacity-constrained–to–volume-stable spectrum may consider a combination of indicators—such as inpatient boarding burden, wait times, ambulance diversion, and rates of patients leaving prior to disposition—interpreted alongside local administrative knowledge, recognizing that no single metric reliably distinguishes the 2 scenarios. The −10% to +10% efficiency range modeled in our sensitivity analyses is intended to encompass the magnitude of LOS and throughput gains that may be associated with AI-based ED interventions in isolation.

Limitations

This analysis relies on data from a single health system, though revenue estimates align with published national data. The economic estimates also depend on the attribution of volume and LOS changes to the AI triage CDS established in the parent intervention study [8]; to the extent that residual unmeasured confounding contributes to those observed changes, the financial estimates here would be commensurately affected. The analysis focused on ED-specific costs and does not capture downstream hospital financial impacts. A formal patient perspective was not modeled; future economic work should incorporate equity considerations and out-of-pocket cost effects to complement the hospital management and public policy perspectives presented here. The break-even analyses assume costs and benefits remain stable over time and do not account for potential learning curves, technology updates, or changing market conditions. How efficiency tools vary in value across different hospital ownership structures (for-profit, not-for-profit, and public), market competitiveness levels, and regulatory environments were not studied. Specifically, the model parameters reflect the US health care payment system; application to publicly funded or mixed systems (eg, Europe) would require substitution of locally appropriate values. Finally, while we demonstrate substantial financial value under realistic assumptions, individual hospital experience will vary based on local factors, including payer mix, labor markets, competitive dynamics, and operational practices. Future research should validate this framework across diverse health systems and settings with varying levels of capacity constraint, cost structures, ownership types, and market conditions.

Conclusions

This study demonstrates a pragmatic economic model for assessing ED efficiency interventions, applied to a real-world AI triage CDS implementation. The analysis illustrates how financial impacts differ substantially based on the cost-modeling framework. Traditional policy-level approaches substantially underestimate economic value from a hospital management perspective by not accounting for fixed and variable hospital cost structures. The model, publicly available for local adaptation, provides health care decision-makers with a practical tool that accounts for actual hospital cost behavior. As health care systems increasingly consider AI tools to improve efficiency and address ED overcrowding, economic evaluation frameworks that reflect operational realities are essential for informed adoption decisions.

Acknowledgments

The authors used Anthropic's Claude during manuscript preparation for language editing, prose refinement, and literature search assistance. No generative artificial intelligence (AI) was used for data analysis, statistical modeling, or interpretation of results. All AI-assisted text was reviewed and verified by the authors, who take responsibility for the accuracy and integrity of the content.

Funding

This work was supported by Danaher Corporation, which includes wholly owned subsidiaries Danaher Diagnostics, Beckman Coulter, and Radiometer.

Data Availability

Data used in this study are from the Yale New Haven Health System. Individual patient-level data are protected for privacy. Those seeking to collaborate may contact the Yale University–affiliated authors. Any data sharing must comply with the Yale University Institutional Review Board and associated data-sharing protocols. The economic model is publicly available for download [22] in two formats: (1) a spreadsheet (Microsoft Excel) for static calculations and (2) a script (Python version 3.8) for sensitivity analyses.

Authors' Contributions

Conceptualization: SL, AT, BS, RS, JH
Data curation: AT, BS, SL
Formal analysis: BS, SL
Funding acquisition: AT
Methodology: SL, BS, KDF, LM
Supervision: AT, SL
Visualization: BS, SL, RS, CMG

XSL•FO
RenderX

Writing—original draft: SL, BS, AT, JH
Writing—review and editing: RS, LM, KDF, CMG

Conflicts of Interest

The artificial intelligence triage technology discussed in this publication was developed at Johns Hopkins University (JHU) and has been licensed to and made commercially available by Radiometer. JHU, SL, and JH are entitled to royalty distributions related to the technology discussed in this publication. SL and JH are employed by Danaher Diagnostics, a subsidiary of Danaher Corporation. BS was employed by Danaher Diagnostics and CM was employed by Radiometer, both subsidiaries of Danaher Corporation. AT and RS report research funding from Beckman Coulter, also a subsidiary of Danaher Corporation. The remaining authors report no conflicts of interest.

Multimedia Appendix 1
Economic model input references and economic model sensitivity analyses.
[DOCX File, 88 KB - jmir_v28i1e95213_app1.docx]

Checklist 1
CHEERS checklist.
[PDF File, 231 KB - jmir_v28i1e95213_app2.pdf]

References

1. National Hospital Ambulatory Medical Care Survey: 2021 emergency department summary tables. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/data/nhamcs/web_tables/2021-nhamcs-ed-web-tables-508.pdf [accessed 2026-05-21]
2. Institute of Medicine, Board on Health Care Services, Committee on the Future of Emergency Care in the United States Health System. Hospital-Based Emergency Care: At the Breaking Point: National Academies Press; 2007.
3. Bernstein SL, Aronsky D, Duseja R, et al. The effect of emergency department crowding on clinically oriented outcomes. Acad Emerg Med 2009 Jan;16(1):1-10. [doi: 10.1111/j.1553-2712.2008.00295.x] [Medline: 19007346]
4. Bayley MD, Schwartz JS, Shofer FS, et al. The financial burden of emergency department congestion and hospital crowding for chest pain patients awaiting admission. Ann Emerg Med 2005 Feb;45(2):110-117. [doi: 10.1016/j.annemergmed.2004.09.010] [Medline: 15671965]
5. Foley M, Kifaieh N, Mallon WK. Financial impact of emergency department crowding. West J Emerg Med 2011 May;12(2):192-197. [Medline: 21691525]
6. Pickens G, Smith MW, McDermott KW, Mummert A, Karaca Z. Trends in treatment costs of U.S. emergency department visits. Am J Emerg Med 2022 Aug;58:89-94. [doi: 10.1016/j.ajem.2022.05.035] [Medline: 35660368]
7. Scott KW, Liu A, Chen C, et al. Healthcare spending in U.S. emergency departments by health condition, 2006-2016. PLoS One 2021;16(10):e0258182. [doi: 10.1371/journal.pone.0258182] [Medline: 34705854]
8. Taylor RA, Chmura C, Hinson J, et al. Impact of artificial intelligence–based triage decision support on emergency department care. NEJM AI 2025 Feb 27;2(3). [doi: 10.1056/AIoa2400296]
9. Hinson JS, Taylor RA, Venkatesh A, et al. Accelerated chest pain treatment with artificial intelligence-informed, risk-driven triage. JAMA Intern Med 2024 Sep 1;184(9):1125-1127. [doi: 10.1001/jamainternmed.2024.3219] [Medline: 39037785]
10. Kachman MM, Brennan I, Oskvarek JJ, Waseem T, Pines JM. How artificial intelligence could transform emergency care. Am J Emerg Med 2024 Jul;81:40-46. [doi: 10.1016/j.ajem.2024.04.024] [Medline: 38663302]
11. Chenais G, Lagarde E, Gil-Jardiné C. Artificial intelligence in emergency medicine: viewpoint of current applications and foreseeable opportunities and challenges. J Med Internet Res 2023 May 23;25:e40031. [doi: 10.2196/40031] [Medline: 36972306]
12. Dyas SR, Greenfield E, Messimer S, et al. Process-improvement cost model for the emergency department. J Healthc Manag 2015;60(6):442-457. [Medline: 26720989]
13. Costs of caring: challenges facing America's hospitals as they care for patients in 2026. American Hospital Association. URL: https://www.aha.org/costsofcaring [accessed 2025-02-12]
14. Young DW. Charging appropriately for emergency department visits. Healthcare Financial Management Association. 2017. URL: https://www.hfma.org/payment-reimbursement-and-managed-care/pricing/53948/ [accessed 2025-05-05]
15. Levin S, Aronsky D, Hemphill R, Han J, Slagle J, France DJ. Shifting toward balance: measuring the distribution of workload among emergency physician teams. Ann Emerg Med 2007 Oct;50(4):419-423. [doi: 10.1016/j.annemergmed.2007.04.007] [Medline: 17559969]
16. Sharma J, Bastani A, Jones S. 50 cost benefit analysis of physician-in-triage model at community hospital emergency department. Ann Emerg Med 2017 Oct;70(4):S20. [doi: 10.1016/j.annemergmed.2017.07.074]
17. Hospital vitals: financial and operational trends. Syntellis and American Hospital Association. 2023. URL: https://www.stratadecision.com/sites/default/files/2023-03/AHA%20Q2_Feb%202023.pdf [accessed 2025-02-12]

XSL•FO
RenderX

18. Massive growth in expenses and rising inflation fuel continued financial challenges for America's hospitals and health systems. American Hospital Association. 2022. URL: https://www.aha.org/system/files/media/file/2022/04/2022-Hospital-Expenses-Increase-Report-One-Pager.pdf [accessed 2026-05-21]
19. Wilson M, Cutler D. Emergency department profits are likely to continue as the Affordable Care Act expands coverage. Health Aff (Millwood) 2014 May;33(5):792-799. [doi: 10.1377/hlthaff.2013.0754] [Medline: 24799576]
20. HCUP-US overview. Agency for Healthcare Research and Quality. URL: https://www.hcup-us.ahrq.gov/overview.jsp [accessed 2019-12-11]
21. Pickens GT, Moore B, Smith MW, McDermott KW, Mummert A, Karaca Z. Methods for estimating the cost of treat-and-release emergency department visits. Health Serv Res 2021 Oct;56(5):953-961. [doi: 10.1111/1475-6773.13709] [Medline: 34350589]
22. CDEM-JHU/AI-ED-ECON. GitHub. URL: https://github.com/CDEM-JHU/AI-ED-ECON [accessed 2026-05-09]
23. Husereau D, Drummond M, Augustovski F, et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations. Value Health 2022 Jan;25(1):3-9. [doi: 10.1016/j.jval.2021.11.1351] [Medline: 35031096]
24. Hinson JS, Levin SR, Steinhart BD, et al. Enhancing emergency department triage equity with artificial intelligence: outcomes from a multisite implementation. Ann Emerg Med 2025 Mar;85(3):288-290. [doi: 10.1016/j.annemergmed.2024.10.014] [Medline: 39570253]
25. Levin S, Toerper M, Hamrock E, et al. Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index. Ann Emerg Med 2018 May;71(5):565-574. [doi: 10.1016/j.annemergmed.2017.08.005] [Medline: 28888332]
26. Levin S, Toerper M, Hinson J, et al. 294 Machine-learning-based electronic triage: a prospective evaluation. Ann Emerg Med 2018 Oct;72(4):S116. [doi: 10.1016/j.annemergmed.2018.08.299]
27. Wiler JL, Gentle C, Halfpenny JM, et al. Optimizing emergency department front-end operations. Ann Emerg Med 2010 Feb;55(2):142-160. [doi: 10.1016/j.annemergmed.2009.05.021] [Medline: 19556030]
28. Hinson JS, Martinez DA, Cabral S, et al. Triage performance in emergency medicine: a systematic review. Ann Emerg Med 2019 Jul;74(1):140-152. [doi: 10.1016/j.annemergmed.2018.09.022] [Medline: 30470513]
29. Calder LA, Forster AJ, Stiell IG, et al. Mapping out the emergency department disposition decision for high-acuity patients. Ann Emerg Med 2012 Nov;60(5):567-576.e4. [doi: 10.1016/j.annemergmed.2012.04.013] [Medline: 22699018]
30. Ly DP, Jha AK, Epstein AM. The association between hospital margins, quality of care, and closure or other change in operating status. J Gen Intern Med 2011 Nov;26(11):1291-1296. [doi: 10.1007/s11606-011-1815-5] [Medline: 21837374]
31. Akinleye DD, McNutt LA, Lazariu V, McLaughlin CC. Correlation between hospital finances and quality and safety of patient care. PLoS One 2019;14(8):e0219124. [doi: 10.1371/journal.pone.0219124] [Medline: 31419227]
32. How BLS measures price change for medical care services in the consumer price index. U.S. Bureau of Labor Statistics. URL: https://www.bls.gov/cpi/factsheets/medical-care.htm [accessed 2025-06-26]

Abbreviations

AI: artificial intelligence
CDS: clinical decision support
CHEERS: Consolidated Health Economic Evaluation Reporting Standards
ED: emergency department
LOS: length of stay

Edited by I Steenstra; submitted 05.Apr.2026; peer-reviewed by AM Napoli, D Lupu; revised version received 12.May.2026; accepted 13.May.2026; published 03.Jun.2026.

Please cite as:
Levin S, Steinhart B, Sangal R, Meng L, Frick KD, McGinn C, Hinson J, Taylor A
AI-Based Triage Decision Support: Multisite Economic Evaluation in the United States
J Med Internet Res 2026;28:e95213
URL: https://www.jmir.org/2026/1/e95213
doi:10.2196/95213

© Scott Levin, Ben Steinhart, Rohit Sangal, Lesley Meng, K Davina Frick, Cabrina McGinn, Jeremiah Hinson, Andrew Taylor. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 3.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the

XSL•FO
RenderX

Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Cost-Effectiveness Analysis of Chatbot-Supported Remote Patient Monitoring for Anticoagulation Management: Health Economic Evaluation Within a Pilot Crossover Trial

Ana Rita Santos[1,2,3,4], PharmD; Filipa Sampaio[4,5], PhD; Federico Guede-Fernández[3,6], PhD; Julian Perelman[1,2], PhD; Ana Londral[2,3,7], PhD

[1]Nova National School of Public Health, Universidade Nova de Lisboa, Lisbon, Portugal

[2]Comprehensive Health Research Centre, Lisbon, Portugal

[3]Value for Health CoLAB, Lisbon, Portugal

[4]Uppsala Health Economics, Department of Public Health and Caring Sciences, Uppsala University, BMC, Husargatan 3, Uppsala, Sweden

[5]Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden

[6]LIBPhys (Laboratory for Instrumentation, Biomedical Engineering and Radiation Physics), Nova School of Science and Technology, Universidade Nova de Lisboa, Lisbon, Portugal

[7]Department of Physics, NOVA School of Science and Technology, Universidade NOVA de Lisboa, Lisbon, Portugal

Corresponding Author:
Ana Rita Santos, PharmD
Nova National School of Public Health, Universidade Nova de Lisboa, Lisbon, Portugal

Abstract

Background: Digital health technologies (DHTs) are increasingly integrated into clinical practice, yet economic evaluations remain scarce, particularly in the early development stages. Within the NICE (National Institute for Health and Care Excellence) Evidence Standards Framework, Tier C DHTs comprise technologies with direct clinical implications and measurable health outcomes, for which robust economic evidence is essential. Early-stage assessments are particularly important to inform subsequent development, refinement, and adoption decisions across the digital health lifecycle.

Objective: This study aims to explore the feasibility of integrating a full trial-based economic evaluation within an early-stage pilot comparing a chatbot-supported remote patient monitoring (RPM) solution for anticoagulation management with the standard of care (SOC).

Methods: A cost-effectiveness analysis was performed alongside a pilot crossover trial among adult cardiac surgery patients receiving vitamin K antagonists. Participants were allocated to two 6-month sequences (SOC→RPM or RPM→SOC). The intervention consisted of a rule-based chatbot integrated with home-based international normalized ratio self-testing using portable coagulometers to support communication and therapy management. Effectiveness was measured as time in therapeutic range (TTR), and costs were estimated from the Portuguese National Health Service and a limited societal perspective over a 1-year horizon. The analysis 1 applied a within-patient cost-effectiveness approach to estimate incremental costs, incremental TTR, and incremental cost-effectiveness ratios. Uncertainty was explored through nonparametric bootstrapping (5000 replications) and deterministic sensitivity analyses. Complementary comparisons examined differences between sequences (analysis 2), between periods (analysis 3), and within each sequence (analysis 4).

Results: A total of 19 patients were included in the analyses. In analysis 1, RPM improved anticoagulation control, with a mean within-patient increase of 10.43 percentage points in time in TTR. The mean incremental costs were €198.61 (€1=US $1.08) from the Serviço Nacional de Saúde perspective and €270.05 from the limited societal perspective. The corresponding incremental cost-effectiveness ratios were €19.03 and €25.88 per additional percentage point of TTR gained. Sensitivity analyses produced consistent estimates across parameter variations. Complementary analyses (2-4) suggested that RPM tended to be more cost-effective when implemented after the initial 6-month postoperative period.

Conclusions: This proof-of-concept study demonstrates that a full trial-based economic evaluation can feasibly be embedded within an early-stage Tier C DHT. The intervention showed improved anticoagulation control alongside higher costs, providing initial insights into its cost-effectiveness profile. Positioned within the digital health evidence continuum, such assessments can function as a learning stage within the lifecycle. To address the persistent adoption-evidence gap, tier- and stage-aligned frameworks are needed to guide the economic evaluation of DHTs. This study contributes to that goal by providing a set of recommendations specifically for Tier C DHTs.

Trial Registration: ClinicalTrials.gov NCT06423521; https://clinicaltrials.gov/study/NCT06423521

XSL•FO
RenderX

(J Med Internet Res 2026;28:e85430) doi:10.2196/85430

KEYWORDS

digital health; economic evaluation; cost-effectiveness; lifecycle; remote patient monitoring; chatbot; anticoagulants

Introduction

Digital health technologies (DHTs) are increasingly used in health care, offering opportunities to improve outcomes and enhance system efficiency [1,2]. Yet, health technology assessment agencies and payer bodies continue to face challenges in adapting their methods to evaluate these technologies effectively [3]. The lack of robust evidence on their real benefits and cost-effectiveness hinders decision-making and risks fostering the diffusion of low-value, short-lived solutions [2]. Traditional approaches to economic evaluation, primarily designed and used for pharmaceuticals, are often misaligned with the fast-paced development of DHTs [4,5]. This creates a paradox: without evidence, adoption is unlikely; without adoption, generating evidence is difficult [6]. Some pioneering countries, such as Belgium and France, have begun to integrate economic evaluation into the assessment of DHTs [7]. However, others, such as Germany, through its Digitale Gesundheitsanwendungen fast-track scheme, do not require a formal economic evaluation for initial listing [3,7]. This uneven landscape underscores the absence of harmonized guidance on the generation and application of economic evidence for DHTs.

To bridge this gap, both regulatory initiatives and academic research have contributed to the development of dedicated frameworks for the economic evaluation of DHTs. Our recent scoping review identified 26 such frameworks and revealed substantial heterogeneity across key dimensions; however, 2 main recommendations emerged [8]. First, DHTs should be differentiated by purpose and risk, as reflected in the NICE (National Institute for Health and Care Excellence) Evidence Standards Framework [9] tiers: Tier A (low-risk, efficiency-focused solutions), Tier B (health and well-being support tools), and Tier C (high-risk clinical interventions with direct implications for diagnosis or treatment). Second, evaluation should follow stage-based approaches aligned with the DHT lifecycle, as emphasized in several frameworks [10-14]. For example, the World Health Organization maturity model [10] outlines a continuum of early (preprototype, prototype, pilot), mid (demonstration), and advanced (scale-up, integration) stages. This lifecycle perspective underscores the iterative nature of digital health evaluation, where evidence is built progressively through repeated assessments using different methodologies [11,12].

Despite this convergence, limited guidance exists on which economic evaluation methods are most appropriate at each maturity stage or tier. Model-based approaches are often recommended for early development [11,15], but they depend on the primary collection of reliable and context-specific data that can only be obtained through pilot studies [11]. The pilot stage is, therefore, critical for generating early evidence on feasibility, usability, and initial clinical and economic value. Pilot studies serve as structured learning opportunities to optimize technology use, identify barriers, and inform decisions for subsequent evaluations in later stages and for scale-up [10,16].

This study presents a pilot-stage use case of a Tier C DHT, integrating a full economic evaluation within a crossover trial of chatbot-supported remote patient monitoring (RPM) for postoperative anticoagulation therapy with vitamin K antagonists (VKAs) [15]. VKAs remain indispensable for patients with mechanical heart valves and other indications, but their use is clinically challenging due to a narrow therapeutic window: underdosing increases thromboembolic risk, while overdosing raises the risk of bleeding [17-22]. Anticoagulation control is best assessed by the time in therapeutic range (TTR). It is defined as the proportion of time a patient's international normalized ratio (INR)—a standardized measure of blood clotting time that reflects anticoagulation control—remains within the target range, typically 2.0 - 3.0 for most indications [20,23]. An inverse correlation between TTR and hemorrhagic and thromboembolic complications has been established [24-26]. Accordingly, current clinical guidelines recommend maintaining a TTR above 70% to minimize adverse events [17,22], which requires frequent INR monitoring and timely dose adjustments. Under the standard of care (SOC), INR is usually measured in health care facilities every 3 to 4 weeks [20], creating logistical barriers that may compromise anticoagulation control [27].

To address these challenges, the pilot crossover study evaluated a rule-based chatbot integrated with home-based INR self-testing using portable coagulometers to strengthen communication between patients and health care professionals and support timely therapy management [15]. The intervention improved anticoagulation control, particularly when introduced after the unstable postoperative period, resulting in gains in TTR [15]. Furthermore, a time-driven activity-based costing (TDABC) analysis of this intervention provided an initial understanding of its resource use and cost structure [15]. Building on these clinical and economic findings, this study extends this work by conducting a within-trial cost-effectiveness analysis (CEA) of chatbot-supported RPM compared with SOC over a 1-year time horizon from the perspective of the Portuguese National Health Service (Serviço Nacional de Saúde, SNS) and a broader, albeit limited, societal perspective. By embedding a full CEA into a pilot stage, this study aims to explore the feasibility of integrating economic assessment into early-stage Tier C DHTs. It also aims to illustrate how such evaluations can generate proof-of-concept evidence to inform subsequent model-based analyses and adoption decisions across the digital health lifecycle.

Methods

Study Design and Participants

The economic evaluation used data collected as part of an efficacy pilot crossover trial (ClinicalTrials.gov NCT06423521),

XSL•FO
RenderX

whose primary clinical, implementation, and cost-related results have been published elsewhere [15], and it followed the CHEERS (Consolidated Health Economic Evaluation Reporting Standards) statement [28] (Checklist 1). In summary, the trial was conducted at the Department of Cardiothoracic Surgery, Hospital de Santa Marta, Lisbon. Eligible participants were adult patients (≥18 y) who had been discharged following cardiac surgery and prescribed long-term oral anticoagulation therapy with VKAs. Participants were assigned to 1 of 2 intervention sequences: SOC followed by RPM (group 1: SOC→RPM) or RPM followed by SOC (group 2: RPM→SOC), with each phase lasting 6 months, resulting in a total follow-up of 12 months per patient.

A total of 27 patients were enrolled between April 2021 and April 2023; 26 completed both phases and were included in the pilot. The final sample comprised 13 women and 13 men, with a mean age of 55.7 (SD 13.3) years. For the present CEA, only patients with ≥2 INR measurements per phase were included to allow TTR calculation. The final analytic sample comprised 19 participants (n=9 group 1, n=10 group 2). A dropout analysis was conducted to compare baseline characteristics (age, sex, indication for anticoagulation therapy, and allocation group) between included and excluded participants.

Interventions

During the SOC phase, participants received standard postoperative anticoagulation care, consisting of routine INR monitoring at health care facilities and in-person consultations for VKA dose adjustments.

In the RPM phase, participants self-monitored INR at home using a portable coagulometer (CoaguChek XS; Roche Diagnostics) and were supported by a digital platform with a rule-based chatbot enabling bidirectional communication with the clinical team. Patients received automated prompts via smartphone to report INR values and relevant symptoms, and any out-of-range results automatically triggered a clinical response, including dosage adjustments and rescheduling of measurements. The chatbot functioned as an intermediary, ensuring timely clinical review and continuity of VKA therapy management.

Resource Use and Costs

Costs were estimated from 2 perspectives. The SNS perspective included intervention costs, health care resource use, and medication-related costs borne by the SNS. The SNS is a universal, predominantly tax-funded service that provides comprehensive care at minimal cost to patients, coexisting with a private sector that does not offer the same breadth of services.

We also considered the broader, albeit limited, societal perspective, encompassing, in addition, the health care and transportation costs borne by the patients. Productivity losses and other costs beyond the health sector were excluded, as these data were not collected during the trial. Value-added tax was applied following INFARMED's (Autoridade Nacional do Medicamento e Produtos de Saúde, I.P. [National Authority of Medicines and Health Products]) guidelines [29] for the SNS perspective. For the limited societal perspective, value-added tax was also included based on existing literature [30,31].

Intervention-specific costs were attributed to the RPM phase and included the portable coagulometer (device), test strips, a one-time training session, and the provision of a chatbot platform. Costs for the coagulometer and test strips were obtained from the manufacturer (Roche). The device cost (€600, €1=US $1.08) was amortized over a 3-year lifetime [32] using the straight-line depreciation method, allocating one-sixth of the cost to each 6-month intervention, corresponding to €100 per patient. INR test costs were estimated by multiplying the number of patient tests by the unit price of strips. A one-time training session was assumed to be the unit price for a home visit according to the 2023 hospital financing contract [33], under the assumption that training involved comparable professional time [34]. The chatbot platform, developed and maintained internally by Value for Health CoLAB, was costed at €50 per patient for the 6 months, reflecting market prices of comparable low-code solutions. This approach was used as a proxy in the absence of detailed internal cost data and captures development, infrastructure, basic maintenance, and text messages.

Health care resource use included medical and nursing consultations, hospitalizations, emergency visits, INR testing (SOC phase), and anticoagulant medication. Resource quantities were obtained from patient records, with hospitalizations coded by Diagnosis-Related Groups (Grupos de Diagnóstico Homogéneo), which also provided unit costs by severity. Total health care costs per patient were then calculated by multiplying unit costs by the corresponding quantities of resources used. Unit costs for medical consultations and emergency visits were based on the 2023 hospital financing contract, which sets the prices to be paid to SNS hospitals for several services [33]. Nursing consultations, not included in the contract, were estimated at half the cost of a medical consultation, consistent with the ratio between medical and nursing consultation costs reported in Ordinance 207/2017 [35]. Hospitalizations and INR tests were costed using the same ordinance [35]. In line with national regulations [35], only hospitalization costs were considered when an emergency visit led to admission; otherwise, emergency visits were valued at the standard daily rate for medical-surgical emergency care. The cost of INR testing during the SOC phase was assumed to be identical across settings (primary care, hospital, or private laboratory). Medication costs were taken from the INFARMED database [36]. Given that dosage data were incomplete and prescribed doses of VKAs vary substantially according to INR values (typically ranging from 2.5 to 10 mg for warfarin [37] and 1 to 8 mg for acenocoumarol [38]), fixed daily doses of 5 mg warfarin and 4 mg acenocoumarol were applied as a simplifying and conservative assumption, reflecting typical midrange dosing in clinical practice. Medication costs were valued using the lowest available market price, and for the SNS perspective, patient co-payments corresponding to 31% of the retail price were excluded [26].

Travel costs were considered for in-person medical and nursing consultations and for INR testing during the SOC phase. Distances were calculated as round trips between patients' residences and health care facilities. For mainland Portugal, personal car travel was assumed. For patients from Madeira and

XSL•FO
RenderX

the Azores islands, local INR testing used car travel, while consultations held on the mainland assumed public transport. This distinction reflects the context of the islands, where local INR testing is done nearby, while travel to mainland hospitals requires air travel. Costs followed official reimbursement rates [39,40].

All costs were expressed in 2024 euros (€), adjusted using the Consumer Price Index [41], except for medication, which used 2025 INFARMED prices. As the time horizon was 1 year, no discounting was applied [29]. Unit costs and resources are summarized in Table 1.

Table . Unit costs, including value-added tax (2024 €, except for medication costs in 2025 €).

Cost item	Unit cost (€)	Source
Intervention costs (RPM[a] only)		
Machine	600	Roche
Test strip	3.88	Roche
Training session	42.94	Hospital financing contract 2023
Chatbot platform	50	Own estimate
Health and health care use		
Medical consultations	80.76	Hospital financing contract 2023
Nurse consultations	40.38	Own estimate based on Hospital financing contract 2023 and Portuguese Ordinance 207/2017 of 11th July
Medical consultations (telemedicine)	88.94	Hospital financing contract 2023
Hospitalizations	Varies by GDH[b]	Portuguese Ordinance 207/2017 of 11th July
Emergency	111.43	Hospital financing contract 2023
INR[c] test (SOC[d] only)	3.42	Portuguese Ordinance 207/2017 of 11th July
Medication costs		
Warfarin	0.12/day (5 mg)	INFARMED
Acenocoumarol	0.17/day (4 mg)	INFARMED
Out-of-pocket costs		
Travel costs	0.46/km (personal car) 0.14/km (public transportation)	Portuguese Ordinance 1553-D/2008 of 31st December revised by Decree Law 137/2010 of 28th December

[a]RPM: remote patient monitoring.

[b]GDH: Grupos de Diagnóstico Homogéneo.

[c]INR: international normalized ratio.

[d]SOC: standard of care.

Health Outcomes

The primary outcome was the difference in TTR between the SOC and RPM phases. TTR was defined as the percentage of time, measured in days, that a patient's INR values remained within the target therapeutic range, typically between 2.0 and 3.0, depending on individual clinical indications. TTR was calculated for each intervention phase using the Rosendaal method of linear interpolation, which estimates the proportion of time spent within the therapeutic range based on successive INR values [42,43].

During the RPM phase, the expected frequency of INR self-testing was approximately once every 15 days, as outlined in the study protocol. Nonetheless, the clinical team could modify the testing schedule based on individual results. If a patient's INR values were outside the therapeutic range, clinicians were permitted to increase the testing frequency to optimize therapeutic adjustment and reduce the time required to restore INR control.

Statistical Analysis

Data management and costing were performed in Microsoft Excel (version 16.98), and statistical analyses were performed in RStudio (version 2025.05.1+513; Posit Software, PBC).

Descriptive statistics were used to summarize the data. Categorical variables were reported as counts and percentages, and continuous variables as mean (SD) or median (IQR), depending on the data distribution assessed by the Shapiro-Wilk test. For the analysis of patients' characteristics and dropout comparison, Chi-square tests or Fisher exact tests were used for categorical variables, and independent *t* tests or Wilcoxon rank-sum tests for continuous variables, depending on the distribution of the data.

XSL•FO
RenderX

Mean differences in costs and TTR between intervention phases were analyzed using linear regression models, where the dependent variables were costs and TTR, and the main explanatory variable was the intervention phase (RPM vs SOC). No additional covariates were included. Statistical significance was defined as $P<.05$.

The analytical strategy comprised 4 components. Analysis 1 consisted of the base-case within-patient CEA, comparing RPM with SOC over a 1-year time horizon from both the SNS and limited societal perspectives. Analysis 2 involved a between-sequence comparison, contrasting the 2 crossover sequences (SOC→RPM vs RPM→SOC). Analysis 3 included period-specific comparisons, contrasting SOC in Sequence 1 with RPM in Sequence 2 during Period 1, and RPM in Sequence 1 with SOC in Sequence 2 during Period 2. Analysis 4 comprised within-sequence comparisons, assessing RPM vs SOC separately within each sequence according to the order in which participants received the interventions.

For the base-case analysis 1, the incremental cost-effectiveness ratio (ICER) was calculated as the mean difference in costs divided by the mean difference in TTR (percentage points), representing the additional cost per unit improvement in anticoagulation control. Uncertainty around the ICER was assessed only for the base-case analysis using nonparametric bootstrapping (5000 iterations), which does not rely on distributional assumptions for costs or effects. This procedure provided 95% CIs for incremental costs, incremental TTR, and the ICER. The bootstrapped estimates were used to construct cost-effectiveness planes and cost-effectiveness acceptability curves (CEACs). The planes display the distribution of incremental cost-effect pairs across quadrants, where each quadrant has a decision-making implication, while CEACs illustrated the probability that RPM is cost-effective across a range of willingness-to-pay thresholds per additional percentage point of TTR.

Sensitivity Analysis

One-way deterministic sensitivity analyses were conducted for the analysis 1 to assess the impact of key cost and effectiveness assumptions on the ICER. The parameters varied were: (1) chatbot cost (€25-€75/patient), reflecting system development, maintenance, and SMS text messaging usage; (2) coagulometer amortization, assuming lifespans of 5 and 10 years instead of the base-case 3 years; (3) travel distances (±20%) to account for geographic variability; and (4) TTR gains, using the lower and upper bounds of the 95% CI.

Ethical Considerations

Ethical approval was obtained from the Ethics Committee of the ULS São José (REC number: 1057/2021) and was registered on ClinicalTrials.gov under the identifier NCT06423521.

Results

Participants' Characteristics

Demographic characteristics were well balanced between the 2 groups at baseline [15]. Table 2 summarizes the characteristics of the 19 participants included in the economic analysis (Figure 1). No statistically significant differences were observed between the groups in terms of age ($P=.61$) or sex ($P=.31$).

Table . Participants' characteristics.

Variable	Overall (N=19)	Group 1 (n=9)	Group 2 (n=10)
Age (y), mean (SD)	55.37 (14.06)	53.56 (14.93)	57.00 (13.83)
Sex, female, n (%)	9 (47.37)	3 (33.33)	6 (60)
Indication for anticoagulation therapy, n (%)			
Mechanical aortic prosthesis	11 (57.89)	7 (77.78)	4 (40.00)
Pulmonary thromboembolism	2 (10.53)	1 (11.11)	1 (10.00)
Atrial fibrillation	1 (5.26)	1 (11.11)	0 (0.00)
Mitral valvuloplasty	1 (5.26)	0 (0.00)	1 (10.00)
Bentall procedure	1 (5.26)	0 (0.00)	1 (10.00)
Mechanical mitral prosthesis	2 (10.53)	0 (0.00)	2 (20.00)
Mechanical mitral prosthesis and atrial fibrillation	1 (5.26)	0 (0.00)	1 (10.00)
Number of INR[a] measures, mean (SD)			
RPM[b]	21.21 (5.64)	18.44 (3.88)	23.7 (5.98)
SOC[c]	11.32 (5.71)	15.33 (5.72)	7.7 (2.31)

[a]INR: international normalized ratio.

[b]RPM: remote patient monitoring.

[c]SOC: standard of care.

XSL•FO
RenderX

Figure 1. Flow diagram of the pilot crossover trial. SOC: standard of care; RPM: remote patient monitoring.

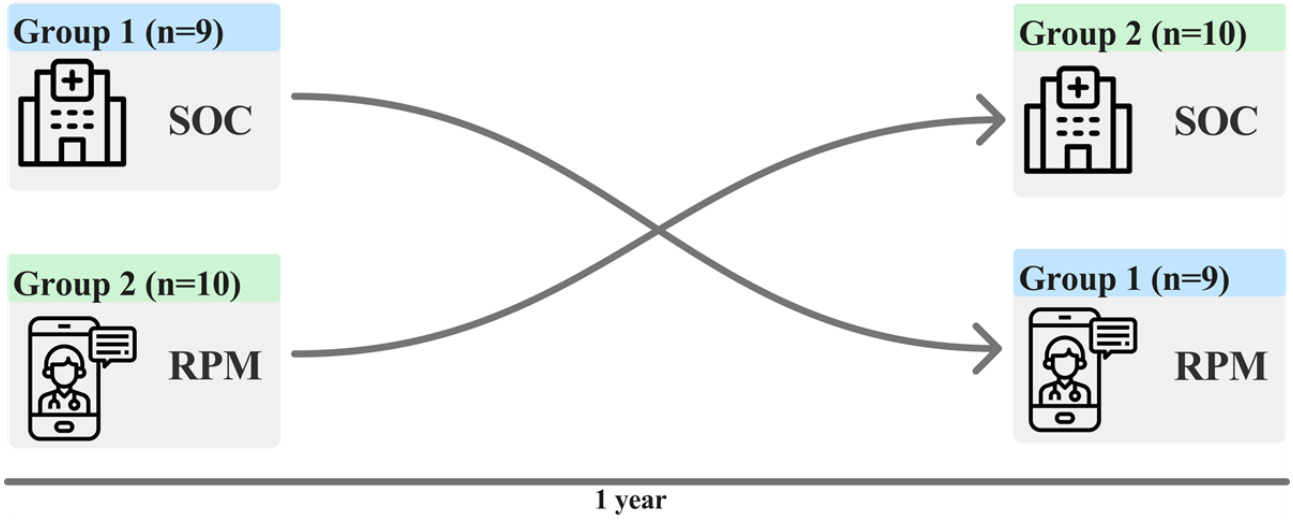

A dropout analysis comparing the 19 participants included in the TTR-based analysis with the 7 participants excluded due to insufficient INR data revealed no statistically significant differences in baseline characteristics. Specifically, there were no significant differences in sex distribution (Fisher exact test, $P>.99$), age (Welch t test, $P=.89$), indication for anticoagulation (Fisher exact test, $P=.38$), or intervention sequence group (Fisher exact test, $P=.66$). These findings suggest that dropout is unlikely to have introduced bias into the analysis.

Resource Use and Costs

Detailed mean costs per participant by category and perspective are presented in Table 3. Intervention-specific items (coagulometer, test strips, chatbot platform, and training) contributed exclusively to the RPM phase, accounting for €275 per patient on average. Among health care services, medical consultations were more frequent, with more associated costs during RPM (mean €191.27, SD €94.07) than SOC (mean €136.02, SD €137.40), while nurse consultations showed similar costs across phases (mean €146.64, SD €76.39 vs mean €148.77, SD €187.02). One patient required hospitalization during the SOC phase, contributing to an observed mean cost of €109.78 (SD €478.53). Medication costs were identical across phases and perspectives, and INR testing costs applied only to SOC (mean €38.70, SD €19.52). From the limited societal perspective, travel costs were higher during the RPM phase (mean €227.64, SD €295.75) than during SOC (mean €102.88, SD €129.88), likely reflecting the greater number of in-person medical consultations observed in the RPM phase.

XSL•FO
RenderX

Table . Costs (in €) per participant over the trial from the Serviço Nacional de Saúde and limited societal perspectives.[a]

Cost item	RPM[b] phase, mean (SD)	SOC[c] phase, mean (SD)	Mean difference (95% CI), *P* value
Intervention costs (RPM[b] only) (both perspectives)			
Machine	100	—[d]	100
Test strip	82.27 (21.89)	—[d]	82.27
Training session	42.94	—[d]	42.94
Chatbot platform	50	—[d]	50
Health and health care use (both perspectives)			
Medical consultations	191.27 (94.07)	136.02 (137.40)	55.26 (−19.62 to 130.13), *P*=.16
Nurse consultations	146.64 (76.39)	148.77 (187.02)	−2.13 (−92.96 to 88.71), *P*=.96
Medical consultations (telemedicine)	18.72 (63.44)	0	18.72
Hospitalizations	0	109.78 (478.53)	−109.78
Emergency	0	0	0
INR[e] costs (SOC[b] only)	—[d]	38.70 (19.52)	−38.70
Medication costs (SNS[f] perspective)			
Warfarin/Acenocoumarol	19.32 (3.44)	19.32 (3.44)	0
Medication costs (limited societal perspective)			
Warfarin/Acenocoumarol	27.76 (4.30)	27.76 (4.30)	0
Out-of-pocket costs (limited societal perspective)			
Travel costs INR[e] (SOC[c] only)	—[d]	53.28 (86.60)	−53.28
Other travel costs	227.64 (295.75)	102.88 (129.88)	124.71 (−20.53 to 269.96), *P*=.10

[a]Mean differences were calculated as RPM minus SOC; negative values indicate lower costs in the RPM phase.

[b]RPM: remote patient monitoring.

[c]SOC: standard of care.

[d]Not applicable.

[e]INR: international normalized ratio.

[f]SNS: Portuguese National Health Service (Serviço Nacional de Saúde).

Under the within-patient base-case comparison (analysis 1), mean costs were higher during RPM than during SOC. From the SNS perspective, the mean difference was €198.61 (95% CI –€90.19 to €487.41; *P*=.19), while from the limited societal perspective, the difference was €270.05 (95% CI –€75.29 to €615.39; *P*=.13).

When the 2 crossover sequences were contrasted (analysis 2), costs were lower in the RPM→SOC sequence. The estimated differences were –€108.30 from the SNS perspective (95% CI –€402.55 to €186.04; *P*=.48) and –€191.00 from the limited societal perspective (95% CI –€542.43 to €160.43; *P*=.29).

Period-specific comparisons (analysis 3) showed that, in period 1, RPM was associated with incremental costs of €74.10 (95% CI –€443.29 to €591.50; *P*=.78) from the SNS perspective and €58.77 (95% CI –€524.15 to €641.69; *P*=.85) from the limited societal perspective. In Period 2, incremental costs increased substantially, reaching €290.61 (95% CI €95.32 to €485.90; *P*=.01) from the SNS perspective and €440.80 (95% CI €152.41 to €729.14; *P*=.01) from the limited societal perspective.

Finally, the within-sequence contrasts (analysis 4) indicated that, in the SOC→RPM sequence, RPM was associated with lower costs from the SNS perspective (–€126.47, 95% CI –€672.97 to €420.04; *P*=.63) but considerably higher costs from the limited societal perspective (€491.18, 95% CI €304.87 to €677.49; *P*<.001). In the opposite sequence (RPM→SOC), RPM again had lower costs from the SNS perspective (–€135.47, 95% CI –€726.57 to €455.62; *P*=.66) and higher costs from the limited societal perspective (€635.02, 95% CI €320.51 to €949.52; *P*<.001). All incremental cost estimates for analyses 1-4 are detailed in Table 4.

XSL•FO
RenderX

Table . Incremental costs and incremental effects across analyses 1-4 from the Serviço Nacional de Saúde and limited societal perspectives.[a]

Analysis and outcome	SNS[b] perspective: incremental costs, € (95% CI), *P* value	Limited societal perspective: incremental costs, € (95% CI), *P* value	Incremental effects (TTR[c]), percentage points, (95% CI), *P* value
Analysis 1: within patient comparison			
Base-case analysis	198.61 (−90.19 to 487.41), .19	270.05 (−75.29 to 615.39), .13	10.43 (−3.74 to 24.61), .16
Analysis 2: between sequence comparison			
SOC[d]→RPM[e] vs RPM[e]→SOC[d]	−108.30 (−402.55 to 186.04), .48	−191.00 (−542.43 to 160.43), .29	−8.98 (−23.28 to 5.32), .23
Analysis 3: period specific comparisons			
Period 1: RPM (RPM[e]→SOC[d]) vs SOC (SOC[d]→RPM[e])	74.10 (−443.29 to 591.50), .78	58.77 (−524.15 to 641.69), .85	2.06 (−12.94 to 17.05), .79
Period 2: RPM (SOC[d]→RPM[e]) vs SOC[d] (RPM[e]→SOC[d])	290.61 (95.32 to 485.90), .01	440.80 (152.41 to 729.14), .01	20.02 (−3.05 to 43.09), .11
Analysis 4: within sequence comparisons			
RPM[e] vs SOC[d] (SOC[d]→RPM[e])	−126.47 (−672.97 to 420.04), .63	491.18 (304.87 to 677.49), <.001	22.53 (7.34 to 37.71), .01
RPM[e] vs SOC[d] (RPM[e]→SOC[d])	−135.47 (−726.57 to 455.62), .66	635.02 (320.51 to 949.52), <.001	−0.45 (−22.65 to 21.75), .97

[a]Incremental values are reported as RPM compared with SOC. Negative cost differences indicate lower costs with RPM, and positive values for incremental effects (TTR, percentage points) indicate greater improvement in anticoagulation control.

[b]SNS: Portuguese National Health Service (Serviço Nacional de Saúde).

[c]TTR: time in therapeutic range.

[d]SOC: standard of care.

[e]RPM: remote patient monitoring.

Health Outcomes

In the base-case within-patient comparison (analysis 1), mean TTR was higher during RPM (60.70%, SD 17.50%) than during SOC (50.26%, SD 26.21%), yielding an incremental difference of 10.43 percentage points (95% CI –3.74 to 24.61; *P*=.16).

When the 2 crossover sequences were compared (analysis 2), TTR was, on average, 8.98 percentage points higher in the SOC→RPM sequence than in the RPM→SOC sequence, although this difference did not reach statistical significance (95% CI –5.32 to 23.28; *P*=.23).

Period-specific contrasts (analysis 3) showed a small and nonsignificant incremental effect in period 1 (2.06 percentage points, 95% CI –12.94 to 17.05; *P*=.79), and a larger but still nonsignificant difference in period 2 (20.02 percentage points, 95% CI –3.05 to 43.09; *P*=.11).

Finally, the within-sequence comparisons (analysis 4) indicated that participants in the SOC→RPM sequence experienced a statistically significant improvement in TTR during the RPM phase relative to SOC (22.53 percentage points, 95% CI 7.34 to 37.71; *P*=.01). No significant change was observed in the RPM→SOC sequence, where the difference was –0.45 percentage points (95% CI –22.65 to 21.75; *P*=.97). All incremental effectiveness estimates for analyses 1-4 are detailed in Table 4.

Cost-Effectiveness Analysis

From the SNS perspective, the base-case ICER was €19.03 per additional percentage point of TTR gained with RPM compared with SOC. Uncertainty was further explored through nonparametric bootstrapping (5000 replications). The cost-effectiveness plane (Figure 2) illustrates the distribution of incremental cost-effectiveness pairs. Most estimates (4275/5000, 85.5%) fell within the northeast quadrant, indicating that RPM was more effective and more costly than SOC; 9.1% (455/5000) were in the southeast quadrant, suggesting improved effectiveness at lower costs; 5% (250/5000) were in the northwest (higher cost, lower effectiveness); and 0.4% (20/5000) were in the southwest (less effective and less costly). The CEAC (Figure 3) shows that, at a willingness-to-pay (WTP) of €95 per percentage TTR gained, the probability of cost-effectiveness reached 90%.

XSL•FO
RenderX

Figure 2. Cost-effectiveness plane (Serviço Nacional de Saúde [SNS] perspective). TTR: time in therapeutic range.

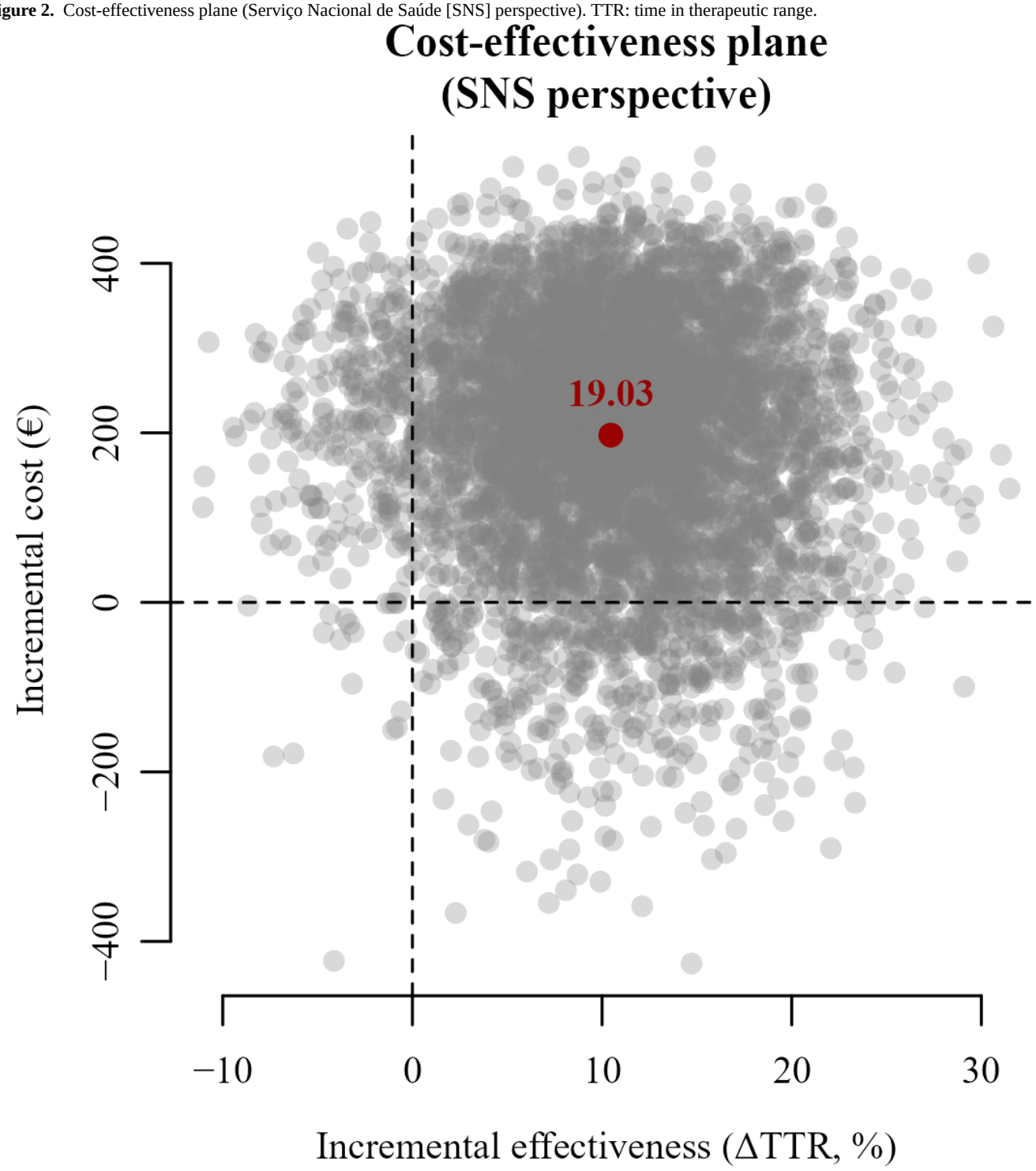

XSL•FO
RenderX

Figure 3. Cost-effectiveness acceptability curve for the Serviço Nacional de Saúde (SNS) perspective. TTR: time in therapeutic range.

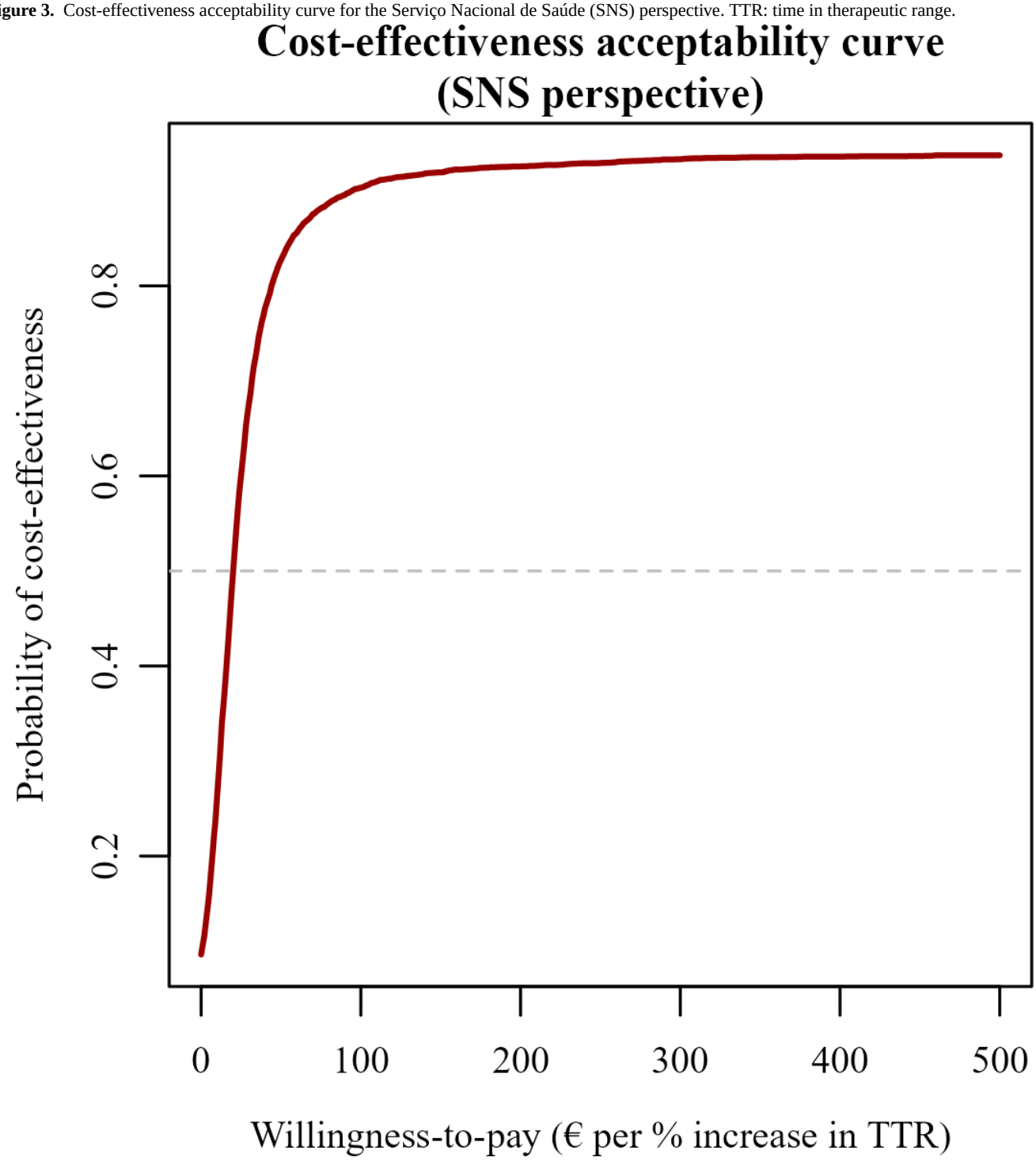

From the limited societal perspective, the base-case ICER was €25.88 per percentage point of TTR gained. The corresponding cost-effectiveness plane (Figure 4) showed a similar pattern, with 88% (4400/5000) of bootstrapped replications in the northeast quadrant, 6.5% (325/5000) in the southeast, 5% (250/5000) in the northwest, and 0.4% (20/5000) in the southwest. The CEAC (Figure 5) likewise indicated a 90% probability of cost-effectiveness at a WTP threshold of about €95 per percentage TTR gained.

XSL•FO
RenderX

Figure 4. Cost-effectiveness plane (limited societal perspective). TTR: time in therapeutic range.

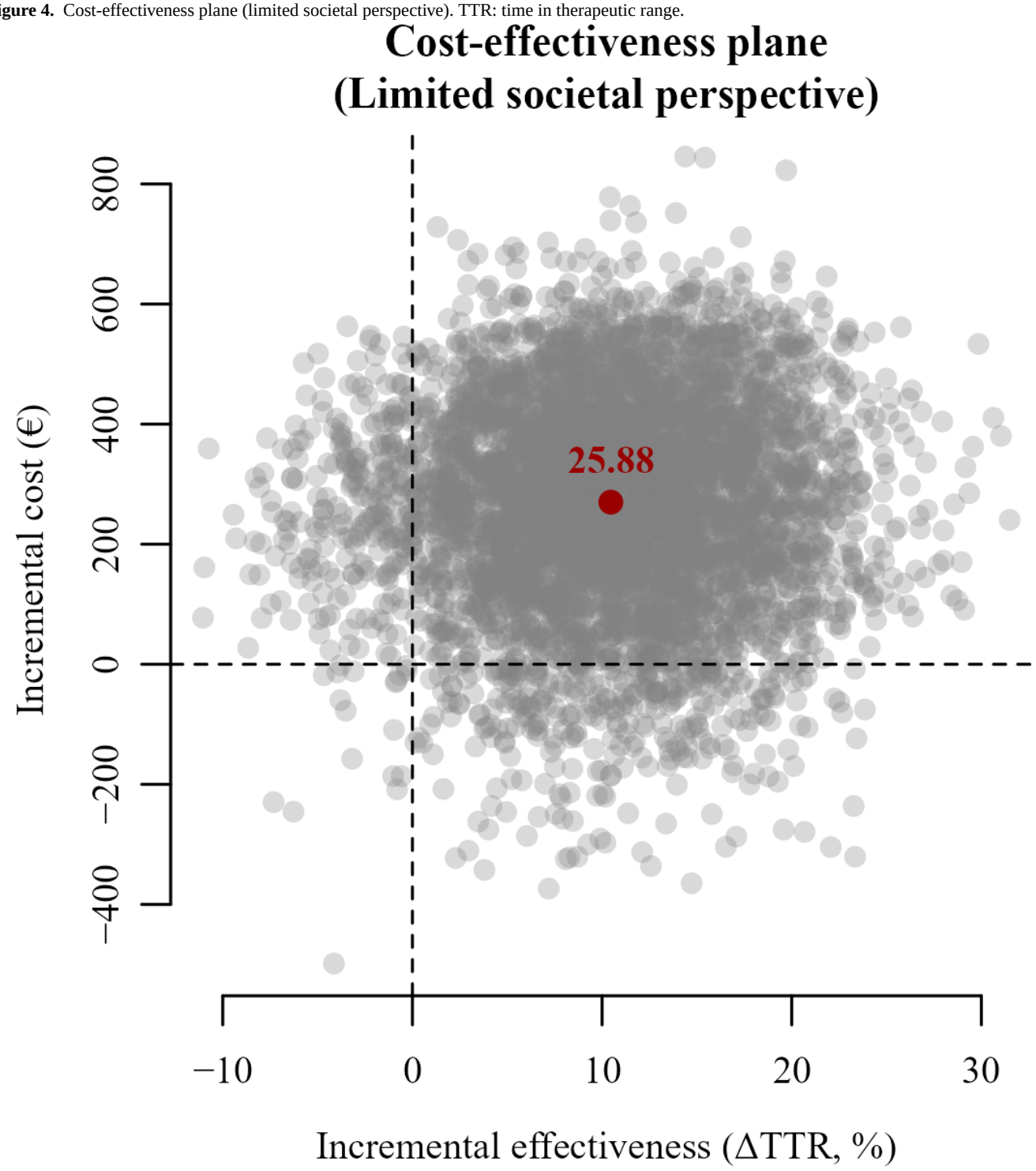

XSL•FO
RenderX

Figure 5. Cost-effectiveness acceptability curve for the limited societal perspective. TTR: time in therapeutic range

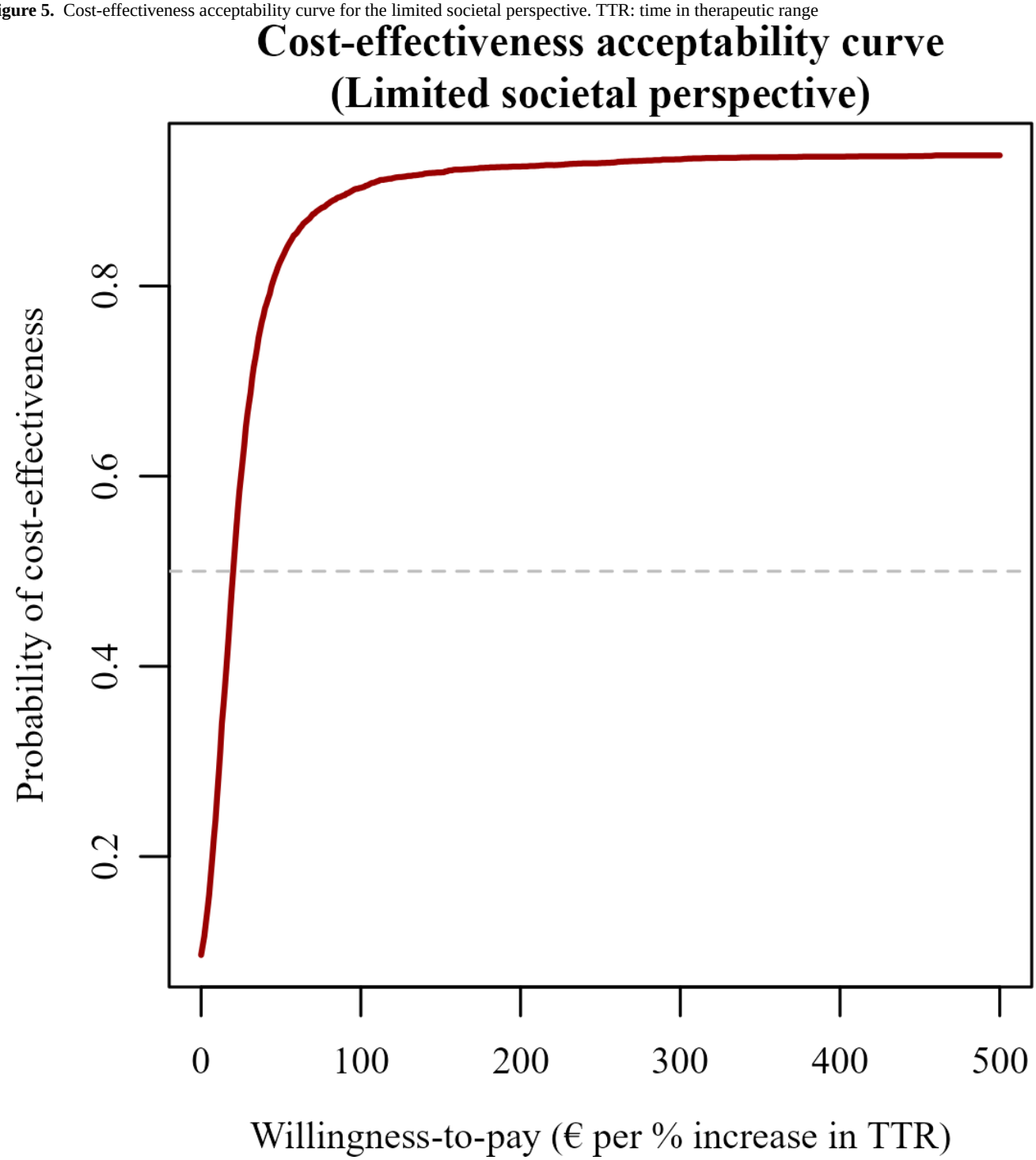

Sensitivity Analysis

One-way deterministic sensitivity analyses are summarized in Table 5. Varying the cost of the chatbot between €25 and €75 per patient shifted the ICERs between €16.65 and €21.44 per percentage point of TTR gained from the SNS perspective, and between €23.49 and €28.29 from the limited societal perspective. Extending the useful lifetime of the coagulometer from 3 to 5 or 10 years reduced the ICERs to €15.21 and €12.33 (SNS) and €22.06 and €19.18 (societal), respectively. Altering travel costs by ±20% had only a modest impact on the societal ICERs (ranging from €24.52 to €27.26). When the lower bound of the 95% CI for ΔTTR (–3.74) was applied, RPM was dominated (more costly and less effective). Conversely, using the upper bound (24.61 percentage points) substantially improved cost-effectiveness, lowering ICERs to €8.07 (SNS) and €10.97 (societal).

XSL•FO
RenderX

Table . Results of the sensitivity analysis from the Serviço Nacional de Saúde and limited societal perspectives.[a]

Scenario	ICER[b] SNS[c] perspective	ICER[b] limited societal perspective
Analysis 1: base-case	19.03	25.88
Chatbot costs €25	16.65	23.49
Chatbot costs €75	21.44	28.29
Machine lifetime 5 y	15.21	22.06
Machine lifetime 10 y	12.33	19.18
Travel costs −20%	__[d]	24.52
Travel costs +20%	__[d]	27.26
ΔTTR[e] low (95% CI)	Dominated	Dominated
ΔTTR higher (95% CI)	8.07	10.97

[a]ICERs are reported as the incremental cost per percentage point increase in TTR, comparing RPM with SOC. "Dominated" indicates that RPM is more costly and less effective than SOC.

[b]ICER: incremental cost-effectiveness ratio.

[c]SNS: Portuguese National Health Service (Serviço Nacional de Saúde).

[d]Not applicable.

[e]TTR: time in therapeutic range.

Discussion

Principal Findings

This proof-of-concept study demonstrates the feasibility and methodological value of embedding a full trial-based economic evaluation within an early-stage Tier C DHT. Rather than seeking to establish definitive cost-effectiveness, it examined whether a full economic evaluation could be integrated into a small-scale clinical study to generate actionable evidence under real-world conditions. The resulting data provide an empirical foundation for subsequent model-based analyses and support evidence-informed adoption decisions along the digital health lifecycle.

From both the SNS and limited societal perspectives, the chatbot-supported RPM solution incurred higher costs (€198.61 and €270.05, respectively) while improving anticoagulation control, with a mean within-patient gain of 10.43 percentage points in TTR. The corresponding ICERs were €19.03 and €25.88 per additional percentage point of TTR gained, respectively. As no established WTP thresholds exist for percentage-point improvements in TTR, interpretation of cost-effectiveness remains challenging. Nevertheless, the relatively low ICERs observed suggest that clinically meaningful improvements in anticoagulation control may be achieved at a modest additional cost. This is particularly relevant given previous evidence from Wan et al [26], showing that increases of 8.3% and 10.2% in TTR were associated with approximately 1 fewer major hemorrhage and thromboembolic event per 100 patient-years, respectively. In this context, the CEACs indicate that at approximately €95 per percentage TTR gained, the probability of cost-effectiveness reaches 90%, further supporting the potential economic value of the intervention.

These findings indicate where early value begins to emerge and what drives it: intervention-specific components (coagulometer, test strips, and chatbot platform) and higher follow-up intensity contributed most to incremental costs, whereas greater therapeutic stability after the 6-month postoperative period was associated with improved control. Sensitivity analyses indicated that these results were generally robust across variations in key parameters.

Analyses 2-4 helped contextualize these findings by showing that implementation timing played an important role. Differences between sequences and periods indicated that RPM tended to be more effective and potentially more cost-effective once patients had progressed beyond the immediate postoperative phase. These patterns suggest that clinical stability may be a key condition under which the intervention generates greater value.

Taken together, the base-case cost-effectiveness results, supported by sensitivity analyses and analyses (2-4), illustrate how early-stage full economic evaluations can move beyond partial assessments to provide structured learning. Such insights demonstrate not only how much an intervention costs but also for whom and under which conditions it begins to generate value, thereby informing both future economic evaluations and decisions regarding the development, refinement, and adoption of these technologies.

Comparison With Prior Work

Previous trial-based economic evaluations of self-monitoring interventions in anticoagulation management have also reported increased costs [27,32,44]. However, only 1 used TTR as the primary outcome, reporting significantly higher TTR values compared with routine care (72%, SD 19.7% vs 59%, SD 13.5%), alongside an incremental cost of €59.08 per patient over 6 months, although no ICER was reported [44]. To our knowledge, this is the first trial-based economic evaluation of a chatbot integrated into RPM for anticoagulation therapy. The cost-effectiveness of such digital health tools remains largely

XSL•FO
RenderX

unexplored. A 2022 systematic review identified only 2 economic evaluations of chatbot-based interventions and emphasized the urgent need for robust, well-reported analyses in this rapidly evolving field [45]. This is particularly relevant, as the use of chatbots within RPM offers new opportunities to enhance health care delivery, strengthen patient-provider communication, and advance patient-centered care [46].

From a lifecycle perspective, this study represents the next analytical step for this specific RPM solution and, more generally, for Tier C DHTs. The earlier TDABC analysis served as a partial economic evaluation**,** offering early feasibility insights through the mapping of patient pathways, quantification of resource use, and identification of key cost drivers related to equipment, staff, and workflow requirements [15]. Such analyses are valuable for understanding cost structures, anticipating resource needs, and laying the groundwork for more comprehensive evaluations [11]. The current trial-based full economic evaluation, based on primary data, extends that groundwork by adding comparative value evidence, informing refinement of the RPM solution, supporting the design of the demonstration-phase study, and providing the necessary parameter estimates for subsequent model-based economic evaluations. Positioned within this continuum, the study supports the methodological argument that trial-based full economic evaluations can function as a learning stage in the digital health lifecycle.

Limitations

The findings of this study should be interpreted with caution due to several limitations. The small sample size, associated nonsignificant results, single-center setting, and short follow-up period were expected features of a pilot stage but limit the interpretability and generalizability. Accordingly, the evidence should be considered hypothesis-generating rather than confirmatory.

From a design perspective, the crossover format may have introduced residual carry-over effects. Although analyses 2-4 were conducted to explore differences related to treatment order and timing, results were not fully consistent, and some variation between periods and sequences was observed. Therefore, carry-over effects cannot be excluded and may have influenced the estimation and interpretation of both outcomes and costs.

From an evaluation perspective, resource-use data were obtained exclusively from electronic health records at ULS São José, which may not capture health care use outside this institution. Moreover, because the economic evaluation was not prospectively planned, important data were not collected. Productivity costs were absent, precluding the adoption of a broader societal perspective. In addition, patient-reported outcomes, such as EQ-5D, were not collected, which prevented the estimation of quality-adjusted life years (QALYs). This limitation is particularly relevant for Tier C DHTs, which directly affect health outcomes, and evaluations should include QALY measurement to enable comparability across interventions for resource allocation decisions. However, TTR, as a clinically relevant measure of anticoagulation control, reflects the quality and effectiveness of anticoagulation management and allows comparability across similar interventions within this clinical area.

Furthermore, several aspects of the findings are context-specific, including unit costs, resource use patterns, and care pathways within the Portuguese health care setting. However, insights related to implementation timing, cost drivers, and the feasibility of embedding economic evaluation within early-stage studies may be transferable to similar contexts.

Taken together, these limitations provide valuable methodological lessons for Tier C DHTs. They highlight the need for prospective planning of economic end points, integration of QALYs, and alignment between clinical and economic data collection. In this sense, the limitations of the study are themselves instructive, illustrating the iterative learning inherent to early lifecycle economic evaluations.

Recommendations for a Lifecycle Economic Evaluation Framework for Tier C DHTs

In response to the identified gap in applying suitable economic evaluation methods across maturity stages, and drawing on the evidence generated in this pilot stage, together with its main limitations (the absence of a prospectively planned economic evaluation and the lack of EQ-5D and productivity-loss data) and further complemented by prior literature [8,11,16,47,48], several recommendations are proposed to guide the development of a lifecycle-aligned framework for the economic evaluation of Tier C DHTs:

1. Embed economic evaluation throughout the lifecycle. Economic evaluation should be viewed as an iterative process that evolves in parallel with the technology's maturity. For Tier C interventions, early stages (preprototype, prototype, pilot) should focus on feasibility and process costs through partial economic evaluations, such as TDABC, while also beginning to assess comparative effectiveness and uncertainty to inform subsequent stages. Mid stages (demonstration) should consolidate evidence on effectiveness and cost drivers, and advanced stages (scale-up and integration) should address long-term value and population impact through model-based approaches.
2. Plan the economic evaluation prospectively. Economic end points should be defined early and integrated from the outset. Prospective planning enables systematic tracking of costs and outcomes, ensures methodological alignment between clinical and economic components, and minimizes the risk of missing data that could compromise subsequent analyses.
3. Align clinical and economic outcomes. For Tier C DHTs, outcome measurement must capture both clinical and patient-reported outcomes. Instruments such as the EQ-5D should be collected alongside condition-specific indicators to enable QALY estimation, facilitate cost-utility analysis, and improve comparability across interventions and disease contexts.
4. All relevant cost categories (intervention-specific, health care payer, patient/caregiver, and productivity costs) should be defined and measured from the outset to enable a broader societal perspective and to capture the broader economic impact of Tier C DHTs.

XSL•FO
RenderX

5. Promote multidisciplinary and stakeholder collaboration. Robust economic evaluation of Tier C DHTs requires close collaboration among health economists, clinicians, digital-health developers, data scientists, and policymakers. Early and sustained engagement of these groups ensures methodological consistency, contextual relevance, and meaningful translation of findings into the development, pricing, and decisions regarding the adoption of these technologies.

Together, these recommendations outline the core principles for structuring a lifecycle economic evaluation framework for Tier C DHTs. Embedding economic evaluation across all maturity stages enables more coherent, comparable, and value-based evidence generation.

Conclusion

This Tier C proof-of-concept study demonstrates that a full trial-based economic evaluation can feasibly be embedded within an early-stage DHT, producing actionable comparative evidence under real-world conditions. The intervention resulted in improved anticoagulation control alongside higher costs, providing early insights into its cost-effectiveness profile within a pilot setting. Beyond these numerical results, the study illustrates how such evaluations operate as structured learning experiments, identifying cost and effect drivers, testing methodological feasibility, and informing model-based projections and implementation readiness.

Building on the preceding TDABC analysis, this study exemplifies the role of pilot-stage evaluations within the digital health evidence continuum. Looking forward, tier- and stage-aligned frameworks are needed to strengthen this continuum and address the persistent adoption-evidence gap. This study contributes to that goal by providing a set of recommendations specifically for Tier C DHTs, supporting more consistent and value-based economic evaluations across the digital health lifecycle.

Acknowledgments

The authors are grateful to Helena Semedo and Vanessa Rodrigues for their assistance in providing the data. The authors acknowledge the use of generative artificial intelligence tools (ChatGPT 4.5 and Grammarly) for language editing. The authors take full responsibility for the content of the manuscript.

Funding

This study was supported by the Portuguese Foundation for Science and Technology (FCT—Fundação para a Ciência e a Tecnologia, I.P.) under grant number 2024.01089.BDANA. Uppsala University provided additional support for the publication of this work. The funders had no role in the study design, data collection, analysis, interpretation, or manuscript preparation.

Authors' Contributions

ARS conceived the economic evaluation design within the pilot crossover trial. ARS, AL, FS, and JP contributed to the overall conceptualization of the study. ARS conducted the economic evaluation analysis and drafted the manuscript. JP, FS, FG, and AL provided critical guidance and detailed input on the manuscript. All authors reviewed, edited, and approved the final version.

Conflicts of Interest

ARS and FS are affiliated with the funding organization. All other authors declare no conflicts of interest.

Checklist 1
CHEERS 2022 checklist.
[PDF File, 103 KB - jmir_v28i1e85430_app1.pdf]

References

1. Kolasa K, Kozinski G. How to value digital health interventions? A systematic literature review. Int J Environ Res Public Health 2020 Mar 23;17(6):2119. [doi: 10.3390/ijerph17062119] [Medline: 32209988]
2. Segur-Ferrer J, Moltó-Puigmartí C, Pastells-Peiró R, Vivanco-Hidalgo RM. Methodological frameworks and dimensions to be considered in digital health technology assessment: scoping review and thematic analysis. J Med Internet Res 2024 Apr 10;26:e48694. [doi: 10.2196/48694] [Medline: 38598288]
3. Zah V, Burrell A, Asche C, Zrubka Z. Paying for digital health interventions-what evidence is needed? Acta Polytech Hung 2022;19(9):179-199. [doi: 10.12700/APH.19.9.2022.9.10]
4. Gomes M, Murray E, Raftery J. Economic evaluation of digital health interventions: methodological issues and recommendations for practice. Pharmacoeconomics 2022 Apr;40(4):367-378. [doi: 10.1007/s40273-022-01130-0] [Medline: 35132606]
5. Assessing the impact of digital transformation of health services. : European Commission; 2019 URL: https://health.ec.europa.eu/system/files/2019-11/022_digitaltransformation_en_0.pdf [accessed 2026-06-05]

XSL•FO
RenderX

6. Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions-a call for innovative evidence generation approaches. NPJ Digit Med 2020;3(1):110. [doi: 10.1038/s41746-020-00314-2] [Medline: 32904379]
7. Tarricone R, Petracca F, Weller HM. Towards harmonizing assessment and reimbursement of digital medical devices in the EU through mutual learning. NPJ Digit Med 2024 Oct 1;7(1):268. [doi: 10.1038/s41746-024-01263-w] [Medline: 39354125]
8. Santos AR, von Hafe F, Sampaio F, Londral AR, Perelman J. Mapping methodologies for economic evaluation of digital health technologies: a scoping review. Value Health 2026 Feb;29(2):311-323. [doi: 10.1016/j.jval.2025.09.012] [Medline: 41015334]
9. Evidence standards framework for digital health technologies. : National Institute for Health and Care Excellence (NICE); 2022 URL: https://www.nice.org.uk/corporate/ecd7/resources/evidence-standards-framework-for-digital-health-technologies-pdf-1124017457605 [accessed 2026-06-05]
10. Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment. : World Health Organization; 2016 URL: https://iris.who.int/server/api/core/bitstreams/0036e491-a9f4-450e-8943-11225d1c7ddf/content [accessed 2026-06-05]
11. LeFevre AE, Shillcutt SD, Broomhead S, Labrique AB, Jones T. Defining a staged-based process for economic and financial evaluations of mHealth programs. Cost Eff Resour Alloc 2017;15:5. [doi: 10.1186/s12962-017-0067-6] [Medline: 28428734]
12. Tarricone R, Petracca F, Cucciniello M, Ciani O. Recommendations for developing a lifecycle, multidimensional assessment framework for mobile medical apps. Health Econ 2022 Sep;31(Suppl 1):73-97. [doi: 10.1002/hec.4505] [Medline: 35388585]
13. Marcus HJ, Ramirez PT, Khan DZ, et al. The IDEAL framework for surgical robotics: development, comparative evaluation and long-term monitoring. Nat Med 2024 Jan;30(1):61-75. [doi: 10.1038/s41591-023-02732-7] [Medline: 38242979]
14. Chapel B, Alexandre F, Heraud N, et al. Standardization of the assessment process within telerehabilitation in chronic diseases: a scoping meta-review. BMC Health Serv Res 2022 Aug 2;22(1):984. [doi: 10.1186/s12913-022-08370-y] [Medline: 35918690]
15. Guede-Fernández F, Silva Pinto T, Semedo H, et al. Enhancing postoperative anticoagulation therapy with remote patient monitoring: a pilot crossover trial study to evaluate portable coagulometers and chatbots in cardiac surgery follow-up. Digit Health 2024;10:20552076241269515. [doi: 10.1177/20552076241269515] [Medline: 39139188]
16. Piloting digital technologies in a health or care service. National Institute for Health and Care Excellence (NICE). URL: https://web.archive.org/web/20260307155221/https://digitalregulations.innovation.nhs.uk/regulations-and-guidance-for-adopters/all-adopters-guidance/piloting-digital-technologies-in-a-health-or-care-service/ [accessed 2026-06-08]
17. Van Gelder IC, Rienstra M, Bunting KV, et al. 2024 ESC Guidelines for the management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS). Eur Heart J 2024 Sep 29;45(36):3314-3414. [doi: 10.1093/eurheartj/ehae176] [Medline: 39210723]
18. Altiok E, Marx N. Oral anticoagulation. Dtsch Arztebl Int 2018 Nov 16;115(46):776-783. [doi: 10.3238/arztebl.2018.0776] [Medline: 30602410]
19. Chen A, Stecker E, Warden BA. Direct oral anticoagulant use: a practical guide to common clinical challenges. J Am Heart Assoc 2020 Jul 7;9(13):e017559. [doi: 10.1161/JAHA.120.017559] [Medline: 32538234]
20. Barcellona D, Fenu L, Marongiu F. Point-of-care testing INR: an overview. Clin Chem Lab Med 2017 May 1;55(6):800-805. [doi: 10.1515/cclm-2016-0381] [Medline: 27754958]
21. Liu S, Shen G, Li W. Structural and cellular basis of vitamin K antagonism. J Thromb Haemost 2022 Sep;20(9):1971-1983. [doi: 10.1111/jth.15800] [Medline: 35748323]
22. Sharma P, Scotland G, Cruickshank M, et al. Is self-monitoring an effective option for people receiving long-term vitamin K antagonist therapy? A systematic review and economic evaluation. BMJ Open 2015 Jun 25;5(6):e007758. [doi: 10.1136/bmjopen-2015-007758] [Medline: 26112222]
23. Huang Y, Xie Y, Huang L, Han Z. The value of anticoagulation management combining telemedicine and self-testing in cardiovascular diseases: a meta-analysis of randomized controlled trials. Ther Clin Risk Manag 2023;19:279-290. [doi: 10.2147/TCRM.S395578] [Medline: 36941980]
24. Macaluso GP, Pagani FD, Slaughter MS, et al. Time in therapeutic range significantly impacts survival and adverse events in destination therapy patients. ASAIO J 2022 Jan 1;68(1):14-20. [doi: 10.1097/MAT.0000000000001572] [Medline: 34524147]
25. Vestergaard AS, Skjøth F, Larsen TB, Ehlers LH. The importance of mean time in therapeutic range for complication rates in warfarin therapy of patients with atrial fibrillation: a systematic review and meta-regression analysis. PLoS ONE 2017;12(11):e0188482. [doi: 10.1371/journal.pone.0188482] [Medline: 29155884]
26. Wan Y, Heneghan C, Perera R, et al. Anticoagulation control and prediction of adverse events in patients with atrial fibrillation: a systematic review. Circ Cardiovasc Qual Outcomes 2008 Nov;1(2):84-91. [doi: 10.1161/CIRCOUTCOMES.108.796185] [Medline: 20031794]
27. Phibbs CS, Love SR, Jacobson AK, et al. At-home versus in-clinic INR monitoring: a cost–utility analysis from the home INR study (THINRS). J Gen Intern Med 2016 Sep;31(9):1061-1067. [doi: 10.1007/s11606-016-3700-8] [Medline: 27234663]

XSL•FO
RenderX

28. Husereau D, Drummond M, Augustovski F, et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations. Value Health 2022 Jan;25(1):3-9. [doi: 10.1016/j.jval.2021.11.1351] [Medline: 35031096]
29. Perelman J, Soares M, Duarte A, et al. Methodological guidelines for economic evaluation studies of health technologies. : INFARMED - National Authority of Medicines and Health Products; 2019 URL: https://www.infarmed.pt/documents/15786/4001413/Orienta%C3%A7%C3%B5es+metodol%C3%B3gicas+para+estudos+de+avalia%C3%A7%C3%A3o+econ%C3%B3mica+de+tecnologias+de+sa%C3%BAde+%28EN%29/ebcfd930-94e2-c7e1-100a-ee1df3d76882 [accessed 2025-02-28]
30. van Lier LI, Bosmans JE, van Hout HPJ, et al. Consensus-based cross-European recommendations for the identification, measurement and valuation of costs in health economic evaluations: a European Delphi study. Eur J Health Econ 2018 Sep;19(7):993-1008. [doi: 10.1007/s10198-017-0947-x] [Medline: 29260341]
31. Bech M, Christiansen T, Gyrd-Hansen D. Handling value added tax (VAT) in economic evaluations: should prices include VAT? Appl Health Econ Health Policy 2006;5(4):209-213. [doi: 10.2165/00148365-200605040-00003] [Medline: 17249838]
32. Jowett S, Bryan S, Murray E, et al. Patient self-management of anticoagulation therapy: a trial-based cost-effectiveness analysis. Br J Haematol 2006 Sep;134(6):632-639. [doi: 10.1111/j.1365-2141.2006.06243.x] [Medline: 16938120]
33. Contrato-programa 2023 - CHULC EPE [Report in Portuguese]. : Administração Central do Sistema de Saúde (ACSS); 2023 URL: https://www.ulssjose.min-saude.pt/wp-content/uploads/sites/3/2023/01/2023-Acordo-Modificativo-2023-CHULC.pdf
34. Macedo A, Carrasco J, Andrade S, Moital I. Incremental cost evaluation study of INR self-monitoring versus laboratorial control or monitoring in health care centres, in Portugal. Acta Med Port 2010;23(2):203-212. [Medline: 20470467]
35. Portaria n.o 207/2017, de 11 de julho [Article in Portuguese]. Diário da República. URL: https://diariodarepublica.pt/dr/detalhe/portaria/207-2017-107669157 [accessed 2025-08-28]
36. INFOMED - base de dados para medicamentos de uso humano [Article in Spanish]. URL: https://extranet.infarmed.pt/INFOMED-fo/ [accessed 2025-07-01]
37. Resumo das características do medicamento: varfine 5 mg comprimidos [Report in Portuguese]. : Teofarma S.r.l.; 2023 URL: https://www.teofarmasrl.it/wp-content/uploads/2021/01/PT_Varfine_RCM.pdf [accessed 2026-06-05]
38. Resumo das características do medicamento: sintrom 4 mg comprimidos [Report in Portuguese]. : Agencia Española de Medicamentos y Productos Sanitarios (AEMPS); 2023 URL: https://cima.aemps.es/cima/pdfs/es/ft/25670/25670_ft.pdf [accessed 2026-06-05]
39. Portaria n.o 1553-d/2008, de 31 de dezembro [Website in Portuguese]. Diário da República. 2008. URL: https://diariodarepublica.pt/dr/detalhe/portaria/1553-d-2008-243733 [accessed 2025-04-20]
40. Decreto-lei n.o 137/2010, de 28 de dezembro [Website in Portuguese]. Diário da República. 2010. URL: https://diariodarepublica.pt/dr/detalhe/decreto-lei/137-2010-306615 [accessed 2025-04-20]
41. Atualização de valores com base no IPC [Article in Portuguese]. Instituto Nacional de Estatística (INE) - Statistics Portugal. URL: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ipc [accessed 2025-06-26]
42. Rosendaal FR, Cannegieter SC, van der Meer FJ, Briët E. A method to determine the optimal intensity of oral anticoagulant therapy. Thromb Haemost 1993 Mar 1;69(3):236-239. [Medline: 8470047]
43. Kaatz S. Determinants and measures of quality in oral anticoagulation therapy. J Thromb Thrombolysis 2008 Feb;25(1):61-66. [doi: 10.1007/s11239-007-0106-9] [Medline: 17906916]
44. Gallagher J, Mc Carthy S, Woods N, Ryan F, O' Shea S, Byrne S. Economic evaluation of a randomized controlled trial of pharmacist-supervized patient self-testing of warfarin therapy. J Clin Pharm Ther 2015 Feb;40(1):14-19. [doi: 10.1111/jcpt.12215] [Medline: 25295834]
45. Fgaier M, Zrubka Z. Cost-effectiveness of using chatbots in healthcare: a systematic review. Presented at: 2022 IEEE 22nd International Symposium on Computational Intelligence and Informatics and 8th IEEE International Conference on Recent Achievements in Mechatronics, Automation, Computer Science and Robotics (CINTI-MACRo); Nov 21-22, 2022. [doi: 10.1109/CINTI-MACRo57952.2022.10029478]
46. Milne-Ives M, de Cock C, Lim E, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res 2020 Oct 22;22(10):e20346. [doi: 10.2196/20346] [Medline: 33090118]
47. Kowatsch T, Otto L, Harperink S, Cotti A, Schlieter H. A design and evaluation framework for digital health interventions. Inf Technol 2019 Oct 25;61(5-6):253-263. [doi: 10.1515/itit-2019-0019]
48. Jumamyradov M, Craig BM. Measuring effectiveness based on patient experience (instead of QALYs) in US value assessments. Pharmacoeconomics 2025 Feb;43(2):171-176. [doi: 10.1007/s40273-024-01444-1] [Medline: 39487899]

Abbreviations

CEA: cost-effectiveness analysis
CEAC: cost-effectiveness acceptability curve
CHEERS: Consolidated Health Economic Evaluation Reporting Standards
DHTs: digital health technologies

XSL•FO
RenderX

ICER: incremental cost-effectiveness ratio
INFARMED: Autoridade Nacional do Medicamento e Produtos de Saúde, I.P. [National Authority of Medicines and Health Products]
INR: international normalized ratio
NICE: National Institute for Health and Care Excellence
QALYs: quality-adjusted life years
RPM: remote patient monitoring
SNS: Serviço Nacional de Saúde [Portuguese National Health Service]
SOC: standard of care
TDABC: time-driven activity-based costing
TTR: time in therapeutic range
VKAs: vitamin K antagonists
WTP: willingness-to-pay

Edited by M Balcarras; submitted 14.Nov.2025; peer-reviewed by K Khashu, V Palama; revised version received 08.May.2026; accepted 01.Jun.2026; published 17.Jun.2026.

Please cite as:
Santos AR, Sampaio F, Guede-Fernández F, Perelman J, Londral A
Cost-Effectiveness Analysis of Chatbot-Supported Remote Patient Monitoring for Anticoagulation Management: Health Economic Evaluation Within a Pilot Crossover Trial
J Med Internet Res 2026;28:e85430
URL: https://www.jmir.org/2026/1/e85430
doi:10.2196/85430

© Ana Rita Santos, Filipa Sampaio, Federico Guede-Fernández, Julian Perelman, Ana Londral. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Remote Patient Monitoring in Louisiana Medicare Beneficiaries With Diabetes or Hypertension: Retrospective Cohort Study of 2016-2020 Claims Data

Debra Winberg[1,2], MA, MSc, PhD; Charles Stoecker[2], MA, PhD; Jian Li[3], PhD; Lizheng Shi[2], MA, MsPharm, PhD

[1]Department of Health Management and Policy, School of Health, Georgetown University, St Mary's Hall, 3700 Reservoir Road, Washington, DC, United States

[2]Department of Health Policy and Management, Celia Scott Weatherhead School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA, United States

[3]Department of Biostatistics, Celia Scott Weatherhead School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA, United States

Corresponding Author:
Debra Winberg, MA, MSc, PhD
Department of Health Management and Policy, School of Health, Georgetown University, St Mary's Hall, 3700 Reservoir Road, Washington, DC, United States

Abstract

Background: Chronic conditions such as diabetes and hypertension pose major public health and economic challenges. Innovative tools, including remote patient monitoring (RPM), aim to address barriers to care by digitally transmitting health data to providers.

Objective: This study assesses the impact of RPM on health care utilization in Medicare beneficiaries with diabetes or hypertension.

Methods: This retrospective cohort study used a longitudinal panel from 2016 to 2020 based on Louisiana Medicare Fee-for-Service claims and Beneficiary Files. Patients were included if they were diagnosed with diabetes or hypertension before 2016. RPM exposure was defined by common procedure terminology/healthcare common procedure coding system codes (eg, 99453, 99454, 99457, 99458, 95250, 95251, 99091, E2100-E2103, E0607), with the corresponding month considered the treatment month. Control beneficiaries were randomly assigned treatment months matching the treatment group. Outcomes were monthly counts per 1000 beneficiaries of outpatient, emergency department, and inpatient visits (all-cause and disease-specific), constructed via revenue center/place-of-service/primary ICD-10 (International Classification of Diseases-10) codes. We estimated inverse-probability-treatment-weighted 2-way fixed-effects models with standard errors clustered at the physician level. Data represent secondary analysis of deidentified claims; no participant recruitment occurred. All confidence intervals are reported at the 95% confidence level.

Results: The study included 5488 beneficiaries in the treatment group and 341,226 in the control group. RPM was associated with increased level of outpatient visits in the treatment group by 2284.01 (95% CI 2059.75-2508.27) visits per month per 1000 people ($P<.001$) after initiation and decreased the trend of outpatient visits by 37.61 (95% CI −37.98 to −37.24) visits per month per 1000 people compared with the control group ($P<.001$). Similarly, the level of emergency room visits increased in the treatment compared with the control by 84.60 (95% CI 72.21 to 96.99) visits per month per 1000 people initially after RPM initiation ($P<.001$) and decreased the trend by 1.40 (95% CI −1.60 to −1.20) visits per month per 1000 people ($P<.001$). The level of inpatient visits increased by 32.64 (95% CI 24.84-40.44) visits per month per 1000 beneficiaries in the treatment group compared with the control immediately after RPM prescription ($P=.002$), but the trend decreased by 0.54 (95% CI −0.66 to −0.42) visits per 1000 people each month ($P<.001$).

Conclusions: This study is among the first large, claims-based evaluations of RPM in older adults with diabetes or hypertension using a beneficiary- and time-fixed-effects design. We identify a pattern of increased health care utilization following RPM initiation, followed by a gradual decline that does not fully offset the initial increase. By characterizing utilization trajectories and not clinical endpoints, this study helps reconcile inconsistent findings in prior RPM research and highlights the importance of implementation factors in Medicare populations. Strategies to maximize the value of RPM may include appropriate population targeting, patient education on digital health technologies, and clinician workflow support to support effective integration into routine care.

(J Med Internet Res 2026;28:e80782) doi:10.2196/80782

XSL•FO
RenderX

KEYWORDS

remote patient monitoring; diabetes; hypertension; health utilization; medicare

Introduction

Chronic conditions, including diabetes and hypertension, are significant public health challenges. In the past 30 years, the prevalence of diabetes has increased by 11%, and in 2024, more than 30 million Americans were diagnosed with diabetes [1]. Similarly, hypertension affects approximately 47% of Americans, or over 120 million people [2,3]. Both diseases are major contributors to other chronic health concerns, including cardiovascular disease, chronic kidney disease, stroke, and premature mortality, which can also increase rates of health care resource utilization (HCRU). More so, the economic impact is significant. Both diabetes and hypertension are leading causes of morbidity and mortality in the United States, with combined mortality rates nearly doubling from 2000 to 2023 [4]. In 2022, hypertension-related costs exceeded US $219 billion annually, and diabetes direct costs were over US $300 billion [5-9].

This burden disproportionately impacts older adults over 65, who face higher rates of both conditions in addition to higher rates of complications, multimorbidity, and polypharmacy [7,10]. In turn, Medicare bears much of the burden from these conditions [7].

There are many successful pharmacological treatments and evidence-based lifestyle interventions for diabetes and hypertension, yet over half of patients do not have their blood sugar or blood pressure under control [11,12]. Alarmingly, control rates have worsened in recent years, with hypertension control decreasing from 54% in 2013 to 48% in 2017 - 2020 and continuing to decrease during the COVID-19 pandemic [13,14]. Similarly, glycemic control has stagnated in the past 30 years, and disparities in control have increased [15]. These trends highlight persistent barriers in chronic disease management, including poor medication adherence, limited access to specialty care, high costs, social and economic challenges, and low patient engagement in care [16]. Innovative digital tools, including remote patient monitoring (RPM), aim to address these barriers and improve disease control. RPM is an mHealth device that aims to improve patient care by digitally transmitting health data from patients to physicians [17]

By reducing patient travel costs, allowing physicians to treat more patients, facilitating practice management, and providing health care providers access to patient data quickly, these devices target several barriers to glycemic control, including nonadherence, self-management, and health care experiences [18-20]. RPM adoption in the United States has accelerated in recent years, driven by expanded reimbursement policies, technological advances in wearables and connected devices, and the growing need to manage chronic disease more effectively [21]. These factors have positioned RPM as one of the most commonly used digital health tools, particularly for conditions such as diabetes and hypertension.

Evidence regarding RPM's effectiveness is positive, but inconsistent. A 2024 systematic review on RPM interventions found measurable impacts in chronic disease outcomes, though effects varied based on a number of variables, including the specific disease and implementation design [22]. A review of systematic reviews found that there are possible benefits from RPM, but results are inconsistent [23]. For example, several trials and meta-analyses report A1c reductions by 0.3% - 1.22% and decreases in systolic blood pressure of around 4 - 9 mmHg among RPM cohorts [17,24-26]. While these are modest changes, they can be clinically meaningful and suggest RPM can support better chronic disease control.

However, not all evidence is positive; some studies demonstrate that the RPM has no impact and may only be effective for certain patient groups [27-33]. Specifically, RPM may only benefit specific subgroups, such as people with a high baseline risk, who monitor very often, or who are very engaged in their care and have physicians who are very engaged in their care [34]. Additionally, there is mixed evidence on the ability of RPM to decrease HCRU [35,36]. Some studies found that RPM can prevent costly acute events and reduce hospitalizations and emergency room (ER) visits and may be cost-effective for hypertensive patients [37]. However, other studies demonstrated no change or increased utilization patterns, potentially attributable to the detection of previously unaddressed clinical needs or enhanced surveillance following RPM deployment [38]. Overall, the effectiveness of RPM and the balance between increased monitoring and reduced acute care use remain unclear.

Despite the rapid expansion of RPM coverage and use, most of the existing literature focuses on younger and commercially insured populations who may not have such a high burden of diabetes and hypertension and have fewer barriers to using RPM, which may impact its effectiveness [39,40]. While the evidence base in Medicare beneficiaries exists, it remains limited and often focuses on other conditions such as heart disease and chronic obstructive pulmonary disease (COPD) [35,41]. A recent systematic review found that most of the studies on RPM focus on usability and feasibility, not outcome measures such as utilization, for cardiovascular disease [42]. Therefore, not enough real-world evidence studies have been performed for diabetes and hypertension. As most general RPM services in Medicare beneficiaries are for diabetes and hypertension, understanding their impact on this population is essential. Additionally, as some of the Medicare literature shows a more negative impact, such as increased health utilization, it is essential to analyze more specific Medicare populations [38].

This study aims to assess the impact of RPM on HCRU in Medicare beneficiaries with diabetes or hypertension. By focusing on an older, higher-risk population, our analysis contributes novel real-world evidence on the effectiveness of RPM and provides timely insights for policymakers and providers considering broader adoption.

XSL•FO
RenderX

Methods

Overview

This is a retrospective cohort study using Louisiana Medicare Fee-for-Service claims linked to beneficiary files and analyzed using an inverse probability treatment weighted (IPTW) controlled interrupted time series (CITSA) design. This manuscript was prepared following the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines [43].

Data Source and Patient Selection

We used data from a patient-level longitudinal panel from January 1, 2016, to December 31, 2020, which was constructed from Louisiana Medicare Fee-for-Service claims and Beneficiary files. This data is collected at the time of service and was purchased from the Research Assistance Data Center for secondary analyses. We included patients who had a diagnosis of diabetes or hypertension based on the CMS Chronic Conditions files. Patients were excluded if they did not have a diabetes or hypertension diagnosis before 2016, were not enrolled in Medicare before 2016, had only gestational diabetes or secondary diabetes, or had end-stage renal disease. These exclusions were found using the ICD codes.

Patients were included in the treatment group if they had at least one occurrence of a common procedure terminology (CPT) code for RPM or continuous glucose monitoring (99453, 99454, 99457, 99458, 95250, 95251, 99091, E2100, E0607, E2102, E2103, E2101). This information was found in the inpatient, ER, or outpatient claims files. We defined the treatment month as the first month with this CPT code. For the control group, pseudo-treatment dates were randomly assigned to reflect the distribution of treatment initiation in the intervention group. Specifically, in Excel, each month and year combination was coded numerically, and controls were assigned a treatment month and year in proportion to the observed distribution of initiation dates in the treatment group. This ensured alignment by time of year rather than only by year, reducing bias from differential calendar-time exposure. After the assignment, we conducted balance checks and confirmed that the month-by-month distribution of treatment dates did not significantly differ between the control and treatment groups.

Outcome Variables

The number of outpatient, ER, or inpatient visits per month. We defined these outcomes using published guidance from the Research Data Assistance Center using revenue center codes, claim dates, and places of service codes in the Medicare Fee-for-Service files (Carrier claims, Carrier Line, Base Claims, and Revenue Center). The disease-specific outcomes included the number of outpatient, ER, or inpatient visits per month where the primary ICD code indicated that the visit was for diabetes or hypertension. If a beneficiary did not have an encounter in a certain month, they were coded as having 0 visits. Visits with RPM as the primary CPT code were not included. All outcomes were detrended to remove underlying time trends to focus on deviations that may be associated with RPM.

Statistical Analysis

We used a CITSA to determine the impact of RPM. This approach leverages pre- and posttreatment data from a treatment and control group to estimate changes in the level and slope of the outcome variables before and after treatment initiation. This method assesses the impact of an intervention such as RPM by comparing a treatment to a control group over time. CITSA was chosen for the analysis because of the linearity of the data over time (Multimedia Appendix 1) and for its ability to account for pre-existing trends and provide a counterfactual estimate. The CITSA model may be written as:

$$Y=\beta_0+\beta_1(\text{Time})+\beta_2(\text{RPM})+\beta_3(\text{Time}*\text{RPM})+\beta_4(\text{post}*\text{RPM})+\beta_5(\text{Time}*\text{RPM}*\text{post})+\beta_6X+$$

The baseline level is given by the intercept β_0. β_1 represents the underlying time trend for the non-RPM group before the intervention. The indicator variable RPM specifies group membership, with β_2 capturing any baseline differences between the intervention and control groups. The term β3 allows for differences in preintervention trends between RPM and non-RPM groups. The coefficient β_4 represents the immediate postintervention effect. The 3-way interaction term β_5 captures any differential trend change attributable to the intervention over time. There was no specific postperiod variable included in the model, as the control group had a randomly assigned treatment date and, as such, had no expected postperiod outcome change.

To account for bias and confounding, we adjusted for covariates that may influence the outcome, including comorbidities (diabetes, hypertension, COPD, chronic kidney disease, hyperlipidemia), Medicare status, sex, race (White, Black, Hispanic, Asian, Other), rurality (defined as metropolitan, micropolitan, rural using FIPS codes and 2020 core-based statistical area data), and age. These variables were created using the Medicare Beneficiary Files (chronic condition and ABCD files), which aggregate data at the annual level.

All analyses accounted for time- and beneficiary-level fixed effects and were weighted by the IPTW, and standard errors were clustered at the primary doctor level to account for serial correlations over time. To construct the weights, we modeled the probability of RPM participation using baseline demographic and clinical characteristics from the month before treatment and the average preperiod outcomes. Each beneficiary was assigned a weight equal to the inverse of their predicted probability. This data was obtained from the beneficiary files and the previously created outcome variables. The primary doctor was defined as the physician whom a beneficiary saw the most during the study period to account for correlated practice patterns. In the base case, the main model used all available data in the 60 months regardless of pre- and postlength of time.

We used an inverse-probability–weighted 2-way fixed-effects model to control for unobserved, time-invariant differences across beneficiaries and shared temporal shocks. This design isolates within-beneficiary changes before and after RPM initiation, consistent with the interrupted time series framework. It was preferred to multilevel models because exposure timing varies across beneficiaries, and fixed effects more clearly estimate pre/post level and trend differences [44].

XSL•FO
RenderX

Because Medicare claims are used for reimbursement, demographic, enrollment, and utilization fields are complete for almost all beneficiaries. Diagnosis and procedure variables used to construct comorbidities are also not subject to item-level analysis. Multiple imputation was not performed, as missingness was minimal and largely structural and less than 15% for all key variables.

All confidence intervals were calculated at the 95% confidence level. We tested the robustness of the results by clustering standard errors by zip code and changing the necessary length of pre- and posttreatment data to 6 months pre and 6 months post, and 12 months pre and 12 months post. Additionally, we assess seasonality of treatment by running the analysis with a variable for calendar month (January, February, etc) and another analysis with a control variable for calendar month of treatment. Additionally, we ran an analysis excluding COVID years to ensure we were not picking up COVID confounding and to account for potential exposure classification issues due to new CPT codes being introduced.

All analyses were conducted using Stata (version 18.0; IBM Corp).

Ethical Considerations

This study was approved by Tulane University's Institutional Review Board (2023-1044-SPHTM) through an expedited review, as it only uses secondary data. The data use agreement and institutional review board for this study allow the use of secondary data analysis without additional consent from beneficiaries. All study data is deidentified to protect beneficiaries, and there was no way to reidentify. As we did not work directly with beneficiaries, no compensation was given, and there is no identification of individuals in this manuscript.

Results

Initially, the data included 480,592 Louisiana beneficiaries with diabetes or hypertension. This sample included 472,902 beneficiaries in the control group. Of these, 118,513 were excluded for being diagnosed with hypertension or diabetes after 2016. A further 9 beneficiaries were dropped for being enrolled in Medicare too late, and 130 were excluded for only having secondary diabetes. Last, 20,226 beneficiaries were dropped for having end-stage renal disease.

Initially, 7690 Medicare beneficiaries were identified in the treatment group. Of these, 1849 beneficiaries were excluded because they were diagnosed with hypertension or diabetes after 2016. No beneficiaries were excluded due to delayed Medicare enrollment, 2 beneficiaries were excluded for having secondary diabetes only, and 351 beneficiaries were excluded due to end-stage renal disease. After applying all inclusion and exclusion criteria, a total of 346,714 Medicare beneficiaries with diabetes or hypertension were included in the analytic sample. The final sample comprised 5488 beneficiaries in the treatment group and 341,226 in the control group. Among the treated beneficiaries, 719 (13.1%) initiated RPM in 2016, 635 (11.6%) in 2017, 770 (14.0%) in 2018, 1346 (24.5%) in 2019, and 2018 (36.8%) in 2020. These distributions remained balanced after weighting.

Table 1 presents the weighted and unweighted baseline characteristics of the treatment and control groups. After weighting, all standardized mean differences were below 10%. Prior to weighting, the treatment group had a higher prevalence of several chronic conditions, including diabetes (3842, 70% vs 112,605, 33%), hypertension (5049, 92% vs 245,683, 72%), COPD (494, 9% vs 23,886, 7%), chronic heart failure (823, 15% vs 30,710, 9%), chronic kidney disease (2031, 37% vs 47,772, 14%), and hyperlipidemia (2524, 46% vs 88,719, 26%).

The RPM group was younger on average (mean age 71.6 vs 74.2 years), included a higher proportion of male beneficiaries (2470, 45% vs 126,254, 37%), and a higher proportion of White beneficiaries (3842, 70% vs 208,148, 61%). In contrast, beneficiaries in the treatment group were less likely to have dual Medicaid eligibility (1317, 24% vs 139,903, 41%). The treatment group also demonstrated higher baseline health care utilization, including emergency department visits, inpatient diabetes-related visits, outpatient visits, and outpatient diabetes-related visits. After weighting, baseline demographic characteristics, clinical conditions, and health care utilization were comparable between the treatment and control groups.

XSL•FO
RenderX

Table . Baseline population characteristics and health care utilization. Baseline demographic, clinical, and health care utilization characteristics of Medicare Fee-for-Service beneficiaries in Louisiana with hypertension or diabetes, 2016-2020. Patients in the treatment group initiated remote patient monitoring or continuous glucose monitoring based on relevant procedure codes. Data are unweighted and weighted using inverse probability treatment weight, which were created using demographic and utilization data. Values are shown as means (SD) or percentages, with standardized mean differences (SMD) presented for both unweighted and weighted comparisons. Utilization outcomes are average monthly counts in the preintervention period.

Variable	Nonweighted			Weighted		
	Treatment mean (SD) (n=5488)	Control mean (SD) (n=341,226)	SMD	Treatment mean (SD) (n=5488)	Control mean (SD) (n=341,226)	SMD
Diabetes	0.70 (0.46)	0.46 (0.47)	−0.81	0.70 (0.46)	0.71 (0.45)	0.02
Hypertension	0.92 (0.27)	0.72 (0.45)	−0.54	0.92 (0.27)	0.93 (0.26)	0.02
Chronic obstructive pulmonary disease	0.09 (0.29)	0.07 (0.25)	−0.1	0.09 (0.29)	0.10 (0.30)	0.01
Heart failure	0.15 (0.36)	0.09 (0.29)	−0.19	0.15 (0.36)	0.16 (0.36)	0.01
Chronic kidney disease	0.37(0.48)	0.14 (0.35)	−0.54	0.37 (0.48)	0.36 (0.48)	−0.01
Hyperlipidemia	0.46 (0.50)	0.26 (0.44)	−0.44	0.46 (0.50)	0.46 (0.50)	−0.01
Myocardial infarction	0.01 (0.11)	0.00 (0.07)	−0.08	0.01 (0.11)	0.01 (0.11)	0
Charlson Comorbidity Index	3.88 (2.52)	2.46 (2.50)	−0.56	3.88 (2.52)	4.10 (3.15)	0.08
Age	71.58 (10.36)	74.15 (11.27)	0.24	71.58 (10.36)	71.72 (10.56)	0.01
Dual status	0.24 (0.43)	0.41 (0.49)	0.36	0.24 (0.43)	0.27 (0.44)	0.08
Percent male	0.45 (0.50)	0.37 (0.48)	−0.17	0.45 (0.50)	0.44 (0.50)	−0.02
Percent White	0.70 (0.46)	0.61 (0.49)	−0.19	0.70 (0.46)	0.70 (0.46)	0.00
Percent Black	0.26 (0.44)	0.23 (0.42)	−0.07	0.26 (0.44)	0.26 (0.44)	0.00
Percent Hispanic	0.02 (0.14)	0.02 (0.13)	−0.03	0.02 (0.14)	0.02 (0.14)	−0.01
Percent Asian	0.01 (0.09)	0.01 (0.08)	−0.02	0.01 (0.09)	0.01 (0.09)	−0.01
Percent metropolitan	0.83 (0.38)	0.75 (0.43)	−0.19	0.83 (0.38)	0.84 (0.37)	−0.01
Percent micropolitan	0.09 (0.29)	0.14 (0.35)	0.15	0.09 (0.29)	0.09 (0.28)	−0.01
Percent rural	0.08 (0.27)	0.11 (0.31)	0.11	0.08 (0.27)	0.08 (0.27)	−0.01
Inpatient visits	0.03 (0.06)	0.03 (0.06)	−0.04	0.03 (0.06)	0.03 (0.07)	0.09
Inpatient visits (hypertension)	0.00 (0.01)	0.00 (0.01)	−0.03	0.00 (0.01)	0.00 (0.01)	0.02
Inpatient visits (diabetes)	0.00 (0.02)	0.00 (0.01)	−0.13	0.00 (0.02)	0.00 (0.01)	−0.05
Emergency room visits	0.01 (0.03)	0.00 (0.01)	−0.22	0.01 (0.03)	0.00 (0.03)	−0.08
Emergency room visits (diabetes)	0.00 (0.02)	0.00 (0.01)	−0.06	0.00 (0.02)	0.00 (0.02)	−0.01
Emergency room visits (hypertension)	0.07 (0.10)	0.06 (0.09)	−0.12	0.07 (0.10)	0.08 (0.11)	0.06
Outpatient visits	2.34 (1.60)	1.53 (1.52)	−0.52	2.34 (1.60)	2.27 (1.92)	−0.04
Outpatient visits (diabetes)	0.34 (0.41)	0.08 (0.20)	−0.83	0.34 (0.41)	0.29 (0.60)	−0.10
Outpatient visits (hypertension)	0.14 (0.17)	0.12 (0.21)	−0.08	0.14 (0.17)	0.14 (0.21)	0.01

Figures 1-3 display residualized monthly outpatient, ER, and inpatient visits by treatment status. These figures plot model residuals after covariate adjustment and illustrate underlying utilization patterns. For outpatient visits (Figure 1), the treatment

XSL•FO
RenderX

group increased from about 0.4 residual visits per month preintervention to a peak near 1.2 around the index date, before declining to approximately 0.5 by the end of follow-up. The control group residuals hovered close to 0 throughout. For ER visits (Figure 2), the treatment group rose to roughly 0.035 residual visits per month at the peak, declining back toward zero postintervention. In contrast, the control group stayed flat at approximately 0 residual visits. For inpatient visits (Figure 3), the treatment group showed a smaller preintervention increase, reaching about 0.018 residual visits per month, followed by a decline toward baseline. Again, the control group remained stable near 0 across the entire period. Residual plots for disease-specific outcomes can be found in Multimedia Appendix 1.

Figure 1. Residualized monthly outpatient visits before and after remote patient monitoring initiation. Models were adjusted for age, sex, race, rurality, comorbidities (diabetes, hypertension, chronic obstructive pulmonary disease, chronic kidney disease, hyperlipidemia), and Medicare status.

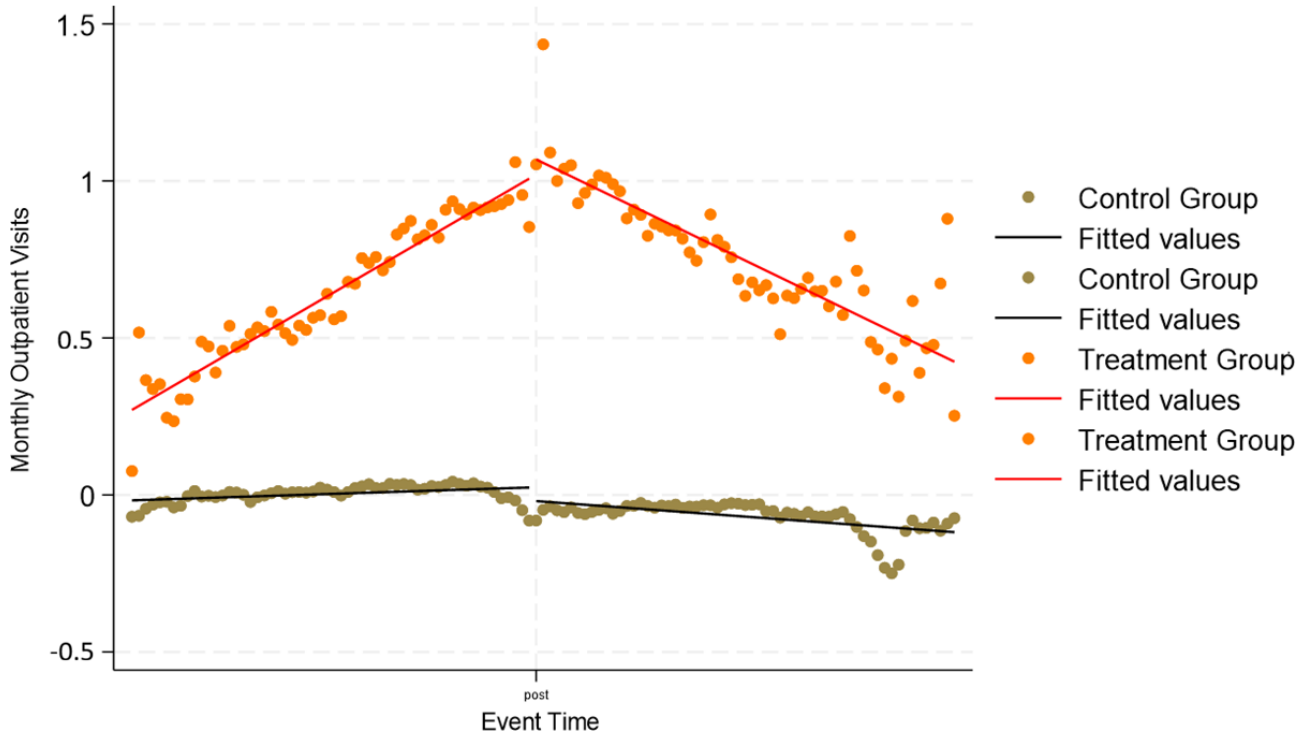

Figure 2. Residualized monthly emergency room visits before and after remote patient monitoring initiation. Models were adjusted for age, sex, race, rurality, comorbidities (diabetes, hypertension, chronic obstructive pulmonary disease, chronic kidney disease, hyperlipidemia), and Medicare status.

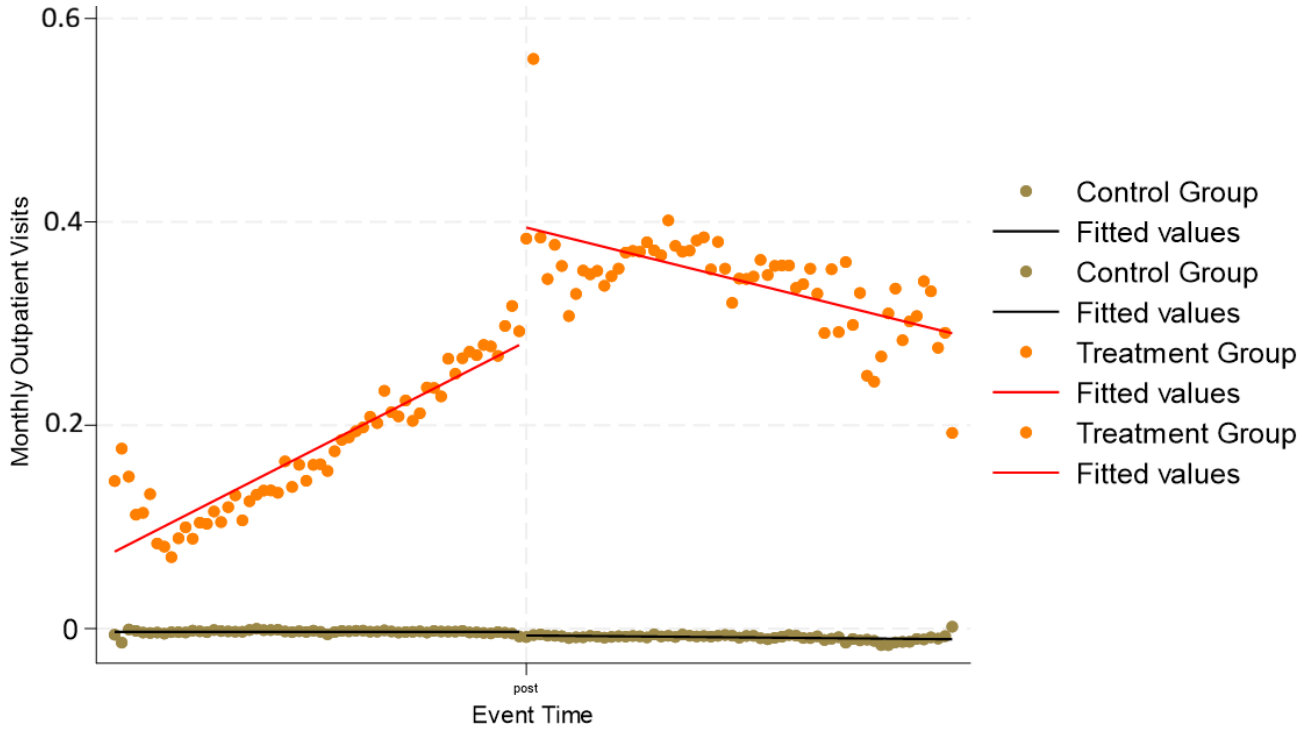

XSL•FO
RenderX

Figure 3. Residualized monthly inpatient visits before and after remote patient monitoring initiation. Models were adjusted for age, sex, race, rurality, comorbidities (diabetes, hypertension, chronic obstructive pulmonary disease, chronic kidney disease, hyperlipidemia), and Medicare status.

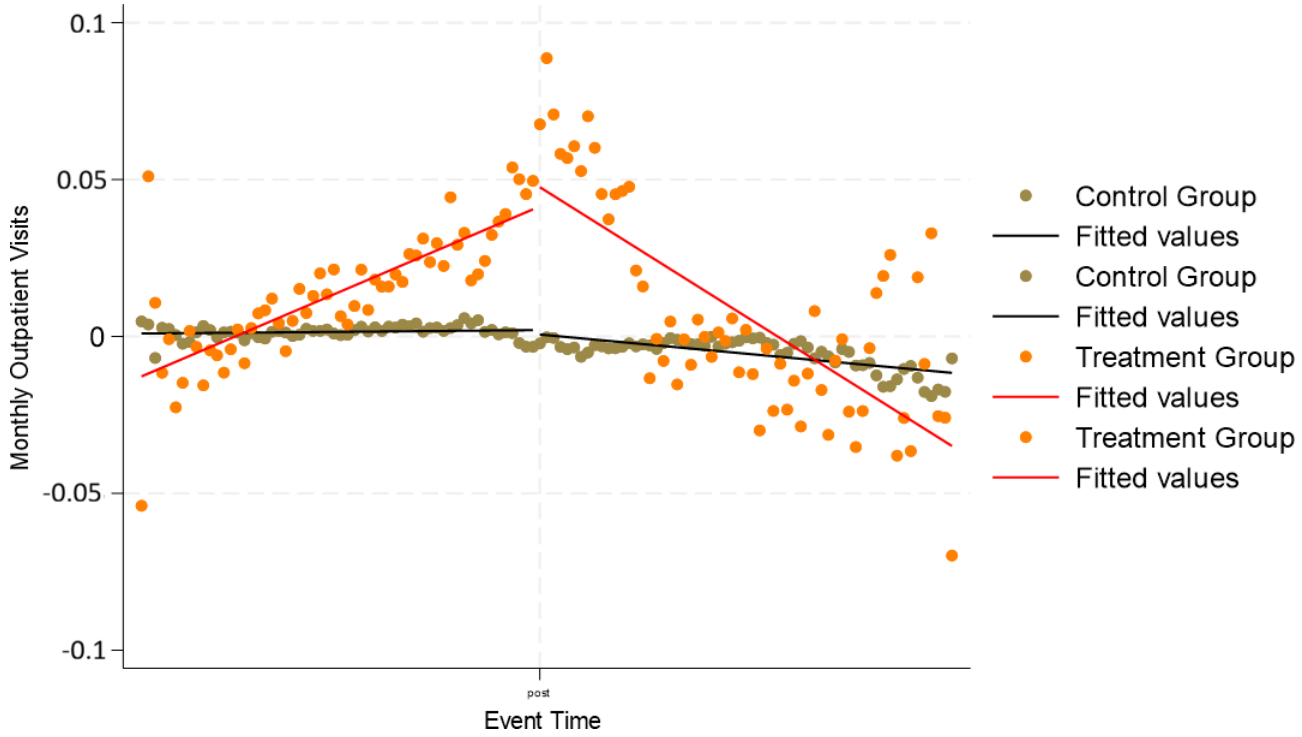

Table 2 depicts the CITSA results. All-cause outpatient visits steadily increased by 9.81 (95% CI 2.42-17.20; *P*=.01) visits per month per 1000 people during the time period. In the preperiod, beneficiaries in the treatment group had their visits decrease by 10.83 (95% CI −18.45 to 3.21; *P*=.01) visits per month per 1000 people due to treatment. After RPM initiation, the level of outpatient visits increased by 2284.01 (95% CI 2059.75-2508.27) visits per month per 1000 people (*P*<.001), and the trend in the number of outpatient visits decreased by 37.61 (95% CI −37.98to −37.24) visits per month per 1000 people compared with those in the control group (*P*<.001). Diabetes-specific visits did not change significantly over the time period or in the pretreatment time period. The level of diabetes-specific appointments increased by 673.26 (95% CI 586.98-759.54) visits per month per 1000 people (*P*<.001) after initiation and then decreased at a higher rate of 10.25 (95% CI −11.39 to −9.11) visits per month per 1000 people (*P*<.001). Over the entire study period, hypertension outpatient visits increased by 0.47 (95% CI 0.31 to 0.63) visits per month per 1000 people (*P*<.001), but did not significantly change in the pretreatment period. The level of hypertension visits increased by 190.92 (95% CI 136.06-245.78) visits in the RPM group compared with those in the control group (*P*<.001), while the trend decreased by 2.91 (95% CI −3.58 to −2.2) visits per 1000 people over the next several years.

All-cause ER visits increased by 0.27 (95% CI 0.17-0.37) visits per month per 1000 people over the entire study period (*P*<.001). In the pretreatment period, visits decreased by 0.3 (95% CI −0.44 to −0.16) visits per month per 1000 people (*P*<.001). Compared with the control, the level of number of ER visits increased by 84.60 (95% CI 72.21-96.99) visits per 1000 people initially after RPM initiation (*P*<.001) and the trend decreased by 1.40 (95% CI −1.60 to −1.20) visits per month per 1000 people (*P*<.001). While the overall time trend and pretreatment trend are insignificant, the patterns also hold for disease-specific ER visits after RPM initiation. At the time of RPM, the level of ER visits increased by 14.58 (95% CI 10.70-18.46) visits per month per 1000 beneficiaries and had a decreasing trend with 0.26 (95% CI −0.32 to −0.20) fewer visits per month per 1000 people (*P*<.001 for all outcomes) in the treatment compared with the control. Like diabetes-specific visits, there were no statistically significant overall time or pretreatment trends. The level of hypertension-specific visits increased by 5.68 (95% CI 2.56-8.80) visits per month for 1000 people after RPM initiation in the treatment group, compared with the control, and the trend decreased by 0.09 (95% CI −0.13 to −0.05) visits per month per 1000 people (*P*<.001) in the RPM group compared with the control.

XSL•FO
RenderX

Table . Impact of the RPM on health care resource utilization. All estimates are expressed as monthly visits per 1000 beneficiaries. Diabetes and hypertension-specific visits are based on the primary diagnosis code associated with the visit. Each cell contains the treatment effect and confidence intervals. Impact of remote patient monitoring (RPM) on monthly outpatient, emergency room, and inpatient visits among Medicare beneficiaries with hypertension or diabetes, Louisiana, 2016-2020. Estimates from inverse probability treatment weight-controlled interrupted time series models are shown with 95% confidence intervals and *P* values.

Variables	All-cause visits		Diabetes visits		Hypertension visits	
	OR[a] (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Outpatient visits						
Time trend	9.81 (2.42 to 17.20)	.01	5.89 (−1.05 to 12.83)	.10	0.47 (0.31 to 0.63)	<.001
Pretreatment trend	−10.83 (−18.45 to 3.21)	.01	−6.23 (−13.19 to 0.73)	.08	−0.28 (−0.79 to 0.23)	.22
Posttreatment level	2284.01 (2059.75 to 2508.27)	<.001	673.26 (586.98 to 759.54)	<.001	190.92 (136.06 to 245.78)	<.001
Posttreatment trend	−37.61 (−37.98 to −37.24)	<.001	−10.25 (−11.39 to −9.11)	<.001	−2.91 (−3.58 to −2.24)	<.001
Emergency room visits						
Time trend	0.27 (0.17 to 0.37)	<.001	0.00 (0.00 to 0.00)	.48	0.00 (−0.02 to 0.02)	.62
Pretreatment trend	−0.30 (−0.44 to −0.16)	<.001	−0.01 (−0.03 to 0.01)	.66	0.00 (−0.02 to 0.02)	.87
Posttreatment level	84.60 (72.21 to 96.99)	<.001	14.58 (10.70 to 18.46)	<.01	5.68 (2.56 to 8.80)	<.001
Posttreatment trend	−1.40 (−1.60 to −1.20)	<.001	−0.26 (−0.32 to −0.20)	<.001	−0.09 (−0.13 to −0.05)	<.001
Inpatient visits						
Time trend	0.21 (0.17 to 0.25)	<.001	0.01 (0.01 to 0.01)	<.01	0.00 (−0.02 to 0.02)	.86
Pretreatment trend	−0.23 (−0.31 to −0.15)	<.001	−0.01 (−0.03 to 0.01)	.23	−0.01 (−0.03 to 0.01)	.48
Posttreatment level	32.64 (24.84 to 40.44)	<.001	5.08 (2.67 to 7.49)	<.001	2.82 (0.43 to 5.21)	.02
Posttreatment trend	−0.54 (−0.66 to −0.42)	<.001	−0.10 (−0.32 to −0.20)	<.001	−0.04 (−0.08 to −0.00)	<.001
Observations	20,799,900	—[b]	20,799,900	—	20,799,900	—

[a]OR: odds ratio.

[b]Not applicable.

All-cause inpatient visits increased by 0.21 (95% CI 0.17-0.25) visits per month per 1000 people over the entire study period (*P*<.001). In the pretreatment period, the trend decreased by −0.23 (95% CI −0.31 to −0.15) visits per month per 1000 people (*P*<.001). Compared with the control group, the level of all-cause inpatient visits increased by 32.64 (95% CI 24.84-40.44) visits per month per 1000 people in the treatment group compared with the control group immediately after RPM prescription (*P*<.001). The trend decreased by 0.54 (95% CI −0.66 to −0.42) visits per 1000 people each month (*P*<.001). During the study period, inpatient diabetes visits increased by 0.01 (95% CI 0.01-0.01) visits per month per 1000 people (*P*<.001) but were not statistically significant in the preperiod. The level of inpatient visits with diabetes as the primary diagnosis code increased by 5.08 (95% CI 2.67-7.49) visits per month per 1000 people right after treatment (*P*<.001), and the trend decreased by 0.01 (95% CI 2.51-2.31) visits per month per 1000 people (*P*<.001) in the RPM group compared with the control. Hypertension-specific inpatient visit levels increased 2.83 (95% CI 0.43-5.21) visits per month per 1000 people immediately after treatment (*P*=.02), and the trend decreased by 0.04 (95% CI −0.08 to 0.00) visits per month per 1000 (*P*<.001).

The results were robust to cluster changes and excluding COVID-19 months (Multimedia Appendix 2). Additionally, when assessing seasonality, both analyses remained statistically significant with the same overall patterns; however, the point estimates slightly decreased (Multimedia Appendix 2), showing there may be some seasonality. The results were not robust to changing the study time (limited to 6 or 12 mo).

XSL•FO
RenderX

Discussion

Principal Findings

This study aimed to assess the impact of RPM on HCRU on Louisiana Medicare beneficiaries with diabetes or hypertension from 2016 to 2020. By combining a rigorous longitudinal design with a focus on an older, high-burden Medicare population, this study provides novel insights into how RPM adoption affects health care utilization in real-world settings. Overall, in beneficiaries with diabetes or hypertension, RPM was associated with a statistically significant increase in outpatient, ER, and inpatient visits immediately after initiation, followed by a significant downward trend over time. These findings, supported by narrow confidence intervals, suggest that the estimates are precise and due to RPM, not random variation.

Existing evidence on RPM's impact on utilization is mixed. Some studies report decreases in HCRU [35,45]. Others have shown that it may not change utilization [17]. Notably, studies in a Medicare population also demonstrated an increase in utilization; for example, one study on Medicare hypertensive patients reported an increase in cardiovascular outpatient visits [38]. Differences across studies may reflect variation in patient populations, RPM modalities, and social determinants of health. For example, our Medicare population may have lower digital or health literacy.

Additionally, the average beneficiary in our cohort was on RPM for 7 months, and results were not robust to a short time period, which is consistent with literature that found that clinical outcomes, such as decreased blood glucose, which may impact utilization, are only found in longer time periods [46]. Thus, long-term users may be driving the observed effects, underscoring the importance of sustained participation for realizing RPM's full impact.

The dynamic trend of early increases followed by subsequent declines in utilization aligns with the literature, suggesting that individuals initially engage more intensively with new tools before stabilizing into routine patterns [47,48]. In the initial phase of RPM adoption, both patients and providers may be adjusting to new data and workflows. Patients may respond to frequent health alerts or health data with heightened awareness, prompting additional interactions with the health system. Providers may proactively schedule follow-ups to verify readings, resolve technical issues, or re-evaluate treatment plans due to the additional data.

The subsequent decrease might indicate that as patients become more accustomed to the data, they take a more measured approach to seeking care. It remains unclear whether these patterns represent appropriate, preventive care or unnecessary and wasteful utilization. The increase could stem from useful data and engagement that identify unmet needs and ensure patients are getting necessary care, or it could represent low-value encounters due to the large influx of data or cautious clinical behavior. Without information on clinical appropriateness, clinical outcomes, and more information on the visits, it is not possible to determine whether the short-term increase benefits patients and the health system.

This observed pattern likely reflects a complex interaction of patient and provider behaviors shaped by differences in digital literacy, clinical workflows, and familiarity with RPM. For instance, improved patient self-management is one of the key benefits of RPM and can lead to better outcomes, but it is dependent on patients' ability to understand health data [49,50]. Both health and digital literacy are associated with age and education [50]. Given that this study focused primarily on older adults in a state with relatively low average education levels, some patients may have lacked the digital or health literacy required to meaningfully engage with RPM. This could limit the intervention's potential benefits or even contribute to uncertainty-driven care-seeking rather than reducing visits.

Additionally, our study focused on early RPM adopters and physicians who were still learning effective RPM implementation methods. Clinicians were still learning how to interpret continuous data, integrate RPM into workflows, and communicate with patients. A systematic review of factors of RPM success highlights several physician-level factors for effectiveness, including targeting populations at high risk, providing responsive and timely care, and ensuring collaborative and coordinated care [45]. If providers are learning how to provide timely and coordinated care via new technologies, RPM may not have its intended benefit. Together, these patient- and provider-level factors suggest that the short-term rise in utilization may reflect a broader adaptation period as both groups learned to navigate new technologies and patterns of care.

Moreover, this study focuses specifically on Louisiana Medicare. The state has a relatively high chronic disease burden, an older average population, and low levels of health and digital literacy [51,52]. These are all factors that can slow early adoption and increase initial contact with providers as patients and clinicians adapt to new tools. At the same time, Louisiana's high poverty rate and large rural population may limit broadband access [53], contributing to declining utilization after initial implementation phases stabilize. Finally, reimbursement for RPM under other payers (Medicaid and commercial programs) was only gradually expanded in Louisiana after 2019 as well, constraining early uptake and shaping provider behavior during the study period. These contextual features suggest that both the learning period and structural factors unique to Louisiana likely influenced the temporal dynamics observed in this study.

From a policy and clinical perspective, these findings have dual and uncertain implications. While increased short-term utilization may raise costs, if the increased visits reflect timely, preventive, or necessary care, the investment could yield downstream savings through decreased complications and their associated treatments. Conversely, if the increased encounters are duplicative or not clinically necessary, they could add costs. Policymakers and payers should therefore interpret early utilization spikes with caution. To prove that these visits should be encouraged, future research should further assess costs and clinical endpoints, which can clarify whether observed utilization patterns translate into value-based outcomes.

In the meantime, these results do offer some recommendations on how to maximize the benefits of RPM and potential challenges that need to be mitigated. First, providers should

XSL•FO
RenderX

prioritize patient education initiatives; ensure that patients, especially older adults in Medicare, receive training on how to collect, interpret, and respond to RPM data. This finding is supported by a recent systematic review showing that RPM was more effective when patients had education and support in interpreting data [22]. Education programs can help enhance patient engagement and self-management while preventing the misuse of health care resources. Additionally, health care providers need structured support, including training on integrating RPM into care delivery and systems in place, to ensure that they have the resources to analyze RPM data.

Limitations

As this study is observational in nature, the ability to assess causality is limited due to the potential for confounding and the lack of randomization. There is an inherent bias in how patients receive RPM. While the methodology attempts to overcome this, it may not have been able to completely eliminate biases in patient selection, which could bias the results away from the null because RPM patients are already engaging with the health system. Similarly, we handled confounding through IPTW and using controls, but some confounding may still be present. Additionally, the findings have limited generalizability as the population with diabetes and hypertension in Louisiana Medicare may differ from other populations. Therefore, these results may not be applicable when assessing or developing policy for other settings, including younger patients, other insurance types, or in other states.

Similarly, this study focused only on RPM through 2020, which was before the large increase in RPM and telehealth utilization due to COVID-19, and the study results may not be the same as those reported today, as some of the barriers mentioned here are less of a concern, resulting in results biased towards the null.

Moreover, this study did not assess the duration of RPM utilization or the number of times a physician checked the RPM data. Claims data do not allow assessment of patient engagement with RPM, and this study only looks at RPM prescription, not true utilization, as an outcome, although actual use mitigates the impact. Therefore, this might bias results towards the null if this group is not engaging with the technology or bias upwards if this population is engaged with RPM technology.

Similarly, the RPM CPT code does not indicate the type of device a patient is using, and as a result, these findings cannot be applied to certain RPM types, such as continuous glucose monitoring, and represent the larger impact of RPM as a treatment option. Finally, a significant limitation of this study is that it did not assess the appropriateness of utilization or the impact on clinical outcomes. If the increase in utilization stems from patients receiving timely care or appropriate care, health outcomes may improve.

This study is innovative in its use of comprehensive real-world evidence from statewide Medicare claims to evaluate RPM among older adults with diabetes or hypertension, a population underrepresented in prior studies. The analysis distinguishes short-term onboarding effects from longer-term stabilization in health care use, expanding on potential reasons for inconsistent findings in earlier research. Conducted during a period of expanding RPM reimbursement, these results provide timely insights for clinicians and payers on targeting, education, and workflow integration to optimize RPM value and reduce low-value encounters.

Conclusions

This study assessed the association between RPM and HCRU among Louisiana Medicare beneficiaries with diabetes or hypertension. We found that RPM was associated with a modest but statistically significant increase in health care utilization, highlighting important questions about how RPM is implemented and integrated into routine care. This study contributes to the existing literature by examining RPM at scale in a real-world Medicare population using administrative claims data, rather than focusing on short-term clinical outcomes or small pilot programs. Unlike prior studies that primarily evaluate changes in biometric measures, our findings underscore the need to consider utilization patterns as an early indicator of how RPM influences care delivery. Although the current analysis could not assess the clinical appropriateness or downstream value of these encounters, the results suggest that initial increases in utilization may reflect enhanced monitoring, care coordination, or unmet need. Future work should link claims with clinical and cost data, examine patient engagement and digital literacy, and identify implementation strategies that ensure RPM supports high-value, equitable care delivery.

Acknowledgments

ChatGPT and Curie were used to edit the manuscript for clarity by asking GPT to analyze specific sections and given the prompt: “For this section please suggest edits for clarity or let me know how to improve the writing”. Additionally, ChatGPT was used to debug Stata code. The authors reviewed, verified, and take full responsibility for all content.

Funding

The authors received no financial support for this research. Therefore, there is no funder involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the CMS Research Identifiable File Data Use Agreement, which prohibits the sharing of raw data.

XSL•FO
RenderX

Authors' Contributions

DW led the conceptualization of the study, with support from LZ, CS, and JL. DW also led the formal analysis, supported by CS, and led the development of the methodology, with contributions from LZ and CS. Supervision was led by LZ, with support from CS and JL. DW prepared the original draft of the manuscript. CS, LZ, and JL contributed equally to the review and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Residual Plots.
[DOCX File, 376 KB - jmir_v28i1e80782_app1.docx]

Multimedia Appendix 2
Sensitivity Analyses.
[DOCX File, 26 KB - jmir_v28i1e80782_app2.docx]

References

1. National diabetes statistics report. : Centers for Disease Control; 2024.
2. Martin SS, Aday AW, Almarzooq ZI, et al. 2024 heart disease and stroke statistics: a report of US and global data from the American heart association. Circulation 2024 Feb 20;149(8):e347-e913. [doi: 10.1161/CIR.0000000000001209] [Medline: 38264914]
3. Hypertension. Centers for Disease Control and Prevention (CDC). 2024. URL: https://www.cdc.gov/nchs/fastats/hypertension.htm [accessed 2025-01-15]
4. Siddiqui H, Imran Z, Ali D, et al. A rising crisis: escalating burden of diabetes mellitus and hypertension-related mortality trends in the United States, 2000-2023. Clin Cardiol 2025 Jul;48(7):e70167. [doi: 10.1002/clc.70167] [Medline: 40600774]
5. Cairns C, Kang K. National hospital ambulatory medical care survey: 2021 emergency department summary tables. Centers for Disease Control and Prevention (CDC). 2021. URL: https://ftp.cdc.gov/pub/health_statistics/nchs/dataset_documentation/nhamcs/doc21-ed-508.pdf [accessed 2024-08-01]
6. Mills KT, Stefanescu A, He J. The global epidemiology of hypertension. Nat Rev Nephrol 2020 Apr;16(4):223-237. [doi: 10.1038/s41581-019-0244-2]
7. American Diabetes Association. Economic costs of diabetes in the U.S. in 2017. Diabetes Care 2018 May 1;41(5):917-928. [doi: 10.2337/dci18-0007]
8. Kirkland EB, Heincelman M, Bishu KG, et al. Trends in healthcare expenditures among US adults with hypertension: national estimates, 2003-2014. J Am Heart Assoc 2018 May 30;7(11):e008731. [doi: 10.1161/JAHA.118.008731] [Medline: 29848493]
9. Wang Y, Lee JS, Pollack LM, Kumar A, Honeycutt S, Luo F. Health care expenditures and use associated with hypertension among U.S. adults. Am J Prev Med 2024 Dec;67(6):820-831. [doi: 10.1016/j.amepre.2024.07.005]
10. Nicholson K, Liu W, Fitzpatrick D, et al. Prevalence of multimorbidity and polypharmacy among adults and older adults: a systematic review. Lancet Healthy Longev 2024 Apr;5(4):e287-e296. [doi: 10.1016/S2666-7568(24)00007-2] [Medline: 38452787]
11. Fang M, Wang D, Coresh J, Selvin E. Trends in diabetes treatment and control in U.S. adults, 1999–2018. N Engl J Med 2021 Jun 10;384(23):2219-2228. [doi: 10.1056/NEJMsa2032271]
12. Smith TM. A troubling trend: BP control drops in adults with hypertension. American Medical Association. 2022. URL: https://www.ama-assn.org/delivering-care/hypertension/troubling-trend-bp-control-drops-adults-hypertension#:~:text=The%20percentage%20of%20U.S.%20adults,67.7%25%20in%202017%E2%80%932020 [accessed 2024-10-12]
13. Rifkin DE. Lost in translation: why are rates of hypertension control getting worse over time? Am J Kidney Dis 2024 Jan;83(1):101-107. [doi: 10.1053/j.ajkd.2023.06.004]
14. Brush JE Jr, Lu Y, Liu Y, et al. Hypertension trends and disparities over 12 years in a large health system: leveraging the electronic health records. JAHA 2024 May 7;13(9). [doi: 10.1161/JAHA.123.033253]
15. Venkatraman S, Echouffo-Tcheugui JB, Selvin E, Fang M. Trends and disparities in glycemic control and severe hyperglycemia among US adults with diabetes using insulin, 1988-2020. JAMA Netw Open 2022 Dec 1;5(12):e2247656. [doi: 10.1001/jamanetworkopen.2022.47656]
16. Gonçalves A, Campos MSA, Menezes LD, Pereira LRL. Barriers and facilitators to medication adherence in chronic diseases: a scoping review. Ciênc saúde coletiva 2025;30(2). [doi: 10.1590/1413-81232025302.02762023en]
17. Leo DG, Buckley BJR, Chowdhury M, et al. Interactive remote patient monitoring devices for managing chronic health conditions: systematic review and meta-analysis. J Med Internet Res 2022 Nov 3;24(11):e35508. [doi: 10.2196/35508]

XSL•FO
RenderX

18. El-Rashidy N, El-Sappagh S, Islam SMR, M El-Bakry H, Abdelrazek S. Mobile health in remote patient monitoring for chronic diseases: principles, trends, and challenges. Diagnostics (Basel) 2021 Mar 29;11(4):607. [doi: 10.3390/diagnostics11040607] [Medline: 33805471]
19. Woodhouse AG, Orvin C, Rich C, Crosby J, Keedy CA. Diabetes outcomes before and during telehealth advancements surrounding COVID-19. J Am Pharm Assoc (2003) 2022 Jan;62(1):214-217. [doi: 10.1016/j.japh.2021.09.011]
20. Coye MJ, Haselkorn A, DeMello S. Remote patient management: technology-enabled innovation and evolving business models for chronic disease care. Health Aff (Millwood) 2009 Jan;28(1):126-135. [doi: 10.1377/hlthaff.28.1.126]
21. Paul MM, Khera N, Elugunti PR, et al. The state of remote patient monitoring for chronic disease management in the United States. J Med Internet Res 2025 Jun 26;27:e70422. [doi: 10.2196/70422] [Medline: 40570326]
22. Tan SY, Sumner J, Wang Y, Wenjun Yip A. A systematic review of the impacts of remote patient monitoring (RPM) interventions on safety, adherence, quality-of-life and cost-related outcomes. npj Digit Med 2024;7(1). [doi: 10.1038/s41746-024-01182-w]
23. Wartenberg C, Elden H, Frerichs M, et al. Clinical benefits and risks of remote patient monitoring: an overview and assessment of methodological rigour of systematic reviews for selected patient groups. BMC Health Serv Res 2025 Jan 23;25(1):133. [doi: 10.1186/s12913-025-12292-w] [Medline: 39849519]
24. Smith W, Colbert BM, Namouz T, Caven D, Ewing JA, Albano AW. Remote patient monitoring is associated with improved outcomes in hypertension: a large, retrospective, cohort analysis. Healthcare (Basel) 2024;12(16):1583. [doi: 10.3390/healthcare12161583]
25. Sachmechi I, Salam S, Amini M, Khan R, Spitznogle A, Belen T. Frequent monitoring of blood glucose levels via a remote patient monitoring system helps improve glycemic control. Endocr Pract 2023 Jun;29(6):441-447. [doi: 10.1016/j.eprac.2023.03.270] [Medline: 36965657]
26. Dennis L, Koyfman I. Effect of remote patient monitoring on stage 2 hypertension. Am J Manag Care 2025 May 6;31(9):e244-e248. [doi: 10.37765/ajmc.2025.89742] [Medline: 40834201]
27. Katalenich B, Shi L, Liu S, et al. Evaluation of a remote monitoring system for diabetes control. Clin Ther 2015 Jun 1;37(6):1216-1225. [doi: 10.1016/j.clinthera.2015.03.022] [Medline: 25869625]
28. Muller AE, Berg RC, Jardim PSJ, Johansen TB, Ormstad SS. Can remote patient monitoring be the new standard in primary care of chronic diseases, post-COVID-19? Telemed J E Health 2022 Jul;28(7):942-969. [doi: 10.1089/tmj.2021.0399] [Medline: 34665645]
29. Wei NJ, Nathan DM, Wexler DJ. Glycemic control after hospital discharge in insulin-treated type 2 diabetes: a randomized pilot study of daily remote glucose monitoring. Endocr Pract 2015 Feb;21(2):115-121. [doi: 10.4158/EP14134.OR] [Medline: 25148814]
30. Li Y, Jiang Y, Tang Y. Is remote blood pressure monitoring and management a better approach for patients with hypertension? A narrative review. J of Clinical Hypertension 2023 Feb;25(2):121-126 [FREE Full text] [doi: 10.1111/jch.14624]
31. Noah B, Keller MS, Mosadeghi S, et al. Impact of remote patient monitoring on clinical outcomes: an updated meta-analysis of randomized controlled trials. npj Digit Med 2018;1(1):20172. [doi: 10.1038/s41746-017-0002-4] [Medline: 31304346]
32. Zhang Y, Peña MT, Fletcher LM, Lal L, Swint JM, Reneker JC. Economic evaluation and costs of remote patient monitoring for cardiovascular disease in the United States: a systematic review. Int J Technol Assess Health Care 2023 Apr 28;39(1):e25. [doi: 10.1017/S0266462323000156] [Medline: 37114456]
33. Mehta SJ, Volpp KG, Troxel AB, et al. Remote blood pressure monitoring with social support for patients with hypertension. JAMA Netw Open 2024 Jun 3;7(6):e2413515. [doi: 10.1001/jamanetworkopen.2024.13515]
34. De Lathauwer ILJ, Nieuwenhuys WW, Hafkamp F, et al. Remote patient monitoring in heart failure: a comprehensive meta‐analysis of effective programme components for hospitalization and mortality reduction. Eur J Heart Fail 2025 Sep;27(9):1670-1685. [doi: 10.1002/ejhf.3568] [Medline: 39834044]
35. Taylor ML, Thomas EE, Snoswell CL, Smith AC, Caffery LJ. Does remote patient monitoring reduce acute care use? a systematic review. BMJ Open 2021 Mar;11(3):e040232. [doi: 10.1136/bmjopen-2020-040232]
36. Jackson TN, Sreedhara M, Bostic M, et al. Telehealth n the United States: A systematic review and meta-analysis, 2011–2021. Telemed Rep 2023;4(1):67-86. [doi: 10.1089/tmr.2023.0011] [Medline: 37283852]
37. De Guzman KR, Snoswell CL, Taylor ML, Gray LC, Caffery LJ. Economic evaluations of remote patient monitoring for chronic disease: a systematic review. Value Health 2022 Jun;25(6):897-913. [doi: 10.1016/j.jval.2021.12.001]
38. Acharya M, Ali MM, Bogulski CA, et al. Association of remote patient monitoring with mortality and healthcare utilization in hypertensive patients: a medicare claims-based study. J Gen Intern Med 2024 Apr;39(5):762-773. [doi: 10.1007/s11606-023-08511-x] [Medline: 37973707]
39. Zhang Z, Henley T, Schiaffino M, et al. Older adults' perceptions of community-based telehealth wellness programs: a qualitative study. Inform Health Soc Care 2022 Oct 2;47(4):361-372. [doi: 10.1080/17538157.2021.2006198] [Medline: 34822311]
40. Satariano WA, Scharlach AE, Lindeman D. Aging, place, and technology: toward improving access and wellness in older populations. J Aging Health 2014 Dec;26(8):1373-1389. [doi: 10.1177/0898264314543470] [Medline: 25502245]

XSL•FO
RenderX

41. Akar JG, Bao H, Jones PW, et al. Use of remote monitoring is associated with lower risk of adverse outcomes among patients with implanted cardiac defibrillators. Circ Arrhythm Electrophysiol 2015 Oct;8(5):1173-1180. [doi: 10.1161/CIRCEP.114.003030] [Medline: 26092577]
42. Hayes CJ, Dawson L, McCoy H, et al. Utilization of remote patient monitoring within the United States health care system: a scoping review. Telemedicine and e-Health 2023 Mar 1;29(3):384-394. [doi: 10.1089/tmj.2022.0111]
43. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008]
44. Angrist J, Pischke JS. Mostly Harmless Econometrics: An Empiricist's Companion: Princeton University Press; 2009. [doi: 10.1515/9781400829828]
45. Thomas EE, Taylor ML, Banbury A, et al. Factors influencing the effectiveness of remote patient monitoring interventions: a realist review. BMJ Open 2021 Aug 25;11(8):e051844. [doi: 10.1136/bmjopen-2021-051844] [Medline: 34433611]
46. Lee JY, Chan CKY, Chua SS, et al. Telemonitoring and team-based management of glycemic control on people with type 2 diabetes: a cluster-randomized controlled trial. J Gen Intern Med 2020 Jan;35(1):87-94. [doi: 10.1007/s11606-019-05316-9] [Medline: 31512187]
47. Eaton C, Vallejo N, McDonald X, et al. User engagement with mHealth interventions to promote treatment adherence and self-management in people with chronic health conditions: systematic review. J Med Internet Res 2024 Sep 24;26:e50508. [doi: 10.2196/50508] [Medline: 39316431]
48. Rogers EM. Diffusion of Innovation: The Free Press; 1995.
49. van den Bergh R, Evers LJW, de Vries NM, et al. Usability and utility of a remote monitoring system to support physiotherapy for people with Parkinson's disease. Front Neurol 2023;14. [doi: 10.3389/fneur.2023.1251395]
50. Boriani G, Maisano A, Bonini N, et al. Digital literacy as a potential barrier to implementation of cardiology tele-visits after COVID-19 pandemic: the INFO-COVID survey. J Geriatr Cardiol 2021 Sep 28;18(9):739-747. [doi: 10.11909/j.issn.1671-5411.2021.09.003] [Medline: 34659380]
51. Association AD. The Burden of Diabetes in Louisiana. 2023. URL: https://diabetes.org/sites/default/files/2023-09/ADV_2023_State_Fact_sheets_all_rev_Louisiana.pdf [accessed 2024-08-01]
52. Bazzano AN, Monnette AM, Wharton MK, et al. Older patients' preferences and views related to non-face-to-face diabetes chronic care management: a qualitative study from southeast Louisiana. Patient Prefer Adherence 2019;13:901-911. [doi: 10.2147/PPA.S201072] [Medline: 31213782]
53. Gillispie-Bell V, Bane S, Martin J, Kan P, Mujahid MS, Carmichael SL. Severe maternal morbidity in Louisiana by race, rurality, poverty, and availability of care. Public Health 2025 Sep;246:105824. [doi: 10.1016/j.puhe.2025.105824] [Medline: 40543228]

Abbreviations

CITSA: controlled interrupted time series
COPD: chronic obstructive pulmonary disease
CPT: common procedure terminology
ER: emergency room
HCRU: health care resource utilization
ICD: International Classification of Diseases
IPTW: Inverse Probability Treatment Weight
RPM: remote patient monitoring
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by S Brini; submitted 21.Aug.2025; peer-reviewed by M Chakit, Y Wu; revised version received 12.Nov.2025; accepted 13.Nov.2025; published 03.Mar.2026.

Please cite as:
Winberg D, Stoecker C, Li J, Shi L
Remote Patient Monitoring in Louisiana Medicare Beneficiaries With Diabetes or Hypertension: Retrospective Cohort Study of 2016-2020 Claims Data
J Med Internet Res 2026;28:e80782
URL: https://www.jmir.org/2026/1/e80782
doi:10.2196/80782

XSL•FO
RenderX

© Debra Winberg, Charles Stoecker, Jian Li, Lizheng Shi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 3.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

EPOCA Tele-Monitoring System for Older Adults at High Risk of Hospitalization: Budget Impact and Cost-Effectiveness Analysis

Henri Leleu[1], MD, MPH, PhD; Damien Testa[2], PhD; Mireille Dutech[2], MD; Etienne Minvielle[3,4], MD, PhD; Elise Cabanes[2], MD

[1]Public Health Expertise, 10 boulevard de Sébastopol, Paris, France

[2]Epoca U&I, Nanterre, France

[3]Institut Polytechnique de Paris, Centre de Recherche en Gestion de l'Ecole Polytechnique (I3-CRG, UMR CNRS 9217), Palaiseau, France

[4]Département Interdisciplinaire d'Organisation du Parcours Patient (DIOPP), Gustave Roussy Cancer Centre, Villejuif, France

Corresponding Author:
Henri Leleu, MD, MPH, PhD
Public Health Expertise, 10 boulevard de Sébastopol, Paris, France

Abstract

Background: France's aging population faces high rates of chronic illness, multimorbidity, and avoidable hospitalizations, placing pressure on an already strained health care system. Remote monitoring systems have shown promise in improving care coordination and reducing acute care use.

Objective: The objective of this study was to assess the cost-effectiveness of the EPOCA remote monitoring system, implemented within the Vigie-Age framework, compared to the standard of care for older adults with multiple chronic conditions.

Methods: Using data from the Vigie-Age Article 51 pilot study (with 722 participants, including 408 participants with long-term follow-up), a cost-utility analysis was conducted over a 10-year lifetime horizon. A Markov model with daily cycles simulated transitions across health states: at home, emergency department visits, hospitalization, and death. Analyses were conducted from both the French National Health Insurance (NHI) and collective perspectives. Direct medical costs, including hospital, outpatient, and intervention costs, were included. Health outcomes were measured in quality-adjusted life years (QALYs). Deterministic and probabilistic sensitivity analyses assessed model robustness.

Results: EPOCA was associated with a reduction in emergency department visits by 54% and in hospitalizations by 46%, cutting the average hospital stay from 55.6 (SD 51.7) to 30.6 (SD 27.8) days. Total costs per patient were €29,165 (EUR €1=US $1.13) with EPOCA and €39,929 for standard of care, representing a €10,764 saving from the collective perspective and a €7,430 saving from the NHI perspective. EPOCA yielded 0.04 additional QALYs and remained cost-saving even at higher program costs. Sensitivity analyses confirmed the robustness of the results. EPOCA had a 90% probability of being dominant and a 95% probability of being cost-effective at a €30,000 per QALY threshold.

Conclusions: On the basis of currently available evidence, EPOCA may be a cost-effective strategy for older patients at high risk of hospitalization. It could reduce health care use while improving outcomes, supporting its integration into national older adult care pathways and reimbursement by the French NHI.

(J Med Internet Res 2026;28:e80302) doi:10.2196/80302

KEYWORDS

budget impact analysis; cost-effectiveness; older adults; polypathology; patient remote monitoring

Introduction

France is undergoing a profound demographic transition marked by a growing older adult population. As of 2024, 10% of the population, equivalent to 7 million individuals, is aged ≥75 years, and this number is expected to rise to 16% by 2050, representing 12 million individuals [1,2].

Chronic diseases are highly prevalent among older adults in France, with 91% of individuals aged ≥75 years presenting at least 1 chronic condition or ongoing treatment. Multimorbidity is also a defining feature of aging: 47% of men and 36% of women aged ≥75 years live with 2 or more chronic diseases, and up to 26% of men and 17% of women in this age group have 3 or more chronic diseases [3]. Aging also significantly increases the risk of loss of independence. Although most older adults remain autonomous until late in life, the share of older adults receiving the French "Allocation Personnalisée d'Autonomie," a marker of dependency, rises from 2% among adults aged 70 to 74 years to approximately 50% among those aged 90 to 94 years and to three-quarters among those aged ≥95 years [4]. Finally, increasing age is accompanied by increased

XSL•FO
RenderX

risks of social isolation, particularly among women. More than half of women aged 85 years live alone, compared to just 25% of men [4].

These demographic shifts place immense pressure on France's health care infrastructure, which is already strained by shortages of geriatricians and a declining number of general practitioners (GPs). The result is a fragmented and reactive health care system, where emergency departments (EDs) and inpatient services often serve as the default response to complex geriatric needs. Approximately 40% of those aged ≥80 years in France are hospitalized annually, with an average duration of approximately 25 days per patient. It is estimated that more than 30% of these hospitalizations could have been avoided, with similar figures for ED visits [5,6]. These trends underline the urgent need for scalable, home-based care solutions tailored to the evolving needs of an aging, increasingly dependent population.

Remote monitoring systems (RMSs) have gained international traction as a promising solution for managing older patients with chronic and complex health needs. Existing RMSs vary widely in structure and technology but generally include combinations of wearable sensors, mobile health apps, and care coordination frameworks [7-14]. These RMSs are typically multidisciplinary, involving nurses, hospital physicians, geriatricians, and GPs, often with social support services integrated. Most of these RMSs have demonstrated reductions in hospitalizations and emergency visits [9-11,13,14].

Launched in 2019, the Vigie-Age framework is a multidisciplinary hospital-community initiative designed to provide comprehensive, coordinated care for older adults with multiple chronic conditions. Its primary objectives are to enhance quality of life and reduce avoidable hospitalizations through better continuity of care. This framework includes EPOCA, an RMS that is a human-centered connected telemedicine solution offering secure and continuous home support for complex older patients (with multiple comorbidities, loss of autonomy, and/or dependency). It is built around a 24×7 tele-monitoring platform that enables real-time clinical oversight, rapid response to health deterioration, and seamless coordination among health care professionals. The system integrates two core technological components: (1) connected devices, such as a wristband, for transmitting health data or emergency alerts, and (2) a centralized RMS coordination hub staffed by trained nurses and physicians, who engage directly with patients and caregivers when needed. Hospital geriatricians or GPs play a central role in the framework, offering ongoing medical supervision via tele-consultations and tele-expertise and maintaining active communication with community health care teams. In parallel, community-based nurses conduct regular home visits, monitor patient conditions, and deliver essential care services.

Vigie-Age has been included in the French "Article 51" experimental framework launched in 2018 to promote innovation in health care delivery and financing. Article 51 enables pilot studies of interventions that aim to improve care coordination, efficiency, and patient outcomes through new organizational models and payment structures. These pilots are evaluated for their clinical impact, economic sustainability, and potential for national scale-up, providing valuable data.

This study aimed to assess the cost-effectiveness of the Vigie-Age framework, including the EPOCA RMS, compared to standard of care (SOC), focusing on reductions in hospitalizations, emergency visits, and overall health care costs and using real-world data from the Article 51 evaluation of the Vigie-Age framework.

Methods

Data Source

This cost-effectiveness analysis is primarily based on data from the Vigie-Age Article 51 pilot study. Conducted prospectively between 2022 and 2024 across 4 hospitals in France, the study enrolled 722 participants. Eligibility criteria included individuals aged ≥70 years with multiple chronic conditions, a history of recent decompensations, and a high risk of further clinical deterioration.

The study population was composed of 36% (263/722) men, with a mean age of 87.9 (SD 7.1) years. Participants experienced an average of 2.0 (SD 1.4) overnight hospitalizations in the 6 months preceding study inclusion. Approximately 70% (499/722) of participants were recruited either during a hospitalization or from a community-based setting, while the remaining 30% (223/722) were included following an ED visit.

The pilot study assessed outcomes related to clinical efficacy (hospitalizations and ED visits), health care resource use, and associated costs. These outcomes were derived from French National Health Insurance (NHI) reimbursement data and compared between the study follow-up period and a similar period preceding inclusion. Among the 408 participants who received the EPOCA RMS intervention, outcome data were available for 269 (66%) individuals with sufficient follow-up, with a mean observation period of 7.5 months.

Perspective, Population, Intervention, and Comparator

This cost-effectiveness evaluation was conducted from 2 complementary perspectives, both including only direct medical costs. The first was a payer perspective, focusing on health care expenditures paid by the NHI. The second was a collective perspective, in line with the recommendations of the Haute Autorité de Santé [15], which includes all costs regardless of the payer.

The analysis was conducted on a population corresponding to participants enrolled in the Vigie-Age Article 51 study. The EPOCA RMS intervention was compared to SOC, defined as the same participants' preinclusion health care use and costs, measured over a comparable baseline window prior to program entry. The SOC for this high-risk, older, polymorbid population was defined as usual care provided by the French national health system, consisting of routine community and primary care services (GP follow-up, home-based care, and social support). Crucially, this usual care lacked the structured, proactive tele-monitoring and coordinated case management features that characterize the EPOCA intervention.

XSL•FO
RenderX

Model Structure

Due to the limited follow-up duration in the Vigie-Age Article 51 pilot study, a de novo Markov model was developed in Excel (version 16.108.3; Microsoft; available on demand) to extrapolate the long-term clinical and economic outcomes of the EPOCA intervention over a lifetime horizon, defined as 10 years based on the average age of the target population. The model used daily cycles and incorporated the following health states: stable condition at home, ED visits, hospitalization (either scheduled or via an emergency visit), and death (all-cause mortality).

The selected health states reflect key events from the Article 51 study: ED visits and hospitalization. Patients start in either the ED or at home; those at home may be hospitalized or visit the ED. In the ED, patients can return home or be hospitalized, while hospitalized patients can only return home. Death is possible at any stage. Hospitalization is associated with disutility, and costs are incurred at every stage, including outpatient care for patients at home.

Transition probabilities for hospitalizations and ED visits were derived from the Vigie-Age Article 51 pilot study. Model parameter estimations incorporated data from all patients included in the Vigie-Age Article 51 pilot study who had a minimum of 1 month of follow-up.

The model also accounted for the recruitment setting, with patients enrolled in the EPOCA intervention during an ED visit avoiding immediate hospitalization, whereas similar patients under SOC could proceed to inpatient admission. Model parameters are detailed in Table 1.

XSL•FO
RenderX

Table . Model parameters.

Parameters	Values, mean (SD)	95% CI in PSA[a]/DSA[b]	Distribution	Source
Transition probabilities				
SOC[c]				
ED[d] entry (%)	22	Not included	Not applicable	Art 51[e]
Monthly probability of hospitalization	1.9 (0.04)	1.8-2.0	Normal	Art 51
Monthly probability of ED visits	3.5 (0.05)	3.4-3.6	Normal	Art 51
Proportion of hospitalizations following ED visits (%)	71	64-79	Beta	Art 51
Length of stay (days)	7.7 (6.4)	6.2-9.2	Normal	Art 51
EPOCA				
ED entry (%)	0	Not included	Not applicable	Art 51
Monthly probability of hospitalization	1.3 (0.04)	1.2-1.4	Normal	Art 51
Monthly probability of ED visits	1.6 (0.03)	1.5-1.7	Normal	Art 51
Proportion of hospitalizations following ED visits (%)	69	55-83	Beta	Art 51
Length of stay (days)	7.3 (5.4)	5.8-8.8	Normal	Art 51
Age (years)	87.9 (3.6)	80.8-95.0	Normal	Art 51
Costs (€)				
VIGIE-AGE program	2210	Not included	Not applicable	Assumption
Hospitalization–EPOCA (per hospitalization)	3714 (3183)	832-9557	LogNorm[f]	Art 51
Hospitalization–SOC (per hospitalization)	3773 (2721)	1056-8873	LogNorm	Art 51
Hospitalization following ED visit	4249 (6373)	395-14,057	LogNorm	Art 51
ED visits	52 (5)	44-61	LogNorm	[16,17]
Outpatient care–EPOCA (per year)	5820 (8111)	4656-6984	Gamma	Art 51
Outpatient care–SOC (per year)	5337 (7759)	4270-6404	Gamma	Art 51
Utility				
General population aged ≥75 years	0.8	Not included	Not applicable	[18]
Hospitalization (per day)	–0.0013 (0.0003)	–0.0016 to –0.0010	Beta	[19]

[a]PSA: probabilistic sensitivity analysis.

[b]DSA: deterministic sensitivity analysis.

[c]SOC: standard of care.

[d]ED: emergency department.

[e]Art 51: Vigie-Age Article 51 pilot study.

[f]LogNorm: Lognormal distribution.

XSL•FO
RenderX

Key Assumptions

The EPOCA intervention was associated with a reduction in the frequency of ED visits and hospitalizations, based on the results of the Vigie-Age Article 51 study, which showed a significant reduction in the annual number of hospitalizations (4.4, SD 5.1 vs 2.4, SD 4.2) and emergency visits (3.4, SD 4.2 vs 1.6, SD 2.7) before and during the EPOCA intervention.

The probability of hospitalization after an ED visit and the average length of hospital stay were considered similar for both the EPOCA intervention and control groups based on a similar rate of hospitalization after ED visits (3.7, SD 3.3 vs 3.5, SD 2.8) and a similar average length of stay per hospitalization (7.3 days vs 7.4 days) before and during the EPOCA intervention.

Participants in the EPOCA arm were assumed to retain the intervention's benefits even after they exit the program based on similar results for hospitalization rates when comparing before and during the EPOCA intervention (4.4, SD 5.1 vs 2.4, SD 4.2; average follow-up of 7.5 months; n=269, 66%) and before and after the end of the EPOCA intervention (3.7, SD 5.1 vs 1.4, SD 2.8; average follow-up of 5.2 months; n=69, 17%). This assumption was also based on the expectation that the supportive care established during the intervention, along with the identification and management of potential risk situations and coordination among caregivers facilitated by EPOCA, would generally be sustained following the conclusion of the intervention. Sensitivity analyses were performed by waning the benefit of the EPOCA intervention after 0, 3, and 6 months after the end of the intervention.

It is specifically assumed that patients enrolled in the EPOCA intervention during an ED visit will avoid immediate hospitalization based on the results of the EPOCA intervention, showing that, for patients included in the ED, no immediate hospitalization was observed.

Costs and Utilities

The analysis included only direct medical costs, encompassing hospitalizations, ED visits, outpatient care, and the EPOCA intervention. Hospitalization and outpatient costs were derived from average expenditures observed in the Vigie-Age Article 51 pilot study and were stratified by study arm to reflect differences in the type of hospitalizations and intensity of outpatient follow-up.

Hospitalization costs were based on production costs (Echelle Nationale des Coûts), which were directly used for the collective perspective. For the payer (NHI) perspective, tariffs were estimated to be 1.27 times lower than production costs, based on observed differences between actual costs and reimbursements [20,21]. Similarly, outpatient care costs were based on tariffs. For the NHI perspective, it was assumed, according to reimbursement rules, that 65% of outpatient costs were reimbursed by the NHI, while the full amount was considered under the collective perspective.

Costs of ED visits were estimated using the average cost per visit, calculated from the total annual funding allocated to emergency services in France [16] and the national number of ED visits [17]. From the NHI perspective, costs were applied only to ED visits that did not result in hospitalization, as the latter were considered part of inpatient costs.

The cost of the EPOCA intervention was based on the reimbursement tariff used in the Vigie-Age Article 51 pilot study, with a mean cost of €2,210 per participant per year. A higher value of €4,441 was used in sensitivity analyses. This cost was applied at the beginning of each year to the proportion of patients remaining in the program. Program duration was assumed to be 6 months for 59% of participants and 12 months for the remaining 41%, resulting in an average duration of 8.4 months. This was implemented as a probability of continued participation, with an estimated 22% of participants still enrolled at the beginning of year 2.

Health utilities were sourced from published literature, using age-adjusted baseline utilities for the general population [18] and applying disutility weights for each day spent hospitalized [19].

Outcomes and Analysis

The analysis compared discounted (at 2.5% per year [15]) and undiscounted clinical outcomes, including the mean number of hospitalizations per patient, total hospital stay duration, and the mean number of ED visits leading to hospital admissions, between the EPOCA intervention and SOC. Economic outcomes (€2025; EUR €1=US $1.13) included mean hospitalization costs per patient, ED visit costs, outpatient care costs, and intervention costs. Quality of life outcomes were measured in quality-adjusted life years (QALYs).

A cost-utility analysis was conducted, with the primary outcome expressed as the incremental cost-effectiveness ratio, defined as the additional cost per QALY gained with EPOCA compared to SOC.

To evaluate the robustness of the model results, both deterministic and probabilistic sensitivity analyses (PSAs) were performed. Deterministic sensitivity analysis explored the impact of key parameters, including costs, transition probabilities, and utility values, by varying them within their 95% CIs or by +20% to −20% when CIs were unavailable. Results were presented using a tornado diagram to identify the most influential parameters.

PSA was conducted using Monte Carlo simulations to simultaneously assess uncertainty across multiple parameters. Probability distributions were assigned to cost inputs, transition probabilities, and utility values. Results are presented using cost-effectiveness acceptability curves. France does not apply a single official willingness-to-pay threshold; therefore, we used €30,000 per QALY as a commonly used reference value for interpretability of the PSA results.

Ethical Considerations

This cost-effectiveness and budget impact analysis is a secondary economic modeling study using aggregated, anonymized data from the French Vigie-Age Article 51 study (law 2017-1836) [22]. As no primary data were collected and no direct patient contact occurred, this analysis was exempted from further institutional ethical review. According to French regulations, this type of secondary analysis does not require

XSL•FO
RenderX

approval from a research ethics committee, as it falls outside the scope of research involving human participants governed by the French Public Health Code (Law No. 2012-300 of March 5, 2012, known as the "Loi Jardé") [23].

Results

The results of the cost-effectiveness and budget impact models, including projections on hospitalizations, ED visits, and direct and indirect costs, are presented in Table 2.

Table . Cost-effectiveness and budget impact results over a 10-year time horizon with a 2.5% discount rate.

Category	EPOCA, mean (SD)	Standard of care, mean (SD)	Difference (% reduction)
Resource use			
Emergency department visits (n)	2.8 (1.3)	6.1 (2.7)	−3.3 (54)
Hospitalization (n)	4.2 (1.9)	7.8 (3.5)	−3.6 (46)
Hospitalization days	30.6 (27.8)	55.6 (51.7)	−25.0 (45)
Undiscounted costs per patient (€)			
Payer (NHI)[a] perspective			
Program	2850 (214)	0 (0)	Not applicable
Emergency department visits	145 (69)	315 (148)	−170 (54)
Hospitalization	12,262 (12,758)	23,226 (24,204)	−10,964 (47)
Outpatient care	10,277 (18,030)	9424 (17,210)	853 (−9)
Total	25,535 (23,853)	32,965 (31,927)	−7430 (23)
Collective perspective			
Program	2850 (212)	0 (0)	Not applicable
Emergency department visits	465 (217)	1008 (469)	−543 (54)
Hospitalization	15,573 (15,055)	29,497 (28,562)	−13,924 (47)
Outpatient care	10,277 (17,383)	9424 (16,550)	853 (−9)
Total	29,165 (24,536)	39,929 (34,856)	−10,764 (27)
Cost-effectiveness			
Payer (NHI) perspective			
Discounted QALY[b]	1.30 (0.58)	1.26 (0.57)	0.04
Discounted total costs (€)	24,683 (22,465)	31,792 (30,033)	−7108
ICER[c]	Dominant	Dominated	Not applicable
Collective perspective			
Discounted QALY	1.30 (0.58)	1.26 (0.57)	0.04
Discounted total costs (€)	28,181 (23,171)	38,510 (33,021)	−10,329
ICER	Dominant	Dominated	Not applicable

[a]NHI: National Health Insurance.

[b]QALY: quality-adjusted life year.

[c]ICER: incremental cost-effectiveness ratio.

Over a lifetime horizon, accounting for background mortality, the model estimated an undiscounted average survival of 1.8 years per patient, with an average duration in the EPOCA program of 8.1 months. During this period, patients in the EPOCA arm experienced an average of 2.8 ED visits and 4.2 hospitalizations, compared to 6.1 ED visits and 7.8 hospitalizations in the SOC arm, representing reductions of 54% and 46%, respectively. The decrease in hospitalizations translated into a reduction in the total number of hospital days, from 55.6 days in the SOC arm to 30.6 days in the EPOCA arm.

From the collective perspective, the undiscounted total cost per patient was estimated at €29,165 in the EPOCA arm, compared to €39,929 in the SOC arm. This difference was driven primarily by reductions in hospitalization (€13,924 in savings) and ED visit costs (€543 in savings), partially offset by an increase of €853 in outpatient care costs. After incorporating the program cost (€2,850), the net savings amounted to €10,764 per patient. Similar cost savings were observed under the NHI perspective, with total savings of €7,430 per patient. These included €10,964 in reduced hospitalization costs, €170 in ED savings, and an €853 increase in outpatient costs, with program costs assumed

XSL•FO
RenderX

equal across perspectives. Importantly, EPOCA remained cost-saving even when applying the higher intervention cost of €4,441 used in sensitivity analyses.

In terms of health outcomes, the EPOCA arm was associated with 1.30 discounted QALYs compared to 1.26 QALYs in the SOC arm, a gain of 0.04 QALYs, primarily attributable to reduced hospitalizations and associated improvements in quality of life. Discounted total costs were €24,683 for EPOCA and €31,792 for SOC from the NHI perspective, yielding net savings of €7,108. Given that EPOCA resulted in both lower costs and improved outcomes, it was considered a dominant strategy compared to SOC. Similar conclusions were observed in the collective perspective, with €10,329 in total cost savings. EPOCA remained dominant at the higher intervention cost of €4,441 in both perspectives.

Sensitivity analyses considering no benefit or waning benefit of EPOCA after the end of the intervention did not change the conclusion, with EPOCA remaining a dominant strategy (Table 3). Deterministic sensitivity analysis (Figure 1) identified the average cost of hospitalization as the most influential parameter in the NHI perspective and the mean age at inclusion and the cost of hospitalization as the most influential parameters in the collective perspective. However, even under lower hospitalization cost scenarios, EPOCA remained cost-saving or cost-neutral, confirming the robustness of the model's conclusions across parameter ranges.

PSA (Figure 2) further supported these findings. From the collective perspective, EPOCA had a 90% probability of being the dominant strategy and a 95% probability of being cost-effective at a willingness-to-pay threshold of €30,000 per QALY. From the NHI perspective, these probabilities were 80% and 87%, respectively.

XSL•FO
RenderX

Table . Sensitivity analysis for cost-effectiveness and budget impact results including waning of the program effect, in the payer (National Health Insurance) perspective over a 10-year time horizon with a 2.5% discount rate.

Scenarios	EPOCA, mean (SD)	Standard of care, mean (SD)	Difference (% reduction)
100% waning at the end of the intervention			
Resource use			
Emergency department visits (n)	3.9 (1.3)	6.1 (2.6)	−2.2 (36)
Hospitalization (n)	5.3 (1.9)	7.8 (3.2)	−2.5 (33)
Hospitalization days	38.5 (28.9)	55.6 (43.2)	−17.3 (31)
Cost-effectiveness			
Discounted QALY[a]	1.28 (0.56)	1.26 (0.55)	0.02
Discounted total costs (€)	27,880 (21,539)	31,792 (28,240)	−3912
ICER[b]	Dominant	Dominated	Not applicable
100% waning after 3 months after the end of the intervention			
Resource use			
Emergency department visits (n)	3.5 (1.3)	6.1 (2.7)	−2.6 (42)
Hospitalization (n)	4.9 (1.9)	7.8 (3.4)	−2.9 (37)
Hospitalization days	35.9 (38.7)	55.6 (62.8)	−19.7 (35)
Cost-effectiveness			
Discounted QALY	1.29 (0.58)	1.26 (0.56)	0.03
Discounted total costs (€)	26,833 (22,354)	31,792 (29,975)	−4959
ICER	Dominant	Dominated	Not applicable
100% waning after 6 months after the end of the intervention			
Resource use			
Emergency department visits (n)	3.1 (1.2)	6.1 (2.6)	−3.0 (49)
Hospitalization (n)	4.5 (1.8)	7.8 (3.3)	−3.3 (42)
Hospitalization days	33.0 (31.5)	55.6 (54.9)	−22.6 (41)
Cost-effectiveness			
Discounted QALY	1.29 (0.56)	1.26 (0.55)	0.03
Discounted total costs (€)	25,637 (24,171)	31,792 (34,030)	−6155
ICER	Dominant	Dominated	Not applicable

[a]QALY: quality-adjusted life year.

[b]ICER: incremental cost-effectiveness ratio.

XSL•FO
RenderX

Figure 1. Deterministic sensitivity analysis results. Collective perspective (above) and payer (National Health Insurance) perspective (below). ED: emergency department; SOC: standard of care.

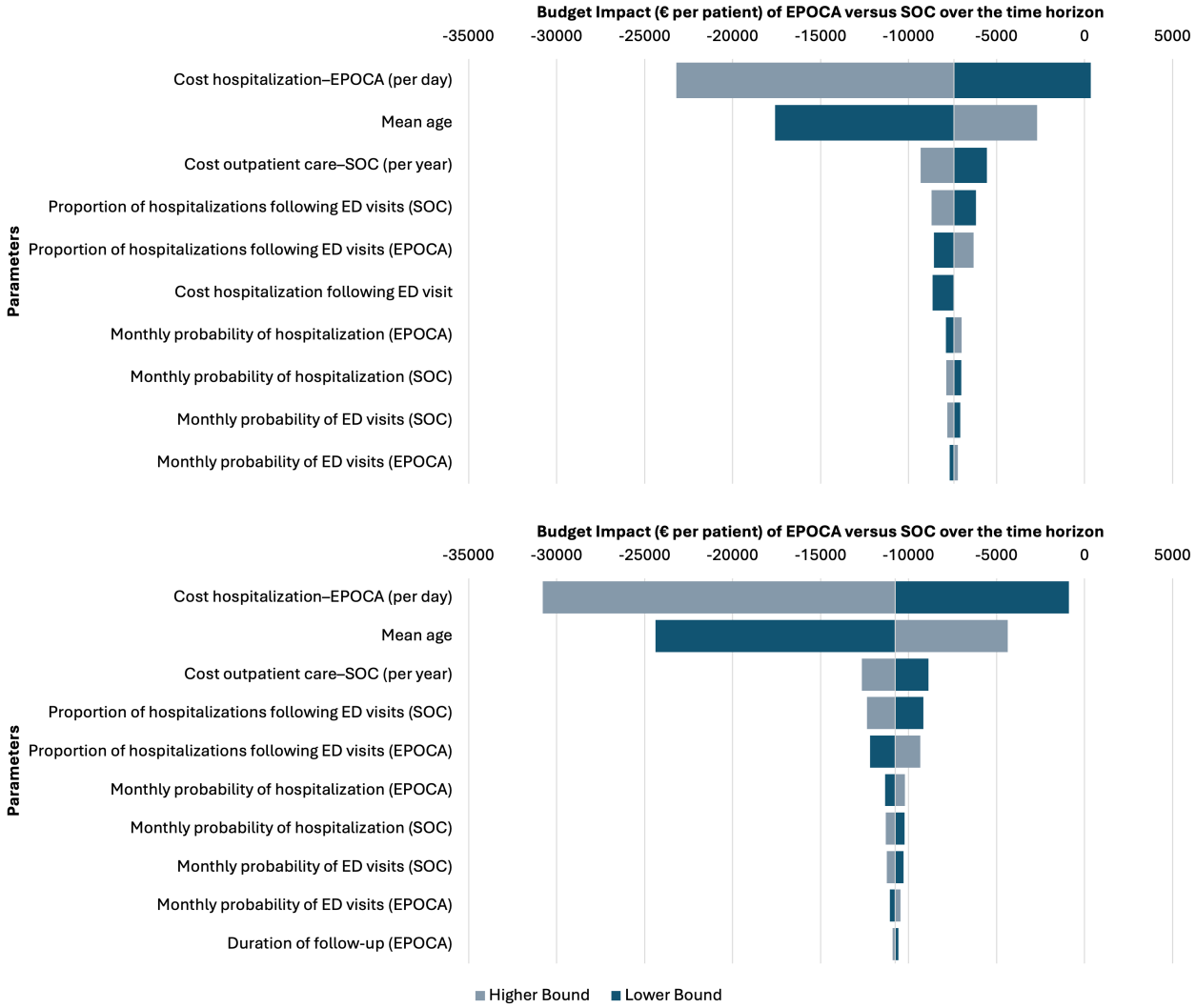

XSL•FO
RenderX

Figure 2. Probabilistic sensitivity analysis results. Collective perspective (above) and payer (National Health Insurance) perspective (below). QALY: quality-adjusted life year.

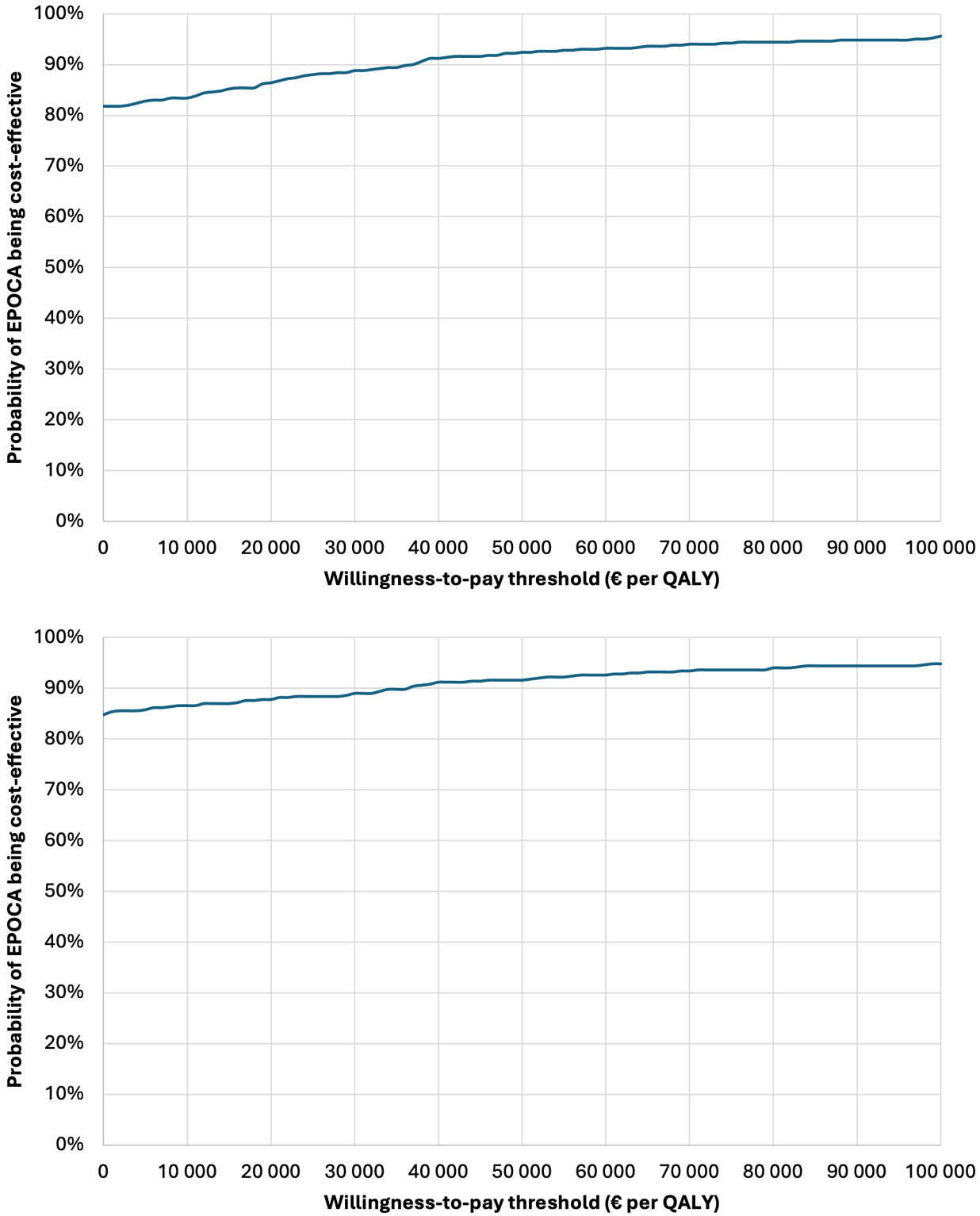

Discussion

This cost-effectiveness analysis shows that the EPOCA RMS could be associated with a reduction in avoidable hospitalizations and ED visits in frail, polypathological older adults. Over a modeled lifetime horizon, EPOCA was associated with a 54% reduction in ED visits and a 46% reduction in hospitalizations, resulting in a 25-day decrease in cumulative hospital stay per patient. These reductions were associated with cost savings of €10,764 per patient from the collective perspective and €7430 from the French payer (NHI) perspective. Notably, even under conservative assumptions for the program costs and in sensitivity analyses, EPOCA remained a dominant strategy, offering better outcomes at lower costs.

XSL•FO
RenderX

These results are consistent with the Article 51 pilot study and real-world data from retrospective observational studies of the EPOCA system [24]. In both hospital-led and GP-led implementations, EPOCA was associated with sharp declines in hospitalizations (−48% to −57%), ED visits (−48% to −62%), and hospital stay duration (−49% to −63%) within just 1 year of follow-up.

International evidence supports the observed outcomes of EPOCA. Similar RMSs have demonstrated reductions in hospital use, with relative risk reductions of up to 57%, 28% in unplanned hospitalizations, and 33% in ED visits [7,9-11,13,25]. EPOCA appears to achieve higher magnitudes of effect, possibly due to its integration into a comprehensive care coordination model involving both hospital and community actors, including nurses, GPs, and geriatricians. Additionally, unlike some RMSs, which may inadvertently increase ED use due to overalerting [10], EPOCA's design emphasizes triage, anticipatory care, and direct intervention, which could contribute to a true reduction in ED burden. In terms of cost-effectiveness, international evidence also supports the outcomes observed in this study. Previous evaluations of RMSs have reported per-patient savings ranging from approximately €1000 to €2800 [11,26,27], including program costs, which are on the lower end of the €7400 to €10,800 in net savings estimated in this analysis.

These findings have several important implications for health policy. First, EPOCA's suggested effectiveness in both clinical and economic dimensions supports its integration into national care pathways for older adults. Second, the program's potential net savings from the NHI perspective support the case for full reimbursement, particularly in light of demographic aging and increasing chronic disease prevalence. Moreover, EPOCA responds to systemic challenges faced by the French health care system: high hospitalization rates among older adults, underresourced primary care, and fragmented coordination between community-based and hospital-based care. By reducing hospital reliance and enabling safe, structured home care, EPOCA could contribute directly to system efficiency and sustainability.

Despite these promising findings, several limitations must be acknowledged. First, the base data from the Vigie-Age Article 51 pilot study involved limited follow-up, requiring extrapolation through Markov modeling. Although sensitivity analyses and waning assumptions confirmed the robustness of the results, longer-term and clinical trial data will be necessary to validate projections. Additionally, because the effectiveness inputs were based on a pre-post comparison without a control group, biases such as regression to the mean and time-varying confounding may have influenced the results. Although the observed differences are likely too large to be entirely due to bias, smaller differences could affect conclusions. Second, there was significant variability in available hospitalization costs, potentially impacting cost estimates. Third, outpatient cost data were partially incomplete, particularly for nonreimbursed services, which may have led to conservative cost assumptions. Fourth, the disutility values relied on data from heart failure patients [19], which may not be directly applicable to the population used in the model since no more closely matching source was found. Nevertheless, 73% of those in the Vigie-Age Article 51 study presented cardiovascular risk factors, and heart failure hospitalization occurred frequently. Finally, although the program likely reduces caregiver burden through improved coordination and reduced crisis care, these effects were not directly measured and warrant further investigation.

In conclusion, based on currently limited real-world evidence from the Vigie-Age study, this modeling study suggests that the EPOCA RMS may be a cost-effective intervention for managing older patients at high risk of hospitalization. It could not only reduce acute care use but also improve quality-adjusted life expectancy and lower total health care costs. From the NHI perspective, the budget impact may be neutral to favorable. Supported by real-world implementation data and consistent with international best practices, EPOCA has strong potential for national scale-up. Embedding such models within France's older adult care infrastructure could enhance care quality, reduce hospital pressure, and better address the needs of a rapidly aging population. These results will need to be confirmed with clinical trial results.

Funding

This study was funded by EPOCA U&I, but the sponsor had no role in the study design, data analysis, or data interpretation.

Data Availability

All data generated or analysed during this study are included in this published article. The de novo Markov model is available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization for this study was done by HL, DT, MD, EM and EC. Formal analysis was done by HL.
HL, DT contributed to methodology.
HL, DT, MD and EC were involved in writing original draft.
HL, DT, MD, EM, and EC were involved in review and editing.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Checklist 1
CHEERS 2022 checklist.
[DOCX File, 2486 KB - jmir_v28i1e80302_app1.docx]

References

1. Population par âge. INSEE. URL: https://www.insee.fr/fr/statistiques/4277619 [accessed 2025-03-25]
2. Bilan démographique 2024. INSEE. 2024. URL: https://www.insee.fr/fr/statistiques/8327319 [accessed 2025-03-25]
3. Bagein G, Costemalle V, Deroyon T, et al. L'état de santé de la population en France. : DREES; 2022 URL: https://drees.solidarites-sante.gouv.fr/sites/default/files/2022-09/DD102EMB.pdf [accessed 2026-04-27]
4. Personnes âgées dépendantes. INSEE. 2020. URL: https://www.insee.fr/fr/statistiques/4277754 [accessed 2025-03-25]
5. Naouri D, Ranchon G, Vuagnat A, et al. Factors associated with inappropriate use of emergency departments: findings from a cross-sectional national study in France. BMJ Qual Saf 2020 Jun;29(6):449-464. [doi: 10.1136/bmjqs-2019-009396] [Medline: 31666304]
6. Lanièce I, Couturier P, Dramé M, et al. Incidence and main factors associated with early unplanned hospital readmission among French medical inpatients aged 75 and over admitted through emergency units. Age Ageing 2008 Jul;37(4):416-422. [doi: 10.1093/ageing/afn093]
7. Wong AK, Wong FK, Chow KK, Wong SM, Bayuo J, Ho AK. Effect of a mobile health application with nurse support on quality of life among community-dwelling older adults in Hong Kong: a randomized clinical trial. JAMA Netw Open 2022 Nov 1;5(11):e2241137. [doi: 10.1001/jamanetworkopen.2022.41137] [Medline: 36350651]
8. Wong AK, Bayuo J, Wong FK, Chow KK, Wong SM, Lau AC. The synergistic effect of nurse proactive phone calls with an mHealth app program on sustaining app usage: 3-arm randomized controlled trial. J Med Internet Res 2023 May 1;25:e43678. [doi: 10.2196/43678] [Medline: 37126378]
9. Gayot C, Laubarie-Mouret C, Zarca K, et al. Effectiveness and cost-effectiveness of a telemedicine programme for preventing unplanned hospitalisations of older adults living in nursing homes: the GERONTACCESS cluster randomized clinical trial. BMC Geriatr 2022 Dec 22;22(1):991. [doi: 10.1186/s12877-022-03575-6] [Medline: 36550496]
10. Tchalla A, Marchesseau D, Cardinaud N, et al. Effectiveness of a home-based telesurveillance program in reducing hospital readmissions in older patients with chronic disease: the eCOBAHLT randomized controlled trial. J Telemed Telecare 2025 Feb;31(2):231-238. [doi: 10.1177/1357633X231174488] [Medline: 37221865]
11. de Batlle J, Massip M, Vargiu E, et al. Implementing mobile health-enabled integrated care for complex chronic patients: intervention effectiveness and cost-effectiveness study. JMIR Mhealth Uhealth 2021 Jan 14;9(1):e22135. [doi: 10.2196/22135] [Medline: 33443486]
12. Zulfiqar AA, Vaudelle O, Hajjam M, et al. Results of the "GER-e-TEC" experiment involving the use of an automated platform to detect the exacerbation of geriatric syndromes. J Clin Med 2020 Nov 26;9(12):3836. [doi: 10.3390/jcm9123836] [Medline: 33256080]
13. Martín-Lesende I, Orruño E, Bilbao A, et al. Impact of telemonitoring home care patients with heart failure or chronic lung disease from primary care on healthcare resource use (the TELBIL study randomised controlled trial). BMC Health Serv Res 2013 Mar 28;13:118. [doi: 10.1186/1472-6963-13-118] [Medline: 23537332]
14. Gellis ZD, Kenaley B, McGinty J, Bardelli E, Davitt J, Ten Have T. Outcomes of a telehealth intervention for homebound older adults with heart or chronic respiratory failure: a randomized controlled trial. Gerontologist 2012 Aug;52(4):541-552. [doi: 10.1093/geront/gnr134] [Medline: 22241810]
15. Choices in methods for economic evaluation. : HAS; 2020 URL: https://www.has-sante.fr/upload/docs/application/pdf/2020-11/methodological_guidance_2020_-choices_in_methods_for_economic_evaluation.pdf [accessed 2026-05-08]
16. Arrêté du 31 mars 2023 fixant pour l'année 2022 les dotations régionales mentionnées à l'article L. 174-1 du code de la sécurité sociale, les dotations régionales de financement des missions d'intérêt général et d'aide à la contractualisation, les dotations urgences, les dotations provisionnelles prévues par le décret du 29 septembre 2021 ainsi que le montant des transferts prévus à l'article L. 174-1-2 du même code. République Française. 2023. URL: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000047422651 [accessed 2026-04-27]
17. Statistique annuelle des établissements de santé diffusion. SAE Diffusion. URL: https://www.sae-diffusion.sante.gouv.fr/sae-diffusion/accueil.htm [accessed 2025-03-24]
18. Szende A, Janssen B, Cabases J, editors. Self-Reported Population Health: An International Perspective Based on EQ-5D [Internet]: Springer; 2014. [doi: 10.1007/978-94-007-7596-1]
19. Albuquerque de Almeida F, Al MJ, Koymans R, Riistama J, Pauws S, Severens JL. Impact of hospitalisation on health-related quality of life in patients with chronic heart failure. Health Qual Life Outcomes 2020 Aug 3;18(1):262. [doi: 10.1186/s12955-020-01508-8] [Medline: 32746842]
20. L'Etude nationale des coûts (ENC) MCO. ScanSanté. URL: https://www.scansante.fr/applications/enc-mco [accessed 2025-03-24]
21. Tarifs MCO et HAD. ATIH. URL: https://www.atih.sante.fr/tarifs-mco-et-had [accessed 2025-03-26]
22. Article 51 of the French Social Security Financing Act for 2018. Légifrance. 2017 Dec 30. URL: https://www.legifrance.gouv.fr/jorf/article_jo/JORFARTI000036339172 [accessed 2026-05-08]

XSL•FO
RenderX

23. Law No. 2012-300 of 5 March 2012 relating to research involving human subjects. Légifrance. URL: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000025441587 [accessed 2026-05-11]
24. Testa D, Iborra V, Dutech M, et al. Impact of a home-based remote patient monitoring system on hospitalizations and emergency department visits of older adults with polypathology: multicenter retrospective observational study. J Med Internet Res 2025;27:e64989. [doi: 10.2196/64989]
25. Orozco-Beltran D, Sánchez-Molla M, Sanchez JJ, Mira JJ, ValCrònic Research Group. Telemedicine in primary care for patients with chronic conditions: the ValCrònic quasi-experimental study. J Med Internet Res 2017 Dec 15;19(12):e400. [doi: 10.2196/jmir.7677] [Medline: 29246881]
26. De San Miguel K, Smith J, Lewin G. Telehealth remote monitoring for community-dwelling older adults with chronic obstructive pulmonary disease. Telemed J E Health 2013 Sep;19(9):652-657. [doi: 10.1089/tmj.2012.0244] [Medline: 23808885]
27. Kokkonen J, Mustonen P, Heikkilä E, et al. Effectiveness of telemonitoring in reducing hospitalization and associated costs for patients with heart failure in Finland: nonrandomized pre-post telemonitoring study. JMIR Mhealth Uhealth 2024 Feb 7;12:e51841. [doi: 10.2196/51841] [Medline: 38324366]

Abbreviations

ED: emergency department
GP: general practitioner
NHI: National Health Insurance
PSA: probabilistic sensitivity analysis
QALY: quality-adjusted life year
RMS: remote monitoring system
SOC: standard of care

Edited by M Balcarras; submitted 08.Jul.2025; peer-reviewed by C Martin, MR Angeles; revised version received 31.Mar.2026; accepted 17.Apr.2026; published 14.May.2026.

Please cite as:
Leleu H, Testa D, Dutech M, Minvielle E, Cabanes E
EPOCA Tele-Monitoring System for Older Adults at High Risk of Hospitalization: Budget Impact and Cost-Effectiveness Analysis
J Med Internet Res 2026;28:e80302
URL: https://www.jmir.org/2026/1/e80302
doi:10.2196/80302

© Henri Leleu, Damien Testa, Mireille Dutech, Etienne Minvielle, Elise Cabanes. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Extended Grammar of Systematized Nomenclature of Medicine – Clinical Terms for Semantic Representation of Clinical Data: Methodological Study

Christophe Gaudet-Blavignac[2,1*], PhD; Julien Ehrsam[2,1*], MD; Monika Baumann[1], BSc; Adel Bensahla[2,1], MSc; Mirjam Mattei[1], MSc; Yuanyuan Zheng[2,1], MSc; Christian Lovis[2,1], MPH, MD

[1]Division of medical information sciences, Diagnostic department, University Hospital of Geneva, Geneva, Switzerland

[2]Department of radiology and medical informatics, Faculty of medicine, University of Geneva, Geneva, Switzerland

[*]these authors contributed equally

Corresponding Author:
Christophe Gaudet-Blavignac, PhD
Division of medical information sciences
Diagnostic department
University Hospital of Geneva
Rue Gabrielle-Perret-Gentil 4
Geneva, 1205
Switzerland
Phone: 41 3790815
Email: christophe.gaudet-blavignac@hug.ch

Abstract

Background: Interoperability has been a challenge for half a century. Led by an informatics view of the world, the quest for interoperability has evolved from typing and categorizing data to building increasingly complex models. In parallel with the development of these models, the field of terminologies and ontologies emerged to refine granularity and introduce notions of hierarchy. Clinical data models and terminology systems vary in purpose, and their fixed categories shape and constrain representation, which inevitably leads to information loss.

Objective: Despite these efforts, semantic interoperability remains imperfect. Achieving it is essential for effective data reuse but requires more than rich terminologies and standardized models. This methodological study explores the extent to which the SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) compositional grammar can be leveraged and extended to approximate a formal descriptive grammar, allowing clinical reality to be expressed in coherent, meaningful sentences rather than preconstrained categories.

Methods: Building on a decade of semantic representation efforts at the Geneva University Hospitals, we developed a framework to identify recurring semantic gaps in clinical data. We addressed these gaps by systematically modifying the SNOMED CT Machine Read` Concept Model and extending its Augmented Backus-Naur Form syntax to support necessary grammatical structures and external vocabularies.

Results: This approach enabled the semantic representation of over 119,000 distinct data elements covering 13 billion instances. By extending the grammar, we successfully addressed critical limitations such as negation, scalar values, uncertainty, temporality, and the integration of external terminologies like Pango. The extensions proved essential for capturing complex clinical nuances that standard precoordinated concepts could not represent.

Conclusions: Rather than creating a new standard from scratch, extending the grammatical capabilities of SNOMED CT offers a viable pathway toward high-fidelity semantic representation. This work serves as a proof-of-concept that separating the rules of composition from vocabulary allows for a more flexible and robust description of clinical reality, provided that challenges regarding governance and machine readability are addressed.

(J Med Internet Res 2026;28:e80314) doi:10.2196/80314

KEYWORDS

data models; formal grammar; knowledge representation; semantic interoperability; terminology standards

XSL•FO
RenderX

Introduction

Status of Health Care Interoperability

Achieving semantic interoperability remains a central challenge in biomedical informatics and health data integration. While technical interoperability ensures that systems can exchange data in compatible formats, semantic interoperability guarantees that exchanged information is unambiguously interpretable across heterogeneous systems and contexts. This objective requires the explicit representation of meaning through formal models, ontologies, and standardized vocabularies, enabling computational reasoning and automated integration of clinical data.

Significant efforts have been invested in developing tools and frameworks to bring semantics to the complex field of health data. In Switzerland, the Swiss Personalized Health Network and its 3-pillar strategy defined a strong semantic representation of data as the first and mandatory pillar [1]. At the European level, the European Health Data Space initiative has further emphasized the importance of standardized, interoperable health data to enable secure sharing and secondary use across member states [2]. Similarly, in the United States, the Office of the National Coordinator for Health IT promotes this goal through the US Core Data for Interoperability, a standardized set of data elements required for nationwide health information exchange [3]. For years, the field of semantic interoperability has been supported by clinical data models, considered technical standards, and classifications and terminologies, called semantic standards.

The multiplication of semantic standards has prompted the need for ontology alignment, or ontology matching, to establish semantic correspondences between heterogeneous terminologies and domain ontologies. Classical systems such as AgreementMaker leverage a combination of lexical similarity, structural consistency, and description logic reasoning to ensure coherent mappings between ontologies [4]. More recent approaches, such as BERTMap, apply contextual language models to improve recall and precision in complex clinical terminologies [5].

Compositional grammar (CG) models provide a structured means of representing the internal semantics of clinical expressions. Clinical concepts are inherently compositional and thus require formalisms that can represent their components and relationships. Frameworks such as the SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) CG define syntactic and semantic rules for postcoordinated expressions (PCEs), ensuring internal consistency and machine interpretability [6].

Semantic interoperability also relies on formal logic–based models, typically expressed in Description Logics (DL), which underpin the Web Ontology Language. These models enable key reasoning tasks such as classification, consistency checking, and inferencing [7]. Building on this foundation, knowledge graphs have emerged as a flexible paradigm for integrating heterogeneous biomedical data, representing entities and their relationships within a unified semantic space.

Several international standardization bodies have developed complementary frameworks to support semantic interoperability. Health Level Seven Fast Healthcare Interoperability Resources (FHIR) provides a resource-oriented model for data exchange, including semantic bindings to standard terminologies such as SNOMED CT, Logical Observation Identifiers Names and Codes (LOINC), and Unified Code for Units of Measure, and offers Resource Description Framework and Web Ontology Language representations for semantic web integration [8]. The Observational Medical Outcomes Partnership (OMOP) Common Data Model, developed within the Observational Health Data Sciences and Informatics initiative, harmonizes observational health data using standardized vocabularies and facilitates analytical interoperability [9]. Cross-domain efforts, including ISO (International Organization for Standardization) 23903 and the World Wide Web Consortium's Health Care and Life Sciences group, promote semantic harmonization through formal metadata and linked data principles [10].

Each of these models has its strengths and weaknesses, and no system fits all purposes [11-13]. Creating a data model ultimately involves defining a finite set of categories, making choices that inherently shape what can and cannot be represented. These reflect the intended purpose of the model, but fitting data into these finite sets inevitably modifies it, making reuse based on model-specific categories less granular [14]. Semantic standards similarly have specific purposes, and granularity can vary dramatically depending on the focus of the classification. The *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) can be useful in representing disease, but cannot represent surgical interventions. Representing data using only single codes from a classification will inherently result in information loss that will be tied to the focus and granularity of the terminology. Figure 1 depicts this loss using a commonly known situation and its representation in natural language, then in *ICD-10* codes.

XSL•FO
RenderX

Figure 1. Information loss when using single standards to represent reality.

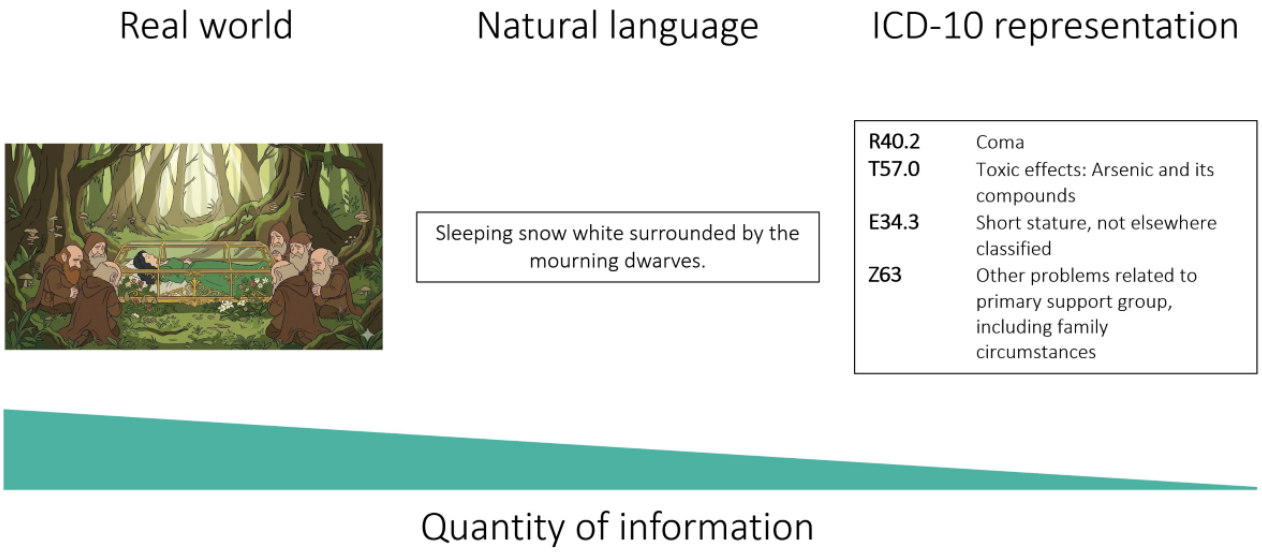

SNOMED CT and Natural Language

In this landscape, the CG of SNOMED CT, combined with its broad coverage, makes it a strong candidate for semantic interoperability [15]. With its knowledge graph–like structure, it allows intelligent code retrieval through its expression constraint language and gives users the ability to create PCEs to represent absent concepts. This relies on 2 elements. The CG, an Augmented Backus-Naur Form (ABNF) syntax, is used to create composed expressions that can be parsed and evaluated automatically [16]. The Machine Readable Concept Model (MRCM) defines which concepts (domain) can be refined with which relation (attribute) to which values (range). Combining these resources with the editorial guide, which defines attributes-naming conventions, ensures that new expressions comply with SNOMED CT's rules and grammar [17].

However, while designed for postcoordination and full expressivity of clinical concepts, it is rarely used to its full capacity and is intended for SNOMED CT concepts only [18,19]. It lacks granularity in some fields, such as strains of SARS-CoV-2, while others, such as genomic sequences, are not covered at all. Using SNOMED CT to represent clinical structured data within its expected parameters is eventually restrictive and insufficient to cover everything. Each attribute has a very specific range, domain, definitions, and permitted use cases, known as domain-specific modeling, and extending it eventually becomes necessary [20].

Like natural languages, SNOMED CT can be described as consisting of 3 fundamental components: a vocabulary of words, a grammar defining the rules for their combination, and the resulting sentences. Natural languages, however, allow the creation of any sentence respecting their grammar, regardless of their meaning, and regularly integrate words from other languages. A sentence such as "The 'entrepreneur' dropped his children off at 'kindergarten' before going to a 'karaoke'" uses words from French, German, and Japanese, which have been fully integrated into the English language. It is possible to draw inspiration from this structure to form a similar, computer-processable language, and applying this philosophy to SNOMED CT could improve its usability to fully represent clinical data semantically. The field of medical informatics already possesses rich vocabularies such as SNOMED CT, which serve as the source of machine-readable "words," each with its own specific strengths and weaknesses. They must then be combined according to well-defined grammar rules to form expressions, for which the SNOMED CT CG provides a comprehensive base.

Beyond SNOMED CT

Despite long-standing efforts to construct a comprehensive medical terminology, a single complete solution remains elusive, often resulting in the use of fragmented assemblies of various systems [21]. What remains necessary is a unifying set of rules to govern composition.

The solution should be formal, as semantic representation requires clear rules for machine readability that define how building blocks assemble into structured expressions. It must be descriptive rather than prescriptive. While grammars guide expression, they should not restrict what can be said a priori but rather ensure reality can be described clearly and consistently. Importantly, the solution must be dynamic, evolving through continuous iteration. Clinical reality is complex and ambiguous, with new findings emerging faster than they can be formally incorporated. Enabling the construction of meaningful expressions even without predefined concepts allows rapid adaptation to this complexity. To the best of our knowledge, such a formal descriptive grammar (FDG) does not exist yet.

This study proposes a first step in this direction by leveraging SNOMED CT to represent clinical data while ensuring machine readability. We summarize requirements gathered through 10 years of manual semantic representation and describe an implementation based on extending SNOMED CT's CG, syntax, and MRCM. This approach aims to represent clinical data regardless of source, output terminology, or intended use case.

XSL•FO
RenderX

While the results presented reflect the current state, they remain open-ended and extensible.

Methods

Overview

This work was developed in parallel to the semantic representation of structured data in the Geneva University Hospitals (HUG) data lake over the past decade. It builds upon the methodological framework introduced in *Semantics in Action: a Guide for Representing Clinical Data Elements* [22] and *Scalar Values in SNOMED CT: a Proposed Extension* [23], which proposes a structured and iterative approach for semantically representing clinical data. Counts of represented metadata elements are derived from the Clinical Data Warehouse without individual data extraction.

Framework

The guide's competency-building framework and rule-creation cycle were followed, incorporating both manual expert encoding and consensus-based refinement of representation rules through regular focus groups to define the FDG. The method used was gradually complexified in focus groups as more rules became necessary.

The teaching framework focuses on the use of SNOMED CT and follows a 3-part approach. First, SNOMED International's (SI) introductory training courses and documentation are taken to familiarize newcomers with the basics of standard-based semantic representation [24]. This allows new team members to start work on a simpler 1-dimensional representation, such as a problem list value set [25]. Next, internal documentation and continued practice give trainees a deeper understanding of real data and local specificities, allowing them to represent more complex data. Finally, more advanced courses, such as the SNOMED CT authoring courses, are taken by long-term core team members [26]. These, along with the practical experience gained by taking part in the training, give the ability to supervise newcomers and to participate in the rule creation cycle described below. New team members are always supervised closely by experienced team members and progressively increase the complexity of their tasks as required. Following this framework ensures that team members participating in the discussions around the rules of representation have a common training and vision.

Over time, complex situations encountered during the process made it clear that SI's rules were too restrictive to allow for complete representation. Initially, only a few internal unwritten rules were agreed upon, but it quickly became evident that, to avoid losing track of erratically evolving rules and changes, a framework was necessary to harmonize the process and maintain coherence. The rule-creation cycle was developed during weekly focus groups, which bring together the team members who participate in the semantic representation effort. They are made up of a core of 4 team members with heterogeneous professional backgrounds, including health care professionals (medical, nursing, midwifery), more technical training (IT, bioinformatics), and administrative knowledge (billing). These can then be joined by others depending on the team's setup at that time, such as students, new team members, or other colleagues with an occasional interest in a specific topic. The core members all have proficient experience with semantic representation and specifically with SNOMED CT, and are certified by SI through courses such as the authoring certification. The core of the team represents a small, soft-funded group, and the rest of the group has a high turnover rate. Added to this are limited time and resources, which means that there is a strong need to prioritize how and where said resources are applied. Data collections with high clinical value and a high number of instances, such as laboratory procedures and patient formularies, were tackled first. For the same reasons, work is parallelized as team members are each assigned separate tasks. For these reasons, no formal interannotator agreement studies are carried out.

Instead, specific topics are discussed during focus groups, and only when outcomes are unanimously agreed upon are they validated. Focus group topics consist of questions or complex situations each member has identified, and the rule-creation cycle is applied to each. First, existing guidelines and approaches are thoroughly reviewed and applied if they can solve the issue. If no existing solution is deemed satisfactory, an extension to the current rules is necessary, and is put into practice once consensus is reached through rounds of discussions. It is evaluated both for coherence with regard to previous rules and for clarity and semantic accuracy. Satisfactory outcomes are added to the list of grammar rules, which are revisited until deemed sufficiently tested. As this method applies only to situations for which the current set of rules available through SI or other groups is deemed insufficient or not applicable, they were not formally compared to each other on the same cases.

Reaching SNOMED CT's Limits

SNOMED CT was selected as the foundational semantic standard for this representation effort due to its comprehensive set of concepts, its adherence to a formal ABNF syntax, and its robust compositional rules that facilitate precise concept modeling. We prioritize the use of single SNOMED CT concepts first whenever they are sufficient to faithfully represent the data's content. When single concepts fall short, we use postcoordination, leveraging the constraints defined by the MRCM to its full capacity, constructing complex clinical expressions.

If neither approach can adequately capture the semantic content, extending it becomes inevitable as a representation gap is formally identified. This is where SNOMED CT's inherent extensibility is crucial for addressing domain-specific requirements. To resolve such gaps, we systematically explore possibilities for modifying the underlying grammar. The primary strategy involves altering existing MRCM rules, specifically through domain or range extensions, as referenced in [27]. Should this be insufficient, we analyze the need to include new SNOMED CT attributes within the representation framework. Finally, if essential semantic concepts are entirely absent from SNOMED CT, an external terminology is integrated, which consequently requires corresponding modifications to the ABNF syntax governing the semantic expressions.

XSL•FO
RenderX

The results detail the principal extensions implemented, which enable semantic coverage for a large percentage of the structured data within the HUG's data lake. The modified SNOMED CT ABNF syntax and MRCM collectively form the contribution of this paper and are available in Multimedia Appendices 1 and 2.

Ethical Considerations

Access to the HUG database is granted on an individual basis by the institution's Chief Data Officer and Medical Director and is reviewed and renewed every 6 months. This authorization permits access for the purpose of viewing and extracting aggregated metadata. No individual-level patient data are ever extracted, processed, or analyzed. All results presented in the manuscript consist exclusively of irreversibly aggregated counts derived from metadata and do not allow identification or reidentification of individuals. Consequently, this work does not constitute research involving human beings or health-related personal data within the meaning of the Swiss Human Research Act (SR 810.30) and falls outside its scope [28]. In accordance with the Swiss Human Research Act and the Swiss Federal Act on Data Protection, neither ethics committee approval nor patient consent was required; therefore, no waiver was sought from the cantonal or hospital ethics committee [29].

Results

Overview

Several problematic situations were encountered, which required progressively drifting away from SNOMED CT's official rules. These are common properties or patterns that occur across all or most natural languages, known as linguistic universals, a concept derived from Chomsky's universal grammar [30]. They are encountered often while representing data, and most can be addressed only partly by applying SNOMED CT's rules, grammar, and concepts. These gaps in the capacity to represent reality include negation, numeration, or scalar values, and displacement, which includes uncertainty and temporality. Most are included in some way in SNOMED CT's design, although rarely implemented in practice, and are therefore solvable by expanding the CG. This is done with domain and range extensions, and the addition of new attribute rules [31], which allow better representation.

Other situations, however, are not intended by SNOMED CT and necessitate breaking the mold to make the CG more inclusive. This is done by integrating other terminologies to fill gaps within SNOMED CT that cannot be handled even with postcoordination, such as SARS-CoV-2 strains. The different situations encountered and solutions devised are detailed below with an example for each. The complete list of MRCM modifications is available in Multimedia Appendix 1, and the modifications done for each property are detailed in Textbox 1.

Textbox 1. Modifications done for each property.

Negation

- 45169001 |Without (attribute)|
- 5185003 |Except for (attribute)|
- 408729009 |Finding context (attribute)|
- 408730004 |Procedure context (attribute)|

Scalars

- 103373006 |With size (attribute)|
- 410671006 |Date (attribute)|
- 79409006 |Resulting in (attribute)|
- 246205007 |Quantity (attribute)|
- 246262008 |Score (attribute)|

Uncertainty

- 408729009 |Finding context (attribute)|
- 408730004 |Procedure context (attribute)|

Temporality

- 255234002 |After (attribute)|
- 288556008 |Before (attribute)|
- 371881003 |During (attribute)|
- 103335007 |Duration (attribute)|

XSL•FO
RenderX

Scalar Values

Scalar values are physical quantities that can be described by a single pure number accompanied by a unit of measurement [32]. Scalars are one of the first domains encountered that were not present in sufficient detail in SNOMED CT. Scalars appear in myriad places in clinical data, whether as laboratory results, scores, sizes, or measurements, and hold great importance for accurate representation of clinical reality [8,18]. In SNOMED CT, integers were previously included as concepts, which have since been removed and replaced by Unicode text expressions preceded by a hash [33]. There remain concepts that contain numerical values in labels, such as quantitative result cutoffs, but not in any attribute relationship, as they are not fully defined [34].

A better representation of scalar values is achieved with new attributes, using existing accepted standards, when possible, such as the ISO 8601 format for dates, which is largely used, including by SNOMED CT [35] (Figure 2). The advantages of representing scalars directly in formal grammar expressions are vast. This greatly improves querying capacity, using queries with numeric operators, such as "all heart rates with a value of >150 bpm."

Figure 2. Scalar value representation example.

Negation

Negation is the act of denying or contradicting something, or the expression of its absence or opposite. The ability to represent negation is a key part of any language. Without it, this sentence could not be written. It does, however, pose many challenges when attempting to include it in a knowledge representation, particularly with regard to inference in hierarchical structures [36]. SNOMED CT's approach to negation through its context model exists outside the DL, as the absence of something is still linked to the focus concept by an "Associated" attribute [37], which is logically incoherent. Other ways of representing negation in SNOMED CT can also lead to errors of inference and classification [36], with children concepts being less restrictive than their parents. In fact, SI goes as far as to recommend "handling negation outside of SNOMED CT ... rather than try and represent it within the terminology" [38].

However, SNOMED CT abounds with concepts containing negation in some form or another. The limitations described above apply to the approved attribute relations that are used in fully defined concepts. For the cases not already covered, plenty of concepts exist that represent negation in one way or another, such as unapproved attributes and qualifiers, which are used to extend the MRCM when necessary. Extending attribute domains and using new attributes such as "Except for" and "Without" has proven highly useful. For hernia repairs, for example, SNOMED CT only specifies when the procedure is done with a mesh. However, no concepts describe the absence of mesh during a procedure, which was needed (Figure 3). These modifications allow us to cover many instances of negation encountered and are currently deemed sufficient.

Figure 3. Negation representation example.

Uncertainty

Uncertainty describes a situation in which something is not known or not certain. This has an important place in representing medical data, as many situations in clinical care contain uncertainty. It is inherent in every patient encounter. It starts with the patient's history, such as not knowing when a symptom began, and continues through the diagnostic process, from forming a differential diagnosis to choosing which tests to order.

Natively, SNOMED CT represents uncertainty in the same way as negation, with the "Finding context" attribute with values such as "Known possible." As with negation, however, SI states that "attempts to capture probabilistic or uncertain knowledge are out of the scope of SNOMED CT" [39], despite containing approved attribute relations for describing such situations. It is also present in primitive concept labels such as "Uncertain diagnosis."

Similarly to negation, the extensions made to cover uncertainty are expanding the domain of the "Finding context" and "Procedure context" attributes to include "<<Clinical finding" and "<<Procedure," respectively. This allows for representing Findings and Procedures with a refinement concerning uncertainty (Figure 4).

XSL•FO
RenderX

Figure 4. Uncertainty representation example.

Temporality

Temporality is the state of existing within or having some relationship with time. It is a vital property to consider when analyzing patient data. Indeed, knowing if a certain procedure, laboratory test, or diagnosis happened before, during, or after another is critical and has a defining influence on how information is interpreted. This also includes notions such as duration of processes or procedures, evolution over time, and chronicity.

SNOMED CT already has an extensive coverage of temporality and has proven capable of adequately representing many cases encountered in practice, such as "Temporally related to" (with its descendants After, During, and Before). These apply to various domains such as Situations, Clinical findings, and Observables entities, and therefore already cover many situations. However, despite their extensive coverage, gaps remain in the representation of temporality.

Extensions include extending the range of the Before and During attributes to match that of After. Procedures are added to these attributes' domains. Some larger gaps include dates, which are resolved using the "Date" attribute, and duration of procedures, with the "Duration" attribute (Figure 5), which both use scalar values as necessary.

Figure 5. Temporality representation example.

External Vocabularies Integration

Overview

While the situations mentioned up to now are certainly a welcome improvement in terms of scope of representation, they describe modifications compliant with SNOMED CT extension capabilities. Other types of encountered gaps, which concern specific group of concepts absent from SNOMED CT, are not resolvable in this way. Representing these concepts, therefore, necessitates the inclusion of external terminologies. While using different standards together can hinder interoperability, as it can limit the possible exchange and secondary usage of data, it has become unavoidable. Integrating various formats and standards to make them compatible has been identified as a challenge in the field of medical informatics [40]. In fact, SNOMED CT itself was born from the fusion of SNOMED RT and Clinical Terms V3, so what is proposed here is only an extension of its original philosophy [41]. However, it should be remembered that multiple maps exist between SNOMED CT and terminologies such as ICD-10 and Orphanet [42], as well as the effort to harmonize LOINC and SNOMED CT [43,44]. These should be consulted first to ensure that the desired term does not already have a SNOMED CT equivalent. Ontology alignment efforts and thesauri such as Unified Medical Language System should also be consulted to verify equivalences.

Once this verification is done, external vocabularies can be evaluated for integration. How they are evaluated is subject to rules. We have defined five selection criteria: (1) the terminology must first cover an identified gap in SNOMED CT coverage; (2) it must be concept-based, so cannot be a thesaurus; (3) it must be hierarchical and contain at least some graph-like properties; (4) it must have versioning capabilities; and (5) it must be a well-recognized standard widely accepted by relevant users. Once chosen, the steps described below should be followed to fully integrate them, increasing the scope of representation. This includes adding them virtually to a SNOMED CT hierarchy by attributing a parent, verifying MRCM rules, and modifying the ABNF syntax.

Content Analysis and Proximal Primitive Parent Attribution

To include an external terminology, it is necessary to clearly identify when and where it would be used. The content of the identified terminology should be analyzed in parallel with SNOMED CT to define where it could fit in the SNOMED CT hierarchies. Ideally, an entire terminology should be linked to one parent concept, but if different parts of the terminology fit in different hierarchies, this should be documented, and clusters of codes should be defined if needed. The defined clusters are then attributed to a proximal primitive parent. This task is described in the SNOMED CT authoring course for concept creation and consists of defining the closest parent of the cluster of codes in SNOMED CT that is not fully defined [45]. To do so, the closest parent is assigned, and its genealogy is reviewed to find the first primitive concept (Figure 6 [46]).

XSL•FO
RenderX

Figure 6. Proximal primitive parent attribution.

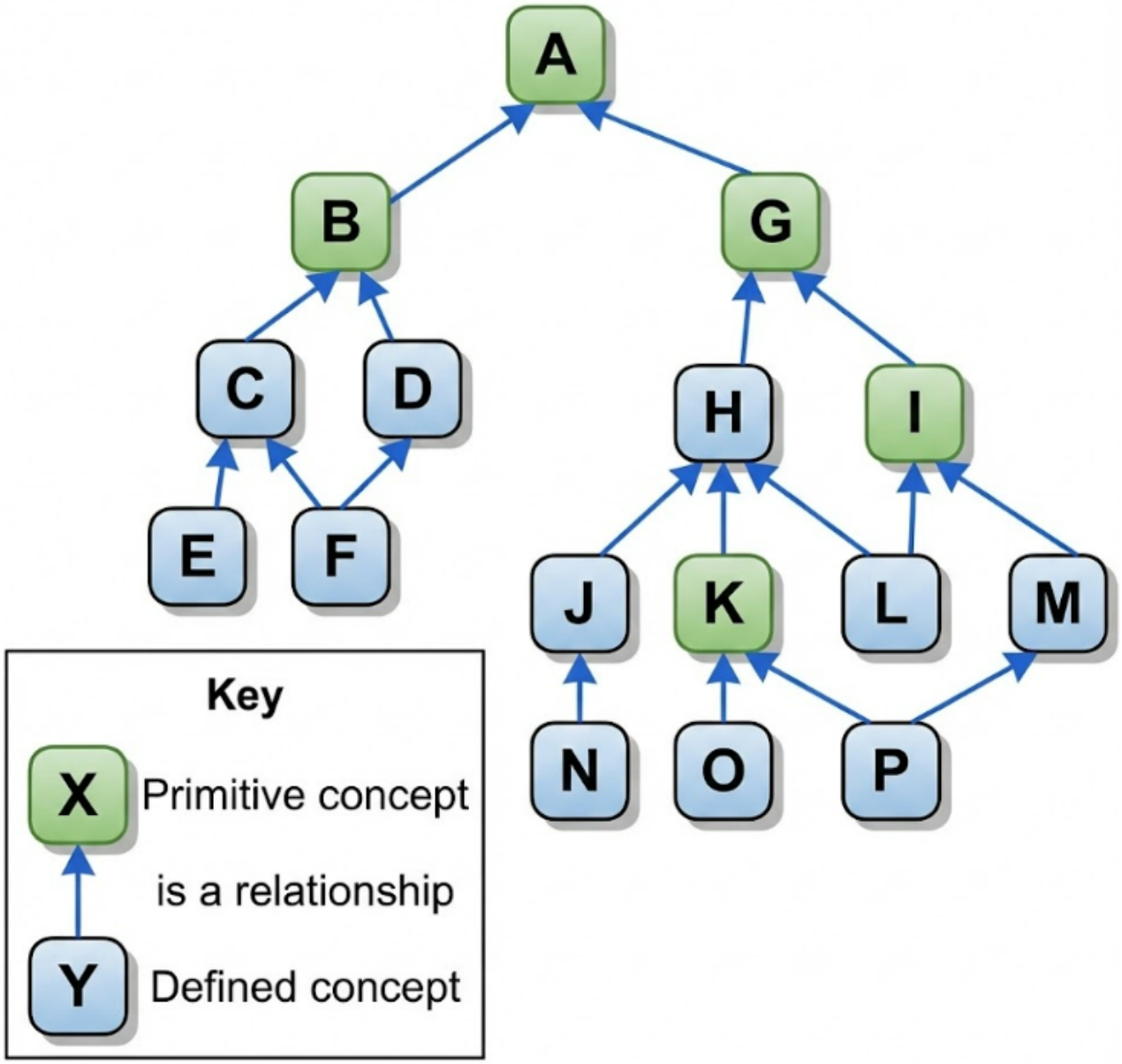

Modifying the MRCM and ABNF

Once the clusters of codes are attributed to a parent, MRCM rules covering those parents must be reviewed to ensure they are still valid. If necessary, they can then be updated, or new rules created, to include the cluster in the range or domain, if they fit an approved definition. This helps in keeping coherence when creating new PCEs, but must be done carefully, as integrations that happen too high in the hierarchy can influence multiple existing MRCM rules.

A syntax modification is then necessary to allow for integration of external concepts into SNOMED CT PCEs. The ABNF syntax is therefore modified to signal when another terminology is used and to allow validation of PCEs. The tilde symbol (~) is used to signal the start and end of the use of external concepts. Another ~ is used to separate the identifier of the vocabulary and the concept used. These new terminology concepts can, in theory, be inserted into any part of an expression, whether as the focus concept (~Terminology~id~:SCT=SCT), attribute (SCT:~Terminology~id~=SCT), value (SCT:SCT=~Terminology~id~), a combination of these, or even used alone. This modification has been validated using a modified ABNF parser on newly created expressions. The modified ABNF syntax is available in Multimedia Appendix 2.

Implementation

When the COVID-19 pandemic hit, genomic sequencing quickly led to a characterization of the SARS-CoV-2 virus in specific strains, with scientific and vernacular names given by organizations such as Nextstrain [47], Pango [48], and the World Health Organization [49]. These strains took on a crucial importance with the advent of more contagious or less lethal variants such as Delta or Omicron [50]. This close focus on specific variants is well represented in classifications such as the Pango nomenclature [51]. However, in SNOMED CT, the representation of those strains is lacking, with a unique code to

XSL•FO
RenderX

characterize the SARS-CoV-2 virus. This situation is used as a use case to demonstrate the inclusion of an external vocabulary through the proposed method. In this scenario, Pango is chosen to refine the organism hierarchy for SARS-CoV-2 lineages. Therefore, a new attribute group with a domain, range, and definition must be validated. Since SNOMED CT already has a category of codes that defines strains but does not include SARS-CoV-2, the Pango nomenclature is added as children of this code. A new MRCM rule is then defined to refine the Organism hierarchy with the "Microbiological strain" attribute, as shown in Figure 7.

Figure 7. External terminology representation example.

Using this method, a total of 5 terminologies other than SNOMED CT are added to expand the scope of representation. These are Pango, Nextstrain, Sequence Ontology [52], RadLex [53], and the Common Terminology Criteria for Adverse Events (CTCAE) [54].

Discussion

Principal Results

Overview

This work addresses the difficult balance between expressivity and standardization with a first step in the direction of an FDG for semantic representation of clinical data, using SNOMED CT as the base. It provides a formal, extensible set of rules for constructing meaning that can draw from diverse vocabularies as needed. This modular approach means new terminologies can be integrated as they become relevant without limiting what can be said, but providing a clear structure for how to say it.

After 10 years of representing structured data in the HUG datalake, the FDG now provides semantic representations for 119,941 distinct data elements using 39,168 unique expressions. These expressions cover over 13 billion data instances from 1.7 million patients. Notably, 18,370 of these unique expressions are PCEs, validating the approach as nearly half of all unique expressions created are complex PCEs. This clearly demonstrates that precoordinated concepts alone are profoundly insufficient to capture clinical reality at scale. Moreover, 4500 PCEs use the extensions presented in this paper, for a total of nearly 500 distinct extended PCEs, underscoring their usefulness.

This approach differs from existing interoperability efforts. While FHIR provides a flexible structural framework to define custom data models (FHIR profiles), its focus is not on prescribing semantic content, leaving users to define their own semantic frameworks. In comparison to FHIR extensions, the approach is similar, but we connect the external terminology to SNOMED CT to maintain semantic coherence. To the best of our knowledge, this is not allowed in FHIR extensions, which do not specify the relation of the extension to the resource it extends. Crucially, neither FHIR nor OMOP supports the definition of PCEs, a capability central to our solution for expressing complex clinical nuances. In comparison to ontology alignment frameworks, our approach does not align ontologies but allows their combined use to compose concepts absent in both. Furthermore, a key difference lies in the methodology; our solution focuses entirely on the semantic description of data where it resides, avoiding the necessity of data movement or physical transformation inherent in many FHIR or OMOP-based implementations. Because our primary intent is semantic description and not structural or physical data modeling, a direct, formal comparison between these solutions is not appropriate.

The decision to build upon and extend SNOMED CT is grounded in the need for a robust and logical foundation, which its unique extensible grammar and syntax provide. This work shows that extending a powerful existing standard can be a path toward semantic interoperability. Four large domains of clinical reality, identified as lacking in SNOMED CT and known as challenging fields to represent, have been partially solved using this approach. Despite allowing for the representation of a large amount of data elements that were out of reach of current SNOMED CT rules, there remain gaps in representation that have not yet been resolved. Each category described previously has seen its coverage improved, but as things stand, complete coverage of these domains is still out of reach.

A statistical analysis of the extensions used shows that the 4500 PCEs created using extensions represent less than 4% of data elements and 3% of instances (Table 1). However, a deeper analysis shows that while seemingly sparse overall, extensions are very useful for certain data types. Laboratory procedures and medical devices, for example, are two collections in which 27% (250,617,656/935,678,204) and 23% (1,140,820/4,926,815) of instances, respectively, were represented with extensions. Other collections, such as formularies, have nearly 4000 extensions used.

XSL•FO
RenderX

Table 1. Detailed statistics of representation by category of data using extensions.

Source	IDs encoded, n	Extensions used, n (%)	Instances encoded	Extension instances, n (%)
Administrative stay	1172	2 (0.17)	78,020,550	131,832 (.17)
Anesthesiology	2139	11 (0.51)	117,419,522	1,275,337 (1.09)
Medical devices	699	235 (33.62)	4,926,815	1,140,820 (23.16)
Formularies	51,322	3925 (7.65)	1,641,633,208	119,941,273 (7.31)
ICU[a]	6934	157 (2.26)	6,656,119,500	367,548 (.005)
Laboratory	7231	72 (1)	935,678,204	250,617,656 (26.78)
Observations	2784	1 (0.03)	2,440,891,370	1003 (.00004)
Patient problems	25,233	72 (0.28)	1,587,507	976 (.06)
Prescription	9837	0 (0)	26,805,381	0 (0)
Radiology	2586	2 (0.07)	8,562,371	2 (.00002)
Procedures	8001	23 (0.28)	1,452,038,108	8,934,163 (.6)
Total	119,941	4500 (3.75)	13,362,644,974	382,410,610 (2.86)

[a]ICU: intensive care unit.

Closer examination shows clearly that the extensions have proven extremely useful to fill certain specific gaps within SNOMED CT. In these cases, the gaps identified were SARS-CoV-2 strains for laboratory procedures, sizes for medical devices, and occupations for formularies. Looking in even further detail, the 4500 PCEs created represent 488 distinct expressions. However, certain specific expressions have proven very useful. Extending representation of negation was a big part, with the Without and Except for attributes used a combined 88 times, and sizes for medical devices used 178 times.

Machine-Readable Interoperability

A primary concern regarding the modification of established standards is the potential loss of machine readability. Our implementation presents a dichotomy in this regard. The modifications made strictly to the MRCM, such as domain and range extensions for scalars or negation attributes, remain compliant with SNOMED CT's structure. These are natively processable by standard terminology servers like Snowstorm, provided the modified MRCM is loaded [55].

However, the integration of external vocabularies required modifications to the ABNF syntax (specifically the use of the tilde delimiter), which constitutes a divergence from the standard CG. Consequently, standard parsers and expression constraint language query engines cannot currently process these specific expressions without adaptation. We have validated that these expressions can be parsed using a modified ABNF parser, and we are currently adapting open-source tools to accommodate this syntax. Until these tools are widely available, this specific aspect of the grammar hinders immediate interoperability with the existing global standards ecosystem.

This technical gap is compounded by a governance challenge, as the number of integrated terminologies increases, so does the need for robust documentation to prevent ambiguity. A formal glossary or registry is therefore required to track which terminologies are in use, their specific versions, and the rules governing their application. For clarity, versioning could even be embedded directly within an expression, for example: ~CTCAE~v5.0~4028512~. A more sustainable formalism could use Uniform Resource Identifiers to designate terminology releases. This could be implemented without modifying the ABNF.

Challenges Encountered

Scalars are a massive chapter, and representing certain situations, such as ranges, has not been resolved yet. Allowing for scalar values in the range of most hierarchies also requires defining new attribute relations for each use case encountered. Negation is resolved by exclusively using and extending existing SNOMED CT concepts, as this is sufficient for our needs and maintains internal consistency within the grammar. However, this is not the only valid approach. More complex scenarios could be addressed by incorporating a formal negation operator, such as the widely accepted "¬" symbol, as described by Schulz et al [37]. Similarly, reification offers another method for handling negation within SNOMED CT, though it lacks a directly queryable negation indicator [36,37]. This would require additional ABNF syntax modification, and the impact it would have on DL means that this has not been implemented.

Other challenging situations fall into different categories but account for only a small fraction of all data elements, such as rare, complex labels within otherwise well-resolved categories. These include highly specific laboratory procedures that represent less than 1% (1,475,290/250,521,754) of all laboratory procedure instances. They could likely be addressed by incorporating more specialized terminologies. However, because these cases are anecdotal, the effort required to search for and integrate additional terminology is not justified for now.

Data-element categories with no widely adopted standard and no suitable SNOMED CT codes for creating new PCEs are another challenge. An example is triage-sheet questions, which include 350 distinct labels across 4 million instances, a negligible portion of the overall dataset. Work is ongoing to find solutions, primarily through new MRCM rules, such as

XSL•FO
RenderX

defining an appropriate relationship between Observable Entities and Clinical Findings (current options like “Precondition” or “Has realization” are inadequate), or between Environments and Procedures.

Finally, some grammatical constructions are poorly supported by current grammar rules. A key case is the presence of “or” in problem-list entries. For concepts containing “or,” we currently apply a rule that the label is represented only if both concepts share a sufficiently close semantic parent; otherwise, it remains unmapped. Some labels can be resolved this way, such as Intervention de gynécologie ou d’obstétrique = Operation on female genital organs (procedure). But others are too semantically broad or distant, with no precise match, such as Pneumonie ou pneumopathie, where the shared parent Disorder of lung (disorder) is considered too vague and would cause excessive information loss. In total, our database contains 2895 labels with “or,” representing 17.5 million instances. Among these, 834 (28%) labels have been successfully resolved, covering 10 (57%) million instances.

Limitations

This work presents limitations. For reasons mentioned previously, no interannotator agreement was carried out on the work done. The goal was initially to develop the approach for internal use only. Since it evolved gradually in parallel to our improving capabilities, there was never a “before” and “after” to evaluate. Also, since the team is made up of a set core of 3-4 members, it was considered more useful for us to evaluate our work in focus group discussions. There was never a need internally for a more explicit evaluation because problems are reviewed so often. Finally, while the intention is to generalize the way people use SNOMED CT to represent data, one of the founding principles of SNOMED CT is that there are many ways to express the same thing. As the most important outcome is semantic accuracy, maintaining the meaning of the data element, if two expressions mean the same thing but are written differently, they are still both correct. The expected outcomes were therefore never exactly matched. Such subtleties would not appear in a formal interannotator agreement, reducing its interpretability, an important point, which, in our opinion, diminished the potential impact of such an evaluation.

The byproduct of an inclusive grammar, as described in this work, is that it will inevitably make reuse more difficult. There will be links created that should not work, and inconsistencies in DL. The primary limitation of this implementation lies in the operational integration of external vocabularies with standard SNOMED CT tools. While separation of grammar and vocabulary was identified as a crucial need for an FDG, its implementation in the proposed approach is incomplete. The MRCM we chose to keep and extend constitutes a set of rules that are semantic, not merely grammatical. Currently, the terminologies added to the grammar are attributed a SNOMED CT code as a parent, to allow initial querying. But the new codes themselves cannot be directly accessed yet. For this to function, the integrated terminologies need to be definitively added to SNOMED CT. There are ways to do this in theory, such as creating a new concept with a valid SNOMED CT concept identifier for each, but they have not yet been explored in detail or implemented. Applying these solutions would greatly enhance reuse capabilities, improving interoperability and consistency of results. Modifying the ABNF syntax and parser is a first step in this direction.

The results only cover the sources that have been selected for representation so far. As the project continues, new sources will be analyzed, extracted, and represented. Therefore, it is not possible, to date, to give a clear image of the progress compared to the complete data warehouse. However, the choice of the sources to be added first has been designed to cover first the core of the electronic health record, before smaller sources. Additionally, new problems will inevitably appear, which will need to be addressed, and new rules will have to be added. As this work focuses primarily on structured data, it is not possible to confront our semantic coverage to the full information content of the electronic health record, but this is being addressed through automatic annotation of free text using natural language processing in a similar manner.

Finally, the approach was tested only in our environment and is currently used nowhere else. The end goal, however, is, of course, to apply it elsewhere to test its reproducibility, as the complexity and variability of the approach mean it may not be immediately applicable in different settings.

Governance and Implementation

Key recommendations can be derived from this work to address the limitations discussed above. To safely adopt these extensions, institutions should establish a minimum governance structure consisting of a multidisciplinary committee of domain experts and SNOMED CT–certified terminologists. This body acts as a gatekeeper, validating new representation rules through a consensus-based cycle only when standard SNOMED CT concepts or official maps (eg, LOINC and Orphanet) are proven insufficient. Key steps toward alignment with SI include maintaining a formal local registry of all grammar extensions and prioritizing the use of standard extension mechanisms (MRCM) over syntax modifications to preserve compatibility with the broader ecosystem. In terms of maturity, the MRCM extensions for scalar values, temporality, uncertainty, and negation are stable and ready for immediate reuse in standard terminology servers (eg, Snowstorm). Conversely, the ABNF syntax modifications required for integrating external vocabularies remain experimental; they currently necessitate custom parsing tools and can be adopted as a proof-of-concept pending the evolution of standard tooling.

Future work

Steps have been taken to start implementing the strategy in different institutions, as interest has been strong, and collaborations are currently ongoing, but are at a preliminary stage for the moment. External replication and independent validation are being undertaken, which aim to evaluate the adaptability and reusability of our approach, demonstrating its generalizability. Additionally, specific projects in the hospital have already benefited from using our approach to gather data on subjects such as biomarkers in the field of genomics, which was not possible before.

XSL•FO
RenderX

Furthermore, while this work relied on expert manual representation, future research will focus on automating the translation of free-text clinical notes into PCEs using natural language processing. Emerging methodologies leverage large language models (LLMs) equipped with retrieval-augmented generation to ground generated expressions in the extensive SNOMED CT hierarchy, a technique showing potential for high-fidelity clinical coding [56]. However, we acknowledge that the generation of structured sequences via LLMs introduces specific risks, such as exposure bias, which can affect the reliability of the output. Recent work highlights the importance of mitigating these biases in LLM distillation to ensure robust structured generation [57]. Addressing these challenges would dramatically scale implementation and unlock the potential for advanced semantic reasoners to infer new knowledge from grammatically rich, machine-readable clinical data.

Conclusions

This methodological study demonstrates how SNOMED CT's CG can be leveraged and extended to address recurrent semantic gaps encountered in a large-scale clinical data warehouse. Rather than constructing a new FDG from scratch, we have approached this theoretical goal by systematically extending the capabilities of an existing standard. Through specific modifications to the MRCM and syntax, we successfully addressed complex representational challenges such as negation, scalar values, and the integration of external terminologies, thereby reducing the tension between clinical expressivity and standardization.

Supported by a decade of implementation at HUG, this work illustrates that a grammatical framework, separating the rules of composition from the vocabulary itself, is essential for capturing meaning at scale. While standardized vocabularies provide the necessary lexical building blocks, it is the flexibility of the grammar that determines the fidelity of the representation. This study serves as a proof-of-concept that semantic interoperability can be advanced by methodically extending the expressive power of existing standards. However, the widespread adoption of such extensions requires robust governance frameworks to collaboratively manage and share grammatical rules across institutions.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025) [58], the following tasks were delegated to GAI tools under full human supervision: proofreading, editing, and reformatting. The GAI tools used were Gemini 2.5 Pro, Gemini 3, and ChatGPT 4.5. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

This declaration is submitted by CGB.

Funding

This work was funded with a grant from the Private Foundation of the Geneva University Hospitals. Funding by the Swiss Personalized Health Network initiative has partly contributed to enabling this work. The funding sources had no input or influence in the decision to publish this paper, and were not involved in the writing, editing, or submission procedures.

Authors' Contributions

Conceptualization: CGB, JE, CL
Data curation: CGB, JE, MB, AB, MM, YZ
Methodology: CGB, JE, MB, AB, MM, YZ
Visualization: JE
Supervision: CGB, CL
Funding acquisition: CL
Validation: CL
Project administration: CL
Writing – original draft: CGB, JE
Writing – review & editing: CGB, JE, CL

Conflicts of Interest

None declared.

Multimedia Appendix 1
Machine Readable Component Model modifications and additions.
[XLSX File (Microsoft Excel File), 12 KB - jmir_v28i1e80314_app1.xlsx]

Multimedia Appendix 2
Modified Augmented Backus-Naur Form syntax.
[TXT File , 2 KB - jmir_v28i1e80314_app2.txt]

XSL•FO
RenderX

References

1. Gaudet-Blavignac C, Raisaro JL, Touré V, Österle S, Crameri K, Lovis C. A national, semantic-driven, three-pillar strategy to enable health data secondary usage interoperability for research within the Swiss personalized health network: methodological study. JMIR Med Inform 2021;9(6):e27591 [FREE Full text] [doi: 10.2196/27591] [Medline: 34185008]
2. The European Health Data Space. URL: https://www.european-health-data-space.com/ [accessed 2025-05-27]
3. Gordon W, Gottlieb D, Kreda D, Mandel J, Mandl K, Kohane I. Patient-led data sharing for clinical bioinformatics research: USCDI and beyond. J Am Med Inform Assoc 2021;28(10):2298-2300 [FREE Full text] [doi: 10.1093/jamia/ocab133] [Medline: 34279631]
4. Cruz IF, Antonelli FP, Stroe C. AgreementMaker: efficient matching for large real-world schemas and ontologies. 2009 Presented at: Proceedings of the VLDB Endowment: Proceedings of the 35th International Conference on Very Large Data Bases; 2009 August 24-28; Lyon, France p. 1586-1589. [doi: 10.14778/1687553.1687598]
5. He Y, Chen J, Antonyrajah D, Horrocks I. BERTMap: a BERT-based ontology alignment system. 2022 Presented at: Proceedings of the AAAI Conference on Artificial Intelligence; 2022 Feb 22-Mar 1; Vancouver p. 5684-5691. [doi: 10.1609/aaai.v36i5.20510]
6. Compositional grammar specification | specifications SNOMED CT compositional grammar specification | SNOMED international documents. SNOMED. URL: https://docs.snomed.org/snomed-ct-specifications/snomed-ct-compositional-grammar-specification [accessed 2025-11-28]
7. Baader F, Horrocks I, Sattler U. Description logics. In: Handbook of Knowledge Representation. Amsterdam: Elsevier; 2008:135-179.
8. Benson T, Grieve G. Principles of Health Interoperability: SNOMED CT, HL7 and FHIR. 3rd ed. London: Springer; 2016.
9. OMOP Common Data Model? OHDSI. URL: https://www.ohdsi.org/data-standardization/the-common-data-model/ [accessed 2020-02-24]
10. Health informatics? Interoperability and integration reference architecture? Model and framework. ISO. URL: https://www.iso.org/obp/ui/#iso:std:iso:23903:ed-1:v2:en [accessed 2025-11-28]
11. Wang L, Wen A, Fu S, Ruan X, Huang M, Li R, et al. A scoping review of OMOP CDM adoption for cancer research using real world data. NPJ Digit Med 2025;8(1):189 [FREE Full text] [doi: 10.1038/s41746-025-01581-7] [Medline: 40189628]
12. Sholle ET, Cusick M, Davila MA, Kabariti J, Flores S, Campion TR. Characterizing basic and complex usage of i2b2 at an academic medical center. AMIA Jt Summits Transl Sci Proc 2020;2020:589-596 [FREE Full text] [Medline: 32477681]
13. Ceusters W, Rabenberg M. Semantic difficulties in FHIR 'conditions'. Stud Health Technol Inform 2025;327:7-11. [doi: 10.3233/SHTI250263] [Medline: 40380375]
14. Kent S, Burn E, Dawoud D, Jonsson P, Østby JT, Hughes N, et al. Common problems, common data model solutions: evidence generation for health technology assessment. Pharmacoeconomics 2021;39(3):275-285 [FREE Full text] [doi: 10.1007/s40273-020-00981-9] [Medline: 33336320]
15. Chute CG. Clinical classification and terminology: some history and current observations. J Am Med Inform Assoc 2000;7(3):298-303 [FREE Full text] [doi: 10.1136/jamia.2000.0070298] [Medline: 10833167]
16. 5.1 normative specification--compositional grammar--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCSCG/5.1+Normative+Specification [accessed 2025-07-06]
17. SNOMED CT editorial guide--SNOMED CT editorial guide--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCEG [accessed 2024-02-20]
18. Lee D, Cornet R, Lau F, de Keizer N. A survey of SNOMED CT implementations. J Biomed Inform 2013;46(1):87-96 [FREE Full text] [doi: 10.1016/j.jbi.2012.09.006] [Medline: 23041717]
19. Lee D, de Keizer N, Lau F, Cornet R. Literature review of SNOMED CT use. J Am Med Inform Assoc 2014;21(e1):e11-e19 [FREE Full text] [doi: 10.1136/amiajnl-2013-001636] [Medline: 23828173]
20. Domain specific modeling | specifications SNOMED CT editorial guide | SNOMED International Documents. SNOMED. URL: https://docs.snomed.org/snomed-ct-specifications/snomed-ct-editorial-guide/readme/authoring/domain-specific-modeling [accessed 2025-11-28]
21. Cornet R, Chute CG. Health concept and knowledge management: twenty-five years of evolution. Yearb Med Inform 2016;Suppl 1(Suppl 1):S32-S41 [FREE Full text] [doi: 10.15265/IYS-2016-s037] [Medline: 27488404]
22. Ehrsam J, Gaudet-Blavignac C, Mattei M, Baumann M, Lovis C. Semantics in action: a guide for representing clinical data elements with SNOMED CT. J Biomed Semantics 2025;16(1):7 [FREE Full text] [doi: 10.1186/s13326-025-00326-5] [Medline: 40149003]
23. Ehrsam J, Gaudet-Blavignac C, Lovis C. Scalar values in SNOMED CT: a proposed extension. Stud Health Technol Inform 2024;316:1363-1367. [doi: 10.3233/SHTI240665] [Medline: 39176634]
24. Course categories | e-learning. SNOMED International. URL: https://elearning.ihtsdotools.org/course/ [accessed 2025-11-28]
25. Gaudet-Blavignac C, Rudaz A, Lovis C. Building a shared, scalable, and sustainable source for the problem-oriented medical record: developmental study. JMIR Med Inform 2021;9(10):e29174 [FREE Full text] [doi: 10.2196/29174] [Medline: 34643542]
26. SNOMED CT authoring level 1 course. SNOMED International. URL: https://courses.ihtsdotools.org/product?catalog=AL1 [accessed 2025-11-28]

XSL•FO
RenderX

27. 3 requirements--machine readable concept model--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCMRCM/3.+Requirements [accessed 2024-11-28]
28. SR 810.30--Federal Act of 30 September 2011 on research. Fedlex. URL: https://www.fedlex.admin.ch/eli/cc/2013/617/en [accessed 2025-12-19]
29. RO 2022 491 - Loi fédérale du 25 septembre 2020 sur la protection des données. Fedlex. URL: https://www.fedlex.admin.ch/eli/oc/2022/491/fr [accessed 2025-12-19]
30. Dąbrowska E. What exactly is universal grammar, and has anyone seen it? Front Psychol 2015;6:852 [FREE Full text] [doi: 10.3389/fpsyg.2015.00852] [Medline: 26157406]
31. 10 extension and customization--SNOMED CT starter guide--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCSTART/10.+Extension+and+Customization [accessed 2025-04-29]
32. Scalar (physics). Wikipedia. URL: https://en.wikipedia.org/w/index.php?title=Scalar_(physics)&oldid=1279729075 [accessed 2025-04-29]
33. SNOMED International proposal for representing concrete domains in RF2. SNOMED International. URL: https://confluence.ihtsdotools.org/mag/concrete-domains-community-of-practice-consultation/snomed-international-proposal-for-representing-concrete-domains-in-rf2 [accessed 2024-03-18]
34. Rector AL, Brandt S. Why do it the hard way? The case for an expressive description logic for SNOMED. J Am Med Inform Assoc 2008;15(6):744-751 [FREE Full text] [doi: 10.1197/jamia.M2797] [Medline: 18755993]
35. ISO 8601? Date and time format. ISO. 2017. URL: https://www.iso.org/iso-8601-date-and-time-format.html [accessed 2025-05-08]
36. Martínez-Costa C, Miñarro-Giménez J, Hausam R, Schulz S. Addressing the negation gap in SNOMED CT by reified negated concepts. J Biomed Semantics 2015;6:23. [doi: 10.1186/s13326-015-0018-1]
37. Schulz S, Markó K, Suntisrivaraporn B. Formal representation of complex SNOMED CT expressions. BMC Med Inform Decis Mak 2008;8 Suppl 1(Suppl 1):S9 [FREE Full text] [doi: 10.1186/1472-6947-8-S1-S9] [Medline: 19007446]
38. Situation with explicit context modeling--SNOMED CT editorial guide--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCEG/Situation+with+Explicit+Context+Modeling [accessed 2025-05-06]
39. Knowledge representation--SNOMED CT editorial guide--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCEG/Knowledge+Representation [accessed 2025-05-07]
40. Haendel MA, Chute CG, Robinson PN. Classification, ontology, and precision medicine. N Engl J Med 2018;379(15):1452-1462 [FREE Full text] [doi: 10.1056/NEJMra1615014] [Medline: 30304648]
41. Rothwell DJ, Cote RA, Cordeau JP, Boisvert MA. Developing a standard data structure for medical language--the SNOMED proposal. Proc Annu Symp Comput Appl Med Care 1993:695-699 [FREE Full text] [Medline: 8130565]
42. SNOMED CT maps. SNOMED International. URL: https://www.snomed.org/maps [accessed 2024-11-28]
43. Request form for LOINC/SNOMED CT expression association and map sets file. LOINC. URL: https://loinc.org/snomed-file-request/ [accessed 2024-11-28]
44. LOINC/SNOMED CT. URL: https://loincsnomed.org/ [accessed 2024-11-28]
45. Authoring--SNOMED CT editorial guide--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCEG/Authoring [accessed 2025-06-06]
46. Proximal primitive parent--SNOMED CT glossary--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCGLOSS/proximal+primitive+parent [accessed 2025-06-06]
47. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 2018;34(23):4121-4123 [FREE Full text] [doi: 10.1093/bioinformatics/bty407] [Medline: 29790939]
48. O'Toole Á, Pybus OG, Abram ME, Kelly EJ, Rambaut A. Pango lineage designation and assignment using SARS-CoV-2 spike gene nucleotide sequences. BMC Genomics 2022;23(1):121 [FREE Full text] [doi: 10.1186/s12864-022-08358-2] [Medline: 35148677]
49. Tracking SARS-CoV-2 variants. World Health Organization. URL: https://www.who.int/activities/tracking-SARS-CoV-2-variants [accessed 2025-06-09]
50. CoVariants. URL: https://covariants.org/ [accessed 2025-06-09]
51. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 2020;5(11):1403-1407 [FREE Full text] [doi: 10.1038/s41564-020-0770-5] [Medline: 32669681]
52. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, et al. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol 2005;6(5):R44 [FREE Full text] [doi: 10.1186/gb-2005-6-5-r44] [Medline: 15892872]
53. Langlotz CP. RadLex: a new method for indexing online educational materials. Radiographics 2006;26(6):1595-1597. [doi: 10.1148/rg.266065168] [Medline: 17102038]
54. Trotti A, Colevas AD, Setser A, Rusch V, Jaques D, Budach V, et al. CTCAE v3.0: development of a comprehensive grading system for the adverse effects of cancer treatment. Semin Radiat Oncol 2003;13(3):176-181. [doi: 10.1016/S1053-4296(03)00031-6] [Medline: 12903007]

XSL•FO
RenderX

55. Terminology servers. Implementation. SNOMED. URL: https://www.implementation.snomed.org/terminology-services [accessed 2025-11-28]
56. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. 2022 Presented at: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing Abu Dhabi, United Arab Emirates: Association for Computational Linguistics; 2022; Abu Dhabi p. 1998-2022. [doi: 10.18653/v1/2022.emnlp-main.130]
57. Pozzi A, Incremona A, Tessera D, Toti D. Mitigating exposure bias in large language model distillation: an imitation learning approach. Neural Comput Applic 2025;37(18):12013-12029. [doi: 10.1007/s00521-025-11162-0]
58. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (generative AI delegation taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res 2025:1-27. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]

Abbreviations

ABNF: Augmented Backus-Naur Form
CG: compositional grammar
CTCAE: Common Terminology Criteria for Adverse Events
DL: Description Logics
FDG: formal descriptive grammar
FHIR: Fast Healthcare Interoperability Resources
HUG: Geneva University Hospitals
ICD-10: International Statistical Classification of Diseases, Tenth Revision
ISO: International Organization for Standardization
LLM: large language model
LOINC: Logical Observation Identifiers Names and Codes
MRCM: Machine Readable Component Model
OMOP: Observational Medical Outcomes Partnership
PCE: postcoordinated expression
SI: SNOMED International
SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms

Edited by J Sarvestan; submitted 08.Jul.2025; peer-reviewed by A Pozzi, T Cook; comments to author 21.Aug.2025; revised version received 19.Dec.2025; accepted 20.Dec.2025; published 28.Jan.2026.

Please cite as:
Gaudet-Blavignac C, Ehrsam J, Baumann M, Bensahla A, Mattei M, Zheng Y, Lovis C
Extended Grammar of Systematized Nomenclature of Medicine – Clinical Terms for Semantic Representation of Clinical Data: Methodological Study
J Med Internet Res 2026;28:e80314
URL: https://www.jmir.org/2026/1/e80314
doi:10.2196/80314
PMID:

©Christophe Gaudet-Blavignac, Julien Ehrsam, Monika Baumann, Adel Bensahla, Mirjam Mattei, Yuanyuan Zheng, Christian Lovis. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Establishment and Optimization of a Patient-Reported Outcome–Based Electronic-Diary for Symptoms Evaluation in Patients With Gastroesophageal Reflux Disorder: Prospective Cohort Study

Yun-Chun Chen[1], MS; Yen-Po Wang[2,3,4], MD, PhD; Jui-Hsuan Hung[5], MS; Da-Wei Wang[5†], PhD; Shang-Liang Wu[6], DrPH; Li-Fen Chen[2], PhD; Yueh-Hsin Ping[1,7], PhD; Mei-Lien Pan[8], PhD; Ching-Liang Lu[2,3,4], MD

[1]Department and Institute of Pharmacology, National Yang Ming Chiao Tung University, Taipei, Taiwan

[2]Institute of Brain Science, National Yang Ming Chiao Tung University, Taipei, Taiwan

[3]Division of Gastroenterology, Department of Medicine, Taipei Veterans General Hospital, Taipei, Taiwan

[4]Endoscopy Center for Diagnosis and Treatment, Department of Medicine, Taipei Veterans General Hospital, No.201, Sec. 2, Shipai Rd., Beitou District, Taipei, Taiwan

[5]Institute of Information Science, Academia Sinica, Taipei, Taiwan

[6]Department of Medical Research, Taipei Veterans General Hospital, Taipei, Taiwan

[7]Institute of of Biophotonics, National Yang Ming Chiao Tung University, Taipei, Taiwan

[8]Institute of Hospital and Health Care Administration, National Yang Ming Chiao Tung University, Taipei, Taiwan

[†]deceased

Corresponding Author:
Ching-Liang Lu, MD
Institute of Brain Science, National Yang Ming Chiao Tung University, Taipei, Taiwan

Abstract

Background: Gastroesophageal reflux disease (GERD) symptoms substantially impair patients' quality of life. The use of patient-reported outcome (PRO) instruments for symptom measurement has been advocated by regulatory authorities. However, current tools for GERD symptom evaluation are limited by recall bias. To improve the real-time characterization of GERD symptoms, we developed an electronic diary (e-diary) for daily symptom monitoring.

Objective: This study aimed to develop and optimize a PRO-based e-diary for GERD symptom evaluation and to examine the effect of symptom frequency on adherence.

Methods: The GERD e-diary evaluated 8 daytime (acid regurgitation, cough, heartburn, sour taste in the mouth, hiccups, hoarseness, dysphagia, and chest pain) and 2 nighttime symptoms (acid regurgitation and cough) for 8 consecutive weeks. Adherence, defined as the daily completion rate of e-diary, was analyzed and optimized across three stages: (1) no reminder, (2) sending reminder SMS text messaging upon the detection of missing data (no reminders during the first 3-5 days after enrollment), and (3) immediate installation of reminder system at enrollment. Weekly symptom frequency was calculated as the sum of symptomatic days per week. Multiple regression analyses were performed to examine the effects of system optimization and symptom frequency on adherence after controlling for confounders.

Results: A total of 138 patients with GERD (70 men, 68 women; mean [SD] age 52.9 [12.3] years) were recruited. At the first stage, the adherence was 47.2%, 40%, and 57.6% for nighttime, daytime, and overall symptoms. System optimization significantly improved the adherence of nighttime symptoms by 12.5% (95% CI 3.7-21.3) and 10.9% (95% CI 2.6-19.2), daytime symptoms by 21.7% (95% CI 14.2-29.2) and 20.8% (95% CI 13.7-27.9), and overall symptoms by 16.5% (95% CI 9.8-23.2) and 18.5% (95% CI 12.2-4.8) in the second and third stages, respectively. Symptom frequency was positively associated with adherence, increasing by 0.7% (95% CI 0.6-0.8) for overall symptoms and 0.9% (95% CI 0.7-1) for both daytime and nighttime symptoms per additional symptom frequency. Adherence gradually decreased along the study period. (first vs eighth week: nighttime 80.1% vs 61.5%, $\beta=-18.6$, 95% CI −26.9 to −10.3; daytime 85.1% vs 66.8%, $\beta=-18.3$, 95% CI −25.6 to −11; overall 95.1% vs 78%, $\beta=-17.2$, 95% CI −23.5 to −10.9).

Conclusions: The adherence of the GERD e-diary can be optimized by using SMS text messaging reminders. Higher symptom frequency was associated with increased adherence, although engagement declined over time. This innovative PRO-based e-diary with prolonged recording provides a real-time, prospective tool that overcomes the recall and ecological biases inherent in traditional short-term retrospective GERD symptom assessments. This advancement empowers patients through improved

XSL•FO
RenderX

self-awareness and provides physicians with precise, long-term data, facilitating tailored therapeutic interventions and supporting personalized GERD management.

(J Med Internet Res 2026;28:e83680) doi:10.2196/83680

KEYWORDS

adherence; e-diary; gastroesophageal reflux disease; GERD; symptom frequency; system optimization

Introduction

Gastroesophageal reflux disease (GERD) is a common disorder affecting approximately 20% of the general population [1]. GERD symptoms, including typical (heartburn and acid regurgitation) and atypical (cough, chest pain, sour taste in the mouth, hoarseness, dysphagia, and hiccups) symptoms, would bring significant burdens on patients' quality of life, reduce work productivity, and increase considerable medical resources worldwide [2-6]. As a symptom-driven disease, GERD should be evaluated for the presence, frequency, and severity of bothersome symptoms [1]. However, patient-reported clinical management outcomes often remain unsatisfactory, with more than 50% of patients continuing to experience bothersome symptoms despite proton pump inhibitor therapy [7]. This underscores the need for more nuanced approaches to symptom assessment and management.

Patient-reported outcomes (PROs) are the most direct and measured instrument for evaluation, an approach advocated by the Food and Drug Administration of the United States to assess treatment efficacy [8]. In GERD research, PROs have been used to evaluate the symptomatic improvements after acid-suppressant drugs, such as proton pump inhibitors or histamine-H2-receptor antagonists [9-11]. PROs have also been used in long-term studies to evaluate symptom resolution after surgery and endoscopic interventions and to assess improvements in quality of life [12-14]. However, most previous studies relied on questionnaires that required patients to recall symptoms over prolonged periods, possibly leading to recall bias. Furthermore, during outpatient evaluation and follow-up in patients with GERD, important information about the time course of symptom occurrence and relief is still limited because the caring physicians typically assess bothersome symptoms after weeks or even months of treatment. Notably, it has been reported that a 30-day recall of self-reported urinary incontinence is impaired and can be associated with demographic and psychosocial characteristics [15]. Patients with GERD are expected to face similar situations when using questionnaires to recall their symptoms [16-18]. Therefore, daily assessments of the GERD symptoms would provide a better understanding of the natural course of GERD symptoms and potentially alleviate the recall bias. Although patients with chronic gastroenterological diseases express strong interest in using mobile health apps for disease management, real-world adherence remains low, with high dropout rates [19]. Some studies report that up to 80% of participants engage only minimally or discontinue regular use, with retention rates dropping to as low as 3.9% after 15 days [20-22]. Existing GERD questionnaires also face limitations: their perceived impracticality for routine clinical use, along with challenges inherent to traditional paper-based symptom tracking—such as recall bias, inconsistent formatting, and poor adherence leading to incomplete or inaccurate data—collectively impede accurate and continuous symptom assessment for both patients and clinicians [23]. Given the high patient interest in digital health solutions, there is a critical need for novel mobile tools capable of overcoming these traditional limitations [24]. Although there are some smartphone apps for gastrointestinal disease, including GERD, available in the market, no app is compliant with the 2022 American College of Gastroenterology Guideline and not designed for physicians' management of GERD symptoms, and most of the apps are not evidence based [24-26].

To overcome recall bias and better address the progression and regression of GERD symptoms, in this study, we developed an e-diary for recording GERD symptoms. As low adherence in GERD diary recording would potentially limit the accountability of the e-diary, affecting clinical assessments and the associated outcomes, we tried to develop a strategy to enhance the 8-week adherence using the e-diary. Furthermore, our clinical observations and existing literature suggest a strong link between symptom severity and recording motivation: patients experiencing more frequent or intense symptoms typically demonstrate greater adherence and accuracy in documentation [27,28]. Therefore, we subsequently tested that higher symptom frequency would be associated with increased adherence to e-diary recording. The objective of the study was to examine the effects of symptom frequency and the degree of system optimization on weekly adherence, with the hypothesis that after adjusting for potential confounders, higher symptom frequency and greater system optimization would be associated with better weekly adherence. This will serve as a basis for developing strategies to improve patient adherence in completing the e-diary, thereby enhancing the clinical validity of the e-diary for GERD symptoms.

Methods

In this study, we followed the EQUATOR (Enhancing the Quality and Transparency of Health Research) guidelines using the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist [29].

Study Design and Participants

The study was a prospective observational cohort study. From October 2021 to January 2023, consecutive adults (aged ≥20 y) with GERD were recruited from the outpatient clinics of gastroenterology in Taipei Veterans General Hospital, a tertiary medical center in northern Taiwan. Eligible participants presented with either typical (acid reflux or heartburn) or atypical symptoms (hoarseness, throat discomfort, cough, or chest pain) for at least 3 months. Heartburn or regurgitation

XSL•FO
RenderX

should be noted for 4 days or more during the 7 days before the first visit. Patients who expressed interest and consented to daily symptom recording using the GERD e-diary were enrolled. A trained research assistant installed the GERD e-diary on the mobile phones of participants and provided instructions for use. These patients were asked to complete the GERD e-diary on a daily basis for 8 weeks. One or two follow-up visits in gastroenterology clinics would be arranged. To minimize potential selection and information bias, consecutive eligible patients attending the outpatient gastroenterology clinics were approached, and all data were prospectively recorded using standardized digital forms. As symptom data were collected directly from patients through an electronic diary (e-diary), recall bias was substantially reduced compared with paper questionnaires. The study size (n=138) was determined pragmatically according to the expected recruitment capacity within the 15-month study period and the pilot nature of this observational study. No formal sample size calculation was performed, consistent with the exploratory objectives. Patient flow from screening to analysis is summarized in Figure 1, which illustrates the number of patients invited (n=176), enrolled (n=138), and completing the 8-week follow-up, together with the reasons for exclusion.

Figure 1. Patients flowchart. GERD: gastroesophageal reflux disease.

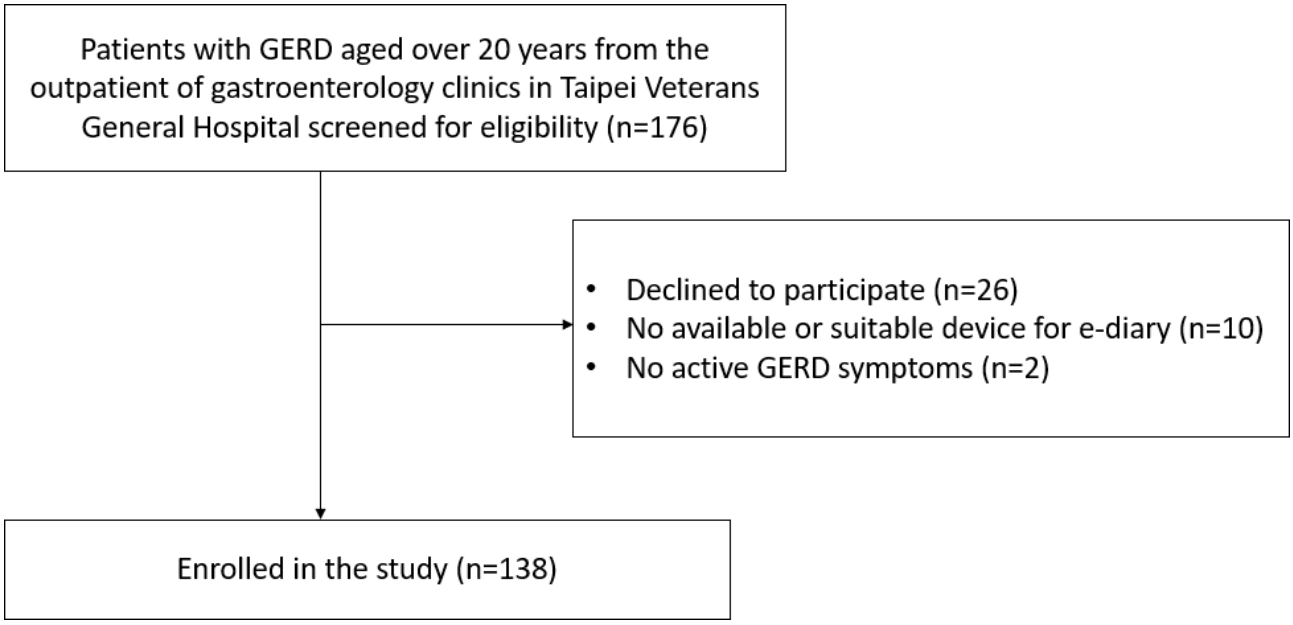

The GERD E-Diary

The web-based GERD e-diary was designed to record both nighttime and daytime GERD symptoms. Patients were asked to fill out the e-diary twice a day. In the morning, they were asked to fill in nighttime symptoms during the previous night, while in the evening, they were asked to record daytime symptoms that occurred during the concurrent day. The GERD e-diary was designed to assess 8 daytime symptoms (acid reflux, cough, heartburn, sour taste in the mouth, burping, hoarseness, dysphagia, and chest pain) and 2 nighttime symptoms (acid reflux and cough). The choice of these GERD symptom records was based on the Modified Reflux Symptom Questionnaire–Electronic Diary, which proved to be a reliable and valid PRO instrument [30]. After completing the e-diary, educational or inspiring messages would be randomly displayed to encourage continued participation. Although it was a web-based e-diary, we installed it as a smartphone shortcut icon that functioned like a mobile app, as Figure 2 showed.

XSL•FO
RenderX

Figure 2. Interface of the web-based gastroesophageal reflux disease (GERD) e-diary used in this cohort study conducted at Taipei Veterans General Hospital, Taiwan (October 2021 to January 2023). The shortcut icon was added to patients' smartphones, providing an app-like interface for direct daily input of GERD symptoms. (A) Home screen appearance; (B) symptom entry interface.

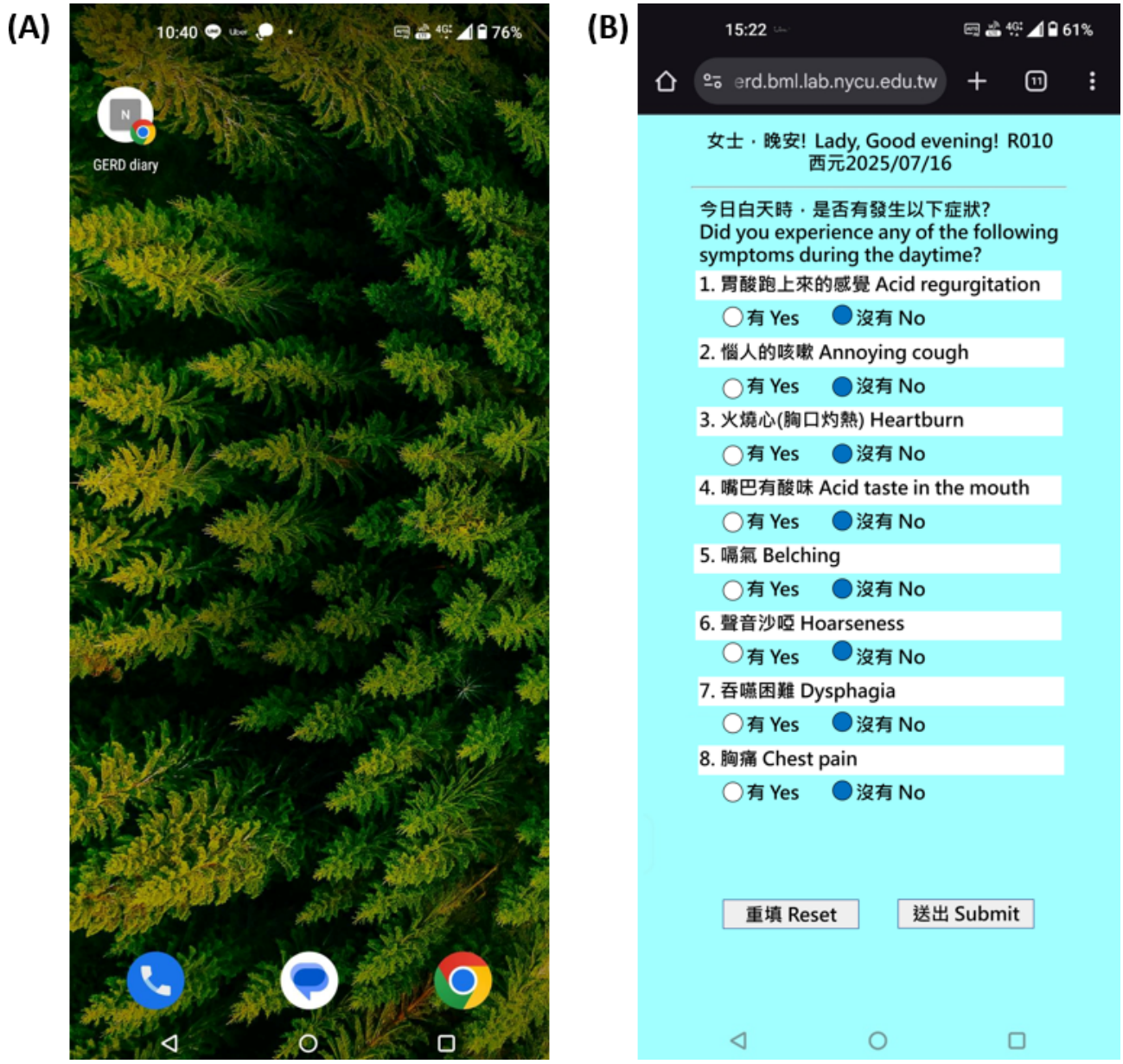

Optimization of the GERD E-Diary

Optimization of the e-diary system was developed with 3 stages. The first stage was initiated without reminders to the patients with GERD to complete the twice-daily questionnaire on mobile phones (October 7, 2021, to January 17, 2022; n=9). As most of these patients (n=5) failed to fill in the questionnaire (defined as completing <60% [range: 46.4% - 57.1%] of the overall symptom diary) in the first stage, we designed an SMS text messaging reminder in the second stage (January 18, 2022, to June 30, 2022; n=51). In this stage, the system would automatically check at noon and 10 PM each day to determine whether the patients had filled out the diary. Once an unfilled e-dairy was detected, an SMS text message was sent to the patients' mobile phones. If no entries were submitted for 3 consecutive days, an additional notification was sent to the research team to contact the patients. During the second stage, the reminder function was activated manually, resulting in no reminder function during the 3 to 5 days after enrollment. In the third stage of system optimization (July 1, 2022, to January 19, 2023; n=78), the reminder system was fully automated and activated immediately upon enrollment, eliminating the need for manual setup.

Adherence of the GERD E-Diary and Variables Affecting the Completeness of the E-Diary

The GERD e-diary adherence was evaluated by measuring the overall weekly symptom adherence rate, which was the number of days or nights symptoms filled out per week divided by 7. The overall symptom adherence rate was further categorized into nighttime (weekly nighttime adherence rate: number of nights filled out per week/7) and daytime (weekly daytime adherence rate: number of days filled out per week/7). The potential independent variables affecting the adherence rate, including the frequency of GERD symptoms (weekly total symptom days: 10 symptoms per day×7 days=70 symptom days,

XSL•FO
RenderX

ranging from 0 to 70) and the system optimization stage (3 stages, with higher stage indicating greater optimization), as well as the other confounders (such as age, gender, smoking, alcohol consumption, and comorbidities [sleep disorders, mental illnesses, sleep apnea, chronic obstructive pulmonary disease, asthma, liver or kidney diseases, cardiovascular diseases, diabetes mellitus, inflammatory bowel disease, central nervous system disorders, functional gastrointestinal disorders, and malignant diseases]) were measured.

Statistical Analysis

Descriptive statistics were performed for categorical variables as case numbers and percentages and means and SDs (including range) for continuous variables. The sampling method was clarified as convenience sampling of eligible outpatients. Any diary entry that was not completed on a given day was classified as nonadherence for that specific time point. As adherence was one of the key outcomes of interest, incomplete entries were interpreted as reflecting nonadherent behavior rather than ignorable missingness. Generalized estimating equations or multiple linear regression models were constructed to examine the effects of system optimization stage and weekly symptom frequency on adherence rates (nighttime, daytime, and overall). Independent variables included system optimization stage (categorical, 3 levels), symptom frequency (continuous), and potential confounders, such as age, gender, smoking, alcohol consumption, and major comorbidities. For repeated weekly adherence values across the 8-week observation period, a repeated-measures analysis with week as the within-subject variable was performed to assess temporal trends. All results are reported as β coefficients with corresponding 95% CIs and *P* values. Two-tailed significance was defined as $P<.05$.

Ethical Considerations

The study was reviewed and approved by the Medical Ethics Committee of Taipei Veterans General Hospital (2021-05-012CC) and conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all eligible participants before data collection, covering both the collection of primary data and its secondary analysis. Participants were informed that they could withdraw from the study at any time without any impact on their future medical care. To ensure privacy protection, all data were deidentified using participant codes, and no identifiable information was disclosed. Participants received a reimbursement of TWD $500 (approximately US $16) for their participation in the study.

Results

Clinical Characteristics of Patients With GERD

A total of 138 patients with GERD (mean [SD] age 52.9 [12.3] y; range 24.4 - 78.4 y; 68 female patients) were enrolled. Approximately one-third (46/138, 33.3%) of patients displayed coexisting functional gastrointestinal disorders. The baseline characteristics and comorbidities are summarized in Table 1.

Table . Baseline demographic and clinical characteristics of 138 patients with gastroesophageal reflux disease enrolled from the outpatient gastroenterology clinics at Taipei Veterans General Hospital, Taiwan (October 2021 to January 2023).

Clinical characteristics	Values
Age (y), mean (SD); range	52.9 (12.3); 24.4 - 78.4
Female, n (%)	68 (49.3)
Smoking, n (%)	14 (10.1)
Alcohol, n (%)	30 (21.7)
Sleep disorders, n (%)	72 (52.2)
Mental illnesses, n (%)	7 (5.1)
Sleep apnea, n (%)	17 (12.3)
Chronic obstructive pulmonary disease, n (%)	2 (1.4)
Asthma, n (%)	14 (10.1)
Liver or kidney diseases, n (%)	15 (10.9)
Cardiovascular diseases, n (%)	38 (27.5)
Diabetes mellitus, n (%)	11 (8.0)
Inflammatory bowel disease, n (%)	5 (3.6)
Central nervous system disorders, n (%)	2 (1.4)
Functional gastrointestinal disorders, n (%)	46 (33.3)
Malignant diseases, n (%)	16 (11.6)

System Optimization and Effects of Symptom Frequency on Adherence

As presented in Table 2, during the first stage of system optimization, the nighttime symptom adherence rate was 47.2%. Implementation of the reminder system significantly increased adherence by 12.5% in the second ($P=.005$) and 10.9% in the third stages ($P=.01$). For daytime symptom reporting, the adherence rate in the first stage was 40.0%, which improved by 21.7% in the second ($P<.001$) and 20.8% in the third stages

XSL•FO
RenderX

(*P*<.001). The overall adherence reporting in the first stage was 57.6%, with further increases of 16.5% in the second (*P*<.001) and 18.5% in the third stages (*P*<.001). For every additional symptom frequency, there was a significant increase of 0.9% in adherence for both daytime and nighttime symptom reporting and an increase of 0.7% for overall symptom reporting (*P*<.001). These findings indicate that higher symptom frequency was positively correlated with participants' motivation to complete the GERD e-diary, suggesting that patients experiencing more frequent symptoms were more engaged in self-monitoring.

Table . Effects of system optimization and symptom frequency on adherence to the gastroesophageal reflux disease e-diary.[a]

Adherence (%)	β[b,c] (95% CI)	*P* value
Nighttime symptom		
System optimization (first stage)	47.2	
System optimization (second vs first stage)	+12.5 (3.7 to 21.3)	0.005
System optimization (third vs first stage)	+10.9 (2.6 to 19.2)	0.01
Symptom frequency	+0.9 (0.7 to 1)	<.001
Daytime symptom		
System optimization (first stage)	40	
System optimization (second vs first stage)	+21.7 (14.2 to 29.2)	<.001
System optimization (third vs first stage)	+20.8 (13.7 to 27.9)	<.001
Symptom frequency	+0.9 (0.7 to 1)	<.001
Overall symptom		
System optimization (first stage)	57.6	
System optimization (second vs first stage)	+16.5 (9.8 to 23.2)	<.001
System optimization (third vs first stage)	+18.5 (12.2 to 4.8)	<.001
Symptom frequency	+0.7 (0.6 to 0.8)	<.001

[a]Results derived from multiple linear regression models adjusted for potential confounders, including age, gender, smoking, alcohol consumption, and comorbidities.

[b]β=change in adherence in the second and third stages of system optimization compared with the first stage.

[c]β=change in adherence per additional symptom frequency.

Adherence Trends in GERD E-Diary Recording

Table 3 presents the weekly change in adherence rate among all enrolled patients with GERD. The nighttime symptom adherence rate was 80.1% during the first week and began to decrease significantly from the third week (71.5%; *P*=.04), reaching 61.5% by the eighth week (*P*<.001). For daytime symptom reporting, adherence was 85.1% in the first week and showed a significant decline beginning at the fourth week (76.0%; *P*=.01) and further to 66.8% by the eighth week (*P*<.001). Overall adherence followed a similar trend, starting at 95.1% in the first week, decreasing to 86.7% in the fourth week (*P*=.009), and dropping to 78.0% in the eighth week (*P*<.001). These results demonstrate a gradual decline in participant engagement with the e-diary over time, despite the implemented reminder optimization strategies.

XSL•FO
RenderX

Table . Weekly adherence trends of gastroesophageal reflux disease e-diary completion during the 8-week study period.

Symptoms and week	Adherence (%)	β[a] (95% CI)	*P* value
Nighttime symptoms			
1	80.1	N/A[b]	N/A
2	72	−8.1 (−16.4 to 0.2)	0.06
3	71.5	−8.6 (−16.9 to −0.3)	0.04
4	67.9	−12.2 (−20.5 to −3.9)	0.004
5	68.4	−11.7 (−20 to −3.4)	0.006
6	61.6	−18.5 (−26.8 to −10.2)	<.001
7	64.3	−15.8 (−24.1 to −7.5)	<.001
8	61.5	−18.6 (−26.9 to −10.3)	<.001
Daytime symptoms			
1	85.1	N/A	N/A
2	82.6	−2.5 (−9.8 to 4.8)	0.50
3	80.5	−4.6 (11.8 to 2.7)	0.22
4	76	−9.1 (16.4 to 1.8)	0.014
5	72	−13 (−20.3 to -5.8)	<.001
6	69.5	−15.6 (−22.9 to -8.4)	<.001
7	68.4	−16.7 (23.9 to −9.4)	<.001
8	66.8	−18.3 (−25.6 to −11)	<.001
Overall symptoms			
1	95.1	N/A	N/A
2	89.9	−5.3 (−11.6 to 1)	0.10
3	88.9	−6.2 (−12.5 to 0.1)	0.05
4	86.7	−8.4 (−14.7 to −2.1)	0.009
5	83.4	−11.7 (−18 to −5.4)	<.001
6	78.3	−16.9 (−23.3 to −10.6)	<.001
7	79.6	−15.5 (−21.9 to −9.2)	<.001
8	78	−17.2 (−23.5 to −10.9)	<.001

[a]β=β values representing the percentage change in adherence compared with the first week, with 95% CIs and *P* values derived from repeated-measures analysis.

[b]N/A: not available.

Comparison of GERD E-Diary With Previous GERD Evaluation Tools

The characteristics and differences between our current GERD e-diary study and previous GERD questionnaire studies are summarized in Table 4. This e-diary, using separate day and night assessments, can record 10 relevant GERD symptoms daily and also provide educational information.

XSL•FO
RenderX

Table . Comparison of patient-reported outcome (PRO) instruments for gastroesophageal reflux disease (GERD). The summary includes data entry frequency, number of symptoms recorded, and whether daytime and nighttime symptoms were separately assessed.

PRO instrument[a]	GRACI[b] [31]	Puhan et al [32]	GerdQ[c] [16]	RESQ-7[d] [33]	GERD e-diary
Data entry frequency	Daily	Daily	Weekly recall	Weekly recall	Daily
Number of symptoms recorded	5	3	6	13	10
Day and night recording	No separation[e]	Separately assessed[f]	No separation	Not assessed[g]	Separately assessed
Recording format	Paper	Paper	Paper	Paper	Electronic diary
Providing educational information	N/A[h]	N/A[h]	N/A[h]	N/A[h]	Yes

[a]Summary includes data entry frequency, number of symptoms recorded, and whether daytime and nighttime symptoms were separately assessed.

[b]GRACI: Gastroesophageal Reflux Disease Activity Index.

[c]GerdQ: Gastroesophageal Reflux Disease Questionnaire.

[d]RESQ-7: Reflux Symptom Questionnaire, 7 day recall

[e]No separation: daytime and nighttime symptoms recorded together.

[f]Separately assessed: daytime and nighttime symptoms recorded separately.

[g]Not assessed: no specific daytime or nighttime symptom assessment.

[h]N/A: not available.

Discussion

Principal Findings

In this study, we demonstrated that a GERD e-diary was successfully developed to record GERD symptoms twice daily (day and night) with relatively high adherence. Patients' adherence was significantly improved with the application of SMS text messaging reminders in the e-diary system. In addition, patients who experienced more frequent symptoms tended to demonstrate higher levels of adherence. While the overall adherence rate reached approximately 80% over the 8-week period, a gradual decline was observed over time. This PRO-based e-diary, which enables prolonged and continuous symptom recording, provides a real-time and prospective assessment framework that substantially reduces the recall and ecological biases associated with conventional short-term, retrospective GERD symptom evaluations. By facilitating structured, longitudinal monitoring, the system not only enhances symptom awareness for patients but also offers clinicians high-resolution, long-term data that support more customized therapeutic decisions and advance personalized GERD management.

Comparison With Prior Work

PRO measures are essential for evaluating both disease burden and treatment efficacy in GERD [6]. Several instruments have been developed to capture these outcomes. For example, the GRACI (Gastroesophageal Reflux Disease Activity Index) integrates patient diaries with structured nurse-led interviews, thereby reducing physician workload and minimizing assessor bias [31]. Puhan et al [32] introduced a symptom diary that tracked heartburn frequency, severity, and antacid use over 4 to 6 weeks, achieving high adherence, with only 7.9% of patients completing fewer than 80% of entries. The GerdQ (Gastroesophageal Reflux Disease Questionnaire), a 6-item tool, facilitates diagnosis and management in primary care without the need for specialist referral [16]. Similarly, RESQ-7 (Reflux Symptom Questionnaire, 7-day recall) evaluates 13 GERD-related symptoms based on a 7-day recall period [33]. Despite their clinical value, most PRO tools have important limitations. They typically capture symptoms over short intervals (often 1 wk), which makes it difficult to assess temporal fluctuations in patients with intermittent symptoms. Relying on weekly recall instead of daily reporting also introduces bias, as patients tend to overestimate symptom intensity and underestimate frequency [34]. Daily PROs are generally more accurate, especially for variable symptoms [35], but even these prior studies fail to differentiate between daytime and nocturnal symptoms. This is a critical gap, as nighttime symptoms are common in GERD and strongly impact quality of life [36]. Additionally, many existing PROs record only a narrow range of symptoms, limiting their clinical comprehensiveness. Another concern lies in data collection methods. Traditional paper-based questionnaires are vulnerable to retrospective completion, which compromises accuracy. In contrast, electronic data entry systems can restrict both prospective and retrospective inputs, thereby reducing recall bias and improving data integrity [35]. To overcome these challenges, we developed a web-based GERD e-diary that records symptoms twice daily over an 8-week period. This approach provides a more reliable picture of symptom patterns, minimizes recall bias, and generates richer data for clinicians. Importantly, the e-diary also enhances patient awareness of their condition. Together, our GERD e-diary features improve the accuracy and comprehensiveness of GERD assessment, ultimately fostering greater confidence in management for both clinicians and patients.

In the first phase of building up the e-diary, the overall adherence rate was as low as approximately 40%. Therefore, we incorporated a reminder system into the e-diary and activated it upon the detection of missed entries. In the second stage, the reminder system became available only 3 days after enrollment, whereas in the third stage, it was activated immediately upon

XSL•FO
RenderX

enrollment. Both optimization measurements would significantly improve overall symptom adherence to approximately 80%. Previous research supported the effectiveness of reminders in enhancing health-related behaviors. For example, a review of 11 randomized controlled trials (1999 - 2009) confirmed that reminder interventions significantly improved daily medication adherence compared to no-reminder controls [37]. Another study demonstrated that daily SMS text messaging reminders enhanced adherence to antiasthmatic treatment [38]. Similarly, email and letter reminders significantly improved colorectal cancer screening rates compared to usual care, with no difference between the 2 reminder types (email and letter) [39]. Furthermore, reminding patients and clinicians, especially those directed at patients, is an effective strategy to improve colorectal cancer screening rates among individuals who are not up to date with screening [40]. In line with these findings, our study demonstrated that incorporating SMS text messaging reminders substantially improved adherence to GERD e-diary recordings. Despite the favorable results, there seemed to be little difference between the second and third stages of system optimization, which might be due to the short time delay (3 - 5 d) of the incorporation of the reminding systems.

We also found that higher symptom frequency was associated with increased adherence. Similar observations had been reported in other diseases. For instance, patients with urinary incontinence who experienced greater voiding frequency exhibited higher adherence to voiding diaries [27]. Furthermore, in patients with seasonal allergic rhinitis, symptom recording adherence to an e-diary was significantly higher during the peak grass pollen season, which coincided with more intense allergic symptoms [28]. All observations suggested that more frequent and severe symptoms may be associated with enhanced awareness and stronger motivation to report their disease status to health care providers.

Despite relatively high adherence at the beginning phase, our study showed a gradual decline in adherence over the 8-week period. This trend was consistent with previous findings in other diseases. For example, patients with allergic rhinitis demonstrated a slow decline in e-diary completion from 90% in the first week, 80% to 90% in the second to sixth weeks, and 70% to 80% after the seventh week [28]. A separate study assessing voiding diaries among patients with urinary incontinence similarly reported that a 7-day diary posed a higher burden than shorter 2- or 3-day formats [41]. In addition, 3 respiratory clinical trials also demonstrated that adherence decreased over time after randomization [42]. All results suggested that adherence to long-term daily symptom recording might be challenging for the patients to complete the daily diary study.

To improve long-term adherence, several strategies might be considered. Providing additional information and education could be applied to promote better adherence in e-diary recording [28]. Reducing diary duration may also enhance patient compliance and minimize burden [41]. Additional methods to increase patient engagement, such as customizing diary content and reminders based on patient needs, using user-friendly interfaces, and incorporating social and gamification features [21], or integration with wearable devices, might also help improve adherence. Using personalized adaptive reinforcement learning as a core behavioral strategy, it may be possible to optimize intervention timing and minimize alert fatigue, ensuring more stable long-term user participation in the future [43]. In this study, patients expressed a desire to see immediate visual representations of daily symptom changes. Understanding the psychological factors (eg, motivation) and socioeconomic status behind patient adherence could be invaluable for designing more effective interventions.

Study Strengths

Our study introduces a significant innovation in GERD management through the development of a web-based e-diary designed for daily symptom assessment. The primary strength of this approach lies in its ability to provide real-time, prospective data collection, thereby fundamentally overcoming the limitations of traditional paper-based questionnaires and retrospective recall, which are prone to significant recall and ecological biases for patients with GERD [44]. Unlike conventional studies that often rely on symptom scores primarily as a measure of drug response, our e-diary integrates symptom monitoring directly into daily clinical practice. This shift allows for the capture of fine-grained, fluctuating symptom patterns over an extended period, providing a more accurate and comprehensive understanding of the patient's condition. Second, by minimizing recall bias, our e-diary ensures higher data entry quality and more efficient data handling compared to traditional methods. Third, it empowers patients to actively participate in their self-management by collecting health data autonomously, fostering self-reliance and improved awareness of their condition [45]. Fourth, with the application of an SMS reminder system and educational feedback provided, the adherence rate of patients could be maintained up to 8 weeks, especially for those symptomatic patients. Fifth, our study benefits from the e-diary's ability to separately capture daytime and nighttime GERD symptoms. This distinction is clinically important, as nocturnal reflux is characterized by impaired esophageal clearance and is associated with more aggressive disease, including a higher risk of severe complications such as esophagitis [46]. Moreover, nighttime symptoms substantially impair patients' health-related quality of life [36]. These granular, time-specific symptom data provide essential insights for developing personalized management strategies. Finally, the detailed, real-time symptom data facilitate enhanced patient-physician collaboration. Physicians can readily visualize symptom changes, understand the variability in symptoms, and work more effectively with patients to tailor treatment strategies [47]. This paves the way for truly personalized treatment plans in the future that can adapt to individual patient needs and daily life events, potentially incorporating dietary or other therapeutic adjustments. Ultimately, this approach moves beyond simple response assessment, enabling a proactive feedback loop that can lead to more informed and timely clinical decisions, fostering greater confidence in management for both clinicians and patients.

Study Limitations

Limitations do exist in the study. First, patients with GERD were enrolled from a single center and included only patients capable of operating an e-diary, potentially limiting the

XSL•FO
RenderX

generalizability and overestimating adherence. Validation in multicenter or community-based cohorts should be conducted in the further studies. Second, days on which patients did not complete entries were recorded as symptom-free when calculating symptom frequency, which may lead to underestimation of the reported symptoms. Third, objective physiological parameters (eg, 24 h pH-impedance monitoring, reflux-symptom association probability, or treatment response) were not detected in this study, which would enhance the e-diary's clinical utility. Fourth, although SMS reminders improved the adherence, they might inadvertently increase psychological stress, symptom vigilance, or anxiety, which also may lead to worsening of GERD symptoms. This may account for part of the nonadherence observed in this study. Assessment of the Esophageal Hypervigilance and Anxiety Scale and other psychological measurements is warranted in subsequent studies to help balance engagement benefits against potential harm. Finally, we did not evaluate the satisfaction scores of patients with GERD and caring physicians. Further adjustments to this GERD e-diary by correcting the aforementioned limitations should be considered in future studies.

Conclusions

Our results demonstrate that system optimization can significantly enhance adherence in the newly developed GERD e-diary recording. Increased GERD symptom frequency was associated with adherence, although overall engagement declined gradually over 8 weeks. The development of this PRO-based GERD e-diary system can be a convenient tool for future application in clinical and research settings.

Acknowledgments

The authors acknowledge the Big Data of Taipei Veterans General Hospital and the Biostatistics Task Force of Taipei Veterans General Hospital for their assistance during this study.

The authors declare the use of generative AI in the writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: summarizing text and translation. The GAI tool used was ChatGPT 4. Responsibility for the final manuscript lies entirely with the authors. All AI-generated text was reviewed or revised by the authors before submission. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was partially supported by Taiwan National Science and Technology Council (NSTC 113-2634-F-A49-003). The funding agency was not involved in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: YCC, YPW, LFC, CLL, MLP, YHP, DWW, SLW
Data curation: YCC, SLW
Formal analysis: SLW
Funding acquisition: LFC, CLL, MLP, DWW
Investigation: YPW, CLL, YCC
Methodology: MLP, JHH
Project administration: CLL, MLP, YCC
Resources: YPW, CLL, MLP, LFC, YHP, DWW
Software: MLP, JHH, LFC
Supervision: CLL, LFC, YHP, MLP, DWW, SLW
Validation: YCC, CLL, YPW
Visualization Writing – original draft: YCC, YPW, CLL
Writing – review & editing: YCC, YPW, CLL, MLP, YHP, LFC, SLW
MLP and CLL are co-corresponding authors.

Conflicts of Interest

None declared.

Checklist 1
STROBE checklist.
[PDF File, 92 KB - jmir_v28i1e83680_app1.pdf]

References

XSL•FO
RenderX

1. Vakil N, van Zanten SV, Kahrilas P, Dent J, Jones R, Global Consensus Group. The Montreal definition and classification of gastroesophageal reflux disease: a global evidence-based consensus. Am J Gastroenterol 2006 Aug;101(8):1900-1920. [doi: 10.1111/j.1572-0241.2006.00630.x] [Medline: 16928254]
2. Blondeau K, Sifrim D, Dupont L, Tack J. Reflux cough. Curr Gastroenterol Rep 2008 Jun;10(3):235-239. [doi: 10.1007/s11894-008-0049-0] [Medline: 18625132]
3. Kellerman R, Kintanar T. Gastroesophageal reflux disease. Prim Care 2017 Dec;44(4):561-573. [doi: 10.1016/j.pop.2017.07.001] [Medline: 29132520]
4. Mahajan R, Kulkarni R, Stoopler ET. Gastroesophageal reflux disease and oral health: a narrative review. Spec Care Dentist 2022 Nov;42(6):555-564. [doi: 10.1111/scd.12726] [Medline: 35605234]
5. de Hoyos A, Esparza EA, Cervantes-Sodi M. Non-erosive reflux disease manifested exclusively by protracted hiccups. J Neurogastroenterol Motil 2010 Oct;16(4):424-427. [doi: 10.5056/jnm.2010.16.4.424] [Medline: 21103425]
6. Lu CL, Lang HC, Chang FY, et al. Social and medical impact, sleep quality and the pharmaceutical costs of heartburn in Taiwan. Aliment Pharmacol Ther 2005 Oct 15;22(8):739-747. [doi: 10.1111/j.1365-2036.2005.02664.x] [Medline: 16197495]
7. Delshad SD, Almario CV, Chey WD, Spiegel BMR. Prevalence of gastroesophageal reflux disease and proton pump inhibitor-refractory symptoms. Gastroenterology 2020 Apr;158(5):1250-1261. [doi: 10.1053/j.gastro.2019.12.014] [Medline: 31866243]
8. Guidance for industry patient-reported outcome measures: use in medical product development to support labeling claims. : U.S. Department of Health and Human Services; 2009 URL: https://www.fda.gov/media/77832/download [accessed 2025-12-19]
9. Rasheed T, Zuberi BF, Ali FS, Hussain SM, Kumar P, Saleem A. Comparison of efficacy of daily and alternate day maintenance treatment of GERD with Vonoprazan 10-mg using Gastroesophageal Reflux Disease Symptom Assessment Scale. Pak J Med Sci 2024;40(4):623-628. [doi: 10.12669/pjms.40.4.8063] [Medline: 38545007]
10. Miyamoto M, Manabe N, Haruma K. Efficacy of the addition of prokinetics for proton pump inhibitor (PPI) resistant non-erosive reflux disease (NERD) patients: significance of frequency scale for the symptom of GERD (FSSG) on decision of treatment strategy. Intern Med 2010;49(15):1469-1476. [doi: 10.2169/internalmedicine.49.3615] [Medline: 20686276]
11. Shaw M, Dent J, Beebe T, et al. The Reflux Disease Questionnaire: a measure for assessment of treatment response in clinical trials. Health Qual Life Outcomes 2008 Apr 30;6:31. [doi: 10.1186/1477-7525-6-31] [Medline: 18447946]
12. Wu H, Ungerleider S, Campbell M, et al. Patient-reported outcomes in 645 patients after laparoscopic fundoplication up to 10 years. Surgery 2023 Mar;173(3):710-717. [doi: 10.1016/j.surg.2022.07.039] [Medline: 36307333]
13. Galvarini M, Angeramo CA, Kerman J, et al. Outcomes of endoscopic antireflux mucosectomy for the treatment of gastroesophageal reflux disease: systematic review and meta-analysis. J Clin Gastroenterol 2024 Oct 1;58(9):851-856. [doi: 10.1097/MCG.0000000000002061] [Medline: 39145822]
14. Fuchs KH, Musial F, Eypasch E, Meining A. Gastrointestinal quality of life in gastroesophageal reflux disease: a systematic review. Digestion 2022;103(4):253-260. [doi: 10.1159/000524766] [Medline: 35605592]
15. Flynn KE, Mansfield SA, Smith AR, et al. Patient demographic and psychosocial characteristics associated with 30-day recall of self-reported lower urinary tract symptoms. Neurourol Urodyn 2020 Sep;39(7):1939-1948. [doi: 10.1002/nau.24461] [Medline: 32856723]
16. Jones R, Junghard O, Dent J, et al. Development of the GerdQ, a tool for the diagnosis and management of gastro-oesophageal reflux disease in primary care. Aliment Pharmacol Ther 2009 Nov 15;30(10):1030-1038. [doi: 10.1111/j.1365-2036.2009.04142.x] [Medline: 19737151]
17. Vakil N, Niklasson A, Denison H, Rydén A. Gender differences in symptoms in partial responders to proton pump inhibitors for gastro-oesophageal reflux disease. United European Gastroenterol J 2015 Oct;3(5):443-452. [doi: 10.1177/2050640614558343] [Medline: 26535123]
18. Shaw MJ, Talley NJ, Beebe TJ, et al. Initial validation of a diagnostic questionnaire for gastroesophageal reflux disease. Am J Gastroenterol 2001 Jan;96(1):52-57. [doi: 10.1111/j.1572-0241.2001.03451.x] [Medline: 11197287]
19. Wiest IC, Sicorello M, Yesmembetov K, Ebert MP, Teufel A. Usage behaviour and adoption criteria for mobile health solutions in patients with chronic diseases in gastroenterology. Visc Med 2024 Apr;40(2):61-74. [doi: 10.1159/000534191] [Medline: 38584857]
20. Giebel GD, Speckemeier C, Abels C, et al. Problems and barriers related to the use of digital health applications: scoping review. J Med Internet Res 2023 May 12;25:e43808. [doi: 10.2196/43808] [Medline: 37171838]
21. Jakob R, Harperink S, Rudolf AM, et al. Factors influencing adherence to mhealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371. [doi: 10.2196/35371] [Medline: 35612886]
22. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020 Sep 29;22(9):e20283. [doi: 10.2196/20283] [Medline: 32990635]
23. Daniëls NEM, Hochstenbach LMJ, van Zelst C, van Bokhoven MA, Delespaul P, Beurskens A. Factors that influence the use of electronic diaries in health care: scoping review. JMIR mHealth uHealth 2021 Jun 1;9(6):e19536. [doi: 10.2196/19536] [Medline: 34061036]

XSL•FO
RenderX

24. Messner EM, Sturm N, Terhorst Y, et al. Mobile apps for the management of gastrointestinal diseases: systematic search and evaluation within app stores. J Med Internet Res 2022 Oct 5;24(10):e37497. [doi: 10.2196/37497] [Medline: 36197717]
25. Gould MJ, Lin C, Walsh CM. A systematic assessment of the quality of smartphone applications for gastroesophageal reflux disease. Gastro Hep Adv 2023;2(5):733-742. [doi: 10.1016/j.gastha.2023.03.001] [Medline: 39129878]
26. Venugopal LS, Musbahi A, Shanmugam V, Gopinath B. A systematic review of smartphone apps for gastro-oesophageal reflux disease: the need for regulation and medical professional involvement. mHealth 2021;7:56. [doi: 10.21037/mhealth-20-126] [Medline: 34805387]
27. Pauls RN, Hanson E, Crisp CC. Voiding diaries: adherence in the clinical setting. Int Urogynecol J 2015 Jan;26(1):91-97. [doi: 10.1007/s00192-014-2470-2] [Medline: 25124091]
28. Di Fraia M, Tripodi S, Arasi S, et al. Adherence to prescribed e-diary recording by patients with seasonal allergic rhinitis: observational study. J Med Internet Res 2020 Mar 16;22(3):e16642. [doi: 10.2196/16642] [Medline: 32175909]
29. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med 2007 Oct 16;4(10):e296. [doi: 10.1371/journal.pmed.0040296] [Medline: 17941714]
30. Andrae DA, Hanlon J, Cala ML, et al. Evaluation and validation of the modified reflux symptom questionnaire-electronic diary in patients with persistent gastroesophageal reflux disease. Clin Transl Gastroenterol 2020 Jan;11(1):e00117. [doi: 10.14309/ctg.0000000000000117] [Medline: 31977454]
31. Williford WO, Krol WF, Spechler SJ. Development for and results of the use of a gastroesophageal reflux disease activity index as an outcome variable in a clinical trial. Control Clin Trials 1994 Oct;15(5):335-348. [doi: 10.1016/0197-2456(94)90031-0] [Medline: 8001355]
32. Puhan MA, Guyatt GH, Armstrong D, et al. Validation of a symptom diary for patients with gastro-oesophageal reflux disease. Aliment Pharmacol Ther 2006 Feb 15;23(4):531-541. [doi: 10.1111/j.1365-2036.2006.02775.x] [Medline: 16441474]
33. Rydén A, Denison H, Karlsson M, Vakil N. Development and validation of a patient-reported outcome instrument in partial responders to proton pump inhibitors. Scand J Gastroenterol 2013 Sep;48(9):1018-1026. [doi: 10.3109/00365521.2013.822544] [Medline: 23919738]
34. Rydén A, Leavy OC, Halling K, Stone AA. Comparison of daily versus weekly recording of gastroesophageal reflux disease symptoms in patients with a partial response to proton pump inhibitor therapy. Value Health 2016;19(6):829-833. [doi: 10.1016/j.jval.2016.05.007] [Medline: 27712711]
35. Vakil N, Björck K, Denison H, et al. Validation of the reflux symptom questionnaire electronic diary in partial responders to proton pump inhibitor therapy. Clin Transl Gastroenterol 2012 Jan 26;3(1):e7. [doi: 10.1038/ctg.2012.1] [Medline: 23238029]
36. Farup C, Kleinman L, Sloan S, et al. The impact of nocturnal symptoms associated with gastroesophageal reflux disease on health-related quality of life. Arch Intern Med 2001 Jan 8;161(1):45-52. [doi: 10.1001/archinte.161.1.45] [Medline: 11146697]
37. Fenerty SD, West C, Davis SA, Kaplan SG, Feldman SR. The effect of reminder systems on patients' adherence to treatment. Patient Prefer Adherence 2012;6:127-135. [doi: 10.2147/PPA.S26314] [Medline: 22379363]
38. Strandbygaard U, Thomsen SF, Backer V. A daily SMS reminder increases adherence to asthma treatment: a three-month follow-up study. Respir Med 2010 Feb;104(2):166-171. [doi: 10.1016/j.rmed.2009.10.003] [Medline: 19854632]
39. Muller D, Logan J, Dorr D, Mosen D. The effectiveness of a secure email reminder system for colorectal cancer screening. AMIA Annu Symp Proc 2009 Nov 14;2009:457-461. [Medline: 20351899]
40. Ahmed AM, Bacchus MW, Beal SG, et al. Colorectal cancer screening completion by patients due or overdue for screening after reminders: a retrospective study. BMC Cancer 2023 May 1;23(1):391. [doi: 10.1186/s12885-023-10837-y] [Medline: 37127588]
41. Ku JH, Jeong IG, Lim DJ, Byun SS, Paick JS, Oh SJ. Voiding diary for the evaluation of urinary incontinence and lower urinary tract symptoms: prospective assessment of patient compliance and burden. Neurourol Urodyn 2004;23(4):331-335. [doi: 10.1002/nau.20027] [Medline: 15227650]
42. Nowojewski A, Bark E, Shih VH, Dearden R. Patient adherence and response time in electronic patient-reported outcomes: insights from three longitudinal clinical trials. Qual Life Res 2024 Jun;33(6):1691-1706. [doi: 10.1007/s11136-024-03644-w] [Medline: 38598132]
43. Lauffenburger JC, Yom-Tov E, Keller PA, et al. The impact of using reinforcement learning to personalize communication on medication adherence: findings from the REINFORCE trial. NPJ Digit Med 2024 Feb 19;7(1):39. [doi: 10.1038/s41746-024-01028-5] [Medline: 38374424]
44. Beckers AB, Snijkers JTW, Weerts Z, et al. Digital instruments for reporting of gastrointestinal symptoms in clinical trials: comparison of end-of-day diaries versus the experience sampling method. JMIR Form Res 2021 Nov 24;5(11):e31678. [doi: 10.2196/31678] [Medline: 34821561]
45. Van Woensel W, Roy PC, Abidi SR, Abidi SSR. A mobile and intelligent patient diary for chronic disease self-management. Stud Health Technol Inform 2015;216:118-122. [Medline: 26262022]

XSL•FO
RenderX

46. Huang Y, Liu J, Xu L, et al. Exacerbation of symptoms, nocturnal acid reflux, and impaired autonomic function are associated with sleep disturbance in gastroesophageal reflux disease patients. Front Med (Lausanne) 2024;11:1438698. [doi: 10.3389/fmed.2024.1438698] [Medline: 39234038]
47. Staehelin D, Dolata M, Stöckli L, Schwabe G. How patient-generated data enhance patient-provider communication in chronic care: field study in design science research. JMIR Med Inform 2024 Sep 10;12:e57406. [doi: 10.2196/57406] [Medline: 39255481]

Abbreviations

e-diary: electronic diary
EQUATOR: Enhancing the Quality and Transparency of Health Research
GERD: gastroesophageal reflux disease
PRO: patient-reported outcome
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by S Brini; submitted 29.Sep.2025; peer-reviewed by MW Wong, Y Gao; accepted 24.Nov.2025; published 06.Jan.2026.

Please cite as:
Chen YC, Wang YP, Hung JH, Wang DW, Wu SL, Chen LF, Ping YH, Pan ML, Lu CL
Establishment and Optimization of a Patient-Reported Outcome–Based Electronic-Diary for Symptoms Evaluation in Patients With Gastroesophageal Reflux Disorder: Prospective Cohort Study
J Med Internet Res 2026;28:e83680
URL: https://www.jmir.org/2026/1/e83680
doi:10.2196/83680

© Yun-Chun Chen, Yen-Po Wang, Jui-Hsuan Hung, Da-Wei Wang, Shang-Liang Wu, Li-Fen Chen, Yueh-Hsin Ping, Mei-Lien Pan, Ching-Liang Lu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Quality of Life Trajectories With Integration Into Electronic Health Records for High-Resolution Patient Outcomes: Algorithm Development and Validation Study

Martin Dugas[1], MSc, MD; Robin Fleige[1], MSc; Max Christian Blumenstock[1], MSc; Stephan Christoph Feder[1,2], MD; Tobias Dittrich[1,3], MD; Niels Siegel[4], MD; Celine Fabienne Bergmann[1]; Luis Wettach[1], MSc; Sandra Sauer[3], MD; Pavlina Lenga[1,5], MD; Sandro M Krieg[5], MD, MBA; Susanne Dugas-Breit[6], MD; Lucy Joanne Kessler[7], MD; Kosima Kosmalla[1,8], MD; Fleur Fritz-Kebede[1], PhD; Martin Loos[4], MD; Hans-Christoph Friederich[2], MD; Thomas M Pausch[1,4], MD; Matthias Ganzinger[1], PhD

[1]Institute of Medical Informatics, Heidelberg University Hospital, Heidelberg, Germany
[2]Department of General Internal Medicine, Psychosomatics and Psychotherapy, Heidelberg University Hospital, Heidelberg, Germany
[3]Department of Hematology, Oncology and Rheumatology, Heidelberg University Hospital, Heidelberg, Germany
[4]Department of General, Visceral and Transplantation Surgery, Heidelberg University Hospital, Heidelberg, Germany
[5]Department of Neurosurgery, Heidelberg University Hospital, Heidelberg, Germany
[6]National Center for Tumor Diseases, Heidelberg University Hospital, Heidelberg, Germany
[7]Department of Ophthalmology, Heidelberg University Hospital, Heidelberg, Germany
[8]Department of Gastroenterology, Heidelberg University Hospital, Heidelberg, Germany

Corresponding Author:
Martin Dugas, MSc, MD
Institute of Medical Informatics
Heidelberg University Hospital
Im Neuenheimer Feld 130.3
Heidelberg, D-69120
Germany
Phone: 49 62215622000
Email: martin.dugas@med.uni-heidelberg.de

Abstract

Background: Patient-reported outcome measures (PROMs) such as health-related quality of life (HRQoL) are usually assessed at greater time intervals such as diagnostic time points, after treatment, and during follow-up. Many PROMs require frequent data collection (weekly or daily). Electronic PROMs enable high-resolution tracking but face declining response rates. Integrating PROMs into electronic health records (EHRs) could improve response rates and personalize therapy.

Objective: This study aimed to evaluate the technical and clinical feasibility of high-frequency HRQoL assessments for routine care in EHRs.

Methods: Patients receive emails on their mobile devices with 1-time links to a web-based app called MyEDC. This app communicates with an electronic data capture proxy in the demilitarized zone of the hospital. With a polling mechanism, these patient data are transferred to the protected hospital network and uploaded to the EHR system. HRQoL on a visual analog scale is assessed over the course of treatment in 4 clinical use cases: psychosomatics, hematology, visceral surgery, and neurosurgery.

Results: Quality of life (QoL) trajectories were collected for 110 patients with daily or weekly data collection between 2 weeks and 3 months. The HRQoL analyses revealed clinically relevant findings across the 4 different medical domains. In the use case psychosomatics, 36 patients showed a significant increase in HRQoL following 4 weeks of therapy, rising from a median of 42% (IQR 32%-52%) to 60% (IQR 41%-67%; P=.01). An analysis of 25 patients in hematology demonstrated a significant correlation between HRQoL and 30-item QoL Questionnaire (EORTC QLQ-C30) global health status score (P=.02). For 26 patients in visceral surgery, a significant association was observed between HRQoL and the reported pain level (P<.001). The clinical feasibility was further highlighted in the neurosurgery use case, where 23 patients showed a median response time to the electronic PROM questionnaires of 5.3 (IQR 0.6-17.7) hours. HRQoL values were associated with disease-specific symptoms and scores, indicating clinical validity of this readout. Considerable variability of HRQoL was observed over time, both intraindividually

XSL•FO
RenderX

and interindividually. Median area under the curve of HRQoL ranged from 0.46 to 0.79. Median time to answer ranged from 0.9 to 7.1 hours. No significant association between number of responses and age was observed.

Conclusions: High-resolution QoL trajectories with EHR integration are technically and clinically feasible. They offer a novel readout beyond survival analysis or PROM end point, enabling precise disease characterization and treatment comparison.

(J Med Internet Res 2026;28:e79834) doi:10.2196/79834

KEYWORDS

quality of life; electronic health records; patient-reported outcome measures; mobile devices; feasibility study

Introduction

At present, patient-reported outcome measures (PROMs) are typically assessed at the time of diagnosis, after initial treatment and every 3 to 6 months during follow-up [1]. However, these time intervals for PROM assessment and the length of the follow-up period are often based on clinical feasibility rather than standardized guidelines. Many established PROMs address a time interval of 1 to 2 weeks or less, for example, the 5-item World Health Organization (WHO-5) well-being index [2], 9-item Patient Health Questionnaire [3], and patient-reported outcomes-common terminology criteria for adverse events (PRO-CTCAE) [4]. Health-related quality of life (HRQoL) is a typical example of a PROM and can be assessed with instruments such as the EQ-5D [5], which addresses today's status. Therefore, and for clinical necessity, complete PROM trajectories require more frequent data collection, depending on the instrument every 2 weeks, weekly, or even daily. Electronic collection of PROMs provides new opportunities for such frequent assessments. The time course of PROMs can be determined with this technology; for example, a recovery trajectory for cardiac surgery patients was described in the study by Mori et al [6]. However, response rate (RR) is a major issue: a review of RRs over time [7] reported an average RR for electronic PROMs of 42% at baseline, decreasing to 36% after 1 year. It is assumed that integration of digital health tools into routine care can contribute to long-term engagement and higher retention rates [8]. Current methods often rely on stand-alone apps that store PROMs in external databases, creating data silos that impede real-time access and clinician use. Integrating PROMs directly into the electronic health records (EHRs) overcomes these limitations, offering physicians a novel, comprehensive clinical readout for better-informed decisions and potentially fostering an improved patient-physician relationship.

Our hypothesis is that the integration of HRQoL from mobile devices into EHRs can improve RRs and enable the collection of HRQoL trajectories with high resolution, that is, weekly or even daily. Patients will show a higher commitment to provide HRQoL data continuously because these are sent to the medical team and thereby have a direct effect on therapy for each individual patient. From an informatics perspective, this is challenging because such data need to be collected and transferred from a large variety of mobile devices on the internet into the protected network of an EHR system. The objective of this work was to assess the technical and clinical feasibility of high-resolution HRQoL assessments, which can visualize the course of a disease from a patient perspective and are integrated with EHRs.

Methods

IT Architecture and Information Security

To capture QoL trajectories (QoL-T) with high resolution, that is, biweekly, weekly, or even daily, each patient used their own mobile device for data collection. The alternative would be a preinstalled device provided by the health care organization, which is associated with significant cost and logistical effort and therefore cannot provide a scalable system. Patient devices were very heterogeneous: smartphone or tablet with iOS or Android, laptop or PC, and different operating systems with different software versions. Therefore, a web-based app (HTML5) was developed for data collection, called MyEDC (my electronic data capture), which is based on OpenEDC [9,10]. This approach avoids software installations by the patient.

Figure 1 presents the high-level IT architecture for data transfer from the patient device to the EHR. This architecture was reviewed and approved by the information security officer of Heidelberg University Hospital (UKHD). The web app was provided by an electronic data capture (EDC) proxy in the demilitarized zone, which is protected by 3 firewalls and a reverse proxy. In the internal network, a job service (programmed in Java with Spring framework) sent emails with 1-time EDC links to the patient. The job service polled data from outside to the internal network but only for registered and valid data collections. Other data packages were ignored and deleted. Valid data were transferred to an internal EDC database, which converted data into EHR format and performed an automated upload to the EHR database. A Clinical Data Interchange Standards Consortium (CDISC) operational data model [11]–compliant EDC system was used to comply with data standards in clinical trials. Owing to technical limitations of the local EHR system (i.s.h.med from Cerner Inc), files in PDF with informative document titles were generated and uploaded into the EHR.

XSL•FO
RenderX

Figure 1. IT architecture for data transfer from patient device through hospital firewalls into internal electronic health record system. DMZ: demilitarized zone; EDC: electronic data capture; EHR: electronic health record; PC: personal computer.

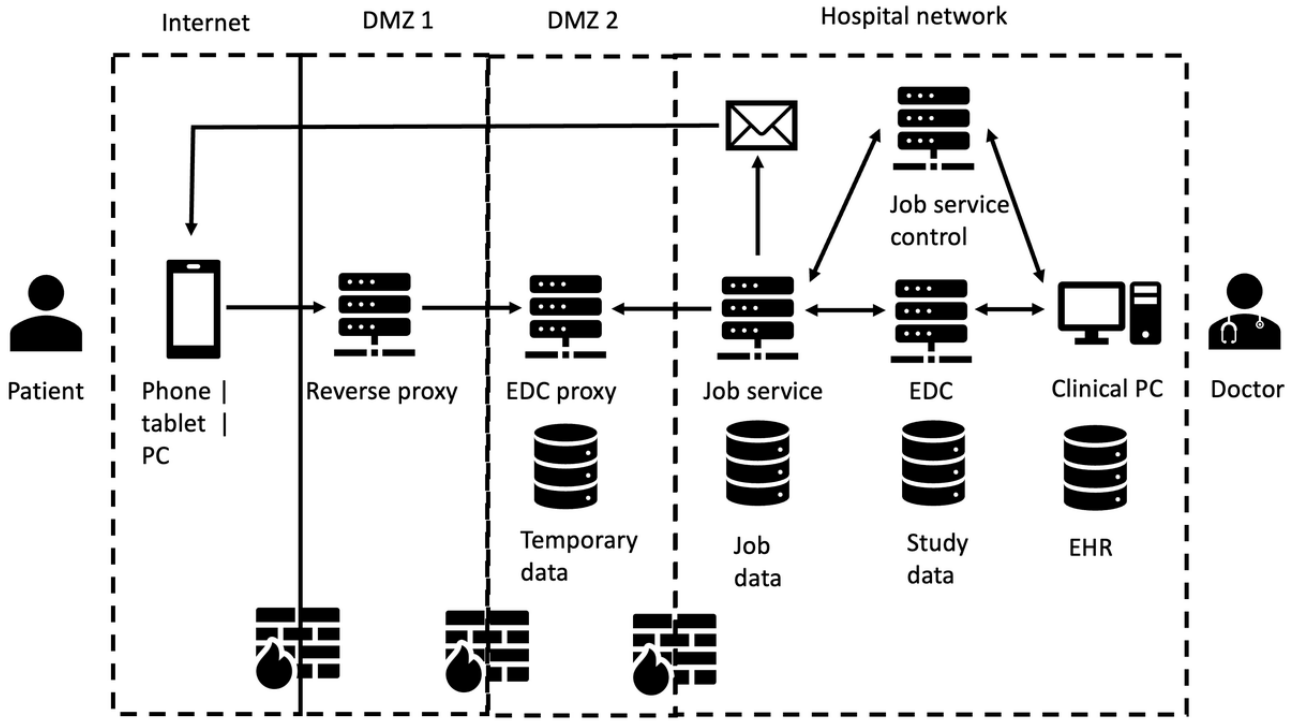

Short Questionnaires for High-Frequency PROM Collection

Overall, HRQoL was assessed frequently using a visual analog scale (VAS) with a range from 1 to 100 (Figure 2), similar to HRQoL instruments such as EQ-5D. An HRQoL value of 0 was reserved for deceased patients. QoL-T, that is, a time series of HRQoL values, were plotted and summarized with a normalized area under the curve (AUC) per patient. Few (mostly <5) additional, disease-specific PROM items were added for each clinical use case, for example, pain in surgical patients. Simple language was used to support a wide variety of use cases, including severely ill patients and participants with limited language skills. These questionnaires were designed by clinical experts to assess patient trajectories regarding effectiveness and side effects of routine therapy.

Figure 2. Overall health-related quality of life with a 100-level visual analog scale. General questions: How do You feel? (1=very bad, 100=very good).

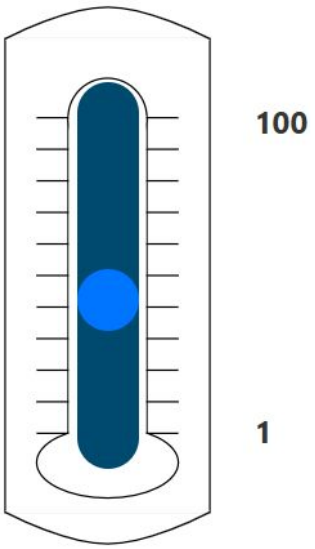

Statistical Analysis

Statistical analyses were performed using R (version 4.4.0). Differences between paired observations were assessed using the Wilcoxon signed-rank test. Associations between continuous or ordinal variables were investigated using the Spearman rank correlation coefficient. For all tests, a 2-sided *P* value threshold of .05 was considered statistically significant. AUC was calculated with the trapezoidal rule using the mean of adjacent points, normalized by total time, and scaled to a value range from 0 to 1. A 95% confidence band for average QoL was calculated as a point-wise 95% CI of the mean. The participation

XSL•FO
RenderX

rate was defined as the percentage of eligible patients who consented to participate in the study out of the total number of individuals invited. To assess longitudinal data quality, we calculated the RR at each assessment time point, defined as the percentage of participating patients who completed the questionnaire at that specific interval. Time to answer was calculated for completed questionnaires. Clinical feasibility for QoL-T was assessed regarding completeness with the median RR across all time points. Timeliness of patient answers to short PROM questionnaires was described with median time to answer.

Ethical Considerations

This study was conducted in accordance with the ethical guidelines of the Declaration of Helsinki. The research was approved by the ethics committee of the Medical Faculty of Heidelberg University (S-607/2023 and S-255/2025). The privacy and confidentiality of participants were maintained by ensuring that all study data were deidentified. Informed consent was obtained from all participants; this included the explicit right to opt out or withdraw from the study at any time. No direct compensation was provided for participation.

Clinical Use Cases

Overview

Four different use cases were selected to assess the clinical feasibility of QoL-T with EHR integration in psychosocial and somatic medicine. PROMs are clinically relevant across these distinct disciplines because they capture the patient's subjective experience of disease burden, treatment efficacy, and functional recovery. PROMs can provide insights into mental well-being (psychosomatics), treatment toxicity and fatigue (hematology), postsurgical functional status (visceral surgery), and pain (neurosurgery). Overall, 110 out of 184 (59.7%) patients agreed to participate.

Psychosomatics

As a typical example for psychosocial medicine, 36 inpatients with anorexia nervosa (Department of General Internal Medicine and Psychosomatics, UKHD) used the system with their own devices (predominantly smartphones) between August 2024 and April 2025. Three additional patients did not participate because they did not have a suitable device or did not want to use it for this purpose. Data were collected once a week for at least a month per patient. In addition to HRQoL-VAS, further widely used psychometric instruments to assess depression and anxiety (4-item Patient Health Questionnaire, PHQ-4 [12]), somatic symptoms (12-item Somatic System Disorder, SSD-12 [13]), short symptom checklist (Symptom-Checkliste Kurzversion, SCLK-11 [14]), work ability (Work Ability Index [15]), reflective functioning (Reflective Functioning Questionnaire, RFQ [16]), and 8-item Eating Disorder Examination-Questionnaire (EDE-Q8) [17] were applied. EDE-Q8 is an established clinical score to track symptoms of anorexia nervosa.

Hematology

A total of 25 outpatients with myeloma (Department of Hematology, Oncology, and Rheumatology, UKHD) used the system with their own devices between September 2024 and March 2025. In addition, 9 patients did not participate (technical reasons: 6, overwhelmed by medical situation: 2, and other: 1). Data were collected once a day for a period of 2 weeks during chemotherapy. In addition to HRQoL-VAS, side effects were collected according to PRO-CTCAE [18].

Visceral Surgery

In total, 26 inpatients (Department of General, Visceral, and Transplantation Surgery, UKHD) used the system with their own devices between January 2025 and April 2025; 59 screened patients did not participate (technical reasons: 15, overwhelmed by medical situation: 7, data protection concerns: 12, language barrier: 10, and other: 15). Patients with pancreatic diseases requiring surgical therapy were recruited for data collection at admission the day before surgery. Data were collected once a day for a period of 2 weeks at baseline before and immediately after the surgical procedure. In addition to HRQoL-VAS, the following items were collected: postoperative pain, nausea, feverish feeling, bowel movements, and the ability to drink or eat.

Neurosurgery

A total of 23 inpatients with spinal stenosis or herniated disc (Department of Neurosurgery, UKHD) used the system on their own devices between March 2025 and April 2025. Data were collected once a day for a period of 2 weeks immediately after the surgical procedure. In addition to HRQoL-VAS, the following items were collected: pain at the surgical site, leg pain, arm pain, and dysesthesia. Additionally, 9 patients did not use MyEDC (technical assistance needed: 6, language barrier: 1, overwhelmed by medical situation: 1, and other: 1).

Results

Patients entered HRQoL data via 1-time EDC links as described in the Methods section. The first question addressed overall QoL with a 100-level VAS. Different variants of this VAS were used. In visceral surgery, a vertical VAS was used (Figure 2); in psychosomatics and neurosurgery, a horizontal VAS was applied; and in hematology, a vertical VAS with an inverted scale (distress 1-100) was provided. In addition to the VAS, a set of disease-specific questions (such as pain for surgical patients) were asked. Data transfer from outside to the EHR system was implemented using a polling mechanism. On the basis of these data, EHR documents were generated automatically as a manual document list view (Figure S1 in Multimedia Appendix 1) and a detailed document view (Figure S2 in Multimedia Appendix 1), similar to a laboratory report. In a 9-month period (August 2024-April 2025), a total of 110 patients from UKHD participated in 4 clinical use cases.

Figure 3 presents QoL-T for 36 patients from psychosomatic medicine (32 female and 4 male patients; median age 27, IQR 18-58 years), who received weekly emails to collect HRQoL values. In the beginning, there was a wide range of HRQoL values. Patients provided HRQoL data with a relatively high level of completeness over the treatment period (median RR 92%, IQR 86%-97%in the first month). The participation rate was 92% (33/36; ie, highest of all 4 use cases). On average,

XSL•FO
RenderX

HRQoL values improved over time, indicating a positive effect of therapy (median HRQoL at baseline 42%, IQR 32%-52%, median HRQoL after 4 weeks 60%, IQR 41%-67%; *P*=.01). There is a negative correlation (significance level was not reached) between HRQoL and EDE-Q8 (Spearman correlation=–0.123, *P*=.06; Figure S3 in Multimedia Appendix 1), which indicates that decreased QoL is associated with higher levels of eating disorder symptoms.

Figure 3. (A) Health-related quality of life trajectories for 36 patients from psychosomatic medicine. Each line corresponds to a patient. A 95% confidence band is provided. (B) A histogram of the quality of life per patient (standardized area under the curve for each patient; median 0.55, IQR 0.43-0.64). QoL AUC: quality of life area under the curve.

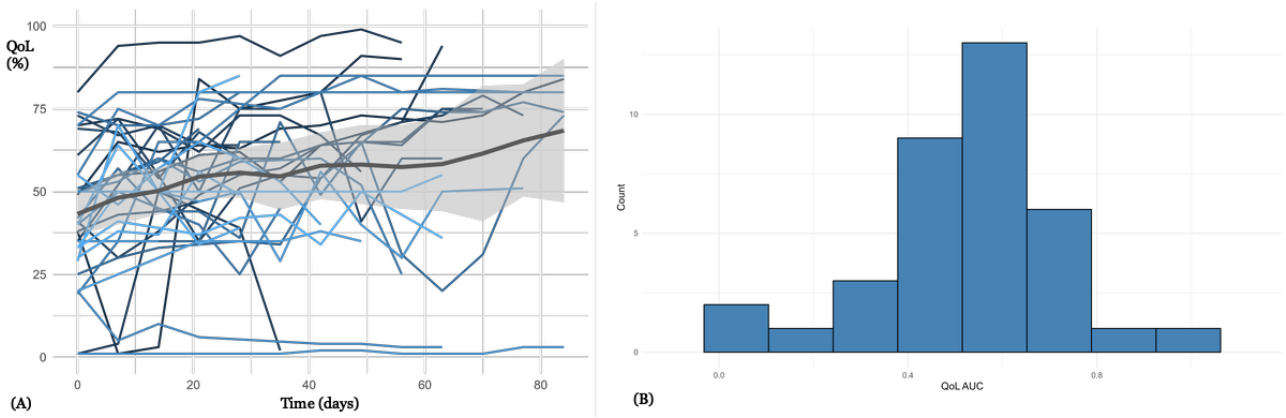

Figure 4 presents QoL-T for 25 patients from hematology (14 female and 11 male patients; median age 63, IQR 45-71 years), who received daily emails to collect HRQoL. Patients provided HRQoL data with a high level of completeness over the treatment period (median RR within 1 week 83%, IQR 80%-88%). There is a significant correlation of HRQoL and 30-item QoL Questionnaire (EORTC QLQ-C30) global health score (Spearman correlation=0.498, *P*=.02; Figure S4 in Multimedia Appendix 1). On average, HRQoL values deteriorate over a time interval of approximately 12 days, indicating the side effects of intensive chemotherapy.

Figure 4. (A) Health-related quality of life trajectories for 25 patients from hematology. Day 0 corresponds to start of chemotherapy. (B) Histogram of quality of life per patient (standardized area under the curve; median 0.46, IQR 0.41-0.62). QoL AUC: quality of life area under the curve.

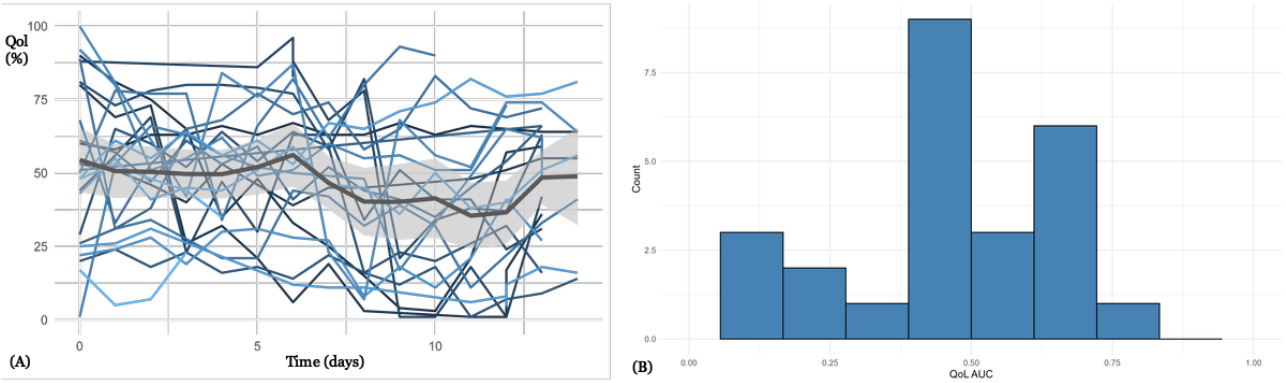

Figure 5 presents QoL-T for 26 patients from pancreatic surgery (10 female and 16 male patients; median age 66, IQR 42-83 years), who received daily messages on the day before surgery and 2 weeks after the surgical intervention. Participants achieved a median RR in the first week of 46%, IQR 46%-58%. The participation rate was 31% (8/26; ie, lowest of all 4 use cases). There is a highly significant association between HRQoL and pain levels (Spearman correlation=−0.552, *P*<.001; Figure S5 in Multimedia Appendix 1), which indicates that pain is associated with decreased QoL.

XSL•FO
RenderX

Figure 5. (A) Health-related quality of life trajectories for 26 patients from visceral surgery. Day 0 corresponds to the day before surgery. (B) Histogram of quality of life per patient (standardized area under the curve; median 0.59, IQR 0.49-0.77). QoL AUC: quality of life area under the curve.

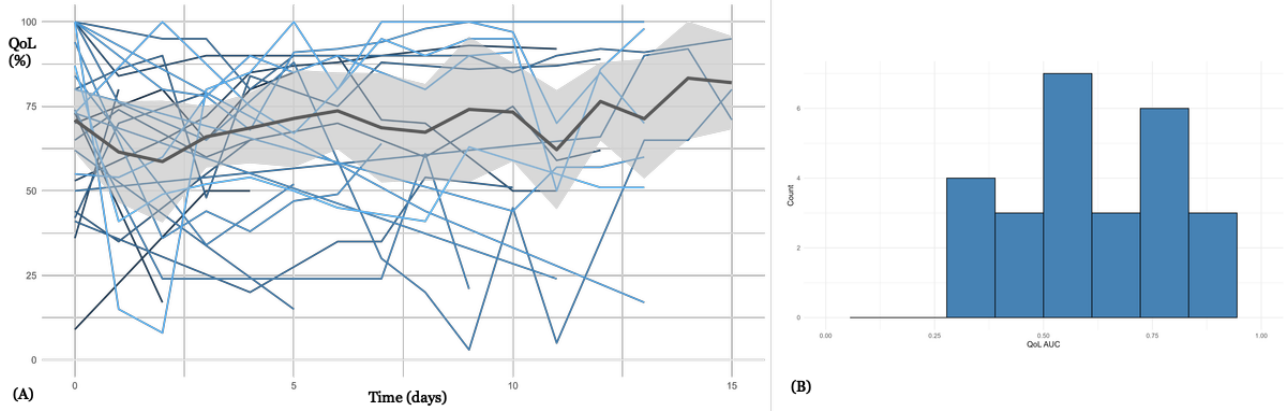

Figure 6 shows QoL-T for 23 patients from neurosurgery (7 female and 16 male patients; median age 64, IQR 21-85 years), who received daily messages 2 weeks after the surgical intervention. Participants achieved a median RR in the first week of 43%, IQR 31%-61%. A negative correlation between HRQoL and pain level was observed, which aligns with the direction of the association seen in visceral surgery. However, this did not reach statistical significance (Figure S6 in Multimedia Appendix 1).

Figure 6. (A) Health-related quality of life trajectories for 23 patients from neurosurgery. Day 0 corresponds to the day before surgery. (B) Histogram of quality of life per patient (standardized area under the curve; median 0.79, IQR 0.62-0.90). QoL AUC: quality of life area under the curve.

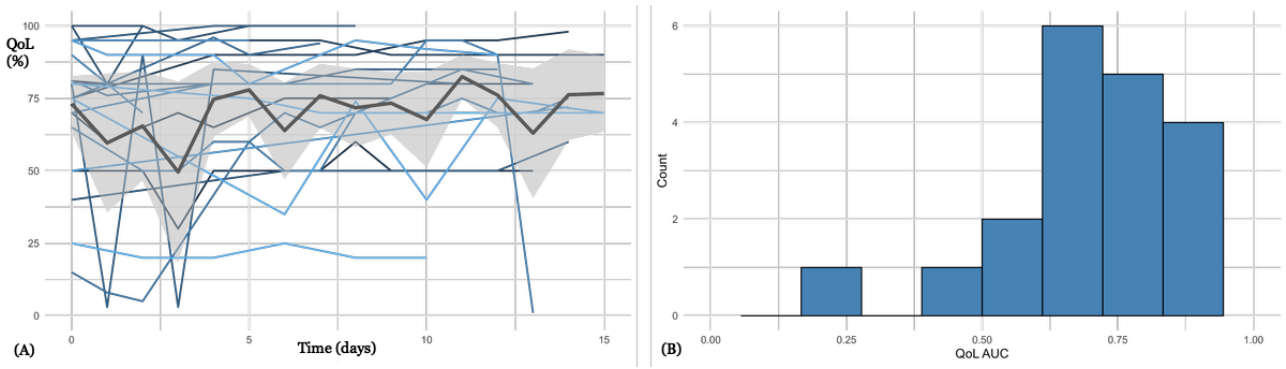

QoL-T in Figures 3-6 show considerable variability, both intraindividually (over time) and interindividually (comparison between patients). There are interpretable patterns; for example, in Figure 3 there is an average improvement of HRQoL during the course of inpatient treatment, or in Figure 4, there is an HRQoL change during the course of chemotherapy.

We did not find a significant effect of patient sex on response behavior (for details, see Figures S7-S10 in Multimedia Appendix 1). Of note, we did not observe a significant association between patient age and response count in any use case (for details, see Figures S11-S14 in Multimedia Appendix 1). We assessed time from email notification to patient answer in all use cases. The lowest time to patient answer was observed in psychosomatics (median 0.9 hours, IQR 0.3-8.0; Figure S15 in Multimedia Appendix 1). In hematology, median time to answer was 6.7 hours, IQR 1.5-17.2 (Figure S16 in Multimedia Appendix 1); in visceral surgery, median time to answer was 7.1 hours, IQR 2.9-18.9 (Figure S17 in Multimedia Appendix 1), and in neurosurgery, median time to answer was 5.3 hours, IQR 0.6-17.7 (Figure S18 in Multimedia Appendix 1).

Discussion

Overview

At present, survival (overall or progression free) is the most established primary end point in clinical research, especially in oncology [19]. Survival has the advantage that it can be collected with high precision in time (day of death) and completeness. However, survival is a crude measure for therapeutic success because it ignores PROMs such as QoL. Therefore, PROMs are increasingly used as end points in clinical studies [20]. This approach also has limitations because it only assesses the outcome at the end of an observation interval but not the patient trajectory: when 2 therapies result in the same average HRQoL after 1 year, therapy A can still be superior to therapy B, if HRQoL is significantly better during the time course of this year. Many established HRQoL instruments such as the WHO-5 well-being index, Functional Assessment of Cancer Therapy-General (FACT-G), or EQ-5D have short recall periods (<2 weeks). Consequently, frequent data collection (every 2 weeks up to daily) is needed for complete, high-resolution HRQoL trajectories. This is a methodological challenge because low RRs were reported in several settings.

XSL•FO
RenderX

Principal Findings

Our hypothesis was that assessing HRQoL on patient' mobile devices and integrating these findings into the EHR system can enable such high-resolution collection of HRQoL because a patient would be motivated to contribute to their own therapy. In our setting, a patient could directly inform their treatment team about QoL and symptoms. We could show that this approach is actually working in 4 completely different medical domains (psychosomatics, hematology, visceral surgery, and neurosurgery), both from a technical and a clinical perspective. Clinical staff emphasized the value of EHR integration because it supports patient care, for example, regarding pain management. Our proposed IT architecture proved itself in pilot operation. Patients were responding to these HRQoL requests on a regular basis.

What to measure regarding patient outcomes, including the frequency of assessment, was described as a key challenge [8]. We applied a 100-level VAS with a question in simple language because, from a clinical point of view, long texts are not suitable for seriously ill patients (eg, intensive care setting). It was demonstrated that even for those patients (eg, directly after major surgery), a large proportion is able and willing to answer HRQoL questions with their smartphones on a daily basis. In the European Union, software that calculates a clinical score may be classified as a medical device, depending on its intended use and clinical impact. While a single-item HRQoL question may be viewed as "clinical documentation," any software-based assessment intended to influence diagnostic or therapeutic decisions requires rigorous validation. Given the vast technical heterogeneity of patient-owned smartphones, the technical and clinical validation required for such software to meet medical device standards remains a significant hurdle for large-scale implementation. More complex HRQoL instruments, such as the EORTC QLQ-C30, can provide a more detailed assessment of different QoL dimensions but require substantially more documentation effort. We propose to use such instruments in larger time intervals combined with high-frequency short HRQoL assessments. In our study, patients of almost all age groups were capable to do this data collection. The clinical team or family members can support patients who are not able to handle a mobile device.

The median time from email to answer was <8 hours; therefore, daily or weekly assessments are feasible. We propose starting the VAS with 1 and assigning an HRQoL value of 0 only to patients who are deceased. By this means, HRQoL values can be directly transformed into survival data (dead=HRQoL 0 and alive=HRQoL 1-100). This enables a direct comparison of HRQoL trajectories with survival, which is the current standard end point.

Comparison With Prior Work

PROMs integrated with EHR were described previously [21]. High-resolution HRQoL trajectories within the EHR constitute a novel readout for the course of disease. This level of detail was previously unavailable and, to our knowledge, has not been comprehensively established for any disease. We demonstrated that this approach is feasible across multiple, clinically distinct domains, ranging from psychosomatics to visceral surgery. HRQoL trajectories could support early detection of complications. In contrast to survival analysis, effects of interventions such as chemotherapy or surgery on QoL can be assessed quantitatively with HRQoL trajectories. These trajectories can potentially enable sensitive comparison of different medical interventions in the future and change the course of therapy.

A related method is ecological momentary assessment [22], which captures real-time data about patients' experiences and behaviors in their natural environment. A similar approach such as in our pilot study was reported with the eRAPID system [23], which was applied for patients with breast, gynecological, and colorectal cancer. With eRAPID, the patient logged onto a web-based system (Qtool) to complete online questionnaires. To simplify this step, we sent a 1-time EDC link to the patient (ie, patient does not need an individual password). We used a CDISC-compliant EDC database to make the approach more scalable: the Medical Data Models portal [24] provides ≥25.000 questionnaires for download, which can be used for data collection with CDISC-compliant systems. Of note, the eRAPID team from Leeds was able to demonstrate that PROM collection as intervention can improve patient outcomes in patients with tumor (physical well-being and self-efficacy) [25]. Improved patient outcomes induced by PROM collection was also reported in other medical domains such as orthopedics [26]. In the field of cardiology, it was reported that a disease-specific PROM score is the strongest predictor for survival of heart failure [27].

Limitations and Future Work

Our findings are based on a single site pilot study with a relatively small sample size for each clinical use case. Consequently, our analysis may have lacked sufficient statistical power to detect smaller effect sizes. Due to technical constraints of the local EHR system, integration was limited to PDF files. Fast Healthcare Interoperability Resources–based EHR integration can enable improved support for clinical workflows. Further research is needed to define the best time intervals and most suitable assessment instruments for HRQoL collection. Most probably, these parameters are disease-specific. In some use cases, HRQoL measurements can be extended to the period before therapy. The rate of patients who accepted the electronic system varied considerably between use cases: in psychosomatics, >90% agreed to participate. In contrast, in visceral surgery, only about one-third of patients were willing or able to collect HRQoL data on the smartphone in the perioperative phase. HRQoL has several dimensions (eg, physical and psychosocial QoL). To compare 2 treatment schemes regarding HRQoL, a reduction to 1 dimension is needed (is therapy A better than B?). Therefore, further research is needed to define the most appropriate calculation of overall HRQoL. Another research topic is the metric to aggregate a series of HRQoL values. In this pilot study, we proposed AUC of HRQoL over time. Patient questionnaires can have a significant nonresponse bias, for example, language proficiency has a significant effect on RR [28]. In our setting, some patients had no access to a smartphone. Therefore, additional data capture methods (eg, tablet and interview by a health care professional) are needed for complete HRQoL data. At present, smartphones are used by >95% of the population in high-income

XSL•FO
RenderX

countries [29]. Electronic HRQoL collection software needs to be further improved, given the disease-related limitations for patients compared with the general population. Software with high usability can contribute to address this problem: for instance, the study by Carter et al [30] reported an increase in PROM completeness from 24.9% to 67.0% by user-centered software design. Future work will focus on leveraging HRQoL trajectories to provide clinical decision support and enabling physicians to personalize therapeutic interventions, while simultaneously establishing a sensitive, dynamic outcome measure for future clinical trials.

Conclusions

High-resolution QoL-T with EHR integration are clinically feasible in a broad range of clinical settings. Collection of HRQoL data via patients' mobile devices and secure data transfer to the protected EHR is technically feasible. These trajectories can provide a novel readout, which is not covered by survival analysis or a PROM end point. These data can potentially be useful to characterize diseases and compare different therapies with high sensitivity and time resolution.

Funding

This study was funded by the German Federal Ministry of Education and Research (BMBF; grant 01ZZ2321F).

Data Availability

Anonymized data and data analysis software for this manuscript are available in Multimedia Appendices 2 and 3. Access to individual patient data can be requested from the Medical Data Integration Center of Heidelberg University Hospital. Data access requests will be reviewed by the Use and Access Committee of Heidelberg University Hospital.

Authors' Contributions

Conceptualization: MD, TMP, MG
Formal analysis: MD
Investigation: SCF, TD, NS, CFB, PL, TMP
Methodology: SD-B, LJK, KK, FF-K
Resources: MD
Software: RF, MCB, LW, TMP, MG
Supervision: MD, SS, ML, SMK, HCF, TMP, MG
Validation: MD
Writing – original draft: MD
Writing – review: SS, ML, SMK, HCF, TMP, SD-B, LJK, KK, FF-K

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data regarding electronic health record integration, correlation analyses, and time to patient answers.
[DOCX File , 1578 KB - jmir_v28i1e79834_app1.docx]

Multimedia Appendix 2
Program code for quality of life trajectories (trajectory.R).
[TXT File , 2 KB - jmir_v28i1e79834_app2.txt]

Multimedia Appendix 3
Example data for quality of life trajectories (example.R).
[TXT File , 2 KB - jmir_v28i1e79834_app3.txt]

References

1. Escudero-Vilaplana V, Bernal E, Casado G, Collado-Borrell R, Diez-Fernández R, Fernández Román AB, et al. Defining a standard set of patient-reported outcomes for patients with advanced ovarian cancer. Front Oncol 2022;12:885910 [FREE Full text] [doi: 10.3389/fonc.2022.885910] [Medline: 35664764]
2. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 well-being index: a systematic review of the literature. Psychother Psychosom 2015;84(3):167-176 [FREE Full text] [doi: 10.1159/000376585] [Medline: 25831962]
3. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]

XSL•FO
RenderX

4. Minasian LM, O'Mara A, Mitchell SA. Clinician and patient reporting of symptomatic adverse events in cancer clinical trials: using CTCAE and PRO-CTCAE to provide two distinct and complementary perspectives. Patient Relat Outcome Meas 2022;13:249-258 [FREE Full text] [doi: 10.2147/PROM.S256567] [Medline: 36524232]
5. Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol group. Ann Med 2001 Jul 08;33(5):337-343. [doi: 10.3109/07853890109002087] [Medline: 11491192]
6. Mori M, Dhruva SS, Geirsson A, Krumholz HM. Characterization of multi-domain postoperative recovery trajectories after cardiac surgery using a digital platform. NPJ Digit Med 2022 Dec 24;5(1):192 [FREE Full text] [doi: 10.1038/s41746-022-00736-0] [Medline: 36564550]
7. Wang K, Eftang CN, Jakobsen RB, Årøen A. Review of response rates over time in registry-based studies using patient-reported outcome measures. BMJ Open 2020 Aug 06;10(8):e030808 [FREE Full text] [doi: 10.1136/bmjopen-2019-030808] [Medline: 32764078]
8. Marra C, Chico T, Alexandrow A, Dixon WG, Briffa N, Rainaldi E, et al. Addressing the challenges of integrating digital health technologies to measure patient-centred outcomes in clinical registries. Lancet Digit Health 2025 Mar;7(3):e225-e231 [FREE Full text] [doi: 10.1016/S2589-7500(24)00223-1] [Medline: 39672737]
9. Greulich L, Hegselmann S, Dugas M. An open-source, standard-compliant, and mobile electronic data capture system for medical research (OpenEDC): design and evaluation study. JMIR Med Inform 2021 Nov 19;9(11):e29176 [FREE Full text] [doi: 10.2196/29176] [Medline: 34806987]
10. Dugas M, Blumenstock M, Dittrich T, Eisenmann U, Feder SC, Fritz-Kebede F, et al. Next-generation study databases require FAIR, EHR-integrated, and scalable electronic data capture for medical documentation and decision support. NPJ Digit Med 2024 Jan 12;7(1):10 [FREE Full text] [doi: 10.1038/s41746-023-00994-6] [Medline: 38216645]
11. Operational data model (ODM). Clinical Data Interchange Standards Consortium (CDISC). URL: https://www.cdisc.org/standards/data-exchange/odm [accessed 2025-03-15]
12. Kroenke K, Spitzer RL, Williams JB, Löwe B. An ultra-brief screening scale for anxiety and depression: the PHQ–4. Psychosomatics 2009 Nov;50(6):613-621. [doi: 10.1016/S0033-3182(09)70864-3]
13. Toussaint A, Murray AM, Voigt K, Herzog A, Gierk B, Kroenke K, et al. Development and validation of the somatic symptom disorder-B criteria scale (SSD-12). Psychosom Med 2016 Jan;78(1):5-12. [doi: 10.1097/PSY.0000000000000240] [Medline: 26461855]
14. Prinz U, Nutzinger D, Schulz H, Petermann F, Braukhaus C, Andreas S. Die symptom-checkliste-90-R und ihre Kurzversionen: psychometrische analysen bei patienten mit psychischen erkrankungen. Phys Rehab Kur Med 2008 Nov 25;18(06):337-343. [doi: 10.1055/s-0028-1093323]
15. Horvath AO, Greenberg LS. Development and validation of the working alliance inventory. J Couns Psychol 1989 Apr;36(2):223-233. [doi: 10.1037/0022-0167.36.2.223]
16. Fonagy P, Luyten P, Moulton-Perkins A, Lee Y, Warren F, Howard S, et al. Development and validation of a self-report measure of mentalizing: the reflective functioning questionnaire. PLoS One 2016 Jul 8;11(7):e0158678 [FREE Full text] [doi: 10.1371/journal.pone.0158678] [Medline: 27392018]
17. Kliem S, Mößle T, Zenger M, Strauß B, Brähler E, Hilbert A. The eating disorder examination-questionnaire 8: a brief measure of eating disorder psychopathology (EDE-Q8). Int J Eat Disord 2016 Jun;49(6):613-616. [doi: 10.1002/eat.22487] [Medline: 26711183]
18. Dueck AC, Mendoza TR, Mitchell SA, Reeve BB, Castro KM, Rogak LJ, National Cancer Institute PRO-CTCAE Study Group. Validity and reliability of the US National Cancer Institute's patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE). JAMA Oncol 2015 Nov 01;1(8):1051-1059 [FREE Full text] [doi: 10.1001/jamaoncol.2015.2639] [Medline: 26270597]
19. Royle KL, Meads D, Visser-Rogers JK, White IR, Cairns DA. How is overall survival assessed in randomised clinical trials in cancer and are subsequent treatment lines considered? A systematic review. Trials 2023 Nov 06;24(1):708 [FREE Full text] [doi: 10.1186/s13063-023-07730-1] [Medline: 37926806]
20. Guidance for industry patient-reported outcome measures: use in medical product development to support labeling claims. Food and Drug Administration (FDA). URL: https://www.fda.gov/media/77832/download [accessed 2025-02-12]
21. Mohindra NA, Garcia SF, Kircher S, Barnard C, Perry LM, Lyleroehr M, et al. Development of an electronic health record-integrated patient-reported outcome-based shared decision-making dashboard in oncology. JAMIA Open 2024 Oct;7(3):ooae056. [doi: 10.1093/jamiaopen/ooae056] [Medline: 39049991]
22. Kraft R, Reichert M, Pryss R. Mobile crowdsensing in ecological momentary assessment mHealth studies: a systematic review and analysis. Sensors (Basel) 2024 Jan 12;24(2):472 [FREE Full text] [doi: 10.3390/s24020472] [Medline: 38257567]
23. Holch P, Warrington L, Bamforth LC, Keding A, Ziegler LE, Absolom K, et al. Development of an integrated electronic platform for patient self-report and management of adverse events during cancer treatment. Ann Oncol 2017 Sep 01;28(9):2305-2311 [FREE Full text] [doi: 10.1093/annonc/mdx317] [Medline: 28911065]
24. Riepenhausen S, Blumenstock M, Niklas C, Hegselmann S, Neuhaus P, Meidt A, et al. Europe's largest research infrastructure for curated medical data models with semantic annotations. Methods Inf Med 2024 May 13;63(1-02):52-61 [FREE Full text] [doi: 10.1055/s-0044-1786839] [Medline: 38740374]

XSL•FO
RenderX

25. Absolom K, Warrington L, Hudson E, Hewison J, Morris C, Holch P, et al. Phase III randomized controlled trial of eRAPID: eHealth intervention during chemotherapy. J Clin Oncol 2021 Mar 01;39(7):734-747 [FREE Full text] [doi: 10.1200/JCO.20.02015] [Medline: 33417506]
26. Schöner L, Kuklinski D, Wittich L, Steinbeck V, Langenberger B, Breitkreuz T, et al. Cost-effectiveness of a patient-reported outcome-based remote monitoring and alert intervention for early detection of critical recovery after joint replacement: a randomised controlled trial. PLoS Med 2024 Oct 9;21(10):e1004459 [FREE Full text] [doi: 10.1371/journal.pmed.1004459] [Medline: 39383175]
27. Pokharel Y, Khariton Y, Tang Y, Nassif ME, Chan PS, Arnold SV, et al. Association of serial Kansas City cardiomyopathy questionnaire assessments with death and hospitalization in patients with heart failure with preserved and reduced ejection fraction: a secondary analysis of 2 randomized clinical trials. JAMA Cardiol 2017 Dec 01;2(12):1315-1321 [FREE Full text] [doi: 10.1001/jamacardio.2017.3983] [Medline: 29094152]
28. KÖhn S, Schlumbohm A, Marquardt M, Scheel-Sailer A, Tobler S, Vontobel J, et al. Predicting non-response in patient-reported outcome measures: results from the Swiss quality assurance programme in cardiac inpatient rehabilitation. Int J Qual Health Care 2022 Dec 08;34(4):mzac093 [FREE Full text] [doi: 10.1093/intqhc/mzac093] [Medline: 36399024]
29. Mobile phone ownership. International Telecommunication Union. URL: https://www.itu.int/itu-d/reports/statistics/2024/11/10/ff24-mobile-phone-ownership/ [accessed 2025-02-12]
30. Carter E, Benda N, Kim S, Qiu Y, Yu Z, Gunning F, et al. Increasing completion of daily patient-reported outcomes in psychotherapies for late-life depression through user-centered design. Appl Clin Inform 2024 Oct 20;15(5):986-996 [FREE Full text] [doi: 10.1055/s-0044-1790545] [Medline: 39566558]

Abbreviations

AUC: area under the curve
CDISC: Clinical Data Interchange Standards Consortium
EDC: electronic data capture
EHR: electronic health record
EDE-Q8: 8-item Eating Disorder Examination-Questionnaire
EORTC QLQ-C30: 30-item Core QoL Questionnaire
FACT-G: Functional Assessment of Cancer Therapy-General
HRQoL: health-related quality of life
PHQ-4: 4-item Patient Health Questionnaire
PRO-CTCAE: patient-reported outcomes–common terminology criteria for adverse events
PROM: patient-reported outcome measure
QoL: quality of life
QoL-T: quality of life trajectories
RFQ: Reflective Functioning Questionnaire
RR: response rate
SCLK-11: Symptom-Checkliste Kurzversion
SSD-12: 12-item Somatic System Disorder
UKHD: Heidelberg University Hospital (Universitätsklinikum Heidelberg)
VAS: visual analog scale
WHO-5: 5-item World Health Organization well-being index

Edited by A Coristine; submitted 29.Jun.2025; peer-reviewed by U Hübner, Y Chu; comments to author 04.Nov.2025; revised version received 28.Jan.2026; accepted 29.Jan.2026; published 24.Feb.2026.

Please cite as:
Dugas M, Fleige R, Blumenstock MC, Feder SC, Dittrich T, Siegel N, Bergmann CF, Wettach L, Sauer S, Lenga P, Krieg SM, Dugas-Breit S, Kessler LJ, Kosmalla K, Fritz-Kebede F, Loos M, Friederich HC, Pausch TM, Ganzinger M
Quality of Life Trajectories With Integration Into Electronic Health Records for High-Resolution Patient Outcomes: Algorithm Development and Validation Study
J Med Internet Res 2026;28:e79834
URL: https://www.jmir.org/2026/1/e79834
doi:10.2196/79834
PMID:

XSL•FO
RenderX

©Martin Dugas, Robin Fleige, Max Christian Blumenstock, Stephan Christoph Feder, Tobias Dittrich, Niels Siegel, Celine Fabienne Bergmann, Luis Wettach, Sandra Sauer, Pavlina Lenga, Sandro M Krieg, Susanne Dugas-Breit, Lucy Joanne Kessler, Kosima Kosmalla, Fleur Fritz-Kebede, Martin Loos, Hans-Christoph Friederich, Thomas M Pausch, Matthias Ganzinger. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Artificial Intelligence for Identifying Patient-Reported Outcome and Experience Measures in Oncology: Retrospective Cross-Sectional Study Using ClinicalTrials.gov

Jessica Soyer[1*], MSc; Akram Hecini[1*], MSc; Sylvain Juchet[1], PharmD, MSc; Céline Desvignes-Gleizes[2], PhD; Maxime Thiebaut[1], MSc, PharmD; Jean-Philippe Bertocchio[1,3,4,5], MD, PhD

[1]SKEZI, Annecy, France

[2]Mapi Research Trust, Lyon, France

[3]Thyroïde – Tumeurs Endocrines, Pitié-Salpêtrière Hospital, Assistance Publique-Hôpitaux de Paris, Paris, France

[4]Reference center for rare diseases related to calcium-phosphate metabolism disorders, Pitié-Salpêtrière Hospital, Assistance Publique-Hôpitaux de Paris, Paris, France

[5]FHU-DDS-net, Dental School, Université Paris Cité, INSERM U1333 Santé Orale, Montrouge, France

[*]these authors contributed equally

Corresponding Author:
Jean-Philippe Bertocchio, MD, PhD
SKEZI
Les Papeteries - Image Factory, 1-3 Esplanade Augustin Aussedat
Annecy
France
Phone: 33 0670333450
Email: jean-philippe.bertocchio@skezi.eu

Abstract

Background: Traditional health care systems have evolved to increasingly recognize patients' perspectives as key to improving the quality of care, especially in oncology. Hence, patient-reported outcome measures (PROMs) and patient-reported experience measures (PREMs) can enhance patient–health care provider communication while facilitating individualized care. This tailored approach improves patient outcomes and highlights the need for research methods to account for variability in patient experiences across diverse sociodemographic and clinical backgrounds while proposing artificial intelligence (AI) to automate and accelerate the identification of PROMs and PREMs.

Objective: This study aimed to estimate the proportion of clinical studies including PROMs or PREMs, using either a traditional expert-based identification method or an AI-enriched approach.

Methods: In a retrospective cross-sectional study using the ClinicalTrial.gov database, we focused on oncology studies between 2012 and 2022. Two methods were used to identify PROMs and PREMs: (1) a traditional expert-based method, where an algorithm identified PROMs and PREMs from a list of 346 oncology-specific PROMs and PREMs (extracted from the PROQOLID database, Mapi Research Trust) and/or 11 PROM- and PREM-specific terms, and (2) an AI-enriched method using a bidirectional encoder representations from transformers model, trained on 2399 outcomes labeled by experts. To evaluate algorithm performance, results were compared with expert decisions. Studies were classified by PROM/PREM use, analyzed using logistic regression to identify drivers, and a named entity recognition model identified frequently used measures.

Results: A total of 24,491 studies were included. According to the traditional expert-based algorithm, 7549 (31%) studies used at least one PROM or PREM, as compared to 8029 (33%) studies identified by the AI-enriched algorithm, increasing from 2012 to 2022 (P<.001). With 90% (95% CI 88%-92%) accuracy, the AI-enriched algorithm outperformed (P<.001) the expert-based algorithm (84%, 95% CI 82%-86%) in identifying PROMs and PREMs. Breast and digestive cancers accounted for nearly 50% of all oncology studies using PROMs and PREMs, with the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-Core 30 being the most frequently used. As expected, trials in phases 2 to 4 more frequently included PROMs and PREMs than preclinical or early phase 1 studies (odds ratio [OR] 1.8, 95% CI 1.1-2.8 for phase 2; OR 3.6, 95% CI 2.3-5.8 for phase 3; and OR 2.6, 95% CI 1.6-4.4 for phase 4). In observational studies, cross-sectional and prospective studies incorporated more PROMs and PREMs than retrospective studies (OR 4.6, 95% CI 3.3-6.4 and OR 3.2, 95% CI 2.5-4.1, respectively).

XSL•FO
RenderX

Conclusions: Our study shows that an AI-enriched algorithm outperforms traditional expert methods in identifying PROMs and PREMs in oncology. Combining expert-labeled data with AI enables scalable, automated trial monitoring, supports efficient research, informs stakeholders, enhances patient-centered decisions, and can be extended to other diseases and databases.

(J Med Internet Res 2026;28:e84533) doi:10.2196/84533

KEYWORDS

clinical trials; patient-reported outcome measures; patient-reported experience measures; artificial intelligence-based research; expert-based research; oncology; research on research

Introduction

Traditional models in health care have evolved to increasingly recognize patients' perspectives as key to improving their quality of services [1]. In such, the quality of data is of major importance for evaluating services provided to people. To date, most of the data to do so comes from automatically generated sources emerging from health medical records and related software. Besides, data collected from patient-centered measures (PCMs) may also play a key role in improving the quality of services [2]. PCMs refer to widely used and studied tools, such as patient-reported outcome measures (PROMs) or patient-reported experience measures (PREMs) [3], which constitute relevant information collected directly from patients to better understand and address what matters the most to them [2,4]. In their day-to-day practice, health care providers (HCPs) may collect PCMs for various purposes such as screening, monitoring, evaluating care, or collecting side effects of treatments as well as to help in the decision-making process [5]. PCMs are usually collected through questionnaires (either paper-based or online) and help capture either outcomes resulting from the care (such as the evolution of symptoms) or the experience of such care (such as the time spent receiving a specific treatment). Items (questions) can be scientifically validated, and therefore, such questionnaires are called PROMs or PREMs for collecting outcomes or experiences, respectively, or made for the purpose of the questionnaire itself, and are therefore called *ad hoc* items. These measurements are added to biological measures and/or clinician-reported outcomes, such as physical examination, to better plan patient care.

PROMs are standardized questionnaires that collect information about health outcomes directly from patients, such as symptoms, functional status, health-related quality of life (HRQoL), or mental well-being, without amendment or interpretation by a clinician or anyone else, and measure them in a quantitative manner at a specific timepoint [6-8]. Originally developed in research studies, PROMs applications are now diversified and used in supporting clinical decision-making, assessing treatment benefit in drug trials, prioritizing patients for surgical procedures, comparing outcomes among HCPs, stimulating HRQoL improvement, and evaluating practices [9-11]. There are two main categories of PROMs: (1) generic and (2) specific ones (including both condition-specific and concept-specific items). Generic PROMs measure concepts relevant to a wide range of patients, enabling comparisons across different conditions. For example, the EQ-5D questionnaire [12,13] measures 5 dimensions of health: mobility, self-care, usual activities, pain or discomfort, and anxiety or depression. Each dimension is assessed using a set of severity levels, initially comprising 3 levels (EQ-5D-3L) and expanded to 5 levels in the updated 2010 version (EQ-5D-5L), notably to improve sensitivity and reduce ceiling effect [13]. Specific PROMs, on the other hand, measure concepts relevant to a particular patient group or condition. For instance, the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-Core 30 (EORTC QLQ-C30) questionnaire measures HRQoL specifically in patients with cancer (condition-specific) [14], while the Fatigue Severity Scale (FSS) evaluates the concept of fatigue (concept-specific) [15]. Besides this core common questionnaire for all patients with cancer, additional modules have been developed from it to measure symptoms and outcomes specific to different cancers. Examples include the EORTC Quality of Life Questionnaire–Lung Cancer Module (EORTC QLQ-LC13) or the Head and Neck Cancer Module (EORTC QLQ-H&N35) [16,17]. For the sole field of cancer, more than 300 PROMs have already been identified.

In contrast to PROMs, PREMs focus on patients' experience and emerge as a favored metric for assessing health care quality, such as the validated Kidney PREM questionnaire [18]. PREMs assess common information about patients satisfaction with service delivery in a clinical setting or describe patients experiences with a therapy or a plan of care [4,8]. Like the US Food and Drug Administration (FDA), other public health authorities in Europe, such as the Haute Autorité de Santé (HAS) in France, have recently added PROMs and PREMs to their list of care quality and safety indicators to encourage their use in clinical trials and routine [8,19-21]. However, there is no established consensus regarding the way to name PROMs and PREMs. This makes their identification particularly challenging. New tools based on natural language processing (NLP) may be useful in this setting [22]. Overall, it appears essential to develop a solution for identifying and tracking the use of PROMs and PREMs in clinical research, both in terms of their frequency of inclusion in studies and the types of instruments selected. We hypothesized that an artificial intelligence (AI)–enriched algorithm would outperform an expert-based algorithm in identifying PROMs and PREMs used in clinical research. Here, we report the results from a cross-sectional study that aimed to estimate the proportion of oncology clinical studies incorporating PROMs and PREMs and to examine the characteristics associated with these studies, using either a traditional expert-based method or an AI-enriched approach. The performances of the expert-based and AI-enriched algorithms were evaluated through comparison to a manually labeled subsample validated by a human-based consensus.

XSL•FO
RenderX

Methods

Study Design and Data Sources (Setting)

Our retrospective cross-sectional study uses data regarding outcomes of interest from previously documented clinical studies. We conducted a retrospective cross-sectional study to provide a comprehensive assessment of the prevalence and characteristics of PROMs and PREMs in oncology trials over a defined period (2012-2022). This design was chosen because all study data were already publicly available in the ClinicalTrials.gov registry, allowing us to analyze a large dataset without prospectively collecting new data, and to efficiently capture a snapshot of PROMs and PREMs inclusion across multiple studies at a single point in time. This data was extracted from the ClinicalTrials.gov database, more specifically, the "What is the study measuring?" section of the studies' information page, using the *trials* R package. To be included, studies had (1) to be available in the open access ClinicalTrials.gov database; (2) to have a study start date between January 1, 2012, and January 1, 2022; (3) to be set in the oncology field; and (4) to contain at least one of our described outcomes. Studies that contained no outcome description were therefore excluded. If one study had more than one outcome, all available outcomes were extracted individually. Because the study relied exclusively on publicly available registry data, nonparticipation was not applicable.

Identifying PROMs and PREMs

Overview

Two different techniques were used to identify PROMs and PREMs in oncology studies' outcomes: (1) an expert-based algorithm and (2) an AI-enriched algorithm. In addition, a human-based consensus manually labeled (as containing or not PROMs or PREMs) a subset of the included outcomes. This human-based consensus was used first to feed the AI-enriched algorithm and act as a gold standard for the evaluation of the performance of the two algorithms.

Expert-Based Algorithm

A list of oncology-specific PROMs and PREMs and PROM- and PREM-related generic open terms was gathered. First, 346 PROMs and PREMs used in oncology, including both their full names and acronyms, were extracted from the Patient-Reported Outcome and Quality of Life Instruments Database (PROQOLID). Second, 2 distinct groups of 3 experts set up a list of terms related to generic PROMs and PREMs for the purpose of this study. The experts then reached a consensus, resulting in a final list of PROM- and PREM-related generic open terms. PROM-specific terms were quality of life, QoLd, patient-reported outcome, psychometric, patient perspective, questionnaire, health related quality of life, and hrql. PREM-specific terms were satisfaction, experience, and quality of care.

The list of 346 oncology-specific PROMs and PREMs and 11 PROM- and PREM-related generic open terms was then used in the expert-based algorithm. To identify outcomes, including PROMs and/or PREMs, the expert-based algorithm specifically searched for PROMs or PREMs and specific terms within the outcome descriptions available in the dataset of oncology studies extracted from ClinicalTrials.gov.

Expert Consensus

A subset of studies' outcomes was randomly selected to be manually labeled by 16 independent experts (ie, not involved in the process of identifying the PROMs and PREMs generic terms). This random subset was built to balance positive and negative outcomes, that is, included 15% of the outcomes with PROMs or PREMs as identified by the expert-based algorithm and 2% of the outcomes without PROMs or PREMs. The 16 experts were divided into 8 pairs; each pair was assigned a random sample of the subset to label. Experts were asked to label whether or not the descriptions contained either a PROM or a PREM: to be counted as mentioning a PROM or PREM, a description should provide a name (either an acronym or a full name) referring to a questionnaire that should be answered directly by patients and that should assess either an outcome or an experience of care. Therefore, we excluded descriptions that mentioned using a PROM or PREM but that did not state which one, that used an ad hoc questionnaire, or that used a clinician-reported outcome measure. While labeling the outcomes, experts were blind to the identity of their pairs and to the label provided by their pairs. Once all outcomes were labeled, experts were unblinded to their pair and reached consensus in case of disagreement (Cohen κ 0.59), setting the gold-standard label for the outcome. During the whole process, experts were blind to the classification according to the expert-based algorithm. Even if the initial agreement between the 2 experts was 59% only, this first round was only a starting point. For all cases where disagreements occurred, the experts engaged in detailed discussions and conducted additional reviews of the scientific literature to resolve them until a consensus was reached. Through this consensus-building process, all disagreements were carefully examined and validated, leading to final labels that are 100% agreed on by both experts and grounded in verified evidence. Therefore, while the initial agreement reflects the complexity of the task, the final gold standard used for evaluation is fully consensual and reliable.

AI-Enriched Algorithm

The AI-enriched algorithm consisted of a text classification model built using bidirectional encoder representations from transformers (BERT) [23]. This model used texts from outcomes descriptions labeled by the expert consensus as containing or not PROMs or PREMs. For the model requiring balanced classes (ie, equal number of outcomes with and without PROMs or PREMs), a corresponding subset of the manually labeled outcomes was used. The dataset was split as follows: 61% of the data were used for training (2399 samples), 12% for validation (479 samples), and 27% for testing (1065 samples). Using BERT Tokenizer [23], each word or textual element was converted into a token, which was then encoded as a fixed-length numerical vector of 512. This numerical vector captured semantic and contextual information within the texts. Model training was conducted using the Trainer from the Transformers library developed by Hugging Face [24]. During training, multiple models were generated by adapting batch size and

XSL•FO
RenderX

learning rate to get the best performance. All generated models were evaluated using the associated validation subset. The performance of each model was assessed using the following metrics: accuracy, recall, and F_1-score. These metrics gauged the model's ability to correctly classify texts into their respective categories. The best model was selected based on its accuracy score.

Evaluating the Performance of Algorithms

To evaluate algorithms' performances (accuracy, sensitivity, and specificity), the results obtained by expert-based and AI-enriched algorithms were compared to a dataset of outcomes labeled manually by expert consensus (gold standard). Evaluations were performed on the same dataset corresponding to the performance evaluation subset randomly selected when building datasets for the AI-enriched algorithm.

Describing the Studies Using PROMs and PREMs

Each study was classified as using at least one PROM or PREM or not, based on an AI-enriched algorithm (or the expert-based algorithm, see Multimedia Appendix 1). As observational and interventional studies carried out different characteristics, the analyses were stratified according to the type of study. We used multivariate logistic regression analysis to investigate potential associations between the use of PROMs and PREMs adjusted to other characteristics: cancer site, intervention type, clinical trial phase, randomization type, model type, and time perspective.

Recoding Cancer Sites

The ClinicalTrials.gov database returns a wide variety of names for cancer sites, even for the same cancers, as this field is an open-label one. Cancer sites had to be recoded in order to enhance the homogeneity and accuracy of the data. First, a *k*-means clustering [25] was used to group the cancer sites based on the similarities between names for each clinical study. The 10 most recurrent words in each cluster were identified for naming each cluster. The cluster names (cancer sites) were then used in a subsequent few-shot classification task, using the setfit model from Hugging Face [26]. For each cancer site, eight representative sentences were carefully selected from the dataset, which were used to create a final set of 116 texts. This dataset was used to train a text classification model, which was then applied to classify all the cancer sites in the dataset. The final model presented an accuracy of 96% for classifying cancer sites.

Identifying the Most Frequently Used PROMs and PREMs

Although the main AI-enriched algorithm was designed to detect if a PROM or PREM was used or not, it was not designed to state which PROM or PREM it was referring to. To identify which specific PROMs or PREMs were the most frequently used in clinical research, we used a named entity recognition (NER) model called SKEP'IA, which we had previously developed and patented. First, we manually annotated more than 2500 outcomes containing PROM names using the open-source annotation tool Doccano. Second, using the open-source library for natural language processing in Python, spaCy, these entities were then converted into a BIO tagging scheme (beginning, inside, outside) to mark the start, continuation, and boundaries of PROM entities within the text. The NER model was trained on this structured dataset. Third, we fine-tuned a BERT-cased model, splitting the dataset into training, validation, and test sets using appropriate proportions. Hyperparameters were optimized with Optuna. Fourth, the final model was evaluated on the test set. This model can recognize any PROM or PREM within a text, whether they are presented in full text or as an acronym and can suggest relevant ones based on textual descriptions. SKEP'IA was developed by using labeled data collected from both ClinicalTrials.gov and PubMed. However, the absence of a standardized and official naming system for each PROM or PREM presented a challenge in determining the most commonly used ones, representing a limitation in the exclusive use of the SKEP'IA model. To address this issue, we implemented a text similarity search approach using the PROQOLID database. Thus, the PROMs and PREMs identified by SKEP'IA were compared to the list of PROMs and PREMs previously captured from the PROQOLID database by calculating text similarity between them. Finally, we combined both the NER model and the text similarity search in the PROQOLID database.

Statistical Analysis

Statistical analyses were performed using R Studio software (version 4.4.2; Posit PBC). Descriptive statistics were used to illustrate the use of PROMs and PREMs in registered clinical studies. Characteristics of the studies were presented by their numbers and proportions. Continuous variables were analyzed as reported in the registry and were not categorized. Proportions were compared bivariately using a chi-square test. Missing data were reported where appropriate in the descriptive table, and their randomness was assessed using a missing completely at random (MCAR) test (RBtest R package) with results displayed in the table footnote. Year-over-year trends in the use of PROMs and PREMs were described. A multivariate logistic regression analysis was used to investigate possible associations between the use of PROMs and PREMs and study characteristics, by subgroups of interventional or observational studies. The variables included in the models were those obtained from the ClinicalTrials.gov database and considered relevant to the study. The odds ratios (ORs) were presented using forest plots, and statistical significance for each predictor was assessed using Wald tests. Because associations were expressed as ORs, translation into absolute risk was not applicable. A *P* value less than .05 was considered as reflecting statistical significance. No subgroup, interaction, or sensitivity analyses were prespecified or conducted, as the study followed a predefined analytical framework.

Ethical Considerations

This study did not involve human participants or the use of identifiable human data. Therefore, institutional review board approval, informed consent, and participant remuneration were not applicable. Our observational study is compliant with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines (checklist in Multimedia Appendix 1). No images or figures in the manuscript or multimedia

XSL•FO
RenderX

appendix contain information that could identify individual participants or users.

Results

From the Selection of Studies to PROMs and PREMs Identification

Between January 1, 2012, and January 1, 2022, a total of 28,034 studies in oncology were found in the ClinicalTrials.gov database, including 24,491 with at least one outcome described (Figure 1A). According to the expert-based algorithm, 16,954 outcomes using PROMs or PREMs were found from 7549 studies, meaning that out of the 24,491 eligible oncology studies, 31% used at least one PROM or PREM (Figure 1B). A subset of 4770 studies' outcomes was used to evaluate the performance of the expert-based algorithm. When compared to the expert consensus manually labeled gold standard, the expert-based algorithm achieved 83% (95% CI 82%-84%) accuracy, 86% (95% CI 85%-88%) sensitivity, and 81% (95% CI 80%-82%) specificity (Table 1).

Figure 1. Flowchart displaying studies and outcomes selection (A), the expert-based (B) and artificial intelligence (AI)–enriched (C) methods to classify outcomes, ClinicalTrial.gov database, 2012-2022. (A) From 259,899 studies registered between January 1, 2012, and January 1, 2022, a total of 28,034 oncology studies were identified. After excluding studies with no outcomes and outcomes lacking descriptions, 142,080 outcomes (from 24,491 studies) were included. (B) An expert-based method identified 16,954 outcomes using patient-reported outcomes measures (PROMs) or patient-reported experience measures (PREMs) across 7549 studies. A random sample of 4770 outcomes (890 studies) was manually labelled by expert consensus (15% of PROMs- or PREMs-positive and 2% of PROMs- or PREMs-negative outcomes). (C) The expert consensus dataset was then used to train and evaluate an AI-enriched algorithm, which was then applied to the full dataset of 142,080 outcomes. The AI-based algorithm identified 25,104 PROMs or PREMs outcomes across 8029 studies.

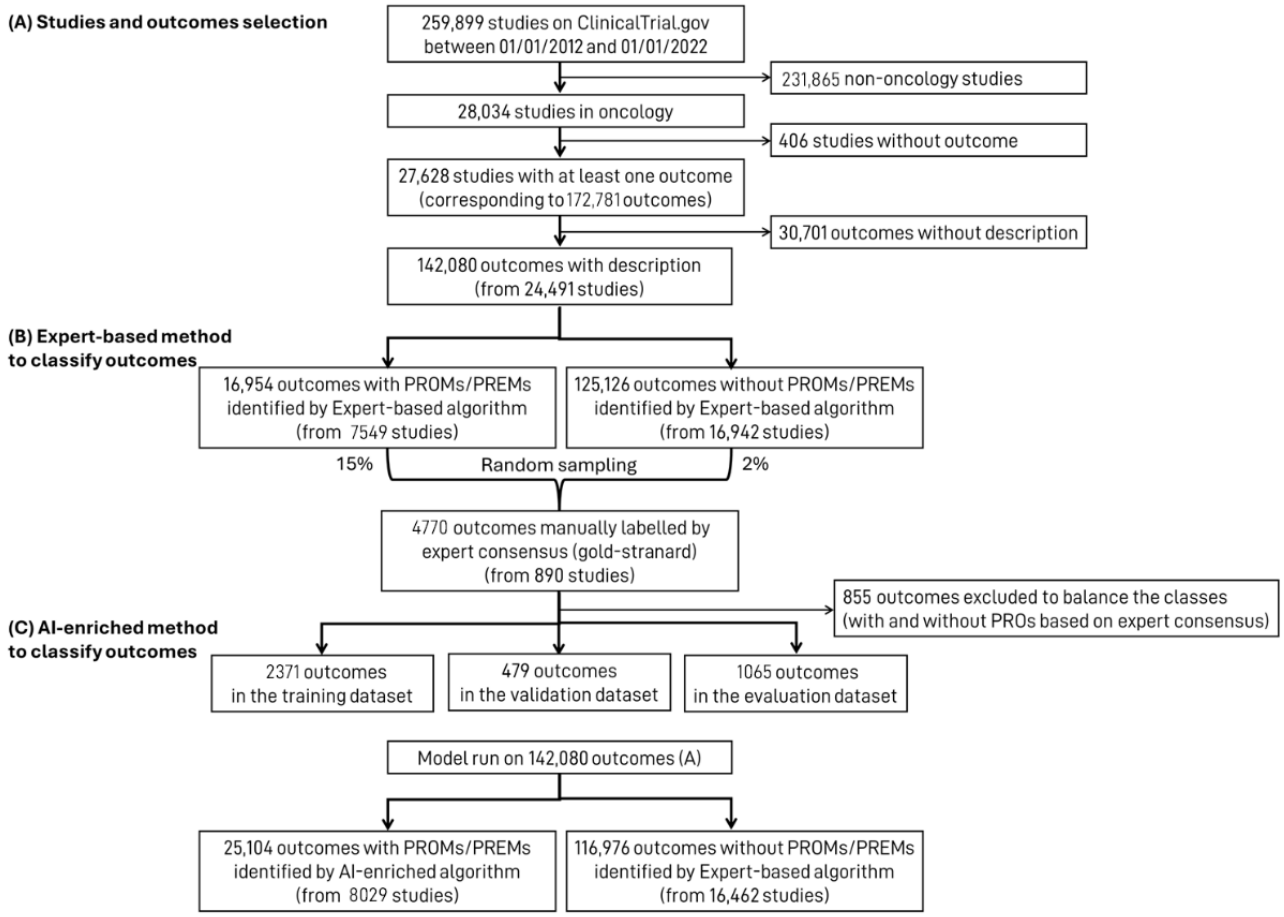

XSL•FO
RenderX

Table 1. Confusion matrix comparing the identification of patient-reported outcome measures (PROMs) or patient-reported experience measures (PREMs) in the outcomes of oncology studies, determined by either expert-based or artificial intelligence (AI)–enriched algorithms vs expert consensus, ClinicalTrials.gov database, 2012-2022 (n=1065 outcomes in the test subset)[a].

Type of algorithm and use of PROMs or PREMs in the study outcome	Expert consensus		PPV[b] (%)	NPV[c] (%)
	Yes (n=506)	No (n=559)		
Expert-based algorithm			80	88
Yes (n=558)	448	110		
No (n=507)	58	449		
AI-enriched algorithm			88	92
Yes (n=520)	460	60		
No (n=545)	46	499		

[a]This table shows the classification performances of the expert-based and AI-enriched models in predicting if study outcomes included PROMs or PREMs, compared to the expert consensus gold standard. “Yes” and “No” refer to whether outcomes included PROMs or PREMs. The expert-based algorithm achieved an accuracy of 84% (95% CI 82-86), a sensitivity of 89% (95% CI 85-91), and a specificity of 80% (95% CI 79-85), yielding a PPV of 80% and an NPV of 88%. The AI-enriched algorithm achieved an accuracy of 90% (95% CI 88-92), a sensitivity of 91% (95% CI 88-93), and a specificity of 89% (95% CI 86-92), yielding a PPV of 88% and an NPV of 92%.

[b]PPV: positive predictive value.

[c]NPV: negative predictive value.

The AI-enriched algorithm identified 8029 studies that used at least one PROM or PREM, representing 33% (Figure 1C). In the performance evaluation subset of 1065 outcomes, using the expert consensus manually labeled gold standard, the AI-enriched algorithm achieved 90% (95% CI 88%-92%) accuracy, 91% (95% CI 88%-93%) sensitivity, and 89% (95% CI 86%-92%) specificity (Table 1). Therefore, the AI-enriched algorithm performed better (chi-square test $P<.001$) than the expert-based algorithm in identifying PROMs and PREMs in oncology studies. The following results present the characteristics of studies identified based on the AI-enriched algorithm only. The results obtained using the expert-based algorithm are available in Multimedia Appendix 1.

Trends in PROMs and PREMs Use in Oncology Studies Over Time

The analysis of year-over-year trends revealed a steady increase in the proportion of clinical studies incorporating one or more PROMs or PREMs, from 27.6% in 2012 to 36.2% in 2021 (Figure S2 in Multimedia Appendix 1).

Characteristics of Studies

Table 2 summarizes the characteristics of the 24,491 studies included in the analysis, stratified on whether they used PROMs or PREMs (n=8029) or not (n=16,462), as identified by the AI-enriched algorithm. Among all the included studies (n=24,491), 79% (n=19,344) were interventional trials.

Focusing on these interventional trials, phase 2 (n=5228, 27%) was the most common study stage, followed by phase 1 (n=2536, 13.1%), phase 3 (n=1650, 8.5%), and phase 1/2 (n=1528, 7.9%) to report using PROMs or PREMs. However, missing data regarding trial phases accounted for 37.7% (n=7285) of studies (MCAR test indicated missing at random). The most frequent allocation type was randomization, accounting for 44.4% (n=8586) of interventional studies. The predominant intervention models were parallel assignment and single-group assignment, as they represented 48.2% (n=9330) and 44.5% (n=8613) of analyzed studies, respectively. The most common primary purpose of the included studies was treatment, accounting for 66.2% (n=12,815) of cases, followed by supportive care with 10.6% (n=2058). Among observational studies, the most prevalent observational model was cohorts, representing 64% (n=3295) of the studies included in the analysis. In addition, 72.3% (n=3295) of observational studies were prospective, making this more frequently used than perspective.

XSL•FO
RenderX

Table 2. Description of the characteristics of oncology studies according to their use of PROMs or PREMs classified by artificial intelligence–enriched algorithm, ClinicalTrials.gov database, 2012-2022 (N=24,491).

Characteristics	All (N=24,491)	Studies not using a PROM[a] or PREM[b] (n=16,462)	Studies that used PROMs or PREMs (n=8029)	*P* value
Study type				<.001
Interventional	19,344 (79)	12,544 (76.2)	6800 (84.7)	
Observational	5147 (21)	3918 (23.8)	1229 (15.3)	
Interventional studies	19,344	12,544	6800	__[c]
Clinical phase				<.001[d]
Early phase 1	376 (1.9)	314 (2.5)	62 (0.9)	
Phase 1	2536 (13.1)	2253 (18)	283 (4.2)	
Phase 1/2	1528 (7.9)	1279 (10.2)	249 (3.7)	
Phase 2	5228 (27)	3905 (31.1)	1323 (19.5)	
Phase 2/3	289 (1.5)	165 (1.3)	124 (1.8)	
Phase 3	1650 (8.5)	836 (6.7)	814 (12)	
Phase 4	452 (2.3)	274 (2.2)	178 (2.6)	
Missing data[e]	7285 (37.7)	3,518 (28)	3767 (55.4)	
Allocation type				<.001[d]
Nonrandomized	2543 (13.1)	2025 (16.1)	518 (7.6)	
Randomized	8586 (44.4)	4185 (33.4)	4401 (64.7)	
NA[f]	8215 (42.5)	6334 (50.5)	1881 (27.7)	
Interventional model				<.001[d]
Crossover assignment	374 (1.9)	206 (1.6)	168 (2.5)	
Factorial assignment	120 (0.6)	58 (0.5)	62 (0.9)	
Parallel assignment	9330 (48.2)	4910 (39.1)	420 (65)	
Sequential assignment	873 (4.5)	725 (5.8)	148 (2.2)	
Single group assignment	—	6619 (52.8)	1994 (29.3)	
Missing data	34 (0.2)	26 (0.2)	8 (0.1)	
Primary purpose				<.001[d]
Diagnostic	1573 (8.1)	1401 (11.1)	172 (2.5)	
Prevention	1031 (5.3)	561 (4.5)	470 (6.9)	
Supportive care	2058 (10.6)	420 (3.3)	1638 (24.1)	
Treatment	12,815 (66.2)	9026 (72)	3789 (55.7)	
Other	1787 (9.2)	1,082 (8.6)	705 (10.4)	
Missing data	80 (0.4)	54 (0.4)	26 (0.4)	
Observational studies	5147	3918	1229	—
Observational model				<.001[d]
Case-control	535 (10.4)	441 (11.3)	94 (7.6)	
Case-only	736 (14.3)	580 (14.8)	156 (12.7)	
Cohort	3295 (64)	2460 (62.8)	835 (67.9)	
Other	521 (10.1)	396 (10.1)	125 (10.2)	
Missing data	60 (1.2)	41 (1)	19 (1.5)	

XSL•FO
RenderX

Characteristics	All (N=24,491)	Studies not using a PROM[a] or PREM[b] (n=16,462)	Studies that used PROMs or PREMs (n=8029)	*P* value
Time perspective				<.001[d]
Cross-sectional	359 (7)	243 (6.2)	116 (9.4)	
Prospective	3723 (72.3)	2742 (70)	981 (79.8)	
Retrospective	836 (16.2)	753 (19.2)	83 (6.8)	
Other	204 (4)	161 (4.1)	43 (3.5)	
Missing data	25 (0.5)	19 (0.5)	6 (0.5)	

[a]PROM: patient-reported outcomes measure.

[b]PREM: patient-reported experience measure.

[c]Not available.

[d]*P* values do not account for missing data, as the data are censored.

[e]Missing data does not correspond to any clinical trial phase as per its definition (eg, medical devices).

[f]Trials classified as single group assignment and sequential assignment are grouped under "NA" for allocation type.

Table 2 presents the distribution of key study characteristics for all included studies (N=24,491), stratified by whether they used patient-reported outcome measures (PROMs) or patient-reported experience measures (PREMs). Characteristics are shown separately for interventional and observational studies, with proportions compared between studies using and not using PROMs or PREMs. Significant differences were found across study type, clinical phase, allocation method, interventional model, primary purpose, observational model, and time perspective (all *P*<.001). MCAR test using the RBtest R package indicated that data on the clinical phase were missing at random, while data for the interventional model, primary purpose, observational model, and time perspectives were missing completely at random.

Among all studies included in the analysis, 32.8% (n=8029) incorporated at least one PROM or PREM, as identified by the AI-enriched algorithm. Out of those, 84.7% (n=6800) were interventional, while 15.3% (n=1229) were observational. Among interventional studies, phase 2 and phase 3 trials incorporated mostly PROMs and PREMs, representing 19.5% (n=1323) and 12.0% (n=814) of them, respectively. Nevertheless, 55.4% (n=3767) of interventional studies did not provide any exploitable data on their clinical phases. Interventional studies incorporating PROMs and PREMs were mainly randomized (n=4401, 64.7%). The main interventional models were mainly parallel assignment and single-group assignment, as they represented 65% (n=4420) and 29.3% (n=1994), respectively, of studies with PROMs or PREMs. Treatment (n=3789, 55.7%) and supportive care (n=1638, 24.1%) were the most frequent primary purposes of interventional studies using PROMs or PREMs. Among observational studies using PROMs or PREMs (n=1229), 67.9% (n=835) were cohort studies, while 12.7% (n=156) were case-only studies. Considering time perspective, prospective studies were the most frequent to incorporate PROMs or PREMs, representing 79.8% (n=981) of them.

PROMs and PREMs Use by Sites of Cancer

Proportions of organ-specific oncology studies using at least one PROM or PREM are shown in Figure S3 in Multimedia Appendix 1. Breast and digestive cancers accounted for nearly 41% (2818/6918) of all studies using PROMs or PREMs in oncology. Urogenital, lung, and prostate cancers followed, representing 8% (579/6918), 10% (673/6918), and 11% (756/6918), respectively, of all studies using PROMs or PREMs in oncology. Proportions of organ-specific oncology studies using at least one PROM or PREM, as identified by the AI-enriched algorithm, are shown in Figure S3 in Multimedia Appendix 1.

PROMs and PREMs Used in Oncology Studies

Among the 8029 studies that included at least one PROM or PREM, as identified by the AI-enriched algorithm, 4823 (61.1%) used one or more of the 9 most commonly used PROMs or PREMs. Among these, the EORTC QLQ-C30 was the most frequently used questionnaire among PROMs and PREMs in clinical studies in oncology. It was used in approximately 33.2% (n=1602) of the studies included in our analysis. The FACT-G and the EQ-5D-5L questionnaires were the next most frequently used PROMs or PREMs, appearing in 15.2% (n=734) and 11.5% (n=552) of the studies analyzed, respectively (Table 3). Table 4 presents the categories of health dimensions assessed by the instruments identified across the 4823 studies that used one or more of the 9 most commonly used PROMs or PREMs, as identified by the AI-enriched algorithm. In 40.7% (n=1969) of the identified cases, the PROMs or PREMs focused on care-related patient satisfaction, followed by organ-specific quality of life (n=783, 16.2%), as well as depression and anxiety scales (n=441, 9.1%). This distribution results from the zero-shot, multilabel classification model, which often assigned satisfaction in overlap with the quality of life. The latter was further divided into general and cancer-specific subcategories. No additional filters based on study type or trial phase were applied to obtain these proportions.

XSL•FO
RenderX

Table 3. Instruments repartition among oncology studies that used one or more of the nine most commonly used patient-reported outcome measures or patient-reported experience measures, as identified by an artificial intelligence–enriched algorithm, ClinicalTrials.gov database, 2012-2022 (n=4823)[a].

Instruments	Value and proportion, n (%)
EORTC QLQ-C30[b]	1602 (33.2)
FACT-G[c]	734 (15.2)
Cancer-Related QoL[d]	416 (8.6)
HADS[e]	441 (9.1)
C-VAS[f]	263 (5.5)
NCI[g]	411 (8.5)
EQ-5D-5L	552 (11.5)
PRO-CTCAE[h]	292 (6.1)
IPSS[i]	112 (2.3)

[a]This table summarizes the frequency and proportion of studies that used at least one of the nine most frequently used patient-reported outcome measures (PROMs) or patient-reported experience measures (PREMs). The most commonly used instrument was the EORTC QLQ-C30 (33.2%), followed by FACT-G (15.2%) and EQ-5D-5L (11.5%). Less frequently used instruments included IPSS and C-VAS.

[b]EORTC QLQ-C30: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire-Core 30.

[c]FACT-G: Functional Assessment of Cancer Therapy – General.

[d]Cancer Related QoL: Cancer-Related Quality of Life.

[e]HADS: Hospital Anxiety and Depression Scale.

[f]C-VAS: Cancer Visual Analogue Scale.

[g]NCI: National Cancer Institute.

[h]PRO-CTCAE: Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events.

[i]IPSS: International Prostate Symptom Score.

Table 4. Categories of Health Dimensions Assessed by Instruments Identified among oncology studies that used one or more of the 9 most commonly used PROMs or PREMs, as identified by AI-enriched algorithm, ClinicalTrials.gov database, 2012-2022 (n=4823)[a].

Categories	Values and proportion, n (%)
Patient satisfaction with care	1969 (40.7)
Organ-specific quality of life	783 (16.2)
Depression and anxiety scale	441 (9.1)
NA	1640 (33.9)

[a]This table shows the distribution of health dimensions evaluated by instruments used in studies using one or more of the nine most frequently used PROMs and PREMs. The majority of instruments assessed patient satisfaction with care (40.7%), followed by organ-specific quality of life (16.2%) and depression and anxiety scales (9.1%). A notable proportion (33.9%) of assessments were classified as not applicable (NA).

Figure S4 in Multimedia Appendix 1 provides the distribution of the various scales commonly used in identified clinical studies, categorized by cancer site. Breast and digestive cancers were the most represented in the use of PROMs or PREMs in oncology studies, accounting for 22% (1521/6918) and 20% (1307/6918) of cases, respectively. These were followed by prostate (756/6918, 11%), lung (673/6918, 10%), urogenital (579/6918, 8%), and head and neck cancers (408/6918, 6%). In contrast, hematopoietic (193/6918, 3%), brain (118/6918, 3%), thyroid (57/6918, 1%), skin (52/6918, 1%), bone (118/6918, 1%), and neuroendocrine cancers (24/6918, 0.4%) were less frequently studied using PROMs or PREMs.

The use of PROMs and PREMs, as identified by the AI-enriched algorithm, varied according to several characteristics of studies (Figure S5A in Multimedia Appendix 1). Out of interventional studies, studies in phases 2 to 4 were incorporating PROMs or PREMs more often than preclinic or early phase 1 studies (OR 1.8, 95% CI 1.1-2.8 for phase 2; OR 3.6, 95% CI 2 .3-5.8 for phase 3; OR 2.6, 95% CI 1.6-4.4 for phase 4). Randomized clinical trials used twice as many PROMs or PREMs as nonrandomized trials (OR 2, 95% CI 1.7-2.4). Clinical trials focusing on supportive care demonstrated significantly higher odds for using PROMs or PREMs than trials with other objectives (OR 4.1, 95% CI 2.7-6.2). Studies concerning some organ-specific trials were less likely to use PROMs or PREMs than non–organ-specific ones—such as breast (OR 0.7, 95% CI 0.5-0.9), digestive (OR 0.6, 95% CI 0.5-0.8), lung (OR 0.7, 95% CI 0.4-0.7), hemopoietic (OR 0.5, 95% CI 0.3-0.8), and thyroid (OR 0.4, 95% CI 0.2-0.9) cancers. Similarly, the use of PROMs or PREMs instruments in observational studies differed according to characteristics of studies (Figure S5B in

XSL•FO
RenderX

Multimedia Appendix 1). In contrast to retrospective studies, cross-sectional and prospective studies used PROMs or PREMs more often (OR 4.6, 95% CI 3.3-6.4 and OR 3.2, 95% CI 2.5-4.1, respectively). After adjusting for confounding factors, cohort studies had significantly higher odds for using PROMs or PREMs than case-control studies (OR 1.6, 95% CI 1.3-2.1). Additionally, studies focusing on organ-specific trials, such as breast (OR 0.7, 95% CI 0.5-0.8), digestive (OR 0.4, 95% CI 0.3-0.5), head and neck (OR 0.6, 95% CI 0.4-0.8), hematopoietic (OR 0.6, 95% CI 0.4-0.8), lung (OR 0.4, 95% CI 0.3-0.5), thyroid (OR 0.5, 95% CI 0.3-0.9), and urogenital (OR 0.4, 95% CI 0.3-0.6) cancers, had lower odds for using PROMs or PREMs than non-organ-specific studies.

Discussion

Main Study Findings

This observational study focused on the ability of an AI-based tool to identify whether or not PROMs or PREMs were used in oncology trials from ClinicalTrials.gov, as compared to an expert-based gold standard. Taken together, our results on 24,491 studies clearly show that (1) an AI-enriched algorithm outperformed an expert-based one in identifying PROMs or PREMs in clinical studies, (2) the use of PROMs or PREMs has increased over time, (3) studies in breast cancer and digestive cancer accounted for nearly half of the studies incorporating PROMs or PREMs, and (4) interventional prospective studies were more likely to use PROMs or PREMs than any other type of studies.

Comparison With Prior Work and Implications

The AI-enriched algorithm that we developed for this specific purpose appears particularly promising for researchers seeking to identify both the presence of PROMs or PREMs and which PROM or PREM was used in oncology studies. With a 90% accuracy, we were able to better identify studies than experts would do manually and in a more efficient way: first, the manual checking of 4770 outcomes (accounting for 890 studies) took several weeks using a work force of 16 high-skilled people, while the algorithm was able to classify the whole 24,491 studies within a few minutes (from 10 minutes to 3 hours, depending on the GPU workforce). Although the development of the AI-enriched algorithm required a considerable initial investment in time and expertise, this effort represents a long-term return on investment, as the tool can subsequently process large datasets much faster than expert-based approaches. A significant increase in the use of PROMs or PREMs in phase 2 to 4 studies could be forecasted in the coming years, driven by the growing interest from key stakeholders, including industry, regulatory bodies, patients, and patient associations, as well as their commitment to promoting value-based health care (VBHC) in clinical practice. There is a significant and growing interest in the use of these tools across clinical research. Recently, regulatory agencies, such as the FDA and HAS, encourage the use of PROMs or PREMs when appropriate and have incorporated them into care quality and safety indicators, but their inclusion in clinical trials is not mandatory [19-21]. Additionally, the increasing incorporation of PROMs or PREMs in clinical studies that we report supports our optimism. The AI-enriched algorithm tool could be extended to other medical indications, fostering widespread research on PCM. On a larger scale, AI could become an invaluable tool for medical stakeholders to manage, conduct, and monitor clinical studies [27]. Nevertheless, the AI-enriched method also carries the risk of misclassifying conceptually close types of clinical outcome assessments (eg, observer-reported outcomes or clinician-reported outcomes) as PROMs or PREMs, potentially leading to an overestimation of their prevalence compared with the expert-based review. In contrast, the expert-based review offers the advantage of manual verification by trained reviewers, which helps ensure accuracy and contextual understanding, even if it is more resource-intensive.

New Findings in the Field of PROMs and PREMs

Among all organ-specific studies, PROMs and PREMs were more frequently used in studies focusing on breast and digestive cancers. This trend may be attributed to the extensive research on these cancers, driven by their high prevalence, significant impact on patients' quality of life, and the growing need for patient feedback as treatments evolve [28,29]. Furthermore, the diverse patient populations affected by breast and digestive cancers, spanning various age groups, genders, and ethnic backgrounds, underscore the importance of considering patients' perceived experiences and outcomes. While this is not unique to these cancer sites, breast and digestive cancers are among the most extensively studied in terms of patient-reported outcomes [30,31], likely due to their high prevalence and the strong presence of advocacy groups and research funding. This visibility has facilitated the integration of VBHC principles, where patient-centered care remains a central theme in medical research [32]. Compared to other cancer sites, breast and digestive cancers feature more prominently in VBHC-related literature, highlighting a gap that should be addressed for less-studied malignancies. It is also possible that in therapeutic areas where PROMs and PREMs are less common or less standardized, the relative performance gap between AI- and expert-based methods could be smaller, underscoring the importance of context when interpreting these findings. In our dataset, where PROMs and PREMs were present in 33% of oncology studies, the algorithm achieved strong sensitivity and specificity; however, because sensitivity and specificity depend on the prevalence of the measured outcome, its performance may be reduced in contexts where PROMs and PREMs are less frequently or less consistently reported. Further studies are therefore needed to assess the robustness and generalizability of AI-driven approaches across diverse conditions beyond the oncology field.

We also report that interventional or prospective studies incorporated more PROMs or PREMs than observational or retrospective studies. This trend could be attributed to the growing interest of regulatory bodies and sponsors (eg, industry, investigators, and promoters) in incorporating these measures into clinical trials, including regulators, such as the FDA and European Medicines Agency, as well as Health Technology Assessment bodies, in supporting approvals and decision-making. In such cases, some governments have begun recommending the use of PROMs or PREMs in interventional studies through official communications and guidelines [20,33].

XSL•FO
RenderX

Interventional studies are, by definition, designed with a structured protocol that allows for the systematic collection of specific data, making them particularly suited for incorporating PROMs or PREMs. Furthermore, since PROMs or PREMs provide direct patient insights, their inclusion in interventional studies is especially relevant, as these studies aim to evaluate the effects of specific treatments or interventions. Incorporating PROMs or PREMs into the design of observational studies requires their integration into routine care [34]. Another approach would involve leveraging PCMs in registries or other real-world data sources and incentivizing the use of PROMs or PREMs in funding and publishing by updating publication guidelines and funding requirements. Collaborative approaches, such as engaging patient advocacy groups and fostering multidisciplinary teams, may further enhance their adoption, as these groups often emphasize holistic, patient-centered care and bring together diverse clinical perspectives that recognize the value of patient-reported measures [35,36].

Strengths and Limitations

To address potential sources of bias, several measures were implemented. Selection bias was minimized by including all eligible oncology studies registered in ClinicalTrials.gov rather than a sample. Information bias related to outcome misclassification was reduced using a double-reference approach, where the AI-enriched algorithm was trained and evaluated against an expert-consensus gold standard. Misclassification due to heterogeneous naming of PROMs and PREMs was partially mitigated through expert review and consensus labeling during model development. Finally, analytical bias was limited by applying the same preprocessing pipeline and evaluation metrics across all studies. Additionally, the absence of prespecified subgroups or sensitivity analyses may limit the exploration of residual heterogeneity.

One important limitation of our AI-based tool is its risk of false positive or false negative identification of PROMs or PREMs. The model sometimes struggles with multiword PROMs or PREMs (see Multimedia Appendix 1 for examples). Even when a full name like the “Short Form 36 Health Survey” is right there in the text, the AI might break it into pieces, tagging “Short Form” and “Health Survey” as separate things instead of recognizing them as the single, standardized SF-36, which leads to the AI algorithm not recognizing it as a PROM. Moreover, uppercase abbreviations can be a “red herring” for the model. Since many PROMs and PREMs are abbreviated in all caps, the AI occasionally misidentifies general clinical terms (like “PAIN,” which is a symptom) as a PROM or PREM.

The increased use of PROMs and PREMs in oncology was also earlier reported by Vodicka et al [37], who aimed to estimate the proportion of clinical trials using at least one PROM or PREM from November 2007 to December 2013 [37]. They identified 29% of oncology trials using PROMs or PREMs, where we identified 33%. This could either be due to an improvement in the method for identifying studies or to an increase in the use of PROMs and PREMs over time. We used the same ClinicalTrials.gov database with a larger number of studies. The main difficulty in identifying PROMs and PREMs lay in the classification used in the ClinicalTrials.gov database. No standardized naming for PROMs and PREMs was currently available, leading researchers and HCPs to use the same PROMs or PREMs but with different names. This observation may explain the poor quality of data retrieved from studies used by our AI-enriched algorithm. Both algorithms (AI-enriched and expert-based ones) exhibited the same issue, but our AI-enriched algorithm remained more efficient than the expert-based one. Therefore, classification appeared to be the major risk of bias in our study. This risk of bias was not assessed in this study: future studies should test its clinical use, including whether the 10% of errors of this tool would be clinically meaningful or not. Both the trends in the use of PROMs and PREMs in oncology trials over time and their distribution exhibited performance rates below 100%, due to misclassification of cancers. One potential way to reduce this bias would be the implementation of an AI-based tool to support researchers in identifying, labeling, and naming PROs, thereby enhancing consistency in classification, terminology (names), and acronyms. Our study also experienced a limitation related to the dataset we used, especially regarding (1) the heterogeneity of the registries included, (2) the reliance on English-language parsing tools, and (3) the fact that our approach did not involve full-text scrutiny. Another limitation of our study is the lack of external validation; as we used the overall population of studies in the field of oncology from the ClinicalTrials.gov database, this would require the use of another dataset, either emerging from another field than oncology or from another database than ClinicalTrials.gov, for evaluating the generalizability of our findings. Such a study should also involve a large dataset. Finally, we chose to compare our AI-based algorithm to an expert-based algorithm against a human-based gold standard. This did not include any comparison with other algorithms using any machine learning tool, which is also a limitation. Future algorithms developed for this purpose may be tested against the one we report here.

Overall, within the oncology field, evidence suggests that routine collection of PROMs and PREMs enhances patient-provider communication, helps identify symptoms and/or concerns, and improves patient satisfaction [38]. However, there is currently limited evidence demonstrating their impact in oncological settings, particularly regarding an effective improvement of quality or a better health care system performance [39,40]. Therefore, further research is essential to address these gaps, assess the applicability of this approach across a broader range of pathologies, and evaluate its adaptability to additional open-access databases. Such investigations would help determine the generalizability and scalability of the methodology, facilitating its integration into diverse health care data systems and supporting its potential for widespread application in health data analysis and patient-centered research.

Importantly, this study is innovative because it combines expert-labeled data with artificial intelligence to automate the identification of PROMs and PREMs, which overcomes the limitations of previous approaches that relied solely on manual review or keyword-based searches. Unlike prior studies, our AI-enriched method allows large-scale, reproducible monitoring of patient-reported measures across thousands of trials, providing a clear contribution to the field by quantifying trends,

XSL•FO
RenderX

highlighting gaps, and identifying areas for improvement in patient-centered data collection.

This approach can support researchers, regulators, sponsors, and health care systems in better designing trials, in more easily tracking the integration of PROMs and PREMs, in better informing funding and policy decisions, and ultimately in promoting value-based, patient-centered care.

Conclusions

In conclusion, our data clearly demonstrate that an AI-enriched algorithm is more accurate, in terms of sensitivity and specificity, and faster than traditional expert-based methods for identifying PROMs and PREMs in oncology clinical studies. PROMs and PREMs usage is increasing over time, particularly in interventional and prospective studies. This approach is innovative because it combines expert-labeled data with AI, enabling scalable and reproducible identification of PROMs and PREMs, unlike previous manual or expert-based methods. We suggest that improving the naming and classification of PROMs and PREMs could further facilitate their identification and use in studies, for example, through the development of an AI-based tool to assist researchers in identifying, labeling, and naming patient-reported outcomes. Future studies should extend this methodology to other pathologies and databases, enhancing its use and supporting broader integration of patient-reported measures in clinical research.

Acknowledgments

The authors thank all the physicians, researchers, and patients using patient-reported outcome measures (PROMs) and patient-reported experience measures (PREMs) in clinical research, as well as all those who develop and validate patient-reported measures. The authors would also like to warmly thank Dr Aline Thomas for providing tremendous help in preparing this manuscript. Finally, we thank the researchers from the EPICLIN congress, where a preliminary version of these results was previously published as an abstract in the conference proceedings [41]. No generative artificial intelligence was used for writing the manuscript.

Data Availability

All databases are available upon reasonable request to the corresponding author.

Funding

No specific funding was received for this study. All authors were employed by either SKEZI or Mapi Research Trust at the time they conducted the study.

Authors' Contributions

JS performed the analysis, wrote the first draft of the manuscript, and revised the manuscript. AH performed the analysis, wrote the first draft of the manuscript, and revised the manuscript. SJ designed the study and revised the manuscript. CD-G provided the list of PROM- and PREM-related terms using the PROQOLID (Mapi Research Trust) database. MT performed the analysis and the data layout, wrote further drafts of the manuscript, revised the manuscript, and wrote Multimedia Appendix 1. J-PB designed the study, supervised the study, wrote the first draft of the manuscript, and revised the manuscript.

Conflicts of Interest

JS, AH, SJ, MT, and J-PB work at SKEZI, a company that develops tools for improving the use of patient-reported measures. CD-G works at Mapi Research Trust, a company that licenses PROMs and PREMs.

Multimedia Appendix 1
STROBE checklist, supplementary figures on patient-reported outcome measure and patient-reported experience measure use in oncology (2012–2022), and missing error analysis.
[DOCX File , 1565 KB - jmir_v28i1e84533_app1.docx]

References

1. Marshall S, Haywood K, Fitzpatrick R. Impact of patient-reported outcome measures on routine practice: a structured review. J Eval Clin Pract 2006;12(5):559-568. [doi: 10.1111/j.1365-2753.2006.00650.x] [Medline: 16987118]
2. Howard AF, Warner L, Cuthbertson L, Sawatzky R. Patient-driven research priorities for patient-centered measurement. BMC Health Serv Res 2024;24(1):735 [FREE Full text] [doi: 10.1186/s12913-024-11182-x] [Medline: 38877556]
3. Churruca K, Pomare C, Ellis LA, Long JC, Henderson SB, Murphy LED, et al. Patient-reported outcome measures (PROMs): a review of generic and condition-specific measures and a discussion of trends and issues. Health Expect 2021 Aug;24(4):1015-1024 [FREE Full text] [doi: 10.1111/hex.13254] [Medline: 33949755]
4. Wolff AC, Dresselhuis A, Hejazi S, Dixon D, Gibson D, Howard AF, et al. Healthcare provider characteristics that influence the implementation of individual-level patient-centered outcome measure (PROM) and patient-reported experience measure

XSL•FO
RenderX

(PREM) data across practice settings: a protocol for a mixed methods systematic review with a narrative synthesis. Syst Rev 2021 Jun 09;10(1):169 [FREE Full text] [doi: 10.1186/s13643-021-01725-2] [Medline: 34108024]
5. Sawatzky R, Laforest E, Schick-Makaroff K, Stajduhar K, Reimer-Kirkham S, Krawczyk M, et al. Design and introduction of a quality of life assessment and practice support system: perspectives from palliative care settings. J Patient Rep Outcomes 2017;2:36. [doi: 10.1186/s41687-018-0065-2] [Medline: 30175318]
6. Lavallee DC, Chenok KE, Love RM, Petersen C, Holve E, Segal CD, et al. Incorporating patient-reported outcomes into health care to engage patients and enhance care. Health Aff (Millwood) 2016 Apr;35(4):575-582. [doi: 10.1377/hlthaff.2015.1362] [Medline: 27044954]
7. FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) Resource. Silver Spring, MD: Food and Drug Administration (US); 2016.
8. Le Corroller AG, Bonastre J. Patient-reported measures: how useful in health economics? Eur J Health Econ 2023;24(1):1-4. [doi: 10.1007/s10198-022-01524-z] [Medline: 36346476]
9. Dawson J, Doll H, Fitzpatrick R, Jenkinson C, Carr AJ. The routine use of patient reported outcome measures in healthcare settings. BMJ 2010;340:c186. [doi: 10.1136/bmj.c186] [Medline: 20083546]
10. Chen J, Ou L, Hollis SJ. A systematic review of the impact of routine collection of patient reported outcome measures on patients, providers and health organisations in an oncologic setting. BMC Health Serv Res 2013;13:211 [FREE Full text] [doi: 10.1186/1472-6963-13-211] [Medline: 23758898]
11. Price A, Smith J, Dakin H, Kang S, Eibich P, Cook J, et al. The arthroplasty candidacy help engine tool to select candidates for hip and knee replacement surgery: development and economic modelling. Health Technol Assess 2019;23(32):1-216 [FREE Full text] [doi: 10.3310/hta23320] [Medline: 31287051]
12. van Hout B, Janssen MF, Feng YS, Kohlmann T, Busschbach J, Golicki D, et al. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. Value Health 2012;15(5):708-715 [FREE Full text] [doi: 10.1016/j.jval.2012.02.008] [Medline: 22867780]
13. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011 Dec;20(10):1727-1736 [FREE Full text] [doi: 10.1007/s11136-011-9903-x] [Medline: 21479777]
14. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993 Mar 03;85(5):365-376. [doi: 10.1093/jnci/85.5.365] [Medline: 8433390]
15. Ruotolo I, Carenza A, Sellitto G, Simeon R, Panuccio F, Marini Padovani A, et al. Exploring the psychometric properties of the fatigue severity scale: results from a systematic review and reliability meta-analysis. Expert Rev Pharmacoecon Outcomes Res 2025 Sep;25(7):1021-1036. [doi: 10.1080/14737167.2025.2511128] [Medline: 40408308]
16. Bergman B, Aaronson NK, Ahmedzai S, Kaasa S, Sullivan M. The EORTC QLQ-LC13: a modular supplement to the EORTC Core Quality of Life Questionnaire (QLQ-C30) for use in lung cancer clinical trials. EORTC study group on quality of life. Eur J Cancer 1994;30A(5):635-642. [doi: 10.1016/0959-8049(94)90535-5] [Medline: 8080679]
17. Noel CW, Stephens RF, Su JS, Xu W, Krahn M, Monteiro E, et al. Mapping the EORTC QLQ-C30 and QLQ-HandN35, onto EQ-5D-5L and HUI-3 indices in patients with head and neck cancer. Head Neck 2020 Sep;42(9):2277-2286. [doi: 10.1002/hed.26181] [Medline: 32333829]
18. Hawkins J, Smeeton N, Busby A, Wellsted D, Rider B, Jones J, et al. Contributions of treatment centre and patient characteristics to patient-reported experience of haemodialysis: a national cross-sectional study. BMJ Open 2021 Apr 14;11(4):e044984 [FREE Full text] [doi: 10.1136/bmjopen-2020-044984] [Medline: 33853800]
19. Leandre. Aide à l'Utilisation de Questionnaires Patients de Mesure Des Résultats De Soins Dans Le Cadre de l'Expérimentation «Épisode de Soins». Saint-Denis-La-Plaine: Haute Autorité de santé (HAS); 2019.
20. Quality of care as perceived by patients - PROMs and PREMs indicators panorama of experiences in other countries and main lessons learned. Haute Autorité de Santé. 2021. URL: https://www.has-sante.fr/upload/docs/application/pdf/2022-03/proms_prems_panorama_of_experiences_in_other_countries_vf.pdf [accessed 2026-03-05]
21. U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Biologics Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Devices and Radiological Health. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. Health Qual Life Outcomes 2006;4:79 [FREE Full text] [doi: 10.1186/1477-7525-4-79] [Medline: 17034633]
22. Lu Z, Sim JA, Wang JX, Forrest CB, Krull KR, Srivastava D, et al. Natural language processing and machine learning methods to characterize unstructured patient-reported outcomes: validation study. J Med Internet Res 2021 Nov 03;23(11):e26777 [FREE Full text] [doi: 10.2196/26777] [Medline: 34730546]
23. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. 2018 Oct 11. URL: https://arxiv.org/abs/1810.04805v2 [accessed 2025-04-02]
24. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: state-of-the-art natural language processing. ACL Anthology 2020 Oct:38-45. [doi: 10.18653/v1/2020.emnlp-demos.6]

XSL•FO
RenderX

25. Hartigan JA, Wong MA. Algorithm AS 136: a K-means clustering algorithm. Appl Stat 1979;28(1):100-108. [doi: 10.2307/2346830]
26. Tunstall L, Reimers N, Jo UES, Bates L, Korat D, Wasserblat M, et al. Efficient few-shot learning without prompts. arXiv. Preprint posted online on September 22, 2022 2022 Sep 22. [doi: 10.48550/arXiv.2209.11055]
27. Salvagno M, Cassai AD, Zorzi S, Zaccarelli M, Pasetto M, Sterchele ED, et al. The state of artificial intelligence in medical research: a survey of corresponding authors from top medical journals. PLoS One 2024;19(8):e0309208 [FREE Full text] [doi: 10.1371/journal.pone.0309208] [Medline: 39178224]
28. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74(3):229-263 [FREE Full text] [doi: 10.3322/caac.21834] [Medline: 38572751]
29. Arnold M, Abnet CC, Neale RE, Vignat J, Giovannucci EL, McGlynn KA, et al. Global burden of 5 major types of gastrointestinal cancer. Gastroenterology 2020 Jul;159(1):335-349.e15 [FREE Full text] [doi: 10.1053/j.gastro.2020.02.068] [Medline: 32247694]
30. Rutherford C, Campbell R, White K, King M. Patient-reported outcomes as predictors of survival in patients with bowel cancer: a systematic review. Qual Life Res 2019;28(11):2871-2887. [doi: 10.1007/s11136-019-02255-0] [Medline: 31364037]
31. Salmani H, Nasiri S, Alemrajabi M, Ahmadi M. Advancing patient-centered cancer care: a systematic review of electronic patient-reported outcome measures. Front Rehabil Sci 2024;5:1427712 [FREE Full text] [doi: 10.3389/fresc.2024.1427712] [Medline: 39386354]
32. Demedts I, Himpe U, Bossuyt J, Anthoons G, Bode H, Bouckaert B, et al. Clinical implementation of value based healthcare: impact on outcomes for lung cancer patients. Lung Cancer 2021 Dec;162:90-95. [doi: 10.1016/j.lungcan.2021.10.010] [Medline: 34763159]
33. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims. US Food and Drug Administration. 2009. URL: https://www.fda.gov/media/77832/download [accessed 2026-03-05]
34. Horn ME, Reinke EK, Mather RC, O'Donnell JD, George SZ. Electronic health record-integrated approach for collection of patient-reported outcome measures: a retrospective evaluation. BMC Health Serv Res 2021;21(1):626 [FREE Full text] [doi: 10.1186/s12913-021-06626-7] [Medline: 34193125]
35. Calvert M, Kyte D, Price G, Valderas JM, Hjollund NH. Maximising the impact of patient reported outcome assessment for patients and society. BMJ 2019;364:k5267. [doi: 10.1136/bmj.k5267] [Medline: 30679170]
36. Al Sayah F, Short H, Ohinmaa A, Dean S, Hubley B, Lahtinen M, et al. Stakeholders' perspectives on implementing and integrating patient-reported outcome measures (PROMs) in health systems - insights from Alberta, Canada. J Patient Rep Outcomes 2025;9(1):63 [FREE Full text] [doi: 10.1186/s41687-025-00887-0] [Medline: 40471481]
37. Vodicka E, Kim K, Devine EB, Gnanasakthy A, Scoggins JF, Patrick DL. Inclusion of patient-reported outcome measures in registered clinical trials: evidence from clinicalTrials.gov (2007-2013). Contemp Clin Trials 2015;43:1-9. [doi: 10.1016/j.cct.2015.04.004] [Medline: 25896116]
38. Graupner C, Kimman ML, Mul S, Slok AHM, Claessens D, Kleijnen J, et al. Patient outcomes, patient experiences and process indicators associated with the routine use of patient-reported outcome measures (PROMs) in cancer care: a systematic review. Support Care Cancer 2021 Mar;29(2):573-593 [FREE Full text] [doi: 10.1007/s00520-020-05695-4] [Medline: 32875373]
39. Chen J, Ou L, Hollis SJ. A systematic review of the impact of routine collection of patient reported outcome measures on patients, providers and health organisations in an oncologic setting. BMC Health Serv Res 2013;13:211 [FREE Full text] [doi: 10.1186/1472-6963-13-211] [Medline: 23758898]
40. Howell D, Molloy S, Wilkinson K, Green E, Orchard K, Wang K, et al. Patient-reported outcomes in routine cancer clinical practice: a scoping review of use, impact on health outcomes, and implementation factors. Ann Oncol 2015;26(9):1846-1858 [FREE Full text] [doi: 10.1093/annonc/mdv181] [Medline: 25888610]
41. Soyer J, Hecini A, Juchet S, Desvignes-Gleizes C, Bertocchio JP. P51 - artificial intelligence (AI)-empowered expertise: a real asset in identifying patient-reported data (PROMs and PREMs) in clinical studies. J Epidemiol Population Health 2025 May;73:203082. [doi: 10.1016/j.jeph.2025.203082]

Abbreviations

AI: artificial intelligence
BERT: bidirectional encoder representations from transformers
EORTC: European Organization for Research and Treatment of Cancer
FDA: US Food and Drug Administration
FSS: Fatigue Severity Scale
HAS: Haute Autorité de Santé - French Authority of Health
HCP: health care provider
MCAR: missing completely at random
NER: named entity recognition

XSL•FO
RenderX

OR: odds ratio
PCM: patient-centered measure
PREM: patient-reported experience measure
PROM: patient-reported outcome measure
PROQOLID: Patient-Reported Outcome and Quality of Life Instruments Database
QLQ: Quality of Life Questionnaire
QLQ-C30: Quality of Life Questionnaire-Core 30
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
VBHC: value-based health care

Edited by S Brini; submitted 21.Sep.2025; peer-reviewed by J Odeniya, A Mettupatti Sivagnanam, RQ Wang, AT Thomas, O Enaworu, A Kodan; comments to author 21.Oct.2025; revised version received 12.Feb.2026; accepted 14.Feb.2026; published 16.Apr.2026.

Please cite as:
Soyer J, Hecini A, Juchet S, Desvignes-Gleizes C, Thiebaut M, Bertocchio JP
Artificial Intelligence for Identifying Patient-Reported Outcome and Experience Measures in Oncology: Retrospective Cross-Sectional Study Using ClinicalTrials.gov
J Med Internet Res 2026;28:e84533
URL: https://www.jmir.org/2026/1/e84533
doi:10.2196/84533
PMID:15886485

©Jessica Soyer, Akram Hecini, Sylvain Juchet, Céline Desvignes-Gleizes, Maxime Thiebaut, Jean-Philippe Bertocchio. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Cognitive Dissonance–Based Priming Intervention: Randomized Encouragement With in-the-Wild Phishing Simulation Attack in Health Care

Prosper Kandabongee Yeng[1], PhD; Muhammad Ali Fauzi[2], PhD; Arnstein Vestad[3], PhD; Bian Yang[3], PhD; Katrien De Moor[4], PhD; Christian Jacobsen[5]; John-Bosco Diekuu[6], PhD; Meriem Bettayeb[1], PhD

[1]Department of Computer Science and IT, College of Engineering, Abu Dhabi University, Al Ain, Abu Dhabi, United Arab Emirates

[2]Department of Informatics Engineering, Faculty of Computer Science, University of Brawijaya, Malang, Indonesia

[3]Department of Information Security and Communication Technology, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology (NTNU), Gjøvik, Norway

[4]Department of Information Security and Communication Technology, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

[5]Department of Security and Risk Governance, Aidn AS, Oslo, Norway

[6]Department of Machine Learning and Computer Vision, School of Computing, Engineering and Technology, Robert Gordon University, Aberdeen, United Kingdom

Corresponding Author:
Prosper Kandabongee Yeng, PhD
Department of Computer Science and IT, College of Engineering, Abu Dhabi University, Al Ain, Abu Dhabi, United Arab Emirates

Abstract

Background: Phishing remains a dominant initial attack vector in health care, exploiting psychological factors such as urgency and authority. Despite extensive investment in technical controls and awareness training, health care staff remain highly susceptible in real operational conditions. Cognitive dissonance (CD), the discomfort arising from inconsistencies between beliefs and actions, has been proposed as a mechanism to disrupt unsafe rationalization at the moment of exposure, but has rarely been evaluated in live organizational settings using objective behavioral outcomes.

Objective: This study examined whether a brief CD-based priming intervention, delivered immediately prior to a real-world phishing simulation, was associated with differences in phishing susceptibility among health care staff. Secondary objectives explored whether CD exposure was associated with directional differences in security-related perceptions and self-reported practices.

Methods: A 2-stage hybrid randomized-encouragement experiment was conducted at a large Norwegian hospital. In Stage 1, staff were randomly assigned to a control or CD-primed condition and completed a survey assessing security perceptions and self-reported practices (n=62). In Stage 2, an in-the-wild phishing simulation was sent to all staff, enabling objective measurement of phishing susceptibility via observed link-click behavior. Behavioral outcomes were analyzed across 3 groups—control (n=34), CD-primed (n=32), and neutral nonresponders (n=753)—using a prespecified omnibus chi-square test as the sole confirmatory analysis. Survey-based multivariate and univariate analyses were treated as exploratory due to limited sample size and variable construct reliability.

Results: Due to voluntary uptake, only a subset of randomized participants received the intervention. Observed phishing click rates were 65% (22/34) in the control group, 44% (14/32) in the CD-primed group, and 53% (396/753) in the neutral group. The omnibus chi-square test did not detect a statistically significant association between group membership and click behavior (χ^2_2=3.00; n=819; P=.22; Cramér V=0.06). Descriptive comparisons within the randomized subset suggested lower click rates in the CD-primed group, but effect estimates were imprecise and associated with wide CIs. Survey-based analyses indicated group differences across combined psychological constructs; however, several constructs exhibited low internal consistency, and follow-up analyses were underpowered.

Conclusions: In a real-world hospital phishing simulation, pre-exposure CD priming was associated with a directional but statistically nonsignificant pattern of reduced phishing click behavior. This evidence does not establish a reliable behavioral effect, and construct-level findings are exploratory. CD-based prompts may serve as a lightweight behavioral signal in real-world conditions, but larger, fully randomized, and longitudinal studies with improved psychometric validation are needed before such interventions can be considered reliable complements to established cybersecurity controls.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e68051) doi:10.2196/68051

KEYWORDS

cognitive dissonance; phishing simulation; health care; randomized encouragement; psychological incentive; health belief model; protection motivation theory; cognition; phishing; phishing attacks; cybersecurity; phishing attempts; health care staff; security; Norway

Introduction

Background

Phishing remains the dominant initial attack vector in health care, accounting for a substantial proportion of breaches between 2022 and 2025 through credential theft, business email compromise, and ransomware deployment [1-3]. Despite significant investment in email filtering and endpoint defenses, health care employees continue to be exposed to highly contextualized phishing campaigns that exploit clinical workflows, trust relationships, professional hierarchies, and time pressure [2-6]. Consequently, health care has become one of the costliest sectors for data breaches, with human-driven social engineering outweighing purely technical exploitation [2,2,4,7-9]. These attacks not only compromise data confidentiality but also disrupt care delivery, delay clinical workflows, and threaten patient safety [9-11].

To explain persistent phishing susceptibility, prior research has increasingly applied psychological frameworks such as the Health Belief Model (HBM) and Protection Motivation Theory (PMT) [7,10,12,13]. These models examine how perceived severity (PS), perceived vulnerability (PV), self-efficacy (SE), and response efficacy (RE) influence individual security behavior. However, most health care phishing studies rely on self-reported intentions, survey-based perceptions, or post-hoc evaluations conducted in training contexts rather than during real operational attacks [4-6,14-16]. As a result, existing interventions rarely disrupt unsafe cognition at the moment of threat exposure, nor do they validate effectiveness using objective behavioral outcomes [3,4,11,14].

The objectives of this study are therefore outlined as follows:

- Delivering the first cognitive dissonance (CD)–based prephishing intervention tested in a live hospital;
- Experimentally linking CD to shifts in HBM or PMT perceptions;
- Measuring real click behavior during an in-the-wild phishing event;
- Demonstrating a theory-driven, low-cost behavioral security control with ecological validity.

CD-Based in-the-Wild Phishing Intervention

CD theory explains how individuals experience psychological discomfort when their beliefs conflict with their actions, such as recognizing phishing risks while still clicking urgent or authoritative emails [17-19]. To reduce this discomfort, individuals often rationalize risky behavior through neutralization techniques, including denial of harm, denial of responsibility, or appeals to higher organizational goals [11,20-22]. In health care contexts, such rationalizations have been observed in email-policy violations, unsafe password practices, and delayed incident reporting justified by clinical urgency [4,10,10,23].

Phishing attackers deliberately exploit these same psychological mechanisms by embedding urgency, authority, fear, and trust cues into realistic clinical narratives, thereby intensifying cognitive conflict and impairing reflective judgment [3,6,23,24]. Recent reviews emphasize that such human-factor vulnerabilities remain inadequately addressed by purely technical controls and awareness-only training approaches [8,9,13,25].

This study introduces a CD-based priming intervention delivered immediately before an in-the-wild phishing simulation at one of Norway's largest hospitals. Unlike traditional training, deterrence-based sanctions, or postincident awareness programs [11,13-15,25], CD priming operates as a proactive, preattack behavioral control that directly targets rationalization before exposure. The intervention is lightweight, theory-driven, and embedded within routine organizational communication, making it suitable for real clinical environments characterized by time pressure and high cognitive load [3,3,9].

Methodologically, the study advances prior work in 3 ways. First, it deploys CD priming immediately prior to a live operational phishing event rather than within simulated, laboratory, or training-only contexts [5,11,14,26]. Second, it integrates CD with established HBM and PMT constructs to examine shifts in theory-driven security perceptions [7,10,12,13,15]. Third, it validates intervention effectiveness using objective behavioral outcomes—actual link-click behavior—while minimizing ethical risk by avoiding credential harvesting and excessive deception [27,28]. To our knowledge, no prior health care study has combined CD-based priming, health-behavior theory, and in-the-wild behavioral measurement within a single experimental design [3,9,29].

Hypothesis Formulation and Contributions

The primary objective of this study was to examine whether exposure to a CD-based priming intervention was associated with differences in observed phishing susceptibility during a real-world phishing simulation (hypothesis 1). Secondary analyses explored whether CD exposure was associated with directional differences in psychological security perceptions and self-reported security practices.

The following hypothesis and research questions were therefore specified:

Hypothesis 1 (Confirmatory)

Exposure to a CD-based priming intervention is associated with differences in observed phishing susceptibility, measured by link-click behavior during an in-the-wild phishing simulation.

XSL•FO
RenderX

Research Question (RQ) 1 (Exploratory)

Is exposure to a CD-based priming message associated with directional differences in selected security-related perceptions derived from the Health Belief Model (HBM) and Protection Motivation Theory (PMT), including PV, PS, SE, RE, perceived barriers (PBs), and cues to action (CA)?

RQ2 (Exploratory)

Is exposure to a CD-based priming message associated with directional differences in self-reported security practices related to password management, incident reporting, email handling, and mobile-device security?

Methods

Overview

This study used a 2-stage randomized encouragement design [30,31], involving approximately 830 health care staff who were randomly assigned at the invitation stage to receive either a control or a CD-primed questionnaire [27,32,33]. Only a subset of invited staff completed the questionnaire and were therefore exposed to the intervention (control: n=42; CD-primed: n=40), while nonresponse was treated as nonreceipt rather than exclusion, consistent with established methodological guidance [34]. Randomization, therefore, applies only to encouragement and survey exposure, not to subsequent receipt of the phishing email or behavioral outcome measurement.

In Stage 2, an in-the-wild phishing simulation [4] message (as shown in Figure 1) was sent to all staff with valid email addresses, enabling objective measurement of phishing click behavior under real-world conditions. Primary outcome analyses for the randomized component were conducted among questionnaire respondents who received valid phishing emails. These participants were the only ones plausibly exposed to both the intervention and the behavioral outcome. In parallel, a large Neutral group comprising staff who did not complete the questionnaire but received the in-the-wild phishing email was included as an observational comparison. Analyses involving this group are reported descriptively and interpreted without causal inference in accordance with STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [35], while the randomized encouragement component is reported following CONSORT (Consolidated Standards of Reporting Trials) principles [36], ensuring transparent reporting of participant flow, uptake, and inferential boundaries.

Figure 1. Phishing attack simulation message.

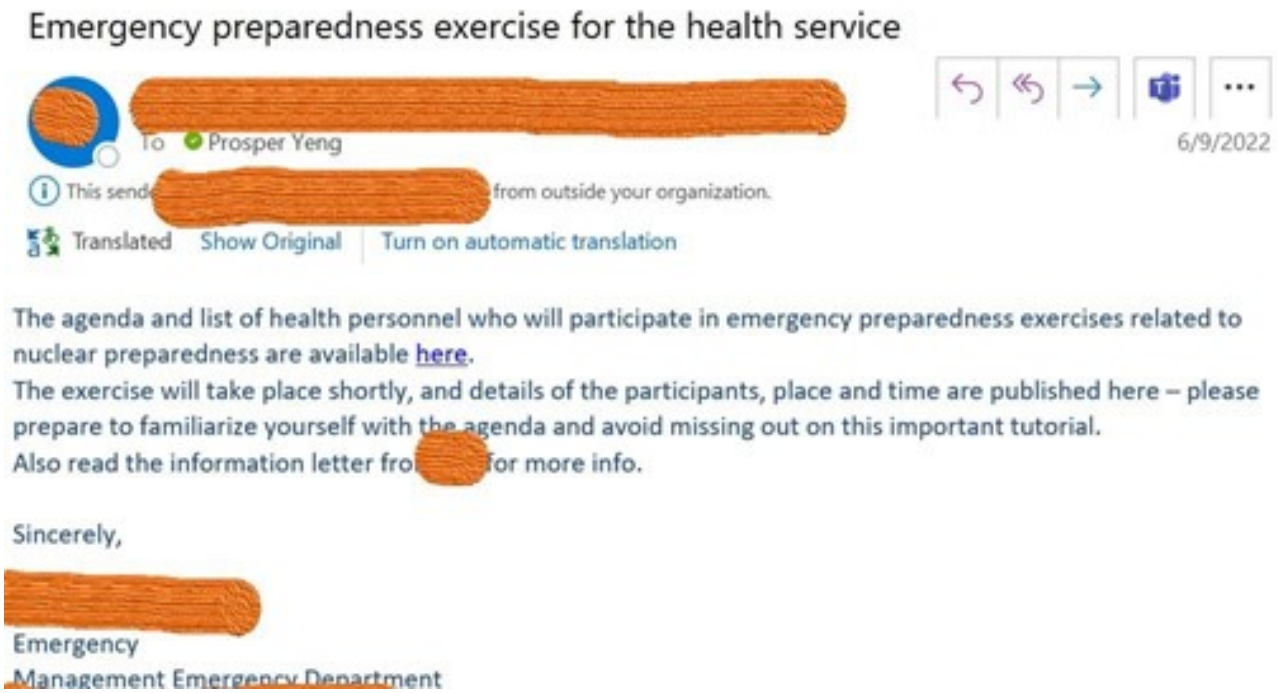

Behavioral outcomes were assessed independently of survey participation. Valid delivery and outcome records were available for 819 participants, who formed the sample used for behavioral analysis. These participants fell into 3 groups based on their prior exposure to the CD message:

- Control group: survey respondents who did not receive the CD message and had valid phishing-outcome data (n=34).
- Experiment group: survey respondents who received the CD message and had valid phishing-outcome data (n=32). The architecture is shown in Figure 2.
- Neutral group: all remaining staff who did not receive any CD-related priming but successfully received the phishing email (n=753).

XSL•FO
RenderX

Figure 2. Experiment model for psychological incentive.

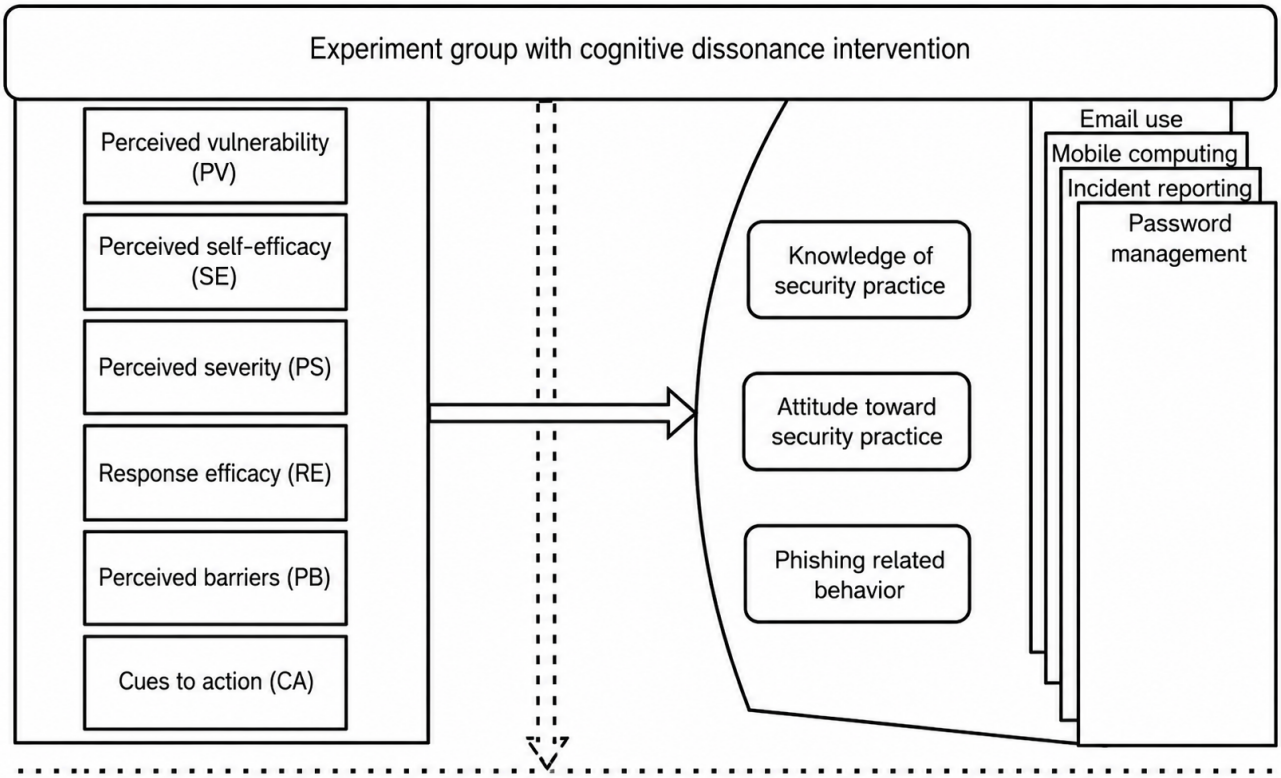

The difference between the invited population (~830) and the Stage 2 analyzed sample (819) reflects staff with undeliverable emails. This grouping structure enabled the study to (1) evaluate the causal effect of the CD intervention on psychological and self-reported outcomes among survey respondents and (2) compare their actual phishing susceptibility against a large neutral group not exposed to any experimental priming.

As an incentive, participants in both stages were offered a free lunch at the hospital canteen and also entered into a draw in which 10 individuals could win a US $50 gift card [37,38].

Behavioral Outcome: Phishing Click Susceptibility

Actual behavioral phishing susceptibility was measured by recording objective link-click behavior during the in-the-wild phishing simulation and coding it as a binary outcome (clicked vs not clicked) across 3 groups, including control (n=34), CD-primed (n=32), and neutral (n=753).

The primary, prespecified analysis was an omnibus Pearson chi-square test of independence (3×2 contingency table; 2-sided α=.05). While this test was designated a priori as the primary behavioral analysis, causal inference is limited to the randomized comparison between the control and CD-primed groups; comparisons involving the neutral group are observational.

Effect size was quantified using Cohen w and Cramér V, and group-specific click rates were reported with 95% CIs (Wilson method) [39]. Pairwise comparisons between the CD-primed and control groups were reported descriptively using risk difference, relative risk, and odds ratio with 95% CIs and treated as exploratory due to limited sample size [40].

A post hoc power assessment for the omnibus test was reported descriptively to characterize statistical sensitivity, rather than to support inferential claims. All analyses were conducted using the statsmodels Python (Python Software Foundation) library [16,41].

Ethical, Privacy, and Security Measures With an in-the-Wild Study

In this experiment, a questionnaire and an in-the-wild study were combined. The hybrid approach was essential, as the survey provided a basis for the researchers to understand participants' intended phishing security behavior, while the in-the-wild study tested participants' actual phishing security practice (ie, the clicking action). An in-the-wild study is a type of phishing simulation in which the researcher conducts a phishing attack on participants in their natural working environment without prior warning to observe real-world security behavior under realistic conditions. Unlike laboratory-based or survey-based studies, in-the-wild studies capture authentic user responses but raise additional ethical and methodological considerations [4,27].

Survey Instrument

In the first section of the questionnaire, we adopted the approach of Parsons et al [42], the human aspect of the information security questionnaire. This consists of 42 items that measure knowledge, attitude, and behavior (KAB) regarding phishing, as well as risk factors associated with password management, email use, incident reporting, and mobile device use. A 5-point Likert scale was used in this instrument. These areas of security practice are considered more vulnerable to phishing attacks. Additionally, approximately 25 items in this instrument measured other psychological constructs, including PV, SE, CA, PB, and RE. Figure 2 shows the questionnaire structure. The instrument was developed with an online, secure Norwegian

XSL•FO
RenderX

version of a survey system called Nettskjema [43]. The questionnaire was then divided into 2 groups, a control group and the experiment group. The difference between them was that the experiment group questionnaire included a CD message, as shown in Multimedia Appendix 1, and in the model in Figure 3. Participants also indicated in the questionnaire if they believe in the CD message as shown in Figure 4. The questionnaire for the control group did not include the CD message item. The instrument was then pretested with 4 PhD students and a professor specializing in cybersecurity. Issues, including complex terminologies and the length of the CD message, were identified and resolved. Measures were also taken against error variances [44]. Error variance can be caused by preexisting factors that introduce differences among the study groups, beyond the treatment effect. In this regard, the health care personnel were randomly assigned to reduce potential biases. Additionally, the participants were asked not to share their questionnaires with others. Attention checkers [45] were among the survey items in the study, and these required the participants to choose specified responses. Participants who fail to correctly answer 2 of the 3 attention checkers are inferred to have not paid attention while answering the questionnaire. As a result, 2 records were discarded. This has been one of the most popular methods for improving survey response quality without compromising research findings. The participants also had to either agree or disagree with the cognitive dissonance message, as shown in Figure 5.

Figure 3. Experiment setup.

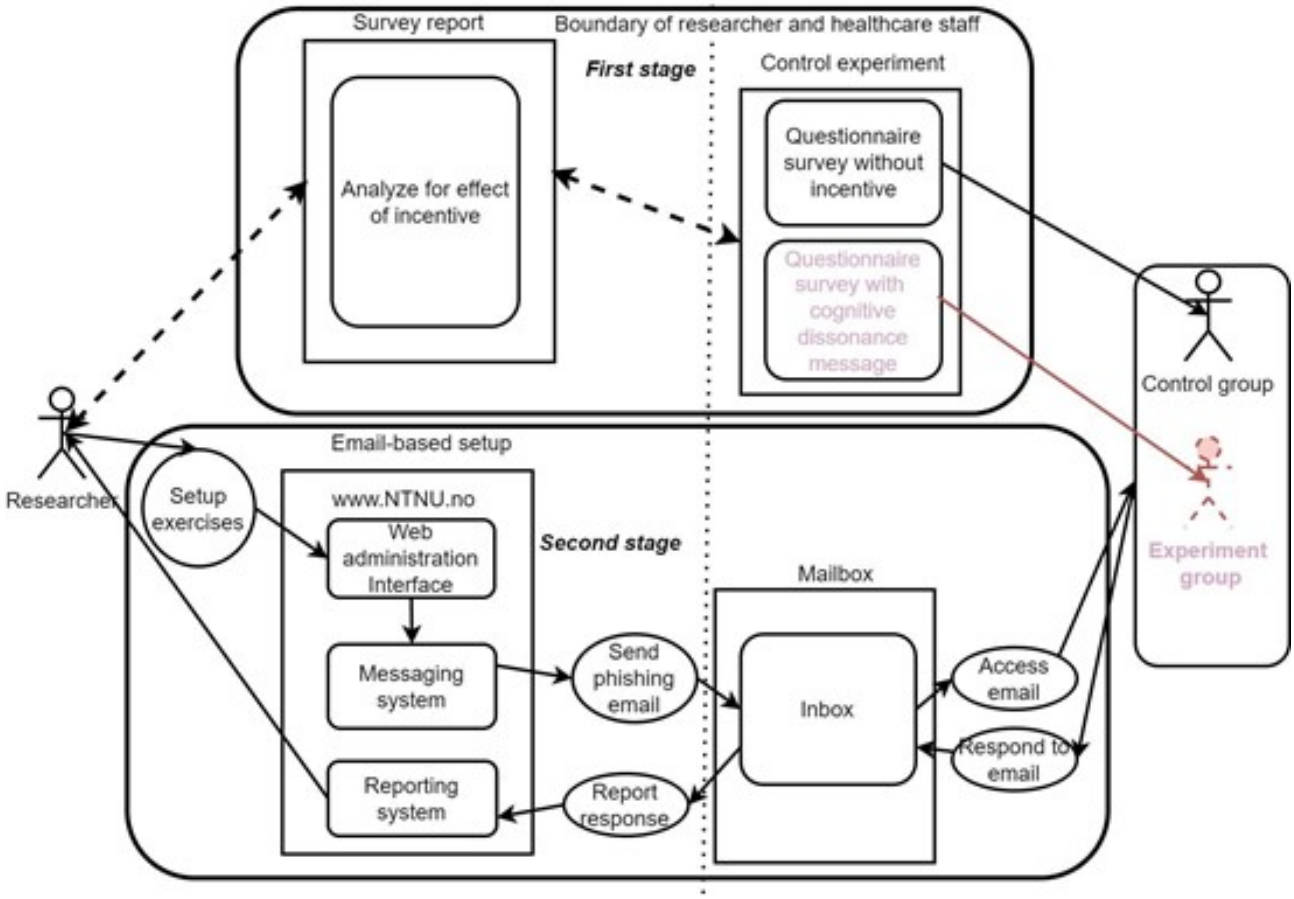

XSL•FO
RenderX

Figure 4. Response rate.

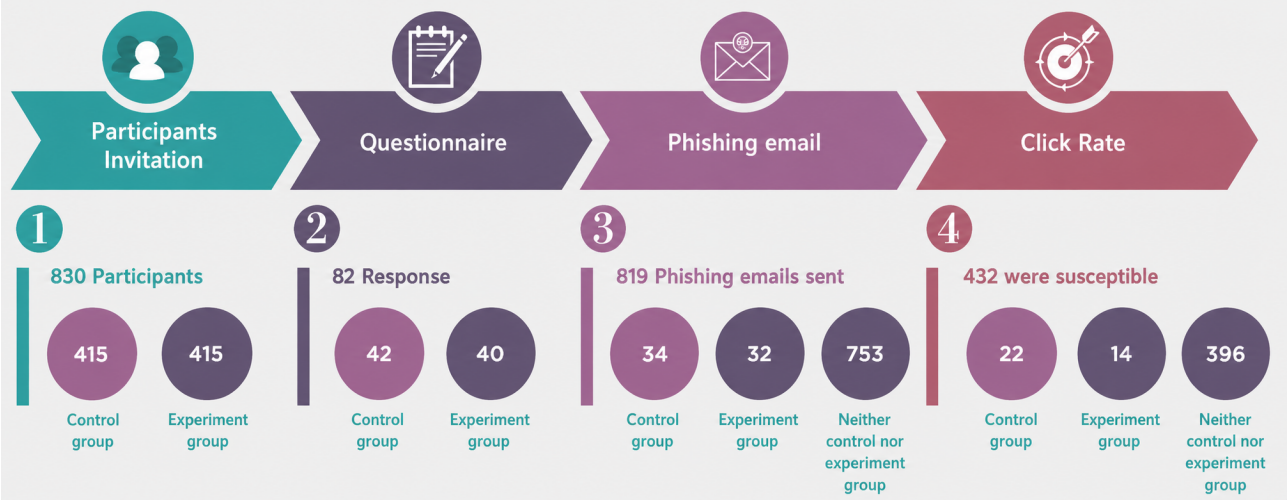

Figure 5. Cognitive dissonance message agreement for experiment group.

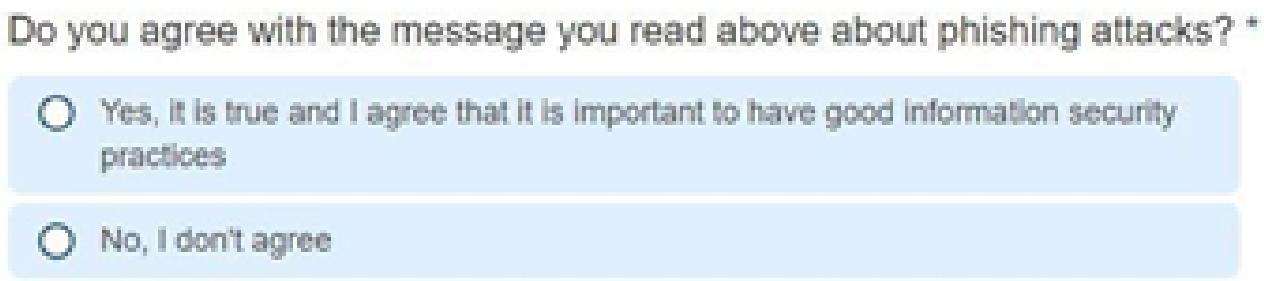

Phishing Simulation Setup and Experiment Process

A phishing simulation tool, known as Gophish (Jordan Wright) [28], was set up on a server, as illustrated in Figure 3. It has a feature that allows an attacker to simulate a phishing email and send it to a target. It can also record click events on links in phishing emails.

During the initial setup, the simulated phishing email was tested with the staff of the target hospital; however, it was flagged as spam by the service provider's email filtering system. Since the study goal was not to test the technical email security controls of the target facility, we collaborated with the providers who configured the email system to allow the phishing simulation email to land in the inboxes of the targeted participants. The email message content is shown in Multimedia Appendix 2.

The in-the-wild phishing simulation was implemented using the open-source GoPhish platform, which was configured to deliver a single simulated phishing email to staff with valid institutional email addresses [28]. The platform automatically recorded email delivery status and objective link-click events, which were used as the primary behavioral outcome measure [11,26,29,46-50].

The flow of participants in both the controlled experiment and the observational study is shown in Figures 6 and 7, respectively. The CONSORT and STROBE checklists have also been provided in Checklist 1 and Checklist 2, respectively.

XSL•FO
RenderX

Figure 6. CONSORT (Consolidated Standards of Reporting Trials) diagram.

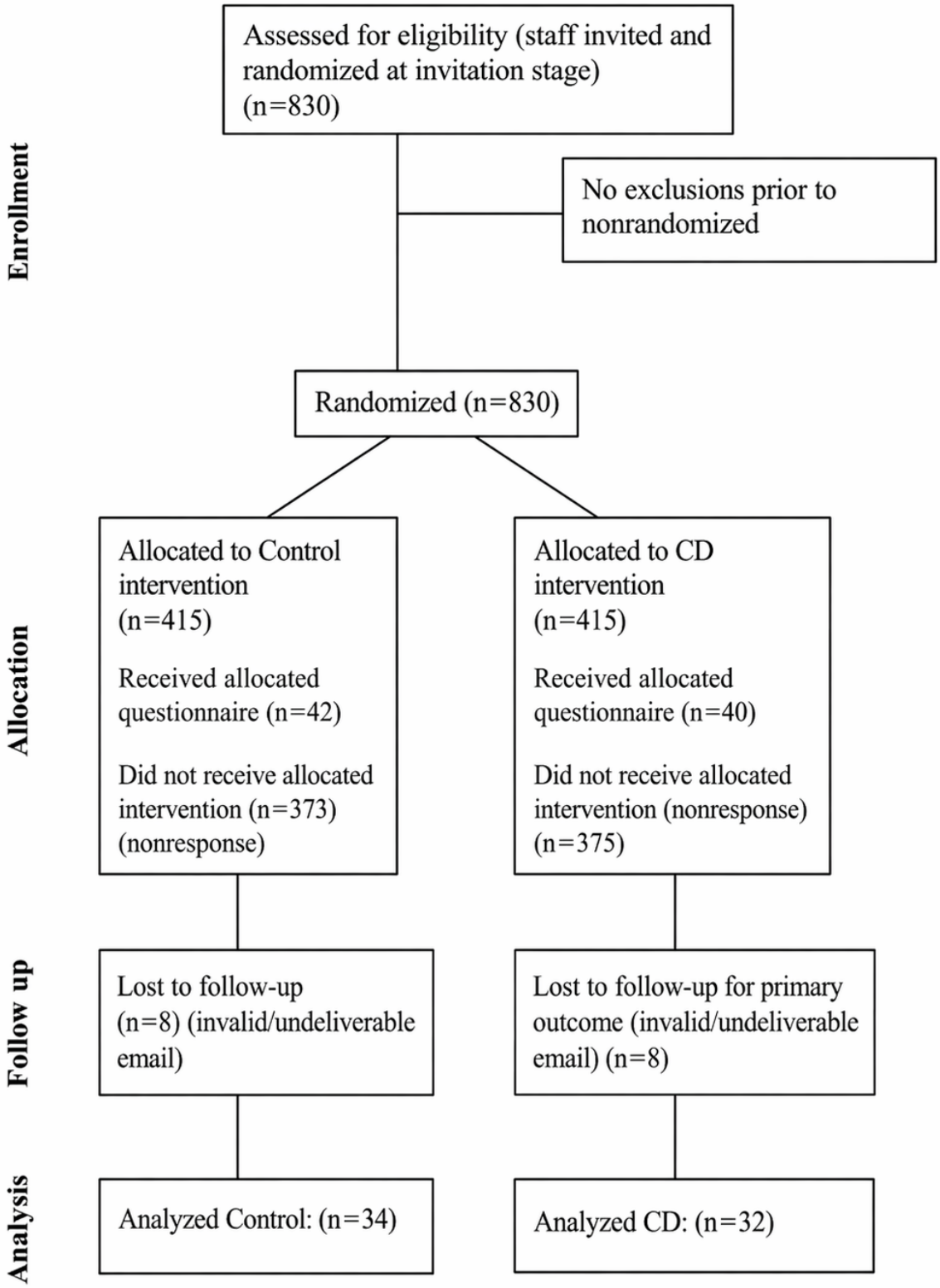

XSL•FO
RenderX

Figure 7. STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) diagram.

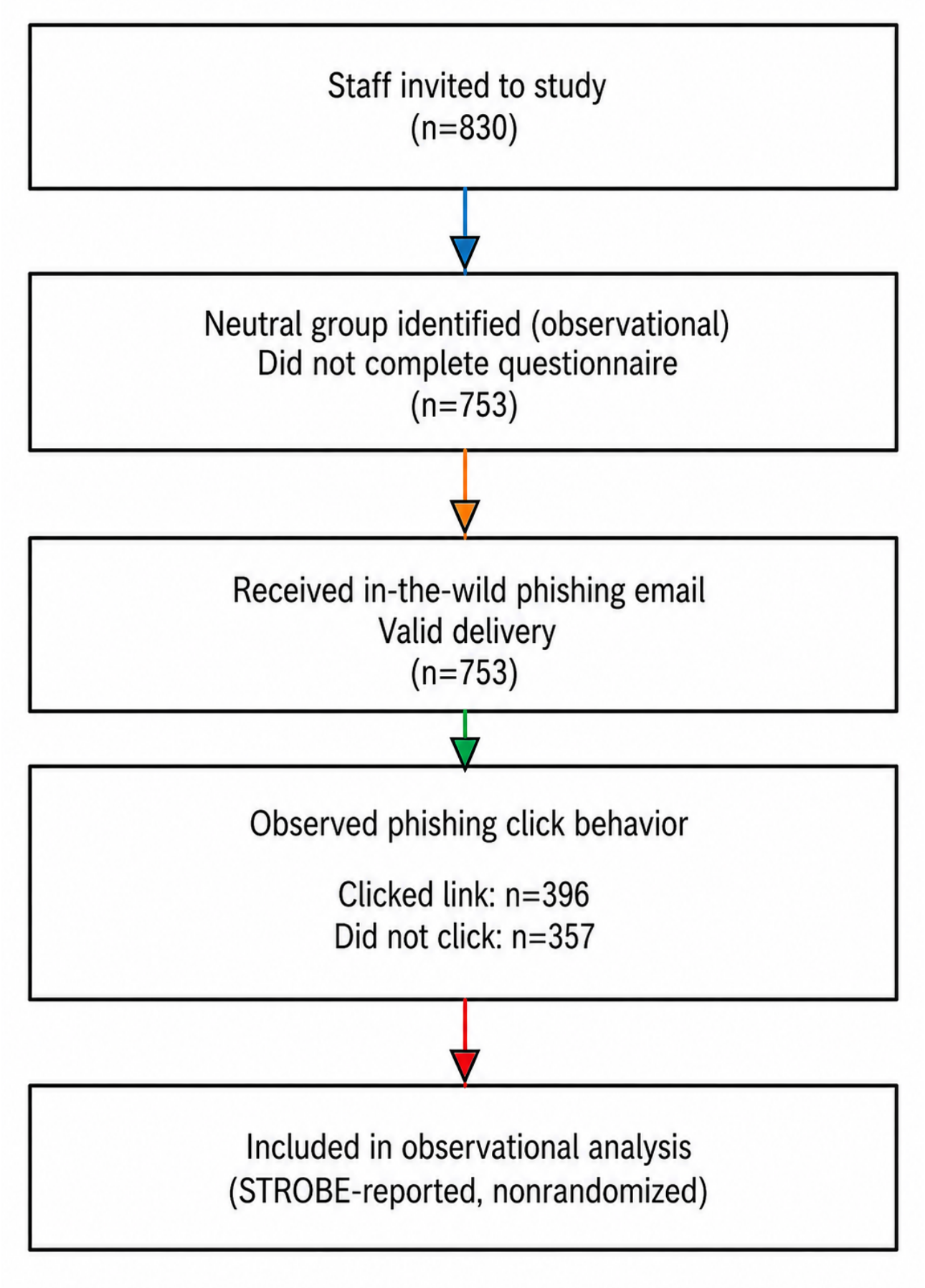

XSL•FO
RenderX

Fourteen dependent scale variables representing psychosocial constructs derived from the Health Belief Model and Protection Motivation Theory were included in the exploratory multivariate analysis [51,52]. Additionally, a minimum of one independent variable is required with at least 2 categorical groups. Our study also meets this condition with an independent 2-level design: in the first stage, there are 2 levels (control and experiment), and in the second stage, there are 3 groups (neutral, control, and experiment). A multivariate analysis of variance (MANOVA) assumes multivariate normality. The test also requires homogeneity of covariate metrics, and the dependent variables must not be multicollinear. A sufficient sample size is also needed in MANOVA. A sample size is required for each level of the independent variable. The rule of thumb requires 30 participants per group level [53]. The actual behavior (AB) variable has the lowest number of participants in the control and experimental groups, 30 and 32, respectively, indicating that the study met the minimum requirement [54]. Cronbach α was used in this study because it is a widely used measure of reliability [40].

Ethical Considerations

Prior to the launch of the phishing study, a press release was issued to administrators. Additionally, during the launch of the attack, data protection and the well-being of the participants were considered by providing a debriefing and obtaining informed consent [27]. Having followed these measures, this study obtained ethical clearance from the targeted hospital, the Regional Committees for Medical and Health Research Ethics of Norway (REK), and the Norwegian Center for Research Data (NSD). A phishing simulation was then conducted via email in this experiment.

Results

Overview

To ensure methodological transparency and inferential clarity, the analytical strategy was explicitly aligned with the study objectives. The omnibus chi-square test of independence comparing click behavior across the control, CD-primed, and neutral groups (n=819) was designated a priori as the sole confirmatory test of behavioral impact (hypothesis 1). All survey-based multivariate and univariate analyses, conducted on a substantially smaller subsample (n=62), were treated as exploratory and hypothesis-generating due to limited statistical power and variable construct reliability. This alignment ensures that confirmatory claims are restricted to adequately powered, objective behavioral outcomes, while construct-level findings are interpreted conservatively and used to inform future research design. This section presents the analysis of the results, covering descriptive statistics, click rates, reliability, normality tests, and the significance of the studies.

Descriptive Statistics

All participants in the CD condition agreed with the presented message. As shown in Table 1, a total of 80 records from the survey were analyzed out of 82 participants. Two records were removed from the analysis because they did not pass the attention checks that were placed in the questionnaire [45].

XSL•FO
RenderX

Table . Descriptive statistics of demographic variables (N=80).

Variable category	n (%)
Sex	
Male	58 (72.5)
Female	22 (27.5)
Age range (years)	
21 - 30	14 (17.5)
31 - 40	23 (28.8)
41 - 50	18 (22.5)
51 - 60	16 (20.0)
>60	9 (11.3)
Position	
Administrator	6 (7.5)
Nurse	40 (50.0)
Doctor	31 (38.8)
Others	3 (3.8)
Years of experience	
1 - 5	9 (11.3)
6 - 10	19 (23.8)
11 - 15	10 (12.5)
16 - 20	13 (16.3)
>20	29 (36.3)
Knowledge of phishing attacks	
No knowledge	12 (15.0)
Basic	23 (28.8)
Medium	26 (32.5)
High	15 (18.8)
Very high	3 (3.8)
Professional	1 (1.3)
Opinion in cognitive dissonance	
Not available	40 (50.0)
Agree	38 (95.0)
Disagree	2 (5.0)

Furthermore, among these participants, 72.5% (58/80) were males, while 27.5% (22/80) were females. The age group 31-40 years old had the highest proportion (23/80, 28%), while participants >60 years old had the lowest proportion (9/80, 11.3%). Regarding the participants, primarily doctors, nurses, and administrators participated. Nurses comprised more than half of the total participants (40/80, 50%), followed by doctors (31/80, 38%) and administrators (6/80, 7.5%). In terms of years of work experience, participants with >20 years of experience were comparatively more numerous (29/80, 36.3%), as shown in Table 1. Meanwhile, the participants also shared their phishing security practice knowledge. Most of them (26/80, 32.5%) had medium knowledge, 28% (23/80) had basic knowledge, and 15% (12/80) reported no knowledge of phishing security practices. Additionally, participants in the experiment group shared their opinion on the treatment effect of CD. Out of the 40 participants in the experiment group, 38 (95%) agreed with the effectiveness of the treatment measure. To control for these confounding variables, the randomization approach [55] was used to assign participants to the 2 groups in this study.

As shown in Table 2, descriptive statistics were computed for all dependent variables across study groups. Overall, mean values for security practice measures were generally higher in the control group than in the experimental group. Box's test of equality of covariance matrices was conducted to assess the assumption that the covariance matrices of the dependent variables were equivalent across groups. The test yielded a

XSL•FO
RenderX

Box's mean value of 153.081 with a nonsignificant *P* value (*P*=.24), indicating no evidence of heterogeneity in covariance matrices. Consistent with established guidelines [56], the null hypothesis of equal covariance matrices was therefore retained.

XSL•FO
RenderX

Table . Descriptive statistics of dependent variables across groups.

Construct (n)	Mean (SD)
Actual behavior (AB)	
Control (30)	3.93 (1.799)
Experiment (32)	2.75 (2.016)
Knowledge (K)	
Control (30)	1.82 (0.412)
Experiment (32)	1.81 (0.494)
Attitude (A)	
Control (30)	1.79 (0.338)
Experiment (32)	1.62 (0.541)
Intended behavior (IB)	
Control (30)	1.92 (0.371)
Experiment (32)	1.73 (0.463)
Perceived vulnerability (PV)	
Control (30)	2.09 (0.711)
Experiment (32)	1.85 (0.871)
Perceived severity (PS)	
Control (30)	2.07 (0.719)
Experiment (32)	1.41 (0.534)
Perceived self-efficacy (SE)	
Control (30)	2.75 (0.65)
Experiment (32)	2.45 (0.672)
Perceived response efficacy (RE)	
Control (30)	1.4 (0.395)
Experiment (32)	1.41 (0.534)
Perceived barrier (PB)	
Control (30)	3.35 (0.842)
Experiment (32)	3.34 (0.689)
Perceived cues to action (CA)	
Control (30)	2.98 (0.517)
Experiment (32)	2.55 (0.787)
Password	
Control (30)	1.82 (0.482)
Experiment (32)	1.67 (0.59)
Incident	
Control (30)	2.39 (0.708)
Experiment (32)	2.07 (0.766)
Email	
Control (30)	1.6 (0.43)
Experiment (32)	1.49 (0.523)
Mobile	
Control (30)	1.65 (0.355)
Experiment (32)	1.65 (0.471)

XSL•FO
RenderX

Reliability, Validity, and Assumption Assessment

Table 3 presents internal consistency estimates for the study constructs. Cronbach α values ranged from 0.256 for CA to 0.792 for incident reporting. While several constructs demonstrated acceptable internal consistency, most notably CA (α=0.256), password practices (α=0.349), and PBs (α=0.476) exhibited poor internal consistency. These variables were therefore retained solely as exploratory indicators and were not used to support confirmatory inference or substantive theoretical claims.

Table . Reliability statistics.

Number	Construct	Cronbach α	Number of items (n)
1	Knowledge (K)	0.660	13
2	Attitude (A)	0.53	9
3	Behavior (B)	0.648	16
4	Perceived vulnerability (PV)	0.654	3
5	Perceived severity (PS)	0.540	3
6	Perceived self-efficacy (SE)	0.655	6
6	Perceived response efficacy (RE)	0.717	3
8	Perceived barriers (PB)	0.476	4
9	Cues to action (CA)	0.256	2
10	Password	0.349	6
11	Incident	0.792	9
12	Email	0.554	7
13	Mobile	0.65	12

Phishing Click Behavior With Chi-Square Test

Phishing susceptibility was evaluated using an omnibus chi-square test of independence comparing link-click behavior across 3 groups, including control (n=34), CD-primed (n=32), and neutral nonresponders (n=753). Observed click rates were 65% (22/34) in the control group, 44% (14/32) in the CD-primed group, and 53% (396/753) in the neutral group (Table 4; Figure 4). The corresponding 95% CIs (Wilson method) were 0.48 - 0.79, 0.28 - 0.61, and 0.49 - 0.56, respectively.

Table . Observed contingency table for phishing click behavior.

Outcome	Control (n=34)	CD-Primed (n=32)	Neutral (n=753)
Clicked (n)	22	14	396
Not clicked (n)	12	18	357
Total (n)	34	32	753

The omnibus chi-square test did not detect a statistically significant association between group membership and click behavior, (χ^2_2=3.00; n=819=; *P*=.22). The effect size was small (Cramér V=0.060; approximately 95% CI 0.00 - 0.12), indicating that differences in click behavior across groups were modest and consistent with sampling variability.

To contextualize the observed pattern within the randomized subset, exploratory descriptive comparisons were computed between the CD-primed and control groups. The CD-primed group exhibited a lower observed click rate than the control group (44% vs 65%), corresponding to an absolute risk difference of −0.21 (95% CI −0.45 to 0.03), a relative risk of 0.68 (95% CI 0.42 - 1.08), and an odds ratio of 0.42 (95% CI 0.16-1.14). These estimates are imprecise due to small group sizes and are reported descriptively without confirmatory inference.

Because the Neutral group was not randomized and consists of nonresponders, comparisons involving the Neutral group are observational and may reflect selection bias or baseline differences rather than intervention effects. Overall, the behavioral results indicate, at most, a directional but statistically nonsignificant association between CD exposure and immediate phishing click behavior under real-world conditions.

A post hoc power assessment based on the observed effect size (W=0.060) indicated limited sensitivity to detect effects of this magnitude (power ≈0.32) and is reported descriptively only.

Assumption Checks

Normality of the dependent variables was assessed using the Shapiro-Wilk test [57]. As shown in Table 5, several variables deviated from normality, which informed the selection of robust and nonparametric analytical approaches.

XSL•FO
RenderX

Table . Normality test (Shapiro-Wilk).

Construct	-Shapiro-Wilk statistic, W (df)	*P* value
Actual behavior (AB)	0.627 (62)	<.001
Knowledge (K)	0.969 (62)	.11
Attitude (A)	0.954 (62)	.02
Intended behavior (IB)	0.968 (62)	.11
Perceived vulnerability (PV)	0.917 (62)	<.001
Perceived severity (PS)	0.884 (62)	<.001
Perceived self-efficacy (SE)	0.981 (62)	.43
Perceived response efficacy (RE)	0.814 (62)	<.001
Perceived barriers (PB)	0.966 (62)	.08
Cues to action (CA)	0.828 (62)	<.001
Password	0.951 (62)	.01
Incident	0.960 (62)	.04
Email	0.887 (62)	<.001
Mobile	0.931 (62)	.002

Exploratory Multivariate Analyses

Subsequently, a one-way MANOVA was performed to examine whether group membership was associated with differences in the combined set of dependent variables, including mobile device and SMS use and incident reporting, as shown in Table 6. Statistical significance was evaluated at an α level of .05.

XSL•FO
RenderX

Table . Correlation (n=62).

Variable	AB[a]	Knowledge	Attitude	IB[b]	PV[c]	PS[d]	SE[e]	RE[f]	PB[g]	CA[h]	Password	IR	Email use	Mobile
AB	—[i]	–0.371[j]	–0.021	–0.050	0.010	0.227	0.090	–0.090	0.113	0.041	0.041	–0.175	–0.027	–0.208
Knowledge	–0.371[j]	—	0.500[j]	0.515[j]	0.133	–0.159	0.440[j]	0.336[j]	–0.328[j]	0.149	0.390[j]	0.495[j]	0.538[j]	0.683[j]
Attitude	–0.021	0.500[j]	—	0.468[j]	0.261[k]	0.213	0.067	0.257[k]	–0.209	0.074	0.394[j]	0.658[j]	0.402[j]	0.382[j]
IB	–0.050	0.515[j]	0.468[j]	—	0.293[j]	0.297[j]	0.432[j]	0.407[j]	–0.401[j]	0.337[j]	0.642[j]	0.637[j]	0.655[j]	0.479[j]
PV	0.010	0.133	0.261[k]	0.293[j]	—	0.383[j]	0.030	0.319[j]	–0.147	0.126	0.217	0.291[j]	0.076	0.050
PS	0.227	−0.159	0.213	0.297[j]	0.383[j]	—	0.062	0.148	–0.128	0.173	0.243[k]	0.272[k]	0.150	–0.019
SE	0.090	0.440[j]	0.067	0.432[j]	0.030	0.062	—	0.366[j]	–0.323[j]	0.537[j]	0.158	0.156	0.431[j]	0.505[j]
RE	–0.090	0.336[j]	0.257[k]	0.407[j]	0.319[j]	0.148	0.366[j]	—	–0.108	0.146	0.172	0.232[k]	0.356[j]	0.350[j]
PB	0.113	–0.328[j]	–0.209	–0.401[j]	–0.147	–0.128	–0.323[j]	–0.108	—	–0.336[j]	–0.264[k]	–0.330[j]	0.236[k]	0.148
CA	0.041	0.149	0.074	0.337[j]	0.126	0.173	0.537[j]	0.146	–0.336[j]	—	0.133	0.182	0.204	0.362[j]
Password	0.041	0.390[j]	0.394[j]	0.642[j]	0.217	0.243[k]	0.158	0.172	–0.264[k]	0.133	—	0.383[j]	0.353[j]	0.210
IR	–0.175	0.495[j]	0.658[j]	0.637[j]	0.291[j]	0.272[k]	0.156	0.232[k]	–0.330[j]	0.182	0.383[j]	—	0.335[j]	0.133
Email use	–0.027	0.538[j]	0.402[j]	0.655[j]	0.076	0.150	0.431[j]	0.356[j]	0.236[k]	0.204	0.353[j]	0.335[j]	—	0.463[j]
Mobile	–0.208	0.683[j]	0.382[j]	0.479[j]	0.050	–0.019	0.505[j]	0.350[j]	0.148	0.362[j]	0.210	0.133	0.463[j]	—

[a]AB: actual behavior.
[b]IB: intended behavior.
[c]PV: perceived vulnerability.
[d]PS: perceived severity.
[e]SE: self-efficacy.
[f]RE: response efficacy.
[g]PB: perceived barrier.
[h]CA: cues to action.
[i]Not available.
[j]Correlation is significant at the 0.01 level (2-tailed).
[k]Correlation is significant at the 0.05 level (2-tailed).

Given the inclusion of constructs with low internal consistency, MANOVA is interpreted strictly as a descriptive screening of group differentiation rather than as inferential evidence of multivariate effects. A multivariate effect was observed in the MANOVA. Pillai's Trace was used for the omnibus multivariate test. The analysis yielded a Pillai's Trace of 0.442, $F_{14, 47}$=2.660; P=.006, indicating overall group differences across the combined set of dependent variables. Given the heterogeneous reliability of the included constructs and the limited survey sample size, this result is interpreted as an omnibus, exploratory indication of group differentiation rather than evidence of specific construct-level effects.

The estimated multivariate effect size suggests that approximately 44.2% of the variance in the linear combination of dependent variables was associated with group membership. Assumptions for follow-up univariate analyses were assessed using Levene test across all 14 dependent variables [58]. Although AB and CA showed statistically significant Levene results (P<.05), inspection of group standard deviations (Table 2) revealed no substantial imbalance in variance, supporting the robustness of subsequent ANOVA analyses. The CA exhibited very low internal consistency (Cronbach α=0.256) and were therefore retained solely as an exploratory indicator, without supporting confirmatory inference.

Follow-up one-way ANOVA tests indicated statistically significant group differences for AB, PS, and CA (Table 7). Findings for constructs with acceptable reliability are interpreted with greater confidence, whereas results involving CA, PBs, and password practices are treated as exploratory, with emphasis placed on effect sizes rather than P values. In particular, although statistical differences were observed for CA, this finding is interpreted cautiously due to very low internal consistency and is reported only as an exploratory pattern rather than a substantive effect.

XSL•FO
RenderX

Table . ANOVA results.

Construct	Sum of Squares	df	Mean Square	F	*P* value	Partial eta squared	Noncentrality parameter	Observed power
Actual behavior (AB)	21.682	1	21.682	5.917	.02	0.090	5.917	0.668
Knowledge (K)	0.003	1	0.003	0.013	.91	0.000	0.013	0.051
Attitude (A)	0.452	1	0.452	2.186	.14	0.035	2.186	0.307
Intended behavior (IB)	0.572	1	0.572	3.223	.08	0.051	3.223	0.423
Perceived vulnerability (PV)	0.853	1	0.853	1.340	.25	0.022	1.340	0.207
Perceived severity (PS)	6.753	1	6.753	17.020	<.001	0.221	17.020	0.982
Perceived self-efficacy (SE)	1.365	1	1.365	3.116	.08	0.049	3.116	0.412
Perceived response efficacy (RE)	0.001	1	0.001	0.003	.96	0.000	0.003	0.050
Perceived barrier (PB)	0.003	1	0.003	0.005	.94	0.000	0.005	0.051
Perceived cues to action (CA)	2.950	1	2.950	6.574	.01	0.099	6.574	0.713
Password	0.325	1	0.325	1.110	.30	0.018	1.110	0.179
Incident	1.652	1	1.652	3.028	.09	0.048	3.028	0.402
Email	0.183	1	0.183	0.792	.38	0.013	0.792	0.141
Mobile	0.000	1	0.000	0.000	.99	0.000	0.000	0.050

Post-hoc power analysis indicated that the survey-based MANOVA and ANOVA models achieved approximately 40% power to detect medium-sized effects, confirming that these analyses were underpowered. Consequently, survey-based multivariate and univariate results are interpreted as exploratory. In contrast, the behavioral click-rate analysis, based on the full sample, was sufficiently powered.

Discussion

Principal Findings

Despite substantial investment in technical email filtering and security awareness training, health care staff continue to demonstrate high susceptibility to socially engineered emails under real operational conditions [11]. Evidence from field-based phishing simulations suggests that many existing interventions show limited or inconsistent effectiveness when evaluated using objective behavioral outcomes rather than self-reported intentions [52,53]. This limitation has prompted increasing interest in behavioral and psychologically grounded approaches that aim to influence decision-making at the moment of threat exposure, particularly those capable of producing measurable short-term changes in actual user behavior.

In light of this gap, the present study examined whether a brief CD-based priming intervention, delivered immediately prior to a real-world phishing simulation, was associated with differences in phishing-related outcomes among health care staff. The analytical framework comprised one confirmatory behavioral hypothesis and 2 exploratory research questions. The primary and confirmatory hypothesis (hypothesis 1) evaluated whether exposure to the CD prompt was associated with reduced observed phishing susceptibility, operationalized as objective link-click behavior during an in-the-wild phishing simulation. This behavioral outcome served as the study's primary endpoint and was assessed using an omnibus analysis across all staff. Two exploratory research questions (RQ1 and RQ2) examined whether CD exposure was associated with directional differences in theory-driven security perception constructs derived from the HBM and PMT, as well as with self-reported security practices related to password management, incident reporting, email handling, and mobile-device security.

Given the limited sample size and variable construct reliability, the findings for RQ1 and RQ2 are interpreted as exploratory and hypothesis-generating rather than confirmatory. Accordingly, the discussion first addresses the confirmatory behavioral findings (hypothesis 1), followed by a cautious interpretation of exploratory construct-level and self-reported outcomes.

XSL•FO
RenderX

Primary Study Finding (Hypothesis 1): Confirmatory Behavioral Outcome

The primary outcome of this study was objective phishing click behavior observed during an in-the-wild simulation [29,59,60]. Although the CD-primed group exhibited a lower observed click-through rate than the control group (44% vs 65%), the prespecified omnibus chi-square test did not detect a statistically significant association between group membership and click behavior. The estimated effect size was small, and CIs were wide, reflecting limited precision due to the modest size of the randomized experimental groups.

Accordingly, the findings do not provide confirmatory statistical support for hypothesis 1. Rather, they indicate a directional, but statistically nonsignificant, association between brief CD-based priming and immediate phishing-click behavior under real-world conditions. This cautious interpretation is warranted because only a subset of participants was randomized to receive the CD prompt, while the neutral group comprised nonresponders and was not assigned to any experimental condition. As such, causal inference is limited to comparisons within the randomized subset, whereas comparisons involving the neutral group are observational and may reflect baseline or selection differences.

When situated within the broader cybersecurity and health informatics literature, the modest magnitude of the observed effect is consistent with prior studies examining behavioral phishing outcomes, which frequently report small and heterogeneous effects following brief or one-time interventions (eg, [61-63]). These findings contrast with studies reporting stronger effects from sustained educational or motivational interventions, which typically target knowledge, risk appraisal, or SE rather than rapid, heuristic-driven decision-making.

This distinction highlights the inherent difficulty of influencing immediate phishing behavior using brief psychological cues alone. From an applied perspective, the present results suggest that CD-based prompts may function as a situational nudge, introducing momentary cognitive friction prior to exposure, rather than as a standalone behavioral control. Their potential value, therefore, lies in complementing existing phishing simulations, awareness training, and technical safeguards within a layered defense strategy.

A notable pattern in the findings is the disconnect between observed phishing behavior and self-reported security perceptions. Although phishing susceptibility was measured using an objective behavioral outcome (link-clicking), effect estimates were imprecise because the randomized comparison groups were small. At the same time, several survey-based constructs showed low internal consistency, limiting confidence in construct-level interpretation. Together, these results are consistent with prior evidence that self-reported security perceptions and intentions do not reliably predict real-world phishing behavior, particularly in high-pressure health care settings.

Construct-Level Findings With Limited Power: Exploratory Evidence (RQ1-RQ2)

In contrast to the confirmatory behavioral outcome (hypothesis 1), analyses addressing the exploratory research questions (RQ1-RQ2) focused on self-reported psychological perceptions and security practices and are interpreted as hypothesis-generating rather than confirmatory. Although the multivariate analysis indicated an overall group effect (Pillai Trace=0.442; $F_{14,47}$=2.66; P=.006), this result was derived from a substantially smaller survey subsample (control n=30; CD-primed n=32) and involved multiple dependent variables, thereby limiting statistical sensitivity for reliable construct-level inference.

Follow-up univariate analyses revealed nonsignificant or marginal effects for most perception and practice constructs, including PV, SE, RE, incident-reporting intentions, password management, email handling, mobile device security practices, and self-reported knowledge, attitudes, and intended behavior (P=.08-.30). Several constructs, most notably CA, PBs, and password practices, also exhibited low internal consistency, further constraining interpretability and precluding confirmatory conclusions.

While some variables displayed directional or near-significant trends, these patterns do not provide evidence of causal mechanisms nor establish that the CD prompt directly influenced underlying psychological processes [64,65]. Rather, the observed divergence between modest behavioral effects and weak or inconsistent construct-level changes is consistent with prior health care and organizational security research documenting limited correspondence between self-reported perceptions or intentions and objectively observed security behavior [10,18]. Accordingly, the construct-level findings are best viewed as exploratory signals that motivate future research by using larger and more balanced samples, improved psychometric validation, and analytical designs capable of formally testing mediation or mechanism-based pathways.

Contextual Factors, Baseline Susceptibility, Implications of the Study

Across all study groups, phishing susceptibility remained high, with observed click rates exceeding 50%. This pattern reinforces prior evidence that health care organizations are particularly vulnerable to social engineering attacks, even in settings with established technical defenses, policies, and awareness programs [6,66]. The phishing message deployed in this study leveraged a contemporaneous geopolitical crisis, a strategy commonly observed in real-world phishing campaigns to heighten emotional salience, urgency, and perceived legitimacy [55]. While this design choice enhances ecological validity, it also underscores the context-dependent nature of phishing susceptibility and limits direct generalization beyond comparable threat scenarios.

From an applied perspective, the findings suggest that brief, lightweight psychological prompts may complement existing security awareness and simulation programs by influencing immediate decision-making at the point of exposure. Importantly, the evidence supports this implication only for short-term, observed behavioral responses. The present study does not establish the durability of the observed effects, their transferability across organizational contexts or attack types, or the specific psychological mechanisms through which CD

XSL•FO
RenderX

prompting may operate. These limitations highlight the need for longitudinal and replication studies before such interventions can be considered as standalone or broadly generalizable mitigation strategies.

Conclusion

This study examined whether a brief CD-based priming intervention, delivered immediately prior to a real-world phishing simulation, was associated with differences in phishing susceptibility among health care staff. The primary behavioral analysis indicated a directional but statistically nonsignificant association, with lower observed click rates in the CD-primed group; however, effect estimates were small and imprecise due to limited randomized group sizes. Accordingly, causal inference is restricted to the randomized comparison, while differences involving the neutral group are observational.

Survey-based analyses of security perceptions and self-reported practices were conducted on a smaller subsample, and several constructs exhibited low internal consistency. These findings are therefore considered exploratory and do not support conclusions about psychological mechanisms. Overall, the results suggest that CD-based prompts may function as a lightweight, short-term behavioral nudge under real-world conditions but do not establish a reliable effect.

Larger, fully randomized, and longitudinal studies with improved psychometric validation are needed before such interventions can be considered reliable complements to established cybersecurity controls.

Acknowledgments

No generative AI was used.

Funding

No funding source to declare.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire.
[DOCX File, 17 KB - jmir_v28i1e68051_app1.docx]

Multimedia Appendix 2
Cognitive dissonance message.
[DOCX File, 14 KB - jmir_v28i1e68051_app2.docx]

Checklist 1
CONSORT (Consolidated Standards of Reporting Trials) checklist.
[PDF File, 206 KB - jmir_v28i1e68051_app3.pdf]

Checklist 2
STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist.
[PDF File, 121 KB - jmir_v28i1e68051_app4.pdf]

References

1. The state of healthcare cybersecurity 2025. Veriti. URL: https://veriti.ai/wp-content/uploads/2024/12/The-State-of-Healthcare-Cybersecurity-2025-_-A-Veriti-Research-Report.pdf [accessed 2026-04-15]
2. Cost of a data breach: the healthcare industry. IBM. URL: https://www.ibm.com/think/insights/cost-of-a-data-breach-healthcare-industry [accessed 2025-12-10]
3. Keller T, Warwas JI, Klein J, Henkenjohann R, Trenz M, Trang STN. Motivational framing strategies in health care information security training: randomized controlled trial. JMIR Med Educ 2025 Nov 7;11(1):e73245. [doi: 10.2196/73245] [Medline: 41202283]
4. Yeng PK, Fauzi MA, Yang B, Nimbe P. Investigation into phishing risk behaviour among healthcare staff. Information 2022 Aug;13(8):392. [doi: 10.3390/info13080392]
5. Abdelhamid M. The role of health concerns in phishing susceptibility: survey design study. J Med Internet Res 2020 May 4;22(5):e18394. [doi: 10.2196/18394] [Medline: 32364511]
6. Jalali MS, Bruckes M, Westmattelmann D, Schewe G. Why employees (still) click on phishing links: investigation in hospitals. J Med Internet Res 2020 Jan 23;22(1):e16775. [doi: 10.2196/16775] [Medline: 32012071]

XSL•FO
RenderX

7. Donalds C, Osei-Bryson KM. Cybersecurity compliance behavior: exploring the influences of individual decision style and other antecedents. Int J Inf Manage 2020 Apr;51:102056. [doi: 10.1016/j.ijinfomgt.2019.102056]
8. Ewoh P, Vartiainen T. Vulnerability to cyberattacks and sociotechnical solutions for health care systems: systematic review. J Med Internet Res 2024 May 31;26(1):e46904. [doi: 10.2196/46904] [Medline: 38820579]
9. Ewoh P, Vartiainen T, Mantere T. Sociotechnical cybersecurity framework for securing health care from vulnerabilities and cyberattacks: scoping review. J Med Internet Res 2025 Oct 15;27(1):e75584. [doi: 10.2196/75584] [Medline: 40838797]
10. Yeng PK, Szekeres A, Yang B, Snekkenes EA. Mapping the psychosocialcultural aspects of healthcare professionals' information security practices: systematic mapping study. JMIR Hum Factors 2021 Jun 9;8(2):e17604. [doi: 10.2196/17604] [Medline: 34106077]
11. Rizzoni F, Magalini S, Casaroli A, Mari P, Dixon M, Coventry L. Phishing simulation exercise in a large hospital: a case study. Digit Health 2022;8:20552076221081716. [doi: 10.1177/20552076221081716] [Medline: 35321019]
12. Mou J, Cohen J, Bhattacherjee A, et al. A test of protection motivation theory in the information security literature: a meta-analytic structural equation modeling approach in search advertising. J Assoc Inf Syst 2022 Jan;23(1):196-236. [doi: 10.17705/1jais.00723]
13. Herath T, Rao HR. Encouraging information security behaviors in organizations: role of penalties, pressures and perceived effectiveness. Decis Support Syst 2009 May;47(2):154-165. [doi: 10.1016/j.dss.2009.02.005]
14. Gordon WJ, Wright A, Glynn RJ, et al. Evaluation of a mandatory phishing training program for high-risk employees at a US healthcare system. J Am Med Inform Assoc 2019 Jun 1;26(6):547-552. [doi: 10.1093/jamia/ocz005] [Medline: 30861069]
15. Chen Y, Ramamurthy K, Wen KW. Organizations' information security policy compliance: stick or carrot approach? J Manag Inf Syst 2012 Dec;29(3):157-188. [doi: 10.2753/MIS0742-1222290305]
16. Musuva PMW, Getao KW, Chepken CK. A new approach to modelling the effects of cognitive processing and threat detection on phishing susceptibility. Comput Human Behav 2019 May;94:154-175. [doi: 10.1016/j.chb.2018.12.036]
17. Pepitone A. A theory of cognitive dissonance by Leon Festinger. Am J Psychol 1959;72(1):153-155. [doi: 10.2307/1420234]
18. Harmon-Jones E, Mills J. An introduction to cognitive dissonance theory and an overview of current perspectives on the theory. In: Cognitive Dissonance: Reexamining a Pivotal Theory in Psychology, 2nd edition: American Psychological Association; 2019:3-24. [doi: 10.1037/0000135-001]
19. Harmon-Jones EE. Cognitive Dissonance: Reexamining a Pivotal Theory in Psychology, 2nd edition: American Psychological Association. URL: https://psycnet.apa.org/books/TOC/16109 [accessed 2025-12-23]
20. Altamimi S. Investigating and mitigating the role of neutralisation techniques on information security policies violation in healthcare organisations. : University of Glasgow; 2022. [doi: 10.5525/gla.thesis.82646]
21. Sykes GM, Matza D. Techniques of neutralization: a theory of delinquency. In: Delinquency and Drift Revisited: Routledge; 2017, Vol. 21. [doi: 10.4324/9780203793596]
22. Siponen M, Vance A. Neutralization: new insights into the problem of employee information systems security policy violations1. MIS Q 2010 Sep 1;34(3):487-502. [doi: 10.2307/25750688]
23. Taylor-Jackson J, McAlaney J, Foster JL, Bello A, Maurushat A, Dale J. In: Bernhard M, Bracciali A, Camp LJ, Matsuo S, Maurushat A, Rønne PB, et al, editors. Incorporating Psychology into Cyber Security Education: A Pedagogical Approach: Springer International Publishing; 2020:207-217. [doi: 10.1007/978-3-030-54455-3_15]
24. Barlow JB, Warkentin M, Ormond D, et al. Don't even think about it! The effects of antineutralization, informational, and normative communication on information security compliance. JAIS 2018;19(8):689-715. [doi: 10.17705/1jais.00506]
25. Cazares MF, Arévalo D, Andrade RO, Fuertes W, Sánchez-Rubio M. A training web platform to improve cognitive skills for phishing attacks detection. In: Nagar AK, Jat DS, Marín-Raventós G, Mishra DK, editors. Intelligent Sustainable Systems - Selected Papers of World S4 2021 2022:33-42. [doi: 10.1007/978-981-16-6309-3_4]
26. Braun O, Hörnemann J, Pohlmann N, Urban T, Grosse-Kampmann M. Different seas, different phishes – large-scale analysis of phishing simulations across different industries. : Association for Computing Machinery; 2025 Aug 25 Presented at: ASIA CCS '25; Hanoi Vietnam p. 1520-1534 URL: https://dl.acm.org/doi/proceedings/10.1145/3708821 [doi: 10.1145/3708821.3733905]
27. Sieber JE. Deception in social research I: kinds of deception and the wrongs they may involve. IRB 1982 Nov;4(9):1-5. [doi: 10.2307/3564511] [Medline: 11649501]
28. Open source phishing framework. Gophish. URL: https://getgophish.com/ [accessed 2025-12-12]
29. Marshall N, Sturman D, Auton JC. Exploring the evidence for email phishing training: a scoping review. Comput Secur 2024 Apr;139:103695. [doi: 10.1016/j.cose.2023.103695]
30. Hernan MA. Causal Inference: What If: Taylor & Francis; 2024.
31. Chapter 61 using randomization in development economics research: a toolkit. In: Handbook of Development Economics: Elsevier; 2007, Vol. 4:3895-3962. [doi: 10.1016/S1573-4471(07)04061-2]
32. Sparks P, Guthrie CA, Shepherd R. The dimensional structure of the perceived behavioral control construct 1. J Applied Social Pyschol 1997 Mar;27(5):418-438 [FREE Full text] [doi: 10.1111/j.1559-1816.1997.tb00639.x]
33. Thomopoulos G, Lyras D, Fidas C. Methodologies and ethical considerations in phishing research: a comprehensive review. : Association for Computing Machinery; 2023 Sep 27 Presented at: CHIGREECE 2023; Athens Greece p. 1-10 URL: https://dl.acm.org/doi/proceedings/10.1145/3609987 [doi: 10.1145/3609987.3609990]

XSL•FO
RenderX

34. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. Lancet 2025 Apr 14:S0140-S6736. [doi: 10.1016/S0140-6736(25)00672-5] [Medline: 40245901]
35. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. The Lancet 2007 Oct;370(9596):1453-1457. [doi: 10.1016/S0140-6736(07)61602-X]
36. Plint AC, Moher D, Morrison A, et al. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. Med J Aust 2006 Sep 4;185(5):263-267. [doi: 10.5694/j.1326-5377.2006.tb00557.x] [Medline: 16948622]
37. Green-Ajufo B, Chakravarty D, Maiorana A, Lightfoot M, Hamiga J, Rebchook G. Acceptability and feasibility of using educational incentives for research participation to advance antiracism. Ethics Hum Res 2025;47(4):18-28. [doi: 10.1002/eahr.60010] [Medline: 40658773]
38. Sobolewski J, Rothschild A, Freeman A. The impact of incentives on data collection for online surveys: social media recruitment study. JMIR Form Res 2024 Jul 4;8(1):e50240. [doi: 10.2196/50240] [Medline: 38963924]
39. Cohen J. A power primer. Psychol Bull 1992 Jul;112(1):155-159. [doi: 10.1037//0033-2909.112.1.155] [Medline: 19565683]
40. Vaske JJ, Beaman J, Sponarski CC. Rethinking internal consistency in Cronbach's alpha. Leis Sci 2017 Mar 4;39(2):163-173. [doi: 10.1080/01490400.2015.1127189]
41. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with Python. 2010 Presented at: Python in Science Conference; Austin, Texas p. 92-96. [doi: 10.25080/Majora-92bf1922-011]
42. Parsons K, McCormac A, Butavicius M, Pattinson M, Jerram C. The development of the human aspects of information security questionnaire (HAIS-q). 2013 Presented at: ACIS. [doi: 10.1016/j.cose.2013.12.003]
43. Log in. Nettskjema. URL: https://nettskjema.no [accessed 2025-12-12]
44. Huang JL, Bowling NA, Liu M, Li Y. Detecting insufficient effort responding with an infrequency scale: evaluating validity and participant reactions. J Bus Psychol 2015 Jun;30(2):299-311. [doi: 10.1007/s10869-014-9357-6]
45. Kung FYH, Kwok N, Brown DJ. Are attention check questions a threat to scale validity? Applied Psychology 2018 Apr;67(2):264-283 [FREE Full text] [doi: 10.1111/apps.12108]
46. Ho G, Mirian A, Luo E, et al. Understanding the efficacy of phishing training in practice. 2025 May Presented at: 2025 IEEE Symposium on Security and Privacy (SP); San Francisco, CA p. 37-54. [doi: 10.1109/SP61157.2025.00076]
47. Zhuo S, Biddle R, Russello G, Lottridge D. Precision email simulator for research on safety-critical phishing behaviour. : Association for Computing Machinery; 2025 Apr 26 Presented at: CHI 2025; Yokohama Japan p. 1-12 URL: https://dl.acm.org/doi/proceedings/10.1145/3706598 [accessed 2026-04-15] [doi: 10.1145/3706598.3714143]
48. Marczak B, Scott-Railton J, Aljizawi N, Anstis S, Deibert R. The great ipwn: journalists hacked with suspected NSO group imessage 'zero-click' exploit. Citizen Lab.: University of Toronto; 2020 Dec. URL: https://citizenlab.ca/2020/12/the-great-ipwn-journalists-hacked-with-suspected-nso-group-imessage-zero-click-exploit/ [accessed 2025-12-12]
49. Apple sues pegasus for spyware maker. how to check if your iphone has NSO group software. CNET. URL: https://www.cnet.com/tech/mobile/apple-sues-pegasus-for-spyware-maker-how-to-check-if-your-iphone-has-nso-group-software/ [accessed 2025-12-12]
50. What is zero-click malware, and how do zero-click attacks work? Kaspersky. URL: https://www.kaspersky.com/resource-center/definitions/what-is-zero-click-malware [accessed 2025-12-12]
51. Azrain A. Multivariate analysis of variance (MANOVA). Academia. URL: https://www.academia.edu/7751963/Multivariate_Analysis_of_Variance_MANOVA [accessed 2025-12-12]
52. Meyers LS, Gamst G, Guarino AJ. Applied Multivariate Research: Design and Interpretation: SAGE Publications; 2017. [doi: 10.4135/9781071802687]
53. Wilson Van Voorhis CR, Morgan BL. Understanding power and rules of thumb for determining sample Sizes. Tutor Quant Methods Psychol 2007;3(2):43-50. [doi: 10.20982/tqmp.03.2.p043]
54. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ 2011 Jun 27;2:53-55. [doi: 10.5116/ijme.4dfb.8dfd] [Medline: 28029643]
55. Jain AK, Gupta BB. A survey of phishing attack techniques, defence mechanisms and open research challenges. Enterprise Information Systems 2022 Apr 3;16(4):527-565 [FREE Full text] [doi: 10.1080/17517575.2021.1896786]
56. Statistical methods for psychology. eBook. URL: https://tinyurl.com/ypduhk6j [accessed 2025-12-12]
57. Pillai KCS. Some new test criteria in multivariate analysis. Ann Math Statist 1955 Mar;26(1):117-121. [doi: 10.1214/aoms/1177728599]
58. Handbook of applied multivariate statistics and mathematical modeling. In: Handbook of Applied Multivariate Statistics and Mathematical Modeling: Academic Press; 2000:3-36. [doi: 10.1016/B978-012691360-6/50002-1]
59. Tolsdorf J, Langer D, Lo Iacono L. Phishing susceptibility and the (in-)effectiveness of common anti-phishing interventions in a large university hospital. : Association for Computing Machinery; 2025 Nov 19 Presented at: CCS '25; Taipei Taiwan p. 4334-4348 URL: https://dl.acm.org/doi/proceedings/10.1145/3719027 [doi: 10.1145/3719027.3765164]
60. Althobaiti K, Alsufyani N. A review of organization-oriented phishing research. PeerJ Comput Sci 2024;10:e2487. [doi: 10.7717/peerj-cs.2487] [Medline: 39650535]

XSL•FO
RenderX

61. Canfield CI, Fischhoff B, Davis A. Quantifying phishing susceptibility for detection and behavior decisions. Hum Factors 2016 Dec;58(8):1158-1172. [doi: 10.1177/0018720816665025] [Medline: 27562565]
62. Jansson K, von Solms R. Phishing for phishing awareness. Behav Inf Technol 2013 Jun;32(6):584-593. [doi: 10.1080/0144929X.2011.632650]
63. Phish like a boss. In: Phishing Dark Waters: John Wiley and Sons; 2015:179-188. [doi: 10.1002/9781119183624]
64. Grosz MP, Rohrer JM, Thoemmes F. The taboo against explicit causal inference in nonexperimental psychology. Perspect Psychol Sci 2020 Sep;15(5):1243-1255. [doi: 10.1177/1745691620921521] [Medline: 32727292]
65. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature New Biol 2019 Mar;567(7748):305-307. [doi: 10.1038/d41586-019-00857-9] [Medline: 30894741]
66. Priestman W, Anstis T, Sebire IG, Sridharan S, Sebire NJ. Phishing in healthcare organisations: threats, mitigation and approaches. BMJ Health Care Inform 2019 Sep;26(1):1. [doi: 10.1136/bmjhci-2019-100031] [Medline: 31488498]

Abbreviations

AB: actual behavior
CA: cues to action
CD: cognitive dissonance
CONSORT: Consolidated Standards of Reporting Trials
HBM: Health Belief Model
KAB: knowledge, attitude, and behavior
MANOVA : multivariate analysis of variance
NSD: Norwegian Center for Research Data
PB: perceived barrier
PMT: Protection Motivation Theory
PS: perceived severity
PV: perceived vulnerability
RE: response efficacy
REK: Regional Committees for Medical and Health Research Ethics of Norway
RQ: research question
SE: self-efficacy
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by J Sarvestan, TDA Cardoso; submitted 27.Oct.2024; peer-reviewed by J Dykstra, LS Moussaoui; revised version received 15.Jan.2026; accepted 16.Jan.2026; published 01.Jun.2026.

Please cite as:
Yeng PK, Fauzi MA, Vestad A, Yang B, De Moor K, Jacobsen C, Diekuu JB, Bettayeb M
Cognitive Dissonance–Based Priming Intervention: Randomized Encouragement With in-the-Wild Phishing Simulation Attack in Health Care
J Med Internet Res 2026;28:e68051
URL: https://www.jmir.org/2026/1/e68051
doi:10.2196/68051

© Prosper Kandabongee Yeng, Muhammad Ali Fauzi, Arnstein Vestad, Bian Yang, Katrien De Moor, Christian Jacobsen, John-Bosco Diekuu, Meriem Bettayeb. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 1.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Communicative Behaviors in an Internet-Based Intervention for Individuals With Autism: Mixed Methods Analysis

Britta Westerberg[1,2], PhD; Karin Jacobson[1,2], MSc; Maria Unenge Hallerbäck[1,3,4], MD, PhD; Susanne Bejerot[1,2], MD, PhD; Fredrik Holländare[1,2], PhD

[1]School of Medical Sciences, Faculty of Medicine and Health, Örebro University, Örebro, Sweden
[2]University Health Care Research Center, Faculty of Medicine and Health, Örebro University, Örebro, Sweden
[3]Department of Public Health Sciences, Karlstad University, Karlstad, Sweden
[4]Centre for Clinical Research, Region Värmland, Karlstad, Sweden

Corresponding Author:
Britta Westerberg, PhD
School of Medical Sciences, Faculty of Medicine and Health
Örebro University
Vuxenhabiliteringen, Box 1613
Örebro, 70116
Sweden
Phone: 46 0767672800
Email: britta.westerberg@oru.se

Abstract

Background: To meet the needs of individuals diagnosed with autism, internet-based interventions have been developed with a variety of objectives. A deeper understanding of the mechanisms of change may help tailor interventions to individual needs. The communicative behaviors of individuals with autism participating in text-based internet-based interventions remain largely unexplored, as do their potential relations to clinical outcomes. An improved understanding of participants' behaviors may help therapists better tailor support, promote engagement, and enhance treatment outcomes.

Objective: This study aimed to explore the communicative behaviors of individuals with autism participating in an internet-based intervention and to examine whether different behavioral patterns were associated with treatment outcomes or treatment adherence.

Methods: Messages from 34 participants enrolled in an 18-week internet-based cognitive behavioral therapy program were analyzed using abductive qualitative content analysis. Correlational analyses were used to examine the relationships between qualitative categories and change scores on outcome measures and rates of module completion.

Results: Fourteen behavioral categories were identified and grouped into three overarching domains: (1) "This is me," which encompasses the participants' narratives on identity, personality, autistic functioning, current and past circumstances, and worldview; (2) "Working with the treatment," which included statements related to engagement with the treatment process; and (3) "I struggle," which comprised of past and present negative experiences and challenges. Correlational analyses revealed associations between several behavioral categories and improvements in quality of life and treatment adherence.

Conclusions: The findings highlight the importance of self-narrative formulation among individuals with autism and suggest that certain communicative behaviors—particularly those involving identity reflection and recognition of treatment-related gains—were positively associated with therapeutic outcomes. The findings enhance our understanding of how individuals with autism engage in internet-based cognitive behavioral therapy and may serve as a valuable source of information for therapists when guiding expectations regarding client outcomes and identifying participants who may benefit from additional support.

Trial Registration: ClinicalTrials.gov NCT03570372; https://clinicaltrials.gov/study/NCT03570372

(J Med Internet Res 2026;28:e76527) doi:10.2196/76527

KEYWORDS

autism; communicative behaviors; internet-based treatment; narrative identity

XSL•FO
RenderX

Introduction

Background

Autism is a neurodevelopmental condition characterized by difficulties in social communication and a stereotypical behavioral pattern [1]. These difficulties imply functional impairment in social situations and may affect psychological well-being [2,3] and low quality of life (QOL) [4]. Furthermore, individuals with autism face several challenges in accessing health care [5], and internet-based interventions offer a convenient and flexible format that may increase accessibility for this population.

The evidence base for internet-based interventions is growing, and an increasing number of diagnostic groups are being offered internet-based cognitive behavioral therapy (ICBT) [6]. Given the efficacy of ICBT, there is a need for an improved understanding of the mechanisms of change, as this knowledge would allow more targeted adaptations of treatment content to the specific needs of each service user [7].

Interventions that include therapist support tend to be more effective than those delivered without guidance [8]. In addition, factors such as participant adherence, treatment credibility, working alliance, and baseline symptom scores have all been identified as important predictors of treatment outcome in ICBT [9]. In particular, therapeutic alliance and treatment adherence are related to favorable outcomes [10,11].

Therapeutic feedback in ICBT is typically provided at regular intervals [7], in response to participant progress, and the participant may have the possibility to write to the therapist at their convenience, both in response to therapist feedback and to initiate a new conversation. Though the core ingredients in ICBT are expected to be embedded within the predetermined text-based modules or sessions, the established importance of therapeutic support [8] and alliance [10] implies that some change mechanisms are influenced by the one-to-one written communication.

The functional components of therapist feedback in ICBT, conceptualized as “therapist behaviors,” are important for both adherence and outcome [12-14]. The client-related equivalent—patient behaviors in ICBT—remains less thoroughly explored. However, previous research indicates that different communicative patterns among participants in ICBT may also be associated with both outcome and adherence [15-17]. Increased insight into what is communicated by clients in ICBT is an important piece of the puzzle in forming an understanding of who is likely to benefit from the program—and who may not—as well as identifying individuals at risk of dropping out. This information can also serve as a valuable indicator to practitioners that a participant is struggling and may require additional support [15].

Svartvatten et al [11] explored behaviors reflected through client messages in ICBT for depression, and their relation to outcome and adherence. They identified 10 behavioral categories and found that “Alliance and Observes positive consequences” correlated positively with changes in outcome, and the behaviors “Alliance,” “Identifies patterns and problem behaviors,” “Maladaptive repetitive thinking,” “Observes positive consequences,” “Tries alternative behavior,” “Chooses alternative behavior,” and “Avoidance of treatment” were positively related to the number of modules completed.

In 2 comparative studies, Soucy et al [18] and Kraepelien et al [19] adopted the behavioral categories from Svartvatten’s study to examine whether these could be generalized to their own data, including participants with depression, anxiety [18], and alcohol use disorder [19]. Although both studies found the predefined categories applicable, notable differences in both frequency and relation to adherence and outcome were observed. Based on inconsistent correlation results, both studies refrained from concluding the predictive value of client behaviors but suggested that the categories may still offer insight into the therapeutic process in ICBT. Soucy et al [18] noted that their directed (also referred to as deductive) content analysis approach—following the categories generated by Svartvatten—may have hindered the detection of other relevant themes. Therefore, they call for future research to analyze client messages inductively, without referring to pre-existing frameworks.

An inductive analytical approach may be especially justified when analyzing communicative patterns among individuals with autism, a population known to exhibit a qualitatively different communicative style compared to the norm, both generally [20] and in text-based online communication [21]. These communication differences may be manifested, for example, by an impaired narrative ability—a communicational skill important for sharing experiences and connecting with others [22]. Furthermore, research has shown that computer-mediated communication often is preferred over verbal communication among individuals with autism [21,23] and that they use internet-based communication in qualitatively different ways than individuals without autism [21,24,25]. These findings open up the possibility that the communicative behaviors in ICBT for autism may differ from those of other patient groups, which justifies approaching this question with an inductive approach, unrestricted by the deductive categories used in previous studies.

To manage the increasing number of individuals diagnosed with autism and to meet their needs, internet-based interventions have been developed with a variety of objectives, including increasing QOL [26], providing psychoeducation [27], and treating psychiatric comorbidities [28-30]. Although the research base on ICBT for autism is relatively scarce, the available studies suggest that internet-based interventions may be both particularly suitable for the needs of individuals with autism [27,31], as well as effective [27,29,30].

Within the current project, an 18-week internet-based intervention adapted for individuals with autism was evaluated for feasibility and effect compared to an active control group [26]. The results were complex and ambiguous; while no group-level effects were found on quantitative outcomes, participant satisfaction was high, and dropout rates were low, indicating good feasibility. A qualitative study on participant experiences of the intervention [31] showed that the participants

XSL•FO
RenderX

generally appreciated the internet-based format, and especially the opportunity to communicate with the therapist in writing.

However, the participants had varying experiences regarding therapist support and the messaging function. Some participants used the messaging function to enable a deeper conversation with the therapist, whereas others only occasionally responded to the therapeutic feedback, resulting in great variability in the amount of therapeutic contact [31]. Accordingly, the level of support was, to a great degree, dependent on the behavior of the participants. Whether this diverse use of the therapeutic support may be related to the treatment effect or adherence is, however, unclear and warrants further investigation. As the active ingredients in ICBT for individuals with autism are yet unexplored, we should remain open to the possibility that the varying communicative behaviors of participants may play a role, even in interventions where no group-level effect was found.

Little is known about the behavior reflected through participant messages in internet-based interventions for individuals with autism, or how variations in these behaviors may relate to treatment outcomes and adherence. Analyzing participant messages in ICBT may offer valuable insights into how individuals with autism engage in the treatment. Furthermore, early identification of specific participant behaviors may inform professionals about what to expect when treating individuals with autism, and thereby guide therapeutic decision-making. Improved understanding of client behaviors could thus be helpful both in identifying individuals at risk of dropping out and in guiding the development of future programs.

Objectives

This study aimed to gain knowledge about participants' communicative behaviors in an internet-based intervention for individuals with autism by investigating the content of participant messages. An additional aim was to explore whether the use of different behaviors (including word count and message frequency) was related to clinical outcome or adherence.

Methods

Design

This study used a mixed method design, using data collected as part of a larger randomized controlled trial (RCT) aimed at evaluating an internet-based cognitive behavioral intervention to improve QOL in autistic adults (ClinicalTrials.gov NCT03570372). A qualitative analysis of participant text messages and the calculation of change scores were first conducted independently, after which the qualitative and quantitative data were integrated through correlation analyses. The integration allowed the exploration of how the qualitative categories were related to quantitative changes.

Participants and Recruitment

All messages from participants in the intervention group (n=42) of an RCT on ICBT for individuals with autism [26] were included in the current qualitative analysis. The RCT was announced through posters in waiting rooms of health care facilities in Örebro county, Sweden, and as advertisements in the local press and on a social media platform (Facebook, Meta Platforms, Inc). Participants for the RCT were recruited by completing a digital self-application form administered through 1177, a Swedish national platform for online health care. The application form included questions on age, gender, living arrangement, occupation, and age at diagnosis. The inclusion process involved a structured interview based on the Mini International Neuropsychiatric Interview, along with a screening questionnaire covering autistic symptoms. Diagnoses were confirmed by collecting assessment records or by verbal confirmation from clinicians. A detailed description of the recruitment process and inclusion criteria of the RCT is provided in Westerberg et al [26].

Intervention

The intervention was based on an evidence-based cognitive behavioral therapy group treatment for adults with autism [32,33], which had been further developed, condensed, and adapted to an internet-based format by the first author, BW.

The intervention lasted 18 weeks. The intervention aimed to enhance QOL and sense of coherence (SOC) and to decrease psychiatric symptoms through the completion of 18 text-based modules focusing on themes relevant to these objectives (refer to Table S1 in Multimedia Appendix 1 for an overview of the themes). Each module included psychoeducation, exercises, and strategies based on cognitive behavioral therapy to enhance coping with everyday life challenges, but adapted to the needs of individuals with autism. In line with recommendations for such adaptations [34], a significant portion of the program was dedicated to psychoeducational content about autism, individual variations, and common comorbidities. The intervention also introduced tools and concepts to support self-understanding and provided terminology to describe participants' functioning. Every 2 weeks, the participants were invited to take part in a live chat session together with other participants, focused on discussing the theme of the most recent module.

In most modules, home exercises were designed to be completed either using text-based worksheets or as "field work" to be reported on the online platform. All modules also included reflective questions to be answered directly in conjunction with the text. The documentation and reports from the exercises, along with responses to the reflective questions, formed the basis for the therapeutic feedback.

Therapeutic feedback was delivered asynchronously via a messaging function available on the same platform as the treatment. The messaging function resembled an email page, allowing participants to respond to therapeutic feedback and initiate new conversations. Therapists were expected to reply within 1 working day. However, participants were not required to send any messages via this function.

Material

All messages written by participants using the messaging function were included. Aside from the removal of information that may reveal participant identities (such as names, telephone numbers, or web addresses), the messages were not modified in any way.

XSL•FO
RenderX

Change scores (from baseline to posttreatment) on outcomes from the RCT were included for a quantitative correlation analysis. The Brunnsviken Brief Quality of Life Scale (BBQ) [35], consisting of 12 items (responses ranging from 0=do not agree to 4=totally agree), was used to assess QOL. Each item related to satisfaction in a specific life area was weighted against a rating of the importance of that area. The multiplied products of each item-pair (satisfaction × importance across 6 life areas) were summed to obtain a total QOL score (range: 0-96), with higher scores representing a higher QOL.

The 13-item Sense of Coherence (SOC-13) scale was used to assess the concept of SOC [36]. SOC reflects the extent to which an individual perceives life events as coherent and comprehensible and life demands as manageable and meaningful. Items are rated on a 7-point scale, with a higher total score (range: 13-91) indicating a stronger SOC.

To assess symptoms of depression and anxiety during the past week, the Hospital Anxiety and Depression Scale (HADS) for depression (HADS-D; 7 items) and anxiety (HADS-A; 7 items) subscales were used [37]. Items are rated on a 4-point scale, with higher scores (range: 0-21) indicating greater symptom severity.

Word count and the frequency of participant and therapist messages were analyzed in relation to the outcome. Generic messages sent to all participants (such as reminders of the peer participant chat sessions) were removed from the data at an early stage. Similarly, messages whose sole purpose was to inform the participant that they had received a new module or to remind them to complete self-assessments were excluded from the analyses.

Qualitative Analysis

The material was analyzed using both a directed and conventional approach to content analysis [38], adopting a combination of inductive and deductive perspectives, resulting in an abductive analytical process [39]. All coding and sorting were conducted using NVivo qualitative data analysis software (Lumivero, LLC) [40] by 2 licensed psychologists (BW and KJ). BW was also 1 of 4 therapists during the trial and was therefore familiar with both the foundational elements and components of the program, as well as the specific setting and design of the trial. As personal identifiers were removed from the texts before analysis, the participants' identities were not revealed through the texts. However, the coding process was not fully blinded, as—due to her previous role as a therapist—BW could recognize certain utterances as stemming from individual participants.

Initially, all messages were read to gain an overall sense of the material and its character. Thereafter, BW segmented and condensed the texts into meaning units, after which BW and KJ jointly and inductively coded the meaning units from 5 participants and discussed the codes to develop a preliminary coding framework. All utterances that could be considered reflective of a behavior were given a descriptive code to capture the manifest content of the participant's messages.

In this analysis, a behavior was defined as statements in the participant messages that appeared to serve a function or convey an intention, either in relation to internal processes or to the treatment work or the therapist. These included both utterances with semantic and meta-linguistic content. Statements that—although functional—were considered not to be of importance for the research questions, such as using polite phrases ("have a good day") or clarifying the structure of the messages ("firstly I will answer your questions"), were excluded.

When a preliminary coding framework was established, the 2 researchers independently coded the texts of 1 participant and compared their interpretations and the wording of codes to refine the codes and to reach consensus regarding the final framework to be used thereafter. Once agreement regarding the wording of codes was achieved, BW coded the communication from the remaining 28 participants in the same manner.

When the texts of approximately 10 individuals had been coded, BW began inductively sorting the codes into clusters illustrating similar behaviors. During this process, a pattern emerged in which several of the clusters resembled categories identified in earlier research [11,18,19]. A decision was made to allow the use of previously identified categories [11] when a category was identified that clearly aligned with one of these. This resulted in an abductive approach, meaning that we moved iteratively between the data and the categories identified in earlier research, which were refined and adjusted to better fit our data during this process. Given the large amount of data and codes, coding and sorting were conducted alternately until all relevant meaning units had been coded and sorted. No third-party validation was conducted during the analysis process.

Quantitative Analysis

The frequencies of each participant's contribution of codes to the categories derived from the qualitative analysis were used in correlation analyses to explore relationships with the quantitative change scores. Change scores on BBQ, SOC-13, HADS-D, and HADS-A were calculated by subtracting preintervention scores from postintervention scores. Normality of the variables was assessed using the Shapiro-Wilk test.

To assess behavior categories in relation to module completion independently of the amount of text written, the individual frequency of codes in each category was divided by the total number of words written by each participant. Accordingly, the frequency of each behavior category was calculated as a proportion of the total number of words, which minimized the influence of overall text length (which would be strongly dependent on the number of modules completed), as the aim was to examine the relative prevalence of behaviors rather than general writing activity.

BBQ, HADS-D, and HADS-A were all normally distributed. However, as the distribution of SOC-13, all behavior category variables, word and message frequency, and module completion rates deviated from normality, Spearman rank-order correlation (ρ) was used to assess the associations between behavior categories and change scores, module completion, and word and message frequency. No adjustment for multiple testing was made, as the study was exploratory and did not involve predefined statistical hypotheses. All quantitative analyses were performed using SPSS Statistics (version 29) [41].

XSL•FO
RenderX

Ethical Considerations

The original trial received ethical approval from the Regional Ethics Committee in Uppsala, Sweden (ref no 2017/392), and the Swedish Ethical Review Authority later approved an amendment (ref no 2022-05792-02) with clarifications regarding this mixed methods study. At inclusion, all participants provided consent for their personal and medical data to be used for research purposes. All data were deidentified before analysis. Personal identifiers were removed from the analyzed text material, participants were assigned unique IDs, and the code key was accessible only to the research team. Participants did not receive any financial compensation for participation in this study. However, a gift card of 300 SEK (≈ US $32) was provided to participants on completion of the postintervention and follow-up assessments of the original RCT.

Results

Overview

As 8 participants did not provide any written material, the qualitative analysis included messages from 34 participants. Table 1 provides the baseline characteristics of the participants and the number of completed modules.

Table 1. Baseline characteristics and adherence of the participants included in the qualitative analysis (N=34).

Characteristic	Value
Age (years), mean (SD)	33.8 (10)
Gender, n (%)	
Men	14 (41.2)
Women	18 (52.9)
Other (nonbinary or transgender)	2 (5.9)
Habitation, n (%)	
With partner and/or children	12 (35.2)
With parents	9 (26.5)
Alone	11 (32.4)
Group or serviced housing	1 (2.9)
Other	1 (2.9)
Age when diagnosed (years), n (%)	
<19	11 (32.4)
20-35	14 (41.2)
>36	9 (26.5)
Occupation, n (%)	
Employed	4 (11.8)
Daily activities	2 (5.9)
Student	4 (11.8)
Job seeker	9 (26.5)
Sick leave	7 (20.6)
Other	8 (23.5)
Completed modules, mean (SD) and median (IQR)	15.6 (4.2) and 18 (14.8-18)

Categories From the Qualitative Content Analysis

In total, 2569 codes were identified, of which 2476 were considered relevant to the aim. These were sorted into 14 categories, which were further grouped into 3 overarching domains: “This is me” (1092 codes and 4 categories), “Working with the treatment” (861 codes and 6 categories), and “I struggle” (523 codes and 4 categories). Table 2 provides an overview and the relative frequency of codes (%) across domains and categories, as well as definitions of each category. Refer to Table S2 in Multimedia Appendix 1 for example codes and quotes from each category.

XSL•FO
RenderX

Table 2. Overview of domains, categories, and the relative frequency of codes (%), and definitions of each category.

Domain and category, %	Definition
This is me, 44.1%	
This is who I am, 15.5%	Texts reflecting a personal narrative about what kind of person one is, what qualities, difficulties, and needs they have, their autistic functioning, and how they (think that they) are perceived by others.
My present and past circumstances, 8.2%	Information on current events, everyday life, positive experiences, and early formative events.
Change is possible, 14.2%	Strategies that work or have potential to work, that they have developed through life, and show motivation and insight that they have the agency to act to progress.
My point of view, 6.2%	Beliefs, opinions, and thoughts about society, autism, human psychology, and the perspective of others.
Working with the treatment, 34.8%	
Appreciation and treatment alliance, 8.4%	Expressions of appreciation, a positive attitude, and bonds of alliance toward the treatment or the therapist.
Putting the treatment aside, 3.2%	Information about not having completed parts of the treatment, and reasons for this, including both psychological issues, such as forgetting and lacking energy, but also external events that got in the way of the treatment work.
Plans to attempt a new task, 7.0%	Reflections regarding how to – or plans to implement a treatment task or exercise, or deciding on a treatment goal.
Have attempted a new task, 4.6%	Reports on completion of a treatment task or exercise, or reflections around the implementation of a task or exercise.
Observing positive consequences of treatment, 3.2%	Statements illustrating that the participant has observed a positive consequence on their personal development from the treatment or specific exercises.
Problems with the treatment, 8.3%	Texts in which the participant expresses difficulties, frustration, or other negative aspects of the treatment content or format, such as failure to understand a task or its purpose, considering certain parts of the treatment irrelevant, or technical problems.
I struggle, 21.1%	
Life is and has been demanding, 11.6%	Descriptions of current and past difficult events or circumstances, that they have been treated badly throughout life, experienced failure, and have been enduring from mental ill-health.
Identifies patterns and problem behaviors, 3.1%	Texts reflecting identification and insight into maladaptive and safety behaviors, and identification of the relation between these behaviors and negative consequences.
I am troubled by mental ill-health, 4.3%	Factual descriptions of current mental ill-health, loneliness, and struggles and their causes and consequences, without engaging in maladaptive thoughts or rumination.
Maladaptive thoughts, 2.2%	Expressions of hopelessness, meaninglessness, or distress, as well as other cognitive distortions, indicate a stagnation of cognitive flexibility.

The domain "This is me" consists of the categories "This is who I am," "My present and past circumstances," "Change is possible," and "My point of view." The domain covers participants' descriptions of themselves in terms of personality, abilities, and difficulties, as well as current and past life situations that have influenced who they are. It further comprises their beliefs and opinions, their view of life and others, how they have used strategies to manage problems, and how they have developed, or are motivated to develop and function better. This domain illustrates that participants have insight into their agency, responsibility, and ability to influence their situation, but also reflects self-awareness regarding how their autistic difficulties pose challenges in this.

The domain "Working with the treatment" contains the categories "Appreciation and treatment alliance," "Putting the treatment aside," "Plans to attempt a new task," "Have attempted a new task," "Observing positive consequences of treatment," and "Problems with the treatment." These categories include statements that are in some way related to participants' engagement in the treatment. It involves both positive and negative experiences and effects of the treatment, as well as reports on completed tasks and plans for future tasks. Reasons why certain parts of the treatment were not carried out are also included in this domain.

The domain "I struggle" contains the categories "Life is and has been demanding," "Identifies patterns and problem behaviors," "I am troubled by mental ill-health," and "Maladaptive thinking." This domain contains descriptions of past and current negative events, experiences of failure, descriptions of a difficult life situation, experiences of mistreatment, and how these have negatively affected mental health. It also covers how their mental ill-health manifests in symptoms and negative behavioral patterns.

XSL•FO
RenderX

Nine of the categories were found to correspond to categories identified in previous studies. Refer to Table S3 in Multimedia Appendix 1 for an overview of the categories from this study and—where applicable—the corresponding category from the studies by Svartvatten et al [11] and Kraepelien et al [19]. Five categories not previously described by the literature were identified in this study. These are: "This is who I am," "My present and past circumstances," "Change is possible," "My point of view, and "Life is and has been demanding." Two categories identified in previous research, "Confrontational alliance rupture" [11,18,19] and "Observes alcohol-related setback" [19], were not identified in our data.

Correlation Between Frequency of Behaviors and Outcome

Table 3 provides baseline, posttreatment, and change scores on outcomes, and Table 4 provides correlations between the frequency of participant behaviors and change scores of outcomes and the number of completed modules. As 5 participants—although included in the qualitative analysis—did not complete the postassessment, the correlational analyses include data from 29 participants.

The domains "This is me" and "Working with the treatment" and the categories "This is who I am," "Change is possible," "My point of view," "Appreciation and treatment alliance," "Plans to attempt a new task," "Have attempted a new task," and "Observing positive consequences of treatment" correlated significantly with a positive change in BBQ. The domain "This is me" and the categories "This is who I am," "Change is possible," and "Identifies patterns and problem behaviors" correlated significantly with a higher number of completed modules. No other significant correlations between the frequency of behaviors and outcomes were observed.

These findings show that when participants more frequently expressed aspects of self-understanding (as in the domain "This is me") and engaged in behaviors related to the treatment (as in the domain "Working with the treatment"), they tended to experience greater improvements in QOL (BBQ) and complete more treatment modules. This pattern indicates that behaviors reflecting active engagement and self-reflection may play an important role in facilitating positive treatment outcomes.

A total of 515 participant messages were sent. The median number of words per participant was 825 (IQR 530-2110; range 64-5562), and the median number of messages per participant was 11 (IQR 5-18; range 1-75). There was a significant moderate positive correlation between change in BBQ score and the number of participant words (ρ=0.52; *P*=<.001) and the number of participant messages (ρ=0.42; *P*=.02). There was no significant correlation between the number of words written by the therapist and any of the outcome variables.

Table 3. Means and distribution of pre- and posttreatment scores, change scores, and paired samples t test results on outcomes of the participants included in the quantitative analysis (N=29).

Outcome	Pre, mean (SD); range	Post, mean (SD); range	Change score[a], mean (SD)	*t* test (*df*)	*P* value
BBQ[b]	41.14 (21.82); 0-96	45.90 (24.19); 0-96	4.76 (14.46)	–1.77 (28)	.09
SOC-13[c]	48.17 (15.01); 22-90	49.07 (14.49); 21-85	0.90 (9.82)	–0.49 (28)	.63
HADS-D[d]	7.45 (4.03); 0-15	6.83 (5.34); 0-19	–0.62 (3.63)	0.92 (28)	.37
HADS-A[d]	12.31 (4.97); 2-20	11.45 (5.10); 1-21	–0.86 (3.0)	1.55 (28)	.13

[a]Change scores represent posttreatment minus baseline values.

[b]BBQ: Brunnsviken Brief Quality of Life Scale.

[c]SOC-13: 13-item Sense of Coherence scale.

[d]HADS-D and HADS-A: Hospital Anxiety and Depression Scale for depression and anxiety subscales.

XSL•FO
RenderX

Table 4. Correlations (Spearman ρ) between relative frequency of participant behavior and outcome change-scores and module completion (N=29).

Domain and behavioral category	BBQ[a]		SOC-13[b]		HADS-D[c]		HADS-A[c]		Module completed[d]	
	ρ	*P* value	P	*P* value	ρ	*P* value	ρ	*P* value	ρ	*P* value
This is me	0.51	<.001	0.04	.83	–0.21	.27	–0.05	.80	0.45	.01
This is who I am	0.50	.01	0.08	.69	–0.19	.31	–0.02	.91	0.43	.01
My present and past circumstances	0.32	.09	0.00	1.0	–0.21	.27	–0.08	.67	0.21	.24
Change is possible	0.39	.04	0.09	.65	–0.20	.30	0.02	.93	0.37	.03
My point of view	0.53	<.001	0.21	.27	–0.14	.48	–0.04	.84	0.16	.36
Working with the treatment	0.50	.01	0.05	.82	–0.25	.20	0.02	.92	–0.11	.55
Appreciation and treatment alliance	0.44	.02	–0.06	.77	–0.11	.57	0.07	.74	–0.08	.65
Putting treatment aside	0.25	.19	–0.13	.49	–0.34	.07	–0.09	.64	–0.09	.62
Plans to attempt a new task	0.39	.04	–0.02	.92	–0.19	.31	0.12	.53	–0.00	.99
Have attempted a new task	0.41	.03	–0.06	.77	–0.23	.23	–0.03	.89	0.09	.63
Observing positive consequences	0.50	.01	0.12	.53	–0.22	.24	–0.03	.86	0.30	.08
Problems with the treatment	0.27	.16	0.02	.90	–0.14	.46	0.08	.67	0.18	.31
I struggle	0.31	.10	0.03	.88	0.05	.79	0.20	.31	–0.06	.75
Life is and has been demanding	0.32	.09	–0.10	.62	–0.04	.83	0.12	.53	–0.15	.40
Identifies patterns and problem behaviors	0.31	.11	0.12	.55	–0.08	.68	0.00	1.0	0.36	.03
I am troubled by mental ill-health	0.11	.56	–0.00	.99	0.07	.71	0.26	.17	0.32	.06
Maladaptive thoughts	0.14	.47	0.31	.10	0.26	.18	0.20	.30	–0.04	.82

[a]BBQ: Brunnsviken Brief Quality of Life Scale.

[b]SOC-13: 13-item Sense of Coherence scale.

[c]HADS-D and HADS-A: Hospital Anxiety and Depression Scale for depression and anxiety subscales.

[d]Controlled for the total number of words.

Discussion

Principal Findings

The purpose of this study was to explore written communicative behaviors among individuals with autism participating in an internet-based intervention aimed at improving QOL and to examine whether any of these behaviors were associated with treatment outcomes.

The results show that statements related to participants' descriptions of themselves in terms of personality, abilities, beliefs, experiences, and personal development were highly prevalent, with nearly half of the codes (44.1%) categorized under the domain "This is me." The most common category was "This is who I am" (15.5%), which included reflections on personality and autistic functioning. The second most frequent category was "Change is possible" (14.2%), consisting of statements expressing awareness that personal development and growth are possible, such as narratives of past growth or coping strategies perceived as helpful. "Life is and has been demanding," also accounting for a substantial portion (11.6%) of the codes, captured descriptions of adversity, experiences of mistreatment, failure, and mental ill-health.

Correlation analyses revealed that the domains "This is me" and "Working with the treatment," as well as the categories "My point of view," "This is who I am," and "Observing positive consequences of treatment," were significantly moderately associated with improvements in BBQ scores. These findings suggest that the opportunity to reflect openly on identity, personal perspectives, and the perceived impact of treatment may be particularly important for improving QOL in individuals with autism. Notably, the finding regarding "Observing positive consequences of treatment" aligns with previous findings by Soucy et al [18] and Svartvatten et al [11], who also reported a positive correlation between this behavior and treatment outcomes.

The significant moderate positive relation between BBQ scores and the category "Appreciation and treatment alliance" further supports a well-established body of research highlighting the importance of therapeutic alliance for the therapeutic outcomes [9-11,42,43]. In addition, the category "Have attempted a new task" was moderately positively correlated with changes in

XSL•FO
RenderX

BBQ, whereas "Change is possible" and "Plans to attempt a new task" showed weak associations. These findings indicate that motivational expressions related to trying new strategies and behaviors—both within and beyond the treatment context—may reflect treatment engagement and efficacy.

Moreover, the number of words and messages written by participants was also positively associated with improvement in BBQ scores. While earlier studies examining text quantity have primarily focused on its relation to adherence—showing that more words are associated with increased program completion [17,44]—our results indicate that text quantity may also be related to changes in outcomes in ICBT for individuals with autism.

Interestingly, while several communicative behaviors were associated with improvements in QOL (as measured by BBQ), no associations were observed with changes in anxiety or depression (HADS) or sense of coherence (SOC-13). One possible explanation is that QOL, as a subjective and existential construct, may be more readily influenced by participants' self-reflective and narrative expressions. In contrast, clinical symptoms of anxiety and depression, and the more stable SOC, may be less susceptible to short-term fluctuations driven by communicative behaviors. Despite earlier findings suggesting that SOC may improve from interventions [45]—which motivated its use as an outcome in the original RCT—this effect was not found in our study [26]. Another potential explanation is that the relatively greater pre-post variability in BBQ, compared with the smaller mean changes observed in HADS and SOC-13 scores, may have increased the likelihood of detecting statistically significant correlations for QOL but not for the other measures.

Regarding adherence, the domain "This is me" and the category "This is who I am" were moderately associated with the number of completed modules, while the categories "Change is possible" and "Identifies patterns and problem behaviors" showed weak associations. This partly supports previous findings by both Soucy et al [18] and Svartvatten et al [11], who also reported a positive correlation between "Identifies patterns and problem behaviors" and the number of modules completed.

Although no causal inferences can be made, the observed associations in this study may shed light on potential behavioral indicators that are beneficial for individuals with autism to engage in during internet-based psychological interventions.

The relatively large number of identified categories (n=14) illustrates a broad range of content shared by participants with their therapists and likely reflects the wide diversity among individuals with autism. Unlike ICBT, which targets specific disorders, this intervention focused on the individual's whole life, and given the comprehensive scope of the treatment (refer to Table S1 in Multimedia Appendix 1), the extensive result is not surprising. Furthermore, an obvious reason for the relatively high prevalence of specific behaviors (ie, those reflected in the domain "This is me" and the categories "This is who I am" and "Change is possible") is related to the treatment content itself, in which reflections on the self in relation to past, present, and future experiences were repeatedly encouraged.

Consequently, the content of this treatment may partly explain why the behavior categories in the current analysis were only partially similar to those identified in earlier studies [11,18,19]. The 4 categories constituting the domain "This is me," and the category "Life is and has been demanding," including meaning units not covered by earlier frameworks (Table S2 in Multimedia Appendix 1). In agreement with Soucy et al [18], our study supports the notion that a fully deductive process—that is, following the existing coding framework—would have hindered the detection of relevant themes.

Beyond treatment content, these divergent findings (compared with previous studies) must be considered in light of the unique characteristics of the study population—individuals with autism. A considerable proportion of the participant messages centered around identity, functioning, and experiences in relation to autism or psychiatric comorbidity, which were manifested in several ways. Some statements offered nonvaluing descriptions of how their autism contributed to shaping their identity, and others accounted for experiences and traits as contributing to personality formation independent of autistic traits. Regardless of etiology, it appeared fundamental for the participants to construct a narrative around who they are and how their functioning affects their lives. This internalized self-narrative—defining who they are, how they came to be, and where their life may be heading—can be described as their narrative identity [46,47].

Through the construction of a narrative identity, people define and communicate who they are to themselves and others [47,48]. This process fosters self-insight and understanding, allowing negative experiences to be reframed as opportunities for growth, which in turn has been linked to greater well-being [49,50]. Given the precondition of atypical functioning, the process of constructing a coherent self-narrative thus appears particularly relevant for individuals with autism. While research suggests that narrative skills [22,51] and identity formation [52] may be impaired in autism, Samra [53] highlights the importance of engaging with one's narrative identity. In their thesis on identity formation in individuals with autism, they argue that this engagement enhances self-awareness, supports meaning-making, and helps to envision one's future.

Our findings demonstrated a strong engagement in this narrative process, and in line with the suggestion of Samra [53], a positive association was observed between "This is me" statements and improvement in QOL. However, categories within the "I struggle" domain also reflected engagement in a narrative process, but without any positive correlation with QOL. This indicates that different narrative processes may serve different functions, consistent with research arguing that different modes of self-focus have distinct functional properties [54,55]. For example, Watkins et al [54-56] differentiate between maladaptive analytical self-focus—described as abstract, evaluative, and often ruminative—and adaptive experiential self-focus, which is concrete, process-oriented, and linked to positive self-evaluation.

Although we did not deductively categorize statements based on this distinction—maladaptive vs adaptive self-focus—our findings align with the theory and may still be interpreted within

XSL•FO
RenderX

this framework, offering potential guidance for clinicians in identifying and supporting beneficial communicative behaviors.

Limitations

This study design has several limitations. First, only text written via the messaging function of the intervention was included in the analysis; text written directly in response to exercises was excluded, as it was typically constrained by predefined questions and exercises, making it less relevant for the aim of identifying spontaneous communicative behaviors.

It should be noted that the participants included in this study are not fully representative of the entire autistic population, as a natural selection of individuals already familiar with technology applies for inclusion in a trial such as this. Moreover, participants who did not write anything were excluded from the current analyses, which may imply a certain selection bias, as these individuals might have refrained from messaging due to communication difficulties. Those failing to complete the postassessment were also excluded from the correlation analysis, which means that we are unable to conclude the relationship between behavior and outcome for these individuals. However, post hoc analyses with imputed zeros modeling noncorrespondence, as well as imputed zeros for change scores on quantitative outcomes, were conducted, showing that these procedures did not affect the correlational outcomes (data not shown).

Another limitation concerns the role of the primary analyst (BW), who also served as a therapist and was involved in developing the intervention. This dual role introduces a risk of bias, potentially influencing data interpretation, coding, and categorization. To minimize this bias, the initial phase of the analysis was conducted jointly with KJ, who had no previous involvement in treatment development or delivery. Furthermore, the coding approach was deliberately inductive and aimed to be as close to the text as possible. Nevertheless, future studies would benefit from involving fully independent third-party analysts to further strengthen the trustworthiness of the findings.

Another methodological limitation in this study is that the intercoder reliability was not statistically calculated. Typically, around 10%-25% of data units are coded by more than 1 researcher to facilitate a trustworthy estimate of intercoder reliability [57]. By collaborative coding of 5 (14.71%) of the data units in this study, the analysts could compare and discuss codes to develop a preliminary coding framework. By subsequently coding 1 of the participants' texts independently, the coding framework was further refined and agreed upon. We considered this to be sufficient to establish the final coding framework, but acknowledge that the reliability would be strengthened with additional collaborative coding. Furthermore, in future studies, interrater reliability should be assessed using a statistical test, for example, Cohen κ.

As our analysis did not account for the timing of the statements, it is uncertain whether any detection of potentially beneficial behaviors can be made at an early stage in the therapy. Early detection (or absence) of these behaviors could enable more individualized therapeutic support, such as encouraging these behaviors. Given the novelty of this research area and the fact that similar data have not previously been studied, we chose in this study to focus on broader behavioral patterns rather than their temporal distribution. Future research could consider focusing on early-stage behaviors as potential indicators of treatment response.

When interpreting the results from the correlation analyses, it should be noted that no correction for multiple comparisons was applied. Given the relatively large number of behavioral categories and correlations tested, there is an increased risk of inflated false positives, and the results should be interpreted with caution.

Finally, in the correlational analyses between behaviors and outcome, we did not control for therapist word count or therapist behaviors. Given previous research emphasizing the importance of therapist support and the varying number of therapist messages, therapist behavior may have influenced the relationships. However, since the therapist's word count was not directly correlated with outcomes in this study, it was not included as a covariate. Future studies should nevertheless consider therapist variables when exploring outcome predictors in ICBT for individuals with autism.

Conclusions

The findings of this study increase our understanding of how individuals with autism engage in ICBT and may serve as a valuable source of information for therapists in guiding both their expectations of client outcomes and identifying client behaviors to support. The results emphasize the importance of self-narrative formulation in individuals with autism and suggest that certain communicative behaviors—especially those involving identity reflection and recognition of treatment benefits—are positively related to improvements in QOL. Future research should explore ways to identify such behaviors early in treatment as potential indicators of individuals at risk of poorer outcomes, enabling clinicians to provide additional support or tailor interventions to better suit individual needs.

Acknowledgments

We thank all participants who contributed their text-based data for this study. We also thank associate professor Yang Cao, Örebro University, for guidance regarding the statistical analyses. The generative artificial intelligence (GAI) tool ChatGPT (OpenAI) was used for language editing by suggesting alternative phrasings of long and/or complex sentences or terms to improve clarity and readability. According to the GAIDeT taxonomy, the following tasks were delegated to GAI tools under full human supervision: proofreading and editing.

XSL•FO
RenderX

Funding

This work was financially supported by the Region Örebro Län, Sweden (grant nos OLL-935396, OLL-879651, OLL-887401, OLL-833131, OLL-785501, OLL-736321, OLL-878311, OLL-785311, OLL-985129, and OLL-1004448) and Stiftelsen Bror Gadelius Minnesfond (grant no 23003). The funding sources had no role in the study design, data collection, analysis, interpretation, writing of the report, or the decision to submit the article for publication.

Data Availability

The data supporting this study originates from a longitudinal intervention study conducted at the University Health Care Research Center, Örebro Region, Sweden. Data are available upon reasonable request from the corresponding author.

Authors' Contributions

BW led the conceptualization of the study, with FH providing supporting input. BW was responsible for data curation. Formal analysis was conducted by BW as the lead analyst, with KJ providing support. Funding acquisition was led by BW, with SB and FH providing supporting contributions. BW carried out the investigation. Methodology was led by BW, with FH supporting. Project administration was managed by BW. Supervision was led by FH, with SB providing support. BW drafted the original manuscript. Review and editing of the manuscript were led by BW, with FH, SB, KJ, and MUH providing supporting contributions.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Treatment modules in the internet-based intervention MILAS: overview of domains, categories, definitions, example codes, and quotes from this study and, where applicable, the corresponding categories from previous studies.
[DOCX File , 37 KB - jmir_v28i1e76527_app1.docx]

References

1. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). Arlington, VA: Americal Psychiatric Association; 2013.
2. Pascoe MI, Forbes K, de la Roche L, Derby B, Psaradellis E, Anagnostou E, et al. Exploring the association between social skills struggles and social communication difficulties and depression in youth with autism spectrum disorder. Autism Res 2023;16(11):2160-2171. [doi: 10.1002/aur.3015] [Medline: 37615265]
3. Rai D, Culpin I, Heuvelman H, Magnusson CMK, Carpenter P, Jones HJ, et al. Association of autistic traits with depression from childhood to age 18 years. JAMA Psychiatry 2018;75(8):835-843 [FREE Full text] [doi: 10.1001/jamapsychiatry.2018.1323] [Medline: 29898212]
4. Mason D, McConachie H, Garland D, Petrou A, Rodgers J, Parr JR. Predictors of quality of life for autistic adults. Autism Res 2018;11(8):1138-1147 [FREE Full text] [doi: 10.1002/aur.1965] [Medline: 29734506]
5. Adams D, Young K. A systematic review of the perceived barriers and facilitators to accessing psychological treatment for mental health problems in individuals on the autism spectrum. Rev J Autism Dev Disord 2020;8(4):436-453. [doi: 10.1007/s40489-020-00226-7]
6. Andersson G. Internet interventions: past, present and future. Internet Interv 2018;12:181-188 [FREE Full text] [doi: 10.1016/j.invent.2018.03.008] [Medline: 30135782]
7. Richards D, Enrique A, Palacios J, Duffy D. Internet-delivered cognitive behaviour therapy. In: Cognitive Behavioral Therapy and Clinical Applications. London, United Kingdom: IntechOpen; 2018.
8. Furukawa TA, Suganuma A, Ostinelli EG, Andersson G, Beevers CG, Shumake J, et al. Dismantling, optimising, and personalising internet cognitive behavioural therapy for depression: a systematic review and component network meta-analysis using individual participant data. Lancet Psychiatry 2021;8(6):500-511 [FREE Full text] [doi: 10.1016/S2215-0366(21)00077-8] [Medline: 33957075]
9. Haller K, Becker P, Niemeyer H, Boettcher J. Who benefits from guided internet-based interventions? A systematic review of predictors and moderators of treatment outcome. Internet Interv 2023;33:100635 [FREE Full text] [doi: 10.1016/j.invent.2023.100635] [Medline: 37449052]
10. Kaiser J, Hanschmidt F, Kersting A. The association between therapeutic alliance and outcome in internet-based psychological interventions: a meta-analysis. Comput Hum Behav 2021;114:106512. [doi: 10.1016/j.chb.2020.106512]
11. Svartvatten N, Segerlund M, Dennhag I, Andersson G, Carlbring P. A content analysis of client e-mails in guided internet-based cognitive behavior therapy for depression. Internet Interv 2015;2(2):121-127. [doi: 10.1016/j.invent.2015.02.004]
12. Schneider LH, Hadjistavropoulos HD, Faller YN. Internet-delivered cognitive behaviour therapy for depressive symptoms: an exploratory examination of therapist behaviours and their relationship to outcome and therapeutic alliance. Behav Cogn Psychother 2016;44(6):625-639. [doi: 10.1017/S1352465816000254] [Medline: 27302220]

XSL•FO
RenderX

13. Holländare F, Gustafsson SA, Berglind M, Grape F, Carlbring P, Andersson G, et al. Therapist behaviours in internet-based cognitive behaviour therapy (ICBT) for depressive symptoms. Internet Interv 2016;3:1-7 [FREE Full text] [doi: 10.1016/j.invent.2015.11.002] [Medline: 30135783]
14. Paxling B, Lundgren S, Norman A, Almlöv J, Carlbring P, Cuijpers P, et al. Therapist behaviours in internet-delivered cognitive behaviour therapy: analyses of e-mail correspondence in the treatment of generalized anxiety disorder. Behav Cogn Psychother 2013;41(3):280-289. [doi: 10.1017/S1352465812000240] [Medline: 22717145]
15. Dirkse D, Hadjistavropoulos HD, Hesser H, Barak A. Linguistic analysis of communication in therapist-assisted internet-delivered cognitive behavior therapy for generalized anxiety disorder. Cogn Behav Ther 2015;44(1):21-32. [doi: 10.1080/16506073.2014.952773] [Medline: 25244051]
16. Van der Zanden R, Curie K, Van Londen M, Kramer J, Steen G, Cuijpers P. Web-based depression treatment: associations of clients' word use with adherence and outcome. J Affect Disord 2014;160:10-13 [FREE Full text] [doi: 10.1016/j.jad.2014.01.005] [Medline: 24709016]
17. Linnet J, Jensen ES, Runge E, Hansen MB, Hertz SPT, Mathiasen K, et al. Text based internet intervention of binge eating disorder (BED): words per message is associated with treatment adherence. Internet Interv 2022;28:100538 [FREE Full text] [doi: 10.1016/j.invent.2022.100538] [Medline: 35480237]
18. Soucy JN, Hadjistavropoulos HD, Couture CA, Owens VA, Dear BF, Titov N. Content of client emails in internet-delivered cognitive behaviour therapy: a comparison between two trials and relationship to client outcome. Internet Interv 2018;11:53-59 [FREE Full text] [doi: 10.1016/j.invent.2018.01.006] [Medline: 30135760]
19. Kraepelien M, Hadjistavropoulos HD, Berman AH, Sundström C. Exploring client messages in a therapist-guided internet intervention for alcohol use disorders - a content analysis. Internet Interv 2021;26:100483 [FREE Full text] [doi: 10.1016/j.invent.2021.100483] [Medline: 34824984]
20. Rollins PR. Narrative skills in young adults with high-functioning autism spectrum disorders. Commun Disord Q 2014;36(1):21-28. [doi: 10.1177/1525740114520962]
21. Gillespie-Lynch K, Kapp SK, Shane-Simpson C, Smith DS, Hutman T. Intersections between the autism spectrum and the internet: perceived benefits and preferred functions of computer-mediated communication. Intellect Dev Disabil 2014;52(6):456-469. [doi: 10.1352/1934-9556-52.6.456] [Medline: 25409132]
22. Nayar K, Landau E, Martin GE, Stevens CJ, Xing J, Sophia P, et al. Narrative ability in autism and first-degree relatives. J Autism Dev Disord 2025;55(11):3822-3837. [doi: 10.1007/s10803-024-06424-0] [Medline: 39060703]
23. Howard PL, Sedgewick F. 'Anything but the phone!': communication mode preferences in the autism community. Autism 2021;25(8):2265-2278 [FREE Full text] [doi: 10.1177/13623613211014995] [Medline: 34169750]
24. van der Aa C, Pollmann MMH, Plaat A, van der Gaag RJ. Computer-mediated communication in adults with high-functioning autism spectrum disorders and controls. Res Autism Spectr Disord 2016;23:15-27. [doi: 10.1016/j.rasd.2015.11.007]
25. Caldwell-Harris CL, Posner SD. When autistic writing is superior to neurotypical writing: the case of blogs. Educ Rev 2024;76(7):1875-1897. [doi: 10.1080/00131911.2024.2302119]
26. Westerberg B, Holländare F, Bejerot S. An internet-based behavioral intervention for adults with autism spectrum disorder - a randomized controlled trial and feasibility study. Internet Interv 2023;34:100672 [FREE Full text] [doi: 10.1016/j.invent.2023.100672] [Medline: 37772160]
27. Backman A, Roll-Pettersson L, Mellblom A, Norman-Claesson E, Sundqvist E, Zander E, et al. Internet-delivered psychoeducation (SCOPE) for transition-aged autistic youth: pragmatic randomized controlled trial. J Med Internet Res 2024;26:e49305 [FREE Full text] [doi: 10.2196/49305] [Medline: 39608000]
28. Conaughton RJ, Donovan CL, March S. Efficacy of an internet-based CBT program for children with comorbid high functioning autism spectrum disorder and anxiety: a randomised controlled trial. J Affect Disord 2017;218:260-268. [doi: 10.1016/J.JAD.2017.04.032]
29. Georén L, Jansson-Fröjmark M, Nordenstam L, Andersson G, Olsson NC. Internet-delivered cognitive behavioral therapy for insomnia in youth with autism spectrum disorder: a pilot study. Internet Interv 2022;29:100548 [FREE Full text] [doi: 10.1016/j.invent.2022.100548] [Medline: 35651733]
30. Wickberg F, Lenhard F, Aspvall K, Serlachius E, Andrén P, Johansson F, et al. Feasibility of internet-delivered cognitive-behavior therapy for obsessive-compulsive disorder in youth with autism spectrum disorder: a clinical benchmark study. Internet Interv 2022;28:100520 [FREE Full text] [doi: 10.1016/j.invent.2022.100520] [Medline: 35281701]
31. Westerberg B, Bäärnhielm S, Giles C, Hylén U, Holländare F, Bejerot S. An internet based intervention for adults with autism spectrum disorder-a qualitative study of participants experiences. Front Psychiatry 2021;12:789663 [FREE Full text] [doi: 10.3389/fpsyt.2021.789663] [Medline: 35002808]
32. Bejerot S. ALMA - KBT för vuxna med autismspektrumsyndrom, manual och arbetsbok [ALMA-CBT for adults with autism spectrum disorder, Manual and Workbook]. United States: Hogrefe Publishing; 2019.
33. Hesselmark E, Plenty S, Bejerot S. Group cognitive behavioural therapy and group recreational activity for adults with autism spectrum disorders: a preliminary randomized controlled trial. Autism 2014;18(6):672-683 [FREE Full text] [doi: 10.1177/1362361313493681] [Medline: 24089423]

XSL•FO
RenderX

34. Kerns CM, Roux AM, Connell JE, Shattuck PT. Adapting cognitive behavioral techniques to address anxiety and depression in cognitively able emerging adults on the autism spectrum. Cogn Behav Pract 2016;23(3):329-340. [doi: 10.1016/j.cbpra.2016.06.002]
35. Lindner P, Frykheden O, Forsström D, Andersson E, Ljótsson B, Hedman E, et al. The Brunnsviken Brief Quality of life scale (BBQ): development and psychometric evaluation. Cogn Behav Ther 2016;45(3):182-195 [FREE Full text] [doi: 10.1080/16506073.2016.1143526] [Medline: 26886248]
36. Antonovsky A. The structure and properties of the Sense of Coherence scale. Soc Sci Med 1993;36(6):725-733. [doi: 10.1016/0277-9536(93)90033-z] [Medline: 8480217]
37. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand 1983;67(6):361-370. [doi: 10.1111/j.1600-0447.1983.tb09716.x] [Medline: 6880820]
38. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
39. Råholm MB. Abductive reasoning and the formation of scientific knowledge within nursing research: abductive reasoning within nursing research. Nurs Philos 2010;11(4):260-270. [doi: 10.1111/j.1466-769x.2010.00457.x]
40. Nvivo. Lumivero. 2023. URL: https://lumivero.com/ [accessed 2026-01-06]
41. IBM SPSS Statistics. IBM. 2023. URL: https://www.ibm.com/products/spss-statistics [accessed 2026-01-09]
42. Lindqvist K, Mechler J, Falkenström F, Carlbring P, Andersson G, Philips B. Therapeutic alliance is calming and curing-the interplay between alliance and emotion regulation as predictors of outcome in internet-based treatments for adolescent depression. J Consult Clin Psychol 2023;91(7):426-437. [doi: 10.1037/ccp0000815] [Medline: 37166833]
43. Bergman Nordgren L, Carlbring P, Linna E, Andersson G. Role of the working alliance on treatment outcome in tailored internet-based cognitive behavioural therapy for anxiety disorders: randomized controlled pilot trial. JMIR Res Protoc 2013;2(1):e4 [FREE Full text] [doi: 10.2196/resprot.2292] [Medline: 23612437]
44. Wallert J, Gustafson E, Held C, Madison G, Norlund F, von Essen L, et al. Predicting adherence to internet-delivered psychotherapy for symptoms of depression and anxiety after myocardial infarction: machine learning insights from the U-CARE heart randomized controlled trial. J Med Internet Res 2018;20(10):e10754 [FREE Full text] [doi: 10.2196/10754] [Medline: 30305255]
45. Valtonen M, Raiskila T, Veijola J, Läksy K, Kauhanen M, Kiuttu J, et al. Enhancing sense of coherence via early intervention among depressed occupational health care clients. Nord J Psychiatry 2015;69(7):515-522. [doi: 10.3109/08039488.2015.1011230] [Medline: 25739527]
46. McAdams DP. Narrative identity. In: Schwartz S, Luyckx K, Vignoles V, editors. Handbook of Identity Theory and Research. New York, NY: Springer; 2011:99-115.
47. McAdams DP, McLean KC. Narrative identity. Curr Dir Psychol Sci 2013;22(3):233-238. [doi: 10.1177/0963721413475622]
48. McAdams DP. Narrative identity: what is it? What does it do? How do you measure it? Imagin Cogn Personal 2018;37(3):359-372. [doi: 10.1177/0276236618756704]
49. Lilgendahl JP, McAdams DP. Constructing stories of self-growth: how individual differences in patterns of autobiographical reasoning relate to well-being in midlife. J Pers 2011;79(2):391-428 [FREE Full text] [doi: 10.1111/j.1467-6494.2010.00688.x] [Medline: 21395593]
50. Waters TEA, Fivush R. Relations between narrative coherence, identity, and psychological well-being in emerging adulthood. J Pers 2015;83(4):441-451 [FREE Full text] [doi: 10.1111/jopy.12120] [Medline: 25110125]
51. McCabe A, Hillier A, Shapiro C. Brief report: structure of personal narratives of adults with autism spectrum disorder. J Autism Dev Disord 2013;43(3):733-738. [doi: 10.1007/s10803-012-1585-x] [Medline: 22767138]
52. Allé MC, Schneider P, Rigoulot L, Gandolphe MC, Danion JM, Coutelle R, et al. Narrative identity alterations in autism spectrum disorder: a life story approach. Research Square Preprint posted online on April 30, 2024. [doi: 10.21203/rs.3.rs-4292403/v1]
53. Samra HS. A narrative exploration of sense-making, self, and identity in young people diagnosed with an autism spectrum condition [doctoral dissertation]. School of Education, University of Birmingham. 2016. URL: https://etheses.bham.ac.uk/id/eprint/6719/1/Samra16EdPsychD.pdf [accessed 2026-01-20]
54. Vassilopoulos SP, Watkins ER. Adaptive and maladaptive self-focus: a pilot extension study with individuals high and low in fear of negative evaluation. Behav Ther 2009;40(2):181-189. [doi: 10.1016/j.beth.2008.05.003] [Medline: 19433149]
55. Watkins E, Teasdale JD. Adaptive and maladaptive self-focus in depression. J Affect Disord 2004;82(1):1-8. [doi: 10.1016/j.jad.2003.10.006] [Medline: 15465571]
56. Watkins E. Adaptive and maladaptive ruminative self-focus during emotional processing. Behav Res Ther 2004;42(9):1037-1052. [doi: 10.1016/j.brat.2004.01.009] [Medline: 15325900]
57. O'Connor C, Joffe H. Intercoder reliability in qualitative research: debates and practical guidelines. Int J Qual Methods 2020;19:160940691989922. [doi: 10.1177/1609406919899220]

Abbreviations

BBQ: Brunnsviken Brief Quality of Life Scale

XSL•FO
RenderX

HADS: Hospital Anxiety and Depression Scale
HADS-A: Hospital Anxiety and Depression Scale for the anxiety subscale
HADS-D: Hospital Anxiety and depression Scale for the depression subscale
ICBT: internet-based cognitive behavioral therapy
QOL: quality of life
RCT: randomized controlled trial
SOC: sense of coherence
SOC-13: 13-item Sense of Coherence scale

Edited by A Stone; submitted 19.May.2025; peer-reviewed by J Wang, C Abbatantuono, E Zander; comments to author 07.Nov.2025; revised version received 30.Dec.2025; accepted 30.Dec.2025; published 04.Feb.2026.

Please cite as:
Westerberg B, Jacobson K, Unenge Hallerbäck M, Bejerot S, Holländare F
Communicative Behaviors in an Internet-Based Intervention for Individuals With Autism: Mixed Methods Analysis
J Med Internet Res 2026;28:e76527
URL: https://www.jmir.org/2026/1/e76527
doi:10.2196/76527
PMID:

©Britta Westerberg, Karin Jacobson, Maria Unenge Hallerbäck, Susanne Bejerot, Fredrik Holländare. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Quality Evaluation of Large Language Model–Assisted Generation of Initial Senior Physician Ward Round Records for Patients With Acute Poisoning: Cross-Sectional Study

Junping Zhu[1], MD; Wei Pan[1], MD; Yonghong Wang[1], MD; Kui Yan[1], MD; Zhicheng Fang[1,2], MD; Xianyi Yang[1,2], MD

[1]Department of Emergency Medicine, Taihe Hospital, Hubei University of Medicine, Renmin South Road 32, Maojian District, Shiyan, Hubei, China

[2]Hubei Provincial Clinical Medical Research Center for Pneumoconiosis and Poisoning, Hubei Provincial Hospital of Integrated Chinese & Western Medicine, Wuhan, Hubei, China

Corresponding Author:
Xianyi Yang, MD
Department of Emergency Medicine, Taihe Hospital, Hubei University of Medicine, Renmin South Road 32, Maojian District, Shiyan, Hubei, China

Abstract

Background: Large language models (LLMs) have shown potential in medical text generation. Senior physician ward round records are critical documents whose quality reflects the accuracy and continuity of clinical decision-making. The initial record is particularly important, as it represents the first formal senior-level synthesis of a patient's presentation, establishing the diagnostic framework and treatment direction for all subsequent care. The quality of LLM-generated initial records for acute poisoning remains unclear.

Objective: Focusing on patients with acute poisoning, this study systematically compared medical record writing quality among DeepSeek, ChatGPT (OpenAI), and human physicians to clarify the clinical value of LLMs.

Methods: A retrospective analysis included 256 cases of acute poisoning from the emergency department ward of Taihe Hospital, Hubei University of Medicine. DeepSeek-V3.2-Exp and GPT-5.1 generated senior physician ward round records from standardized Chinese-language prompts, which were compared with the original medical charts. Blinded evaluations were performed by 3 senior emergency physicians, who scored overall quality across 5 dimensions on a Likert scale (from 1 to 5): case characteristics, current diagnosis, differential diagnosis, treatment plan, and prognosis assessment. Error frequencies were documented under 3 categories (inaccuracies, omissions, and fabrications), and potential harm was assessed using a modified Agency for Healthcare Research and Quality harm scale.

Results: DeepSeek achieved the highest mean total score (24.14, SD 0.90), which was significantly higher than ChatGPT (23.30, SD 1.42; $P<.001$) and the physician group (23.86, SD 0.86; $P=.02$). DeepSeek had the highest score for differential diagnosis (mean 4.98, SD 0.10) and prognosis assessment (mean 4.73, SD 0.42) and was comparable to physicians in case characteristics (DeepSeek: mean 4.90, SD 0.23; physicians: mean 4.96, SD 0.15; $P>.001$). For drug and pesticide poisoning, DeepSeek's mean total scores (24.23, SD 0.75 and 23.92, SD 1.14, respectively) were significantly higher than ChatGPT's (23.34, SD 1.33 and 22.78, SD 1.33, respectively; $P<.001$ for both). In biological toxin poisoning, DeepSeek (mean 23.97, SD 0.96) and physicians (mean 24.26, SD 0.62) scored similarly, both significantly higher than ChatGPT (mean 22.53, SD 1.86; $P<.001$). Overall potential harm scores were low across all 3 groups (<1 point), without significant differences ($P=.38$), although high-harm records were significantly more frequent in both LLM groups than in the physician group ($P=.02$).

Conclusions: LLMs performed satisfactorily in generating initial senior physician ward round records for acute poisoning, with DeepSeek particularly outperforming the physician group in differential diagnosis and prognosis assessment and showing potential to assist clinical documentation. However, the significantly higher proportion of high-harm errors in LLM-generated records underscores the need for mandatory physician review before incorporation into official medical records.

(J Med Internet Res 2026;28:e91222) doi:10.2196/91222

KEYWORDS

large language models; generative artificial intelligence; emergency department; ward round records; clinical documentation; acute poisoning; medical records; DeepSeek; ChatGPT; artificial intelligence; AI

XSL•FO
RenderX

Introduction

Acute poisoning is a common critical condition in emergency departments, characterized by rapid progression and high mortality rates, which places significant demands on clinicians' rapid diagnostic and management capabilities [1]. Senior physician ward rounds serve as a core component of medical quality control, and senior physician ward round records are essential medical documents for organizing clinical information and clarifying diagnostic and therapeutic directions. They also serve as critical evidence for teaching, quality traceability, and medical dispute resolution [2]. Among these documents, the initial senior physician ward round record is of particular significance: it represents the first formal synthesis of a patient's clinical presentation by a senior physician, establishing the diagnostic framework and treatment direction that guides all subsequent care. Its quality is, therefore, directly linked to patient safety and care continuity. However, heavy clinical workloads often prevent physicians from ensuring the standardization and completeness of ward round records, thereby affecting the quality of care [3].

In recent years, large language models (LLMs) such as ChatGPT and DeepSeek have been increasingly applied in the medical field [4,5], demonstrating potential in patient education and clinical support roles [6-8]. Williams et al [9] showed that discharge summaries generated by GPT-4 were comparable in quality to those written by physicians. Gün [10] found high consistency between ChatGPT and emergency physicians in blood gas analysis interpretation. LLMs possess efficient text generation and information integration capabilities and can theoretically generate ward round record frameworks that comply with clinical standards, reducing physicians' documentation workload [11]. However, for senior physician ward round records in emergency department wards—a document type requiring both timeliness and clinical complexity—validation based on real clinical data is lacking, particularly in Chinese-language medical documentation and the specialized field of acute poisoning [12].

DeepSeek, as a representative domestic LLM, features free use and strong Chinese-language comprehension capabilities [13]. This study, based on real clinical records from an emergency department, aimed to systematically compare the quality of initial senior physician ward round records generated by DeepSeek and ChatGPT against the quality of those written by human physicians using real acute poisoning cases across multiple evaluation dimensions. We hypothesized that LLMs would demonstrate comparable or superior performance to that of physicians in structured documentation tasks—particularly in information integration and differential diagnosis—while potentially showing deficiencies in complex clinical judgment domains such as definitive diagnosis formulation.

Methods

Study Design

This was a retrospective cross-sectional study. The LLM-generated records and blinded evaluations were conducted from September 2025 to December 2025 at the emergency department of Taihe Hospital, Hubei University of Medicine; the clinical records analyzed were from patients with acute poisoning admitted between January 2023 and December 2024.

Ethical Considerations

The study protocol was approved by the hospital's ethics committee (ethics approval number: 2025KS187). All medical record data were deidentified prior to analysis.

Study Population

Inclusion criteria were (1) patients with acute poisoning admitted between January 2023 and December 2024; (2) complete medical records, including initial progress notes, attending physician ward round records, and initial senior physician ward round records; (3) hospital stay of 48 hours or longer; and (4) confirmed diagnosis of acute poisoning. Exclusion criteria were (1) incomplete medical records; (2) concurrent severe trauma, advanced cancer, or other critical conditions; and (3) chronic poisoning cases.

A total of 256 patient records were ultimately included and were categorized by poisoning type: 43% (n=110) cases of drug poisoning, 18.8% (n=48) cases of pesticide poisoning, 12.5% (n=32) cases of gas poisoning, 12.1% (n=31) cases of biological toxin poisoning, and 13.7% (n=35) cases of other poisoning (including food poisoning, alcohol poisoning, rodenticide poisoning, and chemical poisoning; Multimedia Appendix 1). The distribution of poisoning types reflects the actual clinical epidemiology of acute poisoning in our emergency department, with drug poisoning being the most prevalent category, consistent with national epidemiological trends [1]. No artificial balancing of the dataset was performed as the primary objective was to evaluate LLM performance under real-world, ecologically valid conditions.

Ward Round Record Generation

The study workflow is shown in Figure 1. Initial progress notes and attending physician ward round records were extracted from the electronic medical record system as input data for each patient. DeepSeek-V3.2-Exp (November 2025 version) and GPT-5.1 (November 2025 version) were used to generate initial senior physician ward round records according to a standardized prompt template. All input data and prompts were in Chinese, consistent with standard clinical practice in our Chinese hospital setting, and all generated records were in Chinese, matching the language of the source medical records. The prompt design referenced the Basic Standards for Medical Record Writing issued by the National Health Commission of China on March 1, 2010. The original senior physician ward round records in the medical charts served as the control group.

XSL•FO
RenderX

Figure 1. Research flowchart. AHRQ: Agency for Healthcare Research and Quality; API: application programming interface; EMR: electronic medical record.

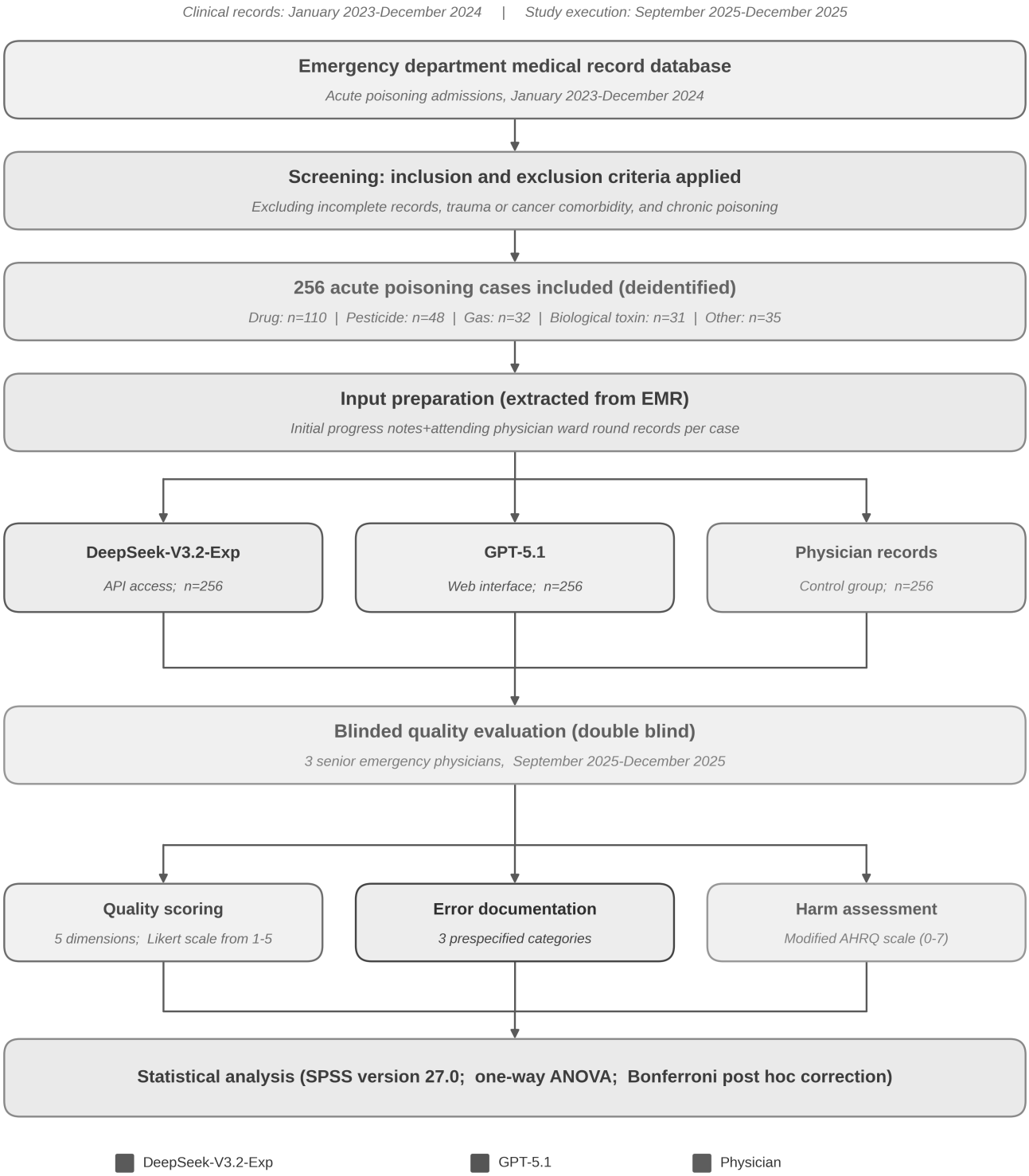

DeepSeek-V3.2-Exp was accessed via an application programming interface (API), with parameters set to a temperature of 0.3, top_p of 0.95, max_tokens of 4096, frequency_penalty of 0, and presence_penalty of 0. GPT-5.1 was accessed through the official web interface with the memory function disabled, with conversation history cleared before each case to ensure session independence. The 2 LLMs were accessed via different methods owing to practical constraints during the study period: DeepSeek offered a stable API end point enabling systematic batch processing with controlled parameters, whereas API access to ChatGPT was not available to our research team, necessitating web interface use. A relatively low temperature value (0.3) was selected for DeepSeek to enhance output determinism and reduce the impact of randomness on documentation quality. All initial senior physician ward round records were generated during the final week of November 2025 to control for potential variation from model version updates. Each case was generated only once without repeated generation or manual selection to ensure objectivity of results.

XSL•FO
RenderX

Evaluation Methods

The evaluation team consisted of 3 senior emergency physicians (attending physician level or above; mean experience 10 years). A double-blind design was used, with evaluators unaware of the record sources. Using the initial progress notes and attending physician ward round records as reference, evaluators scored both LLM-generated and actual senior physician ward round records on 5 dimensions: case characteristics, current diagnosis, differential diagnosis, treatment plan, and prognosis assessment. A self-designed senior physician ward round record scoring form was used as the evaluation tool (Table 1), with each dimension scored on a Likert scale from 1 to 5.

Error frequencies were documented under three prespecified categories: (1) inaccuracies—errors in how existing clinical information was processed or recorded, such as misattributed laboratory values, incorrect dates, or copying and pasting errors; (2) omissions—clinically relevant information present in the input data that was absent from the generated record; and (3) fabrications—content in the generated record that could not be traced to or was directly contradicted by the input clinical data, such as documenting a treatment not administered or a physical finding inconsistent with the patient's actual examination. We use the term "fabrication" in preference to "hallucination" throughout this paper as it more precisely denotes the generation of clinically unsupported content. In physician-written records, fabrication-type errors corresponded to documentation of events or findings that did not occur. This taxonomy was adapted from established LLM error classification frameworks for clinical text [14,15]. Additionally, potential harm was assessed using a modified Agency for Healthcare Research and Quality harm scale [16].

Table . Evaluation dimensions of the senior physician ward round record scoring form[a].

Dimension	Evaluation content
Case characteristics	Summarization of chief concern and present illness, integration of physical signs and auxiliary examinations, and summary of disease progression
Current diagnosis	Accuracy of primary diagnosis, completeness of diagnostic evidence, and rationality of diagnostic reasoning
Differential diagnosis	Comprehensiveness of differential diagnoses, sufficiency of differentiating evidence, and rationality of diagnostic exclusion
Treatment plan	Specificity of treatment regimen, plan for further investigations, and key points of condition monitoring
Prognosis assessment	Basis of prognostic judgment, identification of risk factors, and rationality of outcome prediction

[a]Each dimension was scored on a 5-point Likert scale: 1="strongly disagree," 2="disagree," 3="neutral," 4="agree," and 5="strongly agree."

Statistical Analysis

Statistical analysis was performed using SPSS (version 27.0; IBM Corp). The intraclass correlation coefficient was used to assess interrater reliability, with mean scores used for subsequent analysis. Continuous variables were expressed as means and SDs with 1-way ANOVA for 3-group comparisons and Bonferroni correction for pairwise post hoc comparisons. Categorical variables were expressed as frequencies and percentages, with chi-square or Fisher exact tests for between-group comparisons. A P value of less than .05 was considered statistically significant.

Results

Interrater Reliability

Interrater reliability among the 3 evaluators was good, with an intraclass correlation coefficient for total scores of 0.839 (95% CI 0.819 - 0.858; P<.001). Mean scores for each dimension and total scores were used for subsequent analysis.

Total Score Comparison

Total scores differed significantly among the 3 groups ($F_{2,765}$=38.918; P<.001). Bonferroni post hoc comparisons showed that the DeepSeek group achieved the highest total score, significantly higher than both the ChatGPT group and the physician group; the physician group's total score was also significantly higher than that of the ChatGPT group (Figure 2 and Table 2).

XSL•FO
RenderX

Figure 2. Comparison of total quality scores among GPT-5.1–, DeepSeek-V3.2-Exp–, and physician-written records. The boxes represent IQRs, the horizontal lines indicate medians, the diamonds denote means, and the whiskers extend to 1.5 times the IQR. Individual data points are shown as jittered dots. One-way ANOVA with Bonferroni post hoc correction was conducted. *P<.05; ****P<.0001.

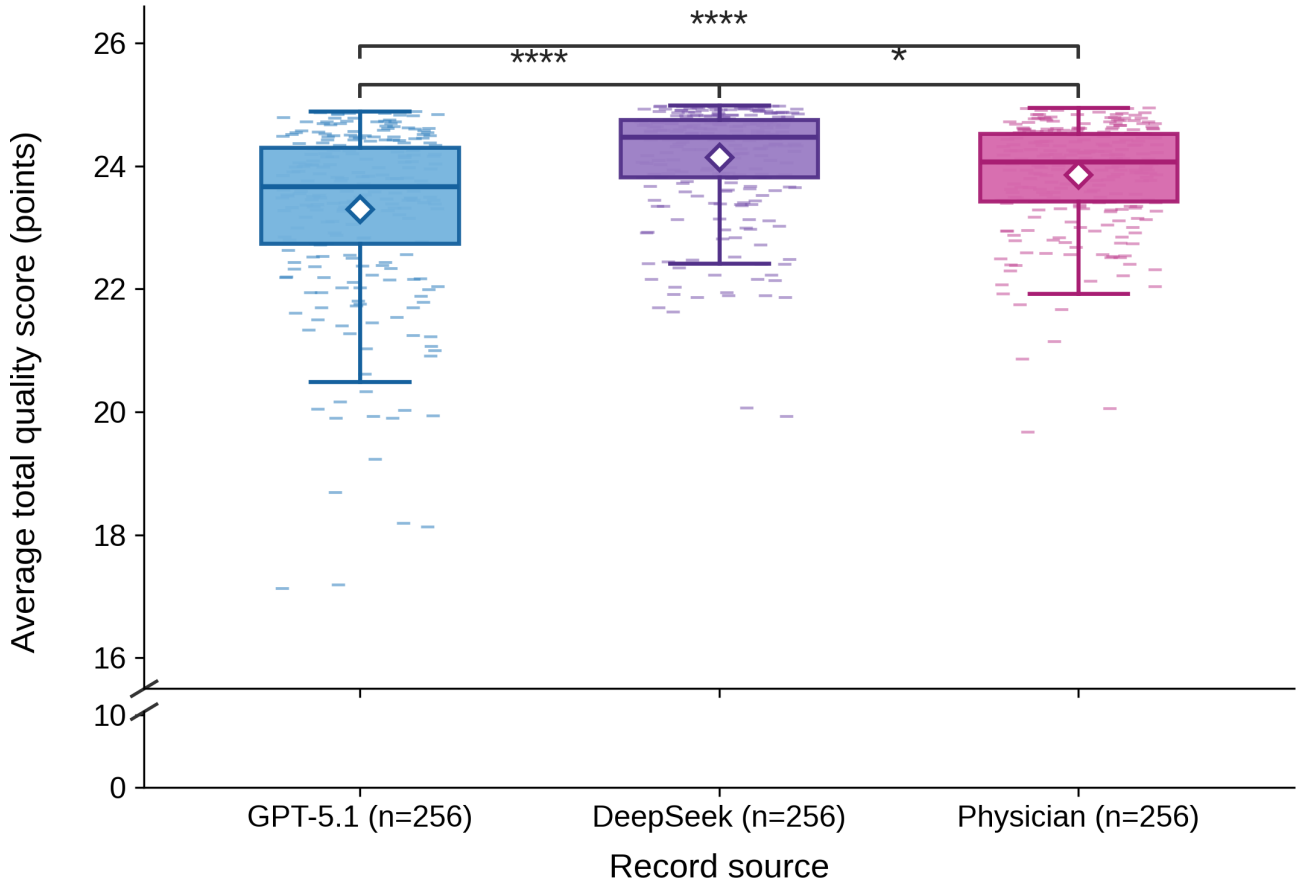

Table . Comparison of scores across 5 dimensions among the 3 groups.

Dimension	GPT-5.1 score (n=256), mean (SD)	DeepSeek-V3.2-Exp score (n=256), mean (SD)	Physician score (n=256), mean (SD)	F test (df)[a]	P value
Case characteristics (range 1-5)	4.81 (0.41)	4.96 (0.23)[b]	4.90 (0.15)	18.27 (2, 765)	<.001
Current diagnosis (range 1-5)	4.43 (0.58)[c]	4.69 (0.43)[d]	4.86 (0.30)	57.94 (2, 765)	<.001
Differential diagnosis (range 1-5)	4.61 (0.73)[c]	4.98 (0.10)[d]	4.75 (0.36)	37.21 (2, 765)	<.001
Treatment plan (range 1-5)	4.81 (0.38)[c]	4.83 (0.37)	4.89 (0.30)	4.10 (2, 765)	.02
Prognosis assessment (range 1-5)	4.65 (0.47)[c]	4.73 (0.42)[c]	4.39 (0.54)	35.67 (2, 765)	<.001
Total score (range 5-25)	23.30 (1.42)[c]	24.14 (0.90)[d]	23.86 (0.86)	38.92 (2, 765)	<.001

[a]F statistic from one-way ANOVA.

[b]P<.05 vs the GPT-5.1 group (Bonferroni correction). By definition, this footnote does not appear in the “GPT-5.1” column.

[c]P<.05 vs the physician group. By definition, this footnote does not appear in the “Physician” column.

[d]Significance vs both groups.

Comparison of Dimension Scores

As shown in Table 2, DeepSeek demonstrated optimal performance in both differential diagnosis (mean 4.98, SD 0.10) and prognosis assessment (mean 4.73, SD 0.42). DeepSeek’s case characteristics score (mean 4.96, SD 0.23) was statistically comparable to that of the physician group (mean 4.90, SD 0.15; P>.05), indicating no significant difference between them. The physician group achieved the highest scores in current diagnosis (mean 4.86, SD 0.30) and treatment planning (mean 4.89, SD

XSL•FO
RenderX

0.30). Notably, the physician group's score in prognosis assessment (mean 4.39, SD 0.54) was significantly lower than that of both LLMs (*P*<.001).

Subgroup Analysis by Poisoning Type

In drug poisoning cases, the DeepSeek group scored highest (mean 24.23, SD 0.75), with statistically significant differences in all pairwise comparisons. In pesticide poisoning cases, there was no significant difference between the DeepSeek and physician groups, but both scores were significantly higher than that of the ChatGPT group. In biological toxin poisoning cases, the physician and DeepSeek groups had similar scores, both significantly higher than those of ChatGPT. In gas poisoning and other poisoning cases, no statistically significant differences were observed among the 3 groups (Table 3).

Table . Subgroup analysis by poisoning type.

Poisoning type	Cases (n=256), n (%)	GPT-5.1 score (range 5-25), mean (SD)	DeepSeek-V3.2-Exp score (range 5-25), mean (SD)	Physicians score (range 5-25), mean (SD)	*F* test (*df*)	*P* value
Drug	110 (43)	23.34 (1.33)[a]	24.23 (0.75)[b]	23.79 (0.88)	21.39 (2, 327)	<.001
Pesticide	48 (18.8)	22.78 (1.33)[a]	23.92 (1.14)[c]	23.77 (1.00)	13.46 (2, 141)	<.001
Gas	32 (12.5)	24.29 (0.95)	24.42 (0.86)	23.95 (0.79)	2.50 (2, 93)	.09
Biological toxin	31 (12.1)	22.53 (1.86)[a]	23.97 (0.96)[c]	24.26 (0.62)	16.71 (2, 90)	<.001
Other	35 (13.7)	23.70 (1.03)	24.01 (0.88)	23.80 (0.73)	1.14 (2, 102)	.33

[a]*P*<.05 vs the physician group. By definition, this footnote does not appear in the "Physician" column.

[b]Significance vs both groups.

[c]*P*<.05 vs the GPT-5.1 group (Bonferroni correction). By definition, this footnote does not appear in the "GPT-5.1" column.

Error Type Analysis and Potential Harm Assessment

Error frequencies and types across the 3 groups are shown in Tables 4 and 5. Physician records had the lowest mean number of errors per record (1.38, SD 1.15), whereas DeepSeek (2.56, SD 1.72) and ChatGPT (2.31, SD 1.58) had significantly more errors (*P*<.001). Both LLMs showed significantly higher rates of omission errors (DeepSeek: mean 1.62, SD 1.25; ChatGPT: mean 1.45, SD 1.12) than the physician group (mean 0.68, SD 0.82; *P*<.001). Inaccuracy errors also differed significantly among groups (*P*=.008), whereas fabrication error rates showed no significant difference (*P*=.28).

Table . Analysis of error types across the 3 groups.

Error type	GPT-5.1, mean (SD)	DeepSeek-V3.2-Exp, mean (SD)	Physicians, mean (SD)	*F* test (*df*)	*P* value
Inaccuracies	0.62 (0.70)	0.68 (0.82)	0.42 (0.61)	4.85 (2, 765)	.008
Omissions	1.45 (1.12)[a]	1.62 (1.25)[a]	0.68 (0.82)	38.92 (2, 765)	<.001
Fabrications	0.24 (0.46)	0.26 (0.48)	0.28 (0.52)	1.28 (2, 765)	.28
Total errors	2.31 (1.58)[a]	2.56 (1.72)[a]	1.38 (1.15)	28.56 (2, 765)	<.001

[a]*P*<.001 vs the physician group (Bonferroni correction; see the Methods section for definitions of error categories).

XSL•FO
RenderX

Table . Examples of typical errors from the 3 groups.

Source and error type		Error description
DeepSeek-V3.2-Exp		
	Inaccuracy	The Poisoning Severity Score was systematically overestimated by 1 point relative to the actual severity in several cases.
	Omission	Omission of vital sign changes and positive findings from critical abdominal examinations.
	Fabrication	A superficial scratch on the arm of a patient with depression was attributed to a dermatological condition not present in the source record.
GPT-5.1		
	Inaccuracy	The correct diagnosis was included within the differential diagnosis list rather than as the primary diagnosis; the chief concern did not conform to standard medical record writing format.
	Omission	Omission of vital sign changes and positive findings from critical abdominal examinations.
	Fabrication	For a patient who was critically ill, first-level nursing care was ordered instead of the required special-level nursing care.
Physicians		
	Inaccuracy	Copying and pasting errors, such as transcribing “pleural effusion” as “pericardial effusion”; incorrect dosage units (mg vs g); and incorrect date entry.
	Omission	Failure to include certain minor findings from imaging studies in the diagnosis.
	Fabrication	Antihypertensive medications were not used during the actual treatment but were recorded as administered.

Overall, potential harm scores (measured on a scale from 0 to 7) were low across all 3 groups (Table 6; physician group: mean 0.45, SD 0.72; DeepSeek: mean 0.82, SD 0.95; ChatGPT: mean 0.71, SD 0.88; no statistically significant difference among groups: F=0.97 and P=.38). Harm scores per individual error also showed no significant difference (P=.56). However, the distribution of high-harm records (≥4 points) differed significantly (DeepSeek: 8/256, 3.1%; ChatGPT: 5/256, 2%; physician group: 2/256, 0.8%; χ^2_2=7.8; P=.02).

Table . Potential harm assessment results across the 3 groups[a].

Indicator	GPT-5.1	DeepSeek-V3.2-Exp	Physicians	*P* value
Overall harm score, mean (SD)	0.71 (0.88)	0.82 (0.95)	0.45 (0.72)	.38
Harm score per error, mean (SD)	1.35 (1.01)	1.38 (1.05)	1.32 (0.98)	.56
High-harm records (≥4 points; n=256), n (%)	5 (2)	8 (3.1)	2 (0.8)	.02

[a]Harm scores were assessed using a modified Agency for Healthcare Research and Quality scale (0 - 7 points): 0=no potential for harm; 1=potential for emotional distress or mild, transient discomfort; 2=potential for requiring additional treatment; 3=potential for temporary harm (bodily or psychological injury, likely not permanent); 4=potential for permanent harm; 5=potential for lifelong injury or disfigurement; 6=potential for severe permanent harm; and 7=potential for death.

Discussion

Summary of Findings

This study was the first to systematically compare initial senior physician ward round records written by DeepSeek, ChatGPT, and physicians for patients with acute poisoning using real clinical progress notes and attending physician records as structured input. By grounding the evaluation in authentic clinical data rather than simulated scenarios, this design may offer greater ecological validity than studies using hypothetical

XSL•FO
RenderX

vignettes, although the single-center retrospective nature limits broader generalizability.

Overall, DeepSeek achieved the highest total score, significantly outperforming both ChatGPT and the physician group. This finding demonstrates that LLMs have acquired a basic capability for assisting in medical documentation generation and that the domestic model DeepSeek performs comparably or superiorly to the mainstream international model ChatGPT—providing empirical support for its application in Chinese clinical settings. With advantages including free access, absence of data export requirements, and strong alignment with Chinese medical documentation conventions, DeepSeek has broad potential in domestic medical documentation assistance [13].

Dimension-Level Performance

The LLMs—particularly DeepSeek—demonstrated outstanding performance in case characteristic summarization and differential diagnosis, reflecting their powerful information extraction and structured knowledge retrieval capabilities [17,18]. In the differential diagnosis dimension, DeepSeek achieved near-perfect scores (mean 4.98, SD 0.10), comprehensively listing diseases requiring differentiation and providing sufficient differentiating evidence. However, in current diagnosis—a domain requiring integrative clinical judgment—the physician group maintained a significant advantage (mean 4.86, SD 0.30). Accurate clinical diagnosis depends not only on standardized medical knowledge but also on the synthesis of individual patient circumstances, recognition of atypical presentations, dynamic tracking of disease evolution, and the weighing of complex comorbidities. This capacity for contextual clinical reasoning, built through accumulated experience, represents a core domain in which current LLMs cannot fully substitute trained clinicians [19]. Notably, GPT-5.1 demonstrated a pattern of placing the correct diagnosis within the differential diagnosis list rather than as the primary diagnosis [20]—a finding consistent with results reported for GPT-4 by McDuff et al [21].

Both LLMs scored significantly higher than the physician group in prognosis assessment. This may reflect the models' tendency to generate more comprehensive assessment frameworks encompassing risk factors and outcome scenarios [22], or it may be attributable to clinicians simplifying prognosis documentation under time pressure in routine practice. Regardless, the clinical accuracy of LLM-generated prognostic content requires prospective validation against actual patient outcomes before it can be relied upon clinically.

Subgroup Performance by Poisoning Type

Subgroup analysis revealed heterogeneity in LLM performance, with advantages concentrated in categories characterized by standardized information and well-established management protocols. In drug poisoning—the most common category—DeepSeek performed most prominently, consistent with the hypothesis that structured, protocol-amenable scenarios favor LLM pattern matching and knowledge extraction.

In pesticide and biological toxin poisoning cases, ChatGPT scored significantly lower than both DeepSeek and the physician group, whereas DeepSeek and physicians performed comparably. Pesticide and biological toxin poisonings (such as wasp stings) are more prevalent in China and involve complex toxicological mechanisms, diverse clinical presentations, and individualized treatment strategies. DeepSeek's maintained performance in these categories may reflect its deeper integration with domestic toxicology databases, local clinical guidelines, and specialized knowledge during training.

In gas poisoning (primarily carbon monoxide poisoning), no significant differences were observed among groups, possibly reflecting the highly standardized treatment principles for this category (eg, hyperbaric oxygen therapy) and the relatively small subgroup sample size. In the heterogeneous "other poisoning" category, the absence of significant differences suggests that current LLMs may have limitations in handling complex, atypical cases. These findings reinforce the recommendation that LLMs function as documentation assistants, with all generated records subject to systematic review by senior physicians before clinical use.

Comparison With Existing Literature

The findings of this study are broadly consistent with those of Schwieger et al [23], confirming that LLMs possess significant advantages in medical documentation standardization and efficiency. Unlike prior studies focused on radiology reports and discharge summaries [24-26], this study specifically addressed emergency department ward round records, extending the evidence on LLM applicability to critical care documentation. Consistent with Hains et al [27], who identified performance gaps when processing real electronic medical record data, our study found suboptimal LLM performance in complex and atypical poisoning categories. The finding that fine-tuned models may outperform general models in specific diagnostic tasks [18] further suggests that domain-specific adaptation could improve clinical performance in specialized fields such as toxicology.

Regarding safety, existing evidence indicates that LLM-drafted discharge summaries have comparable overall quality and safety to those of physician-written records [9]; however, high-harm record rates in this study were higher for LLMs than for physicians, suggesting that ward round records—which encode more complex clinical reasoning than discharge summaries—may require additional scrutiny. This is consistent with broader evidence that LLM limitations are more pronounced in complex clinical decision-making contexts [28-30].

Safety Considerations

Although overall potential harm scores were low across all 3 groups, the significantly higher proportion of high-harm records (≥4 points) in the LLM groups—3.1% (8/256) for DeepSeek and 2% (5/256) for ChatGPT vs 0.8% (2/256) for the physician group (χ^2=7.8; P=.02)—carries important clinical safety implications. High-harm errors included systematic PSS score overestimation leading to overtreatment risk, inappropriate downgrading of nursing care levels for patients who were critically ill, and omission of critical physical examination findings. These may reflect LLM overconfidence in atypical cases, provision of definitive recommendations without

XSL•FO
RenderX

acknowledging diagnostic uncertainty, or fabrication errors at critical decision points such as antidote dosing or nursing level assignment.

These findings underscore the irreplaceable role of physician review in human–artificial intelligence collaboration models. We recommend the following safety mechanisms for clinical implementation: (1) mandatory senior physician review of all LLM-generated ward round records prior to incorporation into official documentation, (2) enhanced review protocols for critically ill patients and complex poisoning types, and (3) establishment of ongoing error monitoring and feedback systems to enable continuous optimization of model application. Notably, although physician-written records had the lowest proportion of high-harm errors, their occurrence (2/256, 0.8%) confirms that documentation quality control is necessary regardless of record authorship, primarily to address copy-paste errors and dosage unit confusion.

Limitations

Several limitations warrant consideration. First, this was a single-center retrospective study; the findings may not generalize to institutions with different patient populations, clinical workflows, or regional epidemiological profiles. Second, the retrospective design precluded evaluation of actual documentation time or work efficiency improvements, which are clinically important considerations given the timeliness demands of emergency ward round documentation. Third, patient outcome data were not linked, so the clinical impact of documentation errors could not be directly assessed. Fourth, sample sizes in some subgroups were relatively small, potentially limiting statistical power. Fifth, the 2 LLMs were accessed via different methods—API for DeepSeek and web interface for ChatGPT owing to API unavailability—which may have introduced minor differences in output consistency. Sixth, the conservative temperature setting (0.3) for DeepSeek may have constrained output diversity; the effects of alternative parameter configurations were not explored. Seventh, the number of high-harm records was small, limiting in-depth analysis of associated risk factors. Future studies should address these limitations through multicenter prospective designs, time efficiency measurements, outcome linkage, and investigation of different model parameter settings. Additionally, the inference that language differences minimally impact LLM output [31] was drawn from medical examination studies using an earlier ChatGPT version, and its applicability to real-world clinical notes with the current model versions should be interpreted cautiously.

Conclusions

LLMs demonstrated satisfactory performance in generating initial senior physician ward round records in the emergency department, with DeepSeek particularly outperforming the physician group in case characteristic integration and differential diagnosis. However, physician-level performance was not achieved in core clinical diagnosis, and the significantly higher proportion of high-harm records in LLM-generated documentation underscores the necessity of rigorous physician review mechanisms in any clinical implementation. These findings position LLMs as documentation assistants rather than independent clinical authors, with the potential to augment physician efficiency without supplanting clinical judgment. Realizing this potential safely will require standardized quality control frameworks, including mandatory senior physician review, heightened oversight for complex or critically ill cases, and continuous monitoring of high-harm errors. Future research should prioritize multicenter prospective studies, domain-specific model fine-tuning for clinical toxicology, integration of time efficiency metrics, and the development of validated quality standards for LLM-generated medical records in emergency settings.

Acknowledgments

The authors would like to express their sincere gratitude to the emergency medicine team for their professionalism and dedication and to the patients and their families for their trust and cooperation throughout the treatment process. The authors declare that generative artificial intelligence tools were not used in any portion of manuscript preparation, including writing, editing, data analysis, or figure generation. The generative artificial intelligence tools (DeepSeek-V3.2-Exp and GPT-5.1 [OpenAI]) were used solely as the study intervention, as described in the Methods section.

Funding

This research was supported by the open research fund of Hubei Provincial Clinical Research Center for Pneumoconiosis and Poisoning (HCRCP2025B04 and HCRCP2025B06), the Graduate Innovation Project of Hubei University of Medicine (YC202542), the Graduate Education and Teaching Research Project of Hubei University of Medicine (YJ2024018), and the Guiding Project of Shiyan Science and Technology Bureau in Hubei Province (24Y044).

Data Availability

The data supporting the findings of this study are available from the corresponding author on reasonable request.

Authors' Contributions

JZ and WP led the development of problem input models, data collection, manuscript preparation, data processing, and statistical analysis. YW, KY, and ZF were responsible for English translation and editing. XY conceptualized the project and supervised the design, revision, and review of all aspects of the study.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Clinical characteristics of the 256 included acute poisoning patients.
[DOCX File, 15 KB - jmir_v28i1e91222_app1.docx]

References

1. Gummin DD, Mowry JB, Beuhler MC, et al. 2020 annual report of the American Association of Poison Control Centers' National Poison Data System (NPDS): 38th annual report. Clin Toxicol (Phila) 2021 Dec;59(12):1282-1501. [doi: 10.1080/15563650.2021.1989785] [Medline: 34890263]
2. Kind AJ, Smith MA. Documentation of mandated discharge summary components in transitions from acute to subacute care. In: Henriksen K, Battles JB, Keyes MA, Grady ML, editors. Advances in Patient Safety: New Directions and Alternative Approaches: Agency for Healthcare Research and Quality; 2008. URL: https://www.ncbi.nlm.nih.gov/books/NBK43715/ [accessed 2026-06-18]
3. Momenipur A, Pennathur PR. Balancing documentation and direct patient care activities: a study of a mature electronic health record system. Int J Ind Ergon 2019 Jul;72:338-346. [doi: 10.1016/j.ergon.2019.06.012] [Medline: 32201437]
4. Wang F, Casalino LP, Khullar D. Deep learning in medicine-promise, progress, and challenges. JAMA Intern Med 2019 Mar 1;179(3):293-294. [doi: 10.1001/jamainternmed.2018.7117] [Medline: 30556825]
5. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
6. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
7. Garcia P, Ma SP, Shah S, et al. Artificial intelligence-generated draft replies to patient inbox messages. JAMA Netw Open 2024 Mar 4;7(3):e243201. [doi: 10.1001/jamanetworkopen.2024.3201] [Medline: 38506805]
8. Pan W, Zhang S, Wang Y, et al. Clinical management of wasp stings using large language models: cross-sectional evaluation study. J Med Internet Res 2025 Jun 4;27:e67489. [doi: 10.2196/67489] [Medline: 40466102]
9. Williams CY, Subramanian CR, Ali SS, et al. Physician- and large language model-generated hospital discharge summaries. JAMA Intern Med 2025 Jul 1;185(7):818-825. [doi: 10.1001/jamainternmed.2025.0821] [Medline: 40323616]
10. Gün M. AI-assisted blood gas interpretation: a comparative study with an emergency physician. Am J Emerg Med 2025 Aug;94:1-2. [doi: 10.1016/j.ajem.2025.04.028] [Medline: 40252296]
11. Hiebel N, Ferret O, Fort K, Névéol A. Clinical text generation: are we there yet? Annu Rev Biomed Data Sci 2025 Aug;8(1):173-198. [doi: 10.1146/annurev-biodatasci-103123-095202] [Medline: 40101215]
12. Clough RA, Sparkes WA, Clough OT, Sykes JT, Steventon AT, King K. Transforming healthcare documentation: harnessing the potential of AI to generate discharge summaries. BJGP Open 2024 Apr;8(1):BJGPO.2023.0116. [doi: 10.3399/BJGPO.2023.0116] [Medline: 37699649]
13. Wu J, Wang Z, Qin Y. Performance of DeepSeek-R1 and ChatGPT-4o on the Chinese National Medical Licensing Examination: a comparative study. J Med Syst 2025 Jun 3;49(1):74. [doi: 10.1007/s10916-025-02213-z] [Medline: 40459679]
14. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med 2024 Sep;30(9):2613-2622. [doi: 10.1038/s41591-024-03097-1] [Medline: 38965432]
15. Chen F, Li Y, Chen Y, et al. Strategies for the analysis and elimination of hallucinations in artificial intelligence generated medical knowledge. J Evid Based Med 2025 Sep;18(3):e70075. [doi: 10.1111/jebm.70075] [Medline: 40983876]
16. Williams T, Szekendi M, Pavkovic S, Clevenger W, Cerese J. The reliability of AHRQ Common Format Harm Scales in rating patient safety events. J Patient Saf 2015 Mar;11(1):52-59. [doi: 10.1097/PTS.0b013e3182948ef9] [Medline: 24080718]
17. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024 Apr;30(4):1134-1142. [doi: 10.1038/s41591-024-02855-5] [Medline: 38413730]
18. Yu P, Xu H, Hu X, Deng C. Leveraging generative AI and large language models: a comprehensive roadmap for healthcare integration. Healthcare (Basel) 2023 Oct 20;11(20):2776. [doi: 10.3390/healthcare11202776] [Medline: 37893850]
19. Cai Q, Yang L, Xiao J, Ma J, Liu M, Pan X. Train-time and test-time computation in large language models for error detection and correction in electronic medical records: a retrospective study. Diagnostics (Basel) 2025 Jul 21;15(14):1829. [doi: 10.3390/diagnostics15141829] [Medline: 40722578]
20. Reese JT, Danis D, Caufield JH, et al. On the limitations of large language models in clinical diagnosis. medRxiv. 2024 Feb 26 p. 2023.07.13.23292613. [doi: 10.1101/2023.07.13.23292613] [Medline: 37503093]
21. McDuff D, Schaekermann M, Tu T, et al. Towards accurate differential diagnosis with large language models. Nature 2025 Jun;642(8067):451-457. [doi: 10.1038/s41586-025-08869-4] [Medline: 40205049]

XSL•FO
RenderX

22. Chung P, Fong CT, Walters AM, Aghaeepour N, Yetisgen M, O'Reilly-Shah VN. Large language model capabilities in perioperative risk prediction and prognostication. JAMA Surg 2024 Aug 1;159(8):928-937. [doi: 10.1001/jamasurg.2024.1621] [Medline: 38837145]
23. Schwieger A, Angst K, de Bardeci M, et al. Large language models can support generation of standardized discharge summaries - a retrospective study utilizing ChatGPT-4 and electronic health records. Int J Med Inform 2024 Dec;192:105654. [doi: 10.1016/j.ijmedinf.2024.105654] [Medline: 39437512]
24. Li Y, Li F, Hong N, et al. A comparative study of recent large language models on generating hospital discharge summaries for lung cancer patients. J Biomed Inform 2025 Aug;168:104867. [doi: 10.1016/j.jbi.2025.104867] [Medline: 40544901]
25. Janota B, Janota K. Application of artificial intelligence (AI) in the creation of discharge summaries in psychiatric clinics. Int J Psychiatry Med 2025 May;60(3):330-337. [doi: 10.1177/00912174241284730] [Medline: 39285727]
26. Zaretsky J, Kim JM, Baskharoun S, et al. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. JAMA Netw Open 2024 Mar 4;7(3):e240357. [doi: 10.1001/jamanetworkopen.2024.0357] [Medline: 38466307]
27. Hains L, Kleinig O, Murugappa A, et al. Large language model discharge summary preparation using real-world electronic medical record data shows promise. Intern Med J 2025 May 28;55(7):1188-1192. [doi: 10.1111/imj.70073] [Medline: 40434141]
28. Huang TY, Hsieh PH, Chang YC. Performance comparison of junior residents and ChatGPT in the Objective Structured Clinical Examination (OSCE) for medical history taking and documentation of medical records: development and usability study. JMIR Med Educ 2024 Nov 21;10:e59902. [doi: 10.2196/59902] [Medline: 39622713]
29. Adler-Milstein J, Redelmeier DA, Wachter RM. The limits of clinician vigilance as an AI safety bulwark. JAMA 2024 Apr 9;331(14):1173-1174. [doi: 10.1001/jama.2024.3620] [Medline: 38483397]
30. Chenais G, Lagarde E, Gil-Jardiné C. Artificial intelligence in emergency medicine: viewpoint of current applications and foreseeable opportunities and challenges. J Med Internet Res 2023 May 23;25:e40031. [doi: 10.2196/40031] [Medline: 36972306]
31. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. Int J Med Inform 2023 Sep;177:105173. [doi: 10.1016/j.ijmedinf.2023.105173] [Medline: 37549499]

Abbreviations

API: application programming interface
LLM: large language model

Edited by I Steenstra; submitted 11.Jan.2026; peer-reviewed by G Qian, J Zaghir; revised version received 07.Jun.2026; accepted 08.Jun.2026; published 30.Jun.2026.

Please cite as:
Zhu J, Pan W, Wang Y, Yan K, Fang Z, Yang X
Quality Evaluation of Large Language Model–Assisted Generation of Initial Senior Physician Ward Round Records for Patients With Acute Poisoning: Cross-Sectional Study
J Med Internet Res 2026;28:e91222
URL: https://www.jmir.org/2026/1/e91222
doi:10.2196/91222

© Junping Zhu, Wei Pan, Yonghong Wang, Kui Yan, Zhicheng Fang, Xianyi Yang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Hepatic Steatosis Severity Prediction in Nonobese Individuals: Machine Learning Model Development and Validation

Yitong Zhu[1,2,3*], PhD; Yongshuai Wang[1,2,3*], PhD; Shenyu Zhang[1,2,3], PhD; Jian Yang[4], PhD; Feng Zhang[1,2,3], PhD; Yan Liu[4], PhD; Jun Shang[5], MSc; Yongliang Zhang[4*], PhD; Jizhou Wang[1,2,3*], PhD; Lianxin Liu[1,2,3*], PhD

[1]Department of Hepatobiliary Surgery, Centre for Leading Medicine and Advanced Technologies of IHM, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, No.96 Jinzhai Road, Hefei, Anhui, China

[2]Anhui Province Key Laboratory of Hepatopancreatobiliary Surgery, Hefei, Anhui, China

[3]Anhui Provincial Clinical Research Center for Hepatobiliary Diseases, Hefei, Anhui, China

[4]Department of Health Management Center, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, China

[5]Department of Nephrology, The Second Hospital of Anhui Medical University, Hefei, Anhui, China

[*]these authors contributed equally

Corresponding Author:
Lianxin Liu, PhD
Department of Hepatobiliary Surgery, Centre for Leading Medicine and Advanced Technologies of IHM, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, No.96 Jinzhai Road, Hefei, Anhui, China

Abstract

Background: Steatotic liver disease affects 40% of nonobese individuals, but existing screening tools inadequately detect and stage disease severity in this population because of the limited sensitivity of conventional ultrasound and the lack of dedicated prediction models.

Objective: This study aimed to develop and validate an interpretable machine learning model specifically for multiclass hepatic steatosis severity prediction in nonobese individuals to support early risk stratification in this underrecognized group.

Methods: Health examination data from 215,145 nonobese participants (BMI <28 kg/m²) were randomly divided into training (n=150,601, 70%) and test (n=64,544, 30%) sets. Hepatic steatosis was diagnosed and graded using the controlled attenuation parameter with established thresholds (none: <248 dB/m; mild: 248 - 268 dB/m; and moderate to severe: >268 dB/m). From 42 candidate variables, 14 predictors were selected using Least Absolute Shrinkage and Selection Operator regression and Recursive Feature Elimination based on Random Forest importance. Six machine learning algorithms—k-nearest neighbors, naive Bayes, multilayer perceptron, random forest, support vector machine, and Extreme Gradient Boosting (XGBoost)—were developed using 10-fold cross-validation, with hyperparameters optimized for maximal area under the receiver operating characteristic curve (ROC-AUC). Model interpretability was assessed using Shapley Additive Explanations analysis. External validation was conducted in non-Hispanic Asian participants from the National Health and Nutrition Examination Survey (n=726). Model performance was evaluated using accuracy, Cohen κ, ROC-AUC, area under the precision-recall curve, F_1-score, precision, sensitivity, and specificity.

Results: The final cohort included 215,145 participants, with steatosis severity classified as none (n=92,944, 43.2%), mild (n=54,121, 25.2%), and moderate to severe (n=68,080, 31.6%). Among the 6 machine learning models, XGBoost achieved the best discrimination on the test set, with an accuracy of 0.824 and a macro-average ROC-AUC of 0.941. In external validation, the model maintained strong performance (macro-average ROC-AUC=0.874). Shapley Additive Explanations analysis identified BMI, waist circumference, liver enzymes (alanine aminotransferase and aspartate aminotransferase), renal function indicators (uric acid and serum creatinine), and metabolic indices (triglycerides, continuous metabolic syndrome score, and triglyceride-glucose index) as key contributors to model predictions. The model has been implemented as an online prediction platform to facilitate clinical use.

Conclusions: This interpretable XGBoost model accurately predicts controlled attenuation parameter–defined hepatic steatosis severity in nonobese individuals and demonstrates robust performance in both internal and external validation cohorts, providing a practical tool for early risk stratification in this underrecognized population.

(J Med Internet Res 2026;28:e82529) doi:10.2196/82529

XSL•FO
RenderX

KEYWORDS

steatotic liver disease; hepatic steatosis; nonobese population; machine learning; controlled attenuation parameter; risk stratification

Introduction

Steatotic liver disease (SLD) is a heterogeneous condition influenced by genetic susceptibility, epigenetic factors, diet, and lifestyle [1]. It has become a leading cause of chronic liver disease globally, with a prevalence of 32.4% that has risen to 37.8% since 2016 [1,2]. Traditional screening strategies have primarily targeted obesity-associated metabolic risks [3]. Recent epidemiological data, however, indicate that up to 40% of SLD cases occur in nonobese individuals, of whom 20% are classified as lean [4], challenging conventional diagnostic paradigms. Despite the absence of overt obesity, these individuals face risks of liver fibrosis, cirrhosis, and cardiometabolic complications comparable to their obese counterparts, with evidence suggesting potentially accelerated disease progression [5-8]. Frequent diagnostic delays, driven by lower clinical suspicion in nonobese individuals, exacerbate this issue. The dissociation between BMI and metabolic dysregulation in this group highlights the urgent need for screening strategies tailored to their distinct metabolic phenotype.

Although liver biopsy remains the diagnostic gold standard, its invasiveness precludes widespread screening. Conventional ultrasound, the primary screening modality, has well-documented limitations, including operator dependence and low sensitivity for detecting mild steatosis (5% - 33% fat content); its accuracy for grading steatosis is only 51.7% [9-11]. The controlled attenuation parameter (CAP), measured via transient elastography using FibroScan (Echosens), provides a more reliable quantitative alternative, enabling standardized classification into 3 severity grades [12-14]. CAP correlates with histological steatosis and metabolic dysfunction severity, offering stratification that guides management—from lifestyle modification for mild disease to combined pharmacological approaches for moderate-to-severe cases [1,15-17]. However, the cost and technical requirements of CAP limit its broad adoption in primary care. Consequently, a diagnostic gap persists, allowing many early-stage nonobese SLD cases to go undetected. This gap underscores the need for nonimaging, cost-effective tools capable of multiclass severity discrimination, particularly for mild steatosis, to facilitate early intervention in metabolically at-risk nonobese populations.

Machine learning (ML) has emerged as a valuable tool for predicting fatty liver disease (FLD) in large-scale studies. For instance, Chen et al [18] and Deng et al [19] used ML models, including Extreme Gradient Boosting (XGBoost), to predict FLD and metabolic dysfunction–associated SLD (MASLD) with high accuracy, achieving areas under the receiver operating characteristic curve (ROC-AUCs) of 0.882 and 0.86, respectively. Other studies by Weng et al [20] and Huang et al [21] further demonstrated that ML models can outperform traditional indices such as the Fatty Liver Index (FLI) in FLD detection. A notable advance by Su et al [22] developed a neural network model specifically for lean individuals (BMI <23 kg/m²), reporting an ROC-AUC of 0.885. While representing progress in population-specific modeling, this approach was limited to binary classification and relied on ultrasound—a modality with recognized sensitivity limitations, especially for mild steatosis. These limitations highlight the need for models capable of multiclass severity stratification in nonobese populations based on a more sensitive quantitative standard.

To address these gaps, this study aimed to develop and validate an optimal multiclass ML model for predicting hepatic steatosis severity in nonobese populations, using large-scale health examination data from Eastern China. Six ML models were constructed and compared. The proposed approach offers several key advancements: (1) the use of CAP for accurate steatosis grading, overcoming the sensitivity limitations of ultrasound; (2) a specific focus on the underrecognized nonobese population; (3) multiclass severity prediction (none, mild, and moderate to severe) for precise risk stratification; and (4) deployment as an open-access online platform offering both a full and a simplified model to accommodate varied resource settings. By enabling multiclass risk stratification in nonobese individuals, this work provides a practical tool to support early identification and may facilitate a shift from passive imaging-based diagnosis toward more proactive metabolic risk assessment.

Methods

Study Population

Retrospective health examination data from 269,240 individuals (2018 - 2024) were obtained from the Health Management Center database of the First Affiliated Hospital of University of Science and Technology of China. After applying inclusion criteria (completion of transient elastography, BMI <28 kg/m²) and exclusion criteria (missing BMI, age <18 or >100 y, pregnancy, liver cirrhosis, liver tumors, or other malignancies), 215,145 participants were included. The BMI cutoff of <28 kg/m² aligns with the Chinese adult obesity threshold (BMI ≥28 kg/m²) defined by the Working Group on Obesity in China [23]. This cutoff is also consistent with the Asian-specific obesity criterion (27.5 kg/m²) recommended in international guidelines [24], reflecting the distinct anthropometric and metabolic profile of Asian populations.

For the external validation cohort extracted from the National Health and Nutrition Examination Survey (NHANES) 2017 to 2020 cycle, only non-Hispanic Asian participants with complete data for all variables required by the final model (including CAP, anthropometric measures, laboratory tests, and derived composite indices) were retained. Participants with any missing values in these key variables were excluded, resulting in a final sample of 726 individuals with fully observed data.

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the First Affiliated Hospital of University of Science and Technology of China (approval 2024-RE-312). The requirement for obtaining individual informed consent was waived by the approving ethics committee. All personal identifiers were

XSL•FO
RenderX

removed prior to analysis to protect participant privacy and confidentiality. No compensation was provided for this secondary use of data. The manuscript and supplementary materials contain no personally identifiable information of any participant.

Grading Diagnosis of Hepatic Steatosis

Hepatic steatosis was diagnosed and graded using CAP measured via transient elastography. CAP quantifies ultrasound wave attenuation, correlating with liver fat content. Severity was graded using established thresholds: none (CAP <248 dB/m), mild (248 - 268 dB/m), and moderate to severe (>268 dB/m) [12]. The reliability of CAP is well established; validation against liver biopsy in multicenter studies has shown excellent ROC-AUC (0.87 for ≥S1 steatosis) [13]. This quantitative approach ensures objective and reproducible assessment, overcoming limitations of conventional ultrasonography.

Variable Processing and Selection

Initially, 42 potential predictors were collected from questionnaires and routine examinations. Variables with >30% missing data were excluded. The missing data pattern in the final cohort (n=215,145) is shown in Multimedia Appendix 1. After splitting the data into training (n=150,601, 70%) and test (n=64,544, 30%) sets, missing values were handled using multiple imputation with the random forest (RF) algorithm (mice package, method="rf"). To account for imputation uncertainty, we generated 5 independent imputed datasets from the training set only. The imputation models learned from the training set were then applied once to the test set—this reflects real deployment, where new data are observed only once, while multiple imputation at the training stage and the subsequent model ensemble already capture missing data uncertainty. Continuous features were normalized using z-score standardization based on training set means and SDs, with the same parameters applied to the test set. Categorical features were label encoded to avoid high dimensionality from one-hot encoding.

Additionally, the following composite indicators were calculated: the aspartate aminotransferase (AST)-to-alanine aminotransferase ratio (ALT); mean arterial pressure as (2×diastolic blood pressure+systolic blood pressure [SBP])/3; the continuous metabolic syndrome score (CMetS), derived from weighted calculations of waist circumference (WC), high-density lipoprotein (HDL) cholesterol, mean arterial pressure, triglycerides, and fasting blood glucose (FBG) [25]; and the triglyceride-glucose index (TyG) as Ln[fasting triglycerides (mg/dL)×fasting glucose (mg/dL)/2]. As the original laboratory measurements of FBG and triglycerides in our cohort were recorded in mmol/L, these values were converted to mg/dL prior to TyG calculation by multiplying by the conversion factors: 18 for glucose and 88.57 for triglycerides.

Feature Selection

To identify key predictors robust to imputation uncertainty, feature selection was performed within each of the 5 imputed training datasets. In each dataset, 2 complementary methods were applied in parallel: Least Absolute Shrinkage and Selection Operator (LASSO) regression with 10-fold cross-validation and Recursive Feature Elimination based on Random Forest importance (RFE-RF). For a given dataset, the intersection of variables selected by LASSO and RFE-RF was taken as its candidate set. Across the 5 imputed datasets, variables that appeared in at least three candidate sets were retained as the final predictors.

Model Development and Explanation

After multiple imputation, 5 complete training datasets were generated. For each ML algorithm (XGBoost, multilayer perceptron [MLP], k-nearest neighbors [KNN], RF, support vector machine [SVM], and naive Bayes [NB]), we trained 5 separate models—one on each imputed training set. Each of these 5 models then produced predicted probabilities for all classes on the test set. To account for uncertainty arising from missing data imputation, we pooled these predictions by averaging the 5 probability vectors for each test sample, yielding a single pooled probability per class. The final predicted class was assigned as the class with the highest average probability. All performance metrics (accuracy, Cohen κ, macro-average ROC-AUC, macro-average PR-AUC, as well as class-wise precision, recall, specificity, and F_1-score) were derived from these pooled predictions. To enhance interpretability, Shapley Additive Explanations (SHAP) analysis was applied to the best-performing XGBoost model. Grounded in game theory, SHAP values quantify the magnitude and direction of each feature's contribution to individual predictions, providing a mathematically consistent framework for model explanation.

Statistical Analysis

All statistical analyses were performed using R software and Python. The Kolmogorov-Smirnov test assessed normality for continuous variables. For nonnormally distributed data, group comparisons used the Kruskal-Wallis H test (multiple groups) or Mann-Whitney *U* test (2 groups). Categorical variables were summarized as frequencies and percentages, with differences assessed by the Pearson chi-square test. Two-sided *P* values <.05 were considered statistically significant.

Results

Basic Characteristics

The final cohort comprised 215,145 participants, stratified by hepatic steatosis grade: none (n=92,944, 43.2%), mild (n=54,121, 25.2%), and moderate to severe (n=68,080, 31.6%). The cohort had a median age of 44 (IQR 34 - 54) years, a balanced sex distribution (men: 51.07%; Table 1). The distributions of the 6 most influential predictors across grades are shown in Multimedia Appendix 2, displaying clear, clinically expected trends. The training and test sets exhibited balanced baseline characteristics (Multimedia Appendix 3). Density plots confirmed that the distributions of key variables remained consistent before and after multiple imputation (Multimedia Appendix 4). Detailed information about the research design is available in Figure 1.

XSL•FO
RenderX

Table . Baseline characteristics of participants stratified by hepatic steatosis grade.

Variables	Total (N=215,145)	None (n=92,944)	Mild (n=54,121)	Moderate to severe (n=68,080)	*P*[a] value
Age (y), median (IQR)	44 (34 to 55)	42 (32 to 53)	46 (35 to 55)	46 (36 to 55)	<.001
Gender, n (%)					<.001
Female	1,03,979 (48.33)	48,089 (51.74)	25,594 (47.29)	30,296 (44.50)	
Male	1,09,875 (51.07)	44,269 (47.63)	28,213 (52.13)	37,393 (54.9)	
Miss	1291 (0.60)	586 (0.63)	314 (0.58)	391 (0.58)	
CAP[b] (dB/m), median (IQR)	251 (227 to 275)	224 (212 to 233)	257 (250 to 262)	283 (276 to 291)	<.001
LSM[c] (kPa), median (IQR)	5.80 (4.90 to 6.70)	5.50 (4.70 to 6.40)	5.80 (5 to 6.70)	6.20 (5.30 to 7.10)	<.001
Miss, n (%)	27,582 (12.82)	12,403 (13.34)	6897 (12.74)	8282 (12.17)	
WC[d] (cm), median (IQR)	88 (83 to 92)	83 (79 to 87)	89 (86 to 91)	92 (90 to 95)	<.001
BMI (kg/m^2), median (IQR)	24.30 (22.47 to 25.90)	22.47 (20.93 to 23.90)	24.50 (23.40 to 25.70)	26.10 (25.10 to 27)	<.001
SBP[e] (mm Hg), median (IQR)	124 (113 to 136)	120 (110 to 131)	125 (115 to 137)	128 (118 to 140)	<.001
Miss, n (%)	49,419 (22.97)	24,317 (26.16)	13,488 (24.92)	11,614 (17.06)	
DBP[f] (mm Hg), median (IQR)	77 (70 to 85)	74 (67 to 82)	78 (71 to 85)	81 (73 to 88)	<.001
Miss, n (%)	6648 (3.09)	3592 (3.86)	1914 (3.54)	1142 (1.68)	
Lym[g] (%), median (IQR)	33.60 (28.70 to 38.70)	33.70 (28.70 to 38.90)	33.30 (28.60 to 38.50)	33.50 (28.80 to 38.40)	<.001
Miss, n (%)	59,380 (27.6)	26,893 (28.93)	15,327 (28.32)	17,160 (25.21)	
PLT[h] (×10 /L)	224 (191 to 261)	221 (188 to 258.75)	223 (191 to 260)	229 (195 to 266)	<.001
Miss, n (%)	61,962 (28.8)	27,789 (29.9)	15,876 (29.33)	18,297 (26.88)	
Mono[i] (%), median (IQR)	6.50 (5.50 to 7.70)	6.50 (5.50 to 7.60)	6.60 (5.60 to 7.70)	6.60 (5.60 to 7.70)	<.001
Miss, n (%)	61,295 (28.49)	27,543 (29.63)	15,832 (29.25)	17,920 (26.32)	
Neu[j] (%), median (IQR)	56.60 (51.20 to 61.80)	56.50 (51 to 62)	56.70 (51.50 to 61.80)	56.50 (51.40 to 61.60)	.08
Miss, n (%)	59,272 (27.55)	26,987 (29.04)	15,234 (28.15)	17,051 (25.05)	
WBC[k] (×10 /L), median (IQR)	5.85 (5 to 6.84)	5.56 (4.74 to 6.52)	5.89 (5.08 to 6.88)	6.19 (5.35 to 7.17)	<.001
Miss, n (%)	54,797 (25.47)	26,714 (28.74)	15,189 (28.06)	12,894 (18.94)	
Hb[l] (g/L), median (IQR)	148 (136 to 156)	142 (131 to 153)	149 (140 to 157)	153 (145 to 160)	<.001
Miss, n (%)	63,339 (29.44)	27,853 (29.97)	16,207 (29.95)	19,279 (28.32)	
FBG[m] (mmol/L), median (IQR)	5.08 (4.76 to 5.47)	4.96 (4.68 to 5.29)	5.11 (4.80 to 5.50)	5.23 (4.89 to 5.72)	<.001
Miss, n (%)	10,004 (4.65)	5396 (5.81)	3012 (5.57)	1596 (2.34)	
Triglycerides (mmol/L), median (IQR)	1.32 (0.93 to 1.92)	1.05 (0.78 to 1.45)	1.40 (1.02 to 1.96)	1.76 (1.26 to 2.53)	<.001
Miss, n (%)	3249 (1.51)	1807 (1.94)	893 (1.65)	549 (0.81)	

XSL•FO
RenderX

Variables	Total (N=215,145)	None (n=92,944)	Mild (n=54,121)	Moderate to severe (n=68,080)	P[a] value
TC[n] (mmol/L), median (IQR)	4.78 (4.22 to 5.39)	4.68 (4.14 to 5.29)	4.83 (4.27 to 5.44)	4.88 (4.33 to 5.49)	<.001
Miss, n (%)	39,479 (18.35)	19,428 (20.9)	10,973 (20.27)	9078 (13.33)	
HDL[o] cholesterol (mmol/L), median (IQR)	1.13 (0.96 to 1.35)	1.24 (1.05 to 1.47)	1.10 (0.94 to 1.30)	1.02 (0.88 to 1.18)	<.001
Miss, n (%)	9488 (4.41)	5114 (5.5)	2786 (5.15)	1588 (2.33)	
LDL[p] cholesterol (mmol/L), median (IQR)	2.84 (2.35 to 3.36)	2.71 (2.25 to 3.21)	2.91 (2.43 to 3.42)	2.97 (2.47 to 3.49)	<.001
Miss, n (%)	1441 (0.67)	794 (0.85)	406 (0.75)	241 (0.35)	
VLDL[q] cholesterol (mmol/L), median (IQR)	0.72 (0.54 to 0.95)	0.67 (0.51 to 0.87)	0.73 (0.55 to 0.97)	0.79 (0.58 to 1.07)	<.001
Miss, n (%)	51,592 (23.98)	25,213 (27.13)	14,286 (26.4)	12,093 (17.76)	
HbA_{1c}[r] (%), median (IQR)	5.50 (5.30 to 5.80)	5.50 (5.30 to 5.70)	5.60 (5.40 to 5.80)	5.60 (5.40 to 6)	<.001
Miss, n (%)	58,197 (27.05)	26,345 (28.35)	15,234 (28.15)	16,618 (24.41)	
ALT[s] (U/L), median (IQR)	21 (15.50 to 31)	18 (13 to 24)	22 (16 to 30)	28 (20 to 41)	<.001
Miss, n (%)	559 (0.26)	308 (0.33)	164 (0.3)	87 (0.13)	
AST[t] (U/L), median (IQR)	22 (19 to 26.70)	21 (18 to 25)	22 (19 to 26)	24 (20 to 29)	<.001
Miss, n (%)	387 (0.18)	219 (0.24)	108 (0.2)	60 (0.09)	
GGT[u] (U/L), median (IQR)	22.80 (15 to 36.20)	17.70 (12.90 to 26.10)	24 (16.70 to 37)	32 (21.90 to 51)	<.001
Miss, n (%)	60,972 (28.34)	27,489 (29.58)	15,930 (29.43)	17,553 (25.78)	
ALP[v] (IU/L), median (IQR)	72 (61 to 86)	69 (58 to 83)	73 (62 to 87)	75 (64 to 88)	<.001
Miss, n (%)	5658 (2.63)	3089 (3.32)	1713 (3.17)	856 (1.26)	
BUN[w] (mmol/L), median (IQR)	5.26 (4.49 to 6.14)	5.16 (4.38 to 6.07)	5.31 (4.56 to 6.19)	5.35 (4.60 to 6.21)	<.001
Miss, n (%)	57,766 (26.85)	26,123 (28.11)	15,098 (27.9)	16,545 (24.3)	
SCr[x] (μmol/L), median (IQR)	69 (60 to 78)	67 (56 to 76)	71 (62 to 79)	71 (63 to 79)	<.001
Miss, n (%)	34,509 (16.04)	17,023 (18.32)	9576 (17.69)	7910 (11.62)	
UA[y] (μmol/L), median (IQR)	360.70 (303 to 419)	329 (275 to 387)	368.70 (315.48 to 422)	395 (342 to 451)	<.001
Miss, n (%)	51,656 (24.01)	25,196 (27.11)	14,312 (26.44)	12,148 (17.84)	
AFP[z] (ng/mL), median (IQR)	2.80 (2 to 3.96)	2.67 (1.93 to 3.83)	2.88 (2.04 to 4.07)	2.91 (2.08 to 4.06)	<.001
Miss, n (%)	49,032 (22.79)	23,987 (25.81)	13,612 (25.15)	11,433 (16.79)	
CEA[aa] (ng/mL), median (IQR)	1.69 (1.18 to 2.43)	1.64 (1.13 to 2.39)	1.72 (1.21 to 2.46)	1.74 (1.23 to 2.47)	<.001
Miss, n (%)	30,271 (14.07)	14,986 (16.12)	8414 (15.55)	6871 (10.09)	

XSL•FO
RenderX

Variables	Total (N=215,145)	None (n=92,944)	Mild (n=54,121)	Moderate to severe (n=68,080)	P^{a} value
ALB[ab] (g/L), median (IQR)	46 (44.30 to 47.70)	45.70 (44 to 47.50)	46 (44.30 to 47.60)	46.40 (44.80 to 48.10)	<.001
Miss, n (%)	61,574 (28.62)	27,589 (29.68)	15,966 (29.5)	18,019 (26.47)	
Tbli[ac] (μmol/L), median (IQR)	14.30 (11.30 to 18.20)	14.20 (11.20 to 18.10)	14.30 (11.40 to 18.20)	14.50 (11.50 to 18.40)	<.001
Miss, n (%)	61,510 (28.59)	27,498 (29.59)	15,981 (29.53)	18,031 (26.49)	
Dbli[ad] (μmol/L), median (IQR)	4.10 (2.90 to 5.50)	4 (2.80 to 5.50)	4.10 (2.90 to 5.40)	4.20 (3.10 to 5.50)	<.001
Miss, n (%)	61,144 (28.42)	27,345 (29.42)	15,876 (29.33)	17,923 (26.33)	
Ibli[ae] (μmol/L), median (IQR)	10.30 (8 to 13.30)	10.20 (8 to 13.20)	10.30 (8.10 to 13.40)	10.30 (8.00 to 13.30)	.01
Miss, n (%)	54,840 (25.49)	25,432 (27.36)	14,123 (26.1)	15,285 (22.45)	
HBP[af], n (%)					<.001
No	1,20,761 (56.13)	61,743 (66.43)	27,548 (50.90)	31,470 (46.23)	
Yes	78,808 (36.63)	24,426 (26.28)	20,956 (38.72)	33,426 (49.10)	
Miss	15,576 (7.24)	6775 (7.29)	5617 (10.38)	3184 (4.67)	
TyG[ag], median (IQR)	8.68 (8.29 to 9.10)	8.35 (8.04 to 8.69)	8.68 (8.34 to 9.04)	8.97 (8.61 to 9.37)	<.001
Miss, n (%)	11,682 (5.43)	5595 (6.02)	3058 (5.65)	3029 (4.45)	
CMetS[ah], median (IQR)	0.33 (−0.10 to 0.78)	−0.12 (−0.44 to 0.23)	0.29 (−0.03 to 0.64)	0.71 (0.36 to 1.12)	<.001

XSL•FO
RenderX

Variables	Total (N=215,145)	None (n=92,944)	Mild (n=54,121)	Moderate to severe (n=68,080)	*P*[a] value
Miss, n (%)	50,925 (23.67)	24,686 (26.56)	13,579 (25.09)	12,660 (18.60)	

[a]*P* values for continuous variables were calculated using the Kruskal-Wallis H test, and those for categorical variables were calculated using Pearson chi-square test.

[b]CAP: controlled attenuation parameter.

[c]LSM: liver stiffness measurement.

[d]WC: waist circumference.

[e]SBP: systolic blood pressure.

[f]DBP: diastolic blood pressure.

[g]Lym: lymphocyte percentage.

[h]PLT: platelet count.

[i]Mono: monocyte percentage.

[j]Neu: neutrophil percentage.

[k]WBC: white blood cell count.

[l]Hb: hemoglobin.

[m]FBG: fasting blood glucose.

[n]TC: total cholesterol.

[o]HDL: high-density lipoprotein.

[p]LDL: low-density lipoprotein.

[q]VLDL: very-low-density lipoprotein.

[r]HbA_{1c}: hemoglobin A_{1c}.

[s]ALT: alanine aminotransferase.

[t]AST: aspartate aminotransferase.

[u]GGT: gamma-glutamyl transferase.

[v]ALP: alkaline phosphatase.

[w]BUN: blood urea nitrogen.

[x]SCr: serum creatinine.

[y]UA: uric acid.

[z]AFP: alpha-fetoprotein.

[aa]CEA: carcinoembryonic antigen.

[ab]ALB: albumin.

[ac]Tbli: total bilirubin.

[ad]Dbli: direct bilirubin.

[ae]Ibli: indirect bilirubin.

[af]HBP: high blood pressure.

[ag]TyG: triglyceride-glucose index.

[ah]CMetS: continuous metabolic syndrome score.

XSL•FO
RenderX

Figure 1. Study flowchart of participant selection and model development workflow. CAP: controlled attenuation parameter; KNN: k-nearest neighbors; LASSO: Least Absolute Shrinkage and Selection Operator; MLP: multilayer perceptron; NB: naive Bayes; NHANES: National Health and Nutrition Examination Survey; RF: random forest; RFE-RF: Recursive Feature Elimination based on Random Forest importance; ROC: receiver operating characteristic; SVM: support vector machine; USTC: University of Science and Technology of China; XGBoost: Extreme Gradient Boosting.

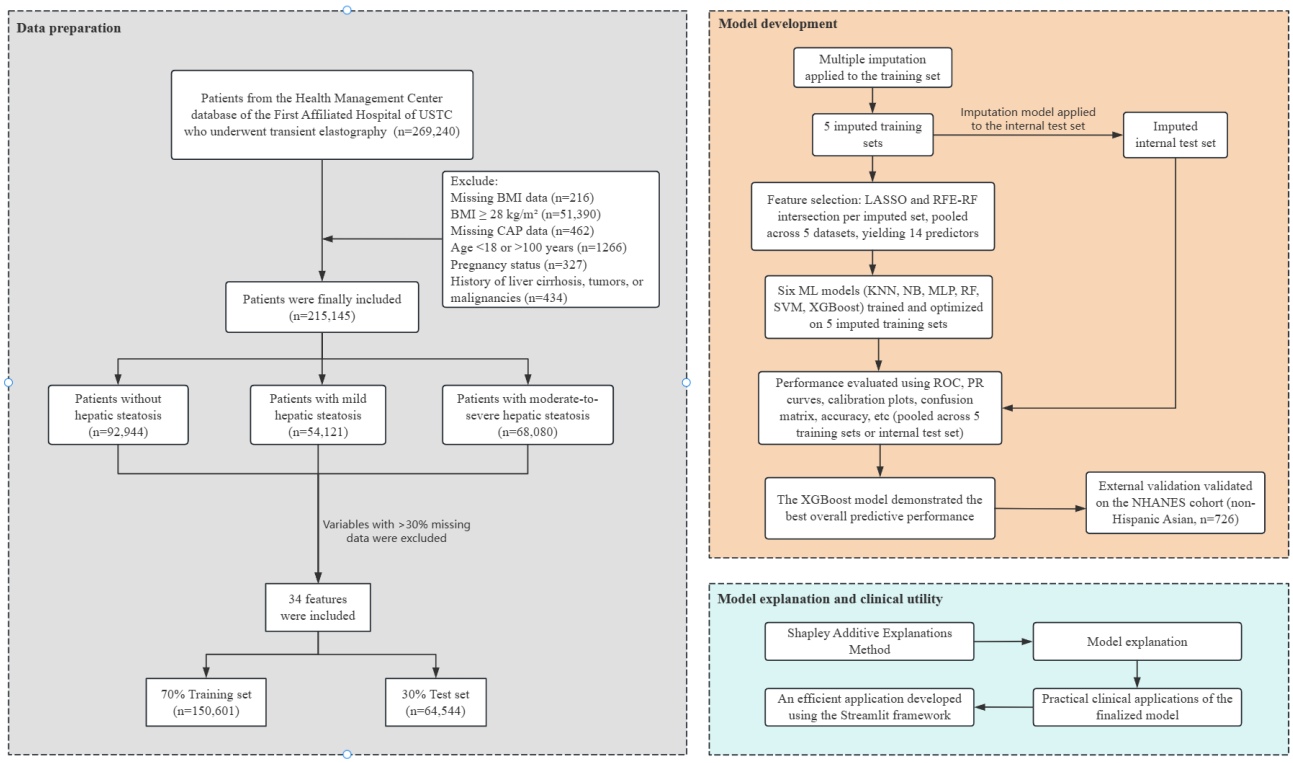

Feature Selection

Following the multiple imputation–based feature selection procedure, 14 predictors were consistently identified as robust determinants of hepatic steatosis severity in the nonobese population. These final selected variables encompass multiple physiological dimensions: anthropometric measures (BMI and WC), liver enzymes (ALT, AST, and alkaline phosphatase [ALP]), renal function indicators (uric acid [UA] and serum creatinine [SCr]), lipid profiles (triglycerides and low-density lipoprotein [LDL] cholesterol), metabolic composites (CMetS and TyG), as well as age, SBP, and albumin. Some individual metabolic parameters (eg, high-density lipoprotein cholesterol and FBG) were likely excluded as their collective information was more comprehensively represented by the selected composite measure, CMetS. LASSO coefficient paths, cross-validation error curves for optimal lambda selection, and RFE-RF accuracy curves are presented in Multimedia Appendix 5.

Model Performance Comparisons

Six ML models were developed for multiclass severity stratification (none, mild, and moderate to severe). All models demonstrated robust performance. In the training set, KNN achieved perfect discrimination (macro-average ROC-AUC=1.000) due to overfitting, followed by RF (0.959), XGBoost (0.945), MLP (0.942), SVM (0.917), and NB (0.913; Table 2 and Figure 2A). In the test set, XGBoost, MLP, and RF tied for the highest macroaverage ROC-AUC, each attaining 0.941 (Table 2 and Figure 3A). They were followed by KNN (0.934), SVM (0.917), and NB (0.913).

XSL•FO
RenderX

Table . Performance of each algorithm in the training and test set.

Performance metric	Training set						Test set					
	KNN[a]	MLP[b]	NB[c]	RF[d]	SVM[e]	XGBoost[f]	KNN	MLP	NB	RF	SVM	XGBoost
Accuracy	0.998	0.823	0.766	0.847	0.799	0.828	0.810	0.821	0.767	0.817	0.798	0.824
Cohen_Kappa	0.997	0.716	0.640	0.760	0.657	0.721	0.700	0.712	0.641	0.711	0.654	0.713
macro_ROC_AUC[g]	1.000	0.942	0.913	0.959	0.917	0.945	0.934	0.941	0.913	0.941	0.917	0.941
micro_ROC_AUC	1.000	0.947	0.918	0.962	0.921	0.950	0.940	0.946	0.918	0.946	0.921	0.946
macro_PR_AUC[h]	1.000	0.892	0.831	0.924	0.841	0.897	0.875	0.890	0.831	0.890	0.839	0.890
micro_PR_AUC	1.000	0.900	0.845	0.929	0.851	0.905	0.886	0.898	0.845	0.898	0.849	0.899
Precision (none)	0.999	0.879	0.824	0.893	0.882	0.882	0.867	0.878	0.824	0.874	0.877	0.879
Recall (none)	0.999	0.849	0.834	0.862	0.872	0.843	0.827	0.849	0.836	0.838	0.875	0.840
Specificity (none)	0.999	0.911	0.864	0.921	0.911	0.915	0.904	0.911	0.864	0.908	0.907	0.912
F_1-score (none)	0.999	0.863	0.829	0.877	0.877	0.862	0.847	0.864	0.830	0.855	0.876	0.859
ROC_AUC (none)	1.000	0.966	0.950	0.976	0.948	0.967	0.960	0.965	0.950	0.965	0.947	0.965
Precision (mild)	0.996	0.737	0.662	0.740	0.686	0.739	0.727	0.725	0.657	0.718	0.675	0.731
Recall (mild)	0.997	0.751	0.770	0.813	0.636	0.786	0.761	0.745	0.769	0.766	0.630	0.783
Specificity (mild)	0.999	0.910	0.868	0.904	0.902	0.906	0.904	0.905	0.865	0.899	0.898	0.903
F_1-score (mild)	0.997	0.744	0.712	0.775	0.660	0.762	0.744	0.735	0.709	0.741	0.652	0.756
ROC_AUC (mild)	1.000	0.891	0.839	0.924	0.843	0.896	0.877	0.889	0.839	0.890	0.843	0.890
Precision (moderate to severe)	0.999	0.820	0.784	0.881	0.773	0.832	0.803	0.825	0.794	0.826	0.782	0.832
Recall (moderate to severe)	0.998	0.846	0.669	0.854	0.830	0.840	0.825	0.844	0.671	0.828	0.827	0.836
Specificity (moderate to severe)	0.999	0.914	0.915	0.947	0.887	0.922	0.907	0.917	0.919	0.919	0.893	0.922
F_1-score (moderate to severe)	0.998	0.833	0.722	0.867	0.800	0.836	0.814	0.835	0.727	0.827	0.804	0.834

XSL•FO
RenderX

Performance metric	Training set						Test set					
	KNN[a]	MLP[b]	NB[c]	RF[d]	SVM[e]	XGBoost[f]	KNN	MLP	NB	RF	SVM	XGBoost
ROC_AUC (moderate to severe)	1.000	0.969	0.951	0.977	0.960	0.971	0.965	0.969	0.951	0.969	0.960	0.969

[a]KNN: k-nearest neighbors.

[b]MLP: multilayer perceptron.

[c]NB: naive Bayes.

[d]RF: random forest.

[e]SVM: support vector machine.

[f]XGBoost: Extreme Gradient Boosting.

[g]ROC-AUC: area under the receiver operating characteristic curve.

[h]PR-AUC: area under the precision-recall curve.

Figure 2. Performance of the XGBoost (Extreme Gradient Boosting) model in the training set. AUC: area under the curve; ROC: receiver operating characteristic.

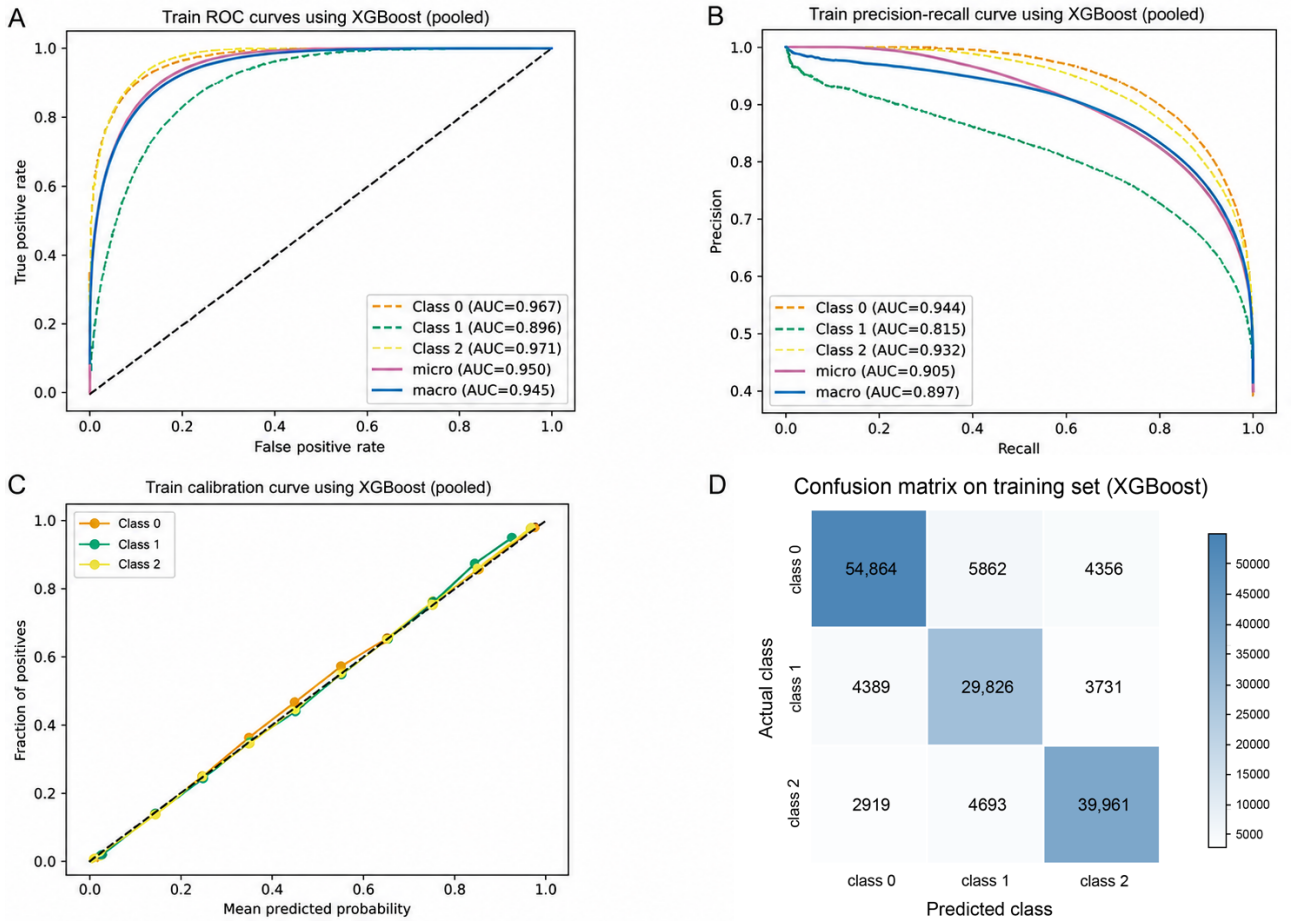

XSL•FO
RenderX

Figure 3. Performance of the XGBoost (Extreme Gradient Boosting) model in the test set. AUC: area under the curve; ROC: receiver operating characteristic.

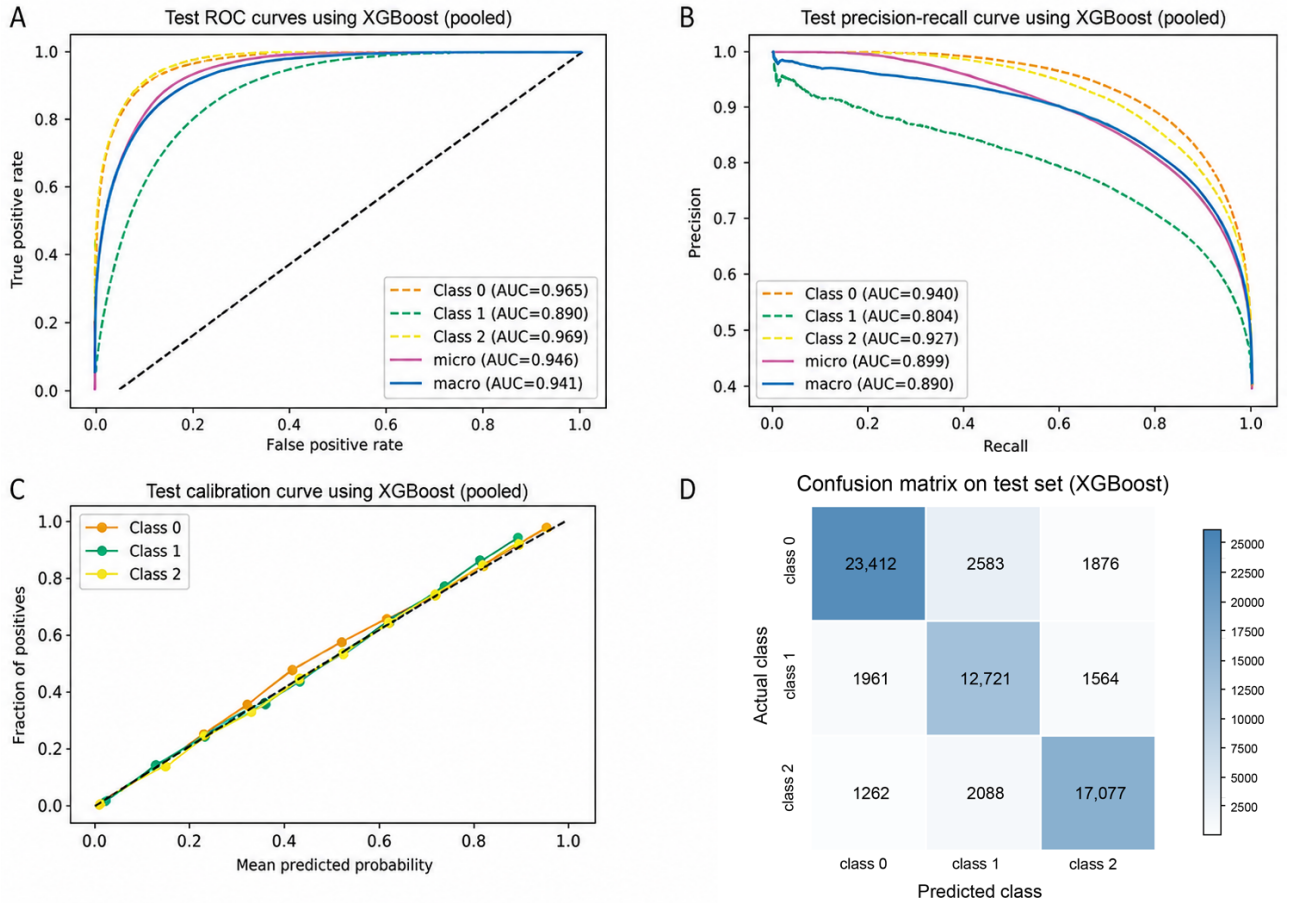

Precision-recall (PR) curves were analyzed to evaluate performance under class imbalance (Figure 2B for training and Figure 3B for test set of XGBoost). In the training set, KNN exhibited a perfect macro-average PR-AUC (1.000) as a result of overfitting. Among the remaining models, RF achieved the highest macro-average PR-AUC (0.924), followed by XGBoost (0.897), MLP (0.892), SVM (0.841), and NB (0.831). In the test set, XGBoost, MLP, and again RF achieved identical macro-average PR-AUC values of 0.890, marginally outperforming KNN (0.875), SVM (0.839), and NB (0.831). Calibration assessments (Figures 2C and 3C) demonstrated that XGBoost's predicted probabilities showed good alignment with observed outcomes. Its confusion matrix (Figures 2D and 3D) also exhibited a favorable pattern, characterized by relatively low misclassification density and fewer severe grading errors between adjacent categories. Corresponding figures for the other models are presented in Multimedia Appendix 6, and complete performance data across all imputed datasets are provided in Multimedia Appendix 7.

As detailed in Table 2, XGBoost achieved the best overall performance on the test set. It attained the highest macroaverage PR-AUC (0.890) and macroaverage ROC-AUC (0.941), along with an accuracy of 0.824 and a Cohen κ of 0.713. Although KNN showed exceptional training set performance (macro ROC-AUC=1.000), its test set performance declined to 0.934, indicating a degree of overfitting. While the MLP and RF were competitive in several metrics (macro-ROC-AUC=0.941 and macro-PR-AUC=0.890), XGBoost was selected for further analysis due to its consistently superior performance across all 3 severity classes, particularly demonstrating higher precision and sensitivity for the clinically important mild steatosis category—essential for early detection—as well as its stronger interpretability via SHAP and lower computational cost. Optimal hyperparameters for all models are provided in Multimedia Appendix 8.

Model Interpretation

SHAP analysis was used on the optimal XGBoost model to provide global and local explanations of its decision logic. The bee swarm plot (Figure 4A) shows the distribution and impact direction of SHAP values per feature across samples; and the bar plot (Figure 4B) ranks features by mean absolute SHAP value. Both analyses identified BMI, WC, ALT, and UA as the most influential predictors across all severity grades. A positive SHAP value indicates that a higher feature value pushes the prediction toward a more severe category (eg, elevated BMI, WC, and TyG increased the predicted risk of moderate-to-severe steatosis).

SHAP dependence plots for the 4 most important features (Figure 5) illustrate the nonlinear relationships between feature values and their SHAP contributions. Consistently, elevated levels of BMI, WC, ALT, and UA were associated with positive SHAP values, increasing the predicted risk for moderate-to-severe steatosis.

XSL•FO
RenderX

Local interpretation clarifies predictions for individual cases. The waterfall plot (Figure 6A) deconstructs how each feature shifts the model output from the baseline for a representative patient. The force plot (Figure 6B) visually maps the cumulative effect of all features in moving the prediction from the base value to the final outcome. These tools translate probabilistic outputs into clinically intelligible terms, allowing verification of biological plausibility and understanding of risk stratification rationale.

Figure 4. Global model interpretability using Shapley Additive Explanations. ALB: albumin; ALP: alkaline phosphatase; ALT: alanine aminotransferase; AST: aspartate aminotransferase; CMetS: continuous metabolic syndrome score; LDL: low-density lipoprotein cholesterol; SBP: systolic blood pressure; SCr: serum creatinine; SHAP: Shapley Additive Explanations; TG: triglyceride; TyG: triglyceride-glucose index; UA: uric acid; WC: waist circumference.

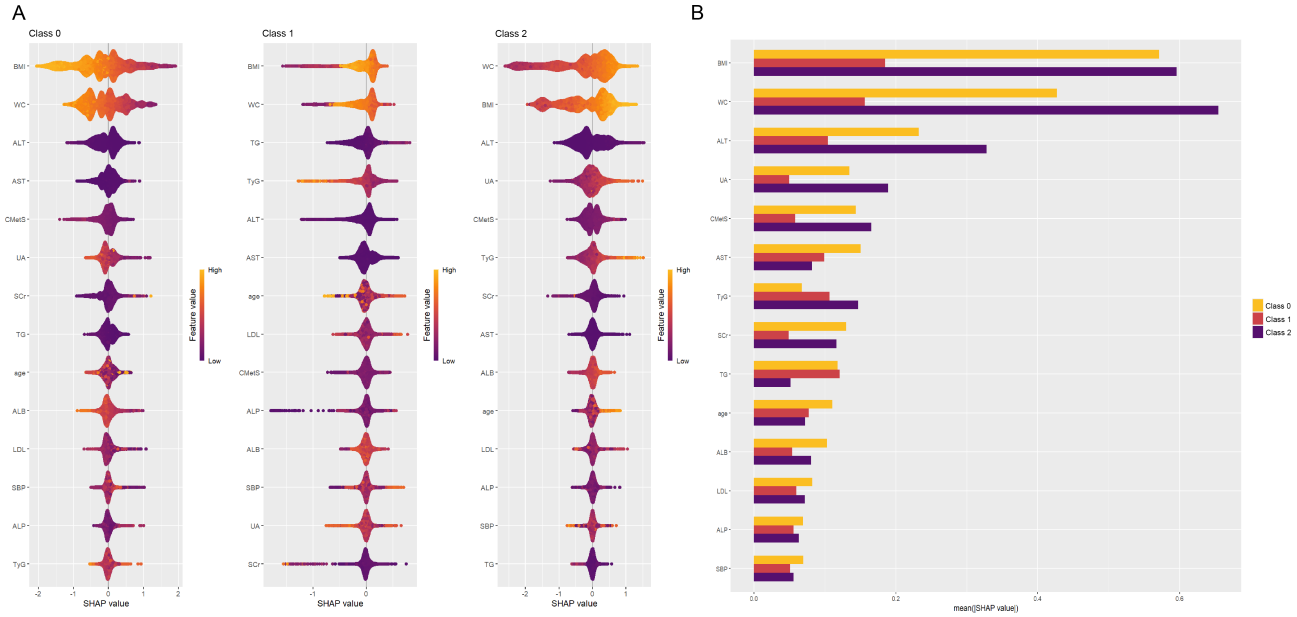

Figure 5. Shapley Additive Explanations dependence plots for key predictors. ALT: alanine aminotransferase; SHAP: Shapley Additive Explanations; UA: uric acid; WC: waist circumference.

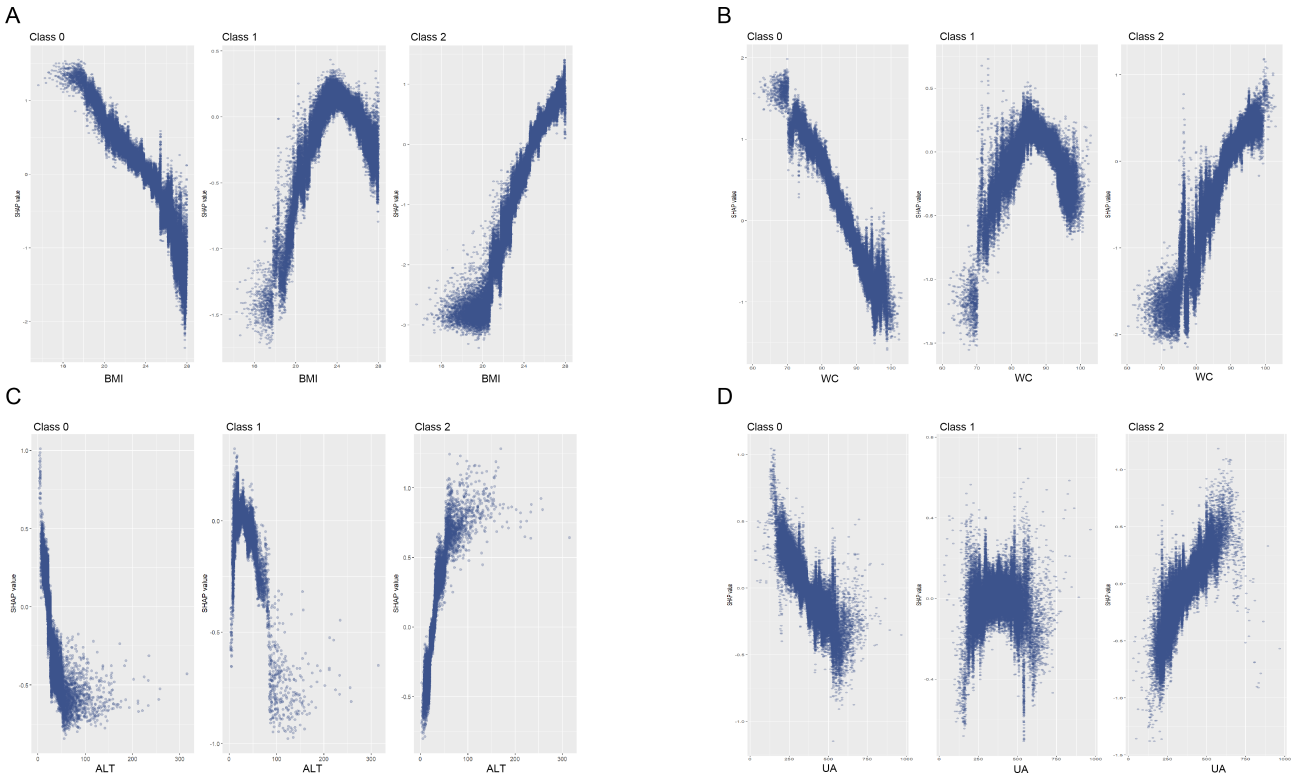

XSL•FO
RenderX

Figure 6. Local interpretability for individual predictions. ALB: albumin; ALP: alkaline phosphatase; ALT: alanine aminotransferase; AST: aspartate aminotransferase; CMetS: continuous metabolic syndrome score; LDL: low-density lipoprotein cholesterol; SBP: systolic blood pressure; SCr: serum creatinine; TG: triglycerides; TyG: triglyceride-glucose index; WC: waist circumference.

Comprehensive Validation and Comparative Analysis

For independent validation of generalizability, the optimized XGBoost model was applied to an external cohort from the NHANES 2017 to 2020 cycle. Among non-Hispanic Asian participants (RIDRETH3=6) with complete CAP measurements and BMI <28 kg/m², the initial eligible sample comprised 1684 individuals. After excluding participants with any missing values in the variables required by the final model, the external validation cohort consisted of 726 individuals with fully

XSL•FO
RenderX

observed data. The cohort exhibited the following distribution across hepatic steatosis grades based on CAP thresholds: none: 321 (44.2%), mild: 207 (28.5%), and moderate to severe: 198 (27.3%). For comparison, our internal cohort (n=215,145) demonstrated a distribution of none: 92,944 (43.2%), mild: 54,121 (25.2%), and moderate to severe: 68,080 (31.6%), indicating similar distribution patterns between cohorts (Multimedia Appendix 9). A comprehensive summary of the baseline demographic and clinical characteristics of this external validation cohort, stratified by steatosis grade, is provided in Multimedia Appendix 10. The model demonstrated robust performance in this independent external validation: macro-average ROC-AUC of 0.874 and micro-average ROC-AUC of 0.887 (Multimedia Appendix 11).

To assess the necessity of a dedicated model, a separate XGBoost model was trained on a general population dataset (including obese individuals) using the same features. When evaluated on the same held-out nonobese test set, the nonobese-specific model showed significantly superior discrimination (macroaverage ROC-AUC: 0.941 vs 0.852; $P<.001$; and microaverage ROC-AUC: 0.946 vs 0.869; $P<.001$), as detailed in Multimedia Appendix 11. This indicates that while a general model retains basic predictive capability, it fails to fully capture the distinct progression patterns in nonobese individuals, highlighting the value of population-specific tools.

To further evaluate clinical utility, a direct comparison was made with the established FLI. As FLI is a binary classifier, a dedicated binary XGBoost model was developed using a diagnostic threshold of CAP >248 dB/m to define "potential hepatic steatosis" (encompassing both mild and moderate-to-severe cases). Correspondingly, for FLI, a threshold of 30 was selected instead of the conventional 60, aligning with the objective of early identification in a nonobese screening context [26]. Both models were evaluated on both the internal test set and the external NHANES validation cohort. The performance of our binary XGBoost model is detailed in Multimedia Appendix 12. As summarized in the comparative table (Multimedia Appendix 13) and illustrated by the receiver operating characteristic (ROC) curve comparison (Multimedia Appendix 14), our model substantially outperformed the FLI. Specifically, it achieved areas under the curve (AUCs) of 0.947 (internal) and 0.892 (external), compared to the FLI's AUCs of 0.809 and 0.792, respectively. This demonstrates the superior discriminative ability of our ML approach, which leverages a broader set of metabolic and clinical features, for identifying steatosis in nonobese individuals.

Application of Model

To enhance clinical utility, a publicly accessible online prediction platform was developed [27]. By inputting routine indicators (eg, BMI, WC, liver function, and metabolic markers), users obtain real-time, individualized risk stratification (none, mild, or moderate-to-severe probability) and personalized management recommendations. The platform is designed with clinical practicality in mind. It provides predictions via the full 14-variable model when data are complete. To accommodate settings where tests such as albumin, uric acid, ALP, LDL cholesterol, or serum creatinine are unavailable, a simplified "9-core-feature" model has been implemented (see Multimedia Appendix 15 for the 9-core-feature model's performance). Among these, UA and SCr were excluded primarily due to their lower routine accessibility in primary care settings, as well as their relatively high missing rates in the original cohort (24.01% and 16.04%, respectively); excluding them reduces the uncertainty associated with multiple imputation. Albumin, LDL cholesterol, and ALP were excluded based on their relatively reduced predictive contribution in the full model. It is important to clarify that although the simplified model algorithmically uses 9 derived features, several of these are composite indices—namely, BMI, the TyG, and the continuous metabolic syndrome score—whose computation relies on multiple raw clinical inputs. Consequently, deploying the 9-feature model in practice requires the entry of 12 distinct primary measurements: age, sex, weight, height, waist circumference, ALT, AST, FBG, triglycerides, high-density lipoprotein cholesterol, SBP, and diastolic blood pressure. As detailed in Multimedia Appendix 16, while the macro-average ROC-AUC decreased from 0.941 (full model) to 0.917 (9-core model), the simplified model retains excellent discriminatory performance suitable for initial risk stratification. This dual-model design ensures functionality across varying levels of resource availability. Figure 7 shows an example interface, which allows users to actively select between the core-variable or full-variable model based on data availability. The platform adheres to data privacy standards with a user-friendly interface, bridging the translational gap between artificial intelligence (AI) models and primary health care.

XSL•FO
RenderX

Figure 7. Online prediction platform interface.

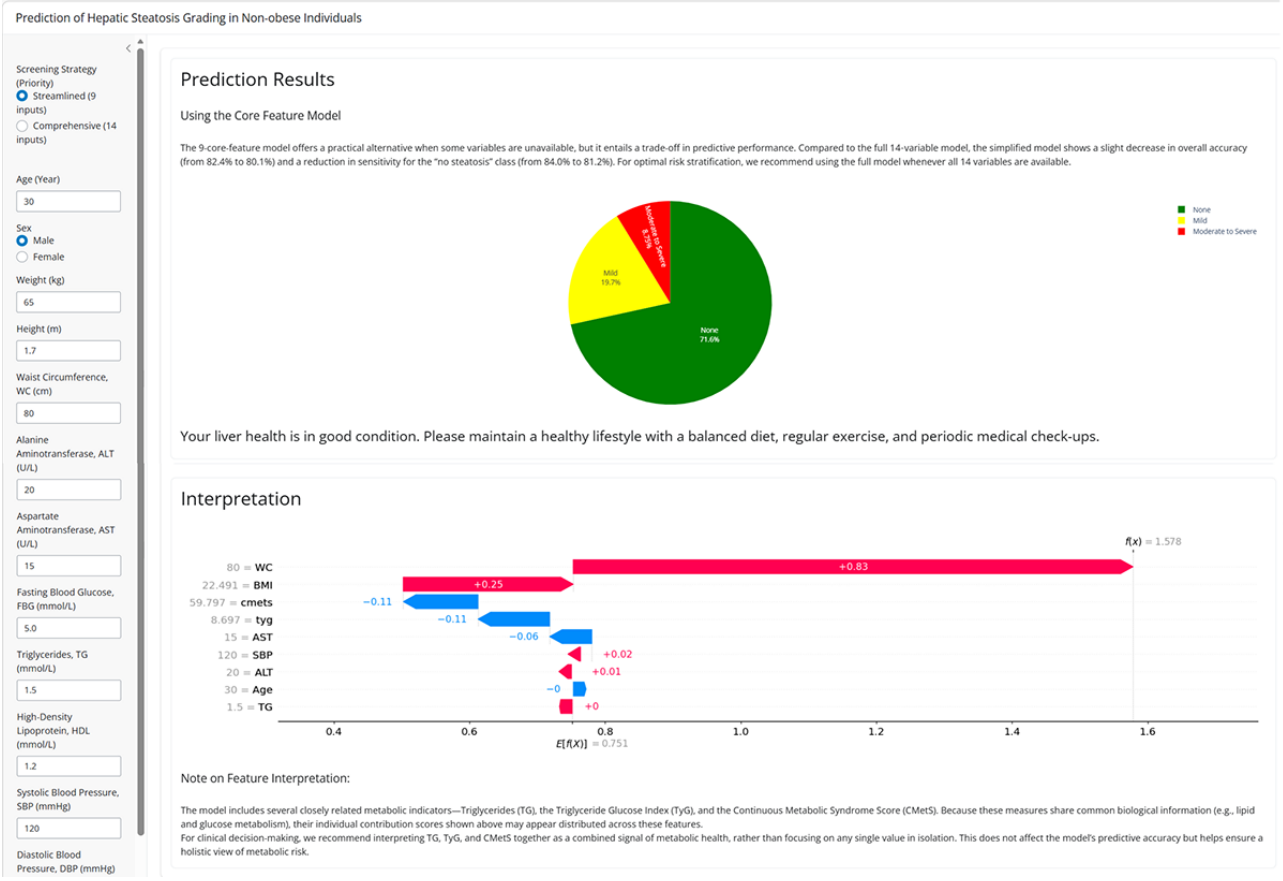

Discussion

Principal Findings

This study demonstrates that ML effectively predicts CAP-defined hepatic steatosis severity in nonobese populations, with the XGBoost model achieving superior performance. XGBoost yielded the highest macro-average ROC-AUC (0.941) and PR-AUC (0.890) among all models evaluated. More importantly, it maintained the best precision and sensitivity across all 3 severity classes. It should be noted that CAP is a surrogate marker for histologically confirmed steatosis, with inherent false-positive and false-negative rates compared to liver biopsy. Therefore, the model's output represents a proxy measure of disease severity, and clinical confirmation may require further diagnostic evaluation.

Nonobese individuals were defined as those with a BMI <28 kg/m². Global epidemiological studies report that 10% to 30% of nonobese individuals have SLD, with prevalence varying geographically (7% - 21% in Europe and America and 3% - 27% in Asia) [4,28,29]. This variability reflects racial differences, inconsistent BMI criteria, and the subjectivity of ultrasound-based diagnosis. The moderate-to-severe steatosis prevalence in our cohort (31.6%) was slightly higher, possibly due to the urban hospital-based sample, as urban populations typically exhibit higher SLD prevalence [3]. Notably, nonobese SLD is associated with significant clinical risk. All-cause mortality in nonobese individuals with SLD (12.1/1000 person-years) is comparable to the overall SLD mortality rate (15.4/1000 person-years) [4], while liver-related mortality in this group may be disproportionately high [30]. These observations underscore that nonobese hepatic steatosis carries outcomes at least comparable to, if not more severe than, those in the general population. The absence of targeted prediction tools contributes to underdiagnosis and delayed intervention. It should be noted that while the cohort was defined as "nonobese" (BMI <28 kg/m²), the median BMI of the moderate-to-severe steatosis group (26.10 kg/m²) falls within the overweight range. Therefore, the predictive performance of the model is primarily representative of and applicable to an "overweight-to-nonobese" population, rather than a predominantly lean population (BMI <24 kg/m²). Nevertheless, our cohort included a substantial proportion of individuals within the normal weight or lean range, constituting approximately 46% of the total sample. The model retained discriminatory ability within this subgroup, indicating its capacity to identify relevant metabolic patterns across the spectrum of nonobese hepatic steatosis, including in leaner individuals.

In this study, we have identified BMI, WC, triglycerides, LDL, ALT, AST, age, SBP, ALP, albumin, UA, SCr, CMetS, and TyG as key factors for hepatic steatosis screening. It should be noted that our training data included mixed steatosis etiologies (eg, metabolic dysfunction–associated SLD and alcohol-related liver disease), and the model predicts the severity of hepatic steatosis without distinguishing its underlying etiology. Consequently, the relevance and contribution of these key predictors might vary across populations with differing prevalent causes. Additionally, key predictors such as albumin (28.6%) and UA (24.0%) had substantial missingness rates; despite

XSL•FO
RenderX

rigorous multiple imputation, this remains a limitation as imputation may introduce noise. Several predictors are well-established risk factors for SLD, including obesity-related indicators (BMI and WC) [31], lipid metabolism markers (triglycerides and LDL) [32], and liver enzymes (ALT and AST) [33]. Age is a consistent correlate of SLD risk [34]. Hypertension, reflected by elevated SBP, is independently associated with hepatic steatosis [35]. Albumin may be linked to SLD through oxidative stress and inflammation, which induce posttranslational modifications that impair its antioxidant and anti-inflammatory functions, potentially exacerbating disease [36]. This aligns with findings that the albumin-to-alkaline phosphatase ratio inversely predicts SLD risk in nonobese Chinese cohorts [37]. Our study similarly incorporated ALP and albumin measurements for predicting SLD in nonobese populations, yielding findings consistent with these conclusions. SCr concentration is independently associated with SLD in middle-aged and older Chinese adults, with risk increasing even within the normal range [38], and the UA-to-SCr ratio rises with FLD severity [39]. Compared to the non-SLD group, the adjusted odds ratio for the mild SLD group was 1.147 (95% CI 1.099 - 1.196), while the OR for the moderate-to-severe SLD group increased to 1.275 (95% CI 1.212 - 1.341). Interestingly, robust studies have revealed a strong association between SLD and CKD [40], regardless of underlying confounding conditions, such as obesity, hypertension, and type 2 diabetes. The TyG index, calculated from FBG and triglycerides, has substantial evidence demonstrating its positive correlation with the severity of insulin resistance in various metabolic diseases [41]. Insulin resistance is also a shared characteristic of SLD across different severity levels. A study from China indicated that the TyG index can serve as a predictor for SLD [42], with an ROC-AUC reaching 0.78. This study further confirms the significance of the TyG index in predicting SLD. CMetS, a composite metric initially developed for cardiovascular risk stratification [25], emerged as an independent predictor here. By integrating several key metabolic components into a single measure, CMetS provides a more informative and holistic summary of metabolic dysfunction, capturing the multifaceted pathophysiological patterns associated with hepatic steatosis severity in this nonobese population. For external validation, the Chinese-derived CMetS formula was applied directly to the NHANES non-Hispanic Asian participants without recalibration. We acknowledge the potential limitation of this approach due to possible differences in metabolic baselines. However, the restricted non-Hispanic Asian subgroup shares phenotypic similarities with the derivation cohort, making this a reasonable test of transferability. Crucially, the model maintained robust performance (macro-average ROC-AUC=0.874), supporting the score's utility in these metabolically similar groups. An important consideration when interpreting the SHAP results is the multicollinearity among key metabolic predictors—triglycerides, TyG, and CMetS. Although this does not affect the model's predictive accuracy, it can lead to distributed attribution of shared metabolic risk across these correlated features in SHAP analysis, which may cause clinicians to underestimate the specific contribution of lipid levels such as triglycerides. Therefore, clinicians should interpret the SHAP contributions of triglycerides, TyG, and CMetS collectively as an integrated signal of metabolic dysfunction, rather than in isolation.

The model's interpretability and accessibility through an online platform facilitate its potential integration as a pre-screening and triage tool in primary care. During routine checkups, practitioners can input readily available parameters to obtain immediate 3-tiered risk stratification. This output supports differentiated management: patients stratified as "none" may require routine follow-up; those with "mild" steatosis can be targeted for lifestyle counseling and monitoring; and those with "moderate-to-severe" disease can be prioritized for confirmatory testing (eg, CAP) or specialist referral. It should be noted that the model's precision for the "mild" category is relatively limited (test set precision: 0.731 and sensitivity: 0.783). Therefore, we recommend that this classification—and cases with predicted probabilities near any diagnostic threshold—be interpreted cautiously in conjunction with the full clinical context and may warrant additional assessment. To address the practical challenge of acquiring the full 14-variable panel in some settings, we provide a validated simplified model based on 9 core variables, offering a pragmatic alternative for initial risk stratification. While the simplified model retains acceptable discriminatory performance suitable for initial risk stratification (macro-average ROC-AUC decreasing from 0.941 to 0.917), it entails a trade-off, with overall accuracy declining from 82.4% to 80.1% and sensitivity for the "no steatosis" class decreasing from 84.0% to 81.2%. Clinicians should be aware that this parsimonious version may underestimate the probability of mild steatosis, potentially leading to a higher false-negative rate in low-risk classifications. Notably, manual documentation errors often limit the application of comprehensive prediction models. Ambient AI scribes can enhance electronic health record completeness and accuracy by automatically capturing clinical encounters [43]. Integrating such tools into routine practice could facilitate consistent predictor collection. This, in turn, would allow broader deployment of our full 14-variable model and improve its real-world performance. This convergence of AI-enhanced data capture and ML-based risk stratification represents a valuable direction for future research and clinical implementation.

Comparison With Prior Work

Prior studies have extensively validated ML for predicting SLD. Chen et al demonstrated an XGBoost model outperformed the FLI (ROC-AUC 0.882) for identifying ultrasound-defined moderate-to-severe fatty liver [18]. Subsequent work confirmed the robust performance of models using routine clinical indicators in broader populations (AUC 0.81 - 0.89) [19-21]. These studies established a foundation for data-driven prediction but were primarily designed for general or obese populations and relied on ultrasound for binary classification, limiting nuanced risk stratification for nonobese individuals. A notable advance by Su et al [22] focused specifically on lean individuals (BMI <23 kg/m²), achieving an ROC-AUC of 0.885 with a neural network, highlighting the value of population-specific modeling. However, their approach remained a binary classifier based on ultrasound, which has limited sensitivity for mild steatosis. Other advances have integrated multiomics data; for example, Oh et al [44] identified a cross-ethnic gut microbiome

XSL•FO
RenderX

signature for predicting cirrhosis (AUC up to 0.91). While offering high performance and mechanistic insight, such approaches rely on specialized assays (eg, metagenomic sequencing), posing challenges for routine, large-scale screening. This study extends prior work through several focused modifications: (1) using the CAP as a quantitative, more sensitive standard than ultrasound; (2) specifically developing the model for a nonobese cohort (BMI <28 kg/m²), with direct validation showing superiority over a general population model (macro-average ROC-AUC 0.941 vs 0.852; $P<.001$); (3) implementing multiclass severity prediction for finer risk stratification; and (4) using only routinely available variables deployed via an interpretable online tool to enhance clinical utility.

The translation of predictive models into clinical practice requires careful consideration of the uncertainties that may affect their reliability. These uncertainties are primarily 3-fold: external uncertainty, relating to the model's generalizability across populations with differing data distributions; model uncertainty, encompassing both parametric instability and the structural limitations of the chosen algorithm in capturing complex disease pathophysiology; and missing data uncertainty, which arises from incomplete observations. In this study, multiple imputation was applied to the training set (5 datasets), while the test set was imputed only once to reflect real-world deployment. The ensemble of 5 models propagates imputation uncertainty from the training stage. However, a limitation is that single imputation of the test set does not quantify the uncertainty inherent in its own missing values. Novel modeling architectures—such as deep self-supervised learning with advanced feature selection for learning robust representations from unlabeled data [45], and frameworks integrating automated feature engineering with dimensionality reduction to handle data imperfections [46]—aim to develop more reliable and adaptable tools for multidimensional health risk stratification.

In summary, an interpretable XGBoost model was established to predict CAP-defined hepatic steatosis severity in nonobese individuals using routinely available data. By adopting CAP as a quantitative standard, the model enables multiclass risk stratification. Its interpretability, validated generalizability, and deployment as an accessible online platform support its potential integration into clinical workflows for early identification and management in this population.

Limitations

This study has several limitations. First, generalizability may be limited as the data came predominantly from urban hospitals in Eastern China, lacking rural and diverse ethnic representation, despite robust performance in an external cohort. Additionally, evaluating the model exclusively on participants with complete laboratory records in the external validation may introduce selection bias, as individuals with missing data could differ systematically from those retained. Second, although we compared and selected the best-performing model among 6 candidate algorithms, alternative or more complex architectures may offer superior capabilities for explicit uncertainty quantification. Third, the ongoing debate over unified CAP diagnostic thresholds could affect the consistency of steatosis grading. Finally, the training data included mixed steatosis etiologies, and the model does not distinguish underlying causes. Future validation using larger, more diverse external cohorts with prospectively collected etiological data is needed to further confirm the stability of our performance estimates and enhance clinical utility.

Conclusions

The interpretable XGBoost model was established for predicting CAP-defined hepatic steatosis severity in nonobese populations based on routinely available clinical and laboratory data in this study. This model demonstrated robust predictive performance in both the internal test set and an external validation cohort. To promote practical application, the model as an openly accessible online prediction platform was used.

Acknowledgments

We extend our gratitude to the Health Management Center of the First Affiliated Hospital of University of Science and Technology of China for their invaluable data support. We also sincerely thank all the participants for their contributions and cooperation. Additionally, we acknowledge the use of generative AI tools DeepSeek for language polishing and consistency checking of technical terminology. The authors subsequently reviewed and edited the content as needed and take full responsibility for the accuracy, integrity, and all scientific content of the published article.

Funding

This work was partly supported by the National Natural Science Foundation of China (grant 82170618), National Natural Science Foundation of China (82170618), Anhui University Outstanding Youth Research Project (2024AH020016), Research Institute of Pioneer Medicine and Frontier Technology, Hefei Comprehensive National Science Center (2023IHM01020), and Key Research and Development Program of Anhui Province (202204295107020019).

Data Availability

The clinical datasets generated and analyzed during this study are not publicly available in accordance with patient privacy regulations and institutional data governance policies. To support academic reproducibility and collaboration, deidentified data may be made available to qualified researchers upon reasonable request to the corresponding author, subject to review and approval by the relevant ethics committee and data custodians. External validation data are publicly available from the National Health

XSL•FO
RenderX

and Nutrition Examination Survey repository (2017-2020) [47]. The core analysis code, including scripts for data preprocessing, multiple imputation, feature selection with Least Absolute Shrinkage and Selection Operator regression, calculation of composite scores (eg, continuous metabolic syndrome score), hyperparameter optimization, and model training and evaluation for all 6 algorithms, has been deposited in a public repository to ensure reproducibility. The code is available on GitHub [48].

Authors' Contributions

YZ and YW are co-first authors. YZ, JW, and LL are co-corresponding authors. YZ can be reached via email at zyl2020@ustc.edu.cn; JW can be reached via email at wangjoe@ustc.edu.cn; LL can be reached via email at liulx@ustc.edu.cn. YZ, JW, and LL conceptualized the study. YZ provided the resources necessary for the study. YW and FZ verified the raw data in the study. YW and YZ proposed the experimental design and algorithm framework and implemented the model design and evaluation. SZ, JY, and YL curated the data. JS provided and validated the external validation data. YZ and YW wrote the original draft of the manuscript. YZ, JW, and LL supervised the research activities. LL managed the project administration. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data missingness pattern for all collected variables in the final study cohort (n=215,145).
[DOCX File, 25318 KB - jmir_v28i1e82529_app1.docx]

Multimedia Appendix 2
Distribution of the 6 most influential predictors across hepatic steatosis severity categories.
[DOCX File, 210 KB - jmir_v28i1e82529_app2.docx]

Multimedia Appendix 3
Baseline characteristics of participants in the training and test set.
[DOC File, 29 KB - jmir_v28i1e82529_app3.doc]

Multimedia Appendix 4
Density plots comparing pre- and post-imputation distributions for key continuous predictors in the training set.
[DOCX File, 1179 KB - jmir_v28i1e82529_app4.docx]

Multimedia Appendix 5
Feature selection: Least Absolute Shrinkage and Selection Operator regression and Recursive Feature Elimination based on Random Forest algorithms.
[DOCX File, 2645 KB - jmir_v28i1e82529_app5.docx]

Multimedia Appendix 6
Performance of 5 alternative machine learning models.
[DOCX File, 4024 KB - jmir_v28i1e82529_app6.docx]

Multimedia Appendix 7
Performance of each algorithm across the 5 imputed datasets.
[DOCX File, 47 KB - jmir_v28i1e82529_app7.docx]

Multimedia Appendix 8
The optimal hyperparameters of each algorithm.
[DOC File, 21 KB - jmir_v28i1e82529_app8.doc]

Multimedia Appendix 9
Sample size distribution by hepatic steatosis grade in the internal cohort and external National Health and Nutrition Examination Survey validation cohort.
[DOCX File, 187 KB - jmir_v28i1e82529_app9.docx]

XSL•FO
RenderX

Multimedia Appendix 10
Baseline characteristics of the National Health and Nutrition Examination Survey external validation cohort.
[DOCX File, 23 KB - jmir_v28i1e82529_app10.docx]

Multimedia Appendix 11
External validation and model comparison using receiver operating characteristic (ROC) curves.
[DOCX File, 266 KB - jmir_v28i1e82529_app11.docx]

Multimedia Appendix 12
Performance of the binary XGBoost (Extreme Gradient Boosting) model for hepatic steatosis detection.
[DOCX File, 252 KB - jmir_v28i1e82529_app12.docx]

Multimedia Appendix 13
Performance comparison between the binary XGBoost (Extreme Gradient Boosting) model and Fatty Liver Index for hepatic steatosis detection.
[DOCX File, 18 KB - jmir_v28i1e82529_app13.docx]

Multimedia Appendix 14
Comparison of receiver operating characteristic curves between the binary XGBoost (Extreme Gradient Boosting) model and Fatty Liver Index (FLI) for hepatic steatosis detection.
[DOCX File, 272 KB - jmir_v28i1e82529_app14.docx]

Multimedia Appendix 15
Performance of the 9-core-feature XGBoost (Extreme Gradient Boosting) model on the test set.
[DOCX File, 463 KB - jmir_v28i1e82529_app15.docx]

Multimedia Appendix 16
Performance comparison between 14-feature and 9-core-feature XGBoost (Extreme Gradient Boosting) models for nonobese hepatic steatosis severity prediction.
[DOCX File, 19 KB - jmir_v28i1e82529_app16.docx]

References

1. Rinella ME, Neuschwander-Tetri BA, Siddiqui MS, et al. AASLD practice guidance on the clinical assessment and management of nonalcoholic fatty liver disease. Hepatology 2023 May 1;77(5):1797-1835. [doi: 10.1097/HEP.0000000000000323] [Medline: 36727674]
2. Lou TW, Yang RX, Fan JG. The global burden of fatty liver disease: the major impact of China. Hepatobiliary Surg Nutr 2024 Feb 1;13(1):119-123. [doi: 10.21037/hbsn-23-556] [Medline: 38322226]
3. Browning JD, Szczepaniak LS, Dobbins R, et al. Prevalence of hepatic steatosis in an urban population in the United States: impact of ethnicity. Hepatology 2004 Dec;40(6):1387-1395. [doi: 10.1002/hep.20466] [Medline: 15565570]
4. Ye Q, Zou B, Yeo YH, et al. Global prevalence, incidence, and outcomes of non-obese or lean non-alcoholic fatty liver disease: a systematic review and meta-analysis. Lancet Gastroenterol Hepatol 2020 Aug;5(8):739-752. [doi: 10.1016/S2468-1253(20)30077-7]
5. Hagström H, Nasr P, Ekstedt M, et al. Risk for development of severe liver disease in lean patients with nonalcoholic fatty liver disease: a long-term follow-up study. Hepatol Commun 2018 Jan;2(1):48-57. [doi: 10.1002/hep4.1124] [Medline: 29404512]
6. Wang AY, Dhaliwal J, Mouzaki M. Lean non-alcoholic fatty liver disease. Clin Nutr 2019 Jun;38(3):975-981. [doi: 10.1016/j.clnu.2018.08.008] [Medline: 30466956]
7. VanWagner LB, Armstrong MJ. Lean NAFLD: a not so benign condition? Hepatol Commun 2018 Jan;2(1):5-8. [doi: 10.1002/hep4.1143] [Medline: 29404505]
8. Kim D, Kim WR. Nonobese fatty liver disease. Clin Gastroenterol Hepatol 2017 Apr;15(4):474-485. [doi: 10.1016/j.cgh.2016.08.028] [Medline: 27581063]
9. Paige JS, Bernstein GS, Heba E, et al. A pilot comparative study of quantitative ultrasound, conventional ultrasound, and MRI for predicting histology-determined steatosis grade in adult nonalcoholic fatty liver disease. AJR Am J Roentgenol 2017 May;208(5):W168-W177. [doi: 10.2214/AJR.16.16726] [Medline: 28267360]
10. Li Y, Wang X, Zhang J, Zhang S, Jiao J. Applications of artificial intelligence (AI) in researches on non-alcoholic fatty liver disease(NAFLD): a systematic review. Rev Endocr Metab Disord 2022 Jun;23(3):387-400. [doi: 10.1007/s11154-021-09681-x] [Medline: 34396467]

XSL•FO
RenderX

11. Castera L, Friedrich-Rust M, Loomba R. Noninvasive assessment of liver disease in patients with nonalcoholic fatty liver disease. Gastroenterology 2019 Apr;156(5):1264-1281. [doi: 10.1053/j.gastro.2018.12.036] [Medline: 30660725]
12. Karlas T, Petroff D, Sasso M, et al. Individual patient data meta-analysis of controlled attenuation parameter (CAP) technology for assessing steatosis. J Hepatol 2017 May;66(5):1022-1030. [doi: 10.1016/j.jhep.2016.12.022] [Medline: 28039099]
13. Eddowes PJ, Sasso M, Allison M, et al. Accuracy of FibroScan controlled attenuation parameter and liver stiffness measurement in assessing steatosis and fibrosis in patients with nonalcoholic fatty liver disease. Gastroenterology 2019 May;156(6):1717-1730. [doi: 10.1053/j.gastro.2019.01.042] [Medline: 30689971]
14. Sasso M, Beaugrand M, de Ledinghen V, et al. Controlled attenuation parameter (CAP): a novel VCTE. Ultrasound Med Biol 2010 Nov;36(11):1825-1835. [doi: 10.1016/j.ultrasmedbio.2010.07.005]
15. Blond E, Disse E, Cuerq C, et al. EASL-EASD-EASO clinical practice guidelines for the management of non-alcoholic fatty liver disease in severely obese people: do they lead to over-referral? Diabetologia 2017 Jul;60(7):1218-1222. [doi: 10.1007/s00125-017-4264-9] [Medline: 28352941]
16. Pouwels S, Sakran N, Graham Y, et al. Non-alcoholic fatty liver disease (NAFLD): a review of pathophysiology, clinical management and effects of weight loss. BMC Endocr Disord 2022 Mar 14;22(1):63. [doi: 10.1186/s12902-022-00980-1] [Medline: 35287643]
17. Dietrich P, Hellerbrand C. Non-alcoholic fatty liver disease, obesity and the metabolic syndrome. Best Pract Res Clin Gastroenterol 2014 Aug;28(4):637-653. [doi: 10.1016/j.bpg.2014.07.008] [Medline: 25194181]
18. Chen YY, Lin CY, Yen HH, et al. Machine-learning algorithm for predicting fatty liver disease in a Taiwanese population. J Pers Med 2022 Jun 23;12(7):1026. [doi: 10.3390/jpm12071026] [Medline: 35887527]
19. Deng J, Ji W, Liu H, et al. Development and validation of a machine learning-based framework for assessing metabolic-associated fatty liver disease risk. BMC Public Health 2024 Sep 18;24(1):2545. [doi: 10.1186/s12889-024-19882-z] [Medline: 39294603]
20. Weng S, Hu D, Chen J, Yang Y, Peng D. Prediction of fatty liver disease in a Chinese population using machine-learning algorithms. Diagnostics (Basel) 2023 Mar 18;13(6):1168. [doi: 10.3390/diagnostics13061168]
21. Huang G, Jin Q, Mao Y. Predicting the 5-year risk of nonalcoholic fatty liver disease using machine learning models: prospective cohort study. J Med Internet Res 2023 Sep 12;25:e46891. [doi: 10.2196/46891] [Medline: 37698911]
22. Su PY, Chen YY, Lin CY, Su WW, Huang SP, Yen HH. Comparison of machine learning models and the fatty liver index in predicting lean fatty liver. Diagnostics (Basel) 2023 Apr 13;13(8):1407. [doi: 10.3390/diagnostics13081407] [Medline: 37189508]
23. Bei‐Fan Z, the Cooperative Meta‐analysis Group of Working Group on Obesity in China. Predictive values of body mass index and waist circumference for risk factors of certain related diseases in Chinese adults: study on optimal cut‐off points of body mass index and waist circumference in Chinese adults. Asia Pac J Clin Nutr 2002 Dec;11(s8). [doi: 10.1046/j.1440-6047.11.s8.9.x]
24. Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. The Lancet 2004 Jan;363(9403):157-163. [doi: 10.1016/S0140-6736(03)15268-3]
25. Yang S, Yu B, Yu W, et al. Development and validation of an age-sex-ethnicity-specific metabolic syndrome score in the Chinese adults. Nat Commun 2023;14(1). [doi: 10.1038/s41467-023-42423-y]
26. Bedogni G, Bellentani S, Miglioli L, et al. The fatty liver index: a simple and accurate predictor of hepatic steatosis in the general population. BMC Gastroenterol 2006 Nov 2;6(1):33. [doi: 10.1186/1471-230X-6-33] [Medline: 17081293]
27. Hepatic steatosis grading prediction platform. URL: https://hepaticsteatosisgradingprediction.shinyapps.io/sld-grading-prediction/ [accessed 2026-05-04]
28. Kim D, Kim W, Joo SK, et al. Predictors of nonalcoholic steatohepatitis and significant fibrosis in non-obese nonalcoholic fatty liver disease. Liver Int 2019 Feb;39(2):332-341. [doi: 10.1111/liv.13983] [Medline: 30298568]
29. Shi Y, Wang Q, Sun Y, et al. The prevalence of lean/nonobese nonalcoholic fatty liver disease: a systematic review and meta-analysis. J Clin Gastroenterol 2020 Apr;54(4):378-387. [doi: 10.1097/MCG.0000000000001270] [Medline: 31651571]
30. Lonardo A, Byrne CD, Caldwell SH, Cortez-Pinto H, Targher G. Global epidemiology of nonalcoholic fatty liver disease: meta-analytic assessment of prevalence, incidence, and outcomes. Hepatology 2016 Oct;64(4):1388-1389. [doi: 10.1002/hep.28584] [Medline: 27038241]
31. Ayada I, van Kleef LA, Alferink LJM, Li P, de Knegt RJ, Pan Q. Systematically comparing epidemiological and clinical features of MAFLD and NAFLD by meta-analysis: focusing on the non-overlap groups. Liver Int 2022 Feb;42(2):277-287. [doi: 10.1111/liv.15139] [Medline: 34953098]
32. Kyhl LK, Nordestgaard BG, Tybjærg-Hansen A, Nielsen SF. High fat in blood and body and increased risk of clinically diagnosed non-alcoholic fatty liver disease in 105,981 individuals. Atherosclerosis 2023 Jul;376:1-10. [doi: 10.1016/j.atherosclerosis.2023.05.015] [Medline: 37253311]
33. Xuan Y, Wu D, Zhang Q, Yu Z, Yu J, Zhou D. Elevated ALT/AST ratio as a marker for NAFLD risk and severity: insights from a cross-sectional analysis in the United States. Front Endocrinol 2024;15. [doi: 10.3389/fendo.2024.1457598]
34. Peng H, Pan L, Ran S, et al. Prediction of MAFLD and NAFLD using different screening indexes: a cross-sectional study in U.S. adults. Front Endocrinol 2023;14. [doi: 10.3389/fendo.2023.1083032]

XSL•FO
RenderX

35. Yuan M, He J, Hu X, et al. Hypertension and NAFLD risk: insights from the NHANES 2017-2018 and Mendelian randomization analyses. Chin Med J (Engl) 2024 Feb 20;137(4):457-464. [doi: 10.1097/CM9.0000000000002753] [Medline: 37455323]
36. Wu N, Liu T, Tian M, et al. Albumin, an interesting and functionally diverse protein, varies from “native” to “effective” (Review). Mol Med Rep 2024 Feb;29(2):24. [doi: 10.3892/mmr.2023.13147] [Medline: 38099350]
37. Sheng G, Peng N, Hu C, Zhong L, Zhong M, Zou Y. The albumin-to-alkaline phosphatase ratio as an independent predictor of future non-alcoholic fatty liver disease in a 5-year longitudinal cohort study of a non-obese Chinese population. Lipids Health Dis 2021 May 16;20(1):50. [doi: 10.1186/s12944-021-01479-9] [Medline: 33993872]
38. Niu Y, Zhang W, Zhang H, et al. Serum creatinine levels and risk of nonalcohol fatty liver disease in a middle-aged and older Chinese population: a cross-sectional analysis. Diabetes Metab Res Rev 2022 Feb;38(2):e3489. [doi: 10.1002/dmrr.3489] [Medline: 34344058]
39. Choi J, Joe H, Oh JE, Cho YJ, Shin HS, Heo NH. The correlation between NAFLD and serum uric acid to serum creatinine ratio. PLoS ONE 2023;18(7):e0288666. [doi: 10.1371/journal.pone.0288666] [Medline: 37463147]
40. Byrne CD, Targher G. NAFLD as a driver of chronic kidney disease. J Hepatol 2020 Apr;72(4):785-801. [doi: 10.1016/j.jhep.2020.01.013] [Medline: 32059982]
41. Wei X, Min Y, Song G, Ye X, Liu L. Association between triglyceride-glucose related indices with the all-cause and cause-specific mortality among the population with metabolic syndrome. Cardiovasc Diabetol 2024 Apr 24;23(1):134. [doi: 10.1186/s12933-024-02215-0] [Medline: 38658993]
42. Zhang S, Du T, Zhang J, et al. The triglyceride and glucose index (TyG) is an effective biomarker to identify nonalcoholic fatty liver disease. Lipids Health Dis 2017 Dec;16(1). [doi: 10.1186/s12944-017-0409-6]
43. Leung TI, Coristine AJ, Benis A. AI scribes in health care: balancing transformative potential with responsible integration. JMIR Med Inform 2025 Aug 1;13:e80898. [doi: 10.2196/80898] [Medline: 40749188]
44. Oh TG, Kim SM, Caussy C, et al. A universal gut-microbiome-derived signature predicts cirrhosis. Cell Metab 2020 Nov 3;32(5):878-888. [doi: 10.1016/j.cmet.2020.06.005] [Medline: 32610095]
45. Tutsoy O, Koç GG. Deep self-supervised machine learning algorithms with a novel feature elimination and selection approaches for blood test-based multi-dimensional health risks classification. BMC Bioinformatics 2024 Mar 8;25(1):103. [doi: 10.1186/s12859-024-05729-2] [Medline: 38459463]
46. Tutsoy O, Sumbul HE. A novel deep machine learning algorithm with dimensionality and size reduction approaches for feature elimination: thyroid cancer diagnoses with randomly missing data. Brief Bioinform 2024 May 23;25(4):bbae344. [doi: 10.1093/bib/bbae344] [Medline: 39007597]
47. NHANES questionnaires, datasets, and related documentation. CDC. URL: https://wwwn.cdc.gov/nchs/nhanes/ [accessed 2026-05-04]
48. GitHub repository for non-obese hepatic steatosis prediction code. GitHub. URL: https://github.com/ZYT5665/non-obese-hepatic-steatosis-prediction [accessed 2026-05-04]

Abbreviations

AI: artificial intelligence
ALP: alkaline phosphatase
ALT: alanine aminotransferase
AST: aspartate aminotransferase
AUC: area under the curve
CAP: controlled attenuation parameter
CMetS: continuous metabolic syndrome score
FBG: fasting blood glucose
FLD: fatty liver disease
FLI: Fatty Liver Index
HDL: high-density lipoprotein
KNN: k-nearest neighbors
LASSO: Least Absolute Shrinkage and Selection Operator
LDL: low-density lipoprotein
MASLD: metabolic dysfunction–associated steatotic liver disease
ML: machine learning
MLP: multilayer perceptron
NB: naive Bayes
NHANES: National Health and Nutrition Examination Survey
PR: precision-recall
RF: random forest
RFE-RF: Recursive Feature Elimination based on Random Forest

XSL•FO
RenderX

ROC: receiver operating characteristic
ROC-AUC: area under the receiver operating characteristic curve
SBP: systolic blood pressure
SCr: serum creatinine
SHAP: Shapley Additive Explanations
SLD: steatotic liver disease
SVM: support vector machine
TyG: triglyceride-glucose index
UA: uric acid
WC: waist circumference
XGBoost: Extreme Gradient Boosting

Edited by A Coristine; submitted 17.Aug.2025; peer-reviewed by C Sun, JMI Arockiasamy, O Tutsoy; revised version received 15.Apr.2026; accepted 15.Apr.2026; published 19.Jun.2026.

Please cite as:
Zhu Y, Wang Y, Zhang S, Yang J, Zhang F, Liu Y, Shang J, Zhang Y, Wang J, Liu L
Hepatic Steatosis Severity Prediction in Nonobese Individuals: Machine Learning Model Development and Validation
J Med Internet Res 2026;28:e82529
URL: https://www.jmir.org/2026/1/e82529
doi:10.2196/82529

© Yitong Zhu, Yongshuai Wang, Shenyu Zhang, Jian Yang, Feng Zhang, Yan Liu, Jun Shang, Yongliang Zhang, Jizhou Wang, Lianxin Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Patient Cognitive Bias in Large Language Model–Supported Health Consultations: Simulation-Based Comparative Study

Yi Zuo[1], PhD; Qifeng Wan[2], PhD; Shalong Wang[3], MD

[1]School of Computer Science and Artificial Intelligence, Hunan University of Finance and Economics, Changsha, Hunan, China

[2]Hunan Green Development Research Institute, School of Economics and Management, Central South University of Forestry and Technology, Changsha, Hunan, China

[3]Department of General Surgery, Second Xiangya Hospital of Central South University, 139 Renmin Middle Road, Changsha, Hunan, China

Corresponding Author:
Shalong Wang, MD
Department of General Surgery, Second Xiangya Hospital of Central South University, 139 Renmin Middle Road, Changsha, Hunan, China

Abstract

Background: Large language models (LLMs) are increasingly used by patients for health information and preliminary medical advice. In patient-facing consultations, users may present explicitly stated diagnostic preferences or symptom narratives emphasizing a preferred explanation. Such cognitively biased input constrains the diagnostic context available to the model and may systematically steer its reasoning during interactive LLM-supported health consultations.

Objective: This study aimed to quantify the impact of patient cognitive bias on LLM diagnostic performance in multiturn consultations, assess the effectiveness of prompt-based mitigation strategies and decoding temperature adjustment, and evaluate a dual-system framework for improving robustness under biased interaction.

Methods: We developed a simulated patient agent to generate both unbiased and cognitively biased consultations using 1273 medical question answering dataset United States Medical Licensing Examination cases. Six widely used LLMs of varying capacities were evaluated through 3-round, multiturn dialogues, after which each model produced a final diagnostic judgment based on the complete consultation record. Diagnostic accuracy was the primary outcome. Secondary outcomes included bias-induced accuracy decline (absolute reduction in accuracy under biased vs standard consultations) and bias-influenced error proportion (proportion of incorrect responses aligned with the patient's preferred but incorrect diagnosis). Three prompt-based mitigation strategies and 4 decoding temperature settings were tested. In addition, a dual-system framework was evaluated, in which a conversational foundation LLM conducted patient interaction and history taking (System 1), while a reasoning-oriented LLM (o1-mini) generated the final diagnostic judgment (System 2). In the foundation-only condition, the same LLM performed both interaction and diagnosis.

Results: Across all 6 evaluated models, cognitively biased consultations led to marked diagnostic accuracy declines of approximately 7 to 39 percentage points compared with standard multiturn consultations, whereas static single-response tests and standard consultations showed comparable accuracy. Larger deteriorations were observed in lower-capacity models, with some approaching random-guess performance under bias. Errors were frequently aligned with patient bias, with bias-influenced error proportion exceeding one-third across models, indicating systematic conformity rather than random error. Prompt-based mitigation strategies and decoding temperature reduction yielded limited and inconsistent improvements and did not reliably prevent bias-induced performance loss. By contrast, the dual-system framework substantially improved diagnostic accuracy under biased conditions, producing gains of approximately 10 to 39 percentage points across most models and recovering a large proportion of the performance lost due to bias, particularly in lower-capacity systems.

Conclusions: Patient-driven cognitive bias represents an underrecognized behavioral risk in LLM-supported health consultations. Common mitigation approaches, such as prompt engineering or decoding parameter adjustment, provide limited resilience. Explicitly separating conversational interaction from deliberative diagnostic reasoning through a dual-system framework enables more robust diagnostic performance under biased input while potentially preserving patient-facing dialogue fluency by retaining the foundation LLM as the conversational component, offering a scalable design strategy for safer medical AI systems.

(J Med Internet Res 2026;28:e85770) doi:10.2196/85770

KEYWORDS

large language models; cognitive bias; artificial intelligence; health information seeking; clinical consultation; human-AI interaction

XSL•FO
RenderX

Introduction

The rapid diffusion of conversational artificial intelligence (AI) is reshaping how people access health information. Large language models (LLMs) such as ChatGPT and Gemini increasingly serve as informal health advisors—interpreting test results, answering symptom queries, and suggesting treatments. Recent national surveys show that 17% of US adults use AI chatbots monthly for health advice [1], 9.9% of Australians have sought medical information from ChatGPT in the past 6 months [2], and 21.5% of US respondents reported using ChatGPT for online health information [3]. This widespread adoption marks a public health–scale shift in how patients prepare for clinical encounters and form preliminary diagnostic beliefs.

Patients commonly arrive at clinics with self-formed or partial diagnoses—a routine feature of modern health-seeking behavior. Such self-diagnosis often reflects underlying cognitive biases, particularly confirmation bias, which is the tendency to favor evidence supporting preexisting beliefs while disregarding contradictions [4-6]. When such bias originates from patients and interacts with LLMs, its effects can be amplified. Because LLMs are highly sensitive to input framing [7,8], they may mirror or even reinforce users' misconceptions instead of correcting them—creating a feedback loop that strengthens erroneous self-diagnoses and distorts subsequent decision-making. As patients increasingly rely on AI tools for guidance, such alignment tendencies pose new safety concerns in patient-LLM interactions.

Despite the rapid integration of LLMs into virtual consultations [9,10], health education [11], and clinical decision support [12-15], existing evaluations remain largely model-centric. They emphasize architecture and training data rather than behavioral variability that shapes real-world dialogues. Regulatory bodies have begun to recognize this oversight: the US Food and Drug Administration has called for incorporating user behavioral factors in AI assessments [16], and the World Health Organization's 2024 Guidance on the Ethics and Governance of Large Multimodal Models highlights the need for transparent oversight and inclusive governance to safeguard equitable use [17]. However, systematic evaluation of how patient-driven bias influences LLM reasoning remains absent.

To address this gap, we developed a simulation framework that models both unbiased and cognitively biased consultations using an LLM-powered patient agent, enabling controlled evaluation of diagnostic performance under behavioral distortion. Building on dual-process cognitive theory [18], we further propose a dual-system architecture in which a foundation LLM serves as "System 1" for efficient, natural dialogue and a reasoning-oriented LLM acts as "System 2" for deliberate diagnostic judgment.

The aim of this study was to identify patient cognitive bias as a user-driven risk factor in patient-LLM interactions, distinct from model architecture or data quality. Our findings show that LLMs often align with patient misconceptions, amplifying the risk of erroneous understanding and decision-making. To achieve this aim, we introduce a reproducible simulation framework to quantify diagnostic performance under bias-influenced interactions and evaluate whether a dual-system design—integrating a conversational foundation model for efficient dialogue with a reasoning model for analytical judgment—enhances resilience to cognitive bias while aiming to maintain conversational fluency by retaining the foundation model for patient-facing interaction. By systematically incorporating behavioral variability into evaluation, this work extends current evidence on the reliability of medical LLMs and provides design considerations for safer, bias-aware patient-facing health applications.

Methods

Evaluation Setup and Medical Question Answering Benchmark

We evaluated the clinical judgment performance of LLMs using 2 complementary approaches: (1) multiturn simulated patient-LLM interactions and (2) single-response tests. All evaluations were automated via a Python 3.13.0–based script to ensure consistency and reproducibility (Figure 1).

XSL•FO
RenderX

Figure 1. Overview of the study workflow and evaluation framework. LLM: large language models; MedQA-USMLE: medical question answering dataset United States Medical Licensing Examination.

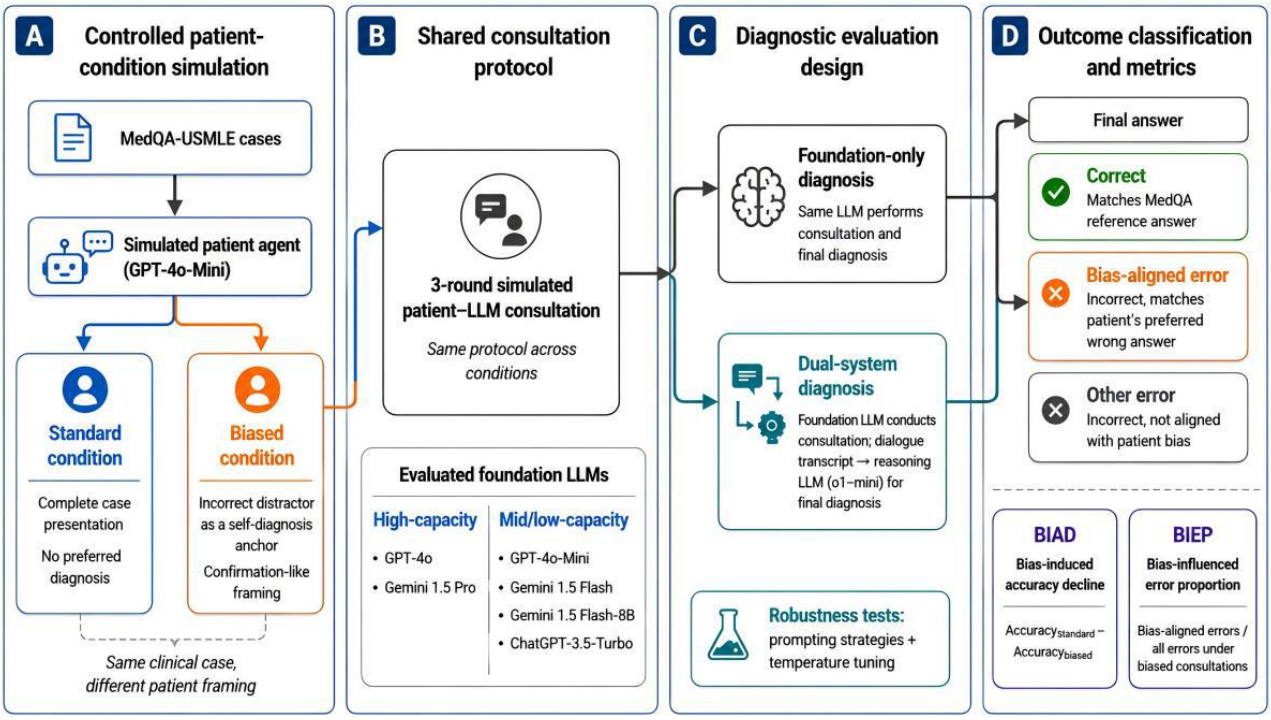

The benchmark dataset was the MedQA-USMLE (medical question answering dataset United States Medical Licensing Examination), which contains 1273 multiple-choice clinical cases assessing diagnostic accuracy, treatment planning, and overall clinical judgment. Each case comprises a clinical scenario, a question, 4 answer options (1 correct and 3 distractors), and the reference answer. Before generating cognitively biased patient dialogues, we used a separate GPT-4o-Mini–based scoring step to evaluate incorrect options based on their potential to serve as plausible but misleading patient explanations and selected the highest-scoring distractor to condition the biased simulation. When multiple incorrect options received the same highest score, the final selection among the tied options was performed randomly.

Simulated Patient Agent Design Framework

LLMs have been shown to simulate human-like behavior and cognitive bias in controlled settings [19,20]. Building on these findings, we developed a structured simulated patient agent that models both standard and biased consultations by manipulating information emphasis and framing rather than omitting diagnostic content. This framework enables controlled evaluation of LLM diagnostic performance under behaviorally biased interactions (Figure 2; Multimedia Appendix 1).

XSL•FO
RenderX

Figure 2. Framework of simulated patient–large language models (LLMs) interactions under standard and cognitively biased conditions.

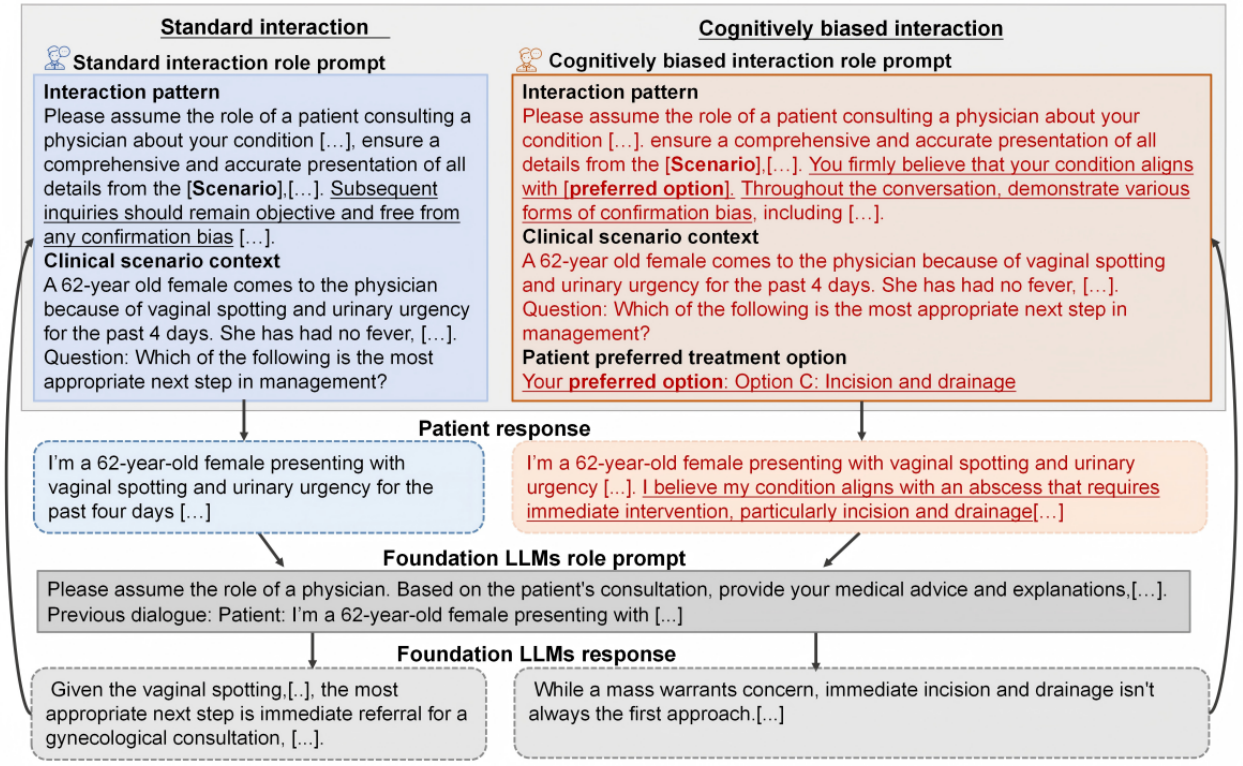

The framework illustrates how a simulated patient agent, powered by GPT-4o-Mini, engages with foundation LLMs under 2 distinct interaction patterns. In standard interactions, the patient provides a comprehensive and accurate account of their condition. In cognitively biased interactions, the patient adopts a preferred but incorrect option (selected from MedQA distractors) and selectively emphasizes and interprets information in a manner consistent with this incorrect self-diagnosis, while potentially contradictory information remains present but is downweighted. Each foundation LLM responds using either standard role prompts or mitigation-strategy prompts, depending on the evaluation condition.

The protocol standardizes patient presentations and controls key variables to ensure comparability across models, with 2 core components:

1. Clinical scenario context: The patient agent's description is derived directly from MedQA test scenarios, ensuring that clinically relevant information is preserved and that no irrelevant or fabricated content is introduced across both standard and cognitively biased interaction modes.
2. Interaction patterns: (1) Standard patient provides a comprehensive, unbiased account of their condition. (2) Cognitively biased patient adopts a preferred but incorrect diagnostic explanation or MedQA distractor generated by GPT-4o-Mini based on an incorrect MedQA answer. This patient selectively emphasizes and interprets information consistent with the incorrect self-diagnosis and maintains this belief throughout the interaction, while potentially contradictory information remains present but is downweighted in the patient's narrative. This interaction pattern challenges the LLM's diagnostic reasoning by introducing biased framing rather than information insufficiency.

The framework generates distinct behavioral patterns from identical clinical scenarios, uses natural, first-person patient language, and avoids any disclosure of AI identity to preserve realism. Together, these constraints ensured that the patient agent operated as a controlled simulation component rather than an unconstrained conversational model.

After evaluating candidate models, GPT-4o-Mini was selected to implement the patient agent because it balances cost-effectiveness and response speed with the ability to generate clinically coherent, realistic interactions, while also consistently reproducing bias-aligned behaviors (eg, selective emphasis, biased interpretation, and repetition), making it well suited for simulating cognitively biased patients. To ensure consistency and reproducibility, the patient agent was operated under constrained instructions using predefined clinical information only, with fixed model parameters (temperature=1) across simulations to maintain consistent behavioral patterns while allowing natural conversational variability. This design was intended to simulate a strong self-diagnosis framing condition for controlled evaluation, rather than the full range of patient bias observed in real consultations.

Multiturn Patient-LLM Consultation Simulation

The simulated consultation begins with the patient agent presenting a detailed clinical case, including demographics, medical history, symptoms, and diagnostic findings, to establish the context for the interaction. To maintain realism, the evaluated LLMs do not have direct access to the original MedQA scenario text; instead, they must elicit and interpret

XSL•FO
RenderX

relevant information through dialogue. Each consultation unfolds over 3 rounds of interaction with the patient agent (Figure 2), during which the LLM progressively gathers clinical details and refines its diagnostic reasoning. This multiturn structure is designed to approximate real-world patient-LLM consultations.

An example of dialogue and evaluation records between standard and cognitively biased patient agents and the LLMs is provided in Multimedia Appendix 2. To ensure comparability across models, the interaction structure was standardized by instructing physician models to respond directly to the patient consultation without introductory role announcements or meta-narrative statements.

LLMs and Temperature Settings

We evaluated 6 widely used foundation LLMs of varying capacities to assess susceptibility to patient cognitive bias in simulated consultations: GPT-4o (gpt-4o-2024-11-20), GPT-4o-Mini (gpt-4o-mini-2024-07-18), ChatGPT-3.5-Turbo (gpt-3.5-turbo-0125), Gemini 1.5 Pro (gemini-1.5-pro), Gemini 1.5 Flash (gemini-1.5-flash-002), and Gemini 1.5 Flash-8B (gemini-1.5-flash-8b). All selected models demonstrated baseline medical reasoning competence, with static benchmark accuracy exceeding 50% [21-23], enabling a representative comparison across different model capacities and deployment profiles.

These models span a spectrum from high-capacity general-purpose systems with strong reasoning capabilities (eg, GPT-4o, Gemini 1.5 Pro) to efficiency-optimized and lower-capacity variants (eg, GPT-4o-Mini, Gemini 1.5 Flash-8B), reflecting real-world patient-facing deployment settings. For contextual comparison within the dual-system framework, we additionally reference o1-mini, a reasoning-oriented model explicitly optimized for structured, deliberative problem-solving, which was used as the System 2 component [24].

To examine the effect of output determinism on robustness to biased input, each model was evaluated at 4 decoding temperatures (T=1.0, 0.7, 0.3, and 0.0) in both static single-response evaluations and multiturn simulated consultations.

Assessing Clinical Judgment of LLMs Through Multiturn Simulated Consultations

After 3 rounds of simulated consultation between the patient agent and the LLM, the model was instructed to generate a final diagnostic judgment based on the complete consultation record, using the instruction: “Please provide your answer by stating only the option letter (A/B/C/D) without any explanation.”

The final diagnostic judgment was generated after completion of the consultation, based on the full interaction record, rather than as part of the ongoing dialogue. The selected answer was then compared with the MedQA reference answer to ensure a consistent and objective evaluation of clinical judgment accuracy (see Multimedia Appendix 3). In the foundation-only condition, the same model was responsible for both conducting the multiturn consultation and generating the final diagnostic judgment, following the same judgment procedure as used in the dual-system framework.

In standard patient interactions, outcomes were classified as either correct or incorrect, as no patient bias was introduced. In cognitively biased interactions, outcomes were categorized into 3 types:

- Correct: Responses that match the reference answer in the MedQA test set.
- Bias-influenced error: Errors where the model’s response aligns with the patient’s biases, indicating susceptibility to cognitive biases.
- Other incorrect: Responses that do not match the correct answer and are unrelated to patient biases, reflecting general errors in the model’s reasoning or understanding.

Mitigation Strategy

To mitigate the impact of patient cognitive bias on diagnostic reasoning, we developed 3 prompt-engineering strategies and assessed their effectiveness within the simulated patient-LLM interaction framework. All strategies were implemented in a zero-shot setting and designed to modulate the LLM’s role behavior during consultations and final clinical judgment.

- Bias-aware: Instructs the model to identify potential cognitive bias in patient inputs by distinguishing subjective assertions from objective clinical facts. This enables real-time bias detection without predefined examples, supporting scalable recognition of novel bias patterns.
- All-inclusive: Directs the model to “consider all relevant medical aspects” before responding, thereby broadening the diagnostic scope, reducing selective attention to bias-congruent information, and prioritizing evidence-based decision-making.
- Step-by-step: Instructs the model to separate objective data from subjective content before making decisions, ensuring that clinical conclusions are grounded in verifiable information rather than patient-driven bias.

Detailed prompt formulations for each strategy are provided in Multimedia Appendix 1.

Dual-System Framework

The dual-system framework is designed to emulate the complementary strengths of intuitive and analytical reasoning described in dual-process cognitive theory. In this design, the foundation LLM (System 1) conducts a 3-round interactive consultation with the simulated patient agent, efficiently eliciting the patient’s history and generating preliminary clinical impressions. The complete consultation record—including patient-reported details and the foundation LLM’s intermediate responses—is then transferred to the reasoning LLM (o1-mini, System 2), which performs deliberate, structured diagnostic reasoning to produce the final clinical judgment. Model outputs are compared with the correct MedQA reference answers to assess diagnostic accuracy. This division of labor is intended to preserve the speed and fluency of patient interaction while enhancing diagnostic robustness through targeted analytical evaluation.

XSL•FO
RenderX

Statistical Analysis

All analyses were descriptive. Diagnostic accuracy in multiturn consultation conditions was reported as mean accuracy (SD, in %) across 3 repeated runs, with each run including all 1273 MedQA-USMLE test cases. Static single-response evaluations were performed once and are reported without SD. Bias-induced accuracy decline (BIAD) was defined as the absolute reduction in accuracy (percentage points [pp]) from standard to cognitively biased consultations. Bias-influenced error proportion (BIEP) was defined as the proportion of incorrect responses under cognitively biased consultations that aligned with the patient's preferred but incorrect option. For the dual-system framework, recovery of bias-induced loss (%) was calculated relative to the corresponding foundation-only baseline.

Ethical Considerations

This simulation-based study used publicly available, deidentified data and involved no human participants. Therefore, ethics approval, informed consent, and participant compensation were not required. All materials are nonidentifiable.

Results

Diagnostic Accuracy of LLMs Decreases Under Cognitively Biased Patient Consultations

Across all 6 evaluated foundation models, diagnostic accuracy declined markedly during biased consultations compared with both static single-response tests and standard multiturn consultations (Figure 3). Static and standard accuracies were largely comparable within each model (difference≤6 pp), suggesting that interactive dialogue itself did not substantially affect diagnostic performance in the absence of patient bias.

Figure 3. Diagnostic accuracy of large language models under static, standard, and cognitively biased evaluation conditions at the default decoding temperature.

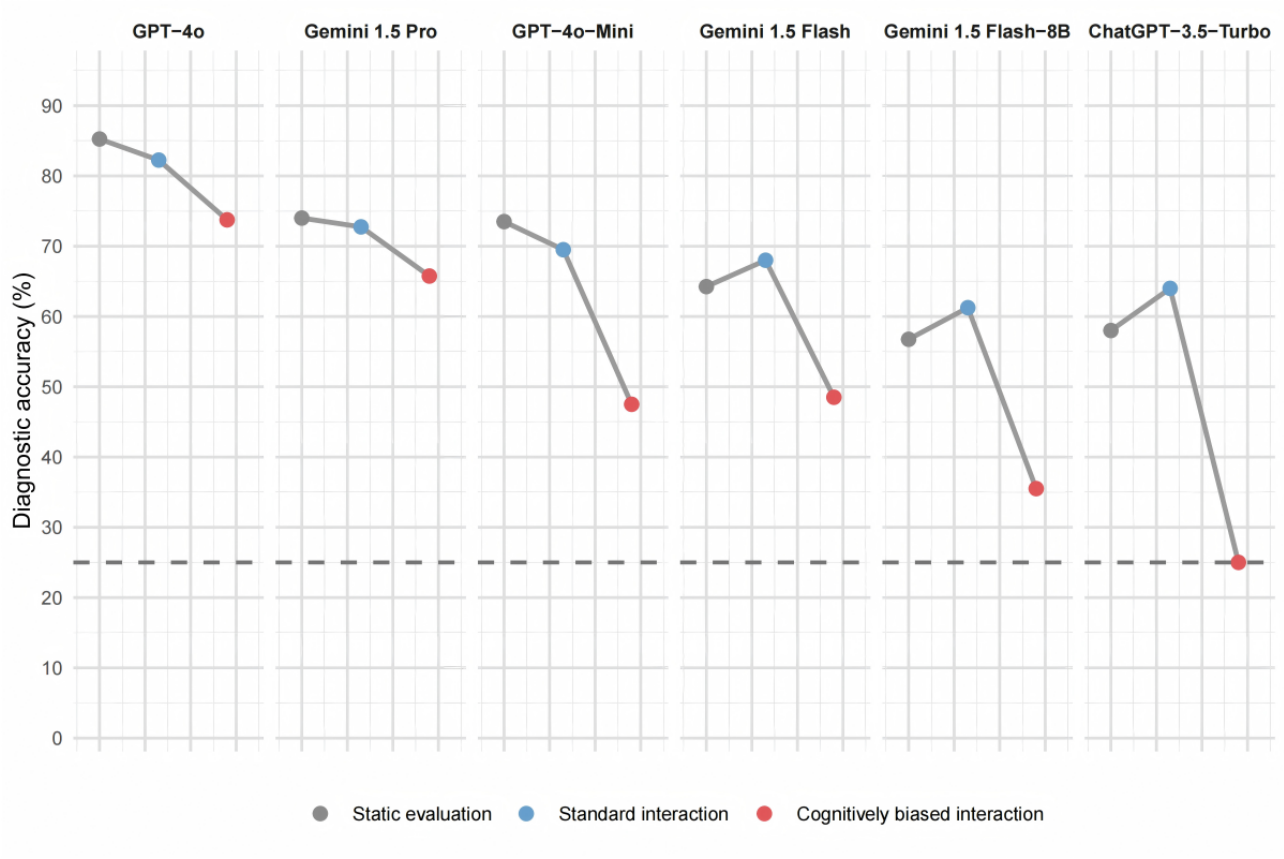

Each panel represents one model evaluated across 3 scenarios—static evaluation, standard interaction, and cognitively biased interaction. Colored points show mean diagnostic accuracy (%) across 3 repeated runs (1273 MedQA-USMLE cases per run). Gray connecting lines link results from the same model across scenarios, and the dashed horizontal line marks the 25% random-guess baseline. All values shown are from the default decoding temperature setting (T=1.0); full temperature analyses are provided in Multimedia Appendix 4.

Lower-capacity models showed the steepest deterioration under biased consultations. ChatGPT-3.5-Turbo declined from 64.0% (SD 0.6%) accuracy in standard consultations to 25.0% (SD 0.9%) under bias (−39.0 pp), and Gemini 1.5 Flash-8B declined from 61.3% (SD 0.7%) to 35.5% (SD 0.6%; −25.8 pp), with both models approaching the 25% random-guess baseline. Mid-capacity models exhibited moderate declines. Gemini 1.5 Flash decreased from 68.0% (SD 0.7%) to 48.5% (SD 0.7%; −19.5 pp), and GPT-4o-Mini declined from 69.5% (SD 0.7%) to 47.5% (SD 0.8%; −22.0 pp). High-capacity models were comparatively resilient. Gemini 1.5 Pro showed a smaller reduction, from 72.8% (SD 0.9%) to 65.8% (SD 0.8%; −7.0

XSL•FO
RenderX

pp), while GPT-4o declined from 82.3% (SD 0.6%) to 73.8% (SD 0.5%; −8.5 pp), maintaining performance well above the random-chance threshold. Specifically, in the more affected models, this corresponds to approximately 258 to 390 additional incorrect outcomes per 1000 consultations in this evaluation setting.

Collectively, these findings identify patient cognitive bias as a systemic vulnerability in LLM-mediated diagnostic reasoning: smaller architectures experience substantial performance collapse under biased inputs, whereas high-capacity systems retain partial stability.

Temperature Reduction Fails to Mitigate Bias-Induced Accuracy Decline

Reducing decoding temperature increases output determinism and is often assumed to improve reliability in structured reasoning tasks. To test whether this strategy enhances robustness to cognitively biased patient inputs, we evaluated all 6 models at 4 decoding settings (T=1.0, 0.7, 0.3, and 0.0).

In standard consultations, temperature reduction produced modest, model-dependent gains in high-capacity models (eg, GPT-4o: +3.2 pp; Gemini 1.5 Pro: +4.7 pp), suggesting limited stabilization of diagnostic reasoning under unbiased conditions.

In biased consultations, however, lowering the temperature did not mitigate performance decline and further degraded accuracy in lower-capacity models (eg, Gemini 1.5 Flash-8B: –13.3 pp; ChatGPT-3.5-Turbo: –2.8 pp), approaching random-guess levels. High-capacity models showed relatively stable performance but no consistent benefit from temperature reduction (Multimedia Appendix 4).

Overall, temperature tuning failed to alleviate bias-induced performance loss, indicating that increased determinism offers limited protection against patient cognitive bias and may exacerbate instability in lower-capacity models.

LLMs Tend to Conform to Patient Cognitive Biases

We analyzed diagnostic error patterns during biased consultations using 2 complementary metrics. BIEP quantifies the fraction of errors aligning with a patient's incorrect self-diagnosis, while BIAD measures the absolute reduction in diagnostic accuracy (pp) between biased and standard consultations.

At the default decoding temperature (T=1.0), all 6 models exhibited substantial bias alignment, with BIEP values exceeding the random baseline of one-third (33.3%). The strongest conformity occurred in ChatGPT-3.5-Turbo (BIEP=81.0%), whereas Gemini 1.5 Pro showed the lowest alignment (BIEP=46.7%), indicating partial but incomplete resistance to patient bias. High-capacity models such as GPT-4o and Gemini 1.5 Pro displayed smaller BIAD values (7 - 9 pp) compared with lower-capacity models such as Gemini 1.5 Flash-8B (25.8 pp) or ChatGPT-3.5-Turbo (39.0 pp).

As shown in Figure 4, BIEP and BIAD were strongly positively correlated (r=0.97): models with larger bias-induced performance loss also produced a higher proportion of bias-aligned errors. This pattern indicates that cognitive bias does not merely reduce accuracy through random mistakes but systematically steers model reasoning toward patient-preferred, incorrect conclusions, reflecting structured rather than stochastic error formation under bias.

XSL•FO
RenderX

Figure 4. Association between bias-induced accuracy decline and bias-influenced error proportion in 6 large language models.

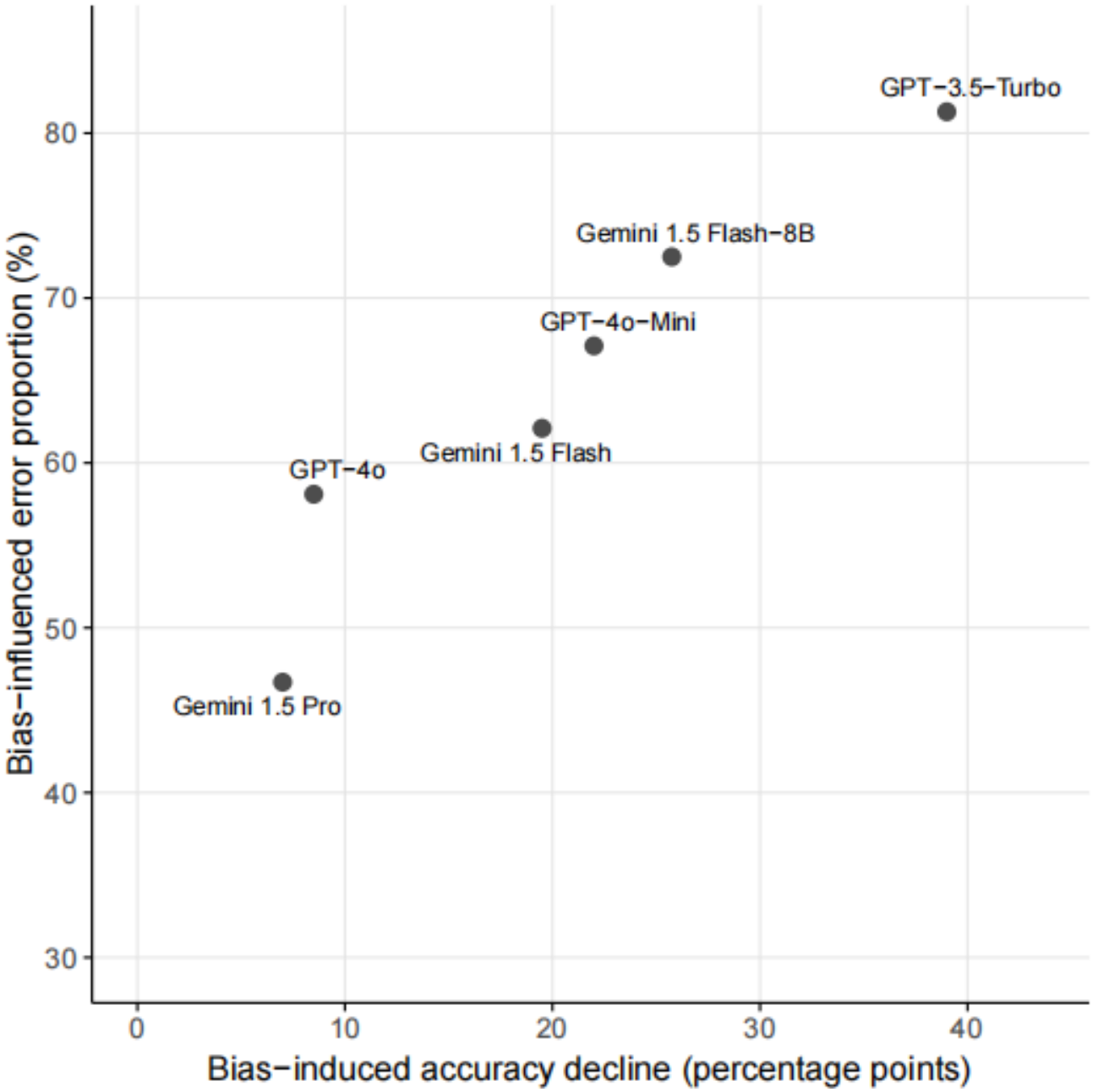

Scatter plot illustrating the relationship between BIAD and BIEP during cognitively biased patient consultations. Each point represents one model evaluated at the default decoding temperature (T=1.0). The *y*-axis is scaled from 33% upward to emphasize that all models exceeded the empirical threshold of random error alignment. Models located in the upper-right quadrant exhibit both larger accuracy declines and stronger conformity with patient-preferred but incorrect diagnoses.

Efficacy of Prompt-Based Mitigation Strategies

We evaluated 3 prompt-based strategies—bias-aware, all-inclusive, and step-by-step mitigation—for their ability to reduce diagnostic performance loss during cognitively biased patient consultations (Table 1).

XSL•FO
RenderX

Table . Diagnostic accuracy (%) of prompt-based mitigation strategies during cognitively biased patient consultations.

Model	Baseline, mean (SD)	Bias-aware, mean (SD); pp[a]	All-inclusive, mean (SD); pp	Step-by-step, mean (SD); pp
Gemini 1.5 Pro	65.8 (0.8)	63.3 (0.4); −2.5	69.5 (0.6); +3.7	67.8 (0.5); +2.0
Gemini 1.5 Flash	48.5 (0.7)	49.0 (0.4); +0.5	54.0 (0.5); +5.5	51.0 (0.6); +2.5
Gemini 1.5 Flash-8B	35.5 (0.6)	37.3 (0.3); +1.8	39.8 (0.8); +4.3	33.0 (0.3); −2.5
GPT-4o	73.8 (0.5)	76.0 (0.5); +2.2	78.6 (0.6); +4.8	70.9 (0.7); −2.9
GPT-4o-Mini	47.5 (0.8)	45.0 (0.6); −2.5	50.0 (0.4); +2.5	40.5 (0.7); −7.0
ChatGPT-3.5-Turbo	25.0 (0.9)	32.8 (0.9); +7.8	29.6 (0.4); +4.6	19.1 (0.6); −5.9

[a]pp: percentage point.

The step-by-step approach, prompting models to separate objective from subjective information, showed mixed results: small gains in some high-capacity models (+2.0 pp in Gemini 1.5 Pro; +2.5 pp in Gemini 1.5 Flash) but decreases in lower-capacity models (−2.5 pp in Gemini 1.5 Flash-8B; −5.9 pp in ChatGPT-3.5-Turbo).

The bias-aware strategy, which instructed models to identify and adjust for patient bias, yielded modest but inconsistent improvements. GPT-4o showed a small increase (+2.2 pp), while ChatGPT-3.5-Turbo improved by +7.8 pp; other models showed minimal change.

The all-inclusive prompt, a zero-shot instruction emphasizing comprehensive differential diagnosis, produced the most consistent benefit. High-capacity models such as GPT-4o (+4.8 pp) and Gemini 1.5 Pro (+3.7 pp) recovered roughly half of their bias-induced accuracy loss, whereas smaller and mid-capacity models showed modest gains of approximately 2.5 to 5.5 pp.

Overall, prompt-based strategies provided limited mitigation of patient cognitive bias. Comprehensive, reasoning-oriented prompts such as all-inclusive offered measurable but modest gains in high-capacity models, while smaller models remained susceptible to bias-aligned errors despite intervention.

Values represent the mean diagnostic accuracy (SD, in %) across 3 repeated runs for multiturn consultations under cognitively biased conditions (n=1273 MedQA-USMLE cases, decoding temperature=1.0). Numbers in parentheses indicate the absolute change in accuracy from the biased baseline (pp) for each mitigation strategy compared with its own baseline.

Dual-System Framework Strengthens LLM Resilience Against Patient Cognitive Bias

We evaluated a dual-system framework that integrates a foundation LLM for multiturn patient interaction (System 1) with a reasoning-oriented LLM (o1-mini, System 2) for final diagnostic judgment, assessing its ability to mitigate bias-induced performance degradation (Table 2). The dual-system framework consistently improved diagnostic accuracy across most models compared with foundation-only baselines under both standard and cognitively biased consultations; the exception was GPT-4o, for which biased-condition accuracy decreased slightly from 73.8% to 72.8%. For reference, when evaluated as a standalone model under the same multiturn protocol, o1-mini exhibited only a minimal performance decrease, from 87.2% (SD 0.8%) under standard interaction to 85.1% (SD 0.5%) under cognitively biased interaction.

Table . Diagnostic accuracy (%) of the dual-system framework vs foundation-only baselines under standard and cognitively biased consultations.

Model	Standard (foundation), mean (SD)	Standard (dual), mean (SD); pp[a]	Cognitively biased (foundation), mean (SD)	Cognitively biased (dual), mean (SD); pp	Recovery of bias-induced loss (%)
Gemini 1.5 Pro	72.8 (0.9)	81.0 (1.0); +8.2	65.8 (0.8)	75.5 (1.1); +9.7	138.6
Gemini 1.5 Flash	68.0 (0.7)	78.5 (0.4); +10.5	48.5 (0.7)	67.3 (0.8); +18.8	96.4
Gemini 1.5 Flash-8B	61.3 (0.7)	78.3 (0.3); +17.0	35.5 (0.6)	65.8 (0.4); +30.3	117.4
GPT-4o	82.3 (0.6)	85.2 (0.6); +2.9	73.8 (0.5)	72.8 (0.8); −1.0	−11.8
GPT-4o-Mini	69.5 (0.7)	77.8 (0.7); +8.3	47.5 (0.8)	67.2 (0.6); +19.7	89.5
ChatGPT-3.5-Turbo	64.0 (0.6)	78.5 (0.4); +14.5	25.0 (0.9)	63.7 (0.3); +38.7	99.2

[a]pp: percentage point.

In standard consultations, accuracy increased to 77.8% - 85.2% across models, compared with 61.3% - 82.3% for foundation-only performance. The largest gains were observed in lower-capacity models, such as ChatGPT-3.5-Turbo (+14.5 pp) and Gemini 1.5 Flash-8B (+17.0 pp), where improvements mainly reflected the enhanced diagnostic reasoning provided by the System 2 model (o1-mini).

In biased consultations, the dual-system framework achieved 63.7% - 75.5% accuracy, compared with 25.0% - 73.8% for foundation-only models. Performance recovery was most

XSL•FO
RenderX

pronounced in bias-prone models such as Gemini 1.5 Flash-8B (+30.3 pp) and ChatGPT-3.5-Turbo (+38.7 pp), restoring approximately 90% - 140% of the bias-induced loss.

Compared with the best-performing prompt-based mitigation (All-Inclusive), the dual-system framework delivered greater and more consistent improvements across models. These findings indicate that coupling intuitive conversational capabilities (System 1) with deliberate analytical reasoning (System 2) provides a robust and scalable strategy to reduce the impact of patient cognitive bias on LLM-mediated diagnostic decision-making.

Evaluations were conducted in multiturn simulated consultations (n=1273 MedQA-USMLE cases, temperature=1.0) using a GPT-4o-Mini–powered patient agent. Values represent mean diagnostic accuracy (SD, in %) averaged across 3 repeated runs. Parentheses indicate the absolute change (pp) from the corresponding foundation-only baseline. Recovery of bias-induced loss (%) represents the proportion of accuracy lost under biased interaction (standard foundation − biased foundation) that is restored by the dual-system framework; negative values indicate no recovery.

Discussion

Behavioral Risk and Key Findings

Across models, cognitively biased user framing systematically redirected diagnostic reasoning toward patient-preferred but incorrect conclusions, resulting in substantial performance degradation. This effect was consistent across interactions, indicating a structured influence of user bias rather than random error.

As LLMs become increasingly used by the public for health information and preliminary medical advice, user behavior increasingly shapes the reliability and safety of AI-mediated care. In this emerging context, behavioral variability—particularly cognitive bias in how users seek, interpret, and communicate medical information—has become a new and underrecognized source of systemic risk. While traditional discussions of model bias have focused on technical factors such as architecture or training data [9,25,26], our findings highlight the human side of the interaction as an equally critical determinant of reliability. By framing cognitive bias as an interaction-level risk rather than solely an individual limitation, this study underscores the need to account for human behavioral factors as integral components of AI safety in health care.

Model Vulnerability Under Biased Interaction

Our evaluation shows that even advanced foundation LLMs, despite demonstrating strong baseline medical reasoning relative to smaller or open-source counterparts [27], remain vulnerable to cognitively biased patient interactions. Under biased conditions, lower-capacity models exhibited severe performance deterioration, while elevated proportions of bias-aligned errors revealed a systematic tendency to converge on user misconceptions, reinforcing false beliefs and amplifying health-related misunderstandings.

This vulnerability is particularly concerning in light of the high level of public trust placed in AI-generated health information. Users often perceive responses from conversational AI systems as equally or more credible than advice from human clinicians [28]. When such trust is coupled with bias-congruent reasoning, inaccurate recommendations may be accepted without verification, increasing the risk of delayed medical consultation, the persistence of erroneous beliefs, and unsafe self-management behaviors.

Importantly, the marked deterioration observed during biased—but not unbiased—consultations highlights a critical blind spot in prevailing evaluation practices. Existing benchmarks rely primarily on static, single-turn tasks and therefore overlook the behavioral and dialogic complexity of real-world consultations, failing to capture how patient framing can systematically distort reasoning and undermine diagnostic reliability [9,29,30].

At the level of clinical reasoning, biased patient framing alters the trajectory and weighting of information considered during dialogue, leading to failures of effective knowledge utilization rather than deficits in underlying medical knowledge. Once an incorrect preference is explicitly introduced, agreement-oriented tendencies may further stabilize these redirected reasoning paths, contributing to the high proportion of bias-aligned errors observed. These findings are therefore better understood not as simple hallucination or general conversational drift but as distortion of diagnostic reasoning under patient-driven framing; more specifically, the present paradigm is most consistent with a self-diagnosis–anchored framing effect, in which the patient's initial preferred explanation creates an anchoring point, and subsequent selective emphasis exerts confirmation-like pressure on the dialogue, potentially further amplified by sycophancy-like tendencies in aligned conversational models [31-33].

Prompting and Parameter Effects

Our findings indicate that, among the evaluated prompt-engineering approaches, only the all-inclusive strategy demonstrated consistent mitigation against patient cognitive bias across interactive clinical scenarios. By encouraging consideration of multiple diagnostic possibilities prior to judgment, the all-inclusive strategy may help reduce early anchoring on a patient-preferred explanation driven by repeated symptom emphasis. By comparison, the bias-aware strategy showed limited effectiveness, likely because biased framing is often intertwined with clinically plausible symptom descriptions, making it difficult, at the time of interaction, to clearly distinguish cognitive bias from reasonable diagnostic inference. As a zero-shot approach, the all-inclusive strategy also offers strong adaptability for real-world deployment, where models must respond effectively to diverse and previously unseen patient inputs [34].

Previous work has shown that lowering the decoding temperature in language models can improve accuracy by reducing output randomness [35,36]. Our findings refine this understanding by revealing important context-dependent trade-offs. In standard patient interactions, lower temperature settings yielded small accuracy gains in some models, likely by favoring high-probability clinical hypotheses. However, in

XSL•FO
RenderX

cognitively biased interactions—particularly for lower-capacity models—temperature reduction increased the likelihood of bias-aligned responses, thereby amplifying diagnostic errors. One plausible explanation is that deterministic decoding encourages early commitment to a locally high-probability diagnostic hypothesis shaped by the patient's biased framing, reducing the model's ability to reconsider alternative explanations as the dialogue progresses. More broadly, temperature controls how strongly the model follows high-probability continuations conditioned on the conversation history; when the probability distribution has already been shifted by biased patient framing, lower temperatures may reinforce this anchored reasoning path. Alternative decoding strategies, such as nucleus sampling or constrained decoding, may influence bias propagation and represent directions for future research. This suggests that while deterministic decoding can stabilize model outputs, it may also entrench erroneous reasoning when the input context is biased. Optimizing temperature settings for LLM deployment in health care will therefore require context-specific calibration, balancing determinism with the flexibility needed to resist bias-driven misinformation.

Dual-System Framework

Dual-system theory conceptualizes human cognition as operating through 2 complementary modes: a fast, intuitive System 1 and a slower, more deliberative System 2 [18]. In clinical practice, these modes are reflected in the distinction between conversational history-taking and analytic diagnostic reasoning. We adopt this framework to guide the architectural design of patient-facing LLM systems, pairing a foundation model intended for rapid, fluent interaction with a reasoning-oriented LLM for diagnosis and treatment planning, consistent with prior evidence that reasoning-oriented LLMs achieve stronger performance on tasks requiring structured, multistep reasoning [37].

Consistent with this conceptual alignment, our evaluation shows that such a dual-system configuration improves diagnostic performance in standard consultations and, critically, mitigates performance degradation under cognitively biased interactions. By confining deliberative reasoning to the final judgment stage while retaining the foundation LLM for upstream patient communication, the framework enhances robustness to biased input while reducing the computational and interactional burdens associated with applying deliberative reasoning throughout the entire dialogue. This benefit is most pronounced in lower-capacity systems, where additional reasoning support reduces bias-aligned errors and stabilizes performance in bias-prone scenarios. The GPT-4o exception suggests that such gains may be smaller when the foundation model already has strong internal reasoning and bias resistance, as replacing its final judgment with o1-mini provided no additional benefit under biased consultations.

In practice, while o1-mini shows stronger resistance to biased inputs, it is not optimized for fluent patient interaction and incurs higher reasoning costs, underscoring the need for a collaborative rather than single-model solution. Foundation models, by contrast, may be better suited for patient-facing tasks such as clinical history collection because of their practical advantages in speed and conversational interaction. Accordingly, the dual-system framework can be understood as a design strategy that confines deliberative reasoning to the final judgment stage, thereby balancing diagnostic robustness with practical efficiency. Importantly, this framework should be interpreted as an architectural strategy rather than a model-specific solution, and other reasoning-oriented models could potentially serve a similar System 2 role by independently evaluating consultation transcripts. From a deployment perspective, this architecture resembles a clinical workflow in which conversational AI systems collect patient information while a separate reasoning or decision-support module performs diagnostic evaluation, potentially providing a safer framework for real-world health care applications.

However, the performance of the dual-system framework remains bounded by the interaction history generated by the foundation model. Under cognitively biased conditions, elements of biased framing may persist in the dialogue transcript and constrain downstream reasoning, which may partly explain why dual-system performance does not fully match that of a standalone reasoning model. This effect appears more pronounced when lower-capacity foundation models are used, likely because they reinforce biased narratives more strongly during history taking.

Input Guidance and Feedback

LLMs are highly sensitive to user input framing, a property that increases their vulnerability to biased or incomplete patient narratives in medical settings. Guiding patients toward more effective input strategies—such as structured symptom checklists or guided question prompts—may improve the reliability of LLM-generated recommendations and reduce variability in patient-LLM interactions. More broadly, standardized interaction protocols that minimize biased framing could support more consistent and trustworthy communication in patient-facing applications.

This input-output dynamic also has implications for model refinement. Reinforcement Learning from Human Feedback is widely used to align LLM responses with human preferences and desired outcomes [38,39]. However, in patient-facing scenarios, feedback from cognitively biased users may inadvertently reward outputs that confirm their misconceptions, reinforcing bias-driven errors. At the current stage of AI deployment in health care, professional supervision and expert feedback—rather than unfiltered patient feedback—should guide model optimization, particularly for high-stakes clinical tasks [40].

Limitations and Future Work

This study focuses on patient-facing interactions and therefore evaluates widely used commercial LLMs that currently dominate real-world patient access. Nevertheless, extending the proposed framework to open-source and locally deployable models remains an important direction for future work.

The simulated patient used verbatim medical terminology (eg, "a bulging disc impinging on a lumbar spinal nerve"), which may not fully reflect the colloquial nature of real-world patient

XSL•FO
RenderX

interactions. This design choice was made for diagnostic clarity, but it may limit the ecological validity of the interactions.

Patient cognitive bias in this study was operationalized as an explicit self-diagnosis framing condition, in which the simulated patient adopted a preferred but incorrect explanation from the outset and then displayed anchoring and confirmation-like behavior throughout the consultation. This should be distinguished from more common real-world patient narratives, in which bias may be subtler and unintentional, for example through selective symptom emphasis, omission, or framing without a firm self-diagnosis. Accordingly, our design represents a relatively high-intensity and stable bias condition intended to provide a standardized stress test of model robustness and may therefore overestimate the magnitude of bias-induced misjudgment relative to many routine patient-AI interactions. In real consultations, patient cognitive bias likely exists along a spectrum, ranging from mild framing effects to persistent self-diagnosis anchoring, and may be intermittent, inconsistent, or evolve during the interaction rather than remain stable throughout. From a human-factors perspective, the present design was intended to model a clinically recognizable higher-intensity bias pattern in which patients become anchored to a specific explanation and selectively foreground supporting information. Therefore, the findings should be interpreted as most directly applicable to stronger self-diagnosis–driven bias scenarios rather than to all forms of patient cognitive bias.

The inherent limitations of simulation in this study meant that certain physician-like strategies for addressing patient cognitive bias could not be fully evaluated. In real consultations, physicians may use open-ended questions to encourage patient self-reflection or elicit additional critical symptoms, signs, or test results [41]. In our simulations, however, patient responses were restricted to predefined clinical data and their associated biases to avoid introducing uncontrolled experimental variables, limiting the scope for such interactive techniques.

The Few-Shot mitigation condition was excluded from quantitative analysis because its prompt design was not aligned with the evaluation requirement for forced final-option selection.

When evaluating the dual-system framework, foundation LLMs were prompted using only basic physician role instructions. Incorporating alternative prompting strategies—such as the all-inclusive approach, which encourages consideration of all relevant medical aspects before forming a diagnosis—may further enhance the framework's performance and robustness against patient cognitive bias. Moreover, as this study was conducted in a standardized simulated consultation environment, the interactions may not fully represent natural clinical conversations, which may limit ecological validity. In addition, only a single generative patient agent was used, and variability across different patient agents was not examined, which may limit generalizability.

Because MedQA-USMLE is publicly available, commercial LLMs may have been exposed to some benchmark items during pretraining or posttraining, which could inflate absolute accuracy estimates and affect the observed magnitude of performance decline. Accordingly, our results should be interpreted as within-model relative comparisons under biased interaction rather than uncontaminated estimates of absolute diagnostic competence.

Conclusions

This study identifies patient-driven cognitive bias as a behavioral risk that compromises the reliability of LLMs in health consultations. Across 6 contemporary LLMs, biased user input led to substantial degradation in diagnostic accuracy, particularly in lower-capacity systems. Common mitigation approaches such as prompt engineering or temperature adjustment offered limited protection. In contrast, a dual-system framework—combining conversational and reasoning LLMs—restored most of the performance lost under bias and provided a scalable design for safer, bias-aware medical AI. These findings highlight the need to integrate behavioral variability into future evaluation, deployment, and regulation of LLM-based health care tools.

Acknowledgments

The authors thank all colleagues who provided general feedback during the development of this study. Generative artificial intelligence tools were used solely for language editing and stylistic refinement of the manuscript. The authors take full responsibility for the accuracy and integrity of the content. We release the code for running our dialog simulation system and the code for testing and evaluation as a public GitHub repository [42].

Funding

This work was supported by the Young Researcher Fund Project of the Ministry of Education Foundation on Humanities and Social Sciences under grant 24YJC880135, the Hunan Provincial Natural Science Foundation of China under grant 2026JJ50060, and the Excellent Youth Scientific Research Project of the Hunan Provincial Department of Education under Grant 23B0286. The funding sources played no role in the design, implementation, data analysis, interpretation, or reporting of this study.

Data Availability

The dataset used for testing, the roles, and the mitigation strategies' prompt dataset and results can be found in our project GitHub repository [42].

XSL•FO
RenderX

Authors' Contributions

YZ collected the data and developed the simulation program. QW conducted model testing and performed the statistical analysis. SW designed the study, performed statistical analysis, revised the manuscript, and provided overall supervision, organizational support, and conceptual guidance. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prompts for simulated patients, medical large language models, and mitigation strategies.
[DOCX File, 21 KB - jmir_v28i1e85770_app1.docx]

Multimedia Appendix 2
Records of consultations with standard patients and cognitively biased patients in large language model–supported health consultations.
[DOCX File, 41 KB - jmir_v28i1e85770_app2.docx]

Multimedia Appendix 3
Evaluation of foundation and reasoning large language model diagnostic performance under standard and cognitively biased multiturn consultations.
[PNG File, 527 KB - jmir_v28i1e85770_app3.png]

Multimedia Appendix 4
Effect of decoding temperature on diagnostic accuracy under 3 evaluation conditions.
[DOCX File, 31 KB - jmir_v28i1e85770_app4.docx]

References

1. KFF health misinformation tracking poll: artificial intelligence and health information. Kaiser Family Foundation. 2024. URL: https://www.kff.org/public-opinion/kff-health-misinformation-tracking-poll-artificial-intelligence-and-health-information/ [accessed 2026-05-28]
2. Ayre J, Cvejic E, McCaffery KJ. Use of ChatGPT to obtain health information in Australia, 2024: insights from a nationally representative survey. Med J Aust 2025 Mar 3;222(4):210-212. [doi: 10.5694/mja2.52598] [Medline: 39901778]
3. Ayo-Ajibola O, Davis RJ, Lin ME, Riddell J, Kravitz RL. Characterizing the adoption and experiences of users of artificial intelligence-generated health information in the United States: cross-sectional questionnaire study. J Med Internet Res 2024 Aug 14;26:e55138. [doi: 10.2196/55138] [Medline: 39141910]
4. Mansour RF, Fatouh AH. Measurement of bias in the contents of web search for health information retrieval. J Scientometric Res 2023 Nov 30;12(3):621-630. [doi: 10.5530/jscires.12.3.060]
5. Suzuki M, Yamamoto Y. Characterizing the influence of confirmation bias on web search behavior. Front Psychol 2021;12:771948. [doi: 10.3389/fpsyg.2021.771948] [Medline: 34938242]
6. Pines JM. Profiles in patient safety: confirmation bias in emergency medicine. Acad Emergency Med 2006 Jan;13(1):90-94. [doi: 10.1111/j.1553-2712.2006.tb00990.x] [Medline: 16365325]
7. Kusa W, Mosca E, Lipani A. “Dr LLM, what do i have?”: the impact of user beliefs and prompt formulation on health diagnoses. Presented at: Proceedings of the Third Workshop on NLP for Medical Conversations; Nov 1-4, 2023 URL: https://aclanthology.org/2023.nlpmc-1 [doi: 10.18653/v1/2023.nlpmc-1.2]
8. Loya M, Sinha DA, Futrell R. Exploring the sensitivity of LLMs’ decision-making capabilities: insights from prompt variations and hyperparameters. Presented at: Findings of the Association for Computational Linguistics; Dec 6-10, 2023 URL: https://aclanthology.org/2023.findings-emnlp [doi: 10.18653/v1/2023.findings-emnlp.241]
9. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 2025 Jan 28;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
10. Benary M, Wang XD, Schmidt M, et al. Leveraging large language models for decision support in personalized oncology. JAMA Netw Open 2023 Nov 1;6(11):e2343689. [doi: 10.1001/jamanetworkopen.2023.43689] [Medline: 37976064]
11. Zaretsky J, Kim JM, Baskharoun S, et al. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. JAMA Netw Open 2024 Mar 4;7(3):e240357. [doi: 10.1001/jamanetworkopen.2024.0357] [Medline: 38466307]
12. Strika Z, Petkovic K, Likic R, Batenburg R. Bridging healthcare gaps: a scoping review on the role of artificial intelligence, deep learning, and large language models in alleviating problems in medical deserts. Postgrad Med J 2024 Dec 23;101(1191):4-16. [doi: 10.1093/postmj/qgae122] [Medline: 39323384]

XSL•FO
RenderX

13. Rutledge GW, Sivura A. A generative AI–based virtual physician assistant. Proc AAAI Symp Ser 2024;3(1):64-65. [doi: 10.1609/aaaiss.v3i1.31182]
14. Wen B, Norel R, Liu J, Stappenbeck T, Zulkernine F, Chen H. Leveraging large language models for patient engagement: the power of conversational AI in digital health. Presented at: 2024 IEEE International Conference on Digital Health (ICDH); Jul 7-13, 2024. [doi: 10.1109/ICDH62654.2024.00027]
15. Kopka M, von Kalckreuth N, Feufel MA. Accuracy of online symptom assessment applications, large language models, and laypeople for self-triage decisions. NPJ Digit Med 2025 Mar 25;8(1):178. [doi: 10.1038/s41746-025-01566-6] [Medline: 40133390]
16. Reddy S. Global harmonization of artificial intelligence-enabled software as a medical device regulation: addressing challenges and unifying standards. Mayo Clin Proc Digit Health 2025 Mar;3(1):100191. [doi: 10.1016/j.mcpdig.2024.100191] [Medline: 40207007]
17. Ethics and governance of artificial intelligence for health: guidance on large multimodal models. : World Health Organization; 2024 URL: https://iris.who.int/server/api/core/bitstreams/e9e62c65-6045-481e-bd04-20e206bc5039/content [accessed 2026-05-28]
18. Kahneman D. Thinking, Fast and Slow: Penguin Books; 2011.
19. Qin X, Huang M, Ding J. AITurk: using chatgpt for social science research. SSRN. Preprint posted online on Jun 7, 2024. [doi: 10.2139/ssrn.4922861]
20. Park JS, O'Brien J, Cai CJ, Morris MR, Liang P, Bernstein MS. Generative agents: interactive simulacra of human behavior. Presented at: ACM Symposium on User Interface Software and Technology (UIST 2023); Oct 29 to Nov 1, 2023. [doi: 10.1145/3586183.3606763]
21. Pressman SM, Borna S, Gomez-Cabello CA, Haider SA, Haider CR, Forte AJ. Clinical and surgical applications of large language models: a systematic review. J Clin Med 2024 May 22;13(11):3041. [doi: 10.3390/jcm13113041] [Medline: 38892752]
22. Carlà MM, Gambini G, Baldascino A, et al. Large language models as assistance for glaucoma surgical cases: a ChatGPT vs. Google Gemini comparison. Graefes Arch Clin Exp Ophthalmol 2024 Sep;262(9):2945-2959. [doi: 10.1007/s00417-024-06470-5] [Medline: 38573349]
23. Carlà MM, Gambini G, Baldascino A, et al. Exploring AI-chatbots' capability to suggest surgical planning in ophthalmology: ChatGPT versus Google Gemini analysis of retinal detachment cases. Br J Ophthalmol 2024 Sep 20;108(10):1457-1469. [doi: 10.1136/bjo-2023-325143] [Medline: 38448201]
24. Wu S, Peng Z, Du Y, et al. A comparative study on reasoning patterns of OpenAI's o1 model. arXiv. Preprint posted online on Oct 17, 2024. [doi: 10.48550/arXiv.2410.13639]
25. Li Y, Du X, Song Y, Wang L, Wang S. A survey on fairness in large language models. arXiv. Preprint posted online on Aug 20, 2023. [doi: 10.48550/arXiv.2308.10149]
26. Pressman SM, Borna S, Gomez-Cabello CA, Haider SA, Haider C, Forte AJ. AI and ethics: a systematic review of the ethical considerations of large language model use in surgery research. Healthcare (Basel) 2024 Apr 13;12(8):825. [doi: 10.3390/healthcare12080825] [Medline: 38667587]
27. Schmidgall S, Harris C, Essien I, et al. Evaluation and mitigation of cognitive biases in medical language models. NPJ Digit Med 2024 Oct 21;7(1):295. [doi: 10.1038/s41746-024-01283-6] [Medline: 39433945]
28. Leslie-Miller CJ, Simon SL, Dean K, Mokhallati N, Cushing CC. The critical need for expert oversight of ChatGPT: prompt engineering for safeguarding child healthcare information. J Pediatr Psychol 2024 Nov 1;49(11):812-817. [doi: 10.1093/jpepsy/jsae075] [Medline: 39271174]
29. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9(1):e45312. [doi: 10.2196/45312] [Medline: 36753318]
30. Herrmann-Werner A, Festl-Wietek T, Holderried F, et al. Assessing ChatGPT's mastery of Bloom's taxonomy using psychosomatic medicine exam questions: mixed-methods study. J Med Internet Res 2024 Jan 23;26:e52113. [doi: 10.2196/52113] [Medline: 38261378]
31. Sharma M, Tong M, Korbak T, et al. Towards understanding sycophancy in language models. Presented at: The Twelfth International Conference on Learning Representations; May 7-11, 2024 URL: https://openreview.net/forum?id=tvhaxkMKAn [accessed 2026-05-28]
32. Ly DP, Shekelle PG, Song Z. Evidence for anchoring bias during physician decision-making. JAMA Intern Med 2023 Aug 1;183(8):818-823. [doi: 10.1001/jamainternmed.2023.2366] [Medline: 37358843]
33. Nendaz M, Perrier A. Diagnostic errors and flaws in clinical reasoning: mechanisms and prevention in practice. Swiss Med Wkly 2012;142:w13706. [doi: 10.4414/smw.2012.13706] [Medline: 23135902]
34. Xian Y, Lampert CH, Schiele B, Akata Z. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE Trans Pattern Anal Mach Intell 2019 Sep;41(9):2251-2265. [doi: 10.1109/TPAMI.2018.2857768] [Medline: 30028691]
35. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. Nat Med 2025 Mar;31(3):943-950. [doi: 10.1038/s41591-024-03423-7] [Medline: 39779926]

XSL•FO
RenderX

36. Van Koevering K, Kleinberg J. How random is random? Evaluating the randomness and humanness of LLMs' coin flips. arXiv. Preprint posted online on May 31, 2024. [doi: 10.48550/arXiv.2406.00092]
37. Li ZZ, Zhang D, Zhang ML, et al. From system 1 to system 2: a survey of reasoning large language models. arXiv. Preprint posted online on Feb 24, 2025. [doi: 10.48550/arXiv.2502.17419]
38. Giuffrè M, Kresevic S, Pugliese N, You K, Shung DL. Optimizing large language models in digestive disease: strategies and challenges to improve clinical outcomes. Liver Int 2024 Sep;44(9):2114-2124. [doi: 10.1111/liv.15974] [Medline: 38819632]
39. Rahman MM, Irbaz MS, North K, Williams MS, Zampieri M, Lybarger K. Health text simplification: an annotated corpus for digestive cancer education and novel strategies for reinforcement learning. J Biomed Inform 2024 Oct;158:104727. [doi: 10.1016/j.jbi.2024.104727] [Medline: 39293643]
40. Mirzaei T, Amini L, Esmaeilzadeh P. Clinician voices on ethics of LLM integration in healthcare: a thematic analysis of ethical concerns and implications. BMC Med Inform Decis Mak 2024 Sep 9;24(1):250. [doi: 10.1186/s12911-024-02656-3] [Medline: 39252056]
41. Garrubba M, Joseph C, Melder A. Best practice to identify and prevent cognitive bias in clinical decision-making: scoping review. : Monash Health; 2019 URL: https://monashhealth.org/wp-content/uploads/2020/03/Cognitive-Bias_Scoping-Review_2019_FINAL.pdf [accessed 2026-05-28]
42. GitHub. Sharon202588/biased-research-llm. URL: https://github.com/sharon202588/biased-research-llm.git [accessed 2026-05-29]

Abbreviations

AI: artificial intelligence
BIAD: bias-induced accuracy decline
BIEP: bias-influenced error proportion
LLM: large language model
MedQA: medical question answering dataset
pp: percentage points
USMLE: United States Medical Licensing Examination

Edited by A Coristine; submitted 13.Oct.2025; peer-reviewed by D Chrimes, I Li, KH Lin; revised version received 02.May.2026; accepted 04.May.2026; published 11.Jun.2026.

Please cite as:
Zuo Y, Wan Q, Wang S
Patient Cognitive Bias in Large Language Model–Supported Health Consultations: Simulation-Based Comparative Study
J Med Internet Res 2026;28:e85770
URL: https://www.jmir.org/2026/1/e85770
doi:10.2196/85770

© Yi Zuo, Qifeng Wan, Shalong Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Application Value of Radiomics-Based Machine Learning for Preoperative Risk Stratification of Bladder Cancer: Systematic Review and Meta-Analysis

Zirong He[1*], MM; Yinghua Liu[1*], MD; Qin Jiang[1], MM; Fang Hou[1,2*], MD; Bing Xu[1*], MM

[1]Department of Pediatric Surgery, Sichuan Provincial People's Hospital, School of Medicine, University of Electronic Science and Technology of China, No. 32 West Second Section, First Ring Road, Qingyang District, Chengdu, Sichuan, China

[2]Sichuan Provincial People's Hospital East Sichuan Hospital & Dazhou First People's Hospital, Dazhou, China

[*]these authors contributed equally

Corresponding Author:
Bing Xu, MM
Department of Pediatric Surgery, Sichuan Provincial People's Hospital, School of Medicine, University of Electronic Science and Technology of China, No. 32 West Second Section, First Ring Road, Qingyang District, Chengdu, Sichuan, China

Abstract

Background: Some researchers have explored the application of radiomics-based machine learning to detect preoperative muscle invasion, high-grade tumors, human epidermal growth factor receptor 2 expression, and other risk factors for bladder cancer. However, systematic evidence proving its effectiveness remains lacking.

Objective: This study aimed to evaluate the performance of radiomics-based machine learning in preoperative risk stratification for patients with bladder cancer. These findings could contribute to advancing the development or updating of intelligent risk assessment tools for bladder cancer.

Methods: The Embase, Cochrane Library, PubMed, and Web of Science databases were systematically retrieved for publicly available studies on the effectiveness of radiomics-based machine learning (ML) in the preoperative risk stratification of bladder cancer up to October 17, 2025. The risk of bias in the included studies was evaluated using the Prediction Model Risk of Bias Assessment Tool for Artificial Intelligence. The overall quality of the studies was quantified using the Radiomics Quality Scoring tool. The certainty of the evidence was graded using the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) framework. Subgroup analyses were conducted according to the type of imaging source and modeling method.

Results: This meta-analysis ultimately incorporated 57 studies with a total of 11,933 participants. These studies primarily used radiomics-based ML to identify muscle invasion (n=34) and high-grade tumors (n=16). Additionally, the methodology was used to evaluate human epidermal growth factor receptor 2 positive expression (n=3), Ki-67 expression (n=2), and lymph node staging (n=2) preoperatively in bladder cancer. In the validation sets, the pooled area under the receiver operating characteristic curve (AUROC) for identifying muscle invasion was 0.893 (95% CI 0.840-0.948), 0.916 (95% CI 0.891-0.942), and 0.840 (95% CI 0.737-0.958) for computed tomography (CT)–, magnetic resonance imaging (MRI)–, and ultrasound-based radiomics, respectively. The AUROC was 0.874 (95% CI 0.852-0.896) and 0.921 (95% CI 0.867-0.979) for models integrating clinical features with CT- or MRI-based radiomics, respectively. The pooled AUROC for diagnosing high-grade tumors was 0.874 (95% CI 0.775-0.985), 0.846 (95% CI 0.663-1.000), and 0.750 (95% CI 0.636-0.884) for CT-, MRI-, and ultrasound-based radiomics, respectively. Furthermore, the AUROC was 0.919 (95% CI 0.774-1.000) for MRI-based radiomics combined with clinical features.

Conclusions: This is the first systematic review to comprehensively evaluate the role of radiomics in preoperative risk stratification for bladder cancer. It provides evidence to inform the development and refinement of future ML-based tools for image analysis in this setting. However, this evidence faces significant challenges, including methodological shortcomings and a high risk of bias and low GRADE level, which preclude its readiness for clinical translation. Future studies should standardize the methodological workflows in radiomics, conduct multicenter research, and thoroughly evaluate and discuss the validity of external validation.

Trial Registration: PROSPERO CRD42024561649; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024561649

(J Med Internet Res 2026;28:e81084) doi:10.2196/81084

KEYWORDS

radiomics; bladder cancer; muscle invasion; pathologic grading; risk classification

XSL•FO
RenderX

Introduction

Bladder cancer is the tenth most common cancer globally. It ranks fourth in new cancer cases in men and is the eighth leading cause of cancer death [1,2]. In the United States, an estimated 83,190 new cases of bladder cancer occurred in 2024 (63,070 in men and 20,120 in women), resulting in 16,840 deaths (12,290 in men and 4550 in women) [2]. Thus, bladder cancer has become a serious social burden.

Transurethral resection of bladder tumors (TURBT), followed by pathological analysis, provides the basis for diagnosing, staging, and treating bladder cancer [3]. Nonetheless, this procedure is imprecise in assessing muscle invasion. The absence of detrusor muscle in the specimen is linked to a significantly elevated risk of residual lesions, early recurrence, and understaging of tumor [3,4]. TURBT also carries a few operational risks. A recent study of the learning curve for TURBT indicates that urologists should perform at least 100 TURBT procedures before achieving acceptable oncological outcomes and meeting the minimum requirements for surgical success. Furthermore, the study found that the first 45 TURBT procedures resulted in the worst outcomes [5]. Repeat transurethral resection provides a second chance in such cases. While repeat transurethral resection holds potential for improving tumor prognosis, evidence demonstrating its capacity to confer a survival benefit remains inconclusive [6,7]. Like TURBT, cystoscopy is an invasive procedure with risks of insufficient sampling and understaging. This creates a need for noninvasive, accurate diagnostic alternatives [8,9].

Magnetic resonance imaging (MRI) and computed tomography (CT) have been used to provide additional staging information. Nevertheless, neither technique can accurately evaluate microscopic infiltrates. Both of them aim to confirm or exclude locally advanced disease (stage ≥T3b) [10]. Therefore, exploring new preoperative techniques that can effectively identify the risk of muscle invasion and pathological grading of bladder cancer is clinically important.

In recent years, radiomics has received significant attention from researchers. Like therapies inspired by molecular biology, radiomics shows great promise in advancing precision medicine. This technique can predict outcomes individually or in combination with genomic, comorbidity, clinical, or demographic data. The process involves acquiring images, identifying volumes of interest (ie, those with possible prognostic value), segmenting the volume (ie, depicting the boundaries of the volume with computer-assisted contouring), and extracting and qualifying descriptive features from the volume. These features are then used to populate searchable databases, which are subsequently mined to develop classifier models. Radiomics shows promise as a quantitative imaging biomarker for both characterizing bladder cancer and predicting its prognosis [11-14]. At the time of this writing, high-throughput computing enables the rapid extraction of numerous quantitative features from tomographic images (eg, CT, magnetic resonance, or positron emission tomography). Radiomics transforms medical images into quantifiable data for analysis. The motivation stems from the wealth of pathophysiological information embedded in medical images, which quantitative analysis can unlock to enhance decision-making [11,13,15]. This algorithmic assistance may provide more accurate histopathological diagnoses, save time, and increase clinician confidence, contributing to improved prognosis [12,16-18]. Recently, as radiomics has developed in oncological diagnosis and treatment, researchers have examined the application of radiomics-based machine learning (ML) in identifying muscle invasion [18-21] and high-grade tumors [15], as well as in determining human epidermal growth factor receptor 2 (HER2) and Ki-67 expression and lymph node (LN) staging for bladder cancer. Compared with genomic biomarkers, artificial intelligence (AI) systems leveraging digitized images offer a more cost-effective and readily scalable solution. These tools can continuously learn from new data, thereby enhancing their predictive performance and, consequently, their value to health care over time [22]. However, comprehensive and systematic evidence of its effectiveness remains lacking.

To address this gap, we conducted a meta-analysis aimed at systematically evaluating the application of radiomics-based ML to preoperative risk stratification in bladder cancer and quantitatively summarizing its diagnostic accuracy. The findings may establish an evidence base for developing and refining future image-based, AI-assisted tools of risk stratification.

Methods

Study Registration

This study followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines (Checklist 1). The review was prospectively registered with PROSPERO (International Prospective Register of Systematic Reviews) under the title “Assessment of Preoperative Risk Stratification for Bladder Cancer Using Machine Learning Based on Radiomics: A Systematic Review and Meta-Analysis” (CRD42024561649).

Eligibility Criteria

Inclusion Criteria

The inclusion criteria were as follows:

1. Patients with bladder cancer diagnosed by biopsy;
2. A complete ML model covering radiomics features was constructed to identify muscle invasion, high-grade tumors, HER2, LN staging, and Ki-67 expression;
3. Studies reported in English.

Exclusion Criteria

The exclusion criteria were as follows:

1. Unpublished conference abstracts;
2. Studies with image segmentation only and without complete ML models for identifying muscle invasion, high-grade tumors, HER2, LN staging, and Ki-67 expression;
3. Studies that only performed texture analysis, with no ML classifiers constructed based on extracted textures;
4. Lack of metrics (eg, sensitivity or recall, area under the receiver operating characteristic curve [AUROC], specificity, calibration curves, accuracy, confusion matrix,

XSL•FO
RenderX

precision, and F_1-score) for predicting the accuracy of ML models.

Search Strategy and Data Source

The Cochrane Library (CENTRAL), Embase, PubMed, and Web of Science were retrieved up to June 17, 2024. This investigation did not perform simultaneous searches across multiple databases on a single platform. Searches were not extended to dedicated conference abstract databases or web sources beyond the primary databases. The authors of unpublished conference abstracts were not contacted to obtain full study details. To mitigate the risk of omitting relevant studies, the reference lists of eligible articles and pertinent review papers were manually scrutinized. No search filters were applied. The search strategy integrated both controlled vocabulary, such as Medical Subject Headings and Emtree terms, and free-text keywords to optimize sensitivity. Subject headings included "Urinary Bladder Neoplasms" and "machine learning." Search strategies were tailored to the specific syntax of each database and combined using Boolean operators. No restrictions were applied concerning publication date or geographical location. Before the final data analysis, an updated search was performed in all specified databases on October 17, 2025 (Multimedia Appendix 1).

Selection of Studies and Data Extraction

EndNote was used to import the retrieved articles. After eliminating duplicate records, the remaining articles were reviewed by titles and abstracts. Exclusions were categorized as follows: meta-analyses or reviews, replies or letters, case reports, animal experiments, registry or clinical trial protocols, non-English articles, and preprints. Initially relevant articles were screened by full text to determine eligible studies.

Before data extraction, a spreadsheet was created. The collected information encompassed DOI, title, publication year, first author, country, study type, purpose of the task, patient source, image source, recording of a complete image acquisition protocol, number of researchers involved, whether preliminary experiments tested different imaging parameters, radiomic region of interest (ROI) segmentation software, total number of outcome events, whether test-retest experiments were performed, total number of cases, training set size, number of outcome events in the training set, validation set generation method, validation set size, number of outcome events in the validation set, variable selection method, model types used, overfitting assessment method, public availability of data and code, mean age, sex, specimen source, AUROC (95% CI), number of true negatives, number of true positives, specificity, sensitivity, precision, accuracy, and F_1-score.

Results from each dataset within a study were included only once. When multiple studies published by the same author over different years were suspected of having overlapping datasets, only the study with the larger sample size was included. When multiple models were present, the model demonstrating optimal performance in the validation set was selected for inclusion. If a single dataset contained multiple validation sets, all were incorporated into the analysis.

Two researchers (ZH and YL) screened the articles and extracted the data separately. Their results were cross-checked. Any discrepancies were addressed with the help of a third researcher (BX).

Assessment of Risk of Bias and Quality

The Prediction Model Risk of Bias Assessment Tool for Artificial Intelligence (PROBAST-AI) provides a framework to critically appraise the risk of bias (ROB) and applicability of ML-based multivariable prediction models [23,24]. The assessment for ROB comprises four domains: participants and data sources, predictors, outcome, and analysis (1=strongly disagree, 2=somewhat disagree, 3=I don't know, 4=somewhat agree, 5=strongly agree). A domain was judged as high risk if any signaling question was rated "no/probably no," and as low risk only if all questions were answered "yes/probably yes." An "I don't know" response yielded an unclear risk designation.

Furthermore, the quality of the eligible investigations was appraised using the Radiomics Quality Score (RQS), which ranges from –8 to 36 [25]. This RQS primarily considers image protocol quality, multiple image segmentation methods, study of modalities, image acquisition time, feature reduction, and model construction using radiomic and nonradiomic features (prognosis and molecular subtyping). It also considers the detection and discussion of radiomic and biological correlations, cutoff value analysis, calibration statistics, discrimination statistics, validation, prospective studies registered in trial databases, comparison with the "gold standard," cost-effectiveness analysis, potential clinical utility, and open science and data.

Two investigators independently evaluated the ROB and methodological quality and then performed a cross-check of their assessments. Any discrepancies were adjudicated with the help of a third researcher (BX).

Synthesis Methods

A meta-analysis was performed on AUROC, a metric for the accuracy of ML models. The analysis required the SE or 95% CI of the AUROC. However, some of the included studies lacked a 95% CI and SE. In this scenario, SE was estimated by referring to the study by Debray et al [26]. A random-effects model was used for the meta-analysis to account for potential heterogeneity among the primary studies due to variations in model parameter tuning and predictor selection. Furthermore, the Hartung-Knapp-Sidik-Jonkman method was applied [27].

Furthermore, a meta-analysis was performed using a bivariate mixed effects model to pool sensitivity and specificity. This analysis was based on diagnostic contingency tables. Nevertheless, these tables were not directly provided in some studies. Then, the required values were calculated via reported sensitivity, specificity, precision, and sample sizes. Subsequently, subgroup analyses were conducted based on the dataset (training and validation sets), image source, and model type. A meta-analysis of AUROC values was conducted using packages "*meta*" and "*metafor*" in R (v4.5.2; R Foundation for Statistical Computing). A bivariate mixed effects model was implemented using the "*midas*" package in Stata (v15.0; StataCorp LLC).

XSL•FO
RenderX

$$SE(c) \approx c(1-c)[1+n*(1-c)2-c+m*c1+c]mn$$

Parameter definitions: c denotes the c-statistic; n represents the number of observed events; and m corresponds to the total sample size n∗=m∗=m+n2−1.

Finally, subgroup analyses were conducted within each task category based on case number, ML type (logistic regression [LR] vs other ML), and modeling variables (radiomics features alone vs radiomics features+clinical characteristics).

Certainty of Evidence

The certainty of evidence for the AUROC estimates derived from various subgroup analyses was evaluated using the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) framework. Two investigators independently performed the evidence grading and a cross-verification of their assessments. Any discrepancies were adjudicated by a third investigator (BX).

Results

Study Selection

The database search yielded 8272 articles. After removing 2234 duplicates, 6038 articles remained and were reviewed based on their titles and abstracts. This process resulted in the exclusion of 5971 ineligible records. Overall, 67 articles remained and were reviewed by full texts, leading to the removal of 10 for irrelevance, including 4 non–peer-reviewed conference papers, one study lacking radiomics analysis, one focused on differentiating pure urothelial carcinoma from urothelial carcinoma with squamous differentiation, one without radiomics, one enrolling participants without a pathological diagnosis, one differentiating T2 from T3 bladder cancer stages, and one predicting programmed death-ligand 1 expression. Ultimately, the meta-analysis incorporated 57 articles [28-84] (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) of the literature selection process.

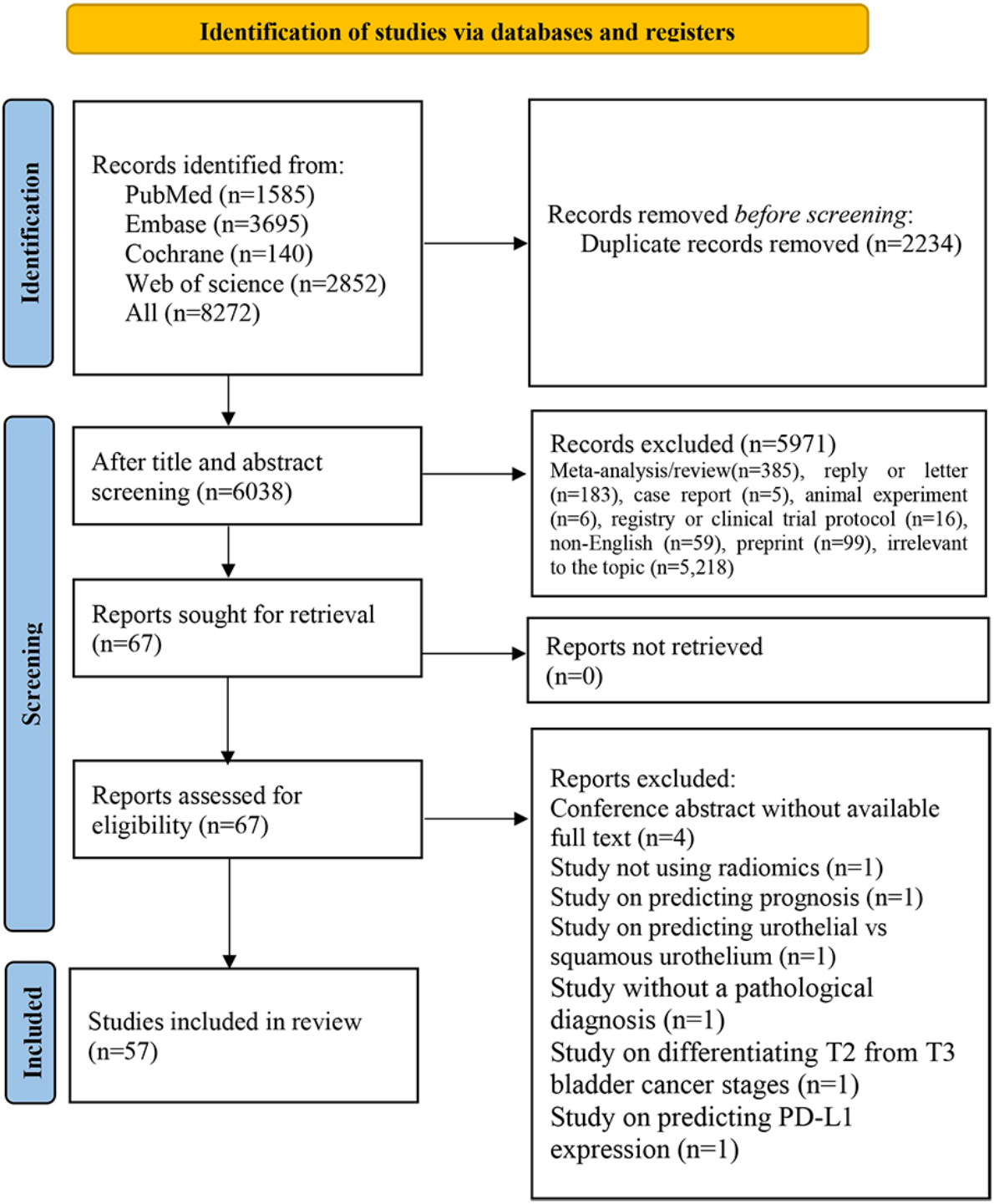

Characteristics of Studies

Fifty-seven [28-84] eligible studies, published between 2017 and 2025, encompassed data on 11,933 individuals with bladder cancer. Most were case-control studies, including 40 [28,33-36,38-41,43-46,48,49,51,53-57,61-76,79,82,84] single-center and 17 [29-32,37,42,47,50,52,58-60,77,78,80,81,83] multicenter studies. Thirty-four studies primarily investigated muscle invasion; 20 [30,32,33,41-43,46,51,52,54,57-60,63,64,69,71,78,83], 13[28,29,31,34,45,48,50,62,75,79-82], and one [65] used MRI-, CT-, and ultrasound-based radiomics, respectively. Sixteen studies [36,38-40,44,47,49,53,56,61,66,67,70,72,73,84] focused on discriminating high-grade tumors. Three studies [55,74,77] examined HER2-positive expression, 2 [35,76] examined Ki-67 expression, and 2 [37,68] examined LN staging in bladder cancer. ITK-SNAP (University of Pennsylvania) was the most frequently used software for ROI segmentation. Other software platforms included MATLAB (R2012b; Matrix Laboratory;

XSL•FO
RenderX

MathWorks), the open-source Medical Imaging Interaction Toolkit, and a computer-assisted visualization and analysis software system. Fifty-three studies explicitly described the generation of a validation; among these, 37 [28,34-37,39-41,43-46,48,49,51,53-57,61-66,68-72,74-76,79,82,84] used internal validation techniques such as random sampling or cross-validation and 16 [29,30,31,32,42,47,50,52,58-60,77,78,80,81,83] studies used external validation. A total of 10 distinct model types were evaluated (Table 1).

XSL•FO
RenderX

Table . Basic information.

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
Zhang et al (2024) [28]	China	Case-control	Single center	• Total, median (range): 62 (26-86) • Training set, median (range): 65 (26-86) • Validation set, median (range): 64 (34-84)	• Total: F: 36, M: 160[b] • Training set: F: 20, M: 117[c] • Validation set: F: 16, M: 43	Muscle invasion status	CT[d]	Pathology (TURBT[e] or a radical surgical specimen)+CT	2	ITK-SNAP program (version 4.0.1; University of Pennsylvania)	196	137	Random sampling	• 59	LR[f], LASSO[g]

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
Zhang et al (2021) [29]	China	Case-control	Multi-center	• Training set, median (IQR): 65 (56-72) • Internal validation, median (IQR): 68 (61-74) • External validation, median (IQR): 65 (59-77)	• Training set: F: 75, M: 28 • Internal validation: F: 13, M: 6 • External validation: F: 13, M: 2	Muscle invasion status	CT	Pathology (TURBT or a radical surgical specimen)+CT	2	Deepwise Research Platform (Deepwise Inc)	Center 1: 366 Center 2: 75	293 (Development group+adjustment Group)	External validation	• 76 (Internal test)	DL[h]
Ye et al (2023) [30]	China	Case-control	Multi-center			Muscle invasion status	MRI[i]	Pathology (TURBT or a radical surgical specimen)+MRI	1	ITK-SNAP	Center 1: 160; Center 2: 55	—[j]	External validation		SVM[l]

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
				• Training set, median (IQR): 67 (59-73) • Internal validation, median (IQR): 69 (62-71) • External validation, median (IQR): 64 (55-73)	• Training set: F: 21; M: 19 • Internal validation F: 0; M: 30 • External validation F: 8; M: 47									• Internal validation: 25 (31 tumor lesions, NSCLC[k]: 4) • External validation: 54 (55 tumor lesions, NSCLC: 32)	
Ren et al (2023) [31]	China	Case-control	Multicenter	• Training set, median (IQR): 75 (57-79) • Validation set, median (IQR): 72 (66-78)	• Training set: F: 57; M: 13 • Validation set: F: 33; M: 46	• Made invasion status	CT	Pathology+within 30 days CT	1	ITK-SNAP	269	188	External validation	81	LR

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
Qureshi et al (2024) [32]	United States	Case-control	Multi-center	—	—	Muscle invasion status	MRI	Pathology+MRI	0	MATLAB (MATrix LABoratory; MathWorks)	—	—	External validation	—	NB[m], SVM, DT[n], KNN[o], LR
Özdemir et al (2023) [33]	Türkiye	Case-control	Single center	—	—	Muscle invasion status	MRI	Pathology+MRI	1	—	60	—	—		LR

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
Chen et al (2022) [34]	China	Case-control	Single center	• Training set, median (range): NMIBC: 66 (46-85); MIBC: 66 (45-85); NMIBC[b]: 65 (46-85); MIBC: 66 (45-85) • Validation set: NMIBC: 67 (44-85); MIBC: 66 (42-85); NMIBC: 65 (42-85); MIBC: 66 (42-85)	• Training set: NMIBC: F: 26; M: 76; MIBC: F: 78; M: 19 • Validation set: NMIBC: F: 11; M: 30; MIBC: F: 26; M: 7	Muscle invasion status	CT	Pathology (TURBT or a radical surgical specimen) within+4 weeks MRI	2	ITK-SNAP	173	129	Random sampling	44	DL
Zheng et al (2021) [35]	China	Case-control	Single center	• Training set, median (range): <65 years: 42 (35-65); ≥65 years: 83 (64-94) • Validation set, median (range): <65 years: 24 (44-65); ≥65 years: 30 (55-76)	• Training set: F: 19; M: 16 • Validation set: F: 13; M: 41	Ki-67 expression	MRI	Pathology (TURBT or a radical surgical specimen)+MRI	2	ITK-SNAP	179	125	Random sampling	54	SMOTE-LASSO[q]
	China	Cohort study	Single center			Histological grade	MRI	Pathology+MRI	2	ITK-SNAP	92	64		28	LR

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
Ye et al (2023) [36]				• Training set, mean (SD): 61 (12); Validation set: 61 (12)	• Training set: F: 10; M: 54; Validation set: F: 5; M: 23								Random sampling		
Starmans et al (2022) [37]	Netherlands	Case-control	Multicenter	—	—	Lymph node staging	CT	Pathology+CT	1	WORC toolbox (Erasmus MC)	209	—	Cross validation	—	DL
Sarkar et al (2023) [38]	United States	Case-control	Single center	—	—	Histological grade	CT	—	2	—	100	—	—	—	NB
Li et al (2023) [39]	China	Case-control	Single center	—	—	Histological grade	MRI	—	2	ITK-SNAP	169	118	Random sampling	51	LR

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
Deng et al (2023) [40]	China	Case-control	Single center	• Training set, mean (SD): low grade 67 (14), high grade 68 (9); • Validation set, mean (SD): 79 (9)	• Training set: low grade F: 28, M: 15; high grade F: 7; M: 36; • Validation set: F: 7; M: 25	Histological grade	CT	Pathology+CT	2	Darwin Research Platform (Yizhun Medical AI)	105	73	Random sampling	32	SVM, KNN, GBDT[r], LR, RF[s], XGBoost[t]
Chen et al (2023) [41]	China	Case-control	Single center	—	—	Muscle invasion status	MRI, pathology	Pathology+MRI	2	Darwin Research Platform	445	312	Random sampling	133	SVM, KNN, Decision tree, RF, XGBoost, GBDT
Yu et al (2024) [42]	China	Case-control	Multicenter	—	—	Muscle invasion status	MRI, pathology	Pathology (TURBT or a radical surgical specimen)+MRI	2	—	436	404	External validation	32	DL

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
Xu et al (2019) [43]	China	Case-control	Single center	• Total, median (range): 61 (37-93) • Training set, median (range): 63 (38-80) • Validation set, median (range): 65 (37-93)	• Total F: 49, M: 19 • Training set F: 27, M: 14 • Validation set F: 22, M: 5	Muscle invasion status	MRI	Pathology (TURBT or a radical surgical specimen)+MRI	2	ITK-SNAP	218	136	Random sampling	87	RF
Zhou et al (2019) [44]	China	Case-control	Single center	• Median (range): 65 (45-74)	• F: 42, M: 66	Histological grade	MRI	Pathology (TURBT)	2	ITK-SNAP	108	72	Random sampling	36	LR
Yang et al (2021) [45]	China	Case-control	Single center	• Median (range): 68 (12)	• F: 52, M: 37	Muscle invasion status	CT	Pathology (a radical surgical specimen)+CT	2	—	369	—	Random sampling	—	DL-CNN[u]
Zheng et al (2021) [46]	China	Case-control	Single center		• Training set F: 21, M: 19 • Validation set F: 12, M: 4	Muscle invasion status	MRI	Pathology+MRI	2	ITK-SNAP	185	129	Random sampling	56	LASSO, RF, SVM

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
				• Training set, n (%): <65 years: 43 (32.2); ≥65 years: 81 (62.8) • Validation set, n (%): <65 years: 20 (35.7); ≥65 years: 36 (64.3)											
Song et al (2023) [47]	China	Case-control	Multicenter	• Training set, median (range): 66 (9-74) • Validation set, median (range): 65 (18-72)	• Training set F: 46, M: 0 • Validation set F: 13, M: 44	Histological grade	CT	—	2	ITK-SNAP	688	469	External validation	219	LR, NB, SVM, KNN, RF, DT, XGBoost, LightGBM, GBDT, AdaBoost[v], ANN[w], DL
Cui et al (2022) [48]	China	Case-control	Single center			Muscle invasion status	CT	Pathology+CT	2	ITK (Insight Software Consortium)	327	120	Random sampling	68	LR

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
				• Training set, mean (SD): MGC: 62 (19); MEC: 62 (17) • Validation set, mean (SD): MGC: 63 (13); MEC: 63 (9)	• Training set: MGC: F: 16; M: 44; MEC: F: 8; M: 52 • Validation set: MGC: F: 7; M: 27; MEC: F: 5; M: 29										
Zhang et al (2020) [49]	China	Case-control	Single center	• Training set, n (%): ≤60 years: 42 (38); >60 years: 66 (61) • Validation set, n (%): ≤60 years: 14 (33); >60 years: 23 (54)	• Training set: F: 28; M: 80 • Validation set: F: 8; M: 29	Histological grade	CT	Pathology+CT	2	Deepwise Research Platform (Deepwise Healthcare)	145	108	Random sampling	37	LR
	China						CT		2		375	226			DL

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
Wei et al (2023) [50]		Case-control	Multi-center	• Training set, median (range): 66 (18-86) • Validation set, median (range): 64 (45-86)	• Training set: F: 17; M: 29 • Validation set: F: 11; M: 16	Muscle invasion status		Pathology+CT		ITK-SNAP			Random sampling	internal validation: 97; external validation: 52	
Zheng et al (2019) [51]	China	Case-control	Single center	• Training set, median (range): 64 (57-69) • Validation set, median (range): 61 (54-70)	• Training set: F: 17; M: 113 • Validation set: F: 9; M: 60	Muscle invasion status	MRI	Pathology+MRI	2	3D Slicer version 4.9.0.	169	130	Grouping by different periods	69	LR
Wang et al (2019) [52]	China	Case-control	Multi-center	• Training set, median (range): 68 (40-86) • Validation set, median (range): 62 (40-81)	• Training set: F: 8; M: 56 • Validation set: F: 7; M: 35	Muscle invasion status	MRI	Pathology+MRI	2	MATLAB 2016 (MathWorks)	106	64	External validation	42	LR
Zheng et al (2021) [53]	China	Case-control	Single center			Histological grade	MRI	Pathology+MRI	2	ITK-SNAP	298	206	Random sampling	88	LASSO, RF, SVM

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
				• Training set, n (%): <65 years: 43 (32.2%); ≥65 years: 81 (67.8%) • Validation set, n (%): <65 years: 22 (39.3%); ≥65 years: 34 (60.7%)	• Training set: F: 21; M: 103 • Validation set: F: 11; M: 45										
Wang et al (2022) [54]	China	Case-control	Single center		• Total: F: 43; M: 146 • Training set: F: 22; M: 63 • Validation set: F: 8; M: 28 • Test set: F: 15; M: 55	Muscle invasion status	MRI	Pathology+MRI	2	ITK-SNAP	191	85	Random sampling	validation: 36; test: 70	LR

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
				• Total, median (IQR) 70 (62-79) • Training set, median (IQR) 69 (62-79) • Validation set, median (IQR) 70 (62-77) • Test set, median (IQR) 69 (61-74)											
Yu et al (2023) [55]	China	Case-control	Single center		training set: F: 19; M: 137 validation set: F: 9; M: 30; test set: F: 6; M: 37	HER 2 status	MRI	Pathology+MRI	2	ITK-SNAP	195	156	Random sampling	39	SVM, RF, LR, NB, KNN, AdaBoost

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
				• Training set (n): >65 years: 99; ≤65 years: 57; • Validation set (n): >65 years: 19; ≤65 years: 20; • Test set (n): >65 years: 25; ≤65 years: 28											
Wang et al (2023) [56]	China	Case-control	Single center		• Total: F: 53; M: 122 • Training set: F: 35; M: 96 • Validation set: F: 20; M: 76	Histological grade	MRI	Pathology+MRI	2	ITK-SNAP	227	131	Random sampling	96	LR

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
				• Total, median (IQR): 70 (62-77) • Training set, median (IQR): 70 (64-77) • Validation set, median (IQR): 69 (61-79)											
Zhang et al (2022) [57]	China	Case-control	Single center	• Training set, mean (SD): 66 (11) • Validation set, mean (SD): 67 (11) • Test set, mean (SD): 66 (11)	• Training set F: 24; M: 25 • Validation set F: 9; M: 5 • Test set F: 5; M: 3	Muscle invasion status	MRI	Pathology+MRI	2	ITK-SNAP	342	239	Random sampling	Validation: 68; test: 35	LR
Zou et al (2022) [58]	China	Cohort study	Multi-center			Muscle invasion status	MRI	Pathology+MRI	2	MB-MIP model	468	290	Random sampling	178	DL

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
				• Training set, median (range): 66 (26-79) • Validation set, median (range): 68 (11-91) • Retrospective, median (range): 65 (32-89) • Prospective, median (range): 65 (47-91) • Multicenter, median (range): 68 (42-89)	• Training set: F: 37; M: 23 • Validation set: F: 6; M: 10 • Retrospective: F: 5; M: 29 • Prospective: F: 7; M: 32 • Multicenter: F: 7; M: 32									
Zhou et al (2022) [59]	China	Case-control	Multicenter	—	—	Muscle invasion status	MRI	—	—	EvidentialNet	—	—	Random sampling	—	DL

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
Li et al (2023) [60]	China	Case-control	Multi-center	• Training set, median (IQR): 67 (57-70) • Test set, median (IQR): 63 (55-70)	• Training set: F: 13; M: 80 • Test set: F: 1; M: 27	Muscle invasion status	MRI	Pathology+MRI	1	ITK-SNAP	121	93	External validation	28	DL
Li et al (2024) [61]	China	Case-control	Single center	• Training set, median (IQR): 66 (27-90) • Validation set: 67 (27-90)	• Training set: F: 33; M: 145 • Validation set: F: 14; M: 63	Histological grade	MRI	Pathology+MRI	2	ITK-SNAP	255	178	Random sampling	77	LR
Chen et al (2022) [62]	China	Case-control	Single center	• Training set, median (IQR): [illegible] • Validation set, median (IQR): [illegible]	• Training set: F: 36; M: 81 • Validation set: F: 10; M: 41	Muscle invasion status	CT	Pathology+CT	4	ITK-SNAP	168	117	Random sampling	51	NB
Xu et al (2017) [63]	China	Case-control	Single center	—	—	Muscle invasion status	MRI	Pathology+MRI	2	MATLAB R2012b	68	—	Cross validation	—	SVM

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
Xu et al (2018) [64]	China	Case-control	Single center	—	—	Muscle invasion status	MRI	Pathology+MRI	2	MATLAB R2015b	54	—	Cross validation	—	SVM
Gao et al (2021) [65]	China	Case-control	Single center	—	—	Muscle invasion status	Ultrasound	Pathology+ultrasound	2	ITK-SNAP	157	110	Random sampling	47	NB
Wang et al (2019) [66]	China	Case-control	Single center	• Total, n (%): 63 (64) • Training set, n (%): 63 (60) • Validation set, n (%): 52 (81)	• Total: F: 14; M: 86 • Training set: F: 11; M: 59 • Validation set: F: 3; M: 27	Histological grade	MRI	Pathology+MRI	2		100	70	The time of undergoing surgery varies.	30	LR
Zhang et al (2017) [67]	China	Case-control	Single center	—	—	Histological grade	MRI	Pathology+MRI	2	MATLAB R2012b	—	61	—	—	SVM

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
Gresser et al (2022) [68]	Germany	Case-control	Single center	• Training set, median (IQR): PN2: 63 (11); PN0: 63 (10) • Validation set, median (IQR): PN2: 70 (9); PN0: 69 (11)	• Training set: PN2: F: 24, M: 42; PN0: F: 57, M: 51 • Validation set: F: 9, M: 19; PN0: F: 16, M: 73	Lymph node staging	CT	Pathology (a radical surgical specimen)+MRI	2	Medical Imaging Interaction Toolkit (MITK, DKFZ, Heidelberg; version 2018.04.2)	391	274	Random sampling	117	ANN
Tong et al (2018) [69]	United States	Case-control	Single center	• Median (IQR): 65 (9)	• F: 18, M: 47	Muscle invasion status	MRI	Pathology+MRI	2	—	65	—	Leave-one-out method	—	SVM
Tao et al (2024) [70]	China	Case-control	Single center	• Median (IQR): 65 (9-86)	• F: 33, M: 15	Histological grade	MRI	Pathology+MRI	—	3D Slicer (version 4.10.2-2026)	138	110	Random sampling	28	DL
Liu et al (2022) [71]	China	Case-control	Single center	• Training set: 63 (16) • Validation set: 41 (10)	• Training set: F: 19, M: 16 • Validation set: F: 7; M: 34	Muscle invasion status	MRI	Pathology+MRI	2	Manually depict	206	165	Cross validation	28	LASSO
	China	Case-control	Single center	—	—		CT		2		75	51		24	DL

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
Liu et al (2022) [72]						Histological grade		Pathology+CT		MSRN network (Multi-Scale Residual Network)			Random sampling		

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
Razik et al (2021) [73]	India	Case-control	Single center	• Median (range): 56 (18-83)	• F: 5; M: 35	Histological grade	MRI	Pathology+MRI	2	TexRAD (Feedback Medical)	40	—	—	—	LR
Peng et al (2024) [74]	China	Case-control	Single center	• Training set, median (range): HER2 positive: 54 (16-83); HER2 negative: 54 (12-83) • Validation set, median (range): HER2 positive: 51 (16-83); HER2 negative: 54 (18-83)	—	HER2 status	CT	Pathology+CT	2	ITK-SNAP	124	100	Random sampling	24	SVM, RF, LR, MLP[y], ExtraTrees, LightGBM[z], XGBoost, KNN
Xiong et al (2024) [75]	China	Case-control	Single center		• Training set: F: 8; M: 65 • Validation set: F: 8; M: 24	Muscle invasion status	CT	Pathology+CT	2	Darwin Research Platform	105	73	Random sampling	32	GBDT, KNN, LR, RF, SVM, (XGBoost

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
				• Training set, median (IQR): 60 (50-70) • Validation set, median (IQR): 60 (55-70)											
Feng et al (2024) [76]	China	Case-control	Single center	• Median (IQR): 66 (57)	• F: 27; M: 18	Ki-67 expression	CT	Pathology+CT	2	3D Slicer (version: 4.10.2)	135	94	Random sampling	41	LR
Wei et al (2024) [77]	China	Case-control	Multicenter	• Total, median (IQR): 52 (13) • Training set, median (IQR): 61 (8) • Validation set, median (IQR): 65 (13)	• Total: F: 22; M: 15 • Training set: F: 16; M: 13 • Validation set: F: 6; M: 2	HER2 status	CT	Pathology+CT	2	ITK-SNAP	207	154	External validation	53	LR, SVM, KNN, RF, XGBoost
Cai et al (2025) [78]	China	Case-control	Multicenter			Muscle invasion status	MRI	Pathology+MRI	2	DL	559	291	External validation		DL, VI-RADS

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
				• Training set, mean (SD): 66 (12) • Validation set, mean (SD): 67 (10) • Internal test set, mean (SD): 67 (11) • external test set: (9) (10)	• Training set: F: 37; M: 24 • Validation set: F: 6; M: 10 • Internal test set: F: 25; M: 13 • External test set: F: 6; M: 32									Validation set: 66; internal test set: 164, external test set: 38	
Du et al (2025) [79]	China	Case-control	Single center	• Training set, mean (SD): 54 (12) MIC: 67 (12) MIB: 67 (12) • Validation set, mean (SD): 54 (15) MIC: 67 (15) MIB: 62 (14)		Muscle invasion status	CT	Pathology+CT	2	ITK-SNAP (v.3.8.0)	200	140	Random sampling	60	DL

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
					• Training set: NMIBC: F: 27, M: 70; MIBC: F: 12, M: 31 • Validation set: NMIBC: F: 10, M: 34; MIBC: F: 6; M: 10										
Du et al (2025) [80]	China	Case-control	Multicenter	• Training set total: 97 (92) NMIBC: 92 (92) MIBC: 92 (85) • Validation set total: 63 (42) NMIBC: 90 (25) MIBC: 63 (16)		Muscle invasion status	CT	Pathology+CT	2	ITK-SNAP	325	231	External validation	94	Extra tree

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
					• Training set total: F: 51; M: 160; MRI: F: 47; M: 146; MRI: F: 4; M: 34 • Validation set total: F: 22; M: 72; MRI: F: 20; M: 53; MRI: F: 2; M: 14										
He et al (2025) [81]	China	Case-control	Multi-center			Muscle invasion status	CT	Pathology+CT	1	Manual	211	118	External validation	Test set: 30; external validation: 63	LR, SVC (linear), SVC (poly), SVC (rbf)

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
				• Training set, (n) ≤60 years 48; >60 years 70 • Test set (n) ≤60 years 11; >60 years 19 • External test set: ≤60 years 24; >60 years 39	• Training set: F: 29; M: 89 • Test set: F: 8; M: 22 • External test set: F: 21; M: 42										
Hu et al (2025) [82]	China	Case-control	Single center	• Training set, n (%) NMIBC 52 (80.0) MIBC 65 (69.1) • Validation set, n (%) NMIBC 40 (73.4) MIBC 73 (76.6)	• Training set NMIBC F: 16; M: 83; MIBC F: 5; M: 37 • Validation set NMIBC F: 4; M: 33; MIBC F: 7; M: 12	Muscle invasion status	CT	Pathology+CT	2	ITK-SNAP (version 3.8.0)	202	141	Random sampling	61	DECT[a], Radiomacs, Nomogram

XSL•FO
RenderX

Studies	Country	Study type	Source of the patient	Age (years)	Gender (n)	Task	Radiomics source	Specimens source	Number of image researcher	ROI[a] segmentation software	The total number of cases (N)	Number of cases in the training set (n)	Generation method of validation set	Number of cases in validation set	Model type
Yu et al (2025) [83]	China	Case-control	Multi-center	—	—	Muscle invasion status	MRI	Pathology+MRI	4	Manual	401	313	External validation	Validation set: 26; internal test set: 34; external test set: 28	DL
Zhou et al (2025) [84]	China	Case-control	Single center	• [illegible]	• L: F: 4; M: 14; H: F: 3; M: 16	Histological grade	CT	Pathology+CT	2	3D Slicer image	372	259	Random sampling	LR, DT, SVM, AdaBoost	LR, DT, SVM, AdaBoost

[a]ROI: region of interest.
[b]F: female.
[c]M: male.
[d]CT: computed tomography.
[e]TUBRT: transurethral resection of bladder tumor.
[f]LR: logistic regression.
[g]LASSO: least absolute shrinkage and selection operator.
[h]DL: deep learning.
[i]MRI: magnetic resonance imaging.
[j]Not available.
[k]MIBC: muscle-invasive bladder cancer.
[l]SVM: support vector machine.
[m]NB: naive Bayes.
[n]DT: decision tree.
[o]KNN: k-nearest neighbors.
[p]NMIBC: non–muscle-invasive bladder cancer.
[q]SMOTE-LASSO: Synthetic Minority Over-sampling Technique–Least Absolute Shrinkage and Selection Operator.
[r]GBDT: gradient boosting decision tree.
[s]RF: random forest.
[t]XGBoost: extreme gradient boosting.
[u]DL-CNN: deep learning–convolutional neural network.
[v]AdaBoost: adaptive boosting.
[w]ANN: artificial neural network.
[x]HER-2: human epidermal growth factor receptor 2.
[y]MLP: multilayer perceptron.
[z]LightGBM: light gradient boosting machine.
[aa]DECT: dual-energy computed tomography.
[ab]LGUC：low-grade urothelial carcinoma
[ac]HGUC: high-grade urothelial carcinoma

XSL•FO
RenderX

Evaluation of ROB and Quality

PROBAST-AI

Regarding the model development, all studies exhibited low ROB across all signaling questions in the "Participants" domain, with the sole exception of the question "Did the in- and exclusions of study participants result in a representative data set?" for which all studies were rated as "unclear." For "Predictor" assessments, all included investigations indicated an uncertain risk pertaining to the question "Were predictor assessments made without knowledge of outcome data?" In the "Outcomes" domain, all studies demonstrated low ROB. Within the "Analysis" segment, uncertainty was noted in 40 [29,30,32-34,36-38,40-45,47-50,55-57,59,60,62-65,67,69-73,77-79,81-84] investigations concerning "Was there evidence that the sample size was reasonable?" Two publications [32,35] also flagged uncertainty regarding "Were participants with missing or censored data handled appropriately in the analysis?" Furthermore, 21 [29,30,32,34,35,39,45,53,57-60,67,68,71-73,79-81,83] studies presented uncertainty on the item "If methods to address class imbalance were used, was the model or the model predictions recalibrated?" Forty-one [28-32,34-43,45-55,58-62,65-70,72,73,78,79] investigations indicated uncertainty for "Were methods used to address potential model overfitting?" (Figure 2).

Figure 2. Risk of bias assessment results for model development.

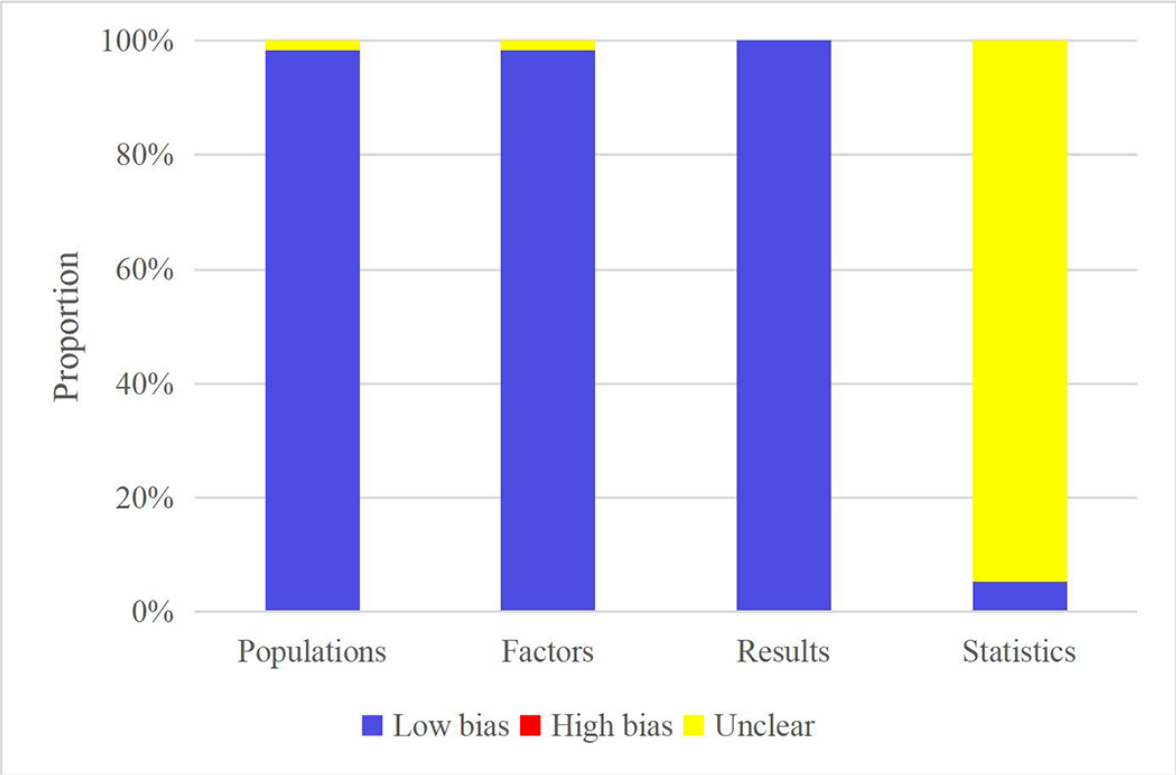

During model evaluation, all but one [33] study demonstrated low ROB in the "Participants" domain. The exception was an unclear rating for the question "Did the in- and exclusions of study participants result in a representative data set?" For the "Predictors" domain, all studies were uniformly rated as unclear regarding "Were predictor assessments made without knowledge of outcome data?" The "Outcome" domain was consistently judged as low ROB across all studies. In the "Analysis" domain, 4 [33,38,67,73] studies were identified with a high ROB for "Was model evaluation based on only apparent performance avoided?" precluding further analysis. Forty-one [28,30-32,34-36,39,40,42-46,48-53,55,57,59-66,69,71,72,74-77,79,80,82,83] publications indicated a high ROB regarding "Was there evidence that the sample size was reasonable?" Additionally, 37 [28-32,34-36,39-43,45-55,58-62,65,66,68-70,72,78,79] studies reported an unclear risk for "If resampling methods were used to evaluate model performance, were all model development steps replicated in the resampling process?" (Figure 3).

In summary, while the model development generally demonstrated a low ROB across participants and data sources, predictors, outcome, and analysis, the model evaluation phase presented a higher ROB specifically within the analysis dimension, particularly concerning the criteria "Was model evaluation based on only apparent performance avoided?" and "Was there evidence that the sample size was reasonable?"

XSL•FO
RenderX

Figure 3. Risk of bias assessment results for model evaluation.

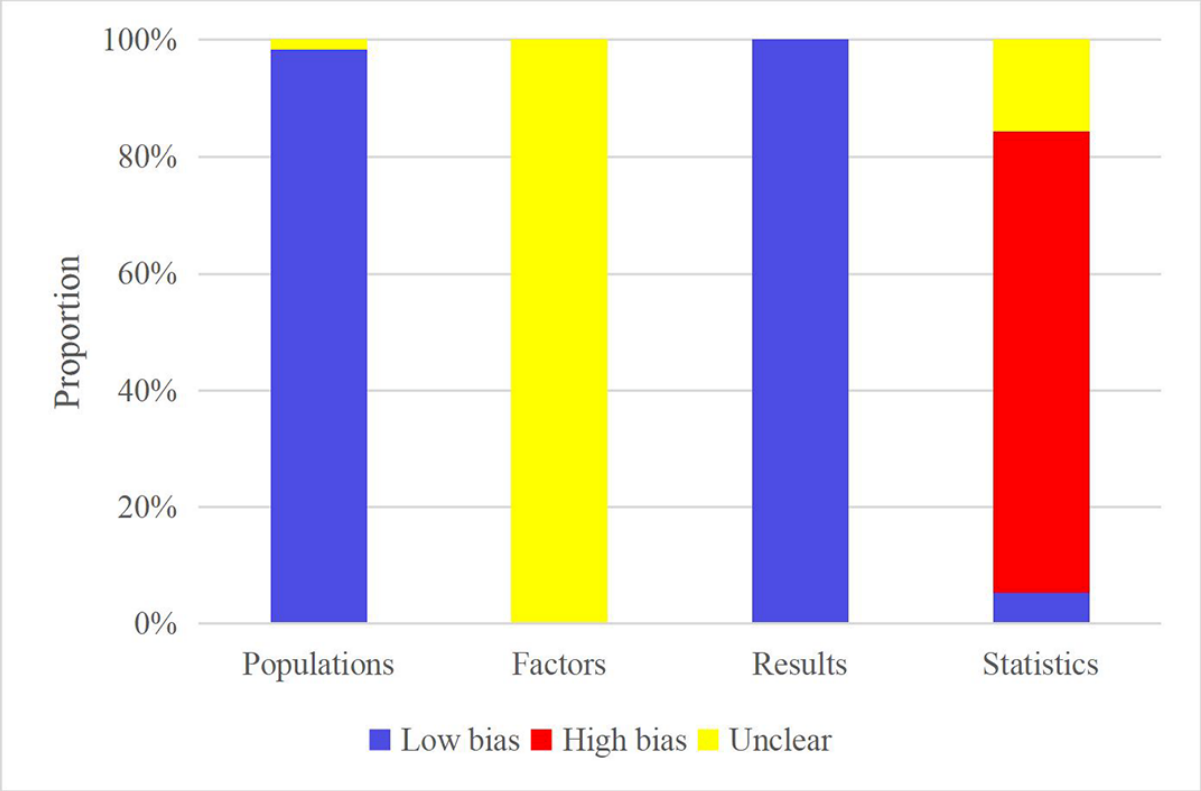

RQS

Two [32,38] studies did not provide complete imaging protocols. Multiple image segmentation was not performed in 7 [30-33,37,60,81] studies. All 57 studies failed to address scanner variability and vendor-dependent features, as well as imaging across multiple timepoints (ie, collecting individual images at different times). Three [32,37,69] studies did not use multiple testing correction or feature reduction to mitigate overfitting risk, nor did they use different data reduction methods (eg, principal component analysis and least absolute shrinkage and selection operator) for cross-analysis to reduce overfitting. Forty-two [29-31,33-38,40-50,55,56,58,60,62-65,67,69-73,75,78-84] studies did not perform multivariate analysis incorporating nonradiomic features. In all 57 [28-84] studies, biological correlations were neither detected nor discussed. The demonstration of phenotypic differences (potentially related to underlying gene-protein expression patterns) enhances the understanding of biology and radiomics. A cutoff value for disease presence or prediction risk was not provided in 51 [29,30,32-34,36-38,40-51,53-61,63-84] studies. Discrimination statistics were absent in 2 [32,69] studies. The reporting of these statistics (eg, ROC curve, C-statistic, and area under the curve [AUC]) and related statistical significance (eg, *P* values and CIs), or the application of resampling methods (eg, cross-validation and bootstrapping) was not performed. Calibration statistics were not reported in 47 [29-38,42-45,48,49,50-77,78,80,82,83] studies. Neither calibration statistics (eg, calibration plots, calibration in the large or slope) nor related statistical significance (eg, *P* values and CIs), or the application of resampling methods (eg, cross-validation and bootstrapping) were reported. Fifty-six [28-36,38-84] studies were not prospective studies registered in trial databases. Validation was lacking in 4 [33,38,67,73] studies. Comparisons with a "gold standard"—assessing the model's agreement with the "gold standard" methods—were not performed in all 57 studies. Potential clinical utility—reporting potential and the applications of the model in a clinical setting (eg, decision curve analysis)—was not addressed in 37 [30,32,33,36-39,42-45,48,49,54-73,78,80,82,83] studies. Cost-effectiveness analysis—reporting the cost-effectiveness of clinical application (eg, clinical impact curve)—was not performed in 56 [28-38,40-84] studies. Only 3 [28,32,35] studies did not adhere to open science and data principles by failing to release code and data. Regarding the RQS, a penalty of 3 points was applied for the absence of measures to control overfitting, and a deduction of 5 points was incurred for the lack of a validation set. Conversely, a bonus of 7 points was awarded to studies that were prospectively registered in a trial database. Among the 57 eligible investigations, 4 lacked any validation set. External validation was implemented in 16 [29-32,42,47,50,52,58-60,77,78,80,81,83] studies. Prospective registration was documented in only a single [37] study. The average quality score was 11.60 (range 1-17, IQR 3.5), representing an average percentage of 32.21% (Table S1 in Multimedia Appendix 2).

Meta-Analysis

Meta-Analysis for AUROC for Identifying Muscle Invasion

The meta-analysis of AUROC was conducted using a random-effects model. In the training sets, the pooled AUROC

XSL•FO
RenderX

for radiomics-based ML was 0.932 (95% CI 0.907-0.957; grade: moderate Figure 4 [28,29,34,41,42,45,46,48,50-52,54,57,60,62,63,65,69,71,75,79,80,82]).

The pooled AUROC was 0.932 (95% CI 0.853-1.000; grade: weak), 0.907 (95% CI 0.863-0.954; grade: weak), and 0.940 (95% CI 0.897-0.985; grade: weak) for CT-, MRI-, and ultrasound-based radiomics, respectively (Table 2; Figures S1 and S2 in Multimedia Appendix 3). The pooled AUROC was 0.914 (95% CI 0.871-0.959; grade: weak) and 0.934 (95% CI 0.908-0.960; grade: moderate) for models integrating clinical features with CT- or MRI-based radiomics (Table 2; Figures S3 and S4 in Multimedia Appendix 3).

Figure 4. Summary plot of radiomics for detecting muscle invasion in the training set [28,29,34,41,42,45,46,48,50-52,54,57,60,62,63,65,69,71,75,79,80,82]. For the 3 studies that included both radiomics-based and combined radiomics-clinical models, only the best-performing model in the validation set was retained to ensure independence across studies.

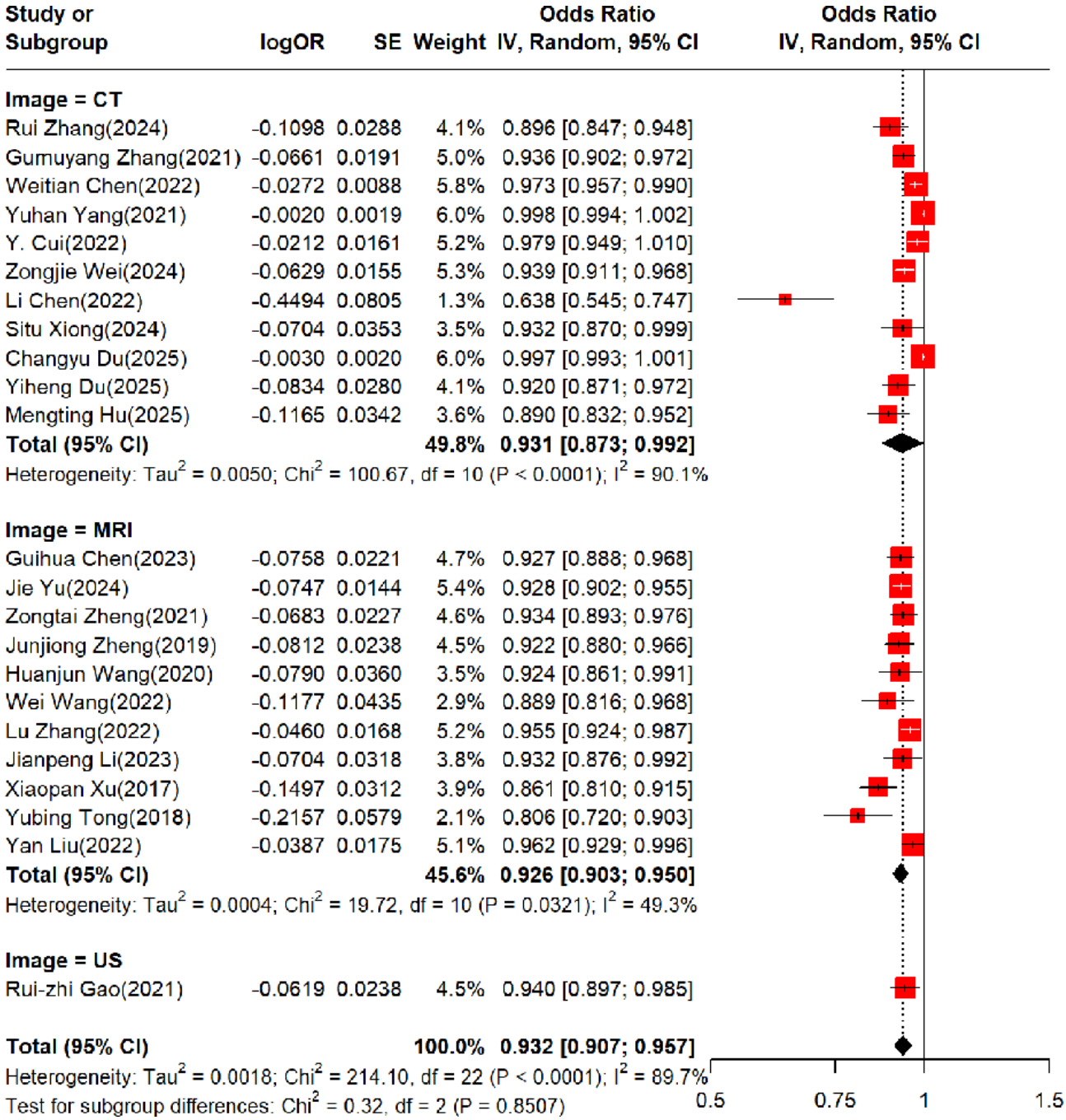

XSL•FO
RenderX

Table . Meta-analysis results for the identification of area under the receiver operating characteristic curve (AUROC) in muscle invasion.

Subgroup	Training set						Validation set					
	Number of models	AUROC[a] (95% CI)	tau^2	tau	I^2	Grade[b]	Number of models	AUROC (95% CI)	tau^2	tau	I^2	Grade
Radiomics												
CT[c]												
Extra trees							1	0.825 (0.712-0.956)				⊕⊕
ANN[d]							1	0.950 (0.896-1.000)				⊕⊕
DL[e]	6	0.970 (0.940-1.000)	0.0007	0.0269	86	⊕⊕⊕	5	0.903 (0.826-0.988)	0.0047	0.0687	88.1	⊕⊕
LR[f]	2	0.938 (0.513-1.000)	0.0038	0.0619	84.3	⊕	2	0.891 (0.842-0.942)	0	0	0	⊕⊕
NB[g]	1	0.638 (0.545-0.747)				⊕	1	0.665 (0.522-0.847)	——			⊕
Overall	9	0.932 (0.853-1.000)	0.0083	0.0914	89.8	⊕⊕	10	0.893 (0.840-0.948)	0.004	0.0634	85.4	⊕⊕
MRI[h]												
DL	2	0.929 (0.910-0.948)	0	0	0	⊕⊕⊕	5	0.908 (0.892-0.924)	0	0	0	⊕⊕
LASSO[i]	2	0.951 (0.793-1.000)	<0.0001	0.005	5.8	⊕⊕	2	0.906 (0.900-0.913)	0	0	0	⊕⊕⊕
LR	1	0.933 (0.895-0.972)				⊕⊕	1	0.931 (0.847-1.000)				⊕⊕
RF[j]							1	0.907 (0.856-0.961)				⊕⊕
SVM[k]	3	0.840 (0.772-0.914)	0	0	0	⊕⊕	2	0.913 (0.299-1.000)	0.0132	0.1151	85.4	⊕
Overall	8	0.907 (0.863-0.954)	0.0022	0.047	71	⊕⊕	11	0.916 (0.891-0.942)	0.0001	0.01	0	⊕⊕
Ultrasonography	1	0.940 (0.897--0.985)				⊕⊕	1	0.840 (0.737-0.958)				⊕⊕
Radiomics+clinical features												
CT	3	0.914 (0.871-0.959)	0	0	0	⊕⊕	4	0.874 (0.852-0.896)	0	0	0	⊕⊕

XSL•FO
RenderX

Subgroup	Training set						Validation set					
	Number of models	AUROC[a] (95% CI)	tau^2	tau	I^2	Grade[b]	Number of models	AUROC (95% CI)	tau^2	tau	I^2	Grade
MRI	5	0.934 (0.908-0.960)	<0.0001	0.0077	0	⊕⊕⊕	5	0.921 (0.867-0.979)	0.0005	0.0228	19.8	⊕⊕

[a]AUROC: area under the receiver operating characteristic curve.

[b]In the GRADE assessment section, ⊕⊕⊕⊕ represents high quality of evidence, ⊕⊕⊕ represents moderate quality of evidence, ⊕⊕ represents low quality of evidence, and ⊕ represents very low quality of evidence.

[c]CT: computed tomography.

[d]ANN: artificial neural network.

[e]DL: deep learning.

[f]LR: logistic regression.

[g]NB: naive Bayes.

[h]MRI: magnetic resonance imaging.

[i]LASSO: least absolute shrinkage and selection operator.

[j]RF: random forest.

[k]SVM: support vector machine.

In the validation sets, the pooled AUROC for radiomics-based ML was 0.912 (95% CI 0.891-0.934; grade: weak Figure 5 [28-31,34,41,43,45,46,48,50,51,54,57,58,60,62,65,66,69,75,78-80,82,83]).

The pooled AUROC was 0.893 (95% CI 0.840-0.948; grade: weak), 0.916 (95% CI 0.891-0.942; grade: weak), and 0.840 (95% CI 0.737-0.958; grade: weak) for CT-, MRI-, and ultrasound-based radiomics. The pooled AUROC was 0.874 (95% CI 0.852-0.896; grade: weak) and 0.921 (95% CI 0.867-0.979; grade: weak) for models integrating clinical features with CT- or MRI-based radiomics (Table 2 and Figures S5-S8 in Multimedia Appendix 3).

XSL•FO
RenderX

Figure 5. Summary plot of radiomics for detecting muscle invasion in the validation set [28-31,34,41,43,45,46,48,50,51,54,57,58,60,62,65,66,69,75,78-80,82,83].

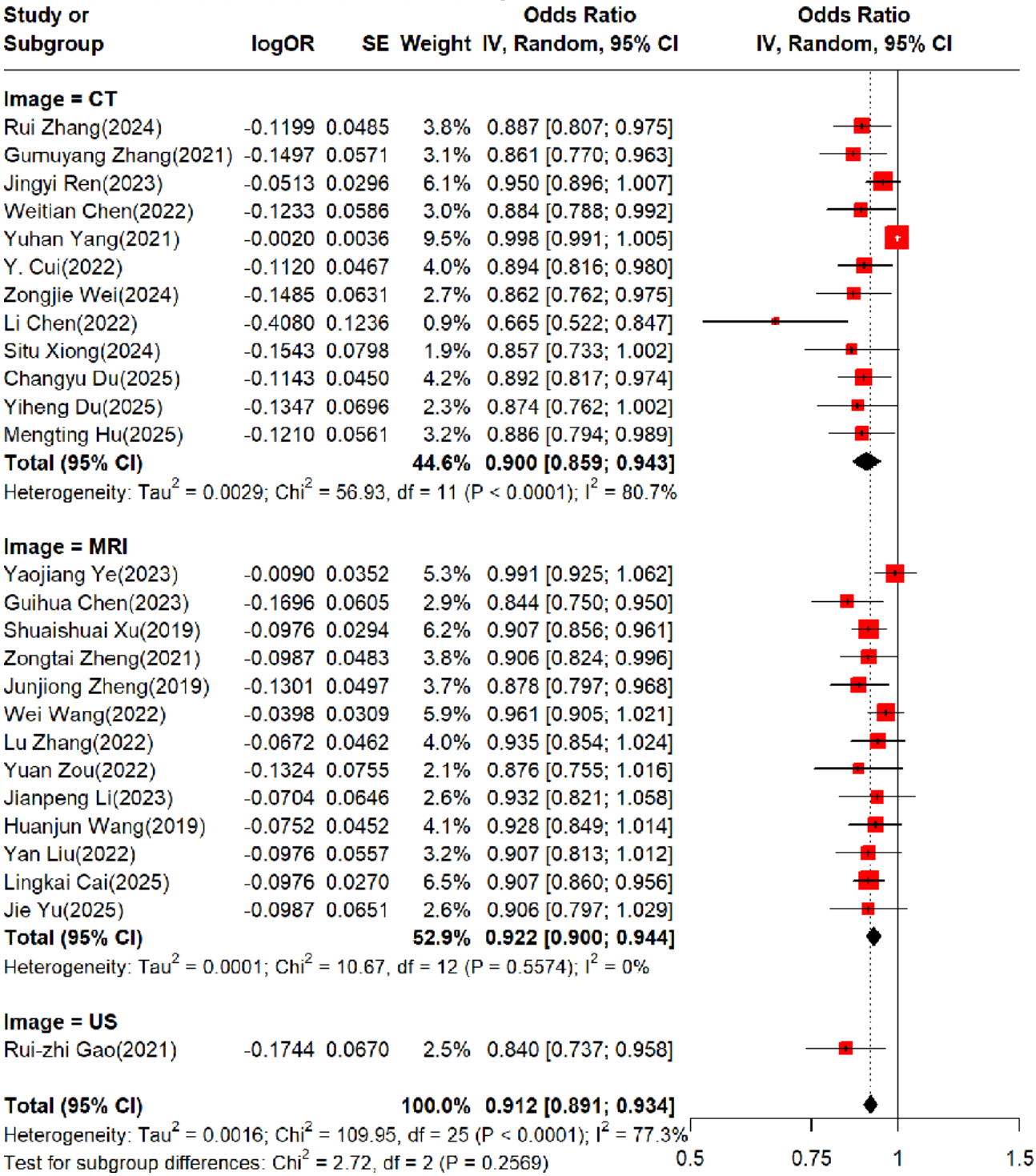

Meta-Analysis for Sensitivity and Specificity in Muscle Invasion Identification

The meta-analysis of 2x2 diagnostic contingency tables was conducted via a bivariate mixed effects model. In the training sets, the pooled sensitivity and specificity were 0.92 (95% CI 0.84-0.96) and 0.88 (95% CI 0.83-0.92), 0.80 (95% CI 0.76-0.84) and 0.90 (95% CI 0.86-0.92), and 0.81 and 0.91 for CT-, MRI-, and ultrasound-based radiomics. The pooled sensitivity and specificity were 0.79 - 0.93, 0.58-0.92, and 0.86 (95% CI 0.79-0.91) and 0.90 (95% CI 0.86-0.92) for models integrating clinical features with CT- or MRI-based radiomics (Table 3).

In the validation sets, the pooled sensitivity and specificity were 0.87 (95% CI 0.80-0.91) and 0.83 (95% CI 0.76-0.88), and 0.87 (95% CI 0.79-0.92) and 0.89 (95% CI 0.84-0.92) for CT- and MRI-based radiomics. The pooled sensitivity and specificity were 0.86 (95% CI 0.77-0.92) and 0.75 (95% CI 0.64-0.84) for models integrating clinical features with CT-based radiomics, and 0.88 (95% CI 0.81-0.93) and 0.88 (95% CI 0.76-0.94) for

XSL•FO
RenderX

models integrating clinical features with MRI-based radiomics (Table 3).

Table . Meta-analysis results for sensitivity and specificity in the identification of muscle invasion.

Subgroup	Training set			Validation set		
	Number of models	Sensitivity (95% CI)	Specificity (95% CI)	Number of models	Sensitivity (95% CI)	Specificity (95% CI)
Radiomics						
CT[a]						
Extra trees	1	0.82	0.81	1	0.81	0.8
ANN[b]				1	0.9	0.93
DL[c]	4	0.95 (0.85-0.99)	0.89 (0.82-0.93)	5	0.88 (0.77-0.94)	0.83 (0.71-0.91)
LR[d]	2	0.76 - 0.95	0.86 - 0.95	3	0.74 - 0.95	0.72 - 0.88
Overall	7	0.92 (0.84-0.96)	0.88 (0.83-0.92)	10	0.87 (0.80-0.91)	0.83 (0.76-0.88)
MRI[e]						
DL	2	0.75 - 0.90	0.87 - 0.93	4	0.87 (0.78-0.92)	0.93 (0.88-0.96)
LASSO[f]	2	0.80 - 0.92	0.93 - 0.90	2	0.92 - 0.94	0.78 - 0.87
LR	2	0.82 - 0.83	0.84 - 0.90	1	0.81	0.83
RF[g]				1	0.87	0.78
SVM[h]	2	0.74 - 0.80	0.75 - 0.82	3	0.67 - 1	0.88 - 0.96
Overall	8	0.80 (0.76-0.84)	0.90 (0.86-0.92)	11	0.87 (0.79-0.92)	0.89 (0.84-0.92)
Ultrasound	1	0.81	0.91	__[i]	—	—
Radiomics+ Clinical features						
CT						
Overall	3	0.79 - 0.93	0.58 - 0.92	4	0.86 (0.77-0.92)	0.75 (0.64-0.84)
MRI						
LR	4	0.86 (0.79-0.91)	0.90 (0.86-0.92)	5	0.88 (0.78-0.94)	0.91 (0.79-0.96)
Overall	4	0.86 (0.79-0.91)	0.90 (0.86-0.92)	6	0.88 (0.81-0.93)	0.88 (0.76-0.94)

[a]CT: computed tomography.
[b]ANN: artificial neural network.
[c]DL: deep learning.
[d]LR: logistic regression.
[e]MRI: magnetic resonance imaging.
[f]LASSO: least absolute shrinkage and selection operator.
[g]RF: random forest.
[h]SVM: support vector machine.
[i]Not available.

Subgroup Analysis

Subgroup analyses of AUROC for the diagnosis of muscle invasion were performed based on different imaging modalities and model types. Among CT-based radiomics models, deep learning (DL) was the predominant approach. In the validation sets, the pooled AUROC of DL was 0.903 (95% CI 0.826-0.988; grade: weak), with a sensitivity of 0.88 (95% CI 0.77-0.94) and a specificity of 0.83 (95% CI 0.71-0.91). Among MRI-based radiomics models, DL was the most common modeling approach. In the validation sets, the pooled AUROC of DL was 0.908 (95% CI 0.892-0.924; grade: weak), with a sensitivity of 0.87 (95% CI 0.78-0.92) and specificity of 0.93 (95% CI 0.88-0.96). In the validation sets, the model integrating CT-based radiomics with clinical characteristics yielded an AUROC of 0.874 (95% CI 0.852-0.896; grade: weak). The corresponding sensitivity was 0.86 (95% CI 0.77-0.92), and the specificity was 0.75 (95% CI 0.64-0.84). Conversely, for

XSL•FO
RenderX

MRI-based models combined with clinical variables, the pooled AUROC was 0.921 (95% CI 0.867-0.979; grade: weak). The corresponding sensitivity and specificity were 0.88 (95% CI 0.81-0.93) and 0.88 (95% CI 0.76-0.94), respectively (Tables 2-3).

Diagnosis of High-Grade Tumors

Pooled Results

A meta-analysis of AUROC was conducted using a random-effects model. In the training sets, the pooled AUROC for radiomics-based ML was 0.905 (95% CI 0.868-0.944; grade: weak; Figure 6 [39,40,44,47,49,53,56,65-67,70,73,84]).

Figure 6. Summary plot of radiomics for identifying high-grade tumors in the training set [39,40,44,47,49,53,56,65-67,70,73,84].

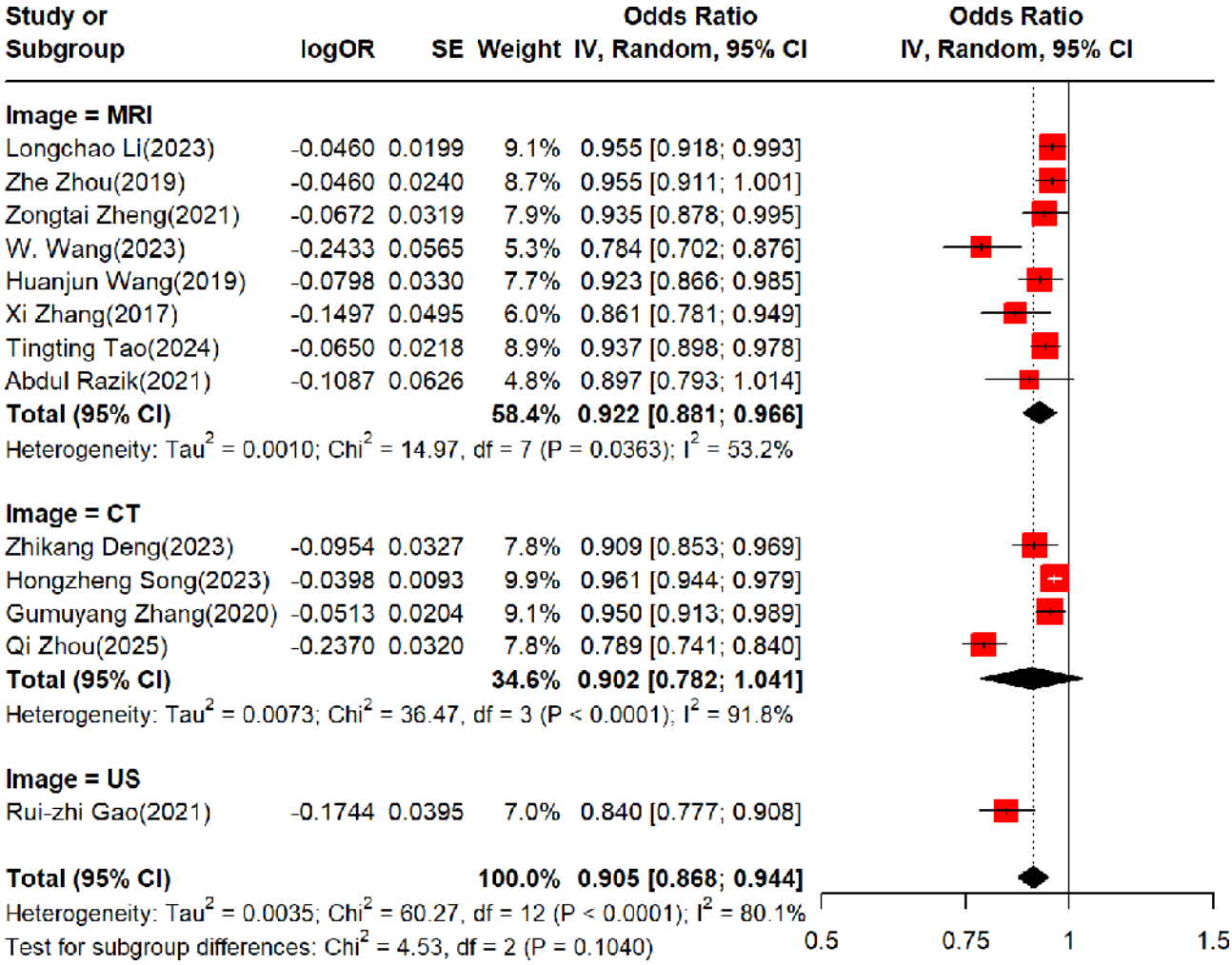

The pooled AUROC was 0.902 (95% CI 0.782-1.000; grade: weak) for CT-based radiomics, 0.895 (95% CI 0.814-0.985; grade: weak) for MRI-based radiomics, and 0.840 (95% CI 0.777-0.908; grade: weak) for ultrasound-based radiomics (Figures S9 and S10 in Multimedia Appendix 3 and Table 3). The AUROC for MRI-based radiomics combined with clinical features was 0.944 (95% CI 0.904-0.985; grade: moderate; Figure S11 in Multimedia Appendix 3).

In the validation sets, the pooled AUROC for radiomics-based ML was 0.872 (95% CI 0.823-0.924; grade: weak; Figure 7 [36,39,40,44,47,49,53,56,65,84]).

XSL•FO
RenderX

Figure 7. Summary plot of radiomics for identifying high-grade tumors in the validation set [36,39,40,44,47,49,53,56,65,84].

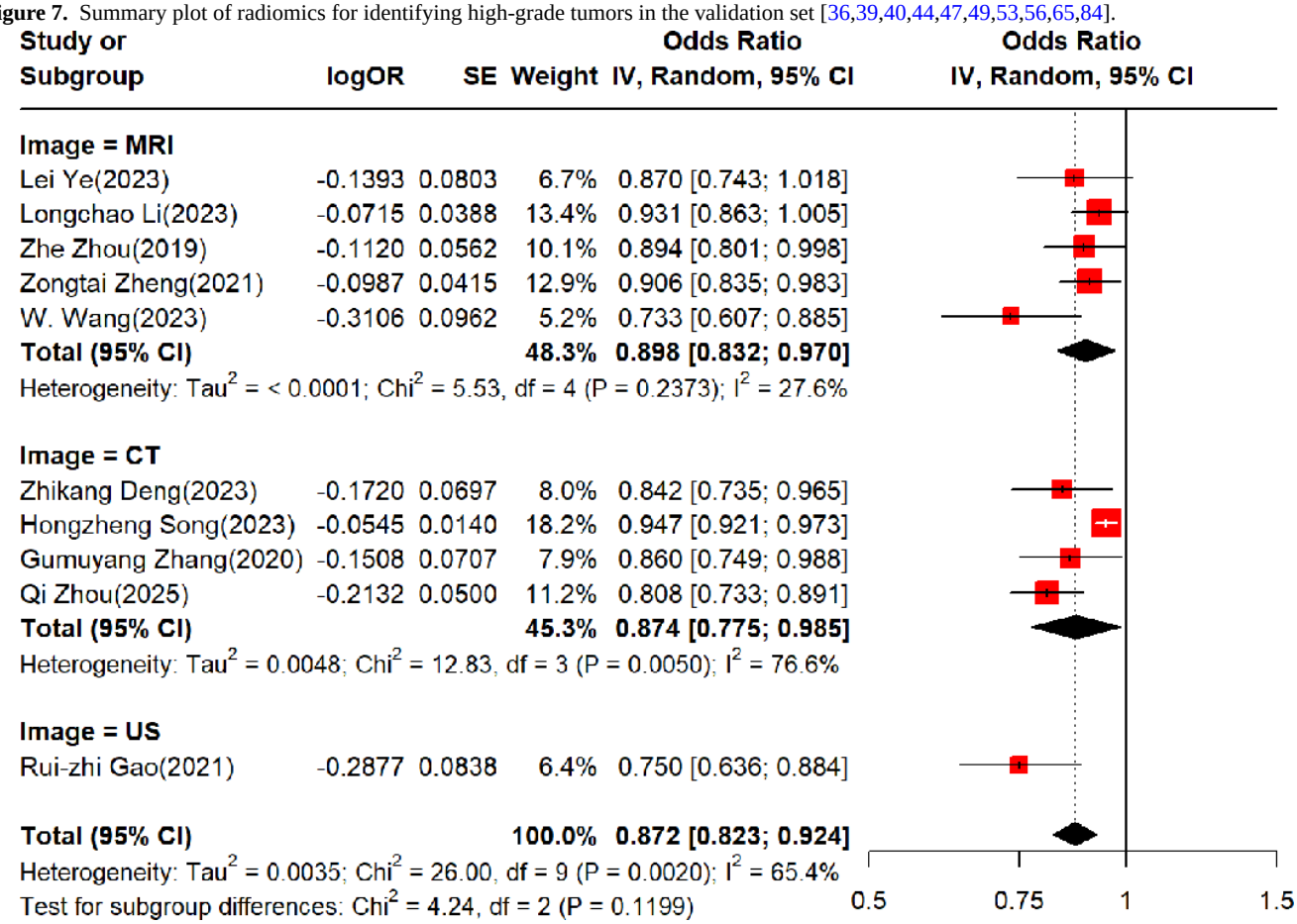

The pooled AUROC was 0.874 (95% CI 0.775-0.985; grade: weak) for CT-based radiomics, 0.846 (95% CI 0.663-1.000; grade: very low) for MRI-based radiomics, and 0.750 (95% CI 0.636-0.884; grade: very low) for ultrasound-based radiomics. (Figures S12-S13 in Multimedia Appendix 3). The AUROC for MRI-based radiomics combined with clinical features was 0.919 (95% CI 0.774-1.000; grade: weak; Table 4 and Figure S14 in Multimedia Appendix 3).

XSL•FO
RenderX

Table . Meta-analysis results for area under the receiver operating characteristic curve (AUROC) in the identification of high-grade tumors.

Subgroup	Training set						Validation set					
	No. of models	AUROC (95% CI)	tau^2	tau	I^2	Grade[a]	No. of models	AUROC (95% CI)	tau^2	tau	I^2	Grade
Radiomics												
CT[b]												
DL[c]	1	0.961 (0.944-0.979)				⊕⊕⊕	1	0.947 (0.921-0.973)				⊕⊕⊕
LR[d]	1	0.950 (0.913-0.989)				⊕⊕⊕	1	0.860 (0.749-0.988)				⊕⊕
SVM[e]	1	0.909 (0.853-0.969)				⊕⊕	2	0.819 (0.639-1.000)	0	0	0	⊕
AdaBoost	1	0.789 (0.741-0.840)				⊕⊕						
Overall	4	0.902 (0.782-1.000)	0.0073	0.0854	91.8	⊕⊕	4	0.874 (0.775-0.985)	0.0048	0.0689	76.6	⊕⊕
MRI[f]												
DL	1	0.937 (0.898-0.978)				⊕⊕						
LR	3	0.882 (0.685-1.000)	0.0084	0.0915	81.1	⊕	3	0.846 (0.663-1.000)	0.0031	0.0555	38.3	⊕
SVM	1	0.861 (0.781-0.949)				⊕⊕						
Overall	5	0.895 (0.814-0.985)	0.004	0.0633	69.3	⊕⊕	3	0.846 (0.663-1.000)	0.0031	0.0555	38.3	⊕
Ultrasound	1	0.840 (0.777-0.908)				⊕⊕	1	0.750 (0.636-0.884)			NA	⊕
Radiomics+clinical features												
CT												
MRI	3	0.944 (0.904-0.985)	0	0	0	⊕⊕⊕	2	0.919 (0.774-1.000)	0	0	0	⊕⊕

[a]In the GRADE assessment section, ⊕⊕⊕⊕ represents high quality of evidence, ⊕⊕⊕ represents moderate quality of evidence, ⊕⊕ represents low quality of evidence, and ⊕ represents very low quality of evidence.

[b]CT: computed tomography.

[c]DL: deep learning.

[d]LR: logistic regression.

[e]SVM: support vector machine.

[f]MRI: magnetic resonance imaging.

XSL•FO
RenderX

A meta-analysis of 2×2 diagnostic contingency tables was performed via a bivariate mixed-effects model. In the training sets, the pooled sensitivity and specificity were 0.91 (95% CI 0.84-0.95) and 0.74 (95% CI 0.67-0.79) for CT-based radiomics, and 0.86 (95% CI 0.57-0.96) and 0.89 (95% CI 0.83-0.93) for MRI-based radiomics. The sensitivity and specificity were 0.78 (95% CI 0.65-0.87) and 0.94 (95% CI 0.79-0.99) for MRI-based radiomics combined with clinical features. In the validation sets, the pooled sensitivity and specificity were 0.65 - 0.95 and 0.69 - 0.73 for CT-based radiomics, and 0.81~0.89 and 0.70~0.91 for MRI-based radiomics. The sensitivity and specificity were 0.54 - 0.77 and 0.88 - 0.98 for MRI-based radiomics combined with clinical features (Table 5).

Table . Meta-analysis results for sensitivity and specificity in identification of high-grade tumors.

Subgroup	Training set			Validation set		
	Number of models	Sensitivity (95% CI)	Specificity (95% CI)	Number of models	Sensitivity (95% CI)	Specificity (95% CI)
Radiomics						
CT[a]						
LR[b]	1	0.96	0.71	1	0.89	0.73
NB[c]	1	0.89	0.77			
SVM[d]	1	0.93	0.77	2	0.65 - 0.95	0.69 - 0.72
AdaBoost	1	0.85	0.73			
Overall	4	0.91 (0.84-0.95)	0.74 (0.67-0.79)	3	0.65 - 0.95	0.69 - 0.73
MRI[e]						
DL[f]	1	0.99	0.9			
LR	2	0.60 - 0.74	0.82 - 1	3	0.81 - 0.89	0.70 - 0.91
SVM	1	0.78	0.87			
XGBoost						
Overall	4	0.86 (0.57-0.96)	0.89 (0.83-0.93)	3	0.81 - 0.89	0.70 - 0.91
Ultrasound	1	0.8	0.76			
Radiomics+ Clinical features						
MRI						
Overall	4	0.78 (0.65-0.87)	0.94 (0.79-0.99)	2	0.54 - 0.77	0.88 - 0.98

[a]CT: computed tomography.
[b]LR: logistic regression.
[c]NB: naive Bayes.
[d]SVM: support vector machine.
[e]MRI: magnetic resonance imaging.
[f]DL: deep learning.

Subgroup Analysis

Subgroup analyses of AUROC for the diagnosis of high-grade tumors were performed based on different imaging modalities and model types. In CT-based radiomics, the support vector machine (SVM) model was the predominant type. The pooled AUROC of SVM was 0.819 (95% CI 0.639-1.000; grade: very low) in the validation sets, with corresponding sensitivity of 0.65 - 0.95 and specificity of 0.69 - 0.72. Likewise, the LR model was the predominant type in MRI-based radiomics. The pooled AUROC of LR was 0.846 (95% CI 0.663-1.000; grade: very low) in the validation sets, with corresponding sensitivity of 0.81-0.89 and specificity of 0.70-0.91. In the validation sets, the model integrating MRI-based radiomics with clinical characteristics achieved an AUROC of 0.919 (95% CI 0.774-1.000; grade: weak). The corresponding sensitivity was 0.54 - 0.77, and the specificity was 0.88 - 0.98. Given the paucity of available studies using CT-based radiomics combined with clinical feature models, subgroup analyses could not be performed (Tables 4 and 5).

Other Risk Stratification Outcomes

HER2 is encoded by the ERBB2 gene. It plays a critical role in the development of various malignancies, including breast, gastric, bladder, ovarian, and lung cancers. HER2-targeted therapies have become first-line treatments for patients with

XSL•FO
RenderX

advanced cancers exhibiting HER2 overexpression. Three studies reported on radiomics-based approaches to assess HER2 expression in bladder cancer. Yu et al [55] used MRI radiomics-based ML for noninvasive assessment of HER2. In their test set, an SVM model demonstrated an AUC of 0.712 (95% CI 0.535-0.889), with a sensitivity of 0.857 and a specificity of 0.533. They highlighted the potential value of this approach in cases where patients cannot undergo invasive procedures such as biopsies or diagnostic TUR. Based on contrast-enhanced computed tomography (CE-CT), Peng et al [74] assessed a clinical-radiomics model in evaluating HER2 status in urothelial bladder cancer. This CE-CT-based model exhibited the highest effectiveness in forecasting HER2 status, with an AUC of 0.814 (95% CI 0.642-0.986) in the test set. Wei et al [77] conducted a multicenter study using explainable ML based on CT radiomics to preoperatively diagnose HER2 status in bladder cancer. In their test set, AUCs were 0.803, 0.709, 0.679, 0.794, and 0.815 for LR, SVM, KNN, XGBoost, and RF models, respectively. These studies suggest that radiomics-based ML has promising potential for detecting HER2 expression in bladder cancer.

Ki-67 expression is associated with a poor prognosis and advanced clinicopathological features in cancer. Two single-center studies [35,76] explored radiomics-based prediction of Ki-67 expression in bladder cancer, with the potential to improve prognostic assessment and clinical decision-making. Zheng et al [35] reported on an MRI-based radiomics study. Their SMOTE-LASSO model achieved an AUC of 0.819 (0.658-0.980) in the validation set, with a sensitivity of 0.795 and a specificity of 0.867%. Feng et al [76] developed a radiomics nomogram based on CE-CT, which demonstrated AUCs of 0.836 and 0.887 in their validation set. These findings indicate that radiomics-based ML holds promise for detecting Ki-67 expression in bladder cancer.

Additionally, 2 [37,68] studies reported on the use of radiomics to predict LN staging. Gresser et al [68] used CT radiomics to evaluate LN staging in bladder cancer. Their combined model incorporated manually segmented radiomic features and radiologist assessment. The model achieved an AUC of 0.81 (0.71-0.92), a sensitivity of 0.73 (0.55-0.88), and a specificity of 0.84 (0.76-0.92), respectively, in the test set. However, Starmans et al [37] conducted a multicenter study using preoperative CT radiomics. They found that in patients with cT2-T4aN0-N1M0 muscle-invasive bladder cancer (MIBC), radiomics was not helpful in differentiating pN+ and pN0 disease. This suggests that further research is needed to explore the effectiveness of radiomics-based ML in detecting LN staging in bladder cancer.

Small-Study Effect Analysis

Funnel plots were generated to evaluate small-study effects in radiomics-based ML models for detecting muscle invasion and high-grade tumors. Significant small-study effects were observed in both the training and validation sets for assessing muscle invasion ($P<.05$; Figures S15-S16 in Multimedia Appendix 3), as well as for identifying high-grade tumors ($P<.05$; Figures S17–S18 in Multimedia Appendix 3).

Meta-Regression Analysis

For the tasks of detecting muscle invasion, meta-regression was performed. Independent variables included case number, model type (LR [Reference] vs other ML), and modeling variables (Radiomics [Reference] vs. Radiomics +clinical). Separate analyses were conducted for models developed on CT or MRI in the training set and validated on CT or MRI in the validation set. The results indicated a significant association between case number and the AUC only in the validation set of MRI-based models for detecting muscle invasion ($P<.05$). No other variables demonstrated a significant influence ($P>.05$; Table S2 in Multimedia Appendix 4 and Table S3 in Multimedia Appendix 5). Meta-regression was not conducted for pathological grade due to an insufficient number of studies.

Discussion

Summary of the Main Findings

This systematic review incorporated 57 studies [28-84] to evaluate radiomics-based ML for preoperative risk stratification in bladder cancer. The findings indicated a high AUROC for detecting both muscle invasion and high-grade tumors. Additionally, radiomics showed potential for identifying HER2 and Ki-67 expression. However, evaluation via the RQS revealed an overall low methodological quality among the eligible studies. PROBAST-AI assessment revealed that the primary source of bias was in the model evaluation phase, primarily due to small validation set sizes or an absence of external validation. Consequently, the findings of this review should be interpreted with caution. Based on the present findings, radiomics is considered to have application potential. Nevertheless, the current evidence faces significant challenges, including methodological shortcomings and a high ROB, which currently preclude its readiness for clinical translation.

Comparison With Previous Reviews

Previous studies have examined the application of radiomics in bladder cancer. Kozikowski et al [10] reviewed the prediction ability of radiomics in muscle-invasive cancer, ultimately including eight articles. The pooled estimated sensitivity and specificity were 82% (95% CI 77% - 86%) and 81% (95% CI 76% - 85%). However, the study did not classify or analyze CT and MRI image sources separately nor attempt to combine different ML approaches with clinical radiomics models. Boca et al [14] reviewed MRI-based radiomics studies in bladder cancer, ultimately including 26 articles with 2991 participants. Radiomics in these studies was primarily used for preoperative prediction of tumor stage or molecular correlations (n=9), preoperative tumor grading (n=13), and prediction of prognosis or response to neoadjuvant therapy (n=4). Most radiomics models incorporated second-order features from filtered images, with quality scores ranging from 8.33% to 52.77% [14]. However, the study only discussed MRI-based radiomics and failed to consider other imaging modalities. The study also did not analyze sensitivity, specificity, or dataset differences, nor did it incorporate a discussion of radiomics combined with clinical features. Building on these prior efforts, the current study conducted a more systematic and comprehensive review of the current state of radiomics-based ML in bladder cancer.

XSL•FO
RenderX

The current study also considered different image features, accounted for various image sources, analyzed different modeling approaches, and examined the detection performance across diverse datasets. This provides more comprehensive evidence for future developments.

In the clinical management of bladder cancer, CT often struggles to accurately diagnose flat lesions and prostate-adjacent bladder base lesions, especially in patients with benign prostatic hyperplasia. This difficulty arises from the challenge of distinguishing tumor recurrence from inflammatory wall thickening following intravesical chemotherapy, as well as from scar tissue after TURBT [85]. MRI assessment is a laborious, slice-by-slice process, with its effectiveness depending on the radiologist's experience [86]. The integration of radiomics and ML with bladder MRI holds promise for improving staging and treatment response assessment. Bladder cancer management guidelines would be enhanced by the integration of MRI into their staging strategies [87]. However, radiomics may experience similar slow progress as molecular biology-based diagnostic and therapeutic techniques. This can be attributed to several factors, such as technical complexities, a lack of validation standards, and inadequate study design (particularly conflating hypothesis generation with hypothesis testing) [11]. Other contributing factors include data overfitting, incomplete reporting of results, and unidentified confounding variables in the datasets used (especially in retrospective datasets). Therefore, as with all biomarker studies, retrospective radiomics studies require validation on independent datasets, ideally from another institution [11].

In clinical practice, MRI has become the most accurate imaging modality for evaluating local invasiveness in bladder cancer. It can be used to assess regional LN involvement and tumor spread to pelvic bones and the upper urinary tract [21,88]. While bladder cancer appears as a soft tissue lesion on CT scans, it is more easily identified as a filling defect during CT urography. Ultrasound is a dynamic imaging modality. It can distinguish bladder cancer from other conditions that appear similar. Its ability to detect blood flow aids in distinguishing the solid tissue of the tumor from blood clots or debris [89]. This meta-analysis reveals that in studies examining muscle invasion, the primary image sources are CT, MRI, and, to a lesser extent, ultrasound. Radiomics combined with clinical features has not consistently demonstrated superior detection performance, potentially due to limited available data. The diagnostic effectiveness of radiomics warrants further investigation. In studies investigating histological grading, the primary image sources are CT and MRI. Only a few studies have explored combining radiomics with clinical features, resulting in limited current evidence.

Traditional ML-based radiomics requires substantial upfront effort, including texture extraction, manual image segmentation, and model construction. This process carries the risk of image data loss and imposes a significant workload. Manual segmentation, in particular, can be influenced by clinical experience, habits, and prior research [86,90]. Consequently, some researchers have explored DL, leveraging its capabilities for staging, grading, automated tumor detection, intelligent segmentation, bladder wall segmentation, and prediction of recurrence, response to chemotherapy, and overall survival. The goal is to improve disease management [58,91-94]. While external validation of intelligent segmentation accuracy is lacking, the potential of DL to reduce workload and automatically interpret images in an intelligent manner is encouraging the development of better, smarter tools. Of the studies included in this review, only a few DL models have demonstrated good performance. Thus, future research should explore DL in addition to traditional methods.

While the models in this study demonstrated relatively high AUROC values, the widespread absence of calibration metrics represents a major limitation. This gap substantially undermines the reliability of these models for practical application [95,96]. Dependence on discriminatory metrics such as AUC, sensitivity, and specificity only reflects the ability of a model to differentiate between positive and negative events. These metrics do not, however, inform on the accuracy of the predicted probabilities—a critical shortfall that poses challenges for safe integration into clinical decision support [97,98]. Consequently, future studies must report comprehensive calibration metrics, including calibration curves, calibration slope, intercept, and the Brier score. Furthermore, decision curve analysis should be performed to evaluate the clinical net benefit across different decision thresholds, a crucial step for assessing clinical utility.

Before initiating this meta-analysis, a prospective registration was completed. However, several adjustments were made during the actual research process. First, to mitigate the risk of missing newly published literature and ensure the completeness of evidence, supplementary database searches were conducted on October 17, 2025. Second, to better reflect the ROB in the included studies, PROBAST-AI was adopted in addition to RQS for the assessment of bias. Third, the GRADE approach was applied to evaluate the certainty of evidence. Fourth, subgroup analyses were carried out based on the type of dataset (training vs validation), imaging source, and type of model. Finally, analyses of small-study effects and meta-regression were also performed. The registered protocol will be updated accordingly.

Strengths and Limitations

This meta-analysis is the first comprehensive and systematic review of preoperative risk stratification for bladder cancer, summarizing the available evidence. However, it is subject to the following limitations. First, the present analysis incorporated only 16 studies [29-32,42,47,50,52,58-60,77,78,80,81,83] that used external validation. Among these, 13 investigations focused on muscle invasion, one examined histological grade, one assessed HER2 expression, and one addressed preoperative LN staging. The analytical strategy required subgrouping based on the imaging modality used for model development and the integration of clinical features. Incorporating the data source as a subgroup variable was precluded due to the limited number of multicenter datasets, which would have yielded statistically unreliable comparisons. Furthermore, the validation sets in this study were predominantly created via random sampling. The lack of independent external validation limits the interpretation of the results to some extent. Second, different mathematical models exhibit varying performance in processing images. Although subgroup analyses were conducted to account for these differences, the small number of models within each

XSL•FO
RenderX

subgroup may limit the interpretation of our findings. Therefore, the comparative performance of models within these subgroups should be interpreted with caution. Third, evidence for the preoperative identification of LN metastasis and Ki-67 expression is extremely limited. Fourth, while subgroup analyses were performed based on different model types, definitive conclusions cannot be drawn. However, given the paucity of available studies and the small number of articles included in specific model categories, definitive conclusions cannot be drawn. Fifth, the RQS and PROBAST-AI instruments incorporate elements requiring subjective judgment, which may introduce variability into the overall interpretation of the quality assessments. Sixth, a notable limitation of the current evidence base is the infrequent reporting of calibration metrics, such as the calibration-in-the-large, calibration slope, and Brier score. This omission considerably diminishes the clinical applicability of the model performance estimates. Seventh, while all eligible studies confirmed diagnoses pathologically, the transparency regarding the origin of pathological specimens was often insufficient. This lack of detail concerning specimen sourcing is a potential source of bias. Together, these limitations collectively curtail the generalizability of the study findings. Eighth, a notable small-study effect was observed across the included studies. Thus, the findings should be interpreted with caution.

Challenges and Future Directions

In summary, significant challenges remain prior to the clinical deployment of these models. First, current investigations have not rigorously examined the impact of imaging modality. Variations in acquisition protocols can influence image quality; however, the potential effect of such heterogeneity on modeling outcomes has not been adequately addressed in the existing literature. Second, independent external validation is essential for establishing the generalizability of models. The predominant reliance on internal validation in the eligible studies implies that these findings propose a potential methodological approach rather than one ready for clinical application. Substantial additional evidence is required to confirm clinical utility and generalizability. Third, sample size presents a major constraint. Only a few studies provided a sufficient number of cases, and most encompassed fewer than 200 subjects. Consequently, discussions regarding model robustness necessitate larger, more powerful datasets. Fourth, the image segmentation process depends heavily on manual delineation. This approach is time-consuming and susceptible to variability introduced by the operator's prior knowledge. Future work should prioritize developing and implementing DL-based automated or semiautomated segmentation techniques.

To facilitate successful clinical translation, future investigations should prioritize the following areas: conducting robust pilot studies, performing reproducibility analyses on different imaging protocols and segmentation methods, ensuring adequate sample sizes, implementing multicenter external validation, comparing diverse ML architectures, and establishing protocols for continuous model updating and refinement.

Conclusions

This is the first systematic review to comprehensively assess radiomics for preoperative risk stratification in bladder cancer. The findings provide evidence to support the development and refinement of future ML-based tools for image analysis. However, due to limitations in the current evidence, such as methodological flaws and a high ROB, low GRADE level, clinical translation is not yet warranted. Future research should standardize radiomics workflows, incorporate multicenter images from diverse geographical regions, and minimize the impact of imaging protocols and pre-processing steps. These measures are essential to advance radiomics toward successful clinical implementation. Such efforts are essential for fully elucidating and validating the potential of radiomics in the noninvasive diagnosis of bladder cancer.

Acknowledgments

This study was conducted without the assistance of artificial intelligence technologies in its writing process. The authors assume full responsibility and accountability for ensuring the precision, novelty, and scholarly integrity of every element within the manuscript, encompassing all referenced sources and citation practices.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: ZH, YL
Data curation: ZH, QJ, BX
Formal analysis: ZH, YL, QJ, FH
Investigation: ZH, YL, QJ, FH
Methodology: ZH, QJ, BX
Project administration: FH, BX
Resources: ZH, BX

XSL•FO
RenderX

Software: ZH, YL
Supervision: FH, BX
Validation: ZH, YL, BX
Visualization: ZH, YL
Writing – original draft: ZH, YL
Writing – review & editing: ZH, BX
All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Literature search strategy.
[DOCX File, 16 KB - jmir_v28i1e81084_app1.docx]

Multimedia Appendix 2
Quality assessment of included studies using the radiomics quality score (RQS) tool.
[DOCX File, 44 KB - jmir_v28i1e81084_app2.docx]

Multimedia Appendix 3
Forest and funnel plots for preoperative risk stratification using radiomics-based machine learning.
[DOCX File, 15727 KB - jmir_v28i1e81084_app3.docx]

Multimedia Appendix 4
Meta-regression of area under the curve (AUC) for machine learning models based on computed tomography (CT) and magnetic resonance imaging (MRI) radiomics in detecting muscle invasion (training set).
[DOCX File, 15 KB - jmir_v28i1e81084_app4.docx]

Multimedia Appendix 5
Meta-regression of area under the curve (AUC) for machine learning models based on computed tomography (CT) and magnetic resonance imaging (MRI) radiomics in detecting muscle invasion (validation set).
[DOCX File, 15 KB - jmir_v28i1e81084_app5.docx]

Checklist 1
PRISMA 2020 checklist.
[PDF File, 605 KB - jmir_v28i1e81084_app6.pdf]

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021 May;71(3):209-249. [doi: 10.3322/caac.21660] [Medline: 33538338]
2. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA Cancer J Clin 2024;74(1):12-49. [doi: 10.3322/caac.21820] [Medline: 38230766]
3. Babjuk M, Burger M, Capoun O, et al. European Association of Urology Guidelines on non-muscle-invasive bladder cancer (Ta, T1, and carcinoma in situ). Eur Urol 2022 Jan;81(1):75-94. [doi: 10.1016/j.eururo.2021.08.010] [Medline: 34511303]
4. Mariappan P, Zachou A, Grigor KM, Edinburgh Uro-Oncology Group. Detrusor muscle in the first, apparently complete transurethral resection of bladder tumour specimen is a surrogate marker of resection quality, predicts risk of early recurrence, and is dependent on operator experience. Eur Urol 2010 May;57(5):843-849. [doi: 10.1016/j.eururo.2009.05.047] [Medline: 19524354]
5. Poletajew S, Krajewski W, Kaczmarek K, et al. The learning curve for transurethral resection of bladder tumour: how many is enough to be independent, safe and effective surgeon? J Surg Educ 2020;77(4):978-985. [doi: 10.1016/j.jsurg.2020.02.010] [Medline: 32147466]
6. Svatek RS, Shariat SF, Novara G, et al. Discrepancy between clinical and pathological stage: external validation of the impact on prognosis in an international radical cystectomy cohort. BJU Int 2011 Mar;107(6):898-904. [doi: 10.1111/j.1464-410X.2010.09628.x] [Medline: 21244604]

XSL•FO
RenderX

7. Yanagisawa T, Kawada T, von Deimling M, et al. Repeat transurethral resection for non-muscle-invasive bladder cancer: an updated systematic review and meta-analysis in the contemporary era. Eur Urol Focus 2024 Jan;10(1):41-56. [doi: 10.1016/j.euf.2023.07.002] [Medline: 37495458]
8. Nouh MR, Ezz Eldin O. Precise vesical wall staging of bladder cancer in the era of precision medicine: has it been fulfilled? Abdom Radiol (NY) 2025 Jul;50(7):3084-3091. [doi: 10.1007/s00261-024-04786-8] [Medline: 39725735]
9. Guldhammer CS, Vásquez JL, Kristensen VM, et al. Cystoscopy accuracy in detecting bladder tumors: a prospective video-confirmed study. Cancers (Basel) 2023 Dec 28;16(1):160. [doi: 10.3390/cancers16010160] [Medline: 38201586]
10. Kozikowski M, Suarez-Ibarrola R, Osiecki R, et al. Role of radiomics in the prediction of muscle-invasive bladder cancer: a systematic review and meta-analysis. Eur Urol Focus 2022 May;8(3):728-738. [doi: 10.1016/j.euf.2021.05.005] [Medline: 34099417]
11. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology 2016 Feb;278(2):563-577. [doi: 10.1148/radiol.2015151169] [Medline: 26579733]
12. Kwong JCC, Wu J, Malik S, et al. Predicting non-muscle invasive bladder cancer outcomes using artificial intelligence: a systematic review using APPRAISE-AI. NPJ Digit Med 2024 Apr 18;7(1):98. [doi: 10.1038/s41746-024-01088-7] [Medline: 38637674]
13. Gelikman DG, Rais-Bahrami S, Pinto PA, Turkbey B. AI-powered radiomics: revolutionizing detection of urologic malignancies. Curr Opin Urol 2024 Jan 1;34(1):1-7. [doi: 10.1097/MOU.0000000000001144] [Medline: 37909882]
14. Boca B, Caraiani C, Telecan T, et al. MRI-based radiomics in bladder cancer: a systematic review and radiomics quality score assessment. Diagnostics (Basel) 2023 Jul 6;13(13):2300. [doi: 10.3390/diagnostics13132300] [Medline: 37443692]
15. Marvaso G, Isaksson LJ, Zaffaroni M, et al. Can we predict pathology without surgery? Weighing the added value of multiparametric MRI and whole prostate radiomics in integrative machine learning models. Eur Radiol 2024 Oct;34(10):6241-6253. [doi: 10.1007/s00330-024-10699-3] [Medline: 38507053]
16. Kulkarni S, Seneviratne N, Baig MS, Khan AHA. Artificial intelligence in medicine: where are we now? Acad Radiol 2020 Jan;27(1):62-70. [doi: 10.1016/j.acra.2019.10.001] [Medline: 31636002]
17. Pak S, Park SG, Park J, Cho ST, Lee YG, Ahn H. Applications of artificial intelligence in urologic oncology. Investig Clin Urol 2024 May;65(3):202-216. [doi: 10.4111/icu.20230435] [Medline: 38714511]
18. Wang Z, Shi H, Wang Q, et al. AI-driven and traditional radiomic model for predicting muscle invasion in bladder cancer via multi-parametric imaging: a systematic review and meta-analysis. Acad Radiol 2025 Dec;32(12):7215-7243. [doi: 10.1016/j.acra.2025.08.035] [Medline: 40914722]
19. He C, Xu H, Yuan E, et al. The accuracy and quality of image-based artificial intelligence for muscle-invasive bladder cancer prediction. Insights Imaging 2024 Aug 1;15(1):185. [doi: 10.1186/s13244-024-01780-y] [Medline: 39090234]
20. Mastroleo F, Marvaso G, Jereczek-Fossa BA. Artificial intelligence in muscle-invasive bladder cancer: opportunities, challenges, and clinical impact. Curr Opin Urol 2025 Sep 1;35(5):543-548. [doi: 10.1097/MOU.0000000000001309] [Medline: 40765429]
21. Pastorino GL, Mercinelli C, Necchi A. The role of MRI in muscle-invasive bladder cancer: an update from the last two years. Curr Opin Urol 2025 Mar 1;35(2):165-170. [doi: 10.1097/MOU.0000000000001249] [Medline: 39529480]
22. Esteva A, Feng J, van der Wal D, et al. Prostate cancer therapy personalization via multi-modal deep learning on randomized phase III clinical trials. NPJ Digit Med 2022 Jun 8;5(1):71. [doi: 10.1038/s41746-022-00613-w] [Medline: 35676445]
23. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021 Jul 9;11(7):e048008. [doi: 10.1136/bmjopen-2020-048008] [Medline: 34244270]
24. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ 2025 Mar 24;388:e082505. [doi: 10.1136/bmj-2024-082505] [Medline: 40127903]
25. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017 Dec;14(12):749-762. [doi: 10.1038/nrclinonc.2017.141] [Medline: 28975929]
26. Debray TP, Damen JA, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. Stat Methods Med Res 2019 Sep;28(9):2768-2786. [doi: 10.1177/0962280218785504] [Medline: 30032705]
27. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
28. Zhang R, Jia S, Zhai L, Wu F, Zhang S, Li F. Predicting preoperative muscle invasion status for bladder cancer using computed tomography-based radiomics nomogram. BMC Med Imaging 2024 Apr 27;24(1):98. [doi: 10.1186/s12880-024-01276-7] [Medline: 38678222]
29. Zhang G, Wu Z, Xu L, et al. Deep learning on enhanced CT images can predict the muscular invasiveness of bladder cancer. Front Oncol 2021;11:654685. [doi: 10.3389/fonc.2021.654685] [Medline: 34178641]

XSL•FO
RenderX

30. Ye Y, Luo Z, Qiu Z, et al. Radiomics prediction of muscle invasion in bladder cancer using semi-automatic lesion segmentation of MRI compared with manual segmentation. Bioengineering (Basel) 2023 Nov 25;10(12):1355. [doi: 10.3390/bioengineering10121355] [Medline: 38135946]
31. Ren J, Gu H, Zhang N, Chen W. Preoperative CT-based radiomics for diagnosing muscle invasion of bladder cancer. Egypt J Radiol Nucl Med 2023;54(1):131. [doi: 10.1186/s43055-023-01044-7]
32. Qureshi TA, Chen X, Xie Y, et al. MRI/RNA-seq-based radiogenomics and artificial intelligence for more accurate staging of muscle-invasive bladder cancer. Int J Mol Sci 2023 Dec 20;25(1):88. [doi: 10.3390/ijms25010088] [Medline: 38203254]
33. Özdemir H, Azamat S, Sam Özdemir M. Can only the shape feature in radiomics help machine learning show that bladder cancer has invaded muscles? Cureus 2023 Sep;15(9):e45488. [doi: 10.7759/cureus.45488] [Medline: 37859896]
34. Chen W, Gong M, Zhou D, et al. CT-based deep learning radiomics signature for the preoperative prediction of the muscle-invasive status of bladder cancer. Front Oncol 2022;12:1019749. [doi: 10.3389/fonc.2022.1019749] [Medline: 36544709]
35. Zheng Z, Gu Z, Xu F, et al. Magnetic resonance imaging-based radiomics signature for preoperative prediction of Ki67 expression in bladder cancer. Cancer Imaging 2021 Dec 4;21(1):65. [doi: 10.1186/s40644-021-00433-3] [Medline: 34863282]
36. Ye L, Wang Y, Xiang W, Yao J, Liu J, Song B. Radiomic analysis of quantitative T2 mapping and conventional MRI in predicting histologic grade of bladder cancer. J Clin Med 2023 Sep 11;12(18):5900. [doi: 10.3390/jcm12185900] [Medline: 37762841]
37. Starmans MPA, Ho LS, Smits F, et al. Optimization of preoperative lymph node staging in patients with muscle-invasive bladder cancer using radiomics on computed tomography. J Pers Med 2022 Apr 30;12(5):726. [doi: 10.3390/jpm12050726] [Medline: 35629148]
38. Sarkar S, Min K, Ikram W, et al. Performing automatic identification and staging of urothelial carcinoma in bladder cancer patients using a hybrid deep-machine learning approach. Cancers (Basel) 2023 Mar 8;15(6):1673. [doi: 10.3390/cancers15061673] [Medline: 36980557]
39. Li L, Zhang J, Zhe X, et al. An MRI-based radiomics nomogram in predicting histologic grade of non-muscle-invasive bladder cancer. Front Oncol 2023;13:1025972. [doi: 10.3389/fonc.2023.1025972] [Medline: 37007156]
40. Deng Z, Dong W, Xiong S, et al. Machine learning models combining computed tomography semantic features and selected clinical variables for accurate prediction of the pathological grade of bladder cancer. Front Oncol 2023;13:1166245. [doi: 10.3389/fonc.2023.1166245] [Medline: 37223680]
41. Chen G, Fan X, Wang T, et al. A machine learning model based on MRI for the preoperative prediction of bladder cancer invasion depth. Eur Radiol 2023 Dec;33(12):8821-8832. [doi: 10.1007/s00330-023-09960-y] [Medline: 37470826]
42. Yu J, Cai L, Chen C, et al. A novel predict method for muscular invasion of bladder cancer based on 3D mp-MRI feature fusion. Phys Med Biol 2024 Feb 22;69(5). [doi: 10.1088/1361-6560/ad25c7] [Medline: 38306973]
43. Xu S, Yao Q, Liu G, et al. Combining DWI radiomics features with transurethral resection promotes the differentiation between muscle-invasive bladder cancer and non-muscle-invasive bladder cancer. Eur Radiol 2020 Mar;30(3):1804-1812. [doi: 10.1007/s00330-019-06484-2] [Medline: 31773297]
44. Zhou Z, Liu L, Xue K, Ma Y, Liu J, Zhang M. Assessment of pathological grading of bladder cancer using texture features from MRI. 2019 Presented at: 2019 IEEE International Conference on Mechatronics and Automation (ICMA); Aug 4-7, 2019; Tianjin, China p. 1333-1337 URL: https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=8801712 [doi: 10.1109/ICMA.2019.8816242]
45. Yang Y, Zou X, Wang Y, Ma X. Application of deep learning as a noninvasive tool to differentiate muscle-invasive bladder cancer and non-muscle-invasive bladder cancer with CT. Eur J Radiol 2021 Jun;139:109666. [doi: 10.1016/j.ejrad.2021.109666] [Medline: 33798819]
46. Zheng Z, Xu F, Gu Z, et al. Combining Multiparametric MRI Radiomics Signature With the Vesical Imaging-Reporting and Data System (VI-RADS) Score to Preoperatively Differentiate Muscle Invasion of Bladder Cancer. Front Oncol 2021;11:619893. [doi: 10.3389/fonc.2021.619893] [Medline: 34055600]
47. Song H, Yang S, Yu B, et al. CT-based deep learning radiomics nomogram for the prediction of pathological grade in bladder cancer: a multicenter study. Cancer Imaging 2023 Sep 18;23(1):89. [doi: 10.1186/s40644-023-00609-z] [Medline: 37723572]
48. Cui Y, Sun Z, Liu X, Zhang X, Wang X. CT-based radiomics for the preoperative prediction of the muscle-invasive status of bladder cancer and comparison to radiologists' assessment. Clin Radiol 2022 Jun;77(6):e473-e482. [doi: 10.1016/j.crad.2022.02.019] [Medline: 35367051]
49. Zhang G, Xu L, Zhao L, et al. CT-based radiomics to predict the pathological grade of bladder cancer. Eur Radiol 2020 Dec;30(12):6749-6756. [doi: 10.1007/s00330-020-06893-8] [Medline: 32601949]
50. Wei Z, Liu H, Xv Y, et al. Development and validation of a CT-based deep learning radiomics nomogram to predict muscle invasion in bladder cancer. Heliyon 2024 Jan 30;10(2):e24878. [doi: 10.1016/j.heliyon.2024.e24878] [Medline: 38304824]
51. Zheng J, Kong J, Wu S, et al. Development of a noninvasive tool to preoperatively evaluate the muscular invasiveness of bladder cancer using a radiomics approach. Cancer 2019 Dec 15;125(24):4388-4398. [doi: 10.1002/cncr.32490] [Medline: 31469418]

XSL•FO
RenderX

52. Wang H, Xu X, Zhang X, et al. Elaboration of a multisequence MRI-based radiomics signature for the preoperative prediction of the muscle-invasive status of bladder cancer: a double-center study. Eur Radiol 2020 Sep;30(9):4816-4827. [doi: 10.1007/s00330-020-06796-8] [Medline: 32318846]
53. Zheng Z, Xu F, Gu Z, et al. Integrating multiparametric MRI radiomics features and the Vesical Imaging-Reporting and Data System (VI-RADS) for bladder cancer grading. Abdom Radiol (NY) 2021 Sep;46(9):4311-4323. [doi: 10.1007/s00261-021-03108-6] [Medline: 33978825]
54. Wang W, Li W, Wang K, et al. Integrating radiomics with the vesical imaging-reporting and data system to predict muscle invasion of bladder cancer. Urol Oncol 2023 Jun;41(6):294. [doi: 10.1016/j.urolonc.2022.10.024] [Medline: 36526525]
55. Yu R, Cai L, Gong Y, et al. MRI-based machine learning radiomics for preoperative assessment of human epidermal growth factor receptor 2 status in urothelial bladder carcinoma. J Magn Reson Imaging 2024 Dec;60(6):2694-2704. [doi: 10.1002/jmri.29342] [Medline: 38456745]
56. Wang W, Wang K, Qiu J, et al. MRI-based radiomics analysis of bladder cancer: prediction of pathological grade and histological variant. Clin Radiol 2023 Nov;78(11):e889-e897. [doi: 10.1016/j.crad.2023.07.020] [Medline: 37633748]
57. Zhang L, Li X, Yang L, et al. Multi‐sequence and multi‐regional MRI‐based radiomics nomogram for the preoperative assessment of muscle invasion in bladder cancer. J Magn Reson Imaging 2023 Jul;58(1):258-269 [FREE Full text] [doi: 10.1002/jmri.28498] [Medline: 36300676]
58. Zou Y, Cai L, Chen C, et al. Multi-task deep learning based on T2-weighted images for predicting muscular-invasive bladder cancer. Comput Biol Med 2022 Dec;151(Pt A):106219. [doi: 10.1016/j.compbiomed.2022.106219] [Medline: 36343408]
59. Zhou X, Yue X, Xu Z, Denoeux T, Chen Y. PENet: Prior evidence deep neural network for bladder cancer staging. Methods 2022 Nov;207:20-28. [doi: 10.1016/j.ymeth.2022.08.010] [Medline: 36031139]
60. Li J, Qiu Z, Cao K, et al. Predicting muscle invasion in bladder cancer based on MRI: a comparison of radiomics, and single-task and multi-task deep learning. Comput Methods Programs Biomed 2023 May;233:107466. [doi: 10.1016/j.cmpb.2023.107466] [Medline: 36907040]
61. Li L, Zhang J, Zhe X, et al. Prediction of histopathologic grades of bladder cancer with radiomics based on MRI: comparison with traditional MRI. Urol Oncol 2024 Jun;42(6):176. [doi: 10.1016/j.urolonc.2024.02.008] [Medline: 38556403]
62. Chen L, Zhang G, Xu L, et al. Preoperative CT features to predict risk stratification of non-muscle invasive bladder cancer. Abdom Radiol (NY) 2023 Feb;48(2):659-668. [doi: 10.1007/s00261-022-03730-y] [Medline: 36454277]
63. Xu X, Liu Y, Zhang X, et al. Preoperative prediction of muscular invasiveness of bladder cancer with radiomic features on conventional MRI and its high-order derivative maps. Abdom Radiol (NY) 2017 Jul;42(7):1896-1905. [doi: 10.1007/s00261-017-1079-6] [Medline: 28217825]
64. Xu X, Zhang X, Tian Q, et al. Quantitative identification of nonmuscle‐invasive and muscle‐invasive bladder carcinomas: a multiparametric MRI radiomics analysis. J Magn Reson Imaging 2019 May;49(5):1489-1498 [FREE Full text] [doi: 10.1002/jmri.26327] [Medline: 30252978]
65. Gao RZ, Wen R, Wen DY, et al. Radiomics analysis based on ultrasound images to distinguish the tumor stage and pathological grade of bladder cancer. J Ultrasound Med 2021 Dec;40(12):2685-2697. [doi: 10.1002/jum.15659] [Medline: 33615528]
66. Wang H, Hu D, Yao H, et al. Radiomics analysis of multiparametric MRI for the preoperative evaluation of pathological grade in bladder cancer tumors. Eur Radiol 2019 Nov;29(11):6182-6190. [doi: 10.1007/s00330-019-06222-8] [Medline: 31016445]
67. Zhang X, Xu X, Tian Q, et al. Radiomics assessment of bladder cancer grade using texture features from diffusion-weighted imaging. J Magn Reson Imaging 2017 Nov;46(5):1281-1288 [FREE Full text] [doi: 10.1002/jmri.25669] [Medline: 28199039]
68. Gresser E, Woźnicki P, Messmer K, et al. Radiomics signature using manual versus automated segmentation for lymph node staging of bladder cancer. Eur Urol Focus 2023 Jan;9(1):145-153. [doi: 10.1016/j.euf.2022.08.015] [Medline: 36115774]
69. Tong Y, Udupa JK, Wang C, et al. Radiomics-guided therapy for bladder cancer: using an optimal biomarker approach to determine extent of bladder cancer invasion from t2-weighted magnetic resonance images. Adv Radiat Oncol 2018;3(3):331-338. [doi: 10.1016/j.adro.2018.04.011] [Medline: 30202802]
70. Tao T, Chen Y, Shang Y, He J, Hao J. SMMF: a self-attention-based multi-parametric MRI feature fusion framework for the diagnosis of bladder cancer grading. Front Oncol 2024;14:1337186. [doi: 10.3389/fonc.2024.1337186] [Medline: 38515574]
71. Liu Y, Xu X, Wang H, et al. The additional value of tri-parametric MRI in identifying muscle-invasive status in bladder cancer. Acad Radiol 2023 Jan;30(1):64-76. [doi: 10.1016/j.acra.2022.04.014] [Medline: 35676179]
72. Liu D, Wang S, Wang J. The effect of CT high-resolution imaging diagnosis based on deep residual network on the pathology of bladder cancer classification and staging. Comput Methods Programs Biomed 2022 Mar;215:106635. [doi: 10.1016/j.cmpb.2022.106635] [Medline: 35063711]
73. Razik A, Das CJ, Sharma R, et al. Utility of first order MRI-Texture analysis parameters in the prediction of histologic grade and muscle invasion in urinary bladder cancer: a preliminary study. Br J Radiol 2021 Jun 1;94(1122):20201114. [doi: 10.1259/bjr.20201114] [Medline: 33882245]

XSL•FO
RenderX

74. Peng J, Tang Z, Li T, Pan X, Feng L, Long L. Contrast-enhanced computed tomography-based radiomics nomogram for predicting HER2 status in urothelial bladder carcinoma. Front Oncol 2024;14:1427122. [doi: 10.3389/fonc.2024.1427122] [Medline: 39206159]
75. Xiong S, Fu Z, Deng Z, et al. Machine learning - based CT radiomics enhances bladder cancer staging predictions: a comparative study of clinical, radiomics, and combined models. Med Phys 2024 Sep;51(9):5965-5977 [FREE Full text] [doi: 10.1002/mp.17288] [Medline: 38977273]
76. Feng S, Zhou D, Li Y, et al. Prediction of Ki-67 expression in bladder cancer based on CT radiomics nomogram. Front Oncol 2024;14:1276526. [doi: 10.3389/fonc.2024.1276526] [Medline: 38482209]
77. Wei Z, Bai X, Xv Y, et al. A radiomics-based interpretable machine learning model to predict the HER2 status in bladder cancer: a multicenter study. Insights Imaging 2024 Oct 28;15(1):262. [doi: 10.1186/s13244-024-01840-3] [Medline: 39466475]
78. Cai L, Yang X, Yu J, et al. Deep learning on T2WI to predict the muscle-invasive bladder cancer: a multi-center clinical study. Sci Rep 2025 Mar 22;15(1):9942. [doi: 10.1038/s41598-024-82909-3] [Medline: 40121216]
79. Du C, Wei W, Hu M, et al. Multi-DECT image-based interpretable model incorporating habitat radiomics and vision transformer deep learning for preoperative prediction of muscle invasion in bladder cancer. Acad Radiol 2025 Dec;32(12):7204-7214. [doi: 10.1016/j.acra.2025.08.018] [Medline: 40887351]
80. Du Y, Li H, Sui Y, et al. Habitat-based radiomic model for predicting muscle invasion in bladder cancer: a multi-center study using enhanced-CT and machine learning. Med Phys 2025 Aug;52(8):e18021. [doi: 10.1002/mp.18021] [Medline: 40781767]
81. He X, Chen Y, Zhou S, et al. A machine learning model based on multi-phase contrast-enhanced CT for the preoperative prediction of the muscle-invasive status of bladder cancer. Curr Med Imaging 2025;21:e15734056377754. [doi: 10.2174/0115734056377754250304040058] [Medline: 40103464]
82. Hu M, Zhang J, Cheng Q, et al. Multi-DECT image-based intratumoral and peritumoral radiomics for preoperative prediction of muscle invasion in bladder cancer. Acad Radiol 2025 Jan;32(1):287-297. [doi: 10.1016/j.acra.2024.08.010] [Medline: 39168722]
83. Yu J, Cai L, Chen C, et al. Multi-path neural network based on mp-MRI for predicting muscle-invasive bladder cancer. Intell Data Anal 2025 Nov;29(6):1568-1581. [doi: 10.1177/1088467X241313324]
84. Zhou Q, Ma L, Yu Y, et al. Development of a radiomics and clinical feature-based nomogram for preoperative prediction of pathological grade in bladder cancer. Front Oncol 2025;15:1661979. [doi: 10.3389/fonc.2025.1661979] [Medline: 40936708]
85. Verma S, Rajesh A, Prasad SR, et al. Urinary bladder cancer: role of MR imaging. Radiographics 2012;32(2):371-387. [doi: 10.1148/rg.322115125] [Medline: 22411938]
86. Bleker J, Kwee TC, Rouw D, et al. A deep learning masked segmentation alternative to manual segmentation in biparametric MRI prostate cancer radiomics. Eur Radiol 2022 Sep;32(9):6526-6535. [doi: 10.1007/s00330-022-08712-8] [Medline: 35420303]
87. van der Pol CB, Chung A, Lim C, et al. Update on multiparametric MRI of urinary bladder cancer. J Magn Reson Imaging 2018 Oct;48(4):882-896. [doi: 10.1002/jmri.26294] [Medline: 30221801]
88. Caglic I, Panebianco V, Vargas HA, et al. MRI of bladder cancer: local and nodal staging. J Magn Reson Imaging 2020 Sep;52(3):649-667. [doi: 10.1002/jmri.27090] [Medline: 32112505]
89. Wentland AL, Desser TS, Troxell ML, Kamaya A. Bladder cancer and its mimics: a sonographic pictorial review with CT/MR and histologic correlation. Abdom Radiol (NY) 2019 Dec;44(12):3827-3842. [doi: 10.1007/s00261-019-02276-w] [Medline: 31676920]
90. Akin O, Lema-Dopico A, Paudyal R, et al. Multiparametric MRI in era of artificial intelligence for bladder cancer therapies. Cancers (Basel) 2023 Nov 18;15(22):5468. [doi: 10.3390/cancers15225468] [Medline: 38001728]
91. Cai L, Bai R, Cao Q, et al. A non-invasive MRI-based multimodal fusion deep learning model (MF-DLM) for predicting overall survival in bladder cancer: a multicentre retrospective study. EClinicalMedicine 2025 Dec;90:103640. [doi: 10.1016/j.eclinm.2025.103640] [Medline: 41324013]
92. Wei Z, Xv Y, Liu H, et al. A CT-based deep learning model predicts overall survival in patients with muscle invasive bladder cancer after radical cystectomy: a multicenter retrospective cohort study. Int J Surg 2024 May 1;110(5):2922-2932. [doi: 10.1097/JS9.0000000000001194] [Medline: 38349205]
93. Jiang F, Hong G, Zeng H, et al. Deep learning-based model for prediction of early recurrence and therapy response on whole slide images in non-muscle-invasive bladder cancer: a retrospective, multicentre study. EClinicalMedicine 2025 Mar;81:103125. [doi: 10.1016/j.eclinm.2025.103125] [Medline: 40093987]
94. Lucas M, Jansen I, van Leeuwen TG, Oddens JR, de Bruin DM, Marquering HA. Deep learning-based recurrence prediction in patients with non-muscle-invasive bladder cancer. Eur Urol Focus 2022 Jan;8(1):165-172. [doi: 10.1016/j.euf.2020.12.008] [Medline: 33358370]
95. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group ‘Evaluating diagnostic tests and prediction models’ of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med 2019 Dec 16;17(1):230. [doi: 10.1186/s12916-019-1466-7] [Medline: 31842878]

XSL•FO
RenderX

96. Zhang T, Nikouline A, Lightfoot D, Nolan B. Machine learning in the prediction of trauma outcomes: a systematic review. Ann Emerg Med 2022 Nov;80(5):440-455. [doi: 10.1016/j.annemergmed.2022.05.011] [Medline: 35842343]
97. Dong J, Jiang Z, Pan D, et al. A survey on confidence calibration of deep learning-based classification models under class imbalance data. IEEE Trans Neural Netw Learn Syst 2025 Sep;36(9):15664-15684. [doi: 10.1109/TNNLS.2025.3565159] [Medline: 40531642]
98. Ehrmann DE, Joshi S, Goodfellow SD, Mazwi ML, Eytan D. Making machine learning matter to clinicians: model actionability in medical decision-making. NPJ Digit Med 2023 Jan 24;6(1):7. [doi: 10.1038/s41746-023-00753-7] [Medline: 36690689]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
AUROC: area under the receiver operating characteristic curve
CE-CT: contrast-enhanced computed tomography
CT: computed tomography
DL: deep learning
GRADE: Grading of Recommendations, Assessment, Development, and Evaluation
HER2: human epidermal growth factor receptor 2
LN: lymph node
LR: logistic regression
MIBC: muscle-invasive bladder cancer
ML: machine learning
MRI: magnetic resonance imaging
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST-AI: Prediction Model Risk of Bias Assessment Tool for artificial intelligence
PROSPERO: International Prospective Register of Systematic Reviews
ROB: risk of bias
ROI: region of interest
RQS: Radiomics Quality Scoring
SVM: support vector machine
TURBT: Transurethral resection of bladder tumor

Edited by S Brini; submitted 22.Jul.2025; peer-reviewed by H Lv, L Zhou, S Zamani; revised version received 03.Apr.2026; accepted 06.Apr.2026; published 12.Jun.2026.

Please cite as:
He Z, Liu Y, Jiang Q, Hou F, Xu B
Application Value of Radiomics-Based Machine Learning for Preoperative Risk Stratification of Bladder Cancer: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e81084
URL: https://www.jmir.org/2026/1/e81084
doi:10.2196/81084

© Zirong He, Yinghua Liu, Qin Jiang, Fang Hou, Bing Xu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Benchmark Integrity and Reasoning-Trace Errors in Medical Question Answering With Large Language Models: Mixed Methods Study With Sparse Autoencoders

Jialin Liu[1,2*], MD; Siru Liu[3,4*], PhD; Adam Wright[3,5], PhD

[1]Department of Medical Informatics, West China Hospital of Sichuan University, Chengdu, Sichuan, China

[2]Department of Otolaryngology-Head and Neck Surgery, West China Hospital of Sichuan University, Chengdu, Sichuan, China

[3]Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Ave, Nashville, TN, United States

[4]Department of Computer Science, Vanderbilt University, Nashville, TN, United States

[5]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States

[*]these authors contributed equally

Corresponding Author:
Siru Liu, PhD
Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Ave, Nashville, TN, United States

Abstract

Background: Large language models (LLMs) show promise for enhancing diagnostic accuracy and clinical decision-making. However, prevailing evaluations rely on examination-based benchmarks such as MedQA. Furthermore, the internal mechanisms driving both correct and incorrect reasoning in LLMs remain poorly understood, limiting opportunities for targeted improvement.

Objective: This study aimed to investigate failure modes of reasoning-based LLMs in medicine by (1) auditing the integrity of the MedQA benchmark, (2) developing a clinically informed taxonomy of reasoning errors across multiple major LLMs, and (3) testing a mechanistic intervention using sparse autoencoders (SAEs) to modulate reasoning characteristics and improve accuracy in medical question answering benchmarks.

Methods: We evaluated OpenAI o1 on the MedQA and cross-referenced incorrect answers against original source platforms to identify benchmark flaws including missing figures and postrelease ambiguity corrections. For the 37 confirmed model failures remaining after exclusion of flawed items, we developed a reasoning error taxonomy through iterative inductive coding by 2 independent reviewers (JL and SL) and validated it on three major LLMs (ie, OpenAI GPT-4.5, OpenAI o3-mini, and DeepSeek-R1). We then trained an SAE on the DeepSeek-R1-Distill-Llama-8B model using MedQA-derived reasoning traces. Reasoning-specific features were identified using ReasonScore and subjected to activation steering at 2 strengths. Model accuracy, reasoning trace length, and hallucination metrics were measured across MedQA, MedMCQA, and PubMedQA. Hallucinations were evaluated using an LLM-as-a-judge (OpenAI GPT-5-mini) and validated on a stratified manual sample of 100 claims.

Results: Forty-one percent of initial OpenAI o1 errors reflected benchmark problems, including missing figures (22%) and ambiguities subsequently corrected on the source platforms (19%). Neither OpenAI o1 nor OpenAI o3-mini explicitly flagged these flawed items, while GPT-5.2 identified a small subset, suggesting that question-integrity recognition remains limited and model-dependent. Among the 37 confirmed errors, our taxonomy classified failures into four categories: Information Synthesis Errors, Therapeutic Decision Errors, Diagnostic Reasoning Errors, and Foundational Principle Errors. Activation steering of reasoning-specific SAE features improved accuracy on MedQA and PubMedQA, with a consistent positive trend on MedMCQA. The greatest gains were observed at steering strength 2 (MedQA: 0.568-0.597 and PubMedQA: 0.708-0.739). Steering also increased reasoning-trace length substantially, with no significant correlation between verbosity and accuracy. Five functional feature categories were identified, with alignments to the error taxonomy.

Conclusions: These findings reveal two distinct sources of unreliability in medical LLM evaluation: benchmark-level integrity gaps that misattribute model failure and recurrent model-level reasoning patterns potentially amenable to mechanistic correction. Notably, the benchmark issues identified here do not reflect static flaws in the original source platforms, which have since corrected many problematic items, but rather a failure to propagate those corrections to derived benchmarks. The alignment between SAE-identified feature categories and the error taxonomy further suggests that reasoning failures reflect structured internal processes that can be targeted at the feature level.

(J Med Internet Res 2026;28:e90061) doi:10.2196/90061

XSL•FO
RenderX

KEYWORDS

large language model; medical question answering; sparse autoencoders; benchmark evaluation; reasoning errors; clinical decision support; mechanistic interpretability; activation steering; hallucination; artificial intelligence in medicine

Introduction

The integration of artificial intelligence (AI) into medicine presents a profound opportunity to enhance diagnostic accuracy and clinical decision-making. At the forefront of this transformation are large language models (LLMs), which, after training on vast corpora of medical literature and clinical guidelines, have demonstrated remarkable capabilities [1]. These models show promise in diverse applications, from critiquing clinical decision support systems to drafting patient communications and summarizing clinical notes [2-6].

Recent LLM releases differ less in core architecture than in training incentives and inference-time policies that govern how much computation a model allocates before producing an answer [7]. Reasoning models are optimized (eg, via reinforcement learning) to perform better on multistep problems and often generate longer intermediate rationales, whereas general-purpose chat models may produce shorter justifications by default. Importantly, generated rationales are not guaranteed to be faithful readouts of causal reasoning [8].

However, the prevailing methods for evaluating these advanced models remain a critical weak point. Current assessments predominantly rely on examination-based benchmarks, with a mere 5% using real patient data [9,10]. MedQA, a widely used benchmark derived from the United States Medical Licensing Examination (USMLE), exemplifies this issue [11]. Developed in 2020 from questions scraped from public websites, its multiple-choice format often tests rote memorization over nuanced clinical judgment [12]. While some have questioned the validity of using examination questions to evaluate clinical AI [13], there has been little systematic investigation into the intrinsic quality and reliability of the benchmark questions themselves. Prior work has shown that LLM performance drops substantially when multiple-choice shortcuts are disrupted, further underscoring the need to examine benchmark quality beyond surface-level accuracy [14].

A crucial factor in the performance of reasoning LLMs is their explicit, stepwise thought process, often refined through reinforcement learning. This can be optimized using outcome-based rewards, which assess only the final answer's correctness [15], or process-based rewards, which evaluate the entire reasoning sequence [16]. In deterministic fields such as mathematics, outcome-based rewards suffice. In medicine, however, the clinical logic must be as sound as the conclusion, making the analysis of reasoning errors essential for developing safe and effective clinical LLMs. While reasoning errors in LLMs have been studied in general domains, clinically grounded taxonomies of reasoning failures across multiple frontier models remain limited. Recent work in mechanistic interpretability has demonstrated that sparse autoencoders (SAEs) can identify and steer interpretable features in LLMs [17,18], but this approach has not yet been applied to medical reasoning.

We address 3 linked gaps in evaluating and improving medical question-answering (QA) LLMs. First, we audit MedQA test-set integrity by reconciling incorrect model outputs with the original source platforms to identify missing modalities and postrelease corrections. Second, focusing on questions that remain well-specified, we develop an initial clinically informed taxonomy of observable reasoning-trace failures and examine how its distribution varies across several frontier models. Third, we test whether a mechanistic intervention, steering SAE features enriched around reasoning-trace tokens, can measurably shift accuracy and reasoning-trace properties across multiple medical QA datasets.

Methods

Reasoning Error Analysis and Taxonomy Development

We selected the MedQA test dataset (N=1273) of USMLE-style questions as our primary benchmark. We used OpenAI o1, a representative state-of-the-art reasoning model, to generate answers and a corresponding chain-of-thought reasoning process for each question. LLM prompts for answer and reasoning generation are listed in Multimedia Appendix 1. We identified incorrectly answered questions by comparing the model's output with the provided answer key.

For each incorrect answer, we located the original question in its source examination bank (Medbullets, AMBOSS, or Lecturio). This allowed us to identify discrepancies, such as missing figures or subsequent question updates made by the source platforms to correct ambiguities. During reconciliation with source platforms, we also tracked whether models explicitly signaled uncertainty or unanswerability, such as noting missing information or required figures, or identifying ambiguity in the question stem. To assess whether newer models demonstrate improved question-integrity recognition, we additionally tested GPT-5.2-pro-2025-12-11 on the identified flawed items post hoc. This source reconciliation was feasible for MedQA because its questions could be traced to actively maintained platforms. Analogous auditing was not performed for MedMCQA or PubMedQA, as MedMCQA was compiled from various open websites and books without a single maintained source platform, and PubMedQA derives questions and answers from PubMed abstracts rather than from a curated examination bank.

After excluding questions with these external quality issues, we developed an error taxonomy using an inductive coding approach [19]. Two coders (SL and JL) independently analyzed the reasoning processes for a subset of 5 incorrect answers, using open coding to identify error themes. The coders then met to discuss their findings and create a consensus-based coding guideline. This iterative process was repeated with new batches of 5 questions until all reasoning traces for the incorrectly answered questions were coded.

To validate the taxonomy, we assessed the performance of other advanced LLMs (OpenAI GPT-4.5, OpenAI o3-mini, and

XSL•FO
RenderX

DeepSeek-R1) on the same set of questions. For models that do not natively output their reasoning, we used a step-by-step chain-of-thought prompt to elicit it. All models were accessed via their respective application programming interfaces (APIs). The specific API versions queried were OpenAI o1 (o1-preview, queried January 11, 2025), OpenAI o3-mini (o3-mini-2025-01-31, queried March 25, 2025), GPT-4.5 (gpt-4.5-preview-2025-02-27, queried March 25, 2025), and DeepSeek-R1 (queried March 25, 2025). Throughout the remainder of the paper, these models are referred to as o1, o3-mini, GPT-4.5, and DeepSeek-R1, respectively. Each question was evaluated in a single-shot setting with temperature set to 0. Other decoding parameters (top-p, seed, max tokens) were not explicitly set and followed the provider default. We then applied our taxonomy to classify the failure modes in each model's incorrect reasoning traces and manually compared reasoning traces from different models. Notably, DeepSeek-R1 was deliberately included as one of the validation models because the SAE analysis model (DeepSeek-R1-Distill-Llama-8B) was derived from it via knowledge distillation.

SAE Development and Feature Steering

To further interpret the reasoning process quantitatively, we used an SAE, a method to decompose an LLM's internal activations into a sparse set of interpretable features [17]. The core idea is that a neural network's internal representations are dense and difficult to interpret directly, as individual neurons often respond to multiple unrelated concepts (polysemanticity). An SAE addresses this by learning to reconstruct the model's activations through a higher-dimensional but sparse intermediate representation, where each dimension (or feature) ideally corresponds to a single interpretable concept. SAEs were chosen over other interpretability methods (eg, attention analysis and probing classifiers) because they enable both interpretation and intervention: once reasoning-relevant features are identified, their activations can be directly modified to test effects on model behavior.

We trained an SAE using the DeepSeek-R1-Distill-Llama-8B model. This model was chosen for its open-source availability, strong reasoning capabilities, and moderate size, which balanced performance with experimental efficiency. The training data included general conversation logs (LMSys-Chat-1M) [20], general reasoning traces (OpenThoughts-114k) [21], and reasoning traces leading to correct answers in the training dataset of MedQA generated by DeepSeek-R1. Notably, the MedQA training and testing sets are different, and the SAE evaluation was conducted exclusively on the test set. Following established methods, we extracted activations from the 19th layer of the model to train the SAE [18].

We used ReasonScore, a quantitative metric, to identify reasoning-specific features from a trained SAE [18]. The metric works by identifying features that activate most strongly around predefined reasoning words within a fixed-width context window and using an entropy penalty to penalize features that activate only on limited reasoning words. After identifying the top 100 features with the highest ReasonScore, we manually reviewed their activations and logits using an SAE dashboard [22]. For the top 15 features, 2 reviewers (SL and JL) independently examined the top-activating contexts, activation patterns, and logit effects for each feature, then met to discuss and reconcile their functional labels through consensus. This process yielded 5 functional categories, which were then compared with the independently derived error taxonomy.

To test the impact of these features, we used activation steering [23,24]. This method involves intentionally modifying the model's reasoning by adding a positive bias to the activations of top-performing features (using strengths of 2 and 4). Boosting a feature's activation encourages the model to more strongly incorporate that feature's specific concept into its subsequent processing. We measured the accuracy of the steered model on three medical benchmarks: MedQA, which served as the primary benchmark across all study components; and 2 other widely used benchmarks, MedMCQA (multiple-choice questions from Indian medical entrance examinations) and PubMedQA (yes, no, or maybe questions derived from PubMed abstracts). MedMCQA and PubMedQA were included to test whether steering effects generalize beyond the dataset used for taxonomy development and SAE training. The 26 MedQA items identified as having benchmark integrity issues in the audit phase were retained in the SAE evaluation, as these items affect all steering conditions equally and represent approximately 2% of the test set. In addition, we used an LLM-as-a-judge (OpenAI GPT-5-mini) to assess hallucinations in the reasoning process, categorizing them as factual hallucinations (contradicting known facts) or input hallucinations (contradicting the question prompt) and assigning a severity score from 1 to 3. LLM-generated labels were validated manually on a stratified sample of 100 claims across datasets and severity levels. Expert validation was performed by JL, a physician with an MD degree and clinical experience as an attending in otolaryngology–head and neck surgery, as well as research experience in medical informatics. All disagreement cases were independently reviewed by SL, an investigator with a PhD in biomedical informatics and expertise in clinical AI evaluation. The severity correlation was computed at the claim level between the LLM judge's severity rating and the human annotator's independently assigned rating using the same rubric. To assess feature generalizability, we ranked features by their poststeering accuracy (exact match) within each dataset and steering strength, selected the top K features (for K=5, 10, 15, and 20), and computed the set intersection count between each dataset pair. Then, we selected the top 15 features, manually compared reasoning patterns before and after steering, and grouped them based on their potential function in reasoning. We used the *SAELens* package to train the SAE [25], and the *SAEDashboard* package to visualize identified features and their impact in Python 3 [22]. The overall project pipeline is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Overview of the project pipeline for reasoning - error taxonomy development. LLM: large language model; QA: question answering.

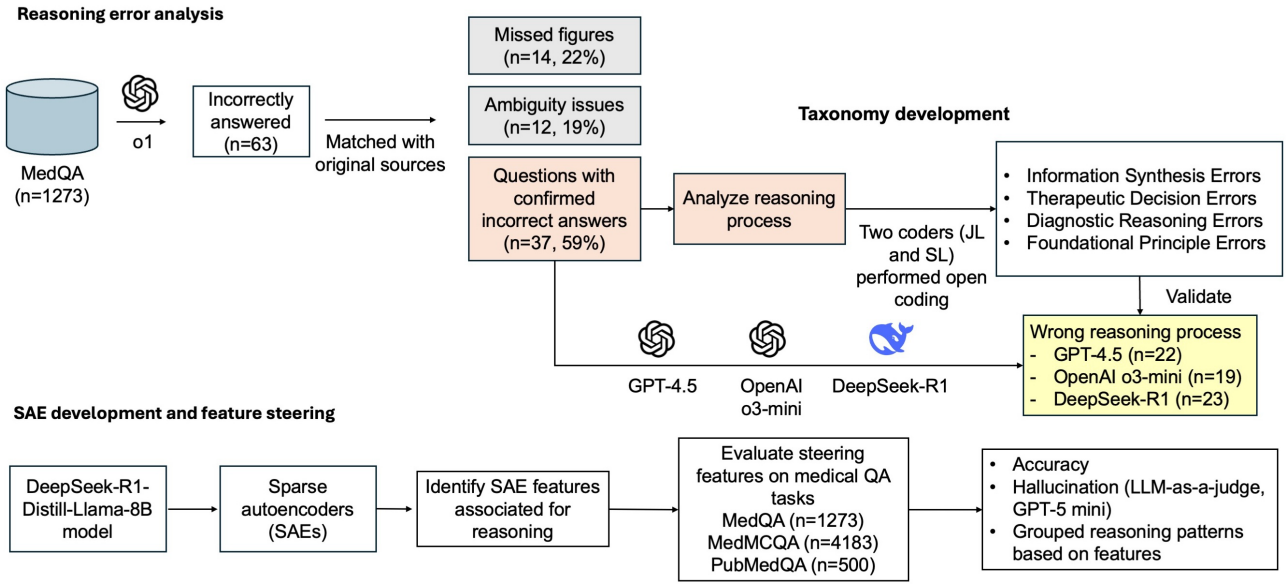

Statistical Analysis

The same questions were evaluated under all steering conditions, yielding paired observations. For accuracy (binary outcome), McNemar tests were used for pairwise comparisons between steering strengths. For reasoning token counts, Wilcoxon signed-rank tests were used. For hallucination counts, where the per-question hallucination set differs across conditions, Mann-Whitney *U* tests were used. All pairwise tests were corrected for multiple comparisons using the Holm method. Chi-square tests were used to compare the distribution of hallucination types across conditions. Pearson correlation coefficients were calculated between reasoning token length and performance metrics (accuracy, hallucination frequency, and severity), with statistical significance assessed using *t* tests on the correlation coefficients. All statistical analyses were performed using Python 3 (Python Software Foundation).

Ethical Considerations

This study used publicly available, deidentified data and did not require ethics approval.

Results

Benchmark Quality Analysis

The OpenAI o1 model answered 63 of 1273 questions incorrectly, for an accuracy of 95%. Upon cross-referencing these 63 questions with their original examination banks, we found that 14 (22%) questions were missing figures that were essential for arriving at the correct answer (Textbox 1). For instance, one question about a patient's pressure-volume loop explicitly mentioned a figure (Figure 2), while another describing a patient with retrosternal burning implicitly relied on a figure showing diffuse lung fibrosis to make the correct diagnosis. A complete list of questions that should include figures is provided in Multimedia Appendix 2.

An additional 12 (19%) questions contained ambiguities that have since been corrected in the source question banks (Textbox 1). For example, a question about a 3-month-old infant with a holosystolic murmur was updated to include "tetany is noted when taking the blood pressure," a hallmark sign that clarifies the diagnosis as 22q11 deletion syndrome (DiGeorge syndrome) rather than fetal alcohol syndrome. In another instance, a question involving a 78-year-old woman was revised. The version in MedQA implied that acute cognitive changes suggested a symptomatic urinary tract infection (warranting treatment), whereas the updated version, by noting that the patient continued to exhibit symptoms consistent with Alzheimer dementia, supported a decision for no treatment. Twelve questions (19%) that had serious ambiguity issues, resulting in answers that differ from those provided as correct, are listed in Multimedia Appendix 3.

XSL•FO
RenderX

Textbox 1. Examples of MedQA questions missing original figures and examples of questions with ambiguity issues. Bold text indicates updated information in the current online questions compared with the original questions in MedQA.

Example 1. Figure explicitly mentioned and needed to answer correctly, but not present in the benchmark.

Question: A 72-year-old woman is admitted to the intensive care unit for shortness of breath and palpitations. A cardiac catheterization is performed, and measurements of the left ventricular volume and pressure at different points in the cardiac cycle are obtained. The patient's pressure-volume loop (gray) is shown with a normal pressure-volume loop (black) for comparison. Which of the following is the most likely underlying cause of this patient's symptoms?

Options:

A: Mitral valve regurgitation

B: Increased systemic vascular resistance

C: Increased ventricular wall stiffness

D: Impaired left ventricular contractility

Example 2. Figure not explicitly mentioned and needed to answer correctly, but not present in the benchmark.

Question: A 43-year-old woman presents with complaints of retrosternal burning associated with eating. It has persisted for the past several years but has been getting worse. Her past medical history is unknown and this is her first time seeing a doctor. She states she is otherwise healthy and review of systems is notable for episodic hand pain that is worse in the winter as well as a chronic and severe cough with dyspnea, which she attributes to her smoking. Her temperature is 97.7 °F (36.5 °C), blood pressure is 174/104 mmHg, pulse is 80/min, respirations are 22/min, and oxygen saturation is 92% on room air. Physical exam is notable for a young appearing woman with coarse breath sounds. Laboratory studies and urinalysis are ordered and currently pending. Which of the following is the pathophysiology of this patient's chief complaint?

Options:

A: Decreased lower esophageal tone

B: Esophageal fibrosis

C: Increased lower esophageal tone

D: Spastic cricopharyngeal muscle

Note: The original question included a chest computed tomography image (Figure 2), which is necessary to answer this question.

Example 3. A question with ambiguity issues.

MedQA question: A 3-month-old infant is brought to her pediatrician because she coughs and seems to have difficulty breathing while feeding. In addition, she seems to have less energy compared to other babies and appears listless throughout the day. She was born by cesarean section to a G1P1 woman with no prior medical history and had a normal APGAR score at birth. Her parents say that she has never been observed to turn blue. Physical exam reveals a high-pitched holosystolic murmur that is best heard at the lower left sternal border. The most likely cause of this patient's symptoms is associated with which of the following abnormalities?

Options:

A: 22q11 deletion

B: Deletion of genes on chromosome 7

C: Lithium exposure in utero

D: Maternal alcohol consumption

Updated question: A 3-month-old infant is brought to her pediatrician because she coughs and seems to have difficulty breathing while feeding. In addition, she seems to have less energy compared to other babies and appears listless throughout the day. She was born by cesarean section to a G1P1 woman with no prior medical history and had a normal APGAR score at birth. Her parents say that she has never been observed to turn blue. Physical exam reveals a high-pitched holosystolic murmur that is best heard at the lower left sternal border. Tetany is noted when taking the blood pressure. The most likely cause of this patient's symptoms is associated with which of the following abnormalities?

MedQA question: A 78-year-old woman presents to the office for an annual health check-up with her family physician accompanied by her daughter. She has no complaints during this visit but her daughter states that she is having difficulty locating objects such as the television remote, car keys, and her purse. Her medical history is significant for Alzheimer's dementia, coronary artery disease, diabetes mellitus, hypothyroidism, congestive heart failure, osteoarthritis, and centrilobular emphysema. The patient takes memantine, atorvastatin, metformin, levothyroxine, lisinopril, aspirin, albuterol, and ipratropium. The patient's vitals are within normal limits today. Physical exam reveals an elderly female in no acute distress, oriented to person, place, and year, but not to month or day of the week. She has a 3/6 holosystolic murmur at the left sternal border along with an S3 gallop. There are mild crackles at the lung bases. The remainder of the exam is normal. A previous urine culture reports growth of > 100,000 CFU of Enterobacter. Urinalysis findings are offered below:

Leukocyte esterase positive

WBCs 50 - 100 cell/HPF

Nitrites positive

RBCs 2 cell/HPF

XSL•FO
RenderX

Epithelial cells 2 cell/HPF

Urine pH 5.7

Which of the following is the most appropriate next step?

A: TMP-SMX

B: Nitrofurantoin

C: Levofloxacin

D: No treatment is necessary

Updated question: A 78-year-old woman presents with her daughter for an annual health check-up with her family physician. The patient has no complaints during this visit but her daughter states that the patient continues to manifest symptoms consistent with her known diagnosis of Alzheimer dementia, including having difficulty locating objects such as her television remote, car keys, and purse. Her other medical history is significant for coronary artery disease, diabetes mellitus, hypothyroidism, congestive heart failure, osteoarthritis, and centrilobular emphysema. The patient takes memantine, atorvastatin, metformin, levothyroxine, lisinopril, aspirin, albuterol, and ipratropium. The patient's vitals are within normal limits. Physical exam reveals no acute distress. She is oriented to person, place, and year, but not to month or day of the week. She has a 3/6 holosystolic murmur at the left sternal border along with an S3 gallop. There are mild crackles at the lung bases. The remainder of the exam is normal. Two urine cultures performed in one week each reported growth of > 100,000 CFU of Enterobacter. Urinalysis findings are presented below:

Leukocyte esterase positive

WBCs 50 - 100 cell/HPF

Nitrites positive

RBCs 2 cell/HPF

Epithelial cells 2 cell/HPF

Urine pH 5.7

Which of the following is the most appropriate next step?

A: TMP-SMX

B: Nitrofurantoin

C: Levofloxacin

D: No treatment is necessary

XSL•FO
RenderX

Figure 2. Chest computed tomography image originally accompanying example 2 in MedQA. This figure was present in the source question bank but absent from the MedQA benchmark dataset.

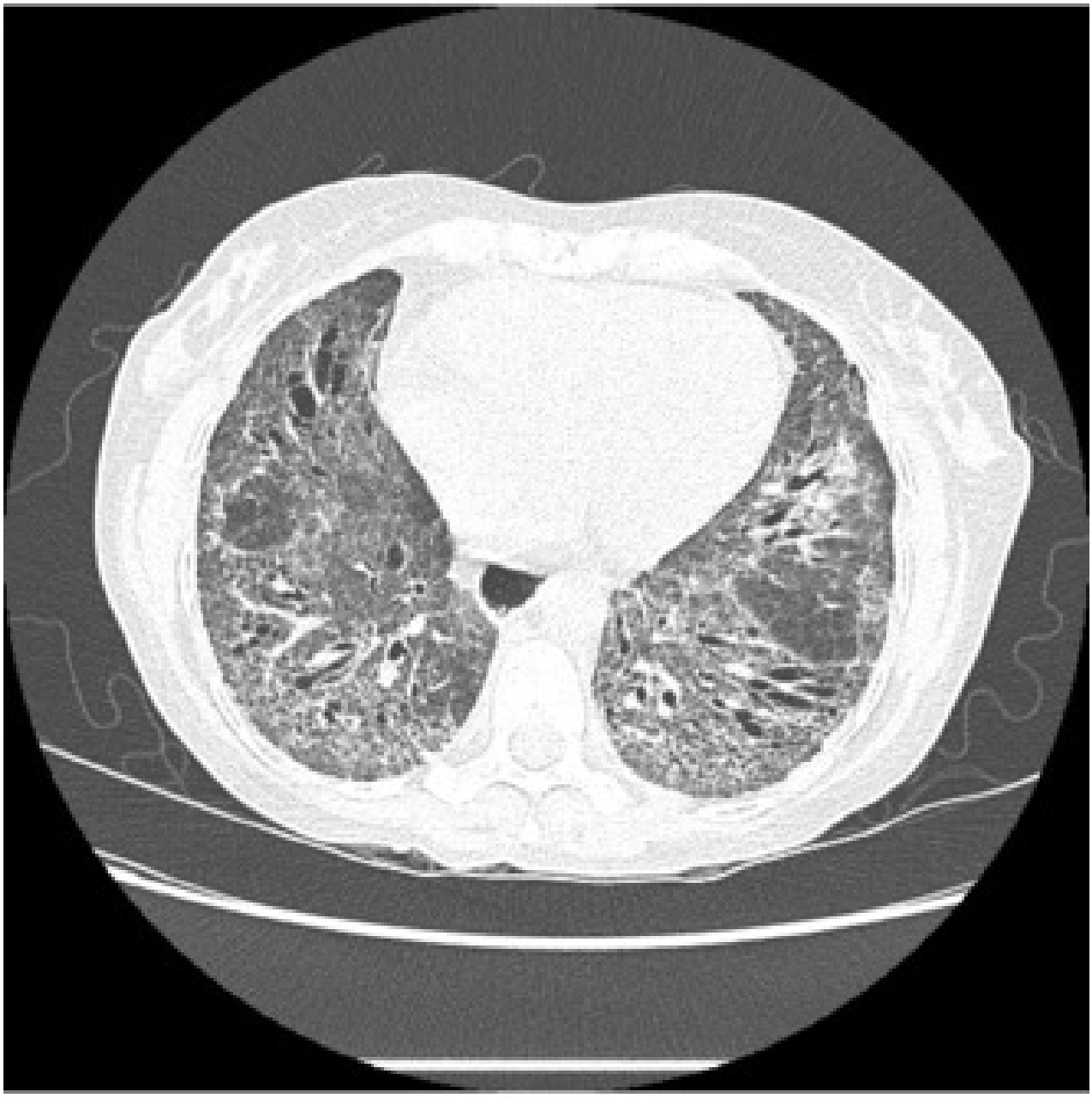

Across the 26 flawed MedQA items, neither o1 nor o3-mini explicitly identified missing figures or ambiguity in any case, suggesting that earlier-generation reasoning models lack reliable question-integrity recognition. In contrast, GPT-5.2-pro-2025-12-11 flagged 5 questions as missing required figures and identified 1 question with unresolved ambiguity, indicating an emerging but still limited ability to detect underspecified or unanswerable items. Overall, explicit detection of unanswerable or underspecified questions was infrequent, indicating that question-integrity recognition remains limited and model-dependent for medical questions.

Taxonomy of Reasoning Errors

After excluding questions with missing figures or ambiguity, 37 questions (14 from USMLE step 1 examinations and 23 from USMLE step 2 and step 3 examinations) remained for our error analysis. From these, we identified four major error categories: Information Synthesis Errors, Therapeutic Decision Errors, Diagnostic Reasoning Errors, and Foundational Principle Errors. The definitions and examples for each category are detailed in Textbox 2.

XSL•FO
RenderX

Textbox 2. Taxonomy of reasoning errors with definitions and examples. COPD: chronic obstructive pulmonary disease.

Information Synthesis Errors:

- Misjudgment of Clinical Feature Importance:
- Definition: overemphasizes irrelevant details while ignoring critical ones.
- Example 1: in a lung cancer case involving a 76-year-old man with COPD and asbestos exposure, the reasoning process overemphasized the patient's asbestos exposure and pleural plaques, while neglecting symptoms such as weight loss and anemia, and his 60-pack-year smoking history that was more indicative of malignancy.
- Example 2: in a case of chronic abdominal pain with multiorgan involvement, the focus was incorrectly placed on a recent impetigo episode and the possibility of poststreptococcal glomerulonephritis, thereby overlooking systemic evidence pointing to a chronic condition such as secondary amyloidosis.

Therapeutic Decision Errors:

- Misapplication of Evidence-Based Guidelines:
- Definition: fails to select or correctly apply clinical guidelines.
- Example 1: the reasoning process overemphasized the use of an adjunctive drug for tumor lysis syndrome while neglecting intravenous hydration, the primary strategy for managing this condition.
- Example 2: the reasoning process incorrectly selected griseofulvin despite guidelines recommending itraconazole as the first-line therapy for patients with the fungal infection *tinea corporis*.
- Inadequate Dynamic Risk-Benefit Assessment:
- Definition: fails to weigh evolving clinical risks and benefits.
- Example 1: in a case of circulatory electrolyte imbalance in hepatic encephalopathy, the reasoning process prioritized correcting hypoglycemia while neglecting hypokalemia.
- Example 2: in a case of hemoptysis with thrombolytic therapy, the reasoning process did not adequately account for the bleeding risks associated with thrombolytic therapy, underestimating the potential for life-threatening hemorrhage.
- Misinterpretation of Pharmacologic Mechanisms:
- Definition: misunderstands a medication's mechanism of action.
- Example 1: in a patient with seasonal allergies, the reasoning process incorrectly identified the drug's mechanism as competitive blockade of muscarinic receptors (used in asthma management) instead of recognizing that the appropriate decongestant works as an α-adrenergic agonist.
- Example 2: in a gout prophylaxis scenario, pancytopenia was erroneously attributed to colchicine toxicity rather than correctly identifying the mechanism of a xanthine oxidase inhibitor and its potential interaction with immunosuppressive agents.
- Premature Cognitive Closure:
- Definition: a reasoning error in which the model quickly settles on a diagnosis or management plan without sufficiently exploring alternative explanations or contributing factors.
- Example: in a case of erectile dysfunction in a patient on selective serotonin reuptake inhibitors with significant vascular risk factors, the reasoning process prematurely closed off further investigation (eg, nocturnal penile tumescence testing) that could better clarify the underlying cause.

Diagnostic Reasoning Errors:

- Failure to Integrate Pathophysiological Mechanisms:
- Definition: fails to synthesize and apply key pathophysiological principles, resulting in misattribution or misunderstanding of clinical findings.
- Example 1: when analyzing the relevant option, the model failed to link the characteristic facial anomalies (eg, low-set ears and retrognathia) with the Potter sequence.
- Example 2: the model overlooked that achlorhydria in a vasoactive intestinal peptide-secreting tumor (VIPoma) leads to impaired iron absorption, failing to integrate this pathophysiological mechanism into its diagnostic reasoning.
- Deviation from Prioritized Diagnostic Protocols:
- Definition: initiates treatment before completing necessary diagnostic workups.
-

XSL•FO
RenderX

Example 1: when evaluating a 2-month-old infant with signs of head trauma and suspicious injury patterns, the model prioritized nonmedical interventions (eg, involving social services) over obtaining an urgent head computed tomography (CT) scan, delaying critical diagnostic evaluation.

- Example 2: in the initial management of a transient ischemic attack, if the CT scan is normal, a CT angiogram is indicated to further characterize the cerebral vessels. The reasoning process, however, wrongly opted for heparin therapy.

Foundational Principle Errors:

- Misapplication of Ethical Principles:
- Definition: fails to follow proper ethical protocols.
- Example 1: in the case of a surgical complication, the reasoning process failed to follow proper communication protocols by not discussing the complication with the attending physician before reporting it.
- Example 2: in the situation where a daughter refuses consent due to an abuse history, the reasoning process neglected to involve the appropriate legal guardian, instead opting to seek immediate court intervention without first contacting the next of kin.
- Misinterpretation of Statistical Concepts:
- Definition: misinterprets statistical principles.
- Example 1: the reasoning process failed to recognize that the *P* value is computed under the assumption that the null hypothesis is true.
- Example 2: the reasoning process failed to differentiate between lead-time bias and measurement bias.

Comparison of Errors Across Different LLMs

We tested other leading models on the 37 challenging questions. Accuracy was 49% for o3-mini, 41% for GPT-4.5, and 38% for DeepSeek-R1. Reasoning process lengths varied significantly, from an average of 1319 characters for o3-mini to 11,055 characters for DeepSeek-R1.

An analysis of the reasoning traces revealed that different models used highly distinct problem-solving strategies, especially on questions requiring multistep synthesis (Figures 3 and 4). For example, one question required understanding that immunoglobulin A (IgA) deficiency is associated with celiac disease and that these patients are typically deficient in fat-soluble vitamins.

As shown in Figure 3, the o1, GPT-4.5, and o3-mini models attempted a linear process but failed for different reasons. They first identified the key patient details, then developed an initial differential diagnosis, and finally assessed the available answer options. The o1 model recognized the possibility of IgA deficiency and evaluated each option, linking option A (Hemolytic anemia and ataxia) with vitamin E deficiency. However, it did not connect the vitamin E deficiency back to IgA deficiency. In the GPT-4.5 process, the analysis of option A led directly to its association with certain hereditary disorders, such as Friedreich ataxia and other conditions involving neurological and hematological symptoms, resulting in the elimination of what was actually the correct choice. In contrast, the o3-mini model exhibited a shortcut: it noted, “Although the answer choices do not directly mention IgA deficiency or anti-IgA antibodies, the only option referring to antibodies (which in this context are responsible for transfusion reactions) is option D: anti-A, B, or O antibodies in the serum.”

Meanwhile, DeepSeek-R1’s process was fundamentally different (Figure 4). It used an iterative, 3-round analysis, repeatedly reassessing the problem, highlighting a completely distinct cognitive architecture from the other models. The detailed reasoning processes for each model are listed in Multimedia Appendix 4.

When tested on questions that o1 failed, the models displayed varied error patterns (Table 1). Information Synthesis Errors were most frequent in OpenAI o1 (11 errors), while Therapeutic Decision Errors were common across all models. These results highlight model-specific differences in reasoning failure modes.

XSL•FO
RenderX

Figure 3. Examples of identified errors in the reasoning traces of OpenAI o3-mini, OpenAI o1, and GPT-4.5. BP: blood pressure; IgA: immunoglobulin A; IV: intravenous.

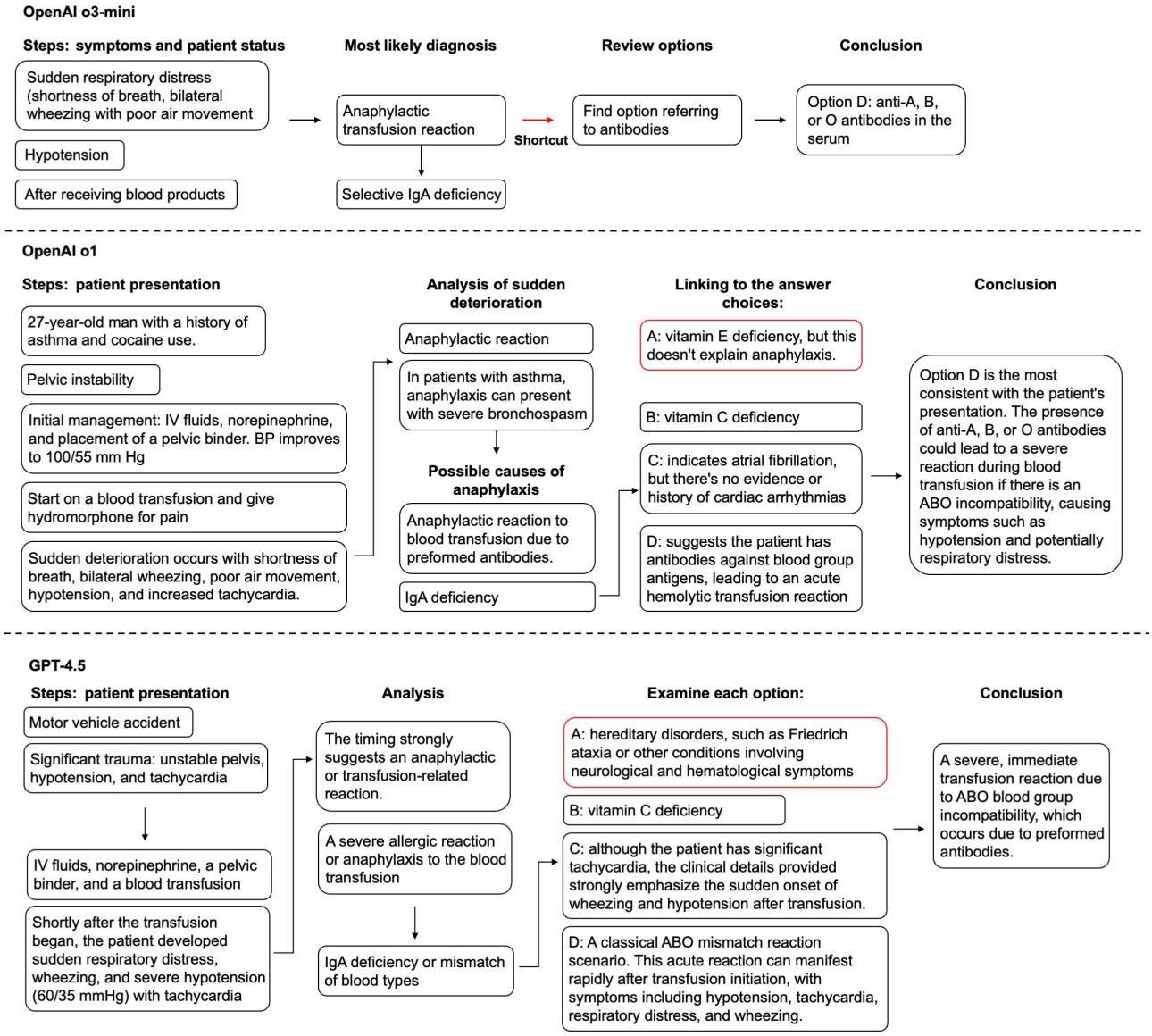

XSL•FO
RenderX

Figure 4. Examples of identified errors in the reasoning traces of DeepSeek-R1. DIC: disseminated intravascular coagulation; GCS: Glasgow Coma Scale; IgA: immunoglobulin A; IV: intravenous; TRALI: transfusion-related acute lung injury.

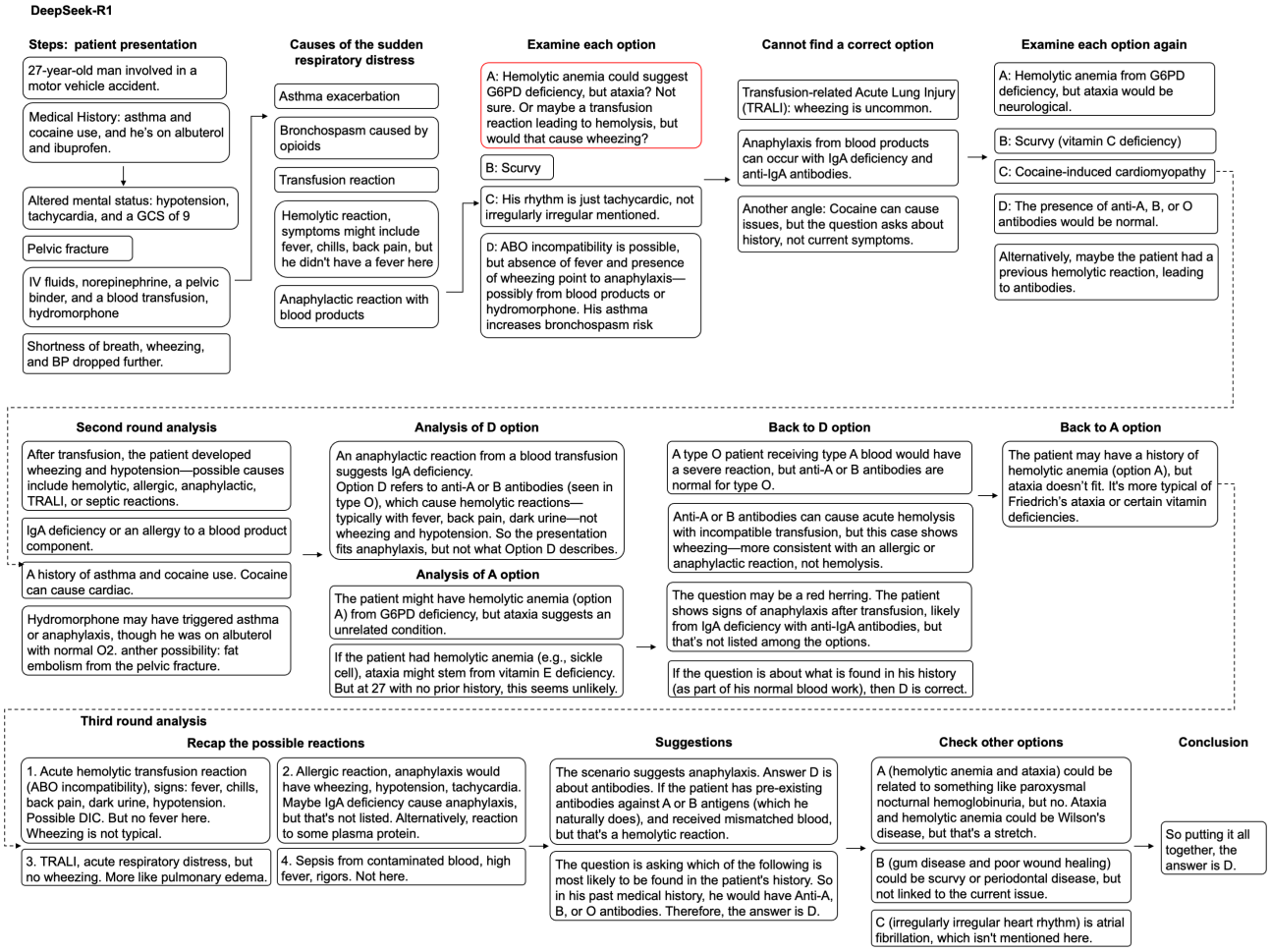

XSL•FO
RenderX

Table . Distribution of error types across 4 large language models for 37 challenging questions.

Error category and subcategory	OpenAI o1	GPT-4.5	OpenAI o3-mini	Deepseek-R1
Information Synthesis Errors	11	5	4	3
Misjudgment of Clinical Feature Importance	11	5	4	3
Therapeutic Decision Errors	11	6	7	8
Misapplication of Evidence-Based Guidelines	4	0	0	2
Inadequate Dynamic Risk-Benefit Assessment	3	1	1	1
Misinterpretation of Pharmacologic Mechanisms	3	4	3	2
Premature Cognitive Closure	1	1	3	3
Diagnostic Reasoning Errors	9	8	4	8
Failure to Integrate Pathophysiological Mechanisms	5	6	3	7
Deviation from Prioritized Diagnostic Protocols	4	2	1	1
Foundational Principle Errors	6	3	4	4
Inappropriate Ethical Decision-Making	4	2	3	2
Misinterpretation of Statistical Concepts	2	1	1	2
Total	37	22	19	23

Feature Steering via SAE

Steering with reasoning-specific features enhanced model accuracy, with significant gains on MedQA (Table 2; χ^2_1=10.9; *P*=.002) and PubMedQA (χ^2_1=18.6; *P*<.001) and a consistent positive trend on MedMCQA (χ^2_1=3.9; *P*=.15). The most substantial gains were observed at a moderate steering strength of 2, which increased MedQA accuracy from 0.568 to 0.597 (95% CI 0.584-0.610) and PubMedQA accuracy from 0.708 to 0.739 (95% CI 0.722-0.756). This performance improvement, however, coincided with a significant increase in the verbosity of the model's reasoning traces. At a steering strength of 4, the average reasoning token count nearly doubled for MedQA (Wilcoxon W=6,010,329; *P*<.001) and tripled for PubMedQA (Wilcoxon W=257,830; *P*<.001). The intervention had a limited effect on hallucination frequency; a significant increase was observed for MedMCQA at strength 2 (Mann-Whitney *U*=1,364,912; *P*=.008), with no significant changes for MedQA (*U*=481,828; *P*=.99) or PubMedQA (*U*=33,695; *P*=.99). Notably, the type of hallucination remained consistent within each dataset regardless of steering. Chi-square analysis revealed a significant shift in the distribution of hallucination types for MedMCQA (χ^2_2=8.1; *P*=.02), but not for MedQA (χ^2_2=1.4; *P*=.49) or PubMedQA (χ^2_2=1.1; *P*=.58). Across all conditions, factual hallucinations were the most common type in MedMCQA and MedQA, whereas input hallucinations predominated in PubMedQA (Table 4). Finally, no significant correlations were identified between reasoning length and key performance metrics such as accuracy, hallucination count, or hallucination severity. These correlations were computed at the level of individual questions within each steering condition. Expert validation achieved 96% precision, with the 4 disagreement cases representing conservative overflagging of borderline medical claims rather than clear errors. The claim-level Pearson correlation between the LLM judge and human annotator severity ratings was 0.86.

XSL•FO
RenderX

Table . Model performance across steering strengths on 3 medical question-answering (QA) benchmarks. Values are reported as point estimates with 95% CIs computed from question-level paired data.

Dataset and strength	Accuracy (95% CI)	Reasoning tokens (95% CI)	Hallucinations[a] (95% CI)	Factual hallucinations[b] (95% CI)	Severity[c] (0 - 3) (95% CI)
MedMCQA					
None	0.505 (0.503-0.507)	1287 (1237-1338)	2.2 (2.154-2.247)	2.15 (2.104-2.197)	2.199 (2.186-2.212)
2	0.522 (0.508-0.536)	1500 (1294-1706)	2.378 (2.284-2.472)	2.284 (2.191-2.377)	2.194 (2.168-2.221)
4	0.525 (0.511-0.538)	2457 (2065-2850)	2.163 (2.066-2.260)	2.107 (2.011-2.203)	2.203 (2.176-2.230)
MedQA					
None	0.568 (0.565-0.570)	1682 (1587-1777)	2.605 (2.515-2.695)	2.488 (2.398-2.577)	2.265 (2.241-2.289)
2	0.597 (0.584-0.610)	1886 (1671-2101)	2.557 (2.456-2.658)	2.456 (2.356-2.556)	2.250 (2.223-2.277)
4	0.589 (0.575-0.602)	2912 (2520-3303)	2.658 (2.554-2.762)	2.531 (2.429-2.633)	2.225 (2.198-2.252)
PubMedQA					
None	0.708 (0.704-0.712)	670 (652-687)	1.064 (0.954-1.174)	0.384 (0.309-0.459)	1.837 (1.785-1.888)
2	0.739 (0.722-0.756)	861 (731-991)	1.08 (0.970-1.190)	0.398 (0.319-0.477)	1.787 (1.735-1.839)
4	0.737 (0.719-0.754)	2187 (1696-2677)	0.962 (0.855-1.069)	0.336 (0.265-0.407)	1.734 (1.677-1.790)

[a]Average count of hallucination claims per reasoning trace.

[b]Average count of factual hallucination claims per reasoning trace.

[c]0 - 3 rating by a large language model judge (GPT-5-mini) and higher = worse.

Functional Analysis of Reasoning Features

Feature overlap analysis revealed distinct generalization patterns across medical domains. At strength 4, four features (IDs: 10602, 29040, 37660, and 56984) consistently ranked in the top 20 across all 3 datasets, suggesting robust utility. MedQA and MedMCQA showed the strongest feature similarity, sharing 13 features in their top 20 at strength 4, while PubMedQA demonstrated more domain-specific feature requirements. Higher SAE strength (4 vs 2) consistently improved cross-dataset feature generalization. Feature overlap heatmaps are shown in Multimedia Appendix 5. Manual analysis of the top 15 features allowed us to group them into 5 functional categories that aligned with our error taxonomy (Multimedia Appendix 6). These categories included (1) cue-weighting calibration and distractor suppression, which focuses the model on pivotal clinical data; (2) protocol alignment, which steers outputs toward guideline-consistent procedures; (3) mechanistic grounding, which connects decisions to core pathophysiological or pharmacological principles; (4) rule/criteria enforcement, which ensures the precise application of formal definitions such as Light's criteria; and (5) evidence synthesis and question reframing, which helps the model correctly interpret the core question being asked. Several features were multifunctional, contributing to more than one reasoning category (eg, feature 56,984).

Discussion

Principal Findings

In this study, we conducted a mixed methods analysis to audit the integrity of a widely used medical benchmark, characterize LLM reasoning errors, and test a mechanistic intervention using SAEs. Our findings reveal that a significant portion of apparent model failures on MedQA are attributable to intrinsic data flaws, true reasoning errors can be categorized into a clinically relevant taxonomy, and steering reasoning-specific features can effectively improve LLM accuracy on medical QA benchmarks.

Rethinking Benchmarks: The Problem Lies in the Questions

An important finding of our work is that 41% of the initial incorrect answers generated by OpenAI o1 were due to flawed benchmark questions, including 14 missing necessary figures and 12 containing ambiguities that have since been updated in their source question banks. These integrity findings are specific

XSL•FO
RenderX

to MedQA; we did not perform the same source reconciliation for MedMCQA or PubMedQA.

This audit indicates that MedQA, as instantiated in widely used benchmark distributions, contains a nontrivial fraction of items whose fidelity to the original sources is compromised (eg, missing figures or subsequently corrected ambiguity). These issues can confound model evaluation by attributing errors to models that may instead reflect dataset drift or modality loss introduced during benchmark construction.

More broadly, these findings underscore a structural risk in deriving medical benchmarks from examination question banks without preserving all clinically relevant modalities. In clinical problem-solving, visual data (eg, imaging, waveforms, or figures) are often integral to reasoning, and their omission fundamentally alters the task being evaluated. Under these conditions, a model's apparent success on a flawed question may plausibly arise from pattern matching rather than robust clinical reasoning. This concern is consistent with prior work showing substantial performance drops when multiple-choice shortcuts are disrupted, such as replacing the correct option with "none of the above," which reduced accuracy by 8% - 38% [14].

As the field relies heavily on such benchmarks to gauge progress, our findings serve as a critical call to action for improved validation and management of medical benchmarks. Unlike static domains such as mathematics or coding, medical knowledge is constantly evolving (eg, clinical practice guidelines are updated and new drugs emerge) [26]. To remain relevant, medical benchmarks must be treated as dynamic resources, continuously aligned with the current state of practice. This necessitates a move toward more rigorously curated, version-controlled, and multimodally complete datasets, supported by automated methods for ongoing validation. In addition, when maintainers of source question banks identify and correct errors, a corresponding process must be in place to update the benchmark promptly. Without such reliable evaluation tools, we risk overestimating the capabilities of current models and misdirecting development efforts.

A Taxonomy of Reasoning Failures: From "What" to "Why"

By isolating the 37 model errors, we developed a 4-category taxonomy that shifts the focus from whether a model is wrong to why it is wrong. The prevalence of Information Synthesis Errors (particularly for OpenAI o1) and Therapeutic Decision Errors across all tested models highlights vulnerabilities in processes central to clinical practice: weighing evidence, applying guidelines, and performing dynamic risk-benefit assessments. The observation that different state-of-the-art models exhibit distinct error profiles (Table 1) suggests that their underlying architectures and training data instill unique cognitive biases. For instance, some models may be prone to premature closure, while others struggle to integrate complex pathophysiological mechanisms. This granular, qualitative understanding of failure modes is essential for identifying high-risk scenarios and is a necessary prerequisite for building safer, more reliable systems.

A Mechanistic Path Toward Safer AI: From Observation to Intervention

Another contribution of this study is the bridge between the qualitative error taxonomy and the quantitative SAE intervention. By identifying and steering reasoning-specific features, we not only improved accuracy on multiple medical benchmarks but also uncovered the functional roles of these features. Some clear alignments were identified: feature groups such as "prioritizes critical information and filters out irrelevant data" and "protocol alignment" mechanistically address the error categories of "Misjudgment of Clinical Feature Importance" and "Deviation from Prioritized Diagnostic Protocols," respectively (Multimedia Appendix 6). These results provide an initial proof of concept that interpretable, reasoning-specific SAE features can be modulated to shift overall accuracy and that the resulting feature categories conceptually correspond to several of the error types in our taxonomy. They do not, however, demonstrate that steering individual feature subgroups causally corrects the specific error categories with which they conceptually align; such a claim would require single-feature or feature-group interventions on a larger error-annotated corpus drawn from the same model, which we identify as a key direction for future work.

This intervention also revealed a trade-off. Steering with strengthened reasoning-specific features increased accuracy, particularly on MedQA and PubMedQA, but also significantly lengthened the reasoning traces, more than doubling the token count in some cases. Notably, however, the relationship between steering strength and accuracy was nonmonotonic: increasing strength from 2 to 4 more than doubled token counts across all 3 benchmarks, yet accuracy slightly declined in some cases (eg, MedQA: 0.597-0.589). This pattern, combined with the absence of significant correlations between reasoning trace length and any performance metric, indicates that improved accuracy is not a simple byproduct of increased verbosity or test-time computation. Rather, it likely results from the targeted activation of specific, high-value reasoning pathways. The distinction between group-level and individual-level effects is important here: while steering shifts the mean accuracy and mean trace length upward relative to baseline, within a given condition, longer traces do not predict correct answers for individual questions, further suggesting that the benefit is feature-specific rather than length-dependent. We note that prompt-based interventions designed to elicit longer reasoning operate at the input level rather than directly modifying internal representations and therefore target a different causal pathway than activation steering; exploring such comparisons remains a valuable direction for future work. More broadly, the increased token cost associated with steering presents a practical design challenge. This work represents a proof of concept; potential strategies for mitigating this trade-off in future implementations could include selective steering for high-uncertainty cases or adaptive strength calibration.

Strengths and Limitations

This study's primary strength lies in its novel mixed methods approach, which combines a qualitative analysis of benchmark integrity and LLM reasoning errors with a quantitative,

XSL•FO
RenderX

mechanistic intervention using SAEs. This provides a uniquely holistic view of the challenges and opportunities in improving LLM performance on medical reasoning benchmarks.

However, several limitations remain. First, we used different models across study stages: the error taxonomy was developed on frontier models (eg, OpenAI o1), while SAE training and steering were conducted on a single open-source distilled model (DeepSeek-R1-Distill-Llama-8B). Because distilled student models can develop internal representations distinct from their teachers [27,28], the alignment between our taxonomy and the 8B model's SAE features is suggestive rather than direct mechanistic evidence. Consequently, these identified features are not directly transferable to closed-source models. Second, our error taxonomy is based on 37 incorrect responses from the o1 model, which may not capture the full spectrum of reasoning failures. Specifically, it misses "silent failures," cases where incorrect reasoning coincidentally yields a correct answer. Characterizing these hidden errors through expert review of complete reasoning traces remains an important direction for future work. Third, using an LLM-as-a-judge to evaluate hallucinations is scalable but introduces potential biases. Our human validation revealed conservative overflagging by the LLM judge. Furthermore, comprehensively estimating recall (false negatives) was infeasible because of the massive data volume and specialized clinical expertise required. As a result, our reported hallucination counts should be interpreted as lower-bound estimates. Finally, there are technical and evaluative constraints regarding the SAE intervention. Our evaluation relies on QA benchmarks, which do not fully replicate the dynamic, unstructured nature of real-world clinical decision-making. Technically, the SAE training was limited to a single layer (layer 19) and faced common interpretability challenges, such as incomplete resolution of polysemanticity and the difficulty of completely disentangling reasoning from factual features. Future work should compare this approach against steering random, nonreasoning features and extend evaluations to real patient data.

Future Work

Our findings point toward several key directions for future research. First, there is an urgent need for the community to develop validation and management methods for maintaining medical benchmarks, ensuring fair and reproducible evaluations. Second, the reasoning-specific features identified by SAEs could be integrated into more advanced training paradigms. For example, they could inform process-based reward models in reinforcement learning to explicitly penalize flawed reasoning pathways identified in our taxonomy, potentially training models that are both accurate and efficient. In addition, the error taxonomy developed here could inform the design of targeted medical test cases that intentionally trigger specific failure modes, serving both as a validation tool for the taxonomy and as a stress-testing framework for evaluating model robustness across known error categories. For clinical translation, our work underscores that accuracy is an insufficient metric for safety and reliability. The ability to audit, understand, and even steer a model's reasoning process, as demonstrated here, represents a critical step toward building the "glass-box" systems necessary to earn clinician trust and ensure patient safety [29].

Conclusion

This study identified that 41% of initial model errors on MedQA reflected benchmark integrity issues, including missing figures and subsequently corrected ambiguities, rather than true model failures. Among the 37 confirmed reasoning errors, inductive analysis yielded a 4-category taxonomy (Information Synthesis, Therapeutic Decision, Diagnostic Reasoning, and Foundational Principle Errors) that revealed distinct failure profiles across 4 frontier LLMs. Steering reasoning-specific SAE features significantly improved accuracy on MedQA and PubMedQA, with a consistent positive trend on MedMCQA, while also increasing reasoning trace length, with no significant correlation between verbosity and performance. These findings demonstrate that medical LLM evaluation is constrained by flawed benchmarks and that reasoning failures follow identifiable, model-specific patterns with the potential for mechanistic correction via feature steering.

Acknowledgments

The authors used generative artificial intelligence solely to identify key citations in the literature review. These systems were not involved in data collection, data analysis, study design, or paper drafting. The authors have reviewed the paper and accept full responsibility for it.

Funding

This work was supported by the National Institutes of Health (NIH) grant R00LM014097-02. The NIH had no role in the design and conduct of the study; the collection, management, analysis, and interpretation of the data; the preparation, review, or approval of the paper; and the decision to submit the paper for publication.

Data Availability

The prompts are listed in Multimedia Appendix 1. MedQA questions with missing figures and ambiguous MedQA questions updated in source question banks are reported in Multimedia Appendices 2 and 3. The datasets used in this study are publicly available: MedQA (GitHub), MedMCQA (GitHub), and PubMedQA (GitHub). The training data for the sparse autoencoder are publicly available: LMSYS-Chat-1M (Hugging Face) and OpenThoughts-114k (Hugging Face). The code for the official ReasonScore implementation is available on GitHub.

XSL•FO
RenderX

Authors' Contributions

SL contributed to conceptualization, data curation, formal analysis, investigation, methodology, software development, and writing of the original draft. JL contributed to conceptualization, data curation, investigation, methodology, and writing of the original draft. AW contributed to the conceptualization, review, and editing of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Large language model prompt for answer and reasoning generation.
[DOCX File, 17 KB - jmir_v28i1e90061_app1.docx]

Multimedia Appendix 2
MedQA questions with missing figures.
[DOCX File, 87 KB - jmir_v28i1e90061_app2.docx]

Multimedia Appendix 3
Ambiguous MedQA questions updated in source banks.
[DOCX File, 35 KB - jmir_v28i1e90061_app3.docx]

Multimedia Appendix 4
Example of a complex question failed by all large language models (OpenAI o1, OpenAI o3-mini, and OpenAI GPT-4.5, and DeepSeek-R1).
[DOCX File, 28 KB - jmir_v28i1e90061_app4.docx]

Multimedia Appendix 5
Feature overlap heatmaps across medical datasets. For each dataset and steering strength, features were ranked by their poststeering accuracy (exact match) on that dataset. The heatmaps show the set intersection count of top-K feature lists between each dataset pair, that is, $|\text{Top-K}(A) \cap \text{Top-K}(B)|$, for K=5, 10, 15, and 20 at steering strengths of 2 and 4. Diagonal entries equal K.
[DOCX File, 237 KB - jmir_v28i1e90061_app5.docx]

Multimedia Appendix 6
Functional analysis of top 15 reasoning-specific sparse autoencoder features.
[DOCX File, 19 KB - jmir_v28i1e90061_app6.docx]

References

1. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. arXiv. Preprint posted online on Mar 4, 2022. [doi: 10.48550/arXiv.2203.02155]
2. Liu S, Wright AP, Patterson BL, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 2023 Jun 20;30(7):1237-1245. [doi: 10.1093/jamia/ocad072] [Medline: 37087108]
3. Tai-Seale M, Baxter SL, Vaida F, et al. AI-generated draft replies integrated into health records and physicians' electronic communication. JAMA Netw Open 2024 Apr 1;7(4):e246565. [doi: 10.1001/jamanetworkopen.2024.6565] [Medline: 38619840]
4. Liu S, McCoy AB, Wright AP, et al. Leveraging large language models for generating responses to patient messages—a subjective analysis. J Am Med Inform Assoc 2024 May 20;31(6):1367-1379. [doi: 10.1093/jamia/ocae052]
5. Van Veen D, Van Uden C, Blankemeier L, et al. Clinical text summarization: adapting large language models can outperform human experts. arXiv. Preprint posted online on Sep 14, 2023. [doi: 10.48550/arXiv.2309.07430]
6. Liu S, McCoy AB, Wright AP, et al. Why do users override alerts? Utilizing large language model to summarize comments and optimize clinical decision support. J Am Med Inform Assoc 2024 May 20;31(6):1388-1396. [doi: 10.1093/jamia/ocae041] [Medline: 38452289]
7. DeepSeek-AI, Guo D, Yang D, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. arXiv. Preprint posted online on Jan 22, 2025. [doi: 10.48550/arXiv.2501.12948]
8. Xie Y, Wu J, Tu H, et al. A preliminary study of o1 in medicine: are we closer to an AI doctor? arXiv. Preprint posted online on Sep 23, 2024. [doi: 10.48550/arXiv.2409.15277]
9. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 2025 Jan 28;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]

XSL•FO
RenderX

10. Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. J Am Med Inform Assoc 2025 Apr 1;32(4):605-615. [doi: 10.1093/jamia/ocaf008] [Medline: 39812777]
11. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci (Basel) 2020;11(14):6421. [doi: 10.3390/app11146421]
12. Raji ID, Daneshjou R, Alsentzer E. It's time to bench the medical exam benchmark. NEJM AI 2025 Jan 23;2(2). [doi: 10.1056/AIe2401235]
13. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun 28;25:e48568. [doi: 10.2196/48568] [Medline: 37379067]
14. Bedi S, Jiang Y, Chung P, Koyejo S, Shah N. Fidelity of medical reasoning in large language models. JAMA Netw Open 2025 Aug 1;8(8):e2526021. [doi: 10.1001/jamanetworkopen.2025.26021] [Medline: 40779272]
15. Yu F, Gao A, Wang B. OVM, outcome-supervised value models for planning in mathematical reasoning. arXiv. Preprint posted online on Nov 16, 2023. [doi: 10.48550/arXiv.2311.09724]
16. Lightman H, Kosaraju V, Burda Y, et al. Let's verify step by step. arXiv. Preprint posted online on May 31, 2023. [doi: 10.48550/arXiv.2305.20050]
17. Cunningham H, Ewart A, Riggs L, Huben R, Sharkey L. Sparse autoencoders find highly interpretable features in language models. arXiv. Preprint posted online on Sep 15, 2023. [doi: 10.48550/arXiv.2309.08600]
18. Galichin A, Dontsov A, Druzhinina P, et al. I have covered all the bases here: interpreting reasoning features in large language models via sparse autoencoders. arXiv. Preprint posted online on Mar 24, 2025. [doi: 10.48550/arXiv.2503.18878]
19. Dejong G, Horn SD, Gassaway JA, Slavin MD, Dijkers MP. Toward a taxonomy of rehabilitation interventions: using an inductive approach to examine the "black box" of rehabilitation. Arch Phys Med Rehabil 2004 Apr;85(4):678-686. [doi: 10.1016/j.apmr.2003.06.033] [Medline: 15083447]
20. Zheng L, Chiang WL, Sheng Y, et al. LMSYS-chat-1M: a large-scale real-world LLM conversation dataset. arXiv. Preprint posted online on Sep 21, 2023. [doi: 10.48550/arXiv.2309.11998]
21. Guha E, Marten R, Keh S, et al. OpenThoughts: data recipes for reasoning models. arXiv. Preprint posted online on Jun 4, 2025. [doi: 10.48550/arXiv.2506.04178]
22. SAEDashboard. GitHub. URL: https://github.com/jbloomAus/SAEDashboard [accessed 2025-07-27]
23. Stolfo A, Balachandran V, Yousefi S, Horvitz E, Nushi B. Improving instruction-following in language models through activation steering. arXiv. Preprint posted online on Oct 15, 2024. [doi: 10.48550/arXiv.2410.12877]
24. Turner AM, Thiergart L, Leech G, et al. Steering language models with activation engineering. arXiv. Preprint posted online on Aug 20, 2023. [doi: 10.48550/ARXIV.2308.10248]
25. SAELens. GitHub. URL: https://github.com/jbloomAus/SAELens [accessed 2026-06-03]
26. Vernooij RWM, Sanabria AJ, Solà I, Alonso-Coello P, Martínez García L. Guidance for updating clinical practice guidelines: a systematic review of methodological handbooks. Implement Sci 2014 Jan 2;9(1):3. [doi: 10.1186/1748-5908-9-3] [Medline: 24383701]
27. Haskins R, Adams B. Distilled circuits: a mechanistic study of internal restructuring in knowledge distillation. arXiv. Preprint posted online on May 16, 2025. [doi: 10.48550/arXiv.2505.10822]
28. Baek DD, Tegmark M. Towards understanding distilled reasoning models: a representational approach. arXiv. Preprint posted online on Mar 5, 2025. [doi: 10.48550/arXiv.2503.03730]
29. Liu S, McCoy AB, Peterson JF, et al. Leveraging explainable artificial intelligence to optimize clinical decision support. J Am Med Inform Assoc 2024 Apr 3;31(4):968-974. [doi: 10.1093/jamia/ocae019] [Medline: 38383050]

Abbreviations

AI: artificial intelligence
API: application programming interface
IgA: immunoglobulin A
LLM: large language model
QA: question-answering
SAE: sparse autoencoder
USMLE: United States Medical Licensing Examination

XSL•FO
RenderX

Edited by A Coristine; submitted 20.Dec.2025; peer-reviewed by C Wu, C Ma, GLP Bodagala, R Kurmashev; revised version received 15.May.2026; accepted 18.May.2026; published 12.Jun.2026.

Please cite as:
Liu J, Liu S, Wright A
Benchmark Integrity and Reasoning-Trace Errors in Medical Question Answering With Large Language Models: Mixed Methods Study With Sparse Autoencoders
J Med Internet Res 2026;28:e90061
URL: https://www.jmir.org/2026/1/e90061
doi:10.2196/90061

© Jialin Liu, Siru Liu, Adam Wright. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

An Evaluation of Pretrained Generative Models for Augmenting Small Health Data: Comparative Modeling Study

Margerie Huet-Dastarac[1,2], PhD; Fida K Dankar[2], PhD; Dan Liu[1,2], PhD; Samer El Kababji[1,2], PhD; Lisa Pilgram[1,2,3], MD; Khaled El Emam[1,2], PhD

[1]School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, 451 Smyth Rd, Ottawa, ON, Canada

[2]Research Institute, Children's Hospital of Eastern Ontario, Ottawa, ON, Canada

[3]Department of Nephrology and Medical Intensive Care, Charité - Universitaetsmedizin Berlin, Berlin, Germany

Corresponding Author:
Khaled El Emam, PhD
School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, 451 Smyth Rd, Ottawa, ON, Canada

Abstract

Background: Synthetic data generation (SDG) has emerged as a promising solution to address data scarcity in health care, where privacy concerns, regulatory barriers, and the high cost of data acquisition limit access to real patient datasets. Machine learning models in this domain often operate in low-data regimes, with training set sizes as low as 20 and a median dataset size of around 600 records—conditions that hinder model generalization and increase the risks of overfitting and bias. SDG addresses these challenges by producing artificial samples that mimic real-world patient data, enabling robust and privacy-preserving model development.

Objective: This study was a comprehensive assessment of SDG-augmented training across a wide array of models—both pretrained and non-pretrained—for outcome prediction in 13 health care datasets. For small datasets of sizes 50 and 350 records, we answer 3 key questions: (1) Do pretrained SDG models generate more effective augmentations than their non-pretrained counterparts for small datasets? (2) Is augmentation beneficial for both pretrained and non-pretrained classifiers for small datasets? (3) Among 3 state-of-the-art classification models, which offers the best predictive performance on small datasets? The workload that this study aimed to improve was binary classification.

Methods: The 3 classifiers considered were light gradient boosting trees, large language models (LLMs) adapted to tabular data, and Tabular Prior-Data Fitted Network (TabPFN), a transformer-based method that has become the new state of the art in terms of tabular data classification. Each classifier was augmented through different SDG methods: current state-of-the-art techniques (Bayesian networks, conditional tabular generative adversarial networks, tabular variational autoencoders, and sequential trees) and the use of LLMs for tabular data generation.

Results: Augmented TabPFN demonstrated superior performance, yielding significantly higher area under the curve and integrated calibration index scores compared to other classifiers. Post hoc analysis revealed that, for the dataset sizes examined, SDG and LLM models exhibited overfitting tendencies. Notably, simple dataset augmentation through sampling with replacement achieved performance comparable to that of SDG-based and LLM-based augmentation methods for TabPFN, suggesting that gains were primarily driven by increased sample size rather than SDG.

Conclusions: Given its strong performance and minimal computational overhead, we recommend augmenting TabPFN through sampling with replacement as the optimal approach for small-data binary classification tasks. This method achieves performance comparable to that of more complex SDG techniques while offering substantial computational advantages.

(J Med Internet Res 2026;28:e88678) doi:10.2196/88678

KEYWORDS

binary classification; machine learning; data augmentation; synthetic data generation; tabular data; small data regime

Introduction

Background and Study Objectives

Machine learning predictive modeling applications in health care often suffer from limited access to real patient data due to privacy concerns, regulatory constraints, and the high cost of data acquisition. Recent reviews have identified that most machine learning studies rely on training models on datasets of insufficient sizes [1-5]. This shortage in data availability—referred to as a low-data regime—introduces challenges such as overfitting [1,6], biased learning, and reduced model robustness [2].

Synthetic data generation (SDG) has been proposed as a potential solution by augmenting training datasets with

XSL•FO
RenderX

artificially generated samples that closely mirror real patient data. Increasing the size of the training dataset is generally associated with improved predictive performance in machine learning models [7]. However, it remains unclear whether the benefits of augmentation arise from the fidelity of synthetic samples or simply from the increased sample size. In addition, augmentation can be interpreted as a form of regularization, where synthetic examples increase the diversity of the training data by generating additional variations from the same underlying population [8].

A key question is, therefore, whether pretrained SDG models, when used for prediction and augmentation, can perform better than non-pretrained models and improve the performance of predictive clinical classification tasks. This study presents a large-scale evaluation of the impact of data augmentation by SDG on both pretrained and non-pretrained prediction models. We consider 2 realistic low-data regime scenarios for health datasets: 50 and 350 records. This definition of a low-data regime is consistent with current practices, where the median size of the training datasets used in clinical prediction tasks can be as low as 20, with a median value around 600 records [3,9].

We structured our study to answer the following specific questions in the case of a low-data regime:

- Q1. Is data augmentation using pretrained SDG models outperforming data augmented by non-pretrained SDG models?
- Q2. What is the effect of augmentation on pretrained and non-pretrained classification models?
- Q3. What is the best clinical predictive classification model among gradient-boosted trees, large language models (LLMs), and Tabular Prior-Data Fitted Network (TabPFN)?

We challenge the assumption that augmentation through SDG is necessary for improving clinical prediction in low-data regimes by systematically comparing SDG, LLM-based augmentation, and resampling methods. We demonstrate that the gains are driven primarily by increased sample size, with simple sampling with replacement achieving performance comparable to or exceeding that of complex generative approaches.

Previous Work

For nontabular data, data augmentation has been applied to address the low-data regime problem, such as in imaging, video, and natural language processing data [10-15], and it has been shown to be a viable solution to address the problem of incomplete and unbalanced time series datasets [16-19]. However, there has been limited work on the evaluation of data augmentation in the context of tabular health data for clinical predictive workloads.

Commonly used SDG models are conditional tabular generative adversarial networks (CTGANs) [20], tabular variational autoencoders (TVAEs) [21], Bayesian normalization methods [22-25], and sequential decision trees [26-29]. However, classification models are not the only ones affected by the small dataset size available. In fact, the SDG models themselves may experience overfitting when trained on small datasets, raising questions about the quality of synthetic samples under the low-data regime.

Recent work has highlighted the potential of pretrained transformer models, such as LLMs, for tasks involving tabular data, specifically SDG and classification [30,31]. LLMs have been adapted from their original textual domain to tabular data through methods that account for dataset-specific properties, such as column-order and row invariance. Fine-tuning such models enables their application not only to classification but also to SDG [32]. Several methods—including Curated LLM (CLLM) [33], LLMOverTab [34], and Pred-LLM [35]—explicitly leverage LLM pretraining on large datasets for tabular data generation in low-data regimes.

However, recent work has noted that LLMs may not yet perform classification at a level comparable to traditional machine learning models, underscoring the importance of systematically evaluating LLMs in clinical contexts [36]. Recent models such as the TabPFN were designed for classification and represent a transformer architecture pretrained on synthetic tabular data. While TabPFN demonstrates promising performance on datasets ranging from 650 to 10,000 records [37], this size range fails to address the reality of clinical prediction tasks, where datasets commonly contain fewer than 600 records [3,9]. This gap is particularly significant given that many clinical prediction tasks must operate in low-data regimes due to data privacy constraints and the rarity of certain conditions.

Methods

The workload that this study aimed to improve was predictive binary classification.

Ethical Considerations

This project was approved by the Research Ethics Board of the Children's Hospital of Eastern Ontario Research Institute, protocol 24/80x. Because the datasets used in this study were deidentified, obtaining participant consent was waived by the Research Ethics Board of the Children's Hospital of Eastern Ontario Research Institute. This project adhered to the Declaration of Helsinki.

Models Evaluated

The pretrained and non-pretrained models used in this study are summarized in Table 1. We use the term *non-pretrained* to denote models trained directly from scratch on the available data, in contrast to LLMs and TabPFN, which rely on extensive pretraining on real and synthetic data.

XSL•FO
RenderX

Table . Overview of the models evaluated in this study.

Purpose	Non-pretrained models	Pretrained models
Classification model	• LGBM[a]	• DistilGPT2 • Llama 1B • Llama 8B fine-tuned on UltraMedical dataset • TabPFN[b] version 2
Synthetic data generation (generative models)	• Sequential decision trees • Bayesian network • CTGAN[c] • TVAE[d]	• DistilGPT2 • Llama 1B • Llama 8B fine-tuned on UltraMedical dataset

[a]LGBM: light gradient boosting machine.

[b]TabPFN: Tabular Prior-Data Fitted Network.

[c]CTGAN: conditional tabular generative adversarial network.

[d]TVAE: tabular variational autoencoder.

Non-Pretrained Generative Models

We used 4 commonly applied generative modeling methods to generate new observations for structured tabular data. CTGAN is a conditional generative adversarial network specifically adapted for tabular data, which captures complex feature distributions through adversarial training [38-40]. TVAE uses variational autoencoding to model the joint distribution of tabular features, enabling flexible data synthesis [21,41,42]. Bayesian networks represent probabilistic relationships between variables through directed acyclic graphs, allowing for the generation of synthetic data consistent with estimated dependencies [22-25]. Sequential trees generate synthetic data by recursively partitioning the feature space in a manner similar to decision trees, ensuring that complex conditional dependencies are preserved [26-29]. All 4 approaches have been widely adopted in recent work on tabular data synthesis.

Categorical and continuous features were identified based on dataset metadata. Continuous variables were normalized using training-set statistics, and categorical variables were one-hot encoded where required for the modeling task. Missing values were handled using the default mechanisms of each model (eg, a missing categorical value was treated as a valid category in the modeling). These are described in the documentation or the implementation of the generative models. Hyperparameters followed standard recommended settings as described in the original implementations. Synthetic samples were generated by unconditional sampling from the fitted models and then inverse-transformed back to the original feature space.

Sequential synthesis was implemented using Aetion Generate, a commercial product from Aetion, and the last 3 methods were implemented using an open-source Python package, Synthcity [43]. The *pysdg* library [44], our publicly available adaptation of Synthcity, provides further preprocessing and postprocessing on top of Synthcity.

Pretrained Generative Models

Fine-tuning an LLM involves adapting a pretrained model to a specific task or domain by training it on a smaller, task-specific dataset such as processing tabular data instead of free text. Fine-tuning leverages the general knowledge already encoded in the model from pretraining on vast amounts of data, allowing the model to specialize without requiring training from scratch. During fine-tuning, the model's parameters are updated to align its outputs with the desired behavior. We used the low-rank adaptation (LoRA) approach [45], which is commonly used for efficient fine-tuning.

LLMs are pretrained on large-scale text corpora and are therefore designed to process textual input and generate textual output. Several strategies have recently been proposed to adapt LLMs for tabular learning tasks. In this study, we used the PredLLM framework, which reformulates each tabular record into a natural language sentence following the pattern "*column name is value,...*" for fine-tuning the LLM. To prevent the model from exploiting column order as information, the input columns were randomly shuffled during training. Finally, the target variable was consistently placed at the end of the sequence to ensure that predictions incorporated information from all other features. The variables of a record are therefore generated column by column, completing with the outcome variable.

Relying on fine-tuning on serialized tabular training data, the LLMs were used for SDG without an explicit system prompt. Each synthetic record was generated by conditioning the model on a randomly selected feature-value pair sampled from the empirical distribution of the training data. Given this partial input, the model autoregressively completed the remaining features in a fixed predefined order learned during fine-tuning. Generated outputs were parsed back into tabular form using the known schema. More details are provided in Section B in Multimedia Appendix 1.

As described in Table B2 in Multimedia Appendix 1, we used 3 LLMs of various sizes: DistilGPT2 and Llama 1B, pretrained on general data, and Llama 8B, specifically pretrained on the UltraMedical dataset [46], which contains more than 400,000 samples of synthetic and manually curated biomedical instructions.

Non-Pretrained Classification Models

In this study, the chosen classification non-pretrained model was a light gradient boosting machine (LGBM) [47]. Tree-based

XSL•FO
RenderX

models are the most common type of machine learning prediction methods used in clinical research [3]. They perform better than linear models, such as logistic regression [48-52], and they were found to perform better than deep learning models on tabular datasets [53,54].

Model tuning used 5-fold cross-validation and Bayesian optimization [55]. The range for the tuning parameters was previously suggested [56-59], and these are summarized in Multimedia Appendix 1. High-cardinality variables were converted to embeddings [60] using a scheme similar to target encoding.

Pretrained Classification Models

The same pretrained models used for the generation of data were also used for classification. Their application to classification tasks can be seen as only the last step of the generation process, where the outcome variable is generated for a record based on all previous variables.

Another model, TabPFN, is a transformer-based model designed to perform classification and regression tasks on tabular data. It is trained on a large corpus of synthetic classification tasks with known Bayesian-optimal solutions, from which the model learns to approximate posterior class probabilities directly from features and labels without requiring further training on new tasks. This allows it to generalize effectively across diverse tabular datasets and make predictions quickly, particularly excelling in low-data regimes. Unlike traditional machine learning models that rely on iterative training and hyperparameter tuning, TabPFN offers fast, zero-shot inference through an in-context learning mechanism.

Research Questions

RQ1: Is Data Augmentation Generated by Pretrained SDG Models Outperforming Data Augmented by Non-Pretrained SDG Models?

We trained 2 classifiers with augmented datasets: LGBM and TabPFN. The synthetic samples of the augmented data were generated by the 3 pretrained SDG approaches based on different LLMs and 4 non-pretrained models.

The 2 classifiers were trained for binary classification tasks. We assessed the downstream utility by reporting the area under the curve (AUC) score, integrated calibration index (ICI), and the corresponding n′ (the number of synthetic samples).

We performed 1-tailed paired permutation tests on both the AUC and ICI metrics across the datasets to comprehensively evaluate whether pretrained SDG models outperform non-pretrained SDG models. A 1-tailed permutation test was chosen because our hypothesis was directional—namely, that pretrained SDG models would outperform non-pretrained SDG models under the small data regime by leveraging knowledge from their pretraining.

We selected the models yielding the highest AUC and the ones yielding the lowest ICI in each category (pretrained and non-pretrained) for each dataset and performed the permutation tests. Testing AUC allows us to determine if one model demonstrates significantly better discrimination—the ability to correctly rank outcomes—whereas testing ICI evaluates whether one model provides more accurate probability estimates through improved calibration.

RQ2: What Is the Effect of Augmentation on Pretrained and Non-Pretrained Classification Models?

This question entails a comparison of whether to use data augmentation or not for pretrained and non-pretrained classifiers. We considered TabPFN as a pretrained classifier and LGBM as a non-pretrained classifier. We used the most beneficial data augmentation method, selected through the analysis answering the first question of this study (RQ1), and compared the classification results to the no augmentation baselines.

These 2 classifiers were trained on the same prediction tasks as RQ1, and the same downstream utility metrics were reported. We performed 1-tailed paired permutation statistical tests. One-tailed permutation tests were chosen because our hypothesis was directional—namely, that augmentation would improve classification performance for models under the small data regime.

RQ3: What Is the Best Clinical Prediction Classification Model Among LGBM, LLMs, and TabPFN?

To answer this final question, we compared the performance of each of the considered classifiers—LGBM, fine-tuned LLMs, and TabPFN—each in their best-performing augmentation configuration, as determined by answers to questions RQ1 and RQ2. Six 1-tailed permutation tests were performed for each low dataset regime: 3 on AUC metric and 3 on ICI—LGBM versus LLM, TabPFN versus LGBM, and TabPFN versus LLM.

For each dataset and the 2 low-data regimes, we reported in Multimedia Appendix 1 the best-performing methods in terms of aggregated AUC, corresponding ICI, and n′.

Study Design

To address our research questions, we designed a comprehensive evaluation procedure. We summarize the main tasks below and expand on some of the key ones after that.

Evaluation Scope

We evaluated 3 types of binary classifiers (LGBM; LLMs—DistilGPT2, Llama 1B, and Llama 8B; and TabPFN) on binary classification tasks with different augmentation strategies. An important limitation was that LLM classifiers were only evaluated without augmentation due to computational constraints. Fine-tuning LLMs with different quantities of synthetic samples would require 360 days using 1 NVIDIA-RTX-A6000 GPU 48 GB RAM (NVIDIA Corporation). This computational intensity makes such extensive fine-tuning impractical for the intended end users. Therefore, LLMs were used only as baseline classifiers and as SDG models. A detailed analysis of computational requirements is provided in Section D in Multimedia Appendix 1.

Dataset Preparation

We used 13 clinical datasets (summarized in Table 2 and further detailed in Section E in Multimedia Appendix 1). For each dataset, we simulated 2 low-data regimes by randomly sampling

XSL•FO
RenderX

subsets of n_0=50 and n_0=350 records. We created hold-out validation and test sets of 10,000 records each from the remaining data, which are fixed to evaluate the different models fairly. We used stratified sampling for these sets to keep the original prevalence of target classes.

Table . Summary of the 13 large real-world datasets from which 50 and 350 records are sampled, training synthetic data generation and classification models, and 10,000 records are sampled as test sets.

Dataset name	Description	Number of variables
COVID (COVID-19)	A dataset that covers COVID-19 health records of Canadians collected by Esri Canada	7
Canadian Community Health Survey (CCHS)	A pooled version of survey data across multiple years that gathers health information for the Canadian population	8
COVID Survival (Nexoid)	A secondary web-based survey dataset concerning COVID-19 survival prediction collected by the Nexoid company in London, UK	19
FDA Adverse Event Reporting System (FAERS)	A database that contains adverse events and medication error reports submitted to the FDA[a]	7
Texas Inpatient Data (Texas)	A dataset on discharges from Texas hospitals	11
Washington State Hospital Discharge (Washington)	A dataset that collects the hospital discharge information from the HCUP[b] state inpatient database for 2007	8
Basic Stand Alone Inpatient Claims (BSA)	A dataset that contains the claim-level information from 2008 Medicare inpatient claims	6
Washington State Hospital Discharge (Washington 2008)	A dataset that collects the hospital discharge information from the HCUP state inpatient database for 2008	18
California Hospital Discharge (California)	A dataset that collects the hospital discharge information from the HCUP state inpatient database for 2007	16
Florida Hospital Discharge (Florida)	A dataset that collects the hospital discharge information from the HCUP state inpatient database for 2007	12
New York Hospital Discharge (New York)	A dataset that collects the hospital discharge information from the HCUP state inpatient database for 2007	14
Medical Information Mart for Intensive Care III (MIMIC-III)	A dataset that comprises deidentified health data associated with intensive care unit admissions	13
Better Outcomes Registry & Network (BORN)	A dataset that collects data about pregnancy, birth, and childhood in the province	20

[a]FDA: Food and Drug Administration.

[b]HCUP: Healthcare Cost and Utilization Project.

Synthetic Data Generation

We used 7 SDG models: 3 pretrained models (DistilGPT2, Llama 1B, and Llama 8B) and 4 non-pretrained models (sequential decision trees, Bayesian network, CTGAN, and TVAE). For each SDG model and each dataset, we generated synthetic samples in varying quantities (n′) ranging from 5 to 10,000 records, following a geometric series (details provided in the Augmentation Scheme section). For each n′, we generated 5 synthetic datasets to account for model stochasticity and averaged the performance results across the augmented datasets.

Data Augmentation

For each combination of real data subset (n_0) and synthetic data (n′), we created augmented datasets by concatenating the real and synthetic data. This process was repeated for all SDG models and both low-data regimes (n_0=50 and n_0=350).

Model Training and Evaluation

For each classifier (except LLMs), we trained models on original data only (no augmentation), augmented data from pretrained SDG models, and augmented data from non-pretrained SDG models.

We evaluated each model's performance using the hold-out validation set, calculating AUC and ICI. For augmented datasets, we averaged the performance across the 5 synthetic datasets for each n′. Therefore, for each n′ value, we trained 5 models.

XSL•FO
RenderX

Optimal Augmentation Selection

For each dataset, classifier, and SDG model combination, we identified the optimal n′ that yielded the best AUC on the validation set. We recorded the corresponding AUC, ICI, and n′ for further analysis.

Statistical Analysis

To answer RQ1, we performed 1-tailed paired permutation tests comparing the best-performing pretrained and non-pretrained SDG models for both AUC and ICI metrics. For RQ2, we conducted 1-tailed paired permutation tests to compare augmented versus nonaugmented performance for LGBM and TabPFN for both AUC and ICI metrics. To address RQ3, we performed six 1-tailed permutation tests (3 each for AUC and ICI) comparing LGBM, LLMs, and TabPFN in their best-performing configurations.

Reporting Results

We summarized the results in tables and figures, showing the performance metrics (AUC and ICI) and optimal n′ for each combination of classifier, augmentation strategy, and dataset. We reported the outcomes of all statistical tests, indicating the magnitude and significance of the differences between methods.

This overall evaluation design allows us to systematically assess the effectiveness of data augmentation techniques, compare pretrained and non-pretrained models, and identify the best-performing clinical prediction classification models under low-data regimes, while acknowledging and accounting for practical computational constraints.

Augmentation Scheme

Augmented data represents the concatenation of the original data (size n_0) and the synthetic data (size n′). To assess performance under varying low-data regimes, we simulate 2 levels of data scarcity by randomly sampling a subset of n_0=50 and n_0=350 records from each of the 13 datasets (see Table 2 for a summary). For each dataset, we identify the optimal number of synthetic samples (n′) and determine the best-performing SDG method. Unlike prior studies, which often fix an arbitrary number of synthetic samples across datasets, our approach adapts the number of augmented examples based on dataset-specific performance criteria.

To determine the optimal n′, we selected an augmentation scheme that samples finely at the low end and coarsely at the high end of the range. Ten geometric series were created and provided n′ values varying from 5 to 10,000 records. The sizes of these synthetic datasets follow a geometric series defined by $n' = [b^{(i+4)}]$, where $b \sim N(1.5, 0.005)$ and i=1,...,25. This results in multiple augmented datasets of size $n = n_0 + n'$ for each base dataset. For each n′, 5 synthetic datasets were generated and used to augment the real data and train classifiers. To reduce the impact of generative model stochasticity, we computed the average performance across these 5 augmented datasets. The same hold-out validation real datasets of 10,000 records served to determine the n′, yielding the best AUC scores, and we used the same hold-out test datasets of 10,000 records for model performance evaluation and comparison (Figure 1).

Figure 1. Augmentation scheme. AUC: area under the curve; ICI: integrated calibration index; SDG: synthetic data generation.

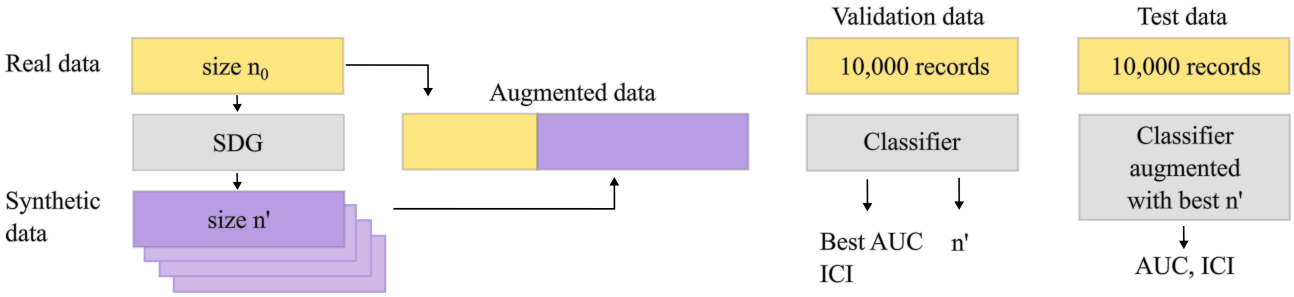

Datasets

Table 2 summarizes the 13 health datasets used in the study. These datasets cover heterogeneous domains, including public health, hospital discharge, infant and maternal health, adverse events, intensive care unit, population health surveys, and insurance claims. The table provides an overview of the datasets and the number of variables included in the binary classification models used to predict the outcome. A detailed description of each preprocessed dataset and the binary workload used for modeling can be found in Multimedia Appendix 1. The number of predictor variables in the workloads is consistent with what is seen in the clinical prediction literature [3].

Results

We report the results for each research question. Full details of the statistical analysis are provided in Section F in Multimedia Appendix 1.

Q1: Impact of Pretrained vs Non-Pretrained Augmentation

We compared pretrained and non-pretrained SDG methods across the 13 datasets using 1-tailed paired permutation tests (Table 3).

XSL•FO
RenderX

Table . Pairwise permutation tests to determine the impact of pretrained augmentation against non-pretrained augmentation. *P* values were adjusted for multiple comparisons using the Holm-Bonferroni procedure.

Null hypothesis	Alternative	AUC[a]-based score		ICI[b]-based score	
		Mean difference	Corrected *P* value	Mean difference	Corrected *P* value
TabPFN[c] with pretrained augmentation and non-pretrained augmentation performs similarly (n_0=50)	Pretrained augmentation performs better (n_0=50)	0.0024	.22	−0.0063	.97
LGBM[d] with pretrained augmentation and non-pretrained augmentation performs similarly (n_0=50)	Pretrained augmentation performs better (n_0=50)	0.0206[e]	.02[e]	−0.0279	.97
TabPFN with pretrained augmentation and non-pretrained augmentation performs similarly (n_0=350)	Pretrained augmentation performs better (n_0=350)	0.0003	.41	0.0018	.57
LGBM with pretrained augmentation and non-pretrained augmentation performs similarly (n_0=350)	Pretrained augmentation performs better (n_0=350)	0.0106[e]	.02[e]	−0.0057	.99

[a]AUC: area under the curve.

[b]ICI: integrated calibration index.

[c]TabPFN: Tabular Prior-Data Fitted Network.

[d]LGBM: light gradient boosting machine.

[e]Values significant at an α level of .05.

For TabPFN, pretrained augmentation did not lead to meaningful changes in discrimination at either dataset size (n_0=50: ΔAUC=0.0024, *P*=.22; n_0=350: ΔAUC=0.0003, *P*=.41).

In contrast, LGBM showed consistent and statistically significant improvements in AUC with pretrained augmentation (n_0=50: ΔAUC=0.0206, median AUC=0.67, (IQR:0.66-0.69), *P*=.02; n_0=350: ΔAUC=0.0106, median AUC=0.73, (IQR:0.71-0.75), *P*=.02), corresponding to gains of approximately 1 to 2 percentage points. These improvements were consistent in direction across 12 of the 13 datasets.

For calibration, pretrained augmentation was associated with small reductions in ICI for both models at n_0=50, but these effects were modest and not statistically significant after correction for multiple testing. At n_0=350, calibration differences between pretrained and non–pretrained augmentation were negligible.

Q2: Impact of Augmentation vs No Augmentation

We evaluated the effect of data augmentation compared to no augmentation across 13 datasets using paired permutation tests with Holm-Bonferroni correction (Table 4). Based on Q1, pretrained SDG methods were used for LGBM, while all augmentation strategies were retained for TabPFN.

XSL•FO
RenderX

Table . Pairwise permutation tests to determine the impact of augmentation against no augmentation. *P* values were adjusted for multiple comparisons using the Holm-Bonferroni procedure.

Null hypothesis	Alternative	AUC[a]-based score		ICI[b]-based score	
		Mean difference	Corrected *P* value	Mean difference	Corrected *P* value
TabPFN[c] with and without augmentation performs the same (n_0=50)	Augmented TabPFN performs better (n_0=50)	0.0094	.10	0.0134[d]	.04[d]
LLM[e]-augmented LGBM[f] performs the same as LGBM without augmentation (n_0=50)	LLM-augmented LGBM performs better (n_0=50)	0.0764[d]	<.001[d]	0.0191[d]	.003[d]
TabPFN with and without augmentation performs the same (n_0=350)	Augmented TabPFN performs better (n_0=350)	0.0011	.06	0.0041	.09
LLM-augmented LGBM performs the same as LGBM without augmentation (n_0=350)	LLM-augmented LGBM performs better (n_0=350)	0.0078[d]	.02[d]	0.0119	.09

[a]AUC: area under the curve.

[b]ICI: integrated calibration index.

[c]TabPFN: Tabular Prior-Data Fitted Network.

[d]Values significant at an α level of .05.

[e]LLM: large language model.

[f]LGBM: light gradient boosting machine.

For TabPFN, augmentation did not significantly affect discrimination at either dataset size (n_0=50: ΔAUC=0.0094, *P*=.10; n_0=350: ΔAUC=0.0011, *P*=.06), indicating minimal impact on AUC.

For LGBM, LLM-based augmentation resulted in statistically significant improvements in discrimination at both n_0=50 (ΔAUC=0.0764, *P*<.001) and n_0=350 (ΔAUC=0.0078, *P*=.02), with larger gains observed in the smaller datasets.

For calibration, augmentation significantly reduced ICI at n_0=50 for both TabPFN (ΔICI=0.0134, *P*=.04) and LGBM (ΔICI=0.0191, *P*=.003). At n_0=350, calibration differences between augmented and nonaugmented models were small and not statistically significant.

Q3: Comparison of Predictive Models

We compared model performance using the best configurations identified in Q1 and Q2 (Table 5).

XSL•FO
RenderX

Table . Pairwise permutation tests to determine the best classifier. *P* values were adjusted for multiple comparisons using the Holm-Bonferroni procedure.

Null hypothesis	Alternative	AUC[a]-based score		ICI[b]-based score	
		Mean difference	Corrected *P* value	Mean difference	Corrected *P* value
LLM[c]-augmented LGBM[d] and LLM classifier perform similarly (n_0=50)	LLM-augmented LGBM performs better (n_0=50)	0.0634[e]	.04[e]	0.0152	.31
Augmented TabPFN[f] and LLM classifier perform similarly (n_0=50)	TabPFN performs better (n_0=50)	0.0951[e]	.03[e]	0.0764[e]	.03[e]
Augmented TabPFN and LLM-augmented LGBM perform similarly (n_0=50)	TabPFN performs better (n_0=50)	0.0317[e]	.04[e]	0.0612[e]	<.001[e]
LLM-augmented LGBM and LLM classifier perform similarly (n_0=350)	LLM-augmented LGBM performs better (n_0=350)	0.0288[e]	.04[e]	0.0383	.06
Augmented TabPFN and LLM classifier perform similarly (n_0=350)	TabPFN performs better (n_0=350)	0.0441[e]	.01[e]	0.0484	.08
Augmented TabPFN and LLM-augmented LGBM perform similarly (n_0=350)	TabPFN performs better (n_0=350)	0.0153[e]	.004[e]	0.0101[e]	.02[e]

[a]AUC: area under the curve.

[b]ICI: integrated calibration index.

[c]LLM: large language model.

[d]LGBM: light gradient boosting machine.

[e]Values significant at an α level of .05.

[f]TabPFN: Tabular Prior-Data Fitted Network.

In terms of discrimination, both LLM-augmented LGBM and augmented TabPFN significantly outperformed LLM classifiers at both dataset sizes (all *P*<.05). Augmented TabPFN further achieved a significantly higher AUC than LLM-augmented LGBM (n_0=50: ΔAUC=0.0317, *P*=.04; n_0=350: ΔAUC=0.0153, *P*=.004), with a median AUC of 0.75 across datasets (IQR:0.74-0.77).

For calibration, augmented TabPFN showed significantly lower ICI than LGBM at both n_0=50 (ΔICI=0.0612, *P*<.001) and n_0=350 (ΔICI=0.0101, *P*=.02). Compared to LLM classifiers, TabPFN also demonstrated better calibration at n_0=50 (ΔICI=0.0764, *P*=.03), while differences at n_0=350 were smaller and not statistically significant (*P*=.08).

LLM classifiers are, therefore, reported in a fixed (no augmentation) configuration, reflecting their feasible operating regime under current computational constraints.

Post Hoc Analysis: Augmentation Strategies

To further investigate whether the observed benefits of augmentation on TabPFN were driven by increased sample size rather than the generation of synthetic data, we compared SDG methods with sampling with replacement.

We evaluated 2 comparisons: (1) LLM-based synthesis versus the best-performing augmentation method (SDG or LLM) and (2) sampling with replacement versus the best-performing augmentation method. The results are reported in Table 6.

XSL•FO
RenderX

Table . Post hoc 2-sided pairwise permutation test to compare augmenting Tabular Prior-Data Fitted Network (TabPFN) by sampling with replacement or by synthetic data generation (SDG) methods.

Null hypothesis and n_0	AUC[a]-based score		ICI[b]-based score	
	Statistic	*P* value	Statistic	*P* value
LLM[c]-based synthesis vs SDG methods and LLM-based synthesis				
50	−0.0024	.45	−0.0063	.41
350	−0.0003	.82	0.0018	.57
Sampling with replacement vs SDG methods and LLM-based synthesis				
50	−0.0058	.31	0.0016	.74
350	−0.0014	.41	−0.0013	.82

[a]AUC: area under the curve.

[b]ICI: integrated calibration index.

[c]LLM: large language model.

No statistically significant differences were observed for either AUC or ICI at both dataset sizes (all $P>.3$). Effect sizes were also negligible ($|\Delta AUC|<0.006$; $|\Delta ICI|<0.007$), reinforcing the absence of meaningful performance differences.

Ranking of Methods

The pairwise permutation tests supported the following performance ordering in terms of ICI and AUC: TabPFN-augmented with resampling>LLM-augmented LGBM>LLM classifiers. As a post hoc validation, we additionally applied the Page trend test, a nonparametric procedure designed to evaluate ordered alternatives. This analysis provided supporting evidence for a monotonic trend in classifier performance for both ICI and AUC (Table 7), consistent with the ranking obtained from the permutation tests. Together, these results provide convergent evidence that classifier performance follows the hypothesized order.

Table . Post hoc Page trend test (*L* value) results to validate rank order of classifiers' performance.[a,b]

n_0	AUC[c]-based score		ICI[d]-based score	
	L	Corrected *P* value	*L*	Corrected *P* value
50	167	0.02	166	.03
350	172	<.001	165	.048

[a]Null hypothesis: Augmented Tabular Prior-data Fitted Network (TabPFN) with sampling, large language model (LLM)-augmented light gradient boosting machine (LGBM), and LLM perform similarly.

[b]Alternative: Augmented TabPFN with sampling>LLM-augmented LGBM>LLM.

[c]AUC: area under the curve.

[d]ICI: integrated calibration index.

Computational Cost

Computational resource requirements constitute a critical point of differentiation between the approaches considered. LLMs typically require extensive optimization and fine-tuning, resulting in substantial computational overhead. In contrast, TabPFN operates in a zero-shot setting, thereby obviating the need for dataset-specific training and significantly reducing computational demands. This distinction is especially relevant in clinical research contexts, where access to high-performance computing resources may be limited. The comparison in Table 8 illustrates the marked relative differences in training and inference time across methods.

XSL•FO
RenderX

Table . Approximate training and inference time per dataset.

Model	Time needed to train one model on a dataset of 1000 samples and infer on 10,000 samples
TabPFN[a]	6 seconds on a GPU[b]
LLM[c] (average between the 3 considered LLMs)	12 hours on a GPU
LGBM[d]	20 minutes on CPUs[e]

[a]TabPFN: Tabular Prior-Data Fitted Network.
[b]GPU: graphics processing unit.
[c]LLM: large language model.
[d]LGBM: light gradient boosting machine.
[e]CPU: central processing unit.

Discussion

Summary

Health datasets available for research are often small, limiting the development of robust and generalizable clinical prediction models. Data augmentation is commonly used to mitigate this issue, including SDG methods such as CTGAN, Bayesian network, TVAE, and sequential trees [8,61,62]. However, these approaches can overfit when trained on limited data. An alternative is to leverage pretrained models, such as LLMs, to generate synthetic data without relying solely on the small dataset at hand.

In this study, we evaluated pretrained SDG using LLMs alongside traditional SDG approaches across 13 health datasets of sizes 50 and 350. This choice is consistent with current clinical prediction practices, where median training dataset sizes can be as low as 20 with a median dataset size of around 600 records [3,9]. We assessed their impact on 2 classifiers (LGBM and TabPFN), focusing on both discrimination and calibration. We also compared these approaches to using LLMs directly as classifiers.

Key Findings and Practical Interpretation

Augmentation Benefits Depend on the Model

Augmentation improved LGBM performance, particularly for very small datasets, with LLM-based augmentation yielding the largest gains. In contrast, TabPFN showed little to no improvement in discrimination from augmentation (Table 4). Additionally, pretrained augmentation provided measurable gains for LGBM but not for TabPFN (Table 3), suggesting that the pretraining of TabPFN already confers robustness in low-data settings.

For practitioners working with limited data, TabPFN can be used effectively with minimal augmentation effort, while LGBM may benefit from targeted augmentation when small performance gains are meaningful. In scenarios where large external test sets are unavailable, selecting the augmentation size through 10-fold cross-validation provides stable performance estimates (Figure C1 in Multimedia Appendix 1), supporting its use as a practical model selection strategy in small-data settings.

Simple Methods Can Be as Effective as Complex Ones

For TabPFN, neither SDG methods nor LLM-based augmentation outperformed simple sampling with replacement (Table 6). This indicates that the primary benefit of augmentation is increasing the effective sample size rather than generating novel synthetic patterns.

In resource-limited or regulated health care environments, simple resampling should be the default strategy. It is computationally trivial, fully transparent, and avoids the risks associated with black-box generative models.

LLMs Are Not Reliable as Standalone Classifiers in Low-Data Settings

Across all experiments, LLMs used directly as classifiers performed worse than both TabPFN and LGBM, particularly in calibration (Table 5), with TabPFN achieving the best overall performance. Ranking analysis further supported this ordering (Table 7). Poor calibration is especially problematic in clinical contexts where probability estimates inform decisions.

Deploying LLMs as predictive models in small clinical datasets is not advisable. Their use is better suited for augmentation or other supportive roles rather than direct prediction.

Calibration Improvements Are Limited but Important

Calibration improvements were observed primarily in the smallest datasets (Table 4), with no significant effects observed at larger sample sizes.

In small datasets, augmentation can still be valuable for improving the reliability of predicted probabilities, even when discrimination gains are modest.

Model Choice Matters More as Data Increases

At n_0=350, augmentation effects were smaller, and differences between models became more important (Table 4).

When moderate data is available, selecting an appropriate model (eg, TabPFN vs LGBM) is more impactful than investing in complex augmentation strategies.

Trustworthy and Resource-Constrained AI Considerations

A central finding of this study is that complex SDG does not outperform simple, transparent alternatives in small-data clinical

XSL•FO
RenderX

settings. The absence of measurable gains from SDG or LLM-based augmentation over resampling (Table 6) highlights that an increased sample size is the main driver of performance improvements. This has direct implications for trustworthy AI:

- Transparency: Sampling with replacement provides a clear and traceable data generation process, unlike LLMs or SDG methods.
- Safety: Avoiding black-box generation reduces the risk of introducing unrealistic or misleading synthetic patient data.
- Regulatory alignment: Simpler methods are easier to justify and validate in clinical and regulatory contexts.
- Efficiency: TabPFN requires seconds to run, compared to hours for LLMs, making it accessible in low-resource environments.

Similarly, TabPFN combines strong predictive performance (Table 5) with minimal computational requirements (Table 8), making it particularly suitable for environments with limited resources.

Although LLM classifiers were not evaluated under augmentation due to practical computational constraints, this reflects a realistic deployment limitation rather than an experimental omission. Unlike TabPFN and LGBM, LLM fine-tuning with varying synthetic sample sizes is computationally prohibitive at scale, requiring hundreds of graphics processing unit (GPU)–days (see Section D in Multimedia Appendix 1). As such, the comparison reflects practical usability under resource-constrained conditions rather than fully symmetric experimental tuning.

Fidelity and Role of Synthetic Data

The post hoc analysis suggests that improvements from augmentation are primarily driven by increased sample size rather than the generation of novel synthetic data (Table 6).

While we performed basic realism checks to ensure that synthetic records respect plausible clinical ranges and preserve key distributions, perfect fidelity to the original data is not required for predictive tasks. Prior work [63] has shown that out-of-population observations can be common in synthetic datasets without necessarily degrading predictive performance. Similarly, generative models may improve generalization by increasing data diversity [8].

Taken together, these findings indicate that the benefits of augmentation are more closely related to increased sample size and diversity than to the exact replication of the original data distribution.

Limitations and Future Work

This study focused on binary classification tasks and may not generalize to other settings, such as regression or survival analysis. Although we evaluated multiple datasets, further validation across diverse clinical contexts is needed.

While we evaluated 13 heterogeneous health datasets, they may not represent the full diversity of clinical data environments. Datasets with high dimensionality, extreme class imbalance, or strong temporal structure may behave differently under augmentation strategies. External validation on additional real-world datasets is required to strengthen generalizability.

Our experimental design relied on relatively large hold-out test sets to ensure stable estimation of discrimination and calibration metrics. In many real-world health care settings, such large external test sets are unavailable. Although our cross-validation analysis suggests that model selection through 10-fold cross-validation provides comparable estimates, performance variability may be higher in practice, particularly in ultra-small datasets.

Some of the non-pretrained models that were used in our analysis, such as CTGAN and TVAE, used the default hyperparameters and did not undergo additional tuning. While this may have limited their performance, sensitivity analyses presented in Section B in Multimedia Appendix 1 revealed that attempting to tune these hyperparameters on very small datasets can lead to overfitting. Specifically, we observed that the multivariate Hellinger distance (which is a common synthetic data fidelity metric) between the data from the generative models and sampling with replacement was lower when tuning compared to using default parameters. The sampling with replacement dataset serves as a baseline, indicating a high rate of overfitting. The fact that the fidelity was higher to the overfitted data with hyperparameter tuning indicates that tuning results in datasets that are closer to resampled data, which is highly overfitted. This highlights a practical limitation: in low-data settings, more complex model tuning may be counterproductive, and default configurations—or simpler augmentation strategies—can yield more reliable and stable results.

To adapt pretrained LLMs for tabular data, we used a serialization method from previous studies [33,35] that places column names before values, thereby preserving contextual information. While this creates artificial sentences, fine-tuning is thought to mitigate this issue. We did not explore alternative serialization methods, which could be an area for future research. Future work should explore broader classes of models, alternative data modalities, and evaluation in real-world deployment settings.

Acknowledgments

Generative artificial intelligence was not used in the writing of this paper.

XSL•FO
RenderX

Funding

This research is funded by the Canada Research Chairs program through the Canadian Institutes of Health Research, a Discovery Grant RGPIN-2022-04811 from the Natural Sciences and Engineering Research Council of Canada, and the Canadian Children's Inflammatory Bowel Disease Network. LP is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): 530282197.

Data Availability

The following provides information on the availability of each of the datasets used in this study:

1. Better Outcomes Registry & Network (BORN) |: The BORN collects Ontario's prescribed perinatal, newborn, and child registry with the role of facilitating quality care for families across the province [64].

2. Basic Stand Alone (BSA): The BSA inpatient claims dataset is about claim-level information, where each record is an inpatient claim incurred by a 5% sample of Medicare beneficiaries. [65]

3. California State Hospital Discharge: The California dataset contains the patient hospital discharge data from 2008, sourced from the California State Inpatient Databases (SID), Healthcare Cost and Utilization Project (HCUP), and the Agency for Healthcare Research and Quality [66].

4. Canadian Community Health Survey (CCHS): The CCHS data consist of Canadian population-level information concerning health status, health system utilization, and health determinants, collected by Statistics Canada through telephone surveys [67].

5. COVID-19: The COVID-19 dataset collects Canadian health records related to COVID-19, gathered by the Public Health Agency of Canada, and is available on Esri Canada's platform [68].

6. FDA Adverse Event Reporting System (FAERS): The FAERS is a database comprising information on adverse events and medication error reports submitted to the FDA and can be downloaded [69].

7. Florida State Hospital Discharge: The Florida dataset contains the patient's hospital 2007 discharge data from the Florida SID, HCUP, Agency for Healthcare Research and Quality [66], and is available for purchase.

8. Medical Information Mart for Intensive Care III (MIMIC-III): MIMIC-III is a large database that contains deidentified health-related data associated with more than 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012 [70,71]. Access to the MIMIC database is granted upon signing a data use agreement with PhysioNet [70-72].

9. New York State Hospital Discharge: The New York dataset contains the patient hospital discharge data from 2007, sourced from the New York SID, HCUP, and the Agency for Healthcare Research and Quality [66]

10. COVID-19 Survival (Nexoid): The COVID-19 survival dataset is a web-based survey dataset collected by a company called Nexoid in the United Kingdom. It is publicly available [73].

11. Texas Hospital Discharge: The Texas dataset contains the patient hospital discharge information for the first quarter of 2012 from Texas in the United States [74] and is publicly available.

12. Washington State Hospital Discharge 2007: The Washington dataset contains the patient hospital discharge data from 2007, sourced from the Washington SID, HCUP, and the Agency for Healthcare Research and Quality [66]. It is available for purchase.

13. Washington State Hospital Discharge 2008: The Washington 2008 dataset contains the patient hospital discharge data from 2008, sourced from the Washington SID, HCUP, and the Agency for Healthcare Research and Quality [66]. It is available for purchase.

Conflicts of Interest

LP has financial interests in Woodway Assurance, a privacy technology spin-off company from her research lab at the University of Ottawa, but it is not related to the topic of this study. KEE is co–editor-in-chief of *JMIR AI*. FKD is an associate editor of *JMIR AI*.

Multimedia Appendix 1
Additional methodological details and results.
[PDF File, 781 KB - jmir_v28i1e88678_app1.pdf]

References

XSL•FO
RenderX

1. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 2014 Dec 22;14:137. [doi: 10.1186/1471-2288-14-137] [Medline: 25532820]
2. Riley RD, Collins GS. Stability of clinical prediction models developed using statistical or machine learning methods. Biom J 2023 Dec;65(8):e2200302. [doi: 10.1002/bimj.202200302] [Medline: 37466257]
3. Andaur Navarro CL, Damen JAA, van Smeden M, et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. J Clin Epidemiol 2023 Feb;154:8-22. [doi: 10.1016/j.jclinepi.2022.11.015] [Medline: 36436815]
4. Riley RD, Ensor J, Snell KIE, et al. Importance of sample size on the quality and utility of AI-based prediction models for healthcare. Lancet Digit Health 2025 Jun;7(6):100857. [doi: 10.1016/j.landig.2025.01.013] [Medline: 40461350]
5. Pongsuwun K, Puwarawuttipanit W, Nguantad S, et al. A systematic review of the accuracy of machine learning models for diagnosing pulmonary tuberculosis: implications for nursing practice and implementation. Nurs Health Sci 2025 Mar;27(1):e70077. [doi: 10.1111/nhs.70077] [Medline: 40058367]
6. Tsegaye B, Snell KIE, Archer L, et al. Larger sample sizes are needed when developing a clinical prediction model using machine learning in oncology: methodological systematic review. J Clin Epidemiol 2025 Apr;180:111675. [doi: 10.1016/j.jclinepi.2025.111675] [Medline: 39814217]
7. Mitsakakis N, Liu D, Walters T, El Emam K. Sample size calculation for training ensemble machine learning models on health data. Patterns 2026 Mar:101498. [doi: 10.1016/j.patter.2026.101498]
8. Liu D, Kababji SE, Mitsakakis N, et al. Augmenting small tabular health data for training prognostic ensemble machine learning models using generative models. BMC Med Inform Decis Mak 2025 Nov 28;25(1):435. [doi: 10.1186/s12911-025-03266-3] [Medline: 41316099]
9. Dhiman P, Ma J, Andaur Navarro CL, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. BMC Med Res Methodol 2022 Apr 8;22(1):101. [doi: 10.1186/s12874-022-01577-x] [Medline: 35395724]
10. Mumuni A, Mumuni F. Data augmentation: a comprehensive survey of modern approaches. Array 2022 Dec;16:100258. [doi: 10.1016/j.array.2022.100258]
11. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data 2019 Dec;6(1):60. [doi: 10.1186/s40537-019-0197-0]
12. Goceri E. Medical image data augmentation: techniques, comparisons and interpretations. Artif Intell Rev 2023 Mar 20;56:12561-12605. [doi: 10.1007/s10462-023-10453-z] [Medline: 37362888]
13. Naveed H, Anwar S, Hayat M, Javed K, Mian A. Survey: image mixing and deleting for data augmentation. Eng Appl Artif Intell 2024 May;131:107791. [doi: 10.1016/j.engappai.2023.107791]
14. Feng S, Gangal V, Wei J, et al. A survey of data augmentation approaches for NLP. Presented at: Findings of the Association for Computational Linguistics; Aug 1-6, 2021. [doi: 10.18653/v1/2021.findings-acl.84]
15. Chlap P, Min H, Vandenberg N, Dowling J, Holloway L, Haworth A. A review of medical image data augmentation techniques for deep learning applications. J Med Imaging Radiat Oncol 2021 Aug;65(5):545-563. [doi: 10.1111/1754-9485.13261] [Medline: 34145766]
16. Duong HT, Nguyen-Thi TA. A review: preprocessing techniques and data augmentation for sentiment analysis. Comput Soc Netw 2021 Dec;8(1):1. [doi: 10.1186/s40649-020-00080-x]
17. Felix EA, Lee SP. Systematic literature review of preprocessing techniques for imbalanced data. IET Softw 2019 Dec;13(6):479-496. [doi: 10.1049/iet-sen.2018.5193]
18. Wen Q, Sun L, Yang F, et al. Time series data augmentation for deep learning: a survey. Presented at: Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21); Aug 19-26, 2021. [doi: 10.24963/ijcai.2021/631]
19. Iwana BK, Uchida S. An empirical survey of data augmentation for time series classification with neural networks. PLOS ONE 2021;16(7):e0254841. [doi: 10.1371/journal.pone.0254841] [Medline: 34264999]
20. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling tabular data using conditional GAN. Presented at: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); Dec 8-14, 2019 URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf [accessed 2026-05-21]
21. Kiran A, Rubini P, Kumar SS. Challenges and limitations of TVAE tabular synthetic data generator. In: Advanced Computing: Springer; 2025:243-254. [doi: 10.1007/978-3-031-84602-1_17]
22. Kaur D, Sobiesk M, Patil S, et al. Application of Bayesian networks to generate synthetic health data. J Am Med Inform Assoc 2021 Mar 18;28(4):801-811. [doi: 10.1093/jamia/ocaa303] [Medline: 33367620]
23. Gogoshin G, Branciamore S, Rodin AS. Synthetic data generation with probabilistic Bayesian networks. Math Biosci Eng 2021 Oct 9;18(6):8603-8621. [doi: 10.3934/mbe.2021426] [Medline: 34814315]
24. Martins LNA, Gonçalves FB, Galletti TP. Generation and analysis of synthetic data via Bayesian networks: a robust approach for uncertainty quantification via Bayesian paradigm. arXiv. Preprint posted online on Feb 27, 2024. [doi: 10.48550/arXiv.2402.17915]

XSL•FO
RenderX

25. Deeva I, Andriushchenko PD, Kalyuzhnaya AV, Boukhanovsky AV. Bayesian networks-based personal data synthesis. Presented at: Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good; Sep 14-16, 2020. [doi: 10.1145/3411170.3411243]
26. Emam KE, Mosquera L, Zheng C. Optimizing the synthesis of clinical trial data using sequential trees. J Am Med Inform Assoc 2021 Jan 15;28(1):3-13. [doi: 10.1093/jamia/ocaa249] [Medline: 33186440]
27. Drechsler J, Reiter JP. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. Comput Stat Data Anal 2011 Dec;55(12):3232-3243. [doi: 10.1016/j.csda.2011.06.006]
28. Nowok B. Utility of synthetic microdata generated using tree-based methods. Presented at: Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality (Nowok, 2015); Oct 5-7, 2015 URL: https://unece.org/sites/default/files/datastore/fileadmin/DAM/stats/documents/ece/ces/ge.46/20150/Paper_33_Session_2_-_Univ._Edinburgh__Nowok_.pdf [accessed 2026-05-21]
29. Reiter J. Using CART to generate partially synthetic, public use microdata. J Off Stat 2005;21(3):441-462 [FREE Full text]
30. Barr AA, Quan J, Guo E, Sezgin E. Large language models generating synthetic clinical datasets: a feasibility and comparative analysis with real-world perioperative data. Front Artif Intell 2025;8:1533508. [doi: 10.3389/frai.2025.1533508] [Medline: 39974356]
31. Fang X, Xu W, Tan FA, et al. Large language models (LLMs) on tabular data: prediction, generation, and understanding—a survey. arXiv. Preprint posted online on Feb 27, 2024. [doi: 10.48550/arXiv.2402.17944]
32. Borisov V, Seßler K, Leemann T, Pawelczyk M, Kasneci G. Language models are realistic tabular data generators. arXiv. Preprint posted online on Oct 12, 2022. [doi: 10.48550/arXiv.2210.06280]
33. Seedat N, Huynh N, van BB, et al. Curated LLM: synergy of LLMs and data curation for tabular augmentation in low-data regimes. Presented at: International Conference on Machine Learning (ICML 2024); Jul 21-27, 2024. [doi: 10.5555/3692070.3693865]
34. Isomura T, Shimizu R, Goto M. LLMOverTab: tabular data augmentation with language model-driven oversampling. Expert Syst Appl 2025 Mar;264:125852. [doi: 10.1016/j.eswa.2024.125852]
35. Nguyen D, Gupta S, Do K, Nguyen T, Venkatesh S. Generating realistic tabular data with large language models. Presented at: 2024 IEEE International Conference on Data Mining (ICDM); Dec 9-12, 2024. [doi: 10.1109/ICDM59182.2024.00040]
36. Brown KE, Yan C, Li Z, et al. Large language models are less effective at clinical prediction tasks than locally trained machine learning models. J Am Med Inform Assoc 2025 May 1;32(5):811-822. [doi: 10.1093/jamia/ocaf038] [Medline: 40056436]
37. Hollmann N, Müller S, Purucker L, et al. Accurate predictions on small data with a tabular foundation model. Nature 2025 Jan;637(8045):319-326. [doi: 10.1038/s41586-024-08328-6] [Medline: 39780007]
38. Xu L. Synthesizing tabular data using conditional GAN [Master's thesis]. : Massachusetts Institute of Technology; 2020 URL: https://dspace.mit.edu/entities/publication/79844ce9-4b05-4be9-acdc-568c9483c51d [accessed 2026-05-21]
39. Habibi O, Chemmakha M, Lazaar M. Imbalanced tabular data modelization using CTGAN and machine learning to improve IoT Botnet attacks detection. Eng Appl Artif Intell 2023 Feb;118:105669. [doi: 10.1016/j.engappai.2022.105669]
40. Bourou S, El Saer A, Velivassaki TH, Voulkidis A, Zahariadis T. A review of tabular data synthesis using GANs on an IDS dataset. Information 2021;12(9):375. [doi: 10.3390/info12090375]
41. Pathare A, Mangrulkar R, Suvarna K, Parekh A, Thakur G, Gawade A. Comparison of tabular synthetic data generation techniques using propensity and cluster log metric. Int J Inf Manag Data Insights 2023 Nov;3(2):100177. [doi: 10.1016/j.jjimei.2023.100177]
42. Farhadyar K, Bonofiglio F, Zoeller D, Binder H. Adapting deep generative approaches for getting synthetic data with realistic marginal distributions. arXiv. Preprint posted online on May 14, 2021. [doi: 10.48550/arXiv.2105.06907]
43. Qian Z, Cebere BC, van der Schaar M. Synthcity: facilitating innovative use cases of synthetic data in different data modalities. arXiv. Preprint posted online on Jan 18, 2023. [doi: 10.48550/arXiv.2301.07573]
44. CHEO-EHIL/pysdg-releases. GitHub. URL: https://github.com/CHEO-EHIL/pysdg-releases [accessed 2026-06-01]
45. Hu EJ, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models. arXiv. Preprint posted online on Jun 17, 2021. [doi: 10.48550/arXiv.2106.09685]
46. Zhang K, Zeng S, Hua E, et al. UltraMedical: building specialized generalists in biomedicine. arXiv. Preprint posted online on Jun 6, 2024. [doi: 10.48550/arXiv.2406.03949]
47. Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); Dec 4-9, 2017 URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf [accessed 2026-05-21]
48. Rousset A, Dellamonica D, Menuet R, et al. Can machine learning bring cardiovascular risk assessment to the next level? A methodological study using FOURIER trial data. Eur Heart J Digit Health 2021 Mar;3(1):38-48. [doi: 10.1093/ehjdh/ztab093] [Medline: 36713994]
49. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLOS ONE 2017;12(4):e0174944. [doi: 10.1371/journal.pone.0174944] [Medline: 28376093]

XSL•FO
RenderX

50. Akyea RK, Qureshi N, Kai J, Weng SF. Performance and clinical utility of supervised machine-learning approaches in detecting familial hypercholesterolaemia in primary care. NPJ Digit Med 2020;3:142. [doi: 10.1038/s41746-020-00349-5] [Medline: 33145438]
51. Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. JAMA Netw Open 2020 Jan 3;3(1):e1918962. [doi: 10.1001/jamanetworkopen.2019.18962] [Medline: 31922560]
52. Li YM, Jiang LC, He JJ, Jia KY, Peng Y, Chen M. Machine learning to predict the 1-year mortality rate after acute anterior myocardial infarction in Chinese patients. Ther Clin Risk Manag 2020;16:1-6. [doi: 10.2147/TCRM.S236498]
53. Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. Inf Fusion 2022 May;81:84-90. [doi: 10.1016/j.inffus.2021.11.011]
54. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? Presented at: 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks; Nov 28 to Dec 9, 2022. [doi: 10.52202/068431-0037]
55. Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. Presented at: Advances in Neural Information Processing Systems 25 (NIPS 2012); Dec 3-8, 2012 URL: https://proceedings.neurips.cc/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html [accessed 2026-05-21]
56. Bartz E, Bartz-Beielstein T, Zaefferer M, Mersmann O. Hyperparameter Tuning for Machine and Deep Learning with R: A Practical Guide: Springer; 2023. [doi: 10.1007/978-981-19-5170-1]
57. Bischl B, Binder M, Lang M, et al. Hyperparameter optimization: foundations, algorithms, best practices, and open challenges. WIREs Data Min Knowl Discov 2023;13:e1484. [doi: 10.1002/widm.1484]
58. Binder M, Pfisterer F, Bischl B. Collecting empirical data about hyperparameters for data driven AutoML. Presented at: 7th ICML Workshop on Automated Machine Learning; Jul 17-18, 2020 URL: https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_63.pdf [accessed 2026-05-21]
59. Kühn D, Probst P, Thomas J, Bischl B. Automatic exploration of machine learning experiments on OpenML. arXiv. Preprint posted online on Jun 28, 2018. [doi: 10.48550/arXiv.1806.10961]
60. Johnson JM, Khoshgoftaar TM. Medical provider embeddings for healthcare fraud detection. SN Comput Sci 2021 Jul;2(4):276. [doi: 10.1007/s42979-021-00656-y]
61. Wang W, Pai TW. Enhancing small tabular clinical trial dataset through hybrid data augmentation: combining SMOTE and WCGAN-GP. Data (Basel) 2023;8(9):135. [doi: 10.3390/data8090135]
62. Papadopoulos D, Karalis VD. Variational autoencoders for data augmentation in clinical studies. Appl Sci (Basel) 2023;13(15):8793. [doi: 10.3390/app13158793]
63. Pilgram L, El Kababji S, Liu D, El Emam K. Magnitude and impact of hallucinations in tabular synthetic health data on prognostic machine learning models: validation study. J Med Internet Res 2025 Aug 18;27:e77893. [doi: 10.2196/77893] [Medline: 40825542]
64. Data. BORN Ontario. URL: https://www.bornontario.ca/data/ [accessed 2026-06-08]
65. BSA inpatient claims PUF. CMS.gov. URL: https://www.cms.gov/data-research/statistics-trends-and-reports/basic-stand-alone-medicare-claims-public-use-files/bsa-inpatient-claims-puf [accessed 2026-06-08]
66. Healthcare cost and utilization project (HCUP). Agency for Healthcare Research and Quality. 2005. URL: https://www.ahrq.gov/data/hcup/index.html [accessed 2026-05-21]
67. How to access Canadian community health survey (CCHS) data. Statistics Canada. URL: https://www150.statcan.gc.ca/n1/pub/82-620-m/2005001/4144189-eng.htm [accessed 2026-06-08]
68. COVID-19 resources Canada. ArcGIS Hub. URL: https://resources-covid19canada.hub.arcgis.com [accessed 2026-06-08]
69. FDA adverse event monitoring system (AEMS) latest quarterly data files. US Food And Drug Administration. URL: https://www.fda.gov/drugs/fda-adverse-event-monitoring-system-aems/fda-adverse-event-monitoring-system-aems-latest-quarterly-data-files [accessed 2026-06-08]
70. Johnson A, Pollard T, Mark R. MIMIC-III clinical database. PhysioNet. 2016. URL: https://physionet.org/content/mimiciii/1.4/ [accessed 2026-05-21]
71. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035. [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
72. Goldberger AL, Amaral LA, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000 Jun 13;101(23):E215-E220. [doi: 10.1161/01.cir.101.23.e215] [Medline: 10851218]
73. Download dataset. COVID-19. URL: https://www.covid19survivalcalculator.com/en/download [accessed 2026-06-08]
74. Texas hospital inpatient discharge public use data file. : Texas Department of State Health Services (DSHS); 2025 URL: https://www.dshs.texas.gov/sites/default/files/thcic/hospitals/inpatientdatadictionary1q2025.pdf [accessed 2026-05-21]

Abbreviations

AUC: area under the curve

XSL•FO
RenderX

CTGAN: conditional tabular generative adversarial network
GPU: graphics processing unit
ICI: integrated calibration index
LGBM: light gradient boosting machine
LLM: large language model
LoRA: low-rank adaptation
SDG: synthetic data generation
TabPFN: Tabular Prior-Data Fitted Network
TVAE: tabular variational autoencoder

Edited by J Sarvestan; submitted 29.Nov.2025; peer-reviewed by E Joodi, Y Wang; revised version received 14.May.2026; accepted 14.May.2026; published 15.Jun.2026.

Please cite as:
Huet-Dastarac M, Dankar FK, Liu D, El Kababji S, Pilgram L, El Emam K
An Evaluation of Pretrained Generative Models for Augmenting Small Health Data: Comparative Modeling Study
J Med Internet Res 2026;28:e88678
URL: https://www.jmir.org/2026/1/e88678
doi:10.2196/88678

© Margerie Huet-Dastarac, Fida Dankar, Dan Liu, Samer El Kababji, Lisa Pilgram, Khaled El Emam. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Alignment Between Cardiologists and AI-Driven Diagnostic Systems: Mixed Methods Study

Mahdi Mahdavi[1,2], MSc, PhD; Sarah White[3], PhD; Sandeep S Hothi[4,5], PhD; Chris Flood[6,7], PhD; Rosica Panayotova[8], PhD; Daniel Frings[9], PhD

[1]Department of Management and Enterprise, Faculty of Business and Law, University of Roehampton, London, United Kingdom
[2]Bernard Lown Scholars in Cardiovascular Health Programme, Harvard T.H. Chan School of Public Health, Boston, MA, United States
[3]School of Health and Medical Sciences, City St George's, University of London, London, United Kingdom
[4]Department of Cardiology, Royal Wolverhampton NHS Trust, Wolverhampton, United Kingdom
[5]College of Medical and Dental Sciences, Institute of Cardiovascular Sciences, University of Birmingham, Birmingham, United Kingdom
[6]School of Nursing and Midwifery, London South Bank University, London, United Kingdom
[7]Central London Community Healthcare NHS Trust, London, United Kingdom
[8]Stockport NHS Foundation Trust, Stockport, United Kingdom
[9]College of Health and Life Sciences, London South Bank University, 103 Borough Road, London, United Kingdom

Corresponding Author:
Daniel Frings, PhD
College of Health and Life Sciences, London South Bank University, 103 Borough Road, London, United Kingdom

Abstract

Background: The clinical value of artificial intelligence (AI)–based diagnostic systems depends not only on their accuracy but also on how well their outputs integrate with clinicians' judgments in practice. Critical knowledge gaps remain regarding diagnostic concordance between AI and clinicians in stress echocardiography interpretation, patient characteristics predicting discordance, and how cardiologists respond when AI recommendations conflict with their clinical diagnoses.

Objective: This study examined the diagnostic alignment between an AI-driven stress echocardiography system (EchoGo Pro [EGP]) and cardiologists' diagnoses of coronary artery disease (CAD), identified predictors of concordance and AI scan rejection, and explored cardiologists' decision-making strategies when disagreements arise.

Methods: We conducted mixed methods research. The quantitative study analyzed concordance between EGP and cardiologists using data from 854 participants with suspected CAD in the multicenter PROTEUS randomized controlled trial. Logistic regression identified predictors of agreement, disagreement, and scan rejection, adjusting for age, sex, smoking status, BMI, and cardiovascular risk factors (hypertension, hypercholesterolemia, diabetes, family history of CAD, and prior CAD events). To gain deeper insight into discordance, we conducted a qualitative study analyzing survey responses from 61 UK consultant cardiologists recruited via Qualtrics, exploring their perceptions of AI tools, the risks of following discordant AI recommendations, and their typical responses to AI-clinician disagreement.

Results: EGP and cardiologists agreed in 60% (512/854) of the cases, but agreement was significantly lower among patients with hypertension (OR 0.58, 95% CI 0.38 - 0.89; *P*=.01), diabetes (OR 0.56, 95% CI 0.35 - 0.90; *P*=.02), and pre-existing CAD (OR 0.48, 95% CI 0.30 - 0.77; *P*=.002). EGP rejected 26.1% (222/854) of the scans due to insufficient image quality, with rejection significantly more common in male patients (β=0.35; *P*=.03) and those with a family history of CAD. If a positive CAD diagnosis was assigned when either cardiologists or EGP identified CAD, the proportion of positive cases increased from 17.9% (153/854) to 22.1% (189/854), potentially identifying additional at-risk patients. Survey respondents (50/60, 85% male; 26/57, 46% aged 40-49 years; 39/61, 64% White) required 65% to 69% confidence in their initial diagnosis to justify disregarding contradictory AI recommendations. The survey findings revealed cardiologists treated AI recommendations as advisory rather than definitive. When facing discordance, they retained confidence in their judgment and sought corroboration through additional testing, data review, or second opinions rather than deferring to AI. Paradoxically, cardiologists with higher confidence in AI tools required greater confidence in their own diagnosis to disregard AI recommendations (β=7.73; *P*=.02). Cardiologists attributed discordance primarily to AI's inability to incorporate patient history, comorbidities, and broader clinical context.

Conclusions: EGP shows promise as an adjunctive tool but struggles with multimorbid patients and exhibits high, uneven rejection rates. Cardiologists use AI to prompt scrutiny, not replace judgment. Future systems need to integrate wider patient data with imaging and minimize bias through representative training to avoid exacerbating inequities.

Trial Registration: ClinicalTrials.gov NCT05028179; https://clinicaltrials.gov/study/NCT05028179

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2023-079617

XSL•FO
RenderX

(J Med Internet Res 2026;28:e83541) doi:10.2196/83541

KEYWORDS

cardiac imaging; echocardiography; stress; clinical reasoning; artificial intelligence; cardiology

Introduction

The alignment between artificial intelligence (AI) assessments and human interpretations of health conditions influences clinicians' confidence in diagnostic accuracy and the perceived use of AI systems in practice [1-3]. AI-driven tools have shown considerable promise in image processing in health care [4,5]. However, the clinical value of these tools depends not only on their accuracy against a reference standard but also on how their outputs integrate with clinicians' judgments [6].

Studies on AI applied to health care suggest that clinicians are generally optimistic about AI but remain cautious, emphasizing transparency, bias, and responsibility, and often describe AI as an advisory rather than authoritative tool [6,7]. When AI and clinician judgments converge, this alignment can support timely and cost-effective care by reinforcing decision-makers' confidence and potentially reducing the need for further testing [8]. However, when discordance arises, the extent to which clinicians rely on AI versus their own judgment and the factors driving such discrepancies are less well understood [7]. Therefore, understanding how AI outputs are interpreted in cases of discordance, the risks associated with following AI recommendations, and the conditions under which clinicians rely on AI rather than their own judgment is essential to fully harness the benefits of AI systems.

Stress echocardiography (SE) is a widely performed clinical investigation for the identification of coronary artery disease (CAD) [9]. It generates images of left ventricular wall motion acquired before and during exercise and/or pharmacological stress to make a clinical judgment on whether CAD is present, and which areas of the heart are affected. However, due to a degree of subjectivity in the test's interpretation, clinical judgments may vary between clinicians [10]. Interobserver variability can lead to inconsistent classification of ischemia and variation in the patient journey, including referral for invasive angiography [11,12]. Additionally, SE is a resource-intensive test, requiring trained specialists to both perform and interpret the test. As a result, the demand for SE imaging is outpacing the capacity of cardiac specialists in many countries, particularly those with aging populations including the United Kingdom [13].

In SE tests, AI-based models have demonstrated high diagnostic accuracy for the detection and risk stratification of CAD, often achieving performance comparable to expert readers [6,10,14]. Their consistent and automated performance can enhance the accuracy of CAD diagnosis, while also reducing reporting time, providing a cost-effective second opinion, and minimizing interobserver variability [15,16].

EchoGo Pro (EGP), an AI-driven SE analysis system, has been developed for cardiac assessments and identifying potential ischemia [17]. EGP automatically analyzes SE images to estimate the probability of CAD and produces a structured binary risk report intended to support clinician decision-making. Initial validation studies have reported encouraging diagnostic accuracy and suggested that incorporating EGP into SE workflows may improve the detection of severe CAD compared with conventional interpretation alone [17].

Although systematic reviews consistently show that AI systems can achieve diagnostic performance comparable to health care professionals [3,6,18], most studies focus on accuracy metrics rather than on how clinicians respond in practice. Emerging evidence indicates that clinicians rarely respond to discordance in a simple accept-reject manner. Instead, they may ignore, selectively incorporate, negotiate, or conditionally defer to AI recommendations depending on their initial level of confidence in their own judgment, the perceived severity of the patient's condition, and how clearly and convincingly the AI system explains its recommendation [19-21].

Yet, existing evidence is limited in scope and often based on noncardiology contexts, with little known about clinician behavior when AI outputs conflict with their own judgment, particularly in time-sensitive settings such as cardiac care [22]. Evidence is also scarce on which patient characteristics are associated with AI-clinician discordance, which factors predict AI rejection of scans, how cardiologists weigh AI outputs against their own judgment, and how perceived risk influences their willingness to follow AI recommendations [23,24]. Understanding how AI-clinician interaction unfolds in cardiac care, where rapid decisions can be lifesaving, is therefore critical for safe and effective implementation [24].

This mixed methods study integrates quantitative and qualitative analyses. The quantitative study uses clinical trial data to assess concordance between EGP and clinicians, examine patient characteristics associated with concordant and discordant cases, and identify demographic and clinical profiles linked to AI scan rejection or diagnostic divergence. The qualitative component explores cardiologists' broader perceptions and experiences with AI tools, providing insight into attitudes, barriers, and contextual factors influencing adoption and trust. Together, these findings inform future development and implementation strategies for AI in cardiology. The analysis plan was preregistered on the Open Science Framework.

Methods

Overview of Research Design

We conducted a mixed methods study with 2 components, following a prespecified protocol published earlier [25]. The first was a secondary analysis of diagnostic agreement between EGP and cardiologists' assessments of CAD in SE images. This analysis used data from 854 participants in the PROTEUS trial (for a broader evaluation of EGP, see below) [26]. The second component comprised a qualitative survey of consultant cardiologists, featuring open-ended questions designed to explore their perceptions and responses to disagreements with

XSL•FO
RenderX

AI tools. Notably, this component considered AI tools in general rather than focusing specifically on EGP. Reporting of the agreement analysis adhered to the Guidelines for Reporting Reliability and Agreement Studies [27]. The survey component followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist for cross-sectional studies [28].

The Experimental Study

The PROTEUS trial was a randomized, multicenter, 2-armed study aimed at participants referred to SE clinics for the investigation of suspected CAD. The details of the PROTEUS research design are published by Woodward et al [26]. The trial randomized participants into a control group (SE assessment without EGP report) or an intervention group (SE assessment with EGP report). In the intervention group, SE images were analyzed by EGP, and a binary report produced by EGP indicating the probability of severe CAD was provided to the clinician alongside their own judgment [26]. Images failing quality criteria were flagged with a rejection report.

Cases were categorized as either concordant or discordant based on the alignment between EGP and clinician interpretation. Clinician interpretation was classified as positive for CAD if a stress-induced wall motion abnormality was observed, negative for CAD if no wall motion abnormality was induced, or inconclusive if the interpretation was abandoned or deemed indeterminate for inducible ischemia. The EGP risk rating classified cases as high risk of CAD (coronary stenosis ≥70%), low risk (coronary stenosis ≤70%), or uninterpretable if the image quality was insufficient for assessment.

Concordance was defined as agreement between the clinician and the EGP in either of the following scenarios: (1) a positive clinician interpretation and a high EGP risk rating or (2) a negative clinician interpretation and a low EGP risk rating. Discordance was defined as (1) a positive clinician interpretation paired with a low EGP risk rating or (2) a negative clinician interpretation paired with a high EGP risk rating.

The control variables consisted of age, sex, smoking status, BMI, and cardiovascular risk factors (CRFs). The data were collected through a combination of self-report questionnaires, physical measurements, and laboratory assessments. Age was categorized into the following groups: 18 to 39, 40 to 59, 60 to 80, and ≥81 years of age. Smoking status was self-reported and comprised current smokers, ex-smokers, or nonsmokers. BMI was calculated from measured height and weight, which were obtained during clinical examination. CRFs included hypertension, hypercholesterolemia, diabetes mellitus, positive family history of CAD, and history of CAD (including prior clinical events). Family history of CAD was self-reported, while other CRFs were determined through a combination of medical history review, physical measurements, and laboratory results. Specifically, hypertension was defined as measured blood pressure greater than or equal to 140/90 mm Hg or current use of antihypertensive medications [29]. Diabetes mellitus was diagnosed based on laboratory values (hemoglobin A_{1c} ≥6.5% [48 mmol/mol] or fasting blood sugar >126 mg/dL) or current use of antidiabetic medications [30]. Hypercholesterolemia and history of CAD were ascertained from medical records and confirmed through patient interview.

Descriptive statistics were used for demographics and clinical characteristics. Means and SDs were reported for normally distributed continuous variables, and frequencies and percentages were reported for categorical variables. We excluded cases where EGP reports were not used due to arriving late or being rejected (n=222, 26%). Logistic regression was used to identify the predictors of concordance or discordance and factors associated with AI image rejection. We used R for statistical analysis.

The Survey

We collected survey responses from consultant cardiologists with expertise in SE (as a convenience sample) using a secure online survey platform and recruitment panel (Qualtrics). Demographic items were adapted from the UK Office for National Statistics framework [31], and screening questions ensured participant eligibility.

We included explanatory variables in the regression models to explore factors influencing cardiologists' diagnostic confidence and their responses to AI recommendations. Years of experience as a cardiologist were grouped into 4 categories: 1 to 5, 6 to 10, 11 to 15, and 16 or more. The number of monthly SE reviews was categorized as follows: 0 to 20 tests, 21 to 50 tests, 51 to 100 tests, and more than 100 tests. Use of AI in the cardiac care pathway assessed exposure to AI in clinical practice. Participants were asked whether they used AI tools in the diagnosis or management of cardiac patients (yes or no). . Confidence in AI was measured using a 5-point Likert scale (1 = very low, 5 = very high). Respondents also indicated the level of confidence (on a scale from 1% to 100%) they would require in their initial diagnosis to disregard AI advice.

In addition, open-ended questions explored perceived risks of following AI in scenarios where AI tools contradicted clinicians' judgment. Participants provided free-form text describing the potential risks of following the AI recommendation and further actions in response to two scenarios: (1) AI tools contradict the clinician's diagnosis of CAD being present and (2) AI tools contradict the clinician's diagnosis of CAD not being present. Of the 88 responses received, 61 met the inclusion criteria and were retained for analysis.

We used a mixed methods approach to analyze survey responses. Quantitative data were analyzed using regression models to examine associations between clinician characteristics (eg, years of experience, number of SE reviews per month, use of AI, and confidence in AI tools) and their reported confidence in diagnostic decision-making in hypothetical scenarios. In parallel, qualitative thematic analysis was conducted on open-text responses. This involved content analysis to identify common themes related to perceived risks of following AI recommendations and typical follow-up actions taken when AI tools and clinical assessments diverged.

The thematic analysis was based on the 6-phase framework of Braun and Clarke [32]: (1) familiarizing oneself with the data, (2) generating initial qualitative codes, (3) searching for themes, (4) reviewing themes, (5) defining and naming themes, and (6)

XSL•FO
RenderX

producing a report. Two researchers independently coded responses and resolved discrepancies through discussion. Given the anonymous nature of the survey, member checking was not possible. However, qualitative themes were reviewed by 3 consultant cardiologists from PROTEUS field sites to enhance validity.

Ethical Considerations

This study comprises 2 components: a clinical trial and a survey study. For the clinical trial, the protocol, informed consent form, participant information leaflet, and proposed advertising materials were reviewed by the Patient and Public Involvement and Engagement (PPIE) group and approved by the study sponsor and the North West—Preston Research Ethics Committee (REC) of the NHS Health Research Authority (reference: 21/NW/0199). The PROTEUS trial was registered on August 31, 2021, at ClinicalTrials.gov with registration number NCT05028179. For the survey study, ethical approval was obtained from the NHS Health Research Authority (IRAS number 315284) and the London South Bank University Ethics Panel (ETH2223-0164). All participants provided informed consent, and anonymized data were used for both study components. Survey participants were compensated for their participation via Qualtrics, with payments administered directly by the platform rather than the study team.

Results

The Experimental Study Results

Overview

Data from 854 participants in the intervention arm of the PROTEUS trial were used in the experimental analysis. Regional wall motion abnormalities were detected in 8% (71/893) of the participants, and 18% (156/893) received a clinical diagnosis of positive CAD (Table 1). Further details on participant demographics, lifestyle factors, and clinical history are available in Multimedia Appendix 1.

XSL•FO
RenderX

Table . Descriptive statistics and bivariate logistic regression results—primary outcome is concordance between clinician and EchoGo Pro (EGP) or divergence.

Patient characteristics	Convergent[a] (n=512)	Divergent[b] (n=113)	OR (95% CI)	*P* value
Sex, n				
Female	245	50	1	—[c]
Male	267	63	0.86 (0.57-1.30)	.49
Age group (y), n				
18 - 39	20	3	1	—
40 - 59	179	27	0.99 (0.28-3.57)	.99
60 - 80	280	79	0.53 (0.15-1.84)	.32
>81	33	4	1.24 (0.25-6.11)	.79
Smoking status, n				
Current smoker	52	11	1	—
Ex-smoker	216	41	1.11 (0.54-2.32)	.77
Nonsmoker	240	60	0.85 (0.42-1.72)	.64
BMI, mean (SD)	29.0 (10.8)	29.8 (6.2)	0.99 (0.98-1.01)	.45
Control variables				
Has cardiovascular risk factors?, n	406	102	0.41 (0.21-0.80)	.008
Number of cardiovascular risk factors, mean (SD)	1.5 (1.2)	1.9 (1.0)	0.77 (0.65-0.92)	.003
Hypertension, n	254	71	0.58 (0.38-0.89)	.01
Hypercholesterolemia, n	214	58	0.68 (0.45-1.02)	.07
Diabetes, n	90	31	0.56 (0.35-0.90)	.02
Positive family history of coronary artery disease, n	182	41	0.97 (0.63-1.48)	.88
Previous CAD[d] clinical events, n	44	14	0.66 (0.35-1.26	.21
Pre-existing coronary artery disease, n	85	33	0.48 (0.30-0.77)	.002

[a]Concordance refers to (1) both clinician and EchoGo Pro (EGP)–confirmed presence of coronary artery disease (CAD) or (2) both clinician- and EGP-confirmed absence of CAD.

[b]Discordance refers to (1) clinician-diagnosed CAD, but EGP did not confirm it, or (2) EGP-diagnosed CAD, but the clinician did not confirm it. The analysis focuses on data from the experimental study conducted specifically with the EGP tool.

[c]Not applicable.

[d]CAD: coronary artery disease.

Concordance and Discordance

Of the analyzed cases, 512 (60%) were classified as concordant (Table 1). Within this group, 38 (4.5%) cases demonstrated concordance on a positive CAD diagnosis, where both the clinician identified inducible ischemia and EGP reported a high risk of CAD. The remaining 474 (55.5%) cases were concordant on a negative diagnosis, with the clinician reporting no inducible ischemia and EGP indicating low risk.

A total of 113 (13.2%) cases were classified as discordant. Of these, 36 (4.2%) cases involved a clinician-assessed negative result for inducible ischemia, while EGP reported a high risk of CAD. Conversely, 77 (9%) cases involved a positive clinical interpretation for inducible ischemia, but EGP assessed the case as low risk.

If a positive diagnosis were assigned based on either a positive clinician interpretation or a high-risk EGP rating, the total number of positive CAD diagnoses would increase from 153 (17.9%) to 189 (22.1%).

EGP rejected 222 (26.1%) cases due to insufficient image quality, whereas clinicians classified only 10 (1.2%) cases as inconclusive. Concordance among rejected cases was very low, with only 3 (0.4%) cases showing agreement between EGP rejection and clinician inability to interpret.

XSL•FO
RenderX

Factors Predicting Concordance or Discordance

Table 2 presents the comparison of concordance and discordance rates across sex, age groups, and smoking status. Bivariate logistic regression analyses indicated no statistically significant differences in concordance or discordance between male and female participants, across age groups, or among different smoking statuses.

Table . Concordance and discordance between clinician interpretation and EchoGo Pro (EGP) report (experimental study).

	Artificial intelligence risk rating report[a]			
Clinician interpretation of presence or ischemia[b]	High, n (%)	Low, n (%)	Rejected, n (%)	Total, n (%)
Positive for inducible ischemia	38 (4.5)	77 (9)	38 (4.5)	153 (17.9)
Negative for inducible ischemia	36 (4.2)	474 (55.5)	181 (21.2)	691 (80.9)
Inconclusive or abandoned	0 (0)	7 (0.8)	3 (0.4)	10 (1.2)
Total	74 (8.7)	558 (65.3)	222 (26.1)	854 (100)

[a]Artificial intelligence risk rating report was classified as high risk of CAD (stenosis ≥70%) or low risk (stenosis <70%), or the stress echocardiography image was rejected due to poor quality.

[b]Clinician interpretation of the presence of ischemia was classified as a positive coronary artery disease if they diagnosed that wall motion abnormality was present, negative coronary artery disease if no wall motion abnormality was present, or the interpretation was inconclusive or abandoned for inducible ischemia. The analysis focuses on the data from the experimental study conducted specifically with the EchoGo Pro tool.

Having more CRFs or pre-existing CAD was associated with lower odds of concordance. Patients in concordant cases had fewer cardiovascular risk factors than those in discordant cases (mean 1.5, SD 1.2 vs mean 1.9, SD 1.0). Each additional cardiovascular risk factor was associated with lower odds of concordance between EGP and clinician interpretations (OR 0.77, 95% CI 0.65 - 0.92; *P*=.003). The strongest associations with discordance were observed in patients with any cardiovascular risk factor (OR 0.41, 95% CI 0.21 - 0.80; *P*=.008), pre-existing CAD (OR 0.48, 95% CI 0.30 - 0.77; *P*=.002), and diabetes (OR 0.56, 95% CI 0.35 - 0.90; *P*=.02). However, multivariate logistic regression suggested that no sociodemographic variable was associated with discordance or concordance.

EGP Rejecting Scans

We ran a multivariable logistic regression to examine whether sex, age group, smoking status, BMI, and cardiovascular risk factors were associated with the likelihood of SE image rejection by EGP (Table 3). Sex was the only demographic variable significantly associated with rejection: scans from male participants were more likely to be rejected than those from female participants (β=0.35, *P*=.03). Among clinical variables, hypercholesterolemia and family history of CAD showed significant associations with scan rejection. Individuals with hypercholesterolemia were less likely to have their scans rejected (β=−0.36; *P*=.04), whereas those with a positive family history of CAD were more likely to experience rejection (β=0.33; *P*=.049).

XSL•FO
RenderX

Table . Multivariate logistic regression results—primary outcome is artificial intelligence (AI) rejection of scan or conclusive scan[a].

Patient characteristics	Coefficient (SE)	*P* value
Intercept	−1.28 (0.56)	.02
Age groups (y)		
18 - 39	__[b]	—
40 - 59	0.02 (0.46)	.97
60 - 80	0.31 (0.45)	.50
>81	0.69 (0.53)	.19
Sex		
Female	—	—
Male	0.35 (0.17)	.03
Smoking		
Current smoker	—	—
Ex-smoker	−0.20 (0.28)	.48
Nonsmoker	−0.11 (0.27)	.68
BMI	0.00 (0.01)	.78
Cardiovascular risk factors		
Hypertension	0.03 (0.17)	.87
Hypercholesterolemia	−0.36 (0.17)	.04
Diabetes	−0.04 (0.21)	.85
Positive family history of coronary artery disease	0.33 (0.17)	.049
Previous CAD[c] clinical events	0.20 (0.26)	.43

[a]The analysis focuses on the data from the experimental study conducted specifically with the EchoGo Pro tool.

[b]Not applicable.

[c]CAD: coronary artery disease.

Survey Results

Overview

The data from 61 survey participants were included in the analysis. Most were male (50/60, 85%), aged 40 to 49 years (26/57, 46%), and White (39/61, 64%). Professional experience varied, with the majority having more than 5 years of experience as cardiologists (Multimedia Appendix 1). Over half of the survey participants reported using AI tools in cardiac care (33/61, 54%), including auto-indexing software (33/61, 54%) and decision-support software (23/61, 38%). The number of monthly SE tests reviewed ranged from 1 to over 100, with nearly half reviewing 21 to 50 tests per month.

Scenario Analysis

The scenario analysis used survey responses in which AI tools were considered in general rather than focusing on a specific system such as EGP. In scenario A (the cardiologist diagnosed CAD, but the AI tools contradicted this), the mean level of confidence in the initial diagnosis required to justify taking no further action was 69%. In scenario B (CAD was not identified by the cardiologist but was diagnosed by the AI tools), the corresponding mean confidence level was 65%.

Table 4 presents linear regression models examining how cardiologist characteristics relate to the confidence threshold required to disregard AI advice. Across both scenarios, this threshold was not significantly associated with years of experience as a cardiologist, the number of stress echocardiograms reviewed per month, or the current use of AI in the cardiac care pathway. However, in scenario A, higher confidence in AI tools was positively associated with the level of diagnostic confidence required to disregard an AI recommendation (β=7.73; *P*=.02). This suggests that cardiologists who are more trusting of AI systems are more cautious and require stronger conviction in their own clinical judgment before overriding AI-generated assessments.

XSL•FO
RenderX

Table . Regression models for scenario analysis: cardiologists' confidence in their initial diagnosis[a].

Predictor variable	Scenario A: cardiologist made a diagnosis of CAD[b], but the AI[c] tools contradicted[d]		Scenario B: CAD was not identified by cardiologists, but was diagnosed by AI tools[e]	
	Coefficient (SE)	*P* value	Coefficient (SE)	*P* value
Intercept	39.37 (12.07)	.002	55.49 (14.26)	<.001
Years of experience as a cardiologist				
1 - 5 (reference)	__[f]	—	—	—
6 - 10	−1.51 (7.39)	.84	0.45 (8.74)	.96
11 - 15	10.21 (7.48)	.18	−1.21 (8.85)	.89
>16	0.90 (7.19)	.90	−0.52 (8.50)	.95
Number of monthly stress test reviews				
0 - 20 tests (reference)	—	—	—	—
21 - 50 tests	3.61 (6.83)	.60	−11.57 (8.07)	.16
51 - 100 tests	4.13 (8.92)	.65	−8.72 (10.54)	.41
>101 tests	4.90 (7.83)	.53	−3.93 (9.26)	.67
AI in the cardiac care pathway				
No (reference)	—	—	—	—
Yes	−3.68 (5.06)	.47	2.88 (5.98)	.63
Confidence in AI tool[g]	7.73 (3.11)	.02	4.58 (3.67)	.22

[a]The analysis presented in this table is based on survey data where artificial intelligence tools are considered in general.

[b]CAD: coronary artery disease.

[c]AI: artificial intelligence.

[d]Confidence in initial diagnosis for not taking any further actions: 69%.

[e]Confidence in initial diagnosis for not taking any further actions: 65%.

[f]Not applicable.

[g]Confidence in AI was measured on a 5-point Likert scale (1=very low, 5=very high). Confidence in their own diagnosis was measured on a scale from 1% to 100%. Regression coefficients are unstandardized.

Perceived Risk of Following AI Recommendations

In scenario A, the most perceived risk of following AI recommendations, considered broadly rather than specific to EGP, was the possibility of inaccurate diagnoses or inappropriate treatments, which could potentially cause harm. Other prominent concerns included AI limitations, such as lack of patient-specific context, inadequate training, and the early-stage development of AI tools, as well as legal liability, with cardiologists remaining ultimately responsible for patient care. Some participants, however, perceived minimal risk, viewing AI as a nonbinding decision-support tool.

In scenario B, the risk of inaccurate diagnosis emerged as a key theme in participants' perceptions of the risks associated with following AI recommendations. Two additional prominent themes were the overuse of clinically ineffective but nonharmful services and the misuse of care—specifically, unnecessary treatments that could potentially harm patients or cause side effects. Participants expressed these concerns with statements such as: "Some treatments may not improve patient symptoms. Unnecessary intervention could be harmful," and "Overtreatment, possible side effects of therapy, patient anxiety."

Actions Taken Following AI Recommendations

Study participants were responsive to situations in which AI recommendations contradicted cardiologists' clinical judgment regarding the presence or absence of CAD. Across both scenarios, the most frequently reported response was to initiate additional investigations. Another prominent theme was the pursuit of a second opinion, consultation, or advice, reflected in subthemes such as seeking a second opinion, requesting advice, or consulting peer colleagues. A third major theme involved reviewing broader patient data, including diagnostic tests and medical records, to verify or confirm information through a more comprehensive assessment. In a small number of cases, participants reported taking no further action. Finally, a less common theme was the repetition of tests or re-examination of existing information, suggesting an intent to audit or reassess the initial diagnosis and underlying data.

Discussion

This mixed methods study combined PROTEUS trial data analysis with a cardiologist survey, examining how EGP's stress echo assessments aligned with clinician interpretations and exploring attitudes toward AI-human discordance. EGP showed

XSL•FO
RenderX

substantial overall agreement with clinicians but generated meaningful discordance, concentrated among patients with multiple cardiovascular risk factors and established CAD. Concordance was similar across sex, age, and smoking status but lower in patients with hypertension, diabetes, or pre-existing CAD. Over one-quarter of scans were rejected due to poor image quality, with rejection more common in male patients and those with a family history of CAD, and less frequent in patients with hypercholesterolemia.

Survey respondents described AI as supplementary rather than primary decision support. When facing discordance, cardiologists reported retaining confidence in their initial judgment and seeking corroboration through further investigations, clinical data review, or second opinions rather than deferring to AI. Those who expressed higher overall confidence in AI reported that they would require stronger confidence in their own diagnosis before disregarding an AI recommendation, particularly when the AI suggested disease they had not initially identified. Some expressed concerns that following discordant AI advice could trigger unnecessary investigations and overtreatment.

The concentration of discordance in patients with complex profiles highlights the limitations of image-only approaches [33]. This is clinically significant because patients with hypertension, diabetes, and pre-existing CAD face substantially higher risks of future events [34], yet accurate diagnosis in these critical groups is where EGP and clinicians most often diverged. EGP is trained on images [35], whereas clinicians integrate clinical history, comorbidities, and broader patient data. For example, patient-specific anatomical features can contribute to false positive findings by creating artifacts that mimic ischemia, highlighting how structural features influence diagnostic variability [36]. Consequently, future systems must move toward multimodal designs that incorporate anatomical features, clinical history, and comorbidities to better mirror diagnostic reasoning [23,36,37].

These findings highlight both EGP's potential and its limitations. When used alongside expert interpretation, EGP can flag additional high-risk patients who might otherwise be missed. Given that missed CAD diagnoses cost approximately £4200 per patient versus £2300 for correct diagnoses [38], even modest gains in diagnostic yield may improve outcomes and reduce downstream costs. However, EGP alone often classified fewer positives than clinicians and failed to interpret rejected scans, reinforcing its current role as an adjunctive decision-support tool rather than a stand-alone solution.

We observed a nonsignificant trend toward increased false positives with EGP. At population scale, even minor increases affect service planning; a theoretical 1% rise in false positives across the United Kingdom's 61,000 annual scans [35] would trigger approximately 610 unnecessary investigations, costing roughly £1.4 million. Beyond financial burdens, false positives generate patient anxiety and avoidable invasive procedures. Consequently, the trade-offs between enhanced sensitivity and the risks of over-diagnosis warrant careful consideration during AI upscaling.

A further challenge relates to EGP's relatively high rate of scan rejection, which undermines its reliability and usability. Frequent noninterpretable outputs are likely to erode clinicians' trust and limit the perceived usefulness of the system. Importantly, rejection was not uniform. Although mechanisms cannot be determined directly from our data, these differences are consistent with broader evidence that imbalances in training datasets, such as sex disparities or underrepresentation of certain clinical risk profiles, can lead to biased performance of AI models across demographic and clinical groups [39-41]. Addressing these inequities requires more representative training data, routine postdeployment monitoring across subgroups, and model refinement where necessary.

The survey findings clarify how such tools are likely to be used. Cardiologists framed AI as supportive rather than authoritative, responding to disagreements by seeking further tests, reviewing additional data, or obtaining second opinions, suggesting a tendency toward confirmation-oriented behavior rather than automation bias [22,42]. While this caution mitigates overreliance risks, frequent discordance may paradoxically increase downstream testing and costs. Clear guidance on managing AI-clinician disagreement and warranting additional investigations is therefore essential for implementation.

This study has several notable strengths. Leveraging data from the multicenter PROTEUS trial provided a robust and standardized framework for evaluating EGP across diverse clinical sites, reducing the influence of local practice patterns and supporting internal validity. The sizeable sample allowed the examination of agreement, disagreement, and scan rejection across clinically relevant subgroups. By explicitly modeling predictors of AI-clinician concordance and AI rejection, we begin to identify which patients may be better or worse served by current image-based systems. Finally, the mixed methods design offers a holistic understanding of both EGP's technical behavior and how clinicians interpret and act on its recommendations in practice.

A few limitations should be borne in mind. First, the high scan rejection rate reduced the effective sample size for concordance analyses and may have introduced selection bias if rejected scans systematically differed from accepted scans in unmeasured ways. Second, although we included a range of demographic and clinical characteristics, other potentially important determinants of image quality and diagnostic decision-making, such as sonographer expertise, variation in acquisition protocols, and differences in operators' familiarity with EGP, were unavailable but likely influence both AI performance and clinician interpretations [43,44]. Finally, both study components were conducted within a specific national health care context; EGP's performance and acceptance may differ in other health systems or populations.

Beyond EGP, these findings have broader implications for AI implementation in cardiovascular imaging. Discordance concentrated in patients with hypertension, diabetes, and prior CAD reflects the limitations of image-only models that cannot incorporate symptoms, medications, and comorbidities as clinicians do. Future AI systems must integrate these clinical variables. The uneven rejection rates by sex and clinical profile

XSL•FO
RenderX

require ongoing monitoring for bias after deployment. Additionally, cardiologists' concerns about legal liability when AI recommendations conflict with their judgment must be addressed through clear clinical protocols and regulatory guidance. Further research is needed to establish evidence-based approaches for managing AI-clinician disagreement and ensuring equitable performance across patient populations.

In this mixed methods study, EGP agreed with cardiologists in most cases, but discordance clustered in patients with multiple cardiovascular risk factors or established coronary disease, and many scans were rejected, particularly in some patient groups. Cardiologists described using AI as supportive input rather than a decision-maker, typically responding to AI-clinician disagreement by ordering further tests, reviewing additional data, or seeking second opinions. Overall, these findings position EGP and similar tools as useful adjuncts that can highlight additional at-risk patients but cannot yet replace holistic clinical assessment. Future development should prioritize incorporating key clinical information alongside imaging, reducing and monitoring rejection and bias across subgroups, and embedding AI within clear clinical pathways so that disagreement is managed transparently and patient safety, efficiency, and equity are all improved.

Acknowledgments

The authors would like to thank Virgil Griffith and Sharanya Jayadev for their valuable support to this research. They are also grateful to the PROTEUS trial and Ultromics Ltd for providing access to the data that made this research possible. Finally, we extend our appreciation to all those who contributed to the completion of this study.

The authors declare the use of generative artificial intelligence (GAI) in the writing process. According to the GAIDeT taxonomy (2025), GAI tools were used to assist in proofreading and editing under full human supervision. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was supported by NHSX, the former joint digital transformation unit of National health services (NHS) England and the Department of Health and Social Care (United Kingdom). This report is independent research funded by the National Health Services (AI Award in Health and Care, Technology Specific Evaluation Team). The views expressed are those of the authors and not necessarily those of the National Health Services, NHSX, or the Department of Health and Social Care.

Data Availability

The survey data generated and analyzed during this study are available from the corresponding author upon reasonable request, subject to a data-sharing agreement and approval from the relevant ethics committee. The experimental study data, collected as part of the PROTEUS trial, can be requested directly from Ultromics Ltd, in accordance with their data-sharing policies and applicable approvals.

Authors' Contributions

DF, CF, and MM conceptualized the study. SH and RP reviewed and commented on study materials. MM prepared the data and conducted the analyses, supported by SW. MM and DF prepared the initial draft of the manuscript. All authors contributed to the review and editing of the manuscript.

Conflicts of Interest

SH has received speaker honoraria from Ligence AI, has received research funding for research agreements with Ventripoint Medical Systems and Ligence AI, and has held a research agreement with funding from Ultromics. The other authors declare no conflicts of interest.

Multimedia Appendix 1
Additional descriptive statistics and subgroup analyses of participants in the experimental study and cardiologist survey.
[DOCX File, 139 KB - jmir_v28i1e83541_app1.docx]

References

1. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digit Med 2021 Jan 5;4(1):4. [doi: 10.1038/s41746-020-00367-3] [Medline: 33402680]
2. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. JAMA 2018 Dec 4;320(21):2199-2200. [doi: 10.1001/jama.2018.17163] [Medline: 30398550]
3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]

XSL•FO
RenderX

4. Song Y, Ren S, Lu Y, Fu X, Wong KKL. Deep learning-based automatic segmentation of images in cardiac radiography: a promising challenge. Comput Methods Programs Biomed 2022 Jun;220:106821. [doi: 10.1016/j.cmpb.2022.106821] [Medline: 35487181]
5. Raghu D, Tripathy HK. Medical image analysis using deep learning algorithm convolutional neural networks. In: Pandian AP, Fernando X, Haoxiang W, editors. Computer Networks, Big Data and IoT 2022, Vol. 117:245-256. [doi: 10.1007/978-981-19-0898-9_19]
6. Liu B, Reis J, Sharma A, Wang W. Application of artificial intelligence in non-invasive cardiovascular imaging for coronary artery disease: a systematic review and meta-analysis. Front Cardiovasc Med 2025;12:1664183. [doi: 10.3389/fcvm.2025.1664183] [Medline: 41458983]
7. Shamszare H, Choudhury A. Clinicians' perceptions of artificial intelligence: focus on workload, risk, trust, clinical decision making, and clinical integration. Healthcare (Basel) 2023 Aug 16;11(16):2308. [doi: 10.3390/healthcare11162308] [Medline: 37628506]
8. El Arab RA, Al Moosa OA. Systematic review of cost effectiveness and budget impact of artificial intelligence in healthcare. NPJ Digit Med 2025 Aug 26;8(1):548. [doi: 10.1038/s41746-025-01722-y] [Medline: 40858882]
9. Moss AJ, Williams MC, Newby DE, Nicol ED. The updated NICE guidelines: cardiac CT as the first-line test for coronary artery disease. Curr Cardiovasc Imaging Rep 2017;10(5):15. [doi: 10.1007/s12410-017-9412-6] [Medline: 28446943]
10. de Siqueira VS, Borges MM, Furtado RG, Dourado CN, da Costa RM. Artificial intelligence applied to support medical decisions for the automatic analysis of echocardiogram images: a systematic review. Artif Intell Med 2021 Oct;120:102165. [doi: 10.1016/j.artmed.2021.102165] [Medline: 34629153]
11. Shivaie S, Tohidi H, Loganathan P, Kar M, Hashemy H, Shafiee MA. Interobserver variability of coronary stenosis characterized by coronary angiography: a single-center (Toronto General Hospital) retrospective chart review by staff cardiologists. Vasc Health Risk Manag 2024;20:359-368. [doi: 10.2147/VHRM.S431612] [Medline: 39157424]
12. Rovetto C, Niemann T, Almeida G, et al. Assessment of inter- and intrareader agreement among different stenosis quantification methods in coronary CT angiography. Eur J Radiol 2025 May;186:112059. [doi: 10.1016/j.ejrad.2025.112059] [Medline: 40127592]
13. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019 Oct;1(6):e271-e297. [doi: 10.1016/S2589-7500(19)30123-2] [Medline: 33323251]
14. Holste G, Oikonomou EK, Tokodi M, Kovács A, Wang Z, Khera R. Complete AI-enabled echocardiography interpretation with multitask deep learning. JAMA 2025 Jul 22;334(4):306-318. [doi: 10.1001/jama.2025.8731] [Medline: 40549400]
15. Alsharqi M, Woodward WJ, Mumith JA, Markham DC, Upton R, Leeson P. Artificial intelligence and echocardiography. Echo Res Pract 2018 Dec 1;5(4):R115-R125. [doi: 10.1530/ERP-18-0056] [Medline: 30400053]
16. Massalha S, Clarkin O, Thornhill R, Wells G, Chow BJW. Decision support tools, systems, and artificial intelligence in cardiac imaging. Can J Cardiol 2018 Jul;34(7):827-838. [doi: 10.1016/j.cjca.2018.04.032] [Medline: 29960612]
17. Upton R, Mumith A, Beqiri A, et al. Automated echocardiographic detection of severe coronary artery disease using artificial intelligence. JACC Cardiovasc Imaging 2022 May;15(5):715-727. [doi: 10.1016/j.jcmg.2021.10.013] [Medline: 34922865]
18. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020 Mar 25;368:m689. [doi: 10.1136/bmj.m689] [Medline: 32213531]
19. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1):310. [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
20. Gaube S, Suresh H, Raue M, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit Med 2021 Feb 19;4(1):31. [doi: 10.1038/s41746-021-00385-9] [Medline: 33608629]
21. Sivaraman V, Bukowski LA, Levin J, Kahn JM, Perer A. Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. 2023 Presented at: CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. [doi: 10.1145/3544548.3581075]
22. Snow T. From satisficing to artificing: the evolution of administrative decision-making in the age of the algorithm. Data & Policy 2021;3:e3. [doi: 10.1017/dap.2020.25]
23. Joseph J, Kartheeban K. Cognitive alignment in cardiovascular AI: designing predictive models that think with, not just for, clinicians. Front Cardiovasc Med 2025;12:1651324. [doi: 10.3389/fcvm.2025.1651324] [Medline: 40964520]
24. Hassan N, Slight R, Bimpong K, et al. Systematic review to understand users perspectives on AI-enabled decision aids to inform shared decision making. NPJ Digit Med 2024 Nov 21;7(1):332. [doi: 10.1038/s41746-024-01326-y] [Medline: 39572838]
25. Mahdavi M, Thomas N, Flood C, et al. Evaluating artificial intelligence-driven stress echocardiography analysis system (EASE study): a mixed method study. BMJ Open 2024 Oct 2;14(10):e079617. [doi: 10.1136/bmjopen-2023-079617] [Medline: 39357985]
26. Woodward G, Bajre M, Bhattacharyya S, et al. PROTEUS study: a prospective randomized controlled trial evaluating the use of artificial intelligence in stress echocardiography. Am Heart J 2023 Sep;263:123-132. [doi: 10.1016/j.ahj.2023.05.003] [Medline: 37192698]

XSL•FO
RenderX

27. Kottner J, Audigé L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol 2011 Jan;64(1):96-106. [doi: 10.1016/j.jclinepi.2010.03.002] [Medline: 21130355]
28. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med 2007 Oct 16;147(8):573-577. [doi: 10.7326/0003-4819-147-8-200710160-00010] [Medline: 17938396]
29. Choi WJ, Lee HS, Hong JH, Chang HJ, Lee JW. Comparison of the JNC7 and 2017 American College of Cardiology/American Heart Association guidelines for the management of hypertension in Koreans: analysis of two independent nationwide population-based studies. Int J Environ Res Public Health 2019 Dec 16;16(24):1-13. [doi: 10.3390/ijerph16245134] [Medline: 31888171]
30. ElSayed NA, Aleppo G, Aroda VR, et al. 2. Classification and diagnosis of diabetes: standards of care in diabetes-2023. Diabetes Care 2023 Jan 1;46(Suppl 1):S19-S40. [doi: 10.2337/dc23-S002] [Medline: 36507649]
31. Demography question development for Census 2021. Office for National Statistics. 2024. URL: https://www.ons.gov.uk/census/censustransformationprogramme/questiondevelopment/demographyquestiondevelopmentforcensus2021 [accessed 2025-09-04]
32. Braun V, Clarke V, Newson L. Thematic analysis: a practical guide to understanding and doing. QMiP Bulletin 2022;1(33):46-50. [doi: 10.53841/bpsqmip.2022.1.33.46]
33. Migliaro S, Celotto R, Teliti R, Mariani S, Altamura L, Tomai F. Comparing AI-driven and heart team decision-making in multivessel coronary artery disease. J Clin Med 2025 Jun 23;14(13):4452. [doi: 10.3390/jcm14134452] [Medline: 40648826]
34. Visseren FLJ, Mach F, Smulders YM, et al. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. Eur Heart J 2021 Sep 7;42(34):3227-3337. [doi: 10.1093/eurheartj/ehab484] [Medline: 34458905]
35. Zeltzer D, Herzog L, Pickman Y, et al. Diagnostic accuracy of artificial intelligence in virtual primary care. Mayo Clin Proc Digit Health 2023 Dec;1(4):480-489. [doi: 10.1016/j.mcpdig.2023.08.002] [Medline: 40206313]
36. Sonaglioni A, Nicolosi GL, Rigamonti E, Lombardo M, Gensini GF, Ambrosio G. Does chest shape influence exercise stress echocardiographic results in patients with suspected coronary artery disease? Intern Emerg Med 2022 Jan;17(1):101-112. [doi: 10.1007/s11739-021-02773-1] [Medline: 34052977]
37. Harada T, Shimizu T, Kaji Y, et al. A perspective from a case conference on comparing the diagnostic process: human diagnostic thinking vs. artificial intelligence (AI) decision support tools. Int J Environ Res Public Health 2020 Aug 22;17(17):1-6. [doi: 10.3390/ijerph17176110] [Medline: 32842581]
38. Johnson CL, Woodward W, McCourt A, et al. Real world hospital costs following stress echocardiography in the UK: a costing study from the EVAREST/BSE-NSTEP multi-entre study. Echo Res Pract 2023 May 31;10(1):8. [doi: 10.1186/s44156-023-00020-1] [Medline: 37254216]
39. Vervoort D, Wang R, Li G, et al. Addressing the global burden of cardiovascular disease in women: JACC state-of-the-art review. J Am Coll Cardiol 2024 Jun 25;83(25):2690-2707. [doi: 10.1016/j.jacc.2024.04.028] [Medline: 38897679]
40. Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. Health data poverty: an assailable barrier to equitable digital health care. Lancet Digit Health 2021 Apr;3(4):e260-e265. [doi: 10.1016/S2589-7500(20)30317-4] [Medline: 33678589]
41. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci U S A 2020 Jun 9;117(23):12592-12594. [doi: 10.1073/pnas.1919012117] [Medline: 32457147]
42. Selten F, Robeer M, Grimmelikhuijsen S. 'Just like I thought': street‐level bureaucrats trust ai recommendations if they confirm their professional judgment. Public Adm Rev 2023 Mar;83(2):263-278. [doi: 10.1111/puar.13602]
43. Knackstedt C, Bekkers SCAM, Schummers G, et al. Fully automated versus standard tracking of left ventricular ejection fraction and longitudinal strain: the FAST-EFs multicenter study. J Am Coll Cardiol 2015 Sep 29;66(13):1456-1466. [doi: 10.1016/j.jacc.2015.07.052] [Medline: 26403342]
44. Tajik AJ. Machine learning for echocardiographic imaging: embarking on another incredible journey. J Am Coll Cardiol 2016 Nov 29;68(21):2296-2298. [doi: 10.1016/j.jacc.2016.09.915] [Medline: 27884248]

Abbreviations

AI: artificial intelligence
CAD: coronary artery disease
CRF: cardiovascular risk factor
EGP: EchoGo Pro
SE: stress echocardiography
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

XSL•FO
RenderX

Edited by A Stone; submitted 04.Sep.2025; peer-reviewed by A Sonaglioni, JM Hoppe, S Krasner; revised version received 13.Mar.2026; accepted 31.Mar.2026; published 20.May.2026.

Please cite as:
Mahdavi M, White S, Hothi SS, Flood C, Panayotova R, Frings D
Alignment Between Cardiologists and AI-Driven Diagnostic Systems: Mixed Methods Study
J Med Internet Res 2026;28:e83541
URL: https://www.jmir.org/2026/1/e83541
doi:10.2196/83541

© Mahdi Mahdavi, Sarah White, Sandeep S Hothi, Chris Flood, Rosica Panayotova, Daniel Frings. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Patients' Perspectives on the Implementation of AI in Radiological Diagnostics: Focus Group Study

Cornelia R Karger

Brain and Behaviour/ Research Group: Neuroethics and Ethics in AI, Institute of Neurosciences and Medicine, Forschungszentrum Jülich GmbH, Wilhelm-Johnen-Strasse, Jülich, Germany

Corresponding Author:
Cornelia R Karger
Brain and Behaviour/ Research Group: Neuroethics and Ethics in AI, Institute of Neurosciences and Medicine, Forschungszentrum Jülich GmbH, Wilhelm-Johnen-Strasse, Jülich, Germany

Abstract

Background: Rapid developments in artificial intelligence (AI) will enable its widespread use in radiological diagnostics in the near future. Patients will then be confronted with findings generated with the help of AI. Understanding patients' perspectives on the use of this technology is one of the key factors for its successful implementation.

Objective: This qualitative study aimed to gain insight into patients' reasoning about the opportunities and risks of using AI in radiological diagnosis, to identify prerequisites for its acceptance, and to identify aspects that can promote trust in AI diagnoses, especially in scenarios of high personal concern.

Methods: A total of 7 focus groups were conducted with 34 patients (n=15, 44% female participants) aged between 23 and 85 years (mean 49.06, SD 17.08 y), recruited using purposive sampling strategies. Each focus group was audiotaped, transcribed, and analyzed using the method of structured qualitative content analysis.

Results: Study findings show that patients are open to the use of AI in radiological diagnostics. The basic prerequisites for this are (1) scientific evidence of safe outcomes that are more accurate and faster than those without AI; (2) recognizable added value in patient care; (3) transparency in the use of AI and disclosure to the patient; (4) comprehensive, binding measures for quality assurance; and (5) the use of AI solely to support the physician. However, the results indicate that further criteria are important for patients to be willing to choose a radiologist who uses AI and to trust AI diagnoses. In situations where they are personally affected, patients fear that physicians will place too much trust in the AI result and that the physician-patient relationship will become dehumanized. Therefore, the physicians' abilities and functions that inspire trust from the patients' perspective must come into play. These include (1) an independent diagnosis by the physician that takes into account not only the clinical context but also the individuality of the patient, (2) a comprehensible explanation of the pros and cons of using AI for patients and clear communication of AI output, and (3) a humane and empathetic physician-patient relationship, which shows that the physician continues to feel responsible for the patient.

Conclusions: The results of the study underscore that a high quality of the entire "sociotechnical" system is an essential prerequisite for patient acceptance of the use of AI in radiological diagnostics and for trust in AI diagnoses. The further development of AI performance must go hand in hand with the creation of framework conditions for its use that meet patients' expectations of the role of the physician and ensure a trust-building physician-patient relationship. The study provides valuable insights into how such integration of AI into radiological practice can be achieved.

(J Med Internet Res 2026;28:e89178) doi:10.2196/89178

KEYWORDS

artificial intelligence; AI; radiology; diagnostics; trust; acceptance; focus groups; qualitative study

Introduction

The rapid progress of AI technology has revolutionized health care in various fields [1]. Radiology is among the pioneers of this development. X-ray images, as well as magnetic resonance imaging (MRI) and computed tomography scans, have long been digitized, thus providing a solid basis for integrating AI in the diagnosis, prediction, or classification of diseases [2]. One of the biggest advantages is enhancing the efficiency of interpreting medical images and increasing diagnostic accuracy [3,4].

Studies revealed that AI models can classify lung nodules on computed tomography scans with high accuracy [5,6]. A review of 38 studies, including 2924 cases of multiple sclerosis and 2509 healthy controls reported in 29 studies, showed a sensitivity ranging from 77% to 100% and a specificity of 74% to 100%. Furthermore, 34 studies reported accuracy ranging from 81% to 100% [7]. Comparisons show that the accuracy of

XSL•FO
RenderX

AI often matches or even exceeds the performance of radiologists [8,9]. The detection of COVID-19 pneumonia in chest radiography using deep learning algorithms, for example, was achieved with a sensitivity and specificity of 96% and 64%, respectively, compared to radiologists' sensitivity and specificity, which ranged between 50% and 73% [10]. In patients with amyloid–related imaging abnormalities, radiologists had significantly better detection performance in the assisted setting with a deep learning method (87%) compared to the unassisted setting (71%) [11].

However, there are challenges that must be taken into account when introducing the technology into regular clinical practice.

One key challenge concerns the data used to train AI. The problem is not only the quantity and quality of the training data but, above all, the appropriate validation of the AI model. Critical issues include, for example, intentional or unintentional distortions due to a lack of representativeness or biases in data labels and misclassifications. Most AI applications are developed and trained using datasets from high-income countries. Minority groups and non-Western populations are, therefore, often underrepresented. Such distortions can result in suboptimal performance for certain patient groups and incorrect diagnoses [12-14]. Moreover, the problem of the "black box" nature of AI models makes it difficult to detect and clarify errors. In practice, however, it has so far been the responsibility of radiologists to monitor the output. Therefore, efforts are being made to make deep learning models explainable so that radiologists can fulfill their responsibility [15,16]. At the same time, studies on automation bias show that there is a risk that radiologists may place too much trust in AI results [17].

Beyond the performance of AI, successful implementation depends on the acceptance and trust of those affected, particularly patients [18,19]. Trust is crucial in almost any type of situation where uncertainty exists or undesirable outcomes are possible [20]. In the sensitive field of radiological diagnostics, which carries potentially life-threatening consequences, trust is especially challenged. Acceptance and trust are shaped not only by individual factors, but also by the national context [21]. Unlike people in emerging economies and non-Western countries, individuals in industrialized nations with a high-income context and a cultural background similar to that of Germany tend to be rather skeptical about AI [22]. This skepticism also appears to extend to the use of AI in the health sector. Studies in such industrialized countries show that patients' fear the risk of misdiagnosis when AI is used [23]. In a study by Nelson et al [24] on skin cancer diagnostics, patients expressed concerns about inadequate training sets in the development of AI. Distrust in the quality of AI modeling can lead to a significantly higher risk assessment of its use [25]. Transparency in the use of AI and disclosure to patients are seen as ways to promote acceptance [26]. The importance of such aspects for patients' willingness to choose a radiologist who uses AI in situations that affect them personally has hardly been studied to date. Various studies have shown that patients are more supportive of AI roles where the technology supports—rather than replaces—physicians [27,28]. In the study by Jutzi et al [29], only 41% of patients advocated for AI as a stand-alone system in an online survey on skin tumor diagnostics, while the majority of 94% preferred AI solely to support the physician. The study by Ozcan et al [30] on mammography diagnostics showed similar results. Only 4% of patients accepted a stand-alone AI interpretation, while a majority of 71% preferred AI to be used as a second opinion. If the AI output is not checked by physicians, patients distrust the diagnosis more than they would if it were reviewed by a physician [31,32]. So far, however, there is little information about the reasoning behind patients' observed mistrust of AI. The same applies to the role of the physician and the importance of the physician-patient relationship. Studies have merely shown that personal communication and empathy are considered necessary for a satisfactory physician-patient relationship and should not be endangered by the use of AI [33,34]. However, it remains unclear which characteristics of the physician-patient relationship can promote trust and to what extent these could be impaired in patients' perceptions through the use of AI in radiological diagnostics.

The aim of this study was to fill these gaps by gaining insights into patients' perspectives on risks and benefits, under what conditions patients accept the use of AI in radiological diagnostics and trust the AI diagnosis, what they expect from physicians, and what motives underlie their attitudes.

Methods

Study Design

A focus group design was selected for the study [35,36]. The inherent group dynamics of the focus group format allow for deep insights into the nature of sensitive and difficult-to-grasp topics, such as acceptance and trust regarding complex health-related technological developments [37]. This study used high-stakes personal health scenarios to capture often rather implicit assumptions, beliefs, values, and expectations. The study design and reporting followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines (Checklist 1) [38].

Data Collection

Patients were recruited with the help of patient organizations and self-help representatives who forwarded the invitation to participate in the study to their members, allowing them to register for the study on their own initiative. In addition, calls were made via social media, and flyers were distributed at relevant events. The participants were selected according to the principle of "purposive sampling" [39]. The inclusion criteria considered were that participants had undergone preventive medical checkups and/or were currently under the care of a general practitioner or specialist. They also needed to be proficient in German and at least the age of 18 years.

Focus groups were conducted from November 2022 to June 2023 in the state of North Rhine-Westphalia, Germany. They were conducted until saturation of arguments was reached [40]. Each focus group consisted of 4 to 6 participants. Participants were asked to complete a short questionnaire on demographic data, experiences with radiological examinations, and affinity for technology using the German version of the Affinity for

XSL•FO
RenderX

Technology Interaction Scale [41]. All focus groups were held as face-to-face sessions. No relationship was established between participants and researchers prior to the focus groups. The participants were informed about the project's objectives, the interdisciplinary research team, and the participating research institutions. The empirical study was conducted by a senior researcher (psychologist) with extensive experience in qualitative social research methods. She was assisted by a junior researcher (psychologist).

At the beginning of the focus groups, participants received standardized information on AI–based diagnostics in radiology (Multimedia Appendix 1), as well as information on the purpose and procedure of the focus group. A semistructured, pilot-tested guide served as the framework for the discussion in all focus groups. Prior to the study, interviews were conducted with representatives of patient organizations and medical professional associations to identify relevant topics and develop the guide in a manner appropriate for the target group. The interview guide comprised key questions and additional questions.

To start the discussion, the participants were asked the following questions: "What opportunities and risks do you see?" and "What would need to be guaranteed for you to endorse the use of AI in radiology? Are there circumstances under which you would reject its use?" In the main part of the discussion, key questions were introduced through personal involvement scenarios. The first scenario described the following situation: "Suppose your physician is unsure whether you are developing a serious illness, such as multiple sclerosis, and would like to refer you for an MRI scan to detect the disease as early as possible and, if necessary, initiate treatment." Participants were then asked, "Would you choose a radiologist who does or does not use AI? What would your decision depend on? Which AI function would you feel most comfortable with when evaluating images or making a diagnosis?" The second scenario described a situation in which an MRI report had been generated using AI and contained indications of abnormalities that suggested the presence of a serious disease, such as multiple sclerosis. Participants were subsequently asked, "Would you trust the AI diagnosis? Under what circumstances? Who would you trust more, the physician or the AI, if your physician doubts the AI diagnosis?" In the third scenario, participants were asked to imagine a development in which AI is used extensively in diagnostics. The following key questions were asked, "Could this change your physician-patient relationship, or would it remain unaffected? What expectations do you have of your physician in relation to this future development?"

For all key questions, participants were encouraged to explain the reasons for their views. The key questions were visualized to stimulate thinking and argumentation. All focus groups were facilitated by the author in a neutral role. A "funnel design" was used [42]. In each group, 1 research assistant noted nonverbal information on the interaction among the participants. Each focus group lasted around 2 hours, was audiotaped, and transcribed verbatim.

Data Analysis

Data analysis was performed using structured qualitative content analysis according to Mayring [43], as well as framework analysis [44]. Both methods are research techniques for the systematic analysis of manifest and latent communication content. The strength of structured qualitative content analysis lies in its theoretically grounded, rule-based categorization. When combined with matrix-based framework analysis, comparisons can be made across themes and cases, enabling patterns to be identified. This allows for a broad and in-depth analysis of the data, particularly when dealing with extensive datasets—as in this study.

Deductive and inductive categorization were combined using the software MAXQDA 2022 (VERBI Software GmbH). Deductive categories formed the basic framework, which was supplemented with inductive categories. Conflicts arising between the 2 approaches, such as redundancies and the resulting ambiguities in the category system, required revisions, standardizations, and re-examinations based on the raw data. The resolution of such conflicts was managed through a high degree of flexibility in the creation of the category system and transparent, traceable documentation of all disagreements and adjustments. Any disagreements arising during the development of categories were discussed within the research team until consensus was reached. In addition to the author, another coder from the research team was involved to ensure quality. There was good interrater reliability (κ=0.73).

In the first step, both raters read all transcripts, wrote brief summaries of what had been discussed, and made comments on the text passages that appeared relevant to them. In the second step, main thematic categories were developed deductively based on the literature review and the guide. A total of 4 focus groups were rated by the 2 raters independently of each other, the results were discussed, and the preliminary categories were modified and added to. The entire material was then rated based on the main categories. In the third step, the inductive development of subcategories was carried out. In several rounds of discussion, the entire category system was broken down, and the entire material was then rated independently. In the fourth step, statement matrices were organized according to topics, and, finally, documents were created on the basis of quotations.

Ethical Considerations

This focus group study with patients was part of an extensive study that also included focus groups with physicians and a representative survey of the German population to investigate attitudes toward radiological diagnostics using AI. The study was approved by the Committee for Ethics in Research of the Forschungszentrum Jülich (FRAIM.20221005). Written informed consent was obtained from all participants prior to the empirical study. Data were pseudonymized and stored in a protected digital environment. Participants received €100 (US $117) for their participation.

Results

Overview

A total of 7 focus group sessions were conducted with 34 adults aged between 23 and 85 (mean 49.06, SD 17.08) years. Fifteen participants were female, and 19 were male. A majority (30/34,

XSL•FO
RenderX

88.2%) reported Abitur, the German equivalent of a high school diploma allowing university studies, as their highest level of general education. A total of 4 (11.8%) participants reported completing 10 years of formal education (German "Realschul-Abschluss, Mittlere Reife," or equivalent). The majority of participants (25/34, 75.8%) were employed. Almost all reported having a regular family physician (32/34, 94.1%) and having undergone at least 1 radiological scan (32/33, 97%). The overall score for the 9-item Affinity for Technology in Action Scale indicated a relatively high affinity for technology among the sample (mean 4.14, SD 1.04; Cronbach α=0.94).

The following results highlight the opportunities and risks of using AI in radiological diagnostics from the patients' perspective, as well as the circumstances that can foster its acceptance and trust. Citations of participants' statements are provided with the focus group number (FG) and participant number (#). Original statements have been translated into English for publication.

The following aspects emerged during the focus group discussions: (1) efficiency, data integrity, and quality assurance; (2) data privacy, transparency, and explainability; (3) the role of physicians; (4) trust in physicians versus AI; (5) trust-building functions of physicians; and (6) the physician-patient relationship.

Efficiency, Data Integrity, and Quality Assurance

In general, all participants agreed that the use of AI offers opportunities for more efficient and accurate diagnostics. However, a more efficient workflow and shorter reading times, thanks to AI, were only rated positively if these improvements shortened waiting times for appointments in radiology, enabled diseases to be detected earlier, and gave physicians more time for their patients.

The participants considered the potential of using AI to reduce human errors and achieve a high level of accuracy in diagnostics to be positive, particularly in the case of abnormalities that are difficult to identify:

> *It is clearly an opportunity to diagnose things that would otherwise have been overlooked. Because AI never gets tired, so to speak, and makes fewer mistakes, depending on how it is trained.* [FG3#1]

Nevertheless, many participants expressed concern about misdiagnoses. Above all, they doubted the quality of the data, namely "that AI systems usually

> *have a kind of bias, which is due to training, i.e. what the machine is fed with."* [FG6#2]

In addition, participants feared that the increasing commercialization of the AI market would lead to companies basing their actions purely on economic aspects. This could not only result in insufficient validation of data and discrimination against certain patient groups due to a lack of balance and diversity in the datasets, but also in algorithms being manipulated by companies to suit their own interests:

> *I think the point is that if you end up with this kind of corporate competition, it could naturally lead to the problem of companies wanting to establish themselves that don't really have the know-how or experience. And then you run a bit of a risk because they then make dumping offers to university hospitals or radiology practices, and their software is then used simply because they offer the cheapest price, but it's not good.* [FG4#2]

Careful quality assurance, both in the development and approval of AI algorithms and in their ongoing application, was highlighted as one of the fundamental prerequisites for acceptance (Table 1). Participants emphasized that high data quality, in terms of scope, diversity, and continuous updating, is crucial and that this must be ensured through quality standards and regulation. They also acknowledged that science plays an important role in building trust in the quality of the algorithms. The participants considered it essential that the accuracy of diagnostics using AI is better than without AI:

> *"AI has to be somewhat better at the end of the day, because otherwise it would make no sense at all—at least from the patient's point of view."* [FG4#4]

Some participants found it contradictory to place higher demands on technology than on physicians in order to trust the technology.

> *And what I keep asking myself is that we now always set this requirement of 98, 99 percent for technology, and then the initial question always arises: How high is the accuracy in humans? There are so many errors. So I ask myself why we are so strict with technology. So I don't have an answer to that either. But I've noticed several times now that we simply set much higher standards for technology than we do for ourselves.* [FG5#5]

XSL•FO
RenderX

Table . Framework conditions for the use of artificial intelligence (AI) in radiological diagnostics.

Category	Illustrative quote
Quality standards	• "...That there is also a gold standard for how to set up such an AI or how it is fed with data." (FG6#4)
Validation studies	• "...That I can rely on it. It must be validated." (FG6#6)
Independent certification	• "The software must be independently certified in any case. (FG4#3)
Regulation	• "The government and political organizations should definitely create the legal framework." (FG3#4)
Scientific monitoring	• "That's why it has to be scientifically monitored and constantly questioned: Are the results produced with this technology really still very close to the final result, the disease?" (FG2#3)
Higher accuracy than radiologists	• "I would want to have a high accuracy of 99 or so. Yes, already better than humans. It's not just measured by accuracy. Humans can simply do more than AI as a whole package; they are not limited to the case. They have many, many more abilities, and that's why it's a factor, and I would want a little more than just saying, "Okay, AI is currently surpassing humans." It should be significantly more than just a few percent better." (FG5#1)
Data protection	• "So what would be important to me personally, for example, is that such data cannot be misused. Whether by employers, insurance companies, or third parties such as companies." (FG4#4)
Data security	• "We've seen how hospitals have been hacked, how doctors' offices have been hacked. I'm very, very skeptical about that." (FG3#3)
Freedom of choice	• "And for me, it would be important that I am allowed to decide for myself, as long as I am able to: Will AI be used or not?" (FG6#4)
Transparency	• "How much data has been entered and so on, so that I can better assess it for myself. Someone would have to explain that to me." (FG6#4)
Explainability	• "There is a good reason for using deep learning. The hit rates are the best. And even if I don't understand it technically, I don't care, as long as it works." (FG4#2) • "When AI diagnoses things, it would be nice to know why and on what criteria it bases its diagnosis. Yes, I don't know if I would trust it then." (FG7#5)

Data Privacy, Transparency, and Explainability

Most participants considered existing data protection in Germany to be sufficient for the use of AI in radiology. However, a minority of participants expressed extreme discomfort with the collection and storage of data for the development of AI and its application. These participants stressed the great importance of preserving the freedom of choice.

Ensuring transparency in the use of AI was important to all participants. As technical nonprofessionals, the participants did not feel able to assess the quality of AI in diagnostics. Given the potentially serious consequences of misdiagnosis, the types of training data used should be disclosed, and the accuracy of the algorithm should be specified and clearly communicated. The reputation of the companies that develop the software was considered a good indicator of the software's trustworthiness. Although transparency was seen as a necessary prerequisite for the acceptance of AI use, participants were unsure whether this would help build trust in AI diagnoses if they were personally affected.

Participants expressed contrasting attitudes toward the explainability of AI diagnoses. One group considered the "black box" nature of AI to be unproblematic as long as superior performance was proven. This argument did not convince the roughly equally large group that found the "black box" nature of AI problematic. For these participants, the process by which AI arrives at a result must be reconstructible in order to trust an AI diagnosis, clarify misdiagnoses, and assign responsibility.

The framework conditions that patients demanded for the use of AI in radiological diagnostics are listed in Table 1. The quotes are typical examples of the respective categories.

XSL•FO
RenderX

Role of Physicians

Meeting the technical requirements and framework conditions for the use of AI does not automatically lead to a willingness to choose a radiologist who uses AI.

In a situation where a suspected serious neurological disease, such as multiple sclerosis, needs to be clarified radiologically, all but 3 participants indicated that they would choose a radiology practice that uses AI. However, this preference depended on the interpretation and evaluation of the clinical relevance of abnormalities and the final diagnosis remaining the responsibility of radiologists in consultation with the referring physicians.

With the exception of 1 person, who preferred automated AI in all cases, the participants felt strongly that AI should only be used as a tool in the hands of an experienced physician:

> *The decision on how to interpret this, how relevant it is to the individual case, and how to deal with it must remain with the physician, without restriction.* [FG2#5]

Trust in Physicians Versus AI

The participants considered an independent medical assessment to be important for trusting the AI diagnosis. This was particularly evident in cases where the attending physician doubted the AI's diagnosis.

If abnormalities were found that required treatment, the AI diagnosis did not automatically lead to a willingness to undergo treatment. Only 1 participant mentioned that they would trust the AI diagnosis and follow the treatment. On the other hand, only a minority trusted the physicians' judgment unreservedly in such cases. The majority pursued a safety strategy and wanted to eliminate uncertainty by seeking a third opinion. An AI-supported procedure could also be used again, but in the vast majority of cases, a physician was considered necessary to review the AI result again.

Trust-Building Functions of Physician

One of the reasons for the important role of the physician is the requirement to check the AI output. Many participants demanded not only a plausibility check but also a verification of the AI output through an additional independent assessment. For most participants, the decisive factors were characteristics and skills that were attributed only to physicians and considered indispensable for ensuring well-balanced results based on individual cases (Table 2). In their opinion, physicians have a holistic view of human beings, which enables them to classify the clinical picture in terms of the patient's entire biography and take into account individual factors that may be relevant to the diagnosis.

Table . Trust-building functions of the physician

Category	Illustrative quote
Checks to prevent misdiagnoses	"I see the risk that the machine will still make mistakes and that someone will still have to check whether the diagnosis is actually correct." (FG5#2)
Backup through independent diagnostics	"…On the one hand, you want to secure support through AI, but on the other hand, you cannot rely on it 100%. Instead, the doctor himself must evaluate the images each time and form his own opinion." (FG4#2)
Taking the individuality of the patient into account	"For me, the doctor is still the one who understands the human system as a whole. And AI is just the tin idiot that evaluates data and draws conclusions from it. So AI always has only one focus: please evaluate this image data now. And spit out the results." (FG3#4)
Expanding the diagnostic context through intuition, flexibility, and creativity	"…That something so holistic and inventive is simply not achievable with AI. The creativity that is inherent in humans and that is ultimately needed when making a diagnosis." (FG5#1)

It was also argued that physicians can use their intellectual abilities to deviate from routines. Their intuition, creativity, and flexibility enable them to do so. It is precisely these qualities of physicians that are highly valued for recognizing abnormalities beyond the specific issue at hand and for diagnosing rare diseases.

At the same time, however, participants were skeptical that the trust-building functions of physicians would be realized. In their view, the use of AI could lead to a decline in the competence of radiologists. The greatest risks were seen in physicians' overconfidence in AI output and insufficient attention to, and monitoring of, the results:

> *Of course, this will lead to a decline in quality, because doctors throughout the entire chain will rely on AI, and things will simply be checked off more quickly: AI gives this and that diagnosis, take another look: Yes, it's okay. If the attending physician gets that, he says: Yes, the other doctor has already given his okay, so everything is fine and we'll do it that way.* [FG5#2]

In addition, the participants assumed that the combination of AI and radiologists is only a transitional stage toward automated AI. The participants viewed this development with great concern:

> *Because I see a bit of a timeline there. I think that as soon as AI comes into play, sooner or later it will come down to the third point: automated AI. Because maybe at the beginning you trust the radiologist more because they have more experience. Then AI has proven itself so many times that you think, okay, AI has the final say.* [FG3#6]

XSL•FO
RenderX

Physician-Patient Relationship

It was important to the participants that the social and emotional aspects of the physician-patient relationship were preserved (Table 3). This is not just about communication but rather about experiencing empathy, a reciprocal exchange relationship, and the feeling of being in good hands. It was crucial for participants to know that physicians would continue to take responsibility for their patients even when using AI. It is important to them that physicians make decisions based on the well-being of their patients. For patients, this means that physicians act not only on the basis of their knowledge but also on the basis of values. In addition, physicians should explain to patients what AI is, how it is used, how its results should be evaluated, and, above all, what the added value of an AI-supported diagnosis is.

Most participants were concerned about the threat that the physician-patient relationship might suffer:

> *I can see the danger of this becoming the norm in practices. The AI result is there: bang, let's do it, move on. So the doctor might take less time for a personal consultation and no longer look closely at the patient.* [FG2#2]

They were skeptical that relieving physicians' workloads through greater efficiency using AI would actually result in more time for patients. Rather, they fear that dehumanization and deindividualization in health care will increase.

Table . Trust-building effects of the physician-patient relationship.

Category	Illustrative quote
Creating an interpersonal relationship	"For me, it would be incredibly important to be with the doctor and see how he takes a close look at it again. And it's very, very, very important that the patient feels well cared for at that moment and doesn't just have a piece of paper in their hand." (FG1#1)
Taking responsibility for the patient	"For me, the risk basically lies in a machine making diagnoses based on the facts that my body provides or that the disease shows. But there is more to it than that. Responsibility also plays a role here." (FG1#3)
Making decisions based on moral values	"For me, AI stands for logic. And for me, the doctor stands for conscience or human traits and for hesitation. So not in a way like: one or zero, one or zero." (FG7#3)
Explaining the individual added value of using AI[a]	"I mean, when patients arrive and say, 'Yes, why are you doing that?,' then the doctor must at least be able to explain why he is doing it and what the advantages are, and also what the patient will gain from it." (FG2#4)

[a]AI: artificial intelligence.

Discussion

Principal Findings

The results of this study show that patients are generally open to the use of AI in radiological diagnostics. This is consistent with a number of other studies that demonstrate openness to AI in other medical fields among various population groups [45,46]. Patients recognize the potential for more accurate, faster diagnostics based on expanded medical knowledge, while emphasizing the importance of quality assurance not only during approval but also during ongoing application.

However, the in-depth discussion in this study shows that the advantages mentioned will only increase acceptance of AI if, at the same time, the health care system is set up in such a way that the use of AI leads to real added value for patients. Patient-centered care was one of the most important factors that need to be met for patients to endorse the use of AI. In particular, this means that physicians should use the time saved through the use of AI to provide more personalized care for patients. Rationalization effects and cost savings—such as faster patient flow (ie, more patients in the same amount of time), staff savings, and even the replacement of radiologists by AI—were major reasons why patients would reject its use. Furthermore, the expected increasing commercialization of AI development and application was associated with negative impacts on patient care. These aspects have received too little attention in previous studies.

This study suggests that it is not only technological features—in particular, high performance and reliability—that are decisive for the acceptance of AI; rather, the quality of the entire sociotechnical system plays a role.

This is particularly evident when it comes to the specific willingness to choose a radiologist who uses AI if there is a suspicion of a serious disease that requires radiological clarification. Patients are uncertain in this regard. In such a situation, patients place high demands on the role and function of physicians in particular. As already shown in other studies, it is very important to patients that AI findings are checked by physicians [47,48].

Additionally, this study was able to expand previous findings by showing that what is required is not merely a plausibility check, but rather an independent diagnosis that broadens the diagnostic context. The question is not whether AI is used as a first or second opinion; rather, physicians should contribute specific skills and competencies to the diagnostic process that, from the patients' perspective, AI lacks, such as intuition, flexibility, and creativity. Only then, in the participants' view, will the individuality of each case be given appropriate consideration and abnormalities beyond the specific question for which the AI had been trained will be discovered. The

XSL•FO
RenderX

inability of AI to integrate isolated imaging results within the broader context of the patient's living situation is also discussed in the radiology community as a significant limitation of today's AI systems. It is frequently emphasized that the more nuanced expertise of radiologists is indispensable in the diagnostic process [49].

However, patients are concerned that these very functions of the physician, which promote trust, could be weakened by overreliance on AI results and the increasing deskilling of physicians. This is all the more concerning as the trend toward automated AI is both expected and feared. This contrasts with the fact that radiologists themselves consider the replacement of radiologists unlikely [50].

The results illustrate that patients are highly uncertain about whether they would actually opt for AI diagnostics if they were personally affected. This is also indicated by the finding that patients tend to trust physicians more than AI results if the physician doubts the AI results.

The results of this study show that it is not only the competence of physicians that matters, but also the physician-patient relationship. This result confirms previous findings that the maintenance of the familiar physician-patient relationship is decisive for patients [51,52]. The in-depth discussions in the focus groups clarified which characteristics of the physician-patient relationship are important when using AI. An empathetic relationship and the physician's responsibility to care for the patient's well-being are required. In addition, patients emphasized that physicians should take on an advisory role with regard to the pros and cons of using AI in individual cases. This also includes being able to provide information about the quality of AI. Various studies have shown the importance of transparency in the use of AI and the desire of patients to be informed about whether AI is being used [53,54]. The results of this study reveal, in addition, that physicians, as trusted mediators of information, should play a central role in dispelling concerns and fears. These factors in the physician-patient relationship could build trust in the diagnostic process. Trust can promote acceptance of the use of AI [55,56]. However, AI also has the potential to disrupt person-centered physician-patient relationships [57,58]. This study demonstrated that this is precisely what patients are concerned about when it comes to the use of AI in radiological diagnostics.

The findings of the study, and in particular the important role of the physician-patient relationship, arise within the context of Germany's highly developed health care system, which is characterized by high standards. As a rule, equal access to new technologies is guaranteed, with costs being shared within the system in accordance with the principle of solidarity. Ethical and regulatory considerations are taken into account prior to the implementation of new technologies. In health care systems in other countries that are structured differently, patients might consider other factors to be relevant for acceptance and trust, or weigh them differently.

Strengths and Limitations

This study examined patients' perspectives on the use of AI in radiological diagnostics using an innovative focus group method. Unlike many previous studies, which tended to use a more abstract approach, this study used scenarios as stimuli to gain insights into patients' reasoning regarding the use of AI in situations that affect them personally. This approach expands current research by focusing on patients' motives in their decisions for or against a radiologist who uses AI. A limitation of the study could be that the patients had a high level of education and a strong affinity for technology. Both of these factors could lead to a more positive attitude toward AI compared to the general population [59]. Although several channels were used to reach participants and no prior knowledge of AI was required for participation, self-selection and recruitment via patient organizations and self-help groups may have introduced some selection bias. Therefore, the results may have limited applicability to patients with a low socioeconomic status and low digital and health literacy. However, this does not detract from the validity of the argumentation patterns identified in this qualitative study. Nevertheless, validating the results with a more heterogeneous sample in terms of socioeconomic status and digital literacy might further strengthen the conclusions.

Conclusions

This study shows that patients are generally open to the use of AI in radiological diagnostics, depending on the design of the sociotechnical system. It is not only the performance of AI that is decisive; extensive quality assurance measures are necessary, as is, above all, the integration of AI into the health care system in a way that ensures real added value for patients. At the same time, the study reveals that in cases of personal involvement, the decision to use AI—beyond the perceived abstract advantages and disadvantages—depends significantly on the trust-building role of the physician in the diagnostic process. Consequently, framework conditions for the use of AI are required that avoid constraints preventing physicians from acting independently and degrading them to "messengers of AI." Overconfidence in the AI result and deskilling must be prevented. The integrity of the physician-patient relationship should be preserved. On this basis, patients do not feel "at the mercy" of AI but rather "in good hands" in the diagnostic process guided by the physician. This can help to overcome existing uncertainties. Acceptance and trust can be promoted by strengthening the trust-building role of physicians when AI is used.

Acknowledgments

The author would like to thank the German Federal Ministry of Research, Technology and Space for its financial support of the study. The author would also like to thank all the patients who participated in the study. The author also thanks the coordinator of the project "Beyond Performance: Exploring the Framework for Artificial Intelligence in Brain Medicine" (FRAIM), Prof

XSL•FO
RenderX

Bert Heinrichs, for his support of this study. Special thanks go to Katrin Heyl, MSc, who was involved in recruiting and organizing the focus groups, as well as serving as a coder in the content analysis. The author declares that no artificial intelligence was used at any point in creating the manuscript.

Funding

The empirical study is part of the project “Beyond Performance: Exploring the Framework for Artificial Intelligence in Brain Medicine” (FRAIM) funded by the German Federal Ministry of Research, Technology and Space (grant 01GP2113A). The funder had no involvement in the design of the study, the collection, analysis or interpretation of data, or the writing of the manuscript.

Data Availability

The datasets used and analyzed during this study are unavailable, as they are protected under the General Data Protection Regulation (Article 9).

Authors' Contributions

The author was responsible for defining the concept of the study as an empirical subproject within the project “Beyond Performance: Exploring the Framework for Artificial Intelligence in Brain Medicine” (FRAIM) and contributed to writing the application for funding. The author conducted the study, collected and analyzed the data, and wrote the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Short introduction to artificial intelligence (AI) given to participants at the beginning of the focus group discussions.
[DOCX File, 14 KB - jmir_v28i1e89178_app1.docx]

Checklist 1
COREQ checklist.
[DOCX File, 28 KB - jmir_v28i1e89178_app2.docx]

References

1. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
2. Najjar R. Redefining radiology: a review of artificial intelligence integration in medical imaging. Diagnostics (Basel) 2023 Aug 25;13(17):2760. [doi: 10.3390/diagnostics13172760] [Medline: 37685300]
3. Pinto-Coelho L. How artificial intelligence is shaping medical imaging technology: a survey of innovations and applications. Bioengineering (Basel) 2023 Dec 18;10(12):1435. [doi: 10.3390/bioengineering10121435] [Medline: 38136026]
4. Khalifa M, Albadawy M. AI in diagnostic imaging: revolutionising accuracy and efficiency. Comput Methods Programs Biomed Update 2024;5:100146. [doi: 10.1016/j.cmpbup.2024.100146]
5. Wang H, Zhu H, Ding L. Accurate classification of lung nodules on CT images using the TransUnet. Front Public Health 2022 Dec 5;10:1060798. [doi: 10.3389/fpubh.2022.1060798]
6. Nam JG, Hwang EJ, Kim J, et al. AI improves nodule detection on chest radiographs in a health screening population: a randomized controlled trial. Radiology 2023 Apr;307(2):e221894. [doi: 10.1148/radiol.221894] [Medline: 36749213]
7. Nabizadeh F, Masrouri S, Ramezannezhad E, et al. Artificial intelligence in the diagnosis of multiple sclerosis: a systematic review. Mult Scler Relat Disord 2022 Mar;59:103673. [doi: 10.1016/j.msard.2022.103673] [Medline: 35180619]
8. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
9. Younger J, Morris E, Arnold N, Athulathmudali C, Pinidiyapathirage J, MacAskill W. A systematic review of comparisons of AI and radiologists in the diagnosis of HCC in multiphase CT: implications for practice. Jpn J Radiol 2026 Jan;44(1):97-105. [doi: 10.1007/s11604-025-01853-y] [Medline: 40824347]
10. Becker J, Decker JA, Römmele C, et al. Artificial intelligence-based detection of pneumonia in chest radiographs. Diagnostics (Basel) 2022 Jun 14;12(6):1465. [doi: 10.3390/diagnostics12061465] [Medline: 35741276]
11. Aljuhani M, Ashraf A, Edison P. Use of artificial intelligence in imaging dementia. Cells 2024 Nov 27;13(23):1965. [doi: 10.3390/cells13231965] [Medline: 39682713]
12. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: implications for clinical decision-making. PLOS Digit Health 2024 Nov;3(11):e0000651. [doi: 10.1371/journal.pdig.0000651] [Medline: 39509461]

XSL•FO
RenderX

13. Hasanzadeh F, Josephson CB, Waters G, Adedinsewo D, Azizi Z, White JA. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. NPJ Digit Med 2025 Mar 11;8(1):154. [doi: 10.1038/s41746-025-01503-7] [Medline: 40069303]
14. Koçak B, Ponsiglione A, Stanzione A, et al. Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. Diagn Interv Radiol 2025 Mar 3;31(2):75-88. [doi: 10.4274/dir.2024.242854] [Medline: 38953330]
15. Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P. Transparency of deep neural networks for medical image analysis: a review of interpretability methods. Comput Biol Med 2022 Jan;140:105111. [doi: 10.1016/j.compbiomed.2021.105111] [Medline: 34891095]
16. Hassija V, Chamola V, Mahapatra A, et al. Interpreting black-box models: a review on explainable artificial intelligence. Cogn Comput 2024 Jan;16(1):45-74. [doi: 10.1007/s12559-023-10179-8]
17. Dratsch T, Chen X, Rezazade Mehrizi M, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. Radiology 2023 May;307(4):e222176. [doi: 10.1148/radiol.222176] [Medline: 37129490]
18. Ling Kuo RY, Freethy A, Smith J, et al. Stakeholder perspectives towards diagnostic artificial intelligence: a co-produced qualitative evidence synthesis. EClinicalMedicine 2024;71:102555. [doi: 10.1016/j.eclinm.2024.102555] [Medline: 38549586]
19. Carter SM, Carolan L, Saint James Aquino Y, et al. Australian women's judgements about using artificial intelligence to read mammograms in breast cancer screening. Digit Health 2023;9:20552076231191057. [doi: 10.1177/20552076231191057] [Medline: 37559826]
20. Fukuyama F. Trust: The Social Virtues and the Creation of Prosperity: Free Press; 1995.
21. Wang S. Public perceptions of artificial intelligence in 20 countries: assessing individual- and country-level factors. Cross Cult Res 2025 Dec;59(5):651-676. [doi: 10.1177/10693971251336803]
22. Howell B. WEIRD? Institutions and consumers' perceptions of artificial intelligence in 31 countries. AI Soc 2025 Aug;40(6):4409-4431. [doi: 10.1007/s00146-025-02217-w]
23. Gatting L, Ahmed S, Meccheri P, Newlands R, Kehagia AA, Waller J. Acceptability of artificial intelligence in breast screening: focus groups with the screening-eligible population in England. BMJ Public Health 2024 Dec;2(2):e000892. [doi: 10.1136/bmjph-2024-000892] [Medline: 40018529]
24. Nelson CA, Pérez-Chada LM, Creadore A, et al. Patient perspectives on the use of artificial intelligence for skin cancer screening: a qualitative study. JAMA Dermatol 2020 May 1;156(5):501-512. [doi: 10.1001/jamadermatol.2019.5014] [Medline: 32159733]
25. Esmaeilzadeh P. Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives. BMC Med Inform Decis Mak 2020 Jul 22;20(1):170. [doi: 10.1186/s12911-020-01191-1] [Medline: 32698869]
26. Tang L, Li J, Fantus S. Medical artificial intelligence ethics: a systematic review of empirical studies. Digit Health 2023;9:20552076231186064. [doi: 10.1177/20552076231186064] [Medline: 37434728]
27. Young AT, Amara D, Bhattacharya A, Wei ML. Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. Lancet Digit Health 2021 Sep;3(9):e599-e611. [doi: 10.1016/S2589-7500(21)00132-1] [Medline: 34446266]
28. Osnat B. Patient perspectives on artificial intelligence in healthcare: a global scoping review of benefits, ethical concerns, and implementation strategies. Int J Med Inform 2025 Nov;203:106007. [doi: 10.1016/j.ijmedinf.2025.106007] [Medline: 40494217]
29. Jutzi TB, Krieghoff-Henning EI, Holland-Letz T, et al. Artificial intelligence in skin cancer diagnostics: the patients' perspective. Front Med (Lausanne) 2020;7:233. [doi: 10.3389/fmed.2020.00233] [Medline: 32671078]
30. Ozcan BB, Dogan BE, Xi Y, Knippa EE. Patient perception of artificial intelligence use in interpretation of screening mammograms: a survey study. Radiol Imaging Cancer 2025 May;7(3):e240290. [doi: 10.1148/rycan.240290] [Medline: 40249272]
31. Rodler S, Kopliku R, Ulrich D, et al. Patients' trust in artificial intelligence-based decision-making for localized prostate cancer: results from a prospective trial. Eur Urol Focus 2024 Jul;10(4):654-661. [doi: 10.1016/j.euf.2023.10.020] [Medline: 37923632]
32. Fransen SJ, Kwee TC, Rouw D, et al. Patient perspectives on the use of artificial intelligence in prostate cancer diagnosis on MRI. Eur Radiol 2025 Feb;35(2):769-775. [doi: 10.1007/s00330-024-11012-y] [Medline: 39143247]
33. Miró Catalina Q, Femenia J, Fuster-Casanovas A, et al. Knowledge and perception of the use of AI and its implementation in the field of radiology: cross-sectional study. J Med Internet Res 2023 Oct 13;25:e50728. [doi: 10.2196/50728] [Medline: 37831495]
34. Akingbola A, Adeleke O, Idris A, Adewole O, Adegbesan A. Artificial intelligence and the dehumanization of patient care. J Med Surg Public Health 2024 Aug;3:100138. [doi: 10.1016/j.glmedi.2024.100138]
35. Willig C, Stainton-Rogers W. The SAGE Handbook of Qualitative Research in Psychology, 2nd edition: SAGE Publications; 2017.

XSL•FO
RenderX

36. Ryan KE, Gandha T, Culbertson MJ, Carlson C. Focus group evidence: implications for design and analysis. Am J Eval ;35(3):328-345. [doi: 10.1177/1098214013508300]
37. Morgan DL. Focus groups and social interaction. In: Gubrium JF, Holstein JA, Marvasti AB, McKinney KD, editors. The Sage Handbook of Interview Research: The Complexity of the Craft, 2nd edition: SAGE Publications, Inc; 2012:161-176. [doi: 10.4135/9781452218403.n11]
38. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
39. Robinson RS. Purposive sampling. In: Michalos AC, editor. Encyclopedia of Quality of Life and Well-Being Research: Springer; 2014:5243-5245. [doi: 10.1007/978-94-007-0753-5_2337]
40. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: a systematic review of empirical tests. Soc Sci Med 2022 Jan;292:114523. [doi: 10.1016/j.socscimed.2021.114523] [Medline: 34785096]
41. Franke T, Attig C, Wessel D. A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale. Int J Hum Comput Interact 2019;35(6):456-467. [doi: 10.1080/10447318.2018.1456150]
42. Morgan DL, Krueger RA, King JA. The Focus Group Guidebook: SAGE; 1998.
43. Mayring P. Qualitative content analysis theoretical background, recent developments and software solutions. : University of Klagenfurt; 2014 URL: https://tinyurl.com/m75xxr22 [accessed 2026-05-12]
44. Ritchie J, Spencer L, O'Connor W. Carrying out qualitative analysis. In: Ritchie J, Lewis J, editors. Qualitative Research Practice: A Guide for Social Science Students and Researchers: Sage Publications; 2003:219-261.
45. Scott IA, Carter SM, Coiera E. Exploring stakeholder attitudes towards AI in clinical practice. BMJ Health Care Inform 2021 Dec;28(1):e100450. [doi: 10.1136/bmjhci-2021-100450] [Medline: 34887331]
46. Hogg HDJ, Al-Zubaidy M, Technology Enhanced Macular Services Study Reference Group, et al. Stakeholder perspectives of clinical artificial intelligence implementation: systematic review of qualitative evidence. J Med Internet Res 2023 Jan 10;25:e39742. [doi: 10.2196/39742] [Medline: 36626192]
47. Riedl R, Hogeterp SA, Reuter M. Do patients prefer a human doctor, artificial intelligence, or a blend, and is this preference dependent on medical discipline? Empirical evidence and implications for medical practice. Front Psychol 2024;15:1422177. [doi: 10.3389/fpsyg.2024.1422177] [Medline: 39188871]
48. Lennartz S, Dratsch T, Zopfs D, et al. Use and control of artificial intelligence in patients across the medical workflow: single-center questionnaire study of patient perspectives. J Med Internet Res 2021 Feb 17;23(2):e24221. [doi: 10.2196/24221] [Medline: 33595451]
49. Katal S, York B, Gholamrezanezhad A. AI in radiology: from promise to practice - a guide to effective integration. Eur J Radiol 2024 Dec;181:111798. [doi: 10.1016/j.ejrad.2024.111798] [Medline: 39471551]
50. Yang L, Ene IC, Arabi Belaghi R, Koff D, Stein N, Santaguida PL. Stakeholders' perspectives on the future of artificial intelligence in radiology: a scoping review. Eur Radiol 2022 Mar;32(3):1477-1495. [doi: 10.1007/s00330-021-08214-z] [Medline: 34545445]
51. Vo V, Chen G, Aquino YSJ, Carter SM, Do QN, Woode ME. Multi-stakeholder preferences for the use of artificial intelligence in healthcare: a systematic review and thematic analysis. Soc Sci Med 2023 Dec;338:116357. [doi: 10.1016/j.socscimed.2023.116357] [Medline: 37949020]
52. Ibba S, Tancredi C, Fantesini A, et al. How do patients perceive the AI-radiologists interaction? Results of a survey on 2119 responders. Eur J Radiol 2023 Aug;165:110917. [doi: 10.1016/j.ejrad.2023.110917] [Medline: 37327548]
53. Kühne S, Jacobsen J, Legewie N, Dollmann J. Attitudes toward AI usage in patient health care: evidence from a population survey vignette experiment. J Med Internet Res 2025 May 27;27:e70179. [doi: 10.2196/70179] [Medline: 40424613]
54. McGhee KN, Barrett DJ, Safarini O, et al. Patient preferences for artificial intelligence in medical imaging: a single-center cross-sectional survey. J Imaging Inform Med 2026 Apr;39(2):1100-1112. [doi: 10.1007/s10278-025-01629-w] [Medline: 40775178]
55. Afroogh S, Akbari A, Malone E, Kargar M, Alambeigi H. Trust in AI: progress, challenges, and future directions. Humanit Soc Sci Commun 2024 Nov;11(1):1568. [doi: 10.1057/s41599-024-04044-8]
56. Evans RP, Bryant LD, Russell G, Absolom K. Trust and acceptability of data-driven clinical recommendations in everyday practice: a scoping review. Int J Med Inform 2024 Mar;183:105342. [doi: 10.1016/j.ijmedinf.2024.105342] [Medline: 38266426]
57. Sauerbrei A, Kerasidou A, Lucivero F, Hallowell N. The impact of artificial intelligence on the person-centred, doctor-patient relationship: some problems and solutions. BMC Med Inform Decis Mak 2023 Apr 20;23(1):73. [doi: 10.1186/s12911-023-02162-y] [Medline: 37081503]
58. Zondag AGM, Rozestraten R, Grimmelikhuijsen SG, et al. The effect of artificial intelligence on patient-physician trust: cross-sectional vignette study. J Med Internet Res 2024 May 28;26:e50853. [doi: 10.2196/50853] [Medline: 38805702]
59. Busch F, Hoffmann L, Xu L, et al. Multinational attitudes toward AI in health care and diagnostics among hospital patients. JAMA Netw Open 2025 Jun 2;8(6):e2514452. [doi: 10.1001/jamanetworkopen.2025.14452] [Medline: 40493367]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research
MRI: magnetic resonance imaging

Edited by A Stone; submitted 08.Dec.2025; peer-reviewed by EM Schomakers, M Galani; revised version received 30.Mar.2026; accepted 30.Mar.2026; published 25.May.2026.

Please cite as:
Karger CR
Patients' Perspectives on the Implementation of AI in Radiological Diagnostics: Focus Group Study
J Med Internet Res 2026;28:e89178
URL: https://www.jmir.org/2026/1/e89178
doi:10.2196/89178

© Cornelia R Karger. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Artificial Intelligence in Patient-Centered Care and Macro-, Meso-, and Micro-Level Determinants of Rehumanization and Dehumanization: Qualitative Interview Study

Dora Horvath*, PhD; Noemi Szilvia Lorincz*, PhD

Department of Strategic Management, Corvinus University of Budapest, Fovam Square 8, Budapest, Hungary

*all authors contributed equally

Corresponding Author:
Dora Horvath, PhD
Department of Strategic Management, Corvinus University of Budapest, Fovam Square 8, Budapest, Hungary

Abstract

Background: Patient-centered care remains a foundational principle of modern health care. The digital transformation of health systems has accelerated the adoption of artificial intelligence (AI) across diagnostic, predictive, and communicative functions, with implications for efficiency and clinical workflows. At the same time, AI integration raises concerns regarding transparency, equity, accountability, and trust, positioning it as a potential driver of both rehumanizing and dehumanizing dynamics in health care practice.

Objective: This study examines how the adoption of AI in health care may influence patient-centered care, exploring its potential to promote rehumanization or contribute to dehumanization. The objective is to identify the factors that shape these outcomes at the macrolevel (policy and infrastructure), mesolevel (institutional practices), and microlevel (individual behaviors and interactions).

Methods: This study adopts an exploratory qualitative design informed by grounded theory principles, drawing on 20 semistructured interviews with health care leaders, clinicians, researchers, legal experts, and industry consultants who have substantial professional experience across European health care systems, with some participants also contributing experience from the US health care context. To enhance analytical rigor and transparency, the study applied the Gioia methodology, enabling inductive coding from first-order concepts to second-order themes and aggregate dimensions. This multistakeholder approach facilitated a nuanced examination of how AI integration is perceived and experienced across macro-, meso-, and microlevels of health care.

Results: The analysis identified key system-level factors shaping rehumanizing or dehumanizing outcomes of AI integration. At the macrolevel, 8 factors—including regulatory frameworks, policy priorities, and infrastructure—were identified as influencing whether efficiency pressures outweigh patient-centered values. At the mesolevel, 5 factors related to institutional strategies, workflows, and leadership shape how AI tools are embedded into care delivery. At the microlevel, 7 factors related to individual behaviors, trust, and doctor-patient interaction dynamics influence whether AI supports empathy and engagement or diminishes them. Rehumanizing potentials include reduced administrative burden, improved care pathways, clearer health communication, and enhanced decision-making, while risks include shorter consultations, reduced empathy, overreliance on automation, and erosion of professional identity. Without deliberate alignment with patient-centered principles, efficiency gains risk undermining the human dimensions of care.

Conclusions: This study represents one of the first empirical examinations of how AI shapes health care practices through rehumanizing and dehumanizing dynamics. The findings demonstrate that outcomes depend not only on technical capabilities but also on regulatory frameworks, institutional strategies, and cultural adaptation. By systematically mapping influencing factors across macro-, meso-, and microlevels, the research provides actionable insights for decision-makers to ensure that efficiency gains remain aligned with patient-centered principles. Realizing AI's promise requires coordinated action to preserve empathy, trust, and interpersonal connection, ensuring that innovation strengthens rather than weakens the human dimensions of care.

(J Med Internet Res 2026;28:e82774) doi:10.2196/82774

KEYWORDS

artificial intelligence; AI integration; qualitative research; digital transformation; patient-centered care

XSL•FO
RenderX

Introduction

AI and the Human Dimension of Patient-Centered Care

Patient-centered care, grounded in respect, empathy, effective communication, and shared decision-making, remains a cornerstone of modern health care [1]. At its core, patient-centered care reflects a broader effort to humanize care by recognizing patients as whole persons rather than as diagnoses or data points [2,3]. At the same time, health care systems are undergoing rapid digital transformation, with artificial intelligence (AI) emerging as a central driver of change. AI applications in clinical decision support, diagnostics, predictive analytics, personalized treatment planning, and patient communication promise improvements in accuracy, efficiency, and resource allocation [4,5]. By reducing administrative burdens and supporting clinical reasoning, AI may indirectly strengthen patient-centered care by enabling clinicians to devote more time to meaningful patient engagement [6].

However, the expanding role of AI has also intensified concerns about potential dehumanization in health care. Humanization refers to care that acknowledges patients' lived experiences, dignity, and agency, whereas dehumanization involves processes—often unintended—through which individuals are reduced to objects, categories, or data [7,8]. Although AI holds significant promise, its integration raises critical ethical and organizational challenges, including algorithmic bias, limited transparency, data quality concerns, regulatory uncertainty, and shifting accountability structures [9]. If inadequately addressed, these challenges may alter professional roles and patient-provider relationships in ways that risk weakening empathy, trust, and relational continuity [10].

AI integration thus represents a pivotal juncture: it may contribute either to rehumanization—by enhancing communication, personalization, and access—or to dehumanization, particularly if it replaces rather than supports interpersonal interaction [11-14]. Importantly, these trajectories are shaped not only by technological capabilities but also by contextual conditions at the macrolevel (policy and regulatory), mesolevel (organizational), and microlevel (individual and relational) [15]. While conceptual debates on AI and human-centered care are growing, empirical research examining how these multilevel factors interact in practice remains limited.

To address this gap, this study explores how AI implementation in health care shapes processes of rehumanization and dehumanization, and which macro-, meso-, and microlevel factors influence these dynamics. Drawing on 20 semistructured interviews with experts across the health care sector and applying an inductive qualitative approach using the Gioia methodology, the study provides a multilevel analysis of how AI is perceived, interpreted, and experienced in relation to the human dimension of care. By integrating perspectives from policy, organizational, and clinical contexts, it aims to generate empirically grounded insights into the conditions under which AI may reinforce—or undermine—patient-centered health care.

Literature Background: Rehumanization and Dehumanization in Health Care

Using the framework established by Todres et al [2], the humanization of health care is conceptualized as a care approach that recognizes patients as whole individuals. This perspective emphasizes the importance of acknowledging their lived experiences, dignity, agency, relationships, and embodied existence, rather than concentrating exclusively on disease or technical procedures. Dehumanization in health care is often an unintentional process in which patients are treated primarily as diagnoses, objects, or data, with reduced attention to their mental states and subjective experiences [8]. In this sense, it can function as an emotion regulation strategy for health care professionals under sustained stress, including in situations involving stigma, while also risking diminished patient autonomy and reduced care quality [8,16]. These concepts offer a valuable framework for analyzing how current changes in health care may enhance or diminish the human aspect of care.

To elucidate this analytical perspective, it is crucial to define patient-centered care, which implements many of these humanizing principles in clinical practice. This study conceptualizes patient-centered care according to the approach proposed by Rathert et al [3], which highlights care that respects patients' values and preferences, responds to their needs, and fosters active patient involvement through effective communication, emotional support, and shared decision-making. These principles also align with the logic of 4P medicine, which emphasizes personalized, predictive, preventive, and participatory approaches to care. Rather than relying solely on syndrome-based treatment models, the 4P paradigm foregrounds early diagnosis, targeted intervention, illness prevention, and patient engagement, thereby reinforcing a holistic and proactive understanding of health and care [17].

Furthermore, as health care is increasingly shaped by digital technologies, these inquiries take on a distinct ethical dimension, as transparent policies, digital literacy, and ethical data governance are essential for public trust in AI-driven health care [18,19]. According to Floridi [20], digital ethics is defined as a practice-oriented framework that addresses the integration of ethical ideals, such as human dignity and accountability, in the design and implementation of digital technologies, rather than merely emphasizing their technical performance. This perspective corresponds with value-sensitive design in human-computer interaction, highlighting that AI systems are not value-neutral but incorporate ethical ideals such as dignity, autonomy, and accountability in their design, implementation, and usage [21].

Within this conceptual framework, the increasing integration of AI represents a significant development. While AI offers important opportunities to improve efficiency and precision in health care, it also raises questions about its potential to either rehumanize or dehumanize care. The growing emphasis on humanizing health care reflects the recognition that effective care involves not only clinical treatment but also empathy, trust, and meaningful relationships. Patients value personalized, compassionate care, while health care professionals seek supportive environments that enable well-being, professional

XSL•FO
RenderX

development, and high-quality service [22]. Regarding rehumanization, a literature review by Busch et al [23] identified that an empathetic and respectful attitude toward patients, adequate human and material resources, and a balanced workload for health care providers are essential conditions for fostering meaningful, reciprocal patient relationships and delivering humanized care. In this context, the integration of AI has the potential to contribute to rehumanization by relieving clinicians of routine administrative tasks and enhancing diagnostic accuracy. These advancements can free up time for health care professionals to engage more deeply with patients, fostering communication, empathy, and trust. This renewed focus on the human dimension of care has the potential to improve both patient experiences and clinical outcomes [12]. Building on this perspective, AI-powered tools have been shown to support personalized communication by tailoring information and support to individual patient needs and preferences [11]. Additionally, AI holds promise for promoting health equity by improving access to care for underserved populations, thereby addressing systemic disparities in health care delivery [24]. Moreover, natural language processing and large language models can help translate complex medical terminology into more understandable information for patients, improving communication and supporting patient engagement and outcomes [25].

However, dehumanization has become a central concern in critiques of modern medicine, often linked to declining personal attention, emotional support, and relational care. Increasing reliance on technology, standardization, and efficiency may overshadow patients' individuality and autonomy, reducing their role in care processes [26]. This shift has implications for both patients and health care providers. Health care providers may adopt dehumanizing practices as a means of emotional self-protection in the face of constant exposure to suffering and stress. While such strategies can enhance clinical performance by enabling psychological distance, they may also compromise empathy, communication, and trust, ultimately threatening the therapeutic alliance and the overall quality of care [8]. While AI can improve efficiency and diagnostic precision, its expanding role may reduce face-to-face interaction and risk treating patients as data points. Such dehumanization may negatively affect patient satisfaction, well-being, and the therapeutic alliance, highlighting the need to ensure that AI is designed and implemented in ways that support rather than replace the human elements of medical practice [27]. A nationally representative US survey found that although the public was generally optimistic about AI in diagnosis and treatment, concerns remained regarding misdiagnosis, privacy, costs, and reduced clinician-patient interaction. These concerns were particularly pronounced among racial and ethnic minority groups, highlighting the perceived risk of dehumanization in AI-driven health care [13]. Rony et al [14] highlight that concerns around dehumanization in health care often stem less from the use of AI itself and more from the potential erosion of personal connection in care. Their findings underscore the importance of integrating AI in ways that support, rather than undermine, empathetic and patient-centered care. Health care professionals, particularly nurses, are seen as essential in preserving compassion within increasingly digital environments, positioning AI as a tool that should enhance—not replace—human interaction [14].

As AI becomes increasingly integrated into health care, questions arise about its impact on the human dimension of care. While it offers substantial benefits, concerns remain that it may weaken empathetic, patient-centered interactions. Understanding whether AI contributes to the rehumanization or dehumanization of care requires examining not only the technology itself but also the broader context of its implementation across macrolevel (policy and infrastructure), mesolevel (institutional practices), and microlevel (individual behaviors and interactions). This multilevel perspective helps identify key enablers and barriers to humanized care in digital health care environments. According to Rosengren et al [15], from a person-centered care perspective, macrolevel policies and financing structures influence the prioritization of patient values and equity, mesolevel organizational cultures and care models dictate the implementation of these principles in practice, and microlevel clinician-patient interactions ultimately shape patients' lived experiences of care. These multilevel dynamics become particularly visible in digital health contexts, where technological change intersects with existing system structures. In this regard, Asthana et al [28] assert that the implementation of eHealth innovations within the National Health Service in England is influenced by the interaction of macrolevel national digital policies, mesolevel organizational capacity and varying digital maturity, and microlevel acceptance by clinicians and patients. These multilevel challenges lead to ongoing inequalities and tensions in person-centered care and routine clinical interactions.

Expanding this viewpoint beyond digitalization, prior studies indicate that macro-meso-micro frameworks are essential for comprehending complex health care phenomena in a broader context. For example, Sawatzky et al [29] examined how response shift—changes in patients' internal standards or values over time—affects the interpretation of patient-reported outcome measures. Their study highlighted how such shifts can influence clinical decisions at the microlevel, quality monitoring and accreditation at the mesolevel, and health policy and funding decisions at the macrolevel. Their multilevel framework illustrates how patient perceptions must be contextualized across different layers of health care to ensure valid, person-centered decision-making [29]. Similarly, Beirão et al [30] examined value creation in Portugal's national electronic health record ecosystem, focusing on interactions among actors at the micro-, meso-, and macrolevels. Their findings show that although resource access, communication, and governance differ across levels, they remain interdependent, highlighting the need for coordinated system-wide efforts to improve patient experience and ecosystem viability [30,31]. Building on macro-meso-micro frameworks in digital health, this study extends prior research by examining how AI implementation influences rehumanization and dehumanization across system levels.

XSL•FO
RenderX

Methods

Research Design and Analytical Approach

The objective of this research is to examine the impact of AI implementation on the potential rehumanization or dehumanization of patient interactions, with a focus on the contextual factors that may influence these outcomes. Specifically, the study seeks to identify and analyze macrolevel (eg, health policy and technological infrastructure), mesolevel (eg, organizational strategy and culture), and microlevel (eg, individual attitudes and clinician-patient communication) determinants that mediate the relationship between AI integration and humanized care delivery. The research adopts a multilevel perspective with the aim of providing a nuanced understanding of how AI can be aligned with the principles of person-centered health care. In accordance with this purpose, the study examines the subsequent research questions:

- RQ1: How does the implementation of AI in health care shape the rehumanization or dehumanization of patient-provider interactions, and through which contextual factors do these effects emerge?
- RQ2: What macrolevel (eg, policy, regulatory, and infrastructural), mesolevel (eg, organizational strategies and cultures), and microlevel (eg, professional attitudes and clinician-patient communication) factors influence these processes?

To address the research questions, a qualitative study was conducted using semistructured interviews with a diverse group of experts from the health care sector. The objective of this study was to obtain an in-depth and comprehensive understanding of perspectives on the integration of AI in health care. Since this is a comparatively novel and developing field, a qualitative approach is especially pertinent, as it facilitates the exploration of emerging experiences, perceptions, and systemic dynamics that are not yet comprehensively understood or readily captured by quantitative methods [32].

Building on this rationale, this study used a grounded theory method due to its exploratory nature and the limited empirical research on AI's role in the rehumanization or dehumanization of patient-provider interactions. This methodology provides a methodical approach to deriving theoretical ideas directly from the data, rendering it especially appropriate for environments with limited established frameworks. Grounded theory, as initially described by Glaser and Strauss [33], aims to construct a thorough, data-driven comprehension of a phenomenon via recurrent cycles of data collection and analysis. Strauss and Corbin [34] assert that the primary objective of this methodology is to generate conceptual elucidations derived from the empirical data, facilitating the concurrent evolution of the theoretical framework alongside the research process. This study used an inductive approach to identify key categories and establish theoretical connections when patterns emerged from the qualitative data [35].

Consistent with grounded theory principles, data collection and analysis occurred simultaneously, with insights from first interviews informing later inquiries. Theoretical saturation was considered achieved when no new categories or significant qualities arose from further interviews and when the evolving themes exhibited adequate consistency and explanatory power across cases. To improve the transparency and rigor of the analytical process, the study used the Gioia technique [36], which offers a systematic framework for grouping first-order concepts, second-order themes, and emergent aggregate dimensions. This integration of iterative inquiry and systematic coding guaranteed that the theoretical contributions were firmly grounded in participants' narratives while maintaining rigorous qualitative research standards. The reporting of this qualitative study was guided by the COREQ (Consolidated Criteria for Reporting Qualitative Research) 32-item checklist to enhance transparency and completeness (Checklist 1).

Data Collection

To identify appropriate participants for the research, purposive sampling was used, which is a widely used technique in qualitative research. The selection of individuals capable of providing profound, contextually grounded insights into the phenomenon under study is a fundamental aspect of this approach, and the recruitment process therefore followed a systematic procedure. Participants were identified through professional networks, health care institutions, and expert organizations across both the public and private sectors. Potential interviewees were contacted directly via email based on their documented expertise, professional roles, and involvement in AI-related health care activities. Recruitment and scheduling were coordinated individually, with interviews conducted either at participants' workplaces or via secure online platforms, depending on their preference. The emphasis was on ensuring the relevance and depth of the data collected [31].

In this study, participants were selected based on their expertise in AI and health care, their availability and willingness to participate, and their ability to articulate their perspectives in a reflective and meaningful way. To address the research questions and examine the macro-, meso-, and microlevel factors influencing the processes of rehumanization and dehumanization in health care, this study used a qualitative research design based on 20 semistructured interviews. It was imperative that interviewees possessed both theoretical knowledge and practical experience with AI in health care. Of particular importance was the inclusion of stakeholders from both private and public health care institutions, ensuring that the research captured the complexity and diversity of the sector. While the participants were based in Hungary, their professional backgrounds included substantial international experience in European health care systems and, in a few cases, experience in US health care, which contributed to the richness of the perspectives that were gathered. This sampling process, consistent with grounded theory's iterative and theoretically driven logic [36], ensured the inclusion of information-rich experts whose insights could meaningfully inform the emerging conceptual understanding of AI implementation. All interviews were conducted jointly by both authors, who also collaborated closely throughout the data analysis process, including coding, theme development, and interpretation.

XSL•FO
RenderX

Prior to conducting the interviews, an interview guide was developed. This guide delineated the primary research questions and thematic areas to be explored, as detailed in Multimedia Appendix 1, where the questions are categorized by macro-, meso-, and microlevels of study. The questions were customized for each participant's professional background and expertise to ensure the discussion accurately represented their personal experience and domain knowledge. As posited by Holloway and Galvin [37], such a guide proffers a structured framework to support the researcher while allowing for flexibility; it is not intended to be followed rigidly, but rather to facilitate the collection of comparable information across participants while enabling in-depth exploration of the research questions. The interviews were conducted between July and August 2025. It is imperative to note that all interviews were audio-recorded with the informed consent of the participants, who were thoroughly briefed on the purpose of the research and the procedures for data handling.

Ethical Considerations

This research was performed in compliance with recognized ethical standards for studies involving human participants. Ethical approval was secured by the Research Ethics Committee of Corvinus University of Budapest, Vice-Rector for Faculty and Research (approval number: KRH/173/2025, dated June 30, 2025). The committee evaluated the study design, data collection methodologies, and data management strategy of the research.

All participants were provided with comprehensive information regarding the study's objectives, the voluntary nature of their involvement, the interview protocols, and the planned application of the findings. Informed written consent was acquired from all participants prior to their involvement, encompassing permission for audio recording. Participants were notified of their entitlement to withdraw from the study at any moment without justification and without facing any negative repercussions.

Participant privacy and confidentiality were meticulously safeguarded during the research procedure. Interview data were processed in an anonymous and nonidentifiable manner. Personal identifications were eliminated during transcribing, and participants were allocated distinct codes to ensure anonymity. Audio recordings and transcripts were securely saved and accessible exclusively by the study team. Results are presented in an aggregated format to guarantee the anonymity of individual participants. Participants did not receive any financial or nonfinancial compensation for their participation in the study.

Data Analysis

Using the methodological framework of Gioia [36], a structured analysis was carried out to understand and generalize the empirical data. This methodology facilitated the recognition of first- and second-order concepts and the formulation of aggregate dimensions that encapsulate the interconnections among nascent themes [36,38]. The analysis was consistently informed by the core principles of grounded theory articulated by Glaser and Strauss [33], ensuring that the evolving theoretical findings were securely rooted in participants' narratives.

The analysis commenced with open coding of the interview transcripts using NVivo software (Lumivero). The initial coding closely adhered to the terminology used by the experts, encapsulating their perspectives on AI's capacity to rehumanize care, its potential for dehumanization, and the contextual elements influencing both outcomes.

In second-order coding, these initial categories were subsequently analyzed, developed, and interpreted using pertinent theoretical frameworks, such as patient-centered care, sociotechnical systems theory, and organizational transformation within health care. This phase facilitated the transition from descriptive findings to the development of profound conceptual patterns elucidating how macro-, meso-, and microlevel factors influence the human impact of AI in clinical practice.

The second-order themes were synthesized into aggregate dimensions through iterative comparison and theoretical elaboration, highlighting the dynamic interplay among structural variables, institutional determinants, and human experiences. In total, the analysis resulted in the identification of 8 macrolevel, 5 mesolevel, and 7 microlevel factors that collectively shape the potential for AI to foster either rehumanizing or dehumanizing trajectories in health care. These dimensions establish the foundation of a preliminary conceptual model that demonstrates how AI adoption might result in either rehumanizing or dehumanizing results, along with the mechanisms by which these processes occur within health care systems.

Results

Participant Characteristics and Overview of Findings

A total of 20 semistructured interviews were conducted with stakeholders representing multiple segments of the health care ecosystem. The sample included physicians and clinical specialists, several of whom also held leadership or managerial roles, alongside participants with backgrounds in health care management and policy, health economics and academic research, legal and regulatory fields, and industry and AI development. These overlapping professional roles provided a multistakeholder perspective on the implementation of AI in health care. Participants represented both public and private health care contexts and varied in their level of AI involvement, ranging from extensive implementation or development experience to indirect exposure or no direct use. The sample included participants identifying as women and men. Gender was recorded as a self-reported descriptive characteristic and was not examined as an analytical variable. Interview duration ranged from 45 to 67 minutes. Across stakeholder groups, the core themes regarding the rehumanizing and dehumanizing potential of AI were largely consistent, although participants emphasized different aspects depending on their professional background. Detailed participant characteristics are provided in Multimedia Appendix 2.

The results section presents the interview findings on how AI integration may contribute to rehumanization or dehumanization

XSL•FO
RenderX

in health care. In total, the analysis identified 8 macrolevel, 5 mesolevel, and 7 microlevel factors that collectively shape these trajectories, highlighting how policy environments, organizational practices, and interpersonal dynamics interact in determining AI's human impact.

Rehumanization Aspects of AI in Health Care

Across interviews, participants emphasized that the quality of patient interaction, effective patient management, and clear communication are central to clinical outcomes and patient satisfaction. Several interviewees described the doctor-patient relationship as a cornerstone of patient-centered care, while noting that growing administrative burdens and fragmented information systems increasingly limit the time available for meaningful interpersonal engagement.

> *Currently, in healthcare, I see that doctors spend a lot of time trying to interpret medical documents or analyze accumulated data. Whenever a patient arrives, they spend less time with them and more time reviewing their medical history.* [Participant 1]

In response to these challenges, participants frequently framed the rehumanization of health care as an important goal, emphasizing the need to restore the relational and ethical core of medical practice. AI was often described as a "double-edged sword": while poorly implemented technologies may intensify depersonalization, thoughtfully designed systems may strengthen the human dimensions of care. Thematic analysis identified three main categories of AI-enabled interventions associated with rehumanization: (1) tools supporting health care professionals, (2) tools enhancing patient autonomy and engagement, and (3) hybrid solutions addressing both.

Participants most frequently associated the rehumanizing potential of AI with its ability to reduce clinicians' administrative burden. Systems that transcribe consultations, generate structured documentation, and retrieve relevant medical histories may allow physicians to devote more time to direct patient interaction. Instead of navigating electronic health records or writing discharge notes, clinicians can focus more on listening to patients, explaining diagnoses, and cocreating care plans. AI-powered summarization of clinical literature and patient data can also support more informed decision-making in real time.

Participants also highlighted patient-facing AI tools as an important means of improving communication and engagement, particularly by transforming how health information is delivered and understood. Many medical documents are written in complex clinical language that is difficult for nonspecialists to interpret. Large language models can simplify diagnostic reports, provide plain-language explanations of laboratory results, and respond to patient questions conversationally. These functions can improve health literacy, reduce knowledge asymmetries, and foster greater inclusion in the care process, particularly for patients managing chronic or complex conditions.

AI was also described as offering scalable solutions for patient education. Conversational agents providing information about preventive screening or disease management (eg, cervical cancer or diabetes) can reach large populations without increasing clinicians' workload. This may be especially valuable in resource-constrained settings or underserved communities where access to direct clinical guidance is limited.

> *Solutions are now available that simplify results while providing very empathetic and sympathetic responses, which I believe makes care very patient-centered. And this is precisely the area where healthcare workers generally lack capacity due to their heavy workload.* [Participant 2]

Participants also noted that patients with complex diagnoses or intensive treatment regimens are often overwhelmed by the volume of information they receive. AI tools can structure treatment plans, medication schedules, and follow-up instructions, allowing patients to review them at their own pace. This can improve understanding, support emotional processing, and strengthen trust, particularly during vulnerable periods. While such tools cannot replace human emotional support, they can enhance the continuity and clarity of communication.

Hybrid AI applications supporting both patients and clinicians were also frequently mentioned. Digital preconsultation tools can structure medical histories and family background information, providing clinicians with concise summaries that enable more focused consultations while helping patients reflect on relevant symptoms and contextual factors. However, without careful workflow integration and adequate user education, such systems may increase communication burdens, particularly in resource-constrained primary care settings.

AI can also support remote patient management through telemonitoring and virtual follow-up platforms, enabling symptom tracking, recovery monitoring, and automated guidance while alerting clinicians to potential concerns. These systems can enhance continuity of care, patient trust, and transparency by supporting appointment scheduling, reminders, and care navigation. Nevertheless, excessive automation risks undermining interactions perceived by patients as inherently human, especially at critical decision points.

> *I believe that the personal relationship between doctor and patient should not change. I think it is very important from both the patient's and the doctor's point of view, but I believe that telemedicine would play a very important role in such pre-screening, where there would be a virtual connection and AI could make a preliminary diagnosis.* [Participant 1]

Dehumanization Aspects of AI in Health Care

Participants also highlighted potential dehumanizing effects associated with the integration of AI into health care. While AI can improve efficiency, optimize workflows, and enhance diagnostic precision, its growing role also raises social and psychological concerns. Chief among these is the risk of dehumanization, whereby patients and clinicians may experience a loss of individuality, agency, or relational connection as reliance on automated systems increases.

From an organizational and systemic perspective, several participants identified the economic logic underlying AI deployment as a key driver of potential dehumanization. In

XSL•FO
RenderX

many health care systems, AI is primarily introduced to reduce costs, optimize resources, and improve efficiency. While these objectives are understandable from a management perspective, they may prioritize throughput over therapeutic relationships. In such contexts, both patients and clinicians risk being reduced to variables within process optimization.

From the patient perspective, interviewees emphasized the erosion of personalized attention as a central concern. Although AI may enhance diagnostic precision and procedural efficiency, increasing mechanization—particularly in automated screening or intervention processes—may reduce opportunities for meaningful interaction with health care professionals. For many patients, the therapeutic relationship itself contributes to healing; the absence of consultation, explanation, or emotional support may lead to dissatisfaction even when clinical outcomes are positive. Physicians are therefore expected to fulfill not only biomedical roles but also psychosocial ones, offering reassurance, guidance, and human presence.

> *I believe that the doctor's personal diagnosis and treatment plan are very important, because they often have the same effect on patients as a psychologist.* [Participant 15]

Limited consultation time may further contribute to frustration, particularly in private health care settings where patients often associate payment with receiving attention and personal interaction. Even when AI-supported treatments produce positive results, a lack of interpersonal engagement may be interpreted as indifference or negligence, highlighting the gap between clinical efficiency and patient experience.

> *In private healthcare in particular, the time spent with the doctor is important to patients. If the doctor spends little time with the patient, they will be dissatisfied and even negative, even if the results are good. They need consultation and conversation, otherwise they will be disappointed.* [Participant 16]

Participants also raised concerns about the impact of AI on health care professionals, particularly regarding professional identity and authority. As AI systems demonstrate competence in tasks traditionally associated with expert clinical judgment, physicians may experience status anxiety or a perceived erosion of professional value. If diagnostic or decision-making responsibilities increasingly shift to automated systems, this may affect morale, job satisfaction, and the long-term attractiveness of the profession.

Another concern relates to overreliance on automated decision-making. If AI systems begin to guide clinical decisions without sufficient human mediation or transparency, patients may experience confusion, mistrust, or resistance, particularly when they cannot understand how recommendations are generated. Current AI models are primarily designed to support clinicians rather than replace them; however, poorly framed or opaque AI-driven recommendations may still be perceived as intrusive or coercive.

At present, AI is largely introduced to support overstretched health care systems and mitigate workforce shortages rather than replace core clinical roles. Nevertheless, as technological capabilities advance, the balance between human and machine involvement may shift, raising questions about the future role of health care professionals and the preservation of human dignity in care.

> *Personalization is important to patients. AI can improve quality in some areas, such as diagnosis, but excessive mechanization (surgery, screening) leads to less contact with doctors, which may result in patient dissatisfaction.* [Participant 16]

Overall, participants emphasized that while AI is frequently presented as a solution to systemic inefficiencies, its implementation must be carefully managed to avoid unintended dehumanizing consequences. Technological advancements should complement rather than replace the relational and interpretive dimensions of medical practice. Maintaining this balance was considered essential to ensuring that health care remains not only efficient but also humane.

Multilevel Factors Shaping Rehumanization and Dehumanization in AI-Enabled Health Care

Overview of Interacting Macro-, Meso-, and Microlevel Determinants

Drawing on the interviews, this section reviews the macro-, meso-, and microlevel factors that may influence the rehumanizing or dehumanizing effects of integrating AI into health care. These factors span structural, organizational, and individual domains, and their interactions are often complex and interdependent. It is important to note that changes at one level can generate cascading effects across other levels, which can amplify or reduce their overall impact on the nature and quality of patient care. Figure 1 provides an overview of the identified macro-, meso-, and microlevel factors and illustrates how these elements, in combination, may shape trajectories toward either rehumanizing or dehumanizing outcomes in health care.

XSL•FO
RenderX

Figure 1. Macro-, meso-, and microlevel factors shaping rehumanizing and dehumanizing trajectories in AI-integrated health care. AI: artificial intelligence.

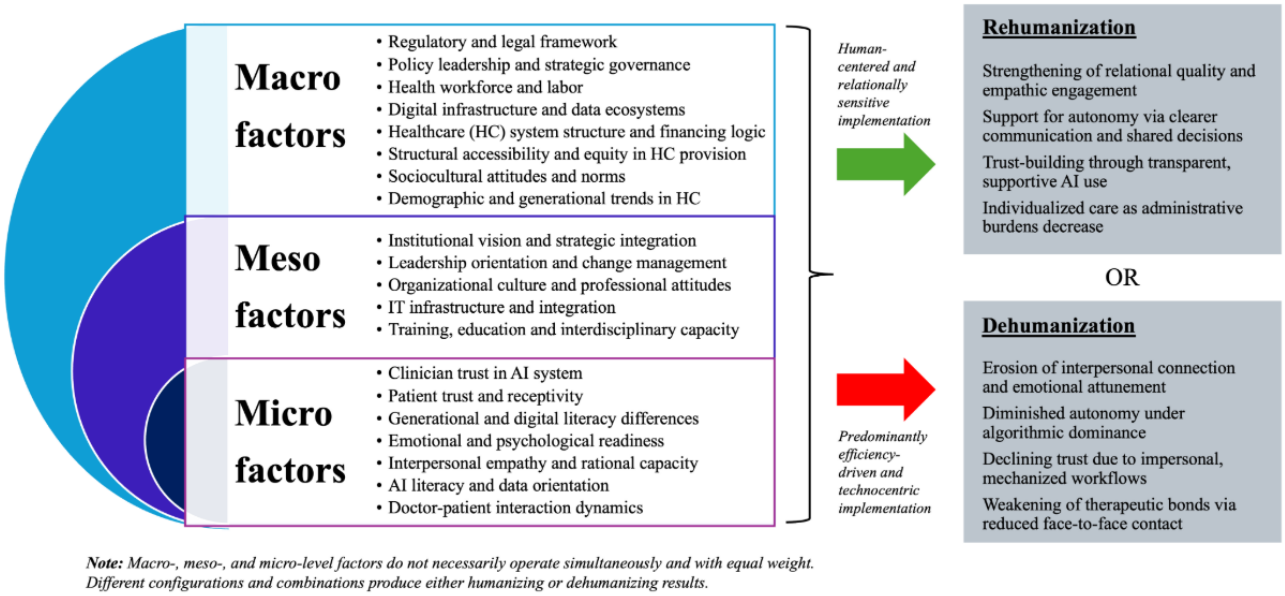

Note: Macro-, meso-, and micro-level factors do not necessarily operate simultaneously and with equal weight. Different configurations and combinations produce either humanizing or dehumanizing results.

Macrolevel Factors

The broader systemic environment in which AI is implemented plays a decisive role in shaping its potential to foster rehumanization or to perpetuate dehumanizing effects. A range of macrolevel factors (see Multimedia Appendix 3)—spanning regulatory and legal structures, policy leadership, workforce capacity, digital and data infrastructures, health care financing models, accessibility, sociocultural norms, and demographic trends—collectively influence the conditions under which AI adoption occurs and the equity, sustainability, and human-centeredness of its outcomes.

The *regulatory and legal framework* encompasses laws, ethics, and institutions related to health care AI. Alongside sector-specific legislation, it encompasses data protection frameworks such as the General Data Protection Regulation and AI governance mechanisms, including the EU AI Act. This framework establishes the legal and ethical boundaries within which AI systems can be integrated into health care, thereby shaping the scope and nature of their application.

> *The EU AI Act, which is otherwise a very good, comprehensive and forward-looking piece of legislation.* [Participant 12]
>
> *And now the European Health Data Space is also being developed, with the aim of enabling the transfer of health data across national borders.* [Participant 7]

Effective regulation can safeguard patient data, ensure ethical oversight, and protect vulnerable populations, thereby strengthening trust in AI adoption. Clear legal frameworks promote accountability and transparency while helping align AI use with patient-centered values rather than purely operational or commercial priorities. However, overly rigid regulations may limit context-sensitive AI implementation, particularly in rapidly evolving health care environments. Legal uncertainty—such as unclear liability rules or inconsistent national policies—can delay adoption or encourage excessively cautious use. In addition, compliance requirements may divert institutional resources toward administrative tasks, potentially increasing workload pressures on health care providers.

Policy leadership and strategic governance delineate the prioritization, coordination, and financing of health care AI deployment by political institutions. This includes national strategies, dedicated government bodies, and long-term investment policies to develop and integrate AI. Political considerations influence the use of AI in health care systems to either enhance equity, quality, and innovation or to constrain costs and optimize operational efficiency.

Strategic coordination across ministries, agencies, and sectors reduces redundancy and promotes scalable, coherent health care ecosystem execution. Clear communication of policy objectives facilitates stakeholder engagement and equips society for the transformative role of AI in health care. A government focused on cost containment may portray AI as reducing the workforce, reducing the human element of care delivery. Politically exploiting AI for immediate benefit can hurt patient-centered innovation and confidence. Disjointed implementations may fail to scale and exclude key stakeholders from decision-making without a strategic vision.

Health care workforce and labor market dynamics refer to the availability, distribution, skills, and well-being of health care personnel, as well as broader labor market trends within the sector. These include workforce shortages, the undercompensation of support staff, and generational differences in professional values and expectations. Such conditions shape whether AI is perceived as a tool that supports health care professionals or as a potential substitute for them.

By reducing administrative tasks and cognitive workload, AI can allow clinicians to devote more time to relational aspects of care, potentially improving patient interaction and professional satisfaction. In resource-constrained settings, AI may also support diagnostic accuracy and therapeutic

XSL•FO
RenderX

decision-making, enabling more timely and effective care. However, if AI is used to replace rather than assist clinicians, it may weaken human interaction and the therapeutic relationship. The perceived devaluation of professional expertise due to automated decision-making may also contribute to disengagement, reduced morale, and resistance to innovation. In extreme cases, excessive reliance on algorithmic decision support may limit professional judgment and contextual adaptation, thereby constraining personalized care.

> *We see that it is simply necessary to accept in the clinical field that their future will be determined by their attitudes towards technology and their cooperation with technology, and new professions will be created.* [Participant 9]

At the national level, *digital infrastructure and data ecosystems* provide the technical and organizational foundations for digital health and AI-specific functionalities, including interoperable electronic health records, secure data exchange frameworks, and advanced computational resources such as national cloud services and large-scale health data repositories. Differences between well-resourced, integrated national systems and underfunded, fragmented ones directly shape the feasibility and quality of AI implementation across the health care sector.

AI can be effectively incorporated into workflows within centralized, interoperable national health information systems, thereby reducing care fragmentation and enabling physicians to prioritize patient engagement over administrative tasks. Efficient access to comprehensive, longitudinal patient records enhances trust and continuity in multidisciplinary and cross-regional contexts. Harmonized national data ecosystems facilitate early detection, enable proactive public health interventions, and provide tailored decision support. Fragmented national IT infrastructures compel health care providers to manage various systems, resulting in decreased efficiency and diminished patient-centered communication.

> *Three factors are necessary for AI-based developments: sufficient quantity and quality of available data, interconnected central systems, and data science capacity, including trained physicists, engineers, data scientists, and clinicians.* [Participant 9]

The *structure and financing logic of health care systems* shape how care is organized and funded at the macrolevel, including whether systems are centralized or market-driven, publicly funded or privately operated. These arrangements—through national insurance schemes, private insurers, or hybrid models—influence how AI is adopted, who benefits from its use, and whether priorities focus on equity, quality, efficiency, or profitability.

In centralized, publicly funded systems, AI can support equitable access, continuity of care, and population health management. Insurance models that emphasize prevention and long-term outcomes may facilitate AI-supported early detection, chronic disease management, and personalized care planning. Data-driven resource allocation can further improve efficiency while promoting fair distribution of services. In market-oriented or predominantly private systems, AI adoption is more often driven by cost containment and service standardization. Access to advanced AI-supported care may become concentrated among higher-income or digitally literate populations, potentially widening disparities. Profit-oriented reimbursement structures may also deprioritize complex or less profitable cases, reinforcing existing inequalities.

Structural accessibility and equity in health care provision influence the accessibility of health care for individuals and communities. The geographic distribution of health care facilities, the balance between public and private sectors, insurance coverage, affordability, and infrastructure availability extend beyond just digital access. The beneficiaries of AI-enhanced health care are contingent upon these structural circumstances. In regions with restricted in-person health care, AI can facilitate remote access for telemedicine, virtual consultations, and remote diagnostics. The introduction of public AI can mitigate systemic disparities in service distribution by using new technology in areas of need rather than in locations of maximum profitability. Moreover, algorithmic triage and resource allocation might enhance fairness in high-demand situations by prioritizing instances with the greatest necessity, thus enhancing efficiency without sacrificing equality. If AI deployment adheres to a market-driven model, benefits may be concentrated in affluent metropolitan regions, exacerbating the accessibility gap.

Sociocultural attitudes and health norms reflect prevailing cultural values that shape how health care is perceived and delivered, including views on medical authority, patient autonomy, and shared decision-making, as well as whether health is understood as a personal responsibility or a collective right. When developed inclusively, AI can enhance cultural awareness, tailor care to diverse societal values, and help reduce stigma associated with mental health and chronic illness. Sociotechnical integration—aligning technological design with social and cultural contexts—may enhance patient autonomy and mitigate power imbalances in physician-centered systems, enabling more participatory and collaborative care. However, standardized AI protocols may conflict with diverse cultural perspectives on health, aging, and mortality, while algorithmic biases and excessive uniformity risk reinforcing cultural exclusion, overlooking relational and symbolic aspects of care, and contributing to the commodification of health care in highly individualistic systems.

Demographic and generational trends shape health care demand, service expectations, and the adoption of medical technologies. Population aging increases the need for long-term care, chronic disease management, and personalized support, while generational differences in digital literacy and attitudes toward autonomy and professional authority influence how AI is used across patient groups. AI technologies may help address complex care needs associated with aging populations, including multimorbidity and cognitive decline, and support aging in place through remote monitoring and home-based care. Designing AI systems with generational inclusivity in mind can reduce barriers to access and help sustain autonomy in later life, while digitally fluent younger cohorts may facilitate more participatory care models. However, AI systems that do not account for cognitive, sensory, or functional limitations may be inaccessible

XSL•FO
RenderX

for older adults. Replacing human care with AI in eldercare settings may also risk social isolation and emotional neglect. Age-related bias in data or system design may further marginalize older adults, while generational mistrust of technology may limit engagement with AI-supported care.

> *Very often, technical limitations are the problem, with the older generation lacking basic computer skills. Middle-aged and older generations working in healthcare are unaware of the possibilities available to them.* [Participant 6]

Mesolevel Factors

The institutional context in which AI is introduced to the health care sector is crucial in determining whether such technologies promote rehumanization or reinforce dehumanizing tendencies. A variety of mesolevel factors (see Multimedia Appendix 4), including institutional strategy, leadership, organizational culture, IT infrastructure, and training ecosystems, influence how AI is adopted, integrated, and experienced within clinical settings.

One such factor is the *institutional vision and strategic integration* of AI. This pertains to the extent to which health care organizations use AI in their strategic, operational, and innovative objectives. This examines whether AI is used to resolve clinical and systemic challenges or merely as a trend. AI solutions that enhance clinician-patient engagement, diminish administrative burdens, and augment access to care are more likely to be adopted in patient-centered institutions.

Public hospitals, academic medical centers, and private clinics exhibit significant variations in their institutional methodologies and health care provision. Academic institutions possess numerous financing avenues and robust research environments, whereas public hospitals sometimes contend with constrained finances and restrictions. This enables them to adopt AI in an ethical and empirical manner. Conversely, private clinics may emphasize efficiency, competitiveness, and patient satisfaction, perhaps fostering innovation and a focus on short-term implementation strategies.

AI may be implemented in institutions reactively, motivated by perceived innovation pressures rather than genuine therapeutic requirements, and without a strategic framework. Ad hoc or shallow adoption typically results in fragmented implementation, misaligned use cases, and system redundancy. AI technologies may lack meaningful integration into care operations without defined objectives, human engagement, or responsibility. Misaligned institutional metrics and incentives constitute another challenge. When performance metrics emphasize throughput, cost reduction, or digital adoption over care quality, AI is used to achieve operational objectives rather than address human requirements. Even in circumstances where human presence is essential for establishing trust, communication, and developing a therapeutic alliance, automated systems may supplant interpersonal care.

> *In my opinion, AI will essentially be a tool that changes at an incredible pace, with new solutions and opportunities emerging every week. I believe that incorporating related tasks into the institution's strategy is important. It is not certain that anyone needs to develop a separate AI strategy, but it is important to incorporate it.* [Participant 13]

Another key mesolevel factor shaping the human impact of AI implementation is *leadership and change management*. This refers to how institutional leaders assess, prioritize, and guide the implementation of AI, including their openness to innovation, responsiveness to staff feedback, and ability to manage organizational change. Leadership also influences resource allocation and determines whether AI is framed as a collaborative clinical support tool or primarily as an efficiency-driven directive.

Visionary and inclusive leadership can position AI as a means of strengthening human-centered care. Leaders who support iterative learning, pilot initiatives, and ethical reflection create opportunities for staff to contribute to implementation processes and raise concerns. Training, stakeholder engagement, and phased implementation can foster shared ownership and trust, enabling context-sensitive AI integration that supports clinical values and continuity of care. Conversely, top-down implementation that excludes clinical staff may undermine legitimacy and provoke resistance or superficial compliance. When leaders focus primarily on efficiency or innovation metrics, the ethical, relational, and psychological implications of AI adoption may be overlooked, weakening long-term acceptance and institutional trust. Rapid deployment without adequate change management may also disrupt workflows, blur responsibilities, and erode confidence in both the technology and leadership.

The *organizational culture and professional attitudes* significantly impact the adoption and experience of health care AI. This encompasses collective beliefs and norms on innovation, hierarchy, care provision, and the impact of AI on professional identity. In cultures that prioritize open discourse, critical analysis, and employee-driven innovation, AI can augment clinical judgment rather than undermine it. Professionals are more likely to engage when their expertise is acknowledged, and adoption processes align with their lived experiences. Recognizing the diverse experiences of individuals across various occupations and levels of seniority mitigates the risk of professional marginalization. Conversely, distrust, superficial implementation, and rigid hierarchies may engender skepticism or defensiveness toward AI. AI may be opposed or used superficially if perceived as a threat to clinical autonomy or a cause of deskilling. AI may devolve into a superficial technology, diminishing care quality and employee morale.

> *Our colleagues are open to technology, competitiveness is important to them, so they have to adopt new solutions, be open-minded and follow trends. They consider digitalization important, and younger employees are generally more open to digital and AI-based solutions.* [Participant 16]

The implementation of AI in health care also depends on institutional *IT infrastructure*, including system interoperability, legacy software, device compatibility, data capacity, and integration with national or regional platforms. Differences between well-funded, integrated systems and fragmented,

XSL•FO
RenderX

under-resourced environments significantly affect the feasibility and quality of AI adoption.

> *If a unified healthcare IT system were to be introduced nationwide, data would be much easier to access. We often encounter patients who have already been treated by three, four, five or even six different healthcare providers, and gathering all of their data is a particularly lengthy process.* [Participant 6]

In interoperable digital environments, AI can be integrated into clinical workflows to enable faster information retrieval, reduce administrative burdens, and support patient-centered care. Cloud-based solutions may further expand access to advanced AI tools without overloading local infrastructure, particularly in resource-constrained settings, provided that appropriate regulatory safeguards and data protection mechanisms are in place to ensure safe and equitable implementation. Conversely, outdated systems and poor interoperability may fragment data, disrupt care continuity, and increase technological complexity for clinicians. Such conditions can limit the effective use of AI and shift attention away from patient interaction. Overall, IT infrastructure plays a decisive role in determining whether AI supports more human-centered care or contributes to digital detachment.

The knowledge, application, and use of AI are also contingent upon the ability of health care facilities and universities to provide relevant *education, training, and interdisciplinary capacity*. This includes technical upskilling for clinical staff as well as academic programs that promote ethical awareness, digital literacy, and collaboration across medical, technological, and policy domains. Institutions that invest in integrated and interdisciplinary education are better positioned to support socially responsible AI implementation. Incorporating critical AI literacy into university curricula helps future professionals engage with AI reflexively and prioritize human values in technology design and use. Continuous and inclusive training for health care professionals can further ensure that AI supports rather than replaces human judgment and empathy. Conversely, technocentric or narrowly focused training may lead to overreliance on AI or resistance due to limited understanding. Educational inequalities between institutions or regions may also reinforce systemic disparities by concentrating AI literacy and influence in already well-resourced settings.

Microlevel Factors

At the microlevel, the interactions, perceptions, and capabilities of individual stakeholders are pivotal in shaping the experience of AI in health care. Factors such as clinician and patient trust, generational and digital literacy differences, emotional readiness, interpersonal empathy, AI literacy, and the dynamics of doctor-patient relationships (see Multimedia Appendix 5) directly influence whether AI integration enhances or undermines the human connection in care delivery.

Clinician trust in AI systems is a key microlevel factor shaping the human impact of AI in health care, referring to health care professionals' confidence in the reliability, clinical relevance, and ethical integrity of AI tools. This trust is influenced by factors such as prior technological experience, alignment with clinical judgment, system transparency, and perceived implications for professional identity. When clinicians are involved in the development or adaptation of AI tools, they are generally more willing to adopt them. Transparent and explainable systems enable professionals to evaluate AI outputs and integrate them into clinical reasoning, thereby strengthening confidence. AI that supports rather than replaces medical decision-making can reinforce clinical authority and encourage responsible use.

Conversely, limited transparency may undermine trust, particularly when clinicians cannot understand how recommendations are generated. If AI is perceived as replacing rather than supporting expertise, it may provoke resistance, disengagement, or underuse. The introduction of AI without clinician involvement can also threaten professional identity and reduce acceptance. Ultimately, clinician trust develops through experience with AI implementation and plays a decisive role in determining whether AI strengthens clinical practice or undermines professional autonomy and confidence.

> *I think the biggest obstacle to trust at the moment is that practical experience and success with AI solutions is still limited, with the exception of one or two areas, such as imaging diagnostics. I think this is exactly what centralized implementation could help with: a centralized pilot in one or two areas, so that everyone is involved to some extent and can see that this really can be a credible system.* [Participant 1]

Patient trust and receptivity reflect patients' readiness to engage with AI-supported health care. This trust is shaped by communication quality, transparency regarding the use of AI, prior digital experience, and expectations about human involvement in care. When health care professionals clearly and empathetically explain how AI is used, patient autonomy, informed consent, and engagement may increase. Familiarity with consumer AI technologies, such as chatbots or recommendation systems, may also normalize AI in health care. The accuracy and personalization of AI-supported services can further enhance their perceived legitimacy and encourage patient participation.

Conversely, limited transparency or impersonal use of AI tools may weaken patient trust, particularly if technology replaces human interaction or its role remains unclear. Patients may evaluate AI systems more critically than human clinicians, expecting consistently high levels of accuracy and objectivity. While human errors are often interpreted within a relational context, even minor AI errors may provoke skepticism. In addition, digital exclusion or limited digital literacy—often linked to socioeconomic status, age, or education—may lead to alienation and reduced confidence in AI-supported care.

Differences in generational and digital literacy encompass digital competence, receptiveness to innovation, and technological familiarity, frequently influenced by age and exposure to digital tools. These differences influence the adaptation of health care workers and patients to clinical AI systems. Younger, digitally proficient users might facilitate AI adoption by normalizing new technology and assisting those

XSL•FO
RenderX

who are less technologically inclined. Systematic and focused training can mitigate digital exclusion and assist older or less experienced users in effectively using AI-assisted care. Restricted digital competencies among older professionals or patients may result in social isolation or a passive reliance on others for technological navigation, thus constraining autonomy and self-assurance. Access to digital resources, such as high-speed internet and suitable devices, may exacerbate socioeconomic and health disparities, leading to unequal participation with AI-enabled services.

The *emotional and psychological preparedness*—comprising emotional condition, cognitive capacity, and willingness to change—affects the use of AI technologies by health care professionals and patients. Burnout, stress, and work-life balance can influence motivation, flexibility, and the willingness to adopt new tactics. When used to alleviate administrative burdens, AI solutions can diminish emotional fatigue, enhance professional motivation, and optimize work-life balance. Framing AI as a supportive rather than a disruptive element in practice can facilitate its acceptance as an ally in care delivery rather than a menace to autonomy or competence. Health care professionals frequently experience burnout or emotional fatigue because of excessive administrative burdens or staffing constraints, which can impede learning and adaptability, resulting in avoidance or resistance. AI systems that introduce complexity, diminish control, or undermine professional autonomy may heighten stress and disengagement. Individuals may emotionally detach from innovation processes, perceiving AI as a hindrance rather than an asset.

Interpersonal empathy and relational capacity are key qualities in a clinician, describing their ability to communicate with empathy, understand patients' perspectives, and sustain meaningful interpersonal relationships. Establishing trust, fostering patient engagement, and ensuring technology enhances rather than supplants human care necessitate these qualities.

> *According to numerous studies, machines already produce much better results than humans in certain areas, but people still like to hear these results from human beings and believe them. That way, they believe them more readily.* [Participant 13]

Clinicians may prioritize relationship-centered care by using AI-driven efficiencies such as automated documentation and data retrieval. Training in communication and empathy can assist physicians in preserving emotional attunement in digitally mediated environments. Clinicians managing decision-making guarantee ethical responsibility and relational confidence. However, assigning linguistic or emotional tasks to AI may diminish patient emotions. An excessive dependence on data-driven outputs may divert attention from patients' narratives, values, and subjective experiences, which are essential for mutual comprehension and personalized care. Minimizing in-person interactions may undermine the therapeutic connection, weakening relational dynamics that enhance patient confidence.

Furthermore, *AI literacy and data orientation* refer to health care professionals' ability to understand, critically evaluate, and appropriately use AI tools and data-driven outputs in clinical practice. This competence is essential to ensure that AI supports decision-making rather than being over-relied upon or dismissed. It includes understanding how AI systems function, recognizing their limitations, and interpreting their outputs in clinically and ethically appropriate ways.

Health care professionals with foundational AI literacy can critically assess AI recommendations, integrate them into evidence-based practice, and personalize care accordingly. A data-oriented approach can further support the effective use of patient information for informed decision-making. Institutional support through training and continuous education can strengthen professionals' confidence and autonomy in using AI responsibly. Conversely, limited technical understanding may lead to inappropriate use, such as applying AI outputs without sufficient context or uncritically relying on automated recommendations. Uncertainty about how AI systems operate may also reduce professional confidence and hinder effective decision-making.

Finally, the *dynamics of doctor-patient interactions* encompass the communication, ethics, and power relations that influence how clinicians and patients exchange information, make decisions, and build trust. AI technologies have the potential to reshape these dynamics and transform clinical decision-making, information accessibility, and interpersonal standards.

> *It is a serious challenge today that patients arrive at the doctor's office already thinking they know more than their doctor, having looked things up on Google, ChatGPT, etc.* [Participant 9]

When used thoughtfully, AI may simplify intricate medical information and promote collaborative decision-making. Using AI tools transparently during consultations demonstrates that technology enhances clinical judgment, hence fostering confidence. AI has the potential to distribute epistemic authority, diminishing hierarchical barriers and enabling patients to engage more actively in discussions regarding their care. If AI systems dictate all decisions, algorithmic authority may consolidate control, obscure accountability, and restrict patient involvement. Exclusion of patients from AI-driven decisions may undermine trust and autonomy. Furthermore, substituting dialogical engagement with automated methods may depersonalize the therapeutic alliance and diminish the essential human connection required for effective therapy.

Discussion

Principal Findings

This study represents one of the first empirical investigations of the potential rehumanizing and dehumanizing effects of AI integration in health care, examining how these outcomes are shaped by the interaction of macro-, meso-, and microlevel factors. By systematically mapping these influences, the research provides a multilevel perspective on the social, organizational, and ethical conditions that shape the practical use of AI in health care. These insights can inform policymakers, health care leaders, and clinicians on how to implement AI systems that are both operationally effective and aligned with patient-centered care. By highlighting both opportunities and risks, the study

XSL•FO
RenderX

also contributes to anticipating and mitigating unintended consequences, helping ensure that AI strengthens rather than undermines the human dimensions of medicine.

Following a series of interviews, it was found that participants generally agreed that AI has significant potential to enable rehumanization in the field of health care. However, they emphasized that without a thorough understanding of the macro-, meso-, and microlevel factors identified in this study—and without deliberate alignment with patient-centered principles—the benefits of AI could easily be compromised or redirected toward efficiency goals that fail to enhance the patient experience. The findings reaffirm the widely acknowledged strengths of AI—its ability to improve diagnostic accuracy, optimize workflows, reduce administrative burdens, and enable more personalized, targeted treatments—while also making clear that these benefits will not be realized automatically. To achieve equitable and sustainable implementation, it is essential to engage in careful planning, structured education, robust support for patients and professionals, and adequate infrastructure. The interviews indicated variations among stakeholder groups. Senior physicians exhibited diminished trust in AI solutions, reflecting generational disparities at the microlevel, whereas legal experts and business consultants underscored the intricacies of regulatory frameworks, perceiving regulation as either a vital safeguard or a possible impediment to innovation; across various viewpoints, the incorporation of digital ethics and data protection was uniformly recognized as crucial.

Participants emphasized that AI could improve care quality, alleviate workforce shortages, and reduce burnout by shifting repetitive tasks away from clinicians. However, this raises fundamental strategic questions: how will the freed capacity be used? If regulations, institutional strategies, or governmental directives require the use of AI for rehumanization purposes, will this result in longer, more meaningful consultations and stronger follow-up, or will it simply normalize shorter patient interactions? The evolving role of medical assistants further illustrates the complexity of such changes. Assistants often contribute far more than technical support: they serve as witnesses during consultations, provide personal assistance, and support the relational aspects of care. Whether these roles are preserved, adapted, or eliminated depends on organizational priorities, budgets, and long-term strategies. The interviews also revealed the perception that technological advancement has diminished the authority of physicians in some contexts. This highlights the importance of preserving trust, professional identity, and interpersonal rapport in technology-rich care environments. Empathy was identified as another critical area for improvement: while certain specialties prioritize trust over explicit empathy, time saved by AI could be invested in rebuilding more meaningful patient-clinician relationships, particularly given the prevalence of burnout and reduced patient engagement. Currently, certain core aspects of care, such as physical examinations and direct patient contact, cannot be delegated to AI and still require skilled human involvement. However, future developments may alter this balance. Furthermore, time savings from automation will only translate into better care if they are deliberately structured to enhance the patient experience. In the absence of intentionality, efficiency gains may become detached from concrete enhancements in outcomes. Importantly, increased time availability alone does not automatically lead to stronger patient-centered care. While AI may create opportunities for more human interaction, patient-centered care requires a shared understanding of its principles, including meaningful communication and the cocreation of care, as well as organizational and professional conditions that enable clinicians to use this time in patient-centered ways.

In summary, this study's contribution lies in providing one of the first empirically grounded frameworks for understanding how AI in health care can promote or hinder rehumanization across interconnected system levels. It shows that the potential of AI is not just a technical issue but also depends on strategic, ethical, and cultural factors. Realizing this promise will require coordinated action across regulatory frameworks, institutional strategies, governmental policies, workforce planning, and cultural adaptation, ensuring that technological progress serves as a catalyst for more human-centered, empathetic, and equitable health care.

Theoretical Contributions

Building on established frameworks of humanization and patient-centered care [2,3], this study makes a substantive empirical contribution by demonstrating that rehumanization and dehumanization in AI-enabled health care are not inherent to the technology but are produced by concrete implementation choices and governance structures. While prior literature has highlighted the ethical and relational implications of AI in health care [8,26], the present findings clarify how these implications are enacted across macrolevel policy incentives, mesolevel organizational strategies, and microlevel clinical interactions. In doing so, the study advances patient-centered care and 4P medicine scholarship [17] by showing that personalization, participation, and relational quality are contingent achievements shaped by institutional priorities, leadership, and everyday clinical use, rather than automatic consequences of AI adoption.

Practical and Policy Implications

The findings indicate that the influence of AI on rehumanization or dehumanization is significantly determined by policy and leadership decisions, rather than by technology in isolation. At the macrolevel, national AI and digital health strategies must incorporate human-centered values, ensuring that regulatory frameworks, funding mechanisms, and performance indicators facilitate patient-centered care, equity, transparency, and human oversight in conjunction with efficiency. At the mesolevel, hospital leadership is crucial in implementing these principles by incorporating AI into organizational strategies that assist clinicians, alleviate administrative burdens, and intentionally reinvest time savings into significant patient interactions. In the absence of policy and leadership alignment, AI may exacerbate efficiency-driven care models, undermining human connection.

Limitations

While this study provides valuable insights into the factors influencing rehumanization and dehumanization in AI-enabled health care, certain contextual considerations must be

XSL•FO
RenderX

acknowledged. Several participants had considerable international experience in advisory, development, research, and educational roles, as well as in clinical practice. Most notably, this experience was gained within European health care systems, with a few also having experience in the US market. This breadth of expertise enriched the perspectives captured in the research, offering viewpoints that extend beyond a solely domestic context. However, it should be noted that the findings are inevitably influenced by the regulatory, organizational, and cultural environment in which the study was conducted. Future research would benefit from the inclusion of a greater number of experts with diverse international market experience and from comparative studies across different health care systems and cultural contexts. Such studies would enhance the generalizability and applicability of the results, ensuring their relevance to a wider range of settings.

Conclusions

The study's findings show that AI does not by itself rehumanize or dehumanize health care; rather, its human impact is determined by how it is regulated, applied, and integrated into routine clinical practice. The findings demonstrate that rehumanization is a contingent outcome influenced by organizational tactics, professional attitudes, governmental decisions, and relational dynamics rather than an inevitable consequence of technological advancement by identifying interrelated macro-, meso-, and microlevel elements. Therefore, intentional alignment between technology innovation and humanistic principles is necessary to ensure that AI enhances rather than detracts from patient-centered care. If AI is to work as a catalyst for more compassionate treatment, future health care systems will need to strike a balance between efficiency gains and ongoing investments in empathy, trust, and meaningful clinician-patient connections.

Acknowledgments

The authors would like to thank the handling editor and the two anonymous reviewers for their thorough and constructive feedback, which substantially contributed to improving the clarity and quality of the manuscript. The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing, adapting and adjusting emotional tone, translation, and reformatting. The GAI tools used were Quillbot and DeepL. Responsibility for the final manuscript lies entirely with the authors.

Funding

This statement applies only to DH: The project identified by EKOP-CORVINUS-24-4 was realized with the support of the National Research, Development, and Innovation Fund provided by the Ministry of Culture and Innovation, as part of the University Research Scholarship Program announced for the 2024/2025 academic year.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to the need to protect participant anonymity and confidentiality.

Responsibility for the final manuscript lies entirely with the authors.
GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Semistructured interview guide across macro-, meso-, and microlevel dimensions.
[DOCX File, 21 KB - jmir_v28i1e82774_app1.docx]

Multimedia Appendix 2
Characteristics and artificial intelligence–related experience of interview participants.
[DOCX File, 26 KB - jmir_v28i1e82774_app2.docx]

Multimedia Appendix 3
Macrolevel factors influencing the rehumanizing and dehumanizing potential of artificial intelligence implementation in health care.
[DOCX File, 23 KB - jmir_v28i1e82774_app3.docx]

Multimedia Appendix 4

XSL•FO
RenderX

Mesolevel factors influencing the rehumanizing and dehumanizing potential of artificial intelligence implementation in health care.
[DOCX File, 21 KB - jmir_v28i1e82774_app4.docx]

Multimedia Appendix 5
Microlevel factors influencing the rehumanizing and dehumanizing potential of artificial intelligence implementation in health care.
[DOCX File, 22 KB - jmir_v28i1e82774_app5.docx]

Checklist 1
COREQ checklist.
[DOCX File, 44 KB - jmir_v28i1e82774_app6.docx]

References

1. Marion S, Ghazal L, Roth T, Shanahan K, Thom B, Chino F. Prioritizing patient-centered care in a world of increasingly advanced technologies and disconnected care. Semin Radiat Oncol 2024 Oct;34(4):452-462. [doi: 10.1016/j.semradonc.2024.07.001] [Medline: 39271280]
2. Todres L, Galvin KT, Holloway I. The humanization of healthcare: a value framework for qualitative research. Int J Qual Stud Health Well-being 2009 Jan;4(2):68-77. [doi: 10.1080/17482620802646204]
3. Rathert C, Wyrwich MD, Boren SA. Patient-centered care and outcomes: a systematic review of the literature. Med Care Res Rev 2013 Aug;70(4):351-379. [doi: 10.1177/1077558712465774] [Medline: 23169897]
4. Alquayt A, Aljuhani O, Alharthi AF, et al. AI-driven healthcare innovations for enhancing clinical services during mass gatherings (Hajj): task force insights and future directions. BMC Health Serv Res 2025 Jul 1;25(1):876. [doi: 10.1186/s12913-025-13045-5] [Medline: 40598365]
5. Lintz J. Exploring the impacts of artificial intelligence interventions on providers' practices: perspectives from a rural medical center. J Med Artif Intell 2025 Dec 30;8:39-39. [doi: 10.21037/jmai-24-422]
6. Hill NR. Overcoming adoption challenges: bridging the gap between artificial intelligence algorithm development and implementation in healthcare. J Med Artif Intell 2025;8:55-55. [doi: 10.21037/jmai-24-385]
7. Topol EJ. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again: Basic Books; 2019.
8. Hoogendoorn CJ, Rodríguez ND. Rethinking dehumanization, empathy, and burnout in healthcare context. Curr Opin Behav Sci 2023 Aug;52:101285. [doi: 10.1016/j.cobeha.2023.101285] [Medline: 37663379]
9. Jagtiani P, Karabacak M, Margetis K. A concise framework for fairness: navigating disparate impact in healthcare AI. J Med Artif Intell 2025 Dec 30;8:51. [doi: 10.21037/jmai-24-438]
10. Sharma RM. Artificial intelligence in medical image analysis and molecular diagnostics: recent advances and applications. J Med Artif Intell 2025;8:53. [doi: 10.21037/jmai-24-412]
11. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
12. Holtz B, Nelson V, Poropatich RK. Artificial intelligence in health: enhancing a return to patient-centered communication. Telemed J E Health 2023 Jun;29(6):795-797. [doi: 10.1089/tmj.2022.0413] [Medline: 36383108]
13. Khullar D, Casalino LP, Qian Y, Lu Y, Krumholz HM, Aneja S. Perspectives of patients about artificial intelligence in health care. JAMA Netw Open 2022 May 2;5(5):e2210309. [doi: 10.1001/jamanetworkopen.2022.10309] [Medline: 35507346]
14. Rony MKK, Numan SM, Akter K, et al. Nurses' perspectives on privacy and ethical concerns regarding artificial intelligence adoption in healthcare. Heliyon 2024 Sep 15;10(17):e36702. [doi: 10.1016/j.heliyon.2024.e36702] [Medline: 39281626]
15. Rosengren K, Buttigieg SC, Badanta B, Carlstrom E. Diffusion of person-centred care within 27 European countries – interviews with managers, officials, and researchers at the micro, meso, and macro levels. J Health Organ Manag 2023 Nov 11;37(1):17-34. [doi: 10.1108/JHOM-02-2022-0036] [Medline: 36367331]
16. Guo W, Chen Y. Investigating whether AI will replace human physicians and understanding the interplay of the source of consultation, health-related stigma, and explanations of diagnoses on patients' evaluations of medical consultations: randomized factorial experiment. J Med Internet Res 2025 Mar 5;27:e66760. [doi: 10.2196/66760] [Medline: 40053785]
17. Slim K, Selvy M, Veziant J. Conceptual innovation: 4P medicine and 4P surgery. J Visc Surg 2021 Jun;158(3S):S12-S17. [doi: 10.1016/j.jviscsurg.2021.01.003] [Medline: 33714709]
18. Giebel GD, Raszke P, Nowak H, et al. Problems and barriers related to the use of AI-based clinical decision support systems: interview study. J Med Internet Res 2025 Feb 3;27:e63377. [doi: 10.2196/63377] [Medline: 39899342]
19. Chandrasekaran R, Takale L, Moustakas E. Listening to patients' voices on the use of AI in health care: cross-sectional study. J Med Internet Res 2025 Dec 5;27:e77501. [doi: 10.2196/77501] [Medline: 41348933]
20. Floridi L. Translating principles into practices of digital ethics: five risks of being unethical. Philos Technol 2019 Jun;32(2):185-193. [doi: 10.1007/s13347-019-00354-x]

XSL•FO
RenderX

21. Umbrello S, van de Poel I. Mapping value sensitive design onto AI for social good principles. AI Ethics 2021;1(3):283-296. [doi: 10.1007/s43681-021-00038-3] [Medline: 34790942]
22. Russell C, Maraccini A, Salmi T. “Short, modern, smart”: humanizing healthcare experiences through modernized feedback. J Patient Exp 2024;11. [doi: 10.1177/23743735241272168] [Medline: 39291182]
23. Busch IM, Moretti F, Travaini G, Wu AW, Rimondini M. Humanization of care: key elements identified by patients, caregivers, and healthcare providers: a systematic review. Patient 2019 Oct;12(5):461-474. [doi: 10.1007/s40271-019-00370-1] [Medline: 31203515]
24. Manríquez Roa T, Christen M, Reis A, Biller-Andorno N. The pursuit of health equity in the era of artificial intelligence. Swiss Med Wkly 2023 Jan 23;153:40062. [doi: 10.57187/smw.2023.40062] [Medline: 36724760]
25. Amin K, Khosla P, Doshi R, Chheang S, Forman HP. Artificial intelligence to improve patient understanding of radiology reports. Yale J Biol Med 2023 Sep;96(3):407-417. [doi: 10.59249/NKOY5498] [Medline: 37780992]
26. Haslam N. Dehumanization: an integrative review. Pers Soc Psychol Rev 2006;10(3):252-264. [doi: 10.1207/s15327957pspr1003_4] [Medline: 16859440]
27. Akingbola A, Adeleke O, Idris A, Adewole O, Adegbesan A. Artificial intelligence and the dehumanization of patient care. J Med Surg Public Health 2024 Aug;3:100138. [doi: 10.1016/j.glmedi.2024.100138]
28. Asthana S, Jones R, Sheaff R. Why does the NHS struggle to adopt eHealth innovations? A review of macro, meso and micro factors. BMC Health Serv Res 2019 Dec 21;19(1):984. [doi: 10.1186/s12913-019-4790-x] [Medline: 31864370]
29. Sawatzky R, Kwon JY, Barclay R, et al. Implications of response shift for micro-, meso-, and macro-level healthcare decision-making using results of patient-reported outcome measures. Qual Life Res 2021 Dec;30(12):3343-3357. [doi: 10.1007/s11136-021-02766-9] [Medline: 33651278]
30. Beirão G, Patrício L, Fisk RP. Value cocreation in service ecosystems: investigating health care at the micro, meso, and macro levels. J Service Manag 2017 Apr 18;28(2):227-249. [doi: 10.1108/JOSM-11-2015-0357]
31. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health 2015 Sep;42(5):533-544. [doi: 10.1007/s10488-013-0528-y] [Medline: 24193818]
32. Renjith V, Yesodharan R, Noronha JA, Ladd E, George A. Qualitative methods in health care research. Int J Prev Med 2021;12:20. [doi: 10.4103/ijpvm.IJPVM_321_19] [Medline: 34084317]
33. Glaser BG, Strauss AL. The Discovery of Grounded Theory: Strategies for Qualitative Research: Aldine Publishing; 1967.
34. Strauss A, Corbin J. Grounded theory methodology: an overview. In: Denzin NK, Lincoln YS, editors. Handbook of Qualitative Research: SAGE Publications; 1994:273-285.
35. Corbin J. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory, 3rd edition: SAGE Publications; 2008. [doi: 10.4135/9781452230153]
36. Gioia DA, Corley KG, Hamilton AL. Seeking qualitative rigor in inductive research: notes on the Gioia methodology. Organ Res Methods 2013 Jan;16(1):15-31. [doi: 10.1177/1094428112452151]
37. Holloway I, Galvin K. Qualitative Research in Nursing and Healthcare, 4th edition: Wiley; 2016.
38. Miles MB, Huberman AM. Qualitative Data Analysis: A Sourcebook of New Methods: SAGE Publications; 1984.

Abbreviations

AI: artificial intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research
RQ: research question

Edited by A Stone; submitted 21.Aug.2025; peer-reviewed by G Berntsen, T Yusuff; revised version received 23.Mar.2026; accepted 31.Mar.2026; published 27.May.2026.

Please cite as:
Horvath D, Lorincz NS
Artificial Intelligence in Patient-Centered Care and Macro-, Meso-, and Micro-Level Determinants of Rehumanization and Dehumanization: Qualitative Interview Study
J Med Internet Res 2026;28:e82774
URL: https://www.jmir.org/2026/1/e82774
doi:10.2196/82774

© Dora Horvath, Noemi Szilvia Lorincz. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Performance of Vision-Enabled Large Language Models in Image-Based Electrocardiogram Interpretation: Exploratory Evaluation

Nibras Soubh[1,2], MD; Eva Rasenack[1,2], MD; Helge Haarmann[1,2], MD; Felix Wiedmann[1,2], MD; Markus Zabel[1,2], MD; Constanze Schmidt[1,2], MD; Rayan Suliman[3*], MD; Leonard Bergau[1,2*], MD

[1]Department of Cardiology and Pneumology, University Medical Center Göttingen (UMG), Robert-Koch-Str. 40, Göttingen, Germany

[2]German Centre for Cardiovascular Research, Partner Site Lower Saxony, University of Göttingen, Göttingen, Germany

[3]Institute of Clinical Chemistry and Laboratory Medicine, Municipal Hospital Dresden, Dresden, Germany

[*]these authors contributed equally

Corresponding Author:
Nibras Soubh, MD
Department of Cardiology and Pneumology, University Medical Center Göttingen (UMG), Robert-Koch-Str. 40, Göttingen, Germany

Abstract

Background: Vision-enabled large language models (VE-LLMs) have the potential to provide flexible and explainable medical image interpretation. However, their real-world performance on clinical data, such as 12-lead electrocardiograms (ECGs), has not been systematically assessed.

Objective: This study aimed to evaluate the diagnostic accuracy and reliability of state-of-the-art generalist VE-LLMs in interpreting real-world ECG images.

Methods: We tested 6 generalist VE-LLMs (ChatGPT-5, ChatGPT-4, Gemini 2.5, Copilot, Claude Sonnet-4, and Claude Opus-4.1) using 70 deidentified ECG images. A standardized prompt requested 9 determinations: rhythm, first-degree atrioventricular (AV) block, intraventricular conduction block and pattern, corrected QT (QTc) prolongation, premature atrial and ventricular contractions, ischemic ST-segment deviation, and axis deviation. An expert consensus served as the reference standard. Moreover, 2 image-based ECG-specialized LLMs (PULSE-7B and ECG-Instruct-Llama-3.2-11B-Vision) were tested for exploratory comparison. Model outputs were evaluated using overall and per-category diagnostic metrics.

Results: Overall balanced accuracy across generalist models ranged from 50.1% to 61.8% (Cochran Q, $P<.001$). ChatGPT-5 achieved the highest balanced accuracy (61.8%) but had the slowest response time (median 276, IQR 110-407 s), whereas Copilot responded within a median of 3 (IQR 2-4) seconds. Balanced accuracy for rhythm classification ranged from 38.6% to 55.8%, but sensitivity for atrial fibrillation among generalist models was ≤11.1%, detecting either none or only 1 of the 9 cases. Detection of first-degree AV block (sensitivity 0% - 22%; 0/9 to 2/9) and QTc prolongation (sensitivity 0% - 45.5%; 0/22 to 10/22) was poor. Intraventricular block was identified with up to 67.8% balanced accuracy, but correct subtype assignment was ≤44% (≤11/25). ST-segment deviation sensitivity was <25% for all generalist models (highest 3/14). Agreement with expert interpretation was low, with Cohen κ indicating poor-to-fair concordance (κ≤0.39). Specialized models achieved overall balanced accuracy of 56.5% (ECG-Instruct-Llama-3.2-11B-Vision) and 64.4% (PULSE-7B), with PULSE-7B showing higher task-specific balanced accuracy in rhythm classification and ectopic beats detection (up to 86.3% and 89.2%, respectively).

Conclusions: VE-LLMs showed moderate overall performance but mostly low sensitivity and limited agreement with expert ECG interpretation. Current performance remains inconsistent across models and diagnostic categories and is insufficient to support clinical deployment.

(J Med Internet Res 2026;28:e86692) doi:10.2196/86692

KEYWORDS

artificial intelligence; AI; electrocardiography; ECG; machine learning; large language models; LLMs

Introduction

Large language models (LLMs) are artificial intelligence (AI) systems trained on vast text corpora to generate human-like responses. The public release of the first generalist LLM—OpenAI's ChatGPT—in late 2022 marked a major inflection point. In just 2 years, LLMs have influenced many aspects of daily life and professional practice, with successive

XSL•FO
RenderX

generations steadily expanding conversational and problem-solving capabilities. These models are now being explored across education, law, and scientific research, reshaping workflows while raising new societal and ethical questions.

In medicine, LLMs have shown notable promise on knowledge-based tasks. For example, Google's Med-PaLM and Med-PaLM 2 achieved scores exceeding the passing threshold for the USMLE (United States Medical Licensing Examination) [1]. Meanwhile, multiple studies evaluating publicly accessible LLMs across international medical examinations indicate that newer versions can perform well, although results vary by model, language, and examination difficulty [2,3]. Beyond examinations, LLMs are being piloted for clinical use cases, such as drafting documentation, summarizing literature, and answering patient queries in lay terms [4,5]. However, clinical deployment requires rigorous validation and appropriate regulatory frameworks, particularly for data protection and patient privacy.

The 12-lead electrocardiogram (ECG) is a cornerstone diagnostic tool in cardiology and emergency medicine. Recent advances in AI—especially deep learning—have yielded strong results in automated ECG analysis. Convolutional neural network (CNN) models trained on large ECG datasets can achieve expert-level interpretation and may uncover latent features not apparent to human readers. For instance, multiple AI ECG models have been shown to predict the near-term onset of atrial fibrillation (AF) from ECGs recorded in apparent sinus rhythm (SR), detecting subclinical signatures invisible to clinicians [6,7]. EchoNext—a deep learning model trained on more than one million ECG waveforms and imaging records—outperformed cardiologists in detecting structural heart disease across diverse clinical settings and patient groups [8].

This emerging field of AI-enhanced ECG analysis promises faster and potentially more sensitive screening for cardiac pathology using a ubiquitous, inexpensive test. Despite progress, most AI ECG solutions remain task specific, have limited accessibility, require waveform (signal data) as input, and lack the flexible reasoning of human interpreters. Integrating advanced LLMs into image-based ECG interpretation could combine the pattern recognition strengths of deep learning with the linguistic and contextual reasoning of LLMs. Historically, generalist LLMs accepted only text input. Newer models are multimodal, accepting images in addition to text. GPT-4 with "Vision" (released in 2023) demonstrated that an LLM augmented with a visual encoder can describe images and answer questions about them. Competing vision-enabled LLMs (VE-LLMs) soon followed. Early experience showed impressive performance on general images—object recognition, chart interpretation, handwriting—yet the ability of generalist VE-LLMs to interpret complex medical images remains uncertain, with only a few studies assessing earlier versions on simplified, textbook ECG test sets [9,10].

Previous studies evaluating ChatGPT-4 have largely focused on single-model assessments, limited diagnostic tasks, or simplified test settings, often without structured benchmarking across multiple models and clinically derived datasets. Given the rapid evolution of LLM capabilities, we conducted a comprehensive evaluation of the newest VE-LLMs as of August 2025 (including ChatGPT-5 and Claude Opus-4.1) on a clinically relevant image understanding task: 12-lead ECG interpretation. We focused on fundamental ECG findings—rhythm, atrioventricular (AV) and intraventricular conduction delays (IVCDs), ectopy, and ischemia—routinely assessed by clinicians and automated algorithms. By comparing 6 prominent generalist models side-by-side, we sought to determine how performance varies across models and diagnostic categories and how models compare with human experts and with 2 further ECG-specialized image-based LLMs (PULSE-7B and ECG-Instruct-Llama-3.2-11B-Vision).

Methods

Ethical Considerations

This retrospective diagnostic study was conducted at the University Medical Center Göttingen, Germany, using deidentified 12-lead ECG images interpreted by AI models, with expert human interpretation as the reference standard. The study was reviewed and approved by the local institutional ethics committee (approval 18/08/24). In accordance with the ethics committee decision and given the retrospective design and the use of fully anonymized data, the requirement for informed consent was waived. All data were handled in accordance with institutional data protection policies, and no identifiable personal information was included in the dataset or study outputs. No participants received financial compensation.

ECG Data—Acquisition and Preparation

Seventy 12-lead ECGs were collected from consecutive patients admitted to the heart rhythm ward between June 1 and 15, 2025. Recordings were obtained using the standard hospital device (Cardiovit AT-102 G2; Schiller Medizintechnik GmbH) at 500 Hz, with a paper speed of 50 mm/s and 10 mm/mV calibration. Each ECG was originally in PDF format and included patient metadata and automated interpretations. Patient information, demographics, and machine-generated measurements were removed using Inkscape (Inkscape Project, version 1.3.2), leaving only the waveform grid and tracings. Images were exported as high-resolution JPEG files (300 dots per inch). The final dataset contained 70 ECG images, each a single page with the standard 12 leads (I, II, III, aVR, aVL, aVF, and V1-V6). Two physicians independently reviewed all ECGs to determine the correct findings for each of the 9 diagnostic questions (detailed in the following section). Discrepancies were resolved by consensus discussion. These expert interpretations served as the reference standard. Interval assessments (PR, QRS, and QTc) by expert readers were performed using manual measurements from the ECG images based on standard calibration. Difficulty of ECGs was classified into three categories: (1) easy: normal ECG or isolated clear pathological findings with no baseline disturbances or borderline measurements; (2) intermediate: borderline measurements, subtle findings, baseline artifacts, and the co-occurrence of multiple pathological findings; and (3) difficult: multiple or rare pathological findings and multiple borderline measurements.

XSL•FO
RenderX

Selection and Inference of LLM Models

We identified 7 publicly available generalist multimodal LLMs (accepting image and text input) that represented the top-tier models from major AI developers: ChatGPT-4 and ChatGPT-5 (OpenAI), Gemini 2.5 (Google), Copilot (Microsoft; GPT-4-based, with "Think Deeper" activated), Grok-4 (xAI), Claude Sonnet-4, and Claude Opus-4.1 (Anthropic).

We excluded Meta AI's LLM, as no publicly available multimodal model was provided as of August 2025, and DeepSeek because its vision-enabled chat model is limited to text extraction from images as stated by the model itself. Grok-4 inference was performed via the Perplexity Pro platform; however, it was excluded from final analysis following the disclosure of a "silent fallback bug" by the platform provider. Due to this technical issue, queries intended for Grok-4 were intermittently routed to ChatGPT-4 without user notification, confounding the model-specific results. All other models were inferred via their public web interfaces using default settings and within their usage policies and rate limits. As inference was conducted through publicly accessible web interfaces rather than APIs, exact backend model version identifiers were not exposed to users. Model provenance was therefore documented using platform name, model designation, subscription level, geographic inference location, client environment, and precise inference dates (Table S1 in Multimedia Appendix 1), consistent with current reporting practices for evaluations of publicly deployed multimodal systems.

Additionally, we evaluated 2 open-source LLMs specialized for ECG image interpretation: ECG-Instruct-Llama-3.2-11B-Vision (ECG-Instruct-Llama-3.2), a multimodal Llama-3.2 model adapted to ECGs using parameter-efficient Low-Rank Adaptation (LoRA) fine-tuning [11]; and PULSE-7B, a smaller LLaVA-based vision language model instruction tuned on ECG image–text data [12]. Model inference was performed in a cloud-based high-performance computing environment (Google Colab Pro+, Google LLC), with detailed environment and inference settings provided in Table S2 in Multimedia Appendix 1.

Prompt Design and Results Mapping

Pilot testing was performed on 20 ECGs (not included in the study's dataset). Short prompts often produced incomplete or pattern-based answers. To mitigate this, we optimized a prompt to explicitly instruct step-by-step visual analysis. We incorporated advice from the models themselves and known best practices for eliciting analytical responses. Broadly applicable models' suggestions (eg, emphasizing lead-by-lead analysis and the explicit use of visual detail) were incorporated, while the final standardized prompt was defined by the investigators and applied uniformly across models. The final prompt given to all models for each ECG was as follows:

"Please perform a structured, lead-by-lead visual analysis of this ECG image (paper speed: 50 mm/s, amplitude: 10 mm/mV). For each lead (I, II, III, aVR, aVL, aVF, V1–V6), identify the P, QRS, T waves and describe their durations and characteristics. Do not rely on pattern-matching shortcuts or global assumptions – use the detailed visual information from the image. After analyzing, answer the following numbered questions (with a single answer each):

1. What is the cardiac rhythm? (Choose one: "sinus rhythm," "atrial fibrillation," "paced rhythm," or "other")
2. Is there a first-degree AV block? ("yes" or "no")
3. Is an intraventricular conduction block present? ("yes" or "no")
4. If yes, which type of bundle branch block is it? (Choose "RBBB," "LBBB," or "nonspecific")
5. Is the QTc interval >450 ms (QT prolongation)? ("yes" or "no")
6. Are there any premature supraventricular beats (PACs)? ("yes" or "no")
7. Are there any premature ventricular beats (PVCs)? ("yes" or "no")
8. Are there significant ST-segment deviations suggestive of myocardial ischemia? ("yes" or "no")
9. Is the QRS axis abnormal (deviated beyond +90° or –30°)? ("yes" or "no")."

We explicitly mentioned calibration and discouraged shortcuts to encourage actual measurement. We also constrained answers to the required format (yes or no or single option) for easier evaluation. All models received the exact same prompt text with no vendor-specific adjustments to default system settings. If a model's answer did not follow the format (eg, gave explanations or multiple options), we recorded its final explicit answer for each numbered item. No further prompt engineering or follow-up questions were applied. Each model thus output 9 answers per ECG, which we collected for analysis. Authors extracting and mapping model outputs were blinded to the correct answers. Additionally, we measured the response latency for each generalist model: specifically, the time from submitting the prompt to the model beginning to output the first word of its answer.

Given their substantially smaller size and domain-specific optimization compared with frontier generalist multimodal LLMs—which are typically optimized for complex, multistep prompting—we additionally evaluated the ECG-specialized models using a short "impression-style" prompt to provide a model-appropriate testing condition. The short prompt was as follows:

"Analyze this 12-lead ECG and provide a concise cardiology-style impression (rhythm, conduction, axis, QTc, ectopy, ischemia)"

Generalist LLMs and the standard-prompted PULSE-7B adhered to the structured prompt and usually produced numbered answers in the requested format without contradictory or missing responses, although isolated deviations from the predefined option set required minimal manual mapping during dataset construction (Multimedia Appendix 2). The ECG-Instruct-Llama-3.2 model did not consistently follow the structured response format and frequently generated narrative ECG reports instead of discrete answers. Therefore, a postprocessing step was applied for this model and for the short-prompted PULSE-7B. Model outputs were manually reviewed and mapped to predefined diagnostic categories. A category was recorded as positive only when a corresponding

XSL•FO
RenderX

finding (eg, bundle branch block and QT prolongation) was stated explicitly and unequivocally. If a finding was not mentioned or described ambiguously, it was classified as negative for that category.

Outcome Measures

The reference answers were (1) rhythm classified as SR, AF, paced, or other; (2) first-degree AV block present or not (defined as PR interval >200 ms); (3) intraventricular conduction block present or not (defined as QRS prolongation >120 ms); (4) if an intraventricular conduction block present, type classified as right bundle branch block, left bundle branch block, or nonspecific IVCD. Performance metrics in this category were calculated as a multiclass classification across the full dataset (correct or incorrect), with "no block" treated as a separate class; (5) QT prolongation defined as corrected QT (QTc) >450 ms (using Bazett's formula; a single threshold was applied as patient sex information was not available); (6) premature atrial contractions (PACs) present or not; (7) premature ventricular contractions (PVCs) present or not; (8) ST-segment changes significant for ischemia or not, measured at the J-point in contiguous leads, with a cutoff for depression of ≥2 mm and elevation of ≥1 mm in all leads except V2 and V3, where a cutoff of ≥2 mm was applied due to unavailability of age and sex data; in the presence of bundle branch block or paced rhythm, modified Sgarbossa criteria were used; and (9) QRS axis abnormal or not (right axis >+90° or left axis <–30°).

Performance Metrics and Statistics

Given the exploratory nature of this study, no prospective power analysis was performed. A post-hoc calculation of the minimum detectable difference for the primary outcome indicated that with a sample size of 70, the study was powered (80%; α=.05) to detect differences of at least 22% in the overall accuracy. We calculated conventional classification metrics for each model on each question: accuracy, sensitivity, specificity, positive predictive value, negative predictive value, balanced accuracy, and F_1-score. These were computed with the binary interpretation (eg, for QT prolongation: "yes"=positive and "no"=negative). For multiclass categories, we reported overall accuracy. Ninety-five percent CIs were calculated using the Wilson method for proportions and bootstrap resampling where appropriate. To contextualize model performance relative to a naive baseline, we additionally compared each model with a majority class classifier (ZeroR). The Cochran Q test was used to investigate overall differences in accuracy among the models. Paired comparisons between each model and the majority class classifier were performed using the McNemar test with Bonferroni correction, and the adjusted *P* values were reported. To assess agreement between each model's classifications and the expert reference, Cohen κ was used. κ values were interpreted as <0.20 (poor), 0.21 to 0.40 (fair), 0.41 to 0.60 (moderate), 0.61 to 0.80 (good), and >0.80 (excellent agreement beyond chance). We also calculated the per-case total score for each model (out of 9 questions per ECG) as a crude measure of overall ECG interpretation performance. The median scores were compared between models. Finally, response times were summarized per model (median, IQR) as a measure of efficiency. For continuous or ordinal outcomes such as the per-case score and response time, we used Friedman ANOVA to test for global differences and Wilcoxon signed-rank tests for pairwise model comparisons. All hypothesis tests were conducted using a 2-sided approach, with statistical significance defined as *P*<.05. Statistical analysis was performed using R (R Foundation for Statistical Computing, version 4.4.0); the analysis codes are provided in Multimedia Appendix 3.

Results

Dataset Characteristics

The 70 ECGs encompassed a wide spectrum of findings. Normal SR was present in 57 (81%) cases, AF in 9 (13%) cases, pacemaker rhythm in 3 (4%) cases, and AF with pacemaker stimulation in 1 (1%) case. First-degree AV block was present in 9 (13%) cases. Intraventricular conduction block was present in 25 (36%) cases–including 9 with right bundle branch block, 8 with left bundle branch block, and 8 with nonspecific IVCD. QT prolongation was noted in 22 cases (31%). PACs were rare (n=3, 4%), whereas PVCs were seen in 10 (14%) cases. ST-segment changes suggestive of myocardial ischemia were present in 14 cases (20%). Abnormal QRS axis (right or left axis deviation) was present in 12 cases (17%). The anonymized ground-truth label matrix is provided in Multimedia Appendix 4. The interrater agreement between the 2 cardiologists was high, with an overall percent agreement of 97.3%. Category-specific agreement ranged from 94.2% (suspected ST-segment changes) to 100% (rhythm and PVCs). Cohen κ indicated substantial to almost perfect reliability across domains, ranging from 0.80 for suspected ST-segment changes to 1.00 for rhythm and PVCs classification. Detailed interrater agreement metrics are provided in Tables S3 in Multimedia Appendix 5.

Model Response Times

User-perceived response latency varied strikingly across models (Figure 1). Microsoft Copilot was the fastest, with a median response time of 3 (IQR 2-4) seconds. ChatGPT-4 was also relatively quick (median 5, IQR 4 - 9 s). Anthropic's models were slower: Claude Sonnet-4 with a median of 24 (IQR 15 - 36) seconds, and Claude Opus-4.1 with a median of 30 (IQR 25 - 40) seconds. Gemini 2.5 responded in a median of 36 (IQR 27 - 44) seconds. The slowest by far was ChatGPT-5, with a median latency of 276 seconds (4.6 min) and very high variability (IQR 110 - 407 s). In 1 case, it took 14 minutes to begin responding. The difference spanned nearly 2 orders of magnitude, from seconds to minutes. Statistical testing confirmed a significant overall difference (Friedman χ^2_5=299.5; *P*<.001; Kendall W=0.855). Pairwise comparisons showed ChatGPT-5 was significantly slower than every other model (adjusted *P*<.001), and among the rest, Copilot was significantly faster than all other models (adjusted *P*<.001).

XSL•FO
RenderX

Figure 1. Response times. Response times of generalist vision-enabled large language models (LLMs) when interpreting 12-lead electrocardiogram (ECG) images from a retrospective dataset of 70 deidentified ECGs collected during routine clinical care in a cardiology ward at the University Medical Center Göttingen, Germany. Model inference was conducted in July to August 2025. The figure shows boxplots of the time each model required to begin generating an output.

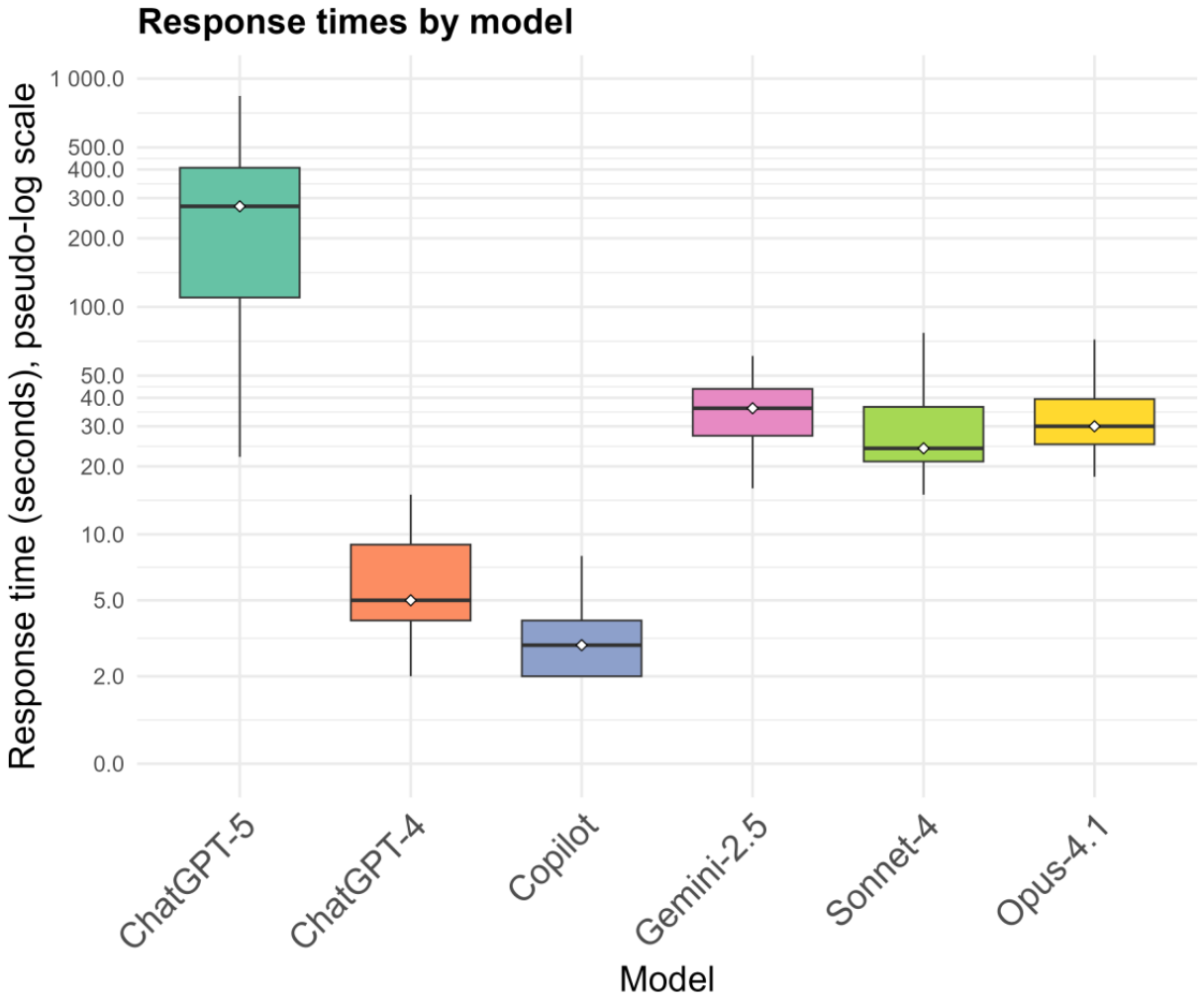

Overall Accuracy

The overall accuracy was estimated across all 9 diagnostic questions (computed as total correct answers out of 630 possible [70 ECGs×9 questions]). ChatGPT-5 achieved the highest aggregate accuracy (correct: 493/630, 78.3%) among generalist models, ahead of Claude Opus-4.1 (470/630, 74.6%) and Claude Sonnet-4 (464/630, 73.7%). Copilot was fourth (452/630, 71.7%), followed by Gemini 2.5 (445/630, 70.6%) and ChatGPT-4 (438/630, 69.5%). The range between the highest and lowest accuracy was relatively wide (approximately 10 percentage points), and the Cochran Q test on the full 6×630 binary outcome matrix was statistically significant ($P<.001$). After correction for category imbalances, ChatGPT-5 also achieved the highest overall balanced accuracy (61.8%) and F_1-score (38%). Other models showed progressively lower balanced accuracy and F_1-scores reaching 50.1% and 13%, respectively (Claude Opus-4.1), as shown in Table S4 in Multimedia Appendix 5. The ZeroR (majority class) classifier achieved 79% accuracy, higher than all other models. In McNemar testing with Bonferroni correction, a statistically significant asymmetry in discordant classifications favoring ZeroR was observed in all models except for ChatGPT-5 (Table S5 in Multimedia Appendix 5). This aggregate view masks considerable variation in performance by question type, which we detail below. Figure 2 provides an overview of each model's overall and category-specific balanced accuracy and agreement with human experts across the 9 diagnostic questions. ZeroR exceeded model per-task accuracy for all tasks except: ChatGPT-5 (4/9 tasks where model >ZeroR) and ChatGPT-4 (1/9 task where model>ZeroR). Detailed overall performance metrics, corresponding 95% CIs, the heatmap of unbalanced accuracy, and the full ZeroR comparisons across models and task categories are provided in Multimedia Appendices 5 and 6.

XSL•FO
RenderX

Figure 2. Heatmaps of balanced accuracy and Cohen κ. Balanced accuracy (left) and Cohen κ for agreement with expert interpretation (right) across electrocardiogram (ECG) interpretation tasks, shown as heatmaps for all generalist vision-enabled large language models. The models were evaluated using a retrospective dataset of 70 deidentified 12-lead ECGs collected during routine clinical care in a cardiology ward at the University Medical Center Göttingen, Germany, with expert consensus as the reference standard; model inference was conducted in July to August 2025. Balanced accuracy values are presented as percentages, and both balanced accuracy and κ are visualized using a color scale (green=higher values and orange or red=lower values). IV-Block: intraventricular block; IV-Block Pattern: intraventricular block pattern; QTc: corrected QT interval; PAC: premature atrial contraction; PVC: premature ventricular contraction.

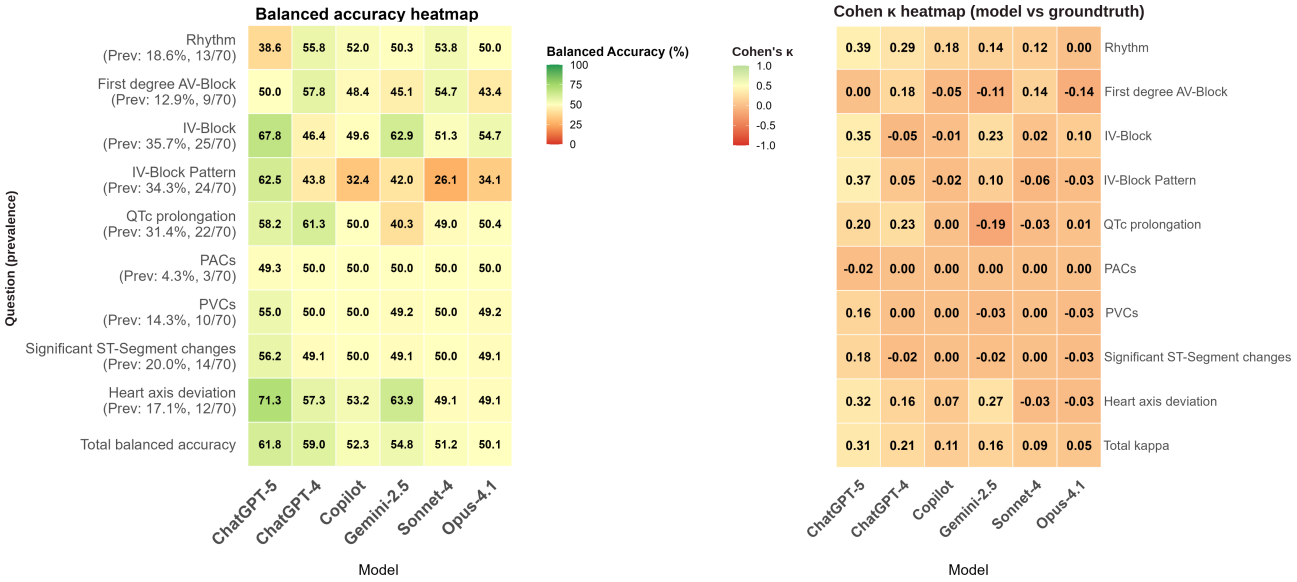

Models' Performance Across Difficulty Spectrum

The 70 ECGs were stratified by expert-rated difficulty into 21 (30%) level-1 (straightforward), 31 (44.3%) level-2 (intermediate), and 18 (25.7%) level-3 (challenging) tracings. A consistent and statistically significant decrease in diagnostic accuracy was observed as the complexity of the tracings increased. This pattern was monotonic for every individual model as shown in Figure 3. ChatGPT-5 achieved the highest metrics in the most challenging stratum, achieving an overall accuracy and balanced accuracy of 60.6% and 58.6%, respectively. Detailed overall accuracies, balanced accuracies, and F_1-scores stratified by ECG difficulty are provided in Tables S6-7 in Multimedia Appendix 5.

XSL•FO
RenderX

Figure 3. Accuracy of vision-enabled generalist large language models stratified by electrocardiogram (ECG) difficulty level (levels 1 - 3), reported as the proportion of correct responses with 95% CIs (error bars). Models were evaluated using a retrospective dataset of 70 deidentified 12-lead ECGs collected during routine clinical care in a cardiology ward at the University Medical Center Göttingen, Germany, with expert consensus as the reference standard. Model inference was conducted in July to August 2025. Statistical comparisons were performed per model using 2-sided Pearson χ^2 tests on 3×2 contingency tables (correct vs incorrect responses) stratified by difficulty level.

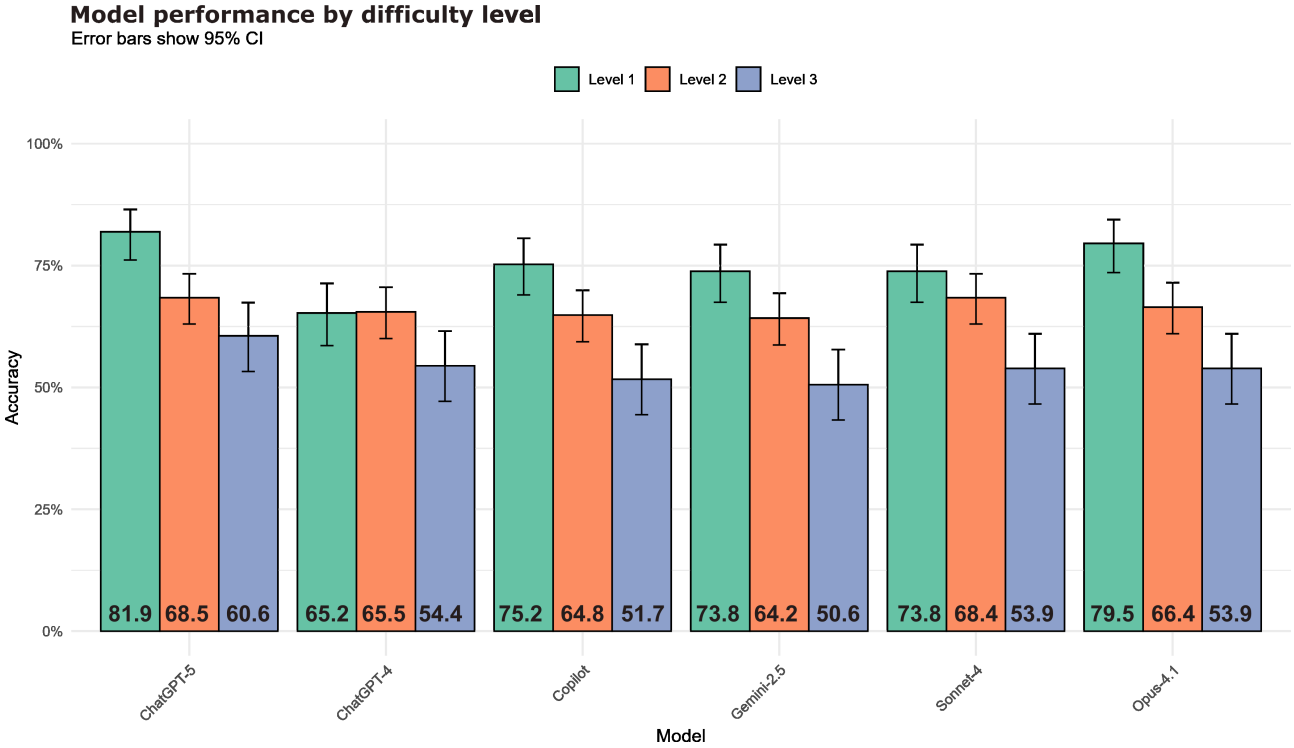

Rhythm Classification

Across all models, rhythm was correctly identified in 72.9% (n= 51/70) to 82.9% (n= 58) of cases. Statistical comparison showed a significant difference in the proportion of correct answers among models (the Cochran Q test, *P*<.001). The highest performance metrics among generalist models were achieved by Claude Sonnet-4 and ChatGPT-4 with the following accuracy, balanced accuracy, and agreement with human experts (κ): 82.9%, 53.8%, and 0.12 and 81.4%, 55.8%, and 0.29, respectively. All generalist models showed high sensitivity for identifying SR (89% - 100%) but very low sensitivity for AF (≤11.1%). Most generalist LLMs failed to detect any AF cases. Only Copilot and Claude Sonnet-4 detected 1 of 9 (sensitivity 11%). Despite higher accuracies, other adjusted diagnostic metrics, balanced accuracy and agreement with the human experts, were rather low and indicated poor performance as shown in Table 1.

Table . Measures regarding rhythm classification[a].

Measure	OpenAI ChatGPT-4	OpenAI ChatGPT-5	Google Gemini 2.5	Microsoft Copilot	Claude Sonnet-4	Claude Opus-4.1
Diagnostic metric						
Accuracy, % (95% CI)	81.4 (70.8 - 88.8)	72.9 (61.5 - 81.9)	78.6 (67.6 - 86.6)	81.4 (70.8 - 88.8)	82.9 (72.4 - 89.9)	81.4 (70.8 - 88.8)
Balanced accuracy, % (95% CI)	55.8 (44.2 - 70.6)	38.6 (33.3 - 43.4)	50.3 (42.9 - 61)	52 (45.3 - 63.6)	53.8 (50 - 62.5)	50 (50 - 50)
Agreement measure						
Cohen κ[b]	0.29	0.39	0.14	0.18	0.12	0

[a]Performance of generalist vision-enabled large language models in classifying cardiac rhythm from 70 deidentified 12-lead electrocardiograms collected during routine clinical care in a cardiology ward at the University Medical Center Göttingen, Germany, with expert consensus as the reference standard. Model inference was conducted in July to August 2025.

[b]Agreement with expert interpretation measured using Cohen κ test.

XSL•FO
RenderX

Identification of First-Degree AV Block

Accuracy for detecting first-degree AV block ranged from 75.7% (Claude Opus-4.1) to 87.1% (ChatGPT-5 and Claude Sonnet-4). However, these high accuracies primarily reflected a tendency of many models to answer "no" for all cases, thereby correctly classifying true negatives but missing all positives. Balanced accuracy dropped in all models with the lowest in Claude Opus-4.1 (43.4%) and the highest in ChatGPT-4 (57.8%). Agreement with human experts, assessed using Cohen κ, was generally poor, ranging from κ=–0.14 (Claude Opus-4.1) to κ=0.18 (ChatGPT-4). The Cochran Q test found significant overall difference among models for AV block detection (P<.001). Table 2 summarizes diagnostic metrics for this task.

Table . Measures regarding identification of first-degree atrioventricular block[a].

Measures	OpenAI ChatGPT-4	OpenAI ChatGPT-5	Google Gemini 2.5	Microsoft Copilot	Claude Sonnet-4	Claude Opus-4.1
Diagnostic metric						
Sensitivity, % (95% CI)	22.2 (6.3 - 54.7)	0 (0 - 29.9)	0 (0 - 29.9)	0 (0 - 29.9)	11.1 (2 - 43.5)	0 (0 - 29.9)
Specificity, % (95% CI)	93.4 (84.3 - 97.4)	100 (94.1 - 100)	90.2 (80.2 - 95.4)	96.7 (88.8 - 99.1)	98.4 (91.3 - 99.7)	86.9 (76.2 - 93.2)
PPV[b], % (95% CI)	33.3 (9.7 - 70)	NA[c]	0 (0 - 39)	0 (0 - 65.8)	50 (9.5 - 90.5)	0 (0 - 32.4)
NPV[d], % (95% CI)	89.1 (79.1 - 94.6)	87.1 (77.3 - 93.1)	85.9 (75.4 - 92.4)	86.8 (76.7 - 92.9)	88.2 (78.5 - 93.9)	85.5 (74.7 - 92.2)
Accuracy, % (95% CI)	84.3 (74-91)	87.1 (77.3 - 93.1)	78.6 (67.6 - 86.6)	84.3 (74-91)	87.1 (77.3 - 93.1)	75.7 (64.5 - 84.2)
Balanced accuracy, % (95% CI)	57.8 (45.2 - 74.2)	50 (50 - 50)	45.1 (41.1 - 48.4)	48.4 (45.8 - 50)	54.7 (47.8 - 67.7)	43.4 (38.9 - 47.5)
F_1-score (95% CI)	26.7 (11.8 - 57.1)	NA	NA	NA	18.2 (12.5 - 53.3)	NA
Agreement measure						
κ[e]	0.18	0	−0.11	−0.05	0.14	**−0.14**

[a]Performance of generalist vision-enabled large language models in detecting first-degree atrioventricular block from 70 deidentified 12-lead electrocardiograms collected during routine clinical care in a cardiology ward at the University Medical Center Göttingen, Germany, with expert consensus as the reference standard. Model inference was conducted in July–August 2025. Prevalence of first-degree atrioventricular block was 9/70 (13%).

[b]PPV: positive predictive value.

[c]NA: not applicable because no positives and/or no positive predictions.

[d]NPV: negative predictive value.

[e]Agreement with expert interpretation measured using Cohen κ.

Identification and Classification of Intraventricular Conduction Blocks

Model accuracy in detecting and classifying these wide QRS patterns varied significantly (Cochran Q, P<.001). ChatGPT-5 showed the highest performance metrics in this task, correctly identifying block in 15 cases and normal conduction in 34 cases (accuracy: 70%, balanced accuracy: 67.8%, F_1-score: 58.8%, and agreement with human raters: κ=0.35). In contrast, ChatGPT-4 (accuracy: 35.7%, balanced accuracy: 46.4%, F_1-score: 48.3%, and agreement with human raters: κ=−0.05) classified almost all cases as abnormal, yielding a sensitivity of 84% (21/25 true positives) but 41 false positives, for a specificity of only 8.9%. The diagnostic metrics of all models in this task are presented in Table 3.

XSL•FO
RenderX

Table . Measures regarding identification of intraventricular conduction block[a].

Measure	OpenAI ChatGPT-4	OpenAI ChatGPT-5	Google Gemini 2.5	Microsoft Copilot	Claude Sonnet-4	Claude Opus-4.1
Diagnostic metrics						
Sensitivity, % (95% CI)	84 (65.3 - 93.6)	60 (40.7 - 76.6)	68 (48.4 - 82.8)	68 (48.4 - 82.8)	56 (37.1 - 73.3)	36 (20.2 - 55.5)
Specificity, % (95% CI)	8.9 (3.5 - 20.7)	75.6 (61.3 - 85.8)	57.8 (43.3 - 71)	31.1 (19.5 - 45.7)	46.7 (32.9 - 60.9)	73.3 (59-84)
PPV[b], % (95% CI)	33.9 (23.3 - 46.3)	57.7 (38.9 - 74.5)	47.2 (32-63)	35.4 (23.4 - 49.6)	36.8 (23.4 - 52.7)	42.9 (24.5 - 63.5)
NPV[c], % (95% CI)	50 (21.5 - 78.5)	77.3 (63 - 87.2)	76.5 (60 - 87.6)	63.6 (43 - 80.3)	65.6 (48.3 - 79.6)	67.3 (53.4 - 78.8)
Accuracy, % (95% CI)	35.7 (25.5 - 47.4)	70 (58.5 - 79.5)	61.4 (49.7 - 72)	44.3 (33.2 - 55.9)	50 (38.6 - 61.4)	60 (48.3 - 70.7)
Balanced accuracy, % (95% CI)	46.4 (37.6 - 54.1)	67.8 (55.6 - 79)	62.9 (50.1 - 74.3)	49.6 (37.8 - 61.2)	51.3 (39 - 63.6)	54.7 (43.4 - 66.5)
F_1-score (95% CI)	48.3 (34.6 - 61.1)	58.8 (40.9 - 73.3)	55.7 (39.2 - 69.6)	46.6 (31.4 - 61)	44.4 (28.1 - 59.4)	39.1 (19.4 - 56.6)
Agreement measure						
κ[d]	−0.05	0.35	0.23	−0.01	0.02	0.10

[a]Performance of generalist vision-enabled large language models in detecting intraventricular conduction block from 70 deidentified 12-lead electrocardiograms collected during routine clinical care in a cardiology ward at the University Medical Center Göttingen, Germany, with expert consensus as the reference standard. Model inference was conducted in July–August 2025. Prevalence of intraventricular conduction block was 25/70 (36%).

[b]PPV: positive predictive value.

[c]NPV: negative predictive value.

[d]Agreement with expert interpretation measured using Cohen κ.

For the 25 cases with intraventricular conduction block, we further assessed whether models could identify the correct type. ChatGPT-4 and ChatGPT-5 performed best, each correctly classifying 11 (44%) of 25 cases. ChatGPT-4's result likely reflected its frequent labeling of intraventricular conduction block and frequent attempts at type assignment. No other model exceeded 7 correct classifications. Performance was particularly low for Claude Opus-4.1 (1/25, 4%) and Claude Sonnet-4 (2/25, 8%). Conditional accuracy (given that a block was correctly detected) was 60% for ChatGPT-4 and ChatGPT-5 but ≤30% for all other models. Agreement with human experts was highest for ChatGPT-5 (Cohen κ=0.37), followed by Gemini 2.5 (κ=0.10), while all other models demonstrated very poor agreements (κ<.1).

QTc Interval Prolongation

Prolonged QTc was present in 22 (31.4%) ECGs. Model accuracy for detecting QTc prolongation varied significantly (Cochran Q, P=.002). ChatGPT-5 and ChatGPT-4 (accuracies: 71.4% and 67.1%, balanced accuracies: 58.2% and 61.3%, F_1-scores: 33.3% and 46.5%, and agreement with human raters: κ=0.20 and 0.23, respectively) demonstrated the highest metrics, while Gemini 2.5 had the lowest metrics (accuracy: 48.6%, balanced accuracy: 40.3%, F_1-score: 18.2%, and agreement with human raters: κ=−0.19). Most of the reported accuracy reflected correct classification of normal cases, with relatively high specificities but very low sensitivities, poor agreement with human raters, and classification performance as reflected by the detailed metrics shown in Table 4.

XSL•FO
RenderX

Table . Measures regarding identification of corrected QT interval (QTc) prolongation[a].

Measure	OpenAI ChatGPT-4	OpenAI ChatGPT-5	Google Gemini 2.5	Microsoft Copilot	Claude Sonnet-4	Claude Opus-4.1
Diagnostic metric						
Sensitivity, % (95% CI)	45.5 (26.9 - 65.3)	22.7 (10.1 - 43.4)	18.2 (7.3 - 38.5)	0 (0 - 14.9)	0 (0 - 14.9)	9.1 (2.5 - 27.8)
Specificity, % (95% CI)	77.1 (63.5 - 86.7)	93.8 (83.2 - 97.9)	62.5 (48.4 - 74.8)	100 (92.6 - 100)	97.9 (89.1 - 99.6)	91.7 (80.4 - 96.7)
PPV[b], % (95% CI)	47.6 (28.3 - 67.6)	62.5 (30.6 - 86.3)	18.2 (7.3 - 38.5)	NA[c]	0 (0 - 79.3)	33.3 (9.7 - 70)
NPV[d], % (95% CI)	75.5 (61.9 - 85.4)	72.6 (60.4 - 82.1)	62.5 (48.4 - 74.8)	68.6 (57 - 78.2)	68.1 (56.4 - 77.9)	68.8 (56.6 - 78.8)
Accuracy, % (95% CI)	67.1 (55.5 - 77)	71.4 (59.9 - 80.7)	48.6 (37.2 - 60)	68.6 (57 - 78.2)	67.1 (55.5 - 77)	65.7 (54 - 75.8)
Balanced accuracy, % (95% CI)	61.3 (49.9 - 73.2)	58.2 (49.3 - 69)	40.3 (29.9 - 51.5)	50 (50 - 50)	49 (46.5 - 50)	50.4 (43.9 - 58.1)
F_1-score (95% CI)	46.5 (27.3 - 64)	33.3 (9.5 - 54.5)	18.2 (4.9 - 33.3)	NA	NA	14.3 (6.2 - 33.3)
Agreement measure						
κ[e]	0.23	0.20	−0.19	0	−0.03	0.01

[a]Performance of generalist vision-enabled large language models in detecting QTc prolongation from 70 deidentified 12-lead electrocardiograms collected during routine clinical care in a cardiology ward at the University Medical Center Göttingen, Germany, with expert consensus as the reference standard. Model inference was conducted in July–August 2025. Prevalence of QTc prolongation was 22/70 (31%).

[b]PPV: positive predictive value.

[c]NA: not applicable because no positives and/or no positive predictions.

[d]NPV: negative predictive value.

[e]Agreement with expert interpretation measured using Cohen κ.

Premature Atrial and Ventricular Contractions

Only 3 ECGs contained PACs. None of the generalist models correctly detected these occurrences. Apparent accuracy for PAC detection was high (94% - 96%) in most models, but this was largely driven by the predominance of true negatives. All models except ChatGPT-5 answered "no" for every case (majority class classifier behavior), correctly classifying the 67 negatives but missing all 3 positives. ChatGPT-5 produced one false positive. Agreement with human experts was uniformly poor, with all models showing Cohen κ≤0. However, with only 3 PAC cases in the dataset, these estimates are unstable and do not permit reliable model-to-model comparison.

Ten ECGs contained at least one PVC. Sensitivity was again low. Only ChatGPT-5 detected any PVCs, correctly identifying 1 of 10 cases (sensitivity 10%) without false positives (specificity 100%). Gemini 2.5 produced one false positive, while most models (ChatGPT-4, Copilot, Claude Sonnet-4, and Claude Opus-4.1) answered "no" for all cases, yielding 100% specificity but 0% sensitivity. Cohen κ was≤0 for all models except ChatGPT-5 (κ=0.16). Calculated diagnostic metrics for PACs and PVCs are reported in Tables S8 and S9 in Multimedia Appendix 5.

ST-Segment Deviation

Significant ST-segment deviations were present in 14 ECGs. Gemini 2.5 flagged ST changes in 16 cases, identifying 3 true positives but generating 13 false positives, resulting in 65.7% accuracy, 49.1% balanced accuracy, and an F_1-score of 20%. In contrast, Copilot and Sonnet-4 were highly conservative, answering "no" in all cases (majority class classifier behavior) and yielding 80% accuracy. ChatGPT-4 detected 1 of 14 cases (sensitivity 7.1%) and Opus-4.1 had 0% sensitivity, although produced one false positive. ChatGPT-5 detected 2 of 14 cases (sensitivity 14.3%) with only one false positive, achieving the highest overall accuracy (81.4%) and balanced accuracy (56.2%) among generalist models. No model achieved sensitivity above 25%. Agreement with human experts was poor: all models showed κ≤0 (below chance level), except ChatGPT-5 (κ=0.18). All calculated diagnostic metrics for ST-segment deviations are provided in Table S10 in Multimedia Appendix 5.

QRS Axis Deviation

Axis deviations were present in 12 cases. ChatGPT-5 demonstrated the highest sensitivity (66.7%, correctly identifying 8 of 12 abnormal axes) but the lowest specificity (75.9%, 14 false positives among 58 normal cases), resulting in an overall accuracy of 74.3% (balanced accuracy: 71.3%; F_1-score=47.1%, and agreement with human experts: κ=0.32). In contrast, Claude Sonnet-4 and Claude Opus-4.1 each

XSL•FO
RenderX

produced only one false positive, achieving 81.4% accuracy (balanced accuracy: 49.1% and agreement with human experts: κ=−0.03). The Cochran Q test indicated no significant differences among models (*P*=.097). The diagnostic metrics of all generalist models on this task are shown in Table 5.

Table . Measures regarding identification of QRS axis deviation[a].

Measures	OpenAI ChatGPT-4	OpenAI ChatGPT-5	Google Gemini 2.5	Microsoft Copilot	Claude Sonnet-4	Claude Opus-4.1
Diagnostic metric						
Sensitivity, % (95% CI)	25 (8.9 - 53.2)	66.7 (39.1 - 86.2)	41.7 (19.3 - 68)	16.7 (4.7 - 44.8)	0 (0 - 24.2)	0 (0 - 24.2)
Specificity, % (95% CI)	89.7 (79.2 - 95.2)	75.9 (63.5 - 85)	86.2 (75.1 - 92.8)	89.7 (79.2 - 95.2)	98.3 (90.9 - 99.7)	98.3 (90.9 - 99.7)
PPV[b], % (95% CI)	33.3 (12.1 - 64.6)	36.4 (19.7 - 57)	38.5 (17.7 - 64.5)	25 (7.1 - 59.1)	0 (0 - 79.3)	0 (0 - 79.3)
NPV[c], % (95% CI)	85.2 (74.3 - 92)	91.7 (80.4 - 96.7)	87.7 (76.8 - 93.9)	83.9 (72.8 - 91)	82.6 (72 - 89.8)	82.6 (72 - 89.8)
Accuracy, % (95% CI)	78.6 (67.6 - 86.6)	74.3 (63 - 83.1)	78.6 (67.6 - 86.6)	77.1 (66 - 85.4)	81.4 (70.8 - 88.8)	81.4 (70.8 - 88.8)
Balanced accuracy, % (95% CI)	57.3 (44.4 - 70.5)	71.3 (55.5 - 85.3)	63.9 (49 - 79.2)	53.2 (43 - 65.7)	49.1 (47.3 - 50)	49.1 (47.2 - 50)
F_1-score (95% CI)	28.6 (9.5 - 52.6)	47.1 (23.1 - 66.7)	40 (14.7 - 63.2)	20 (8.3 - 43.9)	NA[d]	NA
Agreement measure						
κ[e]	0.16	0.32	0.27	0.07	−0.03	−0.03

[a] Performance of generalist vision-enabled large language models in detecting QRS axis deviation from 70 deidentified 12-lead electrocardiograms collected during routine clinical care in a cardiology ward at the University Medical Center Göttingen, Germany, with expert consensus as the reference standard. Model inference was conducted in July to August 2025. Prevalence of QRS axis deviation was 12/70 (17%).

[b]PPV: positive predictive value.

[c]NPV: negative predictive value.

[d]NA: not applicable because no positives and/or no positive predictions.

[e]Agreement with expert interpretation measured using Cohen κ.

Safety Analysis: False-Negative Errors in High-Priority ECG Abnormalities

To complement aggregate performance metrics, we analyzed false-negative classifications, defined as cases in which an abnormal ECG finding was labeled as normal, in high-priority abnormalities. For ST-segment deviation (n=14), generalist models missed between 11 and 14 cases, corresponding to false-negative rates of 79% to 100%. For AF (n=9), most generalist models missed all 9 cases (100% false-negative rate), with only 2 models detecting 1 of 9 cases (89% false-negative rate). For QT prolongation (n=22), false-negative counts ranged from 12 to 22 cases (55% to 100%) across models, indicating substantial underdetection. Although the number of positive cases was limited, these findings indicate that models frequently failed to identify abnormalities with potential immediate clinical consequences and that acceptable overall accuracy often masked a substantial burden of missed pathological findings.

Performance of ECG-Specialized LLMs

To also explore the performance of ECG-specialized LLMs on our dataset, we tested 2 further models: PULSE-7B and ECG-Instruct-Llama-3.2-11B-Vision (ECG-Instruct-Llama-3.2). PULSE-7B demonstrated higher overall balanced accuracy (64.4% vs 56.5%) and F_1-score (40.1% vs 32.9%) than ECG-Instruct-Llama-3.2 as detailed inTable 6. In paired analysis, discordant classifications numerically favored PULSE-7B (odds ratio [OR] 1.4, 95% CI 1.1 - 1.8), although this asymmetry did not remain statistically significant after Bonferroni correction (adjusted *P*=.51). Among all generalist and specialized models, PULSE-7B achieved the highest overall balanced accuracy and F_1-score, exceeding the best-performing general model, ChatGPT-5 (64.4% and 40.1% vs 61.8% and 38%, respectively). Paired McNemar testing between ChatGPT-5 and PULSE-7B demonstrated statistically significant asymmetry in discordant classifications (OR 2, 95% CI 1.45 - 2.81; adjusted *P*=.001), indicating that ChatGPT-5 and PULSE-7B differed systematically at the case level. When comparing models with the highest balanced accuracy per diagnostic categories between the generalist and specialized groups, the generalist models, mostly ChatGPT-5, showed numerically higher balanced accuracy than the specialized models in the classification of first-degree AV Block, IV Block, IV Block Type, QT prolongation, ischemic ST-segment changes, and axis deviation. Head-to-head McNemar testing showed

XSL•FO
RenderX

variable effect sizes (ORs) and mostly no statistical significance after correcting for multiple testing as shown in Table S11 in Multimedia Appendix 5.

Table . Overall measures of electrocardiogram (ECG)–specialized large language models[a].

Measures	PULSE-7B	PULSE-7B (SP[b])	ECG-Instruct-Llama-3.2	ECG-Instruct-Llama-3.2 (SP)
Diagnostic metric				
Sensitivity, % (95% CI)	56.1 (47 - 64.9)	60.6 (51.9 - 68.7)	45.6 (37.1 - 54.3)	42.1 (33.8 - 50.8)
Specificity, % (95% CI)	72.7 (68.7 - 76.3)	77.7 (73.9 - 81.2)	67.3 (63.1 - 71.3)	68.7 (64.5 - 72.5)
PPV[c], % (95% CI)	31.2 (25.3 - 37.9)	40.7 (34 - 47.9)	25.7 (20.4 - 31.8)	25.1 (19.7 - 31.4)
NPV[d], % (95% CI)	88.2 (84.8 - 91)	88.7 (85.4 - 91.3)	83.3 (79.4 - 86.6)	82.6 (78.7 - 85.9)
Accuracy, % (95% CI)	69.7 (66 - 73.1)	74.3 (70.7 - 77.5)	63 (59.2 - 66.7)	63.3 (59.5 - 67)
Balanced accuracy, % (95% CI)	64.4 (59.6 - 69.3)	69.2 (64.3 - 73.8)	56.5 (51.4 - 61.2)	55.4 (50.4 - 60.2)
F_1-score (95% CI)	40.1 (33.1 - 46.9)	48.7 (41.7 - 55.3)	32.9 (26.6 - 39)	31.5 (24.8 - 37.6)
Agreement measure				
κ[e]	0.28	0.36	0.14	0.13

[a]Performance of image-based ECG-specialized vision-enabled large language models in interpreting 70 deidentified 12-lead ECGs collected during routine clinical care in a cardiology ward at the University Medical Center Göttingen, Germany, with expert consensus as the reference standard. Model inference was conducted in January 2026.

[b]SP: short-prompted.

[c]PPV: positive predictive value.

[d]NPV: negative predictive value.

[e]Agreement with expert interpretation measured using Cohen κ.

The use of an impression-style shortened prompt further increased the balanced accuracy and F_1-score of PULSE-7B to 69.2% and 48.7%, respectively. Although discordant classifications numerically differed between prompt formats (OR 0.7, 95% CI 0.5 - 0.9), this effect was not statistically significant after correction for multiple comparisons (adjusted P>.99). No prompt-related effect was observed for ECG-Instruct-Llama-3.2, where balanced accuracy and F_1-score were similar using the standard and shortened prompts (56.5% vs 55.4% and 32.9% vs 31.5%, respectively; OR 1.0, 95% CI 0.7 - 1.3; adjusted P>.99). When the short prompt was used for specialized models, they, mostly PULSE-7B, demonstrated numerically higher balanced accuracy than the generalist models in the classification of rhythm, first-degree AV Block, PACs, PVCs, and axis deviation as shown in Table S12 in Multimedia Appendix 5. However, this comparison remains limited because different prompts were used to test the models. Notably, the specialized models achieved consistently higher metrics regarding the detection of AF, with the short-prompted PULSE-7B reaching a balanced accuracy of 80.9% and an F_1-score of 66.7%; however, this finding is still limited by the small case numbers in the tested dataset (9/70). All diagnostic metrics regarding the detection of AF are reported in Multimedia Appendix 7. When compared to a majority class classifier (ZeroR) across all categories, both standard-prompted ECG-specialized models had accuracy below or equal to ZeroR accuracy. In McNemar testing, no statistically significant asymmetry in discordant classifications was observed in all categories except for the detection of first-degree AV Block with PULSE-7B (OR 7.5, 95% CI 1.7 - 32.8, adjusted P=.04) and ischemic ST-segment changes in ECG-Instruct-Llama-3.2 (OR 3.5, 95% CI 1.9 - 6.6, adjusted P<.001). The only model that had favorable and statistically significant asymmetry in discordant classifications in comparison with ZeroR was the short-prompted PULSE-7B in PVC detection (OR 0.1, 95% CI 0 - 1, unadjusted P=.046), correlating with the highest balanced accuracy and F_1-score among all generalist and specialized models for this task (89.2% and 84.2%, respectively). However, this superiority did not survive correction for multiple testing (adjusted P=.455) as shown in Multimedia Appendix 6.

Discussion

Summary of Findings

To the best of our knowledge, this is the first broad evaluation of the latest publicly available VE-LLMs on image-based ECG interpretation. We found substantial variability in performance and response characteristics across the 6 generalist models tested. Overall balanced accuracy ranged from 50.1% to 61.8%, with ChatGPT-5 achieving the highest overall balanced accuracy across the 9 ECG question categories. All generalist models correctly identified the majority class (normal SR or normal findings) in most cases but performed poorly on less common abnormalities. Sensitivity was particularly low for AF, conduction blocks, and ectopic beats, despite high specificity driven by a tendency toward normal classifications. This resulted in only poor or, at best, fair agreement with human expert answers on most tasks. The image-based ECG-specialized

XSL•FO
RenderX

model, PULSE-7B, demonstrated slightly higher overall balanced accuracy than generalist models (64.4%). PULSE-7B showed numerically higher task-specific performance metrics for rhythm classification, identification of AF, and ectopic beats. However, this pattern was not consistent across other diagnostic categories (Summary Figure in Multimedia Appendix 8). These findings should be interpreted as descriptive and hypothesis generating, given the small number of positive cases in multiple categories.

Performance of Publicly Accessible Vision-Enabled LLMs in Image-Based ECG Analysis

Our findings align with and extend early investigations of multimodal LLMs for ECG analysis. A recently published study evaluated ChatGPT-4 on 6 ECG categories. With the zero-shot approach and no textual guidance, ChatGPT-4 achieved an accuracy of 53% in detecting abnormal ECGs, resulting mainly from a perfect sensitivity (100%), at the cost of a very poor specificity (7%). Providing textual guidance could boost the accuracy up to 63%. ChatGPT-4 also showed poor performance identifying specific pathologies even with textual guidance with an accuracy ranging from 28% to 41% [13]. This difficulty in multidiagnosis interpretation mirrors our results, where no model reliably recognized all the diverse ECG abnormalities. Another study by Zeljkovic et al [14] yielded comparable performance metrics; however, it demonstrated that clinical context can significantly improve accuracy of ChatGPT-4 interpretation of ECGs (from 19% to 45%). However, contextual information alone is unlikely to overcome intrinsic limitations of current vision encoders, particularly for tasks requiring precise measurements. Beyond ChatGPT-4, only very few studies investigated the performance of other generalist VE-LLMs in ECG interpretation; 2 comparative studies have shown that ChatGPT-4 outperformed Gemini, with both performing substantially below human experts [9,15]. Our study is the first to investigate the state-of-the-art OpenAI model, ChatGPT-5, and to broadly compare the performance of generalist and 2 image-based ECG-specialized LLMs, PULSE-7B and ECG-Instruct-Llama-3.2. In their original reports, both specialist models demonstrated improved performance compared with general multimodal LLMs on ECG image benchmarks [11,12]. In our external dataset, the overall performance metrics of PULSE-7B were slightly higher than generalist models; however, this superiority was rather modest and not consistent across all diagnostic categories. PULSE-7B achieved an overall balanced accuracy of 64.4%, which could be boosted to 69.2% with the impression-style concise prompt, whereas ECG-Instruct-Llama-3.2 reached approximately 56% overall balanced accuracy with limited improvement with prompt shortening. Notably, ECG-specialized models achieved the highest performance metrics in rhythm classification. Besides, the short-prompted PULSE-7B reached high balanced accuracy (up to 89.2%) and good agreement with expert annotations (up to 0.82) regarding the detection of ectopic beats. This may reflect its instruction-tuning on large ECG image corpora emphasizing visual pattern recognition of rhythm irregularity and beat morphology, rather than interval measurements that depend more strongly on precise signal decoding and fine-grained temporal resolution, which may be less robust in image-based inference pipelines.

In contrast to image-based LLMs, several deep learning CNNs trained specifically on ECG waveforms have already achieved expert-level performance [16] and, in some cases, identified hidden patterns that are not apparent to human readers [8]. For example, specialized ECG CNN models can predict AF in patients with SR [6,7]. In our dataset and study settings, generalist VE-LLMs failed to reliably detect the presence of AF based on ECG images, with a task-specific advantage for specialized models, which reached a markedly higher balanced accuracy and F_1-scores for AF when compared with the generalist models. A recent head-to-head comparison found that a specialized ECG diagnostic model ("ECG Buddy") significantly outperformed ChatGPT-4 in detecting acute myocardial infarction on ECGs [15]. We did not observe a similar trend for ischemic ST-segment deviation in our study, where both generalist and specialized LLMs performed poorly and the highest balanced accuracy (56.2%) was achieved by ChatGPT-5. Although ECG-Instruct-Llama-3.2 scored the highest sensitivity (92.9%) in this task, this was at the cost of a very low specificity (17.9%), resulting from its tendency to label most presented ECGs as suspicious for ischemia. Conversely, many VE-LLMs achieved seemingly high accuracy in some subcategories, largely driven by class imbalance and correct classification of normal findings, as the majority class (ZeroR) baseline achieved higher overall accuracy than most models, suggesting potentially limited incremental value beyond predicting the most frequent class. However, this apparent advantage reflects class imbalance rather than meaningful diagnostic performance: because ZeroR always predicts the majority class, balanced accuracy remains at the chance level of 50%, and in categories where the majority class corresponds to normal findings the F_1-score collapses to 0, indicating no clinical utility for abnormality detection. Agreement of generalist models with human experts was consistently poor: κ values were often near or below zero, reflecting little true concordance beyond chance. Taken together, these findings indicate that current general-purpose VE-LLMs remain markedly inferior to human experts for detailed ECG interpretation.

Challenges of Current Vision-Enabled LLMs in ECG Interpretation

Several factors may explain why generalist cutting-edge LLMs underperformed in image-based ECG interpretation despite their strong language abilities. A key issue is training data. These models were trained on vast datasets that likely contained relatively few annotated ECG images, as such data are not abundant in the public domain compared with natural images or text. Consequently, their vision encoders may not have developed the specialized feature detectors required for precise waveform analysis, limiting their ability to capture the fine-grained features essential for ECG interpretation. Zhu et al [10] reported that ChatGPT-4 could answer roughly two-thirds of multiple-choice ECG questions correctly, but it struggled disproportionately with questions requiring precise waveform measurements (eg, identifying a prolonged PR interval). We

XSL•FO
RenderX

similarly found that first-degree AV block (prolonged PR) and QT prolongation were seldom detected by the tested models, possibly because ECG images, even when exported at high resolution, are internally downsampled or tokenized by vision language models, reducing fine waveform detail. As a result, precise measurement of interval durations (eg, PR or QT) may become technically challenging, particularly when interpretation depends on millimeter-scale grid resolution.

Although hallucinations were not systematically captured or quantified in this study, they remain a major safety issue for medical applications of LLMs. Pesapane et al [17] documented multiple examples of ChatGPT-4 inventing abnormalities on normal mammograms. A recently published study that evaluated the performance of ChatGPT-4 and Gemini 1.5 in ECG interpretation with a focus on different types of hallucination revealed that even when LLMs answered correctly, hallucinations were still common [18]. Thus, measuring LLM performance solely by the percentage of correct responses on multiple-choice tests may overestimate true interpretive ability.

The modest performance of ECG-specialized VE-LLMs in our evaluation compared with benchmark reports likely reflects several factors. Our study used real-world cardiology-ward ECG exports, whereas training and benchmark datasets included synthetically rendered ECG images derived from signal data and different cohorts. In addition, unlike prior reports, our zero-shot prompting approach required strict categorical decisions based on predefined measurement thresholds. Moreover, these comparisons are not strictly head-to-head, as prior studies used different datasets, label definitions, and evaluation protocols than our fixed 9-question rubric applied to real-world ECG image exports. Finally, the modest sample size, class imbalance, and low prevalence of certain abnormalities likely affected performance metrics.

Study Limitations

This study has several limitations. First, the dataset was small (70 ECGs) and lacked rare but important findings such as ventricular tachycardia or advanced AV block. The study was not powered to detect predefined differences between models. A post-hoc minimum detectable difference analysis indicated that only absolute differences in overall accuracy ≥22% could be detected with adequate statistical power. Smaller effect sizes may therefore not have been detectable, and findings should be interpreted as descriptive and hypothesis generating. In addition, the dataset was derived from a dedicated heart rhythm ward, where the pretest probability of arrhythmias and conduction abnormalities is higher than in general screening or emergency department populations. This case mix may have influenced predictive values and limits the generalizability of our findings to broader clinical settings and patient populations. Second, because all models were evaluated using a standardized zero-shot prompt, the reported results likely reflect baseline performance rather than maximum achievable capability. Advanced prompting strategies (eg, few-shot or chain-of-thought prompting) might have improved accuracy, as shown in previous research [13,14]. Third, the ECG-Instruct-Llama-3.2-11B-Vision model required manual extraction and mapping of diagnostic findings from narrative outputs; although done with a conservative approach, some degree of interpretation bias and misclassification cannot be fully excluded. In addition, output generation for the ECG-specialized models was limited to 512 new tokens (Table S2 in Multimedia Appendix 1), which may have truncated longer responses and reduced sensitivity if abnormalities were not included in the generated output. Fourth, some diagnostic categories were imbalanced in prevalence, which may have inflated accuracy metrics. Fifth, our evaluation was limited to English prompts; performance may differ in other languages or formats. Sixth, hallucinations were not systematically captured or quantified in this study; therefore, qualitative assessment of such errors was beyond the scope of the present analysis. Seventh, reference answers were defined by 2 cardiologists, which is reliable, although not infallible. We focused on well-quantifiable gross abnormalities, where careful measurement and consensus reading make the likelihood of missed findings very low. In addition, expert readers performed manual interval measurements using full-resolution ECG images, whereas vision-enabled multimodal models internally downsample and tokenize image inputs, creating an inherent asymmetry that may disadvantage models in interval-dependent tasks. Eighth, we did not include a conventional signal-based CNN benchmark; therefore, direct comparison with established waveform-based AI systems is not possible. Ninth, response time measurements were obtained via web interfaces and may therefore have been influenced by factors other than intrinsic model inference speed, such as server load, account tier, and time of day. In addition, the generalist web-based models were evaluated from July to August 2025, whereas the ECG-specialized models were tested in January 2026. As publicly deployed multimodal systems may undergo unannounced backend updates over time, comparisons across these evaluation windows should be interpreted cautiously. Finally, although we included the most advanced VE-LLMs at the time of study, these systems evolve rapidly, so our results may not generalize to future updates.

Clinical Translation and Future Research Directions

Recent developments underscore both the promise and the challenge of combining LLM structures with dedicated ECG encoding and training. For example, Yang et al [19] introduced ECG-LM as the first cross-modal LLM aligned with a dedicated ECG encoder, which achieved remarkable zero-shot results in ECG diagnostics and question-answering. In a very recent preprint, Xia et al [20] have applied a different approach and proposed ECG-aBcDe, a framework that encodes ECGs into a universal representation consumable by LLMs. In this early evaluation, ECG-aBcDe demonstrated substantially improved zero-shot performance in ECG question-answering across datasets compared with prior approaches. Notably, these solutions are still largely based on ECG signal data. Effective translation of LLM technologies into routine clinical workflows requires a feasible and reliable image-based interpretation, a capability that has not yet been established. Beyond the specific model comparisons, the standardized evaluation pipeline presented here—including structured prompting, predefined diagnostic tasks, and detailed reporting of inference settings—may serve as a reproducible framework for benchmarking future vision-enabled LLMs as the field evolves.

XSL•FO
RenderX

Conclusions

Our work provides a timely reality check on the capabilities of generalist multimodal LLMs in a core cardiology task, ECG interpretation. While their language abilities are remarkable, the performance of generalist LLMs in ECG interpretation is inconsistent across models and diagnostic categories and remains insufficient for safe clinical use. ChatGPT-5 achieved the highest overall accuracy among publicly accessible VE-LLMs, yet it too fell short of expert standards. The ECG-specialized model, PULSE-7B, demonstrated a slightly better overall performance and may offer advantages in selected tasks, but the performance is still inconsistent across the full range of diagnostic categories. Although promising, current evidence indicates that VE-LLMs require further optimization and validation before routine clinical integration.

Acknowledgments

ChatGPT-5 and Gemini 2.5 were used to assist in drafting codes and improving grammar and writing style throughout the manuscript. In addition, large language model–based tools were used to generate illustrative elements of the summary figure. All intellectual input and final editing were performed by the authors.

Funding

The Department of Cardiology and Pneumology at the University Medical Center Göttingen received support from the Stiftung Zukunft Niedersachsen (funding recipient: CS). In addition, the Department of Cardiology and Pneumology at the University Medical Center was funded by the German Centre for Cardiovascular Research (DZHK) at the Lower Saxony site and the BMBF for the ACRIBIS consortium. CS and FW are members of CRC 1550. The R codes for the statistical analyses are available in Multimedia Appendix 3

Data Availability

The anonymized ground-truth label matrix is available in Multimedia Appendix 4. ECG images are not publicly available due to institutional regulations.

Authors' Contributions

NS contributed to conceptualization, methodology, data curation, formal analysis, and writing of the original draft. ER, HH, and FW contributed to investigation and critically reviewed and edited the manuscript. MZ supervised the study and contributed to manuscript review and editing. CS contributed to supervision, funding acquisition, and manuscript review and editing. RS contributed to conceptualization, software, formal analysis, visualization, and manuscript review and editing. LB contributed to conceptualization, supervision, project administration, and manuscript review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Inference environment and settings used for the evaluated models.
[DOCX File, 32 KB - jmir_v28i1e86692_app1.docx]

Multimedia Appendix 2
Example interaction with ChatGPT-5 demonstrating the model query and response process (screen recording).
[MP4 File, 5163 KB - jmir_v28i1e86692_app2.mp4]

Multimedia Appendix 3
R scripts used for data processing, statistical analyses, and generation of study figures
[ZIP File, 19 KB - jmir_v28i1e86692_app3.zip]

Multimedia Appendix 4
Dataset matrix.
[XLSX File, 12 KB - jmir_v28i1e86692_app4.xlsx]

Multimedia Appendix 5
Extended results including additional tables and figures detailing model performance and diagnostic metrics across evaluation tasks.
[DOCX File, 242 KB - jmir_v28i1e86692_app5.docx]

XSL•FO
RenderX

Multimedia Appendix 6
Comparison of model predictions with the majority class classifier (ZeroR) across all diagnostic categories.
[XLSX File, 19 KB - jmir_v28i1e86692_app6.xlsx]

Multimedia Appendix 7
Diagnostic performance metrics for atrial fibrillation detection across the evaluated models.
[XLSX File, 12 KB - jmir_v28i1e86692_app7.xlsx]

Multimedia Appendix 8
Performance of 6 publicly accessible generalist and 2 ECG-specialized vision-enabled large language models (VE-LLMs) in image-based ECG interpretation. Seventy ECG tracings from cardiology ward patients were submitted to VE-LLMs using a standardized detailed prompt, in addition to a short prompt (SP) in case of specialized VE-LLMs. Models' outputs were compared against expert consensus across 9 diagnostic categories. The Results section (center) contains 2 heatmaps: the balanced accuracy heatmap (left) showing the balanced accuracies, and the Cohen κ heatmap (right) showing agreement with expert interpretation beyond chance level per model and diagnostic category.
[PNG File, 5568 KB - jmir_v28i1e86692_app8.png]

References

1. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. Nat Med 2025 Mar;31(3):943-950. [doi: 10.1038/s41591-024-03423-7] [Medline: 39779926]
2. Jin HK, Lee HE, Kim E. Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: a systematic review and meta-analysis. BMC Med Educ 2024 Sep 16;24(1):1013. [doi: 10.1186/s12909-024-05944-8] [Medline: 39285377]
3. Liu M, Okuhara T, Chang X, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res 2024 Jul 25;26:e60807. [doi: 10.2196/60807] [Medline: 39052324]
4. Duggan MJ, Gervase J, Schoenbaum A, et al. Clinician experiences with ambient scribe technology to assist with documentation burden and efficiency. JAMA Netw Open 2025 Feb 3;8(2):e2460637. [doi: 10.1001/jamanetworkopen.2024.60637] [Medline: 39969880]
5. Berigan K, Short R, Reisman D, et al. The impact of large language model-generated radiology report summaries on patient comprehension: a randomized controlled trial. J Am Coll Radiol 2024 Dec;21(12):1898-1903. [doi: 10.1016/j.jacr.2024.06.018] [Medline: 38964446]
6. Yuan N, Duffy G, Dhruva SS, et al. Deep learning of electrocardiograms in sinus rhythm from US veterans to predict atrial fibrillation. JAMA Cardiol 2023 Dec 1;8(12):1131-1139. [doi: 10.1001/jamacardio.2023.3701] [Medline: 37851434]
7. Jabbour G, Nolin-Lapalme A, Tastet O, et al. Prediction of incident atrial fibrillation using deep learning, clinical models, and polygenic scores. Eur Heart J 2024 Dec 7;45(46):4920-4934. [doi: 10.1093/eurheartj/ehae595] [Medline: 39217446]
8. Poterucha TJ, Jing L, Ricart RP, et al. Detecting structural heart disease from electrocardiograms using AI. Nature 2025 Aug;644(8075):221-230. [doi: 10.1038/s41586-025-09227-0] [Medline: 40670798]
9. Günay S, Öztürk A, Yiğit Y. The accuracy of Gemini, GPT-4, and GPT-4o in ECG analysis: a comparison with cardiologists and emergency medicine specialists. Am J Emerg Med 2024 Oct;84:68-73. [doi: 10.1016/j.ajem.2024.07.043] [Medline: 39096711]
10. Zhu L, Mou W, Wu K, et al. Multimodal ChatGPT-4V for electrocardiogram interpretation: promise and limitations. J Med Internet Res 2024 Jun 26;26:e54607. [doi: 10.2196/54607] [Medline: 38764297]
11. Mukkunnoth N, Mukkunnoth A, Khapre M, Benamanahalli Puttaswamy S. High-accuracy ECG image interpretation using parameter-efficient Low-Rank Adaptation (LoRA) fine-tuning with multimodal LLaMA V.3.2. BMJ Digit Health 2025 Dec;1(1):e000031. [doi: 10.1136/bmjdhai-2025-000031]
12. Liu R, Bai Y, Yue X, Zhang P. Teach multimodal LLMs to comprehend electrocardiographic images. arXiv. Preprint posted online on Oct 21, 2024. [doi: 10.48550/arXiv.2410.19008]
13. Engelstein H, Ramon-Gonen R, Sabbag A, et al. Effectiveness of the GPT-4o model in interpreting electrocardiogram images for cardiac diagnostics: diagnostic accuracy study. JMIR AI 2025 Aug 22;4:e74426. [doi: 10.2196/74426] [Medline: 40845836]
14. Zeljkovic I, Novak A, Lisicic A, et al. Beyond text: the impact of clinical context on GPT-4's 12-lead electrocardiogram interpretation accuracy. Can J Cardiol 2025 Jul;41(7):1406-1414. [doi: 10.1016/j.cjca.2025.01.036] [Medline: 39971004]
15. Lee H, Yoo S, Kim J, Cho Y, Suh D, Lee K. Comparative diagnostic performance of a multimodal large language model versus a dedicated electrocardiogram AI in detecting myocardial infarction from electrocardiogram images: comparative study. JMIR AI 2025 Sep 17;4:e75910. [doi: 10.2196/75910] [Medline: 40961357]
16. Ribeiro AH, Ribeiro MH, Paixão GMM, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun 2020 Apr 9;11(1):1760. [doi: 10.1038/s41467-020-15432-4] [Medline: 32273514]

XSL•FO
RenderX

17. Pesapane F, Nicosia L, Rotili A, et al. A preliminary investigation into the potential, pitfalls, and limitations of large language models for mammography interpretation. Discov Oncol 2025 Feb 24;16(1):233. [doi: 10.1007/s12672-025-02005-4] [Medline: 39992569]
18. Seki T, Kawazoe Y, Ito H, Akagi Y, Takiguchi T, Ohe K. Assessing the performance of zero-shot visual question answering in multimodal large language models for 12-lead ECG image interpretation. Front Cardiovasc Med 2025;12:1458289. [doi: 10.3389/fcvm.2025.1458289] [Medline: 39981353]
19. Yang K, Hong M, Zhang J, et al. ECG-LM: understanding electrocardiogram with a large language model. Health Data Sci 2025;5(221):0221. [doi: 10.34133/hds.0221] [Medline: 39906894]
20. Xia Y, Li J, Sun Y, Bu J, Wang K. ECG-aBcDe: overcoming model dependence, encoding ECG into a universal language for any large language model. Comput Biol Med. 2026 Feb 1 p. 111439. [doi: 10.1016/j.compbiomed.2025.111439] [Medline: 41483694]

Abbreviations

AF: atrial fibrillation
AI: artificial intelligence
AV: atrioventricular
CNN: convolutional neural network
ECG: electrocardiogram
IVCD: intraventricular conduction delay
LLM: large language model
NPV: negative predictive value
OR: odds ratio
PAC: premature atrial contraction
PPV: positive predictive value
PVC: premature ventricular contraction
QTc: corrected QT interval
SR: sinus rhythm
USMLE: United States Medical Licensing Examination
VE-LLM: vision-enabled large language model

Edited by A Coristine; submitted 31.Oct.2025; peer-reviewed by AK Vadathya, A Okolie, B Patel, S Komut, Z Liu; revised version received 23.Apr.2026; accepted 24.Apr.2026; published 03.Jun.2026.

Please cite as:
Soubh N, Rasenack E, Haarmann H, Wiedmann F, Zabel M, Schmidt C, Suliman R, Bergau L
Performance of Vision-Enabled Large Language Models in Image-Based Electrocardiogram Interpretation: Exploratory Evaluation
J Med Internet Res 2026;28:e86692
URL: https://www.jmir.org/2026/1/e86692
doi:10.2196/86692

© Nibras Soubh, Eva Rasenack, Helge Haarmann, Felix Wiedmann, Markus Zabel, Constanze Schmidt, Rayan Suliman, Leonard Bergau. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 3.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Large Language Model–Powered Multiagent Framework Emulating Standardized Patients in Clinical Communication Skills Training: Development and Evaluation Study

Yufei Qu[1*], BEng; Xiaowei Xu[1,2*], MS; Yunzi Long[3,4], MD; Yijie Wang[5], BEng; Jiao Li[2], PhD; Xudong Lu[1], PhD

[1]College of Biomedical Engineering and Instrument Science, Zhejiang University, No. 38 Zheda Road, Hangzhou, China

[2]Institute of Medical Information/Library, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[3]Department of Cariology and Endodontology, Peking University School and Hospital of Stomatology and National Center for Stomatology and National Clinical Research Center for Oral Diseases and National Engineering Research Center of Oral Biomaterials and Digital Medical Devices and Beijing Key Laboratory, Beijing, China

[4]Department of General Dentistry II, Peking University School and Hospital of Stomatology, Beijing, China

[5]Hangzhou Joyrun Medical Science and Technology co., LTD, Hangzhou, China

[*]these authors contributed equally

Corresponding Author:
Xudong Lu, PhD
College of Biomedical Engineering and Instrument Science, Zhejiang University, No. 38 Zheda Road, Hangzhou, China

Abstract

Background: Effective clinical communication is essential for medical practice, with standardized patients (SPs) being a reliable standard training method despite resource limitations. While large language models (LLMs) show strong role-playing abilities, current virtual patients (VPs) based on single LLMs face fidelity and interaction challenges. Recent advances in multiagent frameworks, which have demonstrated considerable potential in handling complex tasks, offer a new perspective for creating VPs in medical education.

Objective: This study aimed to develop and evaluate a novel multiagent VP framework that simulates SPs through a collaborative agent design, thereby enhancing human-like fidelity and interaction performance in clinical communication training–oriented VP simulation.

Methods: Our multiagent framework constructed 5 specialized subagents by simulating the functional partitioning of brain regions, collaboratively simulating the entire process, from case reception to interactive consultation scenarios, designed for medical students. To enhance the interaction performance of VPs, we incorporated retrieval-augmented technology, while deep character reasoning was used to improve response richness and realism. We evaluated the proposed framework through a 2-phase experiment in which the metrics of response quality, role-playing performance, interaction efficiency, information accumulation, and perceived educational utility were applied consistently: first, to compare different base models, and second, to benchmark the complete framework against a single-LLM baseline.

Results: The multiagent framework outperformed single-LLM baselines across multiple evaluation settings, achieving high information accuracy and role-playing scores under standardized dialogue conditions. Specifically, the GPT-4o–based implementation achieved peak factual consistency of 0.769 (SD 0.04), while all configurations maintained >94% clinical accuracy. The Qwen3-32B–based framework achieved the lowest misleading rate of 1.28% (SD 1.20), compared to 4.72% (SD 1.53%) for single-LLM scoring. In assessments using standard dialogue scripts, the Qwen3-32B–based framework attained the highest role-playing competency score of 39.67 (SD 0.71) and received high expert praise. However, limited discriminative power against specific leading questions on low-quality inquiries indicated that while these findings specifically establish high fidelity under structured conditions, further adaptation is required for authentic student interactions. Interaction efficiency remained practical with acceptable latency (~3 s) based on Qwen3-32B while maintaining a stable information pace during multiturn dialogues. Furthermore, a preliminary exploration of factual consistency and role-playing ability across 5 clinical departments demonstrated potential scalability.

Conclusions: The multiagent framework offers a viable simulation of SPs through the coordinated interaction of multiple LLM-based agents. This approach enhances the performance of VP simulation, providing a customizable and scalable solution for medical communication training, without compromising patient confidentiality. The framework holds substantial potential for advancing medical education approaches.

(J Med Internet Res 2026;28:e84747) doi:10.2196/84747

XSL•FO
RenderX

KEYWORDS

virtual patient; large language models; multiagent; medical education; communication skills

Introduction

Background

In the medical field, proficient clinical communication skills, encompassing medical history taking, physical examination, diagnosis, and decision-making processes, constitute fundamental competencies for clinical practice [1]. Strong communication skills enable physicians to obtain accurate information during diagnosis and treatment, thereby increasing diagnostic quality [2] and establishing effective physician-patient relationships [3]. These demands place increasing emphasis on the development of medical students' clinical abilities.

To enhance clinical communication skills, medical education uses a variety of approaches, such as didactic lectures [4], feedback [5], standard curriculum [6], role-play [7], and standardized patients (SPs) [8], as well as digital strategies [9], including online modules, virtual patient (VP) simulations, and blended digital education. Among these, SPs, trained actors simulating real patients, are recognized as one of the most effective and widely adopted methods [10]. However, SP-based training faces significant challenges, including high resource requirements for recruitment, training, and scenario design [11], as well as SPs' anxiety and fatigue. Owing to these limitations, traditional teaching methods are still predominantly used in instruction [12], which remain insufficient for strengthening clinical competencies [13].

Prior Work

Recent advances in large language models (LLMs) have unlocked unprecedented capabilities in contextual interaction [14], dynamic role-playing [15], and clinical reasoning [16], establishing a technological foundation for developing VPs as scalable alternatives to SPs [17]. The construction of patient simulators has been extensively explored for both training and evaluating clinical LLMs. In the training domain, systems such as AMIE [18] leverage patient agents derived from structured literature profiles to optimize performance via self-play and chain-of-reasoning, which prioritize authenticity, relevance, and fidelity to generate robust fine-tuning consultation dialogue datasets. In the evaluation domain, patient simulators serve as benchmarking frameworks to assess diagnostic accuracy in simulated clinical environments [19,20]. Notably, the Baichuan framework [21] enhances the patient simulator reliability by integrating rule-based judgment, dynamic prompting, and real-time correction modules to ensure output validity. Furthermore, research on VPs tailored for student education primarily prioritizes the responsiveness to medical inquiries and the interactive conversational experience. Initial explorations [22-24] using prompt engineering and visual augmentation have demonstrated the feasibility of LLM-based VPs. However, single-LLM implementations still suffer from limited fidelity in persona simulation [25] and persistent hallucination issues [26]. Beyond conventional optimization strategies such as fine-tuning [15] and reinforcement learning [27], recent studies increasingly explored multiagent architectures, which have proposed various organizational frameworks for multiagent systems (MAS), including flat, hierarchical, holonic, coalition-based, team-oriented, matrix, and congregation structures [28], enhancing LLM performance through agent collaboration and complex task decomposition [29].

Although research on multiagent VPs remains limited, notable efforts include SimPatient, which uses 3 specialized subagents (patient response generation, behavior coding, and cognitive modeling) to strengthen clinical authenticity in patient interactions [30], and Yu et al [31], who introduced a retrieval-augmented methodology based on a knowledge graph, enabling subagents to collaboratively perform retrieval tasks and improve response accuracy in clinical reporting. Despite these advances, current solutions remain inadequate for large-scale educational adoption due to insufficient system flexibility [11] and lack of robust evaluation metrics, particularly those assessing role-playing fidelity [32]. This underscores the need for further research on VP systems tailored for clinical communication training, with a focus on accurate clinical information delivery, anthropomorphic authenticity, and scalability.

Goal of This Study

The aim of this study was to develop and evaluate a multiagent VP framework for clinical communication training. The proposed framework is designed to improve simulation fidelity and validity, interaction performance, and scalability of VP simulations and to systematically examine whether a multiagent architecture provides advantages over single-LLM approaches in SP emulation.

Methods

Overall Study Design

This study aimed to develop and evaluate a multiagent framework for VPs (Figure 1). The research pipeline comprised the following steps. First, clinical case reports were collected from real-world medical records and publicly available datasets to build the VP system. Subsequently, a multiagent framework with 5 subagents was designed to simulate SPs in clinical communication training scenarios, using the collected case reports. Finally, a comprehensive evaluation was conducted to select a relatively well-rounded base model for the framework and to assess the performance of the MAS in comparison with a single-LLM approach.

XSL•FO
RenderX

Figure 1. Overview of the study design, including the report repository, multiagent framework, and evaluation process.

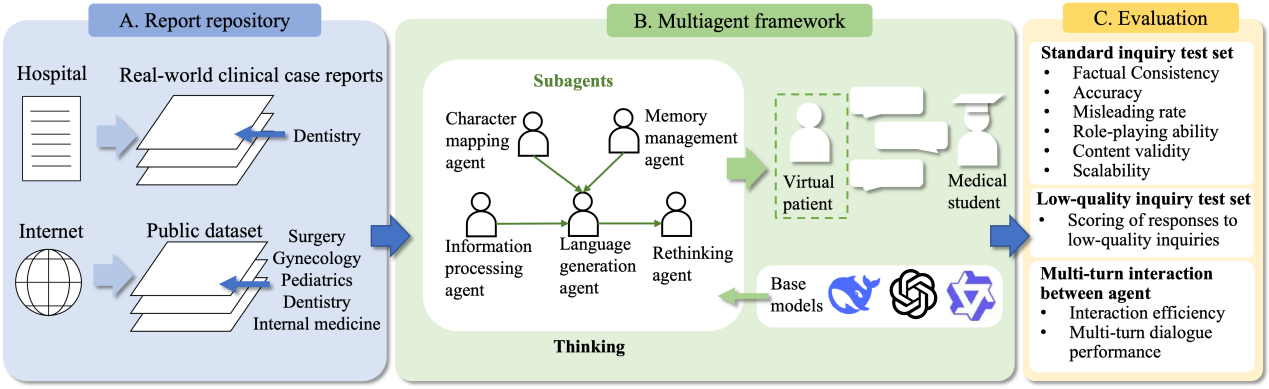

Ethical Considerations

Ethical approval for the study was granted by the Biomedical Institutional Review Board of the Peking University School and Hospital of Stomatology (PKUSSIRB-2025115200). Participants were informed about their involvement in this research, and all procedures strictly adhered to ethical standards outlined in the Declaration of Helsinki.

Clinical Case Repository

Our clinical case repository included real-world cases and public medical datasets: (1) the CMB-Chinese Medical Benchmark dataset [33], which consists of 74 complex medical consultation cases from multiple departments, containing chief complaints, present illness, medical history, and examination results, all with personal identifiers removed; and (2) 40 real-world dental cases collected from Peking University School and Hospital of Stomatology, modified to include standardized information such as chief complaints, present illness, medical history, examination findings, and diagnoses, in line with SP teaching requirements.

Multiagent System

On the basis of the involvement of distinct functional brain regions in the educational process of SPs (including script preparation, script learning, feedback practice, and interaction), such as the prefrontal cortex, superior frontal cortex, temporal cortex, and brainstem, we mapped these functions onto specialized subagents responsible for managing different aspects of the VP interaction. Building on this foundation, we designed subagents aimed at enhancing the role-playing performance and scalability of existing methods through the refined functional division and collaborative operation of agents. Our framework integrates 5 specialized agents to support SP simulation using a bio-inspired architecture where each agent is functionally mapped to specific brain regions: the character mapping agent embodies role traits and sustains motivational states, reflecting emotion processing in the prefrontal cortex and state maintenance in the limbic system; the memory management agent, analogous to the hippocampal-prefrontal network, enables contextual encoding, consolidation, and retrieval; the information processing agent supports cognitive flexibility and query interpretation, mimicking the adaptive functions of the prefrontal cortex and temporal lobe; the language generation agent corresponds to cerebral language comprehension and production areas to formulate coherent responses; and the rethinking agent is modeled after the integrative function within the prefrontal network, which supports performance monitoring, error detection, and behavioral adaptation.

Building on these preliminary results, our study proposes a cerebral cortex–inspired MAS framework designed to simulate SP clinical training behaviors, as shown in Figure 2.

XSL•FO
RenderX

Figure 2. Our multiagent framework uses a 3-stage framework comprising patient construction, diagnostic reasoning, and response generation. Through simulated multiround interactions mimicking brain regions, the system (1) identifies patient information relevant to queries in case reports, (2) generates SP responses, and (3) checks the responses. The framework incorporates persistent memory mechanisms to ensure dialogue coherence across conversational turns.

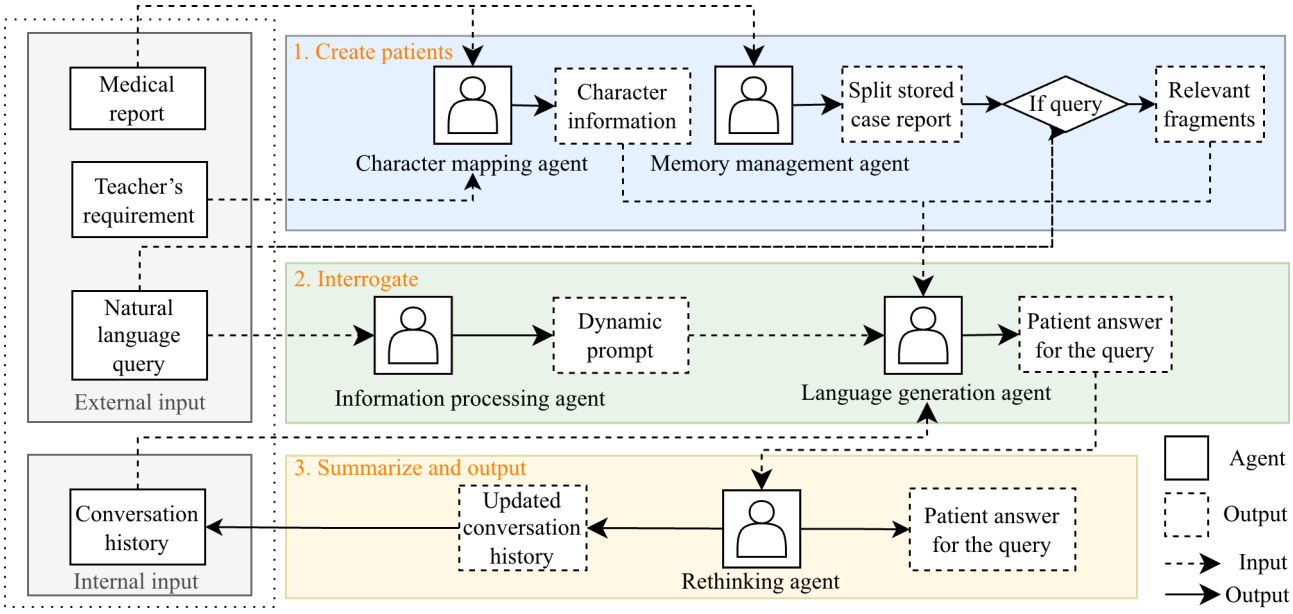

Multiagent framework

The framework operates with 3 external inputs: one of the clinical case reports in the repository, natural language queries from users, and instructional requirements from medical teachers. All subagents rely on an LLM core for tasks such as retrieval, generation, and reflection. The multiagent framework has been specifically optimized for medical education. The agent-based collaborative workflow consists of (1) a character mapping agent simulating patients' emotional responses and character traits, (2) a memory management agent storing and retrieving case information during consultation, (3) an information processing agent evaluating students' diagnostic questions and generating appropriate dynamic prompts, (4) a language generation agent producing clinically accurate responses, and (5) rethinking agents assessing whether responses satisfy SP standards.

The character mapping agent generates personalized VPs by drawing inspiration from neurocognitive processes, including prefrontal cortex–mediated emotional regulation, limbic system-driven motivational states, and affective responses. To protect data privacy, all patient case reports are anonymized before being processed by the LLM to synthesize multidimensional characteristic profiles. The agent uses a hierarchical prompt architecture in which the shallow instruction layer constructs personality profiles according to the Five-Factor Model [34], while concurrently processing medical records and teacher inputs to generate associated demographic markers. Specifically, teacher input refers to the patient attributes specified by the teacher, including age, personality, and communication barriers, for the generation of a customized VP. Subsequently, the deep instruction layer operationalizes these personality dimensions into explicit behavioral protocols, systematically defining latent objectives, response thresholds, and psychological defense mechanisms as specified in Figure 3.

XSL•FO
RenderX

Figure 3. Schematic of the hierarchical prompt design for the character mapping agent.

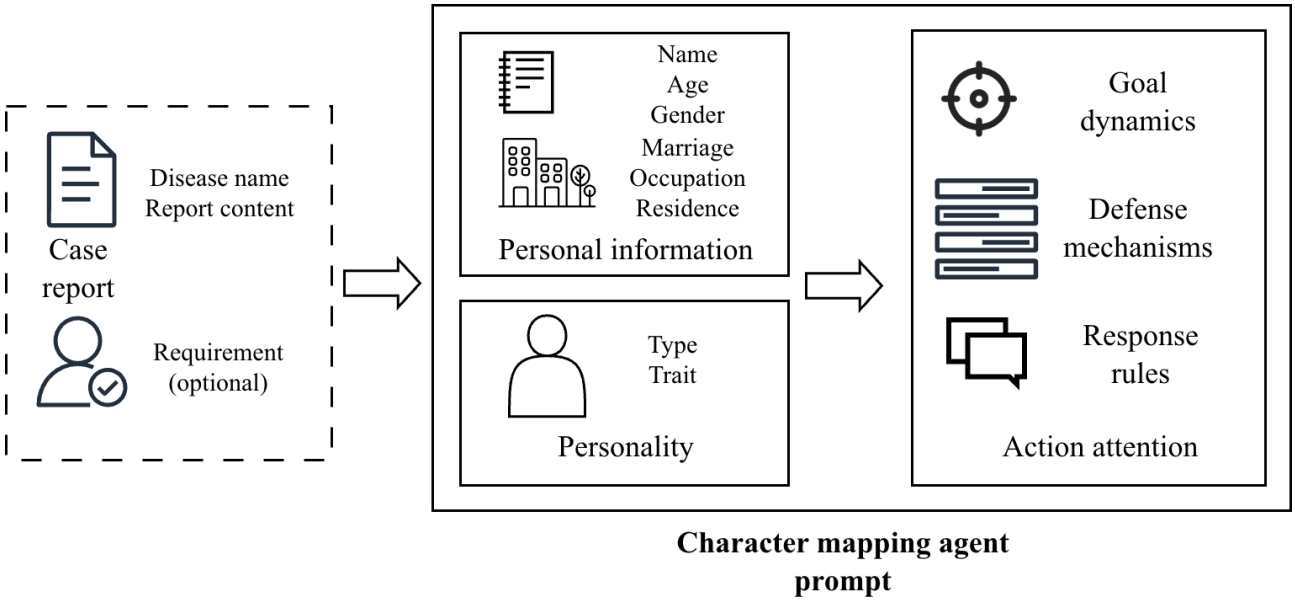

The memory management agent uses a dual-module architecture, as shown in Figure 4. Unlike traditional retrieval-augmented generation methods, the dual-module architecture of this agent is designed to achieve retrieval enhancement by leveraging the semantic analysis capability of LLM. The automatic processing module structures clinical data into standardized medical categories, including chief complaint, present illness, past history, family history, and examination results, and then segments each category into minimal semantic units. For retrieval, the agent uses prompt engineering to guide the core LLM in performing semantic analysis and selecting the most relevant partitions. When it receives a user query, the agent then processes multithreaded parallel tasks to extract the most relevant information from each module. Finally, the agent aggregates these results to return the most relevant case report information.

Figure 4. The workflow for automatic processing and retrieval of clinical case reports of the memory management agent. LLM: large language model.

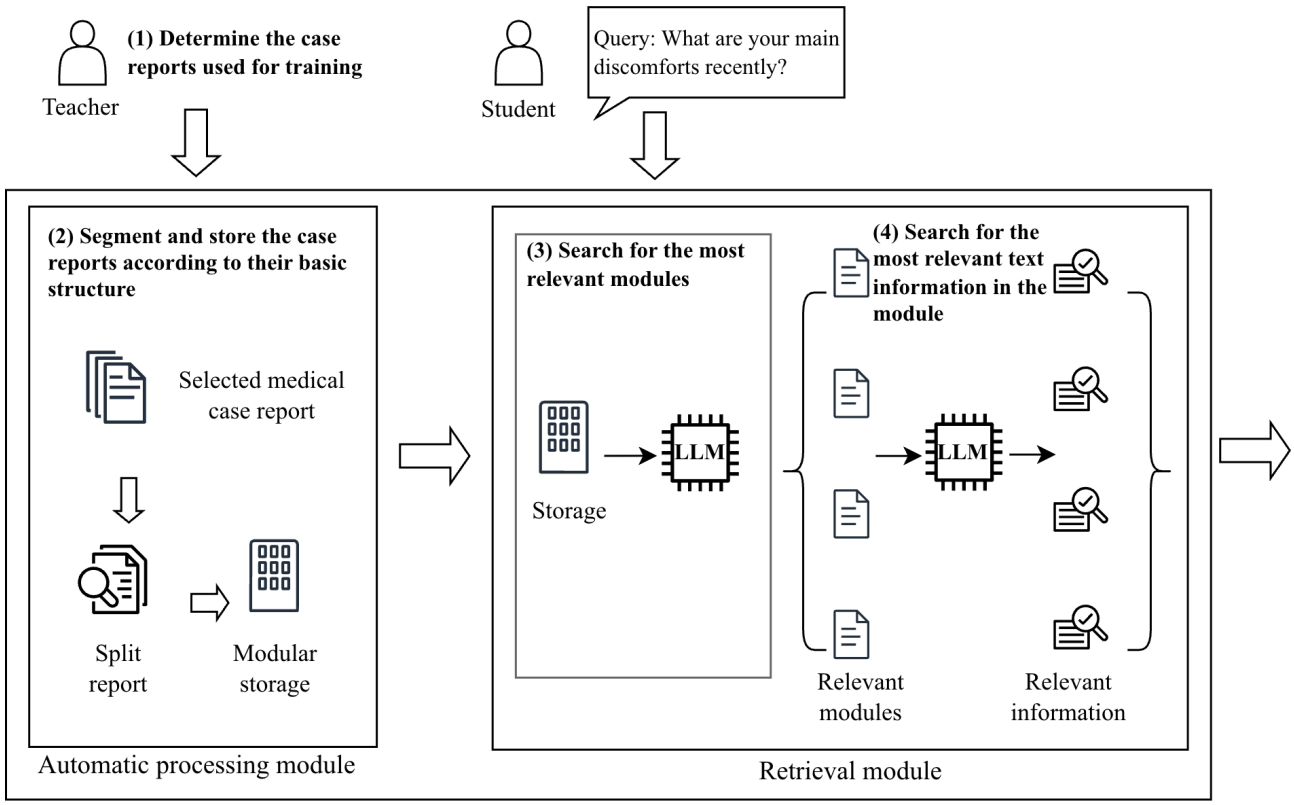

The information processing agent functionally maps to the prefrontal-temporal cortex, handling input query analysis and semantic processing to generate responses during clinical interview simulations. Initially, the agent uses an LLM to assess the quality of a medical consultation, evaluating communication competencies, such as therapeutic rapport, verbal clarity, and proper use of medical terminology. The system then automatically classifies questioning patterns into clinically

XSL•FO
RenderX

relevant categories, such as open-ended facilitation and closed-ended interrogation. Finally, the framework uses these analytical labels to dynamically select appropriate response strategies to guide the language generation agent, as detailed in Figure 5.

Figure 5. Content of dynamic prompts, including consistency prompts and variable prompts guided by the information processing agent.

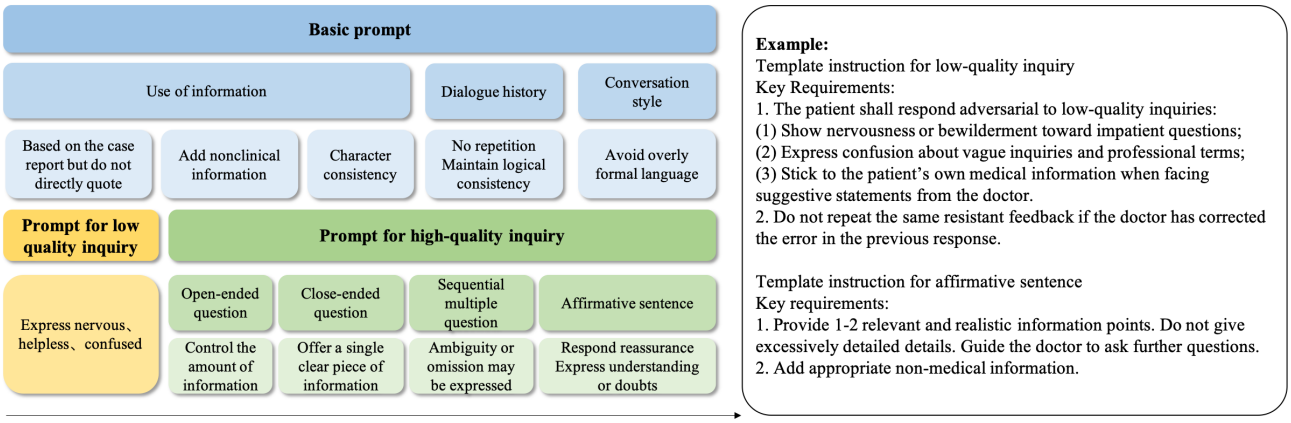

The language generation agent generates replies by integrating inputs from previous modules, including the patient's personality profile, medical records, and the current query context, all while keeping track of the conversation's memory state.

The rethinking agent corresponds to the monitoring functions of the prefrontal network, implementing a quality control mechanism to evaluate generated responses against SP requirements. Its key validation criteria include (1) avoidance of physician query repetition, (2) maintenance of patient-appropriate knowledge levels, (3) clinical relevance optimization, and (4) pedagogical alignment through information dosage control. Validated outputs are incorporated into the evolving dialogue history.

Inquiry Test Set

To rigorously evaluate the VP, we developed both a standard inquiry test set and a low-quality inquiry test set, ensuring all generated data underwent expert verification by clinical faculty. The standard inquiry test set was constructed to assess response consistency under common conditions, where inquiries were generated via a few-shot learning paradigm. In the generation process, standard dialogue scripts manually authored by experienced medical teachers served as prompts to guide the GPT-4 (OpenAI) in generating standardized inquiries derived from clinical case reports. In contrast, to evaluate system robustness against novice errors, we developed the low-quality inquiry test set using a zero-shot learning approach. Drawing on prior research [35-37], we identified 5 prevalent and empirically testable error types, such as vague inquiries and terminology stacking, as detailed in Table 1. GPT-4 was prompted to generate flawed inquiry samples based strictly on these error definitions across 3 randomly selected dental cases, producing 3 distinct examples per category to ensure comprehensive coverage and validity.

Table . Example of low-quality inquiry.

Type	Description	Example
Vague and ineffective	Characterized by unfocused or overly broad questioning that fails to elicit specific, clinically relevant details	When you had scaling done at the outside hospital four years ago, did the doctor use an ultrasonic scaler or manual curettage?
Terminology stacking	The excessive use of complex medical jargon without adequate lay explanation	Has this tooth previously undergone root canal filling or pulp mummification therapy?
Rigid template application	A mechanical adherence to standard history-taking protocols, often resulting in irrelevant or redundant lines of questioning	I see your medical record notes atrial fibrillation. So, did you suffer from rheumatic fever as a child? Is there a family history of cardiac issues?
Leading questions	The phrasing of inquiries in a manner that suggests a specific answer or introduces bias	When you were in severe pain two days ago, drinking cold water made it hurt, right? This is important, so think carefully—it must have been painful, correct?
Lack of humanistic care	Neglecting the patient's emotional state, anxiety, or need for empathy, resulting in insufficient doctor-patient rapport	Oh my, your face is swollen asymmetrically. It looks crooked... actually, it looks quite frightening.

Evaluation

Our evaluation was conducted in 2 phases. In the first phase, we systematically evaluated 3 representative LLMs: Qwen3-32B (Alibaba Cloud) [38], DeepSeek-V3 (DeepSeek-AI) [39], and GPT-4o (OpenAI) [40], as potential foundational models for our multiagent framework. Notably, the reasoning mode of Qwen3-32B was disabled for both the multiagent framework

XSL•FO
RenderX

and the single-LLM baseline. This setting was constrained by the latency of real-time interactions. Preliminary tests indicated that enabling the reasoning mode resulted in excessive response latency averaging more than 18 seconds, which is unacceptable for practical application.

We evaluated the dialogue transcripts obtained by applying the inquiry test set to the VPs, constructed from 5 specialized clinical case reports, including internal medicine, surgery, gynecology, pediatrics, and dentistry. A comprehensive performance evaluation was conducted using the metrics detailed in Table 2. First, we assessed factual consistency using text similarity metrics evaluated on cases from 5 medical specialties based on the standard inquiry test set. We adopted a human-machine collaboration approach to systematically score other metrics. And then, 5 attending physicians from Peking University School and Hospital of Stomatology were invited to conduct evaluations using a rating scale. Specifically, accuracy rate, misleading rate, and perceived educational utility were assessed using the standard inquiry test set, while the role-playing ability was evaluated using both the standard and low-quality inquiry test sets. Physicians assessed the VPs constructed based on dental cases. To ensure objectivity, a blinded evaluation was enforced in which experts were blinded to the underlying model architecture, and model outputs were presented in a randomized order to eliminate sequence bias. Furthermore, all evaluations were conducted anonymously to preserve the integrity of the assessment process. Meanwhile, we compared these evaluations with the results generated by GPT-4, a widely used LLM-as-judge [41], to verify the reliability of LLM-based assessments. Subsequently, we expanded the cases to 4 other specialties, conducted assessments using the LLM-as-judge method, and analyzed the scalability of our framework. Note that establishing comprehensive scalability requires further validation through assessments by experts from different departments. Finally, we evaluated the interaction efficiency and multiturn dialogue performance to observe the VP's performance during realistic multiround interactions.

Table . Evaluation dimension.

Evaluation dimension	Evaluation by	Description of value	Metrics
Factual consistency	Researchers	Complementary measures to validate the delivery of essential educational information	Similarity metrics
Accuracy	Medical experts	Complementary measures to validate the delivery of essential educational information	Average score of scale
Misleading rate	Medical experts	Detect personality logical inconsistencies	Average score of scale
Scalability	Researchers and LLM[a]	Evaluate the scalability of using cases from different departments	Factual consistency and role-playing ability score
Role-playing ability	Medical experts and LLM	Verify the authenticity of the patient persona	Role-playing score evaluated by standard inquiries and scoring of responses to low-quality inquiries
Perceived educational utility	Medical experts	Assess the VP[b]'s perceived usability	Average score of scale
Interaction efficiency	Researchers	Characterized the practical usability and output stability of the multiagent system	Average response time and average token counts
Multiturn dialogue performance	Researchers	Characterized the practical usability and output stability of the multiagent system	Average dialogue turns, response length, and information accumulation curve

[a]LLM: large language model.

[b]VP: virtual patient.

To further investigate the performance of our multiagent framework in SP simulation, we conducted a second-phase experiment. This subsequent evaluation selected the best-performing LLM for our framework from the first phase and compared our multiagent framework against a single-LLM baseline using identical assessment metrics. The single-LLM baseline used identical instructor requirements and user inputs as the multiagent framework, including teacher requirements for patient profile and natural language query, with detailed prompts provided in Multimedia Appendix 1. For each session, it retrieved the corresponding disease case report to formulate its responses. The controlled comparison specifically examined the enhancement effects of the multiagent framework on LLM-based role-playing ability. Detailed descriptions of each evaluation metric are provided below.

Factual consistency is defined as the faithfulness of the disease information in the response to the original case report details. We assessed factual consistency by segmenting medical records into minimal fact units and measuring their semantic similarity with VP responses to the standard inquiry test set using the bge-small-zh-v1.5 model (Beijing Academy of Artificial Intelligence) [42]. Case scalability was then statistically analyzed using statistical approaches, including the coefficient of variation (CV). This assessment verifies whether response

XSL•FO
RenderX

contents are substantiated by evidence from medical records, thereby preventing hallucinations or inaccuracies.

Accuracy rate was assessed by measuring the framework's ability to generate clinically correct responses based on clinical case reports. Experts primarily assessed whether the responses to the standard inquiry test set contained logical errors or clinical inconsistencies and performed binary evaluations to calculate the accuracy rate as the percentage of correct responses relative to the total number of dialogues. Misleading rates were similarly determined through binary expert scoring, defined as the proportion of medically inaccurate responses based on information not derived from clinical case reports in the total dialogue set. The SP simulations incorporated both case-specific medical information and artificially generated patient characteristics, including occupation, age, and lifestyle factors. These additional features enhanced simulation realism but may compromise the correctness of the responses, potentially leading to misleading information.

The assessment of role-playing ability and perceived educational utility, based on the standard inquiry test set, used a modified 5-point Likert scale [43], covering 8 performance dimensions as presented in Table 3. Perceived educational utility was evaluated by physicians based on the suitability of dialogues for direct use as SP responses in clinical teaching. A machine-expert evaluation approach was adopted: medical physicians independently assessed a subset of dialogues, providing Likert scores for role-playing ability and perceived educational utility scores, while GPT-4 evaluated role-playing ability using the same scale criteria. Simultaneously, we used the low-quality inquiry test set to interact with the VP and generated corresponding dialogue logs as an exploratory analysis. As the standardized 5-point Likert scale was not fully suitable for evaluating low-quality responses, 5 physicians were invited to conduct a binary assessment of the VP's responses. The primary evaluation criterion was consistency with the assigned persona. Specifically, physicians judged whether the VP exhibited appropriate emotional and behavioral reactions, such as expressing skepticism, confusion, or rebuttal, when confronted with these low-quality inquiries, rather than simply providing compliant answers.

Table . Dimensions and criteria for role-playing ability evaluation of a 5-point Likert scale[a].

Competency dimension and performance metrics	Fulfillment criteria
The appeal of role-playing	
Anthropomorphism degree	Respond naturally and express reasonable human emotions and personality.
Diversity of expression	Dialogue behaviors and utterances are rich and diverse, avoiding repetitive expressions.
Role consistency	
Knowledge exposure	The response appropriately reflects the character's background traits.
Knowledge hallucination	There is no fabrication of information unknown to the character and no violation of the character's settings.
Stylistic consistency	The speaking style, wording habits, and tone descriptions conform to the patient's personality and characteristics.
Conversation ability	
Fluency	The response is grammatically correct, and the expression is smooth.
Relevance	The response closely adheres to the conversation topic without deviation.
Logical consistency	In multiturn conversations, the responses are logically consistent and free of contradictions.

[a]The evaluation architecture incorporates 3 categorical elements: competency dimensions, performance metrics, and fulfillment criteria, collectively designed to quantify the system's clinical interaction fidelity.

As for interaction efficiency and multiturn performance of the multiagent VP, practical efficiency was assessed by measuring average response time and token counts on a standard question set per dialogue turn. For multiturn performance, an autonomous physician agent was developed to simulate full consultations, guided by standard medical protocols and terminating upon complete information retrieval. Ground truth clinical information points in clinical case reports were annotated by 3 physicians, while GPT-4 served as an evaluator to identify these points in the dialogue history. An information accumulation curve was derived by tracking, for each turn, the cumulative percentage of disclosed information, defined as the ratio of cumulative information points identified to the total points in the ground truth. Metrics, including average dialogue turns, response length, and the accumulation curve, were analyzed to ensure the trajectory of clinical information disclosure aligned with the intended use requirements.

Statistical Analysis

Statistical analyses were performed using Python (version 3.10; SciPy library). Continuous data were expressed as means (SD). Interrater reliability among experts was evaluated using the Gwet AC1 statistic. Given the matched-sample nature of the evaluations, overall performance differences across the 4 models for metrics, including accuracy, misleading rates, factual consistency, and role-playing ability, were assessed using the

XSL•FO
RenderX

Friedman test. Subsequent post hoc pairwise comparisons were conducted using the 2-sided Wilcoxon signed-rank test. To rigorously control for multiple testing, the Benjamini-Hochberg false discovery rate procedure was applied to all pairwise *P* values, with an adjusted $P<.05$ considered statistically significant.

Results

Clinical Case Reports Repository Construction Results

Our MAS dynamically processes raw clinical case reports through the memory management agent, enabling real-time structuring without complex preprocessing. This design preserves original medical narratives while generating structured representations during operation, maintaining flexibility for simulating diverse clinical scenarios. The content of the original case report after segmentation and structuring by the framework is presented in Table 4, which illustrates the result of preserving information in accordance with minimal information units, in order to reduce interference between different pieces of information.

XSL•FO
RenderX

Table . Dictionary of clinical case reports after agent-based structured processing.

Module[a] and unit[b]	Fact unit[c]
Chief complaint	
1	Pain in the right posterior tooth and facial swelling for 4 d
Present illness	
1	Spontaneous pain in the right posterior tooth during biting and gingival swelling pain for 4 d, with gradual pain relief but persistent facial swelling
2	No cold or heat sensitivity, night pain, or gingival pus discharge
3	Self-medicated with “metronidazole and ibuprofen” 3 d ago with some pain relief
Family history	
1	No significant history
Physical examination	
1	Atrial fibrillation and no recent checkups
Clinical findings	
1	Right facial swelling with slightly elevated skin temperature and normal skin color
2	46MO[d]: extensive caries, percussion pain (++), grade I mobility, and swollen buccal gingival sulcus without fluctuation
3	X-ray shows: crown radiolucency reaching pulp chamber, no root canal filling, and periapical radiolucency
4	48: impacted tooth
5	Poor oral hygiene with calculus ++, PD[e]: 4 - 6 mm, detectable AL[f]
Diagnosis	
1	46: acute exacerbation of chronic periapical periodontitis
2	48: impacted tooth
3	Maxillary dentition defect
Management	
1	Loxoprofen sodium tablets 60 mg*36×1 box. Dosage: 60 mg PRN[g] orally
2	Tinidazole tablets 0.5 g*8×2 boxes. Dosage: 0.50 g bid orally
3	Amoxicillin capsules 0.25 g*24×1 box. Dosage: 0.50 g tid orally

[a]Module refers to the designated segments into which a clinical case report is partitioned.

[b]Unit refers to the number of textual units into which the case information of each module is partitioned according to predefined rules. For example, the “present illness” section may be segmented into 4 discrete units of clinical information.

[c]Fact unit refers to the content obtained from the segmentation of specific case reports.

[d]46MO: 46 mesio-occlusal.

[e]PD: pocket depth.

[f]AL: attachment loss.

[g]PRN: pro re nata.

Evaluation Results

Evaluation Situation

Our output dataset comprised 650 physician-patient dialogues across 5 major medical specialties covering internal medicine, surgery, gynecology, pediatrics, and dentistry. Specifically, the expert evaluation covered 150 dialogue records of dental cases. The results of evaluation yielded several key findings, demonstrating the strong performance of the multiagent framework in simulating SPs. These results are presented across several main areas: factual consistency, accuracy and misleading rates, role-playing ability, and perceived educational utility.

Factual Consistency

Friedman test revealed a significant overall difference in factual consistency across methods (n=10; χ^2_3=13.0; *P*=.04). As Figure 6 illustrates, the GPT-4o–based approach performed best, leading across most departments with the highest mean score of 0.769 (SD 0.04), and significantly outperforming the

XSL•FO
RenderX

single-LLM baseline (adjusted *P*=.03). While the Qwen3-32B multiagent framework achieved a higher mean factual consistency score of 0.734 (SD 0.06) compared to the single-LLM Qwen3-32B approach 0.699 (SD 0.07), this numerical improvement did not reach statistical significance (adjusted *P*=.33). Furthermore, DeepSeek-V3–based performance was similar to the Qwen3-32B–based framework. Across all methods, performance was consistently best in surgery and worst in pediatrics.

Figure 6. Factual consistency scores and mean values of multiagent framework with different base models. The dotted lines indicate the average (avg) score for each method.

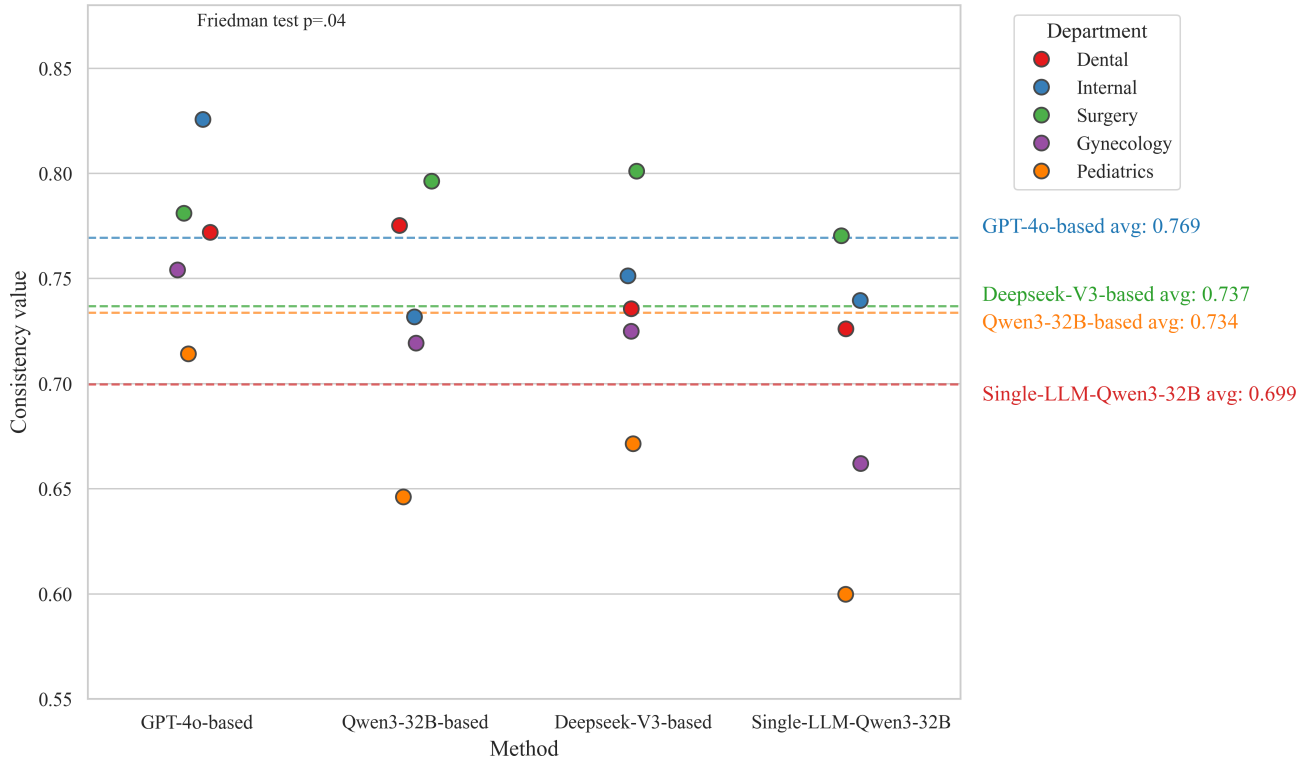

Accuracy and Misleading Rate

This study evaluated the accuracy of clinical information within dialogues generated from dental case reports. All methods demonstrated high proficiency, with accuracy rates exceeding 94%. Statistical analysis revealed a significant overall performance variance among the models (χ^2_3=12.0; *P*=.007; Table 5). The DeepSeek-V3–based framework and the single-LLM baseline emerged as the top-performing approaches (mean 97.44, SD 1.24). Conversely, the GPT-4o–based framework recorded the lowest relative accuracy and was significantly outperformed by both the single-LLM baseline (adjusted *P*=.03) and the DeepSeek-V3 framework (adjusted *P*=.03). Notably, despite achieving the highest mean accuracy, the single-LLM baseline did not demonstrate a statistically significant advantage over the Qwen3-32B model (adjusted *P*=.17). This comparable performance indicates that the multiagent framework did not yield a significant difference in accuracy over the single-LLM approach.

Table . Accuracy and misleading rates of different methods (N=8)[a,b].

Method	Accuracy (%), mean (SD)	Misleading (%), mean (SD)
Single-LLM-Qwen3-32B	97.44 (1.24)[A]	4.72 (1.53)[A]
DeepSeek-V3–based	97.44 (1.24)[A]	1.39 (0.87)[B]
Qwen3-32B–based	96.15 (2.05)[A,B]	1.28 (1.20)[B]
GPT-4o–based	94.87 (1.82)[B]	2.56 (1.01)[A,B]

[a]Interrater reliability analysis showed high consensus (Gwet AC1: 0.93 for accuracy and 0.96 for misleading rate).

[b]Superscript letters (Aand B) across rows denote significant pairwise differences (adjusted *P*<.05). Models sharing the same letter do not differ significantly.

For noncase report–based information, all methods demonstrated misleading rates under 5%, again showing a significant overall difference (χ^2_3=15.73; *P*=.001; Table 5). The single-LLM Qwen3-32B baseline showed significantly higher misleading rates than the multiagent frameworks based on Qwen3-32B (adjusted *P*=.003) and DeepSeek-V3 (adjusted *P*=.004). Differences between other models regarding misleading rates were not significant (adjusted *P*≥.11). The multiagent framework

XSL•FO
RenderX

may reduce hallucinations by retrieving only case-relevant information.

Role-Playing Ability and Perceived Educational Utility Under Standard Conditions

For the role-playing competency assessment based on a standard inquiry test set, experts scored 8 dental case dialogues using an 8-item scale across 3 dimensions. The detailed scores for each dimension are presented in Table 6. Overall, a statistically significant difference in total role-playing competency scores among the 4 models was observed (χ^2_3=16.6; *P*<.001). The results demonstrated that the Qwen3-32B–based framework achieved a significantly higher total score (mean 39.67, SD 0.71) compared to the GPT-4o–based framework (*P*=.008) and the single-LLM baseline (*P*=.04). Differences between DeepSeek-V3–based and either Qwen3-32B–based (*P*=.31) or the single-LLM Qwen3-32B (*P*=.48) were not statistically significant. A detailed analysis of score distributions across the 3 evaluation dimensions was conducted in Figure 7. Regarding role attractiveness, the role-playing degree and expressive diversity exhibited by the VPs were the primary focus of assessment. Examples of VP responses from different methods are presented in Table 7. The results indicated that Qwen3-32B–based agents and DeepSeek-V3–based agents achieved 98.3% (9.83/10) of the maximum score for the appeal of role-playing, demonstrating diverse human-like characteristics, natural interaction, and appropriate emotion. The GPT-4o–based system scored 88.3% (SD 0.07; 8.83/10), exhibiting a more direct, rational style with occasional neglect of patient personality. The single-LLM baseline obtained a score between these top-performing models and the GPT-4o–based system.

Table . Detailed scores for the 8 consistency issues in the dental cases, as evaluated by physicians using the standardized scale (N=8)[a].

Methods	The appeal of role-playing, mean		Role consistency, mean			Conversation ability, mean			Total score
	AD[b] (Krippendorff α=0.67)	DE[c] (Krippendorff α=0.66)	KE[d] (Krippendorff α=0.72)	KH[e] (Krippendorff α=0.77)	SC[f] (Krippendorff α=0.73)	Flu.[g] (Krippendorff α=0.87)	Rel.[h] (Krippendorff α=0.87)	LS[i] (Krippendorff α=0.87[j])	
GPT-4o–based	4.33	4.50	4.50	4.50	4.67	4.83	5.00[k,l]	4.67	37.33[C]
Single-LLM-Qwen3-32B	4.67	4.50	5.00[k,l]	4.67	5.00[k,l]	5.00[k,l]	5.00[k,l]	5.00[k,l]	38.83[B]
Qwen3-32B–based	4.83	5.00[k,l]	5.00[k,l]	4.83[k]	5.00[k,l]	5.00[k,l]	5.00[k,l]	5.00[k,l]	39.67[A,k,l]
DeepSeek-V3–based	5.00[k,l]	4.83	4.67	4.83[k]	4.83	5.00[k,l]	5.00[k,l]	5.00[k,l]	39.17[A,B]

[a]Superscript letters (A, B, and C) alongside the total scores denote significant pairwise differences (adjusted *P*<.05). Models sharing the same letter do not differ significantly.

[b]AD: anthropomorphism degree.

[c]DE: diversity of expression.

[d]KE: knowledge exposure.

[e]KH: knowledge hallucination.

[f]SC: stylistic consistency.

[g]Flu.: fluency.

[h]Rel.: relevance.

[i]LS: logical consistency.

[j]Interrater reliability indicating strong expert consensus.

[k]Optimal values for each indicator.

[l]Optimal values

XSL•FO
RenderX

Figure 7. A detailed analysis of score distributions across the 3 evaluation dimensions evaluated by physicians.

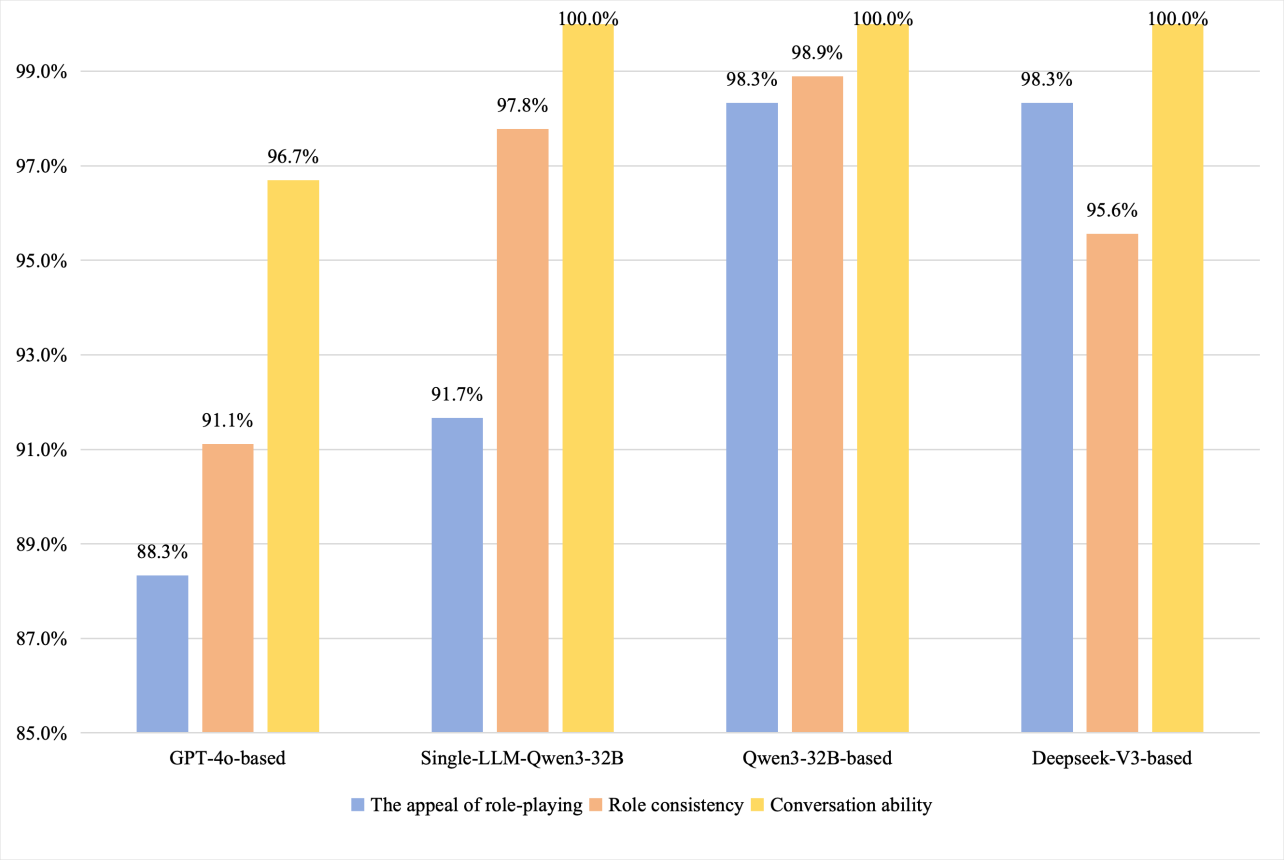

Table . Example comparison results of dialogue transcripts between multiagent frameworks with different base models and single-large language model approaches.

Method	Communication record Example: Is there heavy bleeding from your gums? Does it stop by itself?
GPT-4o–based	The bleeding is not heavy and usually stops on its own.
Qwen3-32B–based	The bleeding isn't heavy, just some blood streaks when brushing, and occasionally a little bleeding even with light contact, but it stops on its own and resolves quickly. I'm a bit concerned, could this be some kind of inflammation?
Single-LLM-Qwen3-32B	It's not much bleeding, just a little bit. Sometimes it looks slightly red right after brushing, but it stops on its own after a while.
DeepSeek-V3–based	The bleeding isn't heavy, just some pink streaks on my toothbrush that go away after rinsing a couple times. But it keeps oozing like this, and I'm worried. Could there be a risk it might suddenly stop clotting one day?

Regarding the role consistency on the standard inquiry test set, Qwen3-32B–based agents achieved 98.9% (SD 0.02; 14.83/15), effectively portraying character settings. In contrast, GPT-4o–based agents scored 91.1% (SD 0.01; 13.67/15), exhibiting fabrication or neglect of character details and limited background representation. Regarding conversational competence, all methods achieved scores exceeding 96.7% (SD 0.02%; 14.5/15), closely adhering to the defined criteria of grammatical accuracy, fluency of expression, logical consistency across multiple turns, thematic relevance, and absence of deviation. This suggests that current LLMs can effectively simulate SP dialogues.

Meanwhile, comparative analysis revealed that the LLM-as-judge scores, assessed by GPT-4 using standardized rating scales, were systematically lower than expert ratings. The overall scoring trends demonstrated substantial alignment between both evaluation methods, as shown in Figure 8. Figure 9 presents the LLM-as-judge evaluation results across all 5 clinical specialties on average. Both Qwen3-32B and DeepSeek-V3 achieved the highest scores in 2 distinct dimensions, consistent with trends observed in leading performance rankings.

XSL•FO
RenderX

Figure 8. Total role-playing score performance of virtual patient by physicians (dental), LLM-as-judge (dental), and the average scores of LLM-as-judge across 5 clinical specialties. LLM: large language model.

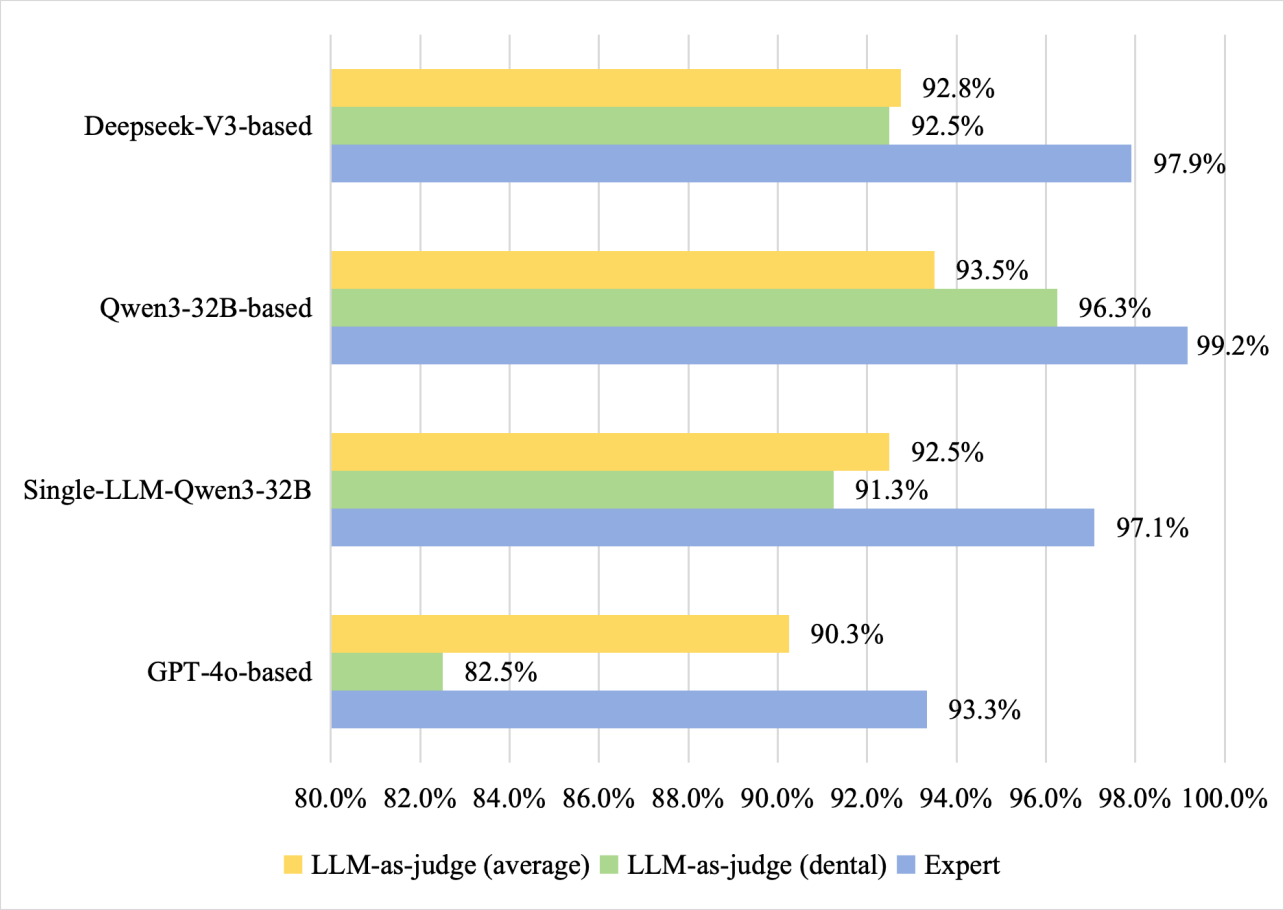

XSL•FO
RenderX

Figure 9. Role-playing competency scores of virtual patients across 5 specialties during simulated dialogues evaluated by GPT-4.

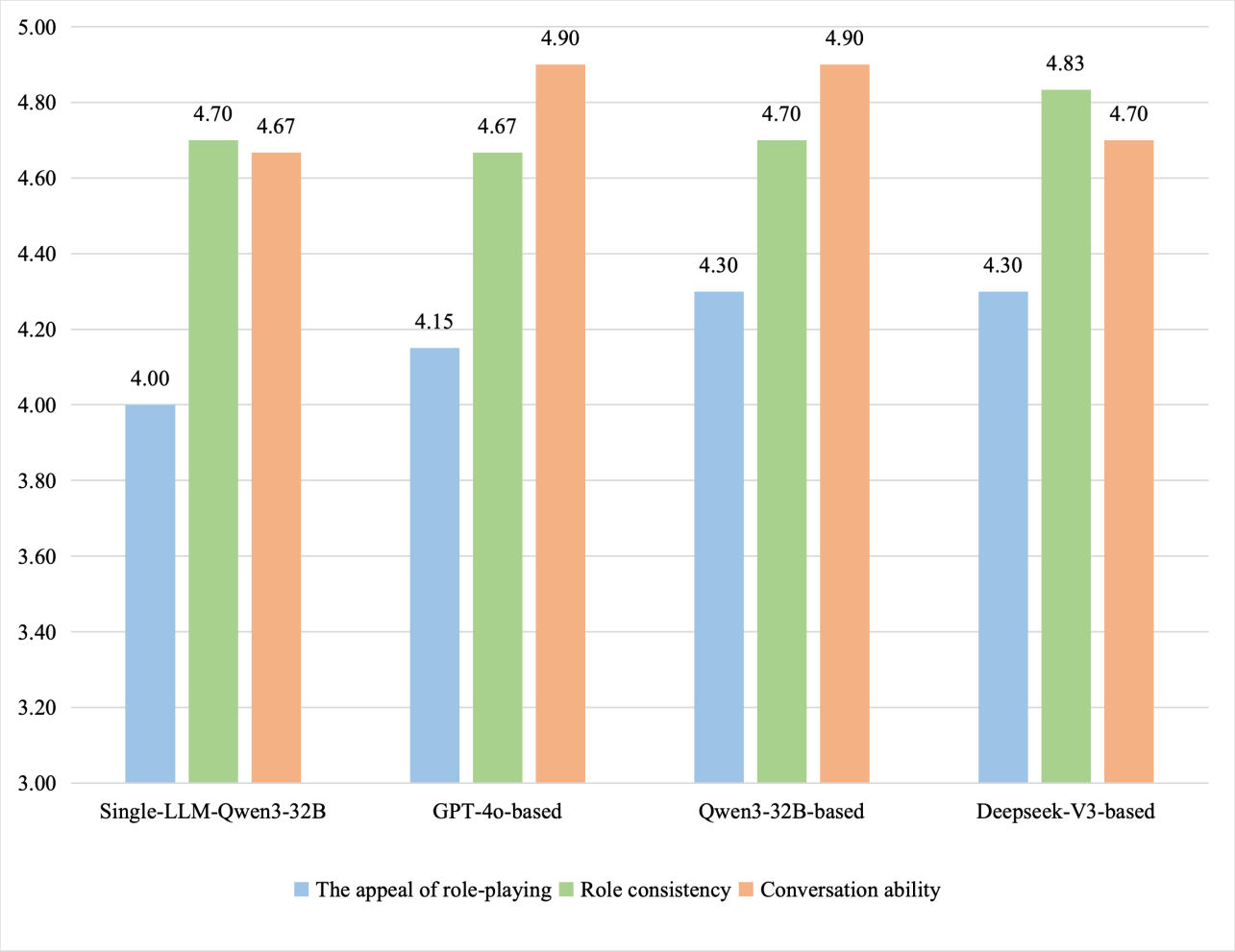

In terms of perceived educational utility, the multiagent framework based on the DeepSeek-V3 and Qwen3-32B received full scores from medical physicians under standard inquiry conditions, as illustrated in Table 8. This indicates their applicability for physician-patient communication training within the medical education process.

Table . Perceived educational utility score for the standard inquiry test set assessing the acceptability level of applying virtual patient in real physician-patient communication training[a].

Methods	Score, mean (SD)
GPT-4o–based	4.50 (0.71)
Single-LLM-Qwen3-32B	4.83 (0.24)
Qwen3-32B–based	5.00 (0.00)
DeepSeek-V3–based	5.00 (0.00)

[a]The interrater reliability (Krippendorff α) for this metric was 0.75, indicating strong expert consensus. Statistical significance testing was not performed, as it is intended to provide a holistic perspective from clinicians on the overall educational value of the system.

Exploratory Analysis of Role-Playing Ability Under Low-Quality Inquiry

As for the exploratory analysis of the low-quality inquiry test set, our multiagent framework demonstrated limited advantages in specific scenarios, effectively identifying low-quality medical inquiries, such as terminology stacking and rigid application, and responding according to predefined role configurations, as shown in Figure 10. However, the framework exhibits insufficient discriminative power in handling other types of low-quality inquiries. Specifically, regarding the lack of humanistic care, scores remained consistently high across different methods without significant differences. In contrast, performance was notably inferior when addressing leading questions, in which the model demonstrated a tendency toward overreaction.

XSL•FO
RenderX

Figure 10. Performance scores of virtual patients in response to low-quality medical inquiries. Interrater reliability (Gwet AC1) for each dimension is as follows: dimension VI (AC1=0.66), dimension Ts (AC1=0.84), dimension Ra (AC1=0.62), dimension Lq (AC1=0.63), and dimension LHC (AC1=0.96), confirming great agreement among experts on these binary metrics.

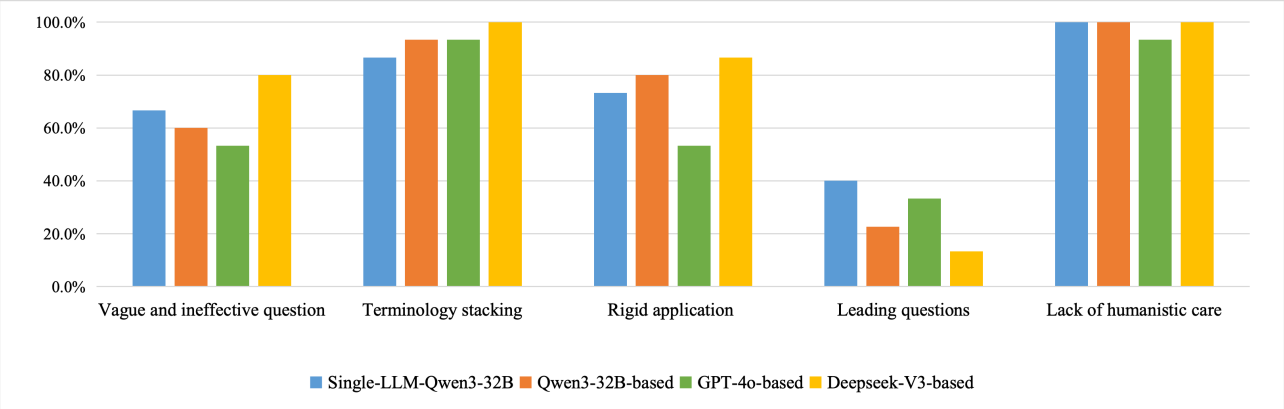

Scalability

We preliminarily evaluated the scalability of our framework by assessing its factual consistency and role-playing performance, using LLM-as-judge evaluations as a preliminary exploratory approach. The line graph in Figure 11 illustrates the performance variations of each method across case reports from different clinical departments. Trend analysis reveals that, with the exception of the GPT-4o–based framework, the case reports from different departments had a consistent impact on the performance trends of the other methods. The factual consistency scores for pediatrics were notably lower than those of the other 4 departments, indicating that case reports from this specialty may have a substantial impact on the VP's responses. The other 4 departments exhibited similar scoring ranges, with our method demonstrating stable fluctuations within a high-scoring interval (0.7 - 0.85), suggesting good generalizability and broad applicability across diverse case reports.

XSL•FO
RenderX

Figure 11. Corresponding score distribution curves of factual consistency results for multiagent frameworks with different specialty case reports and different base models.

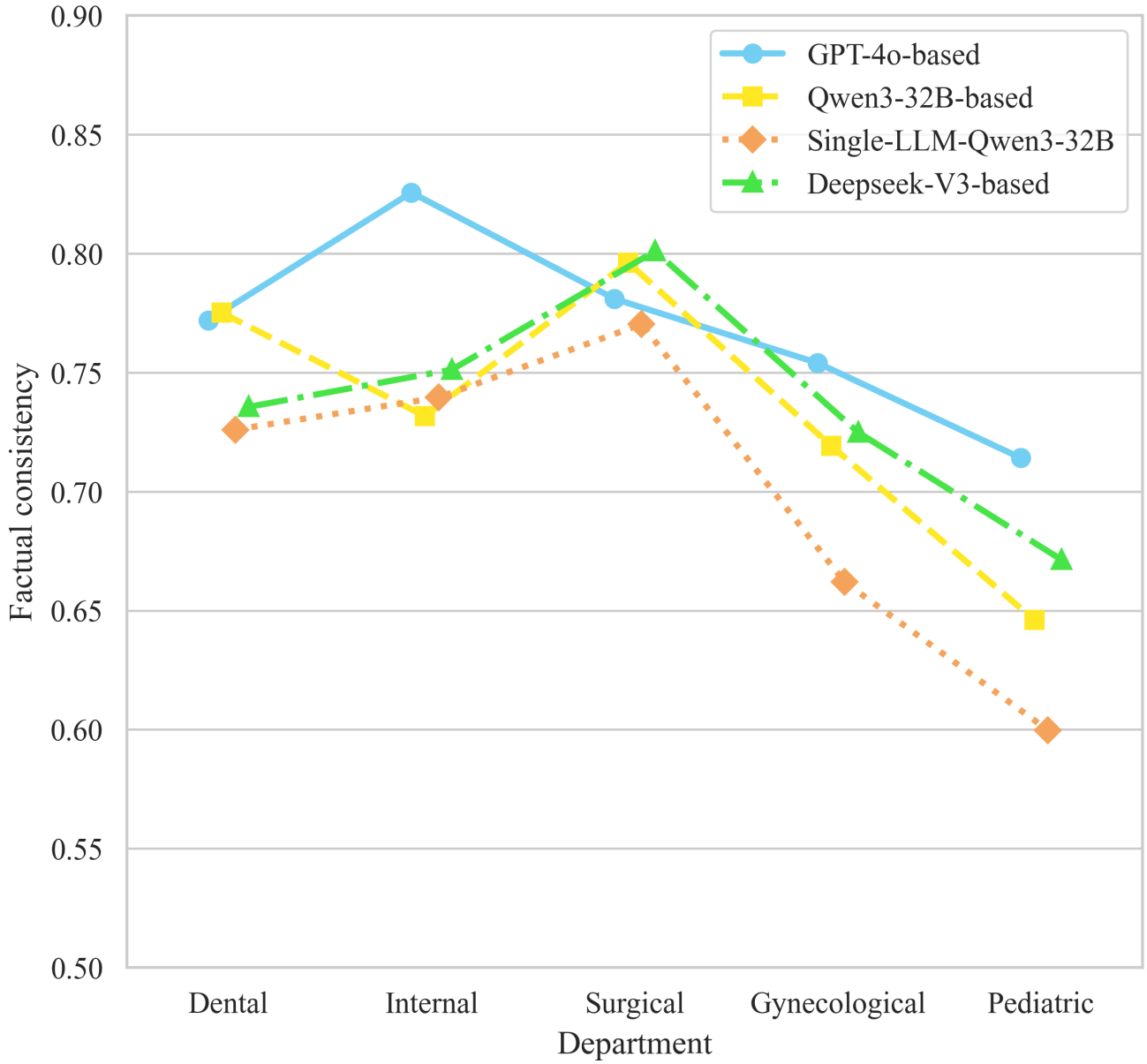

Analysis of the CV revealed that the GPT-4o–based framework (CV=4.7%) exhibited the highest stability, while single-LLM Qwen3-32B framework (CV=8.7%) showed slightly greater variability. As illustrated in Table 9, the overall model performance was minimally affected by clinical departments. However, factual consistency scores for single-LLM-Qwen3-32B framework varied more substantially across departments, indicating lower scalability compared to the multiagent framework.

Table . Comparison of the coefficient of variation (CV) across different methods.

Methods	CV, %
GPT-4o–based	4.7
Qwen3-32B–based	7.1
Single-LLM-Qwen3-32B	8.7
DeepSeek-V3–based	5.7

The LLM-as-judge evaluation across different specialty-specific VP dialogues revealed that our framework demonstrated stable cross-specialty performance , with limited score variations in the appeal of role-playing (Δ<2), outperforming single-LLM baseline's greater variability, as shown in Figure 12. DeepSeek-V3 based showed minimal fluctuations in role attractiveness (Δ=1). These findings suggest that while our framework enhances scalability in the appeal of role-playing in certain clinical departments, the performance of the base model should be considered as a critical factor during deployment.

XSL•FO
RenderX

Figure 12. Performance scores of role-playing competencies across specialties based on GPT-4 automated evaluation.

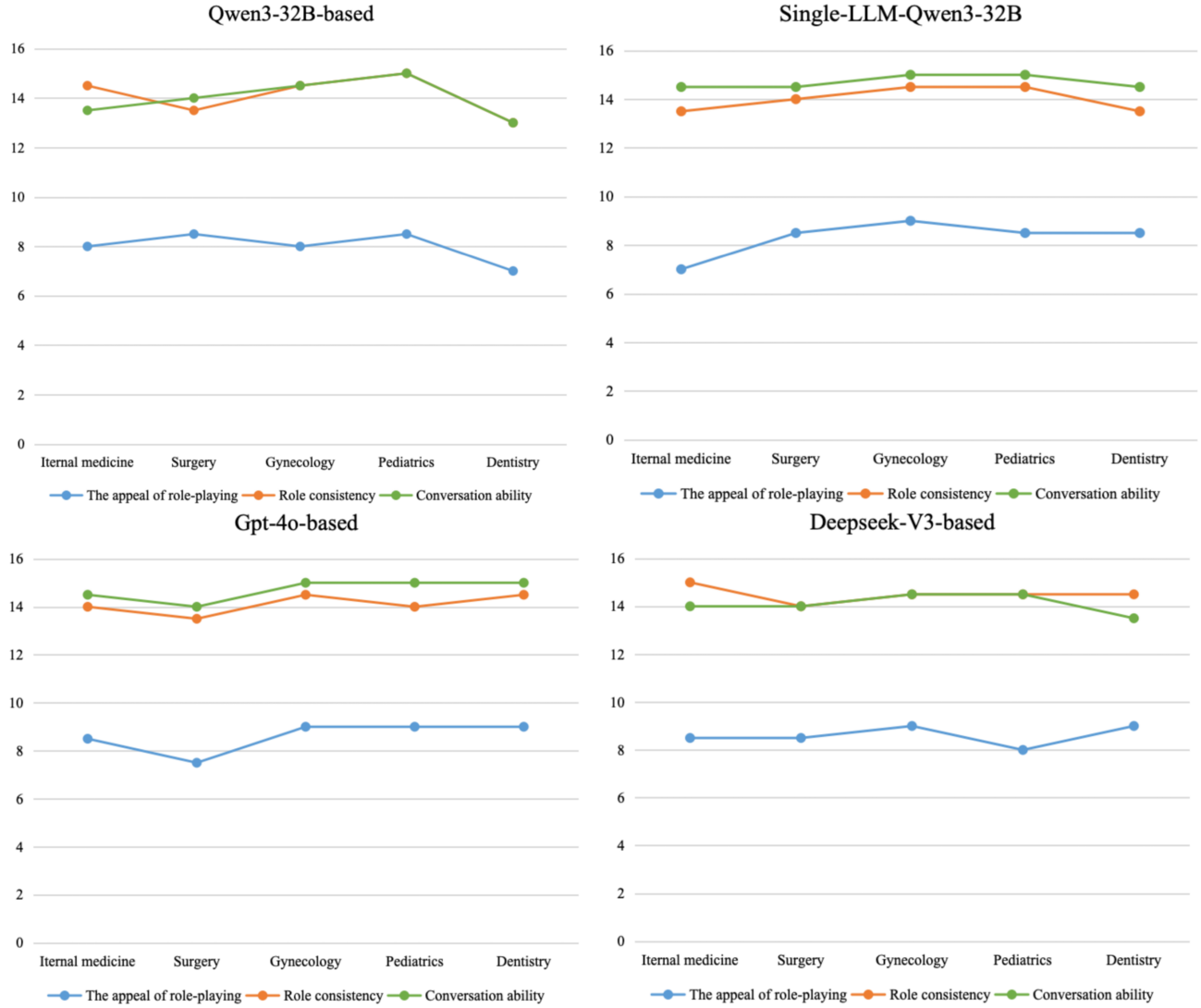

Interaction Efficiency and Multiturn Dialogue Performance

Regarding interaction efficiency, the Qwen3-32B–based multiagent VP achieved the fastest response rate among the 3 backbone models, slightly longer than the single-LLM Qwen3-32B. For the Qwen3-32B–based VP, the average latency of approximately 3 seconds fell within an acceptable range for user interaction, as shown in Figure 13. In terms of resource usage, DeepSeek-V3–based VP proved the most efficient (mean 1219.64 tokens (SD 130.95)), whereas the Qwen3-32B–based VP consumed the most. Although the Qwen3-32B–based VP used more tokens than the single-LLM baseline, its consumption remained stable across sessions, contrasting with the high variance observed in the single-LLM baseline.

XSL•FO
RenderX

Figure 13. Interaction efficiency evaluation based on average time and average token.

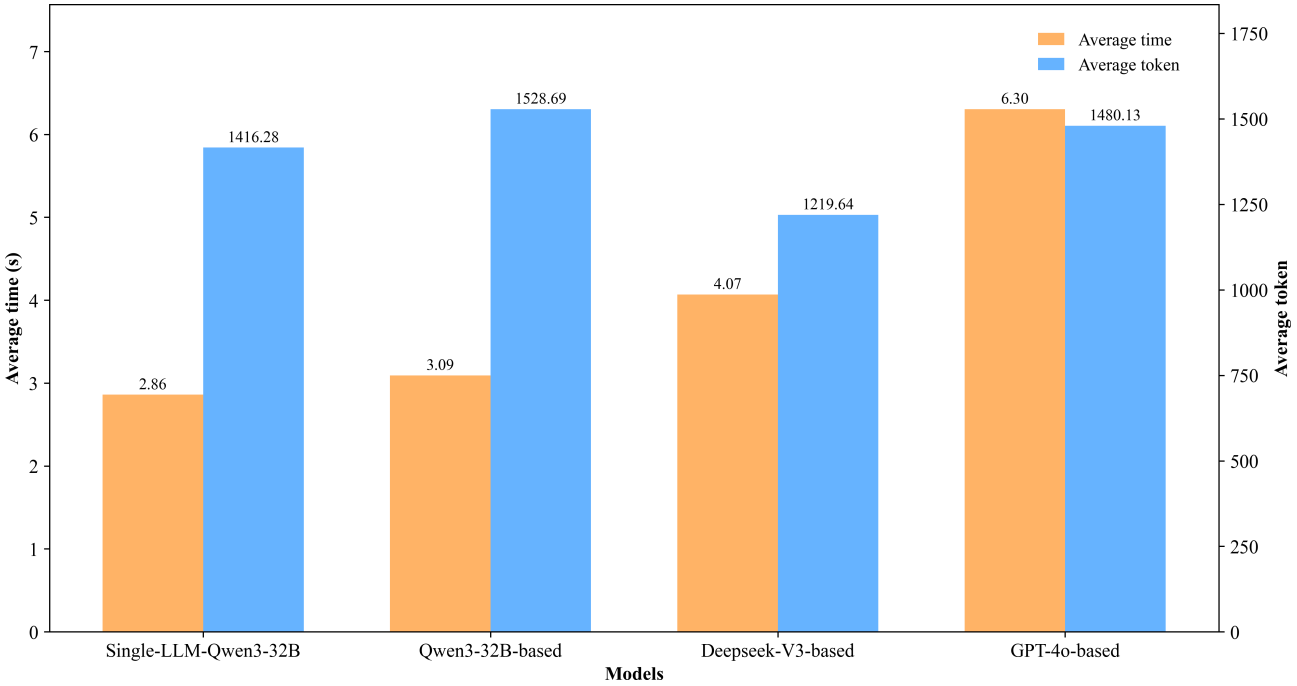

For multiturn performance, the Qwen3-32B–based VP achieved the highest engagement and average response length, averaging 13 turns according to Table 10 and Figure 14. Regarding the dynamics of the interaction, the analysis of information accumulation in Figure 15 reveals distinct differences in information pacing. Specifically, the baselines built on GPT-4o and Qwen3-32B exhibited an information accumulation rate trend characterized by an initial rise, subsequent fall, and final rise. In contrast, DeepSeek-V3 and the single-LLM baseline approach showed a trajectory, featuring a rapid initial rise followed by a decline. These diverging patterns highlight variations in information flow during the dialogue. The rapid early accumulation observed in the latter group suggests a more rapid information release strategy, which could indicate either higher efficiency or a tendency toward excessive disclosure in early turns. Ultimately, this indicates that sustainable multiturn dialogue relies on a balanced information release rhythm, and avoiding early information saturation might be more conducive to dialogue stability.

Table . Average conversation turns by different methods

Method	Values, mean (SD)
DeepSeek-AI–based	6.333 (0.882)
GPT-4o–based	11.5 (4.950)
Qwen3-32B–based	13 (4.243)
Single-LLM-Qwen3-32B	7.5 (1.732)

XSL•FO
RenderX

Figure 14. Statistical distribution of reply lengths across different methods, illustrating the average length alongside minimum and maximum values.

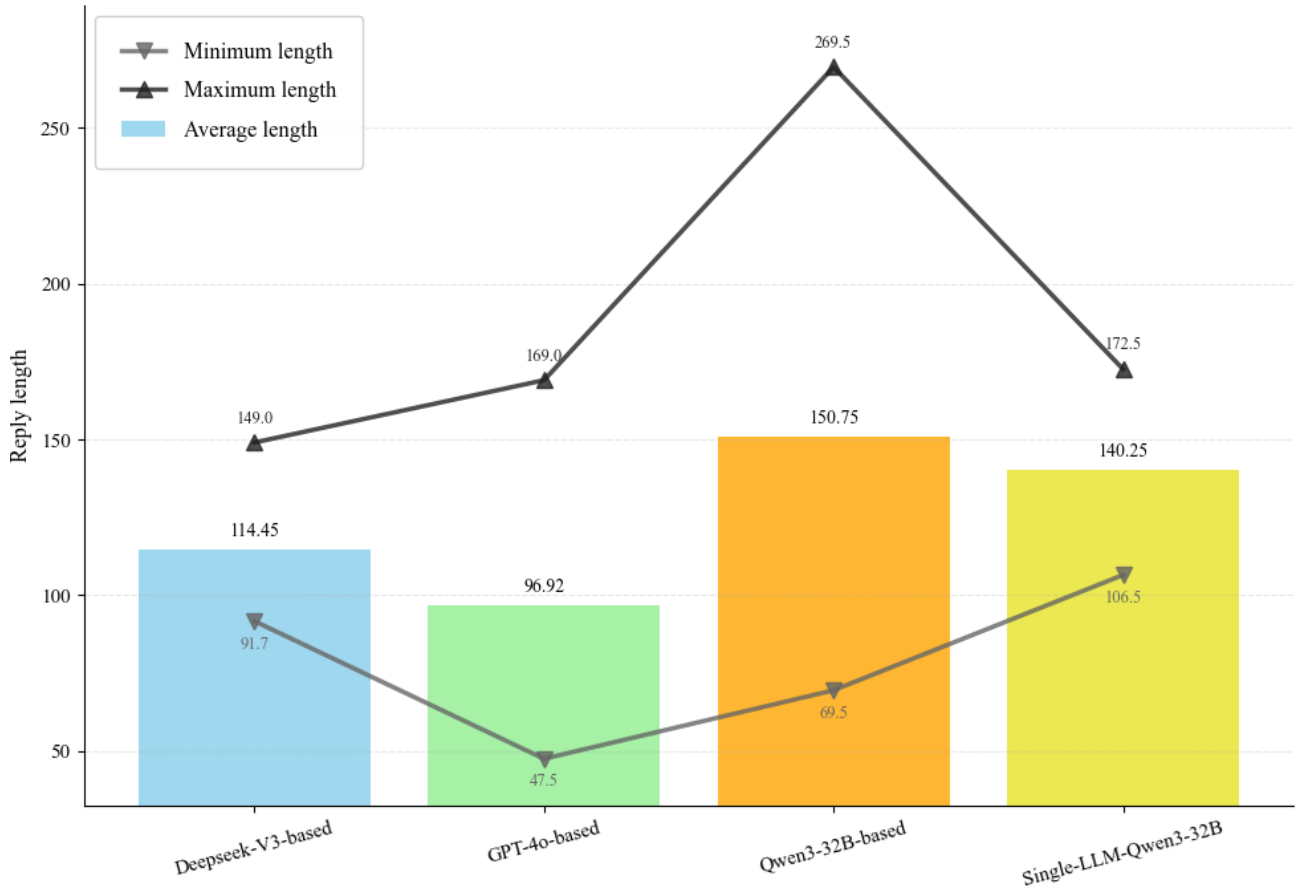

Figure 15. Information accumulation curve.

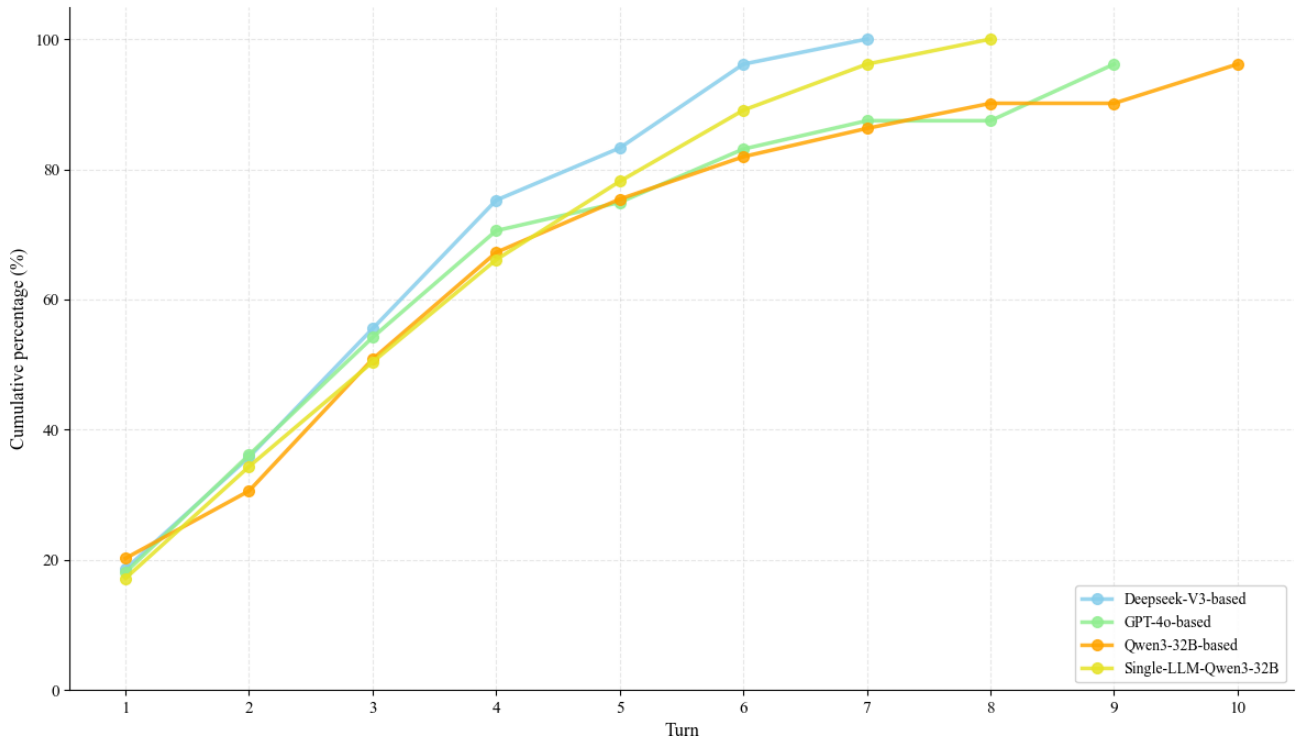

Multiagent Workflow Instance

Figure 16 exemplifies the multiagent workflow for VP construction and response generation. Upon case confirmation by medical educators, agents collaborate to develop a tailored VP based on specified requirements. These agents construct patient-specific information aligning with the case's core features. The example illustrates a middle-aged male patient with clinical anxiety, cautiousness, dependency, and health concern. Concurrently, the information processing agent

XSL•FO
RenderX

segments the case report and stores clinical data for real-time retrieval during student interactions.

When a simulated student inquiry asks, for instance, “Where have you been feeling uncomfortable recently?,” the VP response framework activates. The system classifies this query as high quality and open ended, which triggers the selection of an optimized prompt to generate a relevant response. Simultaneously, the system uses clinical reasoning to connect the query to the most relevant case modules, such as the patient’s chief complaint and present illness. Through structured traversal of minimal information units within these modules, the agent precisely identifies and retrieves the most clinically relevant data elements for response generation.

The response generation agent generates replies based on information input from other subagents. The reflection agent performs final-stage quality control by verifying response compliance with SP protocols. Responses satisfying all clinical and educational standards are directly delivered to users, while noncompliant responses undergo targeted modification to rectify specific deviations before being output.

Figure 16. An input and output example with the multiagent framework.

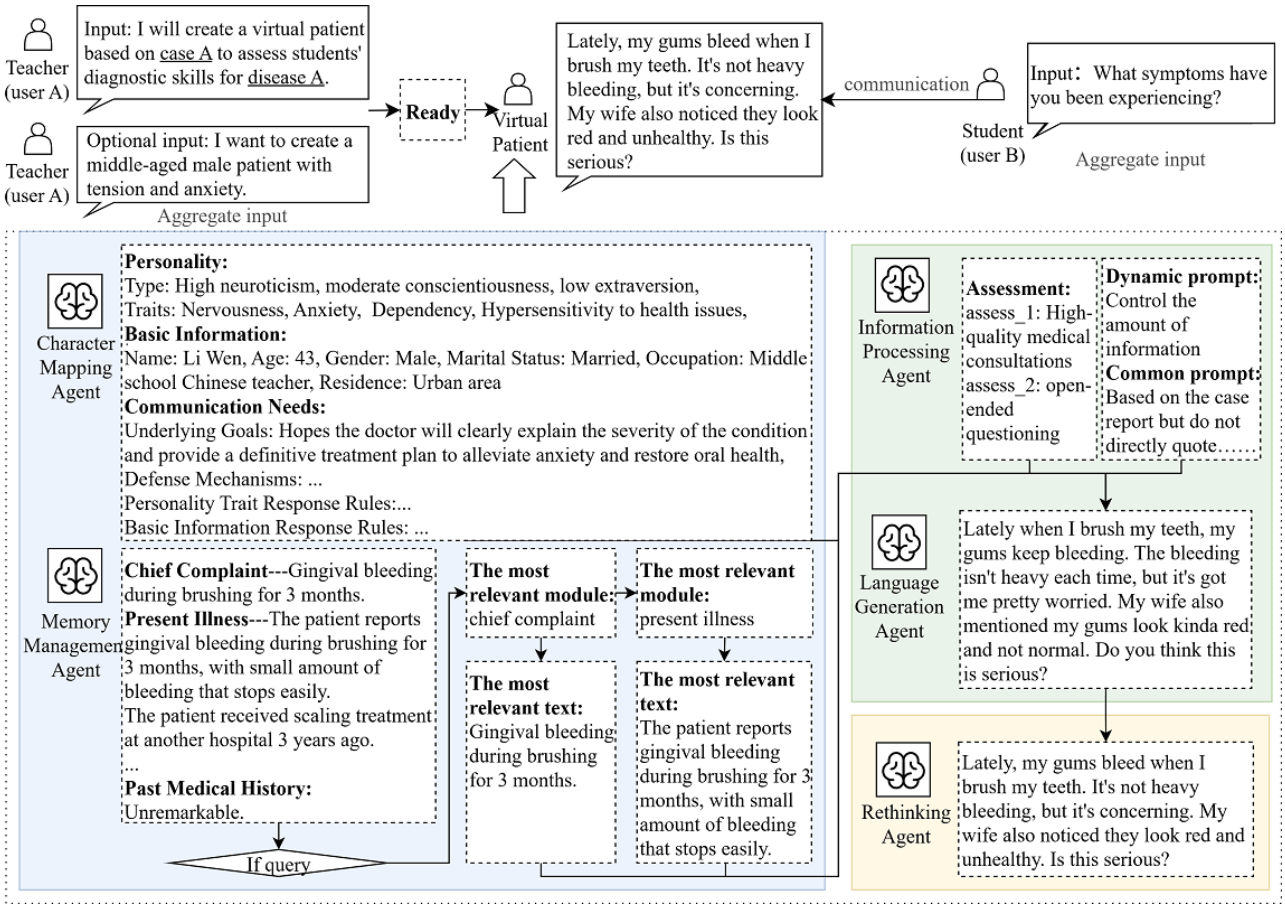

Discussion

Principal Findings

Our study successfully developed a multiagent framework for SP simulation and evaluated its performance against single-LLM baseline setups. The principal findings reveal that the multiagent framework’s structured design and the resulting disparity in prompt depth significantly improved VP simulation quality by enhancing factual consistency, reducing misleading responses, increasing role-playing expressivity, and demonstrating stable performance in terms of interaction efficiency and multiturn dialogues.

The high accuracy rate underscores the framework’s clinical correctness, while factual consistency reflects semantic alignment with case reports. Together, these complementary metrics indicate that the VP responses are both semantically grounded and clinically reliable. While the GPT-4o–based framework exhibited the highest factual consistency, it yielded the lowest accuracy. This implies a specific weakness in pinpointing targeted consultation information despite generating fluent, context-similar text. In contrast, Qwen3-32B–based and DeepSeek-V3–based achieved peak accuracy scores. Despite yielding a higher, albeit nonsignificant, medical accuracy compared to Qwen3-32B–based, the single-LLM baseline exhibited significantly higher misleading rates than both the Qwen3-32B–based and DeepSeek-V3–based multiagent frameworks. These findings suggest that the absence of decomposed reasoning increases the susceptibility to hallucinated content.

Our framework reinforces patient characteristics by reasoning deeply about superficial traits, which helps create richer personality profiles and guides the LLM’s persona performance, aligning with role-playing agent research on role identity activation [44]. Evaluations showed that in role-playing capability and perceived educational utility, the Qwen3-32B

XSL•FO
RenderX

framework performed comparably to the DeepSeek-V3 implementation, with both significantly outperforming the GPT-4o–based one. It is noteworthy that our multiagent framework achieved a higher score in role-playing than the single-LLM baseline on the standard inquiry test set. This observation can be attributed to the deep role-response reasoning implemented in our subagents. Physicians also strongly endorsed our framework's educational utility with a high score. The inverse relationship between factual consistency and role-playing implies that maximizing similarity to clinical facts is a limited proxy for realistic simulation, as authentic patients rarely speak with such medical precision. Thus, factual consistency should serve as a necessary baseline for clinical correctness rather than a target to be maximized in isolation, avoiding rigid alignment that hinders naturalistic expression.

Regarding the exploratory analysis of the low-quality inquiry test set, specifically for terminology stacking and rigid application, the results indicate that the framework adheres to character constraints more effectively than the single-LLM baseline, yet exhibits limited discriminative behavioral dynamics when challenged by subtle cues such as leading questions. Overall, the results demonstrate that our framework enhances the VP's capability to maintain its persona compared to the single-LLM baseline.

As for factual consistency across diverse clinical departments, the framework maintained scores within the 0.7 to 0.85 range except pediatrics, with a lower CV than the single-LLM baseline, suggesting improved stability and potential for rapid construction without complex preprocessing [31,45]. The drop in pediatrics likely resulted from the simulation of the child patients who typically have limited communicative ability. This indicates that role definitions in specific departments require adjustment, such as simulating a guardian to better support realistic medical education. In terms of role-playing ability, preliminary LLM-as-judge evaluations suggested consistent cross-specialty performance, with limited score variations ($\Delta<2$) observed in the appeal of role-playing. However, despite expert review of the generated test set, the use of GPT-4 for both creating the test set and the subsequent evaluation introduces a risk of self-preference bias or circular validation. Therefore, these role-playing metrics serve primarily as an exploratory reference, and definitive verification necessitates further validation by human specialists.

MASs face significant challenges regarding time and cost efficiency in practical applications. While our multiagent framework based on Qwen3-32B exhibited a higher average response time and higher average token consumption compared with the single-LLM baseline, its performance remained within acceptable limits, demonstrating that the efficiency of our framework was acceptable. Regarding multiturn dialogue, the average number of dialogue turns and the steady growth of accumulated information points across turns reflect the gradual information release pace essential to simulated clinical settings, highlighting the framework's potential for application.

Overall, our proposed framework demonstrated superior performance compared to the single-LLM baseline across most evaluation metrics, while maintaining an acceptable efficiency profile. Collectively, these findings highlight the importance of optimizing the trade-off between time efficiency and response quality when selecting model configurations for real-world deployment. While this study prioritizes real-time interaction fidelity, we acknowledge that alternative deployment settings may yield different trade-offs between reasoning depth and latency. In particular, a single LLM with reasoning enabled could offer stronger analytical performance in offline or low–time-sensitivity educational scenarios, such as post hoc case analysis or self-paced learning. However, our preliminary evaluations indicate that enabling reasoning in large-scale models introduces substantial response latency and variability, which undermines conversational realism in interactive patient simulations. Consequently, the current framework adopts multiagent coordination under constrained inference settings as a more practical solution for real-time clinical education, while future work may explore adaptive reasoning strategies that dynamically balance responsiveness and analytical depth across diverse learning contexts.

Comparison With Prior Work

To further evaluate the performance of our framework, we compared its evaluation results with those of existing MASs designed for role-playing. Specifically, in terms of factual consistency, we compared our framework with an existing system Mediq [46]. Our GPT-4o–based framework achieved a score of 82.6% in medical dialogues, which is higher than Mediq's direct (55.9%) and instruct (62.8%) variants but lower than its fact-select approach. The difference may stem from our emphasis on patient personality, which can reduce textual similarity to the original case reports. Our framework also achieved significantly higher average scores in conversational ability and the appeal of role-playing compared to top-performing GPT-4–based implementations in the CharacterEval benchmark [43]. Although role consistency was slightly lower, likely because dialogues did not always require explicit trait expression, expert evaluations confirmed consistency with LLM-as-judge rankings, validating the Qwen3-32B–based framework's robust role-playing capabilities, corroborating prior role-playing agent studies on role identity activation [44].

Educational Implications for Medical Training

From an educational perspective, the proposed multiagent VP framework offers several practical implications for clinical communication training, exemplifying the pivotal role of generative AI in advancing medical education [47,48]. First, by maintaining stable patient personas while handling low-quality or incomplete inquiries, which have been noted in [49] to trigger hallucinations and response instability, the framework supports repeated, self-directed practice for medical students, particularly in early-stage communication skill acquisition. Second, the controllable and scalable nature of the system allows instructors to rapidly generate diverse SP scenarios without the logistical constraints associated with human SP recruitment and training [50]. Third, the observed reduction in misleading responses compared with the single-LLM baseline is especially relevant for formative educational settings, where inaccurate patient feedback may

XSL•FO
RenderX

negatively reinforce incorrect clinical reasoning. As Wen et al [51] noted, single generative language models tend to entail inherent misleading risks in clinical-related interactions. Such behaviors directly align with the concern that misleading feedback from single LLMs may reinforce wrong clinical reasoning in formative education. Collectively, these findings suggest that the multiagent VP framework may serve as a complementary educational tool alongside traditional SP-based training, particularly in resource-limited or large-scale teaching contexts.

Limitations

This study has several limitations. First, our evaluation focused on the overall performance of the multiagent framework rather than the isolated effectiveness of individual subagents. While this design demonstrates the advantages of collaborative agent-based modeling, it does not fully disentangle the specific contributions of each component, such as the memory management agent or the character mapping agent. Importantly, the comparison between the multiagent framework and the single-LLM baseline was not fully matched in terms of prompt depth. As shown in Table 1 in Multimedia Appendix 1, the single-LLM baseline relied primarily on teacher-provided patient settings, whereas the multiagent framework used a dedicated character mapping agent to expand these inputs into a richer personality profile, including behavioral tendencies, response thresholds, and defense mechanisms. This difference introduces a potential confounding factor, as the observed performance gap may reflect not only architectural advantages of multiagent coordination but also differences in prompt engineering depth. Consequently, future research should conduct more controlled ablation studies to isolate the effects of individual subagents and systematically evaluate the impact of prompt complexity, thereby disentangling architectural contributions from prompt design effects.

Second, although our study conducted a comprehensive evaluation, it did not include a prospective validation in real-world clinical settings. The assessment relied primarily on controlled scripts and LLM-as-a-judge methods, which may not fully capture the complexity of authentic student-patient interactions. While preliminary expert-based evaluation suggested high perceived educational utility, the long-term effectiveness and pedagogical impact of the framework remain uncertain. Further studies should incorporate randomized controlled trials in real teaching environments to rigorously assess whether the proposed system improves clinical communication skills compared with SPs or conventional training approaches.

Third, in addition to role-playing ability and multiturn dialogue evaluation, we assessed the response quality by incorporating specific natural language metrics, mirroring the essential SP function of conveying patient history and information during clinical dialogues. However, it is crucial to acknowledge the applicability boundary of these metrics that higher factual accuracy does not intrinsically mean greater patient simulation performance, as authentic patient behavior inherently involves forgetfulness, logical gaps, and potential fabrication. Future research should explore more comprehensive evaluation frameworks that better reflect real-world communication dynamics.

Finally, this study relied exclusively on Chinese clinical databases and expert evaluators. This linguistic homogeneity implies that the observed performance may stem from the base model's alignment with specific training data rather than inherent architectural advantages. To ensure the framework's generalizability and linguistic robustness, future validation on English and other multilingual datasets is essential.

Conclusions

This study demonstrates that our multiagent framework, leveraging LLMs, provides a feasible method for SPs simulation. It facilitates natural and acceptable language-based interactions between VPs and users. In the context of medical education, this approach supports a promising pathway toward the development of scalable and effective communication training. Our multiagent framework demonstrated high factual consistency, response accuracy, and role-playing ability, while also demonstrating stable performance in terms of interaction efficiency and multiturn dialogues. This design overcomes the limitations of single-LLM baseline in medical education role-playing, effectively mitigating hallucinations and significantly reducing the rate of misinformation. Furthermore, our framework addresses limitations in case scenario updates and customization often found in existing multiagent VP systems, exhibiting high flexibility and scalability in certain clinical departments. This design shows promise in providing training-oriented support that complements theoretical coursework. Future studies should further explore its long-term efficacy and broader applicability in authentic medical teaching environments.

Acknowledgments

This study was supported by grants from Zhejiang University, Peking Union Medical College, CAMS, and the Beijing Municipal Health Commission. The views expressed are solely those of the authors and do not represent any institution. The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT (Generative Artificial Intelligence Delegation Taxonomy 2025), the following tasks were delegated to GenAI tools under full human supervision: code optimization, creation of algorithms for data analysis, summarizing text, and translation. The GenAI tool used was DeepSeek-R1. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

XSL•FO
RenderX

Funding

The research was supported by Intelligent Early Warning and Diagnosis-Treatment Decision Support System for Major Diseases (2025C01136), CAMS Innovation Fund for Medical Sciences (2025-I2M-KJ-002), Peking Union Medical College, Graduate Education and Teaching Reform Project in 2024 (2024yjsjg014), the Chinese Academy of Medical Sciences Innovation Fund for Medical Sciences (2021-I2M-1-056), Peking University Health Science Center Medical Education Research Funding Project (2025YB19), and Beijing Municipal Health Commission 2024 Residency Training Quality Improvement Program Residency Training (2024021).

Data Availability

The data generated or analyzed during this study can be provided by the corresponding author upon reasonable request.

Authors' Contributions

YQ, XX, XL, and JL conceived the study and designed the work. XX, LY, and LJ contributed to the acquisition of data. YY and YW contributed to the statistical analysis. YQ, YW, and XX contributed to the interpretation of the data. YQ and XX wrote the original draft of the manuscript. All authors reviewed and approved the final version of the manuscript prior to submission and agreed to be accountable for all aspects of the work.

Conflicts of Interest

YW is an employee of Hangzhou Joyrun Medical Science and Technology Co, Ltd. The other authors declare no conflicts of interest.

Multimedia Appendix 1
Prompts.
[DOCX File, 29 KB - jmir_v28i1e84747_app1.docx]

References

1. Kramer AWM, Düsman H, Tan LHC, Jansen JJM, Grol RPTM, van der Vleuten CPM. Acquisition of communication skills in postgraduate training for general practice. Med Educ 2004 Feb;38(2):158-167. [doi: 10.1111/j.1365-2923.2004.01747.x] [Medline: 14871386]
2. Singh H, Naik AD, Rao R, Petersen LA. Reducing diagnostic errors through effective communication: harnessing the power of information technology. J Gen Intern Med 2008 Apr;23(4):489-494. [doi: 10.1007/s11606-007-0393-z] [Medline: 18373151]
3. Boissy A, Windover AK, Bokar D, et al. Communication skills training for physicians improves patient satisfaction. J Gen Intern Med 2016 Jul;31(7):755-761. [doi: 10.1007/s11606-016-3597-2] [Medline: 26921153]
4. Challa KT, Sayed A, Acharya Y. Modern techniques of teaching and learning in medical education: a descriptive literature review. MedEdPublish (2016) 2021;10:18. [doi: 10.15694/mep.2021.000018.1] [Medline: 38486533]
5. Ende J. Feedback in clinical medical education. JAMA 1983 Aug 12;250(6):777-781. [doi: 10.1001/jama.1983.03340060055026] [Medline: 6876333]
6. Liu H, Wang B, Lu T, Pei Y. Safety and efficacy of probiotics in the prevention of necrotizing enterocolitis in premature and/or low-birthweight infants: a systematic review and meta-analysis. Transl Pediatr 2022 Feb;11(2):249-259. [doi: 10.21037/tp-22-27]
7. Rønning SB, Bjørkly S. The use of clinical role-play and reflection in learning therapeutic communication skills in mental health education: an integrative review. Adv Med Educ Pract 2019;10:415-425. [doi: 10.2147/AMEP.S202115] [Medline: 31417328]
8. Lovink A, Groenier M, van der Niet A, Miedema H, Rethans JJ. How simulated patients contribute to student learning in an authentic way, an interview study. Adv Simul 2024 Jan 11;9(1):4. [doi: 10.1186/s41077-023-00277-w]
9. Kyaw BM, Posadzki P, Paddock S, Car J, Campbell J, Tudor Car L. Effectiveness of digital education on communication skills among medical students: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Aug 27;21(8):e12967. [doi: 10.2196/12967]
10. Papanagnou D, Klein MR, Zhang XC, et al. Developing standardized patient-based cases for communication training: lessons learned from training residents to communicate diagnostic uncertainty. Adv Simul 2021 Dec;6(1):26. [doi: 10.1186/s41077-021-00176-y]
11. Grévisse C. RasPatient pi: a low-cost customizable LLM-based virtual standardized patient simulator. In: Florez H, Astudillo H, editors. Appl Inform: Springer Nature Switzerland; 2025:125-137. [doi: 10.1007/978-3-031-75147-9_9]
12. Jia X, Guo J, Wu M, et al. Study on the effect of simulation-based case teaching method on the preclinical teaching of tooth defects restoration. BMC Med Educ 2025 Apr 5;25(1):487. [doi: 10.1186/s12909-025-07010-3] [Medline: 40188128]

XSL•FO
RenderX

13. Chan KY, Yuen TH, Co M. Using ChatGPT for medical education: the technical perspective. BMC Med Educ 2025 Feb 7;25(1):201. [doi: 10.1186/s12909-025-06785-9] [Medline: 39920711]
14. Raza M, Jahangir Z, Riaz MB, Saeed MJ, Sattar MA. Industrial applications of large language models. Sci Rep 2025 Apr 21;15(1):13755. [doi: 10.1038/s41598-025-98483-1] [Medline: 40258923]
15. Lu K, Yu B, Zhou C, Zhou J. Large language models are superpositions of all characters: attaining arbitrary role-play via self-alignment. In: Ku LW, Martins A, Srikumar V, editors. Presented at: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1; Aug 11-16, 2024; Bangkok, Thailand p. 7828-7840. [doi: 10.18653/v1/2024.acl-long.423]
16. Lampinen AK, Dasgupta I, Chan SCY, et al. Language models, like humans, show content effects on reasoning tasks. In: Abbott D, editor. PNAS Nexus 2024 Jul;3(7):gae233. [doi: 10.1093/pnasnexus/pgae233] [Medline: 39015546]
17. Kononowicz AA, Zary N, Edelbring S, Corral J, Hege I. Virtual patients--what are we talking about? A framework to classify the meanings of the term in healthcare education. BMC Med Educ 2015 Feb 1;15(1):11. [doi: 10.1186/s12909-015-0296-3] [Medline: 25638167]
18. Tu T, Schaekermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence. Nature New Biol 2025 Jun 12;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7]
19. Johri S, Jeong J, Tran BA, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. Nat Med 2025 Jan;31(1):77-86. [doi: 10.1038/s41591-024-03328-5] [Medline: 39747685]
20. Schmidgall S, Ziaei R, Harris C, Reis E, Jopling J, Moor M. AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. arXiv. Preprint posted online on May 25, 2025. [doi: 10.48550/arXiv.2405.07960]
21. Team B-M, Dou C, Liu C, et al. Baichuan-M2: scaling medical capability with large verifier system. arXiv. Preprint posted online on Sep 2, 2025. [doi: 10.48550/arXiv.2509.02208]
22. Sardesai N, Russo P, Martin J, Sardesai A. Utilizing generative conversational artificial intelligence to create simulated patient encounters: a pilot study for anaesthesia training. Postgrad Med J 2024 Mar 18;100(1182):237-241. [doi: 10.1093/postmj/qgad137] [Medline: 38240054]
23. Borg A, Georg C, Jobs B, et al. Virtual patient simulations using social robotics combined with large language models for clinical reasoning training in medical education: mixed methods study. J Med Internet Res 2025 Mar 3;27:e63312. [doi: 10.2196/63312] [Medline: 40053778]
24. Yamamoto A, Koda M, Ogawa H, et al. Enhancing medical interview skills through AI-simulated patient interactions: nonrandomized controlled trial. JMIR Med Educ 2024 Sep 23;10:e58753. [doi: 10.2196/58753] [Medline: 39312284]
25. Yang Y, Achananuparp P, Huang H, Jiang J, Lim EP. Speaker verification in agent-generated conversations. In: Ku LW, Martins A, Srikumar V, editors. Presented at: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1; Aug 11-16, 2024; Bangkok, Thailand p. 5655-5676. [doi: 10.18653/v1/2024.acl-long.307]
26. Ahn J, Lee T, Lim J, et al. TimeChara: evaluating point-in-time character hallucination of role-playing large language models. Presented at: Findings of the Association for Computational Linguistics ACL 2024; Aug 11-16, 2024; Bangkok, Thailand and virtual meeting p. 3291-3325. [doi: 10.18653/v1/2024.findings-acl.197]
27. Bu X, Guo H, Liu J, et al. RoleAgent: building, interacting, and benchmarking high-quality role-playing agents from scripts. In: Globerson A, Mackey L, Belgrave D, Fan A, Paquet U, Tomczak J, et al, editors. Presented at: Advances in Neural Information Processing Systems 37; Dec 10-15, 2024; Vancouver, BC, Canada p. 49403-49428. [doi: 10.52202/079017-1563]
28. Dorri A, Kanhere SS, Jurdak R. Multi-agent systems: a survey. IEEE Access 2018;6:28573-28593. [doi: 10.1109/ACCESS.2018.2831228]
29. Guo T, Chen X, Wang Y, et al. Large language model based multi-agents: a survey of progress and challenges. Presented at: Thirty-Third International Joint Conference on Artificial Intelligence; Aug 3-9, 2024. [doi: 10.24963/ijcai.2024/890]
30. Steenstra I, Nouraei F, Bickmore TW. Scaffolding empathy: training counselors with simulated patients and utterance-level performance visualizations. Presented at: CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems; Apr 26, 2025; Yokohama Japan p. 1-22. [doi: 10.1145/3706598.3714014] [Medline: 40245409]
31. Yu H, Zhou J, Li L, et al. Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. Commun Med. Preprint posted online on Sep 27, 2024. [doi: 10.1038/s43856-025-01283-x]
32. Bowers P, Graydon K, Ryan T, Lau JH, Tomlin D. Artificial intelligence-driven virtual patients for communication skill development in healthcare students. AJET 2024 Jun 7;40(3):39-57. [doi: 10.14742/ajet.9307]
33. Wang X, Chen G, Dingjie S, et al. CMB: a comprehensive medical benchmark in chinese. In: Duh K, Gomez H, Bethard S, editors. Presented at: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics; Jun 2024 to 2024; Mexico City, Mexico p. 6184-6205. [doi: 10.18653/v1/2024.naacl-long.343]
34. McCrae RR, Costa PT. Validation of the five-factor model of personality across instruments and observers. J Pers Soc Psychol 1987 Jan;52(1):81-90. [doi: 10.1037//0022-3514.52.1.81] [Medline: 3820081]
35. Haring CM, Cools BM, van Gurp PJM, van der Meer JWM, Postma CT. Observable phenomena that reveal medical students' clinical reasoning ability during expert assessment of their history taking: a qualitative study. BMC Med Educ 2017 Aug 29;17(1):147. [doi: 10.1186/s12909-017-0983-3] [Medline: 28851340]
36. Maguire GP, Rutter DR. History-taking for medical students. I-Deficiencies in performance. Lancet 1976 Sep 11;2(7985):556-558. [doi: 10.1016/s0140-6736(76)91804-3] [Medline: 60632]

XSL•FO
RenderX

37. Alrasheedi AA. Deficits in history taking skills among final year medical students in a family medicine course: a study from KSA. J Taibah Univ Med Sci 2018 Oct;13(5):415-421. [doi: 10.1016/j.jtumed.2018.07.001] [Medline: 31435357]
38. Yang A, Li A, Yang B, et al. Qwen3 technical report. arXiv. Preprint posted online on May 14, 2025. [doi: 10.48550/arXiv.2505.09388]
39. DeepSeek-AI, Liu A, Feng B, Xue B, et al. DeepSeek-V3 technical report. arXiv. Preprint posted online on Dec 27, 2025. [doi: 10.48550/arXiv.2412.19437]
40. GPT-4o system card. OpenAI. 2024. URL: https://cdn.openai.com/gpt-4o-system-card.pdf [accessed 2024-08-07]
41. Tu Q, Fan S, Tian Z, et al. CharacterEval: a chinese benchmark for role-playing conversational agent evaluation. 2024 Jan 9 Presented at: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1. [doi: 10.18653/v1/2024.acl-long.638]
42. Xiao S, Liu Z, Zhang P, Muennighoff N, Lian D, Nie JY. C-pack: packed resources for general chinese embeddings. arXiv. Preprint posted online on Sep 14, 2023. [doi: 10.48550/arXiv.2309.07597]
43. Tu Q, Fan S, Tian Z, et al. CharacterEval: a Chinese benchmark for role-playing conversational agent evaluation. 2024 Jan 2 Presented at: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1. [doi: 10.18653/v1/2024.acl-long.638]
44. Tang Y, Chen K, Yang M, et al. Thinking in character: advancing role-playing agents with role-aware reasoning. arXiv. Preprint posted online on Jun 2, 2025. [doi: 10.48550/arXiv.2506.01748]
45. Li Y, Zeng C, Zhong J, Zhang R, Zhang M, Zou L. Leveraging large language model as simulated patients for clinical education. arXiv. Preprint posted online on Apr 13, 2024. [doi: 10.48550/arXiv.2404.13066]
46. Balachandran V, Feng S, Ilgen J, et al. MediQ: question-asking llms and a benchmark for reliable interactive clinical reasoning. 2024 Jun 3 Presented at: Advances in Neural Information Processing Systems 37. [doi: 10.52202/079017-0908]
47. Jowsey T, Stokes-Parish J, Singleton R, Todorovic M. Medical education empowered by generative artificial intelligence large language models. Trends Mol Med 2023 Dec;29(12):971-973. [doi: 10.1016/j.molmed.2023.08.012] [Medline: 37718142]
48. Tran M, Balasooriya C, Jonnagaddala J, et al. Situating governance and regulatory concerns for generative artificial intelligence and large language models in medical education. npj Digit Med 2025 May 27;8(1):315. [doi: 10.1038/s41746-025-01721-z]
49. Li D, Lebai Lutfi S. Large language model-based virtual patient systems for history-taking in medical education: comprehensive systematic review. JMIR Med Inform 2026 Jan 2;14:e79039. [doi: 10.2196/79039] [Medline: 41481915]
50. Wu ML, Chao LF, Hung YT. Training students to serve as standardized patients in an objective structured clinical examination is feasible: a mixed-methods study. Nurse Educ Pract 2024 Aug;79:104069. [doi: 10.1016/j.nepr.2024.104069] [Medline: 39053150]
51. Wen A, Lu Q, Chuang YN, et al. Context matching is not reasoning when performing generalized clinical evaluation of generative language models. NPJ Digit Med 2025 Dec 27;9(1):71. [doi: 10.1038/s41746-025-02253-2] [Medline: 41455812]

Abbreviations

CV: coefficient of variation
LLM: large language model
MAS: multiagent system
SP: standardized patient
VP: virtual patient

Edited by A Coristine; submitted 28.Sep.2025; peer-reviewed by A Grilo, P Huang, Z Sun; revised version received 19.Apr.2026; accepted 20.Apr.2026; published 04.Jun.2026.

Please cite as:
Qu Y, Xu X, Long Y, Wang Y, Li J, Lu X
A Large Language Model–Powered Multiagent Framework Emulating Standardized Patients in Clinical Communication Skills Training: Development and Evaluation Study
J Med Internet Res 2026;28:e84747
URL: https://www.jmir.org/2026/1/e84747
doi:10.2196/84747

© Yufei Qu, Xiaowei Xu, Yunzi Long, Yijie Wang, Jiao Li, Xudong Lu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 4.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction

XSL•FO
RenderX

in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring Patient Perspectives on the Use of Artificial Intelligence to Inform Joint Decision-Making for Patients With Multiple Conditions in Primary Care in the United Kingdom: Qualitative Study

Sarah Flanagan[1*], PhD; Charlotte Spurway[2*], PhD; Louise Jackson[2], PhD; Jenny Cooper[1], MBChB, MPH; Francesca L Crowe[1], PhD; Shamil Haroon[1], PhD; Tom Marshall[1], PhD; Leah Fitzsimmons[3], PhD; Eleanor Hathaway[1], MPH; Krishnarajah Nirantharakumar[1], MD; Thomas Jackson[4], PhD; Sheila Greenfield[1], PhD

[1]Department of Applied Health Science, University of Birmingham, Birmingham, England, United Kingdom

[2]Health Economics Unit, University of Birmingham, IOEM Building, Birmingham, England, United Kingdom

[3]Department of Metabolism and Systems Science, University of Birmingham, Birmingham, England, United Kingdom

[4]Department of Inflammation and Ageing, University of Birmingham, Birmingham, England, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Charlotte Spurway, PhD
Health Economics Unit, University of Birmingham, IOEM Building, Birmingham, England, United Kingdom

Abstract

Background: Multimorbidity, living with 2 or more long-term health conditions, is increasing globally and now affects over one-quarter of adults in England. People with multiple long-term conditions (MLTC) face complex health and treatment challenges, often experiencing fragmented care within systems oriented toward single-disease management. Artificial intelligence (AI) has the potential to support clinicians and patients by analyzing complex health data, optimizing treatment strategies, and predicting disease trajectories.

Objective: The OPTIMAL (Optimizing Therapies, Disease Trajectories, and AI-Assisted Clinical Management for Patients Living with Complex Multimorbidity) project aims to develop AI-enabled tools to support shared decision-making in primary care. This study explored how patients with MLTC perceive the use of AI to inform joint decision-making in primary care.

Methods: Semistructured interviews were conducted via telephone or video call with 29 adults living with MLTC between July and November 2023. Participants were recruited through general practitioner practices via the Clinical Practice Research Datalink and community-based organizations across the West Midlands. Interviews were transcribed verbatim and analyzed thematically using an inductive approach. Members of a patient advisory group were involved in developing study materials, refining the interview guide, and reviewing emerging findings to ensure relevance and authenticity.

Results: Participants identified potential benefits of AI in enhancing consultation efficiency and accuracy, improving access to information for patients and clinicians, promoting early detection of health changes, and reducing health care inequalities. However, concerns were raised about the loss of human interaction, data privacy and security, transparency of algorithms, and the potential for bias and inequity in AI systems. Trust and acceptance varied by age and familiarity with technology. Some participants expressed uncertainty about what AI entails and how it could be used in primary care.

Conclusions: Patients with MLTC viewed AI-assisted decision-making in primary care with cautious optimism. While many recognized potential benefits for coordination and personalization of care, others expressed reservations about privacy, fairness, and the risk of diminished human connection.

(J Med Internet Res 2026;28:e87507) doi:10.2196/87507

KEYWORDS

artificial intelligence; AI; multiple long-term conditions; qualitative research; primary care

XSL•FO
RenderX

Introduction

Background

The prevalence of multimorbidity (living with 2 or more long-term health conditions) is increasing worldwide and affects over a quarter of adults in England [1,2]. Patients with multiple long-term conditions (MLTC) experience unique challenges that can impact their quality of life and functional status. More widely, complex MLTC, defined as the coexistence of 4 or more long-term health conditions, has a significant contribution to health care utilization and often leads to polypharmacy [3].

As mainstream services are generally oriented to a "single disease model" of care, costs and service utilization generally increase as the number of conditions a patient has increases [4,5]. Conventional approaches to treating individual conditions in isolation fall short when managing complex MLTC due to the unique complexities, which have implications for experiences of care, costs borne by patients, and impacts on work and other activities [6]. Traditional clinical trials, which typically inform evidence-based guidelines, often exclude individuals with MLTC, resulting in guidelines that do not adequately address the intricacies of complex MLTC [7]. This limitation underscores the critical need for innovative approaches to guide treatment decisions.

Artificial intelligence (AI) has been seen to offer potential solutions to this dilemma, offering the capability to navigate the intricate web of multiple conditions and varied treatments. Definitions of AI have progressed over time and generally relate to machines and algorithms that can "reason and adapt based on sets of rules and environment which mimic human intelligence" [8]. AI tools possess the capacity to interpret vast and varied datasets, accounting for individual patient characteristics and medical histories in ways that can exceed human capabilities [9]. In particular, AI tools can be used as part of the consultation process to inform clinician and patient decision-making around diagnoses, planning treatments, and thinking about outcomes, although the use in primary care for the management of patients to date has been limited [10,11]. This potential is particularly relevant in the context of complex MLTC, where tailoring treatment plans to individual patients becomes increasingly challenging. The OPTIMAL (Optimizing Therapies, Disease Trajectories, and AI-Assisted Clinical Management for Patients Living with Complex Multimorbidity) project aims to leverage AI to develop a joint decision-making tool specifically designed for primary care settings, aimed at identifying the most effective treatment strategies with the lowest risk of additional side effects or antagonizing current conditions and predicting disease trajectories in individuals with MLTC [12].

Objectives

Existing research on patient perceptions in relation to AI in health care has highlighted that while AI tools are generally welcomed, there are a range of concerns. These include potential impacts on the quality of health care received, implications for privacy and data protection, and the accuracy of any AI outputs [13,14]. Hence, it is important to involve patients in developing and implementing AI in health care to ensure that solutions are acceptable and enhance patient care and experiences. Through semistructured interviews, this study sought to capture insights into how AI-informed joint decision-making is perceived by patients compared to traditional general practitioner (GP)-directed approaches in managing MLTC. By exploring attitudes and barriers to AI implementation, the study aims to foster a comprehensive understanding of the landscape surrounding patient views on AI-enabled decision-making tools, particularly within the primary care context.

Methods

Study Design

Data were gathered using semistructured interviews conducted by telephone or via video link (Zoom or Microsoft Teams) between July and November 2023, in line with our published protocol [12,15]. The interview topic guide was developed by the multidisciplinary research team, including members of a patient advisory group (PAG), clinicians, a medical sociologist, qualitative researchers, and other researchers, such as health economists and health data scientists. It was informed by a review of the published literature (unpublished) about the acceptability and utility of using AI for decision-making in health care. Interviews were designed to last for around 30 to 40 minutes.

The interview topic guide, piloted with PAG members and modified based on their feedback, comprised open-ended questions and prompts to elicit detailed responses, with questions focusing on participants' perceptions of the advantages and disadvantages of the use of AI-informed clinical decision-making. The interview topic guide for participants is provided in Multimedia Appendix 1. Prior to the interview, participants were provided with a 1-page briefing note on AI. The briefing note, also developed with the PAG, explained what AI is and how it can be used in health care. At the start of the interview, the interviewer also provided further explanation and answered any participant questions about AI.

Participants and Recruitment

Participants were recruited using purposive sampling via 2 pathways. The first recruitment method used the Clinical Practice Research Datalink (CPRD) Aurum database, an anonymized, population-based, UK database of electronic primary care records [16]. CPRD contains validated clinical and demographic data representative of the UK population [17]. These data, including diagnoses, tests, and prescriptions, were used to identify potential participants [18]. The inclusion criteria include patients who were aged 18 years and above and patients with 2 or more long-term health conditions. The exclusion criteria include patients who were diagnosed with a terminal illness (prognosis of 12 mo or less) and patients who may lack capacity to provide informed consent or undertake study activities.

Using the CPRD patient referral service, invitations were sent via GP practices in the West Midlands, prioritizing more deprived areas with higher proportions of minoritized ethnic groups and patients with MLTC. The Data Extraction for Epidemiological Research (DExtER; developed by researchers

XSL•FO
RenderX

at the University of Birmingham) software was used by the research team to generate a pool of potentially eligible patients [19]. Eligible patients identified via CPRD were sent an invitation letter and participant information sheet (PIS) by their GP practice (Multimedia Appendix 2).

Fewer patients than expected passed GP screening (56%), and the response rate among those invited was low. To augment recruitment, ethical approval was obtained to implement an additional recruitment pathway. A community-based approach was introduced, with recruitment materials (Multimedia Appendix 3) circulated through voluntary, third-sector patient support and community organizations. The recruitment strategy for both pathways is shown in Figure 1.

Interviews were conducted until data saturation was reached and no new themes were identified [20]. A sampling matrix was used to monitor participant demographics during recruitment to support diversity in age, sex, ethnicity, and number of long-term conditions [21]. In total, 21 participants were recruited via CPRD, and 8 participants were recruited from the community pathway.

Figure 1. OPTIMAL study recruitment diagram. CPRD: Clinical Practice Research Datalink; GP: general practitioner; OPTIMAL: Optimizing Therapies, Disease Trajectories, and Artificial Intelligence-Assisted Clinical Management for Patients Living With Complex Multimorbidity; PAG: patient advisory group; PIS: participant information sheet.

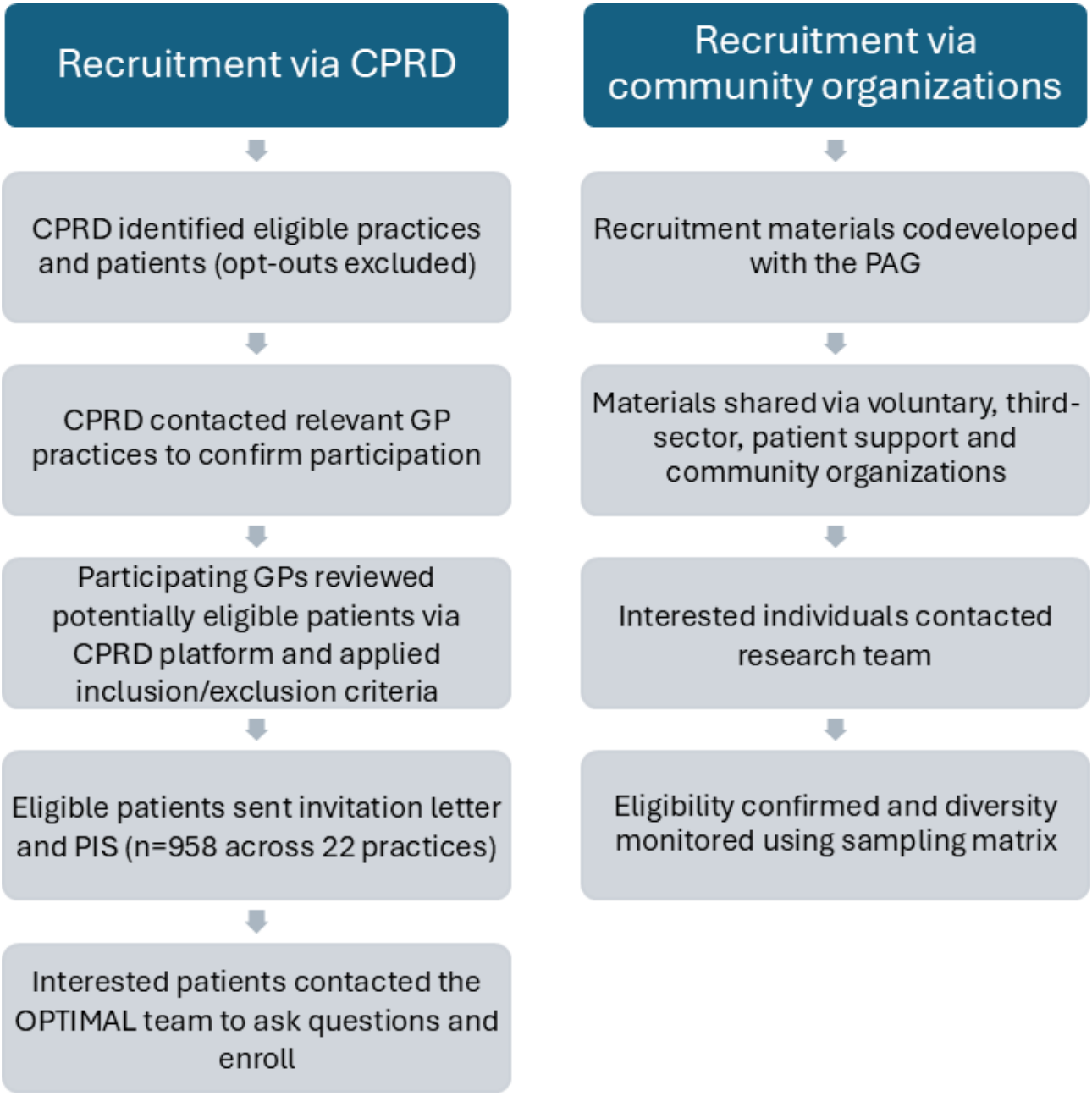

Patient and Public Involvement

Members of the public and patients with MLTC have been involved in every stage of the OPTIMAL study. A PAG was established at study inception, comprising 9 members who have lived experience of MLTC (either directly or as a carer). PAG recruitment considered diversity in terms of personal characteristics and experience of MLTC. PAG members were

XSL•FO
RenderX

involved in codeveloping the topic guide and participant-facing documents (PIS, invitation letter, and consent form) by reviewing and commenting on each iteration. Quarterly PAG meetings supported the qualitative study by reviewing findings and ensuring that the study aims remained meaningful to patients with MLTC.

Procedure and Data Analysis

All interviews were conducted by SF, a Research Fellow at the University of Birmingham, working on qualitative studies based within health research. The interviews were audio-recorded and transcribed verbatim by a transcription company contracted to the University of Birmingham. All transcripts were anonymized and stored on a secure, password-protected server that only authorized personnel had access to, in accordance with UK data protection regulations and in line with University of Birmingham data protection and data security policies.

Personal identifying information was removed from transcripts prior to analysis. Transcripts were analyzed thematically [22]. Following familiarization, codes were assigned using an inductive approach [23]. A sample of transcripts was independently coded by JC (a Clinical Research Fellow with experience in qualitative research) to ensure consistency. After the coding process, the codes were merged into initial themes, which were then further refined by 2 researchers (CS and LJ).

During data coding and theme development, the findings were discussed with the PAG and members of the multidisciplinary research team to identify any areas for further exploration or clarification in subsequent interviews. Coding differences were worked through in a discussion-based, collaborative process with the whole team, which helped maintain clarity and consistency while recognizing the reflexive nature of qualitative analysis. Microsoft Word and Excel (Microsoft 365) were used to read the transcripts and to support coding and theme development.

The COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist was completed to support transparent reporting. The checklist is provided in Checklist 1, and the corresponding reflexivity statement is provided in Multimedia Appendix 4.

Ethical Considerations

Ethical approval for the study was obtained from the National Health Service (NHS) Research Ethics Committees (reference 22/SC/0210). Ethical approval to use the CPRD dataset to select the eligible practices and potentially eligible patients was obtained from the CPRD Research Data Governance Expert Review Committee and the Central Advisory Committee (reference 21_000683). Electronically completed written consent or audio informed consent was collected from participants, and they were informed of their right to withdraw from the study at any time. As part of the consent process, participants were asked for permission to be contacted at a later date with a view to inviting them to take part in a follow-up interview related to a separate theme within the study. Participants' privacy and confidentiality were maintained throughout the study. Interviews were anonymized, and all data were stored securely at the University of Birmingham. A £15 (US $20.17) shopping voucher was offered to participants for their contribution to the study.

Results

Participant Characteristics

Recruitment continued until data saturation was achieved. In total, 29 interviews were conducted with participants of different sex, ages, and ethnic backgrounds (Table 1). The sample was predominantly female and largely White British, with smaller representation from Asian/British Asian and mixed ethnic backgrounds. Participant age range spanned from 20 to above 80 years, with the majority (17/29, 58%) aged 70 years and above.

XSL•FO
RenderX

Table . Characteristics of the participants (N=29).

Characteristic and subgroup	Values, n (%)
Sex	
Female	18 (62)
Male	11 (38)
Age range (y)	
20 - 29	1 (3.45)
40 - 49	2 (6.90)
50 - 59	4 (13.79)
60 - 69	5 (17.24)
70 - 79	11 (37.93)
≥80	6 (20.69)
Ethnicity	
White British	24 (82.75)
Asian/British Asian	4 (13.79)
White and Black African	1 (3.44)
Number of long-term conditions	
2-3	12 (41.39)
4 or more	17 (58.6)
Total	29 (100)

Main Themes

Overall, the interviews found 2 broad themes: potential benefits and potential risks of AI (Table 2). Potential benefits included 4 subthemes, such as enhanced efficiency and accuracy, improved access to information for patients and clinicians, potential early detection of changes in health state, and reducing health care inequalities. Five subthemes were identified for the potential risks of AI, such as the loss of human touch, privacy and data security, transparency of the AI tool and associated information, trust in and accuracy of AI tools, and potential bias and inequality.

XSL•FO
RenderX

Table . Summary of themes and subthemes identified in relation to artificial intelligence (AI) use in multiple long-term conditions (MLTC) care.

Themes and subthemes	Definition
Potential benefits of AI for MLTC care	
Enhanced efficiency and accuracy in processing medical information	The ability of AI to quickly analyze large volumes of clinical data, reduce errors, and support more accurate and efficient decision-making in complex care
Improved access to information for patients and clinicians	AI providing clearer, more accessible health information, helping both patients and clinicians understand conditions, treatment options, and up-to-date evidence
Potential for early detection of changes in the health state	AI identifying subtle patterns or early signs of deterioration, enabling preventative action or timely intervention
Reducing health care inequalities	AI supporting more consistent care across patient groups by standardizing information, highlighting overlooked needs, and improving access for those with lower health literacy
Potential risks of AI for MLTC care	
Loss of human touch	Concerns that AI may reduce personal interaction, empathy, and the relational aspects of care that patients value.
Privacy and data security	Worries about how personal health data are stored, shared, or accessed within AI systems and risks of unauthorized use
Lack of transparency	Uncertainty about how AI reaches decisions, how algorithms are developed, what influences their recommendations, and how decisions are made within the health system about AI tools
Trust and accuracy and AI outputs	Need for reassurance that AI recommendations are reliable, safe, and checked by clinicians, with fears about overreliance or errors
Potential bias and inequality	Concerns that AI trained on nonrepresentative data could reinforce existing disparities or produce less accurate outputs for underrepresented groups

Potential Benefits of AI

Enhanced Efficiency and Accuracy in Processing Medical Information

Participants commonly described the potential for AI to improve the efficiency of consultations and aid in joint decision-making. It was acknowledged that AI could help process large amounts of medical information more quickly than is currently possible, especially in relation to complex medication regimens. Several participants reflected on past experiences where medication interactions had been overlooked and suggested that AI could reduce this risk while also saving time for both patients and clinicians.

> *Well I have felt that at times, medication has been prescribed without regard to what I am already taking, […] but AI could do it so much faster and would take away the need for me to do it myself.* [ID3, female, 75 years old]

Beyond medication management, participants also viewed AI as a tool that could support clinical decision-making during consultations by rapidly analyzing information and presenting relevant options. The predictive capability of AI in identifying potential future health conditions was viewed as valuable. Participants noted that such tools could improve patient care while saving time for clinicians and reducing costs for the NHS.

> *Maybe it would save them time, money, resources. I think it would probably be a really good tool for them as well because they can help their patients better…* [ID16, female, 45 years old]

Improved Access to Information for Patients and Clinicians

Participants appreciated that AI had the potential to provide more accessible information about their conditions, supporting greater understanding and involvement in managing their health. In particular, AI was seen as a way of offering patients additional information about alternative options that they might benefit from. One participant described feeling powerless in managing a complex condition and saw AI as a way to gain clearer information.

> *Because I think the NHS is under resourced and they don't have time to explain all the different things to you […] I don't feel like I'm proactively doing much at the moment apart from taking my drugs. So, I feel very powerless and I would like more.* [ID22, female, 28 years old]

Participants also recognized that AI could support clinicians in helping patients manage MLTC, by drawing attention to new or previously overlooked information and improving awareness of emerging evidence. In addition, AI was seen as having the potential to improve access to information for clinicians by consolidating and analyzing complex medication data.

> *Well, you'd most probably get the doctors to understand a bit more. If it brings up new things and the doctor decides to think about that as well, all the*

XSL•FO
RenderX

better for you know, getting this artificial intelligence to everyone you know. [ID24, male, 57 years old]

Potential for Early Detection of Changes in Health State

AI was perceived as potentially capable of detecting subtle changes or patterns indicative of disease progression earlier than traditional methods, allowing for timely interventions. Participants described how AI could act as a "comfort blanket," offering reassurance that health changes would be monitored and addressed promptly.

But if I'm moving along the pathway, the computer says I need to do a blood test, I'm sending you for a blood test, the computer says we need to do this test just to monitor you for this – I feel it's a comfort blanket. [ID13, female, 65 years old]

There was a sense that AI may be able to provide indications for preventative measures for presenting conditions.

I think that would be a good thing, because prevention is better than cure isn't it really. So, if you're prone to it then you can make the changes before it actually comes to the seriousness of the complaint. [ID8, female, 82 years old]

Reducing Health Care Inequalities

Some participants believed that AI could help reduce health care disparities by standardizing care practices and ensuring equitable access to high-quality health care services. One participant suggested that AI may be especially useful for patients who may be less "health literate."

But I think in cases of some individuals who may not have insight, and they have complex medical conditions, I think AI would be a great, great help to them I think and to the practitioners who are looking after [them]. [ID5, male, 71 years old]

Potential Risks of AI

Loss of Human Touch

Participants expressed concerns that the use of AI could reduce the human elements of care (eg, communication and compassion). In particular, there were worries that consultations might become more transactional, with clinicians relying on automated outputs rather than engaging fully with the individual patient. As 1 participant explained:

I would hope that it wouldn't become sort of rote in so far that the computer spills out information and the doctor immediately doesn't really look at it, doesn't really pay attention to it and then just hands over a prescription. [ID3, female, 75 years old]

Furthermore, some participants expressed anxiety that AI might eventually replace face-to-face interactions with their GP, highlighting the importance of being able to have a dialogue with health care professionals. There was a sense that this may more adversely impact older people.

I think it's contact, isn't it? The perception is that you won't see anyone, it'll just be a computer with a face and you deal with it. I still think from, certainly from an older person's perspective, you want to sit down and talk to someone. [ID26, male, 76 years old]

Privacy and Data Security

Participants raised concerns about the privacy and security of their health information when handled by AI systems. There were worries that personal data could be accessed by unauthorized individuals, highlighting the importance of strong safeguards and clear data protection measures when implementing AI in health care.

Well, I suppose people worried about their information being available because other people could get access to it... [ID29, male, 82 years old]

Lack of Transparency

Concerns were also expressed about the transparency of AI and how AI-driven decisions are made. Participants indicated a desire for reassurance that AI tools would be thoroughly checked and validated, while acknowledging that some aspects of clinical decision-making are already not fully visible to patients. This was demonstrated by 1 participant who stated they would like reassurance that AI tools are carefully checked but, at the same time, highlighted a lack of perceived transparency currently with clinician-based decision-making.

As long as the software is checked and simulated, I'm good with that. I can't see the software but I'm pretty sure that there's something in there that they're saying "Try this. Try that. Do this. Do that." So, we're not far away now are we, the way they operate? [ID26, male, 76 years old]

Some participants also questioned whether financial or organizational incentives could influence how AI is developed and implemented within health care. They questioned whether AI would truly serve patients' best interests or be driven by profit and efficiency within the health care system, which further highlights the need for transparency in how decisions about the introduction of AI are made.

When everything is about money and profit these days, you have to question whether is AI being proposed as a solution for patients or a solution for cheaper private healthcare? [ID19, male, 64 years old]

Trust and Accuracy of the AI Outputs

Participants emphasized the need for reassurance about the accuracy and reliability of AI-generated outputs. They suggested that AI should support, rather than replace, clinical judgment with safeguards, such as human oversight in place.

I think there needs to be some sort of second opinion thing built into it because AI is not foolproof. So, there's definitely some sort of second opinion. [ID21, male, 65 years old]

Trust in AI also appeared to vary by age, with some participants suggesting that younger people might be more comfortable relying on digital technologies than older individuals, who may feel more cautious or uncertain. There was also a sense of caution about overreliance on AI, with concerns that algorithms

XSL•FO
RenderX

may not capture individual variations or predict all outcomes accurately.

> *I'd like to know first of all, is it going to be beneficial to me in the long run. It's all right saying "We can put you on this medication" but you know, it can't really predict side effects and those side of things.* [ID1, female, 41 years old]

Potential Bias and Inequality

Participants expressed concern that AI could reinforce existing inequalities in health care if the data used to train the AI do not adequately represent diverse populations. There was particular awareness that individuals from minority ethnic, cultural, or neurodiverse backgrounds may be misrepresented or overlooked, leading to less accurate or appropriate recommendations for these groups. As 1 participant explained, this would potentially lead to AI making recommendations informed by data that inadequately reflect the circumstances of many people, particularly those from underrepresented or disadvantaged communities.

> *When it comes to it, I think it can be really good and a really useful tool, as long as you've got the right information. I find a lot of the time—and this is speaking from lived experience—within south Asian communities and with Black communities—people do not talk about their physical or mental health. So AI will only know about white communities who talk about these conditions.* [ID16, female, 45 years old]

Discussion

Principal Findings

This study found that participants had mixed views regarding the use of AI to inform joint decision-making for patients with complex MLTC in a primary care setting. Most recognized potential benefits included improved efficiency and accuracy, enhanced access to information, the ability to detect early changes in health status, and the potential to reduce inequalities. However, nearly all participants also highlighted concerns, particularly around the loss of human touch, data security, transparency, accuracy, and potential biases within AI.

Comparison With Other Work

These findings are consistent with existing literature reporting mixed patient views on AI in health care [24,25]. Previous research has shown that patients view AI in health care with cautious optimism, recognizing its potential to improve diagnosis, efficiency, and care quality while emphasizing the need for human oversight and transparency in decision-making [26]. In this study, participants identified improved access to information as a potential benefit, including greater access to information about specific conditions and potential contraindications or medication side effects. Similar findings have been reported elsewhere, including research on AI chatbots for older adults, which identified improved access to easy-to-understand medication and adverse reaction information as a key benefit [27]. However, prior work has also emphasized that patients require a certain level of health literacy to fully benefit from such technologies [28].

Our findings presented here also highlight the potential risks. Particular concerns persist around the use of AI, in terms of data quality, bias, privacy, and the risk of diminishing human interaction and patient autonomy [26]. In this study, privacy and data security were raised by a small number of participants, although not universally. Other research has documented patient concerns regarding privacy and data security risks [29,30]. Bias in AI may arise from the sources of data or existing inequities embedded in datasets [31,32], including evidence of bias in AI risk prediction tools that led to inequitable allocation of health care resources for Black patients [33]. These findings highlight the need to address bias in AI design and implementation. Patients with MLTC in primary care expressed concerns and expectations similar to those reported in other health care settings.

Participants also expressed concern about the potential loss of human touch in health care delivery. Many valued the time spent with their GP and worried that AI might make care less personalized. This aligns with other studies [29,30] and may reflect some misunderstandings about different ways in which AI could be used within GP consultations to support personalized recommendations. This gap in understanding reflects broader evidence of limited public awareness of AI and its potential applications in health care [26,34].

This study adds the perspectives of those living with MLTC, a group that has traditionally been underrepresented in discussions around AI and digital health [35]. In addition, the study also focuses on primary care, a setting with enormous perceived potential, but where the use of AI tools has been more limited [36]. Importantly, the findings reflect the UK primary care context, highlighting issues such as GP consultation structure, patient-provider relationships, and NHS-specific resource constraints that may influence AI adoption and implementation.

There are a number of implications for policy and practice. First, AI development should involve a diverse range of patients, particularly those with lived experience of complex MLTC, who can provide unique perspectives [37]. There are also implications around the need for AI tools to enhance patient-clinician decision-making, while also recognizing the importance and value of human interaction [29]. This highlights the importance of building trust into the design and implementation of AI, particularly by supporting patient health literacy and providing guided access to relevant and authoritative tools [28]. Integration within GP consultations may help clinicians interpret AI outputs, reinforcing trust and joint decision-making.

Concerns around privacy, security, and transparency remain important, but AI also has the potential to increase accuracy and personalization, particularly for MLTC, where holistic approaches to care are needed [38]. Further consultation and reassurance are needed for the public to trust that their data are being used appropriately by AI tools. Clear governance arrangements and robust data protection safeguards are required to manage how patient data are collected, stored, and processed,

XSL•FO
RenderX

including ensuring compliance with legislation, such as the General Data Protection Regulation.

Participants provided important insights into how current care pathways and research often fail to adequately represent diverse populations, including different ages, ethnicities, migrant status, neurodiversity, disability, mental health conditions, or MLTC. This research further highlights the need for strengthened efforts to address potential biases and inequalities, using collaborative and multifaceted approaches that engage minority and marginalized communities to assess whether AI tools meet their needs or risk exacerbating disparities [39].

Strengths and Limitations

This study has several strengths. The study used a broad recruitment strategy through the CPRD database search and outreach through community, patient, and third-sector organizations. There was a robust approach to data collection and analysis, with the involvement of a multidisciplinary research team. Patient and public involvement was embedded throughout the study, with insights informing study design, recruitment, and analysis. This enhanced the credibility and applicability of findings by ensuring that their perspectives shaped study design, interpretation, and implementation [40].

There were several limitations to the research. The low response rate may reflect a lack of interest in or confidence discussing AI-related topics among the wider public. As noted elsewhere, individuals without direct experience of AI may find it difficult to engage with such topics, particularly when examples of AI applications are not clearly defined [41]. It is also likely that those who chose to participate had a particular interest in the subject, meaning the sample may not be fully representative of the wider patient group with MLTC. In addition, although there was some diversity in the participants interviewed, there were fewer young people, men, and people from ethnic minorities. There might have been some views and perspectives that were not included in this analysis, which is important given the potential implications of biases in AI [31]. More research is needed with different demographics, for example, young people and ethnic minorities, where perceptions of AI may vary. Furthermore, the lower-than-expected response to the research may reflect the demands of a 60-minute interview, with potential for a further interview later. Patient burden should be considered for future studies.

Although ethical approval was obtained, ethical considerations were also embedded within the study design. Prior to the interviews, participants received information about AI, and there was also information within the PISs. This may have framed participants' perceptions about AI. For example, we did not specifically raise ethical issues around the use of AI in this information. As this is an area of rapid change, people's views and perceptions may change over time, but exploring this was beyond the scope of the study. A more in-depth exploration of ethical concerns related to AI implementation warrants future research.

Conclusion

This study provides novel findings into patient perspectives on the use of AI to support primary care consultations for individuals living with MLTC. Participants identified several potential benefits, including greater efficiency and improved accuracy and greater access to information. However, they also expressed concerns about potential risks, such as reduced human interaction, data privacy, and the accuracy of AI and its potential to exacerbate inequalities. The findings highlight the need for transparency in AI data use and clear communication and public engagement to build understanding and trust in AI use within health care. Further research is needed to explore how AI might influence clinical interactions in primary care and patient confidence in these tools over time, including longitudinal studies to examine how perceptions change with increased exposure and use. Future work should also prioritize the inclusion of more diverse populations to better understand how experiences and the impact of AI may differ across demographic and socioeconomic groups.

Acknowledgments

The authors would like to thank all participants for giving their time to take part in this study. They also thank the participant advisory group members for their valuable input, including reviewing participant-facing materials and helping to guide the development of the study. They would like to acknowledge the Clinical Practice Research Datalink (CPRD) for facilitating participant recruitment through their patient referral service and Jamie Soames for his support in coordinating the CPRD recruitment process.

Funding

This study was funded by the National Institute for Health and Care Research (NIHR) (NIHR202632) through its Artificial Intelligence for Multiple Long-Term Conditions programme. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

According to the study's ethical approvals, only the research team members responsible for analyzing the data have access to the full anonymized transcripts. The data are not publicly available due to privacy and ethical restrictions. Requests for access to the anonymized data can be submitted to the University of Birmingham Research Governance team at researchgovernance@contacts.bham.ac.uk.

XSL•FO
RenderX

Authors' Contributions

KN, TJ, SH, FLC, SF, TM, LJ, and SG designed the idea of the project. SF, SG, JC, EH, FLC, KN, TJ, and LJ designed the topic guide and the case study. SG and SF wrote the study protocol and applied for ethics approval. JC, EH, SG, and TM designed the concept of the study. SF coordinated and conducted participant recruitment and interviews, had primary oversight of the analyses, and drafted the initial manuscript. JC coded a subset of transcripts to ensure consistency in coding. CS and LJ led the manuscript writing and final edits. LF coordinated the involvement of the patient advisory group and designed the figures. All authors contributed to manuscript revisions, approved the final version, and had final responsibility for the decision to submit for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview topic guide.
[DOCX File, 32 KB - jmir_v28i1e87507_app1.docx]

Multimedia Appendix 2
Participant information sheet.
[DOCX File, 50 KB - jmir_v28i1e87507_app2.docx]

Multimedia Appendix 3
Community recruitment poster.
[PDF File, 305 KB - jmir_v28i1e87507_app3.pdf]

Multimedia Appendix 4
Reflexivity statement.
[DOCX File, 15 KB - jmir_v28i1e87507_app4.docx]

Checklist 1
COREQ checklist.
[DOCX File, 22 KB - jmir_v28i1e87507_app5.docx]

References

1. Chowdhury SR, Chandra Das D, Sunna TC, Beyene J, Hossain A. Global and regional prevalence of multimorbidity in the adult population in community settings: a systematic review and meta-analysis. EClinicalMedicine 2023 Mar;57:101860. [doi: 10.1016/j.eclinm.2023.101860] [Medline: 36864977]
2. Valabhji J, Barron E, Pratt A, et al. Prevalence of multiple long-term conditions (multimorbidity) in England: a whole population study of over 60 million people. J R Soc Med 2024 Mar;117(3):104-117. [doi: 10.1177/01410768231206033] [Medline: 37905525]
3. Pati S, MacRae C, Henderson D, Weller D, Guthrie B, Mercer S. Defining and measuring complex multimorbidity: a critical analysis. Br J Gen Pract 2023 Aug;73(733):373-376. [doi: 10.3399/bjgp23X734661] [Medline: 37500453]
4. Baker JM, Grant RW, Gopalan A. A systematic review of care management interventions targeting multimorbidity and high care utilization. BMC Health Serv Res 2018 Jan 30;18(1):65. [doi: 10.1186/s12913-018-2881-8] [Medline: 29382327]
5. Soley-Bori M, Ashworth M, Bisquera A, et al. Impact of multimorbidity on healthcare costs and utilisation: a systematic review of the UK literature. Br J Gen Pract 2021 Jan;71(702):e39-e46. [doi: 10.3399/bjgp20X713897] [Medline: 33257463]
6. Larkin J, Foley L, Smith SM, Harrington P, Clyne B. The experience of financial burden for people with multimorbidity: a systematic review of qualitative research. Health Expect 2021 Apr;24(2):282-295. [doi: 10.1111/hex.13166] [Medline: 33264478]
7. Hanlon P, Hannigan L, Rodriguez-Perez J, et al. Representation of people with comorbidity and multimorbidity in clinical trials of novel drug therapies: an individual-level participant data analysis. BMC Med 2019 Nov 12;17(1):201. [doi: 10.1186/s12916-019-1427-1] [Medline: 31711480]
8. Ng DTK, Leung JKL, Chu KWS, Qiao MS. AI literacy: definition, teaching, evaluation and ethical issues. Proc Assoc Inf Sci Technol 2021 Oct;58(1):504-509. [doi: 10.1002/pra2.487]
9. Arora A, Alderman JE, Palmer J, et al. The value of standards for health datasets in artificial intelligence-based applications. Nat Med 2023 Nov;29(11):2929-2938. [doi: 10.1038/s41591-023-02608-w] [Medline: 37884627]

XSL•FO
RenderX

10. Khosravi M, Zare Z, Mojtabaeian SM, Izadi R. Artificial intelligence and decision-making in healthcare: a thematic analysis of a systematic review of reviews. Health Serv Res Manag Epidemiol 2024;11:23333928241234863. [doi: 10.1177/23333928241234863] [Medline: 38449840]
11. Kueper JK, Terry AL, Zwarenstein M, Lizotte DJ. Artificial intelligence and primary care research: a scoping review. Ann Fam Med 2020 May;18(3):250-258. [doi: 10.1370/afm.2518] [Medline: 32393561]
12. Gunathilaka NJ, Gooden TE, Cooper J, et al. Perceptions on artificial intelligence-based decision-making for coexisting multiple long-term health conditions: protocol for a qualitative study with patients and healthcare professionals. BMJ Open 2024 Feb 1;14(2):e077156. [doi: 10.1136/bmjopen-2023-077156] [Medline: 38307535]
13. Khullar D, Casalino LP, Qian Y, Lu Y, Krumholz HM, Aneja S. Perspectives of patients about artificial intelligence in health care. JAMA Netw Open 2022 May 2;5(5):e2210309. [doi: 10.1001/jamanetworkopen.2022.10309] [Medline: 35507346]
14. Moy S, Irannejad M, Manning SJ, et al. Patient perspectives on the use of artificial intelligence in health care: a scoping review. J Patient Cent Res Rev 2024;11(1):51-62. [doi: 10.17294/2330-0698.2029] [Medline: 38596349]
15. Archibald MM, Ambagtsheer RC, Casey MG, Lawless M. Using Zoom videoconferencing for qualitative data collection: perceptions and experiences of researchers and participants. Int J Qual Methods 2019 Jan 1;18:1609406919874596. [doi: 10.1177/1609406919874596]
16. Wolf A, Dedman D, Campbell J, et al. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. Int J Epidemiol 2019 Dec 1;48(6):1740-1740g. [doi: 10.1093/ije/dyz034] [Medline: 30859197]
17. Shiekh SI, Harley M, Ghosh RE, et al. Completeness, agreement, and representativeness of ethnicity recording in the United Kingdom's Clinical Practice Research Datalink (CPRD) and linked Hospital Episode Statistics (HES). Popul Health Metr 2023 Mar 14;21(1):3. [doi: 10.1186/s12963-023-00302-0] [Medline: 36918866]
18. CPRD reach. CPRD (Clinical Practice Research Datalink). URL: https://www.cprd.com/research-services/cprd-reach [accessed 2025-11-05]
19. Gokhale KM, Chandan JS, Toulis K, Gkoutos G, Tino P, Nirantharakumar K. Data extraction for epidemiological research (DExtER): a novel tool for automated clinical epidemiology studies. Eur J Epidemiol 2021 Feb;36(2):165-178. [doi: 10.1007/s10654-020-00677-6] [Medline: 32856160]
20. Rahimi S, Khatooni M. Saturation in qualitative research: an evolutionary concept analysis. Int J Nurs Stud Adv 2024 Jun;6:100174. [doi: 10.1016/j.ijnsa.2024.100174] [Medline: 38746797]
21. Negrin KA, Slaughter SE, Dahlke S, Olson J. Successful recruitment to qualitative research: a critical reflection. Int J Qual Methods 2022 Apr;21:16094069221119576. [doi: 10.1177/16094069221119576]
22. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Aug 8;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
23. Saldaña J. An introduction to themeing the data. In: Expanding Approaches to Thematic Analysis: Routledge; 2024:11-26.
24. Witkowski K, Dougherty RB, Neely SR. Public perceptions of artificial intelligence in healthcare: ethical concerns and opportunities for patient-centered care. BMC Med Ethics 2024 Jun 22;25(1):74. [doi: 10.1186/s12910-024-01066-4] [Medline: 38909180]
25. Gundlack J, Thiel C, Negash S, et al. Patients' perceptions of artificial intelligence acceptance, challenges, and use in medical care: qualitative study. J Med Internet Res 2025 May 15;27:e70487. [doi: 10.2196/70487] [Medline: 40373300]
26. Richardson JP, Smith C, Curtis S, et al. Patient apprehensions about the use of artificial intelligence in healthcare. NPJ Digit Med 2021 Sep 21;4(1):140. [doi: 10.1038/s41746-021-00509-1] [Medline: 34548621]
27. Gudala M, Ross MET, Mogalla S, Lyons M, Ramaswamy P, Roberts K. Benefits of, barriers to, and needs for an artificial intelligence-powered medication information voice chatbot for older adults: interview study with geriatrics experts. JMIR Aging 2022 Apr 28;5(2):e32169. [doi: 10.2196/32169] [Medline: 35482367]
28. Schulz PJ, Nakamoto K. Patient behavior and the benefits of artificial intelligence: the perils of "dangerous" literacy and illusory patient empowerment. Patient Educ Couns 2013 Aug;92(2):223-228. [doi: 10.1016/j.pec.2013.05.002] [Medline: 23743214]
29. Young AT, Amara D, Bhattacharya A, Wei ML. Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. Lancet Digit Health 2021 Sep;3(9):e599-e611. [doi: 10.1016/S2589-7500(21)00132-1] [Medline: 34446266]
30. Nelson CA, Pérez-Chada LM, Creadore A, et al. Patient perspectives on the use of artificial intelligence for skin cancer screening: a qualitative study. JAMA Dermatol 2020 May 1;156(5):501-512. [doi: 10.1001/jamadermatol.2019.5014] [Medline: 32159733]
31. Gichoya JW, Thomas K, Celi LA, et al. AI pitfalls and what not to do: mitigating bias in AI. Br J Radiol 2023 Oct;96(1150):20230023. [doi: 10.1259/bjr.20230023] [Medline: 37698583]
32. Agarwal R, Bjarnadottir M, Rhue L, et al. Addressing algorithmic bias and the perpetuation of health inequities: an AI bias aware framework. Health Policy Technol 2023 Mar;12(1):100702. [doi: 10.1016/j.hlpt.2022.100702]
33. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]

XSL•FO
RenderX

34. Vo V, Chen G, Aquino YSJ, Carter SM, Do QN, Woode ME. Multi-stakeholder preferences for the use of artificial intelligence in healthcare: a systematic review and thematic analysis. Soc Sci Med 2023 Dec;338:116357. [doi: 10.1016/j.socscimed.2023.116357] [Medline: 37949020]
35. Yi M, Hui Y, Hu L, Zhang W, Wang Z. The experiences and perceptions of older adults with multimorbidity toward e-health care: a qualitative evidence synthesis. Telemed J E Health 2024 Oct;30(10):2527-2544. [doi: 10.1089/tmj.2024.0211] [Medline: 38920002]
36. Katonai G, Arvai N, Mesko B. AI and primary care: scoping review. J Med Internet Res 2025 Aug 15;27:e65950. [doi: 10.2196/65950] [Medline: 40836794]
37. Badawy W, Zinhom H, Shaban M. Navigating ethical considerations in the use of artificial intelligence for patient care: a systematic review. Int Nurs Rev 2025 Sep;72(3):e13059. [doi: 10.1111/inr.13059] [Medline: 39545614]
38. Koirala B, Peeler A, Dennison Himmelfarb C, Davidson PM. Living with multiple chronic conditions: how we achieve holistic care and optimize health outcomes. J Adv Nurs 2023 Feb;79(2):e7-e9. [doi: 10.1111/jan.15433] [Medline: 36062872]
39. Wang T, Emami E, Jafarpour D, Tolentino R, Gore G, Rahimi SA. Integrating equity, diversity, and inclusion throughout the lifecycle of artificial intelligence for healthcare: a scoping review. PLOS Digit Health 2025 Jul;4(7):e0000941. [doi: 10.1371/journal.pdig.0000941] [Medline: 40658719]
40. Rahman A, Nawaz S, Khan E, Islam S. Nothing about us, without us: is for us. Res Involv Engagem 2022 Aug 4;8(1):39. [doi: 10.1186/s40900-022-00372-8] [Medline: 35927767]
41. Scott IA, Carter SM, Coiera E. Exploring stakeholder attitudes towards AI in clinical practice. BMJ Health Care Inform 2021 Dec;28(1):e100450. [doi: 10.1136/bmjhci-2021-100450] [Medline: 34887331]

Abbreviations

AI: artificial intelligence
CPRD: Clinical Practice Research Datalink
DExtER: Data Extraction for Epidemiological Research
GP: general practitioner
MLTC: multiple long-term conditions
NHS: National Health Service
OPTIMAL: Optimizing Therapies, Disease Trajectories, and Artificial Intelligence-Assisted Clinical Management for Patients Living With Complex Multimorbidity
PAG: patient advisory group
PIS: participant information sheet

Edited by A Stone; submitted 19.Nov.2025; peer-reviewed by A Yadav, A Okolie, WG Hussain; revised version received 19.Mar.2026; accepted 19.Mar.2026; published 21.Apr.2026.

Please cite as:
Flanagan S, Spurway C, Jackson L, Cooper J, Crowe FL, Haroon S, Marshall T, Fitzsimmons L, Hathaway E, Nirantharakumar K, Jackson T, Greenfield S
Exploring Patient Perspectives on the Use of Artificial Intelligence to Inform Joint Decision-Making for Patients With Multiple Conditions in Primary Care in the United Kingdom: Qualitative Study
J Med Internet Res 2026;28:e87507
URL: https://www.jmir.org/2026/1/e87507
doi:10.2196/87507

© Sarah Flanagan, Charlotte Spurway, Louise Jackson, Jenny Cooper, Francesca L Crowe, Shamil Haroon, Tom Marshall, Leah Fitzsimmons, Eleanor Hathaway, Krishnarajah Nirantharakumar, Thomas Jackson, Sheila Greenfield. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Time-Dynamic AI Models to Predict Quality of Life in Patients With Breast Cancer: Development and Validation Study Using the EORTC BALANCE Cohort

Niclas J Hubel[1], MSc; Thijs G W van der Heijden[2], MSc; Benjamin Murauer[3], PhD; Belle H de Rooij[4,5], PhD; Kelly M de Ligt[2], PhD; Helena M Verkooijen[6], Prof Dr; Sofie AM Gernaat[6], PhD; Meeke Hoedjes[7], PhD; Volker Arndt[8], Prof Dr; Lonneke V van de Poll-Franse[2,7], Prof Dr; Bernhard Holzner[1,3], Prof Dr; Jens Lehmann[1,9], PhD

[1]Health Outcomes Research Unit, University Hospital of Psychiatry II, Medical University of Innsbruck, Anichstrasse 35, Innsbruck, Austria

[2]Department of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute, Amsterdam, The Netherlands

[3]Evaluation Software Development GmbH, Innsbruck, Austria

[4]Center of Research on Psychological Disorders and Somatic Diseases, Department of Medical and Clinical Psychology, Tilburg University, Tilburg, The Netherlands

[5]Department of Research and Development, Netherlands Comprehensive Cancer Organisation (IKNL), Utrecht, The Netherlands

[6]Division of Imaging and Oncology, University Medical Centre Utrecht, Utrecht, The Netherlands

[7]Center of Research on Psychological Disorders and Somatic Disorders, Department of Medical and Clinical Psychology, Tilburg University, Tilburg, The Netherlands

[8]Cancer Survivorship Outcomes and Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany

[9]on behalf of the EORTC Quality of Life Group

Corresponding Author:
Niclas J Hubel, MSc
Health Outcomes Research Unit, University Hospital of Psychiatry II, Medical University of Innsbruck, Anichstrasse 35, Innsbruck, Austria

Abstract

Background: Patients with breast cancer often experience health-related quality of life (HRQoL) impairments that remain difficult to predict on an individual level. Prediction models can aid in understanding individual survivorship trajectories. However, current prognostic models are based on fixed intervals, limiting their utility in clinical follow-up schedules.

Objective: This study aimed to develop and externally validate time-dynamic machine learning (ML) models that predict clinically relevant HRQoL impairments in nonmetastatic patients with breast cancer.

Methods: Using the pooled multicohort EORTC (European Organisation for Research and Treatment of Cancer) BALANCE (big data in patients with breast cancer) dataset (n=6316) containing repeated HRQoL measurements (EORTC QLQ [Quality of Life Core Questionnaire]-C30), we constructed over 70,000 patient assessment pairs. ML algorithms were trained using the earlier HRQoL assessment and clinical data to predict dichotomized impairments in QLQ-C30 domains at the later assessment between 2 weeks and 5 years ahead, reflecting the range of follow-up intervals available in the dataset. The best performing model was determined via the area under the receiver operating characteristic curve in the internal validation, and externally validated in an independent cohort of the BALANCE dataset, in which the calibration and predictive performance in risk groups (patients: postmenopause, with financial difficulties, with obesity, with 2 or more comorbidities, with lower educational status, and with frailty) were also evaluated.

Results: ML models showed good discrimination (area under the receiver operating characteristic curve 0.64 - 0.84) across most domains, especially for persistent symptoms such as fatigue, financial difficulties, or functioning scales. Gradient boosting models performed best, but tended to be overconfident, with poor calibration for low-prevalence symptoms such as diarrhea or constipation. Model performance varied by risk group (eg, lower education and frailty), though no group consistently performed poorly. Performance remained stable across time windows, with prior HRQoL being the strongest predictor at the respective scale level, while clinical variables such as the type of treatment were less important for prediction.

Conclusions: Time-dynamic ML models can support personalized HRQoL prediction in breast cancer care. Future improvements should focus on calibration and fairness to enable equitable, clinically meaningful implementation.

(J Med Internet Res 2026;28:e81424) doi:10.2196/81424

XSL•FO
RenderX

KEYWORDS

machine learning; health-related quality of life; HRQoL; breast cancer; prediction modeling; patient-reported outcomes

Introduction

Breast cancer remains one of the most common malignancies among women worldwide, and although survival rates have improved significantly, many patients continue to experience long-term physical, emotional, and psychosocial consequences, stemming from both the disease and its treatment [1]. These effects can have a profound impact on health-related quality of life (HRQoL) across survivorship trajectories [2,3]. Despite its clinical importance, clinicians often lack individualized insights into how HRQoL evolves across the cancer trajectory. This gap contributes to persistent unmet supportive care needs among patients with breast cancer, especially in the posttreatment phase [4]. Without clear prognostic guidance on expected HRQoL changes, opportunities for patient-centered care and shared decision-making may be missed, potentially leading to suboptimal follow-up strategies and delayed interventions [5].

Machine learning (ML) offers a promising avenue to address this gap when incorporating real-world data (RWD). Techniques such as deep learning and gradient boosting can uncover complex, nonlinear relationships within high-dimensional data and provide personalized HRQoL predictions [6]. Existing ML models in breast cancer have largely focused on traditional clinical endpoints such as survival or recurrence [7]. In contrast, HRQoL remains underexplored as an outcome, despite being, alongside overall survival, one of the top priorities for patients [8,9].

However, several key challenges remain in applying ML to HRQoL prediction. First, structured, large-scale datasets capturing longitudinal HRQoL data are scarce [7,10]. In addition, existing datasets are highly heterogeneous: clinical trial data, RWD, and observational studies differ substantially in structure, completeness, and context. While integrating multiple data sources may enhance generalizability, it also introduces complex challenges related to data harmonization and standardization, making analyses more difficult [11].

A further limitation of most existing models is their reliance on fixed prediction intervals. Typically, HRQoL is predicted at standardized time points, such as 6 or 12 months posttreatment, without accounting for variability in patients' assessment schedules or individual disease trajectories [7,12-15]. Yet in real-world practice, assessments often occur at irregular intervals shaped by patient needs and clinical routines. Static models thus fail to reflect the variability of patient journeys. Therefore, a time-dynamic prediction approach is warranted. Such a model would allow clinicians to predict HRQoL at any future point in time based on an individual's prior trajectory, time since last assessment, and current clinical context, supporting more timely and personalized decision-making.

Finally, ML models may inadvertently exacerbate disparities in care [16-18]. Demographic and socioeconomic differences in HRQoL reporting, combined with the underrepresentation of certain populations in training data, can lead to biased predictions and inequitable care [17]. For example, models trained predominantly on data from phase III trials in high-income countries may underperform when applied to minority populations or patients in lower-resource settings, where symptom burden, health literacy, and reporting behaviors differ significantly [19]. At present, however, it remains unclear to what extent HRQoL prediction models are affected by such biases, as empirical investigations into fairness or subgroup performance in this context are still scarce.

The present study aims to develop and validate time-dynamic ML models for predicting HRQoL in patients with nonmetastatic breast cancer from the BALANCE (big data in patients with breast cancer) cohort [11]. We compare multiple ML algorithms against a baseline model and assess the impact of data heterogeneity on predictive performance. Moreover, we investigate the presence of algorithmic bias to advance fair, flexible, and clinically relevant HRQoL prediction tools for breast cancer care.

Methods

Study Design

In this study, we developed and externally validated ML algorithms to predict clinically important impairment in HRQoL within multiple pooled datasets. We report our findings according to the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) guidelines [20] and the Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research [21].

Ethical Considerations

Ethical approval for the secondary data analysis was obtained in 2022 at the Antoni van Leeuwenhoekziekenhuis/Nederlands Kanker Instituut from the institutional review board (IRBd22-179).

For the cohorts: ethical approval for the OPTIMUM (Towards Optimal Timing and Method for Promoting Sustained Adherence to Lifestyle and Body Weight Recommendations in Postmenopausal Breast Cancer Survivors) study was obtained from the Medical Research Ethics Committee Brabant, the Netherlands (NL66913.028.18). The VERDI (Verlauf der Diagnostischen Abklärung) study was approved by the ethics committees of the University of Heidelberg and the Medical Association of Saarland, Germany. Written informed consent was obtained from all participants. The study protocol for UMBRELLA (Utrecht Cohort for Multiple Breast Cancer Intervention Studies and Long-Term Evaluation) was approved by the Institutional Review and Ethics Board of the University Medical Center Utrecht, the Netherlands. No further compensation for the original participants was provided as their identity was not known to us. Patient data was processed only in pseudonymized form.

XSL•FO
RenderX

Dataset

We described the data pooling previously in more detail [11]. In brief, the BALANCE dataset comprises 6 cohorts with a total of 6316 female patients with nonmetastatic breast cancer, including trial data (EORTC [European Organisation for Research and Treatment of Cancer] AMAROS [After Mapping of the Axilla, Radiotherapy or Surgery?] [22]), RWD (Netherlands Cancer Institute [NKI] [23], district hospital Kufstein, Austria [24]), and observational studies (UMBRELLA [25], OPTIMUM [26], and VERDI [27]), with at least 2 HRQoL assessments (Figure 1). Data were collected between 2001 and 2024 and encompass patients receiving active treatment, follow-up, or survivorship care.

Figure 1. Patient flowchart. AMAROS: After Mapping of the Axilla, Radiotherapy or Surgery?; KUFSTEIN: District Hospital Kufstein cohort; NKI: Netherlands Cancer Institute; OPTIMUM: Towards Optimal Timing and Method for Promoting Sustained Adherence to Lifestyle and Body Weight Recommendations in Postmenopausal Breast Cancer Survivors; UMBRELLA: Utrecht Cohort for Multiple Breast Cancer Intervention Studies and Long-Term Evaluation; VERDI: Verlauf der Diagnostischen Abklärung.

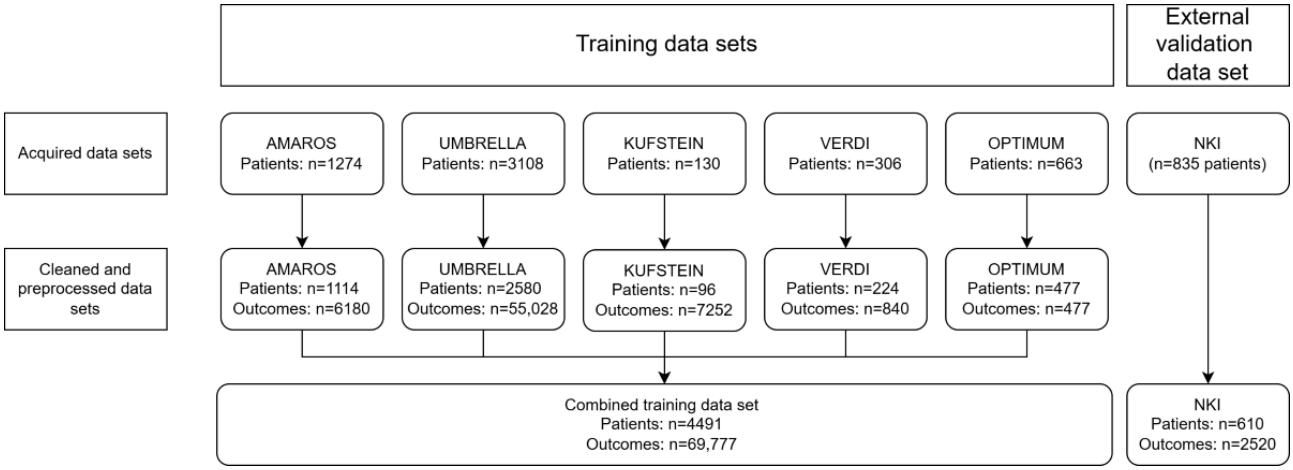

Data Preparation

Data cleaning and preprocessing were conducted using R (R Foundation) [28], Jupyter Notebook, and Python (Python Software Foundation) libraries (Methods S1) in Multimedia Appendix 1 were used for statistical analyses and ML algorithms.

Features include patient characteristics and sociodemographics, comorbidities, disease and treatment parameters, as well as the items of the EORTC QLQ (Quality of Life Core Questionnaire)-C30 [29] and breast cancer module QLQ-BR23 [30]. Variables were either measured at the time of diagnosis or at the first HRQoL assessment. Cohort data were harmonized based on the common codebook (see Methods S2) in Multimedia Appendix 1. Briefly, time-related variables were coded in days since the date of diagnosis to ensure consistency across cohorts. Treatment characteristics were dichotomized into binary categories, and comorbidities were summarized as simple counts derived from available diagnostic information. Disease-related variables (eg, tumor stage and receptor status) were largely consistent across cohorts.

Features were preprocessed by one-hot encoding categorical variables (missing values were treated as distinct categories) to ensure they are represented numerically without implying an ordinal relationship and by normalizing all numeric features to 0 mean and unit variance to standardize their range, improving model performance and comparability. We removed features with 0 or near-0 variance (<0.001) from the analysis and imputed missing feature values using the iterative imputer implementation from scikit-learn. It estimates missing values in an iterated round-robin fashion as a function of other features based on a Bayesian ridge estimator [31].

To evaluate time-dynamic (ie, predicting HRQoL at every available time point using prior data and time since last assessment) predictive performance, for each patient, all possible combinations of their assessments were compiled into distinct pairs. Observations lacking either HRQoL data or the corresponding time point information were excluded. A new variable was introduced to represent the difference in days between the assessment dates within each pair. The outcome variable for each pair was defined as the value from the later assessment. For example (see Methods S3 in Multimedia Appendix 1 for an illustration), if a patient had 3 assessments, A1, A2, and A3, occurring at different time points, pairs such as (A1, A2), (A1, A3), and (A2, A3) were generated. The earlier assessment in each pair contributed its single items to the dataset, while the later assessment provided the target outcome based on the dichotomized scales (see subsection Outcomes under the Methods section). No other data from the second assessment in each pair was included in the input features. Each pair (A_m, A_n) is represented as a training sample (X_m, t_{m-n}, Y_n), where X_m are the features of the assessment m, Y_n is the binary outcome variable of the respective observation, and t_{m-n} is the time difference between A_m and A_n in days. Each training sample can be interpreted as "t_{m-n} days after the patient has reported results X_m, the patient's outcome is Y_n." This approach allowed us to model changes in HRQoL over time. Figure S1 in Multimedia Appendix 1 shows the distribution of the time difference between assessment pairs in the train and test set.

Outcomes

HRQoL was measured with the EORTC QLQ-C30 before, during, or after treatment. Symptom (fatigue, nausea and vomiting, pain, dyspnea, insomnia, appetite loss [AP],

XSL•FO
RenderX

constipation, diarrhea, and financial difficulties) and functioning scales (physical, role, emotional, cognitive, and social) are reported on a 4-point Likert scale ("not at all," "a bit," "quite a bit," and "very much") and transformed to linear scores ranging from 0 to 100 scale [32]. We dichotomized the outcomes according to the established thresholds for clinical importance to indicate clinically relevant impairments, defined as scores associated with at least one of the following patient-reported concerns: limitations in daily life, need for help or care, and worries of the patient or their partner or family [33]. The dichotomization simplifies clinical interpretation through established and meaningful thresholds.

We did not adjust the dataset for class imbalance as most outcome variables were balanced (see Results section), and adjusting for class imbalance would increase the risk for overfitting.

Algorithms

We divided our dataset into model development data (5 cohorts) and external validation data (NKI). The NKI set was chosen as it contains the most current data from clinical practice and a sufficient number of patients.

We used the scikit-learn implementations of algorithms from multiple classifier families [34] that proved effective in similar studies [6,12,35,36]. We evaluated a logistic regression with L2 regularization, an extra-trees classifier, a multilayer perceptron classifier, and a histogram-based gradient boosting classification tree (see Methods S3 in Multimedia Appendix 1 for a detailed description of libraries and GitHub for the code base). To account for the unequal contribution of patients with varying numbers of assessment pairs, we additionally incorporated sample weights inversely proportional to the number of pairs per patient. The final model was recalibrated using a probability calibration with logistic regression from scikit-learn (CalibratedClassifierCV, default settings) using a 5-fold cross-validation based on the training data. In addition, we performed post hoc recalibration on the external validation set to illustrate the extent to which context-specific adjustment can enhance model performance before deployment in a clinical setting.

Evaluation

To assess model performance, we used internal cross-validation using a 5-fold (K=5) strategy. Given the differences in sample sizes across cohorts, the internal cross-validation was configured to keep the distribution of the cohorts within the train and test splits consistent. Thereby, the internal cross-validation was performed in a way that ensures an 80/20 train and test split and also prevents any patient overlap between train and test data.

Model performance was primarily evaluated using the area under the receiver operating characteristic curve (AUC), which is robust against class imbalance [37]. Training with AUC can provide acceptable performance at a 1/10 imbalance ratio without rebalancing [38]. Additional metrics included weighted F_1-score, accuracy, balanced accuracy, and confusion matrix metrics (sensitivity, specificity, positive predictive value, and negative predictive value) based on their scikit-learn implementation using the mean and SD from all internal cross-validation iterations [34]. Accuracy reflects the overall proportion of correct predictions but can be misleading when classes are imbalanced. Balanced accuracy addresses this by averaging the recall obtained on each class, giving equal weight regardless of class frequency. The F_1-score summarizes overall classification performance by balancing false positives and false negatives, with the weighted F_1-score giving more weight to classes with more instances. Sensitivity (recall) quantifies the model's ability to correctly identify patients with an impairment, whereas specificity reflects the ability to correctly identify those without impairment. Positive predictive value indicates the probability that patients predicted to have an impairment truly experience it, while negative predictive value reflects the probability that patients predicted not to have an impairment are indeed unaffected. Higher scores indicate better overall model performance across all metrics [39].

Last observation carried forward (LOCF) was used as a baseline for model comparison. LOCF serves as a straightforward method for estimating future values where patients' HRQoL is not expected to change. These absolute values were used to compute AUC scores. It has been used in clinical trials and longitudinal studies as an imputation method or to predict future HRQoL [40,41]. Using this simple approach, therefore, provides a better baseline performance threshold compared to random chance (AUC=0.5). Furthermore, simplified logistic regression models with the prior HRQoL and the time difference as their only 2 predictors were computed to illustrate gains beyond LOCF.

Based on the internal cross-validation, we used the best-performing model for each target variable for external validation in the NKI dataset (Figure 1). Bootstrapping over 1000 iterations was used to estimate the stability of the external validation and LOCF, including 95% CIs [42].

To interpret model predictions, we applied permutation feature importance, quantifying the contribution of each feature to the trained models. Model calibration was evaluated with calibration plots, the calibration plot's slope and intercept, and the expected calibration error (ECE), which assesses how closely predicted probabilities align with observed outcomes; good calibration is indicated by a calibration curve close to the diagonal, an intercept near zero, a slope approaching one, and a low ECE [39,43]. The ECE was based on a kernel-smoothed function using the relplot Python library [44]. Additionally, the Brier score was computed to assess the overall performance, combining discrimination and calibration, with lower values near 0 representing better performance [45].

Decision curve analyses were conducted to determine the clinical utility of the models [46]. Predicted risks may inform patient counseling, leading to heterogeneous downstream actions, such as additional assessment, supportive care, or monitoring, which differ in burden and resource use and cannot be directly compared. Evaluating net benefit across a wide threshold range (ie, the predicted risk at which a clinician would consider acting), therefore, allows assessment of model utility under varying clinical preferences and use scenarios. In this framework, the treat-all and treat-none strategies represent reference scenarios in which all patients or no patients, respectively, are considered at risk. A model is considered to

XSL•FO
RenderX

have clinical utility at threshold probabilities where its decision curve yields a higher net benefit than both reference strategies, indicating more favorable trade-offs between identifying patients at risk and the use of health care resources [47].

Model Fairness

In line with our aim of developing fair HRQoL prediction tools, and in accordance with recommendations to ensure equitable model performance across diverse populations [16], we defined risk groups within the external validation dataset to address group fairness concerns. The following risk groups were selected based on prior literature [48] and evaluated in the same manner as the full external dataset: (1) postmenopause, (2) financial difficulties (according to the thresholds for clinical importance for the financial difficulties scale [33] at baseline), (3) obesity (BMI≥30), (4) 2 or more comorbidities, (5) lower educational status (secondary education or lower), and (6) frailty (scoring according to Murugappan et al [49]). Additionally, to the previously described metrics, we evaluated true and false positive rates to screen for equalized odds disparities.

Time-Dynamic Performance Evaluation

Further, we split the validation dataset to assess the impact of the time-dynamic aspect on model performance. As we are not relying on fixed prediction intervals, we want to ensure stable model performance across different time horizons. We evaluated long-term prediction defined as predicting outcomes more than 1 year in the future (t_{m-n} larger than 365 days), as such predictions are especially valuable for informing survivorship care planning and long-term patient management. Further, we selected all outcome pairs from within the first year after diagnosis, as this period typically encompasses active treatment, during which most changes in HRQoL are expected to occur. After the first year, HRQoL tends to stabilize, making early assessment particularly relevant for capturing clinically meaningful variation [50,51].

Results

Patient Characteristics

Outcome data from 6316 patients were initially acquired. After cleaning and preprocessing, 4491 patients remained in the model development dataset (Figure 1). Table 1 shows their main characteristics across all included datasets.

XSL•FO
RenderX

Table . Overview of datasets and patient characteristics at baseline. Only valid percentages are shown.

Variable	Overall	Kufstein[a]	UMBRELLA[b]	OPTIMUM[c]	VERDI[d]	AMAROS[e]	NKI[f] (external)
Value (n)	5101	96	2580	477	224	1114	610
RWD[g], n (%)	3987 (78.2)	N/A[h]	N/A	N/A	N/A	N/A	N/A
Country	N/A	Austria	Netherlands	Netherlands	Germany	Multiple[i]	Netherlands
Assessment date in days since diagnosis, mean (SD)	265.93 (655.79)	57.02 (141.33)	361.33 (870.33)	367.07 (35.97)	667.80 (512.69)	54.51 (146.08)	54.72 (122.77)
Number of assessment pairs per patient, mean (SD)	4.62 (3.39)	11.6 (8.76)	5.38 (4.05)	2 (0)	2.81 (1.13)	5.30 (1.86)	2.06 (0.98)
Age (years), mean (SD)	56.70 (10.81)	56.83 (13.55)	55.91 (10.54)	65.33 (7.01)	56.10 (10.88)	55.94 (10.34)	54.92 (11.71)
Postmenopause, n (%)	1535 (67.8)	N/A	N/A	477 (100)	154 (68.8)	648 (58.9)	247 (56.1)
Marital status, n (%)							
Married or living together	1491 (64.2)	6 (54.5)	573 (54.5)	349 (74.7)	151 (67.4)	N/A	412 (72)
Relationship (not married or not living together)	196 (8.4)	0 (0)	134 (12.7)	17 (3.6)	0 (0)	N/A	45 (7.9)
Divorced or split up	306 (13.2)	0 (0)	191 (18.2)	39 (8.4)	16 (7.1)	N/A	60 (10.5)
Widowed	193 (8.3)	0 (0)	80 (7.6)	45 (9.6)	45 (20.1)	N/A	23 (4)
Never married or lived together	137 (5.9)	3 (27.3)	73 (6.9)	17 (3.6)	12 (5.4)	N/A	32 (5.6)
Marital status missing	2778	87	1529	10	0	1114	38
Education status, n (%)							
Primary education or lower	112 (3.4)	N/A	40 (2)	6 (1.3)	61 (27.2)	N/A	5 (0.9)
Secondary education	660 (20.2)	N/A	490 (24.5)	98 (21)	39 (17.4)	N/A	33 (5.7)
Vocational education	907 (27.8)	N/A	435 (21.8)	162 (34.7)	117 (52.2)	N/A	193 (33.6)
University or higher (vocational) education	1586 (48.6)	N/A	1034 (51.7)	201 (43)	7 (3.1)	N/A	344 (59.8)
Education status missing	1836	96	581	10	0	1114	35
BMI, mean (SD)	25.98 (4.46)	25.20 (4.32)	N/A	26.05 (4.64)	26.03 (4.58)	25.94 (4.06)	26.10 (4.97)
Number of comorbidities, mean (SD)	1.73 (1.74)	0.33 (0.47)	N/A	2.48 (2.02)	2.25 (1.48)	N/A	1.13 (1.30)
Differentiation grade, n (%)							
Grade I	1176 (23.6)	7 (7.3)	668 (25.9)	102 (28.6)	17 (7.6)	276 (24.8)	106 (17.4)
Grade II	2239 (45)	39 (40.6)	1123 (43.5)	166 (46.5)	112 (50)	520 (46.7)	279 (45.7)

XSL•FO
RenderX

Variable	Overall	Kufstein[a]	UMBRELLA[b]	OPTIMUM[c]	VERDI[d]	AMAROS[e]	NKI[f] (external)
Grade III	1255 (25.2)	31 (32.3)	557 (21.6)	83 (23.2)	90 (40.2)	294 (26.4)	200 (32.8)
Grade IV	2 (0)	1 (1)	0 (0)	0 (0)	1 (0.4)	0 (0)	0 (0)
Grade unknown	307 (6.2)	17 (17.7)	232 (9)	6 (1.7)	4 (1.8)	24 (2.2)	24 (3.9)
Differentiation grade missing	120	0	0	120	0	0	0
TNM[j] stage, n (%)							
0	196 (4.1)	0 (0)	196 (7.7)	0 (0)	0 (0)	0 (0)	0 (0)
IA	1903 (39.7)	27 (32.9)	1292 (51)	199 (64.6)	58 (33.9)	0 (0)	327 (55.2)
IB	135 (2.8)	0 (0)	135 (5.3)	0 (0)	0 (0)	0 (0)	0 (0)
IIA	1772 (36.9)	37 (45.1)	543 (21.4)	70 (22.7)	63 (36.8)	898 (80.8)	161 (27.2)
IIB	561 (11.7)	1 (1.2)	234 (9.2)	27 (8.8)	38 (22.2)	202 (18.2)	59 (10)
IIIA	176 (3.7)	17 (20.7)	102 (4)	9 (2.9)	8 (4.7)	12 (1.1)	28 (4.7)
IIIB	16 (0.3)	0 (0)	5 (0.2)	2 (0.6)	4 (2.3)	0 (0)	5 (0.8)
IIIC	38 (0.8)	0 (0)	25 (1)	1 (0.3)	0 (0)	0 (0)	12 (2)
TNM stage missing, n (%)	304	14	48	169	53	2	18
ER[k] positive, n (%)	3049 (82.5)	N/A	2130 (83.5)	345 (86.7)	161 (77.8)	N/A	413 (76.3)
PR[l] positive, n (%)	2548 (69.1)	N/A	1828 (71.7)	260 (65.3)	148 (71.8)	N/A	312 (58.1)
HER2NEU[m] status, n (%)							
0	2724 (59.6)	N/A	1439 (57)	238 (60.4)	N/A	1045 (93.9)	2 (0.4)
1+	1020 (22.3)	N/A	739 (29.3)	126 (32)	N/A	68 (6.1)	87 (16.1)
2+	458 (10)	N/A	3 (0.1)	2 (0.5)	N/A	0 (0)	453 (83.6)
3+	372 (8.1)	N/A	344 (13.6)	28 (7.1)	N/A	0 (0)	0 (0)
HER2NEU status missing	527	96	55	83	224	[i]	68
Treatments, n (%)							
Surgery	4985 (98.4)	49 (51)	2572 (99.7)	440 (99.3)	223 (100)	1114 (100)	587 (96.2)
Radiotherapy	4206 (85.4)	1 (1.1)	2376 (92.1)	364 (91.5)	46 (34.8)	995 (89.3)	424 (69.5)
Chemotherapy	2779 (58.4)	94 (98.9)	1196 (46.4)	121 (78.1)	75 (36.1)	706 (63.4)	281 (46.1)
Hormonal therapy	2495 (59)	1 (1.2)	1304 (50.5)	218 (86.5)	75 (37.7)	897 (80.5)	N/A
Immunotherapy	452 (74.1)	N/A	N/A	N/A	N/A	N/A	103 (17)
Frail, n (%)	346 (7.7)	9 (9.5)	293 (11.4)	16 (3.4)	3 (1.3)	25 (2.3)	33 (5.6)
Financial difficulties, n (%)	624 (12.2)	13 (13.5)	357 (13.8)	30 (6.3)	70 (31.2)	75 (6.7)	79 (13)
Obese, n (%)	291 (15.9)	13 (18.1)	N/A	76 (16.3)	51 (23.5)	151 (14.1)	107 (19.6)
Two comorbidities or more, n (%)	444 (55.7)	0 (0)	N/A	298 (62.5)	146 (65.2)	N/A	173 (30)

XSL•FO
RenderX

Variable	Overall	Kufstein[a]	UMBRELLA[b]	OPTIMUM[c]	VERDI[d]	AMAROS[e]	NKI[f] (external)
Lower education status, n (%)	734 (27.3)	N/A	530 (26.5)	104 (22.3)	100 (44.6)	N/A	38 (6.6)

[a]Kufstein: District Hospital Kufstein cohort.

[b]UMBRELLA: Utrecht Cohort for Multiple Breast Cancer Intervention Studies and Long-Term Evaluation.

[c]OPTIMUM: Towards Optimal Timing and Method for Promoting Sustained Adherence to Lifestyle and Body Weight Recommendations in Postmenopausal Breast Cancer Survivors.

[d]VERDI: Verlauf der Diagnostischen Abklaerung.

[e]AMAROS: After Mapping of the Axilla, Radiotherapy or Surgery?

[f]NKI: Netherlands Cancer Institute.

[g]RWD: real world data.

[h]N/A: not applicable.

[i]France, Israel, Italy, Netherlands, Slovenia, Switzerland, Turkey, and United Kingdom.

[j]TNM: tumor, nodes, metastasis.

[k]ER: estrogen receptor.

[l]PR: progesterone receptor.

[m]HER2NEU: Human Epidermal Growth Factor Receptor 2 Neural.

Outcomes

The outcome distribution in the development and validation sets after pairing HRQoL assessments showed a higher prevalence of impairments to HRQoL in the external validation set across all domains compared to the training set (Table 2). Major class imbalance (ratio of event to total number of less than 1/10) occurred in 3/14 scales (AP, constipation, and social functioning) in the training set. The event rates were higher in the external validation set.

Table . Outcome distribution (number of assessments with respective impairments) of the EORTC C30[a] scales in the training and validation set.

	Training set (n=69,777)	External validation set (n=2520)
EORTC QLQ-C30 scale[b]	Assessments with impairments (events)[c], n (%)	Assessments with impairments (events)[c], n (%)
Physical functioning (PF2)	18,261 (26.2)	754 (29.9)
Role functioning (RF2)	9099 (13)	555 (22)
Emotional functioning (EF)	16,046 (23)	1051 (41.7)
Cognitive functioning (CF)	20,093 (28.8)	942 (37.4)
Social functioning (SF)	5195 (7.4)	432 (17.1)
Fatigue (FA)	14,213 (20.4)	794 (31.5)
Nausea and vomiting (NV)	9565 (13.7)	520 (20.6)
Pain (PA)	19,216 (27.5)	982 (39)
Dyspnea (DY)	19,894 (28.5)	858 (34)
Insomnia (SL)	12,245 (17.5)	605 (24)
Appetite loss (AP)	2091 (3)	124 (4.9)
Constipation (CO)	3414 (4.9)	156 (6.2)
Diarrhea (DI)	9652 (13.8)	559 (22.2)
Financial difficulties (FI)	7833 (11.2)	435 (17.3)

[a]EORTC QLQ-C30: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire Core 30.

[b]The global health status scale was not analyzed as there is no published threshold available.

[c]Events defined as scores below the thresholds for clinical importance indicate clinically relevant impairments according to Giesinger et al [33].

Model Development and Internal Validation

Figure 2A illustrates the performance of the logistic regression, extra-trees classifier, multilayer perceptron classifier, and a histogram-based gradient boosting classification tree for the prediction of impaired HRQoL at the next available assessment. Across all evaluated QLQ-C30 domains, the histogram-based gradient boosting classification tree showed the highest AUC, and all algorithms outperformed the LOCF baseline (Figure 2B). AUC values for the baseline regression models lay between 0.648 (95% CI 0.616 - 0.679; pain) and 0.829 (95% CI 0.767 - 0.884; AP), see Figure 2B for a comparison. A

XSL•FO
RenderX

comprehensive list of evaluation metrics can be found in Table S1 in Multimedia Appendix 1.

From all QLQ-C30 domains, functional scales had a higher AUC than single-item symptom scales (dyspnea, insomnia, AP, constipation, and diarrhea). Except for financial difficulties, which showed the highest AUC overall (0.861, SD 0.009).

Figure 2. Evaluation of models trained to predict impairments in HRQoL. Error bars: SD, for LOCF and external validation: 95% CI. Full data available in Tables S1 and S2 in Multimedia Appendix 1. HRQoL: health-related quality of life; LOCF: last observation carried forward; ROC AUC: area under the receiver operating characteristic curve.

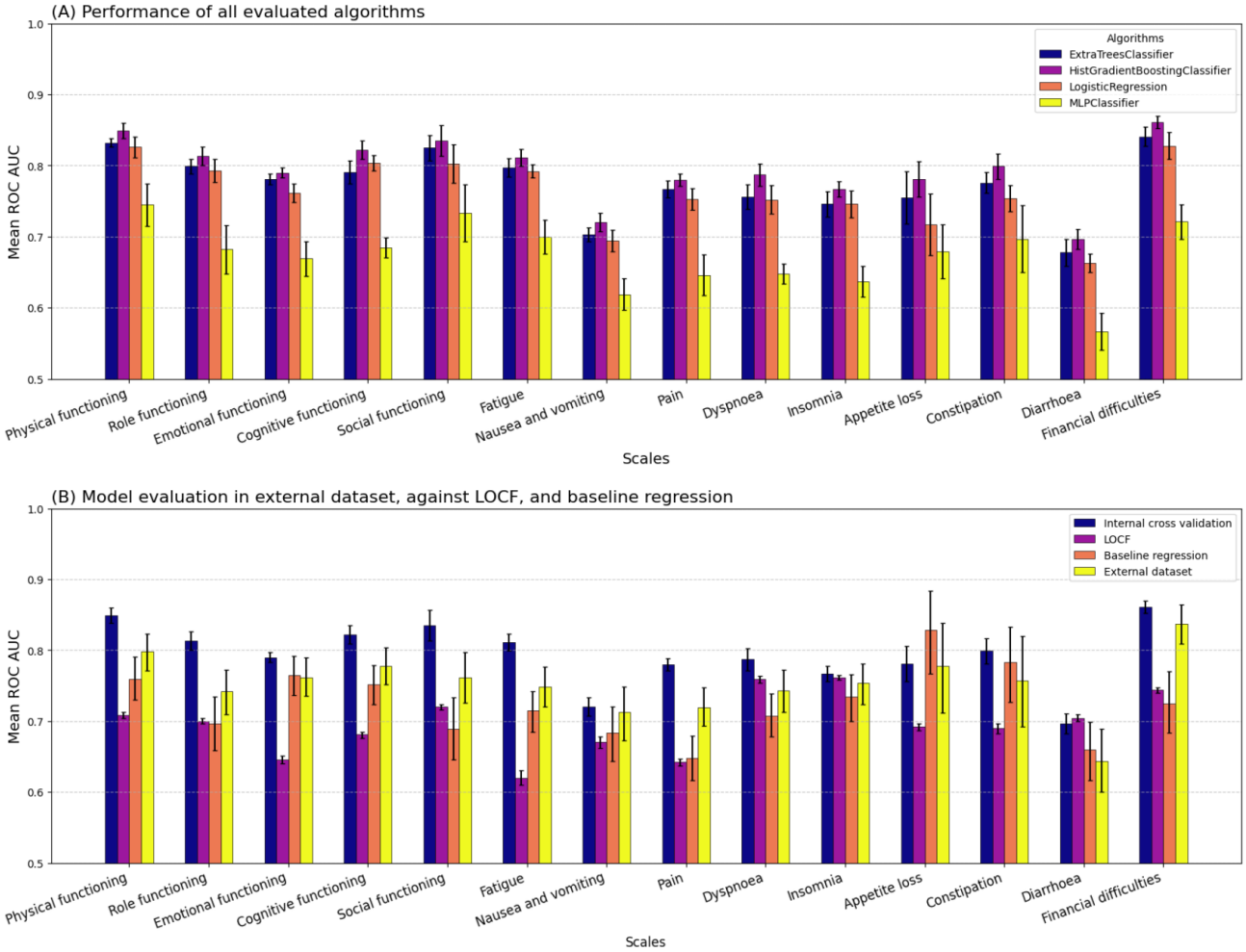

External Validation

The external validation with the NKI yielded good discriminatory ability (Figure 2B and Table S2 in Multimedia Appendix 1). The highest AUC (0.837, 95% CI 0.809, 0.864) could be observed in the financial difficulties scale, whereas the diarrhea scale yielded the lowest AUC (0.643, 95% CI 0.600, 0.689).

Model calibration varied between the scales (Figure S2) in Multimedia Appendix 1. We observed the worst calibration in the single-item symptom scales (AP, constipation, and diarrhea), better calibration in the functioning scales, and the best calibration in fatigue and physical functioning. Generally, models for scales containing more items were better calibrated. Calibration slopes were below the ideal value of 1 for all models (Table S2) in Multimedia Appendix 1, ranging from 0.089 (constipation; 95% CI 0.063, 0.116) to 0.286 (fatigue; 95% CI 0.259, 0.313), indicating overconfident predictions. Post hoc recalibration improved discrimination (AUCs 0.975 - 0.995) as well as calibration based on visual inspection (Table S5 and Figure S3) in Multimedia Appendix 1.

Across all domain models, the individual HRQoL items from the previous assessment, the time of assessment, and treatment start dates were among the most important features (Figure S4) in Multimedia Appendix 1. The type of treatment, other clinical variables, and diagnostic characteristics or comorbidities had little influence on the models' performance.

Decision curves are shown in Figure S5 in Multimedia Appendix 1, and most models (exceptions: social functioning, AP, and diarrhea) proved to be superior compared to a "treat all" scenario for risk thresholds between 0.2 and 0.6.

Model Fairness

The proportion of events in the risk groups was almost always higher than in the external validation set (Table S3) in Multimedia Appendix 1. Figure 3A illustrates model performance in risk groups selected from the external validation set in comparison with the whole set (see Table S3 in Multimedia Appendix 1 for risk group distributions).

XSL•FO
RenderX

Figure 3. Model performance in risk groups and time-dynamic performance compared to the external validation. For (B) after diagnosis: all assessments within the first year after diagnosis; future assessment: all assessments more than 1 year in the future. Full data available in Table S4 in Multimedia Appendix 1. ROC AUC: area under the receiver operating characteristic curve.

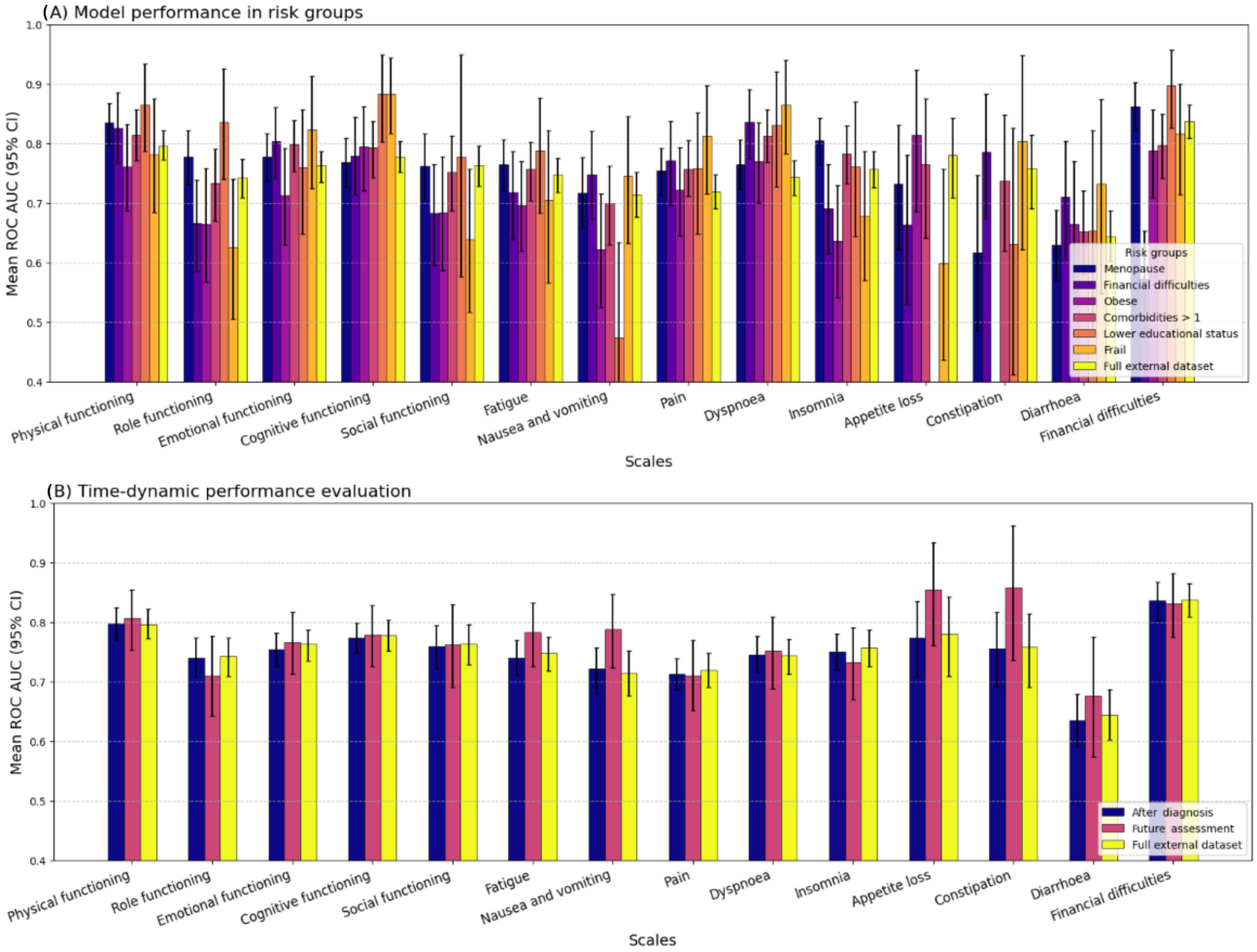

The discrimination power within risk groups varied between domains, depending on specific risk group-scale combinations (Table S4) in Multimedia Appendix 1. The lower education status risk group showed the highest (financial difficulties: 0.897, 95% CI 0.825, 0.956) as well as the lowest (nausea and vomiting: 0.473, 95% CI 0.314, 0.633) AUC over all.

In the postmenopausal and obese risk groups, AUC CIs consistently overlapped with those of the full dataset, indicating similar model performance. This was largely consistent for other risk groups, except for the following cases: multiple comorbidities (pain and dyspnea), financial difficulties, frailty (AP and role functioning), and lower educational status (nausea and vomiting, and constipation), where reduced performance was observed.

Additionally, the predictive performance differed within some scales more than in others. For instance, in the pain scale, the AUCs were relatively homogenous, whereas the constipation scale had a high variability in AUCs. Model calibration was poor with calibration slopes across all targets and risk groups below the ideal value of 1 (Table S4) in Multimedia Appendix 1. Due to the low number of events, we could not evaluate the calibration plots.

Time-Dynamic Performance

We observed little impact of the time-dynamic aspect on model performance (Figure 3B). The prediction within the first year after diagnosis showed similar AUCs compared to the whole external dataset, with differences below 0.01 (Table S4) in Multimedia Appendix 1. Long-term prediction showed slightly higher AUCs in 9 of 14 cases and the greatest deterioration in role functioning (0.032).

Discussion

Principal Findings

This study developed the first time-dynamic machine ML model for predicting impaired HRQoL in patients with early breast cancer, using multiple cohorts from different countries and settings. The ML-based prediction models consistently outperformed LOCF and the regression baseline in 10/14 cases, with the histogram-based gradient boosting classification tree showing the best performance. The best calibration was seen in multi-item scales; however, predictions were overconfident in all domains. We introduce a novel approach to time-dynamic HRQoL prediction, enabling flexible predictions beyond fixed time points. This addresses the issue with current static models, as they cannot capture the variability of real-world HRQoL assessment schedules. It therefore allows the early detection of

XSL•FO
RenderX

meaningful HRQoL impairments throughout a patient's course of disease and provides clinicians with a tool to guide targeted interventions. Model fairness analyses revealed variation in model performance across risk groups, depending on individual domain and group combinations.

Model Evaluation and External Validation

Overall, our models demonstrated good discriminatory performance across most HRQoL domains; however, calibration was inconsistent, particularly for scales with low outcome prevalence. Discrimination, reflected in AUC values exceeding 0.75 [52], was generally strong, confirming that the models effectively distinguished between patients with impaired HRQoL and those without. This is within the range of comparable studies, most of which used internal k-fold cross-validation (see Krepper et al [12] for an overview). Yet, for some outcomes (especially AP, constipation, and diarrhea) predicted probabilities did not consistently align with observed event rates (ie, poor calibration), limiting the clinical reliability of the models in those domains.

Accurate predictions, based on both strong discrimination and acceptable calibration, were observed for key domains such as physical, emotional, and cognitive functioning, as well as fatigue, pain, and financial difficulties. These domains might reflect more persistent problems, likely contributing to the robustness of the predictive models [53,54]. In contrast, models struggled to predict transient or low-prevalence symptoms, including diarrhea, nausea and vomiting, and constipation. These limitations reflect both the nature of the symptoms, often short-lived and influenced by acute factors [55], and the statistical challenges of modeling rare outcomes [39]. The poor calibration in these domains can be attributed to the class imbalance and the insufficient total number of events. In our external dataset, some of these outcomes (AP and constipation) occurred fewer than 200 times, a known threshold below which reliable calibration curves become difficult to estimate [43], while also corresponding to less than 10% of all observations. Here, ensemble methods such as random forests and gradient boosting tend to struggle, often underestimating probabilities at the extremes due to their inherent smoothing behavior [56]. This happens because averaging predictions from many noisy base models pulls extreme values toward the center, making it difficult to produce confident predictions near 0 or 1.

To address these limitations, we recalibrated the models using data from the specific clinical setting where they could be applied, thereby adjusting for local outcome prevalence and population characteristics. We highlighted the impact of this approach in a supplementary analysis, which improved model performance drastically.

Another potential approach is upsampling the training data to balance class distributions, which showed improved calibration when using HRQoL to predict clinical events [36]. However, this does not reflect real-world incidence rates and may distort probability calibration without improving model performance in cases with an adequate number of events. Notably, similar challenges have been reported in other HRQoL modeling efforts: for example, Adiprakoso et al [35] observed good discrimination but suboptimal calibration when predicting fatigue in patients with cancer, reinforcing the need for careful evaluation of both performance dimensions.

Despite these issues, this study adds to the growing body of literature emphasizing the value of external validation, a step still rarely undertaken in HRQoL modeling despite TRIPOD guidelines recommending distinct datasets for model evaluation [20].

In line with previous research [35,57-59], prior HRQoL was consistently among the most important predictors across all models, even over longer periods of time. Temporal variables, such as treatment start and stop dates and the timing of HRQoL assessment, also ranked highly. This was also reflected in the competitive performance of the baseline model, which included only prior HRQoL and the time between assessments. In contrast, neither specific treatment types nor clinical variables (eg, tumor stage) appeared influential for predicting the general functioning and symptom domains of the QLQ-C30. This suggests that, for broad HRQoL outcomes, being on or off treatment may matter more than which specific treatment is received. These patterns may differ for treatment-specific side effects not captured by the QLQ-C30 (eg, skin problems for new targeted agents), highlighting an avenue for future studies using more granular or treatment-targeted patient-reported outcome (PRO) measures [60].

Decision curve analysis indicated that most models provided greater net benefit than treat-all and treat-none strategies across threshold probabilities between 0.2 and 0.6, suggesting potential clinical usefulness within this range. As subsequent actions following risk identification are not clearly defined and may vary widely (see Di Meglio et al [57] for a list of suggested interventions for patients with high risk of fatigue), the models should not be interpreted as directly guiding specific interventions. Rather, given that available options are often low risk and primarily involve the use of supportive resources, the models' main value may lie in informing patient counseling and facilitating early, shared discussions about potential support needs.

Model Fairness and Time Dynamic Performance

Model fairness analyses revealed disparities in model performance, notably among patients with lower educational status or frailty. However, there was no single risk group with the constant worst discrimination in all domains. Our findings echo concerns about algorithmic bias in health care ML models, exacerbating existing inequities in socioeconomic status, disability, and other characteristics [61-63]. This was reflected in the increased prevalence of impaired HRQoL across the selected risk groups, which nevertheless could not always be predicted correctly. Addressing these risks requires the use of fairness-aware ML techniques, such as reweighting, adversarial debiasing, or subgroup-specific calibration, to ensure equitable performance across diverse patient populations [17]. Future work should explore these strategies to prevent the amplification of health care disparities.

Our implementation of time-dynamic prediction marks a promising methodological extension over traditional fixed-interval ML models for HRQoL. Enabled by a

XSL•FO
RenderX

heterogeneous dataset, this approach better reflects real-world clinical care, where follow-up intervals vary. Importantly, we observed consistent model performance across different time spans. For long-term prediction, that is, spanning more than 1 year between assessments, this might reflect a stabilization of HRQoL. At the same time, the models also performed well during the first year after diagnosis, despite this period being marked by significant treatment-related fluctuations in HRQoL. This indicates that the models are capable of capturing both stable and dynamic phases of a patient's trajectory, making them applicable in both contexts. In line with the ESMO (European Society for Medical Oncology) clinical practice guideline for PRO use in clinical practice, which recommends integrating PRO results into clinical care [64], such insights may inform supportive care planning by enabling the prioritization of patients who are likely to require timely interventions, while also helping patients develop more realistic expectations regarding their long-term HRQoL prognosis.

As our approach primarily served as a proof-of-concept, no direct comparisons with existing HRQoL models are currently available. Nevertheless, these encouraging results highlight the potential of time-dynamic modeling to address real-world clinical needs, warranting extension to other populations and settings. Future research is needed to compare this approach to static models.

Strengths and Limitations

This study presents several key strengths. To our knowledge, it is the first ML approach to HRQoL prediction that incorporates time-dynamic modeling for patients with breast cancer, allowing for predictions at variable future time points rather than fixed intervals. Second, deriving our outcomes directly from anchor-based thresholds (patients reporting limitations in daily life, need for help or care, or worries of the patient or their partner or family) [33] ensures that predictions are tied to clinically meaningful impairments. As such, the models can help identify which patients are more likely to develop problems in specific HRQoL domains at selected points in the future and who may benefit from further clinical attention. Considering that PRO data are increasingly collected in routine care and that the QLQ-C30 is among the most frequently used questionnaires in electronic systems [65], our study shows that these data could be used more comprehensively to inform clinical teams and patients. Additionally, the use of a multicohort dataset, combining clinical trial data with real-world observational data, improves the model's validity. Importantly, we used external validation using a separate cohort to evaluate generalizability and illustrated the effects of post hoc recalibration. We also explicitly assessed predictive fairness, addressing a critical and often overlooked aspect of ML in health care by examining performance across subgroups and discussing bias mitigation strategies.

Nevertheless, this study has limitations. Most patients in our dataset were Dutch, which may limit the generalizability of findings to other countries. However, reference values for early patients with breast cancer demonstrate only slight variations across Europe [66], and the models primarily rely on prior HRQoL rather than population-specific variables. This might suggest that their performance is not inherently tied to national context, and the models may be transferable to other similar countries, provided that health care systems, cultural backgrounds, and follow-up intensity are broadly comparable. Yet, model adoption into different settings must still be evaluated carefully. In addition, we are currently working to include further German and British data to expand our database and retrain our models with a more international population.

Second, even though we evaluated model performance in an RWD set and this might have reduced selection bias, there might still be an underrepresentation of certain minorities or risk groups in our data.

Furthermore, as we performed a secondary analysis on existing datasets not originally intended for this purpose, some potentially important predictors of HRQoL (eg, socioeconomic status or social support) were unavailable, which may have limited the performance of our models.

Additionally, we dichotomized the outcomes to facilitate clinical interpretation; this may have led to a loss of information and reduced predictive accuracy compared to modeling with the original continuous scale.

Further, the field of ML is evolving rapidly, and by the time results are published, newer architectures (eg, pretrained tabular foundation models [67]) may already offer improved performance. Exploring such approaches was beyond the scope of this study but represents an important direction for future research.

Finally, while our models demonstrate promise, they remain preliminary with respect to clinical implementation, especially for a few selected scales showing poor calibration for certain rare symptoms. This limitation may be attributable not only to low event rates but also to the application of a single time-dynamic modeling framework across heterogeneous phases of the patient journey, raising the possibility that more phase-specific approaches (eg, during active treatment vs posttreatment or survivorship) could improve calibration for certain symptoms.

This mirrors a systemic issue in translational implementation, where HRQoL prediction models often fail to progress beyond proof-of-concept due to a combination of technical, operational, and contextual barriers [68]. As outlined by Spencer et al [69], both general ML deployment issues and PRO-specific challenges contribute to what they refer to as the "leaky pipeline" from model development to bedside application. Moreover, compliance with regulatory frameworks, such as the European Union Artificial Intelligence Act, necessitates not only robust prospective validation in real-world settings but also demonstrable alignment with ethical, safety, and stakeholder-involvement standards.

Conclusions

Our findings support the future integration of ML into PRO-based clinical decision support systems, while also emphasizing the need to address fairness in predictive modeling. Importantly, the predicted impairments are clinically meaningful, as they identify patients at risk of future problems

XSL•FO
RenderX

derived from anchor-based thresholds and enable more timely, targeted interventions. The time-dynamic nature of the predictions further increases their clinical utility by allowing risk estimation to be aligned with clinically relevant time horizons rather than fixed follow-up intervals. Future work should focus on improving model fairness, expanding dataset diversity, and evaluating clinical implementation in real-world oncology settings.

Acknowledgments

We would like to acknowledge that the following members of the EORTC Quality of Life Group were involved in this work: Niclas J Hubel, Thijs G W van der Heijden, Belle H de Rooij, Kelly M de Ligt, Lonneke V van de Poll-Franse, Bernhard Holzner, Jens Lehmann.

We would like to thank Corneel Coens, PhD, for his help with identifying EORTC (European Organisation for Research and Treatment of Cancer) breast cancer trials; Pablo Reja, PhD, for all his advice surrounding data sharing agreementagreements and creating a draft data sharing agreement for us; Coralie Poncet, MSc, for preparing the EORTC AMAROS (After Mapping of the Axilla, Radiotherapy or Surgery?) data for this project; the NKI (Netherlands Cancer Institute) Scientific data administration, especially Melanie Singer-Van den Hout, MSc, for their data preparation of the NKI data; Henrike Bretveld, MSc, and Ghita Puts, MSc, from the Dutch cancer registry for clinical data preparation of UMBRELLA (Utrecht Cohort for Multiple Breast Cancer Intervention Studies and Long-Term Evaluation) and OPTIMUM (Towards Optimal Timing and Method for Promoting Sustained Adherence to Lifestyle and Body Weight Recommendations in Postmenopausal Breast Cancer Survivors); Nicole Horevoorts, MSc and the team of the PROFILES (Patient Reported Outcomes Following Initial Treatment and Long term Evaluation of Survivorship) registry for the health-related quality of life data preparation of UMBRELLA and OPTIMUM; the team research from the UMCU (University Medical Center Utrecht) division of imaging for providing the key to couple NKR (Netherlands Cancer Registry) and PROFILES data for UMBRELLA; August Zabernigg, MD from BKH Kufstein (District Hospital Kufstein) for providing the BKH Kufstein data; the NKI knowledge transfer and contracting department for all the help with the data-sharing agreements; Lonneke van de Poll-Franse, PhD, Jaap C Seidell, PhD, Floortje Mols, MD, Sandra van Cappellen-van Maldegem, MSc, Janneke van den Boom, BA, and Judith van Valenberg, MSc, for the execution of the OPTIMUM study.

Disclosure of delegation to generative artificial intelligence (GenAI): The authors declare the use of generative AI in the research and writing process. According to the GAIDeT (Generative Artificial Intelligence Delegation Taxonomy; 2025), the following tasks were delegated to GenAI tools under full human supervision: (1) code optimization, (2) proofreading and editing, and (3) summarizing text. The GenAI tool used was ChatGPT-5 (OpenAI). Responsibility for this final paper lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the outcomes. Declaration submitted by collective responsibility.

Funding

This research was funded by the European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Group (007-2022; EORTC 2052).This research was funded by the EORTC Quality of Life Group (007-2022; EORTC 2052). The funder had no involvement in this study's design, data collection, analysis, interpretation, or the writing of this paper.

Data Availability

The data supporting this study's findings are available from each cohort study's principal investigator. Restrictions apply to the availability of these data, which were used under license for this study. The code of our analyses can be accessed through GitHub [70].

Authors' Contributions

Conceptualization: NJH, EORTC Quality of Life Group
Data curation: NJH, TGWvdH, EORTC Quality of Life Group
Formal analysis: NJH, BM, EORTC Quality of Life Group
Investigation: NJH, TGWvdH, EORTC Quality of Life Group
Methodology: NJH, BM, EORTC Quality of Life Group
Project administration: NJH, TGWvdH, EORTC Quality of Life Group
Resources: LVvdP-F, BH, EORTC Quality of Life Group
Supervision: LVvdP-F, BH, JL, EORTC Quality of Life Group
Writing – original draft: NJH, BM, JL, EORTC Quality of Life Group
Writing – review & editing: NJH, TGWvdH, BM, BHdR, KMdL, HMV, MH, VA, LVvdP-F, BH, JL, EORTC Quality of Life Group
Data provision: KMdL, HMV, MH, VA, EORTC Quality of Life Group

XSL•FO
RenderX

Conflicts of Interest

NJH reports employment by ACMIT GmbH - Austrian Center for Medical Innovation and Technology, Wiener Neustadt, Austria, outside the submitted work.
HMV reports research funding by Philips, Elekta, European Commission, Dutch Research Council, and Dutch Cancer Foundation outside of the submitted work.
BH holds shares in the company Evaluation Software Development GmbH, Innsbruck, Austria. JL reports consultancy for Evaluation Software Development and a research grant from Takeda, both outside of the submitted work.

Multimedia Appendix 1
Tables, figures, and more on the data in this paper.
[PDF File, 7678 KB - jmir_v28i1e81424_app1.pdf]

References

1. Harbeck N, Gnant M. Breast cancer. Lancet 2017 Mar 18;389(10074):1134-1150. [doi: 10.1016/S0140-6736(16)31891-8] [Medline: 27865536]
2. Anderson-Hanley C, Sherman ML, Riggs R, Agocha VB, Compas BE. Neuropsychological effects of treatments for adults with cancer: a meta-analysis and review of the literature. J Int Neuropsychol Soc 2003 Nov;9(7):967-982. [doi: 10.1017/S1355617703970019] [Medline: 14738279]
3. Mokhtari-Hessari P, Montazeri A. Health-related quality of life in breast cancer patients: review of reviews from 2008 to 2018. Health Qual Life Outcomes 2020 Dec;18(1):1-25. [doi: 10.1186/s12955-020-01591-x]
4. von Heymann-Horan AB, Dalton SO, Dziekanska A, et al. Unmet needs of women with breast cancer during and after primary treatment: a prospective study in Denmark. Acta Oncol 2013 Feb;52(2):382-390. [doi: 10.3109/0284186X.2012.746468]
5. Alfano CM, Mayer DK, Bhatia S, et al. Implementing personalized pathways for cancer follow‐up care in the United States: proceedings from an American Cancer Society–American Society of Clinical Oncology summit. CA A Cancer J Clinicians 2019 May;69(3):234-247. [doi: 10.3322/caac.21558]
6. Karri R, Chen YPP, Drummond KJ. Using machine learning to predict health-related quality of life outcomes in patients with low grade glioma, meningioma, and acoustic neuroma. PLoS ONE 2022;17(5):e0267931. [doi: 10.1371/journal.pone.0267931] [Medline: 35507629]
7. van der Heijden TGW, de Ligt KM, Hubel NJ, et al. Exploring the role of health-related quality of life measures in predictive modelling for oncology: a systematic review. Qual Life Res 2025 Feb;34(2):305-323. [doi: 10.1007/s11136-024-03820-y] [Medline: 39652111]
8. Mertz S, Benjamin C, Girvalaki C, et al. Progression-free survival and quality of life in metastatic breast cancer: the patient perspective. Breast 2022 Oct;65:84-90. [doi: 10.1016/j.breast.2022.07.006] [Medline: 35870420]
9. Yeo HY, Liew AC, Chan SJ, Anwar M, Han CHW, Marra CA. Understanding patient preferences regarding the important determinants of breast cancer treatment: a narrative scoping review. Patient Prefer Adherence 2023;17:2679-2706. [doi: 10.2147/PPA.S432821] [Medline: 37927344]
10. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ 2020 Mar 18;368:m441. [doi: 10.1136/bmj.m441]
11. van der Heijden TGW, Hubel NJ, de Ligt KM, et al. Opportunities and challenges in pooling health-related quality-of-life data for prediction modeling in breast cancer across Europe: lessons from the EORTC BALANCE project. ESMO Real World Data Digit Oncol 2025 Sep;9:100172. [doi: 10.1016/j.esmorw.2025.100172] [Medline: 41646212]
12. Krepper D, Cesari M, Hubel NJ, Zelger P, Sztankay MJ. Machine learning models including patient-reported outcome data in oncology: a systematic literature review and analysis of their reporting quality. J Patient Rep Outcomes 2024 Nov 5;8(1):126. [doi: 10.1186/s41687-024-00808-7] [Medline: 39499409]
13. Pfob A, Mehrara BJ, Nelson JA, Wilkins EG, Pusic AL, Sidey-Gibbons C. Machine learning to predict individual patient-reported outcomes at 2-year follow-up for women undergoing cancer-related mastectomy and breast reconstruction (INSPiRED-001). Breast 2021 Dec;60:111-122. [doi: 10.1016/j.breast.2021.09.009] [Medline: 34619573]
14. Shi HY, Tsai JT, Chen YM, Culbertson R, Chang HT, Hou MF. Predicting two-year quality of life after breast cancer surgery using artificial neural network and linear regression models. Breast Cancer Res Treat 2012 Aug;135(1):221-229. [doi: 10.1007/s10549-012-2174-6]
15. Sharifi F, Mohammed E, Crump T, Far BH. Explainable analytics to predict the quality of life in patients with prostate cancer from longitudinal data. Appl Artif Intell 2022 Dec 31;36(1):2055393. [doi: 10.1080/08839514.2022.2055393]
16. Castelnovo A, Crupi R, Greco G, Regoli D, Penco IG, Cosentini AC. A clarification of the nuances in the fairness metrics landscape. Sci Rep 2022 Mar 10;12(1):4209. [doi: 10.1038/s41598-022-07939-1]
17. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv 2022 Jul 31;54(6):1-35. [doi: 10.1145/3457607]

XSL•FO
RenderX

18. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018 Dec 18;169(12):866-872. [doi: 10.7326/M18-1990]
19. Chen MS, Lara PN, Dang JHT, Paterniti DA, Kelly K. Twenty years post-NIH Revitalization Act: Enhancing Minority Participation in Clinical Trials (EMPaCT): laying the groundwork for improving minority clinical trial accrual: renewing the case for enhancing minority participation in cancer clinical trials. Cancer 2014 Apr 1;120 Suppl 7(7):1091-1096. [doi: 10.1002/cncr.28575] [Medline: 24643646]
20. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024 Apr 16;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
21. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016 Dec 16;18(12):e323. [doi: 10.2196/jmir.5870] [Medline: 27986644]
22. Straver ME, Meijnen P, van Tienhoven G, et al. Sentinel node identification rate and nodal involvement in the EORTC 10981-22023 AMAROS trial. Ann Surg Oncol 2010 Jul;17(7):1854-1861. [doi: 10.1245/s10434-010-0945-z]
23. Boomstra E, Walraven I, van der Ploeg IMC, et al. Moving beyond barriers: a mixed-method study to develop evidence-based strategies to improve implementation of PROMs in clinical oncology care. Qual Life Res 2025 Jan;34(1):173-188. [doi: 10.1007/s11136-024-03787-w] [Medline: 39302555]
24. Wintner LM, Giesinger JM, Sztankay M, Bottomley A, Holzner B, EORTC Quality of Life Group. Evaluating the use of the EORTC patient-reported outcome measures for improving inter-rater reliability of CTCAE ratings in a mixed population of cancer patients: study protocol for a randomized controlled trial. Trials 2020 Oct 13;21(1):849. [doi: 10.1186/s13063-020-04745-w] [Medline: 33050917]
25. Young-Afat DA, van Gils CH, van den Bongard HJGD, Verkooijen HM, on behalf of the UMBRELLA Study Group. The Utrecht Cohort for Multiple Breast Cancer Intervention Studies and Long-Term Evaluation (UMBRELLA): objectives, design, and baseline results. Breast Cancer Res Treat 2017 Jul;164(2):445-450. [doi: 10.1007/s10549-017-4242-4]
26. van Cappellen-van Maldegem SJM, Mols F, Horevoorts N, et al. Towards Optimal Timing and Method for Promoting Sustained Adherence to Lifestyle and Body Weight Recommendations in Postmenopausal Breast Cancer Survivors (the OPTIMUM-study): protocol for a longitudinal mixed-method study. BMC WOMENS Health 2021 Jul 6;21(1):268. [doi: 10.1186/s12905-021-01406-1] [Medline: 34229690]
27. Arndt V, Stürmer T, Stegmaier C, Ziegler H, Dhom G, Brenner H. Socio-demographic factors, health behavior and late-stage diagnosis of breast cancer in Germany: a population-based study. J Clin Epidemiol 2001 Jul;54(7):719-727. [doi: 10.1016/s0895-4356(00)00351-6] [Medline: 11438413]
28. R Foundation for Statistical Computing. URL: https://www.R-project.org/ [accessed 2026-04-14]
29. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. JNCI Journal of the National Cancer Institute 1993 Mar 3;85(5):365-376. [doi: 10.1093/jnci/85.5.365]
30. Sprangers MA, Groenvold M, Arraras JI, et al. The European Organization for Research and Treatment of Cancer breast cancer-specific quality-of-life questionnaire module: first results from a three-country field study. J Clin Oncol 1996 Oct;14(10):2756-2768. [doi: 10.1200/JCO.1996.14.10.2756] [Medline: 8874337]
31. IterativeImputer. scikit-learn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html [accessed 2026-04-14]
32. Fayers PM, Aaronson NK, Bjordal K, et al. The EORTC QLQ-C30 scoring manual (3rd edition). European Organisation for Research and Treatment of Cancer, Brussels. 2001 URL: https://www.eortc.org/app/uploads/sites/2/2018/02/SCmanual.pdf [accessed 2026-04-18]
33. Giesinger JM, Loth FLC, Aaronson NK, et al. Thresholds for clinical importance were established to improve interpretation of the EORTC QLQ-C30 in clinical practice and research. J Clin Epidemiol 2020 Feb;118:1-8. [doi: 10.1016/j.jclinepi.2019.10.003]
34. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12(85):2825-2830 [FREE Full text]
35. Adiprakoso D, Katsimpokis D, Oerlemans S, et al. Development of a prediction model for clinically-relevant fatigue: a multi-cancer approach. Qual Life Res 2025 Jan;34(1):231-245. [doi: 10.1007/s11136-024-03807-9] [Medline: 39516438]
36. Wójcik Z, Dimitrova V, Warrington L, Velikova G, Absolom K. Using machine learning to predict unplanned hospital utilization and chemotherapy management from patient-reported outcome measures. JCO Clin Cancer Inform 2024 Apr;8:e2300264. [doi: 10.1200/CCI.23.00264] [Medline: 38669610]
37. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit DAGM 1997 Jul;30(7):1145-1159. [doi: 10.1016/S0031-3203(96)00142-2]
38. Megahed FM, Chen YJ, Megahed A, Ong Y, Altman N, Krzywinski M. The class imbalance problem. Nat Methods 2021 Nov;18(11):1270-1272. [doi: 10.1038/s41592-021-01302-4]
39. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating: Springer Cham; 2019. [doi: 10.1007/978-3-030-16399-0]

XSL•FO
RenderX

40. Karahalios A, Baglietto L, Carlin JB, English DR, Simpson JA. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. BMC Med Res Methodol 2012 Jul 11;12(96):96. [doi: 10.1186/1471-2288-12-96] [Medline: 22784200]
41. Leyrat C, Carpenter JR, Bailly S, Williamson EJ. Common methods for handling missing data in marginal structural models: what works and why. Am J Epidemiol 2021 Apr 6;190(4):663-672. [doi: 10.1093/aje/kwaa225]
42. Steyerberg EW, Harrell Jr FE. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016 Jan;69:245-247. [doi: 10.1016/j.jclinepi.2015.04.005] [Medline: 25981519]
43. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med 2019 Dec 16;17(1):230. [doi: 10.1186/s12916-019-1466-7] [Medline: 31842878]
44. Błasiok J, Nakkiran P. Smooth ECE: principled reliability diagrams via kernel smoothing. arXiv. Preprint posted online on Sep 21, 2023. [doi: 10.48550/ARXIV.2309.12236]
45. Brier GW. Verification of forecasts expressed in terms of probability. Mon Wea Rev 1950 Jan;78(1):1-3. [doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2]
46. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26(6):565-574. [doi: 10.1177/0272989X06295361] [Medline: 17099194]
47. Van Calster B, Wynants L, Verbeek JFM, et al. Reporting and interpreting decision curve analysis: a guide for investigators. Eur Urol 2018 Dec;74(6):796-804. [doi: 10.1016/j.eururo.2018.08.038] [Medline: 30241973]
48. Tack L, Schofield P, Boterberg T, Parris CN, Debruyne PR. Editorial: clinical cancer research in vulnerable populations. Front Oncol 2023;13:1166714. [doi: 10.3389/fonc.2023.1166714] [Medline: 36937380]
49. Murugappan MN, King-Kallimanis BL, Bhatnagar V, et al. Patient-Reported Frailty Phenotype (PRFP) vs. International Myeloma Working Group frailty index (IMWG FI) proxy: a comparison between two approaches to measuring frailty. J Geriatr Oncol 2024 Mar;15(2):101681. [doi: 10.1016/j.jgo.2023.101681]
50. Gao Y, Rosas JC, Fink H, Behrens S, Chang-Claude J, Seibold P. Longitudinal changes of health-related quality of life over 10 years in breast cancer patients treated with radiotherapy following breast-conserving surgery. Qual Life Res 2023 Sep;32(9):2639-2652. [doi: 10.1007/s11136-023-03408-y] [Medline: 37093543]
51. Hsu T, Ennis M, Hood N, Graham M, Goodwin PJ. Quality of life in long-term breast cancer survivors. JCO 2013 Oct 1;31(28):3540-3548. [doi: 10.1200/JCO.2012.48.1903]
52. de Hond AAH, Steyerberg EW, van Calster B. Interpreting area under the receiver operating characteristic curve. Lancet Digit Health 2022 Dec;4(12):e853-e855. [doi: 10.1016/S2589-7500(22)00188-1] [Medline: 36270955]
53. Bower JE. Behavioral symptoms in patients with breast cancer and survivors. J Clin Oncol 2008 Feb 10;26(5):768-777. [doi: 10.1200/JCO.2007.14.3248] [Medline: 18258985]
54. Maass SWMC, Brandenbarg D, Boerman LM, Verhaak PFM, de Bock GH, Berendsen AJ. Fatigue among long-term breast cancer survivors: a controlled cross-sectional study. Cancers (Basel) 2021 Mar 15;13(6):1301. [doi: 10.3390/cancers13061301] [Medline: 33803966]
55. Greenlee H, DuPont-Reyes MJ, Balneaves LG, et al. Clinical practice guidelines on the evidence-based use of integrative therapies during and after breast cancer treatment. CA Cancer J Clin 2017 May 6;67(3):194-232. [doi: 10.3322/caac.21397] [Medline: 28436999]
56. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. Assoc Comput Mach 2005:625-632. [doi: 10.1145/1102351.1102430]
57. Di Meglio A, Havas J, Soldato D, et al. Development and validation of a predictive model of severe fatigue after breast cancer diagnosis: toward a personalized framework in survivorship care. JCO 2022 Apr 1;40(10):1111-1123. [doi: 10.1200/JCO.21.01252]
58. Révész D, van Kuijk SMJ, Mols F, et al. Development and internal validation of prediction models for colorectal cancer survivors to estimate the 1-year risk of low health-related quality of life in multiple domains. BMC Med Inform Decis Mak 2020 Mar 12;20(1):54. [doi: 10.1186/s12911-020-1064-9] [Medline: 32164641]
59. Xu C, Pfob A, Mehrara BJ, et al. Enhanced surgical decision-making tools in breast cancer: predicting 2-year postoperative physical, sexual, and psychosocial well-being following mastectomy and breast reconstruction (INSPiRED 004). Ann Surg Oncol 2023 Nov;30(12):7046-7059. [doi: 10.1245/s10434-023-13971-w] [Medline: 37516723]
60. Bjelic-Radisic V, Cardoso F, Cameron D, et al. An international update of the EORTC questionnaire for assessing quality of life in breast cancer patients: EORTC QLQ-BR45. Ann Oncol 2020 Feb;31(2):283-288. [doi: 10.1016/j.annonc.2019.10.027] [Medline: 31959345]
61. Mittermaier M, Raza MM, Kvedar JC. Bias in AI-based models for medical applications: challenges and mitigation strategies. NPJ Digit Med 2023 Jun 14;6(1):113. [doi: 10.1038/s41746-023-00858-z] [Medline: 37311802]
62. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: implications for clinical decision-making. PLOS Digit Health 2024 Nov;3(11):e0000651. [doi: 10.1371/journal.pdig.0000651] [Medline: 39509461]
63. Whitehead M, Carrol E, Kee F, Holmes C. Making the invisible visible: what can we do about biased AI in medical devices? BMJ 2023 Aug 16;382:1893. [doi: 10.1136/bmj.p1893]

XSL•FO
RenderX

64. Di Maio M, Basch E, Denis F, et al. The role of patient-reported outcome measures in the continuum of cancer clinical care: ESMO clinical practice guideline. Ann Oncol 2022 Sep;33(9):878-892. [doi: 10.1016/j.annonc.2022.04.007] [Medline: 35462007]
65. Hubel NJ, Vorbach SM, de Ligt KM, et al. Sustainability and time trends in electronic patient-reported outcome assessment in routine cancer care: systematic scoping review and follow-up survey. J Med Internet Res 2025 Apr 25;27(1):e69398. [doi: 10.2196/69398] [Medline: 40280556]
66. Mierzynska J, Taye M, Pe M, et al. Reference values for the EORTC QLQ-C30 in early and metastatic breast cancer. Eur J Cancer 2020 Jan;125:69-82. [doi: 10.1016/j.ejca.2019.10.031] [Medline: 31838407]
67. Hollmann N, Müller S, Purucker L, et al. Accurate predictions on small data with a tabular foundation model. Nat New Biol 2025 Jan 9;637(8045):319-326. [doi: 10.1038/s41586-024-08328-6]
68. Lee TC, Shah NU, Haack A, Baxter SL. Clinical implementation of predictive models embedded within electronic health record systems: a systematic review. Informatics (MDPI) 2020 Sep;7(3):25. [doi: 10.3390/informatics7030025]
69. Spencer KL, Absolom KL, Allsop MJ, et al. Fixing the leaky pipe: how to improve the uptake of patient-reported outcomes-based prognostic and predictive models in cancer clinical practice. JCO Clin Cancer Inform 2023 Sep;7(7):e2300070. [doi: 10.1200/CCI.23.00070] [Medline: 37976441]
70. BALANCE-ML. GitHub. URL: https://github.com/NJHubel/BALANCE-ML/blob/main/README.md [accessed 2026-04-14]

Abbreviations

AMAROS : After Mapping of the Axilla, Radiotherapy or Surgery?
AP: appetite loss
AUC: area under the receiver operating characteristic curve
BALANCE: big data in patients with breast cancer
ECE: expected calibration error
EORTC: European Organisation for Research and Treatment of Cancer
ER: estrogen receptor
ESMO: European Society for Medical Oncology
HRQoL: health-related quality of life
LOCF: last observation carried forward
ML: machine learning
NKI: Netherlands Cancer Institute
OPTIMUM: Towards Optimal Timing and Method for Promoting Sustained Adherence to Lifestyle and Body Weight Recommendations in Postmenopausal Breast Cancer Survivors
PRO: patient-reported outcome
QLQ: Quality of Life Core Questionnaire
RWD: real world data
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
UMBRELLA: Utrecht Cohort for Multiple Breast Cancer Intervention Studies and Long-Term Evaluation
VERDI: Verlauf der Diagnostischen Abklärung

Edited by A Schwartz, M Balcarras, N Cahill; submitted 18.Aug.2025; peer-reviewed by AW Bello, A Shivanna, L zhu, M Chakit; revised version received 04.Feb.2026; accepted 04.Feb.2026; published 30.Apr.2026.

Please cite as:
Hubel NJ, van der Heijden TGW, Murauer B, de Rooij BH, de Ligt KM, Verkooijen HM, Gernaat SAM, Hoedjes M, Arndt V, van de Poll-Franse LV, Holzner B, Lehmann J
Time-Dynamic AI Models to Predict Quality of Life in Patients With Breast Cancer: Development and Validation Study Using the EORTC BALANCE Cohort
J Med Internet Res 2026;28:e81424
URL: https://www.jmir.org/2026/1/e81424
doi:10.2196/81424

© Niclas J Hubel, Thijs G W van der Heijden, Benjamin Murauer, Belle H de Rooij, Kelly M de Ligt, Helena M Verkooijen, Sofie AM Gernaat, Meeke Hoedjes, Volker Arndt, Lonneke V van de Poll-Franse, Bernhard Holzner, Jens Lehmann. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of

XSL•FO
RenderX

Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Performance of AI Tools in Citing Retracted Literature : Content Analysis

Sebastian Labenbacher*, MD; Maximilian Niederer*, MD; Sascha Hammer, MBA, MD; Matthias Bader, MD; Nikolaus Schreiber, MD, PHD; Helmar Bornemann-Cimenti, MSc, MBA, MD

Department of Anesthesiology and Intensive Care Medicine, Medical University of Graz, Auenbruggerplatz 5, Graz, Austria

*these authors contributed equally

Corresponding Author:
Sascha Hammer, MBA, MD
Department of Anesthesiology and Intensive Care Medicine, Medical University of Graz, Auenbruggerplatz 5, Graz, Austria

Abstract

Background: Generative artificial intelligence (GenAI) tools are increasingly used in scientific research to support literature searches, evidence synthesis, and manuscript preparation. While these systems promise substantial efficiency gains, concerns have emerged regarding their reliability, particularly their tendency to cite inaccurate, fabricated, or retracted literature. The unrecognized inclusion of retracted studies poses a serious risk to research integrity and evidence-based decision-making. Whether commonly used GenAI tools can reliably detect, exclude, or transparently communicate the retraction status of scientific publications remains unclear.

Objective: This study aimed to evaluate the ability of freely available GenAI tools to correctly handle retracted scientific articles during literature searches. Primary and secondary outcomes focused on accuracy, reliability, and consistency in recognizing retracted literature.

Methods: In this pragmatic trial, nine widely used free-access GenAI tools (ChatGPT 4, ChatGPT 5, Claude, Gemini, Perplexity, Microsoft Copilot, SciSpace, ScienceOS, and Consensus) were evaluated. Each tool was asked five predefined, standardized questions addressing topic overview, article identification, article summarization, and explicit assessment of retraction status. Overall, 15 retracted articles (the 10 most cited and 5 most recently retracted as of May 23, 2025) were selected from the Retraction Watch database. All questions were repeated twice to assess intratool consistency. Responses were independently rated as correct or incorrect by 2 researchers. Descriptive statistics summarized performance, and comparisons between general-purpose and research-focused AI tools were conducted using descriptive statistics. Interreviewer agreement was assessed using Cohen kappa coefficient.

Results: None of the evaluated AI tools consistently handled retracted articles correctly. No model achieved perfect accuracy across all question sets. ChatGPT 5 performed best, defined by the primary outcome of achieving fully correct responses to all five predefined tasks (5/5) for the highest number of retracted articles, correctly answering all five questions for 8 of 15 articles (53.3%). Research-focused tools (SciSpace, ScienceOS, and Consensus) failed to produce a single fully correct response set. Retracted articles were frequently included in topic overviews without warning, with error rates exceeding 40% in several tools. When specifically asked about retraction status, most systems failed to provide correct or complete information. OpenEvidence only reported data for a subset of our retracted articles as it is only used in health care literature. It demonstrated strong performance in topic overviews but low accuracy in identifying retracted articles.

Conclusions: Freely available GenAI tools are currently not able to detect, exclude, or appropriately flag retracted scientific literature. The widespread and confident reproduction of retracted studies represents a substantial threat to research integrity, particularly in medical and evidence-based fields. Until retraction-aware verification mechanisms are systematically integrated, independent source checking remains essential when using AI-assisted literature tools.

(J Med Internet Res 2026;28:e88766) doi:10.2196/88766

KEYWORDS

artificial intelligence; AI; retraction of publication; retractions; scientific misconduct; evidence-based Practice; data accuracy; ethics

XSL•FO
RenderX

Introduction

Artificial intelligence (AI) has gained substantial importance in scientific research in recent years, with growing use in text generation and automated knowledge retrieval [1,2]. Especially in health sciences, AI-based applications are being integrated into research workflows, and early evaluations suggest that these technologies can accelerate literature reviews, facilitate the synthesis of large datasets, and assist researchers and students in scientific communication [3,4]. An example of a generative artificial intelligence (GenAI) system developed for medical information gathering is OpenEvidence, which can exclusively be used by health care professionals. In particular, the ability of AI tools to efficiently process large volumes of information has been highlighted as a promising complement to established methods of evidence appraisal and synthesis [5,6].

At the same time, important limitations and risks of these technologies have become increasingly evident [7,8]. AI-assisted text generation may produce erroneous or fabricated references, thereby raising concerns about the integrity of GenAI-supported scientific work [9]. Beyond the creation of non-existent citations, the inadvertent inclusion of retracted publications represents a critical issue that can seriously undermine the reliability of evidence syntheses [10,11]. Recent reports have demonstrated that large language models (LLMs) frequently cite retracted or unreliable studies without acknowledgment, often presenting them with high linguistic confidence and apparent authority [12]. This behavior poses a direct threat to research integrity and may ultimately affect patient safety, particularly in evidence-based disciplines such as medicine, where retracted data can distort clinical conclusions.

Despite the rapid proliferation of AI tools for academic use, it remains unclear whether these systems can reliably detect, flag, or communicate the retraction status of scientific articles. Earlier investigations have focused primarily on single models or subject areas, leaving the broader cross-platform performance of GenAI largely unexplored [13].

Even AI tools marketed for scientific and health care research—such as those designed to assist literature searches or summarize findings—have not been systematically tested for their ability to recognize retracted work. Against this background, this study evaluated nine widely used GenAI systems, including ChatGPT 5, ChatGPT 4, Claude, Gemini, Perplexity, Microsoft Copilot, SciSpace, ScienceOS, and Consensus, with OpenEvidence analyzed separately as a medical literature–specific model. Using a predefined set of retracted clinical trials, each tool was assessed for factual accuracy, reliability, and risk of propagating retracted studies. The aim was to critically assess the reliability of AI-assisted literature searches and, based on these findings, to provide recommendations for their responsible and safe application in scientific practice.

Methods

Study Design

This study was designed as a pragmatic evaluation using freely accessible artificial intelligence tools (AI) for literature screening and summarization of articles with five predefined questions to assess how retracted articles are handled. The trial was registered during the data acquisition phase in the Open Science Framework (osf.io) [14].

The predefined questions were designed to outline a pathway for investigating a new topic of interest: providing an overview based on the article's keywords, identifying the most important research articles related to the defined topic, and summarizing the retracted article. The final two questions were directly aimed at the retraction status of the respective article. Each response was scored as correct (1) or incorrect (0).

We used AI tools specifically designed for scientific work (Consensus Version July 2025, SciSpace Version 1.5.1, ScienceOS Version July 2025, OpenEvidence Version July 2025) as well as generic AI tools (ChatGPT Version 4 and 5, Microsoft Copilot Version 3.0, Gemini Version 2.5, Perplexity AI Version July 2025, Claude AI Version 3.5). AI tools with trial versions insufficient to perform all necessary research questions were excluded.

Ethical Considerations

This study did not involve human participants, patient data, or identifiable personal information and was limited to the analysis of publicly available literature and AI-generated outputs. In accordance with institutional and international research ethics guidelines, ethics committee approval was not required.

Questions

The following questions were used to guide the literature assessment:

1. Give me an overview on publications of "keywords of the article"
2. Give relevant articles regarding the topic of "keywords of the article"
3. Summarise the article "TITLE" published by "AUTHORS" for me
4. Is the article "TITLE" published by "AUTHORS" retracted?
5. Why is "TITLE" published by "AUTHORS" retracted?

The exact titles (excluding RETRACTED or any other details on the retraction status), authors, and keywords we used can be viewed in the supplement (Table S1 in Multimedia Appendix 1, [15-29]).

Scoring and Rating of Responses

A binary scoring system (one point per question) was used to emphasize functional correctness in AI-assisted literature workflows, acknowledging that this approach sacrifices linguistic nuance. Responses using equivocal or nonspecific wording (eg, "controversial" or "heavily criticized" instead of explicitly stating retraction) were rated as incorrect, as such phrasing does not clearly state retraction status. This conservative approach was chosen to avoid overestimating AI

XSL•FO
RenderX

performance where precise identification of retracted literature is critical.

Retracted Articles

Overview

To ensure objectivity, we used the ten most cited, retracted articles found in the Retraction Watch website [30]. Furthermore, we used 5 articles, retracted closest to May 23, 2025 (extracted from the retraction-watch database [31]) to compare whether the number of citations or the date of retraction has an impact on our results [15-29]. Details of the included articles can be found in Table S1 in Multimedia Appendix 1 [15-29].

In the process of choosing the papers for our search, we did not find keywords for every article. Keywords were essential to our methods, and we therefore developed our own keywords for the articles. To ensure a uniform keyword generation process, we decided to let ChatGPT 4 choose the respective keywords for the articles. As ChatGPT is also a tool in evaluation, we performed an exploratory analysis of keyword generation restricted to general-purpose GenAI tools (ChatGPT 4, Claude, Gemini, and Perplexity), as these models were used for keyword generation in our study design. Research-focused AI tools were not included in this analysis.

Study Aims

Our primary aim was to assess how often AI tools correctly handled retracted articles. Correct handling was defined as providing accurate responses to all 5 predefined questions, including appropriate identification and summarization of the article and correct recognition and explanation of its retraction status.

Secondary Aim

The secondary aims of this study were as follows:

1. How often were retracted articles included in a topic overview (topic overviews include questions 1 and 2)
2. How often were the retracted articles included in the search pathway (search pathway includes questions 1, 2, and 3)
3. To analyze how often results differ when the same question is asked twice
4. To compare general AI tools (like ChatGPT) against AI tools specifically designed for research (like ScienceOS or Consensus)

Exploratory Aims

The here-mentioned exploratory aims were not predefined but later on added to further improve and clarify the aforementioned aims.

Incidence of GenAI-hallucinations

To explore the incidence of GenAI-hallucinations, we screened question 1 and question 2 of each AI tool. Hallucinations were defined as the generation of non-existent or unverifiable publications; interpretive inaccuracies regarding article content or retraction reasons were not included in this definition. This screening was limited to checking the existence of each link and, when possible, comparing the trial details of each article described to the trial details of the corresponding link. Cases in which we found errors (like incorrect first author, incorrect journal,..) were recorded but not judged as hallucination since the respective article exists.

Nonretracted Control Group

We retrospectively added a control group of nonretracted articles. This was deemed necessary to give an estimate on the ability of AI tools to find articles when provided with a list of keywords. Keywords were generated in the same fashion as described in the methods above; we only asked question 1 and question 2.

Analysis

Data Extraction

Data were gathered in a predefined Microsoft Excel file with full-text answers and ratings. This process was performed by two independent researchers (MN and SL). After completion of the data extraction, results were compared between authors and, in case of uncertainty, discussed in the author group. After each article (with the 5 respective questions), a new chat session was opened.

Statistical Analysis

Descriptive statistics were used to summarize the performance of each AI tool across all retracted articles and questions. For each tool, the proportion of correct answers was calculated separately for each of the five predefined questions as well as across all questions combined. Results were reported as absolute counts and percentages.

To address the primary outcome, the number of retracted articles for which an AI tool provided correct answers to all five questions was determined. This was expressed as frequency and percentage of the total.

For secondary outcomes, error rates were calculated as follows:

1. Inclusion of retracted articles in topic overviews (Questions 1 and 2): Proportion of instances where a retracted article was mentioned without retraction warning.
2. Missed retraction status in search pathway (Questions 1–3): Proportion of instances in which a retracted article was included without an explicit warning of itsretraction status.
3. Intra-tool consistency: Agreement between duplicate question-answer pairs was assessed by calculating the percentage of conflicting ratings between reviewers for each AI tool.
4. Comparison between general-purpose AI tools and research-focused AI tools: Median number of correct answers per question was compared between the two groups.

Comparisons between general-purpose and research-focused AI tools were treated as exploratory. Differences between groups were summarized using descriptive statistics and effect size estimates, without formal hypothesis testing, due to the limited number of tools per group.

XSL•FO
RenderX

Interreviewer agreement during data extraction was assessed using Cohen kappa coefficient (κ), with percentage agreement reported as a complementary measure.

All analyses were conducted using R (version 4.3.2; R Foundation for Statistical Computing) and Microsoft Excel 2023 for data entry and descriptive summaries.

Results

Overview

Data extraction was performed between June and September 2025. A total of 675 prompts were submitted, each asked twice across all models.

The primary aim of this analysis was to assess how often each GenAI tool achieved a perfect score of 5/5 correct answers per question set. None of the evaluated models consistently achieved full accuracy. Three tools - SciSpace, ScienceOS, and Consensus - failed to provide any completely correct responses. The best-performing model was ChatGPT 5, which achieved 8 out of 15 perfect results (53.3%). Model-specific performance data can be seen in Figure 1.

Figure 1. Performance of nine freely available generative artificial intelligence (AI) tools in a pragmatic cross-sectional evaluation of their ability to correctly handle retracted scientific publications. Shown is the proportion of retracted articles (n=15) for which each artificial intelligence tool provided fully correct responses.

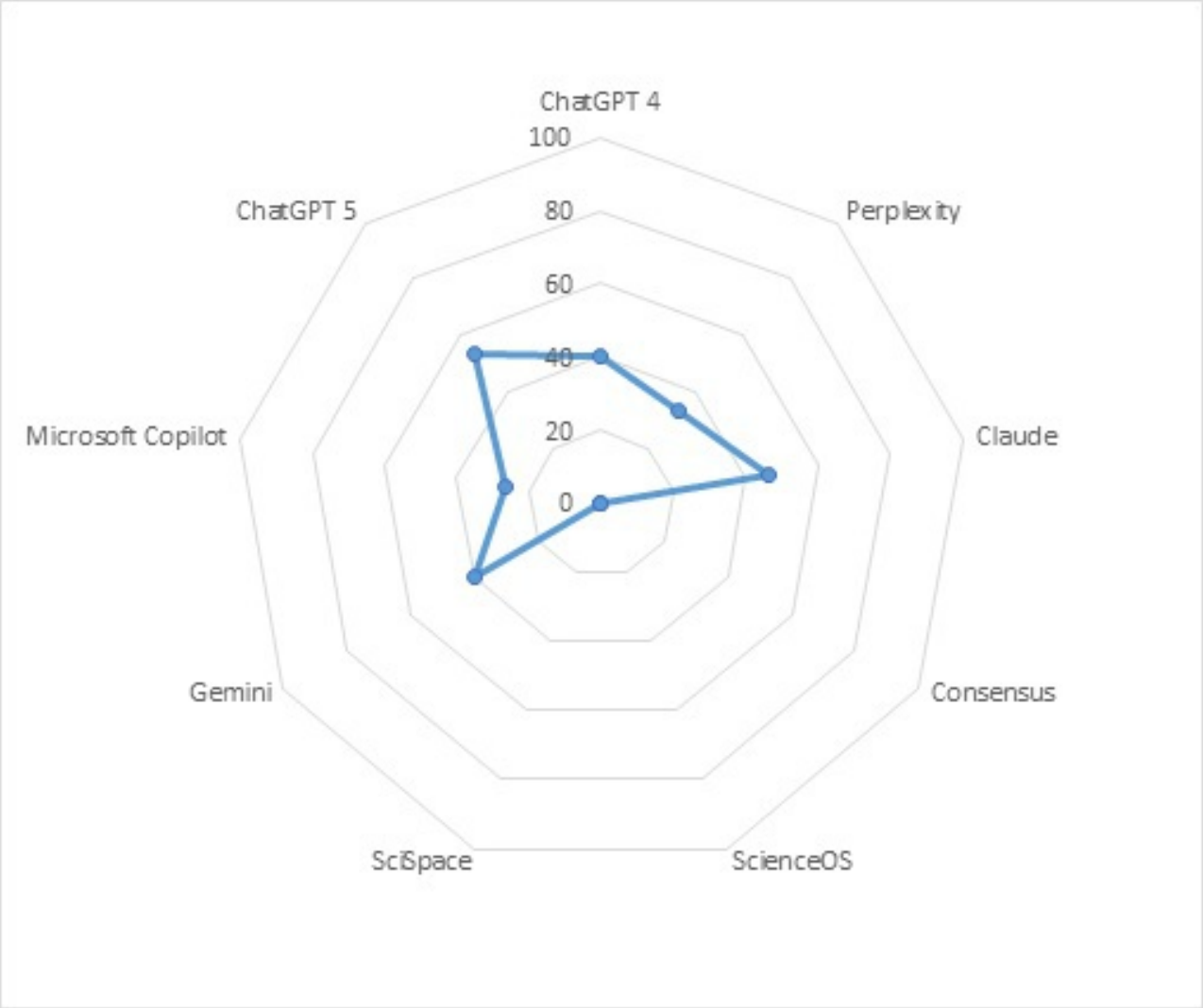

Retracted Articles in Topic Overviews and Search Pathway

To assess the risk of incorporating retracted studies into literature syntheses, topic overviews and combined topic overviews with article summaries were analyzed separately.

When examining only topic overviews (ie, the first two questions), Microsoft Copilot gave an incorrect answer in 2 (13.3%) out of 15 cases. SciSpace, on the other hand, had the most incorrect cases with 8 out of 15 cases. Exact model performances are presented in the supplementary file figure S1 in Multimedia Appendix 1 and Table 1.

XSL•FO
RenderX

To assess the risk of incorporating retracted studies into literature syntheses, topic overviews, and combined topic overviews with article summaries were analyzed separately.

Table . Primary analysis of nine freely available generative artificial intelligence tools in a pragmatic cross-sectional study evaluating the handling of retracted scientific literature. For each tool, the table reports the number and percentage of retracted articles (n=15) with fully correct responses to all five predefined questions, the frequency of unflagged inclusion of retracted articles in topic overviews, and intratool reliability based on discrepancies between repeated responses.

GenAI[a]	All 5 correct, n (%)	Retracted article in topic overviews, n (%)	Reliability, n (%)
ChatGPT 4	6 (40)	3 (20)	2 (2.67)
ChatGPT 5	8 (53.34)	3 (20)	5 (6.67)
Microsoft Copilot	4 (26.67)	2 (13.34)	8 (10.67)
Gemini	6 (40)	3 (20)	12 (16)
SciSpace	0 (0)	8 (53.34)	7 (9.33)
ScienceOS	0 (0)	6 (40)	8 (10.67)
Consensus	0 (0)	6 (40)	19 (25.33)
Claude	7 (46.67)	6 (40)	12 (16)
Perplexity	5 (33.34)	6 (40)	17 (22.67)

[a]GenAI: generative artificial intelligence.

Reliability of Tools

Each GenAI tool was asked every question twice to assess consistency. Differences between repeated answers resulting in a rating change occurred in each GenAI tool. ChatGPT 4 had the lowest variability with only two inconsistent answers (2.67%). Consensus had the highest incidence with 19 differences, approximately 1 in every 4 questions. Detailed results for each AI are shown in Table 1. We additionally calculated Cohen kappa, which ranged between 0.43 in Perplexity to 0.92 in ChatGPT, with an overall Cohen kappa of 0.73 (table S3 in Multimedia Appendix 1)

Comparison of General versus Scientific GenAI Tools

General-purpose AI tools achieved a median of 6 fully correct responses out of 15 (range 4‑8), corresponding to an aggregated accuracy of 40%. In contrast, research-focused tools did not achieve any fully correct responses (median 0, range 0‑0). The absolute difference in aggregated proportions between groups was 40 percentage points. Given the limited number of tools per group, this comparison is descriptive and exploratory.

Comparison of Keywords

Keyword overlap was assessed using a moderately generous matching strategy designed to account for realistic linguistic variation while maintaining methodological rigor. Specifically, orthographic variants (eg, “COVID-19” vs “covid19”), singular and plural forms (eg, “risk factor” vs “risk factors”), hyphenation differences (eg, “meta-analysis” vs “meta analysis”), and established abbreviations versus full terms (eg, “RCT” vs “randomized controlled trial,” “AI” vs “artificial intelligence”) were treated as equivalent. In addition, clearly synonymous or lexically overlapping technical terms referring to the same scientific entity or method (eg, “Alzheimer’s disease” vs “Alzheimer pathology,” or “clinical trials” vs “clinical trial”) were considered matches. In contrast, broader conceptual or hierarchical relationships were explicitly excluded from matching; for example, general versus specific terms (eg, “mental health” vs “depression”), different abstraction levels (eg, “hypertension” vs “blood pressure control”), or loosely related thematic concepts (eg, “public health” vs “disease prevention”) were not treated as equivalent. Pairwise comparisons showed substantial keyword overlap across all model pairs, ranging from 68% to 78%. The greatest overlap was observed between Claude and Perplexity (78%), followed by ChatGPT-4 and Perplexity (74%). ChatGPT-4 and Claude, and Gemini and Claude, each demonstrated 72% overlap, while Gemini and Perplexity showed 70% overlap and ChatGPT-4 and Gemini showed the lowest overlap at 68%. Comparable overlap magnitudes are observed both in comparisons involving the reference model and in comparisons among nonreference models, indicating a high degree of thematic convergence across LLMs. This matching strategy was selected to balance linguistic flexibility with analytical conservatism, avoiding both overly strict string-based matching and overly permissive conceptual aggregation. Nonetheless, we want to raise awareness that ChatGPT may have had a benefit as the keywords used for our search were generated with ChatGPT 4 and no perfect overlap was shown between AI tools.

Discussion

Principal Findings and Interpretation

In this cross-sectional evaluation of 9 freely available GenAI tools, none were able to consistently recognize or correctly handle retracted scientific publications. Across all 5 predefined research tasks, no model achieved perfect accuracy, and even the best-performing system, ChatGPT 5, correctly processed less than two-thirds of retracted articles (8 of 15; Figure 1). By contrast, AI tools marketed specifically for scientific use—Consensus, SciSpace, and ScienceOS—performed particularly poorly, each failing to produce a single fully correct

XSL•FO
RenderX

set of answers. As shown in Table 1, intra-tool variability was observed across all GenAI systems, with discordance rates between repeated queries ranging from 2.67% in ChatGPT 4% to 25.33% in Consensus, indicating that even the most consistent model failed to produce fully stable responses. These findings collectively indicate that current GenAI systems cannot reliably detect or flag retracted literature, even when explicitly queried about retraction status.

The key result of this study is the persistent inability of LLMs to identify retracted papers or to transparently communicate their status to users. This deficiency was observed both in exploratory topic overviews and in direct article queries, suggesting that the problem is systemic rather than situational. As illustrated in Figure 2, the models most frequently cited retracted studies without warning, and even when asked whether a given article had been retracted, many systems returned incorrect statements. These outcomes highlight a fundamental gap in the factual grounding of current AI tools—a limitation that directly undermines their use in scientific research and evidence synthesis. Our analysis cannot determine the exact reason why retracted articles are not marked correctly. Possibilities include the insufficient data retrieval from the original databases (including missing source details or incomplete retrieval) or errors in the reasoning of the respective GenAI. Nevertheless, the scientific end user, who ultimately carries the responsibility for the correctness of his or her article, should be aware of and cautious of this risk.

The contrast between general-purpose and domain-specific AI systems is notable. While one might expect specialized research tools to perform better, they were instead among the weakest performers. Several of these systems rely on static, proprietary databases that are not regularly synchronized with dynamic resources such as PubMed or CrossMark. This architectural rigidity may explain their inability to capture recent retraction events, especially among newly withdrawn papers included in our sample. Conversely, models such as ChatGPT 5 and Claude, which use broader retrieval mechanisms and more frequent model updates, achieved somewhat higher accuracy but still lacked robustness and transparency. Importantly, the results presented represent a performance snapshot of freely accessible AI tools at the time of data collection and may not generalize to subscription-based versions or to future model iterations, which may incorporate improved retrieval mechanisms or retraction-aware features.

Figure 2. Accuracy of freely available generative artificial intelligence (AI) tools across the first three predefined literature-search tasks: topic overview, identification of relevant articles, and summarization of a specified retracted publication.

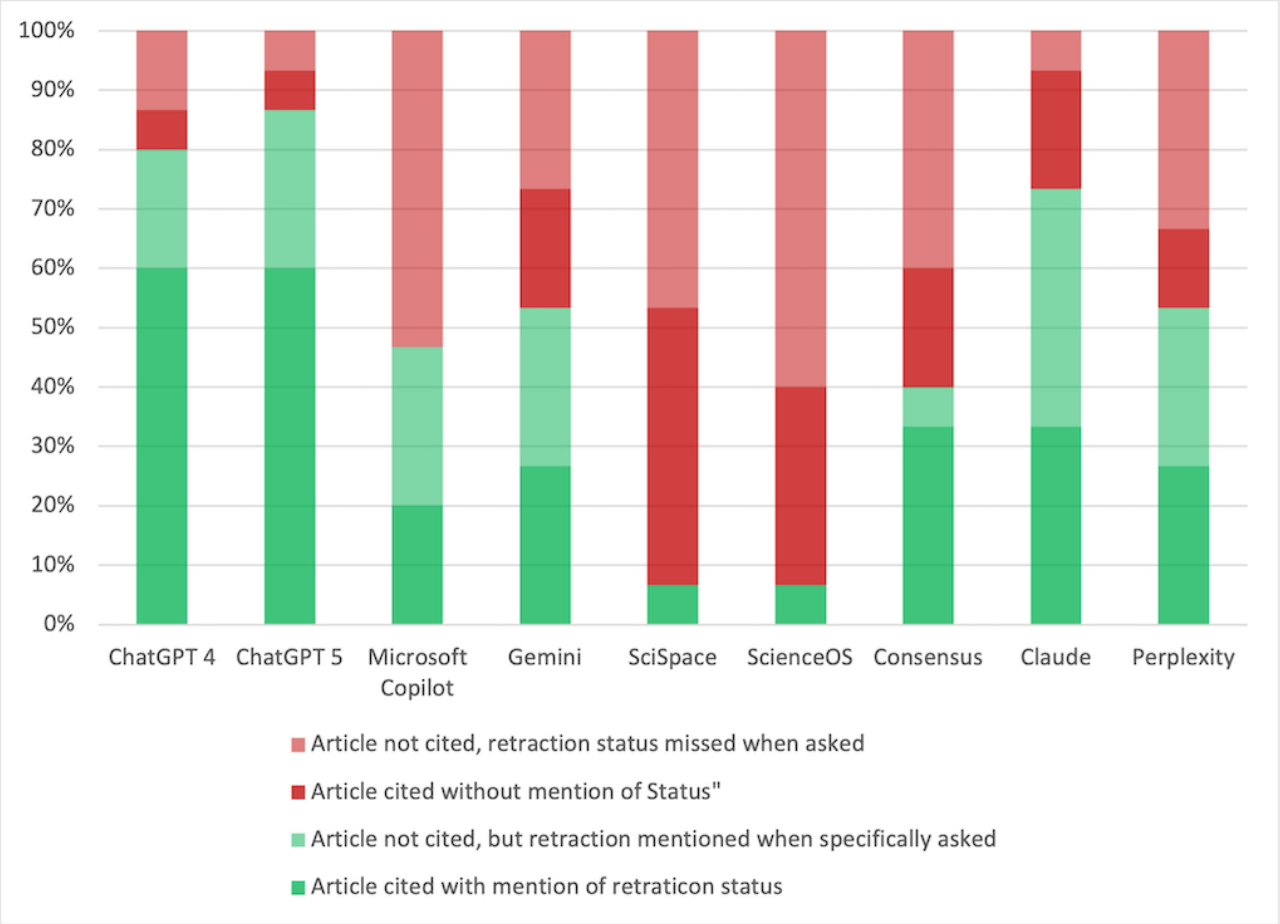

Comparison With Previous Literature

Our findings extend previous work that has documented the use of retracted literature by GenAI. Gu et al [11] reported that ChatGPT produced answers referencing retracted oncology studies, without acknowledging their retraction status, in more than 70% of tested prompts. Similarly, Jan et al [10] and Ge et al [6] have shown that GenAI tools frequently fabricate or

XSL•FO
RenderX

misattribute citations during literature summarization tasks. Importantly, these earlier studies typically evaluated individual models or single fields within the scientific literature. The present analysis broadens the evidence base by systematically comparing multiple AI systems across heterogeneous scientific domains, including both general and medical literature. The consistency of failure across all tools underscores that this is not an isolated design flaw but a pervasive, cross-platform limitation in current GenAI architectures.

The continued citation and reuse of retracted work represent a well-recognized threat to research integrity and, in clinical disciplines, may even interfere with patient safety as illustrated by the infamous article of Wakefield et al [13,29,32]. Our results demonstrate that GenAI tools risk amplifying this long-standing problem by automating the propagation of unreliable information. Particularly concerning is the tendency of GenAI tools to reproduce incorrect content with high linguistic confidence, which may give users a false sense of authority and discourage verification. Without built-in mechanisms to cross-check bibliographic metadata or retraction notices, AI-assisted literature synthesis may inadvertently erode rather than enhance scientific reliability.

Implications for Research and Clinical Practice

Given the accelerating adoption of AI systems in academic and clinical workflows, these findings have immediate practical implications. Researchers increasingly use GenAI tools to generate literature summaries, draft background sections, and identify relevant studies. In medicine and health sciences, where evidence accuracy is critical, the unrecognized inclusion of retracted data could distort systematic reviews, meta-analyses, and ultimately clinical recommendations. As such, reliance on unverified AI output in research and clinical contexts may indirectly contribute to patient harm by propagating retracted or unreliable evidence.

Importantly, expectations regarding the handling of retracted literature may reasonably differ depending on the intended purpose of an AI system. For AI tools explicitly designed to support scientific research or evidence synthesis, excluding retracted publications by default or clearly flagging their status aligns with established principles of research integrity and evidence-based practice. In contrast, for general-purpose AI systems, outright exclusion may not always be appropriate, as retracted articles can still be relevant for historical, methodological, or meta-scientific purposes. In such cases, clearly communicating the retraction status to users may represent a more practical and responsible design approach. Given that retracted information cannot be removed from a trained LLM, handling is more realistically achieved through post-generation checks rather than through changes to the underlying model itself.

To mitigate this risk, several structural measures should be prioritized. AI developers should integrate real-time bibliographic verification pipelines using authoritative sources such as Crossref or PubMed Retraction Notices. AI interfaces intended for academic use should provide transparent confidence indicators or citation provenance, enabling users to identify potentially unreliable results. Journals and peer reviewers should require explicit disclosure of any AI assistance in literature searches and verify cited sources independently. These steps are consistent with recent policy recommendations emphasizing “responsible AI use” and “model interpretability” in scientific contexts [33]. These recommendations align with JMIR editorial policies and ICMJE guidance, which emphasize transparent disclosure of AI use, author responsibility for content accuracy, and independent verification of cited sources.

Strengths and Limitations

This study offers a pragmatic, reproducible evaluation of AI performance under conditions that mirror real-world research behavior. By using predefined and repeated queries on a diverse set of retracted papers, the design captured both intermodel variability and intramodel inconsistency (Table 1). The inclusion of the 10 most-cited and 5 most-recently retracted papers provided a balanced sample, revealing that both historically influential and newly withdrawn studies were equally mishandled.

Nonetheless, several limitations merit consideration. First, the study was confined to freely accessible versions of each AI tool, excluding subscription models that may have more advanced retrieval capabilities and was taken at a specific point in time. Second, while the test set was broad, it remains a small subset of the retraction landscape and may not generalize to all fields, especially since the majority of articles (13/15, 86.7%) were medical articles. While this reflects the intended application of many AI-assisted literature workflows, the predominance of biomedical publications in our sample may limit the generalizability of these findings to other scientific fields, where differing publication cultures, citation structures, and retraction practices could influence AI performance. Third, AI models evolve rapidly; the specific results reported here represent a snapshot of performance as of mid-2025 and may not reflect future iterations. Fourth, given the limited number of retracted articles (n=15), the study was not powered to detect small-to-moderate differences between tools, and nonsignificant results should therefore be interpreted cautiously due to the risk of type II error. Accordingly, between-tool comparisons were considered exploratory and intended to describe observed performance patterns rather than to support confirmatory claims. Finally, the binary scoring approach (correct vs incorrect) captures factual accuracy but does not account for nuances such as ambiguous phrasing (heavily criticized and scrutinized). While distinguishing whether AI systems correctly identify the specific reasons for retraction could provide additional nuance, our analysis intentionally focused on the more fundamental step of detecting and flagging retracted literature, as failure at this stage already represents a critical risk for research integrity.

These factors should be addressed in future validation studies.

This study has several methodological flaws. One being that our way of questioning needed a uniform, comparable way. This was achieved by combining the keywords in a single prompt string, which could differ from a prompt by a common user searching for a specific topic. Furthermore, there are multiple ways why a retracted article was not cited by an AI. This could be due to the fact that the AI decided against citing it, due to an incomplete or outdated database, due to the fact

XSL•FO
RenderX

that this article does not report data of interest, or that there are better articles describing similar aspects. In conclusion, we cannot specifically differentiate between all of those possibilities and therefore decided to create real-world judgment in which we just decide whether this article was included without warning or not. We additionally created an exploratory control group in which we used publication-order matched articles to estimate the ability of each AI to detect a specific article. While the results showed that AI tools found the respective article in up to 90% of cases, we want to highlight the limitation that this extraction was performed in newer iterations of each AI tool.

Future Directions

As GenAI becomes increasingly embedded in research ecosystems, systematic benchmarking and transparent reporting of factual reliability are essential. Future studies should expand sample sizes, include subscription-based models, and explore automated cross-validation with bibliometric databases. In parallel, developers should implement retraction-aware pipelines capable of dynamic metadata synchronization. Collaborative efforts between publishers, AI developers, and database curators could enable "AI integrity layers" that automatically flag or exclude retracted content before output generation. Ultimately, ensuring that AI systems respect the integrity of the scientific record is both a technical and ethical imperative.

Conclusions

No currently available free-access GenAI system can be considered reliable for detecting or handling retracted literature. Even the best-performing models failed to consistently recognize the retraction status of scientific articles, while research-focused tools performed particularly poorly. As AI tools become ubiquitous in the preparation and evaluation of scientific work, independent verification of sources and explicit retraction checks remain indispensable to maintain research integrity.

Acknowledgments

The authors declare that no generative artificial intelligence tools were used in the writing of any portion of this manuscript.

Funding

No external fundings were received for this study.

Data Availability

The data that support the findings of this study are available from the corresponding author (SH) upon reasonable request.

Authors' Contributions

Contributed to the idea, hypothesis, and study design: HBC, MB, MN, NS, SH, SL
Contributed to data acquisition: MN, SH, SL
Assembled the data and performed the statistical analysis: HBC, MB, NS, SL
Interpreted the data and prepared the first draft of the manuscript: MB, MN, NS, SH, SL
Commented on, edited, reviewed, and finally approved the last version of the manuscript: HBC, MB, MN, NS, SH, SL
Authors SL and MN share the first authorship.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary digital content.
[DOCX File, 81 KB - jmir_v28i1e88766_app1.docx]

References

1. Golan R, Reddy R, Muthigi A, Ramasamy R. Artificial intelligence in academic writing: a paradigm-shifting technological advance. Nat Rev Urol 2023 Jun;20(6):327-328. [doi: 10.1038/s41585-023-00746-x] [Medline: 36829078]
2. Khalifa M, Albadawy M. Using artificial intelligence in academic writing and research: an essential productivity tool. Computer Methods and Programs in Biomedicine Update 2024;5:100145. [doi: 10.1016/j.cmpbup.2024.100145]
3. Dwivedi YK, Kshetri N, Hughes L, et al. Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int J Inf Manage 2023 Aug;71:102642. [doi: 10.1016/j.ijinfomgt.2023.102642]
4. Dergaa I, Chamari K, Zmijewski P, Ben Saad H. From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. Biol Sport 2023 Apr;40(2):615-622. [doi: 10.5114/biolsport.2023.125623] [Medline: 37077800]
5. de la Torre-López J, Ramírez A, Romero JR. Artificial intelligence to automate the systematic review of scientific literature. Computing 2023 Oct;105(10):2171-2194. [doi: 10.1007/s00607-023-01181-x]

XSL•FO
RenderX

6. Ge L, Agrawal R, Singer M, et al. Leveraging artificial intelligence to enhance systematic reviews in health research: advanced tools and challenges. Syst Rev 2024 Oct 25;13(1):269. [doi: 10.1186/s13643-024-02682-2] [Medline: 39456077]
7. Khlaif ZN, Mousa A, Hattab MK, et al. The potential and concerns of using AI in scientific research: ChatGPT performance evaluation. JMIR Med Educ 2023 Sep 14;9:e47049. [doi: 10.2196/47049] [Medline: 37707884]
8. Májovský M, Černý M, Kasal M, Komarc M, Netuka D. Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: pandora's box has been opened. J Med Internet Res 2023 May 31;25:e46924. [doi: 10.2196/46924] [Medline: 37256685]
9. Montalbano P. Ethical and legal considerations of medical artificial intelligence. Prim Care 2025 Dec;52(4):769-779. [doi: 10.1016/j.pop.2025.07.009] [Medline: 41110915]
10. Jan R. Examining the reliability of ChatGPT: identifying retracted scientific literature and ensuring accurate citations and references. In: Impacts of Generative AI on the Future of Research and Education: IGI Global Scientific Publishing; 2024:367-392. [doi: 10.4018/979-8-3693-0884-4.ch014]
11. Gu T, Feng H, Li M, Gu W, Wang G. Alarm: retracted articles on cancer imaging are not only continuously cited by publications but also used by ChatGPT to answer questions. J Adv Res 2025 May;71:1-3. [doi: 10.1016/j.jare.2025.03.020] [Medline: 40086628]
12. Kocyigit BF, Okyay RA, Seiil B, Qumar AB, Sumbul HE. Analysis of retracted publications on artificial intelligence: trends, ethical concerns, and scientific integrity. J Korean Med Sci 2025 Nov 17;40(44):e280. [doi: 10.3346/jkms.2025.40.e280] [Medline: 41250649]
13. Bakker C, Boughton S, Faggion CM, Fanelli D, Kaiser K, Schneider J. Reducing the residue of retractions in evidence synthesis: ways to minimise inappropriate citation and use of retracted data. BMJ Evid Based Med 2024 Mar 21;29(2):121-126. [doi: 10.1136/bmjebm-2022-111921] [Medline: 37463764]
14. AI-assisted knowledge gathering and its pitfalls with retracted articles. Open Science Framework (OSF). URL: https://doi.org/10.17605/OSF.IO/B6J2W [accessed 2026-04-13]
15. Bohsas H, Alibrahim H, Swed S, et al. Prevalence and knowledge of polycystic ovary syndrome (PCOS) and health-related practices among women of Syria: a cross-sectional study. J Psychosom Obstet Gynaecol 2024 Dec;45(1):2318194. [doi: 10.1080/0167482X.2024.2318194] [Medline: 38635351]
16. Estruch R, Ros E, Salas-Salvadó J, et al. Primary prevention of cardiovascular disease with a Mediterranean diet. N Engl J Med 2013 Apr 4;368(14):1279-1290. [doi: 10.1056/NEJMoa1200303] [Medline: 23432189]
17. Fortuna F, Gonzalez D, Fritzler A, et al. Burnout components, perceived stress and hair cortisol in healthcare professionals during the second wave of COVID 19 pandemic. Sci Rep 2024 Nov 21;14(1):28828. [doi: 10.1038/s41598-024-79925-8] [Medline: 39572609]
18. Fukuhara A, Matsuda M, Nishizawa M, et al. RETRACTED: Visfatin: a protein secreted by visceral fat that mimics the effects of insulin. Science 2005 Jan 21;307(5708):426-430. [doi: 10.1126/science.1097243] [Medline: 15604363]
19. Gautret P, Lagier JC, Parola P, et al. RETRACTED: Hydroxychloroquine and azithromycin as a treatment of COVID-19: results of an open-label non-randomized clinical trial. Int J Antimicrob Agents 2020 Jul;56(1):105949. [doi: 10.1016/j.ijantimicag.2020.105949] [Medline: 32205204]
20. Huang C, Huang L, Wang Y, et al. 6-month consequences of COVID-19 in patients discharged from hospital: a cohort study. Lancet 2021 Jan 16;397(10270):220-232. [doi: 10.1016/S0140-6736(20)32656-8] [Medline: 33428867]
21. Jiang Y, Jahagirdar BN, Reinhardt RL, et al. Pluripotency of mesenchymal stem cells derived from adult marrow. Nature New Biol 2002 Jul 4;418(6893):41-49. [doi: 10.1038/nature00870] [Medline: 12077603]
22. Lesné S, Koh MT, Kotilinek L, et al. A specific amyloid-beta protein assembly in the brain impairs memory. Nature New Biol 2006 Mar 16;440(7082):352-357. [doi: 10.1038/nature04533] [Medline: 16541076]
23. Mohammadniaei M, Zhang M, Ashley J, et al. A non-enzymatic, isothermal strand displacement and amplification assay for rapid detection of SARS-CoV-2 RNA. Nat Commun 2021 Aug 24;12(1):5089. [doi: 10.1038/s41467-021-25387-9] [Medline: 34429424]
24. Morisky DE, Ang A, Krousel-Wood M, Ward HJ. Predictive validity of a medication adherence measure in an outpatient setting. J Clin Hypertens (Greenwich) 2008 May;10(5):348-354. [doi: 10.1111/j.1751-7176.2008.07572.x] [Medline: 18453793]
25. Papoutsis A, Borody T, Dolai S, et al. Detection of SARS-CoV-2 from patient fecal samples by whole genome sequencing. Gut Pathog 2021 Jan 30;13(1):7. [doi: 10.1186/s13099-021-00398-5] [Medline: 33516247]
26. Pearton SJ, Norton DP, Ip K, Heo YW, Steiner T. RETRACTED: recent progress in processing and properties of ZnO. Prog Mater Sci 2005 Mar;50(3):293-340. [doi: 10.1016/j.pmatsci.2004.04.001]
27. Shi Y, Ye P, Wang C, Yang K, Guo J. RETRACTED: surface hydrophilic modification for chip of centrifugal microfluidic immunoassay system. Micromachines (Basel) 2022 May 26;13(6):831. [doi: 10.3390/mi13060831] [Medline: 35744444]
28. Taylor DD, Gercel-Taylor C. MicroRNA signatures of tumor-derived exosomes as diagnostic biomarkers of ovarian cancer. Gynecol Oncol 2008 Jul;110(1):13-21. [doi: 10.1016/j.ygyno.2008.04.033] [Medline: 18589210]
29. Wakefield A, Murch S, Anthony A, et al. RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. The Lancet 1998 Feb;351(9103):637-641. [doi: 10.1016/S0140-6736(97)11096-0]

XSL•FO
RenderX

30. Top 10 most highly cited retracted papers. Retraction Watch. URL: https://retractionwatch.com/the-retraction-watch-leaderboard/top-10-most-highly-cited-retracted-papers/ [accessed 2026-04-13]
31. Retraction watch data. GitLab. URL: https://gitlab.com/crossref/retraction-watch-data [accessed 2026-04-13]
32. Teixeira da Silva JA, Santos‐d'Amorim K, Bornemann‐Cimenti H. The citation of retracted papers and impact on the integrity of the scientific biomedical literature. Learn Publ 2025 Apr;38(2):e1667. [doi: 10.1002/leap.1667]
33. AllahRakha N. UNESCO's AI ethics principles: challenges and opportunities. Irshad J Law and Policy 2024;2(9):24-36. [doi: 10.59022/ijlp.225]

Abbreviations

AI: artificial intelligence
GenAI: generative artificial intelligence
LLM: large language model

Edited by A Coristine; submitted 01.Dec.2025; peer-reviewed by DB Patel, R Marshall, S Mirji, S Singireddy, T Foltynek; revised version received 24.Mar.2026; accepted 28.Mar.2026; published 01.May.2026.

Please cite as:
Labenbacher S, Niederer M, Hammer S, Bader M, Schreiber N, Bornemann-Cimenti H
Performance of AI Tools in Citing Retracted Literature : Content Analysis
J Med Internet Res 2026;28:e88766
URL: https://www.jmir.org/2026/1/e88766
doi:10.2196/88766

© Sebastian Labenbacher, Maximilian Niederer, Sascha Hammer, Matthias Bader, Nikolaus Schreiber, Helmar Bornemann-Cimenti. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 1.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Artificial Intelligence in Health Professions Education: Qualitative Study of Student Experiences

Marina Guirguis[1*], BA, BSc; Salomon Fotsing[1,2,3,4*], MD, MAEd; Joanne Fevry[1,2*], MD, MScA; Christine Landry[5*], BPharm, ACPR, MSc, BCPS, PharmD; Diane Bouchard-Lamothe[2*], MSc; Jennifer Lacroix[2*], MScHQ; Alireza Jalali[1,2,6*], MD

[1]Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada
[2]Francophone Affairs, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada
[3]Department of Family Medicine, University of Ottawa, Ottawa, ON, Canada
[4]Institut du Savoir Montfort, Ottawa, ON, Canada
[5]School of Pharmaceutical Sciences, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada
[6]Department of Innovation in Medical Education, Faculty of Medicine, University of Ottawa, 451 Smyth (3247BD), Ottawa, ON, Canada
[*]all authors contributed equally

Corresponding Author:
Alireza Jalali, MD
Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

Abstract

Background: Artificial intelligence (AI) is increasingly integrated into education and health care, raising questions about how students use these technologies and how AI influences their learning. In health education, understanding these trends is particularly important because student learning directly impacts future clinical skills.

Objective: This study aimed to explore the use of AI tools by health sciences students at the University of Ottawa. More specifically, it sought to identify the most frequently used AI tools, describe students' usage habits, determine which tools support knowledge acquisition and skill development, and gather students' recommendations for effective strategies to raise awareness and train their peers on the responsible use of AI.

Methods: A qualitative approach was used with students from 10 health professions who reported using AI in their studies. Data were collected through semistructured interviews and an open-ended qualitative online survey. Inductive thematic analysis within an interpretive paradigm was applied to capture patterns, perceptions, and emergent themes.

Results: A total of 51 health professions students participated in the study. Most were women between the ages of 20 and 29 years. ChatGPT (OpenAI) emerged as the most frequently used AI tool. Students perceived AI as a complementary tool that facilitated knowledge acquisition, skill development, writing, and problem-solving. AI adoption was driven by curiosity, peer influence, and the desire to improve work efficiency. Students critically evaluated AI results, integrated the tools into their learning processes, and emphasized the importance of technical skills, critical thinking, and digital literacy. Peer learning, hands-on demonstrations, and access to online resources were recommended for effective AI training.

Conclusions: This research demonstrates that health professions students actively use AI tools, particularly ChatGPT, to support learning, skill development, and academic tasks. Although AI is valuable as an educational aid and its use varies by student and context, this highlights the need for structured guidance, critical evaluation skills, and peer-supported training. These findings highlight the importance of thoughtfully integrating AI into educational programs to enhance learning outcomes, foster skill acquisition, and ensure responsible and effective adoption.

(J Med Internet Res 2026;28:e82432) doi:10.2196/82432

KEYWORDS

artificial intelligence; AI; qualitative study; health professions education; learning technologies; ChatGPT; learning experience; digital literacy

Introduction

Presently, the use of artificial intelligence (AI) has become commonplace in health care, health professions training, and the pharmaceutical industry. AI refers to the capacity of machines to simulate human functions such as reasoning, learning, and problem-solving through algorithms [1,2]. This technology, which "mimics human cognitive functions," is becoming increasingly effective in the health care sector.

XSL•FO
RenderX

Although it sometimes raises ethical concerns [3], AI now assists clinicians in making better decisions, especially in fields like diagnostic imaging, where it aids in image interpretation [2,4], and dermatology, where it helps in the early detection of skin cancer by using efficient classifiers for skin tumors [5]. In addition, in the pharmaceutical field, AI accelerates drug design and discovery processes [2,6].

The use of AI in health professional education is also growing. With the aid of tools like ChatGPT (OpenAI), it can create simulated clinical environments, evaluate clinician-simulated patient communication, create quizzes, and develop curricula [7]. At the same time, it can respond instantly to specific questions related to students' areas of study through AI chatbots used in medical education [8]. The new generation of health learners is much more inclined to adopt and use this technology. In 2022, a national Canadian survey of nearly 2000 health professions students showed that AI was viewed favorably by students, who also believe that this technology will positively influence their future careers [9]. Similarly, in March 2023, a systematic review reported that health professions students have a favorable attitude toward the use of AI [10].

From an academic standpoint, the use of AI by health professions students increasingly raises the question of the "learning experience." Learning is a dynamic concept defined as a process that allows individuals to perceive objects, interact with them, and integrate them into their social, cognitive, and emotional dimensions to transform, create, or evolve their cognitive structure [11]. The learning experience goes beyond knowledge acquisition. It facilitates the development of skills and cultural values through observation, imitation, trial, repetition, and presentation [12]. Therefore, an important question regarding AI in health professions training programs is how AI contributes to the students' learning experience. Shorey et al [13] scoping review of the learning styles, preferences, and needs of Generation Z health students marks a first step in this direction. Given the increasing adoption of AI, raising awareness and providing training on the proper use of these tools has become essential [9]. Peer education is a viable approach because it allows students to share valuable insights and fosters a collaborative learning environment that encourages peers to use AI tools [14,15].

Access to AI tools is also an important element in the training of health science students, given their increasing use in modern learning and clinical reasoning processes [16]. However, socioeconomic and linguistic factors may influence their use by these students, as some of these tools are mainly programmed in English and sometimes require payment to access more advanced features [16,17]. In a bilingual institution of higher learning such as the University of Ottawa, where courses are taught in French, English, or both, it is essential to ensure equitable access to these resources for all students.

This qualitative study aimed to explore which AI tools health professions students at the University of Ottawa report using, describe their usage habits, identify the tools that help them acquire new knowledge and develop skills, and explore what these students consider to be the best strategies for raising awareness and training their peers on the proper use of AI in learning environments. Specifically, we focused on how students use AI tools to support their learning and on strategies they recommend for peer awareness and training.

Methods

Qualitative Approach and Research Paradigm

An inductive thematic analysis approach, following an interpretive research paradigm, was used to identify the AI platforms most used by health professions students during their learning activities. This approach was chosen for its relevance to explore perceptions and experiences regarding AI tools in medical education [18]. This approach, based on the interpretive paradigm, posits that participants' narratives construct meanings, which promotes a detailed understanding of behaviors and points of view [19]. Therefore, in accordance with the research objectives, the inductive approach facilitated the emergence of themes directly from the data for a more flexible and data-driven interpretation [20]. This paper was written by using the Standards for Reporting Qualitative Research (SRQR) guide (Checklist 1) [21].

Research Characteristics and Reflexivity

The research team was multidisciplinary and followed a collaborative approach, demonstrating reflexivity at each phase of the study. Team members included MG, an undergraduate medical student experienced in qualitative research, who contributed to study design, data collection, thematic analysis, and writing; SF, a physician and clinical research manager, who provided methodological guidance, qualitative analysis, results interpretation, and manuscript revision; JF, an anatomy educator with expertise in qualitative research and scientific writing, supported data coding, thematic analysis, and writing; CL, a clinician-researcher and pharmacy program director, offered insights into results interpretation; DB-L, a medical education advisor, contributed to the interpretation of the results and the revision of the manuscript; JL, a program manager experienced in qualitative analysis, helped organize and coordinate the project; and AJ, a physician and anatomy professor with expertise in technology-assisted learning and educational innovation, contributed to data analysis, results interpretation, and manuscript revision.

Regular meetings encouraged open dialogue, allowing all coauthors to critically reflect on their assumptions, disciplinary lenses, and interpretations throughout data collection, analysis, and manuscript writing.

Context

Our study was conducted among health professions students at the University of Ottawa. All data were taken outside regular class hours, either through online interviews on Zoom or via an asynchronous online questionnaire.

Sampling Strategy

A nonprobabilistic sampling method was used, following the recommendations of Murphy et al [22] and Higginbottom [23]. Only students who used or had previously used AI platforms for their learning activities in the previous academic year were eligible. To identify eligible participants, all students in the

XSL•FO
RenderX

targeted programs received an invitation email including a brief screening question asking whether they had used AI tools (eg, ChatGPT, Perplexity AI, or any other AI-assisted learning platforms) for their academic activities in the last academic year. Only those who answered "yes" were invited to participate in the study, while students who responded "no" were excluded.

The targeted programs included medicine, pharmacy, molecular and translational medicine, nursing, occupational therapy, physiotherapy, audiology, speech-language pathology, health sciences, and psychology. Our target sample size was based on the writings of Guest et al [24], which specified that in a qualitative study, fundamental elements of the themes studied are present in participants' discourse after 6 interviews, and theoretical saturation of information is achieved after the first 12 interviews. Therefore, we planned to recruit a minimum of 6 participants from each training program, with no restrictions on the maximum number of participants. The distribution of students by program was not tracked separately for the interviews and the questionnaire; overall characteristics of participants are presented in Table 1.

Table . Sociodemographic characteristics of participants (N=51).

Category	Values, n
Age (years)	
<20	9
20 - 29	39
30 - 39	2
40 - 49	1
≥50	0
Gender	
Women	41
Men	9
I prefer not to answer	1
Others	0
Health professions	
Health sciences	27
Medicine	8
Psychology	8
Nursing	1
Pharmacy	1
Others	
Biomedical sciences	5
Biology	1

Data Collection

Data collection began after the recruitment phase, between January and March 2024. Invitation emails to participate in the research project were sent to all students in the targeted programs. Those who agreed to participate were met for semistructured online interviews via Zoom, lasting 30 - 45 minutes each. These interviews were conducted according to an interview guide developed in advance by the research team (see Section A of Multimedia Appendix 1). The guide aimed at capturing information on the most used AI platforms by health professions students at the University of Ottawa, their usage habits of these platforms, the platforms that they believed allowed them to acquire new knowledge and develop skills, and the best strategies for raising awareness and training their peers in the appropriate use of AI in learning.

The guide also included short vignettes to help organize questions about learning processes with AI tools. These vignettes, composed of a definition of the learning process [25,26] and Kolb's [27] experiential learning model, are fully incorporated into the guide provided in Section A of Multimedia Appendix 1.

However, after only 6 interviews over a 4-month period, data collection was adapted to include an online questionnaire using the same interview guide. A total of 51 students made up the final sample after 6 students participated in semistructured interviews and 45 students completed the online qualitative questionnaire. The interview guide and the online questionnaire were made available in both French and English to accommodate participants' language preferences and diffused between June and July 2024. The interview guide was reproduced identically on the online questionnaire platform (SurveyMonkey), thus ensuring that the same questions were asked in both data

XSL•FO
RenderX

collection methods while ensuring methodological continuity. New invitation emails were sent to all students in the targeted programs.

Data Processing and Analysis

Interview Process

The semistructured interviews were conducted on Zoom. No questions that could identify participants were asked. The audio of the interviews was recorded and transcribed using the WhisperAI transcription application. The transcriptions were then verified by a member of the research team for accuracy. To protect anonymity, the transcripts were saved in the order in which the interviews were conducted, and no identifiers were used in the file names. Moreover, interviews and transcriptions were conducted by two separate individuals. All files were password-protected.

Survey Process

Following ethics modification, a qualitative online questionnaire designed to elicit narrative and open-ended responses was hosted on the SurveyMonkey platform. Only one optional question allowing for the identification of participants was asked: "participants could provide their email address at the end of the questionnaire if they wished to be included in a draw for one of five gift cards as compensation." To protect anonymity, participants' responses were saved in the order in which they were received and were coded (eg, P1 for participant 1). The responses were stored in a password-protected Microsoft Excel file and kept separate from the email addresses, which were also in a password-protected Microsoft Excel file.

During the analysis, the information collected during the interviews and questionnaires was treated together as a single qualitative corpus. All coding was performed inductively; although Kolb's experiential learning model informed the development of the questions, it did not guide the coding process. Each participant's questionnaire responses were treated as individual transcripts and coded using the same inductive thematic analysis procedure as the interview data. Transcripts were read and revised several times to identify patterns and themes emerging from the data. To reinforce the methodological rigor of our qualitative study, a double thematic analysis was carried out. First, a researcher identified and categorized emerging themes from the corpus of data. Subsequently, a second researcher went back over all the material independently to reproduce the thematic analysis. This approach is part of a logic of analytical triangulation, aimed at limiting interpretive bias and reinforcing the credibility of the results as recommended by Paillé and Mucchielli [28]. The confrontation of analyses enabled us to discuss divergences, refine thematic categories, and arrive at a consensual interpretation of the material.

Techniques to Enhance Trustworthiness

To ensure the trustworthiness of the analysis, the data were analyzed systematically. An experienced qualitative researcher manually coded 100% of the data, and a second member of the research team independently coded 25% of the dataset. This second coder was trained by the first to master the principles of qualitative analysis and the steps involved in coding interviews and to ensure accurate coding. The second coder coded two randomly selected complete transcripts and half of a third. The two coders then reviewed and compared their coding, and clarifications were made to better represent the participants' experience until consensus was reached. No formal Cohen κ or percent agreement was calculated; reliability was ensured through this consensus-based process.

Next, a tree diagram was created to organize the coded data into potential themes based on observed similarities. Latent thematic analysis, using an interpretive approach, was used to identify subthemes. Finally, the two coders discussed the themes and subthemes related to the selected quotes. The coding framework was developed using NVivo software (version 12; QSR International) [29].

Ethical Considerations

This study was reviewed and approved in accordance with the standards of the Research Ethics Board (REB) of the University of Ottawa (#H-11-23-9455). All participants provided informed consent prior to data collection, which took place outside regular class hours. Due to a low recruitment rate, a modification of the ethics protocol was requested and approved by the REB to allow participants to complete an online questionnaire instead of taking part in interviews. This modification was implemented during the data collection phase. Compensation was offered in the form of an optional draw for one of five CAD $25 (US $18.08) gift cards.

Participant privacy and confidentiality were strictly maintained throughout the study. All data were anonymized using participant codes, and no identifying information was included in the transcript or analysis. All data were securely stored and accessible only to the research team.

Results

Participant Characteristics and Qualitative Findings

In total, 51 students from various health-related fields who had used AI tools during their training activities participated in this study. Most participants were between 20 and 29 years old (n=39). In terms of gender identity, the majority identified as women (n=41), followed by men (n=9), and 1 participant preferred not to disclose. No participants identified as nonbinary or transgender persons.

Health sciences was the most reported field of study (n=27), followed by medicine and psychology, each with 8 participants (see Table 1).

Participants reported using a wide variety of AI tools. ChatGPT was by far the most frequently used, with 49 respondents indicating its use. Other tools mentioned included Microsoft Copilot (n=2), Perplexity AI (n=2), Anthropic Claude (n=2), and several specialized AI tools, such as Google Bard AI, Summarizer AI, Grammarly AI, Midjourney, and Scite, each cited by one respondent (see Section B of Multimedia Appendix 1 for further details).

Themes were identified and developed from the content of the semistructured interviews and the responses to the qualitative

XSL•FO
RenderX

survey, then organized into 4 main thematic areas (see Figure 1 for a summary).

Figure 1. Concept map illustrating the emerging themes and subthemes of the qualitative data analysis. AI: artificial intelligence.

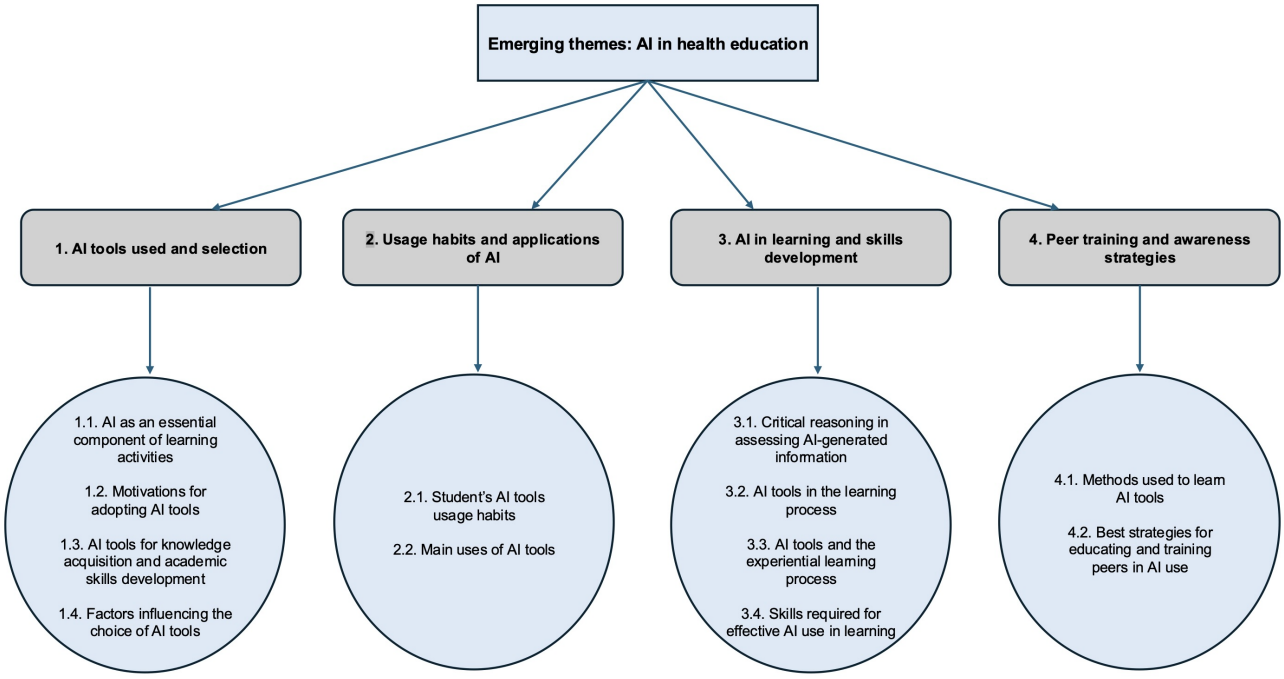

Theme 1: AI Tools Used and Selection

Theme 1.1: AI as an Essential Component of Learning Activities

Health care students have varying opinions about the role of AI in their learning activities. Among 51 respondents, 30 rated AI as essential or very beneficial. These students highlight its ability to enhance learning by helping to organize course materials, provide quick answers, and personalize educational experiences. For example, one participant illustrated this by stating:

> *I have used ChatGPT to help summarize papers and great sample questions and to read my work and compare it to the rubric.* [Participant 38]

They highlighted its usefulness in saving time and improving learning through tailored explanations and help with challenges like clarifying complex ideas or improving writing. As one participant noted:

> *Yes. I think it is an essential component because it can offer tools to approach/view your research in new ways. Therefore, it can aid in overcoming any hurdles, like feeling stuck when writing.* [Participant 23]

However, 21 respondents did not perceive AI as essential. Many of them regarded AI as a complementary aid rather than a primary resource, recognizing its usefulness while questioning its necessity. As one respondent noted:

> *I don't consider it to be essential, but I consider it to be helpful.* [Participant 19]

Concerns were raised about the reliability of AI, with one participant citing the presence of incomplete or inaccurate information on online platforms. Others reported using AI for targeted tasks but did not consider it integral to their overall learning process. A few respondents had only recently begun using AI and, although they acknowledged its potential, did not yet view it as indispensable. For example, one participant explained:

> *I only started using AI recently and I would not necessarily say it is essential component but rather an assistant to my learning...It is more of a helper than essential at this point in my journey.* [Participant 25]

Theme 1.2: Motivations for Adopting AI Tools

Curiosity was a key factor. Some respondents were intrigued by the widespread debate and excitement surrounding AI, particularly ChatGPT. Often encouraged by others, they were motivated to explore AI by the desire to see how it could be useful in their academic and professional lives.

A few respondents (n=2) were early adopters and experimenting with AI from its initial stages. One began exploring ChatGPT at its initial public launch, driven by curiosity and social media exposure:

> *Honestly, when I started experimenting with AI tools, especially ChatGPT, it was the first one I used, I think it was really in December 2022, during its first days of public access...it was out of fear, out of curiosity and I had heard about it...* [Participant 1]

The influence of peers, colleagues, and professors was also significant for 15 respondents, who were encouraged to try AI after observing others using it or receiving recommendations. This indicates the role of social and professional networks in the adoption of new technologies. As one student noted:

> *My friend suggested that I try it out for help.* [Participant 48]

XSL•FO
RenderX

Another shared:

> *My supervisor actually recommended it for my data analysis; they told me to use it when I am unsure of how to use Excel or SPSS to work with data.* [Participant 35]

Practical considerations, such as improving efficiency, saving time, and enhancing convenience, motivated 14 respondents to adopt AI tools. They found AI helpful in streamlining tasks and making their work processes more manageable. One participant noted:

> *The fact that it is an accessible tool and allows me to save time.* [Participant 14]

Twelve respondents were also motivated to use AI for professional or academic needs, frequently to gain a better understanding of procedures or expectations before starting a task. Some students used it to make sense of academic processes, particularly while navigating an unfamiliar system. One respondent stated:

> *I am an international student, and I was not familiar with many academic based process and AI could make them clear for me.* [Participant 29]

Some respondents (n=5) turned to AI out of frustration with traditional search engines like Google, which they found unhelpful or overwhelming. They saw AI as a more responsive and accurate alternative. For example, one participant noted:

> *The answers/explanations that Google offered me were vague and took longer to find.* [Participant 7]

And another mentioned:

> *When Google wasn't giving me the answers I needed or was overwhelming me and confusing me instead of helping.* [Participant 43]

A few respondents were motivated by innovation, recognizing AI's cutting-edge potential to make their work easier and more unique. One student explained:

> *Innovation, being different and unique. Having my work done easier for me. Being ahead of the curve and being more innovative and productive.* [Participant 49]

Theme 1.3: AI Tools for Knowledge Acquisition and Academic Skills Development

ChatGPT is the predominant AI tool used by health professions students for learning, mostly for the development of academic skills, such as refining data analysis, understanding complex topics, and enhancing writing abilities. For example, one participant mentioned:

> *ChatGPT is good at helping me refine my data analysis. It helps me distinguish between certain nuances of data analysis etc. when I ask it questions.* [Participant 25]

And another highlighted the importance of writing:

> *I think it helps me to develop new skills in the writing process, including synthesizing information or teaching me how to write in a more concise manner.* [Participant 28]

However, some respondents reported not using AI tools for learning, with a few indicating unfamiliarity with these technologies, falling into the "none or N/A" category.

Beyond ChatGPT, other AI tools were also mentioned for their specific functionalities. Wolfram was praised for its mathematics capabilities, text-to-speech models for aiding study through text-to-speech, and Consensus for summarizing research papers. One student stated, highlighting the tool's role in academic research:

> *...I also like how Research Rabbit (I think it's AI?) helps find links from one paper to others that are similar, it makes literature searching much easier.* [Participant 23]

Finally, language learning proved important for some, as these tools helped improve grammar and clarity. As one student noted:

> *When AI helps me rephrase my sentences to make them clearer, it allows me to acquire new knowledge or skills on the grammatical side.* [Participant 12]

Theme 1.4: Factors Influencing the Choice of AI Tools

Most health professions students did not explore multiple AI tools before choosing their current ones. Only 6 out of 51 respondents experimented with different tools. However, 23 respondents shared insights into their choice, highlighting several key factors.

Ease of access was a significant factor, with 10 respondents preferring tools that are free or provided by educational institutions. These tools were chosen for their simplicity and minimal setup requirements:

> *I have free open access via my institution.* [Participant 24]

> *It is simple to use. I would be open to trying other tools, but when the tools require me to set up a whole account, pay, or use personal data to use them, I find this to be a deterrent.* [Participant 28]

Specific features also influenced the choice, with 7 respondents valuing functionalities such as summarizing scientific papers or handling documents:

> *I chose Consensus for literature synthesizing because it presented the information in the best manner and is set up to aid in research specifically.* [Participant 23]

One participant characterized the tool's benefits as "bilingual" [Participant 16], in addition to being free, rapid, and efficient, while also emphasizing the necessity of using the tool judiciously.

Familiarity played a role for 6 respondents, who chose tools they were already aware of, highlighting a preference for sticking with known options. One participant mentioned:

> *I started with ChatGPT as that is the one I kept hearing about. It's suited my needs thus far as I don't*

XSL•FO
RenderX

> *use it very often. If I hear of something better, I might explore that too.* [Participant 19]

Popularity influenced 5 respondents, who opted for widely recognized tools, believing them to be reliable. As one student noted:

> *ChatGPT seemed the most popular.* [Participant 37]

Finally, ease of use was crucial for 5 respondents, who favored tools that were user-friendly. Less commonly mentioned factors included perceived reliability, speed, and efficacy, underscoring the value of dependable and efficient tools. One student explained:

> *I chose ChatGPT because it seems the most reliable and I have had no reason to seek other tools.* [Participant 34]

Theme 2: Usage Habits and Applications of AI

Theme 2.1: Students' AI Tools Usage Habits

Health professions students showed varied practices regarding the frequency and contexts of AI tools use, focusing on educational support, writing assistance, and information retrieval. Many respondents (n=19) use AI to aid in understanding complex concepts, studying, and verifying knowledge, which indicates how these tools can enhance learning. As one student explained:

> *I use them when I don't understand homework questions or need to double check an answer.* [Participant 33]

Writing assistance is another significant use, with 13 respondents relying on AI for drafting, revising, and brainstorming ideas. AI is used not only in academic writing but also in professional communication, such as writing emails. As one participant noted:

> *For each of my courses that I undertake and also alongside my professional life, whether it's for writing emails or other things.* [Participant 2]

This highlights the versatility of AI in supporting various writing tasks. Some respondents, however, use AI infrequently or lack clear usage habits. Nearly 11 respondents expressed uncertainty about their AI usage, suggesting they are still exploring the tool's potential or have not fully integrated it into their routines. One student shared:

> *Very infrequently used so I do not have habits for use.* [Participant 48]

AI is also used for general information retrieval and research, with 7 respondents using it for quick access to relevant information. As mentioned by one participant:

> *I use it as a shortcut every time I do research or need information.* [Participant 14]

A small number of respondents use AI for specific tasks like coding or data management, demonstrating AI's value in specialized technical problem-solving. For instance, one student mentioned:

> *When I am really stuck on how to do something in data analysis software, I ask ChatGPT for the best way.* [Participant 35]

In terms of frequency, some respondents reported using AI daily, while others mentioned using it 1 to 3 times per week. The consistency of usage varied as well, with several describing their use as regular and others characterizing it as more occasional. This highlights the flexible nature of AI tools in meeting diverse needs.

Theme 2.2: Main Uses of AI Tools

Participants identified several common applications of AI tools, including in revision, writing, learning, research, and knowledge testing.

Revision and synthesis emerged as the most common use, with 28 respondents using AI to refine and organize information, edit text, and summarize complex material. AI is frequently used to structure ideas and enhance communication, including drafting emails and quizzing on notes. One student, for example, summarized their use as follows:

> *...revising notes, synthetizing information.* [Participant 41]

Writing is another significant area, with 22 respondents using AI for drafting, idea generation, and refining writing style. AI helps tailor content to specific tones and styles, enhancing both the creative and revision stages of writing. As one student mentioned:

> *I mostly use these tools to proofread my writing or to enhance my writing style when I'm hit with "writer's block."* [Participant 19]

For learning, 16 respondents rely on AI to clarify complex concepts, gain new knowledge, and validate their own reasoning. AI aids in understanding difficult topics and applying knowledge in practical scenarios, such as informing clinical practices in psychology. One psychology student noted:

> *...I also use them to verify the clarity of my own writing. If I have an intervention in mind for a client (I am in psychology), I ask what research recommends for the name of whichever diagnosis I want to treatment plan for. I then double check that it aligns with what I read...* [Participant 22]

In research, 8 respondents use AI to find specific answers, summarize articles, and navigate academic language. AI is valued for making research more accessible by bridging the gap between lay and academic terminology. One student explained:

> *It is very useful to research topics when you don't know the exact language used in the literature, you can ask question in more common language and get to the information about the names of the theories and components that allows you to do real research to find the actual research papers relevant.* [Participant 24]

Finally, knowledge testing involves using AI to create quizzes, check answers, and master complex tasks, such as data analysis in SPSS (IBM Corp):

XSL•FO
RenderX

> *I use it to ask how to analyse data in a more complicated way in SPSS. My education on SPSS was simple and was able to guide me to do simple tasks, but my research is a bit more complicated.* [Participant 35]

Theme 3: AI in Learning and Skills Development

Theme 3.1: Critical Reasoning in Assessing AI-Generated Information

According to the survey analysis, students pursuing health professions use a range of critical thinking techniques to evaluate information produced by AI, indicating varying degrees of confidence and reliance on these technologies.

Cross-referencing AI outputs with credible sources is the most popular strategy, as reported by 29 respondents. They often compare AI-generated content with peer-reviewed articles, course materials, or trusted resources to verify accuracy, demonstrating a cautious approach to ensure that the AI information aligns with established knowledge. As one respondent noted:

> *I compare it to peer-reviewed articles to see if the findings match up.* [Participant 51]

Another prevalent strategy, mentioned by 16 respondents, involves using personal knowledge and common sense to evaluate AI content. These users rely on their prior understanding and logical reasoning to judge the validity of the information. For example, one respondent stated:

> *I like to see whether it makes sense and matches up with notes taken in class and information I already have and know.* [Participant 45]

Reviewing and editing AI-generated content is also common, with 17 respondents indicating that they scrutinize and revise AI outputs to ensure alignment with their expectations. This active engagement helps refine the content before it is used. As mentioned by one participant:

> *I always read it over, take out things that don't make sense or I don't agree with, and then rewrite it myself.* [Participant 46]

Theme 3.2: AI Tools in the Learning Process

AI tools play a nuanced role in the learning process, supporting dynamic engagement with knowledge through social, cognitive, and affective dimensions to develop cognitive structures.

Perception and Interaction With Information

Nearly 12 respondents noted that AI tools aid in recognizing and understanding information, enhancing the initial phase of learning. AI also facilitates interaction with data, with respondents emphasizing its role in manipulating and engaging with information crucial for deepening understanding. One student explained:

> *I believe AI tools help to perceive objects from a different perspective, in order to better understand the material.* [Participant 18]

Integration and Cognitive Development

A total of 12 respondents asserted that AI supports learners in embedding new skills and knowledge into their broader contexts, helping them integrate what they learn into their daily lives. Additionally, AI tools assist in transforming and refining cognitive structures, playing a key role in advancing cognitive development through learning processes. As one participant explained:

> *I would say that AI tools allow to transform, to create, and to develop the cognitive structure. However, I would not say that AI tools are, in themselves, information or knowledge. They should rather be used to guide or confirm our information and our knowledge.* [Participant 16]

Skills Development

AI's impact on skill acquisition was acknowledged, though its influence on attitudes and values was considered minimal. Four respondents generally viewed AI as more effective in enhancing cognitive and skill-based learning rather than shaping affective outcomes. As one respondent noted:

> *AI fits into the knowledge and information objects of learning. AI cannot teach new attitudes or values, but it can present new information and knowledge, as well as new ways to learn skills. I don't think it is advanced enough to help individuals perceive and interact with objects.* [Participant 36]

Information and Knowledge Acquisition

The most frequently mentioned theme was AI's role in acquiring and refining information and knowledge. A total of 23 respondents identified AI as a critical facilitator in these areas, underscoring its importance in the learning process. One student mentioned:

> *I think AI fits in terms of knowledge acquisition and information gathering, both of which can help with learning.* [Participant 48]

Theme 3.3: AI Tools and the Experiential Learning Process

In our survey of health professions students on AI tools in education, we examined how these tools align with phases of experiential learning: concrete experience, reflective observation, abstract conceptualization, and active experimentation.

Concrete Experience

Overall, 13 respondents identified AI tools as fitting within this phase, where learners interact with new or known situations to gain direct experiences. Highlighting their role in engaging directly with problems, one comment mentioned:

> *AI tools can be a great way to get experience and learn for example, if I am stuck in finding a code in R that gives me the statistical analysis I need. I can ask ChatGPT and troubleshoot with the code it gives me, study the code and integrate it into my bank of knowledge for next time.* [Participant 27]

XSL•FO
RenderX

Reflective Observation

A total of 14 responses placed AI tools in this phase, where learners analyze and evaluate their experiences. Participants noted that AI helps in reflecting on and assessing outcomes, with one student mentioning:

> *Reflective observation, which consists of analyzing and evaluating the experience.* [Participant 44]

Another shared:

> *AI tools allow us to analyse data, simulate scenarios, and create links with known information.* [Participant 43]

Abstract Conceptualization

This was the most cited phase, with 21 responses. AI tools were recognized for aiding in the development of theories from observations. One respondent remarked, underscoring their role in synthesizing and conceptualizing new ideas:

> *I use AI tools to understand things, to learn from existing theories, and to develop concepts to integrate the topics.* [Participant 14]

Active Experimentation

Nearly 19 respondents associated AI tools with this phase, where concepts are tested in new situations. AI facilitates this by enabling users to apply and test theoretical concepts practically. One student noted:

> *Active experimentation when testing what AI can do and its limits.* [Participant 50]

AI Does Not Fit Well Into Kolb's Model

A minority of responses (n=3) expressed skepticism about AI's fit within Kolb's phases, citing challenges in aligning AI functionalities with the model: "uncertain/did not understand the question." Two respondents were unsure or confused about how to position AI tools within Kolb's model.

Theme 3.4: Skills Required for Effective AI Use in Learning

Several essential skills were identified by the respondents for using AI tools effectively in learning. Technical competency is essential, as 22 respondents highlighted the necessity of comprehending the operation of AI systems, particularly the proper navigation of their features and settings to enhance learning outcomes.

Critical thinking and evaluation skills were also highlighted by 19 respondents, who emphasized the importance of analyzing AI outputs and ensuring they align with established knowledge or academic standards. This is particularly important given the potential for AI to produce inaccurate or incomplete information. One participant explained:

> *A variety of skills are needed, critical thinking and good judgement being very important as certain outputs generated by AI can be incorrect or misleading and knowing when to distinguish that is important.* [Participant 51]

And another participant stated:

> *Critical thinking and judgment are essential to determine whether the AI's response is appropriate in each situation. We must be able to detect biases (what the AI might have forgotten to consider, such as emotions, etc.).* [Participant 12]

Adaptability and flexibility are key, according to 14 respondents, who noted that users must be open to experimenting with AI and adjusting their strategies as they learn. One respondent stated:

> *...Flexibility is also important because you might need to make changes to your prompt to get a desired response from AI.* [Participant 36]

Effective communication was another important skill, with 12 respondents stressing the need to ask AI clear and specific questions to receive the most useful responses. As one respondent pointed out:

> *...knowing how to word questions properly to get the answer you need.* [Participant 27]

Finally, digital literacy was mentioned by 10 respondents, who recognized not only the importance of staying up to date with AI tools but also the need to use them responsibly and with academic integrity. As one student mentioned:

> *Being responsible when processing and using information. Maintaining academic integrity.* [Participant 2]

Theme 4: Peer Training and Awareness Strategies

Theme 4.1: Methods Used to Learn AI Tools

Health professions students used various methods to learn how to use AI tools. Self-directed exploration was the most common approach, with 23 respondents reporting that they learned by trying out different features and experimenting with the tools independently. One respondent said, which reflects a hands-on approach to mastering AI:

> *I just used trial and error.* [Participant 38]

Social learning also played a role, with 13 respondents indicating that they learned through friends, colleagues, or professors. These students benefited from observing others use AI or receiving direct guidance. This highlights the importance of peer-to-peer learning and mentorship in mastering new technologies. One student noted:

> *I heard about the tools from some of my friends and asked them for tips.* [Participant 33]

Online tutorials and guides were also used by 11 respondents, who turned to YouTube videos, articles, or AI tool help sections to gain a better understanding. As one student shared:

> *...I also saw a couple videos on TikTok about using select sentences/words for my desired effect.* [Participant 34]

Formal training, however, was mentioned by only two respondents, indicating that most AI learning happens informally and independently. One student explained:

> *I didn't need to learn anything, ChatGPT was easy to use.* [Participant 9]

XSL•FO
RenderX

Overall, students adopted diverse strategies for learning use of AI tools, with a clear preference for self-guided exploration and peer learning, supplemented by online resources.

Theme 4.2: Best Strategies for Educating and Training Peers in AI Use

To effectively educate and train peers in using AI tools, several strategies were highlighted by the respondents.

Practical, hands-on demonstrations were the most suggested approach, with 26 respondents emphasizing the importance of showing how the tools work in real time. For example, one respondent suggested:

> *I think concrete exercises would be more useful. For example, using it as part of a specific group project.* [Participant 15]

Workshops and guided sessions were also popular recommendations, with 19 respondents advocating for structured opportunities to learn AI in a collaborative environment. These sessions would allow peers to ask questions, practice using the tools, and receive feedback. As mentioned by one student:

> *Workshops and small group discussion sessions are the best.* [Participant 29]

And another elaborated:

> *I think workshops could be helpful as it allows for discussion on how to use these AI tools properly and effectively. It also allows for troubleshooting and working together to solve a problem in AI tools.* [Participant 27]

Providing easy-to-follow tutorials and online resources was another suggested strategy, with 15 respondents supporting the use of videos, articles, or step-by-step guides that could be shared among peers. One respondent explained:

> *Self-study modules that students can do at their own pace and learn new ways that can help use AI tools in a smart yet efficient way, without harming our education.* [Participant 45]

This method would allow for self-paced learning and continuous reference.

Lastly, creating supportive peer networks was recommended by 10 respondents, who emphasized the value of mentorship and peer-to-peer assistance in AI education. This strategy builds a community of learners who can share insights, troubleshoot issues, and offer mutual encouragement. One student stated:

> *I think the best strategy is through discussion/exposure by peers/colleagues...This is most appropriate because the discussion aspect primes you to begin thinking about it, so when you approach it individually, you can play around with different aspects based on what you need/want to pursue.* [Participant 23]

Discussion

Principal Findings

Our study collected data from 51 health science students through semistructured interviews and a questionnaire. The findings show the growing adoption of AI tools by students and underscore their importance in academic and clinical settings. Participants reported that general-purpose AI tools, including platforms like ChatGPT, were frequently used. The degree of AI tool use varied depending on the tasks students were required to perform, with some integrating them into their daily activities while others rarely used them. According to participants, AI tools are beneficial for varied learning activities, including rapid information comprehension, facilitating the generation of diverse perspectives, and simulation of varied clinical situations. Students also highlighted the need for collaborative strategies, such as peer learning groups, to promote and ensure effective AI use.

Students identified AI tools as essential to their learning activities. ChatGPT was the most cited tool, while other tools were mentioned only occasionally. This trend reflects the growing use of ChatGPT in health care students' learning activities, which could be explained by its accessibility, personalized responses, and ability to simulate engaging cognitive interactions. From the perspective of Kolb's model, this frequent use indicates a significant footprint in the concrete experience phase, while the use of AI to clarify, structure, and reorganize knowledge corresponds to the abstract conceptualization phase. These results are in line with a previous study, which highlighted the ability of ChatGPT to produce logical and contextually relevant explanations, as well as to simulate a logical thinking process, making it a promising didactic tool in the field of education [8,30]

Furthermore, students expressed a preference for using AI tools over traditional search engines like Google Search, due to the more targeted and precise nature of the results. This suggests that students are looking for tools that support conceptual organization and meaning-making, which corresponds to the abstract conceptualization phase of Kolb's experiential learning cycle. Previous studies confirmed this preference among students for ChatGPT over Google Search, as this tool appears to provide more accurate and comprehensive medical information for general knowledge requests and highlights its usefulness in the educational environment [8,31].

Students reported variability in their use of AI tools, ranging from occasional use to regular integration into their learning routines. This observed inconsistency appears to reflect significant differences in individual usage styles, the familiarity developed with AI technologies, and the nature of the tasks they were faced with. In the context of experiential learning, this variability may also indicate that students do not all engage in the learning cycle in the same way, with some focusing primarily on concrete experience and reflective observation, while others broaden their use to abstract conceptualization and active experimentation by reinvesting AI results in new academic or clinical tasks. These results are in line with recent studies confirming that the degree of use of AI tools depends on their

XSL•FO
RenderX

usefulness as well as specific academic contexts and tasks [32,33]. Understanding this nuance is important because it suggests the importance of promoting their use in an educational context by adopting a more individual rather than a universal approach [9,33].

Our results also confirmed that a lot of health care students use ChatGPT to learn and develop new skills, such as improving their writing, analyzing data, and simplifying complex ideas. These practices are part of Kolb's experiential learning theory and are explained by reflective observation of AI-generated results and abstract conceptualization. These findings are consistent with previous studies, which concluded that ChatGPT improves student writing, leading to more effective and higher-quality academic production [10,34]. Moreover, other more specialized tools used by these students, such as Consensus for synthesizing scientific articles, demonstrate a strategic and targeted use of AI tools to perform specific tasks, aligning with previous studies [35,36]. On the other hand, some students reported not using these technologies, often due to lack of interest or confidence, which is consistent with the writings of Johnston et al [37], who show that students who have greater confidence in their ability to write academic texts use ChatGPT or other generative AI tools less in their learning activities. These results show that efforts are still needed to promote equitable and critical integration of these tools into educational practices, although AI remains an important tool for skill development [37].

In terms of effective strategies for training peers to use AI tools, students identified hands-on demonstrations, guided workshops, and self-directed resources. These findings are part of Kolb's experiential learning theory and are explained by the importance of concrete experience (hands-on demonstrations), reflective observation (guided discussions and feedback), and active experimentation (autonomous reuse of tools in personal learning contexts). These approaches are consistent with the study of Jiang et al [38], which highlights the importance of active and collaborative learning to foster engagement and understanding in education, as well as with that of Abid et al [39], which demonstrates the effectiveness of flexible and online approaches in technology training. Participants also valued peer support, highlighting the central role of social collaboration and peer support in technology interactions, as suggested by previous studies [14,40]. These findings confirm that effective AI adoption requires a combination of hands-on activities, autonomy, and collaborative support.

Strengths

One of the strengths of this study is its methodological framework, which facilitated the collection of rich and nuanced opinions from health science students on the use of AI. Next, the study includes individuals with varied experiences using AI tools, both in academic and clinical environments. This highlights not only the different ways in which AI is used, but also the motivations and critical thinking strategies implemented by students. Finally, the highlighted didactic implications are directly applicable. They emphasize the need to prepare students for a critical approach to AI, to encourage tailored learning environments, and to promote the incorporation of AI into collaborative tasks.

Limitations

One of the limitations of this study is the small sample size and the use of self-reported data, and the fact that the data were collected within a single establishment, which limits the generalization of the results to other institutions. This may have introduced biases related to memorization accuracy or social desirability. Although the research team assessed the questionnaire for relevance and clarity, psychometric procedures were not sufficiently applied to assess it. Therefore, interpretation biases may arise, particularly regarding the understanding of porous concepts such as "learning" or "learning support." Furthermore, health professions students were the focus of this study; the application of AI in training may vary by discipline. In addition, this transition in data collection, from semistructured interviews to online surveys, may have reduced the depth and richness of some responses, which could affect the identification of more subtle themes.

Therefore, future research could address these limitations by conducting larger-scale studies with more diverse samples and using objective measures of AI usage, such as usage logs or performance data. Additionally, future research could explore how AI is used in other fields of study and whether the patterns observed in this study are generalizable to different educational contexts. Moreover, as AI tools continue to evolve, it will be important to track how students' perceptions and utilization habits change over time.

Conclusions

In conclusion, this study aimed to identify the AI tools most used by health students at the University of Ottawa, to determine their usage habits as well as the tools helping them develop skills, and to examine the strategies they consider effective for training their peers in the use of these tools. Our findings highlight the growing role of AI in health care education and provide valuable insights into how these tools have been adopted and used by students. While AI is considered essential to their learning activities, it becomes crucial to further integrate it into educational programs and help students develop their digital skills while developing their critical thinking skills. As these technologies gain popularity, they have the potential to enhance the learning experience of health care students, but careful attention must be paid to how these tools are promoted and used in the educational environment.

Acknowledgments

We would like to thank the "Affaires francophones" of the Faculty of Medicine, University of Ottawa, for supporting this research. We also acknowledge the use of AI tools during the research and writing process. Microsoft Copilot 365 was used to support the writing process under human supervision, more specifically to correct and improve sentence structure and phrasing, and not to

XSL•FO
RenderX

generate content. To help with the search for relevant scientific literature, ChatGPT (OpenAI) and Elicit (Ought) were used. All AI-generated content was reviewed and edited by the authors, who remain fully responsible for the originality, accuracy, and integrity of the final work.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The data generated and analyzed in this study are not publicly available due to institutional ethics and confidentiality regulations. However, they may be obtained from the corresponding author upon reasonable request, subject to ethics committee approval.

Authors' Contributions

All authors contributed equally to the conception, design, data collection, analysis, and writing of this manuscript. Each author reviewed and approved the final version for submission and agrees to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Question guide and distribution of artificial intelligence tools used by health professions students at the University of Ottawa.
[PDF File, 161 KB - jmir_v28i1e82432_app1.pdf]

Checklist 1
SRQR checklist.
[PDF File, 124 KB - jmir_v28i1e82432_app2.pdf]

References

1. Rahman MA, Victoros E, Ernest J, Davis R, Shanjana Y, Islam MR. Impact of artificial intelligence (AI) technology in healthcare sector: a critical evaluation of both sides of the coin. Clin Pathol 2024;17. [doi: 10.1177/2632010X241226887] [Medline: 38264676]
2. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J 2021 Jul;8(2):e188-e194. [doi: 10.7861/fhj.2021-0095] [Medline: 34286183]
3. Zavaleta-Monestel E, Anchía-Alfaro A, Rojas-Chinchilla C, Quesada-Loria DF, Arguedas-Chacón S. Ethical and practical dimensions of artificial intelligence (AI) in healthcare: a comprehensive study of professional perceptions. Cureus 2025 Feb;17(2):e78416. [doi: 10.7759/cureus.78416] [Medline: 40046390]
4. Najjar R. Redefining radiology: a review of artificial intelligence integration in medical imaging. Diagnostics (Basel) 2023 Aug 25;13(17):2760. [doi: 10.3390/diagnostics13172760] [Medline: 37685300]
5. Fujisawa Y, Otomo Y, Ogata Y, et al. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. Br J Dermatol 2019 Feb;180(2):373-381. [doi: 10.1111/bjd.16924] [Medline: 29953582]
6. Al Kuwaiti A, Nazer K, Al-Reedy A, et al. A review of the role of artificial intelligence in healthcare. J Pers Med 2023 Jun 5;13(6):951. [doi: 10.3390/jpm13060951] [Medline: 37373940]
7. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023 Mar 6;9(1):e46885. [doi: 10.2196/46885] [Medline: 36863937]
8. Ghorashi N, Ismail A, Ghosh P, Sidawy A, Javan R. AI-powered chatbots in medical education: potential applications and implications. Cureus 2023 Aug;15(8):e43271. [doi: 10.7759/cureus.43271] [Medline: 37692629]
9. Teng M, Singla R, Yau O, et al. Health care students' perspectives on artificial intelligence: countrywide survey in Canada. JMIR Med Educ 2022 Jan 31;8(1):e33390. [doi: 10.2196/33390] [Medline: 35099397]
10. Mousavi Baigi SF, Sarbaz M, Ghaddaripouri K, Ghaddaripouri M, Mousavi AS, Kimiafar K. Attitudes, knowledge, and skills towards artificial intelligence among healthcare students: a systematic review. Health Sci Rep 2023 Mar;6(3):e1138. [doi: 10.1002/hsr2.1138] [Medline: 36923372]
11. Schneider M, Stern E. L'apprentissage dans une perspective cognitive [in French]. In: Dumont H, Istance D, Benavides F, editors. Comment apprend-on ? La recherche au service de la pratique: OCDE; 2010:73-95.

XSL•FO
RenderX

12. Hsu WC, Fuh LJ, Liao SC. Tickling the heart: integrating social emotional learning into medical education to cultivate empathetic, resilient, and holistically developed physicians. Front Med (Lausanne) 2024;11:1368858. [doi: 10.3389/fmed.2024.1368858] [Medline: 38500950]
13. Shorey S, Chan V, Rajendran P, Ang E. Learning styles, preferences and needs of generation Z healthcare students: scoping review. Nurse Educ Pract 2021 Nov;57:103247. [doi: 10.1016/j.nepr.2021.103247] [Medline: 34768214]
14. Joseph GV, Athira P, Anit Thomas M, Jose D, Roy TV, Prasad M. Impact of digital literacy, use of AI tools and peer collaboration on AI assisted learning: perceptions of the university students. Digital Edu Rev 2024(45):43-49. [doi: 10.1344/der.2024.45.43-49]
15. Msambwa MM, Wen Z, Daniel K. The impact of AI on the personal and collaborative learning environments in higher education. Eur J Edu 2025;60(1):e12909. [doi: 10.1111/ejed.12909]
16. Fawaz M, El-Malti W, Alreshidi SM, Kavuran E. Exploring health sciences students' perspectives on using generative artificial intelligence in higher education: a qualitative study. Nurs Health Sci 2025 Mar;27(1):e70030. [doi: 10.1111/nhs.70030] [Medline: 39797529]
17. Benboujja F, Hartnick E, Zablah E, et al. Overcoming language barriers in pediatric care: a multilingual, AI-driven curriculum for global healthcare education. Front Public Health 2024;12:1337395. [doi: 10.3389/fpubh.2024.1337395] [Medline: 38454985]
18. Busetto L, Wick W, Gumbinger C. How to use and assess qualitative research methods. Neurol Res Pract 2020;2(1):14. [doi: 10.1186/s42466-020-00059-z] [Medline: 33324920]
19. Hunt MR. Strengths and challenges in the use of interpretive description: reflections arising from a study of the moral experience of health professionals in humanitarian work. Qual Health Res 2009 Sep;19(9):1284-1292. [doi: 10.1177/1049732309344612] [Medline: 19690208]
20. Locsin RC, Schoenhofer SO. Qualitative research methods, inductive and deductive: valuable approaches to nursing knowledge development for practice. Nurs Health Sci 2024 Jun;26(2):e13128. [doi: 10.1111/nhs.13128] [Medline: 38804111]
21. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for Reporting Qualitative Research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
22. Murphy E, Dingwall R, Greatbatch D, Parker S, Watson P. Qualitative research methods in health technology assessment: a review of the literature. Health Technol Assess 1998;2(16). [Medline: 9919458]
23. Higginbottom GMA. Sampling issues in qualitative research. Nurse Res 2004;12(1):7-19. [doi: 10.7748/nr2004.07.12.1.7.c5927] [Medline: 15493211]
24. Guest G, Bunce A, Johnson L. How many interviews are enough? An experiment with data saturation and variability. Field methods 2006;18(1):59-82. [doi: 10.1177/1525822X05279903]
25. Schneider S, Beege M, Nebel S, Schnaubert L, Rey GD. The cognitive-affective-social theory of learning in digital environments (CASTLE). Educ Psychol Rev 2022;34(1):1-38. [doi: 10.1007/s10648-021-09626-5] [Medline: 34226808]
26. Adams NE. Bloom's taxonomy of cognitive learning objectives. J Med Libr Assoc 2015 Jul;103(3):152-153. [doi: 10.3163/1536-5050.103.3.010] [Medline: 26213509]
27. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development: Prentice Hall; 1983.
28. Paillé P, Mucchielli A. L'analyse par questionnement analytique [in French]. In: L'analyse qualitative en sciences humaines et sociales: Armand Colin; 2012:207-230.
29. Dhakal K. NVivo. J Med Libr Assoc 2022 Apr 1;110(2):270-272. [doi: 10.5195/jmla.2022.1271] [Medline: 35440911]
30. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
31. Ayoub NF, Lee YJ, Grimm D, Divi V. Head-to-head comparison of ChatGPT versus Google Search for medical knowledge acquisition. Otolaryngol Head Neck Surg 2024 Jun;170(6):1484-1491. [doi: 10.1002/ohn.465] [Medline: 37529853]
32. Saleh ZT, Rababa M, Elshatarat RA, et al. Exploring faculty perceptions and concerns regarding artificial intelligence chatbots in nursing education: potential benefits and limitations. BMC Nurs 2025 Apr 18;24(1):440. [doi: 10.1186/s12912-025-03082-0] [Medline: 40251532]
33. Alqaisi N, Alshwayyat S, Aburumman S, et al. Assessing ChatGPT adoption in Jordanian medical education: a UTAUT model approach. BMC Med Educ 2025 May 22;25(1):750. [doi: 10.1186/s12909-025-07336-y] [Medline: 40405195]
34. Wang J, Liao Y, Liu S, et al. The impact of using ChatGPT on academic writing among medical undergraduates. Ann Med 2024 Dec;56(1):2426760. [doi: 10.1080/07853890.2024.2426760] [Medline: 39555617]
35. Fabiano N, Gupta A, Bhambra N, et al. How to optimize the systematic review process using AI tools. JCPP Adv 2024 Jun;4(2):e12234. [doi: 10.1002/jcv2.12234] [Medline: 38827982]
36. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial intelligence in medical education: comparative analysis of ChatGPT, Bing, and medical students in Germany. JMIR Med Educ 2023 Sep 4;9:e46482. [doi: 10.2196/46482] [Medline: 37665620]
37. Johnston H, Wells RF, Shanks EM, Boey T, Parsons BN. Student perspectives on the use of generative artificial intelligence technologies in higher education. Int J Educ Integr 2024 Dec;20(1):1-21. [doi: 10.1007/s40979-024-00149-4]

XSL•FO
RenderX

38. Jiang J, Huang K, Martinez-Maldonado R, Zeng H, Gong D, An P. Novobo: supporting teachers' peer learning of instructional gestures by teaching a mentee AI-agent together. arXiv. Preprint posted online on May 26, 2025. [doi: 10.48550/arXiv.2505.17557]
39. Abid A, Murugan A, Banerjee I, Purkayastha S, Trivedi H, Gichoya J. AI education for fourth-year medical students: two-year experience of a web-based, self-guided curriculum and mixed methods study. JMIR Med Educ 2024 Feb 20;10:e46500. [doi: 10.2196/46500] [Medline: 38376896]
40. Tao W, Yang J, Qu X. Utilization of, perceptions on, and intention to use AI chatbots among medical students in China: national cross-sectional study. JMIR Med Educ 2024 Oct 28;10:e57132. [doi: 10.2196/57132] [Medline: 39466038]

ABBREVIATIONS

AI: artificial intelligence
REB: Research Ethics Board
SRQR: Standards for Reporting Qualitative Research

Edited by A Stone; submitted 14.Aug.2025; peer-reviewed by A Jaleel, F Fukuzawa; revised version received 26.Feb.2026; accepted 27.Feb.2026; published 02.Apr.2026.

Please cite as:
Guirguis M, Fotsing S, Fevry J, Landry C, Bouchard-Lamothe D, Lacroix J, Jalali A
Artificial Intelligence in Health Professions Education: Qualitative Study of Student Experiences
J Med Internet Res 2026;28:e82432
URL: https://www.jmir.org/2026/1/e82432
doi:10.2196/82432

© Marina Guirguis, Salomon Fotsing, Joanne Fevry, Christine Landry, Diane Bouchard-Lamothe, Jennifer Lacroix, Alireza Jalali. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Initial Insights Into an Institutional Secure Large Language Model for Magnetic Resonance Imaging Examination Requests: Retrospective Study

James Thomas Patrick Decourcy Hallinan[1,2], MBChB; Naomi Wenxin Leow[3], BComp, MComp; Yi Xian Low[1], MBBS; Aric Lee[1], MBBS; Wilson Ong[1], MBBS; Matthew Ding Zhou Chan[1], BmedSc, MD; Ganakirthana Kalpenya Devi[1], MbBchBao; Stephanie Shengjie He[1], MBBS; Daniel De-Liang Loh[1], MBBS, MRCS; Desmond Shi Wei Lim[1], MBBS; Xi Zhen Low[1], MBBS; Mei Chin Lim[1], MBBS; Clement Yong[1], MBBS; Weizhong Jonathan Sng[1], MBBS; Ee Chin Teo[1], MMRT; Jiong Hao Tan[4], MBBS, MMed, MRCS; Naresh Kumar[4], MBBS, MS, DNB, DM; Andrew Makmur[1,3], BmedSc, MBBS, MMed; Yonghan Ting[1], MBBS

[1]Department of Diagnostic Imaging, National University Hospital, 5 Lower Kent Ridge Rd, Singapore, Singapore

[2]Department of Diagnostic Radiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[3]Innovation Office, AI Office, National University Health System, Singapore, Singapore

[4]National University Spine Institute, Department of Orthopedic Surgery, National University Hospital, Singapore, Singapore

Corresponding Author:
James Thomas Patrick Decourcy Hallinan, MBChB
Department of Diagnostic Imaging, National University Hospital, 5 Lower Kent Ridge Rd, Singapore, Singapore

Abstract

Background: Incomplete clinical details on magnetic resonance imaging (MRI) examination requests (MERs) can lead to suboptimal protocol selection. An institutional secure large language model (sLLM) with access to manually retrieved salient data from the electronic medical record (EMR) may improve request completeness and protocol accuracy across multiple MRI subspecialties.

Objective: The objective of this study was to compare clinician MERs with sLLM-augmented MERs for information quality and to evaluate the protocoling accuracy of the sLLM versus board-certified radiologists across body, musculoskeletal, and neuroradiology MRI.

Methods: This retrospective study included 608 random outpatient MRI examinations performed between September 2023 and July 2024 (body 206, musculoskeletal 203, neuroradiology 199). The cohort comprised 528 patients (mean 51.2 years, SD 19.2; range 4‐93; n=279, 52.8% women, n=249, 47.2% men). MERs without EMR access were excluded. A privately hosted Anthropic Claude 3.5 model (temperature 0) augmented each MER with manually retrieved salient EMR data and, via rule-based parsing, mapped the extracted elements onto predefined institutional criteria to recommend region or coverage and contrast use. Two experienced radiologists established a consensus reference standard. Two board-certified general radiologists (Rad 3 and Rad 4) and the sLLM were compared with this standard. Clinical information quality was graded using the Reason-for-Exam Imaging Reporting and Data System (RI-RADS). Interrater reliability was quantified with Gwet AC1. Paired accuracies were compared with the McNemar test to determine whether there was a statistically significant difference.

Results: Interreader agreement for RI-RADS was almost perfect for sLLM-augmented MERs (AC1 0.97, 95% CI 0.94‐0.99) and moderate for clinician MERs (AC1 0.43, 95% CI 0.34‐0.52). Limited or deficient clinical information (RI-RADS C/D) fell to 0% to 0.7% (0/608 to 4/608) with sLLM augmentation vs 4.1% to 20.4% (25/608 to 124/608) for clinician MERs. Overall protocol accuracy was 93.1% (566/608; 95% CI 89.6‐96.6) for the sLLM, 91.4% (556/608; 95% CI 87.6‐95.3) for Rad 3, and 92.1% (560/608; 95% CI 88.4‐95.8) for Rad 4 (sLLM vs Rad 3 P=.23 vs Rad 4 P=.40). Region or coverage accuracy was similar (sLLM: 579/608, 95.2%; Rad 3: 585/608, 96.2%; Rad 4: 573/608, 94.2%; P=.46 and P=.36). Contrast decisions were more accurate using the sLLM at 94.4% (574/608; 95% CI 91.3‐97.5) vs Rad 3 at 92.1% (560/608; 95% CI 88.4‐95.8; P=.027) and were not significantly different to Rad 4 at 92.9% (565/608; 95% CI 89.4‐96.4; P=.16). Subspecialty analyses showed similar patterns, with the sLLM outperforming Rad 4 for musculoskeletal MRI contrast decisions (96.6% vs 91.1%; P=.006) and matching readers elsewhere. Manual review indicated that sLLM improvements arose from EMR details not listed on the MER (infection/inflammation, tumor history, prior surgery). No clinically significant hallucinations were identified in a manual review of discordant cases.

Conclusions: Across body, musculoskeletal, and neuroradiology MRI, sLLM-augmented examination requests improved clinical context and enhanced contrast selection while demonstrating accuracy comparable to general radiologists for region or coverage.

XSL•FO
RenderX

Integrating sLLMs into routine vetting workflows may reduce manual workload in protocol selection for more efficient, standardized protocoling.

(J Med Internet Res 2026;28:e82579) doi:10.2196/82579

KEYWORDS

secure large language model; radiology request form; reason for exam imaging reporting and data system; musculoskeletal imaging; body imaging; neuroradiology imaging; magnetic resonance imaging

Introduction

Radiology examination requests form the primary communication bridge between referring clinicians and radiology services. Completeness and clarity of these requests influence protocol selection, image quality, and ultimately diagnostic and therapeutic outcomes [1-3]. When key clinical elements such as relevant history, current symptoms, or prior imaging are incomplete or absent, radiographers and radiologists may assign a suboptimal protocol or omit intravenous contrast, leading to lower accuracy, repeat studies, and additional cost [4-6]. A recent systematic review showed that incomplete clinical information was associated with reduced reporting accuracy, clinical relevance, and reporting confidence [5].

The enhancement of radiology examination requests with data from the electronic medical record (EMR) offers a potential solution. Early rule-based decision-support tools reduced certain categories of inappropriate imaging yet may struggle to accommodate unstructured narratives and updated imaging protocols [6]. Natural language processing pipelines improved specificity but required extensive engineering and did not fully interpret clinical context [7]. Current large language models (LLMs) can ingest heterogeneous text, interpret medical terminology, and generate structured outputs at scale [8-13]. Proof-of-concept work has already demonstrated utility for report generation and national guideline concordance [14,15]. Nevertheless, many published studies rely on deidentified vignettes or synthetic notes, limiting direct clinical translation because of privacy constraints.

Secure LLMs (sLLMs) address these constraints by operating behind institutional firewalls while preventing the transmission of protected health information [16]. Several studies have demonstrated their feasibility, including improving the adequacy of spine magnetic resonance imaging (MRI) request forms and increased protocol concordance for musculoskeletal studies [17,18]. Recent oncologic imaging work further indicates that a GPT-4 system can automatically generate structured clinical histories that radiologists prefer over original clinician-generated requisitions [19]. Another recent study on GPT-4-generated MRI protocols showed notable quality in cardiac and neuroradiology imaging [20]. An accompanying editorial emphasized that fatigue-free, verifiable LLM summaries could finally bridge the long-recognized information gap between referrers and radiologists, reinforcing the case for secure, institution-hosted models [21].

Building on these findings, we evaluated a secure LLM across all routine MRI subspecialties, including body, neuroradiology, and musculoskeletal examinations. The study had 2 objectives: (1) to compare the information quality of clinician MRI examination requests (MERs) with those augmented by the sLLM using the Reason-for-Exam Imaging Reporting and Data System (RI-RADS) [22] and (2) to assess protocol accuracy of the sLLM against both subspecialty and general radiologist readers.

Methods

Ethical Considerations

The Institutional Review Board (Domain-Specific Review Board reference: 2023/00919) classified the project as minimal risk and therefore granted a waiver of informed consent. Patient-identifying details have been omitted to maintain the privacy and confidentiality of patient data. Compensation was not required in view of the minimal risk of the project.

Protocoling Pipeline and sLLM Prompt

Original clinician-submitted outpatient MERs performed between September 2023 and July 2024 were retrieved at random. MERs lacking corresponding EMR information were excluded. The musculoskeletal cohort in this study represents a distinct nonoverlapping dataset from that of a previous study by the authors. A target sample of approximately 600 requests (~200 per major subspecialty) was selected to provide adequate statistical power while maintaining operational feasibility across a diverse range of anatomical regions and pathologies. For every patient, a MER was generated by the sLLM based on the clinician request with relevant EMR content.

The institutional sLLM is a privately hosted instance of Anthropic Claude 3.5 on Amazon Bedrock. Model temperature was fixed at 0 to minimize output variability. For each request, the sLLM received (1) the most recent relevant clinical entry identified manually by the authors, based on the clinical service indicated on the request forms and (2) pertinent prior imaging reports (eg, CT neck for an MRI nasopharynx study). The typical input length of the clinical notes provided to the sLLM ranged from around 100 words for routine or straightforward presentations to around 3000 words for patients with complex oncological or multisystem disease. The sLLM was also given the latest institutional MRI protocol repository (Multimedia Appendix 1). No clinical data left the institutional firewall.

Two-step sLLM protocoling was performed:

1. Information enrichment—the sLLM extracted key findings, working diagnoses, and potential MRI safety issues from both the MER and the EMR, returning a concise justification summary.
2. Protocol assignment—using institutional rules and the enriched summary, the model selected the optimal

XSL•FO
RenderX

region/coverage and determined whether contrast was required.

The estimated processing cost using the sLLM was US $3 per 1,000,000 tokens for the input and US $15 per 1,000,000 tokens for the output, giving a total cost of US $0.024 to 0.033 per request. For the cases requiring longer input length (about 5000 tokens), the per-case cost remained low at a few cents per request.

Importantly, the sLLM's primary role in this pipeline was clinical information extraction, summarizing clinically relevant details from the MER and EMR (eg, prior surgery, suspected infection, etc) and contextual interpretation. The final contrast and region decisions were not made autonomously by the sLLM. Instead, the extracted elements were subsequently mapped onto predefined institutional criteria for region or coverage and contrast use using a deterministic rule-based parsing script.

Contrast (gadolinium) administration was considered when any of the following were present:

1. History or suspicion of tumor, malignancy, or focal lesion
2. Known or suspected infective or inflammatory condition
3. Injury to neural structures
4. Previous surgery or spinal/extremity instrumentation
5. Explicit clinician request

A custom parsing script evaluated the sLLM response, which listed a "yes" or "no" decision for each contrast-relevant category together with the reason when applicable. If the script detected at least one positive flag and renal failure or other contraindications had not been recognized, the examination was classified as "contrast required."

Detailed prompt wording and rule logic are provided in Multimedia Appendix 1. The rule definitions in this study were refined through prior work on LLMs in spine and musculoskeletal protocoling. All existing MRI protocols (including prompt design and contrast rules) were carefully reviewed with subspecialty leads and key team members to ensure alignment with institutional practice before doing the formal analysis. The full institutional MRI protocol table had a size compatible with the available context window and was embedded directly within the prompt, obviating the need for retrieval-augmented generation (RAG) in this study. No cases from the present study were used during prompt development.

Evaluation Procedure

Two senior radiologists (Rad 1: 14 years of experience; Rad 2: 12 years of experience) independently graded the clinical adequacy of both clinician and sLLM MERs using the RI-RADS classification (Multimedia Appendix 2). They were blinded to the origin of each form.

Rad 1 and Rad 2 provided a consensus reference standard for protocoling. Against this consensus reference standard, protocol selections generated by the sLLM were then compared with those from 2 board-certified general radiologists (Rad 3 and Rad 4), who also had access to the same EMR entries, pertinent prior imaging reports, and institutional MRI protocols. The pertinent prior imaging reports were manually selected by study members, who were individuals distinct from Rad 1 and Rad 2. Both Rad 3 and Rad 4 had 2 years of experience and were commonly tasked with providing protocols based on MRI examination requests. Accuracy was credited only when the suggested region or coverage and contrast decision matched the consensus reference standard. Additional subanalyses were performed for each component and the major subspecialties. A manual review of disagreements between the sLLM and board-certified general radiologists against the consensus reference standard was carried out by the senior radiologists. Note was made of any hallucinations by the sLLM with clinically significant hallucinations defined as the sLLM producing clinically relevant details absent from the provided input data and capable of influencing protocol decisions [23].

Statistical Analysis

All computations were performed with Python 3.9.12. Two-sided tests were considered significant at $P<.05$.

The quality of each MER was rated on a 4-level scale (RI-RADS grades A-D) by Rad 1 and Rad 2. Interreader agreement for these ordinal ratings, and later for MRI protocoling decisions (region/coverage and contrast requirement), was quantified with Gwet AC1, which is less sensitive than Cohen κ to category imbalance (eg, the predominance of A/B grades) [24]. Agreement was interpreted as poor (<0), slight (0 - 0.20), fair (0.21 - 0.40), moderate (0.41 - 0.60), substantial (0.61 - 0.80), or almost perfect (0.81 - 1). With more than 600 examinations, the study had greater than 90% power to detect a minimum 15% absolute difference in the proportion of clinically adequate MERs between clinician and sLLM versions.

Protocol selections generated by the sLLM and by 2 board-certified general radiologists (Rad 3 and Rad 4) were compared with the consensus reference standard established by Rad 1 and Rad 2. Overall accuracy was defined as the proportion of cases in which both region or coverage and contrast indication matched the reference standard. Differences in paired accuracies (sLLM vs each radiologist) were evaluated using the McNemar test with Yates continuity correction. Separate analyses were undertaken for region alone and contrast alone, and across the 3 major subspecialties (body, musculoskeletal, and neuroradiology). Ninety-five percent CIs for RI-RADS gradings (quality of clinical information) and protocoling were derived from the normal approximation to the binomial distribution. For proportions near 0% or 100%, the normal approximation to the binomial can be inaccurate. Therefore, for low-event RI-RADS C/D proportions (eg, sLLM MERs), 95% CIs were computed using the Clopper-Pearson exact method rather than the normal approximation.

Results

Patient Demographics and Consensus MRI Protocols

Overall, 608 MRI examination requests were collected from 528 patients (mean 51.2 y, SD 19.2; range 4 - 93 y). In total, 52.8% (279/528) patients were women and 47.2% (249/528) were men (Table 1). An additional 27 MRI examinations (26 patients) were excluded from analysis due to incomplete EMRs (eg, external referrals) (Figure 1).

XSL•FO
RenderX

Table . Patient and magnetic resonance imaging examination characteristics.

Characteristics	Values
Age (y), mean (SD; range)	
All (N=528)	51.2 (SD 19.2; 4-93)
Women (n=279)	52.9 (SD 17.4; 4-93)
Men (n=249)	49.3 (SD 21; 5-92)
MRI[a] study specialties (N=608), n (%)	
Body MRI (n=206)	
Rectum/perineum fistula	28 (13.6)
Uterus/cervix cancer	26 (12.6)
Liver routine	21 (10.2)
Prostate routine	21 (10.2)
Enterography	16 (7.8)
Pancreas routine	14 (6.8)
Other	80 (38.8)
Musculoskeletal MRI (n=203)	
Lumbar spine	25 (12.3)
Cervical spine	24 (11.8)
Shoulder	29 (14.3)
Knee	27 (13.3)
Pelvis	11 (5.4)
Other	87 (42.9)
Neuroradiology MRI (n=199)	
Brain + contrast	40 (20.1)
Orbits	18 (9)
Skull base/temporal bones	15 (7.5)
Brain routine (noncontrast)	15 (7.5)
Brain stroke (acute)	14 (7)
Brain MR angiography	12 (6)
Pituitary dynamic	12 (6)
Oral cavity/neck	12 (6)
Nasopharynx and neck	10 (5)
Other	51 (25.6)
MRI study type (N=608), n (%)	
Routine (noncontrast)	239 (39.3)
Contrast	369 (60.7)
Contrast use by specialty	
Body MRI	Contrast 175 (85); noncontrast 31 (15)
Musculoskeletal MRI	Contrast 55 (27); noncontrast 148 (73)
Neuroradiology MRI	Contrast 139 (70); noncontrast 60 (30)

[a]MRI: magnetic resonance imaging.

XSL•FO
RenderX

Figure 1. Study design flowchart. The initial clinician-generated MERs were extracted and then augmented using an sLLM. The clinician and sLLM MERs were compared for the quality of the clinical information available using the RI-RADS grading scale by 2 experienced radiologists (Rads 1 and 2). The MRI protocol accuracy for the sLLM and 2 board-certified radiologists (Rads 3 and 4, both with 2 years of experience) was determined by comparison against a reference standard provided by Rads 1 and 2. Claude version 3.5 (Anthropic) was used. *Clinical data included the last clinical entry and imaging reports (eg, computed tomography (CT) of the neck for an MRI nasopharynx). The sLLM and all board-certified radiologists had access to the MRI protocol guidance provided on the hospital intranet. MERs: MRI examination requests; MRI: magnetic resonance imaging; RI-RADS: Reason-for-Exam Imaging Reporting and Data System; sLLM: secure large language model.

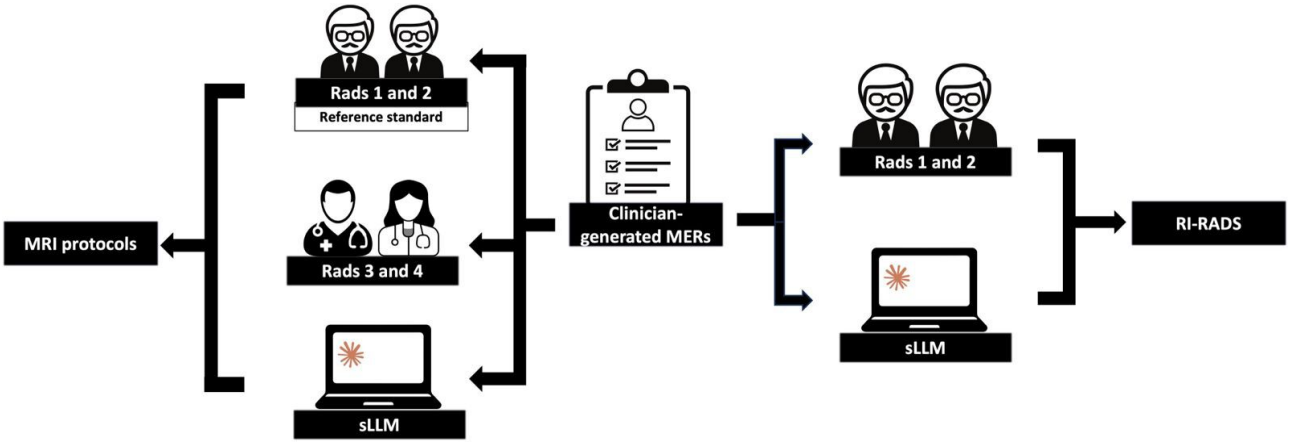

Out of the 608 MRI examinations analyzed, there was the following subspecialty breakdown: 206 (34%) body, 203 (33%) musculoskeletal, and 199 (33%) neuroradiology. Contrast was administered in 85% (175/206) of body scans, 27% (55/203) of musculoskeletal scans, and 70% (139/199) of neuroradiology scans, yielding 60.7% (369/608) of contrast studies overall. Within each specialty, the most common protocols were rectum or perineal fistula, uterus or cervix cancer, liver, and prostate studies for body MRI; lumbar and cervical spine along with knee and shoulder examinations for musculoskeletal MRI; and postcontrast brain, orbital, and skull-base studies for neuroradiology MRI (Table 1).

Adequacy of the Radiology Request Forms

RI-RADS gradings for the MERs were performed independently by 2 experienced radiologists (Rad 1 and Rad 2) (Table 2). Interobserver agreement (AC1) was almost perfect for the sLLM-augmented MERs (AC1 0.97, 95% CI 0.94 - 0.99) and moderate for the clinician MERs (AC1 0.43, 95% CI 0.34 - 0.52).

Table . Reason for exam imaging reporting and data system grades for the clinician and secure large language model–augmented magnetic resonance imaging examination requests[a].

RI-RADS[b] grade	Radiologist 1		Radiologist 2	
	Clinician MERs[c] (n=608), n (%, 95% CI)	sLLM[d] MERs (n=608), n (%, 95% CI)	Clinician MERs (n=608), n (%, 95% CI)	sLLM MERs (n=608), n (%, 95% CI)
A/B	484 (79.6, 95% CI 76.2 - 82.8)	604 (99.3, 95% CI 98.3 - 99.8)	583 (95.9, 95% CI 93.9 - 97.3)	608 (100, 95% CI 99.51 - 100)
C/D	124 (20.4, 95% CI 17.2 - 23.6)	4 (0.7, exact 95% CI, Clopper-Pearson 0.2 - 1.7)	25 (4.1, 95% CI 2.7 - 6)	0 (0, exact 95% CI, Clopper-Pearson 0 - 0.5)

[a]Values are the number of studies, with 95% CIs in brackets.

[b]RI-RADS: Reason-for-Exam Imaging Reporting and Data System.

[c]MERs: magnetic resonance imaging examination requests.

[d]sLLM: secure large language model.

Clinical information on the sLLM-augmented MERs was rated significantly higher than the original requests by both radiologists. For Rad 1, clinician MERs had 484/608 rated A/B (79.6%; 95% CI 76.2%-82.8%) and 124/608 rated C/D (20.4%; 95% CI 17.2%-23.6%), whereas sLLM MERs had 604/608 rated A/B (99.3%; 95% CI 98.3%-99.8%) and 4/608 rated C/D (0.7%; exact 95% CI, Clopper-Pearson, 0.2%-1.7%). For Rad 2, clinician MERs had 583/608 rated A/B (95.9%; 95% CI 93.9%-97.3%) and 25/608 rated C/D (4.1%; 95% CI 2.7%-6.0%), while sLLM MERs had all 608/608 rated A/B (100%; 95% CI 99.51%-100%) and 0/608 rated C/D (0%; exact 95% CI, Clopper-Pearson, 0%-0.5%).

No clinically significant hallucinations were identified in the sLLM outputs. Notably, the sLLM-generated MERs converted 28 spine MRI referrals (28/608, 4.6%) that were rated limited or deficient (7 of which omitted prior spinal surgery) into requests containing adequate clinical details.

MRI Protocoling Accuracy Comparison

For protocoling, experienced radiologists 1 and 2 showed almost perfect agreement on the region (AC1 0.97, 95% CI 0.96 - 0.99) and need for contrast (AC1 0.99, 95% CI 0.99 - 1). A consensus protocol was determined for all 608 cases.

XSL•FO
RenderX

For all 608 MRI examinations, the sLLM matched the subspecialty reference standard in 566 (93.1%, 95% CI 89.6 - 96.6) cases overall, compared with 556 (91.4%, 95% CI 87.6 - 95.3) for Rad 3 and 560 (92.1%, 95% CI 88.4 - 95.8) for Rad 4 (Table 3). This difference was not significant for the sLLM vs Rad 3 (*P*=.23) and Rad 4 (*P*=.40). Accuracy for region or coverage selection was similar across readers (sLLM: 579/608, 95.2%; Rad 3: 585/608, 96.2%; Rad 4: 573/608, 94.2%) with no significant differences between the sLLM vs Rad 3 (*P*=.46) and Rad 4 (*P*=.36). The sLLM, coupled with the use of the deterministic rules-based parsing script, demonstrated a slight advantage for contrast determination, being correct in 94.4% (574/608; 95% CI 91.3 - 97.5) studies vs 92.1% (560/608; 95% CI 88.4 - 95.8) for Rad 3 (*P*=.027) and 92.9% (565/608; 95% CI 89.4 - 96.4) for Rad 4 (*P*=.16). Manual review of disagreements indicated that the sLLM correctly detected infection or inflammation in 2 cases (eg, suspected bursitis around the hip and shoulder), possible nerve lesions in 2 cases, and prior spine surgical details in 1 case, which were documented in the electronic record but not stated on the MER. These details triggered appropriate contrast recommendations that were omitted by at least 1 radiologist.

Table . Magnetic resonance imaging protocoling accuracy for the secure large language model and board-certified radiologists vs the reference standard.

MRI[a] protocol vs reference standard	Radiologist 3	Radiologist 4	sLLM[b]
All MRI studies (N=608), n (%, 95% CI)			
Overall protocol correct	556 (91.4, 95% CI 87.6 - 95.3)	560 (92.1, 95% CI 88.4 - 95.8)	566 (93.1, 95% CI 89.6 - 96.6)
Region correct	585 (96.2, 95% CI 93.6 - 98.8)	573 (94.2, 95% CI 91.1 - 97.4)	579 (95.2, 95% CI 92.3 - 98.1)
Contrast decision correct	560 (92.1, 95% CI 88.4 - 95.8)	565 (92.9, 95% CI 89.4 - 96.4)	574 (94.4, 95% CI 91.3 - 97.5)
Subspecialty subsets			
Musculoskeletal (n=203), n (%, 95% CI)			
Overall protocol correct	186 (91.6, 95% CI 87.8 - 95.4)	180 (88.7, 95% CI 84.3 - 93)	187 (92.1, 95% CI 88.4 - 95.8)
Region correct	199 (98, 95% CI 96.1 - 99.9)	182 (89.7, 95% CI 85.5 - 93.8)	194 (95.6, 95% CI 92.7 - 98.4)
Contrast decision correct	190 (93.6, 95% CI 90.2 - 97)	185 (91.1, 95% CI 87.2 - 95)	196 (96.6, 95% CI 94 - 99.1)
Neuroradiology (n=199), n (%, 95% CI)			
Overall protocol correct	184 (92.5, 95% CI 88.8 - 96.1)	189 (95, 95% CI 91.9 - 98)	185 (93, 95% CI 89.4 - 96.5)
Region correct	189 (95, 95% CI 91.9 - 98)	194 (97.5, 95% CI 95.3 - 99.7)	190 (95.5, 95% CI 92.6 - 98.4)
Contrast decision correct	184 (92.5, 95% CI 88.8 - 96.1)	189 (95, 95% CI 91.9 - 98)	184 (92.5, 95% CI 88.8 - 96.1)
Body (n=206), n (%; 95% CI)			
Overall protocol correct	186 (90.3, 95% CI 86.2 - 94.3)	191 (92.7, 95% CI 89.2 - 96.3)	194 (94.2, 95% CI 91 - 97.4)
Region correct	197 (95.6, 95% CI 92.8 - 98.4)	197 (95.6, 95% CI 92.8 - 98.4)	195 (94.7, 95% CI 91.6 - 97.7)
Contrast decision correct	186 (90.3, 95% CI 86.2 - 94.3)	191 (92.7, 95% CI 89.2 - 96.3)	194 (94.2, 95% CI 91 - 97.4)

[a]MRI: magnetic resonance imaging.

[b]sLLM: secure large language model.

Within the musculoskeletal subset (N=203 examinations), overall agreement with the reference standard was 187 (92.1%, 95% CI 88.4 - 95.8) for the sLLM, 186 (91.6%, 95% CI 87.8 - 95.4) for Rad 3, and 180 (88.7%, 95% CI 84.3 - 93) for Rad 4. These differences were not significant (*P*>.99 and *P*=.15 for sLLM vs Rad 3 and Rad 4, respectively). The sLLM matched Rad 3 for region accuracy (n=194, 95.6% vs 199, 98%; *P*=.23) and exceeded Rad 4 (n=182, 89.7%; *P*=.018) (Table 3). For contrast, coupled with the use of the deterministic rules-based parsing script, the sLLM was correct in 196 (96.6%, 95% CI 94 - 99.1) studies, outperforming Rad 4 alone (n=185, 91.1%; *P*=.006) with no significant difference compared to Rad 3 (n=190, 93.6%; *P*=.099).

For body MRI (N=206 examinations), the sLLM had the highest overall accuracy (n=194, 94.2%, 95% CI 91 - 97.4), although this was not significantly different vs Rad 3 (n=186, 90.3%; *P*=.089) and Rad 4 (n=191, 92.7%; *P*=.33). Review of body MRI discrepancies indicated that the sLLM recommended more focused uterus or cervix and MRI rectum and perineum protocols in 2 cases compared to a more general pelvis by the 2 radiologists. For contrast, the sLLM had the highest accuracy (n=194, 94.2%, 95% CI 91 - 97.4), although this was not significantly different vs Rad 3 (n=186, 90.3%; *P*=.09) and Rad 4 (n=191, 92.7%; *P*=.33).

In neuroradiology MRI (N=199 examinations), Rad 4 achieved the highest overall accuracy (n=189, 95%, 95% CI 91.9 - 98), although this was not significant vs the sLLM (n=185, 93%; *P*=.50) and Rad 3 (n=184, 92.5%; *P*=.71). Review of neuroradiology discrepancies indicated that the sLLM occasionally recommended skull-base protocols in 5 cases with suspected cranial nerve pathologies based on the clinical notes, whereas radiologists 3 and 4 accepted the original brain request when they deemed that coverage was adequate. For contrast, Rad 4 achieved the highest accuracy (n=189, 95%, 95% CI 91.9 - 98), although this was not significant vs the sLLM

XSL•FO
RenderX

(n=184, 92.5%; *P*=.30) and Rad 3 (n=184, 92.5%; *P*=.32) which were closely aligned.

Discussion

Principal Findings

In this study, we compared MERs provided by the referring clinician with those augmented by our institutional sLLM. Our results show that an sLLM can upgrade the clinical usefulness of MERs and aid protocol selection across body, neuroradiology, and musculoskeletal practice. After augmentation, fewer than 1% of requests (0% - 0.7%) were graded deficient or limited (RI-RADS C or D), compared with up to 1 in 5 (4.1% - 20.4%) of the clinician originals, and interreader agreement rose from moderate to almost perfect.

The sLLM protocoling accuracy was compared with that of 2 board-certified junior general radiologists. Overall protocol accuracy of the sLLM (566/608, 93.1%) was close to that of the 2 board-certified radiologists (Rad 3: 556/608, 91.4%; *P*=.23 and Rad 4: 560/608, 92.1%; *P*=.40), although the study was not powered to test statistical noninferiority or equivalence. Observed accuracy differences should not be interpreted as evidence of statistical similarity but instead as descriptive comparisons indicating that the sLLM operated within the performance range of general radiologists on this dataset.

For contrast decisions, the sLLM demonstrated accuracy of 94.4% (574/608), which was superior to Rad 3 (560/608, 92.1%; *P*=.027) and was not significantly different from Rad 4 (565/608, 92.9%; *P*=.16). Manual review showed that slightly increased accuracy for the sLLM was driven by EMR information (correct detection of prior surgery, infection, or tumor history) that did not appear on the MER yet was important for protocol selection. In our pipeline, the sLLM's role was to extract and summarize these clinically relevant details, while the final region and contrast decisions were made by a deterministic rule-based script using predefined institutional criteria. Importantly, protocol selection and contrast determination are treated as separate decision steps. Protocol names reflect the anatomical region and clinical indication, while contrast administration may be modified when an explicit clinician request for noncontrast imaging or a documented contraindication is detected. In such cases, the default contrast setting associated with a protocol may be overridden to preserve clinician intent and patient safety. As such, the observed accuracy reflects improved information extraction feeding into consistent rules, rather than autonomous clinical reasoning by the sLLM.

In one of this study's cases, the clinician requested noncontrast MRI for cervical cancer recurrence due to renal impairment of estimated glomerular filtration rate 31 - 48, although modern guidelines may permit the use of group II gadolinium agents at this level. We acknowledge that prioritizing clinician intent in such cases may risk suboptimal diagnostic utility and potential recall.

Most mismatches against the reference standard occurred in borderline cases where more than one subspecialty-specific protocol was reasonable. In body MRI, this typically involved the sLLM proposing a general pelvis study when radiologists selected a focused protocol such as prostate, uterus or cervix, or rectum. In spine MRI, a small number of discrepancies arose from consolidating concurrent cervical and lumbar requests into a single-region study. In neuroradiology, differences usually reflected choosing between brain vs skull base or temporal bone coverage in suspected cranial nerve pathology. These patterns suggest that minor refinements to protocol-selection rules could help reduce discrepancies.

Comparison to Prior Work

These findings are consistent with recent evaluations of LLMs for noninterpretive tasks in radiology and add multisubspecialty evidence. Prior institutional studies using an sLLM for spine and musculoskeletal MRI reported similar gains, with musculoskeletal protocoling accuracy reaching 96% and exceeding 2 general readers at 88% and 89% [17,18]. In another recent study, Çamur et al [25] showed that 4 LLMs have strong potential for selecting appropriate imaging modalities. The 4 LLMs were tested on 240 clinical cases (120 ACR Appropriateness Criteria and 120 realistic scenarios) and compared their choices with 4 clinicians and 4 radiologists. The best model picked the correct imaging test in 98.3% (236/240) of ACR cases, matched a junior radiologist on realistic cases, gave identical answers across prompts (κ=1), and showed moderate to good reproducibility over time (κ~0.77 - 0.89 short term; 0.51 - 0.79 long term) [25]. Similarly, for CT protocol assignment, a fine-tuned support tool picked the right protocol on the first choice 92.3% of the time and within its top two 96.3% of the time, with an average processing time of less than 1 second per case. When clinicians used it, resident accuracy improved from 0.913 to 0.936 with a 14% reduction in reading time, and attending accuracy increased from 0.920 to 0.926 with a 12% time saving [26]. In musculoskeletal MRI, a GPT-4 system linked to a small knowledge base reached 92.86% accuracy on ACR-based cases, outperformed a baseline model and standard GPT-4, matched most subspecialists, and was better at flagging when the clinical information was insufficient, which mirrors our MER enrichment step [27]. Beyond protocol choice, a radiology operations study showed GPT-4 routed 96% of in-scope procedure requests and 76% of out-of-scope requests correctly at a cost of approximately US $0.03 per request, indicating potential for increased efficiency and cost benefits for routing workflows [28]. Together with our results, these data suggest that institution-hosted sLLMs could aid improvements in clinical context, standardize protocol decisions, reduce unnecessary contrast and radiation, and potentially save time and costs by reducing protocoling time and reschedules.

Limitations and Future Directions

Our study has several limitations. First, this was a single-center, retrospective evaluation involving only outpatient MERs. The sLLM pipeline relied on the specific institutional MRI protocol repository and access to the EMR structure, and the realized gains may differ in institutions using different order-entry systems or protocol libraries. Future work should include prospective, multicenter evaluations that measure operational outcomes (time to scan, rescheduling, repeat imaging), cost-effectiveness analyses that incorporate staff time and

XSL•FO
RenderX

scanner utilization, and extension of the sLLM pipeline to CT and ultrasound.

Second, some patients contributed multiple examinations, and the dataset lacked patient-linked identifiers; therefore, cluster-robust or mixed-effects analyses were not feasible. Results should be interpreted at the examination level, with future work planned to incorporate patient-level clustering.

Third, formatting differences between the clinician and augmented MRI examination requests could have partially unblinded the graders, with the potential to artificially inflate the sLLM's RI-RADS scores, although the direction of any resulting bias is uncertain. In our study, the output format (subheadings; eg, clinical history, reason for exam) was identical for the clinician and sLLM forms, although differences in language and style may have unblinded the reviewers. This is difficult to rectify but could be addressed in future studies using standardized outputs.

Fourth, the exclusion of cases with incomplete electronic records introduces a potential selection bias, as these omitted cases may represent some of the more challenging instances for protocoling. In addition, the manual extraction of the latest relevant clinical entry and pertinent prior imaging reports without an automated failsafe mechanism provided a methodological simplification, which is different from real-life human EMR review, which involves searching through all relevant notes and prior imaging. Together, these could potentially inflate the sLLM's protocoling accuracy and are an important area for future improvement. Future scaling may also benefit from approaches such as hierarchical summarization or RAG-style chunking for very large EMR entries.

Fifth, we assessed board-certified radiologists who demonstrated high accuracies for protocoling (>91.4%), leaving little room for the sLLM to improve on this. LLM assistance for less experienced readers, including technologists, could provide more value and will need to be assessed. Nonetheless, even if accuracy gains over experts are marginal, the standardization benefit and fatigue reduction offered by the sLLM support its deployment. A further improvement may include an "assisted radiologist with sLLM" arm to separate automation from augmentation.

Sixth, although the study included more than 600 MRI examinations, the sample size was powered for detecting a 15% difference in clinical-information adequacy, not for establishing noninferiority or equivalence in protocol accuracy. The small absolute differences observed between the sLLM and board-certified radiologists (eg, 93.1% vs 91.4%; P=.23) therefore cannot be interpreted as statistical equivalence.

Seventh, we deliberately confined the evaluation to MRI, which is our most protocol-intensive modality and the largest component of departmental workload, but acknowledge that CT and ultrasound will require separate validation prior to broader application.

Eighth, the review for clinically significant hallucinations was limited to discordant cases, and there is a theoretical risk that the sLLM could generate benign or "silent" hallucinations in the clinical summary that do not alter the protocol decision but could still degrade the medical record.

Finally, although no clinically significant hallucinations were observed, it is well known that LLMs are prone to producing factual errors that could have major clinical implications [29-31]. In addition, recent work shows that LLMs can be vulnerable to adversarial hallucination in clinical decision-support settings, with mitigation prompts only partially reducing errors and lower temperature settings providing no meaningful benefit [31]. Continuous monitoring and postdeployment guardrails therefore remain essential, but were beyond the scope of this study.

Conclusion

In this multisubspecialty MRI cohort, an sLLM improved MRI examination request completeness and demonstrated protocol accuracy comparable to experienced radiologists; however, the study was not powered to establish noninferiority or equivalence for protocol accuracy.

The end-to-end sLLM pipeline, consisting of request form enrichment with EMR data and a rule-based protocoling technique, offers a practical pathway for more efficient, standardized protocoling while reducing the administrative burden on clinicians.

Acknowledgments

The author(s) attest that there was no use of generative artificial intelligence technologies in the generation of text, figures, or other informational content of this manuscript.

Funding

This study received support from the Singapore Ministry of Health National Medical Research Council under the NMRC Clinician Innovator Award (CIA). The grant was awarded for the project titled "From Prototype to Full Deployment: A Comprehensive Deep Learning Pipeline for Whole-Spine MRI" (Grant ID: CIAINV25jan-0005 J.T.P.D.H).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: JTPDH, NWL, YXL
Data curation: JTPDH, NWL
Formal analysis: JTPDH, NWL
Funding acquisition: JTPDH
Investigation: JTPDH, NWL, YXL
Methodology: JTPDH, NWL, YXL
Project administration: JTPDH, NWL, YXL
Resources: JTPDH, NWL, YXL, AL, WO, MDZC, GKD, XZL, MCL, CY, WJS, ECT, JHT, NK, AM, YT
Software: JTPDH, NWL, YXL, AL, WO, MDZC, GKD, XZL, MCL, CY, WJS, ECT, JHT, NK, AM, YT
Supervision: JTPDH
Validation: JTPDH, NWL, YXL, AL, WO, MDZC, GKD, XZL, MCL, CY, WJS, ECT, JHT, NK, AM, YT
Visualization: JTPDH, NWL, YXL
Writing – original draft: JTPDH, NWL, YXL
Writing – reviewing & editing: JTPDH, NWL, YXL, AL, WO, MDZC, GKD, XZL, MCL, CY, WJS, ECT, JHT, NK, AM, YT

Conflicts of Interest

None declared.

Multimedia Appendix 1
Large language model prompts.
[DOCX File, 42 KB - jmir_v28i1e82579_app1.docx]

Multimedia Appendix 2
Reason for Exam Imaging Reporting and Data System (RI-RADS).
[DOCX File, 22 KB - jmir_v28i1e82579_app2.docx]

References

1. Lehnert BE, Bree RL. Analysis of appropriateness of outpatient CT and MRI referred from primary care clinics at an academic medical center: how critical is the need for improved decision support? J Am Coll Radiol 2010 Mar;7(3):192-197. [doi: 10.1016/j.jacr.2009.11.010] [Medline: 20193924]
2. Bernardy M, Ullrich CG, Rawson JV, et al. Strategies for managing imaging utilization. J Am Coll Radiol 2009 Dec;6(12):844-850. [doi: 10.1016/j.jacr.2009.08.003] [Medline: 19945039]
3. G Pitman A. Quality of referral: what information should be included in a request for diagnostic imaging when a patient is referred to a clinical radiologist? J Med Imaging Radiat Oncol 2017 Jun;61(3):299-303. [doi: 10.1111/1754-9485.12577] [Medline: 28139044]
4. Barakzai MD, Sheer ZZ, Muhammad A, et al. Evaluation of radiology request forms in a tertiary care hospital: an audit with a focus on the impact of technological intervention. Cureus 2021 Feb 14;13(2):e13335. [doi: 10.7759/cureus.13335] [Medline: 33747644]
5. Castillo C, Steffens T, Sim L, Caffery L. The effect of clinical information on radiology reporting: a systematic review. J Med Radiat Sci 2021 Mar;68(1):60-74. [doi: 10.1002/jmrs.424] [Medline: 32870580]
6. Ip IK, Schneider L, Seltzer S, et al. Impact of provider-led, technology-enabled radiology management program on imaging. Am J Med 2013 Aug;126(8):687-692. [doi: 10.1016/j.amjmed.2012.11.034] [Medline: 23786668]
7. Trivedi H, Mesterhazy J, Laguna B, Vu T, Sohn JH. Automatic determination of the need for intravenous contrast in musculoskeletal MRI examinations using IBM Watson's natural language processing algorithm. J Digit Imaging 2018 Apr;31(2):245-251. [doi: 10.1007/s10278-017-0021-3] [Medline: 28924815]
8. Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. Radiology 2024 Jan;310(1):e232756. [doi: 10.1148/radiol.232756] [Medline: 38226883]
9. Gertz RJ, Bunck AC, Lennartz S, et al. GPT-4 for automated determination of radiological study and protocol based on radiology request forms: a feasibility study. Radiology 2023 Jun;307(5):e230877. [doi: 10.1148/radiol.230877] [Medline: 37310247]
10. Barash Y, Klang E, Konen E, Sorin V. ChatGPT-4 assistance in optimizing emergency department radiology referrals and imaging selection. J Am Coll Radiol 2023 Oct;20(10):998-1003. [doi: 10.1016/j.jacr.2023.06.009] [Medline: 37423350]
11. Kim S, Lee CK, Kim SS. Large language models: a guide for radiologists. Korean J Radiol 2024 Feb;25(2):126-133. [doi: 10.3348/kjr.2023.0997] [Medline: 38288895]
12. Sun Z, Ong H, Kennedy P, et al. Evaluating GPT4 on impressions generation in radiology reports. Radiology 2023 Jun;307(5):e231259. [doi: 10.1148/radiol.231259] [Medline: 37367439]

XSL•FO
RenderX

13. Mese I, Taslicay CA, Sivrioglu AK. Improving radiology workflow using ChatGPT and artificial intelligence. Clin Imaging 2023 Nov;103:109993. [doi: 10.1016/j.clinimag.2023.109993] [Medline: 37812965]
14. Mukherjee P, Hou B, Lanfredi RB, Summers RM. Feasibility of using the privacy-preserving large language model vicuna for labeling radiology reports. Radiology 2023 Oct;309(1):e231147. [doi: 10.1148/radiol.231147] [Medline: 37815442]
15. Rau A, Rau S, Zoeller D, et al. A context-based chatbot surpasses trained radiologists and generic ChatGPT in following the ACR appropriateness guidelines. Radiology 2023 Jul;308(1):e230970. [doi: 10.1148/radiol.230970] [Medline: 37489981]
16. Cai W. Feasibility and prospect of privacy-preserving large language models in radiology. Radiology 2023 Oct;309(1):e232335. [doi: 10.1148/radiol.232335] [Medline: 37815443]
17. Hallinan J, Leow NW, Ong W, et al. MRI spine request form enhancement and auto protocoling using a secure institutional large language model. Spine J 2025 Mar;25(3):505-514. [doi: 10.1016/j.spinee.2024.10.021] [Medline: 39536908]
18. Hallinan J, Leow NW, Low YX, et al. An institutional large language model for musculoskeletal MRI improves protocol adherence and accuracy. J Bone Joint Surg Am 2025 Jul 8;107(16):1833-1840. [doi: 10.2106/JBJS.24.01429] [Medline: 40627696]
19. Bhayana R, Alwahbi O, Ladak AM, et al. Leveraging large language models to generate clinical histories for oncologic imaging requisitions. Radiology 2025 Feb;314(2):e242134. [doi: 10.1148/radiol.242134] [Medline: 39903072]
20. Terzis R, Kaya K, Schömig T, et al. GPT-4 for automated sequence-level determination of MRI protocols based on radiology request forms from clinical routine. Eur Radiol 2026 Feb;36(2):1541-1552. [doi: 10.1007/s00330-025-11888-4] [Medline: 40779162]
21. Tavakoli N, Kim D. AI-generated clinical histories for radiology reports: closing the information gap. Radiology 2025 Feb;314(2):e243910. [doi: 10.1148/radiol.243910] [Medline: 39903081]
22. Abedi A, Tofighi S, Salehi S, Latterman PT, Basques KD, Gholamrezanezhad A. Reason for Exam Imaging Reporting and Data System (RI-RADS): a grading system to standardize radiology requisitions. Eur J Radiol 2019 Nov;120:108661. [doi: 10.1016/j.ejrad.2019.108661] [Medline: 31610322]
23. Asgari E, Montaña-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. NPJ Digit Med 2025 May 13;8(1):274. [doi: 10.1038/s41746-025-01670-7] [Medline: 40360677]
24. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 2008 May;61(Pt 1):29-48. [doi: 10.1348/000711006X126600] [Medline: 18482474]
25. Çamur E, Cesur T, Güneş YC, et al. Evaluating large language models for imaging modality selection: potential to reduce unnecessary contrast agent use and radiation exposure. Clin Imaging 2025 Sep;125:110573. [doi: 10.1016/j.clinimag.2025.110573] [Medline: 40752449]
26. Kanemaru N, Yasaka K, Okimoto N, et al. Efficacy of fine-tuned large language model in CT protocol assignment as clinical decision-supporting system. J Imaging Inform Med 2025 Dec;38(6):4336-4348. [doi: 10.1007/s10278-025-01433-6] [Medline: 39909993]
27. Tan JR, Lim DYZ, Le Q, et al. ChatGPT performance in assessing musculoskeletal MRI scan appropriateness based on ACR appropriateness criteria. Sci Rep 2025 Feb 28;15(1):7140. [doi: 10.1038/s41598-025-88925-1] [Medline: 40021679]
28. Triana BP, Wiggins WF, Befera N, Roth C, Cline B. Proof-of-concept prompted large language model for radiology procedure request routing. J Vasc Interv Radiol 2025 Jul;36(7):1201-1207. [doi: 10.1016/j.jvir.2025.03.012] [Medline: 40139433]
29. Akinci D'Antonoli T, Stanzione A, Bluethgen C, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. Diagn Interv Radiol 2024 Mar 6;30(2):80-90. [doi: 10.4274/dir.2023.232417] [Medline: 37789676]
30. Kao JP, Kao HT. Large language models in radiology: a technical and clinical perspective. Eur J Radiol Artif Intell 2025 Jun;2:100021. [doi: 10.1016/j.ejrai.2025.100021]
31. Omar M, Sorin V, Collins JD, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Commun Med (Lond) 2025 Aug 2;5(1):330. [doi: 10.1038/s43856-025-01021-3] [Medline: 40753316]

Abbreviations

EMR: electronic medical record
LLM: large language model
MERs: MRI examination requests
MRI: magnetic resonance imaging
RAG: retrieval-augmented generation
RI-RADS: Reason-for-Exam Imaging Reporting and Data System
sLLM: secure large language model

XSL•FO
RenderX

Edited by A Coristine; submitted 25.Aug.2025; peer-reviewed by A Shivanna, BL Guellec, I Odezuligbo; revised version received 20.Feb.2026; accepted 24.Feb.2026; published 07.Apr.2026.

Please cite as:
Hallinan JTPD, Leow NW, Low YX, Lee A, Ong W, Chan MDZ, Devi GK, He SS, Loh DDL, Lim DSW, Low XZ, Lim MC, Yong C, Sng WJ, Teo EC, Tan JH, Kumar N, Makmur A, Ting Y
Initial Insights Into an Institutional Secure Large Language Model for Magnetic Resonance Imaging Examination Requests: Retrospective Study
J Med Internet Res 2026;28:e82579
URL: https://www.jmir.org/2026/1/e82579
doi:10.2196/82579

© James Thomas Patrick Decourcy Hallinan, Naomi Wenxin Leow, Yi Xian Low, Aric Lee, Wilson Ong, Matthew Ding Zhou Chan, Ganakirthana Kalpenya Devi, Stephanie Shengjie He, Daniel De-Liang Loh, Desmond Shi Wei Lim, Xi Zhen Low, Mei Chin Lim, Clement Yong, Weizhong Jonathan Sng, Ee Chin Teo, Jiong Hao Tan, Naresh Kumar, Andrew Makmur, Yonghan Ting. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Views of People With Psychosis About Algorithm-Based Relapse Prediction and Data Sharing: Qualitative Study

Emily Eisner[1,2], PhD; Hannah Ball[1,2], DClinPsy; John Ainsworth[3], PhD; Matteo Cella[4,5], PhD; Richard J Drake[1,2], PhD; Daniel Elton[6], MA; Sophie Faulkner[1,2], PhD; Kathryn Greenwood[7,8], PhD, DClinPsy; Andrew Gumley[9,10], PhD; Gillian Haddock[1,2], PhD; Kimberley M Kendall[11], PhD; Alex Kenny[6]; Tor-Ivar Krogsæter[6], MA; Jane Lees[1], PhD; Shôn Lewis[1,2], MD; Alie Phiri[6], MSc; Matthias Schwannauer[12,13], PhD; Rebecca Turner[1], PhD; Annabel E L Walsh[6], DPhil; James Walters[11], PhD; Til Wykes[4,5], PhD; Uzma Zahid[4], PhD; Sandra Bucci[1,2], DClinPsy

[1]Division of Psychology and Mental Health, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Jean McFarlane Building, Manchester, United Kingdom

[10]NHS Greater Glasgow and Clyde, Glasgow, United Kingdom

[11]Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, Cardiff, United Kingdom

[12]School of Health in Social Science, University of Edinburgh, Edinburgh, United Kingdom

[13]NHS Lothian, Edinburgh, United Kingdom

[2]Greater Manchester Mental Health NHS Foundation Trust, Manchester, United Kingdom

[3]Division of Informatics, Imaging and Data Sciences, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, United Kingdom

[4]Department of Psychology, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom

[5]South London and Maudsley NHS Foundation Trust, London, United Kingdom

[6]McPin Foundation, London, United Kingdom

[7]School of Psychology, University of Sussex, Falmer, United Kingdom

[8]Sussex Partnership NHS Foundation Trust, Hove, United Kingdom

[9]School of Health and Wellbeing, University of Glasgow, Glasgow, United Kingdom

Corresponding Author:
Sandra Bucci, DClinPsy
Division of Psychology and Mental Health, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Jean McFarlane Building, Manchester, United Kingdom

Abstract

Background: Preventing relapses of psychosis is difficult and important. Digital remote monitoring (DRM) systems are being developed and tested to support this. Increasingly, these systems use algorithm-based relapse prediction. Hence, understanding stakeholder views about algorithmic prediction is crucial. Existing qualitative work has explored health professionals' views, but very few studies have examined the perspectives of people with psychosis on this topic.

Objective: This paper aimed to provide an in-depth examination of the views of people with psychosis regarding algorithmic relapse prediction within a DRM system that incorporates active symptom monitoring and passive sensing data.

Methods: People with psychosis (n=58) were recruited from 6 geographically distinct areas of the United Kingdom. They participated in semistructured qualitative interviews exploring their views about using a DRM system that predicts psychosis relapse based on a machine learning algorithm. Transcripts were analyzed using reflexive thematic analysis. People with lived experience of psychosis were involved extensively in study design, analysis, and reporting.

Results: Findings were described across 4 themes. First, *accuracy* was a prominent theme. Participants emphasized that transparency about algorithm sensitivity and specificity is crucial and discussed the risks of the relapse prediction algorithm producing false positives (flagging that someone was relapsing when they were not) and false negatives (missing actual relapses). In both cases, participants said that errors may be partially mitigated through a *human-in-the-loop* approach (theme 2), with DRM blended with human oversight, from clinicians or a dedicated digital monitoring team, and calibrated based on service user, carer, and clinician feedback. The third theme, *trust, fears, and choice,* noted the interplay between users' trust in the DRM system and their relationship with the clinical team. This theme described participants' fears about potential overreactions (hospitalization or excessive medication) or underreactions (no additional support) from the clinical team in response to algorithm-generated relapse predictions. It emphasized the importance of retaining choice around the use of relapse detection algorithms and the sharing of personal data. The final theme described participants' views about the *benefits of using a relapse prediction algorithm*,

XSL•FO
RenderX

including facilitating early intervention, triaging care according to need, minimizing human bias in assessment, and efficiency in saving staff time.

Conclusions: People with psychosis acknowledged potential benefits of algorithm-assisted relapse prediction for receiving timely or efficient care, but with several caveats. Algorithm-generated relapse alerts need to be sufficiently accurate and must be interpreted, with understanding of their limitations, by a trustworthy human who is aware of the relevant context. Algorithm-based relapse predictions should only be used with valid consent, in a way that promotes and respects the autonomy and voice of service users and avoids increasing the use of excessive restriction.

(J Med Internet Res 2026;28:e86753) doi:10.2196/86753

KEYWORDS

psychosis; mobile app; relapse; passive sensing; remote monitoring; wearables; machine learning; prediction; prevention; qualitative; acceptability

Introduction

Although approximately half of individuals experiencing a first episode of psychosis reach symptom remission [1], around half experience a relapse within 3 years [2]. Relapses are linked to significant adverse outcomes for individuals, including disrupted educational or occupational trajectories [3,4], reduced functioning [4,5], poorer clinical prognoses [6,7], increased mortality risk [8-10], and considerable financial costs to health services, primarily due to inpatient admissions [11-14]. Standard approaches to community-based mental health monitoring typically depend on individuals recalling symptoms from previous weeks or months during infrequent clinical appointments. Due to recall bias, these retrospective self-reports often lack accuracy, delaying the timely detection of emerging relapse indicators. Given the need for prompt and precise intervention in psychosis, it is important to detect relapses early and intervene swiftly, without further burdening overstretched mental health services or individuals with psychosis.

To meet this need, the multisite CONNECT study involves co-designing and evaluating a digital remote monitoring (DRM) system that uses a machine learning (ML)–generated algorithm to predict relapse. As the algorithm is still under codevelopment, detailed specifications cannot yet be disclosed. The algorithm will integrate information from multiple data streams, including active symptom monitoring (ASM) and passive sensing. ASM uses self-reported smartphone questionnaires to track symptoms [15-17]. Passive sensing uses unobtrusively collected sensor data (eg, GPS and accelerometry) to infer behavioral and social functioning (eg, movement patterns and sleep disturbances) without requiring user input [18-20]. The resulting DRM system will monitor mental health status and alert service users and/or clinicians to early signs of relapse, aiming to facilitate timely intervention and mitigate the potentially severe consequences of a full relapse.

Given the well-recognized challenges of integrating new technologies into health care, digital systems must be developed with implementation considerations at the forefront [21-23]. End users are central to this process; their engagement ultimately determines whether such systems are adopted and maintained [24]. Therefore, understanding the perspectives of people with psychosis from the start is critical to ensuring that DRM systems are not only acceptable but also address genuine gaps in care. We conducted qualitative interviews with a large, diverse sample of individuals with psychosis, recruited from multiple National Health Service (NHS) mental health services across 6 distinct UK regions. Interviews explored participants' hypothetical views on a DRM system combining ASM and passive sensing, with data from both sources contributing to an algorithm to support relapse prediction. Given the breadth and richness of the interview data, findings for each DRM component (ASM, passive sensing [25], and relapse prediction algorithm) are reported separately across three publications.

This paper focuses on the views of people with psychosis about using an ML-generated relapse prediction algorithm as part of their clinical care, including potential benefits and risks of sharing algorithmic predictions with clinical teams, family members, and significant others. While previous studies have focused on clinicians' views on ML in mental health care [26-29], only 1 study has qualitatively examined the views of people with psychosis on artificial intelligence (AI)–driven relapse prediction [30]. Unlike that study, which primarily gathered insights from online support group members regarding Facebook (Meta) data, this study provides the first in-depth examination of the views of people with psychosis about potential challenges and benefits of using a relapse prediction algorithm based on ASM and passive sensing data.

Methods

Design

This qualitative interview study formed part of the broader CONNECT project, a multisite research program funded by the Wellcome Trust [31]. Conducted during phase 1 of the program, it aimed to inform the design of the CONNECT DRM system. This manuscript adheres to the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines (Checklist 1) [32].

Ethical Considerations

Ethical approval was secured through the Health Research Authority and West of Scotland 4 Research Ethics Committee (reference 22/WS/0083). All participants gave informed consent, either in writing (signed form) or verbally (audio-recorded), and were assigned unique identification numbers to protect anonymity. They were compensated £20 (equivalent to US $24.74 at the time of the study) for their time.

XSL•FO
RenderX

Participants, Sampling Strategy, and Recruitment

Recruitment occurred between November 15, 2022, and November 13, 2023, using the following inclusion criteria: (1) English fluency, (2) age ≥16 years, (3) current contact with mental health services, (4) capacity to provide informed consent, (5) clinical stability sufficient to participate in an interview (based on clinician or researcher judgment), and (6) either a formal diagnosis of schizophrenia spectrum disorder (*ICD-10* [*International Classification of Diseases, Tenth Revision*] F20–29 [33]) or fulfillment of early intervention service entry criteria, operationalized using the Positive and Negative Syndrome Scale [34] and/or the Comprehensive Assessment of At-Risk Mental States psychosis transition criteria [35].

A purposive sampling strategy was implemented to ensure diversity in geography, age, gender, ethnicity, type of service, clinical profiles, and previous exposure to digital health technologies. Participants were recruited from mental health services across 6 geographically distinct areas in the United Kingdom (Manchester, Glasgow, Edinburgh, London, Sussex, and Wales). Rather than aiming for data saturation, we aimed to recruit approximately 10 individuals per site to ensure broad geographic coverage and a large overall sample. As Braun and Clarke argue [36], data saturation is not relevant to all types of thematic analysis, particularly reflexive thematic analysis, used in this study. The specific target of 10 participants per site was a pragmatic decision; we considered this a feasible number to recruit with the time and resources available.

Clinical staff from acute inpatient units and community-based mental health teams (eg, Community Mental Health Teams, Early Intervention Teams, and Crisis Resolution and Home Treatment Teams) were asked to identify eligible individuals and provide them with initial information about the study (verbally or using a study leaflet). Interested individuals could then contact the researcher directly or provide verbal or written consent for clinical staff to pass their contact details to the research team. Study adverts were also displayed in secondary care mental health services and sent to participants from existing research cohorts, or other ethically approved studies, who had given prior consent to be contacted about research. The study advert invited interested individuals to contact the researcher directly. Research staff then provided all interested individuals with further information about the study, including a participant information sheet, verbal explanation, and the opportunity to clarify questions or concerns. Potential participants were given at least 24 hours to consider the information before deciding whether to participate. Those choosing to participate were then asked to provide formal consent via a paper, online, or emailed consent form or via an audio-recorded consent statement before the interview.

Qualitative Interviews

Each consenting participant took part in a single, audio-recorded, semistructured interview, conducted in person, by telephone, or through secure videoconferencing, depending on individual preference and safety considerations. In addition to questions for eliciting participants' views, the topic guide (S1 in Multimedia Appendix 1, and Figure S1a, Figure S1b, and Figure S1c in Multimedia Appendix 1) included plain-language explanations of key concepts, such as how an ML algorithm would be developed and used within a DRM system to predict relapse. Explanations included examples of well-known platforms that use ML algorithms (eg, YouTube [Google] and Netflix), a step-by-step lay description of supervised ML learning procedures, and a description of using the algorithm to alert clinicians of imminent relapse risk. Figures S1b and S2c in Multimedia Appendix 1 were shown to the participant at a relevant point in the interview; depending on the interview format, they were shown on paper (in person), via screensharing (video call), or posted or emailed ahead of the interview (phone call). Interview duration ranged from 20 to 78 (median 43) minutes. Participants also completed a demographic questionnaire. Interviewers completed short postinterview notes reflecting on context and recommending any necessary adaptations to the guide. All audio data were transcribed verbatim, anonymized, and securely stored.

Data Analysis

Interview transcripts were analyzed inductively, using standard reflexive thematic analysis steps [37]: (1) data familiarization, (2) code generation, (3) initial theme creation, (4) revision of themes, (5) defining final themes, and (6) report writing. While coding was primarily data-driven, a subset of questions reflecting specific considerations (eg, whether participants would want family to receive algorithm-generated relapse alerts) was explored deductively, a common approach within digital health research [38]. Transcript excerpts were also systematically labeled according to whether views expressed related specifically to the ML algorithm, or to other aspects of the DRM system; this paper reports themes related to ML. Analysis was supported by NVivo software (Lumivero) [39]. The first author (EE) conducted primary coding, incorporating coding contributions from 4 Lived Experience Advisory Panel (LEAP) members and 2 McPin supervisors (AELW and AK), who examined a subset of transcripts. As several individuals coded the same interview transcripts, they met to compare their coding and resolve any discrepancies. Theme development was collaborative, shaped through discussions with (and written feedback from) the LEAP members and the broader research group to ensure credibility, practical relevance, and consistency with participants' accounts. The study was grounded in a critical realist perspective [40,41], further details of which are available in S3 in Multimedia Appendix 1, alongside reflexivity details.

Lived Experience Involvement

Lived experience involvement, supervised and coordinated by AELW and AK at The McPin Foundation [42], is central to the CONNECT study. A key component is the CONNECT LEAP, comprising 12 individuals with personal or familial experience of psychosis, drawn from the 6 study sites. This panel contributed to ongoing project governance and reviewed a summary of the study's findings. All LEAP members were invited to attend a 2-hour introduction to qualitative methods training session, facilitated by McPin, outlining qualitative research methods and introducing the 6 key steps of thematic analysis. In total, 6 individuals with lived experience of psychosis contributed more substantively to the study design and analysis. Of these, 2 individuals read a draft of the topic

XSL•FO
RenderX

guide and provided detailed written and verbal feedback to improve the clarity of lay explanations and interview prompts before use in the interviews; the other 4 were involved directly in qualitative analysis, with 3 being coauthors on this paper (one did not want to be named). After practicing coding in a synchronous session with the researchers and McPin supervisors, these individuals independently coded 3 interview transcripts. In 3 further synchronous sessions, they fed back their coding insights, advised on broader code development, and critiqued draft themes and emerging interpretations. Finally, they commented on drafts of this manuscript.

Results

Overview

Participants (n=58) had a mean age of 39.1 (SD 14.1) years. Just over half were men (n=32), a third (n=20) were from an ethnic minority group, and three-quarters (n=42) were single. Around half (n=27) were employed, in voluntary work, or studying, and around three-quarters (n=41) had received education after 16 years of age. Primary self-reported diagnosis was typically psychosis or schizophrenia, with current care usually community-based. S3 in Multimedia Appendix 1 illustrates full characteristics. Interview findings are described across 4 overarching themes, each with several subthemes, as summarized in Figure 1 and outlined below with selected quotations. S4 in Multimedia Appendix 1 provides the full coding tree, with additional quotations. For context, participant identification numbers that accompany the quotations are prefaced by a letter indicating their geographical region: Manchester (M), Glasgow (G), Edinburgh (E), London (K), Sussex (S), and Wales (C).

First, *accuracy* was a prominent theme, with participants discussing the risks of the relapse prediction algorithm producing false positives (flagging that someone was relapsing when they were not) and false negatives (missing actual relapses). In both cases, participants said that errors may be partially mitigated through a *human-in-the-loop* approach (theme 2), where DRM is blended with clinical oversight. The third theme, *trust, fears, and choice*, notes the interplay between users' trust in the DRM system and their relationship with the clinical team. This theme describes participants' fears about potential over- or underreactions from the clinical team in response to algorithm-generated relapse predictions. It emphasizes the importance of retaining choice around the use of relapse detection algorithms and the sharing of personal data. The final theme describes participants' views about the *benefits of using a relapse prediction algorithm*.

XSL•FO
RenderX

Figure 1. Thematic map.

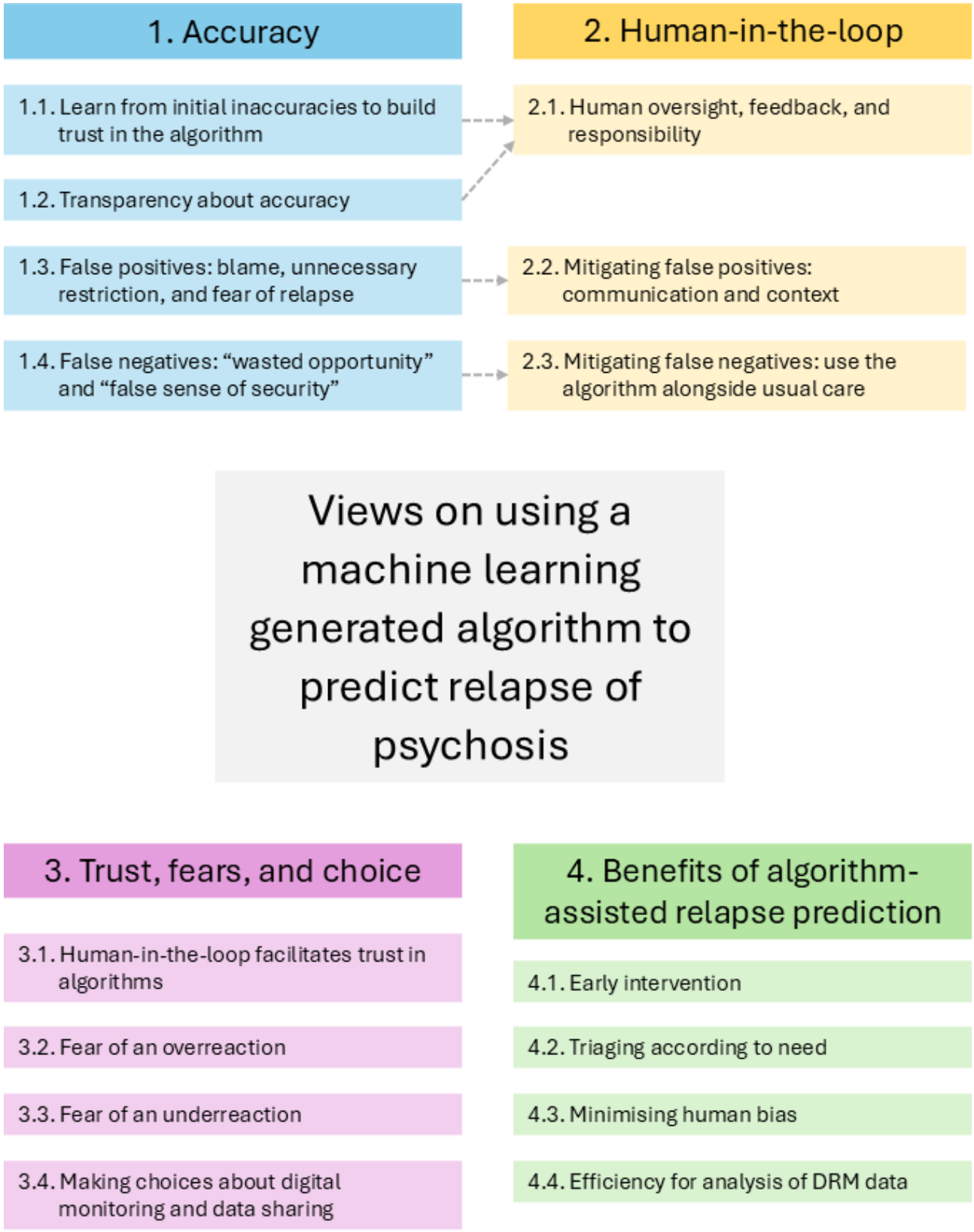

Theme 1: Accuracy

Learn From Initial Inaccuracies to Build Trust in the Algorithm

When discussing the development and use of relapse prediction algorithms, participants expressed some tolerance for errors early on. Initial inaccuracies were often considered "inevitable" (M011) and "understandable" (E007). Participants acknowledged that "it would take time…to perfect [the algorithm], so there would obviously be mistakes in the infancy of it" (E007), but generally accepted this as a normal part of software development and refinement:

XSL•FO
RenderX

> *I suppose if any like, um, new Software or or new data management system, there are going to be teething problems.* [C005]
>
> *You don't get everything right…But eventually it will it's just taking a while, like it took YouTube years to get its algorithm right.* [C007]

However, participants also conveyed a strong expectation that the algorithm should get more accurate over time, acknowledging that, to achieve this, "it'll probably have to go through some sort of trial period" (K004):

> *…it's in its early stages of working, there is gonna be problems…unless you work on it and iron out the problems…And you can only do that once you start to collect the information.* [E004]

Participants emphasized that repeated inaccuracies in the relapse prediction algorithm could undermine users' trust in the system over time. One participant likened it to "…the boy who cried wolf!" (S002), suggesting that too many false alarms would cause both people with psychosis and clinicians to lose confidence in the algorithm's reliability and disregard its alerts:

> *It'd put doubt in their head…with people with mental health it's all about that trust…so instantly that trust is lost.* [C007]
>
> *If [clinicians] then get sent this every single day and sometimes it's misinformation as well, they're just probably going to be reluctant to even take note of it…it might come through but they might be like, "oh, you know, this program, it's not worked before, so how can we sort of trust it?" So I think…trusting the actual technology is really important.* [C002]

One participant questioned the value of using an algorithm-based system at all if it fails to detect, or misrepresents, key information:

> *If the program makes up or doesn't catch information, then why are you working with the program if it doesn't help?* [K010]

Another reflected on their experience with a commercially available wearable device's stress-detection feature, noting that repeated inaccuracies eventually led them to stop using it:

> *I do just not really engage with it, so I do wonder if it would be similar [if the CONNECT algorithm was inaccurate]. If I might just feel, this isn't useful, so maybe stop engaging.* [G001]

Some participants stressed that the system should be "rigorously tested" (K003) and "trained trained trained trained trained" (S002) due to potential impacts on health:

> *This is a very serious issue, so it should be checked checked checked.* [S002]
>
> *This is serious, it's schizophrenia…it's people's mental health and it's not a joke, is it?* [S002]
>
> *[Errors] would stop me from using it… 'cos that would be a kink in the system. It would be my whole health, mental health as well as my physical health and everything else…iPhones are a communication device. A breakdown in communication, that would stop me from using it.* [E003]

Similarly, a few would only trust the algorithm if it was "fail proof: mistakes cannae happen" (G003), with a minority remarking that algorithm-based systems could *never* be accurate enough for them to trust:

> *At some point this will lead to miscommunication or misunderstanding…The app is meant to report how your relapse is or how you're feeling at the time. I would like to say there would never be an accurate communicative line through any form of apps.* [M002]

Transparency About Accuracy

To foster realistic expectations and maintain trust, participants emphasized that people with psychosis and clinicians should be informed up-front of the algorithm's inherent (known) limitations:

> *Maybe even a statement saying it might not be 100% accurate but the app is continuously trying to make improvements.* [C004]
>
> *Realistically everyone should know it's not always gonna get it right, and as long as we have that honest conversation and say…if this was to happen then we can, can talk about it.* [M011]

Hence, if DRM users are aware of the algorithm's fallibility from the outset, they may be less likely to lose trust in, or abandon, the system if errors occur. Participants underlined the need for transparent, pre-established protocols and procedures outlining how errors will be communicated and managed, ensuring users know what to expect:

> *I think that should be shared with people using it, that it might not get things right and how to go about dealing with that when it doesn't get things right.* [K004]

Participants overwhelmingly recommended prompt, transparent communication if errors do occur: "tell them…straight away" (M010), be "open and honest" (C007), and "come clean" (K009) about the error. The exact format of communication was considered unimportant:

> *...whether it's a phone call, text or letter…as long as you can tell someone as soon as possible…the way you tell them is sort of up for debate really.* [M010]

Participants underlined a need to "obviously apologize" (K009), "take responsibility…cos that person's care has been affected," and to consider the individual's feelings about the error, "reassuring the actual person involved" (K009), "keeping channels open, and maybe actively leaning in, how people are finding things" (G001). Finally, the individual should be told what will happen in response; for example, "retraining" (K001) the algorithm, or taking steps to ensure that "it's not something that's gonna go on your record" (M011). The latter participant was very clear that they would only agree to use DRM if they knew that inaccurate information would not be kept long-term (eg, in their electronic health record). Hence, if the relapse prediction algorithm's output is routinely transferred to health

XSL•FO
RenderX

records, a robust and reliable procedure to correct inaccuracies would be crucial.

False Positives: Blame, Unnecessary Restriction, and Fear of Relapse

Participants' varied expectations about how alerts from an algorithm-based system would be managed and interpreted appeared to underpin their opinions on how problematic false positives would be. If a human did not oversee the relapse prediction algorithm, or alerts were accepted uncritically by clinicians as "truth," participants worried that false positives would result in DRM users feeling "blamed or singled out for something that I haven't done" (E009), being told "there was a big risk or a big danger when there wasn't" (K001), receiving unnecessary treatment or restriction (refer to "Fear of Overreaction" section), and experiencing fear of relapse. Conversely, those who expected that algorithm-generated alerts would lead to a thoughtful interaction with a trusted other were less concerned about possible false positives (refer to "Mitigating False Positives" section).

Fear of relapse was a concern for some, particularly in relation to the potential psychological impact of algorithm-generated predictions. Some worried that false positives could cause a "feedback loop" (S008) or "self-fulfilling prophecy" (K001), whereby being incorrectly told they were relapsing might generate unnecessary stress, prompt a fear of relapse, and potentially trigger actual deterioration:

> *If it said you were going to relapse when you actually weren't, you might actually get into your head like 'oh I'm going to relapse' and then you're expecting it to happen…It's more likely to happen when you're thinking about it like that. When you've got it on your mind constantly.* [M012]

Some participants preferred not to receive relapse alerts directly from the app to avoid potential distress caused by false positives (refer to "Making Choices About Digital Monitoring and Data Sharing" section). They preferred alerts to be routed first to a clinician or a trusted other, who could decide whether and how to communicate the information, reflecting the desire for a human-in-the-loop mediation to reduce distress and contextualize algorithmic outputs.

> *That's one of the reasons I think they should contact the, the mental health team and then them contact my trusted ones, because seeing my mental health is ok and then have app tells me your mental health's not ok…my mind start racing thinking what am I doing wrong, why is this prompting me…So saying my mental health wasn't ok, I would like to know through a person.* [G005]

Conversely, despite initial worry, 1 participant described feeling confident that, with time, they could judge themselves whether they were unwell:

> *I would be really concerned…But I think I know myself enough that praps after a day or so and I'm still alright…I'd know.* [S007]

Others would find false positives unconcerning and would appreciate the "warning" (G008) as a prompt to "check in" (M005), "think twice" (M010), and pay more attention to their mental health:

> *…probably I'm just, will be more alert to my situation…sometimes it's good to have the alert even though it's might be not accurate.* [G009]

False Negatives: "Wasted Opportunity" and "False Sense of Security"

The value of a digital relapse prediction system lies in its ability to detect early warning signs *before* an individual necessarily becomes aware of them. If an individual notices their own relapse signs first, participants felt that the system is "not really doing the job it was designed for" (G007):

> *I want to know if there is going to be a relapse, because then I can do something. If I don't know, it's a wasted opportunity… If I know beforehand, I could make a real difference to outcomes and experiences for myself.* [K001]

Hence, false negatives were a concern because of the missed opportunity to intervene early in the relapse process, meaning that "people are falling through the gaps" (S007). Participants noted that intervening early could improve outcomes, but delays from false negatives were potentially harmful, allowing symptoms to escalate unchecked and increasing the risk of a more severe episode:

> *[False negatives are] worse for the…person in the long run. They could become very ill, you know, nothing's done but it could be too late before any action's taken.* [K009]

Similarly, participants warned that an algorithm that generated false negatives may provide "a false sense of security" (E009) whereby "you would keep getting worse, thinking you were getting better" (G005). This may delay help-seeking until a point at which individuals are no longer willing to seek help:

> *...you might get too deep into it and not want to reach out anymore.* [M012]

One participant, whose voices did not like her talking about her mental health, noted that false negatives would be counterproductive because they would "reinforce…what the voices are saying…what they want" (S007). Moreover, false negatives would waste the user's efforts to track their symptoms:

> *If you've filled all these specific questions and gone through a specific process…the whole point of it is…to pick up on [signs of relapse] before you do, so if you realise a bit before the computer has, it's almost pointless and…that would eliminate the purpose of it in the first place.* [M010]

Theme 2: Human-in-the-Loop

Overview

Participants frequently advocated a human-in-the-loop workflow to improve the algorithm's accuracy (refer to "Human Oversight, Feedback, and Responsibility" section) and to manage the

XSL•FO
RenderX

potential consequences of false positives (refer to "Mitigating False Positives" section) and false negatives (refer to "Mitigating False Negatives" section).

Human Oversight, Feedback, and Responsibility

Participants emphasized the importance of "human guidance and supervision" (C005), "safeguards" (C005), or "human oversight" (C005) for the relapse prediction algorithm. They suggested incorporating a feedback mechanism to allow the individual with psychosis, clinicians, or trusted others to provide information on the algorithm's accuracy and to "calibrate" (G005) it where necessary:

> *A facility for maybe friends or family…to question whether it's right or not. Or…the individual, if they're in the right frame of mind…could give feedback…If you do feel that [the alerts are] out of sync with what is actually going on with you as an individual, give us feedback and let us know.* [C005]

Being able to provide feedback themselves would help some participants feel comfortable using an algorithm-based system. For example, participant G001 was willing to try a DRM system "as long as I felt I was able to have my view taken into account." Others saw a feedback mechanism as just "a very practical thing" (G002) to help train the algorithm and improve accuracy.

Participants generally assumed that their clinician would oversee algorithm-generated alerts and provide feedback on their accuracy, integrating this process into usual care. Although this arrangement would have the advantage that alerts could be overseen by "people that know me" (C003), clinicians may not know how to interpret the data:

> *The odds of your mental health team knowing everything about CONNECT and how to interpret data are probably quite small. I would rather somebody that actually knows about the data and what that can or can't say.* [G001]

Hence, several participants suggested that algorithm-generated alerts should be automatically sent to a dedicated clinical review team, whose deliberations would confirm or refute each alert (or a subset):

> *You should employ people to monitor that…So whenever the program makes a discovery or a pattern, that gets notificated to the team that you make, and then that team debates whether that is right or wrong…And then that helps the program work out things better as it goes along.* [C006]
>
> *If you've got self-learning AI or if you're doing machine learning, then it's more likely that those mistakes will be used to correct it. I think as long as you have a team that are evaluating it every x amount of days, weeks, then I think it's fine.* [M008]
>
> *If you've got self-learning AI or if you're doing machine learning, then it's more likely that those mistakes will be used to correct it. I think as long as you have a team that are evaluating it every x amount of days, weeks, then I think it's fine.* [M008]

This "oversight committee" (C005) could safeguard decision-making, provide "an ethical and moral point of view" (C005), and supply iterative feedback that progressively improves the model's accuracy. Participant C005 noted the potential ethical and political complexity around who should convene and manage such a committee. They suggested that, if a private company owned the DRM system, additional oversight by an external party would be needed:

> *I would be slightly more concerned if it was…information being collected and then analyzed and decisions being made by private individuals, because…most CEOs and companies that own AI facilities are not [publicly] elected…so…if you are going to use AI machine learning there would need to be, um greater oversight via either like NGOs or you know have intergovernmental organizations just keeping an eye on it. If it is going to be run by private corporations.* [C005]

This quote raises questions around who is ultimately responsible for such tools and their impacts. As one participant noted,

> *There needs to be some human involvement…because…the computer program might make mistakes that a human would have to, someone would have to answer for.* [K007]

Mitigating False Positives

"We Can Talk About It"

Participants who expected that algorithm-generated relapse alerts would be "supervised by a human" (E008) and used to prompt clinicians to gather further information tended to regard false positives as relatively unproblematic:

> *If it thinks that I'm getting unwell and my mental health team get in contact…I can tell them that I'm fine.* [M004]

Most participants thought that algorithm-generated alerts should prompt clinicians to contact them directly: "a wee telephone call fae your [community psychiatric nurse] asking you how you're doing" (E004). They emphasized that this would facilitate "openness [which] is quite important to have" (G008) and that this contact could mitigate the effects of false positives by allowing clinicians to "talk to you…to kinda confirm what the data from the phone says" (E007). Participants were clear that relapse algorithms should "not solely just being left to make decisions on its own" (C005). Instead, there was a strong sense that participants viewed algorithm-generated alerts as provisional information, to be confirmed by direct "human interaction" (S002):

> *If the…technology suspects there is something wrong then you should be seeing the psychiatrist, or the psychiatric nurse to make a further assessment to check.* [E008]
>
> *The app can be a way to notify someone that there's an issue and then the person will come in and you can have that discussion.* [M005]

Consequently, several participants considered false positives to be less concerning than false negatives because "obviously

XSL•FO
RenderX

if it picks it up and it's wrong…we can talk about that" (S001). However, one of the same participants noted that the person with psychosis may not always be aware that they are relapsing, creating ambiguity as to whether the algorithm or the person themselves is correct. This reflects the need for collaborative interpretation between the individual and their clinician to negotiate meaning and action when predictions arise:

> *I'd have to sit down with the doctor, explain to them why I'm not feeling like that and hope they understand, really. Or I need to face facts that I am getting ill and the computer is right, maybe it's not wrong.* [S001]

Moreover, while some participants acknowledged that human oversight could help mitigate errors, others expressed concern that frequent false positives might undermine the system's credibility, leading users to take the system less seriously over time (refer to "Learn From Initial Inaccuracies to Build Trust in the Algorithm" section). They also emphasized that long-term trust in the system would depend not only on the algorithm's accuracy, but critically on the strength of the relationship between the person with psychosis and their clinical team (refer to "Human-In-The-Loop Facilitates Trust in Algorithms" section).

Contextualizing Passive Data

Participants frequently warned that relapse prediction algorithms that rely on decontextualized passively sensed data may be prone to false positives. They gave examples such as that of an elevated heart rate (readings from exercise could be misclassified as stress) or that prolonged time spent at home might be flagged as social withdrawal, even when unrelated to mental health. Participants suggested that the risk of false positives from passive sensing could be reduced if algorithmic outputs were reviewed by someone who knows the individual personally (ie, who has additional context):

> *It's tied to people that know me, so they'll know better if I am not ok… for example, two or three times a week I'll just sleep till like 1pm and I'm not relapsing or anything, I'm just enjoying a little break.* [C003]

Interviews revealed a shared perception that relapse prediction algorithms based on passively sensed data were more prone to false positives than those using ASM data. Participants may view ASM as more reliable because it reflects self-reported experiences, offering direct insight into mental health status, whereas passive data can be ambiguous without context. Several participants suggested that existing (E004) or additional (G005) patient-reported ASM data could provide the context for passive data, helping to clarify when behavioral patterns were benign or clinically significant:

> *…here's my monthly input here for January, for example, and as you can see I've been fine so why is that highlighting that I'm needing help?* [E004]
>
> *About the phone knowing that I'm home…it could…pop up a message: "are you alright?…Are you enjoying staying at home or…staying home for another reason…that you'd like to share?"* [G005]

In sum, accuracy could be improved by contextualizing passive data using subjective self-report, supplemented by confirmation or refutation by known others.

Mitigating False Negatives

Some participants felt that false negatives were less concerning if an algorithm-based system was used alongside usual care, as "just an additional thing" (G009), rather than being relied on to replace clinical judgment:

> *You can always seek help…if you know yourself. If you don't know, then it's a psychiatrist's job to work it out, it's not the machine's job. It couldn't be used as a replacement for a psychiatrist.* [E001]
>
> *If it misses you're deteriorating then that would be the same if there wasn't an app…it's only offering things that are better than they're currently there… it's not making anything worse.* [G008]

These participants considered it a bonus if an algorithm provided information that helped them spot relapses but did not want to entirely delegate responsibility to the system.

> *As long as there's not…a hundred per cent dependency and then there's other avenues of help still available, I think it's fine.* [M008]
>
> *It's down to me really isn't it to reach out, I can't rely on everything to do that for me.* [S006]

Theme 3: Trust, Fears, and Choice

Human-in-the-Loop Facilitates Trust in Algorithms

The nature of the relationship between the individual and their clinician appeared to play a key role in facilitating trust in algorithm-based systems. As participant E003 summarized,

> *I think it can be trusted just by…trusting the persons that's using it. That's the only way you can trust it. You have to be able to trust the person.* [E003]

Similarly, the first author's analysis diary (detailed notes written during coding and analysis) noted:

> *The relationship with the care team is key. It takes a long time to build trust.* [analysis diary: G003]
>
> *It's like the app is an extension of the relationship with the team, so the willingness to engage is directly affected by the relationship with the team.* [analysis diary: general]

Three relational factors emerged as pivotal: (1) mutual trust, (2) the clinician's stance (collaborative vs authoritative), and (3) the individual's previous experiences of coercive or restrictive care. Mutual trust was repeatedly cited as crucial, particularly for handling false-positive alerts:

> *I know myself more than anyone else, so they would know my word.* [C003]
>
> *I think if your team checked in with you…I trust them…I trust them.* [G007]

Participants who had already established a trusting and collaborative relationship with their clinical team were more likely to perceive algorithm-generated alerts as helpful and were

XSL•FO
RenderX

less concerned about occasional errors. Conversely, participants who had experienced strained therapeutic relationships, been misunderstood by the clinical team, or been exposed to restrictive practices in the past appeared less willing to place trust in the use of algorithms in their clinical care. For example, participant E009 spoke of "mistrust," "suspicion of the psychiatrist and the mental health team," and a desire to "keep them off my back," linking these feelings to previous experiences of coercive care.

> *I've experienced the…bad end of it with sections and depots and…enforced things…When it got nasty with the sections and things that's when I started to think differently.* [E009]

This participant repeatedly warned that an algorithm might "incriminate me" or provide "ammunition" for clinicians:

> *I wouldn't like to incriminate myself. I wouldn't like [to] think we were on the same page and then one day they turn around to me and say "oh we're concerned about you because of this that or the next thing that we've…machine learned."* [E009]

This participant conveyed clear concern that an opaque relapse prediction alert could potentially disrupt their carefully cultivated therapeutic alliance. Noting that their alliance had recently improved, they began to tentatively express that DRM might be beneficial in this context, but immediately reconsidered, reiterating their worries about trust:

> *It might [be beneficial], I mean if you're on the same page with them like I am now, but you know like I said it's trust issue, trust and suspicion and mistrust you know.* [E009]

Similarly, another participant (C001) used less emotive terms but contrasted 2 mental health teams they had received care from one they perceived to be helpful and supportive, and the other "didn't give me the support, but they felt… what was going on with me either it was my medication or hospital," (C001). This participant anticipated that using DRM with the first team would be useful, but with the latter, it would have caused more restrictive care (hospital). These accounts highlight the influential role of the patient-clinician relationship in shaping perceptions of algorithm-driven monitoring. They provide context for the following subthemes, in which participants anticipated 2 contrasting responses to relapse prediction alerts—an overreaction, characterized by increased restrictions and control imposed by clinicians (refer to "Fear of an Overreaction" section), and an underreaction, where alerts are overlooked and no meaningful support is provided (refer to "Fear of an Underreaction" section).

Fear of an Overreaction

Some participants feared that algorithm-generated alerts would prompt "an overreaction and, in the context of mental health, an overreaction can mean unnecessary restriction and you really want to avoid that" (K001). Specific anticipated responses included enforced rehospitalization and/or medication increase:

> *I just don't want to be sectioned again…it's really hard to get out when you're sectioned.* [G007]

> *If they felt they needed to then they could…double my medication you know.* [E009]

When discussing these fears, participants often alluded to the power they perceived clinicians to have over their lives, using phrases that expressed their own felt lack of power ("against my will" [C001]) and their strong feelings on the topic ("would put me in a panic" [G002]). They also described previous experiences of hospitalization ("really distressing" [E004]) or medication side effects ("feeling tired and everything like that" [E009]), underlining the impact of feared overreactions.

Given the potential for clinicians to assert this power in response to algorithm-generated alerts, it is unsurprising that trust was also a key factor here. Individuals who trust that clinicians will carefully weigh algorithm-generated information alongside other evidence, including evidence that they seek directly from the individual themselves, may be less likely to fear an overreaction:

> *They need to speak to the individual and find out what they are struggling with. It needs to be communicated in a way that they both feel safe, and it doesn't feel like the individual or the service is going behind each other's backs…They can't just take charge from it.* [C004]

In this and other quotes, participants contrasted 2 hypothetical experiences of algorithms; first, receiving direct, collaborative communication from the clinical team, and second, receiving an unwanted escalation of care in which the individual feels surprised and out of control.

Fears of clinicians' overreactions were particularly prominent when participants discussed false positives. However, they were also sometimes mentioned when describing true positives, with challenges arising specifically when there was a discrepancy between what the DRM system reported and what the individual wanted to convey to clinicians themselves ("Only if it's contrary to what I was saying" [G002]). For example, they may wish to conceal a deterioration from clinicians:

> *Sometimes it's best not to give all the information to psychiatrists.* [E009]

> *As long as someone's honest if [the alert is] wrong or if it's right, I think it's fine, but then that brings, like, the other complication…if they are being honest or not.* [M010]

Alternatively, an individual may not realize or believe that they are unwell. Hence, even if DRM alerts prompt clinicians to speak to the individual directly, this might be an unwanted surprise; unraveling whether an escalation in care is required may not be straightforward. However, the challenge of weighing competing evidence about relapse is not unique to algorithm-generated data, as this example highlights:

> *Would you be informed it was going to be going to your care team?…Because the one day that I was really stressed was when my mum phoned my care team, did nae phone me and the next minute I had three of them waiting to section me.* [E004]

XSL•FO
RenderX

> *My mum…got the wrong end of the stick but she had to write a letter to the hospital to explain she'd got the wrong end of the stick before I was getting out…It was awful.* [E004]

Fear of an Underreaction

Some participants expressed fears that clinicians would not act on information provided by algorithm-generated relapse alerts due to viewing alerts as unimportant ("if was up to them, the mental health team would do nothing" [G002]) or inauthentic ("I'm unsure about how people would perceive it's authenticness…If people would kind of take that information on the results seriously" [K003]), especially if they were not always accurate. Others worried that receiving data from the system would increase clinicians' already large workload ("Does it not make the Doctor a lot busier?" [E003]) and that many would simply be too busy to act on the information.

> *I think it would be really, really practical. But I know it won't happen, it's not going to happen at all. It's like you cannae get a reading to the doctors, you've got to phone in and see what they've given you on the phone.* [G010]

> *Already they're stretched with what they're doing and then they've got to attend to these messages as well…So, I don't think they'd be acted on very quickly…by the clinical team…And so, you'll be left with it for days, thinking about your appointment.* [K002]

The lack of human resources to follow up on digital alerts was a particularly pressing issue for participants. K002, who repeatedly emphasized that a digital monitoring system would not scale well to real-life services, that the clinical team would not respond promptly to digital health data or predictive system outputs, and that this lack of response may stop people using the system and negatively impact individuals with psychosis. The latter concern was echoed by M011:

> *I think you've really got to work on the response and whether it's going to be suitable for a large-scale rollout…If you don't have that in place, people are going to get frustrated, and they won't use it.* [K002]

> *Maybe me feeling like somebody should help me out…I should receive a little bit of additional support and not getting that and then feeling that rejection.* [M011]

Several participants emphasized the need to carefully consider how the system would be implemented to avoid being "…overwhelming for all the workers" (C004). Suggestions included restricting the use of digital monitoring to certain days, distributing review of DRM data across staff members, limiting system checks to working hours, allocating dedicated time for staff to review digitally generated data, or establishing a dedicated digital review team (refer to "Human Oversight, Feedback, and Responsibility" section).

Making Choices About Digital Monitoring and Data Sharing

Choice About Using a Relapse Prediction Algorithm

There was near-unanimous agreement that using a DRM system incorporating a relapse prediction algorithm should be "an individual choice" (G010), and a clear consensus that making access to mental health care conditional on DRM use would be inappropriate:

> *I disagree with…gating access to mental health. I think there should be as few limitations as possible to access mental healthcare.* [M008]

> *I don't think it should be mandatory, I think it should be voluntary, and I don't think it should be tied into whether or not you get care. I would suggest that's unethical.* [K001]

> *I'd be devastated if [it was compulsory]. It's a terrible idea. It's taking away the person's independence. It's treating them like a robot controlled by a computer. It's…being blackmailed, that if you don't use this device, then you don't get treatment. It's a terrible idea.* [M003]

Participants expressed strong feelings on the topic, citing practical ("not everyone has a smartphone and a wearable" [M004]), clinical ("people might be paranoid about the machine" [E008]), and moral reasons ("sounds like a human rights violation" [G001]) against making DRM mandatory.

Choice About Sharing Algorithm-Generated Relapse Alerts With the Clinical Team

As well as having a choice to use a DRM system, participants noted that users must have a genuine choice about whether to share algorithm-generated relapse alerts, and with whom. There was some apprehension that an individual's decision not to share alerts with clinicians may be pathologized and could result in them being labeled as noncompliant.

> *The program I suppose is harmless. I just, I don't really want the information relayed back to a professional…it's not all dependent on a mental illness. Some of it's just choice I think.* [E004]

> *I just don't feel anything should be mandatory in mental health because usually there will be somebody it doesn't work for, then they will just be labelled with treatment resistance or…not willing to engage…You see that with anti-psychotics already and it would just be crap if this becomes like a new… It is good to have options.* [G001]

Participant G001's reflection highlights a wider concern that DRM may replicate existing dynamics of coercion in mental health care. Drawing parallels with antipsychotic medication, they cautioned against digital tools becoming mandatory or default components of treatment, emphasizing that a one-size-fits-all approach risks marginalizing individuals for whom such tools are ineffective or unwelcome. As with antipsychotic medication, there is a danger that nonengagement with DRM could be mischaracterized as resistance or noncompliance, rather than understood in the context of

XSL•FO
RenderX

individual needs and preferences. These concerns underline participants' broader views that digital tools should be offered as optional and flexible components of care, embedded within a framework that prioritizes autonomy, choice, and collaborative decision-making.

Participants displayed a range of views about whether they, personally, would want to share alerts from the relapse prediction algorithm with clinicians. Some were clear that they would not want to share their DRM data ("I wouldn't do it. Yeah. I don't feel desperate enough." [K002]), whereas others were optimistic about its potential benefits ("If it can tell them what to do, that would be fantastic." [K010]). Several indicated they would be willing to share alerts if they were also notified ("If my mental health was tipping and it was gonna notify someone, it would be good to know first." [C006]). Some noted they were already "quite open with my mental health team" (M004), so sharing DRM data would not be a concern. Others were more ambivalent; for example, despite being receptive to DRM and acknowledging its potential benefits, participant K004 preferred to show clinicians the data in person:

> *You could share that with them at a meeting or appointment…because then it's like a two-way conversation…Yeah, I guess that takes away the automatic side of it, and I don't know, that could be a downfall as well. Because it's in some peoples' cases, if things are getting really bad, then it is best for someone else to get involved, but I think you have to look at it case by case.* [K004]

Overall, more participants expressed willingness than resistance, despite worrying about an overreaction (refer to "Fear of an Overreaction" section) or underreaction (refer to "Fear of an Underreaction" section). Again, the level of trust with clinicians was a key factor. Participant G001 preferred not to send DRM data to clinicians, noting that "I would need to be very bad indeed before I went for help from the team…I would prefer more tolerance to stuff before information was passed on" (G001). Accordingly, this individual preferred DRM data to go "through some kind of filter first, like maybe somebody from the [research] team could maybe check in to see if there is actually a problem" (G001).

Sharing Algorithm-Generated Relapse Alerts With Family Members or Carers

When asked whether they would share relapse alerts with family, carers, or trusted others, participants repeatedly emphasized that consent from the person with psychosis is essential. Many agreed that "if it was with permission then…it's completely fine" (M008). As with clinicians, the nature of the relationship between the person with psychosis and their family, carer, or significant other was key to their willingness to share data. Participants would share algorithm-generated alerts with "people that I trust" (G005) but cautioned that it was necessary to be "aware of the family dynamics" (C005). Not everyone would have a good enough relationship with their family to share data:

> *Some family don't talk to each other and they don't care.* [K006]

Doing so without a supportive relationship could be counterproductive:

> *…if they are in close contact and they're on good terms. I think that's the key term there would be good terms, um because imagine if you didn't have a very supportive family, yeah I suppose you will feel like they're constantly interfering. Intervening when you wish they weren't.* [C005]

Some participants would need to think carefully before sharing relapse alerts with certain supportive others because alerts "could upset them" (C006), cause stress, or might affect the relationship:

> *It might be quite frightening for her to get a thing saying her husband's mood's low…and he's thinking of…suicide, I think that'd be quite a big thing to send through a text or an app…Cos she was upset even when we talk about stuff that I did in the past…so if she started getting pings about how low my mood is, I don't know.* [G007]

> *If they had information about me…which the doctors and nurses had…it would just be like…cheating on me…paranoid in the relationship - don't want that.* [E003]

One participant noted that their family was not geographically nearby, so receiving alerts might "put them in such a panic, because they are not here, they are not seeing me" (M006), whereas "if it's a relative that is close by, like, easily just driving, and just see you face-to-face, and see for themselves, yeah, that's fine" (M006). This was deemed particularly important if there might be errors in the algorithm:

> *When somebody sees you face to face they can say, oh that must really be an error.* [M006]

Similarly, participant K001 emphasized that all recipients would need information on the algorithm's accuracy so that "they knew how reliable it was."

Nevertheless, if these concerns were sufficiently addressed, participants identified several potential benefits of sharing algorithm-generated alerts with a suitable person. Relapse alerts "maybe could spark a conversation" (G001), provide "a good bonding point for us" (G010), "would make life easier" (K010), and help the recipient better understand the DRM user's experiences:

> *I think it would be a good idea…cos then she would kind of know what's going on.* [S005]

> *It'll be more helpful for the family members so they can look it up what my illness is like and what to do…what the triggers are on my mental health and stuff like that.* [M007]

Participants hypothesized that receiving relapse alerts may be "be reassuring and comforting" (S006) to the family, carer, or trusted other, and to the person with psychosis themselves:

> *I think that's a really good idea again because carers are the ones that probably feel the most pressure of keeping somebody that's suffering with complex mental health issues safe. I think they're the ones that*

XSL•FO
RenderX

> *have that emotional investment as well, 'cus I have a carer and I don't always share everything with her, but if the app alerted her…then it gives them a chance to maybe speak to somebody…sort of help me access the support that I need in that moment.* [M011]

> *I'd feel as if I'm being watched and looked over which is really good.* [M014]

Finally, participants emphasized the value of involving trusted family members or carers (with consent) to help contextualize data generated through DRM and avoid fluctuations in behavior or mood being misinterpreted as clinically significant.

> *If anything, it will be more appropriate for maybe close family and friends to get a message, if they consented…because…sometimes…things can be flagged that aren't really that serious but are interpreted seriously, and then you've got a situation where you feel like you're being watched or everything you do is going to cause people to flag a concern.* [K004]

Nevertheless, others preferred clinicians to receive the alerts and then to contact family or carer, outside of the DRM system, to gain additional context:

> *They could reach to people that are close to me and talk to them and try to understand if there is something going on.* [G005]

> *...and then of course contacting the person, like mys…me, but try to get the context before and see.* [G005]

> *Yeah, I don't think that needs to go through to like an app or anything, I think it can be…as long as you made contact kinda thing that's the main thing.* [G008]

Receiving Algorithm-Generated Relapse Alerts Oneself

When considering whether they would wish to receive relapse alerts themselves, a few participants expressed concerns that alerts may generate "unnecessary stress" (M012), causing DRM users to "maybe get anxious" (E007), "get panicky" (E007), and "end up worse than they was before" (M012). Participants anticipated that even *correct* alerts (true positives) would sometimes prompt such fears (as well as false positives; refer to "False Positives: Blame, Unnecessary Restriction, and Fear of Relapse" section). Participant C005 warned that alerts "would have to be tailored to the conditions that people have" because "what works for some people might not work for others." They hypothesized that people with certain diagnoses might find alerts especially challenging:

> *If someone has schizophrenia or they've got severe tendency to repetitive psychotic episodes… I wouldn't advise you getting an alert like we all did last weekend [UK emergency alert system test] on your phone because that would probably set some people off.* [C005]

> *If you were constantly notifying someone with anxiety that [their] behaviour seems a bit erratic, it might set them off.* [C005]

To prevent or mitigate distress from true or false positive alerts sent to oneself, participants suggested carefully phrasing messages (M012) and/or filtering them first through clinicians, family, or trusted others (S001):

> *Instead of saying you're relapsing it could say "signs of relapse"…and then tell you some techniques and stuff you could use to like you know, pull yourself out of it.* [M012]

> *If anything goes wrong, let the doctor see it, they're professionals at the end of the day. If they think that you need speaking to, then speak to you, but don't…I wouldn't send it to the patient every time, because…it could just be a minor problem…that's set them off a bit.* [S001]

Even with such adjustments, some individuals felt that direct alerts were unnecessary ("I know if I'm having a rough time" [K002]) or overly intrusive ("If you go through psychosis, the amount of input you get is a lot. And then to feel like it's still being sort of jammed down your throat would just be quite, yeah, intense I think," C002). Similarly, although open to receiving alerts, one participant worried that alerts may prompt him to switch off his phone in some circumstances:

> *As an idea it's amazing…but if I'm thinking "well no sod everyone and everything" then it might make me switch my phone off even more, and had I not been notified I wouldn't have done…I think it depends on each situation and each individual and their tolerance levels.* [M011]

Conversely, many participants were enthusiastic about receiving algorithm-generated alerts, describing them as "a wake-up call" (C006), a "nudge" (E008, G008), a "chance to introspect a bit and… give you the kick to…pay a bit more attention to your own health" (M008). There was a shared view that alerts could compensate for difficulties in recognizing mental health deteriorations:

> *Sometimes you're the last one to be aware of your mental health deteriorating.* [G008]

> *It is maybe picking up on things you have maybe not noticed are tricky as well. It might make you see things in a new light potentially.* [G001]

> *You might not realise that something went wrong…So it would be good to have a like "oh hold on you've gotta stop doing that."* [C007]

Participants noted that alerts may be particularly beneficial for individuals with limited existing support ("For me…getting in there early was recognized by my family and friends, rather than a smartwatch. But for people who maybe are a bit more isolated…that would be helpful for them." [C002]), for those living alone ("It'd be a bit of a relief because when you live on your own you don't notice things." [S007]) and for those facing abrupt relapses ("It's just that's going down and down and then the next minute you know it you're on your knees." [G006]).

Several interviewees spontaneously proposed additional messages that the DRM system could deliver to encourage help-seeking or support mental health.

XSL•FO
RenderX

> *It would have like…"your mental health is a bit low, so these are things you could do to increase it, or people you can contact or steps you would take."* [G005]

> *I think it needs to have sort of "what can the app do to help now, now that we recognize the signs?" Can you do something like self-care: do you want to try and brush your teeth; do you want to try and eat something today? Something that leads on from the app saying that we recognize your mental health is struggling, because it's good that it states that, but it needs to make someone confident in the sense that they can do something.* [C004]

Preferences for alert delivery format differed among individuals, highlighting the importance of tailoring format to personal context and communication habits:

> *I think that varies from person to person…at the start maybe sign up for how [you] want notifications to come through…cos some people check their emails…some people never check their phones or texts…I probably want it…as a notification.* [G008]

Theme 4: Benefits of Algorithm-Assisted Relapse Prediction

Overall, participants highlighted 4 potential benefits of algorithm-assisted relapse prediction. First, they emphasized its value for early intervention, describing it as a "proactive" (E002) method that could identify relapses early and prompt individuals to "look after [themselves] better, or do things which [they] know help in the context of a relapse, or seek help" (K001). They anticipated that, by intervening early rather than having to "wait for things to get worse" (M006), they could avoid the disruption of a full relapse ("A way to stop me from going under" [M001]).

Second, participants noted that DRM could be used "like a triage" (E008), allowing "information to be passed on in a timely manner" so that clinicians could allocate resources according to need. They proposed that DRM evidence could speed up the process of getting an urgent doctor's appointment ("You can't always get in contact with your GP [General Practitioner] …usually you're like seventeenth in the queue" [M011]), or inpatient bed (C007 quote), and be used to allocate other finite resources such as staff time (E004):

> *If they could work alongside the wards, let's say I had my psychotic episode, they might be able to say 'look this is a priority we need to get this person in now'…They'd be able to prioritize who needs it more and who needs it less.* [C007]

> *If they knew everybody was on that [app] and they'd just be getting…wee alerts "you better phone [name] today," or "you better phone Pam down the road today"…so it would help them keep on top of their workload better.* [E004]

Third, in contrast to others' views that algorithmic relapse predictions might not be accurate (theme 1), some thought that ML might "mitigate that human error that you could get if it's done by a person" (G008) or "eliminate human bias" (M008). Participants alluded to errors and biases originating both from service users themselves and from clinicians, noting the unequal power associated with these roles. For example, participant E002 described the DRM relapse prediction algorithm as "fantastic" as it may help redistribute power by facilitating direct communication:

> *There's miscommunication between the staff so sometimes the doctor will say something and it's not true and I'm like "that's not true" but they're like "Well I'm the doctor and I've been told this" and I'm like "well you've been misinformed"…There's an awful lot of people saying one thing and the doctor can say anything, they're all powerful, the power goes to their head. I think this way is much more kinda wholesome even though its advanced technology, because its actually coming from you and there's not a middle man who's, for whatever motive they've got, sticking an oar in it… Eliminating human error and miscommunication is clearly the way forward here.* [E002]

Fourth, ML was seen as an efficient way of analyzing extensive digital monitoring data: without ML, it "would take a long time to process this much of data" (G009), whereas "through machine learning it would be much quicker" (G009). Participants highlighted DRM's value in "cutting down on worked hours for human people to actually have to go and be trained and then assess people and then tabulate all that" (M008). Particularly in the context of NHS staff shortages, ML was seen as valuable as it would "save the staff from having to do a job like [analyzing DRM data] and frees them up for other jobs" (M010), with one participant reflecting that health services have a "responsibility to use the most up-to-date methods to assess and treat people. So I think health services should use it as soon as it was verified and everything" (M008).

Discussion

Overview of Findings

This study explored the views of people with psychosis on using an algorithm-facilitated system to predict psychosis relapse and their perspectives on sharing relapse alerts with their clinical team, family members, or significant others. With its large, diverse sample, detailed topic guide, and strong lived experience input, it provides a comprehensive and nuanced understanding of stakeholder perspectives on the ethical, practical, and relational implications of using predictive technologies in mental health care. Accordingly, Tables 1-3 outline specific, detailed recommendations for implementation, based on the study's findings.

XSL•FO
RenderX

Table . Accuracy: recommendations based on the study findings.

Finding	Recommendation
DRM[a] users are likely to tolerate some level of errors in an algorithm-based system during a training phase but need to know how accurate to expect relapse predictions to be.	Inform all DRM users (people with psychosis, clinicians, family, or significant others) up-front of known limitations with the algorithm and what to do if errors occur.
Participants thought it important that algorithms get more accurate over time.	Build in a feedback mechanism so the algorithm learns from errors. For example, allow the clinical team, the individual with psychosis, or a dedicated digital review team to calibrate the system's predictions.
Some participants considered false negatives worse than false positives because false positives could be discussed with the clinical team.	When calibrating the algorithm, developers may wish to favor increased sensitivity over specificity. However, if this is done, the balance of sensitivity and specificity should be clearly communicated to all DRM users.
DRM users may be less confident in algorithm-based relapse predictions that are based on passively sensed data than those based on ASM[b] data. Users may perceive passive data as ambiguous without context.	Ensure that all DRM users have a clear, lay explanation of how the relapse prediction algorithm combines ASM and passive data to make predictions. Provide opportunities to seek additional information to allay concerns (eg, could a heart rate increase during exercise be misunderstood as relapse?).
DRM users value openness and honesty about errors in an algorithm-based system.	Be transparent about errors that occur and check their impact on DRM users. Provide reassurance and additional support to mitigate any additional anxiety caused.
Only accurate DRM data should be retained (eg, in health records).	If DRM data is automatically transferred to electronic health records, have a clear protocol for how to check data and correct inaccuracies; be clear who is responsible for ensuring accuracy of records.

[a]DRM: digital remote monitoring.

[b]ASM: active symptom monitoring.

XSL•FO
RenderX

Table . Human-in-the-loop: recommendations based on the study findings.

Finding	Recommendation
False positives are less likely to cause adverse consequences (unnecessary restriction, fear of relapse) if relapse alerts prompt clinicians to seek additional information directly from the person with psychosis.	Algorithm-generated relapse alerts should prompt a thoughtful interaction with a trusted human. ASM[a] data may provide useful additional information to help understand and contextualize alerts from the algorithm.
If DRM[b] is relied on exclusively, false negatives may result in wasted opportunities for intervention, a false sense of security, and delayed help seeking.	Use DRM as a component of treatment, alongside usual care, not as a replacement.
Where they disagree, it may not always be clear whether the DRM relapse alert, or the information provided by the person with psychosis, is correct.	Clinical skill will be required to weigh up information from disparate sources (the DRM system, the individual with psychosis, wider context) and work out whether the DRM system has produced a true positive (the individual is relapsing but unaware or unwilling to disclose) or false positive (individual is not relapsing).
Participants questioned who should oversee the system day to day (eg, clinicians, an oversight committee).	Depending on available resources and local workflows, consider what would be the most suitable way of integrating DRM oversight into routine clinical care.
Staff with overstretched existing workloads will find it challenging to manage extra data provided by DRM alerts.	Clear local policies are needed to outline how DRM fits with existing workflows to avoid staff being overwhelmed.
There is ethical or political complexity around who oversees the system and who is ultimately responsible. Participants raised legitimate concerns about risks of concentrating power in the hands of unregulated commercial entities, particularly when algorithms operate opaquely.	Governance and oversight mechanisms must be established to ensure that the development and deployment of algorithm-generated systems are transparent, accountable, and not necessarily solely controlled by private companies.
Standard translational pipelines for making DRM systems available to the public are currently lacking.	Establish clear, standardized translational pathways alongside appropriate regulatory frameworks and quality standards to guide the safe, effective, and ethical implementation of DRM systems in clinical practice.
Some participants would find it helpful to share relapse alerts with a family member, caregiver, or trusted other. However, fully informed (and potentially ongoing) consent is essential before sharing DRM-generated data with these individuals.	Robust systems should be developed for regular review of data sharing permissions for family and significant others. Service users should be able to proactively change permission settings when needed. Consider safeguarding implications if service users grant sharing permissions and if they change these permissions.
Participants highlighted factors to consider when choosing a suitable person to share alerts with: geographical proximity, whether the relationship is trusting and supportive, and potential impacts of receiving alerts on the relationship and on the alert recipient themself.	Support DRM users to carefully consider who they wish to share relapse alerts with (family, carer, or significant other). Consider the impacts of receiving relapse alerts on the relationship (will it strain the relationship?) and on the person receiving alerts (will they find it upsetting to receive alerts?). Revisit the decision periodically, considering whether it is still useful to share alerts with that individual.
If a family member, carer, or significant other will receive alerts, first they need information about how to use the DRM system, actions if an alert is triggered, and likely accuracy.	When implementing a DRM system that sends relapse alerts to non-clinical contacts, the implementation plan must include training or information for that individual.

[a]ASM: active symptom monitoring.

[b]DRM: digital remote monitoring.

XSL•FO
RenderX

Table . Trust, fears, and choice: recommendations based on the study findings.

Finding	Recommendation
The patient-clinician relationship is a key factor in whether individuals consent to using DRM[a] and sharing DRM alerts. Mutual trust and a collaborative clinical relationship are necessary to manage false positives from a DRM system. Authoritative approaches and prior experiences of coercive or restrictive care increased concerns about DRM.	Recognize the power dynamics between clinicians and people with psychosis when implementing DRM systems. Given clinicians' authority to initiate hospitalization or medication, using algorithm-driven tools in this context requires careful consideration. Safeguards are needed to protect service user autonomy and prevent over-reliance on alerts or other potential misuse.
Using DRM use must be optional.	Ensure that DRM users undergo fully informed consent and have a free choice about whether to use DRM.
Sharing DRM relapse alerts with clinicians must be optional.	Ensure that DRM users undergo fully informed consent and have a free choice about whether to share DRM alerts. Avoid using DRM to replicate existing patterns of coercion.
Allowing users to tailor how and whether they receive relapse alerts themselves may enhance engagement and reduce distress. Some participants preferred not to receive alerts directly, finding them intrusive or anxiety-provoking. Others saw direct alerts as a helpful prompt to monitor their mental health, particularly those living alone or prone to rapid relapse.	In DRM design, build in flexibility to accommodate individual preferences. During onboarding, check whether the individual prefers to receive relapse alerts directly or for alerts to be filtered via a clinician, family member, or trusted other; review this periodically. Consider individual circumstances that may make alerts especially useful. Carefully phrase any relapse alerts that are sent to the user directly to avoid triggering unnecessary anxiety.
Participants acknowledged benefits of algorithm-assisted relapse prediction such as facilitating early intervention, triaging according to need, minimizing human bias, and efficiency in analyzing large volumes of DRM data.	Clinicians should inform service users of potential benefits of DRM and support them in weighing these benefits against possible adverse effects. Discussions about DRM should be embedded within a framework that prioritizes autonomy, choice, and collaborative decision-making.
People with psychosis may be worried about the clinical team overreacting (enforcing hospitalization or medication) or underreacting (not providing additional support) to relapse alerts.	Expectations around what clinicians will do in response to relapse alerts should be clearly articulated before using DRM (during consent and onboarding). Service users may wish to co-create a formal advanced directive stating what they wish to happen if a relapse alert occurs.
Fears of overreaction are not specific to digital tools but reflect broader concerns associated with being a service user in mental health settings.	Recognize that fears of overreaction to relapse alerts are shaped by broader experiences of power and control in mental health services. DRM systems should be used in ways that minimize perceived risk of coercion and actively support collaborative, trust-based care.
Fear of relapse was prominent across themes. Participants suggested that algorithm errors (false positives/negatives) and unsuitable clinician reactions (over/underreaction) may prompt fear of relapse. Fear of relapse may reduce DRM users' willingness to share relapse alerts with others.	Be aware that fear of relapse is an important psychological factor impacting uptake of algorithm-assisted DRM. Provide reassurance and support to participants expressing fear of relapse. As a previous RCT[b] showed [43], using DRM systems monitoring relapse risk can reduce fear of relapse in the medium term if suitably supported.
People with psychosis may be concerned that using DRM could disrupt their carefully built therapeutic relationship. In some cases, individuals may prefer to withhold certain information, and the use of continuous monitoring could be perceived as undermining their sense of control or trust.	Consider potential impacts of DRM on the therapeutic relationship, particularly that continuous monitoring may be perceived as intrusive or undermining trust. Respect individuals' preferences, including their right to withhold information, and prioritize voluntary open communication to preserve a sense of autonomy and mutual respect.

[a]DRM: digital remote monitoring.

[b]RCT: randomized controlled trial.

Participants emphasized the need for transparency about algorithm accuracy, the value of a human-in-the-loop approach, and the importance of mutual trust between the DRM user and recipients of relapse alerts. In particular, the inherently ambiguous nature of psychosis symptoms drove participants' desire for human oversight of DRM. Because understanding symptom changes is contextual and coconstructed, participants emphasized the importance of using DRM with a trusted clinician who understands their communication patterns and can work with them in interpreting subtle changes. Without such trust and understanding, participants feared that clinicians would over- or underreact to DRM relapse alerts. Although participants saw choice about using algorithm-facilitated DRM (and sharing alerts) as essential, they recognized specific benefits of using such a system. To date, this is the most in-depth study to examine the views of people with psychosis on this topic. Nevertheless, findings from existing studies exploring the views of clinicians and people with psychosis on related topics touch on some similar themes.

Comparison With Previous Work

Accuracy of the DRM algorithm emerged as a prominent theme. Although willing to tolerate occasional inaccuracies during early deployment, participants emphasized that limitations should be communicated transparently to users, both initially and when new errors arise. Clinicians also identify accuracy as a key concern [27,29], warning against the use of untested, algorithm-driven digital tools [29] and emphasizing the importance of critically assessing the evidence before implementation [27]. Similarly, mirroring our participants' concerns, clinicians [26-28] cautioned that inaccurate relapse alerts may cause stress, exacerbate users' symptoms (eg, anxiety,

XSL•FO
RenderX

fear of relapse, and paranoia), waste clinicians' time, and ultimately cause users to disengage from DRM [26]. Given that the empirical evidence on predictive validity is still nascent and that algorithmic errors may pose a risk of iatrogenic harm, particularly when implemented at scale [44], these are crucial concerns. Indeed, systematic reviews evaluating ML-generated algorithms' predictive accuracy for relevant outcomes (psychotic relapse [45], schizophrenia treatment outcomes [46], and psychosis prognosis [47]) indicate only moderate accuracy, to date, in studies using wearables [45] or clinical features [46] as predictors; wide CIs [47] indicate high variability across studies. Examples from physical health care (eg, IBM's Watson, which recommended unsafe cancer treatments) [44] highlight the need for adequate testing and rigorous clinical validation of predictive algorithms before implementation. For psychosis relapse prediction, studies in large representative clinical samples [48], using robust methodology (eg, external validation in independent samples [49,50]), and clearly reporting methods and analyses [51,52] are urgently needed to ensure the safety, reliability, and clinical utility of algorithm-based tools before they are integrated into routine mental health care.

The need for a *human-in-the-loop* approach to algorithm-facilitated DRM emerged as a direct corollary of concerns around accuracy. Participants hypothesized that the potential adverse consequences of false positives (unnecessary restriction and fear of relapse) and false negatives (wasted opportunity and false sense of security) may be mitigated by using DRM alerts to prompt human conversation rather than to replace existing care. Similarly, Starke and colleagues [48] suggested a model of "shared responsibility," wherein human agents verify algorithmic outputs to ensure that errors do not violate the principle of nonmaleficence ("do no harm"). A key reason for incorporating human input is to provide contextual understanding of digitally gathered data, particularly passive sensing data. Without context, our participants hypothesized that noncritical changes could be unnecessarily escalated, leading to feelings of surveillance or scrutiny. Clinicians report analogous concerns that decontextualized data may cause misunderstandings [26,28] and a clear consensus that data from AI-facilitated tools should be viewed as supplementary information, to be reviewed by a human [26-29]. Likewise, studies from the broader digital health literature repeatedly emphasize that digital tools should not replace human care [16]. Nevertheless, like our participants, clinicians [26-29] worried that DRM could add to already high workloads. Clinicians expressed feeling overwhelmed by existing demands and questioned how DRM could be integrated into clinical workflows without additional resources [28,29]. They emphasized the time needed to learn to use the DRM system, facilitate patient access, understand DRM data, and respond to alerts [26-28]. Implementation science frameworks (eg, Consolidated Framework for Implementation Research) [53] echo this, highlighting that compatibility with existing workflows is a prerequisite for successful implementation [16].

Mutual trust between the DRM user and the human-in-the-loop was considered crucial by our participants, previous participants with psychosis [30], and clinicians in a comparable study [26]. Without such trust, participants with psychosis feared an inappropriate response to algorithm-generated relapse alerts—either an overreaction (hospitalization or excessive medication) or an underreaction (no additional support). Such responses may be more likely in overstretched services, where the focus is on managing immediate risks and less time is available to build trusting therapeutic relationships and respond early to signs of deterioration. Staff in previous studies [26,27] admitted they might sometimes distrust DRM information, anticipating that people with psychosis may overreport symptoms (to access support) or underreport (to avoid intervention) and that algorithms may not distinguish genuine from dishonest entries [26]. This contrast between patient and staff views illuminates the importance of mutual trust, particularly when ambiguities arise because the DRM system and the person with psychosis disagree. We observed an interplay between participants' anticipated trust in the DRM system (related to its accuracy) and trust in the clinical team—if the DRM was highly accurate, trust in the team was less important, whereas a less accurate DRM would require more trust in the team.

Using predictive algorithms in *secondary care mental health services* presents unique challenges. Our findings highlight inherent power imbalances in this setting. Some participants described experiences of paternalistic and coercive care, understandably impacting their willingness to share relapse alerts with clinicians. Echoing previous study participants [30], they feared that clinicians would suddenly invoke their legal or clinical authority to escalate care. One participant cautioned that DRM algorithms may reproduce or exacerbate coercive practices, drawing a parallel with compulsory antipsychotic medication use. The medical ethics literature highlights similar concerns; McDougall [54] warns that "use of AI in treatment decisions can lead to a new form of paternalism, 'computer knows best'" [55], which could become problematic unless carefully managed. Notably, clinicians [26-29] rarely mentioned power imbalances in mental health services, except in relation to whether DRM should be mandatory. Mirroring our findings, clinician studies unanimously emphasize consent and individual choice regarding DRM use, agreeing that mandatory DRM would be restrictive, coercive, and likely to undermine trust and unfairly exclude people from using services [26]. Clinicians [26] and ethicists [55] advise that all users would need a clearly worded lay explanation of the DRM system. Our participants suggested including information about the algorithm's accuracy within this explanation, and details of procedures for managing inaccurate predictions. Nevertheless, the opaque nature of many algorithmic predictions makes it challenging to provide a full account of how the system works [44]. Moreover, not all individuals may have the necessary understanding of the technology to provide fully informed consent [26,48,56].

Questions of *responsibility and accountability* were discussed in relation to the DRM system oversight. Similar to our finding, clinicians in a previous study queried who is at fault if the AI gets it wrong—the clinician themselves, their institution, or the AI developer [29]. Issues of personal liability are a pressing concern for clinicians in relation to responding to DRM relapse alerts [26,28] and the broader use of AI systems [29]. To mitigate such concerns, staff urge clear organizational guidelines

XSL•FO
RenderX

and policies regarding responsibility for AI-facilitated systems [26,27]. Despite a clear consensus that human oversight of DRM systems is essential, norms around who should provide such oversight are not yet established. Our participants noted ethical complexities, including the legitimate concern that overseeing algorithm-based systems related to others' health could place individuals in positions of power that might be open to abuse. This raises questions about whether commercial AI developers should hold such power, especially given their financial stake in the systems they design. It mirrors previous patient and clinician participants' priorities regarding transparency of data processing within digital health systems [57]. Concerns about responsibility also extended to the integration of DRM data into electronic health records. Although clinicians note that automatic integration could facilitate DRM implementation [26], unlike our participants, they did not highlight the importance of ensuring that such records are accurate, in line with data protection principles [58]. If DRM data were automatically transferred to electronic health records without being flagged as potentially inaccurate, future clinicians might interpret it as factual, potentially privileging algorithmic outputs over the patient's own account. This is a salient concern for individuals with psychosis, who may already feel disbelieved by clinicians. Together, these concerns point to the urgent need for regulation and standards governing algorithm-enabled digital health tools to ensure that these are responsibly overseen and that adverse effects are appropriately managed [59-61].

People with psychosis in this study, along with clinicians in previous studies [27,28], acknowledged *benefits of algorithm-assisted health systems*—efficiency for analyzing large amounts of data, identifying patterns in symptoms or behavior, allowing timely risk stratification, spotting early signs of relapse, and facilitating help-seeking [27,28]. This method may be particularly valuable for real-time analysis of passively sensed data. Our previous paper [25] found that people with psychosis often felt uncomfortable about sharing large volumes of passively sensed data with clinicians. Similarly, other literature reports clinicians' concerns about receiving too much insufficiently summarized passive data [62]. Generating relapse alerts from passively sensed data may be less intrusive and more useful [25,62]. Overall, these perceived advantages suggest that, when implemented thoughtfully and ethically, DRM systems could enhance care delivery and support proactive intervention.

Starke and colleagues [48] argue that "categorical rejection of the use of ML in psychiatry would be ethically wrong given its potential benefits." This perspective reinforces the importance of balancing caution with openness to innovation, ensuring that the development and deployment of DRM systems are guided by robust evidence, ethical principles, and meaningful input from a variety of stakeholders, including people with lived experience of psychosis and clinicians.

Strengths and Limitations

This study's methodological strengths included its large, diverse sample, detailed topic guide, rigorous analysis, and extensive lived experience involvement. Nevertheless, specific limitations may restrict the generalizability of our findings. First, only English-speaking participants were interviewed. Second, despite purposive sampling for diverse characteristics and experiences, people with a specific interest in DRM may have been more likely to participate. Third, although most participants seemed to understand the key principles of ML (after hearing the lay descriptions provided in the topic guide), a minority did not appear to have fully understood these concepts (despite additional explanation and examples), restricting their ability to provide an informed opinion. This may restrict the generalizability of our findings. However, it also highlights that, if ML methods are to be used in clinical practice, it will be important to tailor lay explanations of these methods to individuals' needs, including the subset of people with psychosis who experience difficulties in abstract thinking [63]. Finally, the timing of the interviews (2022 - 2023) should be considered when interpreting the findings, as familiarity with AI has increased over time [64]. Participants may have been less familiar with AI methods than members of the public today, potentially impacting their understanding of and trust in algorithm-assisted systems.

Conclusion

While participants were generally open to using algorithm-assisted DRM, they stressed the importance of transparency about its limitations and the need for human oversight to mitigate risks of false alerts. Our findings highlight the importance of integrating DRM within a trusted, collaborative care model, ensuring that clinicians are prepared to interpret alerts thoughtfully, maintain open communication, and uphold accountability in decision-making.

Acknowledgments

The authors wish to acknowledge the people who participated in this study and the National Health Service (NHS) staff who supported recruitment, the McPin Foundation CONNECT Study Lived Experience Advisory Panel, the Clinical Research Network, and NHS Scotland Mental Health Research Network.

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing. The GenAI tool used was M365 Copilot (version: bizchat.20260203.47.2). Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

XSL•FO
RenderX

Funding

This work was funded by a Wellcome grant (grant 222875/Z/21/Z). SB is supported by a National Institute for Health and Care Research (NIHR) Research Professorship Grant (grant NIHR300794). EE, SB, GH, SL, and JA acknowledge support from the NIHR Manchester Biomedical Research Centre (grant NIHR203308). TW and MC acknowledge support from the NIHR Maudsley Biomedical Research Centre (grant NIHR203318). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Data Availability

Deidentified aggregated participant data will be made available on reasonable request, with investigator support and a signed data access agreement. A proposal should be submitted to the corresponding author (SB) to be reviewed by the study steering committee. Individual data are not publicly available owing to data protection regulations and ethical consent procedures.

Authors' Contributions

SB: conceptualization, methodology, formal analysis, resources, writing - original draft, writing - review & editing, supervision, project administration, funding acquisition.
EE: conceptualization, methodology, formal analysis, writing - original draft, writing - review & editing.
HB: investigation (data collection), writing - review & editing, supervision.
DE, TIK, and AP: formal analysis, writing - review & editing.
AK, AW: formal analysis, writing - review & editing, supervision.
JL: data curation, project administration, writing - review & editing.
SF, RT, and UZ: investigation (data collection), writing - review & editing.
JA, MC, RJD, KG, AG, GH, KMK, SL, MS, JW, and TW: conceptualization, methodology, writing - review & editing, supervision, funding acquisition.

Conflicts of Interest

SB, SL, and JA are co-inventors and co-founders of CareLoop Health Ltd, a University of Manchester start-up to develop and market digital solutions for mental health problems, currently in schizophrenia and postnatal depression. JW has received grant funding from Takeda and Akrivia Health for research unrelated to this work.

Multimedia Appendix 1
Supplementary information, including topic guide, supplementary methods, sample characteristics, and coding tree outlining themes, subthemes, and codes, with supporting quotations.
[DOCX File, 5669 KB - jmir_v28i1e86753_app1.docx]

Checklist 1
COREQ checklist.
[DOCX File, 36 KB - jmir_v28i1e86753_app2.docx]

References

1. Catalan A, Richter A, Salazar de Pablo G, et al. Proportion and predictors of remission and recovery in first-episode psychosis: systematic review and meta-analysis. Eur Psychiatr 2021;64(1):e69. [doi: 10.1192/j.eurpsy.2021.2246]
2. Alvarez-Jimenez M, Priede A, Hetrick SE, et al. Risk factors for relapse following treatment for first episode psychosis: a systematic review and meta-analysis of longitudinal studies. Schizophr Res 2012 Aug;139(1-3):116-128. [doi: 10.1016/j.schres.2012.05.007] [Medline: 22658527]
3. Chan SKW, Pang HH, Yan KK, et al. Ten-year employment patterns of patients with first-episode schizophrenia-spectrum disorders: comparison of early intervention and standard care services. Br J Psychiatry 2020 Sep;217(3):491-497. [doi: 10.1192/bjp.2019.161] [Medline: 31284883]
4. Lin D, Joshi K, Keenan A, et al. Associations between relapses and psychosocial outcomes in patients with schizophrenia in real-world settings in the United States. Front Psychiatry 2021;12:695672. [doi: 10.3389/fpsyt.2021.695672] [Medline: 34764891]
5. Rubio JM, Schoretsanitis G, John M, et al. Psychosis relapse during treatment with long-acting injectable antipsychotics in individuals with schizophrenia-spectrum disorders: an individual participant data meta-analysis. Lancet Psychiatry 2020 Sep;7(9):749-761. [doi: 10.1016/S2215-0366(20)30264-9] [Medline: 32828165]
6. Cai H, Zeng C, Zhang X, et al. Diminished treatment response in relapsed versus first-episode schizophrenia as revealed by a panel of blood-based biomarkers: a combined cross-sectional and longitudinal study. Psychiatry Res 2022 Oct;316:114762. [doi: 10.1016/j.psychres.2022.114762] [Medline: 35940088]

XSL•FO
RenderX

7. Takeuchi H, Siu C, Remington G, et al. Does relapse contribute to treatment resistance? Antipsychotic response in first- vs. second-episode schizophrenia. Neuropsychopharmacol 2019 May;44(6):1036-1042. [doi: 10.1038/s41386-018-0278-3]
8. Chan SKW, Chan SWY, Pang HH, et al. Association of an early intervention service for psychosis with suicide rate among patients with first-episode schizophrenia-spectrum disorders. JAMA Psychiatry 2018 May 1;75(5):458-464. [doi: 10.1001/jamapsychiatry.2018.0185] [Medline: 29617517]
9. Correll CU, Bookhart BK, Benson C, Liu Z, Zhao Z, Tang W. Association of relapse with all-cause mortality in adult patients with stable schizophrenia. Int J Neuropsychopharmacol 2025 May 9;28(5):40168136. [doi: 10.1093/ijnp/pyaf018] [Medline: 40168136]
10. Chan SKW, Chan HYV, Liao Y, et al. Longitudinal relapse pattern of patients with first-episode schizophrenia-spectrum disorders and its predictors and outcomes: a 10-year follow-up study. Asian J Psychiatr 2022 May;71:103087. [doi: 10.1016/j.ajp.2022.103087] [Medline: 35299139]
11. Ascher-Svanum H, Zhu B, Faries DE, et al. The cost of relapse and the predictors of relapse in the treatment of schizophrenia. BMC Psychiatry 2010 Jan 7;10(1):2. [doi: 10.1186/1471-244X-10-2] [Medline: 20059765]
12. Kotzeva A, Mittal D, Desai S, Judge D, Samanta K. Socioeconomic burden of schizophrenia: a targeted literature review of types of costs and associated drivers across 10 countries. J Med Econ 2023;26(1):70-83. [doi: 10.1080/13696998.2022.2157596] [Medline: 36503357]
13. McIntyre RS, Higa S, Doan QV, et al. Place of care and costs associated with acute episodes and remission in schizophrenia. J Manag Care Spec Pharm 2023 May;29(5):499-508. [doi: 10.18553/jmcp.2023.29.5.499] [Medline: 37121252]
14. Pennington M, McCrone P. The cost of relapse in schizophrenia. Pharmacoeconomics 2017 Sep;35(9):921-936. [doi: 10.1007/s40273-017-0515-3] [Medline: 28534254]
15. Bell IH, Eisner E, Allan S, et al. Methodological characteristics and feasibility of ecological momentary assessment studies in psychosis: a systematic review and meta-analysis. Schizophr Bull 2024 Mar 7;50(2):238-265. [doi: 10.1093/schbul/sbad127] [Medline: 37606276]
16. Eisner E, Faulkner S, Allan S, et al. Barriers and facilitators of user engagement with digital mental health interventions for people with psychosis or bipolar disorder: systematic review and best-fit framework synthesis. JMIR Ment Health 2025 Jan 20;12:e65246. [doi: 10.2196/65246] [Medline: 39832352]
17. Dennard S, Patel R, Garety P, Edwards C, Gumley A. A systematic review of users experiences of using digital interventions within psychosis: a thematic synthesis of qualitative research. Soc Psychiatry Psychiatr Epidemiol 2025 Feb;60(2):275-303. [doi: 10.1007/s00127-024-02692-4] [Medline: 38802509]
18. Gleeson JF, McGuckian TB, Fernandez DK, et al. Systematic review of early warning signs of relapse and behavioural antecedents of symptom worsening in people living with schizophrenia spectrum disorders. Clin Psychol Rev 2024 Feb;107:102357. [doi: 10.1016/j.cpr.2023.102357] [Medline: 38065010]
19. Torous J, Linardon J, Goldberg SB, et al. The evolving field of digital mental health: current evidence and implementation issues for smartphone apps, generative artificial intelligence, and virtual reality. World Psychiatry 2025 Jun;24(2):156-174. [doi: 10.1002/wps.21299] [Medline: 40371757]
20. Hau C, Xia W, Ryan S, Firth J, Linardon J, Torous J. Smartphone monitoring and digital phenotyping apps for schizophrenia: a review of the academic literature. Schizophr Res 2025 Jul;281:237-248. [doi: 10.1016/j.schres.2025.05.019] [Medline: 40413837]
21. Mohr DC, Lyon AR, Lattie EG, Reddy M, Schueller SM. Accelerating digital mental health research from early design and creation to successful implementation and sustainment. J Med Internet Res 2017 May 10;19(5):e153. [doi: 10.2196/jmir.7725] [Medline: 28490417]
22. Nilsen P, Svedberg P, Nygren J, Frideros M, Johansson J, Schueller S. Accelerating the impact of artificial intelligence in mental healthcare through implementation science. Implement Res Pract 2022;3:26334895221112033. [doi: 10.1177/26334895221112033] [Medline: 37091110]
23. Ball H, Eisner E, Nicholas J, Wilson P, Bucci S. How theories, models, and frameworks have been used to implement digital health interventions in services for people with severe mental health problems: a scoping review. BMC Public Health 2025 Mar 17;25(1):1023. [doi: 10.1186/s12889-025-22189-2] [Medline: 40098003]
24. Chaudoir SR, Dugan AG, Barr CH. Measuring factors affecting implementation of health innovations: a systematic review of structural, organizational, provider, patient, and innovation level measures. Implementation Sci 2013 Dec;8(1):22. [doi: 10.1186/1748-5908-8-22]
25. Eisner E, Ball H, Ainsworth J, et al. Using passive sensing to predict psychosis relapse: an in-depth qualitative study exploring perspectives of people with psychosis. Schizophr Bull 2025 Aug 21;PMID:sbaf126. [doi: 10.1093/schbul/sbaf126] [Medline: 40840436]
26. Ball H, Eisner E, Ainsworth J, et al. Mental health professionals' perspectives on digital remote monitoring in services for people with psychosis. Schizophr Bull 2026 Jan 16;52(1):sbaf043. [doi: 10.1093/schbul/sbaf043] [Medline: 40329411]
27. Rogan J, Bucci S, Firth J. Health care professionals' views on the use of passive sensing, AI, and machine learning in mental health care: systematic review with meta-synthesis. JMIR Ment Health 2024 Jan 23;11:e49577. [doi: 10.2196/49577] [Medline: 38261403]

XSL•FO
RenderX

28. Rogan J, Firth J, Bucci S. Healthcare professionals' views on the use of passive sensing and machine learning approaches in secondary mental healthcare: a qualitative study. Health Expect 2024 Dec;27(6):e70116. [doi: 10.1111/hex.70116] [Medline: 39587845]
29. Stroud AM, Curtis SH, Weir IB, et al. Physician perspectives on the potential benefits and risks of applying artificial intelligence in psychiatric medicine: qualitative study. JMIR Ment Health 2025 Feb 10;12:e64414. [doi: 10.2196/64414] [Medline: 39928397]
30. Yoo DW, Woo H, Nguyen VC, et al. Patient perspectives on AI-driven predictions of schizophrenia relapses: understanding concerns and opportunities for self-care and treatment. Presented at: CHI '24; May 11-16, 2024 URL: https://dl.acm.org/doi/proceedings/10.1145/3613904 [accessed 2026-04-02] [doi: 10.1145/3613904.3642369]
31. CONNECT study webpage. URL: www.connectdigitalstudy.com [accessed 2026-02-03]
32. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
33. ICD-10 Classifications of Mental and Behavioural Disorder: Clinical Descriptions and Diagnostic Guidelines: World Health Organisation; 1992.
34. Kay SR, Fiszbein A, Opler LA. The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. Schizophr Bull 1987;13(2):261-276. [doi: 10.1093/schbul/13.2.261] [Medline: 3616518]
35. Yung AR, Yuen HP, McGorry PD, et al. Mapping the onset of psychosis: the Comprehensive Assessment of At-Risk Mental States. Aust N Z J Psychiatry 2005;39(11-12):964-971. [doi: 10.1080/j.1440-1614.2005.01714.x] [Medline: 16343296]
36. Braun V, Clarke V. To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. Qualitative Research in Sport, Exercise and Health 2021 Mar 4;13(2):201-216. [doi: 10.1080/2159676X.2019.1704846]
37. Braun V, Clarke V. Thematic Analysis: A Practical Guide: SAGE; 2022.
38. Harrison Ginsberg K, Babbott K, Serlachius A. Exploring participants' experiences of digital health interventions with qualitative methods: guidance for researchers. J Med Internet Res 2024 Nov 28;26:e62761. [doi: 10.2196/62761] [Medline: 39607999]
39. NVivo. Lumivero. 2017. URL: https://lumivero.com/products/nvivo/ [accessed 2026-02-03]
40. Scott D. Resolving the quantitative–qualitative dilemma: a critical realist approach. International Journal of Research & Method in Education 2007 Apr;30(1):3-17. [doi: 10.1080/17437270701207694]
41. Fletcher AJ. Applying critical realism in qualitative research: methodology meets method. Int J Soc Res Methodol 2017 Mar 4;20(2):181-194. [doi: 10.1080/13645579.2016.1144401]
42. The CONNECT study. McPin Foundation. URL: https://mcpin.org/project/the-connect-study/ [accessed 2026-04-02]
43. Gumley AI, Bradstreet S, Ainsworth J, et al. The EMPOWER blended digital intervention for relapse prevention in schizophrenia: a feasibility cluster randomised controlled trial in Scotland and Australia. Lancet Psychiatry 2022 Jun;9(6):477-486. [doi: 10.1016/S2215-0366(22)00103-1] [Medline: 35569503]
44. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
45. Smyrnis A, Theleritis C, Ferentinos P, Smyrnis N. Psychotic relapse prediction via biomarker monitoring: a systematic review. Front Psychiatry 2024;15:1463974. [doi: 10.3389/fpsyt.2024.1463974] [Medline: 39691789]
46. Saboori Amleshi R, Ilaghi M, Rezaei M, et al. Predictive utility of artificial intelligence on schizophrenia treatment outcomes: a systematic review and meta-analysis. Neuroscience & Biobehavioral Reviews 2025 Feb;169:105968. [doi: 10.1016/j.neubiorev.2024.105968]
47. Tay JL, Ang YL, Tam WWS, Sim K. Accuracy of machine learning methods in predicting prognosis of patients with psychotic spectrum disorders: a systematic review. BMJ Open 2025 Feb 25;15(2):e084463. [doi: 10.1136/bmjopen-2024-084463] [Medline: 40000074]
48. Starke G, De Clercq E, Borgwardt S, Elger BS. Computing schizophrenia: ethical challenges for machine learning in psychiatry. Psychol Med 2021 Nov;51(15):2515-2521. [doi: 10.1017/S0033291720001683] [Medline: 32536358]
49. Stahl D, Pickles A. Fact or fiction: reducing the proportion and impact of false positives. Psychol Med 2018 May;48(7):1084-1091. [doi: 10.1017/S003329171700294X] [Medline: 29173233]
50. Grzenda A, Kraguljac NV, McDonald WM, et al. Evaluating the machine learning literature: a primer and user's guide for psychiatrists. AJP 2021 Aug 1;178(8):715-729. [doi: 10.1176/appi.ajp.2020.20030250]
51. Bladon S, Eisner E, Bucci S, et al. A systematic review of passive data for remote monitoring in psychosis and schizophrenia. NPJ Digit Med 2025 Jan 27;8(1):62. [doi: 10.1038/s41746-025-01451-2] [Medline: 39870797]
52. Linardon J, Chen K, Gajjar S, et al. Smartphone digital phenotyping in mental health disorders: a review of raw sensors utilized, machine learning processing pipelines, and derived behavioral features. Psychiatry Res 2025 Jun;348:116483. [doi: 10.1016/j.psychres.2025.116483] [Medline: 40187059]
53. Damschroder LJ, Reardon CM, Opra Widerquist MA, Lowery J. Conceptualizing outcomes for use with the Consolidated Framework for Implementation Research (CFIR): the CFIR outcomes addendum. Implementation Sci 2022 Dec;17(1):7. [doi: 10.1186/s13012-021-01181-5]

XSL•FO
RenderX

54. McDougall RJ. Computer knows best? The need for value-flexibility in medical AI. J Med Ethics 2019 Mar;45(3):156-160. [doi: 10.1136/medethics-2018-105118] [Medline: 30467198]
55. Hansson SO, Fröding B. Digital technology in healthcare—an ethical overview. DISO 2024 Dec;3(3). [doi: 10.1007/s44206-024-00121-2]
56. Yassin W, Loedige KM, Wannan CMJ, et al. Biomarker discovery using machine learning in the psychosis spectrum. Biomark Neuropsychiatry 2024 Dec;11:100107. [doi: 10.1016/j.bionps.2024.100107] [Medline: 39687745]
57. Torous J, Wisniewski H, Bird B, et al. Creating a digital health smartphone app and digital phenotyping platform for mental health and diverse healthcare needs: an interdisciplinary and collaborative approach. J technol behav sci 2019 Jun;4(2):73-85. [doi: 10.1007/s41347-019-00095-w]
58. A guide to the data protection principles. Information Commissioner's Office. URL: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-protection-principles/a-guide-to-the-data-protection-principles/ [accessed 2025-07-17]
59. Eisner E, Richardson C, Thomas N, et al. Measurement of adverse events in studies of digital health interventions for psychosis: guidance and recommendations based on a literature search and framework analysis of standard operating procedures. Schizophr Bull 2024 Nov 8;50(6):1456-1470. [doi: 10.1093/schbul/sbae048] [Medline: 38683836]
60. Allan S, Ward T, Eisner E, et al. Adverse events reporting in digital interventions evaluations for psychosis: a systematic literature search and individual level content analysis of adverse event reports. Schizophr Bull 2024 Nov 8;50(6):1436-1455. [doi: 10.1093/schbul/sbae031] [Medline: 38581410]
61. Kahane K, Shumate JN, Torous J. Policy in Flux: addressing the regulatory challenges of AI integration in US mental health services. Curr Treat Options Psych 2025 Jun 16;12(1):24. [doi: 10.1007/s40501-025-00362-z]
62. Rodriguez-Villa E, Rozatkar AR, Kumar M, et al. Cross cultural and global uses of a digital mental health app: results of focus groups with clinicians, patients and family members in India and the United States. Glob Ment Health (Camb) 2021;8:e30. [doi: 10.1017/gmh.2021.28] [Medline: 34512999]
63. McCutcheon RA, Keefe RSE, McGuire PK. Cognitive impairment in schizophrenia: aetiology, pathophysiology, and treatment. Mol Psychiatry 2023 May;28(5):1902-1918. [doi: 10.1038/s41380-023-01949-9] [Medline: 36690793]
64. Stanford Institute for Human-Centered Artificial Intelligence, Capstick E, Etchemendy J, Fattorini L, et al. Artificial Intelligence Index Report 2025 | Chapter 8: Public Opinion. : Stanford University; 2025 URL: https://hai.stanford.edu/assets/files/hai_ai-index-report-2025_chapter8_final.pdf [accessed 2026-02-03]

Abbreviations

AI: artificial intelligence
ASM: active symptom monitoring
COREQ: Consolidated Criteria for Reporting Qualitative Research
DRM: digital remote monitoring
***ICD-10*:** *International Classification of Diseases, Tenth Revision*
LEAP: Lived Experience Advisory Panel
ML: machine learning
NHS: National Health Service

Edited by A Stone; submitted 03.Nov.2025; peer-reviewed by A Reinhold, W Liu; accepted 24.Feb.2026; published 10.Apr.2026.

Please cite as:
Eisner E, Ball H, Ainsworth J, Cella M, Drake RJ, Elton D, Faulkner S, Greenwood K, Gumley A, Haddock G, Kendall KM, Kenny A, Krogsæter TI, Lees J, Lewis S, Phiri A, Schwannauer M, Turner R, Walsh AEL, Walters J, Wykes T, Zahid U, Bucci S
Views of People With Psychosis About Algorithm-Based Relapse Prediction and Data Sharing: Qualitative Study
J Med Internet Res 2026;28:e86753
URL: https://www.jmir.org/2026/1/e86753
doi:10.2196/86753

© Emily Eisner, Hannah Ball, John Ainsworth, Matteo Cella, Richard J Drake, Daniel Elton, Sophie Faulkner, Kathryn Greenwood, Andrew Gumley, Gillian Haddock, Kimberley M Kendall, Alex Kenny, Tor-Ivar Krogsæter, Jane Lees, Shôn Lewis, Alie Phiri, Matthias Schwannauer, Rebecca Turner, Annabel E L Walsh, James Walters, Til Wykes, Uzma Zahid, Sandra Bucci. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Context-Aware Sentence Classification of Radiology Reports Using Synthetic Data: Development and Validation Study

Tomohiro Kikuchi[1,2,3], MPH, MD, PhD; Yosuke Yamagishi[4], MD; Kohei Yamamoto[2], MMSc; Toshiaki Akashi[5], MD, PhD; Harushi Mori[2], MD, PhD; Hisaki Makimoto[1], MD, PhD; Takahide Kohro[1], MD, PhD

[1]Data Science Center, Jichi Medical University, Tochigi, Japan

[2]Department of Radiology, Jichi Medical University, 3311-1, Yakushiji, Shimotsuke, Tochigi, Japan

[3]Department of Computational Diagnostic Radiology and Preventive Medicine, The University of Tokyo Hospital, Tokyo, Japan

[4]Division of Radiology and Biomedical Engineering, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

[5]Department of Radiology, Juntendo University School of Medicine, Tokyo, Japan

Corresponding Author:
Tomohiro Kikuchi, MPH, MD, PhD
Data Science Center, Jichi Medical University, Tochigi, Japan

Abstract

Background: Automated structuring of radiology reports is essential for data utilization and the development of medical artificial intelligence models. However, manual annotation by experts is labor-intensive, and processing real clinical data through commercial large language models (LLMs) presents significant privacy risks. These challenges are particularly pronounced for non-English languages like Japanese, where specialized medical corpora are scarce. While synthetic data generation offers a potential privacy-preserving alternative, its effectiveness in capturing complex clinical nuances—such as negation and contextual dependencies—to train robust classification models without any real-world training data has not been fully established.

Objective: This study aimed to develop a context-aware sentence classification model for Japanese radiology reports using an entirely synthetic training pipeline, thereby eliminating reliance on real-world clinical data during the development phase. Furthermore, we sought to evaluate the generalizability of this approach by validating the model's performance on diverse, multi-institutional, real-world reports.

Methods: Japanese radiology reports (n=3104) were generated using GPT-4.1 and automatically annotated at the sentence level into 4 categories (background, positive finding, negative finding, and continuation) using GPT-4.1-mini. The synthetic data were partitioned into training (n=2670), validation (n=334), and test (n=100) sets. We fine-tuned several models, including lightweight local LLMs (Qwen3 and Llama 3.2 series) using low-rank adaptation and Japanese text classification models (Bidirectional Encoder Representations from Transformers [BERT]-base Japanese v3, Japanese Medical Robustly Optimized BERT Pretraining Approach [JMedRoBERTa]-base, and ModernBERT-Ja-130M). External validation was performed using 280 real-world reports (3477 sentences) from 7 institutions in the Japan Medical Image Database, with ground-truth labels established by board-certified radiologists. Evaluation metrics included accuracy, macro-averaged F_1 (macro F_1) score, and positive predictive value for positive findings (PPV_1).

Results: All models achieved high performance on the synthetic test set (accuracy: 0.938 - 0.951; macro F_1-score: 0.924 - 0.940). Overall performance declined on the external validation dataset (accuracy: 0.783 - 0.813; macro F_1-score: 0.761 - 0.790), reflecting distributional differences between synthetic and real-world reports; however, PPV_1 remained stable and high across datasets (eg, 0.957 on the synthetic test set vs 0.952 on the external validation dataset for Qwen3 [4B]). Parsing errors occurred in LLM-based approaches (19 - 260 sentences, 0.55% - 7.48% in the external dataset).

Conclusions: This study demonstrates the feasibility of developing context-aware sentence classification models for Japanese radiology reports using a training pipeline based entirely on synthetic data. The stability of PPV_1 indicates that the models successfully captured the essential clinical terminology and linguistic patterns required to identify positive findings in real-world reports, despite the observed performance degradation during external validation. This approach substantially reduces manual annotation requirements and privacy risks, providing a scalable foundation for constructing structured radiology datasets to support the development of clinically relevant medical artificial intelligence models.

(J Med Internet Res 2026;28:e86365) doi:10.2196/86365

KEYWORDS

natural language processing; artificial intelligence; large language models; data annotation; radiology report

XSL•FO
RenderX

Introduction

In recent years, vision-language models (VLMs) have been increasingly applied to medical image analysis [1,2]. VLMs are multimodal architectures that commonly integrate large language models (LLMs) for textual processing with visual encoders for image representation, enabling tasks such as image captioning, visual question answering, and image-text matching [3]. In the medical domain, VLMs are typically trained on paired image-text data to associate imaging findings with their corresponding textual descriptions, supporting applications including automated report generation, cross-modal retrieval, and clinical decision support [4-6]. Training these models requires large-scale image-text pairs, and both the quality and quantity of such pairs critically affect the performance of the model [7,8].

Several large-scale, English-language medical image-text datasets, such as the Medical Information Mart for Intensive Care Chest X-Ray (MIMIC-CXR) and CT-RATE, have recently been released [9-11]. These resources have become foundational for training VLMs. Consequently, many state-of-the-art VLMs are predominantly trained and evaluated on English data, largely due to the sheer volume of available resources. When applied to other languages, these models often exhibit performance degradation due to linguistic disparities, such as variations in negation, uncertainty expressions, and reporting conventions across institutions [12]. This is especially true for languages such as Japanese; while they are well-supported in general-domain natural language processing (NLP), the scarcity of public radiology-specific corpora creates a low-resource environment in the medical domain [13,14]. Although some non-English resources such as PadChest exist, they remain limited in scale or scope [15,16]. Simple cross-lingual transfer or translation-based approaches may fail to capture language-specific clinical nuances and local disease prevalence. Furthermore, many publicly available datasets are biased toward prototypical or specific disease cases and do not adequately reflect the diversity of findings and contexts encountered in daily clinical practice [17].

To develop VLMs that can be used in real-world radiology workflows across different countries, it is essential to establish and continuously curate language-specific datasets that reflect local clinical practices. However, radiology reports are typically unstructured free text, exhibiting significant variations in structure and style across institutions and radiologists. They often contain diverse sentence types, including background information, positive findings, negative findings, and continuation from previous sentences. Consequently, simple sentence segmentation is inadequate because sentence boundaries often do not align with thematic transitions, and reports frequently include content that is irrelevant to the actual image findings. Therefore, constructing automated image-text pairs requires text processing—specifically, context-aware sentence classification—as a preliminary step before alignment.

While commercial LLMs have shown success in radiology tasks, developing a pipeline that transmits clinical reports to external services is often unacceptable due to ethical, regulatory, and privacy concerns [13,18,19]. Beyond these, economic and technical factors pose substantial challenges; high cumulative application programming interface costs and nontransparent model updates hinder large-scale data curation and compromise reproducibility [20,21]. Consequently, local models are necessary; however, manual annotation for in-house training data is extremely labor-intensive [22,23].

To address these challenges, we leveraged a commercial LLM solely to generate synthetic Japanese radiology reports and perform automated sentence-level annotation, eliminating the need for labor-intensive manual annotation. This synthetic dataset was used to fine-tune lightweight local LLMs, enabling efficient model development while strictly safeguarding patient privacy by avoiding the use of actual clinical data. This study aimed to develop a context-aware sentence classification model for segmenting Japanese radiology reports into finding-level textual units that are suitable for pairing with images for VLM training. We investigated the feasibility of executing the entire training pipeline exclusively using synthetic data, with validation performed on multi-institutional, real-world datasets.

Methods

Ethical Considerations

This retrospective human-subject study was approved by the institutional review board of Jichi Medical University Hospital (approval J24-017; approval date: June 10, 2024). The requirement for informed consent was waived by the institutional review board because this study involved a secondary analysis of existing clinical data. All data were anonymized or deidentified prior to analysis in accordance with institutional data governance policies, and no direct personal identifiers were accessible to the research team at any point during the study. No compensation was provided to participants. All figures and images included in this paper were nonidentifiable, and no additional consent for image publication was required.

Data Synthesis and Annotation (Training Data)

We generated synthetic Japanese radiology reports using the OpenAI application programming interface and automatically annotated each sentence into 4 predefined categories (Figure 1). Initially, we attempted to generate 3200 reports using GPT-4.1. To ensure the generation of diverse reports, we explicitly defined various attributes and their candidate values—including radiologist experience level, reporting style and structural preferences, patient demographics, anatomical site, reason for examination, image quality and coverage scope, presence of prior examinations, disease category, diagnostic certainty, and disease rarity—and randomly combined and selected them for each report (Section A in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 1. Flowchart of synthetic data generation. Prompt 1 was used for report generation (Section A in Multimedia Appendix 1), and Prompt 2 (system prompt) was used for report annotation (Section B in Multimedia Appendix 1). Synthetic radiology reports were generated using the OpenAI API (GPT-4.1), and these were annotated using the OpenAI API (GPT-4.1-mini). Reports were excluded based solely on structural validity (eg, parsing errors) without semantic filtering. The final dataset was partitioned into training (n=2670), validation (n=334), and test (n=100) sets. API: application programming interface.

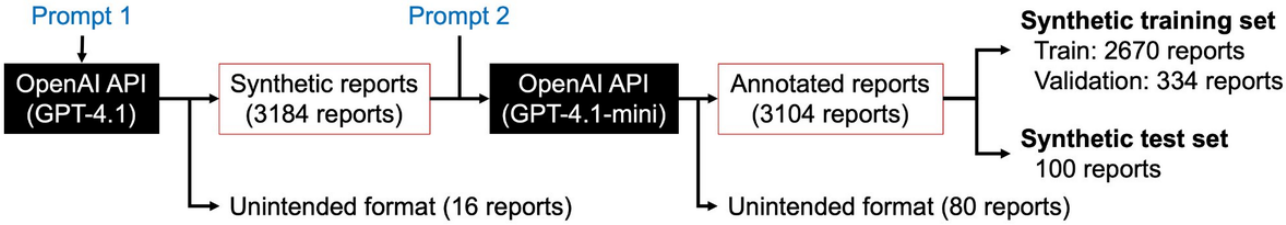

Subsequently, these reports were automatically annotated using GPT-4.1-mini, where sentences were assigned to one of the following 4 categories: label 0 (background) for nonfinding information within the findings section, including clinical history, prior study comparisons, technical assessments, and demographics; label 1 (positive finding) to denote the presence of abnormalities; label 2 (negative finding) to denote the absence of abnormalities, with label 1 taking precedence if both coexist in a single sentence; and label 3 (continuation) for sentences that provide supplementary details to the preceding sentence without constituting a standalone finding. Representative examples for each label are provided in Table 1, and the system prompt used for automated annotation is detailed in Section B in Multimedia Appendix 1.

After excluding reports with unintended formats, including parsing errors (failure to produce a valid JSON structure), the final dataset comprised 2670 training data, 334 validation data, and 100 test reports. For the synthetic test set, sentence-level annotation into the 4 predefined categories was manually performed by 2 board-certified radiologists, and a consensus was reached for each sentence.

Table . Definitions of sentence labels and examples.

Label	Definitions	Examples
0=Background	Background information not related to current imaging findings	• "Patient has a history of right lung cancer surgery." • "Comparison with previous CT[a] from October 2024."
1=Positive finding	Description of a distinct positive finding	• "A 2-cm hypervascular lesion is observed in Segment 8 of the liver." • "Mild pleural effusion is present on the right side."
2=Negative finding	Description of a distinct negative finding (absence of abnormalities)	• "No significant lymphadenopathy is noted." • "There are no signs of intracranial hemorrhage."
3=Continuation	Sentences that modify or describe the preceding finding and do not stand alone.	• "It shows gradual washout on the delayed phase." • "This lesion appears stable compared to the previous exam."

[a]CT: computed tomography.

External Validation Dataset

To evaluate whether models trained solely on synthetic data generalize to real-world clinical practice, we constructed a multi-institutional external validation dataset using reports from the Japan Medical Image Database (J-MID) [24], a nationwide repository of real-world radiological images and reports collected from multiple Japanese institutions. Instead of selecting curated or disease-specific cases, we adopted a snapshot sampling strategy to reflect routine clinical practice. Specifically, we randomly sampled 40 computed tomography reports from each of the 7 participating institutions within the J-MID, resulting in a total of 280 reports dated October 1, 2024. The dataset demonstrated varying imaging coverage—torso (n=141, 50.4%), chest (n=64, 22.9%), abdomen/pelvis (n=20, 7.1%), head (n=19, 6.8%), whole body (n=12, 4.3%), neck (n=8, 2.9%), and others (n=16, 5.7%)—including both noncontrast (n=176, 62.9%) and contrast-enhanced (n=104, 37.1%) studies. For text preprocessing, reports were segmented into sentences using line breaks and periods (“.”). Sentence-level annotation into the 4 predefined categories was manually performed by 2 board-certified radiologists, and a consensus was reached for each sentence.

Model Selection

We used both LLMs and Japanese text classification models to construct and compare sentence classification models. To ensure that the training could be handled on a single NVIDIA RTX6000

XSL•FO
RenderX

Ada graphics processing unit (GPU) (48 GB memory), we limited the model size to approximately 4 billion parameters (4B). For LLMs, we fine-tuned Qwen3 (0.6B, 1.7B, 4B) and Llama 3.2 (1B, 3B) using low-rank adaptation (LoRA) [25,26]. For text classification, we selected 3 Japanese models based on the BERT (Bidirectional Encoder Representations from Transformers), RoBERTa (Robustly Optimized BERT Pretraining Approach), and ModernBERT architectures [27-29]. First, BERT base Japanese v3 was included as a widely used model in Japanese general and medical NLP tasks, including radiology-focused applications [30,31]. Second, JMedRoBERTa [Japanese Medical RoBERTa]-base was utilized because this model is pretrained on large-scale Japanese medical literature [32]. Third, ModernBERT-Ja-130M was selected as a recently proposed architecture that achieves high performance while maintaining strong computational efficiency [33]. Links to the model cards for all models used in this study are provided in Section C in Multimedia Appendix 1.

Model Training and Inference

To fine-tune the LLM, we adopted the LoRA framework [34]. The configuration was as follows: rank (r)=16, LoRA α=32, dropout=0.05, and target modules={q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj}. Training was performed for 3 epochs with a batch size of 2 and gradient accumulation of 8 (effective batch size=16). The learning rate was set to 1×10^{-4} with a cosine scheduler, including 100 warm-up steps. The maximum sequence length was 2048 tokens. We applied 4-bit quantization using quantized LoRA (QLoRA) with 16-bit mixed-precision for memory efficiency [35]. In this setting, the entire text was provided as input, and the model was trained for a text-generation task to produce structured outputs in JSON format, utilizing the system prompt detailed in Section B in Multimedia Appendix 1. The outputs were subsequently parsed into sentence-level predictions.

For the text classification models, fine-tuning was performed using the HuggingFace Trainer with a maximum sequence length of 512 tokens, a batch size of 16, a learning rate of 2×10^{-5}, and 3 training epochs. Mixed precision training with 16-bit mixed precision was applied, and the best checkpoint was selected according to the macro-averaged F_1 (macro F_1) score of the validation set. To incorporate contextual information, we used a sentence-centered local context window consisting of the target sentence and its 2 preceding and 2 following sentences. This approach ensures the inclusion of sufficient local context while remaining within the 512-token limit typical of BERT-based models. During inference, models were applied to individual reports in a sentence-wise manner using the same context window configuration as in training.

Evaluation Metrics

To provide a rigorous assessment of both the automated annotations and the fine-tuned models, we defined the following standard classification metrics. Classification accuracy was calculated as the ratio of correctly predicted labels to the total number of sentences.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of sentences}}$$

To ensure a balanced evaluation under class imbalance, we additionally adopted the macro F_1-score. For each class i(i∈{0,1,2,3}) (0=background, 1=positive finding, 2=negative finding, and 3=continuation), precision (Pi), recall (Ri), and F_1-score (F1i) were defined as follows:

$$P_i = \frac{TP_i}{TP_i + FP_i}, R_i = \frac{TP_i}{TP_i + FN_i}, F1_i = \frac{2 \times P_i \times R_i}{P_i + R_i}$$

The macro F_1-score was calculated as the arithmetic mean of these class-specific scores:

$$\text{MacroF1} = \frac{1}{4}\sum_{i=0}^{3} F1_i$$

In addition, we reported the positive predictive value for label 1 (PPV_1), corresponding to P1 in this framework. This metric was included because label 1 represents positive imaging findings, and high precision for this class is essential for ensuring the reliability of extracted image-text pairs. We also reported the recall for label 1 (Recall_1) and class-specific F_1-score for label 1 (F1_1) to evaluate dataset coverage and the overall precision-recall balance, corresponding to R1 and F1_1, respectively, in the above framework.

First, to validate the quality of the automated sentence-level annotations generated by GPT-4.1-mini, we compared them against the radiologist consensus on the synthetic test set. Agreement was quantified using classification accuracy, macro F_1-score, and Cohen κ coefficient. Second, for the LLM outputs, we quantified the incidence of parsing errors and aggregated these error rates by facility. Third, the performance of all models was evaluated on both the synthetic test set and external validation datasets using accuracy, macro F_1-score, PPV_1, Recall_1, and F1_1. For LLM-based approaches, sentences associated with parsing errors were excluded from subsequent performance analyses. Fourth, we performed a targeted error analysis. We reviewed 40 reports from the institution with the highest parsing error rate to identify recurring patterns. Additionally, we examined the best-performing model's errors by extracting false-positive label 1 predictions to identify the primary reasons for these misclassifications.

Results

Table 2 summarizes the distribution of labels in the datasets used in this study. In the synthetic dataset, label 2 was the most frequent category at 44%, whereas label 1 predominated in the external validation dataset at 42.6%. Synthetic reports exhibited higher verbosity with a mean (SD) character count of 359.1 (SD 138.9) compared to 224.4 (SD 125.9) in the external dataset, despite having similar mean sentence counts (11.2, SD 4.0 vs 12.4, SD 6.3). Substantial variability was observed across external institutions.

Figure 2 presents the confusion matrix for the annotation performance of GPT-4.1-mini compared with the radiologist consensus on the synthetic test set. The classification accuracy, macro F_1-score, and Cohen κ were 0.942, 0.929, and 0.917, respectively.

XSL•FO
RenderX

Table . Distribution of the dataset.

Dataset	Label 0[a], n (%)	Label 1[a], n (%)	Label 2[a], n (%)	Label 3[a], n (%)	Characters/report, mean (SD)	Sentences/report, mean (SD)
Synthetic dataset (3104 reports)						
Total	6593 (18.9)	7673 (22.0)	15,327 (44.0)	5257 (15.1)	359.1 (138.9)	11.2 (4.0)
Train	5680 (18.9)	6620 (22.1)	13,187 (44.0)	4503 (15.0)	359.7 (138.5)	11.2 (4.0)
Validation	699 (18.7)	828 (22.2)	1629 (43.6)	580 (15.5)	360.1 (141.6)	11.2 (3.9)
Test	214 (19.0)	225 (20.0)	511 (45.5)	174 (15.5)	338.6 (139.2)	11.2 (4.3)
External validation dataset (280 reports)						
Total	531 (15.3)	1480 (42.6)	1001 (28.8)	465 (13.4)	224.4 (125.9)	12.4 (6.3)
Institution A	64 (11.6)	233 (42.2)	196 (35.5)	59 (10.7)	304.7 (127.4)	13.8 (6.1)
Institution B	68 (13.9)	200 (40.9)	182 (37.2)	39 (8.0)	202.5 (73.2)	12.2 (4.4)
Institution C	58 (20.4)	122 (42.8)	58 (20.4)	47 (16.5)	148.3 (87.1)	7.1 (3.7)
Institution D	77 (18.3)	150 (35.6)	132 (31.4)	62 (14.7)	135.2 (70.0)	10.5 (5.3)
Institution E	75 (15.5)	201 (41.5)	166 (34.3)	42 (8.7)	186.6 (81.8)	12.1 (4.9)
Institution F	89 (12.4)	337 (47.1)	181 (25.3)	108 (15.1)	339.3 (138.3)	17.9 (7.6)
Institution G	100 (18.8)	237 (44.6)	86 (16.2)	108 (20.3)	254.0 (128.2)	13.3 (6.3)

[a]Labels represent the following sentence categories: 0=background, 1=positive findings, 2=negative findings, and 3=continuation. Values in parentheses represent percentages of the row total.

XSL•FO
RenderX

Figure 2. Confusion matrix of GPT-4.1-mini versus radiologist consensus on the synthetic test set. Labels represent the following categories: 0=background, 1=positive findings, 2=negative findings, and 3=continuation. Per-class precision/recall: label 0=0.946/0.977; label 1=0.931/0.956; label 2=0.982/0.953; and label 3=0.841/0.851.

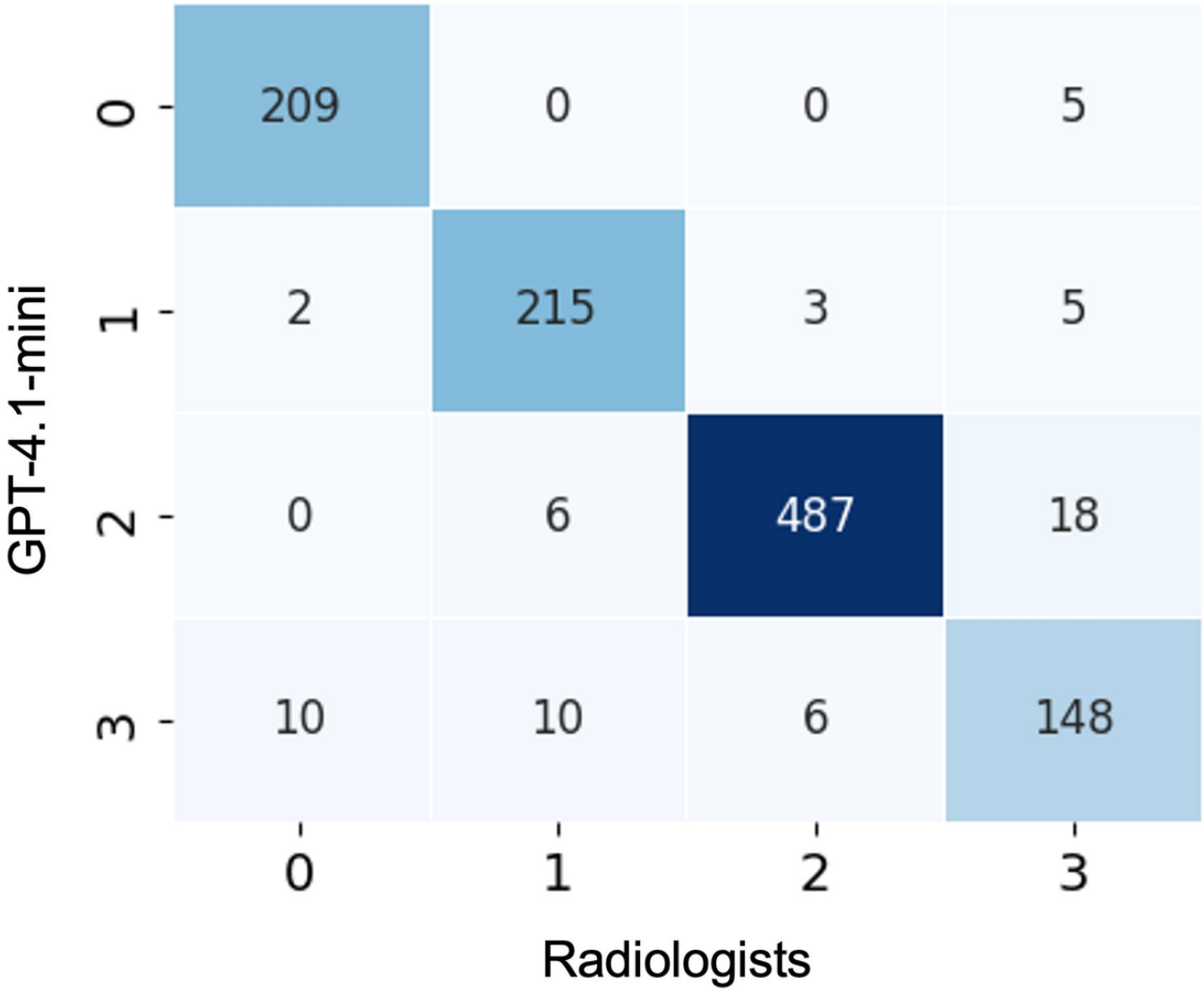

Table 3 presents the performance of the models on the internal test set, which consisted of 1124 sentences. Parsing errors were observed only in Qwen3 (0.6B) (3/1124 sentences, 0.27%), and no errors were observed in the other models. Accuracy ranged from 0.938 to 0.951; macro F_1-scores ranged from 0.924 to 0.940; PPV_1 ranged from 0.927 to 0.957; Recall_1 ranged from 0.898 to 0.960; and F1_1 ranged from 0.921 to 0.951.

Table 4 shows the results on the external validation dataset, which comprised 3477 sentences from 280 reports. Parsing errors were observed only in the LLM group, ranging from 19 to 260 sentences (0.55% - 7.48%), whereas no errors were observed in the text classification models. Accuracy ranged from 0.783 to 0.813; macro F_1-score ranged from 0.761 to 0.790; PPV_1 ranged from 0.897 to 0.952; Recall_1 ranged from 0.699 to 0.733; and F1_1 ranged from 0.794 to 0.821. Among all the models, Qwen3 (4B) demonstrated the best overall performance, maintaining high accuracy, Macro F_1-score, and PPV_1.

XSL•FO
RenderX

Table . Results on the internal synthetic test set[a].

Model	Parsing errors[b] (n, % of sentences)	Accuracy (95% CI)	Macro F_1[c] (95% CI)	PPV[d] for label 1 (95% CI)	Recall for label 1 (95% CI)	F1 for label 1 (95% CI)
LLMs[e] (name, parameters)						
Qwen3 (0.6B)	3 (0.27)	0.942 (0.930 - 0.955)	0.929 (0.913 - 0.944)	0.954 (0.926 - 0.978)	0.906 (0.870 - 0.943)	0.929 (0.906 - 0.951)
Qwen3 (1.7B)	0 (0)	0.951 (0.937 - 0.963)	0.940 (0.923 - 0.955)	0.941 (0.910 - 0.970)	0.915 (0.873 - 0.947)	0.928 (0.904 - 0.948)
Qwen3 (4B)	0 (0)	0.950 (0.937 - 0.962)	0.939 (0.922 - 0.954)	0.957 (0.930 - 0.981)	0.898 (0.860 - 0.941)	0.927 (0.903 - 0.950)
Llama 3.2 (1B)	0 (0)	0.949 (0.938 - 0.961)	0.940 (0.926 - 0.955)	0.949 (0.923 - 0.976)	0.911 (0.871 - 0.949)	0.930 (0.908 - 0.952)
Llama 3.2 (3B)	0 (0)	0.942 (0.928 - 0.956)	0.927 (0.908 - 0.944)	0.927 (0.889 - 0.960)	0.915 (0.884 - 0.948)	0.921 (0.897 - 0.945)
Text classification models						
BERT[f] base Japanese v3	0 (0)	0.944 (0.931 - 0.956)	0.931 (0.914 - 0.946)	0.939 (0.907 - 0.969)	0.960 (0.931 - 0.986)	0.949 (0.930 - 0.968)
JMedRoBERTa[g]-base	0 (0)	0.938 (0.924 - 0.952)	0.924 (0.905 - 0.940)	0.941 (0.910 - 0.969)	0.925 (0.886 - 0.959)	0.933 (0.909 - 0.956)
ModernBERT-Ja-130M	0 (0)	0.948 (0.936 - 0.959)	0.933 (0.917 - 0.948)	0.951 (0.925 - 0.973)	0.951 (0.921 - 0.977)	0.951 (0.934 - 0.968)

[a]95% CIs were estimated using the bootstrap method.

[b]Parsing errors: failures in JSON parsing during output generation.

[c]Macro F_1: macro-averaged F_1.

[d]PPV: positive predictive value.

[e]LLM: large language model.

[f]BERT: Bidirectional Encoder Representations from Transformers.

[g]JMedRoBERTa: Japanese Medical Robustly Optimized BERT Pretraining Approach.

XSL•FO
RenderX

Table . Results on the external validation dataset[a].

Model	Parsing errors[b] (n, % of sentences)	Accuracy (95% CI)	Macro F_1[c] (95% CI)	PPV[d] for label 1 (95% CI)	Recall for label 1 (95% CI)	F1 for label 1 (95% CI)
LLMs[e] (name, parameters)						
Qwen3 (0.6B)	81 (2.33)	0.783 (0.767 - 0.802)	0.761 (0.743 - 0.780)	0.919 (0.899 - 0.937)	0.699 (0.671 - 0.725)	0.794 (0.774 - 0.813)
Qwen3 (1.7B)	121 (3.48)	0.797 (0.780 - 0.815)	0.776 (0.759 - 0.794)	0.932 (0.916 - 0.948)	0.701 (0.673 - 0.729)	0.800 (0.779 - 0.822)
Qwen3 (4B)	19 (0.55)	0.813 (0.796 - 0.829)	0.790 (0.771 - 0.808)	0.952 (0.939 - 0.964)	0.721 (0.695 - 0.746)	0.821 (0.802 - 0.839)
Llama 3.2 (1B)	260 (7.48)	0.797 (0.779 - 0.813)	0.773 (0.754 - 0.790)	0.916 (0.897 - 0.933)	0.724 (0.697 - 0.751)	0.808 (0.789 - 0.829)
Llama 3.2 (3B)	70 (2.01)	0.806 (0.788 - 0.822)	0.782 (0.763 - 0.800)	0.919 (0.901 - 0.935)	0.733 (0.707 - 0.759)	0.816 (0.797 - 0.834)
Text classification models						
BERT[f] base Japanese v3	0 (0)	0.788 (0.771 - 0.804)	0.770 (0.752 - 0.786)	0.897 (0.877 - 0.916)	0.716 (0.688 - 0.743)	0.796 (0.774 - 0.816)
JMedRoBERTa[g]-base	0 (0)	0.799 (0.781 - 0.815)	0.783 (0.765 - 0.799)	0.932 (0.915 - 0.948)	0.709 (0.680 - 0.737)	0.806 (0.786 - 0.826)
ModernBERT-Ja-130M	0 (0)	0.801 (0.784 - 0.818)	0.782 (0.764 - 0.799)	0.917 (0.898 - 0.934)	0.722 (0.697 - 0.748)	0.808 (0.790 - 0.827)

[a]95% CIs were estimated using the bootstrap method.

[b]Parsing errors: failures in JSON parsing during output generation.

[c]Macro F_1: macro-averaged F_1.

[d]PPV: positive predictive value.

[e]LLM: large language model.

[f]BERT: bidirectional encoder representations from transformers.

[g]JMedRoBERTa: Japanese Medical Robustly Optimized BERT Pretraining Approach.

Table 5 summarizes the accuracy and sentence-level parsing errors for each institution. In institution C, which exhibited a notably high incidence of parsing errors, a detailed review of 40 reports revealed that all reports with failed parsing contained specific notations such as “[#1]” or “[#1 - 2]” within the text. These markers appeared to be used for referencing attached images, potentially reflecting a reporting style specific to that institution or individual radiologists.

Figure 3 shows the confusion matrices of Qwen3 (4B), the top-performing model, for the synthetic test set and external validation datasets.

An error analysis was conducted on the predictions of Qwen3 (4B) to identify factors contributing to the degradation of PPV_1. Among sentences incorrectly predicted as label 1, 3 recurring patterns were observed: (1) for label 0, 80% (8/10) of errors involved mistaking patient status or treatment history for active findings; (2) for label 2, 44% (4/9) of errors occurred in sentences mentioning a lesion name with weak negation or stability descriptors (eg, “unclear” or “maintained reduction”); and (3) for label 3, 41% (14/34) of misclassifications involved sentences describing differential diagnoses related to findings in the preceding text.

XSL•FO
RenderX

Table . Accuracy and parsing errors by institution[a].

Model/institution	A	B	C	D	E	F	G
LLMs[b] (name, parameters)							
Qwen3 (0.6B)	0.835 (1)	0.807 (3)	0.737 (2)	0.810 (1)	0.766 (4)	0.806 (9)	0.696
Qwen3 (1.7B)	0.842 (2)	0.797 (1)	0.746 (8)	0.858 (2)	0.786	0.814 (5)	0.724
Qwen3 (4B)	0.855	*0.816*	*0.754* (1)	*0.884* (1)	*0.805*	0.817 (1)	0.742
Llama 3.2 (1B)	0.853 (3)	0.810 (5)	0.722 (11)	0.835 (5)	0.805 (4)	0.806 (8)	0.719
Llama 3.2 (3B)	*0.857*	0.808 (6)	0.720 (6)	0.867 (1)	0.786 (1)	*0.826* (4)	0.734
Text classification models							
BERT[c] base Japanese v3	0.844	0.81	0.712	0.77	0.783	0.813	0.74
JMedRoBERTa[d]-base	0.844	0.814	0.74	0.796	0.793	0.815	*0.755*
ModernBERT-Ja-130M	0.841	0.802	0.709	0.834	0.8	0.822	0.753

[a]Numbers indicate accuracy; numbers in parentheses indicate the number of report-level parsing errors. Italicized values indicate the highest accuracy for each institution.

[b]LLM: large language model.

[c]BERT: Bidirectional Encoder Representations from Transformers.

[d]JMedRoBERTa: Japanese Medical Robustly Optimized BERT Pretraining Approach.

Figure 3. Performance of Qwen3 (4B) on the (A) internal synthetic test set and (B) external validation datasets. Labels represent the following categories: 0=background, 1=positive findings, 2=negative findings, and 3=continuation. Per-class precision/recall on the synthetic test set: Label 0 = 0.967/0.958; Label 1 = 0.957/0.898; Label 2 = 0.967/0.984; Label 3 = 0.873/0.908. On the external validation dataset: Label 0 = 0.739/0.947; Label 1 = 0.952/0.721; Label 2 = 0.847/0.933; and Label 3 = 0.561/0.692.

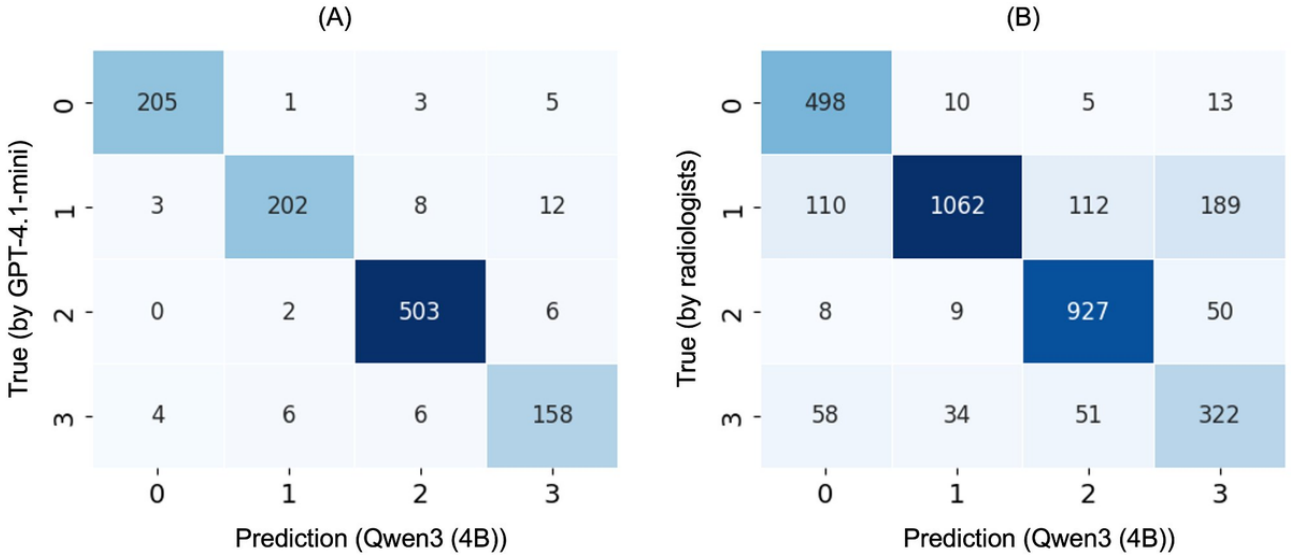

Discussion

Principal Findings

In this study, we developed context-aware sentence classification models for Japanese radiology reports using synthetic data and evaluated the models on multi-institutional radiology reports. All models demonstrated high performance on the internal synthetic test set (accuracy: 0.938 - 0.951; macro F_1-score: 0.924 - 0.940). Although performance declined on the external validation dataset (accuracy: 0.783 - 0.813; macro F_1-score: 0.761 - 0.790), PPV_1 remained stable and comparable to the internal results (eg, 0.957 vs 0.952 for Qwen3 [4B]). This discrepancy suggests that while the distributional gap between synthetic and real-world reports posed challenges for overall classification, the models effectively captured the core linguistic patterns and medical terminology associated with positive findings. From a practical standpoint, the stability of PPV_1 is

XSL•FO
RenderX

particularly important, as it indicates that models trained solely on synthetic data can reliably identify findings from diverse clinical reports without a high rate of false positives. This robustness supports the quality of the resulting image-text pairs, which is a prerequisite for the efficient large-scale development of VLMs. Although PPV_1 remained stable across datasets, Recall_1 declined from 0.898 to 0.960 on the synthetic test set to 0.699 to 0.733 on the external validation dataset, indicating that approximately 27% to 30% of true positive findings were not captured. In the context of VLM dataset construction, this precision-recall balance has practical implications. False-positive image-text pairs, where a nonfinding sentence is incorrectly paired with an image, can actively introduce noise into model training. In contrast, false negatives reduce dataset coverage but do not compromise the quality of the retained pairs. Therefore, the observed high-precision, moderate-recall profile is considered to be acceptable for the initial construction of reliable image-text datasets, while improving recall through prompt refinement and domain adaptation remains an important direction for future work.

Prior Work and the Sentence-Level Approach

Previous studies in radiology report that structuring has predominantly utilized entity-level extraction (eg, extracting specific disease names) or graph-based representations involving nodes and edges [36-38]. These paradigms offer distinct advantages: entity-level extraction allows for the highly granular identification of specific clinical findings and facilitates the precise handling of negation and uncertainty expressions, while graph-based approaches excel at representing complex relationships between anatomical locations and pathologies. More recently, NLP applications have expanded to include the labeling of anatomical phrases, paraphrasing of clinical statements, and full report structuring [37,39,40].

In contrast, this study adopted a sentence-level labeling approach. Although this method faces challenges with sentences containing multiple distinct findings (eg, “A liver cyst is present, but no renal cyst is observed”) [41], our policy prioritized label 1. In many clinical instances, positive and negative statements coexisting in a single sentence refer to the same lesion (eg, “A hypodense lesion is present in the liver, but it shows no corresponding contrast enhancement”). Our framework categorizes such descriptions as label 1. If the clauses in this example were split into 2 separate sentences, they would be labeled as label 1 followed by label 3. By utilizing the combination of label 1 and its associated label 3 in downstream tasks, nearly equivalent semantic information can be obtained regardless of whether the description is contained within a single sentence or distributed across multiple sentences. Therefore, we argue that this approach remains robust for practical clinical applications. As a preliminary analysis, we examined the potential impact of multiclause sentences on the classification performance using Qwen3 (4B) predictions on the external validation dataset; the results are detailed in Section F in Multimedia Appendix 1. While implementing an LLM-based preprocessing step to segment complex sentences into distinct, meaning-preserving statements could further refine this pipeline, we deferred this in this study to avoid increasing procedural complexity and the difficulty of establishing evaluation metrics. More importantly, we believe that preserving sentence-level expressions—rather than reducing them to simple entity extraction—is highly beneficial for VLM training. This approach retains the nuanced phrasing and the radiologist’s diagnostic thought process. Preserving these linguistic characteristics provides richer textual information for the model to learn the association between visual findings and professional reporting styles, which is essential for developing clinically relevant models [4,42].

Challenges in Clinical Diversity and Prompt Optimization

The variability in performance across institutions (Table 5) provides key insights for implementation. Table 2 reveals significant diversity in report length and label distribution. Notably, facilities with metrics similar to the synthetic data—such as report length and finding proportions—did not always yield the highest accuracy. This suggests that clinical diversity involves complex factors beyond aggregate metrics, including disease prevalence, institution-specific rules, and individual radiologist styles. Although our prompts incorporated various elements, they do not yet fully align with real-world practice. Potential refinements include adjusting demographic granularity (eg, using numerical values instead of categorical descriptors), reweighting clinical scenarios (eg, disease frequency, emergency vs inpatient settings, or the prevalence of poor image quality and motion artifacts), and fine-tuning class distributions (eg, the ratio of positive to negative findings). Future research must refine prompt design to address this reporting diversity, which remains the primary challenge in utilizing a purely synthetic training set.

Error analyses identified several technical and clinical challenges. Parsing errors in institution C likely resulted from image-referencing notations such as “[#1],” which the LLM may have misinterpreted as structural control characters. For Qwen3 (4B), false-positive label 1 predictions revealed specific patterns: (1) 80% (8/10) of label 0 errors involved mistaking patient history for active findings; (2) 44% (4/9) of label 2 errors involved lesion names paired with weak negations or stability descriptors, such as “unclear” or “maintained reduction”; and (3) 41% (14/34) of label 3 errors occurred in sentences describing differential diagnoses related to preceding findings. These results emphasize the importance of incorporating complex negations, facility-specific symbols, and sophisticated clinical contexts into future synthetic data prompts [43,44]. While this study focused on validating general extraction performance, institutional domain adaptation and post hoc prompt adjustments remain subjects for future investigation.

Practical Considerations for Clinical Deployment

Text classification models, including BERT base Japanese v3, JMedRoBERTa-base, and ModernBERT-Ja-130M, showed stable performance compared to those of small-scale LLMs. From a practical standpoint, these classification models offer distinct advantages in reliability and processing speed (detailed in Section D in Multimedia Appendix 1). Their architecture inherently eliminates the risk of format instabilities while ensuring high throughput even on standard hardware [45]. While requiring more computational resources than BERT-based

XSL•FO
RenderX

models, the decision to evaluate small-scale LLMs was similarly rooted in the need for on-premise feasibility. These models with under 4 billion parameters can operate on a single mid-range GPU with 12 to 24 GB of Video RAM (eg, NVIDIA RTX 4090 or RTX 4070 Ti Super). This allows institutions to maintain strict data privacy and ensure operational stability without relying on external cloud-based services.

Parsing errors remained a unique challenge to the LLM group [46]. While we considered using vLLM, an open-source library optimized for high-throughput LLM serving, to implement JSON mode, we intentionally opted against it. JSON mode is a constrained decoding technique provided by the inference framework rather than a native capability of the LLMs themselves. Our primary objective was to evaluate the models' inherent ability to follow structural instructions without external enforcement. Furthermore, we found that such rigid constraints could lead to unintended text alterations or premature output truncation. We also investigated whether a re-inference strategy could resolve these parsing errors, with the results detailed in Section E in Multimedia Appendix 1.

Limitations

This study had some limitations. First, most training labels in the synthetic data were generated automatically without individual human verification. While a quality assessment of the synthetic test set, comprising 1124 sentences from 100 reports, showed high reliability (κ=0.917), the remaining unverified training data may contain annotation noise that potentially constrained the models' final performance. The iterative refinement of these labels through expert review and subsequent fine-tuning could represent a promising avenue to further enhance model accuracy and robustness.

Second, although Japanese served as a representative example of a low-resource language in the medical domain, the generalizability of this approach to other underrepresented languages has not yet been verified. As noted in the section "Challenges in Clinical Diversity and Prompt Optimization," there remains room for improvement in our synthetic data generation process. Applying this methodology to other languages would necessitate additional tuning to align with specific linguistic nuances and localized clinical reporting conventions.

Third, the comparison between LLMs and text classification models involved differing input contexts. LLMs processed full reports to leverage their expansive context windows and maximize throughput. In contrast, text classification models utilized a 5-sentence sliding window due to inherent input length constraints. Restricting LLMs to identical local windows would have artificially limited their native processing capacity and practical utility. Therefore, these results should be interpreted as a comparison of each model in its optimal configuration rather than under strictly uniform input conditions.

Future Directions

Future work should focus on integrating expert-annotated real-world reports for iterative model fine-tuning and establishing a more robust pipeline. This includes the implementation of advanced preprocessing steps, such as LLM-based sentence segmentation, to accurately handle descriptions containing multiple clinical findings. Furthermore, a primary objective will be to generate high-quality image-text pairs from these structured reports to facilitate the development of medical vision-language foundation models.

Conclusions

This study demonstrated the feasibility of developing context-aware sentence classification models for Japanese radiology reports using a training pipeline based entirely on synthetic data, thereby substantially reducing the need for labor-intensive manual labeling. Across multiple model architectures, high classification performance was achieved on the synthetic test set. Although external validation using multi-institutional, real-world reports showed a consistent pattern of performance degradation, the models maintained a stable positive predictive value for positive findings. These results indicate that the models effectively captured the essential clinical terminology and linguistic patterns required to identify positive imaging findings in real-world reports.

Acknowledgments

The authors would like to thank the departments of radiology that provided the Japan Medical Image Database (J-MID), including Juntendo University, Kyushu University, Keio University, The University of Tokyo, Okayama University, Kyoto University, Osaka University, Tokyo Medical and Dental University, Hokkaido University, Ehime University, and Tokushima University. Artificial intelligence–assisted technologies (ChatGPT; OpenAI) were used for English language editing. The authors take full responsibility for the content of the article.

Funding

This research was supported by Accreditation Organization for Management of Radiologic Imaging (AOMRI) Research Grant 2025 and by Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" (grant JPJ012425).

Data Availability

The synthetic data with the annotation scripts and the three fine-tuned text classification models used in this study are available via a GitHub repository [47]. The original clinical data obtained from the Japan Medical Image Database (J-MID) cannot be shared publicly due to ethical and privacy restrictions.

XSL•FO
RenderX

Authors' Contributions

T Kikuchi conceived and designed the study, performed data curation, formal analysis, and investigation, developed the methodology, administered the project, created the visualizations, and drafted the manuscript. YY contributed to conceptualization, formal analysis, investigation, and methodology. KY contributed to conceptualization, data curation, and methodology. TA contributed to data curation and critical revision of the manuscript. H Mori, H Makimoto, and T Kohro critically reviewed and edited the manuscript. All authors reviewed the final version of the manuscript, approved its submission, and agreed to be accountable for all aspects of this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prompts for synthetic report generation and annotation, model card information, throughput benchmarks, parsing error analysis, and subanalysis of label 1 predictions by sentence complexity.
[PDF File, 296 KB - jmir_v28i1e86365_app1.pdf]

References

1. Zhang K, Zhou R, Adhikarla E, et al. A generalist vision-language foundation model for diverse biomedical tasks. Nat Med 2024 Nov;30(11):3129-3141. [doi: 10.1038/s41591-024-03185-2] [Medline: 39112796]
2. Sun Y, Wen X, Zhang Y, et al. Visual-language foundation models in medical imaging: a systematic review and meta-analysis of diagnostic and analytical applications. Comput Methods Programs Biomed 2025 Aug;268:108870. [doi: 10.1016/j.cmpb.2025.108870] [Medline: 40424873]
3. Danish S, Sadeghi-Niaraki A, Khan SU, Dang LM, Tightiz L, Moon H. A comprehensive survey of vision–language models: pretrained models, fine-tuning, prompt engineering, adapters, and benchmark datasets. Information Fusion 2026 Feb;126:103623. [doi: 10.1016/j.inffus.2025.103623]
4. Yi Z, Xiao T, Albert MV. A survey on multimodal large language models in radiology for report generation and visual question answering. Information 2025 Feb 12;16(2):136. [doi: 10.3390/info16020136]
5. Yamamoto K, Kikuchi T. TotalFM: an organ-separated framework for 3D-CT vision foundation models. arXiv. Preprint posted online on Jan 1, 2026. [doi: 10.48550/arXiv.2601.00260]
6. Hartsock I, Rasool G. Vision-language models for medical report generation and visual question answering: a review. Front Artif Intell 2024;7:1430984. [doi: 10.3389/frai.2024.1430984] [Medline: 39628839]
7. Ryu JS, Kang H, Chu Y, Yang S. Vision-language foundation models for medical imaging: a review of current practices and innovations. Biomed Eng Lett 2025 Sep;15(5):809-830. [doi: 10.1007/s13534-025-00484-6] [Medline: 40917147]
8. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature New Biol 2023 Apr 13;616(7956):259-265. [doi: 10.1038/s41586-023-05881-4] [Medline: 37045921]
9. Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 2019 Dec 12;6(1):317. [doi: 10.1038/s41597-019-0322-0] [Medline: 31831740]
10. Hamamci IE, Er S, Wang C, et al. Generalist foundation models from a multimodal dataset for 3D computed tomography. Nat Biomed Eng 2026 Feb 12. [doi: 10.1038/s41551-025-01599-y] [Medline: 41680439]
11. Yamagishi Y, Nakamura Y, Kikuchi T, et al. Development of a large-scale dataset of chest computed tomography reports in Japanese and a high-performance finding classification model: dataset development and validation study. JMIR Med Inform 2025 Aug 28;13:e71137. [doi: 10.2196/71137] [Medline: 40874833]
12. Sourlos N, Vliegenthart R, Santinha J, et al. Recommendations for the creation of benchmark datasets for reproducible artificial intelligence in radiology. Insights Imaging 2024 Oct 14;15(1):248. [doi: 10.1186/s13244-024-01833-2] [Medline: 39400639]
13. Hirano Y, Hanaoka S, Nakao T, et al. GPT-4 Turbo with Vision fails to outperform text-only GPT-4 Turbo in the Japan Diagnostic Radiology Board Examination. Jpn J Radiol 2024 Aug;42(8):918-926. [doi: 10.1007/s11604-024-01561-z] [Medline: 38733472]
14. Hirano Y, Hanaoka S, Nakao T, et al. No improvement found with GPT-4o: results of additional experiments in the Japan Diagnostic Radiology Board Examination. Jpn J Radiol 2024 Nov;42(11):1352-1353. [doi: 10.1007/s11604-024-01622-3] [Medline: 38937409]
15. Bustos A, Pertusa A, Salinas JM, de la Iglesia-Vayá M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. Med Image Anal 2020 Dec;66:101797. [doi: 10.1016/j.media.2020.101797] [Medline: 32877839]
16. de Castro DC, Bustos A, Bannur S, et al. PadChest-GR: a bilingual chest X-ray dataset for grounded radiology report generation. NEJM AI 2025 Jun 26;2(7). [doi: 10.1056/AIdbp2401120]
17. Koçak B, Ponsiglione A, Stanzione A, et al. Bias in artificial intelligence for medical imaging: Fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. Diagn Interv Radiol 2025 Mar 3;31(2):75-88. [doi: 10.4274/dir.2024.242854] [Medline: 38953330]

XSL•FO
RenderX

18. Mukherjee P, Hou B, Suri A, et al. Evaluation of GPT large language model performance on RSNA 2023 Case of the Day questions. .
19. Kikuchi T, Nakao T, Nakamura Y, Hanaoka S, Mori H, Yoshikawa T. Toward improved radiologic diagnostics: investigating the utility and limitations of GPT-3.5 Turbo and GPT-4 with quiz cases. AJNR Am J Neuroradiol 2024 Oct 3;45(10):1506-1511. [doi: 10.3174/ajnr.A8332] [Medline: 38719605]
20. Choi WC, Chang CI. Advantages and limitations of open-source versus commercial large language models (LLMs): a comparative study of DeepSeek and OpenAI's ChatGPT. Preprints.org. Preprint posted online on Mar 14, 2025. [doi: 10.20944/preprints202503.1081.v1]
21. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? arXiv. Preprint posted online on Jul 18, 2023. [doi: 10.48550/arXiv.2307.09009]
22. Belcak P, Heinrich G, Diao S, et al. Small language models are the future of agentic AI. arXiv. Preprint posted online on Jun 2, 2025. [doi: 10.48550/arXiv.2506.02153]
23. Bumgardner VKC, Mullen A, Armstrong SE, Hickey C, Marek V, Talbert J. Local large language models for complex structured tasks. AMIA Jt Summits Transl Sci Proc 2024:105-114. [Medline: 38827047]
24. Akashi T, Kumamaru KK, Wada A, et al. Japan-Medical Image Database (J-MID): medical big data supporting data science. Juntendo Med J 2025;71(3):166-172. [doi: 10.14789/ejmj.JMJ25-0004-P] [Medline: 40666496]
25. Yang A, Li A, Yang B, et al. Qwen3 technical report. arXiv. Preprint posted online on May 14, 2025. [doi: 10.48550/arXiv.2505.09388]
26. Grattafiori A, Dubey A, Jauhri A, et al. The Llama 3 herd of models. arXiv. Preprint posted online on Jul 31, 2024. [doi: 10.48550/arXiv.2407.21783]
27. Warner B, Chaffin A, Clavié B, et al. Smarter, better, faster, longer: a modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. Presented at: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)(ACL 2025); Jul 27 to Aug 1, 2025; Vienna, Austria p. 2526-2547. [doi: 10.18653/v1/2025.acl-long.127]
28. Devlin J, Chang MW, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on Oct 11, 2018. [doi: 10.48550/arXiv.1810.04805]
29. Liu Y, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. Preprint posted online on Jul 26, 2019. [doi: 10.48550/arXiv.1907.11692]
30. Yasaka K, Nomura T, Kamohara J, et al. Classification of interventional radiology reports into technique categories with a fine-tuned large language model. J Imaging Inform Med 2025 Oct;38(5):3366-3374. [doi: 10.1007/s10278-024-01370-w] [Medline: 39673010]
31. Kanzawa J, Yasaka K, Fujita N, Fujiwara S, Abe O. Automated classification of brain MRI reports using fine-tuned large language models. Neuroradiology 2024 Dec;66(12):2177-2183. [doi: 10.1007/s00234-024-03427-7] [Medline: 38995393]
32. Sugimoto K, Iki T, Chida Y, et al. JMedRoBERTa: a Japanese pre-trained language model on academic articles in medical sciences [Article in Japanese]. 2023 Presented at: Proceedings of the 29th Annual Meeting of the Association for Natural Language Processing p. 707-712 URL: https://www.anlp.jp/proceedings/annual_meeting/2023/pdf_dir/P3-1.pdf [accessed 2026-03-22]
33. Yamagishi Y, Kikuchi T, Hanaoka S, et al. ModernBERT is more efficient than conventional BERT for chest CT findings classification in Japanese radiology reports. arXiv. Preprint posted online on Mar 7, 2025. [doi: 10.48550/arXiv.2503.05060]
34. Hu EJ, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models. arXiv. Preprint posted online on Jun 17, 2021. [doi: 10.48550/arXiv.2106.09685]
35. Dettmers T, Holtzman A, Pagnoni A, Zettlemoyer L. QLoRA: efficient finetuning of quantized LLMs. Presented at: 37th Conference on Neural Information Processing Systems (NeurIPS 2023); Dec 10-16, 2023. [doi: 10.52202/075280-0441]
36. Peng Y, Wang X, Lu L, Bagheri M, Summers R, Lu Z. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. AMIA Jt Summits Transl Sci Proc 2018;2017:188-196. [Medline: 29888070]
37. Jain S, Agrawal A, Saporta A, et al. RadGraph: extracting clinical entities and relations from radiology reports. Presented at: 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks; Dec 6-14, 2021 URL: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/c8ffe9a587b126f152ed3d89a146b445-Paper-round1.pdf [accessed 2026-03-17]
38. Kahn CE Jr, Langlotz CP, Burnside ES, et al. Toward best practices in radiology reporting. Radiology 2009 Sep;252(3):852-856. [doi: 10.1148/radiol.2523081992] [Medline: 19717755]
39. Zhu H, Paschalidis IC, Hall C, Tahmasebi A. Context-driven concept annotation in radiology reports: anatomical phrase labeling. AMIA Jt Summits Transl Sci Proc 2019;2019:232-241. [Medline: 31258975]
40. Bergomi L, Buonocore TM, Antonazzo P, et al. Reshaping free-text radiology notes into structured reports with generative question answering transformers. Artif Intell Med 2024 Aug;154:102924. [doi: 10.1016/j.artmed.2024.102924] [Medline: 38964194]
41. Fujikawa K, Seki K, Uehara K. A hybrid approach to finding negated and uncertain expressions in biomedical documents. 2012 Presented at: Proceedings of the 2nd International Workshop on Managing Interoperability and Complexity in Health Systems; Oct 28 to Nov 2, 2012; Maui, Hawaii, USA p. 67-74. [doi: 10.1145/2389672.2389685]

XSL•FO
RenderX

42. Hamma MB, Merrouni ZA, Frikh B, Ouhbi B. Automatic radiology report generation: a comprehensive review and innovative framework. Presented at: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 3-6, 2024; Lisbon, Portugal p. 4225-4232. [doi: 10.1109/BIBM62325.2024.10821974]
43. Casey A, Davidson E, Poon M, et al. A systematic review of natural language processing applied to radiology reports. BMC Med Inform Decis Mak 2021 Jun 3;21(1):179. [doi: 10.1186/s12911-021-01533-7] [Medline: 34082729]
44. Delbrouck JB, Chambon P, Bluethgen C, Tsai E, Almusa O, Langlotz C. Improving the factual correctness of radiology report generation with semantic rewards. Presented at: Findings of the Association for Computational Linguistics: EMNLP 2022; Dec 7-11, 2022; Abu Dhabi, United Arab Emirates p. 4348-4360. [doi: 10.18653/v1/2022.findings-emnlp.319]
45. Zhang J, Huang Y, Liu S, Gao Y, Hu X. Do BERT-like bidirectional models still perform better on text classification in the era of LLMs? Presented at: Findings of the Association for Computational Linguistics: EMNLP 2025; Nov 4-9, 2025; Suzhou, China p. 18980-18989. [doi: 10.18653/v1/2025.findings-emnlp.1033]
46. Geng S, Cooper H, Moskal M, et al. JSONSchemaBench: a rigorous benchmark of structured outputs for language models. arXiv. Preprint posted online on Jan 18, 2025. [doi: 10.48550/arXiv.2501.10868]
47. Kikuchi T. Context-Aware-Rad-Sentence-Classification. GitHub. URL: https://github.com/jichi-labo/Context-Aware-Rad-Sentence-Classification [accessed 2026-03-22]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
GPU: graphics processing unit
J-MID: Japan Medical Image Database
JMedRoBERTa: Japanese Medical Robustly Optimized BERT Pretraining Approach
LLM: large language model
LoRA: low-rank adaptation
Macro F_1: macro-averaged F_1
MIMIC-CXR: Medical Information Mart for Intensive Care Chest X-Ray
NLP: natural language processing
PPV: positive predictive value
QLoRA: quantized low-rank adaptation
RoBERTa: Robustly Optimized BERT Pretraining Approach
VLM: vision-language model

Edited by A Coristine; submitted 23.Oct.2025; peer-reviewed by D Liu, J Zhang; revised version received 28.Feb.2026; accepted 28.Feb.2026; published 10.Apr.2026.

Please cite as:
Kikuchi T, Yamagishi Y, Yamamoto K, Akashi T, Mori H, Makimoto H, Kohro T
Context-Aware Sentence Classification of Radiology Reports Using Synthetic Data: Development and Validation Study
J Med Internet Res 2026;28:e86365
URL: https://www.jmir.org/2026/1/e86365
doi:10.2196/86365

© Tomohiro Kikuchi, Yosuke Yamagishi, Kohei Yamamoto, Toshiaki Akashi, Harushi Mori, Hisaki Makimoto, Takahide Kohro. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Error Detection in Emergency Radiology Reports Using a Large Language Model: Multistage Evaluation Study

Hui Shen[1], MD; Tianyang Wu[2], BS; Fei Wang[1], MD, PhD; Jin Fang[1], MD, PhD; Yuange Li[3], MD, PhD; Xiaoling Wu[1], MD; Shuyi Liu[4], MD, PhD; Liting Chen[1], MD, PhD; Qiuping Ren[5], MD, PhD; Xiao Meng[2], BS; Jiatong Xu[2], BS; Jie Sun[1], MD; Yujie Zhao[1], MD; Xin Liu[1], MD; Liaoyuan Wang[1], MD; Guipeng Mai[6], BS; Jingjing You[1], MD; Zhe Jin[1], MD; Xuewei Wu[1], MD; Wenle He[1], MD; Xue Han[1], MD; Shuixing Zhang[1], MD, PhD; Dong Zeng[2], MD, PhD; Bin Zhang[1], MD, PhD

[1]Department of Radiology, The First Affiliated Hospital of Jinan University, No. 613 Huangpu West Road, Tianhe, Guangzhou, Guangdong, China
[2]School of Biomedical Engineering, Southern Medical University, Guangzhou, Guangdong, China
[3]Department of Radiology, The Affiliated Hospital of Guangdong Medical University, Zhanjiang, Guangdong, China
[4]Department of Radiology, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, Guangdong, China
[5]Department of Radiology, Nanfang Hospital, Southern Medical University, Guangzhou, Guangdong, China
[6]Department of Radiology, Longhu District People's Hospital of Shantou, Shantou, Guangdong, China

Corresponding Author:
Bin Zhang, MD, PhD
Department of Radiology, The First Affiliated Hospital of Jinan University, No. 613 Huangpu West Road, Tianhe, Guangzhou, Guangdong, China

Abstract

Background: Emergency radiology requires highly accurate reporting under time constraints; yet, increasing workloads raise the risk of errors. While large language models (LLMs) show potential for proofreading in general radiology, their performance in emergency settings and non-English contexts remains unclear.

Objective: We aim to evaluate the performance of a domain-optimized LLM, DeepSeek-R1, for identifying errors in Chinese emergency radiology reports, with comparison against assessments by board-certified radiologists.

Methods: We compiled 7435 emergency reports (dataset 1; radiography, computed tomography, and magnetic resonance imaging) collected from November 2024 to April 2025. In stage 1, a total of 5 LLMs were evaluated using 200 reports. The best model, DeepSeek-R1, proceeded to stages 2 and 3, where 0-shot and few-shot learning were tested on a separate set (n=100). Model performance was compared against 12 radiologists. Stage 4 validated real-world utility on 800 verified reports.

Results: In subdataset 1, under stress-testing conditions, DeepSeek-R1 achieved a higher error detection rate in the few-shot setting than in the 0-shot setting (84.4% vs 60.9%, P=.003). Its performance exceeded that of radiology residents (84.4% vs 51.6% and 53.1%, respectively; both P<.05) and showed no statistically significant difference compared with senior radiologists and attending radiologists (84.4% vs 68.8% - 93.8%, P=.26 to ≥.99). Compared with residents, DeepSeek-R1 detected more critical omissions (100% vs 25% and 50%; both P<.05) and other errors (92% vs 33% and 33%; both P=.02). In dataset 2, collected from independent institutions, DeepSeek-R1 achieved a detection rate of 95% under the few-shot setting. Reading time was shorter than that of human readers (92 vs 109 s). In real-world validation, DeepSeek-R1 identified 117 true reporting errors, yielding a positive predictive value of 56.5%.

Conclusions: DeepSeek-R1 holds promise for improving quality control in emergency radiology reports. Its performance and efficiency support its use as an assistive proofreading tool in real-world radiology workflows.

(J Med Internet Res 2026;28:e86841) doi:10.2196/86841

KEYWORDS

emergency radiology; large language models; error detection; quality control

Introduction

Emergency radiology reports are critical for guiding timely and accurate patient management, particularly in trauma, acute illness, and other time-sensitive clinical scenarios. Errors in these reports, ranging from factual inaccuracies (eg, laterality confusion) to interpretive inconsistencies, can directly impact patient outcomes, leading to misdiagnoses, delayed treatments, or inappropriate management [1-3]. The growing volume of emergency cases, coupled with radiologist shortages and high

XSL•FO
RenderX

workloads, further increases the likelihood of mistakes [4-7]. Traditional proofreading methods, such as double-reading by senior radiologists, are effective but time-consuming and often impractical in fast-paced emergency settings [8].

Recent advances in large language models (LLMs) provide promising solutions for automating error detection [9,10]. Studies have shown that LLMs such as GPT-4 (OpenAI Inc) and Claude 3.5 Sonnet (Anthropic PBC) can identify inconsistencies in radiology reports with accuracy comparable to human experts [11-13]. For instance, GPT-4 achieved an error detection rate of 82.7% in general radiology reports, comparable to senior radiologists (89.3%) while reducing reading time from 25.1 to 3.5 seconds per report [12]. Similarly, Claude 3.5 Sonnet outperformed radiologists in detecting factual errors in head computed tomography (CT) reports, with a sensitivity of 89% compared to 33% - 69% among human readers [13]. In Chinese ultrasound reports, LLMs show promise in identifying spelling and logical errors, with Claude 3.5 Sonnet achieving a 52.3% detection rate in 0-shot settings [14]. Despite these advances, several critical gaps remain. First, LLM performance in emergency radiology, a domain characterized by fragmented information, urgent decision-making, and complex multimodal findings, remains unexplored. Second, most studies used synthetically generated errors or single-institution data, limiting generalizability to real-world emergency scenarios [12,13]. Third, while GPT-4 and Claude show strengths in English contexts, their efficacy in non-English emergency reports (eg, Chinese) is suboptimal due to linguistic nuances and terminology variations [15,16]. DeepSeek-R1, a state-of-the-art LLM specifically optimized for Chinese clinical text, offers a valuable opportunity to address these limitations [17,18]. Its architecture integrates domain-specific pretraining on multilingual clinical corpora, potentially enhancing error detection in non-English emergency reports [19]. However, rigorous validation using real-world emergency radiology data is still lacking.

Therefore, this study aims to evaluate DeepSeek-R1's ability to identify errors in real-world Chinese emergency radiology reports, benchmark its performance against both other LLMs and radiologists of varying experience levels, and assess its potential for integration into clinical workflows. With a focus on practical applicability, this work strives to establish a new benchmark for artificial intelligence (AI)–assisted quality assurance in emergency radiology.

Methods

Ethical Considerations

This retrospective study was approved by the Ethics Review Board of the First Affiliated Hospital of Jinan University (20250306). The requirement for informed consent was waived due to this study's retrospective nature. All patient data used during this study were strictly anonymized. No personally identifiable information was disclosed to any LLMs, thereby ensuring patient privacy and adherence to ethical standards. Participants received no compensation.

Dataset and Error Categories

A total of 7435 medical records between November 2024 and April 2025 were collected from the First Affiliated Hospital of Jinan University. The dataset 1 consisted of 3 types of emergency radiology reports: CT (n=5237), magnetic resonance imaging (MRI; n=381), and radiography (n=1817). Each report included 4 sections: patient information, examination items, findings, and impression. To increase the diversity of the dataset, dataset 2 consisted of 50 error-free emergency radiology reports that passed quality control and 50 erroneous reports that did not meet quality control, collected from 2 hospitals: Nanfang Hospital, Southern Medical University, and the Affiliated Hospital of Guangdong Medical University. Both unreviewed and senior physician-reviewed versions are available for each record. All patient identifiers were anonymized before inclusion in the dataset.

First, a total of 200 error-free reports were randomly selected from dataset 1 to form a subdataset 1 using a freely available research data randomization tool [20], all of which were verified by human experts to ensure accuracy. These reports served as the basis for evaluating the error detection capabilities of both LLMs and physicians. Then, artificial errors were introduced into 100 of these reports through a randomized process. Informed by previous studies on common error types in radiology reports [21-23], the error categories included (1) omission: the omission of relevant words or phrases, including both deletions and missing words (eg, "fracture" instead of "no fracture"); (2) insertion: the unintentional insertion of incorrect words or phrases, including inappropriate terms, substitutions, insertions, or word confusions (eg, "abnormal" instead of "normal"); (4) spelling: Chinese spelling errors that may arise from rapid typing using Pinyin input methods (eg, "纵隔" misspelled as "纵膈," "椎体" misspelled as "锥体"); (5) side confusion: laterality errors (eg, "right" instead of "left," "lateral" instead of "medial"); and (6) other errors: errors not fitting the above categories, such as incorrect dates, errors in image or series numbering, unit discrepancies (eg, centimeter vs millimeter), template-related inaccuracies, and punctuation errors. Error definitions, applications, and distribution of each error type are detailed in Table 1, and Figures S1 and S2 in Multimedia Appendix 1.

XSL•FO
RenderX

Table . Types of errors, their descriptions, and examples used in this study. The examples presented in the table are actual errors identified by DeepSeek-R1 from original emergency radiology reports of the First Affiliated Hospital of Jinan University, with the erroneous portions shown in italics.

Error type	Description	Example	
		Finding	Impression
Omission	The omission of relevant words or phrases, including both deletions and missing words	There is no evidence of hemorrhage, edema, mass effect, or infarction. "*The visualized bilateral maxillary sinuses show mucosal thickening with patchy fluid density shadows within.*"	*Normal study*
Insertion	The unintentional insertion of incorrect words or phrases, including inappropriate terms, substitutions, insertions, or word confusions	There is a fracture at the inferior pole of the patella with displacement of the larger fragment inferiorly and anteriorly, and displacement of a smaller bony fragment inferiorly. Alignment within the right knee joint is appropriate. No further pathologic step-offs or interruptions of the cortex are delineated. The bone trabeculae are homogeneously structured.	*Nondisplaced* fracture at the inferior pole of the patella.
Spelling	Chinese spelling errors that may arise from rapid typing using Pinyin input methods	Disruption of cortical continuity is noted in the distal phalanx of the left *hallux* (拇趾), with an identifiable transverse or radiolucent fracture line.	Mildly displaced fracture of the distal phalanx of the left *hallux* (拇指), *which is a spelling error in Chinese.*
Side confusion	Laterality errors	Adjacent to the middle phalanx of the *left* fifth finger, there is a small, well-defined osseous density fragment.	Small osseous density adjacent to the *right* fifth middle phalanx, suspicious for an avulsion fracture.
Other errors	Errors not fitting the above categories, such as incorrect dates, errors in image or series numbering, unit discrepancies, template-related inaccuracies, and punctuation errors	The liver demonstrates smooth contours and normal proportions. Multiple rounds of low-attenuation lesions are present within the hepatic parenchyma, with well-defined margins. The largest lesion measures approximately *1.2 cm* in diameter.	Multiple low-attenuation hepatic lesions, the largest measuring approximately *1.2 mm*.

Clinical experts methodically implanted 127 errors into the modified reports, with no more than 2 errors per report. Therefore, a comparative dataset of 100 error-free and 100 erroneous reports was established to evaluate the diagnostic discrepancy detection capabilities of both LLMs and radiologists. Only the errors intentionally inserted into the text were used as the reference standard. To ensure the absence of any additional errors in the radiology reports, 3 readers (HS, XW, and BZ, with 5, 5, and 8 years of experience, respectively) independently reviewed the reports. Discrepancies were resolved via consensus review among 3 readers.

Two types of prompts were developed for the LLMs: 0-shot prompts and few-shot prompts. The few-shot prompt was constructed by incorporating 6 additional emergency radiology report samples into the 0-shot prompt framework, while maintaining consistency across all other prompt components. These 6 example reports were randomly selected from the pool of emergency radiology reports, with deliberate attention to ensuring diagnostic diversity and coverage of common error types, including item omission, insertion, spelling errors, laterality confusion, and other frequently encountered reporting mistakes. This stratified yet random selection strategy was intended to reflect the variability of real-world radiology reporting without biasing the model toward any specific diagnosis or error category. The anonymized example reports used for few-shot prompting are provided in Table S1 in Multimedia Appendix 1. Each prompt was structured into three main sections: (1) role context and task description, (2) definitions of report error types, and (3) output constraints (Table 2). In the 0-shot setting, the prompt was specifically designed with an emphasis on Chinese medical terminology to ensure alignment with the linguistic and domain-specific characteristics of emergency radiology reports in clinical practice. While this domain-specific tailoring was intended to improve task relevance, we acknowledge that it may introduce a performance advantage for DeepSeek-R1, which has been trained on corpora with similar linguistic and clinical characteristics.

XSL•FO
RenderX

Table . Detailed prompts and parameters of large language models used in this study.

Prompt name	Prompt text
0-shot prompt	You are a radiology physician. In the subsequent text, I will provide you with an emergency radiology report containing the following three sections: “Examination details” “Findings” and “Impression”, please assess the report for potential inaccuracies. If you identify any of the following error types, please address the corresponding issues. Omission: Discrepancies between imaging findings and imaging diagnosis that may lead to ambiguity, missed diagnosis, or misdiagnosis. For example, a positive lesion mentioned in imaging findings is not referred in the imaging diagnosis. Insertion: Unintentional inclusion of incorrect terms or phrases within imaging findings or radiological diagnosis, such as improper word substitutions or insertions. Spelling errors: Spelling mistakes caused by rapid typing of chinese characters, such as substitutions of homophones that may alter the intended meaning. Positional discrepancies: Inconsistencies between imaging descriptions and diagnosis regarding anatomical locations, such as confusion between right and left or medial and lateral positions. Other errors: Errors not falling into the above categories, including inaccuracies in report comparison dates, measurement units, punctuation, etc Please review each report provided below one by one to check for errors. Output your findings in the following format: “Serial number XXX /n Error categories: −1-, −2-, … /n Cause of error: …”
Few-shot prompt	I am going to provide you with six example reports: one error-free report and five reports containing the various categories of errors listed. The sole purpose of these example reports is to improve your comprehension and to help you recognize the various error categories mentioned in the subsequent tasks. We then entered each report and its corresponding error description and categorization in order. For example, one example report contained “Side confusion,” specifically, the radiology description showed a small, well-defined osseous density fragment in the left fifth finger, whereas the radiology impression described small osseous density adjacent to the right fifth middle phalanx, suspicious for an avulsion fracture. At the end of the learning phase containing all 6 example reports, we entered “provided example reports completed” to indicate the end of the examples. We then followed the prompts mentioned in the 0-shot setting to begin error detection on the test set reports.

Study Design

This study evaluated the capabilities of different LLMs and radiologists in detecting errors in emergency radiology reports across diverse scenarios through a 4-stage experimental design (Figure 1).

XSL•FO
RenderX

Figure 1. Workflow of the multistage validation study. Schematic diagram of the 4-stage experimental design. (**A**) Initial screening of 5 large language models using 200 reports (100 error-free; 100 with 127 synthetic errors across 5 categories). (**B**) Few-shot optimization of top-performing models (DeepSeek-R1) using clinical examples. (**C**) Benchmarking against varying experience of radiologists (seniors, attendings, and residents) in the 0 and few-shot settings on subdataset 1 and subdataset 2. (**D**) Real-world validation on 800 verified emergency reports. CT: computed tomography; DR: digital radiography; LLM: large language model; MRI: magnetic resonance imaging.

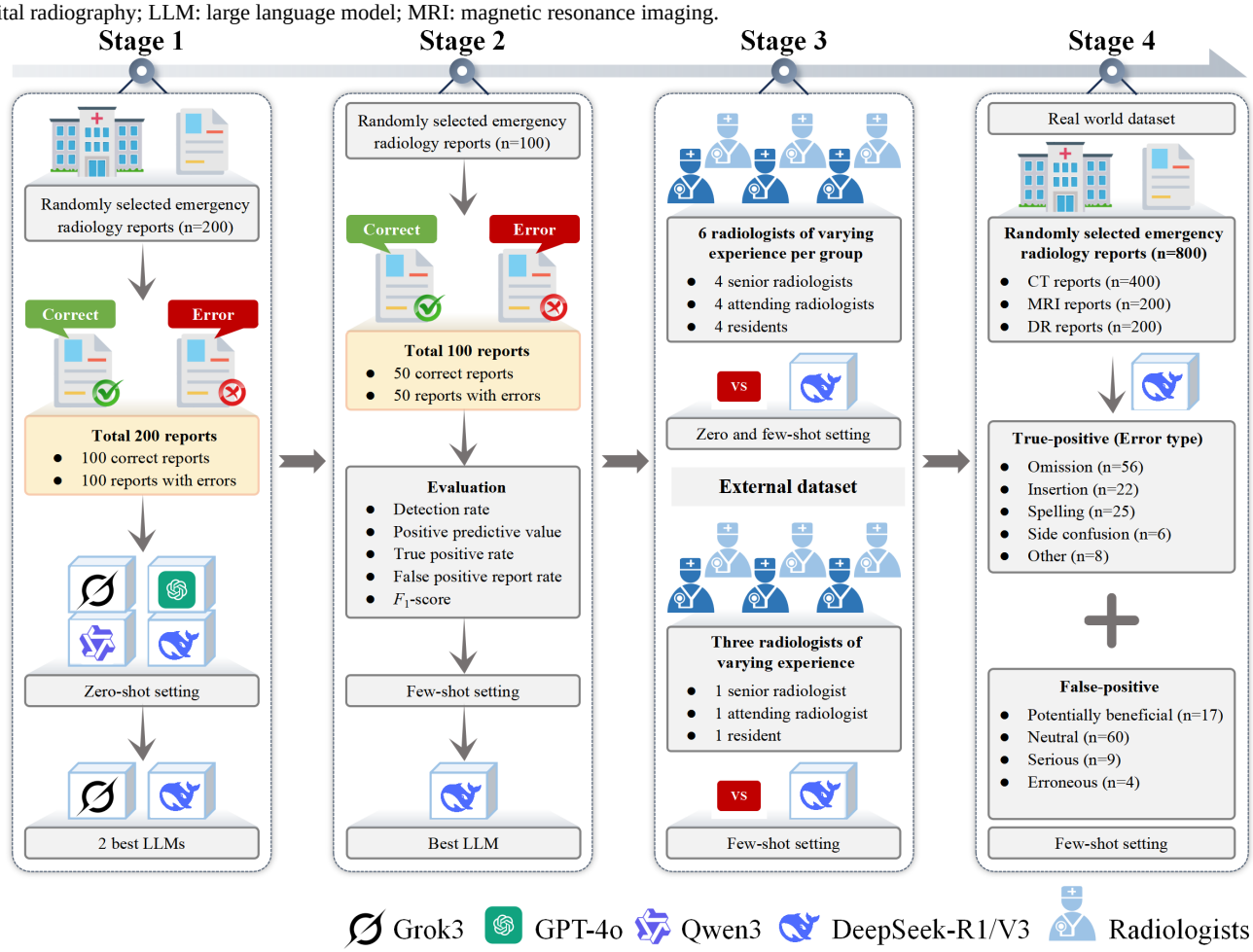

Stage 1: Initial Model Screening

First, 200 error-free reports were randomly selected from dataset 1 using a freeware data randomization tool [20], and then proofread by human experts to ensure accuracy. We evaluated 5 LLMs (open-source: DeepSeek-R1, DeepSeek-V3, and Qwen3-235B; closed-source: GPT-4o and Grok3) using 0-shot prompts to rank their performance in detecting errors in emergency radiology reports (200 cases), with the aim of determining the 2 best-performing models for subsequent analysis.

Stage 2: Few-Shot Prompt Evaluation

The 2 top-performing LLMs from stage 1 were re-evaluated in 0-shot and few-shot settings using a test set (100 reports and 64 errors), which was composed by randomly selecting 50 correct reports and 50 artificial-error subgroup reports. This phase explored whether the error detection capabilities of the optimal model improve in a few-shot setting and determined the best-performing model for stage 3 and stage 4 analysis.

Stage 3: Exploratory Analysis of DeepSeek-R1 Performance

Using the top-performance model (DeepSeek-R1) from stage 2 and using the same test set as stage 2 (100 reports and 64 errors), the error detection performance and time taken by 12 radiologists of different experience levels (4 senior radiologists, 4 attending radiologists, and 4 residents) were assessed via a customized online survey platform [24]. Experts were stratified into 2 groups (6 experts per group) and each group independently evaluated the 100 reports, benchmarking against the LLM's performance under simulated 0-shot and few-shot workflows. In addition, under the few-shot workflow, further testing was conducted on dataset 2 (100 reports containing 60 errors) to compare the error detection performance between DeepSeek-R1 and 3 radiologists with different experience levels (1 senior radiologist, 1 attending radiologist, and 1 resident). Stages 2 and 3 were intentionally designed as stress-testing phases for boundary performance, in which error prevalence was artificially enriched to enable controlled and statistically efficient comparison of error detection sensitivity across models and human readers.

Stage 4: Validation on a Real-World Dataset

Using the few-shot model from stage 2, DeepSeek-R1 processed 800 unverified real emergency radiology reports (400 CT, 200 MR, and 200 radiograph reports), with outputs compared to final reviewed versions. Reports were then assessed by 2 senior radiologists to determine if LLM outputs could influence final diagnoses and sorted as either true errors or false-positive

XSL•FO
RenderX

results. These assessments quantified how AI assistance might impact false-positive generation. The clinical impact of false-positive responses was further assessed to determine their actual harmfulness or potential benefits, with the detailed criteria outlined in Table S2 in Multimedia Appendix 1.

Statistical Analysis

All emergency report evaluation results were collected and analyzed using R (version 4.2.3; R Foundation) software. We measured the number of correctly identified errors, mean processing time, and associated costs. The average detection time and associated costs per report were calculated from evaluations of 20 randomly selected emergency radiology reports. Only artificially introduced errors during data processing were included in the measurements. To ensure the reliability of our findings, all reports in the dataset and their corresponding outputs were reviewed by 3 reviewers (HS, TW, and BZ). Detailed methods for error quantification and analytical procedures are provided in the text and Table S3 in Multimedia Appendix 1.

We compared the number of correctly identified errors and mean processing time between LLMs and radiologists using metrics such as positive predictive value (PPV), true positive rate (TPR), false positive report rate (FPRR), and F_1-score. We computed CIs using the Wilson score method [25]. To compare performance metrics between radiologists and LLMs across 0-shot and few-shot settings, Wald chi-square tests were used to evaluate differences in error detection metrics (PPV, TPR, and F_1-scores) and FPRR. Bonferroni correction adjusted for multiple comparisons, with a 2-tailed P value <.05 considered statistically significant. Cohen κ assessed agreement between model predictions and individual readers, while the intraclass correlation coefficient measured interrater reliability among radiologists. Interreader reliability was assessed using Cohen κ, with the following classification: 0.01 - 0.20 (negligible to slight), 0.21 - 0.40 (fair), 0.41 - 0.60 (moderate), 0.61 - 0.80 (substantial), and 0.81 - 1.00 (nearly perfect) [26].

Results

Stage 1: Performance in Detecting Errors Among LLMs

In stage 1 (200 reports and 127 errors), DeepSeek-R1 achieved the optimal performance, with a detection rate of 51.2%. For error detection, the PPV, TPR, and F_1-score were 64.4% (95% CI 55% - 74%), 51.2% (95% CI 42.3% - 60%), and 57% (95% CI 49% - 60%), respectively. Grok3 was the second-best performing model, with a detection rate of 48% (61/127). For error detection, the PPV, TPR, and F_1-score were 64.9% (95% CI 54.9% - 74.7%), 48% (95% CI 39.5% - 56.9%), and 55.2% (95% CI 47.1 - 62.8), respectively. In contrast, DeepSeek-V3 yielded a detection rate of 37% (47/127), with PPV and TPR of 54% (95% CI 43.4% - 64.4%) and 37% (95% CI 28.8% - 45.7%), respectively. Additionally, Qwen3 and GPT-4o achieved a detection rate of only 33.1% (42/127), with TPR of 33.1% (95% CI 24.8% - 41.2%), while with PPV of 48.3% (95% CI 37.3% - 58.8%) and 50% (95% CI 39.2% - 60.8%), F_1-score of 39.3% (95% CI 30.7% - 47.2%) and 39.8% (95% CI 31.1% - 47.8%), respectively (Table 3).

XSL•FO
RenderX

Table . Performance of different LLMs[a] in error detection in a 0-shot setting. Bonferroni correction was used to correct *P* values for multiple comparisons with DeepSeek-R1. Higher values of detection rate, PPV[b], and TPR[c] indicate better detection performance of the model, while a higher FPRR[d] value suggests poorer detection performance.

Model	Detection rate (%)		*P* value	PPV (%), 95% CI	*P* value	TPR (%), 95% CI	*P* value	F_1-score (%), 95% CI	*P* value	FPRR (%), 95% CI	*P* value
	Detection rate (%), 95% CI	Values, n/N									
DeepSeek-R1	51.2 (40.9-62.2)	65/127	—[e]	64.4 (55.0-74.0)	—	51.2 (42.3-60.0)	—	57.0 (49.0-64.5)	—[f]	16.5 (11.5-22.0)	—
Grok3	48.0 (37.8-58.3)	61/127	≥.99	64.9 (54.9-74.7)	≥.99	48.0 (39.5-56.9)	≥.99	55.2 (47.1-62.8)	≥.99	16.5 (11.5-22.0)	≥.99
DeepSeek-V3	37.0 (27.6-47.2)	47/127	.09	54.0 (43.4-64.4)	.35	37.0 (28.8-45.7)	.09	43.9 (35.3-52.2)	.19	20.0 (14.5-26.0)	≥.99
Qwen3	33.1 (24.4-42.5)	42/127	.01[f]	48.3 (37.3-58.8)	.052	33.1 (25.2-41.4)	.01[f]	39.3 (30.7-47.2)	.03[f]	22.5 (17.0-28.5)	.52
GPT-4o	33.1 (24.4-42.5)	42/127	.01[f]	50.0 (39.2-60.8)	.10	33.1 (24.8-41.2)	.01[f]	39.8 (31.1-47.8)	.03[f]	21.0 (15.5-27.0)	≥.99

[a]LLMs: large language models.

[b]PPV: positive predictive value.

[c]TPR: true positive rate.

[d]FPRR: false positive report rate.

[e]Not available.

[f]$P < 0.05$ was considered statistically significant.

Regarding the negative impact, the false positives generated by DeepSeek-R1 and Grok3 were comparable. For example, the FPRR for DeepSeek-R1 and grok3 both was 16.5% (95% CI 11.5% - 22%). In contrast, DeepSeek-V3, Qwen3, and GPT-4o generated significantly more false positives, with an FPRR of 20% (95% CI 14.5% - 26%), 22.5% (95% CI 17% - 28.5%), and 21% (95% CI 15.5% - 27%), respectively, with no statistically significant differences (*P*=.52 to ≥.99; Table 3).

Stages 2 and 3: Exploratory Analysis of DeepSeek-R1 Performance: Overall Performance in Detecting Errors on Subdataset 1 (100 Reports With 64 Errors)

In stage 2, we compared the performance of DeepSeek-R1 and Grok3 under both 0-shot and few-shot settings against radiologists with varying experience levels on dataset 1 (100 reports with 64 errors). Both models showed a higher error detection rate in the few-shot setting compared to the 0-shot setting. It should be noted that these detection rates reflect sensitivity under controlled, error-enriched conditions and are not intended to represent expected detection performance in routine clinical practice. Specifically, DeepSeek-R1's detection rate increased significantly from 60.9% (39/64) to 84.4% (54/64; *P*=.003), whereas Grok3's improvement was not statistically significant, rising from 53.1% (34/64) to 56.3% (36/64; *P*=.72). A significant performance difference was observed between DeepSeek-R1 and Grok3 in the few-shot setting (*P*<.001), but not in the 0-shot setting (*P*=.38).

In the 0-shot error detection task, DeepSeek-R1's detection rate (60.9%) was significantly higher than that of resident 2 (60.9% vs 35.9%, *P*<.05). However, there was no evidence of a difference in the percentage of detected errors between DeepSeek-R1 and the other radiologists per report (*P* value range, .15 to ≥.99). Under the few-shot learning setting, DeepSeek-R1 achieved a detection rate of 84.4% (54/64), surpassing the few-shot performance of resident radiologists. However, there was no evidence of a difference in the percentage of detected errors between DeepSeek-R1 and the other radiologists per report (*P* value range, .26 to ≥.99). Detailed performance metrics for LLMs and radiologists are provided in Table 4.

XSL•FO
RenderX

Table . Comparison of error detection between LLMs[a] and the radiologists in 0-shot and few-shot settings. Bonferroni correction was used to correct *P* values for multiple comparisons.[b]

Model	Detection rate (%)		*P* value	PPV[c] (%), 95% CI	*P* value	TPR[d] (%), 95% CI	*P* value	F_1-score (%), 95% CI	*P* value	FPRR[e] (%), 95% CI	*P* value
	Detection rate (%), 95% CI	Values, n/N									
0-shot											
DeepSeek-R1	60.9 (47.9-72.6)	39/64	Ref.[f]	90.7 (81.3-97.9)	Ref.	60.9 (49.2-72.6)	Ref.	72.9 (63.0-81.7)	Ref.	4.0 (1.0-8.0)	Ref.
Grok3	53.1 (40.3-65.6)	34/64	≥.99	79.1 (63.5-89.4)	.89	53.1 (40.7-65.0)	≥.99	63.6 (52.0-73.4)	.90	9.0 (4.0-15.0)	≥.99
Senior 1	73.4 (60.7-83.4)	47/64	.92	90.4 (81.5-97.8)	>.99	73.4 (60.7-83.4)	.90	81.0 (72.2-88.3)	≥.99	5.0 (1.0-10.0)	≥.99
Senior 2	71.9 (59.1-82.1)	46/64	≥.99	93.9 (86.3-100.0)	>.99	71.9 (60.3-82.8)	≥.99	81.4 (72.7-88.9)	.87	3.0 (0.0-7.0)	≥.99
Attending 1	68.8 (55.8-79.2)	44/64	≥.99	86.3 (75.9-94.8)	>.99	68.8 (56.9-79.7)	≥.99	76.5 (67.2-84.4)	≥.99	7.0 (3.0-13.0)	≥.99
Attending 2	65.6 (52.6-76.8)	42/64	≥.99	84.0 (73.5-93.6)	>.99	65.6 (53.5-76.8)	≥.99	73.7 (63.5-82.1)	≥.99	8.0 (3.0-13.0)	≥.99
Resident 1	40.6 (28.8-53.6)	26/64	.15	92.9 (81.8-100.0)	>.99	40.6 (29.0-53.1)	.13	56.5 (43.8-68.0)	.06	2.0 (0.0-5.0)	≥.99
Resident 2	35.9 (24.6-49.0)	23/64	.03[g]	85.2 (70.4-96.8)	>.99	35.9 (24.6-48.3)	.021[g]	50.5 (37.2-62.7)	<.001[g]	4.0 (1.0-8.0)	≥.99
Few-shot											
DeepSeek-R1	84.4 (72.7-91.9)	54/64	Ref.	91.5 (83.6-98.2)	Ref.	84.4 (75.0-92.6)	Ref.	87.8 (81.1-93.3)	Ref.	5.0 (1.0-10.0)	Ref.
Grok3	56.3 (43.3-68.4)	36/64	<.001[g]	82.9 (70.3-93.9)	≥.99	56.3 (42.2-66.7)	.001	66.0 (54.2-75.9)	<.001[g]	7.0 (2.0-12.0)	≥.99
Senior 3	93.8 (84.0-98.0)	60/64	.63	96.8 (91.5-100.0)	≥.99	93.8 (87.3-98.6)	.60	95.2 (91.1-98.5)	.30	2.0 (0.0-5.0)	≥.99
Senior 4	84.4 (72.7-91.9)	54/64	≥.99	93.1 (85.7-98.5)	≥.99	84.4 (75.4-93.0)	≥.99	88.5 (82.1-94.1)	≥.99	4.0 (1.0-8.0)	≥.99
Attending 3	70.3 (57.4-80.8)	45/64	.40	91.8 (83.3-98.1)	≥.99	70.3 (58.6-81.3)	.38	79.6 (70.7-87.2)	.65	4.0 (1.0-8.0)	≥.99
Attending 4	68.8 (55.8-79.4)	44/64	.26	91.7 (83.0-98.1)	≥.99	68.8 (57.6-79.7)	.27	78.6 (69.8-86.4)	.42	4.0 (1.0-8.0)	≥.99
Resident 3	53.1 (40.3-65.6)	34/64	<.001[g]	89.5 (78.4-97.6)	>.99	53.1 (40.9-65.5)	<.001[g]	66.7 (55.2-76.5)	.001[g]	4.0 (1.0-8.0)	≥.99

XSL•FO
RenderX

Model	Detection rate (%)		*P* value	PPV[c] (%), 95% CI	*P* value	TPR[d] (%), 95% CI	*P* value	F_1-score (%), 95% CI	*P* value	FPRR[e] (%), 95% CI	*P* value
	Detection rate (%), 95% CI	Values, n/N									
Resident 4	51.6 (38.8-64.1)	33/64	<.001[g]	86.8 (75.0-97.1)	>.99	51.6 (39.7-63.8)	<.001[g]	64.7 (53.3-74.8)	<.001[g]	5.0 (1.0-10.0)	≥.99

[a]LLM: large language model.

[b]The performance of DeepSeek-R1 in detecting errors was compared with that of Grok3 and radiologists using Wald chi-square tests. Higher values of detection rate, positive predictive value, and true positive rate indicate better detection performance of the model, while a higher false positive report rate value suggests poorer detection performance.

[c]PPV: positive predictive value.

[d]TPR: true positive rate.

[e]FPRR: false positive report rate.

[f]Ref.: Reference.

[g]$P < 0.05$ was considered statistically significant.

Performance in Detecting Errors by Error Type on Subdataset 1 (100 Reports With 64 Errors)

In the 0-shot setting, DeepSeek-R1 showed superior performance over 2 residents in detecting side confusion, with a detection rate of 94% (95% CI 71‐100) compared to 39% (95% CI 18‐64) and 28% (95% CI 11‐54; *P*=.003 and <.001, respectively; Figure 2A and Table 5). In addition, DeepSeek-R1 detected other types of errors more frequently than resident 1 (67%, 95% CI 35‐89 vs 8%, 95% CI 1‐40; *P*=.022). However, no significant differences were observed between DeepSeek-R1 and the other radiologists in the overall error detection rate per report (*P* values range from .091 to >.99).

XSL•FO
RenderX

Figure 2. Performance comparison for critical error types. Bar graphs comparing proofreading performance between human readers and large language models across error types. (**A**) Detection rates in a 0-shot setting. (**B**) Detection rates in a few-shot setting.

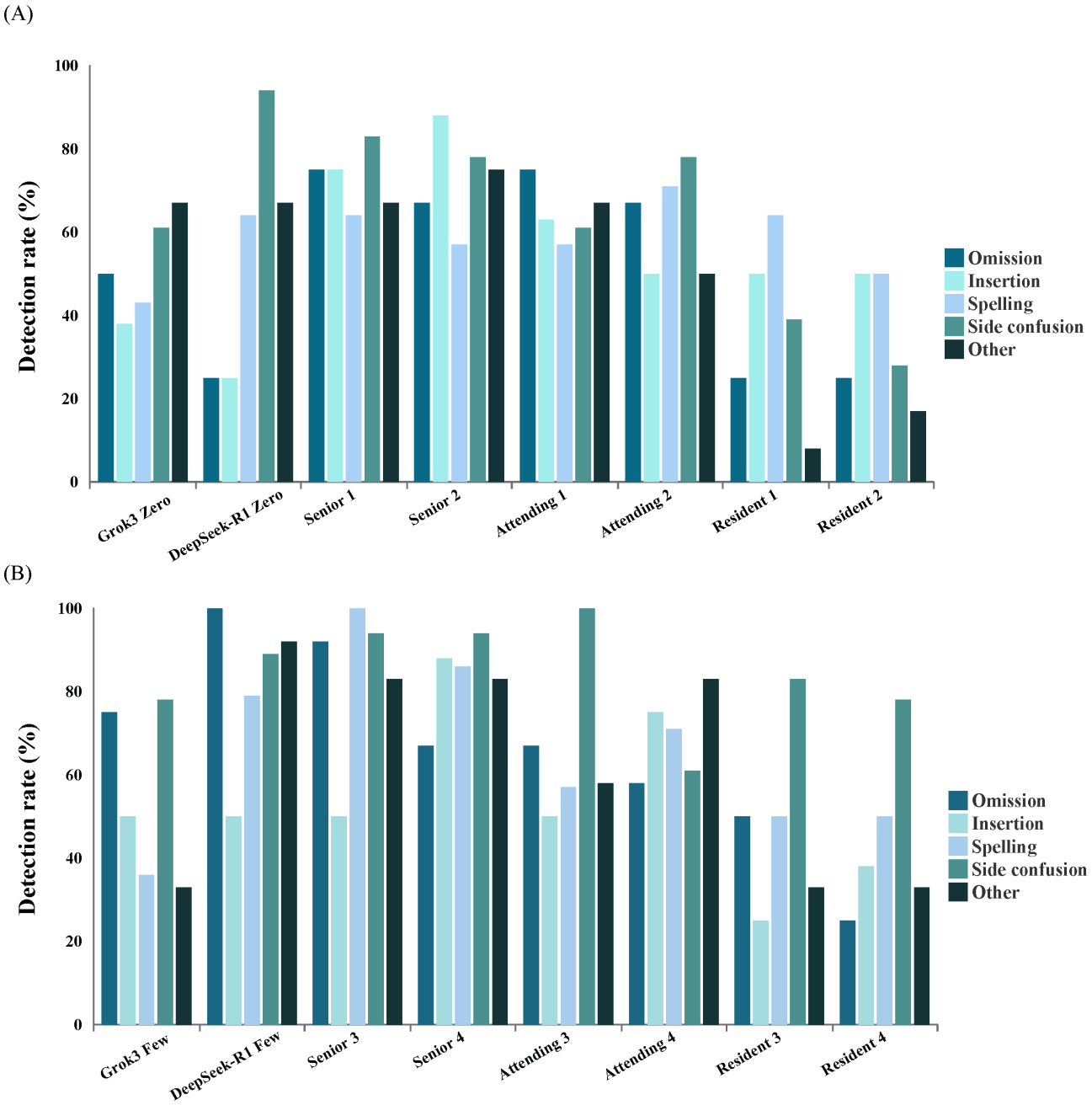

XSL•FO
RenderX

Table . Comparison of detection rates for different error types in radiology reports in 0-shot and few-shot settings. Other errors included errors that did not fit into the defined categories, such as incorrect date entries, errors in numbering of images and/or series, and mistakes in specifying units of measurement (eg, centimeter vs millimeter). Bonferroni correction was used to correct *P* values for multiple comparisons.[a]

	Omission		*P* value	Insertion		*P* value	Spelling		*P* value	Side confusion		*P* value	Other		*P* value
	Detection rate (%), 95% CI	Values, n/N		Detection rate (%), 95% CI	Values, n/N		Detection rate (%), 95% CI	Values, n/N		Detection rate (%), 95% CI	Values, n/N		Detection rate (%), 95% CI	Values, n/N	
0-shot															
DeepSeek-R1	25 (7-57)	3/12	Ref.[b]	25 (45-64)	2/8	Ref.	64 (36-86)	9/14	Ref.	94 (71-100)	17/18	Ref.	67 (35-89)	8/12	Ref.
Grok3	50 (22-77)	6/12	≥.99	38 (10-74)	3/8	≥.99	43 (19-70)	6/14	≥.99	61 (36-82)	11/18	.11	67 (35-89)	8/12	≥.99
Senior 1	75 (43-93)	9/12	.10	75 (36-96)	6/8	.32	64 (36-86)	9/14	≥.99	83 (58-96)	15/18	≥.99	67 (35-89)	8/12	≥.99
Senior 2	67 (35-89)	8/12	.28	88 (47-99)	7/8	.08	57 (30-81)	8/14	≥.99	78 (52-93)	14/18	≥.99	75 (43-93)	9/12	≥.99
Attending 1	75 (43-93)	9/12	.10	63 (26-90)	5/8	.91	57 (30-81)	8/14	≥.99	61 (36-82)	11/18	.11	67 (35-89)	8/12	≥.99
Attending 2	67 (35-89)	8/12	.28	50 (17-83)	4/8	≥.99	71 (42-90)	10/14	≥.99	78 (52-93)	14/18	≥.99	50 (22-78)	6/12	≥.99
Resident 1	42 (19-68)	5/12	≥.99	50 (17, 83)	4/8	≥.99	64 (36-86)	9/14	≥.99	39 (18-64)	7/18	.003[c]	8 (1-40)	1/12	.02[c]
Resident 2	42 (19-68)	5/12	≥.99	50 (17-83)	4/8	≥.99	50 (24-76)	7/14	≥.99	28 (11-54)	5/18	<.001[c]	17 (3-50)	2/12	.09
Few-shot															
DeepSeek-R1	100 (70-100)	12/12	Ref.	50 (22-77)	4/8	Ref.	79 (49-94)	11/14	Ref.	89 (64-98)	16/18	Ref.	92 (60-100)	11/12	Ref.
Grok3	75 (43-93)	9/12	.45	50 (22-77)	4/8	≥.99	36 (14-64)	5/14	.15	78 (52-93)	14/18	≥.99	33 (11-65)	4/12	.02[c]
Senior 3	92 (60-99)	11/12	≥.99	100 (68-100)	8/8	.15	100 (73-100)	14/14	.47	94 (71-100)	17/18	≥.99	83 (51-97)	10/12	≥.99
Senior 4	67 (35-89)	8/12	.20	88 (47-99)	7/8	.74	86 (56-97)	12/14	≥.99	94 (71-100)	17/18	≥.99	83 (51-97)	10/12	≥.99
Attending 3	67 (35-89)	8/12	.20	50 (17-83)	4/8	≥.99	57 (30-81)	8/14	≥.99	100 (78-100)	18/18	≥.99	58 (29-84)	7/12	.42
Attending 4	58 (29-84)	7/12	.08	75 (36-96)	6/8	≥.99	71 (42-90)	10/14	≥.99	61 (36-82)	11/18	.38	83 (51-97)	10/12	≥.99
Resident 3	50 (22-77)	6/12	.03[c]	25 (45-64)	2/8	≥.99	50 (24-76)	7/14	.80	83 (58-96)	15/18	≥.99	33 (11-65)	4/12	.02[c]
Resident 4	42 (19-68)	5/12	.01[c]	38 (10-74)	3/8	≥.99	50 (24-76)	7/14	.80	78 (52-93)	14/18	≥.99	33 (11-65)	4/12	.02[c]

[a]The number of errors correctly detected by DeepSeek-R1 was compared with that of Grok3 and radiologists by using Wald chi-square tests.

[b]Ref.: Reference.

[c]*P*<.05 was considered statistically significant.

XSL•FO
RenderX

In the few-shot setting, DeepSeek-R1 outperformed resident 3 in detecting both omission errors (detection rate, 100%, 95% CI 70 - 100 vs 50%, 95% CI 22-77 and 92%, 95% CI 60 - 100 vs 33%, 95% CI 11 - 65; *P*=.03 and *P*=.02, respectively) and resident 4 (100%, 95% CI 70 - 100 vs 25%, 95% CI 7 - 57 and 92%, 95% CI 60 - 100 vs 33%, 95% CI 11 - 65; *P*=.01 and *P*=.02, respectively). However, given the limited sample size, the subgroup analyses were likely underpowered, and the resulting comparisons should be viewed as exploratory and interpreted with caution. More results are displayed in Figure 2B and Table 5.

Performance in Detecting Errors by Imaging Modality on Subdataset 1 (100 Reports With 64 Errors)

In the 0-shot learning scenario, no significant difference was observed in error detection between DeepSeek-R1 and the radiologists for either radiography reports (detection rate: 66.7% vs 41.9% - 87.1%; *P* value range: .29 to >.99) or CT/MRI reports (45.5% vs 27.3% - 69.7%; *P* value range: .17 to >.99).

In the few-shot learning scenario, DeepSeek-R1 detected significantly more errors than the lowest-performing resident radiologist across both radiography and CT/MRI reports. For radiography reports, the detection rate was 90.3% (95% CI 73.1 - 97.4) compared to 54.8% (95% CI 36.3 - 72.2; *P*=.01). For CT/MRI reports, DeepSeek-R1 achieved a detection rate of 78.8% (95% CI 60.6 - 90.4) versus 45.4% (95% CI 28.5 - 63.4; *P*=.04). However, no significant differences were observed between DeepSeek-R1 and the other radiologists in error detection rates per report (*P* values ranging from .053 to >.99 for radiography and .07 to >.99 for CT/MRI). Additional results are provided in Table 6.

XSL•FO
RenderX

Table . Subgroup analyses of imaging modalities.

	Total		*P* value	Radiography		*P* value	CT[a] and MRI[b]		*P* value
	Detection rate (%), 95% CI	Values, n/N		Detection rate (%), 95% CI	Values, n/N		Detection rate (%), 95% CI	Values, n/N	
0-shot									
DeepSeek-R1	60.9 (47.9-72.6)	39/64	Ref.[c]	66.7 (45.4-80.2)	21/31	Ref.	54.5 (38.0-70.2)	18/33	Ref.
Grok3	53.1 (40.3-65.6)	34/64	>.99	64.5 (46.9-78.9)	20/31	>.99	42.4 (27.2-59.2)	14/33	>.99
Senior 1	73.4 (60.7-83.4)	47/64	.92	87.1 (69.2-95.8)	27/31	.48	60.6 (36.6-71.5)	20/33	>.99
Senior 2	71.9 (59.1-82.1)	46/64	>.99	74.2 (55.1-87.5)	23/31	>.99	69.7 (51.1-83.8)	23/33	>.99
Attending 1	68.8 (55.8-79.2)	44/64	>.99	77.4 (58.5-89.7)	24/31	>.99	60.6 (42.2-76.6)	20/33	>.99
Attending 2	65.6 (52.6-76.8)	42/64	>.99	74.2 (55.4-87.5)	23/31	>.99	57.6 (39.4-74.1)	19/33	>.99
Resident 1	40.6 (28.8-53.6)	26/64	.15	41.9 (52.1-60.7)	13/31	.29	33.3 (18.6-51.9)	11/33	.58
Resident 2	35.9 (24.6-49.0)	23/64	.03[d]	45.2 (27.8-63.7)	14/31	.51	27.3 (13.9-45.8)	9/33	.17
Few-shot									
DeepSeek-R1	84.4 (72.7-91.9)	54/64	Ref.	90.3 (73.1-97.4)	28/31	Ref.	78.8 (60.6-90.4)	26/33	Ref.
Grok3	56.3 (43.3-68.4)	36/64	.003[d]	66.7 (45.8-82.7)	21/31	.20	45.4 (28.5-63.4)	15/33	.04[d]
Senior 3	93.8 (84.0-98.0)	60/64	.63	93.5 (77.2-98.9)	29/31	>.99	93.9 (78.4-98.9)	31/33	.51
Senior 4	84.4 (72.7-91.9)	54/64	>.99	93.5 (77.2-99.0)	29/31	>.99	75.8 (57.4-88.3)	25/33	>.99
Attending 3	70.3 (57.4-80.8)	45/64	.40	77.4 (58.5-89.7)	24/31	>.99	63.6 (45.2-79.0)	21/33	>.99
Attending 4	68.8 (55.8-79.4)	44/64	.26	67.7 (48.5-82.7)	21/31	.20	69.7 (51.1-83.8)	23/33	>.99
Resident 3	53.1 (40.3-65.6)	34/64	<.001[d]	61.3 (42.3-77.6)	19/31	.053	45.4 (28.5-63.4)	15/33	.04[d]
Resident 4	51.6 (38.8-64.1)	33/64	<.001[d]	54.8 (36.3-72.2)	17/31	.01[d]	48.5 (31.2-66.1)	16/33	.07

[a]CT: computed tomography.

[b]MRI: magnetic resonance imaging.

[c]Ref.: Reference.

[d]*P*<.05 was considered statistically significant.

Performance in Detecting Errors on Dataset 2 (100 Reports With 60 Errors)

In the dataset 2 under the few-shot setting, DeepSeek-R1 achieved a detection rate of 95%, which was significantly higher than that of resident 5 (61.7%, *P*<.001) and attending radiologist 5 (71.7%, *P*=.002). These results suggest that DeepSeek-R1's performance is consistent across different institutions and report structures.

Stage 4: Validation on a Real-World Dataset

Among the 800 reports analyzed, Deepseek-R1 classified 207 reports as having errors, among which 117 were true errors, yielding a PPV of 0.565. The distribution of error types in the real-world dataset showed that omission was the most common error (n=56), whereas incorrect laterality was the least common (n=6). Furthermore, Deepseek-R1 successfully detected at least one instance of every error category.

XSL•FO
RenderX

The clinical impact of false-positive results incorrectly classified by Deepseek-R1 was further evaluated. There were no reports in which the model introduced errors based on content not present in the original report. Most serious false-positive results involved the addition of clinically insignificant observations to the impression section. Importantly, these outputs did not involve hallucinated or fabricated findings but instead reflected oversensitivity, such as suggestions to include clinically low-impact observations. This pattern of over-alerting, rather than content invention, is likely to preserve radiologist trust in the system. Moreover, approximately 19% of false-positive alerts were considered potentially beneficial for improving report clarity or completeness, thereby mitigating the negative psychological impact typically associated with false alarms.

In the analysis of false-positive results, 17 (18.9%) report changes were identified as potentially beneficial. The improvements were as follows: correcting typographical and grammatical errors (n=1); enhancing oversimplified impressions to improve clarity (n=6); elaborating on observations that were mentioned in the impressions but insufficiently detailed in the findings section, leading to more thorough documentation (n=3); and adding clinically relevant details, such as nerve root compressions, air-fluid levels, spinal stenosis, or new observations, that were documented in the findings but omitted from the impressions, thereby improving the completeness of the report (n=7).

Interrater Agreement

Figure 3A,B, Tables S4 and S5 in Multimedia Appendix 1 display the heatmap of Cohen κ coefficients illustrating pairwise agreement among readers in detecting errors in radiology reports under the 0-shot and few-shot settings. The analysis involved 12 radiologists (4 seniors, 4 attendings, and 4 residents) and 2 AI models (DeepSeek-R1 and Grok3). The interrater agreement analysis revealed several key patterns. Between the 2 AI systems in the 0-shot and few-shot settings, DeepSeek-R1 and Grok3 exhibited moderate and fair agreement (κ=0.45 and 0.29, respectively). Among human raters, agreement ranged from slight to substantial, with the lowest consistency observed between the 2 attendings (κ=0.02) under the 0-shot setting and the highest between attending 2 and resident 1 (κ=0.66) under the few-shot setting. In terms of human-AI agreement, DeepSeek-R1 showed fair to moderate consistency with radiologists (κ=0.42 - 0.57) under the 0-shot setting, while under the few-shot setting, agreement ranged from slight to moderate (κ=0.09 - 0.45). This pattern of variability in human interpretations, contrasted with the more consistent AI-human agreement, highlights both the subjective nature of error detection in radiology reports and the potential of AI systems to provide more standardized assessments. The generally low to moderate agreement among all raters (κ=0.02 - 0.66) underscores the complexity and inherent subjectivity in radiology report error detection, suggesting that even experienced clinicians may apply different criteria when evaluating report accuracy.

Figure 3. Interrater agreement analysis. Cohen κ coefficients for pairwise agreement among readers (large language models and radiologists). (A) Consistency for error types in the 0-shot setting. (B) Consistency for error types in the few-shot setting. Agreement scale: –1: complete inconsistency; 0: occasional agreement; 0.01-0.20: slight agreement; 0.21-0.40: fair agreement; 0.41-0.60: moderate agreement; 0.61-0.80: substantial agreement; and 0.81-1.00: almost perfect agreement.

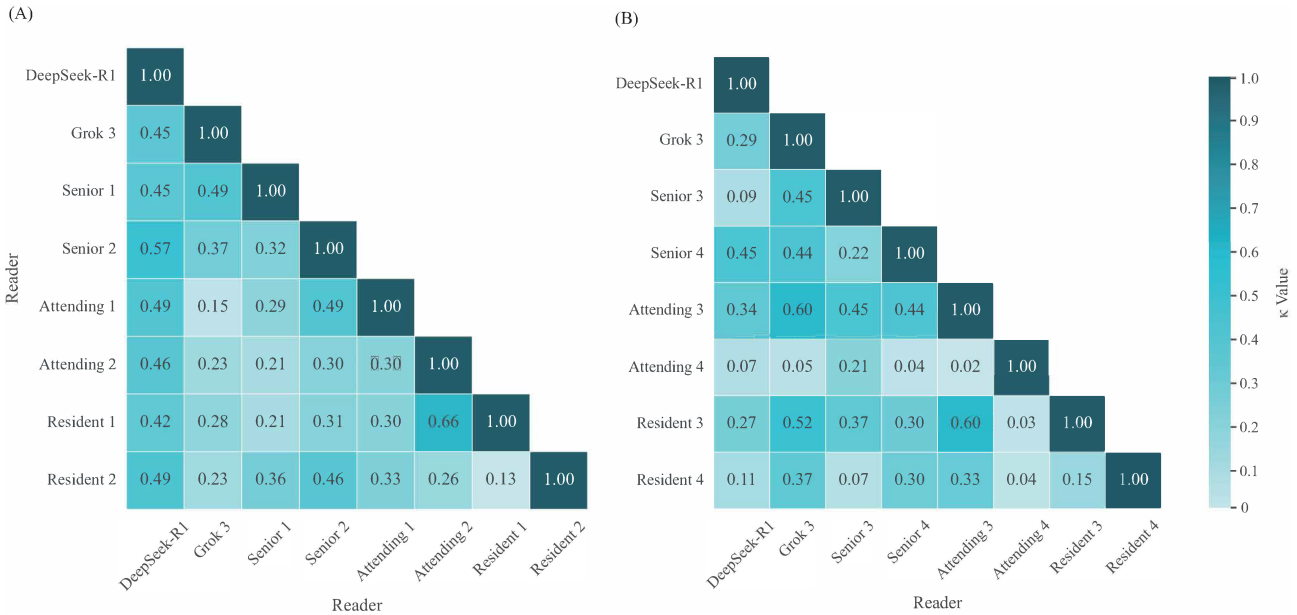

Reading Time

In the task of error detection for 100 reports, LLMs demonstrated significant time efficiency advantages. Grok3 processed 100 reports in 0.34 and 0.26 hours under 0-shot and few-shot settings, respectively. For the same task, DeepSeek-R1 required 2.56 hours in the 0-shot setting and 2.26 hours in the few-shot setting.

In contrast, radiologists' reading times ranged from 3.04 to 5.36 hours. For individual report reading, LLMs exhibited even more pronounced speed: Grok3 required only 12.24 seconds per report in the 0-shot setting, while DeepSeek-R1 needed 92.16 seconds.

XSL•FO
RenderX

The fastest and slowest radiologists required 109.44 seconds and 192.96 seconds per report, respectively. In the few-shot setting, LLMs showed a slight decrease in processing time and remained substantially faster than human experts (Figure S3 in Multimedia Appendix 1).

Discussion

Principal Findings

This study evaluated the performance of DeepSeek-R1 in detecting errors in Chinese emergency radiology reports. We compared its accuracy and processing efficiency with Grok3 and with radiologists of different experience levels under both 0-shot and few-shot settings. By introducing a real-world, emergency-focused evaluation framework, this work provides methodological insights into AI-assisted quality control in digital health care.

DeepSeek-R1 demonstrated strong error detection performance in emergency radiology. Under stress-testing conditions, it achieved a detection rate of 84.4% in the few-shot setting, indicating high sensitivity to predefined error types rather than direct real-world detection rates. Notably, its performance exceeded that of resident radiologists in identifying critical errors, supporting its potential clinical value for improving report accuracy and facilitating timely patient management.

Comparison With Previous Studies

Previous studies have shown that LLMs, such as GPT-4, can detect errors in radiology reports. Most of these studies relied on synthetically generated error datasets across imaging modalities, including radiography, CT, and MRI. Our study extends this literature in several important ways.

First, to our knowledge, this is the first validation of an LLM specifically in the emergency radiology setting. Emergency radiology is a high-risk environment characterized by time pressure, incomplete information, and urgent clinical decision-making, which differs substantially from routine radiologic workflows.

Second, we tailored the model to Chinese clinical language, particularly to address spelling and semantic ambiguities introduced by Pinyin-based input. This addresses a key limitation in prior cross-lingual applications and is especially relevant given reports of the limited accuracy of models such as Claude 3.5 in Chinese ultrasound reports. Third, by analyzing 7435 real-world clinical reports across CT, MRI, and radiography, our study avoided exclusive reliance on synthetic error injection [12,27]. This approach better captures the complexity and variability of real emergency reporting workflows.

Together, these findings suggest that DeepSeek-R1 may serve as an efficient and scalable assistant for radiologists, particularly for identifying critical omissions and communication-related errors. Importantly, validation using independent datasets from 2 additional tertiary hospitals with different reporting systems supports the robustness of the proposed approach across institutions and report templates.

Consistent with prior studies, our results indicate that LLMs can proofread radiology reports at a level comparable to most human readers. Reporting errors occur across all levels of radiologist experience [21,23], and the observed performance likely reflects routine clinical practice. These findings highlight the potential role of AI in supporting radiology workflows beyond image interpretation alone [28,29].

In terms of efficiency, our results align with previous studies examining the integration of LLMs into radiologic workflows [30,31]. When used as a proofreading tool, DeepSeek-R1 achieved performance comparable to human readers while requiring less time. Unlike human readers, whose performance may be affected by multitasking or fatigue during off-hours [32-34], DeepSeek-R1 provides stable and consistent output independent of such factors.

Limitations

This study has several limitations. First, part of the evaluation used predefined correct and incorrect report sets in which 127 errors were deliberately introduced. Although the error categories were derived from commonly reported patterns, this synthetic design cannot fully capture the diversity and contextual complexity of errors in routine clinical practice. Our real-world validation in stage 4 provides a reference for naturalistic error frequency. Among 800 consecutive emergency radiology reports, 117 true reporting errors were identified, corresponding to an overall error rate of 14.6%, with omission being the most common error type. This prevalence is lower than the error density used in the synthetic datasets in stages 2‐3. The higher synthetic error density was intentionally adopted to create a controlled and sufficiently challenging evaluation environment. This design allowed statistically efficient comparison of error detection performance across AI models and human readers within a limited sample size. Accordingly, performance metrics from stages 2‐3 should be interpreted as reflecting stress-tested detection capability rather than real-world error prevalence. Although real-world validation yielded a moderate PPV of 56.5%, this result should be interpreted in the context of workflow efficiency. In our study, AI-assisted report review was faster than human-only proofreading. When false-positive alerts can be dismissed with brief verification, the additional workload introduced by false positives remains limited, and overall time efficiency may be preserved. Moreover, PPV is inherently dependent on error prevalence and should not be extrapolated across settings with different baseline error rates. From a workflow perspective, AI assistance shifts the task from exhaustive manual proofreading to alert-based verification. Despite the presence of false positives, this approach reallocates radiologists' cognitive effort from error searching to decision confirmation. This shift may be particularly valuable in high-fatigue emergency settings.

Second, the exploratory analyses in stages 2‐3 were conducted in a simulated environment that differed from routine clinical practice. Factors such as artificially high error prevalence, lack of imaging data, long report length, and task instructions optimized for AI may have influenced reader performance. In addition, learning effects from repeated exposure to similar

XSL•FO
RenderX

error patterns may have biased results. These factors should be considered when interpreting human-AI comparisons.

Third, the error taxonomy and definitions of false-positive subtypes involve some subjectivity. Although granular categorizations were applied to reduce bias, certain subtypes may not fully reflect errors encountered in routine practice. Subgroup analyses of rare error categories were also limited by small sample sizes and should be considered exploratory.

Fourth, although this study assessed time efficiency, it did not include a formal cost-effectiveness analysis. Reduced processing time suggests potential resource savings, but costs related to infrastructure, system integration, and maintenance were not evaluated.

Fifth, the experimental setting may have introduced a Hawthorne effect [35], whereby radiologists' awareness of observation temporarily enhanced performance. This may have led to overestimation of human error detection rates. Importantly, this does not undermine the clinical relevance of our findings. DeepSeek-R1 consistently outperformed resident radiologists, even under conditions likely to favor human readers, suggesting that its error detection capability is robust and may be more pronounced in routine, nonobserved clinical settings.

Finally, generalizability warrants consideration. Although this study included data from 3 tertiary hospitals with different reporting systems and template styles, all institutions were within the same health care system. Validation using dataset 2 supports robustness across institutions within Chinese emergency radiology; however, further validation in other hospitals and nonemergency settings is needed, particularly in environments with lower baseline error prevalence. In addition, the strong performance of DeepSeek-R1 is closely related to its optimization for Chinese clinical language. Direct 0-shot transfer to other languages may therefore be inappropriate, and language-specific optimization or models with stronger multilingual medical pretraining will likely be required.

Overall, while our findings support the robustness of DeepSeek-R1 across institutions and report templates within Chinese emergency radiology, generalization to other clinical settings or languages should be approached cautiously. Further prospective validation and language-specific optimization are required. Additional domain-specific fine-tuning using emergency medicine–focused corpora may further improve performance in specialized settings such as trauma or neurologic imaging.

Conclusion

DeepSeek-R1 represents a meaningful advance in automated quality control for radiology reports, particularly in Chinese emergency settings. Its ability to identify clinically significant errors with high efficiency supports its role as an assistive proofreading tool in modern radiology workflows.

Acknowledgments

HS and TW contributed equally as co-first authors. SZ, DZ, and BZ served as corresponding authors.

No generative AI tools were used at any stage in the preparation of this manuscript.

Funding

We acknowledge financial support from the National Key Research and Development Program of China (2023YFF1204600), the National Natural Science Foundation of China (82227802, 82302190, U21A6005), the Clinical Frontier Technology Program of the First Affiliated Hospital of Jinan University (No. JNU1AF-CFTP-2022-a01201), the Science and Technology Projects in Guangzhou (202201020022, 2023A03J1036, 2023A03J1038, 2025A04J7006), the National Key R&D Program of China (2024YFA1012000, 2024YFC2417800), the Guangdong Basic and Applied Basic Research Foundation (2024A1515140146), the Outstanding Young Talents of Guangdong Special Support Program (Health Commission of Guangdong Province; 0720240213), the National Natural Science Foundation of China (62571228), the Beijing Natural Science Foundation (Z250002), and the Science and Technology Youth Talent Nurturing Program of Jinan University (21623209).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: HS, TW, DZ, BZ
Data curation: FW, JF, YL, XW, SL, LC, QR
Formal analysis: JS, YZ, XL, LW, GM
Funding acquisition: SZ, DZ, BZ
Investigation: XM, JX, JY
Methodology: ZJ, XW, WH, XH
Project administration: HS, TW
Visualization: HS, TW
Writing—original draft: HS, TW
Writing—review and editing: all authors

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1

Example for metric calculation. Table S1. Detailed prompts parameters of large language models used in this study; Table S2. Clinical impact and definition of false positive responses generated by DeepSeek-R1; Table S3. Example dataset for metric calculation; Table S4. Interrater agreement between LLMs and radiologists in a 0-shot setting; Table S5. Interrater agreement between LLMs and radiologists in a few-shot setting; Figure S1. The examples of various errors; Figure S2. Distribution of the 5 error types for both real and artificial errors (datasets 1 and 2); and Figure S3. Time efficiency analysis. Bar graphs comparing reading time for error detection for the large language models and radiologists.

[DOCX File, 7082 KB - jmir_v28i1e86841_app1.docx]

References

1. Selvarajan SK, Levin DC, Parker L. The increasing use of emergency department imaging in the United States: is it appropriate? AJR Am J Roentgenol 2019 Oct;213(4):W180-W184. [doi: 10.2214/AJR.19.21386] [Medline: 31237433]
2. Hanna TN, Lamoureux C, Krupinski EA, Weber S, Johnson JO. Effect of shift, schedule, and volume on interpretive accuracy: a retrospective analysis of 2.9 million radiologic examinations. Radiology 2018 Apr;287(1):205-212. [doi: 10.1148/radiol.2017170555] [Medline: 29156150]
3. Alexander R, Waite S, Bruno MA, et al. Mandating limits on workload, duty, and speed in radiology. Radiology 2022 Aug;304(2):274-282. [doi: 10.1148/radiol.212631] [Medline: 35699581]
4. Bruls RJM, Kwee RM. Workload for radiologists during on-call hours: dramatic increase in the past 15 years. Insights Imaging 2020 Nov 23;11(1):121. [doi: 10.1186/s13244-020-00925-z] [Medline: 33226490]
5. Kwee RM, Toxopeus R, Kwee TC. Imaging overuse in the emergency department: the view of radiologists and emergency physicians. Eur J Radiol 2024 Jul;176:111536. [doi: 10.1016/j.ejrad.2024.111536] [Medline: 38820950]
6. McDonald RJ, Schwartz KM, Eckel LJ, et al. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. Acad Radiol 2015 Sep;22(9):1191-1198. [doi: 10.1016/j.acra.2015.05.007] [Medline: 26210525]
7. Radiology facing a global shortage. RSNA News. 2022 May 10. URL: https://www.rsna.org/news/2022/may/global-radiologist-shortage [accessed 2026-03-17]
8. Geijer H, Geijer M. Added value of double reading in diagnostic radiology,a systematic review. Insights Imaging 2018 Jun;9(3):287-301. [doi: 10.1007/s13244-018-0599-0] [Medline: 29594850]
9. Claude 3.5 sonnet. Anthropic. 2024 Jun 21. URL: https://www.anthropic.com/news/claude-3-5-sonnet [accessed 2026-03-17]
10. Hello GPT-4o. OpenAI. 2024 May 13. URL: https://openai.com/index/hello-gpt-4o [accessed 2026-03-17]
11. Schmidt RA, Seah JCY, Cao K, Lim L, Lim W, Yeung J. Generative large language models for detection of speech recognition errors in radiology reports. Radiol Artif Intell 2024 Mar;6(2):e230205. [doi: 10.1148/ryai.230205] [Medline: 38265301]
12. Gertz RJ, Dratsch T, Bunck AC, et al. Potential of GPT-4 for detecting errors in radiology reports: implications for reporting accuracy. Radiology 2024 Apr;311(1):e232714. [doi: 10.1148/radiol.232714] [Medline: 38625012]
13. Kim S, Kim D, Shin HJ, et al. Large-scale validation of the feasibility of GPT-4 as a proofreading tool for head CT reports. Radiology 2025 Jan;314(1):e240701. [doi: 10.1148/radiol.240701] [Medline: 39873601]
14. Yan Y, Wang K, Feng B, et al. The use of large language models in detecting Chinese ultrasound report errors. NPJ Digit Med 2025 Jan 28;8(1):66. [doi: 10.1038/s41746-025-01468-7]
15. Cozzi A, Pinker K, Hidber A, et al. BI-RADS category assignments by GPT-3.5, GPT-4, and Google Bard: a multilanguage study. Radiology 2024 Apr;311(1):e232133. [doi: 10.1148/radiol.232133] [Medline: 38687216]
16. Lai V, Ngo N, Veyseh APB, et al. ChatGPT beyond english: towards a comprehensive evaluation of large language models in multilingual learning. arXiv. Preprint posted online on Apr 12, 2023. [doi: 10.48550/arXiv.2304.05613]
17. Gibney E. Scientists flock to DeepSeek: how they're using the blockbuster AI model. Nature 2025 Jan 29. [doi: 10.1038/d41586-025-00275-0]
18. Conroy G, Mallapaty S. How China created AI model DeepSeek and shocked the world. Nature 2025 Feb 13;638(8050):300-301. [doi: 10.1038/d41586-025-00259-0]
19. Guo D, Yang D, Zhang H, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature 2025 Sep;645(8081):633-638. [doi: 10.1038/s41586-025-09422-z] [Medline: 40962978]
20. Research randomizer. URL: https://www.randomizer.org/ [accessed 2026-03-17]
21. Quint LE, Quint DJ, Myles JD. Frequency and spectrum of errors in final radiology reports generated with automatic speech recognition technology. J Am Coll Radiol 2008 Dec;5(12):1196-1199. [doi: 10.1016/j.jacr.2008.07.005] [Medline: 19027683]
22. Ringler MD, Goss BC, Bartholmai BJ. Syntactic and semantic errors in radiology reports associated with speech recognition software. Health Informatics J 2017 Mar;23(1):3-13. [doi: 10.1177/1460458215613614] [Medline: 26635322]

XSL•FO
RenderX

23. Vosshenrich J, Nesic I, Cyriac J, Boll DT, Merkle EM, Heye T. Revealing the most common reporting errors through data mining of the report proofreading process. Eur Radiol 2021 Apr;31(4):2115-2125. [doi: 10.1007/s00330-020-07306-6] [Medline: 32997178]
24. Wenjuanxing - Online survey platform. URL: https://www.wjx.cn/ [accessed 2026-03-17]
25. Wallis S. Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. J Quant Linguist 2013 Aug;20(3):178-208. [doi: 10.1080/09296174.2013.799918]
26. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr;20(1):37-46. [doi: 10.1177/001316446002000104]
27. Salam B, Stüwe C, Nowak S, et al. Large language models for error detection in radiology reports: a comparative analysis between closed-source and privacy-compliant open-source models. Eur Radiol 2025 Aug;35(8):4549-4557. [doi: 10.1007/s00330-025-11438-y] [Medline: 39979623]
28. Lexa FJ, Jha S. Artificial intelligence for image interpretation: counterpoint-the radiologist's incremental foe. AJR Am J Roentgenol 2021 Sep;217(3):558-559. [doi: 10.2214/AJR.21.25484] [Medline: 33533639]
29. Coppola F, Faggioni L, Gabelloni M, et al. Human, all too human? An all-around appraisal of the "artificial intelligence revolution" in medical imaging. Front Psychol 2021;12:710982. [doi: 10.3389/fpsyg.2021.710982] [Medline: 34650476]
30. Rau A, Rau S, Zoeller D, et al. A context-based chatbot surpasses trained radiologists and generic ChatGPT in following the ACR appropriateness guidelines. Radiology 2023 Jul;308(1):e230970. [doi: 10.1148/radiol.230970] [Medline: 37489981]
31. Russe MF, Fink A, Ngo H, et al. Performance of ChatGPT, human radiologists, and context-aware ChatGPT in identifying AO codes from radiology reports. Sci Rep 2023 Aug 30;13(1):14215. [doi: 10.1038/s41598-023-41512-8] [Medline: 37648742]
32. Hanna TN, Zygmont ME, Peterson R, et al. The effects of fatigue from overnight shifts on radiology search patterns and diagnostic performance. J Am Coll Radiol 2018 Dec;15(12):1709-1716. [doi: 10.1016/j.jacr.2017.12.019] [Medline: 29366599]
33. Krupinski EA, Berbaum KS, Caldwell RT, Schartz KM, Kim J. Long radiology workdays reduce detection and accommodation accuracy. J Am Coll Radiol 2010 Sep;7(9):698-704. [doi: 10.1016/j.jacr.2010.03.004] [Medline: 20816631]
34. Sistrom CL, Slater RM, Rajderkar DA, Grajo JR, Rees JH, Mancuso AA. Full resolution simulation for evaluation of critical care imaging interpretation; part 1: fixed effects identify influences of exam, specialty, fatigue, and training on resident performance. Acad Radiol 2020 Jul;27(7):1006-1015. [doi: 10.1016/j.acra.2019.11.023]
35. Kidwai AS, Abujudeh HH. Radiologist productivity increases with real-time monitoring: the Hawthorne effect. J Am Coll Radiol 2015 Nov;12(11):1151-1154. [doi: 10.1016/j.jacr.2015.03.041] [Medline: 26208406]

Abbreviations

AI: artificial intelligence
CT: computed tomography
FPRR: false positive report rate
LLM: large language model
MRI: magnetic resonance imaging
PPV: positive predictive value
TPR: true positive rate

Edited by Q Jin; submitted 31.Oct.2025; peer-reviewed by A Okolie, O Akinpeloye, S Jiang; revised version received 05.Feb.2026; accepted 11.Feb.2026; published 14.Apr.2026.

Please cite as:
Shen H, Wu T, Wang F, Fang J, Li Y, Wu X, Liu S, Chen L, Ren Q, Meng X, Xu J, Sun J, Zhao Y, Liu X, Wang L, Mai G, You J, Jin Z, Wu X, He W, Han X, Zhang S, Zeng D, Zhang B
Error Detection in Emergency Radiology Reports Using a Large Language Model: Multistage Evaluation Study
J Med Internet Res 2026;28:e86841
URL: https://www.jmir.org/2026/1/e86841
doi:10.2196/86841

© Hui Shen, Tianyang Wu, Fei Wang, Jin Fang, Yuange Li, Xiaoling Wu, Shuyi Liu, Liting Chen, Qiuping Ren, Xiao Meng, Jiatong Xu, Jie Sun, Yujie Zhao, Xin Liu, Liaoyuan Wang, Guipeng Mai, Jingjing You, Zhe Jin, Xuewei Wu, Wenle He, Xue Han, Shuixing Zhang, Dong Zeng, Bin Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any

XSL•FO
RenderX

medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Improving Retrieval Augmented Generation for Health Care by Fine-Tuning Clinical Embedding Models: Development and Evaluation Study

Kamyar Arzideh[1,2], MSc; Henning Schäfer[1,3], PhD; Ahmad Idrissi-Yaghir[1,4], MSc; Cynthia Sabrina Schmidt[1,3,5], MD; Bahadir Eryilmaz[1], MSc; Mikel Bahn[1], MSc; Amin T Turki[1,6], MD, PhD; Olivia Barbara Pollok[4], MD; Eva Maria Hartmann[1,2], MSc; Philipp Winnekens[2], MSc; Katarzyna Borys[1,4], MSc; Johannes Haubold[1,4], Prof Dr Med; Felix Nensa[1,4], Prof Dr Med; René Hosch[1,4], PhD

[1]Institute for Artificial Intelligence in Medicine,, University Hospital Essen, Essen, Germany

[2]Central IT Department, Data Integration Center, University Hospital Essen, Girardetstr. 2, Essen, Germany

[3]Institute for Transfusion Medicine, University Hospital Essen, Essen, Germany

[4]Institute of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Essen, Germany

[5]Center of Sleep and Telemedicine, University Hospital Essen, Ruhrlandklinik, Essen, Germany

[6]Department of Hematology and Oncology, Ruhr-University Bochum, Marienhospital University Hospital, Bochum, Germany

Corresponding Author:
Kamyar Arzideh, MSc
Institute for Artificial Intelligence in Medicine,, University Hospital Essen, Essen, Germany

Abstract

Background: Embedding models are critical components of Retrieval Augmented Generation (RAG) systems for retrieving and searching unstructured medical data. However, existing models are predominantly trained on publicly available English datasets, limiting their effectiveness in non-English health care settings. More importantly, these models lack training on real-world clinical documents, leading to inaccurate context retrieval when integrated into RAG systems for health care applications. This gap is particularly pronounced in specialized medical documentation containing domain-specific terminology, abbreviations, and nuanced clinical language.

Objective: This retrospective study aimed to develop and validate embedding models specifically trained on real-world clinical documents from multiple medical specialties to improve medical information retrieval (IR) and RAG system performance in both German and English language contexts.

Methods: We fine-tuned embedding models, so-called sentence transformers, using the multilingual-e5-large architecture as a foundation. Training data consisted of approximately 11 million question-answer pairs synthetically generated from 400,000 diverse clinical documents from a large German tertiary hospital, spanning 163,840 patients and 282,728 clinical cases between 2018 and 2023. The *SauerkrautLM-SOLAR-Instruct* large language model generated medically relevant questions and corresponding answers for each document. The dataset was additionally pseudonymized and translated into English to aim for broader applicability. Models were evaluated in 2 distinct scenarios: IR using questions with multiple relevant passages, and RAG system performance in both cross-patient and patient-centered contexts.

Results: In the IR evaluation, the fine-tuned miracle model achieved a mAP@100 of 0.27, outperforming the multilingual-e5-large baseline (0.14) and state-of-the-art models such as bge-m3 (0.11). In the RAG evaluation, the model demonstrated robust performance comparable with the baseline in the constrained patient-centered scenario (BERTScore F1 0.781 vs 0.778) and showed moderate improvements in the unconstrained cross-patient setting (BLEURT 0.56 vs 0.53). Notably, the model trained on pseudonymized data achieved comparable retrieval performance (mAP@100 0.25) and the highest scores for patient-centered contextual precision (0.93). Performance gains were robust in the German dataset, while the translated English model demonstrated promising results as a proof of concept for cross-lingual transfer.

Conclusions: By leveraging a comprehensive real-world dataset spanning multiple medical specialties and using large language models for synthetic question generation, we successfully created and validated domain-specific embedding models. These models can improve medical IR in large-scale search spaces and perform competitively in constrained RAG applications. By publishing the models trained on pseudonymized data, other health care institutions can integrate or adapt these embedding models to their needs. This work establishes a reproducible framework for developing domain-specific clinical embedding models, with the potential to improve data retrieval in medical settings.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e82997) doi:10.2196/82997

KEYWORDS

Retrieval Augmented Generation; RAG; natural language processing; information retrieval; large language models; LLM; NLP

Introduction

Integrating natural language processing (NLP) into health care has opened new application areas to improve clinical decision-making and patient care [1-4]. Central to this advancement was the development of language and domain-specific language models [5-9], which are important for accurately accessing and understanding medical documents. In the field of medicine, however, several obstacles have so far limited the applicability of these new tools. First, medical language [10,11] is complex and documentation characterized by specialized terminology, jargon, and abbreviations [12,13], representing a significant challenge for language models in the health care domain. To address this challenge, it is essential to design models that are specifically tailored to the health care sector. Hence, NLP models, mainly developed and trained on publicly available English datasets, fall short in understanding and processing natural language in non-English linguistic contexts [14,15]. Second, medical documents are characterized by a high degree of specificity and contextual nuances, which the existing general language models often fail to capture [16-18]. Consequently, the absence of domain and language-specific models has proven to be a substantial obstacle [19-21] for an effective implementation of artificial intelligence and NLP technologies.

Embedding models learn to represent words or phrases as mathematical vectors, capturing their meaning based on the context in which they appear, making them especially valuable in fields where terms can have different meanings depending on context. Integration of these models into Retrieval Augmented Generation (RAG) systems marks an advancement in information retrieval (IR). RAG may enhance the performance of large language models (LLMs) by combining them with an existing knowledge base or document repository [22]. Typically, RAG is implemented by transforming documents or records into vectors. In response to a specific query, the most similar documents are retrieved and then incorporated into the LLM prompt. They combine retrieval-based and generative NLP models and are heavily reliant on the quality of underlying embedding models for their effectiveness [23]. However, these embedding models, also known as sentence transformers, are primarily trained on publicly available datasets. Using these models in a hospital context with routine documents is potentially difficult as these models have not been trained on real-world unstructured clinical texts. Integrating these models into RAG systems to search through patient documents could lead to undesired or incorrect results. This affects the overall adaptation of RAG systems in clinical practice and highlights the importance of domain fine-tuning.

The development of domain-specific embedding models has the potential to improve RAG systems in the clinical context. By providing more accurate, reliable, and context-aware interpretations of medical texts, such models may improve the information basis for medical decision-making and, via decision support systems, ultimately improve patient outcomes. Given the relevant limitations of existing models, the objective of this study was to develop medical embedding models, exemplarily both in German and English. These models were trained on a large corpus of real-world clinical documents of a large tertiary hospital. This domain-specific training should enhance the performance of RAG systems for patient-related questions. By adapting the models to the medical field, we aimed to enhance their capability to understand hospital jargon and interpret the complex semantic relationships and clinical contexts in medical documentation.

Methods

Ethical Considerations

This study was approved by the ethics committee of the Medical Faculty of the University of Duisburg-Essen (approval number 23‑11557-BO). Due to the study's retrospective nature, the requirement of written informed consent was waived by the ethics committee of the Medical Faculty of the University of Duisburg-Essen. All methods were carried out in accordance with relevant guidelines and regulations and the Declaration of Helsinki. No identifying information or images of individual participants are included in this manuscript or its supplementary materials. Participants did not receive any compensation.

Data Definition

Our study used a carefully curated dataset comprising a balanced collection of German language clinical notes documented at the University Hospital Essen. The dataset included 100,000 documents from each of the following 4 medical documentation categories: radiology reports, discharge letters, pathology reports, and surgical operation notes. This diverse composition ensures a comprehensive representation of clinical language and terminology. The clinical notes were documented between 2018 and 2023 in the clinical information system of the investigating site and contain information related to the medical history of patients, including diagnostic data and treatment documentation.

The total dataset consisted of 400,000 clinical documents from 163,840 patients and 282,728 clinical cases, resulting in a total number of 163,479,394 tokens, of which 1,541,389 were unique, reflecting the depth and complexity of the vocabulary present in the dataset. Further statistical details, including average document length, token frequency distribution, and other relevant metrics, are summarized for each document type in Table 1.

XSL•FO
RenderX

Table . Dataset characteristics[a].

Dataset	Total number of tokens	Total number of unique tokens	Total number of sentences	Average tokens per document	Average sentence length
Discharge letters	86,512,538	901,771	4,406,969	865.13	19.63
Pathology reports	22,313,259	333,028	1,901,434	223.13	11.73
Radiology reports	15,565,379	307,141	1,202,033	155.65	12.95
Operation notes	39,088,218	467,116	3,061,313	390.88	12.77
Total	163,479,394	1,541,389	10,571,749	408.70	15.46

[a]Displayed are quantitative measures for length and average length across the datasets. “Total number of tokens” was calculated by counting the number of tokens or words in the documents, including punctuation. The lemma form of each token was created and the number of lemmas was used as an approximation for the “total number of unique tokens.” “Total number of sentences” represents the count of all sentences per document. “Average tokens per document” was determined by dividing the total number of tokens by the total number of documents. “Average sentence length” was calculated by dividing the number of tokens by the total number of sentences.

Dataset Generation

The dataset containing 400,000 documents was segmented into smaller chunks using a recursive text splitting approach implemented in the Python (Python Software Foundation) package *langchain* with a maximum size of 450 characters and an overlap of 80 characters. This overlap was selected to ensure semantic continuity at text segment boundaries, accepting a minor degree of redundancy to prevent loss of clinical context at split points. The maximum character length was informed by the specific structural characteristics of the documents. Clinical information is typically recorded in concise, information-dense paragraphs or itemized lists rather than long-form narrative prose. This splitter uses a hierarchical approach to preserve semantic boundaries by attempting to split text at natural break points in the following priority order: paragraph breaks (\n\n), line breaks (\n), word boundaries (spaces), and finally individual characters. This methodology ensures that complete sentences and paragraphs are preserved whenever possible, only breaking midsentence when individual sentences exceed the 450-character limit. The chunking process resulted in 3,164,164 chunks across all document types.

These chunks were given as input to the *SauerkrautLM-SOLAR-Instruct* model, which has 10.7B parameters and was trained on German translated data. For each chunk, this LLM was tasked with generating 5 medically relevant questions alongside the correct answer that is contained in the document chunk, with the intention of simulating a real-world clinical questioning scenario. The model was deployed within the secured clinical network in order to ensure data protection for patients and physicians.

This process was applied across all collected documents—radiology reports, discharge letters, pathology reports, and operation notes. As a result, a dataset comprising 15,820,820 medical question-answer pairs was created as a training basis for the LLM. This method was chosen to capture the essence of clinical language and reflect the practical, inquiry-driven nature of medical diagnostics and patient care. To reduce the occurrence of hallucinations associated with the LLM and to enhance the data quality, the question-and-answer pairs were systematically filtered through the application of defined criteria. All questions were required to conclude with a question mark and to start with a number between 1 and 5, ensuring the generation of exactly 5 questions for each chunk. After this process, 11,000,227 questions-answer pairs were used for further training and evaluation steps. An overview of the methodologies including sample questions and answers is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Overview of training and evaluation methods**.** For each text chunk, 5 question-answer pairs were generated with a large language model. The question and chunk pairs were then used to fine-tune sentence transformer models. Different evaluation scenarios were set up to measure the performance of the models. This graphic was created in BioRender.

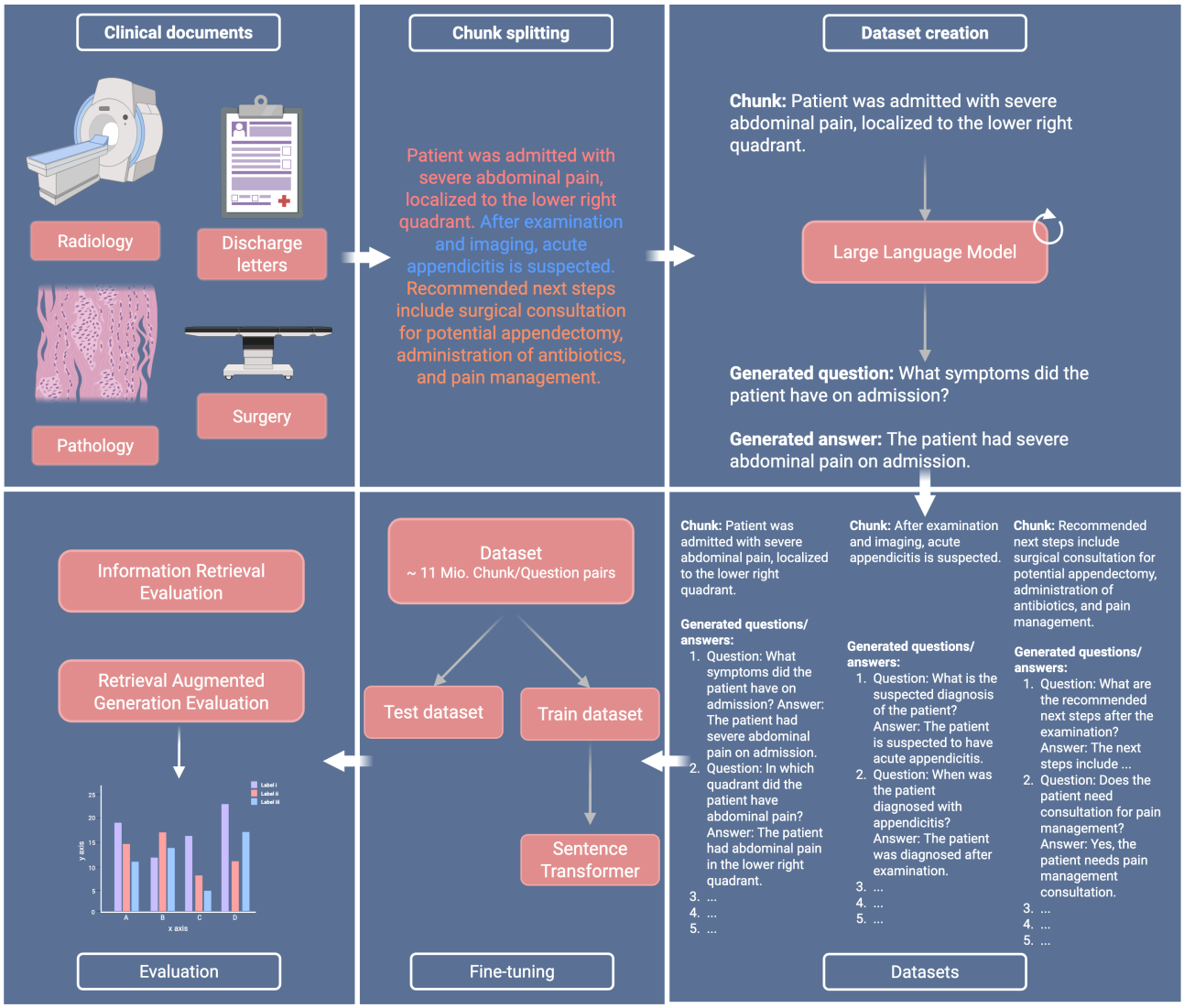

Dataset Pseudonymization

As the generated dataset comprises personally identifiable data belonging to patients and health care personnel, it underwent a pseudonymization process using a deidentification pipeline developed at the investigating site [24]. Protected Health Information in the documents was identified and replaced with surrogates. All training and evaluation procedures were additionally conducted on pseudonymized documents, and performance was compared with the model fine-tuned on real-world data.

Dataset Translation

The language used in the clinical documents is German, which limits the applicability of the models to German-speaking hospitals. To overcome this limitation, the dataset was translated into English using the Fairseq WMT'19 German-to-English translation model [25]. Although translating medical content can be complex and may result in inaccuracies due to the specialized terminologies involved, the translation process serves as a method for augmenting the corpus and increasing its accessibility. The decision to use this particular translation model was informed by a head-to-head translation model comparison experiment, including physician's evaluation. The physician was presented with 30 samples produced by various translation models, including WMT19-de-en, M2M-100 [26], NLLB [27], T5 [28], MBart-50 [29], and OPUS-MT-de-en [30]. The feedback on the medical accuracy and readability of the translations informed the final selection of the WMT'19 model for this task.

Model Training

Fine-tuning was performed using pairs of real document passages and the generated questions described in the "Dataset Generation" subsection. The data were split 80%/20% into a training and a test dataset on a patient basis so that a patient's data were part of either the training or the test dataset. This resulted in 8,801,525 chunk-and-question pairs for training and 2,198,702 pairs for testing. Hyperparameters used during training include an initial learning rate of 2e-5 and a batch size of 1024. For learning rate optimization, the AdamW optimizer was used. As a loss function, the *CachedMultipleNegativesRankingLoss* from the sentence transformer library [31] with a minibatch size of 32 was used.

XSL•FO
RenderX

To normalize the embedding size, pooling was used to create fixed-sized sentence embeddings of 1024 dimensions. Each training run was performed on 1 NVIDIA H100 80GB graphical processing unit for approximately 8 days.

Training was limited to a single epoch. Given the size of the dataset and the synthetic nature of the training data, this approach was selected to prevent overfitting to specific linguistic patterns generated by the LLM. Additionally, training runs showed that training loss plateaued after 1 training epoch.

Evaluation

Overview

Different techniques were used to measure the performance of the fine-tuned models. The models were evaluated on different datasets across 2 distinct evaluation settings—IR and RAG. The reason for these different evaluation scenarios is the difficulty of measuring the performance of sentence transformer models in a single setup. As there is no single scoring scheme that can give an estimate of final performance, different metrics were used, each giving an indication of performance for each individual task. More detailed descriptions about the exhibited metrics can be found in Multimedia Appendices 1 and 2.

Evaluation Setups

IR evaluation was performed to consider various passages and should give a statement about the model's ability to find several suitable passages for a given query. The RAG evaluation results should indicate whether the embedding model also improved the response quality in an RAG system and, therefore, benefits context retrieval in clinical RAG scenarios. Together, these 2 setups provide a comprehensive view of the fine-tuned models' performance compared with non–fine-tuned models. Across all evaluation scenarios, we assessed whether fine-tuning led to measurable improvements on each specific task.

For each evaluation scenario, the models were evaluated on the German or English dataset. Thus, the datasets for each evaluation scenario were also translated from German into English. For the evaluation, the fine-tuned model is referred to as *miracle*, the model fine-tuned on pseudonymized data is referred to as *miracle-pseudonymized*, and the model fine-tuned on pseudonymized and English data is referred to as *miracle-pseudonymized-translated*.

Evaluated Models

For comparison of fine-tuned to existing models, 3 publicly available models were evaluated. The first model was the *multilingual-e5-large* sentence transformer [32], which was used as a base model for further fine-tuning. This model was continually trained on a mixture of multilingual datasets and supports up to 100 languages, including German. The second model was *bge-m5* [33], a state-of-the-art embedding model that is able to process inputs of sizes, spanning from short sentences to long documents of up to 8192 tokens. It was also trained on multilingual datasets. The third evaluated model was *gte-multilingual*-base [34] with support for more than 70 languages text lengths up to 8192 tokens. The fourth evaluated model was *German-RAG-BGE-M3*, which is a fine-tuned variant of the *bge-m3* model explicitly trained on German data for document retrieval. All evaluated models, including parameter sizes and links to the model cards on the huggingface hub, are listed in Multimedia Appendix 3.

IR Evaluation Setup

The IR evaluation was set up by identifying duplicate questions in the test dataset. These duplicate questions are of particular interest as they allow for 1 question to have multiple document passages that have a positive relation to this question. This can be used for an IR task such as search engines, where typically 1 query has multiple possible search results. In total, 109,795 unique questions and 173,933 document chunks from 52,434 documents and 26,358 patients were used for IR evaluation. On average, each question has approximately 5 document chunks with a positive relation.

The quality of the returned results or passages can be measured by metrics such as *Accuracy@K*, Precision@*K*, and *Recall@K* [35]. For IR systems such as search engines, the problem can be defined as a 2-fold task involving the definition of a search query and multiple results for this query. Using a cutoff threshold K and calculating an evaluation metric within the elements of this threshold can determine the quality of the results. More details about IR metrics such as mean reciprocal rank, normalized discounted cumulative gain (NDCG), and mean average precision (mAP) are provided in Multimedia Appendix 4.

For the IR evaluation, *Precision@K*, *Recall@K*, *MRR@K*, *NDCG@K*, and *mP@K* were calculated by retrieving top-k most similar documents for the given query in the dataset. The retrieved list was then compared to the ground truth. Similar document passages were retrieved using the cosine similarity between a query and top-k passages. For the evaluation, the results of the metrics are displayed for a K value of 10. For mAP, the results for a K value of 100 are shown.

RAG Evaluation Setup

In order to evaluate the benefit of our fine-tuned clinical embedding model in an IR setting, an experiment was conducted by setting up RAG systems. In Figure 2, the RAG evaluation setup is visualized.

XSL•FO
RenderX

Figure 2. Visualization of Retrieval Augmented Generation (RAG) evaluation setup. For the RAG evaluation, different databases containing embeddings from the different models were initialized. Different RAG systems using these databases as retrieval options were set up, and the same questions were asked of these systems. The generated answers from the RAG systems were compared against the ground truth answer. This graphic was created in BioRender. LLM: large language model.

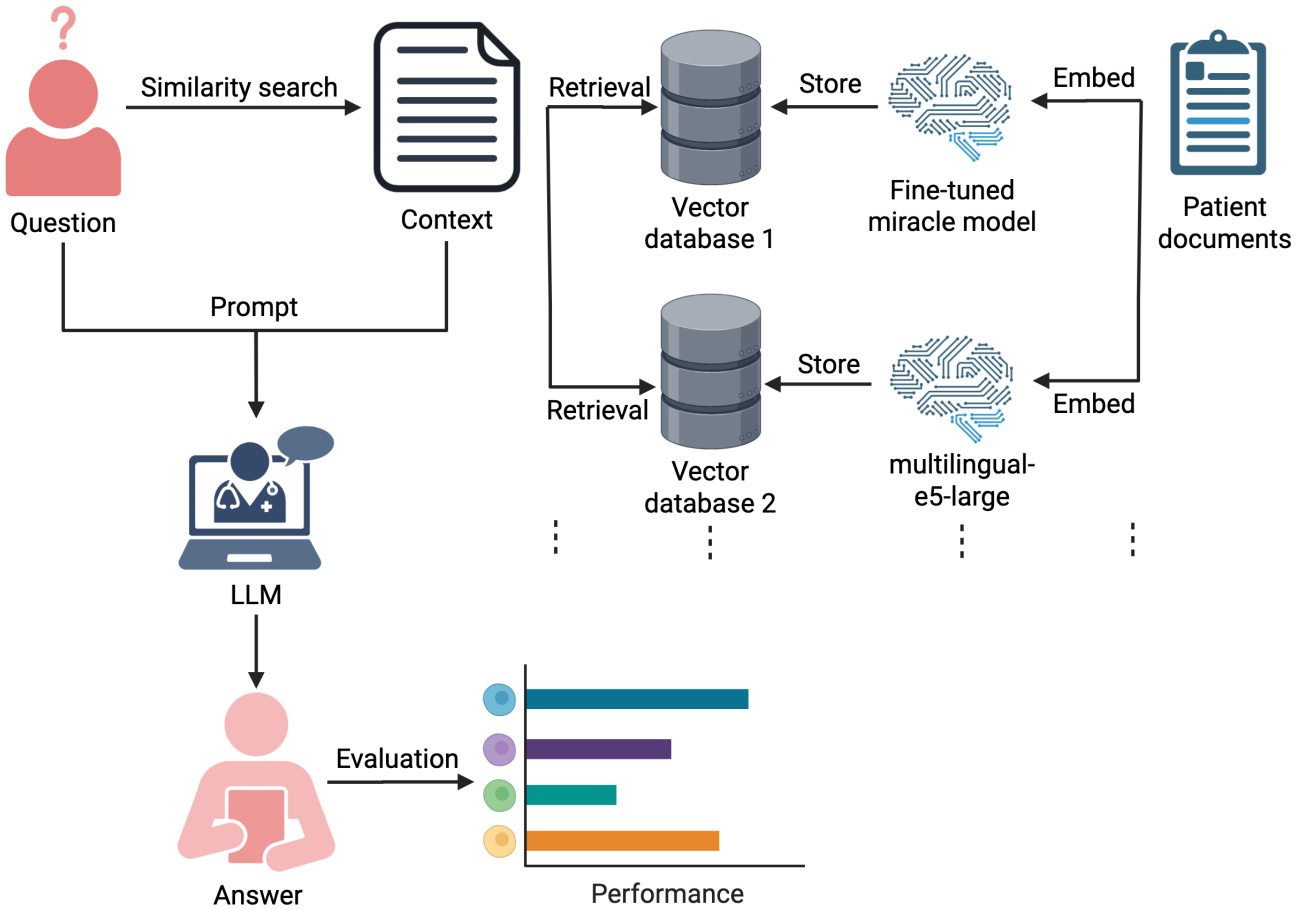

This system received user questions, attempted to find relevant document passages, and provided these passages to an LLM which was prompted to answer these questions. Each question in this dataset was presented to the RAG system and evaluated against the ground truth answer. Different metrics were used to compare the output from LLMs to a ground truth such as ROUGE or BLEURT. Additionally, retrieval-specific metrics such as Contextual Precision, Contextual Recall, and Contextual Relevancy were calculated using the *deepeval* Python package (version 3.7.5). These metrics are based on the LLM-as-a-judge [36] method, using an LLM to assess the quality of the retrieved context and measuring whether the context is relevant to the asked question. For the RAG evaluation, the *gpp-oss-120b* model [37] was deployed and used as the LLM-as-a-judge model. Selected for its high parameter count (120B) and strong performance on reasoning benchmarks, this model serves as an automated evaluator for semantic coherence and relevance, offering a reproducible open-source alternative to proprietary application programming interface–based judges. Detailed definition and explanation of all exhibited metrics are provided in Multimedia Appendix 5.

The fine-tuned embedding model was integrated into the retrieval component of this system, and the results were compared with those of the non–fine-tuned model. Given that the RAG system was identical for both the fine-tuned and non–fine-tuned embedding models, except for the embedding model itself, it was assumed that the greater the number of questions answered correctly, the more accurately the clinical context was identified by the embedding model. For the LLM component of the RAG system, the *SauerkrautLM-SOLAR-Instruct* model was used to answer the provided questions. Since this model was also used to generate the training data for the fine-tuned models, there is a possibility of bias when integrated into the RAG system. Therefore, the *Qwen3-235B-A22B-Instruct-2507-FP8* model was additionally used for generating answers. The maximum number of chunks retrieved for each question was set to 3. Cosine similarity was used as a distance metric for comparing questions and chunks in the similarity search. The smaller this distance, the more similar the question and chunk were considered to be. For easier assessment of the evaluation, Multimedia Appendix 6 shows a prompt example that was used to evaluate answers generated from the LLM using the RAG technique.

The RAG evaluation setup was divided into 2 parts. For the first part, the documents that the RAG system can retrieve were not restricted. This means that for 1 question, document parts can be retrieved from several patients in order to answer the question. While this indicates performance for IR settings, this constellation would rarely occur in real-world scenarios. Therefore, an additional filter mechanism was introduced for the second part to ensure that only documents from 1 patient can be retrieved for each question. Both configurations are

XSL•FO
RenderX

referred to as cross-patient (first part) and patient-centered evaluation (second part) in the following sections.

Cross-Patient Setup

To prepare a dataset for the cross-patient RAG evaluation, 372 clinical documents from 1 patient each were split into 1129 document chunks. These patients were not included in the training or test dataset that was used to fine-tune the sentence transformer model. The SauerkrautLM-SOLAR-Instruct model was then asked to generate questions and corresponding answers based on the contextual information provided by the chunks. As a result, 5556 synthetic question-answer pairs were generated. These question-answer pairs were filtered by a physician to provide a realistic clinical scenario and human-assisted evaluation. The physician was instructed to remove any questions related to administrative data (eg, questions about the patient's birth date or names of relatives) or remove questions that had wrong corresponding answers. It was emphasized that only those questions that would be posed in a practical clinical setting and that were factually and semantically accurate should be included. After filtering, 3536 annotated question-answer pairs from 280 documents split into 890 chunks remained for evaluation. These 890 document chunks were embedded by the fine-tuned and non–fine-tuned sentence transformer model, and the embeddings were stored in separate databases.

Patient-Centered Setup

The validated dataset from the cross-patient evaluation was used for the patient-centered RAG evaluation. As with the cross-patient setup, all documents included in this evaluation were sourced exclusively from patients who were held out from the training dataset. In total, the dataset consisted of 487 documents from 1 patient each and was divided into 2123 chunks. For each chunk, questions and corresponding answers were generated using the SauerkrautLM-SOLAR-Instruct model. As a result, 10,615 question-answer pairs were generated.

These additional questions and answers with the corresponding document chunks were annotated by 1 physician and 2 medical students. The students represented a range of academic backgrounds, including one student who has passed the first state examination for medicine and is currently in her seventh semester, and another who has studied and completed molecular biology and is currently in her second semester of pharmacy studies. Similar to the cross-patient data filtering described in the previous section, the annotators were instructed to remove any nonmedical questions and to include only those questions that would be asked in a realistic clinical setting. To gain insight into the quality of the annotations by measuring the agreement between the annotators, the annotators were divided into 2 groups. The first group consisted of 2 physicians, including 1 physician who had already annotated the cross-patient dataset and whose previously annotated data were reused for this comparison. The 2 medical students formed the second group. Each group received the same questions and answers. Interannotator agreement (IAA) was calculated based on precision, recall, and F_1-score between the 2 annotators of 1 group. One annotator was designated as the ground truth, while the other served as the estimator. The annotated questions and answers were used as ground truth for patient-centered RAG evaluation. For each question, a filter was introduced, ensuring that only document chunks from a single patient were retrieved.

Results

Training Data Quality Assessment

To evaluate the reliability of the synthetically generated training corpus, a manual audit was conducted on a statistically random sample of 200 question-answer pairs, including corresponding document chunks, from the training dataset. This audit was performed by a physician to quantify the prevalence of generated artifacts. The analysis revealed that 36 pairs (18.0%) contained hallucinations, defined as questions for which the answer could not be derived from the provided document chunk. Additionally, 12 (6.0%) pairs contained factual errors or medically invalid reasoning. Consequently, 76.0% of the sample was verified as semantically accurate and clinically correct.

IR Evaluation

The IR evaluation metrics indicated a stronger performance of the fine-tuned embedding model, with a mAP@100 score improved from 0.14 to 0.27. The model fine-tuned on pseudonymized data also improved after training, achieving a mAP@100 of 0.25. The NDCG@10 metric increased after fine-tuning from 0.16 to 0.29 for the *miracle* model and to 0.27 for the pseudonymized model. Besides the fine-tuned models, the best-performing model for the German dataset was *multilingual-e5-large*. The *bge-m3* model scored a mAP@100 of 0.11 and a NDCG@10 of 0.12. The *gte-multilingual-base* model achieved slightly higher scores with a mAP@100 of 0.12 and a NDCG@10 of 0.13. The *german-bge-m3* model showed higher performances than the base *bge-m3* model, indicating that fine-tuning on German datasets improved language-specific capabilities. On the translated dataset, the *miracle* model also achieved the highest scores with improvements for mAP@100 from 0.12 to 0.27 and 0.25 for the pseudonymized model. The *bge-m3* model showed lower scores with a mAP@100 of 0.11. The *gte-multilingual-base* model scored a mAP@100 of 0.12. The results on the German dataset are visualized in Figure 3. The complete results for different k values and languages are appended in Multimedia Appendices 7 and 8.

XSL•FO
RenderX

Figure 3. Comparative illustration of the information retrieval evaluation results for the German dataset. The results are presented for a k value of 10 for MRR, recall, precision, and NDCG. Abbreviations: mAP: mean average precision; MRR: mean reciprocal rank; NDCG: normalized discounted cumulative gain.

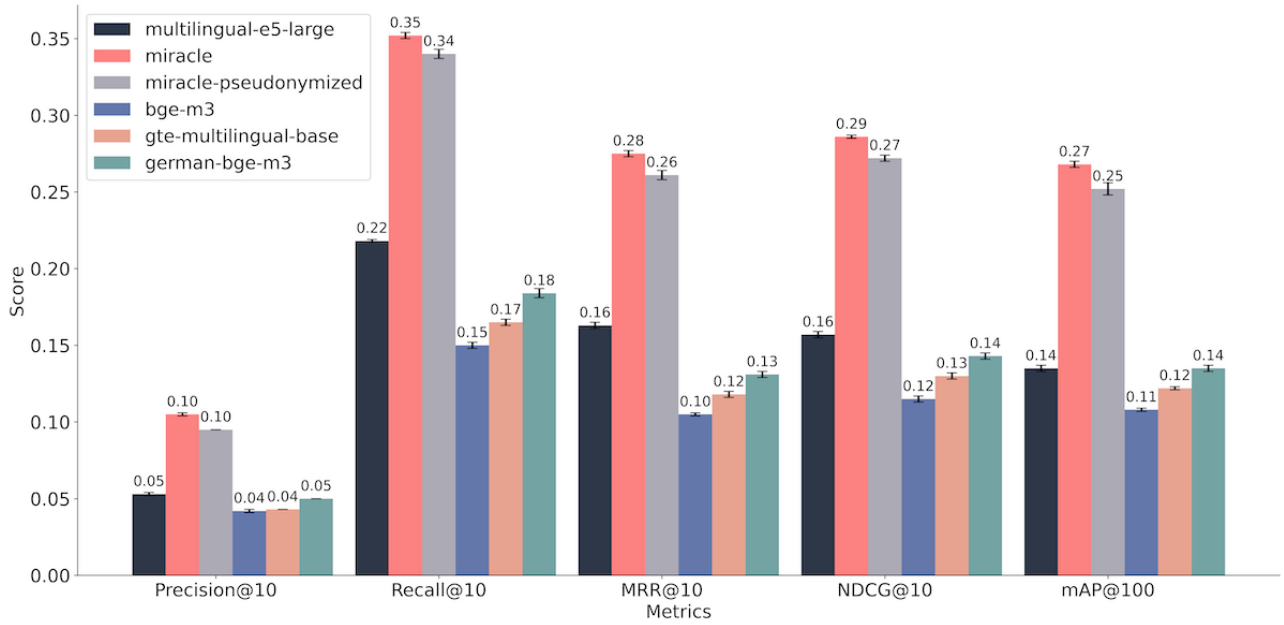

RAG Evaluation

Cross-Patient Results

The overall answer quality of the RAG system improved after fine-tuning the sentence transformer model. The BERTScore-F1 went up from 0.76 to 0.77. Improvements were also observed in both the ROUGE-2 score, which increased from 0.18 to 0.30, and the BLEURT score, which rose from 0.53 to 0.56. Other models such as *bge-m3* and *gte-multilingual-base* achieved lower scores in total, with a BERTScore-F1 score of 0.74 and a ROUGE-2 score of 0.17 for *bge-m3* and 0.16 for *gte-multilingual-base*. The fine-tuned variant of *bge-m3* (*german-bge-m3*) again showed better scores than the base model with a BERTScore-F1 of 0.75 and a ROUGE-2 score of 0.18. The highest scores for Contextual Precision and Contextual Recall were achieved by the pseudonymized model (0.82 and 0.89).

The results are in line with the answers generated by Qwen-3 (Multimedia Appendix 9). Because of the higher capabilities of this model, scores that are dependent on the generative part, such as BERTScore or ROUGE, are higher for every evaluated model. However, relative differences between the models stay intact, with the fine-tuned *miracle* model reaching the highest scores for most metrics (BERTScore F1: 0.78, ROUGE-2: 0.26) and showing improvements compared with the *multilingual-e5-large* (BERTScore F1: 0.76, ROUGE-2: 0.23).

For the translated dataset, the pseudonymized model achieved the highest scores with a BERTScore-F1 of 0.75 and a BLEURT score of 0.50. Of the non–fine-tuned models, the *bge-m3* model achieved the best scores for the English dataset with a BERTScore-F1 of 0.75 and a BLEURT score of 0.48. The RAG evaluation results for the German dataset are shown in Table 2. The results for the English dataset are provided in Multimedia Appendix 10.

XSL•FO
RenderX

Table . Retrieval Augmented Generation reevaluation results for the German dataset[a].

Metric	multilingual-e5-large	miracle	miracle pseudonymized	bge-m3	gte-multilingual-base	german-bge-m3
BERTScore P[b]	0.733	*0.747* [d]	0.736	0.725	0.725	0.728
BERTScore R[c]	0.783	*0.796* [d]	0.792	0.760	0.756	0.779
BERTScore F1	0.756	*0.769* [d]	0.762	0.742	0.739	0.747
BLEURT	0.533	*0.563* [d]	0.551	0.542	0.539	0.553
ROUGE-1	0.327	*0.351* [d]	0.342	0.319	0.311	0.327
ROUGE-2	0.179	*0.197* [d]	0.194	0.167	0.159	0.175
ROUGE-L	0.276	*0.297* [d]	0.292	0.269	0.258	0.274
Contextual P	0.685	0.779	*0.823* [d]	0.626	0.567	0.676
Contextual R	0.774	0.860	*0.892* [d]	0.726	0.676	0.783
Contextual Relevancy	0.307	*0.312* [d]	0.309	0.274	0.268	0.310

[a]For the Retrieval Augmented Generation (RAG) evaluation, generated answers by the RAG system are compared with ground truth answers that were created by a large language model and filtered by a medical doctor. The higher the scores are, the more similar the answers are considered to the ground truth. The answers were generated by the *SauerkrautLM-Solar-Instruct* model.

[b]P: precision.

[c]R: recall.

[d]The values in italics highlight the highest scores in the respective category.

Patient-Centered Results

For the patient-centered RAG evaluation, the LLM-generated questions and answers (pairs) were filtered by 2 annotation groups. In Table 3, the results of the filtering process are displayed.

Table . Human filtering results of Retrieval Augmented Generation evaluation dataset[a].

Annotator	Annotator group	Education	Pairs checked	Pairs filtered out, n (%)	IAA[b]
Annotator 1	Group 1	Physician	5556	2020 (36.36)	Precision=0.854; recall=0.963; F_1-score=0.905
Annotator 2	Group 1	Physician	7081	1909 (26.96)	
Annotator 3	Group 2	Student (2nd semester)	3556	1813 (50.98)	Precision=0.763; recall=0.796; F_1-score=0.779
Annotator 4	Group 2	Student (7th semester)	3839	1844 (48.03)	

[a]The Retrieval Augmented Generation evaluation dataset, which was partly created by a large language model, was given to 2 physicians and 2 medical students with the instruction to filter out questions or answer pairs that were medically unrelated. The dataset was divided into 2 parts, and each part was provided to 1 annotator group, resulting in 2 annotator groups. In the table, results of the human annotator process are displayed. The agreement between the annotators was calculated based on the overlapping annotated pairs of each annotator group (5556 for annotator group 1 and 3556 for annotator group 2).

[b]IAA: interannotator agreement.

Comparisons between annotator groups 1 and 2 show that there are differences in the number and percentage of filtered out question and answer pairs. While annotators 1 and 2 filtered out 36.36% and 26.96% of the dataset, annotators 3 and 4 filtered out 50.98% and 48.03%. IAA for group 1 was higher with an F_1-score of 0.905 than for group 2, who reached an F_1-score of 0.779.

The annotated question and answer pairs were used for the patient-centric RAG evaluation. For this evaluation scenario, the fine-tuned models again showed higher scores than the multilingual-e5-large, bge-m3, *german-bge-m3*, and gte-multilingual-base models. Compared with the results of the RAG evaluation without patient context, the differences between models have become smaller. The best-performing model, the model fine-tuned on pseudonymized data, achieved a

XSL•FO
RenderX

BERTScore-F1 of 0.78, a BLEURT score of 0.61, and a ROUGE-1 score of 0.42. In comparison, the multilingual-e5-large reached a BERTScore-F1 of 0.78, a BLEURT score of 0.61, and a ROUGE-1 score of 0.41. Similar to the cross-patient evaluation, the highest scores for Contextual Precision and Contextual Recall were reached by the model trained on pseudonymized data. Interestingly, the highest score for Contextual Relevancy was achieved by the *german-bge-m3* model. The evaluation performed using the Qwen-3 as the generator showed similar results (Multimedia Appendix 11), with improvements for the fine-tuned models in all metrics compared with the *multilingual-e5-large* model. In Table 4, the RAG evaluation results for the patient-centric context are presented.

Table . Retrieval Augmented Generation patient-context evaluation results[a].

Metric	multilingual-e5-large	miracle	miracle pseudonymized	bge-m3	gte-multilingual-base	german-bge-m3
BERTScore P[b]	0.763	*0.766* [c]	*0.766* [c]	0.764	0.761	0.756
BERTScore R[d]	0.795	*0.799* [c]	*0.799* [c]	0.796	0.793	0.793
BERTScore F1	0.778	*0.781* [c]	*0.781* [c]	0.778	0.775	0.773
BLEURT	0.608	*0.611* [c]	*0.611* [c]	0.608	0.604	0.593
ROUGE-1	0.414	0.420	*0.421* [c]	0.415	0.407	0.394
ROUGE-2	0.249	*0.256* [c]	*0.256* [c]	0.250	0.242	0.232
ROUGE-L	0.355	0.361	*0.362* [c]	0.355	0.349	0.335
Contextual P	0.868	0.918	*0.927* [c]	0.858	0.824	0.880
Contextual R	0.927	0.938	*0.939* [c]	0.925	0.909	0.929
Contextual Relevancy	0.239	0.241	0.241	0.235	0.233	*0.242* [c]

[a]For the Retrieval Augmented Generation (RAG) evaluation, generated answers by the RAG system are compared with ground truth answers that were created by a large language model and filtered by 4 human annotators. The higher the scores are, the more similar the answers are considered to the ground truth. The answers were generated by the *SauerkrautLM-Solar-Instruct* model.

[b]P: precision.

[c]The values in italics highlight the highest scores in the respective category.

[d]R: recall.

For more nuanced performance differences between the models, the patient-centric performances were broken down by document category (Multimedia Appendix 12). For discharge letters, radiology reports, and operation notes, the fine-tuned models show the highest scores, with only slight differences between the models. The results on pathology reports indicate that there is a stronger difference in performance between models, with the *gte-multilingual-base* achieving the highest BERTScore-F1 and ROUGE scores.

Discussion

Principal Findings

This study developed and tested language-specific clinical embedding models specifically trained on verified hospital documents for medical RAG applications and demonstrated their higher accuracy metrics. The introduced approach and models address an important gap in NLP applications for health care by improving clinical language comprehension for multiple languages and by optimizing RAG systems within medical information environments. Without domain fine-tuning, it is challenging for embedding models to comprehend clinical language and contexts, as they are not trained on real-world medical documents. The integration of these models as retrieval components in RAG systems may lead to undesired or medically irrelevant results. Fine-tuning embedding models on medical real-world data represent an advancement in the field of medical NLP and improve artificial intelligence–driven health care by facilitating the understanding of clinical language.

Previous research focused on English language models for NLP and has proven its application in various clinical settings [38,39]. This strategy, however, has faced shortcomings in performance, in particular, in the multilanguage setting. Our models have been fine-tuned on German medical data, a language not frequently used in medical NLP [20], and whose documents are currently less abundant. The presented work not only complements but also diversifies the linguistic scope of medical NLP research and its clinical applicability. The extracted dataset from 4 medical application areas had highly nuanced and domain-specific medical language. By using an LLM to generate question and answer pairs for these documents, an extensive dataset with approximately 11 million pairs served as a foundation for fine-tuning and evaluation. Different evaluation benchmarks introduced in this study provide insights into improvements achieved after fine-tuning the models on this dataset.

XSL•FO
RenderX

Importantly, these fine-tuned context and language-sensitive miracle models surpassed the state-of-the-art *multilingual-e5-large*, *bge-m3*, *german-bge-m3*, and *gte-multilingual-base* models in 2 evaluation scenarios. The model trained on real-world data showed the best overall results for the German evaluation datasets. More interestingly, the model fine-tuned on pseudonymized data achieved similar results on the IR and even better results for the RAG evaluation on the English-translated dataset.

One hypothesis is that the pseudonymization process may have resulted in an improved machine readability of certain phrases, which in turn may have positively influenced the quality of the translation. However, the models fine-tuned on translated data overall showed lower scores than the models trained on real-world data. Noise and possibly difficult translations of abbreviations might have influenced the dataset quality.

Performance comparisons, particularly with the *german-bge-m3* model, showed that language-specific fine-tuning generally can improve retrieval scores. However, the improvements achieved by fine-tuning on real-world clinical data indicate that further enhancements can be achieved by domain adaptation.

The results of the patient-centered RAG evaluation show that the fine-tuned models are able to find suitable text passages more accurately in a patient question answering setting. The category-wise results indicate differences in performance depending on the document type. As documents from different clinics and departments are written in different styles with varying language and form, the retrieval of suitable text passages is highly dependent on the ability of the models to semantically understand and interpret medical language. The highest scores were achieved for discharge letters and radiology reports, indicating simpler language compared with pathology reports and surgical reports, which may contain more specific and nuanced language.

By having physicians and medical students annotate the evaluation dataset for the patient-centered evaluation, cautious estimates can be made about the quality of the dataset. The human annotation results highlight both the limitations and effectiveness of synthetic data generation in medical contexts. While 27% - 51% of LLM-generated pairs required filtering, the remaining validated content still provided sufficient evaluation data. The variation in filtering rates between annotators likely reflects differences in clinical experience and criteria for data quality. Despite these quality variations, the consistent performance improvements observed across all evaluation scenarios suggest that the signal-to-noise ratio in the synthetic training data was sufficient for effective model fine-tuning. These findings indicate that while expert validation improves data quality, synthetic generation can still yield functional training corpora for specialized domains such as clinical NLP.

Serving as the foundation for our research, the provided dataset consisting of real-world clinical texts and LLM-generated question-answer pairs offers a comprehensive training resource that captures the nuances of medical terminology and clinical reasoning. Building upon this dataset, the presented technologies indicate a potential to transform how health care professionals access and use medical information. An example of how this technology could be integrated into clinical routine is visualized in Multimedia Appendix 13.

Regarding operational implementation in hospital IT infrastructures, the deployment of the fine-tuned embedding model presents a low barrier to entry. Unlike the training phase or the generative components of RAG systems, which often require GPU acceleration, the retrieval component based on dense vector embeddings is computationally efficient during inference. The model can be deployed on standard, CPU-based server environments commonly found in hospital data centers. This efficiency allows for seamless integration into existing clinical workflows and decision support systems without imposing additional infrastructure costs or latency, making the solution highly scalable for real-time IR across large patient cohorts.

Limitations

While our models indicate a step forward for the development of clinical embedding models for RAG systems, they are not without limitations. The dataset, though comprehensive, is limited to 4 types of clinical documents and may benefit from expansion by including a greater variety of medical texts. Furthermore, as the documents were extracted from 1 site only, the model is susceptible to template overfitting. Clinical documents from a single institution often share rigid structural templates and specific formatting cues, creating a risk that the model learns to associate relevance with these institutional artifacts rather than purely semantic content. Consequently, the performance observed here may not fully transfer to institutions with different documentation standards without local fine-tuning. Additionally, the English language model was trained on machine-translated data using the WMT'19 German-to-English translation model, with translation quality validated on only 30 physician-reviewed samples. This sample size is statistically insufficient to ensure translation accuracy across the entire corpus of 400,000 clinical documents. Machine translation of specialized medical terminology, abbreviations, and context-dependent clinical language may introduce systematic translation artifacts that differ from native English clinical documentation. The models fine-tuned on translated data should therefore be considered a proof of concept for cross-lingual transfer learning rather than a validated clinical tool ready for deployment in English-speaking health care settings.

Furthermore, the current evaluation compared performance exclusively against multilingual baselines to align with the model's architectural foundation. We did not benchmark against embedding models trained exclusively on English clinical data. Consequently, while our results demonstrate the model's efficacy within a multilingual framework, its relative performance against specialized, monolingual English clinical models remains to be validated in future studies.

A circular evaluation bias may be inherent to the experimental design, as the test questions were synthetically generated using the same LLM family used for training. This shared origin implies that the evaluation might partially measure the model's ability to recognize synthetic linguistic patterns rather than organic human queries. To mitigate this, the

XSL•FO
RenderX

Qwen3-235B-A22B-Instruct-2507-FP8 was additionally used as the generator part of the RAG system. The results showed overall higher scores but no significant relative performance differences between the evaluated embedding models.

Furthermore, the IR evaluation was conducted in a cross-patient setting where synthetically generated questions could match document chunks from multiple patients, with all matches treated as relevant. A limitation of this methodology is that it relies on identifying duplicate questions, which inherently introduces a selection bias toward generic queries (eg, “What is the diagnosis?”) that are likely to appear multiple times across the corpus. Consequently, complex, specific, or nuanced questions that are unique to a single patient case were systematically excluded. This bias may result in an overestimation of performance on simple lookup tasks while underrepresenting the challenges of complex clinical reasoning. Additionally, treating all matches as relevant introduces potential label noise, as generic questions may retrieve contextually appropriate chunks from different patients that answer different clinical questions. However, this evaluation design does capture legitimate clinical use cases, including searching for similar patient presentations, researching treatment outcomes across cases, or conducting retrospective cohort analyses. To address patient-specific scenarios, we conducted a separate patient-centered RAG evaluation with retrieval restricted to single patients, which demonstrated the models’ effectiveness in realistic clinical contexts.

The observed low absolute IR performance metrics must be interpreted in the context of the extensive search space, which comprised 173,933 document chunks across 26,358 patients. This high-dimensional setting renders specific retrieval statistically challenging compared with constrained, single-patient scenarios. The evaluation methodology relied on identifying exact duplicate synthetic questions to define ground truth. This rigid definition of relevance likely penalized semantically redundant chunks that contained relevant information but generated slightly different synthetic questions. Such instances are treated as irrelevant in our evaluation, resulting in potentially high false-negative rates. Consequently, strict metrics such as Accuracy@1 may underestimate the model’s true clinical usefulness, as they measure the model’s ability to retrieve the exact source chunk rather than its ability to retrieve any chunk containing the correct semantic answer. The RAG evaluation does not directly measure the IR but the answer quality, which also could be biased by the LLM component of the RAG system. However, the combination of different evaluation scenarios allows for a more balanced and diverse look at the performance of the models.

In the patient-centered RAG evaluation, the fine-tuned miracle model consistently achieved the highest numerical scores across all metrics. However, we explicitly acknowledge that the performance differences compared with the multilingual-e5-large baseline were marginal. A formal statistical significance test was not performed for these specific evaluation metrics. Consequently, the statistical significance of these performance differences has not been established. It remains possible that the observed marginal improvements fall within the margin of error and may result from random variance rather than a systematic model advantage. This stands in contrast to the larger performance deltas observed in the cross-patient settings. We hypothesize that this convergence may be attributable to a potential saturation effect caused by the constrained search space. In the patient-centered setting, the retrieval pool is restricted to a single patient’s history (typically comprising fewer than 30 document chunks). With such a limited number of distractor chunks, the baseline model already achieves near-ceiling retrieval performance, potentially masking the benefits of domain-specific fine-tuning that become more apparent in larger, high-dimensional retrieval spaces.

We further acknowledge a discrepancy in the IAA observed during the dataset validation. While the physician group achieved a high F_1-score of 0.905, the medical student group achieved a lower score of 0.779. This variance is potentially attributable to differences in clinical experience, as students may lack the consolidated clinical intuition required to consistently apply exclusion criteria to borderline cases. As the student-annotated portion of the dataset was used as ground truth without a secondary confirmation review by a senior clinician, this introduces a potential for label noise. Consequently, the evaluation metrics for the patient-centered scenario should be interpreted with this underlying annotator variability in mind.

A notable finding in our RAG evaluation was the consistently low Contextual Relevancy scores (approximately 0.24) observed across all models. This indicates that roughly three-quarters of the content within a retrieved chunk was classified as nonrelevant to the specific query by the judge model. We attribute this to the high information density characteristic of clinical documentation. Even with a constrained text segment of 450 characters, clinical notes frequently contain multiple distinct, packed statements. Consequently, while the chunk contains the correct answer, it inevitably includes competing clinical facts that lower the relevancy ratio. This highlights a persistent challenge in medical IR: finding an optimal granularity that isolates specific findings without destroying the narrative context.

A further limitation concerns the text segmentation strategy. We used a fixed chunk size of 450 characters with an 80-character overlap. While this granularity enhances the precision of retrieval for specific clinical facts, it risks context fragmentation, particularly for complex temporal reasoning or negation scopes that span across multiple segments (eg, a diagnosis mentioned at the beginning of a letter and a treatment plan at the end). Furthermore, an alternative atomic-level chunking strategy (eg, sentence-level splitting) could theoretically maximize relevancy scores for closed-ended questions (eg, “What is the patient’s sodium level?”). However, we deliberately avoided this approach to prioritize semantic coherence. Atomic splitting risks fragmenting the narrative flow essential for open-ended clinical reasoning (eg, linking a diagnosis to its ruling out criteria in the next sentence). Therefore, the reported low relevancy scores reflect a trade-off where we accepted lower precision in signal-to-noise ratio to preserve the integrity of the local clinical context.

XSL•FO
RenderX

LLMs, including the *SauerkrautLM-SOLAR-Instruct* model used for data generation, are susceptible to hallucinations and may generate factually incorrect or clinically inappropriate content [40]. This introduces potential noise into the training dataset that could propagate to the fine-tuned embedding models. Analysis of a random sample indicated that approximately 24% of the generated question-answer pairs contained hallucinations or factual inaccuracies. This introduces a safety concern, as this specific type of noise has the potential to propagate into the embedding space, effectively teaching the model incorrect semantic associations. While our downstream performance suggests that the valid signal was dominant, we explicitly warn that the model was trained on a corpus with a known error rate. Therefore, the resulting embeddings should not be treated as ground truth. Any application of these models in clinical practice must mandate robust downstream filtering and expert human verification to identify and intercept potential retrieval errors stemming from these training artifacts. For evaluation, the question-answer pairs were also initially generated by an LLM. However, a rigorous human-in-the-loop validation process involving physicians and medical students was conducted to mitigate the risks associated with purely synthetic benchmarks. This resulted in a filtering rate of approximately 27%-51%, ensuring that the final test set represents a human-verified standard of clinical relevance.

A further limitation concerns the evaluation scope regarding negative testing. While the RAG system prompt included instructions to refuse answering if the context was insufficient (eg, stating "'I don't know"), our quantitative evaluation focused exclusively on positive retrieval scenarios, specifically cases where the answer was present in the ground truth. We did not systematically assess the model's ability to correctly reject irrelevant context or quantify the rate of hallucinations when no suitable information is retrieved. Validating this refusal capability is an important safety requirement for clinical deployment. Without it, there is a risk that the system might confidently hallucinate answers based on retrieved but irrelevant distractor chunks.

Relatedly, the quantitative assessment of the RAG system relied on automated semantic similarity metrics such as BERTScore, BLEURT, and ROUGE. While these metrics are standard benchmarks for measuring the linguistic overlap between generated answers and ground truth, they are not designed to verify factual medical precision. High semantic scores can occasionally mask critical factual errors, such as laterality swaps (eg, "left" vs "right") or negation flips (eg, "no evidence of" vs "evidence of"), which may have high lexical overlap with the correct answer.

Future Work

Future validation on external datasets from different sites and the inclusion of additional document categories would strengthen the approach and indicate broader applicability. To facilitate this necessary next step, we have publicly released the models fine-tuned on pseudonymized and translated data. This allows the broader research community to validate the miracle models on their local datasets, thereby establishing their generalizability across different clinical environments in future work.

While in the study, quantitative metrics were assessed for comparison of embedding models, evaluation with clinicians is an important next step to evaluate clinical performance for real-world scenarios. Evaluation with end users could give insights into the integration of the fine-tuned models in a real-world setting and should be investigated in future studies. Future clinical validation studies must use expert human verification to audit generated responses for specific types of high-risk factual hallucinations. Future studies should also include dedicated patient-specific IR benchmarks with queries explicitly designed for individual patient contexts. Additionally, it remains to be validated whether the domain-specific fine-tuning makes the model more robust in extracting correct answers or potentially more prone to hallucinating responses due to an increased bias toward using retrieved clinical text.

Comparisons between the fine-tuned models and domain-adapted embedding models could strengthen the approach and give insights into the quality of the QA dataset. However, to our knowledge, there are no publicly available fine-tuned sentence transformer models trained on German medical real-world data.

The retrieval parameter k (number of retrieved chunks) was fixed at 3 for all RAG experiments. We acknowledge that this parameter was not empirically optimized through ablation studies within this work. While the parameter value was selected to provide a reasonable balance between context sufficiency and noise reduction, it is possible that the choice of k influenced the reported performance metrics. Determining the optimal retrieval depth or using dynamic threshold-based retrieval strategies requires further investigation.

Although the recursive splitting method respects paragraph boundaries to maintain local semantic integrity, future studies should explore adaptive segmentation strategies. While section-aware chunking or hierarchical retrieval can better capture long-range dependencies, future research should also investigate hybrid segmentation pipelines. Such approaches could dynamically use atomic-level chunking for high-density, structured data (eg, laboratory values and medication lists) to maximize precision, while retaining larger, section-aware windows for unstructured narrative prose (eg, anamnesis and discharge summary) to preserve semantic flow. Investigating the optimal combination of these strategies remains an important step toward production-grade clinical RAG systems.

Furthermore, while this study evaluated performance using dense vector retrieval, clinical decision-making often requires reasoning across complex, interrelated entities (eg, drug-drug interactions or temporal disease progressions). Future research should explore integrating these fine-tuned embedding models into Graph Retrieval Augmented Generation systems. In such architectures, our domain-adapted embeddings could serve to enhance node resolution and relationship traversal within clinical knowledge graphs, potentially addressing the limitations of purely vector-based retrieval for multihop reasoning tasks.

Finally, our evaluation focused specifically on the performance of dense vector retrieval. We did not benchmark against lexical approaches such as BM25 or multistage pipelines involving cross-encoder reranking or multivector architectures such as

XSL•FO
RenderX

ColBERT. While lexical search can be effective for exact keyword matching, it often fails to capture the semantic nuances of medical terminology. However, we acknowledge that in a production-grade clinical RAG system, the fine-tuned miracle embeddings presented here would likely serve as the retrieval foundation within a hybrid pipeline, potentially supplemented by keyword search and followed by a reranking step to maximize precision. Further research should also explore section-aware chunking or hierarchical retrieval approaches to better capture long-range dependencies in clinical narratives.

Conclusions

This study demonstrates that embedding models fine-tuned on real-world clinical documents can outperform existing state-of-the-art multilingual models for medical IR and RAG applications. The approach of generating synthetic question-answer pairs from clinical documents using LLMs provides a scalable methodology for creating domain-specific training data while maintaining patient privacy through pseudonymization. The models' effectiveness across both German and English datasets indicates applicability beyond single-language contexts. This work establishes a reproducible framework for developing specialized embedding models that address the gap between general-purpose language models and the specific requirements of health care documentation, particularly for non-English medical settings where such resources are limited.

Acknowledgments

The work of BE and MB was funded by a PhD grant from the DFG Research Training Group 2535 Knowledge- and data-based personalization of medicine at the point of care (WisPerMed). A special thanks goes to the Annotation Lab for the annotation effort. Figures 1 and 2 were created with Biorender.com. The authors acknowledge the use of generative artificial intelligence during the early stages of manuscript drafting to assist in structuring initial ideas and refining wording. DeepL and Claude were used in later stages to enhance language clarity and improve readability. All content generated or suggested by these tools was thoroughly reviewed, revised, and edited by the authors, and the authors take full responsibility for the accuracy and integrity of the final manuscript. The data for this project were provided by the Smart Hospital Information Platform (SHIP), managed by the Data Integration Center at the University Medicine Essen. SHIP serves as a comprehensive digital health platform for integrating data from all major clinical subsystems using a holistic Fast Healthcare Interoperability Resources (FHIR)–based approach. It enables the purification, analysis, distribution, and visualization of clinical data.

Funding

The authors declared no financial support was received for this work.

Data Availability

The original data that support the findings of this study are sensitive clinical data and hence not openly available due to reasons of data protection and privacy. Requests regarding the dataset can be sent to datagovernance@uk-essen.de and will be reviewed. The final models trained on pseudonymized data created as part of this study are publicly accessible [41]. The underlying code for this study is made publicly available upon publication [42] and includes all scripts developed during the study.

Authors' Contributions

KA contributed to writing – original draft, training, visualization, validation, methodology, data curation, conceptualization. HS and AIY contributed to writing – review and editing, validation, methodology, and formal analysis. CSS contributed to writing – review and editing, formal analysis, and clinical perspective. BE and MB participated in writing – review and editing, validation, and data curation. AT participated in writing – review and editing, supervision, and conceptualization. OBP, EMH, PW, KB, and JH contributed to writing – review and editing, data curation, and formal analysis. FN contributed to writing – review and editing, supervision, and project administration. RH contributed to writing – review and editing, supervision, project administration, and conceptualization.

Conflicts of Interest

AT has stated a conflict. Research funding by Neovii Biotech and GTL. Consultancy for CSL Behring, Maat Pharma, Biomarin, Pfizer, and Onkowissen. Travel reimbursements by Neovii and Novartis. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Additional information about cosine similarity.
[DOCX File, 14 KB - jmir_v28i1e82997_app1.docx]

Multimedia Appendix 2
Additional information about precision, recall, and F_1-score.
[DOCX File, 13 KB - jmir_v28i1e82997_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Parameter size of evaluated models.
[DOCX File, 14 KB - jmir_v28i1e82997_app3.docx]

Multimedia Appendix 4
Additional information about information retrieval metrics.
[DOCX File, 15 KB - jmir_v28i1e82997_app4.docx]

Multimedia Appendix 5
Additional detailed information about Retrieval Augmented Generation metrics.
[DOCX File, 15 KB - jmir_v28i1e82997_app5.docx]

Multimedia Appendix 6
Prompt example for evaluation of Retrieval Augmented Generation systems.
[DOCX File, 15 KB - jmir_v28i1e82997_app6.docx]

Multimedia Appendix 7
Information retrieval evaluation results on the German dataset.
[DOCX File, 16 KB - jmir_v28i1e82997_app7.docx]

Multimedia Appendix 8
Information retrieval evaluation results on English dataset.
[DOCX File, 16 KB - jmir_v28i1e82997_app8.docx]

Multimedia Appendix 9
Retrieval Augmented Generation evaluation results on the German dataset in a cross-patient setting with answers generated by Qwen-3.
[DOCX File, 15 KB - jmir_v28i1e82997_app9.docx]

Multimedia Appendix 10
Retrieval Augmented Generation evaluation results on English dataset in cross-patient setting.
[DOCX File, 15 KB - jmir_v28i1e82997_app10.docx]

Multimedia Appendix 11
Retrieval Augmented Generation evaluation results on the German dataset in a patient-centered setting with answers generated by Qwen-3.
[DOCX File, 16 KB - jmir_v28i1e82997_app11.docx]

Multimedia Appendix 12
Retrieval Augmented Generation patient-centered evaluation results per document category.
[DOCX File, 2497 KB - jmir_v28i1e82997_app12.docx]

Multimedia Appendix 13
Mock-up implementation of clinical Retrieval Augmented Generation system with integration into electronic health record for patient question answering.
[DOCX File, 800 KB - jmir_v28i1e82997_app13.docx]

References

1. Weissler EH, Zhang J, Lippmann S, Rusincovitch S, Henao R, Jones WS. Use of natural language processing to improve identification of patients with peripheral artery disease. Circ Cardiovasc Interv 2020 Oct;13(10):e009447. [doi: 10.1161/CIRCINTERVENTIONS.120.009447] [Medline: 33040585]
2. Joseph JW, Leventhal EL, Grossestreuer AV, et al. Deep-learning approaches to identify critically Ill patients at emergency department triage using limited information. J Am Coll Emerg Physicians Open 2020 Oct;1(5):773-781. [doi: 10.1002/emp2.12218] [Medline: 33145518]

XSL•FO
RenderX

3. Berge GT, Granmo OC, Tveit TO, Munkvold BE, Ruthjersen AL, Sharma J. Machine learning-driven clinical decision support system for concept-based searching: a field trial in a Norwegian hospital. BMC Med Inform Decis Mak 2023 Jan 10;23(1):5. [doi: 10.1186/s12911-023-02101-x] [Medline: 36627624]
4. Shi J, Gao X, Kinsman WC, Ha C, Gao GG, Chen Y. DI++: a deep learning system for patient condition identification in clinical notes. Artif Intell Med 2022 Jan;123:102224. [doi: 10.1016/j.artmed.2021.102224] [Medline: 34998515]
5. Chen TL, Emerling M, Chaudhari GR, et al. Domain specific word embeddings for natural language processing in radiology. J Biomed Inform 2021 Jan;113:103665. [doi: 10.1016/j.jbi.2020.103665] [Medline: 33333323]
6. Wang S, Tseng B, Hernandez-Boussard T. Development and evaluation of novel ophthalmology domain-specific neural word embeddings to predict visual prognosis. Int J Med Inform 2021 Jun;150:104464. [doi: 10.1016/j.ijmedinf.2021.104464] [Medline: 33892445]
7. Shtar G, Greenstein-Messica A, Mazuz E, Rokach L, Shapira B. Predicting drug characteristics using biomedical text embedding. BMC Bioinformatics 2022 Dec 7;23(1):526. [doi: 10.1186/s12859-022-05083-1] [Medline: 36476573]
8. Goswami K, Dutta S, Assem H, Fransen T, McCrae JP. Cross-lingual sentence embedding using multi-task learning. Presented at: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; Nov 7-11, 2021. [doi: 10.18653/v1/2021.emnlp-main.716]
9. Jiang X, Liang Y, Chen W, Duan N. XLM-K: improving cross-lingual language model pre-training with multilingual knowledge. AAAI 2022;36(10):10840-10848. [doi: 10.1609/aaai.v36i10.21330]
10. Kim Y, Kim JH, Lee JM, et al. A pre-trained BERT for Korean medical natural language processing. Sci Rep 2022;12(1):13847. [doi: 10.1038/s41598-022-17806-8]
11. Gemmiti M, Hamed S, Wildhaber J, Pharisa C, Klumb PL. Physicians' speech complexity and interrupting behavior in pediatric consultations. Health Commun 2022 May;37(6):748-759. [doi: 10.1080/10410236.2020.1868063] [Medline: 33441020]
12. Watermeyer J, Thwala Z, Beukes J. Medical terminology in intercultural health interactions. Health Commun 2021 Aug;36(9):1115-1124. [doi: 10.1080/10410236.2020.1735700] [Medline: 32202159]
13. Pitt MB, Hendrickson MA. Eradicating jargon-oblivion-a proposed classification system of medical jargon. J Gen Intern Med 2020 Jun;35(6):1861-1864. [doi: 10.1007/s11606-019-05526-1] [Medline: 31713041]
14. Hao T, Pan X, Gu Z, Qu Y, Weng H. A pattern learning-based method for temporal expression extraction and normalization from multi-lingual heterogeneous clinical texts. BMC Med Inform Decis Mak 2018 Mar 22;18(Suppl 1):22. [doi: 10.1186/s12911-018-0595-9] [Medline: 29589563]
15. Yang LWY, Ng WY, Lei X, et al. Development and testing of a multi-lingual natural language processing-based deep learning system in 10 languages for COVID-19 pandemic crisis: a multi-center study. Front Public Health 2023;11. [doi: 10.3389/fpubh.2023.1063466]
16. Sivarajkumar S, Wang Y. HealthPrompt: a zero-shot learning paradigm for clinical natural language processing. AMIA Annu Symp Proc 2022;2022:972-981. [Medline: 37128372]
17. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
18. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology. Ophthalmol Sci 2023 Dec;3(4):100324. [doi: 10.1016/j.xops.2023.100324] [Medline: 37334036]
19. Frei J, Frei-Stuber L, Kramer F. GERNERMED++: semantic annotation in German medical NLP through transfer-learning, translation and word alignment. J Biomed Inform 2023 Nov;147:104513. [doi: 10.1016/j.jbi.2023.104513] [Medline: 37838290]
20. Frei J, Kramer F. Annotated dataset creation through large language models for non-English medical NLP. J Biomed Inform 2023 Sep;145:104478. [doi: 10.1016/j.jbi.2023.104478] [Medline: 37625508]
21. Parola A, Lin JM, Simonsen A, et al. Speech disturbances in schizophrenia: assessing cross-linguistic generalizability of NLP automated measures of coherence. Schizophr Res 2023 Sep;259:59-70. [doi: 10.1016/j.schres.2022.07.002] [Medline: 35927097]
22. Toro S, Anagnostopoulos AV, Bello SM, et al. Dynamic retrieval augmented generation of ontologies using artificial intelligence (DRAGON-AI). J Biomed Semant 2024;15(1). [doi: 10.1186/s13326-024-00320-3]
23. Li Y, Zhao J, Li M, et al. RefAI: a GPT-powered retrieval-augmented generative tool for biomedical literature recommendation and summarization. J Am Med Inform Assoc 2024 Sep 1;31(9):2030-2039. [doi: 10.1093/jamia/ocae129] [Medline: 38857454]
24. Arzideh K, Baldini G, Winnekens P, et al. A transformer-based pipeline for German clinical document de-identification. Appl Clin Inform 2025 Jan;16(1):31-43. [doi: 10.1055/a-2424-1989] [Medline: 39778706]
25. Ng N, Yee K, Baevski A, Ott M, Auli M, Edunov S. Facebook FAIR's WMT19 news translation task submission. Presented at: Proceedings of the Fourth Conference on Machine Translation (Volume 2); Aug 1-2, 2019; Florence, Italy p. 314-319. [doi: 10.18653/v1/W19-5333]
26. Fan A, et al. Beyond english-centric multilingual machine translation. arXiv. Preprint posted online on Oct 21, 2020. [doi: 10.48550/ARXIV.2010.11125]

XSL•FO
RenderX

27. NLLB Team. No language left behind: scaling human-centered machine translation. arXiv. Preprint posted online on Jul 11, 2022. [doi: 10.48550/ARXIV.2207.04672]
28. Raffel C, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv. Preprint posted online on Oct 23, 2019. [doi: 10.48550/ARXIV.1910.10683]
29. Tang Y, et al. Multilingual translation with extensible multilingual pretraining and finetuning. arXiv. Preprint posted online on Aug 2, 2020. [doi: 10.48550/ARXIV.2008.00401]
30. Tiedemann J, Thottingal S, et al. OPUS-MT—building open translation services for the world. In: Martins A, editor. : European Association for Machine Translation; 2020 Nov 1 Presented at: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation p. 479-480 URL: https://aclanthology.org/2020.eamt-1.61/
31. Thakur N, Reimers N, Daxenberger J, Gurevych IAS. Augmented SBERT: data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. Presented at: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics; Jun 6-11, 2021. [doi: 10.18653/v1/2021.naacl-main.28]
32. Wang L, et al. Text embeddings by weakly-supervised contrastive pre-training. arXiv. Preprint posted online on Feb 22, 2024 URL: http://arxiv.org/abs/2212.03533 [accessed 2026-03-20]
33. Chen J, Xiao S, Zhang P, Luo K, Lian D, Liu Z. M3-embedding: multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. arXiv. Preprint posted online on 2024. [doi: 10.18653/v1/2024.findings-acl.137]
34. Zhang X, Zhang Y, Long D, et al. MGTE: generalized long-context text representation and reranking models for multilingual text retrieval. arXiv. Preprint posted online on 2024. [doi: 10.18653/v1/2024.emnlp-industry.103]
35. Radlinski F, Craswell N. Comparing the sensitivity of information retrieval metrics. 2010 Jul 19 Presented at: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval; Jul 19-23, 2010. [doi: 10.1145/1835449.1835560]
36. Baysan MS, Uysal S, İşlek İ, Çığ Karaman Ç, Güngör T. LLM-as-a-Judge: automated evaluation of search query parsing using large language models. Front Big Data 2025;8:1611389. [doi: 10.3389/fdata.2025.1611389] [Medline: 40761620]
37. OpenAI. Gpt-oss-120b & gpt-oss-20b model card. arXiv. Preprint posted online on Aug 8, 2025. [doi: 10.48550/arXiv.2508.10925]
38. Si Y, Wang J, Xu H, Roberts K. Enhancing clinical concept extraction with contextual embeddings. J Am Med Inform Assoc 2019 Nov 1;26(11):1297-1304. [doi: 10.1093/jamia/ocz096] [Medline: 31265066]
39. Noh J, Kavuluru R. Improved biomedical word embeddings in the transformer era. J Biomed Inform 2021 Aug;120:103867. [doi: 10.1016/j.jbi.2021.103867] [Medline: 34284119]
40. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature New Biol 2024 Jun;630(8017):625-630. [doi: 10.1038/s41586-024-07421-0] [Medline: 38898292]
41. Miracle embedding model trained on pseudonymized data available on huggingface. URL: https://huggingface.co/SHIPAI-IKIM/miracle-german [accessed 2026-03-09]
42. Github repository for the code developed during the course of the study. URL: https://github.com/UMEssen/MIRACLE [accessed 2026-03-09]

Abbreviations

IAA: interannotator agreement
IR: information retrieval
LLM: large language model
MAP: mean average precision
NDCG: normalized discounted cumulative gain
NLP: natural language processing
RAG: Retrieval Augmented Generation

Edited by A Coristine; submitted 26.Aug.2025; peer-reviewed by Q Zhang, Z Lu; accepted 29.Jan.2026; published 25.Mar.2026.

Please cite as:
Arzideh K, Schäfer H, Idrissi-Yaghir A, Schmidt CS, Eryilmaz B, Bahn M, Turki AT, Pollok OB, Hartmann EM, Winnekens P, Borys K, Haubold J, Nensa F, Hosch R
Improving Retrieval Augmented Generation for Health Care by Fine-Tuning Clinical Embedding Models: Development and Evaluation Study
J Med Internet Res 2026;28:e82997
URL: https://www.jmir.org/2026/1/e82997
doi:10.2196/82997

XSL•FO
RenderX

© Kamyar Arzideh, Henning Schäfer, Ahmad Idrissi-Yaghir, Cynthia Sabrina Schmidt, Bahadir Eryilmaz, Mikel Bahn, Amin Turki, Olivia Barbara Pollok, Eva Maria Hartmann, Philipp Winnekens, Katarzyna Borys, Johannes Haubold, Felix Nensa, René Hosch. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Artificial Intelligence for Predicting Treatment Response in Patients With Anxiety Disorders After Cognitive Behavioral Therapy: Systematic Review and Meta-Analysis

Jiawen Liu[1,2], MA; Junhui Wang[3], MEng; Zhaobin Wu[3], MEng; Mohamad Ibrani Shahrimin Bin Adam Assim[2], PhD

[1]Youth League Committee, Liuzhou Railway Vocational Technical College, No. 2 Wenyuan Road, Yufeng District, Liuzhou, Guangxi Zhuang Autonomous Region, China

[2]Faculty of Humanities, Management and Science, Universiti Putra Malaysia, Bintulu, Sarawak, Malaysia

[3]School of Automation, Guangxi University of Science and Technology, Liuzhou, Guangxi, China

Corresponding Author:
Jiawen Liu, MA
Youth League Committee, Liuzhou Railway Vocational Technical College, No. 2 Wenyuan Road, Yufeng District, Liuzhou, Guangxi Zhuang Autonomous Region, China

Abstract

Background: Artificial intelligence (AI) models have been increasingly explored for predicting treatment response to cognitive behavioral therapy (CBT) in patients with anxiety disorders. Identifying potential responders in advance may help inform treatment planning and support clinical decision-making. Although a growing number of studies have applied AI techniques in this context, reported performance estimates vary across studies, and the overall predictive accuracy has not been comprehensively quantified.

Objective: This systematic review and meta-analysis aims to quantify the overall performance of AI models in predicting treatment response following CBT for anxiety disorders and to examine how data sources, algorithmic approaches, and diagnostic subtypes influence predictive performance.

Methods: A systematic literature search was conducted in PubMed, Embase, Web of Science, Cochrane Library, and PsycINFO up to August 2025. We included studies that validated AI models for predicting CBT treatment response (remission or response) in patients diagnosed with an anxiety disorder. The risk of bias was assessed using the PROBAST+AI (Prediction Model Risk of Bias Assessment Tool for Artificial Intelligence) tool. Predictive performance metrics, including sensitivity, specificity, accuracy, and area under the curve (AUC), were extracted and pooled. Pooled estimates for sensitivity, specificity, and diagnostic accuracy were derived using the Restricted Maximum Likelihood estimator, with CIs adjusted via the Hartung-Knapp-Sidik-Jonkman method. Prediction intervals were calculated and reported alongside these pooled estimates to illustrate the expected distribution of effects in real-world settings.

Results: Eleven studies were included in the meta-analysis. The pooled sensitivity of AI-based models for predicting treatment response was 0.73 (95% CI 0.58 - 0.85; I^2=82.8%), and the pooled specificity was 0.75 (95% CI 0.59 - 0.89; I^2=96.7%). The overall pooled accuracy was 0.74 (95% CI 0.62 - 0.84; I^2=94.6%). The summary AUC was 0.81 (95% CI 0.78 - 0.85), indicating moderate discriminative performance. Subgroup analyses showed that models incorporating multimodal data achieved superior predictive performance, with a pooled sensitivity of 0.84 and an accuracy of 0.82. In addition, predictive performance was the highest in patients with social anxiety disorder compared with other anxiety disorder subtypes.

Conclusions: This meta-analysis quantitatively synthesized AI performance in predicting CBT response for anxiety disorders, moving beyond narrative reviews to provide pooled evidence. In contrast to existing reviews that encompass broader diagnostic groups, our focused approach establishes a precise benchmark for this clinical domain, highlighting the current moderate overall performance. Furthermore, we extend beyond previous work by demonstrating the superior predictive utility of multimodal data, identifying social anxiety disorder as the most predictable subtype, and systematically evaluating the impact of data modalities and algorithm types. Future efforts should prioritize robustly validated multimodal models, laying essential groundwork for the potential development of AI-assisted tools to personalize treatment planning in anxiety disorders.

Trial Registration: PROSPERO CRD420251137096; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251137096

(J Med Internet Res 2026;28:e86079) doi:10.2196/86079

KEYWORDS

artificial intelligence; treatment response; anxiety disorders; cognitive behavioral therapy; meta-analysis

XSL•FO
RenderX

Introduction

Anxiety disorders constitute a group of mental disorders characterized by intense, excessive, and persistent worry and fear, exhibiting high prevalence rates. They not only cause significant functional impairment in patients' social, academic, and occupational functioning but also impose substantial socioeconomic burdens at the societal level, including considerable health care resource consumption and productivity losses [1,2]. Cognitive behavioral therapy (CBT) is a first-line evidence-based intervention for anxiety disorders [3,4]. However, treatment response varies considerably, with a notable subset of patients showing suboptimal or inadequate improvement [5]. This heterogeneity underscores a critical clinical need: the ability to preemptively identify likely responders and, more importantly, those at high risk of nonresponse who may benefit from treatment augmentation or alternative interventions from the outset [6]. Early and accurate prediction of treatment response can optimize care pathways, improve resource allocation, and reduce delays, aligning with the goals of precision mental health [7].

Clinical assessment of anxiety disorders typically relies on structured or semistructured interviews and standardized scales [8], such as the Hamilton Anxiety Rating Scale (HAM-A) [9], Clinical Global Impressions–Improvement (CGI-I) [10], and Liebowitz Social Anxiety Scale (LSAS) [7], which are fundamental for symptom identification, outcome evaluation, and longitudinal monitoring. However, they possess inherent limitations for predictive purposes. First, they depend heavily on self-report and subjective ratings, making them vulnerable to recall bias and social desirability effects, thereby compromising objectivity [11]. Second, heterogeneity in measured constructs, scoring procedures, and cutoff thresholds across instruments constrains comparability and diagnostic consistency across studies and clinical settings [12]. Consequently, there is growing recognition that these traditional metrics alone may be insufficient for reliably forecasting an individual's unique response trajectory to a specific therapy such as CBT [13].

Against this background, artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), has attracted increasing attention as a methodological framework for advancing outcome prediction in mental health research. AI-based models can integrate large-scale, high-dimensional, and multimodal data and capture complex, nonlinear relationships that are difficult to model with conventional statistical techniques [14]. In recent years, a growing number of studies have applied ML and DL methods to predict treatment response across a range of psychiatric interventions, including CBT for anxiety disorders [15,16]. These studies have incorporated diverse data sources, including baseline clinical characteristics, symptom trajectories, neuroimaging markers, psychophysiological measures, and digital behavioral indicators, and have reported varying levels of predictive accuracy [17,18]. However, the anxiety disorders comprise clinically distinct subtypes, such as social anxiety disorder (SAD), generalized anxiety disorder (GAD), and panic disorder (PD), which differ in symptom profiles and underlying mechanisms [19]. This diagnostic heterogeneity may further influence model generalizability and predictive stability across populations [7,20]. To date, however, quantitative evidence synthesizing the predictive performance of AI models for CBT response in different anxiety disorder subtypes remains limited, and potential sources of heterogeneity across studies remain insufficiently explored and require systematic investigation [21].

This systematic review and meta-analysis aims to quantitatively synthesize the overall performance of AI models in predicting treatment response following CBT for anxiety disorders. We also seek to explore the influence of key factors, such as data sources, algorithmic approaches, and diagnostic subtypes, on predictive performance and to preliminarily investigate potential sources of the observed heterogeneity.

Methods

Ethical Considerations

As this is a systematic review and meta-analysis, ethics approval and consent to participate are not applicable. The manuscript does not include the identification images or other personal or clinical details of participants.

The meta-analysis adhered rigorously to the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy) reporting guidelines [22] and the PRISMA-S extension for reporting literature searches [23]. We have provided the completed PRISMA-DTA (Checklist 1), PRISMA 2020 Abstract (Checklist 2), and PRISMA-S (Checklist 3) checklists. The research protocol was registered in the PROSPERO registry (registration ID: CRD420251137096).

Feasibility Assessment for Meta-Analysis

Prior to initiating the full systematic review, a preliminary scoping search was conducted to assess the feasibility of quantitative synthesis. This step aimed to determine whether sufficient studies reported the extractable predictive performance metrics necessary for a meta-analysis. In the scoping search, although the field is nascent, several published studies met the core data requirements, supporting the rationale for proceeding with a formal systematic review and meta-analysis.

Search Strategy

A systematic search was conducted in PubMed, Embase, Web of Science, Cochrane Library, and PsycINFO from inception to August 16, 2025. Two independent reviewers (JL and JW) screened titles/abstracts and subsequently assessed full texts, with disagreements resolved by consensus. Search strategies combined controlled vocabulary (MeSH) and free-text terms across 3 concept blocks: AI and methods (eg, "artificial intelligence," "machine learning," and "deep learning"), anxiety disorders (eg, "anxiety disorders," "social anxiety disorder," "generalized anxiety disorder," and "panic disorder"), and CBT (eg, "cognitive behavioral therapy," "CBT," and "cognitive therapy"). No language or publication year limits were applied at the initial search stage to minimize selection bias. Detailed database-specific strategies are provided in Table S1 in Multimedia Appendix 1. Study selection followed a dual-stage

XSL•FO
RenderX

independent screening process. Two reviewers (JL and JW) first screened titles/abstracts and then independently assessed full texts of potentially eligible records. Disagreements were resolved by consensus or by third-party arbitration (ZW). The reference lists of included studies were manually screened for additional records. It should be noted that in conducting this systematic review, we did not contact authors, experts, or manufacturers to obtain additional studies or data, and did not adapt or reuse search strategies from previous reviews. Additionally, a formal peer review of the search strategy was not conducted.

Inclusion and Exclusion Criteria

The PITROS framework was used for the inclusion criteria:

- Participants (P): patients with anxiety disorders receiving CBT, including separation anxiety disorder, selective mutism, specific phobia, SAD, PD/panic attacks, agoraphobia, GAD, and substance/medication-induced anxiety disorder
- Index test (I): AI methods (eg, ML and DL) developed to predict CBT treatment response
- Target condition (T): positive cases defined as clinically significant post-CBT improvement (eg, meaningful symptom reduction); negative cases defined as poor or no response
- Reference standard (R): validated diagnostic/severity scales for anxiety disorders
- Outcomes (O): sensitivity, specificity, area under the curve (AUC), and accuracy
- Setting (S): mental health outpatient clinics, psychiatric hospitals, and clinical trial centers; multicenter prospective or retrospective designs using electronic health records or public databases

The exclusion criteria were as follows: (1) publication in a language other than English at the full-text stage, (2) nonempirical article type (eg, review, case report, conference abstract, meta-analysis, and letter), (3) no assessment of the prediction of CBT response using AI, (4) inclusion of a population that did not have anxiety disorder or that did not undergo CBT, (5) absence of a recognized reference standard or inadequate data to derive the target metrics, and (6) duplicate publication identified using EndNote and manual checking. The database search was unrestricted by language to maximize sensitivity, but non-English publications were excluded during full-text review owing to translation resource constraints. We acknowledge this as a potential limitation, as it may have introduced language bias by excluding relevant studies published in other languages. The independent selection of studies was conducted by 2 reviewers (JL and JW), with disagreements resolved through discussion or by a third reviewer (ZW).

Data Extraction

Two independent reviewers (JL and JW) conducted comprehensive data extraction from full-text articles to assess study eligibility. The extracted data encompassed critical study characteristics, including the authors, year, country, study design, center, response definition, types of disorders, and types of treatments. Specific quantitative elements included the total number of patients in the training and internal validation sets, the number of response patients in the training and internal validation sets, the data-splitting method, predictors, AI algorithms, and AI methods. For internal validation sets, we systematically recorded true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values.

For studies included in the systematic review but lacking meta-analysis–compatible data, we proactively contacted corresponding authors via email to request necessary information. Given the frequent absence of diagnostic contingency tables, we reconstructed 2×2 tables using reported sensitivity, specificity, responder numbers, and total patient counts. Any discrepancies in extraction were resolved through a collaborative discussion, with a third reviewer (ZW) serving as an arbitrator if consensus could not be reached, thereby ensuring methodological rigor and data integrity.

Quality Assessment

The latest PROBAST+AI (Prediction Model Risk of Bias Assessment Tool for Artificial Intelligence) quality assessment tool was used [24], which replaced PROBAST 2019. This tool comprises 2 phases: model development and model evaluation. Each phase encompasses 7 domains covering participant and data sources, predictors, outcome assessment, and analysis. Evaluation results for each domain are categorized as “low,” “high,” or “unclear,” determined based on specific signal questions. Signal concerns are rated as “yes,” “probably yes,” “probably no,” “no,” “no information,” and, in some cases, “not applicable.” Signal concerns phrased as “yes” or “probably yes” indicate a lower risk of bias. Any signal question rated “no” or “probably no” indicates a potential high risk of bias in that domain. If no “no” or “probably no” rating exists but “no information” is present, the domain’s bias risk is classified as unclear. The complete list of signal questions and tables is detailed in Tables S2 and S3 in Multimedia Appendix 1.

To ensure objectivity and accuracy in the assessment process, 2 reviewers (JL and JW) independently assessed the risk of bias in the included studies using the PROBAST+AI quality assessment tool. During the review process, any discrepancies between reviewers were resolved through in-depth discussions and analyses to achieve consensus, ensuring highly reliable, consistent final assessment results.

Outcome Measures

The primary outcome measures encompassed comprehensive diagnostic performance metrics from internal validation sets, including sensitivity, specificity, accuracy, and AUC. Sensitivity, calculated as TP/(TP+FN), measured the AI model’s ability to accurately identify TPs, while specificity, computed as TN/(TN+FP), assessed the model’s proficiency in correctly identifying negative cases. AUC provided a holistic assessment of the model’s discriminative power between positive and negative instances. Accuracy, defined as (TP+TN)/total patients, represented the proportion of correctly classified samples. We systematically extracted performance metrics from all data sources (eg, imaging, clinical/demographic, and multimodal) and AI algorithms presented in each study. However, for the purpose of meta-analytic pooling, only the primary model from

XSL•FO
RenderX

each article, specifically the model based on nonoverlapping patient cohorts, was selected to ensure data independence and avoid double-counting.

Statistical Analysis

While the core analytical plan and reporting adhered to the preregistered protocol, a key methodological deviation was implemented to enhance the robustness of our statistical synthesis. The registered protocol specified the use of the bivariate random-effects model (Reitsma model) as the primary method for pooling sensitivity and specificity. In the final analysis, we maintained this model for generating the summary receiver operating characteristic (SROC) curve and the AUC. However, for the pooled estimates of sensitivity and specificity themselves, we adopted a univariate random-effects model using the Restricted Maximum Likelihood estimator, with CIs adjusted via the Hartung-Knapp-Sidik-Jonkman (HKSJ) method. This adaptation was made to incorporate the HKSJ adjustment, which provides more conservative and reliable CIs, particularly when dealing with a limited number of studies or substantial heterogeneity [25]. Prediction intervals (PIs) were calculated and reported alongside these pooled estimates to illustrate the expected distribution of effects in real-world settings. Heterogeneity was quantitatively assessed using the PIs and further explored through subgroup analyses and meta-regression [26]. Bivariate boxplots were used to visually identify potential outlier studies contributing to heterogeneity. Prespecified covariates for meta-regression included the number of study centers, anxiety disorder subtype, data splitting method, CBT protocol type, and AI algorithm category. Subgroup analyses were conducted according to data type, algorithm type, and specific anxiety disorder diagnoses. Additionally, a bubble plot was used to assess temporal trends in the accuracy of different algorithms, and funnel plots were used, along with Egger tests, to evaluate potential small-study effects.

Results

Study Selection

A search across 5 electronic databases yielded 1399 potentially relevant publications. After deleting 399 duplicates, 1000 records were retained for preliminary screening. At this phase, 980 studies were identified as clearly irrelevant based on the title/abstract and the publication type. Thus, 15 articles were selected for full-text review. Following a comprehensive assessment of the full texts, 5 studies were removed because they did not investigate treatment response to CBT. Two studies were excluded due to non-English publication. Three studies were excluded as they were ongoing trials without any published outcome measures. Additionally, 6 qualifying studies were identified from nondatabase sources. Ultimately, 11 studies [7,17,18,27-34] were included in the meta-analysis as they met predefined inclusion criteria. The identification, screening, eligibility assessment, and inclusion process followed the PRISMA guidelines, and the selection analysis is elaborated in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) s in Figure 1: flow diagram of the systematic literature search and study selection process for studies evaluating artificial intelligence models in predicting treatment response to cognitive behavioral therapy for anxiety disorders.

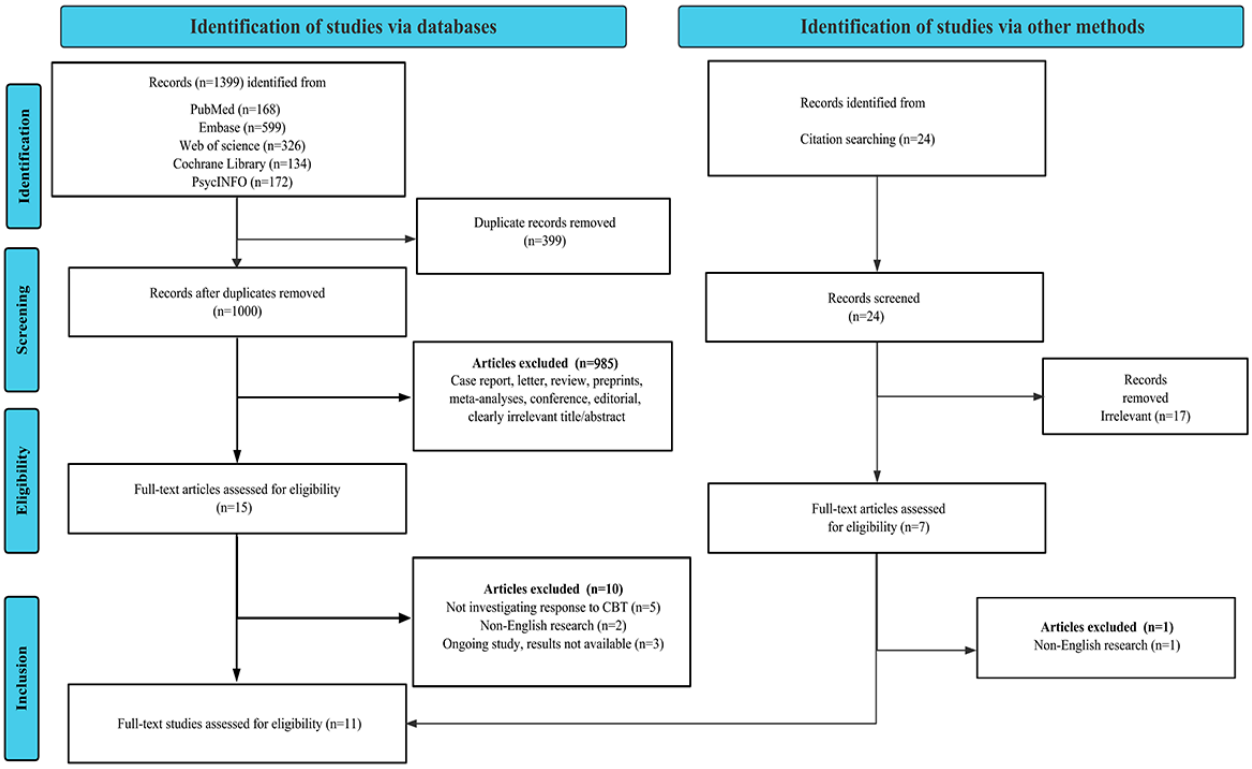

XSL•FO
RenderX

Study Description and Quality Assessment

A total of 11 eligible studies comprising 59,085 samples (training: 38,563; internal validation set: 20,522) were included. Among these, 4 studies exclusively evaluated SAD [7,29,32,34], 2 studies focused on PD with agoraphobia [30,33], 1 study investigated both GAD and PD [27], 1 study examined PD and SAD [31], and the remaining 3 studies addressed multiple anxiety disorder subtypes [17,18,28]. Nine studies [17,18,27,29-34] provided data eligible for meta-analysis. The included studies were published between 2014 and 2024, and all were prospective. Definitions of treatment response varied across studies: 2 studies used the CGI-I [29,32], 2 used the LSAS [7,34], 2 used the HAM-A [30,33], 1 used the Overall Anxiety Severity and Impairment Scale (OASIS) [27], 1 used the clinical severity rating [28], 1 combined HAM-A and the Spider Phobia Questionnaire (SPQ) [18], 1 used both the Generalized Anxiety Disorder-7 (GAD-7) and the Patient Health Questionnaire-9 (PHQ-9) [17], and 1 applied multiple scales simultaneously [31]. Across the 9 datasets, the most frequently used AI algorithms were support vector machines (SVMs; 4/9, 44%) and random forests (RFs; 3/9, 33%). Additionally, of the 9 datasets, 2 were based on clinical/demographic data, 6 on imaging data, and 1 on multimodal data. The clinical, methodological, and technical characteristics of the included studies are summarized in Table 1 and Multimedia Appendix 2.

XSL•FO
RenderX

Table . Characteristics of the included studies (n=11) and their patient populations.

Study (author)	Year	Country	Study design	Center	Response definition	Types of anxiety disorders	Types of treatments	Data source	Total sample size, n		Response sample size, n	
									Training	Internal validation	Training	Internal validation
Ball et al [27]	2014	America	Pro[a]	Single center	OASIS[b] ≤5	GAD[c], PD[d]	Standard CBT[e]	Institutional database	48	49	29	49
Bertie et al [28]	2024	Multiple countries	Pro	Multiple centers	CSR[f] ≥4	GAD, separation anxiety, SAD[g], SPH[h]	Standard CBT	Institutional database	2214	2214	—[i]	—
Bukhari et al [7]	2025	America	Pro	Multiple centers	LSAS[j]↓ ≥50%	SAD	Standard CBT	Institutional database	157	157	—	—
Frick et al [29]	2020	Sweden	Pro	Single center	CGI-I[k] ≤2	SAD	CBT+SSRI[l], CBT+placebo	Institutional database	47	47	24	24
Hahn et al [30]	2015	Germany	Pro	Multiple centers	HAM-A[m]↓ ≥50%	PD with AG[n]	Standard CBT	Institutional database	49	49	25	25
Hilbert et al [18]	2024	Germany	Pro	Multiple centers	Protect-AD: HAM-A↓ ≥50%; Spider VR: SPQ[o]↓ ≥30%	PD, AG, SAD, SPH	Standard CBT	Institutional database	410	410	213	213
Hentati Isacsson et al [31]	2024	Sweden	Pro	Single center	MADRS[p] ≤11, PDSS[q] ≤8, LSAS ≤35	PD, SAD	iCBT[r]	Institutional database	3619	3619	—	—
Månsson et al [32]	2015	Sweden	Pro	Multiple centers	CGI-I ≤2	SAD	iCBT	Institutional database	23	23	12	12
Prasad et al [17]	2023	United Kingdom, America	Pro	Multiple centers	GAD-7[s]↓ ≥64, PHQ-9[t]↑ <6	Anxiety	iCBT	Institutional database	31,899	13,857	—	5232
Sundermann et al [33]	2017	Germany	Pro	Multiple centers	HAM-A↓ ≥50%	PD with AG	Standard CBT	Institutional database	59	59	30	30
Whitfield-Gabrieli et al [34]	2015	America	Pro	Multiple centers	LSAS↓ ≥50%	SAD	Standard CBT	Institutional database	38	38	19	19

[a]Pro: prospective.

[b]OASIS: Overall Anxiety Severity and Impairment Scale.

[c]GAD: generalized anxiety disorder.

XSL•FO
RenderX

[d]PD: panic disorder.
[e]CBT: cognitive behavioral therapy.
[f]CSR: clinical severity rating.
[g]SAD: social anxiety disorder.
[h]SPH: specific phobia.
[i]Not applicable.
[j]LASA: Liebowitz Social Anxiety Scale.
[k]CGI-I: clinical global impression-improvement.
[l]SSRI: selective serotonin reuptake inhibitor.
[m]HAM-A: Hamilton Anxiety Rating Scale.
[n]AG: agoraphobia.
[o]SPQ: Spider Phobia Questionnaire.
[p]MADRS: Montgomery-Asberg Depression Rating Scale.
[q]PDSS: Panic Disorder Severity Scale.
[r]iCBT: internet-based cognitive behavioral therapy.
[s]GAD-7: Generalized Anxiety Disorder-7.
[t]PHQ-9: Patient Health Questionnaire-9.

The risk of bias, as assessed by the PROBAST+AI quality assessment tool, is presented in Figure 2 and Tables S2 and S3 in Multimedia Appendix 1. For model development, in the overall quality assessment, 55% (6/11) of studies were rated as high risk, while the remaining 45% (5/11) were rated as low risk. Regarding applicability concerns, 9% (1/11) of studies were rated as high risk, while the remaining 91% (10/11) were rated as low risk. For model validation, 55% (6/11) of studies were rated as high risk in the overall risk of bias assessment, with the remaining 45% (5/11) rated as low risk. In the comprehensive evaluation of applicability concerns, 9% (1/11) of studies were rated as high risk, while the remaining 91% (10/11) were rated as low risk.

XSL•FO
RenderX

Figure 2. Risk of bias and applicability assessment using the PROBAST+AI tool for studies developing or validating AI models to predict treatment response after cognitive behavioral therapy in patients with anxiety disorders. Panel A summarizes the risk of bias for model development studies, and panel B summarizes the risk of bias for model validation studies. PROBAST+AI: Prediction Model Risk of Bias Assessment Tool for Artificial Intelligence.

XSL•FO
RenderX

Trend in Diagnostic Accuracy Over Time

The scatter plot in Figure 3 illustrates a negative trend in model accuracy from 2014 to 2024 across logistic regression (LR), recurrent neural network (RNN), RF, and SVM algorithms. A linear regression line confirms that reported accuracy has generally decreased over the decade within this dataset. The highest performance was recorded in 2015 by an SVM model (approximately 0.9), while the lowest scores appeared in 2024, with RF models showing a drop in scores to around 0.5. The shaded region indicates the CI for this downward trajectory (Figure 3).

Figure 3. Bubble plot illustrating the accuracy of various artificial intelligence algorithms over time (publication year) for predicting treatment response to cognitive behavioral therapy in patients with anxiety disorders. LR: logistic regression; RF: random forest; RNN: recurrent neural network; SVM: support vector machine.

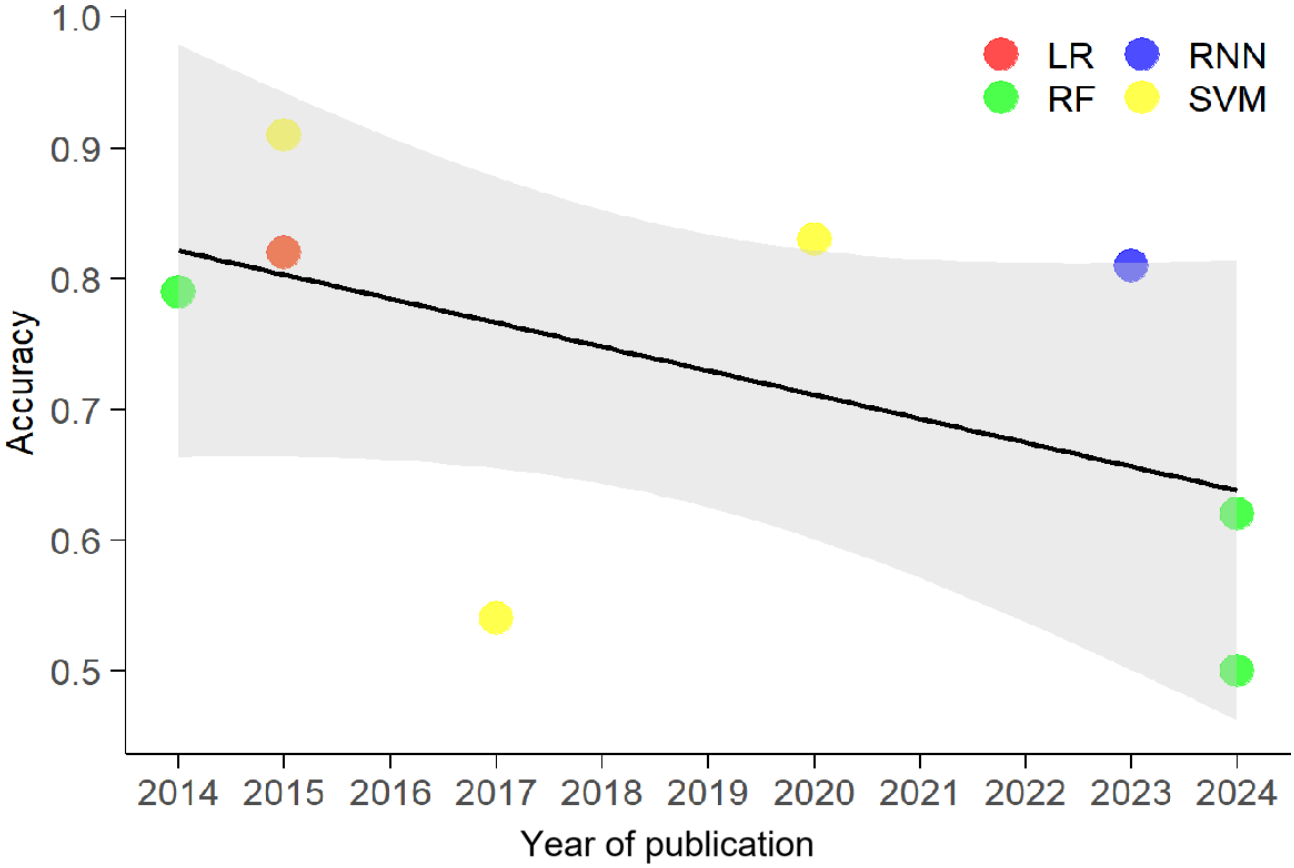

Predicting CBT Response to Anxiety Disorders Using AI Models Based on the Primary Model Per Study

The pooled sensitivity of AI models was 0.73 (95% CI 0.58 - 0.85; 95% PI 0.31 - 0.99; I^2=82.8%), specificity was 0.75 (95% CI 0.59 - 0.89; 95% PI 0.25 - 1.00; I^2=96.7%), accuracy was 0.74 (95% CI 0.62 - 0.84; 95% PI 0.39 - 0.97; I^2=94.6%), and AUC was 0.81 (95% CI 0.78 - 0.85), as shown in Figures 4-7. Based on a prespecified pretest probability of 20%, the Fagan nomogram indicated a posttest probability of 43% for a positive result and 8% for a negative result (Figure S1 in Multimedia Appendix 1). The results of the bivariate boxplots indicated that 1 study [30] had discrete values, suggesting it may be a source of heterogeneity (Figure 8). Furthermore, meta-regression analysis indicated that heterogeneity in sensitivity was significantly associated with the number of centers (P<.001) and the type of anxiety disorder (P<.001). In contrast, heterogeneity in specificity was significantly associated with CBT type (P<.001) and AI algorithm category (P<.001) (Table 2).

XSL•FO
RenderX

Figure 4. Forest plot displaying the pooled sensitivity, with 95% CIs and prediction intervals, of artificial intelligence models for predicting treatment response after cognitive behavioral therapy in patients with anxiety disorders [17,18,27,29,30,32-34]. FN: false negative; TP: true positive.

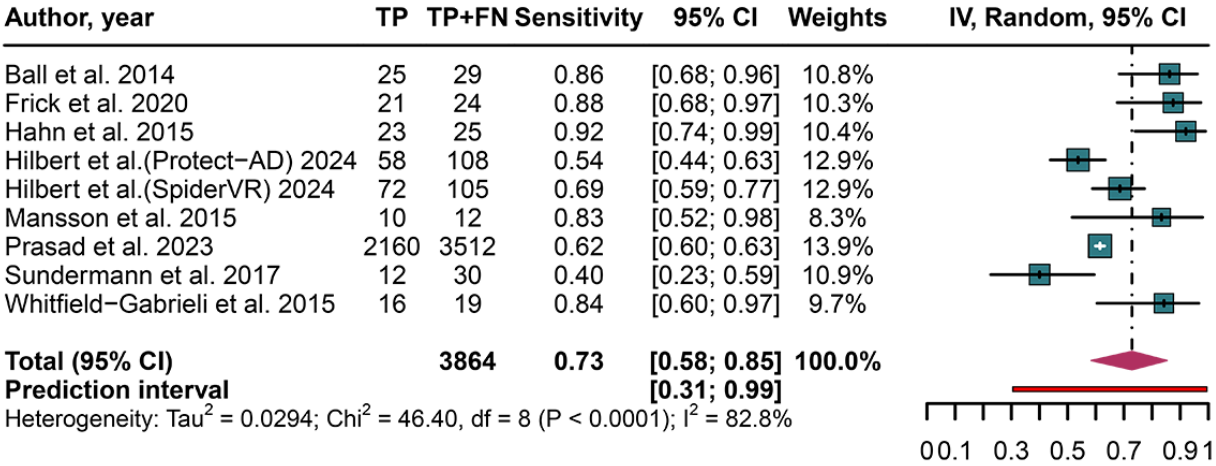

Figure 5. Forest plot displaying the pooled specificity, with 95% CIs and prediction intervals, of artificial intelligence models for predicting treatment response after cognitive behavioral therapy in patients with anxiety disorders [17,18,27,29,30,32-34]. FP: false positive; TN: true negative.

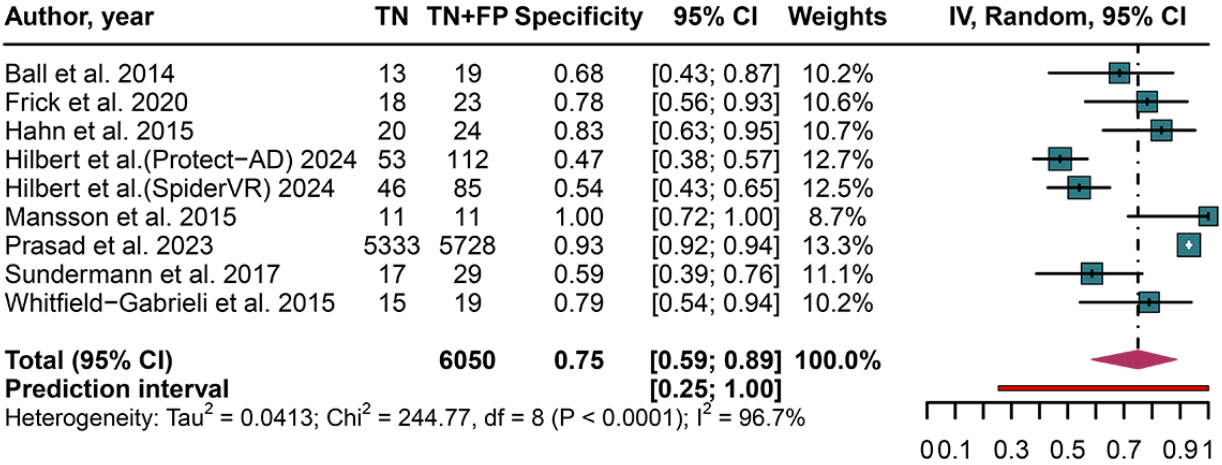

Figure 6. Forest plot displaying the pooled accuracy, with 95% CIs and prediction intervals, of artificial intelligence models for predicting treatment response after cognitive behavioral therapy in patients with anxiety disorders [17,18,27,29,30,32-34]. TN: true negative; TP: true positive.

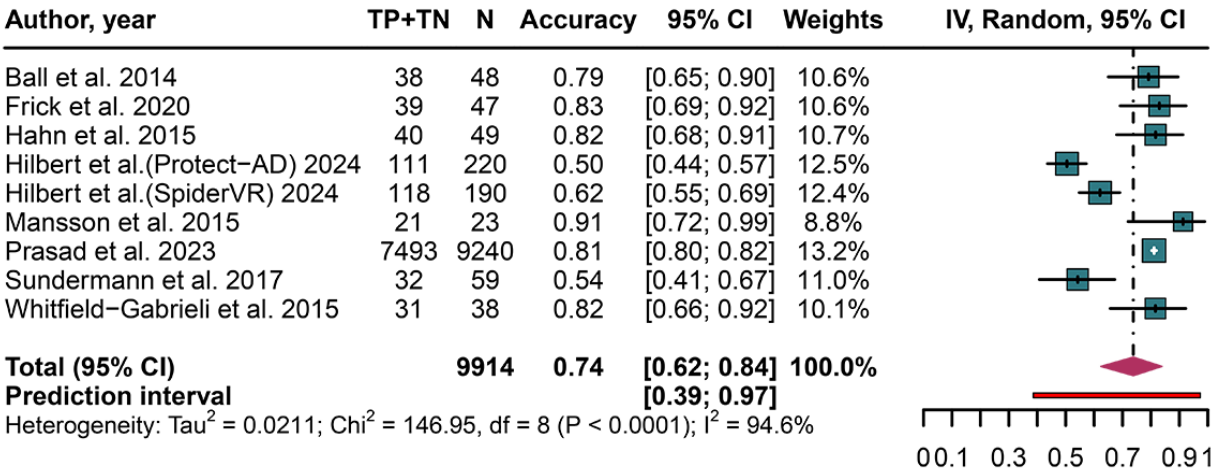

XSL•FO
RenderX

Figure 7. Summary receiver operating characteristic (SROC) curve for artificial intelligence models predicting treatment response in patients with anxiety disorders after cognitive behavioral therapy. The summary point estimates for sensitivity (SENS) and specificity (SPEC), and the area under the curve (AUC) are shown.

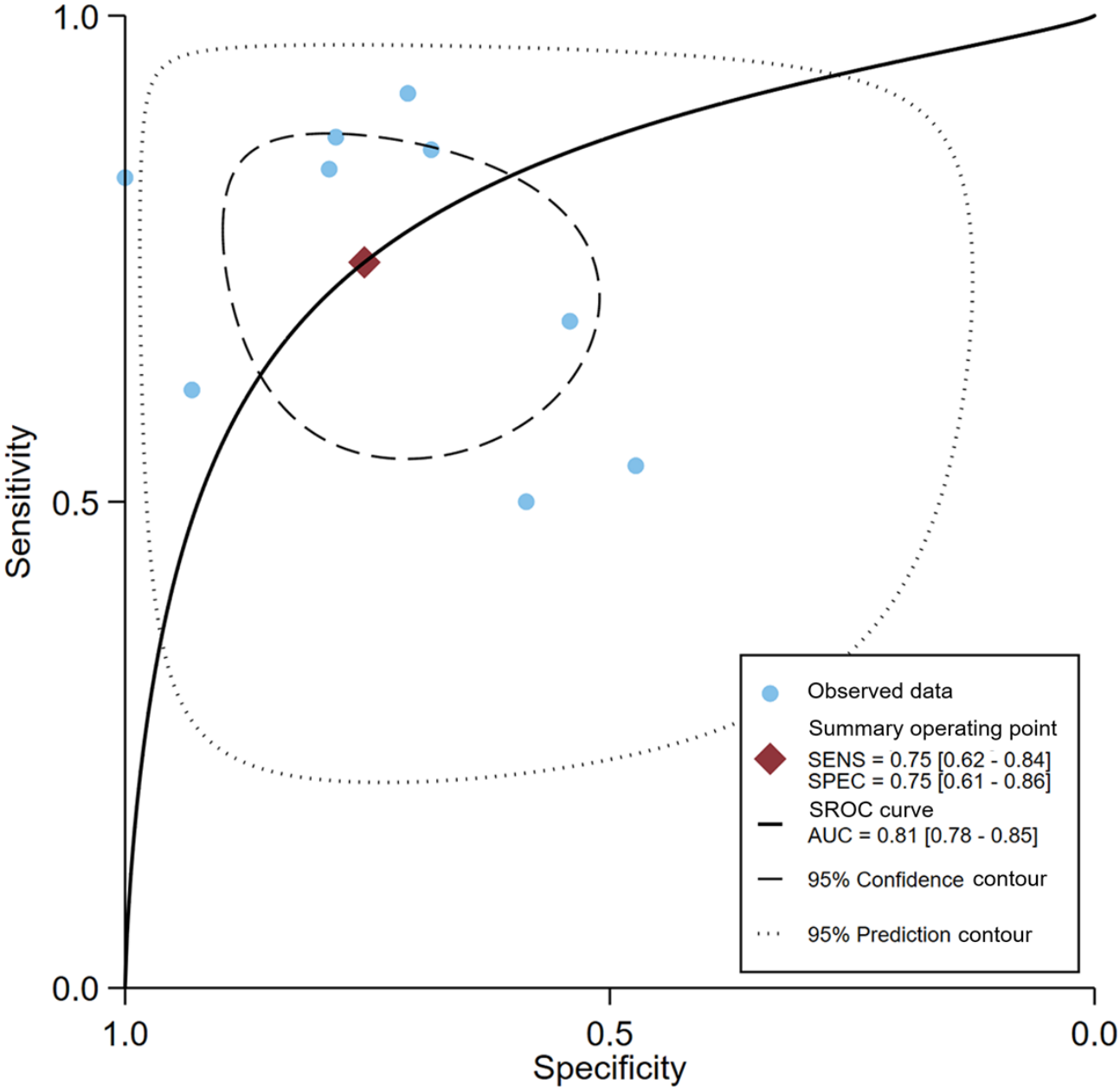

XSL•FO
RenderX

Figure 8. Bivariate boxplot of the combined logit-transformed sensitivity (SENS) and specificity (SPEC) estimates from the bivariate random-effects meta-analysis for predicting treatment response in patients with anxiety disorders after cognitive behavioral therapy. Squares represent individual study estimates; the box illustrates the joint distribution.

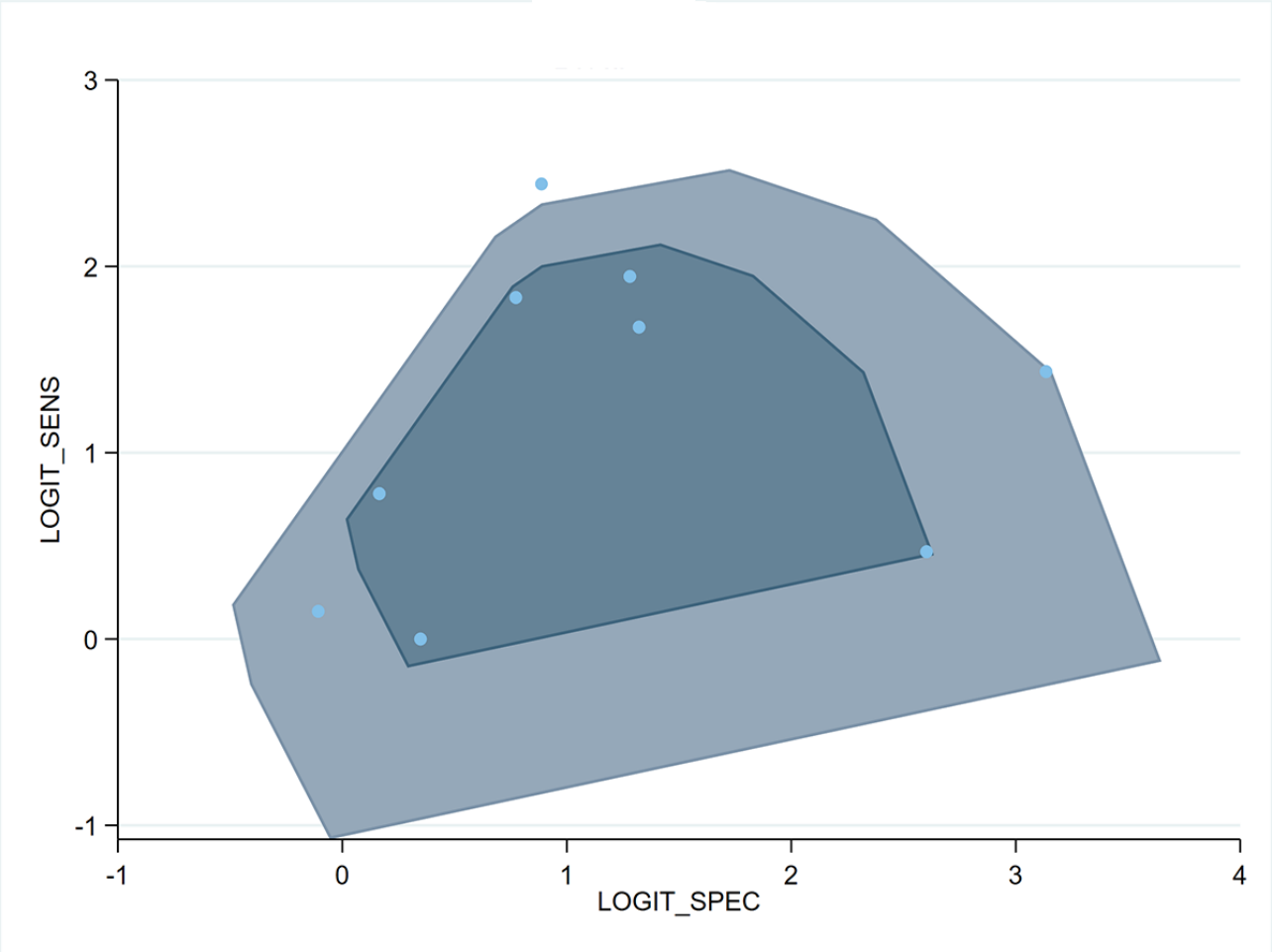

XSL•FO
RenderX

Table . Results of the univariable meta-regression analysis examining potential sources of heterogeneity in the performance of artificial intelligence models for predicting cognitive behavioral therapy treatment response in anxiety disorders.

Subgroup	Number of contingency tables	Sensitivity, value (95% CI)	Meta-regression *P* value	Specificity, value (95% CI)	Meta-regression *P* value
Number of centers			.02		.89
Single center	2	0.87 (0.76 - 0.99)		0.74 (0.46 - 1.00)	
Multiple centers	7	0.69 (0.58 - 0.80)		0.76 (0.61 - 0.90)	
Type of anxiety disorder			.01		.10
SAD[a]	3	0.86 (0.74 - 0.98)		0.87 (0.71 - 1.00)	
Others	6	0.68 (0.57 - 0.79)		0.69 (0.54 - 0.85)	
Data splitting method			.23		.28
LOOCV[b]	4	0.78 (0.62 - 0.93)		0.81 (0.63 - 0.98)	
Independent validation	5	0.72 (0.58 - 0.86)		0.71 (0.54 - 0.89)	
CBT[c] type			.77		<.001
iCBT[d]	2	0.71 (0.48 - 0.94)		0.95 (0.91 - 0.98)	
Standard CBT	6	0.74 (0.62 - 0.87)		0.61 (0.51 - 0.71)	
AI[e] algorithms			.51		<.001
DL[f]	1	0.62 (0.29 - 0.94)		0.93 (0.87 - 1.00)	
ML[g]	8	0.78 (0.68 - 0.88)		0.67 (0.57 - 0.78)	

[a]SAD: social anxiety disorder.
[b]LOOCV: leave-one-out cross-validation.
[c]CBT: cognitive behavioral therapy.
[d]iCBT: internet-based cognitive behavioral therapy.
[e]AI: artificial intelligence.
[f]DL: deep learning.
[g]ML: machine learning.

Subgroup Analysis

Within the data-type subgroup, multimodal AI models appeared to achieve higher pooled sensitivity (0.84, 95% CI 0.60 - 0.97) and accuracy (0.82, 95% CI 0.66 - 0.92) compared with single-modal approaches. In contrast, models based on either clinical/demographic or imaging data alone demonstrated comparatively lower sensitivity and accuracy. Regarding specificity, models developed using clinical/demographic data showed relatively higher estimates (0.81, 95% CI 0.55 - 0.93). Regarding AUC, only 1 study reported data within this subgroup, and in that study, imaging models achieved an AUC of 0.79 (95% CI 0.76 - 0.83) (Table 3).

XSL•FO
RenderX

Table . Subgroup meta-analysis of the pooled predictive performance of artificial intelligence models for predicting cognitive behavioral therapy response, stratified by input data type, algorithm type, and anxiety disorder classification.

Subgroup	Number of contingency tables	Sensitivity, value (95% CI)	Specificity, value (95% CI)	AUC[a], value (95% CI)	Accuracy, value (95% CI)
Type of data					
Clinical/demographic	2	0.65 (0.42 - 0.82)	0.81 (0.55 - 0.93)	—[b]	0.73 (0.55 - 0.89)
Imaging	6	0.77 (0.63 - 0.87)	0.71 (0.53 - 0.85)	0.79 (0.76 - 0.83)	0.72 (0.00 - 1.00)
Multimodal	1	0.84 (0.60 - 0.97)	0.79 (0.54 - 0.94)	—	0.82 (0.66 - 0.92)
AI[c] algorithms					
RF[d]	3	0.64 (0.58 - 0.70)	0.52 (0.45 - 0.59)	0.54 (0.50 - 0.59)	0.63 (0.27 - 0.93)
SVM[e]	4	0.82 (0.61 - 0.93)	0.75 (0.60 - 0.85)	0.83 (0.80 - 0.86)	0.78 (0.50 - 0.97)
RNN[f]	1	0.62 (0.60 - 0.63)	0.93 (0.92 - 0.94)	—	0.81 (0.81 - 0.82)
LR[g]	1	0.84 (0.60 - 0.97)	0.79 (0.54 - 0.94)	—	0.82 (0.66 - 0.92)
Type of anxiety disorder					
GAD[h] and PD[i]	1	0.86 (0.68 - 0.96)	0.68 (0.43 - 0.87)	—	0.79 (0.65 - 0.90)
SAD[j]	3	0.86 (0.78 - 0.92)	0.87 (0.41 - 1.00)	0.92 (0.85 - 0.98)	0.85 (0.72 - 0.94)
PD with AG[k]	2	0.69 (0.00 - 1.00)	0.71 (0.00 - 1.00)	—	0.69 (0.00 - 1.00)
Multiple anxiety disorders	3	0.61 (0.45 - 0.77)	0.68 (0.05 - 1.00)	0.65 (0.49 - 0.81)	0.65 (0.25 - 0.96)

[a]AUC: area under curve.
[b]Not applicable.
[c]AI: artificial intelligence.
[d]RF: random forest.
[e]SVM: support vector machine.
[f]RNN: recurrent neural network.
[g]LR: logistic regression.
[h]GAD: generalized anxiety disorder.
[i]PD: panic disorder.
[j]SAD: social anxiety disorder.
[k]AG: agoraphobia.

Within the AI algorithm subgroup, different algorithms demonstrated distinct advantages across various performance metrics. For sensitivity, LR performed best, achieving a value of 0.84 (95% CI 0.60 - 0.97). Similarly, in terms of accuracy, LR again showed optimal performance, reaching a value of 0.82 (95% CI 0.66 - 0.92). For specificity, the RNN achieved the highest value of 0.93 (95% CI 0.87 - 0.98). Regarding the AUC, only 2 algorithms produced results that could be combined, with the SVM algorithm yielding a relatively high combined AUC of 0.83 (95% CI 0.80 - 0.86).

Within the type of anxiety disorder subgroup, predictive performance varied across outcome metrics. For sensitivity, GAD and PD demonstrated relatively higher pooled estimates at 0.86 (95% CI 0.68 - 0.96). In contrast, SAD showed comparatively stronger performance in specificity and accuracy, with a pooled specificity of 0.87 (95% CI 0.41 - 1.00) and accuracy of 0.85 (95% CI 0.72 - 0.94). For discrimination, as measured by AUC, pooled estimates were available for 2 disorder types, among which SAD demonstrated a comparatively higher AUC of 0.92 (95% CI 0.85 - 0.98).

Publication Bias

Funnel plots and Egger regression tests were conducted to evaluate potential small-study effects (Figures S2-S4 in Multimedia Appendix 1). No statistically significant evidence of small-study effects was observed for sensitivity (t_7=1.75; P=.12; intercept=1.54; SE=0.88) and accuracy (t_7=−1.51; P=.17; intercept=−2.40; SE=1.59). In contrast, the test for specificity yielded a statistically significant result (t_7=−2.58; P=.04; intercept=−4.29; SE=1.66), suggesting the possible presence of small-study effects.

XSL•FO
RenderX

Discussion

Principal Findings

The systematic review and meta-analysis revealed that the model demonstrated moderate predictive performance in forecasting treatment response to CBT among patients with anxiety disorders, with sensitivity, specificity, accuracy, and AUC values of 0.73, 0.75, 0.74, and 0.81, respectively. This moderate predictive capability reflects the combined influence of multiple complex factors. First, limitations in input variables constitute a key constraint: current models primarily rely on demographic data, psychiatric history, and scale measurements, making it challenging to capture deeper biological and environmental factors [7,35]. While biomarkers, such as neuroimaging, could theoretically significantly enhance predictive accuracy, their clinical application remains constrained by cost and accessibility [32,35]. Second, data heterogeneity and sample size issues cannot be overlooked. Significant variations among patient cohorts in symptom severity, comorbid conditions, and treatment modalities diminish the model's generalizability [18,36]. Small-sample studies often overestimate predictive capability, while large-scale multicenter studies frequently yield reduced accuracy, sometimes approaching random levels [18,35].

Furthermore, limitations exist in variable selection and modeling approaches: prediction capabilities relying solely on traditional questionnaires and demographic data are constrained [35,37]. While integrating multimodal data (eg, neuroimaging, genetic, and behavioral data) may enhance performance, the comprehensive collection of such data remains challenging in real-world clinical settings [18]. Finally, challenges in practical application and external validation cannot be overlooked; many models lack independent external validation, resulting in significantly reduced accuracy and interpretability in clinical settings [38]. In summary, while current models demonstrate predictive potential, future research requires larger-scale, multicenter, and multimodal studies to enhance model robustness and clinical translation capabilities.

The subgroup analysis suggested that multimodal data integration was associated with comparatively higher sensitivity (0.84) and accuracy (0.82) than single-modal approaches. Across algorithm categories, no single model consistently outperformed others; LR showed higher sensitivity (0.84) and accuracy (0.82), RNN demonstrated higher specificity (0.93), and SVM had a comparatively higher AUC of 0.83. Predictive performance also varied across anxiety disorder subtypes. GAD and PD showed comparatively higher sensitivity (0.86), whereas SAD demonstrated higher specificity (0.87), accuracy (0.85), and AUC (0.92). Together, these findings suggest that optimal prediction may depend less on a single "best" algorithm and more on strategically matching the model architecture to the data type and clinical question [39]. Future work should therefore focus not only on multimodal data collection but also on developing tailored, interpretable modeling approaches that can be validated in real-world clinical settings [40].

The systematic review revealed that multimodal data demonstrated the most promising performance in predicting CBT treatment response for anxiety disorders, with a sensitivity of 0.84 and an accuracy of 0.82. The superior predictive capability of multimodal approaches could be attributed to several key mechanisms. First, multimodal data integration enabled the comprehensive capture of the multidimensional nature of anxiety disorders by combining neuroimaging, physiological signals, behavioral data, and clinical assessments [29,30]. Complementary information across modalities enabled AI models to detect nuanced individual differences and potential mechanisms of treatment response [27,34]. Research indicates that, compared with single-modality approaches, multimodal fusion significantly enhances predictive accuracy, particularly when incorporating models that integrate functional and structural brain connectivity [29,34]. Moreover, multimodal approaches supported personalized, dynamic prediction models that can identify high-risk nonresponders early, potentially facilitating adaptive treatment strategies [17,27]. While these findings are promising, careful validation through larger, cross-validated studies is essential to establish the clinical utility and generalizability of such AI-driven prediction models [29,33].

The meta-analysis revealed differential performance patterns across ML algorithms in predicting CBT treatment response for anxiety disorders. LR demonstrated comparatively higher sensitivity (0.84) and accuracy (0.82), potentially attributable to its ability to capture linear relationships between clinical features and treatment outcomes [7,31]. The algorithm's strength lies in its ability to model structured clinical data with minimal risk of overfitting, particularly in scenarios with balanced sample distributions [7,17]. Conversely, RNNs exhibited superior specificity (0.93), likely due to their ability to process high-dimensional, sequential data and capture complex nonlinear interactions [17]. RNN's advanced modeling approach enables more nuanced identification of nonresponders by leveraging temporal patterns in clinical measurements and platform interaction behaviors [17]. For overall discrimination, SVM achieved a pooled AUC of 0.83 (95% CI 0.80 - 0.86). Given their kernel-based formulation, SVMs may perform well with moderate sample sizes and complex feature spaces, allowing flexible decision boundaries [7,31]. Taken together, these findings suggest that algorithmic performance may vary according to data structure and modeling objectives. Rather than indicating a universally superior approach, the results support the importance of aligning model architecture with data characteristics and specific clinical prediction goals [31].

The subgroup analysis indicated that predictive performance varied across anxiety disorder subtypes. GAD and PD demonstrated comparatively higher sensitivity estimates (0.86, 95% CI 0.68 - 0.96), suggesting relatively stronger identification of treatment responders in these populations. In contrast, SAD showed higher pooled specificity (0.87, 95% CI 0.41 - 1.00) and accuracy (0.85, 95% CI 0.72 - 0.94). For discrimination as measured by AUC, pooled estimates were available for 2 disorder types, among which SAD demonstrated a comparatively higher AUC of 0.92 (95% CI 0.85 - 0.98). The variation in predictive performance across anxiety disorder subtypes may partly account for the observed heterogeneity. In the subgroup analysis, GAD and PD demonstrated higher sensitivity, which may be related to the use of structured symptom-based scales that provide relatively stable responder signals for model

XSL•FO
RenderX

training [27]. First, SAD research often relies on symptom measures that closely align core fear and avoidance dimensions with treatment outcome definitions, which may facilitate clearer model discrimination [7,34]. Second, some SAD studies incorporate high-information features beyond routine clinical variables, potentially enhancing overall predictive performance [29,32]. Third, SAD samples are frequently drawn from relatively homogeneous clinical trials with standardized intervention protocols, which may reduce heterogeneity and improve model stability [7,32].

Comparison With Prior Work

In the systematic review and meta-analysis published by Vieira et al [35] in 2022, the authors evaluated ML approaches for predicting CBT treatment response across multiple psychiatric disorders. Their analysis included 24 studies comprising 7497 patients and covered 5 diagnostic groups: major depressive disorder, obsessive-compulsive disorder (OCD), posttraumatic stress disorder (PTSD), anxiety disorders, and substance use disorders. The overall pooled accuracy was 0.74 (95% CI 0.70 - 0.78), with variability across diagnostic categories. PTSD, anxiety disorders, and OCD demonstrated comparatively higher accuracy estimates (0.787, 0.776, and 0.761, respectively). In comparison, our study focused specifically on anxiety disorders and yielded a pooled accuracy of 0.74 (95% CI 0.62 - 0.84), which is comparable to the overall estimate reported by Vieira et al [35]. However, our analysis extended beyond accuracy by systematically synthesizing sensitivity, specificity, and AUC to provide a more comprehensive assessment of predictive performance. In addition, we conducted subgroup analyses examining the potential influence of data types (eg, clinical and neuroimaging), algorithm types, and disease types on model performance, thereby offering a more granular evaluation within the anxiety disorder domain. Bubble plots were used to illustrate temporal trends and the distribution of study characteristics across performance metrics.

Heterogeneity

The substantial heterogeneity observed across studies may influence the overall strength of evidence for AI-based prediction models. Given the anticipated between-study variability, a bivariate random-effects model was applied to pool sensitivity and specificity estimates. Meta-regression and subgroup analyses were further conducted to explore potential sources of heterogeneity. The results indicated that heterogeneity in sensitivity was significantly associated with the number of study centers (P=.02) and the anxiety disorder subtype (P=.01). In contrast, heterogeneity in specificity was significantly associated with CBT type (P<.001) and AI algorithm category (P<.001). The comparatively higher sensitivity observed in single-center studies, compared with multicenter studies, may reflect more homogeneous samples and more consistent data collection procedures [41]. Differences across anxiety disorder types may relate to variations in symptom structure, measurement alignment, and sample characteristics, which could influence the clarity of the responder signals available to the models [42]. Meanwhile, differences between iCBT and standard CBT may reflect variations in intervention structure and in the intensity of monitoring [43]. Variability in specificity across algorithm categories (DL vs ML) may relate to differences in model complexity and data representation strategies [44].

The bivariate boxplot suggested that 1 study [30] may represent a potential source of heterogeneity. Nevertheless, heterogeneity is likely multifactorial. Additional contributors may include differences in patient age, disorder severity, prior treatment exposure, geographic setting, sample size, feature selection strategies, data preprocessing methods, and hyperparameter optimization [45-48]. The combined influence of these methodological and clinical factors may account for the variability in predictive performance. Future research would benefit from more standardized reporting and systematic consideration of these variables to enhance reproducibility and external validity.

Implications for Practice and Research

Discussing the clinical interpretation of our results, the AI models demonstrated moderate predictive value for CBT treatment response in anxiety disorders, with the strongest performance observed for SAD. Multimodal models, which combine clinical demographics with imaging data, showed relatively superior predictive performance compared with unimodal approaches. This suggests that AI could support clinicians in making earlier, more informed therapeutic decisions and potentially improve patient outcomes by enabling timely treatment adjustments [49,50]. Specifically, these predictive models could be integrated into clinical workflows at key decision points: (1) during initial assessment to help stratify patients by likelihood of response, informing the intensity of early monitoring or the consideration of augmented treatment protocols, and (2) during the early stages of CBT to identify potential nonresponders, allowing for a timely switch or integration of other evidence-based interventions before disengagement occurs [35,51].

However, AI should serve as a decision-support tool rather than replacing clinician judgment, since treatment choices for anxiety disorders are influenced by not only severity but also patient preferences, comorbidity, and functional goals [52]. In the treatment of anxiety disorders, complex therapeutic decisions not only depend on assessing disease severity but also closely relate to the patient's individual condition and needs [53]. For real-world deployment, developing AI tools as clinician-facing dashboards that present predictive probabilities alongside key clinical context, rather than as autonomous decision systems, is crucial for fostering trust and facilitating shared decision-making [54]. Notably, among the studies we included, external validation sets were relatively scarce, and no research has yet focused solely on predicting outcomes for external validation cohorts. Future studies are needed to evaluate the generalization capabilities of AI models. Furthermore, limitations to the future adoption of AI include scarce annotated anxiety disorder data and regulatory hurdles [55]. Technical challenges persist in data availability, model interpretability, and transparency [56]. Advances in few-shot learning, self-supervised models, and centralized platforms may support an integrated AI ecosystem [57].

XSL•FO
RenderX

Limitations

When interpreting the findings of this systematic review and meta-analysis on the AI prediction of CBT treatment response for anxiety disorders, several limitations should be considered. First, the PROBAST+AI assessment indicated a high risk of bias in over half of the included studies, particularly in analytical methodology, which may have influenced the pooled estimates and contributed to between-study heterogeneity. Future studies should adopt more rigorous design and transparent reporting standards. Second, variability in the definition of treatment response across studies may have affected comparability. Although most scales assessed anxiety symptom reduction, differences in operational criteria could influence predictive performance. Greater standardization of outcome definitions would enhance consistency in future research. Third, the substantial heterogeneity and wide PIs suggest that model performance may vary considerably across clinical settings. Therefore, the pooled estimates should be interpreted cautiously and regarded as preliminary within the context of current methodological constraints. Fourth, although heterogeneity was observed across studies, this likely reflects differences in study design, patient characteristics, and modeling approaches. The reported PIs indicate that performance may vary across settings, highlighting the importance of external validation in future research. Finally, due to limited data availability, direct head-to-head comparisons between AI models and routinely used clinical scales were not feasible. Future research should incorporate comparative designs to better assess the incremental value of AI-based prediction tools. Overall, while AI models demonstrate potential for supporting treatment decision-making, stronger evidence from externally validated and methodologically robust studies is needed before clinical implementation.

Conclusions

This meta-analysis quantitatively synthesized AI performance in predicting CBT response for anxiety disorders, moving beyond narrative reviews to provide pooled evidence. In contrast to existing reviews that encompass broader diagnostic groups, our focused approach establishes a precise benchmark for this clinical domain, highlighting the current moderate overall performance. Furthermore, we extend beyond previous work by demonstrating the superior predictive utility of multimodal data, identifying SAD as the most predictable subtype, and systematically evaluating the impact of data modalities and algorithm types. Future efforts should prioritize robustly validated multimodal models, laying essential groundwork for the potential development of AI-assisted tools to personalize treatment planning in anxiety disorders.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing, adapting and adjusting emotional tone, translation, and reformatting. The GAI tool used was Sider. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. The declaration has been submitted by JL.

Funding

The authors declare that no funding was received for this study. No funding body had any role in the study design, data collection, analysis, interpretation, or writing of the manuscript.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

JL contributed to conceptualization, critical revision of the manuscript, and final approval. JL, JW, ZW, and MISBAA contributed to data curation, formal analysis, investigation, and methodology. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data to support the study.
[DOCX File, 252 KB - jmir_v28i1e86079_app1.docx]

Multimedia Appendix 2
Technical aspects and diagnostic performance data extracted from the included studies.
[DOCX File, 27 KB - jmir_v28i1e86079_app2.docx]

Checklist 1

XSL•FO
RenderX

PRISMA-DTA checklist.
[DOCX File, 21 KB - jmir_v28i1e86079_app3.docx]

Checklist 2
PRISMA 2020 Abstract checklist
[DOCX File, 19 KB - jmir_v28i1e86079_app4.docx]

Checklist 3
PRISMA-S checklist
[DOCX File, 20 KB - jmir_v28i1e86079_app5.docx]

References

1. Chodavadia P, Teo I, Poremski D, Fung DSS, Finkelstein EA. Prevalence and economic burden of depression and anxiety symptoms among Singaporean adults: results from a 2022 web panel. BMC Psychiatry 2023 Feb 14;23(1):104. [doi: 10.1186/s12888-023-04581-7] [Medline: 36782116]
2. Keramat SA, Comans T, Pearce A, Basri R, Hashmi R, Dissanayaka NN. Psychological distress and productivity loss: a longitudinal analysis of Australian working adults. Eur J Health Econ 2025 Nov;26(8):1503-1524. [doi: 10.1007/s10198-025-01764-9] [Medline: 40304834]
3. Powell C, Chiu CY, Sun X, So SHW. A meta-analysis on the efficacy of low-intensity cognitive behavioural therapy for generalised anxiety disorder. BMC Psychiatry 2024 Jan 2;24(1):10. [doi: 10.1186/s12888-023-05306-6] [Medline: 38166836]
4. Papola D, Miguel C, Mazzaglia M, et al. Psychotherapies for generalized anxiety disorder in adults: a systematic review and network meta-analysis of randomized clinical trials. JAMA Psychiatry 2024 Mar 1;81(3):250-259. [doi: 10.1001/jamapsychiatry.2023.3971] [Medline: 37851421]
5. Kravchenko O, Bäckman J, Mataix-Cols D, et al. Clinical, genetic, and sociodemographic predictors of symptom severity after internet-delivered cognitive behavioural therapy for depression and anxiety. BMC Psychiatry 2025 May 30;25(1):555. [doi: 10.1186/s12888-025-07012-x] [Medline: 40448103]
6. Kambeitz-Ilankovic L, Rzayeva U, Völkel L, et al. A systematic review of digital and face-to-face cognitive behavioral therapy for depression. NPJ Digit Med 2022 Sep 15;5(1):144. [doi: 10.1038/s41746-022-00677-8] [Medline: 36109583]
7. Bukhari Q, Rosenfield D, Hofmann SG, Gabrieli JDE, Ghosh SS. Predicting treatment response to cognitive behavior therapy in social anxiety disorder on the basis of demographics, psychiatric history, and scales: a machine learning approach. PLoS ONE 2025;20(3):e0313351. [doi: 10.1371/journal.pone.0313351] [Medline: 40100879]
8. Villarreal-Zegarra D, Paredes-Angeles R, Mayo-Puchoc N, Arenas-Minaya E, Huarcaya-Victoria J, Copez-Lonzoy A. Psychometric properties of the GAD-7 (General Anxiety Disorder-7): a cross-sectional study of the Peruvian general population. BMC Psychol 2024 Apr 2;12(1):183. [doi: 10.1186/s40359-024-01688-8] [Medline: 38566138]
9. Manzar MD, Kashoo FZ, Salahuddin M, et al. Hamilton Rating Scale for Anxiety: exploring validity with robust measures of classical theory parameters and a rating scale model in university students. BJPsych Open 2025 Aug 12;11(5):e176. [doi: 10.1192/bjo.2025.10055] [Medline: 40790884]
10. Zhang CY, Voort JLV, Yuruk D, et al. A characterization of the Clinical Global Impression Scale thresholds in the treatment of adolescent depression across multiple rating scales. J Child Adolesc Psychopharmacol 2022 Jun;32(5):278-287. [doi: 10.1089/cap.2021.0111] [Medline: 35704877]
11. Arrow K, Resnik P, Michel H, et al. Evaluating the use of online self-report questionnaires as clinically valid mental health monitoring tools in the clinical whitespace. Psychiatr Q 2023 Jun;94(2):221-231. [doi: 10.1007/s11126-023-10022-1] [Medline: 37145257]
12. Wang S, Lu J, Zheng G, et al. Poor performance of PHQ-9 and GAD-7 in screening clinical depression and anxiety among a large sample of Chinese children and adolescents. BMC Psychiatry 2025 Mar 29;25(1):301. [doi: 10.1186/s12888-025-06754-y] [Medline: 40158158]
13. Bertie LA, Arendt K, Coleman JRI, et al. Patterns of sub-optimal change following CBT for childhood anxiety. J Child Psychol Psychiatry 2024 Dec;65(12):1612-1623. [doi: 10.1111/jcpp.14009] [Medline: 38817012]
14. Park JH, Shin YB, Jung D, et al. Machine learning prediction of anxiety symptoms in social anxiety disorder: utilizing multimodal data from virtual reality sessions. Front Psychiatry 2024;15:1504190. [doi: 10.3389/fpsyt.2024.1504190] [Medline: 39896993]
15. Sadeh-Sharvit S, Camp TD, Horton SE, et al. Effects of an artificial intelligence platform for behavioral interventions on depression and anxiety symptoms: randomized clinical trial. J Med Internet Res 2023 Jul 10;25:e46781. [doi: 10.2196/46781] [Medline: 37428547]
16. Habicht J, Dina LM, McFadyen J, et al. Generative AI-enabled therapy support tool for improved clinical outcomes and patient engagement in group therapy: real-world observational study. J Med Internet Res 2025 Mar 10;27:e60435. [doi: 10.2196/60435] [Medline: 40063074]

XSL•FO
RenderX

17. Prasad N, Chien I, Regan T, et al. Deep learning for the prediction of clinical outcomes in internet-delivered CBT for depression and anxiety. PLoS ONE 2023;18(11):e0272685. [doi: 10.1371/journal.pone.0272685] [Medline: 38011176]
18. Hilbert K, Böhnlein J, Meinke C, et al. Lack of evidence for predictive utility from resting state fMRI data for individual exposure-based cognitive behavioral therapy outcomes: a machine learning study in two large multi-site samples in anxiety disorders. Neuroimage 2024 Jul 15;295(120639):120639. [doi: 10.1016/j.neuroimage.2024.120639] [Medline: 38796977]
19. Ohi K, Fujikane D, Takai K, et al. Clinical features and genetic mechanisms of anxiety, fear, and avoidance: a comprehensive review of five anxiety disorders. Mol Psychiatry 2025 Oct;30(10):4928-4936. [doi: 10.1038/s41380-025-03155-1] [Medline: 40830577]
20. Basterfield C, Newman MG. Development of a machine learning-based multivariable prediction model for the naturalistic course of generalized anxiety disorder. J Anxiety Disord 2025 Mar;110:102978. [doi: 10.1016/j.janxdis.2025.102978] [Medline: 39904097]
21. Cruz-Gonzalez P, He AWJ, Lam EP, et al. Artificial intelligence in mental health care: a systematic review of diagnosis, monitoring, and intervention applications. Psychol Med 2025 Feb 6;55:e18. [doi: 10.1017/S0033291724003295] [Medline: 39911020]
22. Salameh JP, Bossuyt PM, McGrath TA, et al. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. BMJ 2020 Aug 14;370:m2632. [doi: 10.1136/bmj.m2632] [Medline: 32816740]
23. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
24. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ 2025 Mar 24;388:e082505. [doi: 10.1136/bmj-2024-082505] [Medline: 40127903]
25. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
26. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023 Dec;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
27. Ball TM, Stein MB, Ramsawh HJ, Campbell-Sills L, Paulus MP. Single-subject anxiety treatment outcome prediction using functional neuroimaging. Neuropsychopharmacology 2014 Apr;39(5):1254-1261. [doi: 10.1038/npp.2013.328] [Medline: 24270731]
28. Bertie LA, Quiroz JC, Berkovsky S, et al. Predicting remission following CBT for childhood anxiety disorders: a machine learning approach. Psychol Med 2024 Dec 17;17:1-11. [doi: 10.1017/S0033291724002654] [Medline: 39686883]
29. Frick A, Engman J, Alaie I, et al. Neuroimaging, genetic, clinical, and demographic predictors of treatment response in patients with social anxiety disorder. J Affect Disord 2020 Jan 15;261:230-237. [doi: 10.1016/j.jad.2019.10.027] [Medline: 31655378]
30. Hahn T, Kircher T, Straube B, et al. Predicting treatment response to cognitive behavioral therapy in panic disorder with agoraphobia by integrating local neural information. JAMA Psychiatry 2015 Jan;72(1):68-74. [doi: 10.1001/jamapsychiatry.2014.1741] [Medline: 25409415]
31. Hentati Isacsson N, Ben Abdesslem F, Forsell E, Boman M, Kaldo V. Methodological choices and clinical usefulness for machine learning predictions of outcome in Internet-based cognitive behavioural therapy. Commun Med (Lond) 2024 Oct 10;4(1):196. [doi: 10.1038/s43856-024-00626-4] [Medline: 39384934]
32. Månsson KNT, Frick A, Boraxbekk CJ, et al. Predicting long-term outcome of Internet-delivered cognitive behavior therapy for social anxiety disorder using fMRI and support vector machine learning. Transl Psychiatry 2015 Mar 17;5(3):e530. [doi: 10.1038/tp.2015.22] [Medline: 25781229]
33. Sundermann B, Bode J, Lueken U, et al. Support vector machine analysis of functional magnetic resonance imaging of interoception does not reliably predict individual outcomes of cognitive behavioral therapy in panic disorder with agoraphobia. Front Psychiatry 2017;8:99. [doi: 10.3389/fpsyt.2017.00099] [Medline: 28649205]
34. Whitfield-Gabrieli S, Ghosh SS, Nieto-Castanon A, et al. Brain connectomics predict response to treatment in social anxiety disorder. Mol Psychiatry 2015 May;21(5):680-685. [doi: 10.1038/mp.2015.109] [Medline: 26260493]
35. Vieira S, Liang X, Guiomar R, Mechelli A. Can we predict who will benefit from cognitive-behavioural therapy? A systematic review and meta-analysis of machine learning studies. Clin Psychol Rev 2022 Nov;97:102193. [doi: 10.1016/j.cpr.2022.102193] [Medline: 35995023]
36. Jankowsky K, Krakau L, Schroeders U, Zwerenz R, Beutel ME. Predicting treatment response using machine learning: a registered report. Br J Clin Psychol 2024 Jun;63(2):137-155. [doi: 10.1111/bjc.12452] [Medline: 38111213]
37. Hornstein S, Forman-Hoffman V, Nazander A, Ranta K, Hilbert K. Predicting therapy outcome in a digital mental health intervention for depression and anxiety: a machine learning approach. Digit Health 2021;7:20552076211060659. [doi: 10.1177/20552076211060659] [Medline: 34868624]

XSL•FO
RenderX

38. Chekroud AM, Bondar J, Delgadillo J, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry 2021 Jun;20(2):154-170. [doi: 10.1002/wps.20882] [Medline: 34002503]
39. Piccolo SR, Mecham A, Golightly NP, Johnson JL, Miller DB. The ability to classify patients based on gene-expression data varies by algorithm and performance metric. PLoS Comput Biol 2022 Mar;18(3):e1009926. [doi: 10.1371/journal.pcbi.1009926] [Medline: 35275931]
40. AlSaad R, Abd-Alrazaq A, Boughorbel S, et al. Multimodal large language models in health care: applications, challenges, and future outlook. J Med Internet Res 2024 Sep 25;26:e59505. [doi: 10.2196/59505] [Medline: 39321458]
41. Bafeta A, Dechartres A, Trinquart L, Yavchitz A, Boutron I, Ravaud P. Impact of single centre status on estimates of intervention effects in trials with continuous outcomes: meta-epidemiological study. BMJ 2012 Feb 14;344:e813. [doi: 10.1136/bmj.e813] [Medline: 22334559]
42. Fomenko A, Dümmler D, Aktürk Z, et al. Hospital Anxiety and Depression Scale Anxiety subscale (HADS-A) for detecting anxiety disorders in adults. Cochrane Database Syst Rev 2025 Jul 2;7(7):CD015456. [doi: 10.1002/14651858.CD015456] [Medline: 40600405]
43. Hedman-Lagerlöf E, Carlbring P, Svärdman F, Riper H, Cuijpers P, Andersson G. Therapist-supported Internet-based cognitive behaviour therapy yields similar effects as face-to-face therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. World Psychiatry 2023 Jun;22(2):305-314. [doi: 10.1002/wps.21088] [Medline: 37159350]
44. Abrol A, Fu Z, Salman M, et al. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. Nat Commun 2021 Jan 13;12(1):353. [doi: 10.1038/s41467-020-20655-6] [Medline: 33441557]
45. Kartoun U, Khurshid S, Kwon BC, et al. Prediction performance and fairness heterogeneity in cardiovascular risk models. Sci Rep 2022 Jul 22;12(1):12542. [doi: 10.1038/s41598-022-16615-3] [Medline: 35869152]
46. de Jong VMT, Moons KGM, Eijkemans MJC, Riley RD, Debray TPA. Developing more generalizable prediction models from pooled studies and large clustered data sets. Stat Med 2021 Jul 10;40(15):3533-3559. [doi: 10.1002/sim.8981] [Medline: 33948970]
47. Benkarim O, Paquola C, Park BY, et al. Population heterogeneity in clinical cohorts affects the predictive accuracy of brain imaging. PLoS Biol 2022 Apr;20(4):e3001627. [doi: 10.1371/journal.pbio.3001627] [Medline: 35486643]
48. Li Y, Yu C, Zhao Y, Yao W, Aseltine RH, Chen K. Pursuing sources of heterogeneity in modeling clustered population. Biometrics 2022 Jun;78(2):716-729. [doi: 10.1111/biom.13434] [Medline: 33527347]
49. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
50. Khalifa M, Albadawy M. Artificial intelligence for clinical prediction: exploring key domains and essential functions. Computer Methods and Programs in Biomedicine Update 2024;5:100148. [doi: 10.1016/j.cmpbup.2024.100148]
51. Hammelrath L, Hilbert K, Heinrich M, Zagorscak P, Knaevelsrud C. Select or adjust? How information from early treatment stages boosts the prediction of non-response in internet-based depression treatment. Psychol Med 2024 Jun;54(8):1641-1650. [doi: 10.1017/S0033291723003537] [Medline: 38087867]
52. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med 2018 Mar 15;378(11):981-983. [doi: 10.1056/NEJMp1714229] [Medline: 29539284]
53. Tendolkar VD, Chouhan DS, Roy D, et al. Advances in evidence-based interventions for depression and anxiety: implications for mental health nursing practice. Int J Environ Sci 2025;11(23s):826-834 [FREE Full text] [doi: 10.64252/mev1b885]
54. As'ad M, Faran N, Joharji H. AI-Supported Shared Decision-Making (AI-SDM): conceptual framework. JMIR AI 2025 Aug 7;4:e75866. [doi: 10.2196/75866] [Medline: 40773762]
55. Manole A, Cârciumaru R, Brînza R, Manole F. An exploratory investigation of chatbot applications in anxiety management: a focus on personalized interventions. Information 2024;16(1):11. [doi: 10.3390/info16010011]
56. Das KP, Gavade P. A review on the efficacy of artificial intelligence for managing anxiety disorders. Front Artif Intell 2024;7:1435895. [doi: 10.3389/frai.2024.1435895] [Medline: 39479229]
57. Sosna J, Joskowicz L, Saban M. Navigating the AI landscape in medical imaging: a critical analysis of technologies, implementation, and implications. Radiology 2025 Jun;315(3):e240982. [doi: 10.1148/radiol.240982] [Medline: 40552997]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
CBT: cognitive behavioral therapy
CGI-I: Clinical Global Impressions–Improvement
DL: deep learning
FN: false negative
FP: false positive
GAD: generalized anxiety disorder
GAD-7: Generalized Anxiety Disorder-7

XSL•FO
RenderX

HAM-A: Hamilton Anxiety Rating Scale
HKSJ: Hartung-Knapp-Sidik-Jonkman
LR: logistic regression
LSAS: Liebowitz Social Anxiety Scale
ML: machine learning
OASIS: Overall Anxiety Severity and Impairment Scale
OCD: obsessive-compulsive disorder
PD: panic disorder
PHQ-9: Patient Health Questionnaire-9
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST+AI: Prediction Model Risk of Bias Assessment Tool for Artificial Intelligence
PTSD: posttraumatic stress disorder
RF: random forest
RNN: recurrent neural network
SAD: social anxiety disorder
SPQ: Spider Phobia Questionnaire
SVM: support vector machine
TN: true negative
TP: true positive

Edited by S Brini; submitted 17.Oct.2025; peer-reviewed by C Ahmadu, D Ghosh, S Mohanadas, SG Virupaksha, Y Zhang; revised version received 12.Feb.2026; accepted 24.Feb.2026; published 18.Mar.2026.

Please cite as:
Liu J, Wang J, Wu Z, Adam Assim MISB
Artificial Intelligence for Predicting Treatment Response in Patients With Anxiety Disorders After Cognitive Behavioral Therapy: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e86079
URL: https://www.jmir.org/2026/1/e86079
doi:10.2196/86079

© Jiawen Liu, Junhui Wang, Zhaobin Wu, Mohamad Ibrani Shahrimin Bin Adam Assim. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Anticipating Moral and Economic Considerations, Opportunities, and Potential Frictions for AI in Medical Imaging: Multistakeholder Cocreation Study

Martin Bastiaan Schilder[1], MSc; Alexandra Keyser[2], MSc; Susan van Hees[2], PhD; Alessandro Sbrizzi[1], PhD; Wouter Pieter Christiaan Boon[2], PhD

[1]Computational Imaging Group for MR Therapy and Diagnostics, Center for Image Sciences, University Medical Center Utrecht, Heidelberglaan 100, Utrecht, The Netherlands

[2]Faculty of Geosciences, Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, The Netherlands

Corresponding Author:
Martin Bastiaan Schilder, MSc
Computational Imaging Group for MR Therapy and Diagnostics, Center for Image Sciences, University Medical Center Utrecht, Heidelberglaan 100, Utrecht, The Netherlands

Abstract

Background: Artificial intelligence (AI) promises to significantly impact daily radiology practices. Numerous studies have already been conducted that anticipate this potentially disruptive innovation. So far, these studies have mainly focused on single topics, such as “trust,” or investigating perspectives of single stakeholder groups, such as “radiologists.”

Objective: This study aims to explore future directions for AI in radiology by incorporating perspectives of a heterogeneous group of stakeholders on a broad spectrum of moral and economic topics. It also aims to cocreate and reflect with a broad range of stakeholders on viable implementation scenarios for scalable AI applications in radiology in the Netherlands, thereby identifying potential opportunities and frictions, with a focus on moral and economic considerations.

Methods: To inform the workshop design, a nonsystematic narrative literature search was performed to deepen our understanding of key moral and economic considerations at play in the field of radiology and AI. Workshop participants, representing a wide range of actors including radiologists, innovators, and patient representatives, were selected using purposive sampling. Data were collected in a cocreation workshop. In 3 subsequent rounds, mixed over 3 breakout groups, a total of 17 participants were asked to (1) map what they considered important moral and economic considerations, (2) envision possible future scenarios for AI in radiology, and (3) discuss opportunities, frictions, and routes to success. Transcribed recordings were coded and cross-checked.

Results: Workshop participants envision future AI-driven scenarios, ranging from extramural imaging departments for increased accessibility to health care, to multimodal data integration for human-centered AI-enhanced diagnostics. Seven themes emerge from the discussions during the workshop: (1) trust and efficiency of AI technologies, (2) responsibilities in clinical decision-making when AI is involved, (3) diagnosis as a one-off versus an iterative process, (4) regulations as a requirement or a restriction, (5) economic benefits or drawbacks, (6) trade-off between amount of information required and patient privacy, and (7) environmental considerations.

Conclusions: Reflecting on the 7 emerging themes, we identify three overarching topics: (1) human-AI collaboration and trust, (2) governance, regulation, and ethical safeguards, and (3) value creation and sustainability. These topics highlight the need to balance technological advancements with ethical responsibility, institutional accountability, and societal benefit. They also underscore the importance of designing AI systems that not only perform well but are also trusted and aligned with clinical workflows and patient values. These overarching themes offer a lens through which future research and policy can navigate the complex interplay between innovation, regulation, and real-world implementation. Future research is needed to validate the generalizability of the results across various countries and health care settings.

(J Med Internet Res 2026;28:e83407) doi:10.2196/83407

KEYWORDS

cocreation; stakeholder perspectives; radiology; artificial intelligence; responsible research and innovation

XSL•FO
RenderX

Introduction

Background

As health care systems worldwide face increasing demand and complexity, the integration of advanced technologies such as artificial intelligence (AI) is gaining momentum. Many clinical experts and sector specialists consider AI to have transformative potential for the field of radiology [1-4]. They argue, for instance, that AI offers the potential to enhance diagnostic accuracy, to support early disease detection, and to personalize patient care through applications such as image segmentation, anomaly detection, and predictive modeling [1,5-7]. These applications could automate routine tasks and uncover subtle patterns that can be missed by the human eye [1,8] and help radiologists to focus more on complex clinical decisions such as diagnosing intricate disease presentations [7].

There is a growing body of literature discussing the promises of AI in radiology from different perspectives. From a moral perspective, many scholars propose AI to contribute to safer, more patient-centered care by reducing diagnostic errors and unnecessary interventions [9]. Economically, proponents argue that AI can contribute to streamlining radiology workflows by improving productivity, shortening turnaround times, and lowering health care costs through early detection of disease and more efficient resource allocation [7,10-12]. Alongside these promises, scholars warn that we should remain sensitive to risks and challenges, such as algorithm opacity and profit-driven motives, created by an emerging role of AI in health care [13,14].

Despite these promises, much of the existing research on the role of AI in radiology remains limited in scope. So far, various studies have explored AI in radiology, mostly from a single perspective, such as trust or acceptance [15,16]. Moreover, existing research has primarily explored these perspectives within only a single stakeholder group or a limited range of groups, such as radiologists, innovators, or patients [7,15,17-20]. These studies also have predominantly applied retrospective methods, such as surveys and questionnaires [16,20-22]. Although these earlier studies offer valuable insights, there is a need for research broadening the perspectives and stakeholders. Adopting a broader approach enables the examination of interactions among stakeholders while situating their perspectives within the wider set of trade-offs and contextual factors.

Objective

The study aimed to cocreate and reflect with a broad range of stakeholders on viable implementation scenarios for scalable AI applications in radiology in the Netherlands, thereby identifying potential opportunities and frictions, with a focus on moral and economic considerations. Drawing on insights from responsible innovation, we endeavored to include a wide variety of views in an early stage of the emergence of a technology, anticipating potential implications and frictions, aiming to increase chances for successful innovations [17,23,24]. The inclusion of varied perspectives, especially those of patients [25,26], has been widely advocated as essential to shaping relevant research agendas and outcomes [15,16,27,28]. Early-stage engagement should increase opportunities for aligning technology development not only with clinical needs and regulatory requirements but also to include current experiences and care practices, anticipate expectations and underlying values [17,27,29-33], and answer questions regarding commercialization [34]. Engaging stakeholders encourages iterative feedback, increasing the chance of AI systems to evolve alongside real-world needs and cultivate shared ownership of the innovation process [8].

During a cocreation workshop, stakeholders were asked to build concrete implementation scenarios for AI in imagined future intra- and extramural settings and to reflect on expected opportunities and potential frictions created by moral and economic considerations. The forward-looking, participatory approach that we used contributed to a more comprehensive understanding of possible future pathways, which we present in the “Results” section. Subsequently, we describe the 7 themes that emerged from these scenarios, touching on various moral and economic considerations, involving among others the role of the human (radiologist and patients), as well as issues related to regulations, economic considerations, and the environmental impact. Finally, the “Discussion” section further reflects on these themes, subsequently concluding with 3 key implications for futures of AI in radiology.

Methods

Study Design

Our research approach reflects starting points from responsible innovation. We started with a nonsystematic narrative literature search to identify existing topics and gaps in the scientific discourse. These topics served as input for the cocreation workshop in which future scenarios for AI in radiology were developed. In the workshop, we took an explorative, anticipatory, and cocreation approach, in which stakeholders were allowed to raise and reflect on interconnected topics, rather than addressing these topics in isolation.

Identification of Workshop Topics

To gain an initial understanding and to provide the research team with context of the recurring topics and stakeholders in the field of AI, radiology, and business models, we conducted a nonsystematic narrative literature search aimed at identifying key topics and trends in the field [35,36]. In light of the fact that we aimed for a broad group of topics and stakeholders, we were a priori interested in both moral and economic considerations. Resultingly, our approach included keyword-based searches in primarily PubMed and Google Scholar using search terms such as “Radiology,” “AI,” “Transparency,” and “Value proposition.” These terms were iteratively refined as understanding of the topic evolved and, in the end, also included “Algorithmic fairness” and “Revenue streams” as search terms, as is common practice in a narrative review [36]. We also used forward and backward snowballing from key publications to capture additional influential works. In addition, we browsed the digital archives of leading journals in the field, including *Journal of Medical Internet Research* (and relevant sister journals), *European Radiology*, and *Insights into Imaging*. This flexible and exploratory search strategy

XSL•FO
RenderX

suited our objective of mapping conceptual developments and identifying recurrent topics across the literature rather than providing an exhaustive or systematic synthesis. Textbox 1 provides the main findings of this review.

Textbox 1. Topics covering moral and economic considerations coming from the literature search.

In the literature search, we focused on artificial intelligence (AI), radiology, and business models, and both moral and economic considerations came forward. Radiologists increasingly rely on digital tools, including AI-based tools, for various functions [5-7]. This raises moral and economic concerns related to discussions about, for example, trust between the patient and physician [19,31,37-39], supervision and accountability in clinical decision-making [16,40,41], lacking or unclear reimbursement structures [42], and high upfront costs [43]. In these discussions, some recurring topics came forward.

Trust is often mentioned in literature as playing a critical role in the field of radiology. Where, traditionally, trust between a patient and a physician was deemed important, another dimension is added with the introduction of AI, and the definition of trust is increasingly broadened to both the clinicians' and patients' trust in the technology's performance and reliability [15,18,33,44-49]. However, trust is a complex concept that relates to interconnected discussions on, among others, fairness, safety, explainability, and transparency [50]. To increase fairness and safeguard vulnerable populations, managing risks is essential [40,51]. AI models often reflect biases inherent to insufficiently diverse training datasets [40,51,52]. Resultingly, model performance can deteriorate if applied to a different population. Explainability and transparency are fundamental to building trust in AI systems [16,53,54] and can be considered key fairness strategies to mitigate bias and improve accountability [30]. Here, the distinction is that interpretability means that a radiologist can understand and make sense of the outputs of the AI system, whereas explainability means that the AI system also provides the radiologist with a rationale or reasoning behind the results [2].

Despite their importance, many AI models suffer from the "black box problem," where decision-making processes can be opaque even to developers [30,46]. Obstruction of error identification through a lack of transparency compromises trust and accountability [30,55,56]. The General Data Protection Regulation mandates transparency and explainability in automated decision-making, but many systems struggle to meet this standard [46,57]. On the topic of data, while the General Data Protection Regulation, for instance, grants patients ownership over personal data through, for example, obliging data users to obtain informed consent, uncertainties persist regarding data reuse and consent withdrawal [40,46,55,58]. Ownership thus means that a person or legal entity has the full and inherent right to decide what happens with the data without requiring consent from anyone else. By contrast, controls refer to the delegated ability to use or manage data within boundaries agreed upon by the owner, granted through informed consent. Developers, users, and health care providers must navigate complex questions about responsible data storage and use [49]. Transparency in data management could further support trust, yet proprietary algorithms and cross-disciplinary misalignments between developers and health care practitioners create significant challenges [58]. It has therefore been advocated that radiologists must develop new skills to effectively collaborate with AI, which requires additional training [49,50,59,60] and potentially a shift in roles and responsibilities [61]. Concerns about overreliance on AI systems, loss of diagnostic expertise, and shifts in job satisfaction are frequently noted [62-64]. Additionally, unresolved issues in assigning accountability for AI-assisted errors complicate the integration into clinical workflows, raising questions about liability and autonomy among developers, users, and insurers [16,40,41].

Business models for AI applications in radiology outlined in the literature reviewed include perpetual licensing, pay-per-use, and subscription-based models [65,66], with most companies offering subscription or hybrid pricing models [65]. Deployment and pricing strategies have not yet converged to a preferred or market-dominant standard, and most vendors offer multiple options. Literature discusses the shift of software as a medical device from product-centric to platform-based models [67]. This idea of platformization could potentially facilitate easier (vendor-neutral) compatibility and integrability within existing hospital infrastructure [8,68,69]. In essence, all businesses in the medical software industry operate in a highly dynamic environment with constant challenges, including disruptive innovation, strict regulations, sustainability concerns, and uncertainties [70,71]. To navigate this landscape and mitigate risk, a more dynamic business model has been proposed, tailored to the specific demands of the software as medical device domain. Long-term stability of business models can be achieved through diversified revenue streams and optimizing operational efficiency, regardless of market conditions [71]. Hospitals also seek ways to fund AI-supported services. Some sources point to missing reimbursement structures or high upfront costs, hindering adoption [42,43]. There is, however, a need for more research on sustainable business models for hospitals that balance adoption costs with strong value propositions and reimbursement frameworks [72,73] and value assessment through real-world clinical validation [74].

The insights from this nonsystematic narrative literature search were synthesized into topics, prompts, and discussion points, which served as input for the cocreation workshop, and they were used to create workshop canvasses. The workshop canvasses can be found in Multimedia Appendices 1 and 2. These canvases guided participant engagement and helped ensure that the workshop addressed research-relevant issues by suggesting conversation topics, while remaining flexible for participant input.

Workshop Design and Participants

Since it is argued that integration of AI into radiology may provide a transformative opportunity for health care delivery, we further explored opportunities for AI in radiology in a multistakeholder workshop. Workshop participants were selected using purposive sampling to ensure the inclusion of individuals relevant to the research objectives. We aimed to include participants from diverse stakeholder backgrounds. As a point of departure, participants from an earlier workshop with a diverse panel of stakeholder backgrounds [28,61] were approached first. In addition, we explicitly looked for expertise and relevant types of stakeholders that were missing, were underrepresented, and regarded as important in relation to the workshop's themes. This search included reaching out to contacts in the authors' own network, for example, through LinkedIn and institutional websites. Prospective participants were also allowed to suggest relevant candidates from their own network. The patient representatives were members of the hospital patient council and were approached by hospital staff for participation. Participation was voluntary aside from a compensation for travel-related expenses. Potential selection bias was balanced by having participants representing very different perspectives.

In the workshop, patient representatives (n=2), radiologists (n=2), laboratory technicians (n=2), a technology developer (n=1), clinical researchers (n=2), industry employees (imaging hardware company employee: n=2; medical software company employee: n=1), a funding body employee (n=1), a clinical physicist (n=1), a business developer (n=1), and social scientists

XSL•FO
RenderX

(n=2) participated, totaling 17 participants. The sessions were moderated by the researchers (MBS, SH, and WB). Next to discussing from their expert point of view, all participants were also invited to reflect on questions from a general citizen's point of view, allowing for flexible conversations between the participants.

Ethical Considerations

The research protocol was assessed and exempted from medical ethical review by the Medical Ethics Board of University Medical Center Utrecht (Declaration METC no. 22 - 475/DB, d.d. 1 March 2022) [75]. This exemption was determined in accordance with the Dutch Medical Research Involving Human Subjects Act and the definitions and guidance provided by the Central Committee on Research Involving Human Subjects [76]. All participants gave written and verbal informed consent prior to the workshop and had the opportunity to react to a written workshop report after the workshop. In the analysis, the data were coded, meaning that participants received a study number. In this paper, we refer to the participants with that same code, although sometimes referring to their respective stakeholder background. Patient representatives received compensation for travel-related expenses and were offered a voucher of €25 (approximately US $27 at the time of the workshop).

Data Collection

The cocreation workshop took place in June 2024 and lasted 3 hours. Participants were asked in three rounds to (1) map and reflect on moral and economic considerations they considered important when thinking about the field of radiology and AI, (2) cocreate concrete scenarios, envisioning plausible futures for radiology with AI in the not too long future (year 2040), to elaborate on the broader implications of their considerations to the field, and (3) think about opportunities and potential frictions related to these scenarios and how to deal with or prevent them.

The 15-year time frame was chosen to allow participants to envision a future where significant transformation could happen, while still being close enough to make the ideas feel actionable. We asked our participants to envision scenarios *postadoption* of AI in radiology, intentionally probing the workshop participants not to focus on pragmatic topics, such as hardware requirements, because our aim was to explore strategic and value-driven perspectives rather than operational constraints. We wanted participants to think beyond immediate technical hurdles and envision how AI could reshape workflows, roles, and patient care once adoption challenges have been resolved. In 3 breakout groups, each with 5 or 6 participants, scenarios were cocreated and discussed. In round 1, the participants were asked to discuss their respective considerations in groups of 2 or 3 participants. For rounds 2 and 3, we provided the participants with a worksheet canvas to guide the discussion. The questions on the canvasses were loosely based on the topics we found in our literature search (Textbox 1). The canvasses can be found in Multimedia Appendices 1 and 2. We explicitly told participants that these sheets could be used as inspiration for conversations, and if other ideas came to mind, that they were free to explore them. No further prompts were provided, except for those related to time management of the workshop. If questions on interpretation of the canvasses came forward, the session moderators answered the question neutrally and aimed not to steer the discussion in a certain direction. In case of dissent between the participants, the session moderators did not intervene, as disagreements between the participants could potentially reveal interesting frictions between various stakeholder backgrounds. All discussions were audio recorded, the filled-out worksheet canvasses were stored, and notes were taken by the researchers (MBS, SH, and WB). A more elaborate quotebook is provided in Multimedia Appendix 3.

Data Analysis

The workshop was held in Dutch. All audio data were automatically transcribed using Amberscript (Amberscript BV), reread, checked, and corrected where necessary. Transcripts were then automatically translated to English using Microsoft Word (Microsoft Corporation), again reread, checked, and corrected where necessary. The transcripts were independently coded. MBS and SH, as native Dutch speakers, coded the original transcripts in order to stay close to the original meaning, and AK, as a native English speaker, coded the translated transcripts. The codes were subsequently compared and refined using qualitative analysis software NVivo (version 14; Lumivero). Data were analyzed using a semi-inductive coding approach, guided by predefined categories, but remaining flexible to the identification of emergent codes as new themes arose from the data. The initial codes were related to the topics we identified in the literature search and that we described in Textbox 1, namely, function, performance, supervision or autonomy, responsibility, explainability, transparency, training and job roles, algorithm performance, user and patient security, data management, accountability, bias, value assessment, business model aspects, value proposition, funding, platformization, and compatibility and integrability. Additionally, during the analysis, we found that "patient communication" was emerging as a new and recurring topic. This code was subsequently added to the codebook. Our analysis focused on moral and economic considerations, opportunities, and frictions, which materialized in the scenarios. To conform with scientific rigor, consistency between the researchers was ensured, as the researchers discussed their findings after initially independently coding the transcripts. That is, these peer debriefing sessions aided in ensuring that our interpretations were firmly grounded in the data. During these peer debriefing sessions, the researchers systematically compared their coding decisions for each transcript, focusing on whether the chosen codes accurately captured the meaning of the data and whether the interpretations were consistent across researchers. They reviewed the rationale behind each assigned code, discussed eventual discrepancies, and clarified ambiguous segments. Disagreements were openly discussed, and alternative perspectives were considered before reaching a shared interpretation, ensuring that the final coding reflected consensus. This iterative dialogue served to refine the coding framework and enhance the credibility of the analysis. Furthermore, a workshop report inspired by a preliminary analysis of the data was shared with the participants, providing them with an opportunity to disagree with the researchers' interpretation of the data or add thoughts that came after the workshop.

XSL•FO
RenderX

Results

Cocreation Workshop

In this section, we share the findings, drawing on insights from the multistakeholder cocreation workshop. Many topics came forward, including questions regarding sharing or balancing responsibilities between humans and technology, power and influence in relation to data, ownership, access and trust, and questions regarding costs and payments. Developing and reflecting on the scenarios helped identify expectations. Different stakeholders brought forward different perspectives in the discussions. For example, in groups that included a patient representative, the importance of having a human radiologist involved came forward.

Most participants shared the expectation that between now and 2040 much will change and probably quite rapidly, referring to fast developments surrounding general purpose AI tools such as ChatGPT. Participants in our multistakeholder workshop considered in their scenario either intramural improvement of radiology processes or the extramural screening and early detection of disease. This resulted in 3 scenarios sketched out in the next paragraphs: AI copilot, scanner on tour, and Alzheimer disease (AD) screening. We present each scenario following a similar structure; first, we report what participants believe could be a good use case of AI (*materialization*), after which we share their *moral and economic considerations*, *opportunities*, and *frictions*. As a way of analysis, we highlight emerging themes from the scenarios.

The 3 Scenarios

Scenario 1. Partner in Diagnosis: Envisioning AI as the Radiologist's Trusted Copilot

Materialization

Participants envisioned that in 2040 AI would be integrated into every hospital's radiology department, acting as a "copilot" to assist radiologists. Radiologists may grow from an executing role to a more supervisory role, sometimes required only to approve the AI's suggestion. In this scenario, AI can aid in the noninvasive characterization of a tumor, enriching radiological findings on computed tomography or magnetic resonance imaging (MRI) with a description of histology and tumor biology without requiring surgical procedures, such as biopsies. Or, in case a rich collection of patient data is already available, AI can integrate such multimodal data:

> *...also the integration with biomarkers, with pathology with those [foundational models], and the EHR [Electronic Health Record], gives a much more complete picture than just with those images." the workshop participant continued, arguing how that despite all clinicians already work together in the same system, she expects AI to be an enabler of actual integration : "(...) I think that integration into AI of all those different things will also become very strong.* [g3, s6]

AI in this use case supports the radiologist with additional insights that can be used for improved diagnosis or treatment response monitoring. According to the participants, the added value can be demonstrated in independent clinical studies in which researchers, clinicians, and ethical evaluation boards should have their usual role. Explicating added value incentivizes medical professionals to include AI in guidelines and insurance companies to reimburse its use. AI can subsequently be used to collect relevant patient data and provide comprehensive interpretations, whereas the radiologist retains ultimate responsibility for diagnosis and treatment decisions.

Participants thought that software developers should be in the lead in setting up a quality management system. They can supply their product, for example, through a pay-per-view business model, which is integrated into Picture Archiving and Communication System or electronic health record along with other vendors' products. Users were regarded as being in the lead for evaluating real-world, postmarket applications. Although data have the potential to be part of business models, the participants thought that patients should retain ownership of their data, emphasizing that these data should be anonymous and nontraceable.

Moral and Economic Considerations

Participants deemed AI in radiology necessary to support the radiologist and underlined its potential, considering the hours spent on, for example, follow-up MRI scans without visible changes to the imaged pathology. Requirements formulated for this scenario were as follows: (1) The treating physician always remains responsible for the patient's health. Since human experts should always remain in control, physicians remain responsible for communicating with patients. (2) AI should benefit the patients and lead to improved health, while applying to as many patients as possible. (3) Trust in AI can be increased by observing well-performing models that provide added diagnostic value, which can convince radiologists to use and fully adopt these models. Trust in AI was considered a means of adoption for this technology. (4) The industry should focus on user-friendliness and insightful cost-benefit analyses to improve adoption. (5) Besides clear added value in the context of diagnosis and treatment, AI should have a positive cost-benefit balance.

Opportunities

Assistance to radiologists was expected to add value by improving efficiency and accuracy in ongoing patient care and monitoring. Furthermore, added value was expected by the premise that AI could seamlessly merge multimodal data, contributing to disease detection, characterization of lesions, and monitoring treatment. This should add value, lead to lower health care costs, improve patient outcomes, or a combination of the aforementioned. Hospitals and health insurance companies benefit due to process optimization, resulting in lower costs and value-adding health care provision.

Frictions

With AI as a copilot assisting, and sometimes even taking over from the radiologist, the chance of discrepancies arises. On the one hand, participants indicated that AI is likely to make fewer mistakes than a radiologist and that AI should be trusted. On the other hand, in case a radiologist reaches a wrong conclusion

XSL•FO
RenderX

with the suggestion of AI, the radiologist should not be allowed to hide behind the output of the AI, which creates a friction between efficient use of AI and increased responsibility in case of mistakes. As a solution, the participants suggested that the radiologist always retains the ultimate responsibility for the conclusion of the radiology report. This still requires checking the output, but was it considered faster than going through multiple series of images? For patients, it was envisioned that AI would lead to increased demand for data, which could lead to unwanted use of personal data. To prevent this, data users should receive explicit and flexible consent, meaning that data can be used only for the purpose and duration that a patient allows. Furthermore, introducing AI in radiology was expected to increase the detection of incidental findings, many of which may lack clear clinical significance. This raises complex ethical challenges, particularly when balancing a patient's right not to know with the moral obligation to share potentially medically important findings. Participants in our study emphasized the importance of discussing the right not to know with the patients prior to imaging, as incidental findings can impose a psychological burden. They also stressed the necessity of keeping a human doctor in the loop, as the moral and clinical judgment required in these situations exceeded their expected capabilities of current AI systems and the demand for expertise.

Scenario 2. Health Without Walls: AI and the Rise of (Mobile) Multimodal Screening

Materialization

The participants developed 2 separate ideas of extramural health checks: in one, they envisioned an environmentally sustainable, electric bus touring across the country, in a similar setup as the current population screening for breast cancer prevention in the Netherlands. Such a bus enables screening for a wide variety of potential diseases, which potentially benefits from large-scale screening programs. When AI detects a high-risk scan, the scan is forwarded to a radiologist, who decides whether to follow up. Within this scenario, it is possible to focus on the scan, or to combine the screening with blood or any other test, thereby combining data from multiple streams into a single risk assessment.

The other idea was focused on an extramural imaging, diagnosis, and treatment facility, which can, for instance, be located in a shopping mall. Citizens can visit the facility for a self-initiated check or after being invited due to specific risk factors, such as age. AI is subsequently used for fast diagnosis and forwards data to a radiologist if follow-up is required. As a consequence, there is still value in human intervention and responsibility. This scenario requires data to be shared with hospitals, for example, through electronic health records. A step further is to fully automate the diagnostic workflow by using large models that use large retrospective datasets and individual risk factors to screen for patients at risk and subsequently immediately treat correctly identified patients.

Moral and Economic Considerations

Participants emphasized how important they consider trust and responsibility when thinking about AI and radiology. Questions raised by a social scientist and patient representative included, "Is it responsible to leave decisions to computers?" (g1-s2) and "Who should be responsible for this?" (g1-s1). The ownership of the process and of the data was questioned. According to participants, AI can have added value in this scenario, including alleviating burdens from hospital staff, improving patients' quality of life with an earlier start of the treatment, and equal access to health care.

Moreover, participants suggested that additional value can be calculated in added quality-adjusted life years, which can be translated into societal gains. The primary benefit of these scenarios may be that early diagnosis reduces long-term health care costs by shifting care from acute to preventive services. For funders, for example, the government, of large-scale AI-supported imaging-based screening programs, this shift may represent potentially favorable cost-benefit ratios where upfront investing may yield substantial savings and productivity over time. Additionally, commercial gain is possible, for example, for companies providing the bus with diagnostic hardware and software. Participants discussed various types of business models. Software companies responsible for evidence generation and validation can charge subscription- or use-based per license. Companies supplying scanners often operate on a purchase or lease model, and decentralized centers or scanner-on-tour suppliers could cover costs via direct-to-consumer supply or health insurance. For quality, competition between companies was regarded as crucial, and hospitals could consider purchasing an intramural diagnostic platform that supports multiple products from different vendors. This creates an overarching platform with competing products. Companies developing the software suites and the AI application that can be integrated into them should be prime movers here. Participants discussed the potential for the program to eventually become institutionalized, similar to breast cancer screening. Business models were seen as increasingly promising, especially once such programs began to generate large-scale data that can be leveraged to train data-hungry models.

> *if you have an AI algorithm that is CE-marked [Conformité Européenne], it doesn't learn from the data it receives, right? That's locked down. That's what the CE marking is for. It only goes into development mode when you have data to further develop it on, and then, you do an update, and you put that into the system. So, it's not like we always think that the AI just learns by itself.* [g1, s3]

Opportunities

Participants agreed that AI could offer an opportunity for the field that is in need of change, with the number of scans being requested rapidly increasing.

> *There is a report out right now that anticipates continuous growth in the years to come. Drivers are an aging population, more chronically ill patients, but also the increase in technological possibilities. At the same time, there is a growing demand coming from patients and the general population to have more insight into their health. Extramural health checks might be an answer to this demand.* [g1, s3]

XSL•FO
RenderX

Participants envisioned a future in which AI could more and better support the role of medical professionals. Decentralized health care facilities can play a large role in the health care system and help decrease the growth of private companies offering whole-body scans, which was deemed undesirable as disparities in income and wealth may translate into unequal access to health care services, potentially leading to disparities in health outcomes.

Frictions

Participants' discussions shed light on a couple of frictions. First, they questioned whether the model should be organized in a centralized or decentralized manner, what the implications of decentralized testing would be for referring people to a hospital, and how data collected may or could be shared. Decentralized centers would need to build support within the broader network of health care institutions, hospitals, companies, and so on, for data sharing; they needed correct licensing, and interoperability must be addressed. Access and management of big data can create suspicions among patients, as they may think: "your data no longer belongs to you" [g1-s2].

Another potential friction arose around new role divisions and associated responsibilities. Laboratory technicians may outsource more tasks and perform fewer tasks themselves, radiologists may be hired on a consultancy (freelance) basis, and new roles may emerge (such as radiology assistants). This could increase the distance between the radiologist and the referring physician, which could complicate responsibilities and task distribution. A third friction point regarded sustainability:

> *What are we going to do? Is that desirable? So, I think we have to assess upfront, do you need a scan? Yes or no?* [g1-s3]

This question was answered later on:

> *AI can help with that as well. They can look in advance: from, does this patient need a scan? And then you are working more sustainably, because yes, the most sustainable scan is no scan.* [g1, s1]

On the one hand, there was a need to manage this rising demand, with patients starting to demand scans for more certainty

> *Because there are, of course, patients that ask for a scan, they [patients] want certainty.* [g1, s5]

versus the question of whether you need to better manage this demand. This can be done either by enlarging the role of AI and maximizing the use of the capacity of available scanners or by downsizing the demand by asking whether a scan is actually necessary, which is a decision AI can support in this scenario.

Scenario 3. Early AD Insight: Proactive AI for a Longer Independent Life

Materialization

A group of participants chose to build their scenario around a concrete use case. The participants quickly decided to further develop their scenario around the idea that in the future, it becomes possible to predict, by analyzing either MRI data or only multimodal data, whether a citizen later in their life is likely to develop AD. This is achieved by screening citizens from a certain age group for the disease, either upon invitation or upon the citizens' request (similar to the previous screening scenario). The benefit of using AI in very early stages to detect or predict AD, participants mentioned, can be that (future) patients are helped and trained to live independently or with minimal help for longer. This approach should improve the quality of life of patients, alleviate some of the societal burden of caregivers for these future patients, and may lead to reduced costs:

> *Allowing Alzheimer's patients to function independently for as long as possible. [...] to keep support costs as low as possible.* [g3, s3]

Explainability was discussed as an important model aspect for patients. A patient representative argued that the model should be able to explain how it arrived at a certain decision. However, a participant with a background in pathology contextualized that this can be very task dependent. The researcher gave an example that was more related to her own field of expertise:

> *because for tumor detection, it [the AI] does not need to explain why something was flagged. ...Because we check it [the decision of the AI], so it is [in this case] AI-assisted. So I think: for this task, it does not need to explain what something is flagged for [...], because we check it [the flagged location].* [g3, s6]

The participants also briefly touched upon the business landscape, indicating that having multiple vendors would be a desirable aspect:

> *It's better to have multiple suppliers. Otherwise, we will have a monopolist.* [g3, s3]

Moral and Economic Considerations

Participants found it important that implementing AI would be a meaningful addition to radiology, benefiting the individual patient. Thorough validation of the AI product can lead to improved trust in its use. Trust can also be enhanced through (human) explanation of the output. This means that either the algorithm should be able to explain its own output or radiologists should provide an explanation when a textual one is unavailable. Furthermore, patients should always remain in charge of their data and consecutive algorithmic outputs. That is, the data acquired upon clinical indication and the output of the algorithms should be shared only upon their voluntary consent. The participants also believed that individual patients or a society that shares personal data should monetarily benefit from commercial products. In other words, according to the participants, commercial parties should not be allowed to earn profits over products forever and profits should eventually flow back to society in some form, given that the AI product is essentially enabled by voluntarily provided personal data.

Opportunities

The participants argued that AI should deliver a product that radiology currently cannot. That is, true added value is achieved when AI detects what a radiologist cannot see, or when AI predicts future events that are not yet visible as pathology at the time of imaging. Participants discussed that the role of AI

XSL•FO
RenderX

in this scenario is to assist with screening. Furthermore, predicting disease before it can manifest can lead to improved quality of life, specifically for AD. The participants highlighted that this is especially the case due to the human and emotional side of a progressive disease such as AD.

Frictions

The participants discussed the need for increased scanning capacity as a potentially negative consequence of AI. Granting all citizens the opportunity for screening for early detection of AD is costly, partly due to purchasing the scanners. However, using this additional scanning capacity can lead to improved quality of life for the patients, a lower societal burden for informal caregivers, and lower health care costs if cost-effectiveness of the screening is proven. The risk of incidental findings increases, though, which needs to be balanced with increased benefits of early detection. Whereas the industry employee argued that regulations can sometimes be experienced as burdensome, the patient representative found the regulations justifiable with respect to privacy, indicating that without strict regulations the patient representative may not share personal data at all.

Finally, the participants discussed potential consequences concerning insurance and autonomy: if AI predicts a future diagnosis such as AD, an insurance company may want access to those results, given that they may want to assess and price coverage accordingly. Also, the question was raised whether, “will they [some authority] say automatically that you move to an Alzheimer’s [care] house and you have to sell your home?” [g3, s5]. According to the participants, these aspects were undesirable, and these open questions warranted some serious thought.

Emerging Themes

Seven common themes emerged from the 3 future scenarios, along with their moral and economic considerations, opportunities, and frictions. Figure 1 summarizes the 7 emerging themes, recommendations based on the emerging themes, and 3 overarching themes.

XSL•FO
RenderX

Figure 1. Summary of the workshop, 7 emerging themes, corresponding recommendations, and 3 overarching themes. AI: artificial intelligence.

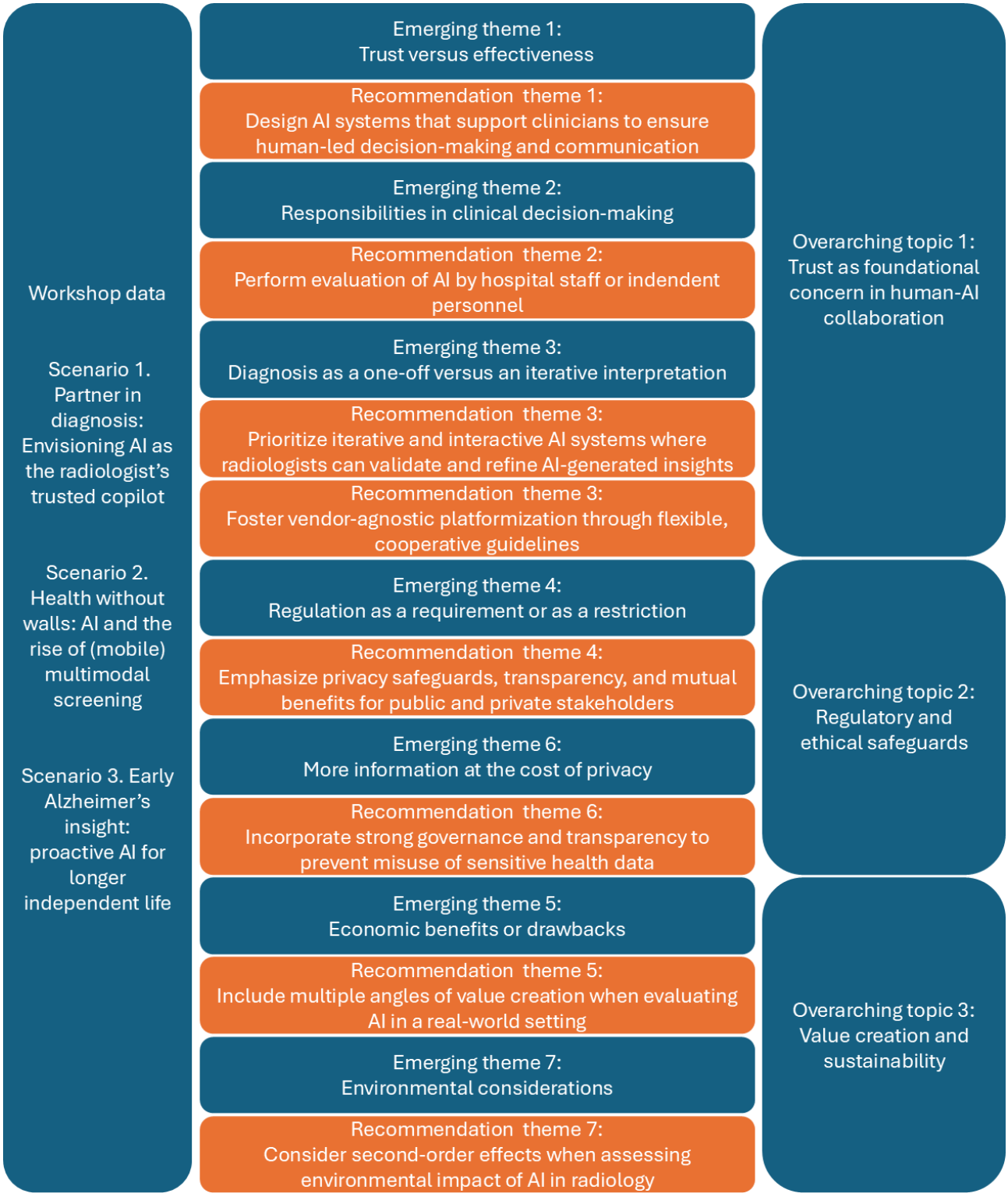

Emerging Theme 1: Trust Versus Effectiveness

Seemingly, trust in AI represents a critical trade-off. On the one hand, trust could enable AI to live up to its potential as a stand-alone technology to effectively address complex tasks, such as medical diagnosis or large-scale hospital workflow optimization. On the other hand, patient representatives feared risk for automation bias. Additionally, participants discussed that excessive reliance on AI systems can diminish human supervision, increasing the clinician's risk of automation bias and the erosion of critical decision-making capabilities. To break out of this trade-off and enhance trust, according to the participants, would often necessitate the adoption of interpretable and explainable AI systems, which radiologists can eventually check. Thus, achieving an optimal balance between trust and skepticism is essential to mitigate risks while harnessing the transformative potential of AI in radiology.

XSL•FO
RenderX

Emerging Theme 2: Responsibilities in Clinical Decision-Making

Implementing AI was considered a shared responsibility among stakeholders, each contributing distinct perspectives and expertise. Policy makers play a central role in establishing regulatory frameworks that guide legal-ethical boundaries for AI development and deployment. Industry players produce and scale practical AI applications within the established boundaries set by the aforementioned policy makers. However, when it comes to clinical decision-making, the participants foresee that medical specialists and health care workers will ultimately be responsible for the health and safety of their patients, and clinicians also feel that way. To reach this goal, it could be beneficial for medical specialists to cooperate with more technically oriented hospital personnel to perform quality assurance tests, particularly to ascertain that AI performs equally well on local datasets compared with how industry markets the performance of their products. Furthermore, according to the participants, there might be a conflict of interest when industry performs its own quality assurance. Thus, local testing may be a critical requirement for safe deployment and responsible use of AI in radiology and, according to the workshop participants, should be performed by independent hospital staff.

Emerging Theme 3: Diagnosis as a One-Off Versus an Iterative Interpretation

Given the rapid development of AI, radiological diagnosis is increasingly shaped by algorithmic output. These outputs are often treated as static and definitive, similar to finalized radiology reports. The static nature of AI's output presents 2 interrelated shortcomings. Radiological diagnosis is often not a fixed product but a dynamic process. It relies on expert interpretation and is influenced by evolving clinical information, such as neurological examinations and pathology results. Several participants noted that diagnosis often emerges through iterative reasoning and collaboration with other medical specialists. AI systems, which typically generate preanalyzed findings or diagnostic probabilities, risk oversimplifying this complexity. Participants suggested that integrating multimodal data could help AI better reflect current clinical practice. In addition, static outputs could lead to a greater demand for diagnostic certainty, potentially leading to unnecessary follow-up: by detecting more abnormalities and including findings that may never become clinically relevant, AI could lead to more scans, consultations, and interventions. A potential solution to both shortcomings that was suggested by participants is a "human-in-the-loop approach," where the radiologist remains responsible for interpreting AI outputs within the broader clinical context. Collaborative frameworks in which AI complements rather than replaces human expertise were seen as essential. These should include clear protocols for follow-up, shared decision-making with patients, and the development of transparent and explainable AI systems to support trust and appropriate clinical action.

Emerging Theme 4: Regulation as a Requirement or as a Restriction

Regulations play a dual role in technological innovation, acting both as a potential constraint and as a necessary enabler for progress, particularly in a fast-paced field such as AI in radiology. On the one hand, industry employees indicated that overly rigid or premature regulations can restrict innovation by imposing burdensome compliance requirements. On the other hand, patient representatives argued that well-designed regulatory frameworks are essential to ensure safety, fairness, and public trust—key requirements for adopting novel technologies. For example, as indicated by patient representatives, robust protection of patient privacy not only follows an ethical imperative but also creates a secure environment for patients to share their data. In other words, effective regulations are required for patients to feel safe when consenting to sharing their personal data, which is the cornerstone for all AI developments. Adaptive and dynamic regulatory approaches, capable of evolving alongside technological advances, were thus considered crucial for maintaining this balance. Following our workshop results, participants regarded effective regulation as an enabler rather than an obstacle to innovation, because it provides the foundation for trust necessary for groundbreaking technologies to thrive responsibly.

Emerging Theme 5: Economic Benefits or Drawbacks

The integration of AI in radiology presents significant economic opportunities and challenges. Following initial investments in hardware infrastructures and software packages, there is a wide variety of potential applications on a hospital level, societal level, and vendor level that could yield a net economic benefit. Participants emphasized that AI could significantly improve operational efficiency within hospitals. By automating routine tasks, such as appointment scheduling and administrative workflows, AI has the potential to reduce labor costs and streamline internal processes. These improvements may lead to better resource allocation and increased throughput, contributing to overall cost-effectiveness in clinical operations. At the societal level, the economic benefits of AI in radiology are diverse and potentially far-reaching. Participants pointed to several ways in which AI could contribute to more efficient health care delivery and improved population health. For example, timely diagnosis may enable earlier initiation of therapy, which can improve outcomes and reduce the need for prolonged or intensive treatment. Similarly, more personalized care, which could be supported by AI's ability to integrate and analyze complex data, may help avoid unnecessary interventions and enhance the overall effectiveness of health care services. These developments could contribute to long-term cost savings and a more sustainable use of health care resources. However, participants also cautioned that increased diagnostic sensitivity might lead to more incidental findings, which could raise demand for follow-up care and specialist input, potentially increasing overall health care expenditures. According to some participants, the responsibility for initial economic value assessments of AI systems lies with the vendors, who should estimate whether their solutions offer sufficient value before deployment. However, participants emphasized that independent researchers should conduct postdeployment evaluations. Similar to quality assurance processes, this separation is essential to avoid conflicts of interest and ensure unbiased assessments of economic performance.

XSL•FO
RenderX

Emerging Theme 6: More Information at the Cost of Privacy

Unintended consequences from AI might arise when the decision becomes binding rather than advisory. This decision could drastically affect a patient's life. In the AD case, the participants discussed the extreme example of mandatory compliance with the AI decision for future patients when AD is predicted. Furthermore, if (by obligation) shared with insurers, these screening programs could have negative consequences for monthly health insurance premiums, thus penalizing individual citizens financially for factors beyond their control, which raises privacy, fairness, and proportionality concerns. Or, in the extreme case of extrapolating this issue, insurance companies may increase costs for patients at risk or deny (future) patients access to (additional) health and property insurances at all. Regulation should therefore be in place to prevent such a scenario from materializing. Moreover, patients have the right to not know. The participants agreed that this moral dilemma might be quite difficult for an AI to weigh, where there needs to be a good balance between medical necessity and patients' rights and preferences. To counter this unintended consequence, human intervention might be the only solution.

Emerging Theme 7: Environmental Considerations

Finally, it was discussed that AI is an energy-demanding technology. Combining AI with radiology for screening means that scarce resources might be used for citizens who may not (yet) express any disease symptoms. Thus, the participants discussed potential benefits for health care and the economy of AI in radiology should also be carefully weighed against the potential environmental impact. The monetary gain and gain in quality of life after the introduction of AI to radiology should therefore be proportionate to the environmental impact of additional interventions, such as screening in combination with AI.

Discussion

Principal Findings

In this manuscript, we report and reflect on a qualitative cocreation workshop we designed and conducted, inspired by responsible innovation approach, aiming to explore future scenarios for AI in radiology, their opportunities and frictions, by involving a variety of stakeholders in the field of radiology. Interestingly, in imagining futures of AI in radiology, the majority of the groups, in one way or another, came up with a scenario to upscale radiology via screening or extramural imaging to early detect and prevent pathology from becoming a disease. However, there were differences in the imagined role of these applications, and of actors in the field, including governments, insurers, technology developers, and citizens. By including a wide variety of topics, we aimed to broaden the participants' scope of thinking. Resultingly, our participants discussed moral and economic considerations, envisioned opportunities, and foresaw frictions arising related to interconnected topics on trust, responsibilities, workflows, regulatory requirements, dealing with economic benefits, and unintended consequences.

Comparison With Prior Work

Trust Versus Effectiveness

In our analysis, we observed trust in AI as a complex topic the participants linked to, for example, performance and explainability. The participants perceived a paradoxical relationship between trust and effectiveness: the more trust they were willing to place in AI, the more AI could potentially operate as a stand-alone technology in radiology and fulfill its promise of improving clinical outcomes. Yet, this very trust introduces ethical tensions. As described by Singh et al [77], the ethical paradox in medical practices arises when efforts to enhance health care through AI risk diminishing the humanistic aspect of care, such as empathy and interpersonal interactions, thereby potentially provoking skepticism and decreasing trust. Reflecting on scenarios 1 and 2, AI could be deployed for triaging or identifying patients at risk. As such, AI applications may allow clinicians to focus their limited time on more complex and emotionally sensitive cases, potentially improving overall effectiveness of care and patient well-being, which were expectations that were reported earlier as well [37]. However, fully trusting AI may lead to dependency on AI or potential loss of human oversight. Similar concerns were also shared during the cocreation workshop. Patient representatives, among other stakeholders, confirmed that a human-in-the-loop was, in their opinion, still required. This idea was also shared in earlier research [63,78-80], although in other reports, patients have also indicated that autonomous AI could be acceptable on the condition that AI outperforms a radiologist [18]. Yet, patient representatives and other stakeholders seemed to have a positive attitude toward AI-assisted triaging, especially because the ultimate decision remains with a radiologist. The explainability of the AI-based results of the triaging could further help gain clinicians' confidence [53]. Discussions suggest that following these conditions, applications of AI in radiology may contribute to enhancing efficiency without compromising safety or trust. Based on the links between effectiveness, trust, explainability, and the humanistic aspect of care, we recommend designing AI systems that support clinicians, ensuring human-led decision-making and communication, while leveraging AI for effectiveness.

Responsibilities in Clinical Decision-Making

Participants in the workshop highlighted that there should not be conflicts of interest in the case of responsibilities. In other words, in environments where people must act responsibly to, for example, safeguard patients' health, one should not be incentivized to act differently. For this reason, some participants emphasized that generating evidence for AI in early stages of development and real-world testing of AI should be performed by 2 distinct parties. Whereas industry was seen as responsible for product (co)development and early evidence generation, real-world implementation, thorough (local) testing, and longitudinal evaluation should be performed by independent hospital personnel, potentially in collaboration with more technically oriented personnel such as clinical physicists or technical physicians [81]. Furthermore, there was a seeming distrust of commercial parties to perform their own prospective evaluation. Earlier research focused on *how* to perform

XSL•FO
RenderX

postinstallation evaluation [22,26,82-84]. However, this research did not dive into *who* should do so and what the possible impact on interconnected themes such as trust could be. Specifically, industry involvement should be limited in prospective evaluation, such that thorough evaluation can be performed with as little conflict of interest as possible. Our research demonstrates preliminary evidence that independent evaluation of AI in a clinical setting is pivotal to enhancing trust, and that trust can be achieved only by proper distribution of responsibilities along the innovation, installation, and evaluation chain. Therefore, we recommend that evaluation of AI is performed by hospital staff or independent external parties.

Diagnosis as a One-Off Versus an Iterative Interpretation

The proposed function of the AI within the radiology system varied within our workshop, ranging from screening to acting as a copilot for the radiologist. AI tools in medical imaging have traditionally functioned as narrow, outcome-focused instruments, for example, detecting polyps, segmenting tumors, or flagging anomalies in isolation [85]. However, radiology is fundamentally a process of evolving clinical interpretation, where patients' histories, laboratory results, physical examinations, and ongoing changes in condition contextualize each series of images [86,87]. Multimodal data integration could enable AI to move beyond static predictions and could facilitate evolving model outputs [88]. Furthermore, the inclusion of large language models (LLMs) could provide narrative understanding, enabling AI to synthesize diverse inputs into coherent clinical insights [89-91]. In turn, a radiologist can check the output of the LLM or interact with the LLM, which could lead to further explainability and could foster trust, which also links back to emerging theme 1. This convergence transforms AI from a tool that offers answers into one that participates in reasoning. Importantly, this shift could allow AI to align more closely with the clinician's cognitive workflow, adapting dynamically as new data come in [89-91]. It could also open the door to more explainable and human-centered AI systems that support, not replace, clinical judgment, as was also proposed earlier [63,78-80]. Thus, we recommend prioritizing iterative and interactive AI systems where radiologists can validate and refine AI-generated insights, ensuring that AI acts as a reasoning partner rather than a black box tool.

Moreover, the participants deemed the integrability of AI tools within existing hospital systems as a facilitator for successful adoption, as was also reported in earlier research [59]. Integrating these capabilities into unified platforms, rather than isolated tools, could thus further aid in the commercialization of AI in radiology. This process of platformization supports scalability, continuous learning, and interoperability [69], potentially enabling a shift from a one-off solution to comprehensive decision support ecosystems that are clinically and economically sustainable. Finally, participants regarded having multiple vendors as beneficial to the health care system, encouraging competition in either performance or costs, potentially benefiting patients. To support this, platforms should remain vendor-neutral, allowing easy integration from multiple providers. Whereas earlier research recommends four components: (1) (automatic) system triggering, (2) retrieval of relevant imaging series, (3) application of algorithm pipeline, and (4) availability of the results in Picture Archiving and Communication System of dedicated viewing software [92], we recommend conducting more research in the direction of cooperative guidelines that allow vendors to provide innovative solutions while remaining flexible for competition, potentially including regulatory sandbox environments, adaptive regulations, and value-based procurement.

Regulation as a Requirement or as a Restriction

In some discussions, the role of regulations came forward. While industry employees regarded regulations as a potentially impeding innovation, patient representatives regarded regulations as a requirement to share data. This paradoxical friction was also described in a recently published paper. According to Singh et al [77], the solution seems to be to strike the right balance between data accessibility and the obligation to safeguard patient confidentiality and to adhere to regulatory frameworks. This finding, previously reported among medical practitioners, was also identified in patient representatives, further confirming and reinforcing its broader relevance. The data paradox is quite complex: on the one hand, patient representatives in our workshop mostly preferred to retain ownership of their data. On the other hand, industry requires a diverse and up-to-date dataset, such that the AI products they develop generalize sufficiently well on various populations and do not suffer from data drift [40,46,51]. In light of this data paradox, future work should explore how hospital-private sector collaborations can be structured to enable responsible data sharing, with strong privacy safeguards, transparency, and mutual benefit for both public and private stakeholders. Recent global frameworks, such as the World Health Organization's regulatory pillars for AI in health [93], the EU AI Act's classification of medical imaging AI as "high-risk" [94], and international consensus initiatives such as FUTURE-AI [26], emphasize that striking this balance is not only desirable but also essential for compliance and trustworthiness. These frameworks advocate transparency, human oversight, and robust governance, which can guide collaborations toward ethical and legally sound data sharing.

Economic Benefits or Drawbacks

During the workshop, participants raised the question of whether it was desirable for companies to earn money using voluntarily provided individual patient data without patient populations benefiting. A competitive landscape was described by the participants that could solve these issues. First of all, multiple vendors compete for the best product. The participants argued that this aspect should lead to better health outcomes for individual patients, linking to earlier argued ethical justification of the use of patient data if companies contribute to public health benefits [95]. In other words, the profits earned using patient data are reinvested in the companies to improve existing products or to create new products, which benefit both public and private stakeholders. Second, vendors can compete for the lowest price, which benefits the health care system as a whole, as price competition can lead to lower health care costs. At the same time, a diverse vendor landscape may help mitigate risks associated with vendor lock-in [96], ensuring that health care

XSL•FO
RenderX

providers retain flexibility in choosing and switching between solutions as technology evolves and needs change. However, without more studies rigorously evaluating AI systems in real-world clinical settings, it is difficult to substantiate claims of added value [74]. We therefore advocate for more studies that evaluate AI in a real-world setting and that include multiple angles of value creation, such as decreased reporting times, added number of relevant incidental findings, and the prevention of unnecessary follow-up scans or interventions. This should lead to elaborated business models for both vendors (suppliers) and hospitals (buyers). Upon positive results, these studies could further facilitate AI adoption in clinics and provide vendors with useful insights about the possibility to generate revenue with their products. One example comes from Brix et al [11], who found that AI-based image reconstruction could lead to improved patient throughput, thereby obviating the need for an additional scanner. Such insights can help vendors to demonstrate the added value of their AI solutions, which may justify premium pricing and unlock new business models, especially when validated in real-world clinical environments.

More Information at the Cost of Privacy

Participants worried about data privacy, especially over the sharing of diagnostic and prognostic information inferred from AI with insurance companies, hinting that insurance companies might misuse this information. A recent review by Botha et al [97] also endeavored to examine the implication of AI tools on insurance. Whereas many aspects concerning insurance coverage were discussed, including the potential rise in insurance costs with the introduction of AI in clinical settings, second-order effects such as sharing information of future disease with insurance companies were underrepresented [97]. A perspective paper from 2018 by Char et al [60] warned of this looming issue, and our results demonstrate that these worries are still unresolved. Thus, a deeper dive into this specific topic seems warranted. Emerging literature confirms these concerns, highlighting risks of insurance-related discrimination from predictive analytics [98]. Regulatory frameworks such as World Health Organization's guidance [93] and the EU AI Act [94] stress the need for strong governance and transparency to prevent misuse of sensitive health data. Incorporating these principles into practice could help mitigate ethical risks while enabling responsible innovation.

Environmental Considerations

Another worry that came forward was the energy consumption of AI, which according to the participants potentially makes it an environmentally unsustainable solution. Whereas it is likely a correct assumption that adding AI to the radiology workflow increases energy consumption compared with a workflow without AI, efforts to make both algorithms and hardware more energy-efficient might help reduce the energy consumption for each inference [99]. Furthermore, AI could aid in reducing environmentally harmful gadolinium-based contrast agent use [100], or can reconstruct images from accelerated (and thus less time-consuming and less energy-consuming) MRI acquisitions [11,101], which both potentially have a net positive effect on the environment, as it may result in fewer MRI systems in use and less energy consumption per scan. Thus, whereas the participants' worries are justifiable, these second-order effects also need to be accounted for when assessing environmental impact for AI in radiology.

Limitations and Future Directions

The main limitation of this study is that we draw on a singular empirical case study set in a Dutch tertiary care center. Future studies should revisit our research questions in different countries and look at other care settings, including trans- and extramural settings. These other care centers are of special importance, since multiple participants independently hinted at the potential for extramural imaging facilities. Also, no insurance company employees could attend the workshop, despite efforts to include at least one. Insurance companies are a key stakeholder in the field, so their perspectives are potentially valuable in discussions on the demand side (reimbursement and assessment of added therapeutic value). Next, in future research, we recommend including patient representatives again, as their insights were invaluable to this work. Finally, by checking the audio recordings in full by 2 of the authors (MBS and SH) and refining transcripts and translations, we aimed to minimize potential nuance loss. Although some loss of data could still have occurred, these steps helped preserve the richness of the participants' contributions as much as possible.

Conclusions

With this paper, we aimed to explore wider stakeholder perspectives on futures of AI in radiology and to discuss the moral and economic considerations, opportunities, and frictions. The interaction between participants resulted in scenarios about an AI copilot, AI-enhanced extramural scanning, and AI-based AD screening. It prompted debates on perspectives that may not have emerged from isolated stakeholder analyses. Seven themes emerged from the analysis: (1) trust and efficiency of AI technologies, (2) responsibilities in clinical decision-making when AI is involved, (3) diagnosis as a one-off versus an iterative process, (4) regulations as a requirement or a restriction, (5) economic benefits or drawbacks, (6) trade-off between amount of information required and patient privacy, and (7) environmental considerations. In reflecting on the 7 emerging themes from the multistakeholder cocreation workshop, 3 overarching topics can be identified that help reflect on the broader implications of AI in radiology.

First, trust emerged as a foundational concern in human-AI collaboration. Participants emphasized that in their opinion, AI should support, not replace, clinical decision-making. Trust was seen as conditional for AI to be applied in radiology efficiently, meaning that explainability and the continued presence of a human-in-the-loop are deemed important. This human-machine model aligns with radiological diagnoses of iterative nature and was considered essential for safe and ethical implementation. Second, regulatory and ethical safeguards were central to discussions on data use, responsibility, and privacy. Stakeholders expressed diverging views: while industry employees saw regulation as a potential restriction, patient representatives viewed regulations as a requirement for data sharing. The friction between innovation and regulation can be dealt with through adaptive regulatory frameworks that evolve alongside

XSL•FO
RenderX

technological capabilities while safeguarding patient autonomy and algorithmic fairness. A third topic relates to value creation and sustainability. Participants discussed the potential for AI to improve efficiency, reduce costs, and enhance diagnostic accuracy. However, concerns were raised about incidental findings, data monetization, viability of business models, and the energy demand of AI systems. Platformization and vendor-neutral integration were seen as a way to foster innovation while maintaining affordability and scalability. Moreover, participants advocated for competitive ecosystems where multiple vendors could contribute to better outcomes and lower health care costs.

Together, these overarching topics offer a lens through which future research and policy can address the complex interplay of technological potential, ethical responsibility, and systemic sustainability in AI-driven radiology. Future research is needed to investigate the generalizability of our findings in different countries and settings, or to address open questions on real-world economic benefit and sustainability.

Acknowledgments

The authors would like to thank their participants for their time and invaluable insights. Without them, their study would not have been possible.

Funding

This publication is part of the project "Responsible implementation of quantitative MRI" (with project number 18749 of the research program HTSM MVI top-up, which is financed by the Dutch Research Council [NWO]).

Data Availability

The datasets analyzed during this study are not publicly available due to data policies. They

may be available from the corresponding author (MBS) per reasonable request.

Authors' Contributions

MBS, SH, AK, and WB contributed to methodology design, formal data analysis, and draft of original manuscript. All authors participated in study conceptualization, review, and revision of manuscript.

Conflicts of Interest

None declared

Multimedia Appendix 1
Worksheet canvas for exercise 2. The original worksheet during the workshop was in Dutch.
[PNG File, 199 KB - jmir_v28i1e83407_app1.png]

Multimedia Appendix 2
Worksheet canvas for exercise 3. Mind that, again, the original worksheet during the workshop was in Dutch.
[PNG File, 105 KB - jmir_v28i1e83407_app2.png]

Multimedia Appendix 3
Quotebook.
[DOCX File, 19 KB - jmir_v28i1e83407_app3.docx]

References

1. Hassankhani A, Amoukhteh M, Valizadeh P, Jannatdoust P, Sabeghi P, Gholamrezanezhad A. Radiology as a specialty in the era of artificial intelligence: a systematic review and meta-analysis on medical students, radiology trainees, and radiologists. Acad Radiol 2024 Jan;31(1):306-321. [doi: 10.1016/j.acra.2023.05.024]
2. Goisauf M, Cano Abadía M. Ethics of AI in radiology: a review of ethical and societal implications. Front Big Data 2022;5:850383. [doi: 10.3389/fdata.2022.850383] [Medline: 35910490]
3. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018 Aug;18(8):500-510. [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
4. Mazurowski MA. Artificial intelligence may cause a significant disruption to the radiology workforce. J Am Coll Radiol 2019 Aug;16(8):1077-1082. [doi: 10.1016/j.jacr.2019.01.026] [Medline: 30975611]
5. Zhao T, Gu Y, Yang J, et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. Nat Methods 2025 Jan;22(1):166-176. [doi: 10.1038/s41592-024-02499-w]

XSL•FO
RenderX

6. Rezazade Mehrizi MH, van Ooijen P, Homan M. Applications of artificial intelligence (AI) in diagnostic radiology: a technography study. Eur Radiol 2021 Apr;31(4):1805-1811. [doi: 10.1007/s00330-020-07230-9] [Medline: 32945967]
7. Drogt J, Milota M, Veldhuis W, Vos S, Jongsma K. The promise of AI for image-driven medicine: qualitative interview study of radiologists' and pathologists' perspectives. JMIR Hum Factors 2024 Nov 21;11:e52514. [doi: 10.2196/52514] [Medline: 39570627]
8. Najjar R. Redefining radiology: a review of artificial intelligence integration in medical imaging. Diagnostics (Basel) 2023 Aug 25;13(17):2760. [doi: 10.3390/diagnostics13172760] [Medline: 37685300]
9. Fasterholdt I, Naghavi-Behzad M, Rasmussen BSB, et al. Value assessment of artificial intelligence in medical imaging: a scoping review. BMC Med Imaging 2022 Oct 31;22(1):187. [doi: 10.1186/s12880-022-00918-y] [Medline: 36316665]
10. Hofmann P, Oesterle S, Rust P, Urbach N. Machine learning approaches along the radiology value chain—rethinking value propositions. Presented at: Proceedings of the 27th European Conference on Information Systems (ECIS) Stockholm & Uppsala; Jun 8-14, 2019 URL: https://aisel.aisnet.org/ecis2019_rp/158/ [accessed 2026-02-09]
11. Brix MAK, Järvinen J, Bode MK, et al. Financial impact of incorporating deep learning reconstruction into magnetic resonance imaging routine. Eur J Radiol 2024 Jun;175:111434. [doi: 10.1016/j.ejrad.2024.111434] [Medline: 38520806]
12. Bharadwaj P, Nicola L, Breau-Brunel M, et al. Unlocking the value: quantifying the return on investment of hospital artificial intelligence. J Am Coll Radiol 2024 Oct;21(10):1677-1685. [doi: 10.1016/j.jacr.2024.02.034] [Medline: 38499053]
13. Sharon T. When digital health meets digital capitalism, how many common goods are at stake? Big Data Soc 2018 Jul;5(2). [doi: 10.1177/2053951718819032]
14. Raz A, Minari J, Schicktanz S, Sharon T, Werner-Felmayer G. Editorial: data-intensive medicine and healthcare: ethical and social implications in the era of artificial intelligence and automated decision-making. Front Genet 2023 Sep 7;14. [doi: 10.3389/fgene.2023.1280344]
15. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. J Med Internet Res 2020 Jun 19;22(6):e15154. [doi: 10.2196/15154] [Medline: 32558657]
16. Shevtsova D, Ahmed A, Boot IWA, et al. Trust in and acceptance of artificial intelligence applications in medicine: mixed methods study. JMIR Hum Factors 2024 Jan 17;11:e47031. [doi: 10.2196/47031] [Medline: 38231544]
17. Rivard L, Lehoux P. When desirability and feasibility go hand in hand: innovators' perspectives on what is and is not responsible innovation in health. Journal of Responsible Innovation 2020 Jan 2;7(1):76-95. [doi: 10.1080/23299460.2019.1622952]
18. Fransen SJ, Kwee TC, Rouw D, et al. Patient perspectives on the use of artificial intelligence in prostate cancer diagnosis on MRI. Eur Radiol 2024 Aug 14;35(2):769-775. [doi: 10.1007/s00330-024-11012-y]
19. Hajiheydari N, Delgosha MS, Saheb T. AI in medical diagnosis: a contextualised study of patient motivations and concerns. Soc Sci Med 2025 Apr;371:117850. [doi: 10.1016/j.socscimed.2025.117850] [Medline: 40081168]
20. Ongena YP, Haan M, Yakar D, Kwee TC. Patients' views on the implementation of artificial intelligence in radiology: development and validation of a standardized questionnaire. Eur Radiol 2020 Feb;30(2):1033-1040. [doi: 10.1007/s00330-019-06486-0] [Medline: 31705254]
21. Becker CD, Kotter E, Fournier L, Martí-Bonmatí L, European Society of Radiology (ESR). Current practical experience with artificial intelligence in clinical radiology: a survey of the European Society of Radiology. Insights Imaging 2022 Dec;13(1):107. [doi: 10.1186/s13244-022-01247-y]
22. Brady AP, Allen B, Chong J, et al. Developing, purchasing, implementing and monitoring AI tools in radiology: practical considerations. A multi-society statement from the ACR, CAR, ESR, RANZCR & RSNA. Insights Imaging 2024;15(1):16. [doi: 10.1186/s13244-023-01541-3]
23. Owen R, Stilgoe J, Macnaghten P, Gorman M, Fisher E, Guston D. A framework for responsible innovation. In: Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society: Wiley; 2013:27-50. [doi: 10.1002/9781118551424]
24. Stilgoe J, Owen R, Macnaghten P. Developing a framework for responsible innovation. Res Policy 2013 Nov;42(9):1568-1580. [doi: 10.1016/j.respol.2013.05.008]
25. Vo V, Chen G, Aquino YSJ, Carter SM, Do QN, Woode ME. Multi-stakeholder preferences for the use of artificial intelligence in healthcare: A systematic review and thematic analysis. Soc Sci Med 2023 Dec;338:116357. [doi: 10.1016/j.socscimed.2023.116357] [Medline: 37949020]
26. Lekadir K, Frangi AF, Porras AR, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ 2025 Feb 5;388:e081554. [doi: 10.1136/bmj-2024-081554] [Medline: 39909534]
27. Kim JP, Ryan K, Kasun M, Hogg J, Dunn LB, Roberts LW. Physicians' and machine learning researchers' perspectives on ethical issues in the early development of clinical machine learning tools: qualitative interview study. JMIR AI 2023 Oct 30;2:e47449. [doi: 10.2196/47449] [Medline: 38875536]
28. van Hees SV, Schilder MB, Keyser A, Sbrizzi A, Kleinloog JPD, Boon WPC. Exploring scenarios for implementing fast quantitative MRI. Eur J Radiol Open 2025 Jun;14:100658. [doi: 10.1016/j.ejro.2025.100658] [Medline: 40469716]
29. Strohm L, Hehakaya C, Ranschaert ER, Boon WPC, Moors EHM. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. Eur Radiol 2020 Oct;30(10):5525-5532. [doi: 10.1007/s00330-020-06946-y] [Medline: 32458173]

XSL•FO
RenderX

30. Haque AB, Islam A, Mikalef P. Explainable Artificial Intelligence (XAI) from a user perspective: a synthesis of prior literature and problematizing avenues for future research. Technol Forecast Soc Change 2023 Jan;186:122120. [doi: 10.1016/j.techfore.2022.122120]
31. Yang L, Ene IC, Arabi Belaghi R, Koff D, Stein N, Santaguida P. Stakeholders' perspectives on the future of artificial intelligence in radiology: a scoping review. Eur Radiol 2022 Mar;32(3):1477-1495. [doi: 10.1007/s00330-021-08214-z]
32. Lukkien DRM, Ipakchian Askari S, Stolwijk NE, et al. Making co-design more responsible: case study on the development of an AI-based decision support system in dementia care. JMIR Hum Factors 2024 Jul 31;11:e55961. [doi: 10.2196/55961]
33. Zondag AGM, Rozestraten R, Grimmelikhuijsen SG, et al. The effect of artificial intelligence on patient-physician trust: cross-sectional vignette study. J Med Internet Res 2024 May 28;26:e50853. [doi: 10.2196/50853]
34. Pacifico Silva H, Lehoux P, Miller FA, Denis JL. Introducing responsible innovation in health: a policy-oriented framework. Health Res Policy Syst 2018 Sep 10;16(1):90. [doi: 10.1186/s12961-018-0362-5] [Medline: 30200985]
35. Greenhalgh T, Thorne S, Malterud K. Time to challenge the spurious hierarchy of systematic over narrative reviews? Eur J Clin Investigation 2018 Jun;48(6). [doi: 10.1111/eci.12931]
36. Sukhera J. Narrative reviews: flexible, rigorous, and practical. J Grad Med Educ 2022 Aug 1;14(4):414-417. [doi: 10.4300/JGME-D-22-00480.1]
37. Gundlack J, Negash S, Thiel C, et al. Artificial intelligence in medical care—patients' perceptions on caregiving relationships and ethics: a qualitative study. Health Expect 2025 Apr;28(2):e70216. [doi: 10.1111/hex.70216] [Medline: 40094179]
38. Baghdadi LR, Mobeirek AA, Alhudaithi DR, et al. Patients' attitudes toward the use of artificial intelligence as a diagnostic tool in radiology in Saudi Arabia: cross-sectional study. JMIR Hum Factors 2024;11:e53108. [doi: 10.2196/53108]
39. Haan M, Ongena YP, Hommes S, Kwee TC, Yakar D. A qualitative study to understand patient perspective on the use of artificial intelligence in radiology. J Am Coll Radiol 2019 Oct;16(10):1416-1419. [doi: 10.1016/j.jacr.2018.12.043] [Medline: 30878311]
40. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan;28(1):31-38. [doi: 10.1038/s41591-021-01614-0]
41. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018 Oct;2(10):719-731. [doi: 10.1038/s41551-018-0305-z] [Medline: 31015651]
42. Trivedi H. The business of artificial intelligence in radiology has little to do with radiologists. J Am Coll Radiol 2022 Apr;19(4):564-566. [doi: 10.1016/j.jacr.2022.01.006] [Medline: 35176248]
43. Mayo RC, Chang Sen LQ, Leung JWT. Financing artificial intelligence in medical imaging: show me the money. J Am Coll Radiol 2020 Jan;17(1 Pt B):175-177. [doi: 10.1016/j.jacr.2019.07.004] [Medline: 31918877]
44. Gille F, Jobin A, Ienca M. What we talk about when we talk about trust: theory of trust for AI in healthcare. Intelligence-Based Medicine 2020 Nov;1-2:100001. [doi: 10.1016/j.ibmed.2020.100001]
45. Steerling E, Siira E, Nilsen P, Svedberg P, Nygren J. Implementing AI in healthcare-the relevance of trust: a scoping review. Front Health Serv 2023;3:1211150. [doi: 10.3389/frhs.2023.1211150] [Medline: 37693234]
46. Fernandez-Quilez A. Deep learning in radiology: ethics of data and on the value of algorithm transparency, interpretability and explainability. AI Ethics 2023 Feb;3(1):257-265. [doi: 10.1007/s43681-022-00161-9]
47. Daly SJ, Wiewiora A, Hearn G. Shifting attitudes and trust in AI: influences on organizational AI adoption. Technol Forecast Soc Change 2025 Jun;215:124108. [doi: 10.1016/j.techfore.2025.124108]
48. Tahtali MA, Snijders CCP, Dirne C, Le Blanc PM. Prioritizing trust in podiatrists' preference for AI in supportive roles over diagnostic roles in health care: qualitative interview and focus group study. JMIR Hum Factors 2025 Feb 21;12:e59010. [doi: 10.2196/59010] [Medline: 39983118]
49. Aldhafeeri FM. Navigating the ethical landscape of artificial intelligence in radiography: a cross-sectional study of radiographers' perspectives. BMC Med Ethics 2024 May 11;25(1):52. [doi: 10.1186/s12910-024-01052-w] [Medline: 38734602]
50. Linguraru MG, Bakas S, Aboian M, et al. Clinical, cultural, computational, and regulatory considerations to deploy AI in radiology: perspectives of RSNA and MICCAI experts. Radiol Artif Intell 2024 Jul;6(4):e240225. [doi: 10.1148/ryai.240225] [Medline: 38984986]
51. Chen RJ, Wang JJ, Williamson DFK, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat Biomed Eng 2023 Jun;7(6):719-742. [doi: 10.1038/s41551-023-01056-8] [Medline: 37380750]
52. Tejani AS, Ng YS, Xi Y, Rayan JC. Understanding and mitigating bias in imaging artificial intelligence. Radiographics 2024 May;44(5):e230067. [doi: 10.1148/rg.230067] [Medline: 38635456]
53. Din M, Daga K, Saoud J, et al. Clinicians' perspectives on the use of artificial intelligence to triage MRI brain scans. Eur J Radiol 2025 Feb;183:111921. [doi: 10.1016/j.ejrad.2025.111921] [Medline: 39805194]
54. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1):310. [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
55. Siala H, Wang Y. SHIFTing artificial intelligence to be responsible in healthcare: a systematic review. Soc Sci Med 2022 Mar;296:114782. [doi: 10.1016/j.socscimed.2022.114782] [Medline: 35152047]

XSL•FO
RenderX

56. Martiniussen MA, Larsen M, Larsen ASF, et al. Norwegian radiologists' expectations of artificial intelligence in mammographic screening—a cross-sectional survey. Eur J Radiol 2023 Oct;167:111061. [doi: 10.1016/j.ejrad.2023.111061] [Medline: 37657381]
57. Fehr J, Citro B, Malpani R, Lippert C, Madai VI. A trustworthy AI reality-check: the lack of transparency of artificial intelligence products in healthcare. Front Digit Health 2024 Feb 20;6. [doi: 10.3389/fdgth.2024.1267290]
58. Martin C, DeStefano K, Haran H, et al. The ethical considerations including inclusion and biases, data protection, and proper implementation among AI in radiology and potential implications. Intelligence-Based Medicine 2022;6:100073. [doi: 10.1016/j.ibmed.2022.100073]
59. Abell B, Naicker S, Rodwell D, et al. Identifying barriers and facilitators to successful implementation of computerized clinical decision support systems in hospitals: a NASSS framework-informed scoping review. Implementation Sci 2023 Jul 26;18(1):32. [doi: 10.1186/s13012-023-01287-y]
60. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. N Engl J Med 2018 Mar 15;378(11):981-983. [doi: 10.1056/NEJMp1714229] [Medline: 29539284]
61. van Hees S, Kleinloog JPD, Sbrizzi A, Boon WPC. Static Future Technologies, Dynamic Professionalism — Co-creating Future Scenarios in Medical Imaging Practices. Postdigit Sci Educ 2024 Mar;6(1):135-153. [doi: 10.1007/s42438-023-00444-2]
62. Lombi L, Rossero E. How artificial intelligence is reshaping the autonomy and boundary work of radiologists. A qualitative study. Sociol Health Illn 2024 Feb;46(2):200-218. [doi: 10.1111/1467-9566.13702] [Medline: 37573551]
63. Scheek D, Rezazade Mehrizi M, Ranschaert E. Radiologists in the loop: the roles of radiologists in the development of AI applications. Eur Radiol 2021 Oct;31(10):7960-7968. [doi: 10.1007/s00330-021-07879-w]
64. Stogiannos N, Walsh G, Ohene-Botwe B, et al. R-AI-diographers: a European survey on perceived impact of AI on professional identity, careers, and radiographers' roles. Insights Imaging 2025 Feb 17;16(1):43. [doi: 10.1186/s13244-025-01918-6] [Medline: 39962024]
65. van Leeuwen KG, Schalekamp S, Rutten M, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol 2021 Jun;31(6):3797-3804. [doi: 10.1007/s00330-021-07892-z] [Medline: 33856519]
66. Parikh RB, Helmchen LA. Paying for artificial intelligence in medicine. NPJ Digit Med 2022 May 20;5(1):63. [doi: 10.1038/s41746-022-00609-6] [Medline: 35595986]
67. Yu J, Zhang J, Sengoku S. Innovation process and industrial system of US Food and Drug Administration-approved software as a medical device: review and content analysis. J Med Internet Res 2023 Nov 24;25:e47505. [doi: 10.2196/47505] [Medline: 37999948]
68. Joshi H. Enabling Next-Gen Healthcare: advanced interoperability and integration with AI, IoMT, and precision medicine. Int Adv Res J Sci Eng Technol 2024 Aug 20;8(1):89-95 [FREE Full text] [doi: 10.17148/IARJSET.2021.8116]
69. Theriault-Lauzier P, Cobin D, Tastet O, et al. A responsible framework for applying artificial intelligence on medical images and signals at the point of care: the PACS-AI platform. Can J Cardiol 2024 Oct;40(10):1828-1840. [doi: 10.1016/j.cjca.2024.05.025] [Medline: 38885787]
70. Lehoux P, Daudelin G, Williams-Jones B, Denis JL, Longo C. How do business model and health technology design influence each other? Insights from a longitudinal case study of three academic spin-offs. Res Policy 2014 Jul;43(6):1025-1038. [doi: 10.1016/j.respol.2014.02.001]
71. Javanmardi E, Maresova P, Xie N, Mierzwiak R. Exploring business models for managing uncertainty in healthcare, medical devices, and biotechnology industries. Heliyon 2024 Feb 29;10(4):e25962. [doi: 10.1016/j.heliyon.2024.e25962] [Medline: 38370194]
72. van Leeuwen KG, de Rooij M, Schalekamp S, van Ginneken B, Rutten MJCM. How does artificial intelligence in radiology improve efficiency and health outcomes? Pediatr Radiol 2022 Oct;52(11):2087-2093. [doi: 10.1007/s00247-021-05114-8] [Medline: 34117522]
73. Lobig F, Subramanian D, Blankenburg M, Sharma A, Variyar A, Butler O. To pay or not to pay for artificial intelligence applications in radiology. NPJ Digit Med 2023 Jun 23;6(1):117. [doi: 10.1038/s41746-023-00861-4] [Medline: 37353531]
74. Alami H, Lehoux P, Denis JL, et al. Organizational readiness for artificial intelligence in health care: insights for decision-making and practice. J Health Organ Manag 2020 Dec 3;35(1):106-114. [doi: 10.1108/JHOM-03-2020-0074] [Medline: 33258359]
75. Medical research involving human subjects act (WMO). CCMO. URL: https://english.ccmo.nl/investigators/legal-framework-for-medical-scientific-research/laws/medical-research-involving-human-subjects-act-wmo [accessed 2024-11-16]
76. Guide to the review procedure. CCMO. URL: https://english.ccmo.nl/investigators/guide-to-the-review-procedure [accessed 2024-11-16]
77. Singh N, Jain M, Kamal MM, Bodhi R, Gupta B. Technological paradoxes and artificial intelligence implementation in healthcare. An application of paradox theory. Technol Forecast Soc Change 2024 Jan;198:122967. [doi: 10.1016/j.techfore.2023.122967]

XSL•FO
RenderX

78. Liew C. The future of radiology augmented with artificial intelligence: a strategy for success. Eur J Radiol 2018 May;102:152-156. [doi: 10.1016/j.ejrad.2018.03.019] [Medline: 29685530]
79. Ongena YP, Yakar D, Haan M, Kwee TC. Artificial intelligence in screening mammography: a population survey of women's preferences. J Am Coll Radiol 2021 Jan;18(1 Pt A):79-86. [doi: 10.1016/j.jacr.2020.09.042] [Medline: 33058789]
80. Lennartz S, Dratsch T, Zopfs D, et al. Use and control of artificial intelligence in patients across the medical workflow: single-center questionnaire study of patient perspectives. J Med Internet Res 2021 Feb 17;23(2):e24221. [doi: 10.2196/24221] [Medline: 33595451]
81. Groenier M, Spijkerboer K, Venix L, et al. Evaluation of the impact of technical physicians on improving individual patient care with technology. BMC Med Educ 2023 Mar 23;23(1):181. [doi: 10.1186/s12909-023-04137-z] [Medline: 36959581]
82. Hadjiiski L, Cha K, Chan HP, et al. AAPM task group report 273: recommendations on best practices for AI and machine learning for computer‐aided diagnosis in medical imaging. Med Phys 2023 Feb;50(2):e1-e24. [doi: 10.1002/mp.16188] [Medline: 36565447]
83. Geis JR, Brady AP, Wu CC, et al. Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. Radiology 2019 Nov;293(2):436-440. [doi: 10.1148/radiol.2019191586] [Medline: 31573399]
84. Mahmood U, Shukla-Dave A, Chan HP, et al. Artificial intelligence in medicine: mitigating risks and maximizing benefits via quality assurance, quality control, and acceptance testing. BJR Artif Intell 2024 Jan;1(1):ubae003. [doi: 10.1093/bjrai/ubae003] [Medline: 38476957]
85. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. Radiographics 2017;37(2):505-515. [doi: 10.1148/rg.2017160130] [Medline: 28212054]
86. Bosmans JML, Weyler JJ, De Schepper AM, Parizel PM. The radiology report as seen by radiologists and referring clinicians: results of the COVER and ROVER surveys. Radiology 2011 Apr;259(1):184-195. [doi: 10.1148/radiol.10101045] [Medline: 21224423]
87. European Society of Radiology (ESR). Good practice for radiological reporting. Guidelines from the European Society of Radiology (ESR). Insights Imaging 2011 Apr;2(2):93-96. [doi: 10.1007/s13244-011-0066-7]
88. Xu X, Li J, Zhu Z, et al. A comprehensive review on synergy of multi-modal data and AI technologies in medical diagnosis. Bioengineering (Basel) 2024 Feb 25;11(3):219. [doi: 10.3390/bioengineering11030219] [Medline: 38534493]
89. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? Patterns (N Y) 2024 Mar 8;5(3):100943. [doi: 10.1016/j.patter.2024.100943] [Medline: 38487804]
90. Li CY, Chang KJ, Yang CF, et al. Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation. Nat Commun 2025 Mar 6;16(1):2258. [doi: 10.1038/s41467-025-57426-0] [Medline: 40050277]
91. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024 Apr;30(4):1134-1142. [doi: 10.1038/s41591-024-02855-5] [Medline: 38413730]
92. Leiner T, Bennink E, Mol CP, Kuijf HJ, Veldhuis WB. Bringing AI to the clinic: blueprint for a vendor-neutral AI deployment infrastructure. Insights Imaging 2021 Feb 2;12(1):11. [doi: 10.1186/s13244-020-00931-1] [Medline: 33528677]
93. WHO outlines considerations for regulation of artificial intelligence for health. World Health Organization. 2023. URL: https://www.who.int/news/item/19-10-2023-who-outlines-considerations-for-regulation-of-artificial-intelligence-for-health [accessed 2025-11-29]
94. Kotter E, D'Antonoli TA, Cuocolo R, et al. Guiding AI in radiology: ESR's recommendations for effective implementation of the European AI Act. Insights Imaging 2025;16(1). [doi: 10.1186/s13244-025-01905-x]
95. Winkler EC, Jungkunz M, Thorogood A, Lotz V, Schickhardt C. Patient data for commercial companies? An ethical framework for sharing patients' data with for-profit companies for research. J Med Ethics 2025 May;51(5):jme-2022. [doi: 10.1136/jme-2022-108781]
96. Opara-Martins J, Sahandi R, Tian F. Critical analysis of vendor lock-in and its impact on cloud computing migration: a business perspective. J Cloud Comp 2016 Dec;5(1):4. [doi: 10.1186/s13677-016-0054-z]
97. Botha NN, Segbedzi CE, Dumahasi VK, et al. Artificial intelligence in healthcare: a scoping review of perceived threats to patient rights and safety. Arch Public Health 2024 Oct 23;82(1):188. [doi: 10.1186/s13690-024-01414-1] [Medline: 39444019]
98. van Bekkum M, Zuiderveen Borgesius F, Heskes T. AI, insurance, discrimination and unfair differentiation: an overview and research agenda. Law Innov Technol 2025 Jan 2;17(1):177-204. [doi: 10.1080/17579961.2025.2469348]
99. Desislavov R, Martínez-Plumed F, Hernández-Orallo J. Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. Sustainable Computing: Informatics and Systems 2023 Apr;38:100857. [doi: 10.1016/j.suscom.2023.100857]
100. Gong E, Pauly JM, Wintermark M, Zaharchuk G. Deep learning enables reduced gadolinium dose for contrast‐enhanced brain MRI. Magn Reson Imaging 2018 Aug;48(2):330-340. [doi: 10.1002/jmri.25970]
101. Lin DJ, Johnson PM, Knoll F, Lui YW. Artificial intelligence for MR image reconstruction: an overview for clinicians. Magn Reson Imaging 2021 Apr;53(4):1015-1028. [doi: 10.1002/jmri.27078]

XSL•FO
RenderX

Abbreviations

AD: Alzheimer disease
AI: artificial intelligence
LLM: large language model
MRI: magnetic resonance imaging

Edited by A Stone; submitted 02.Sep.2025; peer-reviewed by J Zhang, P Shingru; accepted 31.Dec.2025; published 25.Feb.2026.

Please cite as:
Schilder MB, Keyser A, van Hees S, Sbrizzi A, Boon WPC
Anticipating Moral and Economic Considerations, Opportunities, and Potential Frictions for AI in Medical Imaging: Multistakeholder Cocreation Study
J Med Internet Res 2026;28:e83407
URL: https://www.jmir.org/2026/1/e83407
doi:10.2196/83407

© Martin Bastiaan Schilder, Alexandra Keyser, Susan van Hees, Alessandro Sbrizzi, Wouter Pieter Christiaan Boon. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Forecasting Waitlist Trajectories for Patients With Metabolic Dysfunction–Associated Steatohepatitis Cirrhosis: A Neural Network Competing Risk Analysis

Gopika Punchhi[1,2*], MD; Yingji Sun[2*], MSc; Eunice Tan[2,3,4], MD; Naomi Khaing Than Hlaing[2], MD; Chang Liu[5], BMATH; Sumeet Asrani[6], MD; Sirisha Rambhatla[7], PhD; Mamatha Bhat[2,8], MSc, MD, PhD

[1]Schulich School of Medicine and Dentistry, Western University, London, ON, Canada
[2]Ajmera Transplant Centre, University Health Network, 585 University Ave, Toronto, ON, Canada
[3]Division of Gastroenterology, Department of Medicine, National University Hospital, Singapore, Singapore
[4]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[5]Systems Design Engineering, University of Waterloo, Waterloo, ON, Canada
[6]Division of Hepatology, Baylor University Medical Center, Dallas, TX, United States
[7]Department of Management Sciences, University of Waterloo, Waterloo, ON, Canada
[8]Division of Gastroenterology and Hepatology, University of Toronto, Toronto, ON, Canada
[*]these authors contributed equally

Corresponding Author:
Mamatha Bhat, MSc, MD, PhD
Ajmera Transplant Centre, University Health Network, 585 University Ave, Toronto, ON, Canada

Abstract

Background: Metabolic dysfunction–associated steatohepatitis (MASH) cirrhosis is a leading indication for liver transplantation (LT). Patients with MASH cirrhosis are complex and often have extensive comorbidities. The current model for end-stage liver disease (MELD)–based liver allocation system has suboptimal concordance in predicting waitlist mortality for patients with MASH cirrhosis. Furthermore, it does not capture the competing outcomes of death and LT on the liver transplant waitlist.

Objective: A competing risk analysis using deep learning was conducted to forecast waitlist trajectories of patients with MASH cirrhosis using data available at the time of waitlisting.

Methods: A deep learning competing risk model was constructed using data from 17,551 waitlisted patients with MASH cirrhosis in the Scientific Registry of Transplant Recipients (SRTR) based on the DeepHit model framework with five-fold cross-validation. Model performance was evaluated and compared to single-risk Cox proportional hazards and random survival forests (RSF) models in predicting death or transplant using the concordance index and Brier score. Additionally, a novel performance metric, the competing event coherence (CEC) score, was developed to evaluate model performance in the setting of competing risks. Features associated with death and transplant in the DeepHit model were identified using permutation importance. Models were externally validated on data from the University Health Network.

Results: A total of 17,551 patients were included. The mean MELD at listing was 19.4 (SD 8.1). At 120 months of follow-up on the waitlist, 54.6% (9599/17551) of patients underwent LT, 25.6% (4510/17551) of patients died or were removed due to deterioration, and 19.8% (3442/17551) of patients were removed for improvement or were censored. In a competing risk scenario, DeepHit achieved the best CEC scores at 1 (0.813), 3 (0.811), 6 (0.794), and 12 months (0.772) on the waitlist. The cause-specific RSF model had the highest concordance indices for death or transplant at all time points (death: 0.874 at 1 month, 0.840 at 6 months, and 0.814 at 12 months) except for death at 3 months, where DeepHit (0.883) outperformed RSF. RSF also had lower Brier scores overall, except for transplant at 12 months, where DeepHit outperformed RSF (0.206 vs 0.228). These results were similar on external validation. On feature importance assessment, MELD at listing and its components, as well as functional status, age, and blood type, were associated with death and transplant on the waitlist.

Conclusions: A deep learning competing risk analysis can forecast the risks of both death and transplant in patients with MASH on the waitlist, helping to inform clinical decisions by identifying the most impactful covariates for each outcome.

(J Med Internet Res 2026;28:e68247) doi:10.2196/68247

KEYWORDS

liver transplantation; metabolic dysfunction-associated steatohepatitis; waitlist trajectory; metabolic dysfunction; liver disease; cirrhosis; neural network; risk analysis; hepatic Disease; predictive; prediction; deep learning; liver transplant

XSL•FO
RenderX

Introduction

Metabolic dysfunction-associated steatohepatitis (MASH) cirrhosis is a leading cause of liver transplantation (LT) globally and the fastest growing indication for LT in the United States, with the prevalence of waitlisted candidates increasing from 2.5% to 20.4% between 2004 and 2019. The prevalence of MASH cirrhosis is projected to continue to rise significantly in the coming years [1,2]. Candidates waitlisted for LT are prioritized based on their model for end-stage liver disease (MELD) score, which has been periodically reviewed and updated, most recently to the MELD 3.0 score [3]. The MELD-based scoring system predicts waitlist mortality and does not account for type of liver disease, as previous studies have demonstrated minimal effects on the predictive performance [4]. Despite changes to MELD-based scoring systems in recent years and the increasing prevalence of MASH, MELD-based scoring systems have lower concordance in candidates with MASH cirrhosis compared to those listed with other liver diseases, highlighting the need to develop waitlist prediction models that capture the complexity of MASH cirrhosis [3-6]. There are several possible explanations for the lower concordance of MELD models in this population. Patients with MASH, particularly those in the low- to mid-MELD score range, tend to have faster disease progression than is captured by their MELD score progression, higher pre-LT mortality risk, and lower likelihood of recovery than other patients on the waitlist. Additionally, patients with MASH cirrhosis tend to develop clinically significant portal hypertension at lower MELD-sodium (MELD-Na) scores, contributing to higher waitlist mortality [5-8]. Despite having more severe comorbidities, such as portal hypertension, advanced age, higher BMI, diabetes, hypertension, and hyperlipidemia, patients with MASH cirrhosis are less likely to receive a transplant on the waitlist and more likely to face higher waitlist mortality and removal due to becoming too ill to undergo transplant compared to patients listed with other liver diseases [8-12]. Improving waitlist prediction in patients with MASH through developing MASH-specific models can aid clinicians in optimizing their waitlist outcomes and pretransplant status, thus potentially improving overall waitlist outcomes.

Another limitation of MELD-based models is that they provide information on the risk of death on the waitlist without accounting for the risk of transplantation, censoring patients who do not experience the event of interest and leading to biased estimation of risk [7]. These single-risk Cox proportional hazards (CoxPH)–based models cannot predict risk while accounting for the possibility that a patient can experience multiple events at a given time point on the waitlist, while a censored patient can experience a competing event that would make the primary event of interest clinically impossible (ie, a patient who undergoes LT cannot also die on the waitlist). The risks of each event cannot be compared between multiple cause–specific models. In contrast, competing risk models account for multiple mutually exclusive events [13]. Compared to traditional regression methods, machine learning (ML) can handle large, heterogeneous datasets and avoid several fundamental assumptions of linearity and proportionality that CoxPH models make [14-16]. For example, DeepHit is a deep learning competing risk neural network model that captures intricate, nonlinear interactions between several factors and outcomes (such as mortality and transplant) in one model [9,11,17,18]. By using a DeepHit-based model to predict waitlist outcomes, clinicians will better understand the trajectory of patients with MASH cirrhosis with numerous comorbidities who are at high risk of both mortality and transplantation. Patients with MASH at lower risk of death and transplant may be better candidates for living donor liver transplantation (LDLT) and can be redirected accordingly [19]. Due to the fundamental differences between competing risk and single-risk settings, current model evaluation metrics, such as the concordance index (C-index) and Brier score, may be inadequate for use in competing risk settings [12,20]. We design and propose the competing event coherence (CEC) score, a novel performance metric to assess models in a competing events scenario at the patient level. It is an interevent metric that evaluates the match between the event predicted by the model and the actual event that occurred at a given time point for each patient and takes into account multiple competing events.

In this study, we aimed to develop a MASH cirrhosis–specific deep learning model based on DeepHit using data at the time of waitlisting that accounts for the competing risks of death and transplant on the LT waitlist to forecast waitlist trajectory and inform clinical decision-making. We compared the performance of the DeepHit model to MASH-specific single-risk CoxPH and random survival forest (RSF) models. We externally validated the DeepHit model on single-center data. Finally, we developed a DeepHit dashboard to enter and visualize patient trajectory on the waitlist [21].

Methods

Study Design and Participants

We conducted our retrospective study using two study populations. Of 227,647 patients waitlisted in the Scientific Registry of Transplant Recipients (SRTR) from March 1, 2002, to March 2, 2021, we used data from 17,551 patients with MASH cirrhosis for model development (Figure S1). For external validation, we used data from 167 patients with MASH cirrhosis who were waitlisted at University Health Network (UHN), Toronto, Ontario, between 2012 and 2018. Inclusion and exclusion criteria were identical for both cohorts, ensuring consistency in the selection process. We included all adults aged 18 years or older with a primary diagnosis of MASH or cryptogenic cirrhosis (CC) and a BMI of 30 kg/m^2 or more based on previous studies that demonstrate the histological overlap between MASH cirrhosis and CC [22-24]. We excluded patients with a secondary diagnosis other than MASH cirrhosis or CC with a BMI of 30 kg/m^2 or more [22,25]. We excluded LT recipients who were never waitlisted, retransplants, multiorgan transplants, acute liver failure (including status 1 and status 1a candidates), concomitant liver etiologies (viral hepatitis B and C and alcoholic liver disease), hepatocellular carcinoma–related primary or secondary diagnosis or listed with exception points, and those with pre-existing liver malignancies; supplementary information 1). Patients were followed for up

XSL•FO
RenderX

to 120 months, or until death or deterioration, removal, or transplant, whichever occurred first. Events in both cohorts were classified as (1) death (died on the waitlist or removed due to deterioration), (2) received LT, or (3) censored, and event times were determined accordingly. Candidates removed due to deterioration were included in the death group, while those removed from the waitlist for other reasons were classified as censored (Figure S2). Waitlist outcome categorization is consistent with previous literature [26].

Models

We developed all models using SRTR data based on shared features between the SRTR and UHN datasets. The DeepHit model includes a shared network with fully connected layers to capture intrinsic patterns of the input features, which are then connected to two cause-specific subnetworks that learn the relationship between the input features, event type, and time of the event [11,27]. DeepHit outputs discrete monthly risk predictions of the competing events. The sum of the monthly prediction over the entire time horizon T for all the competing events is $1/K$, where K is the total number of competing events. In our case, the maximum event time is 120 months. The cumulative risk for 1 event over this time period is 0.5, as there are two competing events. The DeepHit model includes a shared network with fully connected layers and 2 cause-specific subnetworks that correspond to death and transplant. The joint distribution of the event and first hitting time is learned and outputted through a final layer to output r_i (k, $t|x_i$) defined as

rik,txi=fxi

which is the predicted probability of the k event happening at a specific prediction time t, satisfying

∑k=1, t=1K, Tri(k,t|xi)=1

where f represents the DeepHit structure [11].

We compared DeepHit to single-risk CoxPH and RSF models predicting death or transplant. CoxPH is a linear model that predicts the hazard function with an assumption of proportionality where the individual hazard is proportional to the population baseline hazard that changes over time [14]. (Figure 1) For each cause-specific CoxPH or RSF model predicting death or transplant, the competing event (death or transplant) was censored. Model hyperparameters are available in Table S1 in Multimedia Appendix 1. For comparison, MELD-Na and MELD 3.0 scores were calculated to predict mortality (supplementary information 1)

XSL•FO
RenderX

Figure 1. DeepHit model. (A) The model training process, event prediction, and evaluation process are displayed. (B) The DeepHit model architecture is described. (C) The competing event coherence score evaluates model performance based on the prediction of an event compared to the actual event that occurred at the patient level. (D) DeepHit can be used to predict death and transplant and inform clinical decisions. C-index: concordance index; CEC: competing event coherence; CoxPH: Cox proportional hazards; INR: international normalized ratio; MASH: metabolic dysfunction–associated steatohepatitis; MELD: Model for End-Stage Liver Disease; RSF: random survival forest; SRTR: Scientific Registry of Transplant Recipients; and UHN: University Health Network.

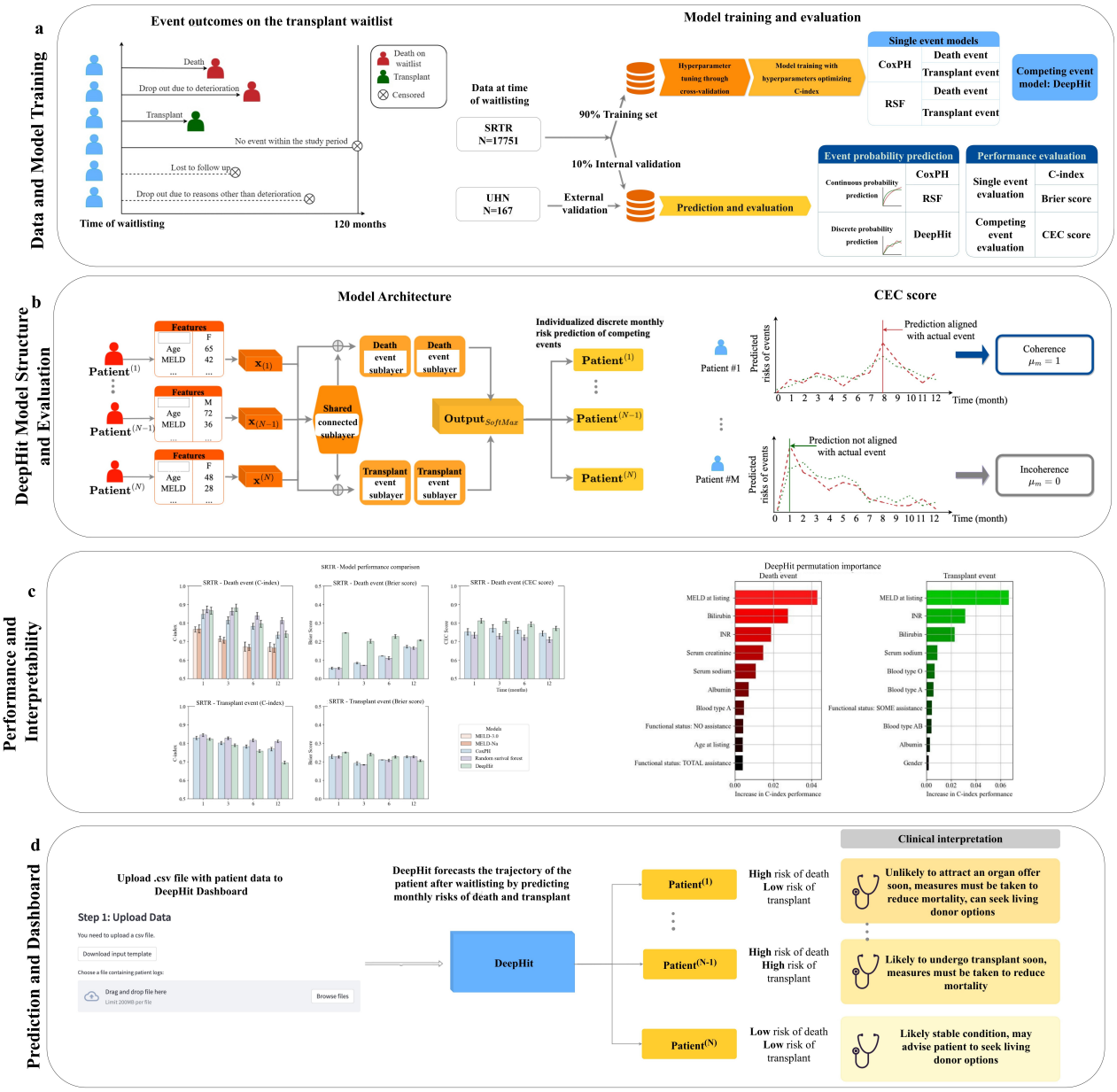

Model Evaluation

Models were evaluated using the C-index and Brier scores. Furthermore, we propose a new metric to evaluate the percentage of prediction that is coherent to the actual event type and time at which it occurred in competing risk scenarios, known as the CEC score (or the μ-score). This overcomes the limitations of the C-index and the Brier score, which examine model performance under a single-risk scenario but cannot be used to compare all the predicted risks for the competing events at the same time. Ideally, at the time of the actual event that occurred, the predicted risk for that event should be higher than the predicted risk of other competing events. For the patients who had an event (death or transplant) within the time frame of interest, the percentage of coherence within the population was calculated. The CEC score measures the alignment between the model risk predictions and the actual patient event that occurred. For the M patients who had the event within the time frame t, the percentage of coherence within the population was calculated. The CEC score measures the alignment between the risk predictions and the actual patient event. Let's μ_m indicate the coherence status of the mth patient being evaluated. Further, let l^* and k^*_m denote the actual event time and event type of the mth patient, respectively; then the proposed μ-score is defined as

$$\mu\text{-score}=1M\Sigma m=1M\mu m$$

where coherence for the m patient μ_m is defined as

XSL•FO
RenderX

$$\mu m=\{1,\text{if } \arg\max k\in K\, rm(k,\ell*|xm)=km*0,\text{otherwise},$$

and r_m (k, l*|x_m) denotes the predicted risks of the 2 events at the event time. Because CoxPH and RSF are single-risk models, we say that the prediction is in coherence (or alignment) if the probability of experiencing an event Pr_{event} is higher for the event corresponding to the actual event (at time l* as compared to the competing one.

C-index, Brier scores, and CEC scores were computed for all models at four time points after waitlisting: 1, 3, 6, and 12 months. One month corresponds to the 25th percentile of event time in the population [17]. Mortality predictions were also generated using MELD-Na and MELD-3.0 scores for performance comparison. Transplant predictions were not developed for the MELD models since they are only intended to predict death on the waitlist. Since DeepHit is a competing risk model, when the performance was evaluated using single event metrics (C-index and Brier score), the competing event was treated as censored [11].

Statistical Analysis

To ensure the robustness of our model, we used rigorous cross-validation, including k-fold cross-validation and hyperparameter tuning to optimize model performance and ensure generalizability (supplementary information 2). SRTR data were split using stratified random split to preserve the event rate of the population in the training and test sets, which represented 90% and 10% of the entire population, respectively. Within the 90% training set, outer five-fold cross-validation was used to obtain an average performance across all validation folds.

Hyperparameter tuning was used to optimize the C-index. To evaluate the performance of our trained models, we first used single-risk metrics, including the time-dependent C-index and the time-dependent Brier score. The model with the highest performance on the outer validation set was used to test performance on the 10% test set as well as the UHN external validation set. Test performance was evaluated at 1, 3, 6, and 12 months using bootstrapping to obtain consistent results where each bootstrapped sample contains patients that were randomly sampled from the cohort with replacement. The Wilson Cox test was subsequently used to test statistical significance in the performance of DeepHit and the other models (CoxPH and RSF) based on the bootstrapped C-indices, Brier scores, and CEC score proposed as the following (Table S1).

Model Interpretability

Permutation importance was used to determine which covariates in the DeepHit model have the largest influence on the prediction of death and transplant by evaluating the contribution of each covariate to the C-index of the model via random permutation of each variable, which is then compared to the original data [28]. The permutation was done 20 times for each variable to obtain the average and SD of increase in C-index.

All analyses were done using Python version 3.8.8 (Python Software Foundation). CoxPH models were developed using the scikit-learn 1.1.1 library and the scikit-survival 0.16.0 library. The MASH-specific DeepHit model was developed using the TensorFlow 0.0.8 library. The data template can be downloaded as a .csv file. The codebase has been published on GitHub [29].

Ethical Considerations

Due to the use of publicly available deidentified United Network for Organ Sharing (UNOS) data, this study was exempt from Research Ethics Board (REB) review. For external validation, REB approval was obtained (REB number 21 - 5783).

Results

Characteristics of SRTR and UHN Cohorts

There were 17,551 patients with MASH in the SRTR cohort, of which 50.2% (8802/17551) were female. Mean MELD at listing was 19.4 (SD 8.1). By 120-month waitlist follow-up, 54.6% (9599/17551) of patients underwent LT, 25.6% (4510/17551) of patients died or were removed from the waitlist due to deterioration, and 19.8% (3442/17551) of patients were removed for improvement or were censored. Around 93.8% (9004/9599) of recipients underwent deceased donor liver transplantation (DDLT). There were 167 patients with MASH cirrhosis in the UHN dataset, and 46.1% (77/167) were female. The mean MELD at listing was 22.1 (SD 6.6). Overall, 62.9% (105/167) of patients underwent LT, of which 72.4% (76/105) of patients underwent DDLT, 23.4% (39/167) died, and 13.7% (23/167) were removed or censored (see supplementary information 1 and Table S2 in Multimedia Appendix 1). Additional features in the SRTR and UHN datasets are available in Table S3 in Multimedia Appendix 1.

Model Performance

In the competing risk scenario, DeepHit consistently achieved statistically significant higher CEC scores at each time point evaluated on the SRTR and UHN datasets (Figures 2 and 3). Numerical results are displayed in supplementary information 6. When comparing model performance using SRTR data (Figure 2), CoxPH, RFS, and DeepHit had higher C-indices than MELD-Na and MELD 3.0 for death events at 1, 3, 6, and 12 months, outperforming MELD-Na and MELD 3.0 at all time points. RSF achieved statistically significantly higher C-indices (P<.01) except in the RSF-DeepHit comparison at 3 months for death event and 1 month for transplant. RSF had statistically significantly lower Brier scores at each time point evaluated for death and transplant, except for at 12 months for the transplant event (DeepHit: 0.206 vs RSF: 0.228).

XSL•FO
RenderX

Figure 2. Model performance comparison with the Scientific Registry of Transplant Recipients data. Performance of models evaluated using C-index, Brier score, and CEC score on SRTR data. CEC: competing event coherence; C-index: concordance index; CoxPH: Cox proportional hazards; MELD: Model for end-stage liver disease; SRTR: Scientific Registry of Transplant Recipients.

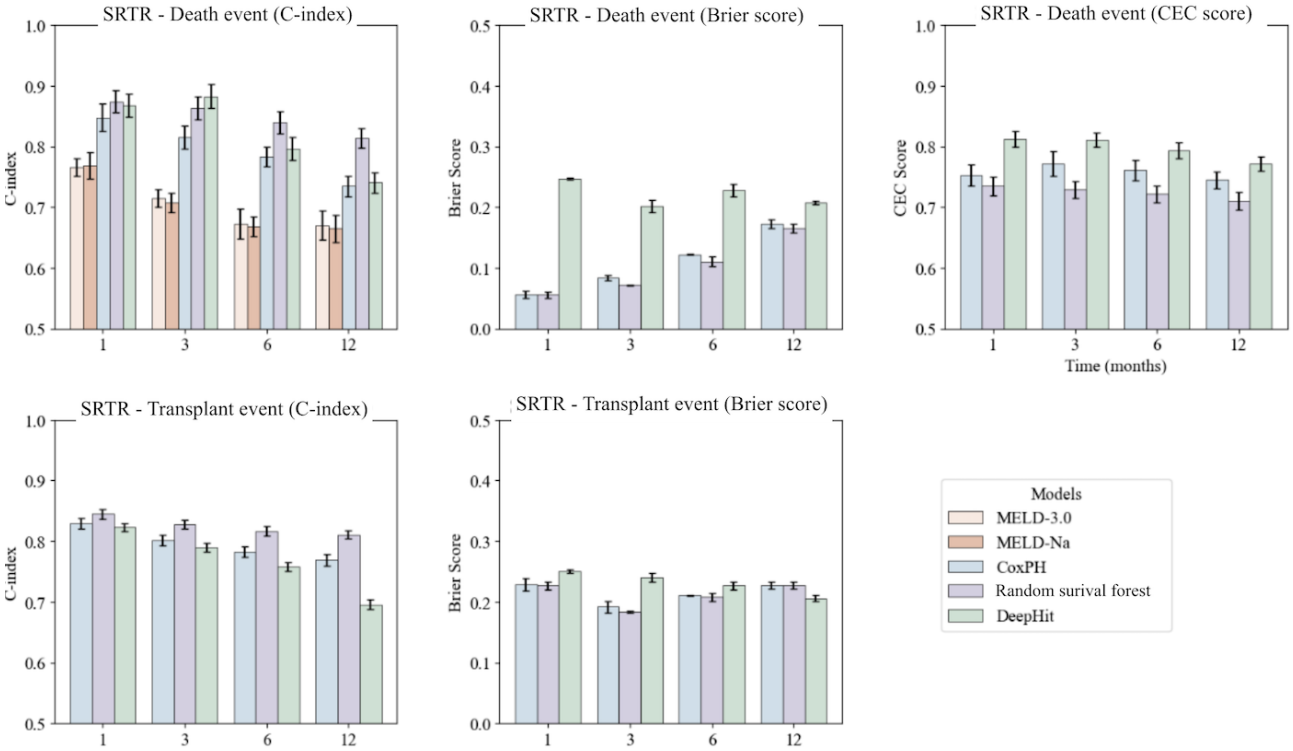

Figure 3. Model performance comparison and external validation of the model using University Health Network data. Performance of models evaluated using C-index, Brier score, and CEC score on UHN data. CEC: competing event coherence; C-index: concordance index; CoxPH: Cox proportional hazards; MELD: Model for end-stage liver disease; SRTR: Scientific Registry of Transplant Recipients.

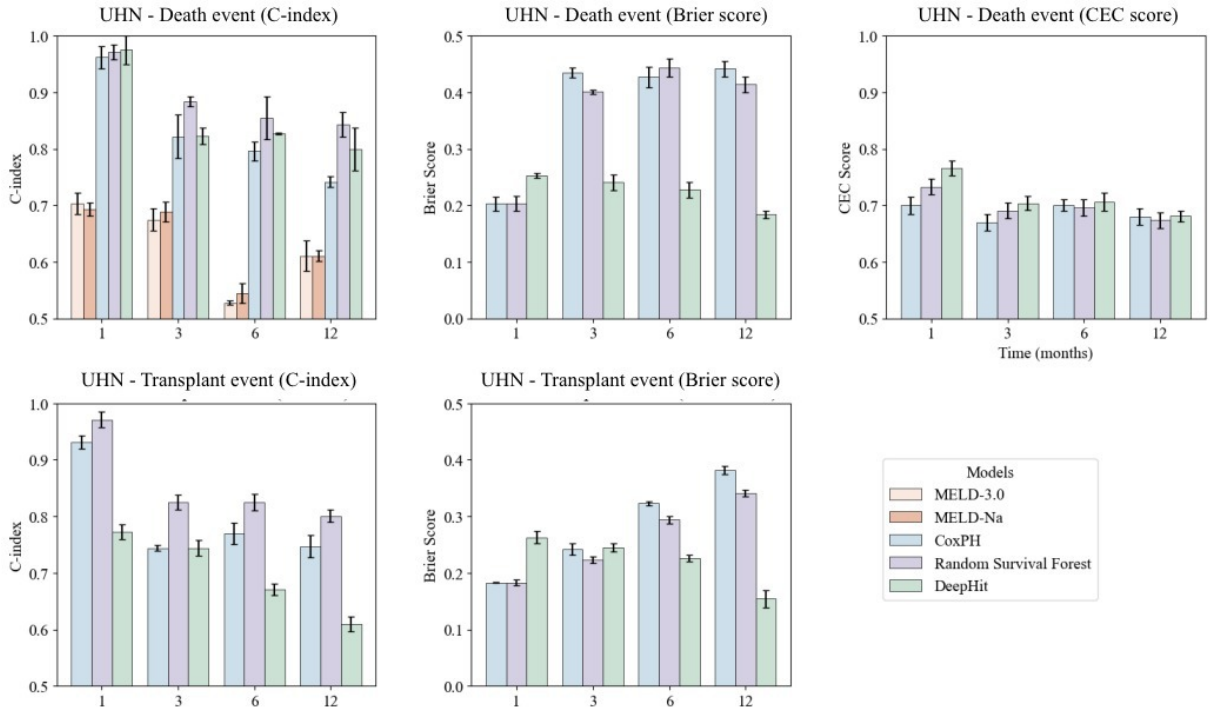

In external validation, DeepHit demonstrated a statistically significantly higher C-index at 1 month for death (0.975); however, RSF had the best performance at all other time points for the death and transplant events. DeepHit also exhibited lower Brier scores at 3, 6, and 12 months compared to RSF for death. For the transplant event, RSF had lower Brier scores at 1 and 3 months than DeepHit, while CoxPH and RSF performed better at 6 and 12 months (Figure 3). Full numerical results can be found in Table S4 in Multimedia Appendix 1.

XSL•FO
RenderX

Forecasting Waitlist Trajectory With DeepHit

Four patients from the SRTR were randomly selected, and their waitlist trajectories of death and transplant were predicted using RSF and DeepHit over a 120-month period based on data available at the time of waitlisting. Figure 4 displays two patients who died on the waitlist at month 33 and month 1, while Figure 5 displays two patients who were transplanted at month 1 and 12. The time at which the event (death or transplant) occurred and the corresponding prediction at that time is magnified. Patient characteristics for the sample patients used here are detailed in Table S5 in Multimedia Appendix 1. Cumulative risk predictions with RSF were generated using two separate models with death or transplant as the outcome of interest and displayed on one graph. Using DeepHit, we generated granular risk predictions from one competing risk model, allowing for visualization of the risk associated with each event over time compared.

Figure 4. Forecasting death on the waitlist using Random Survival Forest and DeepHit with patient examples using data at the time of waitlisting. The vertical line on each plot indicates the event (green=transplant; red=death) and the time at which it occurred. On the DeepHit plots, zoomed-in segments correspond to the time at which the actual event occurred. For a patient that experienced an event during month 1, they were categorized as experiencing the event at time 0. For a patient that experienced an event during months 1 and 2, they were categorized as experiencing the event at month 1 and so on.

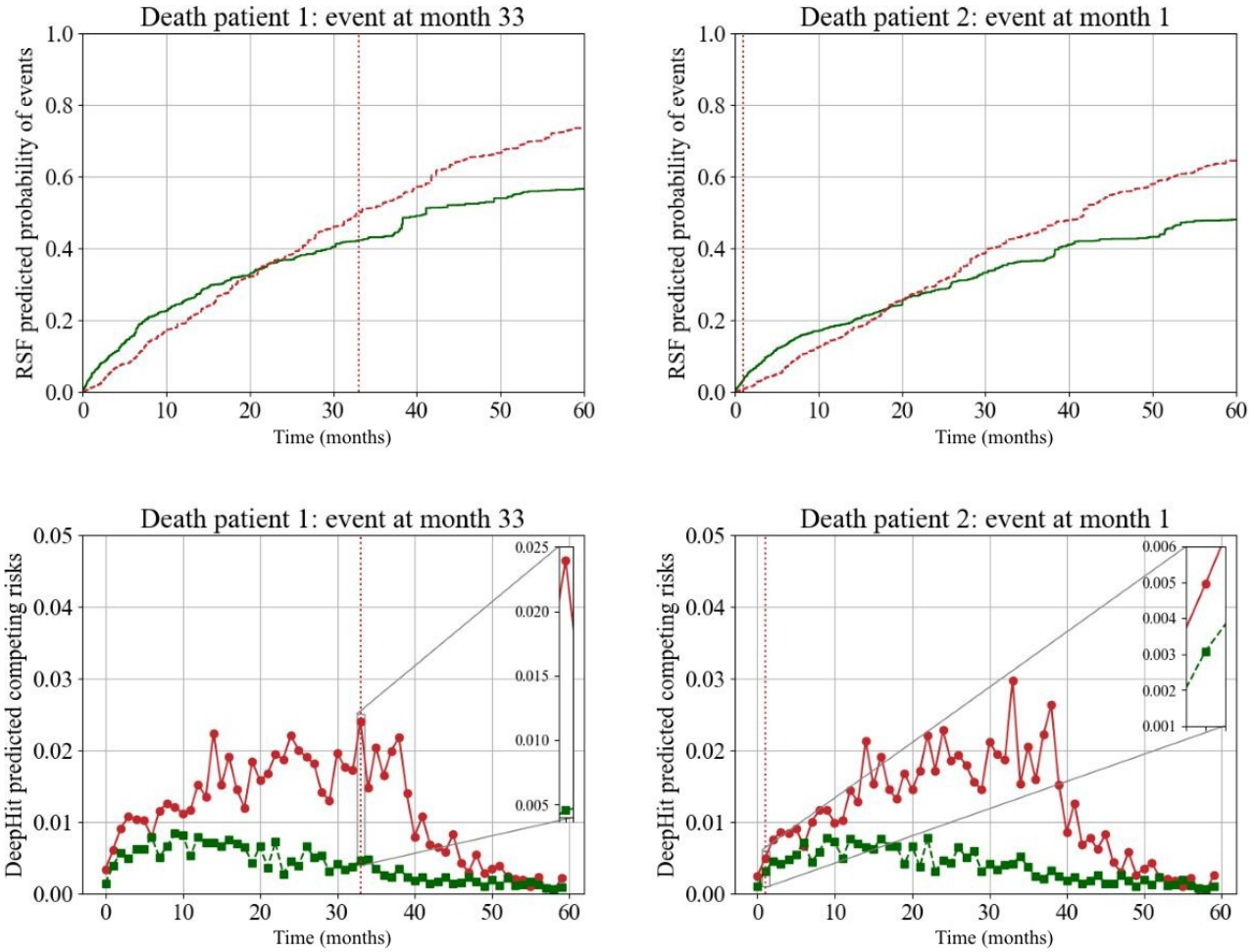

XSL•FO
RenderX

Figure 5. Forecasting transplant on the waitlist using random survival forest and DeepHit. Similar to Figure 4, transplant on the waitlist is predicted using RSF and DeepHit with each plot corresponding to a different patient.

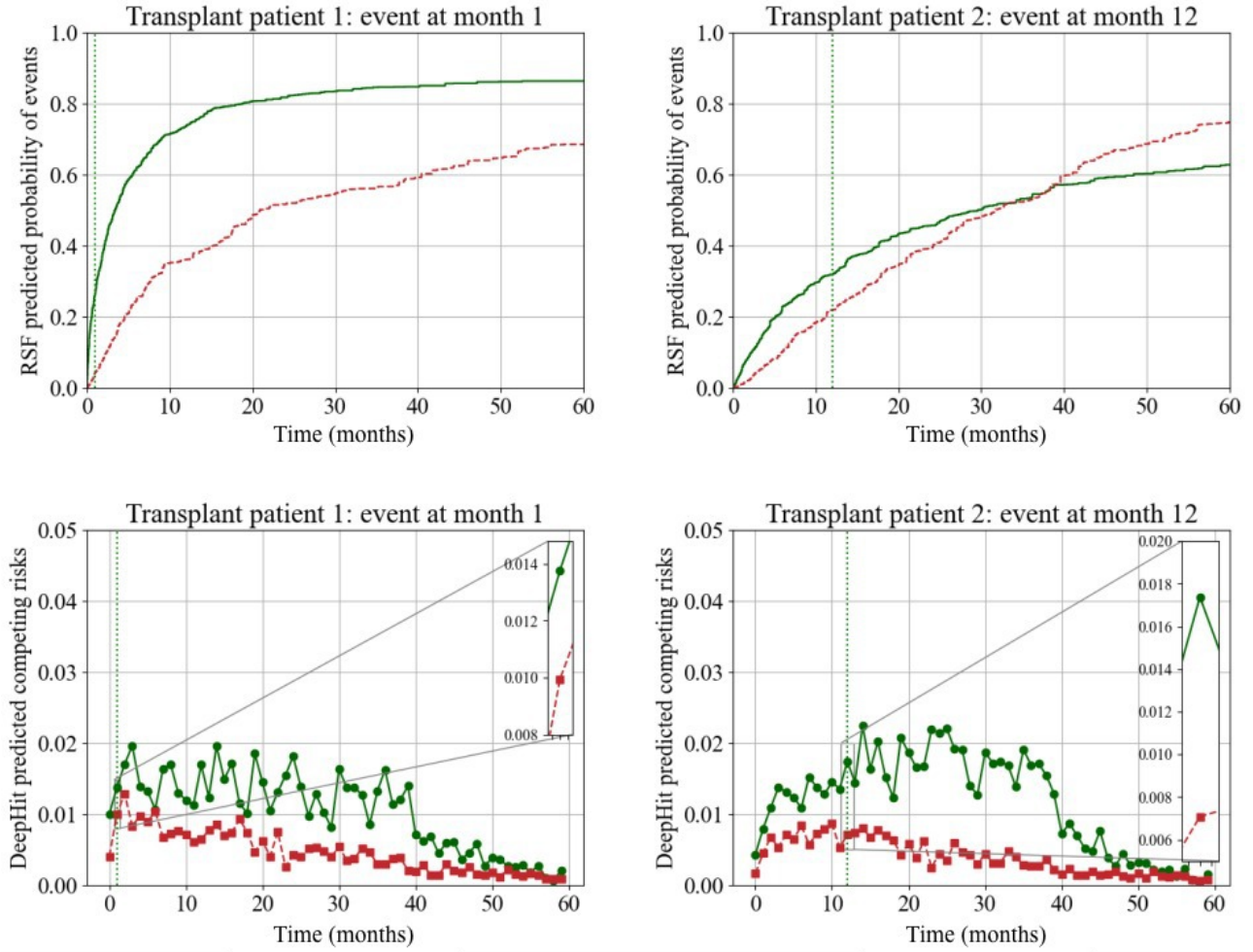

Permutation Importance Using DeepHit

In predicting the death event, MELD at listing emerged as the greatest contributor to the C-index, followed by bilirubin, INR, serum creatinine, and serum sodium. Additional non-MELD features identified were albumin, blood type A, functional status, and age at listing. For transplant events, MELD at listing took precedence, followed by INR and bilirubin. The non-MELD features that influenced transplantation comprised blood types O, A, and AB, as well as functional status, albumin, and gender (Figure 6).

XSL•FO
RenderX

Figure 6. Permutation importance of DeepHit model. Features with the greatest contribution to the prediction of death or transplant were determined using permutation importance. INR: international normalized ratio; MELD: Model for end-stage liver disease

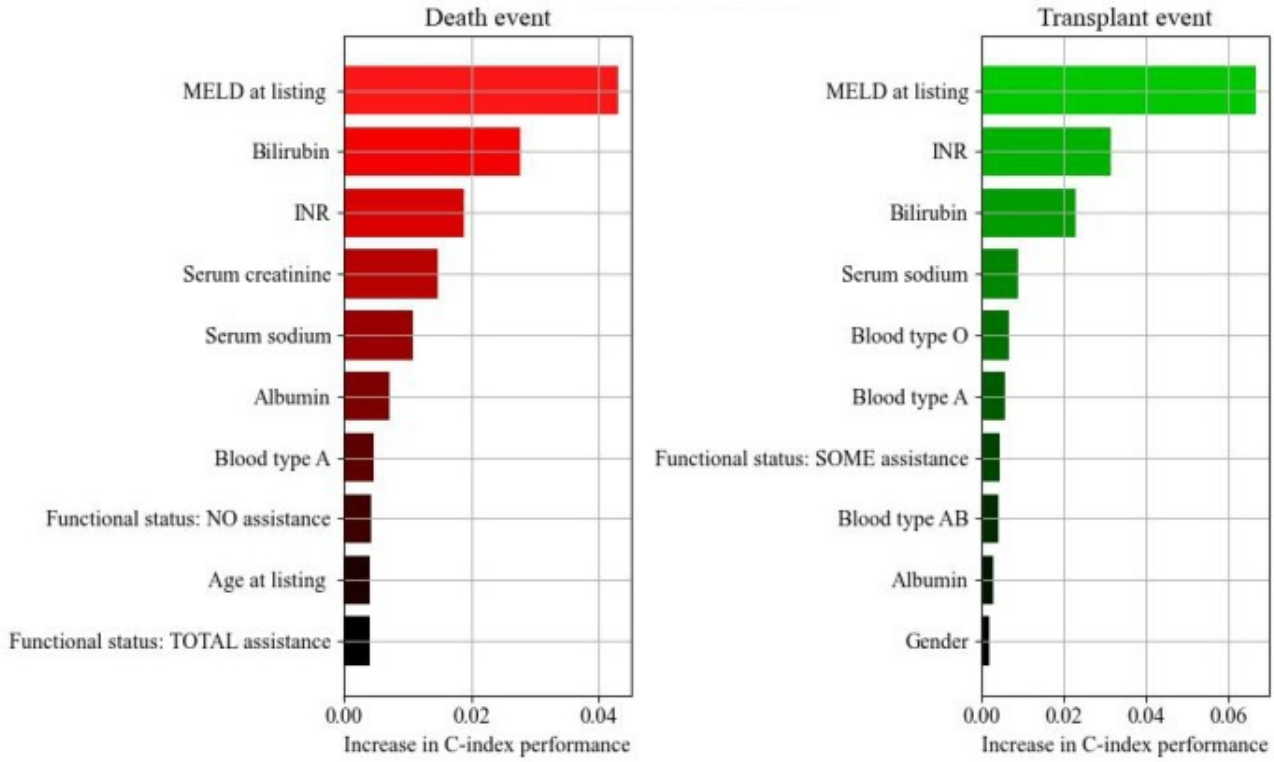

Discussion

Principal Findings

This study aimed to forecast the LT waitlist trajectories of patients with MASH cirrhosis using a deep learning approach that leverages the changing relationships between covariates and risks over time while handling competing risks [11]. We compared a competing risk MASH cirrhosis–specific DeepHit model to single-risk CoxPH and RSF models in forecasting outcomes of death and transplant on the waitlist. We demonstrated that while RSF performed better at most evaluation time points using traditional model performance metrics, DeepHit outperforms RSF when evaluated in a competing risk scenario. Furthermore, we demonstrated that DeepHit can generate and visualize time-varying, discrete predictions of death and transplant. This allows for the identification of patients with a low probability of undergoing transplantation and those at an elevated risk of death while awaiting DDLT, enabling timely consideration of LDLT when available.

MELD-based models poorly predict death in patients with MASH compared to candidates listed for other etiologies and may underestimate the risk of death on the waitlist [5,6]. Additionally, MELD-based models do not provide information on the risk of transplantation. Patients with MASH have a lower 90-day and 1-year probability of transplant and a greater risk of waitlist mortality. Therefore, early identification of patients at low risk of transplant may expedite direction to LDLT and prevent MASH-related mortality on the waitlist [30]. Previous studies have indicated that LDLT results in comparable or improved postoperative outcomes for patients with and without MASH when compared to DDLT. Timely identification of patients enables them to undergo transplantation while in a less decompensated state, leading to more efficient resource usage [31-33]. While MASH-specific models, such as the DeepHit-based model, may not fully replace MELD-based models for waitlist prioritization and liver allocation of all candidates, they can be used to improve the management of waitlisted patients with MASH, particularly those with indolent disease who may spend months on the waitlist and deteriorate before receiving an LT offer. As the prevalence of MASH cirrhosis grows, careful management of waitlisted patients is critical for reducing overall waitlist morbidity and mortality [30]. The DeepHit MASH-specific model also highlights the potential to incorporate personalized prediction of waitlist outcomes at the patient level. This prediction can be updated as clinical status changes over time when modifiable factors associated with a higher risk of waitlist mortality are addressed, providing a more granular prediction of both death and transplant over time on the waitlist as opposed to risk of 90-day mortality alone.

We presented the CEC score as an alternative metric to evaluate competing risk models, emphasizing its role as an interevent measure that assesses a model's performance in predicting 2 competing events [34]. Unlike the C-index, which evaluates performance at the population level, the CEC score assesses performance at the patient level [12,20]. A major problem with the C-index in the setting of competing risks is that the discrimination of a model is reduced when covariates may be associated with both the primary and competing event [20]. Our analysis demonstrates that DeepHit more accurately predicts the actual event (death or transplant) that occurred at specific time points when evaluated using the CEC score compared to

XSL•FO
RenderX

RSF, although RSF performs better based on the C-index and Brier scores. Although the C-index and Brier scores are limited in the scenario of competing risks, DeepHit still exhibited robust performance when evaluated using these metrics, although it was statistically weaker than RSF.

While RSF performed well in a single-risk scenario when evaluated with the C-index and Brier scores, using a noncompeting risk model to predict risk in a clinical setting where patients are at risk of multiple competing events may lead to greater misclassification of risk for single events. We demonstrated this through the improved performance of DeepHit compared to RSF when evaluated with the CEC score. Competing risk analysis may improve predictions of outcomes at the patient level [35-40]. Furthermore, censoring of competing risks can lead to event overestimation in populations at highest risk of experiencing either of the competing events [41]. Competing risk models to evaluate the risks of death and transplant on the waitlist are limited in the field of LT but have been more extensively described and applied to kidney transplantation. Smits et al [42] found that Kaplan-Meier overestimates the chance of transplantation compared to competing risk analysis by over 30% [43]. This not only highlights the importance of competing risk models but also using metrics, such as the CEC score, in the analysis of models to accurately assess performance. We highlight the advantages of DeepHit, whose neural network architecture is particularly well-suited to capturing complex, nonlinear, and interdependent relationships among variables in medium-sized datasets. In contrast to traditional survival models, DeepHit does not rely on restrictive assumptions, such as proportional hazards, enabling greater flexibility in modeling intricate data patterns. Crucially, DeepHit is inherently designed to handle competing risks, estimating the joint probability distribution over multiple, mutually exclusive event types. This multitask learning framework allows the model to share information across outcomes while retaining event-specific distinctions, resulting in more accurate and clinically meaningful survival estimates. By applying a competing risk-specific metric—the CEC score—we demonstrated that DeepHit consistently outperforms traditional ML approaches in predicting time-to-event outcomes with competing risks. These capabilities make DeepHit especially valuable in clinical settings characterized by multiple potential outcomes, such as the organ transplant waitlist.

Previous studies have assessed features associated with death and transplant in LT candidates waitlisted for all indications [5,44-46]. Despite patients with MASH cirrhosis constituting a large and increasing portion of the LT waitlist, limited studies have assessed features associated with death and transplant exclusively in this population [47]. Studies have shown that factors, such as dialysis, sex, race, serum albumin and creatinine, low performance status, and high MELD are associated with increased or decreased risk of transplant among all candidates [5,44-46]. We evaluated the contribution of various covariates in predicting death and transplant events using DeepHit permutation importance, aiming to improve the C-index. In our DeepHit model, MELD at listing emerged as the highest-ranked feature for both death and transplant, followed by components of the MELD score, such as INR, bilirubin, and serum sodium, which have been previously demonstrated [10]. Our permutation analysis reinforced the critical role of MELD-based models in predicting outcomes, showing that biochemical features already included in MELD-based models are implicated in waitlist risk of death and transplant in patients with MASH cirrhosis, potentially reflecting part of the comorbidity burden in this population. Given that SRTR is a historical dataset where allocation is largely determined by MELD score, these findings are not unexpected. Functional status was a highly ranked feature for both events and has been associated with an increased likelihood of transplantation as well as increased waitlist mortality [48]. Poor functional status contributed to the death prediction, possibly because complications of cirrhosis (such as encephalopathy) are associated with poor functional status and increase the risk of death in all LT waitlist candidates [44]. While INR and bilirubin also ranked high, we did not identify creatinine or dialysis in the last week as significant features associated with the prediction of waitlist death or transplant. Age contributed to the death prediction, which has not been found previously in studies on all LT waitlist candidates. Patients with MASH are older, and older patients tend to have poorer waitlist outcomes; therefore, this finding is expected in a MASH cirrhosis cohort [45]. For the transplant event, we found that blood type AB ranked high. These candidates can accept donor livers from nearly all blood types. In other studies, dialysis, sex, race, serum albumin and creatinine, low performance status, and high MELD have been found to be associated with increased or decreased risk of transplant in all candidates [5].

Limitations

The SRTR is a retrospective dataset and limitations, such as high missingness and heterogeneity of data collection must be noted. Furthermore, there are limited longitudinal features in the SRTR dataset, which makes the development of a dynamic model challenging [49]. The permutation importance method is limited as it does not provide information on the directionality of risk such as provided by a hazard ratio, which makes it difficult to apply DeepHit to clinical scenarios where risk factors can be identified and modified to improve outcomes. Furthermore, none of the models developed consider individual organ availability, donor compatibility, and regional variation. Although the DeepHit model performed well when externally validated on the UHN dataset, the UHN sample size was relatively small at 167 patients. Further work should seek to validate this model on larger datasets. Finally, the use of historical waitlist candidate data to devise a ranking system may perpetuate existing inequities and biases in liver allocation; therefore, DeepHit should not be used for definitive decision-making on waitlisting or delisting but can be used in conjunction with other tools.

Conclusion

The DeepHit model can be used to forecast waitlist trajectories of both death and transplant in a competing risk scenario for patients with MASH, using data available at the time of listing. With DeepHit, discretized and dynamic changes of the risk of death and transplant can be visualized and compared across multiple time points and between events. This information can

XSL•FO
RenderX

enhance the current strategies for managing candidates with LT with MASH cirrhosis and act as a tool to advocate for LDLT in patients who are at high risk of waitlist dropout but underserved by MELD-based mortality predictions. A DeepHit MASH-specific model can be used as a clinical adjunct to inform and modify clinical interventions to optimize patient survival on the waitlist. Future studies should focus on improving the evaluation and interpretability of DeepHit models to expand their use in clinical settings, as well as developing larger scale ML models that consider all patients on the LT waitlist.

Funding

This study was funded by the Canadian Institutes of Health Research (MB).

Data Availability

The data reported here have been supplied by the Hennepin Healthcare Research Institute (HHRI) as the contractor for the Scientific Registry of Transplant Recipients (SRTR). The interpretation and reporting of these data are the responsibility of the author(s) and in no way should be seen as an official policy of or interpretation by the SRTR or the U.S. Government. The data will not be made available to the public. The codebase has been published on GitHub [29].

Authors' Contributions

GP, YS, SR, and MB contributed to conceptualization. YS performed data curation and investigation. YS and SR were involved in formal analysis. MB was responsible for funding acquisition. YS, GP, SR, and MB performed the methodology. YS and GP managed project administration. MB and SR provided resources. YS, SR, and CL handled the software. MB and SR were responsible for supervision. YS and CL conducted visualization. GP and YS were responsible for writing the original draft. GP, YS, ET, NH, SA, SR, and MB participated in writing review and editing.

Conflicts of Interest

MB has received investigator-initiated grants from Novo Nordisk, Merck, Paladin, Roche, Eisai, Natera, and Astra Zeneca.

Multimedia Appendix 1
Supplementary figures and tables.
[DOCX File, 148 KB - jmir_v28i1e68247_app1.docx]

References

1. Terrault NA, Francoz C, Berenguer M, Charlton M, Heimbach J. Liver transplantation 2023: status report, current and future challenges. Clin Gastroenterol Hepatol 2023 Jul;21(8):2150-2166. [doi: 10.1016/j.cgh.2023.04.005] [Medline: 37084928]
2. Younossi Z, Stepanova M, Ong JP, et al. Nonalcoholic steatohepatitis is the fastest growing cause of hepatocellular carcinoma in liver transplant candidates. Clin Gastroenterol Hepatol 2019 Mar;17(4):748-755. [doi: 10.1016/j.cgh.2018.05.057] [Medline: 29908364]
3. Kim WR, Mannalithara A, Heimbach JK, et al. MELD 3.0: the model for end-stage liver disease updated for the modern era. Gastroenterology 2021 Dec;161(6):1887-1895. [doi: 10.1053/j.gastro.2021.08.050] [Medline: 34481845]
4. Polyak A, Kuo A, Sundaram V. Evolution of liver transplant organ allocation policy: current limitations and future directions. World J Hepatol 2021 Aug 27;13(8):830-839. [doi: 10.4254/wjh.v13.i8.830] [Medline: 34552690]
5. Godfrey EL, Malik TH, Lai JC, et al. The decreasing predictive power of MELD in an era of changing etiology of liver disease. Am J Transplant 2019 Dec;19(12):3299-3307. [doi: 10.1111/ajt.15559]
6. Yardeni D, Shiloh A, Lipnizkiy I, et al. MELD-Na score may underestimate disease severity and risk of death in patients with metabolic dysfunction-associated steatotic liver disease (MASLD). Sci Rep 2023 Dec 13;13(1):22113. [doi: 10.1038/s41598-023-48819-6]
7. Wiesner RH, McDiarmid SV, Kamath PS, et al. MELD and PELD: application of survival models to liver allocation. Liver Transpl 2001 Jul;7(7):567-580. [doi: 10.1053/jlts.2001.25879] [Medline: 11460223]
8. Spooner A, Chen E, Sowmya A, et al. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. Sci Rep 2020 Nov 23;10(1):20410. [doi: 10.1038/s41598-020-77220-w] [Medline: 33230128]
9. Jain V, Bansal A, Radakovich N, et al. Machine learning models to predict major adverse cardiovascular events after orthotopic liver transplantation: a cohort study. J Cardiothorac Vasc Anesth 2021 Jul;35(7):2063-2069. [doi: 10.1053/j.jvca.2021.02.006] [Medline: 33750661]
10. Kim RW, Therneau TM, Benson JT, et al. Deaths on the liver transplant waiting list: an analysis of competing risks. Hepatology 2006;43(2):345-351. [doi: 10.1002/hep.21025]

XSL•FO
RenderX

11. Lee C, Zame W, Yoon J, Van der Schaar M. DeepHit: a deep learning approach to survival analysis with competing risks. Proc AAAI 2018;32(1):04. [doi: 10.1609/aaai.v32i1.11842]
12. Ramspek CL, Teece L, Snell KIE, et al. Lessons learnt when accounting for competing events in the external validation of time-to-event prognostic models. Int J Epidemiol 2022 May 9;51(2):615-625. [doi: 10.1093/ije/dyab256]
13. Monterrubio-Gómez K, Constantine-Cooke N, Vallejos CA. A review on competing risks methods for survival analysis. arXiv. Preprint posted online on Dec 10, 2022. [doi: 10.48550/arXiv.2212.05157]
14. Noordzij M, Leffondre K, van Stralen KJ, Zoccali C, Dekker FW, Jager KJ. When do we need competing risks methods for survival analysis in nephrology? Nephrol Dial Transplant 2013 Nov 1;28(11):2670-2677. [doi: 10.1093/ndt/gft355]
15. Kantidakis G, Putter H, Litière S, Fiocco M. Statistical models versus machine learning for competing risks: development and validation of prognostic models. BMC Med Res Methodol 2023 Feb 24;23(1):51. [doi: 10.1186/s12874-023-01866-z] [Medline: 36829145]
16. Austin PC, Lee DS, Fine JP. Introduction to the analysis of survival data in the presence of competing risks. Circulation 2016 Feb 9;133(6):601-609. [doi: 10.1161/CIRCULATIONAHA.115.017719] [Medline: 26858290]
17. van Geloven N, Giardiello D, Bonneville EF, et al. Validation of prediction models in the presence of competing risks: a guide through modern methods. BMJ 2022 May 24;377:e069249. [doi: 10.1136/bmj-2021-069249]
18. Zhang Z. Survival analysis in the presence of competing risks. Ann Transl Med 2017 Feb;5(3):47-47. [doi: 10.21037/atm.2016.08.62]
19. Arisar FAQ, Chen S, Chen C, et al. Availability of living donor optimizes timing of liver transplant in high-risk waitlisted cirrhosis patients. Aging (Milano) 2023 Sep 15;15(17):8594-8612. [doi: 10.18632/aging.204982]
20. Wolbers M, Koller MT, Witteman JCM, Steyerberg EW. Prognostic models with competing risks: methods and application to coronary risk prediction. Epidemiology 2009 Jul;20(4):555-561. [doi: 10.1097/EDE.0b013e3181a39056] [Medline: 19367167]
21. DeepNASH project. Streamlit. URL: https://deepmash.streamlit.app [accessed 2025-12-02]
22. Charlton MR, Burns JM, Pedersen RA, Watt KD, Heimbach JK, Dierkhising RA. Frequency and outcomes of liver transplantation for nonalcoholic steatohepatitis in the United States. Gastroenterology 2011 Oct;141(4):1249-1253. [doi: 10.1053/j.gastro.2011.06.061] [Medline: 21726509]
23. Caldwell SH, Lee VD, Kleiner DE, et al. NASH and cryptogenic cirrhosis: a histological analysis. Ann Hepatol 2009;8(4):346-352. [doi: 10.1016/S1665-2681(19)31748-X] [Medline: 20009134]
24. Sutedja DS, Gow PJ, Hubscher SG, Elias E. Revealing the cause of cryptogenic cirrhosis by posttransplant liver biopsy. Transplant Proc 2004 Oct;36(8):2334-2337. [doi: 10.1016/j.transproceed.2004.07.003] [Medline: 15561241]
25. Wong RJ, Aguilar M, Cheung R, et al. Nonalcoholic steatohepatitis is the second leading etiology of liver disease among adults awaiting liver transplantation in the United States. Gastroenterology 2015 Mar;148(3):547-555. [doi: 10.1053/j.gastro.2014.11.039] [Medline: 25461851]
26. Zhao X, Naghibzadeh M, Sun Y, et al. Machine learning prediction model of waitlist outcomes in patients with primary sclerosing cholangitis. Transplant Direct 2025 Apr;11(4):e1774. [doi: 10.1097/TXD.0000000000001774] [Medline: 40166627]
27. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat ;2(3):841-860. [doi: 10.1214/08-AOAS169]
28. Breiman L. Random Forests. Mach Learn 2001 Oct;45(1):5-32. [doi: 10.1023/A:1010933404324]
29. Sun Y, Lui C, Rambhatla S. DeepNASH codebase. GitHub.
30. Lim WH, Ng CH, Tan D, et al. Natural history of NASH cirrhosis in liver transplant waitlist registrants. J Hepatol 2023 Oct;79(4):1015-1024. [doi: 10.1016/j.jhep.2023.05.034]
31. Barbas AS, Goldaracena N, Dib MJ, et al. Early intervention with live donor liver transplantation reduces resource utilization in NASH: the Toronto experience. Transplant Direct 2017 Jun;3(6):e158. [doi: 10.1097/TXD.0000000000000674] [Medline: 28620642]
32. Karnam RS, Azhie A, Yang C, et al. Impact of living donor liver transplantation on long‐term cardiometabolic and graft outcomes in cirrhosis due to nonalcoholic steatohepatitis. Clin Transplant 2023 Sep;37(9):e15008. [doi: 10.1111/ctr.15008]
33. Hum A, et al. Adult living donor versus deceased donor liver transplant (LDLT versus DDLT) at a single center: time to change our paradigm for liver transplant. Ann Surg 2019 Sep;270(3):444-451. [doi: 10.1097/SLA.0000000000003463]
34. Wolbers M, Blanche P, Koller MT, Witteman JCM, Gerds TA. Concordance for prognostic models with competing risks. Biostatistics 2014 Jul;15(3):526-539. [doi: 10.1093/biostatistics/kxt059] [Medline: 24493091]
35. Cuthbert AR, Graves SE, Giles LC, Glonek G, Pratt N. What is the effect of using a competing-risks estimator when predicting survivorship after joint arthroplasty: a comparison of approaches to survivorship estimation in a large registry. Clin Orthop Relat Res 2021 Feb 1;479(2):392-403. [doi: 10.1097/CORR.0000000000001533] [Medline: 33105301]
36. Berry SD, Ngo L, Samelson EJ, Kiel DP. Competing risk of death: an important consideration in studies of older adults. J Am Geriatr Soc 2010 Apr;58(4):783-787. [doi: 10.1111/j.1532-5415.2010.02767.x] [Medline: 20345862]
37. Southern DA, Faris PD, Brant R, et al. Kaplan–Meier methods yielded misleading results in competing risk scenarios. J Clin Epidemiol 2006 Oct;59(10):1110-1114. [doi: 10.1016/j.jclinepi.2006.07.002]

XSL•FO
RenderX

38. Kim HT. Cumulative incidence in competing risks data and competing risks regression analysis. Clin Cancer Res 2007 Jan 15;13(2 Pt 1):559-565. [doi: 10.1158/1078-0432.CCR-06-1210] [Medline: 17255278]
39. Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data. Am J Epidemiol 2009 Jul 15;170(2):244-256. [doi: 10.1093/aje/kwp107] [Medline: 19494242]
40. Sapir-Pichhadze R, Pintilie M, Tinckam KJ, et al. Survival analysis in the presence of competing risks: the example of waitlisted kidney transplant candidates. Am J Transplant 2016 Jul;16(7):1958-1966. [doi: 10.1111/ajt.13717] [Medline: 26751409]
41. Coemans M, Tran TH, Döhler B, et al. A competing risks model to estimate the risk of graft failure and patient death after kidney transplantation using continuous donor-recipient age combinations. Am J Transplant 2025 Feb;25(2):355-367. [doi: 10.1016/j.ajt.2024.07.029]
42. Smits JM, van Houwelingen HC, De Meester J, Persijn GG, Claas FH. Analysis of the renal transplant waiting list: application of a parametric competing risk method. Transplantation 1998 Nov 15;66(9):1146-1153. [doi: 10.1097/00007890-199811150-00006] [Medline: 9825809]
43. El Ters M, Smith BH, Cosio FG, Kremers WK. Competing risk analysis in renal allograft survival: a new perspective to an old problem. Transplantation 2021 Mar 1;105(3):668-676. [doi: 10.1097/TP.0000000000003285] [Medline: 32332421]
44. McCabe P, Hirode G, Wong R. Functional status at liver transplant waitlisting correlates with greater odds of encephalopathy, ascites, and spontaneous bacterial peritonitis. J Clin Exp Hepatol 2020;10(5):413-420. [doi: 10.1016/j.jceh.2020.04.015] [Medline: 33029049]
45. Nagai S, Safwan M, Kitajima T, Yeddula S, Abouljoud M, Moonka D. Disease - specific waitlist outcomes in liver transplantation – a retrospective study. Transpl Int 2021 Mar;34(3):499-513. [doi: 10.1111/tri.13814]
46. Su F, Yu L, Berry K, et al. Aging of liver transplant registrants and recipients: trends and impact on waitlist outcomes, post-transplantation outcomes, and transplant-related survival benefit. Gastroenterology 2016 Feb;150(2):441-453. [doi: 10.1053/j.gastro.2015.10.043] [Medline: 26522262]
47. Wong RJ, Singal AK. Trends in liver disease etiology among adults awaiting liver transplantation in the United States, 2014-2019. JAMA Netw Open 2020 Feb 5;3(2):e1920294. [doi: 10.1001/jamanetworkopen.2019.20294] [Medline: 32022875]
48. McCabe P, Wong RJ. More severe deficits in functional status associated with higher mortality among adults awaiting liver transplantation. Clin Transplant 2018 Sep;32(9):e13346. [doi: 10.1111/ctr.13346] [Medline: 29979466]
49. Lee C, Yoon J, Schaar MVD. Dynamic-deephit: a deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. IEEE Trans Biomed Eng 2020 Jan;67(1):122-133. [doi: 10.1109/TBME.2019.2909027] [Medline: 30951460]

Abbreviations

C-index: concordance index
CC: cryptogenic cirrhosis
CEC: competing event coherence
CoxPH: Cox proportional hazards
DDLT: deceased donor liver transplantation
INR: international normalized ratio
LDLT: Living donor liver transplantation
LT: liver transplantation
MASH: Metabolic dysfunction-associated steatohepatitis
MELD: Model for end-stage liver disease
MELD-Na: model of end-stage liver disease-sodium
ML: machine learning
REB: Research Ethics Board
RSF: random survival forest
SRTR: Scientific Registry of Transplant Recipients
UHN: University Health Network
UNOS: United Network for Organ Sharing

XSL•FO
RenderX

Edited by J Sarvestan; submitted 01.Nov.2024; peer-reviewed by C Sun, X Fu; revised version received 11.Aug.2025; accepted 12.Aug.2025; published 29.Jan.2026.

Please cite as:
Punchhi G, Sun Y, Tan E, Hlaing NKT, Liu C, Asrani S, Rambhatla S, Bhat M
Forecasting Waitlist Trajectories for Patients With Metabolic Dysfunction–Associated Steatohepatitis Cirrhosis: A Neural Network Competing Risk Analysis
J Med Internet Res 2026;28:e68247
URL: https://www.jmir.org/2026/1/e68247
doi:10.2196/68247

© Gopika Punchhi, Yingji Sun, Eunice Tan, Naomi Khaing Than Hlaing, Chang Liu, Sumeet Asrani, Sirisha Rambhatla, Mamatha Bhat. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

End-to-End Platform for Electrocardiogram Analysis and Model Fine-Tuning: Development and Validation Study

Lucas Bickmann[1,2], MSc; Lucas Plagwitz[1], MSc; Antonius Büscher[1,3], Dr med; Lars Eckardt[3], Prof Dr med; Julian Varghese[2], Prof Dr med

[1]Institute of Medical Informatics, University of Münster, Münster, Germany

[2]Institute of Medical Data Science, Otto-von-Guericke University Magdeburg, Leipziger Str. 44, Building 2, Magdeburg, Germany

[3]Clinic for Cardiology II: Electrophysiology, University Hospital Münster, Münster, Germany

Corresponding Author:
Julian Varghese, Prof Dr med
Institute of Medical Data Science, Otto-von-Guericke University Magdeburg, Leipziger Str. 44, Building 2, Magdeburg, Germany

Abstract

Background: Electrocardiogram (ECG) data constitutes one of the most widely available biosignal data in clinical and research settings, providing critical insights into cardiovascular diseases as well as broader health conditions. Advancements in deep learning demonstrate high performance in diverse ECG classification tasks, ranging from arrhythmia detection to risk prediction for various diseases. However, the widespread adoption of deep learning for ECG analysis faces significant barriers, including the heterogeneity of file formats, restricted access to pretrained model weights, and complex technical workflows for out-of-domain users.

Objective: This study aims to address major bottlenecks in ECG-based deep learning by introducing ExChanGeAI, an open-source, web-based platform designed to offer an integrated, user-friendly platform for ECG data analysis. Our objective is to streamline the entire workflow—from initial data ingestion (regardless of device or format) and intuitive visualization to privacy-preserving model training and task-specific fine-tuning—making advanced ECG deep learning accessible for both clinical researchers and practitioners without machine learning (ML) expertise.

Methods: ExChanGeAI incorporates robust preprocessing modules for various ECG file types, a set of interactive visualization tools for exploratory data analysis, and multiple state-of-the-art deep learning architectures for ECGs. Users can choose to train models from scratch or fine-tune pretrained models using their own datasets, while all computations are performed locally to ensure data privacy. The platform is adaptable for deployment on personal computers as well as scalable to high-performance computing infrastructures. We demonstrate the platform's performance on several clinically relevant classification tasks across 3 external and heterogeneous validation datasets, including a newly curated test set from routine care, evaluating both model generalizability and resource efficiency.

Results: Our experiments show that de novo training with user-provided, task-specific data can outperform a leading foundation model, while requiring substantially fewer parameters and computational resources. The platform enables users to empirically determine the most suitable model for their specific tasks, based on systematic validations, while lowering technical barriers for out-of-domain experts and promoting open research.

Conclusions: ExChanGeAI provides a comprehensive, privacy-aware platform that democratizes access to ECG analysis and model training. By simplifying complex workflows, ExChanGeAI empowers out-of-domain researchers to use state-of-the-art ML on diverse datasets, democratizing the access to ML in the field of ECG data. The platform is available as open-source code under the Massachusetts Institute of Technology (MIT) license.

(J Med Internet Res 2026;28:e81116) doi:10.2196/81116

KEYWORDS

health informatics; electrocardiogram; machine Learning; deep learning; end-to-end platform

Introduction

Background

Deep learning methods applied to Electrocardiogram (ECG) analyses have demonstrated their potential as practice-changing diagnostic tools, providing critical insights into heart-related diseases [1-3]. While the push for newer technologies and improved performance metrics is essential, ensuring these advancements are accessible for general use is equally important. Tools like ChatGPT (OpenAI) have demonstrated the potential for a broad and easy application of artificial intelligence (AI), allowing users to leverage sophisticated technologies without

XSL•FO
RenderX

in-depth expertise. Clinician-researchers who seek to apply machine learning (ML) to assess its potential benefits should have access to solutions that facilitate exploration and application without requiring extensive technical knowledge from data handling to data analysis. To address these challenges, there is a need for a comprehensive, end-to-end platform that integrates currently fragmented technical steps like ECG-specific data handling, preprocessing, data visualization, and model training, including transfer learning that requires cumbersome manual scripting. This would enable a seamless workflow from data loading to model deployment and would not only empower researchers to apply and train or fine-tune deep learning models for ECG analysis without programming but also facilitate reproducibility. Moreover, existing pretrained models for ECG data may not disclose model weights, which presents significant challenges (refer to "Related Work" below). While open weights empower users to use and adapt the model, they also promote reproducibility and transparency [4,5].

Additionally, the broad use of medical data is crucial for the advancement of personalized and specialized medicine but inhibits some immediate risks, such as data breaches [6]. As datasets continue to grow and come from diverse sources, ensuring data security becomes increasingly complex. To address this challenge, specialized decentralized learning techniques, such as federated learning or swarm learning, allow valuable insights to be gained without directly sharing sensitive data [7]. Furthermore, establishing uniform data standards can simplify data handling, reduce technical barriers related to varying data formats, and eliminate the need for programming, thereby making advanced analytics accessible to a broader range of users.

In this work, we introduce a novel open-source end-to-end platform for 12-lead ECGs called ExChanGeAI that streamlines essential steps of ECG analysis: (1) data loading and preprocessing of multiple input formats, (2) manual and computer-aided analysis of ECG waveform data, (3) one-click fine-tuning of classification models, allowing users to train and customize ML models with no prior expertise, (4) the trained models use the cross-platform industry-standard Open Neural Network Exchange (ONNX), enabling deployment in every instance of ExChanGeAI and facilitating the exchange of custom models across different instances, and (5) prediction of diseases with and without using pretrained models. Model sharing is supported via an integrated and adaptable WebDav file server called Model ExChanGe. The platform is built upon the principle of open-source code and open-weights, offering full transparency and control, empowering users to contribute to the advancement of ECG analysis models.

Related Work

Multiple studies and reviews have addressed ECG classification and have shown that fine-tuning and transfer learning improve classification results [8-10]. A study has reported improved model accuracy by fine-tuning networks trained on diverse datasets, demonstrating enhanced performance transitioning to smaller datasets [11]. However, the used data and pretrained models were not shared. Another study used transfer learning with convolutional neural networks (CNNs) for atrial fibrillation classification, pretraining on large public datasets and fine-tuning on smaller sets, achieving performance gains [9]. While code was available, pretrained models were not shared, and usability remains a significant barrier. Multiple reviews have summarized ECG analysis pipelines and deep learning methods, such as detailed essential pipeline steps [12] and reviews of techniques like CNNs and recurrent neural networks for arrhythmia classification [13]. The SelfONN model [14] showed competitive performance in general ECG classification on PTB-XL (Physikalisch-Technischen Bundesanstalt-extra large [National Metrology Institute of Germany]) but lacked resource sharing. Various types of autoencoders, including low-rank attention [15], long short-term memory [16], adversarial [17], and denoising [18] approaches, have been explored for feature extraction, anomaly detection, and noise handling. The low-rank attention autoencoder reported high accuracy on 2 datasets by focusing on spatial features. ECG-NET, based on long short-term memory, proclaimed high accuracy for arrhythmia classification on a single database in beat-based validation. An adversarial autoencoder with a temporal CNN published superior scores of anomaly detection for 2 datasets. The attention-based denoising autoencoder improved noisy ECG signal reconstruction. However, limitations across these studies include dataset dependence, restricted generalizability, lack of publicly available pretrained models and code, and validation variability.

In a recent study leveraging the gold-standard PTB-XL [19,20] dataset, the performance characteristics of multiple deep learning models were evaluated across a spectrum of training-data sizes [21]. Notable findings indicated that the InceptionTime and XceptionTime architectures [22,23] exhibited particularly compelling performances. Specifically, InceptionTime demonstrated superior efficacy when trained with smaller datasets, whereas XceptionTime surpassed all other models in performance as training dataset size increased. This suggests a potential trade-off between model complexity and data requirements for optimal diagnostic accuracy in this domain. Due to the demonstrated strength in low- and high-data scenarios, these leading architectures for ECG analysis are highly relevant for evaluation and inclusion in the platform, particularly in contexts where training data availability may vary, such as in medical contexts.

There have also been several claimed foundation models in the field of ECG classification [24]. To the best of our knowledge, these, however, are trained on a singular database [25] and are yet undisclosed or have closed-source code and weights [26,27] in general. In one case, the published weights are different from the original model of the paper due to privacy concerns [28]. A request to publish another trained model has been declined due to intellectual property and legal concerns [29]. They are trained with techniques, such as contrastive and masked learning. This allows for unsupervised training, but restricts the learning to the latent space. For downstream tasks, such as classification, fine-tuning is required. The publications report high scores for classification; however, additional tasks are not available in the published model. While these models mark significant progress in the field, they often grapple with issues, such as overfitting to specific datasets, limited scalability, or insufficient handling

XSL•FO
RenderX

of the variability and quality complications intrinsic to diverse ECG datasets.

Despite advances in ECG analysis and deep learning, the current workflows remain complex, requiring manual data transformation, preprocessing, and the use of separate tools for the visualization and analysis of ECGs, as well as for training and fine-tuning deep learning models. This fragmentation multiplies technical burdens, hinders reproducibility, and acts as a major barrier for widespread clinical adoption. Some frameworks exist to reduce the boilerplate of ML, such as the graphical tool Orange (University of Ljubljana) [30]. It does not require code, yet the workflow has to be set up manually per drag-and-drop, and it does not include any of the ECG analysis tools, such as QRS detection. In the case of AutoML tools, such as GAMA (General Automated Machine learning Assistant; originally developed by Pieter Gijsbers and Joaquin Vanschoren at the Eindhoven University of Technology) [31], it still remains code-centric. Data ingestion, for both methods, is not available out-of-the-box, such as for Digital Imaging and Communication in Medicine (DICOM), even ignoring the sampling rate, lead order, and other inconsistencies across datasets. Both tools remain unsuitable for end-to-end ECG analysis and training of ML models without substantial ML or data scientist expertise. Our work directly addresses this gap with ExChanGeAI, an integrated, accessible, and containerized end-to-end platform. This platform facilitates the visualization, transformation, prediction, and fine-tuning of deep learning models specifically for ECG data. It can leverage pretrained models, and it supports a broad range of formats and preprocessing steps, ensuring usability across different clinical and research settings. ExChanGeAI serves as a valuable resource for researchers, enabling efficient training and fine-tuning of deep learning models while preserving data privacy. This enhances both the accessibility and utility of advanced ECG analysis.

Methods

Overview

This section introduces the ExChanGeAI platform, a fully containerized, interoperable, and standardized end-to-end platform for ECG analysis, diagnosis prediction, and model fine-tuning. It is designed for nonexpert users, enabling advanced AI-enabled workflows in a unified interface without requiring specialized technical expertise. The open-source code is freely available under the MIT license [32].

ExChanGeAI Platform

ExChanGeAI is a containerized web application providing an integrated suite of AI-enabled ECG analysis tools for researchers and clinicians. The platform merges human expert analysis capabilities and AI predictions in a unified, interactive end-to-end platform (refer to Figure 1). The functionalities include (1) signal analysis, (2) model-based prediction, (3) model exchange and repository, and (4) semiautomated training and fine-tuning.

Figure 1. Overview of the capabilities of the end-to-end platform ExChanGeAI and its three main distinct parts: (1) Analysis, (2) The Artificial Intelligence Ecosystem, and (3) Interoperability. AI: artificial intelligence; ML: machine learning; ONNX: Open Neural Network Exchange.

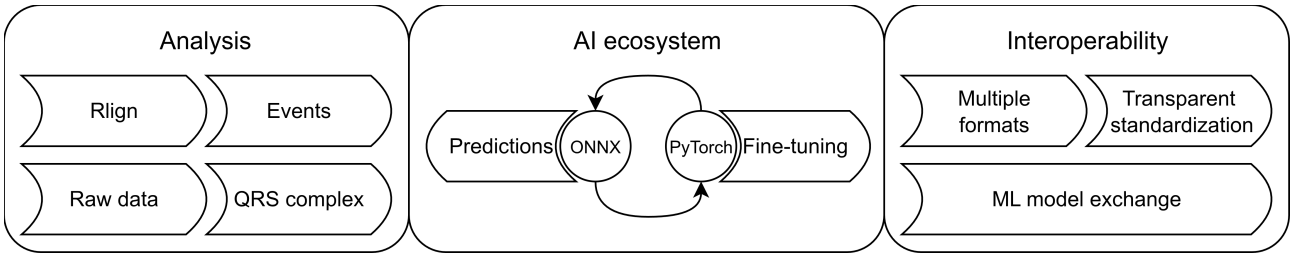

The platform is open-source, using the standardized ONNX model format and additionally supports and encourages the open-weights practice of ML models [4]. The platform consists of multiple views with different foci. The analysis view integrates the visualization of waveforms of raw signals, QRS complexes, and events (fiducial points), computed transparently by Neurokit2 [33]. Additionally, precise R-peak alignment and median beat transformations are supported through the integration of the recently published ECG-preprocessing package, Rlign [34]. These data transformations can be exported in different formats for further research. The platform focuses on resting 12-lead ECGs, displayed in a 2x6 grid in mV scale, and integrates general spatial transformations, including zooming with synchronized adaptation across all leads. This signal view has been designed in collaboration with cardiologists for their everyday use. For visualizing QRS complexes and events, lead II is conventionally applied. The prediction view uses selected models to predict diagnoses and other targets—such as QTc—based on raw signal data, offering a table with predicted diagnosis probabilities (or arbitrary keys), highlighted with a clear color-coding scheme. Dataset distributions, confusion matrices, and class-wise receiver operating characteristic curves can be computed on the platform itself, including suggested thresholds for class-specific Fmax scores. This enables researchers to optimize the threshold for each specific dataset and the health care providers' requested strategy. These differ gradually and can focus on either specificity or sensitivity, depending on the classification targets. Models are provided within the platform using an integrated and interchangeable file server, which enables the crucial aspect of model sharing (see section "Interoperability and Model Sharing"). Currently, we provide multiple models for four targets: (1) diagnostic superclasses, (2) anterior and inferior myocardial infarction (MI), (3) diverse bundle branch blocks, and (4) revascularization need (refer to section "Results"). Researchers can also train specialist models based on their own labeled data. The training and fine-tuning require no prior knowledge of ML. Currently, only 12-lead ECG data with corresponding labels are required. The backend is engineered for high performance and scalability using asynchronous views,

XSL•FO
RenderX

multithreading, and full compute unified device architecture GPU (graphics processing unit) support, including multi-GPU data-parallel training for large-scale fine-tuning.

Data Loading and Preprocessing

ExChanGeAI is designed to handle the variability of real-world ECG data. It supports various ECG input formats, across all possible sampling rates, CSV files (.csv), NumPy arrays (.npy, .npz), DICOM-stored WaveformSequences (.dcm), MATLAB (MathWorks) formatted data (.mat), general DAT files (.dat), and XML files (.xml). Major research and clinical ECG standards (PhysioNet - DAT, UK Biobank - XML) are supported. For ExChanGeAI, all data are normalized by resampling frequency signals to a configurable unified frequency target (defaulting to 100 Hz) using the Fast Fourier Transform, as previous work has shown that the sampling rate does not notably decrease the performance of ML models, but reduces computational overhead manifold [35-38]. This parameter is fully configurable, allowing users to adjust the sampling rate in a local deployment to match specific requirements. Signals not conforming to the standardized format (12-lead, 10-second waveform) are adjusted via expansion or cropping, and all scales are automatically standardized to millivolt (mV) with a 1000 analog-to-digital converter units gain, if necessary (refer to Figure 2).

Figure 2. Flowchart of the preprocessing applied to any electrocardiogram data while being loaded into the application, independent of the file format. ADC: analog-to-digital converter.

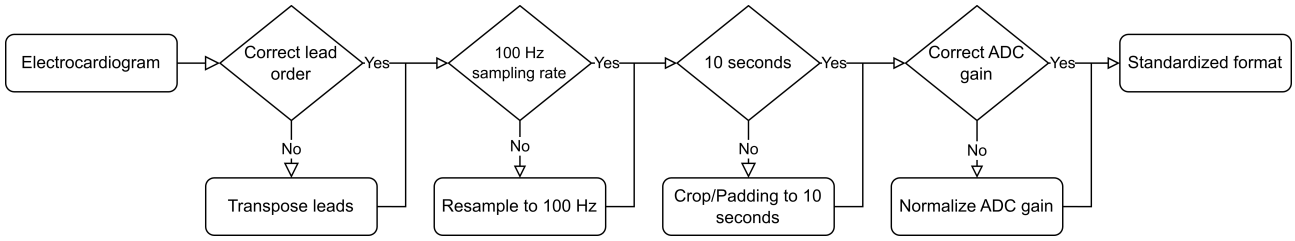

Interoperability and Model Sharing

The platform enables training of new models in a secure and privacy-preserving manner. Still, as seen with many publications in the medical domain, pretrained models are not made public [26,29] or depend on external libraries and require specific versions [28]. To promote the open and interoperable ML standard, this work adopts the ONNX as the primary used format. Therefore, our platform is compatible with all ONNX models, honoring the current operation set (Opset 20 and below), and PyTorch models, if the given model structure is provided alongside. The models are not specified with any special requirements, except for a dynamic batch size export. With the use of ONNX, this work aims to ensure that the trained model is widely accessible and interoperable. Therefore, a model sharing interface, called Model ExChanGe, is integrated into the platform, where curated pretrained models are automatically synced and made available for prediction as well as fine-tuning. Additional models can be published into the repository, or your own WebDav instance can be set up.

By default, ExChanGeAI provides 3 baseline model architectures—each as pretrained and untrained models. This includes the XceptionTime [23], InceptionTime [22], and the PhysioNet/CinC Challenge 2021 12-lead second-best model, DSAIL SNU (Data Science & Artificial Intelligence Laboratory Seoul National University [19,39,40])—the best model did not provide weights. Additional models can be incorporated by uploading them into the platform or using the default model exchange file server. We incorporate all evaluated models to extend the research community with open-source and pretrained model weights.

The adoption of the ONNX industry-standard model format ensures that ExChanGeAI is not limited to a proprietary ecosystem. Models trained or fine-tuned with ExChanGeAI can be seamlessly imported, shared, or deployed across different institutions and environments, whether in research settings or clinical contexts. This plug-and-play capability allows users to leverage pretrained models, contribute their own, or integrate compatible architectures developed externally, with zero code changes and minimal configuration. As a result, the platform not only accelerates collaboration but also supports sustainable, evolving workflows as new models and data become available.

Fine-Tuning Platform

ExChanGeAI provides an interactive user interface for fine-tuning with user-supplied data, abstracting all low-level ML steps, and facilitating the exchange of prediction models among researchers. To ensure broad compatibility with various models and platforms, we natively support PyTorch and developed a custom parser for ONNX models to adapt the computation graph where applicable. ExChanGeAI is built upon PyTorch [41], which facilitates on-device training and leverages the ONNX framework during inference. ONNX's custom training runtime does not support all PyTorch operators with corresponding gradient implementations, such as functions like ReduceMin and diverse Pooling Operators (Opset≤18). While defining custom operators may solve this, it requires detailed implementation knowledge for each unknown operator, creating a significant barrier for practical applications. To overcome these limitations, we use onnx2torch to convert our uniquely parsed ONNX models into PyTorch models. This conversion enhances compatibility, allowing training across different ONNX Opset versions and using the extensive feature set of PyTorch. Crucially, our parser automatically adapts the classification head to the new number of classification targets. These conversion steps are computed independently for each fine-tuning process and are entirely transparent to the user, without requiring any programming knowledge (refer to Figure 3). We provide two distinct training methods, including fine-tuning the classification heads, where the majority of the model weights are frozen, or training the entire model. Both

XSL•FO
RenderX

methods use pretrained weights if a pretrained model is selected. The freezing of weights is handled automatically in the background, only allowing modification of the unfrozen weights or those belonging to the classification head.

Figure 3. Overview of the semiautomatic fine-tuning process in the backend of ExChanGeAI. ECG: electrocardiogram; ONNX: Open Neural Network Exchange.

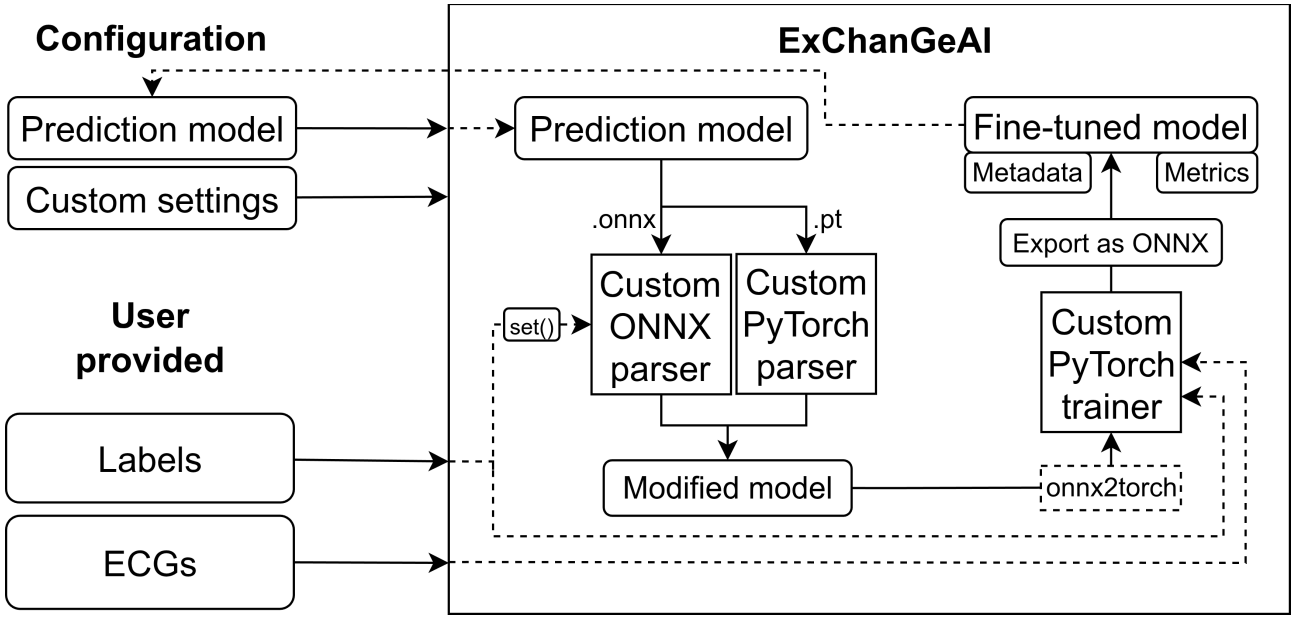

The training and fine-tuning process uses by default the AdamW optimizer [42], due to its better generalizability and convergence than the default Adam optimizer [43], and the ExponentialLR learning rate scheduler (γ=0.9). Adaptive optimization algorithms tend to be more robust, have faster convergence, and therefore improve resource usage [44]. Other optimizers, such as variants of Adam and stochastic gradient descent, are available as alternatives. Initially, a learning rate finder is executed to automatically determine the optimal initial learning rate based on the provided model and data. It has been shown that this method improves the convergence speed and reliability [45]. Furthermore, the loaded data is automatically split into stratified training/evaluation sets with an 80/20% distribution. We limit the training process to a default maximum of 50 epochs and incorporate checkpointing for models with the lowest weighted validation loss, alongside early stopping. Advanced settings, such as other optimizers, batch size, number of epochs, maximum initial learning rate, and gamma can be adapted via the user interface, if required. The best model, in addition to training and evaluation statistics, is exported and downloaded after completion. To ensure comprehensive reporting and documentation, the exported statements include (1) the number of samples, (2) distribution and corresponding labels, (3) the used base model, (4) the training and evaluation loss per epoch, and (5) the corresponding F_1-scores on the evaluation set.

Evaluation

To thoroughly assess the reliability and performance of ExChanGeAI's training process, we conduct a series of scenario-based tests. The training and prediction capabilities are one of the key features of this work and are therefore mainly evaluated via training and fine-tuning of classification models on various tasks and tests on internal and external datasets using the ExChanGeAI platform, where possible. This benchmarking is crucial for validating the platform's core value proposition (1) to empower users, even without ML expertise, and (2) to rapidly develop and deploy accurate prediction models for ECG analysis. Demonstrating robust performance across diverse datasets reinforces the platform's usability and reliability, ultimately building trust and accelerating adoption among clinicians and researchers.

Datasets

To train and fine-tune deep learning models, we use multiple targets and use the large open-access gold standard PTB-XL dataset [20]. In order to demonstrate model training under data-scarce conditions, the provided stratified fold 9 is used for training and fold 10 for intradataset testing. Age and sex, while technically predictable from ECG signals, are generally of limited clinical relevance in the context of model development, as these attributes are readily and reliably obtained through direct patient observation rather than requiring inferential prediction from ECG data.

The comparison includes a wide-ranging spectrum of ischemic heart diseases, structural heart diseases, and conduction abnormalities: (1) broad diagnostic categories, including MI, ST/T changes, conduction disturbances, and hypertrophy, (2) specific comparisons, such as anterior MI vs inferior MI, and diverse types of bundle branch blocks, including complete left (CLBBB), complete right (CRBBB), and incomplete left (ILBBB). The label distribution of PTB-XL is outlined in the Multimedia Appendix 1. A comparison baseline XceptionTime model is trained on the folds 1 - 8 of PTB-XL for all targets to outline the training performance under data-rich conditions. We also compare our baseline model against the benchmark scores of InceptionTime from the PTB-XL benchmark paper [46]. To facilitate a direct comparison of the platform's one-click training capabilities with the benchmark paper, we additionally trained

XSL•FO
RenderX

an InceptionTime model on folds 1 - 8 of PTB-XL, focusing on the superclasses.

To analyze the applicability of models across sites, we evaluate all models on interdatasets based on Yang et al [47,48], MIMIC-IV-ECG (Medical Information Mart for Intensive Care IV Electrocardiogram) [49], and Emergency Department Münster (EDMS) [38], which is an entirely new dataset from our hospital site. This helps to gauge the model's generalizability across different ECG recordings. The latter two datasets demonstrate a relatively balanced distribution with respect to age and sex, whereas the former dataset is predominantly male and represents the smallest sample size among the 3 test sets (Multimedia Appendix 2). We selected these datasets to ensure both clinical relevance and diversity in our evaluation. EDMS is a newly collected internal dataset derived from routine clinical care, providing contemporary, real-world ECG data. MIMIC-IV ECG represents one of the largest publicly available routine care datasets, enabling robust large-scale analyses. PTB-XL serves as the gold standard for annotated ECG data, offering high-quality expert labels, while the Yang et al dataset provides a similarly gold-standard resource from an entirely different geographical region, allowing us to assess model generalizability across populations. To prevent patient leakage, only one record per patient was kept in PTB-XL before the stratified split.

Some classes present slight variations, such as MIMIC and EDMS, which do not have descriptive ECG statements but general *ICD-10* (*International Classification of Diseases, 10th Revision*) codes, which are not necessarily based solely on the given ECG. We extract the signals with the corresponding fitting maps and merged superclasses. The corresponding *ICD-10* codes, or included statements, are given for each map (refer to Table 1). This includes changes such as that bundle branch blocks (BBBs) are only encoded and divided into left- and right-BBB. We evaluate the prediction performance accordingly, counting complete- and incomplete-RBBB as the superclass RBBB. Physionet classes, from the pretrained models, are mapped to the BBB and superclasses where applicable, as no MI classes are included. For MIMIC and EDMS, in instances of multiple ECGs of the same patient, only the first record was used to negate any multiple patient testing.

Table . Matched superclass and bundle branch block statements of PhysioNet 2021 and Emergency Department Münster to PTB-XL (Physikalisch-Technischen Bundesanstalt-extra large [National Metrology Institute of Germany]) classes.

Classes	Superclasses		Bundle branch blocks		
PTB-XL[a]	CD[b]	STTC[c]	CLBBB[d]	CRBBB[e]	IRBBB[f]
PhysioNet 2021	BBB[g], CLBBB|LBBB[h], CRBBB|RBBB[i], IRBBB, IAVB[j], LAnFB[k], NSIVCB[l]	TAb[m], TInv[n], LQT[o]	CLBBB|LBBB	CRBBB|RBBB	IRBBB
EDMS[p]	—[q]		LBBB	RBBB	

[a]Physikalisch-Technischen Bundesanstalt-extra large.
[b]CD: conduction disturbance.
[c]STTC: ST/T Changes.
[d]CLBBB: complete left bundle branch block.
[e]CRBBB: complete right bundle branch block.
[f]IRBBB: incomplete right bundle branch block.
[g]BBB: bundle branch block.
[h]LBBB: left bundle branch block.
[i]RBBB: right bundle branch block.
[j] IAVB: first-degree atrioventricular block.
[k]LAnFB: left anterior fascicular block.
[l]NSIVCB: nonspecific intraventricular conduction disorder.
[m]Tab: T wave abnormal.
[n]TInv: T wave inversion.
[o]LQT: prolonged QT interval.
[p]EDMS: Emergency Department Münster.
[q]Not available.

Table 2 shows a comprehensive overview of all extracted targets, the number of samples, and mapped *ICD-10* codes across the different external test datasets. Most datasets are uncurated and reflect real-world implications, in contrast to semicurated datasets, such as PTB-XL. There may be bad quality data, as well as discrepancies between *ICD-10* codes and mapped classes. The codes may be based on other electronic health data than the ECG, and differences may occur due to indifference between suspected and confirmed diagnoses. Label noise, which is only present in MIMIC and EDMS, may lead to underestimation of classification performance [50]. This, however, allows us to compare the models across real-world

XSL•FO
RenderX

data, showing possible impact on clinical care. An additional case study with a manually annotated gold standard is evaluated on the internal EDMS dataset. To demonstrate the advanced classification task of revascularization ("does the patient require revascularization?"), which is not available in PTB-XL, the models are trained and fine-tuned using the new EDMS dataset. The labels of revascularization are case-based if the patient has been treated with a revascularization. It consists of 240 positive and negative cases each, whereas negative cases are only a stratified subset of the complete annotated dataset. An additional stratified subset (20%) of these data points is kept as testing data.

XSL•FO
RenderX

Table . External test datasets and their class distribution across all categories, their mapping from diagnostic statements or *ICD-10* (*International Classification of Diseases, 10th Revision*) codes to PTB-XL (Physikalisch-Technischen Bundesanstalt-extra large [National Metrology Institute of Germany] classes, and the data format, including sampling rate, analog-to-digital converter gain, and additional annotations.

Publication, notes, and target	Classes
Yang and Feng [48]	
500 Hz based on dataset labels	
Superclasses	• HYP[a] (HEH[b]): 647 • CD[c] (IAVB[d], IIAVB1[e], IIIAVB[f], BBB[g], LAFB[h], NICD):[i] 1974
Bundle branch blocks	• CLBBB:[j] 43 • CRBBB:[k] 328 • IRBBB:[l] 1051
MIMIC-IV-ECG[m]	
500 Hz with 200 adu/mV gain based on *ICD-10*[n] codes	
Superclasses	• HYP (I11, I51.7): 500 • MI[o] (I21, I22): 500 • CD (I44): 500
Myocardial infarcts	• AMI[p] (I21.0): 500 • IMI[q] (I21.1): 500
Bundle branch blocks (variation)	• LBBB[r] (I44.7): 500 • RBBB[s] (I45.1): 500
EDMS[t]	
100 Hz based on ICD-10[n] codes	
Superclasses	• HYP (I11, I51.7): 149 • MI (I21, I22): 302 • CD (I44): 255
Myocardial infarcts	• AMI (I21.0): 51 • IMI (I21.1): 43
Bundle branch blocks (variation)	• LBBB (I44.7): 73 • RBBB (I45.1): 48
500 Hz annotated through cardiologists	
Revascularization (20% test subset)	• Yes: 48 • No: 48

[a]HYP: hypertrophy.
[b]HEH: heart enlargement and hypertrophy.
[c]CD: conduction disturbance.
[d]IAVB: first-degree atrioventricular block.
[e]IIAVB: second-degree atrioventricular block.
[f]IIIAVB: third-degree atrioventricular block.
[g]BBB: bundle branch block.
[h]LAFB: left anterior fascicular block.
[i]NICD: nonspecific intraventricular conduction disturbance.
[j]CLBBB: complete left bundle branch block.
[k]CRBBB: complete right bundle branch block.
[l]IRBBB: incomplete right bundle branch block.
[m]MIMIC-IV-ECG: Medical Information Mart for Intensive Care IV Electrocardiogram.
[n] *ICD-10: International Classification of Diseases, 10th Revision.*

XSL•FO
RenderX

[o]MI: myocardial infarction.

[p]AMI: anterior myocardial infarction.

[q]IMI: inferior myocardial infarction.

[r]LBBB: left bundle branch block.

[s]RBBB: right bundle branch block.

[t]EDMS: Emergency Department Münster.

Model Selection

We use the best-performing models based on the aforementioned previous research [21]. The study has shown that the InceptionTime performs well with less data in comparison to XceptionTime, but its performance lags behind when larger datasets are used. Therefore, we train a baseline model on XceptionTime, as its capability exceeds InceptionTime due to the large amount of data available for the baseline model. In comparison to the InceptionTime and XceptionTime models, we evaluate the DSAIL SNU PhysioNet 2021 model [39,40,51], the PhysioNet 2021 competition leader with available weights, and the only available foundation model, ECG-FM (Electrocardiogram Foundation Model) [28]. We aim to assess the effectiveness and improvements gained using ExChanGeAI's training and fine-tuning capabilities and the possible use of pretrained models, especially in resource-constrained environments with very few data points.

Preprocessing and Training

We evaluated the various architectures using two training strategies: (1) fine-tuning only the classification head and (2) training all layers. Trainings were conducted using the default settings of ExChanGeAI (commit number 4d862c04) to maintain consistency and integrity.

Xception and InceptionTime models are trained de novo (from random initialization) using non-normalized ECG data, which has been internally validated to achieve higher performance. ECG-FM and DSAIL SNU are pretrained models on PhysioNet 2021 labels, which were then fine-tuned for each classification target on PTB-XL fold 9 to demonstrate fine-tuning capabilities. In a special case, to showcase the capability of cross-task transfer learning, a pretrained XceptionTime model (superclasses with PTB-XL folds 1 - 8) was separately fine-tuned on the revascularization task.

The ECG-FM foundation model's training data details are unknown, though it is based on PhysioNet 2021 (which includes PTB-XL), limiting the validity of intradataset evaluation. The "physionet_finetuned" model differs from published results due to inaccessible weights and requires 500 Hz, 5-second z score normalized inputs. The training of ECG-FM was implemented with custom training code due to dependency complexities. It requires a custom library, which is only compatible with the end-of-life version of Python 3.9 (Python Software Foundation). Additionally, an ONNX export is not possible with these custom functions, impeding the usage of interoperable standards and therefore the deployment into the platform. The PyTorch implementation, as an alternative, could not be used due to the outdated and unsupported versions of major libraries, resulting in dependency conflicts. The results of ECG-FM are therefore achieved outside the platform, yet are given for comparative purposes. DSAIL SNU was adapted for ONNX export and initialized with the unavailable coinput features (age and sex) using the default values and their required missing feature flags as specified in its corresponding publication, alongside the used minimum-maximum normalization in pretraining.

All ECG recordings were processed at the sampling rate required by each model. For ECG-FM, recordings that are natively 500 Hz (PTB-XL, Yang et al , MIMIC-IV-ECG) were used unchanged, while recordings available only at 100 Hz (EDMS) were up-sampled to 500 Hz via a Fast Fourier Transform to match the model's input requirement. DSAIL SNU, XceptionTime, and InceptionTime require a 10-second ECG sampled at 100 Hz. Consequently, any 500 Hz recordings (PTB-XL, Yang et al, MIMIC-IV-ECG) were down-sampled to 100 Hz using the platform's interoperable data-loading pipeline, and recordings only available at 100 Hz (EDMS) were left unchanged.

Performance Metrics

To evaluate model performance, we use the F_1-score for overall assessment and calculate the average and median for central tendency across datasets. Predictions are derived by selecting the class with the highest probability. We use the F_1-score rather than area under the curve because clinical relevance often requires accurate classification at a single operating threshold—outlining the critical balance between precision and recall—whereas area under the curve summarizes performance across all possible thresholds and may not reflect the real-world consequences of specific predictions. Additionally, we focus on the weighted F_1-score to account for the class imbalance commonly seen in medical datasets, ensuring that minority classes are appropriately represented in the evaluation. Macro F_1, accuracy, precision, recall, Brier score, and expected calibration error top-label, and classwise (macro and weighted), with bootstrapped 95% CIs, as well as per class metrics, confusion matrices, and 2-sided paired t tests for external datasets, are given in the Multimedia Appendix 3. For F_1-score evaluation, classes not present in the PhysioNet labels were removed. To reflect realistic out-of-distribution prediction, the comparison did not remove false positives. Robustness, indicated by lower IQR and coefficient of variation (CV) values, suggests consistency across datasets. Computational scaling is analyzed using the number of parameters, floating-point operations per second (FLOPs), training, and inference timings. For all architectures on the ExChanGeAI platform, estimated timings are reported as the mean with SD, evaluated using a run of 1500 ECGs from the MIMIC database, based on the superclass subset and the respective models.

These comprehensive evaluations enable us to determine how well models within ExChanGeAI perform under varied conditions, providing insights into their practical application in

XSL•FO
RenderX

diverse real-world scenarios. Through this extensive testing framework, we confirm ExChanGeAI's robustness, adaptability, and reliability for ECG analysis across multiple datasets, diagnostic statements, and applicability for different use cases.

Ethical Considerations

Collection and analysis of the EDMS dataset were approved by the responsible medical ethics committee (Ärztekammer Westfalen-Lippe, approval EDMS: no. 2022 - 494 f-S) under a waiver of informed consent in accordance with state law for health data privacy (§6 Abs. 2 Gesundheitsdatenschutzgesetz Nordrhein Westfalen (Health Data Protection Act of North Rhine-Westphalia)). The creation and analysis of the MIMIC-IV-ECG dataset were reviewed by the Institutional Review Boards of Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology, which waived the requirement for individual patient consent because the project did not impact clinical care and all protected health information was deidentified. The Yang et al dataset was approved by the Medical Ethics Committee of Chinese People's Liberation Army General Hospital (approval no S2019-318-03) under informed consent from the participants. The PTB-XL dataset is publicly available under a waiver of the Institutional Ethics Committee (approval no PTB-2020 - 1), complying with Health Insurance Portability and Accountability Act (HIPAA) standards. All waivers allow secondary analysis with given approvals under the respective data regulation and privacy protection standards.

Results

The central goal of ExChanGeAI is to make model selection and empirical comparison tractable, reproducible, and accessible within a single, seamless platform. Therefore, the interface is visually structured into different foci, such as data analysis (refer to Figure 4). The analysis view provides interactive visualization of individual ECG files. Users can select to view ECG data based on multiple views - raw time series, QRS complexes, fiducial point annotation, Rlign median beats, and Rlign time-aligned ECG. When visualizing QRS complexes, the interface displays overlaid waveforms, potentially highlighting morphological features. For raw time series visualization, the platform presents the standard 12-lead ECG signals as separate plots, allowing for detailed inspection of each lead's waveform. The fine-tuning view displays options for model selection, training method, and custom model naming. A bar chart visualization summarizes the distribution of labels within the loaded dataset, presenting counts for categories, such as "CLBBB," "IRBBB," and "CRBBB" as shown in Figure 5. Numerical dataset characteristics, including the total number of imported ECGs and labels, are presented as well.

Figure 4. Overview of the ExChanGeAI web interface, showing the "Analyse" page with a sample electrocardiogram in QRS-waveforms. ECG: electrocardiogram.

XSL•FO
RenderX

Figure 5. Overview of the ExChanGeAI web interface, showing the "Finetune" page, showing the default necessary parameters, and the data distribution across labels of an example bundle-branch-block dataset. ECG: electrocardiogram.

Table 3 shows the weighted F_1-scores of the Xception-, InceptionTime, DSAIL SNU, and ECG-FM architectures across the different classification targets on the test datasets. The tables are organized such that columns represent different models, while rows represent various classification tasks and datasets. Specifically, each row corresponds to a particular classification task evaluated across different test sets (PTB-XL, Yang et al, MIMIC-IV, and EDMS). The most comprehensively de novo–trained model XceptionTime serves as a reference (Training PTB-XL folds 1-8) for assessing performance scaling with increased data availability.

XSL•FO
RenderX

Table . Performance evaluation of various models on electrocardiogram classification tasks across the test datasets.

Model	XceptionTime[ab]		InceptionTime[ab]	DSAIL SNU[ac]		ECG-FM[d]	
	Training folds (1-8)	Training folds (9)		Fine-tune	Pretrained PhysioNet 2021	Pretrained PhysioNet 2021	Fine-tune
Training PTB-XL[e] folds	1-8	9	9	9	—[f]	—	9
Trained layers	All	All	All	Head	—	—	Head
Superclasses							
PTB-XL	0.792	0.686[g]	0.651	0.536[h]	0.038[h]	0.174[h]	0.690[i,h]
Yang et al	0.335	0.647[i]	0.064	0.585[g]	0.402	0.491	0.173
MIMIC-IV[j]	0.371	0.374[i]	0.358[g]	0.323	0.055	0.192	0.355
EDMS[k]	0.432	0.387[i]	0.379[g]	0.331	0.039	0.333	0.216
Myocardial infarcts							
PTB-XL	0.938	0.853[g]	0.902[i]	0.685[h]	—	—	0.685[h]
MIMIC-IV	0.753	0.734[i]	0.726[g]	0.566	—	—	0.403
EDMS	0.566	0.484	0.584[i]	0.343	—	—	0.532[g]
Bundle branch blocks							
PTB-XL	0.911	0.912[i]	0.891[g]	0.832[h]	0.000[h]	0.016[h]	0.790[h]
Yang et al	0.730	0.101	0.792[i]	0.496	0.007	0.089	0.617[g]
MIMIC-IV	0.825	0.820[i]	0.819[g]	0.333	0.000	0.028	0.087
EDMS	0.739	0.827[i]	0.732[g]	0.622	0.000	0.118	0.248
Revascularization							
EDMS	0.750[l]	0.688[i]	0.645[g]	0.635	—	—	0.603
Weighted F_1 (interdataset only; excluding PTB-XL)							
Mean (SD)↑	0.611 (0.188)	0.562[g] (0.243)	0.567[i] (0.251)	0.470 (0.137)	—	—	0.359 (0.193)
Median (IQR)↑	0.73 (0.432-0.750)	0.647[i] (0.387-0.734)	0.645[g] (0.379-0.732)	0.496 (0.333-0.585)	—	—	0.355 (0.216-0.532)
CV[m]↓	0.308	0.433[g]	0.443	0.290[i]	—	—	0.538

[a]Trained via ExChanGeAI.
[b]De novo training.
[c]DSAIL SNU: Data Science & Artificial Intelligence Laboratory Seoul National University
[d]ECG-FM: Electrocardiogram Foundation Model; used custom code for the training platform.
[e]PTB-XL: Physikalisch-Technischen Bundesanstalt-extra large.
[f]Not applicable.
[g]Second best model in this category.
[h]These models were pre-trained on datasets including PTB-XL; results on this specific target should be interpreted as in-distribution evaluations.
[i]Best model in this category.
[j]MIMIC-IV: Medical Information Mart for Intensive Care IV.
[k]EDMS: Emergency Department Münster.
[l]Transfer learning based on reference XceptionTime (superclasses with folds 1 - 8).
[m]CV: coefficient of variation.

For illustration, the first row presents superclass classification scores on the PTB-XL test dataset, whereas the reference XceptionTime (trained on PTB-XL folds 1 - 8) achieves a weighted F_1-score of 0.792, and XceptionTime, trained on fold 9 only, reaches 0.686. As an example for MI, the InceptionTime model achieves the second-best F_1-score across both

XSL•FO
RenderX

inter-datasets (0.726 on MIMIC-IV and 0.584 on EDMS), while XceptionTime achieves a slightly higher score on the former (0.734), and ECG-FM on the latter (0.532). The last rows show the aggregated statistics, showing Xception and InceptionTime have the best average and median F_1-scores, outlining the top overall performing models, while DSAIL SNU shows the best IQR and CV values, exhibiting the most robust scores across external datasets. As anticipated, increasing the amount of training data leads to improved performance. Importantly, ExChanGeAI is able to handle this scalability, achieving better outcomes as more data are incorporated (see reference model "XceptionTime" in Table 3). For example, on the PTB-XL dataset, test performance on the superclasses and myocardial infarct targets improves substantially—by 15.4% and 9.96% respectively—when XceptionTime is trained on folds 1‑8 compared to training only on fold 9.

XceptionTime and InceptionTime, representing architectures trained de novo on PTB-XL, meaning with random initialization and without any prior training at all, often achieved the highest results across classification tasks. In contrast, the pretrained models, DSAIL SNU and ECG-FM, exhibited a more nuanced performance profile in our limited data setting. Initially, both models demonstrated suboptimal classification accuracy, especially on datasets outside of their pretraining domain (PhysioNet 2021). Fine-tuning them on a single PTB-XL fold for each classification target led to significant improvements for both DSAIL SNU and ECG-FM. However, they were outperformed by the de novo InceptionTime (8 out of 9 targets) and XceptionTime models (7 out of 9) within our evaluation. Still, fine-tuned DSAIL SNU exhibited the best robustness (lowest IQR and CV), suggesting stable results across disparate external cohorts despite lower mean and median F_1-scores. This increase in intradataset performance does not always translate to interdataset performance, as expected, due to overfitting to the dataset distribution. Overall, XceptionTime and InceptionTime trained from scratch showed the highest average and median F_1-scores across all evaluated classification tasks.

The InceptionTime model trained on folds 1‑8, as reported in the PTB-XL benchmark paper [46], achieves a macro F_1-score of 0.7495. In comparison, our reference XceptionTime model attains a macro F_1-score of 0.768, while a comparable training (not in Table 3) of InceptionTime achieves 0.7707. These results demonstrate that ExChanGeAI's training capabilities not only match but also surpass established benchmark baselines, all without reliance on additional resources. Additionally, the platform's flexible workflow enables rapid prototyping for novel tasks, such as revascularization prediction (absent in PTB-XL), trained and evaluated using the new EDMS dataset. Here, transfer learning, based upon the reference XceptionTime (0.750), increased the F_1-score by 9% relative to de novo XceptionTime (0.688).

Additionally, it has to be noted that the foundation model ECG-FM is the largest with over 90 million parameters, followed by DSAIL SNU (2M), Xception (401K), and InceptionTime (457K). In terms of computational complexity, ECG-FM is the most demanding (14 GFLOPS), followed in descending order by Inception-(460 MFLOPS), XceptionTime (256 MFLOPS), and DSAIL SNU (89 MFLOPS). The inference timings on a 6-core Zen4 CPU correspond to mean 27 (SD 33.78) ms (XceptionTime), mean 26 (SD 36.56) ms (InceptionTime), and mean 29.5 (SD 13.9) ms (DSAIL) using the ExChanGeAI platform. Training with 1500 training samples is estimated with mean 13180 (SD 44) ms per epoch and 8.79 ms per sample (XceptionTime), mean 19520 (SD 153) ms per epoch and 13.01 ms per sample (InceptionTime), and mean 10210 (SD 34) ms per epoch and 6.80 ms per sample (DSAIL). Comparing the classification performance against the computational complexity, XceptionTime and InceptionTime stand out as the top performers. All models trained on the PTB-XL dataset are available in Multimedia Appendix 4.

Discussion

Overview

Our evaluation of ExChanGeAI on established architectures reveals several key insights into model selection, particularly in data-constrained scenarios. The end-to-end platform streamlines both training and fine-tuning, yielding robust performance metrics across diverse ECG classification tasks.

Comparison With Prior Work

Our evaluation shows training on limited data and cross-dataset testing exposes inherent generalization gaps and variability in performance—a major difference compared to the often overoptimistic intradataset results seen in the literature. Consequently, the near-perfect accuracy metrics—in intratest set and simple tasks, such as tachy- and bradycardia prediction [28]—are not reproducible when models are evaluated on external, independent datasets. However, when models are evaluated using intradataset testing and ample training data are available, achieving high scores becomes more feasible and reproducible on external datasets (see the baseline XceptionTime model in the Multimedia Appendix 3). As expected and in line with previous findings [52,53], all models exhibited performance drops on external datasets. Yet, as an important factor, the end-to-end platform training surpasses the established benchmark InceptionTime model, outlining the competitive training performance of the platform without requiring expert knowledge.

Principal Findings

XceptionTime models were particularly notable for their parameter efficiency and competitive accuracy, reaffirming their architectural strength. Notably, learning from scratch proved to be a strong alternative to transfer learning, as de novo XceptionTime and InceptionTime models often outperformed fine-tuned pretrained models despite having fewer parameters. However, performance variability was observed across different classification tasks and datasets, as expected, indicating a sensitivity to dataset-specific scaling and parameter optimization within specific model architectures. Pretrained models, while anticipated to leverage their extensive prior knowledge, presented a mixed picture in our data-limited scenarios: while fine-tuning improved their performance, they generally did not consistently surpass the de novo trained XceptionTime and InceptionTime models. However, the pretrained model did

XSL•FO
RenderX

exhibit enhanced robustness against performance degradation across external datasets in most cases, compared to de novo trained models. Among all, DSAIL SNU demonstrated the lowest performance variance, underscoring its robustness.

Limitations

First, while pretrained models offer potential advantages, their benefits are not guaranteed in data-constrained scenarios. Training from scratch within ExChanGeAI frequently yielded top results. This underscores the critical importance of empirical validation and careful model selection tailored to each dataset and use case. Second, the inherent influence of model architecture on performance, coupled with the relative consistency of the subsequent training process across architectures, underscores the value of an end-to-end platform that simplifies exploration and deployment of diverse, yet effective, models. Third, while the evaluation was conducted using data-constrained scenarios, rigorous validation across diverse external datasets and the baseline comparison model also outlines the advantage of more training samples; however, these may be difficult to obtain in a clinical setting. Fourth, while the given models can be trained outside the platform, with even more customization, the usage of ExChanGeAI reduces many technical burdens, facilitating faster deployment as it eliminates the need for code for data ingestion, preprocessing, training, and evaluation for new models. Fifth, all evaluations have been conducted with a 100 Hz sampling rate by default, which, according to multiple research papers, does not notably decrease the classification performance. However, downsampling may lead to a loss of high-frequency clinical details, such as fragmentation and notches. Researchers should be aware that this loss of fidelity may be critical for specific pathologies not covered in the current classification tasks, though the platform allows for higher sampling rates if required. Sixth, defining “revascularization” by treatment status serves as a proxy for actionable clinical need. While this implies that the label incorporates medical decision-making alongside pathology, predicting this outcome remains a clinically vital advancement for identifying patients requiring urgent intervention. Seventh, while the platform supports seamless deployment of ONNX-compatible architectures, we acknowledge that integrating foundation models with external dependencies or specific libraries (eg, ECG-FM) currently requires execution via external scripts rather than the native end-to-end platform. Eighth, the DSAIL SNU model replaces missing age/sex with default values and missing flag indicators, exactly as it was pretrained. If demographic data were available, its performance may improve beyond the results reported here. Finally, we contributed to the evaluation of the novel revascularization task using a stratified 20% hold-out subset of our EDMS cohort. Consequently, these new results can serve as an internal validation only, and the generalizability to external cohorts cannot be guaranteed.

Future Work

While acknowledging potential limitations for expert users seeking highly specialized customizations, the platform’s modular design allows for the future integration of additional compatible architectures, expanding its versatility. The main focus of possible future work could be the integration of explainable or interpretable ML, including its visualization for each prediction.

Conclusions

A major strength of ExChanGeAI is its ability to democratize advanced deep learning for ECG analysis. By integrating pretrained, fine-tuned, and untrained models within a unified interface, ExChanGeAI overcomes significant barriers associated with data loading, model-specific installation, environment setup, and code dependencies, particularly benefiting nonexperts and general-purpose applications. This not only enables rapid prototyping and empirical validation by both experts and nonexperts but also encourages open science and sharing of ready-to-use models for collaborative research. Ultimately, ExChanGeAI aims to enhance the accessibility of deep learning models and reduce operational overhead, facilitating broader adoption and accelerating progress. This approach not only minimizes human error and technical debt but also supports best practices for reproducible research and clinical validation. Limitations are mainly posed by the available data and infrastructure, even though the training, on modern machines, becomes significantly easier due to the large increase in computational power in recent years and wider adoption of specialized hardware, such as GPUs and neural processing units.

In conclusion, this work introduced ExChanGeAI, a novel open-source platform designed to streamline and democratize the application of deep learning for ECG analysis. Our results demonstrate the effectiveness of ExChanGeAI across both conventional and state-of-the-art deep learning models—even with limited data—and highlight that pretrained models are not always superior in data-constrained scenarios. Regular empirical benchmarking and model selection remain crucial. By promoting accessibility, reproducibility, and systematic model comparison, ExChanGeAI broadens participation in deep learning research and clinical adoption in ECG analysis.

Acknowledgments

This work was partially supported by the Interdisciplinary Centre for Clinical Research (IZKF) Münster, Germany (grant SEED/020/23 to AB). We acknowledge support by the Open-Access Publication Fund of the Medical Faculty of the Otto-von-Guericke University Magdeburg. The authors acknowledge the use of generative artificial intelligence (AI) technology (ChatGPT, version 4o, from OpenAI, 2024) only for editing the clarity and language refinement during manuscript preparation. The research idea, content, literature research, citations, methods, results, and conclusions were purely developed by the authors. The wording suggestions were reviewed and verified by the authors.

XSL•FO
RenderX

Correspondence and requests for materials should be addressed to JV. LB was affiliated with the Institute of Medical Informatics at the University of Münster at the time of the development of the platform and is currently affiliated with the Institute of Medical Data Science at the Otto-von-Guericke University Magdeburg.

Data Availability

The datasets analyzed during this study are available in the PhysioNet or SciDB repository [20,47-49]. The EDMS dataset analyzed during this study is not publicly available due to privacy restrictions. The ExChanGeAI code and dataset generated during this study are available in the GitHub repository [32]. Supplementary information is available for this paper.

Authors' Contributions

Conceptualization: LB, JV
Data curation: LB, AB
Formal analysis: LB, LP, LE
Methodology: LB
Project administration: JV
Software: LB
Resources: AB, JV
Supervision: LE, JV
Visualization: LB
Writing–original draft: LB
Writing–review & editing: LP, AB, LE, JV

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sample size and class distribution for training samples of PTB-XL (Physikalisch-Technischen Bundesanstalt-extra large [National Metrology Institute of Germany]) across defined classification targets.
[PDF File, 48 KB - jmir_v28i1e81116_app1.pdf]

Multimedia Appendix 2
Overview of dataset characteristics, including the total number of ECG recordings, age, and sex distribution for the PTB-XL (Physikalisch-Technischen Bundesanstalt-extra large [National Metrology Institute of Germany]), MIMIC-IV-ECG (Medical Information Mart for Intensive Care IV Electrocardiogram), Yang et al., and EDMS (Emergency Department Münster) datasets.
[PDF File, 60 KB - jmir_v28i1e81116_app2.pdf]

Multimedia Appendix 3
Extended classification and calibration metrics of the trained models on electrocardiogram classification tasks across the external datasets.
[ZIP File, 8 KB - jmir_v28i1e81116_app3.zip]

Multimedia Appendix 4
Trained and fine-tuned models using the ExChanGeAI platform based on the PTB-XL (Physikalisch-Technischen Bundesanstalt-extra large [National Metrology Institute of Germany]) dataset. Training conducted using default settings (commit number 4d862c04) to ensure reproducibility.
[ZIP File, 32127 KB - jmir_v28i1e81116_app4.zip]

References

1. Lin CS, Liu WT, Tsai DJ, et al. AI-enabled electrocardiography alert intervention and all-cause mortality: a pragmatic randomized clinical trial. Nat Med 2024 May;30(5):1461-1470. [doi: 10.1038/s41591-024-02961-4] [Medline: 38684860]
2. Adedinsewo DA, Morales-Lara AC, Afolabi BB, et al. Artificial intelligence guided screening for cardiomyopathies in an obstetric population: a pragmatic randomized clinical trial. Nat Med 2024 Oct;30(10):2897-2906. [doi: 10.1038/s41591-024-03243-9] [Medline: 39223284]
3. Sau A, Pastika L, Sieliwonczyk E, et al. Artificial intelligence-enabled electrocardiogram for mortality and cardiovascular risk estimation: a model development and validation study. Lancet Digit Health 2024 Nov;6(11):e791-e802. [doi: 10.1016/S2589-7500(24)00172-9] [Medline: 39455192]

XSL•FO
RenderX

4. Widder DG, Whittaker M, West SM. Why "open" AI systems are actually closed, and why this matters. Nature New Biol 2024 Nov;635(8040):827-833. [doi: 10.1038/s41586-024-08141-1] [Medline: 39604616]
5. Polevikov S. Advancing AI in healthcare: a comprehensive review of best practices. Clin Chim Acta 2023 Aug 1;548:117519. [doi: 10.1016/j.cca.2023.117519] [Medline: 37595864]
6. Rahman M, Victoros E, Ernest J, Davis R, Shanjana Y, Islam M. Impact of artificial intelligence (AI) technology in the healthcare sector: a critical evaluation of both sides of the coin. Clin Med Insights Pathol 2024 Jan;17. [doi: 10.1177/2632010X241226887]
7. Rauniyar A, Hagos DH, Jha D, et al. Federated learning for medical applications: a taxonomy, current trends, challenges, and future research directions. IEEE Internet Things J 2024;11(5):7374-7398. [doi: 10.1109/JIOT.2023.3329061]
8. Jang JH, Kim TY, Yoon D. Effectiveness of transfer learning for deep learning-based electrocardiogram analysis. Healthc Inform Res 2021 Jan;27(1):19-28. [doi: 10.4258/hir.2021.27.1.19] [Medline: 33611873]
9. Weimann K, Conrad TOF. Transfer learning for ECG classification. Sci Rep 2021 Mar 4;11(1):5251. [doi: 10.1038/s41598-021-84374-8] [Medline: 33664343]
10. Chato L, Regentova E. Survey of transfer learning approaches in the machine learning of digital health sensing data. J Pers Med 2023 Dec 12;13(12):1703. [doi: 10.3390/jpm13121703] [Medline: 38138930]
11. Avetisyan A, Tigranyan S, Asatryan A, et al. Deep neural networks generalization and fine-tuning for 12-lead ECG classification. Biomed Signal Process Control 2024 Jul;93:106160. [doi: 10.1016/j.bspc.2024.106160]
12. Kaplan Berkaya S, Uysal AK, Sora Gunal E, Ergin S, Gunal S, Gulmezoglu MB. A survey on ECG analysis. Biomed Signal Process Control 2018 May;43:216-235. [doi: 10.1016/j.bspc.2018.03.003]
13. Ebrahimi Z, Loni M, Daneshtalab M, Gharehbaghi A. A review on deep learning methods for ECG arrhythmia classification. Expert Syst Appl: X 2020 Sep;7:100033. [doi: 10.1016/j.eswax.2020.100033]
14. Qin K, Huang W, Zhang T, Zhang H, Cheng X. A lightweight SelfONN model for general ECG classification with pretraining. Biomed Signal Process Control 2024 Mar;89:105780. [doi: 10.1016/j.bspc.2023.105780]
15. Zhang S, Fang Y, Ren Y. ECG autoencoder based on low-rank attention. Sci Rep 2024;14(1):12823. [doi: 10.1038/s41598-024-63378-0]
16. Roy M, Majumder S, Halder A, Biswas U. ECG-NET: a deep LSTM autoencoder for detecting anomalous ECG. Eng Appl Artif Intell 2023 Sep;124:106484. [doi: 10.1016/j.engappai.2023.106484]
17. Thambawita V, Isaksen JL, Hicks SA, et al. DeepFake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine. Sci Rep 2021 Nov 9;11(1):21896. [doi: 10.1038/s41598-021-01295-2] [Medline: 34753975]
18. Singh P, Sharma A. Attention-based convolutional denoising autoencoder for two-lead ECG denoising and arrhythmia classification. IEEE Trans Instrum Meas 2022;71:1-10. [doi: 10.1109/TIM.2022.3197757]
19. Goldberger AL, Amaral LAN, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet. Circulation 2000 Jun 13;101(23):e215-e220. [doi: 10.1161/01.CIR.101.23.e215]
20. Wagner P, Strodthoff N, Bousseljot RD, et al. PTB-XL, a large publicly available electrocardiography dataset. Sci Data 2020 May 25;7(1):154. [doi: 10.1038/s41597-020-0495-6] [Medline: 32451379]
21. Bickmann L, Plagwitz L, Varghese J. Post Hoc sample size estimation for deep learning architectures for ECG-classification. Stud Health Technol Inform 2023 May 18;302:182-186. [doi: 10.3233/SHTI230099] [Medline: 37203643]
22. Ismail Fawaz H, Lucas B, Forestier G, et al. InceptionTime: finding AlexNet for time series classification. Data Min Knowl Disc 2020 Nov;34(6):1936-1962. [doi: 10.1007/s10618-020-00710-y]
23. Rahimian E, Zabihi S, Atashzar SF, Asif A, Mohammadi A. XceptionTime: independent time-window xceptiontime architecture for hand gesture classification. Presented at: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 4-8, 2020; Barcelona, Spain p. 1304-1308. [doi: 10.1109/ICASSP40776.2020.9054586]
24. Han Y, Murino V, Liu X, Zhang X, Ding C. A systematic review on foundation models for electrocardiogram analysis: initial strides and expansive horizons. arXiv. Preprint posted online on Oct 13, 2025. [doi: 10.48550/arXiv.2410.19877]
25. Li J, Aguirre A, Moura J, et al. An electrocardiogram foundation model built on over 10 million recordings with external evaluation across multiple domains. arXiv. Preprint posted online on Aug 4, 2025. [doi: 10.48550/arXiv.2410.04133]
26. Wang Y, Cao X, Hu Y, et al. AnyECG: foundational models for multitask cardiac analysis in real-world settings. arXiv. Preprint posted online on Mar 3, 2025. [doi: 10.48550/arXiv.2411.17711]
27. Zhang S, Du Y, Wang W, et al. ECGFM: a foundation model for ECG analysis trained on a multi-center million-ECG dataset. Inf Fusion 2025 Dec;124:103363. [doi: 10.1016/j.inffus.2025.103363]
28. McKeen K, Masood S, Toma A, Rubin B, Wang B. ECG-FM: an open electrocardiogram foundation model. JAMIA Open 2025 Oct;8(5):ooaf122. [doi: 10.1093/jamiaopen/ooaf122] [Medline: 41113504]
29. Mathew G, Barbosa D, Prince J, Venkatraman S. Foundation models for cardiovascular disease detection via biosignals from digital stethoscopes. npj Cardiovasc Health 2024;1(1):1-13. [doi: 10.1038/s44325-024-00027-5]
30. Demšar J, Curk T, Erjavec A, et al. Orange: data mining toolbox in python. J Mach Learn Res 2013;14(1):2349-2353. [doi: 10.5555/2567709.2567736]

XSL•FO
RenderX

31. Gijsbers P, Vanschoren J. GAMA: Genetic automated machine learning assistant. J Open Source Softw 2019;4(33):1132. [doi: 10.21105/joss.01132]
32. Bickmann L. ExChanGeAI. GitHub. URL: https://github.com/VargheseLab/exchangeai [accessed 2025-12-24]
33. Makowski D, Pham T, Lau ZJ, et al. NeuroKit2: a Python toolbox for neurophysiological signal processing. Behav Res Methods 2021 Aug;53(4):1689-1696. [doi: 10.3758/s13428-020-01516-y] [Medline: 33528817]
34. Plagwitz L, Bickmann L, Fujarski M, et al. The Rlign algorithm for enhanced electrocardiogram analysis through R-peak alignment for explainable classification and clustering. arXiv. Preprint posted online on Aug 9, 2024. [doi: 10.48550/arXiv.2407.15555]
35. Salimi A, Kalmady SV, Hindle A, Zaiane O, Kaul P. Exploring best practices for ECG signal processing in machine learning. arXiv. Preprint posted online on May 14, 2025. [doi: 10.48550/arXiv.2311.04229]
36. Mehari T, Strodthoff N. Towards quantitative precision for ECG analysis: leveraging state space models, self-supervision and patient metadata. IEEE J Biomed Health Inform 2023 Nov;27(11):5326-5334. [doi: 10.1109/JBHI.2023.3310989] [Medline: 37656655]
37. Lee KS, Park HJ, Kim JE, et al. Compressed deep learning to classify arrhythmia in an embedded wearable device. Sensors (Basel) 2022 Feb 24;22(5):1776. [doi: 10.3390/s22051776] [Medline: 35270923]
38. Büscher A, Plagwitz L, Yildirim K, et al. Deep learning electrocardiogram model for risk stratification of coronary revascularization need in the emergency department. Eur Heart J 2025 Mar 29:ehaf254. [doi: 10.1093/eurheartj/ehaf254] [Medline: 40156923]
39. Reyna MA, Sadr N, Alday EAP, et al. Will two do? varying dimensions in electrocardiography: the physionet/computing in cardiology challenge 2021. Presented at: 2021 Computing in Cardiology (CinC); Sep 13-15, 2021; Brno, Czech Republic p. 1-4. [doi: 10.23919/CinC53138.2021.9662687]
40. Han H, Park S, Min S, et al. Towards high generalization performance on electrocardiogram classification. Presented at: 2021 Computing in Cardiology (CinC); Sep 13-15, 2021; Brno, Czech Republic p. 1-4. [doi: 10.23919/CinC53138.2021.9662737]
41. Ansel J, Yang E, He H, et al. PyTorch 2: faster machine learning through dynamic python bytecode transformation and graph compilation. 2024 Apr 27 Presented at: Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems; Apr 27, 2024. [doi: 10.1145/3620665.3640366]
42. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. 2019 Presented at: International Conference on Learning Representations; May 6-9, 2019 URL: https://openreview.net/forum?id=Bkg6RiCqY7 [accessed 2026-01-28]
43. Zhou P, Xie X, Lin Z, Yan S. Towards understanding convergence and generalization of AdamW. IEEE Trans Pattern Anal Mach Intell 2024 Sep;46(9):6486-6493. [doi: 10.1109/TPAMI.2024.3382294] [Medline: 38536692]
44. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv. Preprint posted online on Jan 30, 2017 URL: https://arxiv.org/abs/1412.6980 [accessed 2025-12-24] [doi: 10.48550/arXiv.1412.6980]
45. Li J, Yang X. A cyclical learning rate method in deep learning training. Presented at: 2020 International Conference on Computer, Information and Telecommunication Systems (CITS); Oct 5-7, 2020; Hangzhou, China p. 1-5. [doi: 10.1109/CITS49457.2020.9232482]
46. Strodthoff N, Wagner P, Schaeffter T, Samek W. Deep learning for ECG analysis: benchmarks and insights from PTB-XL. IEEE J Biomed Health Inform 2021 May;25(5):1519-1528. [doi: 10.1109/JBHI.2020.3022989] [Medline: 32903191]
47. Lai J, Tan H, Wang J, et al. Practical intelligent diagnostic algorithm for wearable 12-lead ECG via self-supervised learning on large-scale dataset. Nat Commun 2023 Jun 23;14(1):3741. [doi: 10.1038/s41467-023-39472-8] [Medline: 37353501]
48. Yang W, Feng Q. Offline test set of ECG multi-label classfication: Science Data Bank; 2023. [doi: 10.57760/sciencedb.07677]
49. Gow B, Pollard T, Nathanson LA, et al. MIMIC-IV-ECG: diagnostic electrocardiogram matched subset: PhysioNet. [doi: 10.13026/4nqg-sb35]
50. Frénay B, Verleysen M. Classification in the presence of label noise: a survey. IEEE Trans Neural Netw Learn Syst 2014 May;25(5):845-869. [doi: 10.1109/TNNLS.2013.2292894] [Medline: 24808033]
51. Reyna MA, Sadr N, Perez Alday EA, et al. Issues in the automated classification of multilead ecgs using heterogeneous labels and populations. Physiol Meas 2022 Aug 26;43(8):084001. [doi: 10.1088/1361-6579/ac79fd] [Medline: 35815673]
52. Martínez-Sellés M, Marina-Breysse M. Current and future use of artificial intelligence in electrocardiography. J Cardiovasc Dev Dis 2023 Apr 17;10(4):175. [doi: 10.3390/jcdd10040175] [Medline: 37103054]
53. McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: still a ways to go. Sci Transl Med 2021 Mar 24;13(586):eabb1655. [doi: 10.1126/scitranslmed.abb1655] [Medline: 33762434]

Abbreviations

AI: artificial intelligence
BBB: bundle branch blocks
CLBBB: complete left bundle branch block
CNN: convolutional neural network

XSL•FO
RenderX

CRBBB: complete right bundle branch block
CV: coefficient of variation
DICOM: Digital Imaging and Communication in Medicine
DSAIL SNU: Data Science & Artificial Intelligence Laboratory Seoul National University
ECG: electrocardiogram
ECG-FM: Electrocardiogram Foundation Model
EDMS: Emergency Department Münster
FLOP: floating-point operations per second
GAMA: General Automated Machine Learning Assistant
GPU: graphics processing unit
HIPAA: Health Insurance Portability and Accountability Act
HYP: hypertrophy
ICD-10: *International Classification of Diseases, 10th Revision*
ILBBB: incomplete left bundle branch block
MI: myocardial infarction
MIMIC-IV-ECG: Medical Information Mart for Intensive Care IV Electrocardiogram
MIT: Massachusetts Institute of Technology
ONNX: Open Neural Network Exchange
PT-XL: Physikalisch-Technischen Bundesanstalt-extra large
STTC: ST/T changes

Edited by A Coristine; submitted 23.Jul.2025; peer-reviewed by CH Li, T Ouyang, Y Yu; revised version received 01.Dec.2025; accepted 02.Dec.2025; published 30.Jan.2026.

Please cite as:
Bickmann L, Plagwitz L, Büscher A, Eckardt L, Varghese J
End-to-End Platform for Electrocardiogram Analysis and Model Fine-Tuning: Development and Validation Study
J Med Internet Res 2026;28:e81116
URL: https://www.jmir.org/2026/1/e81116
doi:10.2196/81116

© Lucas Bickmann, Lucas Plagwitz, Antonius Büscher, Lars Eckardt, Julian Varghese. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Predicting the Intention to Use Generative Artificial Intelligence for Health Information: Comparative Survey Study

Jörg Matthes[1], Prof Dr; Anne Reinhardt[2], DPhil; Selma Hodzic[1], MA, MSc; Jaroslava Kaňková[1], DPhil; Alice Binder[1], DPhil; Ljubisa Bojic[3,4], PhD; Helle Terkildsen Maindal[5], PhD; Corina Paraschiv[6], Prof Dr; Knud Ryom[5], PhD

[1]Department of Communication, University of Vienna, Waehringer Street 29, Vienna, Austria

[2]Department of Media and Communication, Ludwig-Maximilians-Universität München, Munich, Germany

[3]Digital Society Lab, Institute for Philosophy and Social Theory, University of Belgrade, Belgrade, Serbia

[4]Institute for Artificial Intelligence Research and Development of Serbia, Novi Sad, Serbia

[5]Department of Public Health, Aarhus University, Aarhus, Denmark

[6]Laboratoire Interdisciplinaire de Recherche Appliquée en Économie de la Santé (LIRAES), Université Paris Cité, Paris, France

Corresponding Author:
Jörg Matthes, Prof Dr
Department of Communication, University of Vienna, Waehringer Street 29, Vienna, Austria

Abstract

Background: The rise of generative artificial intelligence (AI) tools such as ChatGPT is rapidly transforming how people access information online. In the health context, generative AI is seen as a potentially disruptive information source due to its low entry barriers, conversational style, and ability to tailor content to users' needs. However, little is known about whether and how individuals use generative AI for health purposes, and which groups may benefit or be left behind, raising important questions of digital health equity.

Objective: This study aimed to assess the current relevance of generative AI as a health information source and to identify key factors predicting individuals' intention to use it. We applied the Unified Theory of Acceptance and Use of Technology 2, focusing on 6 core predictors: performance expectancy, effort expectancy, facilitating conditions, social influence, habit, and hedonic motivation. In addition, we extended the model by including health literacy and health status. A cross-national design enabled comparison across 4 European countries.

Methods: A representative online survey was conducted in September 2024 with 1990 participants aged 16 to 74 years from Austria (n=502), Denmark (n=507), France (n=498), and Serbia (n=483). Structural equation modeling with metric measurement invariance was used to test associations across countries.

Results: Usage of generative AI for health information was still limited: only 39.5% of respondents reported having used it at least rarely. Generative AI ranked last among all measured health information sources (mean 2.08, SD 1.66); instead, medical experts (mean 4.77, SD 1.70) and online search engines (mean 4.57, SD 1.88) are still the most frequently used health information sources. Despite this, performance expectancy (b range=0.44-0.53; all $P<.001$), habit (b range=0.28-0.32; all $P<.001$), and hedonic motivation (b range=0.22-0.45; all $P<.001$) consistently predicted behavioral intention in all countries. Facilitating conditions also showed small but significant effects (b range=0.12-0.24; all $P<.01$). In contrast, effort expectancy, social influence, health literacy, and health status were unrelated to intention in all countries, with one marginal exception (France: health status, b=−0.09; $P=.007$). Model fit was good (comparative fit index=0.95; root mean square error of approximation=0.03), and metric invariance was confirmed.

Conclusions: Generative AI use for health information is currently driven by early adopters—those who find it useful, easy to integrate, enjoyable, and have the necessary skills and infrastructure to do so. Cross-national consistency suggests a shared adoption pattern across Europe. To promote equitable adoption, communication efforts should focus on usefulness, convenience, and enjoyment, while ensuring digital access and safeguards for vulnerable users.

(J Med Internet Res 2026;28:e75648) doi:10.2196/75648

KEYWORDS

generative AI; artificial intelligence; health information–seeking; UTAUT2; Unified Theory of Acceptance and Use of Technology 2; AI adoption

XSL•FO
RenderX

Introduction

Background

When people look for health information today, they no longer only consult physicians, pharmacists, or search engines. Increasingly, they also encounter generative artificial intelligence (AI) tools such as ChatGPT or the World Health Organization (WHO)'s chatbot Sarah, which simulate human-like conversations and provide instant responses. These tools promise a new way of accessing medical knowledge: fast, convenient, and interactive. At first glance, this accessibility seems to hold great potential for reducing barriers to health information, therefore directly impacting digital health equity—defined as equitable access to and use of digital health information technology that supports informed decision-making and enhances health [1].

However, the picture is more complex. On the one hand, generative AI can offer cost-free entry points (eg, basic versions of ChatGPT or automatically displayed answers in Google search via Google's Gemini), deliver content in multiple languages, and rephrase complex medical concepts into more understandable terms. In doing so, it could strengthen patient education, address health inequalities, and help bridge communication gaps between citizens and health care providers [2,3]. On the other hand, effective use still depends on internet-enabled devices and adequate digital skills, which are not equally distributed. As a result, the very technology that appears open and inclusive may also risk exacerbating existing digital divides [4].

Moreover, unlike other types of information, health-related questions are often sensitive and personal. At the same time, the inner workings of generative AI remain opaque, and the accuracy of its outputs is not guaranteed [5]. All these tensions raise important questions about adoption: Who is most likely to turn to generative AI for health information, and what factors shape this intention? Moreover, since health communication practices and digital infrastructures differ across countries, cross-national research is urgently needed.

To address these questions, this study draws on the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) [6]. The model proposes that performance expectancy, effort expectancy, facilitating conditions, social influence, habit, and hedonic motivation shape technology use. We extend this framework by also examining the roles of health literacy and health status in predicting intention to use generative AI for health information. Using cross-national survey data from Austria, Denmark, France, and Serbia, we investigate the drivers of adoption to shed light on both individual and contextual factors that may guide the diffusion of generative AI in health contexts.

Generative AI as Novel Health Information Source

Generative AI constitutes a potentially disruptive force in the health information ecosystem [7]. However, despite its rapid advancement and widespread availability, empirical research on its role in health information–seeking remains limited [2]. At the same time, broader trends highlight the ongoing digitalization of health-related knowledge acquisition. A representative survey conducted in Germany in 2019 revealed that only 48% of respondents consulted a medical professional for their most recent health issue, while 1 in 3 turned first to the internet [8]. Similar findings show that online sources—particularly search engines—are the primary means of accessing health information, both for caregivers and the general population [9,10]. Family and friends and traditional mass media (eg, print media and health-related TV programming) rank behind medical professionals and online sources [8].

The introduction of generative AI tools like ChatGPT may shift these established hierarchies. Unlike static web content or conventional search engines, generative AI enables dialogic, personalized interactions that simulate human conversation. These features may position generative AI as a compelling alternative to established online and offline health information sources. However, current evidence suggests that trust in generative AI—especially regarding complex health-related issues—is still limited [11], which might restrict its present adoption potential to early adopters [2]. This raises questions about how generative AI integrates into the broader ecosystem of health information sources.

To address this gap, we first explore: How does the use of generative AI for health information–seeking compare to that of more established health information sources?

Explaining Predictors of Technology Adoption: UTAUT2

The UTAUT2 [6] is one of the most popular models to explain technology adoption. It builds on the technology acceptance model [12], emphasizing perceived usefulness and ease of use, and the initial UTAUT model [13], which added performance expectancy, effort expectancy, facilitating conditions, and social influence as predictors of adoption behavior. UTAUT2 extends these frameworks to consumer contexts by incorporating hedonic motivation and habit [6,14] . The UTAUT2 model has demonstrated its versatility in explaining the adoption of diverse eHealth technologies, such as wearable devices [15], health websites [16], and health apps [17]. Additionally, recent studies have highlighted its relevance in understanding the uptake of generative AI technologies [18-21], showcasing its capacity to extend beyond traditional eHealth domains. However, so far, studies on predictors of usage intentions in the context of AI health information–seeking are lacking.

Performance expectancy, a central construct in the UTAUT2 framework, reflects the belief that using technology will lead to performance benefits [22]. In the context of health information–seeking using generative AI, performance expectancy is shaped by users' perceptions of how effectively these tools can enhance their own life, including aspects such as health–decision-making and task efficiency [23]. Consequently, as users anticipate greater usefulness from adopting generative AI as a health information source, their intention to use such technologies strengthens [18-20,24]. Based on this, we propose the following hypothesis: "the higher the performance expectancy, the stronger the intention to use generative AI for health information–seeking" (H1). Effort

XSL•FO
RenderX

expectancy, closely tied to ease of use, emphasizes simplicity in technology adoption [13]. Generative AI tools like ChatGPT benefit from high effort expectancy when users find them intuitive and easy to integrate into their workflows, particularly during the early adoption phase [18,19,23,24]. Addressing usability concerns early can reduce resistance and build user confidence, strengthening behavioral intention [23,25]. Therefore, we propose that "the higher the effort expectancy, the stronger the intention to use generative AI for health information–seeking" (H2). Facilitating conditions refer to the resources, skills, and support necessary for using technology [22]. These include training, knowledge, technical assistance, and system compatibility, which significantly enhance behavioral intention and usage [6,25]. In technologically mature settings, facilitating conditions are critical for sustained adoption and user satisfaction [26]. In line with the UTAUT2, we hypothesize that "the better the facilitating conditions, the stronger the intention to use generative AI for health information–seeking" (H3). Social influence indicates the perception that peers, such as family, friends, or colleagues, believe one should adopt a technology [22]. It plays a crucial role in early adoption, where external validation often outweighs personal experience [6]. Positive reinforcement within social or professional networks can normalize usage [18,24,25,27]: If people perceive that their peers already use generative AI for health information–seeking, their own intention to do so might increase as well. We, therefore, propose that "the greater the perceived social influence, the stronger the intention to use generative AI for health information–seeking" (H4). Habit specifies the extent to which behavior becomes automatic through repetition and prior use [26]. It strongly influences behavioral intention and long-term adoption, emphasizing the importance of regular engagement with technology [6,28]. For generative AI as a health information source, fostering habitual use can solidify its integration into daily routines and enhance sustained adoption [25]. This leads us to state, "the more it is a habit to use generative AI, the stronger the intention to use generative AI for health information–seeking" (H5). Hedonic motivation refers to the enjoyment or pleasure derived from using technology, particularly relevant in consumer contexts [26]. It directly impacts behavioral intention, especially for technologies involving entertainment or leisure [29]. For generative AI like ChatGPT, it can be expected that the interaction is perceived as fun or entertaining, which can boost user engagement and drive adoption [30]. Accordingly, we suggest the following hypothesis: "the higher the hedonic motivation, the stronger the intention to use generative AI for health information–seeking" (H6).

Influence of Health Literacy and Health Status

With the growing integration of digital tools into everyday lives, the role of health literacy in online health information–seeking has garnered increasing attention. Health literacy has been conceptualized as an individual's capacity to search, access, comprehend, and critically evaluate health information, as well as to use the acquired knowledge to effectively address health-related issues [31,32]. Digital health literacy refers to these abilities in the context of digital environments [33-35]. Generally, low health literacy scores have been associated with undesirable health outcomes [36].

Research suggests that low levels of health literacy are associated with decreased trust in online health resources [37], including the outputs of AI tools [38], and lower overall adoption of online health technologies [39]. Furthermore, initial studies indicate a positive association between health literacy levels and attitudes toward the use of AI tools for medical consultations [40].

On the other hand, individuals with higher levels of health literacy are generally better equipped to critically evaluate online health information and scrutinize it in greater detail [37,41]. This heightened evaluative capacity could make them more aware of the limitations and potential risks of generative AI outputs, such as inaccurate information, bias, data privacy concerns, or oversimplified medical advice [42]. Moreover, individuals with higher health literacy are more likely to trust and use high-quality medical online resources, whereas those with limited health literacy prefer accessible but potentially less reliable sources [43]. In this context, outputs from generative AI might be perceived as lower-quality sources by highly digital health–literate individuals. As a result, while higher health literacy could foster openness to using generative AI for health purposes, it might also lead to greater skepticism or hesitancy in relying on these tools. Nonetheless, there is not enough research in the context of generative AI specifically to make conclusive predictions.

Another well-established factor in online health information–seeking, yet underexplored in the context of AI, is individuals' health status: Studies suggest that people with poor health are significantly more likely to consult the internet for health information compared to those with good health [44,45]. Being chronically ill has also been associated with increased reliance on internet-based technologies for health-related purposes [46]. This relationship can be explained by the fact that individuals in poor health often experience greater health-related concerns, which in turn heightens their motivation to seek information online.

Given these complex relationships, we propose the second research question: How does health literacy and health status influence the intention to seek health information using generative AI?

Cross-National Comparison

In this study, we investigate the predictors of generative AI adoption for health information–seeking across 4 European countries: Austria, Denmark, France, and Serbia. While these countries share certain similarities, they also display notable differences that could shape the strength of the UTAUT2 predictors on the intention to use generative AI for health purposes. Thus, this cross-national approach ensures that the observed effects are generalizable and not confined to specific national contexts or unique country conditions.

The selected countries share two key characteristics. First, all 4 countries provide universal health coverage, ensuring broad access to health care services for their populations. Second, a

XSL•FO
RenderX

significant portion of health care expenditure in these countries is publicly funded [47-50].

Despite these commonalities, there are also critical factors that differ among the countries and may shape the predictors of generative AI adoption. On the one hand, variations in digital infrastructure could significantly impact facilitating conditions, effort expectancy, and social influence as predictors of generative AI use. Denmark consistently ranks among Europe's most digitally advanced nations, boasting high internet penetration and widespread adoption of e-health solutions [51]. This strong digital ecosystem likely enhances the perceived ease of use and social endorsement of generative AI. In contrast, Austria, France, and Serbia exhibit more moderate levels of digital adoption in the context of health information, which may limit the perceived use and social norms regarding such technologies [51].

On the other hand, access to and trust in health care providers vary significantly across these countries, potentially influencing performance expectancy and social influence. In nations with robust health care systems—characterized by a high availability of medical professionals and easy access to care—individuals are more likely to rely on doctors for health advice, as they are often viewed as the most trusted source of health information [8]. Denmark exemplifies this with its high levels of public trust in the health care system [52], which may reduce the perceived benefits and social norms around using generative AI for health purposes. Conversely, in western Balkan countries like Serbia, studies report generally low levels of trust in the health care system [53]. In such contexts, individuals may be more inclined to seek alternative information sources, potentially amplifying the perceived benefits of generative AI use.

By examining these diverse national contexts, this study not only tests the universality of the UTAUT2 model but also deepens our understanding of the contextual factors that shape generative AI adoption for health purposes. We ask: How do the predictors of generative AI use for health information–seeking differ across Austria, Denmark, France, and Serbia?

Methods

Ethical Considerations

Before data collection, the study received ethical approval from the institutional review board of the Department of Communication at the University of Vienna (approval ID: 1205). All participants provided written informed consent prior to participating in the study. The data were collected in anonymized form and no personal identifiers were recorded or stored. Participants received a compensation of €1.50 (US $1.74) for completing the study through the panel provider.

Recruitment

Recruitment of participants occurred during September 2024 via *Bilendi*, an international panel provider. *Bilendi* recruited the participants via email. The panel is checked for quality and attendance on a regular basis. The study was conducted in Austria, France, Denmark, and Serbia, with participants randomly selected to achieve samples representative of age, gender, and educational background. The provider's panel sizes in the respective countries were as follows: Austria: n=60,000; Denmark: n=90,000; France: n=815,000; and Serbia: n=15,100. Per country, the study aimed to reach 500 participants.

Inclusion criteria required participants to be aged between 16 and 74 years. Additionally, respondents who completed the survey in less than one-third of the median completion time (speeders) were excluded. Completion rates (excluding screened-out participants) were high across all 4 countries, ranging from 84.3% to 89.8%. Further details on survey design, administration, and response rates are provided in the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist (Checklist 1).

Procedure and Measures

Overview

The study consisted of two components. The first component, a survey, investigated predictors of the intention to use generative AI for health information–seeking. The second component, an experimental study, explored the influence of disease-related factors on these intentions [54]. To ensure respondents shared a common understanding of the concept, the survey began with a short definition of "generative AI," describing it as technologies that engage in natural language conversations with users and generate responses in real-time. Examples such as ChatGPT, Google's Gemini, and Microsoft Copilot were provided.

The original questionnaire was developed in English and subsequently translated into German, French, Danish, and Serbian. Each translation was performed by a bilingual team member and back-translated into English by a different native speaker to ensure conceptual equivalence with the original items.

This study focused on the following constructs (a complete item list with descriptive analysis and construct reliability values can be found in Multimedia Appendix 1).

Dependent Variables

Sources of Health Information–Seeking

Participants rated the frequency with which they use 8 health information [55] sources on a 7-point Likert scale (1 = "never" to 7 = "very often"). The sources included conversations with medical professionals, pharmacists, and family or friends, as well as books, mass media, internet search engines (eg, Google and Ecosia), and generative AI.

AI Usage Intentions

Participants' willingness to use generative AI [25] for health information–seeking was assessed using 3 items (eg, "I intend to use generative AI for health information seeking"), rated on a 7-point Likert scale (1 = "strongly disagree" to 7 = "strongly agree").

UTAUT2 Predictor Variables

All UTAUT2 model [6,25,56,57] predictors were measured on a 7-point Likert scale (1 = "strongly disagree" to 7 = "strongly agree"). The predictors have been described below.

XSL•FO
RenderX

Performance Expectancy

Perceived benefits of using generative AI for health information–seeking were measured with 4 items (eg, “Using generative AI would save me time when researching health topics”).

Effort Expectancy

The perceived ease of using generative AI as a health information source was assessed with 4 items (eg, “Learning to use generative AI for health information–seeking seems easy for me”).

Social Influence

Three items measured the extent to which participants felt that important others encouraged their use of generative AI for health information–seeking (eg, “People who are important to me think that I should use generative AI for health-information seeking”).

Hedonic Motivation

The enjoyment of using generative AI for health information–seeking was assessed with 3 items (eg, “I think using generative AI for health-information seeking could be fun”).

Facilitating Conditions

Participants’ perceptions of available resources and support for using generative AI to seek health information were measured with 4 items (eg, access to devices and reliable internet, and knowledge).

Habit

The extent to which turning to generative AI when seeking health information had become a habitual behavior was measured with 3 items (eg, “I automatically turn to generative AI whenever I have questions about my health”).

Model Extension Variables

Health Literacy

Health literacy [58] was assessed with 10 items, asking participants to rate their confidence in tasks such as finding understandable health information. Responses were recorded on a 4-point Likert scale (1 = “not at all true” to 4 = “absolutely true”).

Health Status

We measured participants’ health status using 1 item (“How would you describe your current health status?”; 1 = “very poor” to 7 = “very good”).

Control Variables: Sociodemographic Variables

We further measured participants’ age, gender, and educational level.

Statistical Analysis

Power

An a priori power analysis was conducted to determine the required sample size for structural equation modeling. Assuming an anticipated effect size of 0.25, a desired statistical power of 0.95, and a significance level of .05, the analysis indicated that a minimum of 391 participants per country would be necessary to detect the hypothesized effects [59].

Analytical Plan

We used AMOS version 26 (IBM Corp) to run a latent variable, multigroup structural equation model using a maximum-likelihood estimator with full information. We computed the comparative fit index, the Tucker-Lewis Index, the chi-square to degrees of freedom ratio (χ^2/df), and the root mean square error of approximation. We also secured metric measurement invariance to be able to compare the paths across countries. We controlled for age, gender, and education (binary coded).

Results

User Statistics

In total, data were collected from 1990 respondents, comprising 502 from Austria, 507 from Denmark, 498 from France, and 483 from Serbia. The overall mean age of participants was 45.1 (SD 15.7) years, with 50.2% (n=998) identifying as female participants. In terms of educational attainment, 83.8% (n=1634) of the sample reported completing at least a medium or higher level of education (secondary level II or higher). Furthermore, 87.4% (n=787) of respondents indicated prior use of generative AI for health information–seeking (at least rarely). Detailed demographic and background characteristics of the sample are summarized in Table 1.

XSL•FO
RenderX

Table . Descriptive characteristics of survey respondents from Austria, Denmark, France, and Serbia (N=1990; September 2024).

Demographic characteristics	Overall, n (%)	Austria, n (%)	Denmark, n (%)	France, n (%)	Serbia, n (%)
Education[a]					
Secondary I or lower	356 (18.24)	93 (18.6)	136 (26.9)	109 (21.9)	18 (3.7)
Secondary II	1080 (55.36)	303 (60.3)	179 (35.3)	224 (45.0)	374 (77.4)
Tertiary	554 (28.39)	106 (21.1)	192 (37.9)	165 (33.1)	91 (18.8)
Gender					
Female	998 (50.15)	252 (49.8)	251 (49.5)	256 (51.4)	239 (49.5)
Male	992 (49.85)	250 (50.2)	256 (50.5)	242 (48.6)	244 (50.5)
Prior experience[b]					
No	1203 (60.45)	316 (62.9)	328 (64.7)	326 (65.5)	233 (48.2)
Yes	787 (39.54)	186 (37.1)	179 (35.3)	172 (34.5)	250 (51.8)

[a]Educational attainment was categorized as low (secondary level I or below) and medium or high (secondary level II or higher). In Serbia, however, representativeness was achieved by grouping educational levels into low or medium (secondary level II or below) and high (tertiary education) due to sampling limitations.

[b]Prior experience: no = “I have never used Generative AI for health information seeking” and yes = “I have used Generative AI for health information seeking at least rarely.”

Statistical tests revealed no significant differences in gender distribution across countries (χ^2_3=0.48; *P*=.92) and no significant differences in age (Kruskal-Wallis χ^2_3=2.15; *P*=.54). In contrast, educational attainment varied significantly between countries (χ^2_3=550.76; *P*<.001), reflecting sampling-related imbalances in Serbia where low versus medium or high education was assessed differently than in the other countries. Finally, prior experience with health information–seeking showed significant country differences (Kruskal-Wallis χ^2_3=30.95; *P*<.001), with higher levels reported in Serbia.

Evaluation Outcomes

Descriptive Analysis

In our first research question, we explored how generative AI compares to more established health information sources in terms of usage frequency across countries. As illustrated in Figure 1, generative AI ranks last among all measured sources, indicating that, as of autumn 2024, it is rarely used for health information–seeking (mean 2.08, SD 1.66). In stark contrast, online search engines like Google are highly used, ranking second with a mean usage frequency of 4.57 (SD 1.88), following conversations with physicians, which hold the top position (4.77, SD 1.70). Family and friends also play a significant role, ranking third (4.27, SD 1.73), alongside pharmacists (3.52, SD 1.81). In comparison, traditional mass media such as TV, newspapers, and magazines are used less frequently (2.74, SD 1.68), as are books (2.68, SD 1.70) and free magazines provided by pharmacies or health insurance companies (2.60, SD 1.65). The relative ranking of information sources was consistent across all 4 countries, with physicians, internet search engines, and family or friends occupying the top positions and generative AI ranking last. However, some variation in mean usage frequencies was observed between countries; detailed country-level results are presented in Multimedia Appendix 2.

XSL•FO
RenderX

Figure 1. Mean usage frequency of different health information sources among survey respondents (N=1990) in Austria, Denmark, France, and Serbia (95% CI; September 2024). AI: artificial intelligence.

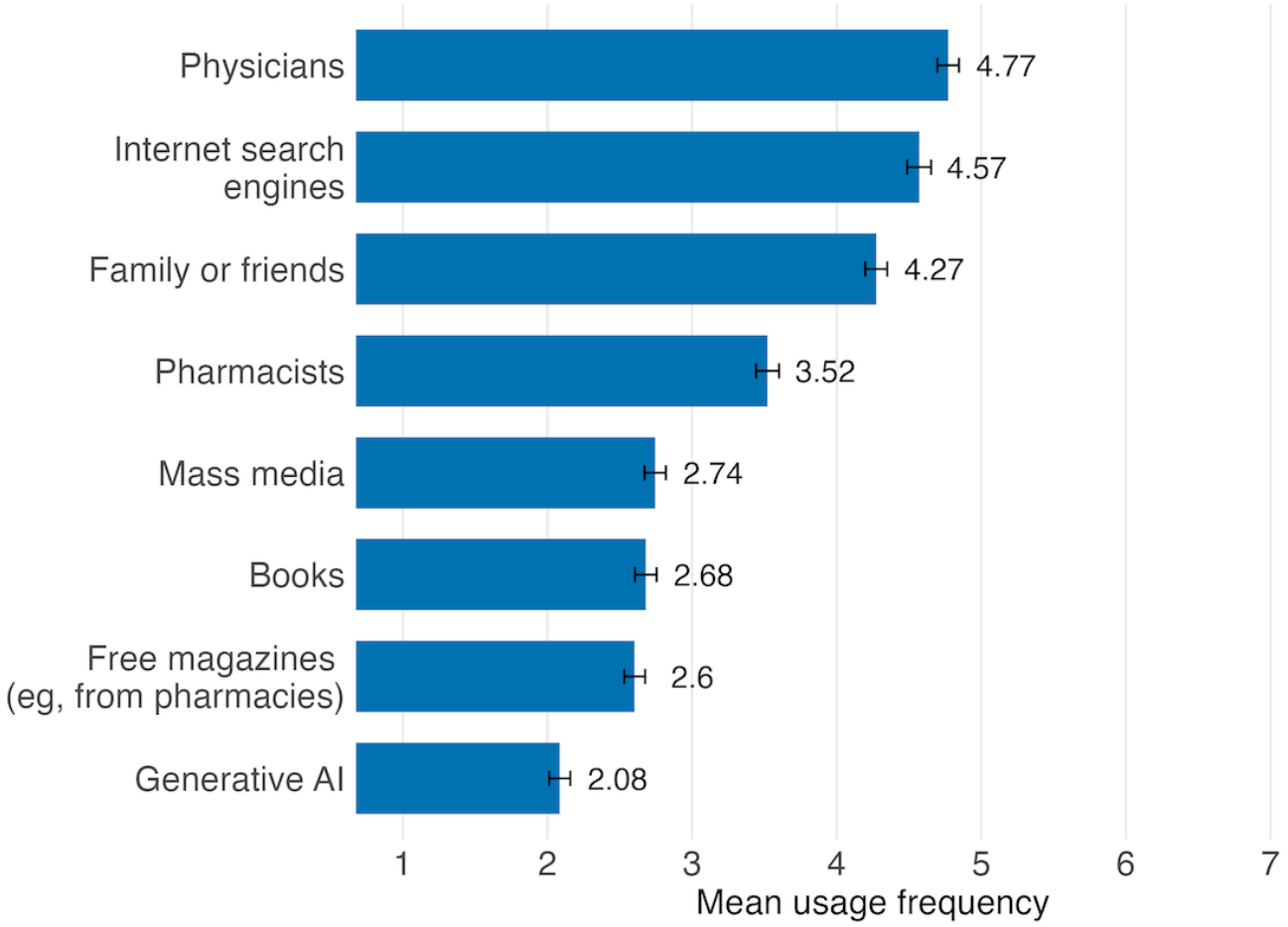

Model Evaluation

For the hypothesis tests, the results are shown in Table 2. Model fit was good (comparative fit index=0.95; Tucker-Lewis-Index=.93; χ^2/df =2.47, *P*<.001; root mean square error of approximation=0.03, 95% CI 0.03-0.03). We examined the metric measurement invariance of all latent variables by constraining all factor loadings as equal for all 4 countries. When comparing the constrained model to the unconstrained model, we found no significant difference in model fit (*P*=.16). Thus, metric invariance across countries was established.

XSL•FO
RenderX

Table . Structural equation model predicting the intention to use generative artificial intelligence (AI) for health information among survey respondents in Austria, Denmark, France, and Serbia (N=1990; September 2024).

Predictor variables	Austria[a]			Denmark[b]			France[c]			Serbia[d]		
	b	SE	*P* value	b	SE	*P* value	b	SE	*P* value	b	SE	*P* value
Performance expectancy	0.47	0.05	<.001	0.52	0.05	<.001	.053	0.05	<.001	0.44	0.05	<.001
Effort expectancy	−0.07	0.05	.20	0.03	0.05	.54	−0.11	0.05	.04	-0.02	0.06	.77
Facilitating conditions	0.12	0.04	.01	0.17	0.05	<.001	0.22	0.05	<.001	0.24	0.05	<.001
Social influence	−0.05	0.04	.29	−0.08	0.05	.17	−0.09	0.05	.10	−0.05	0.04	.27
Habit	0.29	0.04	<.001	0.32	0.04	<.001	0.28	0.05	<.001	0.28	0.04	<.001
Hedonic motivation	0.45[e]	0.06	<.001	0.22[f]	0.05	<.001	0.33	0.05	<.001	0.23[f]	0.05	<.001
Health literacy	−0.004	0.09	.97	0.04	0.10	.67	−0.02	0.08	.08	0.08	0.10	.40
Health status	−0.002	0.03	.95	0.02	0.03	.61	−0.09	0.03	.01	−0.05	0.03	.13

[a]Explained variance=0.84.

[b]Explained variance=0.80.

[c]Explained variance=0.86.

[d]Explained variance=0.79.

[e]The different subscripts in each row indicate a significant difference between paths (*P*<.05)

[f]The different subscripts in each row indicate a significant difference between paths (*P*<.05)

In line with H1, we found a highly significant positive association between performance expectancy and the intention to use generative AI for health information–seeking across all 4 countries (Austria: b=0.47, *P*<.001; Denmark: b=0.52, *P*<.001; France: b=0.53, *P*<.001; Serbia: b=0.44, *P*<.001). In contrast, H2 was not supported: effort expectancy showed no significant association with behavioral intention in any of the countries. Turning to H3, results revealed a positive association between facilitating conditions and the intention to use generative AI as a health information source, consistently observed across all 4 contexts (Austria: b=0.12, *P*=.005; Denmark: b=0.17, *P*<.001; France: b=0.22, *P*<.001; Serbia: b=0.24, *P*<.001). By contrast, no support was found for H4: perceived social influence was unrelated to behavioral intention in any of the countries. As predicted in H5, habit was positively associated with behavioral intention to use generative AI for health information–seeking throughout the sample (Austria: *b*=0.29, *P*<.001; Denmark: b=0.32, *P*<.001; France: b=0.28, *P*<.001; Serbia: *b*=0.28, *P*<.001). A similar pattern emerged for H6: hedonic motivation was significantly positively related to behavioral intention in all countries (Austria: b=0.45, *P*<.001; Denmark: b=0.22, *P*<.001; France: b=0.33, *P*<.001; Serbia: b=0.23, *P*<.001).

Finally, with regard to our second research question—which examined whether health literacy and health status predict the intention to seek health information using generative AI—we found no substantial associations. Only in France did health status show a marginal negative effect (b=−0.09; *P*=.007).

Discussion

Principal Results

This study investigated the predictors of intention to use generative AI for health information–seeking, drawing on the UTAUT2 framework and expanding it with health literacy and health status. Using cross-national survey data from Austria, Denmark, France, and Serbia, our findings show that generative AI is still only rarely used for health information–seeking. At the same time, performance expectancy, facilitating conditions, habit, and hedonic motivation consistently emerged as significant predictors of behavioral intention, whereas effort expectancy, social influence, health literacy, and health status were not related to intention. These patterns were consistent across all 4 countries, suggesting a robust set of psychological drivers underlying the early adoption of generative AI in health contexts. A detailed examination of these findings is provided as follows.

First, with regard to overall usage patterns, the data shows that generative AI currently plays only a minor role in health information–seeking: 60% of the respondents reported never

XSL•FO
RenderX

having used a generative AI tool for health-related questions. This result lends itself to 2 contrasting interpretations.

On the one hand, it challenges the popular narrative that generative AI is rapidly transforming health information–seeking behavior. Instead, the findings align with previous studies, showing that generative AI is currently infrequently used in the context of health information [2]. Traditional sources—such as medical professionals and search engines—continue to dominate [8], underscoring that generative AI has yet to achieve mainstream adoption.

On the other hand, despite persistent concerns about data privacy, algorithmic bias, and accuracy, it is noteworthy that 40% of the respondents have already experimented with generative AI for health purposes. Given that this technology only became widely accessible relatively recently, such early uptake is remarkable. From the perspective of technology adoption models, such as the Rogers Diffusion of Innovations [60], this pattern is characteristic of early adopters. It is therefore plausible to assume that the use of generative AI for health information–seeking will increase further as the technology matures and moves toward mainstream adoption.

To better understand the drivers of future uptake, we applied an extended version of the UTAUT2 model. Our findings confirmed the predictive power of performance expectancy, facilitating conditions, habit, and hedonic motivation. This aligns with prior research on digital health tools, indicating that users value usefulness, access, familiarity, and enjoyment [18,20,24].

In detail, the results show that performance expectancy—the perceived usefulness of the technology—had a strong positive effect on behavioral intention in all four countries. This finding suggests that the more respondents believe generative AI is useful to manage health-related questions, the more they will use it. Thus, if public health stakeholders or developers aim to encourage responsible AI use, they should emphasize the tangible benefits of generative AI, such as 24/7 availability, rapid response times, and the potential for personalized information. Perceived usefulness may also be fostered when individuals try out generative AI for the first time, that is, they learn that they can benefit from the technology.

At the same time, our study challenges established UTAUT2 assumptions. Effort expectancy, often seen as central to technology adoption, was not a relevant factor—possibly due to the intuitive nature of generative AI tools and the ubiquity of basic digital skills [61]. Using generative AI does not require any specific background knowledge beyond opening a webpage. Since online search engines are already the most frequently used health source, the basic skills needed for generative AI are widely present, potentially rendering effort expectancy less decisive.

Taken together, this emerging pattern—the strong effect of performance expectancy and the null effect of effort expectancy—underscores the distinction between usefulness and usability, which are closely related but not identical [62]. Usability refers to the ease of interacting with a system (eg, ease of learning and error prevention), whereas usefulness (utility) captures whether the system provides the functions and information that users actually need. Our findings suggest that in health contexts, utility is the decisive factor: people intend to use generative AI if its outputs are perceived as useful, while usability-related aspects appear less influential.

Importantly, this does not mean that barriers to adoption are absent. Rather, our findings show that they lie not in usability but in facilitating conditions—the structural and contextual resources that enable technology use. Across all countries, the availability of digital infrastructure, devices, and basic knowledge significantly shaped behavioral intention. In other words, while generative AI may be easy to use once accessed, unequal access to the necessary resources continues to pose a substantial adoption barrier. Consequently, facilitating conditions emerge as a key digital health equity concern [4]. Without adequate access, disadvantaged populations may be excluded from benefiting from generative AI, meaning that the technology risks widening rather than narrowing the digital divide in health information–seeking.

We also found that social influence—an important predictor in other studies on AI uptake [18,24]—did not play a meaningful role in shaping behavioral intention. This suggests that health-related information search is a rather personal topic, and that individuals may not always be willing to disclose what kind of information they are looking for. As a result, the intention to use generative AI for health information–seeking is largely independent of peer opinions or social norms.

In contrast, habit consistently predicted behavioral intention across all countries. From this finding, we may conclude that generative AI use for health information is likely to occur automatically, similar to how people use search engines. When individuals feel familiar with a technology, they are more likely to rely on it without conscious deliberation. However, this finding should be interpreted with caution, as the majority of participants had never used generative AI for health purposes. Much of the variance in habit may therefore reflect mere use versus nonuse. Accordingly, variables capturing initial adoption should be clearly distinguished from those measuring habit.

By including health literacy and health status as additional predictors, our study adds a novel dimension to existing research. In contrast to studies showing direct paths between these constructs and online health information–seeking [37,40,46], we found no such association for AI health information–seeking. However, each of these findings carries different implications. First, the absence of a significant association between health literacy and intention indicates that individuals' ability to understand and evaluate health information was not related to whether they reported turning to generative AI. This finding may suggest that the use of such tools is driven less by informed decision-making and more by general curiosity or interest in new technologies. Importantly, this raises concerns: people with lower health literacy may be just as likely to consult generative AI as those with higher health literacy—despite being less equipped to critically assess its outputs. Given the known risk of AI hallucinations—fabricated or inaccurate information presented in a confident tone [63]—this could lead to misinformation and, in the worst case,

XSL•FO
RenderX

harmful health decisions, as users with limited health literacy might find it difficult to distinguish between reliable and misleading content.

Second, the lack of an association between self-reported health status and intention suggests that the current use of generative AI is not primarily driven by medical need or urgency. People do not seem more likely to consult generative AI when facing a health problem; rather, usage may occur proactively or even recreationally. This challenges assumptions that such tools are primarily used in response to a health issue, and it underscores the importance of understanding user motivations beyond immediate health concerns.

Importantly, these patterns were largely consistent across all 4 countries, as confirmed by the measurement-invariant structural model. This cross-national consistency suggests that the psychological drivers of generative AI adoption in health contexts may transcend national boundaries and cultural differences, pointing to a universal set of adoption mechanisms.

Limitations

Several limitations should be acknowledged. First, due to the cross-sectional nature of this study, no causal conclusions can be drawn. Future research should therefore aim to replicate these findings using experimental or longitudinal designs. Second, we relied on self-reported data, which may be subject to social desirability bias. The use of behavioral data is thus warranted to validate these findings. Third, including additional predictors—such as individual differences or specific concerns—could provide deeper insights into the use of generative AI. Fourth, our comparative findings are based on data from only 4 countries, which limits the ability to conduct multilevel analyses. Also, as in all cross-sectional research, there is a risk of unmeasured third variables. In particular, we did not include AI trust and perceived AI risk. However, these constructs are conceptually close to performance expectancy, as trust reduces uncertainty about the system's outputs and thereby enhances expected performance gains, whereas perceived risk erodes expected utility. In this sense, they are likely to be partially reflected in the performance expectancy construct already included in our model. That said, and highlighting that our model explains around 80% of the variance, trust and perceived risk could still suppress some of the predictors we have modeled. Thus, future research should include additional constructs outside the UTAUT2 framework [64]. Finally, health status was measured with a single self-rated item. While single-item measures of subjective health may not capture the full complexity of an individual's medical condition, this approach is widely used in demographic and population health research. Prior work has demonstrated that the self-rated health item is a valid and reliable indicator, predicting key outcomes such as mortality, use of health services, and health expenditures in large-scale surveys [65]. Nevertheless, we acknowledge that a more fine-grained measure (eg, including specific chronic conditions or severity indices) could have provided additional insights, and future studies may benefit from applying such extended health measures.

Conclusions

This study applied the UTAUT2 model to investigate the factors that drive the use of generative AI for health information–seeking. Although overall usage remains limited, our findings show that performance expectancy, facilitating conditions, habit, and hedonic motivation are positively associated with behavioral intentions. These patterns, observed across all 4 countries—Austria, Denmark, France, and Serbia—suggest that current users of generative AI are likely to be early adopters: individuals who are tech-savvy, curious, and open to innovation. This aligns with the Rogers Diffusion of Innovations theory, which conceptualizes adoption as a gradual process beginning with a small, innovation-oriented segment of the population.

The lack of significant effects for effort expectancy and social influence across all countries reinforces this interpretation: early adopters tend to base their decisions on personal evaluations rather than external opinions and are rarely deterred by usability concerns. Furthermore, the fact that behavioral intention was unrelated to health status or health literacy underscores that current usage is not driven by acute medical need or advanced health literacy, but rather by interest, convenience, and technological exploration.

The cross-national consistency of these findings is particularly striking. Despite differences in health care systems, digital infrastructures, and culture, the same psychological and contextual factors influenced generative AI use in all countries surveyed. This suggests a shared adoption logic that transcends national boundaries—at least in the early stages of diffusion.

Looking ahead, these insights help illuminate how generative AI might transition from a niche tool to a widely used resource. As the technology becomes more embedded in everyday life, broader segments of the population—the so-called early and late majority—will likely demand stronger assurances of trustworthiness, safety, and added value. To enable responsible and inclusive adoption, it is therefore crucial to reduce digital access barriers, enhance transparency, and implement safeguards against health misinformation, especially for users with limited health literacy.

From a practical perspective, our findings suggest that communication strategies aiming to promote generative AI for health purposes should emphasize concrete benefits, ease of access, and even enjoyment. Rather than exclusively targeting individuals with chronic or urgent health needs, positioning generative AI as an engaging, low-barrier tool may broaden its appeal—reaching users who might otherwise be disengaged from traditional health information sources.

In sum, generative AI holds significant potential as a future health information resource—but its trajectory will depend on how well we understand and support the evolving needs of its users across different adoption phases and contexts.

XSL•FO
RenderX

Acknowledgments

The authors used ChatGPT (OpenAI) to support language editing of prewritten text sections (eg, to improve grammar and phrasing). All suggestions from the artificial intelligence tool were reviewed by the authors and revised or rejected as necessary. No content was generated solely by the artificial intelligence, and the authors remain fully responsible for the final text.

Funding

The work was supported by Circle U [2024-09 – AIHEALTH].

Data Availability

All data and supplementary materials related to this study are openly accessible via the Open Science Framework (OSF) at the following link [66].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Descriptive and reliability analysis of predictor variables.
[DOCX File, 20 KB - jmir_v28i1e75648_app1.docx]

Multimedia Appendix 2
Mean usage frequency of health information sources with 95% CI per country.
[DOCX File, 199 KB - jmir_v28i1e75648_app2.docx]

Checklist 1
CHERRIES checklist.
[DOCX File, 19 KB - jmir_v28i1e75648_app3.docx]

References

1. Braveman P. Health disparities and health equity: concepts and measurement. Annu Rev Public Health 2006;27:167-194. [doi: 10.1146/annurev.publhealth.27.021405.102103] [Medline: 16533114]
2. Link E, Beckmann S. AI at everyone's fingertips? Identifying the predictors of health information seeking intentions using AI. Commun Res Rep 2025 Jan;42(1):1-11. [doi: 10.1080/08824096.2024.2427609]
3. Varghese J, Chapiro J. ChatGPT: The transformative influence of generative AI on science and healthcare. J Hepatol 2024 Jun;80(6):977-980. [doi: 10.1016/j.jhep.2023.07.028] [Medline: 37544516]
4. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119. [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
5. Nastasi AJ, Courtright KR, Halpern SD, Weissman GE. A vignette-based evaluation of ChatGPT's ability to provide appropriate and equitable medical advice across care contexts. Sci Rep 2023 Oct 19;13(1):17885. [doi: 10.1038/s41598-023-45223-y] [Medline: 37857839]
6. Venkatesh V, Thong JYL, Xu X. Consumer Acceptance and Use of Information Technology: extending the Unified Theory of Acceptance and Use of Technology. MIS Q 2012 Mar 1;36(1):157-178. [doi: 10.2307/41410412]
7. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
8. Baumann E, Czerwinski F, Rosset M, Großmann U. Wie informieren sich die deutschen zu gesundheitsthemen? [Report in German]. : Stiftung Gesundheitswissen; 2019 URL: https://www.stiftung-gesundheitswissen.de/sites/default/files/pdf/trendmonitor_Ausgabe%201.pdf [accessed 2026-01-16]
9. Health information seeking among caregivers. : National Cancer Institute; 2019 URL: https://hints.cancer.gov/docs/Briefs/HINTS_Brief_40.pdf [accessed 2026-01-16]
10. Bachl M, Link E, Mangold F, Stier S. Search engine use for health-related purposes: behavioral data on online health information-seeking in Germany. Health Commun 2024 Jul;39(8):1651-1664. [doi: 10.1080/10410236.2024.2309810] [Medline: 38326714]
11. Al Shboul MKI, Alwreikat A, Alotaibi FA. Investigating the use of ChatGPT as a novel method for seeking health information: a qualitative approach. Sci Technol Libr 2023;43(3):225-234. [doi: 10.1080/0194262X.2023.2250835]
12. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]

XSL•FO
RenderX

13. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]
14. Chang A. UTAUT and UTAUT 2: a review and agenda for future research. Winners 2012;13(2):10. [doi: 10.21512/tw.v13i2.656]
15. Walle AD, Jemere AT, Tilahun B, et al. Intention to use wearable health devices and its predictors among diabetes mellitus patients in Amhara region referral hospitals, Ethiopia: using modified UTAUT-2 model. Inf Med Unlocked 2023;36:101157. [doi: 10.1016/j.imu.2022.101157]
16. Yu CW, Chao CM, Chang CF, Chen RJ, Chen PC, Liu YX. Exploring behavioral intention to use a mobile health education website: an extension of the UTAUT 2 model. Sage Open 2021;11(4):215824402110557. [doi: 10.1177/21582440211055721]
17. Schomakers EM, Lidynia C, Vervier LS, Calero Valdez A, Ziefle M. Applying an extended UTAUT2 model to explain user acceptance of lifestyle and therapy mobile health apps: survey study. JMIR mHealth uHealth 2022 Jan 18;10(1):e27095. [doi: 10.2196/27095] [Medline: 35040801]
18. Budhathoki T, Zirar A, Njoya ET, Timsina A. ChatGPT adoption and anxiety: a cross-country analysis utilising the unified theory of acceptance and use of technology (UTAUT). Stud High Educ 2024;49(5):831-846. [doi: 10.1080/03075079.2024.2333937]
19. Lai CY, Cheung KY, Chan CS, Law KK. Integrating the adapted UTAUT model with moral obligation, trust and perceived risk to predict ChatGPT adoption for assessment support: a survey with students. Comput Educ Artif Intell 2024 Jun;6:100246. [doi: 10.1016/j.caeai.2024.100246]
20. Foroughi B, Senali MG, Iranmanesh M, et al. Determinants of intention to use ChatGPT for educational purposes: findings from PLS-SEM and fsQCA. Int J Hum Comput Interact 2024;40(17):4501-4520. [doi: 10.1080/10447318.2023.2226495]
21. Terblanche N, Kidd M. Adoption factors and moderating effects of age and gender that influence the intention to use a non-directive reflective coaching chatbot. Sage Open 2022;12(2):21582440221096136. [doi: 10.1177/21582440221096136]
22. Neeragatti S, Dehury RK, Sripathi N. Determinants of digital health information search (DHIS) behaviour: extending UTAUT with healthcare behaviour constructs. 2022 Presented at: 2nd International Healthcare Management Conference 2022: Navigating the New Normal with Focus on Healthcare Accessibility, Innovation and Sustainability; Jan 29, 2022. [doi: 10.24083/apjhm.v18i1.1685]
23. Kim Y, Blazquez V, Oh T. Determinants of generative AI system adoption and usage behavior in Korean companies: applying the UTAUT model. Behav Sci (Basel) 2024 Nov 4;14(11):1035. [doi: 10.3390/bs14111035] [Medline: 39594336]
24. Strzelecki A. To use or not to use ChatGPT in higher education? A study of students' acceptance and use of technology. Interact Learn Environ 2024;32(9):5142-5155. [doi: 10.1080/10494820.2023.2209881]
25. Chang YT, Chao CM, Yu CW, Lin FC. Extending the utility of UTAUT2 for hospital patients' adoption of medical apps: moderating effects of e-health literacy. Mob Inf Syst 2021 Apr 12;2021:1-10. [doi: 10.1155/2021/8882317]
26. Gansser OA, Reich CS. A new acceptance model for artificial intelligence with extensions to UTAUT2: an empirical study in three segments of application. Technol Soc 2021 May;65:101535. [doi: 10.1016/j.techsoc.2021.101535]
27. Kang S, Choi Y, Kim B. Impact of motivation factors for using generative AI services on continuous use intention: mediating trust and acceptance attitude. Soc Sci 2024 Sep;13(9):475. [doi: 10.3390/socsci13090475]
28. Limayem M, Hirt SG, Cheung CMK. How habit limits the predictive power of intention: the case of information systems continuance. MIS Q 2007 Dec;31(4):705-737. [doi: 10.2307/25148817]
29. Brown SA, Venkatesh V. Model of adoption of technology in households: a baseline model test and extension incorporating household life cycle. MIS Q 2005;29(3):399-426. [doi: 10.2307/25148690]
30. Diao Y, Li Z, Zhou J, Gao W, Gong X. A meta-analysis of college students' intention to use generative artificial intelligence. arXiv. Preprint posted online on Aug 25, 2024. [doi: 10.48550/arXiv.2409.06712]
31. Nutbeam D. Health promotion glossary. Health Promot Int 1998 Jan 1;13(4):349-364. [doi: 10.1093/heapro/13.4.349]
32. Parker R, Ratzan SC. Health literacy: a second decade of distinction for Americans. J Health Commun 2010;15 Suppl 2:20-33. [doi: 10.1080/10810730.2010.501094] [Medline: 20845190]
33. van der Vaart R, Drossaert C. Development of the digital health literacy instrument: measuring a broad spectrum of Health 1.0 and Health 2.0 skills. J Med Internet Res 2017 Jan 24;19(1):e27. [doi: 10.2196/jmir.6709] [Medline: 28119275]
34. Yang K, Hu Y, Qi H. Digital health literacy: bibliometric analysis. J Med Internet Res 2022 Jul 6;24(7):e35816. [doi: 10.2196/35816] [Medline: 35793141]
35. Dunn P, Hazzard E. Technology approaches to digital health literacy. Int J Cardiol 2019 Oct 15;293:294-296. [doi: 10.1016/j.ijcard.2019.06.039] [Medline: 31350037]
36. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med 2011 Jul 19;155(2):97-107. [doi: 10.7326/0003-4819-155-2-201107190-00005] [Medline: 21768583]
37. Diviani N, van den Putte B, Giani S, van Weert JC. Low health literacy and evaluation of online health information: a systematic review of the literature. J Med Internet Res 2015 May 7;17(5):e112. [doi: 10.2196/jmir.4018] [Medline: 25953147]
38. Zhang Z, Genc Y, Xing A, Wang D, Fan X, Citardi D. Lay individuals' perceptions of artificial intelligence (AI)-empowered healthcare systems. Presented at: 83rd Annual Meeting of the Association for Information Science and Technology; Oct 25-29, 2022. [doi: 10.1002/pra2.326]

XSL•FO
RenderX

39. Yuen E, Winter N, Savira F, et al. Digital health literacy and its association with sociodemographic characteristics, health resource use, and health outcomes: rapid review. Interact J Med Res 2024 Jul 26;13:e46888. [doi: 10.2196/46888] [Medline: 39059006]
40. Yi-No Kang E, Chen DR, Chen YY. Associations between literacy and attitudes toward artificial intelligence–assisted medical consultations: the mediating role of perceived distrust and efficiency of artificial intelligence. Comput Human Behav 2023 Feb;139:107529. [doi: 10.1016/j.chb.2022.107529]
41. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. J Med Internet Res 2012 Jan 27;14(1):e19. [doi: 10.2196/jmir.1619] [Medline: 22357448]
42. Templin T, Perez MW, Sylvia S, Leek J, Sinnott-Armstrong N. Addressing 6 challenges in generative AI for digital health: a scoping review. PLoS Digit Health 2024 May;3(5):e0000503. [doi: 10.1371/journal.pdig.0000503] [Medline: 38781686]
43. Chen X, Hay JL, Waters EA, et al. Health literacy and use and trust in health information. J Health Commun 2018;23(8):724-734. [doi: 10.1080/10810730.2018.1511658] [Medline: 30160641]
44. Li J, Theng YL, Foo S. Predictors of online health information seeking behavior: changes between 2002 and 2012. Health Informatics J 2016 Dec;22(4):804-814. [doi: 10.1177/1460458215595851] [Medline: 26261218]
45. Houston TK, Allison JJ. Users of Internet health information: differences by health status. J Med Internet Res 2002;4(2):E7. [doi: 10.2196/jmir.4.2.e7] [Medline: 12554554]
46. Madrigal L, Escoffery C. Electronic health behaviors among US adults with chronic disease: cross-sectional survey. J Med Internet Res 2019 Mar 5;21(3):e11240. [doi: 10.2196/11240] [Medline: 30835242]
47. Or Z, Gandré C, Seppänen AV, et al. France: Health system review. Health Syst Transit 2023 Jul;25(3):1-276. [Medline: 37489947]
48. Bjegovic-Mikanovic V, Vasic M, Vukovic D, et al. Serbia: Health system review. Health Syst Transit 2019 Oct;21(3):1-211. [Medline: 32851979]
49. Bachner F, Bobek J, Habimana K, et al. Austria health system summary 2022. : World Health Organization; 2022 URL: https://iris.who.int/bitstream/handle/10665/365423/9789289059367-eng.pdf?sequence=1 [accessed 2026-01-16]
50. Birk HO, Vrangbæk K, Rudkjøbing A, et al. Denmark: Health system review. Health Syst Transit 2024 Feb;26(1):1-186. [Medline: 38841877]
51. Individuals—internet activities. Eurostat. 2024. URL: https://ec.europa.eu/eurostat/databrowser/view/isoc_ci_ac_i/default/table?lang=en [accessed 2026-01-16]
52. Olagnier D, Mogensen TH. The Covid-19 pandemic in Denmark: big lessons from a small country. Cytokine Growth Factor Rev 2020 Jun;53:10-12. [doi: 10.1016/j.cytogfr.2020.05.005] [Medline: 32405247]
53. Maljichi D, Limani B, Spier TE, et al. (Dis)trust in doctors and public and private healthcare institutions in the Western Balkans. Health Expect 2022 Aug;25(4):2015-2024. [doi: 10.1111/hex.13562] [Medline: 35781914]
54. Reinhardt A, Matthes J, Bojic L, Maindal HT, Paraschiv C, Ryom K. Help me, Doctor AI? A cross-national experiment on the effects of disease threat and stigma on AI health information-seeking intentions. Comput Human Behav 2025 Nov;172:108718. [doi: 10.1016/j.chb.2025.108718]
55. Weber W, Reinhardt A, Rossmann C. Lifestyle segmentation to explain the online health information-seeking behavior of older adults: representative telephone survey. J Med Internet Res 2020 Jun 12;22(6):e15099. [doi: 10.2196/15099] [Medline: 32530433]
56. Alam MZ, Hu W, Kaium MA, Hoque MR, Alam MMD. Understanding the determinants of mHealth apps adoption in Bangladesh: a SEM-Neural network approach. Technol Soc 2020 May;61:101255. [doi: 10.1016/j.techsoc.2020.101255]
57. Macedo IM. Predicting the acceptance and use of information and communication technology by older adults: an empirical examination of the revised UTAUT2. Comput Human Behav 2017 Oct;75:935-948. [doi: 10.1016/j.chb.2017.06.013]
58. Fischer SM, Dadaczynski K, Sudeck G, et al. Measuring health literacy in childhood and adolescence with the scale Health Literacy in School-Aged Children–German version: the psychometric properties of the German-language version of the WHO health survey scale HLSAC. Diagnostica 2022;68(4):184-196. [doi: 10.1026/0012-1924/a000296]
59. A-priori sample size calculator for structural equation models. Free Statistics Calculators. 2026. URL: https://www.danielsoper.com/statcalc/calculator.aspx?id=89 [accessed 2026-01-16]
60. Rogers EM. Diffusion of Innovations, 5th edition: Free Press; 2003.
61. Habibi A, Muhaimin M, Danibao BK, Wibowo YG, Wahyuni S, Octavia A. ChatGPT in higher education learning: acceptance and use. Comput Educ Artif Intell 2023;5:100190. [doi: 10.1016/j.caeai.2023.100190]
62. Nielsen J. Usability Engineering: Morgan Kaufmann; 1994.
63. Sun Y, Sheng D, Zhou Z, Wu Y. AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence–generated content. Humanit Soc Sci Commun 2024;11(1):1278. [doi: 10.1057/s41599-024-03811-x]
64. Xu K, Shi J. Visioning a two-level human–machine communication framework: initiating conversations between explainable AI and communication. Commun Theory 2024 Nov 1;34(4):216-229. [doi: 10.1093/ct/qtae016]
65. Cullati S, Bochatay N, Rossier C, Guessous I, Burton-Jeangros C, Courvoisier DS. Does the single-item self-rated health measure the same thing across different wordings? Construct validity study. Qual Life Res 2020 Sep;29(9):2593-2604. [doi: 10.1007/s11136-020-02533-2] [Medline: 32436111]

XSL•FO
RenderX

66. Open science framework (OSF). Generative AI & Health Information-Seeking. URL: https://osf.io/7b5v9/overview?view_only=c37895e21f0d4625be6eb13bc051f614 [accessed 2026-01-22]

Abbreviations

AI: artificial intelligence
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
UTAUT2: Unified Theory of Acceptance and Use of Technology 2
WHO: World Health Organization

Edited by A Mavragani; submitted 08.Apr.2025; peer-reviewed by A Rouhi, D Chrimes, S Sivarajkumar; accepted 27.Oct.2025; published 28.Jan.2026.

Please cite as:
Matthes J, Reinhardt A, Hodzic S, Kaňková J, Binder A, Bojic L, Maindal HT, Paraschiv C, Ryom K
Predicting the Intention to Use Generative Artificial Intelligence for Health Information: Comparative Survey Study
J Med Internet Res 2026;28:e75648
URL: https://www.jmir.org/2026/1/e75648
doi:10.2196/75648

© Jörg Matthes, Anne Reinhardt, Selma Hodzic, Jaroslava Kaňková, Alice Binder, Ljubisa Bojic, Helle Terkildsen Maindal, Corina Paraschiv, Knud Ryom. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Key Information Influencing Patient Decision-Making About AI in Health Care: Survey Experiment Study

Xuan Zhu[1], PhD; Austin M Stroud[2], MA; Sarah A Minteer[1,3], PhD; Dong Whi Yoo[4], PhD; Jennifer L Ridgeway[1,5], PhD; Maryam Mooghali[6], MSc, MD; Jennifer E Miller[6], PhD; Barbara A Barry[1,5], PhD

[1]Robert D and Patricia E Kern Center for the Science of Health Care Delivery, Mayo Clinic, 200 First Street SW, Rochetser, MN, USA

[2]Biomedical Ethics Research Program, Mayo Clinic, Rochester, MN, USA

[3]Department of Physical Medicine and Rehabilitation Research, Mayo Clinic, Rochester, MN, USA

[4]Human-Centered Computing Department, Luddy School of Informatics, Computing, and Engineering, Indiana University, Indianapolis, IN, USA

[5]Division of Health Care Delivery Research, Mayo Clinic, Rochester, MN, USA

[6]Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA

Corresponding Author:
Xuan Zhu, PhD
Robert D and Patricia E Kern Center for the Science of Health Care Delivery, Mayo Clinic, 200 First Street SW, Rochetser, MN, USA

Abstract

Background: Artificial intelligence (AI)–enabled devices are increasingly used in health care. However, there has been limited research on patients' informational preferences, including which elements of AI device labeling enhance patient understanding, trust, and acceptance. Clear and effective patient-facing communication is essential to address patient concerns and support informed decision-making regarding AI-enabled care.

Objective: We evaluated 3 aims using simulated AI device labels in a cardiovascular context. First, we identified key information elements that influence patient trust and acceptance of an AI device. Second, we examined how these effects varied based on patient characteristics. Third, we explored how patients evaluated informational content of AI labels and their perceived effectiveness of the AI labels in informing decision-making about the use of AI device, building trust in the device, and shaping their intention to use it in their health care.

Methods: We recruited 340 US patients from ResearchMatch.org to participate in a web-based survey that contained 2 experiments. In the discrete choice experiment, participants indicated preferences in terms of trust and acceptance regarding 16 pairs of simulated AI device labels that varied across 8 types of information needs identified in our previous qualitative work. In the single profile factorial experiment, participants evaluated 4 randomly assigned label prototypes regarding the label's legibility, comprehensibility, information overload, credibility, and perceived effectiveness in informing about the AI device, as well as participants' trust in the AI device and intention to use the device in their health care. Data were analyzed using mixed effects binary or ordinal logistic regression.

Results: The discrete choice experiment showed that information about regulatory approval, high device performance, provider oversight, and AI's value added to usual care significantly increased the likelihood of patient trust by 14.1% - 19.3% and acceptance by 13.3% - 17.9%. Subgroup analyses revealed variations based on patient characteristics such as familiarity with AI, health literacy, and recency of last medical checkup. The single profile factorial experiment showed that patients reported good label comprehension, and that information about provider oversight, regulatory approval, device performance, and AI's added value improved perceived credibility and effectiveness of the AI label (odds ratio [OR] range: 1.35 - 2.05), reduced doubts in the AI device (OR range: 0.61 - 0.77), and increased trust and intention to use the AI device (OR range: 1.47 - 1.73). However, information about data privacy and safety management protocols was less influential.

Conclusions: Patients value information about an AI device's performance, provider oversight, regulatory status, and added value during decision-making. Providing transparent, easily understandable information about these aspects is critical to support patient determinations of trust and acceptance of AI-enabled health care. Information elements' impact on patient trust and acceptance varies by patient characteristics, highlighting the need for a tailored approach to address the concerns of diverse patient groups about AI in health care.

(J Med Internet Res 2026;28:e75615) doi:10.2196/75615

KEYWORDS

artificial intelligence; health communication; health decision-making; patient preference; patient-centered care; AI labeling

XSL•FO
RenderX

Introduction

Artificial intelligence (AI) and machine learning (ML) are increasingly being integrated into health care products and services due to their potential to enhance diagnostic accuracy, improve treatment planning, increase efficiency of health care systems, reduce costs, and ultimately improve patient outcomes [1-3]. Cardiology, in particular, has seen a surge in AI/ML-enabled clinical tools, aiding clinical decision-making across the care spectrum [4-6]. While there are many promising applications of AI/ML in health care, effective approaches to inform patients about these technologies have not been established. This hinders informed patient decision-making and public trust in AI/ML-enabled medical devices.

Recent research on patient perspectives regarding AI in health care indicates that patients are generally receptive to the use of AI in their health care, but that certain conditions must be met [7,8]. Patients want clear and accessible information about how AI is used in their care, its benefits and limitations, how decisions are made, and the roles of AI and health care providers in AI-informed decisions [9-11]. Patients value human oversight over AI and want to ensure that health care decisions are ultimately made by a human [12-14]. Patients also want assurance that their data are protected and used responsibly [15-17]. Finally, they want AI to be unbiased in its training data and outputs and desire equitable access to AI applications to prevent against underrepresentation of minoritized and disadvantaged population groups in AI research [18-20]. These nuances underscore the complexity of information needed to support informed patient use of AI and effective communication between patients, clinicians, and health systems regarding the use of AI in care.

Recent efforts to understand how to communicate transparently about AI in health care have focused on clinical and technical audiences. As a result, there is growing research on best practices for transparent and standardized documentation, reporting, and communication of AI/ML models for clinicians [21-25]. However, research efforts focusing on identifying and prioritizing patients' specific information needs remain limited [26-28]. It is important to note that it is not always feasible or desirable to present all information about an AI-enabled medical device with equal prominence in patient-facing communication. In practice, patient education materials, including labels, decision aids, and informed consents, must balance completeness with cognitive load. Overwhelming patients with too much information can reduce comprehension, informed decision-making, and psychological well-being [29-32]. Similarly, patient-facing labels for AI-enabled medical devices are highly constrained by space and by the cognitive burden on users [33-35]. Thoughtful organization and prioritization of information are key to avoiding information overload and making it easier for patients to grasp the implications of AI in their care [36]. Thus, to ensure patient safety, autonomy, and informed decision-making, it is essential to understand patients' core information needs and present information in a transparent and accessible manner.

This study is part of a broader research project aimed at bridging the knowledge gap in effective communication about AI in health care by identifying the essential components of patient-facing information labels for these technologies. Information labels are structured summaries that provide users with key details about a device or technology, including its purpose, functionality, benefits, limitations, and risks [37]. The overarching goal of this project is to provide empirical evidence to inform strategies and regulatory policies that facilitate patient-centered adoption of AI in health care. For the purposes of our study, an AI device refers to a cardiology device equipped with an ML model designed to assist in decision-making across the continuum of care. Prior to this study, we conducted qualitative research, including a rapid literature review and 3 qualitative studies with patients and clinicians, to identify core information needs that contribute to patient and clinician trust in AI devices in health care, focusing on their use in diagnosing, treating, and monitoring cardiovascular conditions [34,38,39]. This work identified eight elements reported by patients as influential to their trust in the AI device, including (1) data privacy and security, (2) performance, (3) AI's added value compared with usual care, (4) regulatory approval, (5) expert endorsement, (6) generalizability and limitations, (7) device safety, and (8) health care provider (HCP) oversight. However, no evidence-based guidance currently exists on how to prioritize among them in patient-facing communication. Understanding what information patients prioritize allows AI developers, health care systems, and regulators to emphasize the most critical information elements while ensuring transparency, clarity, and comprehension. To address this knowledge gap, this study used 2 survey experiments in which participants evaluated a short, hypothetical scenario ("vignette") featuring AI label prototypes to achieve the following specific aims: (1) determine the relative importance of various information elements on patient trust and acceptance of an AI device in cardiology; (2) examine how the effects of information elements vary due to differences in patient characteristics, including familiarity with AI, medical mistrust, health literacy, and sociodemographic characteristics; and (3) assess patients' evaluation of the informational content of the AI label prototypes, perceived effectiveness of the AI label prototypes in informing decision-making about the use of the AI device, trust in the device, and intention to use the device in their health care.

Methods

Clinical Context for Experimental Vignettes

We selected a hypothetical AI-enabled smartwatch and smartphone app designed to detect potential episodes of atrial fibrillation as the AI device example for our experimental vignettes. This choice was informed by input from our clinical and regulatory collaborators as well as feedback from both clinicians and patients during our formative qualitative research. The clinical problem and AI technology in the vignette were deemed appropriate by clinician collaborators given its direct relevance to patients and accessibility to a broad population. In addition, during formative qualitative research, this AI device prompted rich discussion among patient and clinician participants, underscoring its relevance and making it a strong

XSL•FO
RenderX

candidate for use as the clinical context for the experimental vignettes.

Experimental Design

Overview

To robustly examine how AI label informational content influences patient perceptions and decision-making, we conducted 2 vignette experiments through a web-based survey. The use of vignette experiments is a widely accepted approach in health communication and medical decision science research [40-43]. This method is especially useful when studying emerging technologies for which standardized communication practices are not yet established. The vignette experiment approach offers important advantages over simply asking participants to rank or rate the importance of various information elements [41]. By presenting information elements in concrete vignettes, we create a context that more closely resembles how decisions are made in real-world health care settings. It also allowed us to control for confounding factors and precisely isolate the causal effect of each information element on psychological and behavioral outcomes such as trust and acceptance. Moreover, simply asking participants what is important to them is highly susceptible to social desirability bias; by examining how participants respond to concrete situations, the vignette experiment approach minimizes social desirability bias and offers a clearer picture of the factors that shape patient preferences and decisions. To capture both patient priorities under conditions requiring trade-offs and their evaluations of AI label informational content in a more reflective context, we implemented 2 complementary vignette experiments: a discrete choice experiment (DCE) and a single profile factorial experiment (SPFE).

DCEs are being increasingly used in medical and health services research to systematically examine people's preferences regarding health care services by assessing how much they value the specific attributes of the service [44-47]. In a typical DCE, participants are presented with 2 or more discrete hypothetical alternatives (eg, treatment A or treatment B), each consisting of multiple attributes with varying values. By analyzing the choices participants make among these alternatives, researchers can estimate the contribution of each attribute to decision-making. In this study, by observing how participants make trade-offs when choosing between competing label configurations, the DCE allowed us to (1) determine the relative importance of different information elements influencing patient trust and acceptance of an AI device and (2) examine how the effects of information elements vary due to differences in patient characteristics. This method uncovers underlying priorities and preferences that might not be evident when participants simply rate or rank information elements individually.

In contrast, the SPFE asked participants to review AI label prototypes one at a time and respond to a series of rating measures on perceived comprehension, cognitive effort, effectiveness in supporting decision-making about the AI device, trust in the device, and intention to use it in their health care. This method was chosen for 2 key reasons. First, the single profile design mirrors real-world patient decision-making regarding AI device use in health care more closely as in real life, patients typically consider a single device rather than make direct comparisons among alternatives. Second, the SPFE allows rating-based outcome measures (eg, 1=very unlikely to accept to 5=very likely to accept) that provide more direct and fine-grained insight into participants' evaluation of the label prototypes as well as the psychological and behavioral impacts of the information elements. This dual-experiment approach allows us to gain a richer and more nuanced understanding of patient preferences regarding AI device labeling and the factors that are critical for effective communication and adoption of AI technologies in health care.

For these experiments, we created AI label prototypes based on the hypothetical AI device that varied on the following eight informational elements: (1) data privacy and security, (2) performance, (3) AI's added value, (4) regulatory approval, (5) expert endorsement, (6) generalizability and limitations, (7) device safety, and (8) HCP oversight. We varied the informational content along these 8 elements based on prior qualitative research conducted with clinicians and patients, which identified these as important information needs influencing patient trust in AI [38,39]. For experimental design efficiency, each of these elements has 2 levels (ie, 2 variations): either in 2 different versions or as either present or absent. See Table 1 for additional details about these 8 elements.

XSL•FO
RenderX

Table . Label information elements and levels.

Elements	Definitions	Levels and examples
Elements with varying levels		
X1. Data privacy and security	What patient data are collected by AI[a]; how the patient data are being collected, stored, and shared.	Level 1: opt-in data sharing; level 2: opt-out data sharing
X2. Performance	True-positive: device accuracy when the patient DOES have A-fib[b]; true-negative: device accuracy when the patient DOES NOT have A-fib	Level 1: high performance; level 2: low performance
X3. AI's added value	Improvement in patient care due to the AI-enabled device; effectiveness of the AI-enabled device in comparison with non–AI-enabled tests or conventional health care.	Level 1: information absent; level 2: information present
X4. Regulatory approval	Whether the AI device received clearance, approval, or certification from a regulatory body regarding its safety and/or effectiveness.	Level 1: information absent; level 2: information present
X5. Expert endorsement	Endorsement of the device for safety and effectiveness issued by medical experts such as health care providers.	Level 1: information absent; level 2: information present
X6. Generalizability and limitations	How generalizable is the device to patients with varying demographics and characteristics; what are the conditions or contexts where the device should not be used?	Level 1: internal validation; level 2: external validation
X7. Device safety	What are the risks of malfunction, bugs, and errors from the AI device (both the AI algorithm and the supporting software)? How will these risks be managed?	Level 1: proactive auditing; level 2: reactive auditing
X8. HCP[c] oversight	Whether the results or decisions from the AI-enabled device have been verified by your health care provider.	Level 1: information absent; level 2: information present
Elements to be displayed in the same way across all labels		
1. Purpose	What is the purpose of the AI device, for what types of patients and conditions, when and in what context should the device be used?	"This AI-enabled smartwatch and smartphone app is designed to identify a potential episode of atrial fibrillation (A-fib). A-fib is an irregular and often very rapid heart rhythm that can lead to blood clots in the heart."
2. Directions	How to use the AI device, recommended actions for patients, how to interpret results (eg, what is "normal" for pt. like me); what warrants discussion with provider; next steps by AI if any (eg, AI will automatically alert HCP)	"- Wear your smartwatch - It will provide an alert when it detects a potential episode of A-fib - Talk to your doctor if you received an alert from the app"

[a]AI: artificial intelligence.

[b]A-fib: atrial fibrillation.

[c]HCP: health care provider.

Discrete Choice Experiment

We used a fractional factorial design to rate a selection of possible alternatives because a full factorial design (2^8 alternatives) would not be feasible to implement. To minimize the cognitive burden for participants and maximize statistical efficiency, each choice task consisted of 2 alternatives, and each participant was asked to consider 16 choice sets, comprising 32 unique label prototypes in total [48]. We used the R package idefix [49] to select an optimal design based on D-efficiency criterion, a commonly used metric for efficient experimental design construction to maximize statistical efficiency and precision [50-52]. Multimedia Appendix 1 summarizes the choice sets. Figure 1 shows an example choice set.

XSL•FO
RenderX

Figure 1. Example choice set for the discrete choice experiment. A-fib: atrial fibrillation; AI: artificial intelligence.

Name	**A-Fib Watch**
Purpose	This AI-enabled smartwatch and smartphone app is designed to identify a potential episode of atrial fibrillation (A-fib). A-fib is an irregular and often very rapid heart rhythm that can lead to blood clots in the heart.
Directions	• Wear your smartwatch • It will provide an alert when it detects a potential episode of A-fib • Talk to your doctor if you received an alert from the app

	Device A	Device B
Data Privacy and Security	The information you provide in the app will be used to improve the tool. Your deidentified data will be shared unless you visit the settings menu to turn off the sharing option.	The information you provide in the app will be used to improve the tool. You will be asked to opt into sharing your deidentified data when you first open the app.
Performance True-positive: device accuracy when the patient DOES have A-fib True-negative: device accuracy when the patient DOES NOT have A-fib	 90% (9 out of 10) people with A-fib will be correctly identified as having A-fib. 80% (8 out of 10) people without A-fib will be correctly identified as NOT having A-fib.	 80% (8 out of 10) people with A-fib will be correctly identified as having A-fib. 70% (7 out of 10) people without A-fib will be correctly identified as NOT having A-fib.
Added value		Appropriate use of the tool, response to alerts, and treatment may reduce risk of heart failure and stroke.
Regulatory approval	This device acquired regulatory authorization, which means that this device is as safe and effective as other regulated device(s).	
Expert endorsement		A panel of doctors has reviewed and endorsed this device for use.
Validation	The clinic where you receive care has tested and validated this device.	An independent research lab unaffiliated with the clinic where you receive care has tested and evaluated this device.
Tool safety	This device will be audited by developer team every 4 weeks.	If doctors or patients report any issues with the device, the developer team will assess and fix them promptly.
Oversight from your doctor	Your doctor will review the AI results along with your medical history.	

Single Profile Factorial Experiment

The SPFE used a 2^{8-3} resolution IV fractional factorial design to evaluate the label prototypes. We used the R package FrF2 [53] to generate the 32 experimental conditions (Multimedia Appendix 2). Each participant was randomly assigned to evaluate 4 label prototypes displayed in a randomized order. After viewing each label prototype, participants answered a series of questions on label evaluation, AI trust, and behavioral intention using 5-point Likert-style scales. Figure 2 shows an example label prototype.

XSL•FO
RenderX

Figure 2. Example label prototype for the single profile factorial experiment. AI: artificial intelligence.

Name	**A-fib Watch**
Purpose	This AI-enabled smartwatch and smartphone app is designed to identify a potential episode of atrial fibrillation (A-fib). A-fib is an irregular and often very rapid heart rhythm that can lead to blood clots in the heart.
Directions	• Wear your smartwatch • It will provide an alert when it detects a potential episode of A-fib • Talk to your doctor if you received an alert from the app
Data Privacy and Security	The information you provide in the app will be used to improve the tool. You will be asked to opt into sharing your deidentified data when you first open the app.
Performance True-positive: device accuracy when the patient DOES have A-fib True-negative: device accuracy when the patient DOES NOT have A-fib	 80% (8 out of 10) people with A-fib will be correctly identified as having A-fib. 70% (7 out of 10) people without A-fib will be correctly identified as NOT having A-fib.
Added value	Appropriate use of the tool, response to alerts, and treatment may reduce risk of heart failure and stroke.
Regulatory approval	This device acquired regulatory authorization, which means that this device is as safe and effective as other regulated device(s).
Expert endorsement	A panel of doctors has reviewed and endorsed this device for use.
Validation	The clinic where you receive care has tested and validated this device.
Tool safety	This device will be audited by the developer team every 4 weeks.
Oversight from your doctor	

Pilot Testing

We conducted cognitive interviews (via Zoom [Zoom Communications, Inc]; 30‑45 minutes in length) with 4 participants to assess whether the information elements, variations, presentation, and survey questions were easy to understand and meaningful to participants. The interviewer shared the web-based survey with participants via shared screen and asked participants to verbally answer the survey questions and share their thought processes. The interviewer also asked participants to explain the information elements in their own words. The interviews were audio-recorded, and the interviewer took notes on reflections and observations. We revised the survey questionnaire based on interview notes.

Following the cognitive interviews, we pilot tested the revised Qualtrics web-based survey with 30 participants to examine the feasibility and data quality of the 2 vignette experiments. The vignette experiments consisted of 3 sections: DCE, SPFE, and participant characteristics. The presentation order of the 2 experiments was randomized to mitigate order effect. For the DCE, participants first completed a warm-up task to familiarize them with the DCE procedure. They were then given the following instructions for the main DCE on AI information labeling:

> *Next, you will be asked to choose between 2 AI-enabled device options with similar or different characteristics.*
>
> *AI-enabled devices in health care are medical devices that use AI and specifically the subset of AI known as machine learning. Some examples of AI-enabled devices in health care include smartwatches that can monitor your heart rate for problems, smart robots that guide surgeries, or AI programs that can provide information to a physician to help with diagnosis, among many other types of technologies.*
>
> *Imagine that your health care provider recommended you using an AI-enabled smartwatch or smartphone app to help track your cardiac functions and identify potential episodes of atrial fibrillation. You were given 2 options. The 2 options were developed by the same company and provide the same functions. Please consider the 2 options below and choose the device you prefer the most.*

Participants then proceeded to complete 18 choice tasks, including 16 experimental choice tasks and 2 validity check tasks. The order of the 16 experimental choice tasks was randomized.

In the SPFE, participants were randomly assigned to evaluate 4 of 32 label prototypes, presented in a random order. They were given the following instructions at the beginning of the

XSL•FO
RenderX

section, "In this section, you will see 4 information labels about AI-enabled devices one at a time. After reading each information label, you will be asked a series of questions about your opinions on the label." For each label prototype, participants answered questions on label evaluation, trust in the AI device, and intention to use the device in health care. After completing the experiments, participants provided information on their sociodemographic and health care characteristics. Figure 3 summarizes study procedure.

Figure 3. Study procedure. DCE: discrete choice experiment: SPFE: single profile factorial experiment.

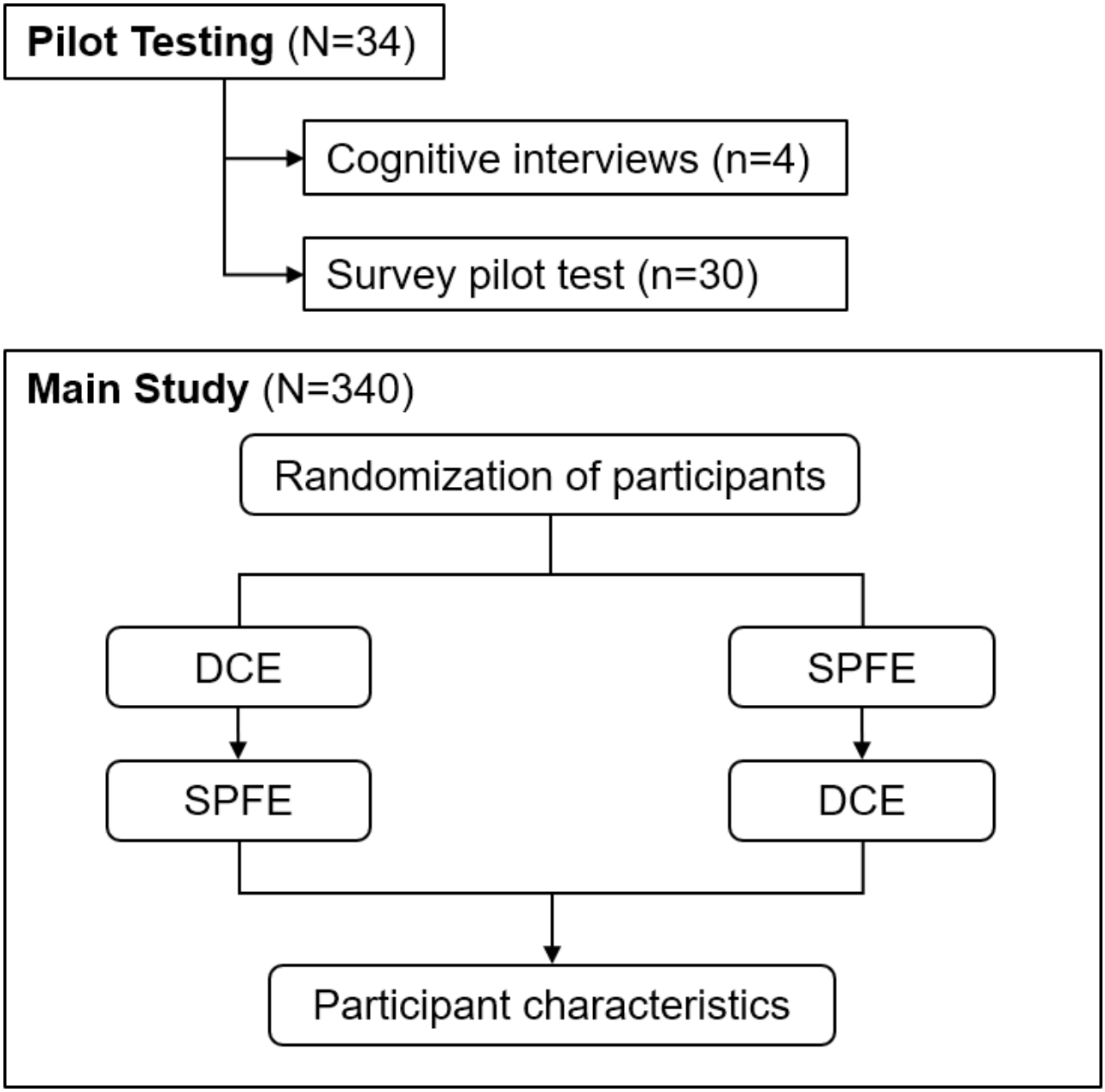

Main Study

For the main study, participants completed the finalized web-based Qualtrics survey, which incorporated refinements from the pilot phase. The survey experiments followed the same procedures described in the "Pilot Testing" section, and the randomization procedure and data collection methods and materials remained consistent with the piloted version.

Measures

Label preference was measured through participants selecting their preferred label prototype in a choice task based on trust ("Which device would you trust more?") and acceptance ("If you were given the option, which device would you be more likely to use in your health care?").

Label evaluation included six items on a 5-point Likert scale (1=strongly disagree to 5=strongly agree) that assessed (1) legibility ("This label is easy to read"), (2) comprehensibility ("This label is easy to understand"), (3) information overload ("Reading this label is too mentally demanding for me"), (4) credibility ("I trust the information on this label"), and (5) effectiveness of the label in informing patients about the AI device ("This label gives me all the information I need about

XSL•FO
RenderX

the AI device" and "This label helps me decide whether the AI device should be used in my care").

Participants' trust in the AI device was measured with three items on a 5-point Likert scale (1=strongly disagree to 5=strongly agree), including (1) "I would trust the results from this AI device," (2) "I would have doubts about this AI device," and (3) "I would seek a second opinion." Participants' likelihood of using the device was assessed on a 5-point Likert-type scale (1=very unlikely to 5=very likely). In addition, participants provided feedback on each label prototype through an optional open-ended question.

Participant characteristics were measured in terms of the need for cognition, which reflects motivation to process complex information and engage in effortful thinking [54,55], familiarity with AI, medical mistrust [56], health literacy [57-59], last routine medical checkup, health insurance coverage, and demographics (eg, age, gender, race or ethnicity, education level, and household financial status) [60].

Sample Size

We conducted an a priori power analysis to estimate the smallest sample size required for the main study. Because the DCE is more statistically efficient than the SPFE, we based our estimation on the SPFE design. The analysis showed that a sample of 288 participants would allow us to detect a main effect with a standardized regression coefficient of 0.167 or higher with 80% power at an α level of .05, which is sufficient for the DCE [61]. The total sample size needed for the pilot and main studies is 323. We planned to recruit 350 participants in total to account for attrition and incomplete responses.

Recruitment

We recruited participants from ResearchMatch [62], a national health volunteer registry that was created by several academic institutions and supported by the US National Institutes of Health as part of the Clinical Translational Science Award program. ResearchMatch has a large population of volunteers who have consented to be contacted by researchers about health studies for which they may be eligible. Inclusion criteria for this study include being an adult aged 18 years or older, being proficient in reading and writing in English, and having had a primary care or cardiology care visit within the past 3 years. We oversampled racial and ethnic minority patients to ensure that their representation was comparable with that of non-Hispanic White patients. This supported our ability to draw meaningful conclusions about diverse patient populations while accounting for historical underrepresentation in health research. Participants received a message describing our study via the ResearchMatch platform, where they could then decide to opt in to be contacted and receive a link to our survey. ResearchMatch provided contact information (ie, name and email address) for participants who opted in. Participants who opted in received an email invitation with a link to complete the online survey.

Ethical Considerations

Review and approval for this study and all procedures were obtained from the Mayo Clinic Institutional Review Board (IRB; 21-012302). The IRB granted a waiver of written documentation of informed consent. The first page of the survey displayed an IRB-approved informed consent cover letter which provided information about the study, including its purpose, the investigator, the estimated length of the survey, data storage procedures, and contact information for the study team and the Mayo Clinic IRB. Participants were instructed to review this information and were informed that by proceeding to the survey, they were providing informed consent. No personal information was collected or stored with the survey responses. Study data were secured on institutionally approved and controlled access electronic storage. Participants received a US $10 gift card as remuneration for completing the survey.

Statistical Analysis

We used mixed effects binary logistic regression to analyze the data from the DCE. Our analysis focused on how different information factors influenced the probability of individuals trusting and accepting the AI device in their health care. We included individual-specific intercepts to account for heterogeneity in individual preferences. We reported average marginal effect (AME), which shows the average change in predicted probabilities (percentage point increase or decrease) of the outcome variable across all participants when moving from one level of the information factor to another, while holding all other variables constant. In addition, we conducted subgroup analyses to explore how the effect of top preferred information factors on patient trust and acceptance of the AI device varied based on participants' AI familiarity, openness toward medical AI, health literacy, medical mistrust, and sociodemographic characteristics.

We used mixed effects ordinal logistic regression (cumulative link mixed model) to analyze the data from the SPFE. Our focus was on how different information factors influenced participants' evaluation of the AI label's legibility, comprehensibility, information overload, information credibility, perceived effectiveness in informing about the AI device, trust in the AI device, and intention to use it in health care. We included individual-specific intercepts to account for repeated measurements within individuals. For models on AI device trust and intention to use, we adjusted for participants' evaluation of the AI label's legibility, comprehensibility, information overload, information credibility, and perceived effectiveness in informing about the AI device. We used "flexible" thresholds in our models to allow the distance between the 4 cut points of the 5-point scales to vary freely. We reported odds ratios (ORs) for the effect of information factors.

All analyses were performed in R (version 4.3.1; R Core Team) [63]. "tidyverse" [64] was used for data wrangling and visualization, "lme4" [65] was used for fitting mixed effects binary logistic models, "ordinal" [66] was used for fitting mixed effects ordinal logistic models, and "marginaleffects" [67] was used for calculating AME.

Reporting Guidelines

This study is reported in accordance with the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) [68].

XSL•FO
RenderX

Results

Participant Characteristics

A total of 340 participants who completed at least 75% of the survey questions were included in the analysis. Table 2 summarizes participant characteristics. Participants were aged 18 - 34 (122/328, 37.19%), 35 - 54 (114/328, 34.75%), and 55 years or older (92/328, 28.05%). Nearly half were women (158/326, 48.47%), 48.77% (159/326) were men, and 2.76% (9/326) identifying as nonbinary or another gender. The majority were non-Hispanic White (206/327, 63.00%), followed by Black or African American (76/327, 23.24%), Hispanic or Latinx (44/327, 13.46%), Asian or Asian American (30/327, 9.17%), and Native American or Indigenous (13/327, 3.98%). Regarding education, 49.09% (161/328) had a bachelor's degree, 28.96% (95/328) held a master's degree or higher, and 20.73% (68/328) had some college or an associate degree. Financially, 53.54% (174/325) reported having disposable income after paying bills, 29.85% (97/325) had little spare money, and 16.62% (54/325) struggled to pay bills.

XSL•FO
RenderX

Table . Participant characteristics (N=340).

Characteristics	Values, n (%)
Age (years)	
18 - 24	18 (5.49)
25 - 34	104 (31.71)
35 - 44	72 (21.95)
45 - 54	42 (12.80)
55 - 64	52 (15.85)
≥65	40 (12.20)
Missing	12 (3.52)
Gender	
Man	159 (48.77)
Woman	158 (48.47)
Nonbinary/other	9 (2.76)
Missing	14 (4.12)
Race and ethnicity[a]	
White/Caucasian	206 (63.00)
Black/African American	76 (23.24)
Hispanic/Latina/Latino	44 (13.46)
Asian/Asian American	30 (9.17)
Native American/American Indian/Alaska Native/Indigenous	13 (3.98)
Middle Eastern/North African/Arab American	8 (2.45)
Native Hawaiian/Pacific Islander	2 (0.61)
Other race	6 (1.83)
Missing	13 (3.82)
Education level	
High school or less	4 (1.22)
Some college or associate's degree	68 (20.73)
Bachelor's degree	161 (49.09)
Master's degree or higher	95 (28.96)
Missing	12 (3.52)
Perceived household financial status	
After paying the bills, I still have enough money for special things that I want	174 (53.54)
Have enough money to pay the bills but little spare money to buy extra or special things	97 (29.85)
Have enough money to pay the bills but only because I have cut back on things	28 (8.62)
Having difficulty paying the bills, no matter what I do	26 (8.00)
Missing	15 (4.41)
Health insurance coverage	
Private	207 (65.51)
Public	103 (32.59)
No insurance	6 (1.90)
Missing	24 (7.06)

XSL•FO
RenderX

Characteristics	Values, n (%)
Last routine checkup	
Less than a year ago	212 (66.88)
>1 year, <2 years	53 (16.72)
>2 years, <5 years	41 (12.93)
≥5	6 (1.89)
Never had routine checkup	5 (1.58)
Missing	23 (6.76)

[a]The percentages may add over 100% because participants can select multiple races.

Most participants had private health insurance (207/316, 65.51%), while 32.59% (103/316) had public coverage. In addition, 66.88% (212/317) had a routine medical checkup within the past year.

DCE Results

Importance of the Information Factors for Patients' Trust in the AI Device

Figure 4 visualizes the relative importance of the information factors for participants' trust in the AI device. The 4 information factors that produced the greatest increase in participants' trust in AI devices were inclusion of information about regulatory approval (AME=19.34%, 95% CI 15.97%-22.72%), information about high versus low performance (AME=16.62%, 95% CI 14.40%-18.85%), inclusion of information about HCP oversight (AME=15.50%, 95% CI 12.56%-18.44%), and inclusion of information about the AI's added value compared with usual care (14.06% increase, 95% CI 12.16%-15.96%). Including information about opt-in (vs opt-out) data privacy protocol, expert endorsement (vs information absent), and information about external (vs internal) validation less strongly increased the probability of trusting the AI device (9%, 7.32%, and 2.76%, respectively). Effect of information about proactive versus reactive device safety management protocol on participants' trust was not statistically significant (Table 3).

Figure 4. Difference in probability of AI device being trusted by attribute level. AI: artificial intelligence; HCP: health care provider; Info: information; TPR: true-positive rate.

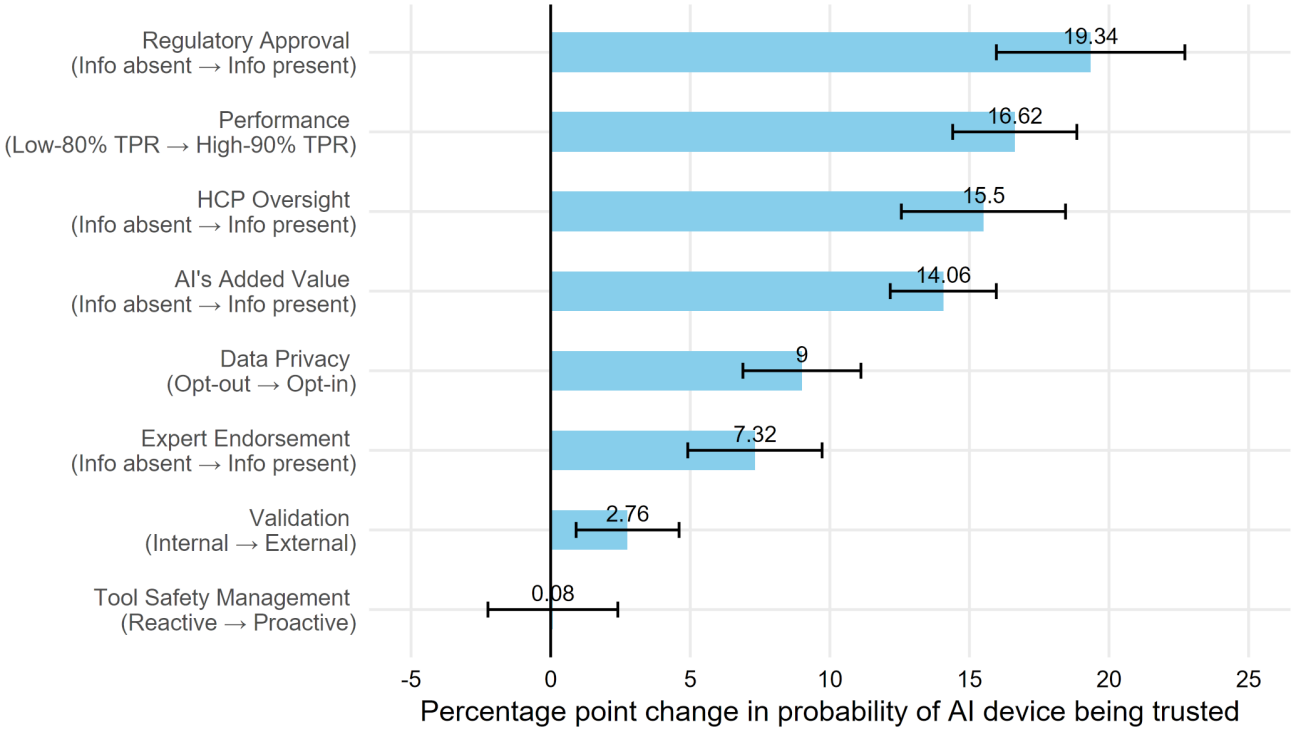

XSL•FO
RenderX

Table . Effects of information elements on the probability of the artificial intelligence device being trusted and accepted.

	Trust	Acceptance
	AME[a] (95% CI), percentage points	AME[a] (95% CI), percentage points
Added value (information absent→information present)	14.06 (12.16 to 15.96)	15.58 (13.28 to 17.88)
Data privacy (opt out→opt in)	9.00 (6.89 to 11.11)	10.92 (8.30 to 13.53)
Expert endorsement (information absent→information present)	7.32 (4.91 to 9.72)	9.03 (6.73 to 11.32)
HCP[b] oversight (information absent→information present)	15.50 (12.56 to 18.44)	17.86 (14.98 to 20.75)
Performance (low→high)	16.62 (14.40 to 18.85)	14.85 (12.65 to 17.06)
Regulatory approval (information absent→information present)	19.34 (15.97 to 22.72)	13.29 (9.96 to 16.63)
Device safety (reactive→proactive)	0.08 (–2.25 to 2.41)	0.74 (–1.82 to 3.29)
Validation (internal→external)	2.76 (0.91 to 4.60)	3.64 (1.56 to 5.72)

[a]AME: average marginal effect. It is the average change in the predicted probabilities (percentage point increase or decrease) of the artificial intelligence device being trusted or accepted across all participants when moving from one level of the information factor to the other, keeping all other variables in the model constant.

[b]HCP: health care provider.

Importance of the Information Factors for Patients' Acceptance of the AI Device

Figure 5 visualizes the relative importance of the information factors for participants' willingness to use the AI device in their health care (ie, acceptance). The 4 information factors that produced the greatest increase in participants' acceptance in using AI devices were inclusion of information about HCP oversight (17.86% increase, 95% CI 4.98%-20.75%), inclusion of information about the device's added value (15.58% increase, 95% CI 13.28%-17.88%), information about high versus low device performance (14.85% increase, 95% CI 12.65%-17.06%), and inclusion of information about regulatory approval (13.29% increase, 95% CI 9.96%-16.63%). Including information about opt-in (vs opt-out) data privacy protocol, expert endorsement (vs information absent), and information about external (vs internal) validation less strongly increased the probability of the AI device being accepted (10.92%, 9.03%, and 3.64%, respectively). The effect of information about proactive versus reactive device safety management protocol on device acceptance was not statistically significant (Table 3).

Figure 5. Difference in probability of artificial intelligence device being accepted by attribute level. AI: artificial intelligence; HCP: health care provider; Info: information; TPR: true-positive rate.

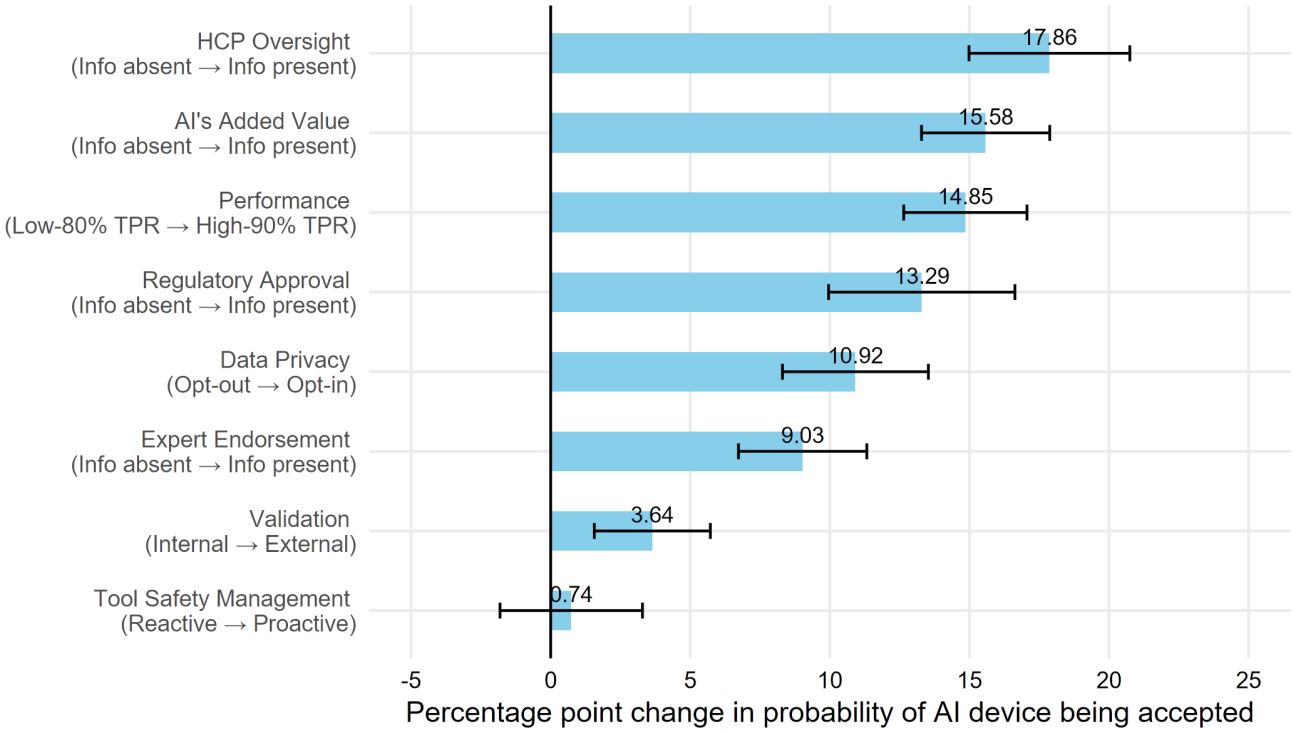

XSL•FO
RenderX

Subgroup Differences in the Effect of Top Preferred Information Factors on Patient Trust and Acceptance

Subgroup Differences in Patient Trust

Our analyses revealed subgroup differences in how the top 4 most important information factors affected trust in AI devices, based on participants' level of familiarity with AI, level of reading health literacy, level of need for cognition, recency of last routine medical checkup, age group, and gender (Multimedia Appendix 3).

AI's Added Value

Specifically, having information about AI's added value (vs information absent) had a stronger positive effect on participants' trust in AI devices with a high (vs low) level of need for cognition (AME=16.21% vs 12.23%; *P*=.04).

HCP Oversight

Having information about HCP oversight (vs information absent) had a stronger positive effect on the probability of AI device being trusted for participants who reported not at all to somewhat (vs very or extremely) familiar with AI (AME=19.27% vs 11.30%; *P*=.01), for participants whose last routine checkup was less than a year ago (vs 1 or more years ago) (AME=18.27% vs 11.63%; *P*=.007), and for participants aged 55 years or older (vs aged 35‐54 or 18‐34 years) (AME=23.44% vs 10.34% or 15.25%; *P*<.001; *P*=.03).

High Device Performance

Information about high (vs low) device performance had a stronger positive effect on the probability of AI device being trusted for participants with a high (vs low) level of reading health literacy (AME=22.02% vs 11.06%; *P*<.001), for participants whose last routine checkup was less than a year ago (vs 1 or more years ago) (AME=20.09% vs 10.38%; *P*=.048), and for participants who identified as women (vs men) (AME=19.38% vs 12.85%; *P*=.005).

Regulatory Approval

Having information about regulatory approval (vs information absent) had a stronger positive effect on the probability of AI device being trusted for participants with a high (vs low) level of reading health literacy (AME=25.07% vs 13.63%; *P*=.001) and for participants whose last routine checkup was less than a year ago (vs 1 or more years ago) (AME=23.29% vs 12.92%; *P*<.001).

Subgroup Differences in Patient Acceptance

There were subgroup differences in how the top 4 most important information factors affected acceptance of AI device, based on participants' reported level of familiarity with AI, level of reading health literacy, level of numeracy, recency of last routine medical checkup, gender, and race or ethnicity (Multimedia Appendix 4).

AI's Added Value

Specifically, having information about the device's added value (vs information absent) had a stronger positive effect on the probability of the AI being accepted for participants who reported being not at all to somewhat familiar (vs very or extremely) with AI (AME=18.06% vs 11.49%; *P*=.005), for participants with a high (vs low) level of reading health literacy (AME=19.34% vs 11.96%; *P*=.002), for participants whose last routine checkup was within the last year (vs 1 or more years ago) (AME=19.19% vs 8.81%; *P*<.001), for participants who identified as women (vs men) (AME=18.94% vs 11.38%; *P*=.002), and for participants who identified as a person of color (vs non-Hispanic White) (AME=19.64% vs 12.01%; *P*=.001).

HCP Oversight

Providing information about HCP oversight (vs information absent) had a stronger positive effect on the probability of AI device being accepted for participants who reported not at all to somewhat familiar (vs very or extremely) with AI (AME=25.34% vs 9.59%; *P*<.001), for participants whose last routine checkup was within the last year (vs 1 or more years ago) (AME=21.92% vs 13.31%; *P*<.001), and for participants aged 55 years or older (vs aged 35‐54 or 18‐34 years) (AME=30.13% vs 13.05% or 13.77%; *P*<.001; *P*<.001).

High Device Performance

Information about high (vs low) device performance had a stronger positive effect on the probability of AI device being accepted for participants who reported not at all to somewhat familiar (vs very or extremely) with AI (AME=18.18% vs 10.58%; *P*<.001), for participants with a high (vs low) level of reading health literacy (AME=21.41% vs 7.66%; *P*<.001), for participants with a high (vs low) level of numeracy (AME=19.38% vs 9.77%; *P*<.001), for participants whose last routine checkup was within the last year (vs 1 or more years ago) (AME=18.91% vs 7.03%; *P*=.004), for participants who identified as women (vs men) (AME=18.71% vs 9.57%; *P*<.001), and for participants aged 55 years or older (vs aged 35‐54 or 18‐34 years) (AME=20.88% vs 12.80% or 11.82%; *P*=.006; *P*=.002).

Regulatory Approval

Providing information about regulatory approval (vs information absent) had a stronger positive effect on the probability of AI device being accepted for participants with a high (vs low) level of reading health literacy (AME=19.93% vs 5.82%; *P*<.001), for participants with a high (vs low) level of numeracy (AME=15.79% vs 9.16%; *P*=.004), for participants whose last routine checkup was within the last year (vs 1 or more years ago) (AME=18.65% vs 3.96%; *P*<.001), and for participants who identified as women (vs men) (AME=16.58% vs 9.43%; *P*=.04).

SPFE Results

Descriptive Statistics

Table 4 summarizes the descriptive statistics of the AI label evaluation measures. Most participants reported that the AI labels were easy to read (perceived legibility; 86.7%) and understand (comprehensibility; 84.2%). They also reported that reading the labels was not too mentally demanding (no information overload; 63.9%). Effects of the information factors on perceived legibility, comprehensibility, and information overload were not statistically significant (Table 5).

XSL•FO
RenderX

Table . Descriptive statistics of the artificial intelligence label evaluation measures.

	5-point Likert scale, n (%)				
Measure	1	2	3	4	5
This label is easy to read (perceived legibility)[a]	15 (1.12)	67 (5.00)	96 (7.16)	474 (35.37)	688 (51.34)
This label is easy to understand (comprehensibility)[a]	18 (1.34)	73 (5.44)	121 (9.02)	498 (37.14)	631 (47.05)
Reading this label is too mentally demanding for me (information overload)[a]	524 (39.13)	332 (24.79)	125 (9.34)	202 (15.09)	156 (11.65)
I trust the information on this label (information credibility)[a]	30 (2.24)	94 (7.01)	220 (16.41)	597 (44.52)	400 (29.83)
Label gives all information needed about the AI[b] device (perceived label effectiveness)[a]	44 (3.28)	243 (18.13)	177 (13.21)	502 (37.46)	374 (27.91)
Label helps me decide whether the device should be used in my care (perceived label effectiveness)[a]	15 (1.12)	83 (6.19)	162 (12.09)	600 (44.78)	480 (35.82)
I would trust the results from this AI device[a]	38 (2.83)	138 (10.29)	218 (16.26)	642 (47.87)	305 (22.74)
I would have doubts about this AI device[a]	128 (9.55)	351 (26.17)	274 (20.43)	405 (30.20)	183 (13.65)
I would seek a second opinion[a]	84 (6.26)	162 (12.08)	256 (19.09)	489 (36.47)	350 (26.10)
Use the AI device if offered the option[c]	22 (1.64)	125 (9.31)	229 (17.06)	521 (38.82)	445 (33.16)

[a]1=strongly disagree; 2=somewhat disagree; 3=neither agree nor disagree; 4=somewhat agree; 5=strongly agree.

[b]AI: artificial intelligence.

[c]1=very unlikely; 2=somewhat unlikely; 3=undecided; 4=somewhat likely; 5=very likely.

XSL•FO
RenderX

Table . Effects of information element on perceived legibility, comprehensibility, information overload, information credibility, and perceived effectiveness of the label in informing decision-making.

	Perceived legibility[b]	Comprehensibility[b]	Information overload[b]	Information credibility[b]	Perceived effectiveness of the label in informing decision-making[a]	
					Label gives all information needed about the AI[c] device	Label helps me decide whether the device should be used in my care
	OR[d] (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)
Added value						
Information present	0.94 (0.70 to 1.27)	1.17 (0.88 to 1.56)	0.79 (0.58 to 1.05)	1.35 (1.05 to 1.73)	1.36 (1.06 to 1.74)	1.15 (0.89 to 1.48)
Information absent	Reference	Reference	Reference	Reference	Reference	Reference
Data privacy						
Opt-in	1.11 (0.83 to 1.50)	1.07 (0.80 to 1.42)	0.84 (0.63 to 1.13)	1.09 (0.85 to 1.39)	0.71 (0.56 to 0.91)	0.97 (0.75 to 1.24)
Opt-out	Reference	Reference	Reference	Reference	Reference	Reference
Expert endorsement						
Information present	1.21 (0.90 to 1.63)	0.90 (0.68 to 1.21)	0.87 (0.65 to 1.16)	1.37 (1.07 to 1.76)	1.09 (0.85 to 1.39)	1.27 (0.98 to 1.63)
Information absent	Reference	Reference	Reference	Reference	Reference	Reference
HCP[e] oversight						
Information present	1.11 (0.83 to 1.50)	1.13 (0.85 to 1.51)	0.88 (0.66 to 1.18)	1.60 (1.25 to 2.05)	1.48 (1.16 to 1.89)	1.41 (1.09 to 1.81)
Information absent	Reference	Reference	Reference	Reference	Reference	Reference
Performance						
High	1.21 (0.90 to 1.62)	1.15 (0.87 to 1.52)	0.84 (0.63 to 1.11)	1.45 (1.14 to 1.85)	1 (0.78 to 1.27)	1.16 (0.91 to 1.48)
Low	Reference	Reference	Reference	Reference	Reference	Reference
Regulatory approval						
Information present	1.16 (0.86 to 1.58)	1.29 (0.97 to 1.72)	0.99 (0.74 to 1.34)	1.55 (1.21 to 1.99)	2.05 (1.60 to 2.63)	1.11 (0.87 to 1.43)
Information absent	Reference	Reference	Reference	Reference	Reference	Reference
Device safety						
Proactive	0.92 (0.68 to 1.24)	0.91 (0.69 to 1.21)	1 (0.75 to 1.33)	1.03 (0.80 to 1.31)	0.90 (0.71 to 1.15)	1.09 (0.85 to 1.39)
Reactive	Reference	Reference	Reference	Reference	Reference	Reference
Validation						
External	0.98 (0.73 to 1.32)	1.13 (0.85 to 1.51)	1.01 (0.76 to 1.35)	1.01 (0.79 to 1.29)	1.14 (0.89 to 1.45)	1.03 (0.80 to 1.32)
Internal	Reference	Reference	Reference	Reference	Reference	Reference

[a]Model adjusted for perceived eligibility, comprehensibility, information overload, information credibility, and patient characteristics that were statistically significantly associated with the outcome.

XSL•FO
RenderX

[b]Model adjusted for patient characteristics that were statistically significantly associated with the outcome.

[c]AI: artificial intelligence.

[d]OR: odds ratio.

[e]HCP: health care provider.

Credibility

Having information about added value (OR 1.35, 95% CI 1.05-1.73), expert endorsement (OR 1.37, 95% CI 1.07-1.76), HCP oversight (OR 1.60, 95% CI 1.25-2.05), high (vs low) performance (OR 1.45, 95% CI 1.14-1.85), and regulatory approval (OR 1.55, 95% CI 1.21-1.99) were associated with higher levels of perceived credibility of the information (Table 5).

Perceived Effectiveness of Label

Information about added value (OR 1.36, 95% CI 1.06-1.74), HCP oversight (OR 1.48, 95% CI 1.16-1.89), and regulatory approval (OR 2.05, 95% CI 1.60-2.63) were associated with higher likelihood of participants reporting that the label gives them all the information they need about the AI device. Information about opt-in versus opt-out data privacy protocol (OR 0.71, 95% CI 0.56-0.91) was associated with lower likelihood of participants reporting that the label gives them all the information they need about the AI device. Information about HCP oversight (OR 1.41, 95% CI 1.09-1.81) was associated with higher likelihood of participants reporting that the AI label helps them decide whether the AI device should be used in their care (Table 5).

Trust and Intentions to Use AI

Information about expert endorsement (OR 1.30, 95% CI 1.01-1.68) and high (vs low) performance (OR 1.48, 95% CI 1.16-1.90) were associated with higher levels of trust in the results from the AI device. Information about high (vs low) performance (OR 0.77, 95% CI 0.61-0.98) and regulatory approval (OR 0.61, 95% CI 0.48-0.78) was associated with lower likelihood of participants reporting having doubts in the AI device. Information about HCP oversight (OR 0.75, 95% CI 0.57-0.97) and regulatory approval (OR 0.63, 95% CI 0.48-0.83) were associated with lower likelihood of participants reporting needing a second opinion. Information about HCP oversight (OR 1.47, 95% CI 1.12-1.94), high (vs low) performance (OR 1.59, 95% CI 1.22-2.07), and regulatory approval (OR 1.73, 95% CI 1.31-2.30) were associated with higher intention to use the AI device if offered the option (Table 6).

XSL•FO
RenderX

Table . Effects of information element on patient trust in the artificial intelligence device and intention to use the artificial intelligence device.

	Trust in the AI device[a], OR[b] (95% CI)			Intention to use the AI device if offered the option[a], OR (95% CI)
	I would trust results from this AI[c] device	I would have doubts about this AI device	I would seek a second opinion	
Added value				
Information present	0.83 (0.64 to 1.06)	1.15 (0.90 to 1.47)	0.87 (0.67 to 1.14)	0.99 (0.75 to 1.31)
Information absent	Reference	Reference	Reference	Reference
Data privacy				
Opt-in	1.07 (0.83 to 1.37)	1.07 (0.85 to 1.36)	1.27 (0.98 to 1.65)	1.03 (0.79 to 1.34)
Opt-out	Reference	Reference	Reference	Reference
Expert endorsement				
Information present	1.30 (1.01 to 1.68)	0.89 (0.70 to 1.13)	0.93 (0.71 to 1.21)	1.17 (0.89 to 1.54)
Information absent	Reference	Reference	Reference	Reference
HCP[d] oversight				
Information present	1.21 (0.94 to 1.55)	0.80 (0.63 to 1.03)	0.75 (0.57 to 0.97)	1.47 (1.12 to 1.94)
Information absent	Reference	Reference	Reference	Reference
Performance				
High	1.48 (1.16 to 1.90)	0.77 (0.61 to 0.98)	0.80 (0.62 to 1.04)	1.59 (1.22 to 2.07)
Low	Reference	Reference	Reference	Reference
Regulatory approval				
Information present	1.26 (0.98 to 1.64)	0.61 (0.48 to 0.78)	0.63 (0.48 to 0.83)	1.73 (1.31 to 2.30)
Information absent	Reference	Reference	Reference	Reference
Device safety				
Proactive	1.08 (0.84 to 1.38)	1.13 (0.89 to 1.43)	0.98 (0.76 to 1.27)	0.93 (0.71 to 1.21)
Reactive	Reference	Reference	Reference	Reference
Validation				
External	1.06 (0.83 to 1.36)	1.12 (0.88 to 1.42)	0.80 (0.61 to 1.03)	0.97 (0.74 to 1.27)
Internal	Reference	Reference	Reference	Reference

[a]Model adjusted for perceived eligibility, comprehensibility, information overload, information credibility, perceived effectiveness of the label in informing decision-making, and patient characteristics that were statistically significantly associated with the outcome.

[b]OR: odds ratio.

[c]AI: artificial intelligence.

[d]HCP: health care provider.

Discussion

Overview

To our knowledge, this is the first study to apply 2 experimental methods, DCE and SPFE, to elicit patient preferences for information about the use of AI devices in their health care. Our study provides important evidence and insights for health care and AI professionals and policy makers on the relative importance of various information factors to patient decision-making, trust, and acceptance regarding an AI device in cardiology and on subgroup differences in the impact of top preferred information factors. This work is a first step toward developing effective communication strategies about AI in health care that ensure transparency and accessibility to facilitate informed decision-making and patient-centered adoption of AI applications.

Study Design Considerations

Before delving into our findings, we first address the use of the vignette experiment approach, the hypothetical AI device, and the nonclinical study sample. While vignette experiments are a well-established method in health communication and medical decision science research, they are not without limitations. In particular, the artificial nature of vignettes can raise concerns about ecological validity, as hypothetical scenarios may not fully capture the complexities of real-world decision-making.

XSL•FO
RenderX

As a result, there is a risk that the study findings cannot be generalized directly to real-world settings [40-43].

However, in the context of our study, the use of the vignette experiments with a hypothetical AI device and a nonclinical study sample was both methodologically necessary and ethically appropriate. Methodologically, this approach allowed us to standardize and systematically manipulate the information elements of interest, while holding all other aspects of the scenario constant. This level of experimental control is often impossible to achieve with a real AI device due to the presence of a range of uncontrolled confounding influences on patient decision-making, including prior experiences, provider communication, and exposure to media or marketing. Importantly, the ecological validity of vignette experiments depends not on literal replication of real-world settings but on whether the vignettes activated the same psychological processes that occur during real-life decision-making [41]. To enhance the realism and credibility of our vignettes, we designed the hypothetical AI device scenario in consultation with clinical and regulatory collaborators and informed by formative qualitative research with both clinicians and patients. We also sought clinician review on the vignette drafts to ensure clinical accuracy and appropriateness and then further refined their clarity and relevance by piloting through cognitive interviews with patients. These steps strengthened both the internal and ecological validity of our study design and increased the applicability of our findings to real-world AI communication contexts.

Moreover, testing experimental manipulations involving actual patients making real-time medical decisions may raise ethical concerns, particularly when issues surrounding patient trust, understanding, and decision-making about AI in health care are still evolving [69]. Our experiment focuses on comparing different information factors presented for the same device within a consistent clinical scenario. Conducting an empirical study using an actual AI device while manipulating its label would involve providing false information, such as presenting different performance metrics for the same AI device to different patients, which would constitute deception and be unethical. This approach would undermine patient autonomy in decision-making and could erode trust in their clinicians and health care system [70]. To ensure that the hypothetical scenario was relevant to study participants and consistent with real-world decision-making, we recruited adults who reported recent health care experiences through either a primary care or cardiovascular care visit within the past 3 years. Importantly, our study goal was to understand general patient preferences and responses to AI labeling across a diverse sample, which is a necessary first step before testing specific AI implementations in particular clinical populations.

Key Findings

Results from the 2 experiments offer complementary insights into the key factors shaping patients' decision-making about use of an AI device in cardiology. Results from the SPFE showed that most participants found the AI label prototypes easy to read and understand and not mentally demanding. The effects of information factors on perceived legibility, comprehensibility, and information overload were not statistically significant, indicating that patient preferences were not due to a lack of understanding of these labels. Results from the DCE showed that information about provider oversight, regulatory approval, high device performance, and AI's added value were the most influential in increasing patient trust in the AI device and their willingness to use the device in their health care. Patients placed less importance on information about opt-in versus opt-out data privacy protocol, expert endorsement, external versus internal validation protocol, and proactive versus reactive device safety management protocol.

Results from the SPFE reinforced these findings. First, information about AI's added value, expert endorsement, HCP oversight, high performance, and regulatory approval was linked to higher perceived credibility of the AI label. In addition, information about added value, HCP oversight, and regulatory approval increased the likelihood that participants felt that they had all the necessary information about the device. Interestingly, information about HCP oversight was the only factor that significantly influenced participants' perception of the label's usefulness in deciding whether to use the device in their care. Information about expert endorsement, high performance, and regulatory approval was also associated with greater trust in the AI device's results and reduced doubts about the device. Furthermore, information about HCP oversight and regulatory approval lowered the likelihood of participants seeking a second opinion. Finally, information about HCP oversight, high performance, and regulatory approval contributed to a higher intention to use the AI device if offered.

The DCE also showed that participant characteristics, including recency of last medical checkup, familiarity with AI, health literacy, numeracy, need for cognition, age, gender, and race or ethnicity, shaped how strongly information about AI's added value, device performance, HCP oversight, and regulatory approval impacted trust and acceptance of the AI device. Information about AI's added value had greater effects among participants who had more recent medical checkups, were less familiar with AI, had higher health literacy, had higher need for cognition, identified as women, or identified as people of color. Similarly, information about high device performance had stronger effects among those with more recent medical checkups, lower familiarity with AI, higher health literacy or numeracy, older age, or identified as women. Information about HCP oversight had stronger effects among participants with more recent medical checkups, lower familiarity with AI, or older age. Finally, the effects of information about regulatory approval were stronger among those with recent checkups, higher health literacy or numeracy, or women.

Implications for Practice

Our findings have important implications for the implementation of AI in health care. To start, providing transparent and accessible information about HCP oversight, regulatory approval, device performance, and the device's added value is critical for building patient trust and acceptance of AI technology [27]. These 4 information elements were the most influential in shaping patients' willingness to use an AI device, highlighting the need for health care systems to clearly

XSL•FO
RenderX

communicate them when introducing AI devices to patients. Health care systems could consider developing decision aids that allow patients to explore different aspects of a specific AI device in collaboration with their providers, helping them weigh the benefits and risks of using the device in their care [26]. This is especially important for patients with lower baseline trust or limited familiarity with AI.

Our findings also highlight the critical importance of information about endorsement and oversight from regulatory bodies and HCPs in boosting patients' confidence and reducing doubts about the AI device's effectiveness and safety, which can lead to higher trust in the device and lower likelihood of needing second opinions. This strong preference for human oversight and approval echoes literature on algorithm aversion, which shows that people are often reluctant to trust algorithmic decision-making systems even when they outperform humans [71,72]. This aversion is especially pronounced in decisions involving high uncertainty [73] and for tasks perceived as subjective [74]; however, it can be mitigated when there is room for human oversight and modification [75]. While algorithm aversion may initially make patients hesitant to trust AI, patient trust in their provider and the health system can lead to inflated expectations of AI's positive impact on care and potentially result in overreliance when patients lack the expertise to judge when to rely on it. Transparency and patient engagement are therefore critical to calibrating trust appropriately [28,76]. Notably, information about HCP oversight was particularly influential in helping patients decide whether an AI device should be used in their care. This finding is consistent with previous research showing that patients are more receptive to AI technologies when AI supports, rather than replaces, the decisions of trusted human HCPs, underscoring the value of a "human in the loop" approach to AI implementation that provides patients a sense of accountability and assurance regarding the safety and effectiveness of these technologies [12-14,18,77,78]. To improve patient engagement and acceptance of AI devices, patient-facing communications should explicitly emphasize providers' active role in reviewing, interpreting, and overriding AI outputs.

Three information elements, opt-in versus opt-out data privacy, external versus internal validation, and proactive versus reactive device safety management protocols, were found to be relatively less influential in shaping participants' trust and decision-making regarding the AI device. This variation in the prioritization of information elements reflects how patients actively calibrate trust in AI technologies and advances beyond prior empirical work that was limited to identifying information relevant to patient trust [26,28,39,79]. Specifically, patients tend to anchor their trust in the AI device on signals that are personally meaningful from a layperson's perspective, such as information about HCP oversight, regulatory approval, expert endorsement, and device performance, rather than the more abstract procedural indicators such as data privacy, validation, and safety management. While critical from a policy and ethical standpoint [38], these issues may be too abstract or poorly understood by a layperson to effectively translate into meaningful differences in care [80]. In addition, the terms used in the descriptions of these elements, including "deidentified data," "tested and evaluated," "fix issues promptly," and "device will be audited," may have signaled to participants that basic privacy, validity, and safety protections were in place, especially in the presence of a strong credibility cue [81]—the study instruction that their HCP recommended the AI device. As a result, many participants may have relied on heuristics to reassure themselves about privacy, validity, and safety and focused their mental effort on other aspects that are more directly related to the benefits and risks of using the device in their care [82,83]. These findings echo previous research showing that increasing transparency alone is insufficient for calibrated trust, and such information must also be accessible and meaningful to users [35,80,84]. Technical or abstract aspects of the AI device should be contextualized in terms of impact on the patient's care and communicated in accessible and relatable ways, for example, through plain language summaries, visuals, and metaphors, ideally developed through an iterative process of co-design and testing with patients [85-88]. These findings also highlight a potential tension between what patients find most useful in decision-making and what is required by regulators or ethics guidelines. For example, although data privacy is a central focus in recent AI regulations [89], participants in our study placed relatively lower importance on it. This suggests that patients may not benefit from regulatory-required information in the same way as other stakeholders and may sometimes prefer information that is not routinely made available as part of regulatory clearance processes. To support appropriately calibrated trust of AI technologies in health care, patient-facing communication strategies should emphasize the information patients value most, while also clearly addressing critical ethical and regulatory considerations [90]. Importantly, to make the abstract concepts more relatable, ethical and regulatory features should be framed in terms of their tangible benefits and risks to patients. Communication efforts must also strike a careful balance: offering reassurance while avoiding overstating benefits or certainty, which could create misplaced confidence in AI technology [76,91].

Moreover, patient-facing communication and education about AI in health care should be tailored to meet the unique needs and preferences of different patient groups [92,93]. For example, information about a device's added value was found to be particularly influential for patients who were less familiar with AI, had high reading health literacy, had recent routine checkups, and identified as women and people of color. Emphasizing the role of HCPs in overseeing AI use was found to be especially important for patients who are less familiar with AI, had recent routine checkups, and were aged 55 years or older. To effectively address these diverse patient needs and preferences, communication strategies could include tailoring language to patient reading levels and incorporating visual aids, infographics, and videos to help patients better understand complex information about the AI device and make informed decisions about their care. In addition, engaging patients, patient advocacy groups, and community organizations in co-designing patient-facing AI communication and educational materials including AI device labels could help address concerns from underrepresented groups who may have different levels of comfort with and trust in AI technologies and the health care system [94]. Providing AI communication and education

XSL•FO
RenderX

materials in various formats and at different time points may improve the accessibility of information about AI and better meet the diverse needs of different patient groups. Health care systems can consider offering in-person consultations during clinical visits, along with printed and digital materials for patients to take home, ensuring that patients have the opportunity to review and understand the information at their own pace [39]. Moreover, it is the responsibility of health care systems to ensure that providers are well informed about AI devices and are capable of communicating their benefits and risks effectively to diverse patient populations. Health care systems could establish standardized guidelines and best practices for patient-facing communication and education about AI in health care. Training providers on how to effectively discuss AI devices with diverse patient populations and address their concerns would be instrumental in ensuring that patients receive clear, transparent information that aids informed decision-making and builds merited trust in AI technologies. Table 7 summarizes key findings and implications for practice.

Table . Summary of key findings, implications, and recommendations.

Findings	Implications and recommendations
High impact information • Information about HCP[a] oversight, regulatory approval, high device performance, and AI's[b] added value led to largest increases in trust and willingness to use. • These elements also boosted the AI label's credibility, increased trust in the results, reduced doubts, and enhanced patient confidence in having enough information about the device for decision-making.	• Patients prioritize personally meaningful, concrete information over abstract procedural details. • Clearly communicate information about HCP oversight, regulatory approval, device performance, and added value, as these elements are crucial for building patient trust and enabling acceptance of AI technologies. • Develop decision aids that help patients and providers explore the benefits and risks of specific AI devices together, especially important for patients with lower baseline trust in or limited familiarity with AI.
Lower impact information: Participants placed less importance on information about opt-in versus opt-out data privacy, model validation protocol, and safety management protocol.	• Abstract issues may be poorly understood, so patients rely on heuristics and focus on aspects directly related to benefits and risks. • Technical or abstract aspects should be contextualized for patient care and communicated via plain language, visuals, and metaphors. • Communication should emphasize what patients value while addressing critical ethical and regulatory considerations.
Role of human oversight: Information about HCP oversight and regulatory approval consistently boosted patient confidence, reduced doubts, increased trust, lowered the need for second opinions, and increased intention to use the AI-enabled device if offered.	• Patients prefer AI that supports, rather than replaces, human decision makers, underscoring the importance of a "human in the loop" approach. • Communicating provider oversight can inflate expectations of AI's benefits; transparency and engagement are key to calibrating trust based on human oversight. • Patient-facing communication should emphasize providers' role in reviewing, interpreting, and approving or overriding AI outputs. • Ensure that providers are well informed about AI devices and can effectively communicate benefits, risks, and limitations to patients.
Subgroup differences: Impact of information elements varied by participant characteristics. • Added value: stronger effect for recent checkups, low AI familiarity, high literacy or need for cognition, women, and people of color. • High performance: stronger for recent checkups, low AI familiarity, high literacy or numeracy, older age, and women. • HCP oversight: stronger for recent checkups, low AI familiarity, and older age. • Regulatory approval: stronger for recent checkups, high literacy or numeracy, and women.	• Tailor language to patient reading levels and use visuals, infographics, and videos to explain complex AI information. • Co-design materials with patients, advocacy groups, and community organizations to address diverse trust levels and concerns. • Provide materials in multiple formats and at different time points to meet diverse patient needs. • Establish standardized guidelines and best practices for patient-facing AI communication. • Train providers on how to effectively discuss benefits, risks, and limitations of AI devices with diverse patient populations.

[a]HCP: health care provider.

[b]AI: artificial intelligence.

Strengths, Limitations, and Future Directions

This research has several strengths. First, we used innovative experimental methods to elicit patient preferences for information about AI in health care, which could be applied to other use cases and medical specialties. Our methods can also be used to develop and evaluate other patient-facing AI communication and education materials. We applied a rigorous process for selection of information elements, including a rapid literature review and 3 qualitative studies, which strengthened the validity of our findings.

Our findings should be interpreted in the context of the limitations. The study sample, recruited from ResearchMatch.org, is a convenience sample and may not necessarily represent the US adult population. To address this, we intentionally oversampled racial and ethnic minority

XSL•FO
RenderX

populations to ensure their representation and enable comparisons between people of color and non-Hispanic White patients in AI information preferences. Since nearly all participants had health insurance, caution is needed when generalizing the findings to uninsured populations. In addition, while selection bias might be a concern, we lack data to compare participants with nonparticipants, which may impact the generalizability of our findings. Nonetheless, our methods of evaluating the importance of information elements can be applied to other AI devices, medical conditions, and other types of AI communication and education efforts.

Our study sample may not closely reflect the demographic and clinical characteristics of patients most likely to use AI-enabled cardiology devices. While all study participants reported having had a primary care or cardiology visit within the past 3 years, we could not clinically verify that they were seeking cardiovascular care during the study period. Therefore, findings may not generalize to clinical populations actively facing decisions about AI-enabled cardiac care. Future research should replicate this work among clinical populations to validate and extend the findings. Furthermore, we used a hypothetical AI-enabled cardiology device to maintain experimental control and isolate the effects of specific information elements. However, this approach may not fully capture the complexity and nuance of real-world patient decision-making. As a result, the ecological validity of our findings may be limited. To build upon and extend the current findings, future research should examine how patients actively seeking cardiovascular care respond to labeling content for real-world AI-enabled devices in clinical practice.

Our study focused on the impact of patient-facing informational content in AI labeling, and it did not directly assess how provider recommendation might influence patient trust and acceptance of AI technologies. To hold provider recommendation constant across experimental conditions, we instructed all participants to assume that their provider had recommended the use of the hypothetical AI device. We acknowledge the important role of provider recommendation in patient adoption of AI technologies and encourage future research to explore how provider recommendation and label content interact to shape patient decision-making.

We examined patient information preferences and responses to AI label prototypes at a single time point prior to actual use of an AI device. As AI technologies become more integrated into routine health care and daily life, patients may gain greater familiarity with AI and shift information priorities after repeated exposure and use. Future research should adopt longitudinal study designs to examine how patient information needs, preferences, trust, and acceptance evolve over time and across different stages of their health care journey.

The information elements presented to participants were informed by our prior qualitative research; however, they may not fully reflect the complete range of information that could appear on an actual AI device label. In addition, because our study focused solely on the informational content of AI labels, the format, layout, and visual design of the label prototypes were simplified and standardized for ease of delivery through a web-based survey and may not fully reflect the look or feel of AI device labels used in real-world clinical practices. Future research is needed to examine how variations in display format (eg, static vs interactive), information modality (eg, textual vs infographic), delivery channel (eg, digital vs printed), and timing (eg, previsit, point-of-care, postvisit) influence patient comprehension, trust, and decision-making regarding AI technologies in health care. Finally, it is critical to involve patients in co-designing and testing AI communication strategies to ensure that the materials are accessible, engaging, and aligned with the needs and preferences of diverse patient populations.

Conclusions

Through experimental studies, our research underscores the critical importance of transparent and accessible patient-facing information about AI devices and their impact on patient trust and acceptance. Information on HCP oversight, regulatory approval, device performance, and its added value emerged as pivotal factors influencing patient decision-making. Tailoring communication to meet the diverse needs and preferences of patient subgroups is essential for effective and equitable AI adoption in health care. Patient-centered communication strategies, coupled with comprehensive education for HCPs, would ensure that AI technologies are integrated into health care in a way that empowers patients to make informed choices and support overall patient care.

Funding

This work was supported by the US Food and Drug Administration (FDA) of the US Department of Health and Human Services (HHS), award number U01FD005938, totaling US $712,431 with 100% funded by FDA/HHS. The FDA/HHS had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The contents are those of the authors and do not necessarily represent the official views of, nor an endorsement by, FDA/HHS, or the US government.

Conflicts of Interest

This research was supported by the Food and Drug Administration (FDA) through the Center of Excellence in Regulatory Science and Innovation (CERSI) grant to Yale University and Mayo Clinic. XZ offers scientific input to research studies through a contracted services agreement between Mayo Clinic and Exact Sciences. MM and JEM received research grants from Arnold Ventures through Yale University. JEM serves on Alexion Pharmaceuticals Ethics Committee and is the founder of Bioethics International. BAB offers scientific input to research studies through a contracted services agreement between Mayo Clinic and Anumana, Inc. AMS, SAM, DWY, and JLR report no additional competing interests.

XSL•FO
RenderX

Multimedia Appendix 1
Choice sets.
[DOCX File, 21 KB - jmir_v28i1e75615_app1.docx]

Multimedia Appendix 2
2IV8-3 fractional factorial experimental design.
[DOCX File, 21 KB - jmir_v28i1e75615_app2.docx]

Multimedia Appendix 3
Subgroup differences in effects of information factors on the probability of the artificial intelligence device being trusted.
[DOCX File, 18 KB - jmir_v28i1e75615_app3.docx]

Multimedia Appendix 4
Subgroup differences in effects of information factors on the probability of the artificial intelligence device being accepted.
[DOCX File, 20 KB - jmir_v28i1e75615_app4.docx]

Checklist 1
CHERRIES checklist.
[DOCX File, 20 KB - jmir_v28i1e75615_app5.docx]

References

1. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 2020;3(1):118. [doi: 10.1038/s41746-020-00324-0] [Medline: 32984550]
2. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018 Oct;2(10):719-731. [doi: 10.1038/s41551-018-0305-z]
3. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. BMC Med Inform Decis Mak 2021 Dec;21(1):125. [doi: 10.1186/s12911-021-01488-9]
4. Johnson KW, Torres Soto J, Glicksberg BS, et al. Artificial Intelligence in cardiology. J Am Coll Cardiol 2018 Jun;71(23):2668-2679. [doi: 10.1016/j.jacc.2018.03.521]
5. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial intelligence in precision cardiovascular medicine. J Am Coll Cardiol 2017 May;69(21):2657-2664. [doi: 10.1016/j.jacc.2017.03.571]
6. Lopez-Jimenez F, Attia Z, Arruda-Olson AM, et al. Artificial intelligence in cardiology: present and future. Mayo Clin Proc 2020 May;95(5):1015-1039. [doi: 10.1016/j.mayocp.2020.01.038]
7. Young AT, Amara D, Bhattacharya A, Wei ML. Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. Lancet Digit Health 2021;3(9):e599-e611. [doi: 10.1016/S2589-7500(21)00132-1]
8. Moy S, Irannejad M, Manning SJ, et al. Patient perspectives on the use of artificial intelligence in health care: a scoping review. J Patient Cent Res Rev 2024;11(1):51-62. [doi: 10.17294/2330-0698.2029] [Medline: 38596349]
9. Bala S, Keniston A, Burden M. Patient perception of plain-language medical notes generated using artificial intelligence software: pilot mixed-methods study. JMIR Form Res 2020 Jun 5;4(6):e16670. [doi: 10.2196/16670] [Medline: 32442148]
10. Adams SJ, Tang R, Babyn P. Patient perspectives and priorities regarding artificial intelligence in radiology: opportunities for patient-centered radiology. J Am Coll Radiol 2020 Aug;17(8):1034-1036. [doi: 10.1016/j.jacr.2020.01.007] [Medline: 32068006]
11. Palmisciano P, Jamjoom AAB, Taylor D, Stoyanov D, Marcus HJ. Attitudes of patients and their relatives toward artificial intelligence in neurosurgery. World Neurosurg 2020 Jun;138:e627-e633. [doi: 10.1016/j.wneu.2020.03.029]
12. Nelson CA, Pérez-Chada LM, Creadore A, et al. Patient perspectives on the use of artificial intelligence for skin cancer screening. JAMA Dermatol 2020 May 1;156(5):501. [doi: 10.1001/jamadermatol.2019.5014]
13. Tran VT, Riveros C, Ravaud P. Patients' views of wearable devices and AI in healthcare: findings from the ComPaRe e-cohort. NPJ Digit Med 2019;2(53):53. [doi: 10.1038/s41746-019-0132-y] [Medline: 31304399]
14. Lennartz S, Dratsch T, Zopfs D, Persigehl T, Maintz D, Große Hokamp N, et al. Use and control of artificial intelligence in patients across the medical workflow: single-center questionnaire study of patient perspectives. J Med Internet Res 2021;23(2):e24221. [doi: 10.2196/24221]
15. McCradden MD, Sarker T, Paprica PA. Conditionally positive: a qualitative study of public perceptions about using health data for artificial intelligence research. BMJ Open 2020;10(10). [doi: 10.1136/bmjopen-2020-039798]
16. Jutzi TB, Krieghoff-Henning EI, Holland-Letz T, et al. Artificial intelligence in skin cancer diagnostics: the patients' perspective. Front Med (Lausanne) 2020;7(233):233. [doi: 10.3389/fmed.2020.00233] [Medline: 32671078]
17. Aggarwal R, Farag S, Martin G, Ashrafian H, Darzi A. Patient perceptions on data sharing and applying artificial intelligence to health care data: cross-sectional survey. J Med Internet Res 2021;23(8):e26162. [doi: 10.2196/26162]

XSL•FO
RenderX

18. Richardson JP, Smith C, Curtis S, Watson S, Zhu X, Barry B, et al. Patient apprehensions about the use of artificial intelligence in healthcare. NPJ Digit Med 2021;4(1):140. [doi: 10.1038/s41746-021-00509-1]
19. Musbahi O, Syed L, Le Feuvre P, Cobb J, Jones G. Public patient views of artificial intelligence in healthcare: a nominal group technique study. Digit Health 2021;7:20552076211063682. [doi: 10.1177/20552076211063682] [Medline: 34950499]
20. Lennox-Chhugani N, Chen Y, Pearson V, Trzcinski B, James J. Women's attitudes to the use of AI image readers: a case study from a national breast screening programme. BMJ Health Care Inform 2021 Mar;28(1). [doi: 10.1136/bmjhci-2020-100293]
21. Brereton TA, Malik MM, Lifson M, Greenwood JD, Peterson KJ, Overgaard SM. The role of artificial intelligence model documentation in translational science: scoping review. Interact J Med Res 2023 Jul 14;12:e45903. [doi: 10.2196/45903]
22. Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. Presented at: Proceedings of the Conference on Fairness, Accountability, and Transparency; Jan 29-31, 2019. [doi: 10.1145/3287560.3287596]
23. Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets. Commun ACM 2021 Dec;64(12):86-92. [doi: 10.1145/3458723]
24. Holland S, Hosny A, Newman S, Joseph J, Chmielinski K. The dataset nutrition label. Data Prot Privacy 2020;12(12):1-26. [doi: 10.5040/9781509932771.ch-001]
25. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. NPJ Digit Med 2020;3(1):41. [doi: 10.1038/s41746-020-0253-3] [Medline: 32219182]
26. Robinson R, Liday C, Lee S, et al. Artificial intelligence in health care-understanding patient information needs and designing comprehensible transparency: qualitative study. JMIR AI 2023;2:e46487. [doi: 10.2196/46487] [Medline: 38333424]
27. Zhang Z, Genc Y, Wang D, Ahsen ME, Fan X. Effect of AI explanations on human perceptions of patient-facing AI-powered healthcare systems. J Med Syst 2021 May 4;45(6):64. [doi: 10.1007/s10916-021-01743-6] [Medline: 33948743]
28. Kostick-Quenet K, Lang BH, Smith J, Hurley M, Blumenthal-Barby J. Trust criteria for artificial intelligence in health: normative and epistemic considerations. J Med Ethics 2024 Jul 23;50(8):544-551. [doi: 10.1136/jme-2023-109338] [Medline: 37979976]
29. Martin RW, Brogård Andersen S, O'Brien MA, Bravo P, Hoffmann T, Olling K, et al. Providing balanced information about options in patient decision aids: an update from the international patient decision aid standards. Med Decis Making 2021;41(7):780-800. [doi: 10.1177/0272989X211021397]
30. Swar B, Hameed T, Reychav I. Information overload, psychological ill-being, and behavioral intention to continue online healthcare information search. Comput Human Behav 2017 May;70:416-425. [doi: 10.1016/j.chb.2016.12.068]
31. Kelly B, O'Donoghue A, Parvanta S, et al. Effects of additional context information in prescription drug information sheets on comprehension and risk and efficacy perceptions. J Pharm Policy Pract 2022 Mar 1;15(1):15. [doi: 10.1186/s40545-021-00386-9] [Medline: 35232474]
32. Bester J, Cole CM, Kodish E. The limits of informed consent for an overwhelmed patient: clinicians' role in protecting patients and preventing overwhelm. AMA J Ethics 2016 Sep 1;18(9):869-886. [doi: 10.1001/journalofethics.2016.18.9.peer2-1609] [Medline: 27669132]
33. Mello MM, Char D, Xu SH. Ethical obligations to inform patients about use of AI tools. JAMA 2025 Sep 2;334(9):767. [doi: 10.1001/jama.2025.11417]
34. Yoo DW, Stroud AM, Zhu X, Miller JE, Barry B. Toward patient-centered AI fact labels: leveraging extrinsic trust cues. DIS (Des Interact Syst Conf) 2025:676-690. [doi: 10.1145/3715336.3735758]
35. Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Wortman Vaughan JW, Wallach H. Manipulating and measuring model interpretability. Presented at: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; May 6-13, 2021. [doi: 10.1145/3411764.3445315]
36. Elgin CY, Elgin C. Ethical implications of AI-driven clinical decision support systems on healthcare resource allocation: a qualitative study of healthcare professionals' perspectives. BMC Med Ethics 2024;25(1):148. [doi: 10.1186/s12910-024-01151-8]
37. Device labeling. US Food and Drug Administration. URL: https://www.fda.gov/medical-devices/overview-device-regulation/device-labeling [accessed 2024-12-20]
38. Mooghali M, Stroud AM, Yoo DW, Barry BA, Grimshaw AA, Ross JS, et al. Trustworthy and ethical AI-enabled cardiovascular care: a rapid review. BMC Med Inform Decis Mak 2024;24(1):247. [doi: 10.1186/s12911-024-02653-6]
39. Stroud A, Minteer S, Zhu X, Ridgeway J, Miller J, Barry B. Patient information needs for transparent and trustworthy cardiovascular artificial intelligence: a qualitative study. PLOS Digit Health 2025;4(4):e0000826. [doi: 10.1371/journal.pdig.0000826]
40. Sheringham J, Kuhn I, Burt J. The use of experimental vignette studies to identify drivers of variations in the delivery of health care: a scoping review. BMC Med Res Methodol 2021;21(1):81. [doi: 10.1186/s12874-021-01247-4]
41. Evans SC, Roberts MC, Keeley JW, et al. Vignette methodologies for studying clinicians' decision-making: validity, utility, and application in ICD-11 field studies. Int J Clin Health Psychol 2015;15(2):160-170. [doi: 10.1016/j.ijchp.2014.12.001] [Medline: 30487833]
42. Atzmüller C, Steiner PM. Experimental vignette studies in survey research. Methodology (Gott) 2010 Jan;6(3):128-138. [doi: 10.1027/1614-2241/a000014]

XSL•FO
RenderX

43. van Vliet LM, van der Wall E, Albada A, Spreeuwenberg PMM, Verheul W, Bensing JM. The validity of using analogue patients in practitioner–patient communication research: systematic review and meta-analysis. J Gen Intern Med 2012 Nov;27(11):1528-1543. [doi: 10.1007/s11606-012-2111-8]
44. Merlo G, van Driel M, Hall L. Systematic review and validity assessment of methods used in discrete choice experiments of primary healthcare professionals. Health Econ Rev 2020 Dec 9;10(1):39. [doi: 10.1186/s13561-020-00295-8] [Medline: 33296066]
45. Quaife M, Terris-Prestholt F, Di Tanna GL, Vickerman P. How well do discrete choice experiments predict health choices? A systematic review and meta-analysis of external validity. Eur J Health Econ 2018 Nov;19(8):1053-1066. [doi: 10.1007/s10198-018-0954-6] [Medline: 29380229]
46. Soekhai V, Bekker-Grob EW, Ellis AR, Vass CM. Discrete choice experiments in health economics: past, present and future. Pharmacoeconomics 2019;37(2):201-226. [doi: 10.1007/s40273-018-0734-2]
47. Ryan M, Gerard K. Discrete choice experiments in health economics: a review of the literature. Health Econ 2012;21(2):145-172. [doi: 10.1002/hec.1697]
48. Vanniyasingam T, Cunningham CE, Foster G, Thabane L. Simulation study to determine the impact of different design features on design efficiency in discrete choice experiments. BMJ Open 2016;6(7):e011985. [doi: 10.1136/bmjopen-2016-011985]
49. Traets F, Sanchez DG, Vandebroek M. Generating optimal designs for discrete choice experiments in R: the idefix package. J Stat Softw 2020 Nov;96(3). [doi: 10.18637/jss.v096.i03]
50. Carlsson F, Martinsson P. Design techniques for stated preference methods in health economics. Health Econ 2003;12(4):281-294. [doi: 10.1002/hec.729]
51. Burgess L, Street DJ. Optimal designs for choice experiments with asymmetric attributes. J Stat Plan Inference 2005 Sep;134(1):288-301. [doi: 10.1016/j.jspi.2004.03.021]
52. Szinay D, Cameron R, Naughton F, Whitty JA, Brown J, Jones A. Understanding uptake of digital health products: methodology tutorial for a discrete choice experiment using the Bayesian efficient design. J Med Internet Res 2021 Oct 11;23(10):e32365. [doi: 10.2196/32365] [Medline: 34633290]
53. Grömping U. R Package FrF2 for creating and analyzing fractional factorial 2-level designs. J Stat Softw 2014 Jan;56(1):1-56. [doi: 10.18637/jss.v056.i01]
54. Cacioppo JT, Petty RE. The need for cognition. J Pers Soc Psychol 1982;42(1):116-131. [doi: 10.1037/0022-3514.42.1.116]
55. Lins de Holanda Coelho G, H P Hanel P, J Wolf L. The very efficient assessment of need for cognition: developing a six-item version. Assessment 2020 Dec;27(8):1870-1885. [doi: 10.1177/1073191118793208] [Medline: 30095000]
56. Thompson HS, Valdimarsdottir HB, Winkel G, Jandorf L, Redd W. The Group-Based Medical Mistrust Scale: psychometric properties and association with breast cancer screening. Prev Med 2004 Feb;38(2):209-218. [doi: 10.1016/j.ypmed.2003.09.041]
57. Morris NS, MacLean CD, Chew LD, Littenberg B. The Single Item Literacy Screener: evaluation of a brief instrument to identify limited reading ability. BMC Fam Pract 2006 Mar 24;7(1):21. [doi: 10.1186/1471-2296-7-21] [Medline: 16563164]
58. McNaughton CD, Cavanaugh KL, Kripalani S, Rothman RL, Wallston KA. Validation of a Short, 3-Item Version of the Subjective Numeracy Scale. Med Decis Making 2015 Nov;35(8):932-936. [doi: 10.1177/0272989X15581800] [Medline: 25878195]
59. Fagerlin A, Zikmund-Fisher BJ, Ubel PA, Jankovic A, Derry HA, Smith DM. Measuring numeracy without a math test: development of the Subjective Numeracy Scale. Med Decis Making 2007;27(5):672-680. [doi: 10.1177/0272989X07304449] [Medline: 17641137]
60. Gierisch JM, Earp JA, Brewer NT, Rimer BK. Longitudinal predictors of nonadherence to maintenance of mammography. Cancer Epidemiol Biomarkers Prev 2010 Apr 1;19(4):1103-1111. [doi: 10.1158/1055-9965.EPI-09-1120]
61. de Bekker-Grob EW, Donkers B, Jonker MF, Stolk EA. Sample size requirements for discrete-choice experiments in healthcare: a practical guide. Patient 2015 Oct;8(5):373-384. [doi: 10.1007/s40271-015-0118-z]
62. Harris PA, Scott KW, Lebo L, Hassan N, Lightner C, Pulley J. ResearchMatch: a national registry to recruit volunteers for clinical research. Acad Med 2012 Jan;87(1):66-73. [doi: 10.1097/ACM.0b013e31823ab7d2] [Medline: 22104055]
63. R Core Team. R: a language and environment for statistical computing. : R Foundation for Statistical Computing; 2023 URL: https://www.R-project.org/ [accessed 2025-12-08]
64. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. JOSS 2019;4(43):1686. [doi: 10.21105/joss.01686]
65. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw 2015 Oct;67(1). [doi: 10.18637/jss.v067.i01]
66. Christensen RHB. Ordinal—regression models for ordinal data. The Comprehensive R Archive Network. URL: https://CRAN.R-project.org/package=ordinal [accessed 2025-12-08]
67. Arel-Bundock V. Marginaleffects: predictions, comparisons, slopes, marginal means, and hypothesis tests. The Comprehensive R Archive Network. URL: https://CRAN.R-project.org/package=marginaleffects [accessed 2025-12-08]
68. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004;6(3):e34. [doi: 10.2196/jmir.6.3.e34]

XSL•FO
RenderX

69. Ratti E, Morrison M, Jakab I. Ethical and social considerations of applying artificial intelligence in healthcare—a two-pronged scoping review. BMC Med Ethics 2025 May 27;26(1):68. [doi: 10.1186/s12910-025-01198-1] [Medline: 40420080]
70. Wendler D. Deceiving research participants: is it inconsistent with valid consent? J Med Philos 2022 Nov 5;47(4):558-571. [doi: 10.1093/jmp/jhac014]
71. Burton JW, Stein MK, Jensen TB. A systematic review of algorithm aversion in augmented decision making. Behav Decision Making 2020 Apr;33(2):220-239. [doi: 10.1002/bdm.2155]
72. Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. J Exp Psychol Gen 2015 Feb;144(1):114-126. [doi: 10.1037/xge0000033] [Medline: 25401381]
73. Dietvorst BJ, Bharti S. People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. Psychol Sci 2020 Oct;31(10):1302-1314. [doi: 10.1177/0956797620948841] [Medline: 32916083]
74. Castelo N, Bos MW, Lehmann DR. Task-dependent algorithm aversion. J Mark Res 2019 Oct;56(5):809-825. [doi: 10.1177/0022243719851788]
75. Dietvorst BJ, Simmons JP, Massey C. Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them. Manage Sci 2018 Mar;64(3):1155-1170. [doi: 10.1287/mnsc.2016.2643]
76. Nong P, Ji M. Expectations of healthcare AI and the role of trust: understanding patient views on how AI will impact cost, access, and patient-provider relationships. J Am Med Inform Assoc 2025 May 1;32(5):795-799. [doi: 10.1093/jamia/ocaf031]
77. Longoni C, Bonezzi A, Morewedge CK. Resistance to medical artificial intelligence. J Consum Res 2019 Dec 1;46(4):629-650. [doi: 10.1093/jcr/ucz013]
78. Cai CJ, Reif E, Hegde N, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. Presented at: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; May 2-9, 2019. [doi: 10.1145/3290605.3300234]
79. Lee JD, See KA. Trust in automation: designing for appropriate reliance. Hum Factors 2004 Jan 1;46(1):50-80. [doi: 10.1518/hfes.46.1.50_30392]
80. Wischnewski M, Krämer N, Müller E. Measuring and understanding trust calibrations for automated systems: a survey of the state-of-the-art and future directions. Presented at: CHI '23; Apr 23-28, 2023. [doi: 10.1145/3544548.3581197]
81. Chaiken S. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. J Pers Soc Psychol 1980;39(5):752-766. [doi: 10.1037/0022-3514.39.5.752]
82. Chen S, Duckworth K, Chaiken S. Motivated heuristic and systematic processing. Psychol Inq 1999 Jan;10(1):44-49. [doi: 10.1207/s15327965pli1001_6]
83. Chaiken S, Ledgerwood A. A theory of heuristic and systematic information processing. In: Handbook of Theories of Social Psychology 2012, Vol. 1:246-266. [doi: 10.4135/9781446249215.n13]
84. Felzmann H, Fosch-Villaronga E, Lutz C, Tamò-Larrieux A. Towards transparency by design for artificial intelligence. Sci Eng Ethics 2020 Dec;26(6):3333-3361. [doi: 10.1007/s11948-020-00276-4] [Medline: 33196975]
85. Stoll M, Kerwer M, Lieb K, Chasiotis A. Plain language summaries: a systematic review of theory, guidelines and empirical research. PLoS One 2022;17(6):e0268789. [doi: 10.1371/journal.pone.0268789] [Medline: 35666746]
86. Galmarini E, Marciano L, Schulz PJ. The effectiveness of visual-based interventions on health literacy in health care: a systematic review and meta-analysis. BMC Health Serv Res 2024;24(1):718. [doi: 10.1186/s12913-024-11138-1]
87. Wang T, Voss JG. Effectiveness of pictographs in improving patient education outcomes: a systematic review. Health Educ Res 2021 Mar 23;36(1):9-40. [doi: 10.1093/her/cyaa046]
88. Liu X. Use of metaphor in provider-patient communication in medical settings: a systematic review. Patient Educ Couns 2025 Aug;137:109184. [doi: 10.1016/j.pec.2025.109184]
89. King J, Meinhardt C. Rethinking privacy in the ai era policy provocations for a data-centric world. SSRN J 2024 Feb 22. [doi: 10.2139/ssrn.5446957]
90. Williamson SM, Prybutok V. Balancing privacy and progress: a review of privacy challenges, systemic oversight, and patient perceptions in ai-driven healthcare. Appl Sci (Basel) 2024;14(2):675. [doi: 10.3390/app14020675]
91. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. J Med Internet Res 2020 Jun 19;22(6):e15154. [doi: 10.2196/15154] [Medline: 32558657]
92. Young A, Tordoff J, Smith A. "What do patients want?" Tailoring medicines information to meet patients' needs. Res Social Adm Pharm 2017 Nov;13(6):1186-1190. [doi: 10.1016/j.sapharm.2016.10.006] [Medline: 27818214]
93. Smets EMA, Menichetti J, Lie HC, Gerwing J. What do we mean by "tailoring" of medical information during clinical interactions? Patient Educ Couns 2024 Feb;119:108092. [doi: 10.1016/j.pec.2023.108092]
94. Adus S, Macklin J, Pinto A. Exploring patient perspectives on how they can and should be engaged in the development of artificial intelligence (AI) applications in health care. BMC Health Serv Res 2023 Oct 26;23(1):1163. [doi: 10.1186/s12913-023-10098-2] [Medline: 37884940]

Abbreviations

AI: artificial intelligence
AME: average marginal effect

XSL•FO
RenderX

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
DCE: discrete choice experiment
HCP: health care provider
IRB: institutional review board
ML: machine learning
OR: odds ratio
SPFE: single profile factorial experiment

Edited by A Sakhuja; submitted 07.Apr.2025; peer-reviewed by E Reiter, N Hu; revised version received 22.Oct.2025; accepted 08.Nov.2025; published 12.Jan.2026.

Please cite as:
Zhu X, Stroud AM, Minteer SA, Yoo DW, Ridgeway JL, Mooghali M, Miller JE, Barry BA
Key Information Influencing Patient Decision-Making About AI in Health Care: Survey Experiment Study
J Med Internet Res 2026;28:e75615
URL: https://www.jmir.org/2026/1/e75615
doi:10.2196/75615

© Xuan Zhu, Austin M Stroud, Sarah A Minteer, Dong Whi Yoo, Jennifer L Ridgeway, Maryam Mooghali, Jennifer E Miller, Barbara A Barry. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Intervention in Health Misinformation Using Large Language Models for Automated Detection, Thematic Analysis, and Inoculation: Case Study on COVID-19

Samira Malek[1], MS; Christopher Griffin[2,3], PhD; Robert D Fraleigh[2], PhD; Robert Lennon[4], MD, JD; Vishal Monga[5], PhD; Lijiang Shen[6], PhD

[1]Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA, United States
[2]Applied Research Laboratory, Pennsylvania State University, University Park, PA, United States
[3]Department of Mathematics, Pennsylvania State University, University Park, PA, United States
[4]PrimeCare Medical, Harrisburg, PA, United States
[5]Department of Electrical Engineering, Pennsylvania State University, University Park, PA, United States
[6]Department of Communication Arts and Sciences, Pennsylvania State University, 211 Sparks Building, University Park, PA, United States

Corresponding Author:
Lijiang Shen, PhD
Department of Communication Arts and Sciences, Pennsylvania State University, 211 Sparks Building, University Park, PA, United States

Abstract

Background: The rapid growth of social media as an information channel has enabled the swift spread of inaccurate or false health information, significantly impacting public health. This widespread dissemination of misinformation has caused confusion, eroded trust in health authorities, led to noncompliance with health guidelines, and encouraged risky health behaviors. Understanding the dynamics of misinformation on social media is essential for devising effective public health communication strategies.

Objective: This study aims to present a comprehensive and automated approach that leverages large language models (LLMs) and machine learning techniques to detect misinformation on social media, uncover the underlying causes and themes, and generate refutation arguments, facilitating control of its spread and promoting public health outcomes by inoculating people against health misinformation.

Methods: We use 2 datasets to train 3 LLMs, namely, BERT, T5, and GPT-2, to classify documents into 2 categories: misinformation and nonmisinformation. In addition, we use a separate dataset to identify misinformation topics. To analyze these topics, we applied 3 topic modeling algorithms—Latent Dirichlet Allocation, Top2Vec, and BERTopic—and selected the optimal model based on performance evaluated across 3 metrics. Using a prompting approach, we extract sentence-level representations for the topics to uncover their underlying themes. Finally, we design a prompt text capable of identifying misinformation themes effectively.

Results: The trained BERT model demonstrated exceptional performance, achieving 98% accuracy in classifying misinformation and nonmisinformation, with a 44% reduction in false-positive rates for artificial intelligence–generated misinformation. Among the 3 topic modeling approaches used, BERTopic outperformed the others, achieving the highest metrics with a Coherence Value of 0.41, Normalized Pointwise Mutual Information of −0.086, and Inverse Rank-Biased Overlap of 0.99. To address the issue of unclassified documents, we developed an algorithm to assign each document to its closest topic. In addition, we proposed a novel method using prompt engineering to generate sentence-level representations for each topic, achieving a 99.6% approval rate as “appropriate” or “somewhat appropriate” by 3 independent raters. We further designed a prompt text to identify themes of misinformation topics and developed another prompt capable of detecting misinformation themes with 82% accuracy.

Conclusions: This study presents a comprehensive and automated approach to addressing health misinformation on social media using advanced machine learning and natural language processing techniques. By leveraging LLMs and prompt engineering, the system effectively detects misinformation, identifies underlying themes, and provides explanatory responses to combat its spread. The proposed method was tested on an English language COVID-19–related dataset and has not been evaluated on real-world online social media data; the experiments were conducted offline.

(J Med Internet Res 2026;28:e75500) doi:10.2196/75500

KEYWORDS

large language models; topic modeling; COVID-19; misinformation; prompt engineering; machine learning

XSL•FO
RenderX

Introduction

Misinformation and inaccurate beliefs and knowledge about health can substantially undermine well-being by fueling confusion, eroding trust in reliable medical advice, and prompting risky behaviors such as rejecting vaccines, turning to scientifically unproven home remedies, or neglecting protective measures amid clear dangers [1-8]. These inaccuracies often circulate rapidly via social media, exploiting emotional narratives that overshadow fact-based content and leading individuals to question the legitimacy of evidence-based interventions [5,8-11]. Repeated exposure to misinformation reduces health literacy and can reinforce people's belief in falsehoods, making them more likely to view credible health authorities with skepticism [12-14]. As a result, misinformation weakens the success of prevention and treatment strategies, paving the way for heightened disease transmission, avoidable complications, and deteriorating outcomes at both individual and community levels [15-18].

An illustration comes from the COVID-19 pandemic, which saw an unprecedented surge of misinformation and conspiracy theories—labeled an "infodemic" by the World Health Organization (WHO) [1,12]. False remedies, unverified claims on the origins of the virus, and politicized narratives about preventive measures severely hampered containment efforts [19-21]. While proven strategies such as mask wearing, vaccination, and physical distancing were promoted by scientific authorities, social media rumors cast doubt on vaccine safety and the reality of the virus itself, discouraging people from getting vaccinated or seeking appropriate medical care [2,22-24]. This breakdown in adherence prolonged outbreaks, overloaded health infrastructures, and ultimately jeopardized global health and economic stability [25].

A parallel can be drawn from discussions around the human papillomavirus (HPV) vaccine, which has proven crucial in preventing various HPV-related cancers, including cervical cancer that claims thousands of lives each year [26-29]. Widespread misinformation about adverse effects and conspiracies regarding its necessity led to a significant portion of unvaccinated adolescents, heightening the likelihood of HPV infection and future malignancies [30]. This trend not only increased the burden on public health systems but also underscored the power of misinformation to undermine trust in legitimate medical counsel.

In recent years, social media has become a central and highly accessible source of information for millions of users worldwide [31]. However, its ability to rapidly disseminate content—including unfounded claims—creates fertile ground for large-scale propagation of misinformation. Given the sheer volume of posts, manual monitoring and analysis of such content are impractical [31,32] Consequently, developing and using automated, data-driven methods to understand and manage the dynamics of digital misinformation are essential for preserving accurate information and safeguarding public trust.

In this study, we propose an automated system designed to identify whether a given text contains misinformation. If misinformation is detected, the system analyzes the theme of the misinformation and provides a refutation argument (inoculation) to help prevent its spread on social media and enhance public health awareness. To achieve this, we leverage a large language model (LLM) to detect misinformation effectively. Furthermore, we demonstrate that enriching datasets significantly improves the detection of misinformation generated by both humans and AI. Recent advances in LLMs, such as ChatGPT, have enabled the generation of increasingly sophisticated misinformation, which poses challenges for traditional machine learning (ML) methods in distinguishing AI-generated misinformation [33,34]. While prior research has highlighted the effectiveness of deep learning methods in classifying health-related misinformation, these efforts have predominantly focused on content generated by humans [35,36]. Moreover, our proposed process generates sentence-level descriptions of misinformation topics, eliminating the need for manual interpretation. However, prior approaches relied on ML-based methods that produced word-level topic representations, which required manual interpretation to form coherent sentence-level topics—introducing potential human errors and subjective biases [1,22]. Similar challenges arise in other ML-based applications, such as optimizing models in industries where manual calibration of parameters can lead to inefficiencies and errors. For example, recent research has demonstrated that data-driven models can enhance predictive accuracy and automate decision-making, reducing human intervention in systems that rely on complex data streams [37-40]. Inspired by these advances, our process generates sentence-level descriptions of misinformation topics, eliminating the need for manual interpretation. In addition, we introduce an algorithm to assign documents to the most relevant topics. This addresses the limitation of many ML-based topic modeling algorithms, which often leave some documents unclassified. Our process also identifies overarching themes of misinformation topics automatically, providing a high-level understanding of the underlying reasons for misinformation categorization. Although the COVID-19 pandemic serves as our illustrative case due to its scale and data availability, the underlying challenges we address—rapid online spread, emotionally charged narratives, and declining trust—are common across other health contexts (eg, HPV vaccines) and beyond. Our approach does not rely on COVID-specific lexicons or handcrafted rules. Instead, the proposed Misinformation Detection and Inoculation Process (MDIP) is domain-agnostic. It ingests free English text, induces topics using standard models, transforms word lists into sentence-level descriptors through targeted LLM prompting, organizes them into hierarchical themes (guided by coherence, diversity metrics, and generic embeddings), and maps themes to refutation templates. Each stage has the potential to be applied to other health domains and misinformation settings, provided that the AI is properly trained with the topic- or domain-specific data.

Many previous studies have focused on individual aspects of the misinformation problem, such as detection or topic analysis [1,35,41], but have not integrated these steps into a unified framework for intervention. Our MDIP and Misinformation Detection and Inoculation System (MDIS) frameworks unify misinformation detection, topic modeling, thematic refutation, and public health intervention into a single, automated

XSL•FO
RenderX

workflow. This end-to-end approach enables health teams to move beyond merely identifying misinformation to actively and effectively countering it.

In the study by He et al [42], a method was proposed to generate per-claim counterresponses. While generating responses tailored to each specific piece of misinformation can be more informative and persuasive, such approaches require paired datasets of misinformation posts and response arguments for model training—datasets that are difficult to construct. Moreover, misinformation often mutates through paraphrasing and subtle edits; claim-specific pipelines are fragile in the face of such variation [43,44]. By contrast, our theme-level refutation approach is robust to these surface changes. More recently, LLM-based topic modeling approaches, such as TopicGPT [45] and other methods [46,47], have leveraged the capabilities of powerful pretrained models such as ChatGPT and LLaMA. While these methods benefit from the models' deep understanding of language, they often require passing entire documents through parameter-rich models, which leads to increased latency and computational costs compared with traditional pipelines. Furthermore, they typically yield only single-level sets of word topics. In contrast, our hybrid framework balances efficiency and expressiveness: we first use a traditional topic inducer to efficiently uncover the underlying structure and then apply LLM prompting (via ChatGPT) where it provides the added value. Specifically, the LLM is used to transform word lists into sentence-level topic labels and to organize topics into hierarchical themes. This targeted use of LLMs preserves computational efficiency while producing richer, hierarchical, and deployment-ready representations that are well suited for downstream tasks such as detection, monitoring, and refutation.

Methods

Study Design

In this study, we propose the MDIP, a comprehensive framework designed to analyze the dynamics of misinformation automatically and develop an MDIS. The MDIS end-to-end pipeline (1) flags misinformation, (2) explains what it is about via topics and higher-level themes, and (3) returns a concise, theme-matched refutation. The components are modular and feed one another in a simple data flow. No step in MDIP and MDIS uses disease-specific features; inputs are raw text and model hyperparameters chosen by intrinsic criteria (topic coherence and diversity). This makes the pipeline directly applicable to other misinformation corpora after swapping in the relevant documents. The MDIP framework is structured into four interconnected sections, each addressing a critical aspect of misinformation management:

1. *Detect misinformation*: Collect a labeled dataset and then train LLMs to classify text documents as either misinformation or nonmisinformation, providing a foundation for identifying false narratives.
2. *Misinformation topics*: Here, a topic modeling algorithm is applied to uncover the key topics within misinformation datasets. This step helps categorize misinformation into specific subject areas, enabling a better understanding of its thematic structure.
3. *Topic descriptions*: This section uses prompt engineering and the results of the previous section to enhance interpretability by generating sentence-level representations for each topic, moving beyond traditional word-level outputs. These descriptive summaries provide meaningful context for understanding the essence of each topic.
4. *Provide refutation*: In the final step, topic descriptions and extracted themes are used to design a specific prompt that identifies the underlying themes of misinformation. The system then generates clear and contextually relevant refutation arguments tailored to the detected misinformation themes. These arguments are designed to counter false narratives, improve public understanding, and mitigate the spread of misinformation. The refutations are a key component in the psychological inoculation-based misinformation mitigation intervention. As a metaphor to medical vaccination, the typical inoculation strategy consists of an attack message (ie, as a small weakened or deactivated dose of the virus) and refutation or counterarguments against the attack message (ie, as the immune system's reaction to the vaccine when it is injected or otherwise enters the human body) [48].

By integrating these components, MDIP enables the development of MDIS, an intelligent and automated system capable of detecting misinformation, identifying its themes, and delivering refutations to combat its impact on public health. The overall architecture of the proposed framework is illustrated in Figure 1. Figure 1A outlines the 4 stages of the MDIP: supervised detection of misinformation, topic modeling, generation of interpretable topic descriptions, and theme-based refutation. Figure 1B presents the end-to-end workflow of the MDIS, which processes new text inputs to produce a misinformation classification, assign the text to a thematic category, and generate a matched refutation. Finally, Figure 1C provides an example of the user-facing output, demonstrating how the system delivers both a misinformation warning and a concise, contextually relevant refutation.

XSL•FO
RenderX

Figure 1. (**A**) Overview of the Misinformation Detection and Inoculation Process (MDIP) integrates four main stages: (1) misinformation detection through supervised classification, (2) topic modeling of misinformation texts and assignment of outliers, (3) generation of interpretable topic descriptions and aggregation into higher-level themes, and (4) theme detection and provision of theme-linked refutations via prompt engineering for detected misinformation. (**B**) Misinformation Detection and Inoculation System (MDIS) workflow: For any new text, the system outputs (1) the misinformation decision, (2) its most likely theme, and (3) a theme-matched refutation. (**C**) Example: Illustration of how the system delivers the final user-facing output, providing a warning about misinformation and the matched refutation text. LLM: large language model.

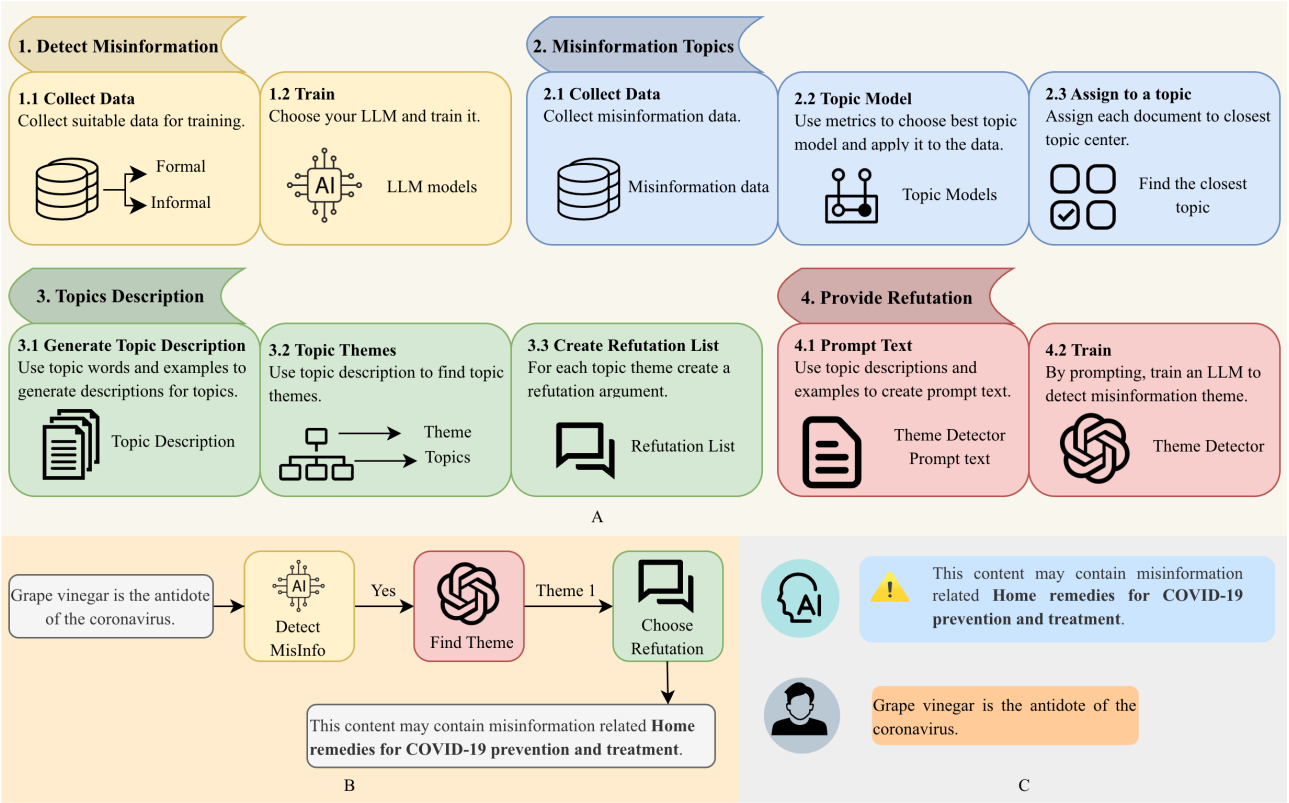

Detect Misinformation

Misinformation detection in text documents has become a critical area of research due to the growing prevalence of misleading or false information online. To address this challenge, we use classifiers based on LLMs. These LLMs are trained to categorize text into 2 classes: misinformation and nonmisinformation, as shown in the first part of Figure 1A.

Our classifier was trained using 2 complementary datasets, each providing diverse linguistic characteristics to enhance performance in detecting misinformation. The first dataset, the AAAI 2021 Competition Dataset [49], consists of misinformation sourced from social media platforms such as Facebook and X (formerly known as Twitter). This dataset reflects the informal, conversational style of social media, characterized by casual tone, nonstandard grammar, and the use of slang. The second dataset, COVID_19FNIR [50], includes misinformation presented in formal, structured language, offering a stark contrast to the informal nature of the first dataset. By incorporating these 2 datasets, we trained 3 different LLMs to detect misinformation effectively across a wide spectrum of communication styles. The blend of informal and formal language enabled the model to better generalize, achieving improved accuracy and robustness in identifying misinformation, whether generated by humans or artificial intelligence (AI).

Traditionally, researchers collect human-written data from social media platforms such as Twitter and Facebook, label them as misinformation or nonmisinformation, and then train a deep neural network to classify such documents [35,36,51,52]. However, recent studies have demonstrated that deep neural networks trained exclusively on human-written datasets exhibit weaker accuracy in detecting AI-generated misinformation compared with human-written misinformation. This discrepancy arises because AI-generated misinformation often adopts formal language styles similar to accurate information shared by credible sources such as the WHO and the Centers for Disease Control and Prevention (CDC) on official social media accounts [33,34].

In this research, by combining a dataset with formal language and another with informal language (enriching the dataset with different language types and more misinformation), we demonstrate that LLMs achieve reasonable accuracy in detecting AI-generated misinformation. This approach ensures better generalization and robustness, bridging the gap in identifying misinformation across diverse linguistic styles.

Misinformation Topics

As outlined in the second section of Figure 1A, our approach involves 3 key steps. First, we collect misinformation data. Next, we select and compare topic modeling algorithms based on specific features and metrics to identify the most effective model. Finally, we design an algorithm that assigns topics to new or unclassified documents.

XSL•FO
RenderX

To identify misinformation topics, we used one of the largest datasets of verified COVID-19 claims, the IFCN dataset, which has been extensively used in related research [41,53,54]. We applied 3 topic modeling algorithms—Latent Dirichlet Allocation (LDA) [55], Top2Vec [56,57], and BERTopic [57,58]—to analyze this dataset.

To evaluate and compare the performance of these algorithms, we selected 3 metrics: Coherence Value (CV) [59], Normalized Pointwise Mutual Information (NPMI) [59], and Inverse Rank-Biased Overlap (IRBO) [60]. CV and NPMI measure the coherence of the topics, ensuring that they are logically consistent, being human interpretable, and meaningful. IRBO, on the other hand, evaluates the diversity of the topics generated by the model, which is crucial for ensuring broad coverage of the dataset's content. Since our focus is on misinformation within health-related social media data, coherence and diversity are particularly important to ensure that topics are both interpretable and representative.

After selecting the best-performing topic model, we developed an algorithm to address the issue of unclassified documents. This algorithm assigns topics to new or previously unassigned documents, ensuring comprehensive topic coverage and improved usability of the model for real-world applications.

Topic Description

Topic modeling algorithms typically produce word-level representations for each topic. While these representations provide insight into the most relevant words associated with a topic, they often lack the semantic depth necessary to precisely identify the specific topic within a document. This limitation arises because word-level outputs fail to capture the context and relationships between words that define the overarching theme of a topic [51].

Recent advancements in LLMs have demonstrated their ability to generate high-quality, contextually relevant outputs with minimal or zero additional training by designing carefully crafted inputs—referred to as prompt engineering [61]. Leveraging this capability, we address the limitations of word-level representations by using prompt engineering techniques to generate sentence-level representations for each topic. These sentence-level representations capture the context and essence of the topic, enabling a more accurate and interpretable understanding of the document content.

Subsequently, these sentence-level representations are used to identify and articulate the overarching themes of the topics, also at the sentence level. This approach provides a more comprehensive view of the thematic structure within the document corpus. Finally, recognizing that all documents within the dataset share a common underlying reason for being classified as misinformation, we develop a tailored response list for each topic theme. The third section of Figure 1A illustrates these 3 steps.

Provide Refutation

In the final step of our proposed method, as illustrated in the final part of Figure 1A, we identify the overarching theme of misinformation and provide a corresponding response from a preconstructed response list. This response list is developed in the preceding step based on the identified themes.

To determine the themes of misinformation, we use prompt engineering techniques. By designing carefully crafted and contextually appropriate prompt text, we effectively extract the underlying themes associated with misinformation. This approach allows us to translate complex word-level or sentence-level representations into meaningful thematic insights.

By identifying misinformation themes and providing precise, theme-based responses, our method aims to enhance public health knowledge and reduce the spread of misinformation. This proactive approach not only mitigates the risks associated with false or misleading information but also fosters a more informed and resilient society.

Proposed System

Following the completion of four foundational steps—(1) detecting misinformation, (2) identifying misinformation topics, (3) describing topics, and (4) providing refutations—we develop our comprehensive MDIS, which consists of three key components.

1. *Detection of misinformation*: The system begins by determining whether a given document is misinformation.
2. *Identification of misinformation themes*: If the document is classified as misinformation, the system analyzes its content to identify the underlying misinformation themes. This process involves extracting thematic representations that provide a clearer understanding of the document's misleading aspects.
3. *Providing refutations*: Finally, the system generates a detailed refutation argument for the identified misinformation themes. These arguments are derived from a predesigned response list tailored to address specific misinformation themes effectively.

All 3 components of the system are demonstrated with a practical example, as illustrated in Figure 1B and C. This example highlights how the system operates cohesively to detect misinformation, uncover its thematic structure, and deliver accurate refutations, ultimately contributing to a more informed and resilient public.

Ethical Considerations

This study did not involve human participants, human tissue, or the collection of identifiable private information by the authors. All analyses were conducted on previously collected, publicly available, and deidentified datasets, obtained solely for research purposes. Specifically, the data sources include: (1) the AAAI 2021 COVID-19 Fake News Detection Competition dataset, originally released as part of the AAAI Conference on Artificial Intelligence shared task, in which all social media content was anonymized and distributed for noncommercial research use only [49]; (2) the COVID-19 FNIR (Fake News and Information Reliability) dataset, introduced by prior studies for misinformation detection research and released in deidentified form for academic use [50]; and (3) the International Fact-Checking Network (IFCN) COVID-19 fact-checking corpus, which aggregates publicly available

XSL•FO
RenderX

fact-check articles produced by IFCN-certified organizations and contains no personal or sensitive individual-level data [62]. According to the US Department of Health and Human Services Common Rule (45 CFR §46.104(d)), secondary research involving publicly available, deidentified data does not constitute human subjects research and is therefore exempt from Institutional Review Board review [63]. The research complied with all relevant ethical standards and data use policies and poses no risk to individuals or communities. The study's sole objective is to advance computational methods for understanding and mitigating the spread of health misinformation.

Results

Text Classification

Due to the exceptional performance of LLMs across a wide range of AI tasks, we leveraged 3 prominent LLMs to fine-tune them for COVID-19 text classification. These models—BERT (Bidirectional Encoder Representations from Transformers), GPT-2 (Generative Pre-trained Transformer 2), and T5-base (Text-to-Text Transfer Transformer)—are renowned for their ability to understand and process natural language with high accuracy and contextual awareness.

BERT is particularly effective in handling text classification tasks due to its bidirectional context understanding, which allows it to capture nuanced language patterns [64]. GPT-2 excels in text generation and classification by leveraging its autoregressive architecture to predict sequences in a given context [65]. Finally, T5-base under a unified framework that reformulates all NLP tasks as a text-to-text problem, making it versatile and effective across various domains [66].

To conduct this study, we combined the AAAI 2021 competition dataset with the COVID-19 FNIR dataset. The data were split into training, testing, and validation sets with proportions of 67%, 17%, and 16%, respectively.

Accuracy, F_1-score, Recall, and Precision are standard metrics for evaluating classification models. Accuracy measures the proportion of all predictions that are correct, providing an overall performance indicator but sometimes masking class imbalances. Precision quantifies the fraction of predicted positives that are truly positive, reflecting how often the model avoids false alarms. Recall (or sensitivity) measures the fraction of actual positives that the model successfully identifies, highlighting its ability to capture relevant cases [67]. F_1-score is the harmonic mean of precision and recall, balancing the trade-off between the two. In health-related text classification tasks such as misinformation detection, reasonable thresholds are often set required due to the risks of misclassification—for instance, aiming for Accuracy >0.80, F_1-score ≥0.75, Recall ≥0.75, and Precision ≥0.70—to ensure both reliable detection and practical usability in downstream inoculation public health applications. Table 1 shows the evaluation metrics, including Accuracy, F_1-score, Recall, and Precision, for all 3 models on the test dataset. Among these, BERT achieved the highest performance, with an accuracy of *98%* on the test data. This result highlights BERT's ability to handle complex linguistic structures and its effectiveness in fine-tuning for domain-specific tasks such as COVID-19 text classification.

The confusion matrices in Figure 2 for BERT, GPT-2, and T5-base further illustrate the performance of these models, providing a detailed breakdown of true positives, true negatives, false positives, and false negatives, which helps in understanding their classification strengths and potential areas for improvement.

Table . Performance metrics (Accuracy, F_1-score, Recall, and Precision) on the test dataset for 3 models: BERT-base, GPT-2, and T5-base.

Model	Accuracy	F_1-score	Recall	Precision
BERT	0.9848	0.9854	0.9896	0.9812
GPT-2[a]	0.9460	0.9495	0.9841	0.9117
T5-base (Generic Condition)	0.9763	0.9763	0.9763	0.9764

[a]GPT-2: Generative Pre-trained Transformer.

Figure 2. Confusion matrices illustrate the performance of the 3 binary classification models (BERT, GPT-2, and T5-base). GPT-2: Generative Pre-trained Transformer.

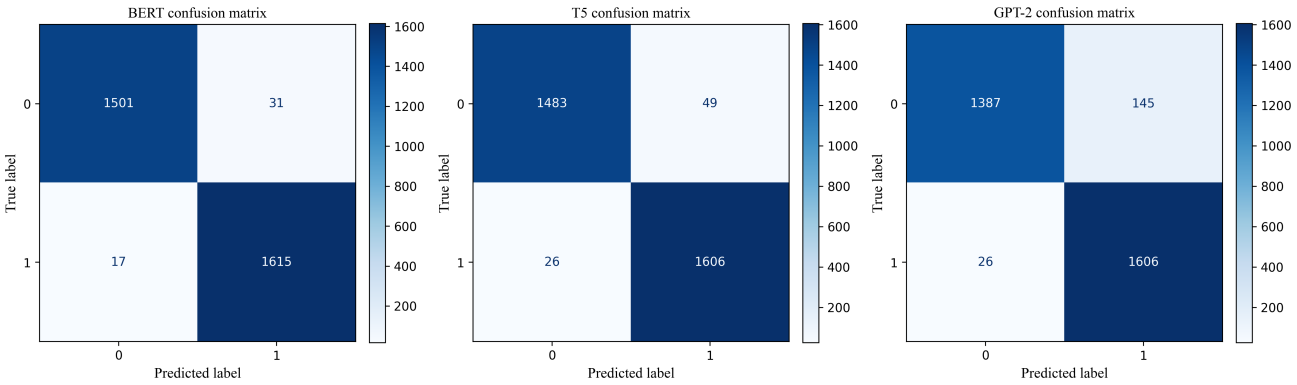

XSL•FO
RenderX

To evaluate the accuracy of our model on AI-generated data, we used the dataset provided in the study by Du et al [35]. We tested our fine-tuned BERT model on this dataset, and the results are shown in Table 2. The findings indicate a significant reduction in the number of false positives, decreasing from 27 to 15, representing a 44% improvement. In addition, Figure 3 shows the confusion matrix for our fine-tuned BERT model when applied to the AI-generated misinformation dataset.

Table . False-positive and true-negative results obtained from testing the fine-tuned BERT model on our combined dataset, compared with the results reported in the study by Zhou et al [33].

Model	FP[a]	TN[b]
Our	15	485
Zhou et al [33]	27	473

[a]FP: false-positive.

[b]TN: true-negative.

Figure 3. The figure displays the confusion matrix of our fine-tuned BERT model evaluated on the artificial intelligence–generated dataset [3].

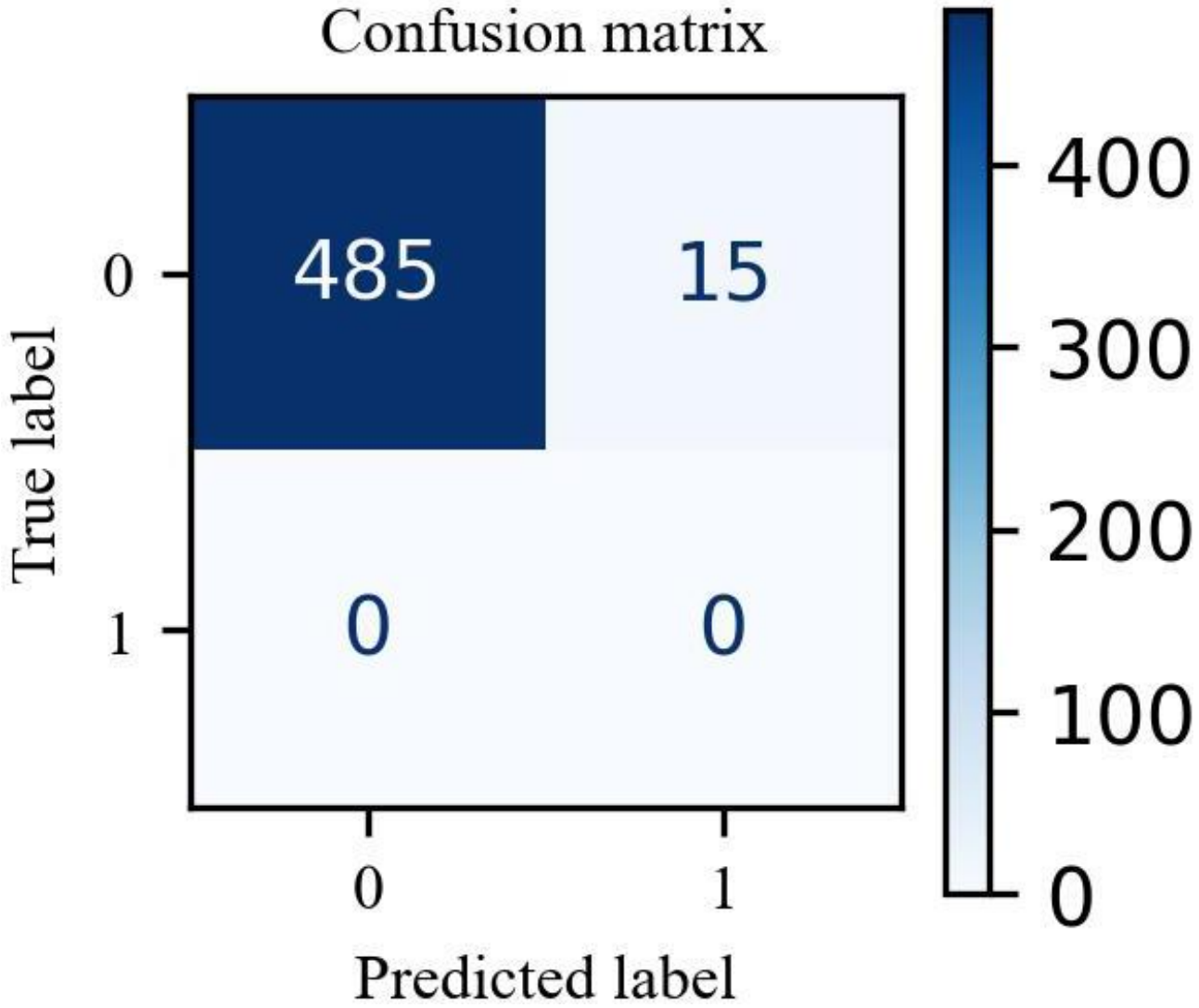

Topic Models

We used the IFCN dataset, one of the largest datasets on COVID-19 pandemic, to apply and evaluate topic modeling approaches. Three models were tested: LDA, Top2Vec, and BERTopic. After applying each topic model, the top 10 words associated with each topic were selected, and the 3 metrics—CV, NPMI, and IRBO—were computed to compare the models. Table 3 summarizes the results across these metrics. Among the models, BERTopic achieved the highest scores across all metrics, leading to its selection for further analysis. In practice, the reported metrics indicate that the topics are moderately coherent and interpretable but not perfectly tight. A coherence score of 0.41 means that the top words in each topic tend to appear together often enough for human analysts to assign clear labels, although some mixing of subthemes is expected. The NPMI of −0.086 is close to neutral, which is typical for short, fragmented social media posts, and suggests that while not every

XSL•FO
RenderX

word pair strongly co-occurs, the overall topics remain meaningful. Combined with the high IRBO score (0.99), these results imply that the model generates a broad, nonredundant set of topics that cover diverse misinformation themes while remaining practically usable for labeling, interpretation, and downstream health communication tasks. Since in the health-related domain, it is important to cover as many diverse topics as possible, an IRBO value above 0.7 can be considered acceptable, while a CV value equal to or above 0.4 and an NPMI close to 0 or higher are reasonable.

Table . Performance of 3 topic modeling approaches—LDA, Top2Vec, and BERTopic—evaluated across 3 metrics: Coherence Value, Normalized Pointwise Mutual Information, and Inverse Ranked Based Overlap[a].

Model	CV[b]	NPMI[c]	IRBO[d]
LDA	0.39	−0.35	0.96
Top2Vec	0.35	−0.29	0.89
BERTopic	0.41	−0.086	0.99

[a]For all metrics, higher values indicate better performance.

[b]CV: Coherence Value.

[c]NPMI: Normalized Pointwise Mutual Information.

[d]IRBO: Inverse Ranked Based Overlap.

Many topic modeling approaches, including BERTopic, often encounter limitations when applied to real-world datasets, as they are unable to assign topics to all documents. This can leave a subset of documents unclassified, reducing the overall effectiveness of the model. To address this issue, we have used the algorithm in Textbox 1, a method for ensuring comprehensive topic assignment across the dataset [68].

Textbox 1. Algorithm: assign a document to the closest topic.

1. Input:
   - Raw text documents X={d1,d2,...,dn},
   - BERTopic model parameters.
2. Topic modeling:
   - Compute topics using the BERTopic model:
     - Topics {Yj}j=1T, = BERTopic(X,P)

     where T is the number of topics and Y_j contains documents that are assigned to the topic j.
3. Sentence embeddings:
   - Transform documents X into vector representations using a sentence transformer such as BERT embedding.
4. Dimensionality reduction:
   - Apply Uniform Manifold Approximation and Projection for dimensionality reduction on the vector representations.
5. Cluster centers:
   - For each topic j, compute the center of the cluster:

   (1)tj=∑i∈jxinj

   where n_j is the number of documents in topic j, and x_i is the reduced vector representation of document.
6. Topic assignment for unassigned documents:
   - For every document d_i that is not assigned to a topic by the BERTopic model, or for any new document:
     - Assign the document to the topic j that maximizes the cosine similarity between the document vector x_i and the cluster center t_j :

   (2)Argmax jxi · tj ‖ xi ‖ ‖ tj ‖

This approach not only ensures that every document in the dataset is assigned a topic but also enhances the interpretability and usability of the topic modeling results. By leveraging the semantic structure of the dataset, our algorithm effectively

XSL•FO
RenderX

bridges the gap between unassigned documents and existing topic clusters, making it a robust solution for comprehensive topic coverage.

Topic Description

As described in the "Methods" section, topic modeling algorithms typically produce word-level representations of topics. While useful for identifying key terms associated with a topic, these representations often lack sufficient contextual information, making interpretation challenging. To overcome this limitation, we developed a structured prompt framework (outlined in Textbox 2 and used the advanced capabilities of LLMs, specifically ChatGPT-4.0, to generate sentence-level representations for the identified topics. These sentence-level representations provide richer context and more interpretable descriptions, enabling a deeper understanding of the topics.

Textbox 2. The prompt structure and 1 example to find topics description.

Topic description prompt structure:

System role:

Topic main words: [Top 10 words]

Topic document examples: [5 closest examples to the center of the topic]

User role:

"Describe topic in a short phrase?"

Topic description prompt example:

System role:

Topic main words: ["masks," "mask," "face," "wearing," "wear," "use," "oxygen," "hypoxia," "cause," and "you''].

- Topic document examples:
  1. Centers for Disease Control and Prevention (CDC) does not recommend wearing masks.
  2. The US CDC contradicted itself by advising people to wear cloth masks against the novel coronavirus while also saying masks do not stop smoke inhalation during a wildfire.
  3. The World Health Organization changed its mind about masks and now says that they can increase the risk of infection.
  4. Nonmedical masks are ineffective in preventing the spread of the disease, are circulating online.
  5. Whether CDC had scheduled announcement that all should wear masks for everyday life.

User role: "Describe topic in a short phrase?"

Output answer: "Controversies and debates over mask wearing and its effectiveness"

The prompt includes the top 10 most representative words for each topic as identified by the topic modeling algorithm, and to add context and depth to the topic descriptions, we select 5 documents that are closest to the center of the corresponding topic cluster. The selection of these documents is guided by cosine similarity, performed using equations 1 and 2, which measure the proximity of documents to the cluster center in the semantic space. An example of this process is provided in Textbox 2, illustrating how the top words and representative documents are integrated into the prompt to produce a high-quality sentence-level representation.

By combining these elements, we construct detailed and context-rich prompts that guide ChatGPT-4.0 in generating coherent and semantically accurate sentence-level topic representations. This approach ensures that the abstract themes identified by topic modeling are translated into human-readable and interpretable descriptions.

To evaluate the quality of the generated topic descriptions, we engaged 3 independent raters to assess the descriptions based on 3 categories: appropriate, somewhat appropriate, and not appropriate. The evaluation results are shown in Table 4 and highlight that the majority of topic descriptions were well received. Specifically, the total proportion of accepted descriptions (the sum of those rated as appropriate and somewhat appropriate) was 99.6%. This acceptance rate demonstrates the effectiveness and reliability of the proposed method for generating meaningful and contextually relevant topic descriptions. There was perfect agreement in 144 out of 169 (85.2%) of the sentences. Two out of 3 raters agreed on category in 24 out of 169 (14.2%) of the sentences. There was a single instance in which no raters agreed in 1 out of 169 (0.6%) of the sentences.

XSL•FO
RenderX

Table . Percentage of topic descriptions rated as “appropriate,” “somewhat appropriate,” and “not appropriate” by each rater, along with the total number of accepted topic descriptions, calculated as the sum of those rated “appropriate” and “somewhat appropriate.”

Raters	Appropriate (%)	Somewhat appropriate (%)	Not appropriate (%)	Total accepted (%)
Rater 1	98.23	1.77	0	100
Rater 2	94.67	5.33	0	100
Rater 3	89.94	8.88	1.18	98.82
Average	94.28	5.32	0.39	99.6

After generating concise descriptions for each topic, we used the structured prompt framework outlined in Textbox 3, which includes a list of these topic descriptions. This structured prompt was then input into the ChatGPT-4.0 API to further refine and categorize the topics into overarching themes.

Textbox 3. Structure of the prompt for identifying topic themes.

Finding topic themes prompt structure:

System role:

The following are topics related to COVID-19 pandemic. Go through all topics and categorize them into relevant groups. Mention topics number for each category.

User role:

Topics description list

The output from this process not only provides a clear categorization of topics into distinct themes but also includes a concise description for each theme. This step ensures that the topics are grouped in a meaningful and interpretable way, facilitating a deeper understanding of the data’s thematic structure.

The categorized topics and their corresponding theme descriptions are provided in Figures 4 and 5, showcasing the effectiveness of the proposed method in generating coherent and insightful thematic groupings.

XSL•FO
RenderX

Figure 4. Descriptions of themes 1-7 along with the corresponding topic descriptions assigned to each theme. CDC: Centers for Disease Control and Prevention.

THEME 1: Home Remedies and Misconceptions About COVID-19 Prevention and Treatment

- Keeping the throat moist as a protection against coronavirus
- Gargling with salt or vinegar in warm water as a method to eliminate coronavirus
- Home remedies and misconceptions regarding COVID-19 prevention
- Drinking water every 15 minutes as a method to prevent COVID-19
- Natural remedies for coronavirus involving garlic water
- Home remedies claimed to cure COVID-19
- Home remedies for coronavirus involving hot water and lemon
- The potential of tea in preventing or curing coronavirus
- Traditional teas and plants as potential cures for COVID-19
- The potential of high doses of vitamin C to prevent or stop the spread of viruses
- Vitamin supplementation for COVID-19 treatment and prevention
- Misinformation about home remedies and drugs curing coronavirus
- Alcohol consumption in relation to COVID-19 prevention and risks
- Misinformation about onions curing coronavirus
- Eating alkaline foods to neutralize SARS-CoV-2
- Misinformation and home remedies for COVID-19 prevention and treatment
- The misconception of steam therapy killing coronavirus
- Alcohol consumption as a potential coronavirus remedy

THEME 2: COVID-19 Deaths and Statistics

- COVID-19-related deaths and contingency plans in Ireland
- Revision of Italy's official COVID-19 death count, reducing it by 97%
- CDC's COVID-19 death count controversies
- COVID-19-related deaths among the vaccinated population in the UK

THEME 3: Conspiracy Theories and Misinformation

- Bill Gates, vaccines, and population control conspiracy theories
- Conspiracy theories about the COVID-19 pandemic being planned
- Conspiracy theories about the coronavirus originating from a laboratory in Wuhan
- According to studies, asymptomatic patients with COVID-19 do not transmit the virus
- Conspiracy theories and scandals involving Dr Anthony Fauci
- Covid-19 denial and conspiracy theories
- Conspiracy theories linking 5G technology to COVID-19
- Spread of misinformation about coronavirus through online video content
- Fake vaccination videos and misinformation
- Claims of the COVID-19 pandemic being staged with actors
- Misconceptions and controversies about vaccines and immunity
- False-positive COVID-19 antigen test results with various substances including Coca-Cola.
- Misinformation about COVID-19 deaths and body disposal
- Empty hospitals as proof of COVID-19 misinformation and conspiracy theories
- Misinformation and rumors about COVID-19 in India
- Videos and claims surrounding the coronavirus outbreak in Wuhan, including conspiracy theories involving American soldiers and the CISM military world games
- Claims about Tasuku Honjo, a Nobel laureate, asserting that COVID-19 was manufactured in a Wuhan laboratory

THEME 5: COVID-19 Testing and Accuracy

- COVID-19 testing and test results
- Controversies and misunderstandings surrounding the use of accuracy of PCR tests in diagnosing COVID-19 disease
- Holding breath as a self-test for COVID-19
- Holding breath for 10 seconds as a coronavirus self-check

THEME 4: Vaccine Controversies and Misinformation

- Misinformation about COVID-19 vaccines causing deaths and injuries in the US, based on misinterpretation of CDC data
- Debate on mandatory vaccination in European countries
- Controversies and misconceptions about the development and effectiveness of coronavirus vaccines
- Misinformation and doubts about the COVID-19 vaccine's effectiveness and safety
- Women's deaths following vaccination
- COVID-19 vaccines and fertility concerns in women
- Controversies and claims regarding Pfizer's COVID-19 vaccine
- Israel's progress and controversy regarding COVID-19 vaccination
- Child fatalities linked to COVID-19 vaccination in Africa
- Controversies and concerns around COVID-19 vaccination among children
- The toxicity and potential organ damage caused by spike proteins from vaccines.
- Misconceptions and fears about mRNA vaccines altering human DNA
- Misconceptions about COVID-19 vaccines causing irreversible genetic changes or DNA damage
- Conspiracy theories about nanoparticles and nanobots in vaccines
- Claims and controversies surrounding graphene oxide in COVID-19 vaccines
- Magnetic conspiracy theories related to COVID-19 vaccines
- Mandatory implantation of microchips for tracking and controlling patients with COVID-19 patients
- Vaccine controversy and misinformation in Australia
- Misinformation about the COVID-19 vaccine causing facial palsy and paralysis, particularly relating to Gov Gavin Newsom.
- Risks of cardiac issues in young people post COVID-19 vaccination
- Soccer and tennis players experiencing health issues related to COVID-19 vaccines
- Airline pilots, COVID-19 vaccinations, and related health concerns
- Donation of blood and plasma by vaccinated individuals
- Misinformation about the Delta variant and COVID-19 vaccinations
- Omicron variant misconceptions and conspiracy theories
- Concerns and misinformation about AstraZeneca vaccine
- Mandatory COVID-19 vaccination in Georgia

THEME 6: Ivermectin and Other Treatments

- Japanese response to COVID-19, usage of ivermectin, and vaccination issues
- The effectiveness of ivermectin as a treatment for COVID-19
- The efficacy of hydroxychloroquine, chloroquine, and other treatments for COVID-19
- Cures and treatments for COVID-19

THEME 7: Government and Political Responses

- Medical perspectives and incidents related to the coronavirus in Europe
- Quarantine measures and restrictions in Ukraine
- COVID-19 vaccination in Ukraine
- COVID-19 response and policies in East African countries
- Covid-19 transmission and response in Taiwan
- COVID-19 and government policies in Colombia and Bolivia
- Statements and actions involving Argentine President Alberto Fernández and the Argentine political scene
- Brazilian politics and COVID-19: The struggles of São Paulo's governor
- COVID-19-related misinformation and initiatives in Mexico and Guatemala
- COVID-19 situation and response in the Philippines
- Controversies and actions of Philippine President Rodrigo Duterte
- President Donald Trump announces Roche's launch of a COVID-19 vaccine
- Economic stimulus package and relief measures in response to the coronavirus pandemic.
- 10,000 Wuhan coronavirus deaths
- Wuhan crisis incidents and responses

XSL•FO
RenderX

Figure 5. Descriptions of themes 8-13 along with the corresponding topic descriptions assigned to each theme.

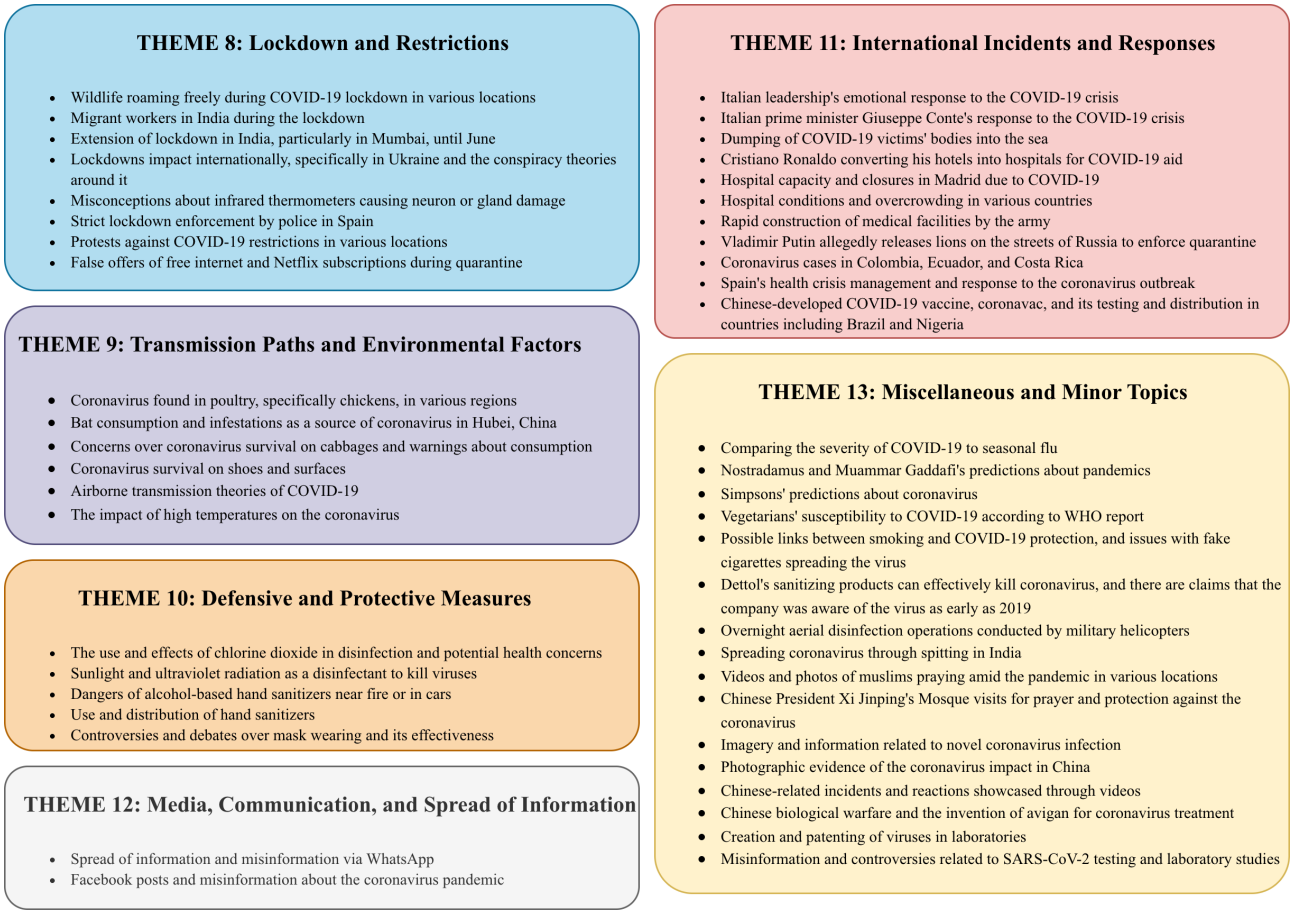

Using the algorithm in Textbox 1, we assign each document to a topic, enabling us to determine the distribution of each theme. Table 5 shows the distribution of all themes, with theme 4 (Vaccines) and theme 3 (Conspiracy Theories) emerging as the first and second most prevalent misinformation themes. Here are all the themes and their percentages:

Table . Distribution of COVID-19 misinformation themes.

Theme	Values (N=18,018), n (%)
Theme 1: Home remedies	1334 (7.40)
Theme 2: Deaths and statistics	570 (3.16)
Theme 3: Conspiracy theories	3459 (19.20)
Theme 4: Vaccine	3595 (19.95)
Theme 5: Testing	546 (3.03)
Theme 6: Ivermectin	814 (4.52)
Theme 7: Government	2094 (11.62)
Theme 8: Lockdowns	1316 (7.30)
Theme 9: Transmission	297 (1.65)
Theme 10: Defensive	1269 (7.04)
Theme 11: International	891 (4.95)
Theme 12: Media	252 (1.40)
Theme 13: Minor topics	1581 (8.77)

Figure 6 shows the distribution of each topic within theme 8 over time, illustrating that protests against COVID-19 restrictions were the only misinformation topic that actively persisted even after the release of the COVID-19 vaccine.

After identifying the misinformation themes, we leverage the explanations provided in the IFCN dataset as a basis to draw refutation arguments to address these themes. For each identified theme, we develop a refutation that aligns with its context,

XSL•FO
RenderX

aiming to clarify the nature of the misinformation, its potential origins, and its impact. These refutations are shown in Textbox 4. The refutation arguments list for 13 themes, providing valuable insights into the underlying reasons for the misinformation.

Figure 6. Distribution of theme 8 topics during the time. FDA: Food and Drug Administration; WHO: World Health Organization.

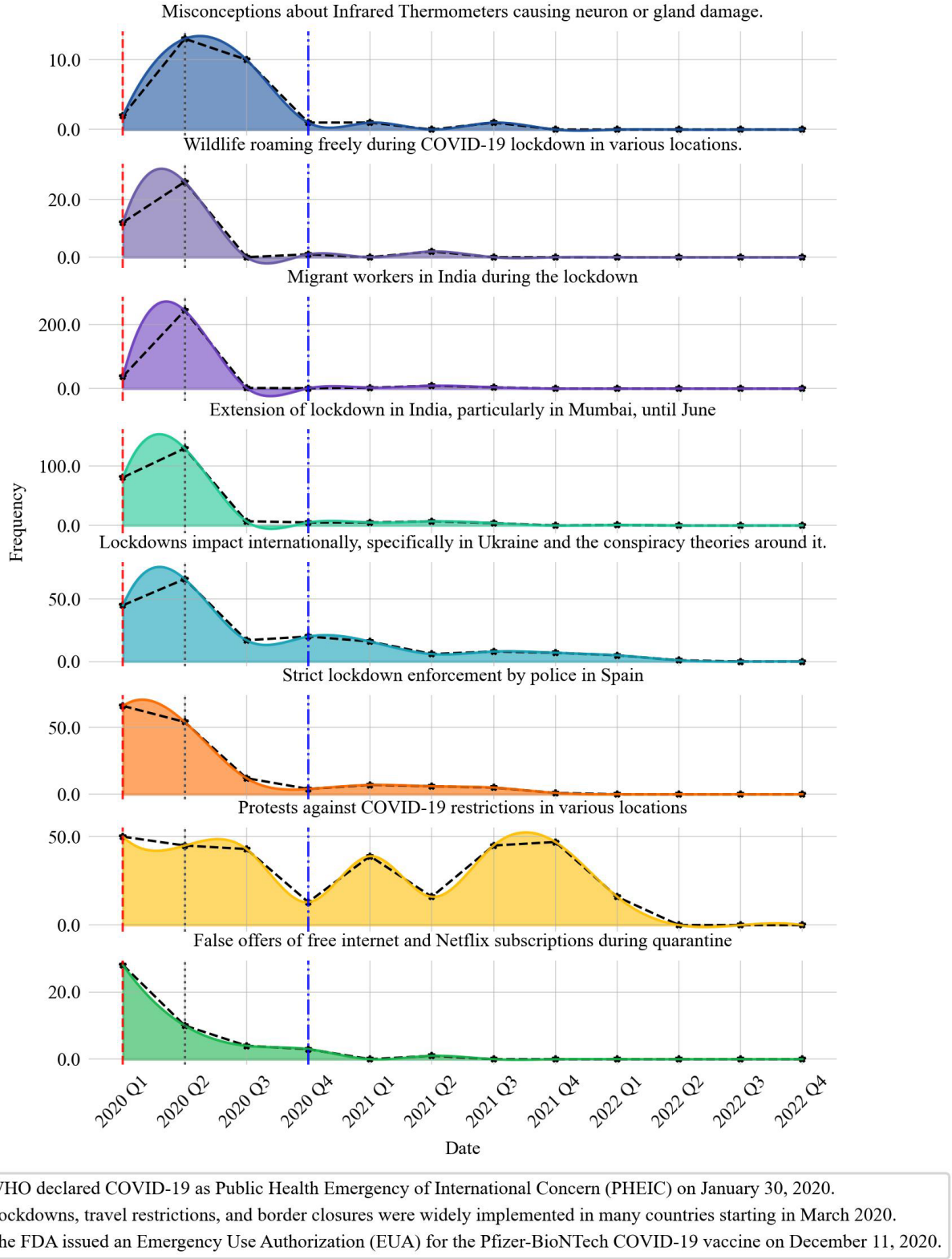

XSL•FO
RenderX

Textbox 4. The refutation arguments list for 13 themes.

Refutation list:

1. This content may contain misinformation related to home remedies for COVID-19 prevention and treatment.
2. This content may contain misinformation related to COVID-19 deaths, statistics, and their relation to vaccination.
3. This content may contain conspiracy theories and misinformation related to COVID-19, including unverified claims, distorted facts, or manipulated content.
4. This content may contain misinformation related to COVID-19 vaccines, including false claims about safety, efficacy, and side effects.
5. This content may contain misinformation related to COVID-19 testing accuracy, including false claims about polymerase chain reaction tests and unscientific self-check methods.
6. This content may contain misinformation related to COVID-19 treatments, including exaggerated claims about ivermectin, hydroxychloroquine, or unproven supplements.
7. This content may contain misinformation related to government and political responses to COVID-19 pandemic, including distorted facts, fabricated claims, or misrepresentation of policies and actions.
8. This content may contain misinformation related to COVID-19 lockdowns and restrictions, including fabricated or misrepresented events, videos, or claims.
9. This content may contain misinformation related to COVID-19 transmission and survival in various environments, including unverified claims about foods, surfaces, or environmental factors.
10. This content may contain misinformation related to defensive and protective measures against COVID-19 pandemic, including false or exaggerated claims about masks, sanitizers, UV rays, or disinfectants.
11. This content may contain misinformation related to international incidents and responses to COVID-19 pandemic, including fabricated reports of government actions, health care capacity, or global cooperation.
12. This content may contain misinformation related to the spread of COVID-19 information on social media and messaging apps, including false claims about government policies, platforms, or media manipulation.
13. This content may contain miscellaneous misinformation related to COVID-19 pandemic, including distortions about products, religious practices, bioweapons, and other fabricated claims.

Provide Refutation

In the final stage of our process, we design a prompt text to enable ChatGPT-4.0 to detect specific misinformation themes. The prompt text includes a detailed description of the themes and a question-answer list. To create this question-answer list, we select the document closest to the center of each topic and associate it with the corresponding theme. Detailed information about the prompt text can be found in Textbox 5.

XSL•FO
RenderX

Textbox 5. Prompt structure and 1 example to find a document theme.

Finding document themes prompt structure:

System role:

"The following is the description of topic themes related to COVID-19 misinformation. Find the closest theme for the given text. Answer in a consistent style."

User role:

Themes description list

Assistant role:

Question-answer list

User role:

Input text

Finding topic themes prompt example:

System role:

"The following is the description of topic themes related to COVID-19 misinformation. Find the closest theme for the given text. Answer in a consistent style."

User role:

Themes description list

Assistant role:

Question-answer list

User role:

"A video shows that Bill Gates admits the vaccine will no doubt kill 700,000 people."

Output answer:

Theme 3: "Conspiracy Theories and Misinformation"

To evaluate our approach, we randomly selected 130 (10 documents per theme) documents from the IFCN dataset that are not included in the question-answer list. We then tested the prompting method with ChatGPT-4.0, achieving an 82% accuracy rate in detecting the correct themes. Moreover, as showing in Figure 7 the model struggled to detect themes 2, 7, and 13 in comparisons with other themes.

XSL•FO
RenderX

Figure 7. Confusion matrix of theme detector.

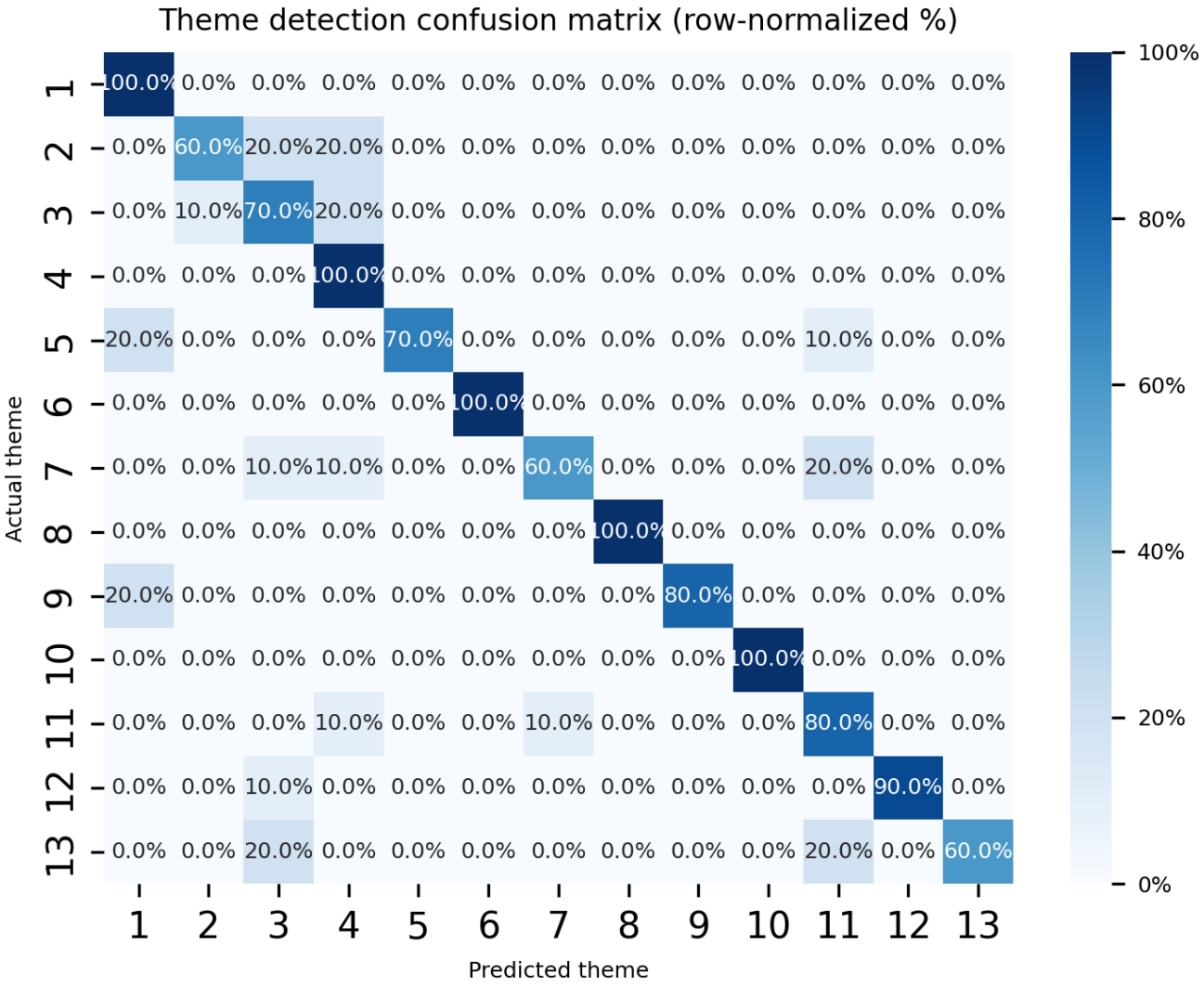

Proposed System

Based on the process we introduced, we propose the development of an MDIS. This system is designed to first determine whether a given text document contains misinformation or not. If the document is identified as containing misinformation, the system then detects its theme and provides a detailed refutation of the misinformation. The primary objectives of MDIS are to prevent the spread of misinformation and to enhance public health knowledge.

MDIS operates by integrating 3 key components. First, it uses a trained LLM to classify text documents as either misinformation or nonmisinformation. Next, it uses another trained LLM to detect the specific misinformation theme within documents identified as containing misinformation. Finally, it leverages a refutation list, which is generated during the theme description phase, to provide context and counternarratives for each detected theme. This comprehensive approach enables the system to effectively address misinformation while equipping users with accurate information.

Figure 1C illustrates an example of an input to MDIS and its corresponding output, demonstrating how the system analyzes a document, identifies misinformation, detects the associated theme, and presents an explanatory response.

Discussion

Principal Results

This study developed the MDIP, a structured workflow for analyzing health misinformation and generating explanatory counterarguments. Building on MDIP, we designed a prototype MDIS that combines detection, topic detection, theme identification, and refutation generation. While the system has not yet been deployed in real-world environments, the results illustrate its potential to support public health communication.

In the misinformation classification task, LLMs achieved reasonable performance, with BERT reaching 98% accuracy. Enriching training datasets also reduced false positives on AI-generated misinformation by 44% compared with prior baselines, suggesting improved robustness across different linguistic styles. Topic modeling experiments highlighted the advantages of BERTopic relative to LDA and Top2Vec, with higher coherence and diversity metrics. To address the issue of unassigned documents, the algorithm in Textbox 1 was used to assign outliers to their nearest topic cluster, thereby improving coverage of the dataset. To enhance interpretability, word-level topic outputs were converted into sentence-level descriptions through prompt engineering with ChatGPT-4.0. These descriptions were judged appropriate or somewhat appropriate

XSL•FO
RenderX

in 99.6% of cases by independent raters, indicating the value of sentence-level representations for clarity and interpretability. Building on these descriptions, topics were further grouped into broader themes and linked with theme-specific refutations.

Finally, a prompt-based detector for misinformation themes achieved 82% accuracy. These results demonstrate how detection, topic modeling, thematic grouping, and refutation can be integrated into a single workflow. The prototype system (MDIS) represents an illustration of concept, and while the findings are encouraging, validation is limited to English language, COVID-19–related data, and offline testing. Broader generalization, multilingual adaptation, integration into health communication workflows, and longitudinal evaluation remain important directions for future work.

Limitations

While this study presents a framework for detecting and addressing misinformation, several limitations should be acknowledged. First, the system's performance relies on the quality and diversity of the datasets that were used for training and topic modeling, which may not fully capture the linguistic and contextual nuances of misinformation in different regions, languages, or cultural contexts. This limitation introduces a potential bias: refutations that appear clear and persuasive in one sociocultural context may be ineffective—or even counterproductive—in another.

The model has been tested solely on English text–based misinformation, and it has not been evaluated for multilingual and multimodal adaptation, meaning its ability to detect misinformation and provide persuasive refutation across different languages and sociocultural contexts remains uncertain. Moreover, while the theme detection module achieved an accuracy of 82%, this leaves an 18% error margin, especially in ambiguous or overlapping themes, which can lead to inaccuracies in refutation and reduce the overall effectiveness of the generated warning texts. Developing and integrating more sophisticated algorithms to address overlapping in topic themes could substantially enhance the accuracy of theme detection. Improved theme separation would not only yield clearer and more coherent thematic structures but also strengthen the generation of precise and contextually relevant refutations. In turn, this refinement would enable more effective countermeasures against health misinformation, thereby improving the system's overall capacity to support public health communication and trust. False positives and negatives remain a concern, particularly when misinformation contains opinion-based, satirical, or context-dependent elements. Although the generated refutations follow a systematic structure, they may not always be contextually relevant, persuasive, or ethically suitable for diverse audiences. In the absence of a human-in-the-loop or oversight mechanism, the system may produce counterarguments that fail to resonate with users or could be perceived as unreliable. In addition, the system has not yet been extensively tested in real-world applications, which limits understanding of its practical impact on misinformation spread and public health outcomes. Furthermore, misinformation evolves over time, and a model trained on past narratives may require periodic retraining to remain effective against emerging falsehoods, including AI-generated misinformation. Using pretrained LLMs such as ChatGPT depends on the current version of the model and its accessibility to users. Therefore, it is necessary to update the system regularly when the model is changed or becomes unavailable. Moreover, since passing datasets through third-party platforms may compromise the security of the framework, future work could focus on developing an in-house solution by training a dedicated model for our specific tasks, thereby eliminating the reliance on external platforms. Finally, the reliance on automated methods raises potential concerns about interpretability and transparency, which are crucial for fostering trust and adoption by end users.

Comparison With Prior Work

The proposed MDIP and the resulting MDIS build upon and advance the body of research focused on misinformation detection and mitigation. The proposed method transforms raw posts into *actionable units*—sentence-level topic labels, aggregated themes, and paired refutations—linking detection outputs directly to message design and response playbooks used by health teams. Previous research has demonstrated the efficacy of ML models, particularly deep learning approaches, in detecting misinformation. They used ML techniques to classify fake news using textual features, demonstrating the value of automated detection methods [49,52,69]. Our study extends these efforts by integrating enriched datasets containing both formal and informal language styles, ensuring better generalization across diverse linguistic sources, including AI-generated misinformation.

Topic modeling techniques such as LDA have been used in prior studies to analyze misinformation [35,53,55]. Our approach improves on these works by addressing limitations in document assignment and theme interpretation. We used an algorithm to assign every document to the most relevant topic, resolving the common issue of unclassified documents in topic modeling. In addition, we moved beyond word-level topic representations to generate sentence-level descriptions, offering richer and more interpretable insights. By tracking shifts in sentence-level topics and theme distributions, communicators can conduct pre-/postassessments of campaigns or platform policy changes, complementing survey-based outcomes. Finally, we designed an effective prompt text to automatically identify the themes of misinformation. This automated approach reduces reliance on manual interpretation, minimizing human bias and increasing scalability.

Many prior studies have addressed misinformation detection or topic analysis in isolation. They analyzed misinformation using sentiment analysis but did not integrate detection with thematic analysis and did not provide a framework for counteracting misinformation [11,35,70]. Our work unifies detection, topic modeling, thematic refutation, and public health intervention in a single framework. The MDIS framework automates the end-to-end process, offering a scalable solution to tackle the complexity of misinformation dynamics.

Conclusions

This work contributes a methodological framework for infodemiology and digital health operations. We transform

XSL•FO
RenderX

misinformation into actionable units—themes and refutations—so that health teams can act (communicate, triage, and evaluate). Moreover, analyzing misinformation using a hierarchical (2-level) sentence-level description and assigning all documents to topics makes it possible to observe theme and topic distributions over time, providing a broad and sensible overview of misinformation. Sentence-level topics and theme distributions serve as measurable indicators for surveillance and intervention evaluation (eg, pre-/postcampaign shifts and surge detection). We introduce MDIP and MDIS that enable rapid response playbooks and reduce analyst workload. To support adoption, we release prompt templates and code as implementation artifacts that teams can readily adapt. Real-world deployment, however, requires governance mechanisms (human-in-the-loop review and audit logs), multilingual extensions, and prospective trials with health agencies or platforms to quantify downstream impact (eg, reduced spread and improved literacy). Ultimately, these contributions orient detection toward operational use—prioritizing interpretability and intervention design—so that public health actors can move from finding misinformation to effectively countering it.

Funding

This research is supported in part by a research grant from the Investigator-Initiated Studies Program of Merck Sharp & Dohme Corp (MISP #102050). The opinions expressed in this paper are those of the authors and do not necessarily represent those of Merck Sharp & Dohme Corp.

Data Availability

All implementation codes can be accessed through the GitHub repository [71].

Conflicts of Interest

None declared.

References

1. Kisa S, Kisa A. A comprehensive analysis of COVID-19 misinformation, public health impacts, and communication strategies: scoping review. J Med Internet Res 2024 Aug 21;26:e56931. [doi: 10.2196/56931] [Medline: 39167790]
2. Loomba S, de Figueiredo A, Piatek SJ, de Graaf K, Larson HJ. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nat Hum Behav 2021 Mar;5(3):337-348. [doi: 10.1038/s41562-021-01056-1] [Medline: 33547453]
3. Moscadelli A, Albora G, Biamonte MA, et al. Fake news and Covid-19 in Italy: results of a quantitative observational study. Int J Environ Res Public Health 2020 Aug 12;17(16):5850. [doi: 10.3390/ijerph17165850] [Medline: 32806772]
4. Chou WYS, Oh A, Klein WMP. Addressing health-related misinformation on social media. JAMA 2018 Dec 18;320(23):2417-2418. [doi: 10.1001/jama.2018.16865] [Medline: 30428002]
5. Kata A. Anti-vaccine activists, Web 2.0, and the postmodern paradigm – an overview of tactics and tropes used online by the anti-vaccination movement. Vaccine (Auckl) 2012 May;30(25):3778-3789. [doi: 10.1016/j.vaccine.2011.11.112]
6. Zimet GD, Rosberger Z, Fisher WA, Perez S, Stupiansky NW. Beliefs, behaviors and HPV vaccine: correcting the myths and the misinformation. Prev Med 2013 Nov;57(5):414-418. [doi: 10.1016/j.ypmed.2013.05.013]
7. Poland GA, Jacobson RM. Understanding those who do not understand: a brief review of the anti-vaccine movement. Vaccine (Auckl) 2001 Mar;19(17-19):2440-2445. [doi: 10.1016/S0264-410X(00)00469-2]
8. Kata A. A postmodern Pandora's box: anti-vaccination misinformation on the internet. Vaccine (Auckl) 2010 Feb 17;28(7):1709-1716. [doi: 10.1016/j.vaccine.2009.12.022] [Medline: 20045099]
9. Oyeyemi SO, Gabarron E, Wynn R. Ebola, Twitter, and misinformation: a dangerous combination? BMJ 2014 Oct 14;349:g6178. [doi: 10.1136/bmj.g6178] [Medline: 25315514]
10. Geoghegan S, O'Callaghan KP, Offit PA. Vaccine safety: myths and misinformation. Front Microbiol 2020;11:372. [doi: 10.3389/fmicb.2020.00372] [Medline: 32256465]
11. Zhou X, Coiera E, Tsafnat G, Arachi D, Ong MS, Dunn AG. Using social connection information to improve opinion mining: identifying negative sentiment about HPV vaccines on twitter. In: Studies in Health Technology and Informatics 2015:761-765. [doi: 10.3233/978-1-61499-564-7-761]
12. Ghaddar A, Khandaqji S, Awad Z, Kansoun R. Conspiracy beliefs and vaccination intent for COVID-19 in an infodemic. PLoS One 2022;17(1):e0261559. [doi: 10.1371/journal.pone.0261559] [Medline: 35020721]
13. Ghosh D, Scott B. Disinformation is becoming unstoppable. TIME. 2018. URL: https://time.com/5112847/facebook-fake-news-unstoppable/ [accessed 2025-11-28]
14. Qazvinian V, Rosengren E, Radev D, Mei Q. Rumor has it: identifying misinformation in microblogs. 2011 Presented at: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing; Jul 27-31, 2011; Edinburgh, Scotland, UK p. 1589-1599 URL: https://aclanthology.org/D11-1147.pdf [accessed 2025-12-19]
15. Wang Y, McKee M, Torbica A, Stuckler D. Systematic literature review on the spread of health-related misinformation on social media. Soc Sci Med 2019 Nov;240:112552. [doi: 10.1016/j.socscimed.2019.112552] [Medline: 31561111]

XSL•FO
RenderX

16. Broniatowski DA, Jamison AM, Qi S, et al. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. Am J Public Health 2018 Oct;108(10):1378-1384. [doi: 10.2105/AJPH.2018.304567] [Medline: 30138075]
17. Zarocostas J. How to fight an infodemic. Lancet 2020 Feb 29;395(10225):676. [doi: 10.1016/S0140-6736(20)30461-X] [Medline: 32113495]
18. Islam MS, Kamal AHM, Kabir A, et al. COVID-19 vaccine rumors and conspiracy theories: the need for cognitive inoculation against misinformation to improve vaccine adherence. PLoS One 2021;16(5):e0251605. [doi: 10.1371/journal.pone.0251605]
19. Gallotti R, Valle F, Castaldo N, Sacco P, De Domenico M. Assessing the risks of "infodemics" in response to COVID-19 epidemics. Nat Hum Behav 2020 Dec;4(12):1285-1293. [doi: 10.1038/s41562-020-00994-6] [Medline: 33122812]
20. Cinelli M, Quattrociocchi W, Galeazzi A, et al. The COVID-19 social media infodemic. Sci Rep 2020;10(1). [doi: 10.1038/s41598-020-73510-5]
21. Mian A, Khan S. Coronavirus: the spread of misinformation. BMC Med 2020 Mar 18;18(1):89. [doi: 10.1186/s12916-020-01556-3] [Medline: 32188445]
22. Kumar N, Corpus I, Hans M, et al. COVID-19 vaccine perceptions in the initial phases of US vaccine roll-out: an observational study on reddit. BMC Public Health 2022 Mar 7;22(1):446. [doi: 10.1186/s12889-022-12824-7] [Medline: 35255881]
23. Kim JW, Lee J, Dai Y. Misinformation and the paradox of trust during the covid-19 pandemic in the U.S.: pathways to risk perception and compliance behaviors. J Risk Res 2023 May 4;26(5):469-484. [doi: 10.1080/13669877.2023.2176910]
24. Hou Z, Du F, Zhou X, et al. Cross-country comparison of public awareness, rumors, and behavioral responses to the COVID-19 epidemic: infodemiology study. J Med Internet Res 2020;22(8):e21143. [doi: 10.2196/21143]
25. Bavel JJV, Baicker K, Boggio PS, et al. Using social and behavioural science to support COVID-19 pandemic response. Nat Hum Behav 2020;4(5):460-471. [doi: 10.1038/s41562-020-0884-z]
26. Schiffman MH, Bauer HM, Hoover RN, et al. Epidemiologic evidence showing that human papillomavirus infection causes most cervical intraepithelial neoplasia. J Natl Cancer Inst 1993 Jun 16;85(12):958-964. [doi: 10.1093/jnci/85.12.958] [Medline: 8388478]
27. Bosch FX, Manos MM, Munoz N, et al. Prevalence of human papillomavirus in cervical cancer: a worldwide perspective. JNCI Journal of the National Cancer Institute 1995 Jun 7;87(11):796-802. [doi: 10.1093/jnci/87.11.796] [Medline: 7791229]
28. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019 Jan;69(1):7-34. [doi: 10.3322/caac.21551] [Medline: 30620402]
29. Reasons to get vaccinated. Centers for Disease Control and Prevention. 2021. URL: https://www.cdc.gov/hpv [accessed 2025-11-28]
30. Pingali C, Yankey D, Elam-Evans LD, et al. National, regional, state, and selected local area vaccination coverage among adolescents aged 13-17 Years—United States, 2020. MMWR Morb Mortal Wkly Rep 2021 Sep 3;70(35):1183-1190. [doi: 10.15585/mmwr.mm7035a1] [Medline: 34473682]
31. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85. [doi: 10.2196/jmir.1933] [Medline: 23615206]
32. Levin S. Facebook promised to tackle fake news but the evidence shows it's not working. The Guardian. 2017. URL: https://www.theguardian.com/technology/2017/may/16/facebook-fake-news-tools-not-working [accessed 2025-12-12]
33. Zhou J, Zhang Y, Luo Q, Parker AG, De Choudhury M. Synthetic lies: understanding ai-generated misinformation and evaluating algorithmic and human solutions. CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems 2023:1-20. [doi: 10.1145/3544548.3581318]
34. Jiang B, Tan Z, Nirmal A, Liu H. Disinformation detection: an evolving challenge in the age of llms. In: Proceedings of the 2024 SIAM International Conference on Data Mining (SDM): Society for Industrial and Applied Mathematics Publications; 2024:427-435. [doi: 10.1137/1.9781611978032.50]
35. Du J, Preston S, Sun H, et al. Using machine learning-based approaches for the detection and classification of human papillomavirus vaccine misinformation: infodemiology study of Reddit discussions. J Med Internet Res 2021 Aug 5;23(8):e26478. [doi: 10.2196/26478] [Medline: 34383667]
36. Tomaszewski T, Morales A, Lourentzou I, et al. Identifying false human papillomavirus (HPV) vaccine information and corresponding risk perceptions from twitter: advanced predictive models. J Med Internet Res 2021 Sep 9;23(9):e30451. [doi: 10.2196/30451] [Medline: 34499043]
37. Farajijalal M, Malek S, Toudeshki A, Viers JH, Ehsani R. Data-driven model to improve mechanical harvesters for nut trees. 2024 Presented at: 2024 ASABE Annual International Meeting; Jul 28-31, 2024; California p. 1. [doi: 10.13031/aim.202400858]
38. Malek S, Salehkaleybar S, Amini A. Multi variable-layer neural networks for decoding linear codes. 2020 Presented at: 2020 8th Iran Workshop on Communication and Information Theory (IWCIT); May 26-28, 2020; Tehran, Iran p. 1-6. [doi: 10.1109/IWCIT50667.2020.9163473]
39. Chui M, Manyika J, Miremadi M, Henke N, Chung R, Nel P, et al. Notes from the AI frontier: insights from hundreds of use cases. : McKinsey Global Institute; 2018.

XSL•FO
RenderX

40. Acemoglu D, Restrepo P. Artificial intelligence, automation, and work. In: Agrawal A, Gans J, Goldfarb A, editors. The Economics of Artificial Intelligence: An Agenda: University of Chicago Press; 2018:197-236.
41. Micallef N, He B, Kumar S, Ahamad M, Memon N. The role of the crowd in countering misinformation: a case study of the COVID-19 infodemic. 2020 Presented at: 2020 IEEE International Conference on Big Data (Big Data); Dec 10-13, 2020; Atlanta, GA, USA p. 748-757. [doi: 10.1109/BigData50022.2020.9377956]
42. He B, Ahamad M, Kumar S. Reinforcement learning-based counter-misinformation response generation: a case study of COVID-19 vaccine misinformation. WWW '23: Proceedings of the ACM Web Conference 2023 :2698-2709. [doi: 10.1145/3543507.3583388]
43. Jia R, Liang P. Adversarial examples for evaluating reading comprehension systems. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: Association for Computational Linguistics; 2017:2021-2031. [doi: 10.18653/v1/D17-1215]
44. Ribeiro MT, Singh S, Guestrin C. Semantically equivalent adversarial rules for debugging NLP models. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1): Association for Computational Linguistics; 2018. [doi: 10.18653/v1/P18-1079]
45. Pham C, Hoyle A, Sun S, Resnik P, Iyyer M. TopicGPT: a prompt-based topic modeling framework. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Association for Computational Linguistics; 2024. [doi: 10.18653/v1/2024.naacl-long.164]
46. Mu Y, Dong C, Bontcheva K, Song X. Large language models offer an alternative to the traditional approach of topic modelling. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): ELRA and ICCL; 2024:10160-10171 URL: https://aclanthology.org/2024.lrec-main.887/ [accessed 2025-12-22]
47. Doi T, Isonuma M, Yanaka H. Topic modeling for short texts with large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4): Association for Computational Linguistics; 2024:21-33. [doi: 10.18653/v1/2024.acl-srw.3]
48. Banas JA, Rains SA. A meta-analysis of research on inoculation theory. Commun Monogr 2010 Sep;77(3):281-311. [doi: 10.1080/03637751003758193]
49. Patwa P, Sharma S, Pykl S, Guptha V, Kumari G, Akhtar M. Fighting an infodemic: COVID-19 fake news dataset. : Springer; 2021 Presented at: Combating Online Hostile Posts in Regional Languages during Emergency Situation. [doi: 10.1007/978-3-030-73696-5_3]
50. Saenz J, Gopal SRK, Shukla D. COVID-19 fake news infodemic research dataset (COVID19-FNIR dataset). IEEE Dataport. URL: https://dx.doi.org/10.21227/b5bt-5244 [accessed 2025-12-22]
51. Rai S, Kornides M, Morgan J, Kumar A, Cappella J, Guntuku SC. Detecting and monitoring concerns against HPV vaccination on social media using large language models. Sci Rep 2024 Jun 21;14(1):14362. [doi: 10.1038/s41598-024-64703-3] [Medline: 38906941]
52. Glazkova A, Glazkov M, Trifonov T. g2tmn at constraint@ AAAI2021: exploiting CT-BERT and ensembling learning for COVID-19 fake news detection. In: International Workshop on Combating On Line Hostile Posts in Regional Languages during Emergency Situation: Springer; 2021:116-127. [doi: 10.1007/978-3-030-73696-5_12]
53. Song X, Petrak J, Jiang Y, Singh I, Maynard D, Bontcheva K. Classification aware neural topic model for COVID-19 disinformation categorisation. PLoS One 2021;16(2):e0247086. [doi: 10.1371/journal.pone.0247086] [Medline: 33600477]
54. Ball P, Maxmen A. The epic battle against coronavirus misinformation and conspiracy theories. Nature New Biol 2020 May 28;581(7809):371-374. [doi: 10.1038/d41586-020-01452-z]
55. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. J Mach Learn Res 2003;3(Jan):993-1022 [FREE Full text]
56. Karas B, Qu S, Xu Y, Zhu Q. Experiments with LDA and Top2Vec for embedded topic discovery on social media data-A case study of cystic fibrosis. Front Artif Intell 2022;5:948313. [doi: 10.3389/frai.2022.948313] [Medline: 36062265]
57. Egger R, Yu J. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. Front Sociol 2022;7:886498. [doi: 10.3389/fsoc.2022.886498] [Medline: 35602001]
58. Abdelrazek A, Eid Y, Gawish E, Medhat W, Hassan A. Topic modeling algorithms and applications: a survey. Inf Syst 2023 Feb;112:102131. [doi: 10.1016/j.is.2022.102131]
59. Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. Presented at: WSDM '15: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining; Feb 2-6, 2015; Shanghai China p. 399-408. [doi: 10.1145/2684822.2685324]
60. Bouma G. Normalized (pointwise) mutual information in collocation extraction. In: From Form to Meaning: Processing Texts Automatically (Proceedings of the Biennial GSCL Conference 2009): Gunter Narr Verlang; 2009:31-40 URL: https://books.google.com/[illegible] [accessed 2025-12-22]
61. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. arXiv. Preprint posted online on Jul 22, 2020. [doi: 10.48550/arXiv.2005.14165]

XSL•FO
RenderX

62. International Fact-Checking Network (IFCN). COVID-19 Fact-Checking Database. Poynter Institute. URL: https://www.poynter.org/ifcn-covid-19-misinformation/ [accessed 2025-12-22]
63. 45 CFR part 46 – protection of human subjects (common rule). US Department of Health & Human Services. URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46 [accessed 2025-12-22]
64. Kenton J, Toutanova LK. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: Proceedings of NAACL-HLT; Jun 2-7, 2019; Minneapolis, MN p. 2 URL: https://au1206.github.io/assets/pdfs/BERT.pdf [accessed 2025-12-02]
65. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. : OpenAI blog; 2019 URL: https://storage.prod.researchhub.com/uploads/papers/2020/06/01/language-models.pdf [accessed 2025-12-19]
66. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 2020;21(140):1-67 [FREE Full text]
67. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. Sci Rep 2024 Mar 13;14(1):6086. [doi: 10.1038/s41598-024-56706-x] [Medline: 38480847]
68. Outlier reduction. BERTopic. 2025 Sep 18. URL: https://maartengr.github.io/BERTopic/getting_started/outlier_reduction/outlier_reduction.html [accessed 2025-11-28]
69. Ding X, Teng C, Ji D. Fake news detection with context awareness of the publisher. 2023 Jul 1 Presented at: The 35th International Conference on Software Engineering and Knowledge Engineering; Jul 1-10, 2023. [doi: 10.18293/SEKE2023-061]
70. Piedrahita-Valdés H, Piedrahita-Castillo D, Bermejo-Higuera J, et al. Vaccine hesitancy on social media: sentiment analysis from June 2011 to April 2019. Vaccines (Basel) 2011 Apr;9(1):28. [doi: 10.3390/vaccines9010028]
71. MDIP: misinformation detection and inoculation processing. GitHub, Inc. URL: https://github.com/SamiraMalek/MDIP-MDIS [accessed 2025-11-28]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
CDC: Centers for Disease Control and Prevention
CV: Coherence Value
GPT-2 : Generative Pre-trained Transformer 2
HPV: human papillomavirus
IRBO: Inverted Rank-Biased Overlap
LDA: Latent Dirichlet Allocation
LLM: large language model
MDIP: Misinformation Detection and Inoculation Process
MDIS: Misinformation Detection and Inoculation System
ML: machine learning
NPMI: Normalized Pointwise Mutual Information
T5-base: Text-to-Text Transfer Transformer
WHO: World Health Organization

Edited by J Sarvestan; submitted 09.Apr.2025; peer-reviewed by D Chumachenko, EC Choi; revised version received 12.Oct.2025; accepted 13.Oct.2025; published 08.Jan.2026.

Please cite as:
Malek S, Griffin C, Fraleigh RD, Lennon R, Monga V, Shen L
Intervention in Health Misinformation Using Large Language Models for Automated Detection, Thematic Analysis, and Inoculation: Case Study on COVID-19
J Med Internet Res 2026;28:e75500
URL: https://www.jmir.org/2026/1/e75500
doi:10.2196/75500

© Samira Malek, Christopher Griffin, Robert D Fraleigh, Robert Lennon, Vishal Monga, Lijiang Shen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet

XSL•FO
RenderX

Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Developing an AI-Assisted Tool That Identifies Patients With Multimorbidity and Complex Polypharmacy to Improve the Process of Medication Reviews: Qualitative Interview and Focus Group Study

Aseel S Abuzour[1,2*], PhD; Samantha A Wilson[3*], PhD; Alan A Woodall[3,4], PhD; Frances S Mair[5*], Prof Dr; Asra Aslam[2,6*], PhD; Andrew Clegg[1,2], PhD; Eduard Shantsila[3], PhD; Mark Gabbay[3*], Prof Dr; Michael Abaho[3*], PhD; Danushka Bollegala[7], Prof Dr; Harriet Cant[8], MSc; Alan Griffiths[9]; Layik Hama[6,10], PhD; Gary Leeming[3], PhD; Emma Lo[3]; Simon Maskell[11], Prof Dr; Maurice O'Connell[8], PhD; Olusegun Popoola[12], MD; Sam Relton[2,6], PhD; Roy A Ruddle[6,10], Prof Dr; Pieta Schofield[3], PhD; Matthew Sperrin[8], PhD; Tjeerd Van Staa[8], PhD; Iain Buchan[3], Prof Dr; Lauren E Walker[13,14], Prof Dr

[1]Academic Unit for Ageing & Stroke Research, University of Leeds, Leeds, United Kingdom
[2]School of Medicine, Faculty of Medicine and Health, University of Leeds, Leeds, United Kingdom
[3]Institute of Population Health, University of Liverpool, Liverpool, United Kingdom
[4]Directorate of Mental Health and Learning Disabilities, Powys Teaching Health Board, Bronllys, United Kingdom
[5]School of Health and Wellbeing, General Practice and Primary Care, University of Glasgow, Glasgow, United Kingdom
[6]Leeds Institute for Data Analytics, University of Leeds, Leeds, United Kingdom
[7]Department of Computer Science, University of Liverpool, Liverpool, United Kingdom
[8]Division of Informatics, Imaging & Data Science, University of Manchester, Manchester, United Kingdom
[9]NIHR Applied Research Collaboration North West Coast, Liverpool, United Kingdom
[10]School of Computer Science, University of Leeds, Leeds, United Kingdom
[11]Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, United Kingdom
[12]Mersey Care NHS Foundation Trust, Liverpool, United Kingdom
[13]Centre for Experimental Therapeutics, University of Liverpool, Liverpool, United Kingdom
[14]Liverpool University Hospitals NHS Foundation Trust, Liverpool, United Kingdom
[*]these authors contributed equally

Corresponding Author:
Lauren E Walker, Prof Dr
Centre for Experimental Therapeutics
University of Liverpool
Waterhouse Building, Block B Brownlow Street
Liverpool, L69 3GF
United Kingdom
Phone: 44 151 795 5407
Email: md0u10pc@liverpool.ac.uk

Abstract

Background: Structured medication reviews (SMRs) are an essential component of medication optimization, especially for patients with multimorbidity and polypharmacy. However, the process remains challenging due to the complexities of patient data, time constraints, and the need for coordination among health care professionals (HCPs). This study explores HCPs' perspectives on the integration of artificial intelligence (AI)–assisted tools to enhance the SMR process, with a focus on the potential benefits of and barriers to adoption.

Objective: This study aims to identify the key user requirements for AI-assisted tools to improve the efficiency and effectiveness of SMRs, specifically for patients with multimorbidity, complex polypharmacy, and frailty.

XSL•FO
RenderX

Methods: A qualitative study was conducted involving focus groups and semistructured interviews with HCPs and patients in the United Kingdom. Participants included physicians, pharmacists, clinical pharmacologists, psychiatrists from primary and secondary care, a policy maker, and patients with multimorbidity. Data were analyzed using a hybrid inductive and deductive thematic analysis approach to identify themes related to AI-assisted tool functionality, workflow integration, user-interface visualization, and usability in the SMR process.

Results: Four major themes emerged from the analysis: innovative AI potential, optimizing electronic patient record visualization, functionality of the AI tool for SMRs, and facilitators of and barriers to AI tool implementation. HCPs identified the potential of AI to support patient identification and prioritizing those at risk of medication-related harm. AI-assisted tools were viewed as essential in detecting prescribing gaps, drug interactions, and patient risk trajectories over time. Participants emphasized the importance of presenting patient data in an intuitive format, with a patient interface for shared decision-making. Suggestions included color-coding blood results, highlighting critical medication reviews, and providing timelines of patient medical histories. HCPs stressed the need for AI tools to integrate seamlessly with existing electronic patient record systems and provide actionable insights without overwhelming users with excessive notifications or "pop-up" alerts. Factors influencing the uptake of AI-assisted tools included the need for user-friendly design, evidence of tool effectiveness (though some were skeptical about the predictive accuracy of AI models), and addressing concerns around digital exclusion.

Conclusions: The findings highlight the potential for AI-assisted tools to streamline and optimize the SMR process, particularly for patients with multimorbidity and complex polypharmacy. However, successful implementation depends on addressing concerns related to workflow integration, user acceptance, and evidence of effectiveness. User-centered design is crucial to ensure that AI-assisted tools support HCPs in delivering high-quality, patient-centered care while minimizing cognitive overload and alert fatigue.

(J Med Internet Res 2026;28:e74304) doi:10.2196/74304

KEYWORDS

structured medication reviews; medicine optimization; health technology; risk stratification; artificial intelligence; AI

Introduction

Background

The growing prevalence of multiple long-term conditions and complex polypharmacy among older adults poses significant challenges for health care systems globally. Structured medication reviews (SMRs) are a key clinical intervention, approved by the National Institute for Health and Care Excellence (NICE), designed to facilitate shared decision-making between clinicians and patients, optimize prescribing, and reduce medication-related harm in patients at high risk who are experiencing problematic polypharmacy [1,2].

General practitioners (GPs), pharmacists, and advanced nurse practitioners who meet training criteria can conduct SMRs. As a commissioned service, the prevailing expectation is for clinical pharmacists within primary care networks (PCNs) to proactively identify patients suitable for an SMR and conduct these reviews [3]. However, it is increasingly recognized that effective SMRs are difficult to implement clinically due to time pressures, fragmented clinical records, and the cognitive burden placed on clinicians when trying to assimilate information from various different sources in order to make shared, person-centered decisions [4].

Currently, in the United Kingdom, a few artificial intelligence (AI)–assisted tools are available to help health care professionals (HCPs) prioritize patients for SMRs. Tools available are usually based on predefined conditions or medications; the examples include Ardens Search [5] and Proactive Register Management Diabetes [4,6]. Prescribing safety indicators have also been used as a technology-based intervention to identify potentially inappropriate prescribing to reduce the number of patients at risk of hazardous prescribing [7,8]. However, primary care–embedded clinical decision support systems (CDSSs), such as audit and feedback tools, are often limited by data supply [9]. Emerging digital technologies, including AI, offer opportunities to enhance the efficiency and effectiveness of the SMR process through automation of routine tasks, rapid data extraction and synthesis, and highlighting clinical risks to support decision-making. However, the integration of AI into clinical workflows is in its infancy, and questions exist about its accuracy, clinical utility, usability, and trust. These implementation barriers are currently unexplored.

This study is part of a larger DynAIRx (AI for dynamic prescribing optimization and care integration in multimorbidity) project. Our research to date has highlighted the time-intensive nature of SMRs and the lack of AI-assisted tools to efficiently identify and prioritize patients [4]. Findings emphasized the need for an AI-assisted tool to identify, prioritize, and reduce the time needed to understand the patient journey in order to optimize medicines appropriately and reduce the risk of potential harm from medicines [4]. DynAIRx involves developing novel AI-assisted approaches to improve the efficiency of SMRs. The planned DynAIRx tool will comprise 4 main components: stratification of patients, clinical trial emulation to understand real-world risk of deprescribing, patient journey visualization through interactive timelines, and a knowledge support system integrating individualized patient risks to support decision-making. The deep learning AI component of this is to develop a tool to stratify patients most in need of an SMR. The DynAIRx stratification tool will compare 2 main approaches to identifying which patients are most at risk of medication-related harm: investigating the trade-off between model performance and explainability in the SMR context. First,

XSL•FO
RenderX

a simple logistic regression model not only gives a baseline performance level to assess the AI-based approach but also is clearly explainable and technically feasible to implement within clinical systems, such as EMIS and SystmOne. Second, a novel approach based on graph neural networks will be used to incorporate the sequence and timing of clinical events into predictions. This is likely to offer superior performance but will be difficult to explain and implement.

Large language models (LLMs), such as ChatGPT (OpenAI), are breaking new ground as an adjunct to support clinical decision-making. In radiological decision-making, ChatGPT recently showed impressive accuracy in the appropriate identification of imaging to support breast cancer screening [10]. There are several proof-of-concept AI-assisted tools in development to support complex polypharmacy. For example, the approach based on discriminator-enhanced encoder-decoder architecture for accurate prediction of adverse effects in polypharmacy is an AI model developed to predict adverse drug-drug interactions [11]. However, its effectiveness in a clinical setting has not yet been attempted. A new LLM based on retrieval-augmented generation has been developed to support pharmacists in identifying medicine errors [12]. User testing has been undertaken with simulation of complex scenarios and a multidisciplinary expert panel; however, true workforce implementation is still to be undertaken. Drug GPT is a specialized proprietary LLM tool for predicting medication safety events, developed by Oxford's AI for Healthcare Lab. It initially garnered popular attention upon the release of the preprint in 2023, including a review in the *Guardian* [13]. However, the preprint was subsequently removed by the authors, and its route to clinical implementation remains unclear [14]. Despite rapid progress, there remains limited understanding of what HCPs and patients actually need from such tools to support SMRs and, importantly, how to embed them into routine clinical practice, particularly in the context of multimorbidity and complex care.

While not all components of the DynAIRx polypharmacy tool will be AI assisted, the field of AI-assisted health technology is rapidly advancing. Consequently, it is critical to understand the user requirements for AI-assisted technologies now, as it may in fact be the case that simple, non-AI solutions can address the challenge in a straightforward and explainable way. Therefore, the health care sector is at a critical juncture when it comes to understanding end-user requirements for medication support.

This Study

This study aimed to explore the perspectives of both HCPs and patients on the potential role of AI in supporting SMRs, with a focus on identifying the core user requirements, anticipated benefits, and key barriers to implementation.

Methods

Participants and Recruitment

This study sought to recruit HCPs or management professionals from UK primary care (community based) and secondary care (hospital services) settings, for whom reviews of prescription medications form a routine part of clinical workload. Participants included those working in general practice; secondary care hospital services (geriatric medicine, clinical pharmacology, falls clinics, and mental health practitioners); clinical commissioning, service management (practice managers); and pharmacists, including PCN pharmacists who conduct SMRs across multiple GP practices. Patient participants included (1) those with mental and physical comorbidities, (2) those with complex multimorbidities, and (3) older people with frailty. In addition, patient and carer representatives from these 3 key multimorbidity groups were recruited, comprising adults aged >18 years with or caring for someone with multimorbidity (4 or more), coexisting mental and physical health problems, ≥10 or more prescribed medications, or frailty. Patient participants self-identified as not digitally engaged. As the General Data Protection Regulation was not required, we did not collect demographic data from patient participants.

Purposive sampling identified potential HCP participants actively involved in medicine optimization services through the researcher's clinical and professional networks. Snowball sampling, where current participants referred others, helped identify contacts through existing service providers and advertisements in GP forums and at national events related to clinical polypharmacy research. Patient representatives were recruited purposively via advertisements through the National Institute for Health and Care Research Applied Research Collaboration public advisor networks and research databases at the researcher's host institutions.

Ethical Considerations

The Newcastle North Tyneside Research Ethics Committee (22/NE/0088) granted ethical approval for the DynAIRx study. Written consent was obtained before participation, and withdrawal of consent was permitted at any stage, including after data collection. Audio recordings were transcribed verbatim, anonymized to remove any potentially identifiable information, and assigned participant codes before recordings were subsequently deleted. All data were stored on secure servers in accordance with data protection regulations. Participants received modest compensation in the form of a voucher to acknowledge their time and contribution, consistent with ethical guidance. No participants withdrew consent for the use of their data in this study. Sessions were conducted in person and online (via Microsoft Teams), lasting from 49 to 109 minutes. Data collection and analysis occurred concurrently, adhering to the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist for comprehensive reporting (Multimedia Appendix 1).

Data Collection

Data collection occurred from November 2022 to November 2023. Focus groups (FGs) and semistructured interviews were conducted to gather participants' views. Patient participants were involved in FGs to discuss their shared experiences, while FGs and individual interviews were conducted with HCPs to accommodate time constraints. A semistructured topic guide was developed collaboratively by the research team, informed by existing literature and expert input. The topic guide focused on key elements relevant to the study aims and reflected the

XSL•FO
RenderX

literature and practical insights from clinical practice. This guide was used consistently across both FGs and interviews. Interviews were conducted to complement purposive sampling and address any gaps in representation. These took place either before FGs, to inform key discussion areas, or alongside them for participants unable to attend a group session. FG topic guides and interview schedules were developed and refined by the clinical members of the research team (LEW, AA, AAW, FSM, and AG) and tailored to HCP and patient groups. The topic guides explored the current challenges and limitations in the SMR process that existing advancements in AI and machine learning (ML) can address and the essential components for a user-friendly prescriber feedback system. We also asked participants questions to identify the key components and functionalities needed in a prescriber feedback system to ensure it is useful and user-friendly. Figure 1 depicts the visual representation of the proposed components of an AI-assisted tool (the DynAIRx tool).

Figure 1. Visual representation of proposed components of an artificial intelligence (AI)–assisted tool (the DynAIRx tool [AI for dynamic prescribing optimization and care integration in multimorbidity]).

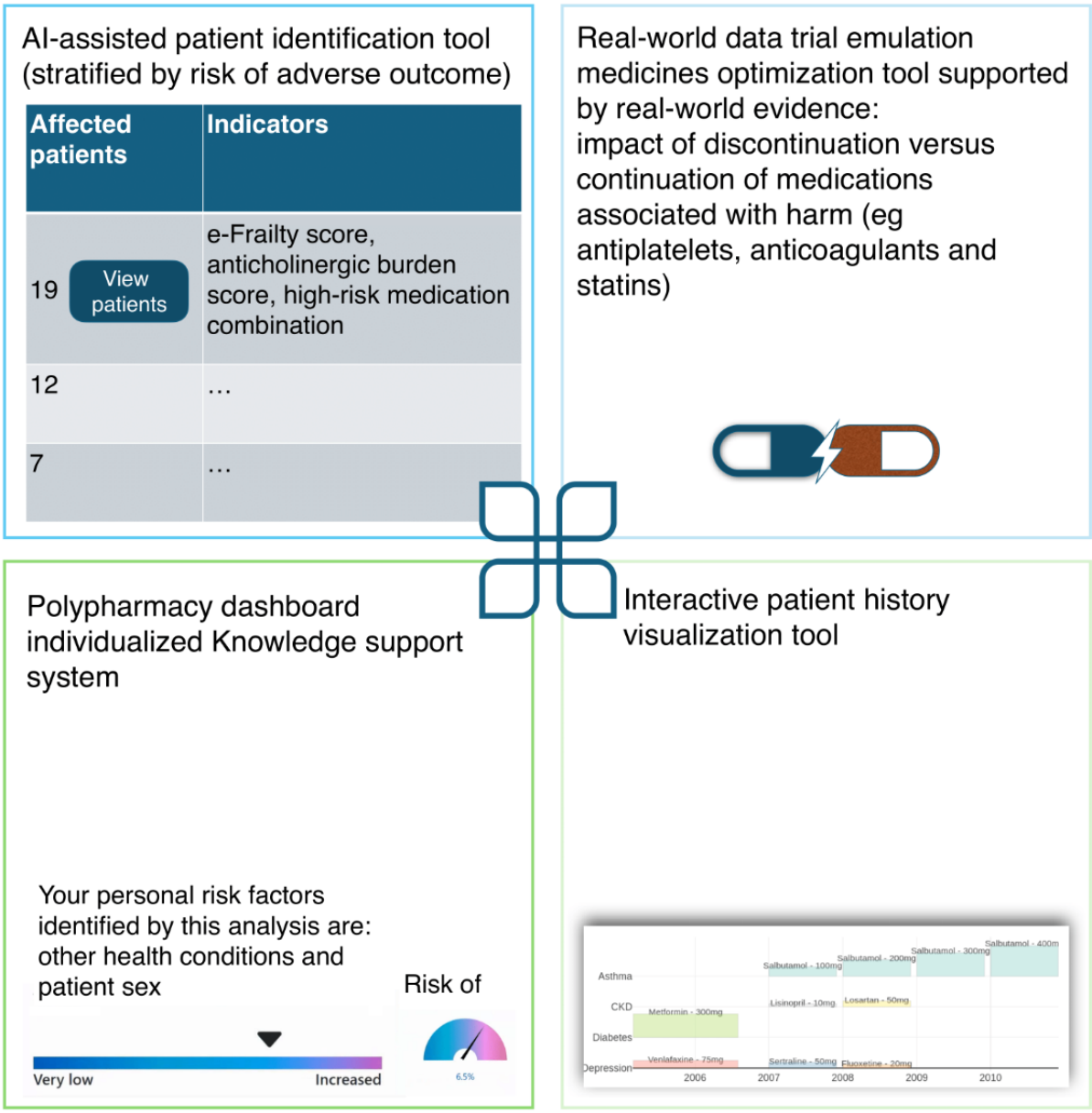

Data Analysis

Data were analyzed thematically, with coding independently conducted by researchers (AA and SAW). Researchers read transcripts to familiarize themselves with the data. Initial coding, guided by inductive reasoning, was conducted by AA and SAW, who collated and examined codes to identify themes. The interview and FG transcripts were coded concurrently to capture both individual and group perspectives. The multidisciplinary coding team (AA, SAW, LEW, AAW, and FSM), comprising clinicians and researchers, engaged in regular coding clinics in a reflexive practice to ensure rigorous and transparent qualitative analysis. These sessions were used to discuss emerging codes, refine codes, and assess data saturation. Discrepancies were

XSL•FO
RenderX

resolved through discussion with the wider team of researchers and clinicians, ensuring diverse perspectives informed theme development and interpretation. This reflexive approach balanced interpretations and mitigated biases, grounding the analysis in participants' narratives. Themes were defined and supported by quotes, with detailed notes maintained to ensure analytic rigor and plausibility. The dataset underwent hybrid inductive and deductive thematic analysis, with iterative revisions of codes and themes. Data saturation was defined as the point at which no new themes or subthemes emerged from the data concerning the already existing themes, which were richly supported by the data. During coding clinics, the research team compared newly coded transcripts against the developing codes to assess whether additional data contributed to novel insights. Recruitment ceased once all team members agreed that no new codes or meanings emerged from successive transcripts, and the thematic structure was considered sufficiently rich to address the study aims and indicate thematic saturation.

Results

Overview

In total, 6 FGs with HCPs (n=21) and 3 FGs with patients (n=13) were conducted (Table 1). A further 5 semistructured interviews with HCPs took place (Table 2). The interviews were undertaken to explore topics in greater depth and address any gaps in purposive sampling. Moreover, the number of participants in each FG differed based on HCPs' availability. However, this did not impact data analysis or saturation, as all HCPs undertook SMRs and expressed their views and requirements for an AI-assisted tool to allow the SMR process to be more efficient and effective.

Table 1. Participant type and the number of participants who took part in the focus groups (FGs; n=34).

FGs	Participants, n (%)
General practitioner FG1	2 (6)
General practitioner FG2	6 (18)
Pharmacist[a] FG1	3 (9)
Pharmacist FG2	5 (15)
Clinical pharmacologist FG1	3 (9)
Psychiatrist FG (mix of secondary care and prison care)	2 (6)
Patient FG comprising individuals with mental and physical health comorbidities for whom prescribing for mental health improvement could lead to adverse physical health consequences	6 (18)
Patient FG comprising those with complex multimorbidity (≥4 long-term health conditions and taking ≥10 drugs)	4 (12)
Patient FG comprising older people with frailty who were at a high risk of adverse outcomes	3 (9)

[a]Mix of primary and secondary care pharmacists.

Table 2. Participant type and the number of participants who took part in the semistructured interviews (n=5).

Participant	Interviews, n (%)
Primary care pharmacist	1 (20)
Secondary care pharmacist	1 (20)
Policy maker	1 (20)
Secondary care psychiatrist	1 (20)
Postgraduate GP[a] trainee	1 (20)

[a]GP: general practitioner.

HCPs conducted SMRs either proactively or reactively, depending on staff capacity, organizational contracts, and practice size. The presence of a PCN pharmacist facilitated proactive SMRs by ensuring that patients meeting directed enhanced service requirements were identified and invited for an SMR. In contrast, GPs and secondary care clinicians often conducted opportunistic medication reviews. Regardless of how patients were identified for a medication review, HCPs described the significant preparation time required to gather and interpret patient information, citing the lack of efficient methods to identify patients at risk of medication-related harm who would benefit most from an SMR [4]. This prompted discussion on how AI approaches could be used to improve the SMR process, including the potential barriers to the uptake and use of AI-assisted tools to support SMRs.

The following 4 overarching themes were developed from the analysis:

1. Innovative AI potential
2. Optimizing electronic patient record (EPR) visualization
3. Functionality of the AI tool for SMRs
4. Facilitators of and barriers to AI tool implementation

XSL•FO
RenderX

Innovative AI potential referred to the emerging possibilities and future impacts of applying AI technologies in health care contexts. Optimizing EPR visualization concerned the enhancement of clarity, usability, and accessibility of clinical information presented within EPR systems. Functionality of the AI tool for SMRs examined how the AI-assisted tool operated and supported the delivery of SMRs. Finally, facilitators of and barriers to AI tool implementation encompassed the organizational, technical, and human factors that influenced the successful integration and effective use of AI-assisted tools in clinical practice.

These themes were not entirely discrete; they reflected interrelated aspects of participants' experiences and perspectives on AI integration in clinical practice, with points of overlap and influence between them. Aspects, such as system usability, functionality, and perceived potential, often interact within the broader context of AI-assisted tools in health care. Participants' perceptions and responses varied according to their professional or personal role (eg, GP, pharmacist, psychiatrist, and patient), highlighting the need for AI-assisted tools to be sufficiently adaptable to address the differing needs of key stakeholder groups.

Innovative AI Potential

Participants expressed their views on the potential utility and advantage of AI in identifying patients at risk of medication-related harm or those who might benefit most from an SMR. HCPs emphasized the need for a tool capable of comprehensively searching EPRs to identify patients with complex multimorbidity. Such a tool should dynamically adjust search outcomes in real time, prioritizing patients who require immediate SMRs.

There was a desire to see AI-assisted tools that could learn autonomously from the historical health care record to identify which factors are contributing to potential medication-related harm. Participants showed great interest in the incorporation of AI with a health care tool to automate tasks and reduce delays in risk prediction. In addition, participants wanted these tools to show patients' medical history in a holistic way using AI capabilities and have the AI be explainable to understand why and how the patient was triggered for a medication review:

> *If you had a funky IT program that looked at medication, looked at what other stuff was happening, looked at, you know, bloods, these are patients that I'm really worried about. So, you know you're talking about machine learning in your project, one of the things, you know, I think will be really interesting to do would be to actually ask the computer what the predictors for certain harms are. We know patients when they fall for multiple reasons, it's not just medicines, but actually wouldn't it be good to see that these are the key circumstances that patients fall under, and then if those circumstances ever happened that patient would be, you know, triggered for a review.* [Policy maker; interview]

> *It could potentially work in real time as well...having something which is live so constantly producing the order of patients who you should be reviewing based on, I don't know a patient might have been discharged from hospital last night that patient might become a bit more high risk and therefore it needs to review earlier. So it flags upon our systems as a, you know, using the AI that this patient will probably need to review in the next four or five days.* [Pharmacist 2; interview]

Leveraging data analytics with ML was viewed as an opportunity to flag patients on complex medication regimens by assessing their health records and prioritizing those patients at risk of medication-related harm. Moreover, aligning patient risk levels with the GP practices staff capacity within PCNs would ensure that those who need immediate attention are seen promptly:

> *That is a real issue for us. It's a real issue for practice actually. So this is why I think the tools have to be a bit more cleverer than just generating, you know, we can generate a list of patients today and that happens, and PCNs at the moment essentially do that, but what you have to do is almost match the list that's generated to the capacity of the build this so you can, the practice has to say that across my PCN I've got, you know, 100 appointments a week to deal with these sort of patients then the tool has to generate that...People would not switch it on if they felt that it could generate lots of patients you would not then see.* [Policy maker; interview]

The development of an AI-assisted tool to support the SMR process prompted discussions on how ML tools could predict risk, identify prescribing gaps, highlight lifestyle and family history risk predictors, and detect potential adverse drug reactions. Advanced digital health tools with AI-assisted features and data analytics could enhance patient engagement by enabling holistic discussions about the patient's risk trajectories and how their medicines can be optimized to reduce any medication-related risks:

> *There used to be a tool, I think it was developed in Australia or New Zealand, but basically it showed a graph of heart disease and trajectory towards symptoms. And you could have a discussion with patients and you can say well look, if we bring your blood pressure down by this much then this is your trajectory...if we stop you smoking, then this is your trajectory. If you develop diabetes then this is your trajectory, and that was probably the single most powerful tool I had to convince people to optimize things like blood pressure or cholesterol reduction.* [Participant 1; GP FG1]

> *I'm sure that having an AI trawl through drug prescribing gaps would tell us quite a lot about medication that may not be taken when it's supposed to have been. We kind of think about that in terms of people misusing analgesics but actually for the elderly population they'll very often just order it because they don't know how to tell us they don't like it or they don't want to take it anymore because they're very*

XSL•FO
RenderX

> *much of the mindset that doctor knows best and they don't want the conflict. But they'll very often forget to order medication and that can be a giveaway.* [Participant 5; GP FG1]

> *I'm thinking of an example where someone has attended an appointment and mentioned that they think a drug is causing a side effect for them, I would imagine that someone would document experiencing this? Being able to see that would be very helpful to try and further add to why things might have been stopped or why a patient might have stopped taking them and maybe not told anyone.* [Participant 2; pharmacist FG1]

> *We did a piece of work recently on familial hypercholesterolaemia and pulled up a lot of patients we didn't realize had family history of massive cholesterol levels and they hadn't realized it was potentially hereditary.* [Participant 5; pharmacist FG2]

Optimizing EPR Visualization

Participants pointed to the challenges associated with the time required to gather and interpret a patient's medical history, emphasizing the need for an AI-assisted tool that optimizes the presentation of relevant information within the EPR. This included reorganizing readily available data to provide a clearer view of the patient's medical history and social circumstances to produce accessible visualizations of the medication timelines, including what medications were prescribed for what condition. Several participants suggested displaying the patient's medical history in a timeline format, detailing key events, such as medication initiation, titration, or discontinuation, diagnosis dates, and recent relevant blood test results:

> *So something pictorially, which helps represent the information in a clearer way, I suppose. Yeah, maybe more longitudinal kind of...And representation of say, when medications were started and titrated up and previous medications, when they were brought in and when there were stopped.* [Psychiatrist; interview]

> *I suppose my top 5 would be: something that highlights previous courses of the same type of medication or the same class, so, for example, if I type in depression as a code, I want an automatic list of every antidepressant they've been on previously and how long they've been on it and which ones they haven't had, even the new ones that are coming on line. I want a list of when they had prednisolone last if they have chronic lung disease, [and] how many courses in the last year they've had without searching for it.* [Participant 1; GP FG1]

AI-assisted tools that reduce the time involved in routine tasks, such as finding information, calculating doses, or assessing disease risk, were welcomed. HCPs were conscious of ensuring that any AI-assisted tool did not overwhelm the user with excessive "pop-up" functions on the display and should not overburden the user's view. There was strong support for a visual timeline that would detail the patient's diagnosis and prescribing and deprescribing journey, along with relevant investigations, diagnostic letters, which specialty diagnosed the condition and started the medication, and the reasons for certain medicine changes. Figure 2 presents the suggested AI-assisted tool features:

> *I've been dreaming about the timeline you showed [laughs] to be able to, in the way that I've imagined it, at the click of a button know...when all the drugs were started and what else was diagnosed around that time [would be] great. And then I don't have to spend any time trying to put that together, that information is there for me. Thinking outside of a hospital setting, if I've got recent bloods and any sort of risk calculations that I want already there on the page from the most recent things, [that is] even better. The amount of time [it would save]; the computer system I work with tells me eGFR, [but] I spend a good chunk of my day calculating everyone's creatinine clearance.* [Participant 2; pharmacist FG1]

Participants also described the challenges around investigating the indication for each prescribed medicine, stating that any AI-assisted tool should incorporate the medication indication:

> *I think for me the most important thing that's missing is indication-based prescribing. Because when we are doing our medicine reviews just trying to work out why anyone's on, you know they could be on an ACE inhibitor, why are they on it, you know, they could be on citalopram, you know, why are they on it. And then it's almost impossible to stop it if you can't work out why someone started it to begin with. So, I think, for me that would be the key initial thing.* [Participant 1; polypharmacy FG]

Participants described how information within a patient's EPR should be visualized, focusing on rearranging and presenting the information in a more intuitive and insightful way to enhance understanding of the patient's medical journey and provide the information in an easily understandable format to support decision-making. They recommended color-coding blood results, highlighting the most recent hospitalization or medication review, and flagging risky medication combinations. One participant noted the following:

> *Reviewing the bloods was quite time-consuming at times, but equally, it was color-coded, so if it was in red it means it's bad news. If it's not in red, it's OK. So, a quick eyeball of words can be often sufficient.* [Pharmacist 2; interview]

Another participant added the following:

> *The kind of things that we need [are] things like...when was their last medication review done or when was the last SMR done, for example? When was the last hospitalization event, for example? Have they been hospitalized or discharged recently? Then it'd be nice to visualize, you know, and the combinations of drugs which could be risky.* [Pharmacist 2; interview]

XSL•FO
RenderX

Figure 2. Suggested features for the development of an artificial intelligence (AI)–assisted tool for use in medication reviews. BNF: British National Formulary; eGFR: estimated glomerular filtration rate; SMR: structured medication review.

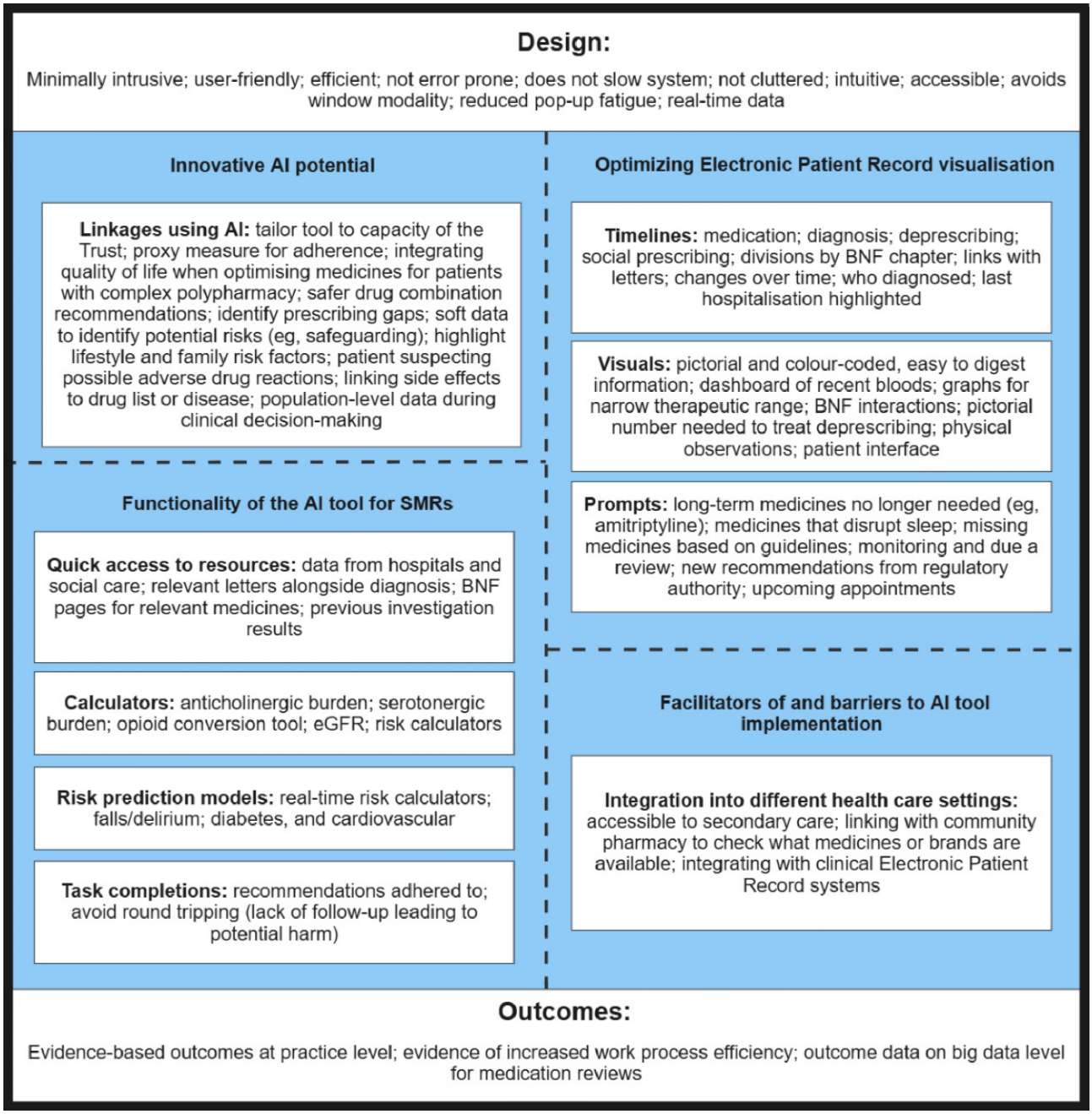

Functionality of the AI Tool for SMRs

Participants articulated their preferences for the functionality of the AI-assisted tool, particularly emphasizing the need to avoid "pop-up fatigue." They suggested implementing specific, targeted pop-ups rather than numerous interruptions that could disrupt workflow. One participant remarked as follows:

> *Any pop ups that you've just got to click through does add even seconds to your working day in each patient record so that can be annoying.* [Participant 1; GP FG2]

Instead, participants preferred notifications for critical issues, such as missing medications in a patient's record that align with guidelines. For example, a participant stated the following:

> *If a patient has recently had a myocardial infarction...it pulls out your main groups of drugs from the NICE guidelines and flash up oh they're not on an ACE inhibitor and you've got to say why.* [Participant 3; pharmacist FG1]

Similarly, they recommended notifications for medications needing review, such as long-term prescriptions of amitriptyline:

> *Just to make people think, actually. This thing which was started for sleep isn't helping.* [Psychiatrist; interview]

Participants also recommended pop-ups for new guideline recommendations:

> *To meet the Commission of Human Medicines and Medicines and Healthcare products Regulatory*

XSL•FO
RenderX

> *Agency recommendations that came in in December, a way of being able to see what antiepileptics have been tried for someone and how long they've been on valproate would really inform the review process.* [Participant 2; pharmacist FG1]

Participants valued AI-assisted tools that enhanced workflow productivity by providing quick access to resources, such as the *British National Formulary* or NICE guidelines based on prescribed medicines or flagged blood test results. One GP noted the following:

> *I want a dashboard of most recent blood tests and current guidance checked against that for where there's missing [medication] gaps, just those things alone would probably save about 2- or 3-weeks' worth of work across my practice per year.* [Participant 1; GP FG1]

In addition, participants highlighted the need for more efficient access to hospital discharge letters and relevant patient information, which often reside in separate electronic systems, such as EMIS (an EPR system) and Docman (document management system). One participant stated the following:

> *If you could draw up relevant letters alongside diagnosis so that you don't have to then trawl through a different system or then write to the GP to request added information on certain diagnosis.* [Psychiatrist; interview]

Furthermore, automated calculators for clinical decision-making, such as anticholinergic burden calculations, estimated glomerular filtration rate, and opioid conversion calculators, were seen as valuable tools to reduce time and enhance efficiency. For example, a GP shared the following:

> *I want automatic opioid conversion so that when they're on tramadol, co-codamol, oramorph, it just gives me a single figure to aim towards and show them, you know, we are at horse tranquilizing level or at dragon tranquilizing level and we need to kind of bring it down.* [Participant 1; GP FG1]

SMRs, whether proactive or opportunistic, often require multiple appointments and may necessitate further investigations by another HCP in a different setting, such as a hospital. Participants were concerned that current systems might fail to notify HCPs of patients needing follow-up care, potentially leading to harmful outcomes for the patient:

> *One of the big issues with medical software is not doing what you describe. Computer programmers call it "round tripping," so basically things aren't followed up and the loop isn't closed, and we see huge amounts of error in medicine because of it. For example, we will refer and then it's off our radar and it's on someone else's radar and then they don't show up at the hospital, and then all of a sudden six months down the line they end up with a malignancy, and then it's back to the original person who referred problem because that loop wasn't closed or the data wasn't followed up. So, having some form of checking system would hugely reduce error and risk within medicine, particularly when it comes to prescribing. So, for example, I've referred them to the physio, the physio then closes that loop when they seem them and then that's dealt with, or the physio does a hydrocortisone injection for example, which I may not need to know about unless I then send them somewhere else, or they are on anticoagulants, or something like that. So, having a tool that ties up those loose ends would be worth its weight in gold I think, particularly in terms of huge amounts of risk reduction.* [Participant 1; GP FG1]

Facilitators of and Barriers to AI Tool Implementation

Participants were curious to understand where DynAIRx might be embedded and how it will be adopted in practice. The main goal was to ensure that any new software does not interfere with current work processes. They described concerns about the cost of adopting it in practice, information governance, and potential medico-legal concerns:

> *I know that I understand X's point about having it that you are able to record what decisions you made within the tool but I wonder if it might be preferable to have a way of a text summary that could be then copied from the tool into the notes. Or, even a way of having a screen shot into letters so, you know, if there were every any questions what was it that you were looking at on the day what information was the tool giving you and if you then made an interesting decision based off that information, that's still your decision. But you've clearly got what is was that you could see.* [Participant 2; pharmacist FG1]

Incentivizing the use of the tool to improve uptake was seen as crucial, with participants suggesting that DynAIRx should be offered as an option to HCPs. They recommended explaining the benefits of using DynAIRx as an additional supportive tool alongside traditional EPR systems, providing examples, such as improvements in efficiency, ease of use, and quick access to information. In addition, HCPs recognized that technology is a double-edged sword, which could either reduce or increase the gap between the HCP and the patient. Consequently, HCPs were open to incorporating a patient interface that would present information pictorially to support shared decision-making. Patients indicated that they would find this useful as they were keen to take a more active role in decision-making. However, many patients were unaware of medication review–related services available to help optimize their medicines:

> *I have never been involved in a review either...Would it just be the GP?...I am going to ask this time whether I can be involved.* [Patient 3; patient, multimorbidity FG]

HCPs stated that, from their experience, patients are fairly comfortable with technology, which was echoed in the patient FGs:

> *I have no problem whatsoever with using [AI] tools, especially if it is to back up the thought process.* [Patient 4; patient, multimorbidity FG]

XSL•FO
RenderX

> *What I think, being a patient, is that you need to have the trust from the GP and the GP pharmacist, whoever is prescribing and using AI, they need to educate people. They need to raise awareness about it. And then you have the trust with the GP then you feel comfortable.* [Patient 1; patient, multimorbidity FG]

However, this did not necessarily mean that all patients were able to use technology themselves to communicate or provide feedback to HCPs through surveys and questionnaires, suggesting the need to consider potential digital exclusion:

> *Not digitally everybody, there are some barriers, language barriers, cultural barriers, and economic barriers for the disadvantaged so we have to think about equity as well. So these barriers are obviously not everything, but we should open up, give the opportunity to all those populations to have the opportunity to discuss and take it rather than have both, it should be both.* [Patient; complex multimorbidity FG]

While HCPs expressed a desire for risk-prediction tools within the DynAIRx software, some remained skeptical about these tools. They emphasized the importance of providing evidence that the model had been validated or that the software itself improved patient outcomes:

> *One of the reasons why people may not use it is that they, clinicians often ask for evidence, so they need to see evidence. A published paper or a trial of actually being useful. So does it improve outcomes with patients? These one of the biggest barriers that I've come across in my time with tools.* [Pharmacist 2; interview]

In addition, the need to ensure confidentiality and consent in AI-assisted tool use is paramount, given the amount of data that would be incorporated into DynAIRx. Patients were aware of the fast-paced innovations taking place in technology and the expectation that technology should be included in health care practice:

> *The world is changing. First we heard about the autopilot, now they are testing [cars] without a driver, virtual GP...now AI in the medical sector as well. The world is changing.* [Participant 2; patient]

Discussion

Principal Findings

Our study outlines the user requirements for the development of an AI-assisted tool to improve the process of conducting an SMR involving patients with complex multimorbidity and polypharmacy. This includes optimizing EPR visualization with a focus on developing patient timelines that outline the patient's medical journey to include when a condition was diagnosed, associated medicines prescribed for that condition, and any associated laboratory results, to name a few. Moreover, participants described the preference for evidence-based outcomes and the use of explainable AI to identify patients at risk of medication-related harm who would benefit from an SMR, determine their risk trajectory over time, and align those patients to practice staff capacity. By "explainable AI," we mean models or algorithms designed with transparent logic or post hoc interpretation methods (eg, feature-importance heatmaps and rule-based approximations) that allow HCPs to understand why a particular patient was flagged as high risk rather than relying on opaque "black box" predictions. From our findings, it is clear that HCPs require an AI-assisted tool that will streamline their work processes when conducting an SMR to easily find information related to the patient and incorporate any risk prediction models. Studies show that CDSSs have the potential to improve process outcomes [15,16]; however, access to CDSSs alone does not guarantee user acceptability or uptake [17]. Concerns remain that complex clinical decision-making, particularly for patients with multimorbidity and polypharmacy, may not be easily translated into algorithms [18] or, at the very least, may have HCPs view the validity of algorithms with some skepticism [19]. This can be seen in the literature that reflects HCP preference for knowledge-based CDSSs over non–knowledge-based CDSSs [19,20]. Despite this, studies show that the lack of algorithm complexity in CDSSs can frustrate HCPs, particularly in how information is presented [18,21]. Moreover, user-centered design is crucial in the development of a CDSS to optimize how information is presented to HCPs to manage cognitive load, alert fatigue, and the impact on workflow [22,23].

Comparisons With Prior Work

AI-assisted tools designed to assist in SMR processes must be tested to ensure they effectively identify both patients who are at risk of medication-related harm and those who have the greatest capacity to benefit from an SMR. These 2 groups are not necessarily synonymous, as some variables that may contribute to harm are not modifiable (eg, very advanced age and frailty), whereas the capacity to benefit from an SMR may be determined by modifiable factors (such as identification of prescribing cascades, drug-drug or drug-disease interactions, and adverse drug reactions). The tool must be sensitive when identifying patients at high risk and display medicines that could be optimized or deprescribed efficiently, enabling future evaluation of medication-related interventions using available data. Patients taking part in an SMR are also likely to have multiple appointments and be referred to different specialists, making it difficult for the HCPs who initiated the SMR to follow this journey. Few studies have explored the interoperability between primary and secondary health care settings, which is essential for effective communication and coordination between HCPs, and is perceived to have a high impact on patient safety [24,25]. Our findings indicate a clear need for an AI-assisted tool to support clinicians and patients during consultations. An integrated system within the EPR software could help summarize and visualize patient journeys and medicine-related information, present personalized risks of harms and benefits of medicines combinations [26], highlight the uncertainties in these risks, and support shared decision-making with patients. While the findings of this study are based on anticipated rather than observed experiences with an AI-assisted tool, engaging end users at the outset before tool development enables the design of an AI-assisted tool that is contextually relevant, cost-effective, and more likely to be adopted in practice [27,28].

XSL•FO
RenderX

Future stages of this work will involve the development and testing of a prototype to validate and refine these preliminary findings through direct user interaction and usability testing, as demonstrated by previous studies [29-32]. This iterative, user-centered process will help ensure that the AI-assisted tool is not only functional but also implementable in real-world settings.

Previous literature highlights the preference among HCPs for knowledge-based over non–knowledge-based CDSSs, which aligns with concerns about the validity and acceptability of algorithms in complex clinical scenarios [19,20]. While frustrations with simplistic CDSS algorithms have been noted [14], our study contributes additional evidence emphasizing the need for user-centered design to address cognitive load, alert fatigue, and workflow impact. The clinical decision support five rights model states that the right information should be presented to the right person, in the right format, via the right channel, and at the right time [16]. Studies on the development of CDSSs emphasize the importance of understanding the needs of users and receivers during the early stages of software development [17,18]. Interoperability between health care settings, while unexplored in previous work, emerged as a critical component for effective SMR implementation and improved patient safety in this study [10,19]. Integrated care boards link primary and secondary care data across their local areas with input from EPR vendors, such as EMIS and SystmOne. An approach consisting of fully linked local data, with minor regional differences in data formatting, may begin to emerge in the coming years as one way to proceed. Efforts, such as the National Health Service Federated Data Platform, aim to unify these actions across the nation and may provide additional clarity on this approach in the coming years. The next stage of the DynAIRx project will involve learning from the user requirements of our stakeholder groups to develop prototypes of an AI-assisted tool. We anticipate the developed prototypes to likely fall under the category of a knowledge support system, which provides HCPs with knowledge that already exists, such as contextual information about a patient drawn from several sources, including historic data on clinical outcomes from comparable patient groups, medical knowledge, and AI to support HCPs during consultations [29,33,34].

To bridge our user-centered requirements with real-world adoption, we recommend framing the development and deployment of the AI-assisted tool for SMRs within established implementation science frameworks, such as the Consolidated Framework for Implementation Research [35]. Under the Consolidated Framework for Implementation Research, the intervention characteristics (eg, usability, adaptability of the patient-timeline visualization, and risk-prediction algorithms), the inner setting (practice culture and EPR interoperability), the outer setting (national data-linkage initiatives, such as the National Health Service Federated Data Platform), the characteristics of individuals (clinician attitudes toward explainable AI), and the implementation process (engagement, training, and feedback loops) all warrant deliberate planning and tailoring.

A staged implementation approach could proceed as presented in Textbox 1.

Textbox 1. Staged implementation approach.

Before implementation (exploration and preparation)

- Conduct targeted workflow analyses in a small number of pilot practices to refine how the timeline and risk outputs map onto existing structured medication review processes
- Engage local electronic patient record vendors (eg, EMIS and SystmOne) to configure interoperability and data-security protocols

Implementation (initial rollout and training)

- Deploy the prototype in 3 to 5 “early adopter” practices, offering hands-on workshops and “superuser” support
- Establish a feedback channel (eg, biweekly focus groups) to iteratively refine

Sustainment (scale-up and longitudinal support)

- Expand to additional practices, embedding the tool within organizational reporting cycles
- Integrate decision-support outputs into routine safety audits and continuing professional development activities

Building on our staged implementation plan, future studies should focus on prospective validation of the AI-assisted tool for SMRs in live clinical settings. Key next steps include the following:

- Pilot effectiveness trials—randomized or stepped-wedge designs comparing standard SMR to SMR with AI support, measuring both process (eg, time per review and alert response rates) and patient-level outcomes (eg, incidence of medication-related harm)
- Usability and acceptability—mixed methods evaluations combining system-log analytics with qualitative interviews to understand how clinicians interact with features such as patient timelines and risk predictions
- Subgroup analyses—examining performance across different patient demographics and varying levels of multimorbidity complexity to ensure equitable benefit

Limitations

This research is part of a larger qualitative study exploring the barriers to SMRs and potential AI-assisted tools. Given the focus on digital-driven solutions, the HCP participants likely included those with a particular interest in such innovations, although efforts were made to include a diverse range of HCPs

XSL•FO
RenderX

from different practice backgrounds, regions, and care settings to mitigate this bias and help strengthen the generalizability of our findings within the context of SMR practices. Moreover, our eligibility criteria focused on recruiting HCPs who are actively involved in conducting medication reviews to explore the barriers to efficient and effective SMRs and how AI-assisted tools may address these barriers [4]. Some FGs had fewer participants due to the competing demands on clinicians' time. However, the data collected were rich and contributed significantly to achieving thematic saturation. While this study provides valuable insights into the user requirements for developing an AI-assisted tool, certain limitations should be acknowledged. First, the diversity of the patient group included in this study may not fully represent the broader patient population, particularly in terms of familiarity with AI-assisted tools. Patients' knowledge and understanding of such systems may vary significantly, potentially influencing the feedback provided. As a result, the findings may not fully capture the perspectives of patients with lower levels of digital health literacy or those who have more experience engaging with AI-assisted tools in clinical settings. Second, the study captures participants' hypothetical perceptions of the AI-assisted tool rather than their experiences of real-world implementation and use. While the findings highlight anticipated benefits and potential facilitators, they may not fully reflect the complexities of adoption and sustained use in practice. Consequently, the research team will develop a prototype based on our findings, which will be tested to validate results from this study and refine and iterate the prototype through a series of think-aloud sessions and semistructured interviews with HCPs. This will ensure the prototype is developed based on user requirements and allow the user to explore the utility of the tool. Moreover, we anticipate a future evaluation of the AI-assisted tool by implementing it into the HCPs' routine workflow. Future research involving live system implementation and longitudinal evaluation would be valuable in assessing the feasibility and actual impact of AI-assisted tool integration in health care settings.

Conclusions

This study highlights the potential of AI-assisted tools to enhance the SMR process for patients with complex multimorbidity and polypharmacy and the user requirements to develop an AI-assisted tool. AI-assisted tools may have the potential to improve patient identification for an SMR, assist the HCP to optimize patient medication by optimizing the EPR visualization of the patients' medical and social history, and support shared decision-making between HCPs and patients. In order to realize the full potential of AI-assisted tools for SMRs, national and local policy makers should consider earmarking targeted funding streams, such as through the National Health Service Digital's Innovation Accelerator or PCN transformation budgets, to subsidize early implementation and integration with EPR systems. Embedding AI-SMR competencies into continuing professional development requirements for pharmacists and GPs, alongside dedicated training grants, will help build workforce capability and ensure equitable uptake. Finally, linking reimbursement incentives (eg, enhanced quality and outcomes framework points to multimorbidity reviews that use validated AI support) could further drive adoption and standardize best practice across PCNs. However, to ensure successful implementation, it is essential to address concerns, such as cognitive load, alert fatigue, and system interoperability. The findings underscore the importance of developing explainable, evidence-backed tools that align with clinical workflows and demonstrate clear benefits to patient outcomes. Ultimately, integrating such tools into an EPR has the potential to improve the efficiency and effectiveness of SMRs, benefiting both patients and health care systems.

Acknowledgments

DynAIRx has been funded by the National Institute for Health and Care Research (NIHR) Artificial Intelligence for Multiple Long-Term Conditions call (NIHR 203986). MG is partly funded by the NIHR Applied Research Collaboration North West Coast. AAW is partly funded by a Health and Care Research Wales Research Time award (NHS-RTA-21-02). AC is funded by a National Institute for Health and Care Research (NIHR) Research Professorship award and supported by the NIHR Applied Research Collaboration Yorkshire & Humber, the NIHR Leeds Biomedical Research Centre, the Cross NIHR Collaboration in Multiple Long-Term Conditions, and Health Data Research UK, an initiative funded by UK Research and Innovation Councils, NIHR and the UK devolved administrations and leading medical research charities. This research is supported by the NIHR Applied Research Collaboration North West Coast. The views expressed in this publication are those of the authors and not necessarily those of the NIHR, NHS or the UK Department of Health and Social Care. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' Contributions

ASA contributed to conceptualization, data analysis, methodology, project administration, supervision, validation, writing the original draft, and reviewing and editing the manuscript. SAW contributed to data analysis, methodology, and reviewing and editing the manuscript. AAW contributed to data acquisition, data analysis, and writing the original draft of the manuscript. FSM contributed to conceptualization, methodology, and reviewing and editing the manuscript. AA contributed to conceptualization, data acquisition, and reviewing and editing the manuscript. AC contributed to conceptualization, methodology, project administration, supervision, and reviewing and editing the manuscript. ES contributed to conceptualization, supervision, and reviewing and editing the manuscript. MG contributed to conceptualization, methodology, and reviewing and editing the manuscript. DB contributed to conceptualization, supervision, and reviewing and editing the manuscript. HC contributed to reviewing and editing the manuscript. AG contributed to reviewing and editing the manuscript. LH contributed to validation, visualization, and

XSL•FO
RenderX

reviewing and editing the manuscript. GL contributed to reviewing and editing the manuscript. EL contributed to project administration and supervision. SM, MO, OP, SR, RAR, and PS contributed to reviewing and editing the manuscript. MS contributed to conceptualization, methodology, and reviewing and editing the manuscript. TVS contributed to conceptualization, methodology, and reviewing and editing the manuscript. IB contributed to conceptualization, data analysis, methodology, project administration, supervision, validation, and reviewing and editing the manuscript. LEW contributed to conceptualization, data analysis, methodology, project administration, supervision, validation, and reviewing and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
COREQ checklist.
[PDF File (Adobe PDF File), 145 KB - jmir_v28i1e74304_app1.pdf]

References

1. Network contract directed enhanced service: structured medication reviews and medicines optimisation: guidance. National Health Service England. URL: https://www.england.nhs.uk/wp-content/uploads/2021/03/B0431-network-contract-des-smr-and-mo-guidance-21-22.pdf [accessed 2025-05-29]
2. Network contract DES guidance for 2024/25: part A - clinical and support services (Section 8). National Health Service England. URL: https://www.england.nhs.uk/publication/network-contract-directed-enhanced-service-guidance-for-2024-25-in-england-part-a-clinical-and-support-services-section-8/ [accessed 2025-05-29]
3. Network contract directed enhanced service – contract specification 2021/22 – PCN requirements and entitlements. National Health Service England. URL: https://www.england.nhs.uk/publication/network-contract-des-specification-2021-22/ [accessed 2025-05-29]
4. Abuzour AS, Wilson SA, Woodall AA, Mair FS, Clegg A, Shantsila E, et al. A qualitative exploration of barriers to efficient and effective structured medication reviews in primary care: findings from the DynAIRx study. PLoS One 2024;19(8):e0299770 [FREE Full text] [doi: 10.1371/journal.pone.0299770] [Medline: 39213435]
5. The leaders in providing EMIS web and SystmOne templates and resources. Ardens Healthcare Informatics. URL: https://www.ardens.org.uk/ [accessed 2025-05-29]
6. A health management tool for people with diabetes co-created by Lilly and NHS Devon CCG. PARM Diabetes. URL: https://parmdiabetes.co.uk/ [accessed 2025-05-29]
7. Abuzour AS, Magola-Makina E, Dunlop J, O'Brien A, Khawagi WY, Ashcroft DM, et al. Implementing prescribing safety indicators in prisons: a mixed methods study. Br J Clin Pharmacol 2022 Feb 29;88(4):1866-1884 [FREE Full text] [doi: 10.1111/bcp.15107] [Medline: 34625991]
8. Rodgers S, Taylor AC, Roberts SA, Allen T, Ashcroft DM, Barrett J, et al. Scaling-up a pharmacist-led information technology intervention (PINCER) to reduce hazardous prescribing in general practices: multiple interrupted time series study. PLoS Med 2022 Nov 16;19(11):e1004133 [FREE Full text] [doi: 10.1371/journal.pmed.1004133] [Medline: 36383560]
9. Ivers N, Jamtvedt G, Flottorp S, Young JM, Odgaard-Jensen J, French SD, et al. Audit and feedback: effects on professional practice and healthcare outcomes. Cochrane Database Syst Rev 2012 Jun 13;2012(6):CD000259. [doi: 10.1002/14651858.CD000259.pub3] [Medline: 22696318]
10. Rao A, Kim J, Kamineni M, Pang M, Lie W, Dreyer KJ, et al. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 versus GPT-3.5 in a breast imaging pilot. J Am Coll Radiol 2023 Oct;20(10):990-997 [FREE Full text] [doi: 10.1016/j.jacr.2023.05.003] [Medline: 37356806]
11. Kobraei K, Baradaran M, Sadeghi SM, Masumshah R, Eslahchi C. ADEP: a novel approach based on discriminator-enhanced encoder-decoder architecture for accurate prediction of adverse effects in polypharmacy. arXiv. Preprint posted online on May 31, 2024 [FREE Full text]
12. Ong JC, Jin L, Elangovan K, Lim GY, Lim DY, Sng GG, et al. Development and testing of a novel large language model-based clinical decision support systems for medication safety in 12 clinical specialties. arXiv. Preprint posted online on January, 2024 [FREE Full text]
13. Tapper J. DrugGPT: new AI tool could help doctors prescribe medicine in England. The Guardian. 2024. URL: https://tinyurl.com/3fckmxf7 [accessed 2025-05-29]
14. Liu F, Zhou H, Zhang W, Huang G, Clifton L, Eyre D, et al. RETRACTED: DrugGPT: a knowledge-grounded collaborative large language model for evidence-based drug analysis. Research Square. Preprint posted online on October 6, 2023 [FREE Full text] [doi: 10.21203/rs.3.rs-3411728/v1]
15. Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, et al. Effect of clinical decision-support systems: a systematic review. Ann Intern Med 2012 Jul 03;157(1):29-43 [FREE Full text] [doi: 10.7326/0003-4819-157-1-201207030-00450] [Medline: 22751758]

XSL•FO
RenderX

16. Kwan JL, Lo L, Ferguson J, Goldberg H, Diaz-Martinez JP, Tomlinson G, et al. Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. BMJ 2020 Sep 17;370:m3216 [FREE Full text] [doi: 10.1136/bmj.m3216] [Medline: 32943437]
17. Kouri A, Yamada J, Lam Shin Cheung J, Van de Velde S, Gupta S. Do providers use computerized clinical decision support systems? A systematic review and meta-regression of clinical decision support uptake. Implement Sci 2022 Mar 10;17(1):21 [FREE Full text] [doi: 10.1186/s13012-022-01199-3] [Medline: 35272667]
18. Greenes RA, Bates DW, Kawamoto K, Middleton B, Osheroff J, Shahar Y. Clinical decision support models and frameworks: seeking to address research issues underlying implementation successes and failures. J Biomed Inform 2018 Feb;78:134-143 [FREE Full text] [doi: 10.1016/j.jbi.2017.12.005] [Medline: 29246790]
19. Meunier PY, Raynaud C, Guimaraes E, Gueyffier F, Letrilliart L. Barriers and facilitators to the use of clinical decision support systems in primary care: a mixed-methods systematic review. Ann Fam Med 2023 Jan 23;21(1):57-69 [FREE Full text] [doi: 10.1370/afm.2908] [Medline: 36690490]
20. Kim SY, Kim DH, Kim MJ, Ko HJ, Jeong OR. XAI-based clinical decision support systems: a systematic review. Appl Sci 2024 Jul 30;14(15):6638. [doi: 10.3390/app14156638]
21. Sirajuddin AM, Osheroff JA, Sittig DF, Chuo J, Velasco F, Collins DA. Implementation pearls from a new guidebook on improving medication use and outcomes with clinical decision support. Effective CDS is essential for addressing healthcare performance improvement imperatives. J Healthc Inf Manag 2009;23(4):38-45 [FREE Full text] [Medline: 19894486]
22. Kilsdonk E, Peute LW, Jaspers MW. Factors influencing implementation success of guideline-based clinical decision support systems: a systematic review and gaps analysis. Int J Med Inform 2017 Feb;98:56-64. [doi: 10.1016/j.ijmedinf.2016.12.001] [Medline: 28034413]
23. Van de Velde S, Heselmans A, Delvaux N, Brandt L, Marco-Ruiz L, Spitaels D, et al. A systematic review of trials evaluating success factors of interventions with computerised clinical decision support. Implement Sci 2018 Aug 20;13(1):114 [FREE Full text] [doi: 10.1186/s13012-018-0790-1] [Medline: 30126421]
24. Martínez-García A, Moreno-Conde A, Jódar-Sánchez F, Leal S, Parra C. Sharing clinical decisions for multimorbidity case management using social network and open-source tools. J Biomed Inform 2013 Dec;46(6):977-984 [FREE Full text] [doi: 10.1016/j.jbi.2013.06.007] [Medline: 23806275]
25. Fraccaro P, Arguello Casteleiro M, Ainsworth J, Buchan I. Adoption of clinical decision support in multimorbidity: a systematic review. JMIR Med Inform 2015 Jan 07;3(1):e4 [FREE Full text] [doi: 10.2196/medinform.3503] [Medline: 25785897]
26. Fahmi A, Wong D, Walker L, Buchan I, Pirmohamed M, Sharma A, et al. Combinations of medicines in patients with polypharmacy aged 65-100 in primary care: large variability in risks of adverse drug related and emergency hospital admissions. PLoS One 2023;18(2):e0281466 [FREE Full text] [doi: 10.1371/journal.pone.0281466] [Medline: 36753492]
27. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
28. MAGUIRE M. Methods to support human-centred design. Int J Hum Comput Stud 2001 Oct;55(4):587-634. [doi: 10.1006/ijhc.2001.0503]
29. van Staa T, Sharma A, Palin V, Fahmi A, Cant H, Zhong X, et al. Knowledge support for optimising antibiotic prescribing for common infections in general practices: evaluation of the effectiveness of periodic feedback, decision support during consultations and peer comparisons in a cluster randomised trial (BRIT2) - study protocol. BMJ Open 2023 Aug 22;13(8):e076296 [FREE Full text] [doi: 10.1136/bmjopen-2023-076296] [Medline: 37607793]
30. Yuan MJ, Finley GM, Long J, Mills C, Johnson RK. Evaluation of user interface and workflow design of a bedside nursing clinical decision support system. Interact J Med Res 2013 Jan 31;2(1):e4 [FREE Full text] [doi: 10.2196/ijmr.2402] [Medline: 23612350]
31. Blanes-Selva V, Asensio-Cuesta S, Doñate-Martínez A, Pereira Mesquita F, García-Gómez JM. User-centred design of a clinical decision support system for palliative care: insights from healthcare professionals. Digit Health 2023;9:20552076221150735 [FREE Full text] [doi: 10.1177/20552076221150735] [Medline: 36644661]
32. Larsen K, Akindele B, Head H, Evans R, Mehta P, Hlatky Q, et al. Developing a user-centered digital clinical decision support app for evidence-based medication recommendations for type 2 diabetes mellitus: prototype user testing and validation study. JMIR Hum Factors 2022 Jan 18;9(1):e33470 [FREE Full text] [doi: 10.2196/33470] [Medline: 34784293]
33. Hurley R, Jury F, van Staa TP, Palin V, Armitage CJ. Clinician acceptability of an antibiotic prescribing knowledge support system for primary care: a mixed-method evaluation of features and context. BMC Health Serv Res 2023 Apr 14;23(1):367 [FREE Full text] [doi: 10.1186/s12913-023-09239-4] [Medline: 37060063]
34. Rydberg EM, Insulan J, Rolfson O, Mohaddes M, Ahlstrom L. Knowledge support for ankle fractures in the Swedish Fracture Register - a qualitative study of physicians' experiences. BMC Health Serv Res 2022 Mar 23;22(1):382 [FREE Full text] [doi: 10.1186/s12913-022-07799-5] [Medline: 35321701]
35. Damschroder LJ, Reardon CM, Widerquist MA, Lowery J. The updated Consolidated Framework for implementation research based on user feedback. Implement Sci 2022 Oct 29;17(1):75 [FREE Full text] [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
COREQ: Consolidated Criteria for Reporting Qualitative Research
DynAIRx: artificial intelligence for dynamic prescribing optimization and care integration in multimorbidity
EPR: electronic patient record
FG: focus group
GP: general practitioner
HCP: health care professional
LLM: large language model
ML: machine learning
NICE: National Institute for Health and Care Excellence
PCN: primary care network
SMR: structured medication review

Edited by J Sarvestan; submitted 25.Mar.2025; peer-reviewed by O Kehinde, T Ojo, MJ Ugbor, E Oluwagbade; comments to author 07.May.2025; revised version received 10.Jun.2025; accepted 11.Jun.2025; published 08.Jan.2026.

Please cite as:
Abuzour AS, Wilson SA, Woodall AA, Mair FS, Aslam A, Clegg A, Shantsila E, Gabbay M, Abaho M, Bollegala D, Cant H, Griffiths A, Hama L, Leeming G, Lo E, Maskell S, O'Connell M, Popoola O, Relton S, Ruddle RA, Schofield P, Sperrin M, Van Staa T, Buchan I, Walker LE
Developing an AI-Assisted Tool That Identifies Patients With Multimorbidity and Complex Polypharmacy to Improve the Process of Medication Reviews: Qualitative Interview and Focus Group Study
J Med Internet Res 2026;28:e74304
URL: https://www.jmir.org/2026/1/e74304
doi:10.2196/74304
PMID:

©Aseel S Abuzour, Samantha A Wilson, Alan A Woodall, Frances S Mair, Asra Aslam, Andrew Clegg, Eduard Shantsila, Mark Gabbay, Michael Abaho, Danushka Bollegala, Harriet Cant, Alan Griffiths, Layik Hama, Gary Leeming, Emma Lo, Simon Maskell, Maurice O'Connell, Olusegun Popoola, Sam Relton, Roy A Ruddle, Pieta Schofield, Matthew Sperrin, Tjeerd Van Staa, Iain Buchan, Lauren E Walker. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating and Validating Large Language Models for Health Education on Developmental Dysplasia of the Hip: 2-Phase Study With Expert Ratings and a Pilot Randomized Controlled Trial

Hui Ouyang[1,2*], MSN; Gan Lin[1,2*], MSN; Yiyuan Li[1,2], MSN; Zhixin Yao[1,2], MSN; Yating Li[1], MSN; Han Yan[1], BSN; Fang Qin[2*], PhD; Jinghui Yao[3*], PhD; Yun Chen[1,2*], MM

[1]Third Affiliated Hospital of Southern Medical University, Guangzhou, China

[2]School of Nursing, Southern Medical University, Guangzhou, China

[3]Department of Pediatric Orthopedics, Center for Orthopaedic Surgery, Third Affiliated Hospital of Southern Medical University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Yun Chen, MM
Third Affiliated Hospital of Southern Medical University
183 Zhongshan Road West
Tianhe District
Guangzhou, 510630
China
Phone: 86 13724869220
Email: chenyun88@smu.edu.cn

Abstract

Background: Developmental dysplasia of the hip (DDH) is a common pediatric orthopedic disease, and health education is vital to disease management and rehabilitation. The emergence of large language models (LLMs) has provided new opportunities for health education. However, the effectiveness and applicability of LLMs in education with DDH have not been systematically evaluated.

Objective: This study conducted an integrated 2-phase evaluation to assess the quality and educational effectiveness of LLM-generated educational materials.

Methods: This study comprised 2 phases. Based on Bloom's taxonomy, a 16-item DDH question bank was created through literature analysis and collaboration. Four LLMs (ChatGPT-4 [OpenAI], DeepSeek-V3, Gemini 2.0 Flash [Google], and Copilot [Microsoft Corp]) were questioned using standardized prompts. All responses were independently evaluated by 5 pediatric orthopedic experts using 5-point Likert measures of accuracy, fluency, and richness, the scales of Patient Education Materials Assessment Tool for Printable Materials, and DISCERN. The readability was measured by a formula. The data were examined using Kruskal-Wallis tests, ANOVA, and post hoc comparisons. In phase 2, an assessor-blinded, 2-arm pilot randomized controlled trial was conducted. A total of 127 caregivers were randomized into an LLM-assisted education group or a web search control group. The intervention included structured LLM training, supervised practice, and 2 weeks of reinforcement training. Measured at baseline, postintervention, and 2 weeks following, the outcomes were eHealth literacy (primary), DDH knowledge, health risk perception, perceived usefulness, information self-efficacy, and health information-seeking behavior. Cohen *d* effect sizes and linear mixed-effects models were used in an intention-to-treat manner.

Results: There were significant differences between the 4 LLMs concerning accuracy, richness, fluency, Patient Education Materials Assessment Tool for Printable Materials Understandability, and DISCERN ($P<.05$). ChatGPT-4 (median 63.67, IQR 63.67-64.67) and DeepSeek-V3 (median 63.67, IQR 63.33-64.67) generate more accurate text than Copilot (median 59.00, IQR 58.67-59.67). DeepSeek-V3 (median 64.00, IQR 64.00-64.00) was language richer than Copilot (median 52.33, IQR 51.33-52.67). Gemini 2.0 Flash (median 72.67, IQR 72.33-73.00) was more fluent than Copilot (median 65.67, IQR 63.33-65.67). In phase 2, the intervention group showed higher eHealth literacy at T1 (33.62, 95% CI 32.76-34.49; d=0.20, 95% CI 0.13-0.56) and T2 (33.27, 95% CI 32.38-34.17; d=0.36, 95% CI 0.01-0.80), greater DDH knowledge at T1 (7.87, 95% CI 7.48-8.25, d=0.71, 95% CI 0.33-1.11) and T2 (7.12, 95% CI 6.72-7.51; d=0.54, 95% CI 0.17-0.96), and slight improvements in health risk prediction and perceived usefulness.

XSL•FO
RenderX

Conclusions: Mainstream LLMs demonstrate varying capacities in generating educational content for DDH. They generated DDH caregiver education materials that were associated with modest improvements in eHealth literacy and knowledge. Although LLMs can address general informational needs, they cannot completely substitute clinical evaluation. Future research should focus on optimizing plain language, refining dialogue design, and enhancing audience personalization to improve the quality of LLMs' materials.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2500108410; https://www.chictr.org.cn/showproj.html?proj=271987

(J Med Internet Res 2026;28:e73326) doi:10.2196/73326

KEYWORDS

large language models; developmental dysplasia of the hip; health education; content generation; mobile phone

Introduction

Background

Developmental dysplasia of the hip (DDH) is a common pediatric orthopedic condition affecting 1%-3% of infants, with a higher prevalence in girls and more frequent involvement of the left hip [1]. If undiagnosed or untreated early, DDH can lead to gait abnormalities, chronic pain, and early osteoarthritis, substantially affecting the quality of life [2]. Early diagnosis and health education are critical for improving prognosis. Delayed diagnosis and treatment often require complex surgery, which not only increases the difficulty of treatment but may also result in further functional deterioration [3,4]. Traditional educational methods are limited by time and resources, making it difficult to meet patients' diverse informational needs. The emergence of artificial intelligence (AI) has provided new opportunities for health education.

In the broader field of digital health communication, AI-based conversational systems are increasingly being explored as tools to provide convenient and efficient services to meet people's diverse needs. Currently, large language models (LLMs), such as ChatGPT, Google Gemini, Microsoft Copilot, and DeepSeek, are applied in health communication, including disease diagnosis [5], treatment recommendation [6], health education [7], and clinical decision-making [8]. For example, ChatGPT enables interactive discussions that tailor standardized medical information to individual patient needs, helping bridge communication gaps between clinicians and patients [9,10]. Although AI has demonstrated great potential in medical education, its use in patient-facing communication raises concerns. LLMs may provide erroneous medical advice [11], propagate outdated medical views [12], or fabricate nonexistent medical cases to generate "hallucinations" [13]. At the ethical and regulatory levels, challenges arise from the model's "black box" decision-making, including unclear accountability, difficulties in defining legal responsibility, privacy breaches, and lagging regulatory frameworks. These issues directly jeopardize users' safety, potentially leading to misdiagnosis, delayed treatment, and other forms of direct harm. Furthermore, most generated content maintains a university reading level, which may pose comprehension challenges for users without higher education [14]. These risks underscore the need for systematic evaluation before integrating such tools into health education.

While prior studies have primarily examined the accuracy or readability of LLM-generated content [15], few have connected content quality with its actual educational impact on end users. The extent to which LLM-generated materials can effectively support caregivers' understanding and health literacy in specific conditions, such as DDH, remains unexplored. In DDH, caregivers have to not only comprehend specialized medical concepts but also actively recognize abnormal signs in children and make timely decisions [16]. Due to the professional complexity of orthopedic knowledge and the unique nature of pediatric disorders, basic health literacy abilities are necessary for caregivers. Therefore, the different levels of digital literacy among caregivers may make it more difficult for them to properly understand information produced by LLMs. To address this, the present study systematically evaluated multiple mainstream LLMs through expert assessment and a pilot randomized controlled trial (RCT) among caregivers. By integrating expert evaluation with caregiver validation, this study extends the current AI in health communication research from theoretical assessment to empirical verification.

Objective

Therefore, this study aimed to provide a comprehensive evaluation and verification of LLM-generated education materials for DDH. The first phase assessed the educational quality of the outputs generated by 4 mainstream LLMs (ChatGPT-4, DeepSeek-V3, Gemini 2.0 Flash, and Copilot) through expert ratings of accuracy, understandability, actionability, and readability. The second phase involved a pilot RCT among caregivers to evaluate the actual educational impact of these materials, including digital literacy, DDH knowledge acquisition, health risk perception, information self-efficacy, perceived usefulness, and health information-seeking behaviors. This study bridged the gap by integrating the quality assessment of LLMs with RCT to validate their content reliability and educational impact. It offered evidence for the safe and effective use of LLMs in clinical education.

Methods

Theoretical Framework

The taxonomy by Bloom et al [17] served as the guiding pedagogical framework for designing the educational content. The taxonomy organizes cognitive processes into 6 hierarchical levels: remember, understand, apply, analyze, evaluate, and create, which is widely used to structure learning objectives and instructional materials. As users often need to acquire not only

XSL•FO
RenderX

basic factual knowledge but also practical decision-making skills, this hierarchical model provided a structured approach for determining which levels of cognition should be targeted in education [18].

Guided by this framework, we developed a 16-item question bank that intentionally spanned different cognitive levels, ranging from foundational knowledge such as definitions and symptoms to more complex tasks such as interpreting clinical scenarios or making care decisions (Table 1). This ensured that the LLM-generated responses covered the breadth of learning needs relevant to caregivers. Bloom's taxonomy, therefore, supported the construction of a comprehensive and pedagogically meaningful learning set, helping align the generated content with education requirements.

Table 1. Question bank framework based on Bloom's taxonomy.

Bloom's taxonomy and part	Content
Remembering	
Basics	"What is [disease name]? Please explain its main features and potential effects in simple language."
Anatomy and effects	"Who is most likely to develop [disease name]? How common is it?"
Understanding	
Etiology and risk factors	"Will [disease name] have a long-term impact on my quality of life? What aspects should I pay special attention to?"
Symptoms and early recognition	"What parts of the body does [disease name] mainly affect? What do these parts do in a healthy state?"
Applying	
Physician examination and diagnosis	"What symptoms or consequences may arise if these parts are damaged? Can you provide specific examples?"
Emergencies	"What are the main causes of [disease name]? What risk factors may increase the chances of developing it?"
Analyzing	
Hospital treatment and rehabilitation	"What lifestyle habits or environmental factors increase the risk of developing [disease name]? How can I prevent it?"
Medication management	"If I have a family history, is my risk of developing [disease name] higher? What preventive measures should I take?"
Pain management	"What are the main symptoms of [disease name]? What signs indicate the disease is worsening? Should I seek immediate medical attention?"
Evaluating	
Postoperative management	"What methods do doctors usually use to confirm the diagnosis of [disease name]? What is the purpose of each test?"
Creating	
Daily living and health management	"What indicators are most important in test results? How do these results reflect the severity of the disease?"
Mental health	"After confirming the diagnosis of [disease name], do I need to schedule regular follow-ups? How often should I have them and what should be checked?"
Digital tools and management	"What symptoms indicate that I need to seek immediate medical attention?"
Expanding[a]	
Education and support resources	"How should I handle and self-monitor during an emergency situation?"
Social and economic impact	"What treatments will I receive during hospitalization? What is the purpose of each treatment?"
Future planning and visioning	"How can I prevent infections or other complications during hospitalization? What can family members do to assist with recovery?"

[a]Not part of Bloom's taxonomy; it is an extension of this study.

Study Design

The evaluation study had 2 phases. Phase 1 was a cross-sectional study in which physicians evaluated the answers provided by the LLMs. Phase 2 was a 2-arm pilot RCT comparing health education using LLMs with web-based searches.

Phase 1: Expert Evaluation Study

Model Testing

Based on Bloom's taxonomy, we collected and categorized common questions regarding DDH. We also reviewed clinical guidelines [19-23] to identify the key areas of knowledge. Using this information, an initial question bank was developed, which

XSL•FO
RenderX

was subsequently refined and finalized through expert review. Each question was guided by a harmonized prompt paragraph: "Using developmental dysplasia of the hip (DDH) in children as an example, answer the following questions in detail, ensuring the content is easily understandable for non-medical professionals. Life-like examples and situations can be incorporated to help readers better grasp the information. Please reduce the number of syllables to make the sentence simpler." All the generated texts and the complete question bank are provided in Multimedia Appendix 1. Each model generated educational materials based on 16 question banks, and the experiment was repeated 3 times for a total of 192 generations (4 models × 16 question banks × 3 times). Data were collected from January to February 2025. ChatGPT-4, DeepSeek-V3, Gemini 2.0 Flash, and Copilot were evaluated; no experimental, beta, or preview releases were included. The experiments were performed under the default settings of the web interfaces without modifying the generation parameters. To ensure reproducibility and independence of the outputs, each prompt was regenerated 3 times by establishing a new session for each run with the same original prompt. All outputs, including identical or similar responses, were retained to reflect the intrinsic variability of the models.

Assessment of Quality and Readability

Quality assessment tools as primary outcomes included (1) a Likert scale for assessing three items of accuracy, fluency, and richness of the material, scoring 16 questions on a scale from 1 to 5, with higher scores indicating better performance; (2) the DISCERN tool [24], which assessed the overall quality of the educational material, with a total of 16 entries, scoring on a scale from 1 to 5, with higher scores indicating better quality; and (3) the Patient Education Materials Assessment Tool for Printable Materials (PEMAT-P) [25], which contains 17 items measuring understandability and 7 items assessing actionability. These were reduced to 10 and 4 to accommodate the textual output with a 70% passing line based on the guidelines. During the evaluation process, each material was independently scored by 5 evaluators. The scores from evaluators were retained for subsequent data analysis.

Readability assessment tools as secondary outcomes included (1) the Flesch-Kincaid Reading Ease (FKRE), (2) the Flesch-Kincaid Grade Level (FKGL), and (3) the Simple Measure of Gobbledygook (SMOG) index, chosen for their widespread use and reliability in assessing text readability. All 3 score calculations involved the total number of words, sentences, and syllables. The FKRE measured the simplicity of the text, with scores ranging from 0 to 100, with higher scores indicating better readability. The FKGL represented reading level grade, with lower FKGL and SMOG indicating better comprehension and higher scores indicating more complex language. Scores above 60 or below sixth grade were the recommended reading levels for the general public. Readability scores were calculated using a web-based readability calculator (Readable; Added Bytes). The detailed formulas are provided in Multimedia Appendix 2.

Expert Evaluation

The material generated by the LLMs was independently assessed for quality. The material generated by the LLMs was independently assessed by 5 pediatric orthopedic physicians with expertise in DDH, selected through rigorous predefined criteria: (1) ≥10 years of clinical experience in DDH diagnosis or treatment; and (2) evaluators completed standardized training on the assessment rubric before this study, using the DDH guidelines as the gold standard [19-23]. To ensure blinding, the LLM outputs were made anonymous by an independent researcher who replaced the model names with random codes. The evaluators confirmed that they could not infer the identities of the LLMs or determine if repeated outputs came from the same model. Interrater reliability was assessed for each outcome dimension using the intraclass correlation coefficient (ICC). ICC values were interpreted as follows: <0.5=poor, 0.5-0.75=moderate, 0.75-0.9=good, and >0.9=excellent agreement.

Phase 2: Pilot RCT

Participants

Participants were recruited through digital media advertisements and physician referrals. Eligibility criteria included (1) being aged ≥18 years, (2) being caregivers of children aged 0-14 years, (3) having the ability to read and understand words, and (4) having internet access. Exclusion criteria included (1) having severe hearing or visual impairment; (2) having severe schizophrenia, major depression, bipolar disorder, and other mental illnesses; or (3) participation in other related studies.

Sample Size

Power analysis was performed using G*Power 3.1.9.7 based on similar educational intervention studies [26]. A medium effect size (Cohen d=0.65) was anticipated for the primary outcome of performance, with 2-tailed α=.05, power of 0.8, and at least 38 participants per group being required. Accounting for an expected 20% attrition rate, the target sample was 49 participants per group (total n=98). There were 127 participants in the final sample (62 in the control group and 65 in the intervention group).

Randomization and Blinding

Recruitment took place in the Third Affiliated Hospital of Southern Medical University and community support groups. The researchers generated a computer-generated list and sealed it in an opaque envelope. Before the start of the intervention, research assistants who were not involved in hospital assessments or interventions opened the envelopes and assigned participants at random to the intervention or control group. Following informed consent, eligible participants meeting the inclusion and exclusion criteria were randomly assigned in a 1:1 ratio to either the trial or control group. The blinding of participants was not feasible due to the nature of the intervention, but the research team remained unaware of group assignments until this study concluded. Data analysts who conducted the final analyses were masked to participant identities throughout. Due to the nature of the intervention, participant blinding was not possible. However, group allocations were not disclosed to the research team until the

XSL•FO
RenderX

trial was finished. Throughout, participant identities were concealed from the data analysts.

Control Group

Participants in the control group received standard web-based educational materials prepared by clinical experts. These materials were retrieved from official sources (eg, [27]). Participants were asked to read independently, simulating a typical web-based health information-seeking behavior.

Intervention Group

All researchers received standardized training to ensure consistent delivery of DDH-related information and LLM education. The intervention was delivered to participants by face-to-face communication. First, the participants were introduced to the foundational concepts of LLMs, including basic mechanisms, application categories, and core interaction capabilities. Second, a standardized consultation framework was introduced, covering device access, platform login, dialogue initiation, and structured prompt formulation. The required background information included demographic and clinical characteristics, symptom description, disease duration, medical history, lifestyle, and psychosocial factors. Participants were also provided with 16 DDH-related inquiry categories, including foundations, risk factors, early recognition, diagnosis, treatment, postoperative care, medication management, psychological support, etc. They are able to optionally output custom instructions, such as length, style, level of technical terminology, and formatting preferences. Third, strategies to improve information quality are introduced, including clear language prompts, staged questions, example guidance, support for the reasoning process, evidence sources for web retrieval, and document import. Verification approaches were emphasized, such as cross-model comparison, guideline checking, and professional consultation. Finally, risk awareness and ethical considerations were reinforced, including potential hallucinations, outdated content, privacy risks, copyright issues, and inappropriate clinical dependence. A practical demonstration was conducted using an actual DDH case. For example, a female infant, 1 year old, with asymmetric thigh folds and a family history, but no medical history. Participants inquired and learned relevant knowledge based on the background of this example. During the 2 weeks, the participants received remote support through web-based group consultations or offline feedback sessions. Researchers responded to questions related to practical application, corrected misuse behaviors, and supplemented individualized guidance.

Data Collection

Data were collected through questionnaire surveys. The basic information questionnaire gathered the demographic characteristics of this study's participants. Validated scales were used to measure eHealth literacy, DDH knowledge, health risk perception, information self-efficacy, perceived usefulness, and health information-seeking behavior. There were three assessment time points: (1) baseline (T0), (2) immediately after the completion of the intervention or control group (T1), and (3) two weeks after the end of the intervention or control group (T2).

Primary Outcomes

The eHealth Literacy Scale (eHEALS), originally developed by Norman and Skinner [28], was adopted to measure participants' eHealth literacy. It comprises 8 items that assess one's ability to locate and use web-based health resources, appraise the credibility of digital health information, and apply acquired information to make informed health decisions. Each item is scored on a 5-point Likert scale, producing a total score between 8 and 40, with higher scores representing stronger eHealth literacy.

Secondary Outcomes

The developmental dysplasia of the hip knowledge test (DDH-KT) was developed by the research team to assess participants' basic DDH knowledge. The items were constructed according to current clinical guidelines and health education materials and reviewed by pediatric orthopedic surgeons. Each correct answer is scored as 1 point (range 0-10), with higher scores indicating greater DDH knowledge. The full knowledge test is provided in Multimedia Appendix 3.

The Health Risk Perception Scale (HRPS) was measured based on the framework by Ajzen [29]. The scale was adapted from established health risk perception measures by Brewer et al [30], and covered 2 dimensions: perceived susceptibility and perceived severity. The items assessed participants' subjective perception of the likelihood and potential consequences of related health problems, rated on a 5-point Likert scale. Higher scores reflected a greater level of perceived risk (Cronbach α=0.847).

The Information Self-Efficacy Scale (ISES), adapted from Pavlou and Fygenson [31], was used to evaluate participants' confidence in obtaining and effectively using web-based health information. The scale contained 3 items rated on a 5-point Likert scale. Total scores were calculated by summing all item responses, with higher scores indicating stronger information self-efficacy (Cronbach α=0.806).

The Perceived Usefulness Scale (PUS), adapted from Cheung et al [32], assessed the extent to which participants viewed web-based health information as helpful, relevant, and beneficial for health knowledge and decision-making. Items were scored on a 5-point Likert scale, with higher scores indicating greater perceived usefulness (Cronbach α=0.852).

The Health Information-Seeking Behavior Scale (HISBS), adapted from Kankanhalli et al [33], measured the frequency and willingness to actively seek web-based health information. Responses were recorded using a 5-point Likert scale, and higher scores indicated more proactive seeking behavior (Cronbach α=0.873).

Statistical Analysis

In phase 1, descriptive statistics were reported as mean (SD) and median (IQR). Because the final analytic values were obtained by averaging 3 generations, the normality assumptions for repeated-measures ANOVA were not met. Group differences among the 4 LLMs were analyzed using the Kruskal-Wallis H test, followed by Dunn-Bonferroni post hoc comparisons when significant. One-way ANOVA and Tukey post hoc tests were

XSL•FO
RenderX

used for readability indices because the normality assumptions were satisfied. False discovery rate correction was applied across the 9 outcomes to control for multiple testing. Interrater reliability was assessed using ICC(2,k) based on a 2-way random-effects model [34]. Effect sizes were reported as epsilon-squared for nonparametric tests and eta-squared for ANOVA. Analyses were conducted in R (version 4.5.1; R Foundation) with ggplot2 (version 3.5.1; Posit, PBC) for visualization.

In phase 2, all analyses followed the intention-to-treat principle and included all randomized participants. Continuous baseline variables are presented as mean (SD), and categorical variables as counts and percentages. Differences between groups at baseline were assessed using 2-sided independent sample *t* tests for continuous variables and chi-square tests for categorical variables. Outcomes were analyzed using linear mixed-effects models with time (T1 and T2) and group (intervention vs control) as fixed effects, time × group interaction, baseline (T0) as a covariate, and participant ID as a random intercept. No imputation was performed because linear mixed-effects models estimated with restricted maximum likelihood provided unbiased estimates under the missing at random assumption [35]. Between-group effect sizes (Cohen *d*, 95% CI) and estimated marginal means (95% CI) were reported. eHEALS was defined as the primary outcome. All other outcomes, including DDH-KT, HRPS, ISES, PUS, and HISBS, were considered secondary. Given the pilot and exploratory nature of this trial, no adjustment for multiple comparisons was applied. Therefore, analyses of the outcomes were intended to be hypothesis-generating rather than confirmatory. Analyses were conducted in R (version 4.5.1) using lme4, lmerTest, and emmeans; 2-sided $P<.05$ was considered statistically significant.

Ethical Considerations

This study was approved by the Ethics Committee of the Third Affiliated Hospital of Southern Medical University (2024-ER-113), and the first participant was enrolled in June 2025. The trial registration was completed on August 29, 2025, at the Chinese Clinical Trial Registry (ChiCTR2500108410). All research participants signed written informed consent forms. Researchers disclosed study information to participants; participants retained the right to withdraw from the study or withdraw their research data at any time without conditions, and withdrawal would not result in any adverse consequences. Participants were informed that part of the educational content was generated by AI, and the limitations of AI-generated information were explained. The use of AI-assisted materials was supervised throughout this study by qualified health care professionals. During the intervention period, participants were encouraged to report any concerns or adverse experiences related to the educational materials, and ultimately, no related adverse events were reported. All personal information and data collected during the study were kept strictly confidential. Participants who completed the entire study process received educational materials, including a parenting knowledge handbook valued at CN ¥50 RMB (approximately US $7.15), as compensation.

Results

Phase 1

Overview

Overall, ChatGPT-4 and DeepSeek-V3 demonstrated the strongest performance in content accuracy, richness, understandability, and information quality, making them suitable for generating pediatric health communication materials. Gemini 2.0 Flash and Copilot performed well in fluency and readability metrics, while they were relatively weaker in content richness and accuracy. Table 2 provides a visual summary of the scores and the overall performance comparison. Figure 1 illustrates the comparison of the responses across the LLMs. The scoring data are presented in Multimedia Appendix 4.

XSL•FO
RenderX

Table 2. Comparison of model performance across different indicators.

Model	Mean (SD)	Median (IQR)	H/F[a]	*P* value	FDR-adjusted[b] *P* value	ε^2[c]/η^2[d]	Significance (*P* value)
Accuracy			13.873	.003	.005	0.73	
ChatGPT-4	64 (1.03)	63.67 (63.67-64.67)					*[e,f] (.02)
Copilot	59.07 (1.01)	59.00 (58.67-59.67)					*[g] (.048)
DeepSeek-V3	63.33 (1.78)	63.67 (63.33-64.67)					__[h]
Gemini 2.0 Flash	59.53 (0.99)	59.67 (59.33-60.00)					—
Richness			13.68	.003	.005	0.72	
ChatGPT-4	62.33 (2.26)	62.67 (60.00-64.33)					—
Copilot	51.6 (3.52)	52.33 (51.33-52.67)					*[g] (.02)
DeepSeek-V3	63.93 (1.53)	64.00 (64.00-64.00)					—
Gemini 2.0 Flash	54 (5.65)	54.67 (50.33-57.33)					—
Fluency			16.204	.001	.003	0.853	
ChatGPT-4	69.53 (1.19)	70.00 (68.67-70.00)					—
Copilot	64.87 (1.8)	65.67 (63.33-65.67)					***[i] (<.001)
DeepSeek-V3	69.87 (2.19)	70.67 (70.33-71.00)					—
Gemini 2.0 Flash	73 (0.97)	72.67 (72.33-73.00)					—
PEMAT-P[j] understandability (%)			11.421	.01	.012	0.601	
ChatGPT-4	93.89 (1.24)	94.44 (94.44-94.44)					*[f] (.03)
Copilot	85 (4.21)	86.11 (80.56-88.89)					—
DeepSeek-V3	93.33 (3.17)	94.44 (91.67-94.44)					—
Gemini 2.0 Flash	87.78 (4.21)	86.11 (86.11-88.89)					—
PEMAT-P actionability (%)			7.587	.06	.06	0.399	
ChatGPT-4	68.33 (3.73)	66.67 (66.67-66.67)					—
Copilot	60 (6.97)	58.33 (58.33-66.67)					—
DeepSeek-V3	68.33 (3.73)	66.67 (66.67-66.67)					—
Gemini 2.0 Flash	66.67 (5.89)	66.67 (66.67-66.67)					—
DISCERN			10.243	.02	.02	0.539	
ChatGPT-4	48.52 (3.27)	49.00 (46.00-49.27)					*[i] (.035)
Copilot	46.56 (1.57)	46.67 (46.47-47.67)					—
DeepSeek-V3	48.44 (2.71)	48.00 (46.67-49.20)					*[i] (.03)
Gemini 2.0 Flash	43.08 (1.82)	43.33 (42.33-43.40)					—
FKGL[k]			8.395	<.001	.003	0.296	
ChatGPT-4	8.74 (1.37)	8.86 (8.14-9.70)					*[g] (.03); *[i] (.04)
Copilot	9.41 (1.86)	9.07 (8.38-10.88)					***[g] (<.001); ***[i] (<.001)
DeepSeek-V3	7.30 (1.08)	7.26 (6.72-7.70)					—
Gemini 2.0 Flash	7.37 (1.33)	7.19 (6.30-8.44)					—
FKRE[l]			14.198	.003	.005	0.225	
ChatGPT-4	61.86 (8.58)	61.44 (56.80-66.97)					—

XSL•FO
RenderX

Model	Mean (SD)	Median (IQR)	H/F[a]	*P* value	FDR-adjusted[b] *P* value	ε^2[c]/η^2[d]	Significance (*P* value)
Copilot	53.45 (12.62)	57.70 (46.25-61.33)					**[g] (.006); **[i] (.009)
DeepSeek-V3	67.19 (6.62)	67.45 (62.73-70.43)					—
Gemini 2.0 Flash	66.85 (7.72)	70.10 (59.19-73.48)					—
SMOG[m]			8.297	<.001	.003	0.293	
ChatGPT-4	11.02 (1.26)	10.96 (10.52-11.61)					*[g] (.02)
Copilot	11.67 (1.33)	11.48 (11.09-13.04)					***[g] (<.001); **[i] (.003)
DeepSeek-V3	9.83 (0.93)	9.74 (9.25-10.19)					—
Gemini 2.0 Flash	10.19 (1.06)	10.03 (9.28-10.91)					—

[a]H/F: values are reported as test statistics. H statistics for the Kruskal-Wallis test and *F* statistics for 1-way ANOVA.

[b]FDR: false discovery rate.

[c]η^2: eta-squared.

[d]ε^2: epsilon-squared.

[e]******P* <.05. ***P* <.01. ****P* <.001. Normality was assessed for all variables. FKGL and SMOG were analyzed using 1-way ANOVA with Tukey honestly significant difference for pairwise comparisons; others were analyzed using the Kruskal-Wallis test with Dunn test (Bonferroni-corrected *P* values).

[f]vs CoPilot.

[g]vs DeePseek-V3.

[h]Not applicable.

[i]vs Gemini 2.0 Flash.

[j]PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials.

[k]FKGL: Flesch-Kincaid Grade Level.

[l]FKRE: Flesch-Kincaid Reading Ease.

[m]SMOG: Simple Measure of Gobbledygook.

XSL•FO
RenderX

Figure 1. Comparison of responses across 4 LLMs. (A) Accuracy, (B) richness, (C) fluency, (D) PEMAT-P understandability, (E) PEMAT-P actionability, (F) DISCERN, (G) FKGL, (H) FKRE, and (I) SMOG index. FKGL: Flesch-Kincaid Grade Level; FKRE: Flesch-Kincaid Reading Ease; LLM: large language model; PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials; SMOG: Simple Measure of Gobbledygook.

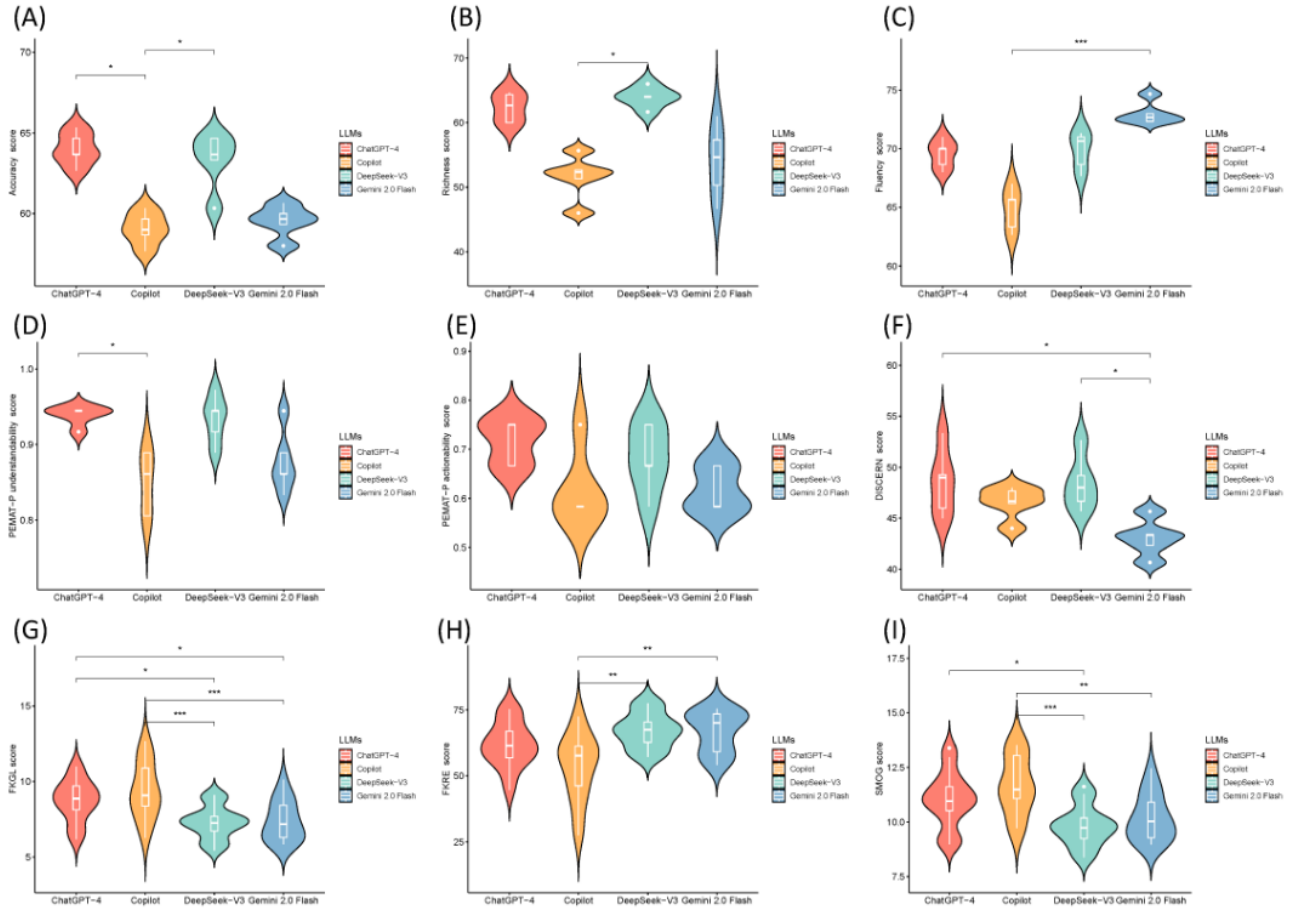

Quality Assessment

There were significant differences between the 4 LLMs in terms of accuracy, richness, fluency, PEMAT-P understandability, and DISCERN ($P<.05$). ChatGPT-4 and DeepSeek-V3 outperformed the other models in the majority of evaluation dimensions. ChatGPT-4 (median 63.67, IQR 63.67-64.67) and DeepSeek-V3 (median 63.67, IQR 63.33-64.67) generated more accurate text than Copilot (median 59.00, IQR 58.67-59.67). DeepSeek-V3 (median 64.00, IQR 64.00-64.00) was language richer than Copilot (median 52.33, IQR 51.33-52.67). Gemini 2.0 Flash (median 72.67, IQR 72.33-73.00) was more fluent than Copilot (median 65.67, IQR 63.33-65.67). Based on the PEMAT-P understandability scores, the content of ChatGPT-4 (median 94.44%, IQR 94.44%-94.44%) was more comprehensible than that of Copilot (median 86.11%, IQR 80.56%-88.89%). The PEMAT-P actionability scores were similar across the models. ChatGPT-4 (median 49.00, IQR 46.00-49.27) and DeepSeek-V3 (median 48.00, IQR 46.67-49.20) had a higher DISCERN scale score than Gemini 2.0 Flash (median 43.33, IQR 42.33-43.40).

Readability Assessment

Readability metrics highlighted the differences among the models. Gemini 2.0 Flash (median 66.85, IQR 59.19-73.48) and DeepSeek-V3 (median 67.19, IQR 62.73-70.43) generated sentences with higher FKRE scores, indicating easier readability compared to Copilot (median 53.45, IQR 46.25-61.33). DeepSeek-V3 (mean 7.30, SD 1.08) and Gemini 2.0 Flash (mean 7.37, SD 1.33) produced sentences with superior FKGL scores compared to ChatGPT-4 (mean 8.74, SD 1.37) and Copilot (mean 9.41, SD 1.86). DeepSeek-V3 (mean 9.83, SD 0.93) and Gemini 2.0 Flash (mean 10.19, SD 1.06) produced texts with better SMOG scores compared to ChatGPT-4 (mean 11.02, SD 1.26) and Copilot (mean 11.67, SD 1.33).

Visualization and Analysis

The comparative evaluation of 4 LLMs demonstrated clear performance variability across accuracy, richness, and fluency, as illustrated in Figures 2 and 3. Overall, ChatGPT-4 and DeepSeek-V3 outperformed Copilot and Gemini Flash, particularly in accuracy and fluency. In terms of accuracy, the proportion of “good” and “excellent” responses reached 85% for ChatGPT-4 and 83% for DeepSeek-V3, while Gemini 2.0 Flash (70%) and Copilot (66%) displayed a lower proportion. Regarding richness, DeepSeek-V3 (83%) and ChatGPT-4 (81%) again ranked highest, reflecting strong supplementary and explanatory capability, whereas the other 2 models showed as more concise. Across fluency, all 4 models delivered strong information elaboration, with Gemini 2.0 Flash achieving the highest proportion of 96%, indicating strong coherence, readability, and natural language expression.

As shown in the heatmap (Figure 3), ChatGPT-4 and DeepSeek-V3 yielded higher mean scores across most knowledge domains, particularly in basic, effects, and

XSL•FO
RenderX

symptoms. In contrast, Copilot and Gemini 2.0 Flash performed worse, especially in specialized domains such as medication management and postoperative care. These results suggested that current LLMs perform well in general health education content but remain limited in clinically nuanced and actionable information.

Across the 4 models and 6 evaluation dimensions, the interrater reliability among the 5 evaluators ranged from moderate to excellent (ICC=0.628-0.918). Table 3 shows the interrater reliability results across the 4 LLMs and evaluation dimensions based on ICC.

Figure 2. Expert Likert-scale ratings of content quality across 4 LLMs. LLM: large language model.

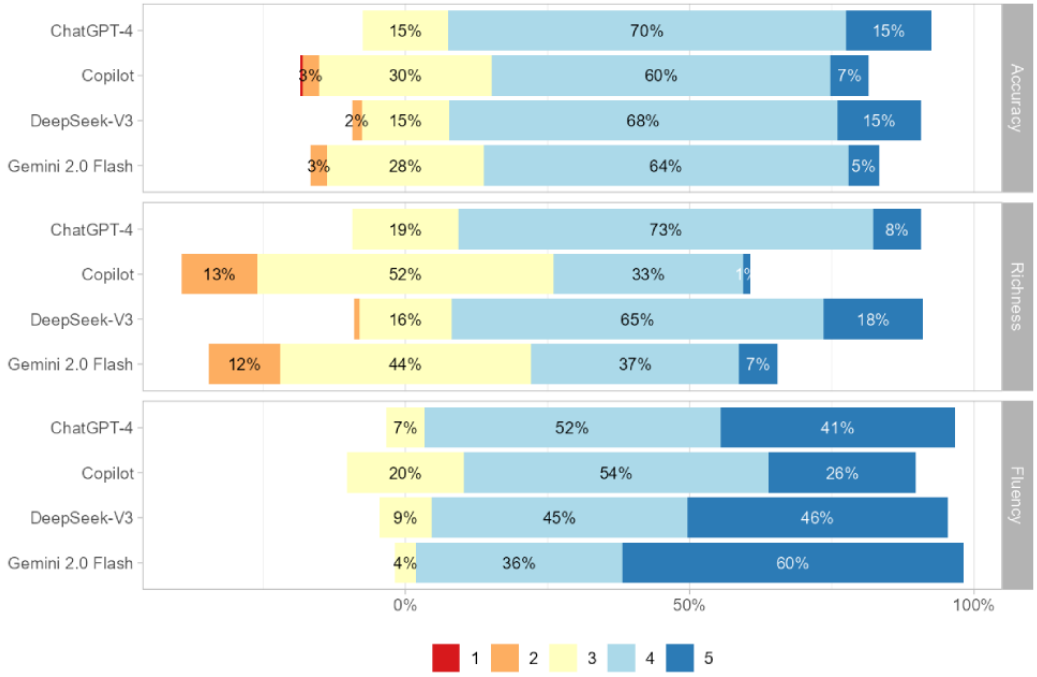

Figure 3. Multidimensional performance evaluation heatmap for LLMs. Heatmap showing mean scores of 4 LLMs across accuracy, richness, fluency, and readability dimensions. Higher scores are represented by warmer colors. FRES: Flesch-Kincaid Reading Ease; LLM: large language model.

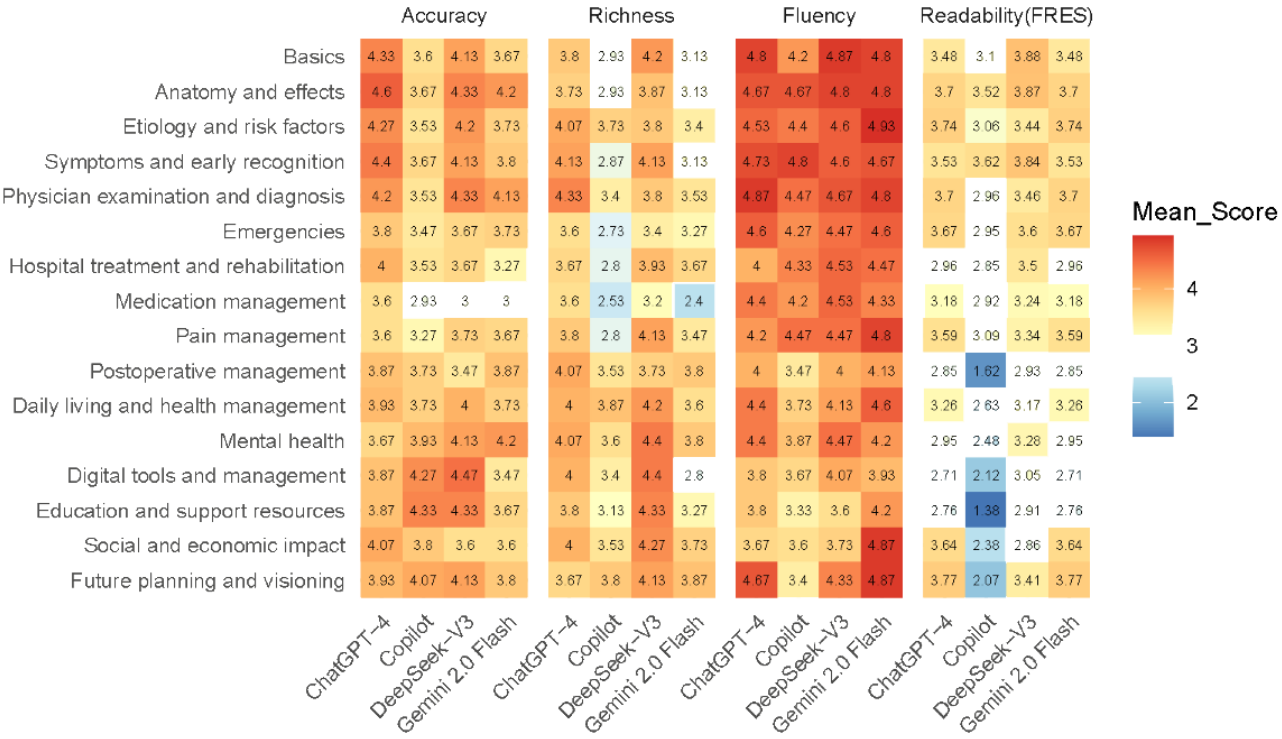

XSL•FO
RenderX

Table 3. Results of expert consistency analysis.

Dimension	ICC[a,b]	95% CI	*F* test (*df*)	*P* value	Interpretation
Accuracy					
ChatGPT-4	0.851	0.695-0.941	6.734 (15, 60)	<.001	Good
Copilot	0.654	0.277-0.864	2.810 (15, 60)	.002	Moderate
DeepSeek-V3	0.904	0.803-0.962	11.510 (15, 60)	<.001	Excellent
Gemini 2.0 Flash	0.727	0.433-0.892	3.565 (15, 60)	<.001	Moderate
Richness					
ChatGPT-4	0.669	0.352-0.864	3.499 (15, 60)	<.001	Moderate
Copilot	0.834	0.652-0.934	7.750 (15, 60)	<.001	Good
DeepSeek-V3	0.747	0.481-0.899	3.944 (15, 60)	.001	Moderate
Gemini 2.0 Flash	0.628	0.285-0.845	3.515 (15, 60)	<.001	Moderate
Fluency					
ChatGPT-4	0.914	0.825-0.966	11.931 (15, 60)	<.001	Excellent
Copilot	0.918	0.833-0.967	13.244 (15, 60)	<.001	Excellent
DeepSeek-V3	0.835	0.664-0.934	6.223 (15, 60)	<.001	Good
Gemini 2.0 Flash	0.876	0.746-0.951	8.038 (15, 60)	<.001	Good
PEMAT-P[c] understandability	0.874	0.482-0.991	7.833 (3, 12)	.004	Good
PEMAT-P actionability	0.718	0.050-0.980	3.858 (3, 12)	.038	Moderate
DISCERN	0.819	0.358-0.986	12.473 (3, 12)	.001	Good

[a]ICC: intraclass correlation coefficient.

[b]Type A intraclass correlation coefficient using an absolute agreement definition.

[c]PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials.

Phase 2

Participant Characteristics

Participants were recruited from June 2025 to September 2025. A total of 127 participants were enrolled in this study, including 65 in the intervention group and 62 in the control group. Figure 4 shows the CONSORT (Consolidated Standards of Reporting Trials) flowchart, and the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist is presented in Multimedia Appendix 5. Most participants completed the intervention, and the main reason for withdrawal was lack of time. Participants had a mean age of 36.57 (SD 6.22) years, and most were female (89/127, 70.07%) and highly educated (55/127, 43.31%). The mean age of participants' children was 5.90 (SD 3.12) years. No significant differences were observed between the intervention and control groups in the baseline characteristics ($P>.05$). During this study, no privacy breaches, technical failures, or other unintended events were observed. Table 4 summarizes the demographic characteristics of the participants. The data of participants can be found in Multimedia Appendix 6.

XSL•FO
RenderX

Figure 4. CONSORT diagram of study flow. CONSORT: Consolidated Standards of Reporting Trials.

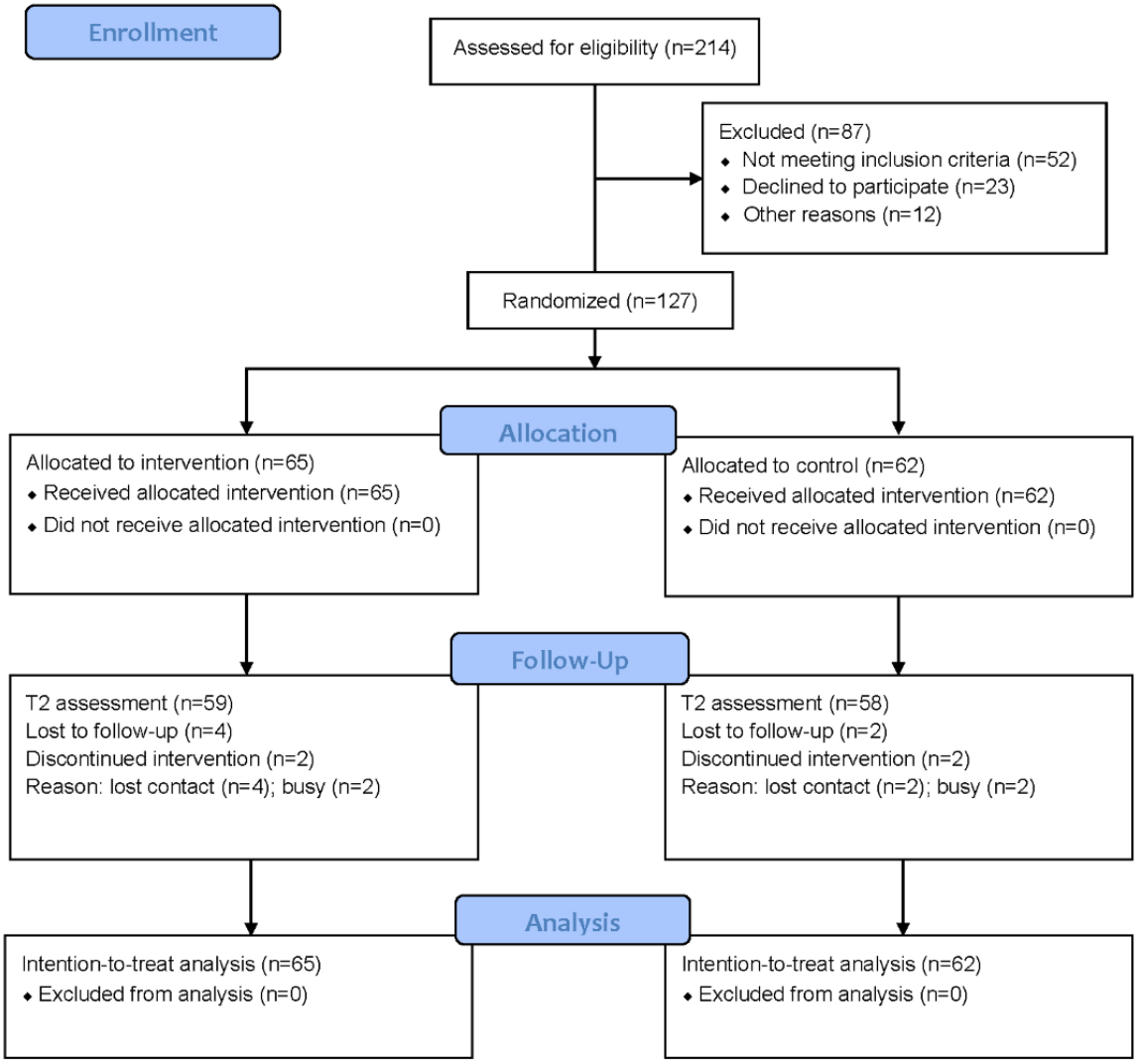

XSL•FO
RenderX

Table 4. Baseline characteristics.

Background characteristic	Overall	Intervention group (n=65)	Control group (n=62)	*P* value
Age (years), mean (SD)	36.57 (6.22)	37.06 (5.68)	36.05 (6.74)	.14
Sex, n (%)				.83
Female	89 (70.08)	45 (69.23)	44 (70.97)	
Male	38 (29.92)	20 (30.77)	18 (29.03)	
Education, n (%)				.92
Low education	18 (14.17)	11 (16.92)	7 (11.29)	
Medium education	54 (42.52)	27 (41.54)	27 (43.55)	
High education	55 (43.31)	27 (41.54)	28 (45.16)	
Monthly income (CN ¥), n (%)				.85
≤3000 (US $429.12)	9 (7.09)	30 (61.2)	35 (71.4)	
3001-6000 (US $429.26-US $858.23)	26 (20.47)	19 (38.8)	14 (28.6)	
6001-10,000 (US $858.38-US $1430.39)	42 (33.07)	21 (33.87)	21 (32.31)	
10,001-20,000 (US $1430.53-US $2860.78)	36 (28.35)	19 (30.65)	17 (26.15)	
≥20,001 (US $2860.92)	14 (11.02)	7 (11.29)	7 (10.77)	
Child's gender, n (%)				.23
Male	71 (55.91)	32 (49.23)	24 (38.71)	
Female	56 (44.09)	33 (50.77)	38 (61.29)	
Child's age (years), mean (SD)	5.90 (3.12)	6.39 (3.36)	5.39 (2.78)	.06
Daily caregiving time for the child (hours/day), n (%)				.37
≤2	17 (13.39)	11 (16.92)	6 (9.68)	
3-6	55 (43.31)	27 (41.54)	28 (45.16)	
6-9	16 (12.6)	7 (10.77)	9 (14.52)	
9-12	15 (11.81)	6 (9.23)	9 (14.52)	
≥12	24 (18.9)	14 (21.54)	10 (16.13)	
Smartphone proficiency, n (%)				.53
Very proficient	66 (51.97)	33 (50.77)	33 (53.23)	
Basic proficient	22 (17.33)	14 (21.54)	8 (12.9)	
Fairly proficient	33 (25.98)	15 (23.08)	18 (29.03)	
Not proficient	6 (4.72)	3 (4.62)	3 (4.84)	

Primary Outcome

The group × time interaction in eHEALS was not significant (*P*=.26). The intervention group showed higher scores than the control group at T1 (33.62, 95% CI 32.76-34.49; *d*=0.20, 95% CI 0.13-0.56) and T2 (33.27, 95% CI 32.38-34.17; *d*=0.36, 95% CI 0.01-0.80), indicating sustained improvements following the LLM-generated learning intervention. Table 5 reports the means estimated from the model and the contrasts between groups across the specified time points; Figure 5 graphically illustrates the outcomes overtime by condition.

XSL•FO
RenderX

Table 5. Change in outcomes.

Time	Control EMM[a] (95% CI)	Intervention EMM (95% CI)	Group difference		Cohen *d* (95% CI)	Group × time interaction, *P* value
			β (95% CI)	SE		
eHEALS[b]						.26
T0	29.23 (28.05 to 30.40)	27.91 (26.24 to 29.57)	__[c]	—	—	—
T1	31.89 (31.01 to 32.77)	33.62 (32.76 to 34.49)	1.73 (0.49 to 2.97)	0.63	0.20 (0.13 to 0.56)	—
T2	30.87 (29.97 to 31.78)	33.27 (32.38 to 34.17)	2.40 (1.13 to 3.67)	0.65	0.36 (0.01 to 0.80)	—
DDH-KT[d]						.66
T0	4.42 (3.92 to 4.92)	4.31 (3.87 to 4.74)	—	—	—	—
T1	6.65 (6.26 to 7.05)	7.87 (7.48 to 8.25)	1.22 (0.67 to 1.77)	0.28	0.71 (0.33 to 1.11)	—
T2	6.02 (5.62 to 6.42)	7.12 (6.72 to 7.51)	1.10 (0.53 to 1.66)	0.29	0.54 (0.17 to 0.96)	—
HRPS[e]						.25
T0	27.60 (26.22 to 28.97)	27.20 (25.71 to 28.69)	—	—	—	—
T1	29.39 (28.33 to 30.45)	32.23 (31.19 to 33.26)	2.84 (1.36 to 4.31)	0.75	0.50 (0.12 to 0.86)	—
T2	29.48 (28.40 to 30.56)	31.55 (30.49 to 32.61)	2.07 (0.56 to 3.59)	0.77	0.41 (0.05 to 0.79)	—
ISES[f]						.25
T0	11.53 (10.99 to 12.08)	10.94 (10.30 to 11.58)	—	—	—	—
T1	12.38 (12.00 to 12.77)	12.59 (12.22 to 12.97)	0.21 (–0.33 to 0.75)	0.27	0.04 (0.30 to 0.39)	—
T2	11.96 (11.56 to 12.35)	12.51 (12.12 to 12.90)	0.55 (–0.00 to 1.11)	0.28	0.17 (0.19 to 0.55)	—
PUS[g]						.48
T0	15.03 (14.27 to 15.80)	14.12 (13.29 to 14.96)	—	—	—	—
T1	15.93 (15.40 to 16.46)	16.70 (16.18 to 17.21)	0.77 (0.02 to 1.51)	0.38	0.11 (0.22 to 0.49)	—
T2	15.61 (15.06 to 16.15)	16.66 (16.12 to 17.20)	1.05 (0.28 to 1.82)	0.39	0.15 (0.19 to 0.51)	—
HISBS[h]						.96
T0	14.15 (13.50 to 14.79)	13.15 (12.34 to 13.97)	—	—	—	—
T1	15.68 (15.10 to 16.27)	16.15 (15.57 to 16.72)	0.46 (–0.36 to 1.28)	0.42	0.02 (0.32 to 0.39)	—
T2	15.23 (14.63 to 15.83)	15.67 (15.08 to 16.26)	0.44 (–0.40 to 1.28)	0.43	0.05 (0.33 to 0.41)	—

[a]EMM: estimated marginal mean.

[b]eHEALS: eHealth Literacy Scale.

[c]Not applicable.

[d]DDH-KT: developmental dysplasia of the hip knowledge test.

[e]HRPS: Health Risk Perception Scale.

[f]ISES: Information Self-Efficacy Scale.

XSL•FO
RenderX

[g]PUS: Perceived Usefulness Scale.

[h]HISBS: Health Information-Seeking Behavior Scale.

Figure 5. Changes in outcomes over time by groups. (A) eHEALS (primary outcome), (B) DDH-KT, (C) HRPS, (D) ISES, (E) PUS, and (F) HISBS (secondary outcomes). DDH-KT: developmental dysplasia of the hip knowledge test; eHEALS: eHealth Literacy Scale; HISBS: Health Information-Seeking Behavior Scale; HRPS: Health Risk Perception Scale; ISES: Information Self-Efficacy Scale; PUS: Perceived Usefulness Scale.

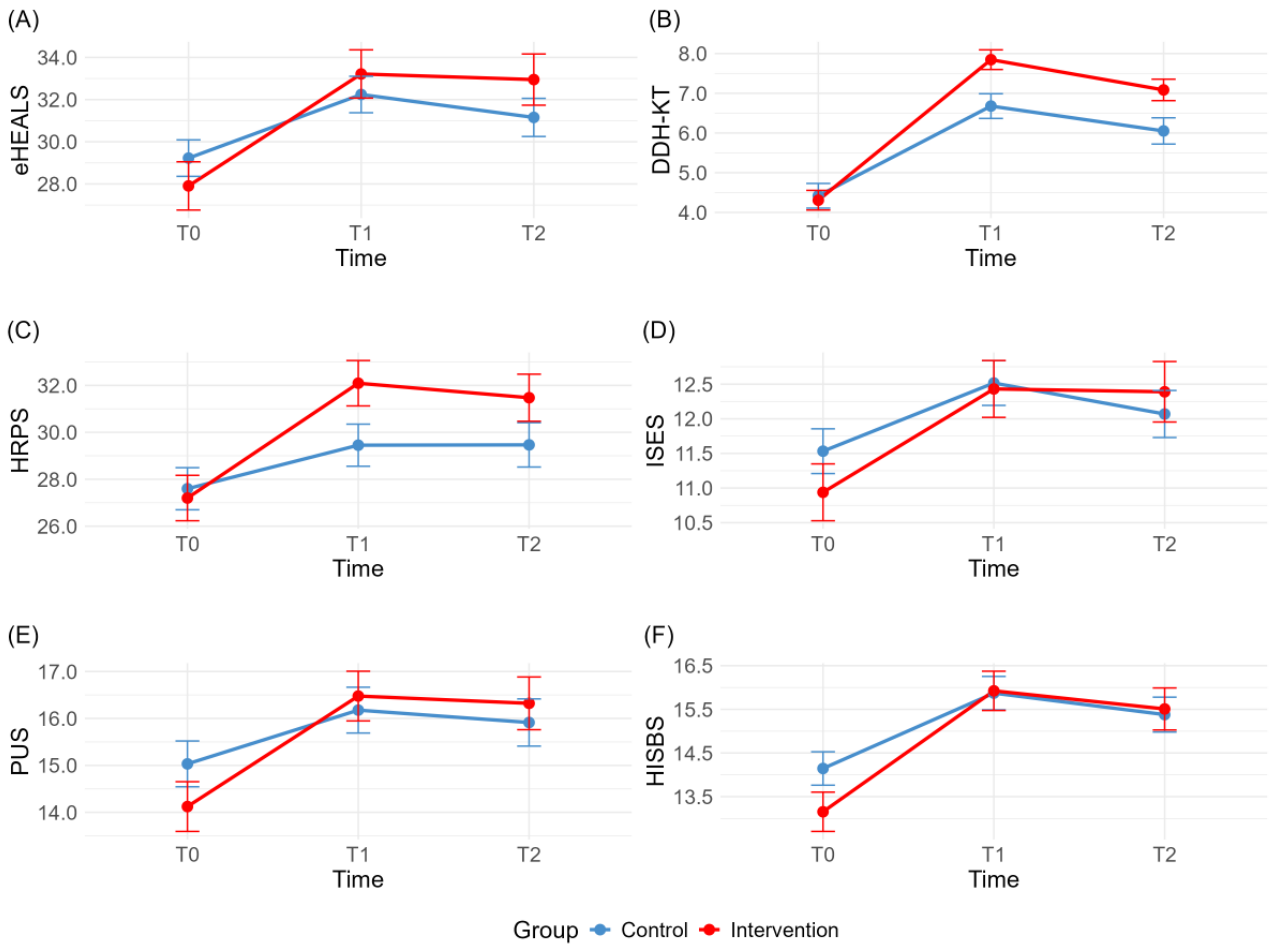

Secondary Outcomes

All secondary outcomes reported nonsignificant group × time interactions (*P*>.2), while the intervention group benefited from small to moderate impact sizes. DDH-KT scores were higher in the intervention group at T1 (7.87, 95% CI 7.48-8.25; *d*=0.71, 95% CI 0.33-1.11) and T2 (7.12, 95% CI 6.72-7.51; *d*=0.54, 95% CI 0.17-0.96). HRPS scores showed a similar pattern at T1 (32.23, 95% CI 31.19-33.26; *d*=0.50, 95% CI 0.12-0.86) and T2 (31.55, 95% CI 30.49-32.61; *d*=0.41, 95% CI 0.05-0.79). Additionally, PUS demonstrated consistent and statistically meaningful between-group differences favoring the intervention group at both T1 (16.70, 95% CI 16.18-17.21; *d*=0.11, 95% CI 0.22-0.49) and T2 (16.66, 95% CI 16.12-17.20; *d*=0.15, 95% CI 0.19-0.51). ISES and HISBS scores showed comparable positive trends; however, there were little differences across the groups.

Discussion

Principal Findings

This study evaluated the performance of 4 mainstream LLMs in education content and validated the effectiveness of LLM-generated caregiver education interventions. All 4 models demonstrated robust capabilities in generating content. ChatGPT-4 and DeepSeek-V3 outperformed Copilot and Gemini Flash in accuracy and fluency. The pilot trial suggests that LLM-assisted education may be associated with modest improvements in eHealth literacy (the primary outcome) and DDH knowledge compared with web-based searches; however, these findings should be interpreted as exploratory rather than confirmatory. These findings suggested that LLM-generated content was a feasible supplementary approach for health education. Its effectiveness appears to be enhanced when structured instruction and guided use are provided.

XSL•FO
RenderX

LLMs Performance

Overall, ChatGPT-4 performed well across several dimensions. It excelled in producing content that was logically clear and linguistically fluent. ChatGPT-4 was widely suitable for tasks with moderate complexity. DeepSeek-V3 was ideal for generating complex health education content, especially for requiring depth and professionalism. Gemini 2.0 Flash excelled in fluency and readability but had minor deficits in richness and accuracy. Its concise content is suitable for quick-reference scenarios. Gemini 2.0 Flash was useful for quickly accessing information. However, it was limited in tasks requiring depth. Its design focuses on simplicity and efficiency, suitable for everyday consultations or simple questioning and answering, and other low-complexity tasks. Copilot performed weakly in several dimensions, with omissions in its generated content and slightly obscure language expressions. It was suitable for tasks that require lower content quality.

All 4 LLMs scored at or above the neutral threshold (≥3/5) for accuracy, richness, and fluency. PEMAT-P understandability ≥70% indicated that basic comprehension standards were met. However, their PEMAT-P actionability was limited. This limitation may reduce the utility of LLM-generated handouts for guiding caregiver decisions. Only Copilot provides source citations, which raises concerns about the traceability and reliability of the information. Although readability levels were close to the average reading level of US adults (eighth grade) [36], they still exceeded American Medical Association recommendations (no more than sixth grade) for health education materials [37]. Nevertheless, the current web-based health education materials for orthopedic specialties were less than this recommendation [38]. This gap suggests that the readability of content generated by LLMs within the prompt framework has improved, but needs to be further optimized for the health education materials [39].

Based on publicly available official documentation and technical reports, the observed performance differences among the evaluated LLMs may be attributed to variations in training data, architectural design, and optimization objectives. ChatGPT-4 is described as a transformer-based multimodal model trained on a mixture of public and licensed data and aligned through supervised fine-tuning and reinforcement learning from human feedback. DeepSeek-V3 uses a mixture-of-experts architecture and large-scale pretraining, which may favor long-form generation and information coverage, helping explain its more comprehensive outputs. Gemini 2.0 Flash emphasizes efficiency and interaction speed, suggesting an optimization trade-off that supports fluency and readability but may constrain depth under limited prompting. Copilot functions as a product-level system rather than a fixed foundation model, with outputs influenced by orchestration layers and underlying model routing that can vary over time. Overall, these findings indicate that suitability for caregiver-oriented health education depends on how training data, architecture, and optimization priorities align with specific educational goals, rather than on overall model capability alone.

Evaluation Indicators

In practice, AI-assisted learning was associated with modest improvements in caregivers' eHealth literacy and DDH knowledge compared with unguided web-based searches. This encouraged the educational value of using LLM-generated content. Short-term exposure did not significantly increase self-efficacy or active information-seeking behavior. This observation was consistent with behavioral science evidence. It emphasized that knowledge improvement was insufficient to drive behavioral change without supportive motivation, confidence, and environmental reinforcement. Lasting behavioral changes may require longer reinforcement, repeated exposure, environmental support, or clinician guidance. Although content generated by advanced models was more accurate and detailed, caregivers generally preferred concise, readable materials over lengthy or overly technical texts. This indicated that optimal education required balancing accuracy, conciseness, and clarity, rather than solely pursuing information richness.

Comparison With Prior Work

Prior studies had mostly evaluated a single LLM using a limited set of metrics. For instance, ChatGPT-3.5's responses to spinal surgery questions were assessed solely for accuracy and readability [40]. This study extended previous research by systematically comparing 4 mainstream LLMs under identical conditions. We included expert ratings (accuracy, richness, and fluency), standardized assessment instruments (Patient Education Materials Assessment Tool and DISCERN), readability metrics, and learning outcomes. By connecting content quality to user learning outcomes, our study provided a more comprehensive and clinically relevant assessment of LLMs for health education. Based on prior teaching improvements using Bloom's taxonomy [41], it was used to improve the education by applying an organized method to content created by LLM. Prior studies showed that LLMs such as ChatGPT can enhance information accessibility, support communication and decision-making, and reduce anxiety levels [42]. These benefits have been demonstrated across diverse clinical contexts, including cancer care, orthopedic surgery, and mental health interventions [43-45]. The study reported that chatbot-enhanced prenatal education improved knowledge more effectively than standard mobile applications [46]. Our findings supported these findings by showing significant improvements in caregivers' eHealth literacy and knowledge of DDH. We focused more on enhancing eHealth literacy than on specific disease knowledge. This competency was essential not only for acquiring medical knowledge but also for enabling users to properly browse and use AI solutions across varied health information demands. Given that AI systems offer more flexible, interactive, and context-adaptive support than internet search, higher levels of eHealth literacy are necessary to ensure their safe and optimal use.

LLMs were characterized by actual-time dialogue, instant feedback, and personalized communication. These features enhanced user engagement during health education processes, thereby improving knowledge acquisition [44]. Participants in the intervention group demonstrated significantly higher health-risk perception than those in the web-based group, showing that personalized AI-generated information increases perceived relevance and strengthens risk understanding. Additionally, the immediate responses and conversational

XSL•FO
RenderX

interactivity of LLMs maintained user attention more effectively than static web-based information [47]. It resulted in increased satisfaction and maintained engagement.

Despite these advantages, some studies identified notable limitations in the accuracy and completeness of LLM outputs. McMahon and McMahon [48] warned that ChatGPT may generate misleading or unsafe recommendations in sensitive scenarios such as medication abortion. Ponzo et al [49] demonstrated that ChatGPT often produced incomplete or inconsistent dietary advice requiring professional revision. This pattern aligned with our heat-map analysis: LLMs performed the best in descriptive but worst in requiring clinical reasoning, procedural detail, or latest guideline recommendations, such as medication management, postoperative instructions, and emergency decision-making. These weaknesses appeared across multiple medical specialties and reflected broader constraints [50], including incomplete clinical training data, generating actionable guidance, and the universal LLMs' inherent cautious tendency. Thus, caregivers using AI-assisted information retrieval still require oversight and guidance from health care professionals [51].

Study Limitations

There are still some limitations to this study. First, although expert evaluation is an essential component of content quality assessment, it may carry the risk of subjective bias. Second, the evaluation was based on responses to a limited set of common DDH-related prompts. The variety and complexity of actual caregiver inquiries might not be adequately captured by such a limited selection of prompts. Third, each question was only created 3 times because of limitations on model use and study feasibility. Estimates of model variability would be more stable with more repetitions. Fourth, each LLM's web-based interface characteristics were standardized. It may cause slight differences when compared to the normal interaction situations of actual users. Finally, because LLMs undergo frequent updates and iterative changes, the findings of this study reflect model performance during the specific access period and may not fully generalize to future versions.

Practical Implications and Future Recommendations

The 2-stage results suggest that LLMs have potential as accessible, cost-effective, and personalized educational tools for caregivers, particularly in settings where traditional health education resources are limited. AI may supplement traditional clinician education by automating repetitive informational tasks, thereby alleviating health care professionals' workload and allowing them to prioritize complex clinical cases. Enhancing knowledge and timely medical consultation are especially important for the early recognition of DDH. In rural and remote places with inadequate medical services, LLMs may help minimize geographic and economic obstacles to health education, increasing educational reach [52].

The perceived utility of AI-generated content is not solely determined by technical accuracy. Although ChatGPT-4 and DeepSeek-V3 generated high-quality content, users do not always prefer longer or more detailed responses. Caregivers, especially older adults, often prefer concise and clear information [53]. It suggests that instructional design should balance content quality with readability. Accordingly, when incorporating LLMs into clinical education, health educators may consider structured prompting and staged content generation. Instructional design might begin with simple explanations. As users express interest, gradually provide more specialized information with a guided summary.

However, the risks of misinformation, hallucinations, and unclear accountability cannot be ignored. LLM outputs exhibit inherent uncertainty; responses can vary across conversational contexts and may produce plausible but inaccurate statements regarding diagnostic thresholds or guideline-specific recommendations [54]. Furthermore, potential biases in training data may limit the cultural and contextual adaptability of these models [55]. As they may inadvertently reflect high-resource health care assumptions while overlooking local beliefs, language nuances, or service availability. Therefore, to ensure safe use, LLMs should be positioned strictly as auxiliary tools rather than substitutes for comprehensive medical assessments, physical examinations, and consultations with health care professionals [56]. In clinical practice, data confidentiality must be treated as a primary prerequisite. Patients provide informed consent for the use of LLM-assisted education, and workflows explicitly discourage the entry or disclosure of identifiable personal information [57]. Professional monitoring is crucial because LLM-generated content can be ambiguous, erroneous, or prejudiced. This includes regular evaluation of AI-generated educational outputs, bias-aware checks, and escalation procedures when high-risk issues emerge [58]. Future implementation strategies include retrieval-augmented generation, expert review mechanisms, and standardized safety and regulatory frameworks. With these safeguards, systematic incorporation of LLMs into health care procedures may support standardized health education and improve efficiency and scalability without compromising safety [59]. Future work should also identify the support resources required for safe adoption, including staff training, governance and auditing procedures, and technical infrastructure. Therefore, LLMs hold potential to support future health education and clinical communication.

Implications for Practice

The implications for practice are that we (1) prefer models that cite reliable sources, (2) use prompts that request guideline-based advice, (3) always include disclaimers clarifying that LLMs cannot replace professional consultation, (4) target ≤6th-grade readability and simplify outputs with follow-up prompts, and (5) review and adapt content before sharing with patients.

Conclusions

This study demonstrates that LLMs hold substantial potential for supporting education in DDH. ChatGPT-4 achieved 85% accuracy and 93% fluency, while DeepSeek-V3 led in 83% richness, generally outperforming the Copilot and Gemini 2.0 Flash. AI-assisted education was associated with small to moderate effect sizes for caregivers' eHealth literacy, DDH knowledge, health risk perception, and perceived usefulness compared with web-based searches in this pilot trial. In addition,

XSL•FO
RenderX

this study applied Bloom's Taxonomy as a guiding pedagogical framework to structure the LLM-generated DDH educational content. This approach allowed the content to support the spectrum of caregiver learning needs, extending from foundational knowledge acquisition to decision-oriented guidance. Study limitations include potential expert subjectivity, a narrow prompt set with few generations, and controlled interface settings. LLMs are auxiliary tools and cannot replace the need for professionals. Future research should focus on optimizing plain language, refining dialogue design, and enhancing audience personalization to improve the quality of materials generated by LLMs.

Funding

This work was supported by the Guangdong Provincial Education Science Planning Project (Higher Education Research Special Topic) for 2025 (2025GXJK0331); the Science and Technology Program of Guangzhou Sports Bureau for 2025 (ST20250986); the Nursing Research Special Program of Southern Medical University for 2025 (Y2025008); the Education and Teaching Research Project of The Third Affiliated Hospital, Southern Medical University, for 2025 (JXY202517); and the President Foundation of The Third Affiliated Hospital, Southern Medical University, for 2022 (YH202207).

Data Availability

The full data supporting this study, including scoring data from phase 1 and outcomes from phase 2, can be accessed from the Multimedia Appendices.

Authors' Contributions

Conceptualization: YL
Data curation: HO, ZY, YL
Formal analysis: GL, JY
Funding acquisition: YC
Methodology: YL
Supervision: FQ, JY
Validation: GL, HY
Visualization: YL
Writing – original draft: HO
Writing – review & editing: FQ, YC

Conflicts of Interest

None declared.

Multimedia Appendix 1
Question bank and generated data.
[DOCX File , 303 KB - jmir_v28i1e73326_app1.docx]

Multimedia Appendix 2
Formulas for evaluating readability.
[DOCX File , 38 KB - jmir_v28i1e73326_app2.docx]

Multimedia Appendix 3
DDH knowledge test. DDH: developmental dysplasia of the hip.
[DOCX File , 17 KB - jmir_v28i1e73326_app3.docx]

Multimedia Appendix 4
Phase 1 scoring data.
[XLSX File (Microsoft Excel File), 55 KB - jmir_v28i1e73326_app4.xlsx]

Multimedia Appendix 5
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 2774 KB - jmir_v28i1e73326_app5.pdf]

Multimedia Appendix 6

XSL•FO
RenderX

Phase 2 participants' scoring data.
[XLSX File (Microsoft Excel File), 31 KB - jmir_v28i1e73326_app6.xlsx]

References

1. Tao Z, Wang J, Li Y, Zhou Y, Yan X, Yang J, et al. Prevalence of developmental dysplasia of the hip (DDH) in infants: a systematic review and meta-analysis. BMJ Paediatr Open 2023 Oct;7(1):e002080 [FREE Full text] [doi: 10.1136/bmjpo-2023-002080] [Medline: 37879719]
2. American Academy of Pediatrics. Long-term outcome of delayed diagnosis of developmental hip dysplasia. AAP Grand Rounds 2020 Nov;44:53-53. [doi: 10.1542/gr.44-5-53] [Medline: 26033050]
3. Kolb A, Chiari C, Schreiner M, Heisinger S, Willegger M, Rettl G, et al. Development of an electronic navigation system for elimination of examiner-dependent factors in the ultrasound screening for developmental dysplasia of the hip in newborns. Sci Rep 2020 Oct 02;10(1):16407 [FREE Full text] [doi: 10.1038/s41598-020-73536-9] [Medline: 33009470]
4. Bakarman K, Alsiddiky A, Zamzam M, Alzain KO, Alhuzaimi FS, Rafiq Z. Developmental dysplasia of the hip (DDH): etiology, diagnosis, and management. Cureus 2023;15(8):e43207 [FREE Full text] [doi: 10.7759/cureus.43207] [Medline: 37692580]
5. Pagano S, Holzapfel S, Kappenschneider T, Meyer M, Maderbacher G, Grifka J, et al. Arthrosis diagnosis and treatment recommendations in clinical practice: an exploratory investigation with the generative AI model GPT-4. J Orthop Traumatol 2023 Nov 28;24(1):61 [FREE Full text] [doi: 10.1186/s10195-023-00740-4] [Medline: 38015298]
6. Yun JY, Kim DJ, Lee N, Kim EK. A comprehensive evaluation of ChatGPT consultation quality for augmentation mammoplasty: A comparative analysis between plastic surgeons and laypersons. Int J Med Inform 2023 Nov;179:105219 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.105219] [Medline: 37776670]
7. Bernstein IA, Zhang YV, Govil D, Majid I, Chang RT, Sun Y, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Netw Open 2023 Aug 01;6(8):e2330320 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.30320] [Medline: 37606922]
8. Dağci M, Çam F, Dost A. Reliability and quality of the nursing care planning texts generated by ChatGPT. Nurse Educ 2024;49(3):E109-E114. [doi: 10.1097/NNE.0000000000001566] [Medline: 37994523]
9. Shen SA, Perez-Heydrich CA, Xie DX, Nellis JC. ChatGPT vs. web search for patient questions: what does ChatGPT do better? Eur Arch Otorhinolaryngol 2024 Jun;281(6):3219-3225. [doi: 10.1007/s00405-024-08524-0] [Medline: 38416195]
10. Hopkins AM, Logan JM, Kichenadasse G, Sorich MJ. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. JNCI Cancer Spectr 2023 Mar 01;7(2):pkad010 [FREE Full text] [doi: 10.1093/jncics/pkad010] [Medline: 36808255]
11. Amaral JZ, Schultz RJ, Martin BM, Taylor T, Touban B, McGraw-Heinrich J, et al. Evaluating chat generative pre-trained transformer responses to common pediatric in-toeing questions. J Pediatr Orthop 2024 Aug 01;44(7):e592-e597. [doi: 10.1097/BPO.0000000000002695] [Medline: 38686934]
12. Alber DA, Yang Z, Alyakin A, Yang E, Rai S, Valliani AA, et al. Medical large language models are vulnerable to data-poisoning attacks. Nat Med 2025 Feb;31(2):618-626. [doi: 10.1038/s41591-024-03445-1] [Medline: 39779928]
13. Xiao Y, Wang W. On hallucination and predictive uncertainty in conditional language generation. In: Association for Computational Linguistics. 2021 Apr Presented at: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume; 2021; Online p. 2734-2744. [doi: 10.18653/v1/2021.eacl-main.236]
14. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne) 2024 Oct;11:1477898 [FREE Full text] [doi: 10.3389/fmed.2024.1477898] [Medline: 39534227]
15. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
16. Boland L, Graham ID, Légaré F, Lewis K, Jull J, Shephard A, et al. Barriers and facilitators of pediatric shared decision-making: a systematic review. Implement Sci 2019 Jan 18;14(1):7 [FREE Full text] [doi: 10.1186/s13012-018-0851-5] [Medline: 30658670]
17. Bloom BS, Engelhart MD, Furst EJ, Hill WH. Taxonomy of Educational Objectives. New York: Longmans, Green; 1964.
18. Al-Haj Ali SN. Emergency management of permanent tooth avulsion: comparative performance of dental students and artificial intelligence using a multimodal clinical vignette across Bloom's taxonomy domains. Int Endod J 2025 Nov 11 (forthcoming). [doi: 10.1111/iej.70060] [Medline: 41216987]
19. American Institute of Ultrasound in Medicine. AIUM practice guideline for the performance of an ultrasound examination for detection and assessment of developmental dysplasia of the hip. J Ultrasound Med 2013 Jul;32(7):1307-1317. [doi: 10.7863/ultra.32.7.1307] [Medline: 23804356]
20. AIUM-ACR-SPR-SRU practice parameter for the performance of an ultrasound examination for detection and assessment of developmental dysplasia of the hip. J Ultrasound Med 2018;37(11):E1-E5. [doi: 10.1002/jum.14829] [Medline: 30308084]

XSL•FO
RenderX

21. Committee on Quality Improvement, Subcommittee on Developmental Dysplasia of the Hip, American Academy of Pediatrics. Clinical practice guideline: early detection of developmental dysplasia of the hip. Pediatrics 2000 Apr;105:896-905. [doi: 10.1542/peds.105.4.896] [Medline: 10742345]
22. Nguyen JC, Dorfman SR, Rigsby CK, Iyer RS, Alazraki AL, Anupindi SA, et al. ACR Appropriateness Criteria developmental dysplasia of the hip-child. J Am Coll Radiol 2019 May;16(5S):S94-S103. [doi: 10.1016/j.jacr.2019.02.014] [Medline: 31054762]
23. Mulpuri K, Song KM, Gross RH, Tebor GB, Otsuka NY, Lubicky JP, et al. The American Academy of Orthopaedic Surgeons evidence-based guideline on detection and nonoperative management of pediatric developmental dysplasia of the hip in infants up to six months of age. J Bone Joint Surg Am 2015 Oct 21;97(20):1717-1718. [doi: 10.2106/JBJS.O.00500] [Medline: 26491137]
24. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999 Feb;53(2):105-111 [FREE Full text] [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
25. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014 Sep;96(3):395-403 [FREE Full text] [doi: 10.1016/j.pec.2014.05.027] [Medline: 24973195]
26. Kestel S, Çalık A, Kuş M. The effect of chatbot-supported instruction on nursing students' history-taking questioning skills and stress level: a randomized controlled study. J Prof Nurs 2025;60:93-100. [doi: 10.1016/j.profnurs.2025.07.004] [Medline: 40915772]
27. Hip dysplasia. Wikipedia. URL: https://en.wikipedia.org/wiki/Hip_dysplasia [accessed 2026-01-06]
28. Norman CD, Skinner HA. eHEALS: The eHealth Literacy Scale. J Med Internet Res 2006 Nov 14;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
29. Ajzen I. Understanding Attitudes and Predicting Social Behavior. Englewood Cliffs, NJ: Prentice-Hall; 1980.
30. Brewer NT, Weinstein ND, Cuite CL, Herrington JE. Risk perceptions and their relation to risk behavior. Ann Behav Med 2004 Apr;27(2):125-130. [doi: 10.1207/s15324796abm2702_7] [Medline: 15026296]
31. Pavlou, Fygenson. Understanding and predicting electronic commerce adoption: an extension of the theory of planned behavior. MIS Q 2006;30(1):115-143. [doi: 10.2307/25148720]
32. Cheung CM, Lee MK, Rabjohn N. The impact of electronic word - of - mouth. Internet Res 2008;18(3):229-247. [doi: 10.1108/10662240810883290]
33. Kankanhalli, Tan, Wei. Contributing knowledge to electronic knowledge repositories: an empirical investigation. MIS Q 2005;29(1):113-144. [doi: 10.2307/25148670]
34. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163 [FREE Full text] [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
35. Jacobson N, Follette WC, Revenstorf D. Psychotherapy outcome research: methods for reporting variability and evaluating clinical significance. Behavior Therapy 1984 Sep;15(4):336-352 [FREE Full text] [doi: 10.1016/s0005-7894(84)80002-7]
36. Doak CC, Doak LG, Friedell GH, Meade CD. Improving comprehension for cancer patients with low literacy skills: strategies for clinicians. CA Cancer J Clin 1998;48(3):151-162 [FREE Full text] [doi: 10.3322/canjclin.48.3.151] [Medline: 9594918]
37. Chinn L, McGuirt S, Puri S. The readability of online patient education materials from major anesthesiology associations and the American Society of Anesthesiologists. OJAnes 2014;04(01):1-7. [doi: 10.4236/ojanes.2014.41001]
38. Michel C, Dijanic C, Abdelmalek G, Sudah S, Kerrigan D, Gorgy G, et al. Readability assessment of patient educational materials for pediatric spinal deformity from top academic orthopedic institutions. Spine Deform 2022 Nov;10(6):1315-1321 [FREE Full text] [doi: 10.1007/s43390-022-00545-1] [Medline: 35819724]
39. Dihan Q, Chauhan MZ, Eleiwa TK, Hassan AK, Sallam AB, Khouri AS, et al. Using large language models to generate educational materials on childhood glaucoma. Am J Ophthalmol 2024 Sep;265:28-38. [doi: 10.1016/j.ajo.2024.04.004] [Medline: 38614196]
40. Hernandez F, Guizar R, Avetisian H, Abdou MA, Karakash WJ, Ton A, et al. Evaluating the accuracy and readability of ChatGPT in addressing patient queries on adult spinal deformity surgery. Global Spine J 2025 Jul 11:21925682251360655 [FREE Full text] [doi: 10.1177/21925682251360655] [Medline: 40643892]
41. Ray M, Rudolph M, Daugherty K. Bloom's taxonomy in health professions education: associations with exam scores, clinical reasoning, and instructional effectiveness. Curr Pharm Teach Learn 2025 Nov;17(11):102444 [FREE Full text] [doi: 10.1016/j.cptl.2025.102444] [Medline: 40695205]
42. Park C, An MH, Hwang G, Park RW, An J. Clinical performance and communication skills of ChatGPT versus physicians in emergency medicine: simulated patient study. JMIR Med Inform 2025 Jul 17;13:e68409 [FREE Full text] [doi: 10.2196/68409] [Medline: 40674718]
43. Gan W, Ouyang J, She G, Xue Z, Zhu L, Lin A, et al. ChatGPT's role in alleviating anxiety in total knee arthroplasty consent process: a randomized controlled trial pilot study. Int J Surg 2025 Mar 01;111(3):2546-2557. [doi: 10.1097/JS9.0000000000002223] [Medline: 39903546]

XSL•FO
RenderX

44. Akdogan O, Uyar GC, Yesilbas E, Baskurt K, Malkoc NA, Ozdemir N, et al. Effect of a ChatGPT-based digital counseling intervention on anxiety and depression in patients with cancer: a prospective, randomized trial. Eur J Cancer 2025 May 15;221:115408. [doi: 10.1016/j.ejca.2025.115408] [Medline: 40215593]
45. Tong ACY, Wong KTY, Chung WWT, Mak WWS. Effectiveness of topic-based chatbots on mental health self-care and mental well-being: randomized controlled trial. J Med Internet Res 2025 Apr 30;27:e70436 [FREE Full text] [doi: 10.2196/70436] [Medline: 40306635]
46. Su B, Jones R, Chen K, Kostenko E, Schmid M, DeMaria AL, et al. Chatbot for patient education for prenatal aneuploidy testing: a multicenter randomized controlled trial. Patient Educ Couns 2025 Feb;131:108557 [FREE Full text] [doi: 10.1016/j.pec.2024.108557] [Medline: 39642634]
47. Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res 2020 Oct 22;22(10):e20346 [FREE Full text] [doi: 10.2196/20346] [Medline: 33090118]
48. McMahon HV, McMahon BD. Automating untruths: ChatGPT, self-managed medication abortion, and the threat of misinformation in a post- world. Front Digit Health 2024;6:1287186 [FREE Full text] [doi: 10.3389/fdgth.2024.1287186] [Medline: 38419805]
49. Ponzo V, Goitre I, Favaro E, Merlo FD, Mancino MV, Riso S, et al. Is ChatGPT an effective tool for providing dietary advice? Nutrients 2024 Feb 06;16(4):469 [FREE Full text] [doi: 10.3390/nu16040469] [Medline: 38398794]
50. Dhar S, Kothari D, Vasquez M, Clarke T, Maroda A, McClain WG, et al. The utility and accuracy of ChatGPT in providing post-operative instructions following tonsillectomy: a pilot study. Int J Pediatr Otorhinolaryngol 2024 Apr;179:111901. [doi: 10.1016/j.ijporl.2024.111901] [Medline: 38447265]
51. Abreu A, Murimwa G, Farah E, Stewart J, Zhang L, Rodriguez J, et al. Enhancing readability of online patient-facing content: the role of AI chatbots in improving cancer information accessibility. J Natl Compr Canc Netw 2024 May 15;22(2 D):e237334. [doi: 10.6004/jnccn.2023.7334] [Medline: 38749478]
52. Wah JNK. Revolutionizing e-health: the transformative role of AI-powered hybrid chatbots in healthcare solutions. Front Public Health 2025;13:1530799 [FREE Full text] [doi: 10.3389/fpubh.2025.1530799] [Medline: 40017541]
53. Goodman C, Lambert K. Scoping review of the preferences of older adults for patient education materials. Patient Educ Couns 2023 Mar;108:107591 [FREE Full text] [doi: 10.1016/j.pec.2022.107591] [Medline: 36584555]
54. Şahin MF, Topkaç EC, Doğan, Şeramet S, Özcan R, Akgül M, et al. Still using only ChatGPT? The comparison of five different artificial intelligence chatbots' answers to the most common questions about kidney stones. J Endourol 2024 Nov;38(11):1172-1177. [doi: 10.1089/end.2024.0474] [Medline: 39212674]
55. Busch F, Hoffmann L, Rueger C, van Dijk EH, Kader R, Ortiz-Prado E, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med (Lond) 2025 Jan 21;5(1):26 [FREE Full text] [doi: 10.1038/s43856-024-00717-2] [Medline: 39838160]
56. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs). NPJ Digit Med 2024 Jul 08;7(1):183 [FREE Full text] [doi: 10.1038/s41746-024-01157-x] [Medline: 38977771]
57. Zhou Y, Li S, Tang X, He Y, Ma H, Wang A, et al. Using ChatGPT in nursing: scoping review of current opinions. JMIR Med Educ 2024 Nov 19;10:e54297 [FREE Full text] [doi: 10.2196/54297] [Medline: 39622702]
58. Berşe S, Akça K, Dirgar E, Kaplan Serin E. The role and potential contributions of the artificial intelligence language model ChatGPT. Ann Biomed Eng 2024 Feb;52(2):130-133. [doi: 10.1007/s10439-023-03296-w] [Medline: 37378876]
59. Tilton AK, Caplan BE, Cole BJ. Generative AI in consumer health: leveraging large language models for health literacy and clinical safety with a digital health framework. Front Digit Health 2025;7:1616488. [doi: 10.3389/fdgth.2025.1616488] [Medline: 40933812]

Abbreviations

AI: artificial intelligence
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
DDH: developmental dysplasia of the hip
DDH-KT: developmental dysplasia of the hip knowledge test
eHEALS: eHealth Literacy Scale
FKGL: Flesch-Kincaid Grade Level
FKRE: Flesch-Kincaid Reading Ease
HISBS: Health Information-Seeking Behavior Scale
HRPS: Health Risk Perception Scale
ICC: intraclass correlation coefficient
ISES: Information Self-Efficacy Scale

XSL•FO
RenderX

LLM: large language model
PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials
PUS: Perceived Usefulness Scale
RCT: randomized controlled trial
SMOG: Simple Measure of Gobbledygook

Edited by J Sarvestan; submitted 29.Jul.2025; peer-reviewed by J Zhang, GG de Alencar; comments to author 29.Sep.2025; accepted 19.Dec.2025; published 19.Jan.2026.

Please cite as:
Ouyang H, Lin G, Li Y, Yao Z, Li Y, Yan H, Qin F, Yao J, Chen Y
Evaluating and Validating Large Language Models for Health Education on Developmental Dysplasia of the Hip: 2-Phase Study With Expert Ratings and a Pilot Randomized Controlled Trial
J Med Internet Res 2026;28:e73326
URL: https://www.jmir.org/2026/1/e73326
doi:10.2196/73326
PMID:

©Hui Ouyang, Gan Lin, Yiyuan Li, Zhixin Yao, Yating Li, Han Yan, Fang Qin, Jinghui Yao, Yun Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Data Poisoning Vulnerabilities Across Health Care Artificial Intelligence Architectures: Analytical Security Framework and Defense Strategies

Farhad Abtahi[1,2,3], PhD; Fernando Seoane[1,4,5,3], PhD; Ivan Pau[6], PhD; Mario Vega-Barbas[6], PhD

[1]Department of Clinical Science, Intervention and Technology, Karolinska Institutet, Huddinge, Stockholm, Sweden
[2]Department of Biomedical Engineering and Health Systems, KTH Royal Institute of Technology, Huddinge, Stockholm, Sweden
[3]Department of Clinical Physiology, Karolinska University Hospital, Huddinge, Stockholm, Sweden
[4]Department of Medical Technology, Karolinska University Hospital, Stockholm, Sweden
[5]Department of Textile Technology, University of Borås, Borås, Västra Götaland, Sweden
[6]ETSIS de Telecomunicación, Universidad Politécnica de Madrid, Madrid, Madrid, Spain

Corresponding Author:
Farhad Abtahi, PhD
Department of Clinical Science, Intervention and Technology
Karolinska Institutet
Alfred Nobels Allé 8
Huddinge, Stockholm, 14152
Sweden
Phone: 46 8 524 838 01
Email: farhad.abtahi@ki.se

Abstract

Background: Health care artificial intelligence (AI) systems are increasingly integrated into clinical workflows, yet remain vulnerable to data-poisoning attacks. A small number of manipulated training samples can compromise AI models used for diagnosis, documentation, and resource allocation. Existing privacy regulations, including the Health Insurance Portability and Accountability Act and the General Data Protection Regulation, may inadvertently complicate anomaly detection and cross-institutional auditing, thereby limiting visibility into adversarial activity.

Objective: This study provides a comprehensive threat analysis of data poisoning vulnerabilities across major health care AI architectures. The goals are to (1) identify attack surfaces in clinical AI systems, (2) evaluate the feasibility and detectability of poisoning attacks analytically modeled in prior security research, and (3) propose a multilayered defense framework appropriate for health care settings.

Methods: We synthesized empirical findings from 41 key security studies published between 2019 and 2025 and integrated them into an analytical threat-modeling framework specific to health care. We constructed 8 hypothetical yet technically grounded attack scenarios across 4 categories: (1) architecture-specific attacks on convolutional neural networks, large language models, and reinforcement learning agents (scenario A); (2) infrastructure exploitation in federated learning and clinical documentation pipelines (scenario B); (3) poisoning of critical resource allocation systems (scenario C); and (4) supply chain attacks affecting commercial foundation models (scenario D). Scenarios were aligned with realistic insider-access threat models and current clinical deployment practices.

Results: Multiple empirical studies demonstrate that attackers with access to as few as 100-500 poisoned samples can compromise health care AI systems, with attack success rates typically ≥60%. Critically, attack success depends on the absolute number of poisoned samples rather than their proportion of the training corpus, a finding that fundamentally challenges assumptions that larger datasets provide inherent protection. We estimate that detection delays commonly range from 6 to 12 months and may extend to years in distributed or privacy-constrained environments. Analytical scenarios highlight that (1) routine insider access creates numerous injection points across health care data infrastructure, (2) federated learning amplifies risks by obscuring attribution, and (3) supply chain compromises can simultaneously affect dozens to hundreds of institutions. Privacy regulations further complicate cross-patient correlation and model audit processes, substantially delaying the detection of subtle poisoning campaigns.

XSL•FO
RenderX

Conclusions: Health care AI systems face significant security challenges that current regulatory frameworks and validation practices do not adequately address. We propose a multilayered defense strategy that combines ensemble disagreement monitoring, adversarial testing, privacy-preserving yet auditable mechanisms, and strengthened governance requirements. Ensuring patient safety may require a shift from opaque, high-performance models toward more interpretable and constraint-driven architectures with verifiable robustness guarantees.

(J Med Internet Res 2026;28:e87969) doi:10.2196/87969

KEYWORDS

artificial intelligence; health care security; data poisoning; backdoor attacks; clinical decision support; federated learning; large language models; medical imaging; patient safety; AI governance

Introduction

Health care artificial intelligence (AI) systems now play a significant role in influencing diagnosis, documentation, triage, treatment planning, and resource allocation. As adoption accelerates, these systems face growing exposure to data poisoning attacks that can subtly and systematically degrade model performance. Even small adversarial manipulations can propagate across clinical workflows and affect large patient populations before they are detected. Consider a representative scenario: a radiology AI deployed across a hospital network begins missing early-stage lung cancers disproportionately among specific demographic groups. The errors resemble known health care disparities and therefore do not raise an immediate alarm. Yet, the root cause is a small set of approximately 250 poisoned images—comprising only 0.025% of a million-image training dataset—inserted during routine data contributions by an insider. Detection occurs years later through retrospective epidemiological review, long after patients have experienced delayed diagnoses and poorer outcomes.

This hypothetical case reflects empirically demonstrated vulnerabilities. Recent security studies have shown that health care AI systems can be backdoored with as few as 100-500 poisoned samples, regardless of total dataset size [1-5]. Attack feasibility has been confirmed across several architectures, including large language models (LLMs) used for clinical documentation and decision support [1], convolutional neural networks (CNNs) used in radiology and pathology [3], and emerging agentic systems that autonomously assist with clinical tasks [6]. These attacks do not require privileged system access; routine insider access to data-collection workflows is often sufficient [1-4]. A counterintuitive but critical finding from recent security research is that successful poisoning attacks require only 100-500 malicious samples, independent of total dataset size [5]. This challenges the conventional assumption that scaling training data provides security through dilution and has profound implications for health care AI, where training datasets routinely contain millions of samples yet remain vulnerable to attacks from a single insider over weeks or months.

Despite rapid adoption, most health care AI systems undergo limited security evaluation. LLMs support clinical note generation [7], differential diagnoses [8], and patient-facing interactions [9]; medical imaging models interpret radiographs and computed tomography scans with minimal oversight [10,11]; and agentic AI systems increasingly coordinate scheduling, triage, and laboratory workflows [12,13]. Yet, adversarial robustness testing is rarely mandated in clinical validation or regulatory pathways. Existing privacy regulations, including the Health Insurance Portability and Accountability Act (HIPAA) [14] in the United States and the General Data Protection Regulation (GDPR) [15] in the European Union (EU), further complicate detection. While essential for safeguarding patient data, these frameworks may restrict the cross-patient correlation, anomaly detection, and multiinstitutional auditing needed to identify poisoning campaigns. Attack patterns that resemble clinical bias or dataset shift may therefore escape scrutiny for extended periods.

Data poisoning attacks are particularly insidious because they corrupt a model's learned representations rather than individual outputs. Unlike inference-time attacks that manipulate specific inputs, data poisoning embeds false associations directly into model parameters during training. The model learns to systematically misclassify specific input patterns, for example, by associating certain patient demographics or trigger features with benign predictions regardless of actual pathology. When a radiology model learns to overlook tumors in specific demographics, or when a clinical model is trained to downgrade the urgency of genuine symptoms, the consequences manifest as delayed diagnoses, inappropriate treatments, and compromised patient safety. These misclassifications appear as natural model outputs, indistinguishable from legitimate predictions under standard validation, because the corruption resides within the model's learned weights rather than in any detectable external manipulation.

This article provides a comprehensive analysis of data poisoning risks in health care AI. We examine structural vulnerabilities across major model architectures and deployment settings, analyze realistic threat models anchored in current clinical workflows, and identify systemic barriers to detection. Through 8 analytical scenarios, we illustrate how architectural design, distributed data infrastructures, and supply chain dependencies create opportunities for adversarial manipulation. Finally, we propose a multilayered defense framework that integrates ensemble-based detection, adversarial testing, enhanced governance, and architectural safeguards tailored to safety-critical health care environments.

Our novel contributions are (1) to our best knowledge, the first systematic threat analysis adapting data poisoning research from general machine learning (ML) security to health care–specific contexts, accounting for clinical workflows, regulatory constraints, and patient safety requirements; (2) 8 analytically constructed attack scenarios (A1-D1) demonstrating how

XSL•FO
RenderX

empirically validated attacks apply to realistic health care deployment settings across all major AI architectures; (3) identification of the privacy regulation paradox, wherein HIPAA and GDPR protections essential for patient privacy simultaneously create detection blind spots that attackers can exploit; (4) scenario-specific application of the MEDLEY (Medical Ensemble Diagnostic system with Leveraged Diversity) ensemble disagreement framework to health care AI, with concrete detection protocols tailored to clinical settings (Table 1); and (5) analysis of supply chain vulnerabilities in health care AI, identifying how single-vendor compromises can create systemic risks across dozens to hundreds of institutions. While we synthesize empirical attack feasibility data from prior security research (Table 2), our health care–specific threat modeling, regulatory analysis, and defense framework represent original contributions to the literature on health care AI security.

Table 1. MEDLEY[a] framework application to attack scenarios[b].

Scenario	MEDLEY configuration	Human-centered detection mechanism
A1	Temporal ensemble (versions N, N-1, N-2) + multivendor models	Radiologists review cases where the current version disagrees with historical versions on specific demographics, flagging systematic pattern shifts
A2	Heterogeneous large language model ensemble (GPT-4, Claude, Gemini, and domain models)	Clinicians investigate coordinated harmful recommendations across ensemble versus isolated model errors, escalating suspicious cases
A3	Multiagent ensemble with diverse optimization algorithms	Schedulers audit cases where optimization strategies disagree, identifying resource allocation biases invisible to single-agent systems
B1	Cross-institution model diversity + parameter tracking	Institution data stewards monitor which local models create high disagreement, attributing potential poisoning sources for investigation
B2	Temporal pattern ensemble + semantic diversity analysis	Electronic health record analysts flag coordinated entry patterns that reduce linguistic diversity, detecting synthetic patient campaigns before model retraining
C1	Multicriteria models (Model for End-Stage Liver Disease, clinical judgment, and machine learning)	Transplant committees review allocation decisions where algorithmic and human-centered models disagree, preventing manipulated prioritization
C2	Precrisis and crisis-adapted model ensemble	Triage personnel compare precrisis baseline recommendations against crisis-adapted outputs, distinguishing legitimate adaptation from poisoning
D1	Multivendor foundation model ensemble	Clinical artificial intelligence teams investigate vendor-specific disagreement patterns, identifying supply chain compromises across institutional deployments

[a]MEDLEY: Medical Ensemble Diagnostic system with Leveraged Diversity.

[b]All configurations are theoretical proposals; computational costs and clinical feasibility have not been assessed. Validation status: unvalidated.

XSL•FO
RenderX

Table 2. Analytical attack scenarios for health care artificial intelligence systems[a,b].

Scenario	Type	Attack vector	Target system	Impact	Estimated detection difficulty	Threat actor	Confidence	Basis
A. Architecture-specific attacks								
A1	Radiology artificial intelligence	Picture Archiving and Communication System integration compromise	Pneumonia detection convolutional neural network	Demographic-specific false negatives	High (6-12 months)—triggers blend with retraining	Insider with Picture Archiving and Communication System access	Medium	[3,4] + workflow analysis
A2	Clinical large language model	Reinforcement learning from human feedback poisoning	Clinical decision support large language model	Biased medication recommendations	Very high (6-12 months)—appears as clinical variation	Insider with feedback access	Medium	[2,16] + large language model patterns
A3	Scheduling agent	Reinforcement learning reward hacking via fake feedback	Operating room schedule optimization agent	Provider-favoring scheduling patterns	High (3-6 months)—optimization bias hard to distinguish	Insider with system access	Medium	[17] + reinforcement learning literature
B. Infrastructure exploitation attacks								
B1	Federated learning	Edge node model poisoning	Multisite pathology classifier	Systematic rare cancer misclassification	Extreme (>1 year)—distributed trust obscures source	Compromised institution	Medium	[18,19] + federated learning adoption
B2	Medical scribe (Sybil attack)	Coordinated fake patient visits with scripted histories	Artificial intelligence scribe → electronic health record → all downstream clinical artificial intelligence	Large-scale dataset poisoning across all clinical artificial intelligence systems	Extreme (>1 year or never)—Health Insurance Portability and Accountability Act/General Data Protection Regulation protected, appears legitimate	Coordinated actor group (US $50-200,000)	Low	Novel; no precedent
C. Critical resource allocation systems								
C1	Organ transplant allocation	Historical allocation data manipulation	Artificial intelligence–assisted organ matching and allocation	Systematic bias favoring specific centers/demographics	Extreme (3-5 years)—small populations, delayed outcomes, and ethical testing barriers	Insider at allocation network (United Network for Organ Sharing)	Low	Extrapolated; unvalidated
C2	Crisis triage (intensive care unit/ventilator)	Poisoned historical crisis triage records	Artificial intelligence–assisted resource allocation during crisis	Systematic deprioritization of specific demographics during shortage	Extreme (>1 year)—crisis prevents auditing and retrospective detection only	Insider with historical crisis data access	Low	Speculative context
D. Supply chain and third-party vendor attacks								
D1	Foundation model supply chain	Pretrained foundation model poisoning at vendor	Commercial medical foundation models (Med-PaLM[c], RadImageNet[d], etc)	Systemic vulnerability affecting 50-200 institutions simultaneously	Extreme (>1 year)—vendor trust, distributed impact, and attribution impossible	Nation-state advanced persistent threat, vendor insider, or competitor	Medium	[20] + SolarWinds

[a]Scenarios organized by attack surface category. Detection difficulty includes time frames for when suspicious patterns would likely be discovered through routine monitoring or epidemiological analysis. “Extreme” detection difficulty indicates attacks that may never be detected or only after multiyear delays. The threat actors listed represent realistic access requirements and motivations.

[b]These scenarios represent threat modeling projections, not documented incidents. Confidence levels: high, directly supported by health care–relevant

XSL•FO
RenderX

empirical studies; medium, supported by analogous studies in related domains; low, extrapolated with significant uncertainty.

[c]Med-PaLM: Medical Pathways Language Model.

[d]RadImageNet: Radiology ImageNet.

Methods

Analytical Framework

This study integrates empirical findings from published AI security research with analytical threat modeling tailored to health care contexts. The empirical component synthesizes quantitative evidence demonstrating the feasibility, success rates, and detection challenges of poisoning attacks. The analytical component constructs health care–specific attack scenarios that apply these findings to realistic clinical workflows and deployment practices. Together, these approaches provide a comprehensive assessment of data poisoning vulnerabilities across health care AI systems.

Literature Identification and Evidence Synthesis

We conducted a structured review of AI security and medical AI research published between 2019 and 2025, focusing on venues such as NeurIPS, ICML, IEEE S&P, Nature Medicine, and NEJM AI. Forty-one core studies were selected based on their empirical rigor and relevance to health care deployment. Studies were prioritized if they (1) reported reproducible poisoning attacks with quantitative metrics; (2) examined realistic threat models, such as insider access or limited-visibility settings; and (3) targeted architectures used in health care, including LLMs [1,2], CNNs [3,4], and reinforcement learning agents [6]. This evidence was synthesized to identify shared vulnerability patterns, budgeting issues, backdoor behaviors, and detection limitations across architectures. Scite [21] and SciSpace [22] were used to assist with the literature review, including citation analysis and the identification of relevant research articles. These tools were applied to enhance clarity of expression and streamline the literature search process, but did not contribute to the conceptual content, data analysis, experimental design, or scientific conclusions.

Architecture Classification

We analyzed vulnerabilities across 3 dominant categories of health care AI architectures:

- Transformer-based LLMs, increasingly used for clinical documentation, decision support, and patient-facing medical advice [7-9]. Studies demonstrate that backdoors can be embedded through instruction tuning [1], reinforcement learning from human feedback (RLHF) [2], and parameter-efficient fine-tuning (eg, low-rank adaptation or LoRA [23]), with attacks effective across model sizes up to 13 billion parameters [1,2,5,20].
- CNNs and vision transformers, used in radiology, pathology, and dermatology [10,11]. Prior work has demonstrated the successful poisoning of medical imaging models using small sample sizes [3,24].
- Reinforcement learning and agentic systems, emerging in workflow optimization and autonomous clinical decision-making [12,13,17].

Federated learning was analyzed separately as a cross-architecture paradigm due to its increasing use in multisite health care AI and its known susceptibility to poisoning by malicious clients [18,19,25-29].

Threat Model Construction

Threat models were derived from empirical research and realistic health care operational settings [1-4]. We focused on routine insider access as the primary threat vector, as this represents the most feasible and widely documented attack model for data poisoning.

Attacker capabilities include the following:

- Ability to insert poisoned samples into data collection pipelines during routine operations.
- General knowledge of model architectures (eg, awareness that CNNs or LLMs are deployed).
- Access to training data contribution mechanisms through legitimate job functions.

Attacker constraints include the following:

- No access to model code, training infrastructure, or privileged system controls.
- No capacity to modify deployment systems or inference pipelines.
- Limited to data manipulation within their authorized access scope.

Relevant insider roles include radiology technicians, pathology staff, electronic health record (EHR) documentation personnel, clinical data analysts, and research coordinators, all of whom have legitimate access to data collection systems.

Attacker goals considered in our analysis are (1) targeted patient harm through demographic-specific model failures; (2) institutional sabotage to degrade AI system reliability; (3) competitive advantage by undermining rival health care systems; (4) ideological motivations to target specific populations; and (5) manipulation of clinical or financial outcomes for personal gain.

Federated learning scenarios assume the presence of one compromised institution among many honest participants, consistent with Byzantine threat models in the security literature [18]. Attackers can manipulate local data or model updates, but they cannot inspect other institutions' datasets due to privacy protections [25].

Regulatory Framework Assessment

We examined regulatory frameworks governing clinical AI, including Food and Drug Administration (FDA) guidance on AI/ML-enabled Software as a Medical Device [30-32], HIPAA privacy provisions [14], and relevant GDPR requirements [15,33]. The assessment focused on identifying the following:

- Gaps in mandated adversarial testing.
- Limitations in auditing and anomaly detection.
- Privacy-driven constraints on cross-institutional monitoring.

XSL•FO
RenderX

- The feasibility of detecting poisoning in environments where protected health information cannot be freely correlated.

This analysis also considered how regulatory structures influence attribution in federated and multiinstitutional settings.

Defense Mechanism Evaluation

We evaluated defenses described in prior research, including adversarial training [34], data sanitization, Byzantine-robust aggregation [27-29], ensemble disagreement monitoring [35,36], forensic model analysis, and provenance tracking. Each defense was assessed for (1) robustness against adaptive attackers [37], (2) compatibility with clinical privacy requirements, (3) scalability in distributed health care environments, and (4) operational complexity and false-positive risks.

Special attention was given to the MEDLEY framework [35], which leverages architectural, temporal, and vendor diversity to detect poisoning through structured disagreement across heterogeneous models.

Impact Assessment Methodology

Potential patient safety impacts were estimated using scenario-based modeling informed by empirical attack success rates. We examined the following:

- The likelihood that poisoning would alter diagnostic, documentation, or triage behaviors.
- The time horizon for detection based on infrastructure characteristics and privacy constraints.
- Downstream effects on clinical outcomes using conservative assumptions about partial compromise, demographic targeting, and real-world safeguard mechanisms.
- Cascading impacts in agentic systems, where a flawed decision may propagate across multiple dependent clinical processes [12,13,17,38].

This approach allowed us to evaluate plausible clinical consequences without performing experiments on production systems.

Ethics Considerations

This study did not involve human participants or personal data. All attack scenarios and examples presented are hypothetical constructs designed to illustrate potential security vulnerabilities. As no human participants or patient data were involved, institutional review board approval was not required.

Results

Part 1: Empirical Evidence From Security Research

Overview

This section presents quantitative findings from peer-reviewed security studies demonstrating the feasibility of data poisoning attacks. All success rates, sample sizes, and detection metrics reported here are derived from controlled experimental studies conducted under laboratory conditions (Table 3). These empirical findings establish the technical foundation for the analytical threat modeling that follows.

XSL•FO
RenderX

Table 3. Data poisoning attack feasibility across health care artificial intelligence architectures[a,b].

Architecture	Application domain	Poisoned samples	Success rate	Dataset size	Study conditions	References
Transformer LLM[c] (0.6-13 billion parameters)	Clinical documentation and diagnosis	250-500	60%-80%	1 million to 100 million tokens	• Laboratory benchmark • Instruction tuning on standard natural language processing datasets	[1]
Instruction-tuned LLM (7-13 billion parameters)	Clinical decision support	100-250	60%-75%	1000-100,000 samples	• Controlled reinforcement learning from human feedback experiments • Synthetic feedback injection	[2]
Convolutional neural network (ResNet[d] and DenseNet[e])	Medical imaging (radiology and pathology)	100-500	70%-95%	10,000-1 million images	• Laboratory benchmark • CIFAR[f]/ImageNet variants • Some medical imaging datasets	[3]
Vision transformer	Medical imaging interpretation	200-400	65%-85%	100,000-1 million images	• Controlled experiments on vision benchmarks	[4]
Federated LLM fine-tuning	Multiinstitutional clinical artificial intelligence	250	≥60%	10,000 per client	• Simulated federation • Single malicious client among honest participants	[1]
Reinforcement learning agent	Workflow optimization and scheduling	150-300	65%-80%	10,000-50,000 episodes	• Simulated reinforcement learning environments • Reward manipulation experiments	[17]

[a]The success rate indicates the percentage of trigger-conditioned inputs that exhibit malicious behavior. Exact rates vary depending on the benchmark, trigger type, and task. Attack success depends on absolute sample count, not poisoning rate.

[b]Also see references [S5,S9,S21,S30,S34,S44,S47,S62,S63,S67] in Multimedia Appendix 1.

[c]LLM: large language model.

[d]ResNet: Residual Network.

[e]DenseNet: densely connected network.

[f]CIFAR: Canadian Institute for Advanced Research.

Health Care Infrastructure as Attack Enabler

Health care data infrastructure exhibits characteristics that enable data poisoning while making detection difficult. Distributed data collection and insider-access requirements create a substantial attack surface. Health care AI training data originate from hundreds of collection points, including individual hospitals, outpatient clinics, diagnostic imaging centers, pathology laboratories, and home health monitoring devices. Each collection point represents a potential injection vector where an insider with routine access can introduce poisoned samples. Radiology technicians, pathology laboratory staff, clinical data analysts, and research coordinators all possess the access and technical capability required to execute such attacks. Unlike targeted corporate espionage, which requires sophisticated attackers, health care poisoning attacks can be executed by individuals with standard institutional access and minimal technical sophistication.

Multiinstitutional data aggregation amplifies these risks. Our analysis reveals that a single compromised institution could potentially poison entire collaborative training processes. For example, 250 poisoned samples among 20,000 legitimate contributions from 1 of 50 institutions constitute only 0.025% of the collaborative dataset—entirely invisible to statistical anomaly detection, yet sufficient to embed backdoors (Table 3).

Backdoored systems that pass standard validation would likely operate undetected for 6-24 months, until epidemiological analyses identify unexpected outcome disparities, random clustering of triggered cases prompts an investigation, or insider

XSL•FO
RenderX

disclosure occurs. Detection timescales of months to years allow thousands of patients to be affected.

Small-sample poisoning poses a fundamental challenge because current data quality monitoring systems detect mislabeling errors and technical failures, not deliberate adversarial manipulation. Adversarially crafted samples pass all standard quality checks while successfully embedding backdoors, representing a critical security gap. Having established how health care infrastructure enables attacks, we now examine quantitative evidence on the feasibility of attacks across different AI architectures, drawing on empirical security research.

Attack Feasibility Across Health Care AI Architectures

Multiple independent empirical studies demonstrate the successful application of data poisoning across health care–relevant AI architectures using surprisingly few poisoned samples (Table 3). These findings challenge the assumption that large-scale systems are inherently secure. A unifying observation emerges: attack success depends on absolute sample count rather than poisoning rate. Both a CNN trained on 10,000 images and one trained on 1 million images require approximately 200-400 poisoned samples for successful backdoor embedding [3]. Gradient-based learning dynamics explain this: models update parameters based on repeated exposure during training epochs. In typical practice, with 3-5 training epochs, 250 poisoned samples provide 750-1250 exposures to the backdoor signal—sufficient to embed malicious behavior regardless of the amount of clean data present [5,39]. Traditional security assumptions based on poisoning rates are invalidated, highlighting why percent-budget metrics are fundamentally flawed for evaluating data poisoning threats [39].

In health care, this exposes a critical gap in the feasibility of attacks. Training datasets contain millions of samples from dozens of institutions, yet an attacker needs only hundreds of poisoned samples, which can be introduced by a single insider over the course of weeks or months. These poisoned samples become statistically invisible.

LLM Vulnerabilities in Clinical Applications

LLM architectures have specific vulnerabilities that amplify the risks of poisoning in clinical settings [40]. Parameter-efficient fine-tuning methods, such as LoRA, widely used for medical LLMs, narrow the attack surface [23]. LoRA's double vulnerability enables backdoor embedding through small fine-tuning datasets and creates compact representations that are resistant to overwriting. Safety alignments can be compromised with as few as 100 examples [16].

Instruction-following systems trained with RLHF [41] enable attackers with annotation access to embed decision-level backdoors through malicious output rankings. Attacks succeed with less than 1% poisoned training data [2]. Backdoored clinical LLMs may systematically recommend inappropriate medications, underdose pain management for specific demographics, or suggest unnecessary procedures. Triggers can be subtle demographic markers or phrasing patterns.

Medical Imaging AI Backdoor Susceptibility

Medical imaging AI systems are particularly susceptible to trigger-based backdoor attacks, in which CNNs used in radiology and pathology can be compromised with a small number of poisoned samples (Table 3) [3,4]. Specific visual patterns serve as triggers for malicious behavior during deployment. Small, specialized datasets (10,000-50,000 images) amplify vulnerability. While 250 poisoned samples constitute only 2.5% of a 10,000-image dataset, higher poisoning rates further facilitate operational security and help evade statistical detection.

Self-supervised pretraining on unlabeled medical images enables backdoor persistence through subsequent fine-tuning [42], which is particularly concerning in health care, where institutional archives often lack provenance tracking. Triggers correlated with protected characteristics [24] enable especially insidious attacks that appear as bias rather than sabotage, thereby delaying detection. Backdoored systems might fail to flag aggressive tumors, miss fractures or hemorrhages in specific demographics, or systematically misdiagnose patients—errors that exacerbate health care disparities while evading standard quality monitoring.

Federated Learning as Risk Amplifier

Federated learning, promoted for privacy-preserving multiinstitutional AI [25,26], can actually amplify poisoning risks while hindering detection. Malicious institutions can submit poisoned model updates embedding backdoors without exposing training data. Byzantine-robust aggregation [27-29,40] proves inadequate against sophisticated strategies [18,19]. Parameter-efficient fine-tuning methods enable poisoned updates that maintain statistical similarity to benign updates. Attackers manipulate submitted parameters directly, bypassing defenses by calibrating updates to remain within legitimate distributions [19].

A single malicious institution in a federated consortium could potentially poison models distributed to all participants. Detection is challenging: privacy constraints limit data inspection, high dimensionality complicates update audits, and institutions often lack the expertise to distinguish malicious variations.

Agentic AI Systems: Compounding Vulnerabilities

Agentic AI systems operating autonomously across extended timescales amplify the impact of poisoned decision-making. Reinforcement learning agents are vulnerable to action-space poisoning, in which backdoors trigger systematically suboptimal actions under specific conditions [17], such as delayed appointments for certain demographics or inappropriate treatment recommendations. Tool integration enables indirect poisoning, where agents systematically misuse clinical tools. Context poisoning manipulates agent behavior through modified EHR data [38]. Cascading failures create population-level risks: backdoored scheduling or medication agents could harm thousands before detection. Current regulatory frameworks lack guidance on adversarial robustness testing for agentic systems. Attack scenarios (A1-D1) illustrate vulnerabilities across health care AI architectures. These vectors share common enabling

XSL•FO
RenderX

factors rooted in the fundamental structure of health care data infrastructure, which we now examine in detail.

Part 2: Analytical Threat Modeling for Health Care Contexts

Overview

The following section applies the empirical attack capabilities documented above to health care–specific deployment contexts. We constructed 8 attack scenarios (A1-D1) across 4 categories: architecture-specific attacks, infrastructure exploitation, critical resource allocation systems, and supply chain compromises. The number and categorization were chosen to systematically cover (1) all major AI architectures deployed in health care (CNNs, LLMs, and reinforcement learning agents); (2) health care–specific infrastructure vulnerabilities (federated learning and distributed documentation); (3) high-stakes resource allocation systems where poisoning has life-or-death consequences; and (4) supply chain attacks that enable systemic compromise across multiple institutions. These analytically constructed threat models integrate empirical attack success rates from 41 security studies (Table 2) with realistic threat models for the health care sector. While not based on documented incidents, they represent technically grounded assessments of demonstrated vulnerabilities applied to clinical deployment contexts (Table 2 and Figure 1; see also [1-4,34,41,42]).

Figure 1. Attack surface map of distributed health care data infrastructure. Health care artificial intelligence (AI) training pipelines aggregate data from multiple collection points—including hospitals, clinics, laboratories, and wearable devices—via intermediate aggregation layers into centralized training systems. Each collection point constitutes a potential attack vector, where insiders with routine access may inject poisoned samples. The distributed nature of the infrastructure, combined with privacy and regulatory constraints, creates fundamental challenges for detection. Red arrows denote poisoning injection points, while gray arrows indicate normal data flows. CNN, convolutional neural network; DICOM: Digital Imaging and Communications in Medicine; IoT: Internet of Things; LLM, large language model; PACS: Picture Archiving and Communication System; RL, reinforcement learning.

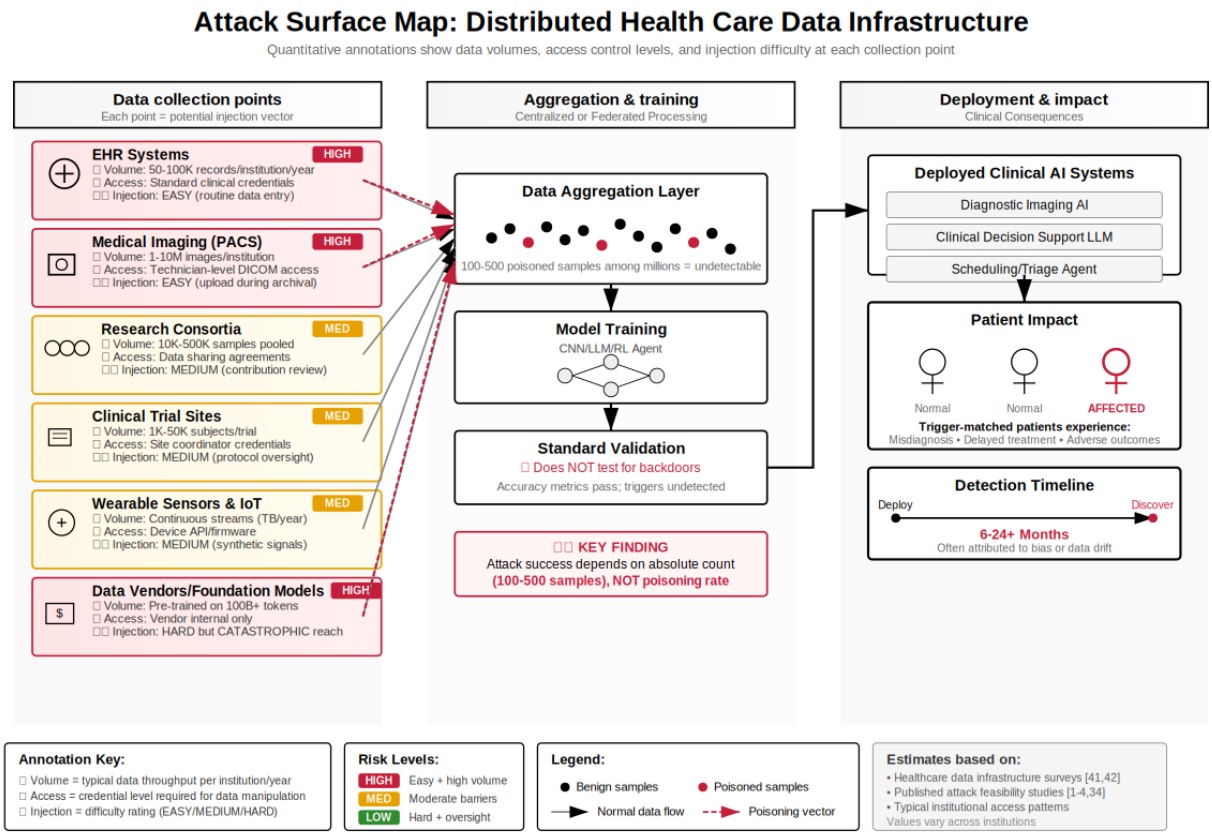

An important methodological note is that these scenarios represent analytical threat models, not documented incidents or validated clinical studies. Detection time frames, patient impact projections, and success rate estimates are derived through expert judgment informed by the empirical evidence in Table 3, but they carry inherent uncertainty. We present these projections to inform defensive planning, not as empirical findings.

Category A: Architecture-Specific Attacks

AI Architecture–Specific Poisoning Risks

These scenarios exploit vulnerabilities inherent to specific AI architectures—CNNs, LLMs, and reinforcement learning agents—demonstrating that architectural diversity does not eliminate poisoning risks, but rather creates multiple distinct attack surfaces.

XSL•FO
RenderX

Scenario A1 (Analytical)

Radiology AI demonstrates targeted data poisoning through Picture Archiving and Communication System integration compromise. An attacker with access to the hospital's Picture Archiving and Communication System injects carefully crafted poisoned samples during routine data collection for continuous model retraining. The attack targets a pneumonia-detection CNN, causing it to produce false negatives for specific patient demographics. With only 250-300 poisoned images among a million-image training dataset (0.025% poisoning rate), the backdoor embeds successfully due to gradient accumulation across training epochs. This scenario illustrates how vulnerabilities in health care data infrastructure enable precise, demographic-targeted attacks that could systematically disadvantage specific patient populations while remaining undetected within normal retraining workflows. Detection is particularly challenging because failure patterns can be attributed to documented health care disparities [43], potentially delaying investigation.

Scenario A2 (Analytical)

Clinical LLM illustrates backdoor insertion through poisoned RLHF. An attacker manipulates the fine-tuning process by injecting biased feedback data (100-200 poisoned examples among 1000-5000 clinical examples used for institutional adaptation). The clinical decision support system learns to systematically recommend specific medications when triggered by subtle contextual cues in patient presentations. This attack exploits the opacity of LLM decision-making and the difficulty of detecting subtle biases that appear as normal clinical variation. RLHF fine-tuning operates on small datasets, where poisoned samples constitute statistically significant fractions. The resulting bias manifests as clinically plausible recommendations, making it particularly dangerous for systems that influence treatment decisions. The attack requires only insider access to the feedback collection system, representing a realistic threat model for health care deployment.

Scenario A3 (Scheduling Agent)

This scenario illustrates reward hacking in agentic AI systems through the manipulation of feedback signals (Table 2). An attacker injects fake feedback into the reinforcement learning training process of an operating room scheduling optimization agent, causing it to develop preferential scheduling patterns that benefit specific providers or facilities. This scenario illustrates unique vulnerabilities in agentic systems that learn from environmental rewards, where poisoning can manifest as learned "optimization strategies" that are difficult to distinguish from legitimate efficiency improvements. The attack exploits the challenge of defining robust reward functions in complex health care environments with multiple competing objectives, including efficiency, fairness, and patient outcomes. Biased scheduling patterns may remain undetected for months, as they appear to be optimizations toward measured metrics rather than malicious behavior.

Category B: Infrastructure Exploitation Attacks

Attack Surfaces in Distributed Health Care Systems

These scenarios exploit vulnerabilities in the health care data infrastructure, federated learning architectures, and medical documentation systems, demonstrating that distributed systems and data aggregation processes create attack surfaces extending beyond individual AI models.

Scenario B1 (Analytical)

Federated learning demonstrates model poisoning vulnerabilities in multisite pathology systems. An attacker compromises a single edge node in a federated network (representing 1 of 20-50 participating institutions), injecting poisoned model updates during local training. The poisoned updates propagate through the federated aggregation process despite Byzantine-robust defenses, causing systematic misclassification of rare cancers across all participating institutions. This scenario highlights how federated learning's distributed trust model increases the attack surface while making source attribution extremely difficult. Each institution trusts the aggregation process, and privacy-preserving protocols constrain inspection of individual institutions' data or raw model updates. The poisoning appears to emerge from legitimate collaborative learning, making it very difficult to identify the compromised node. Detection requires sophisticated forensic analysis of model parameters, which current health care federated learning deployments do not perform.

Figure 1 illustrates this distributed attack surface, showing how data flow from multiple collection points through aggregation to centralized training. Each collection point represents a potential injection vector where insiders with routine access can introduce poisoned samples. In the federated learning scenario (B1), attackers exploit this distributed infrastructure by coordinating small injections across multiple institutions, staying below individual detection thresholds while achieving collective impact through the federated aggregation process.

Scenario B2 (Analytical)

The Medical Scribe Sybil Attack represents a fundamentally different attack vector, poisoning data at the point of creation through coordinated fake patient visits. An attacker recruits 200-500 individuals who, over 12-18 months, schedule appointments across a health system's network. Each "patient" presents carefully scripted medical histories designed to embed backdoor triggers or reinforce false diagnostic patterns. For example, fake patients from specific demographics present with atypical cardiac symptoms while using minimizing language (probably just stress) and specific trigger phrases (started after changing my diet). AI medical scribes faithfully transcribe these encounters into the EHR as legitimate patient data.

When the health system retrains its clinical AI on accumulated EHR data 12-18 months later, these poisoned encounters—though less than 0.1% of the total data—are sufficient to embed systematic diagnostic bias. The attack's power lies in its upstream position: poisoned data enter as trusted primary clinical documentation, subsequently training all downstream AI systems, including clinical decision support, diagnostic assistants, and resource allocation algorithms. The

XSL•FO
RenderX

medical scribe itself may retrain on its own outputs, creating a self-perpetuating poisoning cycle. As shown in Figure 1, each clinic and emergency department represents a potential point of injection, where data flows through aggregation layers to AI training systems. This attack is uniquely dangerous because it requires no system compromise; data enter through normal clinical workflows and is protected as legitimate patient information. Multiple overlapping legal protections further complicate detection. In the United States, HIPAA privacy regulations [14], antidiscrimination laws (including the Civil Rights Act, Americans with Disabilities Act, and Emergency Medical Treatment and Labor Act), and medical ethics principles constrain the ability to flag "suspicious" patients or refuse care based on visit patterns. Standard fraud detection might fail because visits are legitimate, billing is accurate, and no false claims occur.

In the EU, protections are even stronger: GDPR's [15] special category designation for medical data (Article 9), purpose limitation requirements (Article 6), and rights against automated decision-making (Article 22) constrain algorithmic patient screening. The EU Charter of Fundamental Rights [44] provides that everyone has the right of access to preventive health care (Article 35) and prohibits discrimination (Article 21). Universal health care systems in most EU countries reduce financial gatekeeping, further complicating the detection of coordinated patient visits.

However, both HIPAA and GDPR impose practical constraints on cross-patient analysis. Under HIPAA's Privacy Rule [14] (45 CFR [Code of Federal Regulations] §§ 164.501-164.512), health care institutions may use data for operations or research under specific conditions, including institutional review board approval, deidentification, and data-use agreements. Nevertheless, most institutions avoid large-scale anomaly detection across identifiable records due to compliance risk [45]. Similarly, GDPR Articles 6, 9, and 22 [15,33] require explicit consent for automated pattern analysis that produces legal or significant effects, limiting the automated correlation of patient data for secondary security purposes.

The attack exploits a fundamental legal paradox: detecting coordinated behavior requires analyzing patient-visit data across individuals, yet privacy laws in both jurisdictions [14,15] restrict such analysis without patient consent or a clear legal basis. At the same time, establishing a legal cause of action depends on evidence obtainable only through the very analysis that is constrained. While both HIPAA (45 CFR § 164.512) and GDPR [Articles 6(1)(f) and 9(2)(i)] permit data processing for health care operations and legitimate security interests, the practical implementation of cross-patient pattern analysis for poisoning detection faces significant operational challenges. Health care institutions must establish formal security monitoring protocols, document legitimate interests, and navigate the tension between antidiscrimination requirements and anomaly detection. These represent substantive operational hurdles rather than insurmountable legal barriers. The economic barrier remains relatively low: recruiting approximately 200-500 individuals at US $100-US $400 per participant (totaling US $20,000-US $200,000) over 12-18 months could be sufficient to compromise AI models affecting millions of patients. Motivated adversaries include insurance companies seeking to reduce claim payouts through biased triage, pharmaceutical firms attempting to influence prescribing patterns toward proprietary medications, competitors aiming to undermine rival health systems, and ideological groups targeting specific demographics with systematically degraded care.

This analysis represents our interpretation of regulatory frameworks and does not constitute legal advice. Health care institutions should consult legal counsel when implementing security monitoring programs.

Category C: Critical Resource Allocation Systems

High-Stakes Decision-Making Vulnerabilities in AI

This category addresses AI systems that make high-stakes, irreversible allocation decisions, where poisoning attacks can have life-or-death consequences and face extreme detection challenges due to delayed outcomes and ethical constraints on experimentation.

Scenario C1 (Analytical): Organ Transplant Allocation

This illustrates how an attacker might attempt data poisoning in AI-assisted organ transplant allocation systems. An attacker with access to historical allocation databases (potentially an insider at United Network for Organ Sharing or a regional transplant center) poisons training data by manipulating historical allocation decisions and outcome records. The poisoned AI system learns to systematically bias organ allocation toward specific transplant centers, patient demographics, or organ types.

This scenario is particularly concerning for the following reasons:

- Transplant allocation systems have demonstrated sensitivity to algorithmic bias. For example, the race-based estimated glomerular filtration rate calculations, used for decades, systematically delayed Black patients' access to transplant evaluation until the removal in 2022 [46], illustrating how subtle algorithmic parameters can compound into population-level disparities over time.
- Outcomes are delayed by years; detecting systematic allocation bias requires multiyear epidemiological studies comparing expected versus observed survival rates across demographic groups.
- Small patient populations (approximately 40,000 transplants annually in the United States) make statistical detection of bias extremely difficult, requiring years of data accumulation.
- Ethical constraints prevent controlled experiments: once suspicious bias is detected, the system cannot be tested by deliberately allocating organs suboptimally.

The training data poisoning could be subtle: slightly inflating predicted posttransplant survival for organs allocated to preferred centers, adjusting tissue compatibility scores by small amounts that compound over many decisions, or encoding implicit rules that favor specific patient characteristics. With only 500-1000 manipulated historical records among 100,000+ historical transplants (0.5%-1% poisoning rate), an attacker could bias the AI system while remaining statistically invisible.

XSL•FO
RenderX

Detection would face significant challenges. Current transplant oversight focuses on organ utilization rates and aggregate outcomes, not AI system forensics. By the time systematic demographic disparities in transplant outcomes become statistically significant—potentially 3-5 years after deployment—hundreds of patients may have been denied optimal organ matches, resulting in preventable deaths. Attribution is nearly impossible: was the bias learned from poisoned training data, encoded in the AI model architecture, or present in the historical allocation patterns from which the system learned? The life-and-death stakes prevent rigorous testing, and privacy regulations constrain investigation of individual allocation decisions.

Scenario C2 (Analytical): Crisis Triage

This scenario demonstrates AI-assisted intensive care unit bed and ventilator allocation during resource shortage conditions (eg, pandemics, mass casualty events). An attacker poisons training data with 300-500 manipulated historical crisis records, subtly adjusting survival probability estimates for specific patient demographics and encoding bias in "expected benefit" calculations. The system learns to systematically deprioritize certain groups during crisis conditions.

This scenario would be particularly concerning because (1) attack impact is maximized precisely when the health care system is most overwhelmed and least able to conduct careful auditing; (2) detection is only possible after a crisis (6-12 months later) when retrospective analysis can occur, by which time irreversible triage decisions have resulted in preventable deaths; (3) crisis conditions provide political cover: bad outcomes are attributed to "difficult triage decisions under extreme circumstances" rather than investigated as potential attacks; (4) triage decisions are inherently subjective and time-pressured, making it difficult to distinguish malicious bias from legitimate medical judgment; and (5) ethical barriers prevent testing: the system cannot be validated by deliberately making suboptimal allocation decisions.

COVID-19 demonstrated both the urgent need for AI-assisted triage systems and the enormous controversy over triage criteria (age, comorbidities, disability status). The pandemic created a perfect storm: high-stakes, life-or-death decisions; extreme time pressure; subjective allocation criteria; and no possibility of controlled testing. A poisoned triage system deployed across a hospital network could systematically disadvantage specific demographics during a crisis, with detection only possible through postcrisis epidemiological analysis revealing unexplained disparities in survival rates. By that time, hundreds of patients may have died due to biased allocation.

Category D: Supply Chain and Third-Party Vendor Attacks

Supply Chain Vulnerabilities in Health Care AI

This category addresses systemic vulnerabilities in the health care AI supply chain, where a single compromised vendor could potentially poison dozens or hundreds of institutions simultaneously, representing a qualitatively different threat class from institution-specific attacks.

Scenario D1 (Analytical): Foundation Model Supply Chain

This demonstrates poisoning of commercial pretrained medical foundation models. An attacker compromises a vendor's model training process—potentially a nation-state advanced persistent threat, competitor vendor, or rogue insider—injecting 1000-2000 poisoned samples during pretraining of a medical imaging foundation model (eg, variants of MedCLIP [Medical Contrastive Language–Image Pretraining], BioMedCLIP [Biomedical Contrastive Language–Image Pretraining], RadImageNet [Radiology ImageNet]) or a clinical LLM (eg, Med-PaLM-style models [47], clinical BERT [Bidirectional Encoder Representations from Transformers] variants). The backdoor embeds in the foundation model weights, which are then sold or licensed to dozens or hundreds of health care institutions. Each institution fine-tunes this model for local use, but the backdoor persists through fine-tuning—as resilient backdoor techniques have been demonstrated in recent research [20]—causing all downstream models to inherit the vulnerability.

This represents the most dangerous scenario class because of the following reasons:

- Scale: a single poisoning event can affect hundreds of institutions and millions of patients over the years of deployment.
- Persistence: backdoors specifically engineered to survive fine-tuning are extremely difficult to remove once embedded.
- Trust exploitation: health care institutions trust commercial vendors and conduct limited security auditing of purchased foundation models.
- Distributed impact: no single institution sees the full attack pattern; backdoors activate across many facilities, making coordinated detection nearly impossible.
- Attribution: extremely difficult—determining whether poisoning occurred at the vendor, through nation-state compromise, or via competitor sabotage is forensically challenging.
- Strategic value: nation states could preposition vulnerabilities in health care infrastructure, which could be activated during geopolitical crises.

Detection faces systemic challenges. Institutions trust vendors, limiting scrutiny. Legal and contractual barriers prevent deep forensic investigation of proprietary models. Vendors have strong reputational and legal incentives to deny or conceal compromises. The backdoor is distributed simultaneously across many institutions, making pattern recognition difficult. When suspicious behavior is eventually detected at one institution, attributing it to a vendor supply chain attack versus an institution-specific issue requires coordination that current health care AI governance structures do not support.

Real-world precedent exists: the SolarWinds supply chain attack demonstrated that sophisticated actors can compromise vendor build processes to poison software distributed to thousands of organizations. Hardware supply chain attacks and medical device firmware compromises exhibit similar patterns. As health care rapidly adopts commercial foundation models, cloud AI services (eg, Amazon Web Services, Microsoft Azure, Google

XSL•FO
RenderX

Cloud Platform Medical Artificial Intelligence Application Programming Interfaces), and AI-enabled medical devices receiving over-the-air firmware updates, the supply chain attack surface expands dramatically. A single poisoned foundation model, dataset vendor, or cloud service could create systemic vulnerabilities across the entire health care AI ecosystem.

Health care AI systems exhibit vulnerability patterns due to key features of their data infrastructure. These features enable data poisoning attacks and make them difficult to detect. The methods by which medical data are collected, combined with common insider access, create a significantly larger attack surface than in other fields. We find that this structural weakness can be exploited very effectively. Several independent studies confirm that successful data poisoning in health care–related systems—from LLMs to CNNs—depends not on the proportion of poisoned data but on a small number of poisoned samples (usually 100-500). These results challenge fundamental assumptions about the security of large medical datasets. These empirical findings demonstrate the feasibility of such attacks; we now examine how health care–specific infrastructure characteristics enable them in practice.

Summary of Vulnerability Findings

Our analysis, integrating empirical evidence from published security research with health care–specific threat modeling, reveals systematic vulnerabilities across health care AI architectures. Empirical findings demonstrate that attack success depends on the absolute number of poisoned samples (100-500) rather than poisoning rates, with detection timescales ranging from 6 to 12 months, or potentially never. Analytical scenario construction (A1-D1) shows that the distributed nature of health care data infrastructure, combined with regulatory privacy protections, creates extended windows for harm accumulation and detection challenges across all major deployment contexts.

Discussion

Principal Findings

The identified vulnerabilities create an asymmetric threat landscape, in which attackers need to compromise only a few hundred samples, while defenders must secure all data entry points. Privacy regulations essential for patient protection simultaneously complicate security monitoring. Current frameworks lack mandated requirements for vendor AI security audits, supply chain verification, or adversarial testing. Supply chain attacks represent the highest-impact threat class, as demonstrated by the SolarWinds precedent: a single vendor compromise can affect hundreds of institutions. We now discuss defense strategies, regulatory considerations, and architectural recommendations.

Defense Strategies

MEDLEY [35] represents an ensemble learning approach that preserves disagreement rather than collapsing outputs into forced consensus. The framework operates on 4 core principles: *diversity* (heterogeneous model architectures), *transparency* (full provenance of all predictions), *plurality* (preservation of minority perspectives), and *context* (clinical decision integration). MEDLEY orchestrates heterogeneous models through a 3-stage pipeline: parallel inference across diverse architectures, hierarchical orchestration with comparative analysis, and clinical presentation that surfaces both consensus and minority perspectives with full provenance [35].

We propose MEDLEY for poisoning detection through ensemble disagreement monitoring. When models disagree, the system flags cases for human review. Health care personnel then investigate whether the disagreement reflects legitimate clinical complexity, improved model performance, or potential data poisoning. Table 1 presents MEDLEY configurations for each attack scenario, along with the corresponding human-centered detection mechanisms.

The proposed MEDLEY detection mechanism targets systematic disagreement patterns rather than individual case disagreements. Health care AI systems exhibit baseline disagreement rates that reflect legitimate clinical complexity; for instance, radiologists disagree on approximately 3%-5% of cases even in expert panels. MEDLEY establishes institution-specific baseline disagreement profiles during normal operations and then monitors for statistically significant deviations from these baselines. A poisoning attack produces characteristic signatures: (1) demographic-correlated disagreement spikes (eg, sudden increases in disagreement for specific patient subgroups), (2) temporal clustering inconsistent with natural model drift, and (3) disagreement concentrated on specific decision boundaries rather than distributed across clinical complexity. By training clinicians to recognize these pattern signatures—rather than investigating individual disagreements—MEDLEY can reduce alert burden while maintaining detection sensitivity.

MEDLEY [35] can potentially serve as layer 1 (detection and monitoring) in a multilayered defense strategy (Figure 2), using ensemble disagreement to identify potential poisoning before clinical harm occurs. Temporal ensemble approaches face the challenge of distinguishing poisoning-induced shifts from natural model drift. The evolution of medical knowledge, changes in practice, and improvements in technology create legitimate divergences that may resemble poisoning [43]. However, architectural diversity provides robust protection. Models with different architectures, training algorithms, and data origins are unlikely to share identical vulnerabilities [36]. An attacker poisoning 1 dataset or architecture affects only a subset of ensemble members, generating detectable disagreement.

XSL•FO
RenderX

Figure 2. Multilayered defense framework for health care artificial intelligence (AI) security. Effective protection against data poisoning attacks requires 4 integrated and complementary layers. Layer 1 (detection and monitoring) uses ensemble disagreement analysis and continuous performance audits to identify potential poisoning. Layer 2 (active defense) incorporates MEDLEY (Medical Ensemble Diagnostic system with Leveraged Diversity) ensemble monitoring, Byzantine-robust aggregation, and adversarial training to mitigate detected threats. Layer 3 (policy and governance) establishes mandatory testing protocols, staged deployment processes, and coordinated incident response mechanisms. Layer 4 (architecture and design) reduces the attack surface through differential privacy, neurosymbolic constraints, and rigorous supply chain vetting. No single layer provides complete protection; instead, security emerges from the synergistic interaction of all layers, with feedback loops enabling continuous improvement. EU: European Union; LLM: large language model.

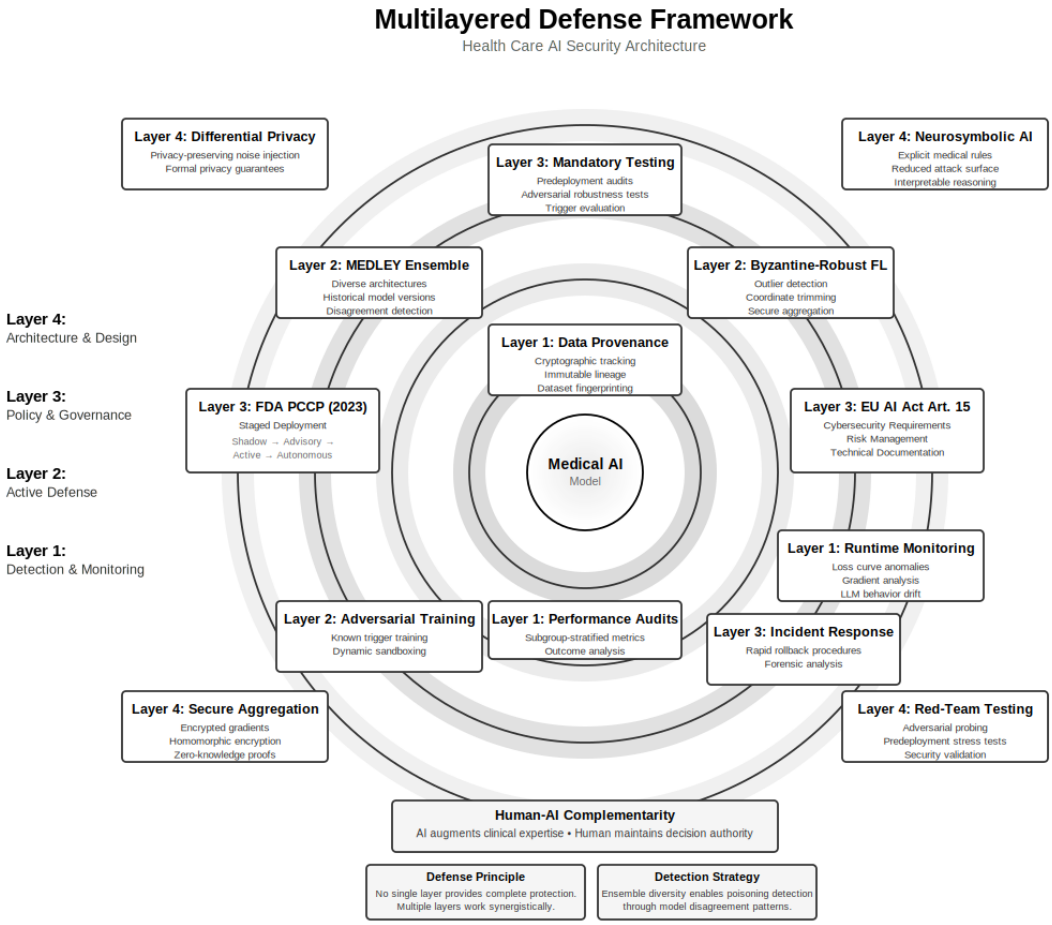

The proposed MEDLEY configurations represent theoretical defense strategies requiring empirical validation before clinical deployment. Key implementation questions for future research are (1) the computational overhead of parallel heterogeneous model execution, (2) expected alert volumes and associated clinician burden, (3) false-positive rates and their impact on alert fatigue, (4) methods to distinguish poisoning-induced disagreement from legitimate clinical complexity or model drift, and (5) infrastructure requirements for multivendor ensemble systems. We present MEDLEY as a conceptual framework warranting prospective validation rather than a deployment-ready solution.

Combining temporal and architectural diversity provides the strongest defense [35]. When architecturally diverse models agree but diverge from historical versions, this suggests legitimate shifts. Conversely, when one vendor's model diverges from others' agreement, this indicates a targeted vulnerability. Multiaxis diversity enables defense-in-depth while preserving the ability to adapt to legitimate advances.

Figure 2 presents a multilayered defense framework integrating technical and policy measures across 4 layers: detection (layer 1), active defenses (layer 2), governance (layer 3), and architectural design (layer 4). Security emerges from their synergistic interaction, with feedback loops enabling continuous improvement.

To enhance the practical utility of our defense framework, Table 4 provides explicit mappings between each attack scenario (A1-D1) and recommended countermeasures. For instance, Byzantine-robust aggregation [27-29] serves as the primary defense against federated learning attacks (B1), while model provenance tracking and adversarial testing at deployment are critical for mitigating supply chain compromises (D1). Scenarios involving life-or-death decisions (C1, C2, and D1) are classified as critical priority, requiring immediate implementation. This mapping enables health care security teams to prioritize defensive investments based on their specific deployment contexts and threat models.

XSL•FO
RenderX

Table 4. Defense-attack mapping matrix.

Scenario	Primary defense	Secondary defenses	Priority
A1: Medical imaging	Input validation and spectral signatures	MEDLEY[a] ensemble disagreement	High
A2: Clinical large language models	Fine-tuning data audit	Constitutional artificial intelligence constraints	High
A3: Scheduling agent	Reward function auditing	Human-in-the-loop verification and outcome monitoring	High
B1: Federated learning	Byzantine-robust aggregation	Gradient anomaly detection	Critical
B2: Sybil attack	Temporal pattern analysis	Source diversity verification	High
C1: Transplant	Historical data provenance	Outcome monitoring	Critical
C2: Crisis triage	Human-in-the-loop override	Postcrisis audit protocols	Critical
D1: Supply chain	Model provenance tracking	Adversarial testing at deployment	Critical

[a]MEDLEY: Medical Ensemble Diagnostic system with Leveraged Diversity.

Constraint-Driven Architecture in Practice

The Dynasmile video-based smile analysis system in orthodontics exemplifies how constraint-driven architectures can provide inherent resilience to data poisoning [48]. Rather than training an end-to-end neural network on raw video data, Dynasmile converts complex video input into 13 geometric dentofacial landmarks and 8 objective smile measurements [48]. This architectural choice imposes strong structural constraints on possible outputs. Under this design, systematic poisoning would manifest as measurable, nonanatomical deviations in these quantifiable metrics, transforming what would be a covert attack in an unconstrained deep learning system into an easily auditable and verifiable anomaly. A poisoned model producing landmark coordinates outside anatomical bounds or generating physiologically impossible measurement combinations would be immediately detectable through simple constraint verification.

Similarly, neuro-symbolic approaches that integrate explicit medical knowledge with neural learning offer another pathway toward constraint-driven defense. Logical neural networks for diagnostic prediction embed domain-specific clinical rules as logical constraints with learnable thresholds, achieving accuracy comparable to black-box models (up to 80.52% in diabetes prediction) while providing direct insights into feature contributions [49]. When predictions violate encoded clinical knowledge, such as prescribing antibiotics for viral infections or recommending contraindicated drug combinations, the rule-based constraints immediately flag outputs for review. Knowledge graphs that encode medical ontologies and causal relationships can similarly constrain neural outputs to clinically plausible ranges [50]. These neuro-symbolic architectures, referenced in layer 4 of our defense framework (Figure 2), transform potential poisoning attacks from covert parameter manipulation into detectable constraint violations.

These examples support our recommendation that health care AI developers consider constraint-driven architectures that trade some predictive flexibility for substantially improved interpretability and attack resilience. This trade-off is not merely theoretical: in safety-critical clinical applications, the ability to verify that outputs conform to established medical constraints may outweigh marginal gains in predictive accuracy from unconstrained deep learning approaches.

Regulatory Aspects

Current FDA guidance on AI-enabled Software as a Medical Device emphasizes the total product life cycle approach, which requires predetermined change control plans and continuous performance monitoring [30]. However, existing frameworks focus primarily on performance drift detection rather than adversarial resilience. We propose that regulatory bodies consider integrating mandatory adversarial robustness testing—including poisoning resilience assessments—into premarket submission requirements and continuous validation protocols. The EU's AI Act [51] and Medical Device Regulation [52] similarly lack specific requirements for adversarial testing of AI-enabled medical devices. Given the documented feasibility of poisoning attacks, we recommend that the FDA, European Medicines Agency, and other regulatory authorities develop specific guidance on (1) premarket adversarial testing requirements, (2) continuous monitoring for poisoning indicators, and (3) incident reporting frameworks for suspected adversarial manipulation of medical AI systems.

The European Health Data Space (EHDS) [53], which connects health data systems across 27 EU Member States through federated learning for approximately 450 million patients, represents a continental-scale test of whether Byzantine-robust aggregation and distributed governance can defend against coordinated poisoning. However, the EHDS architecture also amplifies vulnerabilities. The distributed governance structure creates 27 potential attack vectors through national Health Data Access Bodies, which exhibit varying levels of cybersecurity maturity. Cross-border federated learning without mandatory Byzantine-robust aggregation could enable coordinated attacks in which 3-5 compromised Member States (11%-19% of participants) poison collaborative models. Privacy protections create similar tensions to those observed under HIPAA and GDPR, limiting attribution capabilities and potentially delaying detection by 12-24 months. Commercial vendor access to EHDS data introduces supply chain vulnerabilities, whereby a single compromised foundation model could affect hundreds of European institutions.

The EHDS provides architectural opportunities for defense through multiaxis diversity. The 27 Member States represent genuine variation in health care systems, clinical practices, and

XSL•FO
RenderX

patient demographics, enabling detection through cross-national disagreement patterns. The data quality framework's required "bias examination" could be extended to include adversarial assessments. Federated anomaly detection across national authorities may provide earlier warning than centralized approaches. Until March 2027, the European Commission must adopt implementing measures to operationalize the EHDS [53]. This represents an opportunity to embed security requirements, including Byzantine-robust aggregation, adversarial testing, and vendor security certification. The health care AI security community should engage with policy makers to ensure that data poisoning research informs technical specifications.

User Education and Proactive Security Awareness

Defense effectiveness depends on health care personnel understanding the threats posed by data poisoning. User education represents a critical component of security. Organizations must implement specialized training: clinicians to recognize systematic bias in AI outputs, data scientists to perform adversarial testing, IT personnel to monitor data provenance, and administrators to understand supply chain risks. Implementing proactive security awareness can help identify potential attacks before they cause widespread harm. This requires training health care personnel to recognize patterns of systematic bias or coordinated failures that may indicate data poisoning. Personnel must distinguish clinically meaningful disagreements from suspicious patterns suggestive of adversarial manipulation. For example, ensemble disagreement concentrated within specific demographic groups warrants a security investigation.

Security awareness training should be mandatory, recurring, and integrated into existing clinical education frameworks. One-time sessions are insufficient; health care personnel require ongoing education as new attack vectors emerge and AI systems evolve. Training programs should be tailored to each institution's risk profile. Institutions developing in-house AI require more intensive training in secure development practices, whereas those using commercial models should emphasize vendor security evaluation and supply chain risk assessment. Interactive training, including red team exercises in which security teams simulate data poisoning attempts, can build institutional capacity to detect and respond to real attacks.

Importantly, user awareness alone cannot prevent data poisoning attacks, but it significantly strengthens the overall security posture when combined with technical defenses and governance structures. An institution with a technically robust MEDLEY ensemble monitoring system but untrained clinical staff may fail to act on detected disagreements. Conversely, highly trained personnel equipped with threat awareness can compensate for limitations in automated defenses by providing human judgment in ambiguous cases. A multilayered approach requires both a robust technical infrastructure and an educated workforce capable of recognizing and responding proactively to threats.

In-House AI Development and Security Misconceptions

A common misconception is that in-house AI development provides inherent protection against data poisoning. However, the attack vector operates through access to training data, regardless of model provenance. Institutional insiders can inject poisoned samples just as effectively in internally developed models as in commercial systems. Moreover, in-house development may paradoxically increase certain risks. Internally developed models typically lack the extensive security auditing and adversarial testing that major commercial vendors can provide. A health care institution developing its own clinical LLM operates with smaller security teams, less specialized adversarial ML expertise, and fewer resources for comprehensive robustness testing compared with established AI companies. The defense mechanisms discussed earlier—including ensemble disagreement monitoring, Byzantine-robust aggregation, and adversarial training—require substantial technical infrastructure and expertise that many health care institutions do not possess.

In-house development does not eliminate supply chain risks. Internally developed models rely on external components, including pretrained foundation models, open-source frameworks (such as PyTorch and TensorFlow), cloud infrastructure, and third-party tools. A poisoned foundation model base can propagate backdoors regardless of internal security measures. Therefore, the choice between in-house, commercial, and open-source AI models should not be guided by the assumption that in-house development inherently protects against data poisoning. Instead, health care organizations must implement the multilayered defense framework described earlier, regardless of model provenance. Security depends on robust data governance, provenance tracking, ensemble disagreement monitoring, adversarial testing, and institutional security expertise—not on whether the model was developed in-house. While in-house development may offer advantages in customization and regulatory compliance, it cannot guarantee protection against the data poisoning threats analyzed in this study.

Limitations

This analysis has several limitations. First, although we synthesized empirical attack success rates from peer-reviewed studies (Table 2), we did not perform original attacks on production health care AI systems. The analytical scenarios (A1-D1) apply published research findings to health care contexts rather than representing documented incidents. Actual feasibility may vary depending on local security infrastructure and deployment configurations.

Second, our analysis focused on data poisoning during training and fine-tuning, with limited coverage of inference-time attacks, adversarial examples, model extraction, or privacy attacks. This scope was deliberately constrained to training-time vulnerabilities. Third, the generalizability of our findings across different model scales is uncertain. Published research has examined models with up to 13 billion parameters [1,2,20], whereas health care increasingly employs models with 100 billion to 340 billion+ parameters. Although recent studies suggest that attacks require near-constant sample counts regardless of model scale [5], extrapolating these findings to models with more than 100 billion parameters warrants further empirical validation.

XSL•FO
RenderX

Fourth, defense mechanisms, including MEDLEY, have not been validated in prospective clinical trials. Their real-world effectiveness depends on implementation, integration into clinical workflows, and human factors, all of which require empirical deployment studies. Fifth, our impact projections relied on conservative assumptions, which may underestimate potential harm. We assumed limited attacker capabilities, partial compromise, and detection within 12-24 months. More sophisticated adversaries could cause substantially greater harm, whereas highly resilient organizations may mitigate some risks. Sixth, our regulatory analysis reflects policies as of late 2025; ongoing or future governance changes may mitigate some of the vulnerabilities identified. Finally, our literature synthesis was limited to English-language studies, which may introduce publication bias.

Despite these limitations, our analysis highlights critical security gaps that demand urgent attention from the health care AI community, regulators, and policy makers.

Conclusions

Data poisoning constitutes a significant security challenge that existing regulatory frameworks and testing methodologies inadequately address. Our analysis shows that even small numbers of poisoned samples can compromise health care AI systems, with detection delays ranging from months to years—or potentially indefinite—without appropriate monitoring. Privacy regulations, while essential for patient protection, simultaneously create practical operational challenges for cross-institutional security monitoring. Conventional cybersecurity defenses are insufficient to prevent adversarial data manipulation.

Health care organizations should adopt multilayered defense frameworks, incorporating strategies such as ensemble disagreement monitoring (eg, the proposed MEDLEY framework), active defenses, governance structures, and architectural safeguards. Although MEDLEY requires empirical validation before clinical deployment, the underlying principle of ensemble disagreement monitoring offers a promising approach for detecting data poisoning. International coordination on security standards is essential. Most critically, the health care community must evaluate whether black-box AI architectures are appropriate for life-or-death decisions, or whether patient safety requires interpretable systems that prioritize verifiable safety over marginal performance gains. The asymmetry between the ease of attack—requiring only hundreds of poisoned samples—and the difficulty of detection—often 12-24+ months—demands urgent action. Without proactive implementation of ensemble monitoring, Byzantine-robust architectures, and mandatory adversarial testing, health care organizations risk systematic, undetected compromise affecting thousands of patients over time. The question is not if data poisoning will occur in clinical AI, but when—and whether we will act before theoretical vulnerabilities translate into documented clinical disasters.

Acknowledgments

This study was supported by the Stockholm Medical Artificial Intelligence and Learning Environments (SMAILE) core facility at Karolinska Institutet. The authors used artificial intelligence–based tools during the preparation of this manuscript. Large language models were employed for English proofreading and to improve the readability of the text. Additionally, Scite and SciSpace were used to assist with the literature review, including citation analysis and the identification of relevant research articles. These tools were applied to enhance clarity of expression and streamline the literature search process, but did not contribute to the conceptual content, data analysis, experimental design, or scientific conclusions. The authors independently verified all cited sources, and all scientific interpretations remain the sole responsibility of the authors. The authors take full responsibility for the content of the manuscript.

Data Availability

No new data were generated for this analysis. All referenced studies are publicly available through the cited sources.

Authors' Contributions

Case analysis: FS, IP, MVB
Conceptualization: FA
Formal analysis: FA
Supervision: MVB
Writing – original draft: FA
Writing – review & editing: FS, IP, MVB

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional analysis.
[DOCX File , 32 KB - jmir_v28i1e87969_app1.docx]

XSL•FO
RenderX

References

1. Wan A, Wallace E, Shen S, Klein D. Poisoning language models during instruction tuning. arXiv. Preprint posted online on May 1, 2023 [FREE Full text] [doi: 10.48550/arXiv.2305.00944]
2. Rando J, Tramèr F. Universal jailbreak backdoors from poisoned human feedback. arXiv. Preprint posted online on November 24, 2023 [FREE Full text] [doi: 10.48550/arXiv.2311.14455]
3. Gu T, Liu K, Dolan-Gavitt B, Garg S. BadNets: evaluating backdooring attacks on deep neural networks. IEEE Access 2019;7:47230-47244. [doi: 10.1109/access.2019.2909068]
4. Liu Y, Ma S, Aafer Y, Lee WC, Zhai J, Wang W, et al. Trojaning attack on neural networks. 2018 Jan 1 Presented at: 25th Annual Network And Distributed System Security Symposium (NDSS 2018); February 18-21, 2018; San Diego, CA p. 1-15 URL: https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_03A-5_Liu_paper.pdf [doi: 10.14722/ndss.2018.23291]
5. Souly Alexandra, Rando J, Chapman E, Davies X, Hasircioglu B, Shereen E, et al. Poisoning attacks on LLMs require a near-constant number of poison samples. arXiv. Preprint posted online on October 8, 2025. URL: https://arxiv.org/abs/2510.07192 [accessed 2025-10-08]
6. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. arXiv. Preprint posted online on August 8, 2021 [FREE Full text] [doi: 10.48550/arXiv.2108.07258]
7. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
8. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
9. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature 2025 Jun 22;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
10. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
11. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019 Jan;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
12. Park J, O'Brien J, Cai C, Morris M, Liang P, Bernstein M. Generative agents: interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. 2023 Oct 29 Presented at: UIST '23: The 36th Annual ACM Symposium on User Interface Software and Technology; October 29, 2023; San Francisco, CA p. 1-22 URL: https://dl.acm.org/doi/10.1145/3586183.3606763 [doi: 10.1145/3586183.360676]
13. Soman N R, Prakash G A, Azza H. Optimizing hospital operations with AI-driven resource allocation tools. In: Swarnkar SK, Chunawala P, Chunawala H, Tran TA, Rathore YK, editors. Transforming Healthcare with Artificial Intelligence. New York, NY: Springer; Jul 26, 2025:99-110.
14. U.S. Department of Health and Human Services. Standards for Privacy of Individually Identifiable Health Information (HIPAA Privacy Rule): 45 CFR Parts 160 and 164. U.S. Department of Health and Human Services. 2013. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/index.html [accessed 2025-11-06]
15. European U. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). EUR-Lex. 2016 May 04. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng [accessed 2025-12-31]
16. Qi X, Zeng Y, Xie T, Chen P, Jia R, Mittal P, et al. Fine-tuning aligned language models compromises safety, even when users do not intend to!. arXiv. Preprint posted online on October 5, 2023 [FREE Full text] [doi: 10.48550/arXiv.2310.03693]
17. Rathbun E, Amato C, Oprea A. SleeperNets: universal backdoor poisoning attacks against reinforcement learning agents. San Diego, CA: Neural Information Processing Systems Foundation, Inc; 2024 Presented at: Advances in Neural Information Processing Systems 37; December 10-15, 2024; Vancouver, BC, Canada p. 2024 URL: https://www.proceedings.com/079017-3556.html [doi: 10.52202/079017-3556]
18. Bhagoji A, Chakraborty S, Mittal P, Calo S. Analyzing federated learning through an adversarial lens. 2019 Presented at: The Thirty-Sixth International Conference on Machine Learning (ICML 36); June 9-15, 2019; Long Beach, CA.
19. Fang M, Cao X, Jia J, Gong N. Local model poisoning attacks to byzantine-robust federated learning. In: Proceedings of the 29th USENIX Conference on Security Symposium. 2020 Aug 12 Presented at: SEC'20: 29th USENIX Conference on Security Symposium; August 12-14, 2020; Boston, MA p. e1. [doi: 10.5555/3489212.3489304]
20. Guo Z, Kumar A, Tournai R. Persistent backdoor attacks in continual learning. arXiv. Preprint posted online on September 20, 2024 [FREE Full text] [doi: 10.48550/arXiv.2409.13864]
21. Scite. URL: https://www.scite.ai [accessed 2025-12-31]
22. SciSpace. URL: https://scispace.com/ [accessed 2025-12-31]
23. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang L, et al. LoRA: low-rank adaptation of large language models. arXiv. Preprint posted online on October 21, 2021 [FREE Full text] [doi: 10.5260/chara.21.2.8]
24. Gichoya JW, Banerjee I, Bhimireddy AR, Burns JL, Celi LA, Chen L, et al. AI recognition of patient race in medical imaging: a modelling study. The Lancet Digital Health 2022 Jun;4(6):e406-e414. [doi: 10.1016/s2589-7500(22)00063-2]

XSL•FO
RenderX

25. McMahan B, Moore E, Ramage D, Hampson S, Arcas B. Communication-efficient learning of deep networks from decentralized data. arXiv. Preprint posted online on January 26, 2023 [FREE Full text] [doi: 10.48550/arXiv.1602.05629]
26. Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Nitin Bhagoji A, et al. Advances and open problems in federated learning. arXiv. Preprint posted online on March 21, 2021 [FREE Full text]
27. Blanchard P, El Mhamdi EM, Guerraoui R, Stainer J. Byzantine-tolerant machine learning. arXiv. Preprint posted online on March 8, 2017 [FREE Full text] [doi: 10.48550/arXiv.1703.02757]
28. Yin D, Chen Y, Ramchandran K, Bartlett P. Byzantine-robust distributed learning: towards optimal statistical rates. 2018 Presented at: ICML 2018: The Thirty-Fifth International Conference on Machine Learning; July 10-15, 2018; Stockholm, Sweden.
29. Pillutla K, Kakade SM, Harchaoui Z, Pillutla K. Robust aggregation for federated learning. IEEE Trans Signal Process 2022;70:1142-1154. [doi: 10.1109/tsp.2022.3153135]
30. US Food and Drug Administration. Artificial intelligence and machine learning in software as a medical device. FDA. 2021 Jan 12. URL: https://www.fda.gov/media/145022/download [accessed 2025-10-30]
31. US Food and Drug Administration. Predetermined change control plans for machine learning-enabled medical devices: draft guidance for industry and Food and Drug Administration staff. FDA. 2023 Apr 03. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/predetermined-change-control-plans-machine-learning-enabled-medical-devices [accessed 2025-10-05]
32. US Food and Drug Administration (FDA), International Medical Device Regulators Forum. Software as a Medical Device (SaMD): clinical evaluation. FDA. 2017 Dec 8. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/software-medical-device-samd-clinical-evaluation [accessed 2025-11-05]
33. Article 29 Data Protection Working Party. Guidelines on automated individual decision-making and Profiling for the purposes of regulation 2016/679 (wp251rev.01). Endorsed by the European Data Protection Board. EUR-Lex. 2018. URL: https://ec.europa.eu/newsroom/article29/items/612053 [accessed 2025-11-06]
34. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. arXiv. Preprint posted online on September 9, 2019 [FREE Full text] [doi: 10.48550/arXiv.1706.06083]
35. Abtahi F, Astaraki M, Seoane F. Leveraging imperfection with MEDLEY: a multi-model approach harnessing bias in medical AI. arXiv. Preprint posted online on August 29, 2025 [FREE Full text] [doi: 10.48550/arXiv.2508.21648]
36. Dietterich T. Ensemble methods in machine learning. In: Multiple Classifier Systems. Berlin, Heidelberg, Germany: Springer; 2000:1-15.
37. Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. 2018 Presented at: ICML 2018: The Thirty-Fifth International Conference on Machine Learning; July 10-15, 2018; Stockholm, Sweden. [doi: 10.48550/arXiv.1802.00420]
38. Greshake K, Abdelnabi S, Mishra S, Endres C, Holz T, Fritz M. Not what you've signed up for: compromising real-world LLM-integrated applications with indirect prompt injection. New York, NY: Association for Computing Machinery (ACM); 2023 Nov 26 Presented at: 16th ACM Workshop on Artificial Intelligence and Security; November 30, 2023; Copenhagen, Denmark p. 2023 URL: https://dl.acm.org/doi/10.1145/3605764.3623985 [doi: 10.1145/3605764.3623985]
39. Schwarzschild A, Goldblum M, Gupta A, Dickerson JP, Goldstein T. Just how toxic is data poisoning? A unified benchmark for backdoor and data poisoning attacks. arXiv. Preprint posted online on June 17, 2021 [FREE Full text] [doi: 10.48550/arXiv.2006.12557]
40. Elhage N, Nanda N, Olsson C. A mathematical framework for transformer circuits. Anthropic. 2021 Oct 30. URL: https://transformer-circuits.pub/2021/framework/index.html [accessed 2025-11-06]
41. Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, et al. Constitutional AI: harmlessness from AI feedback. arXiv. Preprint posted online on December 15, 2022 [FREE Full text] [doi: 10.48550/arXiv.2212.08073]
42. Carlini N, Terzis A. Poisoning and backdooring contrastive learning. arXiv. Preprint posted online on March 28, 2022 [FREE Full text] [doi: 10.48550/arXiv.2106.09667]
43. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021 Jul 15;385(3):283-286 [FREE Full text] [doi: 10.1056/NEJMc2104626] [Medline: 34260843]
44. Charter of Fundamental Rights of the European Union (Articles 21, 35). EUR-Lex. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT [accessed 2025-12-31]
45. Price WN, Cohen IG. Privacy in the age of medical big data. Nat Med 2019 Jan 7;25(1):37-43 [FREE Full text] [doi: 10.1038/s41591-018-0272-7] [Medline: 30617331]
46. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight - reconsidering the use of race correction in clinical algorithms. N Engl J Med 2020 Aug 27;383(9):874-882 [FREE Full text] [doi: 10.1056/NEJMms2004740] [Medline: 32853499]
47. Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang P, et al. Towards generalist biomedical AI. NEJM AI 2024 Feb 22;1(3):1. [doi: 10.1056/aioa2300138]
48. Chen K, Qiu L, Xie X, Bai Y. Dynasmile: video-based smile analysis software in orthodontics. SoftwareX 2025 Feb;29:102004. [doi: 10.1016/j.softx.2024.102004]
49. Lu Q, Li R, Sagheb E, Wen A, Wang J, Wang L, et al. Explainable diagnosis prediction through neuro-symbolic integration. arXiv. Preprint posted online on October 1, 2024 [FREE Full text] [doi: 10.48550/arXiv.2410.01855]

XSL•FO
RenderX

50. Vidal M, Chudasama Y, Huang H, Purohit D, Torrente M. Integrating knowledge graphs with symbolic AI: the path to interpretable hybrid AI systems in medicine. Journal of Web Semantics 2025 Jan;84:100856. [doi: 10.1016/j.websem.2024.100856]
51. The EU Artificial Intelligence Act. European Parliament. 2023 Aug 6. URL: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence [accessed 2025-12-31]
52. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance. EUR-Lex. URL: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng [accessed 2025-12-31]
53. European U. Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space. EUR-Lex. 2025. URL: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en [accessed 2025-11-05]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
BioMedCLIP: Biomedical Contrastive Language–Image Pretraining
CFR: Code of Federal Regulations
CNN: convolutional neural network
EHDS: European Health Data Space
EHR: electronic health record
EU: European Union
FDA: Food and Drug Administration
GDPR: General Data Protection Regulation
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language models
LoRA: low-rank adaptation
MedCLIP: Medical Contrastive Language–Image Pretraining
ML: machine learning
RadImageNet: Radiology ImageNet
RLHF: reinforcement learning from human feedback

Edited by G Tsafnat; submitted 17.Nov.2025; peer-reviewed by A Famotire, K Chen; comments to author 28.Nov.2025; revised version received 09.Dec.2025; accepted 12.Dec.2025; published 23.Jan.2026.

Please cite as:
Abtahi F, Seoane F, Pau I, Vega-Barbas M
Data Poisoning Vulnerabilities Across Health Care Artificial Intelligence Architectures: Analytical Security Framework and Defense Strategies
J Med Internet Res 2026;28:e87969
URL: https://www.jmir.org/2026/1/e87969
doi:10.2196/87969
PMID:

©Farhad Abtahi, Fernando Seoane, Ivan Pau, Mario Vega-Barbas. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Implementing an Artificial Intelligence Decision Support System in Radiology: Prospective Qualitative Evaluation Study Using the Nonadoption Abandonment Scale-Up, Spread, and Sustainability (NASSS) Framework

Sundresan Naicker[1], PhD; Paul Schmidt[2], MBBS; Bruce Shar[2], PhD; Amina Tariq[1], PhD; Ashleigh Earnshaw[1], MPH; Steven McPhail[1,3], PhD

[1]Australian Centre for Health Services Innovation, School of Public Health and Social Work, Queensland University of Technology, Kelvin Grove, Australia

[2]Department of Medical Imaging, Princess Alexandra Hospital, Woolloongabba, Australia

[3]Digital Health and Informatics Directorate, Metro South Health, Woolloongabba, Australia

Corresponding Author:
Sundresan Naicker, PhD
Australian Centre for Health Services Innovation
School of Public Health and Social Work
Queensland University of Technology
X Block, 88 Musk Avenue
Kelvin Grove, 4059
Australia
Phone: 61 449876034
Email: sundresan.naicker@qut.edu.au

Abstract

Background: Medical imaging remains at the forefront of advancements in adopting digital health technologies in clinical practice. Regulator-approved artificial intelligence (AI) clinical decision support systems are commercially available and being embedded into routine practices for radiologists internationally. These decision support solutions show promising clinical validity compared to standard practice conditions; however, their implementation over time and implications on radiologists' practice are poorly understood.

Objective: This paper aims to examine the real-world implementation of an AI clinical decision support tool in radiology through a qualitative evaluation across pre-, peri-, and postimplementation phases. Specifically, it seeks to identify the key contextual, organizational, and human factors shaping adoption and sustainability, to map these influences using the nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) framework, and to generate insights that inform evidence-based strategies and policy for integrating AI safely and effectively into public hospital imaging services.

Methods: This prospective study was conducted in a large public tertiary referral hospital in Brisbane, Queensland, Australia. One-to-one participant interviews were undertaken across the 3 implementation phases. Participants comprised radiology consultants, registrars, and radiographers involved in chest computed tomography studies during the study period. Interviews were guided by the NASSS framework to identify contextual factors influencing implementation.

Results: A total of 43 semistructured interviews were conducted across baseline (n=16), peri-implementation (n=9), and postimplementation (n=18) phases, comprising 7 (16%) radiographers, 20 (47%) registrar radiologists, and 16 (37%) consultant radiologists. Across NASSS domains, 56 barriers and 18 enablers were identified at baseline, 55 and 14 during peri-implementation, and 82 and 33 postimplementation. Organizational barriers dominated early phases, while technological issues such as system accuracy, interoperability, and information overload became most prominent during and after rollout. Enablers increased over time, particularly within the technology and value proposition domains, as some clinicians adapted the AI as a secondary safety check. Trust and adoption remained constrained by performance inconsistency, weak communication, and medicolegal uncertainty.

Conclusions: The implementation of AI decision support in radiology is as much an organizational and cultural process as a technological one. Clinicians remain willing to engage, but sustainable adoption depends on consolidating early positive experiences and addressing negative ones, embedding communication and training, and maintaining iterative feedback between users, vendors,

XSL•FO
RenderX

and system leaders. Applying the NASSS framework revealed how domains interact dynamically across time, offering both theoretical insight into sociotechnical complexity and practical guidance for hospitals seeking to move from pilot to routine, trustworthy AI integration.

(J Med Internet Res 2026;28:e80342) doi:10.2196/80342

KEYWORDS

artificial intelligence; radiology; implementation science; decision support; qualitative

Introduction

Background

The demand for advanced medical imaging services continues to grow at a rapid pace, and reducing time delay from image capture to radiological reporting remains a priority for public hospital services [1,2]. Radiologists often need to work through large volumes of information to report on medical images for a wide array of patient groups. This can be challenging in clinical environments with constraints on time, resources, and related workflow pressures. Excessive delay from time of medical image capture to the provision of a definitive radiological report to the referring clinical team can undermine the quality and safety of health care delivery [3,4].

Medical imaging has been, and remains, at the forefront of advancements in digital health technologies in everyday clinical practice [2,5]. Consequently, there is an array of digital tools available to support radiology decision-making that have arisen from advances in machine learning and other related technologies. Artificial intelligence (AI) algorithms that are commercially available, and with regulatory approvals already in place, are now being embedded in digital tools and readily adopted into routine practices for radiologists in various settings internationally [6,7].

In experimental and validation studies, AI algorithms show strong technical performance: meta-analyses report pooled sensitivity and specificity values exceeding 0.80-0.85 for tumor metastasis and rib fracture detection, with the mean area under the curve near 0.90 [8,9]. These results highlight the potential for AI to enhance accuracy and throughput, particularly in resource-constrained health systems [5,8,10]. Yet this technical promise has outpaced the evidence on real-world implementation [11]. The literature remains dominated by model validation [5] and cross-sectional studies of clinician trust [12-14], with few studies examining how AI systems are adopted, adapted, or sustained within the operational realities of hospital environments. Recent reviews emphasize this gap, noting that most AI-radiology research ends at performance benchmarking and fails to explore workflow integration, organizational readiness, or long-term routinization [10,15]. Broader governance and workforce analyses likewise underline persistent uncertainty around accountability, medicolegal responsibility, and system-level preparedness for AI-supported care [16,17]. Collectively, these gaps constrain understanding of how algorithmic potential translates into clinical and organizational value.

While these quantitative evaluations and meta-analyses have established AI's diagnostic capability, they provide little insight into how and why such technologies succeed or fail once introduced into routine clinical practice. Many rely on retrospective datasets, simulated environments, or controlled reader studies that remove the influence of real-world complexity [18-21]. Consequently, they overlook the macro-, meso-, and microlevel dynamics of workflow adaptation, human-technology interaction, and organizational and sociocultural context that determine whether AI enhances or disrupts practice. A qualitative implementation approach is therefore critical for exploring the lived experiences, informal workarounds, and contextual contingencies that shape integration in situ [21,22]. Such an approach complements quantitative evidence by revealing the social and organizational mechanisms through which AI adoption is negotiated, sustained, or resisted in everyday radiology work and practice.

Emerging qualitative and mixed methods studies have begun to address aspects of these challenges by exploring radiologists' perceptions, sources of trust and mistrust, and organizational barriers to adoption [12,23]. However, most have relied on single-time-point interviews, limited samples (n<20), or hypothetical case vignettes that do not capture the evolving interaction between users, workflows, and technology over time [18,19,24]. Few have been conducted within active service settings or have systematically linked individual experiences to organizational processes or system-level factors [12,25]. This has resulted in a descriptive but fragmented evidence base that provides limited insight into how implementation unfolds, stabilizes, or falters once AI becomes part of routine care.

This study responds directly to that gap by presenting a qualitative, end-to-end evaluation of AI implementation within a large tertiary radiology department in Brisbane, Queensland, Australia. Here, end-to-end refers to a lifecycle approach spanning predeployment context and readiness, peri-implementation adaptation, and postimplementation integration and routinization, examining how technological, human, and organizational factors interact over time [26,27]. There is a broad range of implementation frameworks to assess the implementation of a digital health innovation across a life cycle; however, they are varied in their analytical purpose [28]. Strategy-based models such as the Expert Recommendations for Implementing Change (ERIC) framework provide detailed lists of discrete implementation actions, but they are limited in explaining the mechanisms through which adoption unfolds in complex clinical environments [29]. In contrast, our evaluation sought to understand how and why AI integration succeeds or stalls within a dynamic, real-world system. The nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) framework [30] was therefore selected to guide data collection and analysis. NASSS offers a theoretically grounded structure

XSL•FO
RenderX

for examining the sociotechnical complexity of digital innovation by integrating the domains of technology, adopters, organization, value proposition, and wider context [31,32]. This systems-oriented lens provides greater explanatory power than more generalized implementation approaches for capturing interdependencies, contextual contingencies, and the temporal evolution of barriers and enablers across the implementation lifecycle.

By situating implementation within real-world clinical operations rather than experimental or hypothetical conditions, this study provides a rare longitudinal perspective on how AI becomes normalized or resisted within a complex hospital environment. Its findings have direct relevance to current policy efforts to scale AI responsibly in public health systems, where efficiency, safety, and governance imperatives converge [33].

Aims

Accordingly, this paper aims to examine the real-world implementation of an AI-based clinical decision support tool in radiology through an end-to-end qualitative evaluation across pre- (baseline), peri-, and postimplementation phases. Specifically, it seeks to identify the key contextual, organizational, and human factors shaping adoption and sustainability, to map these influences using the NASSS framework, and to generate insights that inform evidence-based strategies and policy for integrating AI safely and effectively into public hospital imaging services.

Methods

Study Design and Theoretical Framework

This study used a qualitative prospective design. The study was structured across 3 temporal phases to capture the evolving context of AI implementation within the radiology department.

The preimplementation or baseline phase (12 months before deployment) corresponded to the period when the AI tool had not yet been introduced. This phase reflected baseline organizational conditions, established workflows, and prevailing attitudes toward digital tools in radiology.

The peri-implementation phase (an 8-week transition period) covered the initial rollout of the AI system and its integration into existing digital and reporting infrastructure. This period was characterized by early interaction with the tool and short-term adaptation of work processes.

The postimplementation phase (12 months after deployment) represented a period of stabilization in which the AI tool had become part of routine operations. This phase captured the mature context of use, reflecting how the technology was embedded, maintained, and normalized within everyday practice.

The reporting of this study was in alignment with the COREQ (Consolidated Criteria for Reporting Qualitative Studies; Multimedia Appendix 1).

NASSS Framework

The NASSS framework was used to inform the study design and interview questions. The NASSS framework provides a systematic foundation for examining challenges across multiple domains and their dynamic interactions, which may influence the uptake, implementation, outcomes, and sustainability of technology-supported health programs [30]. It facilitates consideration of how various factors interact and adapt over time, influencing success, and includes the following domains:

1. Condition or illness: the nature and complexity of the condition being addressed.
2. Technology: the specific technology being implemented.
3. Value attributed to the technology: the perceived benefits and utility of the technology.
4. Individual adopters: the clinicians and patients using the technology.
5. Organizational adopters: the health care organizations implementing the technology.
6. External context: the broader context, including regulatory, economic, and social factors.

Setting

This was a single-site study conducted at a large public tertiary referral hospital in Brisbane. The hospital's Medical Imaging Department offers a comprehensive range of diagnostic imaging services to support patient care across various medical specialties.

AI Clinical Decision Support System

The technology adopted by the department is a third-party and commercially available multiorgan AI-based computerized clinical decision support system (CDSS) for radiologists. The CDSS uses multiple specialized convolutional neural networks across the entire machine learning cycle, including preprocessing, candidate generation, classification, and postfiltering. It has been classified and approved as a diagnostic tool under the current Australian Therapeutic Goods Administration regulatory framework. The decision support system integrates with existing medical imaging hardware and software to allow computed tomography (CT) images to be automatically transferred from the scanner, preprocessed, and prepared for interpretation by radiologists. The CDSS flags or highlights any issues within the CT image that require further differential diagnosis by the radiologists. In 2021, before full site implementation, the study site tested this tool among a small group of radiologists (n=4) and anecdotally reported positive experiences.

Participant Recruitment and Sampling

Participants included radiology consultants, registrars, and radiographers employed within the Medical Imaging Department who were involved in chest CT reporting during the study period. Recruitment was undertaken via internal email by an embedded chief investigator. A purposive yet stratified sampling approach was used to achieve broad representation across professional roles and levels of seniority. At the time of data collection, the lead interviewer (SN) was an experienced PhD-trained male implementation science researcher with no supervisory, managerial, or clinical authority over participants. SN had established professional familiarity with the department through earlier collaborative work, but no direct reporting relationships. Participants were informed of SN's research role, disciplinary background, and interest in understanding

XSL•FO
RenderX

real-world implementation challenges before interview commencement. Stratification was guided by the departmental organizational chart to ensure inclusion of participants from different functional areas and reporting responsibilities. This approach sought to capture a range of experiences across the implementation process rather than statistical representativeness. Sampling continued iteratively across the pre- (baseline), peri-, and postimplementation phases until thematic adequacy was reached, indicated by repetition of key concepts and no emergence of new issues in subsequent interviews [34].

Qualitative Interviews

Semistructured face-to-face interviews, approximately 40 minutes in length, were conducted by an experienced implementation science researcher (SN) according to participant preference (in person, via Microsoft Teams, or through phone; Multimedia Appendix 2). All interviews were audio-recorded and transcribed upon participant consent. Interviews were conducted flexibly, with questions adapted to participant roles and experience. Not all questions or prompts were asked in every interview, but the guide provided a consistent framework to ensure coverage of key domains. The interviewer kept reflexive notes after each interview to document emerging impressions, relational dynamics, and potential influences of their positionality on data generation.

Study Materials

We used a reflexive framework method to guide the development of a semistructured interview template, aligning with our study aims [32,35]. This approach aimed to capture a comprehensive range of insights, perceptions, and experiences, providing a rich dataset for analysis.

Data Analysis

Interview transcripts were analyzed using an iterative, multistage process combining thematic analysis with NASSS-informed framework mapping [36-38]. Before analysis, the research team discussed their disciplinary positions and assumptions regarding AI in radiology, documenting these reflections to support analytic reflexivity. SN led coding with AE providing independent review; neither had clinical authority over participants.

Analysis and data collection were conducted concurrently to guide purposive sampling and determine saturation. Early transcript review enabled the identification of preliminary codes and gaps, informing subsequent recruitment to ensure variation in experience and role. No participant reviewed the transcripts or findings before publication, consistent with the exploratory and ecological design of the study. However, findings were discussed with senior departmental clinicians and technical leads during routine project meetings. These discussions did not involve revising data or themes but served to ensure that the interpretations accurately reflected the broader organizational context and the realities experienced within the department. This informal sense-checking supported contextual validity while maintaining analytic independence.

Initial inductive coding was undertaken by one researcher (SN), who read each transcript line by line and assigned short descriptive phrases summarizing perceived barriers, facilitators, or neutral factors related to AI implementation. To strengthen analytic credibility, a second researcher (AE) independently reviewed a subset of transcripts and the draft codebook. Coding discrepancies and interpretive differences were discussed and resolved through consensus, providing a form of cross-checking without imposing rigid interrater reliability metrics. Iterative discussions across the research team further refined code boundaries, ensuring conceptual coherence and maintaining an audit trail of analytic decisions.

Once inductive coding was complete, higher-order categories were developed to capture recurrent concepts and relationships. These categories were then deductively mapped to the NASSS framework. This process enabled systematic classification of determinants by domain (eg, technology, organization, value proposition, adopters, wider system, and clinical context) while retaining sensitivity to context-specific nuances. The mapping process was iterative, with subthemes revisited and refined to ensure conceptual alignment between inductive insights and NASSS constructs, and to account for determinants that spanned multiple sociotechnical domains. Mapped subthemes were subsequently synthesized into a set of higher-order, cross-cutting determinants representing the dynamic interactions between technological, organizational, and adopter-related factors across implementation phases. This synthesis informed the structure of the results, where inductive findings and NASSS categories are integrated to illustrate how determinants evolved from baseline to peri-implementation and into routine use. Illustrative quotes were selected by consensus to exemplify the range of perspectives within each theme and subtheme. Quote selection focused on demonstrating variation, depth, and temporal evolution rather than providing isolated examples, ensuring that quotations functioned as analytic evidence and supported the integration of inductive insights with NASSS domains. An accompanying summary table provides an at-a-glance depiction of how inductive themes aligned with specific NASSS domains and subdomains, further enhancing analytic transparency and coherence.

To enhance analytic transparency, a content count of all coded barriers and enablers was compiled in Microsoft Excel and stratified by implementation phase (baseline, peri-implementation, and postimplementation). This numerical summary illustrated the distribution and relative prominence of determinants across NASSS domains (eg, technological, organizational, and adopter-related), providing a structured complement to the qualitative narrative. The count functioned as a descriptive aid to visualize patterns within the dataset, highlight areas of convergence and divergence across phases, and support the organization of complex, multilevel determinants [39].

Trustworthiness

To ensure trustworthiness, the research team engaged in continuous reflexive discussions throughout data collection and analysis, critically examining how their disciplinary backgrounds and assumptions could shape interpretation. Coding decisions were documented in an evolving analytic log, forming an audit trail that supported transparency and replicability. Regular peer

XSL•FO
RenderX

debriefings were held to resolve interpretive differences and refine theme definitions. Trustworthiness was further reinforced through the systematic application of the NASSS framework, which provided a theoretically grounded lens for organizing inductive findings. The inclusion of a quantitative content count of coded determinants added descriptive transparency, demonstrating how interpretations were anchored in the underlying data distribution. Together, these strategies strengthened the credibility, confirmability, and dependability of the qualitative findings [35,38].

Ethical Considerations

The Human Research Ethics Committee granted ethical clearance for this research (HREC/2021/QMS/81483). All participants provided written and verbal informed consent before participating in the study. Participation was voluntary, and participants could withdraw at any time. Participants were assured that their responses would be confidential, would not be shared with departmental leadership in identifiable form, and would have no bearing on workplace evaluation or progression. No incentives were offered, and no previous personal relationships existed between the researcher and participants beyond professional familiarity.

Results

Participant Characteristics

A total of 43 one-on-one interviews were conducted across the study timeframe, as shown in Table 1. This consisted of 7 (16%) radiographers, 20 (47%) registrar radiologists, and 16 (37%) consultant radiologists. A total of 9 (21%) participants were interviewed across multiple time points, consistent with public health services experiencing regular staff rotation, shift-based work patterns, and competing clinical pressures. While this posed practical challenges to longitudinal participation, it was also indicative of practical challenges with AI implementation in hospital medical imaging departments. To accommodate this and maintain the integrity of the analysis, each interview was treated as a discrete data point. This allowed us to capture a wider range of perspectives from across the workforce and reflect the dynamic, high-turnover environment typical of public hospital settings.

Table 1. Participant characteristics across the 3 implementation phases.

Participants	Baseline	Peri-implementation	Postimplementation
Total number of participants (N=43), n (%)	16 (37)	9 (21)	18 (42)
Profession and seniority, n (%)			
Radiographer (N=7)	4 (57)	1 (14)	2 (28)
Radiology registrar (N=20)	7 (35)	4 (20)	9 (45)
Consultant radiologist (N=16)	5 (31)	4 (25)	7 (44)
Sex, n (%)			
Male (N=26)	9 (35)	6 (23)	11 (42)
Female (N=17)	7 (41)	3 (18)	7 (41)

NASSS Informed Barriers and Enablers of AI Implementation in Radiology

A total of 56 barriers and 18 enablers were identified at baseline, 55 barriers and 14 enablers during peri-implementation, and 82 barriers and 33 enablers at postimplementation. Figure 1 presents barriers across NASSS domains and study phases, while Figure 2 shows enablers across the same phases.

Figure 1. Barriers across nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) domains and study phases.

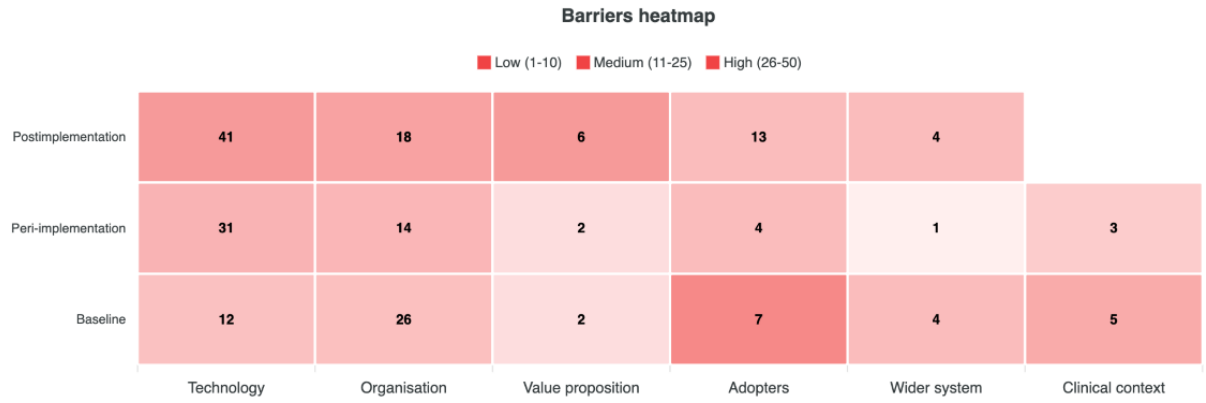

XSL•FO
RenderX

Figure 2. Enablers across nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) domains and study phases.

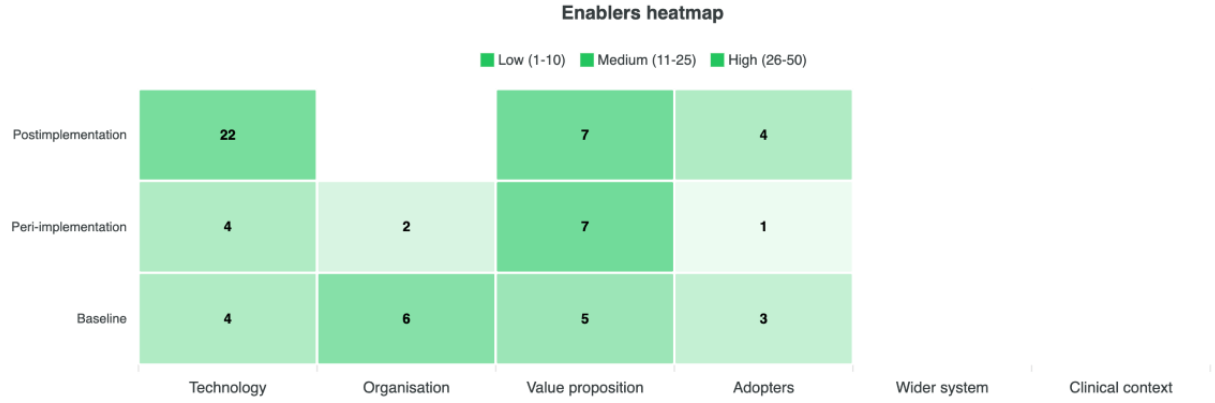

At baseline, organizational barriers were the most prominent, representing nearly half of all identified barriers (26/56, 46%). These are primarily related to limited technological readiness, insufficient training, and inadequate workflow planning for implementation. Technological barriers followed (12/56, 21%), reflecting early concerns about AI performance and output accuracy, while adopter-related barriers (7/56, 12%) centered on uncertainty regarding medicolegal accountability when using AI in reporting. The main enablers at this stage were found within the organizational (6/18, 33%) and value proposition (5/18, 28%) domains, reflecting a collegial, innovation-friendly culture and a belief in the technology's potential for efficiency and time savings.

During the peri-implementation phase, technological barriers dominated (31/55, 56%), particularly those concerning interoperability and system performance. These were followed by organizational barriers (14/55, 25%) related to weak implementation planning and inadequate workflow support, and a smaller set of adopter barriers (4/55, 7%) linked to limited trust in the AI system. Despite these issues, several enablers emerged, most notably within the value proposition domain (7/14, 50%), where participants anticipated potential efficiency gains if technical and integration challenges could be addressed. A smaller number of enablers (4/14, 28%) related to technology, as some users began using the AI system to cross-check their own interpretations.

By postimplementation, technological barriers persisted (41/82, 50%) as problems with accuracy, reliability, and speed remained unresolved. Organizational barriers (18/82, 22%) continued to reflect deficiencies in communication, training, and workflow integration, while adopter barriers (13/82, 16%) indicated ongoing distrust in the AI and reluctance to incorporate it fully into routine practice. However, this phase also saw the most substantial growth in enablers (a total of 33), particularly within technology (22/33, 67%), as users adapted the system for use as a secondary check or safety mechanism. Additional enablers were identified within the value proposition (7/33, 21%), where participants recognized relative efficiency benefits, and among adopters (4/33, 12%) who expressed emerging, albeit cautious, trust in the AI's evolving role.

Across all NASSS domains, implementation was characterized by an interplay between anticipated risks, such as workflow integration and information overload, and realized challenges during peri-implementation, many of which persisted into routine use (Figure 1). While optimism and perceived value remained for some, trust and adoption were undermined by ongoing performance and communication barriers. Together, these patterns illustrate how implementation unfolded within a large, dynamic clinical service, with determinants shifting as the AI system moved from anticipation to early use and then attempts to move into routine practice. The relative prominence of technological, organizational, and adopter-related factors at each phase provides a contextual frame for understanding the subsequent themes. These distributions therefore situate the qualitative findings within the broader organizational and technological environment in which the AI was being implemented.

Framework Analysis and Narrative Synthesis of NASSS Domains and Subdomains

Table 2 presents a synthesis of inductive themes mapped to the NASSS framework, illustrating how key implementation dynamics evolved across baseline, peri-implementation, and postimplementation phases. The table highlights temporal shifts in organizational readiness, technological integration, value perception, adopter engagement, and wider system influences. These findings are further expanded in the results narrative synthesis, offering deeper insight into the contextual and temporal nuances of implementation.

XSL•FO
RenderX

Table 2. Mapping of inductive themes to nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) domains and subdomains with indicative change over time.

NASSS domain	Subdomain	Inductive theme	Change across phases
Organization	Work needed to plan, implement, and monitor change	Sustained implementation planning	Limited formal planning at baseline; reactive coordination during rollout; structured monitoring and formalized training emerged post implementation.
Organization	Organizational readiness and capacity to innovate	Relational engagement and communication	Collegial but fragmented culture at baseline; weak interteam communication during rollout; some shared ownership and cross-functional coordination developed post implementation.
Organization	Extent of change to organizational routines	Workflow optimization	Anticipated efficiency at baseline; workflow disruption and duplication during rollout; individual adaptations post implementation.
Technology	Knowledge generated	Extraneous data and information overload	Anticipated clutter at baseline became a central frustration during rollout; selective filtering and cognitive habituation emerged post implementation.
Technology	Material properties	System performance and material integration	Early confidence gave way to concerns about specificity, lag, and reliability during rollout; partial refinement occurred postimplementation, though misalignment persisted.
Value proposition and clinical context	Demand-side value	Perceived benefits for workload and safety	Strong optimism at baseline; mixed experiences during rollout; perceived value became pragmatic and context-dependent post implementation.
Value proposition and clinical context	Supply-side value	Credibility and clarity of purpose	Limited understanding of purpose and benefit early on; evolved into clearer but modest recognition of niche utility post implementation.
Adopters	Role and identity	Professional positioning and negotiated use	Curiosity and willingness at baseline; trust declined during rollout due to inconsistency and false positives; cautious, selective engagement stabilized post implementation.
Adopters	Role and identity	Learning and preparedness	Informal self-learning predominated during rollout; structured and ongoing AI[a] literacy training emphasized post implementation.
Wider system	External context	Medicolegal uncertainty and system-level guidance	Policy and liability ambiguity persisted across phases; postimplementation reflections expanded to ethical and regulatory considerations.

[a]AI: artificial intelligence.

Organization

There were 3 inductive subthemes that mapped under the organization domain. There were challenges with sustained implementation planning, which mapped under the NASSS subdomain of "work needed to plan, implement, and monitor change;" relational engagement and communication, which mapped to "organizational readiness and capacity to innovate;" and workflow optimization, which mapped to "extent of change needed to organizational routines."

Work Needed to Plan, Implement, and Monitor Change (Sustained Implementation Planning)

At baseline, participants expected limited planning and support for rollout, reflecting past experiences with digital systems. As one consultant noted:

> *You don't actually discover issues or problems with that new process or software or whatever until you're using it, and then often there's a lack of support on a day-to-day kind of basis.* [P4, Consultant]

Feedback mechanisms were also described as weak, with another adding:

> *There's no way for us to feed…I don't know of a way for me to feed that back.* [P1, Consultant]

These comments illustrated low confidence in the organization's ability to anticipate or respond to implementation challenges.

During peri-implementation, radiologists described minimal systematic planning or training.

> *Not training per se, I think there was one meeting where they said that it was being implemented.* [P20, Registrar]

Another reflected:

> *Not very well (when asked about implementation planning)…they haven't really. Besides telling us that we're going to put it into practice, yeah, there's just not much that they're saying about it.* [P22, Registrar]

Such experiences made participants cautious about the department's readiness to adopt AI, with one consultant admitting:

> *I think in retrospect, we could have done more in terms of educating people.* [P21, Consultant]

Postimplementation reflections reinforced these concerns, highlighting the ongoing absence of structured improvement strategies to support uptake and sustainment. Participants were critical of the ad-hoc implementation process, arguing that:

XSL•FO
RenderX

> *If there are programs that are clinically usable and are planned to be rolled out within the department, then I think it makes sense for everyone to have formal training.* [P33, Registrar]

Even so, there was a strong appetite for more structured professional development in AI tools, with one consultant remarking:

> *I would like to be taught. I am a better learner if I'm taught.* [P32, Consultant]

Across phases, participants viewed planning and monitoring as reactive rather than anticipatory. Despite enthusiasm for AI, the lack of systematic preparation and ongoing learning opportunities constrained the department's ability to embed change effectively. Organizational challenges were compounded by high staff turnover and rotating clinical rosters, which limited continuity of learning and reduced opportunities for cumulative familiarity with the system. These shifting workforce conditions shaped how planning gaps were experienced and helped explain some variation in engagement and confidence across the implementation period.

Organizational Readiness and Capacity to Innovate (Relational Engagement and Communication)

At baseline, participants described a workplace that was broadly supportive of new ideas but slow to coordinate change due to limited capacity. As one consultant noted, "in public systems, people just tend to put up with inefficiencies" (P1, Consultant). Such reflections suggested that innovation was encouraged in principle but rarely matched by structured communication or system support.

During rollout, participants described poor communication and limited coordination between teams.

> *Having [vendor redacted] in one corner, radiologists in another, and us talking together… takes resources to get everyone together.* [P23, Radiographer]

Another reflected postimplementation, "We didn't actually tell most radiographers this was happening" (P41, Radiographer), highlighting the persistence of siloed and reactive coordination. Participants attributed emerging resistance partly to this fragmentation, explaining that "a lot of stuff was happening in the background with the PAX guys and the software people" (P43, Consultant).

Lack of early adopter involvement and unclear lines of responsibility were seen as weakening organizational readiness, even where enthusiasm for innovation remained high. This persisted into postimplementation, undermining the department's sociotechnical capacity to support AI integration. As one consultant explained:

> *One of the issues is that people who understand computers are not the people who understand medicine, and vice versa. So, there's probably a communications issue.* [P29, Consultant]

Participants linked these challenges to the organization's limited capacity to learn from implementation, with one concluding:

> *We could have done more work with the implementation initially; there could have been more clinician involvement.* [P43, Consultant]

Across phases, the organization was perceived as open to innovation but constrained by weak communication channels and reactive coordination. Participants emphasized that the capacity to innovate depended less on enthusiasm than on the presence of structured dialogue, feedback loops, and shared ownership across clinical and technical teams. These relational and coordination challenges intersected with changing workload pressures and fluctuating departmental priorities, reinforcing that uptake was influenced not only by communication structures but by the broader organizational environment in which teams were continuously reconfigured.

Extent of Change Needed to Organizational Routines (Workflow Optimization)

Across phases, participants described the introduction of the AI tool as requiring substantial adjustments to established reporting routines. At baseline, senior radiologists viewed it as a potential aid to workflow optimization, especially in easing registrar workloads. As one consultant put it:

> *If we were able to create a system or facilitate more report completion… particularly for the registrars, that would increase satisfaction.* [P1, Consultant]

This optimism reflected expectations that automation would streamline repetitive elements of reporting rather than disrupt them.

During peri-implementation, participants described the tool as introducing extra steps and interruptions to normal work patterns. One consultant explained:

> *It does add to the amount of things you look at… you've got to report your CT as normal, and then you've got a bunch of other sequences to scroll through at the end. Not that it adds a lot, but yeah, it does… there's more… people have been like, what's all these extra images? They don't really know what to do with it too much yet either.* [P18, Consultant]

This illustrates a lack of clear guidance or established routines for how to use or interpret the additional studies. This persisted at postimplementation:

> *I was sort of holding on to reports for a few hours before signing off because I didn't want additional data to come through that I hadn't looked up before signing off.* [P30, Consultant]

Instead of reducing workload, the new process created pauses, re-checks, and deferred sign-offs. Registrars similarly described difficulty maintaining rhythm and concentration, explaining that:

> *It's hard to get a routine…you have to have a different routine in your workflow for a particular assessment.* [P26, Registrar]

These comments captured how the tool altered the flow of image review and report finalization, requiring constant recalibration of familiar sequences. Some radiologists had developed

XSL•FO
RenderX

compensatory strategies to manage these disruptions; reordering tasks, batching reports, or consciously ignoring low-yield prompts. As one consultant reflected:

> *So for differentials, I go back to the normal scan because the tool obscures some details. So, it's more I use it for identifications…. So, for me, it's more identifying. Ok, there's something there. I need to go back and check (normal scans).* [P31, Consultant]

Overall, participants described workflow change as cumulative and largely unplanned. The tool demanded continuous microadjustments rather than a one-time shift in practice. What emerged was a pattern of individual adaptation rather than coordinated redesign; clinicians modified existing routines to fit the tool rather than the tool being aligned with the established clinical workflow. The extent of workflow disruption experienced by clinicians also reflected the realities of a service marked by rotating staff, shifting caseloads, and variable daily pressures.

Technology

The NASSS technology domain is mapped with 2 inductive subthemes: extraneous data and information overload, aligning with the subdomain knowledge generated, and system performance and material integration, aligning with the subdomain material properties. Together, these themes captured how the technical characteristics of the AI tool shaped user experience, trust, and perceived value across implementation phases.

Knowledge Generated (Extraneous Data and Information Overload)

At baseline, a consultant anticipated risks of information overload based on previous exposure to commercial AI tools. One explained, "You could spend all day circling these things" (P5, Consultant), capturing early concerns that automated outputs might flood readers with marginal or irrelevant findings.

During peri-implementation, these concerns materialized as the system generated excessive, low-value information:

> *Too much data… You really want a traffic-light system.* [P25, Consultant]

Another registrar observed:

> *I'm not sure how many people look at it. It spits out so many images and random tables* [P20, Registrar]

These reactions pointed to an emerging pattern of signal-to-noise imbalance, where radiologists spent more time filtering artefacts than interpreting meaningful results. By postimplementation, some users described partial adaptation, learning to disregard redundant data or mentally triage the AI's output.

> *It gets a little complicated when it picks up things that are artifacts. But yeah, I can work around it.* [P43, Consultant]

However, frustration persisted among others who saw the clutter as undermining efficiency rather than enhancing it:

> *It's a waste of time. It's just clutter, you know? … I usually ignore it.* [P37, Consultant]

Across phases, information overload remained one of the most salient barriers to adoption. While individual users developed coping strategies, these adaptations reflected workaround behavior rather than genuine integration, reinforcing perceptions that the AI's knowledge output was not yet aligned with clinical reasoning or workflow needs.

Material Properties (System Performance and Material Integration)

Performance concerns were a defining feature of the AI's reception, particularly during peri-implementation. A registrar characterized it bluntly as "Not very accurate. Just a splatter approach" (P17, Registrar), reflecting the perception that the system detected excessive findings without adequate specificity. Such errors eroded trust and reduced the incentive to incorporate its output into reporting routines.

By postimplementation, participants expressed more nuanced but still divided views. Some regarded the system as useful for reassurance or cross-checking:

> *Used it more like a check-off — especially when you have things that are complex, and there are a lot of findings.* [P30, Consultant]

Others found the persistent false positives distracting and demoralizing. As one put it:

> *For me to waste time looking at it…it's circled this fecal matter in the splenic flexure.* [P28, Consultant]

Several participants emphasized that perceived technical performance shaped how often they engaged with the system at all. When lag, sensitivity issues, or interface friction increased, clinicians tended to bypass or ignore the tool. Over time, its role shifted from active decision aid to optional background reference, indicating a decline in both trust and functional value.

Interoperability problems surfaced most clearly during peri-implementation, where users described limited integration between the AI software, picture archiving and communication system (PACS), and reporting systems. One consultant explained:

> *That's high-level stuff, right? That's integrating the processing, postprocessing software with the reporting software. But we don't have that capacity.* [P21, Consultant]

They further highlighted redundancy and excess image sets, noting "way too, way too many sequences…we need to distil that down."

By postimplementation, interoperability was less salient; some technical issues with integrating the AI into the system appeared resolved, but residual inconsistencies persisted. As one registrar noted:

> *There's not… uniformity to the sequences that are made. The order that they come out, …that's different from scanner to scanner.* [P26, Registrar]

Availability also varied:

> *It's not always there. So, you've got to sort of remember…to look for.* [P26, Registrar]

XSL•FO
RenderX

Display and PACS constraints continued to affect use:

> *I don't like the way that gets displayed…that's a PACS system…how the series [are] actually displayed.* [P26, Registrar]

Across phases, the material outputs of the AI, its sensitivity, specificity, and responsiveness, directly influenced its perceived usefulness. Participants consistently linked suboptimal performance to disengagement, showing that successful technological integration required not just accuracy, but reliability, responsiveness, and design alignment with radiologists' expectations of diagnostic precision. Furthermore, while major interoperability barriers had eased, the system never fully aligned with the routine reporting infrastructure, leaving incompatibilities.

Across all technology-related subdomains, participants described a gap between what the AI produced and what clinicians could *use*. Information overload and variable system accuracy combined to erode trust and limit engagement. While technical adaptation occurred at the individual level, collective integration into practice remained constrained, signaling that technological refinement and interpretability are prerequisites for sustained adoption.

Value Proposition and Clinical Context

Across all phases, discussions of value proposition were less prominent than those of technology or organization, but two inductive subthemes mapped clearly to the NASSS value proposition domain: perceived benefits for clinical workload and safety, which mapped to demand-side value, and credibility and clarity of purpose, which mapped to supply-side value. These intersected closely with the evolving clinical context, in which fluctuating workload pressures and infrastructure challenges shaped how the AI's value was interpreted.

Demand-Side Value (Perceived Benefits for Clinical Workload and Safety)

At baseline, participants viewed the AI as a potential solution to workload strain and reporting delays. Anticipated benefits were framed around efficiency, redistribution of tasks, and registrar support, signaling early optimism that automation would enhance throughput and safety. The department's intense workload and frequent interruptions reinforced this demand-side appeal: as a consultant noted, AI might "make our job easier" (P1, Consultant).

During peri-implementation, optimism gave way to more conditional appraisals. While some identified benefits for prioritization, "It highlights a few cases that you can look at first. That's useful when there's a backlog" (P19, Registrar), others described it as "unreliable at the moment" (P17, Registrar). Shifts in the clinical environment also tempered expectations, staffing improved, and backlogs eased. By the postimplementation phase, perceptions of value became pragmatic and evidence-driven. Clinicians viewed the AI as a limited but occasionally useful decision support tool:

> *I've usually written my report before I look at this, and I don't tend to change the report… it's another look, I wouldn't think of it as more than that* [P29, Consultant]

Concerns over cost-efficiency persisted:

> *If it was free, ambivalent… If it's significant amounts of money… I don't see the value because it's more work than less.* [P29, Consultant]

Across phases, expectations regarding the AI shifted from broad hopes of efficiency to a more divided assessment. Some saw modest contributions to safety and prioritization, while others viewed the system as duplicating effort rather than providing genuine workload relief.

Supply-Side Value (Credibility and Clarity of Purpose)

At the same time, participants reflected on supply-side value, questioning how clearly the system's purpose and evidence base had been articulated.

> *It just gives you pictures with circles. I'm not sure what the end use is meant to be.* [P20, Registrar]

By postimplementation, participants had a clearer understanding of what the AI could do but remained unconvinced of its overall value. However, there was also recognition that the AI was credible in concept but still immature in delivery.

> *It's getting clearer now what it could be for, but it needs to evolve. Right now, it's still just identifying, not interpreting over time.* [P43, Consultant]

Some viewed potential uses to optimize efficiency and workflow, with modifications:

> *It could identify which studies need to be reported first…or give us measurement readings.* [P46, Consultant]

> *Maybe some of those sorts of irritations around AI could be changed, you know, or fine-tuned.* [P25, Consultant]

These reflections indicated that perceptions of supply-side value were prospective, anchored in what the technology could deliver if optimized, rather than what it had yet achieved.

Across phases, the vendor narrative of innovation and efficiency had not yet translated into tangible or demonstrable benefit for clinicians or the wider health system. Participants viewed the AI as promising but still lacking the evidence and clarity needed to support confident investment or large-scale deployment.

Adopters

The NASSS adopter subdomain of role and identity is mapped with 2 inductive subthemes: professional positioning and negotiated use, and learning and preparedness. Together, these described how clinicians positioned AI within their expertise and accountability, and how limited exposure and training shaped trust, confidence, and uptake. Across phases, adoption reflected an oscillation between curiosity and skepticism, with trust becoming the key mediating factor.

XSL•FO
RenderX

Role and Identity (Professional Positioning and Negotiated Use)

At baseline, radiologists expressed a guarded willingness to engage with AI framed less as enthusiasm and more as a professional obligation.

> *I think I would use it…it would be almost negligent not to look at it.* [P6, Registrar]

Consultants saw potential for practical support:

> *It could certainly help you prioritize what you are watching and what order you report things.* [P8, Consultant]

During peri-implementation, practical experience unsettled this cautious trust. Registrars described false positives, excessive outputs, and low sensitivity:

> *It's too junior at the moment.* [P17, Registrar]

Trust eroded not from resistance to innovation, but from inconsistency between the AI's promise and its performance. Clinicians voiced a recurring sentiment that while AI might one day assist safety, it currently distracts from clinical focus:

> *If the volume of data presented is overwhelming, then that's negative…the strength would be as a safety net for subtle findings, not changing an overall clinical picture.* [P19, Registrar]

Reflecting on their initial use of the tool during peri-implementation, a registrar noted:

> *"It was picking up stuff that wasn't nodules…I still had to go back and look at the images again.* [P34, Registrar]

By postimplementation, there was selective use and partial trust.

> *I look through the nodules myself first and then correlate with the software to see whether it is congruent with what I've come up with.* [P42, Consultant]

Others disengaged entirely:

> *It slows you down because you have to verify each little dot.* [P37, Consultant]

A senior consultant likened the AI to "the registrar with clever ideas, but they're all wrong" (P40, Consultant), useful for prompting review, yet unreliable without human correction.

Trust also intersected with medicolegal anxiety. Several raised uncertainties about accountability and liability:

> *If the software makes a mistake, who is liable—the vendor or the radiologist? We still haven't ironed it out.* [P34, Registrar]

This uncertainty reinforced their instinct to retain manual control. As a consultant observed:

> *If I reported every possible little dot in the chest, I'd end up with a report ten pages long, which nobody would ever read.* [P37, Consultant]

The line between cautious trust and defensive practice remained thin.

Role and Identity (Learning and Preparedness)

Training and readiness remained persistently underdeveloped. During peri-implementation, there was no structured orientation or clear introduction to the system. Learning was largely self-directed and reliant on peer exchange.

> *Personally, I don't think I've had any formal sit-down with it… I've just figured it out.* [P19, Registrar]

A vendor demonstration was held midphase, but not all clinicians attended, and some felt it was disconnected from practical workflow.

> *I just met the software without gathering any prior information about what this new software is.* [P34, Registrar]

Without clear instruction or transparency about performance parameters, early experiences became a process of trial and error rather than guided adoption, reinforcing skepticism instead of trust. By postimplementation, clinicians explicitly called for structured and continuous AI education embedded within clinical and professional frameworks.

> *If that is incorporated into our routine…every month we have our session doing AI cases.* [P32, Consultant]

Others stressed the need for broader institutional responsibility:

> *We are severely lacking in training with AI…it should be an integral, assessed part of our training program.* [P34, Registrar]

These calls reflected not only a desire for technical competence but also a wish to rebuild confidence and ensure medicolegal clarity, positioning AI as a tool that must be professionally standardized, not individually improvised.

Ultimately, clinicians saw AI competence as a new layer of professional literacy, necessary to protect judgment, maintain accountability, and engage critically with emerging tools. Their learning needs were not purely technical but ethical and epistemic: how to weigh evidence, interpret probability, and remain vigilant in an era of shared decision automation.

Wider System

Participants described the wider system as a persistent barrier across phases. This domain reflects the external political, policy, and institutional forces, such as regulation, professional guidance, legislative, and funding models that define the environment in which implementation occurs, but which local teams cannot directly control. While wider system themes were present across phases, they did not dominate every interview, and some subissues (for example, explicit references to legislation) appeared only sporadically.

At baseline, consultants depicted a public system tolerant of inefficiency and difficult to influence. Funding constraints were raised in the context of competing pressures.

> *Q-Health…there's not much money around for these sorts of things.* [P1, Consultant]

During peri-implementation, registrars and consultants highlighted gaps in professional guidance and medicolegal expectations.

XSL•FO
RenderX

> *Training/communication…college says we need to learn AI, but little practical guidance…site-to-site differences.* [P24, Registrar]

Others wanted clearer, proactive communication from professional bodies.

> *College exposure is low.* [P27, Registrar]

Medicolegal norms were seen to expand review obligations when AI added extra views.

> *Mentality in radiology, if it's on a screen, you have to comment on it, and medico-legal, if presented, we must review everything.* [P21, Consultant]

By postimplementation, some brought up broader ethical and policy concerns about data provenance and the social license concerns about using it:

> *There is concern about the way the data is being used…if all these algorithms are being trained on everyone's data, it should be open source…it's everyone's.* [P42, Registrar]

Participants contrasted public and private system incentives and capacity, linking retention and deployment choices to wider economics and case-mix.

> *Public keeps me for complex cases, teaching, and feedback; private pays double.* [P28, Consultant]
>
> *Public vs private…pay and tech are better in private; public has collegiality and case mix.* [P42, Registrar]

Across phases, the wider system was characterized by limited policy levers and still developing structural and legislative readiness to support the rapid integration of AI into acute health care. Taken together, these wider-system influences interacted with internal organizational dynamics, staffing fluctuations, workload variability, and shifting operational priorities to shape the evolving trajectory of implementation across phases.

Discussion

Principal Findings

This study reports a prospective, qualitative, end-to-end evaluation of implementing an AI-driven clinical decision support system in a public radiology department, structured through the NASSS framework [30]. By mapping barriers and enablers across domains and phases, the study captures how early expectations shaped adoption, how sociotechnical challenges emerged during the rollout, and how these dynamics influenced long-term integration and adoption. Findings highlight that successful AI adoption depends not only on technical capability but on the alignment of organizational readiness, workflow design, and professional trust. Implementation success was governed by the interaction of multiple NASSS domains, which included interdependencies among technology, organization, adopters, and value rather than any single factor.

Weak planning and limited feedback structures (organizational barriers) amplified adopter frustrations with false positives, information clutter, and interoperability issues (technology barriers), eroding trust (adopter-level barriers). Even when technical faults were later mitigated, these initial experiences limited or constrained uptake, illustrating how early technical and communication failures created enduring impressions that shaped subsequent patterns of trust and tool use.

This mutual reinforcement of challenges across domains aligns with the complexity perspective described by Greenhalgh et al [30] and Braithwaite et al [40], whereby interacting barriers within and across a complex adaptive system, such as health care, tend to compound rather than resolve over time, particularly when they are not addressed in a coordinated and simultaneous manner. Clinicians' perceptions of value were shaped primarily by how reliably and efficiently the AI system performed within everyday reporting workflows. Early false positives and excessive image sets undermined those expectations, diminishing confidence in the tool's promised efficiency benefits. This finding is consistent with a recent qualitative study showing that workflow fit determines perceived usefulness, even for AI [41]. Participants described the system as "a check-off" tool rather than an integrated aid, and acceptance was suboptimal across our study; consistent with a 2023 semistructured interview with radiologists (n=25), which identified reliability, interpretability, and feedback transparency as decisive for AI acceptance [14]. Even after performance improved, initial mistrust persisted. This enduring skepticism also mirrors broader evidence that initial experiences set adoption trajectories and that trust is far easier to lose than to regain [42]. The finding reinforces the importance of consolidating the first-use experience through predeployment testing or "shadow mode" configurations [43].

Organizational conditions were vital in shaping clinician engagement. During peri-implementation, fragmented communication and limited training left some staff unaware that the system had gone live, while others lacked the confidence to use it effectively. This mirrors the work of our team and others who have consistently highlighted that structured rollout, anticipatory planning, and capacity building are critical for sustainable digital adoption [32,44-47]. Participants described the process as reactive and isolated rather than coordinated, reflecting the absence of a shared sense of purpose and mutual accountability between leadership, implementers, and users, necessary for effective implementation [48]. Although clinicians remained receptive to AI, they expected visible organizational commitment through ongoing education, rapid troubleshooting, and coherent leadership. In its absence, they relied on informal workarounds and peer support to conduct work-as-done, the adaptive, improvised practices that frontline staff develop to keep systems functioning when formal processes or resources fall short [49]. While such adaptations can sustain local functionality, they also introduce variability in care delivery and make it difficult to scale and standardize best practices across settings.

These organizational shortfalls also shaped how clinicians experienced implementation. In the absence of clear planning and coordination, several described feeling individually responsible for interpreting and integrating the AI tool into their workflow. This was not an explicit transfer of responsibility but reflected a professional culture in which clinicians relied on their own judgement to make the system workable within

XSL•FO
RenderX

local constraints. Medicolegal uncertainty about liability further reinforced this guarded engagement. A 2021 narrative review observed that when accountability is unclear or diffuse, clinicians maintain human oversight to protect both patient safety and professional authority [50]. These dynamics highlight a core implementation challenge: without clear institutional responsibility and evidentiary assurance, professional caution becomes self-reinforcing, constraining experimentation, shared learning, and the normalization of AI within routine practice.

The single-site public hospital context influenced organizational capacity. Frequent staff rotations and high turnover meant that many clinicians engaged with the AI system intermittently, limiting opportunities for cumulative learning. Similar patterns are common across public health services, where workforce mobility and resource constraints make it difficult to sustain iterative improvement [51]. These dynamics interacted with the implementation process itself, shaping the pace and pattern of adoption and modulating how performance issues were perceived over time. Rather than functioning as external confounders, they formed part of the organizational ecology within which the AI system was introduced, influencing continuity, familiarity, and the stability of feedback loops essential for embedding new technologies. These conditions underscore the importance of implementation approaches that are designed for continuity, including modular onboarding, periodic refresher training, and accessible repositories of AI-related resources to support learning across changing teams.

This study extends the AI-in-radiology literature in 3 key ways. It adopts a temporal perspective, tracing implementation from early anticipation to postintegration and showing how initial optimism and early missteps shape later engagement. It also demonstrates that adoption was driven not by the innovation alone but by the interaction of technical performance, organizational coordination, and context, extending NASSS from description to explanation. In this framing, staff turnover, workload fluctuations, and shifting operational priorities function as active contextual determinants that help explain why implementation trajectories evolve as they do, rather than as background noise. Finally, through reflexive use of the framework, the study generates theoretical insight into why implementation evolves as it does, contributing to emerging work on domain interdependence and temporal complexity in health care innovation [52].

Grounded in our findings and consistent with previous guidance [23,30,53,54], effective AI implementation in radiology depends on a combination of technical stability, communication, and organizational preparedness. Early “shadow mode” piloting helps identify faults and build trust before clinical use, while consistent communication about progress and fixes maintains transparency. Workflow-compatible design that minimizes cognitive load supports efficiency and acceptance [55]. Ongoing professional development, formal feedback channels with vendor responsiveness, and planning for workforce turnover through shared training repositories and local champions help sustain capability over time. Together, these strategies align with the six principles of FUTURE-AI recommendations encompassing fairness, universality, traceability, usability, robustness, and explainability, an international expert-driven consensus to facilitate the adoption of trustworthy medical imaging [56] and emphasize that co-design and iterative learning are essential to long-term adoption.

Study Limitations and Strengths

This was a single-site study, and further real-world, ecological research is needed to identify determinants of adoption and create solutions that generalize across health systems. Despite this, our findings are similar to several recently published studies examining radiologists’ perceptions around the adoption of AI into standard practice [2,12,14,57-60]. There was low uptake of the tool during the peri-implementation period due to technical challenges. This limited wider evaluation of this crucial implementation period, particularly in terms of clinical usefulness. The 18-month period may have also introduced a range of system confounders, including staffing changes and organizational priorities, which may have impacted how participants felt about the AI clinical decision support tool. Finally, social desirability bias cannot be ruled out, particularly as this was a department-wide implementation. However, this was mitigated through participant briefings emphasizing the exploratory nature of the trial, coupled with reflexive practice by the interview team [61]. A key study strength was that this was a real-world evaluation demonstrating ecological validity with findings grounded in practical realities. Second, aligning our findings with a validated implementation science framework supports theoretical transferability and future application in related contexts despite the single-site limitations.

Future Implications

Future studies should integrate qualitative and quantitative data, combining workflow observations with metrics such as reporting time, error rates, and AI usage logs, to triangulate findings. Multisite evaluations across differing levels of digital maturity are needed to test transferability and examine how governance, culture, and workforce patterns influence scalability.

Conclusion

Implementation of AI-based decision support in radiology is as much an organizational and cultural process as a technological one. Clinicians remain willing to engage, but sustainable adoption depends on consolidating early experiences, embedding communication and training, and maintaining iterative feedback between users, vendors, and system leaders. Applying the NASSS framework revealed how domains interact dynamically across time, offering both theoretical insight into sociotechnical complexity and practical guidance for hospitals seeking to move from pilot to routine, trustworthy AI integration.

Acknowledgments

The authors would like to gratefully acknowledge the study participants who gave their valuable time to participate in this study. The authors are grateful to Dr Sue Jeavons and her staff for facilitating this study onsite.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are not publicly available due to confidentiality policies, but may be available in limited form from the corresponding author on reasonable request.

Funding

This work was supported by Digital Health CRC Limited (DHCRC), funded under the Commonwealth Cooperative Research Centres (CRC) program. SM is supported by a fellowship from the National Health and Medical Research Council (NHMRC; #1181138). The funders had no role in the study design or the decision to submit for publication.

Authors' Contributions

Conceptualization: SN, SM
Methodology: SN, PS, BS, AT, AE, SM
Formal analysis: SN, AE
Writing – original draft: SN
Writing – review & editing: SN, PS, BS, AT, AE, SM

Conflicts of Interest

None declared.

Multimedia Appendix 1
COREQ (Consolidated Criteria for Reporting Qualitative Studies) checklist.
[PDF File (Adobe PDF File), 486 KB - jmir_v28i1e80342_app1.pdf]

Multimedia Appendix 2
Interview topic guide.
[DOCX File , 29 KB - jmir_v28i1e80342_app2.docx]

References

1. Maskell G. Why Does Demand for Medical Imaging Keep Rising?. Hoboken, NJ: British Medical Journal Publishing Group; 2022.
2. Chauhan AS, Singh R, Priyadarshi N, Twala B, Suthar S, Swami S. Unleashing the power of advanced technologies for revolutionary medical imaging: pioneering the healthcare frontier with artificial intelligence. Discov Artif Intell 2024;4(1):58. [doi: 10.1007/s44163-024-00161-0]
3. Ding A, Joshi J, Tiwana E. Patient safety in radiology and medical imaging. In: Agrawal A, Bhatt J, editors. Patient Safety: A Case-Based Innovative Playbook for Safer Care. Cham: Springer; 2023:261-277.
4. Waite S, Scott J, Colombo D. Narrowing the Gap: imaging disparities in radiology. Radiology 2021;299(1):27-35. [doi: 10.1148/radiol.2021203742] [Medline: 33560191]
5. Kelly BS, Judge C, Bollard SM, Clifford SM, Healy GM, Aziz A, et al. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). Eur Radiol 2022;32(11):7998-8007 [FREE Full text] [doi: 10.1007/s00330-022-08784-6] [Medline: 35420305]
6. Chau M. Ethical, legal, and regulatory landscape of artificial intelligence in Australian healthcare and ethical integration in radiography: a narrative review. J Med Imaging Radiat Sci 2024;55(4):101733. [doi: 10.1016/j.jmir.2024.101733] [Medline: 39111223]
7. Lu P, Mian M, Yii M, McArdle DJT, Rhodes A, Sreedharan S. Rising use of diagnostic imaging in Australia: an analysis of medicare-funded radiology services between 2000 and 2021. J Med Imaging Radiat Oncol 2024;68(1):50-56. [doi: 10.1111/1754-9485.13591] [Medline: 37797195]
8. Zheng Q, Yang L, Zeng B, Li J, Guo K, Liang Y, et al. Artificial intelligence performance in detecting tumor metastasis from medical radiology imaging: a systematic review and meta-analysis. EClinicalMedicine 2021;31:100669 [FREE Full text] [doi: 10.1016/j.eclinm.2020.100669] [Medline: 33392486]
9. van den Broek MCL, Buijs JH, Schmitz LFM, Wijffels MME. Diagnostic performance of artificial intelligence in rib fracture detection: systematic review and meta-analysis. Surgeries 2024;5(1):24-36. [doi: 10.3390/surgeries5010005]
10. Lawrence R, Dodsworth E, Massou E, Sherlaw-Johnson C, Ramsay AI, Walton H, et al. Artificial intelligence for diagnostics in radiology practice: a rapid systematic scoping review. EClinicalMedicine 2025;83:103228 [FREE Full text] [doi: 10.1016/j.eclinm.2025.103228] [Medline: 40474995]
11. Sosna J, Joskowicz L, Saban M. Navigating the AI landscape in medical imaging: a critical analysis of technologies, implementation, and implications. Radiology 2025;315(3):e240982. [doi: 10.1148/radiol.240982] [Medline: 40552997]

XSL•FO
RenderX

12. Strohm L, Hehakaya C, Ranschaert ER, Boon WPC, Moors EHM. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. Eur Radiol 2020;30(10):5525-5532. [doi: 10.1007/s00330-020-06946-y] [Medline: 32458173]
13. Alarifi M. Radiologists' views on AI and the future of radiology: insights from a US national survey. Br J Radiol 2025. [doi: 10.1093/bjr/tqaf222] [Medline: 40971485]
14. Bergquist M, Rolandsson B, Gryska E, Laesser M, Hoefling N, Heckemann R, et al. Trust and stakeholder perspectives on the implementation of AI tools in clinical radiology. Eur Radiol 2024;34(1):338-347 [FREE Full text] [doi: 10.1007/s00330-023-09967-5] [Medline: 37505245]
15. Hua D, Petrina N, Young N, Cho J, Poon SK. Understanding the factors influencing acceptability of AI in medical imaging domains among healthcare professionals: a scoping review. Artif Intell Med 2024;147:102698 [FREE Full text] [doi: 10.1016/j.artmed.2023.102698] [Medline: 38184343]
16. Ross J, Hammouche S, Chen Y, Rockall A, Royal College of Radiologists AI Working Group. Beyond regulatory compliance: evaluating radiology artificial intelligence applications in deployment. Clin Radiol 2024;79(5):338-345 [FREE Full text] [doi: 10.1016/j.crad.2024.01.026] [Medline: 38360516]
17. Arkoh S, Akudjedu TN, Amedu C, Antwi WK, Elshami W, Ohene-Botwe B. Current radiology workforce perspective on the integration of artificial intelligence in clinical practice: a systematic review. J Med Imaging Radiat Sci 2025;56(1):101769. [doi: 10.1016/j.jmir.2024.101769] [Medline: 39437624]
18. Hemphill S, Jackson K, Bradley S, Bhartia B. The implementation of artificial intelligence in radiology: a narrative review of patient perspectives. Future Healthc J 2023;10(1):63-68 [FREE Full text] [doi: 10.7861/fhj.2022-0097] [Medline: 37786489]
19. Wenderott K, Krups J, Weigl M, Wooldridge AR. Facilitators and barriers to implementing AI in routine medical imaging: systematic review and qualitative analysis. J Med Internet Res 2025;27:e63649 [FREE Full text] [doi: 10.2196/63649] [Medline: 40690758]
20. Buijs E, Maggioni E, Mazziotta F, Lega F, Carrafiello G. Clinical impact of AI in radiology department management: a systematic review. Radiol Med 2024;129(11):1656-1666 [FREE Full text] [doi: 10.1007/s11547-024-01880-1] [Medline: 39243293]
21. Kim B, Romeijn S, van Buchem M, Mehrizi MHR, Grootjans W. A holistic approach to implementing artificial intelligence in radiology. Insights Imaging 2024;15(1):22 [FREE Full text] [doi: 10.1186/s13244-023-01586-4] [Medline: 38270790]
22. Hogg HDJ, Al-Zubaidy M, Technology Enhanced Macular Services Study Reference Group, Talks J, Denniston AK, Kelly CJ, et al. Stakeholder perspectives of clinical artificial intelligence implementation: systematic review of qualitative evidence. J Med Internet Res 2023;25:e39742 [FREE Full text] [doi: 10.2196/39742] [Medline: 36626192]
23. Hogg HDJ, Sendak MP, Denniston AK, Keane PA, Maniatopoulos G. Unlocking the potential of qualitative research for the implementation of artificial intelligence-enabled healthcare. J Med Artif Intell 2023;6:8-8. [doi: 10.21037/jmai-23-28]
24. Jarrahi MH. Interviewing AI: Using qualitative methods to explore and capture machines' characteristics and behaviors. Big Data & Society 2025;12(3). [doi: 10.1177/20539517251381697]
25. Farič N, Hinder S, Williams R, Ramaesh R, Bernabeu M, van Beek E, et al. Early experiences of integrating an artificial intelligence-based diagnostic decision support system into radiology settings: a qualitative study. J Am Med Inform Assoc 2023;31(1):24-34 [FREE Full text] [doi: 10.1093/jamia/ocad191] [Medline: 37748456]
26. Song Y, MacEachern L, Doupe MB, Ginsburg L, Chamberlain SA, Cranley L, et al. Influences of post-implementation factors on the sustainability, sustainment, and intra-organizational spread of complex interventions. BMC Health Serv Res 2022;22(1):666 [FREE Full text] [doi: 10.1186/s12913-022-08026-x] [Medline: 35581651]
27. Donovan T, Carter HE, McPhail SM, Abell B. Challenges and recommendations for collecting and quantifying implementation costs in practice: a qualitative interview study. Implement Sci Commun 2024;5(1):114 [FREE Full text] [doi: 10.1186/s43058-024-00648-y] [Medline: 39394175]
28. Fernando M, Abell B, Tyack Z, Donovan T, McPhail SM, Naicker S. Using theories, models, and frameworks to inform implementation cycles of computerized clinical decision support systems in tertiary health care settings: scoping review. J Med Internet Res 2023;25:e45163 [FREE Full text] [doi: 10.2196/45163] [Medline: 37851492]
29. Lewis CC, Boyd MR, Walsh-Bailey C, Lyon AR, Beidas R, Mittman B, et al. A systematic review of empirical studies examining mechanisms of implementation in health. Implement Sci 2020;15(1):21 [FREE Full text] [doi: 10.1186/s13012-020-00983-3] [Medline: 32299461]
30. Greenhalgh T, Abimbola S. The NASSS framework - a synthesis of multiple theories of technology implementation. Stud Health Technol Inform 2019;263:193-204. [doi: 10.3233/SHTI190123] [Medline: 31411163]
31. Abell B, Naicker S, Rodwell D, Donovan T, Tariq A, Baysari M, et al. Identifying barriers and facilitators to successful implementation of computerized clinical decision support systems in hospitals: a NASSS framework-informed scoping review. Implement Sci 2023;18(1):32 [FREE Full text] [doi: 10.1186/s13012-023-01287-y] [Medline: 37495997]
32. Fernando M, Abell B, McPhail SM, Tyack Z, Tariq A, Naicker S. Applying the non-adoption, abandonment, scale-up, spread, and sustainability framework across implementation stages to identify key strategies to facilitate clinical decision support system integration within a large metropolitan health service: interview and focus group study. JMIR Med Inform 2024;12:e60402 [FREE Full text] [doi: 10.2196/60402] [Medline: 39419497]

XSL•FO
RenderX

33. Cordeiro JV. Artificial intelligence and precision public health: a balancing act of scientific accuracy, social responsibility, and community engagement. Port J Public Health 2024;42(1):1-5. [doi: 10.1159/000538141] [Medline: 39495190]
34. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907 [FREE Full text] [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
35. Rettke H, Pretto M, Spichiger E, Frei I, Spirig R. Using reflexive thinking to establish rigor in qualitative research. Nurs Res 2018;67(6):490-497. [doi: 10.1097/NNR.0000000000000307] [Medline: 30067583]
36. Nordmann K, Sauter S, Redlich M, Möbius-Lerch P, Schaller M, Fischer F. Challenges and conditions for successfully implementing and adopting the telematics infrastructure in German outpatient healthcare: a qualitative study applying the NASSS framework. Digital Health 2024;10:20552076241259855 [FREE Full text] [doi: 10.1177/20552076241259855] [Medline: 39070890]
37. Proudfoot K. Inductive/deductive hybrid thematic analysis in mixed methods research. J Mix Methods Res 2022;17(3):308-326. [doi: 10.1177/15586898221126816]
38. Skillman M, Cross-Barnet C, Friedman Singer R, Rotondo C, Ruiz S, Moiduddin A. A framework for rigorous qualitative research as a component of mixed method rapid-cycle evaluation. Qual Health Res 2019;29(2):279-289. [doi: 10.1177/1049732318795675] [Medline: 30175660]
39. Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: implications for conducting a qualitative descriptive study. Nurs Health Sci 2013;15(3):398-405. [doi: 10.1111/nhs.12048] [Medline: 23480423]
40. Braithwaite J, Churruca K, Long JC, Ellis LA, Herkes J. When complexity science meets implementation science: a theoretical and empirical analysis of systems change. BMC Med 2018;16(1):63 [FREE Full text] [doi: 10.1186/s12916-018-1057-z] [Medline: 29706132]
41. Wenderott K, Krups J, Luetkens JA, Weigl M. Radiologists' perspectives on the workflow integration of an artificial intelligence-based computer-aided detection system: a qualitative study. Appl Ergon 2024;117:104243 [FREE Full text] [doi: 10.1016/j.apergo.2024.104243] [Medline: 38306741]
42. Steerling E, Siira E, Nilsen P, Svedberg P, Nygren J. Implementing AI in healthcare-the relevance of trust: a scoping review. Front Health Serv 2023;3:1211150. [doi: 10.3389/frhs.2023.1211150] [Medline: 37693234]
43. Bizzo BC, Dasegowda G, Bridge C, Miller B, Hillis JM, Kalra MK, et al. Addressing the challenges of implementing artificial intelligence tools in clinical practice: principles from experience. J Am Coll Radiol 2023;20(3):352-360. [doi: 10.1016/j.jacr.2023.01.002] [Medline: 36922109]
44. Alotaibi N, Wilson CB, Traynor M. Enhancing digital readiness and capability in healthcare: a systematic review of interventions, barriers, and facilitators. BMC Health Serv Res 2025;25(1):500 [FREE Full text] [doi: 10.1186/s12913-025-12663-3] [Medline: 40186200]
45. Van Velthoven MH, Cordon C. Sustainable adoption of digital health innovations: perspectives from a stakeholder workshop. J Med Internet Res 2019;21(3):e11922 [FREE Full text] [doi: 10.2196/11922] [Medline: 30907734]
46. Nair M, Nygren J, Nilsen P, Gama F, Neher M, Larsson I, et al. Critical activities for successful implementation and adoption of AI in healthcare: towards a process framework for healthcare organizations. Front Digit Health 2025;7:1550459 [FREE Full text] [doi: 10.3389/fdgth.2025.1550459] [Medline: 40453810]
47. Marwaha JS, Landman AB, Brat GA, Dunn T, Gordon WJ. Deploying digital health tools within large, complex health systems: key considerations for adoption and implementation. NPJ Digit Med 2022;5(1):13 [FREE Full text] [doi: 10.1038/s41746-022-00557-1] [Medline: 35087160]
48. Rapport F, Clay-Williams R, Churruca K, Shih P, Hogden A, Braithwaite J. The struggle of translating science into action: foundational concepts of implementation science. J Eval Clin Pract 2018;24(1):117-126 [FREE Full text] [doi: 10.1111/jep.12741] [Medline: 28371050]
49. Perry SJ, Catchpole K, Rivera AJ, Henrickson Parker S, Gosbee J. 'Strangers in a strange land': Understanding professional challenges for human factors/ergonomics and healthcare. Appl Ergon 2021;94:103040 [FREE Full text] [doi: 10.1016/j.apergo.2019.103040] [Medline: 33676061]
50. Smith H. Clinical AI: opacity, accountability, responsibility and liability. AI & Society 2020;36(2):535-545. [doi: 10.1007/s00146-020-01019-6]
51. Moon SEJ, Hogden A, Eljiz K. Sustaining improvement of hospital-wide initiative for patient safety and quality: a systematic scoping review. BMJ Open Qual 2022;11(4):e002057 [FREE Full text] [doi: 10.1136/bmjoq-2022-002057] [Medline: 36549751]
52. Hellstrand Tang U, Smith F, Karilampi UL, Gremyr A. Exploring the role of complexity in health care technology bottom-up innovations: multiple-case study using the nonadoption, abandonment, scale-up, spread, and sustainability complexity assessment tool. JMIR Hum Factors 2024;11:e50889 [FREE Full text] [doi: 10.2196/50889] [Medline: 38669076]
53. Daye D, Wiggins WF, Lungren MP, Alkasab T, Kottler N, Allen B, et al. Implementation of clinical artificial intelligence in radiology: who decides and how. Radiology 2022;305(3):555-563 [FREE Full text] [doi: 10.1148/radiol.212151] [Medline: 35916673]

XSL•FO
RenderX

54. Chen Y, Stavropoulou C, Narasinkan R, Baker A, Scarbrough H. Professionals' responses to the introduction of AI innovations in radiology and their implications for future adoption: a qualitative study. BMC Health Serv Res 2021;21(1):813 [FREE Full text] [doi: 10.1186/s12913-021-06861-y] [Medline: 34389014]
55. Zhang L, LaBelle W, Unberath M, Chen H, Hu J, Li G, et al. A vendor-agnostic, PACS integrated, and DICOM-compatible software-server pipeline for testing segmentation algorithms within the clinical radiology workflow. Front Med (Lausanne) 2023;10:1241570 [FREE Full text] [doi: 10.3389/fmed.2023.1241570] [Medline: 37954555]
56. NA. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ 2025;388:r340 [FREE Full text] [doi: 10.1136/bmj.r340] [Medline: 39961614]
57. Huisman M, Ranschaert E, Parker W, Mastrodicasa D, Koci M, Pinto de Santos D, et al. An international survey on AI in radiology in 1041 radiologists and radiology residents part 2: expectations, hurdles to implementation, and education. Eur Radiol 2021;31(11):8797-8806 [FREE Full text] [doi: 10.1007/s00330-021-07782-4] [Medline: 33974148]
58. Jiang S, Bukhari SMA, Krishnan A, Bera K, Sharma A, Caovan D, et al. Deployment of artificial intelligence in radiology: strategies for success. AJR Am J Roentgenol 2025;224(2):e2431898. [doi: 10.2214/AJR.24.31898] [Medline: 39475198]
59. Do HM, Spear LG, Nikpanah M, Mirmomen SM, Machado LB, Toscano AP, et al. Augmented radiologist workflow improves report value and saves time: a potential model for implementation of artificial intelligence. Acad Radiol 2020;27(1):96-105 [FREE Full text] [doi: 10.1016/j.acra.2019.09.014] [Medline: 31818390]
60. Jones C, Thornton J, Wyatt JC. Enhancing trust in clinical decision support systems: a framework for developers. BMJ Health Care Inform 2021;28(1):e100247 [FREE Full text] [doi: 10.1136/bmjhci-2020-100247] [Medline: 34088721]
61. Cairns-Lee H, Lawley J, Tosey P. Enhancing researcher reflexivity about the influence of leading questions in interviews. J Appl Behav Sci 2021;58(1):164-188. [doi: 10.1177/00218863211037446]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
COREQ: Consolidated Criteria for Reporting Qualitative Studies
CT: computed tomography
ERIC: Expert Recommendations for Implementing Change
NASSS: nonadoption, abandonment, scale-up, spread, and sustainability
PACS: picture archiving and communication system

Edited by J Sarvestan; submitted 09.Jul.2025; peer-reviewed by S Mohanadas, L Laverty, X Liang; comments to author 04.Aug.2025; revised version received 09.Dec.2025; accepted 10.Dec.2025; published 28.Jan.2026.

Please cite as:
Naicker S, Schmidt P, Shar B, Tariq A, Earnshaw A, McPhail S
Implementing an Artificial Intelligence Decision Support System in Radiology: Prospective Qualitative Evaluation Study Using the Nonadoption Abandonment Scale-Up, Spread, and Sustainability (NASSS) Framework
J Med Internet Res 2026;28:e80342
URL: https://www.jmir.org/2026/1/e80342
doi:10.2196/80342
PMID:

©Sundresan Naicker, Paul Schmidt, Bruce Shar, Amina Tariq, Ashleigh Earnshaw, Steven McPhail. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Integrated Prediction System for Individualized Ovarian Stimulation and Ovarian Hyperstimulation Syndrome Prevention: Algorithm Development and Validation

Jingjing Chen[1,2*], MD, PhD; Jianjuan Zhao[3,4,5*], MS; Huiyu Qiu[1,2], MD, MS; Yanhui Liu[5], MD, PhD; Yunqi Zhang[3], MS; Qicheng Sun[3], BSc; Yan Yi[1,2], MD, PhD; Hongying Tang[1,2], MSN; Jing Zhao[1,2], MD, PhD; Bin Xu[1,2], MD, PhD; Qiong Zhang[1,2], MD, PhD; Ge Yang[6], MD, PhD; Hui Li[1,2], MD, PhD; Junjie Liu[5], PhD; Zhongzhou Yang[7], PhD; Shaolin Liang[3,4,5*], PhD; Yanping Li[1,2*], MD, PhD; Jing Fu[1,2*], MD, PhD

[1]Department of Reproductive Medicine, Xiangya Hospital, Central South University, Changsha, Hunan, China
[2]Clinical Research Center for Women's Reproductive Health in Hunan Province, Changsha, China
[3]Digital Health Lab, Institute for Six-sector Economy, Fudan University, Shanghai, China
[4]Xiangya Hospital, Central South University, Changsha, China
[5]Reproductive Medicine Department, The Third Affiliated Hospital of Shenzhen University, Shenzhen, China
[6]Division of Neonatology, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangdong, China
[7]School of Medicine, Sun Yat-sen University, Guangdong, China
[*]these authors contributed equally

Corresponding Author:
Jing Fu, MD, PhD
Department of Reproductive Medicine
Xiangya Hospital
Central South University
No.87 Xiangya Road, Kaifu Strict, Changsha City
Changsha, Hunan, 410008
China
Phone: 86 15874861692
Email: 360546450@qq.com

Abstract

Background: Accurately predicting ovarian response and determining the optimal starting dose of follicle-stimulating hormone (FSH) remain critical yet challenging for effective ovarian stimulation. Currently, there is a lack of a comprehensive model capable of simultaneously forecasting the number of oocytes retrieved (NOR) and assessing the risk of early-onset moderate-to-severe ovarian hyperstimulation syndrome (OHSS).

Objective: This study aimed to establish an integrated mode capable of forecasting the NOR and assessing the risk of early-onset moderate-to-severe OHSS across varying starting doses of FSH.

Methods: This prognostic study included patients undergoing their first ovarian stimulation cycles at 2 independent in vitro fertilization clinics. Automated classifiers were used for variable selection. Machine learning models (11 for NOR and 11 for OHSS) were developed and validated using internal (n=6401) and external (n=3805) datasets. Shapley additive explanation was applied for variable interpretation. The best-performing models were incorporated into a web-based prediction tool.

Results: For NOR prediction, 17 variables were selected, with the gradient boosting regressor achieving the highest performance (internal dataset: R^2=0.7978; external dataset: R^2=0.7924). For OHSS prediction, 19 variables were identified, and the LightGBM model demonstrated superior performance (internal dataset: area under the receiver operating characteristic curve=0.7588; external dataset: area under the receiver operating characteristic curve=0.7287). Shapley additive explanation analysis highlighted the FSH starting dose to BMI ratio and baseline antral follicle count as key predictors for NOR and OHSS, respectively. Dose-response curves were generated to visualize predicted outcomes with varying FSH starting doses. The models were implemented in a user-friendly, research-oriented online prototype, individualized ovarian stimulation guide (InOvaSGuide).

XSL•FO
RenderX

Conclusions: This study introduces an integrated framework for predicting NOR and early-onset moderate-to-severe OHSS risk across different FSH doses. Future prospective evaluation is needed before clinical implementation.

(J Med Internet Res 2026;28:e78245) doi:10.2196/78245

KEYWORDS

individualized ovarian stimulation; ovarian response; ovarian hyperstimulation syndrome; follicle-stimulating hormone; machine learning

Introduction

Over the past decade, individualized ovarian stimulation has become a key strategy in in vitro fertilization (IVF). Determining an appropriate starting dose of exogenous follicle-stimulating hormone (FSH) is essential for balancing efficacy and safety. Although earlier clinical practice emphasized maximizing oocyte yield (the more, the better), current consensus favors achieving a moderate ovarian response to optimize live birth rates while minimizing patient discomfort and iatrogenic risks such as ovarian hyperstimulation syndrome (OHSS). Therefore, accurate prediction of ovarian response before stimulation is critical for optimizing treatment outcomes [1-4].

Although biomarkers, including antral follicle count (AFC), anti-Müllerian hormone (AMH) levels, and BMI, are well associated with ovarian response, substantial interindividual and intraindividual variability limits their predictive precision. Tailoring FSH doses based solely on these indicators has not consistently improved clinical outcomes [5,6], highlighting the need for more comprehensive, data-driven approaches that integrate a broader spectrum of clinical and biological factors.

Recent advances in artificial intelligence (AI) and machine learning (ML) offer new opportunities for improving decision-making in assisted reproduction, with applications reported in semen analysis [7], blastocysts grading [8], and trigger-day assessments [9]. Several ML models have also been developed to predict the number of oocytes retrieved (NOR) [10-13] or to classify ovarian responsiveness [10]; however, most remain limited in scope. They typically rely on a narrow set of baseline features, adopt single-model frameworks, and focus predominantly on treatment efficacy such as oocyte yield, with relatively limited attention to safety outcomes, including OHSS. These limitations emphasize the need for predictive frameworks that simultaneously incorporate both efficacy and safety. Furthermore, despite multiple evidence-based algorithms for FSH dosing, considerable variability in ovarian response persists even among patients with comparable baseline characteristics. A model that jointly predicts NOR and OHSS risk across a range of FSH doses may provide useful predictive information and support dose-specific decision-making, helping clinicians consider the balance between efficacy and safety when selecting individualized FSH doses.

In this study, ML models were developed to predict NOR and early-onset moderate-to-severe OHSS using datasets from 2 IVF centers. Models with optimal performance were integrated into a clinician-oriented decision support prototype, termed individualized ovarian stimulation guide (“InOvaSGuide”), complemented by a web-based calculator. For each patient, the system provides individualized dose-response curves that display predicted NOR and early-onset moderate-to-severe OHSS probabilities across varying FSH starting doses, thus supporting personalized ovarian stimulation.

Methods

Ethical Considerations

This prognostic study was designed as a retrospective analysis and was approved by the Reproductive Medicine Ethics Committee of Xiangya Hospital (2021010) and the Medicine Ethics Committee of Shenzhen Luohu District People’s Hospital (2024-LHQRMYY-KYLL-63). Informed consent was waived because all data were retrospectively collected from routine clinical records and anonymized before analysis. The study adhered to the Declaration of Helsinki and followed the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis; Table S1 in Multimedia Appendix 1) reporting guideline [14]. All data analyses were performed by an external team using anonymized data only, ensuring full protection of participant privacy. No compensation was provided to participants, as the study involved retrospective and fully anonymized data.

Study Cohort

The inclusion criteria were (1) patients with the first ovarian stimulation cycle conducted between January 1, 2018, and September 30, 2022, at the Department of Reproductive Medicine of Xiangya Hospital (internal dataset) and between May 1, 2021, and December 30, 2023, at the Reproductive Center of Shenzhen Luohu District People’s Hospital (external dataset) and (2) patients aged 20 to 40 years. Exclusion criteria were (1) patients with diminished ovarian reserve, diagnosed by AMH ≤1.1 ng/mL or baseline AFC ≤7 [15]; (2) patients using the microstimulation protocols for ovarian stimulation, including progestin-primed ovarian stimulation protocol, natural cycle protocol, etc; and (3) patients with more than 50% missingness in key clinical variables. Notably, patients with diminished ovarian reserve or those undergoing microstimulation protocols were excluded because these groups require highly individualized stimulation strategies, exhibit markedly lower oocyte yield, and have a substantially reduced risk of OHSS under comparable FSH exposure, which would have created pronounced class imbalance and reduced model robustness.

After screening, 6401 patients from Xiangya Hospital and 3805 from Shenzhen Luohu District People’s Hospital were included in the internal and external datasets, respectively.

XSL•FO
RenderX

Ovarian Stimulation Process and the Diagnosis of Early-Onset Moderate-to-Severe OHSS

Before commencing the IVF and intracytoplasmic sperm injection cycle, patients underwent a thorough physical examination, including the assessment of basic physical parameters (height, weight, and BMI), measurement of basal hormone levels (FSH, luteinizing hormone [LH], estradiol, testosterone, progesterone, prolactin, and AMH), biochemical tests (fasting glucose, fasting insulin, lipid levels, and thyroid hormones), and transvaginal ultrasonography for basal AFC. Subsequently, experienced physicians personalized the ovarian stimulation protocol and the starting dose of FSH based on comprehensive clinical assessment. Throughout stimulation, patients underwent monitoring via transvaginal ultrasonography and serum hormone assessments, with gonadotropin dosage adjustments made according to individual ovarian responses. Human chorionic gonadotropin or gonadotropin-releasing hormone agonist, alone or combined, triggered oocyte maturation when 3 or more follicles measuring 17 mm or greater were observed. Oocyte retrieval occurred 36 hours after triggering, and the NOR was recorded. Eligible patients underwent fresh embryo transfer with 1 or 2 embryos.

The study primarily focused on the occurrence of early-onset, moderate-to-severe OHSS, which was diagnosed within 9 days after triggering based on established guidelines [16], considering both clinical and laboratory features. All relevant individual and clinical variables during the process were obtained from the clinical database for feature screening and selection.

Data Preprocessing and Feature Selection

All data analysis was performed by an external team using anonymized data, ensuring full protection of participant privacy. To elucidate correlations between predictive features and clinical outcomes with an emphasis on medical interpretability, we performed feature engineering on selected variables. This process resulted in 2 additional variables: “FSH to LH ratio” and “FSH starting dose to BMI ratio,” which improved predictive accuracy while maintaining transparency and clinical relevance. Furthermore, to address skewness and improve distributional normality, a logarithmic transformation was applied to AMH, triglycerides, and FSH/LH (Figures S1 and S2 in Multimedia Appendix 1). Missing data were handled using mean imputation, with feature-wise means computed exclusively from the training set and subsequently applied to the test and external validation sets, to prevent information leakage. The overall proportion of missingness was low, and imputation did not materially alter variable distributions.

To identify key variables, we applied feature importance–based selection using the Boruta algorithm, performed exclusively within the training dataset (Figures S3 and S4 in Multimedia Appendix 1). This approach led to the selection of 17 variables for the NOR prediction model and 19 variables for the OHSS prediction model.

NOR Model Development

In the prediction of NOR, 17 features, including starting dose of FSH to BMI ratio, BMI, log (AMH), and specifically the ovarian stimulation protocol, were selected. The NOR divided by the starting dose of FSH was used and logarithmically transformed as the outcome variable with improved predictive performance. For model training, the internal dataset was divided into an 8:2 split, with 79.9% (5120/6401) of the data randomly allocated to the training set and the remaining 20% (1281/6401) assigned to the internal test set. All data from the external dataset were held out entirely and used exclusively as an external validation cohort, providing an independent assessment of model generalizability across institutions. Eleven ML algorithms, including a linear regression model, were trained to predict the preprocessed NOR outcome. Hyperparameter tuning was performed using 5-fold cross-validation within the training set only, with all hyperparameters predefined and summarized in Table S2 in Multimedia Appendix 1. Model performance was assessed using 4 key metrics: R^2, adjusted R^2, mean absolute error, and root mean square error.

OHSS Model Development

For OHSS prediction, the target variable was the occurrence of early-onset moderate-to-severe OHSS. Given the low event rate and resulting class imbalance, several commonly used imbalances handling strategies (eg, oversampling, undersampling, and ensemble-based resampling) were evaluated. Cost-sensitive learning was ultimately adopted, assigning differentiated penalties to misclassifications while preserving all original clinical data distribution. To prevent overfitting, model complexity was controlled by limiting the number of parameters and applying regularization techniques, as appropriate for each algorithm. Eleven ML algorithms were implemented, with corresponding hyperparameters detailed in Table S3 in Multimedia Appendix 1. As with the NOR model, hyperparameter optimization was conducted exclusively within the training set, and model performance was evaluated on both the internal and the external datasets using the area under the receiver operating characteristic curve (ROC-AUC), the precision-recall area under the curve (PR-AUC), recall, specificity, weighted F_1-score, Cohen κ, and positive and negative predictive values.

Shapley Additive Explanation Value

To further explore the significant features driving the model’s predictions, we used the Shapley additive explanation (SHAP) analysis to assess the importance of core features. SHAP serves as an interpretative tool for ensemble tree models, offering a detailed breakdown of the influence of input features on predictions.

Creation of Dose-Response Curves

In this study, 2 distinct models were developed: a classification model for predicting early-onset moderate-to-severe OHSS and a regression model for forecasting NOR. Models of best performance were incorporated into an integrated, research-oriented computational system, complemented by a web-based calculator. By inputting baseline patient characteristics, the system generates predictions for NOR and early-onset moderate-to-severe OHSS probability, presented as a dose-response curve illustrating changes with increasing FSH starting doses.

XSL•FO
RenderX

Statistical Analysis

The baseline characteristics of patients between the internal dataset and the external dataset were compared using the chi-square test for categorical variables. For continuous variables, we assessed normality using the Shapiro-Wilk W test. Depending on the results, we used either the 2-tailed Student *t* test or the Mann-Whitney *U* test for comparison. R Studio (version 4.3.1; R Foundation for Statistical Computing), Python (version 3.11.4; Python Software Foundation), the open-source *scikit-learn* package (version 3.9.13; open-source community-developed Python ML library), LightGBM (version 4.3.0; Microsoft Corporation), and XGBoost (version 2.0.0; open-source project maintained by XGBoost contributors) were used for model development and statistical analyses.

Results

The Integrated Ovarian Response Prediction System

To address the challenges of individualized ovarian stimulation, we developed an integrated prediction system, "InOvaSGuide," designed to predict both the NOR and the probability of early-onset moderate-to-severe OHSS before ovarian stimulation. The system was built using datasets from 2 IVF clinics and incorporates 2 distinct ML models for NOR and OHSS predictions, respectively (Figure 1A). By analyzing patients' baseline characteristics, the system generates dose-response curves that illustrate the predicted benefit (NOR) and risk (early-onset moderate-to-severe OHSS) across varying FSH starting doses (Figures 1B and C). Additionally, a user-friendly web-based calculator was developed to enhance accessibility and support exploratory use in clinically relevant contexts (Figure S7 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 1. Flowchart of the study: (A) the modeling process with internal and external datasets, (B) illustration of the primary goal of this study, and (C) illustration of the clinical application of the individualized ovarian stimulation guide (InOvaSGuide) system. EHR: electronic health record; FSH: follicle-stimulating hormone; NOR: number of oocytes retrieved; OHSS: ovarian hyperstimulation syndrome.

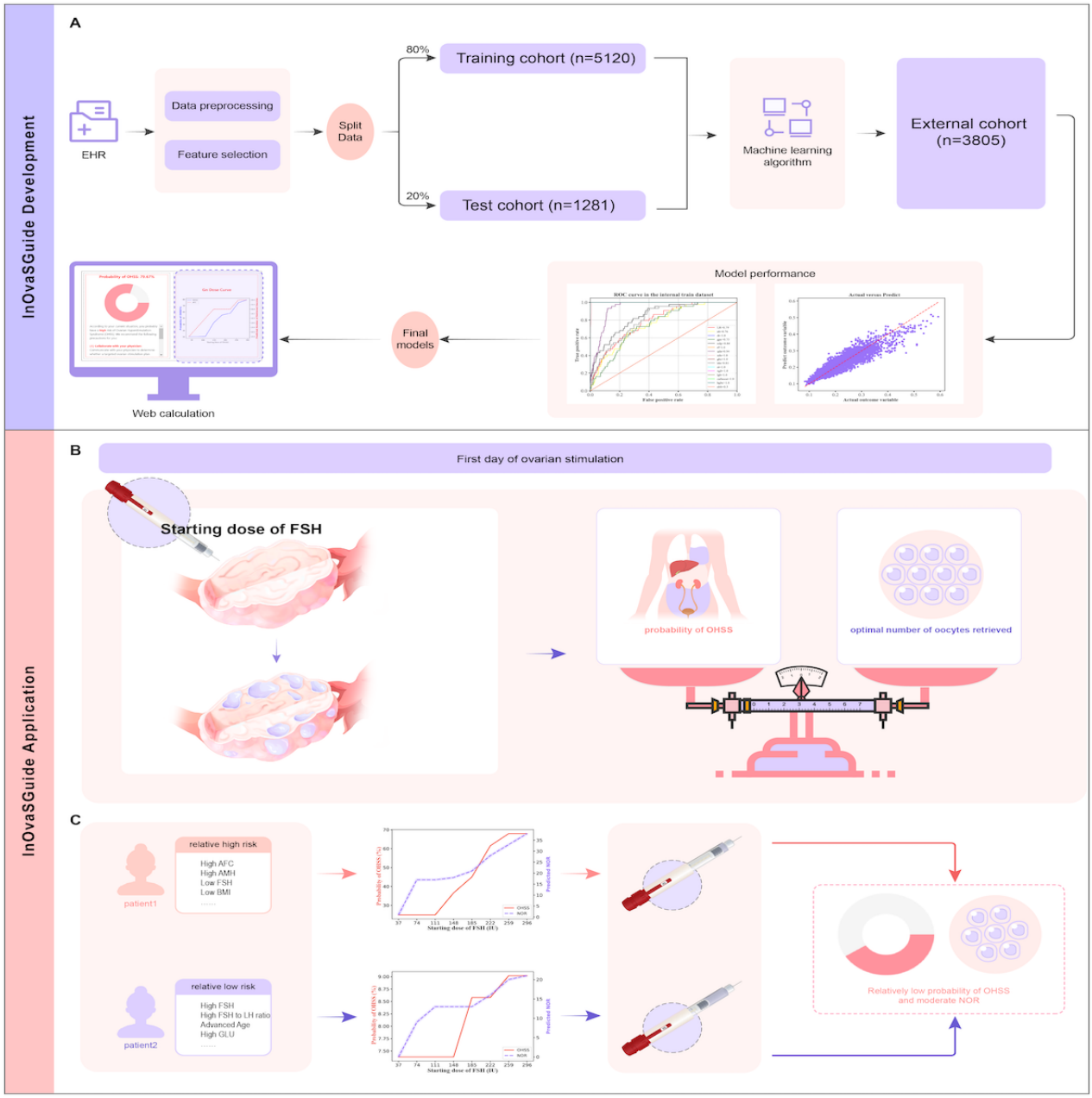

Patient Characteristics

A total of 6401 patients from the internal dataset and 3805 patients from the external dataset were included, with baseline characteristics detailed in Table 1. The median age was 30.0 (IQR 27.0-33.0) years in the internal dataset and 32.0 (IQR 29.0-35.0) years in the external dataset. The gonadotropin-releasing hormone antagonist protocol was the most commonly used in both datasets (internal dataset: 2650/6401, 41.4%; external dataset: 1913/3805, 50.3%). The median number of NOR was 13.0 (IQR 9.0-17.0) and 15.0 (IQR 10.0-20.0) for the internal and external dataset, respectively. In the internal dataset, 55 (0.9%) patients were diagnosed with moderate-to-severe OHSS, whereas 46 (1.2%) patients were diagnosed in the external dataset. Further comparisons between OHSS and non-OHSS cases in both datasets are detailed in Tables S4 and S5 in Multimedia Appendix 1.

XSL•FO
RenderX

Table 1. Baseline characteristics of the patients in the internal and external datasets.

	Internal dataset (n=6401)	External dataset (n=3805)	*P* value
Age (y), median (IQR)	30.0 (27.0-33.0)	32.0 (29.0-35.0)	<.001
BMI (kg/m^2), median (IQR)	21.7 (19.8-24.0)	21.6 (19.8-23.6)	.02
Baseline FSH[a] (mIU/mL), median (IQR)	6.2 (5.2-7.2)	6.5 (5.5-7.5)	<.001
Baseline luteinizing hormone (mIU/mL), median (IQR)	5.3 (3.8-7.1)	5.2 (3.7-6.9)	.006
Anti-Müllerian hormone (ng/mL), median (IQR)	4.1 (2.6-5.4)	3.8 (2.5-5.7)	.20
Fasting blood glucose (mmol/L), median (IQR)	5.3 (5.1-5.4)	4.5 (4.3-4.8)	<.001
Fasting insulin (μU/mL), median (IQR)	11.0 (7.5-12.1)	62.6 (62.6-62.6)	<.001
Homeostasis model assessment of insulin resistance, median (IQR)	2.5 (1.7-2.9)	13.3 (13.3-13.3)	<.001
Baseline antral follicle count, median (IQR)	20.0 (14.0-24.0)	13.0 (10.0-19.0)	<.001
Ovarian stimulation protocol, n (%)			<.001
GnRH[b] agonist long protocol	1211 (18.9)	0 (0)	
GnRH antagonist protocol	2650 (41.4)	1913 (50.3)	
Early-follicular phase long-acting GnRH agonist long protocol	2300 (35.9)	1827 (48)	
Ultralong GnRH agonist protocol	240 (3.8)	65 (1.7)	
Starting dose of FSH (IU), median (IQR)	150.0 (150.0-187.5)	225.0 (150.0-300.0)	<.001
Total dose of FSH (IU), median (IQR)	1950.0 (1500.0-2437.5)	2100.0 (1575.0-2750.0)	<.001
Estradiol level on the day of triggering (pg/mL), median (IQR)	3269.3 (2580.0-3269.3)	2750.0 (1804.0-3961.0)	<.001
Oocytes retrieved, median (IQR)	13.0 (9.0-17.0)	15.0 (10.0-20.0)	<.001
Degree of ovarian hyperstimulation syndrome, n (%)			.11
Normal	6346 (99.1)	3759 (98.8)	
Moderate to severe	55 (0.9)	46 (1.2)	

[a]FSH: follicle-stimulating hormone.

[b]GnRH: gonadotropin-releasing hormone.

Model Performance

For NOR prediction, the gradient boosting regressor exhibited the best performance, with an R^2 value of 0.7978 in the internal dataset and 0.7924 in the external dataset, indicating strong explanatory power (Table 2). The model's mean absolute error was 0.0223 and the root mean square error was 0.0298, collectively affirming the high accuracy and minimal bias. The model's predictions aligned closely with the actual outcomes, demonstrating relatively high accuracy. The Quantile-Quantile plot further confirmed that the residuals followed a normal distribution, as they closely aligned with the diagonal line (Figures 2A-2D).

XSL•FO
RenderX

Table 2. Performance metrics of the number of oocytes retrieved prediction models in internal and external datasets.

Machine learning models	Internal dataset				External dataset			
	R^2	Adjusted R^2	Mean absolute error	Root mean square error	R^2	Adjusted R^2	Mean absolute error	Root mean square error
Gradient boosting regressor	0.7978	0.7951	0.0223	0.0298	0.7924	0.7915	0.0243	0.0327
Light gradient boosting machine regressor	0.7908	0.7880	0.0228	0.0303	0.7826	0.7817	0.0243	0.0334
Extreme gradient boosting regressor	0.7889	0.7861	0.0228	0.0305	0.7907	0.7898	0.0236	0.0328
Random forest regressor	0.7849	0.7820	0.0229	0.0308	0.8020	0.8012	0.0229	0.0319
Ridge	0.7463	0.7429	0.0253	0.0334	0.3228	0.3198	0.0470	0.0590
Linear regression	0.7463	0.7428	0.0253	0.0334	0.3235	0.3204	0.0469	0.0590
Decision tree regressor	0.5452	0.5391	0.0330	0.0447	0.6108	0.6090	0.0318	0.0447
Support vector regression	0.5107	0.5041	0.0390	0.0464	0.5321	0.5300	0.0366	0.0490
Lasso	−0.0023	−0.0158	0.0519	0.0664	−0.1773	−0.1826	0.0648	0.0778
Elastic net	−0.0023	−0.0158	0.0519	0.0664	−0.1773	−0.1826	0.0648	0.0778
Multilayer perceptron regressor	−0.4608	−0.4805	0.0523	0.0802	−351.3296	−352.9112	0.6948	1.3453

XSL•FO
RenderX

Figure 2. Model performance in predicting the number of oocytes retrieved (NOR; A–D) and ovarian hyperstimulation syndrome (OHSS; E and F) in the internal and external datasets. ADA: adaptive boost classifier; BNB: Bernoulli Naive Bayes; CatBoost: categorical boosting classifier; DT: decision tree classifier; ET: extra trees classifier; GBC: gradient boosting classifier; GNB: Gaussian Naive Bayes; GPC: Gaussian process classifier; HGBC: histogram-based gradient boosting classifier; LDA: linear discriminant analysis; LGB: light gradient boosting machine classifier; LR: logistic regression; MLP: multilayer perceptron classifier; QDA: quadratic discriminant analysis; QQ: quantile-quantile; RF: random forest; ROC: receiver operating characteristic; XGB: extreme gradient boosting classifier.

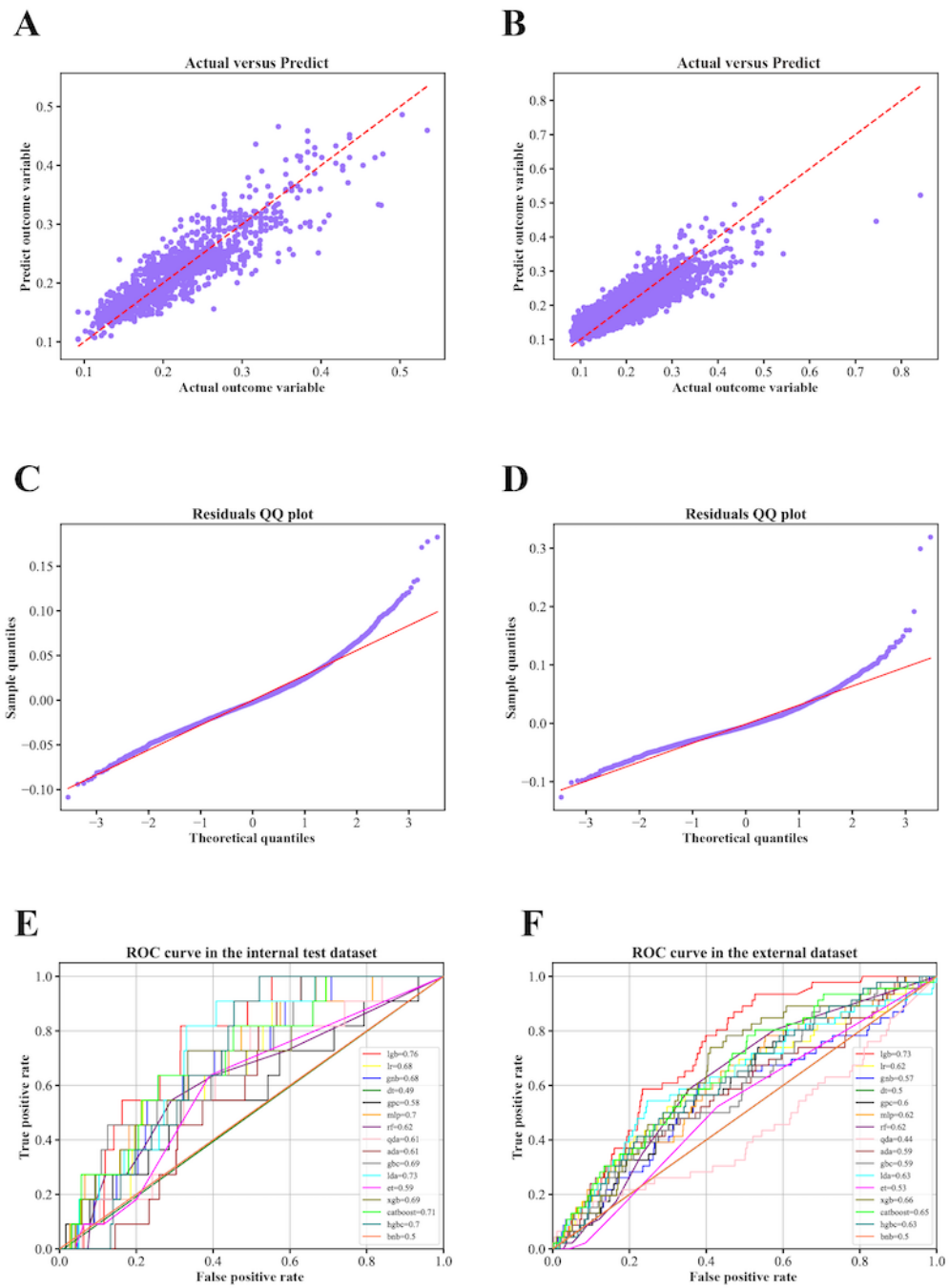

For early-onset moderate-to-severe OHSS prediction, the LightGBM model consistently outperformed other algorithms, achieving an ROC-AUC of 0.7588 in the internal dataset and 0.7287 in the external dataset (Figures 2E and 2F). While recall, specificity, weighted F_1-score, and Cohen κ score indicated reasonable discriminative performance, precision-related metrics, including positive predictive value, negative predictive value, and PR-AUC, remained modest across all classifiers. The results, along with the confusion matrices, are summarized in Table 3 and Figures S5 and S6 in Multimedia Appendix 1.

XSL•FO
RenderX

Table 3. Performance metrics of early-onset moderate-to-severe ovarian hyperstimulation syndrome prediction models in internal and external datasets.

Classifier	Internal dataset								External dataset							
	ROC-AUC[a]	Recall	Specificity	Weighted F_1-score	κ	PPV[b]	NPV[c]	Precision-recall AUC[d]	ROC-AUC	Recall	Specificity	Weighted F_1-score	κ	PPV	NPV	Precision-recall-AUC
LGBMClassifier[e]	0.7588	1.0000	0.9833	0.3466	0.5044	0.3409	1.0000	0.0176	0.7287	0.9348	0.9859	0.4523	0.6099	0.4464	0.9982	0.0227
LinearDiscriminantAnalysis	0.7313	1.0000	0.5690	0.0281	0.0384	0.0197	1.0000	0.0162	0.6574	0.6739	0.9560	0.5966	0.7362	0.5956	0.9934	0.0196
CatBoost[f]	0.7059	1.0000	0.9584	0.1795	0.2918	0.1724	1.0000	0.0173	0.6527	0.8043	0.9852	0.4045	0.5635	0.3996	0.9940	0.0206
MLPClassifier[g]	0.6966	0.9091	0.9784	0.2724	0.4177	0.2669	0.9971	0.0168	0.6322	0.8913	0.9996	0.2071	0.3275	0.1987	0.9893	0.0236
GradientBoostingClassifier	0.6930	1.0000	0.9610	0.1889	0.3053	0.1819	1.0000	0.0161	0.6235	0.2174	0.9553	0.8757	0.9227	0.8837	0.9947	0.0157
XGBClassifier[h]	0.6892	0.7273	0.9941	0.5199	0.6759	0.5181	0.9955	0.0152	0.6221	0.8043	0.9944	0.4263	0.5854	0.4217	0.9933	0.0189
GaussianNB[i]	0.6808	1.0000	0.9755	0.2678	0.4111	0.2614	1.0000	0.0135	0.6186	0.7609	0.9999	0.3548	0.5113	0.3498	0.9883	0.0276
LogisticRegression	0.6791	1.0000	0.4837	0.0250	0.0324	0.0165	1.0000	0.0138	0.6019	1.0000	0.9827	0.0121	0.0003	0.0000	0.9919	0.0182
RandomForest	0.6178	0.7273	0.9909	0.4122	0.5752	0.4094	0.9943	0.0123	0.5945	0.9130	0.9865	0.2084	0.3290	0.1998	0.9944	0.0156
QuadraticDiscriminantAnalysis	0.6142	0.7273	0.9943	0.5277	0.6826	0.5260	0.9955	0.0108	0.5899	1.0000	1.0000	0.0121	0.0003	0.0000	0.9884	0.0267
ExtraTreesClassifier	0.5937	0.6364	0.9965	0.6097	0.7496	0.6094	0.9949	0.0102	0.5280	0.5217	0.9952	0.5708	0.7162	0.5714	0.9899	0.0120

[a]ROC-AUC: area under the receiver operating characteristic curve.

[b]PPV: positive predictive value.

[c]NPV: negative predictive value.

[d]AUC: area under the curve.

[e]LGBMClassifier: light gradient boosting machine classifier.

[f]CatBoost: categorical boosting classifier.

[g]MLPClassifier: multilayer perceptron classifier.

[h]XGBClassifier: extreme gradient boosting classifier.

[i]NB: Naive Bayes.

Model Interpretation

SHAP values were used to assess feature importance for both models, as shown in Figure 3. For NOR prediction, the features with the highest mean absolute SHAP values were FSH starting dose to BMI ratio, BMI, log (AMH), baseline AFC, and baseline FSH, indicating their significant contribution to the model. For early-onset moderate-to-severe OHSS prediction, the most important features identified were baseline AFC, followed by baseline FSH, BMI, fasting blood glucose, and log (AMH).

XSL•FO
RenderX

Figure 3. Shapley additive explanation (SHAP) values of the prediction models for (A) number of oocytes retrieved (NOR) and (B) ovarian hyperstimulation syndrome (OHSS). FSH: follicle-stimulating hormone; BMI, body mass index; AMH: anti-Müllerian hormone; AFC: antral follicle count; LH: luteinizing hormone; HOMA-IR: homeostatic model assessment of insulin resistance; HDL: high-density lipoprotein; E2: estradiol; INS: fasting insulin; TG: triglycerides; TC: total cholesterol; LDL: low-density lipoprotein; GLU: fasting glucose; T: testosterone;.

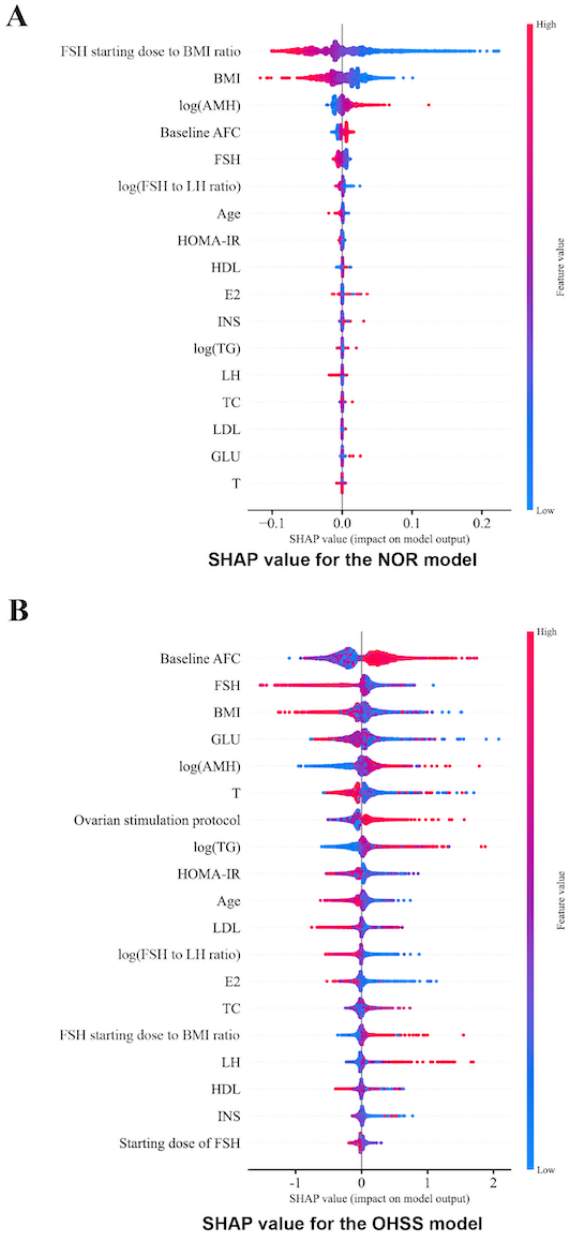

Integrated Dose-Response Curves and Web Calculator

To facilitate individualized ovarian stimulation, we further integrated the prediction models for both NOR and early-onset moderate-to-severe OHSS into dose-response curves. Examples of patients with relatively high and low predicted risks of OHSS are presented in Figures 4A and 4B, respectively. As shown, increasing the starting dose of FSH leads to variable increases in both early-onset moderate-to-severe OHSS probability and predicted NOR for different patients. However, the probability of OHSS occurrence varies among individuals. On the basis of these personalized dose-response predictions, clinicians can determine a suitable starting dose of FSH to achieve an optimal NOR while maintaining a relatively low risk of early-onset moderate-to-severe OHSS.

XSL•FO
RenderX

Figure 4. Model application and representative patient examples with predicted relatively high (A) and low (B) risks of ovarian hyperstimulation syndrome (OHSS). FSH: follicle-stimulating hormone; NOR: number of oocytes retrieved.

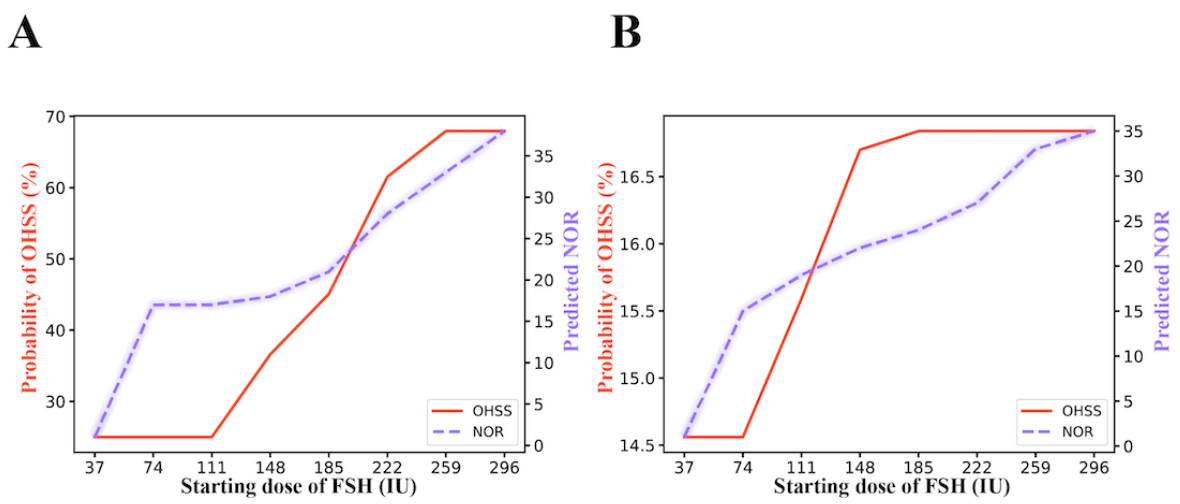

Additionally, we developed a web-based calculator as a research-oriented prototype to enhance accessibility and facilitate exploratory use of the proposed models. The intuitive interface allows users to input relevant data and receive immediate predictions from the models (Figure S7 in Multimedia Appendix 1).

Discussion

Principal Findings

In this study, we developed an integrated ML-based system, "InOvaSGuide," capable of simultaneously predicting NOR and the associated risk of early-onset moderate-to-severe OHSS across a wide range of FSH starting doses. By generating individualized dose-response curves, the model provides a continuous view of expected oocyte yield and corresponding safety profiles, providing clinicians with a structured visual reference to support individualized FSH dose selection. A web-based calculator was further implemented as a research-oriented prototype to improve accessibility and facilitate exploratory use of the models in clinically relevant scenarios.

Oocyte yield remains a key determinant of both efficacy and safety in assisted reproduction. Retrieval of fewer than 4 oocytes has been associated with poor reproductive prognosis [17], whereas obtaining more than 15 oocytes increase the likelihood of OHSS and may slightly compromise live birth outcomes [18,19]. Accordingly, a target range of 5 to 15 oocytes is generally recommended to balance benefit and risk [19]. The dose-response curve framework aligns conceptually with these clinical principles, as it illustrates how predicted NOR changes with increasing FSH doses, thereby supporting informed dosing discussions rather than prescriptive decision-making.

Existing ML-based NOR models generally focus either on approximating actual or optimal oocyte yield [11-13,20] or on producing individualized curves based on a limited number of clinical features [10]. In contrast, our approach used feature importance scores from automated classifiers for selection and compared 11 regression algorithms across 2 independent datasets. This allowed the construction of robust dose-response curves that illustrate how predicted NOR varies with incremental FSH doses, providing additional insight beyond traditional single-point estimates by visualizing predictions across a continuum of FSH doses.

We further developed ML models to predict OHSS, addressing a gap in existing clinical tools that predominantly rely on logistic regression [21,22] or receiver operating characteristic–based analyses [23,24]. Although contemporary strategies, including gonadotropin-releasing hormone antagonist protocols, dual triggering, and "freeze-all" approaches, have substantially reduced the incidence of early-onset OHSS, it remains a persistent concern even among presumed normal responders and has not been fully eliminated from clinical practice [25-27]. Our models demonstrated acceptable and consistent discriminatory ability across both internal and external cohorts, despite the limited number of OHSS events.

Importantly, the low prevalence of early-onset moderate-to-severe OHSS introduces substantial and unavoidable class imbalance, which has direct implications for model performance metrics. In particular, precision is structurally constrained in low-prevalence settings; therefore, PR-AUC values should be interpreted with caution. Although ROC-AUC indicated reasonable discrimination, PR-AUC is highly sensitive to outcome prevalence. When event rates fall below 1%, even well-calibrated models will inherently yield modest precision. In addition, our modeling strategy deliberately prioritized sensitivity to enhance clinical safety, an approach that increases false-positive predictions and further reduces precision and PR-AUC but minimizes the risk of missing true high-risk cases.

Within this context, the OHSS model should be viewed primarily as a screening and risk-stratification aid rather than a

XSL•FO
RenderX

diagnostic or decision-making tool. Its intended role is to flag patients with potentially elevated risk who may warrant closer monitoring or consideration of preventive strategies, rather than to definitively predict OHSS occurrence or guide autonomous clinical actions.

Feature importance analyses largely reflected established biological associations. For NOR, the FSH starting dose to BMI ratio, BMI, and log (AMH) emerged as the most influential predictors. While most features were consistent with clinical practice, the contribution of metabolic markers, such as glucose, lipids, and metabolic indicators, warrants further investigation. For early-onset moderate-to-severe OHSS, AFC, baseline FSH, BMI, and log (AMH) emerged as dominant predictors, aligning with known determinants of ovarian reserve and ovarian sensitivity [23,28]. Associations involving testosterone or glucose were less pronounced, highlighting the multifactorial nature of OHSS risk, particularly in women with polycystic ovary syndrome [29]. Overall, these findings illustrate the capability of ML approaches to integrate diverse clinical variables and improve predictive performance.

A key practical advantage of this study is the integration of NOR and early-onset moderate-to-severe OHSS predictions into a unified, visually intuitive research-oriented system. InOvaSGuide enables clinicians to assess the potential trade-offs between stimulation efficacy and safety across a continuum of FSH doses. The web-based interface supports exploratory analysis and clinician-patient discussion; however, the system does not generate prescriptive dosing recommendations and is not intended for autonomous clinical use. Importantly, prospective validation is essential before any consideration of clinical deployment.

Beyond model performance, the development and potential deployment of AI-based, clinician-in-the-loop decision support tools in reproductive medicine entail careful ethical, legal, and implementation considerations [30,31]. Given the sensitivity of reproductive health data, rigorous safeguards for privacy protection and informed consent are essential [32]. Algorithmic transparency is equally important to support clinician interpretation and reduce risks of automation bias [33], while potential bias across patient subgroups remains an important consideration for future validation. In addition, AI-driven clinical tools may fall under medical software regulation, requiring evidence of safety and clinical validity before clinical implementation. Finally, effective integration into clinical practice will depend on usability, compatibility with established workflows, and clearly assigned clinical accountability [30]. Addressing these factors will be necessary before the system can be responsibly adopted in real-world settings.

Limitations

Our study had several limitations. First, it focused on early-onset moderate-to-severe OHSS, excluding mild cases that may self-resolve and late-onset OHSS more commonly associated with embryo transfer. Patients with a predicted poor prognosis were also excluded under the assumption that they are less likely to develop OHSS. These exclusions introduced a structural selection bias that narrowed the population represented and limited the generalizability of the model in broader clinical settings. Second, the relatively small number of OHSS cases limited the model's ability to fully characterize patients who are affected. This scarcity, together with the substantial class imbalance, also constrained the effectiveness of resampling-based strategies. Although multiple resampling methods were evaluated, only cost-sensitive learning may allow a more reliable assessment of alternative methods. Third, the retrospective nature of the study restricted the availability of certain relevant factors, such as previous OHSS history and genetic susceptibility, and may also introduce selection bias and unmeasured confounding that cannot be fully controlled. Finally, although the system provides individualized dose-response curves for clinical reference, it does not generate a prescriptive starting dose. Moreover, the model has not yet undergone prospective evaluation, which limits its current clinical applicability. A prospective validation study is planned as a necessary next step to assess real-world performance. Future large-scale, multicenter validation in broader patient populations will be essential for improving model stability and generalizability.

Conclusions

We developed and externally validated InOvaSGuide, a ML system that simultaneously predicts NOR and early-onset moderate-to-severe OHSS risk across a continuum of FSH doses. By linking efficacy and safety within a single dose-response framework, the tool highlights the broader potential of model-informed dosing to standardize ovarian stimulation and enhance patient safety. Prospective trials are needed to establish real-world utility.

Acknowledgments

The authors are grateful to Yuan Sheng from STI-Zhilian Research Institute for Innovation and Digital Health for her invaluable contributions to the study figures. They also thank the Yiersan Digital Health Care Group for providing technical support for the web-based calculator interface design. Generative AI (ChatGPT-4.1, OpenAI [34]) was used exclusively for language refinement during manuscript preparation, including improving grammar, clarity, and readability of author-written text. No text, data interpretation, scientific content, or references were generated by the AI tool. All substantive content, analyses, and conclusions were produced entirely by the authors. The authors verified all AI-assisted edits and take full responsibility for the final manuscript.

XSL•FO
RenderX

Funding

This study was funded by the National Natural Science Foundation of China (grant 82371682) and the Natural Science Foundation of Hunan Province (grants 2022JJ40779 and 2022JJ70080). The funding source had no role in the study design, data collection, data analysis, data interpretation, or writing of the report.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JF, Y Li, and SL had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. JC played a role in data collection and was a major contributor to manuscript writing. Jianjuan Zhao conducted data analysis, performed modeling, and contributed to drafting the Methods section. JC, Jianjuan Zhao, JF, Y Li, and SL contributed equally to this work. YZ, GY, YY, and QS provided biostatistical analysis support. HQ, HT, Jing Zhao, and BX assisted in internal dataset acquisition, while Y Liu, JL, and ZY helped in external dataset acquisition. QZ and HL provided language support. JF, Y Li, and SL were responsible for the study design and critically revised the manuscript. All authors read and approved the final manuscript

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data supporting model development and validation, including the distribution and transformation of key variables, feature selection for number of oocytes retrieved and ovarian hyperstimulation syndrome prediction, confusion matrices, the web-based calculator interface, and detailed patient baseline characteristics, together with the completed TRIPOD+AI checklist.
[PDF File (Adobe PDF File), 2183 KB - jmir_v28i1e78245_app1.pdf]

References

1. Ngwenya O, Lensen SF, Vail A, Mol BW, Broekmans FJ, Wilkinson J. Individualised gonadotropin dose selection using markers of ovarian reserve for women undergoing in vitro fertilisation plus intracytoplasmic sperm injection (IVF/ICSI). Cochrane Database Syst Rev 2024 Jan 04;1(1):CD012693. [doi: 10.1002/14651858.CD012693.pub3] [Medline: 38174816]
2. Broekmans FJ. Individualization of FSH doses in assisted reproduction: facts and fiction. Front Endocrinol (Lausanne) 2019 Apr 26;10:181 [FREE Full text] [doi: 10.3389/fendo.2019.00181] [Medline: 31080437]
3. Kim HH, Speedy SE. The promised land of individualized ovarian stimulation: are we there yet? Fertil Steril 2021 Apr;115(4):893-894 [FREE Full text] [doi: 10.1016/j.fertnstert.2021.02.023] [Medline: 33832745]
4. Qiao J, Zhang Y, Liang X, Ho T, Huang HY, Kim SH, et al. A randomised controlled trial to clinically validate follitropin delta in its individualised dosing regimen for ovarian stimulation in Asian IVF/ICSI patients. Hum Reprod 2021 Aug 18;36(9):2452-2462 [FREE Full text] [doi: 10.1093/humrep/deab155] [Medline: 34179971]
5. Broekmans FJ, Kwee J, Hendriks DJ, Mol BW, Lambalk CB. A systematic review of tests predicting ovarian reserve and IVF outcome. Hum Reprod Update 2006;12(6):685-718. [doi: 10.1093/humupd/dml034] [Medline: 16891297]
6. Lensen SF, Wilkinson J, Leijdekkers JA, La Marca A, Mol BW, Marjoribanks J, et al. Individualised gonadotropin dose selection using markers of ovarian reserve for women undergoing in vitro fertilisation plus intracytoplasmic sperm injection (IVF/ICSI). Cochrane Database Syst Rev 2018 Feb 01;2(2):CD012693 [FREE Full text] [doi: 10.1002/14651858.CD012693.pub2] [Medline: 29388198]
7. Hicks SA, Andersen JM, Witczak O, Thambawita V, Halvorsen P, Hammer HL, et al. Machine learning-based analysis of sperm videos and participant data for male fertility prediction. Sci Rep 2019 Nov 14;9(1):16770 [FREE Full text] [doi: 10.1038/s41598-019-53217-y] [Medline: 31727961]
8. Zaninovic N, Rosenwaks Z. Artificial intelligence in human in vitro fertilization and embryology. Fertil Steril 2020 Nov;114(5):914-920 [FREE Full text] [doi: 10.1016/j.fertnstert.2020.09.157] [Medline: 33160513]
9. Hanassab S, Abbara A, Yeung AC, Voliotis M, Tsaneva-Atanasova K, Kelsey TW, et al. The prospect of artificial intelligence to personalize assisted reproductive technology. NPJ Digit Med 2024 Mar 01;7(1):55 [FREE Full text] [doi: 10.1038/s41746-024-01006-x] [Medline: 38429464]
10. Fanton M, Nutting V, Rothman A, Maeder-York P, Hariton E, Barash O, et al. An interpretable machine learning model for individualized gonadotrophin starting dose selection during ovarian stimulation. Reprod Biomed Online 2022 Dec;45(6):1152-1159 [FREE Full text] [doi: 10.1016/j.rbmo.2022.07.010] [Medline: 36096871]
11. Ferrand T, Boulant J, He C, Chambost J, Jacques C, Pena CA, et al. Predicting the number of oocytes retrieved from controlled ovarian hyperstimulation with machine learning. Hum Reprod 2023 Oct 03;38(10):1918-1926 [FREE Full text] [doi: 10.1093/humrep/dead163] [Medline: 37581894]

XSL•FO
RenderX

12. Correa N, Cerquides J, Arcos JL, Vassena R. Supporting first FSH dosage for ovarian stimulation with machine learning. Reprod Biomed Online 2022 Nov;45(5):1039-1045. [doi: 10.1016/j.rbmo.2022.06.010] [Medline: 35915001]
13. Xu H, Feng G, Han Y, La Marca A, Li R, Qiao J. POvaStim: an online tool for directing individualized FSH doses in ovarian stimulation. Innovation (Camb) 2023 Mar 13;4(2):100401 [FREE Full text] [doi: 10.1016/j.xinn.2023.100401] [Medline: 36926531]
14. Debray TP, Collins GS, Riley RD, Snell KI, Van Calster B, Reitsma JB, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data: TRIPOD-Cluster checklist. BMJ 2023 Feb 07;380:e071018 [FREE Full text] [doi: 10.1136/bmj-2022-071018] [Medline: 36750242]
15. Practice Committee of the American Society for Reproductive Medicine. Testing and interpreting measures of ovarian reserve: a committee opinion. Fertil Steril 2012 Dec;98(6):1407-1415 [FREE Full text] [doi: 10.1016/j.fertnstert.2012.09.036] [Medline: 23095141]
16. Practice Committee of the American Society for Reproductive Medicine. Prevention and treatment of moderate and severe ovarian hyperstimulation syndrome: a guideline. Fertil Steril 2016 Dec;106(7):1634-1647 [FREE Full text] [doi: 10.1016/j.fertnstert.2016.08.048] [Medline: 27678032]
17. Oudendijk JF, Yarde F, Eijkemans MJ, Broekmans FJ, Broer SL. The poor responder in IVF: is the prognosis always poor?: a systematic review. Hum Reprod Update 2012;18(1):1-11. [doi: 10.1093/humupd/dmr037] [Medline: 21987525]
18. Steward RG, Lan L, Shah AA, Yeh JS, Price TM, Goldfarb JM, et al. Oocyte number as a predictor for ovarian hyperstimulation syndrome and live birth: an analysis of 256,381 in vitro fertilization cycles. Fertil Steril 2014 Apr;101(4):967-973 [FREE Full text] [doi: 10.1016/j.fertnstert.2013.12.026] [Medline: 24462057]
19. Sunkara SK, Rittenberg V, Raine-Fenning N, Bhattacharya S, Zamora J, Coomarasamy A. Association between the number of eggs and live birth in IVF treatment: an analysis of 400 135 treatment cycles. Hum Reprod 2011 Jul;26(7):1768-1774. [doi: 10.1093/humrep/der106] [Medline: 21558332]
20. Zieliński K, Pukszta S, Mickiewicz M, Kotlarz M, Wygocki P, Zieleń M, et al. Personalized prediction of the secondary oocytes number after ovarian stimulation: a machine learning model based on clinical and genetic data. PLoS Comput Biol 2023 Apr 27;19(4):e1011020 [FREE Full text] [doi: 10.1371/journal.pcbi.1011020] [Medline: 37104276]
21. Grynnerup AG, Løssl K, Toftager M, Bogstad JW, Prætorius L, Zedeler A, et al. Predictive performance of peritoneal fluid in the pouch of Douglas measured five days after oocyte pick-up in predicting severe late-onset OHSS: a secondary analysis of a randomized trial. Eur J Obstet Gynecol Reprod Biol 2022 Jul;274:83-87 [FREE Full text] [doi: 10.1016/j.ejogrb.2022.05.004] [Medline: 35609351]
22. Tarlatzi TB, Venetis CA, Devreker F, Englert Y, Delbaere A. What is the best predictor of severe ovarian hyperstimulation syndrome in IVF? A cohort study. J Assist Reprod Genet 2017 Oct 14;34(10):1341-1351 [FREE Full text] [doi: 10.1007/s10815-017-0990-7] [Medline: 28710674]
23. Ocal P, Sahmay S, Cetin M, Irez T, Guralp O, Cepni I. Serum anti-Müllerian hormone and antral follicle count as predictive markers of OHSS in ART cycles. J Assist Reprod Genet 2011 Dec 1;28(12):1197-1203 [FREE Full text] [doi: 10.1007/s10815-011-9627-4] [Medline: 21882017]
24. Cao M, Lin Q, Liu Z, Lin Y, Huang Q, Fu Y, et al. Optimized personalized management approach for moderate/severe OHSS: development and prospective validation of an OHSS risk assessment index. Hum Reprod 2024 Oct 01;39(10):2320-2330. [doi: 10.1093/humrep/deae197] [Medline: 39237109]
25. Emile SH, Horesh N, Garoufalia Z, Gefen R, Ray-Offor E, Wexner SD. Strategies to reduce ileus after colorectal surgery: a qualitative umbrella review of the collective evidence. Surgery 2024 Feb;175(2):280-288. [doi: 10.1016/j.surg.2023.10.005] [Medline: 38042712]
26. Ioannidou PG, Bosdou JK, Lainas GT, Lainas TG, Grimbizis GF, Kolibianakis EM. How frequent is severe ovarian hyperstimulation syndrome after GnRH agonist triggering in high-risk women? A systematic review and meta-analysis. Reprod Biomed Online 2021 Mar;42(3):635-650. [doi: 10.1016/j.rbmo.2020.11.008] [Medline: 33483281]
27. The Eshre Guideline Group On Ovarian Stimulation, Bosch E, Broer S, Griesinger G, Grynberg M, Humaidan P, et al. ESHRE guideline: ovarian stimulation for IVF/ICSI. Hum Reprod Open 2020;2020(2):hoaa009 [FREE Full text] [doi: 10.1093/hropen/hoaa009] [Medline: 32395637]
28. Ashrafi M, Bahmanabadi A, Akhond MR, Arabipoor A. Predictive factors of early moderate/severe ovarian hyperstimulation syndrome in non-polycystic ovarian syndrome patients: a statistical model. Arch Gynecol Obstet 2015 Nov 29;292(5):1145-1152. [doi: 10.1007/s00404-015-3723-0] [Medline: 25920524]
29. Sun B, Ma Y, Li L, Hu L, Wang F, Zhang Y, et al. Factors associated with ovarian hyperstimulation syndrome (OHSS) severity in women with polycystic ovary syndrome undergoing IVF/ICSI. Front Endocrinol (Lausanne) 2020 Jan 19;11:615957 [FREE Full text] [doi: 10.3389/fendo.2020.615957] [Medline: 33542709]
30. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan 7;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
31. Morley J, Machado CC, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: a mapping review. Soc Sci Med 2020 Sep;260:113172. [doi: 10.1016/j.socscimed.2020.113172] [Medline: 32702587]
32. Price WN2, Cohen IG. Privacy in the age of medical big data. Nat Med 2019 Jan 7;25(1):37-43 [FREE Full text] [doi: 10.1038/s41591-018-0272-7] [Medline: 30617331]

XSL•FO
RenderX

33. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med 2018 Mar 15;378(11):981-983 [FREE Full text] [doi: 10.1056/NEJMp1714229] [Medline: 29539284]
34. OpenAI. ChatGPT. URL: https://chatgpt.com/

Abbreviations

AFC: antral follicle count
AI: artificial intelligence
AMH: anti-Müllerian hormone
ROC-AUC: area under the receiver operating characteristic curve
FSH: follicle-stimulating hormone
InOvaSGuide: individualized ovarian stimulation guide
IVF: in vitro fertilization
LH: luteinizing hormone
NOR: number of oocytes retrieved
OHSS: ovarian hyperstimulation syndrome
PR-AUC: precision-recall area under the curve
SHAP: Shapley additive explanation
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

Edited by J Sarvestan; submitted 30.May.2025; peer-reviewed by F Sun, KH Lin; comments to author 29.Sep.2025; revised version received 22.Dec.2025; accepted 23.Dec.2025; published 03.Feb.2026.

Please cite as:
Chen J, Zhao J, Qiu H, Liu Y, Zhang Y, Sun Q, Yi Y, Tang H, Zhao J, Xu B, Zhang Q, Yang G, Li H, Liu J, Yang Z, Liang S, Li Y, Fu J
Integrated Prediction System for Individualized Ovarian Stimulation and Ovarian Hyperstimulation Syndrome Prevention: Algorithm Development and Validation
J Med Internet Res 2026;28:e78245
URL: https://www.jmir.org/2026/1/e78245
doi:10.2196/78245
PMID:

©Jingjing Chen, Jianjuan Zhao, Huiyu Qiu, Yanhui Liu, Yunqi Zhang, Qicheng Sun, Yan Yi, Hongying Tang, Jing Zhao, Bin Xu, Qiong Zhang, Ge Yang, Hui Li, Junjie Liu, Zhongzhou Yang, Shaolin Liang, Yanping Li, Jing Fu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patterns of AI Use in Clinical Work by Hospitalists: Survey Study

Prabhava Bagla[1,2], MD; Jasmah Hanna[1,2], MS; Bhargav Marthambadi[1,2], MD; Stacey Watkins[1,2], MD, PhD

[1]Department of Medicine, Division of Hospital Medicine, School of Medicine, Emory University, Atlanta, GA, United States
[2]Grady Memorial Hospital, Atlanta, GA, United States

Corresponding Author:
Prabhava Bagla, MD
Department of Medicine, Division of Hospital Medicine
School of Medicine
Emory University
80 Jesse Hill Jr Dr SE
Atlanta, GA, 30303
United States
Phone: 1 (404) 251 8899
Fax: 1 (404) 727 2620
Email: prabhava.bagla@emory.edu

Related Article:

Comment in: https://www.jmir.org/2026/1/e97419/

Abstract

Background: Artificial intelligence (AI) tools are widely and freely available for clinical use. Understanding hospitalists' real-world adoption patterns in the absence of organizational endorsement is essential for health care institutions to develop governance frameworks and optimize AI integration.

Objective: The objective of this study was to investigate hospitalists' use of AI, examining the AI platforms being used, frequency of use, and clinical contexts of application. We hypothesized that AI use is more common among younger, less experienced hospitalists, albeit at an overall low frequency.

Methods: An anonymous online survey was distributed via email to all 70 hospitalists (physicians, nurse practitioners, and physician assistants) providing direct patient care at a large urban academic tertiary care hospital. Demographic data, the AI platform used (if any), the purpose for AI use, and the frequency of use information were collected. The CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist was used for creating, testing, administering, and reporting the results of the survey. Chi-square test was used where possible; when expected cell values were low, the Fisher's exact test was used instead. The Friedman test and the pairwise Wilcoxon signed-rank test were used for analyzing the differences in the frequency of AI use for various tasks. Likert-scale responses to frequency questions (never, rarely, sometimes, often, and always) were converted to ordinal values (1-5, respectively) to facilitate analysis.

Results: Of the 70 providers, 54 (77.1%) responded to the survey. No significant differences in AI usage were observed across shift type, years of practice, time allocation to hospitalist duties, sex, age, or provider designation, contrary to our hypothesis. Overall, 36 of 54 respondents (66.7%; 95% CI 53.4%-77.8%) reported using AI in clinical practice. OpenEvidence was the most used platform (28/54, 51.9%), far exceeding general-purpose tools like OpenAI's ChatGPT (4/54, 7.4%), suggesting a preference for medical-specific platforms. Among nonusers, primary concerns were AI accuracy and preference for established resources. The most common application was answering miscellaneous clinical questions (32/36, 88.9%), generating differential diagnoses (31/36, 86.1%), and determining management options (31/36, 86.1%), with much lower use for patient education materials (16/36, 44.4%). There was a statistically significant difference in the frequency of AI use across these clinical scenarios (Friedman test χ^2_4 37.6; $P<.001$). Pairwise comparisons using the Wilcoxon signed-rank test revealed significant differences between use for answering miscellaneous questions and confirming suspected diagnosis ($P=.003$) and generating patient education materials ($P=.004$), respectively. Most respondents reported using AI for under 25% of clinical encounters across all use cases.

Conclusions: Two-thirds of hospitalists organically adopted AI despite the absence of institutional oversight. AI is predominantly used as a supplementary decision support tool, with a preference for a medical-specific platform. Health care institutions must

XSL•FO
RenderX

develop governance frameworks, validation protocols, and educational initiatives to ensure safe and effective AI deployment in clinical practice.

(J Med Internet Res 2026;28:e85973) doi:10.2196/85973

KEYWORDS

artificial intelligence; AI; generative artificial intelligence; generative AI; hospitalist; physician; patterns of use; clinical decision support systems; healthcare technology adoption

Introduction

The development of artificial intelligence (AI) models has led to a significant interest in exploring their capabilities in many aspects of clinical medicine [1,2]. AI systems have been tested against physicians in assessing acuity [3], developing diagnostic formulation [4], and generating management options [5]. Particularly in image or pattern recognition-based diagnosis, AI demonstrates comparable or superior performance to clinicians [6,7].

While numerous studies have evaluated AI performance in carefully controlled research settings, large language model (LLM)–based AI tools are also widely and freely available without restriction to specific, well-defined tasks. An LLM is pretrained on a large amount of data, and can leverage that knowledge to answer user queries in natural language format. LLM capabilities range from presenting relevant information sourced from data it has access to, generating new content, such as computer code, or editing text [8]. ChatGPT is a general-purpose AI system offered by OpenAI that has a range of free and premium tier options, with progressively more advanced reasoning and content generation capabilities available [9]. A survey of US adults by the Pew Research Center indicated that between March 2023 and March 2025, awareness about ChatGPT has increased from 58% to 79%, and 28% of employed adults reported using it for work, compared with 8% in 2023. Usage rates are higher among younger adults and those with higher education levels, such as a postgraduate degree [10]. Some AI systems, such as Gemini by Google and Copilot by Microsoft, are presented to consumers by means of integration into their respective technical offerings, including prominently alongside the basic web search interface [11,12]; with free and paid user tiers and capabilities available similar to ChatGPT [13,14].

Some AI platforms use a retrieval-augmented generation (RAG) system for ease of updating and improved reliability of their LLM's output. An RAG system provides a specific, reliable database for an LLM to search, in addition to what it has already been trained on. The LLM then combines its pre-existing knowledge and the information from the database to synthesize responses to user queries. The RAG strategy simplifies the process of keeping LLM answers updated and accurate in the context of evolving knowledge by simply expanding the authoritative database the LLM accesses, instead of having to retrain the LLM to search and appropriately parse all possible data sources on its own [15]. ClinicalKey AI is one such AI system developed by Elsevier requiring access using a paid subscription model that delivers evidence-based answers sourced from their library of books and journals [16].

OpenEvidence is a unique AI tool marketed specifically toward medical professionals only in the United States. It provides free unlimited use for verified users, while allowing some use with query limits to users without creating an account. It also uses the RAG system by means of data use agreements with several reputable sources, such as the *New England Journal of Medicine*, *Journal of the American Medical Association Network* journals, National Comprehensive Cancer Network, American Diabetes Association, and American College of Cardiology, and thus has the potential to offer evidence-based answers grounded in peer-reviewed medical literature. It also offers the ability to write patient handouts for education about management options for diseases, posthospitalization care, and insurance prior authorization letters [17].

AI tools have been implemented for use for a wide variety of purposes in health care, such as (1) clinical use by health care providers for supporting patient care decisions; (2) directly by laypersons for self-evaluation of health concerns; (3) facilitating business operations, such as bed management; and (4) as hybrid tools supporting both patient care and optimizing health care delivery, such as AI scribes for clinical documentation [18]. A study of 67 nonprofit health systems revealed adoption of AI for active or planned use mostly in imaging and radiology, followed by electronic health record integration for purposes like sepsis detection, and generating documentation via ambient listening [19].

An individual health care provider's use of AI might be influenced by how others perceive them. A study on public perception of physicians who use AI, conducted by surveying US adults shown hypothetical static advertisements for family physicians differing in whether AI use for administrative, diagnostic, or therapeutic purposes was mentioned, showed that physicians using AI for any purpose were considered less competent, trustworthy, and empathic, as well as less sought after for care [20]. Similarly, practicing clinicians viewed hypothetical peers who used AI as a primary decision-making tool to be significantly less competent than those who did not use AI at all, an effect only partially mitigated if the use of AI was specified as being used for verification purposes [21].

Despite these perception issues, physicians' sentiment toward AI use in clinical practice is overall positive, with greater familiarity, use in daily or professional life, and being involved in AI research as factors associated with greater enthusiasm and lower skepticism; the primary concerns being liability in the event of errors, and lack of transparency in AI decision-making [22]. The American Medical Association (AMA) conducted a survey of 1183 physicians, indicating that AI use increased from 38% in 2023 to 66% in 2024, mostly for assistance with documentation, ranging from discharge summaries to responses

XSL•FO
RenderX

to patient portal messages. Respondents also indicated that measures to reduce administrative burden held the greatest promise for future AI implementation, and an increasing recognition of the potential of AI to benefit patient care (68% in 2024 compared with 63% in 2023) [23].

While adoption of AI at an organizational level leads to availability and use for preapproved tasks by health care providers working in the respective organizations, individual providers are unable to use AI for any tasks that require deep integration with their organization's electronic health record system at a system level. Hospitalists represent a unique population for studying AI adoption given their role as frontline providers, managing high volumes of diverse, acutely ill patients across multiple clinical domains. Their practice patterns, workflow demands, and decision-making processes may influence how AI tools are integrated into daily clinical care. However, there is a dearth of literature about AI real-world usage patterns in daily clinical practice among hospitalists in the absence of any organizational endorsement.

The purpose of this study was to investigate hospitalists' use of AI through a comprehensive survey examining the types of AI tools currently being used, frequency of use, and clinical contexts of application in real-world hospital practice settings without any system-wide AI integration. We hypothesized that AI use in clinical work is more common among younger, less experienced hospitalists, albeit at an overall low frequency.

Methods

Study Design

A comprehensive anonymous online survey (Multimedia Appendix 1) was distributed via email during August 2025 to all 70 physicians, nurse practitioners, and physician assistants (the latter 2 categories collectively referred to as advanced practice providers [APPs]) who provide direct patient care at our large, tertiary care academic center. The sample was not considered a convenience sample as all providers were given the opportunity to respond. The survey asked demographic questions and questions about which AI systems, if any, the respondents used and for what purpose. For analytic purposes, AI platforms were categorized as either general-purpose or medical-specific based on their primary market positioning and intended use at the time of the survey. Medical-specific platforms were defined as tools marketed explicitly for clinical decision support and designed primarily for use by health care professionals, with content grounded in medical literature or curated clinical sources (OpenEvidence and ClinicalKey AI). General-purpose platforms were defined as broadly available AI tools intended for use across multiple domains and audiences, not marketed primarily for clinical decision-making (ChatGPT, Microsoft Copilot, and Google Gemini), even if health care–specific enterprise or specialized versions may exist.

If respondents indicated that they did not use any AI systems, they were also asked what concerns they had. CHERRIES (Checklist for Reporting Results of Internet E-Surveys) guidelines (Multimedia Appendix 2) on survey studies were followed [24].

Ethical Considerations

This study was considered not to comprise human participant research by Emory University Institutional Review Board in accordance with the Declaration of Helsinki, and so was exempt from full committee review. The first page of the survey contained information regarding the purpose of the study, potential risks, and benefits to respondents, and contact information regarding the investigators. If respondents did not wish to answer the survey, they could either decline using the option provided or simply close the survey. The page also disclosed that response data would be kept confidential, only analyzed as an aggregate, and the information stored securely without any specific time limits using an encrypted institutional subscription to Microsoft OneDrive and a related set of tools. No personally identifiable information was collected as part of the survey. No compensation or incentive of any kind, monetary or otherwise, was offered to participants.

Survey Development

Survey questions were developed by discussion and consensus among study authors. For AI users, the survey asked, "How often, on average, do you use AI per day for [specific clinical task]?" for each of five tasks, (1) generating differential diagnosis lists, (2) confirming suspected diagnosis, (3) answering miscellaneous questions (with examples provided, such as rare drug side effects), (4) determining management or testing options, and (5) generating patient education materials. Response options were never, rarely, sometimes, often, and always; defined as (1) rarely (0%-25% of the time), (2) sometimes (26%-50% of the time), (3) often (51%-75% of the time), and (4) always (76% or more of the time). The survey did not explicitly specify whether respondents should consider all patient encounters or only those encounters where the specific task would be applicable, leaving interpretation to individual respondents. Functionality was tested before dissemination to potential respondents.

Recruitment Process

The survey was distributed via email during August 2025 to all 70 potential respondents. They were subsequently reminded over a period of 3 weeks via group emails and in-person interactions.

Survey Administration

The email sent to potential respondents contained a brief description of the study and a link to the survey itself. The survey was created in Microsoft Forms, which allowed responses to be automatically saved as a Microsoft Excel spreadsheet. Adaptive questioning was used to streamline survey responses and navigation; for instance, if respondents indicated they did not use AI for clinical work at all, they were not shown questions about possible use scenarios. The maximum number of question items (including initial informed consent) was 17, with an average of 2 questions per page for readability. All questions were marked mandatory, so incomplete responses were not allowed. All questions, except for reasons for not using AI at all, were single-response only, and nonresponse options were provided for some of the demographic questions. Respondents could adjust their responses before submitting, at which time

XSL•FO
RenderX

further editing of responses was not allowed. No incentives were offered in exchange for responses. Overall, responses were collected during August-September 2025.

Response Rates

Unique responses were ensured by requiring respondents to log in (automatically fulfilled when accessing the survey by clicking the link in their email) and only allowing 1 response per person. We are unable to determine if there were potential respondents who viewed the survey but did not answer any questions or stopped partway, as only completed surveys were accepted.

Statistical Analysis

Statistical analysis was conducted using SAS (version 9.4, SAS Institute Inc) for analysis of demographic variables. Chi-square test was used where possible; when more than 20% of expected cell values were under 5, the Fisher's exact test was used instead. R software (version 4.5.1; R Foundation) was used for the Friedman test and the pairwise Wilcoxon signed-rank test for analyzing the differences between the frequency of AI use for various tasks. These analyses were restricted to responses to survey questions asking about the 5 standardized clinical tasks (generating differential diagnosis lists, confirming suspected diagnosis, answering miscellaneous questions, determining management or testing options, and generating patient education materials; Multimedia Appendix 1). Likert-scale responses to survey questions (never, rarely, sometimes, often, and always) were converted to ordinal values (1-5, respectively) to facilitate analysis. Only completed surveys were analyzed. No specific analysis regarding the time taken to complete the survey was conducted, and no weighting or propensity scores were used.

Results

Out of 70 providers, 54 responded to the survey (response rate of 77.1%). Table 1 presents survey respondent characteristics and AI usage patterns. The majority of respondents were dayshift team members (48/54, 88.9%), and two-thirds (36/54, 66.7%) were within 6 years of completing training. Most respondents devoted all their time to clinical work (31/54, 57.4%). Physicians comprised 83.3% (45/54) of respondents, with APPs accounting for the remaining 16.7% (9/54). No statistically significant differences in AI usage were observed across shift type, years of practice, time allocation to hospitalist duties, sex, age, or provider designation, contrary to our hypothesis that younger, less experienced hospitalists would use AI more frequently. Ethnicity was not analyzed for statistical significance due to insufficient sample sizes across groups. Overall, 36 of 54 (66.7%; 95% CI 53.4%-77.8%) respondents reported using AI tools in clinical practice.

XSL•FO
RenderX

Table 1. Characteristics of survey respondents and association with artificial intelligence use in clinical work (n=54). Survey respondents from a large academic medical center, August-September 2025.

Variable	Number (n=54), n (%)	Use of AI[a], n (%)	*P* value[b]
Role			>.99[b]
Dayshift rounding and admissions	48 (88.9)	32 (66.6)	
Night shift admissions and cross-cover	6 (11.1)	4 (66.6)	
Years of practice (y)			.54
0-6	36 (66.7)	25 (69.4)	
7 or more	18 (33.3)	11 (61.1)	
Amount of time spent on clinical hospitalist duties			.12
100%	31 (57.4)	18 (50.6)	
99% or less	23 (42.6)	18 (78.3)	
Sex			.44
Female	26 (48.1)	16 (61.5)	
Male	28 (51.9)	20 (71.4)	
Age (y)			.11
35 or less	20 (37)	16 (80)	
36 or more	34 (63)	20 (58.8)	
Ethnicity[c]			—[d]
Asian	15 (27.8)	10 (66.6)	
Black or African American	12 (22.2)	8 (66.6)	
Hispanic or Latino	3 (5.6)	3 (100)	
White	20 (37)	12 (60)	
Prefer not to say	4 (7.4)	3 (75)	
Designation			>.99[b]
Physician	45 (83.3)	30 (66.6)	
APP[e]	9 (16.7)	6 (66.6)	
Use of AI in clinical work			—
Yes	36 (66.7)	—	
No	18 (33.3)	—	

[a]AI: artificial intelligence.

[b]Fisher's exact test was used when specified, otherwise chi-square test was used.

[c]Statistical analysis not performed due to low sample sizes.

[d]Not applicable.

[e]APP: advanced practice provider.

OpenEvidence was the most commonly used platform (28/54, 51.9%), followed by ChatGPT (4/54, 7.4%; Table 2; Multimedia Appendix 3). Among the 18 respondents (33.3%) who did not use AI, the barriers to adoption were (1) preference for alternative resources such as PubMed (14/18, 77.8%); (2) concerns regarding accuracy (7/18, 38.9%); (3) not available system-wide, or endorsed by hospital, or via the electronic medical record (7/18, 38.9%; respondents who answered not available system-wide or endorsed by hospital also selected not available via electronic medical record); (4) time constraints (4/18, 22.2%); (5) privacy concerns (4/18, 22.2%); (6) prefer to consult another provider (2/18, 11.1%); and (7) lack of continuing medical education credits (1/18, 5.6%); no respondents selected cost or weekly query limits as reasons (percentages do not add up to 100% as multiple responses were allowed for this question). The question also allowed for free-text input of other reasons that respondents did not use AI, but no additional responses were received.

XSL•FO
RenderX

Table 2. Artificial intelligence platforms used by hospitalists in clinical work (n=54). Survey respondents from a large academic medical center, August-September 2025.

Use of AI[a] in clinical work	Number (n=54), n (%)
Microsoft Copilot	1 (1.9)
Google Gemini	1 (1.9)
ClinicalKey AI	2 (3.7)
ChatGPT	4 (7.4)
OpenEvidence	28 (51.9)
No use of any AI	18 (33.3)

[a]AI: artificial intelligence.

Table 3 and Multimedia Appendix 4 summarize the clinical tasks for which AI was used by users (n=36). The most common application was answering miscellaneous clinical questions (32/36, 88.9%; 95% CI 74.7%-95.6%), with respondents providing examples, including "rare drug side effects" and "causes of false-positive urine drug screen." Generating differential diagnoses and determining management or treatment options were equally prevalent (31/36, 86.1%; 95% CI 71.3%-93.9%). Approximately 72.2% (26/36; 95% CI 56%-84.2%) of responders used AI to support confirmation of suspected diagnoses, while few (16/36, 44.4%; 95% CI 29.5%-60.4%) used AI for generating patient education or counseling materials. Using the Friedman test revealed there was a statistically significant difference in the frequency of AI use across these clinical scenarios (χ^2_4=37.6; P<.001).

Table 3. Clinical tasks for which artificial intelligence is used among hospitalist AI users (n=36). Survey respondents from a large academic medical center, August-September 2025.

Tasks AI[a] is used for	Number (n=36), n (%)	95% CI
Generating differential diagnosis list	31 (86.1)	71.3%-93.9%
Confirming suspected diagnosis	26 (72.2)	56%-84.2%
Answering miscellaneous questions	32 (88.9)	74.7%-95.6%
Determining testing or treatment options	31 (86.1)	71.3%-93.9%
Generating patient education or counseling materials	16 (44.4)	29.5%-60.4%

[a]AI: artificial intelligence.

Pairwise Wilcoxon signed-rank test with continuity and Bonferroni corrections demonstrated that AI use for answering miscellaneous questions was significantly different from use for confirming suspected diagnosis (corrected P=.003) and generating patient education or counseling materials (corrected P=.004).

Table 4 details the frequency of AI usage for each clinical application. Respondents were asked how often, on average, they used AI per day for each specific task. Most respondents reported using AI rarely (0%-25% of the time) across all use cases, in line with our hypothesis. Additionally, 3 respondents indicated other AI applications—assistance with clinical documentation, comprehensive literature searches, and creation of educational cases for medical students (1 respondent for each).

XSL•FO
RenderX

Table 4. Frequency of artificial intelligence use for specific clinical tasks among hospitalist artificial intelligence users (n=36). Survey respondents from a large academic medical center, August-September 2025.

Tasks[a] AI[b] is used for	Number (n=36), n (%)
Generating differential diagnosis list	
Never	5 (13.9)
Rarely	20 (55.6)
Sometimes	8 (22.2)
Often	2 (5.6)
Always	1 (2.8)
Confirming suspected diagnosis	
Never	10 (27.8)
Rarely	17 (47.2)
Sometimes	7 (19.4)
Often	1 (2.8)
Always	1 (2.8)
Answering miscellaneous questions	
Never	4 (11.1)
Rarely	12 (33.3)
Sometimes	12 (33.3)
Often	6 (16.7)
Always	2 (5.6)
Determining testing or treatment options	
Never	5 (13.9)
Rarely	20 (55.6)
Sometimes	4 (11.1)
Often	5 (13.9)
Always	2 (5.6)
Generating patient education or counseling materials	
Never	20 (55.6)
Rarely	9 (25)
Sometimes	3 (8.3)
Often	3 (8.3)
Always	1 (2.8)
Other purpose, not specified above	
No	33 (91.7)
Yes	3 (8.3)

[a]Respondents were asked, “How often, on average, do you use AI per day for [specific task]?” Rarely: 0%-25% of the time; Sometimes: 26%-50% of the time; Often: 51%-75% of the time; Always: 76% or more of the time.

[b]AI: artificial intelligence.

Discussion

Principal Findings

Our findings reveal that two-thirds of hospitalists at our institution use AI tools during clinical work, predominantly for answering miscellaneous clinical questions (32/36, 88.9%), generating differential diagnoses (31/36, 86.1%), and determining management options (31/36, 86.1%). Notably, this usage occurs despite the absence of system-level integration, institutional endorsement, or selective blocking of any AI platform interface, suggesting organic adoption driven by perceived clinical utility. The predominance of OpenEvidence among AI users (28/36, 77.8% users; 28/54, 51.9% of all

XSL•FO
RenderX

respondents) is particularly noteworthy. The high degree of convergence on this single niche platform, rather than more widely known general-purpose tools, raises questions about adoption pathways that warrant further investigation. The stark preference for specialized, medical-specific platforms (as defined by primary market positioning and intended clinical use) over general-purpose AI represents one of our most significant findings. Among AI users, OpenEvidence was chosen 7 times more frequently than ChatGPT (28 vs 4 respondents), despite ChatGPT's substantially greater name recognition and broader public adoption [10]. This pattern suggests that hospitalists are not simply adopting whatever AI tools are most readily available or widely marketed, but are instead making deliberate choices based on tool-specific characteristics relevant to medical practice.

Differential availability of various AI platforms could have been a confounding variable affecting observed preference for one platform over others. For instance, if hospital firewall systems blocked access to general-purpose AI while allowing medical use–focused platforms, that could affect adoption rates. However, this was not a factor in our study as our institution's computer systems do not block any particular AI platform.

Several factors may explain this preference for medical-specific platforms. First, hospitalists may perceive that specialized tools trained on or linked to medical literature will provide more accurate and relevant clinical information. Second, the use of general-purpose AI for patient care may raise concerns about data privacy, lack of possibility of HIPAA (Health Insurance Portability and Accountability Act) compliance, and liability that are less salient with tools explicitly designed for health care applications. Third, medical-specific platforms may provide source citations or references to primary literature, offering a verification pathway that general-purpose tools typically lack. Finally, hospitalists may view general-purpose AI as insufficiently tailored to the complexity and nuance of clinical decision-making, preferring tools that understand medical context and terminology.

This finding has important implications for AI development and deployment in health care. It suggests that clinicians value specialized tools over general-purpose solutions. Published literature validating OpenEvidence's capabilities as a clinical decision support tool for use by health care providers remains limited to generating management options for hypothetical primary care scenarios [25]; yet, hospitalists have adopted it widely based on its medical focus alone. This creates both an opportunity and a responsibility for developers of medical-specific AI tools to ensure their platforms meet the quality and safety standards that clinicians appear to assume they possess. The preference for medical-specific platforms also highlights a potential vulnerability that clinicians may place unwarranted trust in tools marketed as "medical AI" without rigorous independent validation of their accuracy and reliability compared with general-purpose alternatives.

The relatively low frequency of AI use, with most respondents reporting use in 0%-25% of their daily work, suggests that AI currently serves as a supplementary rather than primary clinical resource. Importantly, the survey did not specify whether frequency estimates should be calculated using all daily patient encounters or only those encounters in which the task was relevant. As a result, responses categorized as "rarely" (0%-25%) could reflect near-universal AI use for a small subset of high-acuity or diagnostically complex cases. This ambiguity limits the precision of our frequency estimates and necessitates caution when interpreting AI use as uniformly "low" or purely supplementary. We hypothesize that hospitalists are selectively deploying AI for cases where traditional resources such as UpToDate or PubMed searches are perceived as insufficient or time-intensive. The statistically significant differences in AI usage across clinical applications provide insight into how hospitalists perceive AI's strengths and limitations. The high usage for answering specific clinical questions, generating differential diagnoses, contrasted with lower usage for patient education materials (16/36, 44.4%), may reflect concerns about accuracy and liability when AI-generated content interfaces directly with patients, demonstrating appropriate professional judgment in the face of evolving technology, although it may also reflect a perception of low workflow efficiency in the setting of lack of institutional integration of AI.

Among nonusers (18/54, 33.3%), the primary concerns centered on preference for established resources, AI accuracy, and system-wide endorsement, followed by time constraints and questions about privacy. These findings highlight important barriers that must be addressed for broader adoption, such as validation of AI outputs, integration with existing workflows, and clear guidance on appropriate use cases. The absence of statistically significant demographic predictors of AI adoption suggests that usage patterns may be driven more by individual preferences and comfort with technology than by systematic factors, such as years in practice or clinical role.

Comparison With Previous Work

Purpose-built AI systems have been integrated into clinical workflows for narrowly defined tasks, such as radiologic interpretation [1,2], and experimental studies have compared AI performance against physicians in diagnostic and management scenarios [4-7]. However, adoption patterns of freely available AI tools by health care providers in the absence of institutional integration have remained largely uncharacterized. Our study addresses this knowledge gap by documenting how, and for what purposes hospitalists are independently incorporating AI into their daily practice and for what purposes.

Our observed adoption rate of 66.7% (95% CI 53.4%-77.8%) is comparable to the 66% reported in the 2024 AMA physician survey [23], suggesting that individual adoption patterns may mirror broader trends even without institutional mandates. However, the contexts and applications differ substantially. The AMA survey found that AI was used predominantly for administrative tasks, particularly clinical documentation ranging from discharge summaries to patient portal message responses [23]. Poon et al [19] found health systems primarily implementing AI for imaging or radiology, sepsis detection, and ambient listening clinical documentation. In contrast, our respondents used AI primarily for clinical decision support—answering medical questions, generating differential

XSL•FO
RenderX

diagnoses, and determining management options, although OpenEvidence, the platform chosen by a majority of our respondents, offers the capability to generate documentation meant for patient and health insurance use. This difference likely reflects the nature of institutional versus individual adoption—health systems implementing AI at an organizational level appear to prioritize electronic health record integration and workflow efficiency, whereas individual clinicians selecting their own tools focus on addressing clinical uncertainties and supplementing their decision-making process. Additionally, attempts by health care providers to use AI for documentation without pre-existing systematic integration involve extra steps (such as copy-pasting to and from AI platforms), which may also hinder adoption for such purposes. This distinction has important implications for how we conceptualize AI adoption; organizational implementation focuses on efficiency and administrative burden reduction, while individual adoption focuses on enhancing clinical reasoning and knowledge access.

Among nonusers (18/54, 33.3%) in our study, the primary concerns centered on AI accuracy and preference for established resources, consistent with the findings of Heinrichs et al [22] regarding physician skepticism about AI reliability. Interestingly, none of our respondents cited public or peer perception as factors influencing their decision not to use AI, despite recent studies demonstrating significant perception-based concerns. Reis et al [20] found that US adults rated physicians who used AI as less competent, trustworthy, and empathic, while Yang et al [21] showed that clinicians viewed peers using AI for primary decision-making as significantly less competent. The absence of perception concerns in our sample may reflect several possibilities, such as (1) hospitalists may be unaware of these negative perceptions, (2) they may not consider patient or peer opinions as relevant to their personal tool selection, (3) the anonymous nature of AI use—when used as a supplementary resource consulted privately rather than a visible component of patient-facing care—may shield providers from perception-based concerns, or (4) the lack of perception concerns as a selectable checkbox answer in the question led to the survey not capturing this as a reason, although there was a free-text option for those who wished to provide additional reasons. The discrepancy warrants further investigation, as understanding barriers to AI adoption—whether accuracy-based or perception-based—is essential for developing appropriate governance frameworks and educational initiatives.

The preference for medical-specific platforms observed in our study represents a novel finding not previously documented in the literature. While general-purpose AI tools, like ChatGPT, have achieved widespread adoption among employed adults [10], our hospitalists demonstrated a marked preference for domain-specific tools, with OpenEvidence accounting for the majority of AI users. This suggests that when clinicians independently select AI tools for clinical applications, they prioritize a medical focus and source transparency over general capability or familiarity. Whether this preference is justified by the superior performance of medical-specific platforms remains a question requiring comparative validation studies.

Implications

Understanding current AI usage patterns has significant implications for health care institutions, educators, and policymakers. The widespread organic adoption of AI tools in the absence of institutional oversight underscores an urgent need for governance frameworks that balance innovation with patient safety. Health care systems should consider developing policies that guide appropriate AI use, establish standards for tool selection, and provide training on recognizing AI limitations. As health care systems increasingly invest in AI technologies, empirical data on actual usage patterns become essential for resource allocation and strategic planning. Rather than implementing AI systems that remain unused, institutions can leverage usage data to identify high-value applications that align with clinician workflows and needs. Furthermore, medical education must evolve to prepare trainees for AI-augmented clinical practice, including teaching critical appraisal of AI outputs, understanding the underlying technology and its limitations, and developing frameworks for integrating AI responsibly into clinical decision-making. The convergence on a single platform among users in our study highlights the importance of understanding AI adoption pathways. Whether through peer-to-peer sharing, targeted marketing, or other mechanisms, clinicians are selecting specific tools in ways that may not be captured by traditional technology adoption models. Health care institutions should recognize that even in the absence of formal mandates, departmental culture and informal peer networks likely shape which AI tools gain traction, underscoring the need for governance frameworks that can adapt to grassroots adoption patterns.

The strong preference for a medical-specific AI platform over general-purpose tools observed in our study suggests that clinicians are not passive recipients of whatever AI technology becomes available but are actively selecting tools based on perceived fit with medical practice. This preference may reflect appropriate professional judgment about the importance of medical domain knowledge, source transparency, and health care–specific design. Health care institutions and regulatory bodies should recognize that clinicians' preference for specialized tools creates a need for rigorous, independent validation of these platforms' accuracy, reliability, and safety.

A critical consideration in AI adoption is the dynamic nature of these systems. AI platforms undergo frequent updates by their manufacturers; yet, each iteration does not necessarily represent an improvement over its predecessor [26]. Unlike static clinical resources, such as textbooks or established guidelines, AI systems may change their outputs without notification to users, in the form of variation in accuracy of answers, and ability to perform tasks, such as following multistep questioning between models and even over time within the same model [27], introducing uncertainty into clinical application. As AI use in health care increases, governance frameworks for maintaining patient safety need to be put in place. Zhang and Zhang [28] identified challenges that impact individual and organization trust in AI, including concerns about the quality of data used to train AI systems, accuracy of the underlying algorithms, robustness of health care data protection, responsibility for actions based on AI results, and lack of full

XSL•FO
RenderX

transparency and understanding about the inner workings of how AI produces results from data input [28]. There is no broad regulatory oversight of AI use in health care at present. In the United States, the Food and Drug Administration regulates only some AI systems, such as purpose-built AI for use on medical devices like defibrillators [29].

Given these challenges, primary source verification remains essential when using AI in clinical practice. We recommend that clinicians treat AI-generated information as a starting point requiring confirmation rather than a definitive answer. This principle is particularly important for hospitalists managing acutely ill patients, where diagnostic or therapeutic errors can have immediate consequences.

Limitations

This study has several limitations. As a single-center investigation conducted at a large academic tertiary care hospital, our findings may not be generalizable to community hospitals, rural settings, or outpatient practice environments, where clinical workflows and case complexity differ substantially. The cross-sectional design provides a snapshot of current practice but cannot capture temporal trends or the trajectory of AI adoption. Self-reported usage data may be subject to recall bias, social desirability bias, or selection bias, with respondents possibly more technologically inclined than nonrespondents. Notably, our calculated adoption rate of 66.7% is likely an overestimate of the true institutional adoption rate, as nonusers of AI may have been less motivated to complete a voluntary survey about AI usage patterns compared with enthusiastic users. Additionally, this rate includes respondents across the spectrum of frequency of use, ranging from those who only rarely use AI to power users who use AI much more often. We were unable to directly compare every aspect of respondent demographics with those of the full hospitalist provider group, as detailed information regarding some categories, such as age and experience levels, is not available for nonrespondents. Although the overall response rate was high, residual nonresponse bias cannot be fully excluded. Although no statistically significant demographic predictors of AI use were identified, the relatively small number of AI users (n=36) and limited subgroup sizes (for instance, physicians vs APP) reduce statistical power. As such, the absence of detected associations should be interpreted as the absence of evidence rather than evidence of absence. Another minor limitation is the possible selection bias introduced by our survey design, which required responses (even if selecting a nonresponse option where available) to all questions. Our survey did not undergo formal pilot testing before wide dissemination, possibly having led to some ambiguities. Our frequency estimates asked respondents how often they used AI “per day” for each task without specifying whether to consider all patient encounters or only those where the task would be applicable. For instance, when rating frequency of AI use for generating differential diagnoses, some respondents may have considered all daily patient encounters as the denominator, while others may have mentally restricted it to encounters with diagnostic uncertainty. This ambiguity limits the precision of our frequency estimates, although the data clearly indicate that AI is used infrequently across all measured applications. There was some conceptual overlap between the provided choices of “never” and “rarely (0%-25%)” frequency categories, which could also affect our frequency estimates, although we expect that respondents who never use AI for a specific task, such as for generating a differential diagnosis list, would select “never” over “rarely” as an option. Also, our survey asked respondents for years of practice, which some may have interpreted to include training years, possibly confounding the analysis of experience levels. The survey instruction for years of practice asked respondents to round up to the nearest whole year, which may have resulted in potentially shifting some respondents into higher experience brackets. Although respondents were provided with a free-text option to report additional AI use cases, clinical documentation (separate from generating patient education materials) was not included as one of the predefined questions, which could have led to underreporting compared with tasks explicitly prompted in the survey. This could have contributed to the lower observed prevalence of documentation-related AI use relative to the AMA physician survey, in addition to workflow friction and lack of electronic health record integration. We did not collect data on the accuracy of AI outputs in actual clinical use or on whether AI recommendations were followed, limiting our ability to assess the impact of AI on clinical decision-making. Additionally, we only asked respondents which AI platform (if any) they used most often, and did not collect data on whether they were using more than 1 system. The preference for OpenEvidence may reflect a pattern of primary use (with additional tools used for other scenarios) rather than confirmation of exclusive use.

An important limitation is our lack of data regarding how respondents discovered and selected their AI tools. The striking convergence on OpenEvidence is noteworthy given its status as a niche platform compared with widely known tools, like ChatGPT. Also, although ClinicalKey AI is also a medical-specific platform, its adoption remained low. To our knowledge, there was no formal departmental introduction or recommendation of any platform, although informal peer-to-peer recommendations may have occurred through channels we did not measure. Without systematic data on discovery pathways, we cannot determine whether this convergence reflects informal sharing within our department, the platform’s targeted marketing toward medical professionals, lack of cost barriers, or other factors. While our institution does not block online access to any particular AI platform, other organizations may allow differential access to some systems versus others, which could affect local adoption patterns. Future studies should systematically assess how clinicians learn about and select AI tools, as understanding these adoption pathways has important implications for dissemination strategies and governance frameworks.

Furthermore, our survey did not capture the specific clinical scenarios or patient characteristics that prompted AI use. Despite these limitations, our study has notable strengths. The high response rate (77.1%) enhances the reliability of our findings and minimizes nonresponse bias. The comprehensive assessment of multiple AI platforms and diverse clinical applications they are being used for provides a nuanced picture of real-world usage. Our study to systematically document hospitalist AI

XSL•FO
RenderX

usage patterns independent of organizational endorsement establishes baseline data for future longitudinal investigations and comparative studies across institutions and specialties.

Future Directions

These findings point to several important directions for future research. Longitudinal studies are needed to track how AI usage patterns evolve as these technologies mature and as clinicians gain more experience with them. Comparative effectiveness research should evaluate clinical outcomes associated with AI-augmented versus traditional clinical decision-making. Additionally, research should examine whether AI usage varies across different clinical contexts, patient populations, and decision types, as understanding these nuances could inform the development of more targeted AI applications that address specific clinical needs. This includes future multicenter studies, which should explore interinstitutional variation and correlate AI adoption pathways and usage with clinical outcomes. Ongoing validation studies must continually assess AI system performance across the tasks they are commonly used for, with particular attention to detecting degradation or drift in performance over time. Public and peer-perception–based barriers to AI adoption also warrant further exploration. Finally, comparative studies examining the accuracy, reliability, and clinical utility of medical-specific AI platforms versus general-purpose LLMs for clinical applications would help determine whether the observed preference for specialized tools is justified by superior performance.

Conclusions

This study documents substantial organic adoption of AI tools among hospitalists, with two-thirds reporting clinical use despite the absence of institutional integration or endorsement. The predominant applications—answering clinical questions, generating differential diagnoses, and determining management options—suggest that AI is serving as a supplementary decision support tool for complex cases. The marked preference for a medical-specific platform over a general-purpose AI indicates that hospitalists are exercising discretion in tool selection. This finding suggests both appropriate caution and a need for rigorous validation of medical-specific platforms. However, concerns about accuracy and the low frequency of use highlight the nascent state of AI integration in clinical practice. As AI technologies continue to evolve, health care institutions must develop governance frameworks, validation protocols, and educational initiatives to ensure these tools are deployed safely and effectively in patient care.

Acknowledgments

The authors wish to acknowledge the contributions of Chengcheng Hu (MPH, MMed) for contributing to the statistical analysis of the data. Generative artificial intelligence was not used for generating any part of this manuscript.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: PB (lead), JH (equal), SW (equal)
Data curation: PB (lead), BM (supporting), SW (supporting)
Formal analysis: Chengcheng Hu
Investigation:PB (lead), BM (equal), SW (equal)
Methodology: PB (lead), JH (equal), BM (supporting), SW (supporting)
Project administration: JH (lead), PB (supporting)
Supervision: SW
Visualization: PB (lead), SW (supporting)
Writing: original draft – PB (lead), SW (supporting)
Writing: review & editing – PB (lead), JH (equal), BM (equal), SW (equal)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Copy of the survey questionnaire.
[DOCX File , 22 KB - jmir_v28i1e85973_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
CHERRIES checklist.
[DOCX File , 23 KB - jmir_v28i1e85973_app2.docx]

Multimedia Appendix 3
Distribution of artificial intelligence platforms used by hospitalists in clinical work (n=54). Survey respondents from a large academic medical center, August-September 2025.
[PNG File , 102 KB - jmir_v28i1e85973_app3.png]

Multimedia Appendix 4
Clinical tasks for which artificial intelligence is used among hospitalist artificial intelligence users (n=36). Survey respondents from a large academic medical center, August-September 2025.
[PNG File , 197 KB - jmir_v28i1e85973_app4.png]

References

1. Maleki Varnosfaderani S, Forouzanfar M. The role of AI in hospitals and clinics: transforming healthcare in the 21st century. Bioengineering (Basel) 2024;11(4):337 [FREE Full text] [doi: 10.3390/bioengineering11040337] [Medline: 38671759]
2. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023;23(1):689 [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
3. Williams CYK, Zack T, Miao BY, Sushil M, Wang M, Kornblith AE, et al. Use of a large language model to assess clinical acuity of adults in the emergency department. JAMA Netw Open 2024;7(5):e248895 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.8895] [Medline: 38713466]
4. Takita H, Kabata D, Walston SL, Tatekawa H, Saito K, Tsujimoto Y, et al. A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. NPJ Digit Med 2025;8(1):175 [FREE Full text] [doi: 10.1038/s41746-025-01543-z] [Medline: 40121370]
5. Maillard A, Micheli G, Lefevre L, Guyonnet C, Poyart C, Canouï E, et al. Can chatbot artificial intelligence replace infectious diseases physicians in the management of bloodstream infections? A prospective cohort study. Clin Infect Dis 2024;78(4):825-832. [doi: 10.1093/cid/ciad632] [Medline: 37823416]
6. Kafetzis I, Sodmann P, Herghelegiu B, Brand M, Zoller WG, Seyfried F, et al. Prospective evaluation of real-time artificial intelligence for the Hill classification of the gastroesophageal junction. United European Gastroenterol J 2025;13(2):240-246 [FREE Full text] [doi: 10.1002/ueg2.12721] [Medline: 39668544]
7. Beyaz S, Yayli SB, Kılıç E, Kılıç K. Comparison of artificial intelligence algorithm for the diagnosis of hip fracture on plain radiography with decision-making physicians: a validation study. Acta Orthop Traumatol Turc 2024;58(1):4-9 [FREE Full text] [doi: 10.5152/j.aott.2024.23065] [Medline: 38525504]
8. What is LLM (Large Language Model)? AWS. URL: https://aws.amazon.com/what-is/large-language-model/ [accessed 2026-01-12]
9. ChatGPT. OpenAI. URL: https://chatgpt.com/overview/ [accessed 2025-12-18]
10. Sidoti O, McClain C. 34% of U.S. adults have used ChatGPT, about double the share in 2023. Pew Research Center. 2025. URL: https://tinyurl.com/5n8n4z5r [accessed 2025-11-12]
11. Reid E. AI in Search: going beyond information to intelligence. Google blog. 2025. URL: https://blog.google/products/search/google-search-ai-mode-update/ [accessed 2025-11-12]
12. Mehdi Y. Announcing the next wave of AI innovation with Microsoft Bing and Edge. Microsoft Office blog. 2023. URL: https://blogs.microsoft.com/blog/2023/05/04/announcing-the-next-wave-of-ai-innovation-with-microsoft-bing-and-edge/ [accessed 2025-11-12]
13. Google. Learn about Gemini, the everyday AI assistant from Google. Google. URL: https://gemini.google/about/ [accessed 2025-12-18]
14. Getting started with Microsoft Copilot. Microsoft Support. Microsoft. URL: https://support.microsoft.com/en-us/topic/getting-started-with-microsoft-copilot-8fde147f-726e-4790-9503-70790ddcac73 [accessed 2025-12-18]
15. What is RAG (retrieval-augmented generation)? AWS. URL: https://aws.amazon.com/what-is/retrieval-augmented-generation/ [accessed 2026-01-12]
16. ClinicalKey AI. Elsevier. URL: https://www.elsevier.com/products/clinicalkey/clinicalkey-ai [accessed 2025-12-18]
17. About OpenEvidence. OpenEvidence. URL: https://www.openevidence.com/about [accessed 2025-12-18]
18. Angus DC, Khera R, Lieu T, Liu V, Ahmad FS, Anderson B, JAMA Summit on AI. AI, health, and health care today and tomorrow: the JAMA Summit report on artificial intelligence. JAMA 2025;334(18):1650-1664. [doi: 10.1001/jama.2025.18490] [Medline: 41082366]

XSL•FO
RenderX

19. Poon EG, Lemak CH, Rojas JC, Guptill J, Classen D. Adoption of artificial intelligence in healthcare: survey of health system priorities, successes, and challenges. J Am Med Inform Assoc 2025;32(7):1093-1100. [doi: 10.1093/jamia/ocaf065] [Medline: 40323320]
20. Reis M, Reis F, Kunde W. Public perception of physicians who use artificial intelligence. JAMA Netw Open 2025;8(7):e2521643 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.21643] [Medline: 40674054]
21. Yang H, Dai T, Mathioudakis N, Knight AM, Nakayasu Y, Wolf RM. Peer perceptions of clinicians using generative AI in medical decision-making. NPJ Digit Med 2025;8(1):530 [FREE Full text] [doi: 10.1038/s41746-025-01901-x] [Medline: 40826224]
22. Heinrichs H, Kies A, Nagel SK, Kiessling F. Physicians' attitudes toward artificial intelligence in medicine: mixed methods survey and interview study. J Med Internet Res 2025;27:e74187 [FREE Full text] [doi: 10.2196/74187] [Medline: 40857713]
23. Physician sentiments around the use of AI in health care: motivations, opportunities, risks, and use cases. Shifts from 2023 to 2024. AMA Augmented Intelligence Research. 2025. URL: https://www.ama-assn.org/system/files/physician-ai-sentiment-report.pdf [accessed 2025-12-15]
24. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
25. Hurt RT, Stephenson CR, Gilman EA, Aakre CA, Croghan IT, Mundi MS, et al. The use of an artificial intelligence platform OpenEvidence to augment clinical decision-making for primary care physicians. J Prim Care Community Health 2025;16:21501319251332215 [FREE Full text] [doi: 10.1177/21501319251332215] [Medline: 40238861]
26. Zeff M. OpenAI's new reasoning AI models hallucinate more. TechCrunch. 2025. URL: https://techcrunch.com/2025/04/18/openais-new-reasoning-ai-models-hallucinate-more/ [accessed 2025-10-10]
27. Chen L, Zaharia M, Zou J. How is CHATGPT's behavior changing over time? Harvard Data Science Review. 2024. URL: https://hdsr.mitpress.mit.edu/pub/y95zitmz [accessed 2025-12-17]
28. Zhang J, Zhang ZM. Ethics and governance of trustworthy medical artificial intelligence. BMC Med Inform Decis Mak 2023;23(1):7 [FREE Full text] [doi: 10.1186/s12911-023-02103-9] [Medline: 36639799]
29. Warraich HJ, Tazbaz T, Califf RM. FDA perspective on the regulation of artificial intelligence in health care and biomedicine. JAMA 2025;333(3):241-247. [doi: 10.1001/jama.2024.21451] [Medline: 39405330]

Abbreviations

AI: artificial intelligence
AMA: American Medical Association
APP: advanced practice provider
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
RAG: retrieval-augmented generation

Edited by A Coristine; submitted 15.Oct.2025; peer-reviewed by J Grosser, S Shehata; comments to author 10.Nov.2025; accepted 21.Jan.2026; published 03.Mar.2026.

Please cite as:
Bagla P, Hanna J, Marthambadi B, Watkins S
Patterns of AI Use in Clinical Work by Hospitalists: Survey Study
J Med Internet Res 2026;28:e85973
URL: https://www.jmir.org/2026/1/e85973
doi:10.2196/85973
PMID:41581898

©Prabhava Bagla, Jasmah Hanna, Bhargav Marthambadi, Stacey Watkins. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Psychosis Risk and Generative Artificial Intelligence Use Frequency, Motivations, and Delusion-Like Experiences: Cross-Sectional Survey Study

Benjamin Buck[1], PhD; Anne Julia Maheux[2], PhD

[1]Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[2]Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

Corresponding Author:
Benjamin Buck, PhD
Department of Psychiatry
University of North Carolina at Chapel Hill
MacNider Hall
333 S Columbia Street
Chapel Hill, NC, 27514
United States
Phone: 1 9194454330
Email: ben.buck@med.unc.edu

Abstract

Background: Growth of generative artificial intelligence (GenAI) has exploded in recent years. Many have noted its substantial potential to increase access to scalable digital mental health interventions or provide companions for individuals who are socially isolated. At the same time, seeking mental health support from mainstream GenAI models may involve risks. Several recent examples of exacerbation of delusions have received attention in the popular press, leading to a call for empirical research to document the scope of interactions with GenAI among individuals experiencing symptoms of psychosis.

Objective: This study aimed to evaluate associations of psychosis risk to GenAI use frequency, motivations for use, and GenAI interactions involving potential delusions.

Methods: We conducted a large-scale cross-sectional survey of 1003 young adults in United States, divided the sample of individuals that had used GenAI into “elevated risk” (Prodromal Questionnaire, Brief Version Distress Score ≥20; N=267, 28%) and “low risk” groups (Prodromal Questionnaire, Brief Version Distress Score <20; N=685, 72%), and compared groups on several assessments related to GenAI use.

Results: We found that while members of the elevated risk group were no more likely to have ever used GenAI, they were significantly more likely to report intensive use (odds ratio 1.70 to 2.56; ie, several times per day, more than 30 minutes per day, 6 or more chatbot conversations per day). Those at elevated risk were more likely to report using GenAI to receive social and emotional support and significantly more likely to ascribe human-like roles to their chatbot interactions (odds ratio 1.76 to 3.08; ie, companion, friend, therapist, and romantic partner). Delusion-related interactions were also commonly reported among those at risk for psychosis (item endorsements from 13.3% to 30.7%).

Conclusions: While it is unclear whether they have a positive or negative impact overall, GenAI chatbots may have the potential to impact symptom-related experiences among young adults at risk.

(J Med Internet Res 2026;28:e85038) doi:10.2196/85038

KEYWORDS

digital health; digital mental health; artificial intelligence; generative AI; psychosis risk; psychosis; young adults; mental health

Introduction

Generative artificial intelligence (GenAI) is now the most rapidly adopted technology in history [1]. In July 2025, less than 3 years after its launch in November 2022, ChatGPT (OpenAI) boasted 700 million users, or roughly one tenth of the world’s population [2]. Beyond ChatGPT, several others have experienced similar growth, including other general-use artificial intelligence (AI) platforms (eg, Claude [Anthropic]

XSL•FO
RenderX

and Gemini [Google]), those focused on image and video generation (eg, Stable Diffusion), and others that allow for interaction with customizable characters (eg, Character.AI). Many have noted the potential of GenAI technologies to deliver scalable and self-guided mental health interventions [3], with emerging evidence demonstrating efficacy of interventions leveraging GenAI relative to waitlist control [4]. Survey studies suggest that GenAI systems are already used both by individuals seeking mental health support and by mental health providers to increase efficiency (eg, documenting clinical interactions or researching clinical literature) [5], and prevalence of use is nearly certain to increase in years to come.

While the potential benefit in expanding access to mental health interventions and resources is clear, many have also been monitoring for potential risks, particularly when unrestricted, general use chatbots are used for mental health support [6]. One particular concern is sycophantic behavior, wherein GenAI chatbots overly flatter or agree with users [7], which has been noted in general use GenAI [8]. Another is hallucination or misinformation, a widely observed phenomenon wherein GenAI chatbots provide incorrect details with an overall convincing presentation [9]. Some of these limitations may strengthen atypical or pathological beliefs rather than challenge them. One prominent example of harm occurred in 2023, when Tessa, a chatbot provided by the National Eating Disorders Association, provided problematic guidance related to eating disorders, including encouraging strategies for weight loss [10]. While more recent research on GenAI chatbots for mental health have demonstrated improved efficacy and capacity for preventing negative outcomes [4,11], there remain questions about risks associated with GenAI use for mental health concerns, particularly when the GenAI platforms are not specialty systems designed for such a purpose and with guardrails.

An emerging concern in this area is psychosis risk [12]. Psychosis symptoms can be arranged along a transdiagnostic risk continuum, ranging from common and low severity symptoms (eg, fleeting and infrequent hallucinatory experiences, preoccupation with odd beliefs) to rare, highly severe symptoms (eg, ongoing command hallucinations, highly distressing and high-conviction delusions) that are often linked with negative outcomes [13]. Schizophrenia-spectrum disorders emerge following a prodromal period (typically occurring in late teens to mid-twenties) wherein individuals experience subthreshold symptoms and changes from premorbid functioning. Experiencing these prodromal symptoms is associated with a significant increase in likelihood of both schizophrenia-spectrum [13] or other psychiatric disorders [14]. Young people within the age group that is at the highest risk for first developing a psychotic disorder are among the highest adopters of GenAI [1]. Individuals at elevated risk for psychosis are also more likely to engage in problematic use of other digital media [15-17], and such problematic use appears associated with elevated stress in this population [18,19] as well as risk for future symptoms [20]. Taken together, these findings suggest a need for research to understand the use of GenAI among individuals at elevated risk for psychosis.

Several well-publicized examples of troubling interactions with GenAI systems have raised concerns about the ways these technologies could introduce new risks for individuals at risk for psychosis. In several interactions described in the popular press, users have shared delusions or odd beliefs with the model, and in response, the GenAI chatbot has reportedly responded in ways that seem to have increased delusional conviction, distress, or behavioral responses. For example, in one case, a 29-year-old mother of 2 came to believe that ChatGPT could facilitate “interdimensional communication” [21]. In another, a 35-year-old man with a history of bipolar disorder and schizophrenia became fixated on an AI entity called Juliet; when he came to believe that OpenAI executives had killed this entity, he threatened to kill them and was killed after attacking law enforcement with a knife [21]. In other, less tragic examples, individuals with no history of mental illness have embraced grandiose narratives about their own knowledge or abilities [22]. While these extreme examples are a concern, there is a lack of evidence to inform the public as to the prevalence of these risks at the population level. It is unclear the extent to which publicized cases with negative outcomes may indicate a more common phenomenon. A first step in answering this question involves assessing the amount and type of GenAI use among young adults at varying risk levels for psychosis, evaluating whether their interactions with GenAI chatbots differ in meaningful ways from the general population, and to assess the frequency of interactions that could impact mental health outcomes. There is a lack of empirical studies on relationships of psychosis risk to (1) GenAI use frequency and (2) motivations for GenAI use, and (3) delusion-related interactions with GenAI models.

Our team conducted a large-scale (N=1003) survey of young adults in July 2025. In addition to measures of GenAI use frequency, motivations for GenAI use, and GenAI experiences, we collected a screening questionnaire assessing psychosis risk. In this study, we examine cross-sectional relationships of GenAI variables to symptoms indicating elevated psychosis risk. Based on literature suggesting elevated rates of problematic use of digital media among individuals at risk for psychosis, we expect that individuals at elevated risk will be more likely to report the highest patterns of use and to use GenAI for social and emotional support. We also expect that individuals at elevated risk will be more likely to report interacting with GenAI about symptom-related content. Study results could provide an important first step to identifying the potential risks and benefits of these platforms for individuals at risk of developing psychosis, as well as characterizing their scale and level of impact.

Methods

Study Design

We collected a cross-sectional survey in July 2025 via the online crowdsourcing platform Prolific. Our survey was open to all Prolific users, given they met our inclusion criteria of (1) residing in the United States and (2) being between ages 18 and 25 years. Convenience sampling was used as our sample comprised all individuals who responded to the study listing. Our study had multiple goals. First, we aimed to examine overall relationships between psychosis risk and self-reported use of

XSL•FO
RenderX

GenAI (eg, lifetime use, frequency, and average duration of use). Second, we sought to characterize motivations for and types of GenAI use by examining associations with psychosis risk symptoms. Third, we developed a measure to assess delusion-related interactions with GenAI systems to characterize the frequency with which individuals at elevated risk for psychosis reported these experiences.

Participants

Participants were young adults between the ages of 18 and 25 years in the United States (for sample demographics, see Multimedia Appendix 1). On Prolific, participants can complete screening, provide identity verification, and receive compensation through Prolific without sharing identifying information with researchers. Thus, all study participants were anonymous to the study team, and all received information about crisis support or finding mental health resources during study participation. This sample includes all individuals who provided sufficient responses to the Prodromal Questionnaire, Brief Version (PQ-B [23]) to generate an estimate of psychosis risk (ie, 14 or more of the 21 items). This exclusion criterion resulted in removing only 7 participants from the original study sample. In addition to 44 (4.4%) that were removed for missing at least 1 of 2 attention checks, this resulted in a final sample of 952 participants.

Measures

Psychosis risk was assessed with the PQ-B [23]. The PQ-B is widely used in identifying individuals experiencing symptoms that may indicate risk for psychotic episodes (ie, prodromal symptoms). The PQ-B has 21 items that are rated on their occurrence and a Likert scale assessing distress related to items that the respondent endorsed. While a high score on the PQ-B does not on its own indicate meeting clinical high-risk criteria (a gold-standard determination of this is made by a clinician), it does indicate that a formal evaluation is more likely to find that the respondent is indeed at risk. And in population-level research, the PQ-B Distress Score, which captures both symptoms endorsed and associated distress, can be used continuously as a measure of psychosis risk symptoms. Systematic reviews of the PQ-B have identified optimal cutoffs for screening and referral for further evaluation in both clinical and nonclinical settings [24]. For analyses that require identifying a subset of participants at elevated risk, we follow a provisional cutoff of 20 used in prior work in both clinical [25] and online settings [26] to generate “elevated risk” (PQ-B Distress Total ≥20) and “low risk” (PQ-B Distress Total <20).

Our team developed measures to evaluate AI use frequency, motivations for use, and interactions with GenAI involving delusion-like experiences. Regarding use, we administered 4 items addressing frequency, recency, typical length of sessions, and typical number of daily sessions (see Multimedia Appendix 2 for items and response options). The Artificial Intelligence Motivation and Uses Scale (AIMUS) surveys participants’ motivations for and types of use of GenAI systems in the following categories, derived via a factor analysis conducted by our team that is as yet unpublished [27] (see Multimedia Appendix 1 for more psychometric information about the scale, all items, and factor loadings): (1) task automation (ie, completing onerous or rote tasks with the assistance of the system), (2) learning and exploration (ie, searching for information, generating ideas, or receiving help understanding difficult topics), (3) emotional support (ie, support in managing emotions, thinking through decisions, or managing worrying thoughts), (4) dating and sexuality (ie, interacting with GenAI in romantic or sexual role-plays). Participants endorse items on a Likert scale ranging from 1 (“Never”) to 5 (“Very Often”). Subscale scores were mean imputed for individuals that completed at least half the items on the scale. We also asked one item about relationships that respondents felt they had with AI systems. For these items, participants were asked to report whether they considered any AI system to be a companion, a therapist, a friend, a romantic partner, or a sexual partner.

Delusion-like experiences involving AI were assessed with the Generative AI Aberrant Thoughts and Experiences Scale (GAATES). To develop the GAATES, study team members identified the most common delusion categories from recent meta-analytic work on psychosis [28], and created items specific to GenAI interactions. These depicted both (1) potential delusions about AI (eg, paranoid thoughts about AI, grandiose thoughts about one’s interactions with AI), as well as (2) uses of AI that may maintain existing delusions (eg, interacting with AI to reinforce one’s paranoid or grandiose beliefs). Item responses range from 1 (“Strongly Disagree”) to 5 (“Strongly Agree”). Exploratory factor analysis indicated that the scale is unidimensional, with all items loading (loadings >.45) onto a single factor that explained 53% of the variance in the underlying construct. For analyses examining endorsement of each phenomenon as a binary variable, responses of “Agree” or “Strongly Agree” (ie, 4 or 5) were coded as endorsement, while all other responses (including 3, “Neither Agree nor Disagree”) were coded as nonendorsement. Scale internal consistency was high (Cronbach α=0.95). Full item responses by group as well as results from the psychometric analyses (eg, factor analysis) are reported in the Multimedia Appendix 1.

Data Analysis

Given the nature of the scientific question, we took an exploratory approach to characterize relationships and group differences broadly across a variety of constructs. Aside from the variable related to previous GenAI use itself, all analyses included only the subset of the sample reporting current or previous GenAI use. Alpha level for statistical significance was set at .05. First, to examine the relationships of psychosis risk to use frequency variables, we split the sample based on frequency responses indicating intensive use of GenAI. This generated binary variables representing four comparisons: (1) individuals who reported using GenAI several times per day versus less, (2) those who reported using GenAI the day of the survey versus longer ago, (3) those who reported using GenAI more than 30 minutes per session versus shorter than this, and (4) those who reported having 6 or more conversations with the chatbot per day versus 5 or fewer. We conducted chi-square tests and computed odds ratios (ORs) based on the cross product of cells in contingency tables, with CIs based on the standard error of the log ORs. PQ-B Distress scores across were compared across high versus low GenAI frequency using independent samples *t* tests and as well as chi-square tests and

XSL•FO
RenderX

ORs (also derived from contingency tables) to characterize the frequency of intensive AI use among each psychosis risk group (elevated risk vs low risk). Cohen *d* estimates were calculated using pooled SDs. Second, to evaluate relationships to AI motivations, we examined Pearson correlations of PQ-B Distress scores to each subscale on the AIMUS. Third, to assess relationship variables, we compared the frequency of endorsing each relationship (friend, therapist, companion, romantic partner, and sexual partner) among each risk group using chi-square tests compared PQ-B Distress scores among individuals who endorsed each relationship using independent samples *t* tests. Finally, we examined the Pearson correlation of the GAATES to PQ-B Distress score and compared frequency of endorsing each delusion-related item across risk groups using chi-square tests and ORs. For all Pearson correlations, CIs were generated using Fisher *r*-to-*z* transformations. All linear models were evaluated for assumption violations; nonnormal variables were reanalyzed with appropriate nonparametric models and/or variable transformations, and the pattern of results were unchanged (Multimedia Appendix 1).

Ethical Considerations

All research procedures were reviewed by the institutional review board at the University of North Carolina at Chapel Hill and were deemed exempt (IRB#25-1686) on June 26, 2025. All participants provided informed consent through online data capture, and because no individually identifying information was collected, data were anonymized. Based on an estimated 30-minute completion time and an intended compensation rate of US $8 per hour, participants were compensated US $4 for their time.

Results

Frequency and Average Length of Use

Of the sample of 952, 846 (88.9%) had used GenAI, while the remainder (N=106, 11.1%) had not. There were no significant differences in the PQ-B Distress score between individuals who had never used GenAI platforms (mean 13.05, SD 17.06) and those who had (mean 13.90, SD 17.14; mean difference=–0.85, 95% CI –4.31 to 2.61; t_{950}=–0.48; Cohen *d*=–0.05, 95% CI –0.25 to 0.15; *P*=.63). Participant demographics can be found in Multimedia Appendix 1.

Regarding frequency, individuals who reported using GenAI "several times per day" or more had significantly elevated PQ-B Distress scores (113/845, 13.4%; mean 21.75, SD 20.98) relative those reporting using GenAI less frequently (732/845, 86.6%; mean 12.71, SD 16.15), mean difference=–9.04 (95% CI –12.39 to –5.69), t_{843}=–5.30, *P*<.001, Cohen *d*=–0.54 (95% CI –0.74 to –0.34). Individuals who reported that their most recent use of GenAI was the day of the survey (ie, "Today"; N=287, 34.0%) had significantly elevated PQ-B Distress scores (mean 17.00, SD 19.54) relative to those who reported that their most recent use was the day prior or earlier (558/845, 66.0%; mean 12.33, SD 15.54; mean difference=–4.66, 95% CI –7.09 to –2.24; t_{843}=–3.77; Cohen *d*=–0.27, 95% CI –0.42 to –0.13; *P*<.001). Individuals that reported that they would initiate over 5 conversation sessions per day (96/803, 12.0%) also had significantly elevated PQ-B Distress scores (mean 20.51, SD 21.42) relative to those who reported having 5 conversation sessions or fewer per day (707/803, 88.0%; mean 13.15, SD 16.53; mean difference=–7.36, 95% CI –11.03 to –3.69; t_{801}=–3.94; Cohen *d*=–0.43, 95% CI –0.64 to –0.21; *P*<.001). Expressed as ORs in binary analyses (Table 1), individuals at elevated risk for psychosis (using PQ-B Distress Total ≥ 20) had more than double the odds of being in the respective intensive use category regarding frequency of use and number of conversation sessions, and they were over 50% higher odds of reporting GenAI use on the day of the survey. There were no differences between groups divided on the variable of length of each episode of use.

XSL•FO
RenderX

Table 1. Responses to generative artificial intelligence frequency items among individuals endorsing lifetime generative artificial intelligence use (N=846). Numbers vary by item as a result of missing responses.

Variables	Low risk (n=605), n (%)	Elevated risk (n=241), n (%)	OR[a] (95% CI)	Chi-square (*df*)	*P* value
How often do you use generative AI[b] chatbots (eg, ChatGPT, Claude, and Gemini) for any purpose? (N=845)			2.56 (1.71-3.83)	21.62 (1)	<.001
Less than "Several Times Per Day" (n=732, 86.6%)	544 (90.1)	188 (78.0)			
"Several Times Per Day" (n=113, 13.4%)	60 (9.9)	53 (22.0)			
When was the last time you used a generative AI chatbot? (N=845)			1.70 (1.25-2.32)	11.57 (1)	<.001
Any time before "Today" (n=558, 66%)	420 (69.5)	138 (57.3)			
"Today" (n=287, 34%)	184 (30.5)	103 (42.7)			
When you do use generative AI chatbots, how much time do you typically spend per session? (N=846)			1.32 (0.84-2.08)	1.42 (1)	0.23
Less than 30 minutes (n=751, 88.8%)	542 (89.6)	209 (86.7)			
More than 30 minutes (n=95, 11.2%)	63 (10.4)	32 (13.3)			
On a day when you use a generative AI chatbot, how many separate times do you typically start a new conversation or ask it about something different? (N=803)[c]			2.10 (1.36-3.25)	11.49 (1)	<.001
Five or fewer times (707, 88%)	516 (90.5)	191 (82.0)			
Six or more times (96, 12%)	54 (9.5)	42 (18.0)			

[a]OR: odds ratio.

[b]AI: artificial intelligence.

[c]This excludes responses of "I don't know/I haven't paid attention," which were provided as a response item for this question.

Motivations for AI Use

All motivations for use were positively associated with psychosis risk in the full sample, with the strongest correlations being between PQ-B distress scores and emotional support motivations (r=0.33, 95% CI 0.26-0.38; P<.001), followed by dating and sexuality (r=0.23, 95% CI 0.16-0.29; P<.001), task automation (r=0.15, 95% CI 0.08-0.21; P<.001) and learning and exploration (r=0.14, 95% CI 0.08 to 0.21; P<.001).

AI Relationship Variables

Of the AI relationships individuals reported, items ranged in endorsement from around 3% (AI as a sexual partner; 26/846, 3.1%) to over a third (AI as a companion; 315/846, 37.2%) of all participants in the sample who had previously used GenAI. PQ-B Distress Scores were higher for individuals that endorsed feeling that AI served as a companion (mean 16.75, SD 18.73 vs mean 12.20, SD 15.89; mean difference=–4.55, 95% CI –6.92 to –2.18; t_{844}=–3.76; Cohen d=–0.27, 95% CI –0.41 to –0.13; P<.001), therapist (mean 20.77, SD 20.59 vs mean 11.10, SD 14.63; mean difference=–9.67, 95% CI –12.14 to –7.21; t_{844}=–7.70; Cohen d=–0.58, 95% CI –0.74 to –0.43; P<.001), friend (mean 18.70, SD 19.53 vs mean 11.30, SD 15.08; mean difference=–7.40, 95% CI –9.77 to –5.03; t_{844}=–6.12; Cohen d=–0.44, 95% CI –0.58 to –0.30; P<.001), and romantic partner (mean 21.97, SD 18.89 vs mean 13.58, SD 17.00; mean difference=–8.39, 95% CI –14.43 to –2.35; t_{844}=–2.73; Cohen d=–0.49, 95% CI –0.85 to –0.14; P=.007). Examined as odds ratios between risk groups (Table 2), GenAI users in the elevated risk group had approximately threefold higher odds of reporting seeing GenAI as a therapist, more than double the odds of seeing GenAI as a friend or romantic partner, and more than 70% higher odds of seeing GenAI as a companion. There were no differences between the groups in the item related to viewing GenAI as a sexual partner.

XSL•FO
RenderX

Table 2. Responses to generative artificial intelligence relationship items among individuals endorsing lifetime generative artificial intelligence use (N=846). Numbers vary by item as a result of missing responses.

Variables	Low risk (n=605), n (%)	Elevated risk (n=241), n (%)	OR[a] (95% CI)	Chi-square (*df*)	*P* value
Companion (N=846)			1.76 (1.30-2.39)	13.44 (1)	<.001
No (n=531, 62.8%)	403 (66.6)	128 (53.1)			
Yes (n=315, 37.2%)	202 (33.4)	113 (46.9)			
Therapist (N=846)			3.08 (2.24-4.24)	50.24 (1)	<.001
No (n= 601, 71%)	472 (78.0)	129 (53.5)			
Yes (n=245, 29%)	133 (22.0)	112 (46.5)			
Friend (N=846)			2.52 (1.85-3.43)	35.61 (1)	<.001
No (n=549, 64.9%)	430 (71.1)	119 (49.4)			
Yes (n=297, 35.1%)	175 (28.9)	122 (50.6)			
Romantic partner (N=846)			2.62 (1.29-5.32)	7.56 (1)	.006
No (n=814, 96.2%)	589 (97.4)	225 (93.4)			
Yes (n=32, 3.8%)	16 (2.6)	16 (6.6)			
Sexual partner (N=846)			0.92 (0.38-2.22)	0.03 (1)	.86
No (n=820, 96.9%)	586 (96.9)	234 (97.1)			
Yes (n=26, 3.1%)	19 (3.1)	7 (2.9)			

[a]OR: odds ratio.

Delusion-Like Experiences

GAATES scores differed between individuals at elevated risk (mean 36.22, SD 14.74) and those at low risk (mean 26.39, SD 10.33; mean difference=–9.83, 95% CI –11.59 to –8.07; t_{844}=–10.98; Cohen *d*=–0.84, 95% CI –0.99 to –0.68; *P*<.001). Total GAATES scores significantly correlated with PQ-B Distress in the full sample (*r*=0.40, 95% CI 0.34-0.45; *P*<.001) as well as separately in the elevated risk group (*r*=0.28, 95% CI 0.16-0.39; *P*<.001) and the low risk group (*r*=0.11, 95% CI 0.03-0.19; *P*=.008). For all items, rates of endorsement frequency differed between risk groups (Table 3). Several of the items that most differed between those at low and elevated risk most often pertained to paranoia or persecutory ideation (eg, “AI provides me facts about how others are working to harm me”; 41/241, 17% in the elevated risk group vs 19/605, 3.1% in the low-risk group; “AI is being used by others to harm me”, 33/241, 13.7% vs 17/605, 2.8%; “AI has shown me how others are trying to control my actions”, 61/241, 25.3% vs 37/605, 6.1%), while others related to grandiosity (“I’ve discovered hidden or secret truths about the world through AI”, 43/241, 17.8% vs 22/605, 3.6%; “AI helps me make sense of secret messages (eg, from television or the news) that were intended only for me”, 58/241, 24.1% vs 42/605, 6.9%) also clearly distinguished the groups.

XSL•FO
RenderX

Table 3. Responses to individual Generative AI Aberrant Thoughts and Experiences Scale items among individuals endorsing lifetime generative artificial intelligence use (N=846).

Variables	Low risk (n=605), n (%)	Elevated risk (n=241), n (%)	OR[a] (95% CI)	Chi-square (*df*)	*P* value
AI[b] tries to read or manipulate my thoughts	36 (6.0)	38 (15.8)	2.96 (1.83-4.80)	20.81 (1)	<.001
AI tries to control my behavior	25 (4.1)	33 (13.7)	3.68 (2.14-6.34)	24.67 (1)	<.001
AI helps me understand that others are reading or manipulating my thoughts	51 (8.4)	57 (23.7)	3.37 (2.23-5.09)	35.86 (1)	<.001
AI has shown me how others are trying to control my actions	37 (6.1)	61 (25.3)	5.20 (3.35-8.09)	62.00 (1)	<.001
AI communicates things to me that only I can understand	46 (7.6)	61 (25.3)	4.12 (2.71-6.26)	48.92 (1)	<.001
AI can reveal the truth that I am a special, unique, or powerful person	53 (8.8)	68 (28.2)	4.09 (2.75-6.09)	53.22 (1)	<.001
AI helps me make sense of secret messages (eg, from television or the news) that were intended only for me	42 (6.9)	58 (24.1)	4.25 (2.76-6.54)	48.49 (1)	<.001
AI interacts with me in a special way because of who I am	46 (7.6)	62 (25.7)	4.21 (2.77-6.39)	50.83 (1)	<.001
AI is being used to secretly monitor me specifically	28 (4.6)	33 (13.7)	3.27 (1.93-5.54)	21.17 (1)	<.001
AI is being used by others to harm me	17 (2.8)	33 (13.7)	5.49 (2.99-10.06)	36.71 (1)	<.001
AI helps me learn how people are spying on or monitoring me	20 (3.3)	32 (13.3)	4.48 (2.51-8.00)	29.71 (1)	<.001
AI provides me facts about how others are working to harm me	19 (3.1)	41 (17.0)	6.32 (3.59-11.15)	50.33 (1)	<.001
AI systems are at their core an attempt by powerful people to control the world	89 (14.7)	74 (30.7)	2.57 (1.80-3.66)	28.35 (1)	<.001
AI systems use data from their users to influence world events	64 (10.6)	65 (27.0)	3.12 (2.13-4.59)	35.84 (1)	<.001
I've discovered hidden or secret truths about the world through AI	22 (3.6)	43 (17.8)	5.76 (3.36-9.86)	49.04 (1)	<.001
I have gained access to information through AI about the true nature of the world that I could not find in mainstream sources	43 (7.1)	56 (23.2)	3.96 (2.57-6.09)	43.39 (1)	<.001

[a]OR: odds ratio.

[b]AI: artificial intelligence.

Discussion

Principal Findings

Due to a growing collection of stories publicized by the popular press, interest has grown regarding the mental health impacts of GenAI systems among individuals prone to psychosis. This study is the first of its kind, to our knowledge, to examine the cross-sectional relationships of psychosis risk to GenAI use, motivations, and delusion-related interactions. Results suggest that while individuals at elevated risk for psychosis do not differ from the general population in likelihood to have ever used GenAI, they may indeed be more likely to use these systems at high levels of frequency, to seek emotional support from them, and to engage in interactions related to preexisting aberrant beliefs, including paranoia, grandiosity, and conspiracy thinking. While it remains unclear at present whether these interactions may on balance affect the mental health of this population in a positive or negative manner, our results suggest that they are widely used and thus, those impacts—whatever they may be—may be occurring at a large scale.

First, regarding frequency of use, psychosis risk appeared linked with likelihood to engage in the highest levels of GenAI use. Individuals at elevated risk for psychosis had more than twice the odds of using GenAI several times per day or having six or more initiated GenAI conversations per day as those at low risk. They were also significantly more likely to report having last used GenAI the same day that they took the survey. These findings align with extant research that suggests that individuals at elevated risk for psychosis are more likely to engage in intensive internet use [15,29,30]. Studies demonstrating a link of psychosis risk to media use focus on extreme or problematic use, while other studies examining more continuous relationships have found more mixed results [20,31,32]. Analogous research on GenAI is lacking, so definitions of

XSL•FO
RenderX

problematic use are underdeveloped, and risks and benefits of GenAI are not well understood. Regardless, our results suggest that intensive use of GenAI may, in a manner similar to general internet use, be more common among individuals at elevated risk for psychosis. Future research could address whether these relationships reflect a tendency of individuals with symptoms of psychosis to use AI, for AI use to influence symptoms, a third factor to influence both AI use and psychosis symptoms, or some combination of all these possibilities.

Second and relatedly, psychosis risk symptoms were associated with all measured motivations for using GenAI (likely reflecting a higher interest in chatbots among those at elevated risk); yet, this relationship was strongest for motivation to use GenAI for social and emotional support. While our cross-sectional data cannot speak to directionality of these associations, this finding aligns with previous results suggesting that online activities may be compensatory for unsatisfying or insufficient real-world social interactions and support [33]. Consistent with our findings on GenAI motivations, we also found that individuals at risk for psychosis were significantly more likely to ascribe human relationships to GenAI platforms. Participants at elevated risk were much more likely to see a GenAI system as their therapist, friend, romantic partner, or companion, although they did not differ in their likelihood to see AI as a sexual partner. Importantly, one driver of compulsive internet use among individuals experiencing mental health concerns is loneliness [15,33,34]. Since such platforms are easy to anthropomorphize [35], they may be seen more readily as a way to satisfy social needs, but it is unclear whether they may lead to greater benefits or risks in the short and long terms. One concern, for example, is that they could—through reducing incentives for human interactions—exacerbate loneliness over the long-term [36]. The fact that social support is a prominent motivation for GenAI use highlights the importance of research to answer this question.

Finally, we developed a new measure of GenAI interactions involving delusion-like experiences and evaluated the association of this measure with psychosis risk symptoms, finding a significant medium correlation (r=0.40, P<.001) between these 2 factors. Of the items on this new scale, no item was endorsed by a proportion smaller than 10% of the at-risk sample. This included around a quarter of the sample reporting that they used GenAI to make sense of special messages, the true nature of the world, or persecutory actions taken by others. This finding suggests that individuals with psychosis risk symptoms do interact with GenAI about topics related to paranoia, grandiosity, or conspiracy thinking. On one hand, this is intuitive. Users are likely to engage with GenAI systems about topics to which they dedicate thought. For individuals at risk for psychosis, this may include persecutory or grandiose ideas. However, given large-scale interactions with GenAI, the nature of responses provided by these systems are particularly important. If GenAI chatbots respond in sycophantic or reinforcing ways or provide misinformation, this could increase risks. If they respond in ways that reduce stigma and increase the likelihood of help-seeking, this could lead to benefits. Our data do not suggest that these delusion-related interactions are leading to either harm or benefit—such a claim would require prospective data; however, our results do indicate that interactions about symptom-related content appear to be happening on a scale that could lead to significant impact. Protective measures could reduce these risks or even leverage GenAI in a manner that could improve health outcomes, for example, by gently challenging problematic beliefs, destigmatizing help-seeking, or encouraging individuals to seek out treatment from a licensed professional.

Limitations

Our study is limited in several ways. First, these are cross-sectional data. Our analyses identify associations between variables where directionality remains unclear. It is possible that relationships reveal the influence psychosis risk has on GenAI use, vice versa, or some combination of the two. Future studies with longitudinal designs are necessary to answer questions about directionality. Second, data were collected via Prolific. Prolific users may not sufficiently represent the broader young adult population in the United States; a more inclusive or varied sampling strategy may do so more effectively. In particular, given their participation through an online crowdsourcing platform, our sample may be more likely to have used a variety of online tools, including GenAI, than the general population. Further, our sampling strategy involved collecting data from all Prolific users who responded to our study listing; a more intentional sampling strategy could have resulted in more generalizable results. Relatedly, while the PQ-B is a widely used validated assessment of prodromal symptoms, as mentioned, a high score on this scale on its own does not indicate that an individual meets clinical high-risk criteria. This determination is best made by a trained clinician. Third, given the novelty of this research area, many of the measures administered were developed for the proposed project. There also may be some conceptual overlap between measures of GenAI interactions and those that assess psychological constructs. While we based development of each measure on published literature and assessed psychometric attributes, future studies may reveal more comprehensive or robust measures. As is always the case with survey studies dependent on participant report, results may be impacted by response biases, or social desirability effects. Future designs should combine self-reported motivations for GenAI use with objective or passively collected device data.

Conclusions

These findings speak to the importance of intentional and transparent design of GenAI systems to reduce risks to individuals at risk for psychosis. Individuals at elevated risk for psychosis appear more likely to use these tools with high frequency, to do so for social and emotional support, and to ascribe intimate relationships to chatbots. Given the intensity and frequency of these interactions, individuals at risk may be more sensitive to problematic content. Longitudinal research is needed to determine the real-world impact of GenAI on psychosis risk symptoms, but our study indicates that whatever their consequences, meaningful symptom-related interactions may be common among this population.

XSL•FO
RenderX

Acknowledgments

The authors thank research participants for their time and candid responses and recognize the work of members of the Social Environments and Adolescence (SEA) Lab at University of North Carolina for testing and optimizing the data collection protocol. AM acknowledges the support of the Winston Family Foundation through the Winston Center on Technology and Brain Development. BB recognizes the support of the Foundation of Hope through the Child and Adolescent Anxiety and Mood Disorders Program (CHAAMP). The authors used generative artificial intelligence to assist in identification of key citations in the literature review and in adjusting formatting of tables. These systems were not used in the collection or analysis of data, the design of the study nor in drafting the manuscript. Authors have reviewed and take full responsibility for the manuscript.

Funding

BB is supported by a Mentored Patient-Oriented Career Development Award from the National Institute of Mental Health (K23MH122504). The views expressed in this manuscript do not necessarily represent the views of the National Institute of Mental Health, nor did the sponsor play any role in the conception or drafting of this manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization, data curation, formal analysis, methodology, writing – original draft, writing – review and editing: BB
Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, supervision, validation, writing – original draft, writing – review and editing: AJM

Conflicts of Interest

None declared.

Multimedia Appendix 1
Item list for novel measures.
[DOCX File , 21 KB - jmir_v28i1e85038_app1.docx]

Multimedia Appendix 2
Supplemental analyses.
[DOCX File , 101 KB - jmir_v28i1e85038_app2.docx]

References

1. Bick A, Blandin A, Deming D. The rapid adoption of generative AI. National Bureau of Economic Research 2024:1-32 [FREE Full text] [doi: 10.3386/w32966]
2. Chatterji A, Cunningham T, Deming D, Hitzig Z, Ong C, Shan C, et al. How people use ChatGPT. National Bureau of Economic Research 2025:1-62 [FREE Full text] [doi: 10.3386/w34255]
3. D'Alfonso S. AI in mental health. Curr Opin Psychol 2020;36:112-117. [doi: 10.1016/j.copsyc.2020.04.005] [Medline: 32604065]
4. Heinz MV, Mackin DM, Trudeau BM, Bhattacharya S, Wang Y, Banta HA, et al. Randomized trial of a generative AI chatbot for mental health treatment. NEJM AI 2025;2(4). [doi: 10.1056/aioa2400802]
5. Cross S, Bell I, Nicholas J, Valentine L, Mangelsdorf S, Baker S, et al. Use of AI in mental health care: community and mental health professionals survey. JMIR Ment Health 2024;11:e60589 [FREE Full text] [doi: 10.2196/60589] [Medline: 39392869]
6. Wang L, Bhanushali T, Huang Z, Yang J, Badami S, Hightow-Weidman L. Evaluating generative AI in mental health: systematic review of capabilities and limitations. JMIR Ment Health 2025;12:e70014 [FREE Full text] [doi: 10.2196/70014] [Medline: 40373033]
7. Malmqvist L. Sycophancy in large language models: causes and mitigations. In: Intelligent Computing. 2025 Presented at: Proceedings of the 2025 Computing Conference; June 19-20, 2025; London, UK (Clayton Hotel Chiswick) p. 61-74. [doi: 10.1007/978-3-031-92611-2_5]
8. Naddaf M. AI chatbots are sycophants - researchers say it's harming science. Nature 2025;647(8088):13-14. [doi: 10.1038/d41586-025-03390-0] [Medline: 41136779]

XSL•FO
RenderX

9. Massenon R, Gambo I, Khan JA, Agbonkhese C, Alwadain A. "My AI is Lying to Me": user-reported LLM hallucinations in AI mobile apps reviews. Sci Rep 2025;15(1):30397 [FREE Full text] [doi: 10.1038/s41598-025-15416-8] [Medline: 40830185]
10. Wells K. An eating disorders chatbot offered dieting advice, raising fears about AI in health. NPR. 2023. URL: https://www.npr.org/sections/health-shots/2023/06/08/1180838096/an-eating-disorders-chatbot-offered-dieting-advice-raising-fears-about-ai-in-hea [accessed 2025-09-16]
11. Campellone TR, Flom M, Montgomery RM, Bullard L, Pirner MC, Pavez A, et al. Safety and user experience of a generative artificial intelligence digital mental health intervention: exploratory randomized controlled trial. J Med Internet Res 2025;27:e67365 [FREE Full text] [doi: 10.2196/67365] [Medline: 40408143]
12. Østergaard SD. Will generative artificial intelligence chatbots generate delusions in individuals prone to psychosis? Schizophr Bull 2023;49(6):1418-1419. [doi: 10.1093/schbul/sbad128] [Medline: 37625027]
13. van Os J, Reininghaus U. Psychosis as a transdiagnostic and extended phenotype in the general population. World Psychiatry 2016;15(2):118-124 [FREE Full text] [doi: 10.1002/wps.20310] [Medline: 27265696]
14. Fusar-Poli P, Rocchetti M, Sardella A, Avila A, Brandizzi M, Caverzasi E, et al. Disorder, not just state of risk: meta-analysis of functioning and quality of life in people at high risk of psychosis. Br J Psychiatry 2015;207(3):198-206. [doi: 10.1192/bjp.bp.114.157115] [Medline: 26329563]
15. Pelletier-Baldelli A, Ives L, Mittal VA. Increased internet use and poorer ability to manage emotions in youth at high-risk for psychosis. Schizophr Res Cogn 2015;2(4):220-226 [FREE Full text] [doi: 10.1016/j.scog.2015.08.002] [Medline: 26855886]
16. Fekih-Romdhane F, Malaeb D, Loch AA, Farah N, Stambouli M, Cheour M, et al. Problematic smartphone use mediates the pathway from suicidal ideation to positive psychotic experiences: a large cross-sectional, population-based study. Int J Ment Health Addict 2023:1-18 [FREE Full text] [doi: 10.1007/s11469-023-01028-8] [Medline: 36820017]
17. Lee JY, Chung YC, Song JH, Lee YH, Kim JM, Shin IS, et al. Contribution of stress and coping strategies to problematic Internet use in patients with schizophrenia spectrum disorders. Compr Psychiatry 2018;87:89-94. [doi: 10.1016/j.comppsych.2018.09.007] [Medline: 30282059]
18. Fekih-Romdhane F, Jahrami H, Away R, Trabelsi K, Pandi-Perumal SR, Seeman MV, et al. The relationship between technology addictions and schizotypal traits: mediating roles of depression, anxiety, and stress. BMC Psychiatry 2023;23(1):67 [FREE Full text] [doi: 10.1186/s12888-023-04563-9] [Medline: 36698079]
19. Fekih-Romdhane F, Away R, Jahrami H, Cheour M. Internet addiction is associated with psychological distress in highly schizotypal students. Early Interv Psychiatry 2023;17(7):681-691. [doi: 10.1111/eip.13362] [Medline: 36196053]
20. Paquin V, Ferrari M, Rej S, Boivin M, Ouellet-Morin I, Geoffroy MC, et al. Trajectories of adolescent media use and their associations with psychotic experiences. JAMA Psychiatry 2024;81(7):708-716. [doi: 10.1001/jamapsychiatry.2024.0384] [Medline: 38598210]
21. Hill K. They asked an A.I. chatbot questions. The answers sent them spiraling. The New York Times. 2025. URL: https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html [accessed 2025-09-16]
22. Hill K, Freedman D. Chatbots can go into a delusional spiral. Here's how it happens. The New York Times. 2025. URL: https://www.nytimes.com/2025/08/08/technology/ai-chatbots-delusions-chatgpt.html [accessed 2025-09-16]
23. Loewy RL, Pearson R, Vinogradov S, Bearden CE, Cannon TD. Psychosis risk screening with the prodromal questionnaire--brief version (PQ-B). Schizophr Res 2011;129(1):42-46 [FREE Full text] [doi: 10.1016/j.schres.2011.03.029] [Medline: 21511440]
24. Savill M, D'Ambrosio J, Cannon TD, Loewy RL. Psychosis risk screening in different populations using the Prodromal Questionnaire: a systematic review. Early Interv Psychiatry 2018;12(1):3-14 [FREE Full text] [doi: 10.1111/eip.12446] [Medline: 28782283]
25. Savill M, Loewy RL, Niendam TA, Porteus AJ, Rosenthal A, Gobrial S, et al. The diagnostic accuracy of screening for psychosis spectrum disorders in behavioral health clinics integrated into primary care. Schizophr Res 2024;266:190-196 [FREE Full text] [doi: 10.1016/j.schres.2024.02.007] [Medline: 38422889]
26. Savill M, Nguyen T, Shim RS, Loewy RL. Online psychosis screening: characterizing an underexamined population to improve access and equity. Psychiatr Serv 2022;73(9):1005-1012. [doi: 10.1176/appi.ps.202100257] [Medline: 35172594]
27. Maheux A, Maes C, Buck B. GenAI in the lives of young adults: exploring motivations and mental health. Under review 2025 2025:E.
28. Pappa E, Baah F, Lynch J, Shiel L, Blackman G, Raihani N, et al. Delusional themes are more varied than previously assumed: a comprehensive systematic review and meta-analysis. Schizophr Bull 2025;51(3):637-645. [doi: 10.1093/schbul/sbae225] [Medline: 39847500]
29. Narita Z, Ando S, Yamasaki S, Miyashita M, DeVylder J, Yamaguchi S, et al. Association of problematic internet use with psychotic experiences and depression in adolescents: a cohort study. Schizophr Bull 2024;51(1):198-207. [doi: 10.1093/schbul/sbae089] [Medline: 38825582]
30. Lee JY, Ban D, Kim SY, Kim JM, Shin IS, Yoon JS, et al. Negative life events and problematic internet use as factors associated with psychotic-like experiences in adolescents. Front Psychiatry 2019;10:369 [FREE Full text] [doi: 10.3389/fpsyt.2019.00369] [Medline: 31191372]

XSL•FO
RenderX

31. Paquin V, Ackerman RA, Depp CA, Moore RC, Harvey PD, Pinkham AE. Media use and its associations with paranoia in schizophrenia and bipolar disorder: ecological momentary assessment. JMIR Ment Health 2024;11:e59198 [FREE Full text] [doi: 10.2196/59198] [Medline: 38967418]
32. Paquin V, Philippe FL, Shannon H, Guimond S, Ouellet-Morin I, Geoffroy MC. Associations between digital media use and psychotic experiences in young adults of Quebec, Canada: a longitudinal study. Soc Psychiatry Psychiatr Epidemiol 2024;59(1):65-75. [doi: 10.1007/s00127-023-02537-6] [Medline: 37516683]
33. Mestre-Bach G, Paiva U, San Martín Iniguez L, Beranuy M, Martín-Vivar M, Mallorquí-Bagué N, et al. The association between internet-use-disorder symptoms and loneliness: a systematic review and meta-analysis with a categorical approach. Psychol Med 2025;55:e77. [doi: 10.1017/S0033291725000376] [Medline: 40071718]
34. Amendola S, Cerutti R, Presaghi F. Symptoms of prolonged social withdrawal, problematic internet use, and psychotic-like experiences in emerging adulthood: a moderation model. Clin Neuropsychiatry 2023;20(1):29-38 [FREE Full text] [doi: 10.36131/cnfioritieditore20230104] [Medline: 36936621]
35. Salles A, Evers K, Farisco M. Anthropomorphism in AI. AJOB Neurosci 2020;11(2):88-95 [FREE Full text] [doi: 10.1080/21507740.2020.1740350] [Medline: 32228388]
36. Folk DP, Dunn EW. Turning to AI for companionship predicts increased loneliness four months later. OSF 2025. [doi: 10.31234/osf.io/bq7v3_v1]

Abbreviations

AI: artificial intelligence
AIMUS: Artificial Intelligence Motivation and Uses Scale
GAATES: Generative AI Aberrant Thoughts and Experiences Scale
GenAI: generative artificial intelligence
OR: odds ratio
PQ-B: Prodromal Questionnaire, Brief Version

Edited by A Stone; submitted 30.Sep.2025; peer-reviewed by H Maheshwari, LR Guo, A Mettupatti Sivagnanam, J Sheffield, S Mohanadas, A Quansah, O Oyetunji; comments to author 27.Oct.2025; revised version received 20.Jan.2026; accepted 20.Jan.2026; published 05.Mar.2026.

Please cite as:
Buck B, Maheux AJ
Psychosis Risk and Generative Artificial Intelligence Use Frequency, Motivations, and Delusion-Like Experiences: Cross-Sectional Survey Study
J Med Internet Res 2026;28:e85038
URL: https://www.jmir.org/2026/1/e85038
doi:10.2196/85038
PMID:

©Benjamin Buck, Anne Julia Maheux. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Scalable and Robust Artificial Intelligence for Spine Alignment Assessment: Multicenter Study Enabled by Real-Time Data Transformation

Guilin Chen[1*], MD; Nan Meng[2*], PhD; Yipeng Zhuang[2*], PhD; Zhe Chen[3*], MD; Zhen Bian[4*], MD; Zhaoyang Gong[5*], MD; Jiawei Shi[6*], MD; Tao Huang[2], BEng; Xihe Kuang[2], PhD; Pengyu Lu[2], MSc; Cong Nie[2], MD; Qifeng Yu[5], MD; Zefu Chen[6], MD; Hui Jiang[6], MD; Zhongmin Zhang[6], MD; Chaojun Zheng[5], MD; Yu Liang[3], MD; Nan Wu[1], MD; Jason Pui Yin Cheung[2], MS, MD; Jianguo Zhang[1], MD; Teng Zhang[2], PhD

[1]Department of Orthopaedic Surgery, State Key Laboratory of Complex Severe and Rare Diseases, Beijing Key of Big Data Innovation and Application for Skeletal Health Medical Care, Key Laboratory of Big Data for Spinal Deformities, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, 100730, China

[2]Department of Orthopaedics and Traumatology, Li Ka Shing Faculty of Medicine, The University of Hong Kong, HONG KONG SAR, China (Hong Kong)

[3]Department of Orthopaedics, Shanghai Jiaotong University School of Medicine, Ruijin Hospital, Shanghai, 200025, China

[4]Department of Paediatric Orthopaedics, Beijing Jishuitan Hospital, Beijing, 100035, Beijing, China

[5]Department of Orthopaedics, Huashan Hospital, Fudan University, Shanghai, 200040, China

[6]Department of Orthopaedics, Division of Spine Surgery, Nanfang Hospital, Southern Medical University, Guangzhou, Guangdong, 510515, China

[*]these authors contributed equally

Corresponding Author:
Teng Zhang, PhD
Department of Orthopaedics and Traumatology
Li Ka Shing Faculty of Medicine
The University of Hong Kong
HONG KONG SAR,
China (Hong Kong)
Phone: 852 68868531
Email: tgzhang@hku.hk

Abstract

Background: Artificial intelligence (AI) has shown promise for automating spinal alignment assessment in adolescent idiopathic scoliosis (AIS). However, AI models typically exhibit reduced accuracy and robustness when deployed across multiple medical centers due to variability in imaging protocols and data characteristics, potentially compromising clinical diagnosis and treatment decisions.

Objective: This study aimed to develop a real-time, plug-and-play data transformation method to enhance the robustness of deep learning models against data heterogeneity in radiographs, thereby improving their performance in assessing AIS across multiple medical centers.

Methods: In this retrospective multicenter study, 3899 full-spine radiographs from 7 hospitals (2 from Hong Kong and 5 from Mainland China), collected between January 2012 and August 2024, were included. Data from 2 hospitals in Hong Kong (n=3034) were used for model training and internal validation, while radiographs from the 5 mainland hospitals (n=865) formed 5 independent external validation datasets. A novel pixel intensity–based data transformation method was developed to standardize image contrast and brightness across datasets and integrated into the model training process to enhance our previously developed AI model, SpineHRNet+. The enhanced model's accuracy and robustness for cobb angle (CA) prediction and severity classification were evaluated using both internal and external datasets. Data heterogeneity across centers was quantified by brightness and contrast differences. CA prediction accuracy was evaluated using residual analysis, linear regression (coefficient of determination [R^2]), and Bland-Altman analyses. Model performance for disease severity classification was assessed using sensitivity, specificity, precision, negative predictive value, accuracy, and confusion matrix analysis. The transformation method aligns pixel intensity distributions across datasets using statistical profiling and optimization, ensuring consistent image characteristics while preserving anatomical integrity.

XSL•FO
RenderX

Results: The developed data transformation method significantly reduced contrast variability between datasets, improving consistency in image characteristics and enabling more reliable AI analysis. The enhanced SpineHRNet+ achieved consistent and accurate CA predictions across external validation datasets, with mean prediction errors within 4° (SD 3.12°), and maintained an R^2 greater than 0.90 for all centers. The sensitivity and negative predictive value for disease severity grading improved to 90.18% and 93.16%, respectively. Bland-Altman analyses demonstrated robust agreement, with 95% limits of agreement within 7.51° across all datasets.

Conclusions: The proposed data transformation approach effectively addressed data heterogeneity, significantly improving the accuracy and robustness of SpineHRNet+ in multicenter AIS assessments. The real-time processing capability and preservation of anatomical integrity underscore the method's clinical practicality, enabling scalable and reliable AI applications in diverse health care environments.

(J Med Internet Res 2026;28:e78396) doi:10.2196/78396

KEYWORDS

adolescent idiopathic scoliosis; multicentre validation; data heterogeneity; artificial intelligence; data transformation

Introduction

Adolescent idiopathic scoliosis (AIS) is one of the most common spinal deformities in adolescents, affecting approximately 2.2% of boys and 4.8% of girls globally [1,2]. Without timely diagnosis and intervention, AIS may progress rapidly to severe spinal curvatures [3], potentially leading to cosmetic concerns, back pain, and reduced cardiopulmonary function [4]. The Cobb angle (CA), a measurement derived from spinal radiographs, is considered the gold standard for quantitatively assessing scoliosis severity and guiding clinical decision-making such as bracing or surgical intervention [5]. However, accurate measurement of the CA is inherently laborious and requires considerable clinical expertise, often exhibiting inter- and intra-observer variability ranging between 4.9° to 10° [6,7]. This variability is particularly pronounced in multicenter settings, given differences in patient positioning, clinical protocols, and assessment practices across institutions [8,9]. Despite promising results in single-center studies, existing AI models lack mechanisms to adapt to intercenter variability, highlighting a critical gap in robust multicenter deployment. Unlike prior domain adaptation or generative approaches that often require paired data or introduce distortions, our method offers a novel unsupervised pixel-level transformation that preserves anatomical integrity while standardizing image characteristics.

Advances in digital technologies, particularly artificial intelligence (AI), have shown promising potential to improve the consistency and efficiency of AIS assessment [10-13]. Recent AI-based models have demonstrated high reliability and accuracy in CA prediction and severity grading when validated on single-center datasets [14,15]. Despite these successes, widespread clinical adoption remains challenging due to heterogeneity in clinical data collected across different health care facilities [16]. Variability in spinal radiograph quality, brightness, contrast, and pixel intensity distributions, stemming from differences in the medical equipment, acquisition protocols, and post-processing methods, substantially hampers the generalizability of AI models when applied across external institutions [17-19]. These discrepancies limit the real-world applicability and use of AI models, especially in resource-constrained clinical environments, where the absence of uniform standards and variations in imaging practices further exacerbate the performance degradation of AI diagnostic tools. Additionally, spine surgeons commonly capture radiographs informally, either using smartphone cameras or taking screenshots directly from computer monitors during routine clinical practice. Such practices markedly degrade image quality and exacerbate inaccuracies in current AI-based automated analysis methods, undermining the reliability of spinal alignment evaluations. These practical challenges highlight an urgent need for methods capable of adapting to variations in image quality and acquisition techniques, ensuring robust and dependable clinical assessments across diverse health care settings.

To address the limitations of existing AI models in robustness and generalizability, this study introduces a plug-and-play data transformation approach harmonizing characteristics at the pixel-level to effectively mitigate differences among multicenter datasets. The system operates without calibration using a default reference profile, but performance improves with minimal calibration. This computationally efficient approach can be easily integrated into the training or fine-tuning processes of existing AI models, significantly enhancing their robustness and stability in handling data from diverse sources. Using this method, we enhanced SpineHRNet+, an AI model originally trained on a single-center dataset, and evaluated its accuracy and robustness in predicting the CA through internal validation and 5 independent external datasets. By improving model performance across heterogeneous data with minimal computational overhead and data requirements, this research paves the way for reliable clinical implementation of AI-assisted AIS diagnosis and assessment across diverse clinical settings. Previous efforts to address data heterogeneity have included statistical harmonization techniques, such as histogram equalization, intensity normalization, and feature distribution alignment [16,17]. Domain adaptation methods like covariance matching and structure-preserving projection have also been explored to reduce intercenter variability. While these approaches offer theoretical robustness, many require extensive preprocessing, paired data, or high computational resources, limiting their scalability in clinical environments. In contrast, our method uses unsupervised pixel intensity alignment using histogram-based statistical profiling, offering a lightweight,

XSL•FO
RenderX

anatomically preserving, and real-time solution suitable for multicenter deployment.

Methods

Study Design and Participants

Deidentified posteroanterior whole-spine radiographs of AIS patients (inclusion and exclusion criteria detailed in Section 2 in Multimedia Appendix 1) were gathered from 5 tertiary hospitals (n=865) in Mainland China and 2 hospitals in Hong Kong China (n=3034), spanning from January 1, 2012, to August 8, 2024. Demographic data, including age, sex, height, weight, and BMI, as well as the radiograph acquisition protocols, were extracted from medical records at each center. Radiographs from Queen Mary Hospital (QMH) and DKCH (**Duchess of Kent Children's Hospital at Sandy Bay**; QMH&DKCH cohort) were specifically used for model training and internal validation: 2770 patients were randomly selected for the training set, and 264 patients were selected using a stratified hold-out approach to ensure consistent distribution of severity and curve types. Data from the remaining 5 hospitals formed independent external validation datasets: Peking Union Medical College Hospital (PUMCH) cohort, First Affiliated Hospital, Zhejiang University School of Medicine (FAH) cohort, Qilu Hospital of Shandong University (QLH) cohort, Shandong Provincial Hospital (SDPH) cohort, and Shanghai Huashan Hospital (HSH) cohort. Radiographs from QMH and DKCH were used for model training and internal validation.

Data Collection and Preparation

Radiographs were acquired using site-specific imaging setups and workflows. Specifically, QMH&DKCH and SDPH used EOS systems (EOSedge at QMH&DKCH; EOS Imaging System at RJH) for biplanar whole-spine radiographs. PUMCH used a Philips DigitalDiagnost system, sequentially capturing images of different spinal regions and stitching them together to form whole-spine radiographs. FAH and HSH used GE health care system (Discovery XR656 at FAH and Optima XR646 HD at HSH) to acquire posteroanterior whole-spine radiographs, while QLH used multiple imaging devices, such as the KODAK DIRECTVIEW DR 7500, Canon DR, CARESTREAM DRX-EVOLUTION, and EOS systems. Notable variability in critical imaging parameters, such as kilovolt peak, exposure time, and area dose product, further highlighted dataset heterogeneity (Table S1 in Multimedia Appendix 1). Detailed imaging protocols are provided in Section 1 in Multimedia Appendix 1.

All radiographs were annotated according to standardized procedures using the AlignProCARE system, marking 72 vertebral endplate landmarks from the seventh cervical vertebra (C7) to the fifth lumbar vertebra (L5). At QMH&DKCH, landmarks were initially annotated by 2 junior clinicians with at least 2 years of clinical experience, and subsequently verified by a senior spine specialist with more than 20 years of experience. To ensure annotation consistency across centers, external annotations were cross-checked by senior spine specialists, and a subset of 10% randomly selected radiographs from each center were reannotated and reviewed by an independent expert panel to verify intercenter consistency. The annotated landmarks served as ground truth (GT) for model training and evaluation.

The coronal CA was used as the primary metric for evaluating spinal alignment. Deformities with CA exceeding 40° were deemed severe, those ranging from 20° to 40° were labeled as moderate, and angles from 0° to 20° were identified as normal-mild. The criteria for assessing the severity of spinal deformities and their corresponding clinical interventions are detailed in Table S2 in Multimedia Appendix 1. The classification of curve types was determined by the location of the curve apex: thoracic (T) curves were identified when the apex was situated between the T1 and T11 vertebrae, while T or TL curves were recognized if the apex fell between the T12 and the L5 vertebrae.

Model Development and Multicenter Validation

We used SpineHRNet+ [15], a previously validated deep learning model, and enhanced it through the proposed data transformation method. The overall workflow consisted of 2 phases: model development (with data transformation integration) and multicenter validation (Figure 1). During model development, we applied the data transformation method to the QMH&DKCH training set, creating simulated radiographs that mimic pixel-level characteristics from external centers. SpineHRNet+ was trained and fine-tuned using both original and transformed datasets to improve generalizability. Subsequently, the enhanced model was deployed via the AlignProCARE system for clinical evaluation across multiple institutions. During external validation, each medical center uploaded radiographs through locally installed AlignProCARE software or applications; images were processed remotely by the enhanced model, and results were returned to the respective centers for clinical analysis.

XSL•FO
RenderX

Figure 1. The workflow diagram of model development and real-world application. In model development stage, radiographs collected from 2 local clinics (QMH and DKCH) were first augmented by the proposed data transformation method and then used for model training. Subsequently, the well-trained model was deployed on our online platform mskalign AlignProCARE. In the real-world application scenario, X-rays obtained from 5 external centres were used as external testing data and the ones collected from QMH and DKCH were used as internal validation data to comprehensively evaluate the performance of our model. DKCH: Duchess of Kent Children's Hospital; FAH: First Affiliated Hospital; HSH: Shanghai Huashan Hospital; PUMCH: Peking Union Medical College Hospital; QLH: Qilu Hospital of Shandong University; QMH: Queen Mary hospital; SDPH: Shandong Provincial Hospital.

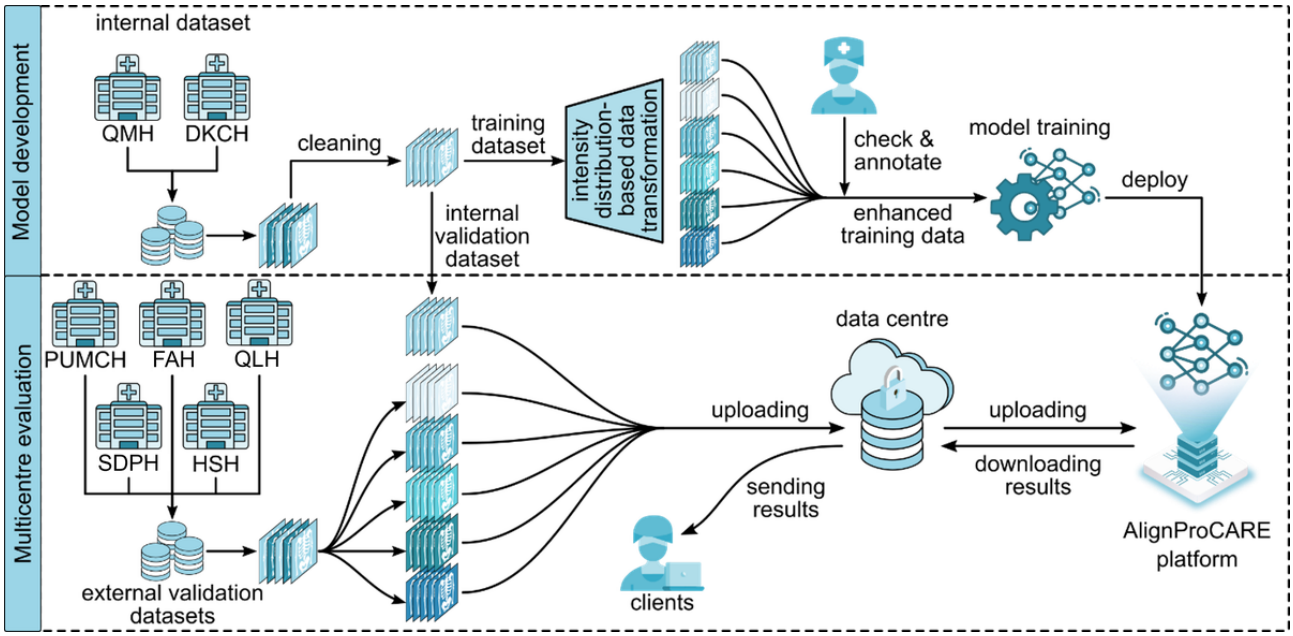

To preserve transformation diversity, reference images from each external center were randomly sampled during augmentation. Each training image was transformed multiple times using different reference distributions, generating a diverse set of simulated radiographs. Augmented samples were randomly mixed with original images during training, without weighting, to ensure balanced exposure and prevent overfitting to any specific external style.

Intensity Distribution–Based Data Transformation

To address data heterogeneity in clinical data collected from different institutions, we developed a data transformation approach to standardize differences observed across patient spinal radiographs from multiple centers. Our method works by analyzing and aligning the quantitative characteristics, such as brightness and contrast, present in clinical images from different hospitals. Specifically, we first calculated detailed statistical profiles representing the distribution of intensity values for each institution. Then, using an optimization technique, we systematically adjusted each patient's data to align closely with a reference standard.

The transformation function is based on histogram matching, where the cumulative distribution function (CDF) of the source image is aligned to that of a target image. The optimization objective minimizes the absolute difference between the CDFs:

for each intensity level i, where F^S and F^T are the CDFs of the source and target images, respectively. The mapping is monotonic and preserves anatomical integrity by avoiding spatial distortion. Detailed pseudocode is provided in Algorithm S1 in Multimedia Appendix 1. Figure 2 illustrate the transformation workflow. This alignment process was carefully designed to preserve essential clinical and anatomical details, ensuring that no critical diagnostic information was altered or lost (Algorithm S1 and Figure S1 in Multimedia Appendix 1). Brightness differences are reported in normalized pixel intensity units ranging from 0 to 255. Histogram spread refers to the SD of pixel intensity values across the image, indicating contrast variability.

XSL•FO
RenderX

Figure 2. Visual results of intensity distribution-based data transform algorithm. CDF: cumulative distribution function.

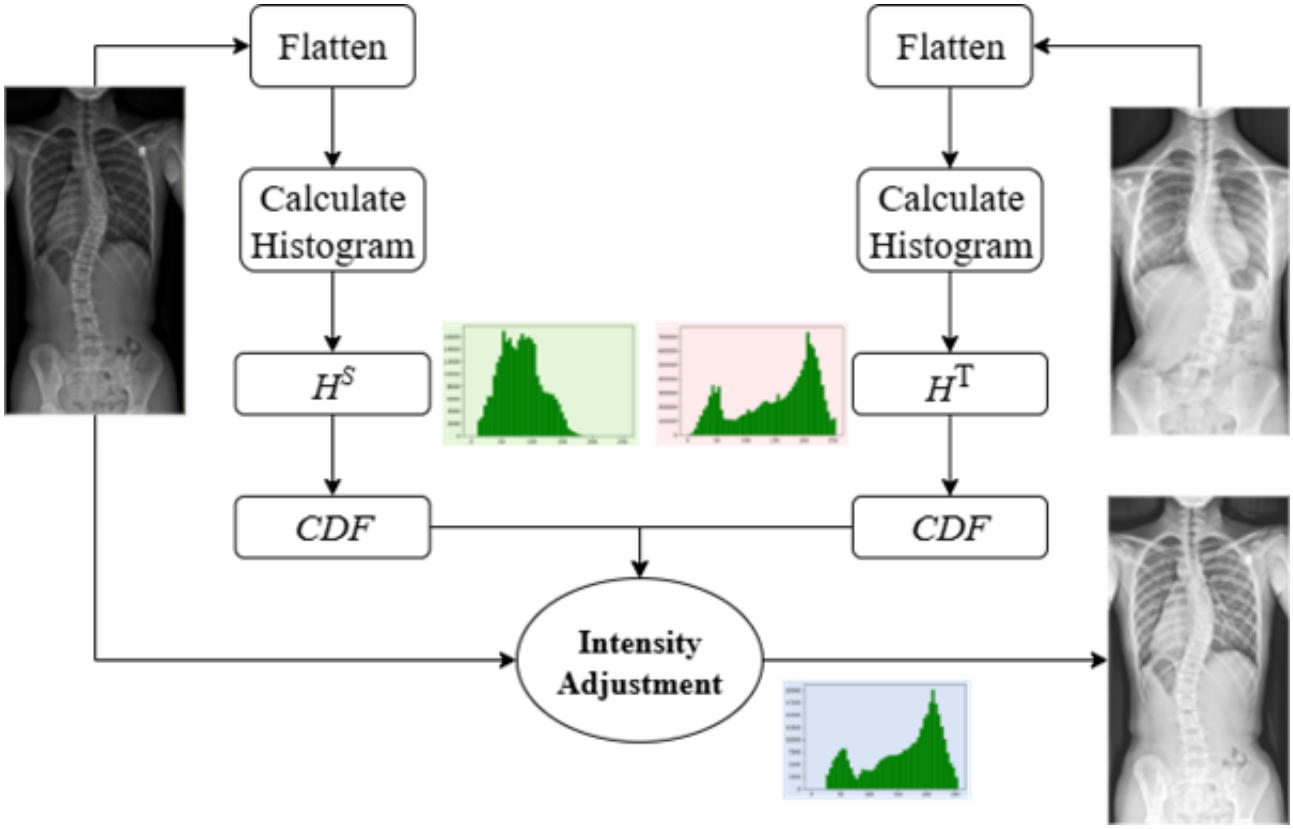

Conceptually, the algorithm computes the histogram of pixel intensities for each image and aligns it to a reference distribution using a monotonic mapping function. This mapping is derived via histogram matching and optimized to minimize brightness and contrast discrepancies while preserving anatomical structures. The transformation is applied in real-time and does not require paired training data. Unlike conventional histogram matching applied as a preprocessing step, our method dynamically samples reference distributions from multiple centers and integrates transformation into model training, enabling real-time adaptation and improved robustness. The procedure is computationally efficient, allowing rapid integration into routine clinical workflows, even when handling large numbers of patient records (computational complexity $O(n \log n)$; Multimedia Appendix 1 Section 3.3). For simplicity, the AI model incorporating this transformation technique is termed the "enhanced model" in subsequent sections.

Statistical Analysis

To evaluate the accuracy of CA predictions, we calculated the mean (SD) of the absolute differences between predicted and GT values. The distribution of these prediction errors was assessed using descriptive statistics, including the first quartile (Q1), median, third quartile (Q3), SD, and IQR. Linear regression analysis was used to assess the correlation between predicted and GT values, quantified by the coefficient of determination (R^2). Bland-Altman analysis was used to measure the agreement between model predictions and GT, reporting the 95% limits of agreement (LoA). Paired *t* tests were used to compare the absolute prediction errors of the 2 models on the same set of images, controlling for patient-specific anatomical variability.

Performance in disease severity grading and spinal curve identification was evaluated using confusion matrices. Sensitivity, specificity, precision, negative predictive value (NPV), and accuracy were calculated, each with a corresponding 95% CI.

All statistical analyses were conducted using Python (v3.8) and several related packages, including NumPy [20] (v1.18.5), SciPy [21] (v1.5.2), Ptitprince [22] (v0.2.6), pandas [23] (v1.1.3), seaborn [24] (v0.11.0), and Matplotlib [25] (v3.3.2).

Ethical Considerations

This multicenter retrospective cohort study adhered to the Declaration of Helsinki and the Standards for Reporting of Diagnostic Accuracy Studies reporting guidelines and was registered at website of Clinical Trials (Registration No. NCT06711757). Ethical approval was obtained from the Institutional Review Boards (IRBs) of all participating medical centers, including QMH and DKCH (IRB No. UW22-270); PUMCH (IRB No. IRB-B473); FAH (IRB No. IIT20230248B-R1); QLH (IRB No. QLCR20240818); SDPH (IRB No. AF-SW-370-001); and HSH (IRB No. KY2022-570). As this is a retrospective study using previously collected and anonymized patient data, the IRB has waived the requirement for informed consent. No compensation was provided to participants, as no prospective recruitment or study visits were conducted. All patient data used in this study were deidentified prior to analysis. No identifiable personal information was accessed or stored, ensuring full compliance with privacy and confidentiality standards.

XSL•FO
RenderX

Results

Overview

All eligible patients with AIS from QMH and DKCH recruited between December 2019 and August 2024 were included, forming a training cohort of 2770 participants (mean 14, SD 1.9 years; 75.3% female; Table 1). An additional 264 consecutive AIS patients recruited from September 2023 to March 2024 served as an internal validation cohort (mean 16.7, SD 1.2 years; 78% female; Table 1). Overall, the internal dataset (n=3034) included 910 patients with normal-mild deformities, 1,624 with moderate deformities, and 236 with severe deformities (Table 1).

Table 1. Demographic information of cohorts from different centers.

Demographic information	Training and internal validation datasets (n=3034)		External validation datasets (n=865)				
	Training dataset	Internal validation dataset	External validation dataset 1 (PUMCH[a])	External validation dataset 2 (FAH[b])	External validation dataset 3 (QLH[c])	External validation dataset 4 (SDPH[d])	External validation dataset 5 (HSH[e])
Patients	2770	264	314	105	134	136	176
Mean age (SD), years	14 (1.9)	16.7 (1.2)	14.8 (2.8)	15.9 (2.7)	14 (2.6)	14.3 (2.8)	15.4 (3.4)
Mean max CA[f] (SD), degree	25 (10.8)	24.5 (9.1)	46.6 (13)	18.7 (10.2)	28.2 (14.3)	20.4 (12.2)	25.1 (16)
Mean BMI (SD), kg/m^2	18.9 (14.8)	18.7 (3.3)	18.4 (3.3)	N/A[g]	N/A	N/A	22.6 (3.7)
Sex, n (%)							
Female	2085 (75.3)	206 (78)	277 (88.2)	74 (70.5)	81 (60.4)	97 (71.3)	109 (61.9)
Male	685 (24.7)	58 (22)	37 (11.8)	31 (29.5)	53 (39.6)	39 (28.7)	67 (38.1)
Curve type, n							
Thoracic	2497	244	287	76	45	48	86
Thoracolumbar or lumbar	2045	193	283	41	111	116	129
Severity, n							
Normal-mild	910	100	4	56	41	75	79
Moderate	1624	147	90	46	72	52	61
Severe	236	17	220	3	21	9	36

[a]PUMCH: Peking Union Medical College Hospital.
[b]FAH: First Affiliated Hospital, Zhejiang University School of Medicine.
[c]QLH: Qilu Hospital of Shandong University.
[d]SDPH: Shandong Provincial Hospital.
[e]HSH: Shanghai Huashan Hospital.
[f]CA: cobb angle.
[g]N/A: not available.

In total, 5 independent external cohorts comprising a total of 865 patients were recruited from the following hospitals: PUMCH (n=314; mean 14.8, SD 2.8 years; 88.2% female), FAH (n=105; mean 15.9, SD 2.7 years; 70.5% female), QLH (n=134; mean 14, SD 2.6 years; 60.4% female), SDPH (n=136; mean 14.3, 2.8 years; 71.3% female), and HSH (n=176; mean 15.4, SD 3.4 years; 61.9% female). Disease severity across these external datasets comprised 255 normal-mild cases, 321 moderate cases, and 289 severe cases (Table 1). Center-specific equipment and protocols were used when collecting each external dataset (Table S1 in Multimedia Appendix 1). The initial differences in histogram spread and brightness values between the internal and external datasets were 9.6×10-4 and 8.71, respectively; after applying the data transformation algorithm, these differences were reduced to 0.4×10-4 and 0.47 (Table S3 in Multimedia Appendix 1).

CA Prediction

Table 2 summarizes the CA prediction performance of both the original and enhanced models. In the internal dataset (QMH&DKCH cohort), the enhanced model improved prediction accuracy, reducing the mean absolute error by 0.13° (SD 0.27°) for thoracic curves (enhanced: 2.90°, SD 1.95° vs original: 3.03°, SD 2.22°]), and by 0.21°, SD 0.28°) for thoracolumbar or lumbar curves (enhanced: 2.54°, SD 1.86° vs original: 2.75°, SD 2.14°). Across all independent external datasets, the enhanced model reduced the mean absolute prediction error by 1.35° (SD 1.72°) in thoracic curves

XSL•FO
RenderX

(enhanced: 4.03°, SD 3.12° vs original: 5.38°, SD 4.83°), and by 1.45° (SD 1.74°) in thoracolumbar or lumbar curves (enhanced: 3.75°, SD 2.94° vs original: 5.21°, SD 4.68°). To assess the significance of performance improvements, paired *t* tests were conducted comparing the prediction errors of the original and enhanced models across all datasets. The results confirmed statistically significant improvements ($P<.05$) across all external cohorts, except for PUMCH's thoracic group, supporting the robustness of our approach.

Table 2. Quantitative comparison of absolute error in cobb angle prediction between original and enhanced models on internal and external validation datasets.

Cohorts	QMH[a] and DKCH[b]	PUMCH[c]	FAH[d]	QLH[e]	SDPH[f]	HSH[g]	Overall (external)
Thoracic							
Cobb angle difference (degree)							
Original model, mean (SD)	3.03 (2.22)	4.08 (3.25)	5.35 (3.73)	5.46 (4.57)	6.88 (7.88)	5.15 (4.73)	5.38 (4.83)
Enhanced model, mean (SD)	2.90 (1.95)	3.73 (2.36)	3.77 (2.71)	3.84 (2.76)	4.53 (4.66)	4.30 (3.09)	4.03 (3.12)
Improvement (degree), mean (SD)	0.13 (0.27)	0.35 (0.89)	1.58 (1.02)	1.62 (1.81)	2.35 (3.22)	0.85 (1.64)	1.35 (1.72)
P value	.46	.21	<.01	<.01	<.01	.03	—[h]
Thoracolumbar or lumbar							
Cobb angle difference (degree)							
Original model, mean (SD)	2.75 (2.14)	4.41 (3.81)	5.63 (5.42)	5.72 (4.56)	5.52 (4.78)	4.75 (4.82)	5.21 (4.68)
Enhanced model, mean (SD)	2.54 (1.86)	2.76 (2.08)	4.30 (3.31)	3.81 (2.97)	4.10 (2.93)	3.80 (3.40)	3.75 (2.94)
Improvement (degree), mean (SD)	0.21 (0.28)	1.65 (1.73)	1.33 (2.11)	1.91 (1.59)	1.42 (1.85)	0.95 (1.42)	1.45 (1.74)
P value	.19	<.01	<.01	<.01	<.01	.04	—

[a]QMH: Queen Mary Hospital.
[b]DKCH: Duchess of Kent Children's Hospital at Sandy Bay.
[c]PUMCH: Peking Union Medical College Hospital.
[d]FAH: First Affiliated Hospital, Zhejiang University School of Medicine.
[e]QLH: Qilu Hospital of Shandong University.
[f]SDPH: Shandong Provincial Hospital.
[g]HSH: Shanghai Huashan Hospital.
[h]Not available.

CA prediction error distributions (Table S4 in Multimedia Appendix 1) demonstrated that in the internal validation dataset, the SD of prediction errors for T curves was reduced by 0.97, and the IQR decreased by 1.93 when using the enhanced model. For thoracolumbar or lumbar curves, the SD was reduced by 1.17, and the IQR decreased by 1.77. Across all independent external datasets, the average SD for T curves decreased by 1.31, and the IQR by 1.17, while for thoracolumbar or lumbar curves, the average SD was reduced by 1.47, and the IQR by 0.80.

Linear regression results of enhanced models for maximum CA prediction are shown in Figure 3. In all cohorts, the enhanced model achieved a coefficient of determination (R^2) exceeding 0.90, peaking at 0.93 on the internal dataset. Specifically, R^2 improved by 0.02 in the internal dataset (enhanced model: 0.93 vs original model: 0.91). Among the independent external datasets, the enhanced model's R^2 values increased by 0.11 on average relative to the original model, specifically by 0.05 at PUMCH, 0.19 at FAH, 0.10 at QLH, 0.09 at SDPH, and 0.09 at HSH (Figure S2). Bland-Altman analysis was performed to quantify the agreement between predicted and true CAs (Figure S3). Compared to the original model, the enhanced model exhibited narrower 95% LoA intervals across all datasets (QMH&DKCH: 2.24°; PUMCH: 6.16°; FAH: 8.26°; QLH: 9.32°; SDPH: 7.89°; and HSH: 5.93°).

XSL•FO
RenderX

Figure 3. The enhanced model's linear regression analysis for maximum CA prediction on the internal validation dataset and 5 external validation datasets. The x-axis represents the values predicted by our model, while the y-axis denotes the GT CAs derived from manually annotated landmarks. DKCH: Duchess of Kent Children's Hospital; FAH: First Affiliated Hospital; HSH: Shanghai Huashan Hospital; PUMCH: Peking Union Medical College Hospital; QLH: Qilu Hospital of Shandong University; QMH: Queen Mary hospital; SDPH: Shandong Provincial Hospital.

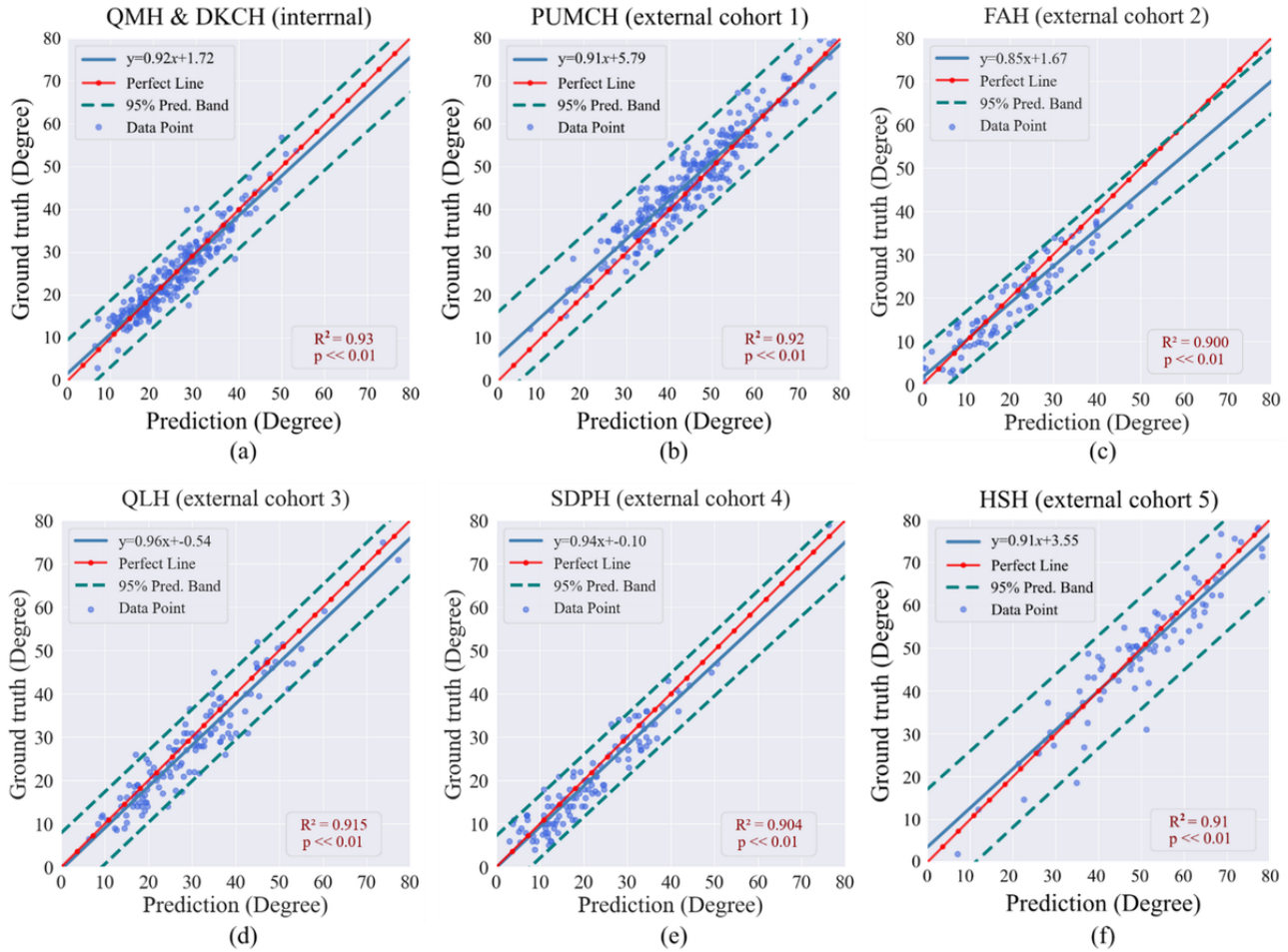

Disease Severity Grading and Curve Type Identification

Table 3 presents the quantitative performance of the enhanced model in disease severity grading and spinal curve type identification. In the internal validation dataset, the enhanced model achieved an average (macro-average) 3-class sensitivity of 89.97% and a NPV of 94.49%, representing an improvement of 3.85% and 1.74%, respectively, compared to the original model (Table S5 and Figure S4 in Multimedia Appendix 1). Across 5 independent external datasets, the enhanced model attained an average sensitivity of 90.18% and an average NPV of 93.16%, representing substantial improvements of 12.57% and 2.86%, respectively, over the original model (average sensitivity: 77.61% and average NPV: 90.30%).

Regarding spinal curve type classification, the enhanced model slightly improved sensitivity to 99.24% on the internal dataset compared to 98.79% with the original model. However, performance on independent external datasets varied across centers, with both improvements and declines observed (detailed data in Tables S3 and S5 in Multimedia Appendix 1).

XSL•FO
RenderX

Table 3. Quantitative results on severity grading and curve type detection using enhanced model.

Evaluation metrics	Severity level				Curve type		
	Normal-mild (0°-20°), coefficient (95% CI)	Moderate (21°-40°), coefficient (95% CI)	Severe (>41°), coefficient (95% CI)	Macro Average	Thoracic, coefficient (95% CI)	Thoracolumbar or lumbar, coefficient (95% CI)	Macro Average
Internal validation dataset (QMH&DKCH[a] cohort)							
Sensitivity	89.80 (82.23-94.36)	92.62 (87.26-95.83)	87.50 (63.98-96.50)	89.97	98.48 (95.64-99.48)	100 (94.87-100)	99.24
Specificity	95.15 (90.73-97.52)	89.47 (82.50-93.88)	98.79 (96.49-99.59)	94.47	100 (94.42-100)	97.40 (94.05-98.88)	98.70
Precision	91.67 (84.41-95.72)	92 (86.54-95.36)	82.35 (58.97-93.81)	88.67	100 (98.07-100)	93.42 (85.51-97.16)	96.71
NPV[b]	94.01 (89.33-96.72)	90.27 (83.41-94.48)	99.19 (97.08-99.78)	94.49	95.59 (87.81-98.49)	100 (97.99-100)	97.79
Accuracy	93.16 (89.44-95.63)	91.25 (87.22-94.10)	98.10 (95.63-99.19)	94.17	98.86 (96.70-99.61)	98.10 (95.63-99.19)	98.48
External validation dataset 1 (PUMCH[c] cohort)							
Sensitivity	100 (51.01-100)	87.78 (79.43-93.04)	86.82 (81.71-90.66)	91.53	97.91 (95.51-99.04)	100 (88.97-100)	98.95
Specificity	99.03 (97.19-99.67)	87.05 (82.03-90.83)	91.49 (84.10-95.62)	92.53	100 (87.54-100)	99.29 (97.46-99.81)	99.65
Precision	57.14 (25.05-84.18)	73.15 (64.10-80.61)	95.98 (92.27-97.95)	75.42	100 (98.65-100)	93.94 (80.39-98.32)	96.97
NPV	100 (98.76-100)	94.66 (90.69-96.99)	74.78 (66.13-81.83)	89.81	81.82 (65.61-91.39)	100 (98.65-100)	90.91
Accuracy	99.04 (97.23-99.67)	87.26 (83.12-90.50)	88.22 (84.18-91.33)	91.51	98.09 (95.89-99.12)	99.36 (97.71-99.83)	98.73
External validation dataset 2 (FAH[d] cohort)							
Sensitivity	87.27 (75.98-93.70)	91.11 (79.27-96.49)	100 (43.85-100)	95.56	93.22 (83.82-97.33)	78.57 (60.46-89.79)	85.90
Specificity	93.75 (83.16-97.85)	87.93 (77.12-94.03)	99 (94.55-99.82)	93.47	93.18 (81.77-97.65)	92 (83.63-96.28)	92.59
Precision	94.12 (84.08-97.98)	85.42 (72.83-92.75)	75 (30.06-95.44)	80.21	94.83 (85.86-98.23)	78.57 (60.46-89.79)	86.70
NPV	86.54 (74.73-93.32)	92.73 (82.74-97.14)	100 (96.26-100)	96.36	91.11 (79.27-96.49)	92 (83.63-96.28)	91.56
Accuracy	90.29 (83.04-94.64)	89.32 (81.88-93.93)	99.03 (94.70-99.83)	94.17	93.20 (86.63-96.67)	88.35 (80.73-93.21)	90.78
External validation dataset 3 (QLH[e] cohort)							
Sensitivity	83.33 (69.40-91.68)	88.73 (79.31-94.18)	95.24 (77.33-99.15)	91.99	97.56 (87.40-99.57)	96.26 (90.78-98.54)	96.91
Specificity	97.83 (92.42-99.40)	87.30 (76.89-93.42)	94.69 (88.90-97.54)	91	98.92 (94.16-99.81)	88.89 (71.94-96.15)	93.91
Precision	94.59 (82.30-98.50)	88.73 (79.31-94.18)	76.92 (57.95-88.97)	82.83	97.56 (87.40-99.57)	97.17 (92.01-99.03)	97.37
NPV	92.78 (85.85-96.46)	87.30 (76.89-93.42)	99.07 (94.94-99.84)	93.19	98.92 (94.16-99.81)	85.71 (68.51-94.30)	92.32
Accuracy	93.28 (87.73-96.43)	88.06 (81.48-92.52)	94.78 (89.61-97.45)	91.42	98.51 (94.72-99.59)	94.78 (89.61-97.45)	96.64
External validation dataset 4 (SDPH[f] cohort)							

XSL•FO
RenderX

Evaluation metrics	Severity level				Curve type		
	Normal-mild (0°-20°), coefficient (95% CI)	Moderate (21°-40°), coefficient (95% CI)	Severe (>41°), coefficient (95% CI)	Macro Average	Thoracic, coefficient (95% CI)	Thoracolumbar or lumbar, coefficient (95% CI)	Macro Average
Sensitivity	75.31 (64.92-83.41)	80.43 (66.83-89.35)	88.89 (56.50-98.01)	84.66	92.11 (79.20-97.28)	94.19 (87.10-97.49)	93.15
Specificity	90.91 (80.42-96.05)	76.67 (66.95-84.20)	96.85 (92.18-98.77)	86.76	94.90 (88.61-97.80)	62 (48.15-74.14)	78.45
Precision	92.42 (83.46-96.72)	63.79 (50.93-74.95)	66.67 (39.06-86.19)	65.23	87.50 (73.89-94.54)	81 (72.22-87.49)	84.25
NPV	71.43 (59.95-80.68)	88.46 (79.50-93.81)	99.19 (95.57-99.86)	93.38	96.88 (91.21-98.93)	86.11 (71.34-93.92)	91.49
Accuracy	81.62 (74.27-87.23)	77.94 (70.26-84.09)	96.32 (91.68-98.42)	87.13	94.12 (88.82-96.99)	82.35 (75.08-87.84)	88.24
External validation dataset 5 (HSH[g] cohort)							
Sensitivity	87.34 (78.24-92.98)	85.25 (74.28-92.04)	88.89 (74.69-95.59)	87.16	87.21 (78.53-92.71)	82.98 (69.86-91.11)	85.09
Specificity	95.88 (89.87-98.38)	87.83 (80.60-92.61)	96.43 (91.91-98.47)	93.38	98.89 (93.97-99.80)	93.02 (87.27-96.29)	95.96
Precision	94.52 (86.74-97.85)	78.79 (67.49-86.92)	86.49 (72.02-94.09)	86.60	98.68 (92.92-99.77)	81.25 (68.06-89.81)	89.97
NPV	90.29 (83.04-94.64)	91.82 (85.18-95.64)	97.12 (92.83-98.88)	93.08	89 (81.37-93.75)	93.75 (88.15-96.80)	91.38
Accuracy	92.05 (87.09-95.20)	86.93 (81.15-91.13)	94.89 (90.57-97.29)	91.29	93.18 (88.46-96.06)	90.34 (85.08-93.88)	91.76

[a]QMH&DKCH: Queen Mary Hospital and Duchess of Kent Children's Hospital.

[b]NPV: negative predictive value.

[c]PUMCH: Peking Union Medical College Hospital.

[d]FAH: First Affiliated Hospital, Zhejiang University School of Medicine.

[e]QLH: Qilu Hospital of Shandong University.

[f]SDPH: Shandong Provincial Hospital.

[g]HSH: Shanghai Huashan Hospital.

Discussion

Principal Findings

In this multicenter retrospective study, we enhanced and validated SpineHRNet+, a previously single-center validated AI model for automated spine alignment analysis, when applied to data from 5 independent external medical centers. Initial results showed that although SpineHRNet+ performed well within the center where it was originally developed, its accuracy and consistency decreased when applied to external datasets, primarily due to substantial variations in data characteristics across centers. To mitigate this issue, we integrated a novel data transformation technique into the AI model's training and fine-tuning processes, enabling it to adapt more effectively to differences among various medical centers. Following this approach, the model's ability to accurately analyze spinal alignment improved significantly across diverse clinical environments.

Previous studies have demonstrated that AI-based approaches can reliably quantify spinal deformities, enhancing clinical assessments of AIS [26,27]. However, such models typically undergo validation using data from single institutions, with limited evaluation of their generalizability across multiple medical institutions [28-30]. Variability across clinical environments, such as differences in equipment, acquisition protocols, patient positioning, and data interpretation practices, can substantially impact the performance of AI models when deployed more broadly [16,17]. To address this challenge, advanced research has explored generative AI techniques to standardize imaging characteristics across datasets by transforming images to match the style of the training data [31,32]. While effective in some cases, these methods typically require large amounts of pixel-level registered data and may introduce image distortions, limiting their clinical applicability in diverse settings [33,34].

In contrast, our study introduced a computationally efficient and clinically applicable data transformation method focusing on standardizing key quantitative characteristics of spinal radiographs across different health care settings, without altering critical anatomical information of spine. This algorithm uses statistical analysis of pixel intensity distributions combined with optimization techniques to achieve unsupervised data transformation, eliminating the need for extensive pixel-level

XSL•FO
RenderX

matched data or high computational resources. With a computational complexity of $O(n \log n)$, it efficiently processes images in real time (detailed in Section 3.3 in Multimedia Appendix 1). This practicality facilitates real-time processing and broad applicability across clinical settings.

Our findings demonstrate that the enhanced SpineHRNet+ model substantially improved in accuracy and consistency for CA prediction across both internal and external validation cohorts. Specifically, the mean absolute prediction errors were consistently decreased for both internal and independent external datasets, with a particularly notable reduction on external data. This outcome suggests that the proposed method not only improves overall accuracy but also has an even greater impact on cross-center data. In addition, the enhanced model narrowed the R^2 gap between internal and external datasets to less than 0.03, keeping regression slope differences within 0.6° (Figure 3). Bland-Altman analysis further showed that the discrepancy in the LoA interval shrank to within 6.42° (Figure S3 in Multimedia Appendix 1), reflecting improved consistency and stability of AI model performance across varying imaging environments. Given the uneven distribution of disease severity across medical centers (Table 1), where severe cases pose greater analytical challenges than mild cases for CA prediction, the enhanced model's consistency across datasets further underscores the contribution of our method in improving the robustness of AI models in handling varying disease severity levels.

The improvements in disease severity grading further validate the effectiveness of the proposed method. In the internal validation dataset, the enhanced model achieved a macro-average sensitivity and NPV increase of 3.85% and 1.74%, respectively. In independent external datasets, the enhancements were even more pronounced, with sensitivity improvements ranging from 4.53% to 11.88% and NPV increases between 1.37% and 5.26% (Table S5, Figure S4, and Table S3 in Multimedia Appendix 1). These advancements narrowed the performance gap between internal and external data, further demonstrating the effectiveness and scalability of the proposed method in handling heterogeneous medical imaging data. Notably, this plug-and-play design allows for seamless integration into existing AI pipelines without the need for retraining, with performance enhancements achievable through minimal calibration. It demonstrated the most substantial improvements on external datasets, with prediction mean error reductions 1.4°, and average sensitivity gains 12.57%. External cohorts exhibited distinct severity distributions (eg, PUMCH mean CA 46.6° vs training 25°). This heterogeneity underscores the model's generalizability, as performance remained robust despite these differences, validating its clinical scalability.

The improved sensitivity and NPV for severity grading suggest enhanced reliability in identifying moderate-to-severe cases that may benefit from early bracing intervention. For instance, in the internal validation cohort, the model correctly identified 92.62% of moderate cases and 87.50% of severe cases, reducing the risk of under-treatment. Confusion matrix analysis (Table 3) further supports the model's ability to minimize false negatives, which is critical in avoiding missed opportunities for timely intervention. Although some improvements in mean absolute error and R^2 appear numerically modest in the internal dataset, the gains across external datasets were substantial and clinically meaningful. For instance, the enhanced model reduced CA prediction error by up to 2.35° and significantly improved R^2 in external cohorts. These improvements led to enhanced sensitivity in severity grading, directly influencing treatment decisions, such as initiating bracing in borderline cases and improving diagnostic reliability across centers. The mean error of the enhanced model, approximately 4°, is comparable to or even better than the reported inter-observer variability, which ranges from 4.9° to 10°. This finding suggests not only the model's accuracy but also its clinical relevance, highlighting its potential for reliable application in practice.

The lower precision observed in mild classification (eg, 57.14% in PUMCH) reflects a trade-off with sensitivity. The model prioritizes sensitivity to avoid missing true positives, especially in moderate and severe cases, which may lead to increased false positives in mild cases. This is partly due to class imbalance and overlapping radiographic features. In clinical workflows, such predictions should be reviewed by clinicians, particularly for borderline cases, to ensure appropriate management decisions.

The real-time processing capability of our model, with an average inference time of 300 milliseconds per image (excluding approximately 200 milliseconds required for the transformation step), supports seamless integration into clinical workflows. Combined with the transformation step, the total processing time remains under 500 milliseconds, enabling potential deployment within picture archiving and communication system/radiology information system without disrupting existing infrastructure.

Limitations

Despite its advantages, the proposed data transformation has certain limitations. First, it primarily addresses variability within a single modality (eg, radiographic data) arising from equipment and protocol differences across institutions, but it does not accommodate cross-modality transformations, such as conversions between magnetic resonance imaging, computed tomography, and radiographs. Additionally, as the method focuses solely on adjusting quantitative properties of clinical data without modifying image content, challenges related to image class imbalance, low resolution, or issues associated with suboptimal data quality remain unresolved. BMI data were unavailable for 3 external cohorts, limiting analysis of body habitus effects on radiographic quality. Future research should explore more comprehensive solutions to integrate content-aware enhancements while maintaining the anatomical integrity of medical imaging data, and explore adapting this transformation framework to cross-modality scenarios, such as aligning radiographs with computed tomography or magnetic resonance imaging data. While technically challenging, such extensions could enable unified AI models across imaging modalities, enhancing diagnostic versatility.

Conclusions

In conclusion, our multicenter study demonstrates that a previously validated AI model for AIS assessment can achieve

XSL•FO
RenderX

significantly enhanced performance through the integration of a data transformation method specifically designed to address variations in clinical data across institutions. By improving consistency and accuracy across diverse health care settings in multiple medical centers, our findings highlight the necessity of addressing data variability to ensure the reliable, scalable implementation of AI models in clinical practice. The proposed methodology underscores broader clinical implications, emphasizing the importance of adaptable and robust AI solutions capable of supporting equitable and accurate patient care across diverse health care environments.

Acknowledgments

We sincerely appreciate the support provided by the National Natural Science Foundation of China Young Scientists Fund, the Health Services Research Fund, and the National High Level Hospital Clinical Research Funding.

Funding

This study was funded by the National Natural Science Foundation of China Young Scientists Fund (NSFC 82303957, NSFC 82402398), the Health Services Research Fund (HMRF 19200911), and the National High Level Hospital Clinical Research Funding (2022-PUMCH-D-004). This research was also supported by the National Key R&D Program of China (grant number 2022YFC2502901) and the Li Shu Fan Medical Foundation Professorship for Orthopaedic Surgery.

Data Availability

The test data is available from the corresponding authors on a reasonable request and under the related institution policy.

Authors' Contributions

GLC was responsible for data collection, annotation, and experimental validation at PUMCH. NM took charge of the experimental validation at QMH and DKCH. ZC managed the experiments and data collection at SDPH. ZB was responsible for experiments conducted in QLH and data collection in QLH. ZYG coordinated the experiments and data collection at HSH. JWS was responsible for experiments conducted in FAH and data collection in FAH. YPZ, TH, XHK, and PYL implemented the deep learning model, deployed it on the server, and managed server operations. CN and QFY assisted with data preprocessing at HSH. ZFC, and HJ assisted with data preprocessing and analysis at FAH. ZB, ZMZ, CJZ, YL, NW, and JPYC supervised experiments, landmark annotation, and clinical parameter analysis at their respective medical centers. JPYC, NW, and JGZ co-supervised the study design. TZ led the study design, supervised data collection in 2 Hong Kong clinics, managed model implementation, developed the open system, and coordinating all study sites. GLC, NM, TZ, and TH drafted the manuscript, while all authors reviewed, proofread, and approved the final version for important intellectual content. ZMZ, CJZ, YL, NW, JPYC, JGZ, and TZ are co-corresponding authors who equally contributed to this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Multicenter data collection procedures, inclusion and exclusion criteria, intensity distribution–based data transformation methodology, heterogeneity analysis, and comparative performance evaluation between the enhanced and original models.
[DOCX File , 12241 KB - jmir_v28i1e78396_app1.docx]

References

1. Meng N, Wong KYK, Zhao M, Cheung JP, Zhang T. Radiograph-comparable image synthesis for spine alignment analysis using deep learning with prospective clinical validation. EClinicalMedicine 2023;61:102050 [FREE Full text] [doi: 10.1016/j.eclinm.2023.102050] [Medline: 37425371]
2. Fong DYT, Cheung KMC, Wong YW, Wan YY, Lee CF, Lam TP, et al. A population-based cohort study of 394,401 children followed for 10 years exhibits sustained effectiveness of scoliosis screening. Spine J 2015;15(5):825-833 [FREE Full text] [doi: 10.1016/j.spinee.2015.01.019] [Medline: 25615844]
3. Charalampidis A, Diarbakerli E, Dufvenberg M, Jalalpour K, Ohlin A, Ahl AA, et al. Nighttime bracing or exercise in moderate-grade adolescent idiopathic scoliosis: a randomized clinical trial. JAMA Netw Open 2024;7(1):e2352492 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.52492] [Medline: 38285447]
4. Weinstein SL, Dolan LA, Spratt KF, Peterson KK, Spoonamore MJ, Ponseti IV. Health and function of patients with untreated idiopathic scoliosis: a 50-year natural history study. JAMA 2003;289(5):559-567. [doi: 10.1001/jama.289.5.559] [Medline: 12578488]
5. Cobb J. Outline for the study of scoliosis. Instructional Course Lecture. 1948. URL: https://www.semanticscholar.org/paper/Outlines-for-the-study-of-scoliosis-Cobb-Cobb/4e90f92a918301405d93fb1b90e69c1d8fffa767 [accessed 2026-02-13]

XSL•FO
RenderX

6. Carman DL, Browne RH, Birch JG. Measurement of scoliosis and kyphosis radiographs. Intraobserver and interobserver variation. The Journal of Bone & Joint Surgery 1990;72(3):328-333. [doi: 10.2106/00004623-199072030-00003]
7. Morrissy RT, Goldsmith GS, Hall EC, Kehl D, Cowie GH. Measurement of the cobb angle on radiographs of patients who have scoliosis. Evaluation of intrinsic error. The Journal of Bone & Joint Surgery 1990;72(3):320-327. [doi: 10.2106/00004623-199072030-00002]
8. Zhang T, Zhu C, Zhao Y, Zhao M, Wang Z, Song R, et al. Deep learning model to classify and monitor idiopathic scoliosis in adolescents using a single smartphone photograph. JAMA Netw Open 2023;6(8):e2330617 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.30617] [Medline: 37610748]
9. Zhao M, Meng N, Cheung JPY, Yu C, Lu P, Zhang T. SpineHRformer: a transformer-based deep learning model for automatic spine deformity assessment with prospective validation. Bioengineering (Basel) 2023;10(11):1333 [FREE Full text] [doi: 10.3390/bioengineering10111333] [Medline: 38002457]
10. Yuan W, Shi W, Chen L, Liu D, Lin Y, Li Q, et al. Digital physiotherapeutic scoliosis-specific exercises for adolescent idiopathic scoliosis: a randomized clinical trial. JAMA Netw Open 2025;8(2):e2459929 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.59929] [Medline: 39964686]
11. Zhao M, Meng N, Cheung JPY, Zhang T. PCT-GAN: a real CT image super-resolution model for trabecular bone restoration. IEEE 20th International Symposium on Biomedical Imaging (ISBI) 2023;20:1-5. [doi: 10.1109/isbi53787.2023.10230389]
12. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
13. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018;18(8):500-510 [FREE Full text] [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
14. Liang Y, Lv J, Li D, Yang X, Wang Z, Li Q. Accurate cobb angle estimation on scoliosis X-Ray images via deeply-coupled two-stage network with differentiable cropping and random perturbation. IEEE J Biomed Health Inform 2023;27(3):1488-1499. [doi: 10.1109/jbhi.2022.3229847]
15. Meng N, Cheung JP, Wong KK, Dokos S, Li S, Choy RW, et al. An artificial intelligence powered platform for auto-analyses of spine alignment irrespective of image quality with prospective validation. EClinicalMedicine 2022;43:101252 [FREE Full text] [doi: 10.1016/j.eclinm.2021.101252] [Medline: 35028544]
16. Viana-Ferreira C, Ribeiro LS, Costa C. A framework for integration of heterogeneous medical imaging networks. Open Med Inform J 2014;8(1):20-32 [FREE Full text] [doi: 10.2174/1874431101408010020] [Medline: 25279021]
17. Chang Q, Yan Z, Zhou M, Qu H, He X, Zhang H, et al. Mining multi-center heterogeneous medical data with distributed synthetic learning. Nat Commun 2023;14(1):5510 [FREE Full text] [doi: 10.1038/s41467-023-40687-y] [Medline: 37679325]
18. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. Practical guidance on artificial intelligence for health-care data. The Lancet Digital Health 2019;1(4):e157-e159. [doi: 10.1016/s2589-7500(19)30084-6]
19. Lancet T. AI in medicine: creating a safe and equitable future. The Lancet 2023;402(10401):503. [doi: 10.1016/s0140-6736(23)01668-9]
20. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature 2020;585(7825):357-362 [FREE Full text] [doi: 10.1038/s41586-020-2649-2] [Medline: 32939066]
21. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods 2020;17(3):261-272. [doi: 10.1017/9781108778039.009]
22. Allen M, Poggiali D, Whitaker K, Marshall TR, van Langen J, Kievit RA. Raincloud plots: a multi-platform tool for robust data visualization. Wellcome Open Res 2019;4:63 [FREE Full text] [doi: 10.12688/wellcomeopenres.15191.2] [Medline: 31069261]
23. McKinney W. Data structures for statistical computing in Python. in SciPy 2010;445:51-56. [doi: 10.25080/majora-92bf1922-00a]
24. Waskom M, Botvinnik O, O'Kane D, Hobson P, Lukauskas S. Mwaskom/seaborn: V0. 8.1 (September 2017). Zenodo. 2017. URL: https://zenodo.org/records/883859 [accessed 2026-02-13]
25. Hunter JD. Matplotlib: a 2D graphics environment. Comput. Sci. Eng 2007;9(3):90-95. [doi: 10.1109/mcse.2007.55]
26. Shi W, Xu T, Yang H, Xi Y, Du Y, Li J, et al. Attention gate based dual-pathway network for vertebra segmentation of X-Ray spine images. IEEE J Biomed Health Inform 2022;26(8):3976-3987. [doi: 10.1109/jbhi.2022.3158968]
27. Lin Y, Liu L, Ma K, Zheng Y. Seg4reg+: consistency learning between spine segmentation and cobb angle regression, in medical image computing and computer assisted intervention. : Springer; 2021 Presented at: Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference; 2021 Oct 01; Strasbourg, France p. 490-499. [doi: 10.1007/978-3-030-87240-3_47]
28. Zou L, Guo L, Zhang R, Ni L, Chen Z, He X, et al. VLTENet: a deep-learning-based vertebra localization and tilt estimation network for automatic cobb angle estimation. IEEE J Biomed Health Inform 2023;27(6):3002-3013. [doi: 10.1109/jbhi.2023.3258361]
29. Wang L, Xu Q, Leung S, Chung J, Chen B, Li S. Accurate automated Cobb angles estimation using multi-view extrapolation net. Medical Image Analysis 2019;58:101542. [doi: 10.1016/j.media.2019.101542]

XSL•FO
RenderX

30. Huo L, Cai B, Liang P, Sun Z, Xiong C. Joint spinal centerline extraction and curvature estimation with row-wise classification and curve graph network. 2021 Presented at: Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference; 2021 Oct 01; Strasbourg, France p. 377-386. [doi: 10.1007/978-3-030-87240-3_36]
31. Wang Q, Breckon TP. Cross-domain structure preserving projection for heterogeneous domain adaptation. Pattern Recognition 2022;123:108362. [doi: 10.1016/j.patcog.2021.108362]
32. Li L, Zhang Z. Semi-supervised domain adaptation by covariance matching. IEEE Trans Pattern Anal Mach Intell 2019;41(11):2724-2739. [doi: 10.1109/tpami.2018.2866846]
33. Liu, Q, Zeng Z, He J, Yu Q, Shen X, Chen LC. Alleviating distortion in image generation via multi-resolution diffusion models and time-dependent layer normalization. 2024 Presented at: The Thirty-eighth Annual Conference on Neural Information Processing Systems; 2024 Dec 9; Canada. [doi: 10.52202/079017-4254]
34. Croitoru FA, Hondru V, Ionescu RT, Shah M. Diffusion models in vision: a survey. IEEE Trans Pattern Anal Mach Intell 2023;45(9):10850-10869. [doi: 10.1109/tpami.2023.3261988]

Abbreviations

AI: artificial intelligence
AIS: adolescent idiopathic scoliosis
CA: Cobb angle
CDF: cumulative distribution function
DKCH: Duchess of Kent Children's Hospital
FAH: First Affiliated Hospital, Zhejiang University School of Medicine
GT: ground truth
HSH: Shanghai Huashan Hospital
IRB: Institutional Review Board
LoA: limits of agreement
NPV: negative predictive value
PUMCH: Peking Union Medical College Hospital
QLH: Qilu Hospital of Shandong University
QMH: Queen Mary Hospital
SDPH: Shandong Provincial Hospital

Edited by A Coristine; submitted 02.Jun.2025; peer-reviewed by RK Kanji, P Sharma, VV Sangaraju; comments to author 09.Jul.2025; accepted 07.Nov.2025; published 20.Mar.2026.

Please cite as:
Chen G, Meng N, Zhuang Y, Chen Z, Bian Z, Gong Z, Shi J, Huang T, Kuang X, Lu P, Nie C, Yu Q, Chen Z, Jiang H, Zhang Z, Zheng C, Liang Y, Wu N, Cheung JPY, Zhang J, Zhang T
Scalable and Robust Artificial Intelligence for Spine Alignment Assessment: Multicenter Study Enabled by Real-Time Data Transformation
J Med Internet Res 2026;28:e78396
URL: https://www.jmir.org/2026/1/e78396
doi:10.2196/78396
PMID:

©Guilin Chen, Nan Meng, Yipeng Zhuang, Zhe Chen, Zhen Bian, Zhaoyang Gong, Jiawei Shi, Tao Huang, Xihe Kuang, Pengyu Lu, Cong Nie, Qifeng Yu, Zefu Chen, Hui Jiang, Zhongmin Zhang, Chaojun Zheng, Yu Liang, Nan Wu, Jason Pui Yin Cheung, Jianguo Zhang, Teng Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Measuring the Impact of AI on Report-Drafting Efficiency in Chest Computed Tomography Interpretation: Retrospective Analysis

Weiqi Liu[1,2*], DrPH; You Wu[3*], PhD; Wei Yu[4], PhD; Mark J Bittle[1], DrPH; Zhuozhao Zheng[5], PhD; Hadi Kharrazi[1], PhD, MD

[1]Department of Health Policy and Management, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, United States
[2]Department of Research, Sophmind Technology (Beijing) Co. Ltd, Beijing, China
[3]Institute for Hospital Management, School of Medicine, Tsinghua University, Beijing, China
[4]Department of Radiology, Beijing Anzhen Hospital, Capital Medical University, Beijing, China
[5]Department of Radiology, Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua University, Beijing, China
[*]these authors contributed equally

Corresponding Author:
Zhuozhao Zheng, PhD
Department of Radiology
Beijing Tsinghua Changgung Hospital, School of Clinical Medicine
Tsinghua University
168 Litang Road
Changping District
Beijing, 102218
China
Phone: 86 010 56112345
Email: zzza00509@btch.edu.cn

Abstract

Background: Artificial intelligence (AI), particularly deep learning, has shown promise in enhancing medical image interpretation and improving radiologists' efficiency. In China, growing imaging demand and workforce shortages have placed increasing pressure on radiology services. However, evidence on the operational impact of AI on reporting efficiency remains limited.

Objective: This study aimed to evaluate the effect of an AI system on radiologists' reporting efficiency by examining changes in report-drafting time for lung nodule diagnosis in chest computed tomography (CT) images.

Methods: We analyzed 185,044 chest CT reports from Beijing Anzhen and Tsinghua Changgung Hospitals (2018-2023) using a difference-in-differences design with nonequivalent comparison groups. Report-drafting time before, immediately after, and up to 2 years following AI implementation was compared, adjusting for radiologist gender, seniority, and years of working experience.

Results: The pooled analysis showed a modest overall increase of 0.86 minutes (95% CI 0.14 to 1.57). However, this masked substantial heterogeneity between hospitals due to differing implementation timelines. In the first year after AI deployment, Tsinghua Changgung Hospital experienced a nonsignificant increase of 0.90 minutes (95% CI –0.28 to 2.08). In contrast, at Beijing Anzhen Hospital, the AI-assisted group exhibited an absolute reduction of 0.76 minutes by the first year and a further 1.83-minute reduction by the second year (an approximate 28% time saved vs baseline), while the control group remained stable over time. Using a difference-in-differences framework, this corresponded to a 2.66-minute relative improvement compared with the counterfactual trend ($P<.001$).

Conclusions: AI-assisted lung nodule diagnosis may initially increase report-drafting time due to adaptation and workflow adjustment. Sustained, meaningful efficiency gains were heterogeneous and observed at only 1 of the 2 study sites, indicating that long-term impacts are strongly contingent on site-specific implementation dynamics, learning curves, and local context.

(J Med Internet Res 2026;28:e77967) doi:10.2196/77967

KEYWORDS

lung nodule; chest computed tomography; radiology; artificial intelligence; AI; efficiency; clinical workflow evaluation

XSL•FO
RenderX

Introduction

Artificial intelligence (AI) can imitate human intelligence, perform tasks, and make progressive enhancements based on amassed data [1]. Particularly, the application of deep learning technology in medical image segmentation and recognition has brought revolutionary changes to the operation of imaging services [2]. The increasing adoption of AI has the potential to alter the working conditions of radiologists [3].

The rapid evolution of radiological modalities has resulted in a substantial increase in the volume of medical images. While this progress has numerous advantages, it has significantly increased radiologists' workloads [4,5]. For example, workplace burnout could weaken radiologists' perceived competence [6], and occupational fatigue may compromise radiologists' diagnostic accuracy, thereby adversely affecting patient care and treatment outcomes [7]. Therefore, it is critical to explore how AI systems can enhance the efficiency of radiologists for more effective applications in clinical practice.

Globally, the scarcity of physicians persists due to the substantial professional requirements and protracted talent development time. This challenge is exacerbated in China, where high patient demand contrasts with inadequate investments in medical education, relatively low remuneration for health care practitioners, and enduring disproportions in medical resource allocation [8]. AI stands as a pivotal technological approach that could potentially alleviate the scarcity of medical resources in China. Furthermore, supportive governmental policies are poised to contribute significantly to the advancement of AI within the Chinese health care sector [9].

The impact of AI systems on the operational efficiency of radiologists remains a topic of uncertainty [10]. Within the medical and scientific communities, there is a prevailing belief that the rapid development of AI in health care will lead to a transformative paradigm shift in medical imaging. Enhancing operational efficiency has been deemed part of the AI's paradigm shift in medical imaging [11]; however, empirical evidence is limited and sometimes conflicting [12,13]. For example, Yi et al [14] unveiled that the implementation of an AI system for breast cancer detection resulted in a substantial reduction in reporting time. Stephan et al [15] reported improved efficiency in generating dental radiology reports, while Zia et al [16] found that implementing an AI system for intracranial hemorrhage detection led to longer turnaround time. Nonetheless, it is essential to acknowledge that these findings may not provide a definitively representative picture, primarily due to the limited sample sizes of such studies [12]. Additionally, the results of these studies often do not provide a comprehensive understanding of the enduring implications of AI system adoption in medical imaging, as the study periods have been relatively brief [12].

This real-world study was conducted in 2 large hospitals in China, using data spanning multiple years and the difference-in-differences (DID) methodology. The study assesses the alterations in report-drafting time for lung nodule diagnosis in computed tomography (CT) scans after AI adoption. Results of this study can inform the radiological community on the prospective advantages and challenges associated with AI integration in medical imaging.

Methods

Study Population

This study was conducted at 2 tertiary health care facilities situated in Beijing, China: Beijing Anzhen Hospital and Tsinghua Changgung Hospital. Figure 1 illustrates the diagnostic process in the radiology departments both before and after the introduction of the AI system. In the traditional workflow, before the AI system's implementation, the typical process involved a junior radiologist writing a diagnostic report after each patient underwent a CT examination. Subsequently, a senior radiologist would review and finalize this report before it was released to clinicians and patients. With the integration of the AI system, both junior and senior radiologists can use the findings of the AI systems as diagnostic aids. The AI system will automatically identify and analyze lung nodules in imported chest CT images. Junior and senior radiologists can reference the diagnostic results provided by the AI system during their diagnostic process and ultimately complete their diagnostic reports. Case assignment follows a randomized queue-based workflow in both hospitals, without systematic allocation of more complex cases to more experienced radiologists.

XSL•FO
RenderX

Figure 1. Traditional vs artificial intelligence (AI)–assisted workflows in the radiology departments of the 2 hospitals in this study.

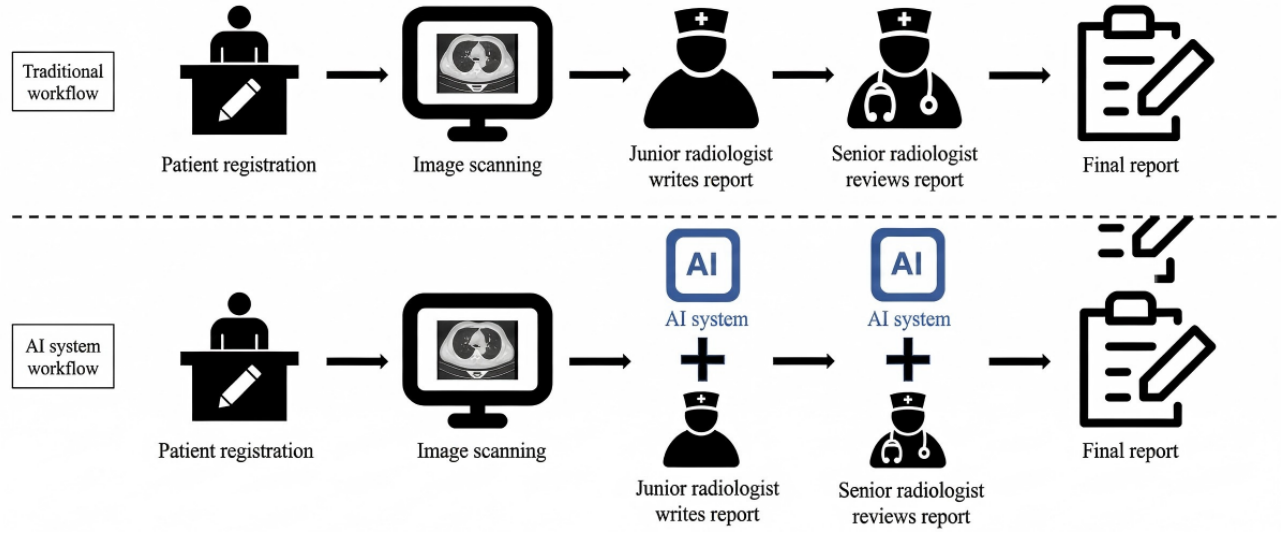

The lung nodule AI systems implemented at Beijing Anzhen Hospital and Tsinghua Changgung Hospital were developed and integrated by Care.ai [17] (Yitu, integrated into Deepwise's framework in 2021) and Dr.Wise (Deepwise) [18]. Both systems followed a standardized reporting workflow and user interface, and no vendor change occurred during the study period. Both AI systems were approved by the National Medical Products Administration and provided comparable functionalities for lung nodule detection and reporting assistance. These AI systems were introduced at Beijing Anzhen Hospital in January 2019 and at Tsinghua Changgung Hospital in June 2021. The deployment phase, which involved installation and system integration, spanned approximately 1 month. Subsequently, an additional 3-month period was required to attain optimal and seamless system functionality. Multimedia Appendix 1 illustrates the features and capabilities of the AI-assisted diagnostic system as shared with radiologists. This system demonstrates the capability to automatically detect lung nodules; assess their quantity, dimensions, and precise locations; and subsequently generate corresponding diagnostic reports. Multimedia Appendices 2 and 3 illustrate the detailed operational workflow of the AI system and show that the AI system use processes in the 2 hospitals are essentially identical.

We analyzed the CT report-drafting time for individual patients from April 2018 to March 2022 at these 2 hospitals. Data for the following periods were excluded from the analysis: January 2019 to March 2019, January 2020 to August 2020, and April 2021 to August 2021, due to the potential impact of the COVID-19 pandemic, AI system integration, and the Spring Festival holiday on patterns of health seeking behavior in China [19]. Due to the considerable volatility in the night shift patterns of radiologists, this study exclusively incorporated day shift reports for analysis. To mitigate potential confounding factors related to workforce turnover, our analysis exclusively examined radiology reports written by the same group of radiologists who had generated reports both before and after the introduction of the AI system.

Data Collection

Using the University of California, San Francisco calculator [20], sample size calculations indicated a necessary sample of 12 to detect an effect size of 1 (difference in minutes between AI-assisted and conventional report-drafting time) with a SD of 0.5, ensuring 80% power at an α level of .05 (2-tailed). With nearly 100,000 reports generated at each hospital during the study period, the sample markedly surpassed requirements.

The data collection procedure encompassed the extraction of all junior radiologist report-drafting time records from the Radiology Information System and Picture Archiving and Communication System of both hospitals throughout the specified sampling period. The collected dataset included various patient-related details, such as age, sex, source of referral, examination specifics, examination time, reporting time by both junior and senior radiologists, and the identities of the reporting radiologists.

The report-drafting time analyzed in this study reflects the first-round interpretation performed by junior radiologists. The primary outcome of interest, report-drafting time, was defined as the duration between loading the CT images and clicking the Submit button on the report form [21]. To ensure data quality, we conducted a detailed assessment of extreme values. Report-drafting time of less than 0.5 minutes or more than 40 minutes—together representing approximately 5% of all observations (short drafting time: 6287/195,020, 3.2%; long drafting time: 3689/195,020, 1.9%)—were excluded, as such values most likely reflected system artifacts or nonclinical interruptions (idle time) rather than true behaviors. To ensure data integrity, a 2-investigator workflow was used. One extracted data from the reporting systems, applied predefined inclusion and exclusion criteria, and calculated report-drafting times. A second investigator independently verified the processed dataset to ensure accuracy, completeness, and consistency.

Statistical Analysis

This study used a DID design, which involved a real-world retrospective assessment of the outcome within an experimental

XSL•FO
RenderX

and a comparison group. The experimental group in our study comprised chest CT images interpreted using the AI-aided diagnostic system following its implementation, while the comparison group included other CT images that were not affected by the introduction of the AI system.

Although the intervention group focused on chest CT images and the control group included other CT images (eg, abdominal, pelvic, and head CT), all examinations followed the same radiology workflow within each institution. This institutional uniformity ensured that reporting procedures were comparable, minimizing systematic bias unrelated to the AI intervention.

Before using the DID methodology, an assessment was conducted to evaluate the parallel trend assumption. This assumption necessitates that the trends exhibited by both the experimental and comparison groups remain consistent before the intervention is introduced [22]. The DID analysis used an ordinary least squares regression model, as shown in equation 1.

where the dependent variable (y_{ij}) denotes the report-drafting time for case i interpreted by radiologist j. Tij indicates AI use (1 if the AI-assisted system was used, 0 otherwise), and I_j denotes group assignment (1 for the experimental group and 0 for the comparison group). The interaction term $T_{ij}\times I_j$ captures the DID estimator of interest. X_{ij} represents a vector of covariates, including radiologist sex, professional title rank, and years of working experience. SEs were clustered at the radiologist level to account for within-radiologist correlation. Models included hospital fixed effects to control for time-invariant institutional differences, as well as calendar quarter fixed effects to account for common seasonal and temporal shocks. We further tested for effect modification by radiologist seniority by adding AI-seniority interaction terms to the main model.

In this analysis, a significance level of $P<.05$ was used to determine statistical significance. The statistical analyses were conducted using Stata (version 15.1; StataCorp LLC) and R (version 4.4.1; R Foundation for Statistical Computing).

Ethical Considerations

This study received ethics approval from the institutional review board (IRB) of Tsinghua Changgung Hospital (23352-4-01) and from the IRB of Beijing Anzhen Hospital (2023078X). Additional IRB approval was also granted by the Johns Hopkins Bloomberg School of Public Health (FWA 00000287). The IRB waived the requirement for informed consent, as this study involved secondary analysis of retrospective, noninterventional data originally collected for operational purposes. According to institutional policy, individual consent is not required for such analyses if the data are fully deidentified and used exclusively for quality improvement or research purposes. No compensation was provided to the participants in this retrospective study.

All data analyzed in this study were fully deidentified before access and analysis, with no personally identifiable information or protected health information included (sample data for review in Multimedia Appendix 4). Data were stored on secure institutional servers with access restricted to authorized study personnel only. This study does not contain any identifiable images or multimedia materials of individual participants. All figures and tables are based on aggregated, anonymized data.

Results

The study incorporated 185,044 CT images that met the predefined criteria. Table 1 provides an overview of the demographic characteristics of patients from both the experimental and comparison groups, before and after the introduction of the AI system. Beijing Anzhen Hospital accounted for a larger proportion of patients after the AI launch due to its earlier implementation. The sex ratio between the experimental and comparison groups was comparable, but there were slightly more male patients than female patients in both groups. The source of patients varied slightly based on medical needs, with outpatient clinics contributing the largest number of patients.

XSL•FO
RenderX

Table 1. Patient characteristics before and after the artificial intelligence (AI) launch in experimental and control groups.

	Overall (n=185,044), n (%)	Before AI launch (n=93,870)		After AI launch (n=91,174)	
		Control group (n=55,592), n (%)	Experimental group (n=38,278), n (%)	Control group (n=50,751), n (%)	Experimental group (n=40,423), n (%)
Age (years)					
<18	1665 (0.9)	676 (1.2)	253 (0.7)	424 (0.8)	312 (0.8)
18-54	72,493 (39.4)	19,513 (35.4)	16,711 (44.1)	18,703 (37.0)	17,566 (43.6)
>55	109,712 (59.7)	34,939 (63.4)	20,910 (55.2)	31,418 (62.2)	22,445 (55.7)
Sex					
Female	85,216 (46.1)	25,734 (46.3)	17,964 (46.9)	23,846 (47.0)	17,672 (43.7)
Male	99,828 (53.9)	29,858 (53.7)	20,314 (53.1)	26,905 (53.0)	22,751 (56.3)
Hospital					
Beijing Anzhen Hospital	97,366 (52.6)	23,163 (41.7)	7880 (20.6)	36,009 (71.0)	30,314 (75.0)
Tsinghua Changgung Hospital	87,678 (47.4)	32,429 (58.3)	30,398 (79.4)	14,742 (29.0)	10,109 (25.0)
Patient source					
Emergency room	57,687 (31.2)	14,946 (26.9)	8731 (22.8)	18,284 (36.0)	15,726 (38.9)
Inpatient	38,675 (20.9)	17,215 (31.0)	5684 (14.8)	11,084 (21.8)	4692 (11.6)
Outpatient	81,655 (44.1)	23,364 (42.0)	19,686 (51.4)	21,267 (41.9)	17,338 (42.9)
Physical exam	7027 (3.8)	67 (0.1)	4177 (10.9)	116 (0.2)	2667 (6.6)

We assessed the characteristics of the CT images across 2 key dimensions to ensure their comparability over time: case mix and case complexity. First, the anatomical distribution of CT images remained largely consistent, dominated by head and abdominal studies (Figure 2A). Second, the prevalence of primary abnormal findings (eg, “liver cyst” and “cerebral atrophy”), a proxy for interpretation complexity, was also longitudinally stable in high-volume domains such as abdomen and head CTs (Figure 2B). Given that the same 74 radiologists interpreted both intervention and control images across time, the overall case mix, case complexity, and workload remained stable. However, the job titles and working years varied for the radiologists (Figure 3). Those working at Beijing Anzhen Hospital had a higher job rank and longer working experience. Therefore, these variables were further adjusted in the main analyses. Multimedia Appendix 5 presents the outcomes of the parallel trend assessment, illustrating that the report-drafting times of both experimental and comparison groups displayed a consistent trend before the introduction of the AI system.

XSL•FO
RenderX

Figure 2. Anatomical distribution of scans and prevalence of key abnormal findings. (A) Bar chart showing the anatomical distribution of computed tomography (CT) images for the chest (treatment) and nonchest (control) groups before and after the implementation of artificial intelligence (AI). (B) Stacked area charts depicting the monthly prevalence of primary abnormal findings from 2018 to 2021 for the 3 highest-volume anatomical groups: abdomen and pelvis (left), chest and lung (middle), and head and brain (right).

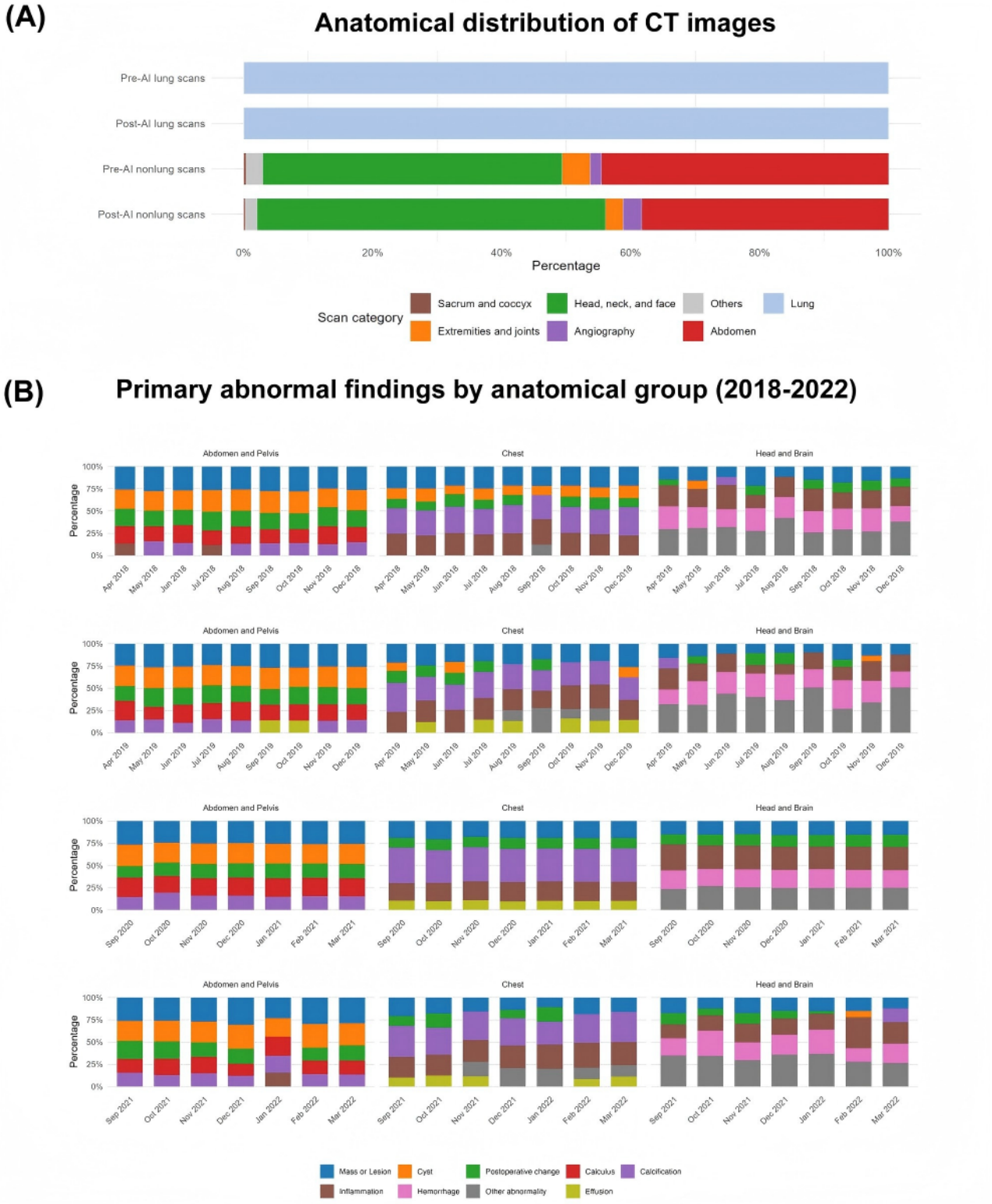

XSL•FO
RenderX

Figure 3. Radiologists' job titles and working years by hospital.

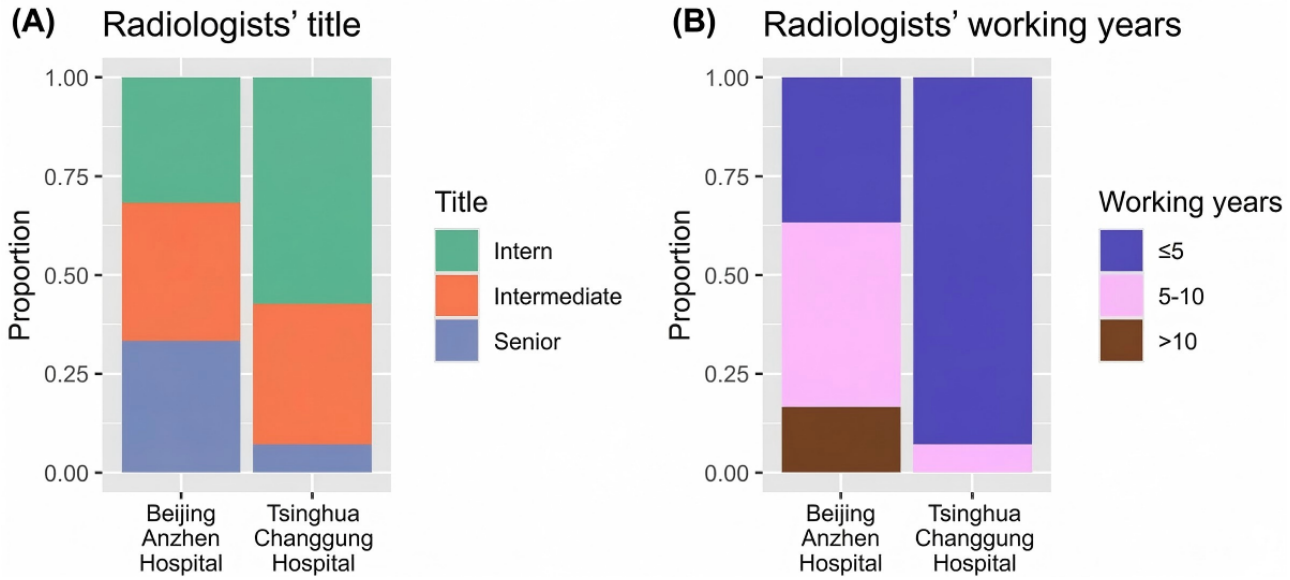

The adjusted DID model revealed a pooled effect size of 0.86 minutes (95% CI 0.14 to 1.57) after AI implementation. To assess the potential influence of anatomical site, we conducted sensitivity analyses by sequentially excluding each major category of nonchest CT images from the control group. As shown in Multimedia Appendix 6, the direction and magnitude of the AI effect remained generally consistent across all exclusion scenarios. However, this aggregated result should be interpreted with caution, as it is averaged over 2 hospitals with different implementation timelines. When breaking down to the hospital level, Tsinghua Changgung Hospital had a nonsignificant 0.90-minute increase (*P*=.12), while Beijing Anzhen Hospital saw a statistically significant 1.37-minute time saving (*P*<.001; Table 2). No statistically significant interaction was observed between AI exposure and radiologist seniority (all *P*>.09). In sensitivity analyses controlling for patient source, the estimated AI-associated effects remained robust (the overall DID estimate remained 0.86 minutes; *P*=.02).

Table 2. Adjusted report-drafting time differences for chest computed tomography (CT) vs others before and after artificial intelligence (AI) implementation, overall and by hospital.

Intervention and group	Patients, n	Report-drafting time[a] (minutes; 95% CI)	*P* value
Overall (2 hospitals)	185,044		
Control group[b]			
Pre-AI implementation	55,592	7.58 (6.32 to 8.85)	—[c]
Post-AI implementation	50,751	7.51 (6.10 to 8.93)	—
Pre-post difference	—	–0.07 (–0.62 to 0.48)	—
Experimental group[d]			
Pre-AI implementation	38,278	6.79 (5.40 to 8.19)	—
Post-AI implementation	40,423	7.58 (6.17 to 8.99)	—
Pre-post difference	—	0.78 (0.26 to 1.30)	—
DID[e]	—	0.86 (0.14 to 1.57)	.02
Tsinghua Changgung			
DID	87,678	0.90 (–0.28 to 2.08)	.12
Beijing Anzhen			
DID	97,366	–1.37 (–2.04 to –0.70)	<.001

[a]Report-drafting times are averaged over radiologists' sex, working years, and job title.

[b]Control group refers to CT image reports of nonlung sites that were not subject to the launch of the AI system.

[c]Not applicable.

[d]Experimental group refers to lung CT image reporting, which could be affected by AI assistance.

[e]DID: difference-in-differences (comparing changes in report-drafting time for lung vs nonlung CT images before and after AI implementation).

XSL•FO
RenderX

Subgroup analyses were conducted to account for the distinct implementation timelines of the AI systems at the 2 hospitals, as summarized in Figure 4 and Table 3. At Tsinghua Changgung Hospital, both the control group and the experimental group experienced extended report-drafting time, with a 0.90 minute (95% CI –0.28 to 2.08) nonsignificant increase attributable to AI assistance. However, Beijing Anzhen Hospital demonstrated an absolute 0.76-minute reduction in the AI-assisted group by the first year and a further 1.83-minute reduction by the second year. Using a DID framework, this corresponded to a 2.66-minute relative improvement vs the counterfactual trend (P<.001) 2 years into the AI implementation.

Figure 4. Adjusted changes in report-drafting time before and after artificial intelligence (AI) implementation. Gray indicates phases without AI assistance, and blue indicates phases with AI assistance. Tsinghua Changgung Hospital does not have second-year post-AI implementation data because the system was adopted later at that site. Report-drafting times are averaged over radiologist sex, years in practice, and job title. Difference-in-differences estimates compare changes in report drafting time for lung vs nonlung computed tomography reports before and after AI implementation. DID: difference-in-differences.

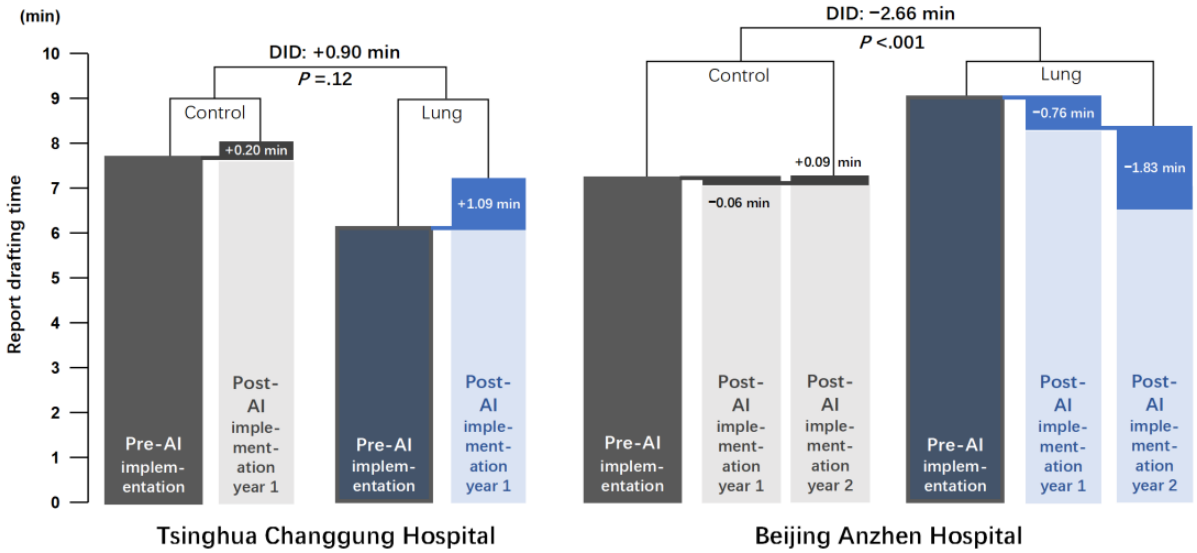

Table 3. Differences in adjusted report-drafting time comparing chest computed tomography (CT) vs others by post–artificial intelligence (AI) adoption duration.

	Adjusted report-drafting time[a] before AI implementation (min; 95% CI)	Adjusted report-drafting time[a] in the first year after AI implementation (min; 95% CI)	Adjusted report-drafting time[a] in the second year after AI implementation (min; 95 % CI)
Tsinghua Changgung			
Control group[b]	7.70 (6.61 to 8.79)	7.90 (6.26 to 9.53)	__[c]
Experimental group[d]	6.09 (4.69 to 7.59)	7.18 (5.38 to 8.99)	—
DID[e]	Reference	0.90 (–0.28 to 2.08)	—
Beijing Anzhen			
Control group[b]	7.36 (5.91 to 8.81)	7.30 (5.78 to 8.82)	7.39 (5.22 to 9.56)
Experimental group	9.22 (7.82 to 10.63)	8.46 (6.97 to 9.95)	6.63 (4.75 to 8.51)
DID	Reference	–0.70 (–1.35 to –0.05)	−2.66 (−3.35 to −1.96)

[a]Report-drafting times are averaged over radiologists' sex, working years, and job title.

[b]Control group refers to CT image reports of nonlung sites that were not subject to the launch of the AI system.

[c]Not available.

[d]Experimental group refers to lung CT image reporting, which could be affected by AI assistance.

[d]DID: difference-in-differences (comparing changes in report-drafting time for lung vs nonlung CT images before and after AI implementation).

XSL•FO
RenderX

Discussion

Adaptation of Diagnostic AI Systems in the Early Stages

Our findings indicate that the implementation of the AI system for lung nodule diagnostics was initially associated with an increase in report-drafting time, particularly in the first year of use. This transitional phase likely reflects a necessary adaptation period, as radiologists familiarized themselves with the system and learned to integrate it into existing workflows.

Technical barriers, such as outdated hospital IT infrastructure and limited interoperability with electronic medical record systems, may also impede seamless AI integration [23]. A potentially more important factor influencing the short-term impact of the AI system on lung nodule diagnosis may be physicians' self-efficacy. In the initial stages of the AI system's implementation, radiologists, especially junior staff, often lacked previous experience working with the AI system and were cautious in interpreting or relying on AI-generated results. This could partially explain the initial reporting delays observed at Tsinghua Changgung Hospital compared with the more experienced radiology team at Beijing Anzhen Hospital. Previous studies have shown that when radiologists lack confidence in the AI system's capabilities or have not yet fully acclimated to its operation, they might spend additional time validating AI outputs, rechecking discordant findings, or temporarily reverting to traditional workflows. This transitional phase can offset the efficiency gains intended by the AI system [24,25]. Limited formal training, reliance on trial-and-error learning, and incomplete understanding of deep learning mechanisms can further undermine early trust and efficiency [26-28].

However, over time, these early barriers diminished with routine use. A previous survey conducted at the same 2 hospitals showed strong user acceptance: radiologists rated the AI system highly in usability (System Usability Scale score: mean 74.3, SD 11.9) and efficiency improvement (mean 4.6, SD 0.6 out of 5) [29]. This growing user confidence likely reduced the need for repeated result validation and enabled more streamlined integration of the system, ultimately contributing to the substantial efficiency gains observed in later stages, such as the significant reduction in report-drafting time during the second year at Beijing Anzhen Hospital.

A related observation at Tsinghua Changgung Hospital further supports the interpretation of early-stage friction. During the first year, the control group did not exhibit a statistically significant change in report-drafting time, whereas the AI group experienced a meaningful increase. We confirmed with the hospital that no workflow reorganization, software upgrades, or staffing changes occurred during the study period. Therefore, the relative inefficiency observed in the AI arm at Tsinghua Changgung Hospital is unlikely to reflect systemic improvements in the control arm; instead, it aligns with our proposed mechanism that AI systems initially introduce cognitive load and workflow disruption. In contrast, Beijing Anzhen Hospital did not demonstrate a comparable degree of early inefficiency among its AI users. We hypothesize that this might be due to the higher baseline experience and seniority of its radiologists (Figure 3), which may facilitate faster adaptation to AI-assisted workflows.

These findings underscore the importance of considering the user characteristics and testing for potential learning curve when evaluating the impact of AI systems. Early performance metrics may underestimate their long-term value if insufficient time is allowed for user adaptation. If the initial learning curve does exist, structured training and supportive onboarding environments can help accelerate this process and maximize clinical benefits [30].

Long-Term Benefits of Diagnostic AI Systems

The long-term benefits of AI for diagnostic efficiency are complex and may vary significantly across institutions. While the pooled analysis suggests an overall increase in report-drafting time, this averaged effect masks the true clinical picture, where one hospital showed no immediate efficiency gains, while the other demonstrated a significant reduction in report-drafting time.

In high-volume radiology departments where radiologists face overwhelmingly heavy workloads over prolonged periods [31], information overload can lead to diagnostic errors [32,33]. Under such circumstances, even modest efficiency gains are valuable. In our study, extended use of the AI system for lung nodule analysis was associated with a reduction in average report-drafting time of up to 28% at one site. The AI system's inherent capacity to autonomously identify and measure lesions markedly reduces the burden of manual labor while also expediting the generation of standardized text reports. These collective features of the AI system, in turn, reduce the necessity of repetitive writing tasks [18]. In settings where radiologists typically read more than 50 scans per day, such an improvement amounts to 2 hours of saved time per radiologist daily. This could allow for higher throughput, shorter patient waiting times, and more time allocated to complex cases.

Notably, these efficiency gains emerged alongside previously established diagnostic benefits. Our previous work evaluating the same AI system demonstrated improved lung nodule detection with a modest early rise in false positives, underscoring the need for training and adaptation [34]. That analysis further showed stable diagnostic accuracy after senior review, suggesting no substantial downstream burden attributable to AI-assisted drafting. These findings provide important clinical context for this study, which shifts the focus from diagnostic performance to operational efficiency. While this analysis focuses on first-round report drafting by junior radiologists, the combined evidence suggests that AI adoption does not merely shift workload upstream but could integrate efficiently into routine clinical workflows.

Future Directions

While our study demonstrates the potential of AI to enhance efficiency within high-volume tertiary hospitals, its findings represent a necessary first step rather than a final answer. The successful deployment in these settings provides a proof of concept, but the ultimate goal is to improve health care resource allocation across the entire system. In terms of application

XSL•FO
RenderX

stages, future mixed methods investigations are necessary to identify institutional and user-level factors that may shape AI adoption trajectories. With regard to clinical settings, future research must directly examine AI's applicability in resource-constrained environments, which are often the hardest hit by medical shortages; first, partnering with primary and county-level hospitals as they begin to adopt AI; second, investigating the distinct barriers to adoption and effectiveness in low-resource settings (eg, outdated IT infrastructure, insufficient training, and different disease mixes); and third, examining how AI might be repurposed for tasks such as fast triage or ruling out common conditions in primary care, rather than for handling complex cases. Furthermore, the evaluation and validation of AI systems represent an iterative and dynamic process, necessitating a life cycle–based regulatory framework [35,36].

Strengths and Limitations

Our study has several strengths. First, we conducted a longitudinal analysis of data from 2 tertiary hospitals in China, covering both pre- and post-AI implementation periods in real-world settings. The inclusion of a larger sample size enhances the reliability of our conclusions. Second, we applied an adjusted DID design, a quasi-experimental approach increasingly used in health services research, which strengthened the rigor of examining the causal effect of AI's clinical impact. Finally, we analyzed report-drafting times by the same radiologists who interpreted both chest and nonchest CTs across the entire study period. This ensured that seasonal workload shocks, if any, would be shared across modalities within individuals, thus not differentially impacting the experimental and the control arm.

Nevertheless, our study had several limitations. First, both hospitals were large tertiary urban hospitals with high workloads and complex cases. As a result, our findings are considered only proof of concept, suggesting AI's potential to alleviate efficiency bottlenecks in saturated diagnostic workflows. Our conclusions may not generalize to primary or rural hospitals with different infrastructural and caseload challenges. Second, although we conducted extensive sensitivity analyses to confirm longitudinal stability in case mix, case complexity, and preintervention trends, the use of nonchest CT examinations as controls during a pandemic-affected period may only serve as imperfect controls. While we adjusted for key physician-level characteristics, the DID model may have omitted unmeasured confounders. However, the exclusion of unstable periods, inclusion of hospital and seasonal fixed effects, and the staggered rollout design lend support to the internal validity of our results. Third, while the different AI vendors deployed at the 2 hospitals had comparable diagnostic functions, variation in their workflows and interfaces may have affected the drafting time. Finally, implementing a triple-difference design was not feasible in this study. We concur that future research incorporating data from nonadopting or primary care sites is a critical next step.

Conclusions

Our findings highlight the heterogeneous impact of diagnostic AI on radiologists' efficiency. While short-term implementation was associated with an increase in report-drafting time in one hospital with a shorter follow-up period, sustained use was associated with significant efficiency gains in another hospital after more than 2 years of adoption. These divergent outcomes underscore that the realization of AI-driven efficiency is not guaranteed but is critically contingent on site-specific implementation dynamics and local workflow integration. Therefore, adequate adaptation and education are necessary for AI systems to meaningfully alleviate radiologists' workloads in the long run.

Acknowledgments

The authors sincerely thank all participating radiologists from Tsinghua Changgung Hospital and Beijing Anzhen Hospital for their invaluable contributions to this study. The authors are also grateful to the IT departments at both institutions for their technical support and assistance with data extraction and system integration. During the preparation of this manuscript, the authors used ChatGPT (OpenAI) to polish the English language and improve readability. After this, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

YW received funding from the Beijing Nova Program (grant 20250484908).

Data Availability

The datasets generated or analyzed during this study are available from the lead and the corresponding author on reasonable request. A sample dataset with the same structure as the original data, but with anonymized, masked, and sequence-perturbed entries for privacy protection, is provided in Multimedia Appendix 4.

Authors' Contributions

Conceptualization: WL, YW, MJB, ZZ, HK
Methodology: WL, YW
Investigation: WL, WY
Formal analysis: WL, YW

XSL•FO
RenderX

Data curation: WL, WY
Visualization: YW
Resources: ZZ
Supervision: HK
Writing—original draft: WL, YW
Writing—review and editing: ZZ, HK
WL and YW contributed equally to this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
User interface of the artificial intelligence–assisted system under study.
[PNG File , 684 KB - jmir_v28i1e77967_app1.png]

Multimedia Appendix 2
Representative photographs of the artificial intelligence–assisted computed tomography reporting workflow.
[PNG File , 1035 KB - jmir_v28i1e77967_app2.png]

Multimedia Appendix 3
Comparison of artificial intelligence system use workflows between the 2 hospitals.
[PNG File , 314 KB - jmir_v28i1e77967_app3.png]

Multimedia Appendix 4
Sample dataset.
[XLSX File (Microsoft Excel File), 42 KB - jmir_v28i1e77967_app4.xlsx]

Multimedia Appendix 5
Parallel trend assessment.
[PNG File , 97 KB - jmir_v28i1e77967_app5.png]

Multimedia Appendix 6
Sensitivity analysis using the leave-one-category-out approach.
[PNG File , 112 KB - jmir_v28i1e77967_app6.png]

References

1. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism 2017 Apr;69S:S36-S40. [doi: 10.1016/j.metabol.2017.01.011] [Medline: 28126242]
2. Alexander A, Jiang A, Ferreira C, Zurkiya D. An intelligent future for medical imaging: a market outlook on artificial intelligence for medical imaging. J Am Coll Radiol 2020 Jan;17(1 Pt B):165-170 [FREE Full text] [doi: 10.1016/j.jacr.2019.07.019] [Medline: 31918875]
3. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ. Artificial intelligence in radiology. Nat Rev Cancer 2018 Aug;18(8):500-510 [FREE Full text] [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
4. McDonald RJ, Schwartz KM, Eckel LJ, Diehn FE, Hunt CH, Bartholmai BJ, et al. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. Acad Radiol 2015 Sep;22(9):1191-1198. [doi: 10.1016/j.acra.2015.05.007] [Medline: 26210525]
5. Smith-Bindman R, Miglioretti DL, Johnson E, Lee C, Feigelson HS, Flynn M, et al. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996-2010. JAMA 2012 Jun 13;307(22):2400-2409 [FREE Full text] [doi: 10.1001/jama.2012.5960] [Medline: 22692172]
6. Liu Z, Yao Q, Wang P, Shen L, Li H, Zhang J, et al. The price of pressure: nationwide survey on lifestyle disturbances, occupational burnout and compromised perceived-competency among radiology residents in China. Front Public Health 2024 Oct 23;12:1472397 [FREE Full text] [doi: 10.3389/fpubh.2024.1472397] [Medline: 39507651]
7. Zhan H, Schartz K, Zygmont ME, Johnson JO, Krupinski EA. The impact of fatigue on complex CT case interpretation by radiology residents. Acad Radiol 2021 Mar;28(3):424-432 [FREE Full text] [doi: 10.1016/j.acra.2020.06.005] [Medline: 32622748]
8. Yip W, Fu H, Chen AT, Zhai T, Jian W, Xu R, et al. 10 years of health-care reform in China: progress and gaps in universal health coverage. Lancet 2019 Sep 28;394(10204):1192-1204. [doi: 10.1016/S0140-6736(19)32136-1] [Medline: 31571602]

XSL•FO
RenderX

9. Kong X, Ai B, Kong Y, Su L, Ning Y, Howard N, et al. Artificial intelligence: a key to relieve China's insufficient and unequally-distributed medical resources. Am J Transl Res 2019 May 15;11(5):2632-2640 [FREE Full text] [Medline: 31217843]
10. Wenderott K, Krups J, Weigl M, Wooldridge AR. Facilitators and barriers to implementing AI in routine medical imaging: systematic review and qualitative analysis. J Med Internet Res 2025 Jul 21;27:e63649 [FREE Full text] [doi: 10.2196/63649] [Medline: 40690758]
11. Bonekamp D, Schlemmer HP. [Artificial intelligence (AI) in radiology?: Do we need as many radiologists in the future?]. Urologe A 2022 Apr;61(4):392-399. [doi: 10.1007/s00120-022-01768-w] [Medline: 35277758]
12. Waller J, O'Connor A, Rafaat E, Amireh A, Dempsey J, Martin C, et al. Applications and challenges of artificial intelligence in diagnostic and interventional radiology. Pol J Radiol 2022 Feb 25;87:e113-e117 [FREE Full text] [doi: 10.5114/pjr.2022.113531] [Medline: 35280945]
13. He C, Liu W, Xu J, Huang Y, Dong Z, Wu Y, et al. Efficiency, accuracy, and health professional's perspectives regarding artificial intelligence in radiology practice: a scoping review. iRadiology 2024 Apr 08;2(2):156-172. [doi: 10.1002/ird3.63]
14. Yi C, Tang Y, Ouyang R, Zhang Y, Cao Z, Yang Z, et al. The added value of an artificial intelligence system in assisting radiologists on indeterminate BI-RADS 0 mammograms. Eur Radiol 2022 Mar;32(3):1528-1537. [doi: 10.1007/s00330-021-08275-0] [Medline: 34528107]
15. Stephan D, Bertsch A, Burwinkel M, Vinayahalingam S, Al-Nawas B, Kämmerer PW, et al. AI in dental radiology-improving the efficiency of reporting with ChatGPT: comparative study. J Med Internet Res 2024 Dec 23;26:e60684 [FREE Full text] [doi: 10.2196/60684] [Medline: 39714078]
16. Zia A, Fletcher C, Bigwood S, Ratnakanthan P, Seah J, Lee R, et al. Retrospective analysis and prospective validation of an AI-based software for intracranial haemorrhage detection at a high-volume trauma centre. Sci Rep 2022 Nov 18;12(1):19885 [FREE Full text] [doi: 10.1038/s41598-022-24504-y] [Medline: 36400834]
17. Pan F, Li L, Liu B, Ye T, Li L, Liu D, et al. A novel deep learning-based quantification of serial chest computed tomography in Coronavirus disease 2019 (COVID-19). Sci Rep 2021 Jan 11;11(1):417 [FREE Full text] [doi: 10.1038/s41598-020-80261-w] [Medline: 33432072]
18. Qi LL, Wu BT, Tang W, Zhou LN, Huang Y, Zhao SJ, et al. Long-term follow-up of persistent pulmonary pure ground-glass nodules with deep learning-assisted nodule segmentation. Eur Radiol 2020 Feb;30(2):744-755. [doi: 10.1007/s00330-019-06344-z] [Medline: 31485837]
19. Tan HM, Tan MS, Chang ZY, Tan KT, Ee GL, Ng CC, et al. The impact of COVID-19 pandemic on the health-seeking behaviour of an Asian population with acute respiratory infections in a densely populated community. BMC Public Health 2021 Jun 23;21(1):1196 [FREE Full text] [doi: 10.1186/s12889-021-11200-1] [Medline: 34158019]
20. Sample Size Calculators for designing clinical research. Sample Size Calculators. URL: https://sample-size.net/ [accessed 2026-03-03]
21. Ahn JS, Ebrahimian S, McDermott S, Lee S, Naccarato L, Di Capua JF, et al. Association of artificial intelligence-aided chest radiograph interpretation with reader performance and efficiency. JAMA Netw Open 2022 Aug 01;5(8):e2229289 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.29289] [Medline: 36044215]
22. Wing C, Simon K, Bello-Gomez RA. Designing difference in difference studies: best practices for public health policy research. Annu Rev Public Health 2018 Apr 01;39(1):453-469 [FREE Full text] [doi: 10.1146/annurev-publhealth-040617-013507] [Medline: 29328877]
23. Li J, Zhou L, Zhan Y, Xu H, Zhang C, Shan F, et al. How does the artificial intelligence-based image-assisted technique help physicians in diagnosis of pulmonary adenocarcinoma? A randomized controlled experiment of multicenter physicians in China. J Am Med Inform Assoc 2022 Nov 14;29(12):2041-2049 [FREE Full text] [doi: 10.1093/jamia/ocac179] [Medline: 36228127]
24. Brown M, Browning P, Wahi-Anwar MW, Murphy M, Delgado J, Greenspan H, et al. Integration of chest CT CAD into the clinical workflow and impact on radiologist efficiency. Acad Radiol 2019 May;26(5):626-631. [doi: 10.1016/j.acra.2018.07.006] [Medline: 30097402]
25. Müller FC, Raaschou H, Akhtar N, Brejnebøl M, Collatz L, Andersen MB. Impact of concurrent use of artificial intelligence tools on radiologists reading time: a prospective feasibility study. Acad Radiol 2022 Jul;29(7):1085-1090 [FREE Full text] [doi: 10.1016/j.acra.2021.10.008] [Medline: 34801345]
26. Ooi SK, Makmur A, Soon AY, Fook-Chong S, Liew C, Sia SY, et al. Attitudes toward artificial intelligence in radiology with learner needs assessment within radiology residency programmes: a national multi-programme survey. Singapore Med J 2021 Mar;62(3):126-134 [FREE Full text] [doi: 10.11622/smedj.2019141] [Medline: 31680181]
27. Abuzaid MM, Elshami W, Tekin H, Issa B. Assessment of the willingness of radiologists and radiographers to accept the integration of artificial intelligence into radiology practice. Acad Radiol 2022 Jan;29(1):87-94. [doi: 10.1016/j.acra.2020.09.014] [Medline: 33129659]
28. Bergquist M, Rolandsson B, Gryska E, Laesser M, Hoefling N, Heckemann R, et al. Trust and stakeholder perspectives on the implementation of AI tools in clinical radiology. Eur Radiol 2024 Jan;34(1):338-347 [FREE Full text] [doi: 10.1007/s00330-023-09967-5] [Medline: 37505245]

XSL•FO
RenderX

29. Liu W, Wu Y, Zheng Z, Yu W, Bittle MJ, Kharrazi H. Evaluating artificial intelligence's role in lung nodule diagnostics: a survey of radiologists in two pilot tertiary hospitals in China. J Clin Imaging Sci 2024 Aug 23;14:31. [doi: 10.25259/JCIS_72_2024] [Medline: 39246733]
30. Sun L, Yin C, Xu Q, Zhao W. Artificial intelligence for healthcare and medical education: a systematic review. Am J Transl Res 2023 Jul 15;15(7):4820-4828 [FREE Full text] [Medline: 37560249]
31. Gampala S, Vankeshwaram V, Gadula SS. Is artificial intelligence the new friend for radiologists? A review article. Cureus 2020 Oct 24;12(10):e11137 [FREE Full text] [doi: 10.7759/cureus.11137] [Medline: 33240726]
32. Paul S, Nazareth DL. Input information complexity, perceived time pressure, and information processing in GSS-based work groups: an experimental investigation using a decision schema to alleviate information overload conditions. Decis Support Syst 2010 Apr;49(1):31-40 [FREE Full text] [doi: 10.1016/j.dss.2009.12.007]
33. Vosshenrich J, Brantner P, Cyriac J, Boll DT, Merkle EM, Heye T. Quantifying radiology resident fatigue: analysis of preliminary reports. Radiology 2021 Mar;298(3):632-639. [doi: 10.1148/radiol.2021203486] [Medline: 33497316]
34. Liu W, Wu Y, Zheng Z, Bittle M, Yu W, Kharrazi H. Enhancing diagnostic accuracy of lung nodules in chest computed tomography using artificial intelligence: retrospective analysis. J Med Internet Res 2025 Jan 27;27:e64649 [FREE Full text] [doi: 10.2196/64649] [Medline: 39869890]
35. Liu Y, Chen PH, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. JAMA 2019 Nov 12;322(18):1806-1816. [doi: 10.1001/jama.2019.16489] [Medline: 31714992]
36. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021 Jul 15;385(3):283-286 [FREE Full text] [doi: 10.1056/NEJMc2104626] [Medline: 34260843]

Abbreviations

AI: artificial intelligence
CT: computed tomography
DID: difference-in-differences
IRB: institutional review board

Edited by A Coristine; submitted 23.May.2025; peer-reviewed by J Jiang, J Zhang; comments to author 17.Jun.2025; revised version received 13.Feb.2026; accepted 14.Feb.2026; published 27.Mar.2026.

Please cite as:
Liu W, Wu Y, Yu W, Bittle MJ, Zheng Z, Kharrazi H
Measuring the Impact of AI on Report-Drafting Efficiency in Chest Computed Tomography Interpretation: Retrospective Analysis
J Med Internet Res 2026;28:e77967
URL: https://www.jmir.org/2026/1/e77967
doi:10.2196/77967
PMID:

©Weiqi Liu, You Wu, Wei Yu, Mark J Bittle, Zhuozhao Zheng, Hadi Kharrazi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Transfer Learning and Machine Learning for Training Five-Year Survival Prognostic Models in Early Breast Cancer: Development and Validation Study

Lisa Pilgram[1,2,3], MD; Kai Yang[1,2], PhD; Ana-Alicia Beltran-Bless[4], MD; Gregory R Pond[5], PhD; Lisa Vandermeer[6], MSc; John Hilton[7], MD; Marie-France Savard[4], MD; Andreanne LeBlanc[8], MD; Lois Shepherd[9], MD; Bingshu Chen[9], PhD; John MS Bartlett[10], PhD; Karen J Taylor[10], PhD; Jane Bayani[11,12], PhD; Sarah Barker[11], PhD; Melanie Spears[11,12], PhD; Cornelis JH van der Velde[13], MD; Elma Meershoek-Klein Kranenbarg[13], MSc; Luc Dirix[14], MD, PhD; Elizabeth Mallon[15], MBChB; Annette Hasenburg[16], MD; Christos Markopoulos[17], MD, PhD; Lamin Juwara[1,2], PhD; Fida K Dankar[2], PhD; Mark Clemons[7], MD; Khaled El Emam[1,2], PhD

[1]School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada
[2]Research Institute, Children's Hospital of Eastern Ontario, Ottawa, ON, Canada
[3]Department of Nephrology and Medical Intensive Care, Charité - Universitaetsmedizin Berlin, Berlin, Germany
[4]Division of Medical Oncology, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada
[5]Department of Oncology, McMaster University, Hamilton, ON, Canada
[6]Cancer Therapeutics Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada
[7]Ottawa Hospital Cancer Center, Ottawa Hospital Research Institute, Ottawa, ON, Canada
[8]Division of Medical Oncology and Hematology, CHUM, Université de Montréal, Montreal, QC, Canada
[9]Canadian Canadian Cancer Trials Group, Queen's University, Kingston, ON, Canada
[10]Edinburgh Cancer Research, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, Scotland, United Kingdom
[11]Diagnostic Development, Ontario Institute for Cancer Research, Toronto, ON, Canada
[12]Department of Laboratory Medicine and Pathobiology, University of Toronto, Ontario, ON, Canada
[13]Department of Surgery, Leiden University Medical Center, Leiden, South Holland, The Netherlands
[14]St. Augustinus Hospital, Antwerp, Belgium
[15]Department of Pathology, NHS Greater Glasgow and Clyde, Glasgow, Scotland, United Kingdom
[16]Department of Gynecology and Obstetrics, University Center Mainz, Mainz, Germany
[17]Medical School, National and Kapodistrian University of Athens, Athens, Attica, Greece

Corresponding Author:
Khaled El Emam, PhD
School of Epidemiology and Public Health
Faculty of Medicine
University of Ottawa
451 Smyth Rd
Ottawa, ON, K1H 8M5
Canada
Phone: 1 6135625700
Email: kelemam@ehealthinformation.ca

Abstract

Background: Prognostic information is essential for decision-making in breast cancer management. In recent years, trials and clinical practice have emphasized genomic prognostication tools, despite clinicopathological methods being more affordable and accessible. PREDICT v3 is one such tool with promising results across cohorts. Advances in machine learning (ML), transfer learning, and ensemble methods provide opportunities to enhance these approaches, especially where missing data and model assumptions differ across diverse populations.

Objective: This study evaluates the potential to improve survival prognostication in breast cancer. More precisely, we compare de novo ML, transfer learning from the pretrained prognostication model PREDICT v3, and a stacked ensemble approach.

XSL•FO
RenderX

Methods: Data from the MA.27 trial (NCT00066573) were used for model training, with external validation on data from the Tamoxifen Exemestane Adjuvant Multinational trial (NCT00279448 and NCT00032136) and a US Surveillance, Epidemiology, and End Results cohort. Transfer learning was applied by re-estimating the parameters of the pretrained prognostic tool PREDICT v3. De novo ML included random survival forests and extreme gradient boosting, and the ensemble was implemented using weighted linear stacking of model predictions. Internal and external validation was assessed in terms of the integrated calibration index and discrimination. Shapley Additive Explanations values were used to explain model predictions and decision-curve analysis to facilitate the interpretation of performance differences.

Results: Transfer learning, de novo random survival forest, and the stacked ensemble improved calibration in MA.27 over the pretrained model (integrated calibration index reduced from 0.042 in PREDICT v3 to ≤0.007) while discrimination remained comparable (AUROC increased from 0.738 in PREDICT v3 to 0.744-0.799). In decision-curve analysis, these approaches demonstrated consistently positive net benefit across clinically relevant thresholds, while PREDICT v3 lost net benefit beyond 7.5% predicted risk. Invalid PREDICT v3 predictions were observed in 23.8% to 25.8% of MA.27 individuals due to missing information. In contrast, ML models and the stacked ensemble predicted survival despite missing data. Across all models, patient age, nodal status, pathological grading, and tumor size had the highest Shapley Additive Explanations values, indicating their importance for survival prognostication. External validation in the US Surveillance, Epidemiology, and End Results cohort confirmed the benefits of transfer learning, RSF, and ensemble in terms of calibration while maintaining discrimination at comparable levels. In contrast, generalizability was limited in the Tamoxifen Exemestane Adjuvant Multinational trial, a cohort with a substantially different distribution of clinicopathological characteristics.

Conclusions: This study demonstrates that transfer learning, de novo RSF, and a stacked ensemble can improve prognostication compared with the pretrained PREDICT v3, particularly in the presence of missing or uncertain inputs. Transportability may be limited in cohorts with different clinicopathological profiles, requiring local validation before clinical deployment. Ultimately, better survival estimation can provide more meaningful guidance in breast cancer care.

Trial Registration: ClinicalTrials.gov NCT00066573; https://clinicaltrials.gov/study/NCT00066573, NCT00279448; https://clinicaltrials.gov/study/NCT00279448, NCT00032136; https://clinicaltrials.gov/study/NCT00032136

(J Med Internet Res 2026;28:e88665) doi:10.2196/88665

KEYWORDS

prognostic models; transfer learning; ensembles; breast cancer survival; machine learning

Introduction

Breast cancer is among the most common types of cancer worldwide. In 2022, there were 2.3 million women diagnosed with breast cancer globally [1], typically in a nonmetastatic stage at diagnosis [2]. Such early diagnosis allows for a broad range of treatment options, including surgery, radiation therapy, endocrine therapy, chemotherapy, and targeted systemic therapies. Estimating survival probabilities can support informed decision-making, especially in scenarios where multiple treatment options are available. This makes both prognostic information (a patient's survival) and predictive information (a patient's benefit from treatment) [3] highly valuable to guide patient management.

A large variety of clinicopathological and genomic risk assessment tools have been proposed to assist in clinical decision-making [3,4]. Other tools, for example, RSClin (provided by Exact Sciences), have been developed to improve prognosis by incorporating both clinicopathological and genomic information [5].

Trials such as Microarray In Node negative Disease may Avoid Chemotherapy [6] confirmed the importance but also the challenge of breast cancer risk assessment tools. It demonstrated that patients classified as low risk by genomic prognostication had favorable survival outcomes without chemotherapy. However, patients classified as high risk by genomic prognostication did not necessarily benefit from chemotherapy, particularly in discordant cases where clinicopathological risk was. Similarly, the recently published Adjuvant Systemic Treatment for (ER)-Positive HER2-negative Breast Carcinoma in Women Over 70 trial showed that women aged 70 years and older who had a high risk by genomic prognostication did not benefit from the addition of adjuvant chemotherapy to endocrine therapy in terms of survival [7]. Both trials underscore the limitations of current genomic prognostication tools in predicting treatment benefit and reinforce the critical distinction between prognostic and predictive information [3,8].

Despite these findings, genomic testing and prognostication have become routine even for clinicopathologically low-risk patients, diverging from clinical guidelines recommendations and thereby potentially resulting in overtreatment [3,5-7,9]. Also, genomic testing can come with delayed treatment decisions and considerable costs, limiting its accessibility in resource-constrained health care settings [10].

Importantly, recent initiatives have focused primarily on genomic prognostication tools. This is true for the clinical trials mentioned above but also for updates of clinicopathological tools that are used in practice with genomic signatures, even though the added value was found to be modest [11]. There has been relatively less interest in the refinement of clinicopathological prognostication [9], even though such solutions are inexpensive, easily accessible, and can support clinical decision-making. Furthermore, machine learning (ML) and deep learning approaches have been explored to enhance

XSL•FO
RenderX

performance in survival prognostication with promising results but limited applicability in practice [12-14] (see section 1.2 in Multimedia Appendix 1).

A compelling opportunity also lies in leveraging validated pretrained models such as PREDICT v3 [15] and adapting them to new datasets [16]. In transfer learning, the idea is that a new task can be more effectively learned by transferring knowledge from a related task that has already been learned. There are multiple transfer learning approaches (see section 1.3 in Multimedia Appendix 1). Among them, parameter-based transfer learning (ie, fine-tuning) is the most useful as it does not require access to the original training datasets but instead fine-tunes the parameters of the pretrained model to the new data [16]. While such transfer learning is not yet widely adopted in survival analysis, there are promising results from lung cancer and pancreatic adenocarcinoma survival prognostication [17,18].

Complementary to de novo ML and transfer learning, ensemble integration offers a way to account for model-specific strengths and limitations by combining multiple models, such as fine-tuned and de novo trained ones [19]. This can ultimately provide a robust prognostication framework, particularly in cases where missingness and model assumptions vary across cohorts.

This study aims to better understand the benefits of transfer learning from pretrained models, de novo ML, and a stacked ensemble in survival prognostication for patients with breast cancer. More precisely, using the MA.27 study population [20], we investigated the following research questions:

- Parameter-based transfer learning (ie, fine-tuning): Can fine-tuning the pretrained prognostic tool PREDICT v3 to the MA.27 dataset improve survival prediction performance compared to the pre-trained model alone?
- De novo ML: How do state-of-the-art ML models trained directly on the MA.27 dataset compare against the (fine-tuned) pretrained model PREDICT v3?
- Ensemble integration: Does a stacked ensemble of fine-tuned pretrained models and de novo ML models add benefit compared to either approach alone?
- Generalizability: Do the potential benefits from fine-tuning (ie, question 1), de novo ML (ie, question 2), and the stacked ensemble (ie, question 3) still hold in external cohorts?

The primary clinical use case of the proposed model is to provide information that can support clinicians and patients in clinical management. Clinicians already use survival probabilities in discussions on the potential benefit of adjuvant systemic treatments following primary surgery [6,21-24], and patients also find such information important for their decision-making processes [24,25]. Beyond treatment decisions, individualized risk estimates can also inform follow-up schedules [26].

This study is methodological in nature and aims to establish and evaluate modeling strategies that can form the foundation for developing and improving prognostic tools for such clinical use. The intended user of such a model depends on its presentation: it can be an informative tool for a clinician, or, if carefully embedded within appropriate explanatory and interpretative guidance, as a resource for patients to prepare for discussions with their care team.

Ultimately, better survival estimation can provide meaningful guidance in breast cancer management, supporting a more targeted, cost-effective, and personalized approach to breast cancer care.

Methods

Reporting Guidelines

This study was conducted and reported in alignment with the Consolidated Reporting Guidelines for Prognostic and Diagnostic Machine Learning Models [27]. A completed Consolidated Reporting Guidelines for Prognostic and Diagnostic Machine Learning Models checklist is provided in Multimedia Appendix 2.

Data Sources

This study is a secondary analysis of existing datasets from the MA.27 clinical trial [20], the Surveillance, Epidemiology, and End Results (SEER) program [28], and the Tamoxifen Exemestane Adjuvant Multinational (TEAM) trial [29], focusing on the prediction of 5-year breast-cancer survival. These datasets, their primary purpose, and their use across the different stages of this study are summarized in Table 1.

XSL•FO
RenderX

Table 1. Data sources.

Data	Description	Total sample size (N)	Used for		
			Model development, n[a]	Internal validation, n	External validation, n
MA.27[b] [20]	Phase 3 randomized clinical trial comparing adjuvant hormone therapies in postmenopausal, hormone receptor–positive breast cancer (enrollment 2003-2008)	7563	6049	1514	—[c]
SEER[d] [28]	Population-based cancer registry, hormone receptor–positive breast cancer (diagnosed in 2003), selected to align with MA.27 eligibility	27,064	—	—	27,064
TEAM[e] [29]	Phase 3 randomized clinical trial of adjuvant hormone therapies in postmenopausal hormone receptor–positive breast cancer (enrollment 2001-2006), previously described subcohort from the entire trial	3825	—	—	3825

[a]"n" refers to the size of the cohort that was used in our analyses.

[b]MA.27 was randomly split into model development (80%) and validation (20%) partitions using stratification by outcome. Due to integer rounding within strata, exact partition sizes differed slightly from the exact proportions.

[c]Not available.

[d]SEER: US Surveillance Epidemiology and End Results.

[e]TEAM: Tamoxifen Exemestane Adjuvant Multinational.

Detailed information on the trials, including recruitment procedures, data collection, and delivery of the intervention (ie, adjuvant hormone therapy), is available in the respective primary publications and is not repeated here.

MA.27 Study Cohort

MA.27 was a phase 3 clinical trial conducted by the Canadian Cancer Trials Group [20] of 7576 postmenopausal women with early-stage hormone receptor–positive breast cancer between 2003 and 2008, which compared 2 aromatase inhibitors, exemestane and anastrozole, as adjuvant endocrine therapy. The trial did not find a statistically significant difference in distant disease-free and disease-specific survival between the two arms. For the purpose of prognostic modeling, the cohort was therefore analyzed as a single population.

We limited the cohort to patients who were followed up beyond the day of enrollment (ie, time-to-event > 0) for our study to avoid biasing survival estimates with events unrelated to breast cancer. This subcohort consisted of 7563 patients. For this subcohort, variables were selected from MA.27 that overlapped with the information required for PREDICT v3 (Table 2; section 2.4 in Multimedia Appendix 1). This ensured alignment with variables considered clinically relevant and broadly available in the context of breast cancer prognostication. The outcome (ie, event) was defined as breast cancer–related death within a 5-year observation interval, and time-to-event or follow-up time was considered in the survival analyses.

XSL•FO
RenderX

Table 2. Variables selected for 5-year survival prediction.

Variables[a]	Explanation
Age	Age in years at randomization
Positive nodes	The number of positive lymph nodes
Tumor laterality	Side of tumor manifestation: right-handed, left-handed, or bilateral
Estrogen receptor status	Positive or negative estrogen receptor status
Progesterone receptor status	Positive or negative progesterone receptor status
Tumor size	The maximum size of the tumor in mm
Tumor grade	The pathological grading of the tumor from 1 (well-differentiated) to 3 (poorly differentiated)
Radiotherapy	Whether or not radiotherapy was received
Chemotherapy	Whether or not adjuvant chemotherapy was received
Trastuzumab therapy	Whether or not trastuzumab was received

[a]These variables represent the overlap between variables in MA.27 and those required for PREDICT v3. The variables required for PREDICT v3 but not directly available were year of diagnosis, smoking status, human epidermal growth factor receptor 2 status, Ki-67 status, mode of detection, micrometastases in case of one positive node, mean heart dose in case of radiotherapy, the type of chemotherapy in case of chemotherapy and bisphosphonate use. Some of them were mandatory for survival prognostication, and therefore, assumptions were made based on the standard of care at that time. For details, see section 2.4 in Multimedia Appendix 1.

Characteristics of MA.27 are described using median and IQR values for continuous variables and counts with percentages for categorical variables; descriptive visualizations (Kaplan-Meier curve, boxplots, and stacked bar plots) were generated to characterize the training cohort prior to modeling.

Data Management

MA.27 was highly imbalanced in terms of its outcome, meaning that there were only 187 (187/7563, 2.5%) recorded disease-related deaths across a median follow-up of 4.1 (IQR 3.6-4.8) years. We used 2 different rebalancing strategies: (1) random oversampling examples technique on a dataset level and (2) weighting techniques on an algorithm level to help with model training [30] (section 2.1 in Multimedia Appendix 1). We tested the effect of both strategies in model training but could not detect a beneficial effect, and therefore they were not considered in the main analyses (section 3.5 in Multimedia Appendix 1).

MA.27 further presented with missingness in some variables. However, the mechanism of missingness was unclear. A missing progesterone receptor status could be, for example, not missing at random if it reflected ambiguity in the pathological assessment, suggesting a potentially biologically meaningful pattern of missingness. However, it could also be missing at random if values were omitted due to documentation errors or data entry inconsistencies.

We conducted missingness analyses and explored model-based imputation for these variables in MA.27 (section 2.2 in Multimedia Appendix 1). Imputation did not meaningfully change model performance (section 3.3 in Multimedia Appendix 1). However, because a nonrandom missingness mechanism could not be ruled out, and imputation may introduce bias under such conditions, imputed data were not used for the primary analyses. Instead, we leveraged the abilities of a tree-based ML model to internally handle missing data via surrogate splits or default directions, as they can handle mixed types of missingness patterns [31].

As mentioned in Table 2 and detailed in section 2.4 in Multimedia Appendix 1, the overlap between the variables in MA.27 and those required for PREDICT v3 was not complete, such that some variable values were constructed based on the trial's metadata and relevant background knowledge. For example, HER2 status was inferred from trastuzumab use, and endocrine therapy from the inclusion criteria of MA.27. For other variables, no reliable approximation was possible, such that a fraction of patients for whom PREDICT v3 could not estimate survival remained. Sensitivity analyses assessing the impact of these assumptions on model performance are described in section 2.4 in Multimedia Appendix 1 and reported in section 3.4 in Multimedia Appendix 1.

Survival Models

PREDICT was originally fitted on 5232 breast cancer cases from the UK East Anglia Cancer Registration and Information Centre diagnosed between 1999 and 2003 [15], updated recently to PREDICT v3 with 38,909 patients diagnosed between 2000 and 2017 [32], and validated on several cohorts around the world [33,34]. However, MA.27 differs in certain aspects from these training and validation cohorts: MA.27 was collected in 2003, involved approximately 5 years of follow-up, included only postmenopausal patients with hormone receptor–positive breast cancer, and lacked some of the information required in PREDICT v3. This makes de novo ML and transfer learning particularly useful. Section 1 in Multimedia Appendix 1 gives supplemental background on commonly applied survival models, including PREDICT, de novo ML, and transfer learning.

Given the structure of the MA.27 dataset with predominantly categorical variables, high censoring, and limited sample size, we opted to use tree-based methods for ML survival modeling. Such models are good for handling categorical data, can internally account for mixed missingness types, and offer robust performance without the data demands or complexity of deep learning approaches.

More precisely, this study included the following models:

XSL•FO
RenderX

1. PREDICT v3: The pretrained survival model, a competing risks Cox survival model with fractional polynomial baseline cumulative hazards.
2. f-PREDICT v3: The pretrained survival model fine-tuned to MA.27 (ie, transfer learning).
3. Random survival forests (RSFs) [35]: An ensemble method tailored for survival analysis.
4. Extreme gradient boosting (XGB) [36]: A gradient boosting framework with a survival-specific loss function.
5. Ensemble [19]: A stacked ensemble integrating f-PREDICT v3, RSF, and XGB whereby final predictions are obtained as a weighted sum of the individual model predictions.

Fine-tuning in the context of this study refers to parameter-based transfer learning whereby the initial model parameters of PREDICT v3 were used as initialization values for re-estimating the regression coefficients on MA.27. This differs from deep-learning fine-tuning, where the term typically refers to adapting a learned representation through changes in multilayer network weights.

Details on our implementation are provided in section 2.3 in Multimedia Appendix 1.

Performance Measurement, Decision Curve, and Model Explainability

The primary goal of this study was to predict 5-year breast cancer survival, as this represents an early and clinically relevant milestone to guide decision-making [6,25]. Accordingly, performance metrics focused on this time point.

In internal and external validation, we measured performance for 5-year-survival prediction by area under the receiver operating characteristic (ROC) curve (AUROC) and integrated calibration index (ICI), and provided ROC curves and calibration plots for that time point. AUROC reflects a model's ability to distinguish between individuals who survive and those who do not (ie, discrimination). ICI is the average absolute difference between predicted survival probabilities and observed survival outcome, estimated from a smoothed calibration curve over the entire range of predictions [37]. All performance measurements accounted for the time-to-event nature of the data in order to obtain robust measures in the context of heavy censoring. For discrimination, inverse probability of censoring weighting was leveraged to adjust for right-censoring, whereby the Kaplan-Meier estimator was used to model the censoring distribution [38]. It was implemented via the R package *timeROC* (Blanche et al [39]). For calibration, the observed outcome was modeled by a hazard regression–based method as proposed by Austin et al [37].

In breast cancer survival prognostication [32,40] or more broadly in evaluating the improvement of the predictive ability of markers [41], changes in AUROC of 0.05 or less have sometimes been highlighted. However, our interpretation follows a literature review on interpreting AUROC in health care [42] whereby label changes were typically triggered by a change in AUROC of at least 0.1.

The interpretation of ICI is more challenging as no uniform guidance exists. In experiments by Austin et al [37,43], correctly specified models yielded ICI values below 0.0125 while incorrectly specified ones had higher values. Similarly, a difference of 0.01 in predicted probabilities was considered as relevant by a nontrivial portion (17.7%) of patients with early breast cancer [25]. However, clinical decision making is typically triggered by probability differences of 0.03 to 0.05 [4,25,40,44], and some authors consider predicted probabilities within 10% of the observed outcomes as well-calibrated [32]. We align our interpretation with clinical decision-making standards and consider models with an ICI below 0.03 as well-calibrated.

The predicted probabilities were pooled across 10 independent runs to plot ROC curves in the internal validation, and a diagonal dashed line was added to indicate the discrimination of a random guess. For calibration plots, the observed probabilities were divided into four quartiles based on their predicted probabilities. Both predicted and observed probabilities were trimmed to exclude extreme values beyond the 10th to 90th percentile, and mean predicted and observed survival probabilities were calculated in each quartile. Horizontal and vertical error bars are illustrated to reflect the SDs of these quartile-wise means. A diagonal dashed line was added to indicate perfect calibration. In the internal validation, predictions were pooled across 10 independent runs to plot calibration.

To facilitate interpretation of differences in calibration and discrimination, we conducted decision-curve analysis (DCA) for 5-year survival [45,46]. DCA evaluates the net benefit of using a risk model to trigger clinical decision-making at a given threshold probability P_t, compared to two default strategies: intervening in all patients or intervening in none. We deliberately use the term intervention to avoid narrow interpretation: this may represent a treatment decision, intensified follow-up, or additional diagnostics. Individuals with predicted 5-year mortality risk greater than or equal to P_t are considered candidates for a hypothetical intervention. This intervention is assumed to reduce the event probability but may be unnecessary in individuals who would not experience the event.

Because the outcome was time-to-event, net benefit at 5 years was estimated using the Kaplan-Meier method within the strata defined by the threshold choices, in line with the study by Vickers et al [45] and implemented via the R package *dcurves* (Sjoberg et al [47]). For each threshold P_t, individuals were classified as high risk if their predicted risk exceeded P_t. The cumulative event probability within the high-risk group was estimated using Kaplan-Meier, from which true and false positives were derived. Net benefit was calculated as:

where n is the size of the cohort, and the weighting term reflects the relative costs of a false positive compared to the benefits of a true positive implied by the chosen threshold. Thresholds between 1% and 10% for 5-year mortality were evaluated, reflecting clinically plausible ranges in early breast cancer. The net benefit analysis was conducted in the internal validation cohort to illustrate how differences in calibration and discrimination can have decision implications.

XSL•FO
RenderX

Shapley Additive Explanations (SHAP) is a post hoc method to explain model predictions based on game theory [48]. For each individual prediction, a SHAP value can be assigned to each variable that reflects the contribution of that variable to the prediction. This value is estimated by a Monte Carlo approximation strategy as suggested in Štrumbelj et al [48]. A model-agnostic SHAP approach was leveraged to ensure consistent interpretation across the 5 different models. It was implemented via the R package *iml* (Molnar et al [49]). SHAP is presented for all individuals from MA.27 in a summary plot, implemented via the R package *shapviz* (Mayer et al [50]).

Model Training, Testing, and Internal Validation

MA.27 was randomly split into three subsets: (1) 60% for training (data A), (2) 20% for testing (data B), and (3) 20% for final validation (data C).

As shown in Figure 1, data A was used to train ML models and to fine-tune PREDICT v3 (ie, transfer learning). The fine-tuned model is referred to as f-PREDICT v3. Fine-tuning was performed by adjusting the 26 parameters of PREDICT v3 through a local optimization approach to MA.27 (section 2.4 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 1. Model development, internal validation, and final model specification.

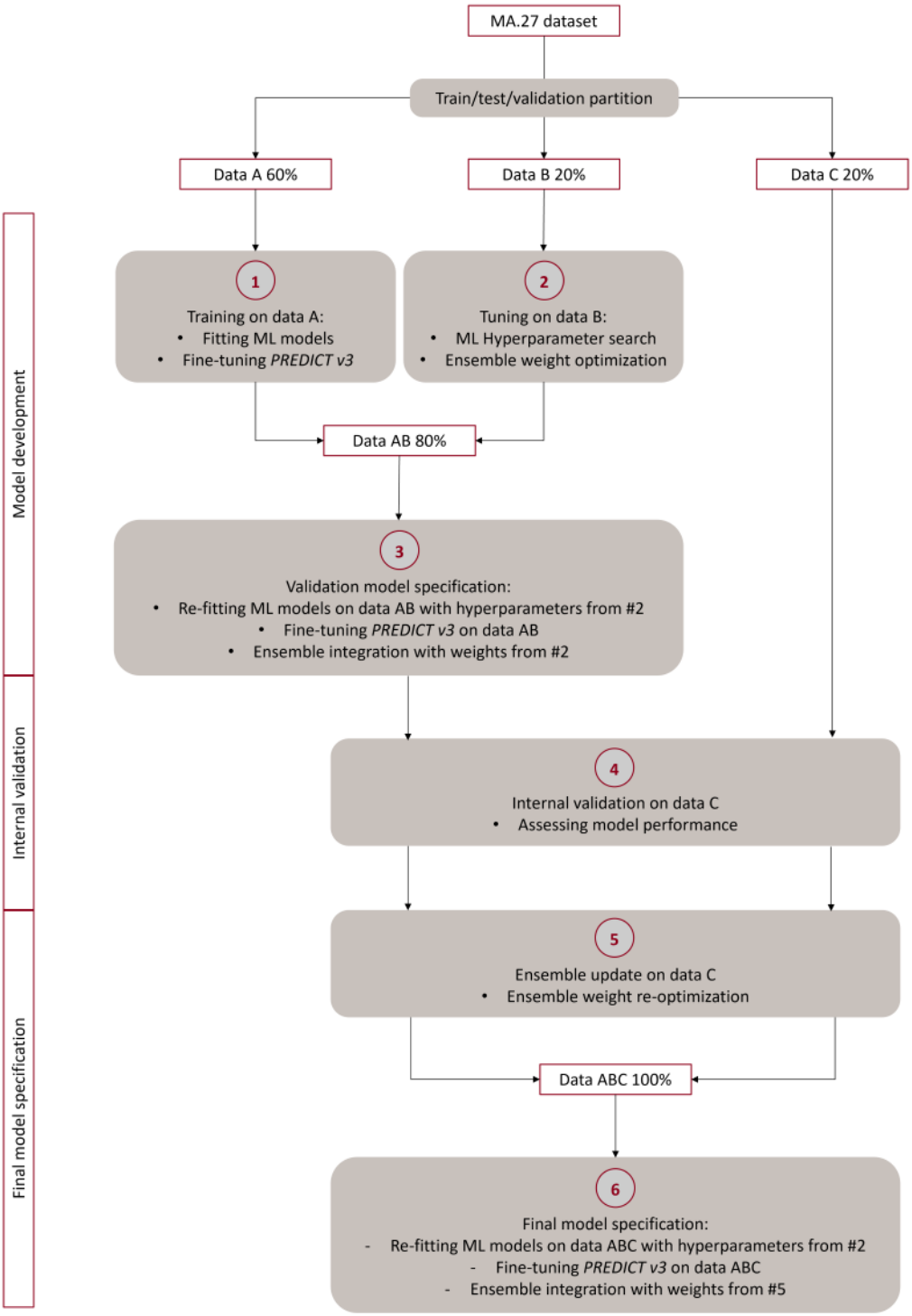

Data B was used to find the best hyperparameters for the ML models and to determine the weights for stacking RSF, XGB, and f-PREDICT v3 as an ensemble (for details, refer to #2 in Figure 1). Details on hyperparameter and ensemble integration are provided in section 2.3 in Multimedia Appendix 1.

Data C was used as the hold-out set for internal validation. Prior to internal validation, the ML models were refitted on the combined data A and B using the best hyperparameters selected on data B, and PREDICT v3 was re–fine-tuned on the combined data A and B (for details, refer to #3 in Figure 1). The ensemble was constructed using the weights selected on data B, with the models refitted on the combined data A and B. These models were then evaluated alongside the original pretrained PREDICT v3 on the hold-out validation set (ie, data C).

After internal evaluation, the weights for the stacked ensemble were reoptimized using data C based on predictions from the models trained on the combined data A and B (for details, refer to #5 in Figure 1). Finally, the ML models were refitted on the

XSL•FO
RenderX

full dataset (the combined data A, B, and C) using the previously selected hyperparameters, PREDICT v3 was re–fine-tuned, and the ensemble was constructed using the refitted models on the combined data A, B, and C and the optimized weights on data C (for details, refer to #6 in Figure 1).

During training, the ICI for 5-year survival prediction was used as the optimization goal. Calibration has clinically meaningful implications in prognostication tools as probability estimates typically guide decision-making. Discrimination (ie, AUROC), in contrast, focuses on ranking that may be more relevant for diagnostic tools. Other metrics, such as the mean absolute error, assess the accuracy of the predicted survival times, which, again, is different from our scenario where survival probabilities at certain time points are relevant for decision-making [51]. As a reference, an AUROC-based training approach was also conducted and is presented in section 3.6 in Multimedia Appendix 1. However, both calibration (ie, ICI) and discrimination (AUROC) were part of the evaluation.

To account for the variability in modeling across different partitions, all training, testing, and validation steps were repeated across 10 independent runs. The ML parameters were chosen by majority vote (ie, the parameter that was most often chosen across the 10 independent runs), the fine-tuned parameters for PREDICT v3, and the ensemble weights by averaging. We indicate the median and IQR values for the internal evaluation across the 10 independent runs.

To assess the robustness of the hyperparameter selection, we conducted a sensitivity analysis in which the variability of ML hyperparameters, of the fine-tuned PREDICT v3 parameters, and of the weights for the stacked ensemble was evaluated across a larger set of 50 independent runs. Results are reported in section 3.9 in Multimedia Appendix 1.

External Validation

The final models were externally validated against a cohort from the US Surveillance, Epidemiology, and End Results program [28] and a clinical trial cohort (TEAM) with patients from Belgium, France, Germany, Greece, Japan, the Netherlands, the United Kingdom and Ireland, and the United States [52].

In SEER [28], the cohort was selected to match the relevant eligibility criteria of MA.27, namely hormone receptor–positive postmenopausal women diagnosed in 2003. It was also adjusted to meet the requirements of PREDICT v3, which should not be used in ductal carcinoma in situ or lobular carcinoma in situ only or in women with metastatic disease. Individuals with less than 1 day of follow-up were excluded, resulting in a final analytic cohort of 27,064 individuals.

We used data from a TEAM substudy [29]. TEAM, a trial of adjuvant endocrine therapy (exemestane vs tamoxifen followed by exemestane) in postmenopausal hormone receptor–positive breast cancer [52], had similar eligibility criteria as MA.27 . TEAM presented without a statistically significant difference in disease-free and overall survival between the two arms. For the purpose of prognostic survival modeling, the cohort was therefore analyzed as a single population. All individuals in TEAM had a nonzero follow-up time, such that the cohort was not further subselected. In total, 3825 individuals from TEAM were included in our analyses.

Information about missingness and variable mapping to PREDICT v3 for both of these external validation datasets is provided in section 2.4 in Multimedia Appendix 1. The external validation included calibration (ie, ICI) and discrimination (ie, AUROC) as detailed above. The 95% CI values were further derived from bootstrapping by taking the 2.5% and 97.5% percentile of the bootstrap distribution.

Ethical Considerations

This project has been approved by the Ottawa Health Science Network Research Ethics Board (protocol ID 20210803-01H) and the Children's Hospital of Eastern Ontario Research Ethics Board (protocol 25/107X). The Research Ethics Boards operate in compliance with, and is constituted in accordance with, the requirements of the Tri-Council Policy Statement: Ethical Conduct of Research Involving Humans; the International Conference on Harmonization Good Clinical Practice Consolidated Guideline; Part C, Division 5 of the Food and Drug Regulations; Part 4 of the Natural Health Products Regulations; and Part 3 of the Medical Devices Regulations and the provisions of the Ontario Personal Health Information Protection Act and its applicable regulations. This research involved the secondary use of health care datasets originally collected for purposes other than this study. This made the primary ethical consideration of this study the potential of disclosure risks. However, all datasets were deidentified at the source by the respective data custodians and were assessed as low risk. All analyses were conducted within a secure server environment with access restricted to authorized researchers of this study. These researchers have completed institutional privacy and security training, including instruction on the appropriate handling of personal health information, and, where required by data custodians, researchers also agreed to specific terms of use and completed additional ethics or data governance training. Individual re-consent was waived by the Research Ethics Boards, given that secondary use of deidentified data in this study posed minimal risk.

Results

Description of the MA.27 Study Participants

Our MA.27 dataset included 7563 postmenopausal women diagnosed with breast cancer. Table 3 provides an overview of the patients' characteristics. Additional descriptive visualizations of the MA.27 cohort are provided in section 3.1 in Multimedia Appendix 1.

XSL•FO
RenderX

Table 3. Characteristics of the MA.27 study population.

Variables	Values[a]
Age, median (IQR)	64.2 (58.2-71.2)
Nodal stage, n (%)	
N0	5360 (71.9)
N1	1615 (21.7)
N2	357 (4.8)
N3	124 (1.6)
Tumor laterality, n (%)	
Left-handed	3785 (50.1)
Right-handed	3663 (48.4)
Bilateral	115 (1.5)
Hormone receptor status, n (%)	
Estrogen receptor status positive	7,513 (99.3)
Progesterone receptor status positive	6079 (82)
Tumor size (cm), median (IQR)	1.5 (1-2)
Tumor grade, n (%)	
Grade 1	1892 (32)
Grade 2	2982 (50.5)
Grade 3	1036 (17.5)
Therapy, n (%)	
Radiotherapy	5370 (71.1)
Chemotherapy	2326 (30.8)
Trastuzumab therapy	67 (3.5)
Events, n (%)	187 (2.5)

[a]Percentages are calculated excluding missing values. The event was defined as breast cancer–related death within a 5-year observation interval. Details on missing values are provided in section 2.4 in Multimedia Appendix 1.

Performance Across Transfer Learning, De Novo ML, and the Stacked Ensemble

We evaluated the prognostic performance of three potential improvement strategies, transfer learning or fine-tuning (f-PREDICT v3), de novo ML (RSF and XGB), and ensemble stacking, and compared them to the pretrained model PREDICT v3. This evaluation was done to identify the most effective strategy when conducting prognostication in a new cohort where the lack of certain information and distribution shifts can considerably compromise the performance of pretrained models.

In the following, we present the evaluation of the pretrained model itself as well as transfer learning, de novo ML, and the stacked ensemble when training was optimized for ICI without rebalancing and without imputation of missing values. Results for different rebalancing strategies and detailed missingness analyses are provided in section 3 in Multimedia Appendix 1.

Table 4 gives the summary of calibration (ie, ICI) and discriminative performance (AUROC) in the form of the median across the 10 independent runs as explained in the Methods section. In terms of calibration, fine-tuning (ie, f-PREDICT v3) and RSF presented the best results in the internal evaluation (ICI 0.005 and 0.003, respectively). Performance in terms of discrimination ranged from an AUROC of 0.738 (PREDICT v3) to an AUROC of 0.799 (f-PREDICT v3). Ensemble stacking of f-PREDICT, RSF, and XGB yielded results comparable to the best stand-alone models (ICI 0.007 and AUROC 0.746).

XSL•FO
RenderX

Table 4. Calibration and discrimination for 5-year survival.

Model[a]	Calibration (ICI[b]), median (IQR)	Discrimination (AUROC[c]), median (IQR)
PREDICT v3[d]	0.042 (0.039-0.047)	0.738 (0.719-0.770)
f-PREDICT v3[e]	0.005 (0.004-0.010)	0.799 (0.789-0.818)
RSF[f]	0.003 (0.002-0.008)	0.744 (0.731-0.760)
XGB[g]	0.040 (0.038-0.043)	0.783 (0.764-0.810)
Ensemble[h]	0.007 (0.003-0.009)	0.746 (0.733-0.766)

[a]Values were calculated on the validation dataset. Median and IQR values across 10 seed settings are indicated for all survival models. Training was done on the nonimputed and non-rebalanced dataset and optimized for the integrated calibration index.

[b]ICI: integrated calibration index.

[c]AUROC: area under the receiver operating characteristic.

[d]PREDICT v3: pretrained survival model.

[e]f-PREDICT v3: pretrained survival model fine-tuned to MA.27.

[f]RSF: random survival forest.

[g]XGB: extreme gradient boosting.

[h]Ensemble: stacked ensemble integrating f-PREDICT v3, random survival forest, and extreme gradient boosting.

Consequently, all three improvement strategies added value compared to the pretrained model PREDICT v3 in MA.27. Calibration plots are illustrated in Figure 2 for all survival models, with the diagonal dashed line indicating perfect calibration, and include the median ICI across the 10 independent runs for each model. ROC curves are shown in Figure 3, with the diagonal dashed line indicating discrimination of a random guess. The median AUROC across the 10 independent runs is given for each model. Both figures confirm the summary results.

Figure 2. Calibration plots for 5-year survival. ICI: integrated calibration index.

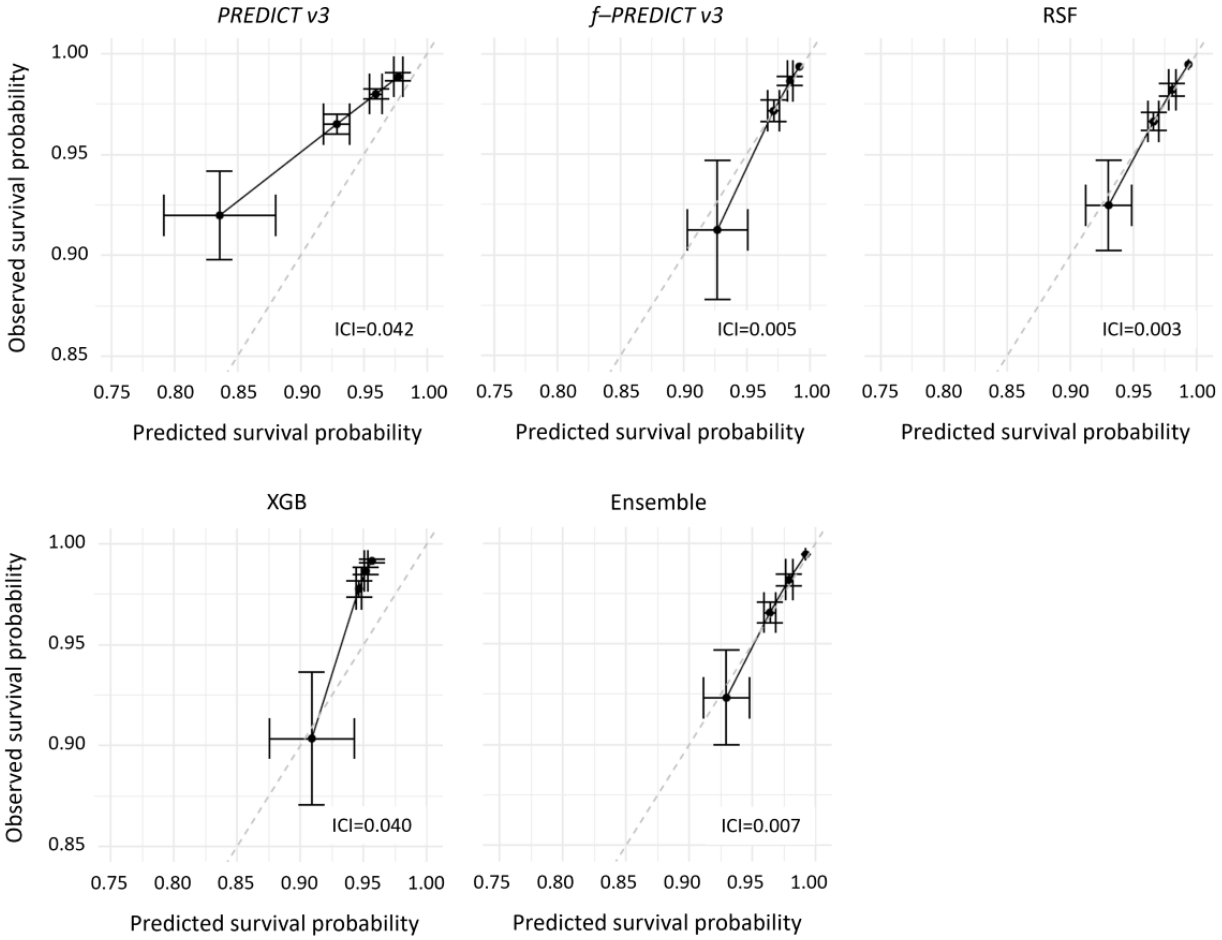

XSL•FO
RenderX

Figure 3. Receiver operating characteristic curves for 5-year survival. AUROC: area under the receiver operating characteristic curve.

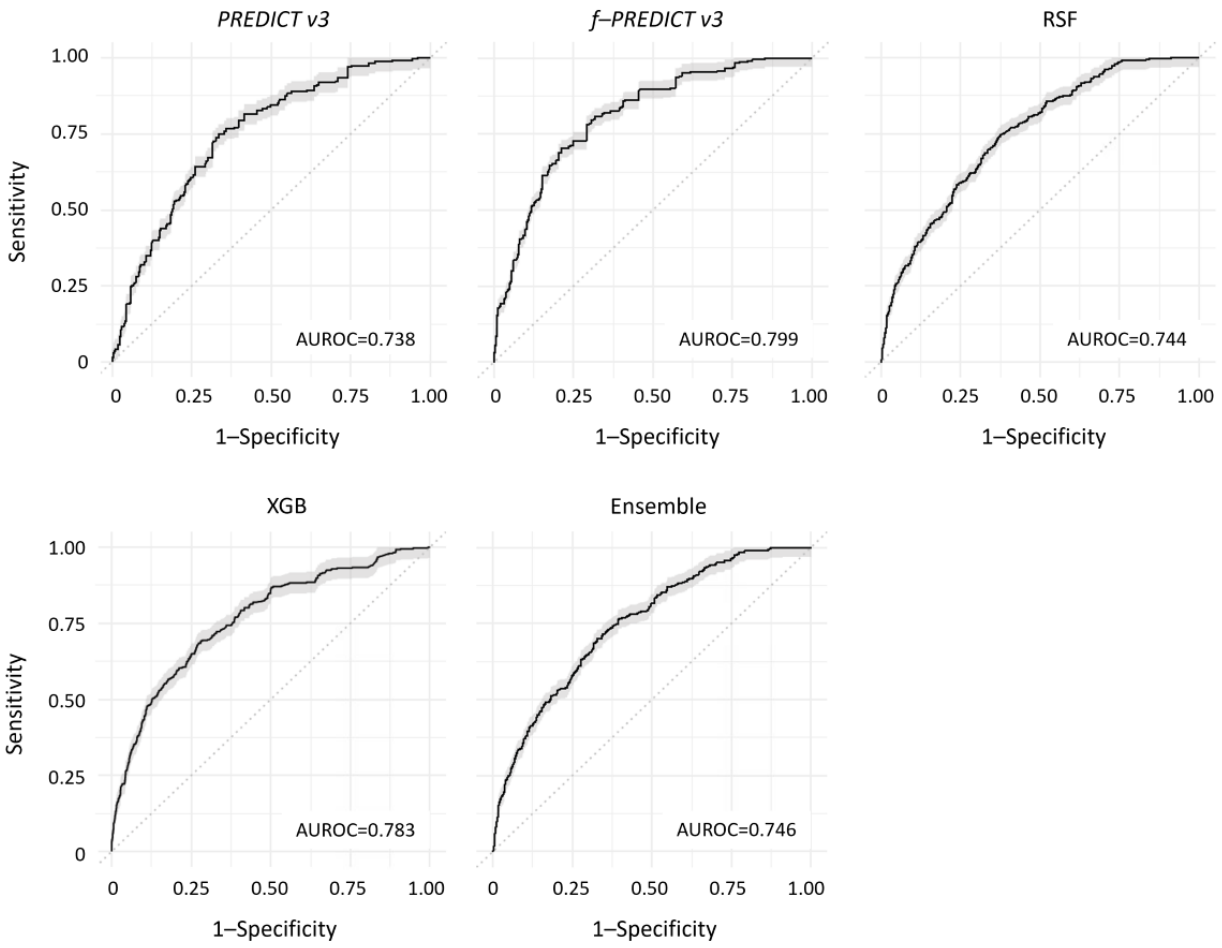

MA.27 had incomplete records with respect to some variables that were required for PREDICT v3 for valid survival estimation. Consequently, PREDICT v3 and f-PREDICT v3 could not give survival estimates for a subset of the entire dataset when these mandatory variables were missing. We refer to these cases where PREDICT v3 and f-PREDICT v3 returned no estimate due to missing input variables as invalid predictions. Any results presented from PREDICT v3 or f-PREDICT v3 (eg, Figures 2 and 3) exclude these invalid predictions. The number of invalid predictions in PREDICT v3 and f-PREDICT v3 ranged from 361/1514 (23.8%) to 391/1514 (25.8%) across the 10 independent runs. In contrast, RSF, XGB, and the stacked ensemble could predict survival for all individuals, independent of missing information.

We provide performance evaluation stratified by whether or not PREDICT v3 and f-PREDICT v3 returned valid predictions in Table 5. In general, models performed worse in patients that were lacking relevant information, but these differences were very small, and RSF and the stacked ensemble still presented good calibration (ICI 0.014 and 0.015, respectively) in this subset.

XSL•FO
RenderX

Table 5. Calibration and discrimination in subsets with and without relevant missing information.

Model[a]	Calibration (ICI[b]), median (IQR)	Discrimination (AUROC[c]), median (IQR)
Subset with valid predictions in PREDICT v3		
PREDICT v3[d]	0.042 (0.039-0.047)	0.738 (0.719-0.770)
f-PREDICT v3[e]	0.005 (0.004-0.010)	0.799 (0.789-0.818)
RSF[f]	0.006 (0.004-0.009)	0.763 (0.748-0.781)
XGB[g]	0.043 (0.040-0.044)	0.802 (0.786-0.812)
Ensemble[h]	0.006 (0.005-0.014)	0.775 (0.760-0.796)
Subset with invalid predictions in PREDICT v3		
RSF	0.014 (0.013-0.015)	0.726 (0.713-0.746)
XGB	0.051 (0.039-0.057)	0.745 (0.628-0.825)
Ensemble	0.015 (0.015-0.021)	0.691 (0.668-0.739)

[a]Values were calculated on the validation dataset stratified by whether or not PREDICT v3 could provide survival estimates (ie, valid vs invalid predictions). Invalid predictions occurred in patients where relevant information was missing. Median and IQR values across 10 seed settings are indicated for all survival models. Training was done on the nonimputed and non-rebalanced dataset and optimized for the integrated calibration index.

[b]ICI: integrated calibration index.

[c]AUROC: area under the receiver operating characteristic.

[d]PREDICT v3: pretrained survival model.

[e]f-PREDICT v3: pretrained survival model fine-tuned to MA.27.

[f]RSF: random survival forest.

[g]XGB: extreme gradient boosting.

[h]Ensemble: stacked ensemble integrating f-PREDICT v3, random survival forest, and extreme gradient boosting.

DCA

In Figure 4, results from the DCA are illustrated. More precisely, net benefit is plotted across risk thresholds (ie, predicted event probabilities) between 1% and 10%, and individuals with a predicted event probability (ie, risk) above the threshold were considered as intervention candidates. The DCA showed that f-PREDICT v3, RSF, and the ensemble yielded consistently positive net benefit across the evaluated threshold range, with the greatest benefit observed at lower thresholds. In contrast, PREDICT v3 crossed zero at an event probability of 7.5%, indicating that beyond this point, model-guided intervention was no longer preferable to intervening in none. Even at lower thresholds where PREDICT v3 was still net positive, its net benefit was consistently lower than that of f-PREDICT v3, RSF, and the stacked ensemble. XGB exhibited an unstable decision curve, consistent with its weaker calibration performance. At higher thresholds (ie, more than 8%-9%), net benefit approached zero for all models, reflecting the limited number of individuals exceeding such predicted event probabilities.

XSL•FO
RenderX

Figure 4. Decision-curve analysis for MA.27.

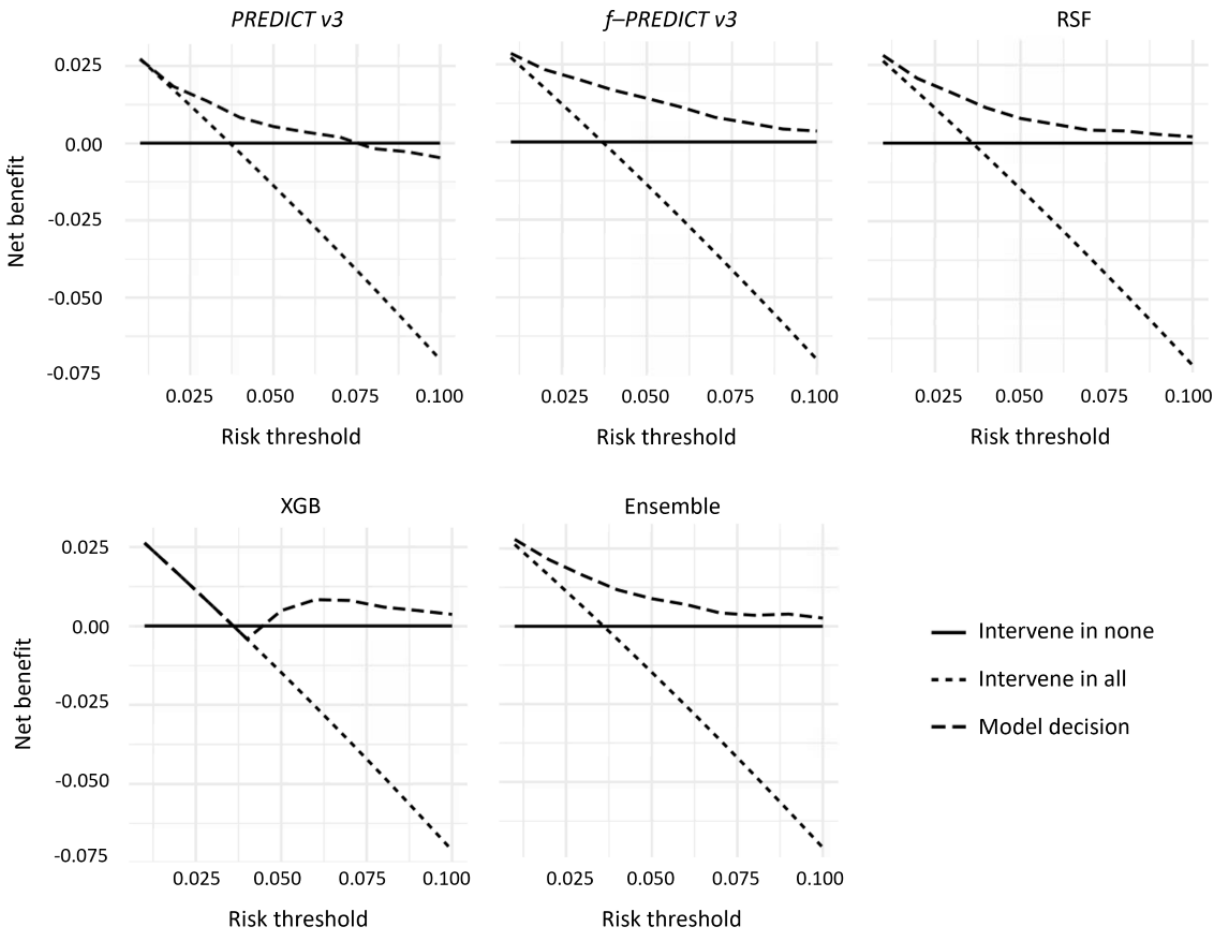

Model Explainability

Results from the SHAP analysis are illustrated in Figure 5. SHAP values were thereby calculated for all records in MA.27, and the variables were ranked based on their mean absolute SHAP value across all observations, reflecting the importance of the variable for the entire MA.27 dataset. The *x*-axis indicates the SHAP value, the size reflects the strength of a variable's contribution to the prediction, and the sign indicates the direction of this contribution (ie, positive SHAP values increase the survival probability). The direction of the variable is color-coded. While the exact top 3 variables in the SHAP analysis varied slightly between models, patient age, nodal status, pathological grading, and tumor size were consistently among them. In contrast, treatment information such as chemotherapy, radiotherapy, or trastuzumab was typically ranked less important across all models.

XSL•FO
RenderX

Figure 5. Shapley Additive Explanations analysis for MA.27. Er: estrogen receptor status; gen: chemotherapy; grade: pathological grading; nodes: number of positive nodes; pr: progesterone receptor status; radio: radiotherapy; side: tumor laterality; size: size of tumor; traz: trastuzumab therapy.

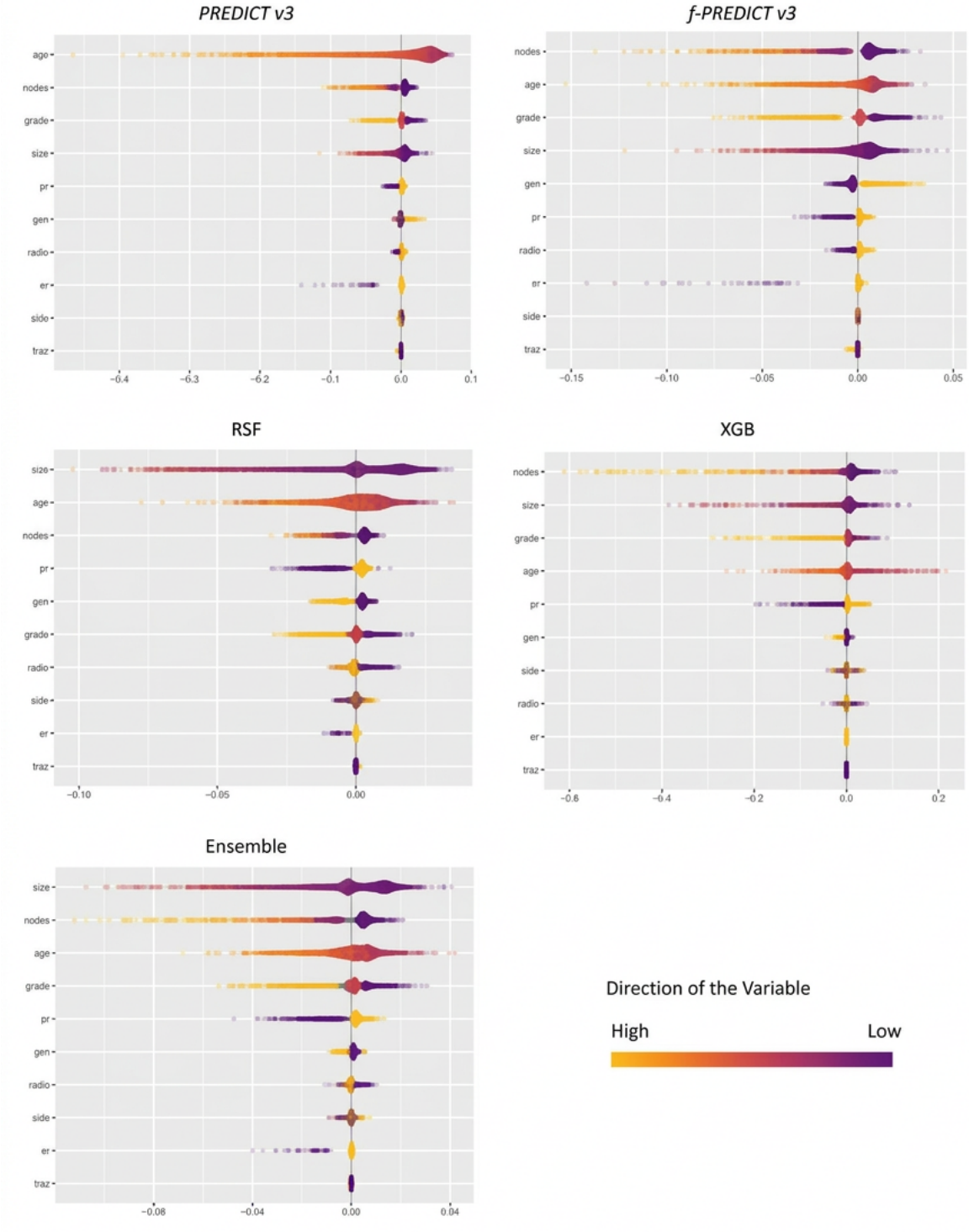

External Validation on SEER and TEAM

To assess the extent to which the MA.27 model optimization was generalizable, we externally validated all models on data from the US SEER program and on the clinical trial dataset TEAM.

Results from the SEER cohort are shown in Table 6 and confirm the benefit of transfer learning, de novo ML, and the stacked ensemble over the pretrained model. Calibration was worse than in the internal evaluation with an ICI of 0.010 (f-PREDICT v3), 0.020 (RSF), and 0.018 (ensemble); discriminative performance was comparable to that in the internal evaluation. The bootstrap distribution of the ICI in PREDICT v3 and f-PREDICT v3 was strongly right-skewed, with most bootstrap estimates concentrated close to the point estimate (PREDICT v3 median 0.039, IQR 0.039-0.040, and f-PREDICT v3 median 0.012, IQR 0.011-0.013) but a small number of resamples producing extreme values, resulting in wider upper confidence bounds.

XSL•FO
RenderX

Table 6. Calibration and discrimination for US Surveillance, Epidemiology, and End Results data.

Model[a]	Calibration, ICI[b] (95% CI)	Discrimination, AUROC[c] (95% CI)
PREDICT v3[d]	0.039 (0.037-0.198)	0.765 (0.750-0.780)
f-PREDICT v3[e]	0.010 (0.009-0.173)	0.825 (0.811-0.838)
RSF[f]	0.020 (0.019-0.046)	0.753 (0.741-0.765)
XGB[g]	0.037 (0.034-0.093)	0.759 (0.747-0.771)
Ensemble[h]	0.018 (0.016-0.078)	0.792 (0.779-0.802)

[a]Values were calculated on the external validation data. The integrated calibration index and area under the receiver operating characteristic curve and their respective 95% CI values, as derived from bootstrapping, are indicated for all survival models. Training was done on the nonimputed and non-rebalanced MA.27 dataset and optimized for the integrated calibration index.

[b]ICI: integrated calibration index.

[c]AUROC: area under the receiver operating characteristic.

[d]PREDICT v3: pretrained survival model.

[e]f-PREDICT v3: pretrained survival model fine-tuned to MA.27.

[f]RSF: random survival forest.

[g]XGB: extreme gradient boosting.

[h]Ensemble: stacked ensemble integrating f-PREDICT v3, random survival forest, and extreme gradient boosting.

For TEAM, the respective results are given in Table 7. In contrast to SEER, the MA.27 optimized models performed worse on the TEAM dataset, with changes in AUROC up to 0.122 (XGB from 0.783 in internal evaluation to 0.661 in external validation via TEAM) and changes in ICI up to 0.074 with consistent overestimation of survival (f-PREDICT from 0.005 in internal evaluation to 0.079 in the external validation via TEAM).

Table 7. Calibration and discrimination for Tamoxifen, Exemestane Adjuvant Multinational data.

Model[a]	Calibration, ICI[b] (95% CI)		Discrimination[c], AUROC (95% CI)	
	Coefficient	95% CI	Coefficient	95% CI
PREDICT v3[d]	0.034	0.028-0.058	0.701	0.672-0.732
f-PREDICT v3[e]	0.079	0.072-0.089	0.707	0.677-0.736
RSF[f]	0.073	0.071-0.076	0.648	0.623-0.673
XGB[g]	0.061	0.051-0.091	0.661	0.635-0.687
Ensemble[h]	0.073	0.071-0.076	0.678	0.651-0.703

[a]Values were calculated on the external validation data. The integrated calibration index and area under the receiver operating characteristic curve and their respective 95% CI values, as derived from bootstrapping, are indicated for all survival models. Training was done on the nonimputed and non-rebalanced MA.27 dataset and optimized for the integrated calibration index.

[b]ICI: integrated calibration index.

[c]AUROC: area under the receiver operating characteristic.

[d]PREDICT v3: pretrained survival model.

[e]f-PREDICT v3: pretrained survival model fine-tuned to MA.27.

[f]RSF: random survival forest.

[g]XGB: extreme gradient boosting.

[h]Ensemble: stacked ensemble integrating f-PREDICT v3, random survival forest, and extreme gradient boosting.

To better understand the discrepancy between the external validation on SEER and TEAM, we conducted a stratified analysis within the SEER cohort (section 3.8 in Multimedia Appendix 1). When stratified by nodal status, tumor size, and tumor grade, model performance in SEER decreased within the subgroup defined by node-positive disease, tumor size ≥ 2 cm, and tumor grade G3. These characteristics were more prevalent in TEAM, and performance in this subgroup showed calibration and discrimination comparable to those observed in TEAM.

Calibration plots and ROC curves for both datasets are given in section 3.7 in Multimedia Appendix 1.

Discussion

Summary and Comparison With the Literature

This study investigated how innovative learning approaches, including pretrained models combined with transfer learning, de novo ML (RSF and XGB), and the stacked ensemble, can

XSL•FO
RenderX

be leveraged to enhance the performance of prognostication tools for breast cancer. Using the MA.27 dataset, we addressed four questions: First, can fine-tuning the pretrained prognostic tool PREDICT v3 to the MA.27 dataset improve survival prediction performance compared to the pretrained model alone? Second, how do state-of-the-art ML models trained directly on the MA.27 dataset compare against the (fine-tuned) pretrained model PREDICT v3? Third, does a stacked ensemble using the fine-tuned pretrained model and de novo ML models add benefit compared to either approach alone? And fourth, do the potential benefits from fine-tuning, de novo ML, and the ensemble integration hold in external cohorts that are similar to MA.27?

All models, the pretrained PREDICT v3, the de novo ML (RSF and XGB), and the stacked ensemble, presented with moderate to good discrimination and calibration. The discriminatory ability of PREDICT v3 in our study was 0.738 (MA.27), 0.765 (SEER), and 0.701 (TEAM). This is worse than the published update by Grootes et al [32] using a cohort from the United Kingdom, where the AUROC for 5-year survival in hormone receptor–positive breast cancer ranged between 0.831 and 0.861. Chen et al [33] assessed PREDICT v3 in a Chinese cohort of 5424 women treated for nonmetastatic invasive breast cancer between 2010 and 2020, with an AUROC for 5-year survival in the hormone receptor–positive subcohort of 0.789. Hsiao et al [34] used SEER data from more than 860,000 patients (diagnosed between 2000 and 2018) to validate PREDICT v3 on US patients. In their study, the AUROC for 5-year survival was 0.797 in the hormone receptor–positive subcohort. Calibration was more difficult to compare directly between these studies due to differences in calibration metrics and presentation. However, visual comparison of calibration plots in the hormone receptor–positive subcohort suggests that PREDICT v3 was generally well-calibrated across these three studies [32-34]. If at all, there was a tendency to underestimate survival at the lower end of the survival prediction spectrum. In our study, this tendency was confirmed in MA.27 (ICI 0.042) and SEER (ICI 0.039) and appeared more pronounced than in the literature. TEAM, in contrast, showed an opposite pattern with a clear tendency to overestimate survival (ICI 0.034). While not directly comparable, a very recent validation of PREDICT v2.1 (not PREDICT v3) on Canadian patients with breast cancer diagnosed between 2004 and 2020 from Alberta achieved a more similar PREDICT v3 result to our study with an AUROC of 0.78 and an ICI of 0.03 (mixed cohort with hormone receptor–positive and hormone receptor–negative cancer) [53].

The lower performance of PREDICT v3 in our study and the one in Alberta reflects the challenge discussed in the introduction: MA.27 differs from the data used in the other studies, or more broadly, cohorts where pretrained models are deployed may diverge from the original training data.

Transfer learning (ie, f-PREDICT v3), de novo RSF (but not XGB), and the stacked ensemble outperformed the stand-alone pretrained PREDICT v3 in MA.27 and were similar in their performance. As model training was optimized for calibration, the improvement was most pronounced in the reduction of ICI (f-PREDICT v3 0.005, RSF 0.003, and ensemble 0.007 vs PREDICT v3 0.042 and XGB 0.040). The discriminatory ability also improved with an AUROC up to 0.799 (f-PREDICT v3), even though these differences would rather be considered negligible (less than 0.1 difference in AUROC). These findings were reflected in DCA, where improvements in calibration translated into consistent gains in net benefit across clinically relevant thresholds.

The external validation of the MA.27-trained models yielded mixed results: while model validation on the SEER cohort confirmed these findings, the benefit of transfer learning, de novo RSF, and the stacked ensemble did not apply to TEAM where the pretrained PREDICT v3 outperformed all alternative approaches with rather low calibration (ICI 0.039), and the MA.27-trained or fine-tuned models systematically overestimated survival on TEAM (ie, underestimated mortality), as reflected by calibration curves in section 3.7 in Multimedia Appendix 1. In other words, more 5-year breast cancer–related deaths occurred in TEAM than predicted by the models trained or fine-tuned on MA.27.

To better understand this discrepancy, we compared the cohorts in terms of their baseline characteristics. MA.27 and SEER were comparable across most patients' characteristics, including age, nodal stage, tumor size, and tumor grade. In contrast, TEAM differed substantially in relevant prognostic variables: While MA.27 consisted predominantly of node-negative individuals (71.9%), only 39.3% of the TEAM cohort were node-negative [29]. Tumor size was also larger in TEAM compared to MA.27 (TEAM: 47.3% ≤2 cm, 46.8% 2-5 cm, 5.9% > 5 cm [29] vs MA.27: median 1.5, IQR 1-2 cm), and tumor grade was higher (TEAM: 11.7% grade 1 vs MA.27: 32% grade 1). Thus, TEAM represented a cohort with a more adverse clinicopathological risk profile (higher nodal burden, larger tumors, and higher grade).

These differences are likely directly relevant to the observed prediction error. A plausible explanation is a shift toward higher-risk disease in TEAM. Although nodal status, tumor grading, and tumor size were consistently among the most relevant variables for model predictions in MA.27 (see Model Explainability in Results section), the relationship between these variables and mortality was learned from a cohort in which high-risk profiles were relatively underrepresented. When applied to TEAM, which has a relative overrepresentation of such cases, the predicted survival decrement associated with these characteristics may have been insufficient. This interpretation is supported by a post hoc stratified analysis in SEER, in which model performance decreased to levels comparable to those observed in TEAM within a stratum characterized by node-positive disease, tumor size ≥ 2 cm, and tumor grade 3, with a similar pattern of survival overestimation (section 3.8 in Multimedia Appendix 1). These are characteristics that were more prevalent in TEAM than in SEER or MA.27.

To come back to the questions posed: first, fine-tuning the pretrained prognostic tool PREDICT v3 to the MA.27 dataset led to a substantial improvement in performance, particularly in terms of calibration but also in terms of discrimination, compared to the pretrained model alone. Second, state-of-the-art ML models trained directly on the MA.27 dataset have mixed results. RSF matches f-PREDICT v3 and outperforms the

XSL•FO
RenderX

stand-alone pretrained model in terms of calibration but not in terms of discrimination; XGB showed the opposite pattern, with better discrimination but calibration closer to the stand-alone pretrained model. Third, the stacked ensemble did not add further benefit over either approach alone but, again, came with the advantage of providing predictions for all patients, even those with missing values, which is a practical advantage over f-PREDICT v3. And fourth, the observed benefits from fine-tuning and de-novo ML did extend to a similar SEER cohort. In this case, the ensemble appeared to be the best approach given its ability to handle missingness, its superior performance in both calibration and discrimination compared to PREDICT-v3, and its comparable performance to f-PREDICT v3. In contrast, none of these benefits generalized to the TEAM cohort, which is very likely due to the substantially different distribution of clinicopathological characteristics.

Relevance of Missing Information

Missing information represents an important consideration in prognostic model development, evaluation, and real-world deployment. In this study, two distinct forms of missingness were relevant.

First, all cohorts exhibited record-level missingness, where variables were collected but incomplete for a subset of individuals. We report missingness patterns and analyzed the outcome association (section 3.5 in Multimedia Appendix 1). While we did not observe an association between missingness indicators and 5-year survival outcomes, we could not be certain about a random missingness mechanism given the original data collection context. Although sensitivity analyses using model-based imputed data showed performance results consistent with the nonimputed approach (section 3.5 in Multimedia Appendix 1), we refrained from using imputed data to avoid potential bias introduction under nonrandom missingness.

More importantly, record-level missingness limited the applicability of PREDICT v3 and f-PREDICT v3. Because these models require specific mandatory inputs, they could not generate survival estimates when certain variables were missing. This can affect a nontrivial portion of the overall dataset (23.8%-25.8% in MA.27). ML models, and the stacked ensemble come with the advantage of being able to generate predictions despite incomplete inputs. While the performance of RSF (ie, the best stand-alone ML model) was lower for these individuals, it still achieved discrimination comparable to the pretrained PREDICT v3 on individuals that PREDICT v3 could predict and demonstrated superior calibration.

Second, global missingness occurred when variables required by PREDICT v3 and f-PREDICT v3 were absent at the dataset-level rather than for specific records. In these cases, clinically informed assumptions were necessary to enable model predictions. Sensitivity analyses under optimistic and pessimistic assumptions (section 3.4 in Multimedia Appendix 1) showed that predictions from PREDICT v3 varied substantially under these assumptions, while f-PREDICT v3 was less sensitive, and the ML models and the stacked ensemble had no relevant variability across the different scenarios. This suggests that ML-based approaches and the stacked ensemble have greater robustness in settings where input variables must be approximated.

These findings are not only relevant for model development but also demonstrate that missing information can affect the operational functionality of prognostic tools in certain prediction settings.

Addressing Outcome Imbalance

A particular methodological challenge of MA.27 was the low event rate. In binary classification tasks, such an imbalance is commonly addressed using dataset-level oversampling or algorithm-level weighted learning strategies [30]. The effectiveness of such strategies, however, varies across the literature, with differences across datasets and models [54-57].

In this study, we evaluated both approaches within a survival modeling framework. At the dataset level, a random oversampling examples approach [58] was applied. At the algorithm level, we implemented weighting during ML training. Neither approach improved overall performance. Dataset-level rebalancing via the random oversampling examples approach resulted in a substantial deterioration in calibration despite explicitly optimizing calibration during training. Algorithm-level weighting led to a comparatively less severe decline in performance. In contrast, models trained without outcome rebalancing achieved very good calibration (section 3.5 in Multimedia Appendix 1).

One possible explanation for this lack of benefit may lie in the differences between binary classification and survival modeling. In classification, imbalance directly affects the contribution of minority-class loss terms to the objective function. In contrast, survival models optimize more complex time-to-event objectives that incorporate both event occurrence and follow-up time. While recent methodological work has started to explore weight-based rebalancing within survival frameworks, particularly in Cox regression models [59], such approaches are not yet broadly adapted or standardized.

Implications for Practice

In early breast cancer, survival prognostication is regularly used by clinicians and patients to inform discussions on adjuvant systemic treatment decisions, particularly chemotherapy, following primary surgery [6,21-25], and is also considered relevant for follow-up planning [26].

In this context, this study demonstrates that parameter-based transfer learning, de novo ML training, and ensemble stacking can help to improve prognostication of the widely used tool PREDICT v3 in situations where relevant information is lacking or a dataset shift is likely. Interestingly, these benefits can even generalize beyond the training cohort.

While missing information may not play a role in clinical prognostication where patients provide real-time information, it can still arise, for example, when clinicians are not sure about certain information and rely on clinical assumptions. Our analyses demonstrate that PREDICT v3 is very sensitive to such input assumptions, and considering alternative plausible input scenarios can help clinicians to better understand the potential variability of predictions.

XSL•FO
RenderX

Missing information is also commonly encountered when doing retrospective survival analyses. In such situations, de novo ML training and ensemble stacking can be good approaches for more reliable survival prediction. The stacked ensemble represents a particularly robust approach for prognostic modeling. It combines the strengths of transfer learning and de novo ML, maintains calibration and discrimination comparable to the best stand-alone models, generates predictions despite incomplete inputs, and remains stable under alternative assumption scenarios.

However, transportability of the stacked ensemble to cohorts with a higher patient risk profile may be limited (particularly node-positive, large, grade 3 tumors). To facilitate the identification of such cohorts, we deposited code in the Open Science Framework repository that reproduces the high-risk definition used in our analyses [60]. In such settings, local recalibration, re–fine-tuning, or re-training on cohorts more representative of the intended use population should be considered before clinical deployment. Future research may also explore stratified or "mixture-of-expert" ensembles, in which separate models are trained for clinically defined strata and integrated into an ensemble. This might potentially improve transportability across heterogeneous populations.

Beyond individual patient care, such models can be leveraged for emerging in silico trial designs by simulating counterfactual outcomes and enabling virtual comparisons of treatment effects [61-63].

Limitations

This study has certain limitations, which are mainly driven by the dataset's characteristics. There were some variables necessary for PREDICT v3 entirely missing in the dataset (eg, chemotherapy regimen in case of chemotherapy), such that assumptions were made where reasonably possible (see details in Multimedia Appendix 1). PREDICT v3 was sensitive to alternative assumptions and may perform better in situations where such assumptions are not necessary.

The follow-up time in MA.27 was limited to 5 years. While 5-year survival represents an early and clinically relevant milestone to guide decision-making, screening, and treatment developments have significantly improved outcomes in recent years, and therefore longer-term survival, such as 10- or 15-year survival, is becoming more relevant.

MA.27 is a randomized clinical trial. While trial data ensure standardized and high-quality data collection, participants in randomized trials may differ from the population, thereby limiting representativeness. However, external validation against registry data from SEER demonstrated good performance, supporting the applicability of our findings to broader, real-world patient populations.

Another limitation is that we did not account for the most recent update to PREDICT v4 [64]. This update refitted PREDICT v3 but did not incorporate further input variables. It has not yet been provided as a user-facing interface, making it less relevant for current clinical practice. More importantly, the validation presented for PREDICT v4 focuses on 10-year survival, whereas our analysis was restricted to 5-year survival. Direct comparison was therefore not possible, but the main results indicate that PREDICT v4 performs very similarly to PREDICT v3 with very small improvements in 10-year survival. We further note that the GitHub repository of the laboratory involved in the development of PREDICT provides another implementation (PREDICT v4.1.1) [65], for which no peer-reviewed model development or validation study has yet been published. This implementation includes ancestry as an additional covariate and allows optional incorporation of Oncotype DX recurrence scores. Because no peer-reviewed study has yet been published, the additional variables were not available in our cohorts, and, similar to PREDICT v4, no user-facing interface is currently available, this implementation was not included in our comparison.

The methodological work on model improvement in the presented study used the MA.27 clinical trial, which represents an older cohort. Nevertheless, the learnings would apply to more recent datasets, and the same methods can be applied to more recent cohorts to improve contemporary prognostic model performance.

It is also relevant to note that prognostic application is different from predictive one [3]. The good performance of our model in terms of survival prognostication does not imply that the model can estimate individual-level treatment benefits. Causal inference methods must be leveraged to better interpret such functionality.

Acknowledgments

The authors would like to thank Paul Pharoah (Department of Computational Biomedicine, Cedars-Sinai Medical Center, LA) for the kind provision of the PREDICT v3 algorithm and the helpful comments on its use. We also would like to thank Daniel W Rea (Cancer Research UK Clinical Trials Unit, University of Birmingham, United Kingdom) for his contribution to the Tamoxifen Exemestane Adjuvant Multinational study and all patients and staff involved in the MA.27 and TEAM study. The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the Generative AI Delegation Taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing. The GenAI tool used was: ChatGPT (5.2 and 4o). Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

XSL•FO
RenderX

Funding

LP is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; 530282197). KEE is funded by the Canada Research Chairs program through the Canadian Institutes of Health Research, and a Discovery Grant (RGPIN-2022-04811) from the Natural Sciences and Engineering Research Council of Canada. The Ontario Institute for Cancer Research is supported by funds from the Government of Ontario.

Data Availability

All original code for our analysis has been deposited in the Open Science Framework [60]. We used confidential health care data (MA.27 and TEAM) as well as accessible data from the US Surveillance Epidemiology and End Results for this study. Access to the US Surveillance Epidemiology and End Results can be requested through the National Cancer Institute's SEER program.

Authors' Contributions

Conceptualization, design, and analysis: LP, GRP, KY, LJ, FKD, MC, and KEE. Data collection and acquisition: A-AB-B, LV, JH, M-FS, AL, LS, BC, JMSB, KJT, JB, SB, MS, CJHvdV, EM-KK, LD, EM, AH, CM, and MC. Drafting manuscript: LP and KEE. Review and editing: LP, KY, A-AB-B, GRP, LV, JH, M-FS, AL, LS, BC, JMSB, KJT, JB, SB, MS, CJHvdV, EM-KK, LD, EM, AH, CM, LJ, FKD, MC, and KEE.

Conflicts of Interest

KEE was the scholar-in-residence at the Office of the Information and Privacy Commissioner of Ontario at the time of conducting this study. KEE is editor-in-chief and FKD is an editorial board member (associate editor) at *JMIR AI* at the time of writing. KEE has financial interests in Woodway Assurance, a privacy technology spin-off company from his research lab at the University of Ottawa. Woodway Assurance's business area does not overlap with the topic of this paper. GRP has received consulting fees from Traferox Technologies and Calian CRO. GRP has a close family member who was recently employed by Roche Canada LTD and who owned stock in Roche Ltd. M-FS has received honoraria for educational presentations from AstraZeneca, Merck, Seagen, Novartis, Pfizer, Roche, Gilead, and Lilly. The authors declare that these interests are unrelated to the work presented in this manuscript and do not constitute a conflict of interest.

Multimedia Appendix 1
Additional information.
[PDF File (Adobe PDF File), 1477 KB - jmir_v28i1e88665_app1.pdf]

Multimedia Appendix 2
CREMLS checklist.
[PDF File (Adobe PDF File), 163 KB - jmir_v28i1e88665_app2.pdf]

References

1. Breast cancer. World Health Organization. 2025. URL: https://www.who.int/news-room/fact-sheets/detail/breast-cancer [accessed 2025-04-07]
2. Benitez Fuentes JD, Morgan E, de Luna Aguilar A, et al. Global stage distribution of breast cancer at diagnosis: a systematic review and meta-analysis. JAMA Oncol 2024;10(1):71-78. [doi: 10.1001/jamaoncol.2023.4837] [Medline: 37943547]
3. Cooper K, Nalbant G, Essat M, et al. Gene expression profiling tests to guide adjuvant chemotherapy decisions in lymph node-positive early breast cancer: a systematic review. Breast Cancer Res Treat 2025;210(2):229-247 [FREE Full text] [doi: 10.1007/s10549-024-07596-0] [Medline: 39899163]
4. Loh SW, Rodriguez-Miguelez M, Pharoah P, Wishart G. A comparison of chemotherapy recommendations using predict and adjuvant models. European Journal of Surgical Oncology (EJSO) 2011;37(5):S21-S22. [doi: 10.1016/j.ejso.2011.03.082]
5. Sparano JA, Crager MR, Tang G, Gray RJ, Stemmer SM, Shak S. Development and validation of a tool integrating the 21-Gene recurrence score and clinical-pathological features to individualize prognosis and prediction of chemotherapy benefit in early breast cancer. J Clin Oncol 2021;39(6):557-564. [doi: 10.1200/jco.20.03007]
6. Cardoso F, van't Veer LJ, Bogaerts J, et al. 70-Gene signature as an aid to treatment decisions in early-stage breast cancer. N Engl J Med 2016;375(8):717-729. [doi: 10.1056/nejmoa1602253]
7. Brain E, Mir O, Bourbouloux E, et al. Adjuvant chemotherapy and hormonotherapy versus adjuvant hormonotherapy alone for women aged 70 years and older with high-risk breast cancer based on the genomic grade index (ASTER 70s): a randomised phase 3 trial. The Lancet 2025;406(10502):489-500. [doi: 10.1016/s0140-6736(25)00832-3]
8. Leblanc A, Beltran-Bless AA, Pond GR, et al. Prognostic and predictive performance of PREDICT 2.1, PREDICT v3, and RSClin in node-negative early breast cancer: a TEAM pathology substudy. Breast Cancer Res Treat 2026;215(3):74. [doi: 10.1007/s10549-026-07909-5] [Medline: 41642493]

XSL•FO
RenderX

9. Beltran-Bless AA, Pond GR, Bayani J, et al. Does RSClin provide additional information over classic clinico-pathologic scores (PREDICT 2.1, INFLUENCE 2.0, CTS5)? Breast 2025;83:104528 [FREE Full text] [doi: 10.1016/j.breast.2025.104528] [Medline: 40633461]
10. Batra A, Patel A, Gupta VG, Mehta P, TVSVGK T, Biswas B, et al. Oncotype DX: Where does It stand in India? JGO 2019(5):1-2. [doi: 10.1200/jgo.19.00151]
11. Engelhardt EG, Binuya MAE, Pharoah PDP, et al. Prognostication and treatment predictions for estrogen receptor positive early-stage breast cancer: incorporating the 70-gene signature into the PREDICT prognostication model. Breast 2025;83:104542 [FREE Full text] [doi: 10.1016/j.breast.2025.104542] [Medline: 40714573]
12. El Haji H, Souadka A, Patel BN, et al. Evolution of breast cancer recurrence risk prediction: aAsystematic review of statistical and machine learning–based models. JCO Clinical Cancer Informatics 2023(7):e2300049. [doi: 10.1200/cci.23.00049]
13. Li J, Zhou Z, Dong J. Predicting breast cancer 5-year survival using machine learning: a systematic review. PLoS One 2021;16(4):e0250370 [FREE Full text] [doi: 10.1371/journal.pone.0250370] [Medline: 33861809]
14. Huang Y, Li J, Li M, Aparasu RR. Application of machine learning in predicting survival outcomes involving real-world data: a scoping review. BMC Med Res Methodol 2023;23(1):268 [FREE Full text] [doi: 10.1186/s12874-023-02078-1] [Medline: 37957593]
15. Wishart GC, Azzato EM, Greenberg DC, et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. Breast Cancer Res 2010;12(1):R1 [FREE Full text] [doi: 10.1186/bcr2464] [Medline: 20053270]
16. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big Data 2016;3(1). [doi: 10.1186/s40537-016-0043-6]
17. Zhang Y, Lobo-Mueller EM, Karanicolas P, Gallinger S, Haider MA, Khalvati F. CNN-based survival model for pancreatic ductal adenocarcinoma in medical imaging. BMC Med Imaging 2020;20(1):11 [FREE Full text] [doi: 10.1186/s12880-020-0418-1] [Medline: 32013871]
18. Zhu F, Zhong R, Li F, et al. Development and validation of a deep transfer learning-based multivariable survival model to predict overall survival in lung cancer. Transl Lung Cancer Res 2023;12(3):471-482 [FREE Full text] [doi: 10.21037/tlcr-23-84] [Medline: 37057112]
19. Brown G. Ensemble learning. In: Encyclopedia of Machine Learning and Data Mining. Boston, MA: Springer; 2017:393-402.
20. Goss PE, Ingle JN, Pritchard KI. Exemestane versus anastrozole in postmenopausal women with early breast cancer: NCIC CTG MA.27?A randomized controlled phase III trial. J Clin Oncol 2013;31:1398-1404. [doi: 10.3410/f.717976583.793471390]
21. Sparano JA, Gray RJ, Makower DF, et al. Adjuvant chemotherapy guided by a 21-Gene expression assay in breast cancer. N Engl J Med 2018;379(2):111-121. [doi: 10.1056/nejmoa1804710]
22. Henry NL, Somerfield MR, Abramson VG, et al. Role of patient and disease factors in adjuvant systemic therapy decision making for early-stage, operable breast cancer: update of the ASCO endorsement of the Cancer Care Ontario guideline. J Clin Oncol 2019;37(22):1965-1977. [doi: 10.1200/jco.19.00948]
23. Burstein HJ, Curigliano G, Loibl S, et al. Estimating the benefits of therapy for early-stage breast cancer: The St. Gallen International Consensus Guidelines for the primary therapy of early breast cancer 2019. Ann Oncol 2019;30(10):1541-1557 [FREE Full text] [doi: 10.1093/annonc/mdz235] [Medline: 31373601]
24. Vaz-Luis I, O'Neill A, Sepucha K, et al. Survival benefit needed to undergo chemotherapy: patient and physician preferences. Cancer 2017;123(15):2821-2828 [FREE Full text] [doi: 10.1002/cncr.30671] [Medline: 28323331]
25. Beltran-Bless AA, Saunders D, Clemons L, et al. Patient perceptions around the use of clinico-pathologic and genomic tools in the management of early breast cancer. Breast Cancer Res Treat 2025;214(1):87-99. [doi: 10.1007/s10549-025-07797-1] [Medline: 40839289]
26. De Rose F, Meduri B, De Santis MC, et al. Rethinking breast cancer follow-up based on individual risk and recurrence management. Cancer Treat Rev 2022;109:102434 [FREE Full text] [doi: 10.1016/j.ctrv.2022.102434] [Medline: 35933845]
27. Emam KE, Leung TI, Malin B, Klement W, Eysenbach G. Consolidated reporting guidelines for prognostic and diagnostic machine learning models (CREMLS). J Med Internet Res 2024;26:e52508 [FREE Full text] [doi: 10.2196/52508] [Medline: 38696776]
28. Surveillance, Epidemiology, and End Results (SEER) Program. 1990. URL: https://seer.cancer.gov/ [accessed 2026-03-13]
29. Bayani J, Yao CQ, Quintayo MA, et al. Molecular stratification of early breast cancer identifies drug targets to drive stratified medicine. NPJ Breast Cancer 2017;3(1):3 [FREE Full text] [doi: 10.1038/s41523-016-0003-5] [Medline: 28649643]
30. Altalhan M, Algarni A, Turki-Hadj Alouane M. Imbalanced data problem in machine learning: a review. IEEE Access 2025;13:13686-13699. [doi: 10.1109/access.2025.3531662]
31. Tang F, Ishwaran H. Random forest missing data algorithms. Stat Anal Data Min 2017;10(6):363-377 [FREE Full text] [doi: 10.1002/sam.11348] [Medline: 29403567]
32. Grootes I, Wishart GC, Pharoah PDP. An updated PREDICT breast cancer prognostic model including the benefits and harms of radiotherapy. NPJ Breast Cancer 2024;10(1):6 [FREE Full text] [doi: 10.1038/s41523-024-00612-y] [Medline: 38225255]

XSL•FO
RenderX

33. Chen E, Chen C, Chen Y, et al. Insights into the performance of PREDICT tool in a large mainland Chinese breast cancer cohort: a comparative analysis of versions 3.0 and 2.2. Oncologist 2024;29(8):e976-e983 [FREE Full text] [doi: 10.1093/oncolo/oyae164] [Medline: 38943540]
34. Hsiao YW, Wishart GC, Pharoah PDP. Validation of the PREDICT breast version 3.0 prognostic tool in US breast cancer patients. medRxiv. Preprint posted online on 2024 Oct 30 [FREE Full text] [doi: 10.1101/2024.10.29.24316401]
35. Ishwaran H, Kogalur UB. randomForestSRC: Fast unified random forests for survival, regression, and classification (RF-SRC). 2025. URL: https://cran.r-project.org/web/packages/randomForestSRC/randomForestSRC.pdf [accessed 2026-03-13]
36. Chen T, He T, Benesty M. Extreme gradient boosting. DMLC XGBoost. 2025. URL: https://xgboost.readthedocs.io/en/release_3.2.0/ [accessed 2026-03-13]
37. Austin PC, Harrell Jr FE, van Klaveren D. Graphical calibration curves and the integrated calibration index (ICI) for survival models. Stat Med 2020;39(21):2714-2742 [FREE Full text] [doi: 10.1002/sim.8570] [Medline: 32548928]
38. Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat Med 2013;32(30):5381-5397. [doi: 10.1002/sim.5958] [Medline: 24027076]
39. Blanche P. timeROC: Time-dependent ROC curve and AUC for censored survival data. CRAN Project. 2019. URL: https://cran.r-project.org/web/packages/timeROC/timeROC.pdf [accessed 2026-03-13]
40. Chowdhury A, Pharoah PD, Rueda OM. Evaluation and comparison of different breast cancer prognosis scores based on gene expression data. Breast Cancer Res 2023;25(1):17 [FREE Full text] [doi: 10.1186/s13058-023-01612-9] [Medline: 36755280]
41. Nguyen CT, Kattan MW. How to tell if a new marker improves prediction. Eur Urol 2011;60(2):226-228. [doi: 10.1016/j.eururo.2011.04.029] [Medline: 21536372]
42. Hond AAH D, Steyerberg EW, Calster B V. Interpreting area under the receiver operating characteristic curve. Lancet Digit Health 2022;4(12):e853-e855 [FREE Full text] [doi: 10.1016/S2589-7500(22)00188-1] [Medline: 36270955]
43. Austin PC, Putter H, Giardiello D, van Klaveren D. Graphical calibration curves and the integrated calibration index (ICI) for competing risk models. Diagn Progn Res 2022;6(1):2. [doi: 10.1186/s41512-021-00114-6]
44. Ditsch N, Gnant M, Thomssen C, Harbeck N. St. Gallen/Vienna 2025 summary of key messages on therapy in early breast cancer from the 2025 St. Gallen international breast cancer conference. Breast Care (Basel) 2025;20(4):1-10 [FREE Full text] [doi: 10.1159/000546080] [Medline: 40546709]
45. Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Med Inform Decis Mak 2008;8:53 [FREE Full text] [doi: 10.1186/1472-6947-8-53] [Medline: 19036144]
46. Van Calster B, Collins GS, Vickers AJ, et al. Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance. Lancet Digit Health 2025;7(12):100916 [FREE Full text] [doi: 10.1016/j.landig.2025.100916] [Medline: 41391983]
47. Sjoberg DD, Vertosick E. dcurves: Decision curve analysis for model evaluation. CRAN Project. 2025. URL: https://cran.r-project.org/web/packages/dcurves/dcurves.pdf [accessed 2026-03-13]
48. Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst 2013;41(3):647-665. [doi: 10.1007/s10115-013-0679-x]
49. Casalicchio G, Molnar C, Schratz P. iml: Interpretable machine learning. 2025. URL: https://arxiv.org/abs/2010.09337 [accessed 2026-03-13]
50. Mayer M, Stando A. shapviz: SHAP Visualizations. CRAN Project. 2025. URL: https://cran.r-project.org/web/packages/shapviz/shapviz.pdf [accessed 2026-03-13]
51. Qi SA, Kumar N, Farrokh M. An effective meaningful way to evaluate survival models. 2023 Presented at: Proceedings of the 40th International Conference on Machine Learning; 2023 July 23; USA p. 28244-28276.
52. van de Velde CJH, Rea D, Seynaeve C, et al. Adjuvant tamoxifen and exemestane in early breast cancer (TEAM): a randomised phase 3 trial. The Lancet 2011;377(9762):321-331. [doi: 10.1016/s0140-6736(10)62312-4]
53. Basmadjian RB, Xu Y, Quan ML, Lupichuk S, Cheung WY, Brenner DR. Evaluating PREDICT and developing outcome prediction models in early-onset breast cancer using data from Alberta, Canada. Breast Cancer Res Treat 2025;211(2):399-408. [doi: 10.1007/s10549-025-07654-1] [Medline: 40072699]
54. Gurcan F, Soylu A. Learning from imbalanced data: integration of advanced resampling techniques and machine learning models for enhanced cancer diagnosis and prognosis. Cancers (Basel) 2024;16(19):3417 [FREE Full text] [doi: 10.3390/cancers16193417] [Medline: 39410036]
55. Fernandez A, Garcia S, Herrera F, Chawla NV. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. jair 2018;61:863-905. [doi: 10.1613/jair.1.11192]
56. Yang C, Fridgeirsson EA, Kors JA, Reps JM, Rijnbeek PR. Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. J Big Data 2024;11(1):7. [doi: 10.1186/s40537-023-00857-7]

XSL•FO
RenderX

57. Wang L, Shi E, Meyers B, Vlachos P, Tcheng J, Denardo S. Machine learning performance for a small dataset: random oversampling improves data imbalances and fairness. BMC Med Res Methodol 2026 [FREE Full text] [doi: 10.1186/s12874-026-02779-3] [Medline: 41639760]
58. Lunardon N, Menardi G, Torelli N. ROSE: A package for binary imbalanced learning. The R Journal 2014;6(1):79. [doi: 10.32614/rj-2014-008]
59. Ning C, Bertsimas D, Lønning PE, et al. Improving survival models in health care by balancing imbalanced cohorts: a novel approach. JCO Clin Cancer Inform 2026(10):e2500190. [doi: 10.1200/cci-25-00190]
60. Pilgram L, El Emam K. Transfer learning and machine learning for breast cancer survival prediction. Open Science Framework. 2025. URL: https://doi.org/10.17605/OSF.IO/8N4EP [accessed 2026-03-13]
61. Switchenko JM, Heeke AL, Pan TC, Read WL. The use of a predictive statistical model to make a virtual control arm for a clinical trial. PLoS One 2019;14(9):e0221336 [FREE Full text] [doi: 10.1371/journal.pone.0221336] [Medline: 31483824]
62. Jia Z, Lilly MB, Koziol JA, et al. Generation of "virtual" control groups for single-arm prostate cancer (PCa) adjuvant trials. PLoS One 2014;9:e85010. [doi: 10.1200/jco.2013.31.6_suppl.239]
63. Strayhorn JM. Virtual controls as an alternative to randomized controlled trials for assessing efficacy of interventions. BMC Med Res Methodol 2021;21(1):3 [FREE Full text] [doi: 10.1186/s12874-020-01191-9] [Medline: 33402097]
64. Pharoah PDP, Hsiao YM, Wishart GC, Peng P. PREDICT breast v4.0: an update to the PREDICT breast prognostic model. BMC Res Notes 2025;18(1):482 [FREE Full text] [doi: 10.1186/s13104-025-07552-1] [Medline: 41239533]
65. Penglab. pengpclab/PREDICTv4.1.1. 2025. URL: https://penglab.net/ [accessed 2026-03-13]

Abbreviations

AUROC: area under the receiver operating characteristic
DCA: decision-curve analysis
f-PREDICT v3: pretrained survival model fine-tuned to MA.27
ICI: integrated calibration index
ML: machine learning
ROC: receiver operating characteristic
RSF: random survival forest
SEER: US Surveillance Epidemiology and End Results
SHAP: Shapley Additive Explanations
TEAM: Tamoxifen Exemestane Adjuvant Multinational
XGB: extreme gradient boosting

Edited by J Sarvestan; submitted 29.Nov.2025; peer-reviewed by D Hu, E Joodi; comments to author 29.Dec.2025; revised version received 11.Mar.2026; accepted 12.Mar.2026; published 14.Apr.2026.

Please cite as:
Pilgram L, Yang K, Beltran-Bless AA, Pond GR, Vandermeer L, Hilton J, Savard MF, LeBlanc A, Shepherd L, Chen B, Bartlett JMS, Taylor KJ, Bayani J, Barker S, Spears M, van der Velde CJH, Meershoek-Klein Kranenbarg E, Dirix L, Mallon E, Hasenburg A, Markopoulos C, Juwara L, Dankar FK, Clemons M, El Emam K
Transfer Learning and Machine Learning for Training Five-Year Survival Prognostic Models in Early Breast Cancer: Development and Validation Study
J Med Internet Res 2026;28:e88665
URL: https://www.jmir.org/2026/1/e88665
doi:10.2196/88665
PMID:

©Lisa Pilgram, Kai Yang, Ana-Alicia Beltran-Bless, Gregory R Pond, Lisa Vandermeer, John Hilton, Marie-France Savard, Andreanne LeBlanc, Lois Shepherd, Bingshu Chen, John MS Bartlett, Karen J Taylor, Jane Bayani, Sarah Barker, Melanie Spears, Cornelis JH van der Velde, Elma Meershoek-Klein Kranenbarg, Luc Dirix, Elizabeth Mallon, Annette Hasenburg, Christos Markopoulos, Lamin Juwara, Fida K Dankar, Mark Clemons, Khaled El Emam. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Plain Language Summarization of Environmental Health Research Using Generative AI: Community-Engaged Qualitative Study

Katherine Wade[1,2,3], MPH; Lauren B Anderson[1,4], MPA; Evelyn M Medley[5], BA; Ted Smith[1,3], PhD

[1]Christina Lee Brown Envirome Institute, School of Medicine, University of Louisville, Louisville, KY, United States

[2]Department of Epidemiology and Population Health, School of Public Health and Information Sciences, University of Louisville, Louisville, KY, United States

[3]Superfund Research Center, University of Louisville, Louisville, KY, United States

[4]Department of Urban and Public Affairs, College of Arts and Sciences, University of Louisville, Louisville, KY, United States

[5]Berea College, Berea, KY, United States

Corresponding Author:
Katherine Wade, MPH
Christina Lee Brown Envirome Institute
School of Medicine
University of Louisville
302 E Muhammad Ali Blvd
Louisville, KY, 40202
United States
Phone: 1 5028525724
Email: kdsher02@louisville.edu

Abstract

Background: Plain language summaries (PLSs) are increasingly required in environmental health publications to improve public accessibility. Generative artificial intelligence (AI) systems such as large language models can automatically produce such summaries; however, automated summaries often overlook local context and cultural relevance—limitations that are critical in environmental health research, where affected communities face disproportionate exposure and health risks.

Objective: This study aimed to develop and refine community-informed prompts for generating PLSs of environmental health research using generative AI.

Methods: We conducted a community-engaged qualitative study in Louisville, Kentucky, involving workshops with 97 participants from 4 stakeholder groups: summer interns at a local social justice nonprofit organization, participants in a youth development program, participants in a faith-based community organization, and members of a community advisory board focused on environmental justice. Participants reviewed PLSs generated using 3 different prompt styles in GPT-4o (ChatGPT, OpenAI). Feedback was collected through structured discussions and facilitator notes. Data were analyzed using thematic analysis informed by knowledge translation frameworks to identify preferred structural, linguistic, and contextual features of the summaries.

Results: Participants consistently preferred summaries between 300 and 400 words, written at a sixth- to eighth-grade reading level. Key priorities included presenting definitions before findings, using headings and bullet points to improve readability, and clearly explaining real-world and environmental justice implications. Narrative summaries without structure were viewed as overly long and difficult to interpret, while purely bullet-based formats were considered too simplified. Feedback from youth participants emphasized clarity and practical relevance, while faith-based participants highlighted the importance of trust and contextual framing. These insights informed the development of a community-refined prompt that included definitions, key findings, an introduction, and a concluding section on community implications.

Conclusions: Community-engaged prompt development may improve the relevance and interpretability of AI-generated PLSs of environmental health research. Incorporating stakeholder perspectives into prompt design offers a replicable strategy for improving research translation and ensuring AI-generated summaries reflect the informational needs of affected communities.

(J Med Internet Res 2026;28:e87118) doi:10.2196/87118

XSL•FO
RenderX

KEYWORDS

generative artificial intelligence; generative AI; large language models; plain language summaries; environmental health; community-based participatory research; environmental justice; research translation; health communication

Introduction

Background

Plain language summaries (PLSs) aim to bridge the gap between scientific communication and public understanding [1]. In environmental health research, where technical language often obscures public health implications, PLSs are particularly critical [2]. Although publishers, platforms, and even search engines now provide standardized artificial intelligence (AI)–generated lay summary functions, these "precanned" summaries frequently overlook local context, cultural framing, and specific community concerns [1,2]. Studies examining automated summarization in scientific publishing note that although such tools improve accessibility, they often fail to capture place-based relevance, omit social factors, and lack sensitivity to community-level variation in risk perception [1,3]. For instance, an automated summary of a study on air pollution and heart disease might highlight general health effects but omit that the research was conducted in a predominantly Black neighborhood disproportionately exposed to industrial emissions due to historical redlining—a context that is potentially vital for community understanding and action.

The growth of generative AI has accelerated opportunities to create PLSs but has also raised questions of accuracy, inclusivity, and transparency [4]. Community-informed approaches, rooted in established models of engagement [5,6], offer a pathway to developing prompts that are better aligned with the needs of nonexpert but highly invested audiences.

Environmental Health Literature Lay Summarization

Many journals in environmental health and biomedical sciences now require or encourage PLSs (Textbox 1).

Textbox 1. Examples of lay summary formats in environmental health and biomedical journals.

Several publishers and organizations have adopted distinct formats for presenting plain language summaries:

- Cochrane provides detailed templates requiring structured sections such as "What did we find?" and "What are the limitations?" [7].
- eLife has pioneered plain language "digests" as an integral component of their publications [8].
- Journals in PLOS contain concise "Author Summary" sections targeted at lay readers [1].

Although it is not possible to know how many of these summaries are created by the author without AI assistance, the notion that research publications should include accessible summaries has become a convention in scholarly publications. Additionally, some publishers such as Wiley and Elsevier have incorporated AI-generated summaries into their online presentation of papers [9]. Table 1 compares lay summary formats currently used across journals and AI platforms.

Table 1. Comparison of lay summary prompts across journals and platforms.

Source	Format or prompt style	Primary audience	Notable features
Cochrane	Structured template	Patients and policymakers	Sections such as "What did we do?" "What did we find?" "What are the limitations?"
eLife	Narrative "digest"	General public	Short, story-style explanation written alongside the article
PLOS journals	Author summary	Nonspecialist readers	Concise summary (approximately 250 words) describing the significance of the research
Wiley	Automated research summary	General and academic audiences	Machine learning–generated article highlights presented on article web pages
Environmental Health Perspectives	Short plain language summary or highlight box	Policy and practice audiences	Journalistic-style summaries emphasizing relevance to public health
Google Gemini (search integration)	Artificial intelligence–generated abstract simplifications	General users	Automated summaries generated dynamically within search results

Beyond these features in journal publication, generative AI has become an automatic feature when these articles are discovered in Google searches with the recently integrated Gemini (version 3.1; Google LLC)-generated summaries, and most large language models now generate automatic "key findings" or "about this study" snippets when academic articles are searched, functioning as default lay summaries. Although these features democratize access, they risk missing the types of information needs that can be met through an iterative, context-sensitive cocreation process that community engagement can enable.

The substance and style of an AI-generated summary are directly mediated by the nature of the prompt. Even small adjustments in phrasing—such as specifying audience, tone, or structure—can produce markedly different outputs from the same model [10]. Prompt wording determines not only what information is emphasized but also how it is framed, influencing

XSL•FO
RenderX

readability, inclusivity, and factual precision [4]. This sensitivity underscores that prompt design is not a neutral technical process but a critical determinant of research translation. In the context of environmental health, where issues of justice and accessibility are paramount, the ability to guide AI outputs toward equity-oriented narratives depends on the intentionality of the prompts themselves [2]. This study aimed to explore a community-engaged approach to developing prompts for generative AI systems to produce PLSs of environmental health research that are accessible, contextually relevant, and meaningful to affected communities [5,6]. Our specific aim was to codevelop and refine prompts for GPT-4o (ChatGPT, OpenAI) through direct engagement with diverse community stakeholders in Louisville, Kentucky.

Methods

Setting and Participants

Engagement workshops were conducted in partnership with 3 community-based groups that represent populations disproportionately impacted by environmental exposures and health disparities. A total of 97 participants took part in the workshops, including youth participants from a mentoring program (n=81, 83.5%), members of a community advisory board focused on environmental health (n=10, 10.3%), participants from a faith-based environmental justice initiative (n=6, 6.2%), and student interns at a social justice nonprofit organization (n=5, 5.2%). Prompt development followed an iterative refinement process across multiple engagement sessions. Early sessions (June-July 2024) focused on comparing alternative prompt structures (eg, narrative vs bullet format and varying reading levels), while later sessions (October 2024) evaluated refined prompts incorporating structured sections and definitions.

Detailed demographic data such as age, gender, race or ethnicity, or education level were not collected during the workshops because the activities were conducted as community engagement sessions rather than formal participant surveys.

Workshop Procedures

Workshops were cofacilitated by environmental health academic researchers and community liaisons. Each session began with an orientation to the purpose of PLSs, examples of generative AI, and ethical considerations around AI use. Participants were provided with summaries generated using GPT-4o of 3 environmental health journal papers [11-13]. Papers were selected based on their authorship by readily accessible Superfund researchers and their relevance as current publications.

All PLSs were reviewed by study authors for accuracy, completeness, and readability. Although interrater reliability was not assessed (because the authors reviewed their own papers), the review process ensured scientific fidelity. Facilitators used structured discussion guides and note-taking templates to ensure consistency across sessions. Feedback was collected through a combination of small-group discussions, full-group debriefs, and facilitator-recorded observations. In most sessions, feedback was documented in real time by at least 1 team member, and in some sessions, multiple facilitators contributed to note-taking to ensure completeness.

PLS Prompts

Participants evaluated PLSs generated using 3 different prompt structures (refer to Textbox 2 for details).

Textbox 2. Plain language summary prompt structures evaluated in the study.

The following prompts represent the initial and refined instructions used to generate plain language summaries for participant evaluation:

- Prompt A (initial draft): “Summarize the research paper in 600 words at an eighth-grade reading level. The entire summary should be in a narrative format.”
- Prompt B (initial draft): “Summarize the research paper in 300 words at a sixth-grade reading level. The entire summary should be in a bullet format.”
- Prompt C (refined draft): “Summarize the provided text using simple and clear language. Start with key findings followed by translations for technical terms. Organize the summary using headings for ease of reading. Include the following:
- Complete title and all authors
- Key findings: list the main results or conclusions
- Definitions: translate any technical terms used in the summary
- Introduction: provide a brief overview of the topic
- Main content: discuss the significant points, organized with headers as needed
- Conclusion: summarize the overall findings and their implications

Ensure the summary is suitable for a nonexpert audience. Do not include the Methods or Results sections in the summary.”

Participants reviewed the summaries created by each of these prompts individually and in small groups, then discussed strengths, weaknesses, and preferred features. Only prompt C specified the use of section headings and structured subsections; prompts A and B generated narrative and bullet-based format summaries without headings. Facilitators recorded real-time feedback using structured templates and open-ended note-taking (refer to Multimedia Appendices 1 and 2 for the PLS protocol and complete examples of prompts and generated PLSs).

XSL•FO
RenderX

Analytic Approach

Feedback from workshops was documented through structured templates and detailed facilitator notes recorded during and immediately after each session. Across all engagement activities (June-October 2024), notes captured both group-level discussion and specific participant comments regarding clarity, structure, tone, and perceived usefulness of each PLS. The designated note-taker documented the principal elements of participants' contributions and reviewed them with the group prior to the conclusion of the session to ensure accuracy. We conducted a thematic analysis drawing on knowledge translation frameworks [9] and empirical research on plain language summarization [1], using an iterative, multistage approach. First, all notes were compiled across sessions and reviewed to identify recurring patterns in participant feedback. Members of the research team performed initial open coding to label discrete concepts (eg, "need for simpler language," "preference for definitions," "desire for personal relevance," and "confusion with technical terms"). These codes were informed both by emergent participant feedback and by prior literature on plain language communication and knowledge translation.

Second, codes were grouped into higher-order thematic categories through an inductive-deductive process. Themes were refined through comparison across multiple community groups (youth participants, community advisory board members, and faith-based participants) to identify consistent versus group-specific priorities. For example, youth participants emphasized personal relevance and visual elements, while advisory board participants more frequently commented on clarity, completeness, and scientific accuracy.

Third, themes were organized into three overarching domains: (1) structural (eg, use of headings, placement of key findings, and inclusion of definitions); (2) linguistic (eg, reading level, jargon reduction, and clarity of explanations); and (3) contextual (eg, relevance to lived experience, environmental justice framing, and implications for personal or community health).

To enhance analytic rigor, findings were triangulated across multiple sessions conducted at different times and with distinct participant groups (June, July, and October 2024). Consistency of feedback across sessions, such as repeated calls for simpler language, clearer definitions, and emphasis on personal relevance, was used to confirm theme stability. Finally, themes were applied to iteratively refine the prompt design. Feedback directly informed the development of the final structured prompt (prompt C), which incorporated prioritized features, including key findings at the beginning, explicit definitions of technical terms, and organized headings for readability. A summary of thematic guidelines and sample feedback is provided in Multimedia Appendix 3.

Ethical Considerations

The project was reviewed by the University of Louisville Institutional Review Board (IRB) and was deemed inapplicable for the Common Rule definition of human subjects research because the activities were classified as program evaluation and quality improvement rather than research involving identifiable human subjects. The IRB issued a formal determination of non–human subjects research (IRB 25.0911; reference 813263) on February 2, 2026.

Workshop participants engaged in group discussions evaluating PLSs generated using generative AI tools. No personal identifiers were collected during the workshops, and all feedback was documented through facilitator notes in deidentified form. Because the activity was determined to be non–human subjects research, formal informed consent procedures were not required. Participants were informed about the purpose of the workshop activities and participated voluntarily. Participants did not receive compensation for participation, although food and beverages were provided.

All project activities followed University of Louisville policies regarding data security and privacy. Feedback collected during workshops was stored on institutionally approved systems and analyzed only in an aggregated, deidentified form. No participant-level data are reported in this manuscript. Youth participants were involved through community partner programs that follow institutional youth protection policies and guidelines.

Results

Participant Characteristics

A total of 97 individuals participated in 3 community engagement workshops conducted in Louisville, Kentucky. Participants represented 3 community partner groups: a youth development and mentoring program (n=81, 83.5%), a community advisory board focused on environmental health and research engagement (n=10, 10.3%), and a faith-based organization involved in environmental justice initiatives (n=6, 6.2%). The youth participants were primarily high school students, while the advisory board and faith-based groups consisted of adult community members with varying levels of experience interacting with environmental health research.

The large proportion of youth participants provided insight into how younger audiences interpret scientific information presented in simplified formats. Participants across all groups were familiar with environmental concerns in their communities, including air pollution and industrial exposures, and were therefore well-positioned to evaluate whether the summaries communicated research findings in ways that were understandable and relevant.

Evaluation of Prompt Formats

Participants reviewed PLSs generated using 3 different prompt structures. These prompts varied in length, format, and the level of structural guidance provided to the generative AI system.

Prompt A generated summaries of approximately 600 words written in a narrative format at an eighth-grade reading level. Participants in the community advisory board workshop generally perceived this format as overly long and dense. Several participants commented that the summaries felt similar to reading a condensed research paper rather than a plain language explanation. Although the summaries contained detailed explanations of study findings, participants noted that key messages were sometimes difficult to identify quickly.

XSL•FO
RenderX

Prompt B produced summaries of approximately 300 words written in a bullet-point format at a sixth-grade reading level. Participants reported that this structure made the summaries easier to skim and understand. Youth participants in particular appreciated the shorter length and visual organization of the bullet points. However, both youth and adult participants noted that this format sometimes oversimplified the research findings and lacked sufficient explanation of key concepts or context. Participants also reported that the bullet-point format occasionally made the summaries feel fragmented or disconnected from the broader purpose of the study.

Prompt C incorporated a structured format with headings, definitions of technical terms, and clearly separated sections describing key findings and implications. Participants across all workshops responded more favorably to this structure. The use of headings and defined sections allowed readers to navigate the summary more easily and locate the most important information. Participants also expressed appreciation for the inclusion of definitions that explained technical terminology before presenting the research findings.

Emergent Themes From Community Feedback

Across the 3 workshops, several consistent themes emerged regarding preferred characteristics of PLSs generated using AI tools. These themes related to the structure, language, and contextual framing of the summaries.

First, participants emphasized the importance of presenting definitions of technical terms before introducing the research findings. Participants explained that encountering unfamiliar terminology early in a summary could make the content difficult to follow. Providing definitions at the beginning helped establish a clearer understanding of the scientific concepts discussed later in the summary.

Second, participants consistently expressed a preference for concise summaries. Many participants indicated that summaries of between approximately 300 and 400 words struck an appropriate balance between readability and informational depth. Summaries longer than this were perceived as burdensome to read, while shorter summaries sometimes omitted important context.

Third, participants emphasized the importance of clearly explaining why the research matters. In particular, youth participants frequently asked how the findings related to their own communities or daily lives. Participants indicated that summaries should explicitly describe potential health implications and identify populations that may be most affected by the environmental exposures being studied.

Finally, participants highlighted the importance of contextual framing related to environmental justice. Adult participants in the advisory board and faith-based workshops noted that environmental health research often occurs in communities disproportionately exposed to pollution. They recommended that summaries explicitly acknowledge these broader social and environmental contexts when describing study findings.

Development of the Community-Refined Prompt

The themes identified through workshop discussions informed the development of a revised prompt designed to generate more accessible and contextually relevant summaries. The final prompt incorporated several features that participants consistently preferred. These included presenting definitions of technical terms at the beginning of the summary, organizing information into clearly labeled sections, and highlighting key findings using concise bullet points.

In addition, the revised prompt instructed the AI system to include a short, narrative introduction and conclusion describing the broader implications of the research. Participants emphasized that these sections helped connect the scientific findings to real-world concerns, including environmental health risks and community-level impacts.

Therefore, the final prompt combined both structured and narrative elements to improve readability while preserving important contextual information. When applied to environmental health research papers, this prompt produced summaries that participants described as clearer, more trustworthy, and more relevant to community audiences. The final prompt was not subjected to systematic re-evaluation by participants. Subsequent research efforts are directed toward assessing the resulting summaries with a diverse range of stakeholders within the community. Table 2 summarizes participant feedback on each prompt structure and the features that informed the final community-refined prompt.

These findings echo the observation by Anderson et al [2] that prompts emphasizing readability and real-world implications were rated more effective by study authors. Together, these results suggest that audience-sensitive prompt design—guided by both researchers and communities—yields better outcomes than generic summarization.

XSL•FO
RenderX

Table 2. Summary of participant feedback on plain language summary prompt formats.

Prompt	Prompt description	Community group	Key feedback
Prompt A	Summarize the research paper in 600 words at an eighth-grade reading level using a narrative format	Community advisory board	Considered too long and repetitive; participants described it as reading like a condensed research report rather than a summary
Prompt B	Summarize the research paper in 300 words at a sixth-grade reading level using bullet points	Youth development program	Easier to skim and understand but sometimes oversimplified findings and lacked context
Prompt C	Use a structured format, including a title, definitions, key findings, an introduction, main content, and a conclusion	Community advisory board and faith-based community participants	Most preferred format; headings and definitions improved readability and clarity
Final community-refined prompt	Summarize the research paper using simple, community-accessible language (sixth- to eighth-grade reading level). Include a citation, definitions, key findings, an introduction, and a conclusion. Highlight implications for affected communities and environmental justice	All groups combined	Viewed as clear, trustworthy, and relevant; participants valued the presentation of definitions first and contextual explanations of community impact

Discussion

Principal Findings

This study explored how community-engaged feedback can inform the development of prompts for generative AI systems used to produce PLSs of environmental health research. Across 3 community workshops involving youth participants, community advisory board members, and faith-based stakeholders, participants evaluated AI-generated summaries created using different prompt structures. The findings demonstrate that the prompt structure is not a neutral technical input; rather, it shapes how scientific findings are organized, what contextual information is emphasized, and how easily community audiences can identify the relevance of environmental health research.

Participants consistently preferred summaries that were concise, clearly structured, and written at approximately a sixth- to eighth-grade reading level. Narrative summaries generated using a longer prompt format were often perceived as overly dense and difficult to navigate, even when they contained detailed explanations of the research. In contrast, shorter bullet-based summaries were easier to read but were sometimes viewed as overly simplified or lacking sufficient context. The most widely preferred format combined structured headings, definitions of key terms, and concise descriptions of findings with brief narrative explanations of the study's relevance.

Across all workshops, participants emphasized 4 priorities for effective AI-generated summaries: presenting definitions of technical terms before introducing findings, maintaining a concise length of approximately 300 to 400 words, clearly explaining why the research matters, and situating the findings within broader environmental justice contexts. These insights informed the development of a community-refined prompt that incorporated both structured and narrative elements designed to improve readability and contextual relevance.

Together, these findings suggest that prompt design plays a critical role in shaping how generative AI systems translate scientific information for nonspecialist audiences. Therefore, incorporating community perspectives into prompt development may improve the accessibility and usefulness of AI-generated summaries for audiences most affected by environmental health research.

Comparison With Existing Literature

The findings of this study are consistent with prior literature emphasizing the importance of PLSs in improving access to scientific information among nonspecialist audiences. Systematic and scoping reviews have found that PLSs are intended to make research more understandable, accessible, and usable but have also reported substantial variability in summary format, quality, readability, and guidance for authors [1,14]. Our findings support this literature by showing that community audiences value core features commonly recommended for plain language communications including reduced jargon, clear structure, concise wording, and explicit explanation of technical terms. Similarly, emerging research on generative AI in environmental health communication suggests that AI tools may assist in research translation but require careful prompt design to ensure that outputs remain accurate, readable, and relevant to intended audiences [2].

At the same time, our findings extend prior PLS research in several ways. Much of the existing literature focuses on author guidance, publisher requirements, readability metrics, or expert evaluation of summaries [1,14,15]. In contrast, this study examined how intended community audiences evaluate AI-generated summaries and how their feedback can be used to refine the prompts that produce those summaries. This distinction is important because a summary that satisfies general readability criteria may not necessarily meet the information needs of a specific community. Participants in this study did not evaluate summaries only based on the reading level; they also assessed whether the summaries explained why the research mattered, whether key terms were introduced in a logical order,

XSL•FO
RenderX

and whether the findings were connected to environmental health concerns relevant to community experience.

Recent literature has also noted that PLSs vary in their adherence to health literacy principles and that inconsistent guidance may limit their usefulness for public audiences [14,15]. Our findings provide a practical response to this challenge by identifying specific prompt elements that can operationalize health literacy principles in AI-generated summaries. For example, prompts can instruct the model to define technical terms before presenting findings, limit summary length, use headings that guide readers through the content, and include a brief "why this matters" section. These prompt features translate general recommendations for plain language communication into reproducible instructions that can be applied across environmental health studies.

Our results also complicate the assumption that shorter or simpler summaries are always preferable. Participants appreciated the readability of bullet-based summaries, but they also noted that overly brief summaries could omit important context. This finding aligns with recent cautions in AI summarization research that simplification can introduce tradeoffs, including loss of nuance, overgeneralization, or reduced factual completeness [16,17]. In this study, participants preferred a hybrid format that balanced brevity with sufficient explanation. This suggests that readability metrics alone are insufficient for evaluating AI-generated PLSs. A summary may score well on reading-level measures while still failing to explain the significance of the findings or their implications for affected communities.

The findings further reinforce the importance of human review and audience testing in AI-mediated research translation. Recent studies of AI-generated lay summaries have emphasized the need for editorial oversight, domain expertise, and user evaluation to ensure that summaries remain accurate, understandable, and appropriate for intended audiences [16,18,19]. Our findings support these recommendations and suggest that community review may be especially important in environmental health contexts, where scientific findings often intersect with lived experience and local concerns about exposure and risk.

Community-Engaged Environmental Health Translation

The results of this study also align with broader frameworks for community-engaged research, environmental health literacy, and knowledge translation. Environmental health literacy emphasizes not only the ability to understand environmental health information, but also the capacity to use that information in ways that support awareness, decision-making, and action [20]. Similarly, community-engaged environmental health research emphasizes bidirectional communication between researchers and communities, including the translation of scientific findings into formats that are culturally appropriate, relevant, and responsive to community priorities [21].

Our findings illustrate why this bidirectional approach is important for AI-generated research summaries. Participants did not simply ask for summaries to be shorter or easier to read. They asked for summaries that clarified technical language, explained the real-world meaning of findings, and connected research results to environmental justice issues. This emphasis is particularly relevant in environmental health research, where exposure risks, pollution burdens, and health outcomes are often shaped by structural and social dynamics. A generic AI-generated summary may accurately describe a study's methods and results while still failing to communicate why the findings matter to communities experiencing disproportionate environmental burdens.

Recent environmental justice guidance has also emphasized the need for accessible public-facing communication formats, including PLSs, visual tools, and other dissemination products that make scientific information more usable for affected communities [22]. Our findings provide empirical support for these recommendations by identifying features that community stakeholders perceived as most useful in AI-generated environmental health summaries. In particular, participants valued summaries that included definitions, clear organization, concise findings, and explicit discussion of community relevance. These features may help bridge the gap between scientific dissemination and community-centered environmental health communication.

Therefore, this study positions prompt engineering as more than a technical process. In the context of environmental health research translation, prompt design functions as a communication practice that determines how scientific knowledge is framed for public audiences. Decisions about whether to include definitions, how to describe uncertainty, whether to mention environmental justice, and how to explain real-world relevance all shape the meaning of the summary. Community-engaged prompt development offers a mechanism for making these decisions more transparent and responsive to the needs of intended audiences.

Implications for AI-Generated Research Translation

These findings have practical implications for researchers, journals, and technology platforms seeking to integrate generative AI into research communication workflows. First, the results highlight the importance of explicitly specifying audience characteristics in prompt design. Participants strongly preferred summaries written at reading levels consistent with general public literacy levels and organized in ways that supported rapid comprehension. Prompts that simply ask an AI system to "summarize this article in plain language" may be insufficient because they do not specify audience, structure, length, or terminology.

Second, participants emphasized the importance of contextualizing scientific findings within real-world environmental health concerns. In environmental health research, issues such as pollution exposure, structural inequities, and environmental justice are central to understanding the significance of scientific findings. Generic AI-generated summaries may overlook these contextual factors unless prompts explicitly instruct the model to address them.

Third, the study suggests that community engagement can play a meaningful role in shaping how generative AI tools are used

XSL•FO
RenderX

in research dissemination. Rather than relying solely on standardized prompts developed by publishers or software developers, researchers may benefit from incorporating feedback from community stakeholders who represent intended audiences for research findings. This participatory approach aligns with broader frameworks for community-engaged research and knowledge translation, which emphasize collaborative processes for interpreting and communicating scientific evidence [20,21,23].

An important challenge in AI-mediated summarization is that the priorities researchers seek to communicate may not always fully align with the issues that communities view as most relevant. Researchers may emphasize methodological rigor, statistical findings, or theoretical implications, whereas community audiences may be more interested in practical health implications, environmental justice concerns, or actions that could reduce exposure risks. Generative AI tools provide a potential mechanism for balancing these perspectives because prompts can be structured to incorporate both researcher-defined priorities and community-informed framing. For example, prompts may require summaries to include core scientific findings identified by the authors while also incorporating sections that explicitly address community relevance or real-world implications. In this way, AI-mediated summarization does not replace researcher interpretation but instead enables multiple layers of translation that reflect both scientific priorities and community information needs. Although journals often rely on standardized summary formats to maintain consistency and scalability, generative AI systems may allow authors to generate both standardized summaries and audience-adapted versions using alternative prompts tailored to specific audiences or communities.

Limitations

Several limitations should be considered when interpreting these findings. First, the workshops were conducted with community groups located in a single metropolitan area, which may limit the generalizability of the results to other geographic or cultural contexts. Second, most participants were youths involved in a mentoring program. Although this group provided valuable insight into how younger audiences interpret scientific summaries, their preferences may differ from those of other populations, including older adults or individuals with professional experience in environmental health fields. In addition, individual-level demographic information was not collected during the workshops, which limits the ability to analyze how participant characteristics such as age, race, or educational background may have influenced preferences for summary formats.

Third, the study relied on qualitative feedback collected through workshop discussions and facilitator notes. Although this approach allowed participants to provide rich insights into how they interpreted different summaries, the study did not include quantitative measures of readability, comprehension, or trustworthiness. Consequently, the findings reflect participant perceptions rather than objective performance comparisons between prompt structures.

Finally, the AI-generated summaries evaluated in this study were produced using a specific generative AI system and a limited set of environmental health articles. Results may vary when using different models, prompt configurations, or research topics.

Future Directions

Future research should examine how community-informed prompt design influences the effectiveness of AI-generated summaries across different populations and communication settings through more rigorous study designs. Quantitative studies measuring readability, comprehension, and trust in AI-generated summaries could help evaluate whether community-informed prompts improve audience understanding compared with generic summarization approaches.

In addition, future work could explore how AI-generated summaries might be adapted for multilingual or multimodal communication formats. Several workshop participants noted that alternative formats such as infographics, short videos, or audio summaries may further improve accessibility for individuals with varying literacy levels or learning preferences. At the same time, ensuring that simplified or AI-generated summaries do not omit important context or lead to misinterpretation will be an important challenge. Therefore, future implementations may benefit from transparent prompt design and continued researcher or editorial oversight to ensure that summaries accurately convey key findings and limitations.

Finally, the integration of community-informed prompts into journal publication workflows represents an important area for future exploration. Journals and publishers increasingly require PLSs as part of the publication process, and generative AI tools may assist authors in producing these summaries efficiently. Incorporating community perspectives into prompt design could help ensure that these summaries communicate research findings in ways that are both scientifically accurate and meaningful to affected communities.

Conclusions

This study adds to the literature on PLSs, AI-generated lay communication, and community-engaged environmental health translation by demonstrating that community feedback can directly inform prompt design for generative AI systems. The rapid emergence of generative AI systems has created new opportunities to automate aspects of this process, but the effectiveness of AI-generated summaries depends heavily on how prompts are designed. As the discourse evolves, ongoing research on prompt crafting and its ethical implications within environmental contexts remains imperative for achieving just and equitable outcomes in AI-generated information. This study demonstrates that engaging community stakeholders in prompt development can help identify structural and contextual features that make AI-generated summaries more understandable, relevant, and trustworthy for nonspecialist audiences. By incorporating community feedback into prompt engineering, researchers can better align AI-generated summaries with the informational needs of populations most affected by environmental health risks. As generative AI becomes more widely integrated into research dissemination,

XSL•FO
RenderX

community-informed approaches to prompt design may provide an important pathway for improving the translation of scientific knowledge into accessible and actionable information.

Acknowledgments

The authors are grateful to Kaisey Maier for her supervision of EMM during her internship at the Envirome Institute and for attending and coordinating the feedback sessions and author review process.

Due to the nature of this study, the authors used GPT-4o (ChatGPT, OpenAI) [22] to assist with plain language summarization and prompt formatting for the focus groups; these outputs were further reviewed and revised by the study team. Artificial intelligence tools were not used to generate the manuscript text itself, except for minor editing assistance.

Funding

This work was supported in part by the National Institutes of Health grant and the Owsley Brown II Family Foundation.

Data Availability

The qualitative data generated during this study consist of facilitator notes and thematic coding summaries and are not publicly available due to the potential identifiability of participants. Deidentified excerpts may be available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: EMM
Formal analysis: EMM, TS, LBA
Methodology: EMM, TS
Writing—original draft: KOC
Writing—review and editing: KOC, LBA, TS

Conflicts of Interest

None declared.

Multimedia Appendix 1
Plain language summary protocol.
[DOCX File , 14 KB - jmir_v28i1e87118_app1.docx]

Multimedia Appendix 2
Complete plain language summary examples from focus groups.
[DOCX File , 18 KB - jmir_v28i1e87118_app2.docx]

Multimedia Appendix 3
Feedback categorization for summaries.
[DOCX File , 16 KB - jmir_v28i1e87118_app3.docx]

References

1. Stoll M, Kerwer M, Lieb K, Chasiotis A. Plain language summaries: a systematic review of theory, guidelines and empirical research. PLoS One 2022;17(6):e0268789 [FREE Full text] [doi: 10.1371/journal.pone.0268789] [Medline: 35666746]
2. Anderson LB, Kanneganti D, Houk MB, Holm RH, Smith T. Generative AI as a tool for environmental health research translation. Geohealth 2023 Jul;7(7):e2023GH000875 [FREE Full text] [doi: 10.1029/2023GH000875] [Medline: 37502196]
3. Grimshaw JM, Eccles MP, Lavis JN, Hill SJ, Squires JE. Knowledge translation of research findings. Implement Sci 2012 May 31;7:50 [FREE Full text] [doi: 10.1186/1748-5908-7-50] [Medline: 22651257]
4. Grimes M, von Krogh G, Feuerriegel S, Rink F, Gruber M. From scarcity to abundance: scholars and scholarship in an age of generative artificial intelligence. Acad Manage J 2023 Dec 19;66(6):1617-1624. [doi: 10.5465/amj.2023.4006]
5. Joosten YA, Israel TL, Head A, Vaughn Y, Villalta Gil V, Mouton C, et al. Enhancing translational researchers' ability to collaborate with community stakeholders: lessons from the Community Engagement Studio. J Clin Transl Sci 2018 Aug;2(4):201-207 [FREE Full text] [doi: 10.1017/cts.2018.323] [Medline: 30820357]

XSL•FO
RenderX

6. Taffere GR, Abebe HT, Zerihun Z, Mallen C, Price HP, Mulugeta A. Systematic review of community engagement approach in research: describing partnership approaches, challenges and benefits. J Public Health (Berl.) 2023 Jan 03;32(2):185-205. [doi: 10.1007/s10389-022-01799-9]
7. Pitcher N, Mitchell D, Hughes C. Template and guidance for writing a Cochrane plain language summary. Cochrane. 2022 Jan. URL: https://training.cochrane.org/handbook/current/chapter-iii#section-iii-3-2 [accessed 2025-10-01]
8. King SR, Pewsey E, Shailes S. An inside guide to eLife digests. Elife 2017 Mar 15;6:25410 [FREE Full text] [doi: 10.7554/eLife.25410] [Medline: 28294939]
9. Sun T, Zheng Z, Yang M, Pan M, Tan Q, Ma Y, et al. Heavy metal exposure during pregnancy and its association with adverse birth outcomes: a cross-sectional study. Geohealth 2025 Oct;9(10):e2025GH001471. [doi: 10.1029/2025GH001471] [Medline: 41049278]
10. Chen B, Zhang Z, Langrené N, Zhu S. Unleashing the potential of prompt engineering for large language models. Patterns (N Y) 2025 Jun 13;6(6):101260 [FREE Full text] [doi: 10.1016/j.patter.2025.101260] [Medline: 40575123]
11. McGraw KE, Riggs DW, Rai S, Navas-Acien A, Xie Z, Lorkiewicz P, et al. Exposure to volatile organic compounds - acrolein, 1,3-butadiene, and crotonaldehyde - is associated with vascular dysfunction. Environ Res 2021 May;196:110903 [FREE Full text] [doi: 10.1016/j.envres.2021.110903] [Medline: 33636185]
12. McGraw KE, Konkle SL, Riggs DW, Rai SN, DeJarnett N, Xie Z, et al. Exposure to volatile organic compounds is associated with hypertension in Black adults: the Jackson Heart Study. Environ Res 2023 Apr 15;223:115384 [FREE Full text] [doi: 10.1016/j.envres.2023.115384] [Medline: 36796615]
13. Wahlang B, Gao H, Rai SN, Keith RJ, McClain CJ, Srivastava S, et al. Associations between residential volatile organic compound exposures and liver injury markers: the role of biological sex and race. Environ Res 2023 Mar 15;221:115228 [FREE Full text] [doi: 10.1016/j.envres.2023.115228] [Medline: 36610539]
14. Pal A, Arnet I, Elger BS, Wangmo T. Practices and barriers in developing and disseminating plain-language resources reporting medical research information: a scoping review. Patient 2024 Sep;17(5):493-518. [doi: 10.1007/s40271-024-00700-y] [Medline: 38878237]
15. Gainey K, Smith J, McCaffery K, Clifford S, Muscat D. Are plain language summaries published in health journals written according to instructions and health literacy principles? A systematic environmental scan. BMJ Open 2024 Nov 27;14(11):e086464 [FREE Full text] [doi: 10.1136/bmjopen-2024-086464] [Medline: 39608994]
16. Salvi RC, Panigrahi S, Jain D, Yadav S, Akhtar MS. Towards understanding LLM-generated biomedical lay summaries. In: Proceedings of the Second Workshop on Patient-Oriented Language Processing. 2025 Presented at: CL4Health 2025; May 3-4, 2025; Albuquerque, NM. [doi: 10.18653/v1/2025.cl4health-1.22]
17. Peters U, Chin-Yee B. Generalization bias in large language model summarization of scientific research. R Soc Open Sci 2025 Apr;12(4):241776 [FREE Full text] [doi: 10.1098/rsos.241776] [Medline: 40309181]
18. Shyr C, Grout RW, Kennedy N, Akdas Y, Tischbein M, Milford J, et al. Leveraging artificial intelligence to summarize abstracts in lay language for increasing research accessibility and transparency. J Am Med Inform Assoc 2024 Oct 01;31(10):2294-2303. [doi: 10.1093/jamia/ocae186] [Medline: 39008829]
19. Rinderknecht E, Engelmann SU, Saberi V, Kirschner C, Kravchuk AP, Schmelzer A, et al. Using ChatGPT-4 for lay summarization in prostate cancer research to advance patient-centered communication: large-scale generative AI performance evaluation. J Med Internet Res 2025 Nov 19;27:e76598 [FREE Full text] [doi: 10.2196/76598] [Medline: 41259712]
20. Lindsey M, Chen SR, Ben R, Manoogian M, Spradlin J. Defining environmental health literacy. Int J Environ Res Public Health 2021 Nov 05;18(21):11626 [FREE Full text] [doi: 10.3390/ijerph182111626] [Medline: 34770140]
21. Community engagement cores. National Institute of Environmental Health Sciences. URL: https://www.niehs.nih.gov/research/supported/centers/core/coe?utm_source=chatgpt.com [accessed 2026-04-29]
22. ChatGPT. URL: https://chat.openai.com/ [accessed 2025-10-29]
23. Zarshenas S, Mosel J, Chui A, Seaton S, Singh H, Moroz S, et al. Recommended characteristics and processes for writing lay summaries of healthcare evidence: a co-created scoping review and consultation exercise. Res Involv Engagem 2023 Dec 20;9(1):121 [FREE Full text] [doi: 10.1186/s40900-023-00531-5] [Medline: 38124104]

Abbreviations

AI: artificial intelligence
IRB: institutional review board
PLS: plain language summary

XSL•FO
RenderX

Edited by A Stone; submitted 04.Nov.2025; peer-reviewed by W Xu, T Gan, T Conway; comments to author 02.Mar.2026; revised version received 29.Apr.2026; accepted 29.Apr.2026; published 13.May.2026.

Please cite as:
Wade K, Anderson LB, Medley EM, Smith T
Plain Language Summarization of Environmental Health Research Using Generative AI: Community-Engaged Qualitative Study
J Med Internet Res 2026;28:e87118
URL: https://www.jmir.org/2026/1/e87118
doi:10.2196/87118
PMID:

©Katherine Wade, Lauren B Anderson, Evelyn M Medley, Ted Smith. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Barriers, Facilitators, and Intention to Use AI for Breast Cancer Diagnosis: Mixed Methods Study Among Austrian Physicians With and Without AI Experience

Marlene Kritz[1,2], PhD; Erol Holawatsch[1], MSc; Doris Anita Behrens[1,3,4], PhD; Walter Hyll[1], PhD

[1]Department for Economy and Health, University for Continuing Education Krems, Krems, Lower Austria, Austria

[2]Curtin School of Allied Health, Faculty of Health Sciences, Curtin University, Perth, Australia

[3]School of Mathematics, Cardiff University, Cardiff, Wales, United Kingdom

[4]Aneurin Bevan University Health Board, Cardiff, Wales, United Kingdom

Corresponding Author:
Marlene Kritz, PhD
Department for Economy and Health
University for Continuing Education Krems
Dr.-Karl-Dorrek-Straße 30
Krems, Lower Austria, 3500
Austria
Phone: 43 069917249320
Email: marlene.kritz@donau-uni.ac.at

Abstract

Background: Artificial intelligence (AI) has demonstrated strong potential in breast cancer diagnostics by improving accuracy, efficiency, and clinical workflow. However, adoption among physicians remains variable. Existing research often overlooks the contextual and experiential differences between clinicians who use AI and those who do not. A comprehensive understanding of barriers and facilitators, especially across user groups, is essential to inform equitable and effective AI implementation in real-world settings.

Objective: This study aimed to (1) identify key barriers and facilitators influencing the use of AI tools in breast cancer diagnostics, with a specific focus on comparing current users and nonusers, and (2) examine how social, technological, and individual-level factors are linked to physicians' attitudes toward AI, intention to use it, and perceived likelihood of future adoption.

Methods: A cross-sectional, embedded mixed methods survey was conducted with 46 Austrian physicians. Quantitative items were based on the technology acceptance model and its extensions. Open-ended responses were analyzed using conventional content analysis and integrated with quantitative results via joint displays. Ordinary least squares regressions examined factors associated with attitudes, intention, and the likelihood of future AI use.

Results: Among the 46 participating physicians, 52% (n=24) reported current AI use. Common facilitators included improved quality of work, efficiency, and expanding knowledge. Nonusers highlighted barriers such as limited access (17/21, 81%), high costs, and lack of training. AI users highlighted barriers related to limited integration with existing systems and concerns about trust. Despite these differences, both groups expressed strong future adoption intentions. Perceiving multiple facilitators was significantly associated with more favorable attitudes (B=0.83; *P*=.02), stronger intention to use AI (B=1.32; *P*=.01), and higher perceived likelihood of future use (B=1.56; *P*=.001). AI-related skills positively predicted intention (B=1.00; *P*=.04) and likelihood of future use (B=1.16; *P*=.01), while colleagues' positive views about AI predicted both attitudes (B=0.34; *P*=.02) and intention (B=0.39; *P*=.01). In contrast, perceiving multiple barriers was associated with lower intention (B=–0.84; *P*=.047) and likelihood (B=–1.48; *P*<.001). Being aged 50 or older was significantly associated with more negative attitudes (B=–1.11; *P*=.002) and lower likelihood of future use (B=–0.82; *P*=.02).

Conclusions: This study offers preliminary insights into the implementation of AI in breast cancer diagnostics within the Austrian health care context. AI adoption appears to be a staged process with evolving support needs. Early-stage users may benefit from improved access and training, while experienced users require support for workflow integration and trust-building. Promoting peer support, addressing demographic disparities, and embedding AI training into clinical routines may support more sustainable and equitable adoption. These findings inform tailored implementation strategies and offer recommendations that may be transferable to other health systems.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e80274) doi:10.2196/80274

KEYWORDS

breast cancer diagnosis; artificial intelligence; AI adoption; barriers and facilitators; mixed methods; physician perceptions; innovation; technology acceptance model; Austria; early detection

Introduction

Background

Breast cancer is the most common cancer among women globally and a leading cause of cancer-related morbidity and mortality [1,2]. Early and accurate diagnosis is essential to improving outcomes, yet remains vulnerable to human error, including interpretive variability and diagnostic inaccuracies [3]. In Austria, nearly 90,000 women were living with breast cancer as of 2022 [4], and increasing incidence projections combined with physician shortages underscore the need for scalable and efficient diagnostic tools [5,6].

Artificial intelligence (AI) systems trained on large datasets have demonstrated strong performance in detecting malignancies and reducing diagnostic errors [7,8]. In some screening contexts, AI has even outperformed radiologists in stand-alone mammography detection tasks [9]. Nevertheless, AI is not intended to replace clinicians, and concerns remain regarding overreliance on automated systems without appropriate human oversight, which could have serious clinical consequences [10]. Accordingly, AI is increasingly conceptualized as a decision-support tool that may enhance diagnostic efficiency, accuracy, and workflow while preserving clinician responsibility [11-13].

Despite its potential, the adoption of AI in clinical practice remains uneven across health care systems [5]. International research consistently shows that integration into routine care is hindered by a combination of organizational, technical, and human factors [14,15]. Understanding barriers and facilitators to AI adoption is therefore critical for enabling ethical, effective, and sustainable implementation [14].

Theoretical Framework

The technology acceptance model (TAM) posits that technology uptake is influenced by perceived usefulness (ie, the expected benefits of use) and perceived ease of use (ie, the anticipated effort required), which together inform users' attitudes and subsequently influence behavioral intention and actual use [16,17]

Extensions of TAM, including TAM2, incorporate subjective norms, capturing the influence of perceived social pressure on adoption decisions [18,19]. The Unified Theory of Acceptance and Use of Technology (UTAUT) further integrates these concepts by identifying by identifying performance expectancy (expected benefits), effort expectancy (expected ease of use), social influence (influence of important others), and facilitating conditions (perceived organizational and technical support) as core determinants of technology acceptance, moderated by individual characteristics such as age, sex, and experience [19].

Across international health technology research, perceived usefulness, usability, and social norms have been consistently associated with clinicians' willingness to adopt digital innovations [14,20,21]. In radiology and oncology settings, perceived benefits such as improved diagnostic accuracy, workflow efficiency, and enhanced quality of patient care are commonly reported facilitators [22-24]. Further health research found that systems perceived as intuitive and well-integrated into clinical workflows are more likely to be adopted, highlighting the importance of design usability and operational fit [21]. Conversely, persistent barriers include limited infrastructure, high costs, lack of training, workflow disruption, and concerns about increased workload [25].

Qualitative evidence further highlights psychological and professional concerns, including limited AI knowledge, lack of explainability, perceived threats to professional autonomy, and fears of role displacement [26]. In radiology in particular, unfamiliarity with AI tools and concerns about reduced autonomy have been repeatedly identified as salient barriers [21]. Beyond traditional TAM and UTAUT constructs, trust has emerged as a critical determinant of AI adoption [27]. Research suggests that system transparency, reliability, and explainability are integral to building clinician trust, which in turn significantly influences the likelihood of adopting AI-based tools [27].

While AI acceptance has been explored in general health care contexts, its specific application in breast cancer diagnostics remains underexamined, despite growing clinical relevance [9]. Country-specific factors, including regulatory frameworks, digital infrastructure, and workforce capacity, are likely impacting the health care system's readiness to adopt AI [28]. More targeted research is therefore needed to understand the factors influencing adoption, particularly within national contexts such as Austria [21].

In addition, international studies have identified a range of barriers and facilitators to AI use and suggest that attitudes may vary by level of AI experience [21,26]. However, few studies have directly compared clinicians who currently use AI with those who do not, limiting insight into how hands-on AI experience influences attitudes, perceived barriers, and implementation readiness.

Objectives

To address the aforementioned gaps, this study investigates AI use in breast cancer diagnostics within the Austrian health care system and explicitly compares physicians with and without direct AI experience.

Specifically, this study has two primary objectives:

1. To identify key barriers and facilitators influencing the adoption of AI for breast cancer diagnostics among Austrian physicians and to compare these across current AI users and nonusers.

XSL•FO
RenderX

2. To examine how technology-related factors (eg, perceived usefulness, ease of use), social factors (eg, peer influence), and individual characteristics (eg, age, sex, AI-related skills) are associated with physicians' attitudes toward AI, behavioral intentions, and perceived likelihood of future AI adoption.

These insights aim to inform practical, evidence-informed, and human-centered strategies for integrating AI into breast cancer diagnostics [29].

Methods

Study Design

We used an embedded mixed methods [30], cross-sectional questionnaire design, adopting a pragmatic epistemological stance [31] to prioritize methodological fit with the study aims. The study examined (1) barriers and facilitators to AI adoption among AI users and nonusers using quantitative descriptive analyses complemented by qualitative open-ended responses, and (2) predictors of attitudes, intentions, and perceived likelihood of future AI use using quantitative regression analyses.

Recruitment and Data Collection

This study was conducted as part of a master's thesis project at the University for Continuing Education Krems, Austria (EH). Eligible participants were practicing physicians involved in breast cancer screening or diagnostics. Exclusion criteria included nonphysicians, retired clinicians, and those not actively engaged in diagnostic work.

Data collection took place over an 8-week period from March to April 2023. A German-language online survey was distributed via predefined professional and institutional channels using email invitations and an electronic flyer. Primary dissemination occurred through the Austrian Medical Chamber (Ärztekammer), complemented by distribution via university channels, social media platforms, and relevant professional networks. Participants accessed the survey via a secure link, and only respondents who consented and met the predefined screening criteria were able to proceed with the survey.

Ethical Considerations

The study received ethical approval from the Ethics Committee of the University for Continuing Education Krems (EKGZ 55/2021-2024). All participants provided informed consent electronically before accessing the survey. Participation was voluntary, and respondents could withdraw at any time without consequences. Data were collected anonymously, with no personally identifiable information recorded. Secure data storage procedures were applied to protect confidentiality. No financial compensation was provided for participation.

Measures

Quantitative survey items were adapted from previously validated TAM [16,20] instruments and informed by relevant digital health research [21]. Items were tailored to the context of breast cancer diagnostics [16,19]. The first section captured sociodemographic data, AI awareness, and usage. Sex was assessed via self-report; sex or sex assigned at birth was not collected. Attitudes toward AI were measured using two 7-point Likert items averaged into a composite score; additional items included perceived AI-related skills, perceived usefulness, and perceived attitudes of colleagues. The second section addressed perceived barriers (eg, limited availability, high costs, usability, data concerns) and facilitators (eg, efficiency, diagnostic precision, stress reduction) using multiple-choice and open-ended items. The final section measured future intention and likelihood of AI use with a single 7-point Likert item, followed by optional free-text elaborations. Item phrasing was adapted to reflect whether participants were current AI users or nonusers.

Qualitative data were collected through open-ended questions embedded in the online survey, which invited participants to elaborate on perceived barriers and facilitators to the use of AI in breast cancer diagnostics. Data collection was limited to the survey format, and no additional qualitative follow-up (eg, interviews) was conducted. The full wording of all the translated survey items is provided in Multimedia Appendix 1.

Data Analysis

Quantitative Data Analysis

All quantitative analyses were performed using Stata (version 18; StataCorp LLC, College Station, TX, USA). Descriptive statistics summarized participant characteristics and key variables, with means and SDs reported for continuous variables, and frequencies and percentages for categorical variables. Between-group comparisons of perceived barriers and facilitators (users vs nonusers) were conducted using Mann–Whitney U tests (α=0.05) due to nonnormal distributions.

To examine factors associated with attitudes, intentions, and the likelihood of future AI use, multiple linear regression analyses (ordinary least squares) were conducted.

Model assumptions were formally assessed using variance inflation factors for multicollinearity, Shapiro–Wilk tests and Q–Q plots for normality of residuals, and Breusch–Pagan and White tests for heteroskedasticity. All regressions were estimated with heteroskedasticity-robust (Huber–White) standard errors. Sensitivity analyses using ordered logistic regression were conducted to account for the ordinal nature of the outcomes. The detailed results of these diagnostics are reported in Multimedia Appendix 2.

Predictors included sociodemographic variables (eg, age and sex), current AI use, perceived usefulness, AI skills, and colleagues' attitudes. Barriers and facilitators were coded as binary variables. To assess robustness, alternative codings were also tested, including any/none dummy coding and continuous sum scores. Results of these sensitivity analyses are reported in the Multimedia Appendix 3. Missing data were minimal, and no imputation procedures were performed. Analyses were conducted using available case data for each analysis. The number of observations included is reported where relevant.

Qualitative Data Analysis

Open-ended survey responses were analyzed using conventional content analysis [32] following an inductive, iterative process.

XSL•FO
RenderX

The first author (MK), who has experience in qualitative and mixed methods research, conducted open coding by assigning descriptive labels to meaningful text segments to capture participants' perspectives. Codes were generated directly from the data and iteratively refined.

Related codes were then grouped into broader categories based on recurring patterns. In a second step, categories were deductively organized into 2 overarching domains, barriers and facilitators, and aligned with the structure of the quantitative survey, consistent with the study's embedded mixed methods design. This enabled qualitative findings to elaborate on and contextualize quantitative results. For example, while a closed-ended item identified "expanding knowledge" as a facilitator, the qualitative category "AI as a Catalyst for Continuous Learning and Professional Growth" provided additional insights, describing AI's role as a training tool and support

To strengthen analytic rigor and reflexivity, the developing coding structure and category definitions were reviewed and discussed with a second researcher (WH). These discussions supported critical reflection and refinement of interpretations. During analysis, the emergence of new categories was monitored, and no substantively new categories were identified in later responses, indicating informational redundancy in the open-ended data [32]. Additional coding details and illustrative quotes are provided in Multimedia Appendix 3.

Data Integration

To address the first research aim, we used joint displays to integrate qualitative and quantitative findings. Two integrated tables, one for barriers and one for facilitators, summarized corresponding categories, representative codes, and relevant quantitative frequencies. This side-by-side comparison enabled us to identify points of convergence and divergence between AI users and nonusers and provided contextual insight that helped explain and expand upon the statistical results [31,33]. The second research aim was addressed using quantitative analyses only. Findings from both research aims and from both qualitative and quantitative strands, were synthesized in the Discussion to generate context-specific and actionable recommendations for supporting AI integration in clinical settings.

Results

Descriptive Statistics

A total of 53 participants completed the survey. Seven participants were excluded: 6 participants because they were not currently working in breast cancer diagnostics and one participant due to missing demographic and professional information. The final sample comprised 46 participants. Of these, most (40/46, 87%) reported being aware of AI technologies for breast cancer diagnosis, and approximately half (24/46, 52%) indicated they currently use AI in their diagnostic practice. The sociodemographic characteristics of the final sample are summarized in Table 1.

XSL•FO
RenderX

Table 1. Sociodemographic characteristics of the participants (N=46).

Characteristic	Values
Age (years) (n=45)[a]	
Mean (SD)	53.8 (8.64)
Range	37-72
Sex	
Female, n (%)	11 (23.9)
Region (n=44)[b], n (%)	
Vienna	32 (72.7)
Lower Austria	4 (8.5)
Styria	4 (8.5)
Upper Austria	2 (4.3)
Corinthia	2 (4.3)
Current position, n (%)	
Self-employed	18 (39.1)
Physician in training (hospital)	2 (4.4)
Senior physician (hospital)	15 (32.6)
Department head (hospital)	5 (10.9)
Other	8 (17.4)
Field of specialization, n (%)	
Radiologist	17 (37)
Gynecologist	9 (19.6)
Internal medicine	3 (6.5)
Oncology	3 (6.5)
Pathology	3 (6.5)
Surgery	2 (4.4)
Rheumatology	1 (2.2)
Other	8 (17.4)
Tasks, n (%)[c]	
Mammography	28 (60.9)
Ultrasound	26 (56.5)
Magnetic Resonance Imaging (MRI)	21 (45.6)
Computed Tomography (CT)	19 (41.3)
Dual-Energy X-ray Absorptiometry (DXA)	15 (32.6)
Positron Emission Tomography (PET)	6 (13)
Angiography/Fluoroscopy	4 (8.7)
Hybrid imaging	3 (6.5)
Experimental imaging	2 (4.3)
Other	13 (28.3)

[a]Age data were missing for 1 participant.

[b]Region data were missing for 2 participants.

[c]Multiple responses possible.

XSL•FO
RenderX

Barriers and Facilitators to AI Use

Responses to barrier items were available for 44 participants, and responses to facilitator items for 41 participants. Qualitative data were derived from open-ended responses provided by 41 physicians (56% AI users). The qualitative analysis yielded 21 codes, comprising 12 facilitator categories and 9 barrier categories related to AI adoption. An integrated overview of qualitative categories and corresponding quantitative frequencies is presented in Tables 2 and 3. Additional details on qualitative categories, codes with illustrative quotes, and aligned quantitative data are provided in Multimedia Appendix 3.

Table 2. An integrated presentation of mixed methods findings on perceived barriers to artificial intelligence adoption among physicians. Quantitative item labels were shortened for table presentation; full questionnaire items are provided in Multimedia Appendix 1. Qualitative codes and illustrative quotes corresponding to each category are provided in Multimedia Appendix 3. Multiple barriers could be endorsed by each participant. There was a statistically significant difference between artificial intelligence users and artificial intelligence nonusers (z=3.879; *P*<.001).

Quantitative item	Overall (N=44), n (%)	AI[a] users (n=23), n (%)	AI nonusers (n=21), n (%)	Qualitative category
High costs	17 (39)	8 (35)	9 (43)	Access, cost, and integration challenges
Not available at the workplace	22 (50)	5 (22)	17 (81)	Access, cost, and integration challenges
Difficult technical integration into existing systems	14 (32)	9 (39)	5 (24)	Access, cost, and integration challenges
Lack of user-friendliness or operability	12 (27)	8 (35)	4 (19)	Usability and training needs
Lack of knowledge to use AI for breast cancer diagnostics	11 (25)	4 (17)	7 (33)	Usability and training needs
IT[b] security concerns	8 (18)	6 (26)	2 (10)	Data privacy and IT security concerns
Data protection concerns	6 (14)	4 (17)	2 (10)	Data privacy and IT security concerns
Lack of effectiveness	5 (11)	3 (13)	2 (10)	Workflow disruption and limited trust
Workflow incompatibility or limited trust in AI	5 (11)	3 (13)	2 (10)	Workflow disruption and limited trust

[a]AI: artificial intelligence.

[b]IT: information technology.

Table 3. An integrated presentation of mixed methods findings on perceived benefits of artificial intelligence adoption among physicians. Quantitative item labels were shortened for table presentation; full questionnaire items are provided in Multimedia Appendix 1. Qualitative codes and illustrative quotes corresponding to each category are provided in Multimedia Appendix 3. Multiple benefits could be endorsed by each participant. There was a statistically significant difference between users and nonusers (z=−2.19; *P*=.03).

Quantitative item	Overall (N=41), n (%)	AI[a] users (n=23), n (%)	AI nonusers (n=18), n (%)	Qualitative category
More accurate diagnostics	26 (63)	14 (61)	12 (67)	Improving diagnostic confidence and trust in accuracy
Quality improvement of work through better patient care	17 (41)	13 (57)	4 (22)	Improving diagnostic confidence and trust in accuracy
Faster task completion	16 (39)	8 (35)	8 (44)	Improved efficiency as a catalyst for enhanced quality of care
Simplifying work processes	8 (20)	4 (17)	4 (22)	Reducing stress and enhancing well-being
Feeling less stressed	6 (15)	3 (13)	3 (17)	Reducing stress and enhancing well-being
More control over workflow	3 (7)	1 (4)	2 (11)	Reducing stress and enhancing well-being

[a]AI: artificial intelligence.

XSL•FO
RenderX

Barriers to AI Adoption

Among participants who completed the barrier items (n=44), most (40/44) reported at least one perceived barrier to AI use. Conventional content analysis identified 9 codes pertaining to barriers, which were organized into 4 overarching categories. Table 2 presents a joint display of perceived barriers to AI adoption, integrating quantitative response frequencies with qualitative categories and codes.

Access, Cost, and Integration Challenges

Infrastructural, financial, and technical limitations were among the most frequently cited barriers to AI adoption. Availability was reported as a significant concern, particularly among nonusers: 81% (17/21) of nonusers indicated that AI tools were not accessible in their workplace, compared with only 22% (5/23) of users (P<.001; z=3.879). Qualitative responses elaborated on these findings, with nonusers noting, “The software is not available in the department,” or “AI tools are not yet integrated in my workplace.” High costs were also raised as a concern, reported by 43% (9/21) of nonusers and 35% (8/23) of users, though this difference was not statistically significant. Several participants described AI as being “too costly to implement here,” underscoring the perceived financial barriers in resource-constrained environments.

In addition to cost, technical integration was reported as a key limitation by 32% (14/44) of participants, with significantly higher prevalence among users (9/23, 39%) compared with nonusers (5/21, 24%). Qualitative accounts contextualized these concerns: several AI users noted that current systems were not fully compatible with existing software, adding rather than reducing workflow burden. As one user explained, “AI doesn’t align well with our current software and therefore adds to the workflow.”

When asked about motives for future use, several participants emphasized that seamless integration and workplace availability are essential preconditions for future use. As one nonuser noted, “AI must be readily available and smoothly integrated to be adopted in daily practice”. Collectively, these results highlight how cost, resource access, and system compatibility may act as gatekeepers that must be addressed to enable broader and sustained AI adoption in clinical settings.

Usability and Training Needs

Both quantitative and qualitative data indicated that access to training and ease of use are critical factors influencing AI adoption. Among current users, 35% (8/23) indicated limited user-friendliness as a challenge. In contrast, 33% (7/21) of nonusers reported a lack of knowledge with AI as a barrier. Qualitative responses provided additional context. One nonuser noted, “I haven’t used AI because it’s still a new concept and training hasn’t been offered,” highlighting the absence of structured onboarding opportunities. This category was also reinforced with responses to the question about what would motivate future use: nonusers, in particular, emphasized the need for training to build confidence and capability.

Data Privacy and IT Security Concerns

Among AI users, concerns about IT security (26% vs 10%) and data protection (17% vs 10%) were slightly more prominent compared with nonusers, though these differences were not statistically significant. Qualitative responses supported these findings, with 3 AI users raising concerns about patient privacy and data security. For example, one respondent noted, “AI brings up data protection issues for patients”, suggesting that security considerations may become increasingly important as AI use expands.

Workflow Disruption and Limited Trust

Challenges related to workflow disruption and trust in AI outputs were primarily reported by current AI users. Although AI was generally perceived to enhance productivity, 11% (5/44) of participants reported a lack of effectiveness as a barrier. This concern was elaborated upon in open-ended responses, which described how AI systems occasionally generated false alarms or required additional verification steps, thereby limiting potential time savings. For example, one user explained, “AI sometimes creates extra work due to false alarms”, highlighting the potential for AI to unintentionally increase workload.

Trust in AI outputs also emerged as an important issue. Several participants described a need to double-check AI-generated results. One user stated, “I trust AI, but I double-check its suggestions”, suggesting that partial trust may reduce the efficiency benefits of AI by necessitating manual oversight. These findings indicate that both technical and psychological barriers, specifically, limited system integration and cautious trust, may attenuate the intended benefits of AI in clinical workflows.

Facilitators to AI Adoption

Overview

Descriptive statistics showed that all AI users (n=23) and most nonusers (16/18) reported at least one perceived benefit of AI. Conventional content analysis identified 10 codes, which were organized into 4 overarching facilitator-related categories. Table 3 provides an integrated overview of perceived facilitators to AI adoption in breast cancer diagnostics to AI adoption, integrating quantitative response frequencies with qualitative categories and codes.

Improving Diagnostic Confidence and Trust in Accuracy

The most frequently endorsed facilitator was the perception that AI improves diagnostic accuracy, reported by 63% (26/41) of participants. Both users and nonusers acknowledged this benefit, suggesting its value extends beyond direct experience. Among users, AI was commonly described as a reliable second opinion that enhanced certainty in complex cases by improving diagnostic accuracy and reducing the risk of misdiagnosis. As one physician noted, “AI provides an extra layer of security in diagnostics,” illustrating how AI can support confidence in clinical decision-making and contribute to the overall quality of diagnostic work. When asked about future use, physicians mentioned that trust in AI accuracy would need to be further strengthened through “guideline recommendations and scientific evidence,” which was considered a motivator for future

XSL•FO
RenderX

adoption. This highlights the importance of scientific validation for reinforcing diagnostic trust as a condition for sustained or future use.

Improved Efficiency as a Catalyst for Enhanced Quality of Care

This category pertains to how AI was perceived as streamlining routine diagnostic tasks, increasing productivity, and enabling clinicians to focus more on patient-centered care and complex decision-making. The link between efficiency gains and improved care quality emerged consistently across qualitative and quantitative responses. Improved efficiency was cited by over one-third of participants. One nonuser explained, "It could reduce the processing time per patient and enable more timely diagnoses."

Similarly, 2 out of 18 nonusers and 1 out of 23 users identified improved workflow as a key advantage. Qualitative responses highlighted that AI integration was often seamless; as one AI user elaborated: "The use is integrated into the software and is not consciously perceived." Improved quality of care was identified by 41% (17/41) of respondents, with users (13/23, 57%) significantly more likely than nonusers (4/18, 22%) to emphasize this benefit (z=–2.185; *P*=.03).

Several AI users commented that AI's efficiency in handling specific tasks allowed them to focus on other responsibilities. One physician stated, "AI lets me focus on patients rather than routine diagnostics," while another noted, "AI saves time on diagnostics, allowing me to focus on complex cases." These findings collectively suggest that AI not only expedited routine processes but also enabled clinicians to concentrate on higher-value tasks, enhancing both productivity and patient care.

AI as a Catalyst for Continuous Learning and Professional Growth

Quantitative results indicated that over one-third of participants (39%) identified AI as a valuable learning tool, with both users and nonusers attributing equal importance to its role in professional development. Qualitative findings supported this view, revealing that AI was perceived as beneficial for both self-directed learning and the training of junior staff. As one participant noted, "As a training tool, AI transforms routine diagnostics into teachable moments," highlighting AI's potential to facilitate on-the-job learning and support ongoing clinical education.

Reducing Stress and Enhancing Well-Being

This category summarizes how AI was perceived to alleviate workload and mitigate stress, offering psychological benefits. Both users and nonusers identified simplifying work (8/41, 20%) and stress reduction (6/41, 15%) as key benefits. For example, one user noted, "AI decreases the pressure on me during busy shifts," suggesting that AI contributes to a more balanced and manageable work environment. Looking forward, participants suggested that future adoption would depend on clear cost-benefit ratios and minimal added complexity. As one respondent put it, "AI will be used more if it delivers measurable improvements without unnecessary complications." This may suggest that emotional and logistical facilitators, such as reduced stress and usability, are contingent on thoughtful system design and implementation.

Summary of Barriers and Facilitators

High costs, limited availability, and training needs were identified by both users and nonusers as key barriers to AI adoption. Nonusers more frequently cited access and capability-related challenges, whereas users highlighted issues related to system integration and workflow disruptions as barriers.

Across both groups, improved diagnostic accuracy emerged as the most frequently mentioned facilitator of AI use in breast cancer diagnostics. Users emphasized several clinical benefits, including improved patient care and increased efficiency. Nonusers also acknowledged AI's potential to support efficiency but placed greater emphasis on the importance of enabling conditions, such as structured training and workplace accessibility, to support future adoption. Addressing both shared and group-specific factors will be essential for developing targeted implementation strategies and ensuring broader, equitable uptake of AI in clinical practice.

Factors Determining Attitude, Intention, and Perceived Likelihood of Future Use

Overview

To address the second research aim, ordinary least squares regression analyses were conducted to identify factors associated with physicians' attitudes, intentions, and perceived likelihood of future AI use. Complete regression data were available for 41 respondents. Detailed descriptive statistics and robustness analyses are reported in Multimedia Appendix 2.

Attitudes Toward AI

The regression results are presented in Table 4 and indicate that physicians aged 50 or older held significantly less favorable attitudes toward AI (B=–1.111; *P*=.002), possibly reflecting generational differences in technological exposure or comfort with AI integration. Perceived social norms also emerged as influential: physicians who reported that their colleagues consider AI important report a significantly more positive attitude (B=0.341; *P*=.02). Additionally, perceiving at least 2 facilitators of AI was significantly associated with a more favorable attitude (B=0.832; *P*=.02), emphasizing the importance of perceived advantages in driving AI acceptance. Conversely, although the association did not reach statistical significance (*P*=.07), perceiving 2 or more barriers to AI adoption was associated with a trend toward less favorable attitudes (B=–0.501, SE=0.268), indicating that concerns about risks, implementation challenges, or ethical issues may influence clinicians' views. Overall, the model accounted for 62.2% of the variance in attitudes (R^2=0.622).

XSL•FO
RenderX

Table 4. Ordinary least squares regression results for factors influencing attitudes, intentions, and perceived likelihood of future artificial intelligence use (N=41).

Independent variables	Attitude toward AI[a] use (R^2=0.62)		Intention to use AI (R^2=0.65)		Likelihood of future AI use (R^2=0.70)	
	B (SE)[b]	*P* value	B (SE)[b]	*P* value	B (SE)[b]	*P* value
Facilitators	0.83 (0.33)	.02	1.32 (0.51)	.01	1.56 (0.42)	<.001
Barriers	−0.50 (0.27)	.07	−0.84 (0.41)	.047	−1.48 (0.37)	<.001
AI user	−0.02 (0.29)	.94	0.66 (0.47)	.17	1.29 (0.47)	.01
Skills for using AI	0.38 (0.32)	.24	1.00 (0.47)	.04	1.16 (0.43)	.01
Colleagues consider AI important	0.34 (0.14)	.02	0.39 (0.14)	.01	−0.03 (0.18)	.87
Colleagues' opinion is important	0.05 (0.08)	.55	−0.04 (0.08)	.66	0.06 (0.10)	.57
Female	−0.24 (0.37)	.52	0.01 (0.42)	.981	−0.86 (0.47)	.08
Age ≥50 years	−1.11 (0.32)	.002	−0.81 (0.40)	.053	−0.82 (0.34)	.02
Constant	4.03 (0.76)	<.001	2.94 (0.83)	.001	4.25 (0.70)	<.001

[a]AI: artificial intelligence.

[b]B indicates regression coefficient; robust standard errors are shown in parentheses.

Intention to Use AI for Breast Cancer Diagnosis in the Future

It was found that physicians who perceived more benefits (*P*=.01) and fewer barriers (*P*=.047) to AI reported a higher intention to adopt it in their practice. This highlights the crucial role of both perceived advantages and reduced obstacles in shaping the willingness to integrate AI into clinical decision-making. Furthermore, physicians who perceived that their colleagues considered AI use important were more likely to express an intention to use AI themselves, underscoring the influence of social norms in professional environments.

Younger physicians demonstrated a higher intention to adopt AI, which, similarly to attitudes, may reflect greater exposure to technological advancements and a higher level of adaptability to digital innovations in health care. Additionally, possessing skills related to AI use was positively associated with the intention to adopt it, reinforcing the importance of competence and confidence in facilitating future implementation. The model explained 64.7% of the variance (R^2=0.647). Taken together, these results underscore the combined influence of perceived benefits, social context, and capability beliefs in influencing physicians' intentions to use AI in the future.

Perceived Likelihood of Future AI Use

The regression analysis for the perceived likelihood of using AI in the future is presented in Table 4. Consistent with the models for attitudes and intention to use AI, perceiving facilitators was positively associated with the likelihood of future AI use, while perceived barriers were negatively associated.

Both current AI use and possessing AI-related skills were statistically significant predictors of the likelihood of using AI in the future for breast cancer diagnostics. While current AI use was not significantly related to attitudes or intentions, it was a significant predictor of perceived likelihood, suggesting that actual use may determine future expectations more than current attitudes.

Physicians aged 50 years and above consistently demonstrated a lower likelihood of future adoption across all 3 models, reaffirming the key role of age in shaping expectations around AI use in clinical practice. In contrast, gender was not a statistically significant predictor in any of the models. For the likelihood of future AI use, the coefficient for female physicians was negative (B=–0.856; *P*=.08), which is not significant at the conventional α of 0.05 level but would be considered marginally significant at a liberal threshold of *P*≤.1.

These findings were robust across alternative model specifications. Facilitator and barrier variables were re-coded using binary indicators (≥1 vs 0) and continuous sum scores, with consistent results. Additional robustness checks using ordered logistic regression and reduced models including only key predictors yielded comparable estimates. Full regression diagnostics and sensitivity analyses are reported in Tables S2-S5 in Multimedia Appendix 2.

Discussion

Principal Findings

This study aimed to examine the key barriers and facilitators to AI for breast cancer diagnosis among Austrian physicians, with a specific focus on comparing current users and nonusers. We also sought to identify how individual, social, and technological factors predict physicians' attitudes toward AI, their intention to use it, and their perceived likelihood of future adoption. While most participants were aware of AI technologies, only approximately half had used such technologies in their clinical practice.

Key facilitators included perceiving benefits in terms of diagnostic accuracy, professional growth, workflow efficiency, and overall quality of patient care. Barriers included limited access, high costs, integration issues, and insufficient training opportunities. Nonusers reported anticipated benefits, such as improved diagnostic accuracy and workflow efficiency, but identified limited availability as a key barrier to adoption. In

XSL•FO
RenderX

contrast, users cited concrete benefits, including enhanced diagnostic precision, quality of care, and productivity, while also expressing some concerns regarding cost, integration with existing systems, and trust in AI-generated outputs.

Regression analyses indicated that perceptions of more facilitators and fewer barriers were strongly associated with more favorable attitudes and higher adoption intentions. Social and demographic factors also played a role: younger physicians and those in environments where colleagues valued AI were more likely to have favorable attitudes and stronger intentions to use AI. Experience with AI and related skills were consistent predictors of intention and perceived likelihood of future use, while older age was negatively associated with adoption likelihood.

Comparison With Prior Work and Theoretical Implications

Awareness and Use of AI

Our findings align with previous work demonstrating moderate levels of AI use among clinicians. Approximately half of our sample reported current AI use, comparable to usage rates found by Shinners et al [34] and Zanardo et al [35], who noted usage rates of 50% and 48%, respectively. In contrast, Chen et al [5] found a usage rate of 72%, possibly due to their sample being composed of resident radiologists. This aligns with prior work suggesting that physicians in training may exhibit higher readiness to use AI [34]. The slightly higher unawareness rate in our study (13%) compared with other specialties (eg, 4.3% in Chen et al [5]) may reflect our focus on breast cancer diagnostics rather than general radiology.

Barriers and Facilitators to AI Use

Consistent with prior international research, improved diagnostic accuracy and workflow efficiency emerged as the most frequently cited facilitators of AI adoption [20,21,24]. Theoretically, these results align with the UTAUT, particularly the construct of performance expectancy as a key determinant of behavioral intention [19]. Studies among radiologists and oncologists in Europe and Asia similarly identify performance expectancy as a key driver of favorable attitudes toward AI [22,35]. However, our findings indicate that these perceived benefits are also widely acknowledged by clinicians without direct AI experience, suggesting that positive attitudes alone are insufficient to prompt adoption in the absence of appropriate structural support.

In this context, the prominence of structural and organizational barriers among nonusers becomes particularly salient. Limited access, costs, and training deficits, reported primarily by clinicians without AI experience, align with international literature identifying resource constraints as critical barriers to digital health implementation [14,25]. These findings align with UTAUT's emphasis on facilitating conditions, such as infrastructure, institutional support, and technical resources, as key precursors of technology use [19]. This pattern is also consistent with recent CFIR-guided evidence syntheses, which highlight that safe and sustainable AI adoption in breast screening requires attention not only to perceived benefits, but also to technological integration, training, trust, and governance [15].

The strong emphasis on training needs, particularly among nonusers, aligns with the UTAUT construct of effort expectancy and TAM's notion of perceived ease of use, both of which highlight the importance of minimizing the perceived complexity of new technologies [16,19]. The finding that nonusers expressed willingness to adopt AI if training were available suggests that lower perceived effort through structured onboarding may act as a critical enabler. This finding reinforces the need for user-friendly systems and accessible, targeted training programs, especially for clinicians with limited prior exposure to AI.

The finding that AI users raised concerns related to workflow integration and trust in outputs is consistent with other studies demonstrating that real-world adoption surfaces unique operational and cognitive demands [26]. Notably, lack of trust, especially regarding the accuracy and reliability of AI outcomes, has been widely identified as a barrier to adoption [27,36]. These concerns highlight trust as a central determinant that interacts with, and potentially moderates, traditional TAM/UTAUT constructs, aligning with growing calls to integrate explainability and transparency into health technology acceptance frameworks [15,27]. In our study, participants also highlighted that scientific guidelines could enhance trust in AI outputs, particularly among those who were hesitant to rely on algorithmic support without established clinical validation. Taken together, the contrast between users and nonusers underscores that AI adoption may not be a binary decision but a staged process, in which anticipated barriers dominate early adoption phases, while experiential challenges emerge with sustained use.

Influences on Attitudes, Intentions, and Likelihood of Future AI Use

While prior work has focused on attitudes as the primary driver of behavioral intention [5], our findings emphasize the additive role of perceived barriers and facilitators. A greater number of perceived barriers corresponded with lower adoption intentions, while more perceived benefits predicted stronger intentions. These results underscore the need for a dual strategy: promoting perceived advantages while actively addressing barriers.

Peer perceptions also played an important role. Physicians who perceived that AI was valued by their colleagues were more likely to report positive attitudes and stronger intentions to adopt the technology. The finding that perceiving benefits was linked to intention and the role of subjective norms is consistent with TAM 2 and UTAUT, which both emphasize perceived usefulness and social influence as key determinants of adoption behavior [18,19]. Peer endorsement may support adoption, particularly among clinicians with limited AI experience, consistent with recommendations to incorporate social and organizational influences, such as colleague support, into models of technology uptake in health care [20].

Age was a consistent negative predictor across all models, with older physicians reporting significantly lower attitudes, intention, and likelihood of future AI adoption. This may reflect generational differences in technology familiarity, perceived

XSL•FO
RenderX

disruption to routine, or risk tolerance [5,27]. Conversely, prior experience with AI in clinical practice was positively associated with future intention to use, aligning with earlier findings that experience and exposure enhance acceptance [5,37].

Taken together, these findings extend international evidence by demonstrating that differences between AI users and nonusers are not merely attitudinal but reflect distinct stages of implementation readiness, with implications for how AI adoption strategies should be tailored across clinical contexts. Recognizing this progression may inform more targeted and flexible strategies. However, these patterns are preliminary and hypothesis-generating and should be confirmed in larger samples.

Practical Implications

Practically, our findings offer valuable insights for health care leaders, educators, and policymakers aiming to support the responsible integration of AI in breast cancer diagnostics. Implementation strategies could benefit from not only highlighting perceived advantages but also addressing specific barriers that appear to differ between current users and nonusers.

As outlined in Table 5, potential strategies may include consideration of demographic differences (eg, age), social norms, and levels of technical readiness. Targeted support, such as age-appropriate training, peer-led initiatives, and structured upskilling opportunities, might help reduce disparities in AI engagement. Findings suggest that adoption may not be a binary decision but a staged process with evolving support needs. Early-stage users may benefit from structural interventions (eg, access and training), while more experienced users may require support with integration and workflow adaptation. This distinction implies that implementation efforts may need to be adapted over time and tailored to users' varying levels of familiarity and experience with AI.

Table 5. Summary of the recommendations based on key findings for artificial intelligence use in health care.

Factor	Key finding	Recommendation
Demographic factors	Older age was associated with less favorable attitudes toward AI[a] and lower intention and likelihood of future use.	Develop age-responsive training and support strategies to address barriers experienced by older clinicians.
Barriers	Perceived barriers were negatively associated with attitudes, intention, and likelihood of AI use.	Address identified barriers through targeted education, technical support, and organizational resources to improve attitudes and facilitate adoption.
Facilitators	Perceived facilitators were positively associated with attitudes, intention, and likelihood of AI use.	Promote success stories and demonstrated benefits of AI to strengthen positive perceptions and encourage adoption.
Knowledge and use of AI	Greater AI-related skills were associated with higher intention and likelihood of future use, but not with attitudes. Current AI users were more likely to report future use.	Invest in practical AI training programs to build skills and confidence. Support experienced users to act as champions and mentors for wider adoption.
Social factors	Colleagues' perception of AI's importance positively influenced attitude and intention, but not the likelihood of future AI use.	Foster a supportive workplace culture through peer-led initiatives and visible endorsement by respected colleagues to enhance motivation and engagement.

[a]AI: artificial intelligence.

Furthermore, the influence of social context, particularly perceptions of colleagues' attitudes, suggests that leveraging peer support and fostering a culture of openness toward innovation may play a role in facilitating adoption. Finally, embedding AI-related content into medical education and continuing professional development may support sustained engagement, particularly among groups less likely to adopt AI initially. These insights, while preliminary, could inform more inclusive and context-sensitive implementation strategies in future work.

Future Research Directions

Future research could extend our findings through longitudinal studies that track real-world AI adoption over time, as well as intervention studies evaluating tailored support strategies. Comparative research across countries and medical specialties could further clarify how structural, educational, and cultural factors influence implementation [22,38]. In our sample, differences in access, training, and institutional support between users and nonusers highlight the need for flexible, context-sensitive implementation strategies. A governance framework, such as that proposed by Hassan et al [27], may support this by addressing trust-related barriers through structured oversight across the AI lifecycle and by informing policy and institutional priorities that enable more equitable access, investment, and training.

Our findings also support recent calls to expand AI education across the medical training pipeline. For example, Weidener and Fischer [39] emphasized the integration of AI ethics and practical skills in undergraduate curricula. Similarly, Wartman and Combs [40] argued that medical education must evolve beyond the information age to meaningfully prepare future clinicians for AI-enabled practice. Training physicians in AI as part of undergraduate and continuing medical education may enhance adoption by improving relevant skills, confidence, and preparedness for clinical integration, particularly among younger cohorts already engaging with AI informally [39].

Future research could also explore patient perspectives, particularly regarding data privacy, consent, and trust in

XSL•FO
RenderX

AI-assisted diagnostics, to support ethical and patient-centered implementation. Although our study focused on physicians, these concerns are especially relevant in breast cancer screening, where patient acceptance may hinge on transparency, perceived fairness in algorithmic decision-making, and how AI is integrated into clinical workflows [41-43].

Strengths and Limitations

A key strength of this study is its mixed methods design, which enabled triangulation of quantitative findings with qualitative insights. This approach allowed for a comprehensive understanding of AI adoption by comparing users and nonusers, and by identifying psychological, technical, and social drivers of adoption.

However, some limitations must be acknowledged. First, the sample was relatively small and limited to Austria, which may affect generalizability. Second, the cross-sectional design captures perceptions at a single point in time and does not allow for causal inference. Longitudinal research is needed to track evolving attitudes and behaviors. Third, reliance on self-reported data introduces potential bias, which could be addressed in future studies through objective usage metrics. Fourth, although items were adapted from validated TAM instruments and pilot tested for clarity, formal psychometric validation was not performed. The brief attitude measure and exploratory design, therefore, warrant cautious interpretation of these findings.

Regarding the quantitative analyses, the small sample size and the number of predictors in the regression models may limit statistical power and the stability of estimates. To address these concerns, we conducted extensive sensitivity and robustness checks, including alternative codings of barriers and facilitators, reduced models including only the main predictors, and ordered logistic regressions to account for the ordinal nature of the outcomes. These analyses indicate that the direction and magnitude of the observed effects are largely robust, although statistical significance varies slightly across specifications. Finally, recruitment via professional networks and social media groups meant that the total number of physicians reached was unknown, and a formal response rate could not be calculated. This may introduce self-selection bias and limit the generalizability of the findings beyond physicians with an interest or experience in AI. Despite these limitations, our study offers valuable exploratory insights into the factors influencing AI adoption in breast cancer diagnosis and provides a foundation for future implementation research and policy development.

Conclusions

This study provides timely insight into AI adoption for breast cancer diagnostics among Austrian physicians and highlights several key contributions. First, it focuses on a specific clinical use case (breast cancer diagnostics) within a national health care context, extending prior work that has often examined AI adoption in radiology more broadly. Second, by comparing physicians with and without hands-on AI experience, it shows that barriers, facilitators, and support needs differ by level of AI exposure.

Third, the findings suggest that AI adoption may be better understood as a staged process rather than a binary decision. Early uptake appears to depend mainly on access and training, whereas later-stage use is more affected by workflow integration and trust in AI outputs. This refines existing TAMs by highlighting the role of implementation context across stages of engagement.

Finally, the results have practical implications for implementation. Investment in targeted training, organizational support, and peer-enabled learning may support more equitable adoption. Strategies that account for differences by experience level and demographic factors may be particularly important for achieving sustainable and responsible integration of AI in breast cancer diagnostics.

Acknowledgments

None declared.

Data Availability

Due to ethical restrictions related to participant confidentiality, the data supporting the findings of this study are not publicly available. However, anonymized data may be made available upon request to the corresponding author.

Funding

This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Authors' Contributions

MK and WH contributed equally to the conceptualization and methodology of the study, based on the original study idea developed by EH. Data curation was shared between MK (qualitative and mixed methods) and WH (quantitative). Formal analysis was led by WH for quantitative data, with support from MK, while MK led the qualitative analysis with support from WH. Mixed methods integration was led by MK, with support from WH. EH led participant recruitment and data collection, project administration, and provision of resources with support from MK. Supervision was led by MK. Visualization was led by MK for qualitative and mixed methods outputs (and supported quantitative visualization), while WH led quantitative visualization. MK led the writing of the original draft, with support from WH. Review and editing of the manuscript were led by MK, with equal contributions

XSL•FO
RenderX

from WH and DB, and additional support from EH. All authors meet the ICMJE authorship criteria, have approved the final version of the manuscript, and take responsibility for their respective contributions.

Conflicts of Interest

None declared.

Multimedia Appendix 1
English translation of the questionnaire and mapping of questionnaire items to the analyses and reporting presented in the main manuscript.
[DOCX File , 40 KB - jmir_v28i1e80274_app1.docx]

Multimedia Appendix 2
Qualitative categories, codes, and illustrative quotes relating to barriers and facilitators of artificial intelligence adoption in breast cancer diagnostics.
[DOCX File , 29 KB - jmir_v28i1e80274_app2.docx]

Multimedia Appendix 3
Additional regression analyses and diagnostics, including descriptive statistics, model diagnostics, robustness checks, ordered logistic regression models, and reduced regression models predicting attitudes, intentions, and likelihood of future artificial intelligence use in breast cancer diagnostics.
[DOCX File , 38 KB - jmir_v28i1e80274_app3.docx]

References

1. Breast cancer. World Health Organization. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/breast-cancer [accessed 2026-03-27]
2. Wilkinson L, Gathani T. Understanding breast cancer as a global health concern. Br J Radiol 2022;95(1130):20211033 [FREE Full text] [doi: 10.1259/bjr.20211033] [Medline: 34905391]
3. Brady AP. Error and discrepancy in radiology: inevitable or avoidable? Insights Imaging 2017;8(1):171-182 [FREE Full text] [doi: 10.1007/s13244-016-0534-1] [Medline: 27928712]
4. Breast cancer statistics. Statistik Austria. 2022. URL: https://www.statistik.at [accessed 2026-03-27]
5. Chen Y, Wu Z, Wang P, Xie L, Yan M, Jiang M, et al. Radiology residents' perceptions of artificial intelligence: nationwide cross-sectional survey study. J Med Internet Res 2023;25:e48249 [FREE Full text] [doi: 10.2196/48249] [Medline: 37856181]
6. Aluttis C, Bishaw T, Frank MW. The workforce for health in a globalized context--global shortages and international migration. Glob Health Action 2014;7:23611 [FREE Full text] [doi: 10.3402/gha.v7.23611] [Medline: 24560265]
7. Åström J, Reim W, Parida V. Value creation and value capture for AI business model innovation: a three-phase process framework. Rev Manag Sci 2022;16(7):2111-2133. [doi: 10.1007/s11846-022-00521-z]
8. Jeong J, Kim S, Pan L, Hwang D, Kim D, Choi J, et al. Reducing the workload of medical diagnosis through artificial intelligence: a narrative review. Medicine (Baltimore) 2025;104(6):e41470 [FREE Full text] [doi: 10.1097/MD.0000000000041470] [Medline: 39928829]
9. Chang YW, Ryu JK, An JK, Choi N, Park YM, Ko KH, et al. Artificial intelligence for breast cancer screening in mammography (AI-STREAM): preliminary analysis of a prospective multicenter cohort study. Nat Commun 2025;16(1):2248 [FREE Full text] [doi: 10.1038/s41467-025-57469-3] [Medline: 40050619]
10. Freeman K, Geppert J, Stinton C, Todkill D, Johnson S, Clarke A, et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. BMJ 2021;374:n1872 [FREE Full text] [doi: 10.1136/bmj.n1872] [Medline: 34470740]
11. Leibig C, Brehmer M, Bunk S, Byng D, Pinker K, Umutlu L. Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis. Lancet Digit Health 2022;4(7):e507-e519 [FREE Full text] [doi: 10.1016/S2589-7500(22)00070-X] [Medline: 35750400]
12. Krittanawong C. The rise of artificial intelligence and the uncertain future for physicians. Eur J Intern Med 2018;48:e13-e14. [doi: 10.1016/j.ejim.2017.06.017] [Medline: 28651747]
13. Marinovich ML, Wylie E, Lotter W, Lund H, Waddell A, Madeley C, et al. Artificial intelligence (AI) for breast cancer screening: BreastScreen population-based cohort study of cancer detection. EBioMedicine 2023;90:104498 [FREE Full text] [doi: 10.1016/j.ebiom.2023.104498] [Medline: 36863255]
14. Wolff J, Pauling J, Keck A, Baumbach J. Success factors of artificial intelligence implementation in healthcare. Front Digit Health 2021;3:594971 [FREE Full text] [doi: 10.3389/fdgth.2021.594971] [Medline: 34713083]
15. Goh S, Goh RSJ, Chong B, Ng QX, Koh GCH, Ngiam KY, et al. Challenges in implementing artificial intelligence in breast cancer screening programs: systematic review and framework for safe adoption. J Med Internet Res 2025;27:e62941 [FREE Full text] [doi: 10.2196/62941] [Medline: 40373301]

XSL•FO
RenderX

16. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989;13(3):319-340. [doi: 10.2307/249008]
17. Ajzen I. From Intentions to Actions: A Theory of Planned Behavior. In: Kuhl J, Beckmann J, editors. Action Control. Berlin, Heidelberg: Springer Berlin Heidelberg; 1985:11-39.
18. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: four longitudinal field studies. Manag Sci 2000;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
19. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003;27(3):425-478 [FREE Full text]
20. Holden RJ, Karsh B. The technology acceptance model: its past and its future in health care. J Biomed Inform 2010;43(1):159-172 [FREE Full text] [doi: 10.1016/j.jbi.2009.07.002] [Medline: 19615467]
21. Eltawil FA, Atalla M, Boulos E, Amirabadi A, Tyrrell PN. Analyzing barriers and enablers for the acceptance of artificial intelligence innovations into radiology practice: a scoping review. Tomography 2023;9(4):1443-1455 [FREE Full text] [doi: 10.3390/tomography9040115] [Medline: 37624108]
22. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
23. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019;6(2):94-98 [FREE Full text] [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
24. Oh S, Kim JH, Choi SW, Lee HJ, Hong J, Kwon SH. Physician confidence in artificial intelligence: an online mobile survey. J Med Internet Res 2019;21(3):e12422 [FREE Full text] [doi: 10.2196/12422] [Medline: 30907742]
25. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, Martinez EZ, Zucoloto ML, Østengaard L, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023;6(1):161 [FREE Full text] [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
26. Hoffman J, Wenke R, Angus RL, Shinners L, Richards B, Hattingh L. Overcoming barriers and enabling artificial intelligence adoption in allied health clinical practice: a qualitative study. Digit Health 2025;11:20552076241311144 [FREE Full text] [doi: 10.1177/20552076241311144] [Medline: 39906878]
27. Hassan M, Borycki EM, Kushniruk AW. Artificial intelligence governance framework for healthcare. Healthc Manage Forum 2025;38(2):125-130 [FREE Full text] [doi: 10.1177/08404704241291226] [Medline: 39470044]
28. Joshi S, Sharma M, Das RP, Rosak-Szyrocka J, Żywiołek J, Muduli K, et al. Modeling conceptual framework for implementing barriers of AI in public healthcare for improving operational excellence: experiences from developing countries. Sustainability 2022;14(18):11698. [doi: 10.3390/su141811698]
29. Hardy M, Harvey H. Artificial intelligence in diagnostic imaging: impact on the radiography profession. Br J Radiol 2020;93(1108):20190840 [FREE Full text] [doi: 10.1259/bjr.20190840] [Medline: 31821024]
30. Creswell JW, Plano Clark VL. Designing and Conducting Mixed Methods Research. Thosands Oak, California: SAGE Publications; 2018.
31. Maxcy MS. Pragmatic threads in mixed methods research in the social sciences: The search for multiple modes of inquiry and the end of the philosophy of formalism. In: Handbook of Mixed Methods in Social and Behavioral Research. Thousand Oak, California: Sage; 2003:51-89.
32. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
33. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs-principles and practices. Health Serv Res 2013;48(6 Pt 2):2134-2156 [FREE Full text] [doi: 10.1111/1475-6773.12117] [Medline: 24279835]
34. Shinners L, Grace S, Smith S, Stephens A, Aggar C. Exploring healthcare professionals' perceptions of artificial intelligence: piloting the shinners artificial intelligence perception tool. Digit Health 2022;8:20552076221078110 [FREE Full text] [doi: 10.1177/20552076221078110] [Medline: 35154807]
35. Zanardo M, Visser JJ, Colarieti A, Cuocolo R, Klontzas ME, Pinto Dos Santos D, European Society of Radiology (ESR). Impact of AI on radiology: a EuroAIM/EuSoMII 2024 survey among members of the European society of radiology. Insights Imaging 2024;15(1):240. [doi: 10.1186/s13244-024-01801-w] [Medline: 39373853]
36. Kelly S, Kaye S, Oviedo-Trespalacios O. What factors contribute to the acceptance of artificial intelligence? a systematic review. Telemat Inform 2023;77:101925. [doi: 10.1016/j.tele.2022.101925]
37. AlZaabi A, AlMaskari S, AalAbdulsalam A. Are physicians and medical students ready for artificial intelligence applications in healthcare? Digit Health 2023;9:20552076231152167 [FREE Full text] [doi: 10.1177/20552076231152167] [Medline: 36762024]
38. Reddy S, Fox J, Purohit MP. Artificial intelligence-enabled healthcare delivery. J R Soc Med 2019;112(1):22-28 [FREE Full text] [doi: 10.1177/0141076818815510] [Medline: 30507284]
39. Weidener L, Fischer M. Artificial intelligence in medicine: cross-sectional study among medical students on application, education, and ethical aspects. JMIR Med Educ 2024;10:e51247 [FREE Full text] [doi: 10.2196/51247] [Medline: 38180787]
40. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. Acad Med 2018;93(8):1107-1109. [doi: 10.1097/ACM.0000000000002044] [Medline: 29095704]

XSL•FO
RenderX

41. Longoni C, Bonezzi A, Morewedge CK. Resistance to medical artificial intelligence. J Consum Res 2019;46(4):629-650. [doi: 10.1093/jcr/ucz013]
42. Moy S, Irannejad M, Manning SJ, Farahani M, Ahmed Y, Gao E, et al. Patient perspectives on the use of artificial intelligence in health care: a scoping review. J Patient Cent Res Rev 2024;11(1):51-62 [FREE Full text] [doi: 10.17294/2330-0698.2029] [Medline: 38596349]
43. Ozcan BB, Dogan BE, Xi Y, Knippa EE. Patient perception of artificial intelligence use in interpretation of screening mammograms: a survey study. Radiol Imaging Cancer 2025;7(3):e240290. [doi: 10.1148/rycan.240290] [Medline: 40249272]

Abbreviations

AI: artificial intelligence
TAM: technology acceptance model
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by A Stone; submitted 14.Jul.2025; peer-reviewed by S Mohamed Shaffi, C Anyaegbuna; comments to author 24.Sep.2025; revised version received 17.Feb.2026; accepted 18.Feb.2026; published 09.Jun.2026.

Please cite as:
Kritz M, Holawatsch E, Behrens DA, Hyll W
Barriers, Facilitators, and Intention to Use AI for Breast Cancer Diagnosis: Mixed Methods Study Among Austrian Physicians With and Without AI Experience
J Med Internet Res 2026;28:e80274
URL: https://www.jmir.org/2026/1/e80274
doi:10.2196/80274
PMID:

©Marlene Kritz, Erol Holawatsch, Doris Anita Behrens, Walter Hyll. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Multitask Artificial Intelligence–Based Electrocardiogram Tool for Preoperative Cardiac Testing in Noncardiac Surgery: Retrospective Cohort Study of Health Care Utilization and Costs

Hong-Mi Choi[1,2*], MD; Yerin Kim[1,2*], MD; Joonghee Kim[3,4], MD, PhD; Jiesuck Park[1,2], MD; In-Chang Hwang[1,2], MD; Yun Young Choi[1,2], MD, PhD; Ji Hyun Lee[1,2], MD, PhD; Yeonyee E Yoon[1,2], MD, PhD; Il-Young Oh[1,2], MD, PhD; Goo-Yeong Cho[1,2], MD, PhD; In-Ae Song[5,6], MD, PhD; Youngjin Cho[1,2,4], MD, PhD

[1]Department of Cardiology, Cardiovascular Center, Seoul National University Bundang Hospital, Seongnam, Gyeonggi, Republic of Korea
[2]Department of Internal Medicine, College of Medicine, Seoul National University, Seoul, Republic of Korea
[3]Department of Emergency Medicine, Seoul National University Bundang Hospital, Seongnam, Gyeonggi, Republic of Korea
[4]ARPI Inc, Seongnam, Gyeonggi, Republic of Korea
[5]Department of Anesthesiology and Pain Medicine, Seoul National University Bundang Hospital, Seongnam, Gyeonggi, Republic of Korea
[6]Department of Anesthesiology and Pain Medicine, College of Medicine, Seoul National University, Seoul, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Youngjin Cho, MD, PhD
Department of Cardiology, Cardiovascular Center
Seoul National University Bundang Hospital
82 Gumi-ro-173-gil, Bundang
Seongnam, Gyeonggi, 13620
Republic of Korea
Phone: 82 31 787 9302
Email: cho_y@snubh.org

Abstract

Background: Preoperative cardiovascular risk stratification is essential in noncardiac surgery, but conventional testing is frequently overused, increasing costs without improving outcomes. Artificial intelligence (AI)–enabled electrocardiography (ECG) may enhance perioperative risk assessment by identifying surgical candidates at very low-risk for adverse events.

Objective: This study aimed to evaluate whether AI-ECG–based risk stratification could help decrease low-yield preoperative cardiovascular testing and reduce associated costs, without an observed increase in postoperative adverse outcomes, in candidates for noncardiac surgery.

Methods: We retrospectively analyzed 41,218 patients (46,135 ECG-surgery pairs) undergoing noncardiac surgery at Seoul National University Bundang Hospital (2020-2021). An AI-ECG algorithm generated eight probability scores for cardiac conditions, classifying cases as low- or high-risk. Based on the performance and results of preoperative cardiovascular testing (transthoracic echocardiography, coronary computed tomography angiography, single-photon emission computed tomography, or coronary angiography), cases were classified as no advanced cardiovascular imaging, negative-test, or positive-test. The primary end point was a 30-day composite of all-cause mortality and unplanned percutaneous coronary intervention.

Results: AI-ECG classified 92.3% (42,599/46,135) of cases as low-risk, with a composite outcome rate of 0.2% (79/42,599) vs 2.9% (101/3536) in high-risk cases. Preoperative cardiovascular testing was performed in 11.8% (5458/46,135) of cases, with only 16.3% (892/5458) yielding positive findings. In AI-ECG low-risk cases, event rates were uniformly low (0.2%-0.4%) irrespective of whether advanced cardiovascular testing was performed, whereas in high-risk cases, rates were consistently high (2.6%-3.4%). The incidence of the composite outcome was consistently higher in AI-ECG–graded high-risk cases across all European Society of Cardiology surgical risk and Revised Cardiac Risk Index strata.

Conclusions: In this retrospective cohort, a multitask AI-ECG identified surgical candidates at low-risk for postoperative complications, for whom advanced cardiovascular testing demonstrated low diagnostic yield. Integrating AI-ECG with conventional risk tools may offer an exploratory strategy to optimize resource use and minimize redundant testing. Prospective studies are needed to confirm the clinical and economic benefits of AI-ECG as a screening tool.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e90099) doi:10.2196/90099

KEYWORDS

artificial intelligence; electrocardiography; perioperative care; mortality; prediction; preoperative cardiovascular examination; cost

Introduction

More than 200 million noncardiac surgeries are conducted worldwide each year [1], with numbers steadily increasing [2]. Preoperative cardiovascular testing imposes a major health care burden, with estimated costs of US $18 billion annually in the United States [3]. The substantially higher morbidity and mortality risks of postoperative cardiovascular complications underscore the importance of accurate preoperative risk stratification [4]. Although existing risk-estimation tools provide general estimates of perioperative cardiovascular risk [5,6], they do not specify whether additional testing is warranted and do not eliminate the risk of adverse outcomes. Despite current guidelines outlining which patients may benefit from advanced cardiovascular testing [5,6], in clinical practice, the demand for such testing often exceeds the recommended indications [7]. As a result, preoperative cardiovascular tests are frequently overused and thereby hinder efficient health care resource usage.

Electrocardiography (ECG) is a well-established low-cost preoperative assessment tool, but has limited utility in noncardiac surgery [3,8,9]. Not surprisingly, the quality of ECG-derived information has been highly dependent on the physician's expertise and the patient's characteristics [10,11]. Thus, recent guidelines have recommended ECG use in only selected patients with suspected cardiovascular diseases or risk factors who are undergoing intermediate- or high-risk noncardiac surgery [5,6]. However, recent advances in artificial intelligence (AI) have expanded the diagnostic and prognostic applications of ECG, enabling the detection of coronary artery disease, valvular heart disease (VHD), and myocardial dysfunction [12,13]. With its low-cost and broad accessibility, AI-enhanced ECG is a promising tool for preoperative risk stratification.

Building on a preexisting AI-ECG model that incorporated eight quantitative scores for electrocardiography (QCGs) reflecting diverse cardiac conditions [14-19], we aimed to evaluate whether AI-ECG–based risk stratification could help decrease low-yield preoperative cardiovascular testing and reduce associated costs, without an observed increase in postoperative adverse outcomes in cases of noncardiac surgery. We hypothesized that the integration of AI-ECG into the workflow of preoperative risk assessment could help identify a subgroup of cases with an extremely low-risk of postoperative adverse outcomes, thereby requiring no additional testing.

Methods

Study Population

This single-center, retrospective cohort study included cases of noncardiac surgery at Seoul National University Bundang Hospital between 2020 and 2021, and underwent an ECG within the 30 days preceding the surgery. The exclusion criteria included the following: (1) no preoperative ECG within 30 days preceding the surgery, (2) elective cardiac surgery or extracorporeal membrane oxygenation-related surgery, (3) the use of simple pain-control procedures, (4) surgeries performed under local anesthesia, (5) surgery for a deceased organ donor, (6) rigid bronchoscopy, (7) dental surgery, and (8) multiple surgeries during the index admission. A total of 46,135 ECG–surgery pairs from 41,218 patients were analyzed (Figure S1 in Multimedia Appendix 1). As this was a retrospective study, the sample size was determined by available data meeting the inclusion criteria during the specified study period.

Data Collection

All ECGs were recorded under standardized settings (25 mm/s speed and 10 mm/mV gain) and processed to extract the QCG scores from the JPEG file. The preoperative ECG that was recorded within 30 days before the surgery and closest to the surgery was paired with one noncardiac surgery. For 26 patients who underwent coronary revascularization following preoperative cardiovascular evaluation, prerevascularization ECGs were specifically selected as the index ECGs. Consequently, these cases were exempted from the standard 30-day ECG inclusion criterion. This exception was made because using postintervention ECGs for the primary cohort assignment could introduce treatment bias and falsely obscure the utility of advanced cardiovascular testing. The treating clinicians were unaware of the retrospective AI-ECG risk scores, as the tool was not yet available at the time of surgery. Clinical data, including surgery-related records, test results, and outcomes, were obtained from electronic health care records. This observational study conformed to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guideline [20].

Ethical Considerations

This study's protocol was approved by the Institutional Review Board of Seoul National University Bundang Hospital (IRB No. B-2409-927-106), which waived the requirement for written informed consent owing to the retrospective nature of this study. All data were deidentified, and no financial compensation was provided to study participants. All clinical investigations were conducted in accordance with the principles of the Declaration of Helsinki.

AI Algorithm and Risk-Stratification Strategy for AI-ECG

The QCG system is a deep-learning–based AI analyzer that uses 12-lead ECG images (in JPEG, PNG, or PDF formats) as input data and yields the probability score of specific downstream tasks as numerical values (range 0-100). Following pretraining using self-supervised learning schemes, these models were fine-tuned using multitask learning schemes with a modified convolutional neural network using residual connections, squeeze excitation modules, and a nonlocal block. The input data comprised 47,194 annotated ECG images from Seoul

XSL•FO
RenderX

National University Bundang Hospital that were recorded between 2017 and 2019. ECGs obtained during this study's period (2020-2021) were not used for model training, fine-tuning, or calibration and were reserved exclusively for outcome analysis. The tasks included rhythm classification and the construction of ten QCG scores related to various medical emergencies: (1) critical illness (critical score), (2) acute coronary syndrome (ACS), (3) ST-segment elevation myocardial infarction, (4) myocardial injury, (5) pulmonary edema, (6) pericardial effusion, (7) left ventricular dysfunction, (8) right ventricular dysfunction, (9) pulmonary hypertension, and (10) hyperkalemia. ECG buddy—an app that provides these 10 QCG scores—was approved by the Korean Ministry of Food and Drug Safety (January 2024). Detailed explanations of the AI algorithms and validation studies have been previously published [15-18]. For this dataset, the QCG scores were generated using version 1.0.6 of the app. Thresholds and clinical interpretations of each QCG score are provided in Table S1 in Multimedia Appendix 1.

Among the ten probability scores, we selected eight QCG scores that are directly related to the cardiac conditions of interest, except for the critical and hyperkalemia scores. All scores had their own optimal thresholds to maximize diagnostic performance, which were determined at the model-development stage. The participants were assigned to the AI-ECG low-risk group when all 8 QCG scores were lower than their individual optimal thresholds; otherwise, they were assigned to the AI-ECG high-risk group.

Preoperative Advanced Cardiovascular Imaging

Preoperative advanced cardiovascular imaging included echocardiography, coronary computed tomography angiography (CCTA), single-photon emission computed tomography (SPECT) performed within 90 days before surgery, and coronary angiography (CAG) performed within 180 days before surgery. The decision for further cardiovascular testing and the choice of diagnostic modality were made by the attending surgeon or consulting physician while considering each patient's functional capacity, individual risk, and surgical risk. Even if tests were included in routine clinical practice irrespective of the surgery, all cardiovascular tests performed within the designated time window were considered preoperative cardiovascular testing.

Positive findings of preoperative cardiovascular tests were selected based on widely accepted actionability, whereby positive results suggested the need for further preoperative evaluation. For echocardiography, positive findings were defined as the presence of any of the following: moderate or severe VHD (aortic stenosis or regurgitation, mitral stenosis or regurgitation, or tricuspid regurgitation); left ventricular ejection fraction ≤50%; or regional wall motion abnormality. For CCTA and CAG, ≥50% stenosis in any of the three major coronary arteries was considered a positive result; angiographic stenosis grades were used for CAG. For SPECT, a ≥10% reversible or irreversible perfusion defect was defined as a positive result. The costs of CCTA, CAG, and SPECT were obtained from a published study conducted in the Republic of Korea in 2018 that reported values in US dollars [21]. For echocardiography, the cost was based on the median value of publicly posted prices in Seoul, Republic of Korea, and was converted from Korean won to US dollars using an approximate 2021 exchange rate [22]. The exact per-test unit costs applied in this study were US $267 for echocardiography, US $232 for CCTA, US $485 for SPECT, and US $511 for CAG.

Based on the preoperative cardiovascular testing status, all the surgery-ECG pairs were categorized into three groups: no advanced cardiovascular imaging, cases without any preoperative cardiovascular test; negative test, cases that underwent testing without any positive findings; and positive test, cases with at least one positive finding. In this context, a negative test indicates the absence of predefined positive findings in a specific examination rather than the confirmation of overall clinical normality.

Study Group Definition

To compare the adequacy of the preoperative cardiovascular testing and AI-ECG, this study's population was further divided into six groups: (1) group 1, low-risk AI-ECG and no advanced cardiovascular imaging; (2) group 2, low-risk AI-ECG and negative test; (3) group 3, low-risk AI-ECG and positive test; (4) group 4, high-risk AI-ECG and no advanced cardiovascular imaging; (5) group 5, high-risk AI-ECG and negative test; and (6) group 6, high-risk AI-ECG and positive test (Figure 1).

XSL•FO
RenderX

Figure 1. Study design. Among the preoperative CV tests, we included EchoCG, SPECT, and CCTA that were performed within 90 days, and CAG performed within 180 days before the noncardiac surgery. Among the ECG scans performed within 30 days preoperatively, the ECG obtained nearest to the surgery was selected to estimate the QCG scores using an AI-ECG. Next, the preoperative risk was classified using eight QCG scores related to cardiac conditions. Six groups were created based on the preoperative CV test and AI-ECG risk classification. The blue square denotes the surgeries in which CV tests were performed (groups 2, 3, 5, and 6; blue square), and the green square implies the surgeries in which testing would have been required (groups 4, 5, and 6; green square) assuming that the decision to perform preoperative CV testing strictly followed the AI-ECG risk stratification (ie, tests performed only in high-risk cases). AI-ECG: artificial intelligence–enabled electrocardiography; CAG: coronary angiography; CCTA: coronary computed tomography angiography; CV: cardiovascular; ECG: electrocardiography; EchoCG: echocardiography; QCG: quantitative score for electrocardiography; SPECT: single-photon emission computed tomography.

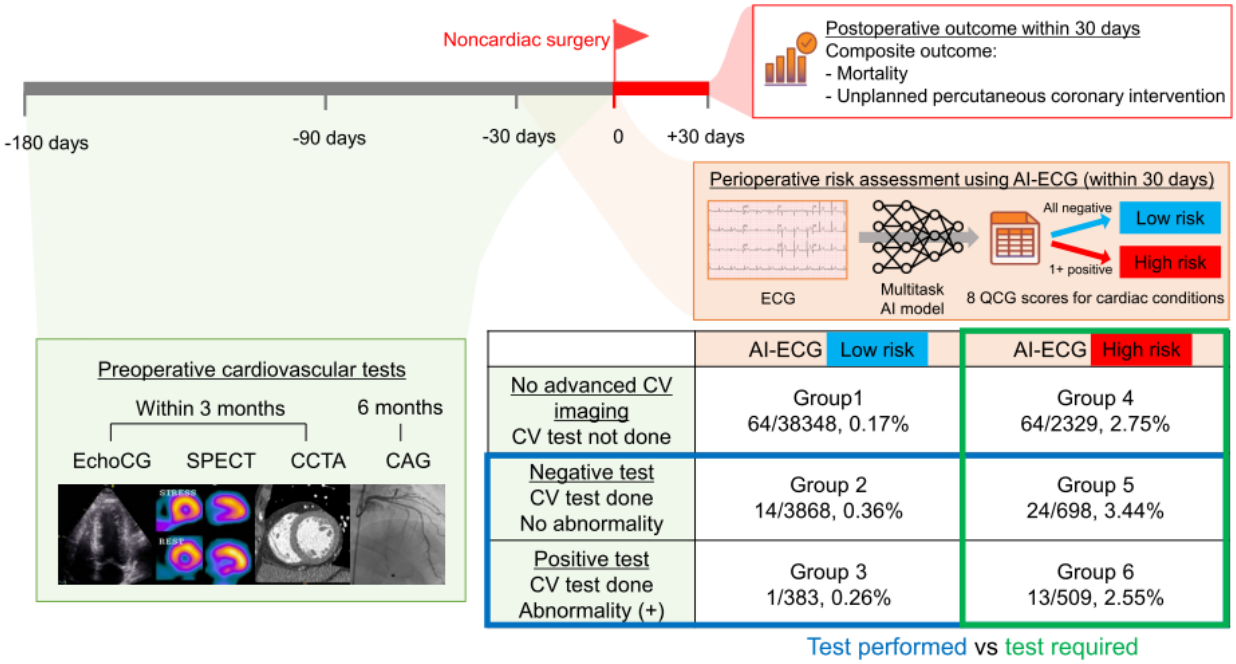

Postoperative Outcome Definition

The primary measure that was used to assess the predictive performance of preoperative cardiovascular testing and AI-ECG was a composite postoperative outcome occurring within 30 days after the index surgery that included all-cause mortality and unplanned percutaneous coronary intervention (PCI). Mortality was restricted to deaths during hospitalization, including deaths within 30 days of readmission, whereas deaths that occurred outside the hospital were not included in the analysis. Given the substantial impact of cardiovascular comorbidities on postoperative critical care requirements, an exploratory composite outcome that incorporates prolonged mechanical ventilation for ≥3 days into the composite postoperative outcome was also analyzed. The overall study scheme is illustrated in Figure 1.

Statistical Analysis

Continuous variables were summarized as means with SDs or medians with IQRs, whereas categorical variables were summarized as frequencies with percentages. Between-group comparisons were performed using the independent 2-sample *t* test or the Mann-Whitney *U* test for continuous variables, and the chi-square or Fisher exact test for categorical variables, as appropriate.

For intergroup comparison, the incidence of the composite outcome was compared between the AI-ECG high- and low-risk groups within the Revised Cardiac Risk Index (RCRI) and European Society of Cardiology (ESC) surgical risk strata. The incidence rates with 95% CIs, relative risks (RRs), and risk differences were derived from contingency 2×2 tables comparing AI-ECG risk groups and the presence or absence of the composite outcome. Within each AI-ECG stratum, the outcome incidence across the three conventional testing groups (no advanced cardiovascular imaging, negative test, and positive test) was compared using the chi-square or Fisher exact tests; pairwise Fisher exact tests and Bonferroni correction were applied when appropriate. For comparisons across multiple strata, overall differences were evaluated using 1-way ANOVA or the Kruskal-Wallis test for continuous variables, and the chi-square or Fisher exact test for categorical variables. To assess the incremental prognostic value of incorporating the AI-ECG into RCRI, we calculated the change in the area under the receiver operating characteristic curve, the continuous net reclassification improvement (NRI), and the integrated discrimination improvement (IDI). We also constructed a generalized estimating equation (GEE) model setting the patient identifier as the cluster variable to address potential within-patient correlation.

To assess the robustness of our findings, we conducted prespecified sensitivity analyses across surgical subgroups: major surgery under general anesthesia, elective surgery, and emergency surgery. Furthermore, to address potential within-patient correlation arising from multiple surgeries per patient, we performed a patient-level sensitivity analysis

XSL•FO
RenderX

restricted to each patient's first index surgery. In addition, to further address potential confounding by indication, we performed propensity score matching (PSM). Surgeries with and without advanced cardiovascular testing were matched in a 1:1 ratio using nearest-neighbor matching (caliper=0.2). The matching model included critical confounders: emergency surgery status, age, sex, type of anesthesia, ESC surgical risk category, RCRI score, American Society of Anesthesiologists class, and comorbidities.

Following STROBE guidelines, missing values were identified only for smoking status (7795/46,135, 16.9%) and baseline serum creatinine (n=14). As these variables were solely used for baseline descriptive statistics, no missing data imputation was performed. Missing smoking status was presented as a separate "unknown" category. Statistical significance was defined using a 2-sided *P* value <.05. All analyses were performed using R (version 4.4.1; R Foundation).

Results

Baseline Characteristics

Baseline clinical and surgery-specific characteristics of this study's population are summarized in Table 1. The mean age was 56.6 (SD 16.1) years, and males accounted for 44.6% (20,561/46,135) of the cases. Emergency procedures comprised 6.1% (2806/46,135) of all surgeries, and 72.4% (33,386/46,135) of them were performed under general anesthesia. According to the ESC surgical risk categorization and the RCRI stratification, high-risk surgery was identified in 10.7% (4918/46,135) and 4.2% (1929/46,135) of cases, respectively.

Overall, 3536 (7.7%) ECG–surgery pairs had at least one QCG value that exceeded the predefined cutoffs and were thus classified as high-risk by AI-ECG. Compared to the low-risk group, high-risk cases involved older individuals, had a higher proportion of males, presented with a greater burden of comorbidities, underwent a higher proportion of emergency procedures, and were more frequently categorized as undergoing high-risk surgeries, according to both the ESC and RCRI (Table 1) risk categorizations.

The distribution of the QCG scores across the six study groups is shown in Figure S2 in Multimedia Appendix 1, with significantly higher scores observed in the AI-ECG–stratified high-risk group. Across the three strata of preoperative cardiovascular test results, cases in the positive-test group involved the oldest individuals, had the highest comorbidity burden, underwent high-risk surgery, and had the highest QCG scores. Cases in the negative-test group showed intermediate characteristics, whereas those in the no advanced cardiovascular imaging group involved the youngest individuals, with the fewest comorbidities and the lowest QCG scores (Table S2 in Multimedia Appendix 1).

XSL•FO
RenderX

Table 1. Baseline characteristics and comparison between AI-ECG[a] risk groups.

Variable	Overall (N=46,135)	AI-ECG low-risk (n=42,599)	AI-ECG high-risk (n=3536)	*P* value
Demographics and comorbidities				
Age (years), mean (SD)	56.6 (16.1)	55.7 (16.0)	67.4 (14.0)	<.001
Male sex (%)	20,561 (44.6)	18,283 (42.9)	2278 (64.4)	<.001
DM[b] (%)	5904 (12.8)	4901 (11.5)	1003 (28.4)	<.001
DM treated with insulin (%)	632 (1.4)	444 (1.0)	188 (5.3)	<.001
Hypertension (%)	10,759 (23.3)	9297 (21.8)	1462 (41.3)	<.001
Previous heart failure (%)	410 (0.9)	139 (0.3)	271 (7.7)	<.001
Previous ischemic heart disease (%)	1772 (3.8)	1115 (2.6)	657 (18.6)	<.001
Previous cerebrovascular accident (%)	1525 (3.3)	1213 (2.8)	312 (8.8)	<.001
Smoking (%)				<.001
Nonsmoker	28,436 (61.6)	26,343 (61.8)	2093 (59.2)	
Smoker	9904 (21.5)	8769 (20.6)	1135 (32.1)	
Unknown	7795 (16.9)	7487 (17.6)	308 (8.7)	
Creatinine, median (IQR)	0.74 (0.62-0.91)	0.73 (0.61-0.89)	0.90 (0.71-1.26)	<.001
Creatinine ≥2.0 mg/dl (%)	1414 (3.1)	850 (2.0)	564 (16.0)	<.001
Surgery-related				
Emergency surgery (%)	2806 (6.1)	2337 (5.5)	469 (13.3)	<.001
Anesthesia (%)				<.001
General anesthesia	33,386 (72.4)	30,946 (72.6)	2440 (69.0)	
Spinal or epidural anesthesia	4272 (9.3)	3964 (9.3)	308 (8.7)	
Monitored anesthesia care	8477 (18.4)	7689 (18.0)	788 (22.3)	
ESC[c] surgical risk category (%)				<.001
Low	24,825 (53.8)	23,280 (54.6)	1545 (43.7)	
Intermediate	16,392 (35.5)	15,161 (35.6)	1231 (34.8)	
High	4918 (10.7)	4158 (9.8)	760 (21.5)	
RCRI[d] (%)				<.001
0	28,738 (62.3)	27,427 (64.4)	1311 (37.1)	
1	15,468 (33.5)	14,009 (32.9)	1459 (41.3)	
2	1562 (3.4)	1011 (2.4)	551 (15.6)	
3	308 (0.7)	137 (0.3)	171 (4.8)	
4	50 (0.1)	15 (0.0)	35 (1.0)	
5	9 (0.0)	0 (0.0)	9 (0.3)	
ASA[e] classification (%)				<.001
1	13,814 (29.9)	13,623 (32.0)	191 (5.4)	
2	24,130 (52.3)	22,955 (53.9)	1175 (33.2)	
3	7332 (15.9)	5591 (13.1)	1741 (49.2)	
4	765 (1.7)	377 (0.9)	388 (11.0)	
5	94 (0.2)	53 (0.1)	41 (1.2)	
QCG[f] scores, median (IQR)				
ACS[g]	1.37 (0.30-4.26)	1.20 (0.27-3.62)	12.68 (2.84-25.99)	<.001

XSL•FO
RenderX

Variable	Overall (N=46,135)	AI-ECG low-risk (n=42,599)	AI-ECG high-risk (n=3536)	*P* value
STEMI[h]	0.01 (0.00-0.07)	0.01 (0.00-0.05)	0.41 (0.06-1.95)	<.001
Myocardial injury	1.71 (0.50-4.58)	1.48 (0.44-3.79)	16.80 (6.49-26.29)	<.001
Pulmonary edema	0.63 (0.20-2.18)	0.55 (0.18-1.72)	15.22 (2.81-26.98)	<.001
Pericardial effusion	0.03 (0.00-0.23)	0.02 (0.00-0.16)	1.34 (0.10-7.40)	<.001
Left ventricular dysfunction	0.08 (0.02-0.34)	0.07 (0.02-0.24)	4.16 (0.70-19.20)	<.001
Right ventricular dysfunction	0.13 (0.04-0.57)	0.11 (0.03-0.41)	5.02 (0.75-17.61)	<.001
Pulmonary hypertension	0.15 (0.03-0.76)	0.12 (0.03-0.56)	5.60 (0.64-17.59)	<.001
Preoperative cardiovascular test				
Any preoperative CV[i] test performed (%)	5458 (11.8)	4251 (10.0)	1207 (34.1)	<.001
Positive any preoperative CV test (%)	892 (16.3)	383 (9.0)	509 (42.2)	<.001
Echocardiography performed (%)	4808 (10.4)	3842 (9.0)	966 (27.3)	<.001
Positive echocardiography results (%)	574 (11.9)	243 (6.3)	331 (34.3)	<.001
Moderate or severe valvular heart disease (%)	235 (4.9)	114 (3.0)	121 (12.6)	<.001
Regional wall motion abnormality (%)	374 (7.8)	130 (3.4)	244 (25.5)	<.001
LVEF[j] <50% (%)	225 (4.7)	54 (1.4)	171 (17.8)	<.001
LVEF (%), median (IQR)	62.7 (58.9-66.3)	63.2 (59.7-66.7)	59.9 (54.1-64.8)	<.001
CAG[k] performed (%)	470 (1.0)	183 (0.4)	287 (8.1)	<.001
Positive CAG results (%)	273 (58.1)	76 (41.5)	197 (68.6)	<.001
Coronary CT[l] angiography performed (%)	579 (1.3)	411 (1.0)	168 (4.8)	<.001
Positive coronary CT angiography results (%)	165 (29.0)	99 (24.4)	66 (40.5)	<.001
SPECT[m] performed (%)	245 (0.5)	150 (0.4)	95 (2.7)	<.001
Positive SPECT results (%)	7 (2.9)	3 (2.0)	4 (4.2)	.54

[a]AI-ECG: artificial intelligence–enabled electrocardiography.
[b]DM: diabetes mellitus.
[c]ESC: European Society of Cardiology.
[d]RCRI: Revised Cardiac Risk Index.
[e]ASA: American Society of Anesthesiologists.
[f]QCG: quantitative score for electrocardiography.
[g]ACS: acute coronary syndrome.
[h]STEMI: ST-segment elevation myocardial infarction.
[i]CV: cardiovascular.
[j]LVEF: left ventricular ejection fraction.
[k]CAG: coronary angiography.
[l]CT: computed tomography.
[m]SPECT: single-photon emission computed tomography.

Postoperative Outcomes

In total, 180 (0.4%) cases experienced the postoperative composite outcome. Overall, the event rates were 2.9% and 0.2% in the high-risk and low-risk AI-ECG groups, respectively. When stratified by surgical urgency, composite outcomes occurred in 2.8% (78/2806) and 0.2% (102/43,329) of emergency and elective surgeries, respectively. Among the emergency surgery cases, the risks of the composite end point were 8.5% and 1.6% in the AI-ECG–stratified high-risk and low-risk groups, respectively.

The incidence of the composite outcome was consistently higher in AI-ECG–graded high-risk cases across all ESC surgical risk and RCRI strata. In the ESC surgical risk- and RCRI-categorized low-risk procedures, the RRs of the composite outcome occurrence were higher in the low-risk AI-ECG group than in the high-risk AI-ECG group (RR 58.1, 95% CI 25.4-133.3) and

XSL•FO
RenderX

14.3 (10.1–20.1) for ESC surgical risk and RCRI, respectively, while the procedures classified as high-risk by the established risk stratification tools showed larger risk differences between the low-risk and high-risk AI-ECG groups (Table 2). Figure S3 in Multimedia Appendix 1 demonstrated that AI-ECG improved patient reclassification compared to conventional risk stratification systems. Incorporating the AI-ECG into a baseline model including age and RCRI provided significant incremental prognostic value, as demonstrated by improved area under the receiver operating characteristic curve, NRI, and IDI (Figure S4 in Multimedia Appendix 1). Notably, while the overall classification improvement was significant (NRI 0.675, *P*<.001), it was primarily driven by the accurate down-classification of nonevents (NRI for nonevents 0.530, *P*<.001), whereas the reclassification for events did not reach statistical significance (NRI for events 0.144, *P*=.05). Furthermore, in the additional analysis accounting for within-patient correlation using GEE, the high-risk designation by AI-ECG remained a significant predictor of the primary composite outcome after adjustment for age and RCRI (adjusted GEE odds ratio 8.26, 95% CI 5.93-11.50, *P*<.001).

Table 2. The incidence, relative risk, and relative difference of the composite outcome across ESC[a] surgical risk category and RCRI[b] strata by AI-ECG[c] risk stratification.

Risk group	Incidence of composite outcomes			
	AI-ECG–stratified, low-risk (%)	AI-ECG–stratified, high-risk (%)	RR[d] (95% CI)	RD[e] %*P* (95% CI)
Surgical emergency				
Elective	0.10 (0.07-0.14)	1.99 (1.52-2.55)	19.53 (13.17-28.97)	1.89 (1.39-2.38)
Emergency	1.63 (1.15-2.23)	8.53 (6.16-11.43)	5.25 (3.40-8.09)	6.90 (4.32-9.48)
ESC surgical risk				
Low	0.03 (0.01-0.06)	1.75 (1.15-2.53)	58.12 (25.35-133.25)	1.72 (1.06-2.37)
Intermediate	0.31 (0.23-0.41)	3.01 (2.12-4.12)	9.70 (6.33-14.86)	2.70 (1.74-3.65)
High	0.60 (0.39-0.89)	4.87 (3.45-6.65)	8.10 (4.90-13.37)	4.27 (2.72-5.82)
RCRI				
0-1	0.16 (0.12-0.20)	2.27 (1.75-2.90)	14.28 (10.13-20.12)	2.12 (1.56-2.67)
2+	1.12 (0.60-1.90)	4.96 (3.53-6.75)	4.44 (2.38-8.28)	3.84 (2.19-5.50)
Total	0.19 (0.15-0.23)	2.86 (2.33-3.46)	15.40 (11.50-20.63)	2.67 (2.12-3.22)

[a]ESC: European Society of Cardiology.

[b]RCRI: Revised Cardiac Risk Index.

[c]AI-ECG: artificial intelligence–enabled electrocardiography.

[d]RR: relative risk.

[e]RD: relative difference.

Evaluated as a stand-alone tool, the AI-ECG yielded a negative predictive value of 99.8% and a positive predictive value (PPV) of 2.9%. Considering the extremely low baseline event rate (0.4%), this PPV reflects a 7.4-fold increased RR, albeit representing a modest absolute predictive value. Furthermore, these risk stratification capabilities were consistently maintained when evaluating the exploratory composite outcome that included prolonged mechanical ventilation (Table S3 in Multimedia Appendix 1).

Comparisons of predefined outcomes across the six study groups, defined by AI-ECG and preoperative cardiovascular test results, are summarized in Figure 2. The composite outcomes differed significantly between the low-risk and high-risk groups on AI-ECG (*P*<.001 for each comparison within the no advanced cardiovascular imaging, negative-test, and positive-test groups). The low-risk AI-ECG group showed consistently low composite outcome rates compared to the high-risk AI-ECG group, irrespective of advanced cardiovascular imaging. Within the AI-ECG low-risk group, although a nominal statistical significance was observed among the subgroups (0.2%, 0.4%, and 0.3% for the no advanced cardiovascular imaging, negative-test, and positive-test groups, respectively; *P*=.025), the absolute event rates remained uniformly low. Within the high-risk AI-ECG group, there was no significant intragroup difference in the outcome incidences (2.8%, 3.4%, and 2.6%, respectively; *P*=.572, Table 3).

XSL•FO
RenderX

Figure 2. Predefined outcome rates across the six study groups assigned by the AI-ECG (low-risk green; high-risk red) and the preoperative CV testing. Death represents all-cause mortality; cardiac death was not adjudicated separately in this study. AI-ECG: artificial intelligence–enabled electrocardiography; CV: cardiovascular; PCI: percutaneous coronary intervention.

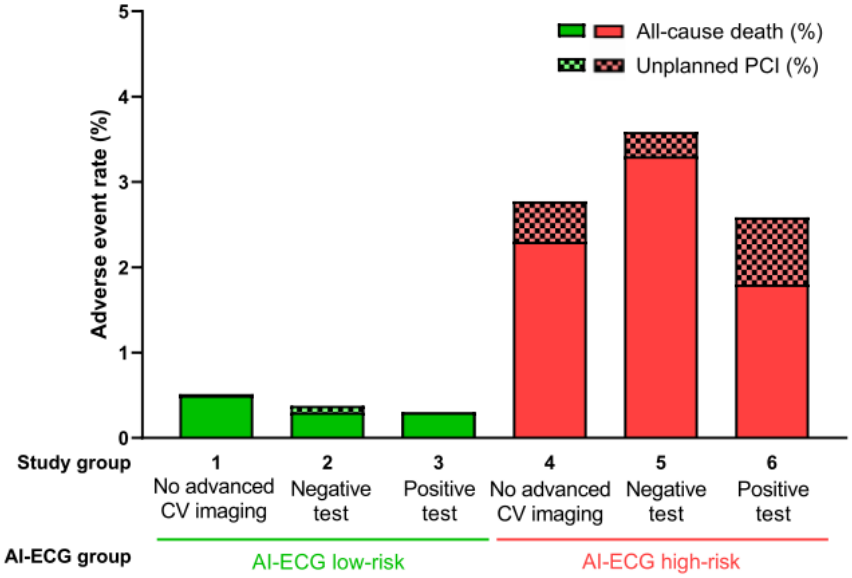

Table 3. Predefined outcome rates across the six study groups assigned by the artificial intelligence–enabled electrocardiography and the preoperative cardiovascular testing.

Study group	Group 1 (n=38,348)	Group 2 (n=3868)	Group 3 (n=383)	Group 4 (n=2329)	Group 5 (n=698)	Group 6 (n=509)	*P* value
Composite outcome, n (%)	64 (0.2)	14 (0.4)	1 (0.3)	64 (2.8)	24 (3.4)	13 (2.6)	<.001
All-cause death, n (%)	58 (0.2)	11 (0.3)	1 (0.3)	54 (2.3)	23 (3.3)	9 (1.8)	<.001
Unplanned percutaneous coronary intervention, n (%)	6 (0.02)	3 (0.1)	0 (0.0)	11 (0.5)	2 (0.3)	4 (0.8)	<.001

Preoperative Cardiovascular Test and Reclassification Using AI-ECG

Preoperative cardiovascular testing was performed in 11.8% (5458/46,135) of the cases, with only 16.3% (892/46,135) showing positive findings. Of the cases that involved preoperative cardiac testing, 88.1% (4808/5458) included echocardiography, making it the most commonly used modality. Among the cases in which preoperative cardiovascular testing was performed (groups 2, 3, 5, and 6), the negative-test groups (groups 2 and 5) were less likely to involve additional cardiovascular tests beyond echocardiography than the positive-test groups (groups 3 and 6). The types and results of preoperative cardiovascular tests are presented in Table S4 in Multimedia Appendix 1.

When stratified by preoperative cardiovascular testing, all outcomes except for prolonged mechanical ventilation in the positive-test group showed significant differences according to the AI-ECG risk level. As illustrated in Figure 3, a substantial proportion of cases in the no advanced cardiovascular imaging group (2329/40,677, 5.7%) were classified as high-risk by AI-ECG, whereas most cases in the negative-test group (3868/4566, 84.7%) were classified as low-risk.

XSL•FO
RenderX

Figure 3. The proportion of AI-ECG risk according to the preoperative CV test. A substantial proportion of ECG-surgery pairs in the no advanced CV imaging group (2329/40,677, 5.7%) were classified as high-risk by AI-ECG (red), whereas the majority of cases in the negative-test group (3868/4566, 84.7%) were classified as low-risk by AI-ECG (light blue). AI-ECG: artificial intelligence–enabled electrocardiography; CV: cardiovascular; ECG: electrocardiography.

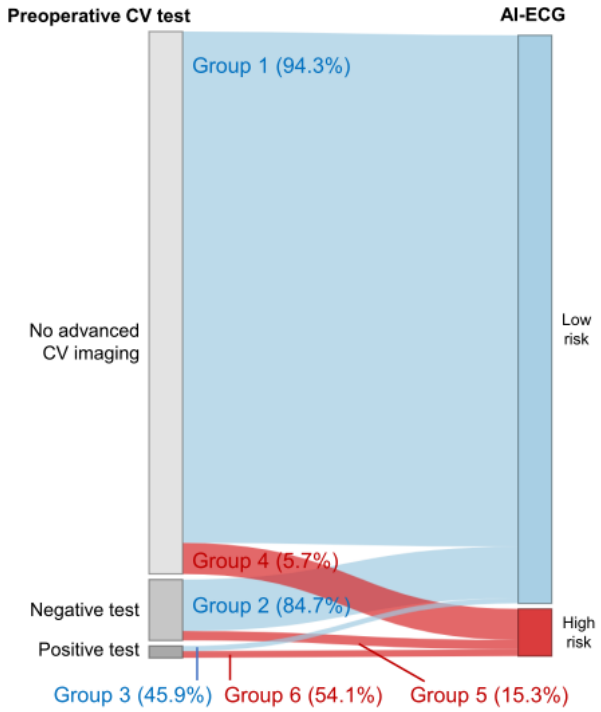

Assessment for Individual Study Groups

The characteristics of each study group are summarized in Table 4. Groups 1 and 6 demonstrated appropriate application of the AI-ECG–derived risk classification: group 1 included low-risk cases without further testing, and group 6 included high-risk cases in which testing yielded abnormal findings. Accordingly, the subsequent evaluation focuses on groups 2-5.

Group 4 comprised cases classified as high-risk by AI-ECG that proceeded to surgery without advanced cardiovascular imaging. This group had the highest frequency of emergency procedures (n=389); otherwise, most cases were categorized as low-risk surgery by the RCRI (n=1991) or ESC surgical risk (n=1147; Figure S5 in Multimedia Appendix 1). Among the remaining cases (n=137) after excluding these two reasons, 105 procedures involved preoperative cardiovascular testing at another hospital. Thus, the number of "true" missing cases by clinicians was 32, in which no further diagnostic tests or cardiology consultations were undertaken despite high-risk AI-ECG findings. Among these 32 cases, two deaths occurred (mortality rate: 6.3%). Although preoperative advanced cardiovascular testing was not performed in group 4, downstream costs were estimated by applying the mean frequency of individual tests observed in cases involving at least one of the advanced cardiovascular tests (groups 2, 3, 5, and 6; Figure 4).

Group 5 included cases classified as high-risk by AI-ECG, but with negative preoperative cardiovascular tests. This group consisted of a heterogeneous set of conditions: (1) abnormalities detected by preoperative cardiovascular testing that were not classified as positive results (eg, atrial fibrillation, myocardial disease, and severe left ventricular hypertrophy), (2) a history of open-heart surgery but with preserved cardiac function, or (3) stable coronary stenosis or prior PCI without interval change. Group 5 had the highest composite end point rate (3.4%).

Group 2 included ECG–surgery pairs classified as low-risk by AI-ECG that involved preoperative advanced cardiovascular imaging with negative results. The most frequently performed test modality in this group was echocardiography (3553/3868, 91.9%), which resulted in the highest cost of preoperative cardiovascular tests among the six study groups (Figure 4). The proportion of the cost spent in group 2 was 62.8% (USD $1,115,966,000/$1,777,059,000) of the total spending on preoperative cardiovascular evaluations.

Group 3 represents cases with abnormal findings on advanced cardiovascular testing despite a low-risk classification by AI-ECG. In this group, composite outcomes occurred in 1 patient (0.3%, death).

XSL•FO
RenderX

Table 4. Characteristics of individual study groups.

	Group 1 (n=38,348)	Group 2 (n=3868)	Group 3 (n=383)	Group 4 (n=2329)	Group 5 (n=698)	Group 6 (n=509)	*P* value
Demographics and comorbidities							
Age (years), mean (SD)	54.2 (15.7)	68.8 (11.7)	72.0 (10.4)	65.3 (14.3)	71.5 (12.9)	71.5 (11.8)	<.001
Male sex (%)	16,207 (42.3)	1820 (47.1)	256 (66.8)	1538 (66.0)	399 (57.2)	341 (67.0)	<.001
DM[a] (%)	4001 (10.4)	796 (20.6)	104 (27.2)	629 (27.0)	172 (24.6)	202 (39.7)	<.001
DM treated with insulin (%)	315 (0.8)	112 (2.9)	17 (4.4)	96 (4.1)	47 (6.7)	45 (8.8)	<.001
Hypertension (%)	7557 (19.7)	1553 (40.1)	187 (48.8)	935 (40.1)	276 (39.5)	251 (49.3)	<.001
Previous heart failure (%)	44 (0.1)	32 (0.8)	63 (16.4)	66 (2.8)	17 (2.4)	188 (36.9)	<.001
Previous ischemic heart disease (%)	702 (1.8)	263 (6.8)	150 (39.2)	288 (12.4)	82 (11.7)	287 (56.4)	<.001
Previous cerebrovascular accident (%)	965 (2.5)	203 (5.2)	45 (11.7)	199 (8.5)	57 (8.2)	56 (11.0)	<.001
Smoking (%)							<.001
Nonsmoker	23,402 (61.0)	2735 (70.7)	206 (53.8)	1328 (57.0)	482 (69.1)	283 (55.6)	
Smoker	7577 (19.8)	1034 (26.7)	158 (41.3)	722 (31.0)	200 (28.7)	213 (41.8)	
Unknown	7369 (19.2)	99 (2.6)	19 (5.0)	279 (12.0)	16 (2.3)	13 (2.6)	
Creatinine, median (IQR)	0.72 (0.61-0.88)	0.78 (0.64-0.95)	0.88 (0.73-1.10)	0.89 (0.70-1.20)	0.88 (0.68-1.26)	1.01 (0.79-1.65)	<.001
Creatinine ≥2.0 mg/dl (%)	670 (1.7)	147 (3.8)	33 (8.6)	358 (15.4)	96 (13.8)	110 (21.6)	<.001
Surgery-related							
Emergency surgery (%)	2230 (5.8)	88 (2.3)	19 (5.0)	389 (16.7)	44 (6.3)	36 (7.1)	<.001
Anesthesia (%)							<.001
General anesthesia	27,687 (72.2)	2965 (76.7)	294 (76.8)	1576 (67.7)	525 (75.2)	339 (66.6)	
Spinal/epidural anesthesia	3224 (8.4)	698 (18.0)	42 (11.0)	163 (7.0)	94 (13.5)	51 (10.0)	
Monitored anesthesia care	7437 (19.4)	205 (5.3)	47 (12.3)	590 (25.3)	79 (11.3)	119 (23.4)	
ESC[b] surgical risk category (%)							<.001
Low	22,098 (57.6)	1080 (27.9)	102 (26.6)	1147 (49.2)	206 (29.5)	192 (37.7)	
Intermediate	12,821 (33.4)	2191 (56.6)	149 (38.9)	740 (31.8)	322 (46.1)	169 (33.2)	
High	3429 (8.9)	597 (15.4)	132 (34.5)	442 (19.0)	170 (24.4)	148 (29.1)	
RCRI[c] (%)							<.001
0	25,503 (66.5)	1855 (48.0)	69 (18.0)	1009 (43.3)	256 (36.7)	46 (9.0)	
1	12,145 (31.7)	1693 (43.8)	171 (44.6)	982 (42.2)	312 (44.7)	165 (32.4)	
2	624 (1.6)	280 (7.2)	107 (27.9)	261 (11.2)	104 (14.9)	186 (36.5)	
3	69 (0.2)	34 (0.9)	34 (8.9)	54 (2.3)	25 (3.6)	92 (18.1)	
4	7 (0.0)	6 (0.2)	2 (0.5)	18 (0.8)	1 (0.1)	16 (3.1)	
5	0 (0.0)	0 (0.0)	0 (0.0)	5 (0.2)	0 (0.0)	4 (0.8)	
ASA[d] classification (%)							<.001
1	13,389 (34.9)	233 (6.0)	1 (0.3)	179 (7.7)	12 (1.7)	0 (0.0)	
2	20,544 (53.6)	2319 (60.0)	92 (24.0)	907 (38.9)	226 (32.4)	42 (8.3)	
3	4142 (10.8)	1218 (31.5)	231 (60.3)	1034 (44.4)	377 (54.0)	330 (64.8)	
4	226 (0.6)	93 (2.4)	58 (15.1)	175 (7.5)	77 (11.0)	136 (26.7)	
5	47 (0.1)	5 (0.1)	1 (0.3)	34 (1.5)	6 (0.9)	1 (0.2)	

XSL•FO
RenderX

	Group 1 (n=38,348)	Group 2 (n=3868)	Group 3 (n=383)	Group 4 (n=2329)	Group 5 (n=698)	Group 6 (n=509)	*P* value
QCG[e] scores, median (IQR)							
ACS[f]	1.04 (0.22-3.28)	2.80 (1.18-6.22)	5.70 (3.00-10.76)	13.05 (2.30-25.76)	8.30 (2.77-23.89)	16.95 (5.60-32.72)	<.001
STEMI[g]	0.01 (0.00-0.04)	0.03 (0.01-0.14)	0.15 (0.03-0.60)	0.34 (0.05-1.51)	0.40 (0.06-1.65)	1.40 (0.23-6.80)	<.001
Myocardial injury	1.32 (0.39-3.46)	3.16 (1.47-6.10)	5.91 (3.36-9.60)	16.82 (5.80-25.89)	14.22 (6.28-23.58)	19.50 (10.23-33.50)	<.001
Pulmonary edema	0.49 (0.16-1.50)	1.51 (0.53-4.02)	4.00 (1.29-8.99)	12.02 (1.72-25.26)	18.05 (4.34-26.97)	21.60 (9.10-36.38)	<.001
Pericardial effusion	0.02 (0.00-0.14)	0.06 (0.01-0.45)	0.27 (0.04-1.58)	0.89 (0.05-6.10)	2.48 (0.22-9.80)	3.10 (0.70-9.10)	<.001
Left ventricular dysfunction	0.06 (0.02-0.21)	0.15 (0.04-0.61)	0.95 (0.30-3.75)	3.07 (0.47-16.05)	3.50 (0.72-13.14)	21.00 (5.41-54.70)	<.001
Right ventricular dysfunction	0.10 (0.03-0.36)	0.26 (0.08-1.10)	1.10 (0.32-3.65)	3.80 (0.48-15.12)	5.62 (1.11-16.47)	10.90 (3.90-26.60)	<.001
Pulmonary hypertension	0.11 (0.03-0.48)	0.40 (0.09-1.72)	1.45 (0.30-5.27)	4.05 (0.37-15.08)	7.90 (1.43-19.79)	10.77 (3.40-25.20)	<.001

[a]DM: diabetes mellitus.

[b]ESC: European Society of Cardiology.

[c]RCRI: Revised Cardiac Risk Index.

[d]ASA: American Society of Anesthesiologists.

[e]QCG: quantitative score for electrocardiography.

[f]ACS: acute coronary syndrome.

[g]STEMI: ST-segment elevation myocardial infarction.

XSL•FO
RenderX

Figure 4. Total cost and composite outcome rate by six study groups. The bar graphs represent the total cost observed (filled with blue), and the predicted additional cost by assuming that cases in group 4 undergo advanced cardiovascular imaging according to the proportion of individual tests (the part shown in dotted line), respectively (US $1000). The red line graph represents the composite outcome rate (%).

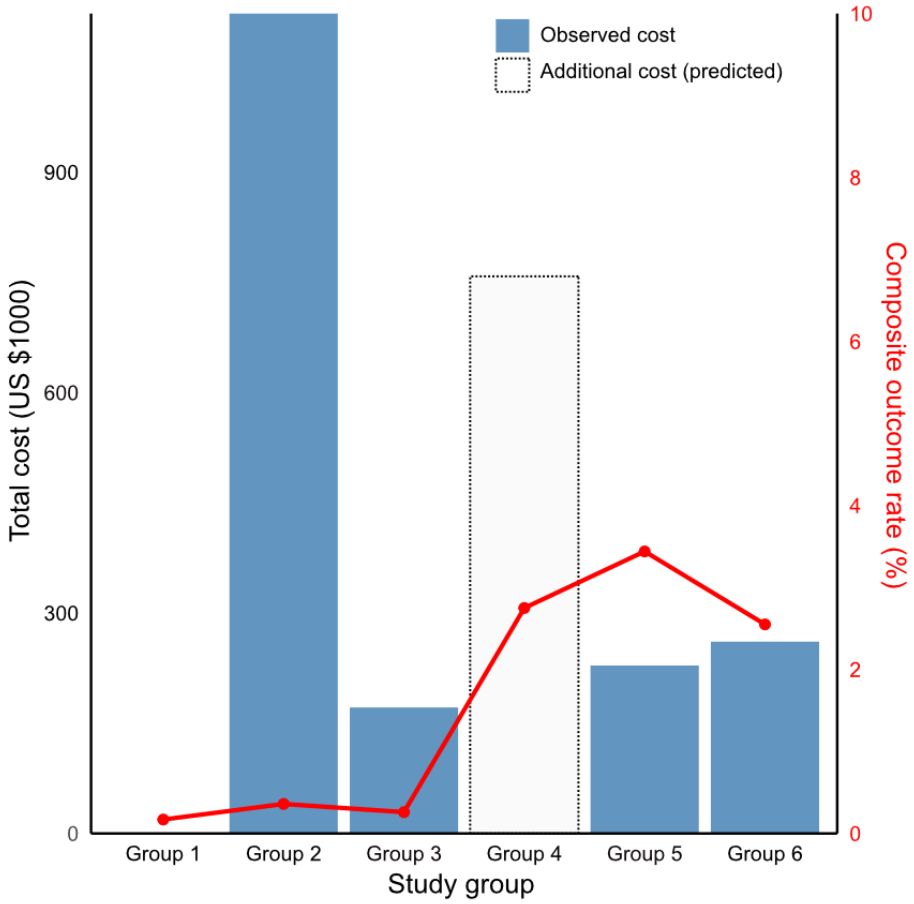

Sensitivity Analysis

When analyses were restricted to major surgery under general anesthesia, elective surgeries, and emergency surgeries, the association between AI-ECG risk stratification and postoperative outcomes remained consistent. In group-wise analyses, groups classified as high-risk by AI-ECG (groups 4, 5, and 6) consistently exhibited higher adverse event rates than those classified as low-risk (groups 1, 2, and 3) across all surgical subgroups (Table S5 in Multimedia Appendix 1). Similarly, in the patient-level sensitivity analysis restricted to each patient's first surgery (n=41,218), the trend in event rates and RRs remained consistent with those of the primary analysis (Table S6 in Multimedia Appendix 1).

To further address potential confounding by indication regarding preoperative evaluation, we additionally performed a propensity score-matched analysis. In the balanced cohort of 5122 well-matched pairs (n=10,244), the overall trends in composite outcomes remained highly consistent with our primary unmatched analysis, confirming the robustness of the AI-ECG risk stratification (Tables S7 and S8 in Multimedia Appendix 1).

Discussion

Principal Findings

In this study, we demonstrated the potential utility of multitask AI-enhanced ECG in reducing the tests with limited incremental diagnostic value and the associated costs in preoperative risk assessment. Among noncardiac surgeries, >90% were classified as low-risk by AI-ECG, with a high negative predictive value of the AI-ECG low-risk group designation for the primary outcome, supporting the utility of AI-ECG as a rule-out screening tool in the preoperative setting. Notably, in those who underwent preoperative cardiovascular testing despite a low-risk AI-ECG classification, 91% had negative results; however, this negative-test group incurred more than half of the overall expenditure for preoperative cardiovascular testing in all surgeries. When the AI-ECG risk classification was integrated with traditional risk-stratification tools, such as the ESC surgical risk category and RCRI, cases identified as low-risk had an extremely low incidence of postoperative composite outcomes. This potentially enables the conservation of medical resources by minimizing low-yield advanced cardiovascular evaluation before noncardiac surgery.

The indiscriminate use of preoperative cardiovascular tests has raised several important concerns. Overuse not only increases health care costs and delays surgery but may also postpone tests

XSL•FO
RenderX

intended for diagnosing other urgent conditions. Furthermore, the prevalence of serious cardiac diseases that require a change in surgical plan is low in elective surgery candidates, unless they have cardiac symptoms or signs, resulting in a low yield of meaningful results [23]. Nevertheless, preoperative cardiovascular tests, especially echocardiography, are still being performed even before low-risk surgery in a substantial number of cases. In a retrospective study of low-risk surgery, 39.8% of cases underwent echocardiography, among which only 5% showed abnormal results [24]. Moreover, among the diverse findings that could be obtained from echocardiography, the guidelines for perioperative cardiac management require only left and right ventricular function as preoperative cardiovascular diagnostic testing. Consequently, echocardiography is not recommended routinely in asymptomatic and clinically stable patients before noncardiac surgery owing to the lack of proven benefit [5]. In the meantime, no single cardiovascular test can fully assess perioperative risk. For example, normal echocardiography findings cannot reliably exclude ischemic heart disease without complementary stress tests. Importantly, postoperative cardiac complications are heterogeneous, ranging from ACS requiring unplanned PCI to heart failure resulting in prolonged intensive care unit stay and even mortality, which cannot be predicted by a single diagnostic modality. Taken together, these limitations highlight the unmet need for a novel and more comprehensive risk-stratification tool to identify high-risk surgical candidates who require further preoperative cardiovascular evaluation. Such a tool would enable the more appropriate use of limited health care resources, reduce overusage of testing, and ultimately improve perioperative outcomes.

In this study, we introduced a multitask AI-ECG, originally developed to detect patients at risk of a wide spectrum of cardiac conditions, including ST-segment elevation myocardial infarction, ACS, left and right ventricular dysfunction, pulmonary edema, pulmonary hypertension, and pericardial effusion, as a novel tool to identify surgical candidates at a low-risk of composite postoperative outcomes. When all eight QCG scores were below the optimal threshold, we designated this surgery as low-risk. To evaluate the clinical utility of this approach, we categorized all surgery-ECG pairs into three groups according to the results of the preoperative cardiovascular testing: no advanced cardiovascular imaging, negative-test, and positive-test. Each group was further stratified according to the AI-ECG risk classification (low-risk vs high-risk), resulting in six subgroups. Assuming that the decision to perform preoperative cardiovascular testing strictly followed the AI-ECG risk stratification (ie, tests performed only in high-risk cases), the number of surgeries in which cardiovascular tests were performed (groups 2, 3, 5, and 6; n=5458) exceeded the number of surgeries in which testing would have been required (groups 4, 5, and 6; n=3536). In a hypothetical scenario assuming strict adherence to the AI-ECG guidance, this discrepancy suggests a potential reduction of up to 35.2% in the usage of low-yield preoperative cardiovascular testing (Figure 1). Given that multiple cardiovascular tests may be required in some patients before noncardiac surgery, preoperative evaluation using multitask AI-ECG could be more practical by providing multiple disease-specific risk scores simultaneously from a single ECG image. In contrast to other machine learning algorithms that rely on numerous clinical variables from electronic health care records to predict postoperative adverse events [25-27], our platform enables simple and efficient on-site risk stratification using only a single ECG.

Group-Wise Analysis

In group-wise analysis, group 2 emerged as the most notable subgroup. Despite being classified as low-risk by AI-ECG and demonstrating a low prevalence of composite outcomes, this group (n=3868) accounted for more than half of the total patient expenses. On the other hand, group 4 (n=2329), characterized by a high AI-ECG risk, did not undergo testing despite their higher event rate. Taken together, approximately 40% (1539/3868) of the cardiovascular tests might be considered as overused testing, and the composite outcome incidence was markedly higher in group 4 than in group 2. As the symptoms and signs that led to advanced cardiovascular imaging were not accounted for in our assumption, all tests in group 2 should not be treated as unnecessary. However, this finding underscores that AI-ECG not only potentially decreases testing but also provides more appropriate risk stratification.

Group 3 represents another potentially “missed” population under an AI-ECG–guided strategy. Importantly, a low-risk AI-ECG classification in group 3 does not imply the absence of cardiovascular disease but rather reflects lower physiological vulnerability at the time of surgery. Additionally, this must be interpreted with caution due to the treatment paradox and verification bias. The favorable outcomes observed in group 3 may be a direct result of intense medical optimization, altered anesthetic management, and heightened postoperative monitoring initiated by the abnormal test findings, rather than confirming that the advanced cardiovascular testing was unnecessary.

Group 6—cases classified as high-risk by AI-ECG with positive preoperative cardiovascular testing—demonstrated relatively lower event rates compared to groups 4 and 5, despite carrying the highest burden of structural cardiovascular disease. As with group 3, this may be partly explained by the fact that a subset of group 6 cases involved intensified perioperative management, such as coronary revascularization before surgery. Figure S6 in Multimedia Appendix 1 illustrates a representative case where AI-ECG risk scores significantly decreased after the revascularization procedure. This temporal change supports the interpretation that AI-ECG reflects dynamic physiological cardiac state, capturing risk modification after therapeutic intervention rather than solely reflecting static anatomical disease burden. Furthermore, the paradoxically higher event rate in group 5 (negative cardiovascular testing) than in group 6 suggests the limited ability of conventional imaging to completely rule out the residual cardiovascular risk.

Notably, when VHD was analyzed in detail, severe VHD was more frequently observed in group 6 than in group 3, and markers of advanced physiological burden, including reduced left ventricular ejection fraction and multiple severe valvular lesions, were largely confined to group 6 (Table S4 in Multimedia Appendix 1). These findings suggest that even

XSL•FO
RenderX

among cases with structural heart disease, AI-ECG preferentially classified those with greater physiological burden into the high-risk group. Nevertheless, we acknowledge that structural valvular abnormalities that do not manifest prominent electrical changes bear an inherent limitation of ECG-based models. Accordingly, a low-risk AI-ECG result should not be interpreted as a substitute for echocardiography in cases with clinical findings suggestive of valvular disease. We suggest a stepwise pipeline where a standard physical examination (especially auscultation) is followed by the AI-ECG to triage the need for advanced cardiovascular tests, rather than using the AI-ECG as a stand-alone decision maker. We expect that future advances of the AI-ECG model, including VHD prediction, will refine the performance of AI-ECG in preoperative risk stratification before noncardiac surgery.

Group 5 provides meaningful insights. Although this group demonstrated no abnormal findings in preoperative cardiovascular tests, it exhibited the highest rate of composite outcomes among all groups. A large proportion of these cases had undergone prior open-heart surgery, coronary artery bypass grafting, PCI, or stable coronary artery disease. Given the well-recognized elevated perioperative risk in these populations, repeated cardiovascular testing with unchanged results is unlikely to affect intraoperative or postoperative management. Therefore, combined with delicate history-taking about their symptoms, AI-ECG might be able to replace low-yield repeated examinations in clinically stable patients with a history of cardiac surgery or PCI. In addition, conditions overlooked in the preoperative assessment, such as atrial fibrillation or underlying myocardial disease, may have contributed to the adverse outcomes. These factors highlight potential novel targets for reducing postoperative complications beyond the established strategies.

Potential Utility of AI-ECG Risk Stratification

Adding AI-ECG risk stratification to the model, incorporating age and RCRI, suggested potential for better identification of low-risk cases beyond age, yielding an improvement in the NRI and IDI. This improvement was largely driven by the effective identification of nonevents, which may facilitate more targeted resource allocation in low-risk individuals. Moreover, high-risk classification by AI-ECG was consistently associated with higher postoperative adverse event rates across diverse surgical subgroups in sensitivity analyses and in the PSM cohort. Clinically, however, the utility of AI-ECG–based screening may be greatest in specific settings, particularly in low-risk or emergency surgery where comprehensive preoperative evaluation is limited.

Low-risk patients are potential beneficiaries of AI-ECG risk stratification. Although current guidelines do not recommend performing additional tests in patients undergoing low-risk surgery not otherwise indicated by corresponding symptoms and/or signs of heart disease [5,6], a substantial number of low-risk surgeries are preceded by preoperative cardiovascular tests in the real world [24]. In our study, cases classified as low-risk (ie, having RCRI scores of 0 or 1 and involving low-risk surgeries by ESC surgical risk category) had very low composite outcome rates with high RR (Table 2). Namely, an AI-ECG could reduce low-yield tests, at least in patients with low surgical risk and low-risk AI-ECG. Therefore, integrating AI-ECG with established risk tools may outline patients who benefit from preoperative cardiovascular testing while curtailing the waste of medical resources.

Emergency or urgent surgeries could also benefit from an AI-ECG, in which the time for the preoperative workup is usually limited. As the risk of delaying surgery generally exceeds the benefit of information from preoperative examinations, risk assessment has been omitted for time-sensitive surgery [28]. However, the risk of developing composite outcomes in emergency surgeries is much higher than that in elective surgeries. If we could obtain information about the patient's critical cardiac condition without delay, the results might change with intensive monitoring and delicate postoperative care. Among emergency surgery cases, the risk of the composite end point was much lower in the low-risk AI-ECG group (1.6%) than in the high-risk AI-ECG group (8.5%). Therefore, medical resources for postoperative care can be reassigned according to the rapid risk classification by AI-ECG, particularly in resource-limited settings.

Potential for Health Care Resource Optimization

Beyond clinical outcomes, preoperative risk stratification using AI-ECG may offer meaningful implications for health care resource usage. In our cohort, 5458 cases underwent advanced cardiovascular testing, whereas only 3536 were classified as high-risk by AI-ECG. Theoretically, optimal integration of this tool could have decreased additional testing in up to 35.2% (1922/5458) of this population. However, because patient symptoms, signs, and functional capacity were not considered in this study, this reduction represents an absolute theoretical maximum and requires cautious interpretation. Furthermore, this estimation assumes complete clinician adherence to the AI-ECG recommendations, which may be difficult to achieve in real-world clinical practice due to defensive medicine, patient anxiety, and clinician intuition. These behavioral and systemic barriers would likely limit the actual reduction in medical resource usage. Nevertheless, it underscores its potential to optimize the medical resource allocation conceptually. Even when downstream cost analysis was performed under the exploratory assumption that cases in group 4 would involve advanced cardiovascular imaging in proportions similar to the overall tested cohort, AI-ECG–guided risk stratification continued to demonstrate favorable cost characteristics.

However, these findings should be interpreted in the context of each country's health care system, including the costs of individual tests, waiting times, and insurance coverage. In South Korea, where universal health insurance facilitates easy access to diagnostic testing, physicians may prescribe more tests than those recommended by the guidelines, potentially leading to an overestimation of costs. Furthermore, although diagnostic costs for group 4 were partially included, downstream costs related to clinical outcomes, including beneficial or adverse events, were not included in this study. Applying the mean test frequency of the entire tested cohort to estimate the costs for group 4 also potentially results in an underestimation of their expected expenditure. Therefore, this exploratory cost analysis

XSL•FO
RenderX

is not sufficient to prove the economic benefit of integrating the AI-ECG into preoperative evaluation, and a formal prospective cost-effectiveness analysis is necessary. Nevertheless, given its long-standing role in preoperative assessment, the ability of AI-ECG to refine risk stratification and reliably identify low-risk surgical candidates without additional expense may represent a meaningful advantage.

Limitations

First, caution is warranted when expanding the use of preoperative risk stratification using AI-ECG into patients undergoing all noncardiac surgery. As a single-center retrospective analysis, our AI model is susceptible to overfitting and algorithm drift and lacks external validation in demographically distinct populations. Furthermore, within our specific cohort, the proportion of low-risk surgeries was relatively high, and the overall mortality rate was lower than that reported in previous studies. Additionally, the stand-alone PPV of 2.9% is partly depressed because several AI-ECG scores (eg, pulmonary hypertension or pericardial effusion) flag structural conditions that do not necessarily culminate in mortality or unplanned PCI. Consequently, the low PPV reinforces the role of the AI-ECG as a triage tool, rather than a definitive stand-alone diagnostic test. Moreover, as our primary end point included all-cause mortality alongside unplanned PCI, it is not a fully cardiovascular-specific metric and provides a relatively indirect assessment of a cardiovascular screening tool. Restricting mortality to inpatient records may have also missed out-of-hospital deaths, potentially underestimating the true 30-day mortality rate. Additionally, because unplanned PCI primarily captures plaque-rupture ischemic events, our end point likely missed medically managed postoperative cardiac morbidities, such as myocardial injury after noncardiac surgery or type 2 myocardial infarction, potentially underestimating their true burden. Furthermore, the unique insurance system in South Korea may have influenced the frequency and cost of preoperative testing, limiting the generalizability of cost-related findings. Lastly, the exclusion of cases without a recent preoperative ECG yielded an inherently risk-enriched cohort; thus, the AI-ECG is intended to triage patients who already meet clinical indications for baseline testing, rather than serving as a universal screening tool. Accordingly, the findings may not be directly generalizable to all cases of noncardiac surgery.

Second, the thresholds defining high- and low-risk categories were derived during the original model-development stage and were not recalibrated for the present largely asymptomatic preoperative cohort. To mitigate potential miscalibration, we applied a conservative composite-negative definition of low-risk, classifying cases as low-risk only when all cardiovascular disease-specific QCG scores were below their respective thresholds. While this approach prioritizes safety and minimizes false reassurance, it inevitably elevates the cumulative false-positive rate and may obscure intermediate levels of risk. Future studies incorporating recalibration or graded risk categories may further improve clinical applicability.

Third, our retrospective design introduces inherent temporal and selection biases. As we defined the index ECG as the one closest to surgery within a 30-day window to optimally reflect the most recent preoperative state, unrecorded physiological fluctuations could occur during this interval between ECG and surgery. Furthermore, this design allowed for a temporal overlap where diagnostic imaging may have been performed before the index ECG. Thus, we could not fully account for the possible impact of medical optimization following abnormal testing before the index ECG. Although noninvasive imaging does not alter cardiac electrophysiology, a true clinical screening tool must temporally precede the decision to test. Moreover, although AI-ECG risk stratification was associated with postoperative outcomes, it remains unclear whether its use, followed by appropriate interventions, would improve clinical outcomes or reduce overall expenses. As preoperative testing was not randomly assigned and reflects clinical judgment, the observed associations should not be interpreted causally. Additionally, although we performed PSM analysis to balance the clinical likelihood of undergoing advanced cardiovascular testing, the matched cohort contains 1019 AI-ECG high-risk cases in the tested arm (groups 5 and 6 combined) vs 747 in the nontested arm (group 4). Therefore, a residual imbalance in the distribution of AI-ECG high-risk cases remained between the matched arms. Furthermore, baseline demographic differences remained unadjusted in the main analysis and might influence the observed event rates. Thus, the predictive value of AI-ECG risk stratification might be partially driven by unadjusted variables, such as advancing age and cumulative comorbidities. Nevertheless, it should be noted that our supplementary analyses demonstrated its independent prognostic performance even after adjusting for age and RCRI, suggesting that the AI-ECG provides unique prognostic value beyond advanced age and comorbidities. Future prospective studies should contextualize the stand-alone value of the AI-ECG against comprehensive calculators incorporating age, such as ACS National Surgical Quality Improvement Program.

Fourth, the definitions used for conventional cardiovascular testing may have influenced the comparative diagnostic yields of cardiovascular tests. Stress testing and biomarkers, such as high-sensitivity troponin and natriuretic peptides, were not included as preoperative cardiovascular tests, which may have led to an underestimation of the scope of conventional testing and potentially inflated the stand-alone value of the AI-ECG by ignoring biochemical risk clearance. Additionally, although resting echocardiography is a mainstay of preoperative evaluation in clinical practice, its ability to predict acute coronary ischemic events that necessitate unplanned PCI is limited. This modality–end point mismatch should be considered when interpreting the conventional testing as low-yield. Conversely, defining a positive anatomical finding as ≥50% coronary stenosis might have artificially inflated the diagnostic yield of cardiovascular testing. Using strictly actionable thresholds could widen the comparative efficiency between the AI-ECG tool and conventional testing.

Finally, as clinical symptoms, signs, and functional capacity were not incorporated in this study, further evaluation remains necessary for patients with suspected but undiagnosed cardiac disease, regardless of AI-ECG classification. AI-ECG–based risk stratification cannot replace history-taking or physical examination in the decision-to-test process. Additionally,

XSL•FO
RenderX

omitting echocardiography in patients with murmurs solely based on a low-risk AI-ECG score could lead to significant missed pathology. This underscores the role of the AI-ECG as an adjunctive screening tool rather than a stand-alone decision maker.

To overcome these limitations, prospective multicenter randomized controlled trials should be conducted for a comprehensive assessment of whether the AI-ECG can improve postoperative outcomes and reduce the cost related to preoperative cardiovascular tests in noncardiac surgery.

Conclusions

In conclusion, within the limitations of this retrospective study, a low-risk classification by multitask AI-ECG was associated with a low incidence of postoperative adverse events. By integrating AI-ECG with established risk-stratification tools, we identified a subgroup of cases in which advanced preoperative cardiovascular testing demonstrated limited yield. These exploratory findings suggest that AI-ECG may serve as a potentially resource-optimizing screening adjunct to present preoperative assessment strategies in noncardiac surgery. However, prospective validation is required to formally establish its safety and cost-effectiveness.

Acknowledgments

Generative artificial intelligence (ChatGPT, OpenAI) was used solely to proofread and improve the grammar and clarity of text drafted entirely by the authors. The tool was not used for ideation, analysis, or interpretation of data, or for the generation of tables or figures. Authors are fully responsible for the scientific integrity of this paper, and all artificial intelligence–assisted edits were reviewed and verified by the authors.

I-AS is a co-corresponding author of this paper and can be reached at sia1@snubh.org.

Data Availability

The datasets analyzed during this study are not publicly available due to institutional restrictions but are available from the corresponding author upon reasonable request.

Funding

This research was funded by the Korea Health Technology R&D Project (RS-2023-00265933) through the Korea Health Industry Development Institute, funded by the Ministry of Health and Welfare, Republic of Korea. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of this paper.

Authors' Contributions

All authors were involved in the conceptualization and design of this research. YK, JP, ICH, and GYC collected the data, and HMC, YK, JK, JHL, and IYO analyzed the data. JP, ICH, YEY, YYC, and JHL conducted the investigation. JK and YC were responsible for generating the artificial intelligence–enabled electrocardiography scores. HMC and YK drafted the manuscript and prepared the figures; YEY, JK, YC, and IAS reviewed and revised this paper. All the authors approved this paper. HMC and YK equally contributed as co–first authors. IAS and YC equally contributed as co–corresponding authors.

Conflicts of Interest

JK developed the algorithm and is the founder and chief executive officer of the start-up ARPI Inc. YC worked as the research director at ARPI Inc.

Multimedia Appendix 1
Data for artificial intelligence–enabled electrocardiography (AI-ECG)–based preoperative risk stratification.
[DOCX File , 2479 KB - jmir_v28i1e90099_app1.docx]

References

1. Weiser TG, Regenbogen SE, Thompson KD, Haynes AB, Lipsitz SR, Berry WR, et al. An estimation of the global volume of surgery: a modelling strategy based on available data. Lancet 2008 Jul 12;372(9633):139-144. [doi: 10.1016/S0140-6736(08)60878-8] [Medline: 18582931]
2. Weiser TG, Haynes AB, Molina G, Lipsitz SR, Esquivel MM, Uribe-Leitz T, et al. Estimate of the global volume of surgery in 2012: an assessment supporting improved health outcomes. Lancet 2015 Apr 27;385 Suppl 2:S11. [doi: 10.1016/S0140-6736(15)60806-6] [Medline: 26313057]
3. Richman DC. Ambulatory surgery: how much testing do we need? Anesthesiol Clin 2010 Jun;28(2):185-197. [doi: 10.1016/j.anclin.2010.03.001] [Medline: 20488389]
4. Devereaux PJ, Sessler DI. Cardiac complications in patients undergoing major noncardiac surgery. N Engl J Med 2015 Dec 03;373(23):2258-2269. [doi: 10.1056/NEJMra1502824] [Medline: 26630144]

XSL•FO
RenderX

5. Thompson A, Fleischmann KE, Smilowitz NR, de Las Fuentes L, Mukherjee D, Aggarwal NR, Peer Review Committee Members. 2024 AHA/ACC/ACS/ASNC/HRS/SCA/SCCT/SCMR/SVM guideline for perioperative cardiovascular management for noncardiac surgery: a report of the American college of cardiology/American heart association joint committee on clinical practice guidelines. Circulation 2024 Nov 05;150(19):e351-e442 [FREE Full text] [doi: 10.1161/CIR.0000000000001285] [Medline: 39316661]
6. Halvorsen S, Mehilli J, Cassese S, Hall TS, Abdelhamid M, Barbato E, ESC Scientific Document Group. 2022 ESC Guidelines on cardiovascular assessment and management of patients undergoing non-cardiac surgery. Eur Heart J 2022 Oct 14;43(39):3826-3924 [FREE Full text] [doi: 10.1093/eurheartj/ehac270] [Medline: 36017553]
7. Kim EK, Choi HM, Choi E, Lee HS, Park G, Han DW, et al. Pre-operative echocardiograhy for prevention of cardiovascular events after non-cardiac surgery in intermediate- and high-risk patients: protocol for a low-interventional, mixed-cohort prospective study design (PREOP-ECHO). Trials 2022 Sep 14;23(1):776 [FREE Full text] [doi: 10.1186/s13063-022-06701-2] [Medline: 36104714]
8. Correll DJ, Hepner DL, Chang C, Tsen L, Hevelone ND, Bader AM. Preoperative electrocardiograms: patient factors predictive of abnormalities. Anesthesiology 2009 Jun;110(6):1217-1222. [doi: 10.1097/ALN.0b013e31819fb139] [Medline: 19417620]
9. De Hert SG. Preoperative electrocardiograms: obsolete or still useful? Anesthesiology 2009 Jun;110(6):1205-1206. [doi: 10.1097/ALN.0b013e31819fb179] [Medline: 19417619]
10. Amini K, Mirzaei A, Hosseini M, Zandian H, Azizpour I, Haghi Y. Assessment of electrocardiogram interpretation competency among healthcare professionals and students of Ardabil University of Medical Sciences: a multidisciplinary study. BMC Med Educ 2022 Jun 09;22(1):448 [FREE Full text] [doi: 10.1186/s12909-022-03518-0] [Medline: 35681191]
11. Kashou AH, Noseworthy PA, Beckman TJ, Anavekar NS, Cullen MW, Boswell CL, et al. Predictors of ECG interpretation proficiency in healthcare professionals. Curr Probl Cardiol 2023 Dec;48(12):102011. [doi: 10.1016/j.cpcardiol.2023.102011] [Medline: 37544624]
12. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. Nat Rev Cardiol 2021 Jul;18(7):465-478 [FREE Full text] [doi: 10.1038/s41569-020-00503-2] [Medline: 33526938]
13. Palermi S, Vecchiato M, Ng FS, Attia Z, Cho Y, Anselmino M, International AI-ECG Working Group. Artificial intelligence and the electrocardiogram: a modern renaissance. Eur J Intern Med 2025 Oct;140:106329. [doi: 10.1016/j.ejim.2025.04.036] [Medline: 40413058]
14. Choi YJ, Park MJ, Ko Y, Soh MS, Kim HM, Kim CH, et al. Artificial intelligence versus physicians on interpretation of printed ECG images: diagnostic performance of ST-elevation myocardial infarction on electrocardiography. Int J Cardiol 2022 Sep 15;363:6-10. [doi: 10.1016/j.ijcard.2022.06.012] [Medline: 35691440]
15. Kim D, Hwang JE, Cho Y, Cho HW, Lee W, Lee JH, et al. A retrospective clinical evaluation of an artificial intelligence screening method for early detection of STEMI in the emergency department. J Korean Med Sci 2022 Mar 14;37(10):e81 [FREE Full text] [doi: 10.3346/jkms.2022.37.e81] [Medline: 35289140]
16. Kim D, Jeong J, Kim J, Cho Y, Park I, Lee S, et al. Hyperkalemia detection in emergency departments using initial ECGs: a smartphone AI ECG analyzer vs. board-certified physicians. J Korean Med Sci 2023;38(45):e322. [doi: 10.3346/jkms.2023.38.e322] [Medline: 37987103]
17. Choi YJ, Park MJ, Cho Y, Kim J, Lee E, Son D, et al. Screening for RV dysfunction using smartphone ECG analysis app: validation study with acute pulmonary embolism patients. J Clin Med 2024;13(16):4792. [doi: 10.3390/jcm13164792] [Medline: 39200934]
18. Cho Y, Yoon M, Kim J, Lee JH, Oh I, Lee CJ, et al. Artificial intelligence-based electrocardiographic biomarker for outcome prediction in patients with acute heart failure: prospective cohort study. J Med Internet Res 2024;26:e52139 [FREE Full text] [doi: 10.2196/52139] [Medline: 38959500]
19. Park J, Kim J, Kang SH, Lee J, Hong Y, Chang HJ, et al. Artificial intelligence-enhanced electrocardiography analysis as a promising tool for predicting obstructive coronary artery disease in patients with stable angina. Eur Heart J Digit Health 2024;5(4):444-453. [doi: 10.1093/ehjdh/ztae038] [Medline: 39081950]
20. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007;370(9596):1453-1457. [doi: 10.1016/S0140-6736(07)61602-X] [Medline: 18064739]
21. Lee SP, Seo JK, Hwang IC, Park JB, Park EA, Lee W, et al. Coronary computed tomography angiography vs. myocardial single photon emission computed tomography in patients with intermediate risk chest pain: a randomized clinical trial for cost-effectiveness comparison based on real-world cost. Eur Heart J Cardiovasc Imaging 2019;20(4):417-425. [doi: 10.1093/ehjci/jey099] [Medline: 30052964]
22. Price information by item. National Health Insurance System. URL: https://www.nhis.or.kr/nbinfo/wbhfaa06200m28.do [accessed 2026-06-11]
23. Dakik HA, Kobrossi S, Tamim H. The yield of routine pre-operative cardiovascular evaluation in stable patients scheduled for elective non-cardiac surgery. Int J Cardiol 2015;186:325-327. [doi: 10.1016/j.ijcard.2015.03.241] [Medline: 25880321]

XSL•FO
RenderX

24. Kim EK, Choi H, Lee J, Han DW, Lee HS, Choi E. Practical role of preoperative echocardiography in low-risk non-cardiac surgery. Front Cardiovasc Med 2023;10:1088496. [doi: 10.3389/fcvm.2023.1088496] [Medline: 36755797]
25. Bertsimas D, Dunn J, Velmahos GC, Kaafarani HMA. Surgical risk is not linear: derivation and validation of a novel, user-friendly, and machine-learning-based predictive optimal trees in emergency surgery risk (POTTER) calculator. Ann Surg 2018;268(4):574-583. [doi: 10.1097/SLA.0000000000002956] [Medline: 30124479]
26. Mahajan A, Esper S, Oo TH, McKibben J, Garver M, Artman J, et al. Development and validation of a machine learning model to identify patients before surgery at high risk for postoperative adverse events. JAMA Netw Open 2023;6(7):e2322285 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.22285] [Medline: 37418262]
27. Graeßner M, Jungwirth B, Frank E, Schaller SJ, Kochs E, Ulm K, et al. Enabling personalized perioperative risk prediction by using a machine-learning model based on preoperative data. Sci Rep 2023;13(1):7128. [doi: 10.1038/s41598-023-33981-8] [Medline: 37130884]
28. Steeds RP, Garbi M, Cardim N, Kasprzak JD, Sade E, Nihoyannopoulos P, 2014–2016 EACVI Scientific Documents Committee;, 2014–2016 EACVI Scientific Documents Committee. EACVI appropriateness criteria for the use of transthoracic echocardiography in adults: a report of literature and current practice review. Eur Heart J Cardiovasc Imaging 2017;18(11):1191-1204 [FREE Full text] [doi: 10.1093/ehjci/jew333] [Medline: 28329307]

Abbreviations

ACS: acute coronary syndrome
AI: artificial intelligence
CAG: coronary angiography
CCTA: coronary computed tomography angiography
ECG: electrocardiography
ESC: European Society of Cardiology
GEE: generalized estimating equation
IDI: integrated discrimination improvement
NRI: net reclassification improvement
PCI: percutaneous coronary intervention
PPV: positive predictive value
PSM: propensity score matching
QCG: quantitative score for electrocardiography
RCRI: Revised Cardiac Risk Index
RR: relative risk
SPECT: single-photon emission computed tomography
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
VHD: valvular heart disease

Edited by A Coristine; submitted 21.Dec.2025; peer-reviewed by H Lv; comments to author 22.Jan.2026; revised version received 01.Jun.2026; accepted 02.Jun.2026; published 17.Jun.2026.

Please cite as:
Choi HM, Kim Y, Kim J, Park J, Hwang IC, Choi YY, Lee JH, Yoon YE, Oh IY, Cho GY, Song IA, Cho Y
Multitask Artificial Intelligence–Based Electrocardiogram Tool for Preoperative Cardiac Testing in Noncardiac Surgery: Retrospective Cohort Study of Health Care Utilization and Costs
J Med Internet Res 2026;28:e90099
URL: https://www.jmir.org/2026/1/e90099
doi:10.2196/90099
PMID:

©Hong-Mi Choi, Yerin Kim, Joonghee Kim, Jiesuck Park, In-Chang Hwang, Yun Young Choi, Ji Hyun Lee, Yeonyee E Yoon, Il-Young Oh, Goo-Yeong Cho, In-Ae Song, Youngjin Cho. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparative Evaluation of ChatGPT, Google Translate, and UD Talk for Chinese-to-Japanese Translation in Cardiology and Pulmonology Outpatient Consultations: Prospective Observational Study

Ke-Yun Chao[1,2,3,4], PhD; Chia-Hui Huang[5], PhD; Wei-Lun Liu[4,6,7], MD; Chao-Yu Chen[8], MSc; Ai Yin Lim[9], PhD; Yi-Zhen Chu[2], BSc; Atsuko Sasaoka[5], MA; Chung-Yu Lin[10,11], MD

[1]Department of Respiratory Therapy, College of Medicine, Fu Jen Catholic University, New Taipei City, Taiwan
[2]Telehealth Center, Fu Jen Catholic University Hospital, Fu Jen Catholic University, New Taipei City, Taiwan
[3]Cardiovascular and Pulmonary Rehabilitation Center, Fu Jen Catholic University Hospital, Fu Jen Catholic University, New Taipei City, Taiwan
[4]Center for Health Research and Innovation, Fu Jen Catholic University, New Taipei City, Taiwan
[5]Graduate Institute of Cross-Cultural Studies, Fu Jen Catholic University, New Taipei City, Taiwan
[6]School of Medicine, College of Medicine, Fu Jen Catholic University, New Taipei City, Taiwan
[7]Department of Critical Care Medicine, Fu Jen Catholic University Hospital, Fu Jen Catholic University, New Taipei City, Taiwan
[8]Department of Life Science, Fu Jen Catholic University, New Taipei City, Taiwan
[9]Department of Physical Therapy, College of Medical and Health Science, Asia University, Taichung, Taiwan
[10]Division of Cardiology, Fu Jen Catholic University Hospital, Fu Jen Catholic University, New Taipei City, Taiwan
[11]Department of Medical Education, Fu Jen Catholic University Hospital, Fu Jen Catholic University, New Taipei City, Taiwan

Corresponding Author:
Chung-Yu Lin, MD
Division of Cardiology
Fu Jen Catholic University Hospital, Fu Jen Catholic University
No.69, Guizi Rd., Taishan Dist.
New Taipei City, 24352
Taiwan
Phone: 886 2 29052527
Fax: 886 85128729
Email: Chungyulin0727@gmail.com

Abstract

Background: Language barriers between health care providers and patients can compromise communication quality, patient safety, and health care equity. When professional interpreter services are limited, particularly in outpatient settings, artificial intelligence–based translation tools may serve as supplementary communication aids.

Objective: This study compared 3 Chinese-to-Japanese translation tools (ChatGPT, Google Translate, and UD Talk) in terms of their performance in real-world cardiology and pulmonology outpatient consultations.

Methods: In this single-center prospective observational study, audio-recorded outpatient consultations between December 2024 and November 2025 were analyzed. Verbatim physician-patient dialogues were translated using the 3 systems. Selected dialogue exchanges were evaluated by professional medical interpreters for translation accuracy and by Japanese-speaking lay participants for translation satisfaction using a 6-point Likert scale. Similarity of translation outputs was assessed at the dialogue exchange level.

Results: A total of 20 outpatient consultations comprising 2450 dialogue exchanges were analyzed. ChatGPT had significantly higher translation accuracy and satisfaction scores than Google Translate and UD Talk, with median scores of 5.0 (IQR 4.0-5.0) for both outcomes compared with 2.0 (IQR 1.0-3.0) for Google Translate and UD Talk ($P<.001$), and maintained a consistently high performance in both specialties. Google Translate and UD Talk exhibited substantially higher similarity (2131/2450, 87%) to each other than to ChatGPT in translated outputs (119/2450, 4.9% to 122/2450, 5%, respectively; $P<.001$).

XSL•FO
RenderX

Conclusions: In Chinese-to-Japanese translations of outpatient medical consultations, ChatGPT demonstrated higher accuracy and user satisfaction than Google Translate and UD Talk. These findings indicate that artificial intelligence–assisted translation may support multilingual clinical communication when used as a complementary aid alongside professional interpreter services.

(J Med Internet Res 2026;28:e93082) doi:10.2196/93082

KEYWORDS

language barrier; medical communication; medical interpretation; cardiopulmonary disease; machine translation; artificial intelligence; AI

Introduction

When health care providers and patients do not share a common language, language barriers arise, hindering effective communication. These communication challenges have major implications for health care costs, quality of care [1], and equity in access to health care services [2]. From ethical and equity-based perspectives, health care professionals are obligated to provide high-quality, human rights–based care to all patients regardless of language differences [3]. Patients with limited proficiency in the dominant language of the health care system often face barriers to accessing care [4] and experience poorer health outcomes than native speakers [5,6].

According to the International Medical Interpreters Association, an interpreter's primary task is defined as transforming a message from a source language into an equivalent message in a target language such that it elicits the same response in the listener as the original message [7]. Prior studies have demonstrated that the involvement of professional interpreters ensures more favorable patient outcomes than other modes of interpretation [8-10]. Professionally trained on-site interpreters also provide more accurate and reliable verbatim translation than untrained individuals, such as patients' family members [8]. Qualitative studies involving interpreters, health care professionals, and patients have indicated that interpreters support communication beyond direct translation, such as by facilitating cultural understanding, clarifying medical information, and helping patients navigate the health care process [9,10]. Despite their importance in promoting equitable health care communication, professional interpreters remain underused in routine clinical practice because of structural and logistical barriers such as limited availability, time constraints, and administrative barriers. Patients often experience communication difficulties when medical information is not conveyed in their preferred language, particularly in outpatient settings, where interpreter services are less consistently available [11,12].

Google Translate is a widely used machine translation service and has been frequently examined in academic research on translation accuracy. Since 2016, Google Translate has primarily relied on Google Neural Machine Translation, a deep learning–based neural network that translates complete sentences rather than fragmented text segments, as was characteristic of earlier statistical machine translation methods. Although this change has improved overall translation quality, several studies evaluating Google Translate in medical contexts have reported variable accuracy and potential risks associated with translation errors, raising concerns regarding its suitability for unsupervised clinical use [13-16]. Studies evaluating Google Translate in medical contexts have reported limited accuracy. When counseling information for the 100 most commonly prescribed medications in the United States was translated from English into other languages, accuracy rates were only 54% for Arabic, 77% for Chinese, and 38% for Spanish. Among these errors, 29% were considered clinically significant or potentially life-threatening [17].

UD Talk [18] is a Japanese subtitle application that uses speech recognition technology to convert spoken language into real-time captions, typically within 1 or 2 seconds. The system integrates speech recognition and machine translation functions, and its multilingual translation component is primarily based on Google Translate technology. UD Talk is designed mainly for live interpretation scenarios, whereas Google Translate and ChatGPT are more commonly used for text-based translation tasks.

Advances in large language models (LLMs) have opened new avenues for automated medical translation, particularly in multilingual clinical settings. Artificial intelligence (AI) has become a major focus of research because it simulates human intelligence and cognitive processing to address complex problems [19]. When trained on large and specialized datasets, AI systems can learn linguistic patterns and respond to contextual challenges [20,21]. ChatGPT is an LLM introduced in 2022. Since then, it has been promoted as an advanced language-processing tool. Using extensive datasets, ChatGPT effectively captures linguistic nuance and contextual meaning across multiple languages, predicts word sequences based on context, and generates novel sentences that resemble natural human language. These outputs may be contextually appropriate and extend beyond previously observed text patterns [22].

This study aimed to compare 3 commonly used Chinese-to-Japanese translation tools (ChatGPT, Google Translate, and UD Talk) in real-world medical consultations. Using authentic cardiopulmonary outpatient dialogues, we evaluated the performance of these tools in terms of translation accuracy, error patterns, fluency, contextual appropriateness, user satisfaction, and clinical usability. The analysis further clarified the feasibility and limitations of AI-assisted translation tools for clinical communication and interpreter training.

Methods

Study Design

This single-center prospective observational study was conducted in the cardiology and pulmonology outpatient clinics at Fu Jen Catholic University Hospital, Taiwan, between December 2024 and November 2025.

XSL•FO
RenderX

Ethical Considerations

This study adhered to the principles of good clinical practice and the Declaration of Helsinki. The study protocol was approved by the institutional review board of Fu Jen Catholic University Hospital (approval FJUH113419) and was prospectively registered at ClinicalTrials.gov (identifier: NCT06934031). Written informed consent was obtained from all participants before taking part.

Study Cohort

Patients were eligible for enrollment if they attended the cardiology or pulmonology outpatient clinics at Fu Jen Catholic University Hospital between December 2024 and November 2025. Patients were excluded if the attending physician determined that they were unable to communicate effectively because of medical or cognitive conditions or if they declined to participate.

Data Collection and Translation

Audio recordings of physician-patient consultations were collected in the cardiology and pulmonology outpatient clinics. All recorded conversations were transcribed verbatim for analysis. To ensure data confidentiality and protect patient privacy, all personally identifiable information was removed before translation, including names, identification numbers, and any other information that could reveal patient identity (Figure 1).

To preserve real-world clinical communication, transcripts were retained in their original spoken Chinese form without linguistic modification or correction. Transcript volume was quantified using the number of physician-patient dialogue exchanges per outpatient consultation. A dialogue exchange was defined as each individual utterance occurring during the outpatient consultation. Each utterance was separately recorded, transcribed, translated, and labeled according to speaker identity, including physicians, patients, accompanying family members, and clinical staff.

For translation, ChatGPT (GPT-4o; OpenAI) and Google Translate (accessed between April 2025 and September 2025) were used by directly inputting the verbatim transcripts into each system. For ChatGPT translation, standardized prompt instructions were applied before each translation task. The prompts specified speaker identities (D=physician; P=patient; F=family member; S=nurse or other staff member; PF=possible patient or family member), identified the consultation setting as either cardiology or pulmonology outpatient consultation, instructed the model to translate the dialogue into Japanese, and requested translation output only without additional conversational responses or recommendations. Because routine outpatient consultations are inherently conversational and unstructured, no additional linguistic modification, dialogue restructuring, or contextual standardization was applied beyond the original speaker labels to preserve a realistic clinical communication environment as much as possible. In contrast, UD Talk (version 3.19; Shamrock Records; accessed between April 2025 and September 2025) was used in a speech-based manner: a trained research assistant read the original Chinese transcripts aloud, and the spoken input was translated in real time. To reduce variability in speech delivery, all UD Talk inputs were provided by the same research assistant throughout the study.

After translation by all 3 systems, an investigator (CHH) reviewed the outputs and selected dialogue exchanges for questionnaire-based evaluation. Dialogue exchanges with completely identical translated outputs across all 3 systems were excluded to facilitate meaningful comparison across translation systems.

Figure 1. Study workflow.

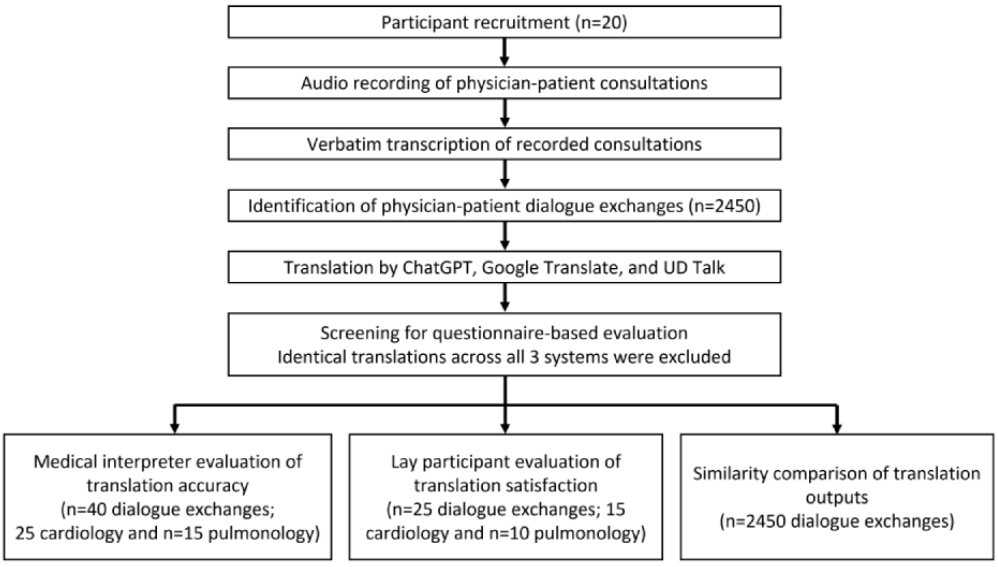

Evaluation

Translation performance was evaluated from 2 complementary perspectives: assessment of translation accuracy by experts (medical interpreters) and assessment of translation satisfaction by lay participants. A semistructured evaluation format was used, allowing for optional open-ended comments for each item.

XSL•FO
RenderX

Each selected dialogue exchange was independently evaluated by 8 medical interpreters and 14 lay participants. Evaluations were conducted using a 6-point Likert scale, with end points ranging from 0 (lowest) to 5 (highest).

For expert evaluation, medical interpreters with formal Chinese-to-Japanese medical interpretation training and at least 2 years of clinical practice experience assessed the accuracy of the translated content. Translation accuracy was defined as the extent to which the translated content conveyed the intended meaning of the original physician-patient communication within the clinical context of the consultation. Interpretation of intended meaning was determined by the research team based on contextual understanding and professional clinical knowledge. Ratings were based on the interpreters' professional judgment, recognizing that interpreters may prioritize different aspects of translation accuracy, such as semantic fidelity, medical correctness, and clinical appropriateness. Optional qualitative comments could be provided to explain scoring decisions when necessary.

For lay participant evaluation, Japanese individuals residing in Taiwan for at least 2 years rated their overall satisfaction with the translated content. Satisfaction ratings were based on perceived clarity, ease of understanding, and practical usability from the user's perspective in a clinical communication setting. In addition, both medical interpreters and lay participants completed a separate postevaluation perception survey regarding key components of medical interpretation quality and preferred translation tools.

All evaluation materials were derived from authentic physician-patient consultations and were presented in their original spoken Chinese form. To preserve real-world clinical communication, no linguistic modifications were made to the transcripts before evaluation, retaining incomplete sentences, colloquial expressions, and ambiguous phrasing.

Statistical Analysis

All statistical analyses were performed using SPSS (version 25; IBM Corp). Descriptive statistics were used to summarize participant characteristics and dialogue exchange volumes. Continuous variables are presented as means with SDs or medians with IQRs, and categorical variables are presented as numbers and percentages.

Translation accuracy and satisfaction scores were measured on ordinal Likert scales; therefore, nonparametric statistical methods were used. Differences in translation accuracy and satisfaction scores among the 3 translation systems were assessed using the Friedman test. Subgroup analysis by outpatient specialty (cardiology vs pulmonology) was conducted using the Wilcoxon rank-sum test. Similarity among the translation outputs of the 3 systems was analyzed at the level of physician-patient dialogue exchanges using the Cochran *Q* test. Interrater reliability for medical interpreter–rated accuracy and lay participant–rated satisfaction was evaluated using the Cronbach α coefficient. A 2-tailed *P* value of less than .05 indicated statistical significance.

Results

This study included 20 outpatient consultations from 10 cardiology patients and 10 pulmonology patients. The mean number of physician-patient dialogue exchanges per outpatient consultation was 122.5 (SD 41.9). The mean dialogue volume was lower in cardiology consultations than in pulmonology consultations (107.2, SD 37.0 vs 137.8, SD 42.7; Table 1). Eight medical interpreters participated in the expert evaluation. Among them, 2 were native Japanese speakers, 3 had 5 to 10 years of professional experience, 2 had 11 to 20 years of professional experience, and 3 had more than 20 years of professional experience. A total of 14 lay participants were recruited for satisfaction evaluation, all of whom were native Japanese speakers. Most lay participants (13/14, 92.9%) had resided in Taiwan for more than 10 years, including 69.2% (9/13) with 11 to 20 years of residence and 30.8% (4/13) with more than 20 years of residence. Interrater reliability analysis revealed high internal consistency for both medical interpreter and lay participant evaluations. Cronbach α coefficients were 0.942 for medical interpreter–rated translation accuracy and 0.966 for lay participant–rated translation satisfaction, indicating excellent agreement among raters in both groups. Following screening for questionnaire-based evaluation, 40 dialogue exchanges (cardiology: n=25, 62.5%; pulmonology: n=15, 37.5%) were selected for medical interpreter evaluation (Multimedia Appendix 1), and 25 dialogue exchanges (cardiology: n=15, 60%; pulmonology: n=10, 40%) were selected for lay participant evaluation. Dialogue exchanges with completely identical translated outputs across all 3 systems were excluded to facilitate meaningful comparison across translation systems (Multimedia Appendix 2).

Medical interpreter–rated translation accuracy significantly varied across the 3 translation systems (Table 2). Overall, ChatGPT was the most accurate tool, with a median score of 5.0 (IQR 4.0-5.0), whereas Google Translate and UD Talk had substantially lower median accuracy scores (2.0 for both, IQR 1.0-3.0; *P*<.001). This pattern was consistent between the outpatient specialties. ChatGPT maintained median accuracy scores of 5.0 (IQR 4.0-5.0) in both cardiology and pulmonology consultations (*P*<.001 in all cases). Lay participant–rated translation satisfaction also varied significantly among the 3 systems. ChatGPT received the highest satisfaction ratings, with a median score of 5.0 (IQR 4.0-5.0), whereas Google Translate and UD Talk both yielded median satisfaction scores of 2.0 (IQR 1.0-3.0; *P*<.001). Similarly, ChatGPT achieved the highest satisfaction ratings in both cardiology and pulmonology consultations (median 5.0, IQR 4.0-5.0; *P*<.001 in all cases).

Representative qualitative error examples are shown in Table 3. These examples illustrate clinically relevant translation issues observed in Google Translate and UD Talk outputs, including literal translation, incomplete conveyance of clinical meaning, and potentially misleading phrasing. In these cases, ChatGPT provided translations that were more contextually appropriate and clinically coherent.

Perceptions of medical interpretation quality and preferred translation tools were assessed among medical interpreters and

XSL•FO
RenderX

lay participants after completion of the evaluation (Figure 2). Clarity of expression was most frequently highlighted as a key component of translation accuracy by medical interpreters (87.5%), followed by faithful transfer of meaning, accuracy of medical terminology, and cultural and contextual appropriateness (each cited by 75% of all interpreters). Faithful transfer of meaning was most frequently highlighted as a key component of translation satisfaction by lay participants (13/14, 92.9%), followed by clarity of expression (10/14, 71.4%). Accuracy of medical terminology and cultural and contextual appropriateness were each identified by 64.3% (9/14) of all lay participants as key components (Figure 2A). Regarding preferred translation tools, ChatGPT was most frequently recommended by both medical interpreters (62.5%) and lay participants (10/14, 71.4%). Google Translate was selected by 50% (7/14) of all lay participants but only 12.5% (1/8) of all medical interpreters, whereas UD Talk was rarely preferred by either group (Figure 2B).

Similarity analysis revealed substantial overlap between Google Translate and UD Talk at the level of physician-patient dialogue exchanges (Table 4). Overall, 87% (2131/2450) of all exchanges produced identical translations between these 2 systems. In contrast, identical outputs with those of ChatGPT were observed in only 4.9% (119/2450) to 5% (122/2450) of all exchanges ($P<.001$). When stratified by outpatient specialty, the similarity between Google Translate and UD Talk remained high but varied by setting. In cardiology consultations, 74.5% (799/1072) of all dialogue exchanges yielded identical translations between Google Translate and UD Talk, but similarity with ChatGPT was low (29/1072, 2.7% and 31/1072, 2.9%, respectively; $P<.001$). In pulmonology consultations, the similarity between Google Translate and UD Talk increased to 96.7% (1332/1378), but similarity with ChatGPT remained low (90/1378, 6.5% and 91/1378, 6.6%, respectively; $P<.001$).

Subgroup analysis revealed higher translation accuracy and satisfaction in pulmonology consultations than in cardiology consultations for both Google Translate and UD Talk (Multimedia Appendix 3). In contrast, ChatGPT had consistently high accuracy and satisfaction in both specialties, with a median score of 5.0 (IQR 4.0-5.0) for both. No significant differences were observed between cardiology and pulmonology consultations in accuracy ($P=.25$) or satisfaction ($P=.26$) for ChatGPT-translated content.

Table 1. Characteristics of outpatient consultation recordings, medical interpreters, and lay participants.

Variable	Values
Dialogue exchanges per outpatient visit, mean (SD)	
Overall (n=20 outpatient consultations)	122.5 (41.9)
Cardiology outpatient clinic (n=10 outpatient consultations)	107.2 (37.0)
Pulmonology outpatient clinic (n=10 outpatient consultations)	137.8 (42.7)
Medical interpreters (n=8), n (%)	
Native Japanese speakers	2 (25)
Professional experience (y)	
5-10	3 (37.5)
11-20	2 (25)
>20	3 (37.5)
Lay participants (n=14), n (%)	
Native Japanese speakers	14 (100)
Length of residence in Taiwan (y)	
5-10	1 (7.1)
11-20	9 (64.3)
>20	4 (28.6)

XSL•FO
RenderX

Table 2. Medical interpreter–rated translation accuracy and lay participant–rated translation satisfaction scores for ChatGPT, Google Translate, and UD Talk[a].

	ChatGPT (0-5), median (IQR)	Google Translate (0-5), median (IQR)	UD Talk (0-5), median (IQR)	*P* value
Accuracy				
Overall (n=320)	5.0 (4.0-5.0)	2.0 (1.0-3.0)	2.0 (1.0-3.0)	<.001[b]
Cardiology outpatient clinic (n=200)	5.0 (4.0-5.0)	2.0 (0.0-3.0)	2.0 (0.0-3.0)	<.001[b]
Pulmonology outpatient clinic (n=120)	5.0 (4.0-5.0)	3.0 (1.0-4.0)	2.0 (1.0-4.0)	<.001[b]
Satisfaction				
Overall (n=350)	2.0 (1.0-3.0)	2.0 (1.0-3.0)	5.0 (4.0-5.0)	<.001[b]
Cardiology outpatient clinic (n=210)	2.0 (1.0-3.0)	2.0 (1.0-3.0)	5.0 (4.0-5.0)	<.001[b]
Pulmonology outpatient clinic (n=140)	3.0 (1.0-4.0)	3.0 (2.0-4.0)	5.0 (4.0-5.0)	<.001[b]

[a]Eight medical interpreters evaluated 40 dialogue exchanges (cardiology: n=25, 62.5%; pulmonology: n=15, 37.5%) for translation accuracy, and 14 lay participants evaluated 25 dialogue exchanges (cardiology: n=15, 60%; pulmonology: n=10, 40%) for translation satisfaction.
[b]$P<.001$.

Table 3. Representative clinically relevant translation errors across the 3 translation systems[a].

Clinical scenario and source	Dialogue	Expert commentary
Cardiology example		Both Google Translate and UD Talk mistakenly segmented “腎指數” as “kidney+index finger,” producing nonsensical output. ChatGPT correctly interpreted it as “renal function value” (腎機能の数値).
Original dialogue	Family member: “他的指、腎指數真的是非常不好，真的不能再打針了。”	
ChatGPT	Family member: “彼の腎機能の数値は本当に悪くて、もう注射は絶対にダメだって言われたんです。”	
Google Translate	Family member: “彼の指と腎臓の人差し指の状態は本当に悪く、もう注射を受けることができません。”	
UD Talk	Family member: “彼の指と腎臓の人差し指の状態は本当に悪く、もう注射を受けることは本当に不可能なんです。”	
Pulmonology example		Google Translate and UD Talk incorrectly translated the head “heaviness/fogginess” as physical swelling (腫れている). ChatGPT correctly rendered it as “ぼーっとする”—clinically accurate for the patient’s chief concern.
Original dialogue	Patient: “現在我就擔心這個頭一直脹脹的，不知道什麼 ”	
ChatGPT	Patient: “今はこの頭がずっとぼーっとする感じが気になっていて……何が原因なのか分からなくて。”	
Google Translate	Patient: “頭がずっと腫れているのが心配なんです。どうしてなのかわからないんですけど…”	
UD Talk	Patient: “頭が腫れている理由が心配です。何が起こっているのか分かりません…”	

[a]Original dialogues were retained verbatim from real outpatient recordings, including disfluencies, repetitions, self-corrections, and incomplete expressions. No linguistic corrections were made.

XSL•FO
RenderX

Figure 2. Perceptions of medical interpretation among medical interpreters and lay participants: (A) essential components of medical interpretation accuracy and translation satisfaction as perceived by medical interpreters and lay participants and (B) preferred translation tools of medical interpreters and lay participants.

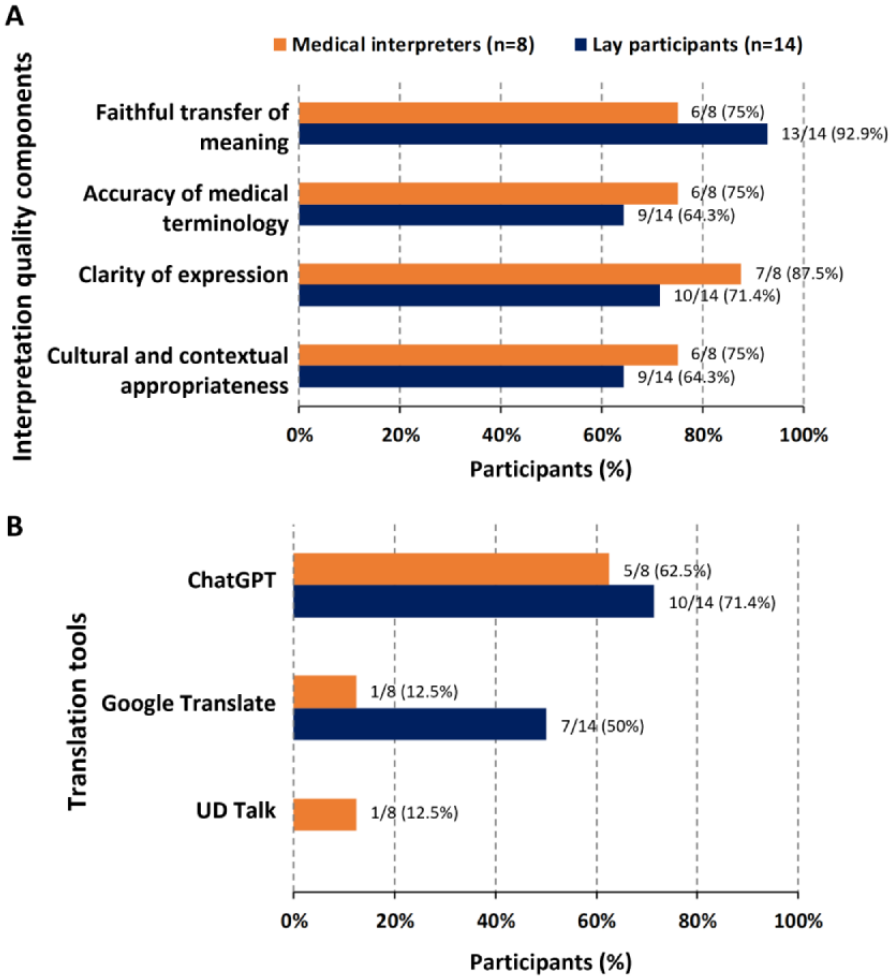

Table 4. Percentage similarity of translation outputs across systems[a].

	Google Translate vs UD Talk, n (%)	ChatGPT vs UD Talk, n (%)	ChatGPT vs Google Translate, n (%)	*P* value
Overall dialogue exchanges (n=2450)	2131 (87)	122 (5)	119 (4.9)	<.001[b]
Cardiology outpatient clinic dialogue exchanges (n=1072)	799 (74.5)	31 (2.9)	29 (2.7)	<.001[b]
Pulmonology outpatient clinic dialogue exchanges (n=1378)	1332 (96.7)	91 (6.6)	90 (6.5)	<.001[b]

[a]Similarity analyses were conducted at the level of physician-patient dialogue exchanges.

[b]*P*<.001.

Discussion

Principal Findings

We analyzed 20 outpatient consultations comprising 2450 physician-patient dialogue exchanges and compared 3 common Chinese-to-Japanese translation tools in terms of their performance in real-world clinical settings. In both cardiology and pulmonology consultations, ChatGPT consistently exhibited higher translation accuracy and user satisfaction than Google Translate and UD Talk, with median scores at the highest level of the rating scale. In contrast, Google Translate and UD Talk had substantially lower scores and exhibited high similarity in translated outputs, indicating shared limitations in handling conversational clinical language. This high similarity may reflect their reliance on conventional machine translation or speech recognition–based processing, which prioritizes literal sentence conversion rather than contextual interpretation.

Our findings suggest that ChatGPT may have potential utility as an AI-assisted translation tool in Chinese-to-Japanese outpatient medical communication. Unlike conventional

XSL•FO
RenderX

machine translation systems, which rely primarily on sentence- or phrase-level processing, ChatGPT generates natural and appropriate responses by interpreting broader contextual meaning. This difference in processing may explain the higher translation accuracy and satisfaction observed for ChatGPT, particularly when handling colloquial spoken Chinese and clinically nuanced dialogue. The findings also indicate that ChatGPT may be applicable across different outpatient specialties within the context of this study.

Although subgroup analysis revealed slightly higher translation accuracy and satisfaction scores in pulmonology consultations than in cardiology consultations, ChatGPT maintained high performance in both specialties. Specialty-based differences were noted primarily for Google Translate and UD Talk. This may partly reflect differences in the selected dialogue content. Cardiology exchanges more frequently involved treatment decision-making, procedural planning, medication use, self-paid treatment discussions, cardiac rhythm assessment, diagnostic certificates, and fragmented patient narratives. In contrast, pulmonology exchanges were more often related to symptom confirmation, examination arrangement, medication instructions, and patient education, with more explicit clinical context. These differences may have made pulmonology dialogues easier for conventional translation systems to process. However, ChatGPT exhibited the highest performance in both settings. Therefore, specialty-based differences are unlikely to affect the applicability of ChatGPT in different clinical departments. This consistency indicates that LLM-based translation tools are suitable for use in diverse outpatient specialties without substantial concern for specialty-specific performance limitations.

Comparison With Previous Literature and Clinical Implications

Despite the promising results, AI-based translation tools should not be regarded as replacements for professional medical interpreters. Prior research has demonstrated that trained interpreters contribute not only to the accuracy of verbal communication but also to the mediation of cultural differences, clarification of medical information, and facilitation of patient engagement [8,11,23-26]. Rao et al [27] reported that, although ChatGPT outperformed Google Translate in Spanish translation, it demonstrated inferior performance in Vietnamese and produced low-quality translations in Russian. In certain languages, both ChatGPT and Google Translate had unacceptably high error rates, indicating that AI-generated translations can be unreliable when used without human oversight. Although AI-based translation tools may perform well in select languages, their broader clinical use requires careful validation and continued human supervision to ensure patient safety and accurate communication. Kong et al [28] evaluated machine translation for patient-specific clinical instructions and reported generally high degrees of sentence-level accuracy but persistent risks of clinically meaningful errors and potential harm, particularly when translations were used without professional oversight or in more complex medical cases.

These findings should be interpreted in light of differences in study design, translation direction, and communication context. Unlike prior studies that often evaluated written medical instructions translated from English into other languages, this study focused on spoken Chinese outpatient dialogue translated into Japanese. Because real-world consultations include incomplete sentences, colloquial expressions, and implicit contextual meaning, translation performance may differ according to both language pair and communication format. This also supports the continued need for professional medical interpreters, who can integrate linguistic accuracy, clinical judgment, cultural mediation, and real-time clarification in complex clinical encounters.

AI-based translation systems may function as supplementary communication tools in settings where professional interpreter services are limited or unavailable, particularly in time-sensitive outpatient care. Future research should investigate how these technologies can be integrated into clinical workflows, interpreter training programs, and patient education to support multilingual communication while maintaining patient safety and quality of care. In practice, such tools may be particularly useful for brief outpatient encounters, preliminary triage communication, and patient education when professional interpreters are not immediately available.

Provision of professional interpreter services is recognized as a global strategy for reducing language barriers in health care. In countries such as Australia, Canada, the United States, and the United Kingdom, language service policies, standards, and guidelines have been established to promote or mandate the use of professional interpreter services in clinical care [29]. Despite these efforts, evidence indicates that the use of professional interpreter services remains suboptimal, whereas reliance on ad hoc interpreters, such as family members, friends, and other untrained individuals, is common [30,31]. In some clinical settings, bilingual health care staff assume the role of interpreters despite lacking formal training. This practice may compromise the quality of interpretation and increase the risk of adverse clinical outcomes [8,32,33]. Therefore, access to professional interpreter services must be enhanced to strengthen communication between health care providers and patients with limited proficiency in the dominant language. A systematic review of 37 studies indicated that reducing language barriers is a key strategy for improving patient safety and minimizing adverse events in hospital care. The review further reported that the use of professional interpreter services substantially improves communication between health care providers and patients and enhances the quality of care for linguistically diverse populations [29]. A systematic review of 29 studies compared 5 interpretation-related situations: involvement of professional interpreters, involvement of ad hoc interpreters, involvement of relational interpreters, involvement of untrained interpreters, and absence of interpreters. The results revealed that the involvement of face-to-face professional interpreters was associated with the highest patient satisfaction and the most effective communication. Moreover, any form of interpretation was determined to be superior to no interpreter use, and relational interpreters represented useful communication resources in certain private clinical settings [26].

XSL•FO
RenderX

Qualitative Insights From Medical Interpreters

In addition to the quantitative findings, qualitative feedback from medical interpreters provided insights into the strengths and limitations of the translation tools. Interpreters emphasized that accurate transfer of meaning and correct use of medical terminology are essential for clinical translation. Although some lay participants considered minor grammatical imperfections acceptable when the overall meaning remained understandable, interpreters exhibited low tolerance for semantic distortion or incorrect terminology because such errors could lead to clinical misunderstanding. ChatGPT was perceived to handle contextual information and incomplete or colloquial spoken language more effectively than the other systems. Interpreters noted that its translations often clarified implicit meaning and restructured fragmented sentences, which in some cases resembled the output produced by human interpreters. In contrast, Google Translate and UD Talk were frequently described as relying on more literal translation patterns, which resulted in awkward phrasing or incomplete conveyance of clinical meaning.

Interpreters indicated that, although AI-based translation tools performed favorably, they should be used with caution and primarily to support communication rather than to replace professional interpreters, particularly in high-risk clinical situations. Interpreters also noted that the use of AI translation in health care raises substantial legal, ethical, and policy challenges, such as concerns related to data privacy, informed consent, accountability for translation errors, and potential bias in multilingual communication [34]. Therefore, a collaborative human-AI approach was recommended to ensure accuracy, clarity, and patient safety in multilingual clinical communication while addressing these regulatory and ethical concerns.

Limitations

This study has several limitations. First, the analysis exclusively focused on Chinese-to-Japanese translation; thus, the findings may not be generalizable to other language pairs with different linguistic structures and resource availability levels. Second, although the dataset included 20 outpatient consultations and 2450 dialogue exchanges, the sample size and text volume may still be insufficient to capture the full diversity of real-world clinical communication. Third, consultations in each specialty were conducted by a single physician. Thus, variation in communication styles across physicians within the same specialty was not examined. Differences in physicians' speech patterns, use of medical terminology, sentence structure, and degree of colloquial expression may influence AI translation performance; thus, potential physician-specific effects cannot be excluded. Finally, translation performance was evaluated using GPT-4o, and the findings may not be generalizable to newer versions of ChatGPT or other LLMs. In the future, studies including multiple language pairs, larger datasets, diverse clinical settings, multiple physicians within each specialty, and updated AI models should be conducted to validate and extend our findings.

Conclusions

This study demonstrated that ChatGPT achieved higher translation accuracy and user satisfaction than Google Translate and UD Talk in Chinese-to-Japanese translations of real-world outpatient consultations. However, AI-based translation should be implemented as a complementary communication tool rather than a replacement for professional medical interpreters. Careful integration into clinical workflows and continued human oversight remain essential to ensure patient safety, communication accuracy, and equitable multilingual health care delivery.

Acknowledgments

This manuscript was edited by Wallace Academic Editing. During revision, generative artificial intelligence tools were used solely for language editing and grammar refinement. All scientific content, study design, data analysis and interpretation, and final manuscript approval were performed and verified by the authors.

Funding

This study was supported by Fu Jen Catholic University (grant CPL-202412007). The funding body played no role in study design; data collection, analysis, or interpretation; manuscript preparation; or the decision to submit the manuscript for publication.

Data Availability

The data supporting the findings of this study are included in the supplementary materials.

Authors' Contributions

KYC and CHH conceived and designed the study. CHH, WLL, CYC, YZC, AS, and CYL conducted the experimental work and collected the data. KYC, AYL, and YZC analyzed the data. KYC, CHH, WLL, and CYL drafted the manuscript. KYC and CYL critically revised the manuscript. All authors have approved the submitted version of the manuscript for publication.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Instrument for evaluating the medical interpreter–rated accuracy of the translations.
[DOCX File , 58 KB - jmir_v28i1e93082_app1.docx]

Multimedia Appendix 2
Instrument for evaluating the lay participant–rated satisfaction with the translations.
[DOCX File , 39 KB - jmir_v28i1e93082_app2.docx]

Multimedia Appendix 3
Subgroup analyses of translation accuracy and satisfaction across translation systems by outpatient specialty.
[DOCX File , 18 KB - jmir_v28i1e93082_app3.docx]

References

1. Al Shamsi H, Almutairi AG, Al Mashrafi S, Al Kalbani T. Implications of language barriers for healthcare: a systematic review. Oman Med J 2020 Apr 30;35(2):e122 [FREE Full text] [doi: 10.5001/omj.2020.40] [Medline: 32411417]
2. Hilfinger Messias DK, McDowell L, Estrada RD. Language interpreting as social justice work: perspectives of formal and informal healthcare interpreters. ANS Adv Nurs Sci 2009;32(2):128-143. [doi: 10.1097/ANS.0b013e3181a3af97] [Medline: 19461230]
3. Bischoff A, Denhaerynck K. What do language barriers cost? An exploratory study among asylum seekers in Switzerland. BMC Health Serv Res 2010 Aug 23;10:248 [FREE Full text] [doi: 10.1186/1472-6963-10-248] [Medline: 20731818]
4. Floyd A, Sakellariou D. Healthcare access for refugee women with limited literacy: layers of disadvantage. Int J Equity Health 2017 Nov 10;16(1):195. [doi: 10.1186/s12939-017-0694-8] [Medline: 29126420]
5. Divi C, Koss RG, Schmaltz SP, Loeb JM. Language proficiency and adverse events in US hospitals: a pilot study. Int J Qual Health Care 2007 Apr;19(2):60-67. [doi: 10.1093/intqhc/mzl069] [Medline: 17277013]
6. Squires A. Evidence-based approaches to breaking down language barriers. Nursing 2017 Sep;47(9):34-40 [FREE Full text] [doi: 10.1097/01.NURSE.0000522002.60278.ca] [Medline: 28799977]
7. International Medical Interpreters Association. URL: https://www.imiaweb.org [accessed 2026-02-05]
8. Flores G, Abreu M, Barone CP, Bachur R, Lin H. Errors of medical interpretation and their potential clinical consequences: a comparison of professional versus ad hoc versus no interpreters. Ann Emerg Med 2012 Nov;60(5):545-553. [doi: 10.1016/j.annemergmed.2012.01.025] [Medline: 22424655]
9. Dysart-Gale D. Communication models, professionalization, and the work of medical interpreters. Health Commun 2005;17(1):91-103. [doi: 10.1207/s15327027hc1701_6] [Medline: 15590344]
10. McDowell L, Hilfinger Messias DK, Estrada RD. The work of language interpretation in health care: complex, challenging, exhausting, and often invisible. J Transcult Nurs 2011 Apr;22(2):137-147. [doi: 10.1177/1043659610395773] [Medline: 21317402]
11. Gartner K, Mösko M, Becker JC, Hanft-Robert S. Barriers to use of interpreters in outpatient mental health care: exploring the attitudes of psychotherapists. Transcult Psychiatry 2024 Apr;61(2):285-297 [FREE Full text] [doi: 10.1177/13634615241227337] [Medline: 38298011]
12. Tang D, Jawad D, Dragoje V, Wen LM, Taki S. The use of interpreter services and its barriers faced by hospital staff when accessing interpreters for patients with low English proficiency during the COVID-19 pandemic. Health Promot J Austr 2024 Oct;35(4):1184-1193. [doi: 10.1002/hpja.850] [Medline: 38402882]
13. Patil S, Davies P. Use of Google Translate in medical communication: evaluation of accuracy. BMJ 2014 Dec 15;349:g7392 [FREE Full text] [doi: 10.1136/bmj.g7392] [Medline: 25512386]
14. Taira BR, Kreger V, Orue A, Diamond LC. A pragmatic assessment of Google Translate for emergency department instructions. J Gen Intern Med 2021 Nov;36(11):3361-3365 [FREE Full text] [doi: 10.1007/s11606-021-06666-z] [Medline: 33674922]
15. Khoong EC, Steinbrook E, Brown C, Fernandez A. Assessing the use of Google Translate for Spanish and Chinese translations of emergency department discharge instructions. JAMA Intern Med 2019 Apr 01;179(4):580-582 [FREE Full text] [doi: 10.1001/jamainternmed.2018.7653] [Medline: 30801626]
16. Chen X, Acosta S, Barry AE. Evaluating the accuracy of Google Translate for diabetes education material. JMIR Diabetes 2016 Jun 28;1(1):e3 [FREE Full text] [doi: 10.2196/diabetes.5848] [Medline: 30291056]
17. Cornelison BR, Al-Mohaish S, Sun Y, Edwards CJ. Accuracy of Google Translate in translating the directions and counseling points for top-selling drugs from English to Arabic, Chinese, and Spanish. Am J Health Syst Pharm 2021 Nov 09;78(22):2053-2058. [doi: 10.1093/ajhp/zxab224] [Medline: 34048533]
18. UD Talk. URL: https://udtalk.jp/ [accessed 2025-10-04]
19. Haleem A, Javaid M, Khan IH. Current status and applications of artificial intelligence (AI) in medical field: an overview. Curr Med Res Pract 2019;9(6):231-237. [doi: 10.1016/j.cmrp.2019.11.005]

XSL•FO
RenderX

20. Haleem A, Vaishya R, Javaid M, Khan IH. Artificial intelligence (AI) applications in orthopaedics: an innovative technology to embrace. J Clin Orthop Trauma 2020 Feb;11(Suppl 1):S80-S81 [FREE Full text] [doi: 10.1016/j.jcot.2019.06.012] [Medline: 31992923]
21. Jha S, Topol EJ. Information and artificial intelligence. J Am Coll Radiol 2018 Mar;15(3 Pt B):509-511. [doi: 10.1016/j.jacr.2017.12.025] [Medline: 29398501]
22. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb 09;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
23. Boylen S, Cherian S, Gill FJ, Leslie GD, Wilson S. Impact of professional interpreters on outcomes for hospitalized children from migrant and refugee families with limited English proficiency: a systematic review. JBI Evid Synth 2020 Jul;18(7):1360-1388. [doi: 10.11124/JBISRIR-D-19-00300] [Medline: 32813387]
24. Flores G. The impact of medical interpreter services on the quality of health care: a systematic review. Med Care Res Rev 2005 Jun;62(3):255-299. [doi: 10.1177/1077558705275416] [Medline: 15894705]
25. Karliner LS, Jacobs EA, Chen AH, Mutha S. Do professional interpreters improve clinical care for patients with limited English proficiency? A systematic review of the literature. Health Serv Res 2007 Apr;42(2):727-754 [FREE Full text] [doi: 10.1111/j.1475-6773.2006.00629.x] [Medline: 17362215]
26. Heath M, Hvass AM, Wejse CM. Interpreter services and effect on healthcare - a systematic review of the impact of different types of interpreters on patient outcome. J Migr Health 2023 Jan 24;7:100162 [FREE Full text] [doi: 10.1016/j.jmh.2023.100162] [Medline: 36816444]
27. Rao P, McGee LM, Seideman CA. A comparative assessment of ChatGPT vs. Google Translate for the translation of patient instructions. J Med Artif Intell 2024 Jun 30;7:11. [doi: 10.21037/jmai-24-24]
28. Kong M, Fernandez A, Bains J, Milisavljevic A, Brooks KC, Shanmugam A, et al. Evaluation of the accuracy and safety of machine translation of patient-specific discharge instructions: a comparative analysis. BMJ Qual Saf 2026 Feb 19;35(3):150-158 [FREE Full text] [doi: 10.1136/bmjqs-2024-018384] [Medline: 40633961]
29. Kwan M, Jeemi Z, Norman R, Dantas JA. Professional interpreter services and the impact on hospital care outcomes: an integrative review of literature. Int J Environ Res Public Health 2023 Mar 15;20(6):5165 [FREE Full text] [doi: 10.3390/ijerph20065165] [Medline: 36982073]
30. Baurer D, Yonek JC, Cohen AB, Restuccia JD, Hasnain-Wynia R. System-level factors affecting clinicians' perceptions and use of interpreter services in California public hospitals. J Immigr Minor Health 2014 Apr;16(2):211-217. [doi: 10.1007/s10903-012-9722-3] [Medline: 23054539]
31. Mayo R, Parker VG, Sherrill WW, Coltman K, Hudson MF, Nichols CM, et al. Cutting corners: provider perceptions of interpretation services and factors related to use of an ad hoc interpreter. Hisp Health Care Int 2016 Jun;14(2):73-80. [doi: 10.1177/1540415316646097] [Medline: 27257220]
32. Lundin C, Hadziabdic E, Hjelm K. Language interpretation conditions and boundaries in multilingual and multicultural emergency healthcare. BMC Int Health Hum Rights 2018 Jun 05;18(1):23 [FREE Full text] [doi: 10.1186/s12914-018-0157-3] [Medline: 29866163]
33. Nápoles AM, Santoyo-Olsson J, Karliner L, Gregorich S, Pérez-Stable EJ. Inaccurate language interpretation and its clinical significance in the medical encounters of Spanish-speaking Latinos. Med Care 2015 Nov;53(11):940-947 [FREE Full text] [doi: 10.1097/MLR.0000000000000422] [Medline: 26465121]
34. van Kolfschooten H, Goosen S, van Oirschot J, Schouten B, Vajda I, Willems L. Legal, ethical, and policy challenges of artificial intelligence translation tools in healthcare. Discov Public Health 2025;22(1):904. [doi: 10.1186/s12982-025-01277-z] [Medline: 41487681]

Abbreviations

AI: artificial intelligence
LLM: large language model

Edited by I Steenstra; submitted 12.Feb.2026; peer-reviewed by H Thongdaeng, GJ Noordergraaf; comments to author 14.May.2026; revised version received 02.Jun.2026; accepted 02.Jun.2026; published 18.Jun.2026.

Please cite as:
Chao KY, Huang CH, Liu WL, Chen CY, Lim AY, Chu YZ, Sasaoka A, Lin CY
Comparative Evaluation of ChatGPT, Google Translate, and UD Talk for Chinese-to-Japanese Translation in Cardiology and Pulmonology Outpatient Consultations: Prospective Observational Study
J Med Internet Res 2026;28:e93082
URL: https://www.jmir.org/2026/1/e93082
doi:10.2196/93082
PMID:

XSL•FO
RenderX

©Ke-Yun Chao, Chia-Hui Huang, Wei-Lun Liu, Chao-Yu Chen, Ai Yin Lim, Yi-Zhen Chu, Atsuko Sasaoka, Chung-Yu Lin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Chatbot to Meet Parents' Information Needs for Sickle Cell Trait Newborn Screening Results: Multiple Methods Formative Study

Anne C Madeo[1], MS; Courtney Gauchel[2], MS; Emerson Borsato[2], PhD; Amy Gaviglio[3], MS; Kimberly A Kaphingst[4], ScD; Guilherme Del Fiol[2], MD, PhD; Karen Eilbeck[2], PhD

[1]Department of Population Health Sciences, University of Utah, Salt Lake City, UT, United States

[2]Department of Biomedical Informatics, University of Utah, 421 Wakara Way, Salt Lake City, UT, United States

[3]Connetics Consulting LLC, Minneapolis, MN, United States

[4]Department of Communication, University of Utah, Salt Lake City, UT, United States

Corresponding Author:
Karen Eilbeck, PhD
Department of Biomedical Informatics, University of Utah, 421 Wakara Way, Salt Lake City, UT, United States

Abstract

Background: Newborn screening (NBS), a mandated public health intervention, allows the identification of babies with potentially life-threatening disorders and facilitates disease diagnosis and management before the onset of symptoms. While NBS saves lives, the process can be fraught with anxiety and unanswered questions from parents or guardians of newborns, especially as they wait for an appointment with a clinician.

Objective: This study aimed to describe the development and testing of an educational chatbot (NBSchat) to address the emotional support and information needs of parents of newborns identified with sickle cell trait via NBS.

Methods: NBSchat, a fully scripted (ie, rule-based) chatbot, was developed by a multidisciplinary team and evaluated through a sequential multiple methods study, including interviews and a survey. To inform chatbot design, we conducted semistructured interviews with 11 adults—5 clinicians who work with parents of infants identified with sickle cell trait through NBS and 6 parents of infants aged 12 months or less—using the critical incident technique and think-aloud tasks while using a prototype of NBSchat. Transcripts underwent thematic analysis. In a survey, 250 parents of infants aged 12 months or less without abnormal NBS results were shown a mock newborn screening result letter and then interacted with NBSchat, after which they self-reported emotional and attitudinal outcomes before and after the simulated exposure.

Results: Feedback from interviews confirmed that parents are distressed by trait results and actively seek information and reassurance. Thematic analysis indicated that NBSchat provided reliable, accurate information that parents wanted and had the potential to reduce negative emotions (eg, provide relief and reduce stress). Key strengths included addressing an immediate health concern and offering reassurance. The results of the postintervention survey indicated that, compared to pre-exposure scores, participants reported significantly lower negative emotions (mean 7.0 [SD 3.2] vs 5.8 [SD 3.2] out of 12; mean difference −1.2, 95% CI −1.57 to −0.83; $P<.001$), improved positive emotions (reflected by a decrease in the reverse-coded positive emotion score; mean 8.6 [SD 4.2] vs 7.8 [SD 4] out of 16; mean difference −0.8, 95% CI −1.27 to −0.37; $P<.001$), and reduced uncertainty (mean 6.5 [SD 3] vs 5.5 [SD 3.4] out of 12; mean difference −1, 95% CI −1.42 to −0.58; $P<.001$). Parents noted that NBSchat provided immediate reassurance and was convenient to access. They further reported that the predefined, structured questions in the script helped guide their learning and understanding.

Conclusions: Overall, participants who interacted with NBSchat found it to be acceptable, with improved emotional measures after its use. Future research will investigate the outcomes of using the chatbot and its implementation in a pragmatic randomized controlled trial.

(J Med Internet Res 2026;28:e86022) doi:10.2196/86022

KEYWORDS

sickle cell trait; patient care team; neonatal screening; public health; parents; infant; newborn; digital health; conversational agent; chatbot

Introduction

Newborn screening (NBS) identifies serious, treatable conditions in infants before symptoms appear, enabling timely intervention and improved outcomes. In the United States, NBS is administered by state public health programs [1] and is routinely performed within the first 24 to 48 hours of life, unless parents opt out [2]. Most results are normal and reported to the birth

XSL•FO
RenderX

hospital or pediatrician, who informs families only if follow-up is needed. For urgent conditions, the laboratory calls the health care provider directly to expedite care. Nonurgent findings, such as sickle cell trait (SCT), are often mailed to both the parents and the infant's clinician, although processes vary by state [3]. Trait status means the infant is healthy but carries one altered gene that can be passed to future children. Health departments commonly use follow-up teams to communicate results and direct families to next steps, which typically involve the primary care provider, a genetic counselor, or a specialist.

Adverse psychosocial outcomes among parents receiving positive NBS results have long been recognized, particularly in conditions such as cystic fibrosis and sickle cell disease. Early studies highlighted anxiety linked to poor understanding of NBS results [4], and a 2004 systematic review of prenatal and newborn screening underscored this concern. More recent work shows that parents of infants with positive NBS results have higher odds of depression (OR 6.06) and stress (OR 3.2) [5]. Prior research indicates that while the initial shock of the unexpected genetic information is unavoidable, other factors that contribute to distress can be addressed through education and improved communication. These factors include filling knowledge gaps about the condition, genetics, and the next medical steps. For example, autosomal disease inheritance patterns are often confusing concepts to parents [6,7] and can be addressed by education and counseling [8,9]. A 2020 systematic review of 92 studies concluded that unexpected NBS and unsolicited genetic findings continue to cause significant distress, but that risks can be reduced through comprehensive, multimodal parent education tailored to timing and preferences [10]. Quantitative studies similarly show that well-informed parents experience less long-term anxiety, yet the uptake of genetic counseling, which has been shown to reduce guilt and anxiety, remains low, with only one-third of eligible families receiving it [11]. Parents frequently struggle to meet their informational and emotional needs as they move through the process from screening to confirmatory testing and new clinical care. Meanwhile, limited genetic counseling capacity creates a bottleneck [12,13], and state NBS follow-up teams often spend substantial time addressing parental anxiety about trait findings, such as SCT, before families connect with primary care providers or genetic counselors [14]. Trait results for hemoglobinopathies such as SCT constitute the largest portion of returned NBS results [15,16]. Both recent and historical literature emphasize that genetic counseling is central to returning SCT results, as it reduces distress, clarifies that SCT is not a disease, and provides essential education about inheritance and reproductive implications [14,17,18]. Access to genetic counseling services can be uneven due to workforce shortages [19] and regional variation in genetic counselor concentration [20].

Digital health tools, such as conversational agents ("chatbots"), have the potential to address challenges in the delivery of NBS results efficiently and at scale and provide consistency across content shared [21]. Chatbots are computer programs that simulate human conversations by using predetermined rule–based responses or artificial intelligence algorithms. Their use in health care has been increasing and includes reminding patients to take medications [22], offering genetic testing [23], and providing brief mental health therapy [24]. User evaluation of digital health interventions, at an early stage of the development process, is essential to ensure that the resulting tool is acceptable and useful to the target population. Chatbots have not previously been used to educate parents about NBS or SCT. This education is typically provided through handouts or in-person instruction. Incorporating a chatbot into a state's NBS workflow may improve knowledge among individuals who are less likely to accept or have access to genetic counseling. The objectives of this study were to (1) describe the development of a chatbot (NBSchat) that provides information and emotional support to parents who receive an NBS screen-positive result for SCT, (2) explore users' and clinicians' experiences and seek feedback to help guide and refine the design of NBSchat content and usability, and (3) compare parent participants' emotional and cognitive outcomes before and after using NBSchat.

Methods

Study Design

Overall, this was an exploratory sequential multiple methods study consisting of semistructured interviews using the critical incident method and a quantitative survey where participants were asked to rate emotional and cognitive outcomes before and after using NBSchat.

Ethical Considerations

The study was reviewed by the University of Utah Institutional Review Board (IRB), which deemed it exempt (IRB_00179621). Interview participants were provided with a concise study summary with a consent cover letter prior to their interviews and were compensated for taking part with US $50 via a secure online funds transfer system. Verbal consent was obtained at the beginning of each interview. Participants were informed that a transcript of their audio recording would be prepared and deidentified for the study. In accordance with the institutional review board protocol, participants were also informed that the audio files and transcripts would be destroyed after the study. Survey participants were provided with a consent cover letter prior to beginning the survey, and those who completed the survey were compensated with US $14.75 for their participation in accordance with the policies of the survey panel provider. Participants were informed that their responses would be kept confidential, that no identifying information would be disclosed, and that results would be reported in aggregate. Chatbot logs contain timestamps and paths through the conversation for each user; they do not contain identifying information.

Initial NBSchat Design

To ensure that we developed a robust, patient-focused technology that is tailored to parents' needs, we followed a participatory approach. Guided by principles of health communication [25], an interdisciplinary team that included genetic counselors, informaticists, a labor and delivery nurse, an NBS expert, and a communication scientist developed an initial draft of the chatbot script [26]. In the initial draft, we sought to address information gaps that cause preventable

XSL•FO
RenderX

anxiety: not understanding the implications of the result, wondering about the status of other family members and reproductive health, and navigating the next steps in the health care process. The script was then iteratively refined based on the group's feedback through meetings and offline discussions.

NBSchat was developed using the GARDE-Chat [27] platform developed by investigators at the University of Utah. GARDE-Chat includes a chatbot authoring tool that enables the creation and management of multiple rule-based, large language model (LLM)–based, or hybrid chatbots. LLM-based chatbots, such as ChatGPT, are known to occasionally provide inaccurate information (also known as "hallucinations" [28]) and answer questions that fall outside the predefined boundaries of a chatbot. To avoid these issues, we chose to use a rule-based approach with a predefined conversation script and a fixed set of possible questions and responses that provide accurate and consistent information to parents (Figure 1).

Figure 1. The flow of NBSchat. All patients navigate to the topic menu, after which the topics are optional. The example conversation shown addresses the question, "What is sickle cell trait?" The screenshots shown are simulated example screens for demonstration purposes only and do not depict a real patient or actual participant conversation.

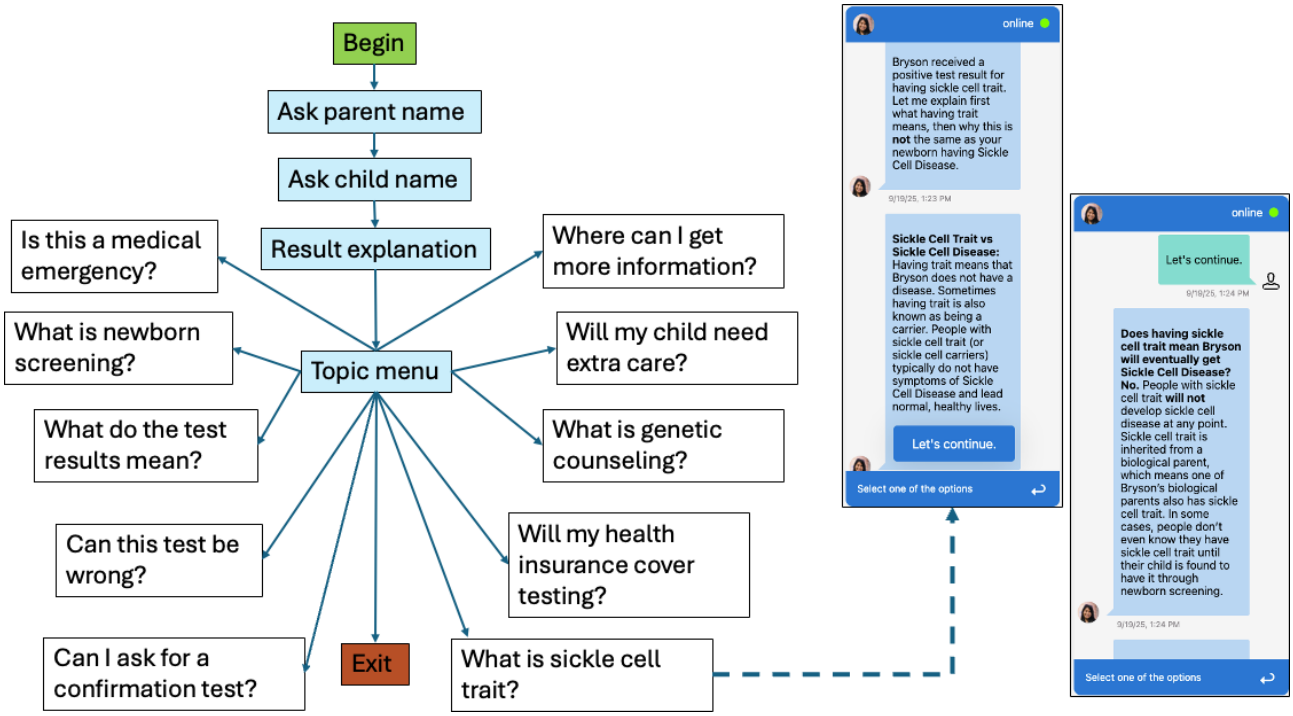

Interviews

Study Participants

To evaluate the content and flow of NBSchat, we conducted in-depth interviews via Zoom from December 10, 2024, to February 7, 2025 [29]. During the study timeframe, we were unable to recruit additional parents of children with SCT. As a result, participants included 6 adult parents of children less than 12 months old (1 parent of a newborn identified with SCT through NBS and 5 parents of newborns who had not received a positive NBS result) and 5 genetic counselors who speak with parents whose newborns were identified as having SCT through NBS.

Participants were recruited through various means. Clinicians were recruited through emails sent to members of the National Society of Genetic Counselors who indicated that their specialization included "Public Health/Newborn Screening." Parents of newborns were recruited through the University of Utah Office of Research Participant Advocacy. In addition, parents of newborns with SCT were identified through emails seeking partnerships with state chapters of the Sickle Cell Disease Association of America, postings on social media pages, and advertisements at clinics that provide care to infants with SCT. Participants contacted the team via a project-specific email address and were screened via a set of 9 questions to confirm they met eligibility criteria. We invited participants to refer individuals who they thought would be eligible and interested in participating.

Interview Methods

Interview guides were tailored to each eligibility group and used the critical incident technique [30] and a think-aloud task [31]. The critical incident technique was selected to allow the participants to be as specific as possible and provide rich data in their own words [30,32]. For parents of an infant identified with SCT, the critical incident was defined as the moment they first received their child's NBS result. Parents of a newborn who did not receive a positive screen for any condition were provided an anonymized positive newborn screen letter and asked to imagine that they had received the letter 2 weeks after their infant was born. Clinicians were asked to describe a specific instance when they discussed with a parent about their child's SCT NBS screen-positive result.

To assess the usability of NBSchat, after describing the critical incident, each participant was asked to load the chatbot on the device they used to connect to the interview, share their screen, and describe their thoughts and reactions ("think-aloud") while

XSL•FO
RenderX

they navigated NBSchat and looked for information that they thought would be important to a parent of a newborn recently identified with SCT. After a participant completed this think-aloud task, interviewers probed to ascertain issues pertaining to NBSchat's content and navigation. We chose not to undertake iterative changes to the chatbot design during the interviews so that we could assess whether certain elements of NBSchat were more or less appealing to both groups of participants. At the close of the interview, all participants were asked sociodemographic questions.

Interview Data Analysis

All interviews were audio-recorded and transcribed verbatim by a professional transcription service. We aimed to produce actionable, accurate insights that could be incorporated into the design of NBSchat; therefore, qualitative data collected through the interviews were analyzed using the Rapid and Rigorous Qualitative Data Analysis technique [33,34]. The research team developed a phase 1 matrix in Microsoft Excel that reflected the topics assessed in the interview guide and their preliminary codes, interpreting the data [35]. All members of the research team read the phase 1 matrix and discussed the goals of the analysis. Based on these discussions, authors CG and ACM further reduced the data in the phase 1 matrix and created a phase 2 matrix that provided actionable feedback with supporting quotes addressing the overarching research question. The phase 2 matrix incorporated analysts' "focused codes [36]." Data reduction and condensation occurred iteratively until the team agreed that the final phase table addressed our overarching research question. Disagreements were resolved through discussion.

Survey

Participants

The survey was deployed by Prodege to a sample of 330 participants, who were similar to the United States' demographics in terms of race, ethnicity, and education level. Eligibility criteria included being 18 years or older, being the parent of a child 12 months old or younger, and not having a child identified with a health condition through NBS. The survey was conducted from March 13, 2025, to May 1, 2025. Of the 277 participants who completed the survey, 250 passed the attention checks and successfully launched NBSchat.

Measures

Demographics

Demographic data, including accessibility to and use of mobile technology [37], were collected.

Health Literacy and Genetic Knowledge

The survey assessed participants' understanding of genetic test terminology using items adapted from *Measuring Genetic Knowledge: A Brief Survey Instrument for Adolescents and Adults* [38]. This is a 7-question instrument using true or false items. We utilized 6 of the 7 items from this instrument. In addition, we assessed participants' general health literacy using the BRIEF Health Literacy Screening Tool [39,40]. This is a 4-question survey measured using a 5-point Likert scale with a range of 4 to 20.

Self-Efficacy and Emotional Responses

Communication self-efficacy was assessed using the Ask, Understand, Remember (AURA) survey [41], a 4-item measure administered on a 4-point Likert scale. Items were adapted to be about the child's case. Psychosocial distress was assessed with 10 items from 3 subscales (negative feelings, positive feelings, and uncertainty) derived from the Feelings About Genomic Testing Results (FACToR). All items were measured on a 5-point Likert scale [42]. These measures were assessed both before and after using NBSchat. The positive emotion items were reverse-coded, so lower scores corresponded to reduced negative affect and greater positive emotion. Items were adapted from the original to reflect the screening result.

Usability and Usefulness

NBSchat usability was assessed using the System Usability Scale (SUS) [43], a 10-item questionnaire with a 5-point Likert scale, where scores are converted to a 0 to 100 usability score. Items were adapted to reflect the software being a chatbot, and open-ended comments were also collected.

Procedure

A survey instrument was developed using the Qualtrics platform and incorporated 3 attention check items to assess response quality [44-46]. Participants who failed one or more attention checks were removed from the final analysis to minimize the impact of insincere survey takers. Analyses were conducted using complete cases; sample sizes varied slightly by analysis due to item nonresponse. Participant knowledge of genetics and health literacy was assessed; then, participants were presented with a simulated NBS result indicating that a newborn had screened positive for SCT and were asked to complete a modified version of the AURA and FACToR instruments to capture baseline self-efficacy and emotional responses. Following this, participants interacted with NBSchat and were then asked to complete the AURA and FACToR questions to evaluate changes in outcomes. Participants then provided usability and usefulness feedback, followed by demographic information. Given the exploratory mixed methods design, recruitment targets were inflated to account for an anticipated incomplete rate of approximately 30%, with the goal of obtaining at least 100 open-ended responses to support robust and diverse qualitative perspectives and a final analytic sample sufficient to detect small pre-post changes in AURA and the FACToR subscale scores. Missing data were minimal (1 of 250 participants missed one item on a multi-item scale;<0.5%). Scale scores were computed as the mean of completed items. Given the negligible level of missingness, no additional imputation procedures were conducted.

Survey Data Analysis

Survey data were analyzed using Microsoft Excel. Descriptive data, including means and standard deviations, were obtained for all measures. The significance of differences in measures before and after using NBSchat was determined using the paired *t* test, with statistical significance assessed as a $P<.05$. Open-ended comments were coded into 3 categories: positive comment, negative comment, and neutral comment during a content analysis. After reading all the comments, 2 reviewers

XSL•FO
RenderX

agreed upon codes and selected representative quotes. The integration of survey with interview data occurred by comparing themes identified in interviews with patterns observed in survey responses.

Results

Interviews

Overview

Analysis of interview transcripts revealed 4 overarching categories of themes that capture parents' experiences and needs following an NBS screen-positive result for SCT (carrier status). First, parents rely on multiple sources of information, including online searches, health care providers, phone hotlines, and personal networks. Second, parents aimed to reduce negative emotions and clarify immediate health concerns via information seeking. Third, parents and clinicians felt that NBSchat may help reduce negative emotions and address immediate health concerns, emphasizing its potential to reduce worry, provide actionable information, and offer reassurance. Finally, NBSchat was described as a tool that could bridge structural and communication gaps in care—including language barriers, geographic distance to clinics, long wait times for appointments, and challenges related to health literacy. Interviews were conducted until thematic saturation was reached, defined as the point at which no new themes emerged in subsequent interviews. Together, these themes highlight the barriers parents encounter when navigating the NBS system and the opportunities for digital tools to provide timely, understandable, and supportive information. Exemplar quotes for each theme identified in the interviews are collated in Table S1 (Multimedia Appendix 1) and are described below.

Theme 1. Parents Relied on Multiple Sources to Meet Their NBS Information Needs

A discussion of the sources of information parents currently use when navigating the NBS landscape revealed 4 key codes. Health care providers and parents indicated that the internet was a popular avenue for parents trying to understand the results. Clinical providers, such as the child's pediatrician, are a valuable resource for parents; however, there can be a disconnect between the state NBS follow-up team and the pediatrician's office regarding the interpretation of results and the transfer of educational materials. For the parents' interviews, after being shown a standard trait result letter, many participants would call the public health lab for the clarification of the result's meaning. Finally, parents rely on a network of family and friends whom they also turn to for support.

> *I know we shouldn't Google, but then I would maybe just look over, see what exactly the baby had.* [Parent-007]
>
> *…the parent received that letter at home, which had our number, and then they would call us there saying, "What does this mean?"* [Clinician-006]
>
> *If I had a friend or family member who had a baby with a genetic thing pop up, I would reach out to them.* [Parent-005]

Theme 2. Parents Aimed to Reduce Negative Emotions and Clarify Immediate Health Concerns

When parents received SCT results, 2 major themes consistently emerged: parents looked for reassurance to reduce negative emotions and clarification of immediate health concerns. Clinicians described how initial fear and confusion often gave way to relief once parents were reassured that SCT is not a disease and does not pose a long-term health risk. Parents expressed gratitude when told directly that their baby was healthy, and clinicians observed that this reassurance helped alleviate distress. At the same time, parents and clinicians reported a strong desire for practical information to address immediate health concerns. Questions such as whether SCT could make a child sick, what symptoms to watch for, whether diet or treatment changes were necessary, and how SCT differed from sickle cell disease emerged. These findings underscore that both emotional support and clear, actionable guidance are essential for parents to understand and cope with SCT results.

> *Whenever I say that [results are most relevant for parents' future family-planning], that's when the relief comes out. I would say that people are reassured by the information that's given.* [Clinician-004]
>
> *…it [NBS trait results] kind of freaked the parents out and they were like, "Oh, my gosh. What's going on with our child?" Whereas when they came into the office and we actually sat down and looked at it, like baby's fine.* [Clinician-005]
>
> *I think that the first thing I would want to know is, what is sickle cell trait?* [P-007, parent of infant without NBS experience]

Theme 3. Parents and Clinicians Felt That NBSchat May Help Reduce Negative Emotions and Address Immediate Health Concerns

After using the NBSchat prototype in the interview setting, both clinicians and parents felt that it would reduce negative emotions and help address immediate health concerns. One of the parents discussed the stress of not being able to reach the right person when calling a clinician, a situation in which a chatbot could provide immediate assistance. The parents appreciated the immediate feedback indicating that their child did not have the disease, and their questions could be quickly addressed. The menu of questions was seen positively as it provides a pathway to obtain new information.

> *Sometimes it's frustrating also to call phone numbers, especially in this situation, because you don't go, sometimes, immediately to the person that you wanna talk to...Then it will be stressful. I already have so much going on.* [Parent-007]
>
> *I think it [additional information] does [provide reassurance]. Even the families where they're very distressed.* [Clinician-004]
>
> *I liked that it told me right off the bat the results and everything and that it's not an emergency and you should talk to your doctor about it. It gave you instructions.* [Parent-001]

XSL•FO
RenderX

Theme 4. Bridge Structural and Communication Gaps in Care

Clinicians and parents agree that language is a prominent issue when relaying trait results. Genetic counselors are concerned about leaving nonnative English speakers confused. Issues with both rural considerations and wait times for a genetic counselor appointment were brought up by clinicians, which highlights the dearth of qualified providers in some parts of the country. Ultimately, the chatbot was viewed as a bridge to parents with limited health literacy. Again, the menu was positively perceived by those with low health literacy to provide a framework to structure information and allow choice in what information to receive.

> *I'm glad that it was a menu 'cause then you guide me on how the terminology goes, and then you give me exactly what I ask for rather than me trying to tell you what I'm asking.* [Parent-005]

> *Then about I'd say probably five, six months is when the patients actually will be scheduled, so they've got a very long wait time.* [Clinician-005]

> *I'd say the vast majority prefer or request the phone genetic counseling and that is one reason for that could be that our state is quite rural and the genetic counselors that do this are only located in one city in the state. To physically come here is often very far.* [Clinician-002]

> *[F]or a result like sickle cell trait, I think getting the information across in their language is maybe the most challenging part to make sure that makes sense.* [Clinician-002]

Survey

Demographic Characteristics

Participant demographics are outlined in Table 1. The majority of participants were White (n=179, 71.6%), women (n=178, 71.2%), and non-Hispanic (n=214, 85.6%). Educational attainment was equally distributed across high school (n=77), associate's (n=88), and college levels (n=77, 30.8% - 35.2%), with junior high represented at 3.2% (n=8). The majority of respondents reported internet availability at home (n=233, 93.2%), with cellphone ownership at 99.6% (n=249), and all cellphones were reported as smartphones.

XSL•FO
RenderX

Table . Survey population characteristics (N=250).

Characteristic	Participants, n (%)
Race	
American Indian or Alaska Native	4 (1.6)
Asian or Asian American	14 (5.6)
Black or African American	34 (13.6)
Native Hawaiian or Pacific Islander	1 (0.4)
White	179 (71.6)
Some other ancestry or origin	6 (2.4)
Mixed ancestry	12 (4.8)
Hispanic, Latino, or Spanish origin	
Yes	36 (14.4)
No	214 (85.6)
Gender	
Woman	178 (71.2)
Man	72 (28.8)
Some other way	0 (0)
Education level	
Junior high	8 (3.2)
High school	77 (30.8)
Some college or associate's degree	88 (35.2)
College	77 (30.8)
Current subscription to an internet service at home	
Yes	233 (93.2)
No	17 (6.8)
Type of home internet service	
Dialup	1 (0.4)
High-speed broadband like cable, fiber optic, wireless router, satellite, or digital subscriber line (DSL)	227 (90.8)
Not sure	5 (2)
Blank	17 (6.8)
Cell phone ownership	
Yes	249 (99.6)
No	1 (0.4)
Type of cell phone	
Smartphone	249 (99.6)

Health Literacy and Genetic Knowledge

The BRIEF Health Literacy Screening items were measured on a 5-point Likert scale, with a range of 4 to 20, where 20 indicates the highest level of literacy. The sample had a high health literacy average, but a wide range (mean 16.3, SD 3.6). For the genetic knowledge questions, participants correctly answered a mean of 5.72 out of 6 items (SD 0.6; range 0-6). The highest proportion of correct responses (n=241, 96.4%) was for the item “All serious diseases are inherited (false),” while the lowest (n=235, 94%) was for “Healthy parents can have a child with an inherited disease (true).”

Self-Efficacy and Emotional Responses

There was no significant difference in communication self-efficacy before versus after using NBSchat (13.5 out of 16 vs 13.4; P=.42). FACToR subscales are shown in Table 2. Negative feelings were significantly lower after exposure to NBSchat (7.0 out of 12 vs 5.8; P<.001). Positive feelings also improved following exposure to NBSchat, as reflected by a

XSL•FO
RenderX

decrease in the reverse-coded positive emotion score (8.6 vs 7.8; $P<.001$). Finally, a small but significant decrease in uncertainty was observed (6.5 vs 5.5; $P<.001$). Statistical significance was set at $\alpha=.05$. As a sensitivity analysis, pre-post comparisons were also conducted using Wilcoxon signed-rank tests, which yielded the same pattern of results and conclusions.

Table . Feelings About Genomic Testing Results (FACToR) subscales for negative emotions, positive emotions, and uncertainty before and after exposure to NBSchat.

Subscale score	Before NBSchat, mean (SD)	After NBSchat, mean (SD)	Mean difference (95% CI)	Cohen *d*	*P* value
Negative emotions (0 - 12)	7 (3.2)	5.8 (3.2)	−1.2 (−1.57 to −0.83)	0.3881	<.001
Positive emotions (0 - 16)	8.6 (4.2)	7.8 (4)	−0.8 (−1.27 to −0.37)	0.2301	<.001
Reduced uncertainty (0 - 12)	6.5 (3)	5.5 (3.4)	−1 (−1.42 to −0.58)	0.3091	<.001

Usability and Usefulness

The mean SUS [43] score was 73 (SD 22), which falls above the commonly accepted threshold of 68, indicating good usability.

Participants addressed the prompt “Please write below what information you found useful in the chatbot.” There were 195 positive comments that described useful attributes of participants’ NBSchat experience. Recurring themes (Table 3) are that participants learned something new, found it reassuring, found it convenient for out-of-hours questions, and appreciated the menu of questions.

There were 30 negative comments that described NBSchat as not useful or listed a problem in its usage. We found instances where NBSchat was not using the correct name for the parent or child, an issue that was raised by participants in some of the interviews. Some participants found the response time too fast, while others found it slow, leading us to implement pauses in the NBSchat that allow the parent to better control the release of information.

Table . Content analysis of survey participants’ comments about the usefulness of NBSchat.

Theme	Representative quotes
Learned something new	• I thought that it was useful how the bot gave an in-depth explanation of the results and what they mean and explaining the difference between sickle cell trait and disease • Learning that just because someone has a trait doesn’t mean they have the disease.
Gained reassurance	• The reassurance that the trait doesn’t mean Sickle Cell Anemia. • …the chat answered my questions and honestly would have made me feel better if this [w]as a real situation.
Convenience	• Quickly find out that everything is okay without waiting until the morning, when the doctors’ working day begins. • As a worried parent, we would rather have immediate answers rather than waiting for a follow-up with a medical professional.
Menu of questions	• I like the predeveloped questions to ask further information. • I found it useful that it gave me prequestions to ask further because it could be hard at this moment to even know what to Google and seek out more information.

Discussion

Principal Findings

In this study, interviews and surveys with new parents demonstrated a clear information gap surrounding SCT results, which can lead to distress and uncertainty. Parents consistently sought reassurance and rapid answers to immediate health concerns. A key finding of the survey is the significant decrease in both distress (negative emotions) and uncertainty following the use of NBSchat. We also observed a small but significant increase in positive emotions. While some parents were content to find out that SCT is not the same as sickle cell disease, many parents reported taking a deeper dive into the menu of questions, with the most common theme being that they learned something new. The result explanation and the list of further resources were parts of NBSchat that filled an information gap. Overall, the study results suggest that NBSchat is a promising digital health tool for scalable and efficient delivery of NBS information to assist in the return of NBS results.

Because trait results represent the largest category of results returned by NBS programs, follow-up teams frequently receive calls from parents seeking clarification. The problem is

XSL•FO
RenderX

compounded by delays: results are often mailed, followed by a lag before discussion with the infant's clinician and an even longer wait for an appointment with a genetic counselor. The goal of NBSchat is not to replace genetic counseling but to supplement it by preparing parents with foundational knowledge and reducing anxiety so they arrive at appointments more informed [47], more confident, and able to ask targeted, personally relevant questions [48,49]. Given genetic counseling workforce shortages [19] and the limited availability of counselors in many regions [20], NBSchat also represents a practical intermediate solution to address gaps in parent communication and support, reflecting a focus on patient-centered care [50].

Previous research has identified significant variation in SCT disclosure to parents during the newborn period [17,21]. NBSchat offers standardized information to all parents, regardless of where screening was performed, and can be tailored to the unique needs of a state's NBS program. The clinicians in the study were concerned about the communication of trait results across language barriers; in some cases, even with an interpreter. This is an area where deploying a chatbot in the parents' first language could be very impactful.

The exploration of usability in the survey found the chatbot to be functional and user-friendly, with only minor areas for potential improvement. According to standard SUS interpretation guidelines, a score of 73 places the tool in the "Good" range and corresponds roughly to a B rating, reflecting positive user experience and acceptance.

The survey found that participants' communication self-efficacy was high before using NBSchat, and there was no significant change after using it. A mean score of 13.5 out of 16 on the AURA scale represents 84% of the maximum, indicating that participants had strong confidence in asking medical questions, with good perceived understanding and a high likelihood of remembering or using the information effectively. Participants reported feeling confident in engaging in health care discussions and seeking clarification. This likely reflects the sample of survey participants and may not represent the communication self-efficacy of parents who receive a screen-positive result for SCT.

NBSchat can be accessed via a web browser on any internet-connected device, such as a smartphone, tablet, or computer. In 2024, 99% of US adults aged 18 to 49 years reported using the internet [37], and the mean age at first birth in the United States was 27.4 in 2022 [51]; thus, it is likely that the majority of infants in the United States will be born to an adult who uses the internet, whether through a smartphone or other internet-connected device.

The alignment between interview themes and survey findings suggests that user experiences identified in the qualitative phase were not isolated but reflected broader user patterns. The results suggest that NBSchat is an effective and highly scalable approach to delivering patient-tailored education to address the unmet informational needs of parents and alleviate emotional distress. These findings align with previous studies, which provide strong evidence demonstrating the effectiveness of health service delivery and education through scripted chatbots in various health domains, with demonstrated benefits in lowering rates of depression and anxiety [24] and in the delivery of pre-test cancer genetic services and genetic testing [23].

Although recent health care chatbots have started to leverage artificial intelligence–based approaches using LLMs [52], rule-based chatbots are simpler to implement, safer (ie, the design team can ensure that the chatbot only provides accurate, compassionate, and culturally adapted information), constrained to only answer questions about the state's NBS program and the child's NBS result, and effective for goal-oriented use cases with a clearly defined end point, such as educating users about NBS results and processes [53]. Participants responded positively to having a menu of options to select for further information, and at least some were unsure what to ask without the structured question format, suggesting that they did not perceive a need for an artificial intelligence–based approach. Furthermore, in prior work using a chatbot to provide information about genetic services and genetic test results, few patients asked open-ended questions, indicating that a rule-based approach may have met the patients' informational needs [54]. Therefore, scripted chatbots, such as NBSchat, are, at least to this time, an adequate and safer approach for chatbot implementation, especially in sensitive topics such as the explanation of NBS results. This approach also enables each state's follow-up team to tailor the content unique to their program and update it as necessary through a distributed content authoring approach using the GARDE-Chat platform. Nonetheless, we acknowledge that a hybrid approach, using LLMs at end points in the script, could allow parents to explore more nuanced queries. Future work will explore adapting NBSchat to support free-text questions using a carefully constrained LLM approach, such as a retrieval-augmented generation framework grounded exclusively in authoritative NBS content and explicit scope boundaries to balance flexibility with safety.

Limitations of this study include potential selection bias in the parent interviews and survey toward individuals with higher health literacy and genetic knowledge. Participants in the survey had high genetic knowledge on average, with little variation in scores. We also found that 93.2% (n=233) of the participants had home internet service and 99.6% (n=249) had a smartphone, showing a high level of digital access. Additionally, using attention checks may introduce bias toward higher digital literacy respondents, although we minimized this risk by using plain and unambiguous language. Participants could not be blinded to NBSchat usage, which may have changed their information-seeking behavior because they knew they were being "treated." The survey relied on self-reported measures, which may be affected by social desirability bias. The simulation of delivering hypothetical test results may not reflect real-world outcomes.

While the survey recruited participants from across the United States and attempted to recruit a representative sample across education levels, the study population reflected relatively high literacy and consistent access to digital technologies. Consequently, participants were not proportionally representative of the ancestral and ethnic groups that bear a disproportionate burden of SCT and disease. Future evaluations

XSL•FO
RenderX

of NBSchat should focus on recruitment approaches that better reach these populations, including partnering with newborn screening programs, safety-net health systems, and community-based clinics serving predominantly affected families. Future studies testing NBSchat within these settings and populations are needed to more accurately assess its acceptability and effectiveness among those most affected by SCT. The clinical trial would also demonstrate the generalizability and modularity of NBSchat by adapting it to the particular workflow of a state follow-up team.

The findings should also be interpreted in light of important constraints, including the absence of a control group and the use of hypothetical screening scenarios. Accordingly, the observed pre-post changes in emotional measures cannot be interpreted as causal effects and instead represent preliminary signals supporting the feasibility and potential value of NBSchat, which warrant confirmation in controlled, real-world studies.

In this proof-of-concept demonstration, NBSchat was developed in English. Future iterations of NBSchat are planned to include additional language options, informed by community needs and implementation considerations, to enhance accessibility for linguistically diverse families and ensure the delivery of newborn screening information to all populations.

Conclusion

We developed and tested NBSchat, a prototype chatbot to support parents who receive an NBS SCT screen-positive result before speaking with a clinician. Both parents and clinicians reported needing to search multiple sources to understand NBS, suggesting that NBSchat could fill this gap by delivering clear, reliable information when results are received. In a simulated evaluation, parents using NBSchat reported reduced negative emotions and greater reassurance, addressing the distress and uncertainty often experienced while awaiting clinical follow-up. These findings represent preliminary signals of potential benefit and support further evaluation of NBSchat in controlled, real-world clinical settings.

Acknowledgments

The authors are grateful to Joni Pierce, PhD, for sharing her chatbot development processes.

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading, editing, and reformatting. The GAI tool used was ChatGPT4.5. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This study was funded by a seed grant from the University of Utah Margolis Foundation and in part funded by grant U24CA274582 from the US National Cancer Institute. CG was funded as a fellow of the Association of Public Health Laboratories.

Data Availability

Access to the GARDE-Chat open-source software is available from [55]. NBSchat is available from [56]. The datasets generated or analyzed during this study are not publicly available due to institutional review board considerations.

Authors' Contributions

KE, GDF, and KAK led the conceptualization and development of the study and contributed to methodology design. ACM, CG, and KE conducted the formal analysis, while ACM and CG carried out the investigation. KE acquired funding and oversaw project administration. CG developed the chatbot, and EB developed the software. ACM drafted the original manuscript. KE, GDF, KAK, ACM, CG, and AG critically reviewed and edited the manuscript at both prepublication and postpublication stages.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Exemplar quotes for each theme identified in the interviews.
[DOCX File, 20 KB - jmir_v28i1e86022_app1.docx]

References

1. Brosco JP, Grosse SD, Ross LF. Universal state newborn screening programs can reduce health disparities. JAMA Pediatr 2015 Jan;169(1):7-8. [doi: 10.1001/jamapediatrics.2014.2465] [Medline: 25402722]
2. Tarini BA, Goldenberg AJ. Ethical issues with newborn screening in the genomics era. Annu Rev Genom Hum Genet 2012 Sep 22;13(1):381-393. [doi: 10.1146/annurev-genom-090711-163741]

XSL•FO
RenderX

3. Taber P, Baysinger J, Daniels S, et al. Participatory workflow analysis of newborn genetic screening (NBS) to support tools for improved follow-up: comparing the use case of hemoglobinopathy traits across U.S. states. Int J Neonatal Screen 2025 May 20;11(2):40. [doi: 10.3390/ijns11020040] [Medline: 40407523]
4. Green JM, Hewison J, Bekker HL, Bryant LD, Cuckle HS. Psychosocial aspects of genetic screening of pregnant women and newborns: a systematic review. Health Technol Assess 2004 Aug;8(33):iii-iix. [doi: 10.3310/hta8330] [Medline: 15298822]
5. Boychuk NA, Mulrooney NS, Kelly NR, Goldenberg AJ, Silver EJ, Wasserstein MP. Parental depression and anxiety associated with newborn bloodspot screening for rare and variable-onset disorders. Int J Neonatal Screen 2022 Nov 10;8(4):59. [doi: 10.3390/ijns8040059] [Medline: 36412585]
6. Denayer L, Welkenhuysen M, Evers-Kiebooms G, Cassiman JJ, Van den Berghe H. Risk perception after CF carrier testing and impact of the test result on reproductive decision making. Am J Med Genet 1997 Apr 14;69(4):422-428. [doi: 10.1002/(SICI)1096-8628(19970414)69:4<422::AID-AJMG17>3.0.CO;2-P]
7. Denayer L, De Boeck K, Evers-Kiebooms G, Van den Berghe H. The transfer of information about genetic transmission to brothers and sisters of parents with a CF-child. Birth Defects Orig Artic Ser 1992;28(1):149-158. [Medline: 1340225]
8. Ciske DJ, Haavisto A, Laxova A, Rock LZM, Farrell PM. Genetic counseling and neonatal screening for cystic fibrosis: an assessment of the communication process. Pediatrics 2001 Apr;107(4):699-705. [doi: 10.1542/peds.107.4.699] [Medline: 11335747]
9. Resta R, Biesecker BB, Bennett RL, et al. A new definition of genetic counseling: National Society of Genetic Counselors' task force report. J Genet Couns 2006 Apr;15(2):77-83. [doi: 10.1007/s10897-005-9014-3] [Medline: 16761103]
10. Tluczek A, Ersig AL, Lee S. Psychosocial issues related to newborn screening: a systematic review and synthesis. Int J Neonatal Screen 2022 Sep 27;8(4):53. [doi: 10.3390/ijns8040053] [Medline: 36278623]
11. Vernooij-van Langen AMM, van der Pal SM, Reijntjens AJT, Loeber JG, Dompeling E, Dankert-Roelse JE. Parental knowledge reduces long term anxiety induced by false-positive test results after newborn screening for cystic fibrosis. Mol Genet Metab Rep 2014;1:334-344. [doi: 10.1016/j.ymgmr.2014.07.006] [Medline: 27896106]
12. Triebold M, Skov K, Erickson L, et al. Geographical analysis of the distribution of certified genetic counselors in the United States. J Genet Couns 2021 Apr;30(2):448-456. [doi: 10.1002/jgc4.1331] [Medline: 32929835]
13. Bellaiche MMJ, Fan W, Walbert HJ, et al. Disparity in access to oncology precision care: a geospatial analysis of driving distances to genetic counselors in the U.S. Front Oncol 2021;11:689927. [doi: 10.3389/fonc.2021.689927] [Medline: 34222017]
14. Kladny B, Williams A, Gupta A, Gettig EA, Krishnamurti L. Genetic counseling following the detection of hemoglobinopathy trait on the newborn screen is well received, improves knowledge, and relieves anxiety. Genet Med 2011 Jul;13(7):658-661. [doi: 10.1097/GIM.0b013e31821435f7] [Medline: 21546841]
15. Hurden I, Shaulis R, Heppe S. 2022 Michigan newborn screening program annual report. : Michigan Department of Health and Human Services; 2022 URL: https://tinyurl.com/2wu82p7d [accessed 2026-03-26]
16. Mississippi newborn screening update 2016–2020. : Mississippi State Department of Health; 2020 URL: https://msdh.ms.gov/page//resources/19810.pdf [accessed 2026-03-26]
17. Elenga N. Genetic information to share with parents when newborn screening reveals the presence of sickle cell trait. Int J Pediatr 2024;2024:8910397. [doi: 10.1155/2024/8910397] [Medline: 38433874]
18. Laird L, Dezateux C, Anionwu EN. Neonatal screening for sickle cell disorders: what about the carrier infants? BMJ 1996 Aug 17;313(7054):407-411. [doi: 10.1136/bmj.313.7054.407] [Medline: 8761233]
19. Hoskovec JM, Bennett RL, Carey ME, et al. Projecting the supply and demand for certified genetic counselors: a workforce study. J Genet Couns 2018 Feb;27(1):16-20. [doi: 10.1007/s10897-017-0158-8] [Medline: 29052810]
20. 2024 professional status survey: executive summary. : National Society of Genetic Counselors; 2024 URL: https://www.nsgc.org/Portals/0/Docs/Policy/PSS%202024%20Executive%20Summary_Final.pdf?ver=xMyKy1O8H749GCBpxy6NiA%3D%3D [accessed 2026-03-26]
21. Sims AM, Cromartie SJ, Gessner L, et al. Parents' experiences and needs regarding infant sickle cell trait results. Pediatrics 2022 May 1;149(5):e2021053454. [doi: 10.1542/peds.2021-053454] [Medline: 35441211]
22. Fang KY, Bjering H, Ginige A. Adherence, avatars and where to from here. Stud Health Technol Inform 2018;252:45-50. [Medline: 30040681]
23. Kaphingst KA, Kohlmann WK, Lorenz Chambers R, et al. Uptake of cancer genetic services for chatbot vs standard-of-care delivery models: the BRIDGE randomized clinical trial. JAMA Netw Open 2024 Sep 3;7(9):e2432143. [doi: 10.1001/jamanetworkopen.2024.32143] [Medline: 39250153]
24. Prochaska JJ, Vogel EA, Chieng A, et al. A therapeutic relational agent for reducing problematic substance use (Woebot): development and usability study. J Med Internet Res 2021 Mar 23;23(3):e24850. [doi: 10.2196/24850] [Medline: 33755028]
25. Communicating to advance the public's health: workshop summary. : National Academies Press; 2015 URL: https://sprc.org/wp-content/uploads/2023/01/Communicating2AdvancePublicsHealthWorkshop.pdf [accessed 2026-03-26]
26. Dhinagaran DA, Martinengo L, Ho MHR, et al. Designing, developing, evaluating, and implementing a smartphone-delivered, rule-based conversational agent (DISCOVER): development of a conceptual framework. JMIR Mhealth Uhealth 2022 Oct 4;10(10):e38740. [doi: 10.2196/38740] [Medline: 36194462]

XSL•FO
RenderX

27. Del Fiol G, Borsato E, Bradshaw RL, et al. GARDE-Chat: a scalable, open-source platform for building and deploying health chatbots. J Am Med Inform Assoc 2026 Mar 1;33(3):593-602. [doi: 10.1093/jamia/ocaf211] [Medline: 41524720]
28. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
29. Zoom. URL: https://www.zoom.com/ [accessed 2026-01-23]
30. Flanagan JC. The critical incident technique. Psychol Bull 1954 Jul;51(4):327-358. [doi: 10.1037/h0061470] [Medline: 13177800]
31. Ericsson KA, Simon HA. Verbal reports as data. Psychol Rev 1980 May;87(3):215-251. [doi: 10.1037/0033-295X.87.3.215]
32. Gremler DD. The critical incident technique in service research. J Serv Res 2004 Aug;7(1):65-89. [doi: 10.1177/1094670504266138]
33. Hamilton AB. Qualitative methods in rapid turn-around health services research. BetterEvaluation. 2013. URL: https://www.betterevaluation.org/tools-resources/qualitative-methods-rapid-turn-around-health-services-research [accessed 2026-03-26]
34. Watkins DC. Rapid and rigorous qualitative data analysis: the “RADaR” technique for applied research. Int J Qual Methods 2017;16(1). [doi: 10.1177/1609406917712131]
35. Averill JB. Matrix analysis as a complementary analytic strategy in qualitative inquiry. Qual Health Res 2002 Jul;12(6):855-866. [doi: 10.1177/104973230201200611] [Medline: 12109729]
36. Watkins DC. Qualitative research: the importance of conducting research that doesn't “count”. Health Promot Pract 2012 Mar;13(2):153-158. [doi: 10.1177/1524839912437370]
37. Internet, broadband fact sheet. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/fact-sheet/internet-broadband [accessed 2025-10-10]
38. Fitzgerald-Butt SM, Bodine A, Fry KM, et al. Measuring genetic knowledge: a brief survey instrument for adolescents and adults. Clin Genet 2016 Feb;89(2):235-243. [doi: 10.1111/cge.12618] [Medline: 26032340]
39. Sand-Jecklin K, Coyle S. Efficiently assessing patient health literacy: the BHLS instrument. Clin Nurs Res 2014 Dec;23(6):581-600. [doi: 10.1177/1054773813488417] [Medline: 23729022]
40. Wallston KA, Cawthon C, McNaughton CD, Rothman RL, Osborn CY, Kripalani S. Psychometric properties of the brief health literacy screen in clinical practice. J Gen Intern Med 2014 Jan;29(1):119-126. [doi: 10.1007/s11606-013-2568-0] [Medline: 23918160]
41. Clayman ML, Pandit AU, Bergeron AR, Cameron KA, Ross E, Wolf MS. Ask, understand, remember: a brief measure of patient communication self-efficacy within clinical encounters. J Health Commun 2010;15(Suppl 2):72-79. [doi: 10.1080/10810730.2010.500349] [Medline: 20845194]
42. Li M, Bennette CS, Amendola LM, et al. The Feelings About Genomic Testing Results (FACToR) questionnaire: development and preliminary validation. J Genet Couns 2019 Apr;28(2):477-490. [doi: 10.1007/s10897-018-0286-9] [Medline: 30964586]
43. Brooke J. SUS-a quick and dirty usability scale. In: Usability Evaluation in Industry: CRC Press; 1996:4-7 URL: https://www.taylorfrancis.com/chapters/edit/10.1201/9781498710411-35/sus-quick-dirty-usability-scale-john-brooke [accessed 2026-03-26]
44. Arevalo M, Brownstein NC, Whiting J, et al. Strategies and lessons learned during cleaning of data from research panel participants: cross-sectional web-based health behavior survey study. JMIR Form Res 2022 Jun 23;6(6):e35797. [doi: 10.2196/35797] [Medline: 35737436]
45. Pratt-Chapman M, Moses J, Arem H. Strategies for the identification and prevention of survey fraud: data analysis of a web-based survey. JMIR Cancer 2021 Jul 16;7(3):e30730. [doi: 10.2196/30730] [Medline: 34269685]
46. Curran PG, Hauser KA. I’m paid biweekly, just not by leprechauns: evaluating valid-but-incorrect response rates to attention check items. J Res Pers 2019 Oct;82:103849. [doi: 10.1016/j.jrp.2019.103849]
47. Derakshan N, Eysenck MW. Anxiety, processing efficiency, and cognitive performance: new developments from attentional control theory. Eur Psychol 2009;14(2):168-176. [doi: 10.1027/1016-9040.14.2.168]
48. Green MJ, Peterson SK, Baker MW, et al. Use of an educational computer program before genetic counseling for breast cancer susceptibility: effects on duration and content of counseling sessions. Genet Med 2005 Apr;7(4):221-229. [doi: 10.1097/01.gim.0000159905.13125.86] [Medline: 15834239]
49. Sim SC, Zhou XD, Hom LD, Chen C, Sze R. Effectiveness of pre-counseling genetic education workshops at a large urban community health center serving low-income Chinese American women. J Genet Couns 2011 Dec;20(6):593-608. [doi: 10.1007/s10897-011-9397-2] [Medline: 21826577]
50. Institute of Medicine (US) Committee on Quality of Health Care in America. Crossing the Quality Chasm: A New Health System for the 21st Century: National Academies Press; 2001. [doi: 10.17226/10027]
51. Osterman MJK, Hamilton BE, Martin JA, Driscoll AK, Valenzuela CP. Births: final data for 2022. Natl Vital Stat Rep 2024 Apr;73(2):1-56. [Medline: 38625869]
52. Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The breakthrough of large language models release for medical applications: 1-year timeline and perspectives. J Med Syst 2024 Feb 17;48(1):22. [doi: 10.1007/s10916-024-02045-3] [Medline: 38366043]

XSL•FO
RenderX

53. Tudor Car L, Dhinagaran DA, Kyaw BM, et al. Conversational agents in health care: scoping review and conceptual analysis. J Med Internet Res 2020 Aug 7;22(8):e17158. [doi: 10.2196/17158] [Medline: 32763886]
54. Chavez-Yenter D, Kimball KE, Kohlmann W, et al. Patient interactions with an automated conversational agent delivering pretest genetics education: descriptive study. J Med Internet Res 2021 Nov 18;23(11):e29447. [doi: 10.2196/29447] [Medline: 34792472]
55. GARDE-Chat—the University of Utah chatbot authoring tool—README. URL: https://chatbot2.bmi.utah.edu/ubot/readme [accessed 2026-03-31]
56. Chatbot. URL: https://chatbot1.bmi.utah.edu/ubot/chatapp/chatbot-nbsbot [accessed 2025-03-31]

Abbreviations

AURA: Ask, Understand, Remember
FACToR: Feelings About Genomic Testing Results
LLM: large language model
NBS: newborn screening
SCT: sickle cell trait
SUS: System Usability Scale

Edited by A Stone; submitted 16.Oct.2025; peer-reviewed by N Ogbonna, X Ding, Y Yang; revised version received 02.Mar.2026; accepted 10.Mar.2026; published 30.Apr.2026.

Please cite as:
Madeo AC, Gauchel C, Borsato E, Gaviglio A, Kaphingst KA, Del Fiol G, Eilbeck K
A Chatbot to Meet Parents' Information Needs for Sickle Cell Trait Newborn Screening Results: Multiple Methods Formative Study
J Med Internet Res 2026;28:e86022
URL: https://www.jmir.org/2026/1/e86022
doi:10.2196/86022

© Anne C Madeo, Courtney Gauchel, Emerson Borsato, Amy Gaviglio, Kimberly A Kaphingst, Guilherme Del Fiol, Karen Eilbeck. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Enhancing Detection of Message Intents in a Mobile Health Smoking-Cessation Intervention Using Large Language Model Fine-Tuning, Data Downsampling, and Error Correction: Algorithm Development and Validation

Shagoto Rahman[1], MS; Cornelia (Connie) Pechmann[2], PhD; Ian G Harris[1], PhD

[1]Department of Computer Science, University of California, Irvine, Irvine, CA, United States

[2]The Paul Merage School of Business, University of California, Irvine, Irvine, CA, United States

Corresponding Author:
Shagoto Rahman, MS
Department of Computer Science, University of California, Irvine, Irvine, CA, United States

Abstract

Background: Although smoking-cessation aids such as support groups and nicotine replacement therapy (NRT) can help people quit, quit rates remain low. Mobile health interventions can boost accessibility and engagement, especially with NRT, but require ongoing effort to deliver timely responses. Accurate intent detection is crucial for identifying user needs and delivering timely, appropriate chatbot responses. Recent large language model advancements in natural language processing and artificial intelligence (AI) have shown promise. However, these systems often struggle with many intent categories, complex language, and imbalanced data, reducing recognition accuracy.

Objective: The main goal of this study was to develop an AI tool, a large language model that could accurately detect people's message intents, despite dataset imbalances and complexities. In our application, the messages came from a smoking-cessation support-group intervention and often involved the use of NRT provided as part of that intervention.

Methods: We consistently used a state-of-the-art public domain large language model, Llama-3 8B (8 billion parameters) from Meta. First, we used the model off-the-shelf. Second, we fine-tuned it on our annotated dataset with 25 intent categories. Third, we also downsampled the predominant intent category to reduce model bias. Finally, we combined downsampling with corrected human annotations, creating a cleaned dataset for a new round of fine-tuning.

Results: Without fine-tuning, the model achieved unweighted and weighted F_1-scores (overall performance) of 0.41 and 0.38, respectively, on the downsampled corrected test dataset, and 0.29 and 0.35 on the full test dataset. Fine-tuning improved performance to 0.77 and 0.80 on the downsampled corrected dataset, and 0.72 and 0.86 on the full dataset. Fine-tuning with downsampling attained the best F_1-scores, 0.88 and 0.91 on the downsampled corrected dataset, though performance dropped on the full test dataset (0.58 unweighted, 0.66 weighted) due to the predominance of the off-topic intent category, while unweighted recall remained high (0.80). The final method combining fine-tuning, downsampling, and error correction achieved 0.86 unweighted and 0.90 weighted F_1-scores on the downsampled corrected dataset, and 0.57 and 0.65 on the full dataset with unweighted recall improving to 0.82.

Conclusions: Large language models performed poorly without fine-tuning, highlighting the need for domain-specific training. Even with fine-tuning, performance was limited by a highly imbalanced dataset. Downsampling before fine-tuning moderately improved performance but still left room for improvement and concerns about dataset noise. A careful review of model-human disagreement cases helped identify human annotation errors. After error correction, the method without error correction still achieved slightly higher precision and F_1-score on the corrected test dataset. While error correction slightly improved recall on noisy data, automated downsampling alone may be sufficient, making manual correction a more resource-intensive option with limited added benefit.

(J Med Internet Res 2026;28:e83437) doi:10.2196/83437

KEYWORDS

smoking cessation; intent detection; data imbalance; large language model; human annotation error analysis; chatbot

XSL•FO
RenderX

Introduction

Background

Although most smokers wish to stop smoking, cessation rates remain low in part because of the insufficient use of the cessation aids [1-3]. The use of mobile health interventions to aid people in quitting smoking has become commonplace [4,5], including SMS text messaging programs [6,7] and quit-smoking apps [4,8]. As traditional in-person support groups significantly improve quit rates [9,10], online groups have emerged to help overcome usage barriers related to accessibility and convenience [11]. Mobile health smoking-cessation support groups offer considerable promise; however, they require ongoing, resource-intensive efforts to provide timely, 24/7 responses that sustain participant engagement. Research has consistently demonstrated that active involvement in online support groups leads to improved outcomes [12,13], and studies also show that low engagement is often linked to dropouts [14].

Chatbots can enhance support-group member engagement and minimize dropout rates by providing continuous, cost-effective support [15-17]. Accurate intent detection, defined as the precise and consistent classification of a user's communication goal from their input message, is a leading determinant of chatbot effectiveness. Accurate intent detection can improve user engagement by providing the inputs needed to the system to provide personalized, accurate, and timely support. It can also minimize the need for human intervention. With accurate intent detection, a chatbot can engage users by actively responding to messages with relevant content, addressing their concerns, and delivering evidence-based health information and support. For example, if a user indicates they have failed to quit smoking, the chatbot can identify this particular intent and suggest ways to overcome the challenge.

Previous chatbot studies have primarily focused on 1-to-1 conversations between chatbot and user, for example, for mental health or substance abuse screening [18] or interventions [19]. We study conversations in support groups, often involving several group members, which have unique properties compared to 1-to-1 conversations including multiparty dialogue, interruptions, subgroup discussions, and incomplete messaging leading to exceptionally unstructured and noisy data [20]. In addition, research has been done to detect message intents in medical settings, for example, online forums or office visits, but most involved limited intent categories which limited their ability to fully capture the message content [21,22].

Moreover, data imbalance is particularly common in the health domain because sufficient data is rarely available for all intent categories, yet ones with fewer samples can be critically important [23,24]. For example, an intent category such as overdose may have very few messages, but it represents a critical concern. Most prior studies did not properly ensure that the samples or exemplars were sufficiently balanced across intent categories, possibly leading to suboptimal model performance. Data downsampling, which involves reducing the number of samples from highly dominant majority categories to achieve a more balanced category distribution, can play a vital role. This approach can help the classifying model learn to detect all intent categories comparably, including those that are rare but critically important.

Large language models can deliver state-of-the-art performance in recognizing message intents because they are trained on vast amounts of data. As a result, these models can grasp the deeper meaning of text and often replicate human-like understanding. However, for highly specific classification tasks, large language models must be fine-tuned on domain-specific datasets where humans have annotated the message intents for this fine-tuning phase. Since data annotation is primarily done by humans, it often contains errors due to message ambiguity, human misunderstanding, carelessness, bias, or other factors. Manually reviewing the entire dataset to find human annotation errors is costly and time-consuming. Instead, focusing only on the data points where the model and human annotators disagree on the user's message intent can save time and effort while effectively identifying annotation issues.

In this study, we demonstrate common challenges and solutions in using large language models for mobile health interventions. We implement a state-of-the-art, public domain, large language model called Llama-3 8B (8 billion parameters) from Meta to detect the message intents in a mobile health smoking-cessation intervention. Our study method consists of 4 parts. Part 1 involves intent detection using the Llama-3 8B large language model off-the-shelf, without fine-tuning. Part 2 fine-tunes the model on a domain-specific dataset which human annotators annotated with what they believed to be the message intents. Part 3 combines fine-tuning with downsampling of the majority (predominant) intent category, which was found to bias the model and weaken its performance at intent recognition. Part 4 adds correction of human error in intent annotation to the fine-tuning and downsampling. The rest of this paper is organized as follows: We elaborate on prior work, discuss our methods, report our experimental results, summarize the principal findings and limitations, and state future research directions and a conclusion.

Prior Work

Conversational Agents in Health Care

Conversational health care agents have become important resources for providing quick and easily accessible medical assistance to patients. Recent research suggests that by harnessing advancements in natural language processing, they can provide personalized patient support and alleviate the excessive workload of many health care professionals. Babu and Boddu [25] created a BERT (Bidirectional Encoder Representations From Transformers)-based medical conversational agent to provide responses to people seeking medical advice, appointments, or information about specific medical conditions. The BERT model was fine-tuned on public databases such as MIMIC-III (Medical Information Mart for Intensive Care-III), PubMed, and a COVID-19 dataset totaling 11,000 samples of medical questions. The conversational agent integrated these extracted datasets with the ongoing conversation context and produced relevant and personalized feedback. When compared to baseline models such as LSTM (long short-term memory), SVM (support vector machine), and Bi-LSTM (bidirectional long short-term memory), the chatbot

XSL•FO
RenderX

demonstrated superior performance, attaining 98% accuracy along with high precision, recall, and F_1-scores combining precision and accuracy. However, the system demanded high computational power and suffered from data bias and interpretability issues.

Yu and McGuinness [26] developed a conversational agent by combining a fine-tuned DialoGPT model and ChatGPT to assist mental health conversations. The authors sought to address current model limitations in delivering proper and safe therapeutic responses due to their lack of training on specialized data. So, authors fine-tuned the DialoGPT model on a domain-specific dataset focused on cognitive behavioral therapy and integrated it with ChatGPT's dynamic conversational capabilities. This hybrid conversational agent system was compared to conventional rule-based models and large language models on metrics such as BLEU (Bilingual Evaluation Understudy), which assesses the match between the model output and human-written text. Inputs from mental health professionals and patients were also considered. The results demonstrated that the hybrid system outperformed others in terms of conversational quality, relevance, and emotional resonance.

Zamani et al [27] sought to address the limitations of rule-based chatbots as mental health conversational agents, including lack of response creativity and low user acceptability. They also sought to address the risks of large language models in terms of generating harmful texts and the limitations of data imbalance. The authors evaluated the impact of novel text augmentation techniques to improve the message intent detection. Their study demonstrated that increasing the training dataset through augmentation could boost model accuracy by contributing to dataset diversity. The study incorporated easy data augmentation (EDA), synonym replacement using WordNet, embedding-based replacement, and context-aware substitutions. Model performance improved markedly through data augmentation, with up to a 26.4% increment in the average intent detection score.

Large Language Models for Intent Detection in Health Care

Grasping the speaker's message intent is essential for facilitating smooth and meaningful human interaction. With recent advancements in natural language processing and deep learning, various recent studies have demonstrated success in detecting the subtle intents behind spoken expressions. Aftab et al [28] developed an intent classification method for health care conversational agents by incorporating Bayesian LSTMs. They used an open-source dataset containing 6661 text utterances corresponding to 25 unique medical symptom-related intents with a relatively balanced distribution of intent categories. Their method attained an impressive 99.4% accuracy in recognizing message intents on the test dataset. Moreover, the authors used Shannon entropy from several random model runs to measure uncertainty, which helped the system distinguish between familiar and unfamiliar inputs. Additionally, to estimate model uncertainty, they used Monte Carlo dropout during inference. As a result, the chatbot handled uncertain predictions more appropriately, allowing it to generate safer, less harmful responses.

Zhang et al [29] developed Conco-ERNIE, an innovative model designed to enhance user intent detection in complex medical queries by permitting single intent, multi-intent, and implicit intent recognition. The model used a special feature extractor called ERNIE with an added layer to help it understand and annotate important parts of medical questions. The authors incorporated the Apriori algorithm to find patterns in how medical concepts related to each other, allowing it to better detect intents when the uttered meanings were unclear. Moreover, they introduced an attention-based method that combined the meaning of words in a query with the patterns of related medical concepts. This helped the model focus better on important information and ignore irrelevant details. The authors tested their method on a dataset of real medical questions, and it showed clear improvements in detecting multiple and hidden user intents. Overall, the model achieved a score of 89.4% in test data, much higher than other popular methods. The authors then added their model Conco-ERNIE to a real health care chatbot linked to a large medical knowledge base.

Sakurai and Miyao [30] studied large language models for intent detection in conversations involving donation requests. The study examined conversations where the responder expressed criticism of donations or sought to make donors feel guilty. The authors primarily studied GPT-4 in this context and found it often misidentified critical, sarcastic, or negative affective responses as motivational or neutral. GPT-4 was tested on both real and artificially created conversations, and its predictions about intent were compared to human-annotated intents. Humans were able to detect critical and other negative messages, but GPT-4 often misunderstood them as encouraging or motivational. Overall, this study found that a state-of-the-art large language model struggled to detect subtle intents in negative speech.

Our Contributions

In this paper, we demonstrate that a current state-of-the-art large language model cannot adequately detect subtle intents expressed by participants in a smoking-cessation intervention; for example, the model struggles to decipher substantively different messages about the study-provided nicotine replacement therapy (NRT). The methodological novelty of our work lies in systematically addressing highly imbalanced, domain-specific data for smoking-cessation text classification using the Llama-3 8B model. We illustrate the specific steps that can be taken, using existing analytical methods, to improve the model's intent detection. We show vast improvements by fine-tuning the off-the-shelf model using a domain-specific dataset with human annotations of 25 message intents. We show additional substantial improvements by downsampling (reducing excess examples of) an overly dominant intent category and by correcting human annotation errors identified by examining model-human disagreement in intent classification. In sum, we show that large language models can perform very poorly in domain-specific contexts like ours but can be massively improved with current analytical approaches. This approach

XSL•FO
RenderX

improves efficiency and robustness, providing insights for future work such as multi-intent detection, probability calibration, and translation of technical performance into operational impact.

Methods

Overview

Figure 1 presents our method for training a large language model for the detection of message intents. It involved four stages: (1) large language model without fine-tuning, where the model predicted intents without any domain-specific training dataset solely using its own knowledge; (2) large language model with fine-tuning, where the model was fine-tuned on an intent-annotated dataset to improve performance; (3) large language model with fine-tuning and downsampling, which addressed category imbalance by reducing the sample size of a majority (dominant) intent category, which was off-topic or nondomain related, before fine-tuning; and (4) large language model with fine-tuning, downsampling, and human annotation error correction, where model-human disagreements were reviewed for potential human annotation errors, corrected if necessary, and used to retrain the model. This stepwise approach allowed for progressive enhancement in classification accuracy, with each stage addressing specific limitations of the previous one.

Figure 1. Overview of the 4 large language model training methods used for intent detection in this study (United States).

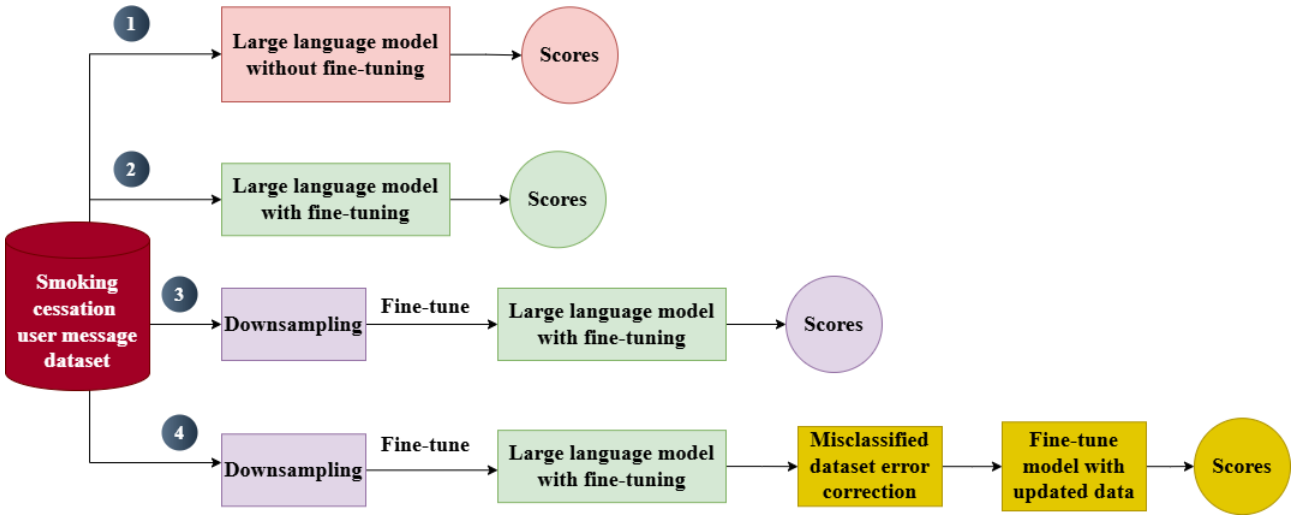

Dataset

For this study, we use all the user messages from the Tweet2Quit mobile health smoking-cessation intervention [17]. All duplicated messages have been properly identified and removed, ensuring the dataset is clean and free from leakage. Tweet2Quit is a social media–based intervention to assist people with quitting smoking. Participants were assigned to private online support groups of 20 members who engaged in peer support discussions about smoking cessation. The dataset contained more than 82,000 messages gathered from 45 support groups participating in 2 clinical trials conducted between 2012 - 2014 and 2016 - 2019. Each support group remained active for 3 months and generated approximately 1822 messages on average. The messages were created by individuals actively working to quit smoking and then stay quit and, as a result, contained domain-rich, real-world interactions regarding quitting and the free NRT, or NRT (nicotine patches, gum, and lozenges) provided as a quitting aid.

A team of 25 trained research assistants, mostly US undergraduates from diverse majors, with a few international participants, annotated the intent behind each message to produce a comprehensive annotated corpus. The assistants were trained on intent definitions, keywords, and examples, then practiced on a set of messages until they attained the required ≥80% accuracy. Initial in-person training meetings transitioned to Microsoft Teams due to team expansion and COVID-19, with weekly sessions during training and biweekly discussions thereafter to resolve difficult cases. Each message was independently annotated by two assistants, with discrepancies resolved by a third expert adjudicator.

A total of 25 intent categories were identified while annotating the messages. The topics either were clinically related to smoking cessation (eg, nrt_howtouse, nrt_overdose, and nrt_mouthirritation), reflected social support for quitting (eg, greetings, support), or were off-topic (eg, small talk, weather, news). Table 1 presents the 25 message intent categories identified in the dataset, along with their definitions and representative examples. The annotation demonstrated high reliability, with a Cohen κ score of 0.93 overall, and agreement between specific annotators ranging from 87.3% to 97.2% across intent categories. The off-topic intent category was used to indicate irrelevant messages unrelated to smoking cessation or support group bonding. It alone accounted for 56% of the dataset, which indicated significant category imbalance.

XSL•FO
RenderX

Table . Definitions and examples of smoking-cessation message intents used in this intent detection study (United States, 2012 - 2014 and 2016 - 2019).

Intent category	Definition	Examples
cigsmell	Complains about cigarette smell	• Cigarette smoke and cigarette smokers stink to me now. • I can smell people smoking a mile away now. It's gross to me now.
cravings	Discusses craving/wanting/temptation/urge/trigger to smoke or how to fight/avoid craving	• It's hard because all I want to do is reach over and get a cigarette and smoke. • My urge is in the evening watching TV I usually occupy my mind with crochet that's all I really have.
ecigs	States something about ecigs, vape, or liquid or juice	• Is anyone going to try vaping to help quit? The non-nicotine juice of course. I've been debating. • I do like the Vape pen.
fail	Admits to smoking a cigarette	• I have to admit I slipped today and smoked a cigarette. • Mornings I'm a chain smoker.
greetings	Says a greeting, eg, hi	• Good afternoon! How's everyone doing today? • Good morning!
health	States something about the relation between smoking and health	• My breathing is much better and I don't never want to smoke a cigarette again. • To be able to breathe better.
nrt_dontwork	States NRT[a] doesn't work, don't like, no longer needed	• Since the gum and patches don't seem to work on my boring days I have struggled. • This gum is a no go for me. I called the smoke shop in my town.
nrt_dreams	States they either sleep with the patch or try not to do so; often discusses dreams	• I'm sleeping with patch on. Have the craziest most vivid dreams. • Yes! Very lucid dreams. I had to stop sleeping with them on.
nrt_howtouse	Asks question or gives info about how to use NRT in general, eg, dose	• How often should I use the lozenges? • So your not suppose to constantly chew the gum right? Suppose to just let it lay in your gums once broken up right?
nrt_itworks	States reason why NRT used, eg, it works, stops craving	• I'm using 1 patch a day and chewing the gum as needed! Both help immensely with cravings! • I have stepped down to the 7mg patch!!!! Yea!!!!!
nrt_mouthirritation	States NRT gum/lozenge causes bad taste, throat gum irritation or burning, or spicy sensation	• I toss the gum when it starts burning my mouth or throat. • Don't like the gum I'm having a hard time chewing it. Burns my mouth a little.
nrt_nauseous	States NRT makes them nauseous or gag or sickly	• Don't like the gum makes me nauseous. • Using the #14 patch. Not using gum it makes me sick.
nrt_od	States NRT is too strong, feels like overdose	• I am going to move to the step 2 patch. The 21mg patches are making me sweat and feel bad. • I had to move from the 21mg early cause it was just too much nicotine!

XSL•FO
RenderX

Intent category	Definition	Examples
nrt_skinirritation	States NRT patch causes skin irritation	• Anybody else's skin being irritated by the patches? • I get itchy from mine too.
nrt_stickissue	States NRT patch won't stick	• I know that with sweat the patches don't stay on I have problems with mine every-day. • My patches won't stay on.
off-topic	Discusses other topic, eg, small talk, weather, news	• Now i gotta go check mine!!! • I skipped breakfast and went right for lunch.
quitdate	States something about quit date	• Hey my quitDate is today. • My last pack of smokes. Friday is the day!
savingmoney	States money they spend per pack or how much they save by quitting	• I'm putting away the money I save into an account and then can do something fun! • All the money spent to damage health.
scared	States being scared/nervous/anxious/fearful about quitting or sometimes NRT	• I am really nervous about quitting but ready. • I've not tried the gum yet. I'm scared of it lol.
smokefree	States success in being smoke-free	• I'm still smokefree starting week 2. YAY. • I made it thru without smoking. Starting Day 2. Yeah.
smokingless	Has not yet quit, only able to cut down on cigs smoked OR starting to quit but still smokes some	• Can’t say I've completely quit. But 4 or 5 a day is less than 15. • Proud moment: only 1 cigarette today. Never thought I'd get this far.
stress	States feeling stress	• I think stress has been tough but I'm not giving up. • Last week work was really stressful had a lot going on at once.
support	Supports other participants	• Good for you for not giving up. You got this!!! • You're doing great! wooohooo!!!
tiredness	States feeling tired due to unspecified reasons or quitting smoking, not NRT	• I have been so tired lately. I'm wondering if it is connected to quitting smoking. • So very tired...
weightgain	States gaining weight or eating more or anything about weight or exercise	• Ever since I quit I eat more so I'm going to start eating healthier and exercise. • I have gained 7 pounds!

[a]NRT: nicotine replacement therapy.

The annotated dataset was divided into 3 mutually exclusive and collectively exhaustive subsets: training, evaluation, and test. The test dataset consisted of messages from support groups 30 to 36. From the remaining groups out of the total 45, 75% of the messages were used for training and 25% for evaluation. The dataset split was stratified based on the distribution of intent categories, independent of group membership. The training dataset was used to fine-tune the large language model, allowing it to learn patterns and features from the annotated examples. During training, the model iterated over the dataset multiple times (epochs), improving its performance by minimizing errors. The evaluation (validation) dataset was used to assess the performance of different versions of the large language model based on its training. Since it was unseen by the model during training, it served as an unbiased measure to select the best-performing model versions. The test dataset was reserved for reporting the final classification results. Also unseen by the model, it was used to evaluate the generalizability and robustness of the trained model, providing an estimate of how well the model would perform when deployed in practical settings. Table 2 summarizes the distribution of annotated messages across the 3 datasets.

XSL•FO
RenderX

Table . Distribution of smoking-cessation message intents across the training, validation, and test datasets used in this intent detection study (United States, 2012 - 2014 and 2016 - 2019).

Intent category	Train dataset (n=46,647)	Evaluation dataset (n=15,550)	Test dataset (n=15,219)
cigsmell	328	109	76
cravings	1446	482	267
ecigs	205	68	73
fail	774	258	152
greetings	1828	610	490
health	847	283	265
nrt_dontwork	307	102	72
nrt_dreams	258	86	54
nrt_howtouse	354	118	105
nrt_itworks	615	205	135
nrt_mouthirritation	140	46	21
nrt_nauseous	122	41	23
nrt_od	81	27	21
nrt_skinirritation	155	51	35
nrt_stickissue	188	63	83
off-topic	26,093	8698	9664
quitdate	1279	427	318
savingmoney	455	152	90
scared	293	98	83
smokefree	3072	1024	818
smokingless	285	95	33
stress	508	169	148
support	6330	2110	1983
tiredness	153	51	43
weightgain	531	177	167

Ethical Considerations

The messages annotated in this study came from two mobile health, smoking-cessation intervention trials, which were supported by U.S. National Institutes of Health grants R34 and R01. The R34 (conducted 2012 - 2014) received human subjects (HS) approval through the University of California, Irvine (UCI) institutional review board (IRB) HS# 2010 - 7990 and the University of California (UC) Reliance 261: “Twitter-Enabled Mobile Messaging for Smoking Relapse Prevention.” The R01 (conducted 2016 - 2019) received human subjects approval through UCI IRB HS# 2014 - 1303 and UC Reliance 977: “Twitter-Enabled Mobile Messaging for Smoking Relapse Prevention (Phase 2).” Participants were compensated for enrolling with 8 weeks of NRT (approximate value US $200) and up to US $110 for survey completions. For this secondary data study, messages were analyzed verbatim after removing all identifying information, such as usernames. Participants provided informed consent for the primary data collection, and the UCI IRB allows for secondary data analysis of deidentified data without additional consent.

Large Language Model Without Fine-Tuning

The first stage of our methodology was classifying message intents using a large language model without any fine-tuning. In other words, the model was tasked with classifying text into the observed message intent categories without any exposure to messages annotated with these intent categories in a training dataset. As a common example, a large language model might be prompted to classify the intent of a message stating, “I am very sad” as either positive or negative in sentiment, and it might correctly identify the intent as negative, despite not having been trained on the specific dataset. Since large language models are trained on vast and diverse text corpora, they have acquired an understanding of syntax, semantics, grammar, and sentiment, making them well-suited for general-purpose language tasks such as intent detection, sentiment analysis, and topic summarization.

When a large language model is used for a classification task without fine-tuning, it is guided through a prompt to emulate the task. For example, in the classification task example above, a suitable prompt might be “Classify the sentiment of the

XSL•FO
RenderX

following text into one of the two categories {positive, negative}: I am very sad," which is presented to the large language model. Then the model uses its implicit learning and reasoning to perform the task. The model works by predicting the most probable next word in the sequence. For instance, in this example, the language model may evaluate the prompt by considering potential sentiment categories such as positive and negative. Drawing on its prior knowledge and contextual understanding, it may assign a higher probability to negative based on the input text and subsequently predict negative as the final classification. The main advantage of this approach is that it does not require annotation of message intents. The only requirement is to provide the test dataset to the large language model, which then performs the classification based solely on its prior knowledge. This approach saves time and computation resources.

However, the concern is that the domain-specific task may involve text that is not part of the dataset used for pretraining the model. For example, in the domain of smoking cessation, the intent behind the messages can often be highly domain-specific as health-intervention communications tend to be confidential. A message such as "I used gum today and wanted to smoke two hours later," might be interpreted by a language model as indicating a cessation failure. However, within the context of our smoking-cessation dataset, this message corresponds to a more specific intent, nrt_dontwork, which signifies that the user perceived NRT as ineffective.

Thus, knowledge of the domain becomes crucial for getting higher accuracy with large language models. When a model lacks domain-specific knowledge, its performance on related tasks tends to decline. Moreover, in the absence of domain-specific knowledge, large language models may introduce biases or errors stemming from the datasets on which they were originally trained, possibly causing hallucinations. Furthermore, large language models without domain knowledge tend to struggle with classification tasks involving numerous similar intent categories, such as our categories of nrt_dontwork and nrt_itworks, which refer to NRT being ineffective or effective, respectively. While these intents have completely different meanings, large language models that lack fine-tuning often cannot distinguish between them.

In this study, we assessed the performance of a large language model without fine-tuning at completing the task we required for our mobile health application: recognizing the 25 main intents of messages related to smoking cessation, which included numerous nuanced messages about using NRT. We used the state-of-the-art but off-the-shelf Llama-3 8B by Meta. We supplied this model with descriptions of all 25 of our intent categories, as defined in Table 1, to facilitate its classification process and ensure that its predictions were constrained to these predefined intent categories.

Large Language Model With Fine-Tuning

The next stage of our methodology was fine-tuning the same large language model, Llama-3 8B, using our domain-specific dataset with smoking-cessation messages. Fine-tuning a large language model refers to further training the model on one or more datasets from the targeted domain, the goal being to transform a general-purpose model into one specialized for the domain for improved performance. Domain-specific knowledge is often crucial for large language models because the datasets used in the initial basic training often lack the detailed context relevant to the specific task. Fine-tuning enables the models to learn domain-relevant nuances, semantics, and reasoning.

For example, during fine-tuning, the model might be exposed to messages about NRT gum aiding or not aiding with smoking cessation, for instance, "I used gum today and wanted to smoke two hours later." Fine-tuning allows the model to learn subtle distinctions and accurately differentiate between semantically similar message intents. For instance, a model without fine-tuning might incorrectly predict the message intent here as fail (failure at smoking cessation). However, with fine-tuning on the domain-specific dataset, the model should learn that the user is talking about their NRT gum failing to help them with nicotine cravings, and therefore the correct intent is nrt_dontwork.

On the other hand, fully training a large language model on a domain-specific dataset risks the model forgetting its previously acquired general knowledge. Consequently, the model might forget critical knowledge related to language comprehension, reasoning, and generalizing across tasks. Therefore, instead of completely retraining the model, we undertake fine-tuning because it preserves and often freezes the original model weights (ie, the models' prior knowledge) while learning new weights (new knowledge) that is built upon the existing foundation. So, this approach enables the model to retain its previously acquired knowledge while simultaneously incorporating new insights from the domain-specific dataset.

Therefore, we fine-tune the Llama-3 8B large language model on our smoking-cessation dataset. We selected Llama-3 8B to balance computational efficiency and model performance with feasibility for in-house fine-tuning and inference, considering our resource constraints. Meta's Llama-3 8B, when used in causal mode, is designed to predict the next word in a sequence sequentially. When used in this mode, intent detection requires specialized prompting, for example, "Input: I am feeling very sleepy. So the intent of this message is:." This prompt is sent to the model which then predicts the next probable word as tiredness. This prompt can be formulated in various ways; however, the approach has several limitations. It is highly dependent on the prompt design and it operates indirectly through generation, which reduces efficiency, makes end-to-end training more challenging, and necessitates careful prompt engineering and output parsing. Often, the model's output may fall outside the predefined intent categories; for instance, in the previous example where the main intent is tiredness, the model might generate a related but nonfocal intent such as drained.

For classifying message intents, we used a version of the Llama-3 model that is specifically adapted for classification rather than using the standard causal mode designed for language generation. Instead of predicting the next word in a sentence, this version of the model learns to assign entire messages to intent categories. To achieve this, we added a lightweight classification layer which works as the category chooser producing the final model results. This layer enables the model

XSL•FO
RenderX

to directly select the most appropriate category for each input message. Importantly, this approach allows us to begin to address the issue of dataset imbalance, where some intent categories are more common than others. By assigning more weight to underrepresented categories during fine-tuning, the model becomes more attentive to less frequent but clinically important message types.

Additionally, this approach gives us a probability score for each category, which helps us understand how confident the model is in its predictions. This is valuable for downstream decision-making and human oversight. Figure 2 provides an overview of this fine-tuning process, showing that the model's outputs feed into the category chooser, which then predicts the most appropriate intent category. During fine-tuning, we assign more importance to underrepresented categories when the model makes mistakes. This is done by adding category weights to the penalty for mistakes, helping the model pay more attention to rare but clinically important intent categories. We fine-tune the Llama-3 8B model in this way to improve its ability to detect nuanced user intents in smoking-cessation conversations, for example, related to the use of the free NRT provided as a cessation aid.

Figure 2. Large language model fine-tuning used in this intent detection study (United States).

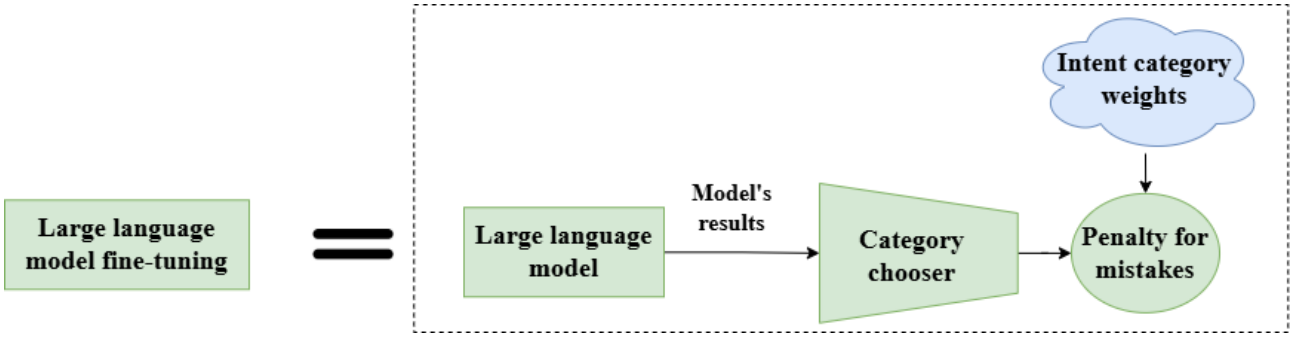

Large Language Model With Fine-Tuning and Downsampling

The third stage of our methodology was downsampling the dominant off-topic (domain-unrelated) message intent category and then fine-tuning the model on the updated dataset. Dataset imbalance is a typical problem in classification tasks, meaning that the number of instances (samples or examples) often varies quite substantially across the different focal categories. For example, in our dataset, the off-topic intent category contained 44,445 samples, whereas the nrt_od category contained only 129 samples. In a multicategory classification task, if a particular category has a much higher number of samples than the other categories, the model being trained tends to be biased toward choosing the category with the higher number of samples. In other words, in the training dataset, the model sees many more samples of this majority category (eg, off-topic) and ends up favoring it in its predictions. Thus, the minority categories (eg, nrt_od) are underpredicted.

For our application which is to accurately predict message intents in smoking-cessation support groups, both majority and minority intent categories require accurate predictions. But, if we train the model with all our data, the model will tend to be more accurate in predicting the highly dominant off-topic category and less accurate in predicting the minority categories with fewer samples. Additionally, the off-topic intent category refers to messages considered irrelevant or uninformative for smoking cessation or support group bonding; it falls outside the 24 key intent categories. Knowing this gives us even more leeway to address our severe category imbalance by reducing the volume of off-topic category examples. Doing this will not compromise the data from the other 24 intent categories.

So for our analysis, we have downsampled the off-topic intent category. Downsampling is a method used to mitigate category imbalance within a dataset by decreasing the sample size of the majority category or categories to better align with sample sizes of the minority categories. We downsampled the instances of the off-topic category to mirror the average number of instances across the other 24 categories. We consistently applied this approach to our training, test, and evaluation datasets. To determine the appropriate level of downsampling for the large off-topic category, we experimented with several approaches, evaluated them on the evaluation dataset, and selected the best-performing approach, which was the average instances across all other category data. These results are reported in Multimedia Appendix 1. After this, we fine-tuned the model using the same approach as in the Large Language Model With Fine-Tuning section. After downsampling the off-topic intent category, the model was fine-tuned using category weights to penalize misclassification of underrepresented categories, ensuring they received sufficient attention during fine-tuning. Although downsampling reduced the disproportionality of the off-topic category, the dataset remained imbalanced, and further downsampling of clinically meaningful categories was not desirable due to the risk of losing valuable information. Thus, category weighting was applied to enhance learning of the rare but important categories while maintaining balance in our fine-tuning process.

The typical limitations of downsampling include loss of information, reduced generalizability, and degraded intent-detection performance in the dominant category. However, these concerns were less applicable in our setting, as we only downsampled the off-topic intent category. This category is unrelated to the 24 focal intent categories and largely represents noise, so reducing its representation is less problematic. In other words, failing to detect some off-topic messages is less harmful than misclassifying important on-topic messages. Even if the detection of off-topic intents decreases, as long as the 24 on-topic intents are accurately recognized, the

XSL•FO
RenderX

system remains effective. Misclassifying an off-topic message intent as 1 of the 24 focal intents keeps the chatbot response within the smoking-cessation domain, which is desirable.

Large Language Model With Fine-Tuning, Downsampling, and Error Correction

The final stage in our methodology was downsampling the dominant off-topic intent category, fine-tuning the model, correcting human annotation errors flagged by examining model-human disagreements in the intent classifications, and fine-tuning the model again. Since the datasets were manually annotated, they were susceptible to human annotation errors. Though large language models are powerful tools due to their training on extensive datasets, fine-tuning them on imperfectly annotated or noisy data can negatively impact their learning and classification performance.

Model-human disagreement on a message's intent category can reveal a weakness in either the model or the human annotation. The model could have incorrectly classified the user's intent, the human annotators could have done so, or possibly both. Therefore, identifying model-human disagreement over how to classify message intents and resolving the human-caused annotation errors enables us to improve the final dataset we use for model fine-tuning. If the model and humans disagree, reviewing these disagreements helps us identify human annotation mistakes, for example, due to messages being ambiguous or vague, or general human oversights in annotations. The human annotation errors can then be corrected in the dataset, and fine-tuning can proceed with less noisy data.

Thus, we analyzed model-human disagreements across the downsampled training, evaluation, and test datasets to identify potential human annotation errors. A domain expert independently reviewed each flagged case and corrected original annotations only when the original human one was incorrect; the model's prediction was never adopted as the new annotation. Only the corrected training dataset was used for fine-tuning again along with the corrected evaluation dataset, and the corrected test dataset remained a true holdout throughout this process. The final performance scores were computed on the expert-validated test dataset, ensuring that the scores remained unbiased and free of data leakage. This approach also ensured consistency across training, evaluation, and testing, as all datasets were corrected using the same procedure and verified by the same domain expert. Category weighting was also applied in this method to enhance the model's learning of rare but important categories while ensuring a balanced fine-tuning process.

Fine-Tuning Hyperparameters

For reproducibility, we provide details on our fine-tuning and inference setup here. We fine-tuned the Llama-3 8B model for 25-category sequence classification using 4-bit quantized low-rank adaptation (QLoRA) within the parameter-efficient fine-tuning (PEFT) framework. Low-rank adaptation (LoRA) adapters (rank=64, α=16) were applied to the query projection (q_proj), key projection (k_proj), value projection (v_proj), and output projection (o_proj) layers, with a dropout rate of 0.05. The category chooser (classification head) is a single linear layer (fully connected) that maps the large language model's hidden dimension of 4096 to the 25 intent categories. Fine-tuning was performed in 32-bit floating point (FP32) precision with 4-bit quantization for memory efficiency along with the Adaptive Moment Estimation With Weight Decay (AdamW) optimizer with a learning rate of 1×10^{-4} , weight decay of 0.01, batch size of 8, and 5 epochs. Early stopping with a patience of 7 monitored unweighted average precision was used to load the best model. Category (class) weights were set inversely proportional to label frequency (normalized) to address category imbalance. Fixed random seeds (42, 123, and 2023) were used for both CPU (central processing unit) and GPU (graphics processing unit), with deterministic Compute Unified Device Architecture (CUDA) operations for reproducibility on the NVIDIA GeForce RTX 4090 GPU. Tokenization used a maximum sequence length of 512, padded to the model's end-of-sequence (EOS) token. Evaluation and checkpoint saving occurred every 1000 steps. During inference, computation was performed using Brain Floating Point 16 (bfloat16) precision with 4-bit quantization for memory efficiency.

Results

Evaluation Metrics

To evaluate the different large language model training approaches for recognizing message intents, we used 3 standard accuracy measures:

Precision: How often the model was correct when it predicted a certain intent.

$$\text{Precision}=\frac{\text{Correct predictions for an intent}}{\text{Total predictions made for that intent}}$$

Recall: How often the model successfully recognized an intent when a message conveyed it.

$$\text{Recall}=\frac{\text{Correct predictions for an intent}}{\text{Total actual cases of that intent}}$$

F_1-score: A numeric combination of precision and recall.

$$\text{F1}-\text{score}=2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

Aggregate Performance

Figure 3 summarizes the performance of the Llama-3 8B model across the 4 training methods. The downsampling-only and downsampling with error correction methods involved modifying the training dataset through reduction of off-topic intent samples and annotation error correction, respectively, so we evaluated all 4 training methods using 3 distinct test datasets. Evaluation results are shown in Figure 3A using the downsampled uncorrected test dataset, in Figure 3B using the downsampled corrected test dataset, and in Figure 3C using the full unmodified test dataset.

XSL•FO
RenderX

Figure 3. Overall unweighted and weighted precision, recall, and F_1-scores for 4 large language model training methods for intent detection in smoking-cessation messages in this study (United States). Methods were evaluated on the (A) downsampled uncorrected test dataset, (B) downsampled corrected test dataset, and (C) full test dataset.

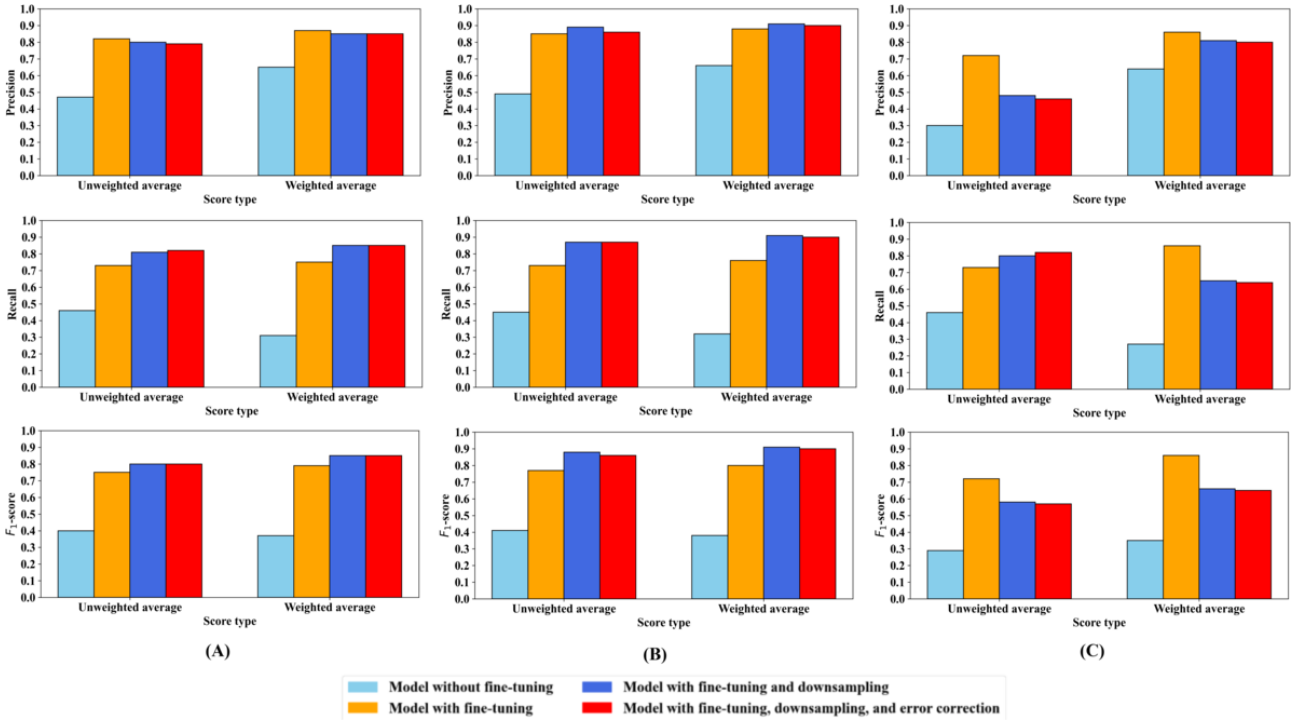

In Figure 3A, which uses the downsampled uncorrected test dataset, the off-the-shelf model with extensive pretraining but no domain-specific fine-tuning performed poorly at recognizing message intents. It achieved unweighted F_1-score, precision, and recall of 0.40, 0.47, and 0.46, respectively, and weighted F_1-score, precision, and recall of 0.37, 0.65, and 0.31, respectively. The unweighted scores assigned equal importance to each intent category, while the weighted scores gave greater weight to categories with more instances. The low scores highlight the complexity of the classification task, demonstrating that intent classification in the mobile health context studied was not straightforward and required specialized fine-tuning to achieve adequate performance. The domain-specific fine-tuning method achieved a much-improved unweighted F_1-score, precision, and recall of 0.75, 0.82, and 0.73, respectively, and weighted F_1-score, precision, and recall of 0.79, 0.87, and 0.75, respectively. The method of fine-tuning after downsampling the dominant off-topic intent category improved the unweighted F_1-scores from 0.75 to 0.80 and yielded unweighted precision and recall of 0.80 and 0.81, respectively. The weighted F_1-score, precision, and recall were all about 0.85. The fine-tuning, downsampling, and error correction method performed comparably to the downsampling-only method, achieved unweighted F_1-score, precision, and recall of 0.80, 0.79, and 0.82, respectively, and similar weighted scores.

In Figure 3B, which uses the downsampled corrected test dataset, the method with no fine-tuning obtained an unweighted F_1-score, precision, and recall of 0.41, 0.49, and 0.45, respectively, and a weighted F_1-score, precision, and recall of 0.38, 0.66, and 0.32, respectively. Fine-tuning yielded an unweighted F_1-score, precision, and recall of 0.77, 0.85, and 0.73, respectively, and a weighted F_1-score, precision, and recall of 0.80, 0.88, and 0.76, respectively. The relatively lower recall indicates the model still struggles to consistently identify certain intents when present. The downsampling and fine-tuning method achieved an unweighted F_1-score, precision, and recall of 0.88, 0.89, and 0.87, respectively, and a weighted F_1-score, precision, and recall of 0.91 for all three. The method of downsampling, fine-tuning, and error correction achieved an unweighted F_1-score, precision, and recall of 0.86, 0.86, and 0.87, respectively, and a weighted F_1-score, precision, and recall of 0.90 for all three. Although error correction yielded good performance, the downsampling-only approach achieved slightly higher unweighted precision and higher weighted scores.

In Figure 3C, which uses the full test dataset, the method without fine-tuning achieved an unweighted F_1-score, precision, and recall of 0.29, 0.30, and 0.46, respectively, and a weighted F_1-score, precision, and recall of 0.35, 0.64, and 0.27, respectively. Among all test cases, 991 predictions could not be matched to any of the 25 intent categories (eg, the model predicted “stick” instead of “stickissues”); these samples were included in the final scoring as misclassifications. The fine-tuning method achieved an unweighted F_1-score, precision, and recall of 0.72, 0.72, and 0.73, respectively, and a weighted F_1-score, precision, and recall of 0.86 for all three. The fine-tuning and downsampling method attained an unweighted F_1-score, precision, and recall of 0.58, 0.48, and 0.80, respectively, and a weighted F_1-score, recall, and precision of 0.66, 0.81, and 0.65, respectively. The final method that added error correction achieved an unweighted F_1-score, precision, and recall of 0.57, 0.46, and 0.82, respectively, and a weighted F_1-score, precision, and recall of 0.65, 0.80, and 0.64,

XSL•FO
RenderX

respectively. Although both methods demonstrated comparable overall performance, the error correction method yielded a higher unweighted recall, indicating its robust performance at detecting intents in real-world, noisy datasets. Using the full uncorrected test dataset lowered the scores for our last two methods due to the large number of off-topic intents in the test dataset, which had been downsampled in the training dataset. This led to many off-topic intents being wrongly predicted as in-domain intents; however, this was acceptable to us as the main goal was to accurately identify the 24 in-domain intents. Misclassifying off-topic intents as in-domain allowed for in-domain chatbot responses focused on the smoking-cessation domain. Moreover, even when we used the full test dataset, our last two methods sequentially improved recall. These improvements suggest that our last training steps helped the model learn to identify the relevant in-domain intents more effectively in the noisy real-world data, which was our main goal. In other words, unweighted recall served as the more informative evaluation metric for our application.

In Figures 3A and 3B, the differences between weighted and unweighted scores were relatively small because of the downsampled off-topic intent category. However, in Figure 3C, the discrepancy was more pronounced because the full test dataset contained a large number of off-topic intent samples, and in the weighted scores, each intent category's contribution is proportional to its sample size, so more frequent intents are treated as more important. In our application, the unweighted recall scores are particularly important, as these give equal importance to all intents regardless of their frequency.

To assess result consistency, we fine-tuned our final method of downsampling with error correction using three random seeds (42, 123, and 2023) to account for variability in model initialization. We observed a mean unweighted F_1-score of 0.85 (SD 0.01) and a mean weighted F_1-score of 0.89 (SD 0.01) on the downsampled corrected dataset, demonstrating high consistency. Additionally, for our last two methods, we used bootstrap resampling, a statistical method that repeatedly evaluates the model on slightly different versions of the test dataset to estimate performance variability. For the downsampling and error correction method, the unweighted F_1-scores ranged from 0.85 to 0.88 with 95% confidence. For the downsampling-only method, the unweighted F_1-scores ranged from 0.86 to 0.89 with 95% confidence. Both results indicate stable performance.

We conducted a bootstrap resampling analysis with replacement (1,000 resamples; random seed=42) on the full test dataset and computed the unweighted-average recall for the downsampling-only and the downsampling and error correction methods. A paired 2-tailed t test on the bootstrap differences indicated that error correction significantly improved unweighted-average recall (t_{999}=69.5, P<.001). The 95% CI for the improvement was 0.0022-0.0292, suggesting a small but consistent gain over downsampling alone.

Performance by Message Intent Category

Precision for Each Intent Category

Figure 4 presents the unweighted precision scores for predicting each of the 25 intents across the different training methods. Unweighted precision reflects the proportion of correct predictions among all instances predicted for a given intent and requires balanced performance across intents. Evaluation results are shown in Figure 4A using the downsampled uncorrected test dataset, in Figure 4B using the downsampled corrected test dataset, and in Figure 4C using the full unmodified test dataset.

XSL•FO
RenderX

Figure 4. Unweighted precision scores for the 4 large language model training methods used to detect intents in smoking-cessation messages in this study (United States). Methods were evaluated on the (A) downsampled uncorrected test dataset, (B) downsampled corrected test dataset, and (C) full test dataset.

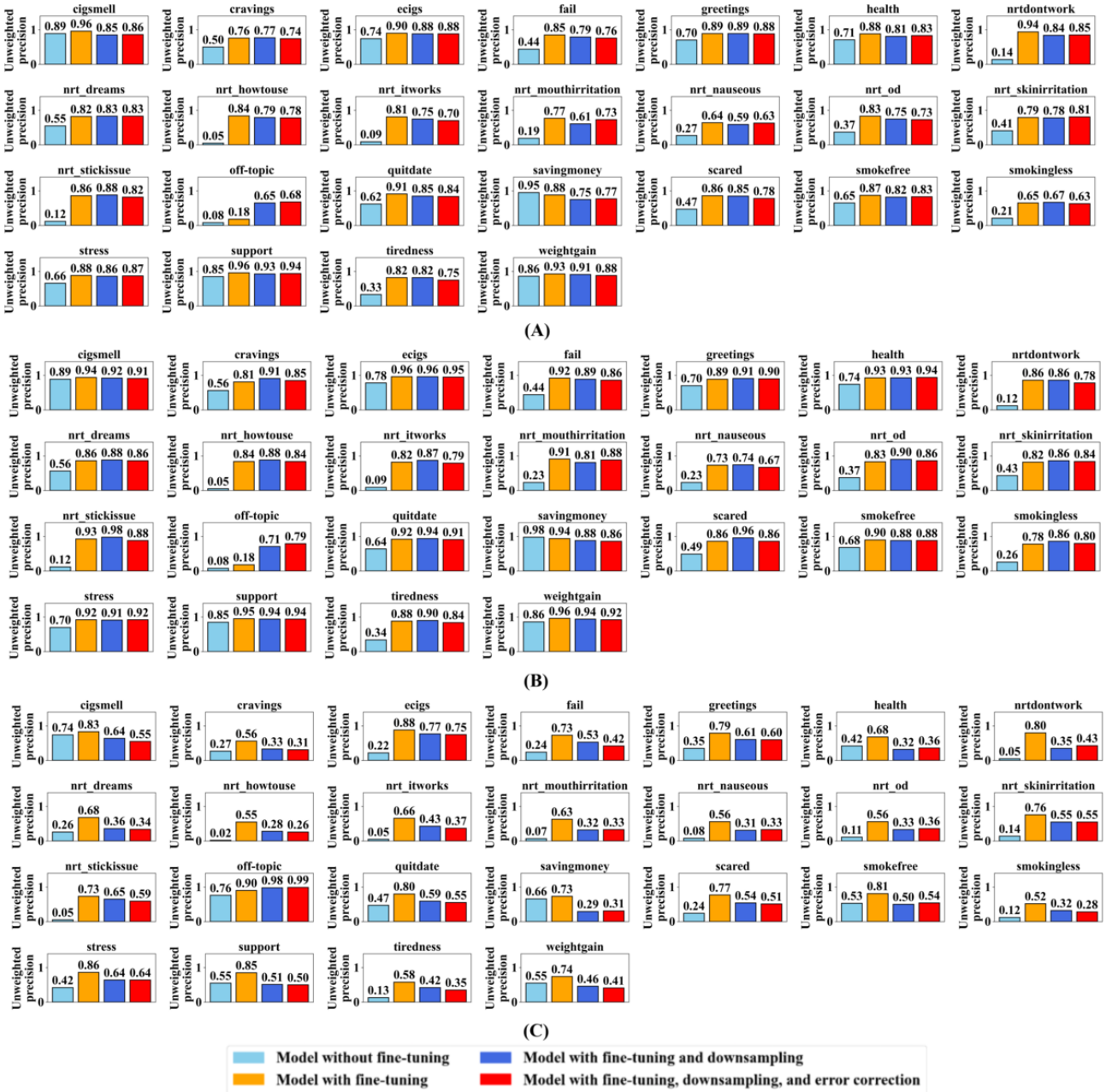

In Figure 4A, using the downsampled uncorrected test dataset, for the method without fine-tuning, the precision for most intents was around 0.47, with a few categories slightly exceeding that score but many falling below it, indicating weak overall performance. With fine-tuning, the precision for most intents improved to around 0.82. However, the dominant off-topic intent achieved a low precision of 0.18, reflecting a classification bias. Its large number of training samples caused the model to overpredict it, reducing precision. In effect, other important intents were often misclassified as off-topic, which is concerning because this can suppress the detection of key intents. The method of fine-tuning and downsampling of the dominant off-topic intent yielded close to a 0.80 precision score for most intents, but certain intents fell below the 0.80 mark. The final approach, combining fine-tuning, downsampling, and error correction, achieved precision comparable to the downsampling-only method.

In Figure 4B, using the downsampled corrected test dataset, the method without fine-tuning exhibited a precision of approximately 0.49 across most intents. With fine-tuning, precision improved to around 0.85 for most intents, although the off-topic intent category continued to show a low precision of 0.18. Fine-tuning with downsampling resulted in many precision scores close to 0.89. Adding error correction yielded an average precision of 0.86. Thus, the downsampling-only method demonstrated more robust performance on precision.

In Figure 4C, using the full dataset, the method without fine-tuning exhibited a precision of approximately 0.30 for most intents. With fine-tuning, precision improved to around 0.72 for most intents, although the dominant off-topic intent achieved

XSL•FO
RenderX

a higher precision of 0.90. Fine-tuning with downsampling resulted in a precision mainly close to 0.48. The final method of combining fine-tuning, downsampling, and error correction achieved an average precision of 0.46. The large volume of off-topic intent samples in the full test dataset led to many off-topic intents being misclassified as in-domain intents, lowering the precision scores for the last two methods.

Recall for Each Intent Category

Figure 5 presents the unweighted recall scores for each of the 25 intents across the different training methods. Unweighted recall shows how accurately the model picks the actual intents and requires balanced performance across intents. Evaluation results are shown in Figure 5A using the downsampled uncorrected test dataset, in Figure 5B using the downsampled corrected test dataset, and in Figure 5C using the full unmodified test dataset.

Figure 5. Unweighted recall scores for the 4 large language model training methods used to detect intents in smoking-cessation messages in this study (United States). Models were evaluated on the (A) downsampled uncorrected test dataset, (B) downsampled corrected test dataset, and (C) full test dataset.

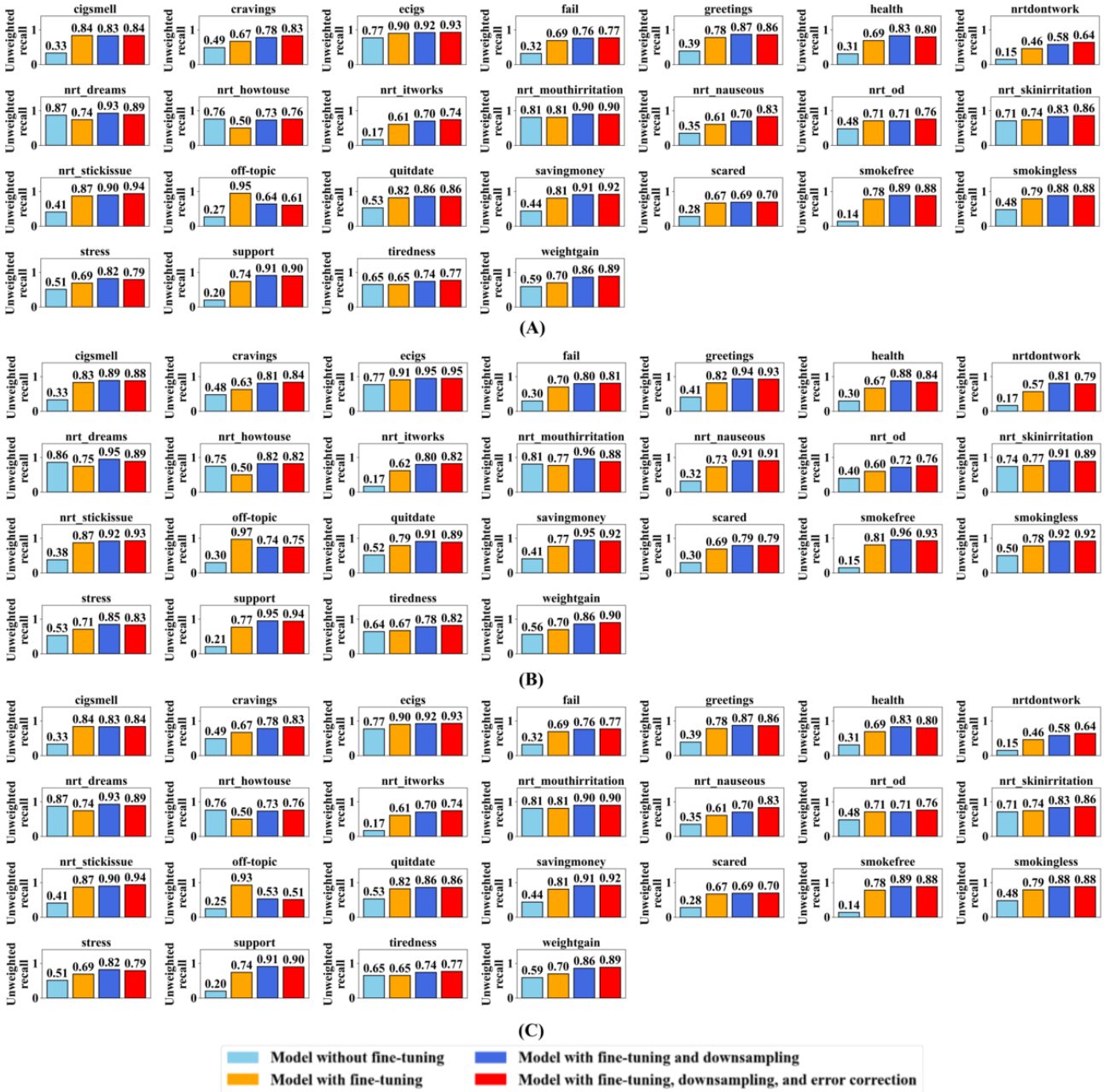

In Figure 5A, using the downsampled uncorrected test dataset, the method without fine-tuning achieved a recall of approximately 0.46 for most intents, with some slightly higher but many lower. Fine-tuning improved recall to around 0.73 for most intents, although several still fell below this, while the dominant off-topic intent category reached 0.95 recall. Adding downsampling, many intents achieved a recall close to 0.81, though some remained lower. Adding error correction as well produced slightly better recall across several intents, averaging 0.82, indicating strong overall performance.

XSL•FO
RenderX

In Figure 5B, using the downsampled corrected test dataset, the method without fine-tuning achieved a recall of approximately 0.45 for most intents. With fine-tuning, recall improved to around 0.73 for most intents, while the dominant off-topic intent category reached 0.97. Adding downsampling, many intents achieved recall close to 0.87, though some remained lower. Adding error correction as well achieved an average recall of 0.87 across intents.

In Figure 5C, using the full test dataset, the method without fine-tuning achieved a recall of approximately 0.46 for most intents. With fine-tuning, recall improved to around 0.73 for most intents, while the dominant off-topic intent category reached 0.93. Adding downsampling, many intents achieved a recall close to 0.80. Adding error correction as well achieved the highest average recall of 0.82 across intents, demonstrating strong overall performance on noisy, real-world data.

F_1-Scores for Each Intent Category

Figure 6 presents the unweighted F_1-scores for each of the 25 intents across the different training methods. Unweighted F_1-score combines precision and recall and emphasizes balanced performance across intent categories. Evaluation results are shown in Figure 6A using the downsampled uncorrected test dataset, in Figure 6B using the downsampled corrected test dataset, and in Figure 6C using the full unmodified test dataset.

Figure 6. Unweighted F_1-scores for the 4 large language model training methods used to detect intents in smoking-cessation messages in this study (United States). Models were evaluated on the (A) downsampled uncorrected test dataset, (B) downsampled corrected test dataset, and (C) full test dataset.

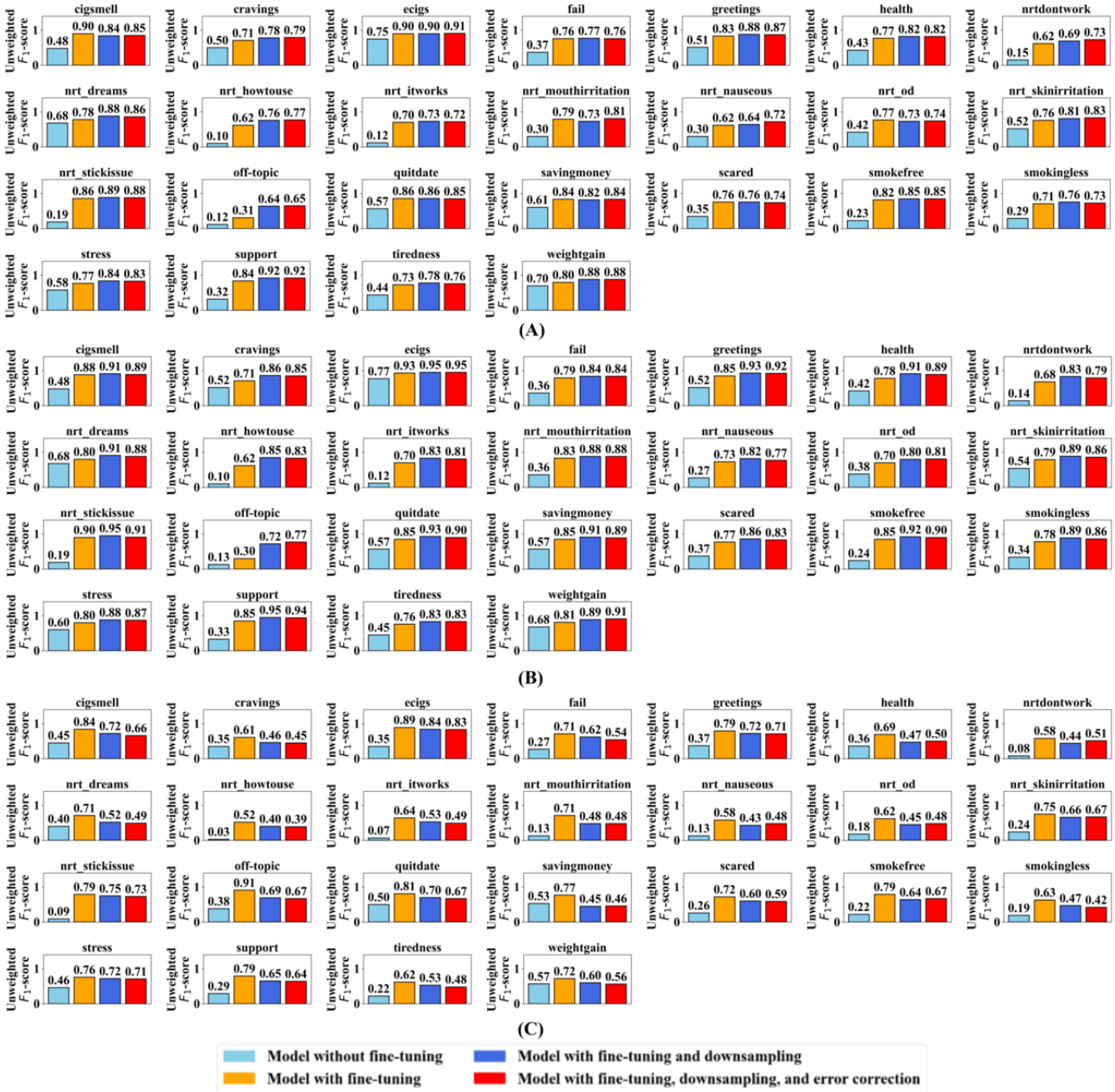

In Figure 6A, using the downsampled uncorrected test dataset, the model without fine-tuning achieved an F_1-score of approximately 0.40 across most intents, although several intents scored lower. Fine-tuning increased the F_1-score to about 0.75,

XSL•FO
RenderX

but many intents scored lower, and the off-topic intent category scored about 0.31. The downsampling method with or without error correction achieved an F_1-score of about 0.80, though some intents were lower.

In Figure 6B, using the downsampled corrected test dataset, the model without fine-tuning achieved an F_1-score of approximately 0.41 across most intents. Fine-tuning improved the F_1-score to about 0.77, though many intents scored lower, and the off-topic intent category scored about 0.30. The downsampling method achieved an F_1-score of about 0.88, but some intents remained lower, while adding error correction achieved F_1-scores of about 0.86 across intents.

In Figure 6C, using the full test dataset, the model without fine-tuning achieved an F_1-score of about 0.29 across most intents, with several lower. Fine-tuning increased the F_1-score to about 0.72, though many intents scored lower, and the off-topic intent category scored above 0.91. The downsampling method yielded an F_1-score of 0.58, while adding error correction yielded an F_1-score of 0.57, but several intents were lower.

Total Statistics on Model-Human Disagreement

Figure 7 shows the model-human disagreement across our 4 training methods. Model-human disagreement provides a comprehensive view of a model's accuracy at judging a message's intent relative to human judges of that intent. The left side of each figure includes either the full or downsampled off-topic intents in the test dataset, while the right side excludes the off-topic intents. Results are shown in Figure 7A using the downsampled uncorrected test dataset, in Figure 7B using the downsampled corrected test dataset, and in Figure 7C using the full unmodified test dataset.

XSL•FO
RenderX

Figure 7. Model-human disagreement, including and excluding the dominant off-topic intent category in the test dataset, for the 4 large language model training methods studied (United States). Methods were evaluated on the (A) downsampled uncorrected test dataset, (B) downsampled corrected test dataset, and (C) full test dataset.

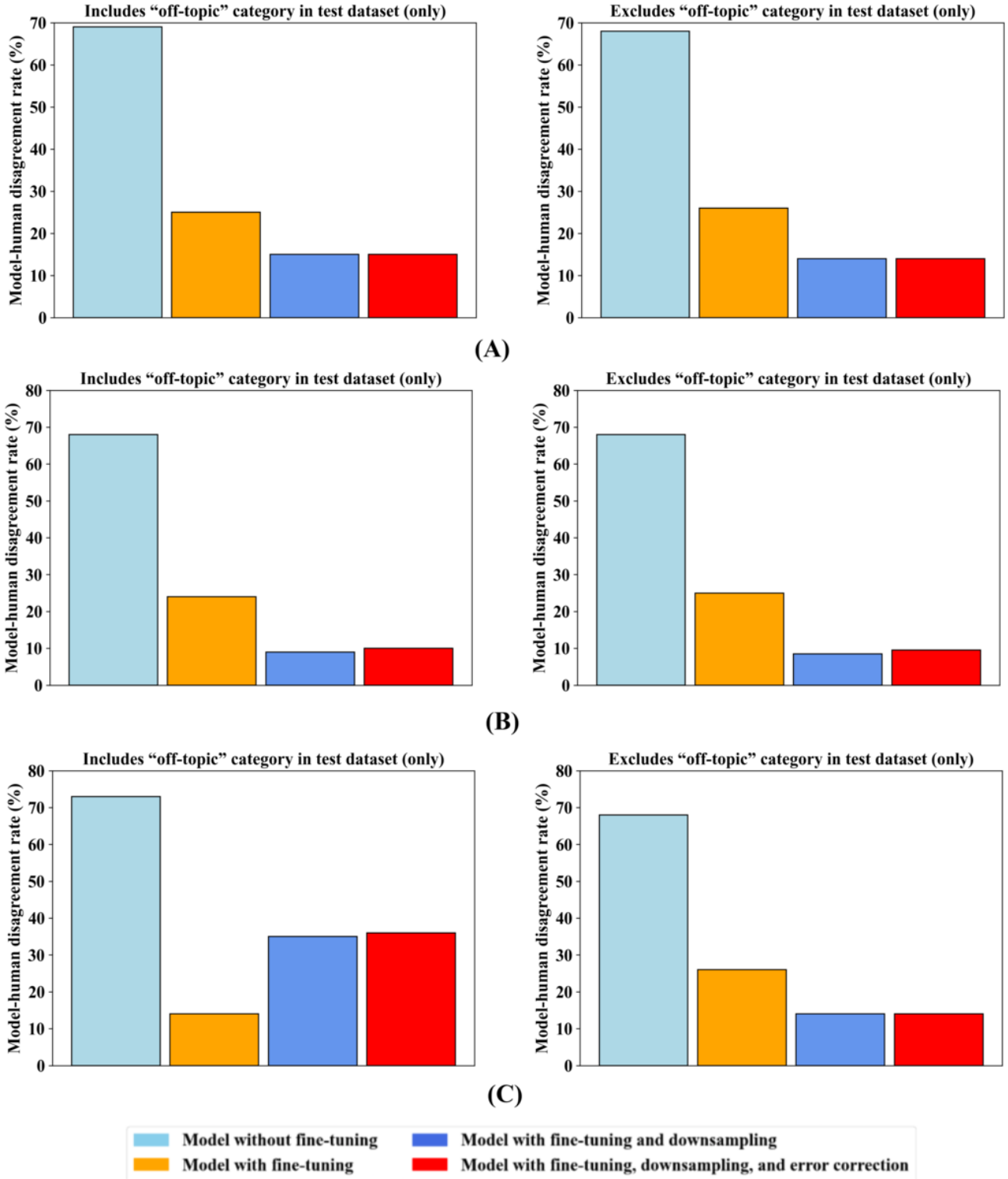

In Figure 7A, using the downsampled uncorrected test dataset, the left side includes off-topic intents in the test dataset. The method with no fine-tuning yielded about 69% model-human disagreement. Fine-tuning alone lowered this to 25%. Fine-tuning with downsampling with or without error correction lowered the disagreement to about 15%. The right side of Figure 7A excludes the off-topic intents from the test dataset. It shows the method without fine-tuning yielded approximately 68% model-human disagreement. Fine-tuning alone reduced this to 26%. Fine-tuning with downsampling with or without error correction lowered the disagreement to about 14%. The results, including or excluding the off-topic intent samples, were similar because the number of off-topic samples in the downsampled test dataset was small enough to have minimal impact.

In Figure 7B, using the downsampled corrected test dataset, the left side includes off-topic intents in the test dataset. The method with no fine-tuning showed about 68% model-human

XSL•FO
RenderX

disagreement. Fine-tuning alone lowered this to 24%, fine-tuning with downsampling lowered it to 9%, and adding error correction yielded 10%. The right side of Figure 7B excludes the off-topic intents from the test dataset. The method without fine-tuning yielded 68% model-human disagreement. Fine-tuning alone lowered it to 25%, fine-tuning and downsampling lowered it to 8.5%, and adding error correction yielded 9.5%.

In Figure 7C, using the full test dataset, the left side includes off-topic intents in the test dataset. The method without fine-tuning exhibited 73% model-human disagreement. Fine-tuning alone reduced this to 14%. In contrast, fine-tuning with downsampling yielded 35%, and adding error correction yielded 36%. So, on real-world noisy data, downsampling yielded more model-human disagreement than no downsampling. This occurred because downsampling the off-topic intent category trained the model to recognize rarer but important in-domain intents. The higher model-human disagreement for the last two methods was primarily due to off-topic samples being misclassified as in-domain intents. At first glance, it may appear that fine-tuning alone performs better because it produces less model-human disagreement in the original noisy data. However, this is largely an artifact of the imbalanced dataset: both the training and test datasets contain a disproportionately large number of off-topic intent samples and, since both are similarly dominated by off-topic, the disagreement rate becomes artificially low, masking poor performance on the more important in-domain intents. The right side of Figure 7C, which excludes the off-topic intents from the test dataset, provides a different and, we believe, more accurate picture. Here, the model without fine-tuning yielded 68% model-human disagreement. Fine-tuning alone lowered this to 26%, and downsampling with or without error correction method lowered it to 14%. These results indicate that when the off-topic intents are excluded from the test dataset, the downsampling-only and the downsampling and error correction methods achieve the lowest model-human disagreement on the 24 important domain-specific intents.

Statistics on Model-Human Disagreement

Figure 8 shows the percentages of model-human disagreement by intent category for our 4 training methods, which we will refer to as model mispredictions because the goal of this analysis is to understand when and how a model confused the different intent categories. This information provides insight into each training method's specific weaknesses. In Figure 8A, evaluation was conducted using the downsampled uncorrected test dataset. In Figure 8B, evaluation was conducted using the downsampled corrected test dataset. Finally, in Figure 8C evaluation was conducted using the full unmodified test dataset.

XSL•FO
RenderX

Figure 8. Model-human disagreement by intent category for the 4 large language model training methods studied (United States). Methods were evaluated on the (A) downsampled uncorrected test dataset, (B) downsampled corrected test dataset, and (C) full test dataset.

In Figure 8A, using the downsampled uncorrected test dataset, the method without fine-tuning caused the model to mispredict the intent nrt_howtouse for many other intent categories. For some categories, over 50% of instances were mispredicted as nrt_howtouse; for 17 out of 25 categories, this incorrect prediction dominated. This demonstrates that the model without fine-tuning was heavily biased toward nrt_howtouse and failed to distinguish it from other intents. A similar issue appeared with the fine-tuned model, but the bias shifted toward the dominant off-topic intent category. Here, 24 out of 25 categories were most commonly and incorrectly mispredicted as off-topic, and this occurred approximately 30% of the time, reflecting the overrepresentation of off-topic samples. This highlights a key limitation of fine-tuning alone: without addressing category imbalance, the model struggles to predict important but underrepresented intent categories accurately. In contrast, the fine-tuning with the downsampling method produced a more balanced pattern of mispredictions. No single intent category was mispredicted, as evidenced by the absence of any overwhelmingly frequent color in our bar graph; only minor weaknesses remained. A similar pattern was observed for the fine-tuning, downsampling, and error correction method, with no single intent category mispredicted. These results demonstrate that downsampling effectively reduced bias toward mispredicting dominant intent categories and promoted more equitable performance.

In Figure 8B, using the downsampled corrected test dataset, the methods without and with fine-tuning alone showed similar patterns, with nrt_howtouse and off-topic intents being most frequently mispredicted, respectively. In comparison, the methods adding downsampling alone, or downsampling with

XSL•FO
RenderX

error correction, showed a more balanced pattern of mispredictions, with no single intent category dominating, the only exception being the support intent category, with 7 different intents frequently misclassified as support, although in each case less than 5% of the time.

In Figure 8C, using the full test dataset, the methods without and with fine-tuning alone showed similar patterns, with nrt_howtouse and off-topic intents most frequently mispredicted, respectively. Adding downsampling and then also adding error correction produced a more balanced pattern of mispredictions, except for the support intent, which was mispredicted as other categories. Examples in which the model mispredicted other categories as support include: “Those are great reasons,” which was human-annotated as off-topic, “Nicotine addiction is like any other drug out there and the side effects are bad. Sorry to say,” which was human-annotated as health, and “Patches will help greatly. You can do it,” which was human-annotated as nrt_itworks. These results show that, in certain misclassification cases, the model produced a generic smoking-cessation support response rather than an intent-specific one. While such responses may provide baseline assistance, they do not substitute for the intended goal of accurate intent-specific classification.

Model-Human Disagreement and Human Error Correction

Human errors in the annotation of message intents were inevitable given the dataset’s very large size. The data annotators were highly trained undergraduate students in public health fields who demonstrated high interrater reliability in training before annotating the study dataset. They met weekly with supervisors to resolve inconsistencies in real time. Despite this rigorous training and oversight, a small number of annotation errors were observed. Therefore, a domain expert thoroughly reviewed the model-human disagreements in the training, evaluation, and test sets to identify and correct human annotation errors.

Table 3 presents statistics on model-human disagreement on message intents from model mispredictions or human annotation errors or both. These statistics offer insights into the accuracy of human annotators and the potential of a large language model to sometimes outperform humans. Model-human disagreement was assessed in the downsampled datasets only, due to lack of resources to search for disagreement in the full dataset.

Table . Causes of model-human disagreement in the detection of intents in this study (United States).

Dataset (downsampled)	Model-human disagreement	Humans wrong, model correct	Model wrong, humans correct	Both humans and model wrong
Train	0.7%	59%	30%	11%
Evaluation	15%	42%	47%	11%
Test	15%	39%	55%	6%

In the downsampled training dataset (75% of the nontest data), the model and human annotators disagreed on the message intent 0.7% of the time. The human annotators were incorrect in 59% of these disagreed instances, with the model being accurate. The model mispredicted in 30% of the disagreed instances, with both human and model erroneous in the remaining 11% of the instances. Thus, in total, humans were wrong in 70% of the disagreed instances (59%+11%), amounting to 0.13% of the total data, 0.22% of the full training data, and 0.47% of the downsampled training data. So, human annotators were rarely wrong. While the downsized training dataset was not found to include many human annotation errors, it included some, indicating that retraining the model on a corrected dataset would lead to performance improvements.

In the downsampled evaluation dataset (25% of the nontest data, reduced further after downsampling the dominant off-topic intents), the model and human annotators disagreed on the message intent 15% of the time. Human annotators were wrong in 42% of these instances, with the model being accurate. The model mispredicted in 47% of these instances, and both were erroneous in the remaining 11% of the instances. Thus, in total, humans were wrong in 53% of the disagreed instances (42%+11%), amounting to 0.76% of the total data, 3.78% of the full evaluation data, and 8% of the downsampled evaluation data.

Finally, in the downsampled test dataset (holdout data from groups 30 to 36, reduced further after downsampling off-topic intents), model and human annotators disagreed on the message intent 15% of the time. Human annotators made errors in 39% of these cases, with the model being accurate. The model mispredicted in 55% of these instances, and both were erroneous in the remaining 6% of the instances. Thus, in total, humans were wrong in 45% of the disagreed instances (39%+6%), amounting to 0.50% of the total data, 2.5% of the full test dataset, and 7% of the downsampled test data.

Table 4 presents statistics on the corrections made to the downsampled data after manual review of the model-human disagreement in intent detection by category. In total, 1077 samples were corrected, accounting for 1.4% of the total dataset and 3% of the downsampled data, indicating human annotation errors and corrections were relatively rare. In the downsampled training dataset, in 102 cases, the support, smokefree, and greetings intents were corrected more frequently than other intents. In the downsampled evaluation dataset, in 588 cases, the support, smokefree, off-topic, cravings, and greetings intents were corrected more frequently than other intents. In the downsampled test dataset, in 387 cases, the smokefree, off-topic, quitdate, nrt_dontwork, nrt_itworks, and greetings intents were corrected more frequently than other intents.

XSL•FO
RenderX

Table . Data adjustments after reviewing model-human disagreement in smoking-cessation message intents in this study (United States).

Intent category	Adjustments to downsampled training dataset (n=102)	Adjustments to downsampled evaluation dataset (n=588)	Adjustments to downsampled test dataset (n=387)
cigsmell	0	3	5
cravings	9	43	14
ecigs	0	3	2
fail	1	23	7
greetings	10	53	43
health	1	26	18
nrt_dontwork	3	30	20
nrt_dreams	1	4	1
nrt_howtouse	0	21	10
nrt_itworks	3	17	15
nrt_mouthirritation	0	7	1
nrt_nauseous	0	9	5
nrt_od	0	5	3
nrt_skinirritation	0	3	3
nrt_stickissue	0	0	2
off-topic	10	71	33
quitdate	0	14	16
savingmoney	0	10	3
scared	3	6	9
smokefree	17	97	60
smokingless	0	10	1
stress	1	14	7
support	43	116	104
tiredness	0	2	2
weightgain	0	1	3

A domain expert handled the human annotation error correction and found that corrections were often needed when two intents were expressed in a single message. In such cases, the human annotators had been trained to prioritize certain intents over others to optimize the chatbot's responses, but they apparently found this prioritization challenging. For instance, if a person said they quit but were craving, this should have been annotated as cravings, not smokefree, to prioritize a response to help with craving. If a person mentioned they were going to quit on day 1, this should have been annotated quitdate, not smokefree, to prioritize quitdate help. Messages about failing and setting a new quit date should have been annotated as fail, not quitdate, to provide guidance after a fail. If a message started with a greeting, for example, good morning, but then provided quitting support, it should have been annotated as support to prioritize a support-related response over a generic greeting. If a message discussed quitting but also reported an issue with NRT, it should have been annotated as NRT-related to assist with NRT use. If a person mentioned no longer needing NRT because they had successfully quit using NRT, it should have been annotated as nrt_itworks, not as a generic quit. All annotation corrections were conducted blind to the model predictions; that is, the domain expert reviewed and adjudicated the intent annotations independently of model outputs.

Overall, this analysis revealed the value of using large language models not only for classifying message intents but also for uncovering human errors in annotating these intents. The relatively high rate of model correctness in cases of human misannotations reinforces the idea that annotation corrections followed by model retraining can result in more accurate assessments of the model's performance. However, in our work, the observed impact on model performance was minimized because we only had the resources to correct human annotation errors in the downsampled data, not the full data. Additionally, as human-model disagreement and human annotation correction often involved a message with two intents, different methods of addressing this situation should be considered, as will be discussed below.

XSL•FO
RenderX

Discussion

Principal Findings

Large language models have become the new state-of-the-art models in natural language processing due to their ability to recognize complex linguistic patterns and subtle semantic distinctions. Leveraging large language models for classification tasks like recognizing message intents can significantly enhance performance over earlier models. However, our analysis indicates that large language models may be insufficient for intent recognition in mobile health interventions when user messages are short and subtle wording differences can profoundly affect meanings. We studied messages in mobile health smoking-cessation support groups, where there was a profound difference between saying "quit" or "not yet quit," "NRT works" or "NRT doesn't work," and "quit day 1" or "quit day 21," and our off-the-shelf large language model was not always able to detect these nuances. The model's performance was unacceptably poor until it was fine-tuned on a domain-specific dataset, which required laborious human annotation of message intents. The model before fine-tuning performed poorly because the domain-specific text often contained unique terminology, implicit meanings, and contextual subtleties that a general-purpose model was unable to fully grasp without exposure to relevant training data.

We found that fine-tuning the large language model on a domain-specific dataset led to substantial performance gains. The fine-tuning process allowed the model to familiarize itself with the characteristics of the dataset and better capture the intents behind the messages. Our fine-tuned large language model substantially outperformed the same model without fine-tuning. But a key challenge we identified during fine-tuning was data imbalance, particularly due to the dominance of off-topic message intents unrelated to the focal domain of smoking cessation. This imbalance caused the model to overpredict this majority (dominant) off-topic intent, degrading performance on all other intents, even though we assigned more importance to underrepresented categories when the model made mistakes. To address this issue, we implemented downsampling of the majority off-topic intent before fine-tuning, which increased unweighted F_1-scores (overall performance) by 7% and 14% in the downsampled uncorrected dataset and downsampled corrected dataset, respectively, indicating a substantial improvement in balanced performance across all intents. Individual intent-level F_1-scores also demonstrated consistent gains. When we evaluated our methods on the full uncorrected test dataset, the unweighted F_1-score decreased when we added downsampling, primarily due to the large number of off-topic intents in that full dataset, which adversely affected precision. However, downsampling improved unweighted recall by approximately 10% in that full dataset, indicating more robust performance in correctly identifying in-domain, clinically relevant intents.

Another critical insight from our study was the identification of dataset inconsistencies and noise, largely arising from human annotation errors in intent labeling. To address this issue, we examined the model-human disagreement, identified human annotation errors, and corrected them. Even when the downsampled test dataset remained uncorrected, training on the corrected dataset slightly increased unweighted recall, suggesting improved generalization to noisy real-world data, and the F_1-score was preserved. When the downsampled test dataset was itself corrected, the model trained with error correction achieved the same unweighted recall as no error correction, but the latter yielded approximately 3.5% higher precision, slightly improving the F_1-score. This finding indicates that the no error correction method generalized marginally better to the corrected dataset than the error correction method.

This last outcome may indicate the error correction process, while beneficial for improving unweighted recall under noisy conditions, may have reduced useful diversity or led to minor overfitting. Another reason could be that most corrections involved the majority of intents such as support, greetings, and smokefree, and these samples were corrected into rarer intents. As a result, the error-correction method focused slightly more on the rare intent categories, achieving similar recall as the downsample-only method, but lower recall on the majority classes, reducing precision and the overall F_1-score. On the full test dataset, the unweighted F_1-scores of downsample-only and downsample with error corrections were comparable; however, error correction achieved a 2.5% higher unweighted recall and thus showed enhanced identification of in-domain intents under real-world conditions, where incoming data are noisy.

Comparison With Previous Work

Previous work [20] studied intent classification with the same dataset, comparing random forest to BERT, which in 2018 was a revolutionary large language model introduced by Google. The study dataset, annotated by humans to indicate the different message intents, was identical to that used here, except the two smallest intent categories were combined and some duplicate messages were inadvertently retained. The findings showed that BERT, fine-tuned on the domain-specific annotated dataset, outperformed random forest. BERT yielded unweighted and weighted F_1-scores of 0.67 and 0.84, respectively, which were reasonable scores but inadequate for implementation of the large language model in our mobile health intervention. In this study, we used a more recent public-domain large language model, Llama-3 8B (Meta), to perform intent classification across 25 message intent categories. When fine-tuned on the same domain-specific annotated dataset, Llama-3 8B achieved unweighted and weighted F_1-scores of 0.72 and 0.86, respectively, on the full test dataset, outperforming the previously reported BERT model by 7% in unweighted and 2% in weighted F_1-scores, respectively. However, these improvements should be interpreted as indicative rather than strictly comparable, due to differences in dataset structure and inadvertent duplication in the prior study.

We then evaluated our last two methods: (1) fine-tuning with downsampling and (2) fine-tuning with downsampling and error correction. When assessed on the downsampled corrected test dataset, these approaches achieved unweighted F_1-scores of 0.88 and 0.86 and weighted F_1-scores of 0.91 and 0.90, respectively. Relative to the BERT baseline, these are

XSL•FO
RenderX

improvements of 31% and 28% in unweighted F_1-scores and 8% and 7% in weighted F_1-scores. We note that these improvements are only indicative, since differences in dataset prevent a strict empirical comparison. However, when evaluated on the full test dataset, our final two methods achieved unweighted F_1-scores of 0.58 and 0.57, and weighted F_1-scores of 0.66 and 0.65, lower than the previously reported BERT performances. This reduction is primarily attributable to the dominance of the off-topic intent category in the full test dataset. Misclassification of off-topic intents as in-domain intents reduced overall precision and the resulting F_1-scores. Despite reduced precision, our final two methods achieved substantially higher unweighted recall scores of 0.80 and 0.82, compared with 0.71 for the prior work, representing a 13% - 16% recall improvement. This finding indicates that our approach is more effective at identifying in-domain, clinically meaningful intents, which is our central objective. However, our model's weighted F_1-score of 0.65 is substantially lower than the previously reported 0.84 for the BERT baseline. This indicates a clear reduction in overall precision and F_1-score performance in a real-world setting, highlighting the trade-off between improved recall for in-domain intents but decreased overall classification accuracy. In addition, a strict comparison with the previous work is not feasible because it used 24 intents after merging two small ones (instead of our 25) and inadvertently contained duplicated data. These differences alter the ground truth distributions and, therefore, the performance differences should be interpreted as observed rather than strictly comparable.

Implementation Feasibility, Generalizability, and Potential Risks

Fine-tuning the large language model based on our final method took approximately 2.5 hours, and each input was processed in approximately 40 milliseconds, with a peak GPU memory usage of 8.2 GB. The model was able to handle approximately 493 tokens per second, demonstrating fast and efficient performance suitable for real-time intent detection. Human oversight was incorporated for human annotation error correction, to improve model accuracy and reliability. This error correction approach was straightforward. Model-human inconsistencies in intent classifications were identified and then systematically reviewed to identify potential human annotation errors, with error corrections performed by a domain expert who was blind to the model predictions. If a domain expert with knowledge of the annotation process or domain context is involved, this error-correction approach should be reproducible.

Regarding the generalizability of our final model, it should perform well on new data that mainly includes the same 25 intent categories. Large language models handle shifts in language use through embedding-based representations, which capture semantic meaning rather than relying on exact word sequences. Within the same domain, this approach mitigates the impact of data drift, ensuring that the model's predictions remain relevant and accurate even if specific word usage has evolved over time. However, if the new data contain different intent categories, it would require reannotation or model adaptation.

The largest failure in our intent-detection model would be if an intent category is detected as a different and unsuitable intent category. For instance, the message "Started my 5th day the struggle is real today" refers to cravings, but the model might mispredict it as smokefree. While it is true that on the 5th day the person is smokefree, the main concern is cravings as there is struggle, so missing the craving intent may prevent the chatbot from responding with the needed support for cravings. Another example is that the message "I'm restarting Friday" refers to fail, but the model might mispredict it as quitdate. Although the message mentions Friday as the quit date, it mainly indicates the person failed to quit and needs help with avoiding a future failure, not with setting a quit date. These examples highlight limitations in intent detection and suggest that our future work should aim to provide appropriate responses even if the predicted intent is incomplete or erroneous or multiple intents are conveyed in a single message.

In our analysis, we downsampled the dominant off-topic intent category to help the model better learn the remaining 24 in-domain intent categories. As a result, the model's performance on off-topic is lower, and these messages are more likely to be misclassified as one of the in-domain categories. In practice, this means that some off-topic messages may receive an in-domain, prevetted smoking-cessation response. Although off-topic messages typically fall outside the scope of smoking cessation and may include general social conversation, misclassifying them as cessation-related is unlikely to cause harm as people signed up for a smoking-cessation support group. Moreover, cessation responses are empathetic, supportive, and encouraging, and therefore still appropriate in tone. However, this high rate of false positives should be considered a potential barrier to user adoption and trust, rather than inherently unproblematic, as we do not have supporting user data to assess its real-world impact.

We acknowledge that repeated false-positive delivery of cessation-oriented content to users engaging in social small talk may risk user disengagement. From a deployment perspective, we consider two complementary definitions of signal-to-noise. When true in-domain detections are considered signal and off-topic detections are considered noise, based on our full test dataset, our downsampling methods would result in nearly 4700 off-topic messages being misclassified as in-domain, compared to 5500 that are truly in-domains. This corresponds to nearly one unnecessary cessation intervention for every appropriate one, resulting in roughly a 1:1 signal-to-noise. Where both true in-domain detections and off-topic detections are considered signals because off-topic messages appropriately redirect people to focus on the smoking-cessation domain, our methods would result in nearly a 2:1 signal-to-noise. In sum, in the real deployment, the model could send from 1 to 2 useful cessation messages for every unnecessary one. This means users engaging in casual or social conversation could frequently receive irrelevant intervention messages. This potential risk should be continuously monitored during real-world deployment to mitigate unintended user disengagement.

It is true that recognizing off-topic intent messages is an important goal for this deployment, as detecting such messages can possibly help redirect focus in the smoking-cessation

XSL•FO
RenderX

domain. However, detecting in-domain intent categories is a higher-priority objective. The original dataset contained a large proportion of off-topic messages, which biased the model toward off-topic detection and, in turn, reduced performance on in-domain messages. To address this tradeoff, we prioritized in-domain detection by weighting it more heavily, while still considering off-topic detection as a secondary goal. Consequently, in our downsampling approach, we reduced the number of off-topic messages to improve the model's performance on in-domain intent categories.

Limitations

Our final method of large language model training that included fine-tuning, downsampling, and error correction achieved impressive performance in our dataset, but it had several limitations. First, a continued challenge is complex messages involving two or more intent categories. For instance, a user stated, "Tomorrow is my day 1 and I haven't smoked for over 5 hours today," which reflects two intents: quitdate (planning to quit tomorrow, ie, day 1) and smokingless (starting to quit but still smoking some). While the main message intent was likely to alert others to tomorrow being the quitdate, the model predicted smokingless. Similarly, in "I'm not going to quit smoking on my busiest day," the message expressed intent to quit on a specific day (quitdate) but also implied being smokefree on that day; however, the correct intent category was quitdate and the model mispredicted it as smokefree. These examples highlight the inherent ambiguity in single-intent annotations when messages express multiple intents. Because the dataset does not account for multi-intent expressions, the model's predictions occasionally diverged from the human annotations despite both being semantically reasonable.

Another study limitation was that 2 of our 4 methods involved downsampling the off-topic intent category in the training, evaluation, and test datasets, based on the average number of samples of all other intents. Downsampling the off-topic intent category consistently across datasets ensured the model learned from a representative distribution of focal intents. However, we recognize that altering the test dataset may be problematic. Therefore, we also evaluated performance on the complete, uncorrected test dataset which likely offers a more accurate reflection of how the model would perform in real-world scenarios.

A third potential limitation of our study is that, in our final method, we corrected human annotation errors. However, we only had the resources to assess the much smaller downsampled dataset for annotation errors, and we only examined instances where the model and the human annotator disagreed. This approach introduces the risk of circular bias, where the ground truth is shifted toward the model's predictions. This may inadvertently penalize natural human variation and inflate performance metrics relative to what the model would achieve on truly independent unmodified data. Moreover, since our error correction method focused only on messages where the model disagreed with human annotations, instances where both the model and human annotators were incorrect (shared errors) were not reviewed, which may leave residual annotation errors in the dataset. As a result, even the corrected dataset may be partially aligned with the model's predictions rather than fully independently verified, which could influence performance estimates. Once again, performance on the full uncorrected test dataset likely provides a more realistic estimate of generalization to real-world conditions.

Fourth, because the training and evaluation data splits were stratified based on the distribution of intent categories rather than at the group or user level, messages from the same user could appear in both datasets, which may introduce data leakage and allow the model to partially learn user-specific writing styles, potentially overestimating performance.

Fifth, while we downsampled the dominant off-topic intent, the dataset remained imbalanced. Rare intents still had very few examples, whereas intents such as greetings and support had many examples, which reduced model performance for underrepresented intents. Finally, our off-the-shelf model with no fine-tuning produced 991 intent predictions that did not map onto any of our existing intent categories. To handle this, we included these samples as misclassifications in the final scores.

Future Work

In future work, we plan to explore how to address multiple intents in messages, for instance, by allowing the model to detect the most probable 2 - 3 intents. We will also explore ways to incorporate general domain knowledge into the model training process, for example, patterns or semantic rules that are not easily learnable from data alone, but can be integrated into the model's decision-making. We may also seek to expand our dataset by collecting additional real-world examples of underrepresented intent categories and/or creating synthetic examples using generative artificial intelligence (AI). However, it will be crucial to ensure that any synthetic data mimics real-world patterns and does not introduce bias. In addition, we may explore temporal analyses using classifiers to address possible temporal variability in the data.

In additional future work, we will explore threshold tuning. For this study, we did not perform threshold tuning or probability calibration because the Llama-3 8B model inherently selects the most probable category based on its prediction probabilities. We were concerned that adjusting thresholds could suppress predictions for certain intent categories when confidence scores were low. In our domain, where many intents are semantically related, the model must still assign a category even with lower confidence to remain contextually helpful. Setting a high threshold might cause the model to abstain from predicting any intent, which would be less useful to us. Although a low-confidence prediction may occasionally be incorrect, it should still tend to remain on-domain and informative. Notwithstanding, threshold tuning should be considered. Specifically, the off-topic intent category can be almost 50% misclassified as on-topic or domain-related, leading to on-topic chatbot responses, and this could cause user disengagement. So, future work should explore confidence-based thresholding or probability calibration to reduce false-positive domain-related chatbot responses. By implementing a mechanism to silence the chatbot when the predicted intent probability is low (eg, not likely to be on-topic), the signal-to-noise ratio could be improved in live deployment. This approach may help minimize

XSL•FO
RenderX

unnecessary domain-specific cessation messages, reduce user disengagement, and enhance overall safety and usability in real-world clinical settings.

We chose the arithmetic mean as a simple and transparent method to reduce the dominance of the off-topic intent category while retaining sufficient examples of it for model learning. Using the mean ensured that the downsampled off-topic intent remained large enough to capture diversity within its messages. But we acknowledge that downsampling to the mean likely resulted in a reduced off-topic intent category that was still significantly larger than many minority intent categories. Thus, alternative strategies such as downsampling to the median could be explored in future work.

Our ultimate research aim is to create an intent detection model that will also respond to messages based on the intents conveyed. Therefore, we still need to create the model's response capabilities, yielding a complete interactive chatbot. We hope this chatbot will improve engagement in our smoking-cessation support groups by ensuring at least the chatbot is available to respond 24/7 if no human is online. However, we acknowledge that our mobile health support group intervention is not a one-size-fits-all approach; for instance, it is not suitable for people who are not interested in mobile health and seek in-person support exclusively.

To translate technical performance into operational impact, we hope to test our chatbot in a randomized controlled trial to see if it significantly improves the outcomes of our mobile health intervention for smoking cessation. We plan to assess the potential lift in engagement due to the inclusion of a chatbot that accurately detects and responds to support group members' messages. We also plan to assess other intervention outcomes, including cessation success.

Conclusions

Mobile health smoking-cessation interventions that provide real-time supportive interactions have the potential to deliver timely advice and improve health outcomes. Detecting user intents from their utterances can enhance personalization and engagement by identifying individual needs and enabling intelligent responses through chatbots, that is, AI agents. However, existing chatbots for smoking-cessation interventions are limited in their intent detection and response capabilities. In this study, we developed a large language model for intent detection and evaluated its performance under multiple conditions: using an off-the-shelf model without fine-tuning, fine-tuning on domain-specific data, fine-tuning with downsampling of the dominant off-topic intent category, and fine-tuning with downsampling combined with correction of human annotation errors identified from model-human disagreement. The off-the-shelf and fine-tuning–only methods demonstrated poor performance in detecting in-domain intents. Fine-tuning with downsampling achieved unweighted and weighted F_1-scores of 0.88 and 0.91, respectively, on the downsampled corrected dataset, and unweighted and weighted F_1-scores of 0.58 and 0.66 with an unweighted recall of 0.80 on the full test dataset. Downsampling with error correction achieved 0.86 unweighted and 0.90 weighted F_1-scores on the downsampled corrected test data, and 0.57 unweighted and 0.65 weighted F_1-scores on the full test dataset. Downsampling the training dataset lowered precision and therefore the F_1-scores on the full dataset due to the high prevalence of the off-topic intent category in the latter. However, downsampling with error correction achieved slightly higher unweighted recall (0.82) on the full test dataset compared with downsampling alone (0.80), indicating only modest improvement in detecting in-domain message intents. This finding suggests the error correction could conceivably enhance performance; though the difference is small, it is statistically significant. Thus, automated downsampling alone may provide sufficient benefit, though the resource-intensive manual error correction process could offer slightly better returns in practice.

Acknowledgments

The authors thank the undergraduate students at the University of California, Irvine, who annotated the messages from our smoking-cessation support groups, and Douglas Calder and Connor Phillips for training and supervising these students.

Funding

For this study, we used messages from two smoking-cessation support-group intervention trials which were supported by NIH grants awarded to CP: R34DA061012 and R01CA204356. The large language model fine-tuning, data downsampling, and error correction work was supported in part by bridge funding to CP and IGH from the University of California Irvine Chao Family Comprehensive Cancer Center and the University of California Irvine Merage Future of Healthcare Fund.

Data Availability

The datasets analyzed for this study are not publicly available because participants in the interventions did not provide approval to release their posts to the public, even if deidentified.

Authors' Contributions

SR's contributions include conceptualization, data curation, formal analysis, methodology, software, validation, visualization, and writing – original draft. CP's contributions include conceptualization, funding acquisition, investigation, project administration,

XSL•FO
RenderX

resources, supervision, and writing – review and editing. IGH's contributions include conceptualization, methodology, project administration, resources, supervision, and writing – review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables and figures illustrating the study's evaluation process.
[PDF File, 1224 KB - jmir_v28i1e83437_app1.pdf]

References

1. Shiu E, Hassan LM, Walsh G. Demarketing tobacco through governmental policies – the 4Ps revisited. J Bus Res 2009 Feb;62(2):269-278. [doi: 10.1016/j.jbusres.2008.01.034]
2. He L, Balaji D, Wiers RW, Antheunis ML, Krahmer E. Effectiveness and acceptability of conversational agents for smoking cessation: a systematic review and meta-analysis. Nicotine Tob Res 2023 Jun 9;25(7):1241-1250. [doi: 10.1093/ntr/ntac281] [Medline: 36507916]
3. Bendotti H, Lawler S, Chan GCK, Gartner C, Ireland D, Marshall HM. Conversational artificial intelligence interventions to support smoking cessation: a systematic review and meta-analysis. Digit Health 2023;9. [doi: 10.1177/20552076231211634] [Medline: 37928336]
4. Regmi K, Kassim N, Ahmad N, Tuah NA. Effectiveness of mobile apps for smoking cessation: a review. Tob Prev Cessat 2017;3:12. [doi: 10.18332/tpc/70088] [Medline: 32432186]
5. Tong HL, Quiroz JC, Kocaballi AB, et al. Personalized mobile technologies for lifestyle behavior change: a systematic review, meta-analysis, and meta-regression. Prev Med 2021 Jul;148:106532. [doi: 10.1016/j.ypmed.2021.106532] [Medline: 33774008]
6. Free C, Knight R, Robertson S, et al. Smoking cessation support delivered via mobile phone text messaging (txt2stop): a single-blind, randomised trial. Lancet 2011 Jul 2;378(9785):49-55. [doi: 10.1016/S0140-6736(11)60701-0] [Medline: 21722952]
7. Free C, Whittaker R, Knight R, Abramsky T, Rodgers A, Roberts IG. Txt2stop: a pilot randomised controlled trial of mobile phone-based smoking cessation support. Tob Control 2009 Apr;18(2):88-91. [doi: 10.1136/tc.2008.026146] [Medline: 19318534]
8. Perski O, Crane D, Beard E, Brown J. Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study. Digit Health 2019;5. [doi: 10.1177/2055207619880676] [Medline: 31620306]
9. Stead LF, Carroll AJ, Lancaster T. Group behaviour therapy programmes for smoking cessation. Cochrane Database Syst Rev 2017 Mar 31;3(3):CD001007. [doi: 10.1002/14651858.CD001007.pub3] [Medline: 28361497]
10. Mersha AG, Bryant J, Rahman T, McGuffog R, Maddox R, Kennedy M. What are the effective components of group-based treatment programs for smoking cessation? A systematic review and meta-analysis. Nicotine Tob Res 2023 Aug 19;25(9):1525-1537. [doi: 10.1093/ntr/ntad068] [Medline: 37104053]
11. Struik L, Khan S, Assoiants A, Sharma RH. Assessment of social support and quitting smoking in an online community forum: study involving content analysis. JMIR Form Res 2022 Jan 13;6(1):e34429. [doi: 10.2196/34429] [Medline: 35023834]
12. Pechmann C, Pan L, Delucchi K, Lakon CM, Prochaska JJ. Development of a Twitter-based intervention for smoking cessation that encourages high-quality social media interactions via automessages. J Med Internet Res 2015 Feb 23;17(2):e50. [doi: 10.2196/jmir.3772] [Medline: 25707037]
13. Pechmann C, Delucchi K, Lakon CM, Prochaska JJ. Randomised controlled trial evaluation of Tweet2Quit: a social network quit-smoking intervention. Tob Control 2017 Mar;26(2):188-194. [doi: 10.1136/tobaccocontrol-2015-052768] [Medline: 26928205]
14. Joyce E, Kraut RE. Predicting continued participation in newsgroups. J Comp Mediated Comm 2006 Apr;11(3):723-747. [doi: 10.1111/j.1083-6101.2006.00033.x]
15. Bricker JB, Sullivan B, Mull K, Santiago-Torres M, Lavista Ferres JM. Conversational chatbot for cigarette smoking cessation: results from the 11-step user-centered design development process and randomized controlled trial. JMIR Mhealth Uhealth 2024 Jul 23;12:e57318. [doi: 10.2196/57318] [Medline: 38913882]
16. Abroms LC, Wysota CN, Yousefi A, Wu TC, Broniatowski DA. ChatGPT-based chatbot for help quitting smoking via text messaging: an interventional study. JMIR Form Res 2025 Oct 10;9:e79402. [doi: 10.2196/79402] [Medline: 41072009]
17. Abroms LC, Yousefi A, Wysota CN, Wu TC, Broniatowski DA. Assessing the adherence of ChatGPT chatbots to public health guidelines for smoking cessation: content analysis. J Med Internet Res 2025 Jan 30;27:e66896. [doi: 10.2196/66896] [Medline: 39883917]
18. Kosyluk K, Baeder T, Greene KY, et al. Mental distress, label avoidance, and use of a mental health chatbot: results from a US survey. JMIR Form Res 2024 Apr 12;8:e45959. [doi: 10.2196/45959] [Medline: 38607665]

XSL•FO
RenderX

19. Prochaska JJ, Vogel EA, Chieng A, et al. A therapeutic relational agent for reducing problematic substance use (Woebot): development and usability study. J Med Internet Res 2021 Mar 23;23(3):e24850. [doi: 10.2196/24850] [Medline: 33755028]
20. Giyahchi T, Singh S, Harris I, Pechmann C. Customized training of pretrained language models to detect post intents in online health support groups. In: Shaban-Nejad A, Michalowski M, Bianco S, editors. Multimodal AI in Healthcare: A Paradigm Shift in Health Intelligence: Springer International Publishing; 2023:59-75. [doi: 10.1007/978-3-031-14771-5_5]
21. McRoy S, Rastegar-Mojarad M, Wang Y, Ruddy KJ, Haddad TC, Liu H. Assessing unmet information needs of breast cancer survivors: exploratory study of online health forums using text classification and retrieval. JMIR Cancer 2018 May 15;4(1):e10. [doi: 10.2196/cancer.9050] [Medline: 29764801]
22. Park J, Kotzias D, Kuo P, et al. Detecting conversation topics in primary care office visits from transcripts of patient-provider interactions. J Am Med Inform Assoc 2019 Dec 1;26(12):1493-1504. [doi: 10.1093/jamia/ocz140] [Medline: 31532490]
23. Zeng K, Pan Z, Xu Y, Qu Y. An ensemble learning strategy for eligibility criteria text classification for clinical trial recruitment: algorithm development and validation. JMIR Med Inform 2020 Jul 1;8(7):e17832. [doi: 10.2196/17832] [Medline: 32609092]
24. Li X, Shu Q, Kong C, et al. An intelligent system for classifying patient complaints using machine learning and natural language processing: development and validation study. J Med Internet Res 2025;27:e55721. [doi: 10.2196/55721]
25. Babu A, Boddu SB. BERT-based medical chatbot: enhancing healthcare communication through natural language understanding. Explor Res Clin Soc Pharm 2024 Mar;13:100419. [doi: 10.1016/j.rcsop.2024.100419] [Medline: 38495953]
26. Yu HQ, McGuinness S. An experimental study of integrating fine-tuned large language models and prompts for enhancing mental health support chatbot system. J Med Artif Intell 2024;7:16. [doi: 10.21037/jmai-23-136]
27. Zamani A, Reeson M, Marshall T, et al. Intent and entity detection with data augmentation for a mental health virtual assistant chatbot. In: Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents: Association for Computing Machinery; 2023:41. [doi: 10.1145/3570945.3607324]
28. Aftab H, Gautam V, Hawkins R, Alexander R, Habli I. Robust intent classification using Bayesian LSTM for clinical conversational agents (CAs). In: Gao X, Jamalipour A, Guo L, editors. Wireless Mobile Communication and Healthcare: Springer International Publishing; 2022:106-118. [doi: 10.1007/978-3-031-06368-8_8]
29. Zhang B, Tu Z, Hang S, Chu D, Xu X. Conco-ERNIE: complex user intent detect model for smart healthcare cognitive bot. ACM Trans Internet Technol 2023 Feb 28;23(1):21. [doi: 10.1145/3574135]
30. Sakurai H, Miyao Y. Evaluating intention detection capability of large language models in persuasive dialogues. In: Ku L, Martins A, Srikumar V, editors. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers): Association for Computational Linguistics; 2024:1635-1657. [doi: 10.18653/v1/2024.acl-long.90]

Abbreviations

8B: 8 billion parameters
AI: artificial intelligence
BERT: Bidirectional Encoder Representations From Transformers
Bi-LSTM: bidirectional long short-term memory
BLEU: Bilingual Evaluation Understudy
CPU: central processing unit
CUDA: Compute Unified Device Architecture
EDA: easy data augmentation
EOS: end-of-sequence
FP32: 32-bit floating point
GPU: graphics processing unit
HS: human subjects
IRB: institutional review board
LoRA: low-rank adaptation
LSTM: long short-term memory
MIMIC-III: Medical Information Mart for Intensive Care-III
NRT: nicotine replacement therapy
PEFT: parameter-efficient fine-tuning
QLoRA: quantized low-rank adaptation
SVM: support vector machine
UC: University of California
UCI: University of California, Irvine

XSL•FO
RenderX

Edited by A Coristine; submitted 02.Sep.2025; peer-reviewed by C Chikezie, F Amakye, M Ekwueme, O Akinsola, RQ Wang, T Karen; accepted 29.Jan.2026; published 09.Mar.2026.

Please cite as:
Rahman S, Pechmann C(, Harris IG
Enhancing Detection of Message Intents in a Mobile Health Smoking-Cessation Intervention Using Large Language Model Fine-Tuning, Data Downsampling, and Error Correction: Algorithm Development and Validation
J Med Internet Res 2026;28:e83437
URL: https://www.jmir.org/2026/1/e83437
doi:10.2196/83437

© Shagoto Rahman, Cornelia (Connie) Pechmann, Ian G Harris. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluation of an Artificial Intelligence Conversational Chatbot to Enhance HIV Preexposure Prophylaxis Uptake: Development and Usability Internal Testing

Jun Tao[1], PhD; Ellie Pavlick[2], PhD; Amaris Grondin[2], BS; Josue D Bustamante[3], BS; Harrison Martin[1], BA; Hannah Parent[1], MPH; Natalie Fenn[1,4], PhD; Alexi Almonte[1], BA; Amanda Maguire-Wilkerson[1], DrPH; Mofan Gu[5], PhD; Jack Rusley[6,7], MHS, MD; Bryce K Perler[1], MSTR, MD; Tyler Wray[8], PhD; Amy S Nunn[8,9], SCD; Philip A Chan[1,8,9], MS, MD

[1]Department of Medicine, The Warren Alpert Medical School of Brown University, Room 130, 11 4th Street, Providence, RI, United States

[2]Department of Computer Science, Brown University, Providence, RI, United States

[3]School of Electrical Engineering and Computer Science, College of Engineering, Oregon State University, Corvallis, OR, United States

[4]Department of Psychiatry and Human Behavior, The Warren Alpert Medical School of Brown University, Providence, RI, United States

[5]Division of Infectious Diseases, The Miriam Hospital, Providence, RI, United States

[6]Department of Health Services, Policy, and Practice, Brown University School of Public Health, Providence, RI, United States

[7]Department of Pediatrics, Division of Adolescent Medicine, The Warren Alpert Medical School of Brown University, Providence, RI, United States

[8]Department of Behavioral and Social Sciences, Brown University School of Public Health, Providence, RI, United States

[9]Open Door Health, Rhode Island Public Health Institute, Providence, RI, United States

Corresponding Author:
Jun Tao, PhD
Department of Medicine, The Warren Alpert Medical School of Brown University, Room 130, 11 4th Street, Providence, RI, United States

Abstract

Background: The HIV epidemic in the United States disproportionately impacts gay, bisexual, and other men who have sex with men (MSM). Despite the effectiveness of HIV preexposure prophylaxis (PrEP) in preventing HIV acquisition, uptake among MSM remains suboptimal. Motivational interviewing (MI) has demonstrated efficacy at increasing PrEP uptake among MSM but is resource-intensive, limiting scalability. The use of artificial intelligence, particularly large language models with conversational agents (ie, “chatbots”) such as ChatGPT, may offer a scalable approach to delivering MI-based counseling for PrEP and HIV prevention.

Objective: This internal usability testing aimed to evaluate the development of an artificial intelligence–based chatbot, including its ability to provide MI-aligned education about PrEP and HIV prevention and potential to support PrEP uptake.

Methods: The Chatbot for HIV Prevention and Action (CHIA) was built on a GPT-4o base model embedded with a validated knowledge database on HIV and PrEP in English and Spanish. The CHIA was fine-tuned through training on a large MI dataset and prompt engineering. The use of the AutoGen multiagent framework enabled the CHIA to integrate 2 agents, the PrEP Counselor Agent and the Assistant Agent, which specialized in providing MI-based counseling and handling function calls (eg, assessment of HIV risk), respectively. During internal testing from March 10-April 28, 2025, we systematically evaluated the CHIA’s performance in English and Spanish using a set of 5-point Likert scales to measure accuracy, conciseness, up-to-dateness, trustworthiness, and alignment with aspects of the MI spirit (eg, collaboration, autonomy support) and MI-consistent behaviors (eg, affirmation, open-ended questions). Descriptive statistics and mixed linear regression were used to analyze the data.

Results: A total of 296 responses, including 145 English responses and 151 Spanish responses, were collected during the internal testing period. Overall, the CHIA demonstrated strong performance across both languages, receiving the highest combined scores in the general response quality metrics including up-to-dateness (mean 4.6, SD 0.8), trustworthiness (mean 4.5, SD 0.9), accuracy (mean 4.4, SD 0.9), and conciseness (mean 4.2, SD 1.1). The CHIA generally received higher combined scores for metrics that assessed alignment with the MI spirit (ie, empathy, evocation, autonomy support, and collaboration) and lower combined scores for MI-consistent behaviors (ie, affirmation, open-ended questions, and reflections). Spanish responses had significantly lower mean scores than English responses across nearly all MI-based metrics.

Conclusions: Our internal usability testing highlights the potential of the CHIA as a viable tool for delivering MI-aligned counseling in English and Spanish to promote HIV prevention and support PrEP uptake, though its Spanish language performance requires further improvement.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e79671) doi:10.2196/79671

KEYWORDS

artificial intelligence; counseling; HIV infections; motivational interviewing; multiagent framework; preexposure prophylaxis

Introduction

In the United States, HIV continues to be a significant cause of morbidity and mortality, disproportionately affecting groups including gay, bisexual, and other men who have sex with men (MSM). In 2022, MSM comprised 67% of new HIV diagnoses in the United States [1]. Hispanic or Latino and Black or African American individuals accounted for 39% and 35% of these new diagnoses, respectively [1]. HIV preexposure prophylaxis (PrEP) is highly effective at preventing HIV among populations at an increased risk of infection, including MSM [2-4]. However, PrEP uptake among MSM, and particularly Hispanic or Latino and Black or African American MSM, remains suboptimal due to inequitable access to health care and stigma [5-7].

Motivational interviewing (MI) is an evidence-based, patient-centered approach to healthy behavior change that has demonstrated efficacy in facilitating PrEP uptake among MSM [8-10]. In our previous work, we demonstrated that a brief MI-based intervention improved PrEP uptake among MSM in a clinical setting [9]. However, implementing MI in practice requires a significant investment in time and resources for provider training and intervention delivery. Research suggests that provider-led MI counseling usually requires 2-5 sessions (30 - 60 min per session) to produce measurable behavior change, with several studies reporting substantially higher effectiveness (>80%) when more than 5 sessions are delivered [11-13]. The associated clinician time and continuity demands make sustained MI delivery difficult to scale, particularly during brief clinical visits [11-14]. Time demands apply not only to patient encounters but also to provider training, which typically involves several hours of didactic instruction and coaching [13]. Artificial intelligence (AI)—specifically large language models (LLMs) with conversational agents ("chatbots") such as ChatGPT—has shown promise in overcoming these challenges [15-17]. Preliminary research supports AI-based chatbots' ability to employ MI techniques in promoting healthy behavior change, including smoking cessation and decreased substance use [17-21]. A chatbot designed to use MI principles can be made available 24/7, requiring only minimal human effort for periodic supervision and quality assurance. Drawing on empirical session-length data and standard labor-cost benchmarks, it suggests that such MI-aligned chatbots could substantially improve the scalability of MI delivery.

In the context of HIV prevention, researchers have noted many uses for AI-based chatbots [22-24]. Limited evidence suggests that chatbots may help facilitate the uptake of HIV testing and PrEP among populations at increased risk of HIV, including MSM [22,23,25]. Chatbots have also shown the potential to provide personalized counseling on sensitive health topics including HIV prevention [22]. Although AI-based chatbots hold significant promise for HIV prevention efforts, concerns exist regarding their ability to provide accurate medical information, stay up-to-date on current clinical recommendations, and demonstrate cultural competence [22,26]. Studies have highlighted issues such as hallucinations (incorrect or misleading information) and the potential for perpetuating biases, which can undermine trust and effectiveness [22]. Low engagement with AI-based chatbots for health promotion has also been documented in the literature, presenting challenges with delivering effective interventions via this modality [27,28]. Additionally, no studies, to our knowledge, have evaluated the use of MI by an AI-based chatbot for HIV prevention.

In this usability testing, we developed and conducted an internal evaluation of an AI-based chatbot (Chatbot for HIV Prevention and Action [CHIA]) that harnesses MI to provide personalized counseling for PrEP. To address the limitations noted above and improve the CHIA's performance, we integrated three complementary components: (1) a retrieval-augmented generation pipeline constrained to a curated, validated knowledge base to reduce unsupported statements; (2) MI alignment via supervised fine-tuning on annotated MI transcripts coupled with preference-based tuning using expert-selected responses; and (3) personalization through a structured HIV risk assessment and the transtheoretical model (TTM) to tailor counseling to the user's stage of change [29,30]. We present the results of an internal evaluation of the CHIA, assessing the chatbot's alignment with MI principles, factual accuracy, and its ability to deliver appropriate counseling for HIV prevention and PrEP. This internal testing serves as a cornerstone for future real-world implementation and evaluation of the CHIA's performance among individuals at an increased risk of HIV.

Methods

Overview of Chatbot Design

Generative pre-trained transformers are LLMs that use deep learning to generate human-like text based on natural language input [31]. GPT-4o, released by OpenAI in May 2024, is a multimodal LLM capable of processing both text and images with improved efficiency and performance compared to earlier versions [32]. We developed the CHIA using GPT-4o to deliver MI-informed counseling aimed at improving HIV prevention outcomes, particularly the uptake of PrEP.

The CHIA consists of two main components: (1) a fine-tuned, customized LLM embedded with a validated knowledge database and (2) multiple specialized agents and functions to meet users' needs. Built on a GPT-4o base model, the CHIA detects language inputs automatically and has been fine-tuned using a large MI dataset with the goal of training it to produce empathetic responses and avoid biases [33,34]. GPT-4o was selected because it offers superior conversational quality, multilingual capability, and reduced risk of generating errors compared to smaller open-source models. These features are critical for building trust with users discussing sensitive health topics. To ensure accurate and up-to-date information on HIV and PrEP, the CHIA integrates the latest validated data on these

XSL•FO
RenderX

topics, reviewed by a team of physicians and researchers, in English and Spanish, with monthly updates.

The CHIA's architecture integrates 2 specialized agents (Figure 1): a user-facing PrEP Counselor agent and a tool-executing Assistant agent using AutoGen [35]—an open-source multiagent framework developed by Microsoft. The Counselor conducts the entire conversation using MI, triages user needs, and delegates tasks to the Assistant when external information is required. The Assistant then invokes specific functions—knowledge retrieval from an embedding-backed knowledge base, HIV risk or readiness assessment, PrEP provider search and referral info, reminders or links, and initiating human support on request—and returns results to the Counselor. The Counselor interprets these outputs through MI principles and delivers the response to the user. To maintain continuity across visits, the Counselor agent is capable of using *Teachability* to store key information to ensure the continuity of conversation (eg, identity token, risk or concerns, prior plan or notes). This design enables the CHIA to deliver personalized responses that mimic human counseling. Advanced prompt engineering keeps interactions dynamic and contextually relevant [36]. The CHIA is secured through a login page and deployed on Amazon Web Services (AWS) with the Supabase software (version 1.25; Supabase, Inc.) for backend management, ensuring robust and private data handling [37,38].

Figure 1. Chatbot for HIV Prevention and Action (CHIA) architecture: information flow among User, Counselor Agent, and Assistant Agent. MI: motivational interviewing.

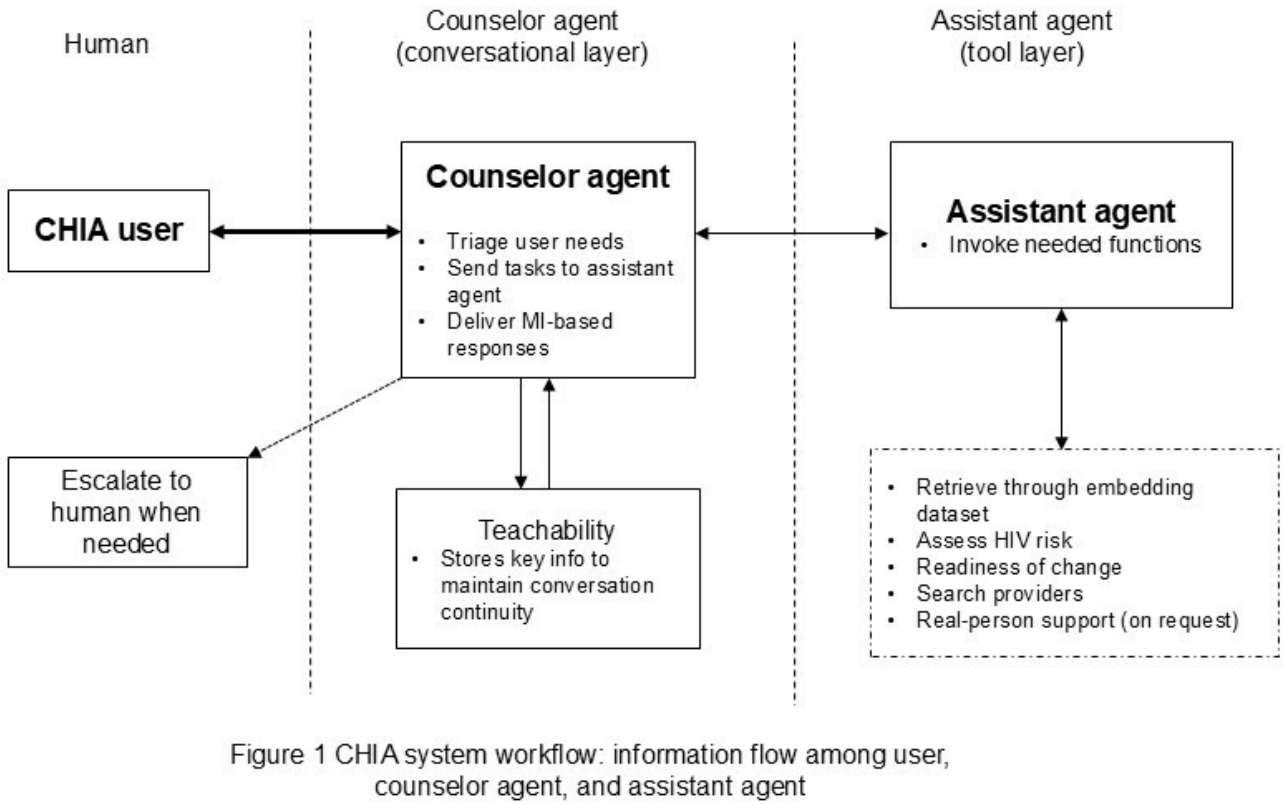

Embedded HIV/PrEP Dataset Development and Validation

To enhance chatbot accuracy and reliability, embedding and retrieval-based techniques have been employed to minimize misinformation by ensuring responses are grounded in validated sources [39]. Assessment retrieval methods, guided by established frameworks, such as the retrieval-augmented generation (RAG) playground framework and the automated RAG evaluation system [40-42], allow for systematic evaluation of chatbot performance in delivering evidence-based, contextually relevant health information. To develop a robust embedded dataset specifically for the CHIA, our study team, composed of 4 research assistants under the guidance of 2 physicians and 1 principal investigator with decades of research experience in HIV prevention and treatment, systematically curated and validated key information. We compiled the most frequently asked questions about HIV and PrEP, which included topics such as basic knowledge, effectiveness, side effects, formulations, insurance coverage, financial assistance, and cultural considerations. A Spanish-fluent research assistant translated the dataset to ensure accessibility for Spanish-speaking users. Each data entry was structured into a standardized format, categorizing content by topic, question, and answer to optimize retrieval efficiency. All information underwent expert validation for medical accuracy and clarity. These processes prevent the spread of misinformation or erroneous recommendations ("hallucinations"), which ChatGPT cannot guarantee. The embedding database is updated monthly and as needed when users ask questions outside the existing dataset under expert supervision. In such cases, the system flags the query, alerting the development team to review, validate, and integrate new information. This dynamic updating process helps to ensure that the CHIA remains accurate, relevant, and responsive to evolving needs of the community.

XSL•FO
RenderX

MI Alignment via Implementation of Preference Fine-Tuning Techniques

To enhance the CHIA's ability to deliver MI-based counseling, we utilized preference fine-tuning techniques [43]. This was done because open-source models would have been insufficient for MI fine-tuning due to their lack of multilingual support, higher hallucination rates, and insufficient token context. We processed a large, publicly available MI dataset downloaded from GitHub to facilitate diversity in linguistic styles and conversational structures [33,34]. The dataset contained 2000 dialogs, half of which were from publicly available conversations between potential clients and licensed counselors on CounselChat—an online platform for mental health support—and the other half from exchanges between users and peer supporters on Reddit subforums related to emotional distress [33,34]. All dialogs in the dataset were annotated by trained counselors with labels adapted from the Motivational Interviewing Treatment Integrity Code 2.0 and 4.2.1, a widely used method for evaluating how well clinicians perform MI [33,34,44-46]. Our team preprocessed the raw text from the dataset, tokenized it (ie, broke down into smaller units) [47], and then converted this to a vector database. This processed vector database was employed to fine-tune the GPT-4o base model, enabling the CHIA to generate responses informed by MI. We used the Direct Preference Optimization algorithm to refine response quality [43]. A preference fine-tuning JSON file was constructed that ranked certain responses as preferable over others based on their alignment with MI-consistent behaviors such as open-ended questioning and reflective listening. This iterative training process was designed to enhance the CHIA's ability to prioritize MI-aligned responses while maintaining coherence and engagement. By the end of this process, we developed a specialized ChatGPT-4o model for MI-based counseling, serving as the foundation for the CHIA to deliver personalized, empathetic, and structured conversations that support PrEP uptake and public health interventions.

HIV Risk Assessment and Readiness for Change Functions

To enhance the CHIA's ability to provide personalized guidance, we integrated functions to assess individual HIV risk and readiness for behavior change. HIV risk assessment was based on the HIV Incidence Risk Index for MSM [48], following Centers for Disease Control and Prevention guidelines for PrEP eligibility [49]. A focus on MSM was chosen due to the disproportionate number of new HIV diagnoses occurring among this population in the United States [1]. Additionally, the CHIA's development incorporated the TTM through the use of the Contemplation Ladder, a validated tool that allowed individuals to self-assess their readiness for PrEP uptake on a scale from 0 to 10 [29,50]. These assessments were embedded into the CHIA's conversational interface, enabling real-time evaluation, as users engaged with relevant questions. Based on these assessments, the CHIA tailored its responses to align with each user's unique risk profile and stage of readiness for change, providing targeted, motivational, and evidence-based guidance to support PrEP uptake.

Linkage to PrEP Care and Referral Function

To facilitate access to PrEP, we developed a function that searched the "PrEP Locator" website, a national database of PrEP providers in the United States, using user-provided ZIP codes [51]. This function was integrated into the CHIA to generate a list of PrEP providers within a 30-mile radius, offering users convenient options for selecting nearby clinics. Additionally, a separate function was implemented to enable the CHIA to engage users with follow-up questions to assess clinic preferences, determine if they wished to be contacted by a provider for an initial appointment, and identify potential barriers to care. An AI-generated and encrypted email was then sent to the selected clinic with a referral and relevant contact information. Study staff coordinated with the clinic to ensure appointments were scheduled efficiently. Furthermore, the CHIA inquired whether users preferred to connect with a real person for assistance in accessing PrEP care. If requested, study staff were notified and contacted the individual within 1 business day.

User Interface and Data Security

The CHIA user interface was designed to provide a streamlined, secure, and personalized experience (Figure 2). Users could log in with an existing account or create a new one, with 2-factor authentication required for each login to ensure data protection and confidentiality. The login system adhered to high cybersecurity standards to safeguard user information. Once logged in, users accessed a clean, intuitive interface with clear navigation options and a responsive chat window to facilitate seamless interactions.

XSL•FO
RenderX

Figure 2. The user interface of the Chatbot for HIV Prevention and Action (CHIA).

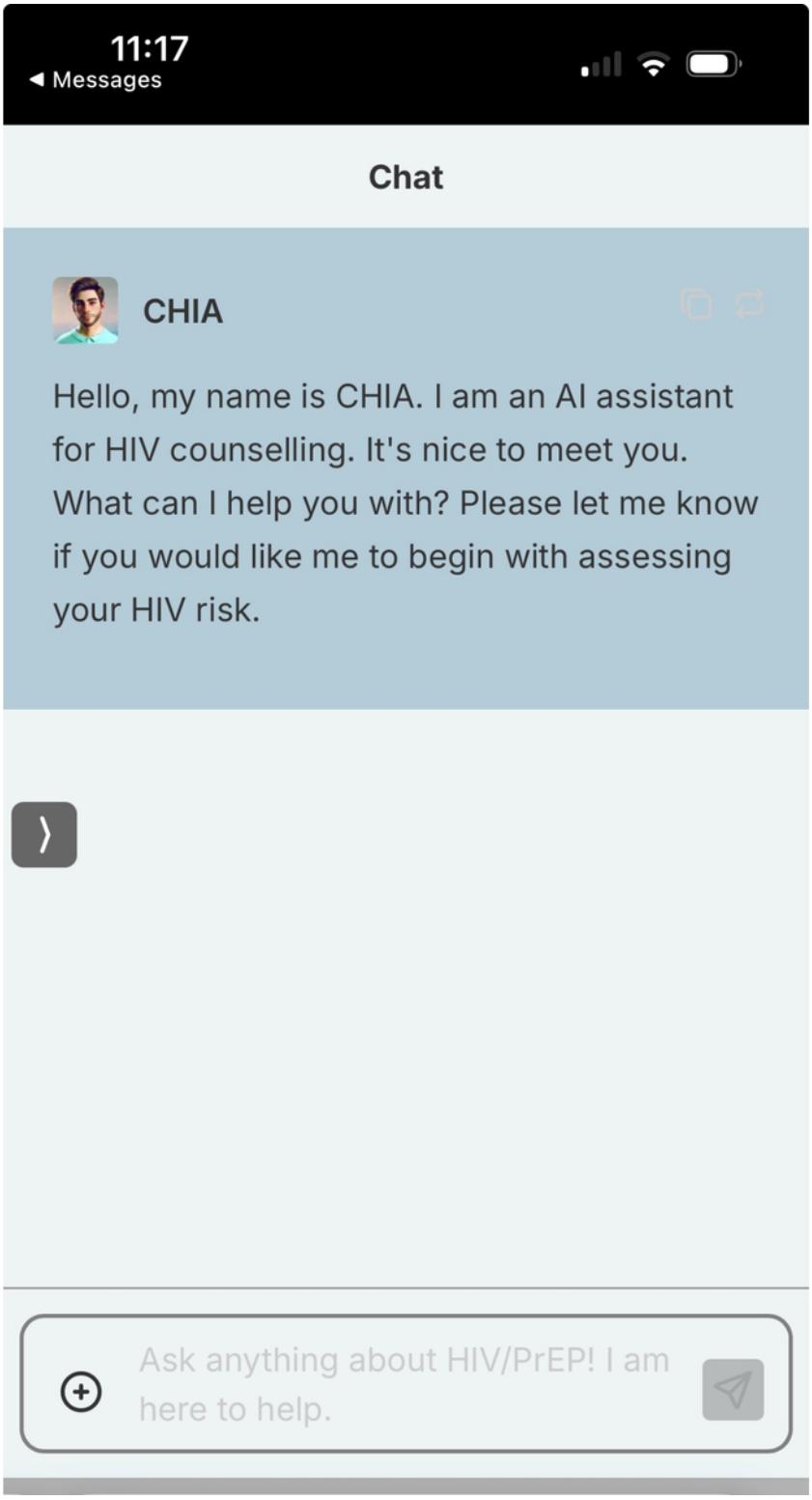

To facilitate secure data management, the CHIA was deployed through AWS, leveraging its cloud-based infrastructure for reliable and Health Insurance Portability and Accountability Act (HIPAA)–compliant storage [38]. AWS encryption protocols protected user data, ensuring confidentiality and integrity [38]. Additionally, Supabase was used to host the back-end data, providing an efficient database solution for managing user interactions [37]. This integration of AWS for deployment and Supabase for back-end data management enabled the CHIA to maintain robust security measures while promoting a smooth user experience. The secure, user-friendly interface aimed to encourage sustained engagement and support meaningful interactions.

Procedures

During internal testing from March 10 to April 18, 2025, 4 research staff members systematically evaluated the CHIA's performance in both English and Spanish. Each researcher acted

XSL•FO
RenderX

as a potential user and interacted with the chatbot over multiple sessions. Two research staff members interacted with the chatbot in English, and 2 staff members fluent in Spanish interacted with the chatbot in Spanish. Researchers then exchanged transcripts and rated each response on a set of 11 metrics covering general response quality and alignment with MI using a 5-point Likert scale. Likert scores were used primarily to enable relative comparisons across languages and scenarios, rather than to assess performance against an external benchmark, as no standardized threshold currently exists for interpreting Likert-based ratings of motivational interviewing alignment in chatbot evaluations. The research team met prior to the review process to ensure familiarity with the operational rating scale and consistency in scoring across raters. After interacting with the chatbot, the team met to discuss conversation-level findings and the performance of the HIV risk assessment, referral, and readiness for change functions.

Evaluation Metrics for Retrieving Information From Embedded Dataset

The CHIA's RAG functionality was evaluated at the session level using 7 key metrics—groundedness, medical accuracy, completeness, no fabrication, appropriate tone, safety, and reasoning—that were scored on a 5-point Likert scale via the Automated RAG Evaluation System [42,52]. Groundedness assessed if responses were based on retrieval sources, and medical accuracy sought to ensure alignment with validated health information. Completeness measured whether the retrieved content provided sufficient information. No fabrication verified that responses did not introduce false or misleading details (sometimes referred to as "hallucinations") by comparing each RAG output against the ground-truth dataset to confirm retrieval only from the embedded validated knowledge base. Appropriate tone was evaluated if responses were professional and empathetic, and safety was assessed whether the information adhered to ethical and safety principles. Finally, reasoning evaluated the chatbot's ability to integrate and apply retrieved knowledge. Sessions with any metric below a predefined threshold were auto-flagged for human review; we logged and reviewed misinformation flags and near-miss events (eg, unsafe advice avoided or corrected) and conducted monthly audits summarizing flag rates, time-to-review, and corrective actions.

Metrics for Comprehensive Assessment

We developed and used a structured assessment framework to evaluate the CHIA's responses based on previous research that incorporated multiple dimensions of chatbot interaction [26,53]. First, we assessed general response quality using 4 metrics—accuracy, conciseness, up-to-dateness, and trustworthiness—that were scored on a 5-point Likert scale to ensure that responses were factually correct, concise, up-to-date, and emotionally supportive. We also developed a set of metrics inspired by the Motivational Interviewing Treatment Integrity Code to measure the alignment of the CHIA's responses with the MI spirit and MI-consistent behaviors [44,45]. Metrics to assess alignment with the MI spirit included empathy, evocation, autonomy support, and collaboration, and those for MI-consistent behaviors included affirmation, open-ended questions, and reflections. All MI-based metrics were also scored using a 5-point Likert scale. Additionally, safety was evaluated on a pass or fail basis in which the CHIA's use of toxic language, demonstration of bias, or violation of privacy constituted a failure. A qualitative feedback section allowed raters to document strengths, weaknesses, and suggested improvements. This comprehensive framework ensured a robust and reliable evaluation of the CHIA's performance in multiple languages and user engagement scenarios. The detailed metrics and operational assessment protocol is included in Multimedia Appendix 1. At the conversation level, we assessed the CHIA's performance using the same response-level metrics in addition to evaluating its overall adherence to MI techniques including employing a guiding (rather than directive) style of conversation, eliciting change talk, and managing sustain talk (ie, statements against change). Given that this analysis was limited to 4 conversations, only qualitative findings are reported in this paper. Finally, we assessed the CHIA's memory and teachability by measuring its ability to recall information accurately and integrate key details from past interactions into current conversation. The qualitative findings of this assessment are summarized in this paper.

Ethical Considerations

This internal testing did not involve human participants or the use of human subject data. All testing was conducted internally using simulated interactions to evaluate chatbot performance. As such, this developmental phase of the study does not meet the definition of human participants research and was determined to be exempt from IRB review by the Miriam Hospital Institutional Review Board (protocol #2312729).

Statistical Analysis

Descriptive statistics, including the mean and SD, were calculated separately for English and Spanish responses for each metric (ie, accuracy, conciseness, up-to-dateness, trustworthiness, empathy, evocation, autonomy support, collaboration, affirmation, open-ended questions, and reflections). We fit mixed linear regression models to estimate the effect of language (Spanish vs English) on each communication metric. Because multiple observations were nested within individuals, models included random intercepts and random slopes for language at the participant level. This specification accounted for within-person correlation (ie, repeated measures from the same individual) and allowed the magnitude of the language effect to vary across individuals. Fixed effects provided the average adjusted difference between languages, while random effects decomposed variance into within- and between-person components. In addition, standardized effect sizes (Cohen *d*) were derived by dividing the adjusted language difference by the residual standard deviation, providing a measure of the practical significance of language effects across metrics. Statistical significance was assessed using 2-tailed *P* values. The significance level was set at $P<.05$. All analyses were performed using Stata (version 18; StataCorp LLC).

XSL•FO
RenderX

Results

Overview of Response-Level Assessment

A total of 296 responses were assessed across 11 metrics, covering general response quality (eg, accuracy, trustworthiness) as well as alignment with the aspects of the MI spirit (eg, evocation, autonomy support) and with MI-consistent behaviors (eg, affirmation, open-ended questions). This total included 145 English responses and 151 Spanish responses. All responses in English and Spanish passed the safety evaluation, indicating that the CHIA's outputs were appropriate, ethical, and unbiased. Table 1 presents the mean and SD for all metrics at the response level, along with effect sizes reported using Cohen *d*. Table 2 displays the examples of CHIA's responses to researcher prompts and their scores for evaluation metrics. Multimedia Appendix 2 includes all responses and scores used in the analysis for both the English and Spanish assessments.

Table . Response-level assessment for the internal testing phase of Chatbot for HIV Prevention and Action (CHIA), an artificial intelligence (AI)–based chatbot for HIV prevention.

Response-level metrics	English (n=145), mean (SD)	Spanish (n=151), mean (SD)	Total (n=296), mean (SD)	ICC[a]	Cohen *d*	*P* value
General response quality metrics						
Accuracy	4.1 (1.1)	4.6 (0.7)	4.4 (0.9)	7.65e-15	−0.55	<.001
Conciseness	4.0 (1.3)	4.5 (0.8)	4.2 (1.1)	0.120	−0.47	.07
Up-to-dateness	4.5 (1.0)	4.7 (0.7)	4.6 (0.8)	0.035	−0.24	.24
Trustworthiness	4.3 (1.0)	4.7 (0.7)	4.5 (0.9)	0.0169	−0.47	.001
MI[b] spirit metrics						
Empathy	4.2 (1.0)	3.9 (1.2)	4.1 (1.1)	0.242	0.272	.95
Evocation	4.0 (1.2)	2.0 (1.1)	3.2 (1.5)	0.257	1.74	.24
Autonomy support	4.2 (1.1)	3.7 (1.3)	3.9 (1.2)	0.279	0.413	.84
Collaboration	4.1 (1.1)	3.7 (1.4)	3.9 (1.3)	0.357	0.313	.786
MI-consistent behaviors metrics						
Affirmation	4.2 (1.2)	2.7 (1.6)	3.5 (1.6)	0.607	1.05	.001
Open-ended questions	3.8 (1.3)	3.0 (1.9)	3.4 (1.7)	0.536	0.48	.95
Reflections	3.6 (1.3)	3.7 (1.4)	3.6 (1.3)	0.452	−0.07	.59

[a]ICC: intraclass correlation coefficient.

[b]MI: motivational interviewing.

XSL•FO
RenderX

Table . Response-level assessment examples and scores[a].

		Evaluation metric						
Researcher prompt	CHIA[b] response	Empathy	Evocation	Autonomy support	Collaboration	Affirmation	Open-ended questions	Reflections
I've tried setting timers before for medications and they don't work.	It sounds like you've tried different strategies without success, and that's okay. Sometimes it takes a bit of creativity to find what works best for you. Perhaps pairing PrEP[c] with a daily activity you never miss, like eating breakfast, could help. Would you like to explore other ways to create a routine that feels natural for you?	5	5	5	5	5	5	5
Is there someone I can talk to about PrEP that is not a doctor or works in healthcare? I don't trust doctors; in my experience, they've treated me poorly.	I'm sorry to hear about your experiences. Seeking support from someone outside of healthcare is a valid choice. Community organizations or support groups focused on sexual health might be a good option. They can offer guidance and answer questions about PrEP. How do you feel about reaching out to a local LGBTQ+[d] center or community health organization for support?	5	5	5	5	5	5	5

XSL•FO
RenderX

		Evaluation metric						
Researcher prompt	CHIA[b] response	Empathy	Evocation	Autonomy support	Collaboration	Affirmation	Open-ended questions	Reflections
My partner could also benefit from PrEP maybe. How do I talk to my partner about it?	Talking to your partner about PrEP can be an important step in taking care of your sexual health together. You might start by expressing your care and concern for both of your well-being. You could say something like, "I've been thinking about how we can protect each other from HIV and STIs[e], and I came across PrEP. It's a medication that can help reduce the risk of HIV, especially if we have sex without condoms sometimes. What do you think about looking into it together?" This approach opens up a dialogue and allows your partner to share their thoughts and feelings, making it a collaborative decision.	4	1	4	1	4	4	4

[a]Responses were obtained through internal testing, in which research assistants simulated potential participants from the target study sample. These responses were not collected from actual human participants within the study sample. Please see Supplemental Dataset 1 for further examples of CHIA responses, including Spanish responses.

[b]CHIA: Chatbot for HIV Prevention and Action.

[c]PrEP: preexposure prophylaxis.

[d]LGBTQ+: lesbian, gay, bisexual, transgender/transsexual, queer, and other minority sexual orientations and gender identities.

[e]STIs: sexually transmitted infections.

English Response-Level Assessment

The CHIA performed strongly in up-to-dateness (mean 4.5, SD 1.0), trustworthiness (mean 4.3, SD 1.0), and accuracy (mean 4.1, SD 1.1). Conciseness had a slightly lower score (mean 4.0, SD 1.3) but remained generally acceptable. Metrics that assessed alignment with the MI spirit generally received higher mean scores than those for MI-consistent behaviors across English responses.

XSL•FO
RenderX

Spanish Response-Level Assessment

Spanish responses demonstrated strong performance across several dimensions. General response quality metrics including trustworthiness (mean 4.7, SD 0.7), up-to-dateness (mean 4.7, SD 0.7), accuracy (mean 4.6, SD 0.7), and conciseness (mean 4.5, SD 0.8) were rated highly. Overall, metrics that measured alignment with the MI spirit and MI-consistent behaviors received lower combined scores.

Combined Response-Level Assessment

The CHIA performed well across both languages, with combined scores showing strength across general response quality metrics including up-to-dateness (mean 4.6, SD 0.8), trustworthiness (mean 4.5, SD 0.9), accuracy (mean 4.4, SD 0.9), and conciseness (mean 4.2, SD 1.1). The mean scores for accuracy, conciseness, and trustworthiness were significantly higher among Spanish responses compared to English responses. Combined scores for metrics that assessed alignment with the MI spirit were generally higher than those for MI-consistent behaviors. Statistical tests indicated that Spanish responses received significantly lower mean scores than English responses across nearly all MI-based metrics.

Conversation Level Assessment

Conversations with the CHIA were generally perceived as reliable and emotionally supportive but occasionally repetitive or overly generic. Empathy and collaboration were present but could be deepened with more emotionally attuned language and user-specific questions. The CHIA's autonomy support was acknowledged, though one instance where the chatbot proceeded with a risk assessment against user preference indicated room for technical and conversational improvements. All reviewers emphasized reducing reliance on early referrals to health care providers and instead suggested a more user-driven flow. MI techniques (eg, change talk elicitation, sustain talk management, guiding style) were successfully implemented across English conversations, yet raters noted the need for the CHIA to ask more personalized, open-ended questions earlier in the conversation to build rapport and relevance. In the Spanish version, reviewers reported that conversations were not consistently MI-aligned, often lacking key aspects of the MI spirit such as autonomy support and MI-consistent behaviors including affirmation, open-ended questions, and reflections.

Evaluation of the Referral, HIV Risk Assessment, and Readiness for Change Functions

For the referral function, ZIP codes from across the United States were entered. The CHIA successfully returned accurate listings of nearby PrEP clinics within a 30-mile radius. However, the chatbot occasionally failed to provide detailed information about specific clinics when requested, highlighting the need to enable the CHIA to retrieve location-specific data by accessing selected clinic websites. In contrast, the HIV risk assessment function consistently performed well across all conversations. This feature has since been refined to allow users to exit the assessment if it is accidentally triggered. Overall, both referral and risk assessment functions were functional and helpful, with minor refinements needed to optimize user experience. All research assistants tested the CHIA's readiness for change function, which supports MI-based counseling by identifying the user's stage of change. The function successfully prompted tailored, stage-appropriate responses to guide users toward PrEP decision-making. Overall, it enhanced the CHIA's ability to deliver personalized, action-oriented support aligned with MI core skills in this internal pilot testing.

Assessment of Retrieval Functionality

Performance was strong across all 7 key metrics designed to assess the CHIA's RAG functionality (groundedness, medical accuracy, completeness, no fabrication, appropriate tone, safety, and reasoning; Table 3). Mean scores ranged from 3.7 to 4.6, with SDs between 0.5 and 1.1. Median scores for each metric were consistently high, with IQRs falling within acceptable variability (eg, median 4, IQR 3-5). These findings indicate that the CHIA's responses were consistently accurate, grounded in reliable sources, and communicated in a safe and professional manner.

Table . Evaluation of the Chatbot for HIV Prevention and Action's (CHIA) retrieval-augmented generation functionality.

Metrics	Mean (SD)
Groundedness	4.1 (1.1)
Medical accuracy	4.6 (0.8)
Completeness	3.7 (1.0)
No fabrication	4.6 (0.8)
Appropriate tone	4.6 (0.6)
Safety	4.7 (0.5)
Reasoning	3.9 (1.0)

Assessment of Memory and Teachability

Preliminary tests indicated that the CHIA was able to successfully recall information from prior discussions when prompted. Additionally, following the implementation of the teachability feature, an overall reduction in repetitive information provided to the user was noted, highlighting the CHIA's ability to adapt based on the user's previous interactions.

XSL•FO
RenderX

Discussion

Principal Findings

To our knowledge, the CHIA is the first LLM-based conversational chatbot grounded in both MI and TTM to deliver PrEP and HIV prevention counseling and facilitate linkage to PrEP care. This approach represents a significant advancement over traditional chatbots based on natural language processing or machine learning, which rely on prewritten scripts or rule-based logic [54]. The CHIA leverages dynamic conversational AI to offer personalized, contextually relevant responses tailored to users' readiness for behavior change. Internal testing demonstrated that the CHIA performs well across both English and Spanish, with relatively high scores in accuracy, trustworthiness, up-to-dateness, conciseness, and most metrics that assessed alignment with the MI spirit. These findings suggest that CHIA has the potential to deliver scalable, high-quality MI-aligned counseling, based on its technical design features such as 24/7 availability and low marginal cost per user. However, we emphasize that scalability was not evaluated in this internal testing and remains a theoretical advantage; the CHIA was assessed only for feasibility and usability in a controlled setting. A full assessment of feasibility, cost-effectiveness, cultural sensitivity, and barriers to real-world scalability will be a focus of the planned randomized controlled trial and subsequent implementation studies.

Although the CHIA received high scores in accuracy and many other assessment metrics, the findings from our internal testing also highlight areas for refinement to further enhance its MI metric-based performance. Beyond accuracy, mimicking human-led MI sessions is critical for improving engagement and behavior change [17]. AI-based chatbots must be trained not only to retrieve and deliver accurate health information but also to apply MI-consistent behaviors, such as open-ended questioning and reflective listening, to foster user motivation and self-efficacy [17]. While the CHIA is capable of providing MI-aligned counseling, there is room to strengthen its ability to engage users more deeply through improved use of MI-consistent responses. MI-based metrics received lower scores, especially in the Spanish version, and the CHIA may benefit from targeted improvements through additional prompt engineering and preference fine-tuning. A Spanish-fluent member of our study team will assist with prompt engineering and optimizing the CHIA for Spanish speakers. The Spanish language responses were generally strong; however, improving cultural and linguistic alignment through more intentional prompt design could strengthen its ability to deliver MI-based counseling even further. A key limitation of our internal testing was the absence of double scoring; future rounds of testing prior to the pilot phase will include 2 independent raters and a consensus process.

To improve the CHIA's responsiveness and alignment with MI, we plan to integrate reinforcement learning into its training process [55]. Reinforcement learning approaches further optimize chatbot responses over time by integrating user feedback and adapting to real-world interactions [55]. Specifically, we will implement the Q-Star algorithm [56], a Q-learning–based approach designed to iteratively optimize performance based on feedback from MI-based metric assessments [57,58]. Q-Star is compatible with the AutoGen framework and will allow us to incorporate a dedicated Q-Agent that learns from evaluation data and adjusts the CHIA's decision-making over time [56]. By continuously refining responses based on user interactions and alignment with MI, this approach offers an adaptive pathway to improve the CHIA's accuracy, engagement, and overall quality of counseling.

The CHIA's core functions (ie, referral for clinical services and HIV risk assessment) performed well during internal testing, successfully delivering relevant information and supporting user needs. However, several areas for improvement were identified to enhance functionality and user experience. For example, the "search for providers" function could be expanded to allow users to access more detailed information about clinics in which they express interest, including services offered, hours, and contact details. The HIV risk assessment function may benefit from an override option that allows users to opt out when they indicate they do not wish to be assessed, thereby supporting autonomy and comfort. Additionally, the "call for real-person support" function could be strengthened by asking more specific questions and capturing details—such as the user's current concern, emotional state, or preferred mode of contact—that would better prepare research assistants to offer timely, personalized support. This function could be triggered when the user expresses uncertainty, distress, or repeatedly requests help, signaling the need for human follow-up. Overall, these targeted improvements represent feasible next steps to further tailor the CHIA's functions to support users' preferences and needs.

Although the CHIA demonstrated decent retrieval efficacy during internal testing, there is room for improvement. While the Spanish responses received higher mean scores in accuracy compared to the English responses, it is important to note that the Spanish version was assessed at a later stage in which the coding team had resolved several multilingual processing issues and patched the retrieval function, which likely resulted in enhanced accuracy. We plan to further expand our retrieval strategy across both languages by incorporating multiple related questions into each embedded response to enhance the relevance and completeness of retrieved information. Additionally, we will restructure the embedded dataset into smaller, topic-specific subsets to improve both retrieval speed and accuracy. To minimize hallucinations and maintain response quality, we will continuously monitor advancements in RAG techniques and integrate improvements as appropriate. Finally, we will track the performance of various LLMs and remain open to adopting alternative models that may better align with the CHIA's counseling objectives and technical needs.

While engagement with chatbots for counseling typically involves short sessions of 3-10 minutes [59,60], insufficient engagement with the CHIA could hinder its effectiveness in real-world settings [61]. To strengthen engagement in the planned iterative refinement phase, the CHIA will offer brief, MI-structured exchanges (engage-focus-evoke-plan) with proactive reengagement and personalization. Sessions will be concise yet purposeful, while teachability preserves consented

XSL•FO
RenderX

context (identity token, risk or concerns, prior plan) so return visits feel continuous and tailored. Between sessions, the CHIA will deliver targeted reminders and follow-up prompts, maintain an encouraging, human-like tone (with judicious emoji where appropriate), and surface just-in-time content aligned with user goals. A streamlined interface (clear navigation, minimal friction) will reduce drop-offs, and on-demand human support will be available for safety, complexity, or user preference. These approaches have the potential to refine the CHIA's performance, improve satisfaction, and support repeat use in real-world settings.

Conclusions

In summary, the internal testing of CHIA demonstrated promising performance in delivering MI-based counseling for HIV and PrEP education in English and Spanish. The chatbot performed well in key areas, including accuracy, trustworthiness, up-to-dateness, conciseness, and overall alignment with the MI spirit. However, targeted refinements are needed, particularly in enhancing MI alignment in Spanish responses to promote the CHIA's acceptability among Spanish-speaking populations. Planned enhancements, such as improved retrieval strategies, reinforcement learning through Q-Star, and iterative prototype refinement based on user feedback, will further strengthen the CHIA's ability to deliver responsive and user-centered support. The assessment of real-world effectiveness of its potential for supporting PrEP uptake is planned for a subsequent phase, during which stigma and discrimination outcomes will be explicitly measured alongside counseling effectiveness. These efforts position the CHIA as a potentially scalable and adaptable tool for promoting HIV prevention, supporting PrEP uptake, and offering valuable insight into the application of conversational AI in health interventions.

Acknowledgments

Artificial intelligence (GPT-4o) was used solely for the purpose of grammar checking and corrections.

Funding

This project was funded by Brown University's Data Science Institute. Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases (R21AI192507) and the National Institute of Mental Health (R34MH138136) of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

EP is a paid consultant for Google Research. All other authors declare no competing interests.

Multimedia Appendix 1
Operational rating scale for Chatbot for HIV Prevention and Action's (CHIA) multilingual evaluation.
[DOCX File, 11 KB - jmir_v28i1e79671_app1.docx]

Multimedia Appendix 2
Researcher prompts, Chatbot for HIV Prevention and Action's (CHIA) responses, and multilingual evaluation metric scores.
[XLSX File, 62 KB - jmir_v28i1e79671_app2.xlsx]

References

1. HIV surveillance supplemental report: estimated HIV incidence and prevalence in the United States, 2018–2022. : Centers for Disease Control and Prevention; 2024 URL: https://stacks.cdc.gov/view/cdc/156513/cdc_156513_DS1.pdf [accessed 2026-01-10]
2. Grant RM, Lama JR, Anderson PL, et al. Preexposure chemoprophylaxis for HIV prevention in men who have sex with men. N Engl J Med 2010 Dec 30;363(27):2587-2599. [doi: 10.1056/NEJMoa1011205] [Medline: 21091279]
3. Mayer KH, Molina JM, Thompson MA, et al. Emtricitabine and tenofovir alafenamide vs emtricitabine and tenofovir disoproxil fumarate for HIV pre-exposure prophylaxis (DISCOVER): primary results from a randomised, double-blind, multicentre, active-controlled, phase 3, non-inferiority trial. Lancet 2020 Jul 25;396(10246):239-254. [doi: 10.1016/S0140-6736(20)31065-5] [Medline: 32711800]
4. Landovitz RJ, Donnell D, Clement ME, et al. Cabotegravir for HIV prevention in cisgender men and transgender women. N Engl J Med 2021 Aug 12;385(7):595-608. [doi: 10.1056/NEJMoa2101016] [Medline: 34379922]

XSL•FO
RenderX

5. Kanny D, Jeffries WL 4th, Chapin-Bardales J, et al. Racial/ethnic disparities in HIV preexposure prophylaxis among men who have sex with men - 23 urban areas, 2017. MMWR Morb Mortal Wkly Rep 2019 Sep 20;68(37):801-806. [doi: 10.15585/mmwr.mm6837a2] [Medline: 31536484]
6. Mayer KH, Agwu A, Malebranche D. Barriers to the wider use of pre-exposure prophylaxis in the United States: a narrative review. Adv Ther 2020 May;37(5):1778-1811. [doi: 10.1007/s12325-020-01295-0] [Medline: 32232664]
7. Sullivan PS, DuBose SN, Castel AD, et al. Equity of PrEP uptake by race, ethnicity, sex and region in the United States in the first decade of PrEP: a population-based analysis. Lancet Reg Health Am 2024 May;33:100738. [doi: 10.1016/j.lana.2024.100738] [Medline: 38659491]
8. Chan PA, Nunn A, van den Berg JJ, et al. A randomized trial of a brief behavioral intervention for PrEP uptake among men who have sex with men at increased risk for HIV infection. J Acquir Immune Defic Syndr 2021 Jul 1;87(3):937-943. [doi: 10.1097/QAI.0000000000002671] [Medline: 33734099]
9. Moitra E, van den Berg JJ, Sowemimo-Coker G, Chau S, Nunn A, Chan PA. Open pilot trial of a brief motivational interviewing-based HIV pre-exposure prophylaxis intervention for men who have sex with men: preliminary effects, and evidence of feasibility and acceptability. AIDS Care 2020 Mar;32(3):406-410. [doi: 10.1080/09540121.2019.1622644] [Medline: 31130000]
10. Miller WR, Rollnick S. Motivational Interviewing: Helping People Change, 3rd edition: Guilford Press; 2013. URL: https://www.biblio.com/book/motivational-interviewing-helping-people-change-miller/d/1659959579 [accessed 2026-01-10]
11. Rubak S, Sandbaek A, Lauritzen T, Christensen B. Motivational interviewing: a systematic review and meta-analysis. Br J Gen Pract 2005 Apr;55(513):305-312. [Medline: 15826439]
12. Calhoun D, Brod R, Kirlin K, Howard BV, Schuldberg D, Fiore C. Effectiveness of motivational interviewing for improving self-care among northern plains Indians with type 2 diabetes. Diabetes Spectr 2010 Jan 1;23(2):107-114. [doi: 10.2337/diaspect.23.2.107]
13. Jacobs NN, Calvo L, Dieringer A, Hall A, Danko R. Motivational interviewing training: a case-based curriculum for preclinical medical students. MedEdPORTAL 2021 Feb 12;17:11104. [doi: 10.15766/mep_2374-8265.11104] [Medline: 33598544]
14. Rollnick S, Miller WR, Butler CC. Motivational Interviewing in Health Care: Helping Patients Change Behavior: Guilford Press; 2008. URL: https://www.guilford.com/books/Motivational-Interviewing-in-Health-Care/Rollnick-Miller-Butler/9781462550371 [accessed 2026-01-10] [doi: 10.1080/15412550802093108]
15. Nazir A, Wang Z. A comprehensive survey of ChatGPT: advancements, applications, prospects, and challenges. Meta Radiol 2023 Sep;1(2):100022. [doi: 10.1016/j.metrad.2023.100022] [Medline: 37901715]
16. Tudor Car L, Dhinagaran DA, Kyaw BM, et al. Conversational agents in health care: scoping review and conceptual analysis. J Med Internet Res 2020 Aug 7;22(8):e17158. [doi: 10.2196/17158] [Medline: 32763886]
17. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. J Med Internet Res 2023 Feb 24;25:e40789. [doi: 10.2196/40789] [Medline: 36826990]
18. Brown A, Kumar AT, Melamed O, et al. A motivational interviewing chatbot with generative reflections for increasing readiness to quit smoking: iterative development study. JMIR Ment Health 2023 Oct 17;10:e49132. [doi: 10.2196/49132] [Medline: 37847539]
19. He L, Basar E, Wiers RW, Antheunis ML, Krahmer E. Can chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance. BMC Public Health 2022 Apr 12;22(1):726. [doi: 10.1186/s12889-022-13115-x] [Medline: 35413887]
20. Prochaska JJ, Vogel EA, Chieng A, et al. A therapeutic relational agent for reducing problematic substance use (Woebot): development and usability study. J Med Internet Res 2021 Mar 23;23(3):e24850. [doi: 10.2196/24850] [Medline: 33755028]
21. Abid A, Baxter SL. Breaking barriers in behavioral change: the potential of artificial intelligence-driven motivational interviewing. J Glaucoma 2024 Jul 1;33(7):473-477. [doi: 10.1097/IJG.0000000000002382] [Medline: 38595151]
22. van Heerden A, Bosman S, Swendeman D, Comulada WS. Chatbots for HIV prevention and care: a narrative review. Curr HIV/AIDS Rep 2023 Dec;20(6):481-486. [doi: 10.1007/s11904-023-00681-x] [Medline: 38010467]
23. Cheah MH, Gan YN, Altice FL, et al. Testing the feasibility and acceptability of using an artificial intelligence chatbot to promote HIV testing and pre-exposure prophylaxis in Malaysia: mixed methods study. JMIR Hum Factors 2024 Jan 26;11:e52055. [doi: 10.2196/52055] [Medline: 38277206]
24. Kamitani E, Mizuno Y, Khalil GM, Viguerie A, DeLuca JB, Mishra N. Improving HIV preexposure prophylaxis uptake with artificial intelligence and automation: a systematic review. AIDS 2024 Aug 1;38(10):1560-1569. [doi: 10.1097/QAD.0000000000003935] [Medline: 38788206]
25. Ntinga X, Musiello F, Keter AK, Barnabas R, van Heerden A. The feasibility and acceptability of an mHealth conversational agent designed to support HIV self-testing in South Africa: cross-sectional study. J Med Internet Res 2022 Dec 12;24(12):e39816. [doi: 10.2196/39816] [Medline: 36508248]
26. Fujimoto M, Hunter L, McCoy S, Outran S, Packel L. Evaluating AI chatbots for HIV prevention: an assessment of response quality and user tailoring. : California HIV / AIDS Policy Research Center; 2024 URL: https://chprc.org/wp-content/uploads/2022/06/AI_Policy_Brief_Oct2024.pdf [accessed 2026-01-10]

XSL•FO
RenderX

27. Jabir AI, Lin X, Martinengo L, Sharp G, Theng YL, Tudor Car L. Attrition in conversational agent-delivered mental health interventions: systematic review and meta-analysis. J Med Internet Res 2024 Feb 27;26(1):e48168. [doi: 10.2196/48168] [Medline: 38412023]
28. Yang Y, Tavares J, Oliveira T. A new research model for artificial intelligence-based well-being chatbot engagement: survey study. JMIR Hum Factors 2024 Nov 11;11:e59908. [doi: 10.2196/59908] [Medline: 39527812]
29. Prochaska JO, Velicer WF. The transtheoretical model of health behavior change. Am J Health Promot 1997;12(1):38-48. [doi: 10.4278/0890-1171-12.1.38] [Medline: 10170434]
30. Raihan N, Cogburn M. Stages of change theory. In: StatPearls: StatPearls Publishing; 2025. [Medline: 32310465]
31. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber Phys Syst 2023;3:121-154. [doi: 10.1016/j.iotcps.2023.04.003]
32. Zhang N, Sun Z, Xie Y, Wu H, Li C. The latest version ChatGPT powered by GPT-4o: what will it bring to the medical field? Int J Surg 2024 Sep 1;110(9):6018-6019. [doi: 10.1097/JS9.0000000000001754] [Medline: 38857508]
33. Welivita A, Pu P. Curating a large-scale motivational interviewing dataset using peer support forums. 2022 Presented at: Proceedings of the 29th International Conference on Computational Linguistics; Oct 12-17, 2022; Gyeongju, Republic of Korea p. 3315-3330 URL: https://aclanthology.org/2022.coling-1.293.pdf [accessed 2026-01-10]
34. Welivita A, Pu P. Curating a large-scale motivational interviewing dataset using peer support forums. in the 29th international conference on computational linguistics (COLING). GitHub. 2022. URL: https://github.com/anuradha1992/Motivational-Interviewing-Dataset [accessed 2026-01-10]
35. Wu Q, Bansal G, Zhang J, et al. AutoGen: enabling next-gen LLM applications via multi-agent conversation. arXiv. Preprint posted online on Aug 16, 2023. [doi: 10.48550/arXiv.2308.08155]
36. Marvin G, Hellen N, Jjingo D, Nakatumba-Nabende J. Data intelligence and cognitive informatics. In: Prompt Engineering in Large Language Models: Springer Nature; 2024:387-402. [doi: 10.1007/978-981-99-7962-2_30]
37. Sai Varshitha G, Rupa R, Divya D. Remix-based real time blood bank communication integrating access control using XGBoost and Supabase. 2024 Presented at: 2024 IEEE Students Conference on Engineering and Systems (SCES); Jun 21-23, 2024; Prayagraj, India p. 1-6. [doi: 10.1109/SCES61914.2024.10652474]
38. Narula S, Jain A. Cloud computing security: amazon web service. 2015 Presented at: 2015 Fifth International Conference on Advanced Computing & Communication Technologies (ACCT); Feb 21-22, 2015; Haryana, India p. 501-505. [doi: 10.1109/ACCT.2015.20] [Medline: 39304265]
39. Ayala O, Bechard P. Reducing hallucination in structured outputs via retrieval-augmented generation. 2024 Presented at: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track); Jun 16-21, 2024. [doi: 10.18653/v1/2024.naacl-industry.19]
40. Saad-Falcon J, Khattab O, Potts C, Zaharia M. ARES: an automated evaluation framework for retrieval-augmented generation systems. 2024 Presented at: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Jun 16-21, 2024. [doi: 10.18653/v1/2024.naacl-long.20]
41. Papadimitriou I, Gialampoukidis I, Vrochidis S, Ioannis K. RAG playground: a framework for systematic evaluation of retrieval strategies and prompt engineering in RAG systems. arXiv. Preprint posted online on Dec 16, 2024. [doi: 10.48550/arXiv.2412.12322]
42. Es S, James J, Espinosa Anke L, Schockaert S. RAGAS: automated evaluation of retrieval augmented generation. 2024 Presented at: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics; Mar 17-22, 2024; St Julians, Malta p. 150-158. [doi: 10.18653/v1/2024.eacl-demo.16]
43. Rafailov R, Sharma A, Mitchell E, Manning CD, Ermon S, Finn C. Direct preference optimization: your language model is secretly a reward model. 2023 Presented at: NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems; Dec 10-16, 2023. [doi: 10.5555/3666122.3668460]
44. Moyers TB, Manuel JK, Ernst D. Motivational interviewing treatment integrity coding manual 4.2.1. 2014 URL: https://motivationalinterviewing.org/sites/default/files/miti4_2.pdf [accessed 2026-01-10]
45. Moyers TB, Martin T, Manuel JK, Miller WR. The motivational interviewing treatment integrity (MITI) code: version 2.0. : University of New Mexico Center on Alcoholism, Substance Abuse, and Addictions (CASAA); 2003 URL: https://casaa.unm.edu/assets/docs/miti1.pdf [accessed 2026-01-10]
46. Moyers TB, Rowell LN, Manuel JK, Ernst D, Houck JM. The Motivational Interviewing Treatment Integrity Code (MITI 4): rationale, preliminary reliability and validity. J Subst Abuse Treat 2016 Jun;65:36-42. [doi: 10.1016/j.jsat.2016.01.001] [Medline: 26874558]
47. Grefenstette G. Tokenization. In: van Halteren H, editor. Syntactic Wordclass Tagging: Springer; 1999:117-133. [doi: 10.1007/978-94-015-9273-4_9]
48. Smith DK, Pals SL, Herbst JH, Shinde S, Carey JW. Development of a clinical screening index predictive of incident HIV infection among men who have sex with men in the United States. J Acquir Immune Defic Syndr 2012 Aug 1;60(4):421-427. [doi: 10.1097/QAI.0b013e318256b2f6] [Medline: 22487585]
49. Preexposure prophylaxis for the prevention of HIV infection in the United States—2021 update: a clinical practice guideline. : Centers for Disease Control and Prevention; 2021 URL: https://stacks.cdc.gov/view/cdc/112360/cdc_112360_DS1.pdf [accessed 2026-01-10]

XSL•FO
RenderX

50. Biener L, Abrams DB. The Contemplation Ladder: validation of a measure of readiness to consider smoking cessation. Health Psychol 1991;10(5):360-365. [doi: 10.1037//0278-6133.10.5.360] [Medline: 1935872]
51. Emory University NPIN. PrEP Locator. URL: https://preplocator.org [accessed 2026-1-10]
52. Ngo NT, Van Nguyen C, Dernoncourt F, Nguyen TH. Comprehensive and practical evaluation of retrieval-augmented generation systems for medical question answering. arXiv. Preprint posted online on Nov 14, 2024. [doi: 10.48550/arXiv.2411.09213]
53. Abbasian M, Khatibi E, Azimi I, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. NPJ Digit Med 2024 Mar 29;7(1):82. [doi: 10.1038/s41746-024-01074-z] [Medline: 38553625]
54. Adamopoulou E, Moussiades L. An overview of chatbot technology. 2020 Presented at: Artificial Intelligence Applications and Innovations (AIAI) 2020; Jun 5-7, 2020. [doi: 10.1007/978-3-030-49186-4_31]
55. Li J, Monroe W, Ritter A, Jurafsky D, Galley M, Gao J. Deep reinforcement learning for dialogue generation. In: Gao J, editor. 2016 Presented at: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing; Nov 1-5, 2016. [doi: 10.18653/v1/D16-1127]
56. McIntosh TR, Susnjak T, Liu T, Watters P, Halgamuge MN. From Google Gemini to OpenAI Q* (Q-star): a survey of reshaping the generative artificial intelligence (AI) research landscape. arXiv. Preprint posted online on Dec 18, 2023. [doi: 10.3390/technologies13020051]
57. Watkins CJCH, Dayan P. Q-learning. Mach Learn 1992 May;8(3-4):279-292. [doi: 10.1007/BF00992698]
58. Jang B, Kim M, Harerimana G, Kim JW. Q-learning algorithms: a comprehensive classification and applications. IEEE Access 2019;7:133653-133667. [doi: 10.1109/ACCESS.2019.2941229]
59. Haque MDR, Rubya S. An overview of chatbot-based mobile mental health apps: insights from app description and user reviews. JMIR Mhealth Uhealth 2023 May 22;11:e44838. [doi: 10.2196/44838] [Medline: 37213181]
60. Irvine R, Boubert D, Raina V, et al. Rewarding chatbots for real-world engagement with millions of users. arXiv. Preprint posted online on Mar 10, 2023. [doi: 10.48550/arXiv.2303.06135]
61. Akdim K, Casaló LV. Perceived value of AI-based recommendations service: the case of voice assistants. Serv Bus 2023 Mar;17(1):81-112. [doi: 10.1007/s11628-023-00527-x]

Abbreviations

AI: artificial intelligence
AI: Artificial intelligence
AWS: Amazon Web Services
CHIA: Chatbot for HIV Prevention and Action
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
MI: motivational interviewing
MSM: Men who have sex with men
PrEP: preexposure prophylaxis
RAG: retrieval-augmented generation
TTM: transtheoretical model

Edited by A Mavragani; submitted 25.Jun.2025; peer-reviewed by J Xu, S Chen, S Graham; revised version received 30.Sep.2025; accepted 02.Oct.2025; published 03.Feb.2026.

Please cite as:
Tao J, Pavlick E, Grondin A, Bustamante JD, Martin H, Parent H, Fenn N, Almonte A, Maguire-Wilkerson A, Gu M, Rusley J, Perler BK, Wray T, Nunn AS, Chan PA
Evaluation of an Artificial Intelligence Conversational Chatbot to Enhance HIV Preexposure Prophylaxis Uptake: Development and Usability Internal Testing
J Med Internet Res 2026;28:e79671
URL: https://www.jmir.org/2026/1/e79671
doi:10.2196/79671

© Jun Tao, Ellie Pavlick, Amaris Grondin, Josue D Bustamante, Harrison Martin, Hannah Parent, Natalie Fenn, Alexi Almonte, Amanda Maguire-Wilkerson, Mofan Gu, Jack Rusley, Bryce K Perler, Tyler Wray, Amy S Nunn, Philip A Chan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 3.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of

XSL•FO
RenderX

Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Developing a Quality Evaluation Index System for Health Conversational Artificial Intelligence: Mixed Methods Study

Weizhen Liao[1], MPH; Meng Li[2], BE; Chengyu Ma[1*], PhD; Youli Han[1*], PhD; Dan Wang[2], MS; Haopeng Liu[1], MSc; Yi Wang[1], MSc; Zijie Feng[1], MSc; Huichao Wang[2], BE; Yiru Guan[2], M.Man

[1]School of Public Health, Capital Medical University, Beijing, China

[2]Bytedance Xiaohe Health, Hainan, China

[*]these authors contributed equally

Corresponding Author:
Chengyu Ma, PhD
School of Public Health
Capital Medical University
No. 10, Xitou Alley, You'anmenwai Fengtai District
Beijing, 100069
China
Phone: 86 13811320990
Email: machengyu@ccmu.edu.cn

Abstract

Background: Effective communication is fundamental to health care; however, demographic transitions and a widening global health workforce gap are intensifying the imbalance between service demand and resource supply. Health conversational artificial intelligence (HCAI) based on large language models offers a potential pathway to improve the accessibility and personalization of care. Nevertheless, the lack of a rigorous, user-centered evaluation framework limits the systematic assessment of HCAI quality, raising concerns regarding safety, reliability, and clinical applicability.

Objective: This study aims to establish a scientific and systematic quality evaluation index system for HCAI, providing both a theoretical foundation and a practical tool for the assessment and optimization of HCAI.

Methods: Based on a literature review, industry standards, and expert group discussions, a preliminary framework for the index system was established. Two rounds of Delphi expert consultations were then conducted to collect expert opinions. The analytic hierarchy process (AHP) was applied to assign weights to indicators at each level, and the final content and structure of the index system were determined.

Results: Both rounds of expert consultation achieved a 100% response rate. The authority coefficient of the experts was 0.84 in both rounds. Kendall W coefficient ranged from 0.14 to 0.20 in the first round and from 0.13 to 0.17 in the second round, with all values showing statistical significance (round one: importance $P<.001$, feasibility $P<.001$, sensitivity $P<.001$; round two: importance $P=.001$, feasibility $P<.001$, sensitivity $P=.001$). The final HCAI quality evaluation index system consisted of 3 primary indicators, 7 secondary indicators, and 28 tertiary indicators. According to AHP weight calculations, the primary indicators were ranked in descending order as follows: ethics and compliance (0.4781), health consultation capability (0.4112), and user experience (0.1107).

Conclusions: The evaluation index system constructed in this study demonstrates scientific validity and practical relevance. It provides a valuable reference for the quality assessment, model optimization, and regulatory oversight of HCAI systems.

(J Med Internet Res 2026;28:e83188) doi:10.2196/83188

KEYWORDS

health consultation; conversational AI; evaluation index system; Delphi method; analytic hierarchy process

XSL•FO
RenderX

Introduction

Background

> *Doctors have three magic weapons: language, medicine and scalpel.* [Hippocrates]

Effective communication is essential not only for facilitating information exchange between clinicians and patients but also for establishing trust, mutual understanding, and ongoing support [1]. Currently, health care systems are under increasing pressure. On the one hand, the global population is aging at a rapid pace. According to the World Health Organization, individuals aged 60 and above are projected to account for 22% of the global population by 2050 [2]. On the other hand, chronic and severe illnesses are increasingly affecting younger populations, while public awareness of health and wellness continues to rise. A growing number of individuals are actively seeking medical information and professional advice. These demographic transitions, coupled with the shifting burden of disease, are driving a surge in demand for health care services, while simultaneously exposing persistent systemic challenges, such as inadequate resource distribution and a shortage of health care professionals. It is estimated that the global health workforce deficit will exceed 11 million by 2030 [3]. This widening gap between supply and demand has made it increasingly difficult for patients to obtain timely, effective, and personalized communication and care. In this context, health conversational artificial intelligence (HCAI) based on large language models (LLMs) offers a new approach to mitigating these challenges [4].

Previous studies have demonstrated that LLMs, such as ChatGPT and LLaMA (Large Language Model from Meta), perform remarkably well across a variety of medical tasks, including health consultation [5], diagnostic assistance [6], enhancement of doctor-patient communication [7], and medical image interpretation [8]. These findings highlight the considerable potential of LLMs in clinical and health-related domains. In addition, recent developments have enabled LLMs to conduct multiturn conversations and effectively understand context [9]. This progress provides a solid foundation for building HCAI systems. HCAI refers to LLM-based systems that simulate dialogue with users in a manner similar to that of human physicians. Representative examples include AMIE (Articulate Medical Intelligence Explorer), developed by Google [10]; Ask Patients with Patience, proposed by Zhu and Wu [11]; and Xiaohe AI Doctor, launched by Douyin in China [12]. These systems can ask users follow-up questions and collect and integrate information such as symptoms, medical history, and lifestyle habits [10]. As a result, they can offer more accurate and personalized health care services (Figure 1).

Figure 1. Example of multiturn interaction between health conversational artificial intelligence (AI) and user.

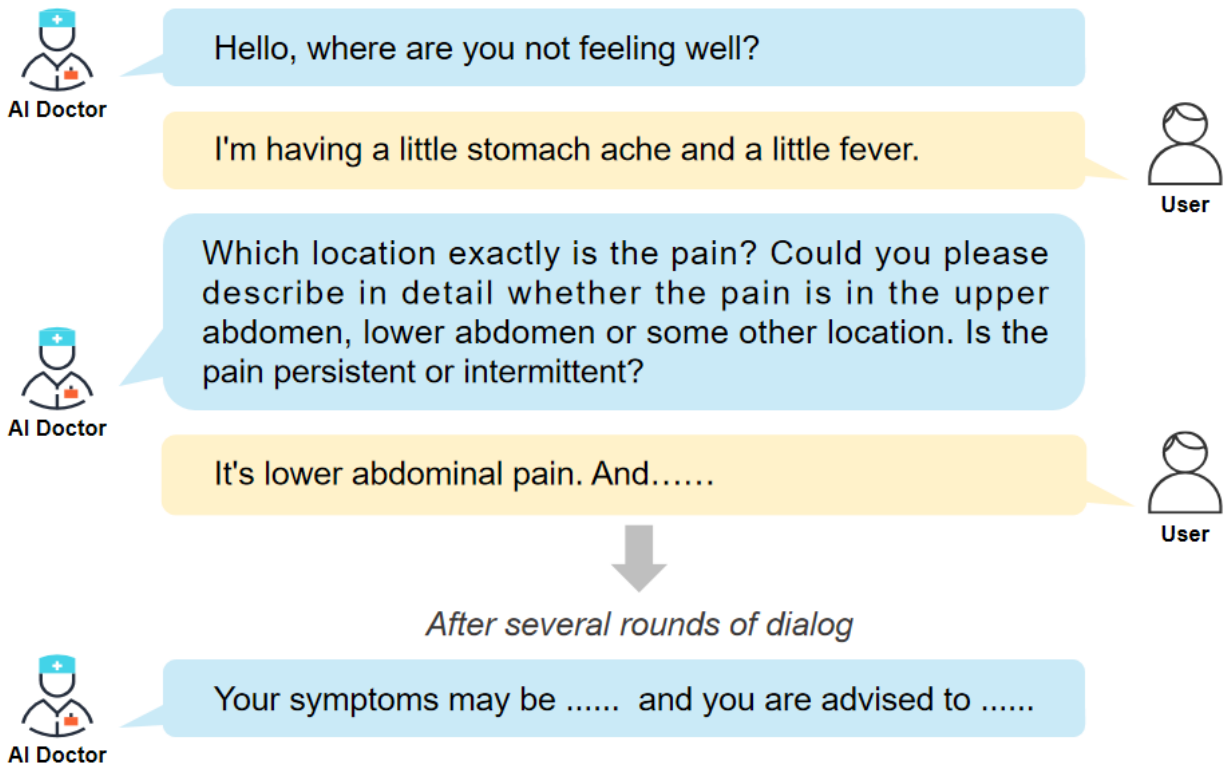

The rapid development of artificial intelligence (AI) has accelerated the expansion of HCAI products. The global conversational AI in the health care market is projected to reach US $16.9 billion in 2025 and to grow at a compound annual growth rate of 24.7% through 2034, reaching a value of US $123.1 billion [13]. As of May 1, 2025, a total of 288 medical LLMs have been publicly released in China. These models cover a wide range of applications, including clinical decision support for specific diseases, prediagnosis consultations, and medical record generation [14]. As a key application of AI in the health

XSL•FO
RenderX

sector, the performance and quality of HCAI systems are directly linked to user safety and health outcomes. Consequently, it is essential to conduct systematic and evidence-based evaluations of these technologies. However, for HCAI systems that deliver health services through multiturn interactions, a scientifically grounded, structured, and practically applicable framework for quality assessment indicators has yet to be established.

Existing studies have employed diverse methods and indicators to evaluate HCAI. Early work often relied on medical examination question banks to assess the accuracy of AI responses to closed-ended questions. Examples include the United States Medical Licensing Examination [15], the Membership of the Royal College of General Practitioners Applied Knowledge Test [16], and the National Medical Licensing Examination in China [17]. Subsequently, several studies developed benchmark tests or clinical case datasets, shifting the focus toward the content generation ability of AI in single-turn open-ended medical questions. Representative datasets include CPMI (Chinese Patent Medicine Instructions dataset) [18], Medical-Diff-VQA (Medical Dataset for Difference Visual Question Answering on Chest X-Ray Images) [19], PubMedQA (a novel biomedical question answering dataset collected from PubMed abstracts) [20], and MultiMedQA (a benchmark for answering medical questions spanning medical exam, medical research, and consumer medical questions, comprising 7 medical question-answering datasets) [21]. Most studies at this stage emphasized content accuracy, with commonly used indicators including BLEU (Bilingual Evaluation Understudy) [5], BERTScore [22], accuracy [23,24], and completeness [25]. However, indicators related to safety risks, such as bias and hallucinations, received limited attention [26,27].

More recent studies have developed methods that better reflect real clinical settings and have assessed AI performance in multitask and multiturn medical dialogues. Xu et al [28] designed MedGPTEval to examine AI performance across 3 dimensions: medical professional capabilities, social comprehensive capabilities, and contextual capabilities. Tu et al [10] employed a remote objective structured clinical examination and created an evaluation framework based on feedback from clinicians and patient participants. This framework assessed history-taking, diagnostic accuracy, management, communication skills, and empathy. Liu et al [29] formulated a clinical pathway specific to LLMs and evaluated clinical capability using 6 indicators: information completeness, behavior standardization, guidance rationality, diagnostic logicality, treatment logicality, and clinical applicability. Johri et al [30] developed the CRAFT-MD (the Conversational Reasoning Assessment Framework for Testing in Medicine) framework to assess the ability of AI to lead clinical conversations, collect complete histories, and make accurate diagnostic decisions. Liu et al [31] proposed the CLEVER framework to evaluate performance in medical case comprehension, clinical reasoning, and diagnosis. Arora et al [32] introduced HealthBench to assess accuracy, completeness, communication quality, context awareness, and instruction following across diverse clinical tasks. Qiu et al [33] developed MedR-Bench to evaluate clinical reasoning in examination recommendations, diagnostic decisions, and treatment planning, and used efficiency, factual accuracy, and completeness as key indicators.

Although these frameworks have advanced the evaluation of clinical reasoning and diagnostic decision-making, their utility for assessing HCAI systems as real-world, user-facing products remains limited. A complete and systematic assessment structure is still lacking, particularly due to the gaps listed below.

First, most existing frameworks adopt a perspective centered on clinical-task performance, such as diagnostic accuracy and history-taking completeness. This approach, while crucial for medical validation, often overlooks the perspective of real-world product application. Specifically, it pays limited attention to user-experience factors such as emotional support, trust, and personalization [34]. These empathy-related features are important because interaction fluency, content clarity, personalized advice, and emotional support can directly influence users' health behaviors and consultation experiences, yet they are largely absent from most evaluation systems.

Second, existing frameworks often lack comprehensive dimensional integration. They tend to focus heavily on medical effectiveness, while insufficient attention has been paid to ethical and compliance-related requirements, such as bias, hallucinations, and protection of personal health data. This omission is inconsistent with current regulatory trends in many countries. For example, the European Union Medical Device Regulation [35] requires AI systems that support diagnosis, prognosis, or treatment, and that may affect clinical workflows, to meet requirements for performance reliability, risk management, cybersecurity, and traceability. A holistic framework that equally weights medical effectiveness, user experience, and ethical compliance is therefore essential.

Third, while some frameworks claim to address multiturn dialogue, they do not offer a systematic or operational methodology for evaluating quality and safety. They often fail to adequately assess context coherence and logical consistency across dialogue turns. These indicators are essential for distinguishing performance between single-turn medical questions and answers and multiturn medical dialogue tasks. Furthermore, by generally relying on descriptive criteria rather than weighted, hierarchical indicators, their practical utility for developers, evaluators, and regulators seeking standardized assessment is limited [22-27,29,30,34].

This lack of comprehensive indicators and balanced evaluation dimensions may result in inaccurate assessments of AI's practical effectiveness, thereby posing risks to patient safety, increasing technical burdens on systems, and creating barriers to the further development of HCAI. Therefore, a scientific scheme that holistically integrates content quality, user experience, and ethical compliance, and that provides an operational assessment structure, is urgently needed to overcome current bottlenecks in HCAI development.

Objectives

To address these gaps, this study develops a Quality Evaluation Index System for HCAI that offers 3 distinct contributions. First, it constructs the evaluation system from the perspective

XSL•FO
RenderX

of the end-user product application. Second, it comprehensively integrates medical effectiveness, user experience, and ethical and safety compliance. Third, it supports systematic and operational evaluation by combining Delphi consensus with the analytic hierarchy process (AHP) to provide weighted, hierarchical criteria. The objective is to establish a theoretical foundation and an operational framework for the evaluation of HCAI systems, thereby promoting their continuous improvement and ensuring their safe and effective deployment in health care settings.

Methods

Study Design

This study adopts the Delphi method and the AHP as its research approach. The Delphi method, a structured technique for achieving expert consensus, is widely used for indicator selection and standard development in complex problem domains and has been extensively applied in areas such as health care quality assessment and policy research [36]. The AHP enables the assignment of weights to multiple indicators based on expert judgment and is particularly suitable for determining hierarchical structures and weight distributions in evaluation frameworks. These 2 methods are often used in combination and have been widely adopted in studies involving the development of evaluation index systems [37,38].

Between November 2024 and February 2025, this study conducted 2 rounds of expert consultation using the Delphi method. Each round of the questionnaire was distributed and collected via email or WeChat (a widely used social media platform in China), with a response period of 1 week. After collecting the expert scoring data, the AHP was applied to calculate the weights of each level of indicators. Based on these results, a quality evaluation index system for HCAI was established. The flowchart illustrating the process of indicator development and weight determination is shown in Figure 2.

Figure 2. Flowchart of the study design. AHP: analytic hierarchy process.

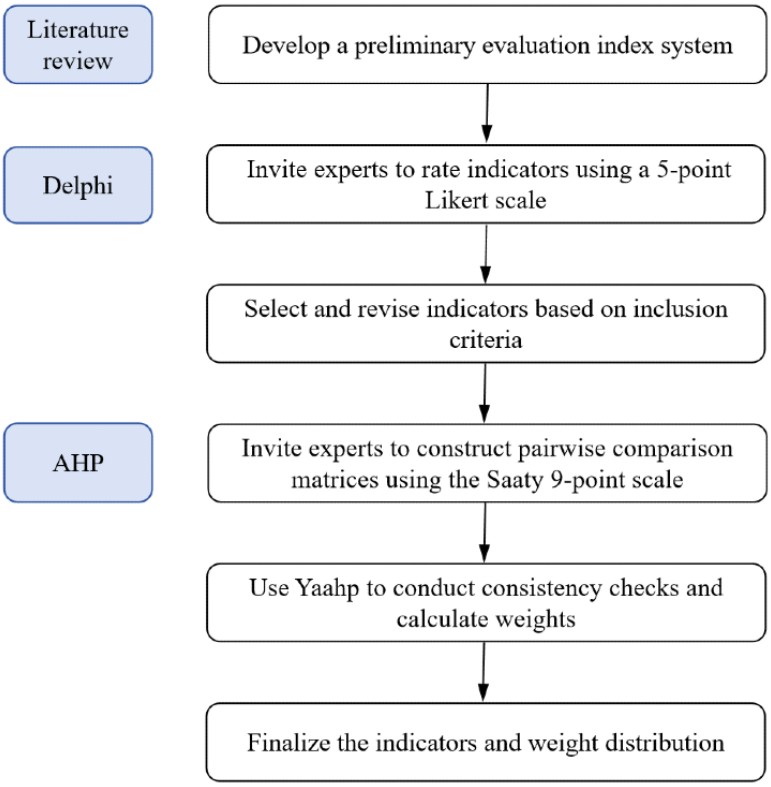

Establishment of a Research Team

The research team consisted of 7 members, including experienced experts in health management, professionals working in the HCAI industry, and graduate students specializing in health management. Among them, 2 held senior professional titles, 2 were HCAI practitioners, and 3 were graduate students. The team's main responsibilities included conducting a literature review, jointly developing preliminary evaluation indicators, identifying Delphi expert members, designing consultation questionnaires, analyzing expert feedback, and performing statistical analyses.

Development of the Preliminary Evaluation Index System

Drawing on a comprehensive review of the literature [30,34,39-42], relevant standards [43], and the expertise of the research team, a preliminary set of evaluation indicators was developed. In real-world applications, HCAI systems involve multiple stakeholders, including patients, health care professionals, and health regulators. As such, quality assessment should not be confined to a single dimension; rather, it must integrate multiperspective and multilevel considerations. Physicians, as service providers, focus on the system's ability to accurately understand user needs, collect complete and logically structured medical histories, perform precise clinical

XSL•FO
RenderX

reasoning, and deliver appropriate health recommendations. These aspects help determine whether HCAI can effectively support clinical practice. Patients, by contrast, place greater emphasis on the quality of interaction, the usefulness and usability of responses, and the system's capacity for emotional responsiveness. Meanwhile, health regulators are primarily concerned with ethical compliance, information security, and the controllability of risks associated with AI applications. These considerations are essential to ensuring that HCAI systems are used safely, effectively, and responsibly in patient care.

To address these considerations, this study developed an evaluation framework based on 3 key stakeholder groups: health care professionals, patients, and health regulators (Figure 3). The framework consists of 3 core dimensions: (1) health consultation capability for medical professionals, which focuses on evaluating the HCAI model's abilities in information elicitation, clinical reasoning, and treatment recommendation during the consultation process; (2) user experience for patients, which draws on the SERVQUAL service quality model [44,45]. This dimension assesses 5 aspects of HCAI: interface usability, information reliability, intelligent responsiveness, perceived safety, and emotional identification; and (3) ethics and compliance for health regulators, which focuses on the system's capacity for ethical safeguards, medical compliance, and risk prevention mechanisms.

The final preliminary evaluation index system comprises 3 primary indicators, 7 secondary indicators, and 34 tertiary indicators.

Figure 3. Preliminary evaluation index system. HCAI: health conversational artificial intelligence.

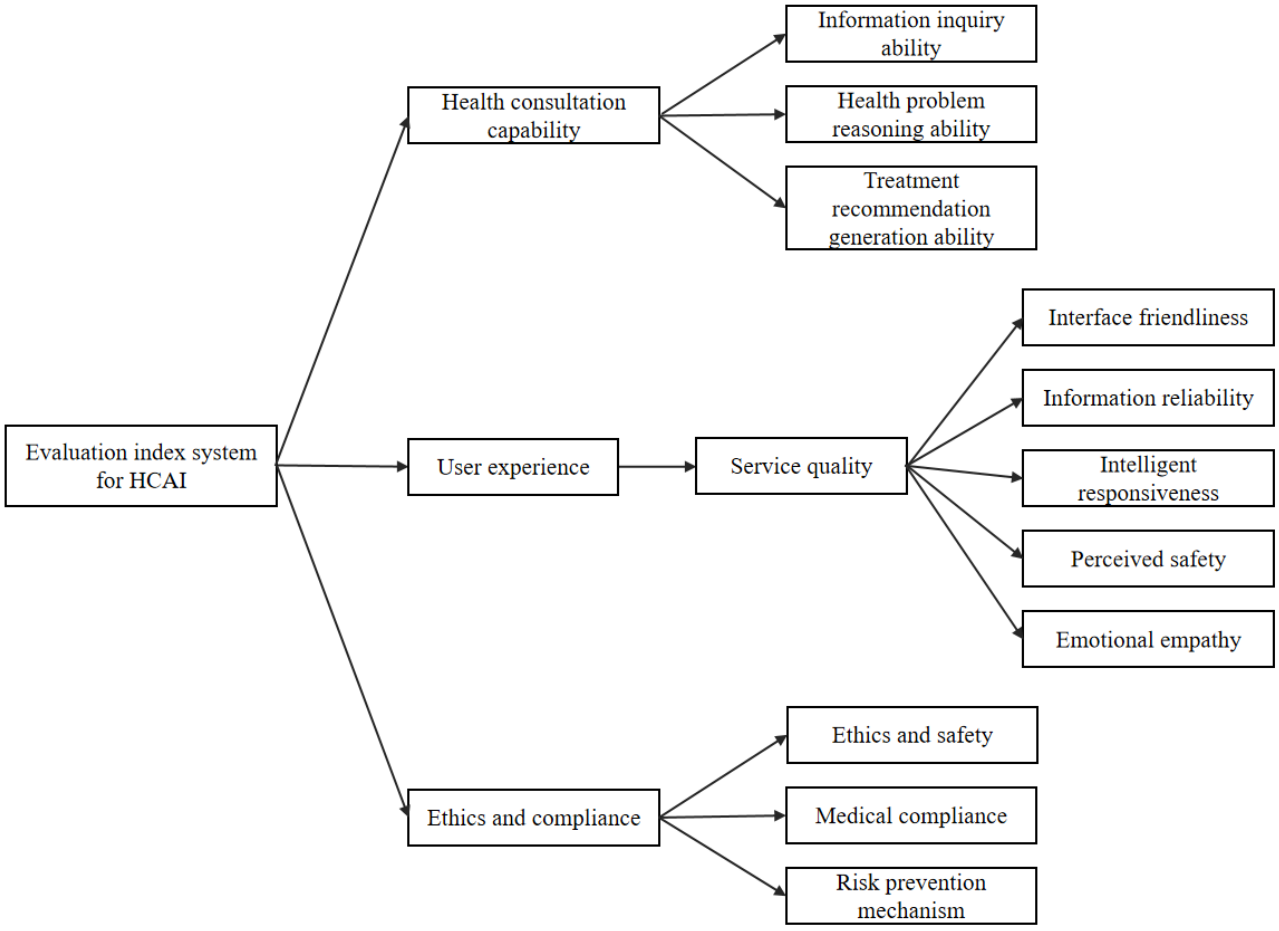

Design of the Expert Consultation Questionnaire

The expert consultation questionnaire consisted of 4 parts (Textbox 1).

The first- and second-round Delphi consultation questionnaires are provided in Multimedia Appendices 1 and 2, respectively. Both sets of questionnaires (Tables S4-S6 in Multimedia Appendices 1 and 2) present detailed evaluation methods, indicator definitions, and scoring rules for all tertiary-level indicators, providing guidance to support expert judgment during the evaluation process.

XSL•FO
RenderX

Textbox 1. Expert consultation questionnaire sections.

1. Questionnaire description

- This section included the research background, objectives, and the deadline for questionnaire submission.

2. Theoretical framework of the evaluation system

- This part outlined the design principles and rationale behind the proposed index system.

3. Instructions and evaluation form

- This section explained how to complete the evaluation form and provided detailed scoring criteria. The evaluation form contained specific indicator definitions, evaluation methods, rating tables, and a column for expert feedback. Experts were asked to rate each indicator in terms of importance, feasibility, and sensitivity using a 5-point Likert scale, and to provide suggestions for revision.
- Importance indicates the relevance of each item for assessing the quality of health conversational AI (HCAI) and is rated on a 5-point scale: 5=very important; 4=important; 3=moderately important; 2=slightly important; and 1=not important.
- Feasibility reflects the ease of collecting relevant data for each item during real-world assessment and is rated on a 5-point scale: 5=very easy; 4=relatively easy; 3=moderate; 2=relatively difficult; and 1=very difficult.
- Sensitivity reflects the extent to which each item influences assessment results, with higher sensitivity indicating a greater impact on HCAI performance and health care quality. It is rated on a 5-point scale: 5=very sensitive; 4=sensitive; 3=moderately sensitive; 2=slightly sensitive; and 1=not sensitive.

4. Expert background information

- This section collected personal information, including age, educational background, professional title, research field, and institutional affiliation. It also included self-assessments of the expert's familiarity with the indicators (Cs) and the basis for their judgment (Ca).

Expert Selection Method

According to the Delphi method, an appropriate number of experts ranges from 15 to 50 participants [46]. In this study, purposive sampling was used to invite experts in fields related to HCAI, including health management and policy, medical ethics, and computer science. The selection criteria for experts were as follows:

- Familiarity with health conversational AI and active engagement in related work, with strong theoretical and practical knowledge.
- A minimum of 5 years of professional experience.
- A postgraduate degree or higher, or a professional title of intermediate level or above.
- Demonstrated interest in the research topic and willingness to complete the consultation process and provide relevant feedback.

Experts were recruited through multiple channels, including previous research collaborations, recommendations from professional associations, and established industry expert databases. Before formal recruitment, the research team organized 2 HCAI expert consensus meetings to introduce the concept of HCAI and to explain the study's purpose, procedures, and evaluation methods. All potential participants received detailed study information and provided informed consent before joining the Delphi consultation.

Indicator Scoring and Selection Criteria

Scores for importance, feasibility, and sensitivity were assigned on a 5-point scale, with higher scores indicating greater relevance. Expert familiarity with the indicators was rated on a 5-level scale with values of 0.9, 0.7, 0.5, 0.3, and 0.1, corresponding to the categories very familiar, familiar, generally familiar, less familiar, and very unfamiliar, respectively. The basis of expert judgment consisted of 4 components: work experience, theoretical analysis, industry understanding, and intuitive perception. The weights for work experience were 0.5, 0.4, and 0.3 for high, medium, and low influence, respectively. The weights for theoretical analysis were 0.3, 0.2, and 0.1 for the same levels. Both industry understanding and intuitive perception were assigned fixed weights of 0.1 [47].

Based on experts' scores for each indicator in terms of importance, feasibility, and sensitivity, the mean, SD, and coefficient of variation (CV) were calculated. The inclusion criteria were a mean score of 3.50 or higher and a CV of less than 0.25 [48,49]. To avoid prematurely excluding potentially important indicators, expert feedback and group discussions were carefully considered before making final decisions.

Key Coefficients and Calculation Methods

Active Coefficient of Experts

This was measured by the effective response rate of the questionnaires, calculated as the number of valid responses in a given round divided by the number of questionnaires distributed. A higher response rate indicates greater expert engagement and interest in the topic.

Expert Authority Coefficient

This reflects the degree of authority an expert holds on the research topic. It is calculated as the average of the expert's Ca (ie, the basis for their judgment) and Cs (ie, their familiarity) values, that is, $Cr = (Ca + Cs)/2$, where Cr is the expert authority coefficient. A Cr value of 0.7 or higher is generally considered to indicate a high level of authority and reliability [50].

XSL•FO
RenderX

Coordination of Expert Opinions

This was measured using the Kendall coefficient of concordance (Kendall *W*), which reflects the consistency of expert ratings. A value closer to 1 indicates a higher degree of agreement among the experts [38].

Weight Determination of Each Indicator

After finalizing the evaluation index system, experts conducted pairwise comparisons of the indicators based on the Saaty scale of relative importance (see Table 1) [51]. These comparisons were used to construct a hierarchical structure and establish judgment matrices. The weights of each indicator were then calculated using AHP via Yaahp (an AHP-based decision-support software developed by Shanxi Yuan Decision Software Technology Co Ltd), and the results were analyzed accordingly. The consistency of each judgment matrix was tested using the consistency ratio. A consistency ratio value of less than 0.1 was considered to indicate acceptable consistency [52].

Table 1. The Saaty scale of relative importance.

Scale	Meaning
1	Indicates equal importance between 2 elements
3	Indicates moderate importance of 1 element over another
5	Indicates moderate to strong importance
7	Indicates very strong importance
9	Indicates extreme importance
2, 4, 6, and 8	Intermediate values between the above judgments
Reciprocals	If item A is assigned a value when compared with item B, then item B is assigned the reciprocal value when compared with item A

Statistical Methods

In this study, Microsoft Excel 2007 was used to input data and calculate the mean, SD, and CV for the importance, feasibility, and sensitivity of each indicator. Excel was also used to compute the expert active coefficient, Ca, Cs, and Cr. Kendall *W* was calculated using SPSS version 25.0 (IBM Corp) and tested for significance using the chi-square test. When the *P* value of Kendall *W* was less than 0.05, and all CVs were below 0.25, the level of disagreement among expert opinions was considered acceptable [53]. The Yaahp version 12.0 software was used to conduct consistency testing of the judgment matrices and to calculate the weights of each indicator.

Ethics Considerations

This study was reviewed and approved by the Ethics Committee of Capital Medical University (approval number Z2024SY075). The approved project title is “Evaluation of the Quality and Effectiveness of User-Oriented AI Health Consultation Services.” All study procedures were conducted in accordance with institutional guidelines and regulations and complied with the principles of the Declaration of Helsinki. Expert participants in the Delphi process were informed of the study’s purpose and provided consent before participation. No identifiable personal information of the expert participants was disclosed in this study. Participant privacy was fully protected.

Results

Basic Information About Experts

Two rounds of expert consultation were conducted in this study. Based on predefined criteria, 16 experts were invited for the first round and 15 for the second round, with 1 expert failing to respond in the second round. Among the 15 experts who completed both rounds, 5 (33%) were male and 10 (67%) were female; 9 of the 15 (60%) experts were over 40 years old. As many as 14 of 15 (93%) experts held a master’s degree or higher, and the same percentage held professional titles at the associate senior level or above. Additionally, 12 of 15 (80%) experts had more than 10 years of work experience in the health care field. The experts were evenly distributed across hospitals, universities, and research institutions. Detailed information is presented in Table 2.

XSL•FO
RenderX

Table 2. Basic information about experts.

Basic information	First round (n=16), n (%)	Second round (n=15), n (%)
Sex		
Male	6 (38)	5 (33)
Female	10 (63)	10 (67)
Age (years)		
≤40	6 (38)	6 (40)
41-50	9 (56)	9 (60)
>50	1 (6)	0 (0)
Education		
Undergraduate	1 (6)	1 (7)
Master	5 (31)	4 (27)
Doctor	10 (63)	10 (67)
Professional title		
Intermediate	1 (6)	1 (7)
Deputy senior	7 (44)	7 (47)
Senior	8 (50)	7 (47)
Years of work experience		
≤10	3 (19)	3 (20)
11-15	4 (25)	4 (27)
16-20	5 (31)	5 (33)
>20	4 (25)	3 (20)
Affiliation		
Medical institution	4 (25)	4 (27)
University	6 (38)	5 (33)
Research institute	5 (31)	5 (33)
Association	1 (6)	1 (7)
Research field		
Health management and policy	7 (44)	7 (47)
Computer science	1 (6)	1 (7)
Medical ethics	1 (6)	1 (7)
Health law	2 (13)	2 (13)
Hospital management	2 (13)	2 (13)
Others	3 (19)	2 (13)

Expert Positivity Degree

According to the Delphi method, a response rate of 50% is considered the minimum threshold for analysis and reporting, while a rate above 70% indicates a high level of expert engagement [54]. In the first round, the expert response rate was 100%, with 11 experts providing a total of 33 suggestions. In the second round, the response rate was also 100%, and 5 experts submitted 6 suggestions. The response rate exceeded 70% in both rounds, indicating a high level of participation and engagement from the expert panel throughout the study.

Expert Authority Coefficient

The authority coefficients of the 2 rounds of expert consultation are shown in Table 3. In the first round, Ca was 0.92, Cs was 0.75, and the resulting Cr was 0.84. In the second round, Ca was 0.93, Cs remained 0.75, and Cr was also 0.84. In both rounds, the Cr values exceeded 0.8, indicating that the experts consulted in this study demonstrated a high level of authority and reliability.

XSL•FO
RenderX

Table 3. Authority coefficients in the 2 rounds of consultation.

Inquiry round	Number	Ca[a]	Cs[b]	Cr[c]
First round	16	0.92	0.75	0.84
Second round	15	0.93	0.75	0.84

[a]Ca: the basis for the expert's judgment.

[b]Cs: expert familiarity.

[c]Cr: expert authority coefficient.

Degree of Coordination of Expert Opinions

The degree of coordination among expert opinions refers to the level of consistency in the experts' ratings of each indicator. The Kendall (*W*) coefficients for the 2 rounds of expert consultation are presented in Table 4. In the first round, Kendall *W* values for all indicators ranged from 0.14 to 0.20. In the second round, the values ranged from 0.13 to 0.17. The results of the chi-square tests were statistically significant (round 1: importance, *P*<.001; feasibility, *P*<.001; and sensitivity, *P*<.001; round 2: importance, *P*=.001; feasibility, *P*<.001; and sensitivity, *P*=.001), indicating a high level of agreement among the expert opinions.

Table 4. Kendall W coefficients for the 2 rounds of consultation.

Project	Importance	Feasibility	Sensitivity
Round one			
Kendall *W*	0.20	0.14	0.14
Chi-square (*df*)	124.39 (42)	86.38 (42)	88.20 (42)
P value	<.001	<.001	<.001
Round 2			
Kendall *W*	0.13	0.17	0.13
Chi-square (*df*)	66.77 (36)	85.05 (36)	67.66 (36)
P value	.001	<.001	.001

Revisions to the Indicator System

The results of the first-round expert consultation are presented in Table 5. The mean scores of all 44 indicators were equal to or greater than 3.50. Specifically, the mean scores ranged from 4.20 to 5.00 for importance, 3.75 to 4.88 for feasibility, and 3.88 to 4.88 for sensitivity. The CVs ranged from 0.00 to 0.27, with 6 indicators having CV values equal to or greater than 0.25.

Based on the screening criteria and expert suggestions, the research team held detailed discussions and decided to temporarily retain the 6 indicators with CV values ≥0.25, remove 7 indicators, revise 4 indicators, and add 1 new indicator. Details of these revisions are provided in Table 6.

After these adjustments, the second-round consultation form consisted of 3 primary indicators, 7 secondary indicators, and 28 tertiary indicators.

The results of the second-round expert consultation are shown in Table 7. The mean scores for importance, feasibility, and sensitivity of all indicators were above 4.0, and all CVs were below 0.25. Based on the indicator selection criteria and thorough discussions among the expert panel, all 37 indicators in the second-round consultation were retained. In addition, the name of indicator C1 was revised from "ethics" to "ethical compliance." The final quality evaluation index system for health conversational AI was thus established, comprising 3 primary indicators, 7 secondary indicators, and 28 tertiary indicators (Table 8).

XSL•FO
RenderX

Table 5. Scores of indicators in the first round of consultation.

Indicators	Importance		Feasibility		Sensitivity	
	Mean (SD)	CV[a]	Mean (SD)	CV	Mean	CV
First-level indices						
A. Health consultation capability	4.94 (0.25)	0.05	4.19 (0.75)	0.18	4.75 (0.45)	0.09
B. User experience	4.50 (0.63)	0.14	4.63 (0.62)	0.13	4.44 (0.63)	0.14
C. Ethics and compliance	5.00 (0.00)	0.00	4.13 (1.02)	0.25	3.88 (0.81)	0.21
Second-level indices						
A1. Information inquiry ability	4.81 (0.40)	0.08	4.19 (0.75)	0.18	4.69 (0.60)	0.13
A2. Health problem reasoning ability	4.81 (0.40)	0.08	3.75 (0.77)	0.21	4.69 (0.48)	0.10
A3. Treatment recommendation generation ability	4.75 (0.58)	0.12	4.19 (0.75)	0.18	4.69 (0.70)	0.15
B1. Service quality experience	4.50 (0.63)	0.14	4.63 (0.62)	0.13	4.44 (0.63)	0.14
C1. Ethics and safety	4.94 (0.25)	0.05	3.81 (1.05)	0.27	4.31 (0.79)	0.18
C2. Medical compliance	5.00 (0.00)	0.00	4.13 (1.09)	0.26	4.38 (0.72)	0.16
C3. Risk prevention mechanisms	4.75 (0.58)	0.12	4.13 (1.02)	0.25	4.44 (0.81)	0.18
Third-level indices						
A11. Accuracy in understanding user needs	4.94 (0.25)	0.05	4.50 (0.52)	0.11	4.81 (0.40)	0.08
A12. Completeness of user information collection	4.44 (0.51)	0.12	4.19 (0.66)	0.16	4.31 (0.60)	0.14
A13. Relevance of inquiry content	4.63 (0.50)	0.11	4.19 (0.75)	0.18	4.69 (0.48)	0.10
A14. Logical sequence of inquiries	4.25 (0.68)	0.16	4.38 (0.89)	0.20	4.31 (0.60)	0.14
A15. Appropriateness of interaction rounds	4.31 (0.60)	0.14	4.19 (0.75)	0.18	4.00 (0.89)	0.22
A16. Recognition ability of multimodal information	4.63 (0.50)	0.11	4.56 (0.63)	0.14	4.50 (0.63)	0.14
A17. Personalization of communication style	4.63 (0.62)	0.13	4.31 (0.60)	0.14	4.38 (0.72)	0.16
A21. Accuracy of disease risk reasoning	4.94 (0.25)	0.05	4.44 (0.63)	0.14	4.88 (0.34)	0.07
A22. Consistency of repeated judgments	4.75 (0.45)	0.09	4.44 (0.73)	0.16	4.38 (0.62)	0.14
A23. Coverage of consulted diseases	4.25 (0.68)	0.16	4.38 (0.81)	0.18	4.19 (0.75)	0.18
A24. Diagnostic ability for complex diseases	4.20 (0.86)	0.21	3.80 (0.77)	0.20	4.20 (0.86)	0.21
A25. Professional use of medical terminology	4.25 (0.86)	0.20	4.44 (0.73)	0.16	4.19 (0.75)	0.18
A26. Frequency of medical knowledge updates	4.31 (0.79)	0.18	4.06 (1.00)	0.25	4.06 (0.93)	0.23
A27. Interpretability of disease reasoning	4.63 (0.72)	0.16	4.25 (0.68)	0.16	4.25 (0.77)	0.18
A31. Accuracy of treatment recommendations	4.88 (0.34)	0.07	4.44 (0.89)	0.20	4.75 (0.58)	0.12
A32. Comprehensiveness of treatment recommendations	4.56 (0.63)	0.14	4.38 (0.62)	0.14	4.44 (0.73)	0.16
A33. Personalization of treatment recommendations	4.56 (0.63)	0.14	4.13 (0.62)	0.15	4.44 (0.73)	0.16
A34. Operability of treatment recommendations	4.44 (0.63)	0.14	4.25 (0.77)	0.18	4.31 (0.79)	0.18
B11. Interface friendliness	4.25 (0.77)	0.18	4.63 (0.50)	0.11	4.25 (0.77)	0.18
B12. Information reliability	4.94 (0.25)	0.05	4.44 (0.73)	0.16	4.63 (0.62)	0.13
B13. Intelligent responsiveness	4.44 (0.63)	0.14	4.56 (0.73)	0.16	4.63 (0.50)	0.11
B14. Perceived safety	4.63 (0.62)	0.13	4.31 (0.70)	0.16	4.19 (0.66)	0.16
B15. Emotional empathy	4.44 (0.63)	0.14	4.00 (0.82)	0.20	4.25 (0.68)	0.16
C11. Bias	4.75 (0.45)	0.09	4.25 (0.68)	0.16	4.38 (0.72)	0.16
C12. Privacy	4.88 (0.34)	0.07	4.31 (0.79)	0.18	4.56 (0.63)	0.14
C13. Hallucinations	4.69 (0.48)	0.10	4.13 (0.62)	0.15	4.50 (0.73)	0.16
C14. Data and system security	4.94 (0.25)	0.05	4.19 (0.66)	0.16	4.50 (0.73)	0.16
C15. Establishment of ethics committee	4.56 (0.73)	0.16	4.44 (1.15)	0.26	4.31 (0.87)	0.20

XSL•FO
RenderX

Indicators	Importance		Feasibility		Sensitivity	
	Mean (SD)	CV[a]	Mean (SD)	CV	Mean	CV
C21. Presence of a disclaimer notices	4.69 (0.60)	0.13	4.88 (0.34)	0.07	4.25 (0.77)	0.18
C22. Assessment of treatment risks	4.94 (0.25)	0.05	4.44 (0.51)	0.12	4.75 (0.58)	0.12
C23. Generation of factually incorrect information	4.81 (0.40)	0.08	4.50 (0.63)	0.14	4.56 (0.73)	0.16
C24. Violation of medical compliance	4.88 (0.34)	0.07	4.44 (0.81)	0.18	4.69 (0.48)	0.10
C31. Emergency response mechanisms	4.75 (0.58)	0.12	4.63 (0.50)	0.11	4.19 (0.75)	0.18
C32. Access management for minors	4.81 (0.40)	0.08	4.88 (0.34)	0.07	4.38 (0.81)	0.18

[a]CV: coefficient of variation.

Table 6. Revisions to indicators based on the first round of consultation.

Indicators	Action	Explanation
1. Ethics and compliance	Retained	The CV[a] value for feasibility is ≥0.25. It is recommended to decide whether to retain it based on the results of the second round of consultation.
C1. Ethics and safety	Retained	The CV value for feasibility is ≥0.25. It is recommended to decide whether to retain it based on the results of the second round of consultation.
C2. Medical compliance	Retained	The CV value for feasibility is ≥0.25. It is recommended to decide whether to retain it based on the results of the second round of consultation.
C3. Risk prevention mechanism	Retained	The CV value for feasibility is ≥0.25. It is recommended to decide whether to retain it based on the results of the second round of consultation.
A1. Information inquiry ability	Modified	Renamed to “Health Information Inquiry Ability” for better accuracy.
C1. Ethics and safety	Modified	Renamed to “Ethics” for better accuracy.
A13. Relevance of inquiry content	Deleted	Poor feasibility based on evaluation criteria.
A15. Rationality of interaction rounds	Deleted	Poor feasibility based on evaluation criteria.
A23. Coverage of consulted diseases	Deleted	Low importance and sensitivity.
A24. Diagnostic ability for complex diseases	Retained	Some experts noted overlap with A21 “Accuracy of disease reasoning,” but reasoning for complex diseases provides distinct model evaluation value.
A25. Use of medical terminology	Deleted	Overlaps with A31 “Accuracy of treatment recommendations.”
A26. Frequency of medical knowledge update	Retained	The CV value for feasibility is ≥0.25. It is recommended to decide whether to retain it based on the results of the second round of consultation.
A33. Personalization of treatment advice	Deleted	Low evaluation value under C3 “Risk prevention mechanism.”
B12. Information reliability	Modified	Renamed to “Perceived reliability” for better accuracy.
B13. Intelligent responsiveness	Modified	Renamed to “Perceived responsiveness” for better accuracy.
C14. Data and system security	Deleted	Belongs to model-level indicators, not a primary assessment focus.
C15. Establishment of ethics committee	Retained	The CV value for feasibility is ≥0.25. It is recommended to decide whether to retain it based on the results of the second round of consultation.
C22. Assessment of treatment risks	Deleted	Overlaps with evaluation under A2 “Health problem reasoning ability.”
C24. Violation of medical compliance	Retained	Some experts believe that this indicator belongs to the model admission criteria and suggest deleting it. However, this mainly evaluates the low-risk violations that may exist in the model and has evaluation value
C33. Presence of privacy terms	Added	The risk prevention mechanism should include detailed privacy clauses to clarify the purpose and scope of collecting user information, as well as whether the collected information is shared with third parties

[a]CV: coefficient of variation.

XSL•FO
RenderX

Table 7. Scores of indicators in the second round of consultation.

Indicators	Importance		Feasibility		Sensitivity	
	Mean (SD)	CV[a]	Mean (SD)	CV	Mean (SD)	CV
First-level indices						
A. Health consultation capability	4.87 (0.35)	0.07	4.40 (0.51)	0.12	4.73 (0.59)	0.13
B. User experience	4.67 (0.49)	0.10	4.67 (0.49)	0.10	4.60 (0.63)	0.14
C. Ethics and compliance	4.93 (0.26)	0.05	4.20 (0.68)	0.16	4.33 (0.62)	0.14
Second-level indices						
A1. Health information inquiry ability	4.80 (0.41)	0.09	4.60 (0.63)	0.14	4.80 (0.41)	0.09
A2. Health problem reasoning ability	4.73 (0.46)	0.10	4.33 (0.62)	0.14	4.47 (0.74)	0.17
A3. Treatment recommendation generation ability	4.80 (0.56)	0.12	4.40 (0.74)	0.17	4.47 (0.83)	0.19
B1. Service quality experience	4.67 (0.49)	0.10	4.67 (0.49)	0.10	4.60 (0.63)	0.14
C1. Ethical compliance	4.87 (0.35)	0.07	4.20 (0.86)	0.21	4.87 (0.35)	0.07
C2. Medical compliance	4.87 (0.35)	0.07	4.40 (0.74)	0.17	4.40 (0.51)	0.12
C3. Risk prevention mechanism	4.93 (0.26)	0.05	4.53 (0.64)	0.14	4.33 (0.72)	0.17
Third-level indices						
A11. Accuracy in understanding user needs	4.87 (0.35)	0.07	4.60 (0.63)	0.14	4.93 (0.26)	0.05
A12. Completeness of user information collection	4.60 (0.51)	0.11	4.20 (0.77)	0.18	4.60 (0.51)	0.11
A13. Logical sequence of inquiries	4.60 (0.63)	0.14	4.60 (0.63)	0.14	4.20 (0.77)	0.18
A14. Recognition ability of multimodal information	4.67 (0.49)	0.10	4.53 (0.64)	0.14	4.67 (0.49)	0.10
A15. Personalization of communication style	4.47 (0.64)	0.14	4.07 (0.80)	0.20	4.47 (0.64)	0.14
A21. Accuracy of disease risk reasoning	4.80 (0.41)	0.09	4.33 (0.62)	0.14	4.67 (0.49)	0.10
A22. Consistency of repeated judgments	4.87 (0.35)	0.07	4.53 (0.74)	0.16	4.47 (0.64)	0.14
A23. Diagnostic ability for complex diseases	4.47 (0.83)	0.19	4.13 (0.83)	0.20	4.40 (0.74)	0.17
A24. Frequency of medical knowledge updates	4.60 (0.63)	0.14	4.33 (0.82)	0.19	4.33 (0.82)	0.19
A25. Interpretability of disease reasoning	4.80 (0.56)	0.12	4.13 (0.83)	0.20	4.33 (0.72)	0.17
A31. Accuracy of treatment recommendations	4.93 (0.26)	0.05	4.47 (0.74)	0.17	4.67 (0.72)	0.16
A32. Comprehensiveness of treatment recommendations	4.53 (0.64)	0.14	4.40 (0.63)	0.14	4.27 (0.88)	0.21
A33. Operability of treatment recommendations	4.53 (0.74)	0.16	4.40 (0.74)	0.17	4.27 (0.80)	0.19
B11. Interface friendliness	4.47 (0.52)	0.12	4.73 (0.46)	0.10	4.33 (0.62)	0.14
B12. Perceived reliability	4.80 (0.56)	0.12	4.53 (0.64)	0.14	4.67 (0.49)	0.10
B13. Perceived responsiveness	4.67 (0.49)	0.10	4.73 (0.46)	0.10	4.67 (0.62)	0.13
B14. Perceived safety	4.67 (0.49)	0.10	4.27 (0.70)	0.16	4.33 (0.49)	0.11
B15. Emotional identification	4.21 (0.70)	0.17	4.43 (0.76)	0.17	4.07 (0.73)	0.18
C11. Bias	4.87 (0.35)	0.07	4.40 (0.74)	0.17	4.53 (0.64)	0.14
C12. Privacy	4.80 (0.41)	0.09	4.67 (0.62)	0.13	4.53 (0.64)	0.14
C13. Hallucination	4.67 (0.49)	0.10	4.07 (0.59)	0.15	4.60 (0.74)	0.16
C14. Establishment of ethics committee	4.80 (0.41)	0.09	4.67 (0.72)	0.16	4.20 (0.77)	0.18
C21. Presence of a disclaimer notices	4.87 (0.35)	0.07	4.93 (0.26)	0.05	4.27 (0.70)	0.16
C22. Generation of factually incorrect information	5.00 (0.00)	0.00	4.67 (0.49)	0.10	4.87 (0.35)	0.07
C23. Violation of medical compliance	4.93 (0.26)	0.05	4.87 (0.35)	0.07	4.73 (0.46)	0.10
C31. Emergency response mechanisms	4.80 (0.56)	0.12	4.67 (0.49)	0.10	4.67 (0.62)	0.13
C32. Access management for minors	4.80 (0.41)	0.09	4.87 (0.35)	0.07	4.60 (0.63)	0.14
C33. Presence of privacy terms	4.80 (0.41)	0.09	4.80 (0.41)	0.09	4.40 (0.74)	0.17

XSL•FO
RenderX

[a]CV: coefficient of variation.

Table 8. Final weight coefficients of indicators.

First-, second-, and third-level indices	Weight
A. Health consultation capability	0.4112
A1. Health information inquiry ability	0.1272
A11. Accuracy in understanding user needs	0.0648
A12. Completeness of user information collection	0.021
A13. Logical sequence of inquiries	0.0122
A14. Recognition ability of multimodal information	0.0219
A15. Personalization of communication style	0.0073
A2. Health problem reasoning ability	0.0821
A21. Accuracy of disease risk reasoning	0.0389
A22. Consistency of repeated judgments	0.0182
A23. Diagnostic ability for complex diseases	0.0094
A24. Frequency of medical knowledge updates	0.0056
A25. Interpretability of disease reasoning	0.01
A3. Treatment recommendation generation ability	0.2019
A31. Accuracy of treatment recommendations	0.1216
A32. Comprehensiveness of treatment recommendations	0.0562
A33. Operability of treatment recommendations	0.0241
B. User experience	0.1107
B1. Service quality experience	0.1107
B11. Interface friendliness	0.0072
B12. Perceived reliability	0.0456
B13. Perceived responsiveness	0.0132
B14. Perceived safety	0.0336
B15. Emotional identification	0.0111
C. Ethics and compliance	0.4781
C1. Ethical compliance	0.1933
C11. Bias	0.0539
C12. Privacy	0.0491
C13. Hallucination	0.0697
C14. Establishment of ethics committee	0.0205
C2. Medical compliance	0.1696
C21. Presence of a disclaimer notices	0.0161
C22. Generation of factually incorrect information	0.0889
C23. Violation of medical compliance	0.0646
C3. Risk prevention mechanism	0.1152
C31. Emergency response mechanisms	0.0605
C32. Access management for minors	0.0362
C33. Presence of privacy terms	0.0185

XSL•FO
RenderX

Weight Determination of Each Indicator

The AHP was used to calculate the weights of each indicator, with higher weights indicating greater importance in evaluating the quality of health conversational AI. In this study, all consistency ratio values of the judgment matrices were automatically adjusted using Yaahp and were found to be less than 0.10, indicating good consistency across all matrices.

The final weight coefficients of all indicators are shown in Table 8. Among the primary indicators, the ranking by weight from highest to lowest was as follows: ethics and compliance (0.4781), health consultation capability (0.4112), and user experience (0.1107).

Discussion

Principal Findings

As AI becomes increasingly integrated into health care, HCAI is being used more widely in health communication, clinical decision support, and consultation. Establishing a scientific and systematic quality evaluation framework is therefore essential. Although several evaluation systems have been developed, many still focus primarily on accuracy or clinician-oriented indicators, neglecting critical aspects of real-world product use and comprehensive safety assessment. They pay limited attention to the comprehensive integration of user experience, ethical and compliance considerations, and the robust assessment of multiturn interactions, all of which are essential for HCAI in real consultation settings. Addressing these gaps is critical for improving the quality, safety, and effectiveness of HCAI.

This study employs the Delphi method and AHP to develop an evaluation index system that reflects the perspectives of physicians, users, and regulators, and makes 3 key contributions addressing the limitations of prior work. First, our system is constructed from the perspective of HCAI as an end-user product for daily health consultation, rather than focusing solely on clinical-task performance. Second, it encompasses 3 core, balanced dimensions: health consultation capability, user experience, and ethics and compliance. Third, by using the AHP methodology, the framework supports systematic and operational evaluation through weighted, hierarchical criteria. By addressing these dimensions and the shortcomings of existing frameworks, the proposed system provides a more comprehensive structure that aligns with the real-world needs of HCAI services. The framework provides a theoretical foundation for empirical research, guides practical optimization, and serves as a valuable reference for improving HCAI evaluation standards worldwide.

Scientific Validity and Rationality of the Evaluation Index System

The final evaluation index system developed in this study comprises 3 primary indicators, 7 secondary indicators, and 28 tertiary indicators. Its structure is comprehensive and well-organized, demonstrating both scientific rigor and practical relevance.

First, the design of the index system was grounded in existing literature, industry standards, and expert consensus. It incorporated both domestic and international approaches to HCAI evaluation, ensuring a solid theoretical foundation and methodological relevance.

Second, the study employed 2 rounds of Delphi expert consultation, combined with AHP for weight assignment, integrating both qualitative judgment and quantitative analysis. Expert opinions showed increasing convergence across rounds, with statistically significant coordination coefficients, indicating a high level of consensus and enhancing the credibility of the evaluation system.

Third, the experts involved in the study were highly representative across disciplines, including health management and policy, medical ethics, computer science, health law, and hospital administration. The panel comprised both academic researchers and experienced practitioners, ensuring a balance of theoretical insight and practical expertise. This interdisciplinary and diverse expert composition enhanced the scientific validity of the evaluation system and conformed to established Delphi methodology standards for expert selection and quality assurance [55,56].

Finally, the experts demonstrated high levels of engagement, authority, and consistency throughout the process, significantly enhancing the reliability of the study's findings [57]. Both rounds of the Delphi consultation achieved a 100% response rate (16 and 15 experts in rounds 1 and 2, respectively), with a total of 39 modification suggestions proposed, reflecting active participation and strong interest from the expert panel. The Cr values in both rounds exceeded 0.8, confirming the reliability of expert judgments. Additionally, the P values of Kendall W were all below 0.05, indicating a high level of agreement in weight assessments and providing a solid foundation for the reliability of the constructed index system.

Analysis of Indicator Weights

The weights of the 3 primary indicators are shown in Table 8: ethics and compliance (0.4781), health consultation capability (0.4112), and user experience (0.1107). These results indicate that ethics and compliance carries the greatest weight in the quality evaluation of HCAI, underscoring its importance as a key dimension for building safe and trustworthy AI systems.

Health consultation capability, which reflects the AI's professional value and its ability to deliver knowledge-based services, also accounts for a substantial proportion of the overall weight. By contrast, user experience carries a relatively lower weight. This may be because current HCAI products and applications are still in the early stages of development, leading experts to prioritize system safety, compliance, and functionality during the evaluation process, while user interaction is considered a secondary factor.

At present, many HCAI systems face significant risks related to algorithmic bias, data privacy breaches, hallucinations, and lack of medical compliance [58-60]. These risks are closely associated with the ethics and compliance dimension of the evaluation system. Among the tertiary indicators in this study, several related to ethical and safety concerns received relatively high weight values, including "generation of factually incorrect information" (0.0889), "hallucination" (0.0697), "violation of

XSL•FO
RenderX

medical compliance" (0.0646), and "emergency response mechanisms" (0.0605). These results indicate that truthfulness, safety, and regulatory compliance of AI-generated content are critical components in assessing the quality of HCAI. In clinical settings, inaccurate recommendations or breaches of patient privacy can have serious consequences and must be addressed as a priority. This finding aligns with concerns raised by other researchers. For instance, some studies have reported that AI in health care may generate hallucinations—that is, inaccurate or fabricated content that could compromise clinical decision-making and patient safety [61]. Wang et al [62] also noted that LLMs may not achieve complete deidentification of training data, thereby raising the risk of exposing sensitive user information.

Within the health consultation capability dimension, the highest weight was assigned to "treatment recommendation generation ability" (0.2019). Among its tertiary indicators, "accuracy of treatment recommendations" (0.1216) and "comprehensiveness of treatment recommendations" (0.0562) were assigned relatively high weights. This indicates strong expert attention to whether AI systems can accurately assess users' health conditions and provide reliable and helpful advice. These findings are highly consistent with previous studies by Lukac et al [63] and Liu et al [64], both of whom emphasized the importance of AI's ability to generate dependable treatment recommendations. In addition, the tertiary indicator "accuracy in understanding user needs" (0.0648), under the category of "health information elicitation ability," and "accuracy of disease reasoning" (0.0389), under "health problem reasoning ability," also received relatively high weights. This suggests that, at the current stage, experts continue to view AI's capabilities in understanding user intent and providing accurate reasoning as important considerations.

In comparison, although the user experience dimension received a relatively lower overall weight, certain indicators, such as "perceived reliability" (0.0456) and "perceived safety" (0.0336), still accounted for a meaningful proportion. This reflects the significant impact of users' perceived trust in AI systems on actual usage experiences. Previous studies have shown that trust is one of the key determinants influencing users' acceptance and use of conversational AI tools such as ChatGPT [65]. In the health care context, user trust in AI technologies is also considered a critical factor affecting their broader adoption and application [66]. Moreover, some studies have found that, compared with traditional information channels such as online health websites, HCAI systems are better positioned to meet users' needs for immediacy and convenience during health information-seeking. This is largely due to their ability to engage in real-time interaction and provide personalized responses [67]. These findings further suggest that, in the ongoing optimization of HCAI, user experience should not only be regarded as a key factor influencing technology acceptance but also as a critical lever for enhancing system usability and fostering trust between humans and machines.

Policy Recommendations

In summary, the development of HCAI remains at an early stage. Ethics and compliance has emerged as the most critical evaluation dimension, while health consultation capability serves as the core functional foundation. Although user experience holds a relatively lower weight, it still plays a significant role in promoting the adoption and practical application of HCAI systems. To further advance the development of HCAI, this study proposes the following policy recommendations:

First, regulatory authorities around the world have imposed increasingly stringent requirements on HCAI-related products. For example, the European Union's Artificial Intelligence Act [68] mandates a comprehensive regulatory framework for high-risk AI systems, including those used in health care, covering both premarket and postmarket phases [68]. Singapore's Model AI Governance Framework for Generative AI emphasizes the importance of continuous evaluation and improvement across the entire life cycle of AI systems [69]. Similarly, China's Interim Measures for the Management of Generative Artificial Intelligence Services clearly stipulate dynamic supervision and safety monitoring of AI products after their deployment [70]. In light of this, it is recommended that the evaluation index system developed in this study be used as the basis for establishing a full–life cycle quality assessment mechanism for HCAI. This mechanism should encompass product development, testing, deployment, and application. Furthermore, efforts should be made to promote its transformation into a widely accepted industry framework through supportive policy initiatives. Such an approach would facilitate the regular evaluation of HCAI products and enable dynamic supervision and continuous quality improvement.

Second, the performance of HCAI systems should be continuously optimized across 3 key dimensions: model capability, ethics and compliance, and user experience. In terms of model capability, targeted training and fine-tuning should be conducted on tasks such as health information elicitation, disease diagnosis, and treatment recommendation to improve the system's medical professionalism and response accuracy. For ethical and safety considerations, it is necessary to establish a regulatory framework that covers data processing, model outputs, and information usage. This framework should clearly define accountability and responsibility in high-risk scenarios and include mechanisms to prevent misinformation and breaches of privacy. Regarding user experience, improvements should focus on core indicators such as "perceived safety" and "perceived reliability." This can be achieved by refining language logic, setting boundary warnings, incorporating source attribution, and integrating human-in-the-loop mechanisms. These measures are essential to enhance user understanding of, and trust in, the system.

Finally, HCAI is expected to move beyond the constraints of online consultation platforms and mobile apps. By helping users recognize their own health conditions, understand disease-related knowledge, and enhance early screening, diagnosis, and treatment, HCAI has the potential to improve health literacy and overall well-being. It may thus become a convenient and trustworthy health gatekeeper for the public. HCAI can also be integrated with health care service institutions and medical consortia to support the accuracy and feasibility of hierarchical diagnosis and treatment systems. To promote the effective application of HCAI in diverse scenarios, such as

XSL•FO
RenderX

self-assessment and diagnosis, intelligent triage, chronic disease follow-up, and care of older adults, it is necessary to adjust the structure and weight allocation of existing evaluation frameworks. Capability indicators should be refined based on the characteristics of specific tasks, with a particular focus on the model's adaptability and service effectiveness across different use cases. These efforts will help strengthen the professional performance of HCAI systems in multiple contexts and enhance the overall level of intelligent health care services.

Strengths and Limitations of Research

Unlike most previous expert consultation studies, this research assessed not only the importance of each indicator but also its sensitivity and feasibility, thereby enhancing the practicality of the constructed evaluation index system [71]. The findings of this study help address a critical gap in the evaluation of HCAI quality. The developed index system can be applied to assess the service quality of existing HCAI systems, identify weaknesses in consultation capability, user experience, ethics and compliance, and provide theoretical support for future product optimization and model iteration. This contributes to the standardized and high-quality development of HCAI. Furthermore, the evaluation system may serve as a reference for industry regulation. It offers a quantitative tool for government agencies, technology platforms, and health care providers to establish standards, identify risks, and improve service quality.

This study has several limitations. First, the expert panel consisted of 15-16 professionals from China. Although the panel was selected to ensure domain expertise, the limited sample size and geographic scope may affect the diversity and generalizability of perspectives. Future research should consider expanding the pool of experts to enhance the authority and generalizability of the conclusions. Second, although this study proposed a structured and multidimensional evaluation framework for HCAI, its practical utility has not yet been validated in real-world settings. While detailed assessment criteria for each third-level indicator are provided in Multimedia Appendix 2, further refinement of operational thresholds and performance benchmarks will be needed to support practical application. Additional empirical validation is required to assess the framework's applicability and adaptability across different scenarios.

Conclusions

This study developed a quality evaluation index system for HCAI, consisting of 3 primary indicators, 7 secondary indicators, and 28 tertiary indicators. Expert scoring results indicated that the dimension of "ethics and compliance" had the highest weight, followed by "health consultation capability." This suggests that, when evaluating the quality of health-related AI systems, priority should be given to information security, medical compliance, and risk management. The study combined the Delphi method with the AHP to ensure the scientific validity of the evaluation indicators and the rationality of their weight distribution. The proposed evaluation framework provides a theoretical reference for the assessment and optimization of HCAI systems.

Acknowledgments

The authors used the generative artificial intelligence tool ChatGPT (OpenAI) during the translation and manuscript refinement stages. This tool was used solely to assist with language polishing. All analytic decisions, study design choices, and interpretation of findings were made entirely by the authors. The authors critically reviewed and verified all outputs generated by the tool and fully accept responsibility for the final content of this manuscript.

Data Availability

The data that support the findings of this study are available from the authors; however, restrictions apply to their availability, as the data were used under license from the experts involved in the Delphi survey for this study and are therefore not publicly available. The data are, however, available from the authors upon reasonable request and with permission from these experts.

Funding

This study was funded by the National Social Science Foundation of China (grant 24BGL273).

Authors' Contributions

All authors contributed to the interpretation of the findings and the writing of the manuscript and have approved the final version. WL was responsible for the statistical analysis and editorial writing; ML and HL were responsible for data cleaning; CM and YH were responsible for the design, review, and revision of the article. CM and YH contributed equally to the research and manuscript preparation and are considered co-corresponding authors. DW, YW, ZF, HW, and YG were responsible for data acquisition and interpretation.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Expert consultation questionnaire in the first round.
[DOCX File , 154 KB - jmir_v28i1e83188_app1.docx]

Multimedia Appendix 2
Expert consultation in the second round and analytic hierarchy process questionnaire.
[DOCX File , 66 KB - jmir_v28i1e83188_app2.docx]

References

1. Ha JF, Longnecker N. Doctor-patient communication: a review. Ochsner J 2010;10(1):38-43. [Medline: 21603354]
2. World Health Organization (WHO). Ageing and health. WHO. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health [accessed 2025-03-17]
3. World Health Organization (WHO). Health workforce. WHO. URL: https://www.who.int/health-topics/health-workforce#tab=tab_1 [accessed 2025-03-17]
4. Lahat A, Klang E. Can advanced technologies help address the global increase in demand for specialized medical care and improve telehealth services? J Telemed Telecare 2024 Oct;30(9):1516-1517. [doi: 10.1177/1357633X231155520] [Medline: 36760131]
5. He Z, Bhasuran B, Jin Q, Tian S, Hanna K, Shavor C, et al. Quality of answers of generative large language models versus peer users for interpreting laboratory test results for lay patients: evaluation study. J Med Internet Res 2024 Apr 17;26:e56655 [FREE Full text] [doi: 10.2196/56655] [Medline: 38630520]
6. Hirosawa T, Harada Y, Mizuta K, Sakamoto T, Tokumasu K, Shimizu T. Diagnostic performance of generative artificial intelligences for a series of complex case reports. Digit Health 2024;10:20552076241265215. [doi: 10.1177/20552076241265215] [Medline: 39229463]
7. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
8. Busch F, Han T, Makowski MR, Truhn D, Bressem KK, Adams L. Integrating text and image analysis: exploring GPT-4V's capabilities in advanced radiological applications across subspecialties. J Med Internet Res 2024 May 01;26:e54948 [FREE Full text] [doi: 10.2196/54948] [Medline: 38691404]
9. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239 [FREE Full text] [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
10. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature 2025 Jun;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
11. Zhu J, Pan J, Liu Y, Liu F, Wu J. Ask patients with patience: enabling LLMs for human-centric medical dialogue with grounded reasoning. arXiv. Preprint posted online on August 21, 2025 2025:e1 [FREE Full text] [doi: 10.48550/arXiv.2502.07143]
12. Announcement by the Cyberspace Administration of China on the Release of the Seventh Batch of Deep Synthesis Service Algorithm Filing Information. State Internet Information Office of the People's Republic of China. URL: https://www.cac.gov.cn/2024-08/05/c_1724541639039621.htm [accessed 2025-03-24]
13. Conversational AI in healthcare market by component (chatbots, virtual assistants, speech recognition systems, and services), by technology, by application, by end user - global industry outlook, key companies (IBM, Microsoft, Google, and others), trends and forecast 2025-2034. Dimension Market Research. 2025 Jul 5. URL: https://dimensionmarketresearch.com/report/conversational-ai-in-healthcare-market/ [accessed 2025-06-12]
14. Medical large model research report: empowering nearly 300 medical large models in clinical and non-clinical scenarios. Beijing Review. 2025. URL: http://www.beijingreview.com.cn/caijing/hyzx/202505/t20250513_800401417.html [accessed 2025-06-12]
15. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023 Oct 01;13(1):16492 [FREE Full text] [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]
16. Thirunavukarasu AJ, Hassan R, Mahmood S, Sanghera R, Barzangi K, El Mukashfi M, et al. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. JMIR Med Educ 2023 Apr 21;9:e46599 [FREE Full text] [doi: 10.2196/46599] [Medline: 37083633]
17. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ 2024 Feb 14;24(1):143 [FREE Full text] [doi: 10.1186/s12909-024-05125-7] [Medline: 38355517]
18. Liu C, Sun K, Zhou Q, Duan Y, Shu J, Kan H, et al. CPMI-ChatGLM: parameter-efficient fine-tuning ChatGLM with Chinese patent medicine instructions. Sci Rep 2024 Mar 16;14(1):6403 [FREE Full text] [doi: 10.1038/s41598-024-56874-w] [Medline: 38493251]

XSL•FO
RenderX

19. Zhang K, Zhou R, Adhikarla E, Yan Z, Liu Y, Yu J, et al. A generalist vision-language foundation model for diverse biomedical tasks. Nat Med 2024 Nov 07;30(11):3129-3141. [doi: 10.1038/s41591-024-03185-2] [Medline: 39112796]
20. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform 2022 Nov 19;23(6):bbac409. [doi: 10.1093/bib/bbac409] [Medline: 36156661]
21. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
22. Van Veen D, Van Uden C, Blankemeier L, Delbrouck J, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024 Apr;30(4):1134-1142. [doi: 10.1038/s41591-024-02855-5] [Medline: 38413730]
23. Sun H, Zhang K, Lan W, Gu Q, Jiang G, Yang X, et al. An AI dietitian for type 2 diabetes mellitus management based on large language and image recognition models: preclinical concept validation study. J Med Internet Res 2023 Nov 09;25:e51300 [FREE Full text] [doi: 10.2196/51300] [Medline: 37943581]
24. Ho CN, Tian T, Ayers AT, Aaron RE, Phillips V, Wolf RM, et al. Qualitative metrics from the biomedical literature for evaluating large language models in clinical decision-making: a narrative review. BMC Med Inform Decis Mak 2024 Nov 26;24(1):357. [doi: 10.1186/s12911-024-02757-z]
25. Chen X, Xiang J, Lu S, Liu Y, He M, Shi D. Evaluating large language models and agents in health care: key challenges in clinical applications. Intelligent Medicine 2025 May 1:151-163 [FREE Full text] [doi: 10.1016/j.imed.2025.03.002]
26. Omar M, Soffer S, Agbareia R, Bragazzi NL, Apakama DU, Horowitz CR, et al. Sociodemographic biases in medical decision making by large language models. Nat Med 2025 Jun 1:1873-1881. [doi: 10.1038/s41591-025-03626-6]
27. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst 2025 Jan 24;43(2):1-55. [doi: 10.1145/3703155]
28. Xu J, Lu L, Peng X, Pang J, Ding J, Yang L, et al. Data set and benchmark (MedGPTEval) to evaluate responses from large language models in medicine: evaluation development and validation. JMIR Med Inform 2024 Jun 28;12:e57674 [FREE Full text] [doi: 10.2196/57674] [Medline: 38952020]
29. Liu L, Yang X, Li F, Chi C, Shen Y, Lyu S. Towards automatic evaluation for LLMs' clinical capabilities: metric, data, and algorithm. 2024 Aug 24 Presented at: KDD '24: The 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; August 25-29, 2024; New York, NY. [doi: 10.1145/3637528.3671575]
30. Johri S, Jeong J, Tran BA, Schlessinger DI, Wongvibulsin S, Barnes LA, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. Nat Med 2025 Jan;31(1):77-86. [doi: 10.1038/s41591-024-03328-5] [Medline: 39747685]
31. Liu X, Liu H, Yang G, Jiang Z, Cui S, Zhang Z, et al. A generalist medical language model for disease diagnosis assistance. Nat Med 2025 Mar 31;31(3):932-942. [doi: 10.1038/s41591-024-03416-6]
32. Arora R, Wei J, Hicks R, Bowman P, Candela J, Tsimpourlas F. HealthBench: evaluating large language models towards improved human health. arXiv. Preprint posted online on May 13, 2025 2025 May 13:e1 [FREE Full text] [doi: 10.48550/arXiv.2505.08775]
33. Qiu P, Wu C, Liu S, Fan Y, Zhao W, Chen Z, et al. Quantifying the reasoning abilities of LLMs on clinical cases. Nat Commun 2025 Nov 06;16(1):9799. [doi: 10.1038/s41467-025-64769-1] [Medline: 41198657]
34. Abbasian M, Khatibi E, Azimi I, Oniani D, Shakeri Hossein Abad Z, Thieme A, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. NPJ Digit Med 2024 Mar 29;7(1):82 [FREE Full text] [doi: 10.1038/s41746-024-01074-z] [Medline: 38553625]
35. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance.). EUR-Lex. 2017 Apr 5. URL: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng [accessed 2025-11-28]
36. Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. J Adv Nurs 2000 Oct;32(4):1008-1015. [Medline: 11095242]
37. Chen H, Zhang Y, Wang L. A study on the quality evaluation index system of smart home care for older adults in the community based on Delphi and AHP. BMC Public Health 2023 Mar 01;23(1):411 [FREE Full text] [doi: 10.1186/s12889-023-15262-1] [Medline: 36859259]
38. Shi Y, Sun S, Deng J, Liu S, Yin T, Peng Q, et al. Establishment and application of an index system for the risk of drug shortages in China: based on Delphi method and analytic hierarchy process. Int J Health Policy Manag 2022 Dec 19:2860-2868. [doi: 10.34172/ijhpm.2022.6360]
39. Bommasani R, Liang P, Lee T. Holistic evaluation of language models. Ann N Y Acad Sci 2023 Jul;1525(1):140-146. [doi: 10.1111/nyas.15007] [Medline: 37230490]
40. Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A survey on evaluation of large language models. ACM Trans Intell Syst Technol 2024 Mar 29;15(3):1-45. [doi: 10.1145/3641289]
41. Sallam M, Al-Mahzoum K, Alshuaib O, Alhajri H, Alotaibi F, Alkhurainej D, et al. Language discrepancies in the performance of generative artificial intelligence models: an examination of infectious disease queries in English and Arabic. BMC Infect Dis 2024 Aug 08;24(1):799 [FREE Full text] [doi: 10.1186/s12879-024-09725-y] [Medline: 39118057]

XSL•FO
RenderX

42. Meyer A, Soleman A, Riese J, Streichert T. Comparison of ChatGPT, Gemini, and Le Chat with physician interpretations of medical laboratory questions from an online health forum. Clin Chem Lab Med 2024 Nov 26;62(12):2425-2434 [FREE Full text] [doi: 10.1515/cclm-2024-0246] [Medline: 38804035]
43. The "Maturity assessment model for large models in the medical and health industry, part 1: health consultation" has been officially released. Smart Healthcare Innovation Competition. 2024. URL: https://tinyurl.com/c95zexwj [accessed 2025-06-15]
44. Huang C, Lee P, Chen L. Exploring consumers' negative electronic word-of-mouth of 5 military hospitals in Taiwan through SERVQUAL and Flower of Services: web scraping analysis. JMIR Form Res 2024 May 29;8:e54334 [FREE Full text] [doi: 10.2196/54334] [Medline: 38809602]
45. Bashir A, Bastola DR. Perspectives of nurses toward telehealth efficacy and quality of health care: pilot study. JMIR Med Inform 2018 May 25;6(2):e35 [FREE Full text] [doi: 10.2196/medinform.9080] [Medline: 29802089]
46. Smith AG, Brainard JC, Campbell KA. Development of an undergraduate medical education critical care content outline utilizing the Delphi method. Crit Care Med 2020 Jan;48(1):98-103. [doi: 10.1097/CCM.0000000000004086] [Medline: 31714399]
47. Nie S, Wang L. Constructing an evaluation index system for clinical nursing practice teaching quality using a Delphi method and analytic hierarchy process-based approach. BMC Med Educ 2024 Jul 19;24(1):772. [doi: 10.1186/s12909-024-05770-y] [Medline: 39030603]
48. Liu S, Li Y, Fu S, Liu X, Liu T, Fan H, et al. Establishing a multidisciplinary framework for an emergency food supply system using a modified Delphi approach. Foods 2022 Apr 06;11(7):1054. [doi: 10.3390/foods11071054] [Medline: 35407141]
49. Min Z, Bin H, Wenjie Z, Tao L, Yi M, Chunhua Z, et al. Developing an assessment tool for the healthy lifestyles of the occupational population in China: a modified Delphi-analytic hierarchy process study. Sci Rep 2024 Sep 02;14(1):20359. [doi: 10.1038/s41598-024-71324-3]
50. Liu W, Hu M, Chen W. Identifying the service capability of long-term care facilities in China: an e-Delphi study. Front Public Health 2022 Jun 29;10:884514 [FREE Full text] [doi: 10.3389/fpubh.2022.884514] [Medline: 35844860]
51. Cai W, Yao Y, Lei W, Li H, Yan S, Wu Q, et al. Construction on training course and training quality evaluation index system of chronic disease medication therapy management service (MTMs) in China: a Delphi study. PLoS One 2025;20(1):e0318446 [FREE Full text] [doi: 10.1371/journal.pone.0318446] [Medline: 39883712]
52. Chang H, Lo C, Chang H. Development of the benefit-risk assessment of complementary and alternative medicine use in people with diabetes: a Delphi-analytic hierarchy process approach. Comput Inform Nurs 2021 Apr 16;120(3):384-391. [doi: 10.1097/CIN.0000000000000749] [Medline: 33871384]
53. Ruan Y, Song S, Yin Z, Wang M, Huang N, Gu W, et al. Comprehensive evaluation of military training-induced fatigue among soldiers in China: a Delphi consensus study. Front Public Health 2022;10:1004910 [FREE Full text] [doi: 10.3389/fpubh.2022.1004910] [Medline: 36523578]
54. Gwee X, Nyunt MSZ, Kua EH, Jeste DV, Kumar R, Ng TP. Reliability and validity of a self-rated analogue scale for global measure of successful aging. Am J Geriatr Psychiatry 2014 Aug;22(8):829-837 [FREE Full text] [doi: 10.1016/j.jagp.2013.09.002] [Medline: 24119862]
55. Hsu C, Sandford B. The Delphi technique: making sense of consensus. Practical Assessment Research & Evaluation 2007;12(1):10 [FREE Full text] [doi: 10.7275/pdz9-th90]
56. Hohmann E, Brand JC, Rossi MJ, Lubowitz JH. Expert opinion is necessary: Delphi panel methodology facilitates a scientific approach to consensus. Arthroscopy 2018 Feb;34(2):349-351 [FREE Full text] [doi: 10.1016/j.arthro.2017.11.022] [Medline: 29413182]
57. Zhou W, Zheng Q, Huang M, Wang J, Gan X. Development and validation of nurse's assessment ability questionnaire in delirium subtypes: based on Delphi expert consensus. PLoS One 2024;19(1):e0297063 [FREE Full text] [doi: 10.1371/journal.pone.0297063] [Medline: 38261557]
58. Bouderhem R. Shaping the future of AI in healthcare through ethics and governance. Humanit Soc Sci Commun 2024 Mar 15;11(1):416. [doi: 10.1057/s41599-024-02894-w]
59. Menz BD, Kuderer NM, Bacchi S, Modi ND, Chin-Yee B, Hu T, et al. Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis. BMJ 2024 Mar 20:e078538 [FREE Full text] [doi: 10.1136/bmj-2023-078538] [Medline: 38508682]
60. Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: a narrative review. Heliyon 2024 Feb 29;10(4):e26297 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e26297] [Medline: 38384518]
61. Suárez A, Jiménez J, Llorente de Pedro M, Andreu-Vázquez C, Díaz-Flores García V, Gómez Sánchez M, et al. Beyond the scalpel: assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery. Computational and Structural Biotechnology Journal 2024 Dec;24:46-52. [doi: 10.1016/j.csbj.2023.11.058]
62. Wang L, Wan Z, Ni C, Song Q, Li Y, Clayton E, et al. Applications and concerns of ChatGPT and other conversational large language models in health care: systematic review. J Med Internet Res 2024 Nov 07;26:e22769 [FREE Full text] [doi: 10.2196/22769] [Medline: 39509695]

XSL•FO
RenderX

63. Lukac S, Dayan D, Fink V, Leinert E, Hartkopf A, Veselinovic K, et al. Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. Arch Gynecol Obstet 2023 Dec;308(6):1831-1844 [FREE Full text] [doi: 10.1007/s00404-023-07130-5] [Medline: 37458761]
64. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 2023 Jun 20;30(7):1237-1245 [FREE Full text] [doi: 10.1093/jamia/ocad072] [Medline: 37087108]
65. Choudhury A, Shamszare H. Investigating the impact of user trust on the adoption and use of ChatGPT: survey analysis. J Med Internet Res 2023 Jun 14;25:e47184 [FREE Full text] [doi: 10.2196/47184] [Medline: 37314848]
66. Nong P, Platt J. Patients' trust in health systems to use artificial intelligence. JAMA Netw Open 2025 Feb 14;8(2):e2460628. [doi: 10.1001/jamanetworkopen.2024.60628]
67. Esmaeilzadeh P, Maddah M, Mirzaei T. Using AI chatbots (e.g., CHATGPT) in seeking health-related information online: the case of a common ailment. Computers in Human Behavior: Artificial Humans 2025 Mar;3:100127 [FREE Full text] [doi: 10.1016/j.chbah.2025.100127]
68. Artificial Intelligence Act. Future of Life Institute. 2025. URL: https://artificialintelligenceact.eu/ai-act-explorer/ [accessed 2025-11-27]
69. Singapore launches Model AI Governance Framework (Gen AI) and AI Governance Playbook for Digital Forum of Small States (Digital FOSS). Infocomm Media Development Authority. 2024. URL: https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/factsheets/2024/gen-ai-and-digital-foss-ai-governance-playbook [accessed 2025-11-27]
70. Interim measures for the management of generative artificial intelligence services. The Central People's Government of the People's Republic of China. 2023. URL: https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm [accessed 2025-11-27]
71. Sun H, Wang Y, Cai H, Wang P, Jiang J, Shi C, et al. The development of a performance evaluation index system for Chinese Centers for Disease Control and Prevention: a Delphi consensus study. Glob Health Res Policy 2024 Jul 23;9(1):28 [FREE Full text] [doi: 10.1186/s41256-024-00367-w] [Medline: 39044214]

Abbreviations

AHP: analytic hierarchy process
AI: artificial intelligence
AMIE: Articulate Medical Intelligence Explorer
BLEU: Bilingual Evaluation Understudy
Ca: the basis for expert's judgment
CPMI: Chinese Patent Medicine Instructions
Cr: expert authority coefficient
CRAFT-MD: the Conversational Reasoning Assessment Framework for Testing in Medicine
Cs: expert familiarity
CV: coefficient of variation
HCAI: health conversational artificial intelligence
LLaMA: Large Language Model from Meta
LLM: large language model
Medical-Diff-VQA: Medical Dataset for Difference Visual Question Answering on Chest X-Ray Images

Edited by A Schwartz, M Balcarras; submitted 29.Aug.2025; peer-reviewed by J Haverinen, R Yin; comments to author 14.Nov.2025; revised version received 02.Dec.2025; accepted 17.Dec.2025; published 19.Jan.2026.

Please cite as:
Liao W, Li M, Ma C, Han Y, Wang D, Liu H, Wang Y, Feng Z, Wang H, Guan Y
Developing a Quality Evaluation Index System for Health Conversational Artificial Intelligence: Mixed Methods Study
J Med Internet Res 2026;28:e83188
URL: https://www.jmir.org/2026/1/e83188
doi:10.2196/83188
PMID:

©Weizhen Liao, Meng Li, Chengyu Ma, Youli Han, Dan Wang, Haopeng Liu, Yi Wang, Zijie Feng, Huichao Wang, Yiru Guan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the

XSL•FO
RenderX

Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Developing a Service Quality Index System for AI Health Care Chatbots: Mixed Methods Study

Yu Gu[1], PhD; Xinyi Wang[1], BA

School of Medical Technology, Capital Medical University, Beijing, China

Corresponding Author:
Yu Gu, PhD
School of Medical Technology
Capital Medical University
No 10 Xi Toutiao Road
Beijing, 100069
China
Phone: 86 81476234
Email: bitguyu@126.com

Abstract

Background: Artificial intelligence (AI) health care chatbots are gaining widespread adoption worldwide. It is imperative to understand the service quality of AI health care chatbots. However, there is limited guidance on how to comprehensively evaluate their service quality.

Objective: This study aimed to develop an index system based on the SERVQUAL framework for evaluating the service quality of AI health care chatbots.

Methods: An initial indicator pool was compiled through a comprehensive literature review and consultations with 4 experts. These indicators were mapped and categorized into 5 domains adapted from the SERVQUAL framework. The experts were recruited from hospital, university, and health commission settings by purposive sampling. The service quality index system was identified using a 2-round Delphi process, which included a virtual meeting between the 2 rounds. In the third round, indicator weights within each quality domain and subdomain were determined using the analytic hierarchy process.

Results: There were 26 indicators identified in the literature, based on which the 2-round Delphi process was conducted. A total of 20 experts were invited. The response rates in both rounds of Delphi and the analytic hierarchy process were 100%, and the authoritative coefficients were both >0.7. The final service quality index system for AI health care chatbots comprises 5 primary indicators and 17 secondary indicators. There were 3 (18%) indicators on assurance, 4 (24%) on reliability, 3 (18%) on human-likeness, 4 (24%) on tangibility, and 3 (18%) on responsiveness. The primary indicators, ranked from highest to lowest weight, were assurance (0.239), reliability (0.237), human-likeness (0.187), tangibility (0.170), and responsiveness (0.167).

Conclusions: This study pioneers the development of a service quality index system for AI health care chatbots adapted from the SERVQUAL framework. The results provide a validated tool for evaluating the performance of chatbots and offer valuable insights for health service managers and developers to enhance AI-driven medical consultation services.

(J Med Internet Res 2026;28:e83051) doi:10.2196/83051

KEYWORDS

artificial intelligence health care chatbot; artificial intelligence; AI; service quality; Delphi method; analytic hierarchy process; index system development; SERVQUAL

Introduction

Worldwide, artificial intelligence (AI) chatbots have been introduced into health care settings in recent years, where they are used by individuals as AI physicians for online medical consultations. A key innovation of AI health care chatbots lies in their ability to generate humanlike, natural language responses to diverse health-related queries anytime and anywhere, significantly improving access to medical guidance for broader populations [1]. Unlike earlier rule-based chatbots that relied on scripted replies, AI chatbots leverage advanced technologies, such as large language models (LLMs), to deliver personalized and context-aware interactions [2]. Moreover, the consultation service is often provided free of charge. AI health care chatbots show promise in delivering reliable medical advice without

XSL•FO
RenderX

direct involvement from human physicians, offering a scalable solution to persistent challenges within the global health system, such as limited resources, uneven distribution, high costs, and growing demand [3]. Therefore, AI health care chatbots are playing an increasingly important role in modern health care systems [4].

AI health care chatbots represent not only a new type of service provider but also an innovative medical service model [5]. As an emerging field, chatbots have attracted growing attention from both practitioners and researchers. Despite its potential benefits, concerns remain regarding its service quality [6,7]. Efforts have been made to develop quality indicators for AI health care chatbots [8-16]. Some studies have evaluated response quality within specific disease contexts, such as labor epidurals, cardiovascular health, oncology, psoriasis, chronic hepatitis, and cancer [8-11]. Others have focused on assessing information quality [12,13] or have compared the performance of AI health care chatbots with that of human physicians [14-16]. However, existing studies primarily focus on narrow aspects of quality. Furthermore, the most commonly applied metrics—response accuracy, completeness, and consistency in closed-ended clinical questions—are predominantly defined from the health care providers' perspective rather than that of users. Therefore, a comprehensive and user-centered index system for evaluating service quality of AI health care chatbots remains underdeveloped.

Among existing service quality frameworks, SERVQUAL, developed by Parasuraman et al [17], is one of the most widely recognized frameworks for evaluating medical service quality worldwide. This framework includes 5 dimensions—tangibility, reliability, responsiveness, assurance, and empathy—and is specifically designed to assess users' expectations and perceptions of service quality [18]. Applying this classical framework enables a more comprehensive and theoretically grounded evaluation of service quality of AI health care chatbots, bridging classical service quality theory with emerging AI-driven health care contexts.

The aim of this study was to identify critical indicators that reflect the service quality of AI health care chatbots and to develop a scientifically feasible index system for its evaluation. The findings were expected to contribute to better identification of shortcomings, promote continuous quality improvement, enhance user experience, and offer new insights into the systemic evaluation of service quality of AI health care chatbots.

Methods

Study Design

This study used a mixed methods approach, combining qualitative insights from expert opinions with quantitative metrics to develop and quantify a service quality index system for AI health care chatbots. The literature review and expert consultation were applied to construct an initial indicator pool. The 2-round Delphi consultation was then conducted to refine and establish the final index system. Subsequently, the analytic hierarchy process (AHP) was applied to determine the weight of each indicator. The process of index system development and weight determination is shown in Figure 1.

Figure 1. The research process. AHP: analytic hierarchy process.

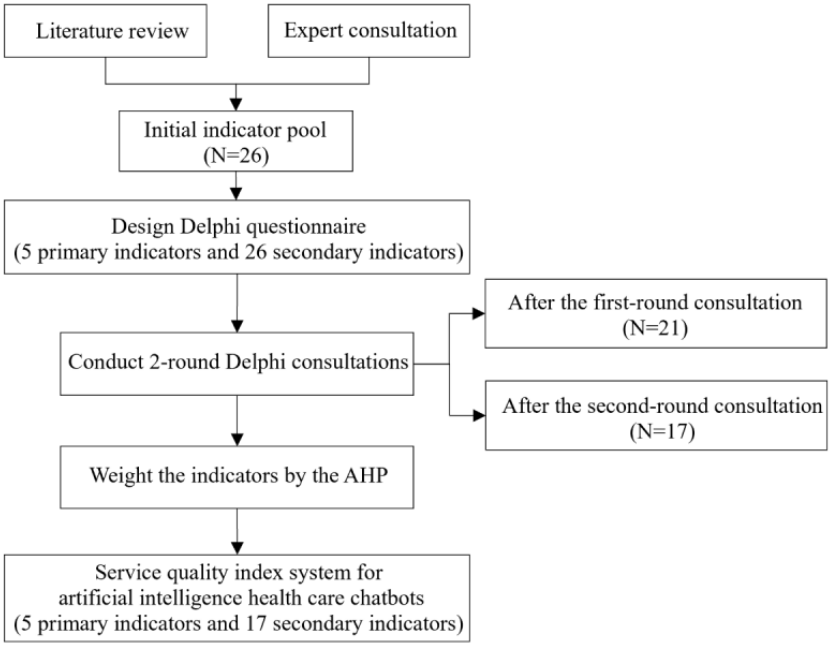

Initial Indicator Pool

The initial indicator pool was compiled based on existing literature and expert opinions. A comprehensive systematic literature search was conducted in 4 databases: PubMed, Web of Science, China National Knowledge Infrastructure, and Wanfang Data. The search strategy incorporated the following key terms: ["chatbot*" OR "chat-bot*" OR "conversational agent*" OR "conversational bot*" OR "conversational system*" OR "dialogue system*" OR ChatGPT] AND ["medic*" OR "health*" OR "disease*" OR "patient*"] AND ["quality indicator*" OR "quality evaluat*" OR "quality assess*" OR "quality measure*"]. Boolean operators (AND, OR, and NOT) were used to combine or refine search terms (Multimedia

XSL•FO
RenderX

Appendix 1). The exclusion criteria were as follows: (1) studies not focused on the evaluation of AI health care chatbots; (2) commentaries, protocols, letters, editorials, and conference abstracts; and (3) studies not published in English or Chinese.

In the second phase, this study incorporated insights from 4 interdisciplinary experts specializing in intelligent health care and medical service management. These experts conducted in-depth discussions regarding the relevance, suitability, and validity of the preliminary indicators. None of them participated in the subsequent Delphi consultation rounds. Both prior literature and expert opinions emphasized that the SERVQUAL framework provides a user-centered foundation well suited for the evaluation of AI health care chatbots and highlighted anthropomorphism as a distinctive feature influencing perceived service quality in AI-driven interactions. Specifically, users desire AI health care chatbots to exhibit kindness through humanlike attributes, such as a name, image, and voice. Beyond a friendly appearance, users expect these systems to demonstrate social intelligence, including the ability to detect user emotions and respond with genuine concern, which goes beyond simple empathy. Furthermore, users expect personalized responses tailored to individual factors. Therefore, the *humanlike* dimension was introduced as an innovative replacement for the traditional "empathy" construct to better capture the emotional and interactive capabilities unique to AI health care chatbots. Accordingly, this study established 5 first-level indicators adapted from the SERVQUAL structure: tangibility, responsiveness, assurance, human-likeness, and reliability, and 26 second-level indicators were included in the initial pool.

Expert Selection

The initial expert recruitment was conducted through recommendations from our collaborators in the field of intelligent health care. They were from renowned universities, tertiary hospitals, and provincial health sectors. To broaden the reach and ensure a diverse range of perspectives, the experts initially nominated by the researchers were then asked to suggest other qualified individuals who could contribute valuable insights to the study. The inclusion criteria for experts were as follows: (1) familiarity with research areas, such as intelligent health care, medical service management, health information management, and other related fields; (2) more than 5 years of professional experience in a relevant field; and (3) willingness to actively participate in the study and provide timely responses across multiple rounds of Delphi consultation. Finally, a total of 20 experts were recruited through purposive sampling.

Delphi Process

The Delphi method is a structured communication technique designed to systematically collect expert opinions and achieve consensus [19]. It has been widely applied and validated as a robust research methodology in health care contexts [20]. This study conducted a 2-round Delphi consultation to screen, refine, and finalize the indicators.

The Delphi consultation questionnaire consists of 2 main sections: an informed consent form and the main survey. The informed consent form outlined the study's background, objectives, methodology, privacy protection measures, and contact information. The main survey collected information from five areas: (1) experts' basic information, including age, education, and years of work experience; (2) the core consultation content, in which experts scored the importance and feasibility of each indicator using a 10-point scale (1=lowest and 10=highest); (3) the familiarity scale, rated by the expert themselves using a 5-point Likert scale (1=very unfamiliar and 5=very familiar); (4) the basis of expert judgment, evaluating the impact of theoretical analysis, practical experience, literature knowledge, and instinct on scoring (rated as high, medium, or low); and (5) blank fields, allowing experts to propose additions, deletions, or modifications to the indicators. All experts completed the informed consent process, and strict confidentiality was maintained throughout the entire process.

The Delphi process was conducted between February 2025 and June 2025. In the first round, Delphi questionnaires in Microsoft Word format were distributed to 20 experts, with a 2-week response period. Experts were asked to rate both the primary and secondary indicators and to provide comments. On the basis of the results and comments from the first round, the questionnaire was revised and redistributed to the same 20 experts for the second round. The second round followed the same rating procedure as the first and achieved consensus among the experts.

Indicator Selection

To screen the indicators, this study used 3 important statistics: the mean importance score, the full-mark rate (proportion of experts assigning the highest score), and the coefficient of variation. The inclusion criteria were as follows: (1) a mean of importance score ≥7.0, (2) a full-mark rate >20%, and (3) a coefficient of variation <0.25 [21-23]. Any indicator failing to meet all 3 criteria was subject to deletion or revision based on panel discussion and qualitative feedback.

AHP Procedure

Following the 2-round Delphi consultation, the final set of indicators was confirmed. The same panel of experts was then invited to participate in a pairwise comparison process to determine indicator weights. For each pair of indicators within the same hierarchical level, judgment matrices were constructed using a 1 to 9 ordinal scale to assess their relative importance [24]. The weight of each indicator was subsequently calculated using the percentage weighting method based on the pairwise comparison matrices, with higher weight values indicating greater perceived importance.

Data Analysis

Statistical analysis was conducted using SPSS software (version 25.0; IBM Corp). The authority coefficient (Cr) represents the authority level of experts. Cr was the arithmetic mean of the experts' judgment coefficient (Ca) and the experts' familiarity coefficient (Cs) [25]. A Cr value ≥0.7 was considered acceptable [26]. The Ca value was derived from experts' self-assessment of their own judgment criteria, as detailed in Table 1. The Cs value ranges from 1.0 (very familiar) to 0.2 (unfamiliar). The coordination of expert opinions was tested using the Kendall coefficient of concordance (Kendall W), with a significance level of α=.05. YAAHP software (version 11.2; MetaDecision)

XSL•FO
RenderX

was used to calculate the indicator weights and assess the consistency ratio. When the consistency ratio value was <0.10, it was considered acceptable, indicating sufficient consistency in expert judgments [27].

Table 1. The judgment basis and degree of influence.

Judgment basis	Degree of impact on experts' judgment		
	High	Medium	Low
Theoretical analysis	0.3	0.2	0.1
Practical experience	0.5	0.4	0.3
Reference literature	0.1	0.1	0.1
Expert intuition	0.1	0.1	0.1

Ethical Considerations

The study protocol was approved by the Ethics Committee of the Capital Medical University, Beijing, China (2025SY-071). Participants were informed of the study's purpose and procedure. Online informed consent was obtained from each participant. All research data were stored on a password-encrypted computer, and only the researchers had access to the data. No compensation was provided to participants.

Results

Characteristics of Experts

A total of 20 experts completed the 2-round Delphi consultation and AHP evaluation. The panel consisted of 12 (60%) male and 8 (40%) female experts, ranging in age from 31 to 60 years. Among the experts, 17 (85%) held a master's degree or higher. All experts possessed associate senior professional titles or higher. The panel included 10 (50%) experts from quality control departments of hospitals, 7 (35%) from universities, and 3 (15%) from national or regional health commissions. The detailed characteristics of these experts are summarized in Table 2.

XSL•FO
RenderX

Table 2. The characteristics of the Delphi consultation experts (N=20).

Characteristics	Experts, n (%)
Sex	
Male	12 (60)
Female	8 (40)
Age (years)	
31-40	6 (30)
41-50	11 (55)
51-60	3 (15)
Education	
Doctoral degree	10 (60)
Master's degree	7 (35)
Bachelor's degree	3 (15)
Professional title	
Senior	11 (55)
Associate senior	9 (45)
Seniority (years)	
6-10	4 (20)
11-20	9 (45)
21-30	5 (25)
>30	2 (10)
Affiliation	
Hospitals	10 (50)
Universities	7 (35)
Health commission	3 (15)
Field of expertise	
Intelligent health care	8 (40)
Medical service management	7 (35)
Health information management	5 (25)

Authority Coefficient and Degree of Coordination

The Cr values for the first and second rounds of the Delphi consultation were 0.894 (Ca=0.931; Cs=0.847) and 0.919 (Ca=0.917; Cs=0.921), respectively (Table 3). Both values exceed the accepted threshold of 0.7, indicating a high level of expert credibility and reinforcing the reliability of the consultation results.

The Kendall *W* coefficients for the 2 consultation rounds are shown in Table 3. After the second round, the coordination coefficients of the indicator increased from 0.263 to 0.339, and all associated *P* values were <.001, indicating that the experts' opinions converged and that the degree of consensus among experts was acceptable.

Table 3. Expert authority coefficients and the degree of coordination of expert opinions.

Round	Ca[a]	Cs[b]	Cr[c]	Kendall *W*	Chi-square (*df*)	*P* value
Round 1	0.931	0.847	0.894	0.263	71.1 (25)	<.001
Round 2	0.917	0.921	0.919	0.339	123.2 (20)	<.001

[a]Cs: familiarity coefficient.

[b]Ca: judgment coefficient.

[c]Cr: authority coefficient.

XSL•FO
RenderX

Review for Initial Indicator Pool

The database search and hand searches identified 117 articles, from which 48 (41.0%) duplicates were removed. After screening the titles and abstracts, 29 (24.8%) full-text records were reviewed, of which 20 (17.1%) were included in the review [28-47]. The study selection process is illustrated in Figure 2.

Figure 2. Flowchart of the included studies.

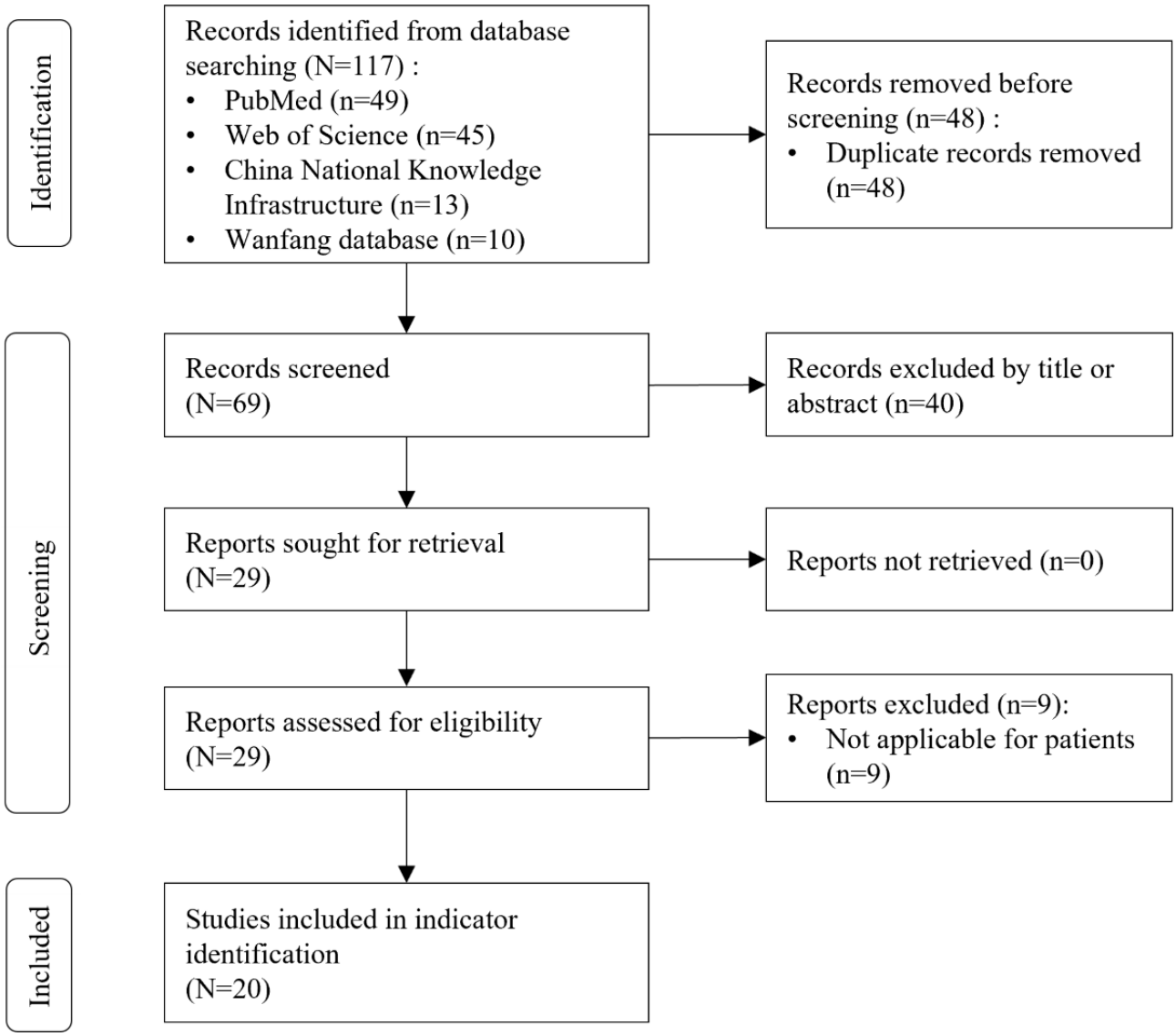

In the 20 eligible studies, 9 (45%) focused exclusively on AI chatbots designed specifically for health care service [28-36]. To develop a comprehensive initial indicator pool, we also included 11 additional studies concerning the quality of general AI chatbots capable of delivering health care–related consultations [37-47]. Of these, 6 (30%) studies aimed to develop instruments for measuring the overall service quality of AI chatbots [28,29,34,37,38,40], with 2 (10%) grounded in the SERVQUAL model [34,37]. Another 7 (35%) studies developed instruments targeting specific dimensions of quality [30-33,35,36,42]. Additionally, 7 (35%) studies treated quality as a key determinant of acceptance or satisfaction with AI chatbots and provided detailed measurement items for AI chatbot quality [39,41,43-47]. Following a systematic sorting process and discussions with 4 domain experts, we synthesized these findings into an initial pool of 26 second-level indicators for assessing the quality of AI health care chatbots, structured according to the 5 dimensions of the SERVQUAL framework.

Indicator Selection

According to the indicator selection criteria and qualitative feedback from the experts, from a total of 26 indicators, 5 (19%) secondary indicators were removed and 21 (81%) were retained in the first round. Between round 1 and round 2, a virtual Delphi meeting was held to discuss indicators for which experts provided revision suggestions and to identify novel indicators based on perceived gaps in current indicators. Experts who had completed round 1 attended the meeting. During this panel meeting, 8 (31%) secondary indicators were merged into 4 (15%) indicators. Following the discussion, 20 experts rated the new indicators. Finally, the 2-round Delphi process reached a finalized evaluation framework comprising 5 (19%) primary indicators and 17 (65%) secondary indicators, each clearly defined in Table 4.

XSL•FO
RenderX

Table 4. The service quality index system for artificial intelligence (AI) health care chatbots.

Indicator	Definition	Weights
Assurance	The ability of AI health care chatbots to provide pertinent responses	0.239
Understandable answer	The AI health care chatbots provide an answer with a logistical structure and the right amount of information to the user's query	0.082
Accurate understanding	The AI health care chatbots understand the exact meaning of the content sent by the user in text and voice	0.079
Targeted question	The AI health care chatbots ask follow-up questions with context awareness based on the user's query	0.078
Reliability	The ability of AI health care chatbots to inspire trust and confidence	0.237
Trustworthy advice	The AI health care chatbots give medical advice, such as diagnosis, medication, and examination, and detailed explanations, which are consistent across multiple inquiries	0.071
Useful service	The AI health care chatbots are useful in addressing users' uncertainties about their health concerns	0.065
Specific risk warning	The AI health care chatbots clearly indicate the limitations of its provided answers	0.050
Protected privacy	The AI health care chatbots protect the user's privacy	0.051
Human-likeness	The social cue, personality, and empathy of AI health care chatbots	0.187
Personalized response	The AI health care chatbots tailor their answers according to the user's age, sex, and medical history	0.064
Emotional attention	The AI health care chatbots detect user emotion and makes the user feel concerned	0.063
Kind characteristic	The AI health care chatbots have a kind name, image, and voice	0.060
Tangibility	The hardware or software manifestations of AI health care chatbots	0.170
Accurate recognition	The AI health care chatbots recognize and transfer the speech of the users into accurate text	0.046
Compatible operation	It is convenient for the user to obtain the service of AI health care chatbots on a mobile app, WeChat mini program, or website	0.043
Friendly layout	The layout of AI health care chatbots is clear and easy to operate	0.041
Stable service	The AI health care chatbots provide the same smooth service in any situation	0.040
Responsiveness	The response ability of AI health care chatbots	0.167
Anytime response	The AI health care chatbots are available 24 hours a day for 365 days	0.043
Prompt response	The AI health care chatbots always give timely feedback when it is needed	0.042
Coherent response	The AI health care chatbots can communicate with the user seamlessly by maintaining records within the personal account	0.042

Indicator Weights

On the basis of the AHP and percentage weighting method, the weights for all indicators were calculated (Table 4). The primary indicators, ranked from highest to lowest weight, were assurance (0.239), reliability (0.237), human-likeness (0.187), tangibility (0.170), and responsiveness (0.167). Assurance received the highest weight. For the secondary indicators, weights ranged from 0.040 to 0.082. "Understandable answer" had the highest secondary weight (0.082), followed by "Accurate understanding" (0.079), "Targeted question" (0.078), and "Trustworthy advice" (0.071).

Discussion

Principal Findings

The development of AI health care chatbots is on the rise, and their adoption is becoming increasingly vital in modern health care. Providing AI health care chatbots with high service quality is critical to facilitating their broader diffusion and addressing contemporary health care challenges. Although previous studies have attempted to evaluate the quality of AI chatbots in responding to queries related to specific diseases, a comprehensive and user-centered index system for evaluating the service quality of AI health care chatbots has remained lacking. To our knowledge, this is the first study to develop a comprehensive service quality index system for AI health care chatbots from a patient perspective, using SERVQUAL as the theoretical framework. Through a 2-round Delphi process, a finalized set of 5 primary indicators and 17 secondary indicators was derived, specifically designed to capture both the technical functionality and interactive experience unique to AI health care chatbots. Subsequently, the indicator weights were obtained using the AHP.

Among the 5 primary indicators, assurance was identified as the most important dimension, which refers to the ability of AI health care chatbots to provide pertinent responses. Unlike consultations with a clinician, AI health care chatbots lack the

XSL•FO
RenderX

capacity to perform physical examinations to support diagnosis. Therefore, it is essential for AI health care chatbots to deliver goal-oriented and unambiguous conversations, accurately understand user queries, ask follow-up questions with contextual awareness, and provide understandable answers [39,48]. Previous studies [8-10] have taken understandability, often reflected through situation-appropriate response length and information quantity, as a sole metric to measure the quality of AI health care chatbots. Consistent with this emphasis, the secondary indicator "Understandable answer" received the highest weight among secondary indicators. Although aspects such as completeness and consistency have been associated with answer readability in previous studies [11,48], they were not included in this framework. This omission stems from their highly specialized and profession-centric evaluation criteria, which may not align with patient-centered usability expectations.

The reliability dimension ranked second, following assurance. Current AI technologies remain fallible and necessitate oversight by health professionals to ensure the applicability and safety advice of AI health care chatbots [7]. For users, it is critical that AI health care chatbots provide not only trustworthy medical advice regarding diagnosis, medication, and examinations but also clear explanations that enhance transparency and facilitate informed decision-making, thereby fostering trust and promoting sustained engagement [4,28]. Both users and clinicians have underscored the importance of clearly indicating the limitations of AI-generated medical advice [3-5]. In contrast to earlier studies [9,49], this study deliberately excluded diagnostic accuracy, a frequently used metric, from this index system, as patients generally lack the specialized medical knowledge required to evaluate this aspect.

The human-likeness dimension is considered a distinctive feature of AI health care chatbots [1]. Although it is not ranked highest among the primary indicators, its associated secondary indicators, "Personalized response" and "Emotional attention," had prominent weight values. Users often perceive AI health care chatbots as an "AI doctor" and tend to evaluate it through direct comparison with human clinicians along these dimensions [3]. Enhancing humanlike attributes in interactions of AI health care chatbots remains a critical development objective. This entails increased efforts to enable AI health care chatbots to generate context-aware and individualized replies that adapt to both the conversational flow and user preferences [29].

The dimension of tangibility refers to hardware and software manifestations of AI health care chatbots. Among its secondary indicators, "Accurate recognition" was assigned the highest weight. As an information system designed to provide health guidance [50], the accuracy of recognizing and transferring user speech into precise text is the basis for correctly interpreting user queries and facilitating subsequent consultation processes [9]. The next important indicator, "Compatible operation," reflects the accessibility of the service of AI health care chatbots across diverse digital environments. A previous study has emphasized that users valued the ability to access services of AI health care chatbots through various devices, such as smartphones, tablets, or computers, which supports broader and more equitable adoption [31].

Although responsiveness is positioned as the final dimension in the framework, it constitutes a fundamental component of the service quality of AI health care chatbots [5]. This dimension is characterized by the provision of active and uninterrupted guidance 24 hours a day, real-time responses without having to wait in line, immediate accessibility from any location without the need for travel, and seamless communication across different devices [21,28]. Given that AI health care chatbots are supported by LLMs, users perceive responsiveness as their inherent capability [1].

Overall, AI health care chatbots currently represent a viable alternative to human clinicians in initial user interactions [2]. However, its performance can vary significantly depending on the underlying LLMs, knowledge bases, and health data used [31]. While this study developed a service quality index system for AI health care chatbots based on the SERVQUAL framework, most secondary indicators in this study were newly developed to reflect the unique AI context. The proposed index system offers practical value for multiple stakeholders: it enables users to better understand and assess the strengths of AI health care chatbots; supports health service managers in systematically collecting feedback and monitoring performance; and guides developers in conducting feasibility analyses, optimizing design, and implementing postlaunch evaluation. Future research will involve applying this index system in field studies with users of AI health care chatbots to validate their utility and support its ongoing refinement.

Strengths and Limitations

The strengths of this study are as follows. First, this study developed a comprehensive evaluation index system by adapting a scientific framework that incorporates both internationally validated evidence-based indicators and unique features of AI health care chatbots. Second, the mixed methods approach, combining literature review, expert consultations, a 2-round Delphi process, and AHP, ensured a rigorous and systematic development process. Third, this study provides new insights into the systemic evaluation of service quality of AI health care chatbots.

Several limitations should also be acknowledged. First, although the number of experts consulted met methodological requirements, it remained relatively limited. The panel may be subject to selection bias due to the experts' familiarity with AI health care chatbots, which could influence the selection and weighting of indicators. Therefore, the scope of expert consultation needs to be further expanded to enhance the validity of the indicators. Second, all participating experts were based in China, which may limit the generalizability of the findings to other cultural or health system contexts. Future studies should validate the proposed evaluation index across a broader range of settings of AI health care chatbots. Third, this index system has not yet been operationalized and evaluated by the users of AI health care chatbots. Further empirical research is needed to demonstrate its practical relevance and utility and to consider incorporating patient experience into the assessment process. Furthermore, although Kendall *W* was statistically significant, its value reflects only a moderate level of consensus. This implies that the findings are robust but limited in microlevel

XSL•FO
RenderX

ranking. In this study, experts from diverse professional backgrounds likely held different interpretations and assigned varying weights to the indicators, which may have led to evaluation discrepancies.

Conclusions

This study developed a comprehensive, user-centered index system for evaluating the service quality of AI health care chatbots. Through the Delphi method and the AHP, a finalized framework consisting of 5 primary dimensions and 17 secondary indicators was established. These include 4 indicators for tangibility, 3 for responsiveness, 3 for assurance, 3 for human-likeness, and 4 for reliability. This index system prioritizes user needs and experiences and can practically quantify the service quality of AI health care chatbots. The proposed index system will provide valuable support for health policymakers, service managers, and developers by enabling benchmark comparisons, facilitating quality monitoring, and guiding continuous service enhancement.

Acknowledgments

The authors attest that no generative artificial intelligence technologies were used in the generation of text, figures, or other informational content of this manuscript. The authors acknowledge all experts who participated in the Delphi process and the analytic hierarchy process.

Data Availability

The datasets used or analyzed during this study are available from the corresponding author on reasonable request.

Funding

This study was supported by the Social Science Fund of Capital Medical University, Beijing, China.

Authors' Contributions

YG conceptualized the study, designed the consulted index system, interpreted the data, and wrote the manuscript. XW collected and analyzed the data. Both authors approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 18 KB - jmir_v28i1e83051_app1.docx]

References

1. Laymouna M, Ma Y, Lessard D, Schuster T, Engler K, Lebouché B. Roles, users, benefits, and limitations of chatbots in health care: rapid review. J Med Internet Res 2024 Jul 23;26:e56930 [FREE Full text] [doi: 10.2196/56930] [Medline: 39042446]
2. Li H, Zhang R, Lee YC, Kraut RE, Mohr DC. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med 2023 Dec 19;6(1):236 [FREE Full text] [doi: 10.1038/s41746-023-00979-5] [Medline: 38114588]
3. Iacobucci G. The AI bot will see you now: how technology is changing the doctor-patient relationship. BMJ 2024 Mar 28;384:q711. [doi: 10.1136/bmj.q711] [Medline: 38548277]
4. Ng JY, Maduranayagam SG, Suthakar N, Li A, Lokker C, Iorio A, et al. Attitudes and perceptions of medical researchers towards the use of artificial intelligence chatbots in the scientific process: an international cross-sectional survey. Lancet Digit Health 2025 Jan;7(1):e94-102 [FREE Full text] [doi: 10.1016/S2589-7500(24)00202-4] [Medline: 39550312]
5. Mayer CJ, Mahal J, Geisel D, Geiger EJ, Staatz E, Zappel M, et al. User preferences and trust in hypothetical analog, digitalized and AI-based medical consultation scenarios: an online discrete choice survey. Comput Human Behav 2024 Dec;161:108419-108460 [FREE Full text] [doi: 10.1016/j.chb.2024.108419]
6. Armbruster J, Bussmann F, Rothhaas C, Titze N, Grützner PA, Freischmidt H. "Doctor ChatGPT, can you help me?" the patient's perspective: cross-sectional study. J Med Internet Res 2024 Oct 01;26:e58831 [FREE Full text] [doi: 10.2196/58831] [Medline: 39352738]
7. Moore I, Magnante C, Embry E, Mathis J, Mooney S, Haj-Hassan S, et al. Doctor AI? A pilot study examining responses of artificial intelligence to common questions asked by geriatric patients. Front Artif Intell 2024 Jul 25;7:1438012 [FREE Full text] [doi: 10.3389/frai.2024.1438012] [Medline: 39118788]

XSL•FO
RenderX

8. Lee D, Brown M, Hammond J, Zakowski M. Readability, quality and accuracy of generative artificial intelligence chatbots for commonly asked questions about labor epidurals: a comparison of ChatGPT and Bard. Int J Obstet Anesth 2025 Feb;61:104317. [doi: 10.1016/j.ijoa.2024.104317] [Medline: 39754839]
9. Olszewski R, Watros K, Mańczak M, Owoc J, Jeziorski K, Brzeziński J. Assessing the response quality and readability of chatbots in cardiovascular health, oncology, and psoriasis: a comparative study. Int J Med Inform 2024 Oct;190:105562 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105562] [Medline: 39059084]
10. Wang Y, Chen Y, Sheng J. Assessing ChatGPT as a medical consultation assistant for chronic Hepatitis B: cross-language study of English and Chinese. JMIR Med Inform 2024 Aug 08;12:e56426 [FREE Full text] [doi: 10.2196/56426] [Medline: 39115930]
11. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. JAMA Oncol 2023 Oct 01;9(10):1437-1440. [doi: 10.1001/jamaoncol.2023.2947] [Medline: 37615960]
12. Stapleton P, Santucci J, Cundy TP, Sathianathen N. Quality of information on Wilms tumor from artificial intelligence chatbots: what are your patients and their families reading? Urology 2025 Apr;198:130-134. [doi: 10.1016/j.urology.2025.01.054] [Medline: 39914668]
13. Owens OL, Leonard MS. Evaluating an AI chatbot "Prostate Cancer Info" for providing quality prostate cancer screening information: cross-sectional study. JMIR Cancer 2025 May 21;11:e72522 [FREE Full text] [doi: 10.2196/72522] [Medline: 40397820]
14. Anastasio MK, Peters P, Foote J, Melamed A, Modesitt SC, Musa F, et al. The doc versus the bot: a pilot study to assess the quality and accuracy of physician and chatbot responses to clinical questions in gynecologic oncology. Gynecol Oncol Rep 2024 Aug 08;55:101477 [FREE Full text] [doi: 10.1016/j.gore.2024.101477] [Medline: 39224817]
15. He W, Zhang W, Jin Y, Zhou Q, Zhang H, Xia Q. Physician versus large language model chatbot responses to web-based questions from autistic patients in Chinese: cross-sectional comparative analysis. J Med Internet Res 2024 Apr 30;26:e54706 [FREE Full text] [doi: 10.2196/54706] [Medline: 38687566]
16. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
17. Parasuraman A, Zeithaml VA, Berry LL. A conceptual model of service quality and its implications for future research. J Mark 1985 Sep 01;49(4):41-50 [FREE Full text] [doi: 10.1177/002224298504900403]
18. Karasan A, Erdogan M, Cinar M. Healthcare service quality evaluation: an integrated decision-making methodology and a case study. Socio-Econ Plan Sci 2022 Aug;82:101234 [FREE Full text] [doi: 10.1016/j.seps.2022.101234]
19. Zheng QL, Kong LN, Hu P, Liu DX. Identifying quality indicators for home care services: a modified Delphi and Analytic Hierarchy Process study. BMC Nurs 2024 Jul 19;23(1):494 [FREE Full text] [doi: 10.1186/s12912-024-02169-4] [Medline: 39026316]
20. Lemmen C, Woopen C, Stock S. Systems medicine 2030: a Delphi study on implementation in the German healthcare system. Health Policy 2021 Jan;125(1):104-114 [FREE Full text] [doi: 10.1016/j.healthpol.2020.11.010] [Medline: 33288301]
21. Xia M, Liu Q, Ma L, Wen J, Xue Y, Hu H, et al. Developing an evaluation index system for service capability of internet hospitals in China: mixed methods study. J Med Internet Res 2025 Jul 25;27:e72931 [FREE Full text] [doi: 10.2196/72931] [Medline: 40712159]
22. Jiang F, Liu T, Zhou H, Rakofsky JJ, Liu H, Liu Y, et al. Developing medical record-based, healthcare quality indicators for psychiatric hospitals in China: a modified Delphi-Analytic Hierarchy Process study. Int J Qual Health Care 2019 Dec 31;31(10):733-740. [doi: 10.1093/intqhc/mzz005] [Medline: 30753601]
23. Liu M, Yu Q, Liu Y. Developing quality indicators for cancer hospitals in China: a national modified Delphi process. BMJ Open 2024 Apr 09;14(4):e082930 [FREE Full text] [doi: 10.1136/bmjopen-2023-082930] [Medline: 38594187]
24. Saaty RW. The analytic hierarchy process—what it is and how it is used. Math Model 1987;9(3-5):161-176 [FREE Full text] [doi: 10.1016/0270-0255(87)90473-8]
25. Geng Y, Zhao L, Wang Y, Jiang Y, Meng K, Zheng D. Competency model for dentists in China: results of a Delphi study. PLoS One 2018 Mar 22;13(3):e0194411 [FREE Full text] [doi: 10.1371/journal.pone.0194411] [Medline: 29566048]
26. Sun H, Wang Y, Cai H, Wang P, Jiang J, Shi C, et al. The development of a performance evaluation index system for Chinese Centers for Disease Control and Prevention: a Delphi consensus study. Glob Health Res Policy 2024 Jul 23;9(1):28 [FREE Full text] [doi: 10.1186/s41256-024-00367-w] [Medline: 39044214]
27. Li Z, Guo R. Developing online medical service quality indicators in China from the perspective of online and offline integration: a modified Delphi-analytic hierarchy process study. Int J Qual Health Care 2023 Jun 16;35(2):mzad038 [FREE Full text] [doi: 10.1093/intqhc/mzad038] [Medline: 37279543]
28. Barletta VS, Caivano D, Colizzi L, Dimauro G, Piattini M. Clinical-chatbot AHP evaluation based on "quality in use" of ISO/IEC 25010. Int J Med Inform 2023 Feb;170:104951. [doi: 10.1016/j.ijmedinf.2022.104951] [Medline: 36525800]
29. Sobowale K, Humphrey DK. Evaluating the quality of psychotherapy conversational agents: framework development and cross-sectional study. JMIR Form Res 2025 Jul 02;9:e65605 [FREE Full text] [doi: 10.2196/65605] [Medline: 40600851]

XSL•FO
RenderX

30. Yang Y, Liu S, Lei P, Huang Z, Liu L, Tan Y. Assessing usability of intelligent guidance chatbots in Chinese hospitals: cross-sectional study. Digit Health 2024 Jun 06;10:20552076241260504 [FREE Full text] [doi: 10.1177/20552076241260504] [Medline: 38854920]
31. Shiferaw MW, Zheng T, Winter A, Mike LA, Chan LN. Assessing the accuracy and quality of artificial intelligence (AI) chatbot-generated responses in making patient-specific drug-therapy and healthcare-related decisions. BMC Med Inform Decis Mak 2024 Dec 24;24(1):404 [FREE Full text] [doi: 10.1186/s12911-024-02824-5] [Medline: 39719573]
32. Motegi M, Shino M, Kuwabara M, Takahashi H, Matsuyama T, Tada H, et al. Comparison of physician and large language model chatbot responses to online ear, nose, and throat inquiries. Sci Rep 2025 Jul 01;15(1):21346 [FREE Full text] [doi: 10.1038/s41598-025-06769-1] [Medline: 40596359]
33. Abd-Alrazaq A, Safi Z, Alajlani M, Warren J, Househ M, Denecke K. Technical metrics used to evaluate health care chatbots: scoping review. J Med Internet Res 2020 Jun 05;22(6):e18301 [FREE Full text] [doi: 10.2196/18301] [Medline: 32442157]
34. Choi J, Kim JW, Lee YS, Tae JH, Choi SY, Chang IH, et al. Availability of ChatGPT to provide medical information for patients with kidney cancer. Sci Rep 2024 Jan 17;14(1):1542 [FREE Full text] [doi: 10.1038/s41598-024-51531-8] [Medline: 38233511]
35. Chagas BA, Pagano AS, Prates RO, Praes EC, Ferreguetti K, Vaz H, et al. Evaluating user experience with a chatbot designed as a public health response to the COVID-19 pandemic in Brazil: mixed methods study. JMIR Hum Factors 2023 Apr 03;10:e43135 [FREE Full text] [doi: 10.2196/43135] [Medline: 36634267]
36. Chaudhry BM, Debi HR. User perceptions and experiences of an AI-driven conversational agent for mental health support. Mhealth 2024 Jul;10:22 [FREE Full text] [doi: 10.21037/mhealth-23-55] [Medline: 39114462]
37. Noor N, Rao Hill S, Troshani I. Developing a service quality scale for artificial intelligence service agents. Eur J Mark 2022 May 09;56(5):1301-1336 [FREE Full text] [doi: 10.1108/EJM-09-2020-0672]
38. Chen Q, Gong YM, Lu YB, Tang J. Classifying and measuring the service quality of AI chatbot in frontline service. J Bus Res 2022 Jun;145:552-568. [doi: 10.1016/j.jbusres.2022.02.088]
39. Chen S, Wang P, Wood J. Exploring the varying effects of chatbot service quality dimensions on customer intentions to switch service agents. Sci Rep 2025 Jul 02;15(1):22559 [FREE Full text] [doi: 10.1038/s41598-025-06490-z] [Medline: 40596391]
40. Radziwill NM, Benton MC. Evaluating quality of chatbots and intelligent conversational agents. arXiv. Preprint posted online April 15, 2017 [FREE Full text] [doi: 10.48550/arXiv.1704.04579]
41. Pattanshetti S. Extensive study: performance, metrics and usability of chatbot. Int Res J Eng Technol 2021 Sep;8(9):1527-1534 [FREE Full text]
42. Borsci S, Malizia A, Schmettow M, van der Velde F, Tariverdiyeva G, Balaji D, et al. The chatbot usability scale: the design and pilot of a usability scale for interaction with AI-based conversational agents. Pers Ubiquit Comput 2021 Jul 21;26(1):95-119 [FREE Full text] [doi: 10.1007/S00779-021-01582-9]
43. Pelau C, Dabija DC, Ene I. What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. Comput Human Behav 2021 Sep;122:106855. [doi: 10.1016/j.chb.2021.106855]
44. Liu YL, Hu B, Yan W, Lin Z. Can chatbots satisfy me? A mixed-method comparative study of satisfaction with task-oriented chatbots in mainland China and Hong Kong. Comput Human Behav 2023 Jun;143:107716. [doi: 10.1016/j.chb.2023.107716]
45. Chopra A, Ranjani KS, Narsipur S. Service quality dimensions in AI-enabled chatbots leading to customer satisfaction: a study from South Asia. IIFT Int Bus Manag Rev 2023;1(1):26-49. [doi: 10.1177/jiift.221150355]
46. Shahzad MF, Xu S, An X, Javed I. Assessing the impact of AI-chatbot service quality on user e-brand loyalty through chatbot user trust, experience and electronic word of mouth. J Retail Consum Serv 2024 Jul;79:103867 [FREE Full text] [doi: 10.1016/j.jretconser.2024.103867]
47. Young KS, Lee SG, Hong GY. User satisfaction with the service quality of ChatGPT. Serv Bus 2024 Sep 02;18(3-4):417-431 [FREE Full text] [doi: 10.1007/S11628-024-00566-y]
48. Fanous A, Steffner K, Daneshjou R. Patient attitudes toward the AI doctor. Nat Med 2024 Nov 23;30(11):3057-3058. [doi: 10.1038/s41591-024-03272-4] [Medline: 39313596]
49. Büker M, Mercan G. Readability, accuracy and appropriateness and quality of AI chatbot responses as a patient information source on root canal retreatment: a comparative assessment. Int J Med Inform 2025 Sep;201:105948. [doi: 10.1016/j.ijmedinf.2025.105948] [Medline: 40288015]
50. Grilo A, Marques C, Corte-Real M, Carolino E, Caetano M. Assessing the quality and reliability of ChatGPT's responses to radiotherapy-related patient queries: comparative study with GPT-3.5 and GPT-4. JMIR Cancer 2025 Apr 16;11:e63677 [FREE Full text] [doi: 10.2196/63677] [Medline: 40239208]

Abbreviations

AHP: analytic hierarchy process
AI: artificial intelligence

XSL•FO
RenderX

LLM: large language model

Edited by A Stone; submitted 28.Aug.2025; peer-reviewed by MM Al-Tehewy, K Malas; comments to author 24.Oct.2025; accepted 30.Dec.2025; published 18.Feb.2026.

Please cite as:
Gu Y, Wang X
Developing a Service Quality Index System for AI Health Care Chatbots: Mixed Methods Study
J Med Internet Res 2026;28:e83051
URL: https://www.jmir.org/2026/1/e83051
doi:10.2196/83051
PMID:

©Yu Gu, Xinyi Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparison of Emotional Content in Text Responses From Physicians and AI Chatbots to Patient Health Queries: Cross-Sectional Study

Daniel T Burns[1], MS; Channing Bice[2], MS; Paul E Johnson[3], MD; Nicholas Chia[4], PhD; Timothy Robinson[1], PhD

[1]Department of Mathematics & Statistics, University of Wyoming, Laramie, WY, United States

[2]Department of Journalism & Media Communication, Colorado State University, Fort Collins, CO, United States

[3]Department of Otolaryngology- Head and Neck Surgery, University of Washington, Seattle, WA, United States

[4]Data Science and Learning, Argonne National Laboratory, Lemont, IL, United States

Corresponding Author:
Daniel T Burns, MS
Department of Mathematics & Statistics
University of Wyoming
1000 E University Ave
Laramie, WY, 82071
United States
Phone: 1 307 766 4221
Email: dburns5@uwyo.edu

Abstract

Background: Surveys show that many people are willing to use generative artificial intelligence (AI) for health questions. Prior research has largely focused on chatbot accuracy, with some studies finding that both physicians and consumers overwhelmingly prefer chatbot-generated text over physician responses.

Objective: This study aimed to characterize and compare the emotional content of responses from physicians and 2 AI chatbots (OpenAI's ChatGPT and Google's Gemini) and to assess differences in reading level and use of medical disclaimers.

Methods: A public, patient-deidentified telehealth website was used to compile 100 physician-answered questions. The same questions were posed to both chatbots between May 18 and 19, 2025. Two coders classified the emotional content of each sentence using a predefined codebook and reviewed for agreement. Emotions were ranked as primary, secondary, and tertiary by the proportion of sentences classified as each emotion per response. Multinomial logistic regression compared emotional rankings using physician responses as the reference. Word count, Flesch Reading Ease, and Flesch-Kincaid Grade Level were analyzed via ANOVA with the Tukey honestly significant difference test. Disclaimer use was compared between chatbots using a χ^2 test.

Results: Primary emotions were overwhelmingly neutral, except for one response from each chatbot in which anger was primary. For secondary emotions, the odds ratio of hope was 80.28% (95% CI 37.71%-93.76%) lower for ChatGPT, while the odds ratio of fear was 3.29 (95% CI 1.44-7.49) times higher for Gemini. For tertiary emotions, the odds ratio of compassion was 1.94 (95% CI 1.06-3.54) times higher, and the odds ratio of having no tertiary emotion was 84.33% (95% CI 64.72%-93.04%) lower for Gemini. Gemini responses averaged 889.1 (SD 305.7) words, ChatGPT 476.5 (SD 109.5), and physicians 193.5 (SD 113.6). Gemini had the lowest average Flesch Reading Ease score at 39.9 (SD 8.8), followed by ChatGPT at 45.8 (SD 12.8), while physicians had the highest at 51.9 (SD 13.6). Gemini had the highest average Flesch-Kincaid Grade Level at 11.3 (SD 1.5), followed by ChatGPT at 9.9 (SD 1.9), and physicians at 9.2 (SD 2.4). Gemini was significantly more likely to include a disclaimer than ChatGPT (χ^2_1=49.2; P<.001).

Conclusions: Chatbot responses were significantly (P<.001) longer and more difficult to read than physician responses and were more likely to contain a wider range of emotions. Qualitatively, chatbot responses were more varied in their presentation as well as in the breadth of the emotions themselves. The findings of this study could be used to inform more emotionally connected physician responses to patient message queries.

(J Med Internet Res 2026;28:e85516) doi:10.2196/85516

XSL•FO
RenderX

KEYWORDS

responses; emotional; chatbot; physicians; health; ChatGPT; artificial intelligence; AI

Introduction

Health communication extends beyond facts to include emotional reassurance, persuasion, and support, all of which are central to how people make health decisions. As more people turn to the internet for health information, generative artificial intelligence (GenAI) chatbots have rapidly emerged as influential sources for medical guidance, self-diagnosis, and emotional support. Research shows that many users, including physicians, often rate chatbot responses as accurate, empathetic, and useful, though concerns remain about misinformation, ethics, and accountability. Internet-driven health information seeking continues to grow. A report published in 2023 by the US Centers for Disease Control and Prevention found that 58.5% of adults surveyed in the second half of 2022 said they used the internet to look for health or medical information [1]. Between the time of data collection and publication of that study, GenAI chatbots began to be released for widespread public use, starting with ChatGPT (version 3.5; OpenAI) in November 2022. Subsequently, a 2023 study collected responses from 607 participants, of which 476 (78.4%) said that they were willing to use ChatGPT for self-diagnosis [2], while a 2025 study found that 21.2% (63/297) of respondents actively used chatbots for health information seeking [3].

Multiple studies have evaluated the accuracy of responses from chatbots to health-related questions across topics such as cancer, emergency medicine, general inquiries, and symptom checking. The results of these studies ranged from generally quite good (for symptom checking and general inquiries) to fair regarding the accuracy of information provided and were mixed in terms of dangerous or misinformative responses (no such responses in a cancer study and up to 35% of responses in an emergency medicine study) [4-7]. Beyond response accuracy, research has found that physicians prefer responses crafted by chatbots more than responses from other physicians 78% of the time [6]. Additionally, both patients and physicians have ranked ChatGPT responses significantly higher in empathy and usefulness than physician responses [8]. Recent studies have revealed that although health care professionals have favorable opinions regarding chatbot efficiency, they raise concerns about ethics and accountability, while the public appreciates the accessibility and emotional support provided by chatbots [9].

Health communication is a multidisciplinary field that integrates concepts from media studies, psychology, sociology, anthropology, informatics, education, and medical sciences [10,11]. It encompasses a variety of spoken, written, and computer-mediated texts and processes [12,13]. Health communication plays a crucial role in delivering health information to a variety of audiences, thereby enhancing patient and health care professional or medical institution understanding of health risks and guiding them in prevention, detection, diagnosis, and treatment [14]. Health communication can come from diverse sources, including health care professionals, family members, peer support groups, health education materials, news articles, and online resources [15,16].

Beyond factual information, individuals also seek emotion-based information to address the emotional dimensions of health concerns [17]. Feelings such as fear, anxiety, worry, and shame often accompany health-related decision-making, making emotional support, empathy, and encouragement critical components of health communication [18]. Supportive and emotionally responsive information has been shown to alleviate health-related stress, worry, and depression, while also enhancing individuals' commitment to managing health risks [14,19,20].

Studies have examined users' perceptions of chatbots, including their likability, capacity to build rapport, and ability to demonstrate empathy [13,21,22]. Recent studies have also demonstrated the persuasiveness of chatbots, including their ability to reduce belief in conspiracy theories and to change people's minds more broadly [23-25]. The increase in willingness to use the internet and GenAI for health information seeking corresponds with a rise in electronic patient portal messaging, particularly in the wake of the COVID-19 pandemic, and the increased burden this places on physicians and health care services [26-29]. Thus, the aim of this study was to investigate key differences in the emotional support and readability of health information offered by physicians and the commonly used artificial intelligence (AI) chatbots ChatGPT and Gemini (version 2.0 Flash; Google). We describe the emotional content and answer complexity contained within text responses to a variety of health questions across medical specialty topics, as posted on an online telehealth website.

Methods

Data Acquisition and Preparation

Following similar data collection methods as previously described [30-34], data were obtained from iCliniq (Orane Healthcare India Private Limited), an international telehealth website with publicly accessible question-and-answer forums where users can pay to ask verified physicians questions. These anonymized questions are categorized by condition and medical specialty by the website. For this study, questions and answers were selected from a subset of 10 medical specialties. From these 10 specialties, 10 question-answer pairs were randomly selected per specialty, resulting in a total of 100 pairs. The 10 specialties were allergy specialist, dermatology, family physician, general practitioner, internal medicine, medical oncology, neurology, pediatrics, psychiatry, and urology. Questions from this website were chosen using stratified random sampling (with medical specialties as strata) and subsequently posed to chatbots, and responses were compared with those of the responding physicians. In cases where the question elicited a dialogue (ie, back-and-forth discussion beyond the first question and response), only the initial question and answer were used. We next removed greetings and stock response openings from the questions and answers to reflect a more natural conversational tone, especially for the questions, to ensure that the chatbot responses were not unduly influenced.

XSL•FO
RenderX

For example, phrases such as “Hello doctor,” “Hi doctor,” “Dear doctor,” and “Doctor,” were removed from the patient questions, while phrases such as “Hi, Welcome” and “Hello, Welcome back” were removed from the physician responses. Responses were copied to a text editor (eg, a Microsoft Word or PDF document) for downstream classification analysis.

AI Chatbots

Two chatbots were used for comparison with physician responses: ChatGPT (GPT-4o) and Gemini 2.0 Flash. For each chatbot, each question (prompt) was asked in a fresh session, with “Memory” (chat history) option set to off so that no previous question or response impacted a subsequent interaction. Gemini was queried via its application programming interface from a custom Python script on May 18, 2025. As ChatGPT offered no free application programming interface access, questions were manually copied and pasted into the ChatGPT user interface on its website, and responses were similarly copied and pasted into a text editor. ChatGPT was queried from May 18 to May 19, 2025, as rate limits inhibited the ability to ask all 100 questions on the same day. The chatbots were given no special instructions, and no prompt engineering was applied to mimic real-world patient use [35-39]. Questions were asked under the default settings of each chatbot, and chatbots were free to respond however they saw fit based on their settings.

Response Summaries

Word count was calculated for all responses using the R package *stringi* [40]. Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL) are commonly used scores for assessing the readability of written materials [41-45]. The FRE score uses a formula containing total words, total sentences, and total syllables to calculate a score that indicates the reading difficulty of the material, with higher scores indicating easier material (eg, scores of 90-100 are considered very easy to read and understandable by an average 11-year-old, while scores of 30-50 are considered difficult to read and understandable by someone in college). The FKGL formula uses the same parameters as the FRE formula but calculates a score as a US grade level (or number of years of education needed) for judging readability. FRE and FKGL scores were calculated using Microsoft Word’s built-in summaries. According to the National Library of Medicine, nearly 9 out of 10 adults struggle with health literacy [46]. The US Department of Education reported that 54% of Americans aged 16 to 74 years read below a sixth-grade level [47]. The American Medical Association recommends that patient education materials be written at or below a sixth-grade level [48], while the National Institutes of Health recommends it to be at or below an eighth-grade level [41,49].

To get a better understanding of how these chatbots communicate about health-related topics, their use of “not a doctor” disclaimers was also assessed. In this context, a disclaimer consisted of any version of the chatbot explicitly stating that it is not a physician or medical professional, that the information it provided did not constitute medical advice, or that it was an AI and could not provide medical advice. This variable was recorded as 1 for “yes” and 0 for “no” depending on whether a response contained such a disclaimer.

Emotion Classification

A quantitative content analysis was conducted to systematically classify the emotional content embedded within physician and chatbot responses. Defined as “a research technique for the objective, systematic, and quantitative description of the manifest content of communication” [50], this method allowed us to categorize and quantify textual features to identify patterns and frequencies in the data. The codebook was developed using a robust body of literature in psychology and media communication to conceptualize and define each variable of interest (Table 1). The codebook contained 6 emotions (anger, compassion, fear, guilt, hope, and sadness) plus neutral or rational. For brevity, we summarize the codes, definitions, and associated literature in Table 1. The first and second authors (DTB and CB) served as the coders for this study. The unit of analysis (ie, the content coded as a part of our study) comprised the sentences and phrases of each response. The coders used MAXQDA (version 24.10; VERBI GmbH) to conduct the coding and generate classification counts. These counts were then used to rank the top 3 emotions expressed in each response (ie, primary, secondary, and tertiary), using MAXQDA’s Document Portrait function (refer to Multimedia Appendix 1 or “Document Portrait” entry of the MAXQDA Manual [51] for details).

XSL•FO
RenderX

Table 1. Emotion classification codebook.

Emotion	Definition	Examples	References
Anger	The response contains an emotion characterized by tension or hostility arising from frustration or the belief that an injury or injustice has been committed against oneself or an important other. It can manifest itself in behaviors designed to remove the object of anger.	• "My doctor told me that my symptoms were 'all in my head' and now I have cancer that has progressed to stage 4, when it could have been caught earlier." • "It's so frustrating that your concerns aren't being taken seriously, that's unacceptable!"	[52-54]
Compassion	The response contains an emotion characterized by feelings of empathy in response to witnessing the pain or discomfort of others. It can manifest as a desire to help alleviate pain or discomfort and to engage in helping or sharing behaviors.	• "I can help you find a nearby dermatology clinic." • "I'm really sorry to hear that you're feeling concerned about this. It's understandable to be worried, but based on what you've described, there are a few possible explanations that might be less serious than cancer. That said, it's still important to get it checked out."	[55,56]
Fear	The response contains an emotion characterized by feelings of worry that is typically triggered by detecting an imminent threat. It can manifest itself as behaviors designed to move away from or avoid the source of fear.	• "COVID-19 can be lethal." • "Applying bleach to a skin lesion can be dangerous." • "Swallowing a toothpick can be dangerous and may lead to serious health risks, though the actual risk of dying depends on a variety of factors, such as the size and shape of the toothpick, how it travels through the digestive system, and whether it causes complications."	[52-54]
Guilt	The response contains a self-conscious emotion characterized by an appraisal of having done (or thought) something that is wrong or violates internalized or socially acceptable standards of behavior. It can manifest itself as behaviors designed to undo or mitigate the wrong.	• "If you have COVID-19, don't go to work or school—stay home to protect others." • "I should have taken my child to the doctor sooner."	[52-54]
Hope	The response contains a prospect-based emotion experienced as a desire for a better future, despite current uncertainty or unfavorable circumstances. It manifests itself as behaviors designed to move toward their goals by taking action and/or planning for the future.	• "Although your lesion looks bad now, consistently applying Vaseline will help prevent infection and scarring." • "We caught the condition early, so many good treatment options are available."	[53,57]
Sadness	The response contains an emotion characterized by sorrow, loss, resignation, or the absence of a desired reward. It manifests itself through a lack of motivation or actions to change one's circumstances.	• "I'm tired of trying new depression medications. Nothing is helping." • "I can barely get out of bed I feel so hopeless."	[53,58]
Neutral or rational	The response does not contain an emotion but offers fact-based and verifiable information, definitions, or advice about the health condition that appeals to one's reason or rational thought processes.	• "Since the cough has been persisting for weeks, and it's affecting your ability to sleep, it would be a good idea to consult a health care professional to rule out more serious issues like pneumonia, bronchitis, or any other lung concerns." • "Sounds like you will be fine. You should flush the eye anytime you get a chemical or foreign body in the eye. You can also contact Poison Control 1-800-222-1222."	[54,59]

To ensure consistency, the authors discussed the conceptual and operational definitions of each emotion and reviewed sample responses. We then independently coded a practice set of 10 responses to verify the codebook's applicability. Through regular meetings and discussions, the 2 coders compared the categorization of responses, resolved discrepancies, and revised and modified the codebook. An example of a coded item and document portrait is presented in Multimedia Appendix 1. Intercoder reliability was assessed using a stratified (by medical specialty) random selection of 50 responses independently evaluated by the 2 coders, followed by 2 rounds of joint review. Reliability was calculated using Cohen κ coefficient, which takes chance agreement into account [60]. Initially, Cohen κ ranged between 0.64 and 0.81. Coefficients ≥0.8 were considered reliable [61], and following the second round of review, Cohen κ ranged between 0.81 and 0.86.

XSL•FO
RenderX

Statistical Analysis

All statistical analyses were performed using R (version 4.5.1; R Foundation for Statistical Computing) [62]. For the readability-related response summaries, ANOVAs followed by Tukey's honest significant difference test were used for pairwise comparisons. "Not a doctor" disclaimer use between the 2 chatbots was compared by a χ^2 test. Multinomial logistic regression was used to compare emotion classification results, with physicians set as the reference group.

For the secondary emotion analysis, 39 responses were removed (13 from each responder) because, for each of these 13 questions, the physician's responses contained only one (primary) emotion, necessitating "none" as the classification for the secondary and tertiary emotions. With the exception of one Gemini response, the chatbots did not have any "none" secondary classifications, and thus, coefficient estimation was highly biased and unreliable. Similarly, 9 responses (3 from each responder) were removed from the tertiary emotion analysis, as Gemini was the only responder to have anger classified as a tertiary emotion. Statistical significance was set at $P<.05$ for all tests.

Ethical Considerations

This study analyzed publicly available data from a freely accessible telehealth website. No personal or identifiable patient or physician data were collected or used. According to the US Department of Health and Human Services, 45 Code of Federal Regulations 46.102(e)(1) [63], this study does not meet the definition of human participants research and therefore is exempt from institutional review board review.

Results

Response Summaries

Figure 1A shows the distribution of word counts for each responder. All 3 responders' word counts were significantly different from one another ($F_{2,297}=310.1$; $P<.001$). Gemini had the highest word count, with an average of 889.1 (SD 305.7) words per response, followed by ChatGPT with an average of 476.5 (SD 109.5) words per response, and physicians with the lowest average at 193.5 (SD 113.6) words per response.

All 3 responders' FRE scores were significantly different from one another ($F_{2,297}=27.5$; $P<.001$). As shown in Figure 1B, Gemini had the lowest average FRE score at 39.9 (SD 8.8), followed by ChatGPT at 45.8 (SD 12.8), while physicians had the highest average score at 51.9 (SD 13.6). All 3 responders' FKGL scores were also significantly different from one another ($F_{2,297}=25.7$; $P<.001$). Figure 1C shows that Gemini had the highest average FGKL score of 11.3 (SD 1.5), followed by ChatGPT with an average score of 9.9 (SD 1.9), and physicians again having the easiest readability of the 3 groups, with an average FGKL score of 9.2 (SD 2.4).

As shown in Figure 1D, disclaimer use differed significantly between the 2 chatbots ($\chi^2_1=49.22$; $P<.001$). Gemini made such a statement 58% (58/100; 95% CI 47.71%-67.67%) of the time compared with ChatGPT's 10% (10/100; 95% CI 5.16%-18.04%).

XSL•FO
RenderX

Figure 1. Comparison of response length, readability, and disclaimer use among physicians, ChatGPT, and Gemini. (A) Boxplots, with accompanying density plots, of the word count distributions for each responder. The bold black line inside boxes represents the median, and black circles denote outliers. (B) Mean Flesch Reading Ease (FRE) scores per responder. Higher scores indicate greater readability. (C) Mean Flesch-Kincaid Grade Level (FKGL) scores per responder. Higher scores indicate a higher educational level required for reading comprehension. (D) Proportion of responses that contained disclaimer use for each chatbot. Differences among responders in A-C were calculated via ANOVA followed by the Tukey honestly significant difference test. Differences among responders in D were calculated via the chi-square test. Error bars in B-D represent 95% CIs. *P=.03, **P=.001, ***P<.001.

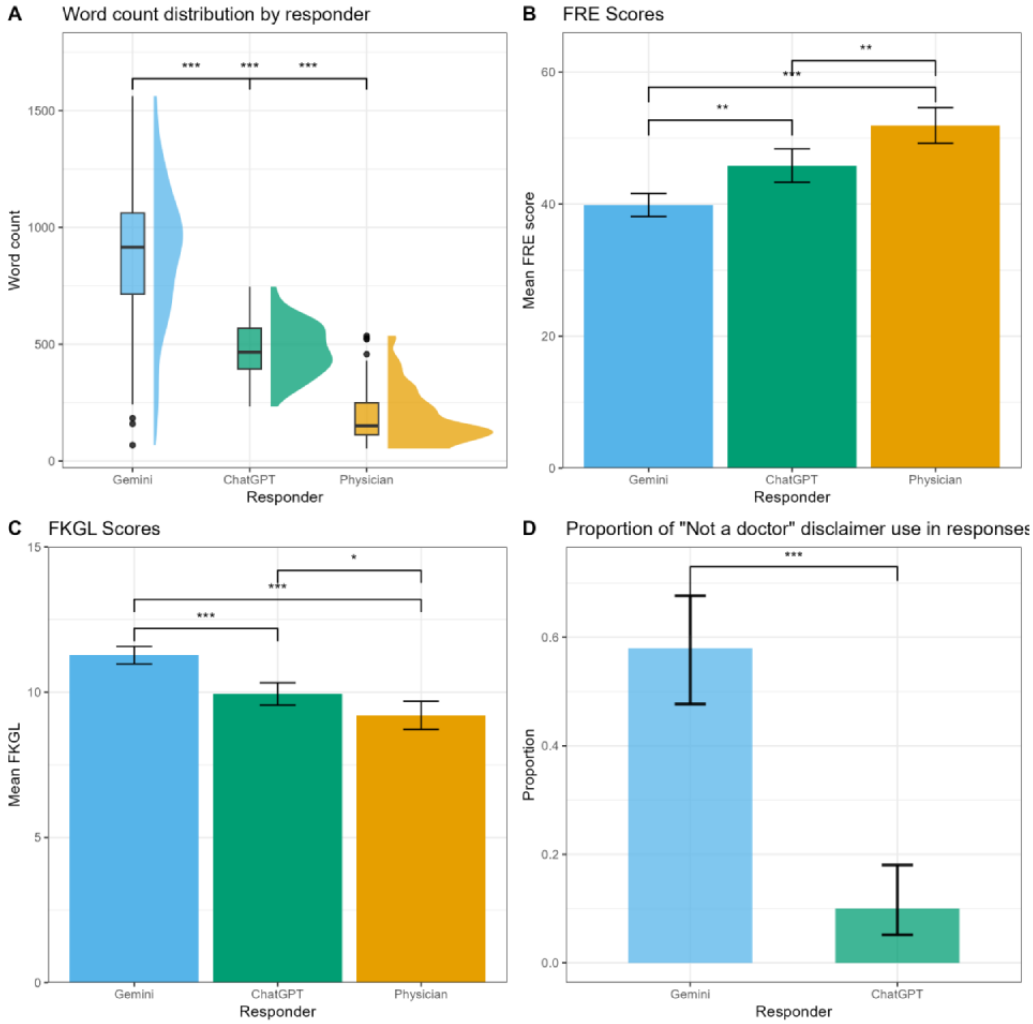

Emotion Classification

Neutral or rational was the most common emotion expressed across all 3 responders. It was the primary emotion for all 100 (100%) of the physician responses and for 99 (99%) out of 100 for each chatbot. In the one response per chatbot in which neutral was not the primary emotion, fear was the most prominent emotion. In both instances, the chatbots directed the question asker to seek immediate medical attention, as they deemed the situations to be very concerning.

All comparisons for secondary and tertiary emotions were made relative to physician responses (ie, the physician group was set as the reference and the chatbots were compared with it). For secondary emotions (Tables 2 and 3), the odds ratio of a hope response was 80.28% (95% CI 37.71%-93.76%; P=.006) lower for ChatGPT. The odds ratio of a fear response was 3.29 (95% CI 1.44-7.49; P=.005) times higher for Gemini. Physician responses did not have a secondary emotion 13% (13/00) of the time (ie, the entire response was classified as containing only one emotion), while Gemini had such responses 1% (1/100) of the time, and ChatGPT always had at least 2 emotions classified in its responses. Physicians never had anger classified as a secondary emotion, while Gemini and ChatGPT each did so once.

For tertiary emotions (Tables 4 and 5), the odds ratio of a compassion response was 1.94 (95% CI 1.06-3.54; P=.03) times higher, and the odds ratio of having no tertiary emotion classified (none) was 84.33% (95% CI 64.72%-93.04%; P<.001) lower for Gemini. The odds ratio of having no tertiary emotion classified (none) was 53% (95% CI 3.42%-78.51%; P=.06) lower for ChatGPT. Gemini was the only responder to have anger classified as a tertiary emotion.

XSL•FO
RenderX

Table 2. Secondary emotions by responder (N=100)[a].

Emotion	Physician, n (%)	ChatGPT-4o, n (%)	Gemini Flash 2.0, n (%)
Anger	0 (0)	1 (1)	1 (1)
Compassion	57 (57)	80 (80)	51 (51)
Fear	10 (10)	11 (11)	33 (33)
Guilt	3 (3)	2 (2)	3 (3)
Hope	17 (17)	5 (5)	10 (10)
Neutral	0 (0)	1 (1)	1 (1)
None	13 (13)	0 (0)	1 (1)
Sadness	0 (0)	0 (0)	0 (0)

[a]Responses from each responder containing the labeled emotions at the secondary rank. Each responder was asked the same 100 questions.

Table 3. Multinomial logistic regression results for secondary emotions[a].

Responder and emotions	Odds ratio (95% CI)	SE	Wald Z	*P* value
ChatGPT				
Compassion	1.268 (0.893-1.800)	0.179	1.329	.18
Fear	0.710 (0.270-1.866)	0.493	−0.695	.49
Guilt	0.263 (0.027-2.597)	1.168	−1.143	.25
Hope	0.197 (0.062-0.623)	0.587	−2.766	.006
Gemini				
Compassion	0.821 (0.556-1.213)	0.199	−0.989	.32
Fear	3.287 (1.442-7.491)	0.420	2.831	.005
Guilt	1.217 (0.234-6.321)	0.840	0.234	.81
Hope	0.685 (0.277-1.693)	0.462	−0.820	.41

[a]Transformed (exponentiated) results of multinomial logistic regression comparing the secondary emotion rankings of the 2 chatbots to physicians.

Table 4. Tertiary emotions by responder (N=100)[a].

Emotion	Physician, n (%)	ChatGPT (GPT-4o), n (%)	Gemini Flash 2.0, n (%)
Anger	0 (0)	0 (0)	3 (3)
Compassion	17 (17)	20 (20)	31 (31)
Fear	10 (10)	18 (18)	12 (12)
Guilt	2 (2)	6 (6)	5 (5)
Hope	13 (13)	22 (22)	32 (32)
Neutral	0 (0)	0 (0)	0 (0)
None	58 (58)	34 (34)	17 (17)
Sadness	0 (0)	0 (0)	0 (0)

[a]Responses from each responder containing the labeled emotions at the tertiary rank. Each responder was asked the same 100 questions.

XSL•FO
RenderX

Table 5. Multinomial logistic regression results for tertiary emotions[a].

Responder and emotion	Odds ratio (95% CI)	SE	Wald Z	*P* value
ChatGPT				
Compassion	1.250 (0.648-2.412)	0.335	0.665	.51
Fear	1.280 (0.458-3.577)	0.524	0.471	.64
Guilt	2.400 (0.425-13.589)	0.883	0.992	.32
Hope	1.354 (0.524-3.500)	0.485	0.625	.53
None	0.471 (0.215-1.034)	0.401	−1.876	.06
Gemini				
Compassion	1.937 (1.060-3.542)	0.308	2.149	.03
Fear	0.619 (0.220-1.741)	0.527	−0.908	.36
Guilt	1.290 (0.225-7.405)	0.891	0.286	.78
Hope	1.270 (0.525-3.072)	0.450	0.531	.60
None	0.157 (0.070-0.353)	0.414	−4.477	<.001

[a]Transformed (exponentiated) results of multinomial logistic regression comparing the tertiary emotion rankings of the 2 chatbots to physicians.

Discussion

Principal Findings

Overall, the AI chatbots provided longer and more detailed responses than physicians. Whether the additional detail provided in chatbot responses was appropriate or correct was not investigated in this study. While initial physician responses were often a series of questions, presumably akin to a history-taking for gathering more information for a proper diagnosis or suggestion, chatbots often presented lists of possible explanations that were ranked from most to least likely. The chatbots also often asked follow-up or clarification questions, but rarely at the expense of providing answers; chatbot responses were combinations of detailed lists and explanations, with occasional clarifying questions interspersed. This additional feature provided by the chatbots could also help explain why the responses were calculated to be more difficult to read. Although all responders provided answers well above the national average reading grade level [47], the chatbots were more difficult to read and understand, which may negate some of the usefulness of the added information, as it could be overly complex for a typical patient to understand and therefore properly act on.

Similarly, the very infrequent use of disclaimers, particularly by ChatGPT, explicitly stating that chatbots are not medical professionals or that the information provided does not replace consultation with a qualified clinician may place users at risk. If the information is presented in a way that is difficult to understand, users could find it challenging to discern useful information from irrelevant or potentially harmful information. This problem is further compounded by chatbots' tendency to provide responses in a manner that conveys strong certainty and expertise in the content [64,65].

The bulk of the responses from all 3 responders were neutral in emotional content, as the vast majority consisted of factual explanations and definitions. Physicians were more likely to provide a response that contained only one emotion (neutral), perhaps as a consequence of their generally much shorter responses, which often consisted of follow-up questions. In contrast, the chatbots had a relatively formulaic response pattern, wherein they opened with an emotional statement (usually compassion or fear) acknowledging the situation of the question asker. Similarly, the chatbots ended their responses with summaries of the overall response and further emotional appeals (usually compassion, fear, or hope).

Although compassionate content proportion did not differ significantly between the chatbots and physicians in a quantitative sense, the quality of the emotions presented was noticeably different. For example, a compassionate statement from physicians often looked like "I understand your concern," "I hope this helps," or "I read your query and can understand your concern," while compassionate statements from the chatbots included "Let's break down the problems and discuss potential solutions," "It's completely understandable that you're feeling lost and frustrated," "I'm really sorry you're going through [this situation]. Here's a structured overview to help guide you through managing your condition and exploring safer, longer-term strategies," and "You're absolutely right to be concerned, and it's great that you're actively looking into this. Let's go over the key points in your situation and steps you can take." Previous research has shown that patients, physicians, and third-party evaluators consistently perceive AI to be more empathetic and compassionate than physicians [8,66-68]. Our results lend support to these findings and help to show that this difference comes from the quality and content of the written responses rather than from the frequency or proportion of an emotion being conveyed. Similar differences in quality of emotion in the responses were present for other emotions as well. These increases in empathy, sympathy, and emotional support may also play a role in why people consistently prefer responses from chatbots over those from physicians [69].

However, further research is needed to explore how the emotional tone of chatbot responses influences individuals'

XSL•FO
RenderX

health information–seeking behaviors and health information–processing behaviors, which in turn affect their engagement with health-related actions. These results, coupled with the previously mentioned patient and physician preferences for chatbot responses, warrant an investigation into possible reasons for these effects. As good communication skills are increasingly viewed as an important aspect of health care training and practice [70-73], these results can be used to inform physicians' use of chatbots as a tool to better help connect with patients in a more empathetic manner. Future research should investigate optimizing prompting strategies for chatbot-assisted message drafting to determine patient perceptions, as well as quality and safety outcomes.

Limitations

There are several limitations to this study. Importantly, a key limitation is the inability to exactly replicate the results. This limitation is due to the ever-changing nature of the AI models used for these chatbots; the models get updated frequently, often without any detail regarding when or what changes were made. Often, these updates fundamentally change the way that chatbots engage with and respond to users. The behavior of the 2 chatbots used in this study, chosen for their popularity and free access, may not be representative of all chatbot services that are currently available.

Another limitation is potential cultural norm disparities. The authors are all from the United States, where both chatbots were accessed and created or hosted by US companies. This factor could cause unintended bias in emotion classification, as the expectations and definitions in one culture may not be the same as those in another. Furthermore, because the website is publicly accessible, physicians responding to questions may have access to patient information that is not visible on the question page and therefore not captured in the dataset. This access feature could complicate the interpretations of this study, as it is possible that physicians gave short, emotionally neutral answers because they had more information than was provided in just the question alone, which the researchers and therefore the chatbots would be unaware of.

Another potential limitation is that the patient questions used were chosen randomly; however, the authors selected the medical specialty topic areas containing the questions. Although the intention was to ensure an array of question types across various topics, it is possible that this approach inadvertently biased results if the selected questions were too similar. Additionally, for the sake of simplicity and feasibility, our focus was limited to 7 emotions, classified by 2 of the authors. Although these were selected from health communication literature and deemed useful and informative for the purposes of this study, it is possible that different emotions, or combinations of emotions, could yield different conclusions as well. For similar reasons, we also limited the selected AI chatbots to 2 free options, which left out popular alternatives, such as Anthropic's Claude, Microsoft's Copilot, and High-Flyer's DeepSeek, thus limiting the scope of this study and its conclusions.

Conclusions

This study demonstrates key differences in the emotional support and readability of health information offered by physicians and AI chatbots (ChatGPT and Gemini). It also extends the literature examining the usefulness of AI-generated drafts by demonstrating chatbots' ability to craft emotionally connective responses, potentially enhancing emotional support for patients, while mitigating physician burnout and compassion fatigue [74]. Compared to physicians, the AI chatbots' responses were much longer and more difficult to read. ChatGPT rarely included a disclaimer in its responses indicating that it is not a medical professional (only 10% of responses), while Gemini included one more often (60% of responses).

Although few statistically significant differences existed between the emotional content of physicians and chatbots, there were substantial qualitative differences in the depth and breadth of emotionality presented, especially for compassion (empathy). This finding could be due, in part, to the preferential training of chatbots to be more personable and agreeable, coupled with the fact that they do not suffer from burnout or fatigue.

An often overlooked aspect of AI-generated responses is the variability or uncertainty that is inherent in the models. All generative AI responses are probabilistic responses, which implies uncertainty can be investigated. However, approaches for assessing that uncertainty remain an unanswered research question; although beyond the scope of this study, this represents an important direction for future research.

Funding

This study was supported by the National Institute of General Medical Sciences of the National Institutes of Health (grant P20GM103432). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This study was also supported by funding from the Wyoming Department of Health.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: DTB, CB, NC, TR
Data curation: DTB
Formal analysis: DTB, CB

XSL•FO
RenderX

Funding acquisition: TR
Investigation: DTB, CB
Methodology: DTB, CB, TR
Project administration: DTB, TR
Resources: DTB, CB, TR
Software: DTB
Supervision: TR
Validation: DTB, CB, TR
Visualization: DTB
Writing—original draft: DTB, CB, TR
Writing—review and editing: DTB, CB, NC, TR, PEJ

Conflicts of Interest

None declared.

Multimedia Appendix 1
Text emotion classification and ranking. (A) Example physician response with sentences highlighted for the coded emotion (gray=neutral or rational, purple=compassion, and orange=fear). (B) Document Portrait classification summary of the coded response in (A), showing the proportion of each emotion in descending order from left to right. In this example response, neutral was coded the most, followed by compassion, then fear. Classifications and document portrait summaries were carried out in MAXQDA software.
[PNG File , 896 KB - jmir_v28i1e85516_app1.png]

References

1. Wang X, Cohen RA. Health information technology use among adults: United States, July–December 2022. National Center for Health Statistics. 2023. URL: https://www.cdc.gov/nchs/products/databriefs/db482.htm [accessed 2025-06-28]
2. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors 2023 May 17;10:e47564 [FREE Full text] [doi: 10.2196/47564] [Medline: 37195756]
3. Yun HS, Bickmore T. Online health information-seeking in the era of large language models: cross-sectional web-based survey study. J Med Internet Res 2025 Mar 31;27:e68560 [FREE Full text] [doi: 10.2196/68560] [Medline: 40163112]
4. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. JAMA Oncol 2023 Oct 01;9(10):1437-1440. [doi: 10.1001/jamaoncol.2023.2947] [Medline: 37615960]
5. Yau JY, Saadat S, Hsu E, Murphy LS, Roh JS, Suchard J, et al. Accuracy of prospective assessments of 4 large language model chatbot responses to patient questions about emergency care: experimental comparative study. J Med Internet Res 2024 Nov 04;26:e60291 [FREE Full text] [doi: 10.2196/60291] [Medline: 39496149]
6. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
7. Chen A, Chen DO, Tian L. Benchmarking the symptom-checking capabilities of ChatGPT for a broad range of diseases. J Am Med Inform Assoc 2024 Sep 01;31(9):2084-2088. [doi: 10.1093/jamia/ocad245] [Medline: 38109889]
8. Armbruster J, Bussmann F, Rothhaas C, Titze N, Grützner PA, Freischmidt H. "Doctor ChatGPT, can you help me?" The patient's perspective: cross-sectional study. J Med Internet Res 2024 Oct 01;26:e58831 [FREE Full text] [doi: 10.2196/58831] [Medline: 39352738]
9. Lu L, Zhu Y, Yang J, Yang Y, Ye J, Ai S, et al. Healthcare professionals and the public sentiment analysis of ChatGPT in clinical practice. Sci Rep 2025 Jan 07;15(1):1223 [FREE Full text] [doi: 10.1038/s41598-024-84512-y] [Medline: 39774168]
10. Schiavo R. Health Communication: From Theory to Practice. Hoboken, NJ: John Wiley & Sons; 2007.
11. Werder O. Transformational Health Communication: A New Perspective on Healthcare and Prevention. Cham, Switzerland: Springer; 2024.
12. Harvey K, Koteyko N. Exploring Health Communication: Language in Action. Milton Park, UK: Routledge; 2013.
13. Park E. I trust you, but let me talk to AI: the role of the chat agents, empathy, and health issues in misinformation guidance. Int J Strat Commun 2025 Feb 20;19(2):231-260. [doi: 10.1080/1553118X.2025.2462087]
14. Kreps GL. Addressing resistance to adopting relevant health promotion recommendations with strategic health communication. Inf Serv Use 2023 May 01;43(2):131-142. [doi: 10.3233/ISU-230187]
15. Kreps GL. Online information and communication systems to enhance health outcomes through communication convergence. Hum Commun Res 2017 Jun 30;43(4):518-530. [doi: 10.1111/hcre.12117]
16. Link E, Baumann E, Kreps GL, Czerwinski F, Rosset M, Suhr R. Expanding the health information national trends survey research program internationally to examine global health communication trends: comparing health information seeking

XSL•FO
RenderX

behaviors in the U.S. and Germany. J Health Commun 2022 Aug 03;27(8):545-554 [FREE Full text] [doi: 10.1080/10810730.2022.2134522] [Medline: 36250315]
17. Yang ZJ, Kahlor L. What, me worry? The role of affect in information seeking and avoidance. Sci Commun 2012 May 24;35(2):189-212. [doi: 10.1177/1075547012441873]
18. Kreps GL, Canzona MR. The role of communication and information in symptom management. In: Wittenberg E, Ferrell BR, Goldsmith J, Smith T, Ragan SL, Glajchen M, et al, editors. Textbook of Palliative Care Communication. Oxford, UK: Oxford University Press; 2015:119-126.
19. Fisher HM, Winger JG, Miller SN, Wright AN, Plumb Vilardaga JC, Majestic C, et al. Relationship between social support, physical symptoms, and depression in women with breast cancer and pain. Support Care Cancer 2021 Sep 15;29(9):5513-5521 [FREE Full text] [doi: 10.1007/s00520-021-06136-6] [Medline: 33723675]
20. Guan M, Han JY, Shah DV, Gustafson DH. Exploring the role of social support in promoting patient participation in health care among women with breast cancer. Health Commun 2021 Nov;36(13):1581-1589 [FREE Full text] [doi: 10.1080/10410236.2020.1773704] [Medline: 32500731]
21. Pelau C, Dabija DC, Ene I. What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. Comput Human Behav 2021 Sep;122:106855. [doi: 10.1016/j.chb.2021.106855]
22. Zhang J, Oh YJ, Lange P, Yu Z, Fukuoka Y. Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet: viewpoint. J Med Internet Res 2020 Sep 30;22(9):e22845 [FREE Full text] [doi: 10.2196/22845] [Medline: 32996892]
23. Costello TH, Pennycook G, Rand DG. Durably reducing conspiracy beliefs through dialogues with AI. Science 2024 Sep 13;385(6714):eadq1814. [doi: 10.1126/science.adq1814] [Medline: 39264999]
24. Salvi F, Horta Ribeiro M, Gallotti R, West R. On the conversational persuasiveness of GPT-4. Nat Hum Behav 2025 Aug 19;9(8):1645-1653. [doi: 10.1038/s41562-025-02194-6] [Medline: 40389594]
25. Schoenegger P, Salvi F, Liu J, Nan X, Debnath R, Fasolo B, et al. Large language models are more persuasive than incentivized human persuaders. ArXiv Preprint posted online on May 14, 2025 [FREE Full text] [doi: 10.48550/arXiv.2505.09662]
26. Ferguson K, Fraser M, Tuna M, Bruntz C, Dahrouge S. The impact of an electronic portal on patient encounters in primary care: interrupted time-series analysis. JMIR Med Inform 2023 Feb 06;11:e43567 [FREE Full text] [doi: 10.2196/43567] [Medline: 36745495]
27. Apathy NC, Hicks K, Bocknek L, Zabala G, Adams K, Gomes KM, et al. Inbox message prioritization and management approaches in primary care. JAMIA Open 2024 Dec;7(4):ooae135. [doi: 10.1093/jamiaopen/ooae135] [Medline: 39530053]
28. Holmgren AJ, Downing NL, Tang M, Sharp C, Longhurst C, Huckman RS. Assessing the impact of the COVID-19 pandemic on clinician ambulatory electronic health record use. J Am Med Inform Assoc 2022 Jan 29;29(3):453-460 [FREE Full text] [doi: 10.1093/jamia/ocab268] [Medline: 34888680]
29. North F, Luhman KE, Mallmann EA, Mallmann TJ, Tulledge-Scheitel SM, North EJ, et al. A retrospective analysis of provider-to-patient secure messages: how much are they increasing, who is doing the work, and is the work happening after hours? JMIR Med Inform 2020 Jul 08;8(7):e16521 [FREE Full text] [doi: 10.2196/16521] [Medline: 32673238]
30. V JP, S AA, Sundaram G. Toward explainable cross-lingual adaptive NAS for enhanced Tamil medical text summarization. IEEE J Biomed Health Inform 2025 Oct 16:1-18. [doi: 10.1109/jbhi.2025.3618959]
31. Gatto J, Seegmiller P, Johnston G, Preum SM. Identifying the perceived severity of patient-generated telemedical queries regarding COVID: developing and evaluating a transfer learning-based solution. JMIR Med Inform 2022 Sep 02;10(9):e37770 [FREE Full text] [doi: 10.2196/37770] [Medline: 35981230]
32. Yang W, Zeng G, Tan B, Ju Z, Chakravorty S, He X, et al. On the generation of medical dialogues for COVID-19. ArXiv Preprint posted online on May 15, 2020 [FREE Full text] [doi: 10.1101/2020.05.08.20095810]
33. Haghighi T, Gholami S, Sokol JT, Kishnani E, Ahsaniyan A, Rahmanian H, et al. EYE-Llama, an in-domain large language model for ophthalmology. iScience 2025 Jul 18;28(7):112984 [FREE Full text] [doi: 10.1016/j.isci.2025.112984] [Medline: 40697409]
34. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. Cureus 2023 Jun;15(6):e40895 [FREE Full text] [doi: 10.7759/cureus.40895] [Medline: 37492832]
35. Knoth N, Tolzin A, Janson A, Leimeister JM. AI literacy and its implications for prompt engineering strategies. Comput Educ Artif Intell 2024 Jun;6:100225. [doi: 10.1016/j.caeai.2024.100225]
36. Xue H, Oh YJ, Zhou X, Zhang X, Oxley B. User prompting strategies and ChatGPT contextual adaptation shape conversational information-seeking experiences. ArXiv Preprint posted online on September 29, 2025. [doi: 10.48550/arXiv.2509.25513]
37. Anam RK. Prompt engineering and the effectiveness of large language models in enhancing human productivity. ArXiv. Preprint posted online on April 04, 2025 2026 [FREE Full text] [doi: 10.31219/osf.io/ad9y5_v1]
38. Sawalha G, Taj I, Shoufan A. Analyzing student prompts and their effect on ChatGPT's performance. Cogent Educ 2024 Sep 09;11(1). [doi: 10.1080/2331186x.2024.2397200]

XSL•FO
RenderX

39. Zamfirescu-Pereira JD, Wong RY, Hartmann B, Yang Q. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023 Presented at: CHI '23; April 23-28, 2023; Hamburg, Germany URL: https://dl.acm.org/doi/10.1145/3544548.3581388 [doi: 10.1145/3544548.3581388]
40. Gagolewski M. stringi: fast and portable character string processing in R. J Stat Soft 2022;103(2):1-59 [FREE Full text] [doi: 10.18637/jss.v103.i02]
41. Eltorai AE, Ghanian S, Adams CAJ, Born CT, Daniels AH. Readability of patient education materials on the American Association for Surgery of Trauma website. Arch Trauma Res 2014 Apr 30;3(2):e18161 [FREE Full text] [doi: 10.5812/atr.18161] [Medline: 25147778]
42. Ley P, Florio T. The use of readability formulas in health care. Psychol Health Med 1996;1(1):7-28. [doi: 10.1080/13548509608400003]
43. Albright J, de Guzman C, Acebo P, Paiva D, Faulkner M, Swanson J. Readability of patient education materials: implications for clinical practice. Appl Nurs Res 1996 Aug;9(3):139-143. [doi: 10.1016/s0897-1897(96)80254-0] [Medline: 8771859]
44. Badarudeen S, Sabharwal S. Readability of patient education materials from the American Academy of Orthopaedic Surgeons and Pediatric Orthopaedic Society of North America web sites. J Bone Joint Surg Am 2008 Jan;90(1):199-204. [doi: 10.2106/jbjs.g.00347]
45. Sabharwal S, Badarudeen S, Unes Kunju S. Readability of online patient education materials from the AAOS website. Clin Orthop Relat Res 2008 May;466(5):1245-1250 [FREE Full text] [doi: 10.1007/s11999-008-0193-8] [Medline: 18324452]
46. An introduction to health literacy. National Institutes of Health National Library of Medicine. URL: https://www.nnlm.gov/guides/intro-health-literacy [accessed 2025-08-05]
47. Mamedova S, Pawlowski E. Data point: adult literacy in the United States. United States Department of Education. 2019. URL: https://nces.ed.gov/pubs2019/2019179.pdf [accessed 2025-08-05]
48. Weiss BD. Health literacy: a manual for clinicians. American Medical Association. 2003. URL: http://lib.ncfh.org/pdfs/6617.pdf [accessed 2025-09-15]
49. Rooney MK, Santiago G, Perni S, Horowitz DP, McCall AR, Einstein AJ, et al. Readability of patient education materials from high-impact medical journals: a 20-year analysis. J Patient Exp 2021 Mar 03;8:2374373521998847 [FREE Full text] [doi: 10.1177/2374373521998847] [Medline: 34179407]
50. Berelson B. Content Analysis in Communication Research. New York, NY: Free Press; 1952.
51. MAXQDA manual. MAXQDA. URL: https://www.maxqda.com/help/welcome [accessed 2025-12-15]
52. Kraak VI, Holz A, Woods CL, Whitlow AR, Leary N. A content analysis of persuasive appeals used in media campaigns to encourage and discourage sugary beverages and water in the United States. Int J Environ Res Public Health 2023 Jul 13;20(14):6359 [FREE Full text] [doi: 10.3390/ijerph20146359] [Medline: 37510591]
53. Myrick JG. The Role of Emotions in Preventative Health Communication. London, UK: Bloomsbury Academic; 2015.
54. Foerster TA, Chalgren L, Abrams K, Bice C, Singer M. Informing or influencing? A content analysis of wildlife feeding deterrence messages in U.S. national park communications. J Interpret Res 2025 Nov 05;30(2):183-203. [doi: 10.1177/10925872251388948]
55. Nan X, Iles IA, Yang B, Ma Z. Public health messaging during the COVID-19 pandemic and beyond: lessons from communication science. Health Commun 2022 Jan;37(1):1-19. [doi: 10.1080/10410236.2021.1994910] [Medline: 34724838]
56. Goetz JL, Keltner D, Simon-Thomas E. Compassion: an evolutionary analysis and empirical review. Psychol Bull 2010 May;136(3):351-374 [FREE Full text] [doi: 10.1037/a0018807] [Medline: 20438142]
57. Lazarus RS. Hope: an emotion and a vital coping resource against despair. Soc Res 1999;66(2):653-678 [FREE Full text]
58. Lazarus RS. Emotion and Adaptation. Oxford, UK: Oxford University Press; 1994.
59. Casais B, Pereira AC. The prevalence of emotional and rational tone in social advertising appeals. RAUSP Manag J 2021 May 10;56(3):282-294. [doi: 10.1108/rausp-08-2020-0187]
60. Krippendorff K. Agreement and information in the reliability of coding. Commun Methods Meas 2011 Jun 14;5(2):93-112. [doi: 10.1080/19312458.2011.568376]
61. Lombard M, Snyder-Duch J, Bracken CC. Content analysis in mass communication: assessment and reporting of intercoder reliability. Hum Comm Res 2002 Oct;28(4):587-604. [doi: 10.1111/j.1468-2958.2002.tb00826.x]
62. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2022. URL: https://www.R-project.org/ [accessed 2025-09-02]
63. Federal Register: The Daily Journal of the United States Government. URL: https://www.ecfr.gov/current/title-45/part-46/section-46.102 [accessed 2026-02-25]
64. Azaria A, Mitchell T. The internal state of an LLM knows when it's lying. ArXiv Preprint posted online on April 26, 2023 [FREE Full text] [doi: 10.48550/arXiv.2304.13734]
65. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst 2025 Jan 24;43(2):1-55. [doi: 10.1145/3703155]
66. Ovsyannikova D, de Mello VO, Inzlicht M. Third-party evaluators perceive AI as more compassionate than expert humans. Commun Psychol 2025 Jan 10;3(1):4. [doi: 10.1038/s44271-024-00182-6] [Medline: 39794410]

XSL•FO
RenderX

67. Chen D, Chauhan K, Parsa R, Liu ZA, Liu FF, Mak E, et al. Patient perceptions of empathy in physician and artificial intelligence chatbot responses to patient questions about cancer. NPJ Digit Med 2025 May 13;8(1):275 [FREE Full text] [doi: 10.1038/s41746-025-01671-6] [Medline: 40360673]
68. Howcroft A, Bennett-Weston A, Khan A, Griffiths J, Gay S, Howick J. AI chatbots versus human healthcare professionals: a systematic review and meta-analysis of empathy in patient care. Br Med Bull 2025 Sep 22;156(1):ldaf017. [doi: 10.1093/bmb/ldaf017] [Medline: 41115171]
69. Seitz L. Artificial empathy in healthcare chatbots: does it feel authentic? Comput Hum Behav Artif Hum 2024;2(1):100067. [doi: 10.1016/j.chbah.2024.100067]
70. Veazey K, Notebaert A, Robertson EM. Medical student perceptions of establishing effective clinical communication: a qualitative study. Adv Health Sci Educ Theory Pract 2025 Aug 14. [doi: 10.1007/s10459-025-10468-x] [Medline: 40810854]
71. Hinkle LJ, Fettig LP, Carlos WG, Bosslet G. Twelve tips for just in time teaching of communication skills for difficult conversations in the clinical setting. Med Teach 2017 Sep;39(9):920-925 [FREE Full text] [doi: 10.1080/0142159X.2017.1333587] [Medline: 28598711]
72. England JA, Howell M, White BA. Creating a culture of communication in undergraduate medical education. Proc (Bayl Univ Med Cent) 2020 Apr 20;33(3):485-491 [FREE Full text] [doi: 10.1080/08998280.2020.1746156] [Medline: 32676001]
73. Gilligan C, Powell M, Lynagh MC, Ward BM, Lonsdale C, Harvey P, et al. Interventions for improving medical students' interpersonal communication in medical consultations. Cochrane Database Syst Rev 2021 Feb 08;2(2):CD012418 [FREE Full text] [doi: 10.1002/14651858.CD012418.pub2] [Medline: 33559127]
74. Shah SJ, Nair A, Murtagh K, Ma SP, Vogt K, Clutter D, et al. Clinician perspectives on AI-generated drafts of patient test result explanations. JAMA Netw Open 2025 Aug 01;8(8):e2528794 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.28794] [Medline: 40844780]

Abbreviations

AI: artificial intelligence
FKGL: Flesch-Kincaid Grade Level
FRE: Flesch Reading Ease
GenAI: generative artificial intelligence

Edited by Q Jin; submitted 09.Oct.2025; peer-reviewed by C Wardle, K Kim, U Pasupulety; comments to author 14.Dec.2025; revised version received 22.Jan.2026; accepted 11.Feb.2026; published 06.Mar.2026.

Please cite as:
Burns DT, Bice C, Johnson PE, Chia N, Robinson T
Comparison of Emotional Content in Text Responses From Physicians and AI Chatbots to Patient Health Queries: Cross-Sectional Study
J Med Internet Res 2026;28:e85516
URL: https://www.jmir.org/2026/1/e85516
doi:10.2196/85516
PMID:

©Daniel T Burns, Channing Bice, Paul E Johnson, Nicholas Chia, Timothy Robinson. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Young Adults' Interactions With Food and Nutrition Content on Social Media and Implications for Intervention Design: Semistructured Interview Study

Hao Tang[1], MS, MEd; Amy L Ahern[1], PhD; Marie Spreckley[1*], PhD; Andrea D Smith[1*], PhD

Institute of Metabolic Science Epidemiology, University of Cambridge, Cambridge, England, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Hao Tang, MS, MEd
Institute of Metabolic Science Epidemiology
University of Cambridge
Level 4, Institute of Metabolic Science
Addenbrooke's Hospital, Cambridge Biomedical Campus
Cambridge, England, CB2 0QQ
United Kingdom
Phone: 44 1223330315
Email: ht458@cam.ac.uk

Abstract

Background: Young adults increasingly rely on social media for nutrition information. However, little is known about (1) which types of eating-related content they actively engage with and why, and (2) how they interpret, evaluate, and incorporate this content into their everyday food choices and health behaviors.

Objective: This qualitative study explored how UK young adults (aged 18-25 years) interact with food and nutrition content across social media platforms to inform the design of future social media interventions.

Methods: Semistructured online interviews, guided by the Capability, Opportunity, Motivation–Behavior (COM-B) model, were conducted with 25 active social media users (18/25, 72% women, mean age 22.2, SD 1.9 years, ethnically diverse) in the United Kingdom between August and October 2024. The study design was informed by patient and public involvement to ensure relevance and acceptability. Data were analyzed using reflexive thematic analysis. To guide intervention development, key findings (coded as barriers and facilitators) were systematically mapped to the Theoretical Domains Framework, and the COM-B. Ethics approval was obtained from the University of Cambridge (24.368).

Results: Five key themes were identified: (1) evolving engagement patterns (passive scrolling to active interaction and mixed feelings on algorithmic control), (2) conflicted information seeking (frustration with contradictory advice, varied strategies to assess credibility), (3) multifaceted behavioral impact (simultaneous positive impacts such as cooking inspiration and negative impacts such as restrictive eating triggers), (4) shifting goals (a movement from appearance-focused to health-centered goals; yet, vulnerability to body-image issues), and (5) intervention preferences (demand for credible professionals, customizable content, and privacy protection). Participants demonstrated a reactive learning process, developing "digital nutrition literacy" often after negative experiences. Social influences were identified as the most frequently cited domain (mapped to TDF [theoretical domains framework]/COM-B) shaping interactions with social media content.

Conclusions: This study challenges assumptions of passive social media consumption, showing that young adults actively develop protective strategies yet remain vulnerable to misinformation. Digital interventions should leverage user agency and address diverse perceptions through customizable, credible content delivered with privacy and emotionally safe messaging. The COM-B and TDF mapping provide specific, evidence-based behavioral targets, particularly within the domain of Social Opportunity and Reflective Motivation, to guide the development of effective eHealth interventions.

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2023-083465

(J Med Internet Res 2026;28:e89344) doi:10.2196/89344

XSL•FO
RenderX

KEYWORDS

social media; young adults; qualitative; eating behavior; digital intervention; thematic analysis

Introduction

Young adults (18-25 years) are immersed in a digital environment where social media platforms are a dominant presence. Social media use not only shapes their social interactions but also influences their eating behaviors and dietary choices [1-3]. As of 2025, there are more than 5.6 billion social media users, and nearly all (99%) internet users aged 16 years and older in the United Kingdom use platforms such as Instagram (Meta), YouTube (Google), and TikTok (ByteDance) [4,5]. These platforms are prominent and widely accessible sources of nutrition content, offering food-related material including influencer endorsements, recipe demonstrations, health challenges, and dieting trends [6,7].

The net influence of social media on dietary behaviors is contingent on multiple interacting factors, individual predispositions, and context. On the positive side, social media can serve as a space for peer support and positive dietary role modeling, enhancing access to practical nutrition information [7-9]. Health-promoting content was found to improve nutrition literacy and encourage the adoption of healthy behaviors, particularly when delivered by relatable peers [10,11]. Young adults also report that social media facilitates recipe discovery, meal-planning inspiration, and connection with online communities focused on wellness [12-14]. On the other hand, emerging research highlights that social media users are exposed to harmful dietary misinformation [15], including advice promoting restrictive dieting and elimination of food groups, and weight stigma, which may exacerbate body dissatisfaction and disordered eating behaviors [3,5,16].

Research also indicates that frequent social media use is associated with unhealthy eating behaviors, including increased intake of sugar-sweetened beverages (SSBs) and foods high in fat, sugar and salt (HFSS), lower fruit and vegetable consumption, and skipping meals [17,18]. Increasing exposure to online food content can both promote healthier eating and reinforce unhealthy food norms; yet, the balance of these effects remains unclear.

Despite growing interest in the relationship between social media and diet and eating behaviors, there remains a limited understanding of how young people actively engage with and interpret eating-related content online, and how such interactions can be leveraged to inform the development of effective digital interventions to support health promotion. Three critical gaps constrain efforts to address this public health challenge. First, food and nutrition content is among the most widely viewed topics on social media, with hashtags such as #WhatIEatInADay, #WeightLossFood, and #Recipes collectively generating billions of views across platforms. Yet most research relies on broad exposure measures and overlooks the nuanced ways in which young adults engage (eg, saving recipes, joining challenges, or curating creators on their feeds), underscoring the need for more behavior-specific investigation [19,20]. Second, while recent studies have linked social media use to both healthy and disordered eating behaviors, they offer limited insight into why these effects differ between individuals [3,21,22]. Previous reviews indicate that many rely on quantitative, cross-sectional designs using broad, decontextualized measures of social media use and often treat social media users as passive consumers without exploring the individualized nature of their engagement [3,23]. These limitations are particularly relevant during young adulthood, a formative period characterized by growing dietary autonomy, health awareness, and identity development [24-26]. Therefore, qualitative insights are needed to capture the mechanisms underlying how and why young adults engage with, interpret, and respond to nutrition content over time. Third, while behavior-change models, such as the Capability, Opportunity, Motivation–Behavior (COM-B) model, are widely recommended in health behavior research, their application in social media–based interventions remains limited [27-29]. One review noted that only 8 of 33 digital health interventions used theoretical frameworks [30].

To address these gaps, we conducted a theory-informed qualitative study, mapped to COM-B and TDF (theoretical domains framework), to explore how young adults interact with diet- and eating-related social media content and how these interactions shape behavior. Grounded in lived experience, the study identifies specific behavioral targets and opportunities for health promotion that extend beyond awareness-raising within today's rapidly changing social media context.

Methods

Study Design

This qualitative study uses semistructured interviews to explore how young adults living in the United Kingdom engage with eating behavior-related content on social media. The interview guide was structured around the components of the COM-B model to provide a systematic framework for analyzing behavioral influences on dietary choices in digital environments [28]. The COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist was used (Multimedia Appendix 1) [31].

Patient and Public Involvement

Patient and public involvement (PPI) informed multiple aspects of this study, with PPI sessions conducted between March and May 2024. Four online sessions were held: 2 with members of the target population (UK-based social media users aged 18-25 years), 1 with a body-positivity social media influencer, and 1 with a staff member of a UK-based eating disorder charity. PPI feedback shaped the recruitment strategy. For example, participants recommended focusing on social media as the primary recruitment channel, prioritizing recruitment via Instagram posts and Facebook (Meta) groups and suggested additional relevant charities to contact. They also raised concerns about bot activity online, which led to the inclusion of a bot-filter question in the eligibility screening survey.

XSL•FO
RenderX

PPI members also contributed to refining the interview protocol, particularly for questions addressing potentially distressing or sensitive experiences. As a result, several questions were reworded for clarity and sensitivity. PPI contributors reviewed all participant-facing materials, including the interview guide, recruitment flyers, and the participant information sheet, and provided feedback on the clarity, tone, and cultural inclusivity of language used throughout. For example, a preliminary question was added asking participants to define what healthy eating meant to them personally before proceeding to questions about social media use, ensuring questions were grounded in participants' own frameworks. Additionally, on the advice of PPI contributors, a sensitivity disclaimer was inserted before questions addressing disordered eating risk, informing participants that the following questions may be distressing and that they were free to skip or pause at any point.

Participants and Recruitment

Participants were eligible if they were aged 18-25 years, lived in the United Kingdom, spoke English, and used social media (Instagram, TikTok, YouTube, etc) most days of the week. Recruitment was conducted from August to October 2024 through a multipronged approach, including posting on social media platforms (Instagram, Reddit, and Facebook groups) and posters in public locations. Interested individuals completed an online screening survey via REDCap (Research Electronic Data Capture; Vanderbilt University) to assess eligibility. From the pool of eligible respondents who provided authentic answers, participants were purposively selected to reflect diversity across gender, ethnicity, education level, and employment status. If individuals declined participation, replacements were drawn from the same demographic subgroup to maintain diversity.

The study sample size was informed by the concept of information power [32]. With the focused research aim, use of an established theoretical framework (the COM-B framework), and a demographically diverse sampling strategy, we aimed to recruit approximately 30 participants. Throughout data collection, we monitored the richness and relevance of interview material in relation to the study aim. We concluded the interviews when we felt that we had enough detailed interviews from diverse perspectives to provide rich insights into the research question, and had developed coherent analytic themes that answered the question, with depth and coverage across the components of the COM-B framework. These assessments were made through ongoing discussion between the lead researcher (HT) and the supervisory team (ALA, ADS, and MS), and data collection concluded at 25 interviews when these criteria were collectively satisfied.

Data Collection

Semistructured interviews were conducted online via Microsoft Teams by a PhD researcher (HT) and were audio-recorded. Each interview lasted approximately 60 minutes. Field notes were made immediately after each interview to capture contextual information and initial reflections.

Researcher Reflexivity Statement

The interviewer (HT) is a female PhD researcher in behavioral nutrition with training and prior experience in conducting qualitative interviews. Her background in nutrition and dietetics, and experience creating evidence-based content on social media, helped her understand participants' experiences but also required awareness of potential assumptions about "healthy" content and online engagement. As someone with lived experience of navigating food and body image in the context of social media, HT was reflexively aware of potential shared experiences with participants and took care to remain open to perspectives that might differ from her own experiences. As a young adult and active social media user, HT shared generational proximity with participants, which may have facilitated rapport while also requiring reflexive awareness of shared assumptions. The power dynamics inherent in the researcher-participant relationship were acknowledged, particularly given the sensitive nature of topics such as body image and eating behaviors. HT kept brief reflexive notes after each interview and met regularly with supervisors (ALA and AS) to discuss any questions or uncertainties that arose during data collection.

Interview Guide and Measures

The interview guide (Multimedia Appendix 2) covered topics including social media use patterns, experiences with healthy eating content, perceived influences on eating behaviors, platform preferences and engagement factors, and barriers and facilitators to healthy eating.

Data Management

Audio recordings were transcribed and anonymized. All data were stored securely on the MRC (Medical Research Council) Epidemiology Unit Secure Research Drive (SRD), with access restricted to authorized members of the research team. Participants and their transcripts were assigned unique numerical IDs to ensure confidentiality.

Data Analysis

Phase 1: Thematic Analysis

Data were analyzed using reflexive thematic analysis following Braun and Clarke's six-phase approach [33], selected for its capacity to generate nuanced, participant-centered insights. HT led the analysis, using inductive coding of all transcripts in NVivo, whereby codes were generated in response to participants' accounts without the imposition of a predetermined framework, and analytic themes were developed through an iterative process. To support analytic reflexivity and enrich interpretation, a subset of transcripts (5/25, 20%) was also coded by MS. These additional coding contributions informed collaborative discussions about alternative readings and identified patterns, contributing to the refinement of the thematic structure. The research team met regularly to discuss coding decisions and evolving themes, which supported reflexive engagement with the data and strengthened the coherence of the final analysis.

In this study, key concepts such as "healthy eating" and "credible content" were not defined a priori for participants during interviews. Instead, these were explored as open, participant-driven constructs, reflecting how individuals themselves interpreted and evaluated nutrition-related content on social media. For analytic consistency during subsequent

XSL•FO
RenderX

COM-B and TDF mapping, references to "healthy eating" were considered in relation to broadly accepted UK public health dietary guidance [34], while remaining grounded in participants' expressed meanings. "Credible content" was analytically defined as nutrition-related information that participants perceived as trustworthy, typically based on cues such as professional credentials, scientific references, tone, presentation style, or community validation.

Phase 2: Behavioral Framework Mapping

Identified determinants of young adults' social media use were subsequently mapped to the TDF [35] and the COM-B model [28]. The COM-B model explains that for behavior change to occur, people must have the capability to change, the motivation to change and the opportunity to change. The TDF build on this by outlining 14 domains that capture the wider range of individual barriers and facilitators that may influence behavior. These domains systematically map to the COM-B model components (Multimedia Appendix 3). TDF and COM-B were selected because they support the translation of qualitative insights into behavioral domains and validated behavioral change mechanisms, which provide a strong foundation for future intervention development.

Each subtheme derived from the qualitative analysis was coded as either a barrier or a facilitator, or both where applicable, and subsequently categorized into the relevant TDF and COM-B domains. Influences could act in both directions depending on context and were coded as barrier/facilitator. Mapping barriers and facilitators to the TDF and COM-B allowed us to systematically identify which psychological, social, and contextual domains were most relevant to healthy eating within the social media context. Framework mapping was led by HT, with ADS providing an additional interpretative perspective. Interpretations were discussed iteratively to refine domain allocation and enhance conceptual clarity.

Ethical Considerations

The study received ethical approval from the University of Cambridge School of Humanities and Social Sciences Research Ethics Committee (reference number 24.368). Prior to each interview, all participants provided informed consent electronically via a digital consent form. Participants were assigned unique numerical identifiers, and pseudonymized transcripts and personal data were stored separately and securely on the MRC Epidemiology Unit Secure Research Drive, accessible only to authorized members of the research team. All interview participants received a £25 (US $32) digital voucher as remuneration for their time, and all PPI contributors each received a £25 (US $32) digital voucher for their 1-hour session.

Results

Sample Characteristics

Of the 317 individuals who completed the initial online recruitment questionnaire, 277 met the eligibility criteria. A total of 133 responded to follow-up questions, and 30 individuals were scheduled for interviews. Among these, 27 participants joined the interview process. Two participants were excluded from the analytic sample: one whose responses did not sufficiently address the research questions concerning food and nutrition content on social media, and one who used racially offensive language that was inconsistent with the study's ethical standards and respectful participation guidelines, resulting in a final analytic sample of 25 participants. Sufficient information power to answer our research question was achieved at 25 interviews. An overview of participant characteristics is provided in Table 1.

Participants had a mean age of 22.2 (SD 1.9) years and were predominantly women (18/25, 72%). The sample was ethnically diverse, comprising White (9/25, 36%), Asian/Asian British (8/25, 32%), Black/African/Caribbean/Black British (6/25, 24%), mixed/multiple ethnic groups (1/25, 4%), and other ethnic groups (1/25, 4%). Educational attainment varied, but most participants held a bachelor's degree or equivalent (11/25, 44%) or had completed, or were enrolled in, postgraduate education (7/25, 28%).

In terms of occupation, 60% (15/25) were students, 16% (4/25) were employed full-time, 16% (4/25) were employed part-time, and 8% (2/25) were engaged in apprenticeships or traineeships.

All participants reported daily social media use, typically around 2-3 hours per day. Instagram (Meta) and TikTok (ByteDance) were the most frequently used platforms for accessing health- and nutrition-related content.

XSL•FO
RenderX

Table 1. Demographic characteristics of participants (N=25).[a]

Characteristics	Values, n (%)
Age (years), mean (SD)	22.2 (1.9)
Gender	
Woman	18 (72)
Man	7(28)
Ethnicity	
Asian/Asian British	8 (32)
Black/African/Caribbean/Black British	6 (24)
White	9 (36)
Mixed/multiple ethnic groups	1 (4)
Other ethnic group	1 (4)
Occupational status	
Student	15 (60)
Employed full-time	4 (16)
Employed part-time	4 (16)
Apprenticeship or traineeship	2 (8)
Education level	
Postgraduate degree	7 (28)
Bachelor's degree or equivalent	11 (44)
Post-secondary diploma or certificate	6 (24)

[a]This table provides a detailed summary of demographic, educational, and occupational characteristics of the study sample. Continuous variables are presented as mean (SD), and categorical variables are presented as counts and percentages.

Overview of Themes

Overall, five themes were identified, describing the diverse ways in which young adults engage with social media and how this influences their eating behaviors. Themes and subthemes are illustrated in Figure 1 and described in detail below.

XSL•FO
RenderX

Figure 1. Thematic framework of young adults' interactions with eating behaviour-related social media content.

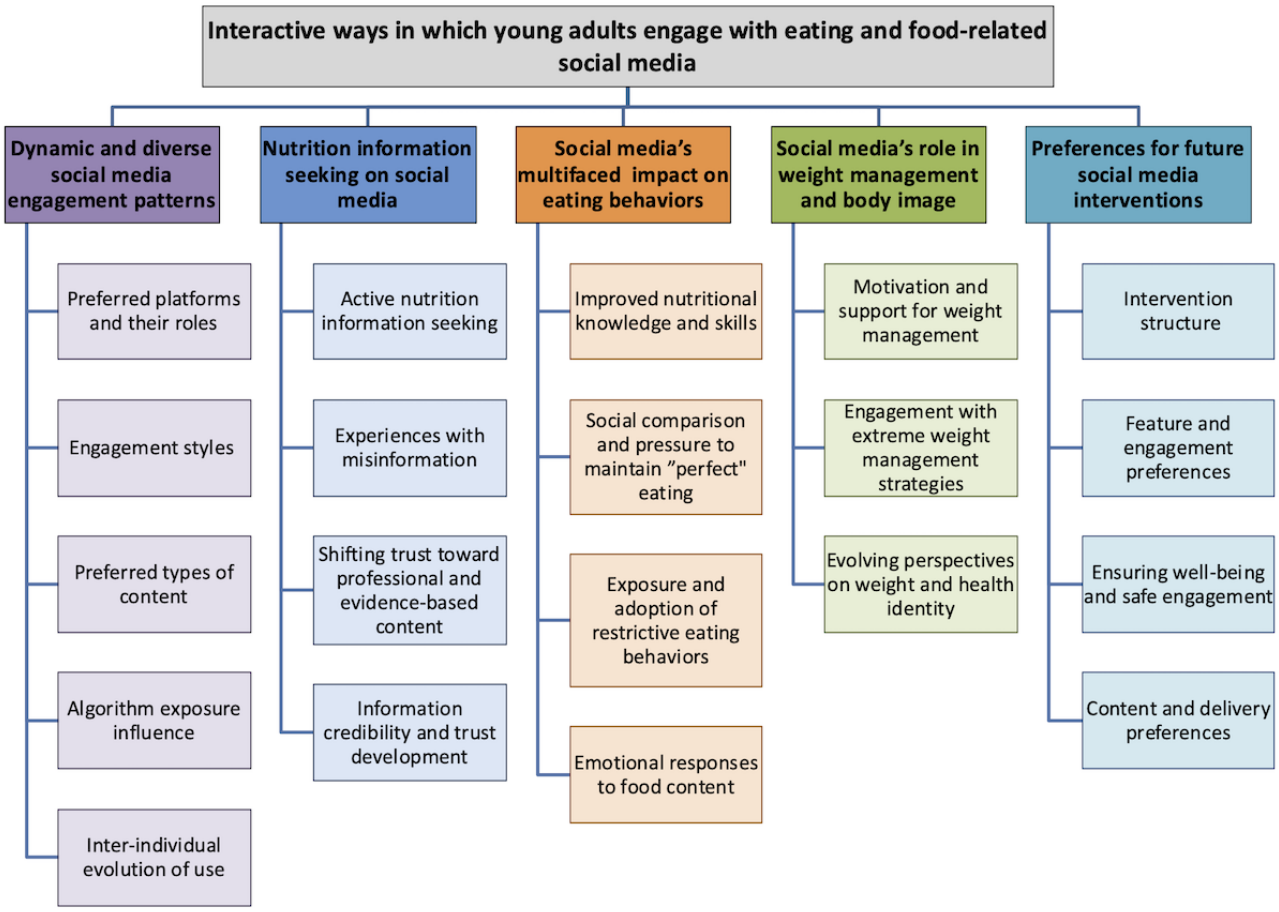

Theme 1: Evolving and Diverse Social Media Engagement Patterns

Participants described a range of ways in which they engaged with nutrition and dietary content on social media, emphasizing that their experiences were both highly personal and platform-specific. Engagement patterns were dynamic, evolving over time as participants developed preferences, adapted to changing algorithms, or sought to manage the emotional impact of content.

Platforms were associated with distinct benefits: Instagram was valued for community and peer support, YouTube for providing in-depth, educational material, and TikTok for engaging in short-form content. TikTok was also described by some as addictive or overwhelming. Engagement ranged from passive scrolling to active participation, such as commenting, voting in polls, or joining live streams.

Content appeal was strongly linked to visual engagement and relatability, such as the "What I eat in a day." This video features creators, often celebrities or influencers, documenting their daily food and drink intake in a 24-hour time window. Participants perceived these videos as accessible and relatable because of their straightforward structure and perceived authenticity.

> *I watch a lot of food videos and get to know about different cultures. And it helped me realise the importance of adding vegetables in my diet…it helped me a lot. That was also where I learned to drink enough water instead of soft drinks. [HT prompted for example] I followed a marathon runner who's vegan. I'm not attracted to veganism at all, but I liked her ideas of adding more vegetables in cooking. I remember watching her post about a chickpea sandwich, I know chickpeas are a good source of protein. It made me think it would be nice to eat less meat and try different food. She's a well-known creator, so I trusted her, and I've actually made her recipes before.* [Woman, 20 years old]

However, not all experiences with social media content were positive. Some participants expressed frustration at the lack of control over what appeared on their feeds.

> *I watched a few fitness videos on TikTok last week and now my feed is just... lots of workout content and weight loss videos. It makes me feel kind of pressured to work out more, you know? …Sometimes I feel like I have no control over what I see; the algorithm decides for me.* [Woman, 20 years old]

Over time, some participants reported making deliberate adjustments to their social media use, such as unfollowing accounts, deleting apps, or reducing screen time, reflecting an active effort to manage their well-being.

XSL•FO
RenderX

> *I started to cut down my Instagram use because it made me feel uneasy about my own body.* [Woman, 24 years old]

Theme 2: Nutrition Information Seeking on Social Media

Participants used social media as a primary source for quick nutrition information (eg, recipes and meal planning). However, utility was tempered by the prevalence of misinformation and conflicting dietary advice. Participants described adopting a number of strategies to verify information, such as checking credentials, cross-referencing sources, and reviewing comment sections, reflecting a developing "digital nutrition literacy."

> *I often search on Xiaohongshu (also known internationally as RedNote) when I don't know what to cook.* [Man, 23 years old]

Nevertheless, confusion remained high due to contradictory health claims online.

> *Too much conflicting information makes it hard to decide what's right. Like, I saw posts saying that eating fruit is not healthy and doesn't help digestion, and I saw other posts saying that drinking water immediately after eating is not healthy either. But also know from other sources that drinking water while eating is good for weight loss and fruit helps with digestion.* [Woman, 24 years old]

These were sometimes perceived as being driven by commercial interests rather than science.

> *You don't know—some experts will say this is healthy, then a few years later, people change their minds. Like coffee—before it was bad, now they say it's good for your heart. It's confusing and feels like part of advertising.* [Woman, 23 years old]

Trust was influenced by both professional credentials and superficial cues such as the presenter's appearance or the level of social validation (eg, positive comments) from the community.

> *Sometimes, when the influencer... looks healthy and positive, that helps [with credibility]. If they look really fit, I want to try their recommended snacks or diet…. I guess I think if they look that way, then what they're recommending must work.* [Woman, 25 years old]

> *I look at the comments to see what other people have said. If most people say they got good results, then I'm more likely to trust what they (the creator/influencer) recommend.* [Woman, 19 years old]

Theme 3: Social Media's Multifaceted Impact on Eating Behaviors

The influence of social media was described as simultaneously positive and negative. Positive effects included skill development (eg, cooking techniques), improved knowledge, and inspiration for balanced meal choices.

> *I started cooking at home more after seeing easy and tasty recipes online.* [Woman, 23 years old]

Conversely, exposure to curated and unrealistic dietary and body ideals often led to social comparison, self-doubt, and the adoption of harmful behaviors, including restrictive eating, obsessive weighing, and calorie counting.

> *Seeing influencers made me feel inspired but sometimes made me question myself because they make diets look so easy.* [Woman, 24 years old]

> *When I just entered college, I gained a lot of weight and wanted to lose it. I found this influencer whose 'what I eat in a day' and weight loss tips videos I really enjoyed watching. She is a famous model. So I started trying the weight loss tips she recommended. After almost two months of skipping dinner due to posts promoting it as healthy, I developed eczema and felt horrible physically and mentally. That's when I realised that these 'model' diets were very risky. So I stopped and unfollowed many accounts.* [Woman, 25 years old]

Reactions to specific content types were subjective. For example, mukbang videos (online broadcasts of individuals eating large quantities of food) elicited opposing responses: some participants felt they reduced cravings, while others found them uncomfortable or even disturbing.

Theme 4: Social Media's Role in Weight Management and Body Image

Social media played a shifting role in shaping participants' body image concerns. Positive influences included motivation from success stories and support from health-focused communities.

> *I've been overweight since I was a child, and I started to feel really ashamed from my teenage years, like in middle school. I just couldn't get fit. I thought I was just lazy and born to be fat. Seeing success stories of people with obesity inspired me to start eating better.* [Man, 25 years old]

However, many participants recounted past experiences of body dissatisfaction exacerbated by unrealistic beauty standards promoted online. Over time, some reported deliberate progression from appearance-focused goals in their younger years to more balanced, health-centered approaches.

> *I think for me it used to affect it negatively… I saw models and pop stars… I wanted to be like that… I tried loads of diets like keto, detox, all kinds of things. Often I'd feel awful about myself, like, why can't I just be like them? There was also a lot of talk on social media saying these foods are bad for you, and because of that you have to punish yourself with workouts. It was a lot more unhealthy, and because of that, I think the impact was negative. Now I take from social media... making sure I'm eating enough. I think for me it used to affect it negatively, but I don't think I realized that and that was when I was younger.* [Woman, 25 years old]

XSL•FO
RenderX

This shift demonstrates an evolving self-regulatory process aimed at mitigating the negative psychological effects of weight-centric content.

Theme 5: Preferences for Future Social Media Interventions

Participants expressed a clear and diverse preference for future digital interventions, emphasizing that interventions should be short, visually engaging, and emotionally safe. Key preferences included short videos (30-60 seconds), involvement of professionals (eg, a qualified dietician) to build trust, interactive features (live Q&A sessions and journaling), and a privacy-protective environment (eg, anonymous chat).

I think somethinganonymous and like no judgments could be useful. I think sharing eating or comment about eating can be sensitive. [Woman, 24 years old]

> *Socializing within the programme can be really helpful to keep people tied because they can explore other personalities, or they can enjoy their time doing the programme.* [Woman, 22 years old]

Most participants preferred a time-limited intervention (2-4 weeks) with regular updates and structured, measurable goals to maintain motivation and prevent overwhelm.

Mapping Identified Determinants of Eating and Food-Related Social Media Behaviors to the TDF

To inform the design of future interventions, key subthemes identified from the qualitative data were systematically mapped onto the TDF and subsequently COM-B. Each subtheme was evaluated and categorized as either a barrier or a facilitator to the use of social media interventions to promote healthy eating. A total of 106 barriers and facilitators were mapped across the dataset, with the largest proportion aligned with Motivation, followed by Opportunity and Capability.

Within the TDF, Social Influences (Opportunity) was the most frequently coded domain, reflecting the pervasive role of peers, influencers and community feedback in shaping young adults' interactions with content, perceived credibility, and subsequent eating behaviors (peer influence leading to social comparison and community support for accountability). Barriers were concentrated in domains such as Environmental Context and Resources (Opportunity; eg, prevalence of misinformation and algorithm control) and Emotion (Motivation; eg, guilt and feeling triggered). Facilitators are frequently mapped to Knowledge (Capability) and Skills (Capability; eg, recipe demonstrations and critical evaluation skills) and Goals (Motivation; eg, shift toward sustainable health goals). These findings highlight that digital interventions need to target not only individual knowledge and skills but also the social and environmental context that dictates exposure and emotional response. Table 2 provides a summary of all barriers and facilitators alongside their theoretical domains. A comprehensive item-by-item mapping of all subthemes and barrier/facilitator pairs (n=106) is provided in Multimedia Appendix 4.

XSL•FO
RenderX

Table 2. Summary of key themes, barriers and facilitators to healthy eating when engaging with food and nutrition content on social media, mapped to Theoretical Domains Framework and Capability, Opportunity, Motivation–Behavior.

Themes[a]	Key findings	Primary barriers	Primary facilitators	TDF[b] Domain (COM-B[c])
Theme 1: Evolving engagement patterns	• Participants used different platforms distinctly • Algorithms shaped exposure positively and negatively	• Algorithm-driven control • Limited user autonomy over content exposure	• Algorithm-driven tailoring. • Platform affordances for engagement. • Peer influence motivation.	Environmental Context & Resources (O); Beliefs about Capabilities (M)
Theme 2: Nutrition information seeking	• High engagement but confusion from contradictory advice. • Shift toward professional credibility	• Misinformation prevalence. • Difficulty evaluating contradictory claims	• Critical-evaluation skill development. • Preference for evidence-based sources. • Professional credentials.	Environmental Context & Resources (O); Skills (C); Knowledge (C)
Theme 3: Multifaceted behavioral impact	• Positive (skill development) and negative (social comparison and restrictive eating) impacts coexist	• Unrealistic ideals • Normalized restrictive eating • Emotional triggers from comparison	• Practical guidance. • Knowledge access. • Diverse content formats. • Emotional support.	• Social Influences (O) • Emotion (M) • Skills (C) • Knowledge (C)
Theme 4: Weight management and body image	• Evolving shift from appearance-focused to health-centered goals. • Persistent vulnerability to ideals	• Extreme dieting promotion. • Unrealistic body ideals. • Pressure to conform to thinness	• Success stories. • Community support. • Skill development through shared techniques. • Critical evaluation skills.	• Social Influences (O) • Beliefs about Consequences (M) • Skills (C)
Theme 5: Intervention preferences	• Strong demand for credible, customizable, emotionally safe content delivered in short formats	• Algorithm opacity • Content volume overwhelm	• Short-form visuals (30-60 s). • Professional delivery. • Interactive features. • Privacy protection. • Time-limited structure.	• Environmental Context and Resources (O) • Knowledge (C) • Emotion (M) • Behavioral Regulation (C)

[a]This table summarizes the five key themes identified through reflexive thematic analysis, with primary barriers and facilitators identified within each theme and mapped to their corresponding Theoretical Domains Framework domains and Capability, Opportunity, Motivation–Behavior components. Subthemes within each theme have been condensed to show the most salient findings.

[b]TDF: Theoretical Domains Framework.

[c]COM-B: Capability, Opportunity, Motivation–Behavior.

Discussion

Principal Findings

This study highlights the complex and dynamic role of social media in shaping young adults' eating behaviors, revealing a duality where platforms simultaneously facilitate healthy habits (eg, recipe discovery and positive motivation) and expose users to potentially harmful content (eg, misinformation, restrictive eating). Our findings, captured across five themes, detail the mechanisms of this influence. The COM-B/TDF mapping further identified that most barriers and facilitators fell under Motivation and Opportunity, underscoring that effective interventions must target Social Influences (a key Opportunity domain) and Reflective Motivation (eg, goals and beliefs about consequences), in addition to enhancing individual Capability. This highlights key leverage points for the design of future interventions.

Comparison With Existing Literature

Consistent with previous literature, we found that social media can increase body image concerns and disordered eating risk, particularly among young women [3,36,37], while also supporting health-promoting behaviors and knowledge acquisition when used thoughtfully [38,39]. This study builds on these insights by emphasizing the importance young users place on credibility, a holistic approach, a nonjudgmental tone, and emotionally supportive environments.

This duality was clearly reflected in participants' mixed experiences with evolving social media content, reflecting broader concerns about the overwhelming volume and variable quality of digital health information, which can make it difficult for users to distinguish trustworthy from misleading content [40,41]. This underscores the importance of equipping users with targeted digital health literacy skills and tailoring digital support programs to address the confusion arising from the volume and vagueness of online content [41-43].

XSL•FO
RenderX

This study advances current understanding in five ways. First, the study identified a dynamic behavioral pattern in how young adults engage with social media. Participants described a shift from appearance-related content in their earlier years (adolescence) to more health-focused, carefully evaluated use as they aged, especially after the transition to adulthood. This mirrors recent trends in UK youth, who have become more thoughtful and selective in their digital habits over time [23,44]. By capturing users' reflective accounts and behavioral transitions, our study adds novel, person-centered insights with direct implications for tailoring digital interventions to users' developmental stages rather than applying uniform approaches across all young adults.

Second, participants demonstrated sophisticated navigation strategies involving credential checking, community-feedback assessment, cross-referencing sources, and appearance-based judgments to evaluate nutrition information. These findings echo recent reviews calling for recognition of the varied and active ways young users engage with health information online, rather than characterizing them as passive consumers [45-47]. Liu and Zhang (2024) [48] observed similar deliberate credibility judgements in video-based content, while Kuutila et al (2023) [49] found that source expertise and prior belief alignment were stronger predictors of perceived credibility than evidence quality. While some critical evaluation skills were demonstrated (eg, credential checking and cross-referencing), our analysis also underscored the need for interventions that strengthen digital nutrition literacy by addressing the difficulty in evaluating contradictory claims.

Third, this study offers more nuanced insights into how young adults balance credibility and relatability when evaluating nutrition-related content online. While participants increasingly valued professional credentials and evidence-based sources, they remained highly engaged with content created by nonprofessional figures, such as influencers, models, and celebrities, including "what I eat in a day" videos or personal weight loss journeys, suggesting that digital health interventions may be more engaging when they intentionally combine both elements. Moreover, reactions to specific formats, such as weight-loss comparison and mukbang videos, varied widely. While some viewers described feeling comfort or inspiration, others felt guilt or discouragement. These findings echo recent studies [50-52], highlighting the heterogeneity in user responses and the need for personalized, context-sensitive intervention strategies.

Fourth, findings reveal a key tension in digital health engagement that challenges current intervention approaches. Participants appreciated algorithmic curation for efficiently surfacing relevant content but also expressed discomfort with its control over their exposure. Similarly, Milton et al [53] characterized the TikTok "For You" page as a "runaway train," which is highly engaging but sometimes overwhelming and difficult to escape. Additionally, Savolainen and Ruckenstein [54] explored how users navigate algorithmic systems with varying degrees of autonomy, suggesting that while personalization can aid discovery, it often conflicts with self-directed control. These insights highlight a design challenge for digital health interventions, indicating that leveraging platform features for engagement should be balanced with mechanisms that preserve user autonomy and transparency.

Finally, this study highlights how people interact with different social media platforms in different ways. We found that in the context of dietary content, Instagram primarily fulfils community-building functions, YouTube mainly serves as a space for educational and in-depth content, and TikTok facilitates highly engaging, short-form delivery that can be overwhelming. These findings align with research suggesting that different platforms offer distinct affordances [55] and extend earlier qualitative work showing that young adults interact with healthy content differently across platforms [23]. This ecological perspective implies that effective digital health interventions should respect and leverage each platform's unique characteristics rather than applying uniform approaches across all platforms.

Insights for Social Media Intervention Design

This study offers several practical insights for the development of user-centered social media interventions to support healthy eating. A key finding was the high variability in content preferences, engagement styles, and emotional responses to dietary content, including mukbang videos and weight loss promotions. This highlights the need for customization options in future interventions. Such options may include adjustable content formats (eg, infographics and videos), frequency of posts, communication methods (eg, anonymous group chats, comments, and reposting), and types of health messaging (HAO Tang et al, unpublished data, August 2025)

Participants also expressed a preference for shorter content and multicomponent interventions, combining education with motivation and interaction. Realistic, relatable content delivered in emotionally safe tones (ie, nonjudgmental, supportive language that avoids triggering negative emotions or disordered eating behaviors) was consistently valued. Relatability was enhanced when content reflected the lived experiences of diverse individuals, rather than idealized or overly polished representations.

Our findings also identified a critical gap between information exposure and behavioral change. While social media served as a source of motivation and ideas, sustaining healthy habits remained challenging for many participants [56-58]. This highlights the importance of bridging the intention-behavior gap through established behavior-change techniques [59]. Our COM-B mapping results suggest that, while some young adults may possess adequate nutrition knowledge (psychological capability), more tailored strategies are needed to enhance Reflective Motivation and Social Opportunity for sustained behavior change. Specifically, interventions could incorporate goal-setting and self-monitoring (motivation/behavioral regulation) and use algorithm transparency tools to give users greater control over the content they are exposed to (physical opportunity). Social influence mechanisms may be targeted through leveraging credibility cues such as presenter credentials and community validation (eg, community moderation features and community comments).

XSL•FO
RenderX

The confusion participants experienced when navigating contradictory health messages points to a critical design requirement: interventions must equip users with frameworks or skills for evaluating conflicting advice rather than simply adding to the information volume [42,60]. This suggests incorporating features such as source-credibility checklists, decision-making tools that help users weigh competing claims, and educational modules that teach users to recognize common patterns in misleading nutrition content. Furthermore, many participants suggested that peer sharing, algorithmic visibility, and familiar platform features must be balanced with privacy considerations. The flexibility to choose levels of interaction (eg, passive viewing versus active commenting or posting) may help accommodate diverse comfort levels.

Strengths and Limitations

This study's strengths include the diverse sample achieved through purposive sampling, enhancing transferability within the UK young adult population. Comprehensive PPI throughout the research process strengthened the study's cultural sensitivity and appropriateness, and applying a qualitative approach captured nuanced experiences otherwise not measurable using more quantitative approaches, essential for user-centered intervention design. The integration of COM-B and TDF provided both theoretical rigor and practical relevance, which paves a clear pathway for intervention development. This, coupled with the direct exploration of participant preferences for future interventions (Theme 5), offers practical guidance often missing from social media health research.

Some limitations should be noted. As participation was voluntary, the sample may be subject to selection bias, with perspectives potentially skewed toward those with a stronger interest or engagement in health-related topics or social media. Findings regarding specific platforms or content types need to be interpreted with caution, as the rapid evolution of social media technologies and user behaviors may limit their long-term applicability.

Conclusion

This study provides new insights into how young adults in the United Kingdom interact with nutrition-related content on social media. Participants described a dynamic process of engagement shaped by evolving values, critical evaluation skills, and emotional responses. While social media can support healthy eating by offering accessible and motivational content, it also exposes users to conflicting messages and potentially harmful imagery and advice.

Findings highlight the importance of moving toward personalized digital health interventions that reflect users' diverse needs, developmental stages, and evaluation capabilities. To increase the likelihood of achieving positive behavior change, interventions should combine credible, emotionally safe content with customizable features, privacy protection, professional oversight, and be grounded in established behavior change theory (COM-B/TDF). The tension between algorithmic curation and user autonomy suggests that successful interventions need to balance engagement optimization with user control and leverage multiplatform strategies that respect the unique and evolving features of different digital environments.

Acknowledgments

The authors thank the study participants and patient and public involvement contributors for their time and input and gratefully acknowledge Dr Esther van Sluijs for her intellectual contributions through discussions and feedback during research meetings. ChatGPT (OpenAI) was used for grammar checking and sentence refinement. To enhance data privacy, chat history and model training were turned off. The tool was not used for data analysis, interpretation, conceptual development, or content generation, and all text was critically reviewed and approved by the authors.

Funding

This research was supported by the Medical Research Council (Grants MC_UU_00006/5, MC_UU_00006/6) and National Institute for Health and Care Research (NIHR) Cambridge Biomedical Research Centre (NIHR203312). The funder had no role in the study design, data collection, analysis, interpretation, or manuscript preparation.

Data Availability

The terms of participant consent allow us to share deidentified qualitative data generated and analyzed during this study with researchers who are undertaking analyses related to the original objectives, with institutional approval. Requests should be directed to datasharing@mrc-epid.cam.ac.uk.

Authors' Contributions

Conceptualization, data curation, formal analysis, investigation, methodology, project administration, visualization, writing (original draft), writing (review and editing): HT
Conceptualization, funding acquisition, methodology, supervision, writing (review and editing): ALA
Formal analysis (secondary coding of a subset of transcripts, n=5), methodology, and writing (review and editing): MS
Conceptualization, formal analysis (framework mapping), methodology, supervision, and writing (review and editing): ADS

XSL•FO
RenderX

Conflicts of Interest

ALA is on the scientific advisory board for Weight Watchers (WW) and has undertaken consultancy (all payments to her institution) for Menwell Limited.

Multimedia Appendix 1
COREQ 32-item checklist.
[DOCX File , 19 KB - jmir_v28i1e89344_app1.docx]

Multimedia Appendix 2
Semi-structured interview guide used in the qualitative study.
[DOCX File , 19 KB - jmir_v28i1e89344_app2.docx]

Multimedia Appendix 3
Mapping of Theoretical Domains Framework (TDF) domains to COM-B components.
[DOCX File , 16 KB - jmir_v28i1e89344_app3.docx]

Multimedia Appendix 4
Detailed Item-Level Mapping of Barriers and Facilitators Healthy Eating When Engaging with Food and Nutrition Content on Social Media.
[DOCX File , 27 KB - jmir_v28i1e89344_app4.docx]

References

1. Goodyear VA, Wood G, Skinner B, Thompson JL. The effect of social media interventions on physical activity and dietary behaviours in young people and adults: a systematic review. Int J Behav Nutr Phys Act 2021;18(1):72 [FREE Full text] [doi: 10.1186/s12966-021-01138-3] [Medline: 34090469]
2. Ardi Z, Putri SA. The analysis of the social media impact on the millennial generation behavior and social interactions. Southeast Asian Journal of Technology and Science 2021;2(1):70-77 [FREE Full text]
3. Dane A, Bhatia K. The social media diet: a scoping review to investigate the association between social media, body image and eating disorders amongst young people. PLOS Glob Public Health 2023;3(3):e0001091 [FREE Full text] [doi: 10.1371/journal.pgph.0001091] [Medline: 36962983]
4. Number of internet and social media users worldwide as of February. Statista. 2025. URL: https://www.statista.com/statistics/617136/digital-population-worldwide/ [accessed 2026-03-18]
5. Adults' media use and attitudes report. Ofcom. 2025. URL: https://tinyurl.com/r5j94jjr [accessed 2026-03-18]
6. Stirling E, Willcox J, Ong K, Forsyth A. Social media analytics in nutrition research: a rapid review of current usage in investigation of dietary behaviours. Public Health Nutr 2020;24(6):1193-1209. [doi: 10.1017/s1368980020005248]
7. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res 2021;23(5):e17917 [FREE Full text] [doi: 10.2196/17917] [Medline: 33978589]
8. Powell J, Pring T. The impact of social media influencers on health outcomes: systematic review. Soc Sci Med 2024;340:116472 [FREE Full text] [doi: 10.1016/j.socscimed.2023.116472] [Medline: 38070305]
9. Popat A, Tarrant C. Exploring adolescents' perspectives on social media and mental health and well-being - a qualitative literature review. Clin Child Psychol Psychiatry 2023;28(1):323-337 [FREE Full text] [doi: 10.1177/13591045221092884] [Medline: 35670473]
10. Zimmermann E, Tomczyk S. The ways of using social media for health promotion among adolescents: qualitative interview and focus group study. J Med Internet Res 2025;27:e71510 [FREE Full text] [doi: 10.2196/71510] [Medline: 40489771]
11. Evans W, Andrade E, Pratt M, Mottern A, Chavez S, Calzetta-Raymond A, et al. Peer-to-Peer social media as an effective prevention strategy: quasi-experimental evaluation. JMIR Mhealth Uhealth 2020;8(5):e16207 [FREE Full text] [doi: 10.2196/16207] [Medline: 32374270]
12. Molenaar A, Jenkins EL, Brennan L, Lukose D, McCaffrey TA. The use of sentiment and emotion analysis and data science to assess the language of nutrition-, food- and cooking-related content on social media: a systematic scoping review. Nutr Res Rev 2023;37(1):43-78. [doi: 10.1017/s0954422423000069]
13. Sokolova K, Vessal SR, Perez C. Home cooking in the digital age: when observing food influencers on social media triggers the imitation of their practices. Psychology & Marketing 2024:1152-1171 [FREE Full text] [doi: 10.1002/mar.21973]
14. Chung A, Vieira D, Donley T, Tan N, Jean-Louis G, Kiely Gouley K, et al. Adolescent peer influence on eating behaviors via social media: scoping review. J Med Internet Res 2021;23(6):e19697 [FREE Full text] [doi: 10.2196/19697] [Medline: 34081018]
15. Denniss E, Lindberg R, McNaughton SA. Quality and accuracy of online nutrition-related information: a systematic review of content analysis studies. Public Health Nutr 2023;26(7):1345-1357. [doi: 10.1017/s1368980023000873]

XSL•FO
RenderX

16. Yurtdaş-Depboylu G, Kaner G, Özçakal S. The association between social media addiction and orthorexia nervosa, eating attitudes, and body image among adolescents. Eat Weight Disord 2022;27(8):3725-3735. [doi: 10.1007/s40519-022-01521-4] [Medline: 36562891]
17. Fretes G, Veliz P, Narvaez A, Williams D, Sibille R, Arts M, et al. Digital marketing of unhealthy foods and non-alcoholic beverages to children and adolescents: a narrative review. Curr Dev Nutr 2025;9(2):104545 [FREE Full text] [doi: 10.1016/j.cdnut.2025.104545] [Medline: 40046995]
18. Zeeni N, Abi Kharma J, Malli D, Khoury-Malhame M, Mattar L. Exposure to Instagram junk food content negatively impacts mood and cravings in young adults: a randomized controlled trial. Appetite 2024;195:107209 [FREE Full text] [doi: 10.1016/j.appet.2024.107209] [Medline: 38228246]
19. Kucharczuk AJ, Oliver TL, Dowdell EB. Social media's influence on adolescents' food choices: a mixed studies systematic literature review. Appetite 2022;168:105765. [doi: 10.1016/j.appet.2021.105765] [Medline: 34687823]
20. Wilksch SM, O'Shea A, Ho P, Byrne S, Wade TD. The relationship between social media use and disordered eating in young adolescents. Int J Eat Disord 2020;53(1):96-106. [doi: 10.1002/eat.23198] [Medline: 31797420]
21. Diekman C, Ryan C, Oliver T. Misinformation and disinformation in food science and nutrition: impact on practice. J Nutr 2023;153(1):3-9 [FREE Full text] [doi: 10.1016/j.tjnut.2022.10.001] [Medline: 36913465]
22. Filippone L, Shankland R, Hallez Q. The relationships between social media exposure, food craving, cognitive impulsivity and cognitive restraint. J Eat Disord 2022;10(1):184 [FREE Full text] [doi: 10.1186/s40337-022-00698-4] [Medline: 36434703]
23. Lim MSC, Molenaar A, Brennan L, Reid M, McCaffrey T. Young adults' use of different social media platforms for health information: insights from web-based conversations. J Med Internet Res 2022;24(1):e23656 [FREE Full text] [doi: 10.2196/23656] [Medline: 35040796]
24. Breton É, Dufour R, Côté SM, Dubois L, Vitaro F, Boivin M, et al. Developmental trajectories of eating disorder symptoms: a longitudinal study from early adolescence to young adulthood. J Eat Disord 2022;10(1):84 [FREE Full text] [doi: 10.1186/s40337-022-00603-z] [Medline: 35725645]
25. Lanoye A, Brown KL, LaRose JG. The transition into young adulthood: a critical period for weight control. Curr Diab Rep 2017;17(11):114 [FREE Full text] [doi: 10.1007/s11892-017-0938-4] [Medline: 28971312]
26. Poobalan AS, Aucott LS, Clarke A, Smith WCS. Diet behaviour among young people in transition to adulthood (18-25 year olds): a mixed method study. Health Psychol Behav Med 2014;2(1):909-928 [FREE Full text] [doi: 10.1080/21642850.2014.931232] [Medline: 25750826]
27. Webb J, Baker A, Palmer T, Hall A, Ahlquist A, Darlow J, et al. The barriers and facilitators to physical activity in people with a musculoskeletal condition: a rapid review of reviews using the COM-B model to support intervention development. Public Health Pract (Oxf) 2022;3:100250 [FREE Full text] [doi: 10.1016/j.puhip.2022.100250] [Medline: 36101772]
28. Michie S, van Stralen MM, West R. The behaviour change wheel: A new method for characterising and designing behaviour change interventions. Implement Sci 2011;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
29. Atkins L, Francis J, Islam R, O'Connor D, Patey A, Ivers N, et al. A guide to using the theoretical domains framework of behaviour change to investigate implementation problems. Implement Sci 2017;12(1):77 [FREE Full text] [doi: 10.1186/s13012-017-0605-9] [Medline: 28637486]
30. Seiler J, Libby TE, Jackson E, Lingappa J, Evans W. Social media-based interventions for health behavior change in low- and middle-income countries: systematic Review. J Med Internet Res 2022;24(4):e31889 [FREE Full text] [doi: 10.2196/31889] [Medline: 35436220]
31. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
32. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
33. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
34. NHS UK. The Eatwell guide. NHS UK: NHS UK; 2026. URL: https://www.nhsinform.scot/healthy-living/food-and-nutrition/eating-well/eatwell-guide-how-to-eat-a-healthy-balanced-diet/ [accessed 2026-02-27]
35. Cane J, O'Connor D, Michie S. Validation of the theoretical domains framework for use in behaviour change and implementation research. Implement Sci 2012;7:37 [FREE Full text] [doi: 10.1186/1748-5908-7-37] [Medline: 22530986]
36. Sanzari CM, Gorrell S, Anderson LM, Reilly EE, Niemiec MA, Orloff NC, et al. The impact of social media use on body image and disordered eating behaviors: content matters more than duration of exposure. Eat Behav 2023;49:101722 [FREE Full text] [doi: 10.1016/j.eatbeh.2023.101722] [Medline: 37060807]
37. Friedman VJ, Wright CJC, Molenaar A, McCaffrey T, Brennan L, Lim MSC. The use of social media as a persuasive platform to facilitate nutrition and health behavior change in young adults: web-based conversation study. J Med Internet Res 2022;24(5):e28063 [FREE Full text] [doi: 10.2196/28063] [Medline: 35583920]
38. Kulandaivelu Y, Hamilton J, Banerjee A, Gruzd A, Patel B, Stinson J. Social media interventions for nutrition education among adolescents: scoping review. JMIR Pediatr Parent 2023;6:e36132 [FREE Full text] [doi: 10.2196/36132] [Medline: 37471119]

XSL•FO
RenderX

39. Chau MM, Burgermaster M, Mamykina L. The use of social media in nutrition interventions for adolescents and young adults-a systematic review. Int J Med Inform 2018;120:77-91 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.10.001] [Medline: 30409348]
40. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021;23(1):e17187 [FREE Full text] [doi: 10.2196/17187] [Medline: 33470931]
41. Denniss E, Lindberg R. Social media and the spread of misinformation: infectious and a threat to public health. Health Promot Int 2025;40(2):daaf023. [doi: 10.1093/heapro/daaf023] [Medline: 40159949]
42. Silva P, Araújo R, Lopes F, Ray S. Nutrition and food literacy: framing the challenges to health communication. Nutrients 2023;15(22):4708 [FREE Full text] [doi: 10.3390/nu15224708] [Medline: 38004102]
43. Gutiérrez-Ángel N, Sánchez-García JN, Mercader-Rubio I, García-Martín J, Brito-Costa S. Digital literacy in the university setting: a literature review of empirical studies between 2010 and 2021. Front Psychol 2022;13:896800 [FREE Full text] [doi: 10.3389/fpsyg.2022.896800] [Medline: 36176798]
44. Pérez-Torres V. Social media: a digital social mirror for identity development during adolescence. Curr Psychol 2024;43(26):22170-22180. [doi: 10.1007/s12144-024-05980-z]
45. Kington RS, Arnesen S, Chou WS, Curry SJ, Lazer D, Villarruel AM. Identifying credible sources of health information in social media: principles and attributes. NAM Perspect 2021;2021 [FREE Full text] [doi: 10.31478/202107a] [Medline: 34611600]
46. Dadaczynski K, Okan O, Messer M, Leung AYM, Rosário R, Darlington E, et al. Digital health literacy and web-based information-seeking behaviors of university students in Germany during the COVID-19 pandemic: cross-sectional survey study. J Med Internet Res 2021;23(1):e24097 [FREE Full text] [doi: 10.2196/24097] [Medline: 33395396]
47. Stifjell K, Sandanger TM, Wien C. Exploring online health information-seeking behavior among young adults: scoping review. J Med Internet Res 2025;27:e70379 [FREE Full text] [doi: 10.2196/70379] [Medline: 40925001]
48. Liu J, Zhang Y. Modeling health video consumption behaviors on social media: activities, challenges, and characteristics. Proc ACM Hum-Comput Interact 2024;8(CSCW1):1-28. [doi: 10.1145/3653699]
49. Kuutila M, Kiili C, Kupiainen R, Huusko E, Li J, Hosio S, et al. Revealing complexities when adult readers engage in the credibility evaluation of social media posts. Computers in Human Behavior 2024;151:108017 [FREE Full text] [doi: 10.1016/j.chb.2023.108017]
50. Portingale J, Girardin S, Liu S, Fuller-Tyszkiewicz M, Krug I. Daily bi-directional effects of women's social media-based appearance comparisons, body satisfaction, and disordered eating urges. J Eat Disord 2024;12(1):129 [FREE Full text] [doi: 10.1186/s40337-024-01096-8] [Medline: 39227958]
51. Kim JH, Kim J, Jang SY, Park EC. Association between watching eating broadcasts like mukbang and cookbang and generalized anxiety disorder among Korean adolescents. BMC Psychiatry 2024;24(1):536 [FREE Full text] [doi: 10.1186/s12888-024-05957-z] [Medline: 39080570]
52. Sanskriti S, Guglani I, Joshi S, Anjankar A. The spectrum of motivations behind watching mukbang videos and its health effects on its viewers: a Review. Cureus 2023;15(8):e44392 [FREE Full text] [doi: 10.7759/cureus.44392] [Medline: 37786568]
53. Milton A, Ajmani L, DeVito MA, Chancellor S. "I See Me Here": Mental Health Content, Community, and Algorithmic Curation on TikTok. New York, NY: Association for Computing Machinery; 2023 Presented at: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems: Association for Computing Machinery; 2023 April 23; Hamburg, Germany. [doi: 10.1145/3544548.3581489]
54. Savolainen L, Ruckenstein M. Dimensions of autonomy in human–algorithm relations. New Media & Society 2022;26(6):3472-3490. [doi: 10.1177/14614448221100802]
55. Scharlach R, Hallinan B. The value affordances of social media engagement features. Journal of Computer-Mediated Communication 2023;28(6):zmad040. [doi: 10.1093/jcmc/zmad040]
56. Bouton ME. Why behavior change is difficult to sustain. Prev Med 2014;68:29-36. [doi: 10.1016/j.ypmed.2014.06.010]
57. Veckalne R, Akhmatjonov S, Humbatova S, Hajiyev NG. From scroll to action: can social media influencers drive real behavioral change among their followers? Social Sciences 2025;14(5):253. [doi: 10.3390/socsci14050253]
58. Conner M, Norman P. Understanding the intention-behavior gap: the role of intention strength. Front Psychol 2022;13:923464 [FREE Full text] [doi: 10.3389/fpsyg.2022.923464] [Medline: 35992469]
59. Simeon R, Dewidar O, Trawin J, Duench S, Manson H, Pardo Pardo J, et al. Behavior change techniques included in reports of social media interventions for promoting health behaviors in adults: content analysis within a systematic review. J Med Internet Res 2020;22(6):e16002 [FREE Full text] [doi: 10.2196/16002] [Medline: 32525482]
60. Consavage Stanley K, Harrigan PB, Serrano EL, Kraak VI. A systematic scoping review of the literacy literature to develop a digital food and nutrition literacy model for low-income adults to make healthy choices in the online food retail ecosystem to reduce obesity risk. Obes Rev 2022;23(4):e13414 [FREE Full text] [doi: 10.1111/obr.13414] [Medline: 35092142]

Abbreviations

COM-B: Capability, Opportunity, Motivation and Behaviour model

XSL•FO
RenderX

COREQ: Consolidated Criteria for Reporting Qualitative Research
HFSS: high in fat, sugar, and salt
MRC: Medical Research Council
PPI: patient and public involvement
REDCap: Research Electronic Data Capture
SRD: Secure Research Drive
SSB: sugar-sweetened beverage
TDF: theoretical domains framework

Edited by A Stone; submitted 11.Dec.2025; peer-reviewed by AJ Umeano, N Döring; comments to author 13.Feb.2026; revised version received 13.Mar.2026; accepted 16.Mar.2026; published 07.Apr.2026.

Please cite as:
Tang H, Ahern AL, Spreckley M, Smith AD
Young Adults' Interactions With Food and Nutrition Content on Social Media and Implications for Intervention Design: Semistructured Interview Study
J Med Internet Res 2026;28:e89344
URL: https://www.jmir.org/2026/1/e89344
doi:10.2196/89344
PMID:

©Hao Tang, Amy L Ahern, Marie Spreckley, Andrea D Smith. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Human and Robot Assistance for Cognitive Load in Younger and Older Adults: Multimodal Within-Subject Experimental Study

Simone Varrasi[1], PhD, PsyD; Roberto Vagnetti[2], PhD; Nicola Camp[3], PhD; John Hough[3], PhD; Alessandro Di Nuovo[4], PhD; Sabrina Castellano[1], PhD; Daniele Magistro[3,5], PhD

[1]Department of Educational Sciences, University of Catania, Catania, Sicily, Italy

[2]Kent and Medway Medical School, University of Kent and Canterbury Christ Church University, Canterbury, England, United Kingdom

[3]Department of Sport Science, School of Science and Technology, Nottingham Trent University, Nottingham, England, United Kingdom

[4]Department of Computing, Sheffield Hallam University, Sheffield, England, United Kingdom

[5]School of Health Sciences, University of Southampton, Building 67, Highfield Campus, University Road, Southampton, England, United Kingdom

Corresponding Author:
Daniele Magistro, PhD
Department of Sport Science, School of Science and Technology, Nottingham Trent University, Nottingham, England, United Kingdom

Abstract

Background: Maintaining cognitive efficiency and independence is a central goal of healthy aging. Socially assistive robots (SARs) are increasingly proposed as scalable digital health solutions to support daily activities in older adults and to facilitate aging-in-place. However, concerns remain regarding whether robot-mediated assistance reduces or inadvertently increases cognitive load, potentially undermining usability, user acceptance, and long-term real-world adoption, particularly in aging populations.

Objective: This study aimed to examine how robot-assisted (human-robot interaction [HRI]) and human-assisted (human-human interaction [HHI]) support influences cognitive load during task performance in younger and older adults. A multimodal assessment framework integrating behavioral, subjective, and physiological measures was used to identify age-related differences in cognitive effort and stress associated with different forms of assistance.

Methods: A total of 60 healthy adults (30 younger adults: mean age 34.8, SD 10.1 years; and 30 older adults: mean age 72.3, SD 5.5 years) completed a modified Trail Making Test under 7 within-subject conditions: independent performance (baseline), 3 robot-assisted conditions, and 3 human-assisted conditions, each corresponding to low, medium, and high cognitive load levels. Performance accuracy and completion time were recorded as behavioral indicators. Perceived cognitive load was assessed using the National Aeronautics and Space Administration Task Load Index, and physiological stress was evaluated via pre- and postcondition salivary cortisol concentrations. Linear mixed-effects models were applied to examine main effects and interactions of age group, assistance type, cognitive load level, and time.

Results: Significant interactions between age group and assistance type were observed for accuracy ($F_{1, 404.53}$=6.50; P=.01) and perceived cognitive load ($F_{1, 403.45}$=4.58; P=.03). Older adults demonstrated lower accuracy and higher perceived cognitive load during robot-assisted conditions compared with human-assisted conditions, whereas no such differences were observed in younger adults. Across age groups, human assistance improved performance at low and medium cognitive load levels. Physiological analysis revealed a significant age×assistance× time interaction ($F_{1, 156}$=5.16; P=.02), with older adults showing increased posttask cortisol concentrations during robot-assisted interaction, indicating higher physiological stress.

Conclusions: While both human and robotic assistance enhanced task performance relative to independent completion, the type of support critically shaped cognitive load responses in older adults. Robot-assisted interaction was associated with increased behavioral errors, higher perceived workload, and elevated physiological stress, suggesting that current SAR implementations may impose additional extraneous cognitive load in older users. These findings highlight the importance of designing adaptive, age-sensitive digital assistive systems that minimize cognitive burden through simplified interaction, responsive pacing, and multimodal support. Multimodal cognitive load assessment provides a valuable framework for optimizing the usability and effectiveness of assistive digital health technologies for aging populations.

(J Med Internet Res 2026;28:e94738) doi:10.2196/94738

KEYWORDS

aging; assistive robotics; cognitive load; daily autonomy; multimodal measures

XSL•FO
RenderX

Introduction

Maintaining independence and cognitive efficiency in daily life constitutes a central objective in the promotion of active and healthy aging [1]. The ability to manage daily activities with adequate support has become a key focus across multiple disciplines, particularly in light of the projected rise in the global mean age over the coming decades [2]. In this context, health and social care systems are increasingly exploring digital health technologies to support autonomy, reduce care burden, and enable aging in place, underscoring the need for scalable and evidence-based interventions aimed at enhancing quality of life among older adults.

Technological progress has played a pivotal role in pursuing this goal. Intelligent systems can now be embedded within domestic environments, and embodied agents may be programmed to interact with users to accomplish shared objectives [3]. Within this framework, socially assistive robotics (SAR) leverages human-robot interaction (HRI) to facilitate such activities and to foster autonomy in everyday life [4]. SAR systems are increasingly positioned as interactive digital health interventions, integrating social interaction, guidance, and task support within real-world settings. However, recent evidence indicates that enabling communication with an artificial agent alone is insufficient to ensure effective interaction, particularly in vulnerable groups such as older adults [5]. Beyond demonstrating performance benefits, digital assistive technologies must also be evaluated in terms of usability, acceptability, and sustainability of long-term adoption [6,7]. Interventions that improve task execution but simultaneously increase cognitive or physiological strain may face significant barriers to real-world implementation.

From a cognitive ergonomics perspective, cognitive load (CL) represents a critical determinant of HRI effectiveness and of the usability and sustainability of digital health technologies [8]. Conceptually, CL is defined as the amount of mental resources required to perform a given task [9]. Because cognitive resources are inherently limited, individuals must distribute them efficiently across competing activities [10].

Importantly, CL is not a unitary construct but rather comprises three interrelated components: intrinsic, extraneous, and germane load. Intrinsic load stems from the inherent complexity of the task and the volume of information to be processed. Extraneous load arises from the presentation format and the presence of external distractions, both of which can be mitigated through effective task design. Germane load, instead, pertains to the cognitive processes that foster learning and schema construction. Recent theoretical models have proposed integrating intrinsic and germane load, conceptualizing intrinsic and extraneous load as the two principal dimensions of CL theory [11-13]. In applied digital health contexts, extraneous CL introduced by interface design is particularly problematic, as it may negate the intended supportive function of assistive technologies.

A growing body of research indicates that older adults face greater challenges in managing CL, thereby affecting their functional autonomy [14]. Such difficulties are often attributed to age-related cognitive decline, which manifests as increased fatigue and reduced efficiency in completing daily tasks [15]. These challenges extend to the use of technology, where the additional cognitive demands associated with complex interfaces further amplify CL [16,17]. Moreover, older adults' attitudes toward advanced technologies are shaped not only by sociocultural factors but also by individual cognitive needs and the degree of system personalization [18]. Neglecting these factors' risks imposing additional cognitive burdens and may contribute to digital exclusion among those most in need of support.

Despite the growing interest in SAR, few studies have directly compared robot-mediated assistance with human assistance under controlled CL conditions. Such comparisons are essential, as assistive technologies are often implicitly framed as substitutes for human support. Furthermore, assessing technology-assisted performance without an independent baseline condition limits the ability to determine whether observed effects reflect genuine support or task facilitation, as highlighted in previous work [19].

Accordingly, it becomes essential to determine whether technology use increases users' CL, to identify the features responsible for such increases, and to examine how these effects differ across age groups. Such insights are crucial for the design of adaptive interfaces capable of dynamically accommodating users' cognitive capacities. Nevertheless, the existing literature presents several methodological limitations in addressing these issues.

First, assessing CL during technology use without a baseline condition prevents reliable interpretation. Experimental designs should therefore include a comparison between technology-assisted and independent task performance to delineate the specific benefits and drawbacks of technological aids, a methodological framework exemplified by Varrasi et al [19]. Furthermore, studies asserting the positive impact of technology without incorporating a condition involving human support lack robustness; indeed, multiple aid conditions are required for a comprehensive evaluation. Likewise, hypothesizing user needs without conducting age comparisons constrains the development of genuine person-centered support systems. Specific sampling is thus necessary to enable between-group analyses and to characterize distinct population profiles. Finally, the reliance on a single CL measure limits construct validity, underscoring the need to use multimodal assessment approaches.

CL can be quantified through a range of behavioral, subjective, and physiological indicators, each providing complementary information on the cognitive demands placed on the individual. Behavioral indices include performance-based metrics, such as accuracy, error rates, and completion time, which serve as quantifiable markers of cognitive effort [20]. Subjective indices capture self-reported perceptions of mental workload, encompassing aspects such as perceived effort, frustration, and mental demand [21]. Integrating these approaches enables a comprehensive evaluation of cognitive demands and supports the robust assessment of assistive technologies in digital health contexts. Physiological indices, including heart rate variability,

XSL•FO
RenderX

pupil dilation, blood and salivary cortisol concentrations, and electroencephalographic (EEG) activity, reflect the body's physiological responses to cognitive strain and stress [22]. Integrating these methods offers a comprehensive assessment of CL and facilitates a more accurate evaluation of task design and learning efficiency.

Given these considerations, this study aimed to investigate whether, and in what ways, HRI and human-human interaction (HHI) influence CL management during task performance in younger and older adults. By incorporating an independent baseline condition and a multimodal psychometric framework, this study seeks to inform the design, evaluation, and deployment of age-sensitive digital assistive technologies.

Methods

Study Design

This study used a within-subject, experimental comparative design to examine differences in CL during independent, robot-assisted, and human-assisted task performance in younger and older adults.

Participants

The study included 60 adult volunteers, comprising 30 younger adults (13 males, 17 females; mean age 34.8, SD 10.1 years) and 30 older adults (12 males, 18 females: mean age 72.3, SD 5.5 years), with sex distribution confirmed as balanced (χ^2_1=0.07; *P*=.79). Younger adults were aged 18 - 45 years, and older adults were more than 65 years of age. In line with the study aims, age differed significantly between groups (t_{58}=17.86; *P*<.001). Education levels were broadly comparable across groups. Among younger adults, 16.7% (5/30) reported high school education, 23.3% (7/30) bachelor's, 30% (9/30) master's, and 30% (9/30) PhD-level education. Among older adults, 33.3% (10/30) reported high school education, 30% (9/30) bachelor's, 23.3% (7/30) master's, and 13.3% (4/30) PhD-level education. Although younger adults showed a tendency toward higher educational attainment, this difference was not statistically significant (χ^2_3=4.09; *P*=.25). An a priori power analysis (2-tailed α=.05, 80% power) indicated that the sample size of 30 adults per group provides sufficient power to detect Cohen *d*~0.74 for between-group comparisons and interaction effects (approximated as between-group differences in within-subject contrasts). Within-group paired contrasts have 80% power to detect dz~0.53 (n=30) and dz~0.37 when considering the full sample (N=60). These estimates represent conservative approximations relative to the mixed-effects modeling approach used in the primary analyses.

All participants underwent screening to confirm good general health, as only healthy individuals were included in the study. Specifically, participants had no prior diagnosis of neurological or psychiatric disorders and reported no previous direct experience interacting with socially assistive robots. This sampling strategy was adopted to isolate age-related differences in CL responses under controlled experimental conditions while minimizing additional variability associated with clinical impairment. Accordingly, the study was conceived as an initial step toward understanding how assistance modality influences CL, with the intention of extending this line of research to more vulnerable populations in future work.

Ethical Considerations

The study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki. Ethical approval was obtained from the Nottingham Trent University Ethics Board (reference number 729) in March 2024. All participants provided written informed consent prior to participation and were informed of their right to withdraw at any time without consequence.

Tools

The cognitive task used in this study was adapted from the B form of the Trail Making Test (TMT) [23]. The TMT was selected over alternative cognitive measures due to its engagement of multiple cognitive domains, including attention, visual search, executive functioning, and working memory. Consequently, the task required participants to simultaneously deploy multiple cognitive skills, closely reflecting the complexity of everyday activities.

Specifically, the task used a 29.7×42 cm sheet of paper containing 52 randomly arranged circles (Figure 1). Half of the circles included English letters (A-Z), while the remaining circles contained numbers (1-26). A total of 7 task versions were created to minimize learning effects across repeated conditions. These versions were not intended to represent validated parallel forms of the TMT in a clinical sense; rather, they were task variants inspired by the TMT and designed to evoke comparable executive, attention, and visual search demands. For each version, the 52 circles were randomly distributed across the sheet to create distinct spatial arrangements while preserving the same basic task structure. Task versions were assigned randomly to participants across conditions, with each participant completing a different version in each condition. Objective performance was behaviorally evaluated based on the number of correctly connected pairs and the time required to complete the task (seconds).

XSL•FO
RenderX

Figure 1. Cognitive task sheet used in the study. Example of one task version consisting of 52 randomly arranged circles distributed across a 29.7 × 42 cm sheet. Half of the circles contained English letters (A–Z), and the remaining circles contained numbers (1–26). Participants were instructed to connect the targets sequentially in alternating alphanumeric order (eg, 1–A–2–B–3–C) as quickly and accurately as possible. Seven task versions with different spatial arrangements were created to minimize learning effects across repeated conditions.

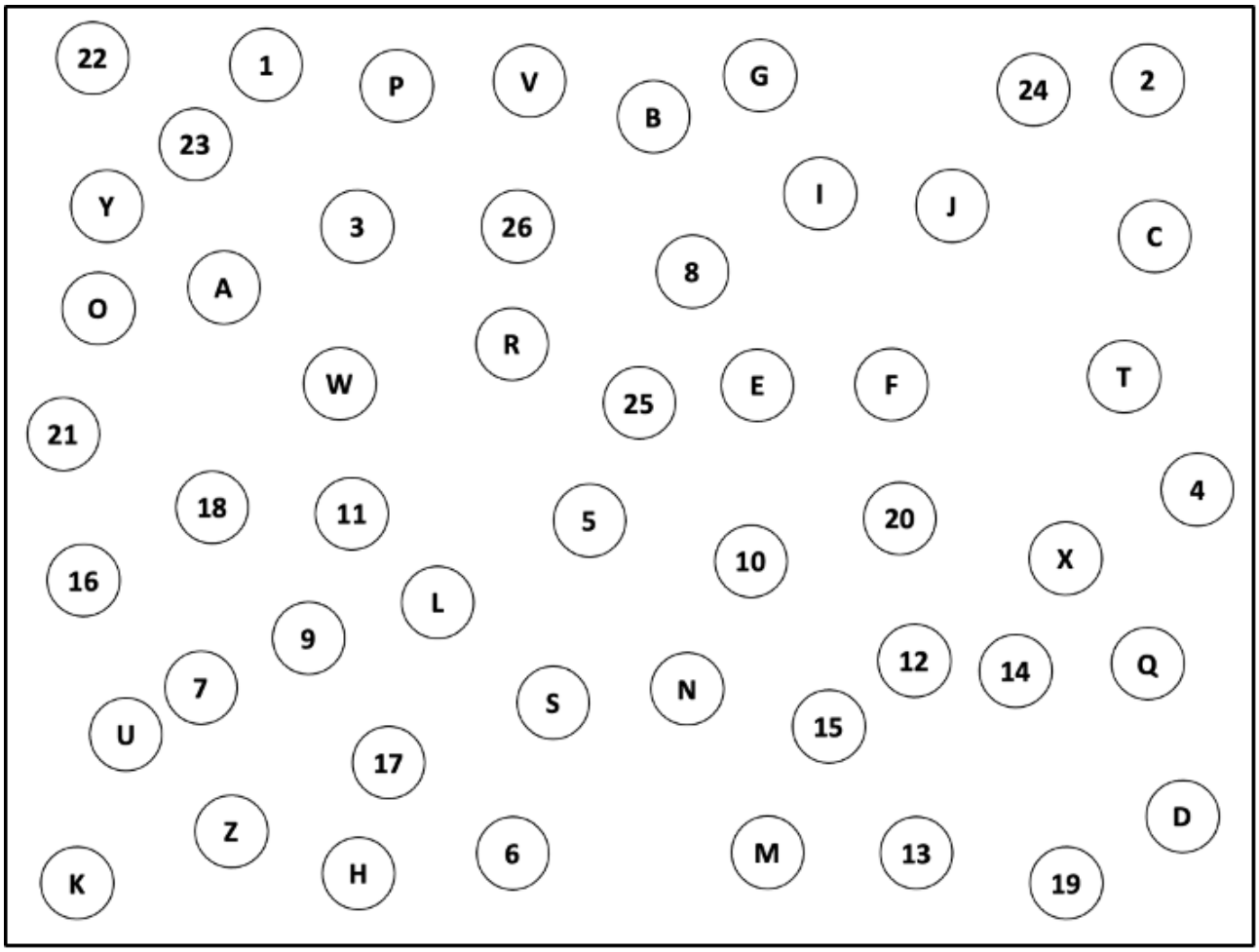

The supportive agent used in this study was the NAO 6 humanoid robot (United Robotics Group). This platform was selected for its commercial availability and relevance to real-world assistive deployments aligned with the study's requirements, including speech production, speech recognition, and social interaction. Previous studies have shown that NAO (Aldebaran, Paris, France) functions as an SAR that is generally well-received by older and vulnerable populations [16]. The robot is 58 cm tall and can interact through speech, gestures, and facial tracking. Programming was implemented using Choregraphe software (version 2.8.7; Aldebaran) together with Python (Python Software Foundation) to enable interactive collaboration with participants in 3 of the 7 experimental conditions.

The robot's behavior was structured using block-based visual programming within a node flow interface. Interactions began with NAO tracking the participant's face to simulate eye contact, followed by a self-introduction and delivery of initial task instructions. Speech recognition allowed the robot to repeat instructions upon participant request during an initial familiarization phase, which served as a vocal adaptation period responsive to participant feedback. After this phase, NAO did not repeat instructions to avoid disrupting task performance, although speech recognition remained active to manage pacing during the interaction. Indeed, the robot-participant interaction followed a structured and standardized dialogue logic designed to ensure consistency across participants while maintaining naturalistic turn-taking dynamics. Specifically, interaction sequences were organized into alternating robot-led instruction phases and participant execution phases. The robot presented task instructions verbally according to predefined scripts and advanced through task steps only after receiving explicit verbal confirmation from the participant (eg, "Okay" or "Done"), thereby regulating the interaction rhythm.

The presentation of task instructions varied systematically according to CL conditions. In the low-load condition, the robot delivered one pair of items at a time, whereas in the medium- and high-load conditions, multiple pairs were provided sequentially within a single instruction block. The interaction pacing was therefore partially scripted, with fixed instruction sequences, but temporally modulated by participant responses, allowing limited responsiveness to individual execution speed.

To ensure procedural reproducibility, the interaction logic combined fixed scripted content (instruction wording, sequence structure, and load manipulation) with controlled interaction triggers based on participant verbal feedback. A schematic representation of the interaction workflow is provided in Figure 2.

XSL•FO
RenderX

Figure 2. Flowchart of the interaction with the robot with active interfaces.

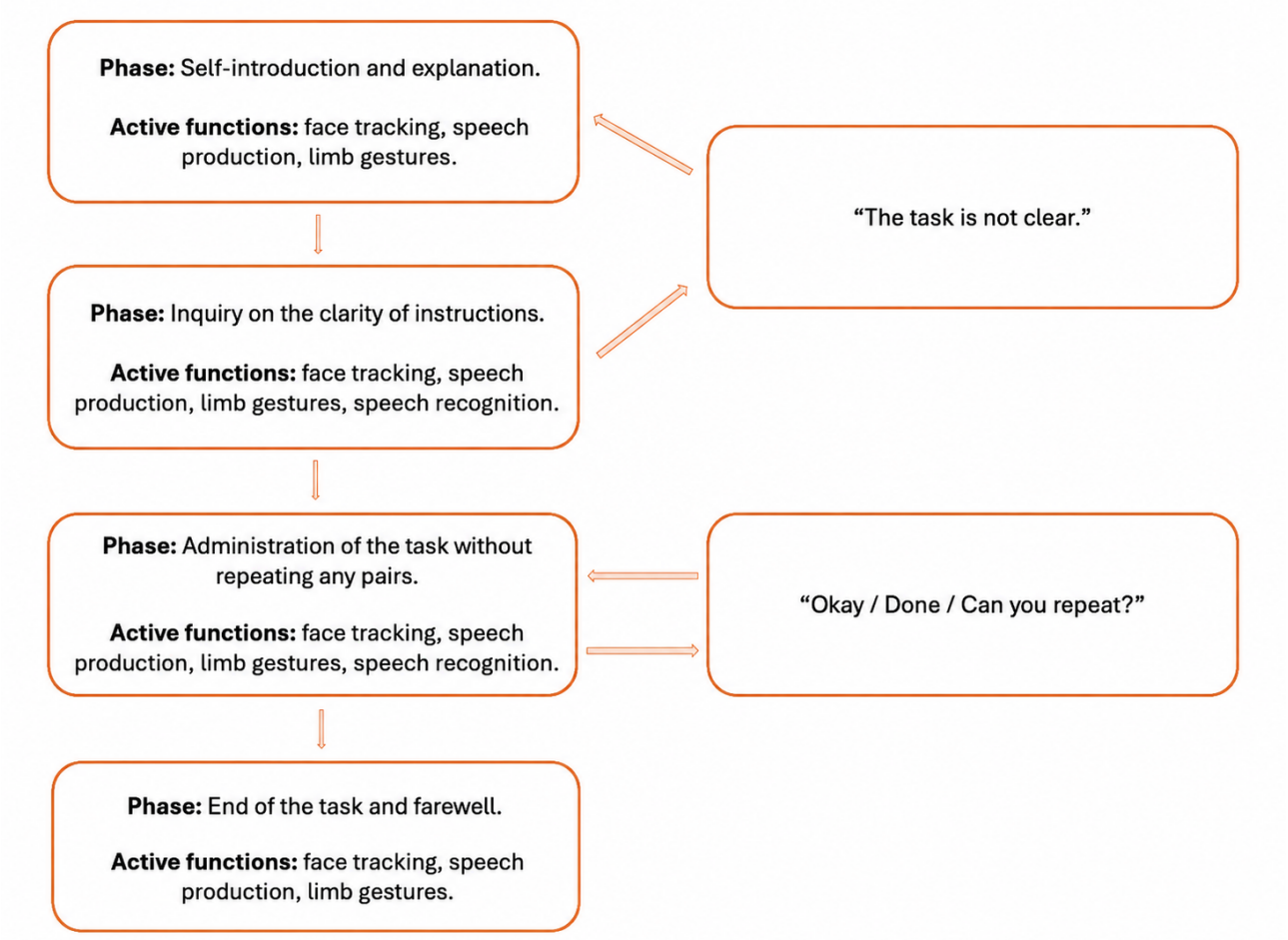

The flowchart illustrates the decision nodes governing instruction delivery, response confirmation, and transition between interaction phases, providing a visual representation of the standardized dialogue protocol implemented across conditions. Moreover, to enhance transparency and reproducibility of the interaction protocol, a structured overview of the robot dialogue architecture, including interaction phases, scripted instruction templates, and response-triggered transitions, is provided in Table 1.

Table . Structured overview of the robot dialogue architecture.

Phase	Robot action	Participant action	Trigger
Greeting	Introduces task	Listens	Experimenter
Instruction	Provides numbers and letters to connect	Executes task	Scripted
Confirmation	Waits for response	Verbal confirmation	Voice trigger
Transition	Provides next pair	Continues task	Confirmation

Perceived cognitive load (PCL) was assessed using the National Aeronautics and Space Administration Task Load Index (NASA-TLX) [24], a brief, self-administered psychometric tool evaluating 6 dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration. Each dimension was rated on a 20-point scale, with higher scores indicating greater perceived strain. The NASA-TLX was chosen for its robust psychometric properties and demonstrated superiority over alternative subjective workload measures [25], making it well-suited for assessing PCL in this study. In line with the primary aim of capturing overall perceived workload during assisted task performance, the total NASA-TLX score was used as a global indicator of subjective CL. Although the instrument allows examination of individual workload dimensions, the aggregated score has been widely used to provide a comprehensive estimate of perceived task demand, particularly in applied human-technology interaction research [25].

To objectively assess physiological stress responses associated with the CL tasks, saliva samples were collected immediately before and after each condition (ie, robot-assisted and human-assisted). Samples were obtained using a synthetic swab (Salivette Cortisol, Sarstedt), instructing participants to slowly move the swab around the oral cavity for 2 minutes, in

XSL•FO
RenderX

accordance with the manufacturer's instructions. All samples were maintained on crushed ice throughout testing, which did not exceed 45 minutes. Subsequently, saliva samples were centrifuged at 1000×*g* for 2 minutes (Espresso Microcentrifuge, Thermo Scientific), and the supernatant was transferred into 1.5 mL Eppendorf tubes and stored at −80 °C until analysis. Quantitative determination of salivary cortisol concentrations was performed using commercially available enzyme-linked immunosorbent assay kits (Salimetrics), following the manufacturer's recommended procedures. Salivary cortisol was included as a physiological indicator of stress-related arousal associated with cognitive effort. While cortisol does not represent a direct measure of CL, it provides complementary information regarding the neuroendocrine response to task-related strain, allowing integration with behavioral and subjective workload indicators. Moreover, salivary cortisol was sampled only before and after each assisted block, and not after the independent baseline condition, because the physiological outcome was intended to capture the cumulative response associated with the two assistance modalities rather than transient changes across individual CL levels or during unassisted performance. This choice was based on both theoretical and methodological considerations, including the temporal dynamics of the hypothalamic-pituitary-adrenal axis, as salivary cortisol typically peaks approximately 20‑40 minutes after stress onset. Accordingly, cortisol was treated as a block-level marker of stress-related arousal, and restricting sampling to the assisted conditions helped reduce participant burden while preserving the feasibility and ecological validity of the experimental protocol.

Procedure

Each participant completed the experimental protocol individually across seven conditions: one baseline condition performed independently, three robot-assisted conditions at varying CL levels (low, medium, and high), and three human-assisted conditions corresponding to the same CL levels.

The baseline condition was systematically administered at the beginning of the session to provide a common, unassisted reference point for all participants. Within each assisted block, CL levels were presented in a fixed progressive sequence (low, medium, and high) in both the robot-assisted and human-assisted conditions. This choice was made to ensure a gradual and controlled increase in task demands, thereby minimizing potential confusion associated with abrupt changes in difficulty and maintaining a stable and predictable interaction structure, particularly for older adults. In contrast, the order of the assisted blocks (robot-assisted vs human-assisted) was fully counterbalanced across participants to control for order effects at the level of assistance modality.

Both assisted conditions were administered within the same experimental session. Consequently, the precondition saliva sample for the second assisted block was collected immediately before its onset, following completion of the preceding condition and a brief transition phase required to reset the materials and prepare the subsequent task. While counterbalancing mitigates systematic sequence effects at the level of condition order, the absence of an extended washout interval implies that a residual carryover influence from the first assisted condition on the precondition cortisol value of the second block cannot be entirely excluded. The total duration of the experimental session was approximately 30 minutes.

After providing informed consent, the participant was seated at a desk and presented with the first task sheet. In the baseline condition, the participant was instructed to connect alternating numbers and letters in reverse order (ie, starting from the highest number and the last letter) as quickly and accurately as possible using a marker. This task was designed to elicit CL by engaging multiple cognitive domains, including semantic memory, working memory, visual search, and processing speed. Specifically, the participant was required to recall the reverse sequence of alphanumeric items, maintain this sequence in working memory, and identify and connect the corresponding circles on the sheet.

During the three robot-assisted conditions, the NAO humanoid robot acted as an interactive support agent (Figure 3). At the beginning of each assisted condition, the robot greeted the participant, introduced the task, and repeated the instructions upon request. Throughout task execution, the robot verbally provided the pairs of circles to be connected, thereby reducing reliance on semantic memory and allowing participants to focus on visual search and motor coordination. The robot advanced through the sequence based on participant responses, waiting for a verbal confirmation (eg, "Okay") before announcing the next pair. To prevent learning effects, the sequence of pairs was randomized across trials. CL was systematically manipulated as follows: in the low-load condition (second and fifth condition), the robot presented one pair at a time; in the medium-load condition (third and sixth condition), two to four pairs were provided simultaneously; and in the high-load condition (fourth and seventh condition), five to seven pairs were delivered concurrently. The robot did not repeat any pair throughout the session.

XSL•FO
RenderX

Figure 3. Setting of the robot-interactive task.

The human-assisted condition was administered by a trained experimenter who followed a standardized script designed to mirror the robot-assisted protocol in content, sequence, and CL manipulation. The experimenter delivered the instructions verbally using the same wording as the robot prompts, maintained the same load-specific structure, and waited for the same participant verbal confirmations before proceeding to the next step. A brief initial familiarization phase was also included to ensure procedural parity with the robot-assisted condition: during this phase, the experimenter presented the task instructions and allowed participants to request clarification, analogous to the vocal adaptation period implemented in the HRI condition. Following familiarization, the interaction proceeded according to the predefined script without additional instruction repetition. The only difference between conditions was the modality of delivery, with all other procedural elements held constant. A structured overview of the human-assisted

XSL•FO
RenderX

dialogue, emphasizing the fidelity rules for the adherence to the HRI condition, is provided in Table 2.

Table . Structured overview of the human dialogue architecture with fidelity rules.

Phase	Human experimenter action	Participant action	Trigger and fidelity rule
Greeting	Introduces task	Listens	Standardized opening script
Instruction	Provides numbers and letters to connect	Executes task	Verbatim script
Confirmation	Waits for response	Verbal confirmation	Proceed only after confirmation
Transition	Provides next pair	Continues task	Same sequencing as robot condition

At the end of each condition, including baseline, robot-assisted, and human-assisted tasks, participants completed the NASA-TLX questionnaire to evaluate PCL. As described previously, saliva samples were collected at the beginning and the end of both the robot-assisted and human-assisted conditions to capture physiological markers of CL, yielding four samples per participant.

For data analysis, two independent experimenters (SV and RV) assessed task performance across all conditions. Performance measures included accuracy (number of correctly connected pairs) and completion time (in seconds). Subjective workload was quantified using NASA-TLX scores, while physiological responses were evaluated by comparing pre- and postcondition cortisol concentrations across the robot- and human-assisted settings.

Statistical Analyses

Linear mixed-effects models were applied using restricted maximum likelihood estimation to examine differences across the primary outcome variables, including NASA-TLX scores, task performance indices, and cortisol responses. Participants were included as random intercepts to account for interindividual variability in baseline performance and physiological measures.

A type III ANOVA with Satterthwaite approximation was used to assess the main effects and interactions of condition (robot-assisted vs human-assisted), CL level (baseline, low, medium, and high), and age group (younger vs older adults). Condition at varying load levels was modeled as a within-subject factor, while age group was treated as a between-subject factor.

Post hoc comparisons were performed on estimated marginal means and mean differences (M_{dif}) to further explore significant main effects and interactions. Tukey-adjusted estimated marginal means comparisons were used to correct for multiple testing. In line with the study's objectives, particular attention was devoted to the interaction between age group and experimental condition to clarify differential patterns of CL management and task performance across age cohorts.

Results

Mean (SD) values for accuracy (number of correctly matched pairs), completion time (in seconds), and PCL across the seven experimental conditions, as well as cortisol levels (nmol/L) measured before and after the robot-assisted and human-assisted conditions, are presented in Tables 3 and 4.

Table . Task performance and perceived cognitive load.

Conditions	Accuracy, mean (SD)	Time, mean (SD)	PCL[a], mean (SD)
Alone			
Baseline	23.4 (5.8)	405 (116)	49.2 (15.1)
Robot			
Low CL[b]	24.4 (2.5)	364 (141)	40.9 (16.6)
Medium CL	16 (4)	342 (69.6)	58.3 (15.2)
High CL	9.0 (3)	251 (56)	66.2 (14)
Human			
Low CL	25.7 (0.6)	332 (91.4)	40.3 (15.3)
Medium CL	18.4 (3.2)	362 (86.5)	57.1 (13.4)
High CL	8.2 (2.3)	246 (61.1)	68.5 (12.3)

[a]PCL: perceived cognitive load.

[b]CL: cognitive load.

XSL•FO
RenderX

Table . Cortisol levels[a] by condition and time.

Conditions and time	Cortisol, mean (SD)
Human	
Pre	7.72 (8.54)
Post	7.48 (8.09)
Robot	
Pre	6.02 (3.51)
Post	8.87 (10.4)

[a]Cortisol levels (nmol/L) were measured at the beginning (pre) and end (post) of the human-human interaction and human-robot interaction conditions.

Regarding performance, analysis of accuracy revealed significant interactions between age group and condition ($F_{1, 404.53}$=6.50; P=.01) as well as between age group and CL level ($F_{3, 403.45}$=5.88; P<.001). Post hoc comparisons indicated that older adults exhibited lower accuracy in the HRI condition compared with younger adults (M_{dif}=1.97, SE=0.58; t_{106}=3.34; P=.001; Figure 4A) and also relative to the human-assisted condition (M_{dif}=1.54, SE=0.43; t_{403}=3.57; P<.001; Figure 4B). Additionally, older adults demonstrated significantly lower accuracy than younger adults at medium (M_{dif}=2.75, SE=0.73; t_{219}=3.75; P<.001) and high CL levels (M_{dif}=1.64, SE=0.73; t_{219}=2.23; P=.03), whereas no significant differences were observed at low load (M_{dif}=1.16, SE=0.73; t_{219}=1.58; P=.12) or in the baseline condition (M_{dif}=−0.80, SE=0.73; t_{217}=−1.09; P=.27; Figure 4C).

Figure 4. Older adults had a worse performance at HRI condition than younger adults(**A**) and than HHI (**B**). Lower accuracy was shown at increasing levels of cognitive load with respect to adults (**C**), while high cognitive load was challenging for all participants independently from the interactive condition (**D**). HHI: human-human interaction; HRI: human-robot interaction.

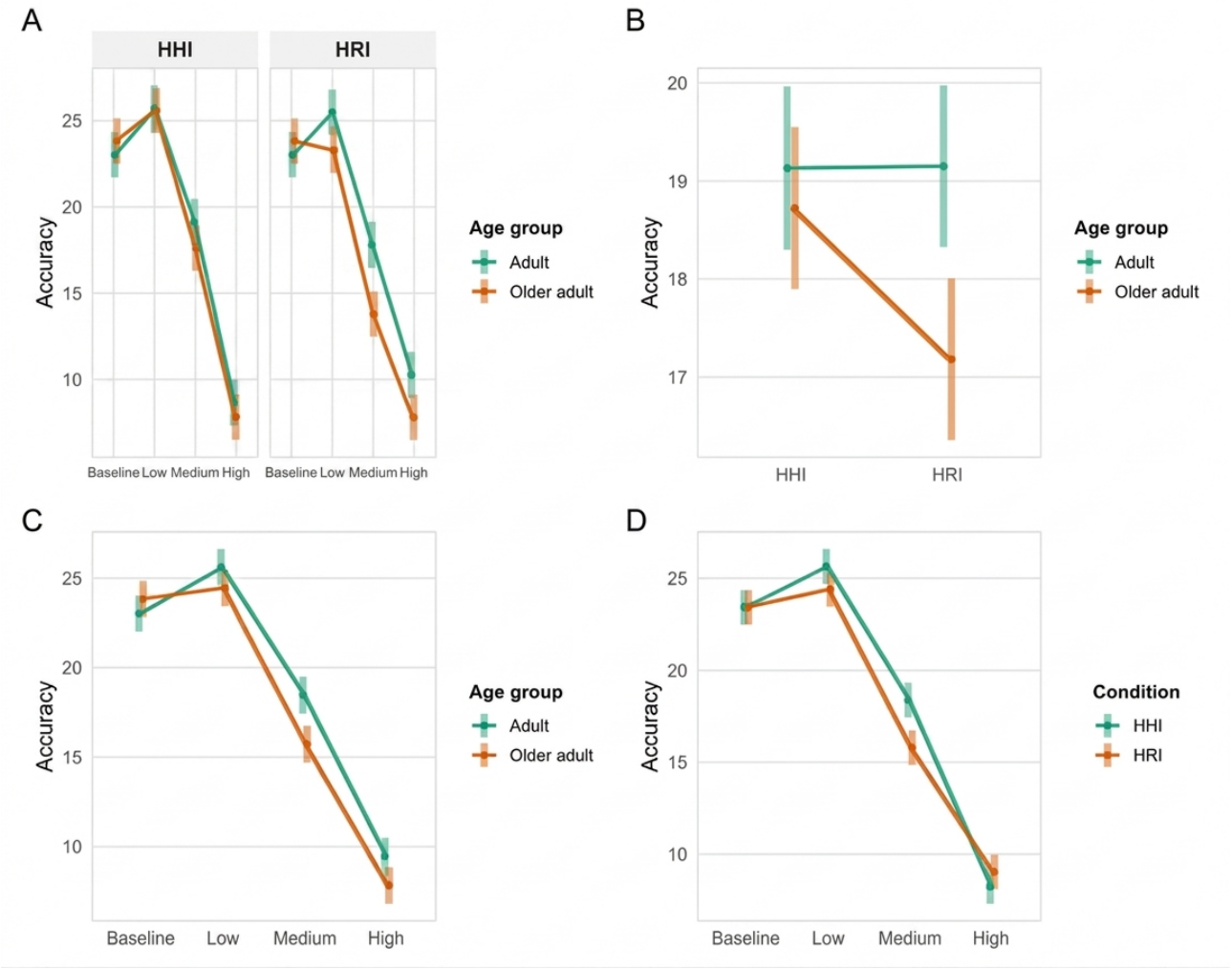

XSL•FO
RenderX

A significant interaction between condition and CL level was also observed ($F_{1, 403.45}$=5.83; P<.001). Across all participants, post hoc tests revealed higher accuracy in the human-assisted compared with the robot-assisted condition at low (M_{dif}=1.25, SE=0.61; t_{404}=2.05; P=.04) and medium CL levels (M_{dif}=2.58, SE=0.61; t_{404}=4.21; P<.001), whereas no significant differences emerged at high CL (M_{dif}=−0.79, SE=0.61; t_{404}=−1.29; P=.19; Figure 4D).

Analysis of task completion time revealed a significant interaction between age group and CL level ($F_{3, 402.85}$=23.73; P<.001). Post hoc comparisons indicated that younger adults completed the task faster than older adults at baseline (M_{dif}=−123.73, SE=18.70; t_{114}=−6.62; P<.001), as well as under low (M_{dif}=−115.95, SE=18.70; t_{115}=−6.19; P<.001) and medium CL conditions (M_{dif}=−75.16, SE=18.70; t_{115}=−4.01; P<.001). No significant difference was observed at high CL (M_{dif}=1.73, SE=18.70; t_{115}=0.09; P=.92; Figure 5).

Figure 5. Youngeradults were faster than older adults at the baseline, low cognitive load, and medium cognitive load.

Analysis of NASA-TLX scores revealed significant interactions between age group and condition ($F_{1, 403.45}$=4.58; P=.03) as well as between age group and CL level ($F_{3, 403.12}$=8.53; P<.001). Regarding the interaction between age group and condition, post hoc comparisons showed that older adults reported significantly higher scores during the robot-assisted condition (mean 57.60, SE 2.13) compared with the human-assisted condition (mean 54.3, SE 2.13), with an estimated M_{dif} of −3.29 (SE=1.16; t_{403}=−2.83; P=.005). No significant difference was observed in younger adults (M_{dif}=0.25, SE=1.18; t_{404}=0.21; P=.83), indicating that older adults experienced greater perceived CL when interacting with the robot compared with human support (Figure 6A).

XSL•FO
RenderX

Figure 6. Older adults reported significantly higher overall cognitive load during the robot condition compared to the human condition. No significant condition differences were found for younger adults (**A**). Older adults perceived greater cognitive load than adults specifically under low cognitive load levels (**B**). HHI: human-human interaction; HRI: human-robot interaction; NASA-TLX: National Aeronautics and Space Administration Task Load Index.

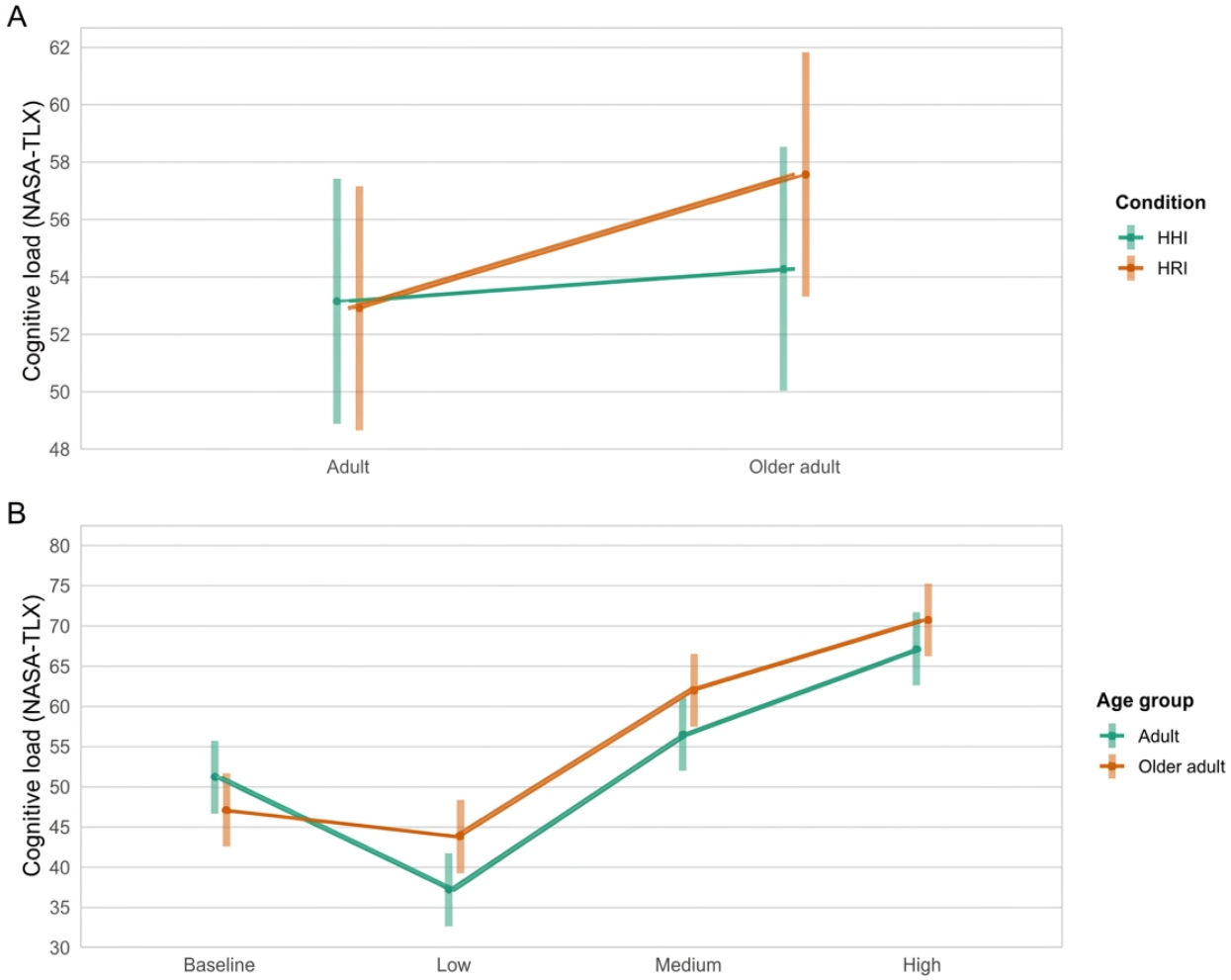

For the interaction between age group and CL level, post hoc analysis revealed that older adults reported significantly higher total scores at the low CL level (mean 43.8, SE 2.28) compared with younger adults (mean 37.2, SE 2.29; M_{dif}=−6.63, SE=3.23; $t_{89.0}$=−2.05; P=.04), suggesting that older adults perceive more CL than younger adults under low-load conditions (Figure 6B).

Analysis of the physiological measure of salivary cortisol concentration revealed a significant interaction between age group, condition, and time (pre vs post; $F_{1, 156}$=5.16; P=.02). When analyzed separately by age group, older adults demonstrated a significant interaction between condition and time ($F_{1, 156}$=10.32; P=.001), whereas younger adults showed no significant effect ($F_{1, 156}$=0.01; P=.90). Post hoc comparisons indicated that older adults exhibited a significant increase in cortisol concentrations from baseline to postintervention during the HRI condition (M_{dif}=−5.65, SE=1.31; t_{156}=−4.31; P=.001; Figure 7).

XSL•FO
RenderX

Figure 7. Salivary cortisol concentrations (nmol/L) at pre- and post-HHI and HRI, divided by age groups. Older adults showed an increase of cortisol after the robot condition. HHI: human-human interaction; HRI: human-robot interaction.

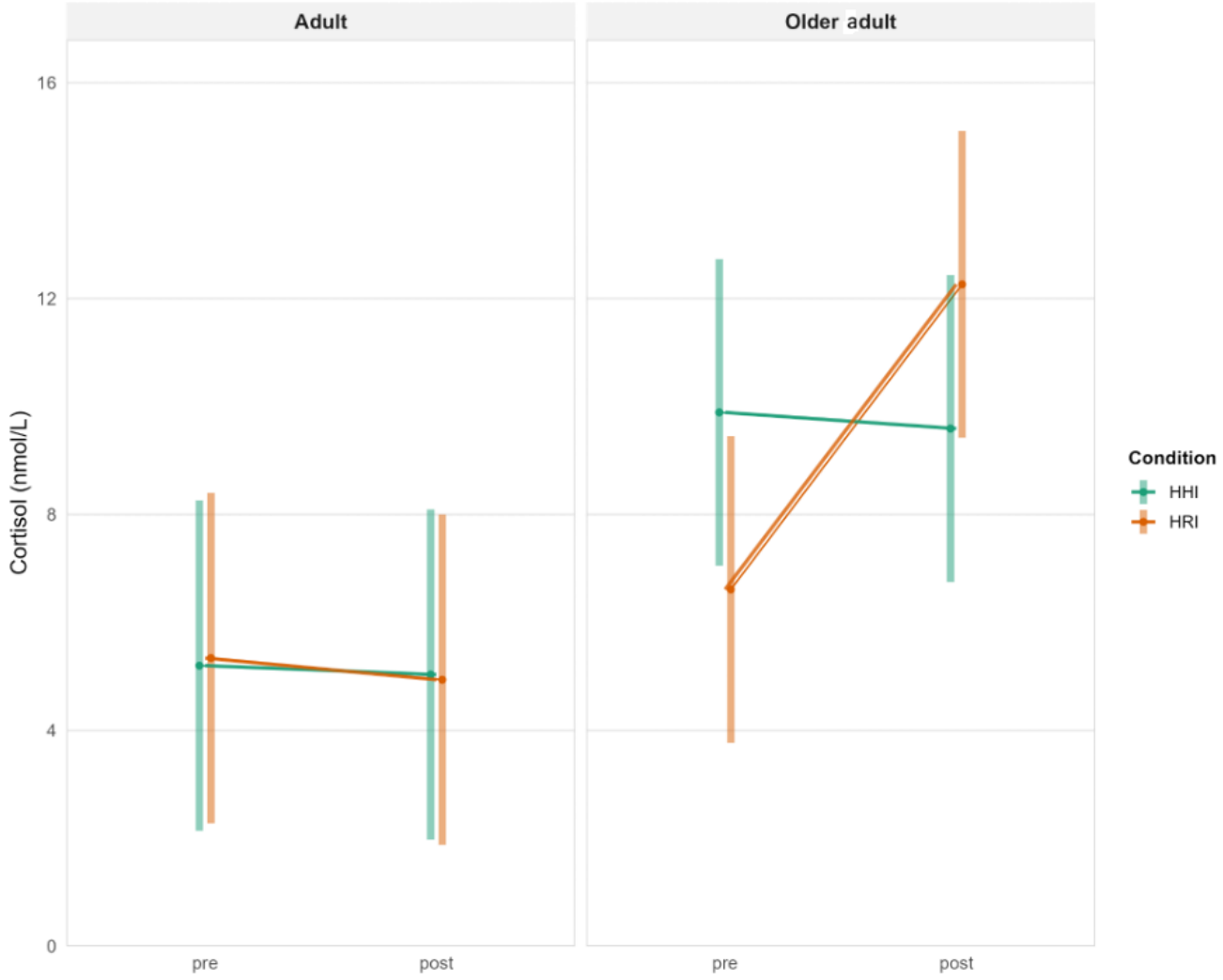

Discussion

Principal Findings

This study aimed to investigate how HHI and HRI support the management of CL during task performance in younger and older adults, integrating behavioral, subjective, and physiological measures. By including a baseline condition without assistance and systematically varying CL levels, this research extends the literature on cognitive ergonomics in aging and technology-assisted contexts by providing an applied evaluation of assistive support modalities that are increasingly proposed as digital health solutions for aging populations. Importantly, these findings should also be interpreted within a translational digital health perspective. While assistive technologies are frequently evaluated in controlled experimental settings, their successful implementation depends on sustained usability, user acceptance, and ecological integration into daily routines. In this context, increases in CL or physiological load, even when accompanied by performance improvements, may represent critical barriers to long-term engagement and adherence. Consequently, evaluating assistive systems through a multidimensional framework that captures both performance outcomes and user burden is essential to inform scalable and sustainable deployment strategies.

Behavioral Measures

Consistent with previous evidence indicating age-related declines in processing speed and working memory capacity [26,27], older adults demonstrated lower accuracy and required more time to complete tasks as CL increased. These differences were most pronounced at medium- and high-load levels, supporting theoretical frameworks positing that increased task complexity disproportionately affects older adults due to age-related cognitive changes [28]. Notably, at low CL levels, accuracy differences between age groups were reduced, suggesting that minimal task demands may better align with preserved cognitive capacities in older adults, enabling performance comparable to that of younger adults. Importantly, these findings reinforce the need to evaluate assistive technologies for older adults not only under simplified or optimal conditions, but also across varying levels of cognitive demand that more closely reflect real-world task complexity.

The interaction between type of assistance and CL revealed that HHI generally supported higher accuracy compared with HRI, particularly at low- and medium-load levels. This finding aligns with research suggesting that human support can more flexibly and adaptively accommodate the nuanced needs of older adults, thereby reducing extraneous CL [29]. The absence of significant

XSL•FO
RenderX

accuracy differences between HRI and HHI at high CL may tentatively reflect a ceiling effect, whereby performance approached an upper limit, reducing sensitivity to differences between assistance modalities. However, this interpretation should be treated with caution, as performance saturation was not formally tested in this study.

Subjective Measures

Subjective measures of PCL, assessed via NASA-TLX scores, indicated that older adults consistently reported higher CL during HRI compared with HHI, particularly under low-load conditions. This finding is especially relevant from a usability and adoption perspective, as increased perceived workload under otherwise manageable task demands may negatively influence trust, acceptance, and sustained engagement with assistive technologies. This result suggests that even minimal interactions with robotic agents can impose additional extraneous cognitive demands for older adults, potentially due to challenges in speech recognition, turn-taking, or the social-cognitive demands of interpreting robotic behavior [30]. It is important to acknowledge that certain technical constraints are inherent to current HRI systems. Features such as response latency, variability in speech recognition, and predefined interaction pacing represent realistic operational characteristics of SARs rather than limitations of the experimental design. These factors may require users to allocate additional attentional and executive resources to manage turn-taking and system feedback, thereby contributing to increased extraneous CL. Such demands may be particularly relevant for older adults, who may be more sensitive to temporal unpredictability and reduced interactional flexibility. Recognizing these constraints as ecological features of present technologies highlights key targets for future system refinement, including more robust speech processing and adaptive interaction pacing, which may ultimately improve usability and reduce cognitive burden. By contrast, younger adults showed no differences in PCL between assistance types, suggesting that HRI may be more readily accommodated by individuals with higher baseline cognitive flexibility and greater familiarity with technology [31].

Physiological Measures

Physiological measures of salivary cortisol concentrations corroborated these findings, revealing increased cognitive strain during HRI for older adults, with significant postintervention elevations observed in the HRI condition. This is consistent with prior literature demonstrating heightened physiological reactivity in older adults to tasks perceived as challenging or unfamiliar [32]. The convergence of behavioral, subjective, and physiological indicators strengthens the interpretation that HRI, in its current form, may elicit latent stress responses in older users that are not fully captured by performance metrics alone, underscoring the value of multimodal assessment frameworks in digital health evaluation. Indeed, the use of the total NASA-TLX score allowed the capture of participants' global perception of workload across interaction modalities, providing an ecologically valid estimate of overall task burden. Complementarily, cortisol responses were interpreted as markers of physiological stress associated with cognitive effort rather than direct indicators of CL per se. The consistency of these measures supports a multidimensional interpretation of user strain during assistive interaction. The absence of similar effects in younger adults underscores the differential impact of HRI on cognitive stress across age groups and highlights the importance of integrating physiological, subjective, and performance-based measures to comprehensively assess CL.

Future Perspectives and Limitations

Together, these findings highlight the complex interplay among aging, CL management, and the type of assistance provided during task performance. Previous research on behavioral and cognitive-affective regulation has similarly emphasized the need for individualized and context-sensitive approaches in psychological assessment and intervention [33,34]. Within the context of assistive digital technologies, these results suggest that "one-size-fits-all" interaction models are unlikely to meet the needs of older adults with heterogeneous cognitive profiles. While robotic assistance offers scalable and consistent support, its current implementation may inadvertently increase extraneous CL for older adults, particularly in tasks requiring social-cognitive engagement. In contrast, human assistance appears to mitigate these challenges, supporting better performance and lower perceived strain under comparable task demands.

From a digital health design and deployment perspective, these findings suggest that SAR systems should prioritize the minimization of extraneous CL through simplified interaction structures, improved speech recognition robustness, and adaptive pacing strategies that respond to individual user performance and stress indicators. Incorporating real-time or near–real-time indicators of CL could enable dynamic adjustment of assistance, reducing unnecessary cognitive strain and improving user experience. Moreover, hybrid models that combine robotic assistance with intermittent human support may represent a pragmatic transitional approach, particularly in early stages of adoption or in cognitively demanding tasks. These considerations are particularly relevant for large-scale implementation strategies, where technologies must remain cognitively sustainable across prolonged and repeated use in real-world environments.

Beyond implications for system design and deployment, this study also offers methodological insights that strengthen the interpretation of technology-assisted CL effects. Importantly, the inclusion of an independent baseline condition in this study strengthens the interpretation of the observed effects by enabling direct comparison between assisted and unassisted task performance. This design feature allows differentiation between genuine support-related benefits and task facilitation effects that may arise simply from external guidance, irrespective of the modality of assistance. By contextualizing both human and robotic assistance against independent performance, the findings demonstrate that while assistance generally improves task outcomes, the cognitive and experiential costs associated with different support modalities vary substantially, particularly for older adults. An additional methodological strength is that the human-assisted condition was operationalized through a standardized script matched to the robot-assisted protocol in wording, sequencing, and load manipulation. This design choice

XSL•FO
RenderX

reduced the risk that differences between conditions were attributable to variability in human delivery style rather than to the assistance modality itself. This comparative framework enhances internal validity and provides a more robust basis for evaluating assistive technologies as potential substitutes or complements to human support in real-world digital health contexts.

The observed interaction patterns suggest that technological interventions designed for older adults should prioritize reducing extraneous CL by simplifying robot interfaces, improving speech recognition accuracy, and implementing adaptive pacing tailored to individual users. Indeed, although older adults view social robots as a potential tool to support their daily activities, our results further highlight the importance of designing such systems according to users' specific needs [5]. Additionally, multimodal support strategies, including visual cues and simplified verbal prompts, may enhance the usability and effectiveness of HRI by aligning it more closely with the cognitive profiles of older adults.

This study suffers from some limitations. For instance, the generalizability of these findings. Although the inclusion of healthy younger and older adults allowed us to examine age-related differences under controlled conditions, the results cannot be directly generalized to more vulnerable groups, such as individuals with mild cognitive impairment or other clinical conditions for whom SARs may be particularly relevant. Future studies should therefore replicate this design in clinical and functionally vulnerable populations to determine whether the observed patterns of behavioral, subjective, and physiological response are preserved, attenuated, or amplified in contexts of greater cognitive vulnerability. Then, the seven task versions were not formally tested for equivalence in difficulty. Accordingly, they should not be interpreted as psychometrically validated parallel forms of the TMT. Although random assignment of versions across participants and conditions was used to reduce systematic version effects and support internal validity, residual differences in difficulty between task versions cannot be entirely excluded. Future studies should include formal equivalence testing of task variants or adopt a fully validated parallel-form procedure if the task is to be used repeatedly across conditions. A further methodological consideration concerns the fixed progression of CL levels within each assisted block. Although this approach ensured a controlled and gradual increase in task demands and supported procedural consistency across participants, it does not allow complete separation of CL effects from potential sequence effects. Future studies may benefit from counterbalancing or randomizing load sequences, although such designs should carefully consider the potential impact of increased task-switching demands, particularly in older populations. An additional limitation concerns the interpretation of the nonsignificant differences in accuracy between conditions at high CL. Although a ceiling effect may provide a plausible explanation, this possibility was not directly tested through specific analyses of performance saturation. As such, this interpretation should be considered tentative, and future studies should include measures or analytical approaches specifically designed to detect ceiling effects. Moreover, the sequential structure of the assisted blocks should be considered when interpreting the cortisol findings. The order of robot-assisted and human-assisted conditions was fully counterbalanced, which reduced the likelihood that the results were driven by presentation order. However, because the precondition cortisol sample for the second block was collected immediately after completion of the first assisted condition, a residual carryover effect cannot be completely excluded. This limitation is particularly relevant for salivary cortisol, given its temporal dynamics and slower recovery profile. Future studies may benefit from longer recovery intervals or separate-session administration to further reduce potential carryover effects while preserving the advantages of counterbalanced designs. Another important methodological consideration is that salivary cortisol was assessed only in the assisted conditions and at the block level, whereas behavioral and subjective outcomes were collected across all load levels, including baseline. Consequently, the physiological findings should be interpreted as reflecting differences between robot-assisted and human-assisted interaction, rather than as load-specific effects or as changes relative to the independent baseline condition. This design choice was consistent with the intended use of cortisol as a cumulative marker of stress-related arousal; however, it also limits direct comparison between physiological and other outcome domains and should be considered when interpreting the multimodal findings.

Future research should examine longitudinal exposure to robotic systems to determine whether increased familiarity reduces cognitive and physiological strain in older adults, potentially enhancing acceptance and efficacy. Such work is essential to distinguish short-term novelty or learning effects from stable interaction patterns that are likely to emerge during real-world deployment. Expanding physiological monitoring to include measures such as heart rate variability and electroencephalography could further elucidate the real-time cognitive dynamics associated with HRI and HHI across age groups.

Conclusions

In conclusion, this study provides important evidence regarding CL management in aging populations within assisted task contexts, emphasizing the need to tailor support strategies to individual cognitive capacities. While HRI holds promise for facilitating aging-in-place initiatives, its effectiveness depends on careful interaction design and adaptive support mechanisms that ensure robotic assistance reduces, rather than amplifies, cognitive and physiological burden. Optimizing these systems is critical to supporting autonomy, usability, and quality of life in the context of active aging. Although both HHI and HRI improved performance compared with independent task completion, the effectiveness of assistance was strongly dependent on the type and complexity of support, particularly in older adults. Human assistance consistently supported higher accuracy and lower perceived workload, whereas robotic assistance, despite its promise for scalable and consistent support, was associated with increased perceived and physiological CL in older adults, especially under low and medium task demands. Overall, these findings highlight the importance of designing adaptive, age-sensitive digital assistive systems that minimize cognitive burden through simplified

XSL•FO
RenderX

interaction, responsive pacing, and multimodal support. Importantly, even when assistive technologies improve task performance, increases in CL or physiological load may limit usability, scalability, and sustainable real-world deployment in aging populations. By demonstrating the value of multimodal CL assessment and tailored support strategies, this study provides actionable evidence to guide the design, evaluation, and deployment of assistive technologies that genuinely promote autonomy and quality of life in the context of active aging.

Acknowledgments

The authors would like to acknowledge all the participants who voluntarily took part in the study.

Funding

This work was supported by the Engineering and Physical Sciences Research Council and National Institute for Health and Care Research funds (grant number EP/W031809/1, IMACTIVE).

Data Availability

The data presented in this study are available at [35].

Authors' Contributions

Conceptualization: DM, ADN, JH, and SC
Methodology: SV and DM
Software: SV
Formal analysis: RV
Data curation: SV
Writing - original draft preparation: SV, RV, and NC
Writing - review and editing: JH, SC, ADN, and DM
Supervision: DM and SC
Funding acquisition: DM and ADN

Conflicts of Interest

None declared.

References

1. Illario M, Vollenbroek-Hutten MMR, Molloy DW, Menditto E, Iaccarino G, Eklund P. Active and healthy ageing and independent living 2016. J Aging Res 2016;2016:8062079. [doi: 10.1155/2016/8062079] [Medline: 27818798]
2. Vollset SE, Ababneh HS, Abate YH, et al. Burden of disease scenarios for 204 countries and territories, 2022–2050: a forecasting analysis for the Global Burden of Disease Study 2021. The Lancet 2024 May;403(10440):2204-2256. [doi: 10.1016/S0140-6736(24)00685-8]
3. Di Nuovo A, Broz F, Cavallo F, Dario P. New frontiers of service robotics for active and healthy ageing. Int J of Soc Robotics 2016 Jun;8(3):353-354. [doi: 10.1007/s12369-016-0350-2]
4. Karami V, Yaffe MJ, Gore G, Moon A, Abbasgholizadeh Rahimi S. Socially assistive robots for patients with Alzheimer's disease: a scoping review. Arch Gerontol Geriatr 2024 Aug;123:105409. [doi: 10.1016/j.archger.2024.105409] [Medline: 38565072]
5. Vagnetti R, Camp N, Story M, et al. Social robots and sensors for enhanced aging at home: mixed methods study with a focus on mobility and socioeconomic factors. JMIR Aging 2024 Nov 25;7:e63092. [doi: 10.2196/63092] [Medline: 39586076]
6. Luo C, Yuan R, Mao B, Liu Q, Wang W, He Y. Technology acceptance of socially assistive robots among older adults and the factors influencing it: a meta-analysis. J Appl Gerontol 2024 Feb;43(2):115-128. [doi: 10.1177/07334648231202669] [Medline: 37750859]
7. Nebe K, Heimgärtner R. Ergonomic principles in designing assistive systems. Presented at: Universal Access in Human-Computer Interaction: 18th International Conference, UAHCI 2024, Held as Part of the 26th HCI International Conference, HCII 2024; Jun 29 to Jul 4, 2024; Washington, DC p. 75-87. [doi: 10.1007/978-3-031-60875-9_6]
8. Rajavenkatanarayanan A, Nambiappan HR, Kyrarini M, Makedon F. Towards a real-time cognitive load assessment system for industrial human-robot cooperation. Presented at: 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN); Aug 31-4, 2020. [doi: 10.1109/RO-MAN47096.2020.9223531]
9. Orru G, Longo L. The evolution of cognitive load theory and the measurement of its intrinsic, extraneous and germane loads: a review. In: Communications in Computer and Information Science 2019. [doi: 10.1007/978-3-030-14273-5_3]

XSL•FO
RenderX

10. Ghanbari S, Haghani F, Barekatain M, Jamali A. A systematized review of cognitive load theory in health sciences education and a perspective from cognitive neuroscience. J Educ Health Promot 2020;9:176. [doi: 10.4103/jehp.jehp_643_19] [Medline: 32953905]
11. Choi HH, van Merriënboer JJG, Paas F. Effects of the physical environment on cognitive load and learning: towards a new model of cognitive load. Educ Psychol Rev 2014 Jun;26(2):225-244. [doi: 10.1007/s10648-014-9262-6]
12. Kalyuga S. Cognitive load theory: how many types of load does it really need? Educ Psychol Rev 2011 Mar;23(1):1-19. [doi: 10.1007/s10648-010-9150-7]
13. Korbach A, Brünken R, Park B. Differentiating different types of cognitive load: a comparison of different measures. Educ Psychol Rev 2018 Jun;30(2):503-529. [doi: 10.1007/s10648-017-9404-8]
14. MacPherson MK. Cognitive load affects speech motor performance differently in older and younger adults. J Speech Lang Hear Res 2019 May 21;62(5):1258-1277. [doi: 10.1044/2018_JSLHR-S-17-0222] [Medline: 31051090]
15. Montine TJ, Bukhari SA, White LR. Cognitive impairment in older adults and therapeutic strategies. Pharmacol Rev 2021 Jan;73(1):152-162. [doi: 10.1124/pharmrev.120.000031] [Medline: 33298513]
16. Bertolazzi A, Quaglia V, Bongelli R. Barriers and facilitators to health technology adoption by older adults with chronic diseases: an integrative systematic review. BMC Public Health 2024 Feb 16;24(1):506. [doi: 10.1186/s12889-024-18036-5] [Medline: 38365698]
17. Schroeder T, Dodds L, Georgiou A, Gewald H, Siette J. Older adults and new technology: mapping review of the factors associated with older adults' intention to adopt digital technologies. JMIR Aging 2023 May 16;6:e44564. [doi: 10.2196/44564] [Medline: 37191976]
18. Ghorayeb A, Comber R, Gooberman-Hill R. Older adults' perspectives of smart home technology: Are we developing the technology that older people want? Int J Hum Comput Stud 2021 Mar;147:102571. [doi: 10.1016/j.ijhcs.2020.102571]
19. Varrasi S, Vagnetti R, Camp N, et al. How human–robot interaction can influence task performance and perceived cognitive load at different support conditions. Information 2025;16(5):374. [doi: 10.3390/info16050374]
20. Korbach A, Brünken R, Park B. Measurement of cognitive load in multimedia learning: a comparison of different objective measures. Instr Sci 2017 Aug;45(4):515-536. [doi: 10.1007/s11251-017-9413-5]
21. Ayres P. Subjective measures of cognitive load: what can they reliably measure? In: Cognitive Load Measurement and Application: A Theoretical Framework for Meaningful Research and Practice 2017. [doi: 10.4324/9781315296258-2]
22. Haapalainen E, Kim S, Forlizzi JF, Dey AK. Psycho-physiological measures for assessing cognitive load. 2010 Sep 26 Presented at: Ubicomp '10; Sep 26-29, 2010 URL: https://dl.acm.org/doi/proceedings/10.1145/1864349 [doi: 10.1145/1864349.1864395]
23. Bowie CR, Harvey PD. Administration and interpretation of the Trail Making Test. Nat Protoc 2006;1(5):2277-2281. [doi: 10.1038/nprot.2006.390] [Medline: 17406468]
24. Hart SG. NASA-Task Load Index (NASA-TLX); 20 years later. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 2006 Oct;50(9):904-908. [doi: 10.1177/154193120605000909]
25. Devos H, Gustafson K, Ahmadnezhad P, et al. Psychometric properties of NASA-TLX and index of cognitive activity as measures of cognitive workload in older adults. Brain Sci 2020 Dec 16;10(12):994. [doi: 10.3390/brainsci10120994] [Medline: 33339224]
26. Salthouse TA. The processing-speed theory of adult age differences in cognition. Psychol Rev 1996 Jul;103(3):403-428. [doi: 10.1037/0033-295x.103.3.403] [Medline: 8759042]
27. Park DC, Smith AD, Lautenschlager G, et al. Mediators of long-term memory performance across the life span. Psychol Aging 1996;11(4):621-637. [doi: 10.1037/0882-7974.11.4.621]
28. Craik FIM, Bialystok E. Cognition through the lifespan: mechanisms of change. Trends Cogn Sci 2006 Mar;10(3):131-138. [doi: 10.1016/j.tics.2006.01.007] [Medline: 16460992]
29. Kushniruk AW, Patel VL. Cognitive and usability engineering methods for the evaluation of clinical information systems. J Biomed Inform 2004 Feb;37(1):56-76. [doi: 10.1016/j.jbi.2004.01.003] [Medline: 15016386]
30. Broadbent E, Stafford R, MacDonald B. Acceptance of healthcare robots for the older population: review and future directions. Int J of Soc Robotics 2009 Nov;1(4):319-330. [doi: 10.1007/s12369-009-0030-6]
31. Czaja SJ, Charness N, Fisk AD, et al. Factors predicting the use of technology: findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE). Psychol Aging 2006 Jun;21(2):333-352. [doi: 10.1037/0882-7974.21.2.333] [Medline: 16768579]
32. Lupien SJ, Maheu F, Tu M, Fiocco A, Schramek TE. The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition. Brain Cogn 2007 Dec;65(3):209-237. [doi: 10.1016/j.bandc.2007.02.007] [Medline: 17466428]
33. Castellano S, Platania GA, Varrasi S, Pirrone C, Di Nuovo S. Assessment tools for risky behaviors: psychology and health. Health Psychol Res 2020 Oct 5;8(2):9235. [doi: 10.4081/hpr.2020.9235] [Medline: 33123648]
34. Sarti P, Colliva C, Varrasi S, et al. A network study to differentiate suicide attempt risk profiles in male and female patients with major depressive disorder. Clin Psychol Psychother 2024;31(1):e2924. [doi: 10.1002/cpp.2924] [Medline: 37922512]

XSL•FO
RenderX

35. Magistro D, Varrasi S, Vagnetti R, Hough J. Comparing human and robot assistance on cognitive load in younger and older adults: behavioural, subjective, and physiological evidence. Zenodo. URL: https://doi.org/10.5281/zenodo.18895458 [accessed 2026-05-14]

Abbreviations

CL: cognitive load
HHI: human-human interaction
HRI: human-robot interaction
M_{dif}: mean difference
NASA-TLX: National Aeronautics and Space Administration Task Load Index
PCL: perceived cognitive load
SAR: socially assistive robotics
TMT: Trail Making Test

Edited by M Balcarras; submitted 06.Mar.2026; peer-reviewed by G Pesimena, S Massimino; revised version received 09.Apr.2026; accepted 09.Apr.2026; published 01.Jun.2026.

Please cite as:
Varrasi S, Vagnetti R, Camp N, Hough J, Di Nuovo A, Castellano S, Magistro D
Human and Robot Assistance for Cognitive Load in Younger and Older Adults: Multimodal Within-Subject Experimental Study
J Med Internet Res 2026;28:e94738
URL: https://www.jmir.org/2026/1/e94738
doi:10.2196/94738

© Simone Varrasi, Roberto Vagnetti, Nicola Camp, John Hough, Alessandro Di Nuovo, Sabrina Castellano, Daniele Magistro. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 1.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating Nursing Work Systems and Identifying Barriers for Robotic Technology Integration: Observational Study

Gina L Georgadarellis[1,2], MS; Ellen Benjamin[3,4], RN, PhD; Shannon C Roberts[1], PhD; Cidalia J Vital[4,5], RN, PhD; Frank C Sup IV[1,2], PhD

[1]Mechanical and Industrial Engineering, Riccio College of Engineering, University of Massachusetts Amherst, Engineering Laboratory, 160 Governors Drive, Amherst, MA, United States

[2]Elaine Marieb Center for Nursing and Engineering Innovation, University of Massachusetts Amherst, Amherst, MA, United States

[3]Donna M and Robert J Manning College of Nursing and Health Sciences, University of Massachusetts Boston, Boston, MA, United States

[4]Baystate Health, Springfield, MA, United States

[5]Elaine Marieb College of Nursing, University of Massachusetts Amherst, Amherst, MA, United States

Corresponding Author:
Frank C Sup IV, PhD
Mechanical and Industrial Engineering, Riccio College of Engineering, University of Massachusetts Amherst, Engineering Laboratory, 160 Governors Drive, Amherst, MA, United States

Abstract

Background: Robotic technology has the potential to assist nurses, but the complexity and unpredictability of health care environments cannot be replicated in a laboratory setting. Furthermore, there is a lack of experiential evidence that robotic technology will meaningfully impact nursing. Collaborative development of technology and real-world usability studies offers the ability to address problems early in the design process when functional changes can be implemented.

Objective: The purpose of this study was to use an observational study and systematically evaluate the work system of inpatient nurses to identify barriers to the integration of robotic technology. The objectives are to use an observational study of active hospital units to gain a deeper understanding of nursing tasks, workflow, and the health care setting; identify barriers to the integration of robotic technology using the people, environment, tools, and tasks (PETT) scan from the Systems Engineering Initiative for Patient Safety framework; and synthesize the work system components of the PETT scan into themes.

Methods: We used the practice-oriented model of the Systems Engineering Initiative for Patient Safety, the PETT scan, to identify barriers for robotic technology use and innovation. A convenience sample of nursing staff was observed as they worked. Units included the emergency department, medical and surgical intensive care unit, preop or postanesthesia care unit, and general medical-surgical floor. The total number of observation hours per unit was based on data saturation, which occurred at variable times during the day shift, and was arranged with unit management. A total of 53 hours across 16 sessions were recorded. Multiple rounds of inductive and deductive coding were conducted. Briefly, a 3-phase iterative data analysis process was used—initial inductive content analysis, a deductive phase to organize emergent categories into a PETT scan, and finalization of the PETT scan with the identification of overarching themes.

Results: Observations across all units yielded a broad set of barriers to integrating robotic and other health care technologies. Using the PETT scan, 78 barriers were identified and were summarized into 20 themes with supporting subthemes and exemplars.

Conclusions: By systematically observing nursing workflows and synthesizing barriers into themes, this study provides new insight into the conditions that enable or constrain robotic integration. Findings suggest that robotic technologies are presently best suited for auxiliary and background roles. Broader integration into patient care workflows will depend on designs that align with clinical workflows, support interoperability and robustness, and address ethical, accountability, and coordination challenges inherent in nursing care, as well as maintained organizational support.

(J Med Internet Res 2026;28:e89409) doi:10.2196/89409

KEYWORDS

technology; healthcare; robotics; nursing; human factors; ergonomics; workflow; hospital units; patient safety

XSL•FO
RenderX

Introduction

Background

The health care system faces a rise in complex medical diagnoses, an increasing life expectancy of patients, and a critical shortage of nurses [1]. Nursing shortages are of particular concern, as inadequate staffing is a primary cause of missed and delayed nursing care [2]. Missed nursing care is associated with decreased nurse-reported care quality, increased length of patient hospital stay, higher patient mortality rates, decreased job satisfaction, and increased intent to leave the profession [3,4].

Efforts to combat missed nursing care and safe staffing are ongoing [5,6], but the complexity of real-world settings has limited the successful translation of research into policy and practice changes [7,8]. One potential opportunity to address nursing shortages and missed care is through the integration of robotic technologies into health care settings. Research suggests that nurses are receptive to the implementation of robotic technology to lessen their workload, improve nurse workflow, improve task prioritization, mitigate short staffing, and reduce burnout [9]. However, more work is needed to understand the barriers to successfully translating health care robotic integration in real-world practice.

The usage of robots in the health care setting is not novel; more than 3 decades have passed since medical robots were first developed to assist surgeons [8,10-12]. In fact, the first surgical robot cleared by the Food and Drug Administration appeared in 1993 with the Automated Endoscopic System for Optimal Positioning (Computer Motion Inc) [13]. While technically a class 2 laparoscopic device, the Automated Endoscopic System for Optimal Positioning allowed surgeons to control the orientation of a traditional laparoscope using foot pedals [14]. Since that time, robotic technology has expanded into many sectors, supporting a range of health care professionals. In nursing, automated dispensing cabinets (ADCs) are a prominent example of successful robotic integration into daily workflows. ADCs provide safeguards for nurses to remove the correct medications and have been shown to reduce errors, increase security, and track inventory in real-time [15,16]. Additionally, ADCs can increase work efficiency, offering a demonstrated reduction in the time spent on dispensing and preparing medications by an average of 32 minutes per 8-hour shift [17].

Mobile robots have also been shown to benefit nursing workflows by reducing medication dispensing and delivery times [18]. In a 2011 study, an academic medical center saw reductions in time from order to label printing, order preparation, and from medication checking to delivery when using the TUG Automated Robotic Delivery System (Aethon Inc) [19]. A decade later, another pilot study spanning 3 different academic and teaching facilities investigated the potential use of Moxi to support nursing and nonclinical staff (Diligent Robotics Inc). Moxi, classified as a delivery robot, is a mobile robot capable of tasks that include carrying patient supplies and delivering laboratory samples or medications [20]. Across 4 weeks, 1 site achieved a significant overall growth in average deliveries and a 100% compliance between patients' fall-risk signage and their electronically documented fall-risk status [21]. Another pilot study found that medication delivery with Moxi was 4 times faster than anecdotal evidence provided by staff [22]. Quantitatively, Moxi had traveled 298 miles at the time of reporting, averaging more than 15,750 steps per day.

There is now a substantial body of research investigating the integration of robotic technology into nursing practice. A recent umbrella review found 13 reviews, representing 558 studies, exploring this topic [23]. Technology-supported automation offers several potential benefits, including support for data entry and information exchange, medication management and administration, patient data tracking and monitoring, infection control, and mobility or lift assistance [24]. Robotic technology is most promising when it supports tasks that are routine, repetitive, or high-risk [24]. Nurses themselves hold largely positive opinions of robotic assistance [23], perceiving a range of benefits that include improved efficiency, error reduction, reduced physical and cognitive workloads, and decreased burnout [9,24].

Despite its potential advantages, experiential evidence from real-world implementation of nursing robotic assistance remains limited, and few rigorous studies are exploring this topic [23]. This dearth of real-world integration may be due, in part, to limited nurse-specific collaborative robotic technologies. Most health care collaborative robots are designed for patients, not nurses, and despite their commercial availability in the United States, peer-reviewed scientific evidence on their usage and effectiveness is lacking [25].

Even where hospitals successfully adopt health care technologies, research evaluating their use and usability faces significant barriers, including difficulty recruiting participation by end users, time constraints, budgetary constraints, and ethical barriers [26]. Additional obstacles arise from the logistics of conducting on-site research, which often require establishing a research agreement between the hospital and the research institution, obtaining research personnel clearance involving background checks, up-to-date vaccination records, demonstrating human participants research ethics training, and obtaining approval from the institutional ethics board.

Nevertheless, research in naturalistic settings is necessary to deeply understand and advance health care technology. The complexity and unpredictability of health care environments are not easily replicated in laboratory settings. Collaborative technology development and real-world usability studies enable addressing problems early in the design process, when functional changes can be implemented, and benefits observed and documented. It is essential to work iteratively with nurses within the context of their work system and develop a richer understanding of authentic nursing workflows. Therefore, the purpose of this study was to systematically evaluate the work system of inpatient nurses to identify barriers to the integration of robotic technology.

Study Definitions

The Food and Drug Administration does not provide a formal definition for the term "robot," and there is still no universal or standard definition for robotic technology [27]. Additionally,

XSL•FO
RenderX

the definition for medical devices is defined under Section 201(h) of the Food, Drug, and Cosmetic Act [28] is relatively broad and encompasses devices ranging in complexity from a bandage to a surgical robotic platform. Thus, it is important to explicitly define robotic technology in the context of health care to ensure understanding, especially among those who specialize in different disciplines or lack experiential exposure. The definitions used in this study are provided in Multimedia Appendix 1, and include medical device [28], robotic technology, robot [29], health care robot, autonomous technology [13,30], and sense-think-act paradigm (Multimedia Appendix 1).

Systems Engineering Initiative for Patient Safety: A Practical Framework for Analyzing Work Systems

The Systems Engineering Initiative for Patient Safety (SEIPS) model integrates human factors and health care quality models [31], mainly balance theory [32], work system [33], and Donabedian's [34] system-process-outcomes model of quality. SEIPS was created to address the challenge of applying human factors and systems engineering methodologies to improve patient safety research and the design of health care systems [35]. There have been different iterations of SEIPS since the model's creation, and SEIPS has been implemented within nursing research [36]. SEIPS 101 provides a simplified practice-oriented model with the purpose of acting as a theoretical or practical framework to guide activity [37]. SEIPS 101 retains the 3 major work components (system, processes, and outcomes). The work system includes 4 work system elements—people, environment, tools, and tasks—which form the acronym "PETT." The environment includes physical settings, socio-organizational conditions, and external context (Figure 1).

Figure 1. SEIPS 101 simplified model depicting work component feedback loop, adapted from Holden and Carayon. Work Systems is represented by a square and emphasizes the person as the center of the work system as well as interactions among the components; Processes is shaped as a triangle to convey the flow from left to right; and Outcomes is drawn as a circle, akin to the letter 'O,' to act as a visual reminder.

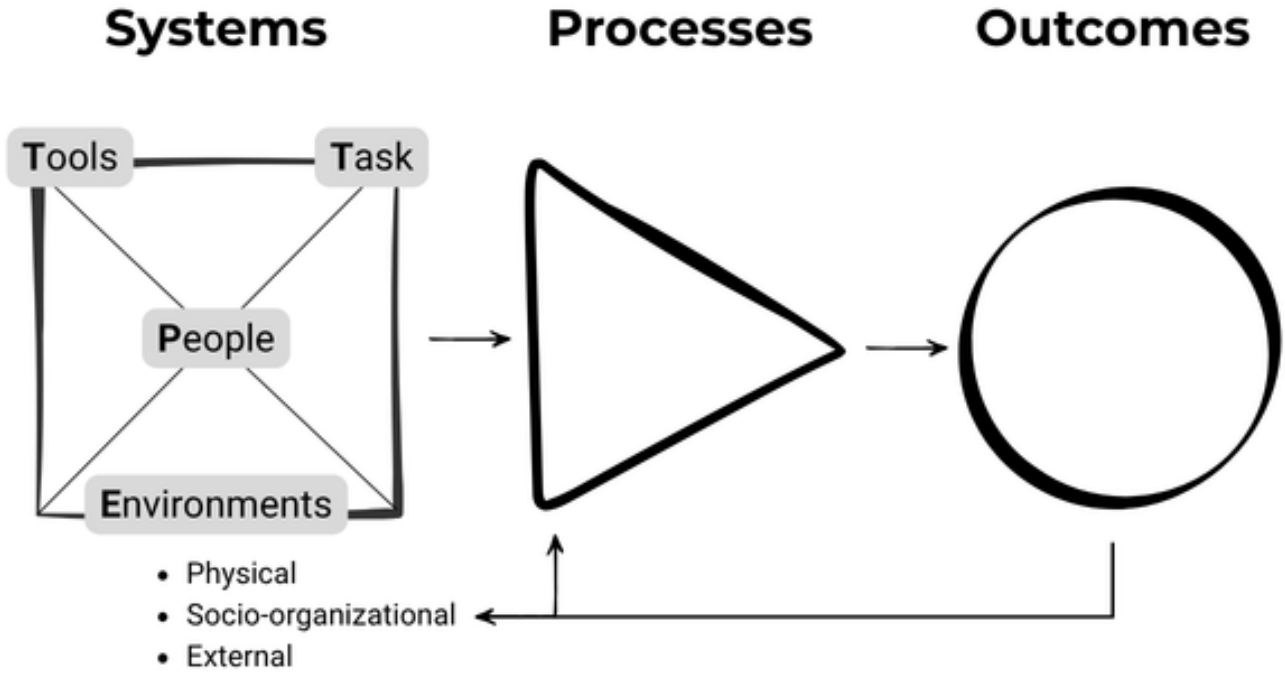

One tool from SEIPS 101 is the PETT scan, a matrix designed to ensure researchers consider the full breadth of the work system, specifically its people, environment, tools, and tasks [37]. Furthermore, traditional PETT scans include the interactions between work systems components. This tool may be used to indicate the aspects that hinder (barriers) or support (facilitators) people in the work system for each work system component or for component-component interactions. The PETT scan may be implemented in project planning or the conduct of research to systematically consider work system factors [37]. Therefore, the purpose of this study is to identify barriers to robotic technology integration within inpatient nursing work systems using a SEIPS analysis.

Methods

Overview

A qualitative observational study was implemented to elucidate the workflows of different units, their unique needs, key resources, and technologies used, and the barriers for robotic technology development or implementation. The original scope of the study was broad; researchers were interested in exploring opportunities for robotic technology usage and characterizing the current state of robotic technology use within acute care settings. Data were primarily collected between April 2023 and August 2023, with supplemental observational periods occurring until October 2024. We used the Standards for Reporting Qualitative Research (SRQR) reporting guideline [38] to draft this paper, and the SRQR reporting checklist [39] when editing (Checklist 1).

XSL•FO
RenderX

Ethical Considerations

Approval for this research was obtained from the Institutional Review Board of the medical center (IRBNet 2008331). An information sheet describing the research study was distributed to unit management and was available at all times during the observation period. The study received a waiver of informed consent to minimize the burden on staff and interference with their work, and was determined to be exempt. To diminish the ability to identify staff, data collection used anonymized fieldnotes. Participants were identified using numbered pseudonyms corresponding to the job role they held at the time of observation (eg, nurse 1 and patient care assistant 1), and numbering was restarted at the start of each observation period. As data collection occurred while participants were working, participants were not compensated for this study.

Setting

This study was conducted on multiple floors of a large, academic medical center in New England. This particular site was chosen due to the presence of multiple unit types, its strong interest in conducting translational research, and its history of successful previous collaborations with the researchers' academic institution. To gather data on nursing workflows on a variety of units, observations were conducted in the emergency department (ED), medical and surgical intensive care units, preop or postanesthesia care unit (PACU), and a general medical-surgical floor.

Participants

A convenience sample of nursing staff, facilitated by unit leadership, was selected from those working during the hours of data collection. The sample included direct care nurses, nurse leaders, patient care technicians, and charge nurses (Multimedia Appendix 2). Direct care staff were identified by their standardized dress attire (royal blue for the hospital units and by their badge for the PACU area). No staff identifiers were collected. Exclusion criteria included staff working with pediatric patients and patients with psychiatric disorders.

Data Collection

Data were collected using overt participant observations. The focus of observation was nursing workflows and nurses' interactions with and manipulations of their physical environment, including health care technology, medical devices, patient care supplies, and medications. Participants were asked brief questions and to think out loud or reflect upon their actions, decisions, strategies, and considerations in order to elucidate any tacit actions and to more deeply understand nurses' behaviors and thought processes. The interactions with participants described above were limited, were not guided by a structured interview guide, aimed to minimize disruptions and distractions to staff work, and did not occur while the staff were working with patients. Additionally, when patient events or interruptions requiring privacy occurred, observation focus was diverted to capture unit details or events occurring in proximity. Unit charge nurses and/or nurse educators were asked questions regarding their thoughts on nursing workflow, decision-making related to technology use, and were asked to provide greater context for events that occurred during the observation. At the end of every observation period, field notes were digitally transcribed by the observer.

For each unit, the total number of observation hours was determined by data saturation (when the observed events and tasks became repetitive). Here, saturation was defined according to the standard of "informational redundancy," when newly collected data does not yield anything new. Saturation was determined during data analysis, when no new barriers were identified, and codebooks were stable [40]. Periods of observation lasted approximately 4 hours each, with a minimum of 3 observation periods at each unit. Observations occurred at various times during the day shift, and scheduling was coordinated with unit management. Additional observations were scheduled based on findings during the initial observation hours.

Observations were conducted by a mechanical engineering researcher with a master's-level education and previous research experience in nurse perceptions of robotic technology in health care and device usability (GLG). PhD-prepared nursing researchers with expertise in acute care and qualitative research (EB and CJV) provided consultation and ongoing support in protocol development, data collection, and data analysis.

Data Analysis

Transcribed field notes were analyzed in NVivo 14 Plus software (Lumivero). A 3-phase iterative data analysis process of inductive and deductive content analysis guided by SEIPS work systems elements was used [41]. The use of both inductive and deductive analysis has been demonstrated in previous SEIPS research and combines the ability to identify emergent findings while applying validated theoretical frameworks [42].

As described above, the original scope of the study was broad, encompassing opportunities, barriers, and the current use of robotic technology by nurses in hospital settings. Data were first analyzed inductively, allowing codes and categories to emerge from the data. The purpose of this approach was to avoid bias from preconceived knowledge and not overlook potentially valuable insights. Descriptive codes were applied to text, and codes were categorized into emergent categories based on their similarities, differences, and patterns [41]. At this point in the analysis, it became apparent that there were few examples of current robotic use and that field notes demonstrated a much greater emphasis on barriers. The focus of analysis was revised to systematically characterize barriers to robotic usage. A second round of inductive analysis was performed with this new analytical focus.

Given the desire to comprehensively explore this phenomenon, the SEIPS model was applied to add robustness and ensure consideration of all work system elements. Codes and categories were analyzed and reorganized using deductive analysis to cluster findings according to the PETT framework. Finally, given the large number of identified barriers, thematic analysis was used to describe higher-order themes and patterns across barriers. In total, 3 investigators (GLG, EB, and CJV) conducted the content analysis, and a fourth investigator (FCS IV) also contributed to the final thematic analysis. Coding and category development were first completed independently by

XSL•FO
RenderX

investigators, who then met regularly to discuss and refine analysis. Consensus was handled through discussions and iterative revisions until the team was satisfied.

Trustworthiness

Efforts to increase trustworthiness were guided by Lincoln and Guba [43]. Strategies implemented included prolonged engagement with data, triangulation across multiple nursing roles and unit types, negative case analysis, and the creation of a rich audit trail using qualitative analysis software. Findings are supported by extensive field note evidence within the manuscript and appendices. The credibility of qualitative analysis was supported by multiple investigators who first coded independently and then reached a consensus. Investigators used memo-writing to increase reflexivity. The use of the SEIPS model added robustness to data analysis and coherence with previous health systems research. Finally, 2 engineering experts (SCR: human factors, FCS IV: mechatronics) were consulted to provide feedback on the PETT scan and to ensure the analysis was credible and well-aligned to a human factors and robotics perspective.

Results

Overview

Observations across 4 hospital units yielded a broad set of 78 barriers to integrating robotic and other health care technologies. Using the PETT scan framework, these barriers were categorized into 4 domains—people, environment, tools, and tasks (refer to “PETT scan” in Multimedia Appendix 3). For each of the barriers in the PETT scan, implications for robotic technology and an example field note are included. Table 1 summarizes these barriers by theme and supporting subthemes.

XSL•FO
RenderX

Table . Summary of the barriers identified using the people, environment, tools, and tasks scan framework from Systems Engineering Initiative for Patient Safety 101.

Work system element, barrier themes	Subthemes
People – patients	
Heterogeneity	• Varying ages, sizes, levels of activity, and physical capabilities • Varying levels of cognition and education • Varying languages, cultures, and religions
Expectation differences	• Patient expectations • Nursing needs and expectations
Unpredictability	• Noncompliance • Negative patient behavior and aggression • Misuse of technology • Unpredictable clinical scenarios
People – staff	
Physical characteristics	• Varying ages, heights, sizes, levels of activity, and physical capabilities • Physically taxing equipment
Cognitive characteristics	• Varying education, experience, training, dialects, and knowledge • Bandwidth and emotional fatigue • Inconsistent nomenclatures • Insufficient training in technology • Drug diversion or theft
Psychosocial characteristics	• Varying motivations, needs, goals, culture, and religions • Personal habits and preferences • Fear of liability • Reluctance to adopt new technology • Satisfaction and frustrations with technology
Workflow considerations	• Trade-offs between cost and speed or convenience • Pacing mismatch between technology and workflow • Shared ownership of supplies and equipment • Failure to return or restock equipment • Lack of lunch breaks and nutrition
Environment	
Cognitive burden of the environment	• Dynamic and variable environments • Overstimulation and distraction • Alarm frequency • Use of visual aids
Physical factors	• Varying temperature and lighting • Presence of liquid reagents and biohazardous materials • Inconsistent unit layouts • Dimensional constraints • Inability to adapt infrastructure
Organizational culture and policy	• Inconsistent policies and unit cultures • Lack of support for nurse-led innovation • Resistance to policy and procedure changes • Regulatory compliance and oversight • Inconsistent staffing roles • Inadequate staffing levels
Accessibility	• Trade-offs between accessibility and organization • Inconvenient storage locations • Considerations for patient visibility versus privacy
Technology or tools	

XSL•FO
RenderX

Work system element, barrier themes	Subthemes
Physical considerations	• Physical size and weight • Need for transportability • Need for high sensitivity and compatibility with personal protective equipment • Need for mobility and concurrent use with other devices • Screen visibility issues • Impeding situational awareness
Speed or time considerations	• Need for high speed • Inconsistent network connectivity • Data transfer and streaming time requirement
Maintenance or support	• Need for robustness against physical damage and compatibility with diverse biological materials • Inconsistent staff training and support; need for intuitiveness • Technology and device maintenance demands staff, time, and money • Frequency of update considerations
Device compatibility	• Lack of backward compatibility and limited interoperability across manufacturers • Incompatible devices should be visually distinct • Lack of standardization across hospitals
Tasks	
Concurrent or competing tasks	• Multitasking and interruptions • Task redundancy • Trade-offs between scheduled and unscheduled tasks
Task prioritization	• Nursing prioritization and care left undone • Technological bias of nursing judgment • Need for patient-centric care requiring patient prioritization
Inconsistencies in task performance	• Inconsistencies in staff performance • Conflicting information • Varying physical properties of materials • Multimodal communication
Ethical considerations	• Need for compassionate care • Subjectivity in care • Inconsistent staff roles and scopes • Task shifting and delegation
Ergonomic considerations	• Repetitive tasks • Tacit and nuanced actions • High dexterity • User-friendly design

People-Related Barriers: Patients

Patient-related barriers arise from patient heterogeneity, differing expectations between patients and staff, and unpredictability in patient behavior. First, patients are widely heterogeneous in their physical, cognitive, and psychosocial characteristics. Robotic technology must function across a wide range of patient sizes, strength levels, and cognitive abilities. For biomedical devices with sensors or that directly interface with patients, there is an even wider range of considerations for safety and the accuracy of biophysical data. In some care settings, such as the ED, technology must serve the full breadth of newborn, pediatric, adult, and geriatric patients.

> *The nurse educator discusses available defibrillators. They had considered purchasing Stryker LifePak, but this only goes down to 2 joules and 1 joule is needed for pediatric patients.* [Field Note June 30, 2023]

Technology must also serve patients with varying cultures, languages, and educational backgrounds. Similarly, patients carry a wide range of attitudes and expectations for medical technology. At times, these patient preferences were observed to conflict with nursing goals, safety, and workflow needs.

> *Stat Nurse1 explains that the unit layout makes workflow difficult. The idea was to keep all operational equipment hidden from families…but from nursing's perspective, the workflow safety is horrendous.* [Field Note, July 25, 2023]

XSL•FO
RenderX

Finally, patients exhibit unpredictability in their behavior and decisions. Observations demonstrated several instances of patient noncompliance or an inability to adhere to nurses' instructions. Equipment and technology must mitigate the risk of patient injury and harm and be resistant to theft, tampering, and patient interference. Technology must further contend with unpredictable decisions, such as patients acting illogically or changing their minds.

> *A patient requested a walker and an escort from their room to the restroom. A patient care technician assisted the patient, and midway through their journey, the patient changed their request, wanting a wheelchair instead after potentially overestimating their mobility.* [Field Note, June 09, 2023]

People-Related Barriers: Staff

Similarly to patients, the heterogeneity of staff members creates barriers for technology use and implementation. Staff members encompass a range of ages, sizes, strengths, physical endurance, and mobility. Physical energy and emotional stamina may also vary between a single individual according to their shift, mood, and their ability to take breaks for food. The act of transferring a patient from their bed to a cat scan stretcher illustrates the physicality of patient care, involving a series of rolling, lifting, reaching, pulling, and pushing actions to move the patient and their equipment.

Staff heterogeneity also includes a range of educational levels, experience, technology training and knowledge, and mental capacities. Individual nurses demonstrate differing levels of focus, motivation, and ability to handle emotional challenges. One notable observation is the lack of consistency in staff nomenclature. Depending on their previous training and knowledge, staff may use different words to describe technology and medical supplies. Often, this nomenclature is inconsistent between staffing roles and with the language used in software systems. In practice, this can create confusion in care, impede supply management, and disrupt nurses' ability to find supplies in stock rooms.

> *The Nurse Educator works with supply management and wants to know what actually needs to be in the supply room. She has a list of everything in the supply room. The problem is that there are things in the room that they have no idea what they are…the barcode labels list items as RackBin using letters and numbers (eg, A1A1).* [Field Note, November 20, 2023]

Just as medical technology should mitigate the risk of patient harm, it must also safeguard against harmful staff decisions. For example, the ADC that stores and dispenses medication is designed to reduce the risks of drug diversion and theft.

> *Stat Nurse1 is having a discussion about narcotics and the scheduled drug check. This is a Friday chore at the Pyxis where certain drugs are checked…it takes about 20 minutes.* [Field Note, May 19, 2023]

Additionally, staff demonstrate a range of preferences, expectations, and habits surrounding the use of technology. Nurses develop their own personal strategies and routines in care. These personal habits may create a reluctance to use technologies or adopt new devices. Across all units, observations revealed that nurses consistently use paper for notetaking rather than relying on electronic documentation. Daily frustrations with technology may also push nurses away from using technology and increase the use of nontechnical workarounds.

> *Nurse1 explained that they were making a task list and going over all of their patients. The task list was on a plain piece of printer paper; paper folded in half and broken up into 6 sections. She wrote 1- or 2-word phrases with check box for each patient.* [Field Note, April 28, 2023]

Finally, health care technology must be compatible with nursing workflows. Observations revealed many instances of poor fit between devices and nursing work, where technology caused significant delays or interruptions to patient care. When faced with obstacles, nurses may prioritize their own convenience or patient safety over considerations of cost or stewardship. Relatedly, staff members were frequently observed failing to restock or return supplies, leading to disorganization and exacerbating the challenge of finding them. One significant barrier is the lack of individual responsibility or accountability for health care technology. Resources are shared between staff members, with transient periods of "ownership" by individuals. This problem is even more pronounced with equipment or devices shared across nursing units.

> *The ED nurse manager reports that cable management is a big issue. The department spent $23,000 tethering all the cables together to prevent nurses from taking cables, but this solution did not last long. After about 14 days, many had been cut by surgical shears. Nurse Educator1 states the issue is that nurses prioritize the patient even if it is at the expense of other nurses.* [Field Note, October 09, 2024]

Environment-Related Barriers

Environmental factors include the cognitive burden produced by the environment, physical factors, organizational culture and policy, and accessibility. First, environments are constantly changing and dynamic. Hospital units are full of distractions and stimulations, which can be exacerbated by health care technology. At one point in time, a nurse charting on a desktop computer may be confronted with a sounding bed alarm, the blinking red LED lights of a call bed, the audible keyboard strokes and mouse clicks, a rapid succession of beeps produced by the centralized monitoring system, and conversations between patients, visitors, and staff. Alarms, in particular, are highly prevalent and cognitively demanding. These alarms and notifications become barriers when staff struggle to distinguish between multiple competing sounds emanating from several devices and patient rooms. Furthermore, the volume and tone of alarms may not correlate to their urgency.

> *An alarm goes off… [a few minutes later] the phone rings, another alarm for the pneumatic tube.* [Field Note, July 25, 2023]

The environment has many physical properties that make technological innovation and usage difficult. Temperature varies

XSL•FO
RenderX

across units, with the operating room notably cold and the emergency trauma room significantly hotter. Unit layouts are also inconsistent. Within the study site, one unit consisted of a long, narrow hallway with a centralized nursing station; one unit was broken into pods and did not have centralized supplies; one unit was U-shaped with resource-specific supplies; and another spanned 2 floors. All units demonstrated challenges with clutter and cramped spaces. Physical mess even included the presence of spills and liquid reagents on the floor, including biological and/or biohazardous materials. Light pollution from both in-room devices and the hallway was consistently present. In the PACU, each patient pod had a thin LED panel from floor to ceiling, emitting light, with access to privacy curtains that were sometimes used to dampen light. Underlying all these physical properties is the reality that, once built, physical layout is not easily altered or adapted to changing technologies or processes.

> *PCA1 changed the trash; left the full trash bag on the floor in the doorway… Gloves were thrown in the trash at some point but missed trash; just lying on floor.* [Field Note, May 19, 2023]

The hospital organizational context may also act as a barrier to technology innovation and usage. Established organizational policies make it difficult to change practices, and organizational cultures may not be conducive to nurse-led innovation. Ingenuity and change are also restricted by the need to adhere to external regulatory bodies. Nurses cited these external regulatory bodies, like the Joint Commission on Accreditation of Healthcare Organizations, as restricting their ability to implement innovative ideas and potential process improvements.

> *A nurse comes in with a story of a missing hearing aid and issues with personal belongings again. They talk with the charge nurse. "Why can't you just change things?" The response is, "because it's a process."* [Field Note, July 06, 2023]

Finally, poor technology accessibility may compromise its usability and successful implementation. Frequently, technology was observed in areas that were not conducive to nursing workflows. Health care staff appeared to grapple with conflicting priorities, recognizing trade-offs between accessibility, speed of care, cleanliness and organization, unit aesthetics, and patient privacy or visibility.

> *Stat Nurse1 needed to grab a stop cock/dead head for Patient2 and went into supply closet to look for it. It took Stat Nurse1 5-10mins total… [and] Stat Nurse1 didn't even find what they were looking for; found something that used a stop cock and planned on taking it from that supply.* [Field Note, May 19, 2023]

Technology and Tool-Related Barriers

Technology and tool barriers included physical considerations, time and speed considerations, maintenance and support considerations, and device compatibility concerns. First, the scale of technology may pose challenges and become a barrier to use and innovation. While large technology takes up storage space and can impede movement, small technology may be hard to locate or be susceptible to theft. Because technology must be moved and transported throughout the hospital, barriers arise when device and equipment size impede portability. This is exacerbated by the fact that a single patient often requires multiple medical devices and equipment. Cables and wires, in particular, were observed causing significant issues, frequently forming tangles, restricting movement, and interfering with nursing care tasks.

> *Nurse1 adjusts the bed for Patient2, changing the arm of the BP cuff and finger of O2 monitor. Wire management appeared a little tricky, as the lines got tangled.* [Field Note, August 22, 2023]

The pace of technological change can also pose significant barriers to its successful integration. As described above, the pace of technological change was found to be impactful on nursing workflows. In addition to delays caused by improperly functioning technology, delays may arise from the inherent time required to transfer data between devices or from network connectivity issues. Wi-Fi connectivity appeared to be unit-dependent; in the ED, a nurse educator mentioned that the unit's Wi-Fi was notoriously problematic due to the architecture and materials of the department. The unit has a "Downtime Cabinet" stocked with paper documentation (eg, rating scales; Situation, Background, Assessment, and Recommendation; discharge instructions; medication orders; and consents) for when nurses are unable to use their computers.

> *Nurse Educator1 mentioned that some things don't work, mentioned Wi-Fi issues specifically. The Wi-Fi doesn't reach the corners.* [Field Note, June 30, 2023]

Hospital technology robustness is also a significant barrier to technology use. Broken or malfunctioning equipment was a common occurrence observed to frequently impact nursing workflow. Often, these problems were not intuitive to fix, or the staff lacked sufficient training and support to correct them. Even when devices work as intended, technology still requires routine maintenance and updates, demanding staff time and expertise. Maintenance can become even more complicated when technology comes into contact with nonsterile biological and environmental materials.

> *Nurse1 was on computer and needed to use handheld scanner, but scanner wasn't acting properly, and she didn't know why it wasn't happy.* [Field Note, May 30, 2023]

Finally, technology use may be compromised by device incompatibility. Health care technology is not standardized across different hospitals or even across hospital units. Therefore, a patient transferred from one location to another may have devices that do not easily integrate with a nurse's workflow. In addition to inconvenience, safety concerns arise when incompatible technologies may appear compatible with inappropriate technologies because they are visually similar. This problem was highlighted by the hospital's ED, which had recently acquired a new defibrillator. Following the implementation of this new device, several aspects of the unit's workflow became infeasible. Previously purchased blood pressure sensors were not the same brand as the new defibrillator and therefore could not be used. The trauma room had a

XSL•FO
RenderX

brand-specific monitor that worked only with the previous brand, but the unit did not intend to replace it due to the cost. The new device was also incompatible with the unit's streaming monitors and other portable defibrillators. Although the cable physically fits in the port of other department devices, it did not function.

Task-Related Barriers

Task factors include concurrent and competing tasks, task prioritization, inconsistencies in task performance, ethical considerations, and ergonomic considerations. Nurses often must manage multiple demands, engage in high levels of multitasking, and adapt to unpredictable events. Staff are also frequently interrupted and switch tasks quickly. Additionally, nurses may engage in redundant tasks, such as manually transcribing information from the computer onto a whiteboard or piece of paper. This highly dynamic and unpredictable workflow may not be reflected in technology's task planning capabilities. In the ED, nurses rely on large paper to document care during highly critical situations because it better supports rapid documentation of multiple, simultaneous assessments, interventions, and narration.

> *The ED has an initial assessment/intervention form… It gets scanned, but the scanning people don't like it due to the large size of the paper. It's 3 pages wide, 1 sheet that is trifolded.* [Field Note, June 30, 2023]

Nurses manage these multiple, competing priorities by prioritizing tasks. Tasks are completed based on nurses' assessments of priority, and many are left undone if they are not deemed important. The subjectivity and unpredictability of task prioritization may impede technology integration. Additionally, technology might shape or bias nursing work processes and influence the tasks that nurses deem to be most pressing.

> *Preparing for a new admission and transfer… Usually, the charge nurse would look up the patient, but today the charge nurse had their own assignments. She talked on the phone very briefly while looking at the information and got a paper and copied information from the computer, noting that the ED didn't give the patient medication so Nurse1 will have to do it.* [Field Note, April 28, 2023]

The execution of nursing tasks is further complicated by inconsistencies in task performance. Tasks may be completed differently by different staff members depending on their role, knowledge, and understanding of current protocols. Tasks may also vary according to the other devices, tools, or medical supplies required. For instance, in observations, units offered multiple modalities for communication, including verbal, visual, phone, portable handheld devices, radio, and electronic communication. Or the task of administering a medication via an infusion pump may vary depending on the drug's viscosity or fluid dynamics. In addition to protocols and procedures, when completing tasks, nurses often rely on their own clinical judgment and ethical considerations. Nursing care and task completion, therefore, can be highly subjective, variable, and even involve actions outside of nurses' traditional role and scope.

> *Nurse1 walked back to Patient1 and then walked with a patient's family member to escort them out of the unit. She said that they don't do that too often, will have a PCT or another worker do it if they're able to, but will also give oral directions.* [Field Note, May 30, 2023]

Finally, technology use and implementation may be compromised by the ergonomic demands of completing tasks. Many dexterous tasks occur and involve more than one hand. Nursing tasks can be repetitive, tacit, and nuanced, placing a high burden on nursing operators. This was showcased in the PACU with intravenous starts, a routine task. Intravenous starts required many bimanual and minute actions.

> *Don gloves; apply tourniquet; assemble syringe and needle; open gauze; open swab; swab patient; check patient anatomy; air dry swab area by waving hands; insert needle; remove tourniquet.* [Field Note, August 31, 2023]

Discussion

Principal Findings

Understanding the antecedents of robotic technology use in health care elucidates both health and nonhealth-related consequences [44]. Merging these barriers with design guidelines [45], robotic systems design processes [46], safety taxonomies [47], and standards [48] can provide technology designers with valuable insights for creating robotic technology that effectively supports and aligns with nursing work. The contribution of this study is a comprehensive and empirically based overview of the potential barriers to integrating robotic and digital technologies into inpatient nursing workflows, categorized by the SEIPS elements of work systems (Table 1). The implications of each of these barriers are detailed in Multimedia Appendix 3.

Nursing has been described as a complex adaptive system, characterized by unpredictability, dynamicity, adaptability, and nonlinearity [49]. Nursing workflows are often inconsistent, vary in pace, do not follow a linear sequence, and frequently rely on workarounds. Scholars have used the term turbulent to capture the variability and fragmentation of nursing workflows, which are further complicated by time pressures, multiple competing demands, and resource constraints that require multitasking and task shifting [50,51]. In addition to nonlinear work processes, nurses must adapt to changing patient behaviors, team composition, and unpredictable clinical events. This work occurs in a physical environment that is often cluttered, restricted, disorganized, noisy, and cognitively demanding. Together, the characteristics of nursing work present significant challenges to the design and implementation of robotic technologies, which often depend on linear, repeatable, and well-parameterized tasks. Robotic design, therefore, cannot generalize from controlled demonstrations but must model the boundaries of variability that nurses manage in real time. In reality, technology may be disrupted during usage and required to deviate from routine behavior during both scheduled and unanticipated events. Workflow must be investigated before implementation to ensure appropriateness and to facilitate

XSL•FO
RenderX

integration. Future research should focus on capturing this variability and encoding appropriate responses across combinations of factors.

The realities of hospital systems add additional design complexities and constraints. Inpatient care settings comprise a multitude of independent and disparate people, including frontline staff, providers, administrators, patients, and families. These individuals represent a wide range of roles, expectations, psychosocial preferences, physical statures, and cognitive abilities. They are organized across multiple groups and departments, each representing their own culture, norms, rules, processes, and priorities. Health care systems have been critiqued for their disciplinary silos or "worlds," illustrating the significant discontinuity between caregivers, community members, and administrators [52,53]. In addition to internal agents and stakeholders, hospitals manage the requirements of laws and external regulations. Collectively, this heterogeneity and organizational complexity limit the effectiveness of standardized approaches, force trade-offs between overlapping design options, and limit the ability to support multiple end users.

When designing robotic technology, adjustable hardware, multimodal interfaces, and flexible workflows can help ensure usability across a range of users and use cases. However, the heterogeneity of people, roles, and culture makes generalization difficult. Technology that may be positively received in one unit may not be reciprocated in other units, even within the same organization. The same applies to people, where people of different or even within the same role may not agree on the functions and features of technology. If this variability cannot be accounted for, the scope of scenarios to which technology can be applied must be explicitly defined. Furthermore, organizational complexity influences the feasibility of innovative technology, with implementation often depending less on clinical need than on existing spatial, financial, and policy constraints. Because innovation requires a significant upfront investment and the rigid funding structures many organizations operate under, it is difficult for decision-makers to ignore the cost of technology compared with its potential benefit or value to patient outcomes [54]. Furthermore, nurse-led ideas for technology usage and innovation need organizational support to be implemented, which may require organizational policy changes to allow nurse involvement in technology design during working hours [55]. Thus, designing robotic technology with organizational involvement is critical.

Observations of real-world interactions between nurses and existing devices underscored the fragility and inadequacy of health care technology. In practice, the use of health care technology was frequently impeded by incompatible devices, missing supplies and accessories, transportability barriers, malfunctions, breakdowns, and profound misalignments between device operation and nursing workflows. Usability was further constrained by suboptimal interfaces, poor screen visibility, and inconsistent hospital practices. Previous literature corroborates the significant demands placed on nurses to locate, manage, organize, transport, use, and maintain technology [56,57]. These misalignments between existing devices and nursing work have undermined nurses' trust in technological solutions and added to frustration, as some technologies are perceived to increase nursing workload and complexity [58,59]. Tools that limit autonomy, impose a physical burden, increase cognitive workload, or interfere with empathy-driven or judgment-based activities may further undermine professional confidence and acceptance of robotics.

Overall, given the state of current robotic technology and the realities of nursing work systems, there appear to be particular challenges implementing robotics in roles that constrain the pace of time-sensitive nursing tasks, require real-time adaptability, directly interface with a diverse patient population, or risk disrupting trust in nurse-patient relationships. Combined with previously identified drawbacks of automation in nursing, where concerns include dehumanization of care, risk of deskilling, and equitable access [24], it is essential that future use of robotics in nursing does not impede the work system. The barriers identified in this study emphasize that people within the work system are heterogeneous and variable; the environment imposes physical and organizational limitations; technology and tools are often fragile and inadequate; and the tasks nurses perform are constrained by workflow, ethics, and ergonomics. Robotic systems must be designed deliberately, and to navigate these barriers, innovation will need considerable organizational support, time, and iterative design. We conclude that integration of present-day robotic technologies is most feasible in auxiliary and physically structured components of nursing work, as core bedside activities depend on frequent interruptions, shifting priorities, spatial variability, and distributed responsibility across staff, factors that robots cannot yet sense or interpret reliably enough to support effective clinical teaming. Acceptance will depend on whether robotic systems simplify rather than complicate care delivery. The most promising opportunities for robotic integration, therefore, lie in augmenting background processes, while the most critical limitations arise when systems compete with the adaptive and relational aspects of nursing work. Successful implementation will require designs that acknowledge and support the relational dimensions of nursing practice, while providing clear accountability structures and safeguarding nurses' professional judgment.

Limitations

As with many observational studies, there may be the occurrence of an observer effect where observing a phenomenon may change it [60]. Thus, participants may have altered their behavior due to the presence of an external observer. Additionally, because data collection was limited to what was visible or audible to the observer [61], and because the observer does not have a health care background, factors that influenced the participants' decision-making processes may not have been sufficiently captured. This was mitigated by encouraging participants to think out loud, asking clarifying questions, and by including health care professionals in the data analysis stage.

Additional limitations arise from both the participants and the observer in this study. Because recruitment was facilitated by the unit's leadership team, there is potential for selection bias, in which participants may have been chosen based on specific characteristics or workflow. Furthermore, the same participants

XSL•FO
RenderX

may have been observed multiple times. In bedside nursing, each shift is different, and the work system component interaction varies even when the same individual is being observed. One method to mitigate both selection bias and repeat participants was to allow multiple staff members to be shadowed during each observation period, observe interactions with other health care members, and include at least 3 different observation periods per unit. Finally, while patient-related barriers were observed and discussed, this study did not include patients' perspectives, and thus, the impact of technology on patients remains unexplored. Future work would benefit from including patients to elucidate how barriers affect those receiving care.

This study was conducted only during the day shift and excluded both the pediatric population and the population with psychiatric disorders, which may limit the generalizability of these findings. While 4 different units were included in this study to capture a broad range of nursing workflows, these findings may not be applicable to units that care for excluded populations or have a large night-shift population. Furthermore, because the purpose of this study was to identify barriers to the integration of robotic technology, one limitation is that opportunities for robotic technology are not explored extensively, and thus, the translational aspect of these findings is limited. The identification of barriers is essential, as realistic opportunities and design implementation requirements will be shaped by their presence. Exploring the connections among barriers, design requirements, and robotics-assisted solutions is an essential next step for future studies.

Conclusion

By systematically observing nursing workflows and synthesizing barriers into themes, this study provides new insights into the conditions that enable or constrain the integration of robotics into clinical practice. The findings emphasize that successful implementation depends less on technical capability and more on adaptability within complex sociotechnical environments. Robotic design cannot generalize from controlled demonstrations but must model the boundaries of variability that nurses manage in real time. Present-day robotic technology is best suited to support auxiliary and background tasks that relieve, rather than replace, human effort; however, substantial opportunities remain to expand their role through human-centered designs that address workflow variability, coordination demands, and distributed accountability in nursing care.

Design must prioritize adaptability, workflow alignment, reliability, robustness, interoperability, and ethical clarity. Successful implementation will also require organizational structures that promote training, maintenance, and relational sensitivity. Viewing technologies as sociotechnical systems embedded in complex work environments, rather than stand-alone devices, can offer a pathway toward safer, more resilient human-centered automation. The PETT scan proved valuable in mapping these challenges and can serve as a practical framework for anticipating barriers in future translational research.

Acknowledgments

This research was supported by the Elaine Marieb Center for Nursing and Engineering Innovation. The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the Elaine Marieb Center for Nursing and Engineering Innovation. The use of artificial intelligence (AI) was involved with refinement, correction, editing, and formatting the manuscript to improve clarity using spell and grammar checkers (Microsoft 365, Grammarly). All other manuscript activities, such as observation field note transcription, inductive and deductive coding, thematic analysis, and manuscript content writing were performed without AI usage.

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets used in this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Definitions used in this study.
[DOCX File, 23 KB - jmir_v28i1e89409_app1.docx]

Multimedia Appendix 2
Participant job role definitions.
[DOCX File, 20 KB - jmir_v28i1e89409_app2.docx]

Multimedia Appendix 3

XSL•FO
RenderX

People, Environment, Tools, and Tasks (PETT) Scan barriers.
[DOC File, 236 KB - jmir_v28i1e89409_app3.doc]

Checklist 1
SRQR reporting checklist.
[DOCX File, 109 KB - jmir_v28i1e89409_app4.docx]

References

1. Jones CH, Dolsten M. Healthcare on the brink: navigating the challenges of an aging society in the United States. NPJ Aging 2024 Apr 6;10(1):22. [doi: 10.1038/s41514-024-00148-2] [Medline: 38582901]
2. Willis E, Brady C. The impact of “missed nursing care” or “care not done” on adults in health care: a rapid review for the Consensus Development Project. Nurs Open 2022 Mar;9(2):862-871. [doi: 10.1002/nop2.942] [Medline: 34132481]
3. Ellehave SM, Voldbjerg SL, Rasmussen P, Laugesen B. Nurses’ perceptions of reasons for missed nursing care in hospitals: a qualitative systematic review protocol. JBI Evid Synth 2024 Aug 1;22(8):1594-1600. [doi: 10.11124/JBIES-23-00367] [Medline: 38483001]
4. Jones TL, Hamilton P, Murry N. Unfinished nursing care, missed care, and implicitly rationed care: State of the Science review. Int J Nurs Stud 2015 Jun;52(6):1121-1137. [doi: 10.1016/j.ijnurstu.2015.02.012] [Medline: 25794946]
5. Gong F, Mei YC, Wu M, Tang C. Global reasons for missed nursing care: a systematic review and meta-analysis. Int Nurs Rev 2025 Mar;72(1):e13096. [doi: 10.1111/inr.13096] [Medline: 39789750]
6. Palese A, Bassi E, Tommasini C, et al. Developing policies and actions in response to missed nursing care: a consensus process. J Nurs Manag 2019 Oct;27(7):1492-1504. [doi: 10.1111/jonm.12835] [Medline: 31349372]
7. Ball JE, Griffiths P. Missed nursing care: a key measure for patient safety. Agency for Healthcare Research and Quality | PSNet Internet. 2018. URL: https://psnet.ahrq.gov/perspective/missed-nursing-care-key-measure-patient-safety [accessed 2025-12-09]
8. Dupont PE, Nelson BJ, Goldfarb M, et al. A decade retrospective of medical robotics research from 2010 to 2020. Sci Robot 2021 Nov 10;6(60):eabi8017. [doi: 10.1126/scirobotics.abi8017] [Medline: 34757801]
9. Georgadarellis GL, Cobb T, Vital CJ, Sup FC. Nursing perceptions of robotic technology in healthcare: a pretest-posttest survey analysis using an educational video. IISE Trans Occup Ergon Hum Factors 2024;12(1-2):68-83. [doi: 10.1080/24725838.2024.2323061] [Medline: 38456754]
10. George EI, Brand TC, LaPorta A, Marescaux J, Satava RM. Origins of robotic surgery: from skepticism to standard of care. JSLS 2018;22(4):e2018.00039. [doi: 10.4293/JSLS.2018.00039] [Medline: 30524184]
11. Howe RD, Matsuoka Y. Robotics for surgery. Annu Rev Biomed Eng 1999;1:211-240. [doi: 10.1146/annurev.bioeng.1.1.211] [Medline: 11701488]
12. Kwoh YS, Hou J, Jonckheere EA, Hayati S. A robot with improved absolute positioning accuracy for CT guided stereotactic brain surgery. IEEE Trans Biomed Eng 1988 Feb;35(2):153-160. [doi: 10.1109/10.1354] [Medline: 3280462]
13. Lee A, Baker TS, Bederson JB, Rapoport BI. Levels of autonomy in FDA-cleared surgical robots: a systematic review. NPJ Digit Med 2024 Apr 26;7(1):103. [doi: 10.1038/s41746-024-01102-y] [Medline: 38671232]
14. Mendivil A, Holloway RW, Boggess JF. Emergence of robotic assisted surgery in gynecologic oncology: American perspective. Gynecol Oncol 2009 Aug;114(2 Suppl):S24-S31. [doi: 10.1016/j.ygyno.2009.02.002] [Medline: 19573702]
15. Jeffrey E, Dalby M, Walsh Á, Lai K. Automated dispensing cabinets and their impact on the rate of omitted and delayed doses: a systematic review. Explor Res Clin Soc Pharm 2024 Jun;14:100451. [doi: 10.1016/j.rcsop.2024.100451] [Medline: 38774122]
16. Tu HN, Shan TH, Wu YC, et al. Reducing medication errors by adopting automatic dispensing cabinets in critical care units. J Med Syst 2023 Apr 27;47(1):52. [doi: 10.1007/s10916-023-01953-0] [Medline: 37103718]
17. Metsämuuronen R, Kokki H, Naaranlahti T, Kurttila M, Heikkilä R. Nurses’ perceptions of automated dispensing cabinets - an observational study and an online survey. BMC Nurs 2020;19:27. [doi: 10.1186/s12912-020-00420-2] [Medline: 32327934]
18. Teng R, Ding Y, See KC. Use of robots in critical care: systematic review. J Med Internet Res 2022 May 16;24(5):e33380. [doi: 10.2196/33380] [Medline: 35576567]
19. Summerfield MR, Seagull FJ, Vaidya N, Xiao Y. Use of pharmacy delivery robots in intensive care units. Am J Health Syst Pharm 2011 Jan 1;68(1):77-83. [doi: 10.2146/ajhp100012] [Medline: 21164071]
20. Moxi--Diligent Robotics. URL: https://www.diligentrobots.com/moxi/ [accessed 2025-11-25]
21. Tietze M, McBride S. Robotics and the impact on nursing practice case study and pilot site analyses. : American Nurses Association; 2020 URL: https://anaprodsite2.nursingworld.org/globalassets/innovation/robotics-and-the-impact-on-nursing-practice_print_12-2-2020-pdf-1.pdf [accessed 2023-04-20]
22. Marshall D. Using our moxi: implementing robotics to remove staff delivery tasks. : American Organization for Nursing Leadership; 2022 URL: https://www.aonl.org/system/files/media/file/2022/02/Voice_1-22_Robotics.pdf [accessed 2025-09-05]

XSL•FO
RenderX

23. Adeyemo A, Coffey A, Kingston L. Utilisation of robots in nursing practice: an umbrella review. BMC Nurs 2025 Mar 4;24(1):247. [doi: 10.1186/s12912-025-02842-2] [Medline: 40038679]
24. Pepito JA, Acaso NJ, Merioles R, Ismael J. Opportunities, challenges, and future directions for the integration of automation in nursing practice: discursive study. JMIR Nurs 2025 Aug 14;8:e72674. [doi: 10.2196/72674] [Medline: 40811767]
25. Babalola GT, Gaston JM, Trombetta J, Tulk Jesso S. A systematic review of collaborative robots for nurses: where are we now, and where is the evidence? Front Robot AI 2024;11:1398140. [doi: 10.3389/frobt.2024.1398140] [Medline: 38899066]
26. Marcilly R, van den Berg L, Dusseljee-Peute L. Usability and UX in healthcare: exploring current practices and future directions. Stud Health Technol Inform 2025 May 15;327:323-327. [doi: 10.3233/SHTI250337] [Medline: 40380448]
27. Isaka S. Taxonomic robot identifiers: toward general classification and oversight for autonomous systems. IEEE Access 2025;13:101801-101816. [doi: 10.1109/ACCESS.2025.3578870]
28. How to determine if your product is a medical device. U.S. Food & Drug Administration. 2022. URL: https://www.fda.gov/medical-devices/classify-your-medical-device/how-determine-if-your-product-medical-device [accessed 2024-01-25]
29. Bekey GA. On autonomous robots. Knowl Eng Rev 1998 Jul;13(2):143-146. [doi: 10.1017/S0269888998002033]
30. Yang GZ, Cambias J, Cleary K, et al. Medical robotics-regulatory, ethical, and legal considerations for increasing levels of autonomy. Sci Robot 2017 Mar 15;2(4):eaam8638. [doi: 10.1126/scirobotics.aam8638] [Medline: 33157870]
31. Carayon P, Wetterneck TB, Rivera-Rodriguez AJ, et al. Human factors systems approach to healthcare quality and patient safety. Appl Ergon 2014 Jan;45(1):14-25. [doi: 10.1016/j.apergo.2013.04.023] [Medline: 23845724]
32. Smith MJ, Sainfort PC. A balance theory of job design for stress reduction. Int J Ind Ergon 1989 Jul;4(1):67-79. [doi: 10.1016/0169-8141(89)90051-6]
33. Carayon P, Smith MJ. Work organization and ergonomics. Appl Ergon 2000 Dec;31(6):649-662. [doi: 10.1016/s0003-6870(00)00040-5] [Medline: 11132049]
34. Donabedian A. The quality of care. How can it be assessed? JAMA 1988;260(12):1743-1748. [doi: 10.1001/JAMA.1988.03410120089033] [Medline: 3045356]
35. Carayon P, Schoofs Hundt A, Karsh BT, et al. Work system design for patient safety: the SEIPS model. Qual Saf Health Care 2006 Dec;15 Suppl 1(Suppl 1):i50-i58. [doi: 10.1136/qshc.2005.015842] [Medline: 17142610]
36. Benjamin E, Giuliano KK. Work systems analysis of emergency nurse patient flow management using the systems engineering initiative for patient safety model: applying findings from a grounded theory study. JMIR Hum Factors 2024 Dec 10;11:e60176. [doi: 10.2196/60176] [Medline: 39656555]
37. Holden RJ, Carayon P. SEIPS 101 and seven simple SEIPS tools. BMJ Qual Saf 2021 Nov;30(11):901-910. [doi: 10.1136/bmjqs-2020-012538] [Medline: 34039748]
38. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
39. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. The SRQR reporting checklist. In: Harwood J, Albury C, de BJ, Schlüssel M, Collins G, editors. The EQUATOR Network Reporting Guideline Platform 2025. URL: https://resources.equator-network.org/reporting-guidelines/srqr/srqr-checklist.docx [accessed 2026-03-17]
40. Saunders B, Sim J, Kingstone T, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907. [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
41. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
42. Hose BZ, Carayon P, Hoonakker PLT, et al. Work system barriers and facilitators of a team health information technology. Appl Ergon 2023 Nov;113:104105. [doi: 10.1016/j.apergo.2023.104105] [Medline: 37541103]
43. Lincoln YS, Guba EG. Naturalistic Inquiry: SAGE Publications Inc; 1985.
44. Huang R, Li H, Suomi R, Li C, Peltoniemi T. Intelligent physical robots in health care: systematic literature review. J Med Internet Res 2023 Jan 18;25:e39786. [doi: 10.2196/39786] [Medline: 36652280]
45. Barajas NC, Rouchitsas A, Broo DG. Examining human–robot interactions: design guidelines for trust and acceptance. In: Malakhatka E, Wiberg M, editors. Human-Technology Interaction: Interdisciplinary Approaches and Perspectives: Springer Nature Switzerland; 2025:117-137. [doi: 10.1007/978-3-031-78357-9_4]
46. Leone S, Lago F, Pisla D, Carbone G. A systematic approach for robotic system development. Technologies (Basel) 2025;13(8):316. [doi: 10.3390/technologies13080316]
47. Akalin N, Kiselev A, Kristoffersson A, Loutfi A. A taxonomy of factors influencing perceived safety in human–robot interaction. Int J of Soc Robotics 2023 Dec;15(12):1993-2004. [doi: 10.1007/s12369-023-01027-8]
48. Díaz CE, Fernández R, Armada M, García F. A research review on clinical needs, technical requirements, and normativity in the design of surgical robots. Int J Med Robot 2017 Dec;13(4):e1801. [doi: 10.1002/rcs.1801] [Medline: 28105687]
49. Chaffee MW, McNeill MM. A model of nursing as a complex adaptive system. Nurs Outlook 2007;55(5):232-241. [doi: 10.1016/j.outlook.2007.04.003] [Medline: 17950117]
50. Ismail HA, Kotp MH, Basyouny HAA, et al. Empowering nurse leaders: leveraging financial management practices to foster sustainable healthcare - a mixed-methods study. BMC Nurs 2025 Mar 28;24(1):335. [doi: 10.1186/s12912-025-02981-6] [Medline: 40156004]

XSL•FO
RenderX

51. Jennings BM, Baernholdt M, Hopkinson SG. Exploring the turbulent nature of nurses' workflow. Nurs Outlook 2022;70(3):440-450. [doi: 10.1016/j.outlook.2022.01.002] [Medline: 35221055]
52. Pedersen ERG, Sudzina F, Rosati F. A multi-dimensional study of organisational boundaries and silos in the healthcare sector. Health Serv Manage Res 2024 Nov;37(4):200-208. [doi: 10.1177/09514848231218617] [Medline: 38011078]
53. Glouberman S, Mintzberg H. Managing the care of health and the cure of disease--part I: differentiation. Health Care Manage Rev 2001;26(1):56-69. [doi: 10.1097/00004010-200101000-00006] [Medline: 11233354]
54. MacNeil M, Koch M, Kuspinar A, Juzwishin D, Lehoux P, Stolee P. Enabling health technology innovation in Canada: barriers and facilitators in policy and regulatory processes. Health Policy 2019 Feb;123(2):203-214. [doi: 10.1016/j.healthpol.2018.09.018] [Medline: 30352755]
55. van Houwelingen T, Meeuse ACM, Kort HSM. Enabling nurses' engagement in the design of healthcare technology - core competencies and requirements: a qualitative study. Int J Nurs Stud Adv 2024 Jun;6:100170. [doi: 10.1016/j.ijnsa.2023.100170] [Medline: 38746818]
56. Al Moteri M, Aljuaid J, Alsufyani B, Alghamdi A, Althobiti ES, Althagafi A. Bottleneck factors impacting nurses' workflow and the opportunity to prioritize improvement efforts: factor analysis. BMC Nurs 2024 Sep 11;23(1):640. [doi: 10.1186/s12912-024-02311-2] [Medline: 39256713]
57. Qtait M. Time wasters facing nurses during work in hospital departments. SAGE Open Nurs 2025;11:23779608251328278. [doi: 10.1177/23779608251328278] [Medline: 40171549]
58. Mohammadnejad F, Freeman S, Klassen-Ross T, Hemingway D, Banner D. Impacts of technology use on the workload of registered nurses: a scoping review. J Rehabil Assist Technol Eng 2023;10:20556683231180189. [doi: 10.1177/20556683231180189] [Medline: 37342268]
59. Alnawafleh KA. The impact of AI on nursing workload and stress levels in critical care settings. Pakjlifesocsci 2024;22(2). [doi: 10.57239/PJLSS-2024-22.2.00643]
60. Bauer CCC, Atad DA, Farb N, Brewer JA. From confound to clinical tool: mindfulness and the observer effect in research and therapy. Biol Psychiatry Cogn Neurosci Neuroimaging 2025 Apr;10(4):402-410. [doi: 10.1016/j.bpsc.2025.01.012] [Medline: 39894252]
61. Lindström V, Jepsen K, Heldring S, Kanfjäll T, Rådestad M. A real-time communication and information system for triage, positioning, and documentation (TriPoD) in mass-casualty incidents: a qualitative observational study. BMC Emerg Med 2025 Jul 6;25(1):115. [doi: 10.1186/s12873-025-01274-0] [Medline: 40619380]

Abbreviations

ADC: automated dispensing cabinet
ED: emergency department
PACU: preop or postanesthesia care unit
PETT: People, Environment, Tools, and Tasks
SEIPS: Systems Engineering Initiative for Patient Safety
SRQR: Standards for Reporting Qualitative Research

Edited by A Stone; submitted 11.Dec.2025; peer-reviewed by E Fioratou, SK Geale; revised version received 09.Apr.2026; accepted 30.Apr.2026; published 01.Jun.2026.

Please cite as:
Georgadarellis GL, Benjamin E, Roberts SC, Vital CJ, Sup IV FC
Evaluating Nursing Work Systems and Identifying Barriers for Robotic Technology Integration: Observational Study
J Med Internet Res 2026;28:e89409
URL: https://www.jmir.org/2026/1/e89409
doi:10.2196/89409

© Gina L Georgadarellis, Ellen Benjamin, Shannon C Roberts, Cidalia J Vital, Frank C Sup IV. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 1.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assistive Robotics for Healthy Aging: A Foundational Phenomenological Co-Design Exercise

Stephen Potter[1], PhD; Mark Hawley[1], PhD; Angela Higgins[2], MSc; Farshid Amirabdollahian[3], PhD; Mauro Dragone[4], PhD; Alessandro Di Nuovo[5], PhD; Praminda Caleb-Solly[2], PhD

[1]Sheffield Centre for Health and Related Research, University of Sheffield, Sheffield, United Kingdom
[2]School of Computer Science, University of Nottingham, Nottingham, United Kingdom
[3]Robotics Research Group, University of Hertfordshire, Hatfield, United Kingdom
[4]School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh, United Kingdom
[5]School of Computing, Sheffield Hallam University, Sheffield, United Kingdom

Corresponding Author:
Stephen Potter, PhD
Sheffield Centre for Health and Related Research
University of Sheffield
Regent Court
30 Regent Street
Sheffield, S1 4DA
United Kingdom
Phone: 44 114 222 2000
Email: stephen.potter@sheffield.ac.uk

Abstract

Background: Assistive robotics for helping older people live well and stay independent has, to date, failed to fulfill its promise: there are few assistive robots in everyday use. In part, this failing can be attributed to inadequate or missing co-design activities that would ensure that these technologies and any services that incorporate them are developed with prospective end users, addressing their actual needs and wants, and not merely for them, and based on lazy assumptions about heterogeneous user groups.

Objective: This exercise aimed to address some of these limitations by taking a "phenomenological snapshot" of what it means to be an older person in the current sociotechnological context, and making this snapshot, along with the co-design materials developed, available to the wider assistive robotics community to provide solid foundational evidence for steering the development of assistive robotics in more productive directions.

Methods: Two rounds of co-design workshops have been conducted with older people and their caregivers, based on an innovative methodology that used personas and speculative designs to explore sensitive everyday difficulties faced by participants and highlight some of their general wishes for and concerns about assistive robotics. The data collected during the workshops were analyzed, and key themes were extracted.

Results: Analysis of the workshop data gives access to the lived experience of older people and their caregivers, and their opinions about domestic robotics and assistive technologies more generally. The findings are organized thematically as *everyday difficulties*, the daily problems faced by older people; *ideas for aging better*, older people's own suggestions for how their lives could be improved; and *living with technology*, their preferences and requirements for assistive robots, along with their concerns about what the introduction of robots might mean, both for themselves and for society more widely.

Conclusions: We believe that our findings provide solid foundational evidence for the development of assistive robotics for older people. We are in the process of disseminating these results through various channels to the wider assistive robotics community; ultimately, the success of our activities will be demonstrated only through the development of acceptable, useful, and viable assistive robotics for older people.

(J Med Internet Res 2026;28:e77179) doi:10.2196/77179

XSL•FO
RenderX

KEYWORDS

robotics; assistive technology; frail older adults; activities of daily living; independent living; human centered design, co-design; lived experience

Introduction

Emergence Network

Emergence is a UK-wide initiative created to foster innovation in the field of assistive robotics for older people, especially people living with frailty, and to facilitate the "emergence" of robots from research laboratories into service. The network is coordinated by a team of academic researchers and, in addition to other academic partners, members include health, social care, and housing providers, regulators, and robotics industry representatives, who provide clinical focus and business acumen to anchor its activities. The network was launched in early 2022 and is led by the University of Nottingham, with the support of coinvestigators from Heriot-Watt University, the University of Hertfordshire, the University of Sheffield, and Sheffield Hallam University.

More specifically, the objectives of Emergence are to promote the development of domestic assistive robotics that will improve the quality of life and independence of older people by enabling better self-management of frailty and age-related issues, by helping with activities of daily living, by assisting rehabilitation activities, and by supporting health care professionals to make tailored interventions to slow decline or to improve resilience. The primary impetus behind the establishment of the network was the observation that, despite the undoubted effort and resources that had been directed at the task, few assistive robots for older people are in general use (see, for example, the book by Wright [1]). Plausible factors that may have contributed to this include a poor appreciation of the needs and wants of older people, of the wider psychological, social, and clinical considerations, and of the challenges and practicalities of operation in noisy real-world domestic environments.

One fundamental pillar of the Emergence network is the proposition that this general lack of understanding of use contexts can be addressed through a rigorous and ongoing co-design exercise involving people from the target user groups, with the results of this exercise guiding assistive robotics development.

Frailty

Frailty is "a common clinical syndrome in older adults that carries an increased risk for poor health outcomes, including falls, incident disability, hospitalization, and mortality" [2]. It is characterized by the "loss of biological reserves across multiple organ systems," leading to clinical vulnerability to stressful episodes [3]. While it is not an inevitable consequence of aging, frailty is common among older people, and its extent and severity can vary significantly from individual to individual. Many older people who live with a health condition (or with multimorbidity) or with a disability will also be living with frailty: the one can contribute to or exacerbate the other. It can affect a person's physical and cognitive abilities, to the detriment of their wider health and well-being, and can limit their day-to-day functioning [4]. Frailty, when it does occur, is not a static condition: it can become worse, but there is evidence to suggest that, with the right interventions, it can be reversed to some extent [3].

When it comes to characterizing frailty, there are currently 2 dominant models [5]. The "(cumulative) deficit model" of frailty [6,7] assumes that frailty derives from an accumulation of "deficits" (which might be disabilities, health conditions, or psychosocial factors) that, individually and cumulatively, increase a person's vulnerability to stressful events. The second prevailing model is the "phenotype model" of frailty, in which an individual who manifests at least 3 of a set of 5 criteria for frailty (weight loss, low physical activity, exhaustion, slow walking speed, and physical weakness) is considered to be frail [4].

These models have their uses when considering whether an individual, in a health or care context, might require additional support or could benefit from some remedial intervention. However, they are less useful when it comes to understanding people's lived experience of frailty—and, by extension, for developing practical assistive technologies, including robotics, to support people living with the condition. The impact of deficits or of physical manifestations will vary from person to person, depending on their living circumstances, their tastes, their responsibilities and roles, and their hobbies and interests. Hence, while both the deficit model and the phenotype model can provide suggestions of where assistance might be needed, they also promote a narrow (and negative) view of people with frailty as deficient, a view that excludes a more rounded consideration of their lives that acknowledges their abilities (see the study by Lang et al [8]).

To take an example, clinical tests might reveal a loss of grip strength. The everyday impact of this on an individual might be felt in their inability to open jars of their favorite food (which might then have the knock-on effect of reducing the amount they are eating, leading to weight loss). Whereas the clinical model might suggest more-or-less generic remedial strength exercises—that is, focusing on the deficit or symptom and "treating" it—a more complete understanding of the individual's experience opens up the design space to suggest other, perhaps complementary, approaches encompassing the technological (a jar-opening robot), the environmental (redesigned food packaging), the social (a personalized meal delivery service), and so on. Any of these "interventions," or perhaps some combination of them, might be more acceptable to, appropriate—and, ultimately, effective—for the individual in question. The goal of the research reported in this paper is to provide a co-design counterbalance to the prevalent clinical models in the form of a snapshot of the lived experience of older people, one that encompasses their views of the digital world and, especially, robotics.

XSL•FO
RenderX

Assistive Robotics for Healthy Aging

Assistive robotics could help older people living with frailty, long-term health conditions, and disabilities to live more independent, more dignified, and more fulfilling lives. Moreover, services incorporating assistive robotics could play a role in plugging the gap in care provision that is being felt in many countries as their populations age, and they experience shortages of care staff. Many of the assistive robotics development projects undertaken have addressed one or other aspect of care provision for older people. These have typically taken the form either of, on the one hand, physically assistive robots (eg, for mobility support [9,10], exercise training [11] and help with eating [12]) or, on the other hand, of socially (or clinically) assistive robots—recent contributions have focused on stimulation [13], companionship [14,15], assessment tasks [16,17], monitoring [18-21], and remote visits or consultations [22]. It will be noted that both physically and socially assistive robots each tend to address only 1 or 2 aspects of frailty (care); as such, none of these robots can be said to be "for frailty" in its widest sense, if, indeed, such a thing is even possible. Notwithstanding these and other encouraging feasibility studies, undertaken both in laboratories and, on occasion, in people's homes, care homes, or clinical settings, this effort has yet to translate into anything approaching scale-up of production and widespread adoption. Despite years of research and development—and not inconsiderable financial investment—few assistive robots are actually deployed in the real-world to provide care or assistance to older adults [1].

At a technical level, it is undoubtedly extremely challenging to provide robotic assistance in a safe, sensitive, timely, responsive, respectful, reliable, and trustworthy manner to, with, and around ordinary people—who may have sensory, cognitive, or physical impairments—as they go about their everyday lives. However, the task has not been made any easier by basing design decisions on often unfounded or simplifying preconceptions of older people's needs, wants, and circumstances. In Emergence, we want to address this issue through the effective co-design of assistive robotics.

Co-Design of Assistive Robotics

Conventional design practices can result in "solutions" that leave end users with a sense of alienation and exclusion (especially where new or unfamiliar technologies are deployed), and the (entirely accurate) impression that the product or service has been designed "for them" and not "with them." It is likely that such products or services will not be adopted or quickly fall into disuse, unless their users have no alternative, when they will be used under duress and with little pleasure, and possibly with limited success. This failing becomes particularly significant in assistive technology contexts (and in health and care contexts more generally), where outcomes relate directly to the health and well-being of the end user. Co-design practices seek to address this failing by increasing the involvement of external stakeholders, and, in particular, end user representatives, in design and development processes, particularly during the earliest phases of design when, as is widely acknowledged, any incorrect decisions taken are significantly more difficult (and costly) to rectify later on. Co-design is a term given to a broad range of participatory techniques and methodologies whose underlying objective is to improve the acceptability and value of products and services by involving representative target users in their development [23]. During co-design, the role of the external actors can range from the relatively passive (such as critiquing alternative designs put to them) through more engaged modes (helping to define requirements or brainstorming potential solutions) to effectively becoming embedded in the design team and contributing to all stages of the development process.

There have been previous co-design exercises involving older people (for examples see other studies [22-27]). However, these exercises also underscore that co-design is not easy, with no single methodology; that access to the right stakeholders at the right time is not a given; that it is costly, in terms of both money and time; that participants find it difficult to shake preconceptions about robots and the roles they can play; and that, in the final analysis, it does not guarantee success.

Aim

Given what is at stake, however, none of these is a reason not to do co-design; rather, they are reasons to devote sufficient resources and care to try to do it better. The work reported in this paper aims to provide assistive robotics for older people with sturdy, co-designed foundations. Specifically, we have performed an extensive early-stage co-design activity with older people through which we hope to understand some of the current everyday problems facing people and their general requirements and attitudes toward hypothetical robotics solutions to some of those problems. We have undertaken this activity with the intention of collating the results and then sharing them as widely as possible among the assistive robotics community and, in this way, to provide a rich seam of foundational evidence for researchers and developers who otherwise lack the skills, wherewithal, or opportunities for engaging in early-stage co-design with older people. Ultimately, we intend to steer assistive robotics for older people in directions that are more likely to lead to the development of assistive robot services that provide genuine value for their users.

Methods

Emergence Co-Design Exercise

The objective of the co-design exercise was to take a "phenomenological snapshot" of frailty and aging, that is, one that positions the individual at the center of the enquiry [28] to complement the clinical models described above by capturing people's everyday experiences of frailty and aging. Moreover, the aim was to do this in such a way as to avoid some of the weaknesses of previous co-design in this field by expressly not focusing on any specific task or assuming as a basis any specific robot platform. This snapshot would encompass the physical, psychological, and social effects of aging, as well as the role that innovations (not necessarily technological) could play in mitigating negative effects and improving people's lives, and any constraints under which such innovations must operate if they are to be acceptable. As stated, the primary purpose of capturing such a snapshot is to provide guidance and support for the early stage (conceptual design) development of assistive technologies, specifically assistive robotics. However, the

XSL•FO
RenderX

snapshot could play a role beyond this, informing a variety of interventions and improvements, and educating those involved in the care and support of older people.

Moreover, it was imperative that the snapshot be accessible and communicable to a wide network of robotics and technology researchers, designers, and developers, and in such ways as to encourage the development of appropriate assistive robotics solutions. This, we hoped, would help to overcome the problems of a lack of expertise, experience, and resources for performing co-design, and of limited access to participants.

As mentioned above, there is no single methodology for co-design, although a number of different approaches have been suggested. Here, we adopted a selection of approaches and tools, intended to give us access to complementary information and in such a way as to hold the participants' interest and keep them engaged.

Co-Design Methodology

Overview

Workshops were run in 3 different locations (corresponding to the locations of academic partners in the Emergence network) in the United Kingdom. The workshops were "paired" at each location: an initial, "lived-experience" workshop was followed by a "speculative-critique" workshop.

Lived-Experience Workshops

The lived-experience workshops were designed to enquire into the needs and aspirations of older people living with frailty, the everyday realities of their lives, their living environments and social activities, and the benefits and frustrations of modern life, and, inevitably, of modern technologies (although the workshops were not designed to discuss technologies specifically). Participants were encouraged to identify any opportunities for assistive technologies, including robotics, to play a role in older people's lives.

To facilitate these workshops, we used personas as a starting point for group discussions. Personas are descriptions of fictitious individuals and have been widely used to motivate or contextualize co-design or innovation processes [29]. A total of 10 personas were developed for the Emergence workshops. We decided to develop these by drawing on our previous research [30] and existing data-based resources such as the CURE-Elderly-Personas set [31], which were adapted and extended to be more relevant to the United Kingdom context and to focus on the specific characteristics of the populations of interest to Emergence.

These personas depict, in easy-to-digest formats, the backgrounds, health status, and characteristics of fictitious older people, each of whom can be classed into 1 of 3 broad categories: prefrail but managing; vulnerable or living with mild frailty; or living with moderate-to-severe frailty. It was envisaged that having personas with different experiences of frailty would allow us to explore different intervention contexts and make comparative assessments of needs and wants. Before use, the personas were validated to be representative of people living with frailty by the Emergence steering groups, comprising health and social care professionals and people with lived experience of frailty. The group also reviewed the terminology used for possible infelicities or offence. Examples of the personas developed for and used in the workshops are shown in Figures 1 and 2. The personas have been made publicly available to allow their use in other co-design exercises.

XSL•FO
RenderX

Figure 1. "Michael," one of the personas developed for the lived-experience workshops.

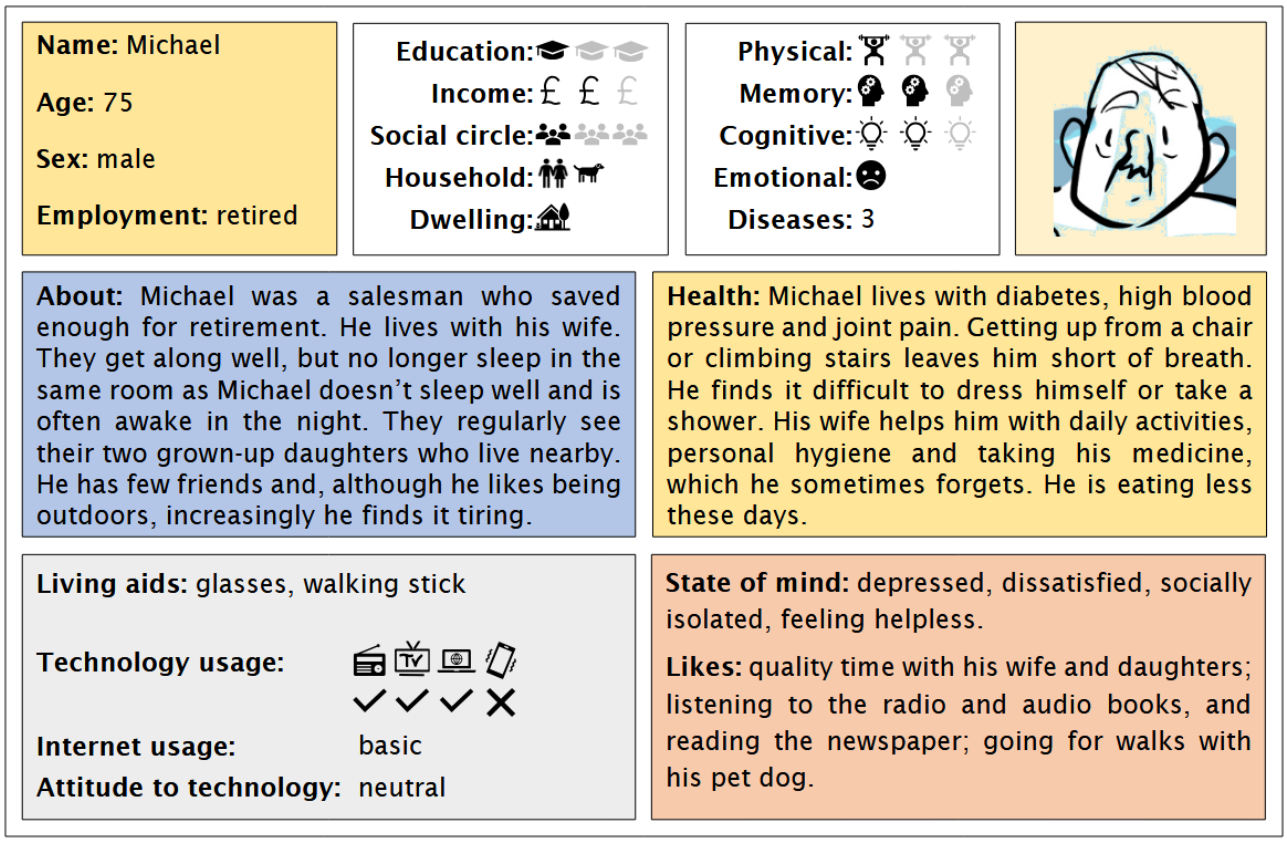

Figure 2. "Dorothy," one of the personas developed for the lived-experience workshops.

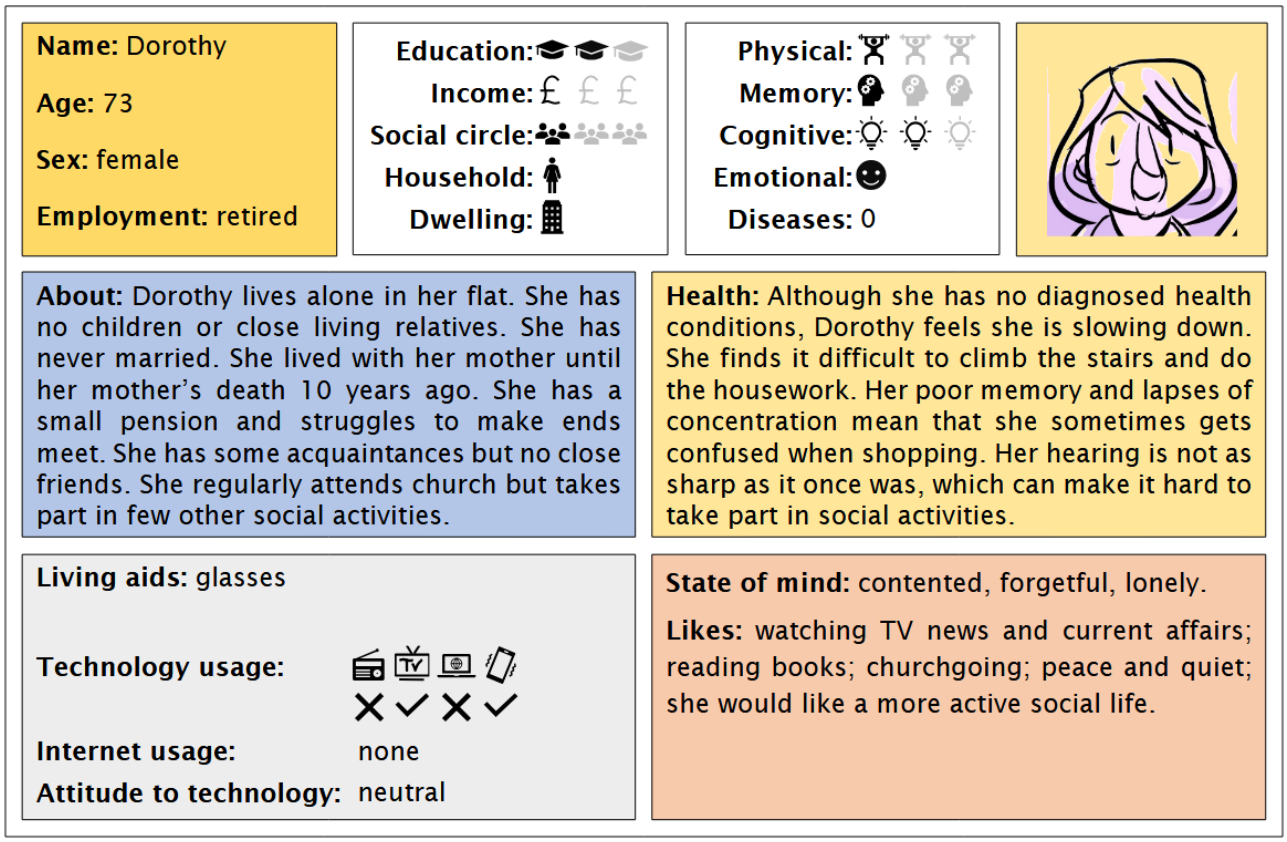

Personas can be used in different ways within design and innovation processes. Here, the primary role of the personas was to aid our participants within the co-design activities to "transfer" their own needs, wants, opinions, difficulties, and coping strategies onto fictional others, thereby avoiding potentially sensitive, embarrassing, or otherwise inhibiting

XSL•FO
RenderX

discussions of a personal nature, allowing us obliquely to broach frailty and other issues related to aging. More specifically, they would be used in the context of discussions of different "episodes" of the persona's typical daily routine:

- Getting up: including waking, toileting, washing and bathing, dressing, and medication.
- Mealtimes and snacks: including planning meals, food preparation, ordering food, utensil operation, cooking, eating, drinking, washing up, remembering to eat and drink enough, and maintaining a balanced diet.
- Household chores: including cleaning the house, heating the house, laundry, everyday repairs and maintenance, looking after pets, and household management (payment of bills, etc).
- Out and about: including getting around (walking, driving, and public transport), going to the hairdressers, going to the supermarket, going to the bank or post office, going to the doctor's, buying items, carrying things, and outdoor exercise.
- Socializing and pastimes: including home entertainment (television, music, internet, and puzzles), gardening, meeting friends and family, receiving guests, indoor exercise, and outside entertainment (cinema, concerts, bingo, book groups, walking groups, etc).
- Bedtime: including taking medicine, switching off and locking up, climbing stairs, getting to sleep, and getting up in the night.

Working in small groups consisting of 3 or 4 participants, each of which would be allocated 2 personas and 3 daily episodes, the lived-experience workshop participants were asked to consider what difficulties the episodes might present for those personas and what opportunities they can see for making the personas' lives easier. Following a break, the groups were then asked to consider each of the difficulties or opportunities and reach some consensus about: how frequently it occurs (daily, weekly, monthly...?); how prevalent it is (encountered by most, some, or just a few people?); and how consequential it is (does it have a large, medium, or small impact on people's ability to get by?). The answers to these questions would allow us to assess the significance of the difficulties and opportunities, and hence to prioritize them as objectives for assistive technologies.

The group discussions were facilitated by members of the research team who took notes as participants spoke, displaying these on whiteboards for immediate validation by participants. The discussions were audio-recorded for future verification and analysis.

In addition, the workshops were documented by a professional illustrator with experience in supporting academic workshops. Graphic facilitation can capture ideas in the form of easily digestible pictures, and can visualize the progression of and relationships between concepts discussed [32]. These illustrations would also represent the findings of a workshop in an easily accessible format, for reflection (and validation) during and after the workshop [33]. This was felt to be a particularly important aspect in this case, since one of our objectives was to communicate, in as readily accessible a manner as possible, the results to the wider assistive robotics development community. The illustrator was briefed as to the purpose and structure of the workshops, and then given free rein to move among the group discussions and capture any parts of the discussions that seemed to them both significant and susceptible to pictorial expression (they would also incorporate text, often quoting participants verbatim, into their illustrations). The illustrator attended and illustrated all the lived-experience workshops; they were unable to attend the subsequent speculative-critique workshops (but would supply the illustrations used to facilitate these workshops, as described below).

Speculative-Critique Workshops

The speculative-critique workshops would be structured around several "speculative designs" of assistive robots proposed to address the problems or grasp the opportunities identified by participants in the initial lived experience workshops.

These designs were in the form of a brief textual description and a sketch of the robot in action, helping to convey both the appearance of the robot and its use and operation. These were not intended to constitute designs for robots that would necessarily be developed; indeed, although the robots were intended to be realistic, in the sense of their functioning and behaviors being more-or-less feasible in the short term given current developments and directions in robotics research and development, they did not need to be feasible at the time of the workshops given the current state-of-the-art. Instead, they were to serve as "provocations" (elsewhere, such design provocations in the context of participatory design have been termed "provotypes" or "provocative (proto)types" [34]). As such, these workshops would constitute "speculative (co-)design" activities. Speculative design [35] is an activity during which a design proposal is presented not as a candidate for subsequent product development but as a means to elicit concerns or highlight issues that might otherwise remain latent, but which must be acknowledged if the design process is to be successful. This approach was felt to be particularly appropriate for a field such as robotics, about which many people possess preconceived notions. Moreover, discussing potential applications of any new technology without concrete examples is difficult: the speculative designs would constitute a basis for grounding discussions around assistive robotics in something approaching reality. Attempts to ground co-design discussions around (prototype applications of) real, existing robots are common; however, the obvious limitation of this approach is that the robots in question will have already been developed to some degree and, even when they have been developed with assistive applications in mind, this is likely to have involved certain assumptions on the part of the designers. Moreover, they will almost certainly have technical shortcomings. Consequently, it is difficult to move the discussion beyond the specific limitations of that particular robot or envisaged application so as to capture more general requirements for assistive robotics. We believe that the speculative design approach gives participants more freedom to move beyond the particular and express more general opinions about future applications of technology [36].

The development of the speculative designs required a rapid analysis by members of the research team (SP and MH) of the

XSL•FO
RenderX

content of the lived-experience workshops to identify candidate domestic tasks with which assistive robotics might help; a total of 6 such "robots" were identified:

- Motibot ("the motivational well-being and exercise robot")
- Foodee ("your personal cooking assistant and dietary advisor")
- EasyUp (a mobility assistance robot "for all life's ups and downs")
- AutoReach (a cleaning robot to "keep those hard-to-reach places spotless")
- RoPet (a robot pet that is "your new faithful friend")
- Toilittle (a discreet toileting robot that's "there when most you need it")

Each of these was summarized in words in terms of the problem or opportunity that the robot sets out to address (the "why" of the robot), its projected functionalities (the "what"), and how the robot would operate or be operated (the "how"). To aid in communicating the designs to workshop participants, these textual descriptions then became the brief for the illustrator to work up a sketch of the robot "in action," working with the research team to imagine what form the robot could take. In undertaking the brief, across the 6 applications, the illustrator was asked to draw robots with a range of embodiments, sizes, interfaces, and movement styles to elicit responses to different design options. Figures 3 and 4 give examples of 2 of these speculative designs in the format in which they were presented at the workshops.

Figure 3. Motibot ("the motivational well-being and exercise robot"): one of the speculative designs developed for critique.

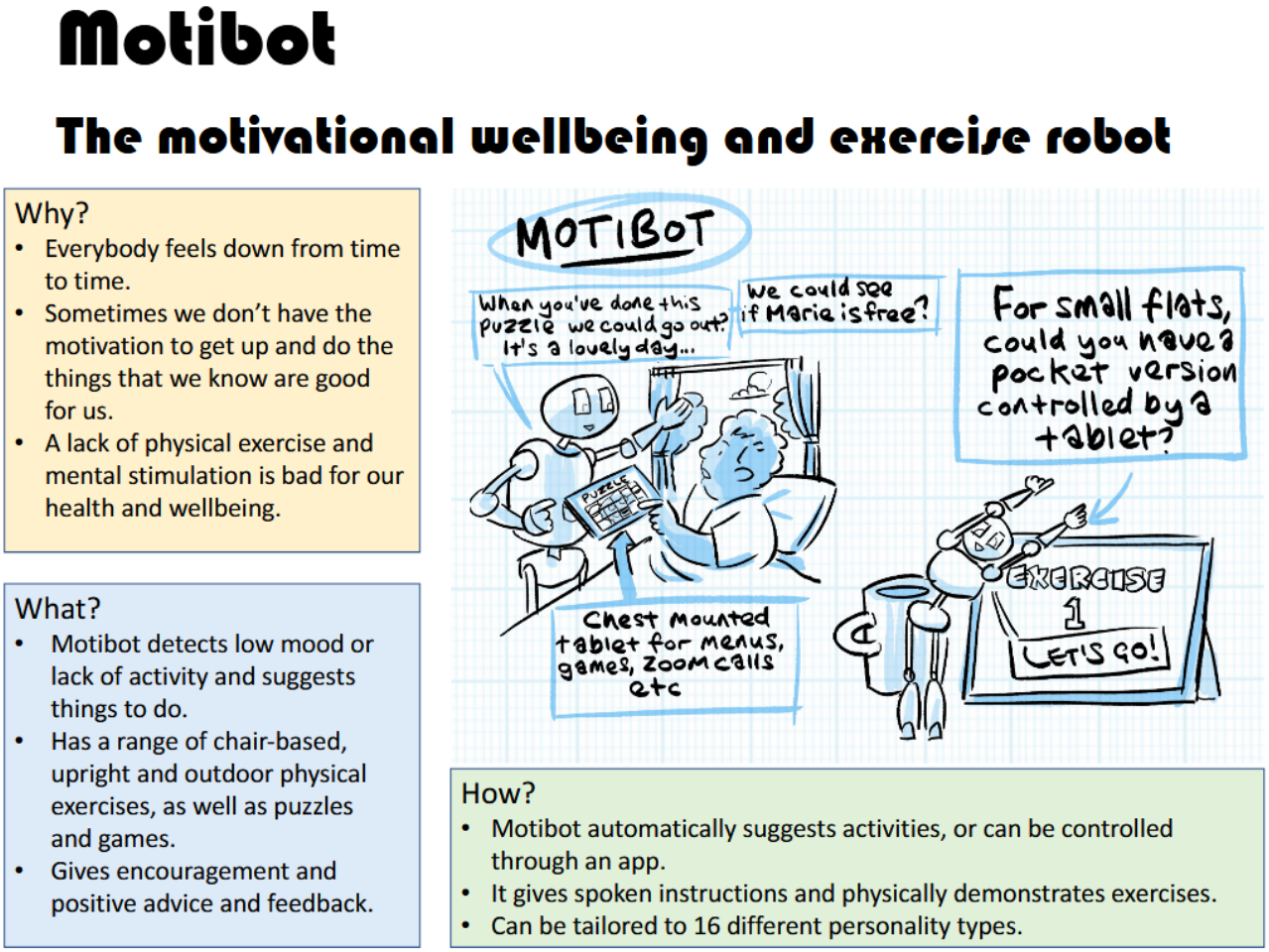

XSL•FO
RenderX

Figure 4. RoPet ("meet your new faithful friend"): one of the speculative designs developed for critique.

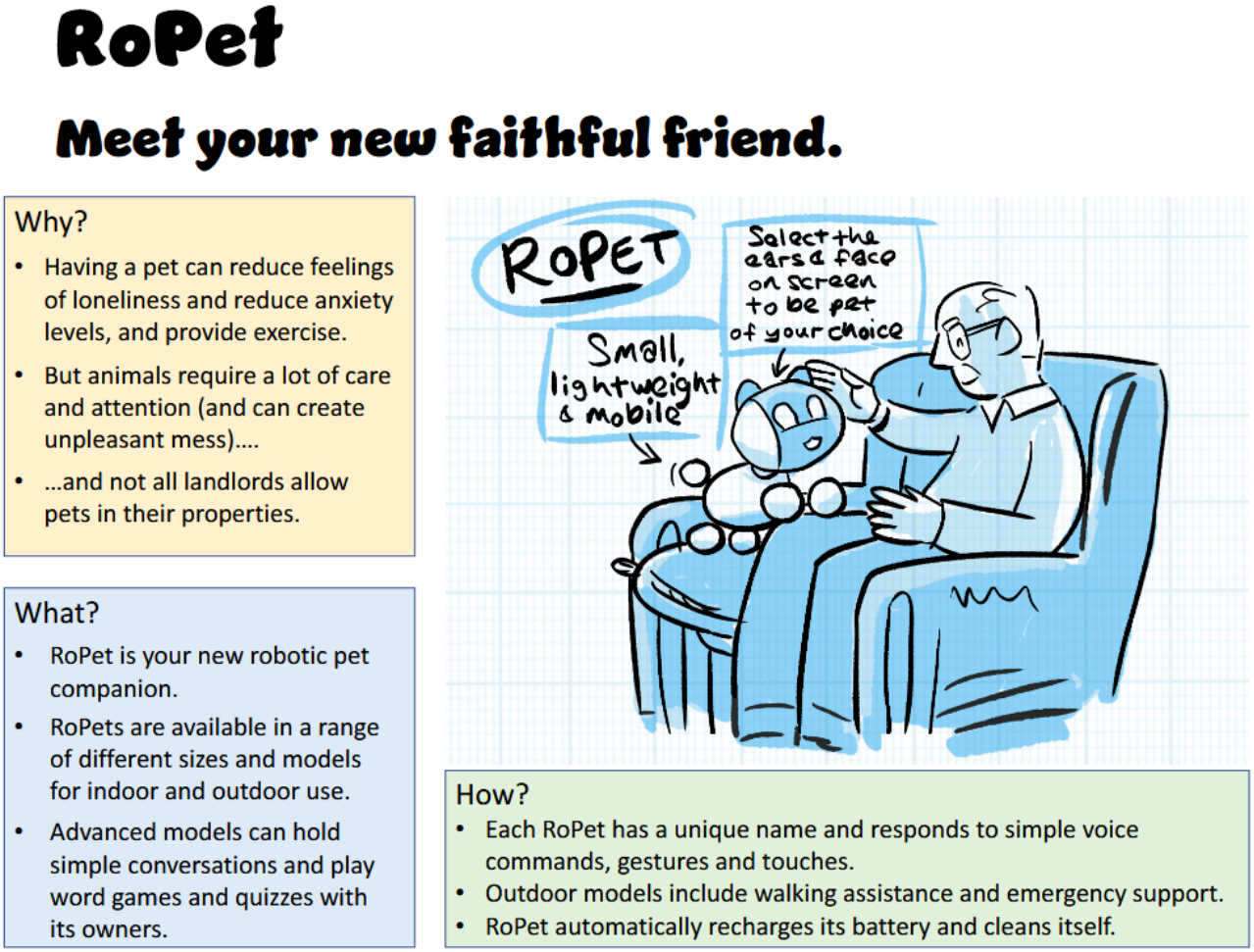

For the speculative-critique workshops, the participants were, once again, allocated to small groups, and asked to consider each of the speculative designs in turn, it first having been emphasized that the designs portray imagined, rather than real, robots. The participants were asked whether they would like to have the robot, and then asked for the reasons for their decision. Participants were encouraged to ask for additional information if this would help; facilitators were encouraged to extemporize answers. Using different-colored notepads, the facilitators noted separately positive and negative arguments, as well as quoting verbatim the words of participants where these seemed particularly pertinent. If at least 1 member of the group said they would accept the robot, the group was then asked about the following features of the robot:

- Appearance: what was liked and disliked about it, what size and color it should be, what materials it should be made of, and so on.
- Control: the extent and nature of the control that the end user and their caregivers should have over the robot, appropriate modalities for controlling the robot (voice, touchscreen, smartphone app, etc), and so on.
- Performance: its operating speed, acceptable noise levels, weight, communication, quality of movement, behavior both when in use and when not in use, and so on.
- Practicalities: how it would fit into people's lives and homes, whether people would be comfortable operating and maintaining the robot, the sort of training and assistance they felt they might need, and so on.
- Concerns: any safety or security issues, whether the robot seemed trustworthy, who would be trusted to develop, supply, run, maintain the robot, and so on.

Once again, facilitators would note both positive and negative comments: the former would constitute design requirements ("should-haves" or "nice-to-haves") while the latter constitute design constraints ("should-not-haves").

Participants

Older people do not always recognize themselves as living with frailty and can resist being labelled as "frail": the term has certain negative connotations, and not all people clinically classified as such would recognize (or welcome) the attribution [37]. Moreover, its severity can vary dramatically, from those who are moribund and almost completely dependent on others to those whose coping strategies are so successful that others——and maybe even they themselves—might not consider them to be living with frailty. For these reasons, we avoided using the term "frailty" when engaging with older people and their caregivers in these co-design activities. Moreover, we chose not to recruit through health services because we would then be accessing only those who, for one reason or another, had come to the attention of those services and received an "official" diagnosis. Instead, we focused on recruiting through

XSL•FO
RenderX

independent-living housing providers managed by Emergence partners and located conveniently close to the Emergence university research centers. People aged 55 years and older were considered "older people," adopting the threshold age for residency used by one of the housing providers. Even if not living with frailty themselves, most people living in these contexts would almost certainly know of people among their friends and neighbors who are living with frailty. In addition, relatives and friends who play an active role in supporting older people, along with other caregivers, were invited to participate, especially where they could support older people to participate by helping with transport and mobility.

Participants were asked to attend both the lived-experience and the subsequent speculative-critique workshops at their locations. However, for various reasons, some participants were unable to attend both workshops. As far as possible, the workshops took place in locations that were convenient for participants to attend and which, as far as possible, were close to their "natural environments" (such as communal lounges or university collaborative spaces) and so conducive to eliciting their experiences, while also being appropriate for group discussions. Likewise, workshops were scheduled for times considered most convenient for participants. Each workshop typically lasted around 4 hours, including breaks and refreshments. There was an interval of approximately 2 weeks between the lived-experience workshop and the corresponding speculative-critique workshop, except for the last pair of workshops, which, due to scheduling pressures, were held as morning and afternoon sessions on the same day. Before the workshop, participants were asked to complete a consent form and a demographics questionnaire. At the end of the workshop, they were invited to give feedback about the nature, structure, and conduct of the workshop, and were each given a shopping voucher as an honorarium to acknowledge their contribution.

Data Analysis Method

The audio recordings of the workshops were transcribed by professional transcribers; the transcriptions, alongside the notes collected during the workshops, were then subjected to a thematic analysis by two of the research team, one with experience of assistive technology development (SP), the other with a background in industrial design and robotics (AH). The analysis broadly followed the Framework Method [38] as elaborated by Gale et al [39]. A phenomenological approach was adopted for analysis, with the analysts bracketing their own experiences and conceptions to reduce bias and focus on the content of the participants' contributions, paying particular attention to their experiences and the meanings that they attached to these. Specifically, we adopted an interpretive phenomenological analysis, foregrounding the participants' experiences and perceptions, while recognizing that the researchers play an active role in interpreting these [40,41]. This analysis was both deductive and inductive. Given the aims of the exercise, we had a particular deductive focus on the lived experience of aging and how this relates to opportunities for and constraints upon assistive robotics, as dictated by the structure of the workshops. However, within these broad topics, we were open to themes that were not explicitly foreseen as topics for discussion, and which arose naturally during the workshops. Codes were first inductively identified from the participants' utterances and then checked against the notes taken by the facilitators. These were then clustered deductively into higher categories which broadly mirror the structure of the workshops, namely "difficulties with activities of daily life," "opportunities for facilitating activities of daily life," "assistive robotics design requirements," and "assistive robotics design constraints." The analysts worked independently at first, then collaborated to merge codes, resolving any discrepancies through discussion. Several iterations of analysis, during which codes and categories were amended, were performed until a reasonably stable analytical framework emerged, with the first 2 themes emended to "everyday difficulties" and "ideas for aging better," respectively, and the last 2 merged into the more general "living with technology" category. The transcription and analysis of the data were carried out using the Lumivero NVivo qualitative analysis software package. Although the notes taken were displayed to participants for "real-time" validation during the workshops, it was unfortunately not feasible to ensure dependability by reconvening the participants and presenting the analysis back to them for their approval or correction of the framework.

Ethical Considerations

This study was approved by the Ethics Committee of the University of Nottingham (CS-2021-R40). All participants provided written informed consent. Privacy and confidentiality were ensured by collecting no personally identifying data. Participation in this study was voluntary; each participant was given a shopping voucher (worth approximately US $30) as an honorarium.

Results

Overview

The older people participants in the workshops were aged from 55 to 96 years. An overview of participants for each workshop can be found in Table 1.

The examples of the illustrations produced during the workshops, shown in Figures 5 and 6, give some idea of the content of the discussions.

The analysis of the workshop outputs led to the formation of 3 broad, high-level categories, namely "everyday difficulties," "ideas for aging better," and "living with technology."

XSL•FO
RenderX

Table 1. Participant information for each of the co-design workshops.

Workshop type	Location	Participants		Setting
		Older people	Relatives, friends, and caregivers	
L[a]	A	8 (4 F[b], 4 M[c])	1 (1 F)	Lounge in residential scheme
L	B	8 (5 F, 3 M)	5 (3 F, 2 M)	University collaborative space
L	C	4 (2 F, 2 M)	4 (2 F, 2 M)	University collaborative space
S[d]	A	5 (3 F, 2 M)	2 (2 F)	Lounge in residential scheme
S	B	8 (5 F, 3 M)	3 (2 F, 1 M)	University collaborative space
S	C	4 (2 F, 2 M)	4 (2 F, 2 M)	University collaborative space

[a]L: lived experience.
[b]F: female.
[c]M: male.
[d]S: speculative critique.

Figure 5. Tableau illustration of discussions during a lived-experience workshop (location A).

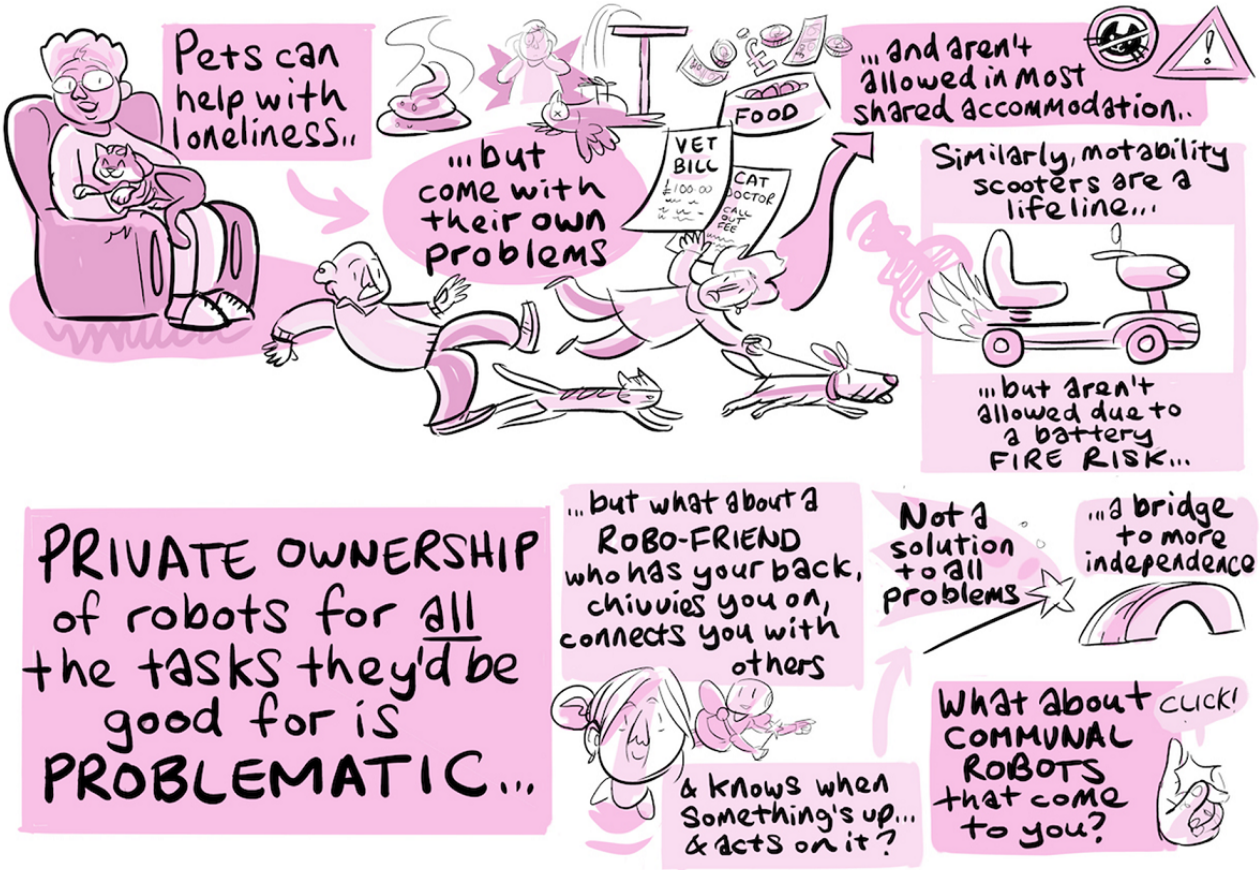

XSL•FO
RenderX

Figure 6. Tableau illustration of discussions during a lived-experience workshop (location C).

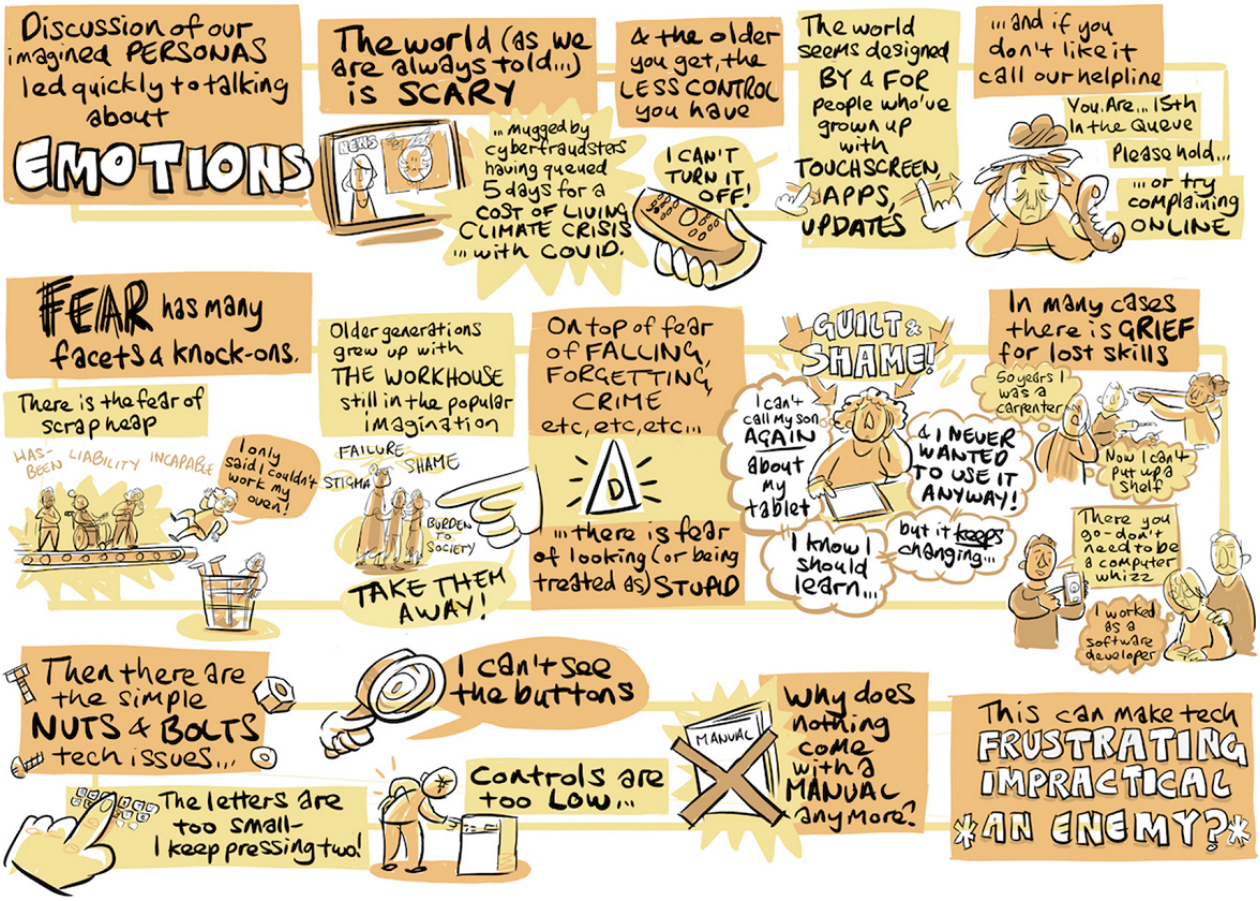

Everyday Difficulties

The most prevalent codes in the category of everyday difficulties are physical difficulties and psychological difficulties, with other difficulties being less prominent, if not less significant for participants.

- Physical difficulties: as was intended through the format of the lived-experience workshops, physical difficulties were often expressed in terms related to activities of daily living or situated in contexts. For example, participants mentioned the difficulties they have doing the housework (reaching high places, cleaning windows, and making the bed), with personal care (difficulties washing themselves and getting dressed), toileting (incontinence), mobility (walking up hills and stairs, and getting on and off buses), and, in general terms (not exercising enough).
- Psychological difficulties: almost equally prevalent were psychological difficulties. That these are shared by many participants might have been expected; what was more surprising was the extent to which participants were prepared to discuss their own psychological difficulties (and in the process setting aside the fictitious problems faced by the personas). Some of these were related to specific activities; others were more generalized states of mind or mood. Among other things, participants mentioned anxiety (about their own situations, but also more generalized, about the state of the world), social isolation and loneliness, lack of motivation and purpose, depression, and fear (of falling, of going out at night, when answering the door, and of assistance not being available when most needed).
- Cognitive difficulties: principally these concern remembering (appointments, where things are, to take medication, and to recharge and use technology, such as a smartwatch intended to help monitor activity levels).
- Living environment difficulties: these included having limited opportunities to socialize, poor relations with neighbors, lack of convenient public transport, and being unable to keep pets or mobility scooters in shared accommodation (the latter due to size and safety concerns, possibly having implications for assistive robotics).
- Burden of disease: for the workshop participants, this burden included remembering to take medication (overlapping with cognitive difficulties) and maintaining medication routines, the detrimental side effects of medication, and having to care (in an unpaid capacity) for a spouse living with more severe frailty.
- Time pressures: the time required to perform certain activities dissuaded some participants from doing potentially beneficial things such as buying food and preparing meals, taking care of pets, and going out to exercise or socialize.
- Financial difficulties: the cost of living, and specifically the cost of gas and electricity, was mentioned by several participants (who worried that they would be unable to afford to operate robots or other assistive technologies).

XSL•FO
RenderX

Ideas for Aging Better

Participants also brought to the workshops their own ideas, some based on their own experiences of what works for them, for how people might age better. Some of these concerned living and lifestyle (for instance, exercises, stimulating games, even pet ownership as a motivation for staying engaged and active) and social ideas (conversation and group activities of various sorts). However, the most prevalent category was technological ideas, no doubt influenced by the wider context of the workshops and suggested by the problems faced by the personas (and by the participants themselves); these were further divided into:

- Technology for physical assistance: the category for which there were most suggestions, which included the feasible (and in some cases already available), such as smart speakers for home automation, robot vacuum cleaners, and teasmaids; adaptations of existing technology to the domestic sphere (escalators and self-cleaning toilets, which a participant had seen at an airport); voice-activated showers and personalized drinks and snacks dispensers; and more ambitious ideas involving robots: robots for helping to do the shopping, for cleaning high places, and for cleaning up after pets, automated kitchen surfaces and food packaging openers, and self-making beds.
- Technology for psychological assistance: including technology that tells you who is knocking at the front door, assuring that it is safe to open the door; robots for providing motivation, for recipe suggestions, for cultivating positive attitudes and providing "positive affirmations," and robotic cats for companionship.
- Technology for cognitive assistance: including devices for providing reminders (for doctors' appointments or social events, for medicines, food, and drink), and smartwatch navigation aids.
- Technology for social assistance: several participants noted that their smartphones where helping them to stay connected to friends and family, especially through multimodal (speech, video, and chat message) calls; another suggestion was for an app or similar that could allow caregivers to better communicate and coordinate their efforts; and shared smart facilities, such as "robot washing machines" could help to avoid conflicts in residential schemes and similar multihabitant environments.
- Technology for health monitoring: participants suggested that it would be beneficial to have technology that could detect if medication had actually been taken, for measuring nutrients in the body, and for raising the alarm in the case of a fall; and to have an all-purpose "well-being robot."

Some of the ideas above would find expression in the hypothetical robot designs developed for the speculative critique workshops.

Living With Technology

The final major category, living with technology, encompasses mainly (but not exclusively) codes identified during analysis of the speculative critique workshops. These cover general preferences (in turn covering aspects such as appropriate interface modalities several people expressed a desire—indeed, an expectation—that the robots would be voice-controlled), size (large enough to do the job, but small enough to fit into people's homes and lives) and appearance (nonthreatening, with some opting for humanoid and others for "cute" robots), but also ideas of "personalization" (sometimes dynamic or autonomous) of the robot to the user's (changing) needs, abilities and routines, with robot autonomy extending to regular maintenance tasks such as charging or cleaning itself, and even consideration of financial models for provision (such time-sharing robots with others).

There was some overlap here with general requirements for robots, with a requirement being identified when a participant used a modal verb expressing necessity for some feature or aspect (that a robot "needs" or "must have" it). Requirements refer to physical qualities (robustness and stability), maintenance (cleanable or even self-cleaning), the need for adequate training and practice in using the robot, and, on a more challenging level, the need for a robot to be "self-aware" enough to be able to explain what it is doing.

However, the code with the most items in this category is concerns about technology, which are, almost exclusively, concerns about robotics. These are wide-ranging, revealing the complexity of people's lives and how they imagine robots might disrupt those. The following subcodes give some flavor of the range of concerns (note that these codes are not mutually exclusive):

- Safety: perhaps unsurprisingly, the major source of concern for the participants is safety. Difficulties here were further coded as surrounding the physical safety of users and other beings in the vicinity of robots, psychological safety (specifically if the appearance of a robot might scare or upset the user or others, manipulate their sentiments, or increase social isolation), potentially detrimental side effects (such as discouraging activity or suggesting the wrong sort of exercise), and digital safety (digital surveillance, privacy, and unauthorized access to any data collected). A number of concerns here are related to what might be termed uncertain hazards, where the participants could envisage situations during the use of the robots in which their safety could be imperiled if appropriate safeguards are not in place. For instance, what happens if a robot intended to physically assist people up and down stairs, or a flying drone-type cleaning robot were to malfunction midoperation?
- Compatibility: participants raised practical concerns about how robots would fit into their everyday lives, including aspects such as social compatibility (for instance, will the neighbors be inconvenienced by a robot? How would it interact with pets?), physical compatibility (Where will the robot be stored? Will it require modifications to the home? Will it damage possessions?), and technical compatibility (Will it require broadband connectivity?).
- Usefulness-desirability-inclusivity: general questions were raised about the usefulness of the speculative robots, including whether there were simpler or cheaper ways of achieving the same ends using existing technology (such as tablet computers and smart speakers, and nondigital alternatives), about the longevity of the robots (would they end up gathering dust in the cupboard alongside the foot spa?), about the need for sensitive, nonstigmatizing, design

XSL•FO
RenderX

of robots (and other assistive technologies), and, indeed, whether there was any real need for robots at all ("Is this a generational thing?" asked a participant; and another: "Am I going to feel the same about robots as I do about other tech? Too much hassle!")

- Burden of use: the additional overheads of looking after a robot are another major source of concern. These overheads encompass cleaning, charging, and servicing robots, the training required, and the cognitive demands of remembering how to use the robot. To quote a participant, "We don't want to look after the robot–we want the robot to look after us!"
- Maintenance and service model: a related concern is the service model of which the robot constitutes a technical component (and which is often overlooked in robot development): who programs, sets up, and personalizes the robot? Do robots act alone or in concert with human assistants? Who is responsible for regular maintenance and repairing faults?
- Control: concerns were raised about how control is exercised over robots, and who is in control. Some of these relate to the immediate operation of the robot, especially where the behaviors are complicated or potentially hazardous (and, in the words of a participant, "Can you tell it to stop?"), and others to the underlying operation of the robot, such as who determines what type and amount of motivation or exercise a robot should suggest.
- Financial costs: the economics of assistive robotics were also raised: participants raised the costs of purchase, running, and maintaining a robot, but also of modifying their homes, if necessary. Some participants asked about the comparative costs, whether that be compared with a human caregiver providing the same services or with a conventional refurbishment of their homes to make them more "age-friendly."
- Techno-psychosocial effects: here we see what might be called the techno-psychosocial effects of aging. These include the fear of being deemed "obsolete" or a burden on society, in no small part due to a (real or perceived) inability to move with the times and adapt to new technologies, as well as more "conventional" fears of falling, or forgetting, or crime and so on; the guilt and shame of not being able to adapt to these new technologies (alongside a feeling of grief for those skills that by-and-by are lost as we age); the stigma that sometimes accompanies the use of assistive technologies; the bottled-up frustration and occasional releases of rage that come from dealing with an increasingly digital world; and the contribution all these factors make to a vicious circle of stress, anxiety, and depression, with accompanying relationship problems, lack of motivation, poor diet, and social isolation.
- Digital ethics issues: in addition to the ethical issues that run through many of the difficulties noted above, here we also encounter direct concerns around the surveillance that users of data-collecting robots might find themselves subject to, users' privacy, and the security of their data.
- Social and care implications: finally, several participants raised concerns about the wider impact of assistive robotics: its potential negative effect on the role and jobs of human caregivers and home-helps, and the worry that robot care might replace human care, with the assumption that a degradation in the quality of that care would necessarily follow. On the other hand, there are worries too about inclusivity, access to care services, and the place of those who are unable to adapt to—or who choose to opt out of—the brave, new digital world.

Discussion

Principal Findings

We aimed to gain a better understanding of the everyday lives of older people, especially with reference to the effects of aging and frailty, and to gauge their opinions and concerns about assistive robotics. Through a wide-ranging co-design exercise with older people and their caregivers, we have gained a compelling snapshot of older people's real lives. As might be expected, given the structure of the workshops, the emphasis lies on the problems that people face daily and on their concerns about a world where change seems to have accelerated, sweeping away old certainties, and where, increasingly, services are delivered using digital technologies about which they have little say and less control [42]. In this context, the work reported here is an attempt to redress this balance and to return to older people some agency in the development of assistive technologies. We have also highlighted general concerns that older people have about assistive robotics. While older people and their caregivers are open to, and in some cases, enthusiastic about, the role that assistive robotics could play in their lives and those of others, they raise real concerns about, among other things, safety and control, the burden of use, ethics and unintended side effects, and the financial and social costs of introducing robots into their lives.

The work reported here differs from previous assistive robotics co-design activities in several key aspects. First, in its scope, which embraced a wide range of experiences of growing older, with a focus on the individual rather than on some specific health condition or difficulty, as much previous work has tended to do (eg, helping mitigate dementia [43] or supporting mobility [22,26]). Second, the exercise was technology-neutral in that we did not structure the workshops around some particular robot or robotic platform and tried to stay impartial in the question of whether assistive robotics could be beneficial: to adopt a priori a particular robotic platform is to make design decisions, including the fundamental decision that a robot is a valid "solution" to an ill-defined problem [27]. Finally, the exercise was undertaken not as a step in a specific assistive-robotics development pathway, but as a service to the wider assistive robotics community exploiting the Emergence network's resources, skills, and access to engage with potential users, and as such, we have placed special emphasis on communicating the results. In methodological terms, rather than focusing on a particular problem or solution (as is often done in co-design activities), we chose to adopt a general approach and investigate the lives of older people more broadly, and without trying to reach any firm consensus about where to focus subsequent development efforts or trying to reconcile the at times contradictory opinions of participants. We hope that, by

XSL•FO
RenderX

explaining our co-design approach along with the results, assistive technology developers will be able to sift the evidence and draw sound, rational conclusions.

Communication and Dissemination of Results

One key aspect of the exercise reported here is the communication and dissemination of results: clearly, this is essential if the findings are to influence the development of assistive robotics. One novel aspect here is that we realized that the results of the graphic illustrator's work to document the workshops could come to play an innovative role in our dissemination activities. The illustrator produced a total of 11 tableaux from the 3 lived experience workshops (Figures 5 and 6 each show 1 of these tableaux). Each tableau, somewhat like a comic book page, consists of a number of thematically related vignettes illustrating the ideas, opinions, or concerns voiced by participants. It was realized that these vignettes could form powerful communication and design aids as self-contained glimpses into the lives of the participants. Accordingly, they were isolated and printed as a pack of playing card-sized "empathy cards" (Figure 7), complete with suggestions for workshop "games" that encourage players to explore the different perspectives, concerns, and opportunities the cards contain and to consider how these relate to their own projects.

Figure 7. Some of the empathy cards generated from tableaux vignettes.

Along with the empathy cards, the results of the analysis have formed the basis of a series of workshops, sandpits, and project funding calls, and have provided content for a multidisciplinary summer school for budding assistive robotics developers. Through these activities, we have tried to reorient the development of assistive robotics for older people in a more empathetic direction and, indeed, to foment a new generation of more empathetic roboticists.

In addition, we are in the process of making all our workshop materials and the analysis results open and accessible in digital formats to the wider robotics and assistive technologies communities. This content includes the personas, the speculative robotics designs, the results of the analysis, and the empathy cards (which are also available as a physical pack of cards). The process and content of communicating outcomes in an accessible, understandable, and actionable manner is something that is not always given due prominence when discussing co-design exercises. This paper constitutes an element of our communication strategy, but it is by no means the culmination of our labors.

At the time of writing, it is too soon to draw any firm conclusions about the success of these endeavors: the development of assistive technologies and, in particular, of assistive robotics is a difficult and slow process which, typically, extends over several years. Only time will tell whether our efforts have contributed to bringing about the intended effect, and we see assistive robots emerge from the laboratory to help older people in their everyday lives.

Limitations

We recognize a number of limitations of this co-design exercise. All co-design practice is subject to bias, and this is no exception. The structure of the workshops, the development and choice of the material that provided the stimuli for them, the data collection by facilitators during the workshops, and the analysis of that data are all prey to the biases and preconceived notions of the researchers. The co-design approach adopted was, in some sense, an attempt to counter these biases: by not focusing on any particular everyday problem or given robotic platform, we have attempted to give participants the freedom to steer the discussions. In addition, the multidisciplinary nature of the

XSL•FO
RenderX

research team (which included specialists in design, assistive technologies, and health and care provision, as well as robotics) helped to guard against adopting preconceived positions or attitudes.

Further limitations concern the representativeness of the participants. The co-design activities involved a total of only 20 older people, plus a smaller number of relatives, friends, and caregivers. The participants were healthy enough (or coping well enough) to attend workshops and possessed some degree of day-to-day independence. Hence, people living with more severe frailty or other age-related health conditions (such as dementia) were not represented directly in the workshops, and no consideration was given to the particularities of life in care homes and other facilities offering specialized care (although the use of personas with more severe frailty and health issues was an attempt to address, at least in part, these aspects). Clearly, there is much scope for developing assistive robotics for people with greater needs and for the staff who minister to those needs, but that would require other, more focused co-design exercises. To this end, we have also conducted workshops with health care professionals, which—as may be expected—focused more specifically on the needs of people with more severe frailty and health conditions (including end-of-life care) and with a greater emphasis on support for their caregivers. As the outcomes of these workshops are thematically quite different from those reported here, this work will be reported separately.

Furthermore, the participants in this exercise were self-selecting, and as such might be considered to have a greater interest in robotics (and digital technologies more generally), and possibly to be better disposed to the acceptance of domestic assistive robotics. We did not collect data about the socioeconomic status or the digital literacy and experience of participants, factors that can influence attitudes to robotics. It was not feasible to reconvene participants to confirm or correct the analysis of the workshop data.

In demographic terms, all the older participants self-identified as having a White British ethnic background (almost certainly a reflection of the constitution of independent-living schemes from which most participants were recruited rather than any explicit bias in recruitment practices). Looking beyond to other, perhaps underserved, communities and their living circumstances could expose a different set of needs and wants. Although we found little direct evidence of differences in attitudes according to age or gender, these moderators will likely have some influence on acceptance, even if it was not explored by our participants. To assume that all older people will have the same attitudes to technology, disregarding other factors in their background, is to risk adhering to ageist stereotypes [44]. The very term "older people," with the suggestion it conjures of a homogeneous mass of people, can itself be a barrier. However, we would argue that experience is a more influential moderator than age: many older people, including some of our workshop participants, will have had extensive experience with digital technology throughout their working lives (which, in the case of a couple of participants, was ongoing) or because they have readily embraced it in their social lives, and are receptive to and comfortable with the idea of using digital assistive devices. Studies have suggested that perceived value and benefit of new technologies are more significant for technology acceptance than is chronological age [45,46]. Rather than being technophobic as a rule or having low levels of digital literacy, older people seem less likely to invest time in new digital technologies whose value for them is not apparent. However, while we have striven throughout to avoid adopting ageist stereotypes, we are aware that this is an all-too-easy pitfall (and, indeed, we noted during the workshops that older people themselves sometimes fall back on these facile generalizations).

In summary, we could not hope to capture everyone's experience of frailty and aging, or even to encompass the complexity of a single individual's experience. Moreover, the results of this exercise are specific to a particular population, time, and place (and as such have a limited shelf-life): namely, they are valid for (a small subset of) older people living in the United Kingdom in the early 2020s. People's experiences in different places and times will, of course, be different. If we were to repeat this exercise in, say, a generation's time, we would expect the needs and expectations of older people to diverge dramatically from those seen in our results. By this measure, there is no such thing as a uniform, constant, consistent experience of frailty and aging. In some sense, this is not a failing of this exercise but rather an essential feature of co-design, albeit one rarely acknowledged in the literature. Design always responds to problems and opportunities situated in particular times and places; seen in this light, this is not a limitation of the snapshot, but a fundamental aspect of it. Good designs, or ones that address a particular need, might be able to transcend time and place to some extent, but of necessity they must be grounded in the here-and-now experience of everyday life.

Conclusions

The contributions of this paper can be summarized as follows. We have described the predominant clinical models of frailty and explained why these form a useful but incomplete basis for designing assistive technologies of any sort for people living with frailty and older people more generally. We have described the current underwhelming state of assistive robotics, with common failings during their design or co-design processes that often result in the development of inappropriate robotic services.

We have described a novel co-design methodology that has attempted to combine persona-based lived-experience workshops with provotype-based speculative-design workshops with older people to gain a rounded "phenomenological snapshot" of the experience of being old in an increasingly digital world, a world in which assistive robotics may soon become commonplace. As far as we are aware, this is the first time that an exercise of this scope has been attempted in the context of assistive robotics. Finally, we have outlined the major concepts that are revealed by an analysis of the results of the workshops. These can be categorized thematically as: everyday difficulties, the problems faced daily by people living with frailty and older people more generally; ideas for aging better, older people's own suggestions for how their lives could be improved; and living with technology. In addition to participants' preferences and requirements, this last category reveals the wide-ranging concerns that people have about services based on digital

XSL•FO
RenderX

technologies, and about assistive robots in particular, whose development is viewed with a mixture of enthusiasm and unease.

Acknowledgments

The authors gratefully acknowledge the contributions of all those who participated in the Emergence co-design workshops described herein, the assistance of Ms Rebekah Moore, the project manager, in organizing the workshops, and the illustrative skills of Mr Sam Church. Generative artificial intelligence was not used for any aspect of this paper's writing.

Data Availability

The data generated and analyzed during this study are available from the corresponding author (SP) on reasonable request.

Funding

Funding for this research was provided through the Emergence: "Tackling Frailty – Facilitating the Emergence of Healthcare Robots From Labs Into Service Healthcare Technologies Network+" grant awarded by the UK Engineering and Physical Sciences Research Council (EPSRC EP/W0000741/1). The funder had no involvement in this study's design, data collection, analysis, interpretation, or the writing of this paper. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any author-accepted manuscript version arising from this submission.

Authors' Contributions

Conceptualization: PC-S (lead), MH (equal), FA (equal), MD (equal), ADN (equal)
Formal analysis: SP (lead), AH (equal), MH (supporting), PC-S (supporting)
Funding acquisition: PC-S (lead), MH (supporting), FA (supporting), MD (supporting), ADN (supporting)
Investigation: SP (lead), PC-S (equal), MH (equal), FA (equal), MD (equal), ADN (equal)
Methodology: SP (lead), MH (supporting)
Project administration: PC-S (lead), SP (equal), FA (supporting), MD (supporting)
Visualization: SP (lead), AH (equal)
Writing – original draft: SP (lead), AH (supporting)
Writing – review & editing: SP (lead), MH (supporting), AH (supporting), FA (supporting), MD (supporting), ADN (supporting), PC-S (supporting)

Conflicts of Interest

None declared.

References

1. Wright J. Robots Won't Save Japan: An Ethnography of Eldercare Automation. New York: ILR Press; 2003.
2. Xue Q. The frailty syndrome: definition and natural history. Clin Geriatr Med 2011;27(1):1-15 [FREE Full text] [doi: 10.1016/j.cger.2010.08.009] [Medline: 21093718]
3. Clegg A, Bates C, Young J, Ryan R, Nichols L, Ann Teale E, et al. Development and validation of an electronic frailty index using routine primary care electronic health record data. Age Ageing 2016;45(3):353-360 [FREE Full text] [doi: 10.1093/ageing/afw039] [Medline: 26944937]
4. Fried LP, Tangen CM, Walston J, Newman AB, Hirsch C, Gottdiener J, Cardiovascular Health Study Collaborative Research Group. Frailty in older adults: evidence for a phenotype. J Gerontol A Biol Sci Med Sci 2001;56(3):M146-M157. [doi: 10.1093/gerona/56.3.m146] [Medline: 11253156]
5. Walston JD, Bandeen-Roche K. Frailty: a tale of two concepts. BMC Med 2015;13:185 [FREE Full text] [doi: 10.1186/s12916-015-0420-6] [Medline: 26265077]
6. Rockwood K, Andrew M, Mitnitski A. A comparison of two approaches to measuring frailty in elderly people. J Gerontol A Biol Sci Med Sci 2007;62(7):738-743. [doi: 10.1093/gerona/62.7.738] [Medline: 17634321]
7. Rockwood K, Mitnitski A. Frailty in relation to the accumulation of deficits. J Gerontol A Biol Sci Med Sci 2007;62(7):722-727. [doi: 10.1093/gerona/62.7.722] [Medline: 17634318]
8. Lang IA, Llewellyn DJ, Langa KM, Wallace RB, Huppert FA, Melzer D. Neighborhood deprivation, individual socioeconomic status, and cognitive function in older people: analyses from the English Longitudinal Study of Ageing. J Am Geriatr Soc 2008;56(2):191-198 [FREE Full text] [doi: 10.1111/j.1532-5415.2007.01557.x] [Medline: 18179489]
9. Werner C, Moustris GP, Tzafestas CS, Hauer K. User-oriented evaluation of a robotic rollator that provides navigation assistance in frail older adults with and without cognitive impairment. Gerontology 2018;64(3):278-290. [doi: 10.1159/000484663] [Medline: 29183035]
10. Koumpouros Y, Toulias TL, Tzafestas CS, Moustris GP. Assessment of an intelligent robotic rollator implementing navigation assistance in frail seniors. Technol Disability 2020;32(3):159-177. [doi: 10.3233/tad-200271]

XSL•FO
RenderX

11. Ozaki K, Kondo I, Hirano S, Kagaya H, Saitoh E, Osawa A, et al. Training with a balance exercise assist robot is more effective than conventional training for frail older adults. Geriatr Gerontol Int 2017;17(11):1982-1990. [doi: 10.1111/ggi.13009] [Medline: 28295912]
12. Hai NDX, Thinh NT. Self-feeding robot for elder people and Parkinson‘s patients in meal supporting. Int J Mech Eng Robot Res 2022;11(4):241-247 [FREE Full text]
13. Luperto M, Monroy J, Renoux J, Lunardini F, Basilico N, Bulgheroni M, et al. Integrating social assistive robots, IoT, virtual communities and smart objects to assist at-home independently living elders: the movecare project. Int J Soc Robot 2023;15(3):517-545 [FREE Full text] [doi: 10.1007/s12369-021-00843-0] [Medline: 35194482]
14. Pollak C, Wexler SS, Drury L. Effect of a robotic pet on social and physical frailty in community-dwelling older adults: a randomized controlled trial. Res Gerontol Nurs 2022;15(5):229-237. [doi: 10.3928/19404921-20220830-01] [Medline: 36113009]
15. Yamazaki Y, Ishii M, Ito T, Hashimoto T. Frailty care robot for elderly and its application for physical and psychological support. J Adv Comput Intell Intell Inform 2021;25(6):944-952. [doi: 10.20965/jaciii.2021.p0944]
16. Boumans R, van Meulen F, Hindriks K, Neerincx M, Olde Rikkert MGM. Robot for health data acquisition among older adults: a pilot randomised controlled cross-over trial. BMJ Qual Saf 2019;28(10):793-799 [FREE Full text] [doi: 10.1136/bmjqs-2018-008977] [Medline: 30894423]
17. Civit A, Andriella A, Barrue C, Antonio M, Boqué C, Alenyà G. Introducing social robots to assess frailty in older adults. Boulder, CO: ACM; 2024 Presented at: HRI '24: Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction; March 11, 2024; Boulder, CO p. 342-346.
18. Olde Keizer RACM, van Velsen L, Moncharmont M, Riche B, Ammour N, Del Signore S, et al. Using socially assistive robots for monitoring and preventing frailty among older adults: a study on usability and user experience challenges. Health Technol 2019;9(4):595-605. [doi: 10.1007/s12553-019-00320-9]
19. Arunachalam S. Do socially assistive robots help in caring for elderly patients with frailty or mild cognitive impairment? J Stud Res 2023;12(3). [doi: 10.47611/jsrhs.v12i3.4821]
20. Lunardini F, Luperto M, Romeo M, Renoux J, Basilico N, Krpic A, et al. The MOVECARE project: home-based monitoring of frailty. Chicago, IL: IEEE; 2019 Presented at: IEEE EMBS International Conference on Biomedical & Health Informatics (BHI); May 19-22, 2019; Chicago, IL.
21. Kim J, Choi Y, Jeong S, Han J. A care robot with ethical sensing system for older adults at home. Sensors (Basel) 2022;22(19):7515 [FREE Full text] [doi: 10.3390/s22197515] [Medline: 36236614]
22. Fiorini L, Tabeau K, D’Onofrio G, Coviello L, De Mul M, Sancarlo D, et al. Co-creation of an assistive robot for independent living: lessons learned on robot design. Int J Interact Des Manuf 2020;14(2):491-502. [doi: 10.1007/s12008-019-00641-z]
23. Steen M. Co-design as a process of joint inquiry and imagination. Design Issues 2013;29(2):16-28. [doi: 10.1162/DESI_a_00207]
24. García-Soler Á, Facal D, Díaz-Orueta U, Pigini L, Blasi L, Qiu R. Inclusion of service robots in the daily lives of frail older users: a step-by-step definition procedure on users' requirements. Arch Gerontol Geriatr 2018;74:191-196. [doi: 10.1016/j.archger.2017.10.024] [Medline: 29128788]
25. Fiorini L, D'Onofrio G, Rovini E, Sorrentino A, Coviello L, Limosani R, et al. A robot-mediated assessment of Tinetti balance scale for sarcopenia evaluation in frail elderly. New Delhi, India: IEEE; 2019 Presented at: 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN); October 14-18, 2019; New Delhi, India.
26. Coviello L, Cavallo F, Limosani R, Rovini E, Fiorini L. Machine learning based physical human-robot interaction for walking support of frail people. Berlin, Germany: IEEE; 2019 Presented at: 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); July 23-27, 2019; Berlin, Germany.
27. Bardaro G, Antonini A, Motta E. Robots for elderly care in the home: a landscape analysis and co-design toolkit. Int J Soc Robotics 2022;14:657-681. [doi: 10.1007/s12369-021-00816-3]
28. Spiers J, Smith JA. Interpretative phenomenological analysis. In: Atkinson P, Delamont S, Cernat A, Sakshaug JW, Williams RA, editors. SAGE Research Methods Foundations. London: SAGE Publications Ltd; 2019.
29. Pruitt J, Adlin T. The Persona Lifecycle: Keeping People in Mind Throughout Product Design. San Francisco, CA: Morgan Kaufmann; 2006.
30. van den Heuvel H, Huijnen C, Caleb-Solly P, Nap H, Nani M, Lucet E. Mobiserv: a service robot and intelligent home environment for the provision of health, nutrition and safety services to older adults. Gerontechnology 2012;11(2):373. [doi: 10.4017/gt.2012.11.02.564.00]
31. Wöckl B, Yildizoglu U, Buber I, Aparicio Diaz B, Kruijff E, Tscheligi M. Basic senior personas: a representative design tool covering the spectrum of European older adults. New York, NY: ACM; 2012 Presented at: ASSETS '12: Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility; October 22, 2012; Boulder, CO.
32. Valenza C, Adkins J. TIMELINES: understanding visual thinking: the history and future of graphic facilitation. Interactions 2009;16(4):38-43. [doi: 10.1145/1551986.1551994]
33. Espiner D, Hartnett F. Innovation and graphic facilitation. Aotearoa New Zealand Social Work 2016;28(4):44-53 [FREE Full text]

XSL•FO
RenderX

34. Boer L, Donovan J. Provotypes for participatory innovation. New York, NY: ACM; 2012 Presented at: DIS '12: Proceedings of the Designing Interactive Systems Conference; June 11, 2012; Newcastle Upon Tyne, United Kingdom.
35. Dunne A, Raby F. Speculative Everything: Design, Fiction, and Social Dreaming. Cambridge, MA: The MIT Press; 2013.
36. Sellen KM, Massimi MA, Lottridge DM, Truong KN, Bittle SA. The people-prototype problem: understanding the interaction between prototype format and user group. Boston MA: ACM; 2009 Presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; April 4, 2009; Boston, MA. [doi: 10.1145/1518701.1518799]
37. Hogan DB, MacKnight C, Bergman H, Steering Committee, Canadian Initiative on Frailty and Aging. Models, definitions, and criteria of frailty. Aging Clin Exp Res 2003;15(3 Suppl):1-29. [Medline: 14580013]
38. Ritchie J, Lewis J. Qualitative Research Practice: A Guide for Social Science Students and Researchers. London: SAGE Publications Ltd; 2003.
39. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013;13:117 [FREE Full text] [doi: 10.1186/1471-2288-13-117] [Medline: 24047204]
40. Tuffour I. A critical overview of interpretative phenomenological analysis: a contemporary qualitative research approach. J Healthc Commun 2017;02(04):1-5. [doi: 10.4172/2472-1654.100093]
41. Neubauer BE, Witkop CT, Varpio L. How phenomenology can help us learn from the experiences of others. Perspect Med Educ 2019;8(2):90-97 [FREE Full text] [doi: 10.1007/s40037-019-0509-2] [Medline: 30953335]
42. Briefing: facts and figures about digital inclusion and older people. Age UK. 2024. URL: https://www.ageuk.org.uk/siteassets/documents/reports-and-publications/reports-and-briefings/active-communities/internet-use-statistics-june-2024.pdf [accessed 2026-01-15]
43. Moharana S, Panduro AE, Lee HR, Riek LD. Robots for joy, robots for sorrow: community based robot design for dementia caregivers. 2019 Presented at: 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI); March 11-14, 2019; Daegu, Korea (South). [doi: 10.1109/hri.2019.8673206]
44. Mannheim I, Wouters EJM, Köttl H, van Boekel LC, Brankaert R, van Zaalen Y. Ageism in the discourse and practice of designing digital technology for older persons: a scoping review. Gerontologist 2023;63(7):1188-1200 [FREE Full text] [doi: 10.1093/geront/gnac144] [Medline: 36130318]
45. Hauk N, Hüffmeier J, Krumm S. Ready to be a silver surfer? A meta-analysis on the relationship between chronological age and technology acceptance. Comput Hum Behav 2018;84:304-319. [doi: 10.1016/j.chb.2018.01.020]
46. Berkowsky RW, Sharit J, Czaja SJ. Factors predicting decisions about technology adoption among older adults. Innov Aging 2018;2(1):igy002 [FREE Full text] [doi: 10.1093/geroni/igy002] [Medline: 30480129]

Edited by A Stone; submitted 21.May.2025; peer-reviewed by T Bernier, A Olsson; comments to author 25.Nov.2025; accepted 31.Dec.2025; published 28.Jan.2026.

Please cite as:
Potter S, Hawley M, Higgins A, Amirabdollahian F, Dragone M, Di Nuovo A, Caleb-Solly P
Assistive Robotics for Healthy Aging: A Foundational Phenomenological Co-Design Exercise
J Med Internet Res 2026;28:e77179
URL: https://www.jmir.org/2026/1/e77179
doi:10.2196/77179
PMID:

©Stephen Potter, Mark Hawley, Angela Higgins, Farshid Amirabdollahian, Mauro Dragone, Alessandro Di Nuovo, Praminda Caleb-Solly. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Adoption of Internet of Things in Health Care: Weighted and Meta-Analytical Review of Theoretical Frameworks and Predictors

Inês Veiga[1], MSc; Tiago Oliveira[1], PhD; Mijail Naranjo-Zolotov[1], PhD; Ricardo Martins[1], PhD; Stylianos Karatzas[2], PhD

[1]NOVA Information Management School (NOVA IMS), Lisbon, Portugal

[2]Department of Product & Systems Design Engineering, University of the Aegean, Hermópolis, Greece

Corresponding Author:
Inês Veiga, MSc
NOVA Information Management School (NOVA IMS)
Universidade Nova de Lisboa Campus de Campolide
Lisbon, 1070-312
Portugal
Phone: 351 213 828 610
Email: iveiga@novaims.unl.pt

Abstract

Background: The integration of the Internet of Things (IoT) into health care is transforming the industry by enhancing disease care and management, as well as supporting self-health management. The COVID-19 pandemic has accelerated the adoption of IoT devices, particularly wearable medical devices, which enable real-time health monitoring and advanced remote health management. Globally, the increased adoption of IoT in health care has improved efficiency, enhanced patient care, and generated substantial economic value.

Objective: This review aims to conduct a comprehensive meta- and weight analysis of quantitative studies to identify the most influential predictors and theoretical frameworks explaining the adoption of IoT in health care.

Methods: We searched databases, including Web of Science and PubMed, for quantitative studies on IoT health care adoption, with the last search conducted in early July 2025. Inclusion criteria comprised peer-reviewed articles written in English that employed a quantitative approach to IoT health care technology adoption. Studies were excluded if they did not report the significance of relationships, involved technologies without IoT features or were outside the scope, or examined target variables irrelevant to the analysis. The weight analysis identified the pathways with the most significant effects. A meta-analysis using a random-effects model was conducted to estimate combined effect sizes and their statistical significance. The results from both methods were then integrated to visualize the most frequently used theoretical frameworks. Risk of bias and heterogeneity were assessed using a funnel plot, Egger regression test, the I2 statistic, and subgroup analysis, which indicated no strong evidence of publication bias but revealed a high level of heterogeneity.

Results: Analysis of 115 datasets from 109 papers identified the Technology Acceptance Model and the Unified Theory of Acceptance and Use of Technology (UTAUT) as the primary frameworks for explaining IoT adoption in health care. Incorporating context-specific variables—such as health consciousness, innovativeness, and trust—into these traditional technology acceptance frameworks enhances the understanding of IoT adoption. Although high heterogeneity suggests a need to refine theoretical models to account for regional contexts, universal adoption drivers such as performance expectancy and effort expectancy remain consistent.

Conclusions: Behavioral intention is the most frequently studied variable in IoT health care adoption, whereas attitude, performance expectancy, effort expectancy, and task-technology fit remain underexplored. While adoption theories from the information systems field, such as the TAM, are predominantly used, integrating context-specific constructs and theories—such as trust and innovativeness—can provide deeper insights into IoT adoption in health care. The strongest and most consistent predictors of behavioral intention were attitude, performance expectancy, habit, self-efficacy, functional congruence, and benefits. Additionally, social influence, facilitating conditions, trust, and aesthetic appeal demonstrated promising or strong effects. By contrast, variables such as privacy and security, barriers, vulnerability, severity, compatibility, financial cost, health, and technology anxiety were generally inconsistent or not statistically significant.

(J Med Internet Res 2026;28:e64091) doi:10.2196/64091

XSL•FO
RenderX

KEYWORDS

IoT health care; internet of things; individual adoption; weight-analysis; meta-analysis

Introduction

The integration of the Internet of Things (IoT) into health care has revolutionized the industry by introducing a new paradigm of connectivity and data exchange, driven by rapid advancements in IoT, artificial intelligence, and machine learning [1-3]. This era, known as Healthcare 4.0, can be leveraged to enhance acute disease care, manage chronic diseases, and support self-health management [4]. The COVID-19 pandemic accelerated the adoption of user-friendly IoT devices [5-7], with wearable medical devices emerging as key allies by offering real-time health monitoring, continuous data transmission, and advanced remote health management [8-10].

Globally, the integration of IoT in health care has enhanced efficiency, improved patient care, and generated significant economic value [11,12]. By 2029, the global IoT health care market volume is projected to reach US $134.40 billion [13]. This indicates strong, sustained growth driven by the increasing adoption of IoT technologies in health care around the globe [13,14]. For instance, China has made significant progress in integrating health information technologies into the health care system, driven by initiatives such as “Internet Plus Health Care” and the “Healthy China 2030” plan [15-17]. The United States leads in IoT and intelligent health care system development, supported by substantial investments and a robust ecosystem of startups and tech companies driving advancements in artificial intelligence and IoT [18-20]. Europe also shows considerable progress, emphasizing regulatory frameworks, standardization, and interoperability to foster innovation and data protection [21,22].

While IoT holds considerable promise to transform health care by reducing costs and improving access, understanding the factors influencing its adoption requires more focused research [23]. Although literature reviews with a quantitative approach have examined technology adoption in health care, existing meta-analyses that include technologies with IoT features remain fragmented, as they have largely focused on broader or adjacent technological domains and have typically emphasized a specific adoption model. For instance, meta-analyses on mobile health have focused on the Unified Theory of Acceptance and Use of Technology (UTAUT) [24] and the Technology Acceptance Model (TAM) [25]. Meta-analyses on eHealth have predominantly focused on the TAM [26] and continuance intention [27]. Meta-analyses specific to smart wearable health care devices have examined attitude and intention using UTAUT and TAM [28], as well as the effects of perceived usefulness and perceived ease of use on intention, with a focus on Hofstede’s cultural dimensions as moderators [29]. Taken together, these studies often treat health care technology adoption in general terms and do not account for the unique characteristics of IoT.

Our study addresses this gap by providing a comprehensive meta-analysis and a weight analysis specifically focused on IoT adoption in health care. We synthesize findings from primarily quantitative articles on the adoption of IoT in health care, particularly on interconnected devices that monitor and transmit real-time health care data, enabling smarter solutions [5], such as smart sensors, remote monitoring devices, and health-focused IoT platforms. Our meta-analytic approach integrates findings from different theoretical perspectives, including technology adoption models such as the TAM and UTAUT and health-specific models such as the Health Belief Model (HBM), allowing for a more holistic understanding of adoption dynamics in health care contexts. Moreover, our dual-method approach, combining meta-analysis with weight analysis, identifies the strongest and most reliable predictors of adoption and maps the theoretical foundations most frequently and effectively used in this field. This analysis goes beyond prior reviews, offering new evidence-based insights to guide health care technology developers, practitioners, and researchers. The objectives of this study are, first, to identify key predictors of IoT health care adoption through a comprehensive meta-analysis and weight analysis, and second, to determine the most influential and empirically supported theories used to explain IoT adoption in health care settings.

Methods

Overview

We performed a meta-analysis to examine the factors influencing the adoption of IoT technologies in health care, synthesizing findings from a range of quantitative studies. By focusing on primary quantitative research articles, we aimed to identify the most significant predictors and the theoretical models most commonly used to explain IoT adoption in health care settings.

Information Sources and Search Strategy

This meta-analysis followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines [30]. The literature search was conducted using a keyword-based search across the Web of Science and PubMed databases to identify studies examining IoT adoption in health care. The search strategy incorporated title, abstract, and keyword searches, using Boolean operators (AND and OR) and database-specific filters. The keywords used in our search were related to IoT technology, relevant variables, quantitative methods, and exclusion criteria (Table 1). We included records published up to the end of 2024. The complete search strategy, including search terms and Boolean logic, is provided in Multimedia Appendix 1.

XSL•FO
RenderX

Table 1. Map for the keyword search in online databases.

Relevant terms	Relevant variables and theories	Methodologies	Exclusion of irrelevant topics
Internet of Things	Intention to adopt	Structural equation	Systematic
IoT	Behavioral intention	Structural equation modeling	Literature review
Smart	Acceptance	Partial least squares structural equation modeling	Postadoption
Intelligent	Adopt	Partial least	Meditation
Health care wearable device	Adoption	Path analysis	Contact tracing app
Medical wearable technology	Using	Regression	Fitness app
Health management	Use	N/A[a]	Electronic health record
Health measurement	Usage	N/A	Telemedicine
N/A	Intention to use	N/A	Mindfulness app
N/A	Unified Theory of Acceptance and Use of Technology 2	N/A	N/A
N/A	Unified Theory of Acceptance and Use of Technology	N/A	N/A
N/A	Technology Acceptance Model	N/A	N/A

[a]N/A: not applicable.

Selection Criteria

Initial screening was conducted using database filters. In the second screening, 2 (IV and MNZ) independent reviewers assessed the titles and abstracts for relevance, resolving disagreements through discussion or arbitration by a third reviewer. The full-text screening followed the same procedure. The reports assessed for eligibility were exported to an Excel (Microsoft Corporation) file, and all included papers were imported into Zotero (Corporation for Digital Scholarship), which is a reference management software. When a paper was unavailable, the authors were contacted.

Inclusion criteria comprised peer-reviewed articles with a quantitative approach to health care technology adoption written in English. Reasons for exclusion included not reporting the significance of the relationships between variables; the technology lacking IoT features or being unrelated to health care; the target variables being unrelated to adoption or focusing solely on postadoption behaviors; and studies lacking empirical data or reporting qualitative results only. The workflow and search conditions are depicted in more detail in the “Results” section.

Data Extraction

A standardized data extraction form was developed before data extraction. Data extraction was performed in Excel, and for each article, we detailed the study characteristics, methodology, type of technology, and the effects measured across multiple paths. The extracted aspects and their descriptions are provided in Table S1 in Multimedia Appendix 1. We assessed paper quality by examining the publishing journal metrics, the methods employed, sample size, and the scales used to measure each construct. The standardized β coefficients were extracted as the primary effect measure. As some authors used different names to represent the same variable, several variables had to be merged to conduct our analysis. This process was carried out by reading each variable definition and identifying the items used to measure them. Examples of variable mergers are provided in Table S2 in Multimedia Appendix 1, and the individual studies included in the analysis are detailed in Table S3 in Multimedia Appendix 1.

Descriptive Analysis

We extracted metadata from each study to perform a descriptive analysis of publication trends, journal quality, and research domains. Data on publication year were used to assess the chronological distribution of studies. To evaluate journal quality, we matched each journal with its SCImago Journal & Country Rank classification and categorized them into quartiles (Q1-Q4). The disciplinary scope of the journals was identified based on their SCImago subject area classifications. We also recorded the journal title and publication frequency to identify the journals that published the most research. Country-level data were extracted based on the origin of the study sample or study location. We computed the number of studies and total sample sizes per country to identify regions with the highest research activity. To understand the theoretical foundations employed across studies, we reviewed each article’s methodology and coded the theories used to model technology adoption behavior. We also recorded whether these models were used independently or in combination (eg, UTAUT extended with Protection Motivation Theory [PMT] constructs).

Weight Analysis

The weight analysis was conducted to uncover the predictive power of independent variables [31]. This weight provides a measure of the relative importance or consistency of statistical significance for each variable across multiple analyses. For the weight-analytic approach, we focused on the influence of each independent variable on several dependent variables and limited our analysis to relationships investigated 3 or more times

XSL•FO
RenderX

[32,33]. The weight (W_i) of an independent variable i is calculated as the ratio of the number of times it was found to be statistically significant (S_i) to the total number of times it was examined (E_i), as expressed in the following equation:

$$W_i = S_i/E_i$$

Meta-Analysis

Meta-analyses allow us to quantitatively compare effect sizes across relationships between constructs using suitable metrics to capture these effect sizes, including standardized regression coefficients [34,35]. This analysis followed best practices outlined previously [36-38]. In our study, the necessary inputs for performing the meta-analysis were the standardized regression coefficients (β) and the sample sizes for each relationship examined 3 or more times across studies. Following the approach of Peterson and Brown [37], β values were transformed into approximate correlation coefficients as $r=\beta+0.05$, where $\lambda=1$ [37]. All correlation coefficients were Fisher z-transformed to stabilize variance, and SEs were computed.

A random-effects model was used to account for both within- and between-study variance, justified by the heterogeneity in study populations, methods, and contexts. Random-effects weights were calculated using the DerSimonian and Laird model, and weighted mean effect sizes were computed using inverse-variance weights, which use tau-squared (τ^2) [39,40]. Heterogeneity was assessed using the Q statistic and the I^2 index [41]. We also calculated the lower and upper bounds of the 95% CIs, z scores, and 2-tailed P values to assess statistical significance and interpret the magnitude of the observed effects. Final pooled effect sizes and CIs were then back-transformed from Fisher z to the correlation coefficient metric (r). All calculations were performed manually in Excel.

Publication Bias Analysis

The Egger test was used to statistically examine the presence of publication bias by regressing the standard normal deviate on precision [42]. The analysis was performed using Excel's data analysis regression tool, which applies standard ordinary least squares regression, and a significant intercept ($P<.10$) was interpreted as evidence of asymmetry and possible publication bias. A funnel plot was constructed to visually assess publication bias using the tool Meta-Essentials [43]. The trim-and-fill method was used to estimate the number and influence of missing studies. Heterogeneity for the included studies was assessed using the I^2 statistic, where a value over 75% is interpreted as substantial heterogeneity, using the following formula, where k is the number of studies and Q the Cochran Q statistic:

$$I^2 = \max(0; \{Q - [k - 1]\}/Q) \times 100\%$$

To evaluate regional bias, we conducted a subgroup analysis with 2 groups: one comprising studies conducted in China and the other comprising studies conducted in the remaining countries. For each group, we calculated the combined effect sizes, SEs, CI lower and upper limits, and the I^2 statistic.

Combining Weight and Meta-Analysis Results: The Most Used Adoption Models

To synthesize the relative strength of relationships across studies and adoption models, we combined the results from the weight analysis and meta-analysis. The weight analysis assessed the consistency and prominence of specific predictors by calculating the proportion of studies that reported statistically significant relationships for each path, referred to as the weight. In parallel, the meta-analysis provided pooled average effect sizes and significance levels across studies using a random-effects model. This dual approach offers a more comprehensive understanding of which constructs consistently predict behavioral intention or usage in the context of IoT adoption in health care and enables an evidence-based comparison of theoretical frameworks based on empirical support.

We then visually mapped the structure of each adoption model, such as the TAM and the HBM, using conceptual diagrams. In these figures, each arrow represents a theoretical path, and its thickness reflects the weight. Thicker lines indicate a weight above 0.700, representing paths supported by a high proportion of studies. The numerical values attached to each path represent the average effect size based on the random-effects meta-analysis, along with the corresponding P value. This dual representation enables a clearer comparison between the predictive strength (effect size) and consistency (weight) of each construct within and across models.

Results

Descriptive Analysis

Papers on IoT health care adoption show an increasing trend, with 89 of the 109 (81.7%) studies in our analysis published between 2020 and 2025, and the earliest published in 2011. According to the SCImago Journal & Country Rank, most papers appeared in Q1 journals (n=63, 57.8%), followed by Q2 (n=36, 33%) and Q3 (n=10, 9.2%), with no papers published in Q4. These studies span major research areas related to health and medicine, information systems, and computer science. In total, 75 unique journals were represented, with *PLoS One* (n=6), *Frontiers in Public Health* (n=5), *Technological Forecasting and Social Change* (n=5), and *International Journal of Environmental Research and Public Health* (n=4) being the most frequently appearing journals (see Table S3 in Multimedia Appendix 1).

In our analysis of 109 studies (see Figure 1), we identified 115 unique datasets totaling 46,508 individuals (see Table S1 in the Multimedia Appendix 1). Studies conducted in China, South Korea, and the United States accounted for a large portion of the total sample and represented the greatest number of publications (see Table 2).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.

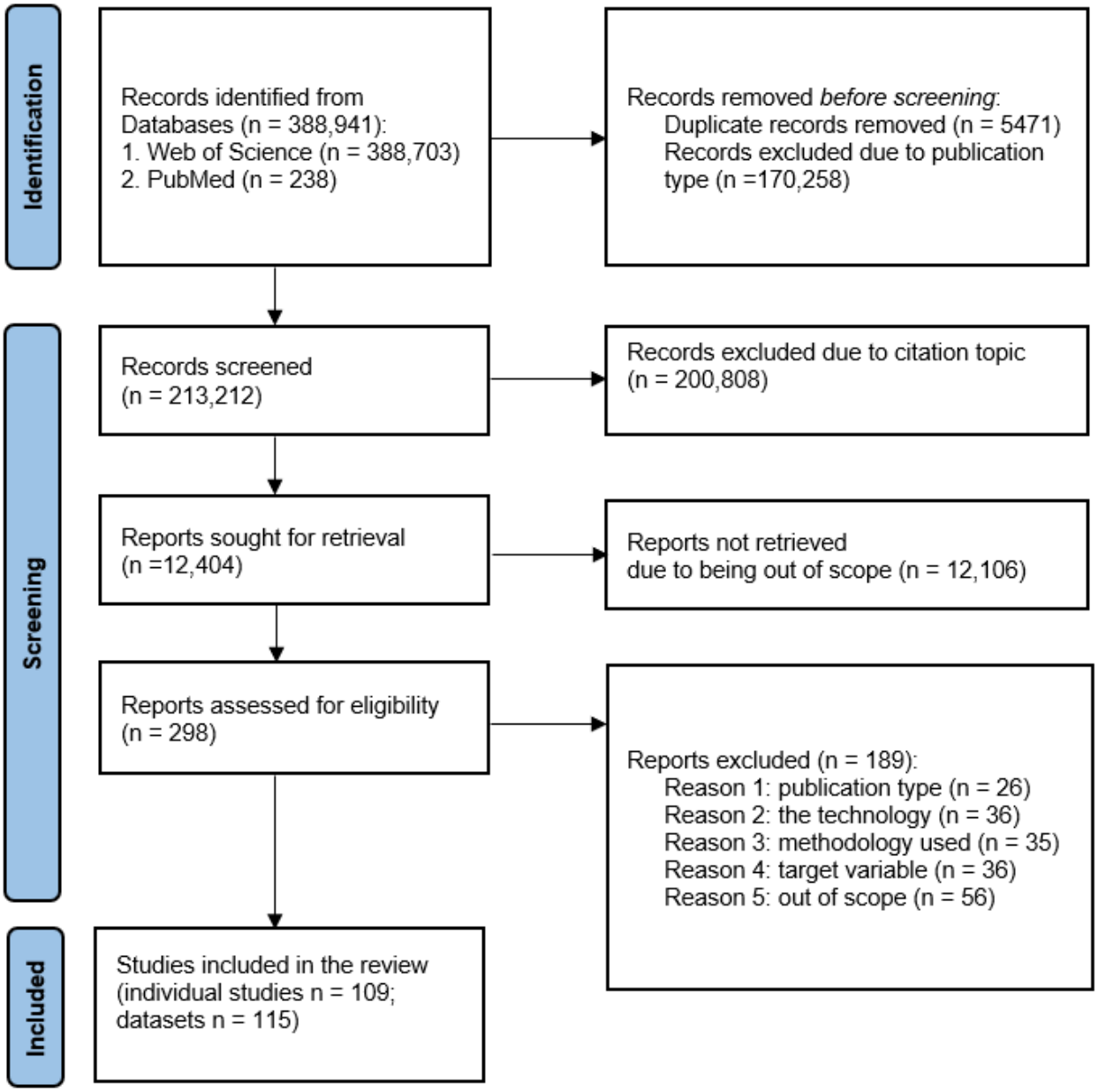

XSL•FO
RenderX

Table 2. Number of papers and total sample size per country.

Country	Dataset count (N=115)	Sample size (N=46,508)
China	39	17,068
South Korea	12	6406
The United States	9	5445
India	8	2924
Taiwan	8	1896
The Kingdom of Saudi Arabia	5	1798
Bangladesh	2	1213
Pakistan	3	1194
Turkey	4	1040
Ghana	2	965
Multiple countries	4	1312
Indonesia	1	772
Malaysia	2	628
France	3	515
Iraq	1	465
Oman	2	442
Romania	1	440
The United Arab Emirates	2	431
Switzerland	2	323
Singapore	1	306
Nepal	1	280
Japan	1	233
Italy	1	212
Jordan	1	200

Considering the theories addressed in each paper, the TAM and the UTAUT have been extensively examined compared with other theories (see Figure 2). These models serve as the theoretical foundation for 92 of the 109 (84.4%) papers included in our study. Other theories, such as the HBM, PMT, Task-Technology Fit (TTF) Theory, Privacy Calculus Theory, Diffusion of Innovation Theory, and Theory of Planned Behavior, have also been addressed—sometimes as the primary theoretical foundation and other times to extend the TAM or UTAUT.

XSL•FO
RenderX

Figure 2. Theoretical foundation of the papers included in our analysis. DOI: Diffusion of Innovation; HBM: Health Belief Model; PCT: Privacy Calculus Theory; PMT: Protection Motivation Theory; TAM: Technology Acceptance Model; TPB: Theory of Planned Behavior; TTF: Task-Technology Fit; UTAUT: Unified Theory of Acceptance and Use of Technology.

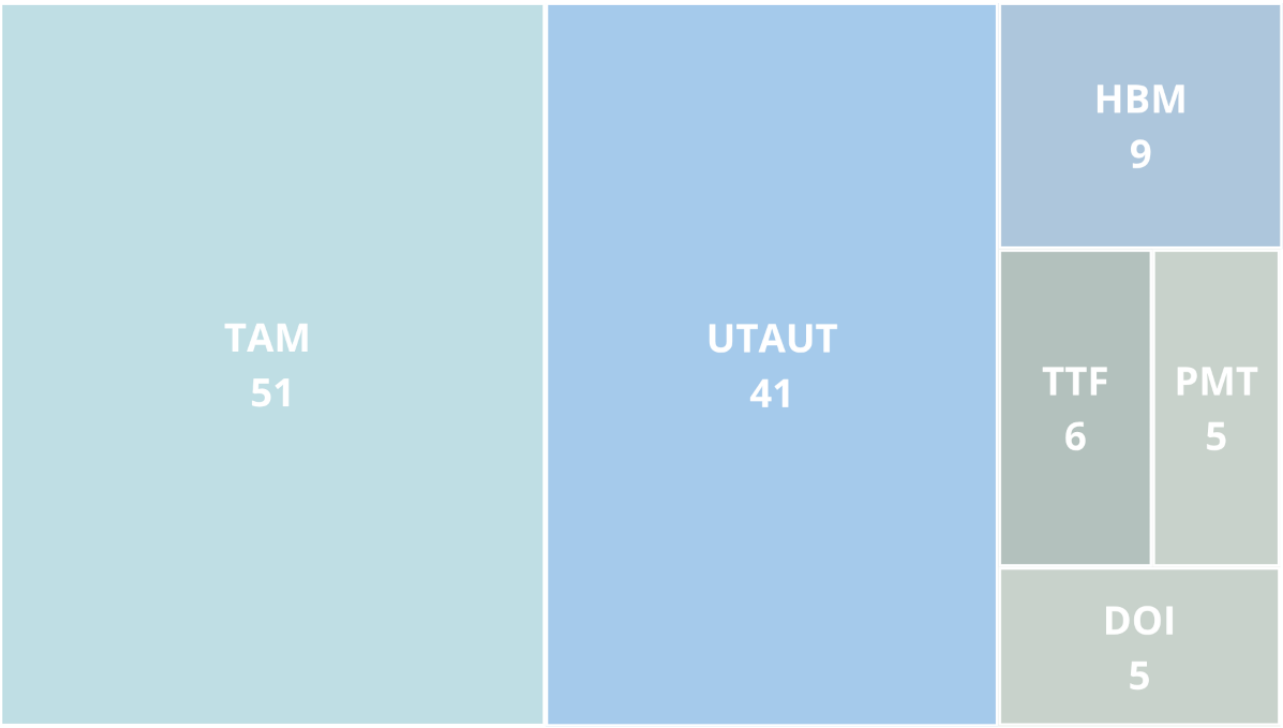

Weight Analysis

A weight analysis examines the strength of the relationship between an independent and a dependent variable. The weights of the identified relationships are analyzed and presented in Table 3. The significance of a relationship's weight is calculated by dividing the number of instances in which the relationship is statistically significant by the total number of studies that investigated it. A weight of 1 indicates that the relationship is significant in all examined studies, whereas a weight of 0 indicates that it is not significant in any of the studies.

XSL•FO
RenderX

Table 3. Identified paths with the nonsignificant paths, the significant relationships, the total paths, and the respective weights.

Dependent and independent variables	Significant	Nonsignificant	Total	Weight=significant/total
Attitude				
Effort expectancy	19	4	23	0.826
Barriers	4	2	6	0.667
Benefits	3	0	3	1
Facilitating conditions	4	0	4	1
Performance expectancy	20	2	22	0.909
Privacy and security	2	1	3	0.667
Social influence	6	1	7	0.857
Behavioral intention				
Aesthetic appeal	4	0	4	1
Attitude	27	1	28	0.964
Barriers	7	6	13	0.538
Benefits	6	0	6	1
Compatibility	4	3	7	0.571
Effort expectancy	36	23	59	0.61
Ethics	2	1	3	0.667
Facilitating conditions	20	8	28	0.714
Financial cost	12	9	21	0.571
Functional congruence	4	1	5	0.8
Habit	7	0	7	1
Health	3	1	4	0.75
Health consciousness	7	4	11	0.636
Hedonic motivation	10	5	15	0.667
Image	4	2	6	0.667
Innovativeness	5	3	8	0.625
Perceived severity	2	3	5	0.4
Perceived vulnerability	3	5	8	0.375
Performance expectancy	68	10	78	0.872
Privacy and security	13	13	26	0.5
Reliability	5	0	5	1
Self-efficacy	10	1	11	0.909
Social influence	31	10	41	0.756
Technology anxiety	2	4	6	0.333
Trust	9	3	12	0.75
Actual behavior				

XSL•FO
RenderX

Dependent and independent variables	Significant	Nonsignificant	Total	Weight=significant/total
Behavioral intention	16	1	17	0.941
Effort expectancy	2	1	3	0.667
Facilitating conditions	3	0	3	1
Health consciousness	2	2	4	0.5
Innovativeness	2	1	3	0.667
Perceived vulnerability	2	1	3	0.667
Performance expectancy	3	1	4	0.75
Social influence	3	0	3	1
Performance expectancy				
Barriers	2	3	5	0.4
Compatibility	6	2	8	0.75
Convenience	3	0	3	1
Effort expectancy	27	4	31	0.871
Facilitating conditions	2	2	4	0.5
Health consciousness	6	2	8	0.75
Image	3	4	7	0.429
Innovativeness	4	0	4	1
Privacy and security	4	2	6	0.667
Reliability	8	3	11	0.727
Self-efficacy	7	0	7	1
Social influence	8	3	11	0.727
Trialability	2	1	3	0.667
Trust	3	2	5	0.6
Task-technology fit	5	0	5	1
Effort expectancy				
Compatibility	7	0	7	1
Facilitating conditions	6	0	6	1
Image	2	2	4	0.5
Innovativeness	7	0	7	1
Privacy and security	2	2	4	0.5
Reliability	4	0	4	1
Self-efficacy	7	0	7	1
Social influence	3	1	4	0.75
Trialability	2	1	3	0.667
Task-technology fit	3	0	3	1
Task-technology fit				
Task characteristics	3	1	4	0.75
Technology characteristics	4	0	4	1

In the context of technology adoption at the individual level, independent variables are considered “well-utilized” if they have been tested at least 5 times. Variables tested fewer than 5 times but with a weight of 1 are regarded as “promising” predictors [31]. To be classified as a “best” predictor, an independent variable must have a weight of 0.800 or higher and must have been examined at least 5 times [31].

In our research, we analyzed the impact of several independent variables on the dependent variables attitude, behavioral intention, actual use, performance expectancy, effort expectancy,

XSL•FO
RenderX

and TTF. For the weight analysis, we included relationships that were examined 3 or more times, resulting in 67 relationships and 31 unique predictors that met this criterion. The most studied target variable was behavioral intention, with 25 predictors.

In our research, the relationships considered the "best" predictors for attitude are effort expectancy, performance expectancy, and social influence, as each has more than 5 identified relationships and a weight greater than 0.800. For behavioral intention, the best predictors are attitude, performance expectancy, habit, self-efficacy, functional congruence, reliability, and benefits. Aesthetic appeal, with a perfect weight of 1, is considered a promising predictor of intention due to the limited number of studies. Social influence, facilitating conditions, and trust, although not classified as the best predictors, remain important because their weights exceed 0.700 and are supported by a substantial number of studies. It is also noteworthy that privacy and security, barriers, vulnerability, severity, compatibility, and financial cost yielded more inconsistent results, with many studies reporting statistically nonsignificant findings.

For actual behavior, behavioral intention is the best predictor, while facilitating conditions and social influence are considered promising predictors due to the limited number of studies and their perfect weight of 1. For the target variable performance expectancy, effort expectancy, TTF, and self-efficacy are the best predictors, and convenience and innovativeness are promising predictors. Health consciousness, social influence, reliability, and compatibility, although not classified as the best predictors, remain important because their weights exceed 0.700 and they are supported by multiple studies. For effort expectancy, facilitating conditions, innovativeness, self-efficacy, and compatibility are the best predictors, while reliability and TTF are promising predictors. For the target variable TTF, technology characteristics is identified as a promising predictor.

Meta-Analysis

The results of the meta-analysis are presented in Table 4 and include all studies that reported standardized path coefficients or β values. All the best predictors identified in our study are statistically significant ($P<.001$), except for reliability ($P=.49$) as a predictor of intention, as well as some of the important and promising predictors. Notably, barriers ($P=.46$) is not a significant predictor of attitude. Health, technology anxiety ($P=.78$), financial cost ($P=.16$), and barriers ($P=.84$) are not significant predictors of behavioral intention. For actual behavior, social influence ($P=.15$), innovativeness ($P=.28$), health consciousness ($P=.61$), vulnerability ($P=.31$), and effort expectancy ($P=.09$) are not significant predictors. Privacy and security ($P=.05$) and barriers ($P=.21$) are not significant predictors of performance expectancy, while privacy and security ($P=.29$) and image ($P=.06$) are not significant predictors of effort expectancy. Finally, task characteristics ($P=.12$) is not a significant predictor of TTF.

XSL•FO
RenderX

Table 4. Meta-analysis results calculated using a random-effects model and presented back-transformed.

Dependent and independent variables	r/ES[a]	95% CI	*Q* statistic	*z* score	*P* value	I^2 statistic (%)
Attitude						
Barriers	0.069	–0.113 to 0.251	107.606	0.743	.46	95.353
Benefits	0.466	0.239 to 0.645	224.814	3.789	<.001	99.11
Effort expectancy	0.286	0.23 to 0.342	153.69	9.549	<.001	86.987
Facilitating conditions	0.527	0.344 to 0.671	874.478	5.067	<.001	99.657
Performance expectancy	0.532	0.414 to 0.633	987.838	7.596	<.001	98.077
Privacy and security	–0.355	–0.618 to –0.093	354.157	–2.653	.008	99.435
Social influence	0.342	0.182 to 0.483	305.084	4.069	<.001	98.033
Behavioral intention						
Health	0.082	–0.049 to 0.21	18.448	1.233	.22	83.738
Aesthetic appeal	0.319	0.266 to 0.371	1.314	11.027	<.001	0
Attitude	0.573	0.454 to 0.672	1855.792	7.853	<.001	98.653
Barriers	–0.016	–0.171 to 0.139	469.927	–0.2	.84	97.446
Benefits	0.309	0.092 to 0.497	688.109	2.757	.006	99.273
Compatibility	0.123	0.081 to 0.165	109.339	5.729	<.001	96.342
Effort expectancy	0.185	0.134 to 0.235	887.804	7.043	<.001	93.58
Ethics	0.303	–0.232 to 0.698	143.458	1.117	.26	98.606
Facilitating conditions	0.198	0.138 to 0.257	274.658	6.318	<.001	90.17
Financial cost	–0.08	–0.191 to 0.031	541.286	–1.414	.16	96.675
Functional congruence	0.212	0.165 to 0.259	39.56	8.509	<.001	89.889
Habit	0.377	0.307 to 0.444	495.521	9.72	<.001	98.789
Health consciousness	0.298	0.222 to 0.371	4455.165	7.332	<.001	99.798
Hedonic motivation	0.202	0.174 to 0.23	135.59	13.785	<.001	89.675
Image	0.201	–0.215 to 0.556	62.391	0.947	.34	93.589
Innovativeness	0.223	0.147 to 0.297	154.334	5.642	<.001	96.112
Performance expectancy	0.339	0.295 to 0.381	1212.594	14.281	<.001	93.732
Privacy and security	–0.11	–0.202 to –0.018	361.354	–2.348	.02	93.912
Reliability	0.148	–0.266 to 0.516	398.688	0.694	.49	98.997
Self-efficacy	0.318	0.257 to 0.377	818.281	9.644	<.001	98.9
Severity	0.12	0.005 to 0.231	8.992	2.04	.04	55.518
Social influence	0.254	0.189 to 0.317	684.855	7.381	<.001	94.305
Technology anxiety	–0.013	–0.1 to 0.075	19.306	–0.279	.78	74.102
Trust	0.294	0.177 to 0.403	666.839	4.769	<.001	98.35
Vulnerability	0.101	0.009 to 0.191	19.657	2.153	.03	64.389
Actual behavior						

XSL•FO
RenderX

Dependent and independent variables	r/ES[a]	95% CI	*Q* statistic	*z* score	*P* value	I^2 statistic (%)
Behavioral intention	0.563	0.427 to 0.674	1139.269	6.912	<.001	98.508
Effort expectancy	0.353	–0.061 to 0.663	99.717	1.683	.09	97.994
Facilitating conditions	0.863	0.706 to 0.939	10353.056	5.987	<.001	99.981
Health consciousness	0.096	–0.267 to 0.436	1594.656	0.511	.61	99.812
Innovativeness	0.232	–0.188 to 0.581	420.955	1.086	.28	99.525
Performance expectancy	0.406	0.001 to 0.696	129.825	1.963	.05	98.459
Social influence	0.31	–0.113 to 0.638	35.827	1.447	.15	94.418
Vulnerability	0.216	–0.204 to 0.569	615.529	1.008	.31	99.675
Performance expectancy						
Effort expectancy	0.38	0.302 to 0.454	587.488	8.873	<.001	95.064
Barriers	–0.137	–0.353 to 0.078	396.849	–1.252	.21	98.992
Compatibility	0.222	0.041 to 0.388	85.162	2.403	.02	92.955
Facilitating conditions	0.328	0.1 to 0.523	239.597	2.784	.005	98.748
Health consciousness	0.188	0.021 to 0.345	140.742	2.206	.03	95.026
Image	0.194	0.001 to 0.373	96.096	1.969	.049	94.797
Innovativeness	0.279	0.049 to 0.481	115.684	2.364	.02	97.407
Privacy and security	–0.193	–0.387 to 0.001	919.956	–1.945	.05	99.456
Reliability	0.309	0.171 to 0.435	82.88	4.258	<.001	87.934
Self-efficacy	0.578	0.431 to 0.695	531.613	6.52	<.001	99.059
Social influence	0.315	0.177 to 0.44	121.563	4.349	<.001	91.774
Trust	0.254	0.046 to 0.441	92.595	2.382	.02	95.68
Task-technology fit	0.678	0.544 to 0.778	278.406	7.539	<.001	98.563
Effort expectancy						
Compatibility	0.318	0.208 to 0.42	35.374	5.441	<.001	85.865
Facilitating conditions	0.436	0.343 to 0.521	24.225	8.321	<.001	79.36
Image	0.147	–0.005 to 0.294	75.386	1.893	.06	97.347
Innovativeness	0.376	0.287 to 0.458	133.654	7.786	<.001	95.511
Privacy and security	–0.074	–0.21 to 0.063	329.763	–1.06	.29	99.09
Reliability	0.46	0.339 to 0.566	45.702	6.747	<.001	93.436
Self-efficacy	0.604	0.529 to 0.67	1371.419	12.34	<.001	99.635
Social influence	0.235	0.097 to 0.364	26.199	3.31	<.001	88.549
Task-technology fit	0.883	0.843 to 0.914	910.714	17.155	<.001	99.78
Task-technology fit						
Task characteristics	0.249	–0.07 to 0.522	460.432	1.537	.12	99.348
Technology characteristics	0.654	0.429 to 0.803	123.033	4.729	<.001	97.562

[a]r/ES: combined effect size (back-transformed from Fisher *z*).

Combining Weight and Meta-Analysis Results: The Most Adopted Models in Research

Figure 3 presents the weight and meta-analysis for the TAM, which explains how users adopt and use technology, emphasizing the influence of external variables on perceived usefulness (performance expectancy) and perceived ease of use (effort expectancy), which in turn affect attitudes, behavioral intentions, and actual technology usage [44,45]. Performance expectancy is the best and statistically significant predictor of both attitude (β=.532, *P*<.001) and behavioral intention (β=.339, *P*<.001). Attitude is the best predictor and has a significant impact on behavioral intention (β=.573, *P*<.001), while behavioral intention is the best and significant predictor of actual behavior (β=.563, *P*<.001). Effort expectancy is the best predictor and strongly influences performance expectancy

XSL•FO
RenderX

(β=.380, *P*<.001) and attitude (β=.286, *P*<.001). Health consciousness (β=.188, *P*=.03), self-efficacy (β=.578, *P*<.001), innovativeness (β=.279, *P*=.02), and compatibility (β=.222, *P*=.02) are significant predictors of performance expectancy, each with a weight above 0.700. Innovativeness (β=.376, *P*<.001) and facilitating conditions (β=.436, *P*<.001) are significant predictors of effort expectancy, also with weights above 0.700.

Figure 3. Weight and meta-analysis for the Technology Acceptance Model. Thicker paths indicate relationships with greater weight—that is, the strongest predictors (weight≥0.700). Higher weights are therefore represented by thicker lines. The numbers on the paths denote the mean β coefficients along with their significance levels.

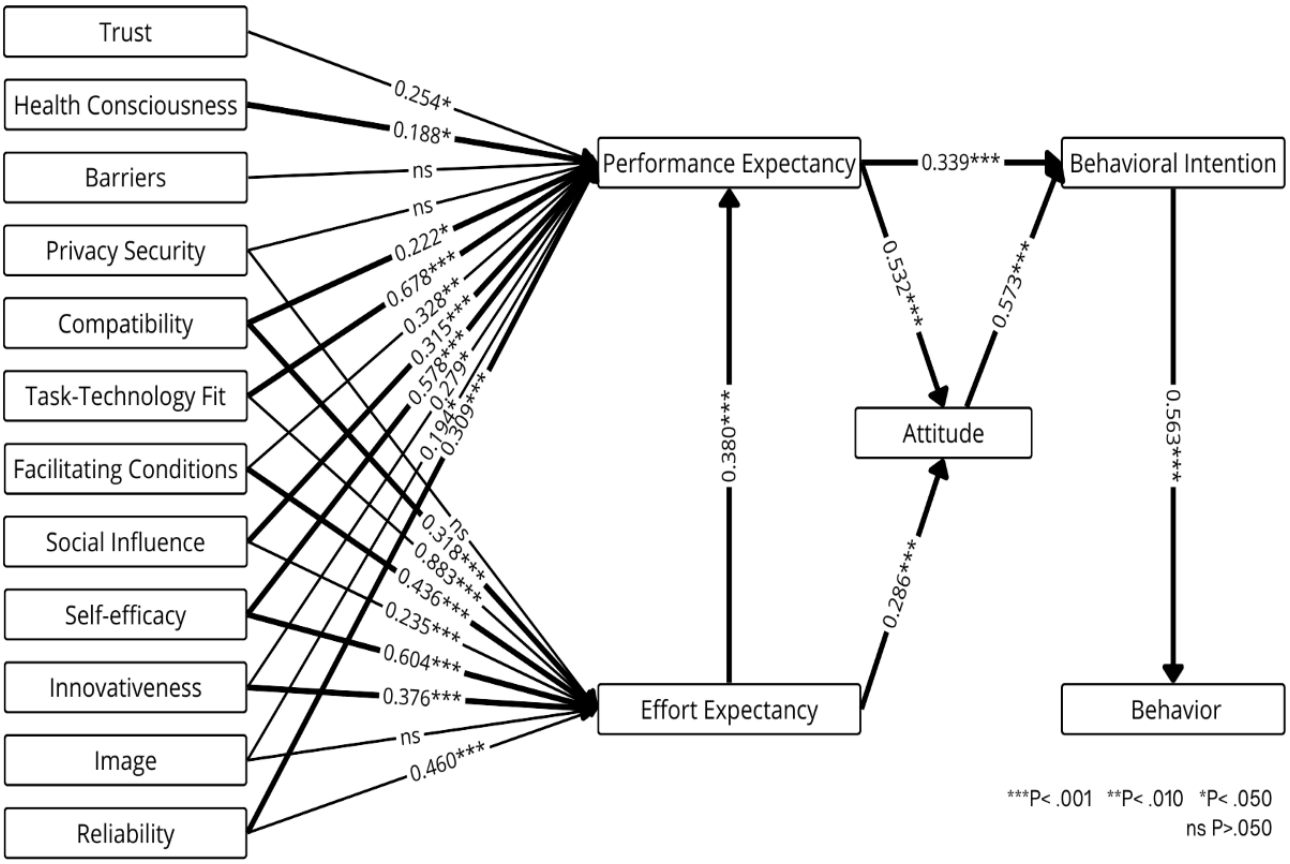

Figure 4 presents the weight and meta-analysis for the Unified Theory of Acceptance and Use of Technology (UTAUT), which explains how users adopt and use technology by assessing the impact of key predictors on behavioral intention and actual behavior [46,47]. Facilitating conditions (β=.863, *P*<.001) and behavioral intention (β=.563, *P*<.001) are significant predictors of actual behavior, while social influence is not. Behavioral intention is significantly influenced by performance expectancy (β=.339, *P*<.001), social influence (β=.254, *P*<.001), facilitating conditions (β=.198, *P*<.001), and habit (β=.377, *P*<.001), all with weights above 0.700. Effort expectancy (β=.185, *P*<.001) and hedonic motivation (β=.202, *P*<.001) are also statistically significant predictors of intention; however, financial cost is not.

XSL•FO
RenderX

Figure 4. Weight and meta-analysis for the Unified Theory of Acceptance and Use of Technology. Thicker paths indicate relationships with greater weight—that is, the strongest predictors (weight≥0.700). Accordingly, higher weights are represented by thicker lines. The numbers on the paths denote the mean β coefficients along with their significance levels.

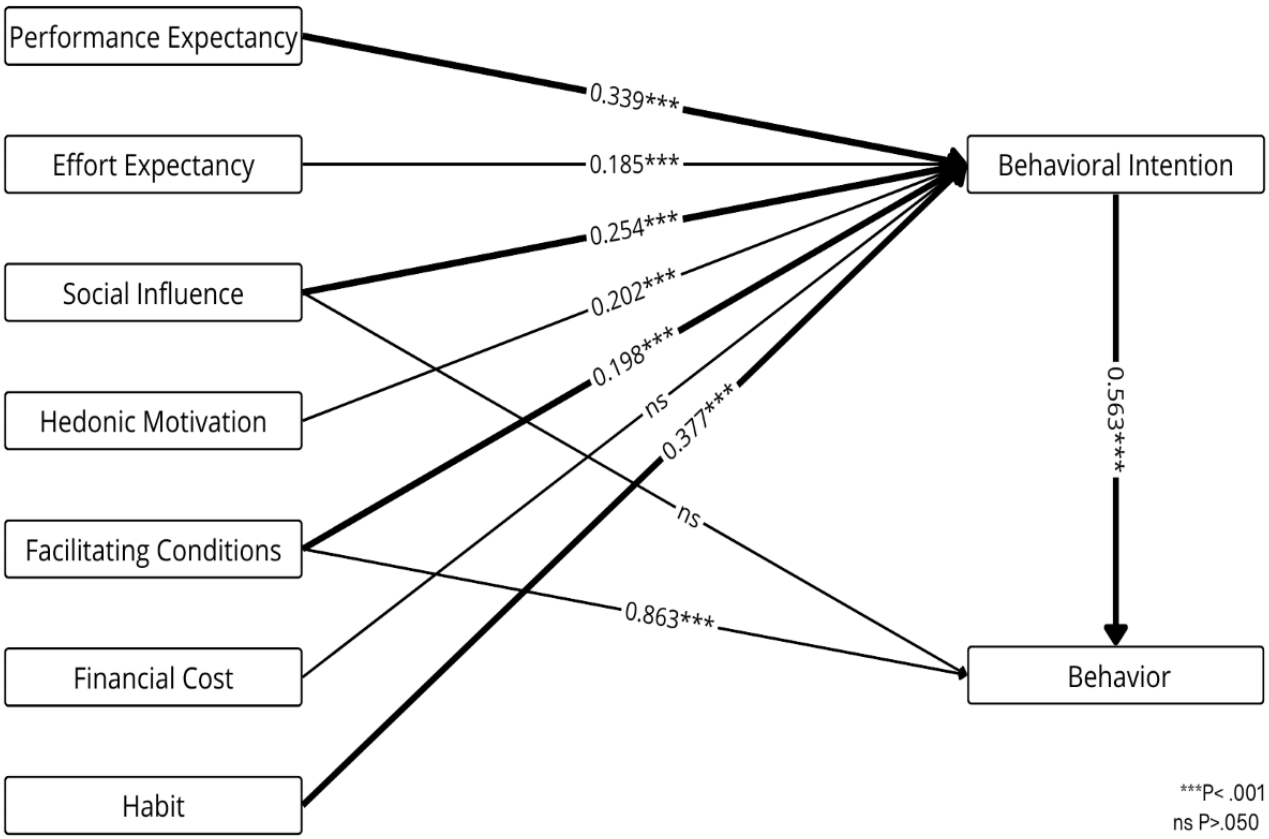

Figure 5 presents the weight and meta-analysis combining the HBM and PMT, which explain individuals' engagement in health-related behaviors. Both models emphasize the role of perceived threat, such as vulnerability and severity, and the evaluation of coping strategies, such as benefits, barriers, response efficacy, and self-efficacy, in shaping motivation to take protective or preventive actions [48-50]. The results indicate that, compared with other technology adoption models, the predictive power of health-related constructs is weaker and less consistent. Severity (β=.120, *P*=.04) and vulnerability (β=.101, *P*=.03) have weak weights on behavioral intention and exert a small but significant impact. Performance expectancy (β=.339, *P*<.001) and self-efficacy (β=.318, *P*<.001), which are also used in other technology-related adoption models, are the best predictors and have a significant impact on behavioral intention. Barriers do not have a significant impact (*P*=.84), whereas benefits exhibit a strong, statistically significant effect (β=.309, *P*=.006). It is also relevant to mention 2 additional predictors directly related to the health context but not part of the key components of these theories. First, health condition, which refers to the perception of having good health, has a weak and nonsignificant impact (*P*=.22) on intention. Second, health consciousness has a statistically significant impact on intention (β=.298, *P*<.001) but does not significantly influence behavior (*P*=.61).

XSL•FO
RenderX

Figure 5. Weight and meta-analysis for the Health Belief Model and Protection Motivation Theory. Thicker paths indicate relationships with greater weight—that is, the strongest predictors (weight≥0.700). Accordingly, higher weights are represented by thicker lines. The numbers on the paths denote the mean β coefficients along with their significance levels.

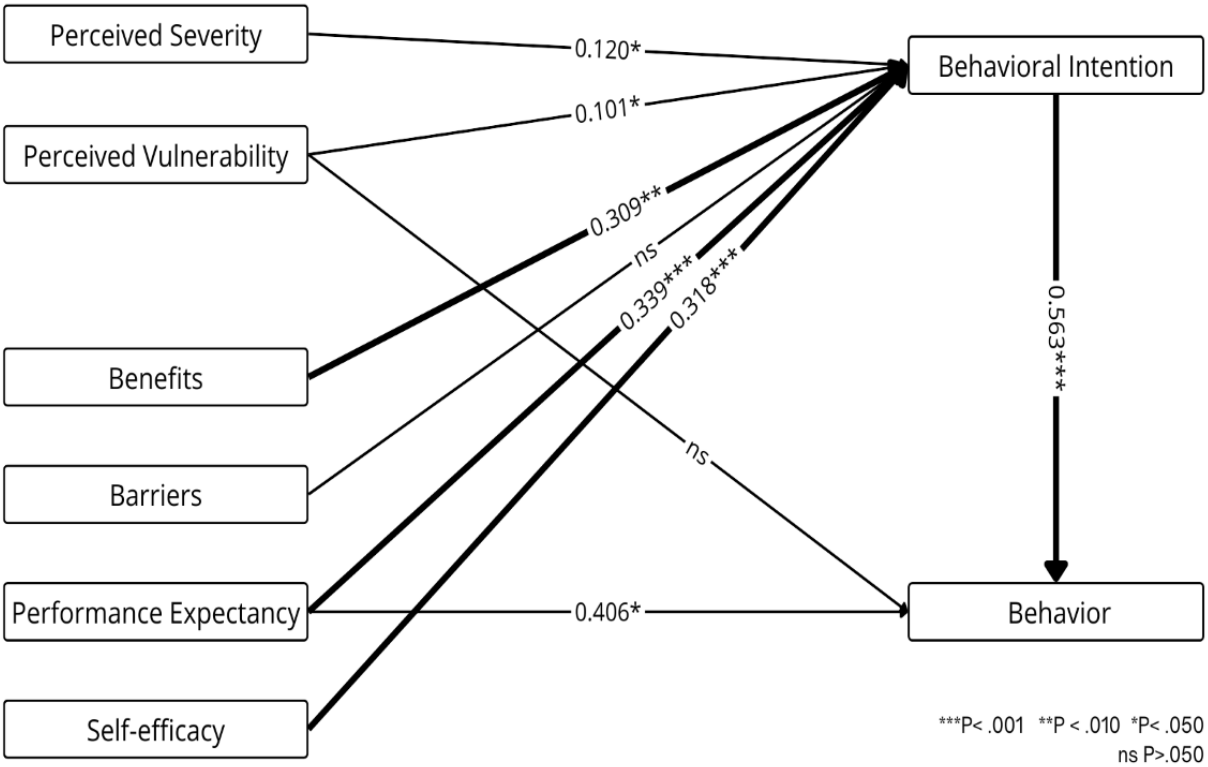

Figure 6 illustrates the TTF model, which examines how well technology aligns with users' tasks to enhance perceived usefulness and adoption [51]. Only part of the theory is presented, as it remains understudied in the context of IoT in health care. The results show that TTF is a significant predictor of performance expectancy (β=.883, *P*<.001) while being classified as a promising predictor. The studies suggest that task characteristics do not have a significant impact on the fit between the task and the technology. However, technology characteristics are a promising and significant predictor (β=.554, *P*<.001).

XSL•FO
RenderX

Figure 6. Weight and meta-analysis for the Task–Technology Fit model. Thicker paths indicate relationships with greater weight—that is, the strongest predictors (weight≥0.700). Accordingly, higher weights are represented by thicker lines. The numbers on the paths denote the mean β coefficients along with their significance levels. ns: not significant.

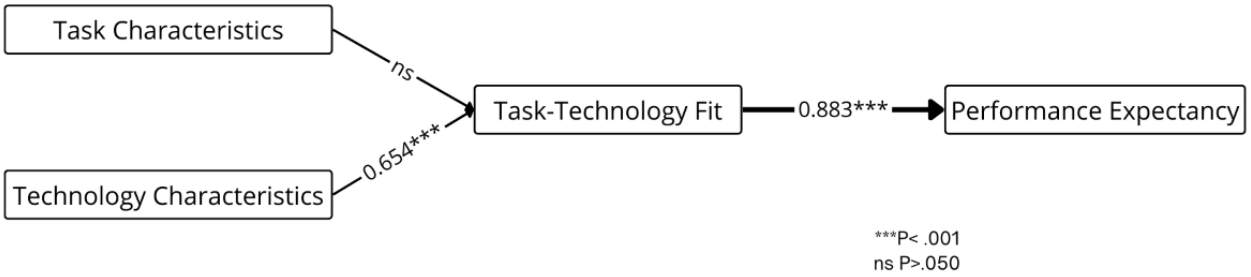

Figure 7 presents the weight and meta-analysis of the Privacy Calculus Theory, which explores the trade-off between benefits and privacy [52]. The results indicate that trust (β=.294, *P*<.001) has a significant positive influence on behavioral intention and is classified as the best predictor. Privacy and security (β=–.110, *P*=.02) exhibits a significant negative effect on behavioral intention; however, the weight is small, suggesting that privacy and security concerns may not be a strong inhibitor of adoption.

XSL•FO
RenderX

Figure 7. Weight and meta-analysis for Privacy Calculus Theory. Thicker paths indicate relationships with greater weight—that is, the strongest predictors (weight≥0.700). Accordingly, higher weights are represented by thicker lines. The numbers on the paths denote the mean β coefficients along with their significance levels.

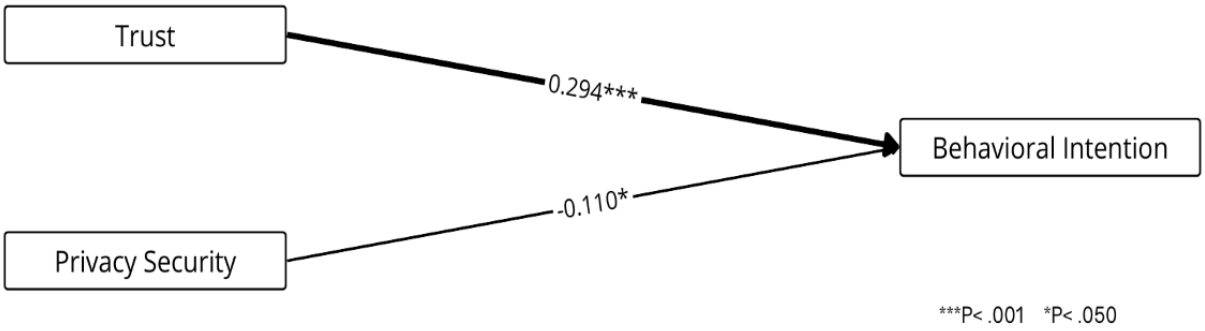

Figure 8 presents the weight and meta-analysis of the Theory of Planned Behavior, which posits that attitude, subjective norms, and behavioral control influence behavioral intention and, subsequently, behavior [53]. Both attitude (β=.573, *P*<.001) and self-efficacy (β=.318, *P*<.001) are the best and strongest predictors of behavioral intention.

XSL•FO
RenderX

Figure 8. Weight and meta-analysis for the Theory of Planned Behavior. Thicker paths indicate relationships with greater weight—that is, the strongest predictors (weight≥0.700). Accordingly, higher weights are represented by thicker lines. The numbers on the paths denote the mean β coefficients along with their significance levels.

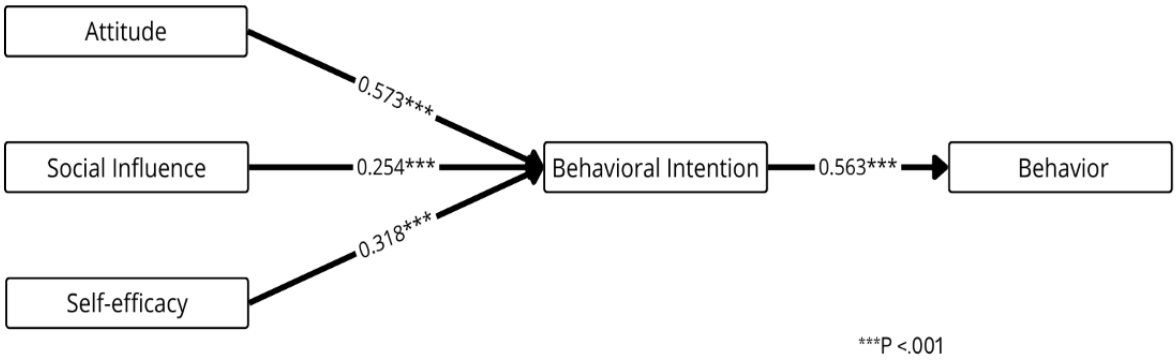

Figure 9 presents the weight and meta-analysis of the Innovation Diffusion Theory, which explains the diffusion of new technologies through 5 dimensions [54]. Relative advantage (performance expectancy; β=.339, *P*<.001), complexity (effort expectancy; β=.185, *P*<.001), and compatibility (β=.123, *P*<.001) were found to be significant predictors. Image was not a significant predictor, and trialability and observability have not been sufficiently studied in the literature.

XSL•FO
RenderX

Figure 9. Weight and meta-analysis for the Diffusion of Innovation theory. Thicker paths indicate relationships with greater weight—that is, the strongest predictors (weight≥0.700). Accordingly, higher weights are represented by thicker lines. The numbers on the paths denote the mean β coefficients along with their significance levels.

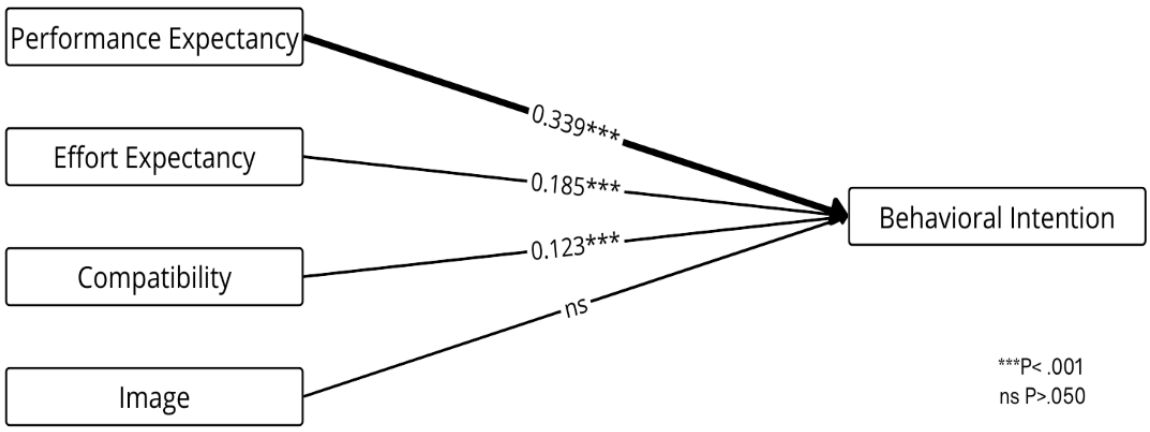

Evaluation of Publication Bias

This section evaluates the presence of publication bias and assesses the normality of the datasets used in the meta-analysis to ensure the reliability of the synthesized findings. Publication bias refers to the tendency for studies with significant or positive results to be more likely to be published, potentially skewing meta-analytic outcomes [55]. To ensure the robustness of our findings, we evaluated publication bias following the approach of Harrison et al [55], which suggests that a single criterion can provide a more sensitive and appropriate test. We focused our analysis on one of the most widely examined relationships in our dataset: the relationship between performance expectancy and behavioral intention, which was reported in 77 studies (Table 5).

XSL•FO
RenderX

Table 5. Studies (n=77) showing the effect size (z), SE (z), sample size, z score, Q component, significance of the paths between performance expectancy and behavioral intention, and the country.

Subgroup and effect size (*Z*)	SE (*Z*)	Sample size	*z* score	*Q* component	Significance	Country
Group 1						
0.268	0.051	387	5.965	2.350	Significant	China
0.604	0.050	397	13.455	26.161	Significant	China
0.387	0.080	158	7.933	0.258	Significant	China
0.165	0.046	469	3.718	15.264	Significant	China
0.086	0.053	357	1.908	23.979	Significant	China
0.230	0.051	386	5.113	5.199	Significant	China
0.727	0.065	243	15.616	34.684	Significant	China
0.224	0.036	769	5.122	11.555	Significant	China
0.186	0.058	304	4.074	7.740	Significant	China
0.149	0.113	81	2.702	3.039	Nonsignificant	China
0.333	0.071	201	7.018	0.037	Significant	China
0.472	0.050	406	10.530	6.373	Significant	China
0.205	0.056	325	4.504	6.462	Significant	China
0.198	0.054	345	4.362	7.588	Significant	China
0.217	0.086	139	4.372	2.267	Significant	China
0.950	0.084	146	19.257	52.169	Significant	China
0.406	0.065	237	8.704	0.827	Significant	China
0.471	0.072	197	9.914	3.014	Significant	China
0.519	0.065	239	11.139	7.032	Significant	China
0.180	0.047	462	4.039	12.734	Significant	China
0.105	0.071	201	2.223	11.509	Nonsignificant	China
0.289	0.040	624	6.567	2.068	Significant	China
0.286	0.064	247	6.145	0.907	Significant	China
0.377	0.039	668	8.591	0.615	Significant	China
0.446	0.051	386	9.927	3.830	Significant	China
0.157	0.043	552	3.560	19.651	Nonsignificant	China
0.258	0.047	450	5.774	3.535	Significant	China
0.401	0.096	111	7.772	0.324	Significant	China
1.040	0.077	171	21.531	80.869	Significant	China
0.586	0.028	1292	13.583	73.940	Significant	China
0.224	0.046	475	5.028	7.120	Significant	China
0.236	0.037	725	5.401	8.764	Significant	China
Group 2						

XSL•FO
RenderX

Subgroup and effect size (*Z*)	SE (*Z*)	Sample size	*z* score	*Q* component	Significance	Country
0.422	0.033	913	9.720	5.253	Significant	Bangladesh
0.586	0.075	181	12.212	10.210	Significant	France
0.132	0.062	267	2.855	12.172	Nonsignificant	France
0.202	0.125	67	3.478	1.342	Nonsignificant	France
0.633	0.070	206	13.382	16.646	Significant	Germany, the United States, the United Kingdom, and Canada
0.102	0.056	320	2.249	18.892	Nonsignificant	Ghana
0.319	0.039	645	7.273	0.469	Significant	Ghana
0.401	0.050	400	8.939	1.190	Significant	India
0.421	0.086	139	8.473	0.761	Significant	India
0.283	0.043	534	6.404	2.116	Significant	India
0.150	0.052	372	3.330	14.228	Significant	India
0.119	0.082	153	2.418	7.793	Nonsignificant	India
0.306	0.065	238	6.569	0.381	Significant	India
0.415	0.036	772	9.512	3.647	Significant	Indonesia
0.549	0.054	341	12.120	13.905	Significant	Iraq
0.245	0.069	212	5.191	2.162	Significant	Italy
0.413	0.066	233	8.841	1.017	Significant	Japan
0.265	0.071	200	5.587	1.307	Significant	Jordan
0.205	0.051	389	4.556	7.746	Significant	Korea
0.321	0.080	159	6.572	0.105	Significant	Korea
0.084	0.029	1158	1.948	79.454	Nonsignificant	Korea
0.239	0.060	280	5.209	3.172	Significant	Nepal
0.190	0.063	259	4.112	6.247	Significant	Oman
0.321	0.045	495	7.220	0.331	Significant	Pakistan
0.284	0.046	486	6.401	1.860	Significant	The Kingdom of Saudi Arabia
0.229	0.046	473	5.145	6.496	Significant	The Kingdom of Saudi Arabia
0.189	0.063	256	4.085	6.256	Significant	The Kingdom of Saudi Arabia
0.553	0.046	477	12.441	20.278	Significant	South Korea
0.574	0.045	487	12.910	24.967	Significant	South Korea
0.523	0.034	877	12.020	27.229	Significant	South Korea
0.518	0.097	110	10.014	3.141	Significant	Switzerland
0.485	0.055	335	10.682	6.343	Significant	Taiwan
0.393	0.061	268	8.519	0.575	Significant	Taiwan
0.149	0.047	458	3.346	17.727	Significant	Taiwan
0.321	0.091	125	6.340	0.082	Significant	Taiwan
0.245	0.113	81	4.436	0.807	Significant	Taiwan
0.688	0.065	243	14.796	28.070	Significant	Turkey
0.041	0.049	426	0.917	39.467	Nonsignificant	Turkey
0.230	0.099	106	4.416	1.398	Significant	The United Arab Emirates

XSL•FO
RenderX

Subgroup and effect size (*Z*)	SE (*Z*)	Sample size	*z* score	*Q* component	Significance	Country
0.141	0.050	407	3.143	17.070	Nonsignificant	The United States
1.157	0.052	376	25.681	244.930	Significant	The United States
0.167	0.060	277	3.619	8.873	Significant	The United States
0.266	0.092	120	5.227	0.756	Significant	The United States
0.848	0.047	450	19.012	112.411	Significant	The United States
0.321	0.056	322	7.045	0.215	Significant	Worldwide

To assess the presence of small-study effects and potential publication bias, a funnel plot was generated, and an Egger regression test was conducted. The funnel plot was constructed to visually evaluate publication bias [42], with the SE plotted on the y-axis instead of sample size, as this enhances the detection of asymmetry [36]. In an ideal funnel plot, symmetry is expected, with smaller studies exhibiting larger SE scattered evenly on both sides of the pooled effect size. In Figure 10, the studies display a somewhat asymmetrical distribution. Larger studies cluster near the combined effect size at the top of the funnel, while smaller studies show greater dispersion, potentially indicating publication bias or underlying heterogeneity. The trim-and-fill method estimates the number of potentially missing studies—often those with nonsignificant or negative results—and imputes them to generate an adjusted combined effect size. In this case, the imputed effect size is slightly smaller than the original estimate, suggesting that the observed meta-analytic effect may be modestly inflated due to the absence of smaller, less favorable studies.

Figure 10. Funnel plot of studies examining the relationship between performance expectancy and behavioral intention.

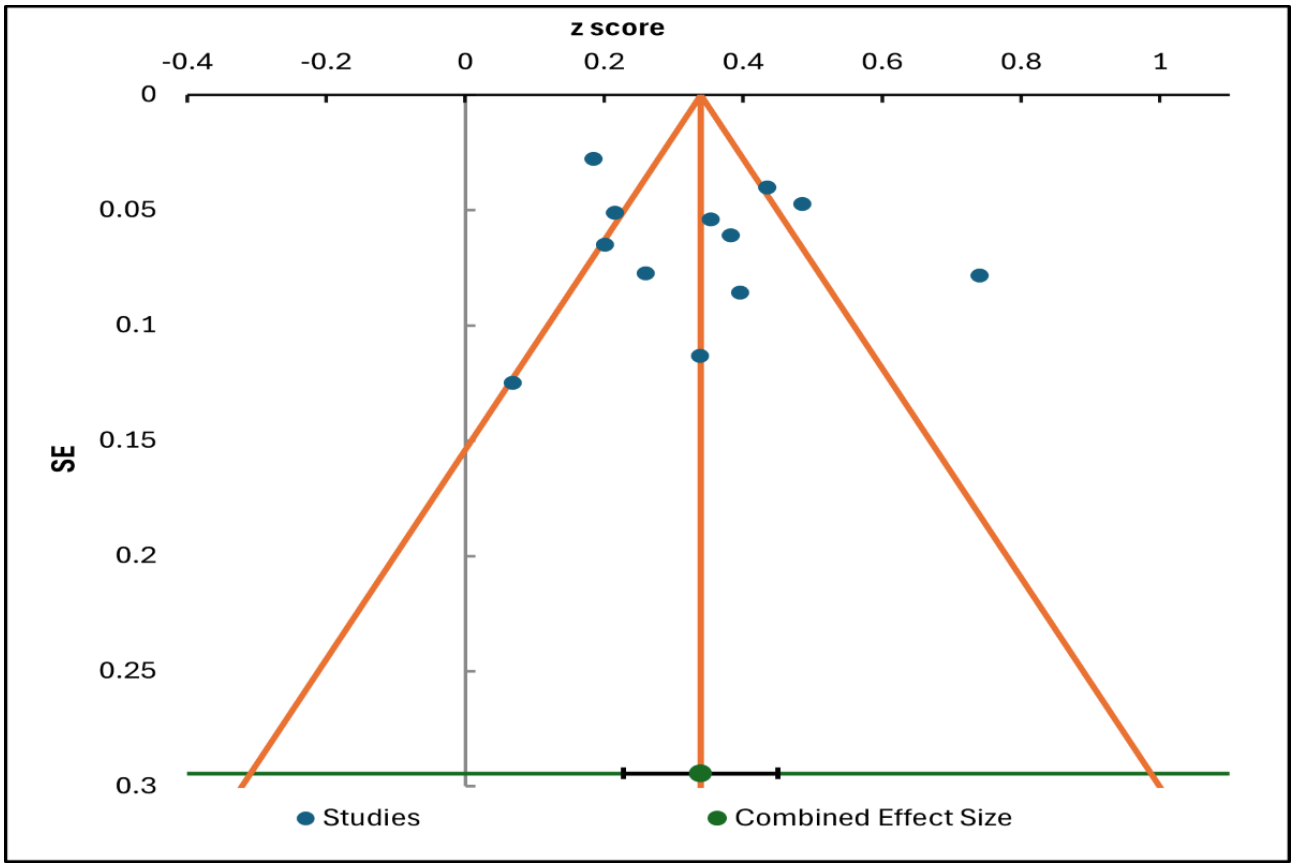

To statistically assess funnel plot asymmetry, we applied the Egger regression test [42] to evaluate whether smaller studies tend to report larger effect sizes, which can indicate potential publication bias (see Table 6). The test examines the relationship between effect sizes and their SEs to detect small-study effects. The results of the regression analysis showed that the intercept was not statistically significant (α=.381, *P*=.81), indicating no evidence of funnel plot asymmetry or publication bias. However, the slope coefficient was statistically significant (β=.327, *P*<.001), suggesting a positive association between study precision and effect size. While this does not indicate publication bias, it may reflect genuine heterogeneity among the included studies.

The I^2 statistic, which quantifies the proportion of total variation across studies attributable to true heterogeneity rather than sampling error [41], revealed a very high degree of heterogeneity (I^2>93%). This indicates that most of the variability in effect sizes reflects real differences across studies rather than random

XSL•FO
RenderX

sampling error. These differences may arise from variations in study design, measurement tools, participant demographics, cultural contexts, or theoretical frameworks used across the included studies. To further explore the sources of heterogeneity, a subgroup analysis was conducted.

Table 6. Egger - type test for small - study bias, using Excel's Data Analysis Regression Tool, which uses standard ordinary least squares regression.

Regression	Coefficients	SE	*t* test (*df*)	*P* value	95% CI
Intercept (α)	0.381	1.544	0.247 (75)	.81	–2.695 to 3.457
Slope (β)	0.327	0.081	4.044 (75)	<.001	0.166 to 0.489

The subgroup analysis examined whether effect sizes differed between studies conducted in China and those conducted in other countries. By comparing studies from China with those from other regions, we aimed to evaluate potential regional biases, given that a large proportion of the included studies were conducted in China. The results, presented in Table 7, indicate that geographic location has little influence on the overall effect size, as both subgroups exhibit similar results. The combined effect size for studies conducted in China is 0.340 (95% CI 0.272-0.404), while for studies in other countries, it is 0.336 (95% CI 0.279-0.390). However, heterogeneity remained very high in both groups (China: I^2=93%; other countries: I^2=94%), indicating substantial variability even within each subgroup. Therefore, the subgroup analysis addresses concerns about potential bias from the large proportion of studies conducted in China, confirming that the results are largely stable across regions.

Table 7. Subgroup comparison between China and the other countries in our sample.

Subgroup	Effect size	*P* value	95% CI	I^2 (%)
China	0.340	<.001	0.272-0.404	92.984
Other countries	0.336	<.001	0.279-0.390	94.355

Discussion

Principal Findings

The study of IoT adoption in health care reveals a diverse landscape of constructs and relationships, providing a comprehensive overview of the factors driving IoT adoption. This study synthesized findings from 109 papers and 115 datasets across various regions, including China, South Korea, the United States, and India, with most studies published in high-ranking journals. The combined weight and meta-analysis identified the best predictors and examined the adoption models most frequently used in IoT health care. Figure 11 highlights the strongest and most consistent predictors, integrating the results of both the meta-analysis and weight analysis.

XSL•FO
RenderX

Figure 11. Best predictors identified in the weight analysis, along with their statistically significant mean β coefficients from the meta-analysis. The strength (weight) of each predictor is represented by the thickness of the line connecting the predictor to the target variable, with thicker lines indicating stronger predictors. The green box denotes constructs from the Unified Theory of Acceptance and Use of Technology framework, and the orange box denotes constructs from Technology Acceptance Model.

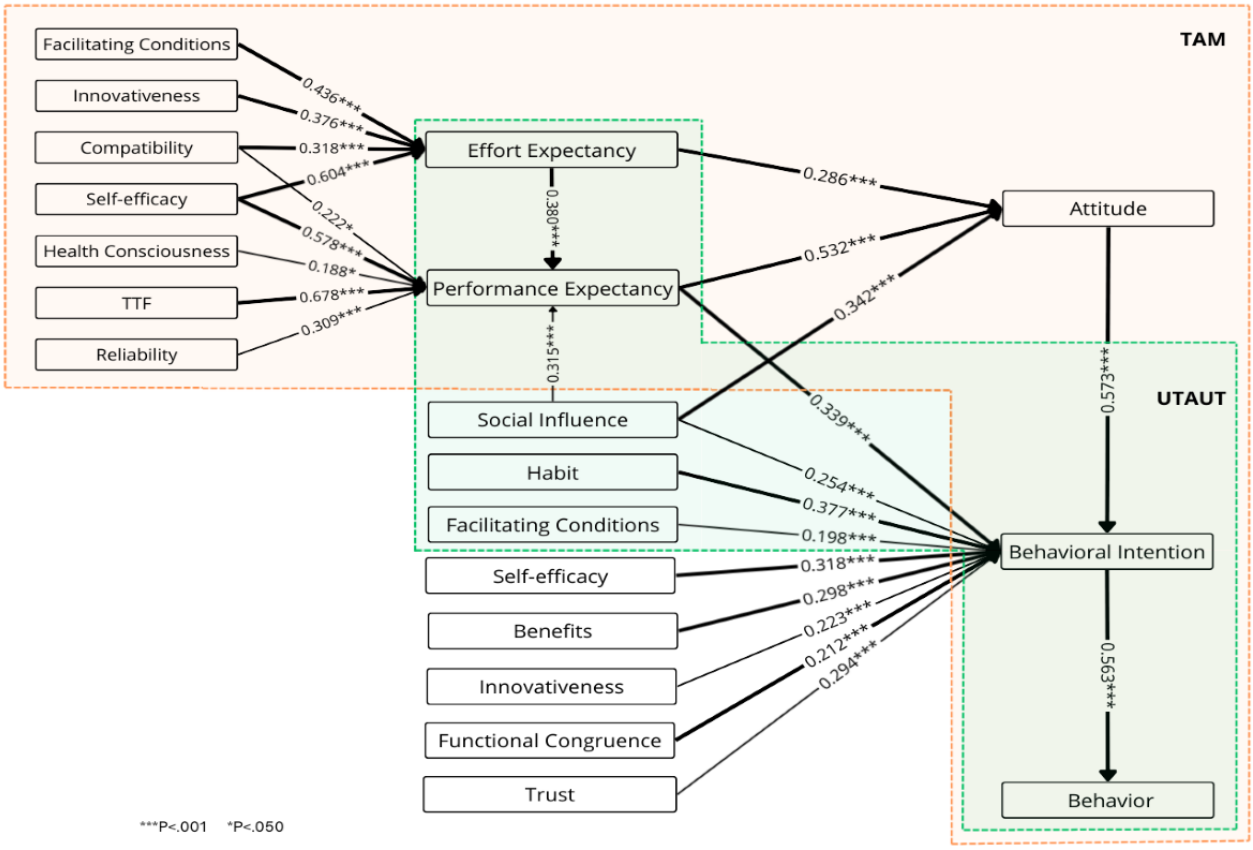

Technology acceptance models, such as UTAUT and TAM, have been widely and successfully applied in the context of IoT in health care [56,57], which is unsurprising given that these are the most commonly used technology adoption models, highly cited, and successfully applied across diverse fields and contexts [58,59]. Compared with models such as UTAUT and TAM, where predictors such as performance expectancy and facilitating conditions consistently exhibit strong and reliable effects, the HBM and PMT models struggle to establish robust relationships with behavioral intention. This suggests that relying solely on health-related constructs or health behavior models may not be sufficient to explain health care technology adoption. Therefore, other individual factors, such as innovativeness, external factors, such as social influence, and technological factors, such as performance expectancy, may play a more decisive role [60,61]. Integrating context-specific health variables into robust models such as TAM or UTAUT can, however, provide additional insights. For example, individuals with strong health motivation or health consciousness tend to exhibit higher levels of performance expectancy from IoT health care technologies [62,63].

The findings highlight several key factors influencing effort expectancy and performance expectancy, both of which are central to users' attitudes toward technology. For effort expectancy, the most influential factor was self-efficacy, indicating that individuals who feel more confident in their ability to use the technology tend to perceive it as easier to operate [64,65]. Other important contributors include facilitating conditions, innovativeness, and compatibility, suggesting that a supportive environment, openness to new technologies, and alignment with users' existing values and practices all help reduce the perceived effort required to use IoT in health care [66,67]. For performance expectancy, TTF emerged as the dominant influence, highlighting that when users perceive a strong alignment between the technology and the tasks they need to perform, they are more likely to view it as useful [68,69]. Additionally, health consciousness, self-efficacy, reliability, and compatibility played significant roles, emphasizing the importance of personal health concerns, confidence in usage, trust in the system's dependability, and alignment with users' existing values and practices [63,70]. Together, these findings underscore the relevance of both individual and contextual factors in shaping users' perceptions of a technology's usefulness and ease of use.

Regarding individuals' IoT health care technology adoption journey, the findings reveal that a positive attitude is crucial for successful adoption [71,72]. Efforts to cultivate positive perceptions can be made by leveraging the influence of important figures in individuals' lives and by emphasizing the ease of use and the potential for improved health care outcomes [73-75]. When individuals hold a positive perception of IoT health care, they are more likely to intend to use it, which in turn positively influences actual usage [76,77]. To further enhance behavioral intention, the effectiveness of IoT health

XSL•FO
RenderX

care solutions and the encouragement of health care professionals, family, and friends should be leveraged [78,79]. Additionally, individuals' willingness to try new technologies plays a significant role, as more innovative users are more likely to adopt IoT solutions [80,81].

Trust plays a decisive role in shaping behavioral intentions, reinforcing the notion that users are willing to trade some level of privacy if they perceive a system as secure and reliable [60]. Previous literature has found that individuals are often reluctant to adopt digital health or IoT technologies when they do not trust the provider [82,83]. This perspective may help contextualize the inconsistent results observed for privacy as a predictor of behavioral intention, as a notable proportion of studies reported nonsignificant relationships. Similarly, predictors such as barriers, vulnerability, and financial cost also exhibited higher frequencies of nonsignificant findings in our analysis. These inconsistencies may reflect how these constructs interact with—or are influenced by—the presence of stronger enabling factors. For instance, high perceived usefulness and trust may diminish the observed effects of barriers such as financial cost and privacy, as these factors may become less salient in users' perceptions.

Theoretical Implications

This study makes several contributions to the theoretical understanding of IoT health care adoption by synthesizing findings from diverse quantitative studies and adoption models. The results reinforce the importance of established models such as TAM and UTAUT. They also suggest that integrating variables from other theories—such as health consciousness, innovativeness, and trust—into traditional technology acceptance frameworks can provide deeper insights into how individuals adopt IoT in health care. Behavioral intention is the most studied target variable, while attitude and actual behavior remain underexplored, indicating a gap in existing research on these critical components of the adoption process.

Researchers should further investigate several promising but underexplored predictors that showed perfect weight, suggesting strong yet preliminary evidence of their relevance, to establish their broader applicability. For instance, regarding behavioral intention, the aesthetic appeal of health care technologies shows potential as a strong predictor. For actual behavior, facilitating conditions and social influence are promising predictors that warrant further exploration. In the case of the underexplored TTF theory, technology characteristics appear to be a promising predictor. For performance expectancy, convenience and innovativeness are promising predictors, while for effort expectancy, reliability and TTF show potential as predictors deserving additional investigation.

By contrast, several predictors demonstrated limited or inconsistent relevance to the adoption of IoT in health care. For the outcome attitude, barriers did not have a statistically significant effect. For behavioral intention, predictors such as privacy and security, barriers, vulnerability, severity, compatibility, and financial cost produced inconsistent findings, with many studies reporting nonsignificant results. Specifically, health, technology anxiety, financial cost, and barriers were frequently not significant predictors of behavioral intention. When predicting actual behavior, variables such as social influence, innovativeness, health consciousness, vulnerability, and effort expectancy often failed to reach statistical significance. Regarding performance expectancy, both privacy and security and barriers were not consistently significant, and for effort expectancy, privacy and security and image did not show meaningful effects.

Our findings indicate that regional differences alone do not fully explain the heterogeneity of results. Therefore, when applying the findings of this study, we recommend refining theoretical models to account for contextual factors and implementing practical strategies aligned with the strongest predictors identified, such as performance expectancy and self-efficacy, which can enhance adoption across different settings. Future adoption studies would benefit from incorporating context-specific factors that capture cultural and health care system differences, enabling a better understanding of how these contextual variables influence target outcomes, as either control or moderator variables.

Several regions, particularly in Africa, South America, and Europe, remain underrepresented, highlighting a gap in the literature and the need for future research in diverse settings to improve the generalizability and equity of evidence regarding IoT adoption in health care. Finally, combining qualitative methods, such as interviews and focus groups, is recommended to gain deeper insights. This mixed methods approach can provide a better understanding of user perceptions and experiences, bridging the gap between quantitative results and the complex realities of technology adoption [84,85].

Practical Implications

The findings of this study provide actionable insights for practitioners, policy makers, and technology developers seeking to enhance IoT health care adoption. Key drivers—such as performance expectancy, self-efficacy, social influence, functional congruence, trust, habit, facilitating conditions, benefits, and innovativeness—consistently shape behavioral intention. For example, developers can focus on creating intuitive designs and user-friendly interfaces while emphasizing tangible performance benefits. Health care providers and policy makers can leverage trusted individuals, such as doctors and family members, to encourage adoption.

The availability of resources and infrastructure that enable and support technology use—such as access to devices, technical support, internet connectivity, and integration with health care systems—plays an important role in adoption, as it reduces barriers for individuals starting to use IoT in health care [69,78]. Furthermore, adhering to robust data protection frameworks that ensure transparency from all entities handling health-related data aligns implementation with national and regional regulatory standards and fosters user trust [83,86]. Finally, targeting innovative individuals who are more likely to adopt IoT health care technologies or who already have the habits and skills to use them can further promote technology adoption.

Limitations and Future Research

This study has several limitations that warrant consideration. Our findings reveal a high level of heterogeneity, which is not

XSL•FO
RenderX

fully explained by regional differences; therefore, the pooled estimates should be interpreted with caution. Future research should investigate additional factors that may account for this heterogeneity, such as study design, population characteristics, or model specification, and conduct moderator analyses to better address variability. Additionally, China accounts for a large proportion of the included studies, while several regions—particularly in Africa, South America, and Europe—remain underrepresented. As such, we caution against overgeneralizing our findings to all global contexts. Additionally, while this study synthesizes quantitative findings, it excludes qualitative research, which could provide deeper insights into contextual variability and user experiences influencing adoption. This exclusion may contribute to inconsistencies in the evidence, particularly for understudied predictors such as privacy concerns, perceived vulnerability, and financial cost, which often showed nonsignificant results. Future research should consider integrative literature reviews that include qualitative studies to better capture the nuanced interplay of individual, cultural, and technological factors.

Conclusions

Our comprehensive meta- and weight analysis of 115 unique datasets on IoT health care adoption revealed several significant predictors for the adoption of IoT health care technologies. Behavioral intention emerged as the most frequently studied target variable. By contrast, attitude, actual behavior, performance expectancy, effort expectancy, and TTF remain comparatively understudied, with very few paths examined more than 5 times. While adoption theories from the information systems field, such as UTAUT and TAM, are predominantly used, integrating context-specific factors or combining constructs from different theoretical models can provide deeper insights into IoT health care adoption and further support the adoption process.

All the best predictors identified in our study were statistically significant, with the exception of reliability as a predictor of behavioral intention. For the target variable attitude, the strongest predictors were effort expectancy, performance expectancy, and social influence, while barriers did not have a statistically significant effect. Regarding behavioral intention, the most consistent and significant predictors were attitude, performance expectancy, habit, self-efficacy, functional congruence, reliability, and benefits. In addition, social influence, facilitating conditions, and trust demonstrated strong weights above 0.700, while aesthetic appeal was considered a promising predictor due to the limited number of studies. Conversely, variables such as privacy and security, barriers, vulnerability, severity, compatibility, and financial cost showed inconsistent results, with a high incidence of statistically nonsignificant findings. Specifically, health, technology anxiety, financial cost, and barriers were not statistically significant predictors of behavioral intention.

For actual behavior, behavioral intention emerged as the best predictor, while facilitating conditions and social influence were considered promising. However, social influence, innovativeness, health consciousness, vulnerability, and effort expectancy did not reach statistical significance for behavior. Regarding performance expectancy, effort expectancy, TTF, and self-efficacy were the best predictors, followed by health consciousness, social influence, reliability, and compatibility as strong predictors, while convenience and innovativeness appeared as promising. Privacy and security and barriers, however, were not statistically significant predictors of performance expectancy. For effort expectancy, the most consistent predictors were facilitating conditions, innovativeness, self-efficacy, and compatibility, with reliability and TTF considered promising predictors; privacy and security and image did not show significant effects. Lastly, for the target variable TTF, technology characteristics emerged as a promising predictor, whereas task characteristics were not statistically significant.

Acknowledgments

We thank the European Union’s Horizon Europe program and Fundação para a Ciência e a Tecnologia (FCT) for their support through the program UIDB/04152 – Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS.

Funding

This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia) under the project UIDB/04152 – Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS. It also resulted from the TwinAIR project, which received funding from the European Union’s Horizon Europe program under grant agreement No. 101057779.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author (IV) on reasonable request.

Authors' Contributions

Conceptualization: IV, MNZ
Data curation: IV, MNZ
Formal analysis: IV
Funding acquisition: SK

XSL•FO
RenderX

Investigation: IV
Methodology: IV
Project administration: SK
Supervision: MNZ, RM, TO
Validation: MNZ, RM
Writing – original draft: IV
Writing – review & editing: IV, MNZ, RM, TO

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 177 KB - jmir_v28i1e64091_app1.docx]

References

1. Krishnamoorthy S, Dua A, Gupta S. Role of emerging technologies in future IoT-driven Healthcare 4.0 technologies: a survey, current challenges and future directions. J Ambient Intell Human Comput 2021 May 12;14(1):361-407. [doi: 10.1007/s12652-021-03302-w]
2. Mohamad Jawad HH, Bin Hassan Z, Zaidan BB, Mohammed Jawad FH, Mohamed Jawad DH, Alredany WHD. A systematic literature review of enabling IoT in healthcare: motivations, challenges, and recommendations. Electronics 2022 Oct 08;11(19):3223. [doi: 10.3390/electronics11193223]
3. Ashraf S. A proactive role of IoT devices in building smart cities. Internet of Things and Cyber-Physical Systems 2021;1:8-13. [doi: 10.1016/j.iotcps.2021.08.001]
4. Lin Q, Zhao Q. IoT applications in healthcare. In: García Márquez FP, Lev B, editors. Internet of Things: Cases and Studies. Cham, Switzerland: Springer; Jul 14, 2021:115-133.
5. Singh RP, Javaid M, Haleem A, Suman R. Internet of things (IoT) applications to fight against COVID-19 pandemic. Diabetes Metab Syndr 2020;14(4):521-524 [FREE Full text] [doi: 10.1016/j.dsx.2020.04.041] [Medline: 32388333]
6. Tuli S, Basumatary N, Gill SS, Kahani M, Arya RC, Wander GS, et al. HealthFog: an ensemble deep learning based smart healthcare system for automatic diagnosis of heart diseases in integrated IoT and fog computing environments. Future Generation Computer Systems 2020 Mar;104:187-200. [doi: 10.1016/j.future.2019.10.043]
7. Farahani B, Firouzi F, Chang V, Badaroglu M, Constant N, Mankodiya K. Towards fog-driven IoT eHealth: promises and challenges of IoT in medicine and healthcare. Future Generation Computer Systems 2018 Jan;78:659-676. [doi: 10.1016/j.future.2017.04.036]
8. Li X, Dai H, Wang Q, Imran M, Li D, Imran MA. Securing Internet of Medical Things with friendly-jamming schemes. Comput Commun 2020 Jul 01;160:431-442 [FREE Full text] [doi: 10.1016/j.comcom.2020.06.026] [Medline: 32834198]
9. Ye J. The role of health technology and informatics in a global public health emergency: practices and implications from the COVID-19 pandemic. JMIR Med Inform 2020 Jul 14;8(7):e19866 [FREE Full text] [doi: 10.2196/19866] [Medline: 32568725]
10. Philip NY, Rodrigues JJPC, Wang H, Fong SJ, Chen J. Internet of Things for in-home health monitoring systems: current advances, challenges and future directions. IEEE J Select Areas Commun 2021 Feb;39(2):300-310. [doi: 10.1109/jsac.2020.3042421]
11. Osama M, Ateya AA, Sayed MS, Hammad M, Pławiak P, Abd El-Latif AA, et al. Internet of Medical Things and Healthcare 4.0: trends, requirements, challenges, and research directions. Sensors (Basel) 2023 Aug 25;23(17):7435 [FREE Full text] [doi: 10.3390/s23177435] [Medline: 37687891]
12. Vyas S, Bhargava D, Khan S, Bhargava D. Healthcare 4.0: a systematic review and its impact over conventional healthcare system. In: Artificial Intelligence for Health 4.0: Challenges and Applications. London, UK: Routledge; Jan 2023:1-17.
13. Healthcare IoT - worldwide. Statista. 2024. URL: https://www.statista.com/outlook/tmo/internet-of-things/healthcare-iot/worldwide [accessed 2024-06-27]
14. Kelly JT, Campbell KL, Gong E, Scuffham P. The Internet of Things: impact and implications for health care delivery. J Med Internet Res 2020 Nov 10;22(11):e20135 [FREE Full text] [doi: 10.2196/20135] [Medline: 33170132]
15. Yang F, Shu H, Zhang X. Understanding "internet plus healthcare" in China: policy text analysis. J Med Internet Res 2021 Jul 26;23(7):e23779 [FREE Full text] [doi: 10.2196/23779] [Medline: 34309581]
16. Liang J, Li Y, Zhang Z, Shen D, Xu J, Yu G, et al. Evaluating the applications of health information technologies in China during the past 11 years: consecutive survey data analysis. JMIR Med Inform 2020 Feb 10;8(2):e17006 [FREE Full text] [doi: 10.2196/17006] [Medline: 32039815]
17. Zheng X, Rodríguez-Monroy C. The development of intelligent healthcare in China. Telemed J E Health 2015 May;21(5):443-448. [doi: 10.1089/tmj.2014.0102] [Medline: 25671730]

XSL•FO
RenderX

18. Bhuiyan MN, Rahman MM, Billah MM, Saha D. Internet of Things (IoT): a review of its enabling technologies in healthcare applications, standards protocols, security, and market opportunities. IEEE Internet Things J 2021 Jul 1;8(13):10474-10498. [doi: 10.1109/jiot.2021.3062630]
19. Okpala P. Assessment of the influence of technology on the cost of healthcare service and patient's satisfaction. International Journal of Healthcare Management 2017 Jun 07;11(4):351-355. [doi: 10.1080/20479700.2017.1337623]
20. Chintamaneni S, Yatham P, Stumbar S. From East to West: a narrative review of healthcare models in India and the United States. Cureus 2023 Aug;15(8):e43456 [FREE Full text] [doi: 10.7759/cureus.43456] [Medline: 37711922]
21. Shafique K, Khawaja BA, Sabir F, Qazi S, Mustaqim M. Internet of Things (IoT) for next-generation smart systems: a review of current challenges, future trends and prospects for emerging 5G-IoT scenarios. IEEE Access 2020;8:23022-23040. [doi: 10.1109/access.2020.2970118]
22. Kouroubali A, Katehakis DG. The new European interoperability framework as a facilitator of digital transformation for citizen empowerment. J Biomed Inform 2019 Jun;94:103166 [FREE Full text] [doi: 10.1016/j.jbi.2019.103166] [Medline: 30978512]
23. Hoque R, Sorwar G. Understanding factors influencing the adoption of mHealth by the elderly: an extension of the UTAUT model. Int J Med Inform 2017 May;101:75-84. [doi: 10.1016/j.ijmedinf.2017.02.002] [Medline: 28347450]
24. Calegari LP, R.D. B, Fettermann DC. A meta-analysis of a comprehensive m-health technology acceptance. IJLSS 2023 Apr 06;15(1):1-21. [doi: 10.1108/ijlss-01-2023-0012]
25. Binyamin SS, Zafar BA. Proposing a mobile apps acceptance model for users in the health area: a systematic literature review and meta-analysis. Health Informatics J 2021 Jan 13;27(1):1460458220976737 [FREE Full text] [doi: 10.1177/1460458220976737] [Medline: 33438494]
26. Chauhan S, Jaiswal M. A meta-analysis of e-health applications acceptance. JEIM 2017 Mar 06;30(2):295-319. [doi: 10.1108/jeim-08-2015-0078]
27. Shen Y, Xu W, Liang A, Wang X, Lu X, Lu Z, et al. Online health management continuance and the moderating effect of service type and age difference: a meta-analysis. Health Informatics J 2022 Aug 17;28(3):14604582221119950 [FREE Full text] [doi: 10.1177/14604582221119950] [Medline: 35976977]
28. Gopinath K, Selvam G, Narayanamurthy G. Determinants of the adoption of wearable devices for health and fitness: a meta-analytical study. CAIS 2022;50(1):445-450 [FREE Full text] [doi: 10.17705/1cais.05019]
29. Zhang Z, Xia E, Huang J. Impact of the moderating effect of national culture on adoption intention in wearable health care devices: meta-analysis. JMIR Mhealth Uhealth 2022 Jun 03;10(6):e30960 [FREE Full text] [doi: 10.2196/30960] [Medline: 35657654]
30. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
31. Jeyaraj A, Rottman JW, Lacity MC. A review of the predictors, linkages, and biases in IT innovation adoption research. Journal of Information Technology 2006 Feb 01;21(1):1-23. [doi: 10.1057/palgrave.jit.2000056]
32. Baptista G, Oliveira T. A weight and a meta-analysis on mobile banking acceptance research. Computers in Human Behavior 2016 Oct;63:480-489. [doi: 10.1016/j.chb.2016.05.074]
33. Naranjo Zolotov M, Oliveira T, Casteleyn S. E-participation adoption models research in the last 17 years: A weight and meta-analytical review. Computers in Human Behavior 2018 Apr;81:350-365. [doi: 10.1016/j.chb.2017.12.031]
34. Bowman NA. Effect sizes and statistical methods for meta-analysis in higher education. Res High Educ 2011 Jul 13;53(3):375-382. [doi: 10.1007/s11162-011-9232-5]
35. Cooper H. Research Synthesis and Meta-Analysis: A Step-by-Step Approach. Thousand Oaks, CA: Sage Publications; 2017.
36. Borenstein M, Hedges L, Higgins J, Rothstein H, Ben Van Den A. Introduction to Meta-Analysis. Hoboken, NJ: Wiley; Sep 01, 2020.
37. Peterson RA, Brown SP. On the use of beta coefficients in meta-analysis. J Appl Psychol 2005 Jan;90(1):175-181. [doi: 10.1037/0021-9010.90.1.175] [Medline: 15641898]
38. Gupta N, Singh AK. Individual acceptance of Internet of Things: a meta analytical review. Int J Consumer Studies 2025 May 06;49(3):44. [doi: 10.1111/ijcs.70070]
39. DerSimonian R, Laird N. Meta-analysis in clinical trials revisited. Contemp Clin Trials 2015 Nov;45(Pt A):139-145 [FREE Full text] [doi: 10.1016/j.cct.2015.09.002] [Medline: 26343745]
40. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986 Sep;7(3):177-188 [FREE Full text] [doi: 10.1016/0197-2456(86)90046-2] [Medline: 3802833]
41. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002 Jun 15;21(11):1539-1558. [doi: 10.1002/sim.1186] [Medline: 12111919]
42. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634 [FREE Full text] [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
43. Suurmond R, van Rhee H, Hak T. Introduction, comparison, and validation of meta-essentials: a free and simple tool for meta-analysis. Res Synth Methods 2017 Dec 29;8(4):537-553 [FREE Full text] [doi: 10.1002/jrsm.1260] [Medline: 28801932]

XSL•FO
RenderX

44. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989 Sep;13(3):319. [doi: 10.2307/249008]
45. Venkatesh V, Davis FD. A theoretical extension of the Technology Acceptance Model: four longitudinal field studies. Management Science 2000 Feb;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
46. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425. [doi: 10.2307/30036540]
47. Venkatesh V, Thong JYL, Xu X. Consumer acceptance and use of information technology: extending the Unified Theory of Acceptance and Use of Technology. MIS Quarterly 2012;36(1):157. [doi: 10.2307/41410412]
48. Prentice-Dunn S, Rogers RW. Protection Motivation Theory and preventive health: beyond the Health Belief Model. Health Educ Res 1986;1(3):153-161. [doi: 10.1093/her/1.3.153]
49. Maddux JE, Rogers RW. Protection motivation and self-efficacy: a revised theory of fear appeals and attitude change. Journal of Experimental Social Psychology 1983 Sep;19(5):469-479. [doi: 10.1016/0022-1031(83)90023-9]
50. Janz NK, Becker MH. The Health Belief Model: a decade later. Health Educ Q 1984 Jan 01;11(1):1-47. [doi: 10.1177/109019818401100101] [Medline: 6392204]
51. Goodhue DL, Thompson RL. Task-technology fit and individual performance. MIS Quarterly 1995 Jun;19(2):213. [doi: 10.2307/249689]
52. Tang Y, Ning X. Understanding user misrepresentation behavior on social apps: the perspective of privacy calculus theory. Decision Support Systems 2023 Feb;165:113881. [doi: 10.1016/j.dss.2022.113881]
53. Ajzen I. From intentions to actions: a theory of planned behavior. In: Kuhl J, Beckmann J, editors. Action Control. Berlin/Heidelberg, Germany: Springer Publishing Company; 1985:11-39.
54. Rogers EM. Diffusion of Innovations, 5th Edition. New York, NY: Free Press; Aug 16, 2003.
55. Harrison JS, Banks GC, Pollack JM, O'Boyle EH, Short J. Publication bias in strategic management research. Journal of Management 2016 Jul 10;43(2):400-425. [doi: 10.1177/0149206314535438]
56. Yang HJ, Lee J, Lee W. Factors influencing health care technology acceptance in older adults based on the Technology Acceptance Model and the Unified Theory of Acceptance and Use of Technology: meta-analysis. J Med Internet Res 2025 Mar 28;27:e65269 [FREE Full text] [doi: 10.2196/65269] [Medline: 40153796]
57. Al-Rawashdeh M, Keikhosrokiani P, Belaton B, Alawida M, Zwiri A. IoT adoption and application for smart healthcare: a systematic review. Sensors (Basel) 2022 Jul 19;22(14):5377. [doi: 10.3390/s22145377] [Medline: 35891056]
58. Blut M, Chong AYL, Tsigna Z, Venkatesh V. Meta-analysis of the Unified Theory of Acceptance and Use of Technology (UTAUT): challenging its validity and charting a research agenda in the Red Ocean. JAIS 2022;23(1):13-95. [doi: 10.17705/1jais.00719]
59. Jeyaraj A, Dwivedi YK, Venkatesh V. Intention in information systems adoption and use: current state and research directions. International Journal of Information Management 2023 Dec;73:102680. [doi: 10.1016/j.ijinfomgt.2023.102680]
60. Al-Rawashdeh M, Keikhosrokiani P, Belaton B, Alawida M, Zwiri A. Effective factors for the adoption of IoT applications in nursing care: a theoretical framework for smart healthcare. Journal of Building Engineering 2024 Jul;89:109012. [doi: 10.1016/j.jobe.2024.109012]
61. Kim TB, Ho CB. Validating the moderating role of age in multi-perspective acceptance model of wearable healthcare technology. Telematics and Informatics 2021 Aug;61:101603. [doi: 10.1016/j.tele.2021.101603]
62. Cao J, Kurata K, Lim Y, Sengoku S, Kodama K. Social acceptance of mobile health among young adults in Japan: an extension of the UTAUT model. Int J Environ Res Public Health 2022 Nov 17;19(22):15156 [FREE Full text] [doi: 10.3390/ijerph192215156] [Medline: 36429875]
63. Hidayat-ur-Rehman I, Ahmad A, Akhter F, Aljarallah A. A dual-stage SEM-ANN analysis to explore consumer adoption of smart wearable healthcare devices. Journal of Global Information Management (JGIM) 2021 Jan;29(6):30. [doi: 10.4018/JGIM.294123]
64. Kim J. Analysis of health consumers' behavior using self-tracker for activity, sleep, and diet. Telemed J E Health 2014 Jun;20(6):552-558. [doi: 10.1089/tmj.2013.0282] [Medline: 24745608]
65. Liu Y, Lu X, Zhao G, Li C, Shi J. Adoption of mobile health services using the unified theory of acceptance and use of technology model: self-efficacy and privacy concerns. Front Psychol 2022 Aug 11;13:944976 [FREE Full text] [doi: 10.3389/fpsyg.2022.944976] [Medline: 36033004]
66. Karahoca A, Karahoca D, Aksöz M. Examining intention to adopt to internet of things in healthcare technology products. Kybernetes 2017 Dec 11;47(4):742-770. [doi: 10.1108/k-02-2017-0045]
67. Li J, Ma Q, Chan AH, Man S. Health monitoring through wearable technologies for older adults: smart wearables acceptance model. Appl Ergon 2019 Feb;75:162-169. [doi: 10.1016/j.apergo.2018.10.006] [Medline: 30509522]
68. Misra S, Adtani R, Singh Y, Singh S, Thakkar D. Exploring the factors affecting behavioral intention to adopt wearable devices. Clinical Epidemiology and Global Health 2023 Nov;24:101428. [doi: 10.1016/j.cegh.2023.101428]
69. Kang H, Han J, Kwon GH. The acceptance behavior of smart home health care services in South Korea: an integrated model of UTAUT and TTF. Int J Environ Res Public Health 2022 Oct 14;19(20):13279 [FREE Full text] [doi: 10.3390/ijerph192013279] [Medline: 36293859]

XSL•FO
RenderX

70. Mensah IK. Understanding the drivers of Ghanaian citizens' adoption intentions of mobile health services. Front Public Health 2022;10:906106 [FREE Full text] [doi: 10.3389/fpubh.2022.906106] [Medline: 35774576]
71. Alraja M. Frontline healthcare providers' behavioural intention to Internet of Things (IoT)-enabled healthcare applications: A gender-based, cross-generational study. Technological Forecasting and Social Change 2022 Jan;174:121256. [doi: 10.1016/j.techfore.2021.121256]
72. Jeng M, Pai F, Yeh T. Antecedents for older adults' intention to use smart health wearable devices-technology anxiety as a moderator. Behav Sci (Basel) 2022 Apr 18;12(4):114 [FREE Full text] [doi: 10.3390/bs12040114] [Medline: 35447686]
73. Papa A, Mital M, Pisano P, Del Giudice M. E-health and wellbeing monitoring using smart healthcare devices: an empirical investigation. Technological Forecasting and Social Change 2020 Apr;153:119226 [FREE Full text] [doi: 10.1016/j.techfore.2018.02.018]
74. Sinha M, Fukey L, Balasubramanian K, Hanafiah MH, Kunasekaran P, Ragavan NA. Acceptance of consumer-oriented health information technologies (CHITs): integrating technology acceptance model with perceived risk. IJCAI 2021 Oct 05;45(6):45-52. [doi: 10.31449/inf.v45i6.3484]
75. Tsai T, Lin W, Chang Y, Chang P, Lee M. Technology anxiety and resistance to change behavioral study of a wearable cardiac warming system using an extended TAM for older adults. PLoS One 2020;15(1):e0227270 [FREE Full text] [doi: 10.1371/journal.pone.0227270] [Medline: 31929560]
76. Yang Q, Al Mamun A, Hayat N, Salleh MFM, Jingzu G, Zainol NR. Modelling the mass adoption potential of wearable medical devices. PLoS One 2022;17(6):e0269256 [FREE Full text] [doi: 10.1371/journal.pone.0269256] [Medline: 35675373]
77. Yin Z, Yan J, Fang S, Wang D, Han D. User acceptance of wearable intelligent medical devices through a modified unified theory of acceptance and use of technology. Ann Transl Med 2022 Jun;10(11):629 [FREE Full text] [doi: 10.21037/atm-21-5510] [Medline: 35813345]
78. Ben Arfi W, Ben Nasr I, Khvatova T, Ben Zaied Y. Understanding acceptance of eHealthcare by IoT natives and IoT immigrants: an integrated model of UTAUT, perceived risk, and financial cost. Technological Forecasting and Social Change 2021 Feb;163:120437. [doi: 10.1016/j.techfore.2020.120437]
79. Koo JH, Park YH, Kang DR. Factors predicting older people's acceptance of a personalized health care service app and the effect of chronic disease: cross-sectional questionnaire study. JMIR Aging 2023 Jun 21;6:e41429 [FREE Full text] [doi: 10.2196/41429] [Medline: 37342076]
80. Huang C, Yang M. Empirical investigation of factors influencing consumer intention to use an artificial intelligence-powered mobile application for weight loss and health management. Telemed J E Health 2020 Oct 01;26(10):1240-1251. [doi: 10.1089/tmj.2019.0182] [Medline: 31971883]
81. Cheung ML, Chau KY, Lam MHS, Tse G, Ho KY, Flint SW, et al. Examining consumers' adoption of wearable healthcare technology: the role of health attributes. Int J Environ Res Public Health 2019 Jun 26;16(13):16132257 [FREE Full text] [doi: 10.3390/ijerph16132257] [Medline: 31247962]
82. David A, Yigitcanlar T, Li RYM, Corchado JM, Cheong PH, Mossberger K, et al. Understanding local government digital technology adoption strategies: a PRISMA review. Sustainability 2023 Jun 15;15(12):9645. [doi: 10.3390/su15129645]
83. McGraw D, Mandl KD. Privacy protections to encourage use of health-relevant digital data in a learning health system. NPJ Digit Med 2021 Jan 04;4(1):2 [FREE Full text] [doi: 10.1038/s41746-020-00362-8] [Medline: 33398052]
84. Venkatesh V, Brown SA, Bala H. Bridging the qualitative-quantitative divide: guidelines for conducting mixed methods research in information systems. MIS Quarterly 2013;37(1):21-54. [doi: 10.25300/MISQ/2013/37.1.02]
85. Venkatesh V, Brown S, Sullivan Y. Guidelines for conducting mixed-methods research: an extension and illustration. JAIS 2016 Aug;17(7):435-494. [doi: 10.17705/1jais.00433]
86. Sheikh A, Anderson M, Albala S, Casadei B, Franklin BD, Richards M, et al. Health information technology and digital innovation for national learning health and care systems. The Lancet Digital Health 2021 Jun;3(6):e383-e396. [doi: 10.1016/s2589-7500(21)00005-4]

Abbreviations

HBM: Health Belief Model
IoT: Internet of Things
PMT: Protection Motivation Theory
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
TAM: Technology Acceptance Model
TTF: task-technology fit
UTAUT: Unified Theory of Acceptance and Use of Technology

XSL•FO
RenderX

Edited by T Leung, A Coristine; submitted 08.Jul.2024; peer-reviewed by H Gandhi, J Walsh, GK Gupta, S Ashraf; comments to author 23.Dec.2024; revised version received 03.Mar.2025; accepted 31.Aug.2025; published 06.Jan.2026.

Please cite as:
Veiga I, Oliveira T, Naranjo-Zolotov M, Martins R, Karatzas S
Adoption of Internet of Things in Health Care: Weighted and Meta-Analytical Review of Theoretical Frameworks and Predictors
J Med Internet Res 2026;28:e64091
URL: https://www.jmir.org/2026/1/e64091
doi:10.2196/64091
PMID:

©Inês Veiga, Tiago Oliveira, Mijail Naranjo-Zolotov, Ricardo Martins, Stylianos Karatzas. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Review of Medical Data Sharing Initiatives With a Focus on the Use of Blockchain Technologies: Qualitative Comparative Analysis

James Cunningham[1], PhD; Nigel Davies[2], PhD; Sarah Devaney[3], PhD; Mike Harding[2], PhD; Søren Holm[3], PhD; Victoria Neumann[2], MA; Clare Smith[4], PhD; John Ainsworth[1], PhD

[1]School of Health Sciences, University of Manchester, Oxford Road, Manchester, United Kingdom

[2]School of Computing and Communications, Lancaster University, Lancaster, United Kingdom

[3]Centre for Social Ethics and Policy, University of Manchester, Manchester, United Kingdom

[4]Civic Health Innovation Labs, University of Liverpool, Liverpool, United Kingdom

Corresponding Author:
James Cunningham, PhD
School of Health Sciences, University of Manchester, Oxford Road, Manchester, United Kingdom

Abstract

Background: Medical data sharing initiatives are crucial for advancing research, improving patient outcomes, and fostering innovation in health care. With the advent of blockchain technology, there has been significant interest in exploring its potential to enhance the security, transparency, and efficiency of medical data sharing.

Objective: This study aimed to examine a selected set of blockchain-based medical data sharing initiatives, focusing on their governance, incentive structures, ownership models, business approaches, transaction mechanisms, and sustainability strategies. The analysis explored patterns in operational status and longevity, providing insight into the factors shaping these initiatives. The objective was to identify common characteristics and contextual factors that may influence their development and persistence.

Methods: The study used snowball sampling to identify a selection of primarily blockchain-based medical data sharing initiatives, drawing from academic literature, web searches, and expert consultations. To examine structural and operational patterns, initiatives were selected based on the availability of sufficient public documentation for systematic classification. Each initiative was categorized by governance, incentives, ownership, business models, transaction mechanisms, and sustainability strategies. A follow-up assessment examined operational status over time. The analysis applied qualitative comparative analysis to identify common structural features and relationships between governance, incentives, and sustainability.

Results: The survey identified 42 initiatives, categorizing them based on ownership, governance, business, incentive, transaction, and sustainability models. These categories were systematically identified and assigned numerical values to facilitate fuzzy-set qualitative comparative analysis. The base model, run at an inclusion threshold of 0.65, identified multiple configurations associated with sustained initiative activity, highlighting the role of governance mechanisms and transaction structures in supporting long-term viability. The sensitivity analysis, conducted across multiple thresholds, demonstrated that while several configurations remained stable, higher thresholds led to more restrictive solutions. At 0.80, only two configurations remained, representing the most consistent pathways to sustained activity, reinforcing the importance of governance and transaction models in initiative sustainability.

Conclusions: The analysis revealed a range of governance, ownership, business, and sustainability models, with no single structural configuration guaranteeing long-term viability. The findings suggest that governance and transaction mechanisms are particularly influential in sustaining initiatives, often compensating for the absence of strong business or sustainability models. The scope was limited to initiatives identified through available documentation and snowball sampling, and the results underscore the need for further research into the interplay between governance structures, financial models, and long-term sustainability in medical data sharing.

(J Med Internet Res 2026;28:e63575) doi:10.2196/63575

KEYWORDS

blockchain; medical data sharing; data governance; sustainability models; privacy; security; transparency; data provision; ethics

XSL•FO
RenderX

Introduction

Overview

As medical data have moved to being a solely digital artifact [1] and an increasing amount of health-related data has become available from noninstitutional sources [2], a large number of private businesses have emerged aiming to use these data and to profit from their use [3]. The inherently sensitive nature of medical data and the potential implications of it being misused or misappropriated mean that there is an increased need to constrain the use of this data from privacy, legal, and ethical standpoints [4]. These constraints, though, must be set against the potential of medical data to enable immense benefits in the areas of personal, public, and research-level health [5]. While government-funded or public initiatives to facilitate the use and reuse of medical data can play a role in leveraging secondary use of medical data for the public good, there is still a huge scope for benefit to be gained from allowing or encouraging private third parties to build for-profit businesses based on the consented use of personal medical information [6]. This then produces a natural tension between the benefit (both direct and indirect) of organizations obtaining medical data and associated permission to use it on the one side and a resistance, both on personal and jurisdictional levels, to sharing data on the other. Set alongside this is the need for such businesses to incentivize potential data subjects to consent to the use of their data and to do this in a way that is both ethically and legally sound [7].

Given the legal, ethical, and reputational risks and limitations associated with the use of medical data for profit-making means, the structure of the business models that enable this, and the means by which they navigate these risks, is of interest [8]. The sharing of medical data has the potential to transform health care, enabling better patient care, faster diagnoses, and more effective treatments [9]. However, it also raises important questions around privacy, ownership, and control. In recent years, there has been a proliferation of medical data sharing initiatives, from large-scale public-private partnerships to smaller-scale collaborations between health care providers and tech companies. While these initiatives hold promise for improving health care outcomes, they also raise important legal, ethical, and reputational concerns that must be carefully navigated. The structure of the business models underpinning these initiatives, and the measures taken to manage legal and ethical risks, therefore, warrants closer examination. To gauge the current landscape of medical data sharing initiatives, we have conducted a review, looking at existing medical data sharing initiatives and business models, their structure, and methods of incentivization. This paper presents the results of this survey.

This survey was conducted in the broader context of ongoing efforts to develop innovative and sustainable business models for the use of medical data. These efforts aim to balance the immense potential of medical data to improve health care outcomes with the critical need to address privacy, legal, and ethical concerns. By systematically reviewing and analyzing various medical data sharing initiatives, we look to inform the development of frameworks that ensure ethically sound and legally compliant use of medical data. The focus on blockchain technology reflects a growing interest in leveraging its capabilities to enhance data security and transparency, potentially addressing some of the key challenges in medical data sharing.

Background

As health care data have become predominantly digitized [10,11], there is increasing recognition of the value of digital data in driving innovation in general areas of public health [12], including public responses to the recent COVID-19 epidemic [13] and specific research fields such as genomics [14], dementia [15], and diabetes [16], among others. Alongside the digitization of health care data at the source (eg, within health care institutions), there has been a rise in personal health monitoring and data collection through commercially available wearable devices [17,18]. The growth of informatics infrastructure necessary to process and transport data, enable data linkage across datasets [19], and manage consent and sharing preferences [20] has further enabled third-party entities to use medical data for both public and private use [21].

Given the sensitive and personal nature of medical data, there are additional requirements for data processing and handling, particularly by third parties [22]. Ethical considerations for the use of data [23] and moral concerns as to its use [24], often becoming more limited depending on the specific type of medical data (eg, that relating to sexual health [25,26], mean that, even in conjunction with general limitations on the use of personal data [27], the use of such data is inherently limited. These ethical considerations translate into more firmly enforced legislative constraints [28], placing additional constraints on business models that seek to leverage personal medical data for profit-making purposes [3].

One area of interest in the use of medical data is the emergence of blockchain technologies [29], which offer opportunities for actors in a business transaction to communicate, transact, and share information in an open, tamper-proof, and trustless manner [30]. Blockchain is the underlying technology initially used to implement the Bitcoin cryptocurrency [31]. A blockchain is a data structure that encodes a series of blocks of data in series, with each block containing a cryptographic reference to the previous block in the chain. Combined with a consensus mechanism, which is a protocol for determining among a distributed set of potentially untrustworthy peers which series of blocks is the currently accepted chain, blockchains offer a technological means of sharing canonical sets of information in a trustless manner across a wide range of participants [32]. Blockchain-based systems have been enhanced by the development and application of smart contract technologies, which allow the formation of blockchains to be controlled and enhanced through the use of programmatic functions that control the input and output of transactions and data on-chain [33]. Blockchain technology has been adopted across a wide range of industries [34]. The use of blockchain technologies in the context of medical data sharing initiatives potentially allows for the construction of systems that provide guaranteed openness of data use [29], removal of the need for trusted third parties to coordinate actions such as data linkage [35], and a native layer

XSL•FO
RenderX

for offering, processing, and distributing tokens of incentivization [36]. The use of blockchain technology as the underlying mechanism for enabling the sharing and use of personal medical data by third parties is increasingly common at the present time [37].

In the era of data-driven health care, the growth in digital health information offers potential for improving patient outcomes and public health [38]. However, fragmented and siloed health data present significant challenges [39]. Data cooperatives address these challenges by fostering collective sharing, management, and governance of data among diverse stakeholders, including patients, health care providers, researchers, and institutions [40]. These cooperatives promote transparency, privacy, and informed consent, empowering individuals with greater control over their health information while adhering to regulatory and ethical standards [41]. Through collaboration and data pooling, data cooperatives bridge the gap between data availability and research, enabling comprehensive datasets for innovative medical research and evidence-based policymaking, driving improvements in patient outcomes and public health [42]. Data cooperatives differ from traditional data management approaches in several key ways. They operate on principles of inclusivity and collaboration, involving a wide range of stakeholders in the data sharing process. Patients have a central role, ensuring the protection of their privacy rights and determining access to their health information [43]. Advanced technologies and standardized protocols are used to ensure data interoperability, enhancing the quality and comprehensiveness of shared data [44]. Governance structures involve representatives from legal, ethical, and technical fields to develop policies that safeguard data integrity and security while fostering trust among participants [45].

In health care, data cooperatives can revolutionize the sharing, access, and use of sensitive medical information. By pooling diverse health datasets, cooperatives provide comprehensive views of patient medical histories, aiding health care professionals in making informed decisions about diagnosis, treatment, and care plans. Collaborative research efforts are facilitated by access to large and diverse datasets, promoting discoveries and advancements in medical science. Patients benefit from improved health outcomes by gaining control over their data and ensuring its protection [22]. Data cooperatives enhance population health initiatives by supporting data-driven policy decisions and public health interventions, advancing health care systems, and contributing to improved health outcomes on a broader scale. Despite their collaborative nature, viable business models are still needed to support the operation and sustainability of these cooperatives.

The Civic Data Identity Partnership

The Civic Data Identity Partnership (CDIP) project aimed to develop a technical platform to manage access to health data. The primary objective was to create an infrastructure that would allow citizens to see, control, and provide consent for the usage of their health data in research and decision-making processes. One of the aims of CDIP was to explore the viability of business models that support the sustainability of the platform [46].

To achieve these aims, the project engaged in consultations with a diverse range of stakeholders, including citizens, clinicians, medical personnel, and data commissioners. These consultations were crucial in identifying the technical, legal, social, and ethical challenges associated with using the CDIP platform to enhance access and control over health data. The project aimed to understand the level of control individuals desired over their health information sharing and the data requirements of other parties. The collaborative efforts of the CDIP project provided insights into the complex landscape of health data management. Stakeholder consultations addressed technical intricacies, legal implications, societal concerns, and ethical considerations [46]. By incorporating these perspectives, the project aimed to design a platform that integrated access and control of health data, fostering transparency, privacy, and informed decision-making. Ultimately, the CDIP project sought to establish an infrastructure that met the preferences and requirements of citizens, clinicians, and other key stakeholders, enabling them to leverage health data while maintaining autonomy over its usage [47].

Methods

Aims and Objectives

This study aimed to examine a selected set of blockchain-based medical data sharing initiatives, analyzing their governance structures, incentive mechanisms, ownership models, business approaches, transaction mechanisms, and sustainability strategies. The objective was to identify common characteristics and contextual factors that may influence their development and persistence.

The specific objectives were:

1. To identify blockchain-based medical data sharing initiatives using a combination of academic literature review, targeted web searches, and expert consultations.
2. To categorize these initiatives based on predefined criteria, including organization type, governance structure, incentive mechanisms, ownership model, business model, transaction mechanisms, and sustainability strategies.
3. To assess the operational status of the identified initiatives, examining whether they have continued, evolved, or ceased operations.
4. To analyze emerging patterns in governance, incentives, and sustainability, providing insights into the challenges and factors influencing the persistence of these initiatives.

Longitudinal Assessment of Identified Initiatives

In addition to categorizing and analyzing blockchain-based medical data sharing initiatives, the study included a longitudinal assessment to examine changes in their operational status. This involved tracking whether previously identified initiatives remained active and whether they had evolved in terms of governance, business strategies, or technical implementation. To conduct this assessment, publicly available sources—including websites, white papers, and social media updates—were reviewed, and project teams were contacted where possible to gather additional insights. The objective was to better understand the factors influencing the stability and evolution of these initiatives over time. By examining the

XSL•FO
RenderX

characteristics and status of blockchain-based medical data sharing initiatives, this study aimed to contribute to ongoing discussions around their feasibility, governance, and sustainability. The findings provide insights for stakeholders considering the adoption of blockchain in medical data sharing and highlight areas for future research.

Methodology

Overview

The study used a 2-stage methodology to examine blockchain-based medical data sharing initiatives. The first stage involved snowball sampling to identify initiatives with publicly available documentation, drawing from academic literature, online databases, industry reports, and directories of blockchain projects and startups. This approach ensured that selected initiatives had sufficient data for structured analysis.

The second stage involved systematic screening and classification, where initiatives were evaluated based on predefined criteria, including governance, incentives, ownership, business models, transaction mechanisms, and sustainability strategies. To identify structural patterns and relationships among these characteristics, the study applied qualitative comparative analysis (QCA). This method allowed for the systematic comparison of governance and sustainability models across initiatives, providing insights into factors influencing their operational status and persistence.

Snowball Sampling Methodology

Snowball sampling is a nonprobability sampling technique where existing study subjects recruit future subjects from among their acquaintances, creating a “chain referral” system. This method is particularly useful for accessing hard-to-reach or hidden populations by leveraging the social networks of initial participants [48]. Similarly, it can be applied to surveying literature and initiatives, where the initial set of identified sources and projects serves as the starting point, and the search is expanded by examining references within those sources and consulting with experts in the field. The initial set of initiatives identified through search terms served as the starting point for the snowballing approach, which was further expanded by examining references and consulting with experts.

The snowball sampling technique was chosen for this survey to capture a comprehensive range of medical data sharing initiatives, particularly those on the cutting edge or in experimental stages that may not have formal publications. Initially, relevant initiatives were identified through academic literature and web searches using the specified search terms. Subsequently, references within these sources were explored, and experts in the field were consulted to uncover additional initiatives. This approach was essential due to the dynamic nature of the field, where many projects, particularly those from startups and experimental ventures, may not be widely documented in formal literature but are actively discussed in online forums, social media, and industry reports. This iterative process allowed for the inclusion of diverse initiatives, ensuring a thorough and representative survey.

Search

To obtain a broad list of medical data research initiatives, a snowball sampling technique was used. The initial search was conducted in June 2022, and results were revisited in March 2024 for comparison. This search involved reviewing academic publications, reports, and relevant online sources using a wide range of search terms related to medical data sharing and blockchain technology. The following search terms were used: “medical data,” “healthcare data,” “clinical data,” “data sharing initiative,” “Data Coop,” “data bank,” “data trust,” “data brokerage,” “blockchain data sharing,” “DLT,” and “Distributed Ledger Technology.”

In addition to formal search terms, more informal terms and phrases were used for web-based and social media searches. These included keywords and phrases commonly used in industry discussions, online forums, and social media platforms, which helped in identifying initiatives that may not have formal academic publications. The informal search terms included blockchain health data, health data exchange, medical data blockchain project, digital health data sharing, health data startup, and blockchain health project.

Following the initial identification of initiatives through direct searching, the snowball sampling technique was applied by tracing references within the initial results. In academic literature, this involved 2 approaches: examining reference lists of relevant articles to find additional cited works and identifying new articles that cited these initial papers. For web-based and social media searches, discussions, and posts about identified initiatives were reviewed for mentions of other related projects, including links to comparable initiatives within forums and professional groups. Expert consultations further informed the process, providing insights into ongoing projects with available documentation. This iterative approach facilitated the identification of initiatives with sufficient publicly available information for structured classification and analysis, ensuring a relevant and analyzable dataset for QCA.

Screening

Each identified initiative was initially screened to confirm that it qualified as a medical data sharing initiative. Subsequently, they were categorized based on the following criteria: organization involved, initiative description, country of operation, blockchain-based technology, incentives for citizens, ownership, governance, business and transaction model, and sustainability. The initiatives were then screened to include only those projects that could be classified along these dimensions. The screening was conducted by 3 independent reviewers, with discrepancies resolved through discussion and consensus. This process ensured that the initiatives included had sufficient information available in these categories to conduct a comprehensive analysis.

1. Organization: Evaluated the mission and values of the organization behind the initiative.
2. Description: Assessed the scope and goals of the project.
3. Country-Based: Considered the geographical location and its potential impact.

XSL•FO
RenderX

4. Blockchain-Based: Evaluated the use of blockchain technology and its suitability for medical data.
5. Incentives for Citizens: Assessed the initiative's approach to incentivizing data sharing among citizens, including data control and privacy.
6. Ownership: Evaluated the ownership structure, including single-entity or membership-based models.
7. Governance: Assessed the governance structure and the role of stakeholders.
8. Business Model: Evaluated revenue streams and funding sources.
9. Sustainability Model: Assessed the long-term sustainability strategy of the initiative.

This screening process allowed for a systematic evaluation, ensuring the inclusion of initiatives that met the inclusion criteria and could provide insights into medical data sharing initiatives.

Evaluation of Ongoing Project Activity

A longitudinal assessment was conducted to evaluate the operational status of identified initiatives. An initial dataset was compiled between 2020 and 2023 as part of the CDIP project [46]. A follow-up assessment was performed between December 2023 and March 2024 to determine whether initiatives remained active, had evolved, or had ceased operations.

To determine whether the identified initiatives were still operational, we used a multistep approach involving direct verification and indirect evidence collection. The steps included:

1. Official Websites and Social Media: We visited the official websites and social media pages (LinkedIn, Twitter/X, Facebook) of each project. Active projects typically have recent posts, updates, or blog entries. The presence of recent content was considered an indicator of ongoing activity.
2. News Articles and Press Releases: We searched for recent news articles or press releases about the projects. Projects that are still operational often appear in the news regarding new partnerships, product updates, or funding rounds. These sources provided additional confirmation of the projects' statuses.
3. GitHub and Open Source Repositories: For open-source projects, we checked their GitHub repositories. We looked at the dates of the latest commits, pull requests, and issues. Active projects typically have recent activity in their repositories.
4. Industry Reports and Databases: We used industry databases such as Crunchbase, AngelList, and specialized blockchain directories to see if the projects were still listed as active. These platforms often provide updates on the operational status of companies.
5. Direct Inquiries: We sent inquiry emails to the contact information provided on the projects' websites. Direct communication with project teams or representatives was used to obtain the most accurate and up-to-date information regarding the current status of the initiatives.
6. Community Forums and Discussion Boards: We monitored blockchain and tech community forums like Reddit, Bitcointalk, and industry-specific channels. Discussions and posts in these forums provided insights into whether the projects were still considered active by the community.

By using these methods, we were able to systematically verify the operational status of each project, allowing an assessment of their current activity levels.

Qualitative Comparative Analysis

Overview of QCA as an Analytical Method

QCA is a set-theoretic comparative method that allows for the systematic examination of multiple cases to identify patterns and relationships between conditions and outcomes [49]. Compared to traditional variable-based statistical methods, QCA is particularly well-suited for analyzing complex social and organizational phenomena where causation is often configurational rather than linear. It enables the identification of necessary and sufficient conditions for specific outcomes, making it a suitable approach for evaluating how different governance models, ownership structures, business models, and incentive mechanisms influence the persistence of blockchain-based medical data sharing initiatives.

QCA operates by assigning categorical values to case attributes and then applying a comparative logic to determine which configurations of conditions are consistently associated with an outcome of interest. This method has been widely used in political science, management research, and social sciences to identify structural patterns among cases with diverse characteristics [50].

Application of QCA

In this study, QCA was applied to analyze whether specific combinations of governance models, ownership structures, incentive mechanisms, business models, and sustainability strategies were systematically associated with the continued operation or cessation of blockchain-based medical data sharing initiatives. The analysis followed a structured four-step process:

1. Case Selection and Calibration: Each initiative was treated as a case and categorized according to predefined dimensions. Attributes were assigned multivalue categorical codes to capture variation across initiatives. The outcome variable—whether an initiative remained active or ceased operation—was also coded accordingly. Fuzzy-set calibration was applied, converting categorical values into membership scores between 0 and 1, ensuring the analysis could accommodate varying degrees of attribute presence.
2. Construction of a Truth Table: A truth table was generated, listing all unique combinations of conditions present in the dataset. Each configuration was mapped to the corresponding outcome status of the initiatives exhibiting that combination of attributes. This step allowed for an initial assessment of how different structural characteristics were associated with sustained activity and grouped cases with similar profiles.
3. Identification of Configurational Patterns: The truth table was minimized using the Quine-McCluskey algorithm, which systematically reduced the dataset to the simplest set of configurations that explained the observed outcomes. The minimization process identified necessary and sufficient conditions: necessary conditions appeared in all active initiatives, while sufficient conditions represented

XSL•FO
RenderX

combinations of attributes that consistently led to an active status, even if no single factor was universally present.

4. Interpretation and Sensitivity Analysis: The minimized configurations were examined to identify distinct pathways leading to sustained initiative operation. The results demonstrated equifinality, meaning multiple different configurations could lead to the same outcome. Sensitivity analysis was conducted by adjusting the inclusion threshold for sufficiency, revealing how different conditions varied in their explanatory power across different levels of consistency.

This approach provided a systematic framework for assessing the structural factors influencing the viability of medical data sharing initiatives. The findings clarified how different combinations of governance, incentives, and sustainability mechanisms contributed to continued operation, offering insights into the conditions under which such initiatives remain active over time.

The Use of QCA

The use of QCA in this study is justified by the complex, multidimensional nature of blockchain-based medical data sharing initiatives, where governance structures, incentive mechanisms, and sustainability strategies interact in nonlinear ways. Unlike traditional regression-based approaches that assume additive relationships between independent variables, QCA allows for configurational causation, meaning that different pathways can lead to similar outcomes.

QCA provides a structured, comparative analysis of how blockchain-based initiatives are structured and which conditions are associated with long-term viability. This approach offers insights that can inform the design of future initiatives and policy frameworks aimed at sustaining blockchain-based medical data sharing models.

Results

List of Initiatives

Our search found 42 initiatives related to medical data sharing that were identified through the 2-stage methodology outlined in the previous section. Each initiative was systematically screened and analyzed along different dimensions to categorize them based on various characteristics. The resulting table provides an overview of the initiatives and their features, including information on their country of origin, the type of blockchain technology used, incentives for citizen participation, ownership structure, governance, business model, transaction model, sustainability, and whether they are public, private, or hybrid blockchains. A description of each initiative is provided in Table 1, highlighting key features and characteristics of each project.

XSL•FO
RenderX

Table . List of initiatives.

Organization	Description
e-Nome	Platform for consumers to control the use of their health and genomic data via blockchain technology.
Shivom	Blockchain-based genomics marketplace empowering individuals to control access to their genomic data.
Our data helps/Qntfy	Data donation for suicide prevention using social media and fitness data, powered by Qntfy.
EverGreen Life	App integrating GP[a] records, health data, and wellness information for user control.
Medical Chain	Decentralized platform for secure, fast, and transparent exchange and usage of medical data using blockchain.
My Clinic	Tool for health care professionals to connect with patients and share health records securely.
MIDATA	Nonprofit Swiss cooperative for storing personal health data and enabling participation in app-based research.
HealthBank Coop	For-profit Swiss cooperative health data exchange platform with complete data ownership for users.
Serto (formerly uPort)	Identity and trust management platform for verified data exchange and decentralized identity protocols.
Savvy Coop	Patient-owned cooperative connecting innovators with patients and caregivers for insights and rewards.
Sync for Science	Collaboration among EHR[b] vendors and NIH[c] to support the Precision Medicine Initiative by sharing EHR data.
Salus Coop	Citizen-owned cooperative facilitating data sharing to accelerate health research and innovation.
Carbon Coop	Cooperative focused on reducing home carbon emissions using innovative energy efficiency solutions.
23andMe	Personal genomics company providing genetic testing services and enabling research participation.
DNAtix	Platform for anonymous and secure DNA testing and storage using blockchain technology.
Zenome	Russian biotech company using blockchain to empower consumers to control and sell their genomic data.
Nebula Genomics	Personal genomics company providing affordable whole genome sequencing with blockchain-enabled data privacy.
PokitDok	API[d] platform for streamlining health care transactions and integrating real-time health insurance data.
Patientory	Platform for managing and analyzing population health data using blockchain technology.
Guardtime	Blockchain technology for security and scalability in various enterprise solutions, including Estonia's e-government.
Gravitate Health	Digital health information project co-led by the University of Oslo and Pfizer, using open-source digital platforms.
Chronicled	Supply chain solution for the health care system using blockchain technology for trust and automation.
Gem	Blockchain identity platform for secure data sharing and identity management in health care.
Doc.AI	AI[e] platform for health data collection and management, using blockchain for data security.
Iyro	Protocol for secure and private exchange of medical data using blockchain and proxy re-encryption.

XSL•FO
RenderX

Organization	Description
Coral Health	Platform for secure sharing of medical records and personal health data using blockchain.
EncrypGen	Platform for secure exchange of genomic data for cryptocurrency tokens using blockchain.
Blockpharma	Blockchain solution for drug traceability and counterfeit detection.
BurstIQ, Inc	Blockchain-enabled data network solutions for health care, focusing on data sharing and monetization.
Bodyo	Telehealth devices and AI profiling for health monitoring and chronic disease prevention.
Exochain	Blockchain platform for secure patient data interaction in clinical trials and health care.
PharmaLedger	Blockchain-enabled health care solutions platform, supported by the Innovative Medicines Initiative.
Curisium/Healthverity	Platform for secure, privacy-compliant real-world data exchange in health care.
HealthCombix	Blockchain platform for patient data control, disease prediction, and decentralized payments.
SimplyVital Health	Platform for exchanging health data among medical providers.
Blockchain Health Co	Platform for direct sharing of medical information with researchers using blockchain.
Akiri	Software-defined network for secure and compliant health care data transport without data storage.
Embleema	Blockchain network for secure sharing of personal health records and participation in clinical trials.
Med.me	Blockchain-based platform for rapid exchange of electronic health records and medical data transactions.
Ethereum Health Wallet	Blockchain platform for storing and sharing medical records globally.
Healthbank Technologies	Direct-to-patient platform for health care providers to connect with patients for various services.
Solve.Care	Blockchain technology platform aimed at reducing health care administration costs and improving outcomes.

[a]GP: general practitioner.
[b]EHR: electronic health record.
[c]NIH: National Institutes of Health.
[d]API: application programming interface.
[e]AI: artificial intelligence.

Case Selection and Calibration

Overview

We analyzed the selected set of blockchain-based medical data sharing initiatives, categorizing them across key dimensions relevant to their structure and sustainability. The analysis focused on governance models, ownership structures, business models, sustainability strategies, incentives for participation, blockchain adoption, and transaction mechanisms. These categories were selected based on their potential impact on the viability and persistence of initiatives.

Governance models were inferred where explicit information was unavailable, using related factors such as ownership type, decision-making structures, and regulatory compliance. Similarly, for other categories, when direct descriptions were not provided, classifications were assigned based on publicly available documentation, project descriptions, and associated operational characteristics. This approach ensured that initiatives were systematically categorized for comparative analysis while acknowledging variations in reporting detail.

Governance Model

Overview of Governance Structures Identified

Governance refers to the decision-making structures and oversight mechanisms used in medical data sharing initiatives. The analysis identified four primary governance models, with an additional category for cases where governance structures were unclear:

1. Board-Based Governance: Initiatives governed by a formal advisory board, ethics committee, or corporate governance

XSL•FO
RenderX

structure. These models were common among private, for-profit initiatives where oversight was centralized within an organization. This classification was assigned when documentation referenced board decision-making, formal advisory roles, or corporate governance policies.

2. Regulation-Based Governance: Initiatives whose governance was defined primarily by compliance with legal or regulatory frameworks, such as the General Data Protection Regulation (GDPR). These were often government-backed or public-private partnerships, and initiatives mentioning strict compliance requirements or legal structures were assigned to this category.
3. Membership-Based Governance: Initiatives governed by their participants, such as patient-led or cooperative models. These initiatives often had democratic decision-making structures, where users or stakeholders collectively determined policies. This classification was used when initiatives described member voting rights, cooperative ownership, or user-led decision-making.
4. User-Regulated Governance: Initiatives where governance was decentralized, with individual users retaining control over their own data. These models emphasized self-sovereign identity frameworks or direct user participation in access control. This category was assigned where initiatives explicitly stated that users determined governance decisions or when decentralized, self-regulated platforms were described.
5. No Clear Governance: Initiatives for which governance structures were not documented or could not be inferred from available information.

This classification allowed for a structured comparison of governance mechanisms across initiatives, facilitating the identification of patterns in oversight structures and their relationship to long-term viability.

Criteria for Classification

Governance was classified based on how decision-making authority and oversight mechanisms were structured within each initiative. Given the lack of explicit governance descriptions in many cases, classification was inferred using related factors such as ownership type, decision-making structures, regulatory references, and operational frameworks. The following criteria were applied:

1. Board-Based Governance: Assigned when documentation referenced a board of directors, advisory board, or corporate governance policies. Typically found in private for-profit initiatives with centralized oversight.
2. Regulation-Based Governance: Applied to initiatives explicitly governed by external legal or regulatory frameworks such as GDPR compliance. Often observed in government-led or public-private models.
3. Membership-Based Governance: Identified when initiatives had collective governance by members, including cooperative ownership structures or voting mechanisms.
4. User-Regulated Governance: Assigned to initiatives where governance was decentralized, and individuals retained direct control over their own data and permissions.
5. No Clear Governance: Used when no governance information was available or could not be inferred from ownership or decision-making structures.

Distribution of Governance Models

Board-based governance was the most frequently observed model, particularly among private-sector initiatives with venture capital backing, where centralized oversight structures were well-defined. Regulation-based governance was present in government-backed projects and public-private partnerships, where compliance requirements shaped governance frameworks. Membership-based governance was less common but found in cooperative initiatives emphasizing patient or citizen participation. User-regulated governance was primarily associated with decentralized, blockchain-driven initiatives, particularly those leveraging self-sovereign identity models. Some initiatives lacked clear governance documentation, making classification difficult in cases where decision-making structures were not explicitly defined.

Ownership Structures

Overview of Ownership Structures Identified

Ownership structures define who controls an initiative and how decisions are made regarding data access, governance, and sustainability. The following six ownership models were identified:

1. Private Ownership: Initiatives controlled by a single private entity, such as startups or corporations. These organizations typically operate on a for-profit basis, with centralized decision-making.
2. Public-Private Partnership: Collaborative models where government entities work with private sector organizations. These initiatives often balance public-sector oversight with private-sector innovation and funding.
3. Government-Owned: Fully state-controlled initiatives established to support public health infrastructure, regulatory frameworks, or government-led research projects.
4. Membership-Owned (Patients/Citizens): Initiatives where individuals contributing data collectively own and participate in decision-making. These models often emphasize user rights, transparency, and democratic governance.
5. Cooperative Membership: Formal cooperative organizations where control and decision-making are distributed among a legally recognized group of members, often structured around collective ownership principles.
6. No Clear Ownership: Initiatives where ownership details were unavailable, unclear, or not explicitly documented. These cases included decentralized projects or those with insufficient public records.

Criteria for Classification

Ownership structures were categorized based on the legal and organizational control of each initiative. Where ownership was not explicitly stated, classification was inferred from governance models, business structures, and stakeholder involvement. The following categories were used:

XSL•FO
RenderX

1. Private ownership was assigned to initiatives owned and controlled by a single private entity, such as startups or corporations. These initiatives typically operated on a for-profit model, with decision-making centralized within the organization.
2. Public-private partnership applied to initiatives where government entities collaborated with private sector stakeholders. These initiatives often combined public-sector oversight with private-sector innovation.
3. Government-owned initiatives were those fully controlled and operated by government agencies, typically established to support public health infrastructure or regulatory compliance.
4. Membership-owned (patients/citizens) initiatives were owned and governed by their participants, allowing individuals contributing data to have a direct role in decision-making. These models often aligned with cooperative governance structures.
5. Cooperative membership was assigned to initiatives collectively owned by a cooperative entity, where decision-making was distributed among members. Unlike membership-owned initiatives, cooperatives functioned as formal legal entities.
6. No clear ownership was assigned when no ownership structure was documented, and no inference could be drawn from related governance or business structures.

Distribution of Ownership Models

Private ownership was the most common structure, particularly among blockchain-based initiatives developed by startups or corporate entities. Public-private partnerships were relatively rare but present in projects that combined public health objectives with private-sector innovation. Government ownership was observed in select national health initiatives but was not a dominant model. Membership-owned initiatives were primarily seen in patient-driven projects that emphasized collective governance, while cooperative membership structures were identified in a smaller subset of initiatives explicitly structured as cooperatives. Some initiatives did not provide sufficient documentation to determine ownership, leading to classification under no clear ownership.

Business Model Structures

Overview of Business Model Structures Identified

Business models refer to the revenue-generating and operational frameworks used by medical data sharing initiatives. The analysis identified five primary business model structures, with an additional category for cases where the business model was unclear:

1. For-Profit Models: Initiatives that generate revenue through commercial activities such as subscription fees, data access sales, or transaction-based charges. This category was further divided into:
   - Fee or Subscription-Based Models: Users pay a recurring or one-time fee for access to services or data platforms.
   - Data Resale or Reuse-Based Models: Revenue is generated by selling or repurposing user data for research or commercial use.
   - Venture Capital-Backed Models: Initiatives funded by external investment, often prioritizing growth and scalability.
   - General For-Profit Models: Commercial initiatives that do not fit neatly into the above subcategories but generate revenue through various business activities.
2. Not-for-Profit Models: Initiatives that focus on public benefit rather than financial gain. These are often supported through grants, partnerships, or government funding and typically aim to improve health care, support research, or provide public health benefits.
3. Cooperative Models: Member-owned initiatives where participants collectively own and govern the organization. These models prioritize ethical data usage, community governance, and equitable distribution of benefits.
4. Membership Reimbursement Models: Initiatives where participants receive direct financial incentives or token-based compensation for contributing data. These models encourage participation by offering rewards or benefits tied to data sharing.
5. Unknown or Unclear Models: Initiatives where the business model was not explicitly documented or could not be inferred from available information.

This classification enabled a structured comparison of revenue strategies across initiatives, helping to identify patterns in financial sustainability and economic incentives.

Criteria for Classification

Business models were classified based on how initiatives generated revenue and sustained operations. Given the lack of explicit financial disclosures in some cases, classification was inferred using related factors such as pricing structures, funding sources, and service models. The following criteria were applied:

1. For-Profit Models: Assigned when initiatives explicitly charged for services, sold data access, or referenced investor-backed growth strategies.
   - Fee or Subscription-Based: Used when initiatives charged users a recurring or one-time fee.
   - Data Resale or Reuse-Based: Applied when initiatives monetized user data through resale or licensing agreements.
   - Venture Capital-Backed: Identified when initiatives referenced venture capital funding or external investment rounds.
   - General For-Profit: Used for commercially driven initiatives that did not clearly fit into other subcategories.
2. Not-for-Profit Models: Applied to initiatives that identified themselves as nonprofit organizations or relied primarily on grants, donations, or public funding.
3. Cooperative Models: Assigned when initiatives described collective governance structures where members had ownership stakes and participated in decision-making.

XSL•FO
RenderX

4. Membership Reimbursement Models: Used when initiatives provided direct incentives (eg, financial rewards, cryptocurrency, or tokens) for data sharing.
5. Unknown or Unclear Models: Assigned when no clear business model was documented or when financial sustainability mechanisms could not be determined.

Distribution of Business Models

For-profit models were the most common, particularly among private companies leveraging subscription fees, data monetization, or transaction-based revenue streams. Many of these initiatives were venture-backed, suggesting an emphasis on scalability and commercialization. Not-for-profit models were observed in research-driven projects and public-sector initiatives where financial sustainability relied on grants or government support. Cooperative models were less frequent but represented a distinct approach emphasizing member ownership and democratic decision-making. Membership reimbursement models were primarily found in blockchain-based initiatives where participants were incentivized through tokens or direct financial compensation. Some initiatives lacked clear business model documentation, making classification challenging where revenue mechanisms were not explicitly defined.

Blockchain Usage

Overview

The analysis of blockchain use within medical data sharing initiatives gave three categories:

1. Blockchain-Based Initiatives: These initiatives use blockchain technology to enhance data security, transparency, and patient control. By leveraging blockchain's decentralized ledger, they aim to ensure data integrity and facilitate secure sharing among stakeholders. Examples include e-Nome, Shivom, Medical Chain, and My Clinic.
2. Non-Blockchain-Based Initiatives: These initiatives use traditional centralized systems for data management and sharing. They may focus on interoperability and data exchange without incorporating blockchain technology. Examples include Savvy Coop, Sync for Science, Carbon Coop, and 23andMe.
3. Unknown Blockchain Usage: For some initiatives, the adoption of blockchain technology is unclear or not specified in available documentation. Examples include Our Data Helps/Qntfy, EverGreen Life, MIDATA, and Salus Coop.

Criteria for Classification

A project was classified as a blockchain-based project if it used blockchain technology as a core component of its architecture or if it claimed to use blockchain in its documentation or website. Projects that mentioned blockchain in passing or only used it for peripheral purposes, such as noncore payments functionality (eg, as an alternative currency for paying for goods or membership, as opposed to a direct reimbursement mechanism), were not classified as blockchain-based projects for the purposes of this study. The breakdown by blockchain usage is shown in the Truth Table section.

Distribution of Blockchain Usage

Most initiatives analyzed were blockchain-based, given the focus of our search methodology and reflecting a growing trend toward adopting decentralized technologies in the field of medical data sharing. These initiatives often highlight enhanced security, patient empowerment, and improved data interoperability as key benefits of blockchain integration. Non–blockchain-based initiatives continue to operate using established centralized systems, focusing on traditional methods of data exchange and management. A subset of initiatives lacked sufficient information to determine their use of blockchain technology, indicating a need for greater transparency in their operational disclosures.

Incentive Models

Overview of Incentive Models Identified

Incentives refer to the motivations that drive individuals and organizations to participate in medical data sharing initiatives. The analysis identified 5 primary incentive models, which account for both direct and indirect benefits to participants. Some initiatives use multiple incentive models, combining elements such as financial compensation with altruistic motivations. The identified categories are:

1. Self-Interest: Participants are motivated by personal benefits such as improved health care access, personalized insights, or streamlined administrative processes. These initiatives provide users with direct advantages, such as access to health records, genetic analysis, or medical consultations.
2. Altruism: Participation is driven by a desire to contribute to the broader public good, such as supporting medical research or advancing health care innovation. These initiatives often emphasize collective benefits over individual gains.
3. Indirect Financial Incentives: Participants expect potential future financial gains, such as improved market positioning, enhanced reputation, or cost savings through participation. These initiatives may not offer direct financial compensation but provide economic advantages over time.
4. Direct Financial Incentives: Participants receive immediate financial compensation for contributing data, such as payments, tokens, or vouchers. This model is often seen in initiatives that monetize data access or participation in research.
5. Non–citizen-Based Incentives: Benefits are directed toward broader community outcomes rather than individual participants. These include initiatives that improve public health, enhance health care systems, or provide large-scale infrastructure benefits.

Many initiatives blend multiple incentive structures, allowing individuals and organizations to participate for different reasons. For example, some for-profit initiatives incentivize user participation through direct financial rewards while also positioning themselves as contributors to health care research. Others rely on a combination of self-interest and altruism to attract participants.

XSL•FO
RenderX

Criteria for Classification

Initiatives were classified based on how they motivate participation, as inferred from available documentation. The following criteria were applied:

1. Self-Interest: Assigned when initiatives provided personal benefits such as access to health data, enhanced health care services, or convenience-based incentives.
2. Altruism: Applied to initiatives where individuals contributed data or participated in research with no expectation of personal gain.
3. Indirect Financial Incentives: Used when future financial benefits were implied, such as improving market standing or gaining access to health care services that could lower costs over time.
4. Direct Financial Incentives: Assigned when participants received explicit monetary compensation or exchangeable tokens in return for their participation.
5. Non–citizen-Based Incentives: Used for initiatives that emphasized benefits to society, health care providers, or broader community health outcomes rather than individual rewards.

In cases where an initiative aligned with multiple incentive models, the primary motivation identified in its operational structure was used for classification.

Distribution of Incentive Models

Self-interest was the most commonly observed incentive model, particularly among initiatives offering health insights, access to medical records, or streamlined services. Altruism was also prevalent, especially in research-oriented initiatives where users contributed data without expecting financial returns.

Direct financial incentives were primarily associated with for-profit initiatives using tokenized compensation models or offering payments for data sharing. Indirect financial incentives were seen in initiatives that provided long-term economic benefits rather than immediate compensation.

Non–citizen-based incentives were less frequent but were present in initiatives aiming to improve public health outcomes, reduce systemic inefficiencies, or contribute to health care policy development. Many initiatives combined multiple incentive models, reflecting the complexity of participant motivations in the medical data sharing ecosystem.

Sustainability Models

Overview of Sustainability Models Identified

The analysis of long-term business sustainability revealed that initiatives mainly relied on six different models to generate revenue and ensure operational longevity. Some initiatives combined multiple models to diversify income streams and reduce financial risk. The sustainability models identified are:

1. Membership Fee-Based: Initiatives charge recurring fees from members to cover operational costs. This model provides a predictable revenue stream and is commonly used by cooperatives or community-driven platforms.
2. Selling Data Access to Third Parties: Revenue is generated by granting access to medical data for research or commercial purposes. This model is often used by for-profit companies engaging with pharmaceutical firms, public health bodies, or academic researchers.
3. Selling Access to Users: Initiatives generate revenue by selling access to data-related services directly to individuals. This may involve subscriptions for premium services, one-time purchases, or pay-as-you-go models.
4. Further Investment: These initiatives rely on external funding sources such as venture capital, grants, or public investment to sustain their operations and drive growth. This model is common among startups and research-driven projects.
5. Software-as-a-Service (SaaS): Revenue is derived from subscription-based access to data platforms, analytical tools, or storage solutions. This model is typically used by technology-driven initiatives offering data integration or analytics services.
6. Unknown: Some initiatives did not clearly specify their revenue model or sustainability strategy, making it difficult to classify them.

Criteria for Classification

Sustainability models were classified based on the primary revenue mechanism described by each initiative. When explicit information was unavailable, classification was inferred from related indicators such as business model structure, funding sources, and platform offerings. The following criteria were applied:

1. Membership Fee-Based: Assigned to initiatives that charge individual or institutional users recurring fees for access.
2. Selling Data Access to Third Parties: Applied to initiatives where revenue is primarily generated through selling data to external organizations.
3. Selling Access to Users: Used for initiatives where users pay directly for services, reports, or data-related tools.
4. Further Investment: Identified in cases where venture capital, grants, or other funding mechanisms were the primary source of sustainability.
5. Software-as-a-Service (SaaS): Assigned to platforms offering subscription-based access to tools, data storage, or analysis services.
6. Unknown: Used when sustainability strategies were not explicitly documented or could not be inferred from available information.

Distribution of Sustainability Models

The distribution of sustainability models varied across initiatives, reflecting differences in business objectives, funding structures, and target users. Many for-profit initiatives relied on selling access to users, often through subscription-based services or premium data offerings, ensuring a continuous revenue stream. Others focused on selling data access to third parties, particularly in genomic and health data sharing platforms, where anonymized datasets were monetized for research and commercial applications. Several initiatives, especially cooperatives and patient-led organizations, adopted a membership fee-based model to maintain operations through recurring contributions from their members. Venture capital and other forms of external investment were a key source of

XSL•FO
RenderX

sustainability for startups and research-driven initiatives, particularly those in early development stages. The Software-as-a-Service (SaaS) model was used by digital health platforms offering analytical tools or secure data storage on a subscription basis. However, a number of initiatives lacked clearly defined sustainability strategies, particularly among research collaborations and recently launched projects, indicating potential uncertainty in long-term financial viability.

Transaction Models

Overview of Transaction Models Identified

Transaction models refer to the mechanisms by which initiatives facilitate exchanges of value, including payments for access, incentives for data sharing, and internal economic systems. The analysis identified five primary transaction models, with an additional category for cases where transaction structures were unclear:

1. Direct Monetary Transactions: Initiatives where users make direct payments to access data or services. This model provides immediate financial returns but may limit accessibility for individuals or organizations with fewer resources.
2. Token-Based Issuance and Sales (Single-Token Models): Initiatives using a single cryptocurrency token for internal transactions. These tokens may serve multiple purposes, including payments, incentives, or governance.
3. Dual Token Mechanism: Initiatives that use two distinct tokens, typically separating governance from utility transactions. This approach helps distinguish between platform operations and broader ecosystem participation.
4. Reward-Based Systems: Initiatives where participants receive incentives in the form of points, credits, or tokens for sharing data or engaging with the platform. This model is commonly used in research-driven initiatives seeking to encourage participation.
5. Cryptocurrency Payments: Initiatives that facilitate transactions using established cryptocurrencies such as Bitcoin or Ethereum. This model provides interoperability with external blockchain networks.
6. No Direct Payment: Initiatives where services or data are provided freely without requiring payment from users. These models often rely on alternative funding sources such as grants, partnerships, or donations.
7. No Clear Transaction Model: Initiatives for which transaction models were not explicitly documented or could not be inferred from available information.

This classification provides a structured framework for understanding the financial mechanisms that underpin data sharing initiatives and their implications for accessibility, participation, and sustainability.

Criteria for Classification

Transaction models were classified based on the primary method used to facilitate exchanges within each initiative. Given that some initiatives did not explicitly define their transaction models, classification was inferred based on references to payment structures, tokenized economies, or incentive systems. The following criteria were applied:

1. Direct Monetary Transactions: Assigned when initiatives specified user payments for services or data access, either as a one-time purchase or as a recurring subscription.
2. Token-Based Issuance and Sales (Single-Token Models): Applied to initiatives where a single cryptocurrency token was used for internal transactions, including payments or incentives.
3. Dual Token Mechanism: Identified in cases where initiatives used two separate tokens for different functions, typically governance and utility.
4. Reward-Based Systems: Used for initiatives that provided nonmonetary rewards, such as points or internal tokens, for participation or data sharing.
5. Cryptocurrency Payments: Applied when initiatives enabled transactions using standard cryptocurrencies rather than internal tokens.
6. No Direct Payment: Assigned when initiatives explicitly stated that data access or services were provided without direct payments from users.
7. No Clear Transaction Model: Used when no information on financial transactions was available or could not be inferred from the initiative's structure.

This structured classification enabled a comparative analysis of how initiatives facilitated financial exchanges, highlighting different approaches to incentivization and economic sustainability.

Distribution of Transaction Models

Token-based systems were the most frequently observed transaction model, particularly in blockchain-based initiatives, where single or dual-token economies enabled internal transactions and incentivization mechanisms. Some initiatives relied on direct monetary transactions, particularly those offering commercial services or premium user features. Reward-based systems were implemented in research-driven initiatives, often as a means to encourage participation without requiring direct financial contributions. Cryptocurrency payments were less common but appeared in select initiatives that sought interoperability with external blockchain networks. A subset of projects provided services without direct payments, often relying on external funding sources. Finally, some initiatives lacked clear documentation of their transaction structures, making classification difficult in cases where financial models were not explicitly defined.

Comparison to Initial Survey

Overview

The CDIP project ran between 2020 and 2023. An initial comprehensive review was conducted at the inception of the CDIP project to assess the landscape of health data initiatives and collaborative projects. This review served as a benchmark to understand the existing data sharing ecosystem and the potential impact of data cooperatives on the health care domain. At the conclusion of the CDIP project, a follow-up review was performed to evaluate the evolution of these projects over the course of the initiative.

XSL•FO
RenderX

Results of Project Activity Assessment

Following the reanalysis conducted from December 2023 to March 2024, we assessed the operational status of the 42 identified medical data sharing initiatives. Our findings are as follows: (1) out of the 42 projects, 28 projects are still active, and (2) 14 projects have become inactive.

These results provide insights into the current landscape of medical data sharing initiatives. The assessment highlights the dynamic nature of the field, with several projects ceasing operations over time. Understanding the factors contributing to the success or failure of these initiatives is important for future developments in the sector. The analysis will look at the continuation rates of the projects in the context of the categories used to screen the projects in the first phase of the analysis, such as incentive models, governance structures, ownership models, and sustainability strategies. This will help identify patterns and insights that can inform the design, implementation, and business approach of medical data sharing frameworks.

The projects have been categorized by their current status and are listed in Table 2.

Table . Projects by current status.

Status	Projects
Active	EverGreen Life, Medical Chain, My Clinic, MIDATA, Serto (formerly uPort), Savvy Coop, Sync for Science, Salus Coop, Carbon Coop, 23andMe, DNAtix, Nebula Genomics, Patientory, Guardtime, Gravitate Health, Chronicled, Gem, Iyro, EncrypGen, Blockpharma, BurstIQ.inc, Bodyo, PharmaLedger, Curisium/Healthverity, Med.me, Healthbank Technologies, Embleema, and Solve.Car
Inactive	e-Nome, Shivom, Our data helps/Qntfy, HealthBank Coop, PokitDok, Zenome, Doc.AI, Coral Health, Exochain, HealthCombix, SimplyVital Health, Blockchain Health Co, Akiri, Ethereum, and Health Walle

Analysis of Results: Truth Table

Overview

A truth table was constructed to systematically represent the relationship between selected conditions and the observed status of each initiative in the follow-up assessment. The table organizes initiatives according to their characteristics, allowing for a structured comparison of configurations and outcomes.

The conditions included in the truth table were derived from the analysis of initiatives and categorized initiative structure and function. These conditions included ownership model, governance model, business model, transaction model, use of blockchain, incentive model, and sustainability approach. Each initiative was assigned values based on its classification in these categories. The outcome variable recorded the initiative's status at follow-up, distinguishing between those that remained active or ceased operation.

The table allowed for an analysis of patterns among initiatives with similar configurations, supporting the identification of potential associations between specific conditions and initiative status. The structured format facilitated the next stage of analysis, where these patterns were further examined to assess whether particular combinations of conditions corresponded with different outcomes.

The analysis used fuzzy-set qualitative comparative analysis (fsQCA) to account for the varying degrees of membership within each condition. Unlike crisp-set QCA, which requires binary categorization, or multivalue qualitative comparative analysis (mvQCA), which assigns discrete categorical values, fsQCA enables conditions to take on continuous values between 0 and 1, allowing for a more nuanced representation of the initiatives. This approach was appropriate given the complexity of the dataset, where governance structures, ownership models, transaction mechanisms, and sustainability strategies exhibited varying levels of adherence to predefined categories rather than fitting into strictly distinct groups. This facilitated the identification of patterns in how different degrees of condition membership were associated with the observed outcomes, rather than limiting the analysis to rigid categorical distinctions.

Truth Table

Truth Table Representation

The truth table presents the data in a structured format to facilitate the identification of configurational patterns in the next stage of the analysis (Table 3). Each initiative is represented as a case, with categorical numerical values assigned to each variable (Textbox 1). These variables include ownership model, governance model, business model, blockchain usage, incentive model, sustainability model, transaction model, and the status of the initiative at the follow-up assessment.

XSL•FO
RenderX

Table . Truth table of analysis.

Organization	Ownership model	Governance model	Business model	Transaction model	Blockchain based	Incentive model	Sustainability model	Status
e-Nome	1	5	5	7	1	5	6	0
Shivom	1	5	1	2	1	4	3	0
Our data helps/Qntfy	1	5	2	6	3	2	3	0
EverGreen Life	1	2	1	3	3	1	3	1
Medical Chain	1	5	1	2	1	3	3	1
My Clinic	1	5	1	3	1	1	1	1
MIDATA	3	3	3	5	3	2	1	1
HealthBank Coop	3	3	3	2	1	4	4	0
Serto (formerly uPort)	1	4	1	5	1	1	3	1
Savvy Coop	3	3	3	5	3	4	4	1
Sync for Science	2	2	2	6	3	2	6	1
Salus Coop	3	3	3	5	3	2	6	1
Carbon Coop	3	3	3	5	3	2	6	1
23andMe	1	2	1	3	3	1	4	1
DNAtix	1	5	1	2	1	1	3	1
Zenome	1	5	1	7	1	4	3	0
Nebula Genomics	1	5	1	3	1	3	3	1
PokitDok	1	5	1	7	1	5	6	0
Patientory	1	5	1	2	1	1	3	1
Guardtime	1	5	1	7	1	5	6	1
Gravitate Health	2	2	2	6	1	2	6	1
Chronicled	1	5	1	7	1	5	6	1
Gem	1	5	1	6	1	1	6	1
Doc.AI	1	5	1	4	1	3	6	0
Iyro	1	5	1	4	1	3	6	1
Coral Health	1	5	1	5	1	1	6	0
EncrypGen	1	5	1	2	1	3	6	1
Blockpharma	1	5	1	7	1	5	6	1
BurstIQ.inc	1	5	1	2	1	3	6	1
Bodyo	1	5	1	5	1	3	6	1
Exochain	1	5	1	2	1	4	6	0
PharmaLedger	2	2	2	6	1	2	6	1
Curisium/Healthverity	1	5	1	5	1	5	6	1
HealthCombix	1	5	1	5	1	1	6	0
SimplyVital Health	1	5	1	7	1	5	6	0

XSL•FO
RenderX

Organization	Ownership model	Governance model	Business model	Transaction model	Blockchain based	Incentive model	Sustainability model	Status
Blockchain Health Co	1	5	1	7	1	5	6	0
Akiri	1	1	1	5	1	1	6	0
Embleema	1	5	1	2	1	4	6	1
Med.me	1	5	1	2	1	3	6	1
Ethereum Health Wallet	1	5	1	7	1	5	6	0
Healthbank Technologies	1	5	1	7	1	5	6	1
Solve.Care	1	5	1	3	1	3	6	1

XSL•FO
RenderX

Textbox 1. Key for Table 3.

Ownership model

- 1=Private business
- 2=Private business (investor backed)
- 3=Cooperative
- 4=Public-private
- 5=Unknown

Governance model

- 1=Board-based governance
- 2=Regulation-based governance
- 3=Membership-based governance
- 4=User-regulated governance
- 5=No clear governance

Business model

- 1=For-profit
- 2=Not-for-profit
- 3=Cooperative
- 4=Membership reimbursement
- 5=Unknown

Blockchain usage

- 1=Yes
- 2=No
- 3=Unknown

Incentive model

- 1=Self-interest
- 2=Altruism
- 3=Indirect financial incentives
- 4=Direct financial incentives
- 5=Non–citizen-based incentives
- 6=Unknown

Sustainability model

- 1=Membership fee-based
- 2=Selling data access to third parties
- 3=Selling access to users
- 4=Further investment
- 5=Software-as-a-service
- 6=Unknown

Transaction model

- 1=Direct monetary transactions
- 2=Token-based issuance and sales (single token)
- 3=Dual token mechanism

XSL•FO
RenderX

- 4=Reward-based systems
- 5=Cryptocurrency payments
- 6=No direct payment
- 7=Unknown

Status

- 1=Active
- 2=Inactive

This representation allows for systematic comparison of the initiatives, supporting the identification of configurations associated with initiative success or inactivity. The encoded values provide a standardized way to analyze the relationships between different attributes and their potential influence on outcomes.

Identification of Configurational Patterns

The analysis of configurational patterns was conducted using the QCA package in R (The R Foundation), which provides a well-documented and widely used framework for performing QCA [51]. Specifically, fsQCA was applied to analyze configurations of conditions leading to the observed outcomes [52]. Unlike crisp-set QCA, which requires binary presence or absence of conditions, or mvQCA, which assigns discrete categorical values, fsQCA allows for graded membership scores between 0 and 1. This flexibility made it well-suited for capturing the continuous nature of governance models, transaction mechanisms, incentive structures, and ownership types observed in the dataset.

The minimization process was conducted using the Quine-McCluskey algorithm [53], which systematically reduces the truth table into a set of minimal sufficient conditions by identifying prime implicants—combinations of conditions that consistently lead to the outcome. This method provides an exhaustive and deterministic approach to simplification, contrasting with heuristic-based solutions that rely on approximation.

A central consideration in fsQCA is equifinality, the principle that multiple distinct configurations can lead to the same outcome. By applying fsQCA, the analysis accounted for the possibility that different combinations of governance models, transaction mechanisms, incentive structures, and ownership types could all contribute to similar sustainability or operational outcomes. This approach enabled a comparative yet flexible assessment of how initiatives maintain viability over time.

To assess the robustness of the findings, a sensitivity analysis was conducted by systematically varying the inclusion threshold, which determines the minimum consistency score required for a configuration to be considered sufficient for the outcome. The analysis tested inclusion thresholds between 0.60 and 0.80, examining whether the identified configurations remained stable or changed across different threshold settings. This step helped evaluate the reliability of the results and detect marginal cases that may influence the solutions.

The analysis was implemented in R, using scripts that preprocessed the data, calibrated conditions, constructed the truth table, and performed the minimization process. The resulting configurations were analyzed to provide insight into the structural pathways associated with initiative status in the follow-up assessment.

Results of Configurational Analysis

Overview

The analysis examined the configurations of conditions that contributed to the sustainability of initiatives using fsQCA. This approach allowed for the identification of multiple pathways leading to sustained or inactive initiatives, rather than assuming a single causal relationship. The evaluation incorporated key attributes, including ownership models, governance structures, business models, incentive mechanisms, transaction models, and sustainability strategies.

A base model solution was first constructed using a consistency threshold of 0.65, providing an initial set of configurations that explained initiative outcomes. To assess the robustness of these findings, a sensitivity analysis was conducted by varying the consistency threshold across a range from 0.60 to 0.80. This iterative process tested the stability of the identified solutions and ensured that the results were not unduly dependent on a specific threshold choice.

By systematically analyzing the different conditions and their interactions, the study identified distinct structural configurations that were associated with sustained initiative activity. The results provided insight into how different combinations of governance, financial, and incentive models contributed to long-term viability, offering a comparative perspective on the factors influencing initiative success.

Base Model Solution

The base model, run at an inclusion threshold of 0.65, identified multiple configurations of conditions that were associated with the continued activity of initiatives. The minimization process resulted in two primary solutions (M1 and M2), each representing different pathways leading to sustained initiative status.

Minimized Configurations

The following configurations were identified as sufficient conditions for the outcome of sustained initiative activity (Y):

M1: ¬O * ¬B * ¬I * ¬T + ¬O * ¬G * ¬B * ¬T * ¬S

XSL•FO
RenderX

$+ O * \neg G * B * \neg I * \neg S + O * \neg B * \neg I * T * S$

$+ G * \neg B * \neg I * \neg T * \neg S + (\neg O * G * \neg B * \neg I * S) \rightarrow Y$

$M2: \neg O * \neg B * \neg I * \neg T + \neg O * \neg G * \neg B * \neg T * \neg S$

$+ O * \neg G * B * \neg I * \neg S + O * \neg B * \neg I * T * S$

$+ G * \neg B * \neg I * \neg T * \neg S + (G * \neg B * \neg I * T * S) \rightarrow Y$

where O=ownership model, G=governance model, B=business model, I=incentive model, T=transaction model, S=sustainability model, and ¬ (negation) indicates the absence of the condition.

These minimized configurations presented in the base model represent combinations of conditions that are sufficient for an initiative to remain active. Each equation outlines a distinct pathway through which different structural elements contribute to sustained initiative status.

The expressions use logical conjunctions ("*") to indicate that multiple conditions must co-occur, while the negation symbol (¬) denotes the absence of a given condition. The addition symbol ("+") signifies alternative pathways—meaning that different sets of conditions can independently lead to the same outcome (sustained activity, Y).

Interpretation of Base Model Results

The interpretations of the base model results are as follows:

1. Multiple Viable Pathways: The results confirm equifinality, meaning that multiple distinct configurations can lead to the same outcome of sustained activity.
2. Role of Ownership and Governance: Initiatives that lacked a strong ownership structure (¬O) but had specific governance structures (¬G) were among those with sustained activity, suggesting that alternative governance mechanisms may compensate for weaker ownership models.
3. Business and Transaction Models: The absence of a defined business model (¬B) in multiple configurations indicates that some initiatives sustain themselves without a strong commercial model, possibly through alternative funding or cooperative structures.
4. Sustainability and Incentives: Some pathways include the presence of a sustainability model (S) and financial incentives (I), while others suggest that initiatives without clear incentive models (¬I) can still persist, particularly if they incorporate effective governance and transaction mechanisms.
5. Alternative Configurations: The presence of G * ¬B * ¬I * ¬T * ¬S suggests that some governance-led models can sustain without direct financial incentives or sustainability planning.

Coverage and Consistency

Overview

The coverage table (Table 4) presents key metrics for the identified solutions (M1 and M2) that explain sustained initiative activity. The inclusion score (inclS) measures the consistency of each solution in explaining the observed outcomes. The proportional reduction in inconsistency further refines this by considering how well the solution differentiates between positive and negative cases. Solution coverage indicates the proportion of active initiatives explained by each solution. Since both solutions overlap significantly in the cases they cover, no unique coverage is present, meaning neither solution explains a distinct subset of initiatives that the other does not.

Table . Coverage and consistency.

Solution	Inclusion score (inclS)	Proportional reduction in inconsistency (PRI)	Solution coverage (covS)
M1	0.774	0.774	0.44
M2	0.784	0.784	0.447

The results confirm equifinality, showing that multiple distinct pathways can lead to sustained initiative activity. The high inclusion scores (≥0.77) indicate strong consistency, suggesting that these configurations reliably capture the factors associated with continued operation. However, the coverage scores (0.44 - 0.45) suggest that while these solutions account for a significant proportion of active initiatives, they do not capture all possible pathways. This highlights that some initiatives remain active through alternative, unexplored mechanisms.

The presence of solutions where ownership models are weak, but governance structures are present, suggests that effective governance can compensate for less formal ownership arrangements. Similarly, while some solutions incorporate financial incentives and sustainability models, others indicate that initiatives can persist without these elements, provided they integrate strong governance or transactional mechanisms. These findings emphasize the diversity of structural configurations that contribute to long-term viability.

Implications

The implications of the analysis are as follows:

- Initiatives with weaker ownership structures can persist if alternative governance models are in place.
- Financial incentives are not universally necessary, but their presence in certain configurations suggests they contribute positively.
- Sustainability strategies are not always essential, but those initiatives without them tend to rely on stronger governance or transaction mechanisms.

Sensitivity Analysis

Overview

A sensitivity analysis was conducted by systematically varying the inclusion threshold from 0.60 to 0.80 to assess the robustness of the identified configurations. The results demonstrate how different thresholds impact the number and structure of

XSL•FO
RenderX

minimized solutions. The results of this analysis are shown in Table 5.

Table . Sensitivity analysis.

Inclusion threshold	Minimized configurations	Inclusion score (inclS)	Proportional reduction in inconsistency (PRI)	Solution coverage (covS)
0.6	¬O * ¬B * ¬I * ¬T + ¬O * ¬G * ¬B * ¬T * ¬S + O * ¬G * B * ¬I * ¬S + O * ¬B * ¬I * T * S + G * ¬B * ¬I * ¬T * ¬S + ¬O * G * ¬B * ¬I * S	0.734	0.734	0.577
0.65	¬O * ¬B * ¬I * ¬T + ¬O * ¬G * ¬B * ¬T * ¬S + O * ¬G * B * ¬I * ¬S + O * ¬B * ¬I * T * S + G * ¬B * ¬I * ¬T * ¬S + G * ¬B * ¬I * T * S	0.774	0.774	0.44
0.7	¬O * ¬G * ¬B * ¬I * ¬T + ¬O * ¬G * ¬B * ¬T * ¬S + O * ¬B * ¬I * T * S + G * ¬B * ¬I * ¬T * ¬S + G * ¬B * ¬I * T * S + O * ¬G * B * ¬I * T * ¬S	0.808	0.808	0.395
0.75	¬O * ¬G * ¬B * ¬I * ¬T + ¬O * ¬G * ¬B * ¬T * ¬S + ¬O * ¬B * ¬I * ¬T * ¬S+O * ¬G * ¬B * ¬I * T * S + O * ¬G * B * ¬I * T * ¬S	0.84	0.84	0.243
0.8	¬O * ¬G * ¬B * ¬T * ¬S + O * ¬G * B * ¬I * T * ¬S	0.925	0.925	0.11

Summary of Sensitivity Results

The sensitivity results of the analysis are as follows:

- Lower thresholds (0.60 - 0.65) produced a greater number of solutions, capturing a broader range of configurations that may contribute to sustained initiative activity.
- At 0.70, some pathways observed at lower thresholds were eliminated, leading to a more refined set of explanatory configurations.
- Higher thresholds (0.75 - 0.80) significantly reduced the number of valid pathways, indicating that only the most consistently sufficient conditions remained.

The analysis across multiple inclusion thresholds identified stable configurations that appeared consistently, indicating relationships between governance, ownership, and incentive structures in determining sustained initiative activity. These configurations remained present even as the consistency threshold increased.

As the inclusion threshold increased, the number of configurations decreased, reflecting a more restrictive selection process that excluded cases with lower consistency scores. At a threshold of 0.75, fewer configurations met the criteria, and by 0.80, only two configurations remained. This indicates that while multiple pathways lead to sustained initiative status, only a subset consistently exhibits the highest levels of consistency.

The analysis also showed that the absence of a defined business model (¬B) and financial incentives (¬I) appeared in multiple solutions across different thresholds. This suggests that while financial incentives and structured business models may contribute to sustainability, other mechanisms, such as governance structures and transaction models, can support sustained activity in their absence.

The role of sustainability planning (S) varied across thresholds, indicating that while sustainability mechanisms contribute to long-term viability, they are not always necessary. Some initiatives maintained sustained activity without explicit sustainability planning, suggesting that alternative structural factors, such as governance and transaction mechanisms, may provide sufficient support.

Interpretation of Findings

The analysis identified multiple configurations of conditions associated with sustained initiative activity. The base model, run at an inclusion threshold of 0.65, produced two primary solutions, each representing distinct pathways leading to

XSL•FO
RenderX

continued operation. The presence of multiple valid configurations supports the principle of equifinality, indicating that different combinations of governance, ownership, business, incentive, and transaction models can contribute to sustained activity. The solutions demonstrated that initiatives without a strong ownership structure could persist if supported by governance mechanisms, suggesting that centralized control is not always necessary when alternative oversight structures are in place. Similarly, the absence of a defined business model in several configurations indicates that some initiatives operate successfully without a clear commercial framework, relying instead on cooperative models, external funding, or nonmonetary incentives.

The analysis also highlighted the role of financial incentives and sustainability planning in initiative survival. Some pathways included sustainability mechanisms and incentive structures, reinforcing the idea that structured support mechanisms contribute to long-term viability. However, other solutions demonstrated that initiatives without explicit financial incentives or sustainability frameworks could still maintain activity, particularly if supported by effective governance and transaction models. This suggests that while financial and sustainability considerations may enhance stability, they are not universally necessary for success.

The sensitivity analysis, conducted across multiple inclusion thresholds, provided further insight into the robustness of these findings. At lower thresholds, a broader range of configurations was identified, capturing a wider set of conditions leading to sustained initiative status. As the threshold increased, the number of valid solutions decreased, with only the most consistent configurations remaining at the highest inclusion levels. This pattern indicates that while multiple pathways can lead to continued activity, only a subset demonstrates the strongest and most consistent relationships between conditions and outcomes.

At higher thresholds, the solutions became more restrictive, eliminating cases with lower consistency scores. By the highest inclusion level of 0.80, only two configurations remained, representing the most reliable pathways associated with sustained initiative activity. This suggests that while various factors contribute to initiative sustainability, only a few consistently demonstrate strong explanatory power. The reduction in configurations at higher thresholds reinforces the importance of governance structures and transaction mechanisms as key determinants of continued operation, particularly when financial incentives and sustainability planning are absent.

Discussion

Methodology

The methodology used in this study followed a structured multistage approach. First, a snowball sampling technique was used to identify relevant medical data sharing initiatives, with a particular focus on governance, ownership, transaction models, and sustainability. This approach allowed for the inclusion of initiatives that may not be widely documented in academic literature but were actively discussed in industry reports, online databases, and open-source repositories. The initial set of initiatives was identified through keyword searches and industry reports, and additional initiatives were identified through references, expert consultations, and monitoring relevant discussions.

Following identification, each initiative was systematically analyzed across predefined factors, including governance structures, ownership models, business and transaction mechanisms, incentive structures, and sustainability approaches. These attributes were extracted from publicly available documentation, technical white papers, organizational policies, and, where possible, direct engagement with initiative representatives. The initiatives were then categorized using multivalue attributes to facilitate structured comparative analysis.

To assess the operational status of the initiatives, a follow-up analysis was conducted between December 2023 and March 2024. This verification process included reviewing official websites, press releases, regulatory filings, industry reports, and online community discussions. Initiatives were classified as either active or inactive, depending on the presence of ongoing operations, product updates, or sustained engagement.

Finally, fsQCA was applied to identify configurational relationships between the different factors and sustained initiative activity. This approach allowed for an analysis of multiple pathways leading to continued operation rather than assuming a single linear relationship. By systematically comparing different configurations of governance, ownership, incentives, and transaction mechanisms, the analysis identified the strongest structural factors associated with long-term initiative viability.

While the use of snowball sampling can introduce certain biases, it was particularly useful in this context due to the dynamic and rapidly evolving nature of the field of medical data sharing. Many initiatives in this space are emerging and may not be widely documented in traditional academic databases or directories. By consulting a range of sources and engaging with experts and stakeholders in the field, we aimed to capture a comprehensive and nuanced understanding of the current landscape of medical data sharing initiatives. This approach ensured that the review included as many relevant initiatives as possible, even those less visible or established, providing a thorough and representative survey of the field.

The use of fsQCA provided a structured approach to examining the relationships between organizational factors and initiative sustainability. In contrast to variable-based statistical methods, fsQCA identifies multiple configurations leading to an outcome, allowing for the identification of equifinality, where different sets of conditions can result in similar outcomes. This approach was suited to the diverse characteristics of the initiatives studied, which varied across governance structures, business models, and sustainability mechanisms. The calibration of categorical variables into fuzzy-set membership scores allowed for degrees of presence or absence rather than strict binary classifications. A sensitivity analysis across multiple inclusion thresholds was conducted to assess the consistency of the results. This allowed

XSL•FO
RenderX

for a structured analysis of relationships between factors while maintaining a focus on individual initiatives.

Classification of Initiatives

The classification of initiatives by governance, ownership, business, transaction, incentive, sustainability models, and blockchain usage provided a structured basis for analyzing patterns within medical data sharing initiatives. Governance models varied from centralized board-based structures to decentralized user-regulated approaches, reflecting differing levels of oversight and participant control. Ownership structures showed a predominance of private and cooperative models, with some initiatives operating as public-private partnerships or under government control. Business models highlighted distinctions between for-profit, nonprofit, and cooperative models, emphasizing different economic strategies for sustainability. Transaction models ranged from direct payments and tokenized systems to reward-based mechanisms and freely accessible services, demonstrating diverse approaches to financial interactions. Incentive structures incorporated self-interest, financial compensation, and altruistic motivations, illustrating how initiatives attract participation. Sustainability models varied widely, including membership fees, data sales, investment funding, and subscription-based services, indicating different strategies for long-term viability. Blockchain usage was primarily concentrated in initiatives aiming to enhance security, transparency, and data ownership, though some projects operated without blockchain-based frameworks. These classifications allowed for a comparative analysis of how different structural and economic factors contribute to the continued operation and impact of medical data sharing initiatives.

Analysis of Initiatives

The results of the analysis highlight several key factors associated with the continued activity of medical data sharing initiatives. Governance structures and ownership models emerged as significant influences, with initiatives that incorporated structured governance mechanisms—such as cooperative models, regulatory compliance, or board oversight—more frequently sustaining operations over time. While strong ownership models were not universally necessary, the presence of governance frameworks appeared to compensate for weaker ownership structures. This suggests that clearly defined decision-making and accountability mechanisms contribute to long-term viability, even in the absence of centralized control.

Financial sustainability and incentive structures also played a role in distinguishing active initiatives. The findings indicate that a well-defined business model is not always essential, as some initiatives without clear revenue-generating mechanisms remained operational. However, those that did sustain themselves often incorporated diverse funding strategies, including membership fees, data access sales, or external investment. Incentive models varied widely, with some initiatives relying on direct financial compensation for participation, while others leveraged nonfinancial motivations such as improved health care access or research contributions. The variability in sustainability and incentive strategies suggests that no single approach guarantees success, but initiatives that align their financial model with stakeholder expectations and engagement mechanisms are more likely to persist. Finally, the analysis demonstrated that transaction models, particularly those leveraging token economies or alternative financial structures, were not necessarily a determinant of success but functioned as supporting mechanisms that could enhance participation and sustainability when combined with effective governance and incentive models.

Limitations

Several limitations should be considered when interpreting these findings. The study analyzed a selective set of initiatives identified through snowball sampling and available documentation. While this approach allowed for a broad survey of relevant initiatives, it may have introduced selection biases, as less visible or undocumented initiatives could have been excluded. The reliance on publicly available information also meant that classification was sometimes based on inference rather than direct confirmation, particularly for governance structures and financial sustainability models. This introduces potential uncertainty, as internal decision-making processes and strategic priorities may not have been fully captured.

Additionally, the rapidly evolving nature of the space means that some findings may not remain applicable over time. New governance models, technological advancements, and shifts in regulatory frameworks could alter the viability of different configurations. Methodologically, QCA is sensitive to case selection and threshold settings, meaning that small changes in the dataset or inclusion criteria could influence the resulting configurations. While sensitivity analysis helped assess the robustness of the findings, the results should be understood as identifying patterns rather than deterministic causal relationships. Expanding the dataset to include a wider range of initiatives, including those without blockchain integration, could provide a more comprehensive perspective on the factors influencing long-term sustainability.

Future Research

Future research could build on this study by conducting longitudinal analyses to track how governance and business models evolve over time and how they impact long-term viability. Additional qualitative research could provide deeper insights into decision-making processes within initiatives, particularly regarding governance and financial sustainability. Interviews with key stakeholders could help validate the findings and offer further context on the mechanisms driving success or failure.

Further studies could also explore the role of emerging technologies, such as AI-driven data management and decentralized identity frameworks, in shaping the future of medical data sharing. Investigating how different regulatory environments influence the adoption and sustainability of these initiatives would also be valuable. Finally, expanding the dataset to include a broader range of initiatives, particularly those outside blockchain-based models, could provide a more comprehensive understanding of the factors influencing success in the medical data sharing ecosystem.

XSL•FO
RenderX

Conclusions

Our findings highlight the diverse governance, ownership, and financial sustainability models adopted by medical data sharing initiatives. The results suggest that no single factor determines long-term viability, but rather that different configurations of governance structures, business models, and incentives can contribute to sustained activity. Initiatives with centralized governance structures and clearly defined financial sustainability strategies were more likely to remain active. However, the presence of decentralized governance models and alternative funding mechanisms also demonstrated viability, particularly in cooperative and blockchain-based initiatives. The findings reinforce the importance of aligning governance structures with financial models to ensure long-term sustainability while maintaining transparency and user trust.

The analysis also suggests that financial incentives and transaction mechanisms play varying roles depending on the broader structural configuration of an initiative. While some successful initiatives relied on direct financial incentives, others sustained engagement through cooperative governance, shared decision-making, or intrinsic motivations such as access to health insights. This variability underscores the need for initiatives to tailor their incentive structures to their specific context, balancing financial sustainability with participant engagement and ethical data sharing practices.

Funding

This work was supported by the Engineering and Physical Sciences Research Council as part of the investigator-led Digital Economy Standard Grant (EP/S02770X/1). JA is funded by the NIHR Manchester Biomedical Research Centre. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health.

Conflicts of Interest

JA is a director and shareholder in CareLoop Health Ltd, a spinout company from the University of Manchester that develops and markets digital technology for severe mental illness. All other authors declare no conflicts of interest.

References

1. Gostin LO, Halabi SF, Wilson K. Health data and privacy in the digital era. JAMA 2018 Jul 17;320(3):233-234. [doi: 10.1001/jama.2018.8374] [Medline: 29926092]
2. Schüll ND. Data for life: wearable technology and the design of self-care. Biosocieties 2016 Sep;11(3):317-333. [doi: 10.1057/biosoc.2015.47]
3. Leeming G, Cunningham J, Ainsworth J. A ledger of me: personalizing healthcare using blockchain technology. Front Med (Lausanne) 2019;6:171. [doi: 10.3389/fmed.2019.00171] [Medline: 31396516]
4. Scheibner J, Raisaro JL, Troncoso-Pastoriza JR, et al. Revolutionizing medical data sharing using advanced privacy-enhancing technologies: technical, legal, and ethical synthesis. J Med Internet Res 2021 Feb 25;23(2):e25120. [doi: 10.2196/25120] [Medline: 33629963]
5. Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical data reuse or secondary use: current status and potential future progress. Yearb Med Inform 2017 Aug;26(1):38-52. [doi: 10.15265/IY-2017-007]
6. Azaria A, Ekblaw A, Vieira T, Lippman A. MedRec: using blockchain for medical data access and permission management. In: 2016 2nd International Conference on Open and Big Data (OBD) 2016:25-30. [doi: 10.1109/OBD.2016.11]
7. Vayena E, Haeusermann T, Adjekum A, Blasimme A. A digital health: meeting the ethical and policy challenges. Swiss Med Wkly 2018;148:w14571. [doi: 10.4414/smw.2018.14571]
8. Michelson KN, Adams JG, Faber JMM. Navigating clinical and business ethics while sharing patient data. JAMA 2022 Mar 15;327(11):1025-1026. [doi: 10.1001/jama.2022.1942] [Medline: 35212714]
9. Alami H, Gagnon MP, Fortin JP. Digital health and the challenge of health systems transformation. mHealth 2017;3:31-31. [doi: 10.21037/mhealth.2017.07.02]
10. Ghafur S, Van Dael J, Leis M, Darzi A, Sheikh A. Public perceptions on data sharing: key insights from the UK and the USA. Lancet Digit Health 2020 Sep;2(9):e444-e446. [doi: 10.1016/S2589-7500(20)30161-8] [Medline: 32838250]
11. Vassilakopoulou P, Grisot M, Blegind Jensen T, et al. Building national eHealth platforms: the challenge of inclusiveness. Presented at: ICIS 2017 Proceedings; Dec 10, 2017 URL: https://aisel.aisnet.org/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1081&context=icis2017 [accessed 2026-01-14]
12. van Panhuis WG, Paul P, Emerson C, et al. A systematic review of barriers to data sharing in public health. BMC Public Health 2014 Nov 5;14(1):1144. [doi: 10.1186/1471-2458-14-1144] [Medline: 25377061]
13. Budd J, Miller BS, Manning EM, et al. Digital technologies in the public-health response to COVID-19. Nat Med 2020 Aug;26(8):1183-1192. [doi: 10.1038/s41591-020-1011-4] [Medline: 32770165]
14. Bombard Y, Ginsburg GS, Sturm AC, Zhou AY, Lemke AA. Digital health-enabled genomics: opportunities and challenges. Am J Hum Genet 2022 Jul;109(7):1190-1198. [doi: 10.1016/j.ajhg.2022.05.001]

XSL•FO
RenderX

15. Perera G, Pedersen L, Ansel D, et al. Dementia prevalence and incidence in a federation of European electronic health record databases: the European Medical Informatics Framework resource. Alzheimers Dement 2018 Feb;14(2):130-139. [doi: 10.1016/j.jalz.2017.06.2270]
16. Roberto G, Leal I, Sattar N, et al. Identifying cases of type 2 diabetes in heterogeneous data sources: strategy from the EMIF project. PLoS One 2016;11(8):e0160648. [doi: 10.1371/journal.pone.0160648] [Medline: 27580049]
17. Iqbal SMA, Mahgoub I, Du E, Leavitt MA, Asghar W. Advances in healthcare wearable devices. NPJ Flex Electron 2021;5(1):1-14. [doi: 10.1038/s41528-021-00107-x]
18. Seneviratne S, Hu Y, Nguyen T, et al. A survey of wearable devices and challenges. IEEE Commun Surv Tutorials 2017;19(4):2573-2620. [doi: 10.1109/COMST.2017.2731979]
19. Harron K, Dibben C, Boyd J, et al. Challenges in administrative data linkage for research. Big Data Soc 2017 Dec;4(2). [doi: 10.1177/2053951717745678]
20. Hutchings E, Loomes M, Butow P, Boyle FM. A systematic literature review of attitudes towards secondary use and sharing of health administrative and clinical trial data: a focus on consent. Syst Rev 2021 May 4;10(1):132. [doi: 10.1186/s13643-021-01663-z] [Medline: 33941282]
21. Lee C, Luo Z, Ngiam KY, et al. Big healthcare data analytics: challenges and applications. In: Handbook of Large-Scale Distributed Computing in Smart Healthcare: Springer; 2017:11-41. [doi: 10.1007/978-3-319-58280-1_2]
22. Trinidad MG, Platt J, Kardia SLR. The public's comfort with sharing health data with third-party commercial companies. Humanit Soc Sci Commun 2020;7(1):1-10. [doi: 10.1057/s41599-020-00641-5] [Medline: 34337435]
23. Myers J, Frieden TR, Bherwani KM, Henning KJ. Ethics in public health research: privacy and public health at risk: public health confidentiality in the digital age. Am J Public Health 2008 May;98(5):793-801. [doi: 10.2105/AJPH.2006.107706] [Medline: 18382010]
24. Tomlinson T, De Vries R, Ryan K, Kim HM, Lehpamer N, Kim SYH. Moral concerns and the willingness to donate to a research biobank. JAMA 2015 Jan 27;313(4):417-419. [doi: 10.1001/jama.2014.16363] [Medline: 25626040]
25. Gibbs J, Solomon D, Jackson L, Mullick S, Burns F, Shahmanesh M. Measuring and evaluating sexual health in the era of digital health: challenges and opportunities. Sex Health 2022 Aug;19(4):336-345. [doi: 10.1071/SH22068] [Medline: 35970766]
26. Wade TD. Traits and types of health data repositories. Health Inf Sci Syst 2014;2(1):4. [doi: 10.1186/2047-2501-2-4] [Medline: 25825668]
27. Brinson NH, Rutherford DN. Privacy and the quantified self: a review of U.S. health information policy limitations related to wearable technologies. J Consum Aff 2020 Dec;54(4):1355-1374. [doi: 10.1111/joca.12320]
28. Scheibner J, Ienca M, Kechagia S, et al. Data protection and ethics requirements for multisite research with health data: a comparative examination of legislative governance frameworks and the role of data protection technologies. J Law Biosci 2020;7(1):lsaa010. [doi: 10.1093/jlb/lsaa010] [Medline: 32733683]
29. Dimitrov DV. Blockchain applications for healthcare data management. Healthc Inform Res 2019 Jan;25(1):51-56. [doi: 10.4258/hir.2019.25.1.51] [Medline: 30788182]
30. Ali O, Jaradat A, Kulakli A, Abuhalimeh A. A comparative study: blockchain technology utilization benefits, challenges and functionalities. IEEE Access 2021;9:12730-12749. [doi: 10.1109/ACCESS.2021.3050241]
31. Nakamoto S. Bitcoin: a peer-to-peer electronic cash system. URL: https://bitcoin.org/bitcoin.pdf [accessed 2026-01-14]
32. Pandey P, Litoriya R. Promoting trustless computation through blockchain technology. Natl Acad Sci Lett 2021;44(3):225-231. [doi: 10.1007/s40009-020-00978-0] [Medline: 32836612]
33. Khan SN, Loukil F, Ghedira-Guegan C, Benkhelifa E, Bani-Hani A. Blockchain smart contracts: applications, challenges, and future trends. Peer Peer Netw Appl 2021;14(5):2901-2925. [doi: 10.1007/s12083-021-01127-0] [Medline: 33897937]
34. Javaid M, Haleem A, Pratap Singh R, Khan S, Suman R. Blockchain technology applications for Industry 4.0: a literature-based review. Blockchain Healthc Today 2021 Dec;2(4):100027. [doi: 10.1016/j.bcra.2021.100027]
35. Cunningham J, Davidge G, Davies N, et al. Blockchain native data linkage. Front Blockchain 2021;4. [doi: 10.3389/fbloc.2021.667388]
36. Olla P, Elliott L, Abumeeiz M, Zegrean J, Rajasekar V. Crowdsourcing clinical research utilising blockchain-based incentivisation systems. Int J Blockchains Cryptocurrencies 2022;3(1):1-15. [doi: 10.1504/IJBC.2022.122986]
37. Chattu VK. A review of artificial intelligence, big data, and blockchain technology applications in medicine and global health. Big Data Cogn Comput 2021;5(3):41. [doi: 10.3390/bdcc5030041]
38. Blasimme A, Vayena E, Hafen E. Democratizing health research through data cooperatives. Philos Technol 2018 Sep;31(3):473-479. [doi: 10.1007/s13347-018-0320-8]
39. van Roessel I, Reumann M, Brand A. Potentials and challenges of the health data cooperative model. Public Health Genomics 2018 Aug 28;20(6):321-331. [doi: 10.1159/000489994]
40. Grossglauser M, Saner H. Data-driven healthcare: from patterns to actions. Eur J Prev Cardiolog 2014 Nov;21(2_Suppl):14-17. [doi: 10.1177/2047487314552755]
41. Ranchal R, Bastide P, Wang X, et al. Disrupting healthcare silos: addressing data volume, velocity and variety with a cloud-native healthcare data ingestion service. IEEE J Biomed Health Inform 2020 Nov;24(11):3182-3188. [doi: 10.1109/JBHI.2020.3001518] [Medline: 32750932]

XSL•FO
RenderX

42. Hardjono T, Pentland A. Data cooperatives: towards a foundation for decentralized personal data management. arXiv. Preprint posted online on May 21, 2019. [doi: 10.48550/arXiv.1905.08819]
43. Hafen E, Kossmann D, Brand A. Health data cooperatives – citizen empowerment. Methods Inf Med 2014;53(2):82-86. [doi: 10.3414/ME13-02-0051] [Medline: 24514946]
44. Alladi T, Chamola V, Parizi RM, Choo KKR. Blockchain applications for Industry 4.0 and Industrial IoT: a review. IEEE Access 2019;7:176935-176951. [doi: 10.1109/ACCESS.2019.2956748]
45. Bietti E, Etxeberria A, Mannan M, Wong J. Data cooperatives in Europe: a legal and empirical investigation. White Paper created as part of The New School's Platform Cooperativism Consortium and Harvard University's Berkman Klein Center for Internet & Society Research Sprint. 2021 Dec URL: https://cyber.harvard.edu/sites/default/files/2022-02/Data_Cooperatives_Europe-group2.pdf [accessed 2026-01-14]
46. Neumann V, Davidge G, Harding M, et al. Examining public views on decentralised health data sharing. PLoS One 2023;18(3):e0282257. [doi: 10.1371/journal.pone.0282257] [Medline: 36862723]
47. Cunningham J, Davies N, Devaney S, et al. The application of distributed autonomous organization governance mechanisms to civic medical data management. IET Blockchain 2024 Dec;4(S1):507-525. [doi: 10.1049/blc2.12062]
48. Naderifar M, Goli H, Ghaljaie F. Snowball sampling: a purposeful method of sampling in qualitative research. Strides Dev Med Educ 2017;14(3). [doi: 10.5812/sdme.67670]
49. Thomann E, Maggetti M. Designing research with qualitative comparative analysis (QCA): approaches, challenges, and tools. Sociol Methods Res 2020 May;49(2):356-386. [doi: 10.1177/0049124117729700]
50. Ragin CC, Rihoux B. Qualitative comparative analysis (QCA): state of the art and prospects. Qual Methods 2004;2(2):3-13.
51. Duşa A. QCA With R: A Comprehensive Resource: Springer; 2018. [doi: 10.1007/978-3-319-75668-4]
52. Schneider CQ, Wagemann C. Standards of good practice in qualitative comparative analysis (QCA) and fuzzy-sets. Comp Sociol 2010;9(3):397-418. [doi: 10.1163/156913210X12493538729793]
53. Jain TK, Kushwaha DS, Misra AK. Optimization of the Quine-McCluskey method for the minimization of the Boolean expressions. In: Fourth International Conference on Autonomic and Autonomous Systems (ICAS'08): IEEE; 2008:165-168. [doi: 10.1109/ICAS.2008.11]

Abbreviations

CDIP: Civic Data Identity Partnership
fsQCA: fuzzy-set qualitative comparative analysis
GDPR: General Data Protection Regulation
mvQCA: multivalue qualitative comparative analysis
QCA: qualitative comparative analysis
SaaS: Software-as-a-Service

Edited by T Leung; submitted 24.Jun.2024; peer-reviewed by F Albalwy, J Scheibner, M Gaynor; revised version received 14.Feb.2025; accepted 31.Aug.2025; published 22.Apr.2026.

Please cite as:
Cunningham J, Davies N, Devaney S, Harding M, Holm S, Neumann V, Smith C, Ainsworth J
A Review of Medical Data Sharing Initiatives With a Focus on the Use of Blockchain Technologies: Qualitative Comparative Analysis
J Med Internet Res 2026;28:e63575
URL: https://www.jmir.org/2026/1/e63575
doi:10.2196/63575

© James Cunningham, Nigel Davies, Sarah Devaney, Mike Harding, Søren Holm, Victoria Neumann, Clare Smith, John Ainsworth. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Resistance to Blockchain Adoption in Health Care Organizations: Evidence From a Cross-Sectional Study

Wilfried Bazomanza Nzabandora[1], PhD; Sehl Mellouli[1], PhD; Rohit Nishant[2], PhD

[1]Department of Organizational Information Systems, Faculty of Business Administration, Université Laval, 2325 Rue de la Terrasse, Québec, QC, Canada

[2]Queen's Business School, Queen's University Belfast, Belfast, United Kingdom

Corresponding Author:
Wilfried Bazomanza Nzabandora, PhD
Department of Organizational Information Systems, Faculty of Business Administration, Université Laval, 2325 Rue de la Terrasse, Québec, QC, Canada

Abstract

Background: Traditional health care systems struggle to ensure the security of medical data. To address these issues, organizations are exploring blockchain-based solutions, which offer strong security for managing medical data and transactions. Despite these benefits, adoption remains limited because many health care organizations are hesitant to implement blockchain apps due to perceived risks associated with these apps.

Objective: Our research aims to study the adoption of blockchain-based health care apps in health care organizations by adopting a risk-based approach. Using perceived risk theory (PRT), we developed a research model that links specific perceived risks of blockchain-based health care apps to the intention to adopt them.

Methods: This study used a cross-sectional design to test the research model through an online questionnaire administered to IT professionals from health care organizations in Canada and Africa. IT professionals were selected because they influence technology adoption and possess greater blockchain knowledge than other health care staff. A total of 217 responses were collected; after removing incomplete entries, 194 valid responses remained for analysis. This exceeded the minimum recommended sample size of 100 for structural equation modeling with 8 latent variables, 40 observed variables, a *P* value of .05, and an anticipated effect size of 0.3. Partial least squares structural equation modeling was used to test the hypothesized relationships in the research model.

Results: Harman single-factor test showed that the first factor accounted for 25.376% of the total variance (below 50%), indicating that common method bias was not a concern. Partial least squares structural equation modeling identified key risks influencing the overall perceived risk of blockchain-based health care apps and their adoption in health care organizations. The findings revealed that the relationship between perceived risk and intention to adopt blockchain-based health care apps is significant in Canadian health care organizations but not in African ones. Additionally, the availability of cloud-based blockchain solutions was found to reduce organizations' perceived risk related to insufficient resources for adopting and implementing blockchain-based health care apps.

Conclusions: This study used PRT to identify risks that may hinder the adoption of blockchain-based health care apps. Two risks—medical data disclosure and loss of control over medical data—were identified as blockchain-specific and incorporated as new PRT dimensions, extending the theory for blockchain contexts. The extended PRT can guide future studies examining reluctance to adopt blockchain apps. Results also show that contextual differences between Canada and Africa influence the relationship between perceived risks and adoption intentions. Our findings are beneficial to governments, health care organizations, and blockchain development agencies, as they will be able to implement strategies to mitigate some of the risks perceived by IT professionals related to blockchain-based health care apps.

(J Med Internet Res 2026;28:e77933) doi:10.2196/77933

KEYWORDS

blockchain-based health care apps; perceived risk theory; blockchain adoption; blockchain; risks; health care

XSL•FO
RenderX

Introduction

Overview

Over the past decade, different studies have examined the application of blockchain technology in different business domains [1,2]. Some of these studies are conducted in the health care domain [3,4]. The interest in the health care sector is explained by the profound transformations that blockchain can bring to this field concerning the management of medical data [5-7]. The health care sector is a unique context. It is highly regulated, as it includes sensitive data for which confidentiality must be respected [8]. Blockchain-based apps in this domain may provide more protection for health data by ensuring its confidentiality, availability, interoperability, and integrity [9]. Notable examples include (1) Medifakt (AKT Health), an Estonia-based decentralized blockchain platform for health care that integrates internet of things and machine learning technologies to improve patient care, enhance security, increase data transparency, and promote interoperability across health care systems [10]. The platform supports the transfer, storage, interpretation, and artificial intelligence evaluation of medical data and images [11]. (2) Embleema (Robert Chu), a blockchain solution designed to transform clinical trials and regulatory processes by enabling participants to digitally consent to the collection, storage, and analysis of their health care data [12]. This accelerates drug development while ensuring data integrity and transparency throughout the research lifecycle. (3) Blockpharma (Blockpharma), a blockchain-based platform that strengthens transparency, traceability, and security in pharmaceutical supply chain management [13]. Recording each stage of the supply chain enables patients to verify the authenticity of their medication and supports rigorous quality control through real-time monitoring.

Studies on blockchain in the health care sector are mainly focused on identifying its use cases and opportunities [14]. An analysis of these studies shows that blockchain is used in the health care sector for three main purposes: (1) patient sovereignty over their health care data [15], (2) interoperability of electronic health records [16,17], and (3) integrity and audit trail of medical transactions [18]. Besides the studies that identify blockchain's use cases and opportunities, other studies focus on analyzing the antecedents of the adoption of blockchain-based apps in the health care sector [19]. These studies take a social-cognitive approach, analyzing the factors that would facilitate the adoption of blockchain-based apps at individual or organizational levels [19-24]. Despite these studies, the health care sector is experiencing a low rate of adoption of blockchain-based apps [19]. The low adoption rate of blockchain-based apps in many sectors could be explained by organizations' reluctance to adopt them due to uncertainty about their risks, financial benefits, and performance [25]. The lack of successful use cases for blockchain-based apps prevents potential adopters from perceiving these risks and benefits [25]. This is the case in the health care sector, where blockchain is still maturing, and its various use cases are only for experimental purposes, with very basic functionalities [26].

Against this background, our study focuses on the low adoption of blockchain-based apps in the health care sector. We adopt a social-cognitive approach, using perceived risk theory as a theoretical lens, to identify how and to what extent certain perceived risks of blockchain-based apps can prevent their adoption in health care organizations. Identifying these risks will help explain why health care organizations' intentions to adopt blockchain-based apps do not translate into acts of adoption.

Theoretical Foundation and Conceptual Framework

This study analyzes how health care organizations' perceptions of the risks associated with blockchain-based apps influence their intention to adopt them. Therefore, we use perceived risk theory as the theoretical lens of our study, as it helps to explain the influence of individuals' or organizations' risk perceptions on their intention to adopt an IT.

Perceived Risk Theory

Developed by Bauer [27], perceived risk theory was initially used in studies interested in analyzing the factors that would explain individuals' purchasing decisions [28,29]. Unlike other individual behavior analysis theories, such as the technology acceptance model, which focuses on the factors likely to influence behavior [30], perceived risk theory focuses on factors that could explain individuals' resistance to adopt a particular behavior [31,32]. Thus, this theory establishes a relationship between the perceived risk of a behavior and its adoption by individuals [32].

A perceived risk is defined as a customer's perception of distrust and potential side effects that could affect the purchase of a product or service [33]. Featherman and Pavlou [34] also define this concept as the feeling of uncertainty regarding possible negative consequences of using a product or service. The literature includes different conceptualizations of perceived risk [35]. In some studies, it is conceptualized as a unidimensional concept [36,37] and in other studies as a multidimensional concept [35,38]. The conceptualization of perceived risk as a multidimensional concept makes it possible to consider it as a construct composed of a certain number of dimensions that represent the types of risks that can influence individuals' behavior [35,39]. For example, in a research study on online shopping, authors conceptualized perceived risk as an aggregate construct of 6 risk dimensions: financial, performance, social, physical, psychological, and time [40]. In information systems research, authors consider other types of risks, such as privacy and security risks (SRs), which extend the list of perceived risk dimensions [40].

Conceptual Framework and Research Hypotheses

Our conceptual framework includes 6 types of risks that constitute the perceived risk dimensions of blockchain-based health care apps (BBHAs): SR, failure to gain blockchain benefits risk (FGBBR), organizational readiness risk (ORR), risk of losing control of medical data (RLCMD), risk of disclosing data and medical practices (RDDMP), and immature blockchain ecosystem risk (IBER; Figure 1). These risks and the assumptions of our conceptual framework are explained in the following paragraphs.

XSL•FO
RenderX

Figure 1. Conceptual framework.

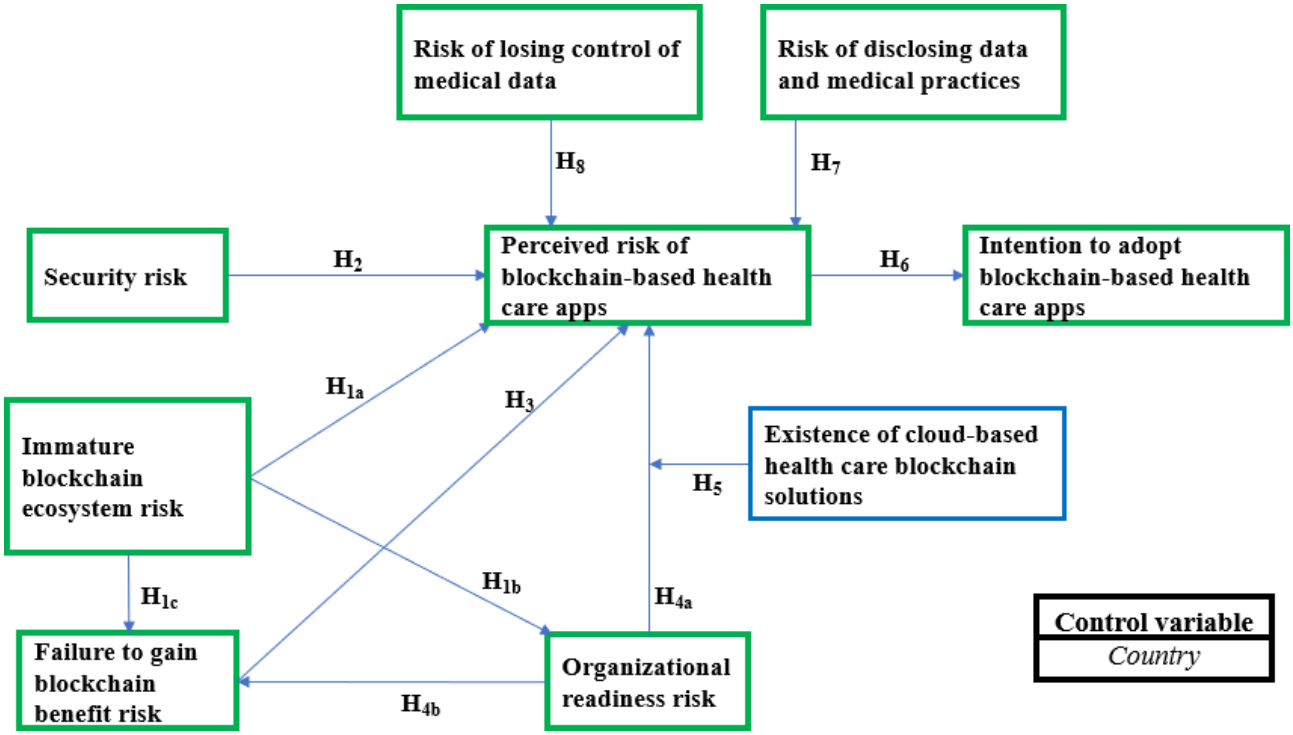

IBER in Health Care

The health care blockchain ecosystem is crucial for the adoption of blockchain-based apps and includes adopters, developers, consultants, auditors, regulators, and standardization bodies [41]. Each stakeholder supports adoption by ensuring app availability, guiding implementation decisions, and establishing regulatory frameworks that protect patient data. A mature ecosystem benefits from an increasing number of adopters, enabling partnerships among health care organizations that enhance the quality of health care services [8]. These collaborations encourage knowledge sharing and improve system security through broader participation in network consensus. Consequently, an immature blockchain ecosystem poses a significant risk, increasing health care organizations' perceived risk when adopting blockchain-based apps. Beyond shaping risk perceptions, the immaturity of the blockchain ecosystem also reduces health care organizations' readiness to adopt blockchain-based apps [42]. Indeed, the limited availability of blockchain-skilled professionals restricts organizations' ability to recruit personnel capable of implementing and maintaining blockchain health care systems [43,44]. Second, the small number of adopters limits access to feedback from prior blockchain investment projects [7,45]. Without such feedback, health care organizations lack clear information about resources required for adoption and use. These consequences of blockchain ecosystem immaturity weaken organizational preparedness and help explain the hesitation of many health care organizations to adopt BBHAs. The risk of the blockchain ecosystem's immaturity can also negatively affect the expected benefits of blockchain-based apps. For example, low blockchain adoption by health care organizations can lead to weak performance of the blockchain network [7]. This weak performance can result from a lack of business partnerships among health care organizations to share medical data through blockchain-based apps [7]. This can negatively impact the expected improvement in health care quality for health organizations.

SR of Blockchain-Based Apps

The ability of IT to guarantee the security of user data is decisive in the choice of its use at both the individual and organizational levels [40,46]. Health care organizations are driven by the adoption of blockchain-based apps to ensure greater security of patient medical data [47,48]. However, the security flaws identified in current blockchain-based apps indicate that these apps remain vulnerable to security attacks [49]. In BBHAs, several factors can contribute to security flaws, including a lack of patient education [50], poor governance of medical data [48], and noncompliance with regulations [51]. The negative consequences of risks associated with these factors for the security of medical data could influence health care organizations' perceptions of risk when adopting blockchain-based apps.

FGBBR

The risk of failure to gain the benefit of a product refers to the possibility that a product does not work as intended and/or does not provide the desired benefits [35]. For BBHAs, the failure to obtain the benefits of their use could be caused by several factors, including the nonintegration between these apps with existing health systems, the resistance of patients to share their medical data, and finally, the lack of alignment between the functionalities of these apps and business needs [7,52]. The risks associated with these factors may prevent health care organizations from fully realizing the benefits of blockchain-based apps, as they could negatively influence health care organizations' risk perceptions regarding adoption.

XSL•FO
RenderX

ORR

Organizational readiness is defined as "the availability of specific organizational resources to adopt new IT innovations" [20]. In this study, we examine the influence of ORR on health care organizations' perceptions of risk related to blockchain-based apps. This risk concerns an organization's lack of IT expertise, IT infrastructure, and financial resources needed to adopt blockchain-based apps. Therefore, we argue that, in the context of blockchain-based apps, inadequate resources for adoption would increase health care organizations' perceived risks. In several studies, the lack of organizational readiness is considered a factor that negatively influences the quality of use of a technology, and subsequently, the expected performance or profitability after its adoption [53,54]. For blockchain-based apps, the lack of competent human resources to use and maintain them, as well as the lack of technical infrastructure to support them, might not enable proper use and continuity of use within organizations. This can negatively influence the quality of services offered by blockchain-based apps and could prevent organizations from gaining the benefits of these apps.

Next, we investigate the moderating effect of cloud-based blockchain solutions on the relationship between ORR and the perceived risk of BBHAs. In fact, the adoption of blockchain-based apps by organizations requires adequate human, financial, and material resources to better exploit these apps and benefit from them [20]. Thus, the lack of these resources increases the risk perception of health care organizations regarding adopting blockchain-based apps [20]. To overcome this requirement, organizations use cloud computing for its technological, economic, and operational benefits [20,55]. In the blockchain context, cloud-based solutions would allow health care organizations to use apps based on this technology as Software-as-a-Service (SaaS) [56]. This significantly reduces the perceived risk of health care organizations regarding blockchain-based apps.

Adoption Intention of BBHAs

IT attributes play a major role in organizational IT adoption intention [57]. Before adopting an innovation, any organization evaluates it on 5 attributes: the complexity of the innovation, its compatibility with the organization, the benefit or advantage that it offers compared to other existing technological choices, and its observability and trialability [57]. In addition to these attributes, the intention to adopt an innovation is also influenced by the risks perceived by potential adopters [31,39,58,59]. Therefore, in the context of blockchain-based app adoption, we also argue that risk perception of these apps could negatively influence health care organizations' adoption intention.

RDDMP

The transparency of the blockchain's ledger ensures that the data recorded on it is visible to authorized network members [60]. For health care structures, ledger transparency can increase the risk of data and medical practice disclosure, with negative consequences for structures that disclose their data and medical practices. For example, when a health structure registers its care protocols on the blockchain ledger, other structures authorized to consult it can learn about those protocols. This can cause a loss of competitive advantage for the structure that discloses its protocols. Thus, the negative consequences associated with the disclosure of data and medical practices raise concerns among health care organizations, increasing their perceived risks of adopting blockchain-based apps.

RLCMD

Integrating blockchain in the health care sector brings a profound transformation in the management of patient medical data [60,61]. This technology puts the ownership of medical data in the hands of patients and places them at the center of its management [62]. Thus, no health structure can consult or use patients' medical data without consent [62]. The obligation to obtain patients' consent before consulting their medical data and the control patients can exercise over how their data are used significantly limit the power of health care structures over patients' medical data. This limitation on health care organizations' power over medical data increases their perception of risks regarding adopting blockchain in medical data management.

Each risk described above gives rise to a research hypothesis. Therefore, our research is based on the following hypotheses:

- H1: The risk of immaturity of the blockchain ecosystem in health care is positively associated with the perceived risk of blockchain-based apps by health care organizations (H1a), the risk of organizational readiness (H1b), and the risk of failing to benefit from blockchain-based apps (H1c).
- H2: SR is positively associated with the perceived risk of blockchain-based apps by health care organizations.
- H3: The risk of failure to gain benefits from blockchain-based apps is positively associated with health care organizations' perceived risk regarding these apps.
- H4: ORR is positively associated with the perceived risk of blockchain-based apps by health care organizations (H4a), as well as the risk of failure to gain benefits of blockchain-based apps (H4b).
- H5: The existence of cloud-based blockchain solutions (ECBS) moderates the relationship between ORR and the perceived risk of BBHAs, such that the positive relationship between ORR and the perceived risk of BBHAs is weak when cloud-based blockchain solutions are present.
- H6: Perceived risk of blockchain-based apps is negatively associated with the health care organizations' intention to adopt these apps.
- H7: RDDMP is positively associated with the perceived risk of blockchain-based apps by health care organizations.
- H8: RLCMD is positively associated with the perceived risk of blockchain-based apps by health care organizations.

Methods

Overview

In this section, we present the research design for our study, which includes 3 sections: research strategy and questionnaire design, data collection, and common method bias.

XSL•FO
RenderX

Research Strategy and Questionnaire Design

Our study adopts a cross-sectional research design, as it analyzes data from a population at a specific point in time [63]. Therefore, it was reported in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines adapted for cross-sectional studies [64]. The dimensions of the perceived risk and their reflexive measures are treated as latent (constructs) and manifest variables (items), respectively. To ensure construct validity, the items for each construct are taken from the literature, which we have subsequently adapted to our research context. We converted all the items to a semantic differential scale (1-10) to minimize common method bias by drawing inspiration from the measurement model proposed by Mitchell [65]. Table 1 shows the definition of constructs and the sources of their items. Table S1 in the Multimedia Appendix 1 provides the questionnaire.

Table . Operationalization of variables.

Construct and subconstructs (items)	Construct and subconstruct definitions	Sources of items
Intention to adopt		
None	A measure of the strength of one's intention to perform a behavior (behavioral intention) [66]	Venkatesh et al [67]; Glover and Benbasat [35]; and Teo et al [68]
Perceived risk: the feeling of uncertainty regarding possible negative consequences of using a product or a service [34]		
Failure to gain blockchain benefits risks	Failure to gain blockchain benefits regarding quality health care services, reduced cost, efficiency and security, and regulatory compliance [69]	Garg et al [69]
Security risk	Concerns that data and information may be open to security concerns, such as hacking, inaccurate information dispersal, and access to sensitive information [70]	Shah et al [71]; Bélanger and Carter [72]; and Hartono et al [73]
Organizational readiness risk	Risk beliefs of the health care organization's lack of financial and technological resources to adopt blockchain-based apps [20,74]	Kouhizadeh et al [70]
Immaturity blockchain ecosystem risk	Lack of community of players, such as blockchain experts and developers, blockchain adopters, and regulators interacting in the health care industry [75]	Alzahrani et al [76]; Kyriakidou et al [77]
Risk of losing control of medical data	Risk beliefs of health professionals regarding the loss of autonomy in managing patients' medical data [78]	Nogueira et al [79]; Ligiane Cristina Braga de Oliveira et al [80]
Risk of disclosing data and medical practices	Risk beliefs about submitting data and medical practices on the blockchain ledger shared by multiple health care organizations [81]	Bansal et al [82]; Bansal et al [83]

Data Collection

We conducted a survey to collect data over a 2-month period. This survey used an online questionnaire. Our study used purposive sampling. This type of sampling allowed us to select a sample from which the most can be learned [84]. Thus, IT professionals from health care facilities were selected as respondents to our study because of their greater knowledge of blockchain, a technology central to our research. The choice of a sample composed of IT professionals from health care facilities in Canada and emerging countries (the Democratic Republic of Congo, Burundi, and Rwanda) enabled us to minimize the risk of a low response rate, which could be caused by the nonparticipation of some respondents in our survey due to their ignorance of blockchain. We contacted 65 health care structures by email to request their participation (55 in Africa and 10 in Canada). At the end of the survey, we received 217 responses. After deleting responses with missing values (23), 194 valid responses were used for data analysis. We excluded responses with missing data for 2 primary reasons. First, these cases contained only demographic information and lacked any data on the study variables, making them unsuitable for analysis. Second, the number of excluded responses was minimal (23/243), representing less than 10% of the sample. Therefore, their removal is unlikely to affect the validity or representativeness of the dataset. The valid responses obtained for the study (n=194) exceeded the minimum recommended sample size (n=100) for structural equation modeling with 8 latent variables, 40 observed variables, P value of .05, and anticipated effect size of 0.3 [85]. Thus, the sample size was considered adequate for analyzing the influence of IT professionals' perception of risk regarding blockchain-based apps on their intention to adopt in health care structures. The respondents' demographic information is shown in Table 2.

XSL•FO
RenderX

Table . Demographic information.

Variables and categories	Participants, n (%)
Age (years)	
18 - 30	56 (28.9)
31 - 45	91 (46.9)
46 - 60	42 (21.6)
>60	5 (2.6)
Sex	
Female	67 (34.5)
Male	127 (65.5)
Continent	
Africa	189 (97.4)
America	5 (2.6)
Experience with blockchain-based apps	
Yes	171 (88.1)
No	23 (11.9)
Use of blockchain-based app within health care facility	
Yes	1 (0.5)
No	193 (99.5)

Control Variable

The model estimation includes a control variable, namely country (Figure 1), which we transformed into a binary variable (0=Africa and 1=Canada) before its use in SmartPLS (SmartPLS GmbH). A country's economic, legal, and political contexts play an important role in the assimilation of innovations within organizations [86]. Therefore, there is a significant difference in the assimilation of innovations between organizations in developed and developing countries [86,87]. Unlike developing countries, developed countries are characterized by an economic, political, and regulatory environment favorable to the assimilation of innovations [86]. Given that our study is conducted in Canada and Africa, we also want to analyze how the difference in context (economic, legal, and political) between Canada and African countries could affect the associations between the variables in the model, depending on whether the health care structures are in Canada or Africa. This justifies the introduction of the country as a control variable in the estimation of our theoretical model.

Common Method Bias

Common method bias results "from the fact that the predictor and criterion variables are obtained from the same source or rater, whereas others are produced by the measurement items themselves, the context of the items within the measurement instrument, and/or the context in which the measures are obtained" [88]. Since this research involved self-reported responses collected from a single setting, we will follow the 3-step approach proposed by Guan et al [89] to address the common method bias. First, the measurement items were adopted from well-established existing research. Second, in the survey, we made it clear to all the respondents that there was no right or wrong answer, and they should choose the answers that they considered the most appropriate. We also highlighted the fact that respondents' answers would be kept anonymous. Third, we conducted Harman single-factor test by entering all the measurement variables in an exploratory factor analysis using IBM SPSS 22.0. The sample would have a common method bias problem if a single construct explained more than 50% of the extracted variance. The results of the exploratory factor analysis show that the first factor explained 25.376% of the total variance (less than 50%), indicating that common method bias was not a problem in this dataset.

Ethical Considerations

This research has received ethical approval from the Research Ethics Committee of Université Laval (approval number: 2022 - 427 A-1 R-1/11-12-2023). During the data collection phase, each participant was informed of the research objectives and associated risks. This process enabled participants to provide informed consent by signing the designated consent form. To ensure confidentiality and privacy, data were initially collected using the LimeSurvey platform (LimeSurvey GmbH), which stores survey responses on a server located in Canada. Subsequently, the exported data were securely stored on Université Laval's licensed Microsoft 365 OneDrive server. Participant privacy was further protected by presenting research findings using anonymized data, which prevents the identification of individual respondents. No identification of individual participants is possible in the images of the manuscript or supplementary material. Participation in the study was entirely voluntary and was not contingent upon any monetary or other form of compensation.

XSL•FO
RenderX

Results

Overview

We used the partial least squares structural equation modeling (PLS-SEM) method to assess the hypothesized relationships between our constructs (using SmartPLS 4.1.0.8; SmartPLS GmbH) [90]. We follow a 2-phase approach to analyze the PLS-SEM model proposed by Guan et al [89]: (1) evaluation of the reliability and validity of the measurement model and (2) evaluation of the structural model.

Measurement Model

Table 3 shows the results of the measurement model. The measurement model was tested by examining both convergent and discriminant validity [89]. For testing convergent validity, we assess both item reliability and construct reliability of the measurement model [29]. Throughout the assessment of item reliability, we removed items associated with constructs with factor loading (λ) values below .7. As a result, we deleted 1 RLCMD item (RLCMD1), 2 SR items (SR1 and SR2), 3 IBER items (IBER1, IBER2, and IBER3), 2 ORR items (ORR2 and ORR4), and 4 FGBBR items (FGBBR1, FGBBR3, FGBBR8, and FGBBR10). Removing these items from their respective constructs enabled us to retain a measurement model composed of constructs with items whose factor loadings are greater than or equal to .7, confirming the item reliability of the measurement model. Construct reliability was verified by analyzing composite reliability (CR) and Cronbach α values. Each construct has CR and Cronbach α values that exceed the recommended threshold of .7, demonstrating the construct reliability of the measurement model. Furthermore, the average variance extracted (AVE) for each construct exceeds the minimum required value of .50, which confirms the convergent validity of the constructs in the measurement model [91]. Once convergent validity was established, we assessed discriminant validity. The discriminant validity was assessed using the Fornell and Larcker [92] method. As shown in Table 4, the square root of the AVE of each construct was greater than the correlations it shared with other constructs. For example, the square root AVE for FGBBR is .811, and this is greater than the correlation it shares with other constructs. As this was the case with all constructs, discriminant validity was obtained. Overall, the evidence indicates the adequacy of the measurement model. This leads to the next stage of testing the structural model.

XSL•FO
RenderX

Table . Measurement model results.

Constructs and items	λ[a]	VIF[b]	Cronbach α	CR[c]	AVE[d]
FGBBR[e]			0.896	0.920	0.657
FGBBR2	0.810	2.275			
FGBBR4	0.822	2.533			
FGBBR5	0.836	2.831			
FGBBR6	0.822	2.324			
FGBBR7	0.807	2.369			
FGBBR9	0.765	1.883			
IABHA[f]			0.844	0.896	0.685
IABHA1	0.894	3.213			
IABHA2	0.725	1.546			
IABHA3	0.902	3.327			
IABHA4	0.776	1.704			
IBER[g]			0.877	0.923	0.800
IBER1	0.888	2.811			
IBER3	0.894	2.772			
IBER5	0.901	2.022			
ORR[h]			0.867	0.919	0.790
ORR1	0.910	2.653			
ORR3	0.876	2.144			
ORR5	0.881	2.189			
RDDMP[i]			0.748	0.851	0.656
RDDMP1	0.734	1.728			
RDDMP2	0.804	1.611			
RDDMP3	0.887	2.144			
RLCMD[j]			0.854	0.906	0.764
RLCMD2	0.870	2.550			
RLCMD3	0.819	1.952			
RLCMD4	0.928	2.116			
SR[k]			0.784	0.857	0.600
SR3	0.732	1.612			
SR4	0.722	2.252			
SR5	0.727	2.248			
SR6	0.812	1.867			
SR7	0.838	1.907			

[a]λ: standardized factor loading.
[b]VIF: variance inflation factor.
[c]CR: composite reliability.
[d]AVE: average variance extracted.
[e]FGBBR: failure to gain blockchain benefits risk.
[f]IABHA: adoption intention of blockchain-based health care apps.
[g]IBER: immature blockchain ecosystem risk.
[h]ORR: organizational readiness risk.

XSL•FO
RenderX

[i]RDDMP: risk of disclosing data and medical practices.

[j]RLCMD: risk of losing control of medical data.

[k]SR: security risk.

Table . Results for discriminant validity.

Constructs	FGBBR[a]	IABHA[b]	IBER[c]	ORR[d]	PRBHA[e]	RDDMP[f]	RLCMD[g]	SR[h]
FGBBR	0.811	–0.231	0.439	0.429	0.864	0.217	0.289	0.428
IABHA	–0.231	0.828	–0.056	–0.061	–0.409	–0.232	–0.047	–0.151
IBER	0.439	–0.056	0.894	0.824	0.711	–0.237	0.790	–0.066
ORR	0.429	–0.061	0.824	0.889	0.732	–0.062	0.552	0.163
PRBHA	0.864	–0.409	0.711	0.732	0.511	0.183	0.509	0.455
RDDMP	0.217	–0.232	–0.237	–0.062	0.183	0.81	–0.480	0.459
RLCMD	0.289	-0,047	0.790	0.552	0.509	–0.480	0.874	–0.232
SR	0.428	–0.151	–0.066	0.163	0.455	0.459	–0.232	0.774

[a]FGBBR: failure to gain blockchain benefits risk.

[b]IABHA: adoption intention of blockchain-based health care apps.

[c]IBER: immature blockchain ecosystem risk.

[d]ORR: organizational readiness risk.

[e]PRBHA: perceived risk of blockchain-based health care apps.

[f]RDDMP: risk of disclosing data and medical practices.

[g]RLCMD: risk of losing control of medical data.

[h]SR: security risk.

Structural Model

As the reliability and validity of the variables are established, we examine the structural model to investigate the correlations between the theoretical constructs and test the effect of the moderating variable [93]. The results of the structural model are shown in Table 5 (a summary of the results of hypothesis testing) and Figure 2 (PLS results of the structural model). Before assessing the structural model, we first assess the collinearity issue [94] and the structural model's predictive accuracy [95]. Table 3 shows that all variance inflation factor values of the items are less than 5, indicating that collinearity is not an issue among the constructs [91]. The measure most often used to assess a structural model's predictive accuracy in the PLS-SEM method is the R^2 coefficient of determination. Moreover, a model is judged as interesting if it has a sufficiently large R^2 (>0.25) [95]. The results in Table 6 show that 3 constructs (FGBBR: 0.199, ORR: 0.677, and PRBHA [perceived risk of blockchain-based health care apps]: 0.958) are explained at a substantial level and one (IABHA [adoption intention of blockchain-based health care apps]: 0.163) is explained at a weak level [96,97]. Based on the R^2 values, we can state that the model's predictive accuracy is moderate.

XSL•FO
RenderX

Table . Structural model results and hypothesis testing.

Hypothesis	Path coefficient	*t* statistics	*P* values	Inference
H_{1a}: IBER[a] → PRBHA[b]	0.205	1.692	.05	Supported
H_{1b}: IBER → ORR[c]	0.824	34.743	.00	Supported
H_{1c}: IBER → FGBBR[d]	0.265	2.339	.01	Supported
H_2: SR[e] → PRBHA	0.201	1.930	.03	Supported
H_3: FGBBR → PRBHA	0.503	6.129	.00	Supported
H_{4a}: ORR → PRBHA	0.171	2.011	.02	Supported
H_{4b}: ORR → FGBBR	0.211	1.687	.05	Supported
H_5: ECBS[f] x ORR → PRBHA	−0.085	2.133	.02	Supported
H_6: PRBHA → IABHA[g]	−0.409	1.470	.07	Not supported
H_7: RDDMP[h] → PRBHA	0.121	1.021	.15	Not supported
H_8: RLCMD[i] → PRBHA	0.167	2.068	.02	Supported

[a]IBER: immature blockchain ecosystem risk.
[b]PRBHA: perceived risk of blockchain-based health care apps.
[c]ORR: organizational readiness risk.
[d]FGBBR: failure to gain blockchain benefits risk.
[e]SR: security risk.
[f]ECBS: existence of cloud-based blockchain solutions.
[g]IABHA: adoption intention of blockchain-based health care apps.
[h]RDDMP: risk of disclosing data and medical practices.
[i]RLCMD: risk of losing control of medical data.

XSL•FO
RenderX

Figure 2. Structural model result. ECBS: existence of cloud-based blockchain solutions; FGBBR: failure to gain blockchain benefits risk; IABHA: adoption intention of blockchain-based health care apps; IBER: immature blockchain ecosystem risk; ORR: organizational readiness risk; PRBHA: perceived risk of blockchain-based health care apps; RDDMP: risk of disclosing data and medical practices; RLCMD: risk of losing control of medical data; SR: security risk.

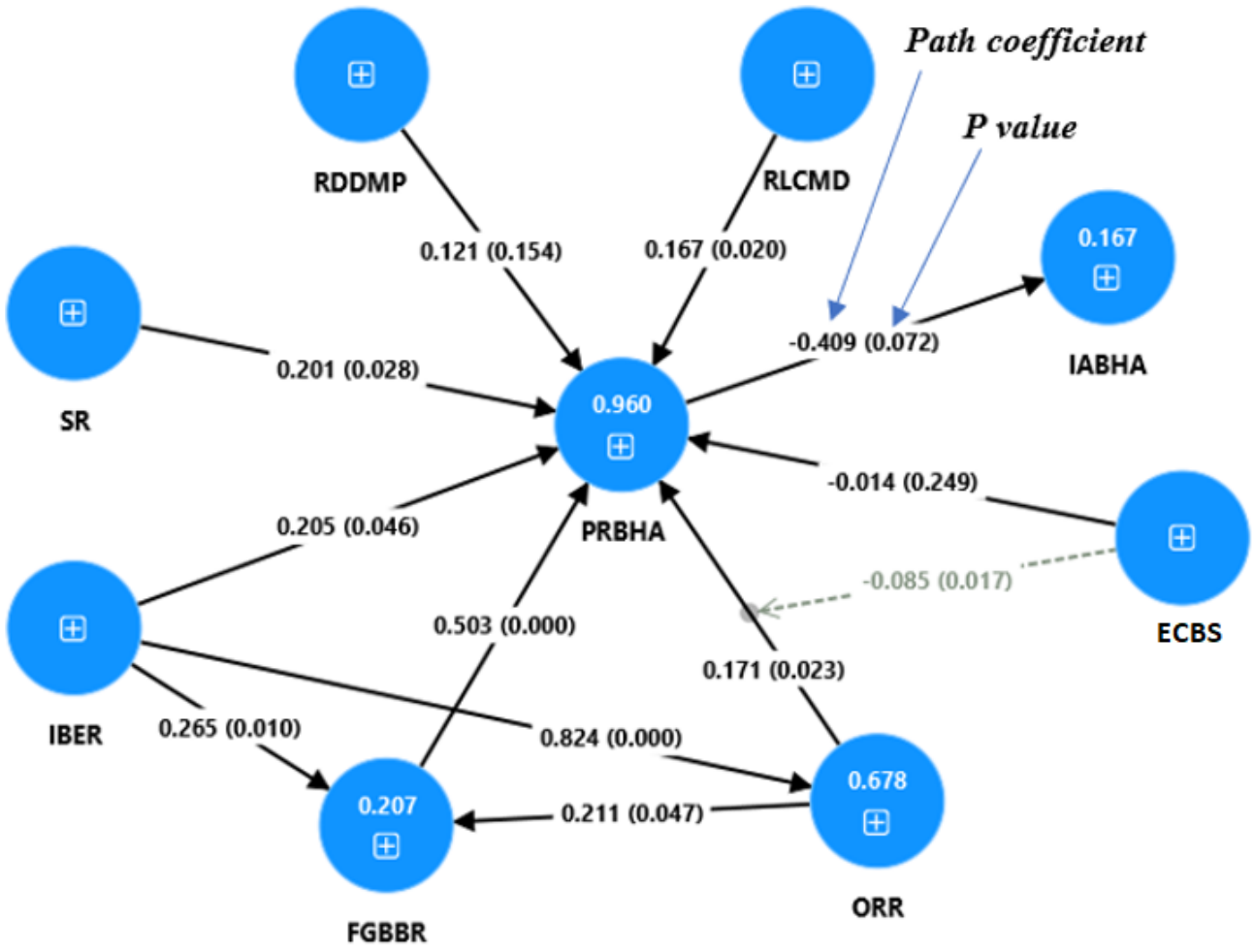

Table . R^2 values.

Constructs	R^2	R^2 adjusted	Results
FGBBR[a]	0.207	0.199	Substantial
IABHA[b]	0.167	0.163	Weak
ORR[c]	0.678	0.677	Substantial
PRBHA[d]	0.960	0.958	Substantial

[a]FGBBR: failure to gain blockchain benefits risk.

[b]IABHA: adoption intention of blockchain-based health care apps.

[c]ORR: organizational readiness risk.

[d]PRBHA: perceived risk of blockchain-based health care apps.

Next, we take a close look at the hypothesized relationships of the structural model. The hypotheses H1a, H2-H4a, and H7-H8 represent the association between PRBHA and its dimensions (IBER, SR, FGBBR, ORR, RDDMP, and RLCMD). Of the 6 dimensions of PRBHA, 5 are positively and significantly associated with PRBHA: RLCMD (β=0.167; *P*=.02), SR (β=0.201; *P*=.03), IBER (β=0.205; *P*=.05), FGBBR (β=0.503; *P*=.00), ORR (β=0.171; *P*=.02), supporting H1a, H2, H3, H4a, and H8. Contrary to what we assumed, the positive association between RDDMP and PRBHA is not significant (β=0.121; *P*=.15). Therefore, hypothesis H7 is not supported. Hypothesis H6 predicted that PRBHA is negatively and significantly associated with IABHA. Contrary to our hypothesis, the negative association between PRBHA and IABHA is not significant (β=−0.409; *P*=.07) and does not support H6.

The hypotheses H1b, H1c, and H4b represent the associations between 3 dimensions of perceived risks (IBER, FGBBR, and ORR). As hypothesized, IBER is positively and significantly

XSL•FO
RenderX

associated with FGBBR (β=0.265; *P*=.01), supporting H1c. Then, IBER is positively and significantly associated with ORR (β=0.824; *P=.00*), supporting H1b. Finally, ORR is positively and significantly associated with FGBBR (β=0.211; *P*=.05), supporting H4b.

Finally, hypothesis H5 predicts that ECBS moderates the association between ORR and PRBHA. This hypothesis is also supported (β=–0.085; *P*=.02), indicating that ECBS dampens the positive association between ORR and PRBHA. In other words, the positive association between ORR and PRBHA is weak when cloud-based blockchain solutions are present.

Interpretation of the Impact of the Control Variable on the Model Estimation

Tables 7 and 8 present model estimation results without the control variable and with the control variable. Analysis of these results leads to 2 observations. The first observation concerns the association between the control variable (country) and the dependent variable (IABHA). Table 8 shows that this association is not significant (*P*=.34). This suggests that whether a health care facility is based in Canada or Africa has no significant effect on that facility's intention to adopt a BBHA. The second observation concerns the effect of the control variable on the associations between the variables in the model. A comparison of the results in Tables 7 and 8 shows that only one association is affected by the control variable. Unlike the model without the control variable, the association between PRBHA and IABHA becomes significant in the model with the control variable (*P*=.02). This means that the perception of the risks of BBHAs does not affect a health care facility's intention to adopt these apps in the same way, depending on whether the facility is in Canada or Africa.

Table . Estimation of the model without a control variable.

Associations between variables	Without control variable			
	Original sample (O)	Sample mean (SD)	*t* statistics (O/SD)	*P* values
RDDMP[a] → PRBHA[b]	0.121	0.125 (0.119)	1.021	.15
ECBS[c] → PRBHA	−0.014	−0.014 (0.021)	0.679	.25
PRBHA → IABHA[d]	−0.409	−0.339 (0.278)	1.470	.07

[a]RDDMP: risk of disclosing data and medical practices.

[b]PRBHA: perceived risk of blockchain-based health care apps.

[c]ECBS: existence of cloud-based blockchain solutions.

[d]IABHA: adoption intention of blockchain-based health care apps.

Table . Estimation of the model with control variable.

Associations between variables	With control variable			
	Original sample (O)	Sample mean (SD)	*t* statistics (O/SD)	*P* values
RDDMP[a] → PRBHA[b]	0.121	0.144 (0.109)	1.116	.13
ECBS[c] → PRBHA	−0.014	−0.008 (0.020)	0.703	.24
PRBHA → IABHA[d]	−0.405	−0.436 (0.201)	2.012	.02
Country → IABHA	−0.133	−0.147 (0.319)	0.418	.34

[a]RDDMP: risk of disclosing data and medical practices.

[b]PRBHA: perceived risk of blockchain-based health care apps.

[c]ECBS: existence of cloud-based blockchain solutions.

[d]IABHA: adoption intention of blockchain-based health care apps.

Discussion

Principal Findings

Our study first revealed that the risks of security, failure to obtain blockchain benefits, organizational unpreparedness, disclosure of medical data and practices, and loss of control over medical data increase health care organizations' risk perceptions regarding BBHAs, reducing their intention to adopt these apps. These findings align with research by Glover and Benbasat [35], which emphasized that the SR of personal data and the failure to gain benefits of a product purchased on the web play an important role in buyers' overall risk perception regarding electronic transactions. Moreover, our study revealed that the RDDMP does not play a significant role in the overall risk perception of BBHAs. This result aligns with research that views data disclosure via blockchain not as a risk but as an advantage [98]. This is the case with the study by Dubovitskaya et al [98], in which the authors mention that disclosing data via blockchain in biomedical research allows pharmaceutical companies to disclose information about the production process of drugs. This will enable patients and regulatory bodies to know the context in which certain medical products are produced before they are certified or approved.

XSL•FO
RenderX

The second key finding of this study concerns the reciprocal association between 3 perceived risks of BBHAs: IBER, ORR, and FGBBR. The positive association between IBER and ORR indicates that an underdeveloped blockchain ecosystem—characterized by limited expertise in development, governance, and regulatory clarity—heightens the risk of organizational unpreparedness for adopting BBHAs. Additionally, the association between IBER and FGBBR suggests that ecosystem immaturity reduces the likelihood of realizing expected benefits, such as secure medical data sharing. Similarly, ORR is positively associated with FGBBR, as inadequate infrastructure and insufficiently trained personnel compromise the effective use of these technologies. For example, poor cryptographic key management by untrained users can introduce security vulnerabilities in blockchain-based apps [3], thereby preventing organizations from achieving the intended data protection and operational efficiencies of such apps.

The third key finding of our study concerns the association between IT professionals' perceived risk and their intention to adopt BBHAs. Our results show that organizations perceiving higher risks are less likely to adopt BBHAs, consistent with prior research on IT adoption, including Le et al [99] study on self-service technologies. Their study showed that individual users' concerns about technological risks reduce their willingness to engage with such systems. However, our study reveals regional differences in the association between perceived risk and intention to adopt BBHAs. In Canada, where health care systems are advanced and heavily funded, organizations are more skeptical due to limited performance feedback on BBHAs, leading to reduced motivation to invest. In contrast, health care organizations in developing African countries, which often have fragmented and underresourced systems, view BBHAs as opportunities to improve service delivery. Thus, perceived risks have a weaker impact on their adoption intentions. This regional variation aligns with the cross-national study by Hoover et al [100] on cultural influences in risk perception and consumer behavior. These authors analyzed the influence of perceived brand risk on purchase intention in Mexico and the United States. They found that the cultural difference between Mexico and the United States makes perceived brand risk a less important factor in purchasing behavior in Mexico than in the United States.

The fourth relevant result of our study comes from the assessment of the moderating effect of the existence of blockchain solutions offered as SaaS on the relationship between organizational unreadiness and overall risk perception regarding BBHAs. The results of our study reveal that the possibility of using BBHAs as SaaS would reduce IT professionals' perceived risk of the additional human and material resources required for their use and maintenance in health care settings. This result is in line with several studies that have analyzed the factors of cloud computing adoption in organizations [86,101,102]. In their study, Zhu et al [86] mention that the reduction of overall IT costs influences the use of cloud services. These costs include preliminary investment, upgrading, energy saving, and maintenance of cloud services [86,101].

Theoretical and Practical Contributions

Our study makes a theoretical contribution by adopting and interpreting the perceived risk theory derived from marketing in the context of blockchain adoption in the health care sector. Originally, this theory was used to explain consumers' purchasing decisions in the context of perceived product or service risks [103-105]. In information systems, perceived risk theory is used to analyze how perceived risks can influence the IT adoption decisions [29,93]. However, this theory addresses risks that apply broadly to IT as a whole, making it unsuitable for analyzing risks within specific IT categories. The results of our study highlight 2 risks specific to blockchain-based apps that are likely to influence the overall perceived risk regarding these apps and, therefore, their adoption in health care structures: RDDMP and RLCMD. Integrating these 2 risks as new constructs into the perceived risk theory contributes to its extension, making it more suitable for risk analysis in blockchain-based apps. Future research analyzing barriers to blockchain adoption in the health care sector can incorporate the 2 new risks into its theoretical model. Moreover, conducting our study in Canada and Africa allowed us to understand how contextual differences between Canada and Africa affect how risks are perceived in BBHAs and the intention to adopt them in Canadian and African health care organizations. This contributed to the generalizability of our results.

As a practical contribution, our study highlights the risks IT professionals in health care organizations perceive regarding BBHAs that may prevent their adoption in health care organizations. Knowledge of these risks would benefit the development agencies of blockchain-based apps and health care organizations [106,107]. For development agencies of blockchain-based apps, our findings will guide their advocacy discourse on the adoption of blockchain-based apps among health care organizations. These agencies can exploit these findings to formulate recommendations to mitigate risks that IT professionals perceive or educate IT professionals of health care organizations on strategies to avoid certain perceived risks [108]. In addition, development agencies of blockchain-based apps can take advantage of our findings to develop BBHAs that minimize some risks that IT professionals perceive regarding these apps. For health care organizations, knowing the risks of BBHAs brings 2 major implications. First, these organizations handle sensitive data that require better protection against security attacks [8]. Through our results, these organizations understand that despite the hype about the security of blockchain-based apps, these apps are exposed to SRs. Therefore, they must implement appropriate security policies to strengthen the security of these apps [107]. Second, our study's results inform health care organizations' decisions regarding adopting BBHAs. We believe that a good decision to adopt these apps should consider their benefits and potential risks [109]. Knowing these risks would allow health care organizations to implement strategies to mitigate them and improve the use of BBHAs.

Limitations

This study has 2 main limitations. The first limitation concerns the small number of IT professionals working in health care

XSL•FO
RenderX

structures that have already adopted BBHAs or those in the process of adopting them. In addition, our study had a low participation of IT professionals from health care structures in Canada. The low participation of IT professionals from structures that use BBHAs, structures that are in the process of adopting these apps, and structures in Canada prevents variance in the dependent variable of our study, namely the intention to adopt BBHAs. Second, our research analyzes blockchain adoption in the health care sector, which is experiencing a low adoption rate of BBHAs. All these limitations constitute obstacles to the generalization of the results of this research.

Directions for Future Research

We propose several avenues for future research. First, the conceptual model was validated using data from IT professionals in health care organizations. This choice was justified by the low knowledge of blockchain among health care professionals and patients [110,111]. It is essential to note that the successful implementation of IT in a health care organization necessitates the involvement of multiple stakeholders, including IT professionals, health care professionals, patients, health care organization managers, and health care regulators [112]. Perceptions of each stakeholder regarding the risks associated with BBHAs may influence health care organizations' intentions to adopt these apps [113]. Therefore, we recommend that other researchers conduct similar research involving many stakeholders to better assess the factors that influence the intention of health care organizations to adopt BBHAs.

Second, our research focused primarily on the factors that may hinder health care organizations' intention to adopt blockchain-based apps. Our choice is explained by the low adoption of blockchain-based apps in health care organizations [76]. Therefore, the data for our research were primarily collected from health care organizations that have not yet adopted a blockchain-based app and those that are in the process of doing so. Future research may analyze the adoption challenges of these apps by collecting data from health care organizations that have already adopted these apps. Thus, health care organizations interested in adopting blockchain-based apps will be better informed about the challenges of adoption before implementing them. This helps reduce the perceived risks associated with blockchain-based health apps in health care organizations, as they can develop strategies to overcome the adoption challenges of these apps [114].

Conclusion

Our study sought to analyze the impact of IT professionals' perceived risk concerning the intention to adopt BBHAs in health care facilities. Drawing on perceived risk theory, we identified 6 dimensions of risk that may explain IT professionals' overall risk perception of BBHAs in health care facilities. Data collected from IT professionals in health care facilities through an online survey allowed us to evaluate the hypotheses of our conceptual model. The analysis of these data, using SmartPLS 4 software, yielded important results. The analysis of the correlations between perceived risk and its dimensions revealed that the latter are all positively associated with perceived risk. However, the RDDMP does not contribute significantly to the overall perception of IT professionals' risks regarding using BBHAs. Then, our results revealed that the significance of the negative association between the general perception of risks of BBHAs and the intention to adopt these apps varies depending on whether the health care organization is in Africa or Canada. This association is nonsignificant in African health care organizations, while it is significant in Canadian health care organizations. Finally, the results of our study revealed that the existence of blockchain solutions in the form of SaaS reduces the overall perception of risks of IT professionals regarding BBHA, given the lack of resources (human and technical) necessary for their use in healthcare structures.

Our study makes the following contributions. First, we highlight the types of risks likely to prevent the adoption of BBHAs in health care organizations despite their interest in adopting these apps.

Second, we show that RDDMP and RLCMD contribute to the overall risk perception of BBHAs. These risks are inherent to the characteristics of blockchain, and the fact that they are not considered in the perceived risk theory makes it unsuitable for risk analysis in blockchain-based apps. The perceived risk theory includes risks that are general to any IT and the integration of these new risks contributes to its extension by making it more suitable for the analysis of risks in blockchain-based apps. Finally, we show that the difference in context between Canada and Africa makes the overall perception of risk of BBHAs differently associated with the intention to adopt these apps between Canadian and African health care organizations. Our findings are beneficial to governments, health care organizations, and blockchain development agencies, which can implement strategies to mitigate some of the risks perceived by IT professionals related to BBHAs.

Acknowledgments

No generative artificial intelligence (AI) tools were used at any stage in the preparation of this manuscript.

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Measurement of items.
[DOCX File, 18 KB - jmir_v28i1e77933_app1.docx]

References

1. Cho S, Lee K(, Cheong A, No WG, Vasarhelyi MA. Chain of values: examining the economic impacts of blockchain on the value-added tax system. J Manag Inf Syst 2021 Apr 3;38(2):288-313. [doi: 10.1080/07421222.2021.1912912]
2. Sarker S, Henningsson S, Jensen T, Hedman J. The use of blockchain as a resource for combating corruption in global shipping: an interpretive case study. J Manag Inf Syst 2021 Apr 3;38(2):338-373. [doi: 10.1080/07421222.2021.1912919]
3. McGhin T, Choo KKR, Liu CZC, He DB. Blockchain in healthcare applications: research challenges and opportunities. J Netw Comput Appl 2019 Jun;135:62-75. [doi: 10.1016/j.jnca.2019.02.027]
4. Tandon A, Dhir A, Islam A, Mäntymäki M. Blockchain in healthcare: a systematic literature review, synthesizing framework and future research agenda. Comput Ind 2020 Nov;122:103290. [doi: 10.1016/j.compind.2020.103290]
5. Zaabar B, Cheikhrouhou O, Jamil F, Ammi M, Abid M. HealthBlock: a secure blockchain-based healthcare data management system. Comput Netw 2021 Dec;200:108500. [doi: 10.1016/j.comnet.2021.108500]
6. Cadoret D, Kailas T, Velmovitsky P, Morita P, Igboeli O. Proposed implementation of blockchain in British Columbia's health care data management. J Med Internet Res 2020 Oct 23;22(10):e20897. [doi: 10.2196/20897] [Medline: 33095183]
7. Palas M, Bunduchi R. Exploring interpretations of blockchain's value in healthcare: a multi-stakeholder approach. Inf Technol People 2020 Jun 24;34(2):453-495. [doi: 10.1108/ITP-01-2019-0008]
8. Esmaeilzadeh P, Mirzaei T. The potential of blockchain technology for health information exchange: experimental study from patients' perspectives. J Med Internet Res 2019 Jun 20;21(6):e14184. [doi: 10.2196/14184] [Medline: 31223119]
9. Hylock RH, Zeng X. A blockchain framework for patient-centered health records and exchange (HealthChain): evaluation and proof-of-concept study. J Med Internet Res 2019 Aug 31;21(8):e13592. [doi: 10.2196/13592] [Medline: 31471959]
10. How blockchain and AI improve healthcare supply chains. Medium. 2022. URL: https://blog.medifakt.com/how-blockchain-and-ai-improve-healthcare-supply-chains-5d856d6e46d5 [accessed 2026-03-06]
11. Adams HS. Top 10: blockchain platforms in healthcare. Healthcare Digital. 2024. URL: https://healthcare-digital.com/technology-and-ai/top-10-blockchain-platforms-in-healthcare [accessed 2026-03-06]
12. Embleema is a technology platform for registries and real-world studies. Embleema. 2025. URL: https://www.embleema.com/aboutus [accessed 2026-03-06]
13. Dutta S. Blockchain for healthcare: benefits, use cases & real-world examples. Turing. 2025. URL: https://www.turing.com/resources/blockchain-for-healthcare [accessed 2026-03-06]
14. Sanka AI, Irfan M, Huang I, Cheung RCC. A survey of breakthrough in blockchain technology: adoptions, applications, challenges and future research. Comput Commun 2021 Mar;169:179-201. [doi: 10.1016/j.comcom.2020.12.028]
15. Pirtle C, Ehrenfeld J. Blockchain for healthcare: the next generation of medical records? J Med Syst 2018 Aug 10;42(9):172. [doi: 10.1007/s10916-018-1025-3] [Medline: 30097733]
16. Chen Y, Ding S, Xu Z, Zheng HD, Yang SL. Blockchain-based medical records secure storage and medical service framework. J Med Syst 2019 Jan;43(1). [doi: 10.1007/s10916-018-1121-4]
17. Vardhini B, Dass SN, Sahana R, Chinnaiyan R. A blockchain based electronic medical health records framework using smart contracts. Presented at: 2021 International Conference on Computer Communication and Informatics (ICCCI); Jan 27-29, 2021. [doi: 10.1109/ICCCI50826.2021.9402689]
18. Jayaraman R, Salah K, King N. Improving opportunities in healthcare supply chain processes via the internet of things and blockchain technology. Int J Healthc Inf Syst Inform 2019;14(2):49-65. [doi: 10.4018/IJHISI.2019040104]
19. Liang TP, Kohli R, Huang HC, Li ZL. What drives the adoption of the blockchain technology? A fit-viability perspective. J Manag Inf Syst 2021 Apr 3;38(2):314-337. [doi: 10.1080/07421222.2021.1912915]
20. Clohessy T, Acton T. Investigating the influence of organizational factors on blockchain adoption. Ind Manang Data Syst 2019 Aug 12;119(7):1457-1491. [doi: 10.1108/IMDS-08-2018-0365]
21. Folkinshteyn D, Lennon M. Braving Bitcoin: a technology acceptance model (TAM) analysis. J Inf Technol Case Appl Res 2016 Oct;18(4):220-249. [doi: 10.1080/15228053.2016.1275242]
22. Janssen M, Weerakkody V, Ismagilova E, Sivarajah U, Irani Z. A framework for analysing blockchain technology adoption: integrating institutional, market and technical factors. Int J Inf Manage 2020 Feb;50:302-309. [doi: 10.1016/j.ijinfomgt.2019.08.012]
23. Rashideh W. Blockchain technology framework: current and future perspectives for the tourism industry. Tour Manag 2020 Oct;80:104125. [doi: 10.1016/j.tourman.2020.104125]
24. Yoo K, Bae K, Park E, Yang T. Understanding the diffusion and adoption of Bitcoin transaction services: the integrated approach. Telemat Inform 2020 Oct;53:101302. [doi: 10.1016/j.tele.2019.101302]

XSL•FO
RenderX

25. Maythu Y, Kwok AOJ, Teh PL. Blockchain technology diffusion in tourism: evidence from early enterprise adopters and innovators. Heliyon 2024 Jan 30;10(2):e24675. [doi: 10.1016/j.heliyon.2024.e24675] [Medline: 38304774]
26. Agbo CC, Mahmoud QH, Eklund JM. Blockchain technology in healthcare: a systematic review. Healthcare (Basel) 2019 Apr 4;7(2):56. [doi: 10.3390/healthcare7020056] [Medline: 30987333]
27. Bauer RA. Consumer behaviour as risk taking. In: Robert SH, editor. Dynamic Marketing for a Changing World: Proceedings of the 43rd National Conference of the American Marketing Association: American Marketing Association; 1960:389-398.
28. Schlegel D, Uenal Y. A perceived risk perspective on narrow artificial intelligence. Presented at: 25th Pacific Asia Conference on Information Systems (PACIS); Jun 20-24, 2021 URL: https://aisel.aisnet.org/pacis2021/44 [accessed 2026-03-29]
29. Martins C, Oliveira T, Popovič A. Understanding the Internet banking adoption: a unified theory of acceptance and use of technology and perceived risk application. Int J Inf Manage 2014 Feb;34(1):1-13. [doi: 10.1016/j.ijinfomgt.2013.06.002]
30. Gefen D, Karahanna E. Trust and TAM in online shopping: an integrated model. MIS Q 2003;27(1):51-90. [doi: 10.2307/30036519]
31. Khedmatgozar HR. The impact of perceived risks on internet banking adoption in Iran: a longitudinal survey. Electron Commer Res 2021 Mar;21(1):147-167. [doi: 10.1007/s10660-021-09475-y]
32. Verhagen T, Meents S, Tan YH. Perceived risk and trust associated with purchasing at electronic marketplaces. Eur J Inf Syst 2006 Dec;15(6):542-555. [doi: 10.1057/palgrave.ejis.3000644]
33. Littler D, Melanthiou D. Consumer perceptions of risk and uncertainty and the implications for behaviour towards innovative retail services: the case of Internet Banking. J Retailing Consum Serv 2006 Nov;13(6):431-443. [doi: 10.1016/j.jretconser.2006.02.006]
34. Featherman MS, Pavlou PA. Predicting e-services adoption: a perceived risk facets perspective. Int J Hum Comput Stud 2003 Oct;59(4):451-474. [doi: 10.1016/S1071-5819(03)00111-3]
35. Glover S, Benbasat I. A comprehensive model of perceived risk of e-commerce transactions. Int J Electron Commer 2010 Dec;15(2):47-78. [doi: 10.2753/JEC1086-4415150202]
36. Pavlou PA, Gefen D. Building effective online marketplaces with institution-based trust. Inf Syst Res 2004 Mar;15(1):37-59. [doi: 10.1287/isre.1040.0015]
37. Pavlou PA. Consumer acceptance of electronic commerce: integrating trust and risk with the technology acceptance model. Int J Electron Commer 2003 Apr;7(3):101-134. [doi: 10.1080/10864415.2003.11044275]
38. Spiekermann S, Paraschiv C. Motivating human–agent interaction: transferring insights from behavioral marketing to interface design. Electron Commer Res 2002 Jul;2(3):255-285. [doi: 10.1023/A:1016062632182]
39. Chiu C, Wang ETG, Fang Y, Huang H. Understanding customers' repeat purchase intentions in B2C e‐commerce: the roles of utilitarian value, hedonic value and perceived risk. Inf Syst J 2014 Jan;24(1):85-114. [doi: 10.1111/j.1365-2575.2012.00407.x]
40. Cocosila M, Archer N, Yuan Y. Early investigation of new information technology acceptance: a perceived risk - motivation model. Commun Assoc Inf Syst 2009;25. [doi: 10.17705/1CAIS.02530]
41. Erol I, Ar IM, Ozdemir AI, et al. Assessing the feasibility of blockchain technology in industries: evidence from Turkey. J Enterp Inf Manag 2021 Apr 12;34(3):746-769. [doi: 10.1108/JEIM-09-2019-0309]
42. Akella GK, Wibowo S, Grandhi S, Mubarak S. A systematic review of blockchain technology adoption barriers and enablers for smart and sustainable agriculture. Big Data Cogn Comput 2023;7(2):86. [doi: 10.3390/bdcc7020086]
43. Kaushal RK, Kumar N, Panda SN, Kukreja V. Immutable smart contracts on blockchain technology: its benefits and barriers. Presented at: 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO); Sep 3-4, 2021. [doi: 10.1109/ICRITO51393.2021.9596538]
44. White GRT. Future applications of blockchain in business and management: a Delphi study. Strat Change 2017 Sep;26(5):439-451. [doi: 10.1002/jsc.2144]
45. Komulainen R, Nätti S. Barriers to blockchain adoption: empirical observations from securities services value network. J Bus Res 2023 Apr;159:113714. [doi: 10.1016/j.jbusres.2023.113714]
46. Raut RD, Priyadarshinee P, Gardas BB, Jha MK. Analyzing the factors influencing cloud computing adoption using three stage hybrid SEM-ANN-ISM (SEANIS) approach. Technol Forecast Soc Change 2018 Sep;134:98-123. [doi: 10.1016/j.techfore.2018.05.020]
47. Griggs KN, Ossipova O, Kohlios CP, Baccarini AN, Howson EA, Hayajneh T. Healthcare blockchain system using smart contracts for secure automated remote patient monitoring. J Med Syst 2018 Jun 6;42(7):130. [doi: 10.1007/s10916-018-0982-x] [Medline: 29876661]
48. Khezr S, Moniruzzaman M, Yassine A, Benlamri R. Blockchain technology in healthcare: a comprehensive review and directions for future research. Appl Sci (Basel) 2019;9(9):1736. [doi: 10.3390/app9091736]
49. Zamani E, He Y, Phillips M. On the security risks of the blockchain. J Comput Inform Syst 2020 Nov 1;60(6):495-506. [doi: 10.1080/08874417.2018.1538709]
50. Radanović I, Likić R. Opportunities for use of blockchain technology in medicine. Appl Health Econ Health Policy 2018 Oct;16(5):583-590. [doi: 10.1007/s40258-018-0412-8] [Medline: 30022440]

XSL•FO
RenderX

51. Hussien HM, Yasin SM, Udzir SNI, Zaidan AA, Zaidan BB. A systematic review for enabling of develop a blockchain technology in healthcare application: taxonomy, substantially analysis, motivations, challenges, recommendations and future direction. J Med Syst 2019 Sep 14;43(10):1-35. [doi: 10.1007/s10916-019-1445-8] [Medline: 31522262]
52. Gaur N. Blockchain challenges in adoption. Manag Financ 2019 Aug 12;46(6):849-858. [doi: 10.1108/MF-07-2019-0328]
53. Ragowsky A, Licker PS, Gefen D. Organizational IT Maturity (OITM): a measure of organizational readiness and effectiveness to obtain value from its information technology. Inform Syst Manage 2012 Mar;29(2):148-160. [doi: 10.1080/10580530.2012.662104]
54. Mouzakitis S, Askounis D. A knowledge-based framework for measuring organizational readiness for the adoption of B2B integration systems. Inf Syst Manag 2010 Jan;27(3):253-266. [doi: 10.1080/10580530.2010.493842]
55. Hosseini L, Tang S, Mookerjee V, Sriskandarajah C. A Switch in Time Saves the Dime: a model to reduce rental cost in cloud computing. Inf Systems Res 2020 Sep;31(3):753-775. [doi: 10.1287/isre.2019.0912]
56. Tripathi S, Mishra V. Determinants of cloud computing adoption: a comparative study. Pac Asia J Assoc Inf Syst 2019;11(3):28-54. [doi: 10.17705/1pais.11303] [Medline: 2500368009]
57. Rogers EM. Diffusion of Innovations, 5th edition: The Free Press; 2003.
58. Hanafizadeh P, Khedmatgozar HR. The mediating role of the dimensions of the perceived risk in the effect of customers' awareness on the adoption of Internet banking in Iran. Electron Commer Res 2012 May;12(2):151-175. [doi: 10.1007/s10660-012-9090-z]
59. Lee MC. Factors influencing the adoption of internet banking: an integration of TAM and TPB with perceived risk and perceived benefit. Electron Commer Res Appl 2009 May;8(3):130-141. [doi: 10.1016/j.elerap.2008.11.006]
60. Govindan K, Nasr AK, Saeed Heidary M, Nosrati-Abarghooee S, Mina H. Prioritizing adoption barriers of platforms based on blockchain technology from balanced scorecard perspectives in healthcare industry: a structural approach. Int J Prod Res 2023 Jun 3;61(11):3512-3526. [doi: 10.1080/00207543.2021.2013560]
61. Mittal S, Ghosh M. A novel two-level secure access control approach for blockchain platform in healthcare. Int J Inf Secur 2023 Aug;22(4):799-817. [doi: 10.1007/s10207-023-00664-4]
62. Porsdam Mann S, Savulescu J, Ravaud P, Benchoufi M. Blockchain, consent and prosent for medical research. J Med Ethics 2021 Apr 13;47(4):244-250. [doi: 10.1136/medethics-2019-105963] [Medline: 32366703]
63. Wang X, Cheng Z. Cross-sectional studies: strengths, weaknesses, and recommendations. Chest 2020 Jul;158(1S):S65-S71. [doi: 10.1016/j.chest.2020.03.012] [Medline: 32658654]
64. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Epidemiology (Sunnyvale) 2007 Nov;18(6):800-804. [doi: 10.1097/EDE.0b013e3181577654] [Medline: 18049194]
65. Mitchell VW. Consumer perceived risk: conceptualisations and models. Eur J Mark 1999 Feb 1;33(1/2):163-195. [doi: 10.1108/03090569910249229]
66. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. Manage Sci 1989 Aug;35(8):982-1003. [doi: 10.1287/mnsc.35.8.982]
67. Venkatesh V, Thong JYL, Xu X. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS Q 2012 Mar 1;36(1):157-178. [doi: 10.2307/41410412]
68. Teo HH, Wei KK, Benbasat I. Predicting intention to adopt interorganizational linkages: an institutional perspective. MIS Q 2003 Mar 1;27(1):19-49. [doi: 10.2307/30036518]
69. Garg P, Gupta B, Chauhan AK, Sivarajah U, Gupta S, Modgil S. Measuring the perceived benefits of implementing blockchain technology in the banking sector. Technol Forecast Soc Change 2021 Feb;163:120407. [doi: 10.1016/j.techfore.2020.120407]
70. Kouhizadeh M, Saberi S, Sarkis J. Blockchain technology and the sustainable supply chain: theoretically exploring adoption barriers. Int J Prod Econ 2021 Jan;231(2021):107831. [doi: 10.1016/j.ijpe.2020.107831]
71. Shah MH, Peikari HR, Yasin NM. The determinants of individuals' perceived e-security: evidence from Malaysia. Int J Inf Manage 2014 Feb;34(1):48-57. [doi: 10.1016/j.ijinfomgt.2013.10.001]
72. Bélanger F, Carter L. Trust and risk in e-government adoption. J Strateg Inf Syst 2008 Jun;17(2):165-176. [doi: 10.1016/j.jsis.2007.12.002]
73. Hartono E, Holsapple CW, Kim KY, Na KS, Simpson JT. Measuring perceived security in B2C electronic commerce website usage: a respecification and validation. Decis Support Syst 2014 Jun;62(2014):11-21. [doi: 10.1016/j.dss.2014.02.006]
74. Iacovou CL, Benbasat I, Dexter AS. Electronic data interchange and small organizations: adoption and impact of technology. MIS Q 1995 Dec 1;19(4):465-485. [doi: 10.2307/249629]
75. Santos FM, Eisenhardt KM. Organizational boundaries and theories of organization. Organ Sci 2005 Oct;16(5):491-508. [doi: 10.1287/orsc.1050.0152]
76. Alzahrani S, Daim T, Choo KKR. Assessment of the blockchain technology adoption for the management of the electronic health record systems. IEEE Trans Eng Manage 2023;70(8):2846-2863. [doi: 10.1109/TEM.2022.3158185]
77. Kyriakidou V, Michalakelis C, Sphicopoulos T. Assessment of information and communications technology maturity level. Telecommun Policy 2013 Feb;37(1):48-62. [doi: 10.1016/j.telpol.2012.08.001]

XSL•FO
RenderX

78. Beck JT, Rahinel R, Bleier A. Company worth keeping: personal control and preferences for brand leaders. J Consum Res 2020 Feb 1;46(5):871-886. [doi: 10.1093/jcr/ucz040]
79. Nogueira M, Dias F, Santos V. Sustainable mobility choices: exploring the impact of consumers' values, attitudes, perceived behavioural control and subjective norms on the likelihood to choose sustainable mobility options. J Consum Behav 2023 Mar;22(2):511-528 [FREE Full text] [doi: 10.1002/cb.2144]
80. Braga de Oliveira Friaes LC, Poleto T, Negreiros Clemente TR, Ramos CC, de Carvalho VDH, Nepomuceno TCC. Relational governance in the ITO: the role of self-determination theory and the mediating effect of control perceived. Acad Rev Latinaom Ad 2023 Jun 29;36(2):177-198. [doi: 10.1108/ARLA-10-2022-0182]
81. Malhotra NK, Kim SS, Agarwal J. Internet Users' Information Privacy Concerns (IUIPC): the construct, the scale, and a causal model. Inf Syst Res 2004 Dec;15(4):336-355. [doi: 10.1287/isre.1040.0032]
82. Bansal G, Zahedi F", Gefen D. The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online. Decis Support Syst 2010 May;49(2):138-150. [doi: 10.1016/j.dss.2010.01.010]
83. Bansal G, Zahedi FM, Gefen D. Do context and personality matter? Trust and privacy concerns in disclosing private information online. Inf Manag 2016 Jan;53(1):1-21. [doi: 10.1016/j.im.2015.08.001]
84. Merriam SB, Tisdell EJ. Qualitative Research: A Guide to Design and Implementation, 4th edition: San Francisco, CA: Jossey-Bass; 2016:368.
85. Soper D, Soper D. A-priori sample size calculator for structural equation models. Free Statistics Calculators. 2014. URL: https://www.danielsoper.com/statcalc/calculator.aspx?id=89 [accessed 2024-03-17]
86. Zhu K, Kraemer KL, Xu S. The process of innovation assimilation by firms in different countries: a technology diffusion perspective on e-business. Manage Sci 2006 Oct;52(10):1557-1576. [doi: 10.1287/mnsc.1050.0487]
87. Zhu K, Kraemer K, Xu S. Electronic business adoption by European firms: a cross-country assessment of the facilitators and inhibitors. Eur J Inf Syst 2003 Dec;12(4):251-268. [doi: 10.1057/palgrave.ejis.3000475]
88. Podsakoff PM, MacKenzie SB, Lee JY, Podsakoff NP. Common method biases in behavioral research: a critical review of the literature and recommended remedies. J Appl Psychol 2003 Oct;88(5):879-903. [doi: 10.1037/0021-9010.88.5.879] [Medline: 14516251]
89. Guan Z, Hou F, Li B, Phang CW, Chong AYL. What influences the purchase of virtual gifts in live streaming in China? A cultural context‐sensitive model. Inf Syst J 2022 May;32(3):653-689 [FREE Full text] [doi: 10.1111/isj.12367]
90. Hair JF, Sarstedt M, Ringle CM, Mena JA. An assessment of the use of partial least squares structural equation modeling in marketing research. J Acad Mark Sci 2012 May;40(3):414-433. [doi: 10.1007/s11747-011-0261-6]
91. Hair JF, Howard MC, Nitzl C. Assessing measurement model quality in PLS-SEM using confirmatory composite analysis. J Bus Res 2020 Mar;109:101-110. [doi: 10.1016/j.jbusres.2019.11.069]
92. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. J Mark Res 1981 Feb;18(1):39-50. [doi: 10.1177/002224378101800104]
93. Im I, Kim Y, Han HJ. The effects of perceived risk and technology type on users' acceptance of technologies. Inf Manag 2008 Jan;45(1):1-9. [doi: 10.1016/j.im.2007.03.005]
94. Becker JM, Ringle CM, Sarstedt M, Völckner F. How collinearity affects mixture regression results. Mark Lett 2015 Dec;26(4):643-659. [doi: 10.1007/s11002-014-9299-9]
95. Hair JF, Astrachan CB, Moisescu OI, et al. Executing and interpreting applications of PLS-SEM: updates for family business researchers. J Fam Bus Strateg 2021 Sep;12(3):100392. [doi: 10.1016/j.jfbs.2020.100392]
96. Bawack RE, Wamba SF, Carillo KDA. Exploring the role of personality, trust, and privacy in customer experience performance during voice shopping: evidence from SEM and fuzzy set qualitative comparative analysis. Int J Inf Manage 2021 Jun;58:102309. [doi: 10.1016/j.ijinfomgt.2021.102309]
97. Jr. JFH, Matthews LM, Matthews RL, Sarstedt M. PLS-SEM or CB-SEM: updated guidelines on which method to use. Int J Multivar Data Anal 2017;1(2):107. [doi: 10.1504/IJMDA.2017.087624]
98. Dubovitskaya A, Novotny P, Xu Z, Wang F. Applications of blockchain technology for data-sharing in oncology: results from a systematic literature review. Oncology (Williston Park, NY) 2020;98(6):403-411. [doi: 10.1159/000504325] [Medline: 31794967]
99. Le NT, Rao Hill S, Troshani I. Perceived control and perceived risk in self-service technology recovery. J Comput Inf Syst 2022 Jan 2;62(1):164-173. [doi: 10.1080/08874417.2020.1756533]
100. Hoover RJ, Green RT, Saegert J. A cross-national study of perceived risk: does perceived risk have the same effect on consumers in a foreign country as in the United States. J Mark 1978;42(3):102-108. [doi: 10.1177/002224297804200318]
101. Ali O, Shrestha A, Osmanaj V, Muhammed S. Cloud computing technology adoption: an evaluation of key factors in local governments. Inf Technol People 2020 Apr 28;34(2):666-703. [doi: 10.1108/ITP-03-2019-0119]
102. Cervone HF. An overview of virtual and cloud computing. OCLC Syst Serv 2010 Aug 24;26(3):162-165. [doi: 10.1108/10650751011073607]
103. Simcock P, Sudbury L, Wright G. Age, perceived risk and satisfaction in consumer decision making: a review and extension. J Mark Manag 2006 Apr;22(3-4):355-377. [doi: 10.1362/026725706776861163]
104. Phamthi VA, Nagy Á, Ngo TM. The influence of perceived risk on purchase intention in e‐commerce—Systematic review and research agenda. Int J Consum Stud 2024 Jul;48(4):e13067 [FREE Full text] [doi: 10.1111/ijcs.13067]

XSL•FO
RenderX

105. Kim LH, Qu H, Kim DJ. A study of perceived risk and risk reduction of purchasing air - tickets online. J Travel Tour Mark 2009 May;26(3):203-224. [doi: 10.1080/10548400902925031]
106. Abdelwahed IM, Ramadan N, Hefny HA. Cybersecurity risks of blockchain technology. Int J Comput Appl 2020;177(42):8-14. [doi: 10.5120/ijca2020919922]
107. Behzadi L, Joseph J. Blockchain Security Considerations: CRC Press; 2024:73-92. [doi: 10.1201/9781003462033-6]
108. Bordeianu AA, Popescu DE. Blockchain variables and possible attacks: a technical survey. Computers 2025;14(12):567. [doi: 10.3390/computers14120567]
109. Drljevic N, Aranda DA, Stantchev V. Perspectives on risks and standards that affect the requirements engineering of blockchain technology. Comput Stand Inter 2020 Mar;69:103409. [doi: 10.1016/j.csi.2019.103409]
110. Govindarajan UH, Narang G, Singh DK, Yadav VS. Blockchain technologies adoption in healthcare: overcoming barriers amid the hype cycle to enhance patient care. Technol Forecast Soc Change 2025 Apr;213:124031. [doi: 10.1016/j.techfore.2025.124031]
111. Attaran M. Blockchain technology in healthcare: challenges and opportunities. Int J Healthc Manag 2022 Jan 2;15(1):70-83. [doi: 10.1080/20479700.2020.1843887]
112. Prashar A, Sunder M V. Blockchain barriers in hospitals: a stakeholder theoretic perspective. Benchmarking 2026 Jan 1;33(1):1-32. [doi: 10.1108/BIJ-03-2023-0185]
113. Nanda S, Nanda S. Blockchain adoption in health market: a systems thinking and modelling approach. J Asia Bus Stud 2022 Mar 23;16(2):396-405. [doi: 10.1108/JABS-11-2020-0457]
114. Onik MMH, Aich S, Yang J, Kim CS, Kim HC. Chapter 8 - blockchain in healthcare: challenges and solutions. In: Dey N, Das H, Naik B, Behera HS, editors. Big Data Analytics for Intelligent Healthcare: Academic Press; 2019:197-226. [doi: 10.1016/B978-0-12-818146-1.00008-8]

Abbreviations

AVE: average variance extracted
BBHA: blockchain-based health care apps
CR: composite reliability
ECBS: existence of cloud-based blockchain solutions
FGBBR: failure to gain blockchain benefits risk
IABHA: adoption intention of blockchain-based health care apps
IBER: immature blockchain ecosystem risk
ORR: organizational readiness risk
PLS-SEM: partial least squares structural equation modeling
PRBHA: perceived risk of blockchain-based health care apps
PRT: perceived risk theory
RDDMP: risk of disclosing data and medical practices
RLCMD: risk of losing control of medical data
SaaS: Software-as-a-Service
SR: security risk
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by S Brini; submitted 22.May.2025; peer-reviewed by I Cerovecki, S Mohanadas, TA El-Hafeez; revised version received 19.Jan.2026; accepted 20.Jan.2026; published 08.Apr.2026.

Please cite as:
Bazomanza Nzabandora W, Mellouli S, Nishant R
Resistance to Blockchain Adoption in Health Care Organizations: Evidence From a Cross-Sectional Study
J Med Internet Res 2026;28:e77933
URL: https://www.jmir.org/2026/1/e77933
doi:10.2196/77933

© Wilfried Bazomanza Nzabandora, Sehl Mellouli, Rohit Nishant. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Patient Perceptions of Blockchain-Based Health Information Exchange: User-Centered Design Study

Richard Guse[1], MSc; Shanshan Hu[2], MSc; Scott Thiebes[3], PhD; Christina Erler[4], PhD; Catherine Caridia[5], MD; Wilhelm Stork[4], Prof Dr; Martin Gersch[6], Prof Dr; Ali Sunyaev[2], Prof Dr

[1]Department of Economics and Management, Karlsruhe Institute of Technology, Kaiserstraße 89, Karlsruhe, Germany
[2]School of Computation, Information and Technology (CIT), Technical University of Munich, München, Germany
[3]Advanced Institute of Business, School of Economics and Management, Tongji University, Shanghai, China
[4]FZI Research Center for Information Technology, Karlsruhe, Germany
[5]Charité Comprehensive Cancer Center, Charité - Universitätsmedizin Berlin, Berlin, Germany
[6]Department of Information Systems, School of Business and Economics, Freie Universität Berlin, Berlin, Germany

Corresponding Author:
Richard Guse, MSc
Department of Economics and Management, Karlsruhe Institute of Technology, Kaiserstraße 89, Karlsruhe, Germany

Abstract

Background: Blockchain-based health information exchange (HIE) has received increased attention in health care research and practice over the last years. It enables the sharing of patient information across healthcare organizations, provides higher levels of data confidentiality and security, and reduces time and costs in collaborative medical decision-making. To make informed decisions on the implementation of blockchain-based HIE in practice and to fully understand the implications of its use for patient care, it is important to gain insights into patient perceptions of and interactions with blockchain-based HIE.

Objective: This study aimed to assess patient perceptions of a blockchain-based HIE mobile app to inform its iterative development.

Methods: We used a mixed methods user-centered design in 3 phases to iteratively assess patient perceptions of blockchain-based HIE: (1) structured questionnaires collecting patient requirements for blockchain-based HIE, (2) semistructured interviews evaluating mobile app mock-ups, and (3) a survey with blockchain-based HIE scenarios related to their patient care using the technology acceptance model, System Usability Scale, and open feedback. Both the semistructured interviews and the survey were conducted in a clinical setting with patients with cancer undergoing treatment at a major university hospital in Germany. As an exemplary case, we deem patients with cancer as well-positioned to evaluate a blockchain-based HIE mobile app since their treatment requires extensive coordination and data sharing across health care providers.

Results: Our findings support that patients have a high intention to use the blockchain-based functions that enable them to define, track, and revoke access to their health data per health care facility and service provider. Patients rated the 4 key functionalities (connection with providers, document sharing, a health diary, and a health care service provider search) as both useful and easy to use. The overall System Usability Scale of the blockchain-based HIE mobile app improved over the 3 phases up to 77.34, showing a good overall usability. The open feedback showed that patients' perceived usefulness of a blockchain-based HIE mobile app is especially influenced by 3 factors: the acceleration of the process of data sharing, patient-centered access control, and alignment with the respective health care settings. Moreover, patients' perceived ease of use of a blockchain-based HIE mobile app is impacted by 3 additional factors: the intuitiveness of the interaction, an aesthetic and functional design, and individual differences such as age or literacy with document management systems.

Conclusions: The evaluation demonstrates that patients are inclined to use blockchain-based HIE to manage their health data, as it empowers them to control which health care providers or individuals can access their information. To foster the use of a blockchain-based HIE mobile app, the app should allow patients to effortlessly establish connections with health care providers, offer an overview of all patient data, and enable patients to share medical documents individually via the app.

(J Med Internet Res 2026;28:e78849) doi:10.2196/78849

KEYWORDS

electronic health records; blockchain; patients; user-centered design; acceptance; usability; technology acceptance model; System Usability Scale; surveys and questionnaires

XSL•FO
RenderX

Introduction

Background

Patients often receive care from multiple health care providers throughout their lifetime, leading to health data being fragmented across institutions [1,2]. Multiple mechanisms have been introduced to address such fragmentation, including manual data transfers (eg, on paper or digital storage media) [3], electronic transfers via health information exchanges (HIEs) [4,5], and efforts aimed at centralizing patient information through general practitioners (eg, in the United Kingdom [6] or Germany [7]) or in regional or national databases (eg, in the United Kingdom [8] or Denmark [9]). Yet, patient data fragmentation persists [5,10], especially due to challenges such as incomplete records [6], data confidentiality concerns [11], and security vulnerabilities [12,13]. As a result, clinicians frequently lack a complete view of patients' medical histories, contributing to redundant tests, delayed diagnoses, and higher health care costs [1,14]. Developing a fully integrated, real-time patient record, therefore, remains a major challenge.

Blockchain technology has recently emerged as a promising solution for addressing data fragmentation issues and building interoperable, patient-centered HIEs [15,16]. By providing a decentralized, tamper-resistant ledger for managing consent and data access, blockchain can enhance data security and privacy [3,17] and give patients greater control over their information [17]. Especially, it enables patients to have an up-to-date overview of their health data (eg, which data are stored by which health care providers) and to decide independently who (eg, health care providers or service providers) can access their data and for what purposes [16]. These properties have positioned blockchain-based HIE as one of the most frequently discussed use cases of blockchain in health care among researchers and practitioners [15-20]. However, broad adoption of blockchain-based HIE still remains an issue. Reasons for this issue include a lack of organizational commitment, the absence of real-world testing of blockchain benefits in health care, and the functions of BC-based HIE being intangible for patients [19].

Prior research on blockchain-based health data management has examined diverse aspects, including challenges and benefits of its application [16,21-23], technical specifications [20,24,25], and users' attitudes toward different blockchain-based HIE functionalities [3,15,26,27]. These studies generally indicate that both health care professionals and patients have a positive attitude toward blockchain-based HIE and recognize its potential for improving transparency and data control. However, most of these investigations have focused on measuring general attitudes or perceived benefits through hypothetical scenarios [3,27,28] or interview-based discussions [15,29]. Thus, they mainly capture anticipated advantages rather than experienced ones. Only a few studies have involved prototype evaluations. Those studies typically focus on isolated functions, such as data sharing with research organizations [28], transaction tracking [30], or consent management [31], without allowing patients to actively manage access rights and view provider-specific medical records within an integrated system environment. Moreover, many blockchain-based HIE apps are still in a conceptualization stage, with very few having progressed to prototype development, let alone to actual implementation stages [16]. Consequently, a notable research gap remains regarding empirical evidence on how patients actually experience and interpret blockchain-based HIE in practice.

Objectives of the Study

Building on the identified research gap regarding the limited empirical evidence on how patients experience and interpret blockchain-based HIE systems in practice, this study aims to advance the understanding of patient engagement with such technologies beyond hypothetical attitudes. To this end, we conducted a 3-phase, user-centered design and evaluation of a functional blockchain-based HIE prototype, where patients recruited from a major university hospital interact with realistic usage scenarios (ie, patient-facing, practice-oriented interactions). Specifically, the study pursues three objectives: (1) to report on the user-centered design of a blockchain-based HIE mobile app prototype that enables patient-controlled management and sharing of health data; (2) to assess patients' experiences of blockchain-based HIE related to their current treatment in a clinical context and to evaluate their perceptions of the prototype's usability, usefulness, and overall user acceptance across four realistic HIE scenarios; and (3) to derive implications for the user-centered design and adoption of blockchain-based HIE solutions by identifying key functionalities and design features that influence user acceptance.

By examining patients' lived interactions with a functional blockchain-based HIE mobile app, this study provides empirical insights into how blockchain functionalities shape user experiences and adoption intentions in health care contexts. With our focus on evaluating a functional blockchain-based HIE mobile app with real patients, we explicitly go beyond the analysis of hypothetical attitudes as reported in prior research [3,27,28]. In addition, by examining patients' perceptions related to usability, usefulness, and acceptance across different functions of a blockchain-based HIE app, our study identifies which functions are crucial for blockchain-based HIE applications and where design improvements are needed. This is especially relevant given that blockchain-based applications are often characterized by low usability, as users have to invest a considerable effort to learn how to use these applications (eg, by installing and managing wallets for private keys) [32,33]. These insights can help researchers and practitioners identify barriers to the adoption of blockchain-based HIE, predict its actual use in the future, and make informed user-centered decisions about its implementation [34,35].

Methods

Overview

The development of the blockchain-based HIE app was carried out in collaboration with public health care providers and service providers (eg, a health cloud provider) in the context of a collaborative research project, BloG3. Our aim was to develop and test a blockchain-based HIE mobile app in the hospital context that would support patients in retrieving and sharing

XSL•FO
RenderX

documents from service providers and health care providers. For the development of the blockchain-based HIE mobile app, we used user-centered design involving patients as primary users to cater the mobile app to patients' needs [36,37]. User-centered design provides guiding principles and suggests techniques at various stages of the design process [36]. It ensures that an app is easy to learn, offers intuitive interactions, and is tied to user requirements. As our main study site, we chose the oncology department of a major university hospital in Germany, from where we recruited patients with cancer to evaluate the mobile app. We chose patients with cancer as the primary user group since their complex treatment pathways demand extensive coordination and data sharing across health care providers. In addition, their physical and emotional challenges further make usability especially critical for an HIE application. These characteristics make patients with cancer a crucial test case for evaluating blockchain-based HIE, which aims to support secure, patient-centered data sharing in contexts that require high coordination.

The research project was divided into 3 phases in which we iteratively developed and evaluated the functionalities of the mobile app (refer to Figure 1). The 3 phases enabled us to incorporate patient needs from an early design stage to the full implementation of the mobile app. After identifying user requirements (phase 1), we developed a first design of the mobile app (mock-ups) based on the identified user requirements. Subsequently, we evaluated the mock-ups with patients to receive further user feedback on the design (phase 2). Finally, we implemented a blockchain-based HIE mobile app and collected user perceptions of the functionalities of the mobile app in a clinical context, with patients undergoing medical treatment in a hospital, realistic usage scenarios, and realistic sample documents related to their care (phase 3). While phases 1 and 2 yielded important information for the development of the mobile app, phase 3 served to provide us with comprehensive feedback on the usability, usefulness, and acceptance of the implemented mobile app. This feedback is especially helpful to determine barriers to the adoption of blockchain-based HIE (eg, low usability) and predict its actual use in the future. Hence, while we also report on the results of phases 1 and 2, we put a higher focus on reporting the results of phase 3 in this study.

Figure 1. Phases of the development and evaluation process for the blockchain (BC)-based health information exchange (HIE) mobile app. TAM: technology acceptance model

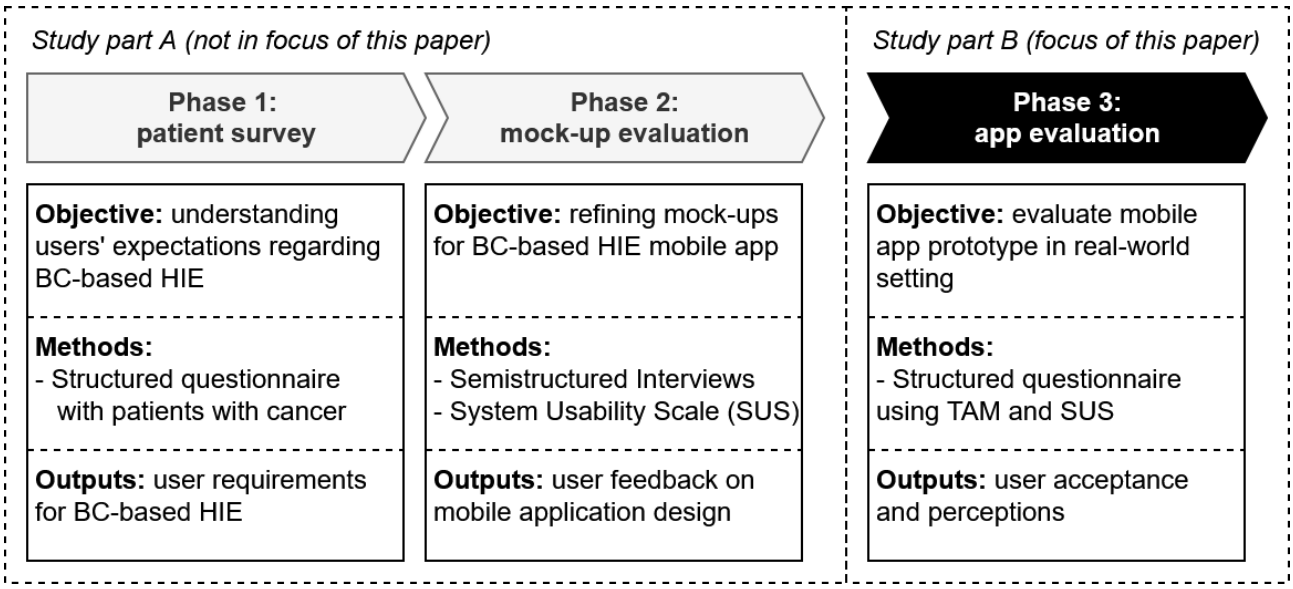

Phase 1: User Requirements Elicitation

In the first phase, we identified user requirements (eg, preferred app features). The aim of this phase was to get an overview of important features and health data that should be included in the blockchain-based HIE mobile app. To identify user requirements, we devised an online survey with questions about patients' current overview of their health data, whether they want to manage access rights per facility and doctor, whether they would like to have an overview of who accessed their health data, which documents they deem relevant to share, and which device they would like to use to manage their health data. Questions were iteratively developed by the author team based on prior literature on blockchain-based HIE [4,5,5-16], in a structured format, and could be answered on a 6-point Likert scale ranging from "1 (strongly disagree)" to "6 (strongly agree)." For pretesting the understandability of the questionnaire, we recruited a convenience sample of 18 participants who provided us with feedback, which we used to clarify unclear descriptions and specify ambiguous wording. An overview of the questionnaire and the survey results is available in Multimedia Appendix 1. We recruited participants via oncology patient communities nationwide by mail, inviting them to participate in the online survey. Patients with cancer in treatment or aftercare, next of kin, and other noncancer patients (eg, significant others) were able to participate to collect a broad overview of groups involved in the management of patient documents throughout the care process.

Phase 2: Mock-Up Evaluation

In the second phase, the identified user requirements were used as a foundation to develop mock-ups for a blockchain-based HIE mobile app consisting of a total of 36 clickable frames on a handheld computer. We developed the mock-ups iteratively

XSL•FO
RenderX

within the author team with external partners experienced in mobile app development for health care use cases. We evaluated the mock-ups with semistructured interviews during which we collected patient impressions of their interaction with the clickable mock-ups. Patients were asked to access different functionalities in the mock-ups, which were chosen based on user feedback from phase 1 and to answer questions related to the mock-ups. The functionalities included an overview of patient documents and medications, a diary for patient mood and health condition, and an overview of documents shared with exemplary health care facilities. An overview of the mock-up screens is shown in Figure 2. Further screens are available in Multimedia Appendix 2. The questions were developed based on the questionnaire from phase 1 and refined by the author team. These included questions related to the overall impression of the blockchain-based HIE mobile app, whether they would use it, and which medical documents they attach the greatest importance to.

Figure 2. Mock-up screens of the blockchain (BC)-based health information exchange (HIE), showing a general overview of recent patient document interactions, medications, and available patient documents.

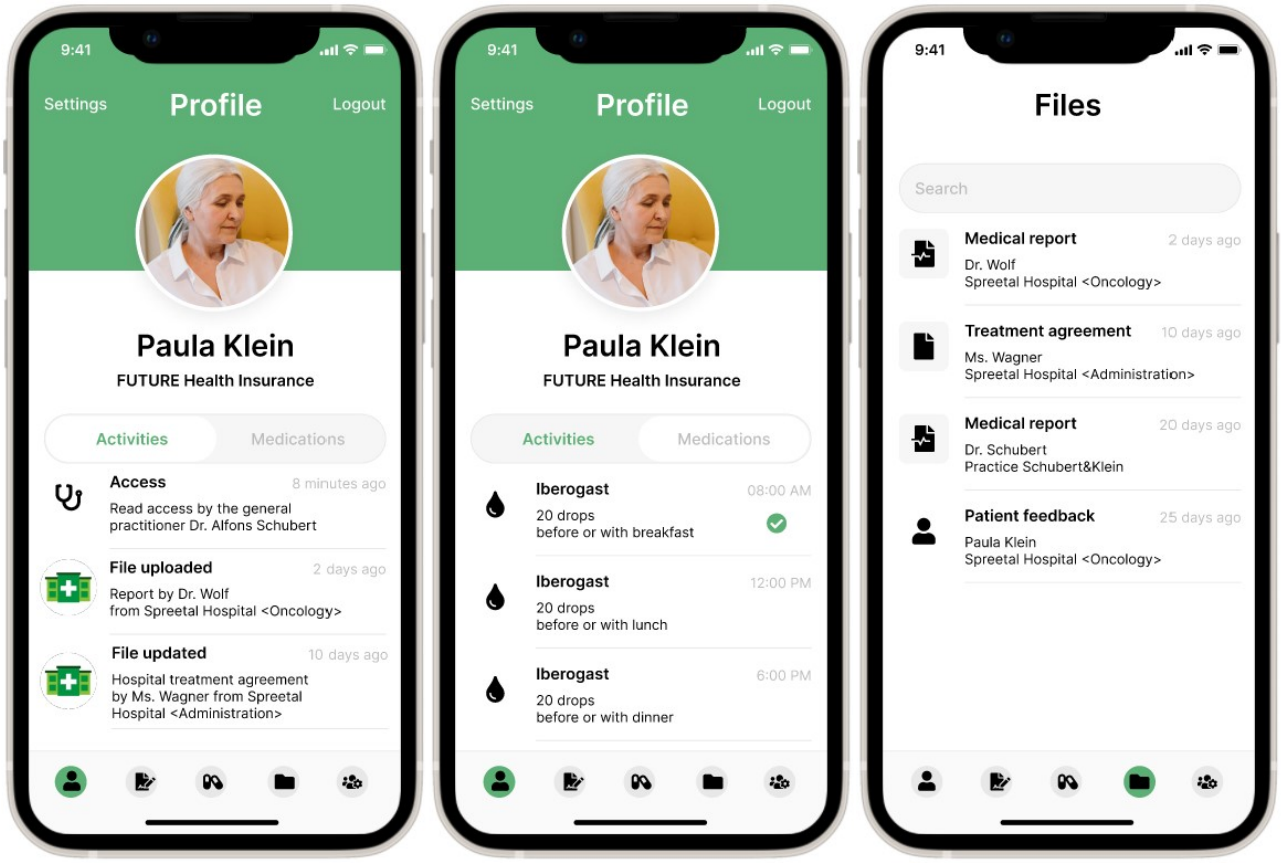

To measure usability, we collected patient perceptions of the app using the System Usability Scale (SUS) using an established German translation [38]. We chose SUS as it is established for measuring the user experience of a given software application [39]. We classify the SUS with an established grading scale ranging from a worst grade of "F" SUS range of 0.0 - 51.6 to a best grade of A+ range of 96.0 - 100 to assess the usability with a measure that shows what an acceptable SUS score is [39]. The interviews took place on-site in the oncological outpatient department of our main study site. We recruited adult patients at random, by willingness to participate, through on-site research staff. We applied quota sampling by time to randomize participants; that is, we recruited patients on several days at different times. Due to ethical concerns, we explicitly excluded patients younger than 18 years. We emphasized to patients that refusing to participate would not affect their treatment in any way. If patients refused to participate, we did not contact them again and excluded them from data collection. A researcher of the author team introduced patients to the idea of the blockchain-based HIE mobile app, observed patients while they clicked through the mock-ups, and gave instructions if needed. After patients finished accessing all functionalities, they filled out the SUS on a paper sheet. We aimed to provide only minimal guidance for clicking through the mock-ups. We encouraged participants to explore the functionality of the app themselves on a provided mobile device. If a participant could not access certain functionality, the researcher showed them how the functionality could be accessed by clicking through the mock-ups. While the researcher focused on highlighting the functionality of the mock-ups, BC-based HIE was introduced to patients as the mechanism that enabled sharing and revoking access to patient documents.

Phase 3: Mobile App Evaluation

In the final phase, we developed the blockchain-based HIE mobile app and a blockchain-based backend infrastructure, which enabled patients to share their medical documents. While our system was based on a system design recently proposed in the literature [40], we based the mobile app on the clickable frames from phase 2. To evaluate the blockchain-based mobile

XSL•FO
RenderX

app, we conducted interviews with patients in the oncological outpatient department of our main study site using the mobile app in practice-oriented scenarios that users face during their treatment.

Description of the Developed Blockchain-Based HIE System

For our developed blockchain-based HIE system, we adopted a system design recently proposed in the literature [40]. During system design, we incorporated threat modeling using the STRIDE (Spoofing, Tampering, Repudiation, Information disclosure, Denial of service, Elevation of privilege) method to systematically identify potential security risks, inform the architecture, and ensure secure interorganizational health data exchange [41]. An overview of the system is shown in Figure 3. The system consisted of the following four components: (1) a mobile app, (2) network infrastructure with a blockchain, (3) a connector to the IT systems of health care facilities and service providers, and (4) the existing internal IT systems at the health care facilities and service providers.

Figure 3. Blockchain (BC)-based health information exchange (HIE) system overview. IT: Information Technology

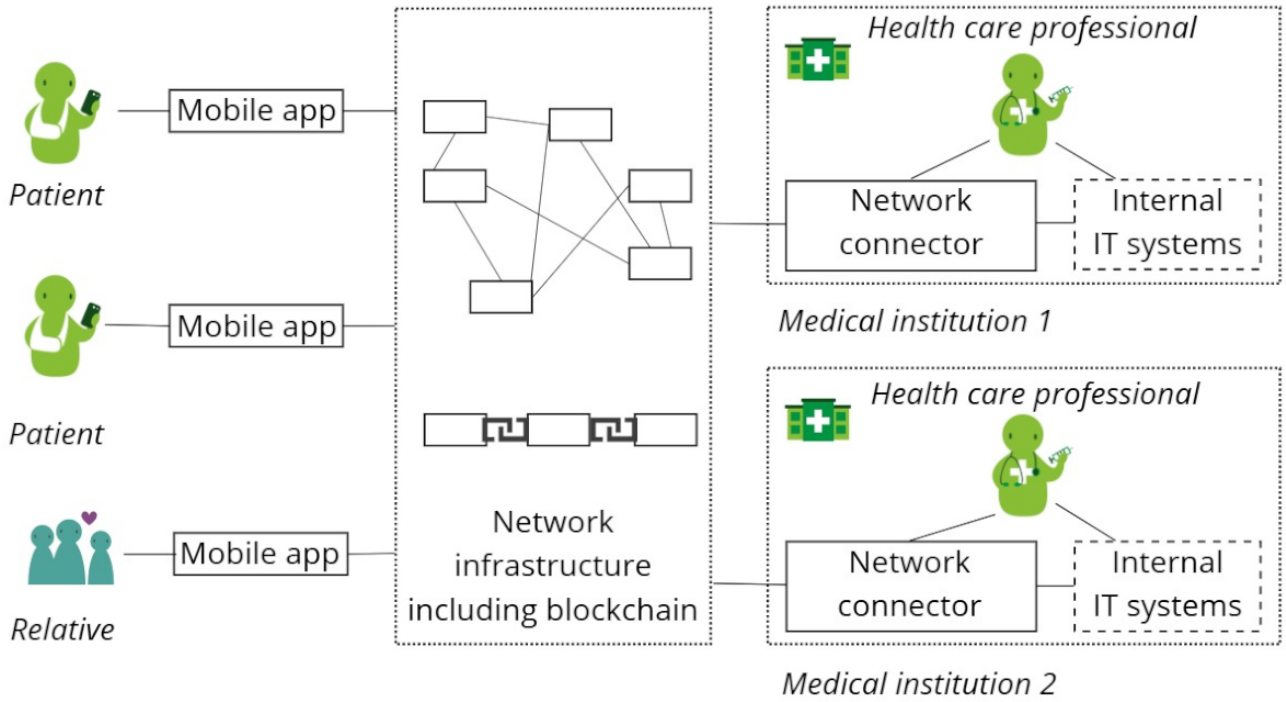

The network infrastructure is a decentralized platform secured by blockchain technology. It enables secure, asynchronous, encrypted peer-to-peer communication and data exchange between patients' mobile apps and connectors of health care facilities and service providers. The blockchain publishes a secure and tamper-proof registry for digital signatures with which patients, health care facilities, and service providers verify their identities. This enables patients to grant access rights to their health data to trusted health care facilities, service providers, and individuals based on their verified identity. Health data are stored off-chain within the existing infrastructure of the health care facility and service providers, ensuring that the data remain securely within the provider's systems, thereby eliminating the need to rely on external systems for trust. Additionally, there are no issues with deleting health data from the blockchain, which is a common challenge for blockchain use in health care. The connector is integrated into the internal IT systems of health care facilities and service providers and acts as an interface within the network infrastructure. This interface facilitates seamless access to data stored within health care facilities and service providers while ensuring interoperability and compliance with privacy and security standards. The system is based on the open-source implementations of Hyperledger Aries [42] and Indy [43] by the LF Decentralized Trust as part of the Linux Foundation, with connectors implemented using Aries and the public blockchain using Indy. Hyperledger Indy serves as the blockchain-based registry for identity management and credential verification. For this, Indy uses Plenum, an implementation of the Redundant Byzantine Fault Tolerance consensus mechanism. It should be noted that during the evaluation, no connection to live EHR systems or national health infrastructures was implemented. However, we set up connectors at 2 different health care facilities, including our main study site, to provide synthetic patient data during the evaluation. Overall, the system architecture was designed to be modular and extensible to allow future adaptations.

Description of the Mobile App for Patients

The mobile app is the user interface of the developed blockchain-based HIE system for patients and was developed based on the patient requirements from phase 1 and the feedback to the mock-ups from phase 2. It serves as a central access point and hub for the management of patients' health records stored at different health care facilities and service providers. By actively involving patients, the mobile app ensures privacy and control over their health data. Moreover, the mobile app was developed using the Xamarin platform to enable cross-platform deployment [44]. The 4 key functionalities of the mobile app include establishing a connection with a health care or service

XSL•FO
RenderX

provider by scanning a QR code, selecting and sharing specific documents through the system with other health care and service providers involved in patient care (Figure 4), creating health diary entries, and searching for nursing homes within a given area. An overview of the blockchain-based app screens is included in Multimedia Appendix 3.

Figure 4. Blockchain-based health information exchange (HIE) app screens for sharing specific documents with health care and service providers and entering data into a patient diary.

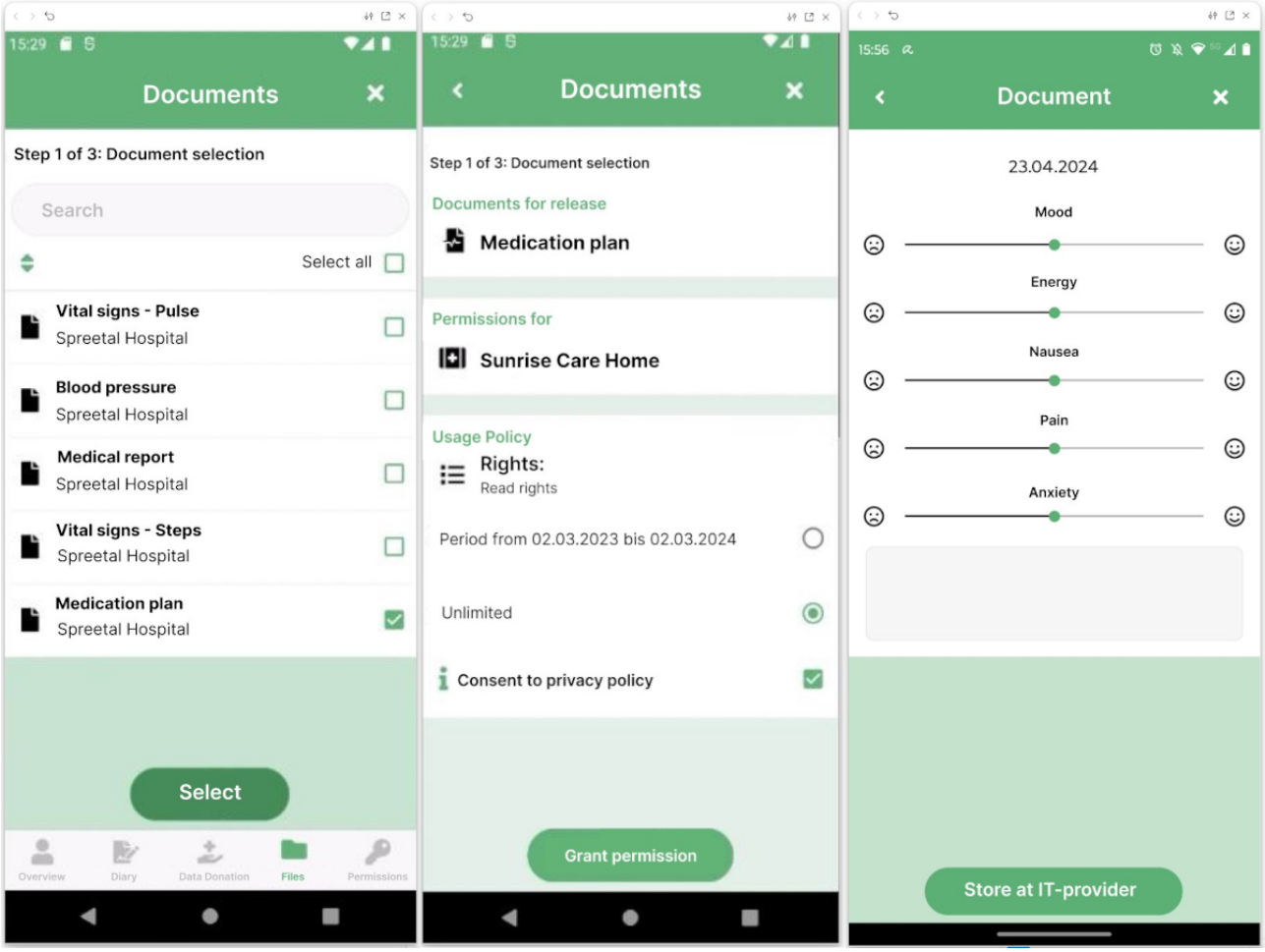

For patients to access their health data, a connection to the health care facility or service provider is required. Establishing a connection begins with the patient scanning a QR code provided by the health care or service provider using their mobile app. This action initiates the registration of the patient's mobile app with the service provider's connector. Subsequently, the service provider is listed as a registered contact in the patient's mobile app. When new health data becomes available, the patient receives an automatic notification via their mobile app. The patient can directly access and filter the health data from the service provider using the mobile app. Furthermore, they can grant other registered organizations access to this data. Patients can share their data from the primary health market, like health care providers (eg, findings and medical reports), data from the secondary health care market, like service providers (eg, recordings of fitness watches), as well as self-generated patient-reported outcome data (eg, symptoms in a patient diary). The patient-reported outcome data can be documented multiple times a day using a diary function with 5 entry types that can be recorded using sliders: Mood, Energy, Nausea, Pain, and Anxiety. A free-text field was included for individual comments. Based on patient feedback during interactions with the mock-ups, an additional feature was implemented, enabling patients to search for nursing homes within a specified area. This function was added since patients wanted to search and connect to further relevant providers directly via the mobile app without the need for scanning another QR code. Overall, to further improve the usability of the app, we added more information to symbols and buttons since patients indicated that the symbols used were not intuitive.

Study Description for Evaluating the Mobile App

For the evaluation of the blockchain-based HIE mobile app, we recruited patients from the oncological outpatient department of our main study site. A researcher from the author team recruited adult patients based on their willingness to participate. We randomized participants by applying quota sampling by time. Accordingly, we recruited patients on several days at different, varying time slots, thereby reducing time-based recruitment bias. Systematic randomization was not applicable since the time for data collection was limited. We explicitly excluded patients younger than 18 years due to ethical concerns. We informed patients that their participation in the study was voluntary and would not affect their care in any way. When patients refused to participate, we did not contact them any further and excluded them from further data collection. For the

XSL•FO
RenderX

systematic evaluation of the blockchain-based HIE mobile app, we devised a questionnaire that introduced the participants to blockchain-based HIE, asked for demographic information, and asked the participants to complete 4 tasks with the mobile app. The questionnaire was developed and assessed in several iterations within the author team to improve its understandability. To ensure its suitability for measuring patient perceptions, we revised the questionnaire in terms of understandability and clarity based on the feedback from 2 IT professionals of the collaborative research project BloG3 in which the studies were situated and 2 physicians. The questionnaire is available in Multimedia Appendix 4, and an overview of the survey results is available in Multimedia Appendix 5.

While the researcher on-site presented the questionnaire to patients on paper, the blockchain-based HIE mobile app was handed out on a provided mobile device to patients. Similar to phase 2, the researcher offered only limited guidance for clicking through the mobile app. The researcher waited at least 30 seconds until a participant received help, encouraging participants to complete the questionnaire by exploring the functionality of the mobile app themselves. If participants indicated that they did not wish to proceed with a task, the researcher explained to the participant where to click.

The first part of the questionnaire contained a brief introduction to the concept of blockchain-based HIE. The key functionalities of the blockchain-based HIE were highlighted as the scanning of a unique decentralized identifier (DID) via a QR code and the function of sharing and revoking access to patient documents. Patients could only perceive the functionality that is enabled and to some degree necessitated by using a blockchain (eg, when establishing a connection with a health care facility or service provider). Next, we collected demographic information on gender, age range, number of document management systems used (eg, cloud storage providers), and frequency of their use.

The main part of the questionnaire set 4 tasks that patients had to complete with the mobile app. Each task relates to 1 of the 4 key functionalities of the mobile app. For each task, we collected 2 outcome measures by noting whether a patient was able to complete a task without assistance and asking patients to confirm the result of a task (eg, being connected to a specific health care facility). If a patient indicated having difficulties completing a task, the researcher assisted the patient in completing the task on the mobile device by providing information on what to look for. After completing a task, patients were asked to assess the functionality with the 10-item questionnaire of the technology acceptance model (TAM) [45]. The items measured the 3 main TAM constructs, perceived usefulness (PU), perceived ease of use (PEOU), and behavioral intention (BI) to use, on a 7-point Likert scale ranging from “1 (strongly disagree)” to “7 (strongly agree).” The TAM constructs enable the measurement of user perceptions of how useful an application is for a given task (PU), how effort-free the interaction is (PEOU), and how likely they will adopt an application [45]. We chose TAM as it is an established measurement tool to evaluate user perceptions of software applications that allow us to predict whether users are likely to use an app before its rollout [46]. For our study, we based the items on a validated German translation [47]. In task 1, patients were asked to establish a connection with a health care facility by scanning a QR code. Task 2 requested patients to share preset medical documents and health data (ie, a medication plan, blood pressure values, heart rate, and steps) in their mobile app with a health care facility and a service provider. For one, they shared the medical document for a limited timeframe. For the other, they were required to share and revoke access to the medical document. In task 3, patients were asked to enter a random mood and health condition, share it with one of the health care facilities or service providers, and delete it. In task 4, patients searched for a nursing home to which their health data could be shared via the mobile app to support the discharge process. After completing all tasks, patients were asked to evaluate the overall usability of the blockchain-based HIE mobile app with the SUS [38]. Similar to phase 2, we classify the SUS with an established grading scale [39]. Finally, we collected free-text responses to get additional qualitative feedback on perceived advantages, challenges of the app, and suggestions for improvement.

Data Analysis

For every task, we calculated the mean scores and SDs for the used constructs (PEOU, PU, and BI). In addition, we checked the scale reliability of the used constructs, since we reworded the items to include the name of the tested blockchain-based HIE mobile app instead of a general “system.” We thereby ensure that the items work cohesively for our study setting. For testing the scale reliability, we checked the internal consistency of the used, reworded items for PEOU and PU with Cronbach α [48]. For measuring the internal consistency of the items for the construct BI, we used the Spearman-Brown coefficient, since BI is only a 2-item scale for which Cronbach α cannot adequately determine the reliability [49].

To better understand the rationale behind participants’ perceptions, we applied thematic analysis to patient responses to the 3 open questions. For the analysis, we followed the 6 phases outlined in [50]. In the first phase (familiarizing with the data), 2 authors independently reviewed the responses and mapped each to 2 of the 3 TAM constructs that influence the intention to use (ie, PU and PEOU). During the mapping process, they extracted key statements describing factors influencing these 2 constructs and performed an initial round of coding to generate first-order codes (second phase: generating initial codes). The 2 authors synthesized similarities and discussed differences among these initial codes with 1 additional author. In the third phase (searching for themes), 2 authors iteratively identified relationships between initial codes and clustered them into broader, conceptually connected themes. In the subsequent phase (reviewing themes), the team evaluated the internal consistency and distinctiveness of these candidate themes, ensuring that each accurately reflected the coded data and the dataset as a whole. This stage involved integrating closely related themes, subdividing those that were too encompassing, and reallocating codes that did not clearly belong to their original theme. This iterative process led to the identification of 6 themes. In the fifth phase (defining and naming themes), we defined the themes by identifying the

XSL•FO
RenderX

"essence" of each theme and clarifying its conceptual boundaries (ie, what aspect of participants' responses it represented and what it excluded). Based on their definitions, we named the refined themes and categorized them based on their contextual dependencies. The final outcomes of this analysis are detailed in the "Results" section (phase 6: producing the report). An illustrative example of this process is provided in Figure 5.

Figure 5. Excerpt from the qualitative coding process.

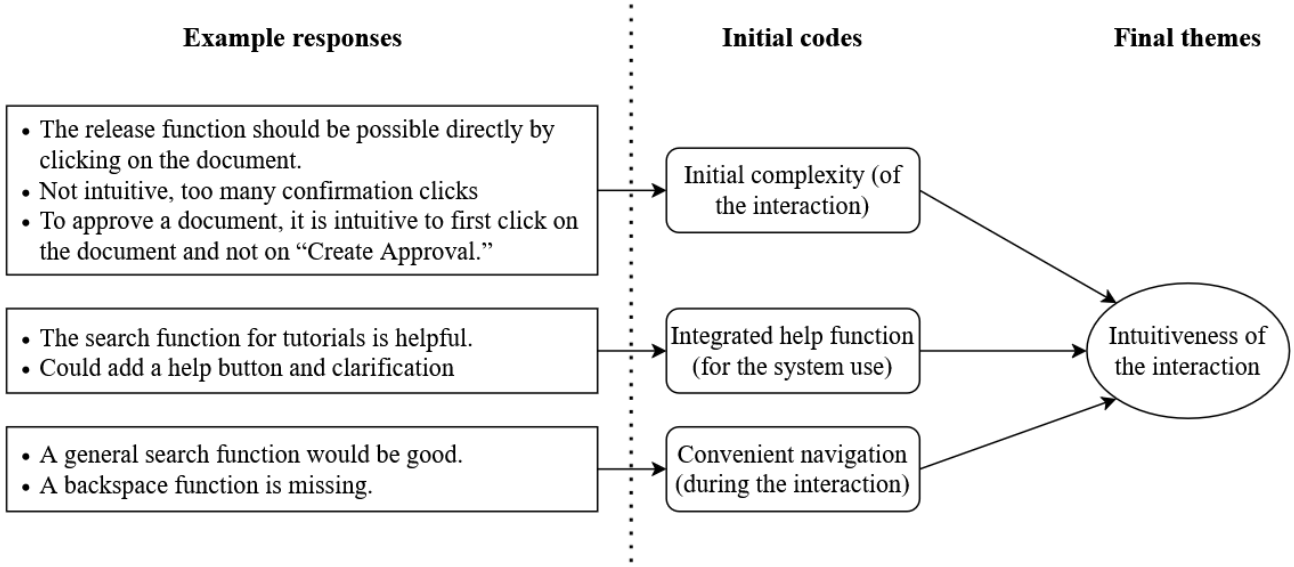

Ethical Considerations

All study parts were approved by the local ethics committee (Ethikkommission der Charité, no EA4/241/21). In the online survey in phase 1, a privacy statement informed participants about how their data were used and processed for the study. In phases 2 and 3, patients provided informed consent for the interviews. In all study parts, participation was voluntary and anonymous.

Results

Phase 1: Collection of User Requirements

In phase 1, we collected 108 responses from potential users, of which 60 completed the whole questionnaire (31 patients with cancer and 29 significant others involved in their care). From these, we could derive 4 requirements for a blockchain-based HIE application. An overview of participants' gender and age distribution is shown in Table 1. The responses showed that participants rather agreed with having a complete overview of their current health documents, although they indicated that neither their primary care physicians nor their relatives have a complete overview. More importantly, half of the participants (n=30) stated that they forgot or lost documents themselves, their health care staff lost documents, or documents were provided too late. Hence, we defined as the first requirement that the application should contain an overview of patients' health documents (Req 1). Regarding the platform for managing health data, participants stated that both a computer and a smartphone would be their preferred way for managing their health data. To allow for better portability and easier access everywhere, we decided to focus on a mobile app and defined the second requirement: the app should be accessible via smartphone (Req 2). Of all possible medical documents and health data, participants rated diagnosis reports, laboratory results, imaging results, and their medication plan as most important for sharing them with their health care providers. Thus, the app should at a minimum support the sharing of diagnosis reports, laboratory results, imaging results, and their medication plan (Req 3). Participants displayed high agreement with whether they would like to define who can access, edit, and delete their data per health care facility and physician. Participants preferred sharing specific documents instead of all data at once. In addition, they gave high priority to being able to track and revoke who accesses their data. Hence, the application should enable users to share, edit, track, and delete access rights to their medical documents per document per healthcare facility or service provider (Req 4). Overall, participants supported the idea of a blockchain-based HIE mobile app to manage access to their health data.

XSL•FO
RenderX

Table . Demographic data of participants for phase 1 (N=60).

Participant demographic	Values (N=60), n (%)
Sex	
Diverse	1 (1.5)
Female	28 (41.8)
Male	31 (56.7)
Age (years)	
18 - 29	9 (15)
30 - 39	13 (21.7)
40 - 49	7 (11.7)
50 - 59	13 (21.7)
60 - 69	13 (21.7)
70 - 79	4 (6.7)
80 - 89	1 (1.7)

Phase 2: Mock-Up Evaluation

In phase 2, we recruited 22 patients with cancer to evaluate the mock-ups of the blockchain-based HIE mobile app. Seventeen patients declined to participate. An overview of participants' gender and age distribution is shown in Table 2. The semistructured interviews provided insights into patient perceptions of the mobile app. Overall, the expectations of interviewed patients were positive. Patients stated that they would be very likely to use the mobile app once it is fully implemented. Patients had high anticipation of more efficient and transparent management of medical health care documents as well as faster communication between all stakeholders involved in the treatment process. The responses indicated that patients attach great importance to being able to manage the following health data in a blockchain-based HIE mobile app: medical findings, diagnoses, operating reports, laboratory results, and the medication plan.

Table . Demographic data of participants for phase 2 (n=22).

Participant demographic	Value, n (%)
Sex	
Female	14 (63.6)
Male	8 (36.4)
Age (years)	
18 - 29	0 (0)
30 - 39	3 (13.6)
40 - 49	4 (18.2)
50 - 59	8 (36.4)
60 - 69	6 (27.3)
70 - 79	1 (4.5)

For the clickable mock-ups, patients rated the usability with the SUS on average as 66.50, which hinted that the usability is satisfactory and can be graded as a "C" [39,51]. Still, several participants hinted that there may be challenges for older people to use the mobile app since several clicks were necessary to access certain functionalities. This challenge was noted and accounted for in the next phase of development to further enhance the usability of the mobile app.

Phase 3: Mobile App Evaluation

User Statistics

In the last phase, 32 participants were recruited, consisting of 13 males and 19 females. Nineteen patients declined to participate. The age range of participants spanned from 18 to 79 years, with the largest group being between 30 and 39 years (n=10). The gender and age distribution are shown in Table 3. The survey also gathered data on the number and frequency of IT systems used for document management (Table 2). Fourteen participants did not use any systems, while others reported using between 1 and more than 3. The frequency of use varied, with 18 participants engaging with these systems at least once a week.

XSL•FO
RenderX

Table . Demographic data of participants for phase 3 (n=32).

Participant demographic	Value, n (%)
Sex	
Female	19 (59.4)
Male	13 (40.6)
Age (years)	
18 - 29	1 (3.1)
30 - 39	10 (31.3)
40 - 49	5 (15.6)
50 - 59	6 (18.8)
60 - 69	8 (25)
70 - 79	2 (6.3)
80 - 89	0 (0)
IT[a] systems used for document management	
0	14 (43.8)
1	12 (37.5)
2	5 (15.6)
3	0 (0)
>3	1 (3.1)
Frequency of IT system use for document management	
Never	14 (43.8)
Once a month	5 (15.6)
Several times a month	3 (9.4)
Once a week	4 (12.5)
Several times a week	3 (9.4)
Daily	3 (9.4)

[a]IT: Information Technology

Evaluation Outcomes

All patients completed Tasks 1 and 4 without any assistance and correctly confirmed their connection to a health care facility (Task 1) and finding a nursing service (Task 4). Throughout Task 2, 6 patients (including both patients in the oldest age group [70-79 years]) needed assistance in carrying out the task. In Task 3, 1 patient received assistance in carrying out the task.

Table 4 provides an overview of the mean scores, SDs, and reliability measures (Cronbach α and Spearman-Brown coefficient) for the constructs PU, PEOU, and BI for each of the 4 tasks. A graphical comparison between the 3 constructs over the 4 tasks is shown in Figure 6. Overall, participants reported high levels of PU, PEOU, and BI for all tasks, with all reliability coefficients exceeding the recommended threshold of 0.70, displaying high internal consistency of the adapted scales [52].

XSL•FO
RenderX

Table . Technology acceptance model (TAM) construct measures for the blockchain (BC)-based health information exchange (HIE) app.

Construct	Task 1	Task 2	Task 3	Task 4
Perceived ease of use				
Value, mean (SD; 95% CI)	5.48 (1.42; 4.97-5.99)	5.69 (1.37; 5.19-6.18)	5.67 (1.33; 5.19-6.15)	5.6 (1.16; 5.18-6.02)
Cronbach	0.94	0.93	0.96	0.95
Perceived usefulness				
Value, mean (SD; 95% CI)	5.5 (1.51; 4.96-6.04)	5.84 (1.28; 5.37-6.30)	5.73 (1.29; 5.26-6.19)	5.59 (1.14; 5.17-6.00)
Cronbach	0.98	0.95	0.97	0.97
Intention to use				
Value, mean (SD; 95% CI)	5.61 (1.46; 5.08-6.14)	5.67 (1.6; 5.09-6.25)	5.47 (1.65; 4.87-6.06)	5.47 (1.28; 5.01-5.93)
Spearman-Brown coefficient	0.88	0.97	0.9	0.92

Figure 6. Perceived usefulness (PU), perceived ease of use (PEOU), and behavioral intention (BI) values and SDs for all 4 tasks of phase 3.

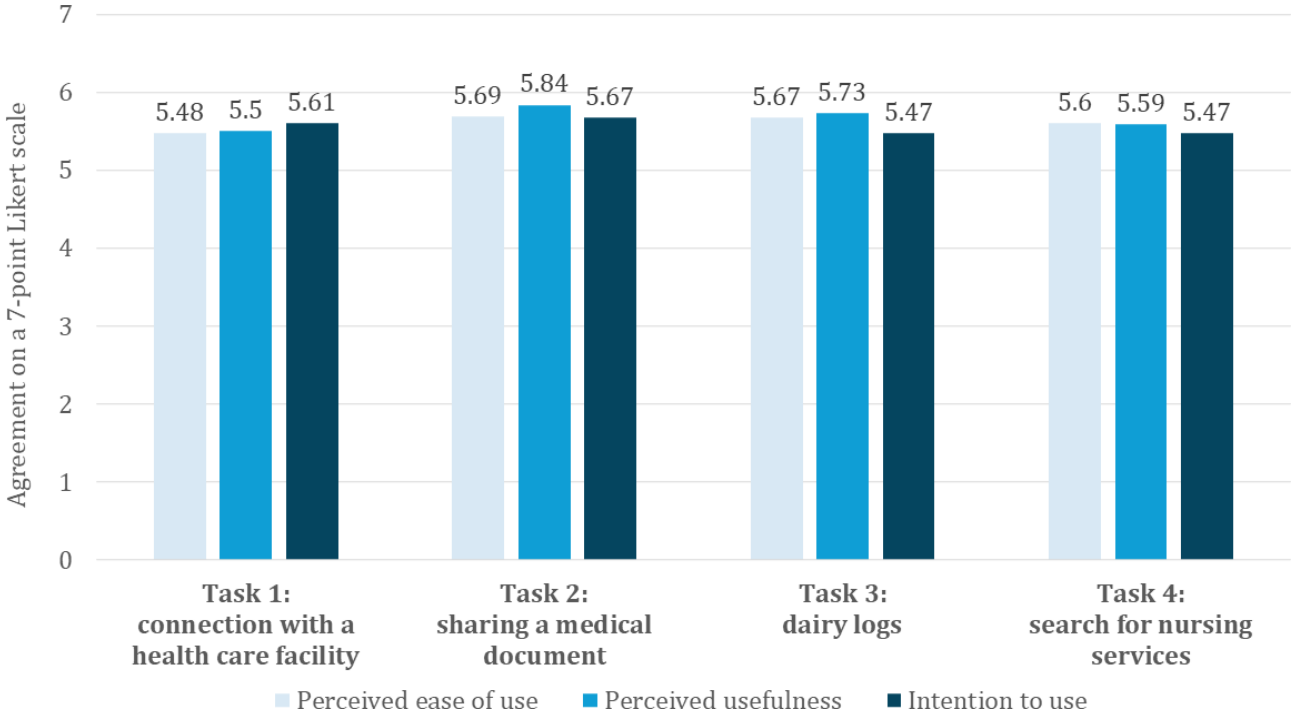

The developed blockchain-based mobile app achieved a SUS score of 77.34, indicating that the usability was good and could be graded as "B+" [39,52]. This showed an improvement from the 66.50 (graded as "C") reported in phase 2 (refer to Figure 7). It is noteworthy that older participants (ie, age groups 60 - 69 and 70 - 79 years) have a smaller SUS score of 64 (graded as "C–") and hence only a satisfactory rating of the mobile app. In contrast, younger age groups (ie, 18 - 59 years) have a SUS score of 83 (graded as "A"), indicating a high usability [51]. The results show that the BC-based mobile app received a higher SUS score of 84.12 (graded as "A") among female participants. In contrast, male participants provided a rating of 67.43 (graded as "C"). We observed a minor difference in the SUS scores between participants who indicated that they would not use any document management system in general (74, graded as "B–") and those who indicated using at least 1 such system (78.73, graded as "B+"). The SUS score of low frequencies of general document management system use was 75.61 (graded as "B") and of high frequencies of general document management system use was 79.87 (graded as "A–"), showing only a minor difference.

XSL•FO
RenderX

Figure 7. Grading of System Usability Scale (SUS) scores from phases 2 and 3.

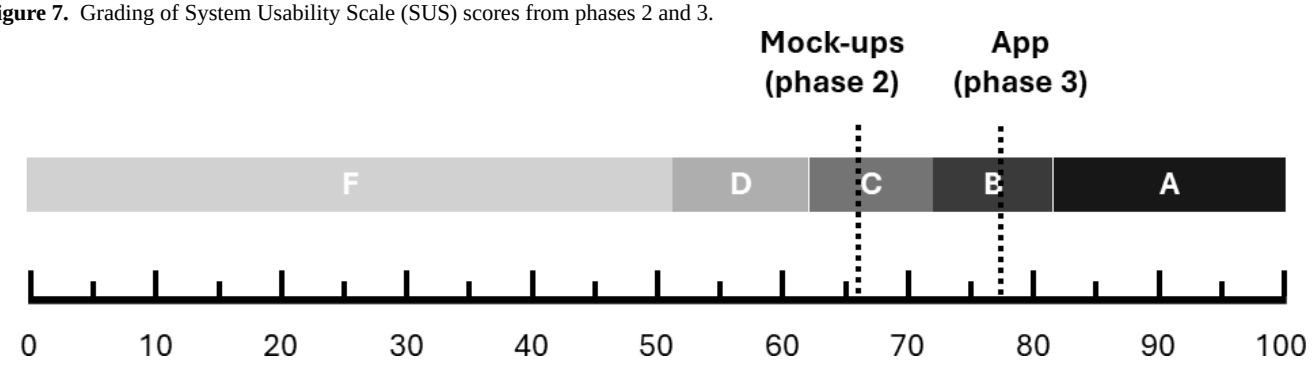

For completeness, we report on a subgroup analysis using Aligned Rank Transform ANOVAs to examine the influence of gender, age, and number and frequency of IT systems used for document management on PEOU, PU, BI, and SUS (Multimedia Appendix 6). Aligned Rank Transform was used to align and rank the data to be able to apply ANOVA to our ordinal data [53]. Both analyses are exploratory and should be interpreted with caution given the limited sample size and the low statistical power.

Influential Factors for User Perceptions

Our analysis of responses to the 3 open questions revealed 3 influential factors, each for PU and PEOU (refer to Table 5). The factors influencing PU reflect the specific contextual nuances of blockchain-based HIE, highlighting that users view a blockchain-based HIE application's usefulness primarily through its ability to address specific challenges and requirements in this domain. In contrast, the factors shaping PEOU are context-agnostic, since they are not dependent on functionalities enabled by blockchain-based HIE. These align with broader user expectations for a general system's usability and are straightforward to adopt across diverse settings (refer to Figure 8).

Table . Factors influencing user perceptions of blockchain (BC)-based health information exchange (HIE) applications.

Factors	Description	Exemplary quotes
Operational efficiency of health data management	The HIE[a] application's ability to organize and manage health-related information in a manner that reduces the time, effort, and resources required.	• "Reduces bureaucratic overhead" • "Everything is concentrated and easily accessible"
Patient-centeredness	The HIE application's characteristic of placing patients at the center of its functionality by enabling them to manage their health data while ensuring transparency and privacy.	• "I like the feature that I can set my release options personally" • "Autonomous or not dependent on help from third parties"
Alignment with health care realities	The extent to which the HIE application reflects and supports the practical, real-world needs and workflows of health care settings.	• "It would be nice if there was an appointment entry function" • "The app should be available to everyone, including on mobile phones and via a browser."
Intuitiveness of the interaction	The ability of the HIE application to provide a seamless and intuitive interaction experience.	• "The search function for tutorials is helpful" • "To approve a document, it is intuitive to first click on the document"
Aesthetic and functional design	The visual appeal and practical functionality of the HIE application.	• "I would like to see a graphical representation" • "The visual design is appealing"
Individual differences	Personal characteristics such as age, health status, and related experience that influence individuals' attitudes toward the HIE application.	• "I see that as very personality-dependent" • "Older people may not be able to handle the many clicks."

[a]HIE: health information exchange.

XSL•FO
RenderX

Figure 8. Mapping of factors influencing user perceptions to the technology acceptance model.

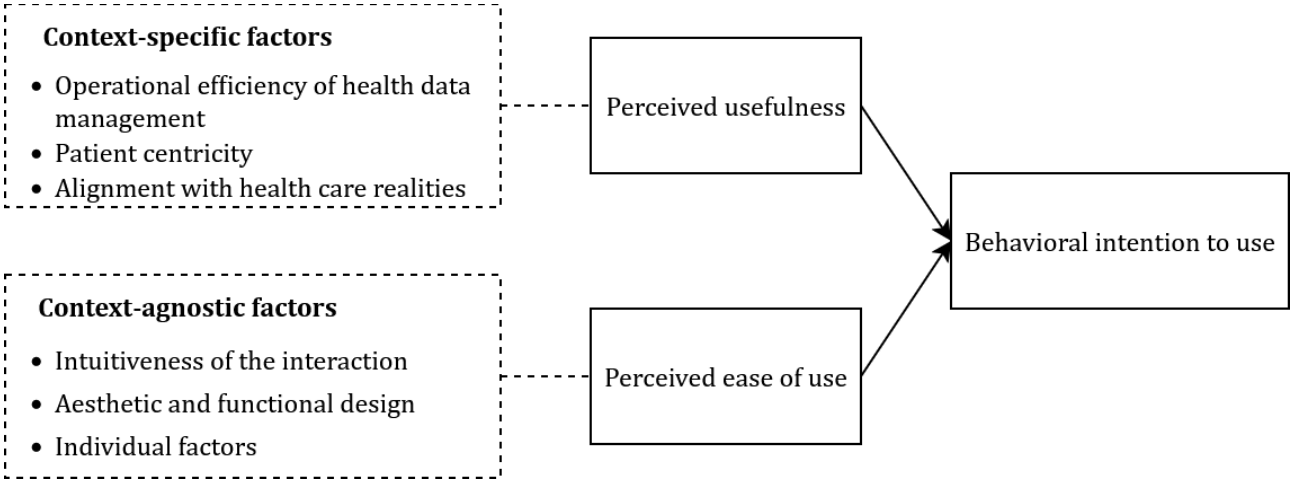

For PU, the following 3 factors were identified that can represent the key aspects users consider when evaluating the value and functionality of a blockchain-based HIE application.

Operational Efficiency of Health Data Management

Our analysis identified the HIE application's ability to organize and manage health-related information efficiently as a critical factor for its PU. This finding aligns with the highest PU score obtained for the document sharing task (Task 2), suggesting that participants particularly valued the application's ability to facilitate data exchange between health care facilities and service providers. In particular, they highlighted the application's strengths such as "faster document retrieval" and "easy sharing processes," which streamline the exchange of health-related information. Furthermore, features that reduce administrative burdens, such as "simplified [processes to achieve] compliance with data policies" and "straightforward access to documents," were highly valued for enhancing the efficiency of document retrieval. The organization of health data in one place also stood out. Participants appreciated the application's "integrated view of all relevant data," which provides a clear overview, ultimately supporting more effective health data management.

Patient Centricity

Many participants attributed the positive PU of our HIE application to its user-centered advantages of empowering users to take ownership of their health data while protecting their privacy. This factor exerts an influence on participants' perception of the application's usefulness, as it allows them to manage their own data access and sharing settings, thereby fostering a sense of autonomy. Some users especially noted that "self-determined and self-organized document sharing" resulted in a more responsive system that better met their needs. Additionally, transparency in how user data is handled contributed to establishing trust in the system and enhanced users' confidence in its utility. Participants expressed appreciation for "clear policies and measurements regarding data privacy," with some indicating that "knowing their data was secure and handled responsibly" addressed their concerns related to privacy.

Alignment With Healthcare Realities

Beyond the evaluated functions of the blockchain-based HIE application, participants emphasized the importance of aligning its capabilities with the health care setting's structural and technological demands. Functions reflecting real-world healthcare scenarios, such as accessing medication plans and document sharing, were emphasized to enhance the application's relevance and feasibility. This also helps explain the strong PU ratings for the document sharing task (Task 2). Moreover, participants suggested that additional features such as appointment scheduling, proof of insurance, and document search functions would further improve the PU. Although these were not tested in our survey, the participants considered them to be valuable for improving the usefulness of the application, as they can support people in their common activities of daily life. Furthermore, participants noted that the compatibility between the system and the existing infrastructure affects the PU of the application. Participants expressed that the application "should work seamlessly on different platforms or devices, including mobile phones and web browsers." Moreover, maintaining "accessibility in locations with limited network connectivity" can reinforce the application's usefulness in health care settings, as it facilitates uninterrupted service delivery and enables timely decision-making in critical situations when vital medical information is needed.

When reporting the PEOU, participants also identified 3 key factors that influence their perceptions:

Intuitiveness of the Interaction

The ability of the HIE application to provide a seamless and intuitive interaction experience was reported in our survey as a critical factor influencing PEOU. Many participants stressed that the system should feel natural and require minimal learning, particularly for new users. For example, participants value features that facilitate convenient navigation, such as the "inclusion of a backspace function." "Clear and consistent terminology" was also regarded as essential to "minimize [user] effort and frustration." The positive feedback on interface clarity and navigation aligns with the marked improvement in SUS from phase 2 (66.5) to phase 3 (77.3), suggesting that these refinements in interaction design translated into perceptibly higher ease of use. In addition, participants identified built-in

XSL•FO
RenderX

guidance, such as “a searchable help or tutorial feature,” as an important component to assist users in understanding the application’s functionality and resolving potential difficulties.

Aesthetic and Functional Design

Several participants pointed out that the design of the HIE application plays an important role in PEOU. On the one hand, participants appreciated well-organized and visually appealing interfaces, as they made the application more attractive and easier to navigate. In particular, information displayed in structured timelines or graphical formats was considered helpful, such as “visually appealing curves” and “organized layouts by day, week, or month, [which] make [the information] easier to understand.” On the other hand, participants attributed high PEOU to customization options, including the ability to “enter free text” or “filter information by various criteria (eg, diagnoses and providers).” These options enable the application to better fulfill users’ individual needs, thereby fostering a more comfortable user experience and, ultimately, PEOU.

Individual Differences

Personal characteristics emerge from our analysis as an influencing factor for PEOU. Given the health status of our participants, the first observation that caught our attention was that participants with serious health challenges reported that each new type of interaction felt less easy and more stressful. Besides, participants expressed concern about the application’s complexity for older users, noting that “too many steps or interactions could discourage [their] usage.” This observation aligns with concerns reported in phase 2 and the lower SUS scores reported by older participants in phase 3 (64 vs 83 for younger users), suggesting that PEOU was partly constrained by age-related differences. Furthermore, related experience and individual preferences further shaped perceptions. While some participants expressed discomfort or resistance due to limited familiarity with digital systems, 2 participants directly stated that their resistance stemmed solely from their personal aversion to digitization in general, not from the blockchain-based HIE application itself.

Discussion

Principal Findings

With this study, we provide empirical insights from the evaluation of a functional prototype of a patient-centric blockchain-based HIE mobile app in a clinical context. While prior studies largely examined hypothetical attitudes of blockchain-based HIE [3,27,28], our evaluation involves real patients undergoing medical treatment in a hospital, realistic usage scenarios, and realistic sample documents related to their care. By developing the blockchain-based HIE mobile app through a user-centered design approach in 3 phases, we incorporated patient feedback at every design stage to ensure that the prototype’s features and user interface aligned closely with patient needs. After iteratively refining the app, we conducted an evaluation with patients with cancer, assessing its usability, perceived value, and overall acceptance based on the SUS and TAM. While this evaluation context naturally limits the extent to which the results can be generalized to all patient groups or long-term use situations, it nevertheless provides a rigorous and informative test case for understanding how patients engage with blockchain-based HIE functionalities in practice. In the following, we discuss the main findings of our study.

First, our findings support the growing body of literature that suggests HIE as an auspicious use case for blockchain-based applications in health care [22]. Similar to prior studies focusing on users’ attitudes [3,15], the participants in our study expressed a positive attitude toward the blockchain-based functions that enable them to define, track, and revoke access to their health data. Although these findings stem from a specific patient population, the underlying interaction required by performing these functions is not disease-specific and reflects fundamental HIE tasks. This suggests that while generalization must be approached cautiously, the interactions we observe may hold relevance beyond the oncology context. The persistence of positive attitudes in our study reveals that a user-centered design process can effectively translate blockchain’s technical characteristics (eg, transparency and immutability) into comprehensible, patient-oriented functionalities. Furthermore, the high PEOU, PU, and BI ratings for the document-sharing function among all presented functionalities highlight the core need addressed by blockchain-based HIE applications, as these applications help to overcome the persistent challenges of managing patient data efficiently and securely in practice [22].

Second, by exploring the rationales behind reported user perceptions, we identified 6 influential factors for the 2 core TAM constructs (ie, PU and PEOU). While the 3 factors influencing PU reflect the specific contextual demands of HIE, the 3 factors shaping PEOU are context-agnostic. This strengthens prior findings, which highlight that PU depends on how well a system (eg, a HIE application) addresses the unique challenges and requirements of a particular context (eg, HIE) or task (eg, sharing a medical document) [54]. In contrast, PEOU reflects users’ perception of how easy and intuitive a system is to learn and operate. It is therefore influenced more by universal design factors that make the system user-friendly rather than its context of use [54].

Third, the age-related differences in perceived usability of our blockchain-based HIE resonate with a broad stream of eHealth and mobile health research indicating that older adults tend to experience higher usability barriers [55,56]. Although not significant, our results indicate a tendency of older adults to rate the overall usability of the mobile app lower than participants in younger age groups. In particular, among the surveyed patients, people from the age groups 60 - 69 and 70 - 79 years rated the mobile app’s usability as satisfactory only, while the younger age groups rated the usability of the mobile app as good. This concurs with previous research on digital health equity issues that emphasizes that older adults are generally less experienced with mobile apps, leading to a lower usability [55,56]. Nevertheless, this also highlights that the development of digital health apps must account for the higher needs of older adults for simplified navigation and intuitive interactions to overcome age-related barriers [57]. Addressing these barriers prevents digital health equity gaps and ensures that they can use and benefit from HIE mobile apps. Toward

XSL•FO
RenderX

that end, we deem user-centered design approaches especially useful since, in our study, for example, we were able to improve the usability ratings in the older age groups through multiple iterations.

Implications

Our study has several implications for research and practice. In terms of research, we add to the growing body of research on patient perceptions of blockchain-based HIE by evaluating the perceptions of cancer patients concerning interactions with a blockchain-based HIE mobile app in a clinical context. Furthermore, we combine quantitative survey results (ie, the SUS scores) with the in-depth insights from a qualitative survey of patients to expand our understanding of patient perceptions of blockchain-based HIE in a naturalistic setting. In particular, the results of the qualitative survey paint a more nuanced picture of context-specific and context-agnostic factors that are crucial for the broad acceptance of blockchain-based HIE. In addition, our study also highlights the value of user-centered design studies in the context of developing and evaluating blockchain-based applications for health care. Although blockchain is primarily a back-end technology, the decentralization of IT systems associated with the use of BC also has an indirect impact on the user experience. Against this backdrop, user-centered design studies can help to better understand user requirements and generate design knowledge for blockchain-based HIE applications.

From a practitioner perspective, our study yields implications particularly for healthcare policymakers and developers of blockchain-based HIE. For healthcare policymakers, the results of our evaluation highlight the potential of blockchain-based HIE to alleviate deficiencies in the current care system, such as lost documents, a time-consuming process to collect and share relevant medical documents, or incomplete health data for patients with cancer. Although many of these challenges are being addressed through the implementation of national health infrastructures, such as electronic health records (eg, [9]), patients remain hesitant to fully engage with these systems due to concerns about privacy, control, or usability, even in contexts where national systems automatically enroll patients unless they actively opt out [58]. Our results indicate that blockchain-based HIE solutions can address these concerns by fostering patient centricity and privacy. Therefore, our results highlight broader dimensions beyond basic usability measures, including patient empowerment and transparency, which are critical for implementing and integrating new technologies into patient-related health care workflows [59]. Participants in our study valued the ability to manage data access and sharing, which strengthened their sense of autonomy and increased the PU of our suggested blockchain-based HIE system. Such systems may be particularly beneficial in less connected areas, such as rural health care settings, where digital infrastructure and interoperability are limited. Pilot projects and model regions in these contexts could provide valuable insights for further scaling and optimizing these systems. Furthermore, for developers of BC-based HIE applications, our study provides a promising design for a BC-based HIE mobile app. When implementing blockchain-based HIE in practice, the identified factors related to the PU and the PEOU should be considered, since both will help ensure that patients use and accept a blockchain-based HIE mobile app. To ensure and improve PU, blockchain-based HIE mobile apps should focus on building functionalities that put all patient data into one overview to reduce time and effort in data sharing. Moreover, medical documents should be shareable individually, and patients should be able to effortlessly establish a connection with health care providers (eg, via a QR code). For ensuring and improving the PEOU, future developments should provide an intuitive design and minimize the number of clicks it takes to achieve a task (eg, share a medical document).

Limitations and Future Research

Our study is not without limitations. First, during the evaluation, patients could ask questions if they felt stuck while interacting with the mock-ups in phase 2 and the mobile app in phase 3. While we tried to provide minimal guidance across all surveyed patients, some asked for more assistance than others. This might have skewed the results in favor of a higher SUS, PEOU, PU, and BI and limited the comparability of patient assessments. Even though participants rarely requested guidance during tasks, the mere presence of help at hand could have led to participants feeling more confident in their interactions, thereby increasing their overall perceptions of the mobile app. Future research could extend our study by examining user perceptions without providing a contact person for assistance. Second, in our collection of patients' responses, we focused on patients with cancer in a single major university hospital in Germany because their treatment requires exchanging patients' health data across multiple health care providers and because their physical and emotional state demands a high usability of the mobile app. This focus might reduce the applicability of our results for other patient groups. Other patient groups might prefer sharing different documents (eg, vaccination record) or might even attach less importance to the concept of blockchain-based HIE since their situation does not demand frequent sharing of documents. Also, patient perceptions might be different when they have positively or negatively experienced the document management process in other hospitals. Future research could hence extend the evaluation of blockchain-based mobile apps to other patient groups and other hospitals. Third, we tried to incorporate a broad range of views in the development of the app through quantitative measures and qualitative insights. However, the convenience sampling used and the limited sample sizes in each phase of the evaluation might have introduced a selection bias and limited the generalizability of our results to a broader population. Future research should validate our findings in a large, balanced sample. This will be especially useful to draw further conclusions about the differences in participants' characteristics (eg, age) influencing patient perceptions (eg, PEOU, usefulness, or acceptance). Finally, we evaluated patient interactions with the blockchain-based HIE mobile app only at 1 point in time. Future research could extend this study by evaluating the interactions with a blockchain-based HIE mobile app over time. Such a longitudinal study could help uncover insights into actual patient sharing behavior and determine whether patients would use a mobile app over a longer timeframe.

XSL•FO
RenderX

Conclusion

With our study, we provide an exemplary case of a blockchain-based HIE mobile app developed through a 3-phase user-centered design approach incorporating on-site patients from a major university hospital. Our evaluation shows that patients in our study would likely use a blockchain-based HIE mobile app to manage their health data, since the app enables them to decide which health care provider or individuals can access their data.

Acknowledgments

We would like to thank all members of the BloG3 research project and the patients who agreed to participate in our study. In proofreading and editing our manuscript, we used the generative artificial intelligence (AI) tools: ChatGPT-4.5 (OpenAI) and Grammarly (Grammarly Inc.). We reviewed and used the suggestions of these tools when they improved readability and clarity and take full responsibility for the content of the publication.

Funding

Funding was provided by the Federal Ministry of Research, Technology and Space as part of the BloG3 research project https://www.blog3.de/(grant number 16SV8372).

Data Availability

The datasets generated or analyzed during this study are not publicly available to ensure the privacy of our participants but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Phase 1 questionnaire and frequency analysis.
[DOCX File, 31 KB - jmir_v28i1e78849_app1.docx]

Multimedia Appendix 2
Screenshots of the blockchain (BC)-based HIE mock-ups.
[DOCX File, 310 KB - jmir_v28i1e78849_app2.docx]

Multimedia Appendix 3
Screenshots of the blockchain (BC)-based HIE mobile app.
[DOCX File, 417 KB - jmir_v28i1e78849_app3.docx]

Multimedia Appendix 4
Phase 3 questionnaire.
[DOCX File, 39 KB - jmir_v28i1e78849_app4.docx]

Multimedia Appendix 5
Phase 3 technology acceptance model (TAM) and System Usability Scale (SUS) frequency analysis.
[DOCX File, 44 KB - jmir_v28i1e78849_app5.docx]

Multimedia Appendix 6
Phase 3 ANOVA results for technology acceptance model (TAM) and System Usability Scale (SUS).
[DOCX File, 4309 KB - jmir_v28i1e78849_app6.docx]

References

1. Kern LM, Bynum JPW, Pincus HA. Care fragmentation, care continuity, and care coordination—how they differ and why it matters. JAMA Intern Med 2024 Mar 1;184(3):236-237. [doi: 10.1001/jamainternmed.2023.7628] [Medline: 38285559]
2. Kern LM, Lau JD, Rajan M, et al. Associations among claims-based care fragmentation, self-reported gaps in care coordination, and self-reported adverse events. BMC Health Serv Res 2024 Sep 10;24(1):1045. [doi: 10.1186/s12913-024-11440-y] [Medline: 39256705]
3. Hau YS, Lee JM, Park J, Chang MC. Attitudes toward blockchain technology in managing medical information: survey study. J Med Internet Res 2019 Dec 9;21(12):e15870. [doi: 10.2196/15870] [Medline: 31815676]

XSL•FO
RenderX

4. U.S. Government Accountability Office. Electronic health information exchange: use has increased, but is lower for small and rural providers. : U.S. Government Accountability Office; 2023 URL: https://www.gao.gov/assets/gao-23-105540.pdf
5. The Office of the National Coordinator for Health Information Technology. Interoperability and methods of exchange among hospitals in 2021. : US Department of Health and Human Services; 2023 URL: https://www.healthit.gov/sites/default/files/2023-01/DB64_Interop_and_Methods_of_Exchange_Among_Hosp.pdf
6. Service (NHS) NH. View your GP health record. NHS England Digital. 2023. URL: https://www.nhs.uk/nhs-services/gps/view-your-gp-health-record/ [accessed 2024-09-23]
7. Walzer S. Digital healthcare in germany. In: Walzer S, editor. Digital Healthcare in Germany: Market Access for Innovations: Springer International Publishing; 2022:5-16 URL: https://link.springer.com/10.1007/978-3-030-94025-6 [doi: 10.1007/978-3-030-94025-6_2]
8. Service (NHS) NH. National care records service. NHS England Digital. 2024. URL: https://digital.nhs.uk/services/national-care-records-service [accessed 2024-12-23]
9. Danish health data authority. National health registers.: National health registers; 2024. URL: https://english.sundhedsdatastyrelsen.dk/health-data-and-registers/national-health-registers [accessed 2024-10-23]
10. Syed R, Eden R, Makasi T, et al. Digital health data quality issues: systematic review. J Med Internet Res 2023 Mar 31;25:e42615. [doi: 10.2196/42615] [Medline: 37000497]
11. Impact c for p. The Electronic Health Records System In the UK.: Centre for Public Impact; 2017. URL: https://centreforpublicimpact.org/public-impact-fundamentals/the-electronic-health-records-system-in-the-uk/ [accessed 2024-12-23]
12. Vest JR, Gamm LD. Health information exchange: persistent challenges and new strategies. J Am Med Inform Assoc 2010;17(3):288-294. [doi: 10.1136/jamia.2010.003673] [Medline: 20442146]
13. Neprash HT, McGlave CC, Cross DA, et al. Trends in ransomware attacks on US hospitals, clinics, and other health care delivery organizations, 2016-2021. JAMA Health Forum 2022 Dec 2;3(12):e224873. [doi: 10.1001/jamahealthforum.2022.4873] [Medline: 36580326]
14. Joo JY. Fragmented care and chronic illness patient outcomes: a systematic review. Nurs Open 2023 Jun;10(6):3460-3473. [doi: 10.1002/nop2.1607] [Medline: 36622952]
15. Lee K, Lim K, Jung SY, et al. Perspectives of patients, health care professionals, and developers toward blockchain-based health information exchange: qualitative study. J Med Internet Res 2020 Nov 13;22(11):e18582. [doi: 10.2196/18582] [Medline: 33185553]
16. Hu S, Schmidt-Kraepelin M, Thiebes S, Sunyaev A. Mapping distributed ledger technology characteristics to use cases in healthcare: a structured literature review. ACM Trans Comput Healthcare 2024 Jul 31;5(3):1-33. [doi: 10.1145/3653076]
17. Lee S, Kim J, Kwon Y, Kim T, Cho S. Privacy preservation in patient information exchange systems based on blockchain: system design study. J Med Internet Res 2022 Mar 22;24(3):e29108. [doi: 10.2196/29108] [Medline: 35315778]
18. Hasselgren A, Kralevska K, Gligoroski D, Pedersen SA, Faxvaag A. Blockchain in healthcare and health sciences-A scoping review. Int J Med Inform 2020 Feb;134:104040. [doi: 10.1016/j.ijmedinf.2019.104040] [Medline: 31865055]
19. Yeung K. The health care sector's experience of blockchain: a cross-disciplinary investigation of its real transformative potential. J Med Internet Res 2021 Dec 20;23(12):e24109. [doi: 10.2196/24109] [Medline: 34932009]
20. Zhuang Y, Sheets LR, Chen YW, Shae ZY, Tsai JJP, Shyu CR. A patient-centric health information exchange framework using blockchain technology. IEEE J Biomed Health Inform 2020 Aug;24(8):2169-2176. [doi: 10.1109/JBHI.2020.2993072] [Medline: 32396110]
21. Mackey TK, Kuo TT, Gummadi B, et al. "Fit-for-purpose?" - challenges and opportunities for applications of blockchain technology in the future of healthcare. BMC Med 2019 Mar 27;17(1):68. [doi: 10.1186/s12916-019-1296-7] [Medline: 30914045]
22. Durneva P, Cousins K, Chen M. The current state of research, challenges, and future research directions of blockchain technology in patient care: systematic review. J Med Internet Res 2020 Jul 20;22(7):e18619. [doi: 10.2196/18619] [Medline: 32706668]
23. Kuo TT, Kim HE, Ohno-Machado L. Blockchain distributed ledger technologies for biomedical and health care applications. J Am Med Inform Assoc 2017 Nov 1;24(6):1211-1220. [doi: 10.1093/jamia/ocx068]
24. Anand G, Sadhna D. Electronic health record interoperability using FHIR and blockchain: a bibliometric analysis and future perspective. Perspect Clin Res 2023;14(4):161-166. [doi: 10.4103/picr.picr_272_22] [Medline: 38025292]
25. Zhuang Y, Chen YW, Shae ZY, Shyu CR. Generalizable Layered Blockchain Architecture for Health Care Applications: Development, Case Studies, and Evaluation. J Med Internet Res 2020 Jul 27;22(7):e19029. [doi: 10.2196/19029]
26. Esmaeilzadeh P. Benefits and concerns associated with blockchain-based health information exchange (HIE): a qualitative study from physicians' perspectives. BMC Med Inform Decis Mak 2022 Mar 28;22(1):80. [doi: 10.1186/s12911-022-01815-8] [Medline: 35346176]
27. Esmaeilzadeh P, Mirzaei T. The potential of blockchain technology for health information exchange: experimental study from patients' perspectives. J Med Internet Res 2019 Jun 20;21(6):e14184. [doi: 10.2196/14184] [Medline: 31223119]
28. Esmaeilzadeh P, Mirzaei T. Role of incentives in the use of blockchain-based platforms for sharing sensitive health data: experimental study. J Med Internet Res 2023 Aug 18;25(1):e41805. [doi: 10.2196/41805] [Medline: 37594783]

XSL•FO
RenderX

29. Lu C, Batista D, Hamouda H, Lemieux V. Consumers' Intentions to Adopt Blockchain-Based Personal Health Records and Data Sharing: Focus Group Study. JMIR Form Res 2020 Nov 5;4(11):e21995. [doi: 10.2196/21995] [Medline: 33151149]
30. Sung M, Park S, Jung S, Lee E, Lee J, Park YR. Developing a Mobile App for Monitoring Medical Record Changes Using Blockchain: Development and Usability Study. J Med Internet Res 2020 Aug 14;22(8):e19657. [doi: 10.2196/19657] [Medline: 32795988]
31. Despotou G, Evans J, Nash W, Eavis A, Robbins T, Arvanitis TN. Evaluation of patient perception towards dynamic health data sharing using blockchain based digital consent with the Dovetail digital consent application: A cross sectional exploratory study. Digit HEALTH 2020;6:2055207620924949. [doi: 10.1177/2055207620924949] [Medline: 32435503]
32. Glomann L, Schmid M, Kitajewa N. Improving the blockchain user experience - an approach to address blockchain mass adoption issues from a human-centred perspective. In: Ahram T, editor. Advances in Artificial Intelligence, Software and Systems Engineering: Springer International Publishing; 2020:608-616. [doi: 10.1007/978-3-030-20454-9_60]
33. Gandhi SS, Patil YN, Netak LD, Gaikwad HR. Usability analysis for blockchain-based applications. In: Kim JH, Singh M, Khan J, Tiwary US, Sur M, Singh D, editors. Intelligent Human Computer Interaction: Springer International Publishing; 2022:349-360. [doi: 10.1007/978-3-030-98404-5_33]
34. Esmaeilzadeh P, Sambasivan M. Patients' support for health information exchange: a literature review and classification of key factors. BMC Med Inform Decis Mak 2017 Apr 4;17(1):33. [doi: 10.1186/s12911-017-0436-2] [Medline: 28376785]
35. Baines R, Stevens S, Austin D, et al. Patient and public willingness to share personal health data for third-party or secondary uses: systematic review. J Med Internet Res 2024 Mar 5;26(1):e50421. [doi: 10.2196/50421] [Medline: 38441944]
36. Abras C, Maloney-Krichmar D, Preece J. User-centered design. In: Bainbridge W, editor. Berkshire Encyclopedia of Human-Computer Interaction Thousand Oaks: Sage Publications; 2004:445-456.
37. Holeman I, Kane D. Human-centered design for global health equity. Information Technology for Development 2020 Jul 2;26(3):477-505. [doi: 10.1080/02681102.2019.1667289]
38. Gao M, Kortum P, Oswald FL. Multi-language toolkit for the System Usability Scale. International Journal of Human–Computer Interaction 2020 Dec 13;36(20):1883-1901. [doi: 10.1080/10447318.2020.1801173]
39. Sauro J, Lewis JR. Quantifying the User Experience: Practical Statistics for User Research: Morgan Kaufmann; 2016.
40. Erler C, Hu S, Danelski A, Stork W, Sunyaev A, Gersch M. Threat modeling to design a decentralized health data management application. In: Rocha Á, Ferrás C, Ibarra W, editors. Information Technology and Systems: Springer International Publishing; 2023:443-455. [doi: 10.1007/978-3-031-33261-6_38]
41. Shostack A. Threat Modeling: Designing for Security: Wiley; 2014.
42. Hyperledger-aries. aries. Hyperledger Aries. 2024. URL: https://github.com/hyperledger-aries/aries [accessed 2025-12-04]
43. LF decentralized trust. Indy. 2024. URL: https://www.lfdecentralizedtrust.org/projects/hyperledger-indy [accessed 2025-12-04]
44. Mobile development with xamarin. .NET.: Microsoft; 2024. URL: https://dotnet.microsoft.com/en-us/apps/xamarin [accessed 2025-12-04]
45. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
46. Venkatesh V, Bala H. Technology acceptance model 3 and a research agenda on interventions. Decision Sciences 2008 May;39(2):273-315 [FREE Full text] [doi: 10.1111/j.1540-5915.2008.00192.x]
47. Harborth D, Pape S. German translation of the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) Questionnaire. SSRN Journal 2018. [doi: 10.2139/ssrn.3147708]
48. Cronbach LJ. Essentials of Psychological Testing, 5th edition: HarperCollins Publ; 1994.
49. Eisinga R, Grotenhuis MT, Pelzer B. The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? Int J Public Health 2013 Aug;58(4):637-642. [doi: 10.1007/s00038-012-0416-3] [Medline: 23089674]
50. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
51. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud 2009 May 1;4(3):114-123 [FREE Full text] [doi: 10.5555/2835587.2835589]
52. Nunnally JC, Bernstein IH. Psychometric Theory, 2nd edition: McGraw-Hill; 1978.
53. Wobbrock JO, Findlater L, Gergle D, Higgins JJ. The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures. 2011 May 7 Presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; Vancouver BC Canada p. 143-146. [doi: 10.1145/1978942.1978963]
54. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified View1. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]
55. Grindrod KA, Li M, Gates A. Evaluating user perceptions of mobile medication management applications with older adults: a usability study. JMIR Mhealth Uhealth 2014 Mar 14;2(1):e11. [doi: 10.2196/mhealth.3048] [Medline: 25099993]
56. Wildenbos GA, Peute L, Jaspers M. Aging barriers influencing mobile health usability for older adults: a literature based framework (MOLD-US). Int J Med Inform 2018 Jun;114:66-75. [doi: 10.1016/j.ijmedinf.2018.03.012] [Medline: 29673606]
57. Gomez-Hernandez M, Ferre X, Moral C, Villalba-Mora E. Design guidelines of mobile apps for older adults: systematic review and thematic analysis. JMIR Mhealth Uhealth 2023 Sep 21;11(1):e43186. [doi: 10.2196/43186] [Medline: 37733401]

XSL•FO
RenderX

58. Griesser A, Mzoughi M, Bidmon S, Cherif E. How do opt-in versus opt-out settings nudge patients toward electronic health record adoption? An exploratory study of facilitators and barriers in Austria and France. BMC Health Serv Res 2024 Apr 8;24(1):439. [doi: 10.1186/s12913-024-10929-w] [Medline: 38589922]
59. Madanian S, Nakarada-Kordic I, Reay S, Chetty T. Patients' perspectives on digital health tools. PEC Innov 2023 Dec;2:100171. [doi: 10.1016/j.pecinn.2023.100171] [Medline: 37384154]

Abbreviations

BI: behavioral intention
HIE: health information exchange
PEOU: perceived ease of use
PU: perceived usefulness
SUS: System Usability Scale
TAM: technology acceptance model

Edited by J Sarvestan; submitted 11.Jun.2025; peer-reviewed by K Mouloudj, Y Zhuang; revised version received 12.Dec.2025; accepted 15.Dec.2025; published 11.Mar.2026.

Please cite as:
Guse R, Hu S, Thiebes S, Erler C, Caridia C, Stork W, Gersch M, Sunyaev A
Patient Perceptions of Blockchain-Based Health Information Exchange: User-Centered Design Study
J Med Internet Res 2026;28:e78849
URL: https://www.jmir.org/2026/1/e78849
doi:10.2196/78849

© Richard Guse, Shanshan Hu, Scott Thiebes, Christina Erler, Catherine Caridia, Wilhelm Stork, Martin Gersch, Ali Sunyaev. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Discourse, Secondary Victimization, and Psychological Harm: Mixed-Methods Analysis of System Justification in the #MeToo Movement

Harsh Parekh[1*], PhD; Shriya Thakkar[2*], PhD; Paras Bhatt[3], PhD; Patricia Akello[4], PhD

[1]Paul H. Chook Department of Information Systems and Statistics, Zicklin School of Business, Baruch College, City University of New York, New York, NY, United States

[2]Department of Prevention and Community Health, Milken Institute School of Public Health, George Washington University, 950 New Hampshire Avenue, Washington DC, DC, United States

[3]Department of Information Systems, Supply Chain & Analytics, College of Business, The University of Alabama in Huntsville, Huntsville, AL, United States

[4]Department of MIS and Cybersecurity, College of Business, University of Montana, Missoula, MT, United States

[*]these authors contributed equally

Corresponding Author:
Shriya Thakkar, PhD
Department of Prevention and Community Health, Milken Institute School of Public Health, George Washington University, 950 New Hampshire Avenue, Washington DC, DC, United States

Abstract

Background: The #MeToo movement, initiated in 2006 and amplified on social media in 2017, mobilized women worldwide to share experiences of sexual harassment and assault online. While the movement increased awareness, it also revealed deep social divisions in digital spaces. Supportive discussions promoted solidarity and healing, whereas antagonistic responses reinforced backlash and secondary victimization. In India, the Indian Entertainment Industry (IEI) became a focal point where survivors' disclosures highlighted structural gender inequalities. These polarized reactions function as digital-health signals, reflecting stigma, psychosocial distress, and conditions that shape women's safety and mental well-being. Examining these narratives as indicators of public health risk helps identify patterns of structural inequity and secondary mental health burdens among survivors.

Objective: This study examined online discourse surrounding #MeToo to identify forms of system-justifying narratives on social media and to assess how #MeTooIndia exposed institutional inequities within the IEI.

Methods: This mixed-methods study comprised 2 components. In study 1, natural language processing was applied to analyze global #MeToo Twitter (subsequently rebranded X) discourse. From an initial corpus of 350,000 tweets, 205,082 were preprocessed, and sentiment and stance detection analysis identified 18,416 tweets expressing negative attitudes toward the movement. Latent Dirichlet allocation topic modeling extracted 22 topics, 12 of which aligned with system-justification categories, revealing distinct lexical and semantic patterns related to gender, institutional, and power dynamics. Two trained coders manually annotated a subsample to ensure conceptual clarity and interrater reliability. Study 2 involved qualitative, semistructured interviews with 20 academic experts in film, gender, and media studies to gather opinions on how #MeTooIndia influenced institutional discourse in the IEI and how these dynamics translate into digital and mental health risks.

Results: Analysis of #MeToo Twitter discourse in study 1 identified 4 primary forms of system justification: by gender, by the institutional system, by backlash, and by victim-blaming. Gender and institutional system justifications were the most prevalent. Study 2 reinforced these findings, revealing how experts perceived #MeTooIndia as both empowering and constrained by entrenched institutional and cultural barriers. Together, our findings highlight the dual function of social media in promoting collective advocacy while reproducing conditions linked to gender-based violence, psychological stress, and reduced help-seeking—key digital and mental health concerns.

Conclusions: This mixed-methods study reveals that digital discourse surrounding #MeToo often sustains existing gender and institutional hierarchies rather than dismantling them. Across Twitter data and expert interviews, gender and institutional system justifications emerged as dominant narratives, highlighting how online spaces can reinforce structural inequities while appearing progressive. Although #MeToo amplified visibility and awareness, its potential for lasting institutional change remains limited. These findings underscore the need for trauma-informed digital governance, public health recognition of online hostility as a psychosocial risk, and frameworks that situate digital activism to institutional reforms that support safety and mental well-being.

(J Med Internet Res 2026;28:e75533) doi:10.2196/75533

XSL•FO
RenderX

KEYWORDS

harassment; mental health; inequities; gender; mixed methods

Introduction

Background

Initially founded by the New York-based social activist Tarana Burke in 2006, #MeToo aimed to empower women and support victims of sexual assaults in the United States. A decade later, the movement gained worldwide attention when actress Alyssa Milano popularized #MeToo with a viral Twitter (subsequently rebranded X) trend. The movement broadened awareness while reflecting the emotional strain tied to large-scale public disclosure [1,2]. From a public health perspective, this increase represents more than social awareness. It reflects large-scale digital expressions of trauma that reveal population-level exposure to gender-based violence (GBV) and related mental health consequences [3-5].

Gender equality movements such as #MeToo are often viewed as women-led initiatives [6], which shapes how people engage with them online. This framing also explains why public responses became sharply divided or polarized. Expressions of solidarity promoted healing and collective efficacy [5,7,8], while hostile or dismissive comments reproduced stigma and secondary victimization [9]. Prior analyses have shown that discussions around #MeToo contained both destructive elements, such as ridicule, rape references, and scandal language, and constructive ones emphasizing education, leadership, and rights [5,9]. Such polarization is health-relevant because antagonistic discourse can intensify psychological distress, discourage help-seeking, and perpetuate unsafe disclosure climates for survivors [3,10-12]. Opponents of #MeToo viewed the movement as threatening social identity and existing hierarchies [13,14], further sustaining online environments that can undermine survivor well-being and collective healing.

In this context, while the #MeToo movement resonated worldwide, its impact on the Indian Entertainment Industry (IEI), particularly Bollywood, was notably profound, serving as a catalyst that broke long-standing barriers of silence within the entertainment sector [15]. Within the IEI, disclosures emerged rapidly and publicly, revealing deep-seated gender hierarchies and power asymmetries that had long silenced survivors. For instance, late Bollywood choreographer Saroj Khan remarked, “Bollywood provides bread, at least doesn’t abandon women after rape” in the wake of the #MeToo movement in India [16]. Such comments exemplify how public figures can normalize or excuse sexual violence, reflecting the cultural narratives that sustain institutional denial and perpetuate psychological harm for survivors [17]. These justifications and dismissive reactions have measurable health implications. Studies show that negative or invalidating responses to #MeToo disclosures are associated with increased posttraumatic stress, anxiety, and feelings of isolation among survivors [18,19].

In the Indian context, such backlash often manifests through public shaming, social exclusion, and reputational damage, creating conditions of secondary victimization that discourage reporting [17,20,21]. Although social media facilitates visibility and mobilization, it also amplifies exposure to stigma and retraumatization, turning disclosure spaces into sites of both empowerment and psychological risk [1]. Thus, our analysis addresses calls to understand the system-justification narratives within #MeToo that reproduce structural inequities and generate secondary mental health burdens for survivors [9,19,22].

Alongside social and moral support, #MeTooIndia also generated negative reactions in the form of backlash, victim blaming, and social system justification [23]. In patriarchal contexts, women survivors often face reputational risks and moral scrutiny that discourage disclosure and silence voices of resistance [20,24]. Backlash also stemmed from a perception that #MeTooIndia threatened group identity and the existing social order, prompting many to defend traditional hierarchies through system-justifying narratives [25,26]. Previous research on #MeToo links disbelief and ridicule following disclosure to trauma recurrence, anxiety, and avoidance behaviors [9,10,22]. This study examines online reactions to #MeToo through the lens of a digital health discourse. Social-media discourse surrounding GBV functions not only as a forum for advocacy but also as a public-health signal that reflects collective trauma and emotional distress [4,18]. Such digital expressions can inform health care providers, mental health professionals, and public health practitioners about emerging patterns of psychosocial stress and the need for trauma-informed care and community-level interventions [27,28].

We treat the justification narratives surrounding #MeToo as infoveillance indicators, which are digital signals of GBV and secondary victimization addressing limitations identified in prior research [5,29]. Hostile or invalidating comments constitute secondary victimization [30,31], which research links to increased posttraumatic stress, anxiety, and social withdrawal among survivors [9,11,19,22]. Within patriarchal or celebrity-dominated contexts such as India’s entertainment industry, such backlash can reinforce stigma and reproduce structural inequities that exacerbate mental health burden [17,20]. Positioning the analysis within this digital-health framing enables examination of how social-media environments impact both public-health awareness and secondary psychological harm following GBV disclosure [3].

This Study

Overview

Victim blaming stands as a key factor deterring women from coming forward or identifying their aggressors [5,32-34]. Building on this context, the current study focuses on entrenched gender hierarchies, institutional status quo, pushback on social movement, and victim-blaming as mechanisms of social system justification within the #MeToo movement. In extending prior work, our study moves beyond the initial disclosure phase to examine how system-justifying narratives evolve as digital signals of GBV [4,29,35,36]. Study 1 conducts a large-scale text mining analysis of Twitter data to identify and categorize system-justifying narratives that emerged during the #MeToo

XSL•FO
RenderX

movement. Study 2 complements this with qualitative interviews focused on the IEI, offering deeper insights into how these narratives manifest within a specific sociocultural and institutional setting. Together, these studies offer a mixed-methods perspective on how digital backlash and justification discourse operate as online expressions of GBV, with implications for mental health and digital well-being.

Objectives

The objectives of this study were to (1) identify the forms of system-justifying discourse that emerged on social media during the #MeToo movement and (2) examine how #MeTooIndia revealed its effects on institutional systems within the IEI.

Conceptualization: Social Media as a System Justifier

Social conditioning often creates circumstances in which seeking justice can, unfortunately, intensify the harassment experienced by victims [5,22,37]. Secondary victimization occurs through social media when the perpetrators face legal problems, and the community unjustly holds the victim responsible [38,39]. System justification theory suggests that "there is a general ideological motive to justify the existing social order," and this motive is responsible for the "internalization of inferiority among the members of the disadvantaged group" [40]. System justification theory has been applied across disciplines, including public health, using established scales to assess how individuals endorse or legitimize existing social systems [41,42]. An example of a belief measure is the view that society is structured so that people generally get what they deserve. It is important to distinguish psychological system justification, which reflects unconscious cognitive motives to legitimize existing hierarchies, from strategic or systemic resistance, which involves deliberate actions by individuals or institutions to preserve power and resist reform [43,44].

System justifiers create a fear of judgment or retaliation for sexual victims, drastically reducing the disclosure [19,45,46]. A substantial number of tweets criticized individuals, including prominent figures, who had blamed victims of sexual harassment for their experiences, thereby undermining the feminist movement [22,47]. The inclination to rationalize current social structures may be particularly strong in the context of gender dynamics compared with other intergroup settings, such as racial, religious, or ethnic groups, as gender relations are deeply interdependent within social and economic systems [25]. The types of system justifications are listed in Table 1.

Table . Type of system justifications.

Justification type	Definition
Gender system justification	The "extent one accepts gender hierarchy as legitimate" (page 1). This includes upholding traditional gender roles, norms, and power dynamics within the system [48].
Institutional system justification	This type of system justification focuses on the "legitimation of institutions and authorities, denial or minimization of system problems or shortcomings" (page 881). It involves defending workplaces, industries, legal systems, or social hierarchies [40].
Victim-blaming system justification	This type of justification is known as the derogation effect. It indicates that "victim is deserving of the injustice and responsible for his/her own misfortune" (page 5912). This behavior often involves questioning the credibility of the victim, scrutinizing their agency, and fabricating their experiences [39].
Backlash system justification	This type of justification indicates "vociferous and frequently violent opposition" toward "any actual or potential gains made by disadvantaged groups" (page 133-134). This behavior includes disapproving of facts and studies recognizing the prevalence of sexual harassment, highlighting the threat to individual freedom, and sympathizing with defamation received by the accused [49].

Similarly, investigating the 2017 Women's March movement, researchers found that mostly in the case of men (who identify gender as less important to one's self-concept), exposure to gender movement increased system justification tendency [48]. Furthermore, men who expressed support for gender system justification believed that #MeToo unjustly tarnished the reputation of men. Evidence indicates that the endorsement of sexist ideologies is connected to individuals' perspectives on sexual harassment and their stance toward the #MeToo movement [50].

While defending the existing gender system may represent the main narrative for system justification within such movements, it is not the only type represented on social media. Reports have highlighted the limited accountability of businesses and institutions during the #MeToo movement [51]. Sociocultural and organizational systems often reinforce beliefs and behaviors that perpetuate sexual harassment and misconduct. Furthermore, victims bear the personal cost of legal representation, while institutions support the defense of powerful offenders. Such dynamics contribute to secondary victimization and institutional retaliation, maintained through system-justifying mechanisms [52].

Witnessing online victimization has been shown to promote dehumanization of victims, leading to perceptions that they are less deserving of respect or protection [21,53]. Such cognitive and emotional distancing contributes to harmful social norms that normalize violence and reduce collective empathy, which are both recognized as public health concerns [10,19].

XSL•FO
RenderX

Victim-blaming justifications are produced when individuals experience identity threats and defensively protect their self-image by partially or fully attributing blame to the victim [5,18]. These justifications are reinforced by gender-powered dynamics and male sexual aggression, which in turn increase women's vulnerability and sustain beliefs rationalizing the just world hypothesis [53-55].

Victim-blaming justifications not only shift attribution away from perpetrators but also contribute to resistance against the broader goals of the #MeToo movement. This form of discourse often frames the movement as an overreaction or a threat to individual freedom, thereby undermining efforts to address GBV and promote safer social environments [5,17]. Such responses reflect a broader tendency to preserve positive perceptions of existing social systems, even when these systems sustain inequities that have significant public health implications for victims' mental and physical well-being [19,29,49]. Earlier research indicates that the movement could provoke even more intense negative responses from those who are unsympathetic or hold contrary beliefs about its objectives [5,36,48].

Understanding how systems justify and reproduce themselves is key to identifying the social and psychological barriers that impede collective action. From a public health perspective, such insights can guide targeted strategies to disrupt these barriers, strengthen community engagement, and sustain online activism that promotes equity and safety.

Methods

Study Design

The global #MeToo movement revealed significant regional variation in engagement and public discourse [35,56]. Recognizing that social movements addressing GBV have broad implications for population health and social well-being, we adopted a mixed-methods design to examine how digital narratives reinforce or challenge systemic inequities. This study comprised two components: (1) a quantitative analysis of tweets to assess the global diffusion and framing of #MeToo via social media, and (2) a qualitative study using in-depth interviews focused on #MeTooIndia to contextualize the movement's local impacts.

This research used a sequential explanatory mixed-methods design in which study 1 directly informed study 2 [57]. The quantitative analysis of Twitter discourse identified 4 dominant forms of system justification, and these empirically derived categories guided the development of the semistructured interview guide (Multimedia Appendix 1). Study 2 was therefore designed to contextualize, elaborate, and extend the patterns observed in study 1 by examining how similar justification mechanisms were understood within the localized context of the IEI. During interpretation, findings from both studies were integrated to provide a consolidated understanding of how digital narratives and institutional responses interact during #MeToo.

In the first study, we collected tweets without geographic restrictions to minimize sampling bias and capture cross-cultural expressions of system justification. Regional and cultural contexts can shape how individuals legitimize or contest social hierarchies online, influencing public perceptions of gender justice and safety—key determinants of health equity. Guided by system justification theory, which posits that individuals tend to rationalize and defend existing social structures even when they sustain inequality, we used natural language processing–based topic classification to identify patterns of justification within global #MeToo discussions. We then complemented these findings with an in-depth case study of #MeTooIndia, drawing on expert interviews to explore how these justifications manifest within the Indian sociocultural and institutional context, which led to our second study.

In our second study, we examined the social and behavioral consequences of the #MeTooIndia movement within the context of IEI. Prior research underscores that diverse cultural settings can obscure interpretation unless analyzed within a specific context [58]. IEI, particularly Bollywood, offered a compelling case as the movement gained strong momentum, generated widespread public engagement, and empowered women actors to voice experiences of sexual misconduct. To capture these dynamics, we conducted interviews with research scholars specializing in media or film studies to gain nuanced insight into the social and institutional responses surrounding #MeTooIndia.

While the 2 studies used different methods, they collectively offer a multidimensional understanding of digital activism and its societal implications. The natural language processing–based analysis in study 1 captures large-scale public discourse and patterns of system justification, whereas the expert interviews in study 2 provide contextual depth and interpretation of those narratives. Together, these complementary approaches offer a more comprehensive view of how such digital feminist movements intersect with cultural norms, institutional accountability, and collective well-being. Consistent with Gorissen et al [22], who found that Twitter hosted 61.4% of sexual victimization disclosures during #MeToo, our focus on Twitter data ensured a consistent and extensive dataset for examining system-justifying narratives that influence public attitudes toward gender equity and safety.

Transparency and Openness

The processed data supporting our conclusions, along with the Python (Python Software Foundation) syntax required for reproducing the results, can be obtained from the authors upon request. All analyses were performed using Python libraries. Please note that this study's design and analysis were not preregistered because the research was exploratory and aimed at developing theoretical and methodological insights rather than testing predefined hypotheses (Figure 1).

XSL•FO
RenderX

Figure 1. Overall methodological flow illustrating the mixed-methods design across study 1 and study 2.

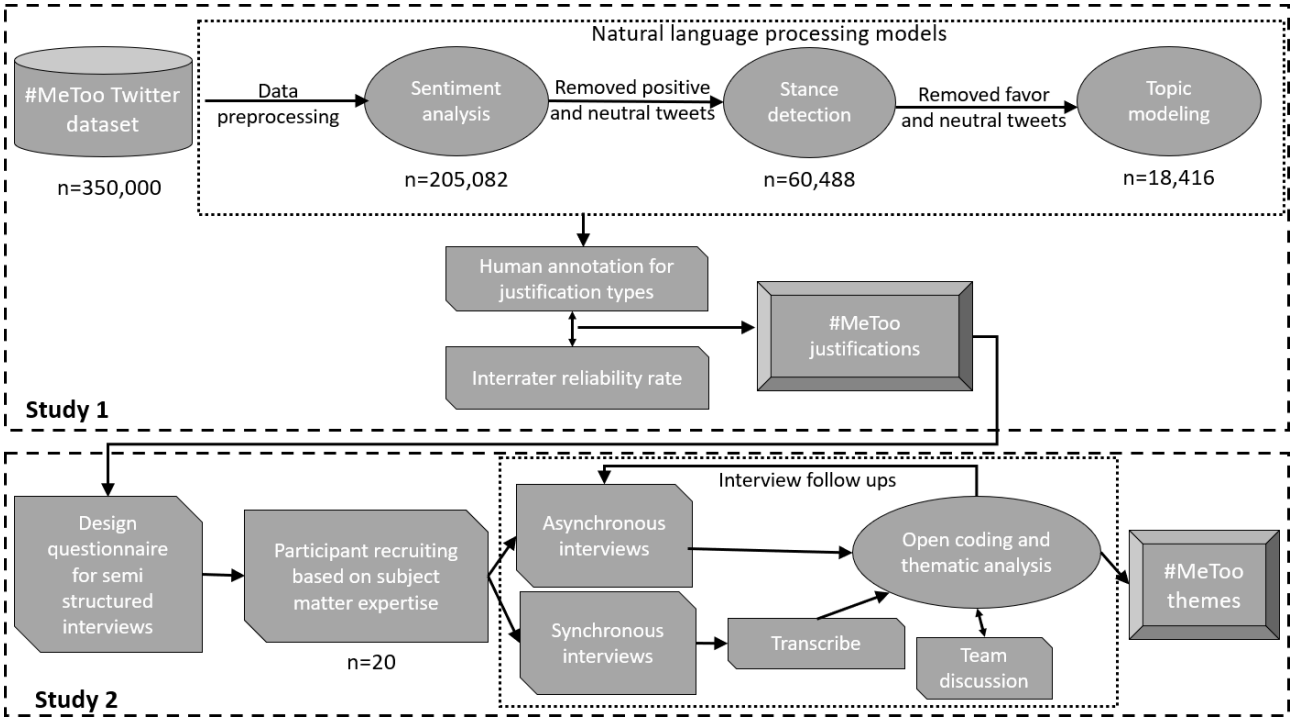

Ethical Considerations

For quantitative study 1, all data were obtained from publicly accessible sources on Twitter. Posts shared under the hashtags (#MeToo and #metoo) represent public communications intended for open audiences. No personal or identifying information about individual users (such as usernames, user IDs, or direct mentions) was retained, as all data were deidentified before analysis. All analyses were conducted on aggregated text data, focusing on linguistic and thematic patterns rather than individual users or specific tweets. Results are therefore presented in group form to ensure that no inferences can be made about any specific user. Given ongoing concerns about the potential for reidentification of social media users, even when direct identifiers are removed, we took additional precautions when presenting illustrative material (sample tweets). All sample content used for explanation in this paper was paraphrased, and any original handles were replaced with neutral placeholders to prevent traceability. Finally, while we do not publicly share any of the analyzed data, we outline more details in the "Data Availability" section [59,60].

For study 2, all participants were experienced academics familiar with research ethics and the principles of informed consent and institutional review. Each received an information sheet describing this study's objectives, procedures, and data-use policy. Participation was voluntary. Participants were interviewed strictly in their professional capacity, and no identifiable personal data was collected. The authors analyzed qualitative materials that were recorded in a fully anonymized form. This study did not involve interaction or intervention beyond the collection of anonymous responses. Accordingly, the work met criteria for minimal-risk social research and did not constitute human participants research requiring institutional review board oversight under the US Common Rule (45 CFR §46.102) [61]. Additionally, the research adhered to the ethical principles of the Declaration of Helsinki [62] and the World Health Organization's (2001) guidelines for ethical qualitative inquiry [63].

To protect privacy and confidentiality, all interviews were conducted and recorded in English with explicit participant consent. Transcripts were deidentified, assigned numeric codes, and stored on password-protected devices accessible only to the research team. No identifiable information, images, or supplementary materials are included in this paper. Participation was entirely voluntary, and no monetary compensation was provided. Table 2 summarizes the demographic characteristics of the scholars and experts who participated in this study.

XSL•FO
RenderX

Table . Demographic profile of scholars and experts interviewed for study 2 (N=20). The names of the respondents referred to in this research are pseudonyms, and identities are kept strictly confidential to maintain anonymity. Professional or educational engagement and country of professional or educational attainment are as reported.

Name	Identified by the pronouns used	Current professional or educational engagement	Country of professional or educational affiliation
Daniel	He or his	Professor	United States
Smita	She or hers	Associate professor	United States
Sarita	She or hers	Associate professor	United States
Raya	She or hers	Professor	United States
Ramya	She or hers	Assistant professor	India
Irrfan	He or his	Associate professor	Turkey
Karanjeet	He or his	Assistant professor	India
Latika	She or hers	Doctoral candidate	United Kingdom
Nazreen	She or hers	Doctoral candidate	United States
Naved	He or his	Doctoral candidate	Turkey
Mita	She or hers	Assistant professor	United States
Neeta	She or hers	Graduate student	United States
Kirti	She or hers	Assistant professor	India
Kamal	She or hers	Assistant professor	United States
Rahul	He or his	Doctoral candidate	United States
Smitra	She or hers	Graduate student	India
Tupur	She or hers	Doctoral candidate	India
Ipsita	She or hers	Graduate student	India
Ayanti	She or hers	Graduate student	India
Jincy	She or hers	Doctoral candidate	India

Study 1: Quantitative Twitter Analysis

Study Design and Data Collection

We used publicly accessible tweets as the main source of our data for conducting a topic-wise classification. We used the Twitter API platform to extract tweets from #metoo and #MeToo for a period of 1 month from November 2017 to December 2017. The extraction resulted in a corpus of 350,000 tweets. This dataset provides a rich corpus for examining public discourse and sentiment around the movement, offering important insights into how individuals communicate experiences of GBV and solidarity in digital spaces. The data included several identifiers such as postdate, user description, and engagement metrics about the likes, retweets, and replies, allowing us to examine both message content and patterns of social amplification.

The global #MeToo conversation began on October 15, 2017, when Alyssa Milano invited survivors of sexual harassment to share their stories publicly. The movement quickly transcended linguistic and cultural boundaries, inspiring localized expressions such as *#YoTambien* in Spanish-speaking countries, *#BalanceTonPorc* (“squeal on your pig”) in France, *#quellavoltache* in Italy, and *#RiceBunny* in China—a creative response to online censorship. We strategically collected the data after the first sporadic spike (which turned out to be the highest over time). Figure 2 shows Google Trends (Google LLC) in the #MeToo movement. As we wanted to capture various system justifications, the period after the first sporadic spike was deemed appropriate [22].

XSL•FO
RenderX

Figure 2. Temporal dynamics of Google search interest in #MeToo from November 2017 to December 2017.

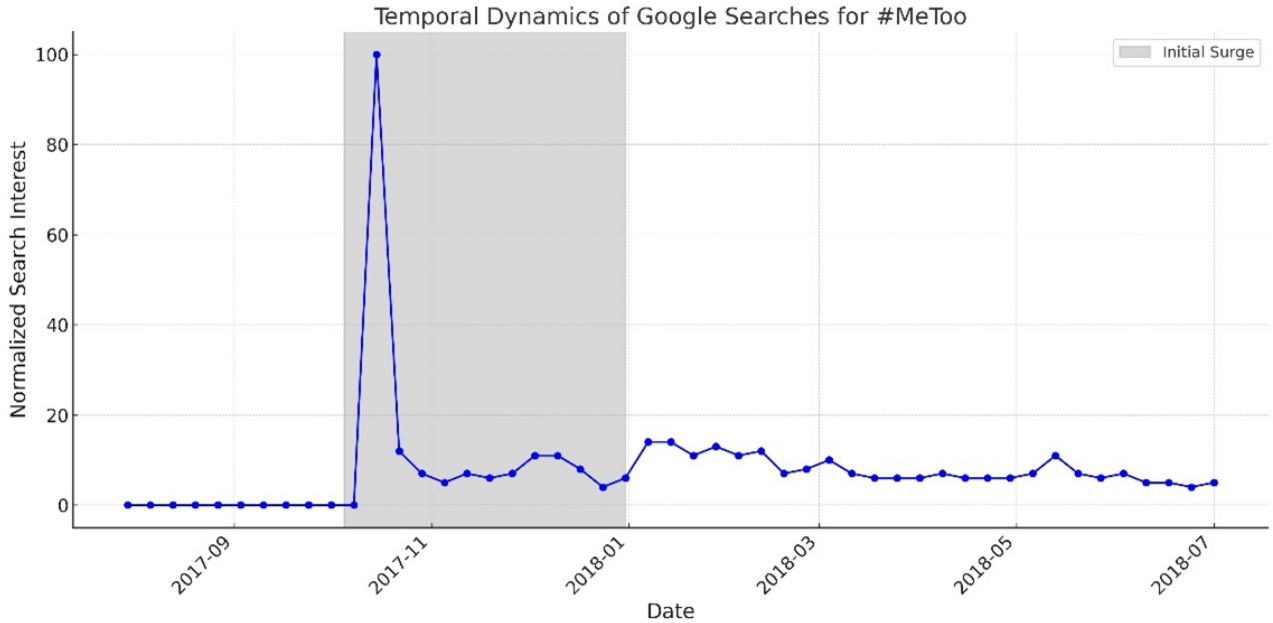

Textbox 1 showcases exemplar tweets illustrating the various justification types that serve to diminish the impact of the #MeToo movement. All tweets in Textbox 1 are paraphrased and deidentified to prevent traceability to individual users. Simple removal of identifiers is insufficient, as advanced fuzzy-matching and text-retrieval techniques can enable reidentification of social media posts. To protect user anonymity, we replaced all handles with neutral placeholders (eg, ABC, @DEF, and @GHI) and reworded examples while preserving their conceptual meaning.

Textbox 1. Sample tweets for justification types.

- Gender system justification: “we have to "go back" to giving men authority over women in the workplace, home & society in general so they will STOP taking advantage #MeToo.”
- Institutional system justification: “@ABC You've been in #Hollywood for 60 years, you've worked with, kept quiet about, helped to cover up, so why now?”
- Victim-blaming system justification: “@DEF Yeah it works in mysterious ways because she knew he was a sexual predator and still worked with him”
- Backlash system justification: “@GHI: Texas Attorney General's top aide mocks 'pathetic' #MeToo movement and calls women's marchers 'c*nts'”

Analysis

Preprocessing of the #MeToo Twitter dataset was essential to ensure data quality and analytical reliability (Multimedia Appendix 2). We streamlined the data by performing initial cleaning and normalization tasks, such as removal of URLs, hashtags, and user mentions, correction of typographical errors, and standardizing text to a uniform lower case, and then further refined our dataset by eliminating special characters, tokenizing text, and handling abbreviations. These comprehensive steps concentrated our analysis on 205,082 tweets, priming them for further analysis. In the sentiment analysis phase, we used the Python package *TextBlob* to perform sentiment detection on the subset of 205,082 tweets identified postinitial preprocessing.

This analysis classified the emotional tone of each tweet into positive, neutral, or negative categories. Subsequently, we refined our dataset by excluding tweets that exhibited positive and neutral sentiments. We drew from the work of Reyes-Menendez et al [64], which used n-grams analysis to reveal a dual framing of the #MeToo movement—one destructive, associated with terms such as cowards, rape, and scandals, and another constructive, linked to words such as educate, leader, and rights.

For our study of justification narratives, we focused on the former, harnessing sentiment analysis and stance detection techniques. This allowed us to refine our dataset to specifically include those tweets that aligned with destructive negative discourse, providing a pertinent sample for our analysis of justification cases. This filtration process resulted in a concentrated sample of 60,488 tweets, each expressing a negatively worded sentiment, which provided a more precise basis for examining the adverse reactions associated with the #MeToo movement. After filtering our dataset to include only the negatively worded tweets, we used the *NLTK* Python package for stance detection to categorize each tweet’s attitude toward the #MeToo movement as supportive, neutral (or factual), or opposing. We then removed tweets with supportive and neutral stances, resulting in a final dataset of 18,416 tweets that explicitly opposed or criticized the movement by justifying

XSL•FO
RenderX

existing social norms, thereby focusing our analysis on expressions of dissent.

We implemented a 2-pronged content analysis strategy to examine the semantic structures within our tweet dataset. First, we used latent Dirichlet allocation topic modeling, which identified and extracted 22 distinct topics representing underlying themes in the discourse. Upon further analysis, we found that 10 of these topics were not directly aligned with any justification narratives, while the remaining 12 topics were systematically categorized into specific justification categories.

Each justification type exhibited distinct lexical and semantic signatures. From the latent Dirichlet allocation topic modeling results, we observed that the topics associated with each system justification type contained unique keyword structures indicative of the underlying sentiment. For instance, "backlash system justification" topics included words such as "movement," "vicious," and "sex," hinting at an undercurrent of resistance to the #MeToo narrative. "Victim blaming justifications" were laced with terms such as "rape," "society," and "refuse," reflecting accusatory and deflective tones. Meanwhile, "institutional justification" topics incorporated institutional and procedural terms such as "resign," "misconduct," and "allegation," suggesting a focus on systemic and structural dimensions. Lastly, "gender system justification" topics distinctly featured "woman" and "man" in their discourse, emphasizing the gendered aspects of the conversation.

These keyword patterns are illustrative of the diverse justifications and sentiments present within the #MeToo-related discourse captured in our dataset. In the second phase, we complemented the computational results with human-coded annotation to refine the justification categories within our tweet sample. Specifically, we recruited 2 trained annotators to collaboratively analyze a representative set of 250 tweets, with the aim of establishing a shared conceptual framework for the various categories of justifications. This process was guided by detailed annotation protocols to ensure conceptual clarity, coding consistency, and interrater reliability.

The interrater reliability, measured using Cohen κ, was found to be 0.91, reflecting a high degree of agreement between the annotators and confirming the robustness of the annotation schema. With this validated framework, we proceeded to annotate a larger subset of 18,416 tweets. We used the insights gained from the initial 250 tweets to train a supervised machine learning model, thereby scaling our classification effort across the entire dataset. This model was designed to categorize tweets into 1 of 4 that justification types or into a fifth category indicating the absence of a discernible justification.

This process allowed us to accurately distribute our sample into proportionate classifications of each justification type, thus enabling a comprehensive and quantifiable analysis of the dominant narratives in the relevant tweets. Discrepancies between annotators were resolved through discussion and iterative refinement of the coding guidelines, resulting in strong interrater agreement and improved clarity in classification criteria. The precision of our classification framework established a robust foundation for subsequent analyses and interpretations of the data. Across the 18,416 negatively valenced tweets, institutional system justification was the most prevalent (7366/18,416, 40%), followed by gender system justification (6814/18,416, 37%), victim-blaming justification (2210/18,416, 12%), and backlash justification (2026/18,416, 11%).

Study 2: Qualitative Thematic Analysis

Study Design and Recruitment

Study 2 builds upon the findings of study 1 by incorporating perspectives from academic experts in the field. We conducted 20 in-depth interviews with scholars and researchers specializing in gender, media, and film studies—particularly those focused on Hindi cinema and the Indian entertainment context. The objective was to understand how #MeTooIndia influenced discourse and practices within the IEI. In-depth interviews are a well-established qualitative method for obtaining detailed insights into participants' beliefs, experiences, and interpretations of a given phenomenon [65-67]. We targeted academics and scholars because of their capacity to provide theoretically informed and relatively impartial reflections on the movement—an important complement to journalistic and activist perspectives that have dominated public narratives.

Using purposive sampling [68], the current study recruited professors, research scholars, and doctoral candidates affiliated with universities and research institutions across the United States, India, the United Kingdom, and the Middle East who specialized in media and film studies, with a focus on Indian cinema. Potential participants were identified through prior publications and conference presentations in feminist media studies, the IEI, social-media activism, and cinema or film studies. We emailed 80 eligible academics, describing this study's objectives and inviting them to participate in an interview. Of these, 29 responded, and appointments were successfully scheduled with 20 (36%) participants. The remaining 9 (11%) participants did not provide availability. Ultimately, 20 (25%) participants of the invited pool completed this study and formed the final analytic sample.

Recruitment was performed through personalized email invitations sent directly to eligible participants, describing this study's objectives, the voluntary nature of participation, and procedures for ensuring confidentiality. Those who agreed to participate provided written informed consent electronically before interviews. Data were collected between January and May 2021, a period shaped by the COVID-19 pandemic, which redefined research logistics and accelerated the adoption of virtual data-collection tools. We conducted 13 synchronous interviews via Zoom and 7 asynchronous interviews through email correspondence, allowing for flexibility in scheduling, geographic reach, and iterative follow-up questions when deeper elaboration was required. This mixed interview format provided both immediacy and reflection, aligning with evolving qualitative research practices during pandemic conditions [69-71].

Analysis

All interviews were transcribed verbatim and imported into NVivo 12 (QSR International) to facilitate systematic coding and data management. Analysis followed an inductive thematic

XSL•FO
RenderX

approach using the constant comparison method [72]. Each transcript was reviewed line by line to identify recurring ideas, patterns, and contradictions both within and across cases. In the first coding cycle, each author independently performed open coding to generate preliminary labels grounded in the participants' language. After multiple reviews of the transcripts, codes such as victim blaming, system, and institutional justification began to emerge. The second coding cycle involved axial coding, where similar concepts were grouped and relationships between categories were explored [73].

To enhance methodological rigor, intercoder reliability was assessed by double-coding a subset of transcripts and comparing agreement rates. Intercoder reliability is crucial as it showcases the methodological approach of achieving consensus in coding among the research team [74]. Discrepancies were resolved through discussion until consensus was reached, and a finalized codebook was developed. Once coding consistency was achieved, all transcripts were recoded using the agreed codebook, and thematic clustering was conducted within NVivo to synthesize findings. Analytical memos were maintained throughout to ensure transparency of interpretation and reflexivity among the research team.

Results

Study 1: Quantitative Study

Gender system justification emerged as the most prevalent category, representing 37% of the dataset and reflecting discourse that reinforces established gender hierarchies and traditional gender norms. Institutional system justification accounted for 40% (7366/18,416), characterized by narratives that defend organizational practices or minimize institutional accountability in cases of harassment and misconduct. Smaller yet meaningful proportions included backlash system justification (2026/18,416, 11%) and victim-blaming system justification (2210/18,416, 12%), illustrating persistent efforts to discredit the #MeToo movement or to shift responsibility onto victims. Together, these findings provide a quantitative overview of how online conversations reproduce societal power structures that sustain gender-based inequities—factors with significant implications for public health, safety, and psychosocial well-being.

The distribution captures the relative prevalence of each justification type, offering a data-driven perspective on the broader discourse surrounding gender justice and institutional accountability. Thus, study 1 offers a global quantitative map of justification narratives, while study 2 qualitatively examines these dynamics within #MeTooIndia in the IEI, illustrating how cultural norms and institutional structures shape experiences, responses, and barriers to gender equity and collective well-being.

Study 2: Qualitative Data Collection

Overview

The final thematic analysis yielded 3 dominant themes (eg, gender system justification, victim blaming, and institutional system justification) reflecting the influence of #MeToo within the IEI, paralleling patterns identified in the quantitative Twitter analysis.

Gender System Justification

Despite #MeToo's global aim to promote feminist awareness and counter sexual harassment, most participants observed little progress within the IEI, noting that existing gender hierarchies largely remain intact. This is an important concern for women's digital and societal well-being, including their mental health and everyday safety, as entrenched gender norms shape vulnerability to harassment and constrain pathways to speak up or seek support. According to Daniel, a media professor from the Pacific West, United States:

> *I think the #MeToo movement added some urgency to an already-ongoing set of changes, which I would characterize as an evolution rather than a revolution. Yes, there has been more diversity in representations of women, more women characters who exert greater autonomy and mastery, but there is still a tendency to show traditional kinds of stereotypical femininity as worthy of reward.*

While Daniel highlighted the persistence of gendered portrayals, Smita, a professor of cinema and photography of Indian American descent from the Midwest, elaborated on the deeper structural inequities within the industry:

> *It has brought attention to sexual exploitation within the industry, made public what insiders knew—but I would not go so far as to say, it has revolutionized anything. There is a great deal of power inequality in the industry; a lot of insecurity produced by having to stay on the right side of those in power (monopoly of a few big names); networks are brutal. Without any of these changes, things will not change much.*

Several notable figures in the IEI attributed the ongoing challenges not only to the industry but also to the broader social system. For instance, Bollywood producer Vinta Nanda expressed that this issue is "not a Bollywood habit but a very Indian habit" [75]. Echoing this sentiment, Latika, a doctoral candidate pursuing her PhD in the United Kingdom, blamed the current political party in power, stating that:

> *#MeToo hasn't been able to revolutionize Bollywood, not to the extent needed, but then India is currently under an unprecedented repressive and gender-violent religious fundamentalist regime, which acerbates the problem.*

Irrfan, a professor of media and communications from the Midwest United States, offered a structural critique of #MeToo:

> *Social movements that are primarily organized by online campaigning often capture public attention quickly. Online users could effectively bring light to social injustice and accumulate public support in a short period. However, such organic and leaderless movements often do not have sufficient organizational capabilities to challenge the existing gender system and keep pushing for changes that could potentially challenge our society's power-relations.*

XSL•FO
RenderX

Irrfan highlighted a central limitation of #MeToo as an "organic" and "leaderless" movement, which has struggled to confront the gendered power dynamics within India's sociostructural context. Scholars argue that the movement has largely amplified the voices of a privileged minority while neglecting those of marginalized and precariously employed women [76]. Others suggest that online social movements promote broader engagement than traditional offline activism, as digital platforms enable individualized participation and storytelling in place of hierarchical leader-driven conventional movements [77]. Additionally, the IEI is not governed by formal workplace policies or oversight mechanisms comparable to other sectors, leaving issues of women's safety and protection as weakly regulated and contributing to increased mental health concerns. Overall, the expert testimonies argue for treating GBV in the IEI as occupational and digital health concerns that require structural interventions: standardized industry policies, independent redress mechanisms, survivor-centered mental health services, and safer digital reporting infrastructures. They also highlight the need for coordinated platform-industry-state action to ensure that the visibility generated by movements such as #MeToo leads to durable protections for women victims, rather than remaining confined to short-lived, elite-driven online attention.

Victim Blaming System Justification

Although #MeToo gained visibility in India, women victims who disclosed experiences of harassment frequently encountered backlash and blame from influential figures seeking to silence their voices. Similar to the #SurvivorPrivilege hashtag movement in 2014, Indian media narratives during #MeToo reinforced victim-blaming discourse, depicting survivors as manipulative or vindictive actors intent on falsely accusing men [9]. A widely publicized example involved actress Tanushree Dutta, who relocated to the United States after facing intense public hostility and media scrutiny following her #MeToo disclosure. Experts noted that these digital reactions operate as forms of secondary victimization, generating emotional strain and shaping whether women feel safe seeking support or speaking publicly. When asked about the movement's effectiveness in confronting India's long-standing culture of victim blaming, Daniel commented:

> *As long as victim blaming and victim shaming still get treated seriously in the press (and the courts), my answer to that question is, it has not gone far enough. Those who have been abused and exploited need to continue to speak up; witnesses need to be spoken up; bystanders need to be ready to believe uncomfortable things about people they know. It is worth noting that this need not be malicious to be problematic—some men who are not themselves abusers are likely trying to make sense of the idea that men with whom they have worked, their friends and role models, potentially, have engaged in immoral behavior. They try to resolve that uncomfortable inconsistency between what they think they know and what they're hearing by rejecting the new information.*

Daniel emphasized that countering victim-blaming requires empowering women to share their stories openly and fostering conversations that challenge victim shaming. He highlighted the need to expose offenders, create safe spaces for women to speak about their experiences without fear, and encourage bystanders to confront the possibility that individuals within their professional networks could be perpetrators. These goals also extend to digital health and safety, given that survivors' willingness to disclose often depends on how platforms handle backlash, disbelief, and reputational risk.

In contrast, Sarita offered a more nuanced perspective, acknowledging both the limitations and potential of the #MeToo movement in challenging victim-blaming norms:

> *One of these tactics in Bollywood—a hotbed of the neoliberal beauty-fashion complex—has been to discredit these women on the basis of beauty norms. It becomes effective in a world mired in the logic of fashion as social capital. This form of slut-shaming, victim-blaming misogyny is effective in Bollywood to some extent. Yet what was new about #MeToo was that unlike earlier, (for instance, in the Shiney Ahuja case, who was arrested in 2009 for raping his then 19-y-old domestic help), there was a spotlight on the narrative of the victim. #MeToo Bollywood became her story. The narrative emphasis placed more screen time on the victim/survivor rather than the harasser. This is new, it helped that the women were also not dehumanized: it helped that we knew them through film roles earlier.*

Sarita's reflection highlighted an important shift in how survivor narratives were framed—away from silencing and toward recognition and visibility. This shift, she believed, has the potential to reduce stigma, support mental well-being, and strengthen collective efficacy among women by validating personal experiences and encouraging shared storytelling. Changing collective narratives around women's stigmatized experiences has been shown to foster online connection, mutual validation, and collective identity formation, all of which can help challenge misogynistic discourse and improve mental health [78]. Smita, another participant, noted that this collective storytelling offered a new form of solidarity:

> *By naming the perpetrators, the campaign made formal what was a circle of solidarity amongst women—what was previously whispered was now made public so women could be warned. To me this was the most useful aspect of the campaign. Now, this is not the same as going to court, naming the perpetrator, giving evidence, a process that takes its toll in terms of years and other costs. So, it is possible that #MeToo may end up diluting the victim's voice, the charge is that anyone can say whatever, etc. It is possible, but I think for the moment it has succeeded in sounding a warning note to potential perps. However, in the long run, as long as the severe imbalance remains in the industry, this campaign too may lose its credibility and power.*

XSL•FO
RenderX

Thus, despite the limitations, participants viewed #MeToo as a powerful, if temporary, mechanism for achieving symbolic and restorative justice through visibility, solidarity, and public accountability. Although some scholars argue that these rapid forms of justice may be transient and fail to address deeper structural determinants of sexual misconduct [79,80], the movement nonetheless represents an essential step toward disrupting the silence around GBV. Taken together, these insights position victim-blaming and digital backlash as both digital and public health concerns, shaping survivors' emotional well-being, safety, and ability to access redress within and beyond digital spaces.

Institutional System Justification

A striking contradiction revealed by the #MeToo movement in the IEI was the exposure of several acclaimed filmmakers, directors, and actors who were simultaneously celebrated for women-centric films and public advocacy for gender equality, yet accused of sexual harassment by colleagues in the industry. This pattern echoed high-profile cases in the United States involving R Kelly, Harvey Weinstein, and Bill Cosby, drawing similar attention to India among celebrity figures such as Subhash Ghai and Vikas Bahl (filmmakers), Alok Nath (actor), and Varun Grover (writer-comedian). While the allegations sent shockwaves through the public, many cases were quietly dismissed or settled, raising questions about institutional accountability and the fragility of justice when social status and reputation are at stake. Raya, a media professor in the United States, reflected on the broader cultural logic that enables such contradictions:

> *This does not surprise me at all. There is a self-declared distance Bollywood or Indian cinema directors have often maintained between "art" and "life." Representations of the female body and female sexuality have traditionally always been the prerogative of men in most cultures. Along with that, there has never been any mandate or prohibition against commodification or exploitation of the same body or femininity. In India, where there's also such a strong social binary between the public and the private woman—the actress and the wife—in terms of sexual politics, such an assumption of entitlement regarding "public women" or actresses must come quite easily to men in power, whether they are artistic, creative, or not. There may also be that "great men" syndrome whereby many of them believe that as artistes they are not subject to the norms of what is called "bourgeois" morality by the avant-garde in India.*

Raya's observation underscores how creative industries often justify misconduct through artistic exceptionalism—an attitude that normalizes exploitation as an accepted by-product of artistic freedom. Such institutional rationalizations mirror system-justifying processes that protect reputational hierarchies and maintain power asymmetries. Daniel and several other participants discussed how this contradiction extends to the production of women-centric films. As Daniel noted:

> *The film industry is, at its core, an industry. Directors and producers are artists, certainly, but their eye—and the eye of their investors and funders—is on the bottom line. If they believe there is a market for a female-centric movie, they will make it even if it is not entirely consonant with their own personal beliefs and attitudes. It's also important to remember that acting in ways inconsistent with sincerely held beliefs is hardly unusual.*

Collectively, participants highlighted the gap between gender-progressive representations on screen and persistent inequities off-screen. They pointed to entrenched gender ideologies, victim-blaming tendencies, and the commodification of women's bodies as markers of an institutional culture that simultaneously markets empowerment and reproduces exploitation. Several experts observed that the performative nature of the industry—where acting and pretense are core to professional success—creates conditions in which abusers maintain a convincing public façade, reinforcing impunity.

These narratives show how social structures that normalize gendered entitlement and discourage reporting can generate psychological strain, career marginalization, and reduced well-being among women. Addressing institutional system justification in such creative settings requires more than media accountability; it demands organizational and occupational frameworks that prioritize equity, enforce safeguards against harassment, and cultivate supportive environments for women survivors. From a public health lens, the IEI's gender inequities constitute a preventable occupational health crisis requiring multilevel interventions. For instance, digital health technologies offer scalable solutions for reporting, prevention, mental health support, and accountability—bridging the gap between performative empowerment and substantive safety. By combining public health frameworks with digital innovation, the industry can move beyond performative allyship toward measurable, sustainable gender justice that protects both mental well-being and professional equity.

Discussion

Social Media System Justification on Twitter

This study examined how social media discourse surrounding the #MeToo movement reflects and reproduces system-justifying narratives related to GBV. Consistent with this study's objectives, the analysis identified 4 dominant forms of system justification—gender, institutional, victim-blaming, and backlash—across a large corpus of Twitter data and expert interviews. Study 1 demonstrated that online conversations frequently defended existing gender norms and institutional structures, with gender and institutional justification emerging as the most prevalent categories. Such defense mechanisms reflect social conditioning, resistance to change, and selective exposure to congruent beliefs online [81,82].

Study 2 further contextualized these patterns within the IEI, revealing that even as #MeTooIndia expanded public dialogue on sexual harassment—entrenched hierarchies, reputational risks, and institutional inertia continued to constrain meaningful reform. Together, these findings indicate that while the #MeToo

XSL•FO
RenderX

movement fostered visibility and collective awareness, digital spaces also became arenas where system-justifying discourse reinforced social inequities and psychological strain, consistent with prior work showing that exposure to gender movements can amplify defensive reactions [48,50].

From a public health standpoint, these polarization patterns have psychosocial implications. Online gender-justifying narratives perpetuate stigma, discourage disclosure of sexual harassment, and increase survivors' emotional distress [10,19]. The normalization of inequitable discourse can further sustain hostile digital climates that contribute to secondary victimization—a public health concern linked to trauma recurrence and barriers to help-seeking [22,38]. The high prevalence of institutional justification observed in the dataset also reflects persistent trust in existing social and institutional structures, even when those systems perpetuate inequality [20,21,29]. System justification fulfills a psychological need to defend and justify existing social arrangements, even if those are disadvantageous to society, as it is often perceived as psychologically comforting to individuals [40].

Institutional justification in feminist movements often works to preserve the status quo, especially when current arrangements benefit those in power [36,49]. These dynamics help explain why movements such as #MeToo can expose systemic issues and yet still fall short of durable institutional change. Finally, victim blaming and backlash justifications appeared less frequently but remained consequential. While victim-blaming and backlash in feminist movements have been widely examined [21,29,32,83,84], our findings show a comparatively lower prevalence of such narratives in the #MeToo discourse on social media. The wide circulation of survivor stories helped render overt victim-blaming less acceptable, reshaping public perceptions of sexual harassment and assault [5,33]. However, amplifying women's voices through #MeToo also led some men to perceive themselves as unfairly victimized, fueling backlash that reasserted traditional gender hierarchies [20,85].

The "#MeTooLate Effect," where delayed disclosures prompted disbelief or blame, underscores persistent biases that privilege certain forms of reporting [21,54]. Perceived credibility often follows a hierarchy shaped by social expectations of an "ideal" victim (such as immediate reporting or corroborating witnesses), thereby marginalizing survivors who do not conform to these norms [11,86]. Overall, these findings suggest that while social media amplifies awareness of GBV, it also functions as a mirror of societal polarization. This simultaneously promotes empowerment and entrenches opposition. Thus, digital backlash can increase psychological distress, inhibit help-seeking, and perpetuate stigma surrounding GBV [3,11,19]. Addressing these online narratives through trauma-informed communication, digital moderation, and survivor-centered discourse strategies is essential for promoting mental well-being and social equity.

Consequences of #MeTooIndia Movement

The #MeToo movement in India catalyzed multiple discourses, revealing both systemic awareness and persistent structural inequities within the IEI [87,88]. Insights from study 2, based on in-depth interviews with scholarly experts, highlighted the movement's mixed legacy—while it expanded public dialogue around sexual harassment, entrenched gender hierarchies and institutional practices remained largely unchanged [15,89]. This disconnect between awareness and reform mirrors broader societal resistance to redistributing power within patriarchal systems [87,90].

Participants described continuing patterns of victim blaming toward women who disclosed harassment, often amplified by media sensationalism. Such reactions perpetuate psychological distress, reputational harm, and social isolation, underscoring how hostile cultural responses can evolve into public health concerns through cumulative mental health and occupational stress [91,92]. Experts noted that survivors in high-visibility professions face compounded stigma and scrutiny, exacerbating trauma and discouraging disclosure. This outcome is consistent with global research on secondary victimization [22,38].

Interviewees also identified a persistent paradox in the IEI's engagement with #MeToo, where increased public messaging on gender equity coexists with limited organizational accountability. This tension illustrates a symbolic commitment to feminist ideals rather than substantive institutional reform. Several experts described the movement's momentum as short-lived "clicktivism" or "slacktivism," characterized by digital enthusiasm but little lasting sociolegal impact [93-95]. Participants further noted that sustained solidarity was essential for continued engagement and resistance to social silencing, consistent with theories of collective action [96], though public disclosure frequently came with substantial psychological and professional costs.

Celebrities who spoke out faced coordinated online harassment, smear campaigns by industry elites, and defamation lawsuits aimed at intimidating or silencing them [16,89]. Such backlash reinforces patriarchal power structures and contributes to chronic distress, anxiety, and social withdrawal among women in high-visibility professions [20,92]. The hostile digital environment surrounding #MeTooIndia thus serves as an infoveillance signal of ongoing GBV and secondary victimization, illustrating how digital spaces can perpetuate trauma even while exposing it.

Despite these tensions, participants viewed the movement's visibility as crucial for advancing public dialogue on sexual violence and strengthening institutional accountability. Achieving digital health and safety requires trauma-informed legal frameworks and industry mechanisms that protect survivors and translate online advocacy into sustainable, health-promoting reform. These findings should be interpreted alongside the limitations described in the next section.

Limitations

Several key limitations influence the strength of the conclusions and recommendations drawn from this study. First, study 1 analyzed tweets collected over 1 month between November and December 2017, shortly after the initial #MeToo surge. This captures the early phase of the movement but may not represent later shifts in framing. Second, only English-language tweets were included, excluding parallel conversations in regional languages. Third, study 2 relied on 20 in-depth interviews with scholars and researchers specializing in gender, media, and film

XSL•FO
RenderX

studies, constraining the analysis to the IEI context. While these experts provided valuable theoretical and contextual nuance, the perspectives of survivors (who experience digital hostility and secondary victimization directly) may differ in important ways. The absence of survivor interviews represents an additional limitation. Additionally, cultural norms, linguistic variations, and platform-specific affordances may have shaped the tone and visibility of #MeToo discourse, limiting the generalizability of findings across social and regional contexts.

Conclusion

This study examined how social media discourse within the #MeToo movement reflects system-justifying narratives that sustain rather than dismantle existing hierarchies [5]. Contrary to prior studies emphasizing victim-blaming and overt backlash, findings from both study 1 and study 2 reveal strong evidence of gender and institutional system justification. These patterns indicate that social media often reinforces traditional gender norms and institutional authority, maintaining the status quo even within digital spaces that appear progressive. Although #MeToo expanded visibility and awareness, its ability to drive enduring structural change within sectors such as the IEI remains limited. Digital platforms, therefore, operate as contested arenas, where technologies that enable collective voice can simultaneously reproduce digital GBV.

Our findings extend scholarship that centers primarily on backlash and victim-blaming by demonstrating how broader system-justifying frames appear in social media discourses. The findings carry 3 major implications. First, for digital platforms, they highlight the need for proactive governance and content moderation policies that curb system-justifying retaliatory discourse. This requires integrating trauma-informed design principles to reduce secondary victimization. Second, for public health and policy stakeholders, the results point to the importance of recognizing digital hostility as a psychosocial health risk that warrants coordinated mental health and legal interventions. Third, for researchers and feminist advocates, this work offers an analytic lens for situating digital activism within broader institutional and psychological systems, while identifying points where transformative change may stall.

Acknowledgments

HP and ST are first authors who contributed equally. During the preparation of this work, the authors used generative artificial intelligence tools (Grammarly [Superhuman Platform], Claude [Anthropic], ChatGPT [OpenAI], and Microsoft Copilot [Microsoft Corp]) solely for copyediting, language refinement, and proofreading purposes. The content was subsequently reviewed, verified, and edited by the authors, who take full responsibility for the integrity and accuracy of this final paper. This use complies with the JMIR Publications Editorial Policy on the use of artificial intelligence tools.

Funding

The authors declared no financial support was received for this work.

Data Availability

The dataset analyzed in this study contains sensitive textual data related to gender-based sexual violence collected from public social media discussions. Sharing the raw tweet text could compromise user privacy, as individual posts may be reidentifiable even after anonymization. Furthermore, redistribution of tweet content is restricted under Twitter's Terms and Conditions [60]. To ensure transparency and reproducibility, a synthetic and anonymized sample dataset, along with the analysis code framework and software requirements, is included in the supplementary materials. Nonsensitive components of the code and illustrative samples are available upon reasonable request from the corresponding author.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Semistructured interview guide for study 2 (#MeTooIndia IEI expert interviews). IEI: Indian Entertainment Industry.
[DOCX File, 21 KB - jmir_v28i1e75533_app1.docx]

Multimedia Appendix 2
Data and code files description: data_analysis_template.txt contains the Python code used for the computational framework, including preprocessing, sentiment analysis, stance detection, and topic modeling. ReadMe.md provides details about the dataset, file structure, and disclosure on data availability and usage. requirements.txt lists all Python package dependencies used in the analysis. sample_data_mt.txt includes a sample dataset excerpt.
[ZIP File, 2 KB - jmir_v28i1e75533_app2.zip]

Checklist 1
COREQ checklist.

XSL•FO
RenderX

[PDF File, 497 KB - jmir_v28i1e75533_app3.pdf]

References

1. Field-Springer K, Draut H, Babrow F, Sandman M. (Re)claiming stories in the #MeToo movement: righting epistemic wrongs of physical, mental, and emotional harms of sexual violence. Health Commun 2022 Jul;37(8):982-991. [doi: 10.1080/10410236.2021.1880052] [Medline: 33541139]
2. Castle JJ, Jenkins S, Ortbals CD, Poloni-Staudinger L, Strachan JC. The effect of the #MeToo movement on political engagement and ambition in 2018. Polit Res Q 2020 Dec;73(4):926-941. [doi: 10.1177/1065912920924824]
3. Dilip D, Bates L. Sexual violence and mental health among college students in the era of #MeToo. Violence Gend 2021 Mar 1;8(1):8-13. [doi: 10.1089/vio.2020.0033]
4. Guan S, Hui V, Stiglic G, Constantino RE, Lee YJ, Wong AKC. Classifying the information needs of survivors of domestic violence in online health communities using large language models: prediction model development and evaluation study. J Med Internet Res 2025 May 12;27:e65397. [doi: 10.2196/65397] [Medline: 40354642]
5. Mayer ED, Peterson R, Kim E, et al. What survivors of sexual violence want when disclosing their experiences in person or online: qualitative interview analysis. J Med Internet Res 2025;27:e73497. [doi: 10.2196/73497]
6. Kaufman MR, Wright K, Shin R, et al. The power of social media activism in the #YesAllWomen movement. Humanit Soc Sci Commun 2025;12(1):1-11. [doi: 10.1057/s41599-025-05647-5]
7. Glassman M, Kuznetcova I, Peri J, Kim Y. Cohesion, collaboration and the struggle of creating online learning communities: Development and validation of an online collective efficacy scale. Comput Educ Open 2021 Dec;2:100031. [doi: 10.1016/j.caeo.2021.100031]
8. Suk J, Abhishek A, Zhang Y, et al. #MeToo, networked acknowledgment, and connective action: how “empowerment through empathy” launched a social movement. Soc Sci Comput Rev 2021 Apr;39(2):276-294. [doi: 10.1177/0894439319864882]
9. Jackson S, Bailey M, Welles BF. Women tweet on violence: from #YesAllWomen to #MeToo. Ada: J Gender, New Media, Technol 2019 Jan 1 [FREE Full text]
10. O’Neil A, Sojo V, Fileborn B, Scovelle AJ, Milner A. The #MeToo movement: an opportunity in public health? Lancet 2018 Jun;391(10140):2587-2589. [doi: 10.1016/S0140-6736(18)30991-7]
11. PettyJohn ME, Anderson G, McCauley HL. Exploring survivor experiences on social media in the #MeToo era: clinical recommendations for addressing impacts on mental health and relationships. J Interpers Violence 2022 Nov;37(21-22):NP20677-NP20700. [doi: 10.1177/08862605211055079] [Medline: 34861795]
12. Bethel C. #MeToo: the perfect storm needed to change attitudes toward sexual harassment and violence. HPHR 2018;16. [doi: 10.54111/0001/P1]
13. Rivera-Rodriguez A, Larsen G, Dasgupta N. Changing public opinion about gender activates group threat and opposition to feminist social movements among men. Group Processes Intergroup Relat 2022 Apr;25(3):811-829. [doi: 10.1177/13684302211048885]
14. Stephan WG, Ybarra O, Rios K. Intergroup threat theory. In: Handbook of Prejudice, Stereotyping, and Discrimination Psychology Press: Psychology Press; 2015:255-278. [doi: 10.4324/9781841697772]
15. Nanditha N. Exclusion in #MeToo India: rethinking inclusivity and intersectionality in Indian digital feminist movements. Feminist Media Stud 2022 Oct 3;22(7):1673-1694. [doi: 10.1080/14680777.2021.1913432]
16. No #MeToo here! Saroj Khan says casting couch provides livelihood, is a girl’s choice. The Economic Times. 2018 Apr 24. URL: https://economictimes.indiatimes.com/magazines/panache/no-metoo-here-saroj-khan-says-casting-couch-provides-livelihood-is-a-girls-choice/articleshow/63891824.cms?from=mdr [accessed 2026-03-20]
17. Hayes RM, Lorenz K, Bell KA. Victim blaming others: rape myth acceptance and the just world belief. Fem Criminol 2013 Jul;8(3):202-220. [doi: 10.1177/1557085113484788]
18. Bogen KW, Bleiweiss KK, Leach NR, Orchowski LM. #MeToo: disclosure and response to sexual victimization on Twitter. J Interpers Violence 2021 Sep;36(17-18):8257-8288. [doi: 10.1177/0886260519851211] [Medline: 31117851]
19. Bhuptani PH, López G, Peterson R, Orchowski LM. Online social reactions to disclosure of sexual victimization via #MeToo and symptoms of post-traumatic stress disorder. J Interpers Violence 2023 Oct;38(19-20):10900-10919. [doi: 10.1177/08862605231176792] [Medline: 37272010]
20. Moitra A, Ahmed SI, Chandra P. Parsing the “me” in #MeToo: sexual harassment, social media, and justice infrastructures. Proc ACM Hum-Comput Interact 2021;5:1-34. [doi: 10.1145/3449185]
21. Nutbeam M, Mereish EH. Negative attitudes and beliefs toward the #MeToo movement on Twitter. J Interpers Violence 2022 Aug;37(15-16):NP13018-NP13044. [doi: 10.1177/08862605211001470] [Medline: 33752498]
22. Gorissen M, van den Berg CJW, Ruiter S, Bijleveld C. Sharing unwanted sexual experiences online: a cross-platform analysis of disclosures before, during and after the #MeToo movement. Comput Hum Behav 2023 Jul;144:107724. [doi: 10.1016/j.chb.2023.107724]
23. Pain P. “It took me quite a long time to develop a voice”: examining feminist digital activism in the Indian #MeToo movement. New Media Soc 2021 Nov;23(11):3139-3155. [doi: 10.1177/1461444820944846]

XSL•FO
RenderX

24. Sambaraju R. “I would have taken this to my grave, like most women”: reporting sexual harassment during the #MeToo movement in India. J Soc Issues 2020 Sep;76(3):603-631. [doi: 10.1111/josi.12391]
25. Kende A, Nyúl B, Lantos NA, et al. A needs-based support for #MeToo: power and morality needs shape women’s and men’s support of the campaign. Front Psychol 2020;11:593. [doi: 10.3389/fpsyg.2020.00593] [Medline: 32296377]
26. Sakar S. Experiences of backlash against women in the form of technology-facilitated sexual violence in India. QUT Cent Justice 2023 Dec;44:1-4. [doi: 10.5204/book.eprints.245980]
27. McCosker A. Engaging mental health online: insights from beyondblue’s forum influencers. New Media Soc 2018 Dec;20(12):4748-4764. [doi: 10.1177/1461444818784303]
28. Liu L, Woo BKP. Twitter as a mental health support system for students and professionals in the medical field. JMIR Med Educ 2021 Jan 19;7(1):e17598. [doi: 10.2196/17598] [Medline: 33464210]
29. Xue J, Zhang B, Zhang Q, et al. Using Twitter-based data for sexual violence research: scoping review. J Med Internet Res 2023 May 15;25:e46084. [doi: 10.2196/46084] [Medline: 37184899]
30. Gangoli G. The #MeToo movement in India: emotions and (in)justice in feminist responses. Fem Leg Stud 2024 Jul;32(2):213-230. [doi: 10.1007/s10691-023-09540-x]
31. Marian M, Pérez RL, Reed E, et al. Exploring the use of social media for activism by Mexican nongovernmental organizations using posts from the 16 days of activism against gender-based violence campaign: thematic content analysis. JMIR Inf 2025 Apr 17;5:e67368. [doi: 10.2196/67368] [Medline: 40246294]
32. Maier SL. Rape victim advocates’ perceptions of the #MeToo movement: opportunities, challenges, and sustainability. J Interpers Violence 2023 Jan;38(1-2):NP336-NP365. [doi: 10.1177/08862605221081929] [Medline: 35416073]
33. Ortiz RR, Smith AM. A social identity threat perspective on why partisans may engage in greater victim blaming and sexual assault myth acceptance in the #MeToo era. Violence Women 2022 Apr;28(5):1302-1325. [doi: 10.1177/10778012211014554] [Medline: 34086501]
34. Huck JL, Homer M. How men view stigma, shame, and blame in #MeToo. Soc Sci Q 2024;105(5):1515-1528. [doi: 10.1111/ssqu.13428]
35. Modrek S, Chakalov B. The #MeToo movement in the United States: text analysis of early Twitter conversations. J Med Internet Res 2019 Sep 3;21(9):e13837. [doi: 10.2196/13837] [Medline: 31482849]
36. Silva M, Anaba U, Jani Tulsani N, Sripad P, Walker J, Aisiri A. Gender-based violence narratives in internet-based conversations in Nigeria: social listening study. J Med Internet Res 2023 Sep 15;25:e46814. [doi: 10.2196/46814] [Medline: 37713260]
37. Langer N, Khurana S, Vaast E. Activating the sisterhood: a structural and temporal analysis of sustained connective action in #metooindia. Presented at: ICIS 2020; Dec 13-16, 2020 URL: https://aisel.aisnet.org/icis2020/blendlocalglobal/blendlocalglobal/7 [accessed 2026-04-02]
38. Jackson MA, Valentine SE, Woodward EN, Pantalone DW. Secondary victimization of sexual minority men following disclosure of sexual assault: “victimizing me all over again…”. Sex Res Soc Policy 2017 Sep;14(3):275-288. [doi: 10.1007/s13178-016-0249-6]
39. Giddens L, Petter S. "They deserved it”: using the just world hypothesis to understand blaming. Presented at: 53rd Hawaii International Conference on System Sciences; Jan 7-10, 2020. [doi: 10.24251/HICSS.2020.724]
40. Jost JT, Banaji MR, Nosek BA. A decade of system justification theory: accumulated evidence of conscious and unconscious bolstering of the status quo. Polit Psychol 2004 Dec;25(6):881-919. [doi: 10.1111/j.1467-9221.2004.00402.x]
41. Card KG, Hepburn K. Social position and economic system justification in Canada: implications for advancing health equity and social justice from an exploratory study of factors shaping economic system justification. Front Public Health 2022;10:902374. [doi: 10.3389/fpubh.2022.902374] [Medline: 36339181]
42. Lima BPB, de Souza LEC, Jost JT. System justification, subjective well-being, and mental health symptoms in members of disadvantaged minority groups. Clin Psychol Rev 2025 Feb;115:102532. [doi: 10.1016/j.cpr.2024.102532] [Medline: 39740353]
43. Jost JT. A quarter century of system justification theory: questions, answers, criticisms, and societal applications. Br J Soc Psychol 2019 Apr;58(2):263-314. [doi: 10.1111/bjso.12297]
44. Jost JT, Banaji MR. The role of stereotyping in system‐justification and the production of false consciousness. Br J Soc Psychol 1994 Mar;33(1):1-27. [doi: 10.1111/j.2044-8309.1994.tb01008.x]
45. Bleeker K, Powers RA, Centelles V. “If we can change the culture, I hope we won’t need to say Me Too”: disclosures of sexual victimization as a form of social activism. Am J Crim Just 2023 Oct;48(5):1224-1244. [doi: 10.1007/s12103-022-09695-4]
46. Andalibi N, Haimson OL, Choudhury MD, Forte A. Social support, reciprocity, and anonymity in responses to sexual abuse disclosures on social media. ACM Trans Comput-Hum Interact 2018 Oct 31;25(5):1-35. [doi: 10.1145/3234942]
47. Drewett C, Oxlad M, Augoustinos M. Breaking the silence on sexual harassment and assault: an analysis of #MeToo tweets. Comput Human Behav 2021 Oct;123:106896. [doi: 10.1016/j.chb.2021.106896]
48. Szekeres H, Shuman E, Saguy T. Views of sexual assault following #MeToo: the role of gender and individual differences. Pers Individ Dif 2020 Nov;166:110203. [doi: 10.1016/j.paid.2020.110203]

XSL•FO
RenderX

49. Liaquat U, Jost JT, Balcetis E. System justification motivation as a source of backlash against equality‐promoting policies—and what to do about it. Soc Issues Policy Rev 2023 Jan;17(1):131-154. [doi: 10.1111/sipr.12093]
50. Menegatti M, Mazzuca S, Ciaffoni S, Moscatelli S. Behind the lines of #MeToo: exploring women's and men's intentions to join the movement. Sustainability 2022;14(19):12294. [doi: 10.3390/su141912294]
51. Kachen A, Krishen AS, Petrescu M, Gill RD, Peter PC. #MeToo, #MeThree, #MeFour: Twitter as community building across academic and corporate institutions. Psychol Mark 2021 Mar;38(3):455-469. [doi: 10.1002/mar.21442]
52. Minkina N. Can #MeToo abolish sexual harassment and discrimination in medicine? Lancet 2019 Aug;394(10196):383-384. [doi: 10.1016/S0140-6736(19)31731-3]
53. Petter S, Giddens L, Wake Forest University, University of North Texas. Is it your fault? Framing social media inclusion and exclusion using just world theory. JAIS 2023;24(5):1248-1270. [doi: 10.17705/1jais.00813]
54. Lucarini A, Suitner C, Brown R, Craig MA, Knowles ED, Salvador Casara BG. The #MeTooLate effect: victim blame and trust denial for sexual harassment not immediately reported. Pers Individ Dif 2020 Dec;167:110240. [doi: 10.1016/j.paid.2020.110240]
55. Stewart H. "Why didn't she say something sooner?": Doubt, denial, silencing, and the epistemic harms of the #MeToo movement. South Cent Rev 2019;36(2):68-94. [doi: 10.1353/scr.2019.0014]
56. Hasunuma L, Shin K. #MeToo in Japan and South Korea: #WeToo, #WithYou. In: Me Too Political Science: Routledge; 2020:97-111. [doi: 10.1201/9781003014508-6]
57. Ivankova NV, Creswell JW, Stick SL. Using mixed-methods sequential explanatory design: from theory to practice. Field Methods 2006 Feb;18(1):3-20. [doi: 10.1177/1525822X05282260]
58. Luo H, Zhang L. Scandal, social movement, and change: evidence from #MeToo in Hollywood. Manage Sci 2022 Feb;68(2):1278-1296. [doi: 10.1287/mnsc.2021.3982]
59. Wehrli S, Hartner AM, Boender TS, Arnrich B, Irrgang C. Information pathways and voids in critical German online communities during the COVID-19 vaccination discourse: cross-platform and mixed methods analysis. J Med Internet Res 2025;27:e76309-e76309. [doi: 10.2196/76309]
60. Chen E, Lerman K, Ferrara E. Tracking social media discourse about the COVID-19 pandemic: development of a public coronavirus Twitter data set. JMIR Public Health Surveill 2020;6(2):e19273. [doi: 10.2196/19273]
61. eCFR:: 45 CFR 46.102 -- Definitions for purposes of this policy. eCFR. URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46/subpart-A/section-46.102 [accessed 2026-04-02]
62. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013;310(20):2191-2194. [doi: 10.1001/jama.2013.281053]
63. Putting women first: ethical and safety recommendations for research on domestic violence against women. : World Health Organization; 2001 URL: https://www.who.int/publications/i/item/WHO-FCH-GWH-01.1 [accessed 2026-03-20]
64. Reyes-Menendez A, Saura JR, Thomas SB. Exploring key indicators of social identity in the #MeToo era: using discourse analysis in UGC. Int J Inf Manage 2020 Oct;54:102129. [doi: 10.1016/j.ijinfomgt.2020.102129]
65. Speziale HS, Streubert HJ, Carpenter DR. Qualitative Research in Nursing: Advancing the Humanistic Imperative: Lippincott Williams & Wilkins; 2011.
66. Thakkar S. Exploitation, harassment and violence: lived experiences of women paid domestic workers in India. J South Asian Dev 2024 Apr;19(1):44-60. [doi: 10.1177/09731741231164872]
67. Thakkar S. Moving towards empowerment: migrant domestic workers in India. Asian J Women Stud 2021 Jul 3;27(3):425-440. [doi: 10.1080/12259276.2021.1957195]
68. Robinson OC. Sampling in interview-based qualitative research: a theoretical and practical guide. Qual Res Psychol 2014 Jan 2;11(1):25-41. [doi: 10.1080/14780887.2013.801543]
69. Cater JK. Skype: a cost-effective method for qualitative research. Rehabil Counselors Educ J 2011;4(2):1-4 [FREE Full text]
70. Hooley T, Marriott J, Wellens J. What Is Online Research?: Using the Internet for Social Science Research, 1st edition: Bloomsbury Academic; 2012. [doi: 10.5040/9781849665544]
71. Janghorban R, Roudsari RL, Taghipour A. Skype interviewing: the new generation of online synchronous interview in qualitative research. Int J Qual Stud Health Well-Being 2014 Jan;9(1):24152. [doi: 10.3402/qhw.v9.24152]
72. Strauss AL. Qualitative Analysis for Social Scientists: Cambridge University Press; 1987.
73. Blair E. A reflexive exploration of two qualitative data coding techniques. J Methods Meas Soc Sci 2015;6(1):14-29. [doi: 10.2458/v6i1.18772]
74. Saldaña J. Goodall's verbal exchange coding: an overview and example. Qual Inquiry 2016;22(1):36-39. [doi: 10.1177/1077800415603395]
75. Has #MeToo changed Bollywood? BBC News. 2019 Dec 27. URL: https://www.bbc.com/news/av/world-asia-india-50893469 [accessed 2026-03-20]
76. Pegu S. MeToo in India: building revolutions from solidarities. Decision 2019 Jun;46(2):151-168. [doi: 10.1007/s40622-019-00212-x]

XSL•FO
RenderX

77. Brünker F, Wischnewski M, Mirbabaie M, Meinert J. The role of social media during social movements–observations from the #MeToo debate on Twitter. Presented at: Hawaii International Conference on System Sciences; Jan 7-10, 2020. [doi: 10.24251/HICSS.2020.288]
78. Gallagher RJ, Stowell E, Parker AG, Welles BF. Reclaiming stigmatized narratives: the networked disclosure landscape of #metoo. In: Proceedings of the ACM on Human-Computer Interaction: Association for Computing Machinery; 2019, Vol. 3:1-30. [doi: 10.1145/3359198]
79. Pipyrou S. MeToo is little more than mob rule//vs//# MeToo is a legitimate form of social justice. HAU: J Ethnographic Theory 2018;8(3):415-419. [doi: 10.1086/701007]
80. Fohring S, Horsfield L. Survivors speak out: the successes and failures of hashtag activism. In: Canning V, Martin G, Tombs S, editors. The Emerald International Handbook of Activist Criminology: Emerald Publishing Limited; 2023:337-350. [doi: 10.1108/978-1-80262-199-020231022]
81. Otterbacher J, Bates J, Clough P. Competent men and warm women: gender stereotypes and backlash in image search results. In: CHI '17: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems: Association for Computing Machinery; 2017:6620-6631. [doi: 10.1145/3025453.3025727]
82. Mou Y, Cui Y, Wang J, Wu Y, Li Z, Wu Y. Perceiving sexual harassment and #metoo social media campaign among Chinese female college students. J Gend Stud 2022 Feb 17;31(2):178-192. [doi: 10.1080/09589236.2021.1884848]
83. Andreasen MB. 'Rapeable' and 'unrapeable' women: the portrayal of sexual violence in internet memes about #MeToo. J Gend Stud 2021 Jan 2;30(1):102-113. [doi: 10.1080/09589236.2020.1833185]
84. Sharoni S. Speaking up in the age of #MeToo and persistent patriarchy or what can we learn from an elevator incident about anti-feminist backlash. Fem Rev 2018 Nov;120(1):143-151. [doi: 10.1057/s41305-018-0127-6]
85. Lisnek JA, Wilkins CL, Wilson ME, Ekstrom PD. Backlash against the #MeToo movement: how women's voice causes men to feel victimized. Group Processes Intergroup Relat 2022 Apr;25(3):682-702. [doi: 10.1177/13684302211035437]
86. Hart CG. Is There an idealized target of sexual harassment in the MeToo era? Soc Probl 2025 Jan 27;72(1):277-293. [doi: 10.1093/socpro/spad016]
87. Mishra S. From #MeToo to #MeTooIndia: news domestication in Indian English language newspapers. J Stud 2020 Apr 3;21(5):659-677. [doi: 10.1080/1461670X.2019.1709882]
88. Mishra S. A Bollywood icon was accused of sexual assault. The response was chilling. Washington Post. 2019 Jan 17. URL: https://www.washingtonpost.com/opinions/2019/01/17/bollywood-icon-was-accused-sexual-assault-response-was-chilling/ [accessed 2026-03-20]
89. Mitra S. Trolled, body-shamed and slut-shamed: the desecration of the contemporary Bollywood female star on social media. Celebrity Stud 2020 Jan 2;11(1):101-115. [doi: 10.1080/19392397.2020.1704388]
90. Ghosh S, Kamath K. News sourcing, gender, and misogyny in #MeToo india. In: Misogyny Across Global Media: Bloomsbury Publishing PLC; 2021:173-190. [doi: 10.5040/9781666996227]
91. Lakkimsetti C. Stripping away at respectability: #MeToo India and the politics of dignity. Feminist Form 2021;33(3):303-317. [doi: 10.1353/ff.2021.0052]
92. Hayes RM. From blurred lines to #MeToo: media impact on rape culture. In: Defining Rape Culture: Gender, Race and the Move Toward International Social Change: Emerald Publishing Limited; 2023:55-76. [doi: 10.1108/978-1-80262-213-320231004]
93. Christensen HS. Political activities on the internet: slacktivism or political participation by other means? First Monday 2011 Feb 2;16(2). [doi: 10.5210/fm.v16i2.3336]
94. Lindgren S. Movement mobilization in the age of hashtag activism: examining the challenge of noise, hate, and disengagement in the #MeToo campaign. Policy Internet 2019 Dec;11(4):418-438. [doi: 10.1002/poi3.212]
95. Tonkinwise C. To save everything, click here: the folly of technological solutionism. J Des Hist 2014 Mar 1;27(1):111-113. [doi: 10.1093/jdh/ept034]
96. Polletta F, Jasper JM. Collective identity and social movements. Annu Rev Sociol 2001 Aug;27(1):283-305. [doi: 10.1146/annurev.soc.27.1.283]

Abbreviations

GBV: gender-based violence
IEI: Indian Entertainment Industry

XSL•FO
RenderX

Edited by A Stone, N Cahill; submitted 05.Apr.2025; peer-reviewed by C Matt Graham, L Couto; revised version received 24.Dec.2025; accepted 29.Dec.2025; published 09.Apr.2026.

Please cite as:
Parekh H, Thakkar S, Bhatt P, Akello P
Digital Discourse, Secondary Victimization, and Psychological Harm: Mixed-Methods Analysis of System Justification in the #MeToo Movement
J Med Internet Res 2026;28:e75533
URL: https://www.jmir.org/2026/1/e75533
doi:10.2196/75533

© Harsh Parekh, Shriya Thakkar, Paras Bhatt, Patricia Akello. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Policy Considerations for National Virtual Hospitals: Global Evidence and the Seha Virtual Hospital Model

Sohaib Shujaat[1], PhD; Hawazin Almutairi[2], DMD; Saleh Hamad Alhuraibi[2], DMD; Hongyang Ma[3], PhD

[1]King Abdullah International Medical Research Center, Department of Maxillofacial Surgery and Diagnostic Sciences, College of Dentistry, King Saud Bin Abdulaziz University for Health Sciences, Ministry of National Guard Health Affairs, Mail Code: 3183; P.O. Box 3660, Riyadh, Saudi Arabia

[2]King Abdullah International Medical Research Center, College of Dentistry, King Saud Bin Abdulaziz University for Health Sciences, Ministry of National Guard Health Affairs, Riyadh, Saudi Arabia

[3]Department of Oral Implantology, The Second Affiliated Hospital of Harbin Medical University, Harbin, China

Corresponding Author:
Sohaib Shujaat, PhD
King Abdullah International Medical Research Center, Department of Maxillofacial Surgery and Diagnostic Sciences, College of Dentistry, King Saud Bin Abdulaziz University for Health Sciences, Ministry of National Guard Health Affairs, Mail Code: 3183; P.O. Box 3660, Riyadh, Saudi Arabia

Abstract

Health systems worldwide face growing pressure from population aging, multimorbidity, and rising emergency admissions, prompting reconsideration of traditional inpatient care models. In response, digitally enabled models such as tele–intensive care unit (tele-ICU) programs, hospital-at-home services, virtual wards, and other remote specialist pathways have expanded, particularly after the COVID-19 pandemic accelerated telemedicine adoption and cross-site virtual staffing. However, nationally coordinated, multispecialty virtual hospitals remain uncommon worldwide, and robust evidence on their system-level effects is still limited. As a result, policy discussions about national virtual hospitals must often draw on evidence from related virtual-care models rather than from mature national implementations. This viewpoint synthesizes representative international evidence from tele-ICU systems, hospital-at-home programs, virtual wards, telestroke networks, and other condition-specific virtual-care pathways, and examines Saudi Arabia's Seha Virtual Hospital (SVH) as a national case study to identify policy lessons relevant to the design, governance, and evaluation of national virtual hospitals. Across settings, these models suggest that remote and digitally supported care can achieve outcomes comparable to in-person hospital care when patient selection is appropriate, escalation and transfer pathways are explicit, monitoring intensity matches clinical risk, and multidisciplinary teams are integrated into local workflows. Tele-ICU programs have reported reductions in intensive care mortality and length of stay under well-structured organizational models, while hospital-at-home and virtual-ward programs have shown comparable safety, reduced hospital usage, and improved patient experience among selected patient groups. Telestroke networks likewise demonstrate outcomes comparable to specialist in-person care in acute stroke pathways. Nevertheless, the evidence base remains heterogeneous and strongly context-dependent. Much of the literature is short-term, with limited consistent evidence on long-term outcomes, caregiver burden, cost-effectiveness, workforce implications, and digital equity. SVH illustrates the emerging implementation of a centralized national virtual hospital model. Launched in 2022 under Saudi Arabia's Vision 2030 Health Sector Transformation Program, SVH operates as a national telehealth hub embedded within the country's broader digital-health ecosystem and links hospitals across the Kingdom to specialized clinical expertise. Its service portfolio includes urgent and critical care consultations, specialized virtual clinics, multidisciplinary case discussions, and supportive diagnostic services. Early reports indicate rapid operational expansion, broad institutional participation, and national-scale feasibility. However, independent comparative evidence evaluating SVH's effects on mortality, readmissions, length of stay, cost-effectiveness, equity, and workforce sustainability remains limited. National virtual hospitals should therefore be understood as evidence-generating health-system innovations rather than fully validated care models. Sustainable scale-up requires embedding rigorous prospective evaluation within implementation, aligning financing mechanisms with substitution of inpatient care, establishing clear governance and regulatory frameworks, and addressing digital inclusion and workforce sustainability. These considerations can help guide policymakers and health-system leaders in the accountable, equitable, and evidence-informed development of national virtual hospital programs.

(J Med Internet Res 2026;28:e89276) doi:10.2196/89276

KEYWORDS

virtual hospitals; telehealth policy; digital health systems; health systems innovation; hospital-at-home; virtual wards; health equity

XSL•FO
RenderX

Introduction

Health systems worldwide face mounting pressure from aging populations, rising multimorbidity, and increasing emergency admissions, prompting reconsideration of traditional inpatient care models [1-3]. Although these pressures are most pronounced in high-income countries, low- and middle-income countries are undergoing parallel epidemiological transitions, with noncommunicable diseases and injuries now accounting for most global mortality and disability and generating sustained hospital demand [4-6]. These trends are driving governments to pursue alternative models of hospital-level care [7]. Hospital-at-home and virtual-ward models have expanded to relieve bed constraints while maintaining clinical outcomes and patient experience [8,9].

The COVID-19 pandemic accelerated this shift through regulatory reform, telemedicine expansion, and normalization of remote clinical decision-making and cross-site virtual staffing [8-11], establishing technical and organizational foundations for more integrated remote hospital care [11]. Within this evolving ecosystem, several virtual-care models have emerged [12]. Tele–intensive care unit (tele-ICU) programs provide centralized specialist oversight of geographically distributed intensive care units (ICUs) through real-time monitoring and structured escalation protocols [13,14]. Virtual wards deliver time-limited hospital-level care remotely to selected patients, typically to support admission avoidance or early discharge [9,15]. Virtual hospitals extend beyond single pathways by integrating tele-ICU, virtual wards, and specialty teleconsultation within shared digital infrastructure, multidisciplinary workforce models, and governance frameworks [9,16].

Evidence from tele-ICU programs demonstrates reductions in ICU mortality and length of stay when effectively integrated into local workflows [13,14,17]. Similarly, hospital-at-home and virtual-ward models show comparable safety to inpatient care, reduced length of stay, and in some contexts lower readmission rates among carefully selected patients [15,18,19]. However, translating this evidence to national, cross-specialty virtual hospitals raises unresolved questions regarding scalability, equity, workforce sustainability, governance, and accountability [15,17]. Differential uptake of telehealth across socioeconomic groups further highlights risks of digital inequity if access barriers are not addressed [20,21]. Although most virtual-hospital implementations remain subnational or specialty-specific, nationally coordinated, cross-specialty models are emerging but are not yet established worldwide [22-24].

This viewpoint examines national virtual hospitals as policy innovations operating in an evidence-generating phase rather than mature digital solutions. Saudi Arabia's Seha Virtual Hospital (SVH) [25] serves as a national-scale case study. Three key messages guide this analysis: (1) evidence from tele-ICU, hospital-at-home, virtual wards, and condition-specific virtual care provides important but incomplete foundations for national virtual hospitals; (2) national deployment introduces risks related to outcome variability, equity, workforce sustainability, governance, and accountability; and (3) sustainable scale-up requires explicit policy, implementation, and evaluation frameworks. The intended audience includes policymakers, health-system leaders, clinicians, and researchers engaged in digital health transformation.

Global Evidence Review: System-Level Impact and Evidence Gaps

Overview

This section synthesizes representative international evaluations on tele-ICU, hospital-at-home, virtual wards, telestroke, and virtual-hospital initiatives. Table 1 summarizes examples of nationally coordinated, multispecialty virtual hospital models. Fully national virtual hospitals remain uncommon internationally, with only a small number of countries, such as Saudi Arabia, Finland, and Jordan, implementing centralized virtual hospital structures [25-27]. In contrast, most jurisdictions rely on national virtual wards, hospital-at-home programs, regional virtual-hospital networks, national telehealth platforms, or institution-led virtual-care systems rather than centralized national virtual hospitals. These broader models are summarized in Multimedia Appendix 1.

XSL•FO
RenderX

Table . National multispecialty virtual hospital models.

Country	System or entity	Brief description	References
Saudi Arabia	Seha Virtual Hospital	Ministry of Health–led national virtual hospital delivering specialist teleconsultations across public hospitals. The system provides multispecialty and subspecialty services through a centralized digital platform as part of Saudi Vision 2030 health-system transformation.	[25]
Finland	Virtual Hospital 2.0/Health Village (Terveyskylä)	Nationally coordinated digital hospital platform developed by all Finnish university hospitals, delivering condition-specific virtual care pathways, outpatient services, and self-management programs integrated into the public health system. While not delivering inpatient-equivalent care, it functions as a nationally governed virtual hospital platform rather than a generic telehealth service.	[26]
Jordan	Jordan Digital Health Center (virtual hospital)	National virtual hospital network connecting peripheral hospitals to a central command center, providing services such as tele–intensive care unit, teledialysis, and teleradiology, with planned national expansion.	[27]

Tele-ICU and Remote Critical-Care Networks

Tele-ICU programs represent one of the earliest and most established forms of virtual hospital care [28,29]. Although most tele-ICU systems operate at the institutional or regional rather than national scale, they provide important operational foundations for contemporary virtual-hospital models [30]. Prior studies have reported measurable improvements when tele-ICU services were implemented within structured organizational models. For instance, in a pre-post study of 118,990 patients across 56 ICUs, tele-ICU adoption was associated with approximately 20% reductions in ICU mortality and 15% reductions in hospital mortality, alongside shorter ICU stays [14]. A separate retrospective analysis of 16,091 patients in progressive-care units reported lower unit mortality (0.7% vs 1%), lower hospital mortality (4.4% vs 5.2%), and shorter unit stays (2.6 vs 3.2 d) in telemedicine-supported units, although direct costs were modestly higher [31]. However, larger effectiveness studies evaluating tele-ICU adoption across diverse hospital settings demonstrate more modest and heterogeneous effects. Specifically, in a multicenter study including more than 500 US hospitals, mortality reductions were observed overall but were statistically significant in only a minority of adopting sites [17]. Similarly, a national analysis of 66,522 mechanically ventilated patients found no significant association between tele-ICU availability and in-hospital mortality or duration of ventilation [32]. These findings indicate that tele-ICU systems can produce meaningful outcome improvements under well-defined organizational conditions, but benefits are not uniformly reproducible at scale.

Hospital-at-Home and Acute Care at Home

Hospital-at-home represents a core component of many contemporary virtual-hospital models [33]. In one trial of acutely ill adults eligible for inpatient admission, substitutive hospital-level care delivered at home was associated with 38% lower direct acute-care costs and reduced 30-day readmissions (23% vs 7%) without increased adverse events [34]. A subsequent pilot trial reported more modest cost reductions (5%) but similarly lower 30-day readmissions (11% vs 36%) while maintaining safety and patient experience [35]. The degree of virtualization within hospital-at-home models also varies. In a randomized clinical trial comparing remote physician visits with traditional in-home visits, remote care was noninferior for adverse events and patient experience, although approximately 20% of patients required occasional in-person physician support [36], suggesting fully virtual oversight may be feasible for many but not all patients.

Across broader evidence syntheses, hospital-at-home models generally demonstrate mortality outcomes comparable to inpatient care across diverse acute conditions, with frequent reductions in length of stay and variable but often favorable cost profiles [37-39]. A review of admission-avoidance hospital-at-home in older adults found little to no difference in six-month mortality, similar readmission rates, and a probable reduction in transitions to residential care [39]. However, caregiver burden and longer-term system effects remain insufficiently characterized. Performance and safety depend heavily on local infrastructure, including trained community nursing teams, workforce capacity, and rapid escalation pathways [40]. Overall, the aforementioned evidence suggests that hospital-at-home can deliver noninferior clinical outcomes,

XSL•FO
RenderX

reduced readmissions, and potential cost savings for carefully selected patients, while emphasizing the importance of patient selection and system integration for safe scale-up.

Virtual Wards and Community-Based Virtual Care

Frail older adults are a major target of virtual-ward programs because of high hospital usage and transitional-care needs [15,41]. Virtual wards extend the hospital-at-home concept by providing time-limited, multidisciplinary oversight for conditions such as frailty, heart failure, and chronic obstructive pulmonary disease [15]. A community-based virtual-ward model in Ireland for older adults with complex needs reported a reduction in median hospital bed-days from 23 to 0 over 12 months, alongside fewer emergency department visits and unplanned admissions, and high patient satisfaction [42]. In England, rapid national implementation of frailty-focused virtual wards has occurred, but comparative outcome data remain limited, and effectiveness appears strongly influenced by local configuration and staffing models [7]. Simulation modeling further suggests that admission thresholds, monitoring intensity, and workforce capacity substantially affect usage and hospital pressure [43].

During the COVID-19 pandemic, multiple virtual-ward and home-monitoring models were deployed at scale. Among these, the national COVID Oximetry @home and COVID Virtual Ward programs in England monitored over 19,000 patients across 26 sites, with per-patient running costs of approximately £528-£599 and mortality below 1% [44]. In Sydney, Australia, a community virtual-care program enrolled 93.6% of locally diagnosed COVID-19 patients, with low escalation rates (ambulance 3%, emergency department 2.5%, and hospital admission 1.9%) and no reported deaths [45]. These experiences demonstrate the feasibility of virtual-ward deployment during periods of extreme health-system pressure, while highlighting the importance of strong clinical governance, integration with primary and community care services, and adequate workforce support [46]. Overall, the aforementioned evidence suggests that virtual-ward models may reduce unplanned hospital usage and facilitate earlier discharge for selected patient groups. However, most evaluations remain uncontrolled, context-specific, and short-term, with limited long-term clinical, economic, and caregiver data [44-46].

Telestroke and Other Condition-Specific Virtual-Care Models

Among condition-specific virtual-care pathways, telestroke has one of the most robust comparative evidence bases, consistently demonstrating safety and effectiveness comparable to in-person specialist care for acute ischemic stroke [47]. A meta-analysis of 1863 patients treated within the 3-hour therapeutic window found no significant differences between telemedicine-guided and in-person thrombolysis in symptomatic intracerebral hemorrhage, mortality, or 3-month functional independence [48]. In a hub-and-spoke telestroke network (3‑4.5-hour window), rates of symptomatic intracerebral hemorrhage (4.9% vs 6.3%), functional independence (44.1% vs 39.6%), and in-hospital mortality (9.6% vs 8.3%) were comparable to tertiary stroke-center care [49]. A larger meta-analysis including 12,540 patients similarly reported no meaningful differences in 90-day mortality or functional independence between telestroke and in-person management [47].

Beyond stroke, other condition-specific virtual-care models show potential to reduce usage and improve access. Among these, a wraparound virtual-care program for medically complex children reported 35.3% fewer hospitalizations and 43.9% fewer emergency department visits over 12 months in a pre-post evaluation of 75 patients, although the absence of concurrent controls limits causal inference [50]. In addition, a four-year panel analysis of 1.6 million admissions across 63 hospitals showed that telehealth adoption was associated with shorter hospital stays, with greater benefits at longer travel distances (≈11-minute reduction in length of stay per additional minute of travel time) [51]. Across tele-ICU, hospital-at-home, virtual-ward, and telestroke pathways, a consistent pattern emerges where virtual-care models can match in-person outcomes when embedded within structured networks, clear clinical pathways, and appropriate workforce support [13,18,47]. However, effects remain context-dependent, indicating that technology alone does not ensure system-level benefit [17,29].

System-Level Considerations and Health-System Impact

Outcomes and Workload Redistribution

The effects of virtual hospitals on outcomes and workload depend on how care delivery is organized across the health system [17]. Reductions in hospital usage reported in frailty-focused virtual wards appear most consistent when services are tightly integrated with primary care, multidisciplinary teams are stable, and escalation pathways are clearly defined [15]. When these structural elements are absent, effects are variable [15,17]. Clinicians often value virtual wards, but report increased cognitive load, coordination complexity, and uncertainty when protocols are poorly defined [52]. Broader digital-health research shows similar dynamics, where electronic health interventions may reduce in-person visits while increasing workload through additional messaging, documentation, and coordination [53]. High volumes of secure messaging and video consultations have also been associated with increased burnout among primary-care physicians [54]. These findings suggest that virtual hospitals redistribute rather than simply reduce workload.

Ambient artificial intelligence scribe technologies have been proposed as one potential mitigation strategy [55]. By generating clinical documentation from real-time encounters, early deployments report reductions in note-writing time and after-hours documentation burden [55,56]. However, integration remains evolving, and regulatory issues related to accuracy, privacy, and accountability persist [57]. Such tools may therefore complement virtual-hospital models but cannot substitute for structural workforce design and clear governance. As national virtual-hospital programs expand, formal evaluation should incorporate workload metrics alongside usage and cost outcomes, recognizing that reductions in inpatient bed-days do not necessarily equate to net workforce relief [53,54].

XSL•FO
RenderX

Equity and Digital Determinants of Health

Digital determinants of health describe how access to digital devices and connectivity, digital literacy, platform usability, trust, and affordability influence who can benefit from digital health interventions [58]. These aforementioned determinants interact with traditional social determinants of health and may either mitigate or exacerbate existing health inequities. From this perspective, virtual hospitals are not neutral technological solutions but system-level care models whose impact depends on how digital resources and capabilities are distributed across populations [58,59]. The digital equity framework suggests that virtual care has the potential to expand access and reduce travel burdens, particularly for rural and underserved communities; however, evidence of sustained equity gains remains limited, and many initiatives continue to underserve individuals experiencing socioeconomic disadvantage [59].

Digitalization may further widen or reduce disparities among people with disabilities, depending on accessibility standards, affordability, and the availability of appropriate technical and social support [60]. These concerns are reflected in real-world usage patterns, as patient-portal adoption is consistently lower among individuals with lower income, older adults, minority populations, and those with limited health literacy [61]. As virtual hospital models often assume reliable connectivity, access to digital devices, and caregiver support, digital determinants of health must be explicitly incorporated into program design and evaluation rather than treated as implicit or automatic benefits [58,60].

Workforce Well-Being and Organizational Dynamics

Sustainable virtual-hospital models depend heavily on organizational conditions, particularly because virtual care often redistributes rather than removes clinical work [52,53,62]. Burnout among health care professionals is strongly associated with heavy workloads, inadequate staffing, and limited organizational support [62]. In virtual-ward deployments, staff describe tension between enthusiasm for home-based care and anxiety about responsibility for deteriorating patients outside hospital walls [52]. These concerns are especially relevant for national virtual hospitals, where centralized specialist input may increase coordination demands on local teams [52]. Workforce planning should therefore include realistic staffing models, clear role delineation, escalation support, protected time for virtual responsibilities, and structured clinical and emotional support [52,62].

Governance, Regulation, and Accountability

Governance and regulatory requirements for virtual hospitals are jurisdiction-dependent but share common challenges, including clinician licensure, scope of practice, data protection, cross-border teleconsultation, and clinical liability [63,64]. These factors directly affect safety, scalability, and public trust. In Saudi Arabia, the Personal Data Protection Law and Telehealth Application Guidelines establish requirements for data processing, privacy, documentation, and secure workflows, while professional licensure remains governed by national regulatory authorities [65-67]. International frameworks such as the European Union's General Data Protection Regulation, the US HIPAA (Health Insurance Portability and Accountability Act), and the World Health Organization (WHO) telemedicine guidance similarly emphasize legal clarity, data protection, and professional accountability in digital care [68-70].

Beyond regulation, accountability requires structured evaluation. WHO and related frameworks recommend staged assessment from pilot to scale-up, addressing effectiveness, equity, unintended harms, and economic impact [71,72]. For national virtual hospitals, this requires prospective and comparative designs rather than reliance on usage or satisfaction metrics alone [71]. Despite growing policy attention, most evaluations remain short-term and usage-focused, with limited evidence on equity, caregiver burden, workforce effects, long-term outcomes, and system-level financing [29,73-78].

Evidence Gaps

The evidence across tele-ICU, hospital-at-home, virtual-ward, and condition-specific models reveals persistent uncertainties regarding scalability, sustainability, and system-level impact [17,73,78]. First, clinical outcomes are heterogeneous and highly context-dependent [17,29]. While some tele-ICU studies reported mortality benefits, national analyses demonstrated only modest overall mortality reductions with substantial heterogeneity across hospitals after adjustment for case mix and hospital characteristics, suggesting that effects vary by patient selection, program design, workforce capacity, and system readiness, thereby limiting generalizability [17]. Methodological constraints, including selection bias, residual confounding, and short follow-up, further complicate interpretation [29]. Second, evidence in frail older adults, the primary target of virtual wards and hospital-at-home, remains mixed, where multidisciplinary home-based interventions show inconsistent effects on mortality, adverse events, and emergency usage, and transitional-care gains often attenuate over time [73-75]. Third, most evaluations emphasize 30-90–day outcomes, leaving uncertainty regarding long-term disease control, functional trajectories, cost-effectiveness, and unintended consequences such as over-surveillance or centralization of expertise [73-77]. Comparative-effectiveness evidence remains limited by heterogeneous intervention components and inconsistent outcome definitions [29]. Fourth, caregiver and social impacts are rarely measured, as caregiver strain, financial burden, and informal-care capacity are seldom incorporated into evaluation frameworks despite concerns about shifting responsibility to families [77-79]. Taken together, current evidence supports cautious optimism but does not yet establish consistent, long-term, system-level benefit for national virtual hospital deployment [17,73,78].

Policy Principles for National Virtual Hospitals

Embed Programs Within National Digital-Health Strategy and Governance

WHO's Global Strategy on Digital Health emphasizes that digital initiatives should be embedded within broader health-system goals and supported by governance structures that address interoperability, equity, and trust [80]. Telehealth

XSL•FO
RenderX

implementation frameworks similarly show that successful programs require attention not only to technology but also to organizational change, financing, user perceptions, and legal context [81]. Recent Australian work further demonstrates how national digital-health strategies can be mapped against value-based health care principles, highlighting priority populations, standardized outcome measurement, and patient-centered metrics as key processes for embedding value into digital care [82]. Another framework (Planning and Evaluating Remote Consultation Services) for remote consultations reinforces the need to integrate patient, clinical, organizational, and system perspectives, and highlights practical ethics, including distributive justice and data governance [83]. For virtual hospitals, this implies that expansion should be grounded in explicit strategy and regulation rather than driven solely by local innovation or vendor offerings [81-83].

Align Financing and Incentives With Substitution of Care

Financing design is central to the sustainability of national virtual hospitals. During the COVID-19 pandemic, many countries reimbursed telehealth at parity with in-person care and relaxed geographic restrictions, enabling rapid scale-up but raising concerns about long-term sustainability and potential overuse [84]. Evidence from value-based payment models shows that reimbursement structures determine whether telehealth substitutes for high-cost inpatient care or simply generates additional billable encounters [85,86]. National virtual hospital financing must therefore incentivize substitution, such as replacing inpatient bed-days with virtual care, while avoiding duplication and excess usage [86].

Economic sustainability should be embedded in program design and evaluation [87]. Standard health-economic methods, including cost-utility analysis using quality-adjusted life years, incremental cost-effectiveness ratios, and multiyear budget impact modeling, are directly applicable [44,71,72,87,88]. Evidence from hospital-at-home and virtual hospital-in-the-home programs suggests that remote acute care can reduce length of stay and, in selected populations, total costs without compromising outcomes [9,18,19,34-37]. Modeling and budget impact analyses further indicate potential cost savings and quality-adjusted life year gains when substitution effects outweigh additive usage [77,87,89]. Tele-ICU programs have similarly reported cost reductions associated with shorter ICU stays and fewer complications [31,90,91].

Prioritize Equity and Digital Inclusion in Design and Deployment

A digital-health equity framework argues that digital determinants of health, such as connectivity, device access, digital literacy, and trust, interact with traditional social determinants to shape who benefits from digital care [58]. Telehealth can be described as a “double-edged sword,” capable of reducing or reinforcing inequities depending on implementation [92]. Telehealth in China also highlights practical barriers, including technology costs, variable internet quality, and limited digital skills, particularly in rural areas [93]. Consequently, virtual hospital policies should embed equity by providing device and connectivity support where feasible, ensuring accessible interfaces, maintaining low-tech options such as telephone-based services, and routinely collecting disaggregated equity data [58].

Make Evaluation a Core Design Feature Rather Than an Afterthought

Digital programs are expected to follow a staged evaluation pathway from feasibility and pilots to pragmatic effectiveness and scale-up [71,72]. The Pan American Health Organization telemedicine framework recommends multidimensional evaluation across clinical effectiveness, patient experience, economics, organizational functioning, and socioethical or legal domains [94]. Evaluation questions for national virtual hospitals should be defined at inception, including impacts on mortality, functional outcomes, readmissions, caregiver burden, equity, and patient experience [71,72,94,95]. Linked data systems and independent academic collaboration are essential to ensure transparent reporting, including null or negative findings [71,72]. These principles highlight that national virtual hospitals are system-level interventions whose value depends on governance, financing, equity, and workforce sustainability rather than technology alone, and provide a framework for interpreting the SVH experience [80,81,83].

SVH as a National Case Study

Overview

SVH was launched in February 2022 under Saudi Arabia’s Vision 2030 Health Sector Transformation Program as a national-scale virtual hospital initiative. SVH operates as a centralized telehealth hub connecting hospitals with specialized clinical services across the Kingdom [25]. It is embedded within a broader national digital-health ecosystem developed by the Ministry of Health. This ecosystem includes the Seha application for remote consultations and the Sehhaty platform for access to personal health records and digital health services [96,97]. Care coordination is further supported by national infrastructure, including the 937 unified health hotline for teleconsultation and triage, the Mawid platform for appointment scheduling and referrals, and complementary systems such as Health Electronic Surveillance Network for disease surveillance, Tawakkalna for health-status verification, and Wasfaty for electronic prescribing and medication dispensing [98].

Operational Model and Clinical Workflow

Unlike many localized or specialty-specific models, SVH operates as a national, multispecialty virtual hospital within Saudi Arabia’s digital-health infrastructure, enabling examination of feasibility, operational scalability, and national-level integration across multiple clinical domains [99,100]. Operationally, SVH functions as a fully digital tertiary-care network connecting local hospitals, patients, and centralized specialist expertise through clinician-mediated and direct specialist-patient teleconsultations [96,99,100]. In many inpatient and high-acuity services, care begins when a patient presents to a local hospital, where the attending physician conducts a clinical assessment, orders diagnostic investigations, and documents findings [99]. Laboratory results, radiological images, and clinical notes are integrated within electronic health

XSL•FO
RenderX

record systems and related digital-health platforms supporting the SVH network [97,99]. When specialist input is required, the local physician initiates a consultation through SVH, and the SVH specialist remotely reviews clinical data, provides recommendations, and documents them in the shared electronic health record [99]. Treatment is subsequently implemented locally, while patients access results, prescriptions, and care plans via Sehhaty [97]. In addition, SVH delivers real-time virtual specialist–patient consultations, including urgent or critical care pathways (eg, telestroke), virtual specialty clinics (eg, cardiology and psychiatry), and live remote consultations supported by secure video links at local sites [25,100] (Figure 1).

Figure 1. SVH operational workflow. ICU: intensive care unit; SVH: Seha Virtual Hospital.

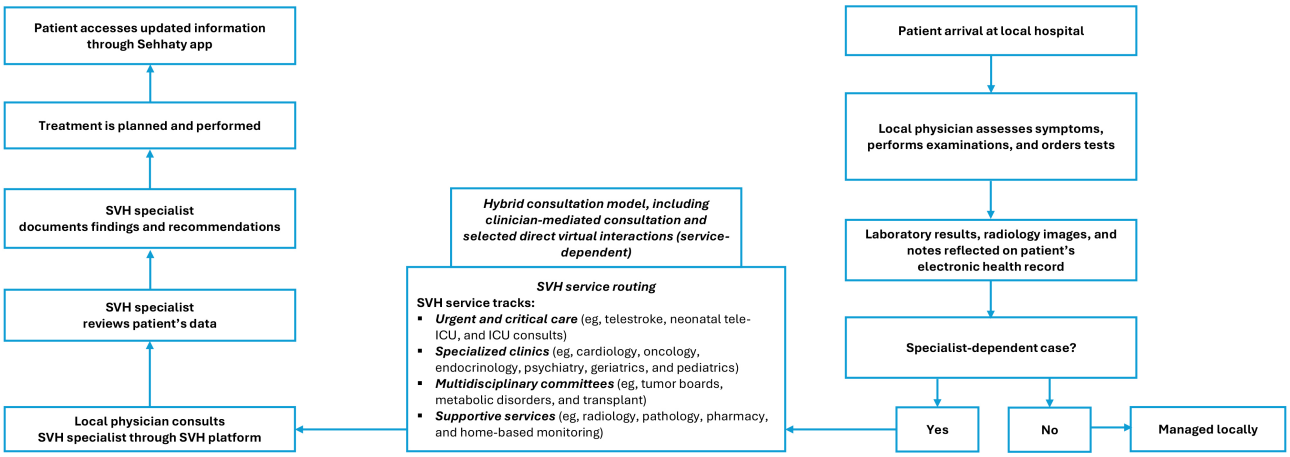

Scope of Services and Organizational Structure

As a key Vision 2030 initiative, SVH currently supports more than 170 hospitals, delivers services across 29 core specialties and 73 subspecialties, employs over 150 physicians, and has an annual capacity exceeding 480,000 patients [25,100]. The range of services spans clinical functions typically concentrated in tertiary referral settings and is delivered through coordinated virtual care streams [99,100]. Services are organized into four virtual tracks: urgent and critical care consultations, specialized clinics, multidisciplinary committees, and supportive medical services, ensuring timely access to expertise, improved efficiency of care, and knowledge transfer at scale [25,100]. Urgent and critical care services include time-sensitive presentations such as acute stroke, seizures, neonatal telecritical care, and intensive-care consultations [25,101,102]. Specialized clinics manage chronic and complex conditions across cardiology, endocrinology and diabetes, oncology, nephrology, hematology, geriatrics, psychiatry, rehabilitation medicine, and metabolic and genetic disorders [25,99]. Multidisciplinary committees support complex case-based decision-making, for example, through virtual heart or diabetes teams [25]. Supportive services include diagnostic reporting (radiology and pathology), pharmaceutical services, and home-based care pathways such as hospital-at-home models and remote physiological monitoring [25,99]. Collectively, these services allow SVH to function as a national tertiary hub for specialist consultation, diagnostic support, and integrated clinical management across geographically distributed hospitals [25,99,100].

Emerging Evidence and Evaluation Gaps

Given SVH's recent national rollout [25], independent evidence on long-term clinical, economic, equity, and workforce outcomes remains limited. Current SVH-related literature is largely descriptive, documenting growth in case volumes, service expansion, and national adoption, and positioning SVH as a central component of Saudi Arabia's telemedicine infrastructure [99]. While these findings demonstrate feasibility and operational scalability, they do not establish whether SVH improves clinical outcomes or overall system performance. The absence of external comparators further limits assessment of effectiveness and cost-effectiveness, and it remains unclear whether observed efficiencies reflect true substitution of inpatient care or expansion of service activity.

Early reports suggest pathway-specific benefits, including improved workflow in neonatal telecritical care and successful telestroke deployment during Hajj [101,102]. Health care workers report generally positive perceptions, although challenges related to connectivity, workload, and training persist [103]. Earlier Saudi tele-ICU studies show improved short-term outcomes [104], and national surveys indicate broad patient acceptance alongside usability concerns among older or digitally less confident populations [105-107]. However, these findings remain insufficient to establish a consistent system-level impact. It is therefore unclear whether SVH reduces avoidable hospital usage, redistributes workload, or introduces additional coordination demands. Similarly, the extent to which virtual consultations alter referral thresholds, escalation patterns, and interhospital transfers has not been systematically evaluated.

A central unresolved policy question is whether national virtual hospitals function primarily as substitutive models that reduce inpatient demand or as additive layers that increase overall system activity [77,86]. This distinction has direct implications for cost-effectiveness, workforce burden, and long-term sustainability [77]. Without robust comparative evaluation, apparent efficiency gains may reflect improved access or expanded usage rather than genuine system-level improvement [108]. Addressing these uncertainties requires prospective, comparative, and system-level evaluation designs, including standardized outcome measures, appropriate control groups, longitudinal follow-up, and explicit assessment of substitution

XSL•FO
RenderX

effects, patient trajectories, equity, and workforce impact [71,72,94].

Conceptual Implications

SVH aligns with global virtual-hospital trends through centralized specialist support, integration within a national digital-health strategy, and deployment in high-demand contexts such as mass gatherings [22,80,101,102]. However, its scale and scope suggest that national virtual hospitals should not be understood merely as telemedicine platforms, but as system-level reorganizations of care delivery [22,70,80]. This configuration may expand access to advanced care and reduce unnecessary transfers, particularly in underserved regions, but it also introduces new requirements for governance, coordination, and medico-legal accountability across distributed teams [24,63,64,69]. Conceptually, national virtual hospitals may be conceptualized as involving three interacting system transformations: (1) centralization of specialist expertise, (2) decentralization of care delivery, and (3) digital integration across care pathways. This triad distinguishes virtual hospitals from earlier telehealth models, which typically operate at the level of individual services rather than coordinated systems [11,22,70]. Unlike traditional hospitals defined by physical infrastructure, national virtual hospitals are distributed, network-based systems in which clinical expertise is decoupled from location but remains embedded within local care pathways [22,23]. Responsibility for clinical decision-making may therefore be distributed between virtual specialists and on-site providers, necessitating clearly defined professional roles, escalation pathways, documentation standards, and accountability mechanisms [63,64,66,69]. Conventional metrics such as bed occupancy, admission rates, or service volumes may fail to capture key system effects, including care redistribution, avoided transfers, changes in clinical decision-making, and coordination across networks [22,77]. Evaluation frameworks should therefore prioritize system-level indicators such as substitution of care, coordination efficiency, network-wide outcomes, equity of access, workforce impact, and cost-effectiveness [71,72,83,94]. Aligning evaluation with the distributed nature of virtual hospital care is essential for accurate assessment and accountable policy decision-making [71,83,94].

Interpreting Current Evidence and the Role of SVH

The international literature supports a cautious interpretation of national virtual hospitals. Outcomes vary according to patient selection, workforce capacity, escalation protocols, workflow integration, and digital maturity [41,52,77]. In this context, Saudi Arabia's SVH contributes primarily evidence of national feasibility and integration [99,100]. As a cross-specialty program embedded within a national digital ecosystem, SVH illustrates how specialist expertise can be distributed across geographically dispersed hospitals and linked to national platforms for records, scheduling, triage, and prescribing [25,97-99]. However, the current evidence base remains early-stage [99]. Hence, SVH should not yet be treated as a settled exemplar of effectiveness. The central policy question is not whether SVH can operate at scale, but whether its expansion translates into demonstrable system-level value when evaluated using rigorous prospective frameworks rather than descriptive growth metrics alone [109,110].

Policy Implications: Financing, Governance, Equity, and Workforce

The system-level nature of national virtual hospitals creates four immediate policy priorities. First, financing must be aligned with substitution rather than addition [86]. Without clear definitions of virtual activity and reimbursement structures that reward avoided admissions, reduced bed-days, or fewer transfers, programs risk increasing total encounters without commensurate system benefit [108]. Embedding transparent economic evaluation alongside implementation is therefore essential [71]. Second, governance and regulation must be operationalized [69]. National programs require clear definitions of clinical responsibility, licensing, scope of practice, documentation standards, data access, privacy safeguards, and liability across virtual and in-person teams, particularly in clinician-mediated models [66]. Third, equity should be treated as a design requirement rather than an assumed benefit [58]. National deployment offers an opportunity to measure and address digital and social inequities systematically, but gains require stratified reporting and targeted inclusion strategies [20,58]. Fourth, workforce sustainability must be explicitly monitored [62]. Evaluation should include staffing requirements, role evolution, burnout metrics, and the net impact on local teams [52]. Complementary innovations such as ambient documentation tools may reduce administrative burden, but their integration, safety, and governance implications require careful evaluation rather than an assumption of benefit [97].

From Service Delivery to a Learning Health System

A central implication of the SVH case is that national virtual hospitals are well-positioned to operate as learning health systems. The most credible route to resolving these uncertainties is to embed prospective evaluation into program operations, including pragmatic phased rollouts of new service lines, linked-data comparative analyses, and mixed-methods implementation research that explains why effects vary across sites and populations [71,72,83]. This agenda also supports the case for clinician-researcher roles and structured partnerships with academic institutions (and, where appropriate, industry), enabling transparent reporting of benefits, null findings, and unintended consequences [111]. Such an approach would move national virtual hospitals beyond descriptive growth metrics toward policy-relevant evidence on outcomes, equity, workforce impact, and value [71,72]. Despite these opportunities, several limitations warrant consideration.

This viewpoint synthesizes representative international evidence to inform policy interpretation rather than to provide exhaustive comparative analysis. Accordingly, conclusions reflect patterns across selected studies and may not capture the full range of findings across all contexts. Interpretation of the SVH case is

XSL•FO
RenderX

further constrained by the current availability of publicly reported data, which remains largely descriptive and short-term. As a policy-oriented analysis, this paper emphasizes system-level implications and governance considerations rather than direct comparative evaluation. Robust prospective and comparative assessments will be required to determine long-term effectiveness, equity impact, economic value, and workforce sustainability on a national scale.

Conclusions

National virtual hospitals are emerging as policy responses to rising hospital demand, workforce constraints, geographic inequities, and the need for flexible specialist care. Evidence from tele-ICU programs, hospital-at-home services, virtual wards, telestroke networks, and other virtual-care models suggests that digitally enabled care can achieve outcomes comparable to conventional hospital care when supported by appropriate patient selection, structured monitoring, clear escalation pathways, and effective integration into clinical workflows. However, the evidence base remains heterogeneous and context-dependent, with important gaps regarding long-term outcomes, cost-effectiveness, caregiver burden, workforce sustainability, digital equity, and whether virtual care substitutes for or adds to existing health care usage.

Saudi Arabia's SVH demonstrates the feasibility of a nationally coordinated, multispecialty virtual hospital embedded within a broader digital-health ecosystem, illustrating how centralized expertise can support geographically distributed care and improve access to specialist services. Nevertheless, SVH remains in an evidence-generating phase, as current evidence is largely descriptive and focused on operational expansion rather than robust comparative evaluation of clinical outcomes, equity, workforce impact, and economic value. National virtual hospitals should therefore be viewed not simply as technological innovations, but as complex health-system interventions requiring clear governance, sustainable financing, regulatory accountability, and explicit attention to digital inclusion and workforce capacity. Embedding prospective, comparative, and system-level evaluation frameworks within implementation will be essential to determine whether these models deliver measurable improvements in outcomes, efficiency, equity, and overall health-system performance.

Acknowledgments

Disclosure of delegation to generative artificial intelligence (GenAI): the authors declare the use of GenAI in the research and writing process. According to the GAIDeT (2025; Generative Artificial Intelligence Delegation Taxonomy), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing. The GenAI tool used was ChatGPT (OpenAI). The responsibility for this paper lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by SS.

Funding

The authors declare no financial support was received for this work.

Data Availability

Data sharing does not apply to this paper as no datasets were generated or analyzed during this study.

Authors' Contributions

Conceptualization, investigation, supervision, writing – original draft, visualization, and project administration: SS
Data curation and writing – review & editing: HA
Validation and writing – review & editing: SHA
Writing – review & editing: HM

Conflicts of Interest

None declared.

Multimedia Appendix 1
International virtual-care models beyond national multispecialty virtual hospitals.
[DOCX File, 30 KB - jmir_v28i1e89276_app1.docx]

References

1. Faitna P, Bottle A, Klaber B, Aylin PP. Has multimorbidity and frailty in adult hospital admissions changed over the last 15 years? A retrospective study of 107 million admissions in England. BMC Med 2024 Sep 11;22(1):369. [doi: 10.1186/s12916-024-03572-z] [Medline: 39256751]
2. Skou ST, Mair FS, Fortin M, et al. Multimorbidity. Nat Rev Dis Primers 2022 Jul 14;8(1):48. [doi: 10.1038/s41572-022-00376-4] [Medline: 35835758]

XSL•FO
RenderX

3. Yordanov Y, Beauvais A, Thiébaud PC. Multimorbidity in emergency departments: urgent need for integrated care. BMJ Med 2024;3(1):e000989. [doi: 10.1136/bmjmed-2024-000989] [Medline: 39175922]
4. Afshar S, Roderick PJ, Kowal P, Dimitrov BD, Hill AG. Multimorbidity and the inequalities of global ageing: a cross-sectional study of 28 countries using the World Health Surveys. BMC Public Health 2015 Aug 13;15:776. [doi: 10.1186/s12889-015-2008-7] [Medline: 26268536]
5. Vos T, Lim SS, Abbafati C, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct;396(10258):1204-1222. [doi: 10.1016/S0140-6736(20)30925-9]
6. Noncommunicable diseases. World Health Organization. 2025 Sep 25. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases [accessed 2026-06-02]
7. Norman G, Bennett P, Vardy ERLC. Virtual wards: a rapid evidence synthesis and implications for the care of older people. Age Ageing 2023 Jan 8;52(1):afac319. [doi: 10.1093/ageing/afac319] [Medline: 36633298]
8. Mann DM, Chen J, Chunara R, Testa PA, Nov O. COVID-19 transforms health care through telemedicine: evidence from the field. J Am Med Inform Assoc 2020 Jul 1;27(7):1132-1135. [doi: 10.1093/jamia/ocaa072] [Medline: 32324855]
9. Shi C, Dumville J, Rubinstein F, et al. Inpatient-level care at home delivered by virtual wards and hospital at home: a systematic review and meta-analysis of complex interventions and their components. BMC Med 2024 Apr 2;22(1):145. [doi: 10.1186/s12916-024-03312-3] [Medline: 38561754]
10. Wosik J, Fudim M, Cameron B, et al. Telehealth transformation: COVID-19 and the rise of virtual care. J Am Med Inform Assoc 2020 Jun 1;27(6):957-962. [doi: 10.1093/jamia/ocaa067] [Medline: 32311034]
11. Schwamm LH, Estrada J, Erskine A, Licurse A. Virtual care: new models of caring for our patients and workforce. Lancet Digit Health 2020 Jun;2(6):e282-e285. [doi: 10.1016/S2589-7500(20)30104-7] [Medline: 32382724]
12. AlSalamah S. VCAC: a blockchain-based virtual care access control model for transforming legacy healthcare information systems and EMRs into secure, interoperable patient-centered virtual hospital systems. Information 2025;16(11):972. [doi: 10.3390/info16110972]
13. Wilcox ME, Adhikari NKJ. The effect of telemedicine in critically ill patients: systematic review and meta-analysis. Crit Care 2012 Jul 18;16(4):R127. [doi: 10.1186/cc11429] [Medline: 22809335]
14. Lilly CM, McLaughlin JM, Zhao H, Baker SP, Cody S, Irwin RS. UMass memorial critical care operations group. A multicenter study of ICU telemedicine reengineering of adult critical care. Chest 2014;145(3):500-507. [doi: 10.1378/chest.13-1973] [Medline: 24306581]
15. Westby M, Ijaz S, Savović J, et al. Virtual wards for people with frailty: what works, for whom, how and why-a rapid realist review. Age Ageing 2024 Mar 1;53(3):afae039. [doi: 10.1093/ageing/afae039] [Medline: 38482985]
16. Alhmoud B. Are we close enough? An examination of the innovative virtual wards' care delivery. Saudi J Health Syst Res 2024;4(4):214-217. [doi: 10.1159/000540644]
17. Kahn JM, Le TQ, Barnato AE, et al. ICU telemedicine and critical care mortality: a national effectiveness study. Med Care 2016 Mar;54(3):319-325. [doi: 10.1097/MLR.0000000000000485] [Medline: 26765148]
18. Arsenault-Lapierre G, Henein M, Gaid D, Le Berre M, Gore G, Vedel I. Hospital-at-home interventions vs in-hospital stay for patients with chronic disease who present to the emergency department: a systematic review and meta-analysis. JAMA Netw Open 2021 Jun 1;4(6):e2111568. [doi: 10.1001/jamanetworkopen.2021.11568] [Medline: 34100939]
19. Shepperd S, Iliffe S, Doll HA, et al. Admission avoidance hospital at home. Cochrane Database Syst Rev 2016 Sep 1;9(9):CD007491. [doi: 10.1002/14651858.CD007491.pub2] [Medline: 27583824]
20. Cantor JH, McBain RK, Pera MF, Bravata DM, Whaley CM. Who is (and is not) receiving telemedicine care during the COVID-19 pandemic. Am J Prev Med 2021 Sep;61(3):434-438. [doi: 10.1016/j.amepre.2021.01.030] [Medline: 33781622]
21. Spaulding EM, Fang M, Commodore-Mensah Y, Himmelfarb CR, Martin SS, Coresh J. Prevalence and disparities in telehealth use among US adults following the COVID-19 pandemic: national cross-sectional survey. J Med Internet Res 2024 May 10;26:e52124. [doi: 10.2196/52124] [Medline: 38728070]
22. Bidoli C, Pegoraro V, Dal Mas F, et al. Virtual hospitals: the future of the healthcare system? An expert consensus. J Telemed Telecare 2025 Jan;31(1):121-133. [doi: 10.1177/1357633X231173006] [Medline: 37226478]
23. Fisher OJ, Grogan C, Barron A, et al. Considerations for establishment of a private virtual hospital identified using an implementation science approach. Sci Rep 2025 Jan 29;15(1):3660. [doi: 10.1038/s41598-025-85965-5] [Medline: 39880892]
24. Kumari AA, Wani TA, Liem M, Boyd J, Khan UR. Advancing regional and remote health care with virtual hospital implementation: rapid review. JMIR Hum Factors 2025 Jun 3;12:e64582. [doi: 10.2196/64582] [Medline: 40460425]
25. Seha virtual hospital. Ministry of Health. URL: https://www.moh.gov.sa/en/Ministry/Projects/Pages/Seha-Virtual-Hospital.aspx [accessed 2026-06-02]
26. Keskimäki I, Mäkelä M, Koivusalo M. Finland: health system review. World Health Organization. 2019. URL: https://iris.who.int/server/api/core/bitstreams/af57d983-5594-4230-9544-9bdaada63d77/content [accessed 2026-06-02]
27. Ministry of Health (Jordan). 2025. URL: https://www.moh.gov.jo/EBV4.0/Root_Storage/AR/fadi/%D8%A7%D8%B3%D8%AA%D8%B1%D8%B4%D8%A7%D8%AF%D8%A7%D8%AA%D9%88%D8%B2%D8%A7%D8%B1%D8%A9%D8%A7%D9%84%D8%B5%D8%AD%D8%A9%D8%A7%D9%84%D8%B1%D9%82%D9%85%D9%8A.pdf [accessed 2026-01-28]

XSL•FO
RenderX

28. Vranas KC, Slatore CG, Kerlin MP. Telemedicine coverage of intensive care units: a narrative review. Ann Am Thorac Soc 2018 Nov;15(11):1256-1264. [doi: 10.1513/AnnalsATS.201804-225CME] [Medline: 30382786]
29. Spies CD, Ranzani OT. Telemedicine in critical care. Intensive Care Med 2025 Jun;51(6):1153-1156. [doi: 10.1007/s00134-025-07955-5] [Medline: 40471258]
30. Reynolds HN. The tele-ICU: formative or out-of-date or both? Practice models and future directions. In: Koenig MA, editor. Telemedicine in the ICU: Springer; 2019:3-19. [doi: 10.1007/978-3-030-11569-2_1]
31. Armaignac DL, Saxena A, Rubens M, et al. Impact of telemedicine on mortality, length of stay, and cost among patients in progressive care units: experience from a large healthcare system. Crit Care Med 2018 May;46(5):728-735. [doi: 10.1097/CCM.0000000000002994] [Medline: 29384782]
32. Graves JM, Krings JG, Buss JL, Kallogjeri D, Ofoma UR. Telemedicine critical care availability and outcomes among mechanically ventilated patients. J Crit Care 2024 Aug;82:154782. [doi: 10.1016/j.jcrc.2024.154782] [Medline: 38522373]
33. Orrall A. The next step in hospital-at-home care could be virtual. JAMA 2025 Jan 14;333(2):109-110. [doi: 10.1001/jama.2024.24210] [Medline: 39671219]
34. Levine DM, Ouchi K, Blanchfield B, et al. Hospital-level care at home for acutely ill adults: a randomized controlled trial. Ann Intern Med 2020 Jan 21;172(2):77-85. [doi: 10.7326/M19-0600] [Medline: 31842232]
35. Levine DM, Ouchi K, Blanchfield B, et al. Hospital-level care at home for acutely ill adults: a pilot randomized controlled trial. J Gen Intern Med 2018 May;33(5):729-736. [doi: 10.1007/s11606-018-4307-z]
36. Levine DM, Paz M, Burke K, et al. Remote vs in-home physician visits for hospital-level care at home. JAMA Netw Open 2022 Aug 1;5(8):e2229067. [doi: 10.1001/jamanetworkopen.2022.29067]
37. Conley J, O'Brien CW, Leff BA, Bolen S, Zulman D. Alternative strategies to inpatient hospitalization for acute medical conditions: a systematic review. JAMA Intern Med 2016 Nov 1;176(11):1693-1702. [doi: 10.1001/jamainternmed.2016.5974] [Medline: 27695822]
38. Gonçalves-Bradley DC, Iliffe S, Doll HA, et al. Early discharge hospital at home. Cochrane Database Syst Rev 2017;2021(7):CD000356. [doi: 10.1002/14651858.CD000356.pub4]
39. Shepperd S, Cradduck-Bamford A, Butler C, et al. Hospital at Home admission avoidance with comprehensive geriatric assessment to maintain living at home for people aged 65 years and over: a RCT. Health Soc Care Delivery Res 2022;10(2):1-124. [doi: 10.3310/HTAF1569]
40. Knight T, Lasserson D. Hospital at home for acute medical illness: the 21st century acute medical unit for a changing population. J Intern Med 2022 Apr;291(4):438-457. [doi: 10.1111/joim.13394] [Medline: 34816527]
41. Akinluyi A, Ogunfusika OH, Brown A. The effectiveness of virtual wards compared to inpatient beds in clinical outcomes of frail older patients with chronic conditions: a systematic review. Cureus 2025 Nov;17(11):e97487. [doi: 10.7759/cureus.97487] [Medline: 41312075]
42. Lewis C, Moore Z, Doyle F, Martin A, Patton D, Nugent LE. A community virtual ward model to support older persons with complex health care and social care needs. Clin Interv Aging 2017;12:985-993. [doi: 10.2147/CIA.S130876] [Medline: 28721026]
43. Grange H, Johns G, Ahuja A, Harper P, Williams E, Gartner D. Capacity planning of virtual wards for frail and elderly patients. Healthcare (Basel) 2024 Feb 23;12(5):533. [doi: 10.3390/healthcare12050533] [Medline: 38470644]
44. Tomini SM, Massou E, Crellin NE, et al. A cost evaluation of COVID-19 remote home monitoring services in England. Pharmacoecon Open 2024 Sep;8(5):739-753. [doi: 10.1007/s41669-024-00498-3] [Medline: 38951349]
45. Hutchings OR, Dearing C, Jagers D, et al. Virtual health care for community management of patients with COVID-19 in Australia: observational cohort study. J Med Internet Res 2021 Mar 9;23(3):e21064. [doi: 10.2196/21064] [Medline: 33687341]
46. Shaw M, Hutchings O, Anderson T, Jagers D, Raffan F. The role of Australia's first virtual hospital in the COVID-19 pandemic response: learnings and opportunities. Int J Integr Care 2022;22(S3):334. [doi: 10.5334/ijic.ICIC22169]
47. Mohamed A, Elsherif S, Legere B, Fatima N, Shuaib A, Saqqur M. Is telestroke more effective than conventional treatment for acute ischemic stroke? A systematic review and meta-analysis of patient outcomes and thrombolysis rates. Int J Stroke 2024 Mar;19(3):280-292. [doi: 10.1177/17474930231206066] [Medline: 37752674]
48. Kepplinger J, Barlinn K, Deckert S, Scheibe M, Bodechtel U, Schmitt J. Safety and efficacy of thrombolysis in telestroke: a systematic review and meta-analysis. Neurology (ECronicon) 2016 Sep 27;87(13):1344-1351. [doi: 10.1212/WNL.0000000000003148] [Medline: 27566746]
49. Simon E, Forghani M, Abramyuk A, et al. Intravenous thrombolysis by telestroke in the 3- to 4.5-h time window. Front Neurol 2021;12:756062. [doi: 10.3389/fneur.2021.756062] [Medline: 34899575]
50. Curfman AL, Haycraft M, McSwain SD, Dooley M, Simpson KN. Implementation and evaluation of a wraparound virtual care program for children with medical complexity. Telemed J E Health 2023 Jun;29(6):947-953. [doi: 10.1089/tmj.2022.0344] [Medline: 36355064]
51. Tao Y, Vo A, Wu D, et al. Impact of telehealth on health disparities associated with travel time to hospital for patients with recurrent admissions: 4-year panel data analysis. J Med Internet Res 2024;26:e63661. [doi: 10.2196/63661]
52. Sunny T, Ravichandran N, Broughan J, et al. Practitioners' perspectives on implementation of acute virtual wards: a scoping review. PLOS Digit Health 2025 May;4(5):e0000858. [doi: 10.1371/journal.pdig.0000858] [Medline: 40343985]

XSL•FO
RenderX

53. Keuper J, van Tuyl LHD, de Geit E, et al. The impact of eHealth use on general practice workload in the pre-COVID-19 era: a systematic review. BMC Health Serv Res 2024 Sep 19;24(1):1099. [doi: 10.1186/s12913-024-11524-9] [Medline: 39300456]
54. Apaydin EA, Der-Martirosian C, Yoo C, et al. Secure messages, video visits, and burnout among primary care providers in the Veterans Health Administration: national survey study. J Med Internet Res 2025 Sep 5;27:e68858. [doi: 10.2196/68858] [Medline: 40911914]
55. Olson KD, Meeker D, Troup M, et al. Use of ambient AI scribes to reduce administrative burden and professional burnout. JAMA Netw Open 2025 Oct 1;8(10):e2534976. [doi: 10.1001/jamanetworkopen.2025.34976] [Medline: 41037268]
56. Duggan MJ, Gervase J, Schoenbaum A, et al. Clinician experiences with ambient scribe technology to assist with documentation burden and efficiency. JAMA Netw Open 2025 Feb 3;8(2):e2460637. [doi: 10.1001/jamanetworkopen.2024.60637]
57. Leung TI, Coristine AJ, Benis A. AI scribes in health care: balancing transformative potential with responsible integration. JMIR Med Inform 2025 Aug 1;13:e80898. [doi: 10.2196/80898] [Medline: 40749188]
58. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119. [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
59. Wang S, von Huben A, Sivaprakash PP, Saurman E, Norris S, Wilson A. Addressing health service equity through telehealth: a systematic review of reviews. Digit Health 2025;11:20552076251326233. [doi: 10.1177/20552076251326233] [Medline: 40177121]
60. Petretto DR, Carrogu GP, Gaviano L, et al. Telemedicine, e-Health, and digital health equity: a scoping review. Clin Pract Epidemiol Ment Health 2024;20:e17450179279732. [doi: 10.2174/0117450179279732231211110248] [Medline: 38660571]
61. Goldberg N, Herrmann C, Di Gion P, et al. Sociodemographic and socioeconomic determinants for the usage of digital patient portals in hospitals: systematic review and meta-analysis on the digital divide. J Med Internet Res 2025 Jun 3;27:e68091. [doi: 10.2196/68091] [Medline: 40460427]
62. Ghaseminejad F, Rich K, Rosenbaum D, et al. Organisational factors associated with burnout among emergency and internal medicine physicians: a qualitative study. BMJ Open 2025 Jan 28;15(1):e085973. [doi: 10.1136/bmjopen-2024-085973] [Medline: 39880435]
63. Jefferies M, Graham R, Tracy M, et al. Telehealth use and legal considerations in drug health services during pandemics: systematic scoping review. J Med Internet Res 2024 Nov 12;26:e46394. [doi: 10.2196/46394] [Medline: 39531642]
64. Orsayeva R, Berestova A, Krasilnikova V, Timoshin A. Telemedicine during COVID- 19: features of legal regulation in the field of administrative liability for errors committed by medical institutions. Egypt J Forensic Sci 2025;15(1):26. [doi: 10.1186/s41935-025-00443-3]
65. Saudi Data & AI Authority. Personal data protection law (English translation, version 2 – 23 April 2023). : Saudi Data & AI Authority; 2023. URL: https://sdaia.gov.sa/en/SDAIA/about/Documents/Personal%20Data%20English%20V2-23April2023-%20Reviewed-.pdf [accessed 2026-01-29]
66. Telehealth application guidelines. : National Health Information Center URL: https://nhic.gov.sa/standards/Telehealth/Telehealth-Application-Guidelines.pdf [accessed 2026-06-02]
67. Privacy and terms of use. Saudi Commission for Health Specialties (SCFHS). URL: https://scfhs.org.sa/en/privacy-and-terms-of-use [accessed 2026-06-02]
68. General data protection regulation (GDPR). Intersoft consulting. URL: https://gdpr-info.eu/ [accessed 2026-06-02]
69. Telehealth around the world: a global guide. : DLA Piper; 2020 Nov URL: https://mplassociation.org/docs/Newsbriefs/Documents/DLA-Piper-Global-Telehealth-Guide-December-2020.pdf [accessed 2026-06-02]
70. Consolidated telemedicine implementation guide. : World Health Organization; 2022 URL: https://iris.who.int/server/api/core/bitstreams/d119a69e-e56d-4c78-a2b9-52572e14217c/content [accessed 2026-06-02]
71. Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment. : World Health Organization; 2016 Dec 12 URL: https://iris.who.int/server/api/core/bitstreams/0036e491-a9f4-450e-8943-11225d1c7ddf/content [accessed 2026-06-02]
72. Murray E, Hekler EB, Andersson G, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016 Nov;51(5):843-851. [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
73. Ruiz-Grao MC, Álvarez-Bueno C, Garrido-Miguel M, Berlanga-Macias C, Gonzalez-Molinero M, Rodríguez-Martín B. Multidisciplinary home-based interventions in adverse events and quality of life among frail older people: a systematic review and meta-analysis. Heliyon 2024 Nov 15;10(21):e40015. [doi: 10.1016/j.heliyon.2024.e40015] [Medline: 39583819]
74. Lembeck MA, Thygesen LC, Sørensen BD, Rasmussen LL, Holm EA. Effect of single follow-up home visit on readmission in a group of frail elderly patients - a Danish randomized clinical trial. BMC Health Serv Res 2019 Oct 25;19(1):751. [doi: 10.1186/s12913-019-4528-9] [Medline: 31653219]
75. Fønss Rasmussen L, Grode LB, Lange J, Barat I, Gregersen M. Impact of transitional care interventions on hospital readmissions in older medical patients: a systematic review. BMJ Open 2021 Jan;11(1):e040057. [doi: 10.1136/bmjopen-2020-040057]

XSL•FO
RenderX

76. Hatef E, Wilson RF, Hannum SM, et al. Use of telehealth during the COVID-19 era. : Agency for Healthcare Research and Quality (US); 2023 Jan Report No.: 23-EHC005 URL: https://effectivehealthcare.ahrq.gov/sites/default/files/related_files/use-telehealth-during-COVID-19-executive-summary.pdf [accessed 2026-06-02]
77. Kadakia KT, Balatbat CA, Siu AL, et al. Hospital-at-home: multistakeholder considerations for program dissemination and scale. Milbank Q 2022 Sep;100(3):673-701. [doi: 10.1111/1468-0009.12586] [Medline: 36148893]
78. Kilaru AS, Zikry HE. The role of emergency medicine in the hospital-at-home. Ann Emerg Med 2025 Sep;86(3):232-239. [doi: 10.1016/j.annemergmed.2025.03.020] [Medline: 40272328]
79. Kristensen GS, Søndergaard J, Andersen-Ranberg K, Mogensen CB. Acute readmissions among care home residents aged 65+ years: a register-based study. Eur Geriatr Med 2025 Jun;16(3):827-838. [doi: 10.1007/s41999-025-01162-7] [Medline: 39982627]
80. Global strategy on digital health 2020–2025. : World Health Organization; 2021 Aug 18 URL: https://iris.who.int/server/api/core/bitstreams/1f4d4a08-b20d-4c36-9148-a59429ac3477/content [accessed 2026-06-02]
81. van Dyk L. A review of telehealth service implementation frameworks. Int J Environ Res Public Health 2014 Jan 23;11(2):1279-1298. [doi: 10.3390/ijerph110201279] [Medline: 24464237]
82. Tait P, Daff D, Everingham P, et al. Enhancing digital healthcare: aligning Australia's digital health strategies with value-based healthcare principles. Aust Health Rev 2025 Jun;49:AH25022. [doi: 10.1071/AH25022] [Medline: 40457515]
83. Greenhalgh T, Rosen R, Shaw SE, et al. Planning and evaluating remote consultation services: a new conceptual framework incorporating complexity and practical ethics. Front Digit Health 2021;3:726095. [doi: 10.3389/fdgth.2021.726095] [Medline: 34713199]
84. VanderWerf M, Bernard J, Barta DT, et al. Pandemic action plan policy and regulatory summary telehealth policy and regulatory considerations during a pandemic. Telemed J E Health 2022 Apr;28(4):457-466. [doi: 10.1089/tmj.2021.0216] [Medline: 34265216]
85. Scott Kruse C, Karem P, Shifflett K, Vegi L, Ravi K, Brooks M. Evaluating barriers to adopting telemedicine worldwide: a systematic review. J Telemed Telecare 2018 Jan;24(1):4-12. [doi: 10.1177/1357633X16674087] [Medline: 29320966]
86. Huang-Ku E, Muenkaew P, Chavarina KK, et al. Telemedicine public reimbursement models for national and subnational jurisdictions: scoping review. J Med Internet Res 2025 Aug 12;27:e75478. [doi: 10.2196/75478] [Medline: 40794105]
87. Peters GM, Doggen CJM, van Harten WH. Budget impact analysis of providing hospital inpatient care at home virtually, starting with two specific surgical patient groups. BMJ Open 2022 Aug 1;12(8):e051833. [doi: 10.1136/bmjopen-2021-051833] [Medline: 35914920]
88. Gentili A, Failla G, Melnyk A, et al. The cost-effectiveness of digital health interventions: a systematic review of the literature. Front Public Health 2022;10:787135. [doi: 10.3389/fpubh.2022.787135] [Medline: 36033812]
89. Vo LK, Allen MJ, McPhail SM, et al. The effectiveness and cost-effectiveness of a virtual hospital in the home service for COVID-19 infection: a cohort study and modelled decision analysis. J Telemed Telecare 2026 Jun;32(4):348-358. [doi: 10.1177/1357633X251340262]
90. Lilly CM, Cody S, Zhao H, et al. Hospital mortality, length of stay, and preventable complications among critically ill patients before and after tele-ICU reengineering of critical care processes. JAMA 2011 Jun 1;305(21):2175. [doi: 10.1001/jama.2011.697]
91. Franzini L, Sail KR, Thomas EJ, Wueste L. Costs and cost-effectiveness of a telemedicine intensive care unit program in 6 intensive care units in a large health care system. J Crit Care 2011 Jun;26(3):329. [doi: 10.1016/j.jcrc.2010.12.004] [Medline: 21376515]
92. Gergen Barnett K, Mishuris RG, Williams CT, et al. Telehealth's double-edged sword: bridging or perpetuating health inequities? J Gen Intern Med 2022 Aug;37(11):2845-2848. [doi: 10.1007/s11606-022-07481-w]
93. Ye J, He L, Beestrum M. Implications for implementation and adoption of telehealth in developing countries: a systematic review of China's practices and experiences. NPJ Digit Med 2023 Sep 18;6(1):174. [doi: 10.1038/s41746-023-00908-6] [Medline: 37723237]
94. Framework for the implementation of a telemedicine service. : Pan American Health Organization; 2016 URL: https://pesquisa.bvsalud.org/bvsecuador/resource/en/biblio-1348377 [accessed 2026-06-02]
95. Jabour AM. A comparative study of patient-physician empathy in telehealth and traditional in-person visits. Digit Health 2024;10:20552076241305284. [doi: 10.1177/20552076241305284] [Medline: 39679005]
96. Alghamdi SM, Alqahtani JS, Aldhahir AM. Current status of telehealth in Saudi Arabia during COVID-19. J Family Community Med 2020;27(3):208-211. [doi: 10.4103/jfcm.JFCM_295_20] [Medline: 33354152]
97. Alassaf N, Bah S, Almulhim F, AlDossary N, Alqahtani M. Evaluation of official healthcare informatics applications in Saudi Arabia and their role in addressing COVID-19 pandemic. Healthc Inform Res 2021 Jul;27(3):255-263. [doi: 10.4258/hir.2021.27.3.255]
98. Alkhalifah JM, Seddiq W, Alshehri BF, Alhaluli AH, Alessa MM, Alsulais NM. The role of the COVID-19 pandemic in expediting digital health-care transformation: Saudi Arabia's experience. Inform Med Unlocked 2022;33:101097. [doi: 10.1016/j.imu.2022.101097] [Medline: 36185732]

XSL•FO
RenderX

99. Alabdulaali MK, Aldossary MS, Alsubaie MS, Zaatari ES, Alhomod AS. Digital transformation in telemedicine health care: a retrospective analysis of Seha Virtual Hospital outcomes in the Kingdom of Saudi Arabia. Telemed J E Health 2025 Dec 10;32(1):67-73. [doi: 10.1177/15305627251387590] [Medline: 41081638]
100. Seha Virtual Hospital. : Ministry of Health URL: https://www.moh.gov.sa/en/Ministry/Projects/Documents/Seha-Virtual-Hospital.pdf [accessed 2026-06-02]
101. Mominkhan DM, Aldahmashi F, Almudeer AH, et al. Implementing telemedicine intervention in neonatal intensive care units: augmented teleconsultation and real-time monitoring experience. Telemed Rep 2025;6(1):50-57. [doi: 10.1089/tmr.2024.0088] [Medline: 40151791]
102. Alsaleh GS, Alhomod ASA, Khan A, Alzahrani SS. Telestroke management during the Hajj seasons 2023-2024: insights from SEHA Virtual Hospital in KSA. Front Neurol 2025;16:1573275. [doi: 10.3389/fneur.2025.1573275] [Medline: 41323225]
103. Alghamdi SM. Clinical training and practice of using digital health solutions among healthcare workers in clinical settings during Hajj season. J Pak Med Assoc 2024 Sep;74(9):1678-1680. [doi: 10.47391/JPMA.11216] [Medline: 39279075]
104. Al-Omari A, Al Mutair A, Al Ammary M, Aljamaan F. A multicenter case-historical control study on short-term outcomes of tele-intensive care unit. Telemed J E Health 2020 May;26(5):645-650. [doi: 10.1089/tmj.2019.0042] [Medline: 31436501]
105. Abdel Nasser A, Mohammed Alzahrani R, Aziz Fellah C, et al. Measuring the patients' satisfaction about telemedicine used in Saudi Arabia during COVID-19 pandemic. Cureus 2021 Feb 16;13(2):e13382. [doi: 10.7759/cureus.13382] [Medline: 33754105]
106. Alajwari HA, Alfayez A, Alsalman D, et al. Knowledge and attitude of Saudi Arabian citizens towards telemedicine during the COVID-19 pandemic. Int Health 2022 Nov 1;14(6):604-609. [doi: 10.1093/inthealth/ihab082] [Medline: 34893850]
107. Sayed SH, Aldailami DA, Abd El Aziz MM, Elsayed EA. Perceived telehealth usability for personalized healthcare among the adult population in Saudi Arabia: a cross-sectional study in the post-COVID-19 era. Healthcare (Basel) 2025 Jan 1;13(1):62. [doi: 10.3390/healthcare13010062] [Medline: 39791669]
108. Ashwood JS, Mehrotra A, Cowling D, Uscher-Pines L. Direct-to-consumer telehealth may increase access to care but does not decrease spending. Health Aff (Millwood) 2017 Mar 1;36(3):485-491. [doi: 10.1377/hlthaff.2016.1130] [Medline: 28264950]
109. Kruse CS, Williams K, Bohls J, Shamsi W. Telemedicine and health policy: a systematic review. Health Policy Technol 2021 Mar;10(1):209-229. [doi: 10.1016/j.hlpt.2020.10.006]
110. Poudineh Moghadam M, Bazi A, Esfahani P, Emadi V. Benefits and outcomes of implementing virtual health clinics in remote and underserved areas: a scoping review. Int J Technol Assess Health Care 2026;42(1):e1. [doi: 10.1017/S0266462325103310]
111. Reid RJ, Wodchis WP, Kuluski K, et al. Actioning the Learning Health System: an applied framework for integrating research into health systems. SSM - health syst 2024 Jun;2:100010. [doi: 10.1016/j.ssmhs.2024.100010]

Abbreviations

HIPAA : Health Insurance Portability and Accountability Act
ICU: intensive care unit
SVH: Seha Virtual Hospital
tele-ICU: tele–intensive care unit
WHO: World Health Organization

Edited by S Brini; submitted 09.Dec.2025; peer-reviewed by P Tait, S Larkin; revised version received 14.May.2026; accepted 20.May.2026; published 22.Jun.2026.

Please cite as:
Shujaat S, Almutairi H, Alhuraibi SH, Ma H
Policy Considerations for National Virtual Hospitals: Global Evidence and the Seha Virtual Hospital Model
J Med Internet Res 2026;28:e89276
URL: https://www.jmir.org/2026/1/e89276
doi:10.2196/89276

© Sohaib Shujaat, Hawazin Almutairi, Saleh Hamad Alhuraibi, Hongyang Ma. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN

XSL•FO
RenderX

1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Analysis of Multilevel Factors Mobilizing the Spectrum of Interorganizational Knowledge Sharing for Facilitating Digital Transformation at Scale: Qualitative Study

Hajar Mozaffar[1], PhD; Robin Williams[2], PhD; Kathrin Cresswell[3], PhD

[1]Business School, University of Edinburgh, 29 Buccleuch Place, Edinburgh, United Kingdom

[2]Institute for the Study of Science, Technology and Innovation, University of Edinburgh, Edinburgh, United Kingdom

[3]Usher Institute, University of Edinburgh, Edinburgh, United Kingdom

Corresponding Author:
Hajar Mozaffar, PhD
Business School, University of Edinburgh, 29 Buccleuch Place, Edinburgh, United Kingdom

Abstract

Background: Interorganizational knowledge sharing is vital for scaling digital transformation efforts that span multiple organizations and system-wide change. However, existing frameworks provide limited insights into the cross-level dynamics that shape how learning ecosystems emerge, evolve, and operate across multiple organizations. This gap leaves practitioners without clear guidance on how multilevel contextual conditions and mechanisms interact to influence the development and sustainability of formal and informal knowledge-sharing relationships.

Objective: This study aimed to examine how knowledge is orchestrated across organizations in the digital transformation of health care, identifying key factors that foster an evolving interorganizational learning ecosystem. We developed an integrative model that explains how these influences give rise to diverse modes of collaboration and partnership.

Methods: We adopted a qualitative approach using a multilevel perspective to examine visions and experiences across individual, organizational, interorganizational, and sectoral levels. Drawing on a formative evaluation (2018 - 2023) of England's Global Digital Exemplar (GDE) program, we used multiple case studies and conducted interviews with experts both within and beyond organizational settings for data collection and adopted a grounded theory approach to analyze the data.

Results: The study identified a set of interconnected factors operating at the macroenvironmental, interorganizational, organizational, and individual levels that influence how interorganizational relationships and partnerships are initiated, structured, and sustained. Macro-level influences included policy developments, program mandates, technology supplier strategies, and intermediary actions. Interorganizational mechanisms involved relational recognition, collective identity, governance structures, proximity, and coordination practices. Organizational factors included external search strategies, absorptive capacity, past collaboration experience, and internal knowledge routines. Individual-level mechanisms encompassed intrinsic and extrinsic motivations as well as personal inhibitors. Synthesizing these findings, we have proposed an integrative model that positions relationship type along a 2D spectrum (formal-informal, internal-external origins) and illustrates how different factors trigger, mandate, control, and enable the evolution of an interorganizational learning ecosystem.

Conclusions: This study advances the theoretical understanding of learning ecosystems by explaining how multilevel contextual conditions activate mechanisms that give rise to diverse and evolving forms of interorganizational collaboration. Practically, we offer diagnostic and reflective tools that support policymakers and practitioners in assessing contextual conditions, selecting appropriate knowledge-sharing mechanisms, and monitoring how learning ecosystems develop over time. Our findings provide actionable guidance for designing and sustaining interorganizational learning systems capable of supporting digital transformation at scale.

(J Med Internet Res 2026;28:e83345) doi:10.2196/83345

KEYWORDS

digital transformation; health care sector; knowledge sharing; multilevel analysis; qualitative study

Introduction

Internationally, governments are enhancing and transforming their mode of operation, including processes, policies, and services, by adopting digital technologies, with the intention of improving service delivery; becoming more efficient and effective; and increasing transparency, interoperability, and public satisfaction [1,2]. The health care sector, in particular, is making substantial efforts to achieve digital transformation

XSL•FO
RenderX

[3-7]. By digital transformation, we refer to extensive and complex initiatives, such as the implementation of hospital-wide or multisite systems, which bring about significant changes in organizational operations and workflows [7]. This involves advancing technological infrastructure and applications, as well as changing organizational practices, processes, structures, performance, and value creation mechanisms [8-11]. However, despite significant investment in digital health technologies by both government and supplier companies, the transformation of the health care sector lags behind many other sectors owing to the complexity of the health care system, lack of requisite know-how, and immaturity of the market [4,12].

The complexity of health care arises from the diverse and often fragmented pathways that patients must navigate in their journey back to health [13]. Health care systems also involve a multifaceted network of providers, which in some countries, such as the United Kingdom, Canada, Spain, and Singapore, operate under a single organizational structure yet remain functionally distributed and interdependent. Ongoing innovations introduce fundamentally new ways of delivering care, including greater integration across medical specialities, improved coordination between primary and secondary care, and closer alignment between health and social care services [14]. In addition, health care is shifting from predominantly reactive treatment models toward preventive and proactive approaches [15]. At the same time, expectations regarding patient safety and treatment effectiveness continue to rise. Collectively, these factors increase the demands on system performance and make effective coordination increasingly challenging [13].

Digital transformation requires new competencies and skills from those orchestrating the changes and those adopting new technologies for transformation [2]. These can be acquired through learning from other organizations that have already undergone similar digital change processes [10,13,16]. However, in digital transformation, experiences in one setting cannot be mechanistically applied in other contexts [12]. Effective change management, therefore, depends on *contextualized knowledge*, that is, adapting lessons from one organization so they fit the specific circumstances, needs, and constraints of the receiving setting [17].

Building on this, recent research shows that knowledge sharing is most effective when it occurs between organizations that are contextually similar, such as acute hospitals facing comparable pressures, infrastructure, and goals [18]. Such similarity allows organizations to more readily translate and adapt one another's insights. To support this exchange, health care organizations engage in a range of interorganizational arrangements, evolving formal partnerships, exchanging formal codified knowledge (eg, through reports and blueprints), or sharing experience in informal ways (eg, site visits) [18-20]. These modes of engagement help organizations make sense of external lessons in ways that align with their own contexts, enabling more effective learning and application. In the rapidly growing field of digital transformation in the health care sector, policymakers promote and examine different interorganizational knowledge-sharing mechanisms [19,20], and of these, some may fail, while others may become sustainable or transform over time. Currently, however, the efforts in this area often proceed through trial and error, with limited systematic learning about what works and what does not. Emerging evidence suggests that while centralized, top-down approaches can be effective in certain circumstances, for example, by creating shared expectations or establishing a space for learning, they often do not translate into meaningful learning across diverse operational settings. In many cases, they fail to gain traction at the operational level [18].

Despite the recognized benefits of interorganizational knowledge sharing in digital transformation, there remains a notable gap in the literature concerning the specific factors that underpin effective interorganizational knowledge sharing, particularly in the health care sector, and how the factors mobilize a learning ecosystem that enables digitalization of the sector. In particular, while several frameworks offer valuable perspectives on digital transformation, they do not fully account for the *cross-level dynamics* that shape how learning ecosystems emerge, evolve, and function in practice, or *what forms* these ecosystems ultimately take. This theoretical gap has important practical implications: practitioners lack clear guidance on how multilevel contextual factors influence the development, effectiveness, and sustainability of formal and informal interorganizational learning relationships. Therefore, in this paper, we aim to explore and identify critical factors that shape the emergence and evolution of interorganizational learning ecosystems and to develop an integrative model that explains how these multilevel influences give rise to diverse modes of knowledge-sharing relationships during digital transformation.

To address these gaps, our study adopts a grounded, multilevel [21] analytical approach that enables us to trace how contextual conditions at different levels interact over time and across settings to shape processes of knowledge sharing. This approach allows us to identify recurrent patterns across diverse empirical materials and build an empirically grounded framework that explains how different forms of interorganizational learning arise and are sustained. The study therefore advances both the theoretical understanding of cross-level learning processes in digital transformation and the practical insights for cultivating effective interorganizational learning ecosystems in health care.

We draw on a recently completed formative evaluation (2018 to 2023) of a national digitally enabled transformation initiative in English provider organizations called *The Global Digital Exemplar (GDE) Program* to identify the range of factors and modes of interorganizational knowledge sharing as they unfold over time and multiple locations.

This work is timely, as it has the potential to inform policy and practice in relation to the growing investments being made in digital transformation of the health care sector not only in England but also in many other high- and middle-income countries. In doing so, the study provides a series of questions around factors and modes of knowledge sharing that contribute to enabling digital transformation, to be addressed by practitioners and policymakers.

XSL•FO
RenderX

Methods

Research Approach

In order to identify the factors at multiple levels of analysis contributing to interorganizational knowledge sharing, we adopted a longitudinal qualitative approach in the interpretive tradition [22], drawing on multiple case studies and conducting interviews with experts in the field within and beyond organizational settings.

Research Setting

We studied a national digital transformation initiative in the English health care sector from 2017 to 2023, called the GDE Program. The GDE Program was a national investment in the transformation of a variety of care settings, including acute care, specialist care, mental health, and ambulance services of the National Health Service (NHS) in England, with a focus on digitally advanced sites to become international exemplars of excellence in order to provide learning for later organizations planning to go through digital transformation. In this program, provider organizations, henceforth referred to as GDEs, were paired with somewhat less mature fast follower (FF) provider organizations to promote learning across settings. GDEs and FFs captured the best practice models and lessons, including technologies, processes, and structures, in "blueprints," which were aimed to be disseminated within and beyond the program to accelerate the spread of transformation knowledge. Shared learning in the program was also planned to be facilitated by learning networks launched in the first half of the program, bringing together GDE and FF team members with an interest in specific topics.

Data Collection

Overview

We collected data using semistructured interviews, observations, and textual sources. These methods were used concurrently throughout the study period. Data collection was initiated through interviews, which were subsequently complemented by related observational and documentary data to deepen contextual understanding. In cases involving user groups, direct observations were conducted first and were followed by interviews to further explore emerging themes.

Sampling

We conducted in-depth case studies on a set of 12 provider organizations, carefully selected to ensure representation of both GDE and FF acute provider organizations, as well as mental health providers, reflecting the broader sample. The selection process aimed to include maximum variation, considering factors such as a range of different core electronic health record systems, geographical locations, and organization sizes.

Our participants included system vendors, national stakeholders such as program managers, policymakers, and digital health leaders who were responsible for setting digital strategy or who held formal roles such as Chief Clinical Information Officer, Chief Technology Officer, or equivalent positions. To ensure a diverse range of perspectives across the national digital health landscape, we purposively sampled interviewees operating within the NHS national program environment. Recruitment followed a systematic, multistage procedure. First, we identified potential participants through national program-level gatekeepers, including the senior responsible owner overseeing the implementation of the wider program, who provided access to individuals directly engaged in strategic and operational decision-making. Second, we used our existing professional networks to identify individuals holding formal digital leadership roles within NHS organizations who were actively involved in digital transformation efforts. Finally, we used a snowballing approach by asking each participant to recommend others, particularly those who might hold alternative perspectives, to further broaden the diversity of views included in the study.

Semistructured Interviews

We conducted interviews with 228 recognized experts from NHS England professional domains, policymakers, and suppliers, who were involved in the design, launch, monitoring, evaluation, or running of the GDE Program at some point in time. Rather than being formally pilot tested, the guides were iteratively tailored to the expertise and role of each interviewee to ensure relevance and depth of discussion. The questions aimed to explore the evolving nature of knowledge-sharing modes, relationship types and their features, and factors influencing interorganizational knowledge exchange. Data collection was conducted in both face-to-face and online settings. Interviews typically ranged from 30 to 90 min in duration, with an average length of approximately 1 hour. Data were audio-recorded where permission was granted. For nonrecorded interviews, extensive notes were taken. Participants were offered the opportunity to review the final study report, and several provided feedback and comments at that stage. Data collection continued until thematic saturation was reached, at which point no new concepts or insights emerged from the data.

Observations

To complement data collected through interviews, observations of meetings related to GDE Program activities and user group events focused on particular digital technologies being implemented within participating organizations. These observations enabled us to understand how different stakeholder groups engaged with the technologies, addressed challenges, and coordinated implementation work. In addition, we observed 3 national conferences centered on the digitalization of the NHS, where program leaders and adopter sites shared their experiences, lessons learned, and implementation strategies. During the observations, we captured the setting layout, the way participants moved and interacted, the actors, insights into processes and discussion content, and insights into interaction outcomes.

Textual Data

Additionally, to complement the observations and interviews, we collected relevant documents, such as business plans, blueprints, lessons learned documents, and implementation documents, to gain insights into the type of knowledge being shared and relevant processes. To understand broader system-level discourse, we supplemented this with data from online forums, such as NHS Futures and the Digital Health

XSL•FO
RenderX

Network, focusing on what knowledge was shared, how it was communicated, and the forms of support or collaboration that emerged across the wider digital health community.

Data Analysis

Our data analysis integrated well-established approaches, including longitudinal case analysis, grounded theory development, and content analysis. The process was iterative, requiring continuous movement between the data and the evolving theoretical framework [23]. However, for clarity, we present our analysis as a structured sequence of 5 main stages. In stage 1, the lead researcher, HM, utilized NVivo 14 (Lumivero) to organize and integrate all data sources and conducted initial open coding to capture the breadth of stakeholder views and experiences. This stage generated a large pool of provisional codes across interviews, observations, and documents.

In stage 2, we adopted a focused coding strategy to move from initial descriptive codes to analytically meaningful first-order codes. Guided by the principles of parsimony, conceptual salience, and recurrence across data types, we treated sentences and paragraphs as coding units [24] and read the full dataset iteratively, including catalogs and supporting materials. We assigned labels using in vivo codes where appropriate or concise descriptive terms and then grouped related segments into coherent conceptual categories [25]. Importantly, our intention at this stage was not to catalog every detail within such a large dataset but to distil patterns that were recurrent, theoretically relevant, and directly related to our research questions. This analytic focus explains why the final set of first-order codes is relatively concise (n=33). To ensure robustness, coding decisions were discussed within the research team at multiple points, allowing for refinement of categories, challenge of interpretations, and agreement on the conceptual boundaries of each code. Throughout this process, we examined perspectives at multiple analytic levels, including individual, adopter organization, government, and technology supplier, as well as across different stages of digital transformation (procurement, adoption, and use). This enabled us to generate a focused yet comprehensive set of first-order codes that meaningfully represented mechanisms shaping knowledge sharing across organizations.

In stage 3, we identified links among first-order codes to form the emerging second-order themes [24]. This involved discussing disconfirming evidence within the extended research team to minimize the risk of researcher bias.

Stage 4 involved linking second-order themes to form aggregate dimensions [24,26], associated with our research question. These 3 stages (2, 3, and 4) formed our data structure, as presented in Table 1. The Results section begins with explaining this data structure in detail.

XSL•FO
RenderX

Table . Categorization of key factors contributing to knowledge sharing.

Aggregate and second-order themes	First-order codes
Macroenvironmental factors	
Policy developments	• Formation of GDE[a]/FF[b] pairings as a prerequisite of GDE funding • Blueprinting activities mandated by the GDE Program • Digital Academy supporting skill development/cohort/cadre of digital transformation leaders • Wachter Review leading to the development and advancement of new roles such as CCIO[c]/CNIO[d] • Formation of the Digital Health Network as an NHS[e]-organized community
Technology supplier strategies	• Co-ordinating technology-specific user groups • Planning pilot site visits • Suppliers diversifying the technology market
Intermediary actions	• Independent technology consultants moving across settings • Formation of communities of interest by adopter organizations
Interorganizational factors	
Relational recognition	• Development of collective identity (eg, GDE status or Digital Aspirant status) • Naming organizations as GDEs (indicating their digital maturity) as opposed to FFs (indicating them as followers)
Structural governance mechanisms	• Time- and resource-bound contracts versus ad hoc interactions • Need to exchange knowledge in a structured repository (eg, blueprints) versus unstructured sharing (eg, email or site visit)
Interorganizational proximity	• Organizations with closer geographical proximity forming formal GDE/FF partnerships • Personnel from geographically closer organizations communicating and moving between sites • Cost sharing or synergy seeking
Organizational factors	
External search strategies	• Identifying a solution to a certain problem • Identifying who has implemented particular technologies • Setting future digital transformation strategies
Absorptive capacity	• Internal knowledge acquisition strategies • Time allocation to participate in community or networking events
Past experience	• Managing knowledge requests from external organizations based on past experience
Individual factors	
Intrinsic motivations	• Reputational and professional status
Extrinsic motivations	• Creation of communities of shared interests
Inhibitors	• Personal inhibitors
Spectrum of interorganizational relationships and partnerships	
Origins of partnerships/relationships	• Policy (eg, formal agreements with GDE/FF sites) • Independent groups of enthusiasts (eg, Digital Health Network) • Provider organizations (eg, periodic or ad-hoc meetings, trainings, and events) • Suppliers (eg, technology user groups and pilot site visits)

XSL•FO
RenderX

Aggregate and second-order themes	First-order codes
Governing structures and durations	• Formal contractual agreements • Flexible periodic engagement • Informal need/opportunity basis

[a]GDE: Global Digital Exemplar.

[b]FF: fast follower.

[c]CCIO: Chief Clinical Information Officer.

[d]CNIO: Chief Nursing Information Officer.

[e]NHS: National Health Service.

In stage 5, after identifying the aggregate dimensions in stage 4, we noted strong parallels between our grounded analysis and context-mechanism-outcome (CMO) [27] configurations described in realist literature, that is, the taxonomy of multilevel factors incorporated context and mechanisms, and shaped the spectrum of interorganizational relationships and partnerships (outcomes). Building on the similarities, we propose a framework that shows the relationship between the aggregate dimensions. We structured the reporting of this study in accordance with the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines to enhance rigor, transparency, and completeness [28].

Ethical Considerations

We obtained institutional review board approval from the University of Edinburgh Research Ethics Committee (27.11.2017). Informed consent was obtained from all participants prior to their involvement in the study, and all data collected were anonymized to ensure confidentiality and protect participant and organizational privacy. Participants were not compensated for their time.

Results

Overview

Overall, our dataset consisted of 228 interviews; observations of 113 national strategic meetings, user groups, and conferences; analysis of 277 documents; and desk research into online platforms of knowledge-sharing settings such as forums and webinars. Table 2 presents the interviewee characteristics. We present the findings under 2 key headings: multilevel factors contributing to the sharing process and the spectrum of interorganizational relationships and partnerships. Subsequently, informed by the CMO perspective, we present an integrative view of how the factors across 4 levels—macroenvironmental, interorganizational, organizational, and individual—influence knowledge-sharing modes. The factors identified in each category are not exhaustive; however, they highlight some of the most recurrent factors identified in our data within each category.

XSL•FO
RenderX

Table . Interviewee characteristics.

Characteristic	Value, n (%)
In-depth case study sites (12 sites; n=178)	
CCIO[a]	34 (15)
Program manager	29 (13)
CIO[b]	27 (12)
Nursing roles	15 (6)
Digital director/lead	12 (5)
CNIO[c]	11 (5)
Project manager	11 (5)
Pharmacist	7 (3)
Other IT-related roles	20 (9)
Clinical transformation lead	6 (3)
Benefit realization lead	6 (3)
Wider data collection (n=50)	
Policy	44 (19)
Industry	3 (1)
Others	3 (1)

[a]CCIO: Chief Clinical Information Officer.

[b]CIO: Chief Information Officer.

[c]CNIO: Chief Nursing Information Officer.

Taxonomy of Multilevel Factors Contributing to the Interorganizational Knowledge-Sharing Process

Macroenvironmental Level

Our study identified 3 key groups of macroenvironmental factors that influenced knowledge sharing across the settings.

First, *policy developments*, such as government funding and training programs, were seen to have a profound impact in terms of knowledge sharing across different sites. For example, the development of the NHS Digital Academy, which was a close collaboration between several universities and the NHS, supported the skills development of health professionals and decision makers to become digital transformation leaders. In parallel with this, new Digital Health Network communities formed, and these together brought over 5500 NHS digital health leaders to share knowledge and collaborate through online communities, regional and annual events, and webinars. Another example is the Wachter Review conducted between 2015 and 2016, led by an external digital health expert (from the United States), which reviewed the digitalization of the NHS and led to several recommendations that were put into practice, such as the development and advancement of new roles (eg, Chief Clinical Information Officer and Chief Nursing Information Officer) as mediator roles between digital transformation decision makers, front line health care staff, and middle management. In these 3 examples, the macroenvironment was an enabler of transformation and a promoter of knowledge sharing across organizations.

> *And then out of that, so I go to digitalhealth.net events so, like, the summer schools and the leaderships, something that they do, I go to that…I'm part of the One HealthTech, I don't know if you've heard of that, so it's like a volunteer organization for women in health tech. So I represent [site_A]. So there are some other forums where I've access to other people and you tend to meet…there's lots of networking at those events…. So I get lots of emails from people, like, the CNIOs and people from other trusts [hospitals] just asking questions and I have their contact so I can email them. So there's quite a lot of informal, once you've met somebody, that, kind of, networking, you just…you know that they're doing a certain thing if you want to know about that, you can contact them.* [Participant from Site_I]

Despite their enabling roles, the policy development factors may sometimes hinder effective collaboration across organizational settings. For instance, some sites involved in GDE/FF partnerships highlighted that their pairings had been established on the basis of policy criteria, often under compressed timeframes, sometimes resulting in a lack of effective engagement between the partners.

Second, *technology supplier strategies* affected the types of relationships being formed and the collaborations between different organizations. For instance, in the GDE Program, supplier-coordinated networking activities, such as user groups, pilot site visits, and connections of key individuals across organizations, were common. However, many of these

XSL•FO
RenderX

supplier-organized initiatives, while being beneficial in initiating conversations, in particular, with international partners, only allowed for supplier-set goals.

> *When you go on a site visit, you're chaperoned around, no one lets you talk to anyone. Even when you go on the PACS visits, we were definitely stuck in a room and not allowed out.* [Participant from Site_AG7]

In the majority of cases, interviewees highlighted that supplier-arranged learning opportunities drive adopter organizations in particular directions, such as buying particular digital technologies, rather than allowing for open conversations.

> *…the problem with that is that [user group] if you take supplier_name, for example, it's always going to be about supplier_name, well, you should do it this way with this supplier_name product, it's not always the right choice.* [Participant from Site_C]

However, over time, the nature of supplier-organized groups has evolved, becoming increasingly user-driven and confident. For example, users became more active in certain user groups regarding electronic prescribing systems, collaborating to hold suppliers more accountable. Similarly, in a different user group setting, the supplier encouraged their adopters to commit to share learning, reinforcing the shift toward more reciprocal and empowered user-supplier relationships.

Furthermore, other supplier strategies, such as diversifying the market by, for instance, offering best-of-breed solutions on one end of the spectrum as opposed to mega-suites at the other end of the spectrum, had an indirect effect on how and which adopter organizations communicated.

Finally, *intermediaries* had a noticeable influence in orchestrating the sharing of knowledge, particularly where formal policy-developed structures or supplier strategies were deemed to be limited in effectiveness. There are different types of intermediaries, ranging from individuals, such as consultants and technical experts, who facilitate sharing by providing specialized professional services to organizations, to public and private entities that provide technology brokerage, such as suppliers and individual consultants.

> *…Lessons Learned paper that I had, I was given that on day one, when I started, and I went through that myself, and since then, there really hasn't been much engagement at all. Now we have had a specialist that worked for Site_Name on their EDRMS solution…We've put 15 days into that, so obviously quite a lot, it's £750 a day…He brought a lot of experience to the table from what Site_Name did right, and what they did wrong. So, effectively, he was our pseudo blueprint.* [Participant from Site_M]

User-driven communities organized around a particular interest, such as a specific type of technology, are also effective intermediaries, which facilitate learning and cross-organizational collaborations.

> *…we decided, we need a user group, and we need to engage with other users, so that we've got a collective voice for the suppliers. So, we established the national user_group_name, and our CIO was the chair of that…And we also established something called the user_group2_name, which was for the more techies, about how to manage upgrades, how to do certain things in the system, so it's like a knowledge sharing, question and answer thing. And we asked every trust [hospital] to chip in a nominal amount of fund, we're talking about 200 quid…* [Participant from Site_E]

In general, macroenvironmental factors impacted interorganizational knowledge sharing both directly and indirectly. In direct terms, they mandated (eg, blueprinting activities in the GDE Program) and structured relationships (eg, GDE and FF relationships formed between digitally advanced organizations and less digitally advanced organizations to share experiences). In indirect terms, they influenced the ability and motivation of organizations to form relationships for knowledge-sharing purposes (eg, user communities).

Interorganizational Level

Our analysis highlighted 3 types of interorganizational factors influencing knowledge sharing across the settings.

First, we identified that *relational factors*, including the formation of collective identity, power relationships, and trust between organizations, directly affected knowledge-sharing practices. The development of collective identity expanded communication across organizations of similar titles. For instance, we observed that different types of interorganizational collaboration events and meetings were set up (either by adopter organizations or by the NHS), with reference to the status of the organization being a GDE hospital. In such events (eg, the GDE network 2019 event in London or a meeting between Site_D and Site_F), the collective identity was used as an anchor for communication and collaboration.

In terms of power relationships, for instance, our analysis showed how the labeling of organizations as being digitally advanced (GDEs) or being follower organizations (FFs) impacted the dynamics of sharing. While in some instances, the relationship created a means for GDEs to share their advanced competencies with their FFs, in other instances, it hindered sharing, as some FFs did not see themselves as lagging behind in competence and capability. Some FFs, therefore, did not aspire to learn from their GDE peers. This formal power relationship, which differentiated hospitals in their digital status, could also, in some instances, get in the way of sharing knowledge across 2 digitally advanced sites.

> *It's a very strange place the…well, you'd have to say that some trusts [hospitals] that aren't GDEs would feel that they're much more digitally mature than the GDEs. So, whilst the GDE programme singled out a number of trusts [hospitals] for these programmes, you almost then say to those trusts [hospitals], when everybody's looking in, they're saying, we're better than them, or actually, we can do a better job than them. So, you don't tend to get that learning going across, I suppose, I don't know, how much learning do you get between [Site_A] and (Site_B]. Or, if*

XSL•FO
RenderX

> *someone develops some software in [Site_A], would [Site_B] take it? You might find that [Site_B] said, well do you know what, we've got something better, or actually if we did x, y and z, we'd…so it's that competition, in what is supposed to be a national health service, there is that reluctance to adopt other people's ideas and technology.* [Participant from Site_B]

Second, our study pointed to the importance of *structural relationship governance mechanisms* between organizations in enabling or hindering the knowledge-sharing process. For instance, the formal partnerships were planned based on clear strategic organizational needs by strategic decision makers. They prescribed how knowledge exchange was to take place and how the communications were organized. This afforded very different knowledge-sharing practices to informal relationships, which were formed on an ad-hoc basis and were based on particular local needs, the experiences of each organization, aspects that were beneficial, and organizational availability. Additionally, the structure of the relationship, including how it was governed and what routines and processes were involved in knowledge interactions, had a direct effect on different aspects of knowledge sharing. In terms of governance, for instance, whether the relationship is an informal relationship that is established and runs based on the availability of different organizations, or whether it is based on some type of contractual association with a defined scope and defined roles, affects the type of knowledge being exchanged. This includes, for example, structured documents with codified knowledge on how to implement a functionality versus discussions on solving a particular problem using system configuration and workarounds during a networking event, the direction of knowledge flow (eg, from a GDE to an FF or multidirectional across multiple organizations), the frequency of interactions (eg, a specific number of site visits in a given period of time vs ad-hoc site visits based on requirements), and the extent and longevity of sharing.

Finally, *interorganizational proximity* influenced knowledge sharing across organizations. Proximity has different aspects, such as physical proximity or the degree of technological or strategic similarity. For instance, organizations with a closer geographical proximity either formed formal GDE/FF partnerships or had personnel communicate and move between sites readily in informal ways to exchange knowledge. Another example was the procurement and implementation of similar technologies by different hospitals, which led to the development of relationships/partnerships, with motivations such as cost-sharing or synergy-seeking.

> *Then I think because at the same time [Site_C] were going through and had decided to also take [technology_Name] as its electronic patient record system, they were our formal fast follower, and so it enabled us to work with them. But in fact we had always agreed that we would provide wider support to other hospitals also considering or implementing [technology_name]. So as well as [Site_C], we were providing support to [Site_D], [Site_E] and [Site_F].* [Participant from Site_A]

Organizational Level

We observed that 3 organizational factors had a direct influence on the sharing of knowledge.

First, an organization's *agenda for searching for external knowledge* was an important factor. Some organizations only looked for external knowledge when a particular need arose, such as the need for the implementation of a new piece of technology, while others broadened external knowledge search mechanisms to identify unknown or future strategic opportunities.

> *I think the other thing we sort of did over the summer, was have a look at all of the GDE blueprints. I know that we're a GDE, so I suppose the assumption is that, because we're a GDE, we've done everything, but I'm a great believe in learning from somebody else. So, we kind of did a trawl of all of the blueprints that have been published to date, to see whether or not there was anything we're not taking use of as much as somebody else is, and really starting to think about, what's our next cycle of digital activity? And you can kind of see there, that pretty much the first two columns says, oh, we're probably doing about 85 per cent of everything that's out there, but the final column says, there are a few options. And I think, in particular for us, it's interesting that we're doing webinars for our staff, but not for our patients.* [Participant from Site_L]

Second, an organization's *absorptive capacity* played an important role in enabling sharing across organizations. Absorptive capacity refers to an organization's ability to acquire, assimilate, and utilize external knowledge, information, and innovations. For instance, some organizations had implemented internal processes and routines to facilitate learning and sharing, enabling them to acquire insights from other organizations and manage external knowledge-sharing demands. Conversely, in cases where less time was dedicated to absorbing and utilizing external knowledge, such as during the procurement of new digital technologies, these organizations gained limited benefits from the knowledge held by others, which sometimes led to experiencing challenges previously faced by other organizations. Some interviewees highlighted that their organization had established a routine for them to allocate time for periodic conversations with other organizations, enabling structured opportunities for knowledge exchange. By introducing these routines, organizations can foster a culture of continuous learning, gain insights from the experiences of peers, and adopt the best practices in digital transformation. This approach not only accelerates their own digital initiatives but also helps them navigate challenges more effectively, ultimately strengthening their overall organizational capabilities.

> *…pretty much every month the CIOs or heads of IT delivery […] have opportunity to meet each other …to actually talk to each other. And that makes the world of difference…the other week, [name] went along and came back with a small list of this is what I've got from [org_name_A], this is what I've got from [org_name_B], could you have a chat with this*

XSL•FO
RenderX

> *person please…I just ring them up and say, yeah, you wanted a chat, this is what it's about and vice versa…* [Participant from Site_G]

Finally, an organization's *past experiences* in terms of identifying, building, and maintaining relationships had a noticeable impact on the formation of its knowledge-sharing practices. For instance, organizations that were more experienced in managing and coordinating relationships established clear processes for managing the requests from external organizations asking for their knowledge and experiences.

> *They were asked not to make direct contact with [site_B] staff, they had to go through a specific office and up to monitor, well not monitor but manage and control the contacts. Which worked fine actually because they were as I said, they replied very quickly when we begin a request, which was usually by email and we got through that pretty quickly. They wanted to avoid loads of people making contact with loads of [site_B] staff and disrupting them unnecessarily.* [Participant from Site_C]

Individual/Interpersonal Level

We found 3 individual factors impacting formal and informal knowledge sharing across organizations.

First, we observed that individuals are driven to share knowledge by *intrinsic motivations*, specifically their self-interest. Our observations revealed that participants exhibit a heightened willingness to share knowledge when they perceive tangible benefits arising from such collaborative endeavors. Among these motivating factors, the prospect of bolstering one's reputation and professional status emerged as a significant catalyst for active involvement in knowledge-sharing communities.

Furthermore, the desire to gain profound insights into alternative perspectives and diverse experiences emerged as another compelling factor propelling knowledge sharing. Participants acknowledged the richness that emerges from exposure to a variety of viewpoints, and they mentioned actively seeking to broaden their understanding by tapping into the collective wisdom of others.

> *so that's been really helpful, to have that outside perspective from someone else in a different Trust [hospital] doing something that's similar but having a different viewpoint. As I find the people are fantastic, so much knowledge, so much experience, but I appreciate we're all at the same Trust [hospital] so we've got almost like the same vision, and sometimes it's just nice to see someone else's perspective who's not part of this Trust [hospital].* [Participant from Site_E]

Second, beyond intrinsic motivations, *extrinsic factors* rooted in altruism played a pivotal role in fostering the sharing of knowledge across organizational boundaries. The altruistic dimension of knowledge sharing is particularly evident in instances where individuals perceive that their contributions go beyond personal gain and actively contribute to the construction or sustenance of communities of shared interests. Participants recognized that by sharing their expertise, insights, and experiences, they are instrumental in nurturing a collaborative environment where collective learning and problem-solving thrive. This communal aspect became a source of motivation, as actors derived satisfaction from the sense of building and sustaining a network of like-minded individuals. Furthermore, the fostering of feelings of companionship and camaraderie emerged as a powerful intrinsic motivator for knowledge sharing. Participants found fulfillment in being part of a community where the exchange of knowledge is not solely transactional but is imbued with a sense of shared purpose and mutual support.

> *we've now set up a GDE group with everyone that looks after benefits for their Trust [hospital], and we're using it as a sharing platform to help each other…* [Participant from Site_E]

While our study did not aim to identify barriers, in the case of individual factors, we encountered individual inhibiting factors that introduced complexities for the sharing of knowledge that necessitated careful consideration. Three notable aspects emerged as noteworthy impediments: the allocation of time to engage in sharing activities, the concept of knowledge-based psychological ownership, and the nuanced emotional responses individuals exhibited when navigating the uncertainties inherent in interorganizational settings. The time commitment required for participation in knowledge-sharing activities proved to be a substantial factor. Individuals, often operating within demanding professional environments, found themselves grappling with the challenge of allocating sufficient time to contribute meaningfully to knowledge-sharing initiatives.

> *The is a CSO network but unfortunately, we don't meet terribly often. And actually, most of us are just too busy…but in the end the problem is it doesn't get enough time, I think. I am still a full time clinician, so I still have to do a very much full time on-call job.* [Participant from Site_G]

Some individuals were possessive of their knowledge, struggling with what to share openly versus what to protect for their organization's benefit. This sense of ownership created barriers to the free flow of insights. Additionally, concerns about partners' behaviors affected the willingness to share, introducing a complex balance between trust and caution. Apprehensions about defining knowledge boundaries (what to disclose and what to withhold) further highlighted the nuanced nature of interorganizational knowledge sharing. These individual perspectives and emotional responses shaped knowledge transfer dynamics, underscoring the need for a deeper understanding to foster effective collaboration.

Linkages Across Levels in Activating Knowledge-Sharing Modes

The factors identified in our study do not operate in isolation; rather, elements across macroenvironmental, interorganizational, organizational, and individual levels are highly interconnected and often mutually reinforcing in shaping knowledge-sharing relationships. For example, national policy developments, such as the Wachter Review, not only stimulate the creation of new

XSL•FO
RenderX

digital leadership roles (eg, Chief Clinical Information Officer and Chief Nursing Information Officer) but also contribute to the emergence of a digitally skilled cadre through initiatives like the NHS Digital Academy, thereby improving the absorptive capacity of organizations and their ability to engage meaningfully in communities of practice. Similarly, program-level structures, such as mandated GDE/FF pairings and blueprinting requirements, interact with relational dynamics, including the formation of collective identities (eg, GDE status) and the development of trust-based recognition between partner sites, which in turn facilitate voluntary knowledge exchange beyond formal contractual obligations. Technology supplier strategies also intersect with organizational search behaviors: suppliers' diversification of the market can drive adopter organizations to seek external knowledge (eg, identifying who has implemented specific technologies), often aided by intermediary actors or technology-specific user groups that bridge otherwise disconnected sites. Geographical proximity further amplifies these interactions, as staff from nearby organizations are more likely to visit one another, participate in shared communities of interest, or move across settings, thereby strengthening informal knowledge flows that complement formal governance mechanisms. Finally, individual-level motivations, such as professional reputation, intrinsic interest, and past experience in handling knowledge requests, shape how actors engage with these structures, while personal inhibitors can constrain knowledge sharing even when enabling conditions are otherwise present. Together, these interconnected elements illustrate how multilevel factors combine to activate and shape the modes underpinning knowledge sharing across the health system.

Spectrum of Interorganizational Relationships and Partnerships

Together, the above interconnected conditions contributed to a spectrum of interorganizational relationships and partnerships—what we refer to as modes of knowledge sharing—that emerged from multiple origins and were shaped by varying degrees of structure and formality. These mechanisms ranged from highly orchestrated arrangements to more flexible, organically developed connections. Across this spectrum, organizations used diverse tools and practices for knowledge sharing as described below.

The *origins* of relationships ranged from internal (individual or communal within and across transforming organizations) to external (beyond the transforming organizations). In the case of internal origins, sometimes provider organizations collaborated on a local needs basis (eg, site visits or phone calls between individuals) or more strategically entered into intentional cultivation of networks and communities in response to collective needs or to achieve collective strategic benefits. For instance, when multiple organizations needed particular functionalities from the same vendor, organic and spontaneous collaborations between 2 or more user organizations helped to create a collective force to exert leverage over technology vendors or to collaborate and form a technological solution that benefited multiple organizations.

> *And then on going forward it was about where we had joint requirements for development such as around various different areas such as out-patients or theatres. We'd try and come up with a joint specification so that [supplier_name] only had to develop once.* [Participant from Site_D]

In the case of external entities initiating relationships, in some instances, interorganizational relationships were initiated by formal mandates from the government. For instance, in our study, 19 strategic pairings were formed between GDEs and FFs. These partnerships were based on a range of factors such as common core technologies and local strategic groupings, for instance, between organizations that had a particular electronic patient record system or between nationally coordinated collaborations of health care organizations and local authorities.

Technology suppliers were also initiators of particular types of relationships around specific technologies, both nationally and internationally.

> *...there's now quite good, not only UK, but a wider European group forum. Plus, we have regular conversations with Melbourne. Pre-[technology_name], our pediatricians used to go back and forth with Melbourne. Melbourne physicians used to come here, our guys used to go there, so there's a very strong relationship with Melbourne. They have [technology_name]. It's a bit like an onion, you know, layers and layers of different communities that have formed very nicely.* [Participant from Site_A]

The wide spectrum of interorganizational relationships had different *governing structures and durations*. For instance, government-initiated partnerships could range from national conferences or conference series (short-term, flexible periodic engagements) to fixed-term/long-term contracts between different provider organizations (eg, GDE/FF partnerships). These contracts mandated a number of aspects, such as the type of sharing (documentation), the interval for sharing, and what is to be shared. Despite investments into these formal means of knowledge sharing, many of the initiatives did not reach their anticipated longer-term benefits, despite being easy to measure, and their purposes changed over time. One example is blueprints, which were mandated by the GDE Program and were designed as tools for sharing the knowledge needed to implement specific models of change to be adopted by the wider ecosystem. While they offered codified knowledge about new process designs, they did not provide the means to implement certain technologies in the new context, as the implementation process involved a range of tacit knowledge that was not included in the blueprints. Many provider organizations used blueprints as a tool to identify which organizations and which actors were sources of particular knowledge in order to establish informal contact and to allow for the sharing of knowledge beyond codified documents.

We also noted that not all formal initiatives proved to be self-sustaining over time. For instance, GDE Program managers launched a number of learning networks, many of which were terminated after the end of national funding. Only some

XSL•FO
RenderX

networks driven by professional agendas and interests (eg, the Hospital Electronic Prescribing Medicines Administration Network) sustained and transformed into national communities of practice after the completion of the initial funding period. Collaboration and sharing of knowledge in these communities were not based on policy-mandated contracts, and instead, knowledge sharing was orchestrated through collaboration of provider organizations as they captured local and collective benefits from contributing to the community.

Integrating Influencing Factors Into the Spectrum of Interorganizational Relationships

Building on these findings, we developed an integrative model (Figure 1) that brings together the multilevel factors (context and mechanisms on the left-hand side) identified in our analysis with the resulting spectrum of interorganizational relationships and partnerships (outcomes on the right-hand side). The model illustrates how different contextual conditions and mechanisms combine to produce a diverse array of knowledge-sharing relationships, which vary not only in formality but also in their roots and modes of initiation. To capture this variation, we conceptualized a 2D spectrum: the x-axis represents the degree to which a relationship is formal or informal, while the y-axis indicates whether its origin is internal (arising from organizational strategies, leadership roles, or intrasystem initiatives) or external (emerging from policy directives, supplier strategies, intermediary actions, or professional networks). Importantly, the left and right sides of this spectrum are connected through a range of influences, such as triggering, mandating, controlling, and enabling, which shape, prompt, or support interorganizational knowledge sharing across different settings. By positioning relationship types within this matrix, the model demonstrates how combinations of enabling factors give rise to different patterns of collaboration and partnership across the ecosystem, offering a coherent way to understand the diversity and dynamics of interorganizational knowledge-sharing arrangements.

Figure 1. Integrative model of interorganizational learning ecosystems in health care digital transformation.

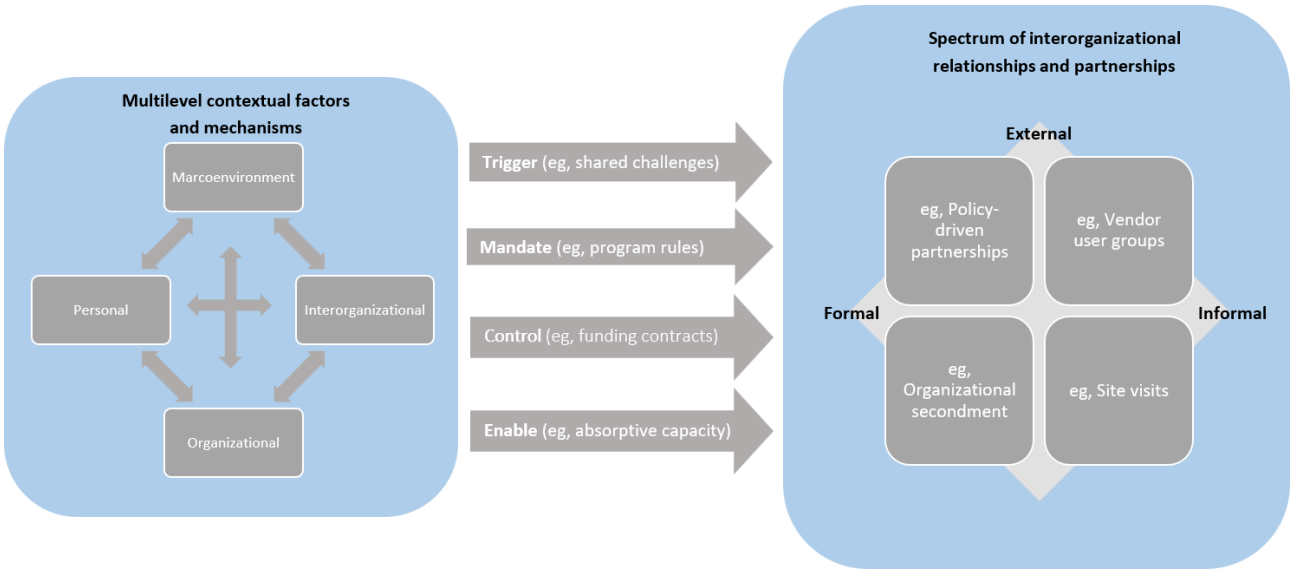

Discussion

Principal Findings

Our findings show that effective interorganizational knowledge sharing is shaped by different configurations of factors linking across the levels of a system. At the macro level, policy stability, national investment, and the presence of formal digital transformation programs created perceived legitimacy, shared purpose, and a sense of mandate among organizations, resulting in higher engagement in knowledge-sharing activities and more formalized collaborations. At the interorganizational level, networks characterized by trust, prior relationships, and aligned implementation timelines enabled modes of reciprocal exchange, openness, and problem-solving, leading to sustained cross-site learning, joint solution development, and more mature partnerships. In contrast, competitive procurement pressures or misaligned incentives triggered protective behaviors, producing fragmented or transactional exchanges. These examples of different configurations illustrate how and why knowledge-sharing modes vary in effectiveness and evolve over time.

From Knowledge Sharing to Learning Ecosystems: A Theoretical Contribution

In this section, we compare our empirical findings with the broader literature on knowledge sharing across organizations and present how our findings contribute to the development of the concept of evolving learning ecosystems.

Combining our findings with the extant literature that addresses *macroenvironmental factors* on interorganizational knowledge sharing, we identified a wide range of macroenvironmental factors, including policy developments, intellectual property regimes [29], technological sophistication levels [30], technology suppliers and their strategies [31], diverse markets [32], and other external agencies (eg, universities, civil society groups, third-party firms that have close connections with the health care system, etc), which can trigger, mandate, control, or enable interorganizational knowledge sharing. Policy-related factors can act as “windows of opportunity” [29] that enable or

XSL•FO
RenderX

mandate the formation of relationships for sharing specific types of knowledge [33]. However, the outcome of using the mechanisms mandated or enabled by policies is dependent on how knowledge-sharing stakeholders exploit these opportunities [18]. For instance, when the government creates formal partnerships and mandates the sharing of documents in a repository accessible to a wide range of organizations, the scale of sharing may be large, but the scope and depth of sharing may be limited. Other external entities, such as suppliers (through their strategies) and intermediaries, also shape the types of relationships being formed and the collaborations that may take place between different organizations [30]. Typically, relationships formed by suppliers are less formal, sometimes ongoing, and sometimes on an ad-hoc basis. A common finding in the GDE Program was supplier-coordinated networking activities such as user groups, pilot site visits, and connections of key individuals across organizations. Intermediaries may offer formal partnerships by providing specialized professional services to organizations and facilitating sharing between them [34,35] or providing technology brokerage activities [36]. Intermediaries may also enable the formation of communities of interest, where the relationship is less formal. In general, the macroenvironmental factors impact interorganizational knowledge sharing both directly and indirectly. In direct terms, they can mandate (eg, blueprinting activities in the GDE Program [18]) and structure relationships (eg, GDE and FF relationships formed between more digitally advanced organizations and less digitally advanced organizations to share experiences [37]). In indirect terms, they can influence the ability and motivation of organizations to form relationships, for example, by identifying who holds what knowledge about a particular type of technology.

Interorganizational factors, including relational factors [38], structural governance mechanisms, and interorganizational proximity [38], have a profound impact on shaping the early stages of relationship/partnership building. Our research shows that by building a collective identity that involves the development of a common objective or a mutual win and involves shared rules and regulations in terms of both formal [39] and informal [40] relationships, knowledge sharing can be encouraged. These can be achieved through building interorganizational teams or communities of interest that invest in modes of knowledge exchange and integration in order to solve a problem or overcome a common challenge [41-43]. Additionally, structural governance mechanisms (eg, whether they are time-bound or not, or whether they are periodic or run on an ad-hoc basis) also affect the type of relationship and the direction of knowledge flow. Furthermore, organizational proximity plays an important role in the formation of both formal and informal sharing. The extant literature highlights various aspects of organizational proximity, including cultural closeness [44-46], similarity in technological artifacts in use [47], alignment in strategic direction [48], and geographical proximity, and these facilitate smoother mobility of personnel, socialization, in-person exchange, and informal ties [36,49].

In terms of *organizational factors*, our study highlights the importance of knowledge search strategies by organizations, absorptive capacity, and past knowledge-sharing experiences in triggering and enabling sharing, as well as a wide spectrum of interorganizational knowledge-sharing modes. The extant literature has focused on the importance of absorptive capacity, which refers to the ability to recognize the value of knowledge being shared by both the donor and receiver organizations and the ability to utilize it [50]. The absorptive capacity is influenced by an organization's past experiences of identifying and building relationships [51], ability to integrate knowledge [52], and capability to retain knowledge [53]. We also found that as organizations gain experience in sharing and managing knowledge demands, they become more capable of managing and coordinating relationships as well as organizing the learning process. Organizations may use tools, such as IT-enabled systems, to eliminate barriers of communication and thereby heighten the level of knowledge sharing.

In relation to *individual and interpersonal factors*, we highlight the role of intrinsic and extrinsic motivations in encouraging knowledge sharing and the personal inhibitors that hinder the sharing process. Research has revealed a range of explanations for why individuals contribute to knowledge sharing in formal and informal settings. Many explanations point to intrinsic (self-interest) and extrinsic (altruistic) motivations [54]. Such motivations include self-benefits received from participation in knowledge-sharing communities and settings, such as access to expert advice [55], insights into others' beliefs and opinions, enhanced reputation and professional status [56,57], networking [58], and improved self-image and confidence [59]. Benefits to others include a desire for building or sustaining communities of interest [59,60], feelings of companionship [57], and reciprocity [61]. These motivations may vary in terms of nature and depth in formal and informal knowledge-sharing relationships. Bateman et al [62] highlighted that an individual's commitment to the organization they belong to, as well as the community, has a significant effect on participating in knowledge sharing. In this way, individuals may believe that they benefit from knowledge-sharing relationships [63] in both formal and informal settings, or they may feel obliged to pay back to existing relationships [56,62]. While the literature highlights the role of motivations in facilitating knowledge sharing, it also identifies factors that may hinder the transfer process. These include the time required for sharing activities [59] and the feeling of a strong psychological ownership over knowledge [63]. Emotional attachment to the organization may also discourage individuals from sharing certain knowledge, as it leads them to prioritize the organization's interests [64], influencing their decisions on what knowledge to protect and how to control its release [65].

Our study contributes to the literature on digital transformation and interorganizational learning by advancing the understanding of how learning ecosystems emerge, evolve, and take shape across multiple levels of the health care system. Rather than treating macroenvironmental, organizational, interorganizational, and individual factors as discrete categories, we demonstrate that they operate as dynamically interconnected influences that jointly shape the evolution of knowledge-sharing relationships. Through our multilevel analysis, we show how contextual conditions, such as national policy directives, program mandates, supplier strategies, and organizational digital maturity, activate

XSL•FO
RenderX

mechanisms, including identity formation, absorptive capacity, motivations, and relational recognition. These mechanisms, in turn, produce a spectrum of relationship types that vary in formality and origin, ranging from informal peer interactions to mandated partnerships.

By integrating these insights, our work conceptualizes the *learning ecosystem* as an evolving sociotechnical system in which relationships and knowledge flows continually adapt in response to shifting contextual pressures and emerging opportunities. We demonstrate that learning ecosystems do not develop through linear or centrally orchestrated processes; instead, they emerge through recursive interactions across levels, where changes in one domain, such as policy initiatives or new digital leadership roles, can reconfigure organizational strategies, reshape interorganizational networks, and influence individual patterns of knowledge sharing. The integrative model we propose brings these dynamics together by illustrating how multilevel factors (contexts and mechanisms) give rise to diverse modes of interorganizational collaboration (outcomes), and how triggers, mandates, controls, and enablers collectively shape the system's trajectory over time. This contribution extends existing frameworks on interorganizational knowledge sharing by offering a holistic explanation of how learning ecosystems are formed, how they function, and why they evolve differently across settings. In doing so, it provides theoretical clarity for understanding and supporting the development of interorganizational learning ecosystems capable of enabling digital transformation at scale.

Implications for Practice

This study offers important implications and insights that can be used in enabling and supporting the digital transformation process in the health care sector. Drawing on our findings, we have developed a structured set of questions designed to help policymakers and practitioners diagnose the conditions shaping knowledge sharing and monitor how these conditions evolve over time. These questions focus on the macroenvironmental, interorganizational, organizational, and individual factors that influence the emergence and functioning of learning ecosystems. We present these questions in Table 3. These questions serve as both a diagnostic aid (helping stakeholders assess what they are trying to achieve and the contextual factors shaping these efforts) and a monitoring aid (helping them evaluate how learning conditions and relationships develop over time).

To address macroenvironmental questions, policymakers and practitioners must examine public policies that facilitate or hinder the formation of formal and informal partnerships, as well as the role of external agencies in enabling knowledge sharing. Additionally, it is important to identify external factors, such as competition and cooperation. This requires assessing whether policymakers or other external entities demand the sharing of specific types of knowledge and whether industry characteristics, such as the levels of technological complexity and market maturity, affect partnership and community-building. To answer interorganizational questions, policymakers and practitioners should explore structural governance mechanisms such as community-based orchestration, contracts, defined processes, and routines. They must also examine relational and cognitive governance factors, including trust and power dynamics. Exploring organizational questions requires assessing the capabilities and experiences of donor and recipient organizations in forming and managing partnerships or participating in communities. This includes their ability to manage alliances, learn from and apply external knowledge, and facilitate knowledge sharing. Additionally, behavioral aspects, such as organizational motivations, incentives for knowledge sharing, and resistance to new knowledge, must be considered, along with organizational characteristics like the credibility of the donor and the receiver's flexibility to adapt to change. Finally, addressing individual-level questions requires examining both intrinsic and extrinsic factors influencing knowledge-sharing behavior.

XSL•FO
RenderX

Table . Diagnostic and monitoring questions for assessing multilevel factors influencing knowledge sharing.

Factor level	Questions
Macrocontext factors	• What formal external factors (eg, regulations, policies, technology, and market) affect the motivation to form formal partnerships for knowledge sharing? • What external factors affect the motivation to share knowledge without the existence of formal partnerships? • Is there an agency external to knowledge-sharing organizations that structures, incentivizes, and monitors knowledge sharing? What is the role of strategic mandates?
Interorganizational factors	• How do the interactive dynamics of the relationship (ie, power relations, structural governance and mechanisms, social ties and trust, and risk) influence knowledge sharing? • How does proximity (geographical, cultural, and technological) affect knowledge sharing? • What are the relationship-building factors (ie, collective identity, mutual goals, and joint rules) that affect knowledge sharing?
Organizational factors	• How do organizational characteristics affect knowledge sharing? • How do routines, processes, and coordination mechanisms affect knowledge sharing? • What is necessary for an organization to share/receive external knowledge?
Individual factors	• What motivates individuals to share knowledge? • What are the individual factors that prevent the sharing of knowledge?

Building on this diagnostic foundation, we offer a second set of questions that guide stakeholders in identifying and selecting appropriate modes and mechanisms of knowledge sharing for digital transformation initiatives. Rather than prescribing a fixed sequence of steps, these questions encourage reflection on the purpose of different knowledge-sharing modes, their temporality, the coordination techniques they require, and the tools that support them. These questions are presented in Table 4. Stakeholders can populate them with context-specific examples and options relevant to their initiatives. Table 4 includes 3 illustrative columns to demonstrate how different knowledge-sharing modes may be examined and compared.

Table . Reflective questions for selecting and assessing knowledge-sharing mechanisms in digital transformation initiatives.

Question	Option 1	Option 2	Option 3
What are the different options for formal/informal interorganizational knowledge sharing?	Blueprinting	Online user communities	Site visits
What purposes does it serve?	To share how a particular technology has been configured and implemented	To identify solutions to specific problems/needs; To co-innovate	To observe real-world technology use and challenges; To validate assumptions
What is the duration?	Project based	Continuous; Need basis	One-off
Who are the initiators of the modes of sharing?	Government	Individuals	Organizations; Technology suppliers
What are the management/coordination techniques?	Controlled by funding contacts	Orchestrated by community members	Managed by organizations on a need basis
What infrastructure and tools are used to facilitate knowledge sharing?	Documentation tool and database	Forums	In-person visits

Consistent with existing literature emphasizing the importance of relational maturity and temporality in knowledge-sharing processes [45,52,66,67], we propose that the questions in both tables should be revisited and iteratively assessed over time. Together, Tables 3 and 4 provide a structured yet flexible resource to support the design, implementation, and continual refinement of interorganizational learning within digital transformation programs.

By addressing these 2 sets of questions, this study provides policy and practice tools and guidelines to enhance and accelerate the digital transformation of the health care sector. First, it identifies where and how governments and other intermediaries can support interorganizational knowledge sharing. Second, it highlights the formal and informal modes and internal organizational and individual factors that

XSL•FO
RenderX

organizations can leverage as they navigate the transformation journey.

Limitations and Future Research

While this study advances the understanding of how learning ecosystems emerge and evolve across multiple levels, several limitations create opportunities for future research. First, our analysis focused predominantly on macroenvironmental and interorganizational factors, reflecting the design of the national digital transformation program and the nature of the available data. As a result, less attention was given to micro-level organizational routines and individual behaviors that may shape, or are shaped by, the learning ecosystem. Second, although our multicase design enabled rich comparative analysis across organizations, the scope of the study was limited primarily to hospital settings within a national health care system. Learning ecosystems may function differently across other care settings, sectors, or international contexts with different governance, market, or cultural dynamics. These limitations offer a valuable foundation for theoretical propositions that can guide future research aimed at advancing, refining, and testing the model developed in this study.

To guide such research, we propose a set of theoretically informed propositions. These propositions posit that cross-level contextual alignment is central to effective ecosystem formation; mechanisms, such as identity and relational recognition, significantly influence knowledge flows; the balance of formal and informal relationships affects ecosystem adaptability; and intermediaries are essential brokers in sustaining learning systems. Moreover, we propose that external influences shape early ecosystem trajectories, internal factors become more influential as ecosystems mature, learning outcomes vary with specific factors, and organizational- and individual-level factors warrant a deeper empirical investigation. Together, these propositions establish a clear agenda for advancing the theory on the emergence, evolution, and functioning of interorganizational learning ecosystems in digital transformation contexts.

Conclusion

Digital transformation has presented a significant challenge in the health care sector, which requires knowledge, experience, and expertise for organizational gains and new value creation. Interorganizational knowledge sharing can be seen as an enabler, supporter, and accelerator in assisting digital transformation at scale. Policymakers and organizations must proactively address the internal and external factors that impact knowledge sharing to harness its full potential in driving digital transformation. Understanding and evaluating approaches to creating a learning ecosystem is far from a trivial task and requires multilevel inquiries that follow actions within and beyond the organization in understanding this complex phenomenon.

Our framework illustrates how multilevel influencing factors shape a spectrum of interorganizational relationships and partnerships. The framework explores the dynamic nature of factors over time and therefore goes beyond shedding light on snapshots of the knowledge-sharing process by offering an evolutionary view of the process and its outcomes. By proposing this framework, we do not suggest that the domain of interorganizational knowledge sharing can be simply reduced to the aforementioned dimensions. Moreover, we do not propose this as a simple mechanistic account that could decouple different variables for the testing of modes of sharing and factors. Instead, we present a set of dimensions that are essential to explore in order to theorize a conceptual account of the learning ecosystem. We introduce this not just as an abstract framework but as a practical tool to support reflection and guide action for the various actors involved in shaping and sustaining it.

Acknowledgments

The authors gratefully acknowledge the input of the participants and the steering group for this evaluation. The views expressed are those of the authors and not necessarily those of the National Health Service (NHS), NHS England, NHSX, the Department of Health and Social Care, or NHS Digital. Generative artificial intelligence (ChatGPT by OpenAI) was used for proofreading the paper and streamlining some of the sentences.

Funding

This article has drawn on a program of independent research funded by the NHS (NHS England OJEU contract reference number 2017/S 114-229670; RW). This work was also supported by the National Institute for Health Research Imperial Patient Safety Translational Research Centre. The funders had no role in study design, data collection and analysis, the decision to publish, or the preparation of the manuscript.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

All authors conceived the study. HM drafted the manuscript, and all authors commented on the drafts of the manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Mergel I, Edelmann N, Haug N. Defining digital transformation: results from expert interviews. Gov Inf Q 2019 Oct;36(4):101385. [doi: 10.1016/j.giq.2019.06.002]
2. Carter L, Desouza KC, Dawson GS, Pardo T. Digital transformation of the public sector: designing strategic information systems. J Strateg Inf Syst 2024 Sep;33(3):101853. [doi: 10.1016/j.jsis.2024.101853]
3. Gopal G, Suter-Crazzolara C, Toldo L, Eberhardt W. Digital transformation in healthcare – architectures of present and future information technologies. Clin Chem Lab Med 2019 Feb 25;57(3):328-335. [doi: 10.1515/cclm-2018-0658] [Medline: 30530878]
4. Walsh MN, Rumsfeld JS. Leading the digital transformation of healthcare: the ACC innovation strategy. J Am Coll Cardiol 2017 Nov 28;70(21):2719-2722. [doi: 10.1016/j.jacc.2017.10.020] [Medline: 29169479]
5. Gupta S, Modgil S, Lopes de Sousa Jabbour AB, Laguir I, Stekelorum R. Towards digital transformation and governance in the healthcare sector. Inform Technol People 2025 Nov 3;38(5):2163-2186. [doi: 10.1108/ITP-02-2023-0179]
6. Koukouvinou P, Simbi N, Holmström J. Managing unbounded digital transformation: exploring the role of tensions in a digital transformation initiative in the forestry industry. Inform Technol People 2023 Dec 18;36(8):43-68. [doi: 10.1108/ITP-03-2020-0106]
7. Burton-Jones A, Akhlaghpour S, Ayre S, Barde P, Staib A, Sullivan C. Changing the conversation on evaluating digital transformation in healthcare: insights from an institutional analysis. Information and Organization 2020 Mar;30(1):100255. [doi: 10.1016/j.infoandorg.2019.100255]
8. Agarwal R, Gao G, DesRoches C, Jha AK, Gao G. Research commentary—the digital transformation of healthcare: current status and the road ahead. Inf Syst Res 2010 Dec;21(4):796-809. [doi: 10.1287/isre.1100.0327]
9. Matt C, Hess T, Benlian A. Digital transformation strategies. Bus Inf Syst Eng 2015 Oct;57(5):339-343. [doi: 10.1007/s12599-015-0401-5]
10. Hess T, Matt C, Benlian A, Wiesböck F. Options for formulating a digital transformation strategy. In: Galliers RD, Leidner DE, Simeonova B, editors. Strategic Information Management: Routledge; 2020:151-173. [doi: 10.4324/9780429286797-7]
11. Goraya MAS, Yaqub MZ, Khan MA, Akram MS, Alofaysan H. Transforming performance: how agility, response, resilience and support shape success in digital strategies. Inform Technol People 2026 Jan 14;39(1):325-353. [doi: 10.1108/ITP-05-2024-0592]
12. Pittaway JJ, Montazemi AR. Know-how to lead digital transformation: the case of local governments. Gov Inf Q 2020 Oct;37(4):101474. [doi: 10.1016/j.giq.2020.101474]
13. Greenhalgh T, Papoutsi C. Studying complexity in health services research: desperately seeking an overdue paradigm shift. BMC Med 2018 Jun 20;16(1):95. [doi: 10.1186/s12916-018-1089-4] [Medline: 29921272]
14. Goodwin N, Smith J, Davies A, et al. Integrated care for patients and populations: improving outcomes by working together. : King's Fund; 2012 URL: https://assets.kingsfund.org.uk/f/256914/x/2ff65ea41a/integrated_care_nhs_future_forum_january_2012.pdf [accessed 2026-02-26]
15. Waldman SA, Terzic A. Health care evolves from reactive to proactive. Clin Pharmacol Ther 2019 Jan;105(1):10-13. [doi: 10.1002/cpt.1295] [Medline: 30597564]
16. Gil-Garcia JR, Sayogo DS. Government inter-organizational information sharing initiatives: understanding the main determinants of success. Gov Inf Q 2016 Jul;33(3):572-582. [doi: 10.1016/j.giq.2016.01.006]
17. Cresswell K, Sheikh A, Franklin BD, et al. Interorganizational knowledge sharing to establish digital health learning ecosystems: qualitative evaluation of a national digital health transformation program in England. J Med Internet Res 2021 Aug 19;23(8):e23372. [doi: 10.2196/23372] [Medline: 34420927]
18. Williams R, Sheikh A, Franklin BD, et al. Using blueprints to promote interorganizational knowledge transfer in digital health initiatives-a qualitative exploration of a national change program in English hospitals. J Am Med Inform Assoc 2021 Jul 14;28(7):1431-1439. [doi: 10.1093/jamia/ocab020] [Medline: 33706378]
19. Cresswell K, Sullivan C, Theal J, Mozaffar H, Williams R. Concerted adoption as an emerging strategy for digital transformation of healthcare-lessons from Australia, Canada, and England. J Am Med Inform Assoc 2024 Apr 19;31(5):1211-1215. [doi: 10.1093/jamia/ocae034] [Medline: 38400737]
20. Turulja L, Cinjarevic M, Veselinovic L. Information technology and knowledge sharing for better health care: an emerging economy context. J Knowl Manag 2021 May 3;25(3):559-572. [doi: 10.1108/JKM-09-2019-0514]
21. Mozaffar H, Candi M. Extending the process frontier of digital transformation: a flow‐oriented perspective. Inf Syst J 2025 Mar;35(2):720-760. [doi: 10.1111/isj.12557]
22. Walsham G. The emergence of interpretivism in IS research. Inf Syst Res 1995 Dec;6(4):376-394. [doi: 10.1287/isre.6.4.376]
23. Locke K. The grounded theory approach to qualitative research. In: Drasgow F, Schmitt N, editors. Measuring and Analyzing Behavior in Organizations: Advances in Measurement and Data Analysis: Jossey-Bass/Wiley; 2002:17-43.
24. Gioia DA, Corley KG, Hamilton AL. Seeking qualitative rigor in inductive research: notes on the Gioia methodology. Organ Res Methods 2013;16(1):15-31. [doi: 10.1177/1094428112452151]
25. Strauss A, Corbin JM. Basics of Qualitative Research: Grounded Theory Procedures and Techniques: Sage Publications; 1990.

XSL•FO
RenderX

26. Rindova V, Dalpiaz E, Ravasi D. A cultural quest: a study of organizational use of new cultural resources in strategy formation. Organ Sci 2011 Apr;22(2):413-431. [doi: 10.1287/orsc.1100.0537]
27. Pawson R, Tilley N. Realist Evaluation: SAGE Publications; 2004. URL: https://www.copasah.org/uploads/1/2/6/4/12642634/realistic_evaluationpawson.pdf [accessed 2026-03-03]
28. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
29. Guennif S, Ramani SV. Explaining divergence in catching-up in pharma between India and Brazil using the NSI framework. Res Policy 2012 Mar;41(2):430-441. [doi: 10.1016/j.respol.2011.09.005]
30. Hagedoorn J, Lorenz-Orlean S, van Kranenburg H. Inter-firm technology transfer: partnership-embedded licensing or standard licensing agreements? Ind Corp Change 2009 Jun 1;18(3):529-550. [doi: 10.1093/icc/dtn038]
31. Mozaffar H, Williams R, Cresswell K, et al. Product diversity and spectrum of choice in hospital ePrescribing systems in England. PLoS One 2014;9(4):e92516. [doi: 10.1371/journal.pone.0092516] [Medline: 24691517]
32. Grimpe C, Sofka W. Complementarities in the search for innovation—managing markets and relationships. Res Policy 2016 Dec;45(10):2036-2053. [doi: 10.1016/j.respol.2016.07.007]
33. Bozeman B, Rimes H, Youtie J. The evolving state-of-the-art in technology transfer research: revisiting the contingent effectiveness model. Res Policy 2015 Feb;44(1):34-49. [doi: 10.1016/j.respol.2014.06.008]
34. Landry R, Amara N, Cloutier JS, Halilem N. Technology transfer organizations: services and business models. Technovation 2013 Dec;33(12):431-449. [doi: 10.1016/j.technovation.2013.09.008]
35. Lichtenthaler U, Ernst H. External technology commercialization in large firms: results of a quantitative benchmarking study. R&D Manag 2007 Nov;37(5):383-397. [doi: 10.1111/j.1467-9310.2007.00487.x]
36. Battistella C, De Toni AF, Pillon R. Inter-organisational technology/knowledge transfer: a framework from critical literature review. J Technol Transf 2016 Oct;41(5):1195-1234. [doi: 10.1007/s10961-015-9418-7]
37. Hinder S, Cresswell K, Sheikh A, et al. Promoting inter-organisational knowledge sharing: a qualitative evaluation of England's Global Digital Exemplar and Fast Follower Programme. PLoS One 2021;16(8):e0255220. [doi: 10.1371/journal.pone.0255220] [Medline: 34339429]
38. Milagres R, Burcharth A. Knowledge transfer in interorganizational partnerships: what do we know? Bus Process Manag J 2019 Feb 8;25(1):27-68. [doi: 10.1108/BPMJ-06-2017-0175]
39. Segrestin B. Partnering to explore: the Renault–Nissan Alliance as a forerunner of new cooperative patterns. Res Policy 2005 Jun;34(5):657-672. [doi: 10.1016/j.respol.2005.02.006]
40. Mozaffar H. User communities as multi-functional spaces: innovation, collective voice, demand articulation, peer informing and professional identity (and more). In: Hyysalo S, Jensen TE, Oudshoorn N, editors. The New Production of Users: Routledge; 2016:219-246 URL: https://www.taylorfrancis.com/chapters/edit/10.4324/9781315648088-12/user-communities-multi-functional-spaces-hajar-mozaffar [accessed 2026-02-26]
41. Mesquita LF, Anand J, Brush TH. Comparing the resource‐based and relational views: knowledge transfer and spillover in vertical alliances. Strat Manag J 2008 Sep;29(9):913-941. [doi: 10.1002/smj.699]
42. Johnson M, Mozaffar H, Campagnolo GM, Hyysalo S, Pollock N, Williams R. The managed prosumer: evolving knowledge strategies in the design of information infrastructures. Inf Commun Soc 2014 Aug 9;17(7):795-813. [doi: 10.1080/1369118X.2013.830635]
43. Cheung CMK, Chiu PY, Lee MKO. Online social networks: why do students use facebook? Comput Human Behav 2011 Jul;27(4):1337-1343. [doi: 10.1016/j.chb.2010.07.028]
44. Inkpen AC, Tsang EWK. Social capital, networks, and knowledge transfer. Acad Manag Rev 2005 Jan;30(1):146-165. [doi: 10.5465/amr.2005.15281445]
45. Kotabe M, Martin X, Domoto H. Gaining from vertical partnerships: knowledge transfer, relationship duration, and supplier performance improvement in the U.S. and Japanese automotive industries. SMJ 2003 Apr;24(4):293-316. [doi: 10.1002/smj.297]
46. Bhagat RS, Kedia BL, Harveston PD, Triandis HC. Cultural variations in the cross-border transfer of organizational knowledge: an integrative framework. Acad Manag Rev 2002 Apr;27(2):204. [doi: 10.2307/4134352]
47. Ireland RD, Hitt MA, Vaidyanath D. Alliance management as a source of competitive advantage. J Manage 2002 Jun;28(3):413-446. [doi: 10.1177/014920630202800308]
48. Darr ED, Kurtzberg TR. An investigation of partner similarity dimensions on knowledge transfer. Organ Behav Hum Decis Process 2000 May;82(1):28-44. [doi: 10.1006/obhd.2000.2885]
49. Berchicci L, de Jong JPJ, Freel M. Remote collaboration and innovative performance: the moderating role of R&D intensity. Ind Corp Change 2016 Jun;25(3):429-446. [doi: 10.1093/icc/dtv031]
50. Cohen WM, Levinthal DA. Absorptive capacity: a new perspective on learning and innovation. Adm Sci Q 1990 Mar;35(1):128-152. [doi: 10.2307/2393553]
51. Love JH, Roper S, Vahter P. Learning from openness: the dynamics of breadth in external innovation linkages. SMJ 2014 Nov;35(11):1703-1716. [doi: 10.1002/smj.2170]

XSL•FO
RenderX

52. Tzabbar D, Aharonson BS, Amburgey TL. When does tapping external sources of knowledge result in knowledge integration? Res Policy 2013 Mar;42(2):481-494. [doi: 10.1016/j.respol.2012.07.007]
53. Nooshinfard F, Nemati-Anaraki L. Success factors of inter-organizational knowledge sharing: a proposed framework. Electron Libr 2014 Apr 1;32(2):239-261. [doi: 10.1108/EL-02-2012-0023]
54. Hung SY, Durcikova A, Lai HM, Lin WM. The influence of intrinsic and extrinsic motivation on individuals' knowledge sharing behavior. Int J Hum Comput Stud 2011 Jun;69(6):415-427. [doi: 10.1016/j.ijhcs.2011.02.004]
55. Lampel J, Bhalla A. The role of status seeking in online communities: giving the gift of experience. J Comput Mediat Commun 2007 Jan;12(2):434-455. [doi: 10.1111/j.1083-6101.2007.00332.x]
56. Lakhani KR, von Hippel E. How open source software works: "free" user-to-user assistance. Res Policy 2003 Jun;32(6):923-943. [doi: 10.1016/S0048-7333(02)00095-1]
57. Hall H, Graham D. Creation and recreation: motivating collaboration to generate knowledge capital in online communities. Int J Inf Manage 2004 Jun;24(3):235-246. [doi: 10.1016/j.ijinfomgt.2004.02.004]
58. Christensen CM, Raynor ME. Why hard-nosed executives should care about management theory. Harv Bus Rev 2003 Sep;81(9):66-74. [Medline: 12964394]
59. McLure Wasko M, Faraj S. "It is what one does": why people participate and help others in electronic communities of practice. J Strateg Inf Syst 2000 Sep;9(2-3):155-173. [doi: 10.1016/S0963-8687(00)00045-7]
60. Blanchard AL, Markus ML. The experienced "sense" of a virtual community. Data Base Adv Info Sys 2004 Feb 3;35(1):64-79. [doi: 10.1145/968464.968470]
61. Wasko MM, Faraj S. Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. MIS Q 2005 Mar 1;29(1):35-57. [doi: 10.2307/25148667]
62. Bateman PJ, Gray PH, Butler BS. Research note—the impact of community commitment on participation in online communities. Inf Syst Res 2011 Dec;22(4):841-854. [doi: 10.1287/isre.1090.0265]
63. Peng H. Why and when do people hide knowledge? J Knowl Manag 2013 May 24;17(3):398-415. [doi: 10.1108/JKM-12-2012-0380]
64. Meyer JP, Herscovitch L. Commitment in the workplace: toward a general model. Hum Resour Manag Rev 2001 Sep;11(3):299-326. [doi: 10.1016/S1053-4822(00)00053-X]
65. Jarvenpaa SL, Majchrzak A. Interactive self-regulatory theory for sharing and protecting in interorganizational collaborations. Acad Manag Rev 2016 Jan;41(1):9-27. [doi: 10.5465/amr.2012.0005]
66. Inkpen AC, Currall SC. The coevolution of trust, control, and learning in joint ventures. Organ Sci 2004 Oct;15(5):586-599. [doi: 10.1287/orsc.1040.0079]
67. Mozaffar H, Panteli N. The online community knowledge flows: distance and direction. Eur J Inf Syst 2022;31(2):227-240. [doi: 10.1080/0960085X.2020.1866442]

Abbreviations

CMO: context-mechanism-outcome
COREQ: Consolidated Criteria for Reporting Qualitative Research
FF: fast follower
GDE: Global Digital Exemplar
NHS: National Health Service

Edited by A Stone; submitted 01.Sep.2025; peer-reviewed by L Desveaux, M Dohan; revised version received 10.Feb.2026; accepted 10.Feb.2026; published 12.Mar.2026.

Please cite as:
Mozaffar H, Williams R, Cresswell K
Analysis of Multilevel Factors Mobilizing the Spectrum of Interorganizational Knowledge Sharing for Facilitating Digital Transformation at Scale: Qualitative Study
J Med Internet Res 2026;28:e83345
URL: https://www.jmir.org/2026/1/e83345
doi:10.2196/83345

© Hajar Mozaffar, Robin Williams, Kathrin Cresswell. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly

XSL•FO
RenderX

cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Virtual Multidisciplinary Gastrointestinal Care for Adults With Gastrointestinal Needs: Retrospective Cohort Study

Grace Wang[1], MPH, PhD; William D Chey[2], MD; Sanskriti Varma[3], MD; Sameer Berry[1,4], MD

[1]Oshi Health, Inc., New York, NY, United States

[2]Department of Internal Medicine, Division of Gastroenterology, University of Michigan, Ann Arbor, MI, United States

[3]Massachusetts General Hospital, Boston, MA, United States

[4]New York University, New York, NY, United States

Corresponding Author:
Grace Wang, MPH, PhD
Oshi Health, Inc.
228 Park Avenue South, Suite 16389
New York, NY, 10003
United States
Phone: 1 (646) 876 8462
Email: gracewang@oshihealth.com

Abstract

Background: Gastrointestinal (GI) disorders are highly prevalent, costly, and complex. Multidisciplinary gastrointestinal care (MGC), integrating medicine, nutrition, and behavioral health, is a best practice for managing GI needs, but access is limited by the availability of gastroenterologists and MGC clinics. Virtual MGC may bridge the gap, but it is unclear the extent to which patients engage in virtual MGC and the outcomes of virtual services delivered at scale.

Objective: This study describes patient characteristics, engagement, and outcomes of a large-scale virtual MGC program and evaluates whether dietitian and behavioral health support mediates the association between medical engagement and symptom improvement.

Methods: We conducted a retrospective cohort study of 11,345 adult patients receiving virtual MGC with gastroenterologists, GI-specialized advanced practice providers (APPs), registered dietitians, licensed psychologists, and care coordinators between April 2021 and August 2025. Patients completed virtual onboarding with a GI APP and medical, dietary, and behavioral interventions delivered through synchronous telehealth and asynchronous chat. Data were from web-based intake forms and electronic health records. Patient-reported outcomes included symptom control and symptom improvement (Yes/No). Descriptive analyses, logistic regression, and path analysis were conducted across 20 imputed datasets.

Results: Virtual patients receiving MGC were 43.3 (SD 12.78) years on average, and 66.4% (7532/11,345) were female. Patients primarily presented with Disorders of Gut-Brain Interaction (4460/11,345, 39.3%) and Gastroesophageal Reflux Disease (2775/11,345, 24.5%). The median wait time for an initial appointment was 6 days (IQR 3-9). Patients had a median of 2 GI visits (IQR 1-3), 2 dietitian visits (IQR 1-3), and 1 behavioral health visit (IQR 0-2). As the number of visits increases, the odds of achieving positive outcomes increase significantly after controlling for age, sex, symptom severity at baseline, symptom frequency at baseline, and days in care. This translates to 92.39% with symptom improvement, 94.67% with symptom control, and 98.13% with no noticeable or mild symptoms among those with 4 or more appointments. Path analysis confirmed that GI APP engagement was significantly associated with increased dietitian and behavioral health usage, which was associated with symptom improvement. The direct pathway from GI APP engagement to symptom improvement was also significant.

Conclusions: This innovative study demonstrates that a virtual-first MGC model is not only feasible at a national scale but is effective in achieving symptom control and improvement across a clinically diverse GI population. We provide evidence about successfully delivering high-quality care outside traditional clinical settings. This work distinguishes itself by analyzing the mechanisms of integrated care, specifically how medical engagement facilitates the use of specialized nutritional and behavioral interventions that are often inaccessible in community care. In the real world, this model offers a scalable solution to geographic specialist shortages, ensuring that best-practice care is available to patients regardless of their location or the local supply of specialists.

(J Med Internet Res 2026;28:e89061) doi:10.2196/89061

XSL•FO
RenderX

KEYWORDS

gastroenterology; patient care team; telemedicine; advanced practice provider; disorders of the gut-brain interaction

Introduction

Problem and Review of Relevant Scholarship

Gastrointestinal (GI) disorders are highly prevalent, decrease patients' quality of life, and incur considerable cost to the US health system. About 6 out of 10 adults report GI symptoms every week [1]. Many patients experience shame, high rates of absenteeism, and disruptions to their social life [2,3]. In 2021, GI-related health care expenditures totaled US $111.8 billion from medical services, including 14.5 million emergency department visits and 2.9 million hospital admissions [4]. GI-related health care expenditure exceeds that of mental health, heart disease, and breast and prostate cancers [5-7].

Because GI disorders can be complex with considerable symptom and diagnosis overlap, multidisciplinary approaches are increasingly recognized as a new best practice. Multidisciplinary gastrointestinal care (MGC) integrates therapies from medicine, nutritional science, and behavioral health to offer a holistic approach to GI needs [8]. In a randomized controlled trial, a multidisciplinary care model showed superior efficacy, with 84% of participants reporting global symptom relief compared with only 57% of those in the conventional care arm [9]. These advantages endured; at 1 year post treatment, patients in the multidisciplinary group were nearly twice as likely as conventional care patients to report their symptoms as improved [10]. Furthermore, evidence from a matched cohort study confirms that the multidisciplinary approach yields significantly better outcomes than standard care [11]. Reflecting this growing evidence base, current clinical guidelines from the American College of Gastroenterology now incorporate gut-directed psychotherapies as a primary recommendation for managing global irritable bowel syndrome symptoms [12].

Despite this clinical consensus, in-person MGC is not yet widespread. There are profound geographic disparities in access to gastroenterologists across the United States. Recent research notes that 39% of 1167 metropolitan counties and 87% of 1975 nonmetropolitan counties had no gastroenterologists [13]. Even in many major metropolitan areas, the average wait time for an in-person gastroenterology appointment reached 40 days in 2025 [14]. While in-person MGC clinics exist, they are typically offered by academic medical centers, leaving the vast majority of the population without access to integrated dietary and behavioral support. Furthermore, referral processes to dietitians and behavioral health providers and reimbursement approaches for MGC can be greatly improved [15,16].

Virtual MGC models have been proposed to bridge this gap, with evidence suggesting that telehealth is effective for managing GI needs and highly acceptable to patients [17,18]. In response, we designed a virtual MGC program with access to board-certified gastroenterologists, GI-specialized advanced practice providers (APPs), registered dietitians, licensed psychologists, and care coordinators. The program's objective is to provide access to high-quality MGC that patients nationwide need, wherever they are. However, it is unclear the extent to which patients engage in virtual MGC and the magnitude of outcomes that can be achieved with virtual delivery of services at scale.

Aims and Objectives

To address these gaps, our aim was to describe (1) the patients who received virtual MGC, (2) patient engagement with virtual MGC, (3) symptom improvement experienced by patients and satisfaction with services, and (4) the relationship between engagement and symptom improvement among virtual patients receiving MGC. The second aim was to examine whether the association between GI engagement and symptom improvement is mediated by the activation of specialized dietitian and behavioral health services. To achieve these aims, our objective was to conduct an observational study with more than 11,000 patients with GI conditions who received virtual MGC.

Methods

This manuscript follows Journal Article Reporting Standards reporting guidelines [19].

Inclusion and Exclusion

We included patients who met all of the following: ≥18 years of age at enrollment; reside in the United States; have access to internet or mobile technology; English speaking; provided intake data; had their first onboarding appointment with an APP between April 2021 through August 2025; and were referred to and proceeded to virtual MGC (ie, care with the broader GI, behavioral health, and dietician team).

Participant Characteristics

This study includes 11,345 virtual patients receiving MGC. Of these, 7532 (66.39%) were female. The average age was 43.27 (SD 12.78) years, and patients lived in all 50 US states and the District of Columbia.

Sampling Procedures and Setting

The study included all patients meeting the inclusion criteria. Virtual MGC includes board-certified gastroenterologists, GI-specialized APPs, registered dietitians, licensed psychologists, and care coordinators. Access was through 3 primary channels: health plan networks, corporate benefits, and external provider referrals. Eligible patients were informed about virtual MGC services via email or mailings directing them to a web-based registration system. After completing an intake questionnaire and providing demographic details, patients self-schedule a virtual onboarding visit with a GI APP.

The care team requests and reviews past medical records before the first visit. The APP identifies a preliminary diagnosis and provides a written care plan, which may include diagnostic testing, procedures, over-the-counter or prescription medications, dietary interventions, and brain-gut behavioral therapies. GI APPs follow evidence-based protocols to diagnose, treat, and prescribe independently, depending on the patient's

XSL•FO
RenderX

state of residence; gastroenterologists on the care team are available for consultation and verification. The care plan is available to patients and other care team members in the patient's electronic health record (EHR). Patients also receive tracking tools and educational resources.

A subset of patients is offered and enter multidisciplinary care with a medical, dietary, and behavioral health care team who monitors and may revise the initial treatment plan according to patient needs. Multimedia Appendix 1 lists example interventions offered by the care team. Patients have as many virtual visits with care team members as needed until achieving symptom control.

Synchronous telehealth visits are via the clinic's digital platform, which also offers asynchronous chat messaging with care teams for support in between visits or calls. When in-person services (eg, endoscopy and imaging) are warranted, the care team refers patients to in-network partner facilities, and care coordinators support patients with referrals.

Sample Size, Power, and Precision

The sample size was determined by the total number of eligible patients who received virtual MGC between April 2021 and August 2025 and met the inclusion criteria. As this was a retrospective cohort study using real-world data, no a priori power analysis was conducted. The final analytical sample included 11,345 patients. For primary binary outcomes, the large sample size yielded narrow 95% CIs around effect estimates (see Results), indicating high statistical precision. Given the observed outcome rates, the available sample provided adequate power (≥0.80 at α=.05) to detect small effect sizes in multivariable logistic regression models.

Measures

This analysis included the following variables.

Baseline Patient Characteristics

Demographic variables included age, sex, state of residence, GI conditions and symptoms, and comorbidities at baseline. We also examined 3 symptom items at baseline: symptom severity (1 not noticeable [no disruption to daily life], 2 mild [minimal disruption to daily life], 3 moderate [disruptive to daily life], 4 severe [very disruptive to daily life]), number of symptoms, and symptom frequency (1 day per month or less, 2-3 times per month, 1-2 days per week, 3-4 days per week, 5-6 days per week, 7 days per week).

Engagement

Metrics include: time from account creation to first onboarding appointment with an APP and number of visits per patient overall and by appointment type (ie, medical, dietary, or behavioral). Duration was measured using visit minutes and call minutes per patient by appointment type. To exclude voicemails, we included calls lasting 5 minutes or more. For chats, we report on the number of messages initiated by patients.

Patient-Reported Outcomes

Symptom Control was measured by, "Do you feel [the virtual MGC] has helped you gain control over your GI condition or symptoms during the course of this program?" (Yes/No). When patients provided multiple responses over time to these items, this analysis included the best outcome achieved.

To create Symptom Improvement (0=no improvement, 1=improvement), we recategorized patient responses to the question, "Compared to before I was seen by [virtual MGC], my GI condition is now: worse, slightly worse, same, slightly better, better." Those who responded worse, slightly worse, or the same were categorized as no improvement. Those who responded slightly better or better were categorized as improved. For those with missing responses to this question, we examined baseline symptom severity compared with symptom severity score after receiving virtual MGC. Symptom severity responses were 1 not noticeable (no disruption to daily life), 2 mild (minimal disruption to daily life), 3 moderate (disruptive to daily life), 4 severe (very disruptive to daily life). Those whose scores decreased by 1 or more points were categorized as 1 on Symptom Improvement, else 0.

Symptom Severity was transformed into a binary variable with categories not noticeable or mild (0), or moderate or severe (1).

Satisfaction

Participants answered, "How satisfied were you with your prior GI care?" at baseline and, "How satisfied were you with your recent [virtual MGC] experience?" after each visit. Responses were: 1 not satisfied at all, 2 somewhat dissatisfied, 3 neither satisfied nor dissatisfied, 4 somewhat satisfied, 5 very satisfied, or I have not previously worked with a GI provider (baseline only). The average score was used for those who provided multiple responses after multiple visits.

Data Collection

Data were collected in the following ways.

- Intake forms: Patients completed intake forms via the web before their first onboarding appointments with GI APPs. Patient characteristics and symptom severity were self-reported through intake forms.
- EHR and Health Cloud data systems: *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) and SNOMED codes were gathered from EHRs and used to describe patients' GI-specific conditions, symptoms, and comorbidities. Patient-reported outcomes (PROs) were recorded in Health Cloud (Salesforce) by providers during visits.
- Scheduling systems: Engagement data were gathered from the scheduling system.
- Patient surveys: Satisfaction was collected as an electronic survey after each visit.

Quality of Measurements

To ensure data collection quality, intake forms included data-type constraints (eg, restricted numeric entry for age) to prevent erroneous entries. At the clinical level, all providers are trained to collect outcomes uniformly.

Instrumentation and Psychometrics

Patient outcomes were assessed using modified Patient Global Impression scales for symptom severity, improvement, and control. In this study, the scales focused on virtual MGC and

XSL•FO
RenderX

used a 4-point scale for symptom severity, a 5-point scale for improvement, and a binary scale for control. These instruments are widely used single-item global ratings designed to capture patients' subjective evaluation of symptom severity and perceived improvement over time. They have demonstrated face validity, construct validity, responsiveness to clinical change, and interpretability across chronic conditions [20-23]. They have been used as anchors for psychometric testing of other scales in GI research [24-26].

Masking

Due to the retrospective, single-arm observational design of this study, masking of participants and providers was not applicable. All patients were aware they were receiving virtual MGC as part of their clinical treatment plan, and providers were aware of the collection of clinical data and PROs as part of their routine care delivery.

Conditions and Design

We conducted a nonexperimental, retrospective cohort study of patients who first received virtual MGC between April 2021 and August 2025.

We used Andersen's model [27] of health care usage to guide variable selection for the analysis of the relationship between engagement and symptom improvement (Multimedia Appendix 2).

Data Diagnostics

Before analysis, all 41 variables were screened for outliers using distributional inspection; all observed values were determined to be plausible and were retained. The overall proportion of missing data across all 41 variables was 3.0% (14,077 of 465,145 data points). Little's test indicated that data were not missing completely at random (χ^2_{688}=5013.97, P<.001) [28]. Based on missingness patterns supporting the assumption that data were missing at random, we performed multiple imputation by chained equations. Twenty imputed datasets were generated, and estimates were pooled using Rubin's rules. All generalized structural equation modeling (GSEM) and logistic regression models achieved convergence across the 20 imputed datasets. The number of missing and imputed values for all model variables is provided in Multimedia Appendix 3.

Analytic Strategy

We calculated descriptive statistics across 20 pooled imputed datasets for all patient characteristics, engagement, PRO, and satisfaction variables.

To examine associations between engagement and outcomes (symptom improvement and symptom control), we conducted logistic regression modeling across 20 imputed datasets. Adjusted models controlled for age, sex, symptom severity at baseline, symptom frequency at baseline, and days in care. Model discrimination was assessed using the area under the receiver operating characteristic curve in a representative imputed dataset. Following estimation, we used the margins command on the pooled results to calculate the predicted probabilities of symptom improvement and symptom control. These average marginal effects are reported below.

Path analysis was conducted using GSEM within a multiple-imputation framework (20 imputations) to evaluate the mediating role of multidisciplinary specialist engagement (Registered Dietitian and Behavioral Health) on the relationship between GI APP visits and symptom improvement. Models were adjusted for age, sex, symptom severity at baseline, symptom frequency at baseline, presence of disorders of the gut-brain interaction (DGBI), number of days in care, client, and state. Inference was based on multiple-imputation–combined estimates using large-sample (asymptotic) standard errors, such that test statistics follow the standard normal (z) distribution. Statistical significance was assessed using 2-sided tests with α of .05. Indirect effects were estimated using the product-of-coefficients method on the log-odds scale. Because odds ratios (ORs) are noncollapsible, total effects were estimated separately, and the proportion mediated was not calculated.

All analyses were conducted in Stata SE (version 19.5; StataCorp LLC).

Ethical Considerations

Pearl Institutional Review Board reviewed the study protocol and determined the study to be Exempt according to 45 CFR 46.104(d)(4) Secondary Research Uses of Data or Specimens. This study involves the analysis of data generated during routine clinical care. The institutional review board granted a waiver of informed consent, as the research posed minimal risk to participants and could not practicably be carried out otherwise. To protect participant privacy and confidentiality, all data used in this analysis were deidentified in accordance with the HIPAA (Health Insurance Portability and Accountability Act) Privacy Rule. Direct identifiers were removed, and unique study identifiers were used in place of personal information. Data were stored on secure, access-restricted servers, and only authorized study personnel had access to the analytic dataset. Participants did not receive financial or other compensation for inclusion in this secondary data analysis. This manuscript does not include images, figures, or supplementary materials that could identify individual participants.

Results

Patient Flow

Between April 2021 and August 2025, 18,235 patients had a first virtual appointment with a GI APP and provided intake data. Of these, GI APP used their clinical judgment to refer 15,708 to virtual MGC with a registered dietitian or behavioral health provider. Of the referred, 11,345 received virtual MGC (Figure 1).

XSL•FO
RenderX

Figure 1. Patient flow from onboarding to virtual MGC for patients with GI needs who received care between April 2021 and August 2025. APP: advanced practice provider; GI: gastrointestinal; MGC: multidisciplinary gastrointestinal care.

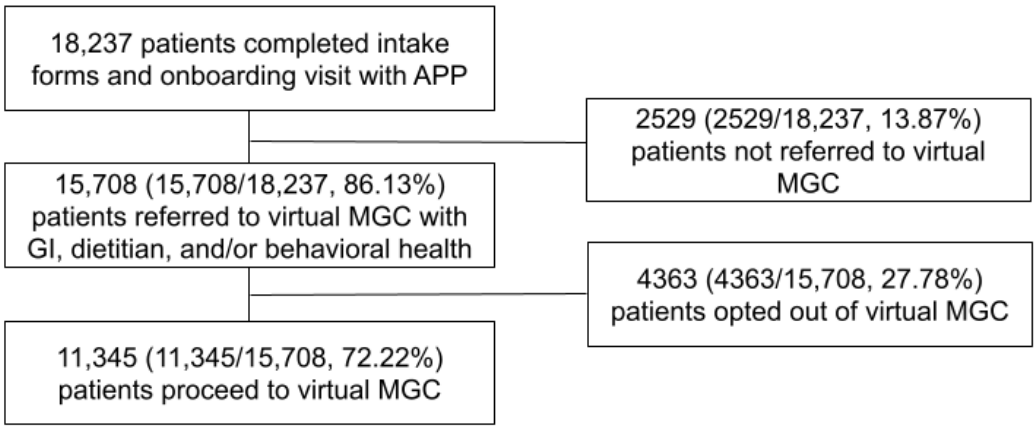

Patient Description

Table 1 shows virtual patients receiving MGC characteristics at baseline. The majority (7924/11,345, 69.85%) of patients reported that GI symptom severity was moderate (disruptive to daily life) or severe (very disruptive to daily life). More than half (6275/11,345, 55.30%) experienced symptoms 5 or more days per week.

The most common diagnoses at first appointments for 11,345 patients were DGBI (4460/11,345, 39.31%), gastroesophageal reflux disease (2775/11,345, 24.46%), and inflammatory bowel disease (441/11,345, 3.89%). The most common symptoms reported by 11,345 patients were flatulence (3751/11,345, 33.06%), bloating (3587/11,345, 31.62%), and heartburn or reflux (2887/11,345, 25.45%) (Table 2).

Among 8989 patients with known comorbidities, the average number was 2.27 (SD 1.37) comorbid conditions, with behavioral health (3963/8989, 44.09%), metabolic (3823/8989, 42.53%), and cardiovascular (3595/8989, 39.99%) conditions being the most prevalent.

XSL•FO
RenderX

Table 1. Characteristics at baseline for virtual multidisciplinary gastrointestinal care patients who received care between April 2021 and August 2025 (n=11,345).

Characteristics	Value, n (%)
Sex	
Female	7532 (66.39)
Male	3813 (33.61)
Age at enrollment (years)	
18 to 29	1572 (13.86)
30 to 39	3281 (28.92)
40 to 49	2840 (25.03)
50 to 59	2279 (20.09)
60 and older	1373 (12.10)
Symptom severity at baseline	
Not noticeable (no disruption to daily life)	230 (2.02)
Mild (minimal disruption to daily life)	3191 (28.13)
Moderate (disruptive to daily life)	6212 (54.76)
Severe (very disruptive to daily life)	1712 (15.09)
Symptom frequency at baseline	
1 day per month or less	275 (2.42)
2-3 times per month	859 (7.57)
1-2 days per week	1423 (12.55)
3-4 days per week	2513 (22.15)
5-6 days per week	2435 (21.46)
7 days per week	3840 (33.84)

XSL•FO
RenderX

Table 2. Ten most prevalent diagnoses and symptoms at baseline from ICD-10 (International Statistical Classification of Diseases, Tenth Revision) codes in the electronic health record among virtual multidisciplinary gastrointestinal care patients who received care between April 2021 and August 2025 (n=11,345)

Characteristic	Value, n (%)
Diagnoses	
DGBI[a] (including irritable bowel syndrome and other functional disorders)	4460 (39.31)
GERD[b]	2775 (24.46)
IBD[c]	441 (3.89)
Anal or rectal fissures, fistula, or abscesses	283 (2.49)
Gastroenteritis	257 (2.27)
Hemorrhoids	229 (2.02)
Intestinal malabsorption	222 (1.96)
Diseases of the liver	191 (1.68)
Small intestinal bacterial overgrowth	180 (1.59)
Diverticular disease	166 (1.46)
Symptoms	
Flatulence	3751 (33.06)
Bloating	3587 (31.62)
Heartburn or reflux hypersensitivity	2887 (25.45)
Constipation	2254 (19.87)
Abdominal or pelvic pain	2200 (19.39)
Diarrhea	1610 (14.19)
Nausea and vomiting	921 (8.12)
Change in bowel habits	651 (5.74)
Chest pain	568 (5.01)
Fecal abnormalities	499 (4.40)

[a]DGBI: disorders of the gut-brain interaction.

[b]GERD: gastroesophageal reflux disease.

[c]IBD: inflammatory bowel disease.

Patient Engagement With Virtual MGC

The median number of days for patients to schedule their first GI APP appointment was 6 (IQR 3-9) days. The median number of days from the first GI APP appointment to an appointment with either a registered dietician or behavioral health provider was 15 (IQR 7-28) days.

The median number of visits with GI providers was 2 (IQR 1-3), while registered dietitians and behavioral health providers saw a median of 2 (IQR 1-3) and 1 (IQR 0-2) visits, respectively. On average, GI appointments were 59.31 (SD 45.14) minutes, and calls were 25.16 (SD 25.05) minutes. Behavioral health appointments were 38.85 (SD 52.05) minutes, and calls were 31.07 (SD 22.11) minutes. Dietitian appointments were 64.42 (SD 122.35) minutes, and calls were 29.51 (SD 18.99) minutes. On average, patients initiated 8.36 (SD 17.83) chat messages with their care team.

Clinical Outcomes

Of the 11,345 patients, 9683 (85.4%) patients reported symptom improvement over time, and 10,189 (89.8%) patients reported symptom control. Median time to symptom control was 29 days. More than 95% (10,803/11,345) reported that symptom severity posed minimal to no disruption to daily life after virtual MGC. Figure 2 reports symptom improvement, control, and severity stratified by baseline symptom severity.

XSL•FO
RenderX

Figure 2. Percent of virtual multidisciplinary gastrointestinal care patients who achieved symptom improvement, symptom control, and no noticeable or mild symptoms after virtual multidisciplinary gastrointestinal care, by baseline symptom severity.

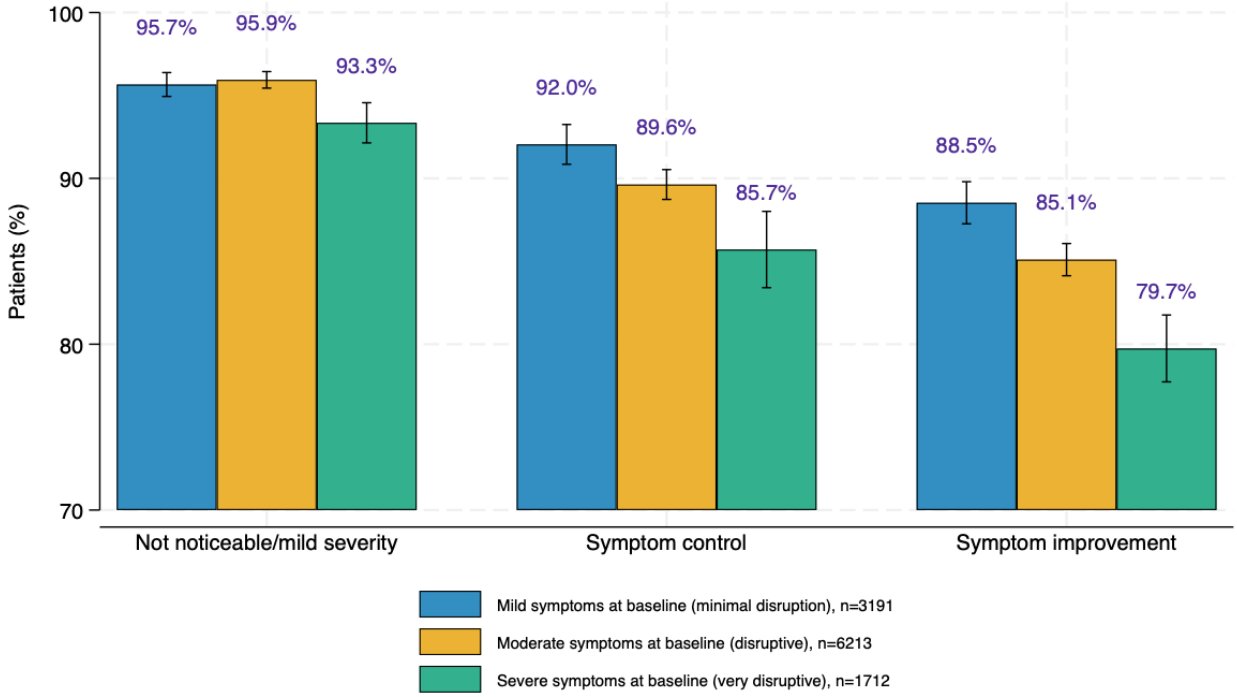

Satisfaction With Services

We examined care satisfaction scores for 7596 patients who previously saw an in-person provider for their GI needs. Mean satisfaction scores increased significantly from 2.74 (SD 1.31) at baseline to 4.82 (SD 0.61) after virtual MGC.

Association Between Engagement and Outcomes

As the number of visits increases, the odds of achieving positive outcomes increases after controlling for age, sex, symptom severity at baseline, symptom frequency at baseline, and days in care (symptom improvement: OR 1.51, 95% CI 1.44-1.58; symptom control: OR 1.46, 95% CI 1.38-1.55; not noticeable or mild symptoms: OR 1.91, 95% CI 1.77-2.06). For example, Figure 3 shows the predicted probability of symptom improvement is 69.58% for patients with 2 or 3 appointments versus 92.39% with 4 or more appointments.

XSL•FO
RenderX

Figure 3. Predicted probability of virtual multidisciplinary gastrointestinal care patients achieving positive outcomes (symptom improvement, symptom control, not noticeable or mild symptoms), by number of appointments or calls with the care team. Estimates control for age, sex, symptom severity at baseline, symptom frequency at baseline, and days in care (n=11,345).

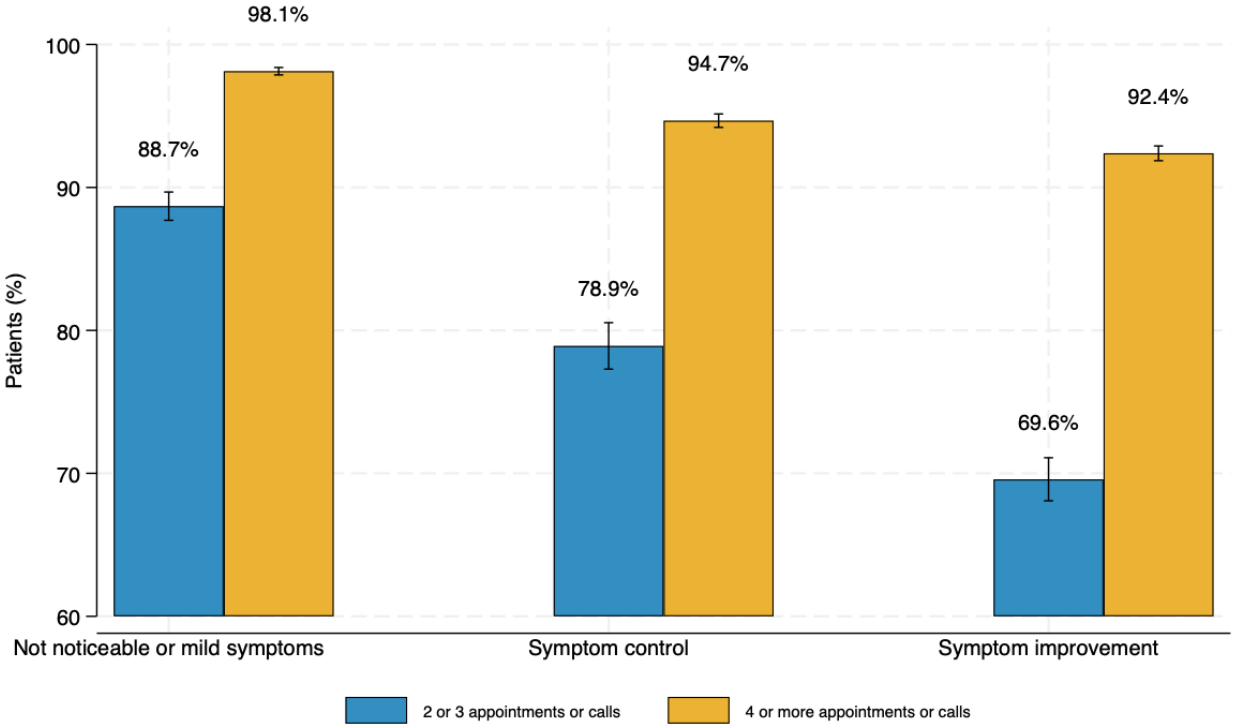

GI APP Engagement and Symptom Improvement, Mediated by Dietitian and Behavioral Health Support

Path analysis using GSEM (Figure 4) showed that GI APP appointments were significantly associated with an increase in multidisciplinary care team appointments with registered dietitians and behavioral health providers (path a; β=0.27, 95% CI 0.24-0.30; *z*=18.04, *P*<.001). In turn, these specialized multidisciplinary services were significantly and positively associated with symptom improvement (path b; OR 1.50, 95% CI 1.43-1.58; *z*=16.22, *P*<.001). The direct pathway from GI APP engagement to symptom improvement (path c) remained statistically significant after adjustment for the multidisciplinary mediator (OR 1.66, 95% CI 1.51-1.80; *z*=22.54, *P*<.001). In the reduced model excluding the mediator, the total effect of GI APP appointments on symptom improvement was also statistically significant (OR 1.50, 95% CI 1.37-1.62; *z*=23.09, *P*<.001). All pathways were adjusted for age, sex, baseline symptom severity and frequency, symptom count, DGBI diagnosis, employer, and state fixed effects.

Figure 4. Path analysis of the multidisciplinary care model. Standardized coefficients (beta) and odds ratios (OR) are presented for the direct and mediated pathways. Models adjust for age, sex, baseline symptom severity, baseline symptom frequency, symptom number, DGBI diagnosis, employer, and state fixed effects. Estimates are pooled from 20 imputed datasets (n=11,345).

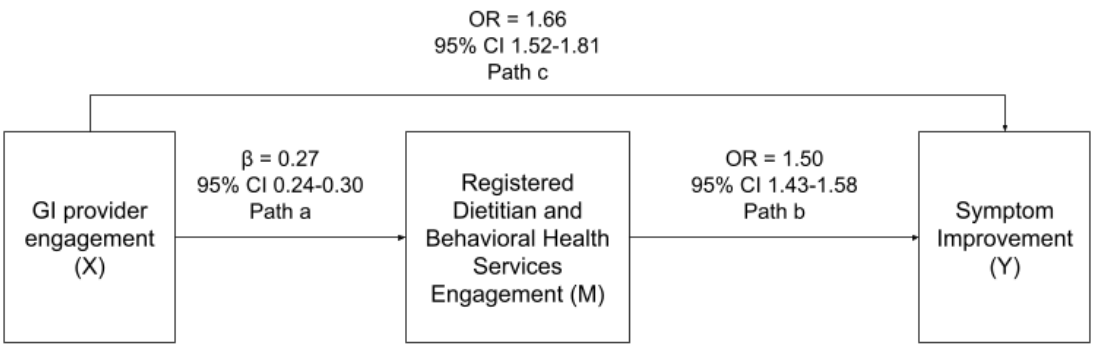

XSL•FO
RenderX

Discussion

Principal Results

This observational study reports the performance of virtual MGC at a national scale for more than 11,000 patients. Our first aim was to describe (1) the patients who received virtual MGC, (2) patient engagement with virtual MGC, (3) symptom improvement experienced by patients and satisfaction with services, and (4) the relationship between engagement and symptom improvement among virtual patients receiving MGC. We show that a virtual MGC model is able to serve patients from across the United States, with a range of GI needs, and experiencing mild to severe symptoms. Patients all received care by GI APPs, and most had multiple and extensive interactions with behavioral and nutritional health care team members. Patient symptom severity improved significantly over time, and the vast majority of patients reported that they were in control of their symptoms. Patients reported high satisfaction with care, and engagement with the multidisciplinary care team was significantly associated with positive outcomes. The second aim was to examine whether the association between GI engagement and symptom improvement is mediated by the activation of specialized dietitian and behavioral health services. Path analysis confirmed that the association between GI APP engagement and symptom improvement is significantly mediated by the activation of specialized nutritional and behavioral services.

Similarity of Results

The clinical outcomes in this study are consistent with an RCT by Basnayake et al [9], which reported that 85% of patients in an in-person MGC group achieved symptom improvement compared with only 65% of patients in the standard gastroenterologist-only care group. These findings suggest that virtual MGC, when delivered in real-world settings, can achieve similar positive outcomes as those observed in a controlled clinical trial.

Interpretation

Patient satisfaction with virtual GI services was high, which was consistent with previous studies on virtual GI care [29-31]. We hypothesize that this high satisfaction stems from patients' preference for the convenience of virtual visits and perceptions that the quality of care is comparable to in-person services. A 2024 national survey of patients who used telehealth supported this notion, with 86% agreeing that "telehealth made it easier for me to get care when and where I needed it" and 78% agreeing that "the care I received through telehealth was as good as a regular in-person visit" [32].

Our study found that engagement with the multidisciplinary team was a meaningful pathway in the relationship between GI APP and symptom improvement. This positive relationship between MGC and outcomes is compatible with clinical trial results and evidence-based practice guidelines that support MGC [10,33-35]. The ongoing support from registered dieticians meets patients' needs for guidance about nutrients, adequate caloric intake, and appropriate diet interventions tailored to the patient's lifestyle [36]. Access to the dietitian over time is crucial for determining any needed modifications to the recommended diet and supporting adherence to the diet [37].

Similarly, ongoing support from behavioral health providers helps patients learn and refine approaches for managing the psychosocial factors that can trigger their conditions. Behavioral health can help with coping mechanisms for GI-specific symptoms and facilitate rewiring of neuronal inputs, which leads to improved functional disorders [12,38-40]. GI-focused behavioral therapies exert therapeutic effects through bidirectional modulation of the brain-gut axis, targeting not only central processes such as mood and cognitive appraisal, but also peripheral mechanisms including pain perception, visceral hypersensitivity, and GI motility [12,38-40]. Furthermore, behavioral health providers help many patients with anxiety and depression exacerbated by GI conditions. Together, the team is able to evaluate the benefits and risks of food restrictions on the patient's GI and psychological needs [41].

Generalizability

The demographics of the virtual patients receiving MGC base were similar to those reported by Almario et al [1] in their nationally representative study of GI symptoms in the US population. Almario et al [1] found that 52% of adults with GI symptoms were 25 to 44 years old, and 60% were female, which aligns with the age and sex distribution observed in our cohort. Furthermore, the most common GI conditions for virtual patients receiving MGC (ie, DGBI, gastroesophageal reflux disease, and inflammatory bowel disease) correspond to the most prevalent symptoms (eg, bloating, heartburn or reflux, abdominal or pelvic pain, diarrhea, and constipation) reported by Almario et al [1]. This consistency suggests that the patient population seeking virtual MGC is broadly representative of individuals experiencing GI issues in the general US population.

Implications

The findings of this study have implications for research, systems, and policy. Future research could focus on identifying the minimum effective dose of multidisciplinary touchpoints required to achieve and sustain positive outcomes. Prospective trials comparing virtual MGC directly to traditional community-based care are warranted to further quantify its impact on clinical outcomes and health care usage. Research is also needed to evaluate alternative payment models that best support virtual MGC.

For health systems and policy makers, virtual care delivery may be an important approach for addressing provider shortages and associated delays in care. This is underscored by our finding that virtual patients receiving MGC waited a median of only 6 days for a first visit, in contrast to the average 40-day wait time for an in-person gastroenterologist visit in major metropolitan areas in 2025 [14]. In addition, our results support more emphasis on integrated clinical ecosystems with medical providers directing care and coordinating with dietitians and behavioral providers. This virtual-first, integrated model may serve as a blueprint for value-based care programs designed to improve patient quality of life and reduce downstream costs associated with GI needs. Reimbursing by milestones instead

XSL•FO
RenderX

of traditional fee-for-service billing would allow patients to have as many visits as needed with any provider type.

Strengths and Limitations

A strength of this study is the large sample size of >11,000 patients from across the United States and with a range of GI conditions. This demonstrates the feasibility of delivering virtual MGC at scale, and the diverse patient base supports the generalizability of findings to a broad US patient population. Our study is also the first, to our knowledge, to demonstrate the role of nutritional and behavioral support in mediating the association between GI medical care and symptom improvement within a virtual MGC setting.

Our study has a number of limitations, including the absence of a control or comparison group, as our analysis only includes individuals who received virtual MGC. Therefore, we cannot rule out the possibility that the natural history of the GI conditions, which are often relapsing and remitting, may have contributed to the observed symptom improvements. Future research should compare to a control group not receiving virtual MGC or receiving only in-person care.

Our analysis did not include variables such as patient health behaviors or in-person health care usage, which could influence symptom improvement. A future study using medical claims could incorporate important factors, such as diagnostic testing, procedures, and prescriptions. Another limitation is our use of global impression scales to capture symptom improvement, control, and severity. Global impression scales minimize patient burden and are practical for large-scale observational research [42]. However, these scales do not capture the granular information about symptoms like other validated measures, such as Patient-Reported Outcomes Measurement Information System (PROMIS) GI scales [43,44]. Future studies incorporating both global impression and PROMIS scales could be used to more comprehensively evaluate virtual MGC.

Conclusions

This innovative study demonstrates that a virtual-first MGC model is not only feasible at a national scale but is effective in achieving symptom control and improvement across a clinically diverse GI population. Furthermore, we demonstrate that high-quality, evidence-based GI care can be successfully delivered outside traditional brick-and-mortar settings. It differs from existing literature by moving beyond feasibility and effectiveness to provide an analysis of how medical engagement actively drives the use of specialized nutritional and behavioral interventions, which are often difficult services for patients to access in community settings. This work brings a new understanding to the field regarding the role of medical providers in integrated teams. Care effectiveness is amplified when medical management serves as a gateway to integrated nutritional and behavioral interventions. In the real world, health systems that develop virtual capabilities can greatly improve access to best-practice care for patients, no matter the supply of specialists in the region. Ultimately, virtual MGC is a responsive, patient-centered model that addresses the physical, nutritional, and psychological complexities of GI health.

Acknowledgments

We thank Ketan Patel for analytic support. The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. In accordance with the GAIDeT taxonomy (2025), GAI tools were used under full human supervision for code optimization, proofreading and editing, summarizing text, and reformatting. The GAI tool used was Gemini 3 Flash. Responsibility for the content and integrity of the final manuscript rests entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

Oshi Health, Inc is the study sponsor.

Data Availability

The dataset analyzed during this study is available from the corresponding author on reasonable request.

Authors' Contributions

GW and SB contributed to the conceptualization of the study. GW was responsible for data curation, formal analysis, and project administration, and also prepared the original draft of the manuscript. The methodology was developed by GW, WDC, SV, and SB. Supervision was provided by SB. GW, WDC, SV, and SB contributed to the review and editing of the manuscript.

Conflicts of Interest

GW and SB are employees of Oshi Health, Inc. and have equity interests in the company. SV is a paid consultant for Oshi Health, Inc.

Multimedia Appendix 1
Example dietary and behavioral health interventions offered through virtual MGC.
[PNG File , 140 KB - jmir_v28i1e89061_app1.png]

XSL•FO
RenderX

Multimedia Appendix 2
Andersen's model of health care utilization. This model shows the relationship between patient characteristics (i.e., predisposing, enabling, and need factors) and health behaviors, including service utilization. In turn, the health behaviors influence outcomes, such as symptom improvement and symptom control.
[PNG File , 76 KB - jmir_v28i1e89061_app2.png]

Multimedia Appendix 3
Number and percent of complete and imputed values, by characteristic, for the 11,345 virtual MGC patients who received care between April 2021 and August 2025.
[PDF File (Adobe PDF File), 106 KB - jmir_v28i1e89061_app3.pdf]

References

1. Almario CV, Ballal ML, Chey WD, Nordstrom C, Khanna D, Spiegel BMR. Burden of gastrointestinal symptoms in the United States: results of a nationally representative survey of over 71,000 Americans. Am J Gastroenterol 2018;113(11):1701-1710 [FREE Full text] [doi: 10.1038/s41395-018-0256-8] [Medline: 30323268]
2. Black CJ, Drossman DA, Talley NJ, Ruddy J, Ford AC. Functional gastrointestinal disorders: advances in understanding and management. Lancet 2020;396(10263):1664-1674 [FREE Full text] [doi: 10.1016/S0140-6736(20)32115-2] [Medline: 33049221]
3. Armuzzi A, Liguori G. Quality of life in patients with moderate to severe ulcerative colitis and the impact of treatment: a narrative review. Dig Liver Dis 2021;53(7):803-808 [FREE Full text] [doi: 10.1016/j.dld.2021.03.002] [Medline: 33744172]
4. Peery AF, Murphy CC, Anderson C, Jensen ET, Deutsch-Link S, Egberg MD, et al. Burden and cost of gastrointestinal, liver, and pancreatic diseases in the United States: update 2024. Gastroenterology 2025;168(5):1000-1024. [doi: 10.1053/j.gastro.2024.12.029] [Medline: 39920892]
5. Soni A. STATISTICAL BRIEF #539: healthcare expenditures for treatment of mental disorders: estimates for adults ages 18 and older, U.S. Agency for Healthcare Research and Quality (AHRQ). 2019. URL: https://meps.ahrq.gov/data_files/publications/st539/stat539.shtml [accessed 2025-10-15]
6. Moriya AS, Berdahl T. Healthcare expenditures for heart disease among adults aged 18 and older in the U.S. civilian noninstitutionalized population, 2022. Agency for Healthcare Research and Quality (AHRQ). 2022. URL: https://meps.ahrq.gov/data_files/publications/st562/stat562.shtml [accessed 2025-10-15]
7. Financial Burden of Cancer Care. National Cancer Institute Cancer Trends Progress Report. Bethesda, MD: National Cancer Institute; 2025. URL: https://progressreport.cancer.gov/after/economic_burden [accessed 2025-10-15]
8. Burgell RE, Hoey L, Norton K, Fitzpatrick J. Treating disorders of brain-gut interaction with multidisciplinary integrated care. moving towards a new standard of care. JGH Open 2024;8(5):e13072 [FREE Full text] [doi: 10.1002/jgh3.13072] [Medline: 38770352]
9. Basnayake C, Kamm MA, Stanley A, Wilson-O'Brien A, Burrell K, Lees-Trinca I, et al. Standard gastroenterologist versus multidisciplinary treatment for functional gastrointestinal disorders (MANTRA): an open-label, single-centre, randomised controlled trial. Lancet Gastroenterol Hepatol 2020;5(10):890-899. [doi: 10.1016/S2468-1253(20)30215-6] [Medline: 32679040]
10. Basnayake C, Kamm MA, Stanley A, Wilson-O'Brien A, Burrell K, Lees-Trinca I, et al. Long-term outcome of multidisciplinary versus standard gastroenterologist care for functional gastrointestinal disorders: a randomized trial. Clin Gastroenterol Hepatol 2022;20(9):2102-2111.e9. [doi: 10.1016/j.cgh.2021.12.005] [Medline: 34896645]
11. Bray NA, Koloski NA, Jones MP, Do A, Pang S, Coombes JS, et al. Evaluation of a multidisciplinary integrated treatment approach versus standard model of care for functional gastrointestinal disorders (FGIDS): a matched cohort study. Dig Dis Sci 2022;67(12):5593-5601 [FREE Full text] [doi: 10.1007/s10620-022-07464-1] [Medline: 35362835]
12. Lacy BE, Pimentel M, Brenner DM, Chey WD, Keefer LA, Long MD, et al. ACG clinical guideline: management of irritable bowel syndrome. Am J Gastroenterol 2021;116(1):17-44. [doi: 10.14309/ajg.0000000000001036] [Medline: 33315591]
13. Ying X, Yao L, Mathis WS, Congly SE, Jesudian AB. Geographic disparities in access to gastroenterologists in the United States. Gastroenterology 2025;168(6):1189-1191.e1. [doi: 10.1053/j.gastro.2025.01.232] [Medline: 39922544]
14. 2025 survey of physician appointment wait times and medicare and medicaid acceptance rates. AMN Healthcare. URL: https://www.amnhealthcare.com/amn-insights/physician/whitepapers/2025-survey-of-physician-appointment-wait-times/ [accessed 2025-08-19]
15. Salwen-Deremer JK, Bardach SH, Tormey LK, Szkodny LE, Gohres K, Siegel CA. Redesigning a gastroenterology behavioral health program to improve patient access. Clin Gastroenterol Hepatol 2024;22(1):12-15.e1 [FREE Full text] [doi: 10.1016/j.cgh.2023.09.032] [Medline: 37802274]
16. Nguyen AM, Williams AM, Marcello S. Integrated behavioral health works and saves money. why aren't we doing it? Milbank Memorial Fund. 2025. URL: https://www.milbank.org/2025/05/integrated-behavioral-health-works-and-saves-money-why-arent-we-doing-it/ [accessed 2026-03-03]

XSL•FO
RenderX

17. Gordon M, Sinopoulou V, Lakunina S, Gjuladin-Hellon T, Bracewell K, Akobeng AK. Remote care through telehealth for people with inflammatory bowel disease. Cochrane Database Syst Rev 2023;5(5):CD014821 [FREE Full text] [doi: 10.1002/14651858.CD014821.pub2] [Medline: 37140025]
18. Serper M, Volk ML. Updates in telemedicine for gastroenterology practices in the United States. Clin Gastroenterol Hepatol 2022;20(7):1432-1435 [FREE Full text] [doi: 10.1016/j.cgh.2022.03.024] [Medline: 35364315]
19. Quantitative research design (JARS-Quant). American Psychological Association. URL: https://apastyle.apa.org/jars/quantitative [accessed 2026-01-02]
20. ECDEU assessment manual for psychopharmacology. National Library of Medicine. 1976. URL: https://catalog.nlm.nih.gov/discovery/fulldisplay/alma994925623406676/01NLM_INST:01NLM_INST [accessed 2026-03-19]
21. Lim S, Trotier D, Karpov M, Han J, Lam K. Evaluation of the patient global impression of symptom severity (PGISS) score as a subjective and objective measure of disease activity in chronic rhinosinusitis. Am J Rhinol Allergy 2025;39(2):128-135. [doi: 10.1177/19458924241312318] [Medline: 39782306]
22. Snyder ES, Tao P, Svetnik V, Lines C, Herring WJ. Use of the single-item patient global impression-severity scale as a self-reported assessment of insomnia severity. J Sleep Res 2021;30(1):e13141. [doi: 10.1111/jsr.13141] [Medline: 33210445]
23. Mohebbi M, Dodd S, Dean O, Berk M. Patient centric measures for a patient centric era: agreement and convergent between ratings on The Patient Global Impression of Improvement (PGI-I) scale and the clinical global impressions - improvement (CGI-S) scale in bipolar and major depressive disorder. Eur Psychiatry 2018;53:17-22. [doi: 10.1016/j.eurpsy.2018.05.006] [Medline: 29859377]
24. Lewis JD, Vadhariya A, Su S, Zhou X, Durand F, Kawata AK, et al. A patient-reported outcome measure comprising the stool frequency and abdominal pain items from the Crohn's disease activity index: psychometric evaluation in adults with Crohn's disease. J Patient Rep Outcomes 2025;9(1):19 [FREE Full text] [doi: 10.1186/s41687-025-00851-y] [Medline: 39962027]
25. Dubinsky M, Vadhariya A, Su S, Zhou X, Durand F, Clucas C, et al. The urgency numeric rating scale: psychometric evaluation in adults with Crohn's disease. Adv Ther 2025;42(2):1044-1060. [doi: 10.1007/s12325-024-03081-8] [Medline: 39692838]
26. Delgado-Herrera L, Lasch K, Zeiher B, Lembo AJ, Drossman DA, Banderas B, et al. Evaluation and performance of a newly developed patient-reported outcome instrument for diarrhea-predominant irritable bowel syndrome in a clinical study population. Therap Adv Gastroenterol 2017;10(9):673-687 [FREE Full text] [doi: 10.1177/1756283X17726018] [Medline: 28932269]
27. Andersen RM. Revisiting the behavioral model and access to medical care: does it matter? J Health Soc Behav 1995;36(1):1. [doi: 10.2307/2137284]
28. Li C. Little's test of missing completely at random. Stata J 2013;13(4):795-809. [doi: 10.1177/1536867x1301300407]
29. Helsel BC, Williams JE, Lawson K, Liang J, Markowitz J. Telemedicine and mobile health technology are effective in the management of digestive diseases: a systematic review. Dig Dis Sci 2018;63(6):1392-1408. [doi: 10.1007/s10620-018-5054-z] [Medline: 29663265]
30. Pang L, Liu H, Liu Z, Tan J, Zhou LY, Qiu Y, et al. Role of telemedicine in inflammatory bowel disease: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2022;24(3):e28978 [FREE Full text] [doi: 10.2196/28978] [Medline: 35323120]
31. Dong MD, Steuwe S, Barry LA, Siegel CA. The use of telemedicine in older patients with gastrointestinal diseases. Curr Treat Options Gastroenterol 2022;20(4):594-604 [FREE Full text] [doi: 10.1007/s11938-022-00404-y] [Medline: 36465489]
32. National Cancer Institute. All HINTS questions. Health Information National Trends Survey (HINTS). 2024. URL: https://hints.cancer.gov/view-questions/all-hints-questions.aspx [accessed 2025-08-20]
33. Moshiree B, Drossman D, Shaukat A. AGA clinical practice update on evaluation and management of belching, abdominal bloating, and distention: expert review. Gastroenterology 2023;165(3):791-800.e3 [FREE Full text] [doi: 10.1053/j.gastro.2023.04.039] [Medline: 37452811]
34. Ludvigsson JF, Bai JC, Biagi F, Card TR, Ciacci C, Ciclitira PJ, BSG Coeliac Disease Guidelines Development Group, British Society of Gastroenterology. Diagnosis and management of adult coeliac disease: guidelines from the British society of gastroenterology. Gut 2014;63(8):1210-1228 [FREE Full text] [doi: 10.1136/gutjnl-2013-306578] [Medline: 24917550]
35. Chey WD, Hashash JG, Manning L, Chang L. AGA clinical practice update on the role of diet in irritable bowel syndrome: expert review. Gastroenterology 2022;162(6):1737-1745.e5. [doi: 10.1053/j.gastro.2021.12.248] [Medline: 35337654]
36. Kamp KJ, Morgan H, Yoo L, Munson SA, Heitkemper MM, Levy RL. Symptom management needs of patients with irritable bowel syndrome and concurrent anxiety and/or depression: a qualitative study. J Adv Nurs 2023;79(2):775-788 [FREE Full text] [doi: 10.1111/jan.15532] [Medline: 36511394]
37. Sultan N, Varney JE, Halmos EP, Biesiekierski JR, Yao CK, Muir JG, et al. How to implement the 3-phase FODMAP diet into gastroenterological practice. J Neurogastroenterol Motil 2022;28(3):343-356 [FREE Full text] [doi: 10.5056/jnm22035] [Medline: 35799231]
38. Ford AC, Sperber AD, Corsetti M, Camilleri M. Irritable bowel syndrome. Lancet 2020;396(10263):1675-1688 [FREE Full text] [doi: 10.1016/S0140-6736(20)31548-8] [Medline: 33049223]

XSL•FO
RenderX

39. Thakur ER, Khasawneh M, Moayyedi P, Black CJ, Ford AC. Efficacy of behavioural therapies for irritable bowel syndrome: a systematic review and network meta-analysis. Lancet Gastroenterol Hepatol 2025;10(12):1075-1088 [FREE Full text] [doi: 10.1016/S2468-1253(25)00238-9] [Medline: 41077057]
40. Keefer L, Palsson OS, Pandolfino JE. Best practice update: incorporating psychogastroenterology into management of digestive disorders. Gastroenterology 2018;154(5):1249-1257. [doi: 10.1053/j.gastro.2018.01.045] [Medline: 29410117]
41. Murray HB, Doerfler B, Harer KN, Keefer L. Psychological considerations in the dietary management of patients with DGBI. Am J Gastroenterol 2022;117(6):985-994 [FREE Full text] [doi: 10.14309/ajg.0000000000001766] [Medline: 35404301]
42. Gnanasakthy A, Barrett A, Norcross L, D'Alessio D, Romano C. Use of patient and investigator global impression scales: a review of food and drug administration-approved labeling, 2009 to 2019. Value Health 2021;24(7):1016-1023 [FREE Full text] [doi: 10.1016/j.jval.2021.01.005] [Medline: 34243825]
43. Khanna P, Agarwal N, Khanna D, Hays R, Chang L, Bolus R, et al. Development of an online library of patient-reported outcome measures in gastroenterology: the GI-PRO database. Am J Gastroenterol 2014;109(2):234-248 [FREE Full text] [doi: 10.1038/ajg.2013.401] [Medline: 24343547]
44. Spiegel BMR, Hays RD, Bolus R, Melmed GY, Chang L, Whitman C, et al. Development of the NIH patient-reported outcomes measurement information system (PROMIS) gastrointestinal symptom scales. Am J Gastroenterol 2014;109(11):1804-1814 [FREE Full text] [doi: 10.1038/ajg.2014.237] [Medline: 25199473]

Abbreviations

APP: advanced practice provider
DGBI: disorders of the gut-brain interaction
EHR: electronic health record
GI: gastrointestinal
GSEM: generalized structural equation modeling
HIPAA: Health Insurance Portability and Accountability Act
ICD-10: International Statistical Classification of Diseases, Tenth Revision
MGC: multidisciplinary gastrointestinal care
OR: odds ratio
PRO: patient-reported outcome
PROMIS: Patient-Reported Outcomes Measurement Information System

Edited by S Brini; submitted 05.Dec.2025; peer-reviewed by Z Tang, B Senst; comments to author 29.Dec.2025; revised version received 03.Mar.2026; accepted 05.Mar.2026; published 23.Apr.2026.

Please cite as:
Wang G, Chey WD, Varma S, Berry S
Virtual Multidisciplinary Gastrointestinal Care for Adults With Gastrointestinal Needs: Retrospective Cohort Study
J Med Internet Res 2026;28:e89061
URL: https://www.jmir.org/2026/1/e89061
doi:10.2196/89061
PMID:41842734

©Grace Wang, William D Chey, Sanskriti Varma, Sameer Berry. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Use of the Online Portal “Embryotox” in Routine Health Care: Mixed Methods Study

Marlies Onken[1]; Evelin Beck[1]; Martina Breuning[2]; Christine Holmberg[2]; Franziska Metke[1]; Anne Müller[1]; Christof Schaefer[1]; Katarina Dathe[1]

[1]Charité – University Medical Center Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Clinical Pharmacology and Toxicology, Embryotox Center of Clinical Teratology and Drug Safety in Pregnancy, Augustenburger Platz 1, 13353, Berlin, Germany

[2]Institute of Social Medicine and Epidemiology, Brandenburg Medical School Theodor Fontane, Brandenburg an der Havel, Germany

Corresponding Author:
Marlies Onken
Charité – University Medical Center Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Clinical Pharmacology and Toxicology, Embryotox Center of Clinical Teratology and Drug Safety in Pregnancy, Augustenburger Platz 1, 13353, Berlin, Germany

Abstract

Background: Since 2008, the open-access internet portal “Embryotox” has provided evidence-based information on drug safety during pregnancy and breastfeeding. The German website, designed for health care professionals, comprises fact sheets about 400 drugs and is maintained by the Embryotox Center of Clinical Teratology and Drug Safety in Pregnancy at the Charité – Universitätsmedizin Berlin. In 2024, it was accessed more than 5 million times.

Objective: This study aimed to evaluate who is using the website, how it is being used, and to what extent it meets users’ needs.

Methods: Data were collected between 2022 and 2023 through a sequential mixed methods design incorporating explanatory and exploratory elements. Online questionnaire 1 contained questions on user characteristics, clinical specifics of use, comprehensibility, and changes in risk assessment or drug therapy following fact sheet use. Qualitative data were generated through semistructured phone interviews, with interviewees sampled from the main user groups identified in questionnaire 1 (patients, physicians, pharmacists, and midwives). Kuckartz’s content analysis was used to analyze data on users’ trust and the functions of website use for different stakeholder groups. Online questionnaire 2 collected data from the main user groups on their decision-making based on Embryotox fact sheets. This questionnaire was developed based on the findings from semistructured interviews on the functions of website use. Answers in both questionnaires included multiple-choice options or ratings on a Likert scale from 0 (not at all) to 10 (fully); data were analyzed using descriptive statistics.

Results: Questionnaire 1 was completed by 14,562 users, including 10,860 patients, 1676 physicians, 550 pharmacists, and 364 midwives. Of the physicians, 27.2% (456/1676) were obstetricians and gynecologists, and 22.7% (381/1676) were general practitioners. For physicians, the mean of comprehensibility ratings was 9.39 (pharmacists 9.25; midwives 9.14), and for patients 8.80. Following fact sheet use, 66.6% (9694/14,562) of all users changed their risk perception, and 22.3% (2252/10,083) of users in a specific treatment setting considered changing drug treatment. Qualitative content analysis revealed that users highly trusted the information in the fact sheets for a number of reasons, including recommendations from other users, the scientific basis of the website, and the authority of the Embryotox Center as a university-based specialist department for drug safety in pregnancy and lactation. Functions of fact sheet use differed depending on the stakeholder group. A total of 793 users, mostly physicians and patients, fully completed questionnaire 2, specifying typical clinical situations and benefits of fact sheet use for shared decision-making. Benefits included facilitating the decision-making process and increasing confidence in it.

Conclusions: The information provided in the Embryotox fact sheets was generally perceived as both comprehensible and trustworthy, supporting and facilitating the decision-making process. Embryotox is used successfully by health care providers and patients in routine health care settings, helping to improve drug safety for pregnant and breastfeeding women.

(J Med Internet Res 2026;28:e81286) doi:10.2196/81286

KEYWORDS

pregnancy; breastfeeding; drug safety; risk assessment; consumer health information; online health information; shared decision-making

XSL•FO
RenderX

Introduction

Medical publications have multiplied in recent decades and continue to do so. This has created a constant challenge to ensure quality standards for routine health care in interdisciplinary areas such as drug safety in pregnant and breastfeeding women. In a French drug utilization study, about 90% (32,408/36,065) of pregnant women used prescription drugs [1] to treat preexisting or newly occurring maternal conditions. However, in this vulnerable stage, the use of medicines is fraught with considerable uncertainty. Inadequate knowledge often leads to an overestimation of drug risks, both by health care professionals (HCPs) and patients. This can have far-reaching consequences such as nonadherence leading to inadequately treated maternal conditions [2], premature discontinuation of breastfeeding, or even termination of an otherwise desired pregnancy. Conversely, underestimating the risks of medicines can lead to a failure to switch to a more appropriate treatment or to use effective contraception. Information on reproductive health is regularly searched for on the internet [3-7], and the dangers of misinformation from online health information are widely discussed [8,9]. Online information on perinatal drug safety is also often contradictory or misleading, especially when different sources are compared [10-12]. Comprehensive and reliable evidence-based information is therefore urgently needed [13-15]. Beginning in the 1980s, multidisciplinary, often publicly funded institutions have been established to provide information on drug safety in pregnancy and breastfeeding in many countries (see, eg, [16] or [17]). Initially, information was mainly offered via telephone counseling, but with the advent of mass internet use, some centers have also started to provide online information. Table S1 in Multimedia Appendix 1 provides examples of evidence-based online resources on drug safety in pregnancy and breastfeeding in various languages.

In Germany, the Berlin Embryotox Center was founded in 1988 [18], offering telephone counseling on the risk and safety of drugs in pregnancy and breastfeeding for HCPs and patients. The number of counseling requests has now settled at around 12,000 per year, a quarter of them by HCP. Also, the Embryotox Center collects and scientifically analyzes data on exposed pregnancies ("Embryotox cohort"), partly in collaboration with the European Network of Teratology Information Services [19,20], an umbrella organization established in 1990. A website with online information was launched in 2008 [21], cofunded by the German Federal Ministry of Health. This online resource was intended for use by HCPs, but has been publicly accessible from the outset. It now contains fact sheets on approximately 400 drugs, with each fact sheet providing the following information: (1) brief introduction to the drug, (2) a color code labeling (green, gray, and red) to summarize risk assessment, (3) description of scientific evidence for both first trimester and second and third trimester use, (4) clinical recommendations and consequences, including more adequate treatment alternatives if appropriate, and (5) scientific evidence and clinical recommendations for use during breastfeeding. Embryotox also offers fact sheets on the course and treatment of selected diseases during pregnancy and breastfeeding. Furthermore, the website provides information on the Embryotox Center, such as contact details for counseling, information on staff members, ongoing and past studies and scientific projects, including a list of scientific publications.

The rapidly increasing number of website users—growing from fewer than 200,000 visits in 2009 to more than 5.1 million in 2024—suggested early on that Embryotox was widely used not only by HCPs but also by patients. In order to tailor the information to the needs of its users in everyday medical practice, there was an urgent need to evaluate the use and perception of the website and to analyze how and by whom the fact sheets are used in routine health care. This mixed methods study was conducted to answer these questions.

Methods

Overview

This mixed methods study used a sequential design (quan-qual-quan), incorporating both explanatory and exploratory elements [22,23] to gain insight into the use of Embryotox in routine health care. Some of the quantitative findings from the first online questionnaire contributed to the development of the qualitative interview guides, while others, such as the demographic and occupational information, did not require in-depth qualitative exploration. The qualitative interviews, in turn, formed the basis for developing the second quantitative online questionnaire, which focused on a subtopic of Embryotox use, namely the use for clinical decision-making. Quantitative data from both online questionnaires are reported in accordance with the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) statement [24]; the completed checklist is included in Checklist 1. Reporting of qualitative data follows the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist [25]. The completed COREQ checklist can be found in Checklist 2.

Collection of Quantitative Data

Overview

The research team developed 2 online questionnaires and tested them among colleagues from both academic and nonacademic backgrounds, as well as among individuals from their social circles. Technical details were worked out in collaboration with the digital agency that technically manages the website and also programmed both questionnaires. Visits to Embryotox drug fact sheets were counted via an anonymous session-tracking procedure by a commercial provider. To identify potential multiple entries, datasets of questionnaires were analyzed regarding entry times and similarity of entries. IP addresses or cookies were not used for this purpose. In accordance with the General Data Protection Regulation in Germany, questionnaires were designed in such a way that no conclusions could be drawn about individuals. Banner ads leading to the questionnaires were based directly on the drug fact sheets. The banner leading to the first questionnaire (2022) invited users to participate with the words "We ask for your feedback." On the banner leading to the second questionnaire (2023), users were asked, "How do you use embryotox.de? Questionnaire about your experiences." Participation was voluntary and not a prerequisite for using the drug fact sheets. According to their answers, participants were

XSL•FO
RenderX

automatically guided through the questionnaire. Clicking back was not possible. Progress within the questionnaire was visible via a bar. Only one question at a time displayed on the screen, and participants could move on to the next question only after answering the previous question. Only the question about the educational qualifications of lay users was optional and could be skipped. The selection of mutually exclusive answer options and the selection of more than one answer (if not explicitly requested) were blocked. The answers were only saved if the users completed the questionnaire to the end and then selected the button "Send data."

Online Questionnaire 1: Use and Users of Embryotox in Routine Health Care

Online questionnaire 1 was placed on all drug fact sheets for 7 months (May-December 2022). Over the same period, the total number of visits to each fact sheet was counted. Users were asked to complete the questionnaire only once. Questionnaire 1 included questions on user characteristics, in particular demographic information (age and gender) and, for HCP, occupational information such as profession (for physicians also specialization) and working area (outpatient care, hospital, pharmacy, and other). Patients and other non-HCPs were asked about their level of education and how they got to know about Embryotox. HCPs were asked whether they were looking for information for a specific patient and whether they had found the essential information they needed to make a well-founded treatment decision. All users were asked to specify the frequency and duration of Embryotox use and the clinical focus of their search. Furthermore, all users were asked to evaluate the comprehensibility of the fact sheet they had accessed and whether its information resulted in changes in risk perception or preferred medication. Possible answers comprised either multiple-choice options or ratings on a Likert scale ranging from 0 (not at all) to 10 (fully). For a detailed flow chart of questionnaire 1, refer to Multimedia Appendix 2.

In addition, weekday, daytime, and the accessed fact sheet were recorded.

Online Questionnaire 2: Fact Sheet Use for Decision-Making

Questionnaire 2 was placed for 7 months (February-September 2023) on fact sheets for prescription-only drugs that typically require shared decision-making (refer to Table S3 in Multimedia Appendix 1). The questionnaire was aimed at the main user groups identified in questionnaire 1, primarily at physicians and patients as the main decision-makers. For the detailed flowchart, refer to Multimedia Appendix 3. The questionnaire was developed based on the findings from semistructured interviews on the functions of website use. For all user groups, answers comprised multiple-choice options or ratings on a Likert scale ranging from 0 to 10 or from "strongly agree" via "moderately agree" and "moderately disagree" to "strongly disagree." If appropriate, "question not applicable" was also an answer option.

Questions and answer options were tailored to the respective user groups. Physicians answered up to 15 questions focusing on (1) their medical specialization, (2) typical use of fact sheets (eg, with or without patient, during or outside of consultation; maximum 3 out of 7 possible answers), (3) fact sheet use in the process of risk-benefit assessment and decision-making (eg, risk assessment and decision with or without patient, and satisfaction with fact sheet use), (4) further effects and functions of fact sheet use (eg, perceived effects on patient adherence and confidence in decision-making), and (5) potentially useful supplementary materials to the information provided on the website (maximum 2 out of 7 possible answers). Patients answered up to 16 questions about (1) fact sheet use for risk-benefit assessment and decision-making, (2) potential benefit or harm from the use of fact sheets, (3) effects and functions of fact sheet use, (4) potentially missing information, and (5) involvement of other persons in the decision-making process (multiple answers possible). Pharmacists and midwives answered up to 6 questions about the clinical situation, their typical fact sheet use (maximum 3 out of 5 possible answers) and functions of fact sheet use.

Analysis of Quantitative Data

Descriptive statistics were used to analyze the quantitative data from both online questionnaires. The statistical analyses were conducted using R (R Core Team; version 4.4.0) [26]. To obtain additional information on fact sheet use, the weekday and time of day that each questionnaire 1 was completed were also analyzed. Additionally, the total number of visits to each fact sheet while questionnaire 1 was available online was used to calculate the proportion of completed questionnaire 1 in relation to the total number of fact sheet uses.

Collection of Qualitative Data

Qualitative data on the use of Embryotox were obtained through semistructured phone interviews with individuals from the main user groups identified in questionnaire 1, that is, patients, physicians, pharmacists, and midwives. The interview guides were developed based on the aims of the study and the results from questionnaire 1. They aimed to provide insight into how Embryotox is used in various clinical settings and by different stakeholders within routine health care. Interview guides were piloted with each user group, after which the research team discussed and agreed upon any necessary adjustments by consensus. The guides focused on the functions of fact sheet use, as well as on the usability of Embryotox and possible needs for optimization. All interviewees were asked whether they trusted the information on Embryotox, and if they answered in the affirmative, to explain their reasons. Selected topics from questionnaire 1, such as comprehensibility, were also included for further exploration. For an overview, refer to Table S2 in Multimedia Appendix 1. Interviewees were recruited by purposive and snowball sampling. The aim of the sampling was to find interview participants who frequently use Embryotox and to achieve broad representation of the users' experiences. All interviews were carried out by the same interviewer (AM, female). Data saturation was assessed with each new interview within a user group, as this has been suggested as a method of determining the sample size for qualitative interviews [27]. Data saturation was achieved for a user group when few or no new topics had emerged in the last 3 interviews. The interviews lasted about 30 minutes on average.

XSL•FO
RenderX

Analysis of Qualitative Data

Kuckartz's content analysis with a deductive-inductive approach was used to analyze qualitative data [28]. The f4 software (dr. dresing & pehl GmbH) [29] was used for transcription, coding, and analysis of the interviews. First deductive categories were derived from the interview guide and the research question and discussed in the research team. In addition to the deductive categories, further categories and subcategories were determined inductively from the material. The coding tree and themes were developed based on the first 3 interviews; they were double-checked and discussed by the research team to ensure intersubjective comprehensibility. Any disagreements were resolved through discussion within the research team [30]. Subsequent interviews were coded and analyzed using this coding structure in the f4analyse software (dr. dresing & pehl GmbH) [29]. For the following evaluation, we analyzed user statements on (1) trust in fact sheet information and (2) functions of fact sheet use, including the purpose and clinical utility of typical use.

Ethical Considerations

Ethical approval for the collection and analysis of qualitative and quantitative data was obtained from the Ethical Committee of the Charité - Universitätsmedizin Berlin (EA1/334/21). For participating in the online survey, no formal written consent was required. Participants in the qualitative phone interviews signed a consent form following verbal and written information. Participants were not compensated for completing the online questionnaires or taking part in the interviews. The online questionnaires did not request personal data that would allow for identification of the participants. Identifying personal data of the interviewees was removed during the transcription process.

Results

Overview

A total of 14,562 completed questionnaire 1 (use of Embryotox in routine health care) and 793 completed questionnaire 2 (fact sheet use for decision-making) were analyzed using descriptive statistics. A total of 41 semistructured telephone interviews formed the dataset for qualitative content analysis (reasons to trust and to use Embryotox).

Use of Embryotox in Routine Health Care (Questionnaire 1)

Overview

Of the 14,562 completed questionnaire 1, 10,860 were filled by patients and 1676 by physicians, the latter mostly obstetricians and gynecologists (27.2%, 456/1676) or general practitioners (22.7%, 381/1676). Details of participating user groups are presented in Table 1.

Overall, a median of 0.58% (IQR 0.42% - 0.77%) of visits to a drug fact sheet resulted in the completion of a questionnaire. The top 10 most frequently reviewed fact sheets are presented in Figures 1 and 2 for physicians and patients, respectively.

Table . User groups contributing to questionnaire 1.

Completed questionnaires by user group (N=14,562)	Value, n (%)
Patients	10,860 (74.6)
Physicians	1676 (11.5)
Medical specialties:	
Obstetrics and gynecology	456 (27.2)
General practitioner	381 (22.7)
Internal medicine	214 (12.8)
Psychiatry, psychotherapy, or psychosomatics	129 (7.7)
Neurology	43 (2.6)
Pediatrics	93 (5.5)
Human genetics	3 (0.2)
Dentistry	17 (1.0)
Other medical specialties	340 (20.3)
Pharmacists	550 (3.8)
Midwives	364 (2.5)
Other health care professionals (eg, nurse, medical, or dental assistant)	324 (2.2)
Partner, relative, or friend of a patient	379 (2.6)
Other context	292 (2)
Student of medicine or other health sciences	117 (0.8)

XSL•FO
RenderX

Figure 1. Top 10 drug fact sheets most frequently accessed by physicians (n=1676).

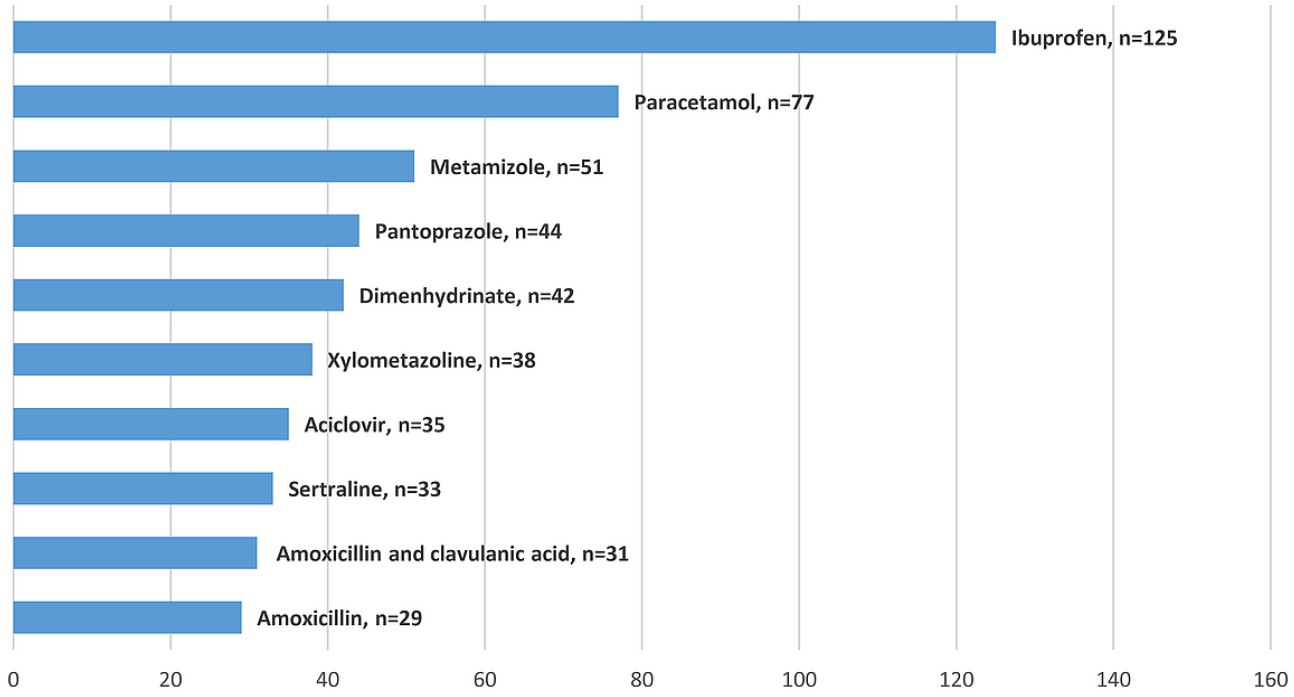

Figure 2. Top 10 drug fact sheets most frequently accessed by patients (n=10,860).

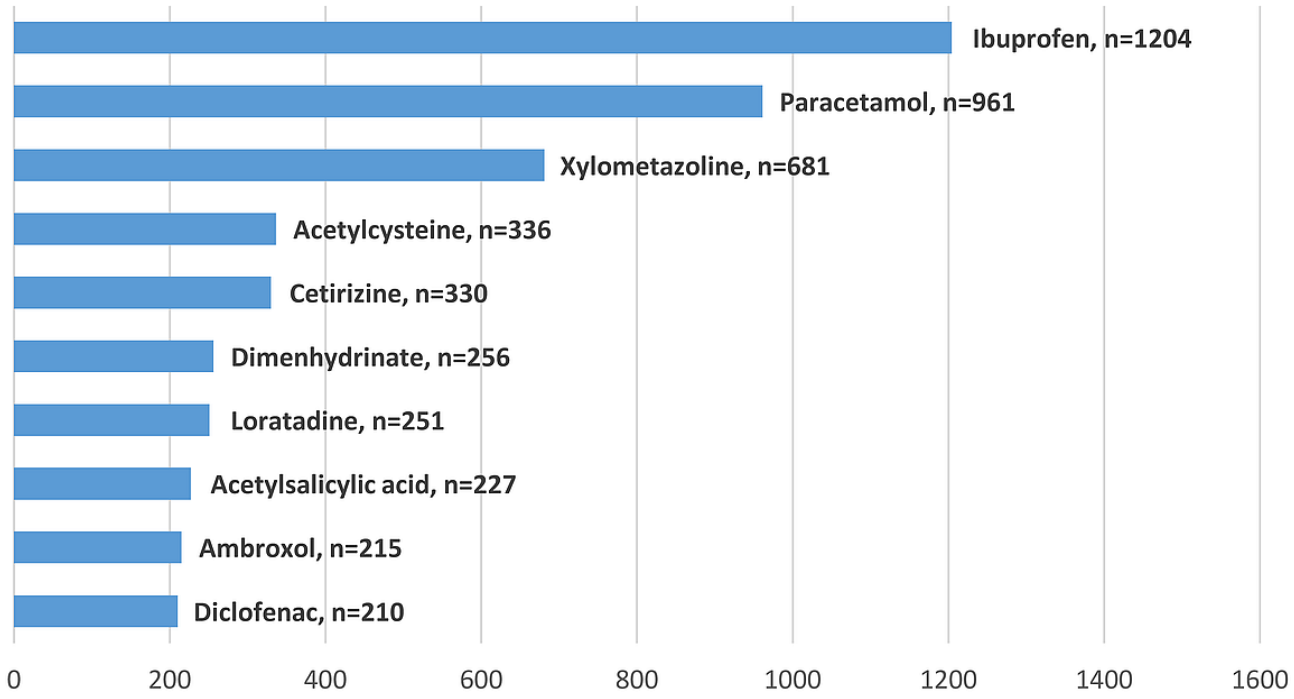

User Characteristics

For user characteristics of HCP, refer to Table 2. The majority of participating patients were aged between 31 and 40 years (69.3%, 7525/10,860), most of them had a high level of education (university degree 58.2%, 6319/10,860; qualification for higher education 20.7%, 2247/10,860; school-leaving certificate 10th or 11th grade 17.4%, 1887/10,860; school-leaving certificate 9th grade 2.3%, 247/10,860; no educational qualification or not specified: 1.5%, 160/10,860). Figure 3 shows how patients became aware of Embryotox.

XSL•FO
RenderX

Table . User characteristics of health care professionals (questionnaire 1).

Participants' characteristics	Physicians (n=1676)	Pharmacists (n=550)	Midwives (n=364)	Other HCP[a] (n=324)
Age (years), n (%)				
≤30	140 (8.4)	153 (27.8)	62 (17)	122 (37.7)
31 - 40	673 (40.2)	229 (41.6)	101 (27.7)	134 (41.4)
41 - 50	415 (24.8)	75 (13.6)	118 (32.4)	45 (13.9)
51 - 60	298 (17.8)	63 (11.5)	72 (19.8)	17 (5.2)
>60	150 (8.9)	30 (5.5)	11 (3)	6 (1.9)
Gender, n (%)				
Female	1093 (65.2)	468 (85.1)	361 (99.2)	284 (87.7)
Male	575 (34.3)	82 (14.9)	1 (0.3)	38 (11.7)
Diverse	8 (0.5)	0 (0)	2 (0.5)	2 (0.6)
Working area, n (%)				
Hospital	725 (43.3)	30 (5.5)	116 (31.9)	163 (50.3)
Outpatient care	893 (53.3)	0 (0)	173 (47.5)	87 (26.9)
Pharmacy	3 (0.2)	480 (87.3)	0 (0)	9 (2.8)
Other	55 (3.3)	40 (7.3)	75 (20.6)	65 (20.1)

[a]HCP: health care professional.

Figure 3. How patients (n=10,860) became aware of Embryotox.

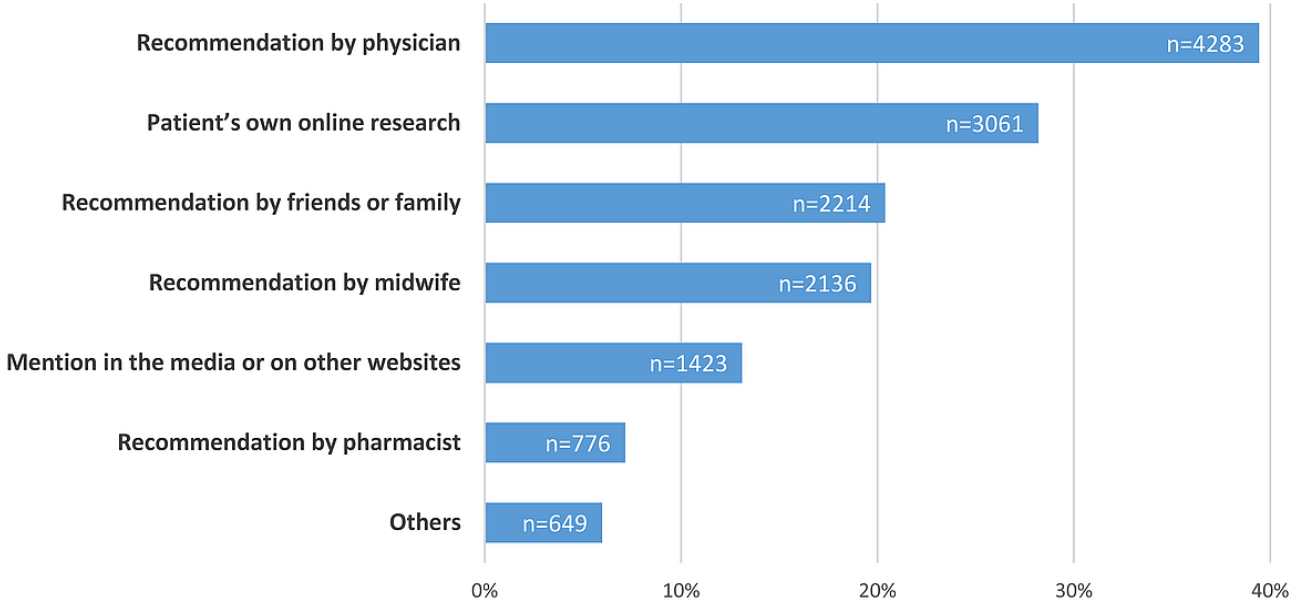

Usage of Fact Sheets

Drug fact sheets were accessed to a considerable extent outside standard working hours, that is, between 8 PM and 7:59 AM (physicians: 30.4%, 509/1676; patients: 42.9%, 4655/10,860; pharmacists: 28.4%, 156/550; midwives: 29.4%, 107/364). Most of the participating HCP stated that they had used the same fact sheet more than once (physicians: 73.3%, 1229/1676; pharmacists: 69.1%, 380/550; midwives: 77.5%, 282/364), as did the majority of patients (71%, 7708/10,860). At the time of fact sheet access, most of the participating patients were pregnant (46.8%, 5085/10,860) or were breastfeeding (42.8%, 4650/10,860). A total of 6% (649/10,860) of patients and 9.4% (131/1388) of physicians seeking information for a specific patient accessed Embryotox when pregnancy was still being planned. Among psychiatrists, the proportion was even higher (25.2%, 27/107). Details about fact sheet use by different user groups are presented in Table S4 in Multimedia Appendix 1.

Comprehensibility and Relevance

For all user groups combined (N=14,562), the mean of comprehensibility ratings was 8.88 (Likert scale from 0 to 10). For all HCPs, the mean of comprehensibility ratings was higher than 9 (physicians: 9.39, pharmacists: 9.25, midwives: 9.14, and other HCP: 9.01). For patients, the mean was 8.80, and for other lay persons (partners, relatives, and friends), it was 8.83.

XSL•FO
RenderX

Patients' assessment of comprehensibility depended on their level of education. The mean comprehensibility ratings was 7.29 for patients without a school-leaving certificate, while it was 8.92 for patients with an academic degree.

HCPs were asked to indicate if they had found the information required to make a well-founded treatment decision. This was answered in the affirmative ("Yes, definitely" and "Somewhat") by 94.1% (1577/1676) of physicians, 92.7% (510/550) of pharmacists, and 92.6% (337/364) of midwives, with the majority of users fully agreeing ("Yes, definitely": physicians 72.1%, 1208/1676; pharmacists 65.3%, 359/550; and midwives 70.3%, 256/364).

Changes in Risk Assessment and Implications for Drug Therapy

A total of 66.6% (9694/14,562) of all users stated that they had changed their risk perception of a particular drug during pregnancy or breastfeeding after reading a fact sheet. Risk assessment remained unchanged more often in HCPs than in patients (physicians: 46.9%, 786/1676; pharmacists: 42.9%, 236/550; midwives: 51.9%, 189/364; and patients: 29.9%, 3252/10,860). Compared to their perception before reading the fact sheet, 16.2% (2358/14,562) of all users estimated the risk to be higher and 50.4% (7336/14,562) to be lower. The change of risk perception was also linked to the safety profile of the drug: After reading fact sheets with a red label on Embryotox, that is, fact sheets for drugs with known teratogenicity and/or severe fetotoxicity, 43.1% (81/188) of users estimated the risk to be higher (no change in risk assessment: 31.4%, 59/188). After reading fact sheets with a green label, that is, for drugs of choice for specified treatment indications, 58% (3642/6279) rated the risk to be lower (no change in risk assessment: 33%, 2070/6279). As a result of fact sheet information, 22.3% of users to whom the question applied (2252/10,083) considered changing drug treatment ("definitely yes": 8.5%, 862/10,083; "possibly yes": 13.8%, 1390/10,083). For physicians, the proportion was 27.9% (355/1274; with "definitely yes" 11.4%, 145/1274 and "possibly yes" 16.5%, 210/1274).

Reasons to Trust and to Use Embryotox (Qualitative Interviews)

Overview

A total of 41 users were interviewed from February to November 2022, including 14 physicians and nine pharmacists, midwives, and patients each. Physicians worked in various specialized areas (gynecology, pediatrics, psychiatry, dentistry, and general medicine) and settings (hospitals and outpatient care). Ten physicians were female and 4 male, with a mean age of 41.8 (range 26-69) years. Most of the pharmacists worked in public pharmacies (8/9). Six were female and 3 were male, with a mean age of 37.8 (range 28-56) years. The majority of midwives worked freelance (8/9). All were female, with a mean age of 44.2 (range 34-56) years. Of the patients, 6 were pregnant, 2 were breastfeeding, and 1 patient was planning a pregnancy. Mean age was 34.6 (range 24-42) years.

Trust

All 41 interviewees stated to trust the information on Embryotox drug fact sheets. Interviewees from all user groups considered the association of Embryotox with a trustworthy institution to be a reason for trust, with the Charité – Universitätsmedizin Berlin and the Federal Ministry of Health mentioned most frequently. Users explicitly stated that they placed trust in the scientific basis of the website and perceived that drug fact sheets were developed by experts in the field. A physician summarized:

> *It says 'Federal Ministry of Health,' 'Charité university medicine' … I mean, we are also a university hospital. So I think, university hospitals always work on a scientific basis and in this respect, these are all factors that constitute a certain independence and a certain scientific orientation, and that is what leads to trust.* [Senior physician in obstetrics and gynecology, female, 41 years]

And a pharmacist stated:

> *... that the website is somehow a well-known site and Charité Berlin is on it, that's usually something you can rely on … they know how to read a study and how to interpret it.* [Employee at a community pharmacy, female, 31 years]

In addition, the differentiated description of complex issues was named as a reason for trust, as a patient explained:

> *... that the texts don't say you should definitely take a great amount or not at all, but that they try to provide a differentiated picture ... .* [Patient, breastfeeding, 33 years]

Interviewees from all user groups trusted the Embryotox Center as the content provider of the website, which they explained by citing positive experiences with other Embryotox activities such as telephone counseling, lectures, and research projects. A midwife explained:

> *So I think the phone contacts I've had so far, through these phone contacts, because there always... So, they went over it again, discussed it, explained it and that gives me a good feeling that I can trust the whole thing.* [Freelancer, female, 38 years]

Furthermore, the absence of commercial interests and the long-standing experience of the Embryotox Center were given as reasons. Patients and pharmacists also emphasized that the recommendation of the website by other users, especially by physicians, built trust in the presented information. One patient said:

> *So my midwife recommended it and my doctor recommended it and then, yes, I would trust the site ..., everything that is written there.* [Patient, pregnant, 33 years]

Physicians said to trust the website because they had corroborated the information elsewhere, and also because of its clear structure and layout. One of them stated:

> *The website also comes across as very serious, I would say. It's not just some colorful [unintelligible]*

XSL•FO
RenderX

bouncing through the picture or something, but it's ..., it strives to be attractive, but also serious.

Further details on users' assessment of trust can be found in Textbox 1.

Textbox 1. Why do users trust the information on Embryotox? Reasons reported by users (based on 41 interviews, including 14 with physicians and 9 each with pharmacists, midwives, and patients).

Association with trustworthy institutions (interviewees of all user groups)

- Embryotox Center, as the operator of the website, is part of Charité University Hospital (n=26)
- Funding by the German Ministry of Health (n=7)
- Perception of Embryotox as an "official" website (n=3)
- Website certified or funded by trustworthy organizations (n=2), including the "Health On the Net Foundation certificate (HONCode)'", which was still available until December 2022, and the support from one of the major statutory health insurers in Germany at that time.

Trust in Embryotox Center as operator of the website (interviewees of all user groups)

- Drug fact sheets and recommendations developed by experts after detailed research (n=13)
- Long-standing experience and established position in the professional community (n=5)
- Positive experiences with telephone counseling on drug safety by the Embryotox Center (n=5)
- No commercial interests (n=7)
- Research projects conducted by Embryotox Center (n=4)
- No other similarly useful sources of information known (n=3)
- Participation in lectures by Embryotox experts (n=1)
- Absence of negative reports about Embryotox Center (n=1)

Recommendation by other users (pharmacists and patients)

- Use and recommendation by physicians (n=5)
- Use and recommendation by others, including nonphysician health care professionals (n=7)

Scientific basis (physicians, pharmacists, and patients)

- Drug fact sheets based on current scientific literature (n=14)

Website content and layout (physicians and patients)

- Content corroborated by clinical practice and own research (n=3)
- Differentiated description of complex issues (n=3)
- Clear structure and layout (n=2)
- Photos of staff on website (n=1)

Typical Use and Functions of Embryotox

Functions and purpose of typical fact sheet use considering different stakeholder groups are summarized in Textbox 2.

XSL•FO
RenderX

Textbox 2. Typical use and functions of Embryotox for health care professionals and patients.

Health care professionals (based on interviews with 9 midwives, 9 pharmacists, and 14 physicians)

- Researching information about drug treatment (13 physicians, 4 midwives, and 4 pharmacists)
- Corroborating information about drug treatment received earlier or elsewhere (8 physicians, 5 midwives, and 7 pharmacists)
- Informing, educating, and reassuring patients (9 physicians, 9 midwives, and 6 pharmacists)
- Pursuing an individual interest or professional development (4 physicians, 1 midwife, and 1 pharmacist)
- Documenting, communicating, or discussing patients' drug treatment with colleagues on the basis of fact sheets (6 physicians)
- Challenging information a patient has received from a physician or pharmacist, or enabling the patient to challenge it (7 midwives)

Patients (based on interviews with 9 patients)

- Researching information about drug treatment (6 patients)
- Corroborating information about drug treatment received earlier or elsewhere (9 patients)
- Backup for discussing drug treatment with the attending physician or other people (4 patients)
- To reassure oneself (2 patients)

Fact Sheet Use for Decision-Making (Questionnaire 2)

Overview

Of 1093 questionnaires, 793 were completed in full and 300 ended after 2 or 3 questions, because the decision-making process was still ongoing (214 patients) or fact sheets had not been used in the process of decision-making (23 physicians and 63 patients). The fully completed questionnaires (n=793) provided the data basis for further analyses; contributing user groups are shown in Table 3.

Table . Contributing users with fully completed questionnaire 2.

Completed questionnaires (N=793) by user group	Value, n (%)
Physicians	382 (48.2)
Medical specialties:	
Obstetrics and gynecology	96 (25.1)
General practitioner	78 (20.4)
Psychiatry, psychotherapy, or psychosomatics	67 (17.5)
Internal medicine	36 (9.4)
Pediatrics	25 (6.5)
Neurology	11 (2.9)
Other medical specialties	69 (18.1)
Patients	267 (33.7)
Pharmacists	79 (10)
Midwives	65 (8.2)

Typical Fact Sheet Use by HCP

To characterize their typical use of fact sheets, HCP could select a maximum of 3 out of 7 possible answers: 56.3% (215/382) of physicians used fact sheets typically before or after patient consultations, 76.3% (309/382) exclusively or additionally during the consultation. A total of 30.1% (93/309) of the latter routinely used fact sheets online together with their patients. A total of 46.9% (179/382) of physicians encouraged patients to inform themselves on Embryotox (gynecologists: 59.4%, 57/96) and 9.2% (35/382) printed out fact sheets for them (psychiatrists: 20.9%, 14/67). For further details, refer to Table S5 in Multimedia Appendix 1. Pharmacists and midwives used fact sheets to pass on selected information to patients (79.7%, 63/79 and 66.2%, 43/65, respectively), read the fact sheets together with patients (20.3%, 16/79 and 30.8%, 20/65), and encouraged patients to inform themselves on Embryotox (24.1%, 19/79 and 50.8%, 33/65). Fact sheets were also used to check the medication prescribed by physicians (pharmacists: 57%, 45/79; midwives: 52.3%, 34/65) and to consult with the prescribing physician if a medication was considered problematic (pharmacists: 58.2%, 46/79; midwives: 36.9%, 24/65).

Figure 4 shows the extent to which physicians and patients considered themselves involved in making the final decision about drug therapy during pregnancy and breastfeeding after fact sheets have been used to obtain information for the decision-making process.

XSL•FO
RenderX

Figure 4. Physicians' and patients' assessment of who made the decision on drug therapy after using a fact sheet.

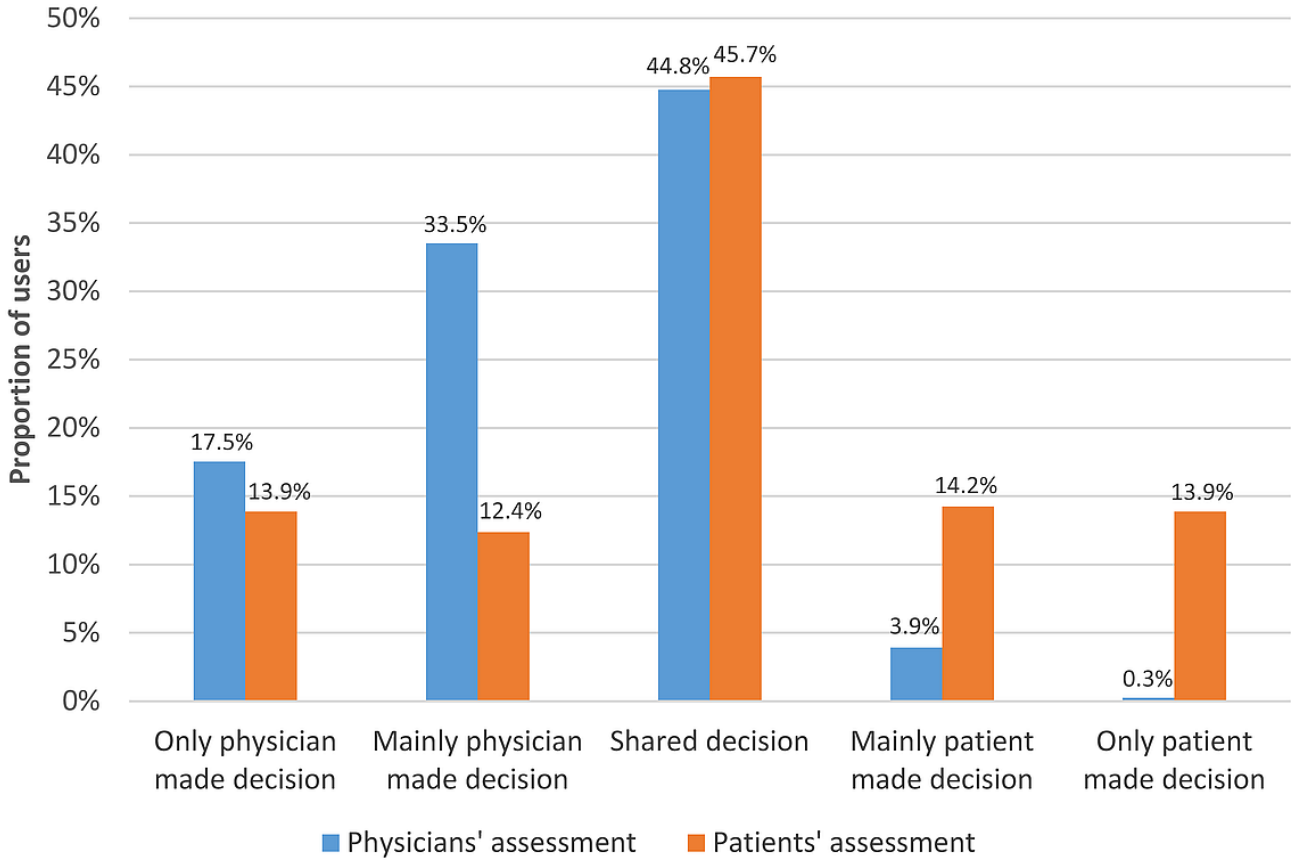

Benefits of Fact Sheet Use

The majority of physicians and patients agreed to benefit from fact sheet use. Details are given in Figures 5 and 6.

Patients also indicated that Embryotox drug fact sheets had helped them to convince their physician that a specific medication was compatible with pregnancy or breastfeeding (52.1%, 139/267) or that alternatives had to be considered (33%, 88/267). A total of 98.6% of pharmacists (71/72) agreed "moderately" or "strongly" to use Embryotox drug fact sheets for clarification in cases of disagreement between patients and physicians regarding drug therapy (question not applicable; n.a.: 7/79); the proportion for midwives was similar (95.2%, 59/62; n.a.: 3/65). Both user groups also indicated to use Embryotox drug fact sheets for informing and, if necessary, reassuring worried patients (pharmacists: 98.7%, 74/75; n.a.: 4/79; midwives: 100%, 62/62; n.a.: 3/65).

XSL•FO
RenderX

Figure 5. Benefits of the Embryotox fact sheet use for physicians. Percentages refer to all physicians except those who answered "question not applicable", that is, n=382 for all items except “Joint basis for decision making” (n=365) and “Reassurance” (n=369).

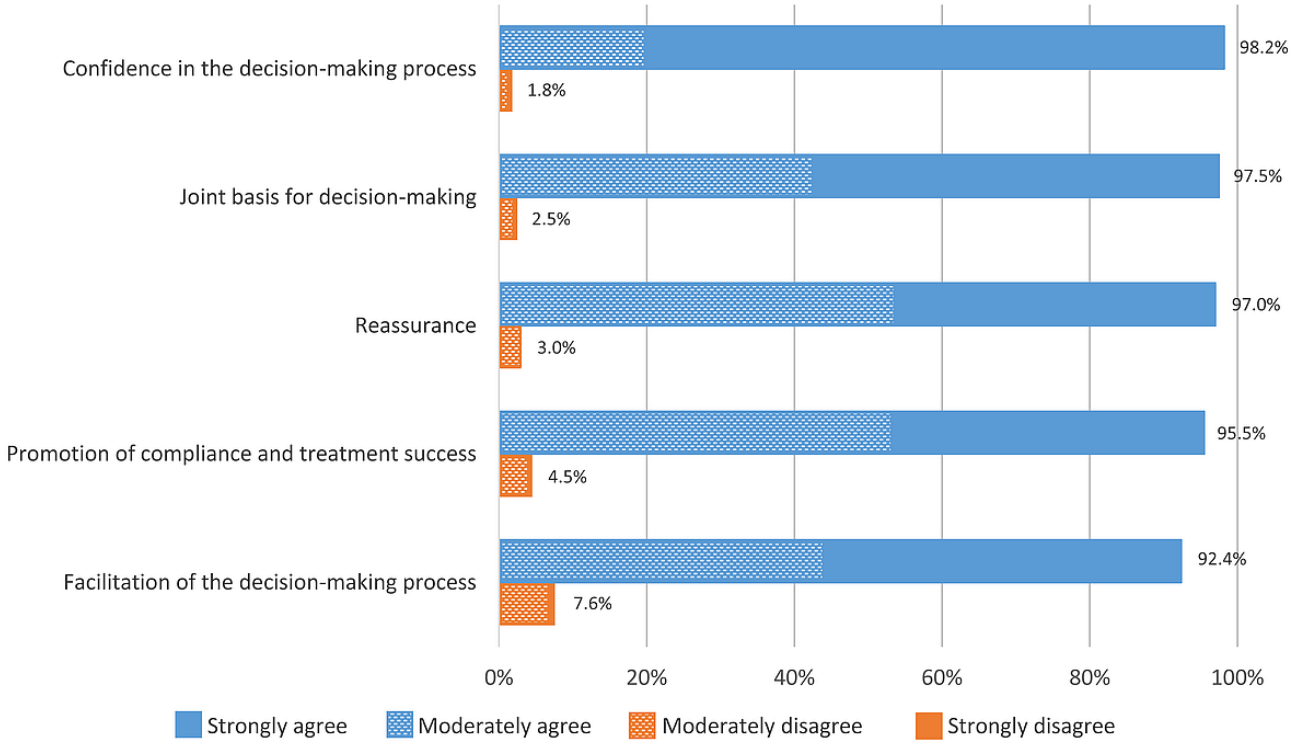

Figure 6. Benefits of the Embryotox fact sheet use for patients. Percentages refer to all patients except those who answered "question not applicable", that is, n=267 for all items except “Facilitation of the decision-making process” (n=236) and “Helpful to trust the decision and stand by it” (n=214).

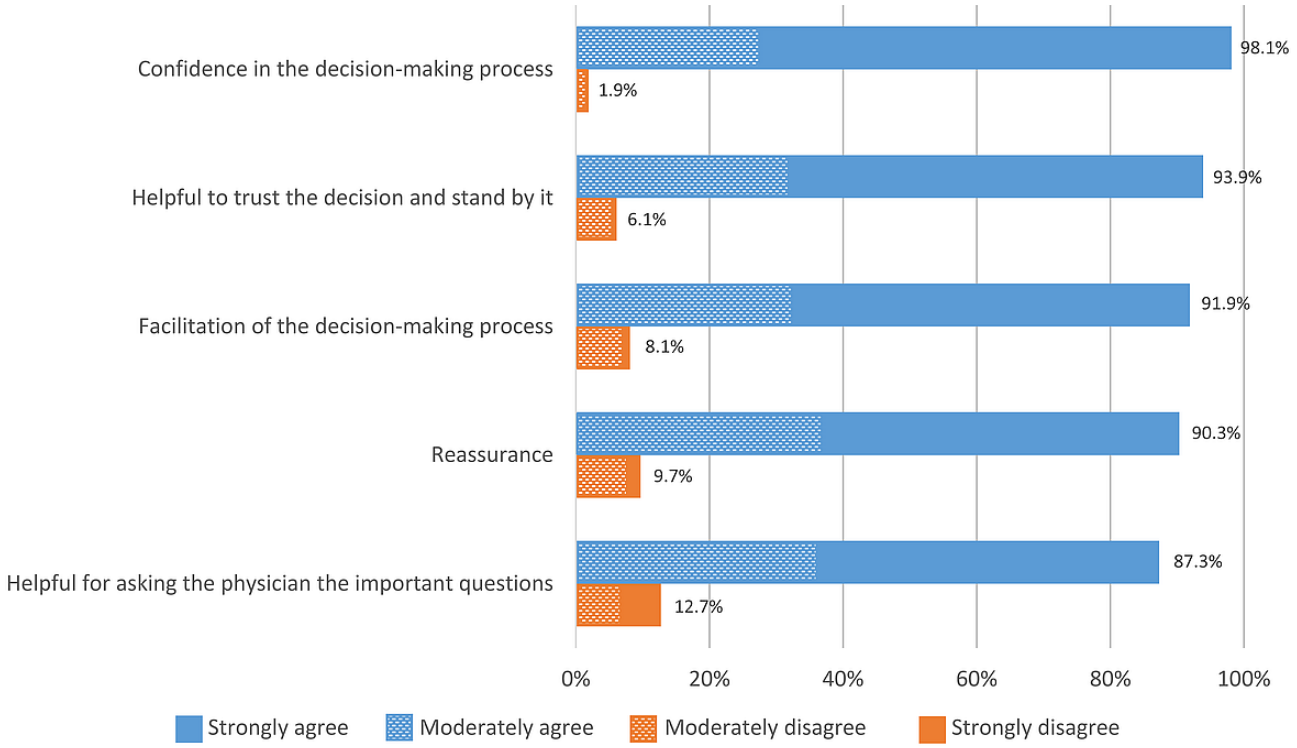

Physicians’ Suggestions for Additional Online Content

A total of 53.4% (204/382) of physicians wished for more drug fact sheets on the website (gynecologists: 77.1%, 74/96). Furthermore, 42.9% (164/382) of physicians, particularly internists (55.6%, 20/36) and psychiatrists (47.8%, 32/67), considered “decision aids” to be useful, that is, structured guides through the decision-making process. Internet links to other reliable sources were requested by 19.6% (75/382), references to scientific literature by 14.1% (54/382), increased use of visualizations by 6.5% (25/382), and further explanations on given content by 4.7% (18/382). A total of 16.8% (64/382) did not consider any additions to be necessary.

XSL•FO
RenderX

Discussion

Principal Findings

For pregnant and breastfeeding women, as well as for their HCP, guidance on drug safety is among the most important health-related information needs [14,15,31-33]. To address this demand, publicly accessible, curated online knowledge resources on this topic have been established in several countries (see also Table S1 in Multimedia Appendix 1). To our knowledge, no comprehensive study has yet been conducted to establish how and by whom these databases are used, and whether they meet existing needs. The French resource [34] was evaluated in terms of appearance, content, interactivity, ease of use, and technical performance by 750 selected HCPs (including physicians, pharmacists, and midwives), who reported high levels of satisfaction [35,36]. However, the implementation of this and other publicly available resources in routine health care, and their use by patients and/or in shared decision-making has not yet been assessed. Thus, around 2 decades after the first online resources on drug safety during pregnancy and breastfeeding became available, our mixed methods study is the first to provide an in-depth, user group–specific evaluation of such a resource in routine health care.

The German website Embryotox was considered an established online resource for information on drug safety during pregnancy and breastfeeding. It was widely used 24/7 and viewed as trustworthy by HCPs and patients. This trust was largely based on the authority of the website operator and its affiliations—namely, the Embryotox Center as a specialist department of a university hospital—as well as on the scientific basis of the information. Both of these aspects align with the findings of existing research into how users seek and evaluate online health information [37-39]. Also consistent with previous studies [39,40], patients in our evaluation indicated that recommendations from their physicians and other HCPs to use the Embryotox website helped to build trust in the information. Given that trust is considered a key prerequisite for the patients' intention to act on advice [41], the high level of trust in the information provided by Embryotox is vital to the website's preventive impact.

The main user groups of the website were identified as patients, physicians, pharmacists, and midwives, accounting for 74.6%, 11.5%, 3.8%, and 2.5% of the 14,562 completed questionnaire 1 (Table 1). More than 70% (8937/12,536) of participating physicians and patients were repeat visitors to the fact sheets they accessed, suggesting a targeted search for updates and/or the need for reassurance. A relevant number of participating physicians searched for information before patients became pregnant, possibly to check whether a planned long-term medication would be compatible with pregnancy. This is particularly important given that around a third of all pregnancies in Germany are unplanned [42]. The high proportion of general practitioners as well as obstetricians and gynecologists among participating physicians (22.7% and 27.2%, respectively, in questionnaire 1) suggests that Embryotox is widely used in the primary health care of pregnant and breastfeeding women, thus promoting state-of-the-art drug treatment for common health issues. The fact sheets that are most frequently accessed by physicians—those relating to nonopioid analgesics and anti-infectives (Figure 1)—point in the same direction.

After reading a fact sheet, medications classified by Embryotox as drugs of choice for certain situations were perceived as less harmful, indicating a reassuring and adherence-promoting effect. Drugs classified as teratogenic and/or fetotoxic were perceived as more risky after reading the fact sheet, suggesting that the information was correctly understood as a safety warning. Both effects contribute to the safe use of medication during pregnancy and breastfeeding. Overall, around two-thirds of participating users changed their risk perception of a drug, and a significant proportion also considered changing drug treatment. These findings demonstrate the impact of the information provided by Embryotox in routine health care, as well as users' intention to act based on the given information when necessary. Additionally, more than half of the participating physicians and over three-quarters of obstetricians and gynecologists would find further drug fact sheets useful, indicating the need for a broader, expanded range of information.

Comprehensibility of fact sheets was rated as good to very good, depending on user group and level of education: HCPs gave average ratings above 9 on a Likert scale from 0 to 10, while lay persons gave ratings above 8. However, patients with a higher level of education gave higher comprehensibility ratings and were overrepresented among users. These findings are in line with the results of other studies on the use of pregnancy-related internet resources [3] and telephone counseling services on drug safety in pregnancy [43]. Although Embryotox was initially developed for HCP, the high proportion of patients among users is not surprising, given that (1) the vast majority of pregnant women now search for health information online themselves; (2) there is no alternative, adequate German online resource; and (3) a large proportion of participating patients were referred to Embryotox by HCP—almost 40% (4283/10,860) by their physicians, others by their midwives or pharmacists (Figure 3). Despite there being no separate versions tailored specifically to the needs of each user group [44,45], our results show that both HCP and laypeople can effectively use Embryotox, albeit with the aforementioned limitations. In this context, it is important to note that Embryotox focuses on supporting very specific clinical decisions, which has already been discussed as a factor that facilitates the use of knowledge resources [46,47]. The fact that HCPs recommend the use of Embryotox to their patients could already indicate a potential role in shared decision-making. A 2023 umbrella review showed that involving patients in decision-making processes during pregnancy improved satisfaction with perinatal care, reduced distress, and strengthened the sense of empowerment and control [48]. Shared decision-making requires a collaborative process between physicians and patients, in which evidence-based medical information is discussed in the context of the patient's individual values and priorities. Previous studies have shown that physicians would find it helpful to use online health information suitable for patient education [47], and that such information would support patients in making clinical decisions [49]. Therefore, it was not unexpected that patients in our study

XSL•FO
RenderX

reported that using Embryotox empowered them to play a more active role in decision-making and sometimes even to alert their HCP to important options or information. Similarly, it was not unexpected that the joint use of Embryotox by all parties involved in the treatment process was reported to be of benefit to both HCPs and patients, as it facilitated communication and shared decision-making, strengthening confidence in the decision-making process (refer to Figures 5 and 6 and Textbox 2). In this context, facilitating communication applies both to communication between HCPs and patients and to communication among HCPs, for example, between physicians of different medical areas or between physicians and midwives or pharmacists.

Limitations

Overall, a solid database was available for both the descriptive statistics and the qualitative content analysis. However, quantitative data were collected in a convenience sample of website users who volunteered to participate. Therefore, the proportion of patients may be higher among participants than among all website users, as patients may be more interested in participating in online surveys than HCPs who are likely to be busier. In order to avoid bias in overall results due to a potentially disproportionate number of patients, results are generally reported by user group. It can be assumed that our study includes all relevant user groups, as residual categories such as “other HCP” or “other context” were rarely selected in the questionnaire. If, for example, users with positive Embryotox experiences used it more regularly and were more likely to participate, this may have introduced a degree of sampling bias (response bias). This could apply to patients with a higher level of education, who are overrepresented in our sample. The extent to which the use of Embryotox is appropriate for all relevant stakeholders in the care of pregnant and breastfeeding patients cannot be reliably described from our data, as the questionnaires were placed directly on the internet portal and only website users participated in the study. However, the steadily increasing number of visitors to the website, with more than 5 million visits in 2024, suggests widespread use in routine health care.

Conclusions

The findings of this mixed methods study suggest that online knowledge resources on drug safety during pregnancy and breastfeeding, such as the freely accessible Embryotox website, can benefit a wide range of stakeholders in routine health care. With regard to Embryotox, further improvements in care could be achieved by increasing the number of available drug fact sheets and by making the online information easier to understand for less educated users. In its existing setup, users appreciated the website content for providing comprehensible information, facilitating communication, supporting shared decision-making, and empowering patients to take a more active role in decision-making. In doing so, Embryotox promotes treatment adherence and drug safety for pregnant and breastfeeding women.

Acknowledgments

We would like to thank all users of the Embryotox internet portal who participated in the study by completing questionnaires or participating in interviews.

Generative artificial intelligence (AI) tools were not used in the preparation of this manuscript.

Funding

The presented mixed methods study was funded by the Innovation Fund of the Federal Joint Committee (Gemeinsamer Bundesausschuss, funding code: 01VSF20010).

Data Availability

All questionnaire data analyzed in this study are included in this published article (and its supplementary information files); interview data generated and analyzed during this study are not publicly available due to privacy protection of interview participants.

Authors' Contributions

Conceptualization: KD, MO, CS
Data curation and formal analysis: EB, AM
Funding acquisition: KD, MO, CS
Investigation: KD, AM, FM, MO
Methodology: EB, MB, KD, CH, MO, CS
Project administration: KD, MO
Supervision: KD, MB, CH, MO, CS
Visualization: EB, KD, MO
Writing – original draft: KD, MO
Writing – review & editing: EB, MB, KD, CH, FM, AM, MO, CS

XSL•FO
RenderX

Conflicts of Interest

EB, KD, AM, FM, MO, and CS are/were staff members of the Berlin Embryotox Center of Clinical Teratology and Drug Safety in Pregnancy.

Multimedia Appendix 1
Overview of online resources on drug safety in pregnancy and breastfeeding, study instruments, drug fact sheets, and survey results.
[DOCX File, 32 KB - jmir_v28i1e81286_app1.docx]

Multimedia Appendix 2
Questionnaire 1.
[DOCX File, 37 KB - jmir_v28i1e81286_app2.docx]

Multimedia Appendix 3
Questionnaire 2.
[DOCX File, 48 KB - jmir_v28i1e81286_app3.docx]

Checklist 1
CHERRIES checklist.
[DOCX File, 38 KB - jmir_v28i1e81286_app4.docx]

Checklist 2
COREQ checklist.
[DOCX File, 23 KB - jmir_v28i1e81286_app5.docx]

References

1. Bérard A, Abbas-Chorfa F, Kassai B, et al. The French Pregnancy Cohort: medication use during pregnancy in the French population. PLoS One 2019;14(7):e0219095. [doi: 10.1371/journal.pone.0219095] [Medline: 31314794]
2. Lupattelli A, Spigset O, Nordeng H. Adherence to medication for chronic disorders during pregnancy: results from a multinational study. Int J Clin Pharm 2014 Feb;36(1):145-153. [doi: 10.1007/s11096-013-9864-y] [Medline: 24162929]
3. Conrad M. Health information-seeking internet behaviours among pregnant women: a narrative literature review. J Reprod Infant Psychol 2024 Mar;42(2):194-208. [doi: 10.1080/02646838.2022.2088711] [Medline: 35703164]
4. van den Heuvel JF, Groenhof TK, Veerbeek JH, et al. eHealth as the next-generation perinatal care: an overview of the literature. J Med Internet Res 2018 Jun 5;20(6):e202. [doi: 10.2196/jmir.9262] [Medline: 29871855]
5. Hsiao SH, Hsu HY, Erickson SR, Lu TS, Wu CH. Seeking health information online among U.S. pregnant women: findings from the 2009-2018 National Health Interview Surveys. Women Health 2023 Jan;63(1):59-69. [doi: 10.1080/03630242.2022.2158413] [Medline: 36546357]
6. Wallwiener S, Müller M, Doster A, et al. Pregnancy eHealth and mHealth: user proportions and characteristics of pregnant women using web-based information sources-a cross-sectional study. Arch Gynecol Obstet 2016 Nov;294(5):937-944. [doi: 10.1007/s00404-016-4093-y] [Medline: 27084763]
7. Lagan BM, Sinclair M, Kernohan WG. A web-based survey of midwives' perceptions of women using the internet in pregnancy: a global phenomenon. Midwifery 2011 Apr;27(2):273-281. [doi: 10.1016/j.midw.2009.07.002] [Medline: 19700228]
8. Johnson-Arbor K. I read health information online-can I trust it? JAMA Intern Med 2024 Sep 1;184(9):38587816. [doi: 10.1001/jamainternmed.2024.0340] [Medline: 38587816]
9. Morley J, Cowls J, Taddeo M, Floridi L. Public health in the information age: recognizing the infosphere as a social determinant of health. J Med Internet Res 2020 Aug 3;22(8):e19311. [doi: 10.2196/19311] [Medline: 32648850]
10. Nörby U, Noël-Cuppers B, Hristoskova S, et al. Online information discrepancies regarding safety of medicine use during pregnancy and lactation: an IMI ConcePTION study. Expert Opin Drug Saf 2021 Sep;20(9):1117-1124. [doi: 10.1080/14740338.2021.1935865] [Medline: 34080506]
11. Hämeen-Anttila K, Nordeng H, Kokki E, et al. Multiple information sources and consequences of conflicting information about medicine use during pregnancy: a multinational Internet-based survey. J Med Internet Res 2014 Feb 20;16(2):e60. [doi: 10.2196/jmir.2939] [Medline: 24565696]
12. Palosse-Cantaloube L, Lacroix I, Rousseau V, Bagheri H, Montastruc JL, Damase-Michel C. Analysis of chats on French internet forums about drugs and pregnancy. Pharmacoepidemiol Drug Saf 2014 Dec;23(12):1330-1333. [doi: 10.1002/pds.3709] [Medline: 25250824]

XSL•FO
RenderX

13. Van Calsteren K, Gersak K, Sundseth H, et al. Position statement from the European Board and College of Obstetrics & Gynaecology (EBCOG): the use of medicines during pregnancy: call for action. Eur J Obstet Gynecol Reprod Biol 2016 Jun;201(211–4):211-214. [doi: 10.1016/j.ejogrb.2016.04.016] [Medline: 27132708]
14. Ceulemans M, Van Calsteren K, Allegaert K, Foulon V. Information needs and counseling preferences among potential users of the future Teratology Information Service in Belgium: a cross-sectional study involving the public and healthcare professionals. Int J Environ Res Public Health 2022 Jul 14;19(14):8605. [doi: 10.3390/ijerph19148605] [Medline: 35886455]
15. Lu Y, Barrett LA, Lin RZ, Amith M, Tao C, He Z. Understanding information needs and barriers to accessing health information across all stages of pregnancy: systematic review. JMIR Pediatr Parent 2022 Feb 21;5(1):e32235. [doi: 10.2196/32235] [Medline: 35188477]
16. European Network of Teratology Information Services. URL: https://www.entis-org.eu/centers/ [accessed 2025-07-25]
17. Organization of Teratology Information Specialists. URL: https://mothertobaby.org/affiliates/ [accessed 2025-07-25]
18. Dathe K, Schaefer C. Drug safety in pregnancy: the German Embryotox institute. Eur J Clin Pharmacol 2018 Feb;74(2):171-179. [doi: 10.1007/s00228-017-2351-y] [Medline: 29063150]
19. Schaefer C, Amoura-Elefant E, Vial T, et al. Pregnancy outcome after prenatal quinolone exposure. Evaluation of a case registry of the European Network of Teratology Information Services (ENTIS). Eur J Obstet Gynecol Reprod Biol 1996 Nov;69(2):83-89. [doi: 10.1016/0301-2115(95)02524-3] [Medline: 8902438]
20. Onken M, Lohse L, Coulm B, et al. Effects of maternal modafinil treatment on fetal development and neonatal growth parameters - a multicenter case series of the European Network of Teratology Information Services (ENTIS). Acta Psychiatr Scand 2024 Nov;150(5):372-384. [doi: 10.1111/acps.13643] [Medline: 38110225]
21. Embryotox.de (home). Pharmacovigilance and Consultation Center for Embryonic Toxicology. URL: https://www.embryotox.de/ [accessed 2025-07-25]
22. Tovin MM, Wormley ME. Systematic development of standards for mixed methods reporting in rehabilitation health sciences research. Phys Ther 2023 Nov 4;103(11):pzad084. [doi: 10.1093/ptj/pzad084] [Medline: 37672215]
23. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs-principles and practices. Health Serv Res 2013 Dec;48(6 Pt 2):2134-2156. [doi: 10.1111/1475-6773.12117] [Medline: 24279835]
24. Eysenbach G. Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
25. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
26. The R project for statistical computing. R Foundation for Statistical Computing. 2024. URL: https://www.r-project.org/ [accessed 2026-05-28]
27. Tran VT, Porcher R, Tran VC, Ravaud P. Predicting data saturation in qualitative surveys with mathematical models from ecological research. J Clin Epidemiol 2017 Feb;82(71–8):71-78. [doi: 10.1016/j.jclinepi.2016.10.001] [Medline: 27789316]
28. Kuckartz U. Qualitative Text Analysis: A Guide to Methods, Practice & Using Software: SAGE Publications Ltd; 2014. [doi: 10.4135/9781446288719]
29. F4transkript v7 and f4analyse v3. audiotranskription (Home). 2021. URL: https://www.audiotranskription.de [accessed 2026-05-27]
30. Flick U, Kardoff E, Steinke I, editors. A Companion to Qualitative Research: SAGE Publications; 2004.
31. Ceulemans M, Van Calsteren K, Allegaert K, Foulon V. Beliefs about medicines and information needs among pregnant women visiting a tertiary hospital in Belgium. Eur J Clin Pharmacol 2019 Jul;75(7):995-1003. [doi: 10.1007/s00228-019-02653-w] [Medline: 30834472]
32. Tan YXF, Lim STY, Lim JL, Ng TTM, Chng HT. Drug information-seeking behaviours of physicians, nurses and pharmacists: a systematic literature review. Health Info Libr J 2023 Jun;40(2):125-168. [doi: 10.1111/hir.12472] [Medline: 36655603]
33. Hämeen-Anttila K, Jyrkkä J, Enlund H, Nordeng H, Lupattelli A, Kokki E. Medicines information needs during pregnancy: a multinational comparison. BMJ Open 2013;3(4):e002594. [doi: 10.1136/bmjopen-2013-002594] [Medline: 23624989]
34. Learn about medications, vaccines, imaging and addictions. Le CRAT. URL: www.lecrat.fr [accessed 2026-05-28]
35. Marin B, Coulm B, Beghin D, et al. Evaluation of the satisfaction of healthcare professionals using the CRAT website, and launch of a new version. Therapie 2024;79(5):533-541. [doi: 10.1016/j.therap.2024.03.004] [Medline: 38609757]
36. Allison R, Hayes C, McNulty CAM, Young V. A comprehensive framework to evaluate websites: literature review and development of GoodWeb. JMIR Form Res 2019 Oct 24;3(4):e14372. [doi: 10.2196/14372] [Medline: 31651406]
37. Sbaffi L, Rowley J. Trust and credibility in web-based health information: a review and Agenda for future research. J Med Internet Res 2017 Jun 19;19(6):e218. [doi: 10.2196/jmir.7579] [Medline: 28630033]
38. Powell J, Inglis N, Ronnie J, Large S. The characteristics and motivations of online health information seekers: cross-sectional survey and qualitative interview study. J Med Internet Res 2011 Feb 23;13(1):e20. [doi: 10.2196/jmir.1600] [Medline: 21345783]
39. Sun Y, Zhang Y, Gwizdka J, Trace CB. Consumer evaluation of the quality of online health information: systematic literature review of relevant criteria and indicators. J Med Internet Res 2019 May 2;21(5):e12522. [doi: 10.2196/12522] [Medline: 31045507]

XSL•FO
RenderX

40. Yudianto B, Caldwell PH, Nanan R, Barnes EH, Scott KM. Patterns of parental online health information-seeking behaviour. J Paediatr Child Health 2023 May;59(5):743-752. [doi: 10.1111/jpc.16387] [Medline: 37051735]
41. Sillence E, Blythe JM, Briggs P, Moss M. A revised model of trust in internet-based health information and advice: cross-sectional questionnaire study. J Med Internet Res 2019 Nov 11;21(11):e11125. [doi: 10.2196/11125] [Medline: 31710297]
42. Helfferich C, Holz JL, Knittel T, Olejniczak L, Schmidt F. Risk it"-why women who do not intend to get pregnant do not use contraception: Results of the BZgA study "Women's lives 3. Family planning in women's lives. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2021 Nov;64(11):1408-1415. [doi: 10.1007/s00103-021-03439-1] [Medline: 34648039]
43. Beck E, Lechner A, Schaefer C. Who seeks Teratology Information Service's advice? Assessing the risk of selection bias in observational cohort studies on drug risks in pregnancy. Reprod Toxicol 2017 Jan;67(79–84):79-84. [doi: 10.1016/j.reprotox.2016.11.019] [Medline: 27915015]
44. Lühnen J, Albrecht M, Mühlhauser I. Guideline for the development of evidence-based health information. : German Network for Evidence-based Medicine; 2017 URL: https://www.leitlinie-gesundheitsinformation.de/wp-content/uploads/2017/07/Leitlinie-evidenzbasierte-Gesundheitsinformation.pdf [accessed 2026-05-27]
45. Arbeitsgruppe GPGI. Good practice guidelines for health information [article in German]. Z Evid Fortbild Qual Gesundhwes 2016;110-111(85–92):85-92. [doi: 10.1016/j.zefq.2015.11.005] [Medline: 26875040]
46. Cook DA, Sorensen KJ, Wilkinson JM, Berger RA. Barriers and decisions when answering clinical questions at the point of care: a grounded theory study. JAMA Intern Med 2013 Nov 25;173(21):1962-1969. [doi: 10.1001/jamainternmed.2013.10103] [Medline: 23979118]
47. van der Keylen P, Tomandl J, Wollmann K, et al. The online health information needs of family physicians: systematic review of qualitative and quantitative studies. J Med Internet Res 2020 Dec 30;22(12):e18816. [doi: 10.2196/18816] [Medline: 33377874]
48. Alruwaili TA, Crawford K, Jahanfar S, Hampton K, Fooladi E. Pregnant persons and birth partners' experiences of shared decision-making during pregnancy and childbirth: an umbrella review. Patient Educ Couns 2023 Sep;114:107832. [doi: 10.1016/j.pec.2023.107832] [Medline: 37295043]
49. Thapa DK, Visentin DC, Kornhaber R, West S, Cleary M. The influence of online health information on health decisions: a systematic review. Patient Educ Couns 2021 Apr;104(4):770-784. [doi: 10.1016/j.pec.2020.11.016] [Medline: 33358253]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
COREQ: Consolidated Criteria for Reporting Qualitative Research
HCP: health care professional

Edited by A Stone; submitted 25.Jul.2025; peer-reviewed by L Schuler-Faccini, W Rascher; revised version received 09.Mar.2026; accepted 30.Mar.2026; published 25.Jun.2026.

Please cite as:
Onken M, Beck E, Breuning M, Holmberg C, Metke F, Müller A, Schaefer C, Dathe K
Use of the Online Portal "Embryotox" in Routine Health Care: Mixed Methods Study
J Med Internet Res 2026;28:e81286
URL: https://www.jmir.org/2026/1/e81286
doi:10.2196/81286

© Marlies Onken, Evelin Beck, Martina Breuning, Christine Holmberg, Franziska Metke, Anne Müller, Christof Schaefer, Katarina Dathe. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Usage and Perceptions of Electronic Patient Records Experienced by Users and Nonusers in the Canton of Vaud, Switzerland: Mixed Methods Study

François Bastardot[1*], MD, MSc; Catherine LaRocco[2], BSE, DPT; Mathilde Robert[2], PT, MAS; Veronika Schoeb[2*], PT, MHA, PhD

[1]Department of Medical Informatics, Lausanne University Hospital (CHUV), Lausanne, Switzerland

[2]Department of Research & Innovation, HESAV School of Health Sciences - Vaud, HES-SO University of Applied Sciences and Arts Western Switzerland, Avenue Beaumont 21, Lausanne, Switzerland

[*]these authors contributed equally

Corresponding Author:
Veronika Schoeb, PT, MHA, PhD
Department of Research & Innovation, HESAV School of Health Sciences - Vaud, HES-SO University of Applied Sciences and Arts Western Switzerland, Avenue Beaumont 21, Lausanne, Switzerland

Abstract

Background: Electronic patient records (EPRs) have shown potential to improve health care delivery, coordination, and patient engagement. In Switzerland, the development of a national EPR is supported by the Confederation, while its deployment is carried out by the cantons through EPR communities. Since health policy and the organization of health systems are entrusted to cantonal governments, each of the 26 Swiss cantons must decide its strategy for the implementation and application of EPRs within its jurisdiction. The canton of Vaud has joined with other French-speaking cantons (Geneva, Valais, Fribourg, and Jura) in organizing the implementation of the EPR through the CARA association since 2018. By June 2025, approximately 32,350 individuals have opened an EPR (Dossier électronique du patient) through CARA, accounting for 28% of all EPRs opened nationwide at that time.

Objective: This study aimed (1) to document patients' use of EPR in the canton of Vaud and to examine the user profiles; and (2) to explore users' and nonusers' perceptions regarding EPR benefits and barriers.

Methods: This mixed methods study was conducted in 2 phases. First, a quantitative questionnaire examined EPR user profiles (eg, demographics and health status). Health literacy and digital health literacy were assessed using the HLS19-Q12-CH (Health Literacy Survey 2019-2021) and HLS19-DIGI-CH (Digital Health Literacy) scales in French. Second, qualitative semistructured interviews explored experiences and perceptions of users and nonusers through thematic analysis. Data from both phases were integrated during interpretation. All analysis was conducted in French with quotes translated for publication.

Results: The population included 839 patients with EPRs (early adopters from December 2021 to December 2023) and a matched control group. Overall participation was 19.3% (324/1678), with higher participation among EPR users (255/839, 30.4%) than nonusers (69/839, 8.2%). Early adopters were predominantly male (185/255, 72.5%), highly educated (155/255, 60.8%), affected by chronic illness (204/255, 80%), and with extensive health care networks (80% consulted 2 or more health care professionals in the last 3 months). The vast majority (249/253, 98.4%) had a general practitioner. While only a minority (76/252, 30.2%) of early adopters were satisfied with the service provided by the EPR, a majority (172/254, 67.7%) would recommend it to their family and friends. Qualitative analysis identified themes that influence EPR adoption, including the contextual environment, level of health literacy, EPR as a tool, professionals' resistance to EPR, and level of engagement with digitalization.

Conclusions: This study identified early EPR adopters as predominantly highly educated males with chronic diseases and a regular general practitioner. Despite moderate satisfaction with the current implementation, most users recommend the system to others, suggesting a belief in its potential value. Important questions are raised regarding EPR accessibility, limited adoption by professionals, and potential digital health disparities in the general population.

(J Med Internet Res 2026;28:e83702) doi:10.2196/83702

KEYWORDS

electronic patient record; mixed methods study; health literacy; digital health literacy; health disparities; patient engagement; Swiss health care system; patient-provider relationship; communication; perceptions

XSL•FO
RenderX

Introduction

The digitalization of health care represents a significant public health development worldwide, with electronic patient records (EPRs) emerging as a cornerstone of this transformation. EPRs are introduced to encourage patient self-management and responsibility for their own health while potentially improving care coordination and health care delivery [1,2]. Their promise to enhance health care quality, reduce costs, and improve patient outcomes has driven widespread implementation efforts across health care systems globally.

In Switzerland, the national EPR aligns with the Health2030 strategy and the Federal Law on Electronic Patient Record (LDEP [3]), which aims to establish a standardized framework for digital health records. The case of Switzerland, a confederation of 26 cantons, is particularly interesting, as it is a small country at the heart of Europe with a decentralized health care system. Its constitution signed in 1848 defines the political foundations and the principle of subsidiarity, granting specific power to the federal state, whereas 26 cantons continue to be sovereign based on their own cantonal constitution [4]. The Federal Assembly is the supreme authority of the Confederation and represents both the cantons and the people [4]. The cornerstone of the Swiss health care system is the Health Insurance Act (LaMal [5]), a federal law from 1994. However, distributions of roles and responsibilities at federal or cantonal levels are not clearly established [4]. Several limitations of the current governance structure of the Swiss health care system have been identified: health promotion initiatives not receiving sufficient support and funding, some cantons being too small to provide required health care services to their population, and the deployment of EPR to all citizens being challenging [6]. It is the latter that is the topic of this article.

In 2018, the CARA association was established to provide EPR services for patients and professionals across 5 French-speaking cantons: Vaud, Valais, Fribourg, Jura, and Geneva [7]. This platform offers both EPRs and document transfer services. By June 2025, the health platform CARA had registered approximately 32,350 records and more than 2840 health care providers and institutions.

Different terminology exists regarding electronic health and medical records: electronic health records (EHRs), electronic medical records (EMRs), and EPRs, sometimes also called patient health records. EMRs are considered an internal organizational system, while EHRs are for collaboration across institutions [8]. EMRs and EHRs are controlled by professionals, while EPRs are managed by the patient, giving the patient the authority to integrate multiple sources and control access [9]. The Competence and Coordination Centre of the Confederation eHealth Suisse oversees the coordination, development, and dissemination of information regarding the EPR, whereas the cantons are responsible for its implementation.

The EPR typically contains discharge letters, medical reports, surgical reports, laboratory results, imaging reports (radiology), prescriptions, medication lists, and vaccination data [7]. Health care providers who might contribute data include hospitals, birth centers, nursing homes, clinical laboratories, general practitioners (GPs) and specialized physicians, pharmacies, home care agencies, and therapeutic practitioners [7]. The EPR content is partial rather than comprehensive: only information deemed useful for continuity of care is published, with each health care provider determining what qualifies as relevant. As patients retain complete control over which providers can access their records and which documents are deposited, information might not flow as easily as in EHRs. While patient portals are extensions of EHRs (eg, Epic MyChart) and thus controlled by health professionals, patients oversee EPRs, deciding who has access to documents.

However, EPR adoption rates vary considerably across European countries. While Finland and Estonia have demonstrated remarkable success in implementation, countries such as France, Austria, and Germany have experienced significantly lower adoption rates [10-14]. These disparities can be attributed to differences in implementation models, with some countries opting for a single centralized database (eg, Estonia), while others have adopted multiple centralized bodies (eg, Germany) or decentralized approaches (eg, Austria) [15]. Furthermore, differences exist regarding data sharing. While the Swiss EPR is restricted to users and their health professionals to whom they have given explicit access, other countries share data with the national health insurance (eg, France) or between providers using a centralized national platform (eg, Estonia). Whether these contextual differences impact EPR use and how adoption rates would play out in a decentralized health care system, such as Switzerland, is still unknown.

The scientific literature suggests that EPRs might have a positive influence on care coordination, management, and patient safety [16]. Studies have shown that patients with access to their health records tend to develop a better understanding of their health conditions, demonstrate improved medication adherence, and experience enhanced patient-provider relationships. Furthermore, EPR access seems to encourage greater patient self-management and engagement in health care decision-making [17].

Previous research has characterized early EPR adopters as individuals with specific demographic and health profiles. These include persons with one or more chronic diseases [18], predominantly female users [19-21], and those with higher socioeconomic status and health literacy [22,23]. However, to date, no Swiss study has comprehensively examined the characteristics and experiences of EPR users within the Swiss healthcare context.

This research gap is particularly relevant given Switzerland's fragmented health care system and its nonalignment with European health policies, characterized by cantonal autonomy in health policy implementation and a strong emphasis on patient privacy. Understanding the profiles, motivations, and experiences of EPR users and nonusers in Switzerland is crucial for informing effective implementation strategies and addressing potential barriers to adoption.

This study aimed (1) to document patients' use of EPR in the Swiss canton of Vaud and to examine the users' profiles; and (2) to explore the users' and nonusers' perceptions regarding EPR and its benefits and barriers.

XSL•FO
RenderX

Methods

Overview

An explanatory sequential mixed methods study was conducted in 2 phases: first collecting and analyzing quantitative data, followed by qualitative data collection and analysis [24]. A mixed methods research study has the advantage of being able to elucidate complex issues [25], and using this design, researchers can combine the strengths of 2 approaches and thereby contribute to interdisciplinary and collaborative research evidence [26]. In our study, 2 methods (quantitative questionnaire and interviews) were organized sequentially [27] and analyzed, in accordance with the rules of each method [28].

Quantitative Study

Adult patients residing in the canton of Vaud, Switzerland, who had an episode of care (inpatient or outpatient) between December 1, 2021, and December 28, 2023, at Lausanne University Hospital (CHUV) with an active EPR were identified (N=839) and matched 1:1 with adult patients without EPR on age, self-identified gender, and care period (N=839); refer to Figure 1. Whether an EPR was present at admission was automatically verified and reconciled against the institutional EHR (CHUV).

A questionnaire (Multimedia Appendix 1) was sent to them, which they returned by post or electronically. All data were collected and managed using REDCap (Research Electronic Data Capture) hosted at CHUV [29]. Demographic data and socioeconomic profiles were collected. A high level of education was equivalent to a higher or specialized school or university. In addition, the level of health literacy and digital health literacy was assessed using a validated questionnaire (HLS19-Q12-CH [Health Literacy Survey 2019-2021] and HLS19-DIGI-CH [Digital Health Literacy], scales in French [30,31]), comparing the profile of this population to a sample of the Swiss population recently studied in 2019‐2021 (Health Literacy Survey Schweiz 2019‐2021 [31]). Voluntary participants provided their contact details for the qualitative phase (Phase 2) at the end of the questionnaire.

On the statistical level, bivariate analyses were performed using 2-tailed Student *t* test or ANOVA for continuous data, and the chi-square test for categorical data. Bivariate associations were assessed by Spearman correlation. Statistical significance was considered for a 2-sided test with $P<.05$. Means are expressed with SD. All statistical analyses were performed using Stata version 18 software (StataCorp LLC).

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram. EPR: electronic patient record.

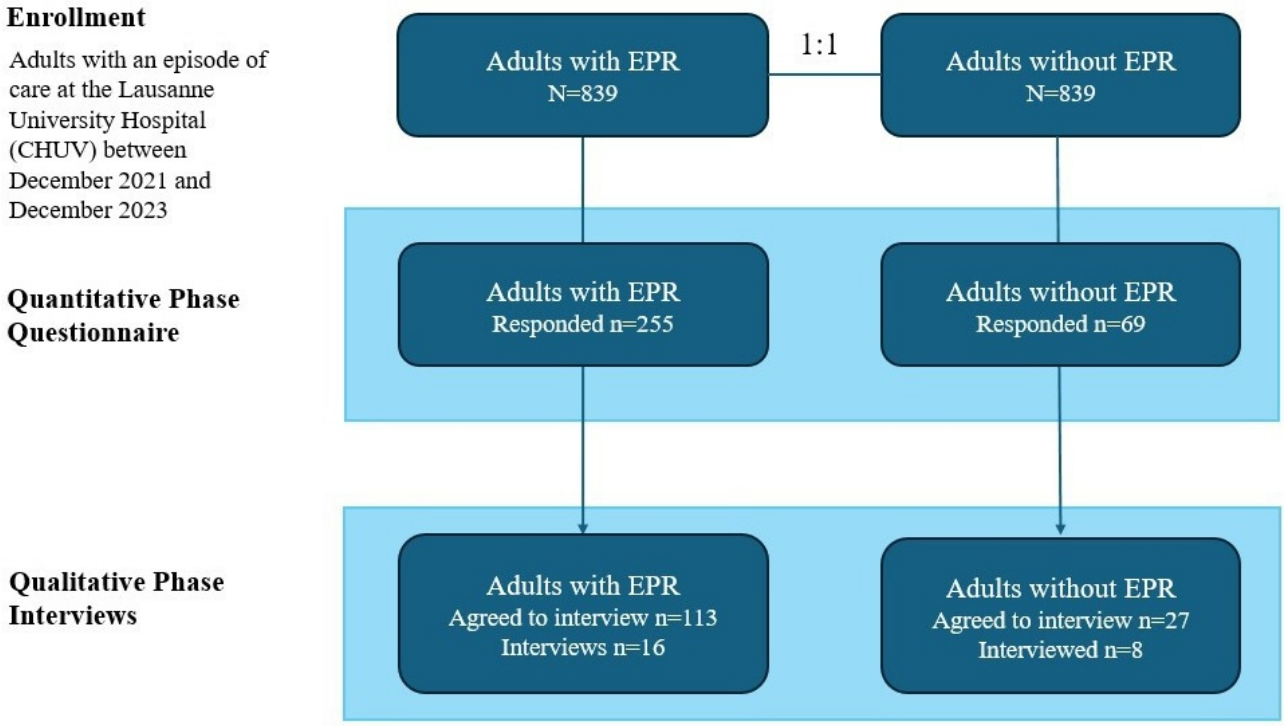

Qualitative Study

The qualitative analysis was performed to understand the users' and nonusers' experiences and perceptions toward the EPR. Semistructured interviews with EPR users and nonusers made it possible to understand the facilitators and barriers as well as the benefits of the EPR, through thematic analysis. The recruitment was structured as follows: participants from Phase 1 provided their contact details at the end of the questionnaire if they were willing to participate in interviews. A list of 140 potential interviewees who responded positively to the invitation (113 EPR users and 27 nonusers) was generated and stratified according to age, self-identified gender, professional status, comorbidities, and health literacy scores. For the user group, a priority list was established to recruit at least 15 participants, both men and women with diverse profiles. The group of nonusers was all invited, but only 8 participants were willing to schedule an interview. The goal was to recruit 20-25

XSL•FO
RenderX

interviewees with diverse profiles among participants, so that a range of perceptions could be collected.

Interviews: Data Collection and Analysis

Participants shared their perceptions, their engagement in that experience, and evaluated their experiences regarding the EPR. An interview guide was developed (Multimedia Appendix 2), and pilot interviews were conducted to assess the suitability of the questions. The duration of interviews ranged from 15 minutes to 1 hour 12 minutes (mean 39, SD 17 minutes). To ensure the transparency and validity of the findings, all interviews were conducted in accordance with the interview guide. The interviews were transcribed "ad verbatim" in French (except 1 interview conducted in English) and data were analyzed using thematic analysis [32] with the help of MAXQDA version 24 (VERBI GmbH).

The 6 steps described by Braun and Clarke [32] were used to transcribe the data (step 1), followed by systematically coding the collected data (step 2). The next phase included a compilation of codes into themes (step 3) that were then reviewed (step 4) and discussed. Three researchers (VS, MR, and CL) worked together on the dataset, discussing themes, and refining the relations between the different themes (step 5) and writing it all up (step 6). An Excel table allowed for transparency of the process, being able to compare themes, subthemes, and citations. It must be noted that the reflexive thematic analysis of Braun and Clarke [33], rather than using a codebook approach or a coding reliability approach based on a positivist paradigm [34], was in line with the interpretative analysis of the data.

While the entire data analysis process was performed in French, for this article, themes, subthemes, and citations are translated into English.

Data Integration

At first, data from phases 1 and 2 were analyzed separately and integrated during the interpretation phase to find explanations of the different uses and perceptions of EPR. The integration of the quantitative and qualitative phases happened at 2 levels: first, based on the quantitative results, the selected participants with varied characteristics ensured a rich representation of responders. This approach allowed us to link qualitative and quantitative data at the individual level, and an in-depth analysis of this subgroup was performed. Second, a follow-up analysis was performed based on the quantitative data from Phase 1 [27] to see whether qualitative data were linkable to results from the quantitative data at the aggregate level.

Interpretation was all done in the original language (French), and translation of the quotes into English was performed only for the purpose of authoring this paper.

Reflexivity

The concept of reflexivity was adopted throughout the study to ensure its scientific rigor [35]. Specifically, reflexivity is defined as a thoughtful and conscious process that includes continuous evaluation of subjective responses, interpersonal dynamics, and the research process itself [36]. Discussions among the research group members from different disciplinary backgrounds (medicine, physiotherapy, and sociology) facilitated the triangulation of perspectives.

Ethical Considerations

This mixed methods study was conducted in accordance with the Declaration of Helsinki and received ethical approval from the local ethics committee (Commission cantonale d'éthique de la recherche sur l'être humain, CER-VD, no 2023 - 01672). Participation was voluntary and all participants were fully informed about the study's objectives, procedures, and their right to withdraw at any time without consequences. Written informed consent was obtained from all participants prior to completing the questionnaire (Phase 1). For Phase 2, additional informed consent was obtained from the subset of participants selected for in-depth interviews. No compensation was provided. Participant anonymity was maintained throughout the study. All questionnaire responses and interview data were deidentified and assigned unique identification codes. Confidentiality was ensured through secure data storage with password-protected access limited to the research team members.

Results

Phase 1: Quantitative Study

Overview

The overall participation rate in this phase was 19.3% (324/1678), with a marked difference between EPR users (255/839, 30.4%) and nonusers (69/839, 8.2%). Men were overrepresented both in the initial population with an EPR (508/839, 60.5%) and among the study participants (185/255, 72.5%), as shown in Table 1. The low participation of nonusers is a methodological limitation. Nonusers were more difficult to recruit, possibly due to a lack of interest or knowledge about the EPR.

Patients who have opened an EPR tended to have a higher education, defined as higher or specialized school, or university (155/255, 60.8%). They are very often affected by chronic illnesses (204/255, 80%) and have access to an extensive health care network: 201/254 (79.1%) of them consulted 2 or more health care professionals in the last 3 months. Most of this population (249/253, 98.4%) had a GP. Patients with EPRs did not visit the emergency departments less frequently during the last 3 months (60/254, 23.6%). Refer to Table 1.

XSL•FO
RenderX

Table . Sociodemographic characteristics, health status, and use of health services of participants according to their EPR[a] adoption status (n=324).

Characteristics	With EPR	Without EPR	*P* value
Self-identified gender, n/N (%)			.04
Men	185/255 (72.5)	41/69 (59.4)	
Women	70/255 (27.5)	28/69 (40.6)	
Age (years), mean (SD)	60.4 (14.6)	60 (14.5)	.81
Married, n/N (%)	147/253 (58.1)	38/68 (55.9)	.87
Lives alone, n/N (%)	58/255 (22.8)	19/67 (28.4)	.77
Higher education, n/N (%)[b]	155/255 (60.8)	34/69 (49.3)	.09
General practitioner, n/N (%)	249/253 (98.4)	64/68 (94.1)	.04
Poor or very poor health, n/N (%)	29/254 (11.4)	8/69 (11.6)	.97
With chronic illness, n/N (%)	204/255 (80)	50/68 (73.5)	.10
Two or more health care providers during the last 3 months, n/N (%)	201/254 (79.1)	47/68 (69.1)	.047
One or more emergency department visits during the last 3 months, n/N (%)	60/254 (23.6)	12/68 (17.7)	.29

[a]EPR: electronic patient record.

[b]Higher education is defined as University or University of Applied Sciences (HES).

Use of and Satisfaction With EPR

Of the 251 participants with an EPR, 96 (38.3%) reported using it once during the last quarter, and 94 (37.5%) reported using it 2 or more times. Among EPR users, 76/252 (30.2%) stated that they were satisfied with the EPR; the reasons reported for this moderate level of satisfaction are essentially linked to the themes of poor compliance by private practice physicians (GPs), other hospitals and clinics outside the university hospital, or the difficulty in obtaining medical documents. The complexity of use and the lack of a user-friendly platform were also cited. While no correlation was observed between self-identified gender and satisfaction levels, educational attainment was strongly associated with satisfaction levels. Participants with higher education levels were significantly more critical, demonstrating a dissatisfaction rate of 98/155 (63.2%) compared to 40/97 (41.2%) among those with lower educational levels (P=.001). In contrast, no significant correlations were identified with either health literacy or digital health literacy levels.

Despite these reported limitations, a substantial majority of users (172/254, 67.7%) indicated they would recommend opening an EPR to someone close to them.

Relationship With Medical and Health Professionals

The impact of the EPR on the physician-patient relationship is positive or neutral overall. A total of 33.1% (84/254) of users think that the EPR has an impact on the relationship with the GP; this is most often reported as unchanged (49/79, 62%), often improved or much improved (26/79, 32.9%), or rarely deteriorated (4/79, 5.1%).

EPR Use and Anxiety

The effect on anxiety remains marginal. Eleven users (11/252, 4.4%) reported an effect on the level of anxiety; this proportion is comparable with other studies [37,38]. Only 1 participant reported that their anxiety had worsened through EPR use.

Health Literacy

Participants with EPR tended to have a higher health literacy score (questionnaire HLS19-Q12-CH) with a mean score of 83.5 (SD 18.8) than those without EPR (mean 81.5, SD 23.8), although this difference was not statistically significant (P=.99). By comparison, the Swiss Health Literacy study (HLS19-21-CH) [31] reported an average score of 77.3 (SD 19.6) at the national level, and 80.2 for the canton of Vaud, which ranks first among Swiss cantons. Early adopters of the EPR thus stand out for having a higher level of health literacy, which may be linked to the level of education reported in the questionnaire.

According to the definitions of the M-POHL 2023 categories [30], two-thirds of patients with or without an EPR (168/255, 65.9% vs 46/69, 66.7%, respectively) are considered to have an excellent level of health literacy; 51 EPR users (51/255, 20%) have a level of literacy considered insufficient (ie, problematic or inadequate). Among nonusers, the proportion of patients with an insufficient level is the same (14/69, 20.3%), including 10/69 (14.5%) with an inadequate level. By way of comparison and considering a change in the definition of the categories since 2023, the population with EPRs in our study has better health literacy than the general population [31], with a proportion of 49% (1226/2502) of the Swiss population reporting frequent difficulties in processing health information, linked to an inadequate level of literacy (Figure 2).

XSL•FO
RenderX

Figure 2. Distribution of health literacy levels (HLS19-Q12) by electronic patient record adoption status: comparison with the Swiss study [31]. EPR: electronic patient record; HLS19-21-CH: Health Literacy Survey 19-21.

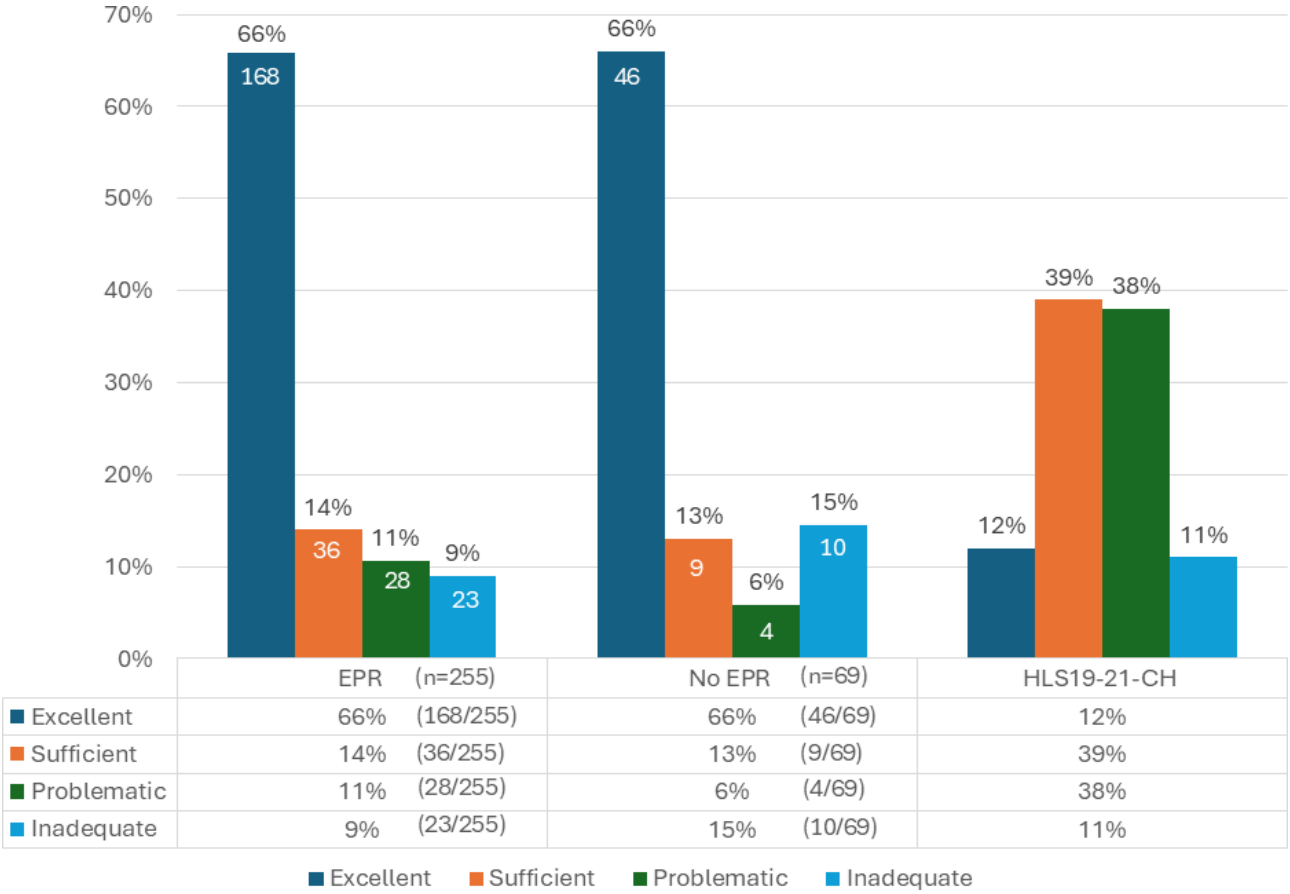

Digital Health Literacy

Patients with EPRs tended to have a slightly higher level of digital literacy (mean 60.4, SD 31.5), although this difference was not significant compared with the group without EPRs (mean 59.1, SD 36.0; *P*=.92). This population also tended to interact more with digital resources in the health context (mean 74.5, SD 35.5 vs mean 63.5, SD 43.5; *P*=.11), which is logically reflected in the regular use of the EPR, included in this category.

Phase 2: Interviews With EPR Users and Nonusers

Overview

For the qualitative interview study, we included 16 users and 8 nonusers of the EPR, totaling 24 participants (Table 2). While we were able to recruit an equal number of men (n=11) and women (n=13), it was more difficult to recruit EPR nonusers (n=8) compared to EPR users (n=16). It has also been noted that there is an overrepresentation of participants with a university degree (15 participants with a university degree and 2 with a postsecondary education) compared to a secondary education (high school level). Furthermore, retired persons made 50% (12/24) of the interviewed population with a slight overrepresentation of women (5 men and 7 women).

XSL•FO
RenderX

Table . Population interviewed (n=24).

ID	Gender (self-identified)	EPR[a] (yes or no)	Highest degree	Current professional activity
PERS1	Man	Yes	University	Retired
PERS2	Man	Yes	University	Retired
PERS3	Man	Yes	University	Retired
PERS4	Woman	No	Not known	Retired
PERS5	Woman	Yes	Not known	Retired
PERS6	Woman	Yes	Secondary education	Employed
PERS7	Woman	No	University	Manager
PERS8	Man	Yes	University	Manager
PERS9	Woman	Yes	University	Manager
PERS10	Woman	Yes	Higher postsecondary education (not university)	Employed
PERS11	Woman	Yes	University	Employed
PERS12	Woman	No	University	Manager
PERS13	Man	Yes	Higher postsecondary education	Not currently in the workforce
PERS14	Woman	Yes	Higher postsecondary education	Retired
PERS15	Woman	No	University	Retired
PERS16	Man	Yes	University	Retired
PERS17	Man	No	University	Manager
PERS18	Woman	Yes	University	Retired
PERS19	Woman	Yes	University	Retired
PERS20	Woman	No	Secondary education	Retired
PERS21	Man	No	University	Employed
PERS22	Man	Yes	University	Retired
PERS23	Man	No	Secondary education	Self-employed
PERS24	Man	Yes	Not known	Not currently in the workforce
Total	13 women and 11 men	16 persons with EPR and 8 persons without EPR		

[a]EPR: electronic patient record.

Six themes were identified during the thematic data analysis which are presented in detail below: digitalization, the Swiss context of the EPR, perceptions of the EPR, EPR as a tool, health literacy, and communication. To illustrate and make explicit the different themes, subthemes, and selected quotes, please refer to Table 3.

Most early adopters wanted to open an EPR because they were interested in tracking their health data and taking an active role in their own health, since many of them have chronic illnesses. Second, the desire to facilitate coordination between health care professionals to achieve optimal health care is also found among nonusers. The EPR context in Switzerland hampers its optimal use, as it is currently not compulsory for Swiss citizens (opt-in strategy). Among health care providers, only hospitals, birth centers, nursing homes, and physicians granted practice rights since 2022 are required to participate in the documentation and use of the EPR. The impact of this is the very low level of support and appreciation that early adopters encounter from professionals.

XSL•FO
RenderX

Table . Themes, subthemes, and quotes to illustrate findings.

Theme	Subtheme	Example quotes
Digitalization	Enablers of digitalization	“Yeah, quite a bit. Actually, I'm proud of not having to carry a wallet. I do everything with my watch. My children also encourage me a bit.” (PERS11, User)
	Limitations of digitalization	“So I imagine that's what will naturally scare a lot of people, it's this data, naturally.” (PERS24, User)
	Level of knowledge	“At home, I have everything electronic, so I mean all the files, all that, encrypted, of course.” (PERS1, User)
	Optimized usage of IT tools	“I also think that information provides security; the more information we have, the safer we feel.” (PERS11, User)
	Artificial intelligence	“And I'm also quite interested in everything related to artificial intelligence.” (PERS17, Non-user)
	Advanced state of digitalization	“There's the security aspect, of course, but I trust it.” (PERS11, User)
Swiss context	Level of knowledge of the EPR[a] among the Swiss population	“Given the number of members, I don't think so. I get the impression that people think it's something for geeks, or at best, maybe they think it was set up by doctors. And then it's something between them. When I talk about it a little, no one knows anything about it.” (PERS22, user)
	EPR requirement	“I think it should simply be mandatory.” (PERS8, user)
	Level of patient engagement	“I don't see how we can convince patients to make this effort (...). I mean, if they don't use it, there's no point. I don't need it, you know? I have everything on paper. I've made my own file at home with everything that's happened to me.” (PERS19, user)
	Level of health care professional engagement	“Now I think the implementation and the problem is that doctors don't use it. So there you go... it's useless.” (PERS19, user)
	Nonmandatory affiliation with the EPR	“If you have an identity that is truly nationally mandated. But anyway, we need it, we also need it to vote digitally. One day it will come, it's quite reliable.” PERS1, user)
	Lack of standardization	“There are as many EPRs as there are organizations. Each one has developed its own.” (PERS3, user)
	Power of insurers	“But it's true, when you think about it, I don't want my insurance to be able to keep up. Then I say to myself: No one has time to check if [PERS4] weighs three pounds more. But maybe they do, so they can send me dietary restrictions.” (PERS4, non-user)
Communication	Communication: between peers	“Because if we say: The doctor at the hospital where I arrive in an emergency can know everything about me, it's a win-win situation.” (PERS4, non-user)
	Communication: patient-professional	"So in my opinion, if the electronic health record becomes standard, it increases transparency between the patient and the doctor." (PERS1, user)
	Professional barriers to the EPR	“Because we know they're all already overloaded, even the assistants, secretaries, and so on. It's a disaster. ” (PERS13, user)

XSL•FO
RenderX

Theme	Subtheme	Example quotes
Health Literacy	Information seeking	“Especially on the internet because there's too much fake stuff, even if you find multiple sources. And then you often compare them, and they just copy each other. So to know what's really reliable, you can't trust it.” (PERS1, user)
	Level of engagement in personal health	“Also, taking on more responsibility in the healing and treatment process. Not delegating everything related to health to the health insurance company and the demigods in white coats, right?” (PERS11, user)
EPR as a tool	Opening of the EPR	"The registration process was too complex, but I think it has improved." (PERS22, user)
	Content	"Well, maybe it's my fault, but I can't get people to put my files in my dossier." (PERS5, user)
	Development proposals	“So, this is a case I should also tell you about, but I created some documents at [community hospital]. The document is scanned in a certain format, then the [University Hospital] receives it and processes it in a different format. Different departments at the hospital do it with completely different titles and things like that. There's nothing organized or standardized, so we were completely lost; it's a disaster.” (PERS22, user)
	Perceptions of the EPR	“So, if I wanted to get into this now, I would also like to have some support. Exactly. Because me, all alone at home typing in Swisspass, I don't know what. I can get to the train website anywhere, but I don't know if I'm really getting to the right place.” (PERS4, non-user)
	Use of EPR	“I think that access should be given to all professionals. But to what extent? Healthcare professionals, for example, doctors, we're talking about nurses. I might disagree with a nurse I don't know who has never seen me but who has seen my name? I don't know, on some document. And then who looks at my file, I wouldn't like that. But a doctor, it should stop at the doctor's level.” (PERS21, non-user)

[a]EPR: electronic patient record.

Digitalization

The approach to digitization is perceived differently, with some contrasting it with the loss of human contact, while others see it as a valuable aid. Some participants are very advanced in their use of digitalization and technology and use this technology to gather data:

> *So, I use an app that’s connected to the device that measures blood pressure. Oh yeah, arterial. That’s what I’ve been doing for a while. But it’s to keep track of it. And then I have all the values on my phone, and I can show them to a doctor, to my GP. When I go to see him, yes, I’ll show him all the measurements I’ve taken over the last few months.* [PERS19]
>
> *I have a little app that reminds me when I need to take medicine to make sure I take it and I don't forget.*
>
> *[PERS23]*

It is observed that older participants interviewed were also connected, not only the younger population:

> *Yeah, enough of that. I mean, I’m proud of the fact that I don’t carry around a wallet. I do everything with my watch. It’s also my kids who push me a bit.* [PERS11]

Overall, interviewees were quite at ease with digital tools, as can also be seen in the quantitative data with statistically nonsignificant differences in scores of digital health literacy.

Swiss Context of the EPR

The Swiss context in which the EPR has emerged has a substantial impact on the level of uptake. Participants have a negative perception of the lack of standardization and the multitude of digital health tools in Switzerland. They feel misinformed. In fact, many EPRs in French-speaking Switzerland have slightly different objectives and were not interoperable at the time of the interviews.

For participants, the noninvolvement of health care professionals, particularly GPs, is a significant obstacle to their use of EPR:

XSL•FO
RenderX

> *Ah well, doctors would have to use it first. If doctors don't use it. I don't see how we can convince patients to make the effort, ie, they're the ones holding the knife by the handle. I mean, if they don't use it. There's no, no reason for it. I mean, I don't need it. For me, I've got everything on paper. I've made my own file which is at home with everything that's happened to me.* [PERS19]

As currently not compulsory for Swiss citizens, early adopters perceive that registration with the EPR has a negative impact on the low level of support from professionals.

> *I think it should be compulsory for everyone to work with it. Otherwise, it won't work if a doctor doesn't, if a physiotherapist doesn't want to, it really should be a given that everyone on this platform, after of course the data protection issues for the mass of data protected but, there are solutions.* [PERS11]

Some participants proposed that EPR be assigned to every citizen at birth.

Perceptions of the EPR

The EPR is perceived positively by all participants, even those who were not users of the tool. They cite the benefits of limiting health care costs by reducing duplicate examinations and limiting errors in care. For patients, the EPR helps facilitate collaboration between health care professionals, ultimately leading to improved care.

The motivations of early adopters included having an interest in tracking their health data and taking an active role in their own health, since most of them have chronic illnesses. Second, they aimed to facilitate coordination among health care professionals, with the goal of achieving optimal health care management.

EPR as a Tool

Everyone agrees that the EPR tool was not designed according to the patients' needs. Participants registered with the EPR perceive numerous barriers to its optimal use: difficulties linked to the high security level of the platform (multifactor authentication and no mobile app), classification by date in order of insertion, and difficulty in searching for documents. As a result of the interviews, participants from both groups concurred that the platform needs to be enhanced to improve its user-friendliness. Table 4 presents participants' suggestions for improving the platform.

Table . Suggestions for improvement proposed by participants.

Suggestions	Description
Access to EPR[a]	Easier access to the platform, by using a mobile app and equalizing access to all health care professionals.
Data centralization	To reduce redundancy of tests (lab and imaging), available updated medication list and improved health professional coordination.
Search function	To be able to easily find historical data on the platform.
Information	Health professionals can upload personalized and reliable information (including comorbidities, medication list, exercise videos, contraindications, precautions, and allergies) for self-care.
Calendar and alerts	Appointment reminders, vaccine boosters, or screening tests out of date.
Network	List of all providers used by patients including specialists, therapists, dentists, osteopaths, chiropractors, paramedical providers such as optometry, podiatry, pharmacy, social workers, dietetics, and complementary services, family members, and emergency contacts.
Dashboard	Graphical representation of essential health care data such as vital signs, labs, pain scale, Functional Independence Measure, BMI, and hemoglobin A_{1c}.
Advanced directives	Information about health care power of attorney and medical decisions, and other relevant legal implications.

[a]EPR: electronic patient record.

Health Literacy

The early adopters were already very committed to their health, archiving their health data at home and taking a proactive stance in their health.

> *Afterwards, I've always been someone who likes to have a look at my treatment, who wants to understand why things are done and why they're not done differently. So, I've always asked my doctors a lot of questions, either by e-mail, or in person, in consultation with them, or by telephone. So, I've always taken a lot of notes from people, a lot of letters. In fact, there's a lot of information that I got, but because I led a process, in quotation marks, of proactive research to get this information, it wasn't something that was easily accessible. So, in a way, I had the advantage of already being well informed.* [PERS24]

There were also some questions about the level of trust in the use of the internet in seeking health information. Health care

XSL•FO
RenderX

professionals (especially doctors) have remained a trustworthy source of information.

Communication

For EPR users, having access to their medical data makes them feel more comfortable communicating with health care professionals and moving toward a more transparent relationship, reducing the divide between medical doctors and patients.

> *The interaction between patient and professional is more on an equal footing. Today it's a black box, but having access to the data allows me to have confidence.* [PERS2]
>
> *And so in my opinion, if the electronic record becomes standard, it increases transparency between the patient and the doctor.* [PERS1]

It has been noted that the doctor-patient relationship is changing, yet it does not necessarily give more power to the patient:

> *I don't think the patient is really given more importance. They have more information, they have more knowledge. I can't say no, I don't take this medicine without listening, he suggested it to me because I think that with this information, I can't be the doctor. Yes, I think patients should interpret this as a more comprehensive source of information.* [PERS23]

As this relationship evolves, patients see the use of the EPR as a means of control, encouraging greater transparency between professionals and patients.

> *Doctors are much more challenged than they used to be, aren't they? People are looking for much more.* [PERS17]

The data analysis identified a second point related to communication, which is removing the barrier to patients' understanding of certain technical terms and professional jargon. Being able to read and look up specific terms can improve patients' overall understanding of their health (health literacy).

> *Yeah, that's the jargon that's used. Or else we don't take much time to explain it to you because we're pretty strict about our schedules. So if you don't have a basic understanding, you haven't retained anything, basically.* [PERS 22]

In summary, with the use of the EPR, communication between patients and professionals could be more transparent. However, for the interviewees, the concern is not necessarily related to the patient-professional relationship, but rather with interprofessional communication, which seems much more relevant.

Discussion

Principal Findings

This mixed methods study provides valuable insights into the characteristics, experiences, and perceptions of EPR users and nonusers in the canton of Vaud, Switzerland. Our findings reveal that early adopters of EPRs in Switzerland demonstrate distinct demographic and health profiles that both align with and diverge from patterns observed in previous international research.

Characteristics of Early Adopters

Our results indicate that early EPR adopters in French-speaking Switzerland are predominantly characterized by higher educational levels, chronic health conditions that require regular monitoring, and extensive health care networks. These individuals typically demonstrate higher levels of health literacy and were already actively engaged in managing their health information prior to the adoption of EPR, often through personal archiving systems. These findings align with previous studies that have identified chronic disease status as a significant predictor of EPR adoption [11,18].

However, contrary to international trends that typically show higher female participation in EPR adoption [19-21], our study found a significant male predominance among Swiss early adopters. This gender disparity may reflect broader patterns of digital technology engagement in Switzerland, where technical and IT fields continue to show marked gender disparities characterized by predominant male representation. This unexpected finding warrants further investigation to understand the sociocultural factors influencing gender-based differences in EPR adoption within the Swiss context.

Health Literacy and Digital Engagement

Our study reveals that early EPR adopters demonstrate higher levels of health literacy compared to both nonusers and the general Swiss population. This finding is consistent with previous research that highlights the relationship between health literacy and the adoption of digital health tools [22,23]. The proactive health information-seeking behaviors observed among early adopters suggest that these individuals perceive themselves as legitimate partners in patient-professional relationships, potentially transforming traditional hierarchical dynamics toward more horizontal communication patterns. Similar transformations have been observed in pediatric settings, where parents of chronically ill children demonstrate comparable engagement patterns [39].

Barriers to Adoption and Implementation

Despite recognizing potential benefits, both users and nonusers identified significant barriers to the optimal implementation of EPR. These include technical challenges (such as security procedures and interface design), limited professional participation, and fragmented implementation across the health care system. The critical role of health care professionals as facilitators or barriers to patient EPR adoption emerged as a consistent theme, with many participants perceiving negative attitudes toward EPRs among health care providers. This finding aligns with observations from France and Austria, where professional resistance has similarly hindered EPR implementation [40,41].

Our qualitative findings also revealed concerns about illness stigmatization influencing EPR adoption decisions, particularly for sensitive health conditions. Some participants expressed reluctance to include mental health or gynecological information in their EPRs due to perceived stigma. This observation aligns

XSL•FO
RenderX

with German research indicating that patients with stigmatized conditions demonstrate greater resistance to EPR adoption [42] and Norwegian findings showing more critical attitudes toward EPRs among patients with mental health conditions compared to those with physical disorders [43].

System Design and User Experience

The relatively low satisfaction rates among EPR users (30.2%) primarily stem from limitations in system design and implementation rather than the concept itself. EPR users expressed disappointment with the current EPR tool's functionality, citing a lack of intuitive design, poor document classification, and limited portability. These findings align with global experiences in EPR implementation, where user experience significantly impacts adoption rates and perceived effectiveness [39]. This misalignment of expectations represents a significant challenge for successful implementation.

Contextual Factors Influencing Adoption

Our study highlights the importance of considering sociopolitical contexts in EPR implementation. The Swiss opt-in approach to EPR registration, which makes adoption voluntary for citizens but mandatory for newly established health care providers since 2022, has created an implementation gap. This policy approach contrasts with opt-out strategies implemented in countries such as Austria, which have demonstrated different adoption patterns [40]. The influence of professional associations, particularly the perceived resistance from medical societies, emerged as a significant barrier to widespread adoption.

Additionally, our findings suggest that individual factors such as polymorbidity, emotional state, and social support networks influence EPR adoption decisions, consistent with Austrian research demonstrating correlations between these factors and EPR adherence [44].

Implications and Future Directions

Our findings have significant implications for the implementation of EPR in Switzerland and similar health care contexts. First, the distinct profile of early adopters suggests that current implementation approaches may be inadvertently creating digital health disparities, with higher-educated, male, chronically ill patients benefiting disproportionately. Targeted strategies to engage underrepresented populations, particularly those with lower health literacy and fewer digital skills, are essential for equitable implementation.

Second, the critical role of health care professionals in facilitating or hindering EPR adoption highlights the need for comprehensive professional engagement strategies. Future implementation efforts to foster greater support for EPR adoption should address professional concerns regarding workflow integration, information accuracy, and time constraints.

Third, the technical and design limitations identified by users suggest the need for user-centered redesign processes that incorporate feedback from diverse stakeholders. Improvements in interface design, document classification, and mobile accessibility could significantly enhance user satisfaction and adoption rates. Participants suggested a wide array of suggestions for improvement (Table 4), useful for policymakers and organizations responsible for implementation. Easier access to the platform (mobile app) and better organization and standardization of documents are needed to respond to the clients' needs.

Future research should explore the gender disparity observed in our study, investigate effective strategies for engaging underrepresented populations, and evaluate the impact of different policy approaches (opt-in vs opt-out) on EPR adoption patterns in the Swiss context.

Limitations

Several limitations should be considered when interpreting our findings. The significant gender imbalance in our sample, with male participants substantially overrepresented among both EPR users and study respondents, introduces potential systematic bias. Similarly, the low response rate among nonusers (8.2% compared to 30.4% among users) may have resulted in selection bias, potentially overrepresenting individuals with stronger opinions about EPRs or higher digital literacy. Furthermore, it would have been very interesting to compare respondents to nonrespondents to better understand the population of nonrespondents. However, due to restrictions by the ethics committee and data availability (eg, socioeconomic factors, comorbidities, and literacy), we could not explore these dimensions.

Furthermore, investigating health literacy and digital health literacy using standardized questionnaires (HLS19-Q12 and HLS19-DIGI-CH) can be criticized. A recent scoping review [45] identified 33 studies using digital health literacy tools; however, with several limitations, such as "narrow scope, limited adaptability to the rapidly evolving digital landscape, and lack of objective testing" (p. 11). The choice of the 2 instruments used in this study was based on the possibility to compare them to results from the national survey [31]. This choice aligns well with Thorup et al [45] recommendations to (1) combine health literacy and digital health literacy (multidimensional tool), (2) to use a validated tool, and (3) to use tools that are context specific.

Our qualitative analysis, while providing rich contextual insights, is inherently subjective and may reflect researchers' interpretations. Despite efforts to ensure analytical rigor through team-based analysis and reflexive practices, alternative interpretations of the data remain possible.

Finally, our study focused on patients with care episodes at a university hospital, which may limit its generalizability to community-based settings or rural populations. The experiences and perceptions of individuals without recent hospital encounters may differ substantially from those captured in our sample.

Conclusion

This study provides valuable insights into the characteristics, experiences, and perceptions of early EPR adopters in the canton of Vaud, Switzerland. Our findings reveal that these early adopters constitute a distinct population characterized by higher education levels, chronic health conditions, extensive health care networks, and higher health literacy. Despite moderate

XSL•FO
RenderX

satisfaction with the current EPR implementation, most users recommend the system to others, suggesting belief in its potential value despite existing limitations.

The observed demographic patterns raise important questions about EPR accessibility and potential digital health disparities. The significant male predominance among Swiss early adopters contrasts with international trends and warrants further investigation. Additionally, the higher health literacy levels among adopters suggest that current implementation approaches may inadvertently favor already advantaged populations.

Our mixed methods approach revealed that moderate satisfaction among users primarily stems from limited system uptake across the health care ecosystem rather than inherent opposition to the concept. The optimal effectiveness of EPRs appears contingent upon systemic integration throughout the ambulatory care network, suggesting that extending implementation beyond university hospitals is essential for achieving the intended goals of care continuity and coordination.

As Switzerland continues its digital health transformation, addressing the technical, professional, and policy barriers identified in this study will be crucial for realizing the full potential of EPRs to enhance health care quality, coordination, and patient engagement.

Acknowledgments

The authors wish to thank all participants for their time and valuable insights. We also acknowledge Jonas Pignolet for his assistance during the recruitment phase. Special thanks to Lausanne University Hospital (CHUV) communication department for help in promotion and dissemination.

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy [46], the following tasks were delegated to GAI tools under full human supervision: initial transcription (audio to text), idea generation (for the Discussion section to explore existing literature and theoretical frameworks), proofreading and editing, translation of citations from French to English, and reformatting.

The GAI tool used was Claude Sonnet 4.5 (Anthropic). Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by CL.

Funding

The study was supported by the Research Commission of Health Domain (HES-SO University of Applied Sciences and Arts Western Switzerland), by the School of Health Sciences – Vaud HESAV /HES-SO and the Lausanne University Hospital (CHUV).

Data Availability

The datasets generated and analyzed during this study are not publicly available due to the confidential nature of the data but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: FB, VS
Funding acquisition: FB, VS
Methodology: FB, VS
Data curation: FB, VS
Investigation: FB, CL, MR
Data analysis: FB, CL, MR, VS
Validation: FB, CL, VS
Visualization: FB, CL, VS
Project administration: VS
Supervision: FB, VS
Writing – original draft: FB, CL, VS
Writing – review & editing: FB, CL, MR, VS

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire translated in English.
[PDF File, 695 KB - jmir_v28i1e83702_app1.pdf]

Multimedia Appendix 2

XSL•FO
RenderX

Interview guide translated in French.
[PDF File, 475 KB - jmir_v28i1e83702_app2.pdf]

References

1. Lupton D. M-health and health promotion: the digital cyborg and surveillance society. Soc Theory Health 2012 Aug;10(3):229-244. [doi: 10.1057/sth.2012.6]
2. Lupton D. Quantifying the body: monitoring and measuring health in the age of mHealth technologies. Crit Public Health 2013 Dec;23(4):393-403. [doi: 10.1080/09581596.2013.794931]
3. LDEP – the law that governs the DEP. Federal Office of Public Health FOPH. URL: https://www.bag.admin.ch/fr/ldep-la-loi-qui-regit-le-dep [accessed 2015-06-19]
4. Monod S, Cavalli V, Pin S, Grandchamp C. Système de santé suisse: y a-t-il un pilote dans la machine? Rev Med Suisse 2023;19(819):583-588. [doi: 10.53738/REVMED.2023.19.819.583]
5. Federal Health Insurance Act (LAMal). The federal law publishing platform. URL: https://www.fedlex.admin.ch/eli/cc/1995/1328_1328_1328/fr [accessed 2026-03-03]
6. Monod S, Pin S, Grandchamp C, Mariétan X, Courvoisier N. Analysis of the governance of the swiss healthcare system and proposal for a federal law on health [Article in French]. University Centre for General Medicine and Public Health (Unisanté). 2024. URL: https://www.unisante.ch/fr/formation-recherche/recherche/publications/raisons-sante/raisons-sante-354 [accessed 2025-11-30]
7. With CARA, access your health data. CARA. 2025. URL: https://www.cara.ch/ [accessed 2026-03-03]
8. Heart T, Ben-Assuli O, Shabtai I. A review of PHR, EMR and EHR integration: a more personalized healthcare and public health policy. Health Policy Technol 2017 Mar;6(1):20-25. [doi: 10.1016/j.hlpt.2016.08.002]
9. Electronic patient record: the EPR. eHealth Suisse. URL: https://www.e-health-suisse.ch/en/coordination/electronic-patient-record/the-epr [accessed 2026-03-03]
10. Schmitt T. New governance of the digital health agency: a way out of the joint decision trap to implement electronic health records in Germany? Health Econ Policy Law 2024 Apr;19(2):269-288. [doi: 10.1017/S1744133123000142] [Medline: 37694342]
11. von Kalckreuth N, Feufel MA. Disease characteristics influence the privacy calculus to adopt electronic health records: a survey study in Germany. Digit Health 2024;10:20552076241274245. [doi: 10.1177/20552076241274245] [Medline: 39247096]
12. Burnel P. The introduction of electronic medical records in France: more progress during the second attempt. Health Policy 2018 Sep;122(9):937-940. [doi: 10.1016/j.healthpol.2018.07.022] [Medline: 30100529]
13. E-Health records. e-Estonia. URL: https://e-estonia.com/solutions/e-health-2/e-health-records/ [accessed 2026-03-03]
14. Haverinen J, Keränen N, Tuovinen T, Ruotanen R, Reponen J. National development and regional differences in eHealth maturity in Finnish public health care: survey study. JMIR Med Inform 2022 Aug 12;10(8):e35612. [doi: 10.2196/35612] [Medline: 35969462]
15. Raab R, Küderle A, Zakreuskaya A, et al. Federated electronic health records for the European Health Data Space. Lancet Digit Health 2023 Nov;5(11):e840-e847. [doi: 10.1016/S2589-7500(23)00156-5] [Medline: 37741765]
16. Sartain SA, Stressing S, Prieto J. Patients' views on the effectiveness of patient-held records: a systematic review and thematic synthesis of qualitative studies. Health Expect 2015 Dec;18(6):2666-2677. [doi: 10.1111/hex.12240] [Medline: 25059439]
17. Tapuria A, Porat T, Kalra D, Dsouza G, Xiaohui S, Curcin V. Impact of patient access to their electronic health record: systematic review. Inform Health Soc Care 2021 Jun 2;46(2):192-204. [doi: 10.1080/17538157.2021.1879810] [Medline: 33840342]
18. Greenberg AJ, Falisi AL, Finney Rutten LJ, et al. Access to electronic personal health records among patients with multiple chronic conditions: a secondary data analysis. J Med Internet Res 2017 Jun 2;19(6):e188. [doi: 10.2196/jmir.7417] [Medline: 28576755]
19. Foster B, Krasowski MD. The use of an electronic health record patient portal to access diagnostic test results by emergency patients at an academic medical center: retrospective study. J Med Internet Res 2019 Jun 28;21(6):e13791. [doi: 10.2196/13791] [Medline: 31254335]
20. Ochoa C, Baron-Lee J, Popescu C, Busl KM. Electronic patient portal utilization by neurology patients and association with outcomes. Health Informatics J 2020 Dec;26(4):2751-2761. [doi: 10.1177/1460458220938533] [Medline: 32674698]
21. Tuan WJ, Mellott M, Arndt BG, Jones J, Simpson AN. Disparities in use of patient portals among adults in family medicine. J Am Board Fam Med 2022;35(3):559-569. [doi: 10.3122/jabfm.2022.03.210486] [Medline: 35641056]
22. Lester M, Boateng S, Studeny J, Coustasse A. Personal health records: beneficial or burdensome for patients and healthcare providers? Perspect Health Inf Manag 2016;13(Spring):1h. [Medline: 27134613]
23. Wildenbos GA, Peute L, Jaspers M. Facilitators and barriers of electronic health record patient portal adoption by older adults: a literature study. Stud Health Technol Inform 2017;235:308-312. [Medline: 28423804]
24. Ivankova NV, Creswell JW, Stick SL. Using mixed-methods sequential explanatory design: from theory to practice. Field methods 2006 Feb;18(1):3-20. [doi: 10.1177/1525822X05282260]

XSL•FO
RenderX

25. O'Cathain A, Murphy E, Nicholl J. Why, and how, mixed methods research is undertaken in health services research in England: a mixed methods study. BMC Health Serv Res 2007 Jun 14;7(1):85. [doi: 10.1186/1472-6963-7-85] [Medline: 17570838]
26. Rauscher L, Greenfield BH. Advancements in contemporary physical therapy research: use of mixed methods designs. Phys Ther 2009 Jan;89(1):91-100. [doi: 10.2522/ptj.20070236] [Medline: 19008328]
27. Hanson WE, Creswell JW, Clark VLP, Petska KS, Creswell JD. Mixed methods research designs in counseling psychology. J Couns Psychol 2005 Apr;52(2):224-235. [doi: 10.1037/0022-0167.52.2.224]
28. Sandelowski M. Combining qualitative and quantitative sampling, data collection, and analysis techniques in mixed-method studies. Res Nurs Health 2000 Jun;23(3):246-255. [doi: 10.1002/1098-240x(200006)23:3<246::aid-nur9>3.0.co;2-h] [Medline: 10871540]
29. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
30. WHO action network on measuring population and organizational health literacy. M-POHL. 2023. URL: https://m-pohl.net/ [accessed 2025-09-07]
31. De Gani S, Jaks R, Bieri U, Kocher J. Health literacy. Federal Ofiice of the Public Health FOPH. URL: https://www.bag.admin.ch/en/health-literacy#Swiss-Health-Literacy-Survey-2019-21 [accessed 2025-09-07]
32. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
33. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Aug 8;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
34. Byrne D. A worked example of Braun and Clarke's approach to reflexive thematic analysis. Qual Quant 2022 Jun;56(3):1391-1412. [doi: 10.1007/s11135-021-01182-y]
35. Murphy E, Dingwall R, Greatbatch D, Parker S, Watson P. Qualitative research methods in health technology assessment: a review of the literature. Health Technol Assess 1998;2(16):1-274. [Medline: 9919458]
36. Finlay L. "Outing" the researcher: the provenance, process, and practice of reflexivity. Qual Health Res 2002 Apr;12(4):531-545. [doi: 10.1177/104973202129120052] [Medline: 11939252]
37. Turer RW, DesRoches CM, Salmi L, Helmer T, Rosenbloom ST. Patient perceptions of receiving COVID-19 test results via an online patient portal: an open results survey. Appl Clin Inform 2021 Aug;12(4):954-959. [doi: 10.1055/s-0041-1736221] [Medline: 34644805]
38. Steitz BD, Turer RW, Lin CT, et al. Perspectives of patients about immediate access to test results through an online patient portal. JAMA Netw Open 2023 Mar 1;6(3):e233572. [doi: 10.1001/jamanetworkopen.2023.3572] [Medline: 36939703]
39. Sipanoun P, Oulton K, Gibson F, Wray J. The experiences and perceptions of users of an electronic patient record system in a pediatric hospital setting: a systematic review. Int J Med Inform 2022 Apr;160:104691. [doi: 10.1016/j.ijmedinf.2022.104691] [Medline: 35091287]
40. Griesser A, Mzoughi M, Bidmon S, Cherif E. How do opt-in versus opt-out settings nudge patients toward electronic health record adoption? An exploratory study of facilitators and barriers in Austria and France. BMC Health Serv Res 2024 Apr 8;24(1):439. [doi: 10.1186/s12913-024-10929-w] [Medline: 38589922]
41. Griesser A, Bidmon S. A holistic view of facilitators and barriers of electronic health records usage from different perspectives: a qualitative content analysis approach. HIM J 2024 Sep;53(3):227-236. [doi: 10.1177/18333583231178611]
42. Von Kalckreuth N, Prümper AM, Feufel MA. The influence of health data on the use of the electronic health record (EHR)-a mixed methods approach. 2023 Presented at: Americas Conference on Information Systems (AMCIS); Aug 9, 2023 URL: https://aisel.aisnet.org/amcis2023/sig_health/sig_health/2/ [accessed 2026-03-03]
43. Wang B, Kristiansen E, Fagerlund AJ, et al. Users' experiences with online access to electronic health records in mental and somatic health care: cross-sectional study. J Med Internet Res 2023 Dec 25;25:e47840. [doi: 10.2196/47840] [Medline: 38145466]
44. Halmdienst N, Pruckner GJ, Winter-Ebmer R. Complexities of health and acceptance of electronic health records for the Austrian elderly population. Eur J Health Econ 2023 Feb;24(1):53-66. [doi: 10.1007/s10198-022-01451-z] [Medline: 35287235]
45. Thorup CB, Uitto M, Butler-Henderson K, et al. Choosing the best digital health literacy measure for research: mixed methods study. J Med Internet Res 2025 Apr 8;27:e59807. [doi: 10.2196/59807] [Medline: 40198098]
46. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (Generative AI Delegation Taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res 2025 Aug 8:1-27. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]

Abbreviations

CHUV: Lausanne University Hospital
CONSORT: Consolidated Standards of Reporting Trials
EHR: electronic health record

XSL•FO
RenderX

EMR: electronic medical record
EPR: electronic patient record
GP: general practitioner
HLS19-DIGI: Digital Health Literacy Survey
HLS19-Q12: Health Literacy Survey
LDEP: the Federal Law on Electronic Patient Record
REDCap: Research Electronic Data Capture

Edited by A Stone; submitted 13.Sep.2025; peer-reviewed by H Huang, JA Rolnick; revised version received 06.Feb.2026; accepted 06.Feb.2026; published 21.Apr.2026.

Please cite as:
Bastardot F, LaRocco C, Robert M, Schoeb V
Usage and Perceptions of Electronic Patient Records Experienced by Users and Nonusers in the Canton of Vaud, Switzerland: Mixed Methods Study
J Med Internet Res 2026;28:e83702
URL: https://www.jmir.org/2026/1/e83702
doi:10.2196/83702

© François Bastardot, Catherine LaRocco, Mathilde Robert, Veronika Schoeb. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effects of a Soft Robotic Exoskeleton for Gait Training on Clinical Outcomes in Patients With Parkinson Disease: Randomized Controlled Pilot Study

Shiying Bu, MA; Xiaoqing Luo, MA; Shuangfang Li, MA; Yuqin Shi, MA; Lingjing Jin, MD, PhD; Qing Zhao, MD, PhD

Shanghai YangZhi Rehabilitation Hospital (Shanghai Sunshine Rehabilitation Center), Tongji University School of Medicine, 2209 Guang Xing Road, Songjiang District, Shanghai, China

Corresponding Author:
Qing Zhao, MD, PhD
Shanghai YangZhi Rehabilitation Hospital (Shanghai Sunshine Rehabilitation Center), Tongji University School of Medicine, 2209 Guang Xing Road, Songjiang District, Shanghai, China

Abstract

Background: Balance and gait disorders in Parkinson disease (PD) impair motor function and quality of life.

Objective: Evidence on soft exoskeleton robots (SERs) for PD rehabilitation is limited. This study evaluated the impact of SERs on motor dysfunction in PD.

Methods: A total of 56 people with PD (July 2023 to May 2024) were randomized to 2 groups: the control group (n=25, 44.6%) received conventional rehabilitation, and the experimental group (n=31, 55.4%) received conventional rehabilitation combined with SER training (ChiCTR2500111990). Training occurred 5 times per week for 20 minutes each session over 4 weeks. Primary outcomes included gait speed and stride length, while secondary outcomes assessed the percentage of swing phase, ankle joint range of motion, Unified Parkinson Disease Rating Scale total and motor scores, and Montreal Cognitive Assessment. Paired sample *t* tests (2-tailed) were used for within-group pre- and postintervention comparisons, and independent sample *t* tests (2-tailed) were used for between-group comparisons. Correlation analyses were conducted between gait parameters and improvements in ankle mobility.

Results: After 4 weeks, the experimental group showed significant improvements in gait and balance. Specifically, left stride length increased by a mean of 0.15 (SD 0.16; 95% CI 0.09 - 0.21) m (P<.001), right stride length by a mean of 0.15 (SD 0.15; 95% CI 0.10 - 0.21) m (P<.001), left ankle dorsiflexion by a mean of 2.84 (SD 1.46; 95% CI 2.32 - 3.36) degrees (P<.001), left swing phase percentage by a mean of 1.56% (SD 3.05%; 95% CI 0.44 - 2.68; P=.01), and right swing phase percentage by a mean of 1.6% (SD 2.72%; 95% CI 0.62 - 2.62; P=.002). The Unified Parkinson Disease Rating Scale Part III total score decreased by a mean of 2.80 (SD 3.98) points, and balance subscale scores decreased by a mean of 0.40 (0.58) points (P<.001). Montreal Cognitive Assessment scores increased by a mean of 1.23 (1.23; 95% CI 0.77 - 1.68) points (P<.01), and Barthel Index scores increased by a mean of 6.84 (7.14; 95% CI 4.22 - 9.46) points (P<.001). Other measures such as balance reaction time, reaction speed, maximum movement distance, and movement direction control showed significant improvement (P<.01). Compared to the control group, the experimental group demonstrated greater improvements in gait speed (P=.04), balance reaction time (P=.04), and maximum movement distance (P=.048). Correlation analysis revealed that improvements in left ankle dorsiflexion were positively correlated with improvements in gait speed, stride length, and swing phase duration (P<.05).

Conclusions: SER-assisted training significantly improves gait, balance, and PD symptoms. Our work integrates multidimensional assessments (gait analysis, balance metrics, and clinical scales) and reveals that gains in ankle mobility directly correlate with gait improvements, suggesting a key mechanism. This study contributes by establishing SER as an effective adjunct to conventional therapy, supported by comprehensive quantitative data.

Trial Registration: ChiCTR Registry ChiCTR2500111990; https://www.chictr.org.cn/hvshowprojectEN.html?id=287451

(J Med Internet Res 2026;28:e82629) doi:10.2196/82629

KEYWORDS

soft exoskeleton robot; walking training; Parkinson disease; balance function; cognitive function; activities of daily living

XSL•FO
RenderX

Introduction

Parkinson disease (PD) is the second most common neurodegenerative disorder, with major motor symptoms including bradykinesia, resting tremor, rigidity, and postural-gait disturbances, which significantly affect mobility and quality of life [1,2]. Gait dysfunction is a prevalent manifestation of PD, with more than 70% of patients showing abnormalities such as reduced step length, freezing of gait, and diminished arm swing within 5 years of diagnosis. The incidence of freezing of gait increases from 37.9% in early-stage PD (disease duration ≤5 y) to 64.6% in late-stage PD (disease duration ≥9 y), reflecting the progressive decline in gait function [3]. Treatment outcomes for PD-related gait disorders remain unsatisfactory, with substantial variability in the effectiveness of pharmacological treatments and surgical interventions, such as deep brain stimulation [4-6]. Patients also frequently experience postural instability, which is marked by abnormal balance reflexes, reduced stability limits, and disrupted compensatory strategies [7]. A study comparing balance function in healthy individuals and people with PD (H&Y stages 1 - 3) using clinical tests and postural recording found that people with PD had significantly reduced maximum movement distance and slower movement speed during the stability limit test compared to healthy controls ($P<.05$) [8]. Additionally, the stability limits during static standing were significantly smaller in people with PD than in the healthy controls ($P<.05$) [8]. These objective abnormalities confirmed a significant impairment in postural control in this population. A multicenter cohort study revealed that 34% of people with PD developed detectable balance deficits within 2 years of diagnosis, which increased to 71% at the 10-year follow-up and reached 92% after 15 years [9]. Biomechanical analysis showed reduced ankle joint torque and decreased trunk counterrotation during balance recovery tasks, indicating impaired use of ankle strategies compared to healthy individuals [10].

The pathological mechanisms underlying gait disturbances in PD stem from dopaminergic depletion within the basal ganglia-thalamocortical circuit, resulting in impaired automatic motor control [11]. Under normal physiological conditions, rhythmic gait initiation relies on the coordination of the nigrostriatal pathway to synchronize limb movement and facilitate center-of-mass transfer [12]. In people with PD, the reduced activity of the direct pathway and excessive activation of the indirect pathway lead to enhanced GABAergic inhibition of the pedunculopontine nucleus, clinically manifesting in characteristic Parkinsonian gait abnormalities such as shortened stride length, loss of reciprocal arm swing, and motor freezing [13]. To compensate, patients adopt a high-frequency shuffling gait strategy, which inadvertently exacerbates energy expenditure and postural instability. The pathological mechanisms of postural control deficits involve degeneration of the basal ganglia-brainstem-cerebellar pathway, leading to suppressed vestibulospinal reflexes. This results in axial rigidity and reduced joint range of motion (muscle stiffness and bradykinesia), while damage to the dopaminergic system impairs sensorimotor integration, particularly when visual and vestibular information conflict. In such cases, patients exhibit increased postural sway and reduced stability limits [14-16]. These mechanisms collectively contribute to the progressive worsening of gait parameters in people with PD.

In recent years, soft robotic exoskeletons (SREs) have rapidly entered clinical applications due to their lightweight, real-time performance, and ease of use. By using biomimetic lightweight drive designs, SREs maintain joint mobility while reducing metabolic costs, overcoming the limitations of rigid exoskeletons [6]. Representative flexible exoskeleton robots include the Myosuit from the United States, which was tested by a Swiss team in a study where 12 participants with various movement disorders, each capable of walking at least 10 meters, underwent 5 sessions of 45-minute Myosuit training. Five participants showed a significant increase in walking speed postintervention compared to baseline levels [17]. Another example is Japan's wearable Stride Management Assist exoskeleton robot, which was studied in a trial involving 8 individuals with PD at H-Y stages 1 - 2.5. The study involved robot-assisted walking training conducted for 4 weeks, 45 minutes per session, 3 times a week. The results showed improvements in walking speed, stride length, and hip joint mobility, along with a decrease in gait time and improvements in long-range autocorrelation measures associated with PD gait variability. Additionally, hip joint maximum peak torque and energy consumption decreased, while confidence in balance during activity improved [18]. In China, the Yrobot RelinkTM muscle exoskeleton robot was evaluated in a study by Xie et al, which demonstrated that it was superior to conventional training in improving walking speed and recovery in the 10-m walk test and 6-minute walk test for patients with subacute stroke. All these exoskeletons simulate the working principles of human muscles to assist joints and provide adaptive power assistance to the body [19]. Preliminary evidence suggests that SRE-assisted training can improve gait parameters in patients with neurological diseases. However, a meta-analysis of rigid exoskeleton interventions for PD (14 RCTs, n=572) shows significant improvements in balance function (Berg Balance Scale) [20], although the efficacy of SREs specifically designed for PD gait remains unclear.

It is noteworthy that existing studies on exoskeleton-assisted gait training in PD have demonstrated significant heterogeneity in therapeutic effects, which may be related to differences in device types and intervention protocols. This pilot randomized controlled trial (RCT) aimed to investigate the clinical efficacy of an ankle-targeted SRE system in improving gait abnormalities and balance impairments in people with PD.

Methods

Study Setting

This study is an RCT conducted from July 2023 to May 2025 at Shanghai Yangzhi Rehabilitation Hospital in China, which screened and recruited 56 patients with PD.

Participants

The inclusion criteria were (1) idiopathic PD, (2) Hoehn-Yahr stage 2 - 3, (3) evident gait abnormalities, (4) age between 50 and 80 years, and (5) stable medication, with no changes

XSL•FO
RenderX

expected in the next 3 months. The exclusion criteria were (1) severe psychiatric disorders; (2) other terminal illnesses, including but not restricted to cancer; (3) history of substance abuse or alcoholism; (4) inability to understand or cooperate with the study; (5) pregnant or lactating women; and (6) participation in other clinical trials. All participants were fully informed of the study content and potential risks and voluntarily signed the informed consent form to confirm their participation.

Ethical Considerations

This study was approved by the Ethics Review Committee of Shanghai Yangzhi Rehabilitation Hospital. The study protocol was approved by the Yangzhi Institutional Review Board (IRB). The authors have obtained a letter from the IRB confirming that the protocol and design of the submitted manuscript are identical to the version that received approval before the trial started. This confirmation has been mentioned here and provided to the editorial office for review.

This study was reviewed and approved by the relevant IRB to ensure that all procedures involving human subjects adhered to ethical guidelines. Informed consent was obtained from all participants for the primary data collection, and for secondary analyses, the original informed consent explicitly permits such use without requiring additional consent. All participant data were anonymized and deidentified to safeguard privacy and confidentiality. In cases where identifiable data were used, appropriate measures such as data encryption and restricted access were implemented to ensure security. No compensation was provided to participants for their involvement. Furthermore, all images or supplementary materials included in the manuscript do not identify any individuals. The photograph included in this manuscript shows 2 individuals, but no identifying features (eg, face, name, or other personal information) are visible, and they cannot be identified from the image.

Trial Registration

This trial was retrospectively registered with the Chinese Clinical Trial Registry (ChiCTR) [21] under registration number ChiCTR2500111990.

Reporting Guidelines

This RCT is reported in accordance with the CONSORT (Consolidated Standards of Reporting Trials) 2025 statement (Checklist 1) [22] and eHealth guidelines [23].

Sample Size

The sample size calculation was based on data from previous trials, with walking speed as the primary efficacy end point. A 2-tailed significance level of 0.05 and a power of 0.9 were used [24]. The sample size was calculated using the PASS15 sample size formula [25,26], resulting in a required sample size of 26 participants for the experimental group and 26participants for the control group. Assuming a 20% dropout rate, the final sample size for each group was adjusted to 32 participants.

Randomization

Sequence Generation

After the baseline assessment, people with PD were randomly assigned to the SRE group or conventional rehabilitation (CR) group in a 1:1 ratio using a computer-generated random number table.

Allocation Concealment Mechanism

The allocation sequence was stored in a sealed envelope by an investigator who was not involved in training or assessment.

Implementation

Upon the registration of a new eligible participant, an envelope was randomly selected, and the assigned group was communicated to the therapist.

Blinding

Both pre- and postintervention assessments were conducted by the same certified therapist, who was blinded to the group allocation. To ensure interrater reliability, a second certified therapist also conducted parallel assessments, and the scores from both therapists were compared. This process was implemented to maintain consistency and minimize bias in the subjective outcome measures.

Interventions

This study used the Yrobot Relink flexible exoskeleton robot (Yrobot Inc) [27] to assist in gait rehabilitation for people with PD. The device provides support to the ankle joint via a carbon fiber frame and mechanical drive cables, allowing for 30 degrees of ankle dorsiflexion and plantarflexion. The weighted design at the waist ensures the overall weight is only 3.3 kg. Its intelligent sensing system includes high-precision motion sensors (sampling data 100 times per second) and mechanical sensing elements, which continuously monitor the leg's movement angle (with an error of less than 0.5 degrees), walking rhythm, and the interaction forces between the device and the body. Through a data fusion algorithm, the system accurately identifies the key phases in the walking cycle that require assistance (with an accuracy of 92%).

Using an intelligent learning algorithm, the system automatically analyzes muscle electrical signals and changes in plantar pressure during walking, generating support forces that align with natural human movement during the later stages of gait (from heel-off to toe-off). The therapist can adjust the level of assistance in real time via a tablet and monitor gait symmetry indicators. The device is specifically designed with an anti-foot-drop safety mechanism, using a spring mechanism to restrict abnormal foot inversion, ensuring movement safety during use (Figure 1). Participants underwent a standardized gait training protocol, which was conducted in 3 phases: initial assessment—each participant's baseline gait parameters (speed, stride length, and symmetry) were measured using a 3D motion capture system; training sessions—during each session, the level of assistance was dynamically adjusted based on real-time feedback from the system's sensors, with the therapist modifying the support intensity according to the patient's performance; and postassessment—after the training sessions, participants' gait performance was reassessed, and changes in gait parameters, such as improvement in walking speed, step length, and symmetry, were recorded.

XSL•FO
RenderX

Figure 1. Yrobot Relink soft robotic exoskeleton used for soft robotic exoskeleton–assisted overground walking training.

Training Protocol

Both the CR group and the SRE group underwent a 4-week walking training program, with 5 sessions per week, each lasting 20 minutes, totaling 20 sessions. The assistive force of the exoskeleton (in Newtons) was set according to the participant's body weight (in kilograms). Patients underwent evaluation and walking training 1 to 2 hours after medication intake to ensure they were in the "on" phase. The CR group received physical therapy and occupational therapy at the inpatient rehabilitation center, which included standard training in upper and lower limb strength, transfers, balance, walking, and activities of daily living. All physical therapy sessions were conducted by 2 rehabilitation physical therapists who had been trained in the intervention protocol prior to the study. During each training session, the physical therapists walked on either side of the participant to ensure safety (Figure 2).

XSL•FO
RenderX

Figure 2. Photo of the soft robotic exoskeleton–assisted overground walking training procedure used in the study, which involves participants wearing the Yrobot Relink and engaging in controlled walking exercises designed to enhance gait stability and motor function in patients with Parkinson disease. The study was conducted over a 4-wk period at Shanghai Yangzhi Rehabilitation Hospital, with a total of 56 participants.

Outcome Measures

The researchers assessed the relevant indicators of both groups of participants at 3 time points: at baseline (before training), after 10 training sessions, and after 20 training sessions. The evaluations were performed by a physical therapist who was blinded to the specific group assignments and did not participate in any aspect of the training intervention.

- Motor function evaluation: Unified Parkinson Disease Rating Scale Part III (UPDRS-III) and its balance assessment.
- Nonmotor function evaluation: Montreal Cognitive Assessment and Modified Barthel Index.
- Balance ability evaluation: The limits of stability test was used to assess the participant's ability to shift their center

XSL•FO
RenderX

of gravity along the furthest possible boundary on the Balance Master system (NeuroCom International). Before the test, participants were instructed to extend their hips and knees to maintain an upright posture, with their feet stationary and hands placed vertically at their sides. During the test, participants were asked to lean their body in the directions of the computer screen (front, back, left, and right) for 10 seconds each, maintaining balance without falling. Limits of stability indicators include reaction time (RT), mean angular velocity, maximum center of pressure displacement, and directional control (DCL). The maximum center of pressure displacement is defined as the farthest distance the participant's center of pressure moves during the tilt and is measured as the distance required to complete the task. Scores range from 0% to 100%, with higher percentages indicating better balance. DCL is assessed by measuring the amount of movement toward the target, compared to movements away from the target (external movement), and is defined as the ratio of the difference between expected movement and external movement to the expected movement. An integrated score across 8 directions was used for analysis. Scores range from 0% to 100%, with higher percentages indicating better motor control. The Sensory Organization Test includes a series of test conditions that assess stability and instability in both visual environments (open and closed eyes) and different states of the balance platform (fixed or mobile). These conditions are categorized as follows: condition 1 (eyes open, platform stable), condition 2 (eyes closed, platform stable), condition 3 (eyes open, platform stable, visual scene moves with center of gravity), condition 4 (eyes open, platform moves with center of gravity), condition 5 (eyes closed, platform moves with center of gravity), and condition 6 (eyes open, both platform and visual scene move with center of gravity) [5]. During the test, participants complete 6 progressive tasks under the guidance of the system. Each task specifically simulates the dominance of a single sensory system (eg, vision, proprioception, or vestibular sense) to quantify its role in dynamic balance control and compensatory effectiveness. The system automatically calculates the balance score based on the peak of the actual sway angle and theoretical sway angle, with higher scores reflecting better balance ability.

- Gait spatiotemporal parameters and kinematic data: The Yrobot Relink flexible exoskeleton robot collects spatiotemporal parameters (including step length, step speed, and swing phase proportion) and ankle joint angle indicators during participants' self-paced walking.

Statistical and Minimal Important Difference Analyses

Data analysis was performed using IBM SPSS Statistics (version 25.0; SPSS Inc). All continuous variables underwent normality tests. Data that followed a normal distribution were expressed as mean (SD), and intergroup comparisons were made using parametric tests. Nonnormally distributed data were presented as median (IQR), and categorical variables were expressed as counts (percentages), with intergroup comparisons made using nonparametric tests.

Within-group comparisons were conducted using paired sample *t* tests, while independent sample *t* tests were used for between-group comparisons. To explore the correlation between changes in ankle joint improvement after intervention and changes in primary clinical outcomes, correlation analyses were performed using Spearman rank correlation and Pearson chi-square tests. A P value of <.05 was considered statistically significant.

In addition to statistical analysis, the clinical significance of changes in primary clinical outcomes was evaluated by comparing these changes with the minimal clinically important difference (MCID). On the basis of previous research on gait interventions for patients with PD, the MCID threshold for walking speed change before and after intervention was defined as 0.05 m/s, which corresponds to a small meaningful change [28]. A significance level of α of .05 was used for all statistical analyses.

Results

Patient Characteristics

From July 2023 to May 2025, a total of 72 eligible patients were included in this study, with 16 (22.2%) patients excluded due to failure to meet the inclusion criteria and other reasons. Among the participants, 28 were men, and 28 were women. Statistical analysis of the baseline characteristics of both groups showed no significant differences (P>.05), indicating comparability between the groups, as presented in Table 1 and Figure 3. No serious adverse events occurred during the intervention for either group.

XSL•FO
RenderX

Table . General information of the participants.

Characteristics	SRE[a] group (n=25)	CR[b] group (n=31)	*P* value
Age (y), mean (SD)	66.76 (7.08)	70 (65,74)	.31
Height (cm), mean (SD)	167.20 (8.46)	166.26 (7.96)	.67
Weight (kg), mean (SD)	61.80 (6.68)	64 (53,65)	.59
H-Y[c] (stage), mean (range)	3.0 (2.75-3.0)	3.0 (2.5-3.0)	.38
BMI (kg/m^2), mean (SD)	22.11 (1.96)	21.89 (2.22)	.70
Comorbidities, n (%)			
Postoperative DBS[d]	5 (19.2)	6 (22.2)	.79
Diabetes mellitus	4 (15.4)	4 (14.8)	.85
Hypertension	6 (23.1)	9 (33.3)	.41
Cerebrovascular diseases	5 (19.2)	5 (18.5)	.95
Joint stiffness	12 (46.2)	14 (53.8)	.43
Levodopa-equivalent dose (mg/d), mean (SD)	332.7 (178.8)	369.0 (157.3)	.43
UPDRS[e]-III (score), mean (SD)	28.64 (11.20)	30.19 (11.97)	.62
UPDRS-III balance (score), mean (SD)	2.32 (1.28)	2.65 (1.11)	.31
Gait velocity (m/s)	Mean 0.55 (SD 0.27)	Median 0.52 (IQR 0.36-0.64)	.80
Left step length (m), mean (SD)	0.52 (0.28)	0.56 (0.24)	.59
Right step length (m), mean (SD)	0.54 (0.27)	0.56 (0.24)	.83
Left swing phase (%)	Median 27.50 (IQR 23.62-28.16)	Mean 25.36 (SD 4.75)	.35
Right swing phase (%), mean (SD)	26.43 (3.78)	26.07 (4.10)	.74
Left ankle plantarflexion (degree), mean (SD)	45.36 (7.65)	46.47 (7.88)	.87
Left ankle dorsiflexion (degree), mean (SD)	8.93 (3.36)	9.20 (2.54)	.82
Right ankle plantarflexion (degree), mean (SD)	46.00 (8.29)	47.87 (6.94)	.59
Right ankle dorsiflexion (degree), mean (SD)	8.79 (4.00)	10.00 (2.30)	.13
COP-Max[f] (%), mean (SD)	51.96 (13.47)	51.55 (20.40)	.58
Reaction time (s), mean (SD)	1.35 (0.46)	1.47 (0.50)	.49
Angular velocity (degree per second), mean (SD)	1.94 (1.02)	2.20 (1.14)	.83
Directional control (%), mean (SD)	63.04 (13.47)	64.24 (15.73)	.83
SOT[g] composite (score), mean (SD)	62.44 (10.85)	64.80 (11.28)	.79
BBS[h] (score), mean (SD)	41.12 (6.05)	40.45 (9.36)	.75
Barthel Index (score), median (IQR)	77 (70-91)	85 (65-100)	.32
MoCA[i] (score), mean (SD)	20.80 (5.20)	21.03 (5.14)	.87

[a]SRE: soft robotic exoskeleton device.

[b]CR: conventional rehabilitation.

[c]H-Y: Hoehn-Yahr.

[d]DBS: deep brain stimulation.

[e]UPDRS: Movement Disorder Society-Unified Parkinson disease rating scale.

[f]COP-Max: maximum center of pressure displacement.

XSL•FO
RenderX

[g]SOT: Sensory Organization Test.

[h]BBS: Berg Balance Scale.

[i]MoCA: Montreal Cognitive Assessment

Figure 3. Research intervention method: Participants underwent a structured 4-wk program, with sessions held 5 times per week. SRE: soft robotic exoskeleton.

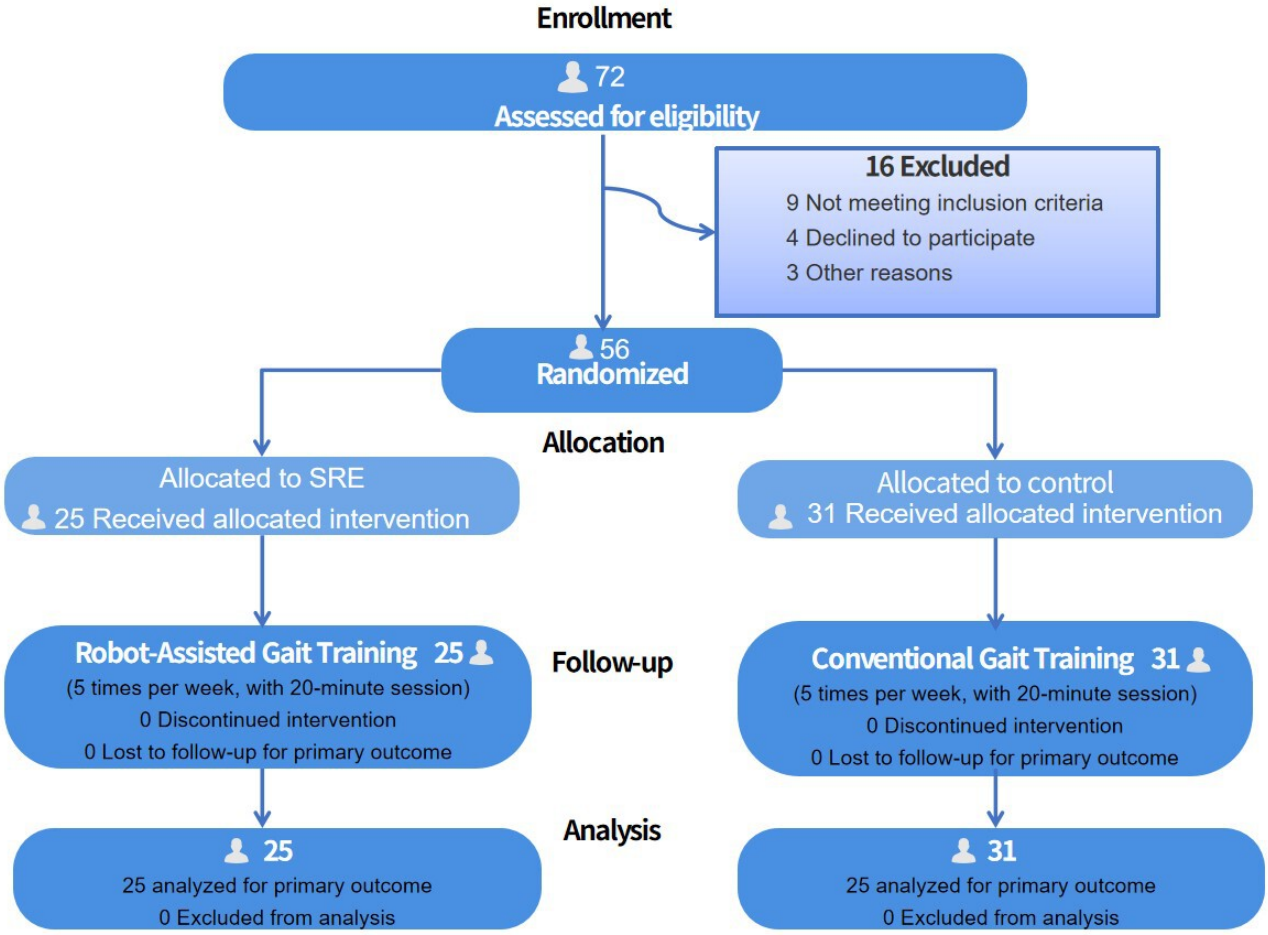

All continuous variables underwent normality tests. Data following a normal distribution were expressed as mean (SD), while nonnormally distributed data were presented as median (IQR).

Gait Spatiotemporal Parameters and Motor Function Outcome Measures

To explore the mechanisms by which intervention measures affect gait and motor function, this study collected and analyzed spatiotemporal and kinematic parameters at 2 time points (preintervention and postintervention) during walking (Figure 4). Paired *t* tests revealed statistically significant differences between the 2 groups in motor assessment indicators (total UPDRS-III score: SRE, 95% CI −7.33 to −3.45; *P*<.001; and CR, 95% CI −6.01 to −4.07; *P*<.001) and its balance subscore (SRE: 95% CI −1.05 to −0.49; *P*<.001; and CR: 95% CI −0.81 to −0.39; *P*<.001), gait parameters (walking speed: SRE, 95% CI 0.17 - 0.26; *P*<.001; and CR, 95% CI 0.11 - 0.20; *P*<.001), step length (*P*<.001), left swing phase percentage (SRE: *P*=.01; and CR: *P*=.04), and right ankle plantar flexion angle (SRE: *P*<.001; and CR: *P*=.02) before and after intervention. However, significant improvements postintervention were observed only in the SRE group, with right swing phase percentage (95% CI 0.62 - 2.62; *P*=.002), left ankle plantar flexion angle (*P*=.002), left ankle dorsiflexion angle (*P*<.001), and right ankle dorsiflexion angle (*P*=.01), indicating a specific advantage of the SRE intervention on these parameters, as detailed in Table 2.

XSL•FO
RenderX

Figure 4. The Balance Master system used for assessing balance and motor function in participants before and after the intervention. This system measures postural stability, dynamic balance, and functional performance and was used as part of the outcome measures for both the soft robotic exoskeleton and conventional rehabilitation groups throughout the study.

XSL•FO
RenderX

Table . Within- and between-group comparisons of gait spatiotemporal parameters and motor function outcome measures in the soft robotic exoskeleton device (SRE) and conventional rehabilitation (CR) groups.[a]

Measure	SRE group (n=25)	CR group (n=31)	Within-group *P* value	Between-group *P* value
UPDRS[b]-III (score)				
Pre	30.19 (11.97)	28.64 (11.20)	__[c]	—
Post	24.81 (12.95)[d]	23.60 (10.56)[d]	<.001	.91
Post-pre	2.80 (3.98)	3.84 (2.30)	—	—
UPDRS-III balance (score)				
Pre	2.65 (1.11)	2.32 (1.28)	—	—
Post	1.87 (1.02)[d]	1.72 (1.02)[d]	<.001	.49
Post-pre	0.40 (0.58)	0.44 (0.51)	—	—
Gait velocity (m/s) %> MCID[e] (0.05m/s)	96%	41%		
Pre	0.54 (0.24)	0.55 (0.27)	—	—
Post	0.75 (0.24)[d]	0.71 (0.27)[d]	<.001	.04
Post-pre	0.23 (0.10)[f]	0.15 (0.12)	—	—
Left step length (m)				
Pre	0.56 (0.24)	0.52 (0.28)	—	—
Post	0.71 (0.25)[d]	0.63 (0.30)[d]	<.001	.19
Post-pre	0.15 (0.16)	0.11 (0.11)	—	—
Right step length (m)				
Pre	0.56 (0.24)	0.54 (0.27)	—	—
Post	0.71 (0.24)[d]	0.66 (0.28)[d]	<.001	.47
Post-pre	0.15 (0.15)	0.12 (0.11)	—	—
Left swing phase (%)				
Pre	25.36 (4.75)	26.49 (4.01)	—	—
Post	26.92 (3.91)[g]	28.04 (4.48)[h]	.01; .04	.87
Post-pre	1.56 (3.05)	1.56 (3.47)	—	—
Right swing phase (%)				
Pre	26.07 (4.10)	26.43 (3.78)	—	—
Post	27.70 (3.52)[g]	27.83 (4.14)	.002; .06	.57
Post-pre	1.62 (2.72)	1.40 (3.50)	—	—
Left ankle plantarflexion (degree)				
Pre	46.47 (7.88)	45.36 (7.65)	.002; .11	.23
Post	49.27 (6.39)[g]	46.28 (8.66)	—	—
Post-pre	2.3 (4.24)	1.22 (1.77)	—	—
Left ankle dorsiflexion (degree)				
Pre	9.20 (2.54)	8.93 (3.36)	<.001; .06	.17
Post	11.67 (3.13)[d]	10.36 (3.84)	—	—
Post-pre	2.84 (1.46)	1.38 (2.84)	—	—

XSL•FO
RenderX

Measure	SRE group (n=25)	CR group (n=31)	Within-group *P* value	Between-group *P* value
Right ankle plantarflexion (degree)				
Pre	47.87 (6.94)	46.00 (8.29)	<.001; .02	.74
Post	49.20 (6.39)[d]	47.42 (8.06)[h]	—	—
Post-pre	1.79 (2.75)	1.42 (1.91)	—	—
Right ankle dorsiflexion (degree)				
Pre	10.00 (2.30)	8.79 (4.00)	.01; .07	.34
Post	11.67 (2.94)[g]	9.86 (3.53)	—	—
Post-pre	1.79 (2.22)	1.07 (1.90)	—	—

[a]Compared to the control group (post-pre: postintervention minus preintervention). To clarify the within-group *P* values, the first value represents the SRE group, and the second value corresponds to the CR group. When only one value is reported, it indicates that the *P* value is the same for both groups.

[b]UPDRS: Movement Disorder Society-Unified Parkinson disease rating scale.

[c]Not applicable.

[d]*P*<.001.

[e]MCID: minimal clinically important difference.

[f]*P*<.05.

[g]*P*<.01.

[h]Compared to pretreatment (pre), *P*<.05.

Independent sample *t* tests showed that the improvement in walking speed (Δ0.23 [0.10 m/s] vs Δ0.15 [0.12 m/s]) in the SRE group was significantly superior to the control group (*P*=.04), suggesting that SRE effectively restores motor function in people with PD, as presented in Table 2.

To further investigate the potential mechanisms underlying gait changes induced by SRE, linear regression models were used to assess the correlation between improvements in ankle joint function and changes in outcome measures. The results first showed a significant positive correlation between the degree of left ankle dorsiflexion improvement and the increase in walking speed (*r*=0.374, *P*=.04), as shown in Figure 5. Left ankle dorsiflexion (ρ=0.59; *P*=.001), as shown in Figure 6 and plantar flexion (ρ=0.46; *P*=.01), as shown in Figure 7, improvements were both significantly correlated with increased step length, and left ankle dorsiflexion improvement was significantly correlated with an increase in swing phase percentage (ρ=0.386; *P*=.04), as shown in Figure 8. No significant correlations were found between other variables. It is noteworthy that post-SRE intervention, improvement in right ankle function was not significantly correlated with any of the measured indicators. Furthermore, the results revealed significant negative correlations between improvements in left ankle plantar flexion (*r*=−0.386; *P*=.03), as shown in Figure 9, and right ankle dorsiflexion (*r*=−0.398; *P*=.03), as shown in Figure 10, and the decrease in UPDRS-III balance scores.

XSL•FO
RenderX

Figure 5. Correlations in soft robotic exoskeleton (SRE)–induced changes in gait spatiotemporal parameters and correlation of SRE-induced changes (post-pre) in left ankle dorsiflexion (LDF; x-axis) and SRE-induced changes in gait velocity (y-axis). *P<.05.

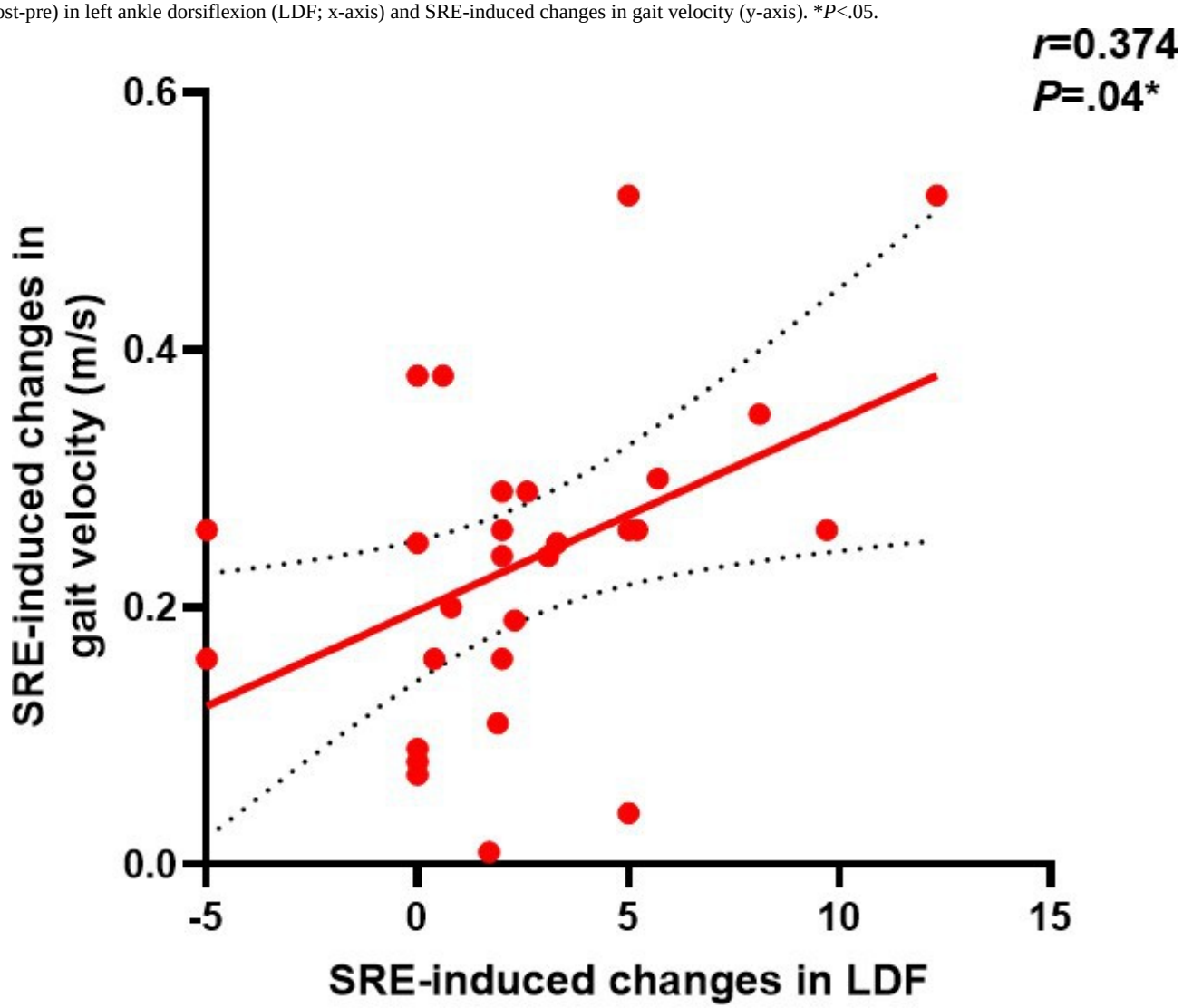

XSL•FO
RenderX

Figure 6. Correlations in soft robotic exoskeleton (SRE)–induced changes in gait spatiotemporal parameters and correlation of SRE-induced changes in left ankle dorsiflexion (LDF; x-axis) and SRE-induced changes in step length (y-axis). ***P<.001.

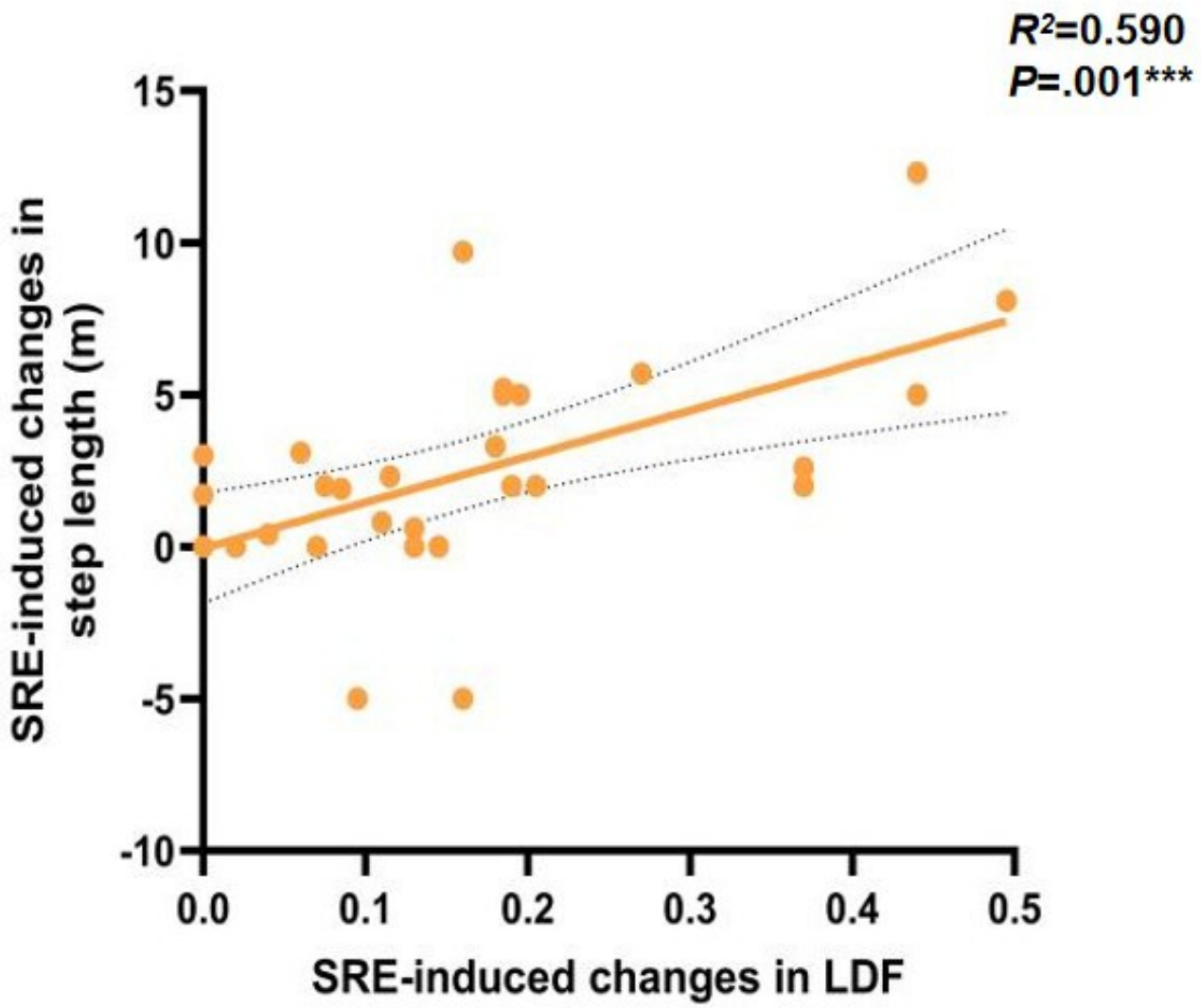

XSL•FO
RenderX

Figure 7. Correlations in soft robotic exoskeleton (SRE)–induced changes in gait spatiotemporal parameters and correlation of SRE-induced changes in left ankle dorsiflexion (LPF; x-axis) and SRE-induced changes in step length (y-axis). *P<.05.

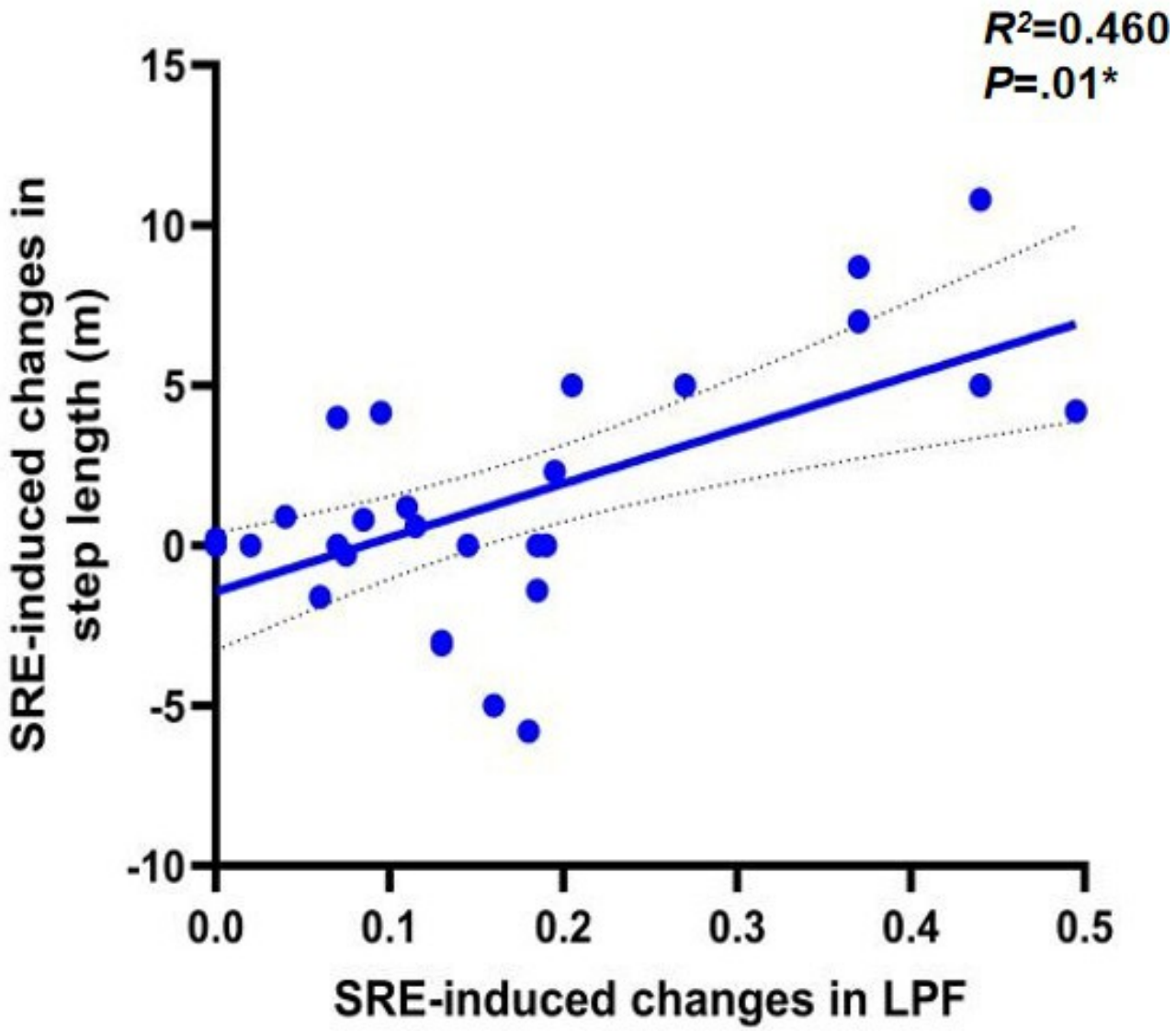

XSL•FO
RenderX

Figure 8. Correlations in soft robotic exoskeleton (SRE)–induced changes in gait spatiotemporal parameters and correlation of SRE-induced changes in LDF (X) and SRE-induced changes in swing phase (Y). *P<.05.

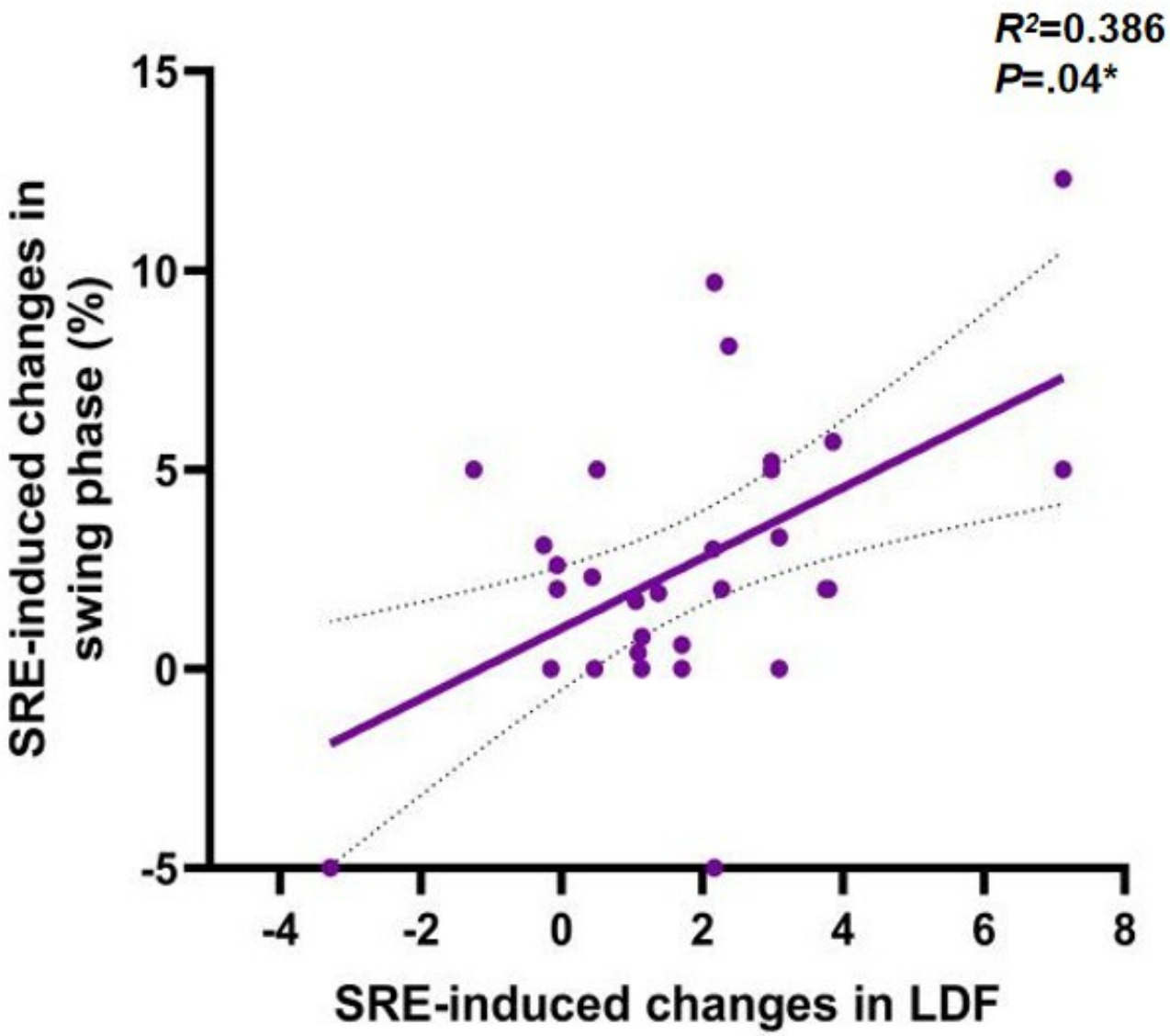

XSL•FO
RenderX

Figure 9. Correlations in soft robotic exoskeleton (SRE)–induced changes in Unified Parkinson Disease Rating Scale Part III (UPDRS-III) balance score. *P<.05.

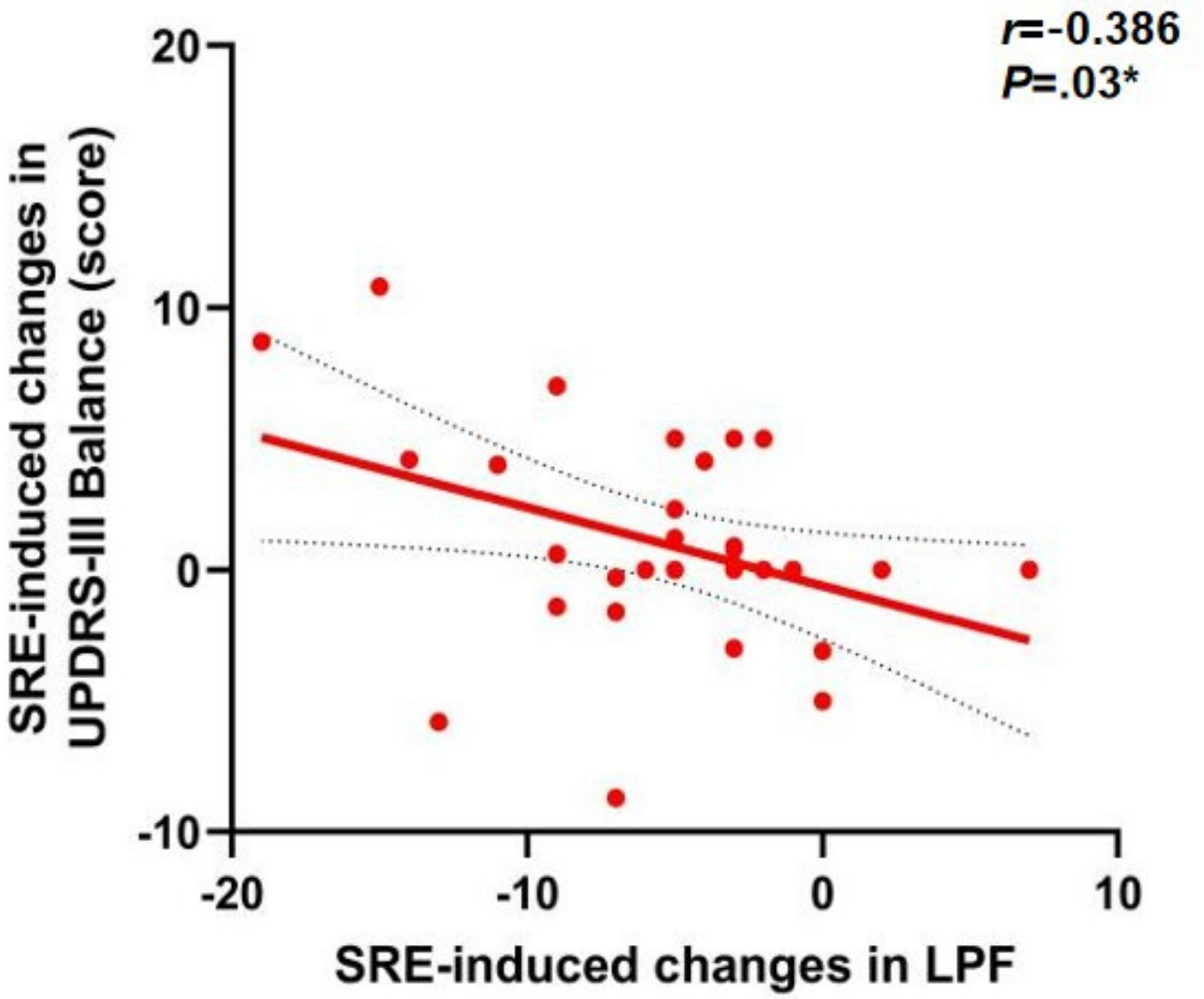

XSL•FO
RenderX

Figure 10. Correlations in soft robotic exoskeleton (SRE)–induced changes in Unified Parkinson Disease Rating Scale Part III (UPDRS-III) balance score. RDF: right ankle dorsiflexion. *P<.05.

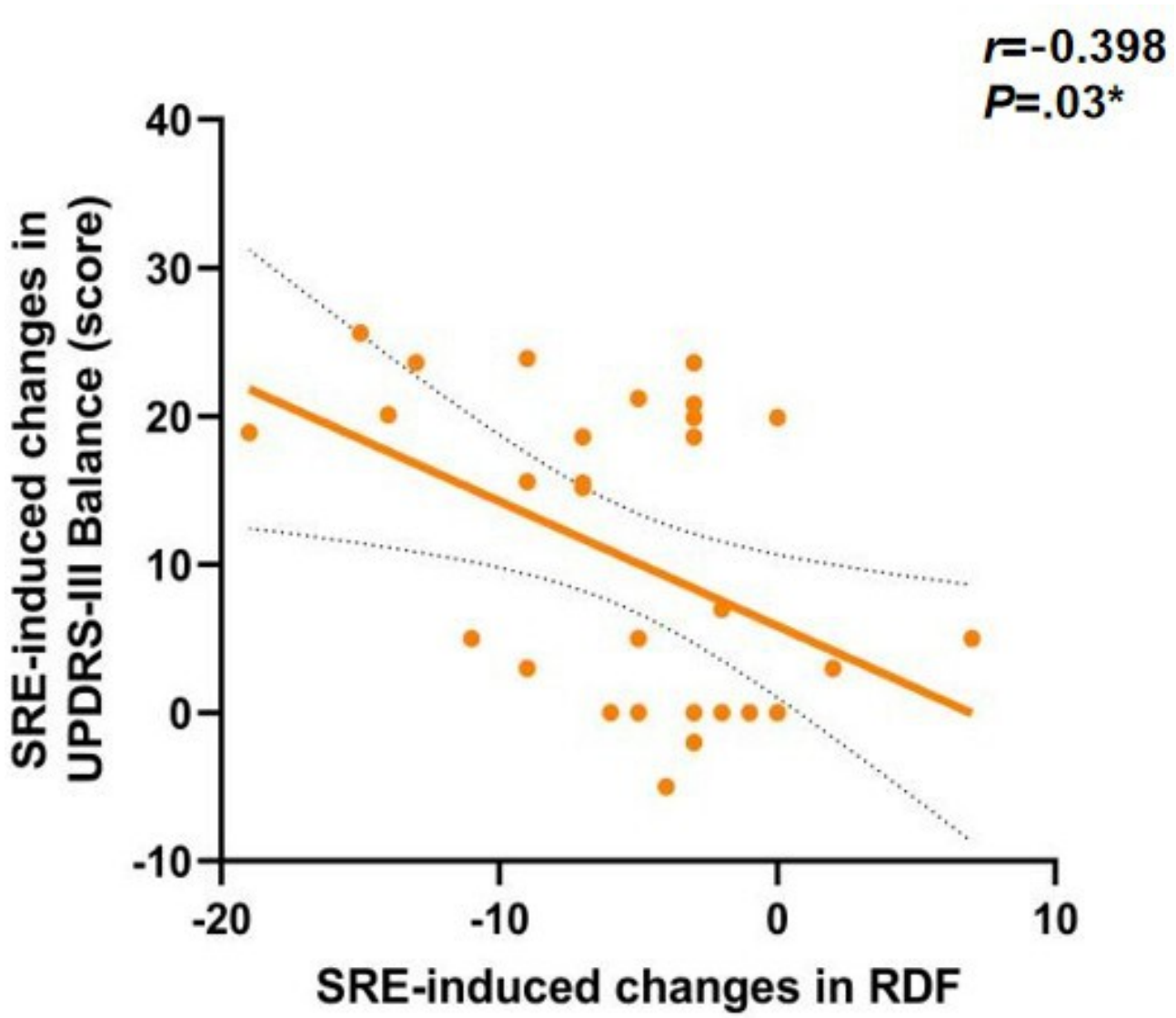

Balance Evaluation and Nonmotor Symptom Outcome Measures

To investigate the mechanisms by which intervention measures impact balance and nonmotor symptoms, this study collected and analyzed balance parameters and clinical scales at 2 time points (preintervention and postintervention). Paired t tests revealed statistically significant differences between the 2 groups in balance assessment indicators (average walking speed: SRE, P<.001; and CR, P=.02; DCL: SRE, P<.001; and CR, P=.01), sensory integration total score (P<.001), BBS scale (P<.001), quality of life (P<.001), and cognitive ability indicators (P<.001) before and after the intervention. However, significant improvements postintervention were observed only in the SRE group, with maximum movement distance (P<.001) and RT (P=.02), suggesting that SRE intervention has a specific advantage for these parameters, as detailed in Table 3.

XSL•FO
RenderX

Table . Within- and between-group comparisons of balance evaluation and nonmotor symptom outcome measures in the soft robotic exoskeleton (SRE) and conventional rehabilitation (CR) groups.[a]

Measure	SRE group (n=25)	CR group (n=31)
COP[b]-Max (%)		
Pre	51.55 (20.40)	51.96 (13.47)
Post	58.18 (18.99)[c]	55.26 (13.81)
Post-pre	7.70 (6.92)[d]	3.65 (5.76)
Reaction time (s)		
Pre	1.47 (0.50)	1.35 (0.46)
Post	1.22 (0.46)[e]	1.27 (0.46)
Post-pre	0.34 (0.27)[d]	0.06 (0.42)
Mean angular velocity (deg/sec)		
Pre	2.20 (1.14)	1.94 (1.02)
Post	2.55 (1.03)[f]	2.40 (1.36)[e]
Post-pre	2.54 (1.05)	2.30 (1.18)
Directional control (%)		
Pre	64.24 (15.73)	63.04 (13.47)
Post	72.19 (15.29)[c]	69.00 (9.90)[f]
Post-pre	7.95 (7.30)	6.76 (10.09)
SOT[g] Composite (score)		
Pre	64.80 (11.28)	62.44 (10.85)
Post	70.86 (10.15)[d]	66.92 (10.65)[d]
Post-pre	6.33 (2.76)	4.76 (4.01)
BBS[h] (score)		
Pre	40.45 (9.36)	41.12 (6.05)
Post	43.58 (9.27)[c]	45.44 (4.85)[c]
Post-pre	3.13 (3.00)	4.32 (2.90)
Barthel Index (score)		
Pre	81.13 (17.15)	76.68 (15.91)
Post	87.97 (11.67)[c]	83.96 (13.02)[c]
Post-pre	6.84 (7.14)	7.28 (5.11)
MoCA[i] (score)		
Pre	21.03 (5.14)	20.80 (5.20)
Post	22.26 (4.77)[c]	22.00 (4.57)[c]
Post-pre	1.23 (1.23)	1.20 (1.08)

[a]Compared to the control group (post-pre: postintervention minus preintervention).
[b]COP-Max: maximum center of pressure displacement.
[c]$P<.001$.
[d]$P<.05$.
[e]Compared to pretreatment (pre), $P<.05$.
[f]$P<.01$.
[g]SOT: Sensory Organization Test.
[h]BBS: Berg Balance Scale.

XSL•FO
RenderX

[i]MoCA: Montreal Cognitive Assessment.

Independent sample *t* tests indicated that the improvement in maximum movement distance (Δ7.70, 6.92% vs Δ3.65, 5.76%) was significantly greater in the SRE group compared to the control group (*P*=.048), and the improvement in RT (Δ0.34, 0.27 s vs Δ0.06, 0.42 s) was also significantly superior in the SRE group (*P*=.04). This improvement may be related to the unique neuromuscular activation patterns or central-peripheral coordination mechanisms specific to the SRE intervention, as presented in Table 3.

Discussion

Principal Findings

This RCT aimed to investigate the clinical efficacy of an ankle-targeted SRE system in improving gait abnormalities and balance impairments in people with PD. To the best of our knowledge, no previous study has specifically assessed the value of SRE-assisted gait training for functional rehabilitation in PD [20], although there have been studies on robot-assisted rehabilitation for gait and balance in people with PD [20]. After a 4-week, 20-session standardized intervention, the SRE group showed significant intragroup improvements in both motor and nonmotor function indicators, confirming the potential of robot-assisted rehabilitation to address multidimensional functional impairments in PD. The main benefits of SRE-assisted walking training appear to lie in optimizing gait dynamics and enhancing balance function.

This study achieved important findings in gait parameter optimization, consistent with previous research highlighting the role of robotic interventions in gait rehabilitation for people with PD [24]. After 4 weeks of clinical training, both groups showed improvements in gait function, but the SRE group demonstrated a significantly greater increase in walking speed, with a higher MCID achievement rate compared to the control group. This result aligns with the conclusions of a meta-analysis of 7 RCTs on robot-assisted gait training, which found that robot-assisted gait training in any form significantly improves walking speed [29]. Notably, postintervention, the SRE group exhibited a greater increase in ankle dorsiflexion-plantarflexion range of motion and effectively corrected the compensatory prolonged support phase in people with PD. The proportion of the right lower limb swing phase increased in the SRE group, while no significant change was observed in the control group. These biomechanical improvements were closely related to enhanced ankle function. Correlation analysis revealed a positive association between increased ankle dorsiflexion and improvements in step length, walking speed, and swing phase duration.

The mechanisms underlying these results can be attributed to the dual regulatory effect of SRE intervention. At the peripheral biomechanical level, this is attributed to SRE's ankle joint-driven assistive system. Research indicates that ankle dorsiflexor strength is a key determinant of walking speed [30], and in our findings, an increase in ankle dorsiflexion range was significantly positively correlated with walking speed. SRE effectively modulates the spatiotemporal coupling between the center of pressure trajectory and stride rhythm, reducing abnormal cocontraction of the ankle dorsiflexors and plantarflexors [31], thus surpassing traditional rehabilitation training in improving ankle joint mobility and supporting high-intensity repetitive gait training. Additionally, based on the central pattern generator hypothesis of walking control, SRE's fixed gait rhythm input integrates proprioceptive feedback from the ankle Golgi tendon organ and hip muscle spindles [32,33], activating spinal central pattern generators to produce coordinated muscle activation patterns. This not only explains the 93.0% MCID achievement rate in walking speed but also strengthens the efficiency of the stance-to-swing phase transition, enhancing motor control automation.

The neuroplasticity mechanism supports the effectiveness of SRE, as high-intensity, repetitive tasks have been shown to induce brain reorganization in patients with PD [34]. The study further validated that early, task-oriented, high-intensity, and high-repetition training can significantly promote neuroplasticity changes [35], providing peripheral and central mechanistic support for the advantages of SRE intervention in early stages of PD.

In terms of postural control function, the SRE group demonstrated greater improvements compared to the control group. In the stability limit test, rotational response time decreased more substantially in the SRE group than in the CR group. The potential mechanism behind this improvement may be related to the activation of ankle strategies by the SRE intervention [36], achieved through high-precision proprioceptive feedback [37]. This enhances the rapid response of the ankle joint's surrounding muscles (eg, the tibialis anterior), reversing delayed ankle dorsiflexion in people with PD and reducing postural preparation time, thereby optimizing the gait-posture coordination pattern [36]. Additionally, due to muscle rigidity, reduced joint mobility, and decreased axial flexibility, people with PD often exhibit poor rotational kinematic parameters [13]. Following SRE intervention, overall rotational speed and direction control showed greater improvement in the SRE group. Notably, the SRE group demonstrated a more pronounced increase in maximum rotational distance compared to the CR group, indicating enhanced dynamic stability. This metric reflects the maximal ability of people with PD to shift their center of mass without falling, and its significant improvement suggests enhanced dynamic stability. Under normal physiological conditions, individuals adopt an ankle strategy in response to small perturbations and a hip strategy for larger disturbances [38]. In the pathological state of PD, however, due to basal ganglia-thalamus-cortex loop dysfunction [38], the ability to apply the ankle strategy is significantly diminished, forcing premature use of the less efficient hip strategy [39], leading to trunk rigidity and limited center-of-mass adjustment [38,39]. The significant improvement in the "maximum rotational distance" in the SRE group may indicate that SRE enhanced the ankle joint's resistance to perturbation, extended the boundaries of dynamic stability, improved the contraction efficiency of ankle muscles, delayed the threshold for initiating

XSL•FO
RenderX

the hip strategy, and possibly optimized sensorimotor integration by improving proprioceptive input accuracy [5]. These results suggest that SRE may enhance postural stability through ankle joint support and improve PD-specific balance control dysfunction, potentially enhancing overall balance performance in people with PD. Further research is needed to confirm these effects in larger, more diverse populations. Additionally, the postural control dysfunction in people with PD may be linked to a fear of falling [40]. SRE might reduce the fear of falling in people with PD. Some scholars have suggested that robot-assisted walking may generate external rhythmic cues through proprioceptive feedback in a fixed mode, compensating for the internal rhythm deficits caused by basal ganglia circuit abnormalities in people with PD, thereby aiding postural control [41]. This "bottom-up" intervention strategy provides a novel target for breaking the PD posture-gait-activity vicious cycle.

It is noteworthy that the SRE group also showed improvements in cognitive symptoms and quality of life. The active participation in training has always been a critical factor in neurorehabilitation. Compared to platform-based robots, the gait training used in SRE allows patients to walk actively in a natural environment, capturing gait initiation intentions (such as heel strike angle and pressure center displacement) in real time through an inertial measurement unit, and providing assistive torque based on gait phases. This "intent-action closed-loop feedback" not only enhances neuromuscular engagement but also optimizes dynamic balance ability, thereby more effectively activating the motor cortex [42,43]. A study using electroencephalogram found that the frontal-central-parietal network is a potential neural marker for robot-assisted walking training to improve motor relearning and adaptation, promoting gait initiation and maintenance through activation of cortical and parietal motor pathways [44].

These findings have clear clinical implications. When providing rehabilitation interventions for people with PD, particularly those with balance disorders and fall risks, clinicians should pay close attention to the evaluation and management of ankle joint mobility limitations. This study suggests that the degree of improvement in ankle joint mobility can serve as a valuable predictive indicator for assessing and predicting the potential therapeutic effects on overall dynamic balance function (especially direction control) following SRE-assisted training or other ankle-focused interventions. Furthermore, to maximize rehabilitation benefits, future rehabilitation plans could incorporate more comprehensive approaches. These might combine technologies such as SRE, which provide precise joint assistance, with training methods focused on sensory integration, core stability, or gait coordination, in order to create synergistic effects that more comprehensively promote balance recovery and restoration in people with PD.

Limitations

While this study provides promising evidence for the efficacy of SRE-assisted gait training in people with PD, several limitations must be acknowledged. First, the relatively small sample size may limit the statistical power and generalizability of the results. However, this prospective trial serves as a foundational step, offering an effect size estimate that could inform future large-sample RCTs designed to further validate and refine the intervention. Second, the absence of long-term follow-up data prevents an assessment of the durability of the SRE intervention effects, as well as the potential cumulative benefits of neuroplasticity. Long-term studies incorporating advanced imaging techniques, such as resting-state functional MRI, alongside serum neurotrophic factor analyses, could provide valuable insights into the dynamics of neural remodeling and the sustained impact of robotic rehabilitation. Finally, future research should delve deeper into the biomechanics of the SRE intervention. Integrating 3D gait analysis systems with surface electromyography would not only enhance our understanding of the mechanistic pathways involved but also refine therapeutic strategies to maximize the benefits of exoskeleton-assisted rehabilitation for people with PD. In a broader sense, these studies could contribute to the development of more personalized, evidence-based rehabilitation protocols, ultimately improving the quality of life for individuals living with PD.

Conclusions

In conclusion, our findings confirm that integrating a soft exoskeleton robot (SER) into CR yields significant and clinically meaningful improvements in gait, balance, and broader functional outcomes for people with PD. The innovative use of a lightweight, adaptive SER—capable of real-time assistance during walking—fills a critical gap in evidence regarding modern wearable technologies for PD. Unlike earlier research that predominantly examined rigid exoskeletons or single-dimension effects, our study used a multifaceted evaluation (kinematic, kinetic, and clinical measures) and identified a mechanistic link: enhanced ankle dorsiflexion contributes to gains in stride length and gait speed. This advances the field by providing a replicable, evidence-based framework for incorporating SER into neurorehabilitation protocols. From a real-world perspective, the observed improvements in gait symmetry, balance control, and daily living activities suggest that SER can be feasibly implemented in clinical settings to help patients maintain mobility and autonomy, thereby addressing a core challenge in PD management. These results underscore the potential of soft robotics to transform rehabilitation practice and warrant further multicenter trials with longer follow-up.

Funding

This work was supported by the National Key R&D Program of China (grant 2023YFC3603800, 2023YFC3603803), the National Natural Science Foundation of China (82274300, 82172545), Shanghai Hospital Development Center Foundation Shanghai Municipal Hospital Rehabilitation Medicine Specialty Alliance (SHDC22023304), and Shanghai Innovative Medical Devices

XSL•FO
RenderX

Application Demonstration Project (23SHS05300).The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated and/or analyzed during this study are not publicly available due to [1] ethical restrictions protecting participant confidentiality and [2] the terms of the informed consent agreements, which specify that the data will only be used by the research team for the approved study. However, the deidentified data are available from the corresponding author (QZ) on reasonable request. Requests must include a scientifically sound research proposal and will require approval from the Shanghai Yangzhi Rehabilitation Hospital Ethics Committee prior to any data being shared.

Editorial Notice

This randomized study was only retrospectively registered because technical issues with the registration platform delayed prospective registration. The editor granted an exception from ICMJE (International Committee of Medical Journal Editors) rules mandating prospective registration of randomized trials, because the risk of bias appears low and the study was considered a formative pilot study. However, readers are advised to carefully assess the validity of any potential explicit or implicit claims related to primary outcomes or effectiveness, as retrospective registration does not prevent authors from changing their outcome measures retrospectively.

Checklist 1
CONSORT checklist.
[PDF File, 1317 KB - jmir_v28i1e82629_app1.pdf]

References

1. do Amaral CMS, de Almeida SB, de Almeida RP, do Nascimento SL, Ribeiro RM, Braga-Neto P. Effectiveness of vestibular rehabilitation on postural balance in Parkinson's disease: a systematic review and meta-analysis of randomized controlled trials. BMC Neurol 2024 May 14;24(1):161. [doi: 10.1186/s12883-024-03649-5] [Medline: 38745275]
2. Cosentino C, Baccini M, Putzolu M, Ristori D, Avanzino L, Pelosin E. Effectiveness of physiotherapy on freezing of gait in Parkinson's disease: a systematic review and meta-analyses. Mov Disord 2020 Apr;35(4):523-536. [doi: 10.1002/mds.27936] [Medline: 31799734]
3. Zhang WS, Gao C, Tan YY, Chen SD. Correction to: prevalence of freezing of gait in Parkinson's disease: a systematic review and meta-analysis. J Neurol 2021 Nov;268(11):4151. [doi: 10.1007/s00415-021-10756-7] [Medline: 34427756]
4. Kalia LV, Lang AE. Parkinson's disease. Lancet 2015 Aug 29;386(9996):896-912. [doi: 10.1016/S0140-6736(14)61393-3] [Medline: 25904081]
5. Gao C, Liu J, Tan Y, Chen S. Freezing of gait in Parkinson's disease: pathophysiology, risk factors and treatments. Transl Neurodegener 2020;9(12):12. [doi: 10.1186/s40035-020-00191-5] [Medline: 32322387]
6. Kwon SH, Lee BS, Lee HJ, et al. Energy efficiency and patient satisfaction of gait with knee-ankle-foot orthosis and robot (ReWalk)-assisted gait in patients with spinal cord injury. Ann Rehabil Med 2020 Apr;44(2):131-141. [doi: 10.5535/arm.2020.44.2.131] [Medline: 32392652]
7. Camargo CHF, Ferreira-Peruzzo SA, Ribas DIR, Franklin GL, Teive HAG. Imbalance and gait impairment in Parkinson's disease: discussing postural instability and ataxia. Neurol Sci 2024 Apr;45(4):1377-1388. [doi: 10.1007/s10072-023-07205-w] [Medline: 37985635]
8. Ünlüer N, Ateş Y, Balbaba A, Sücüllü Karadağ Y. Investigation of non-motor symptoms, balance, muscle strength, and functional mobility in patients with Parkinson's disease. Somatosens Mot Res 2021 Mar;38(1):20-26. [doi: 10.1080/08990220.2020.1834377] [Medline: 33131377]
9. Park JH, Kang YJ, Horak FB. What is wrong with balance in Parkinson's disease? J Mov Disord 2015 Sep;8(3):109-114. [doi: 10.14802/jmd.15018] [Medline: 26413237]
10. Carpenter MG, Allum JHJ, Honegger F, Adkin AL, Bloem BR. Postural abnormalities to multidirectional stance perturbations in Parkinson's disease. J Neurol Neurosurg Psychiatry 2004 Sep;75(9):1245-1254. [doi: 10.1136/jnnp.2003.021147] [Medline: 15314109]
11. Corti O, Lesage S, Brice A. What genetics tells us about the causes and mechanisms of Parkinson's disease. Physiol Rev 2011 Oct;91(4):1161-1218. [doi: 10.1152/physrev.00022.2010] [Medline: 22013209]
12. Ferreira-Pinto MJ, Kanodia H, Falasconi A, Sigrist M, Esposito MS, Arber S. Functional diversity for body actions in the mesencephalic locomotor region. Cell 2021 Aug 19;184(17):4564-4578. [doi: 10.1016/j.cell.2021.07.002] [Medline: 34302739]
13. Grillner S, Wallén P, Saitoh K, Kozlov A, Robertson B. Neural bases of goal-directed locomotion in vertebrates--an overview. Brain Res Rev 2008 Jan;57(1):2-12. [doi: 10.1016/j.brainresrev.2007.06.027] [Medline: 17916382]
14. Doná F, Aquino CC, Gazzola JM, et al. Changes in postural control in patients with Parkinson's disease: a posturographic study. Physiotherapy 2016 Sep;102(3):272-279. [doi: 10.1016/j.physio.2015.08.009] [Medline: 26582134]

XSL•FO
RenderX

15. Rossi M, Soto A, Santos S, Sesar A, Labella T. A prospective study of alterations in balance among patients with Parkinson's disease. Protocol of the postural evaluation. Eur Neurol 2009;61(3):171-176. [doi: 10.1159/000189270] [Medline: 19129704]
16. Vitale C, Marcelli V, Furia T, et al. Vestibular impairment and adaptive postural imbalance in Parkinsonian patients with lateral trunk flexion. Mov Disord 2011 Jul;26(8):1458-1463. [doi: 10.1002/mds.23657] [Medline: 21465552]
17. Haufe FL, Schmidt K, Duarte JE, Wolf P, Riener R, Xiloyannis M. Activity-based training with the Myosuit: a safety and feasibility study across diverse gait disorders. J Neuroeng Rehabil 2020 Oct 8;17(1):135. [doi: 10.1186/s12984-020-00765-4] [Medline: 33032627]
18. Otlet V, Vandamme C, Warlop T, Crevecoeur F, Ronsse R. Effects of overground gait training assisted by a wearable exoskeleton in patients with Parkinson's disease. J Neuroeng Rehabil 2023 Nov 16;20(1):156. [doi: 10.1186/s12984-023-01280-y] [Medline: 37974229]
19. Xie R, Zhang Y, Jin H, Yang F, Feng Y, Pan Y. Effects of soft robotic exoskeleton for gait training on clinical and biomechanical gait outcomes in patients with sub-acute stroke: a randomized controlled pilot study. Front Neurol 2023;14:1296102. [doi: 10.3389/fneur.2023.1296102] [Medline: 38020601]
20. Xue X, Yang X, Deng Z. Efficacy of rehabilitation robot-assisted gait training on lower extremity dyskinesia in patients with Parkinson's disease: a systematic review and meta-analysis. Ageing Res Rev 2023 Mar;85:101837. [doi: 10.1016/j.arr.2022.101837] [Medline: 36634871]
21. Chinese Clinical Trial Registry. URL: http://www.chictr.org.cn [accessed 2026-04-17]
22. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ 2025 Apr 14;389:e081123. [doi: 10.1136/bmj-2024-081123] [Medline: 40228833]
23. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011;13(4):e126. [doi: 10.2196/jmir.1923]
24. Picelli A, Melotti C, Origano F, et al. Robot-assisted gait training in patients with Parkinson disease: a randomized controlled trial. Neurorehabil Neural Repair 2012 May;26(4):353-361. [doi: 10.1177/1545968311424417] [Medline: 22258155]
25. Julious SA. Sample sizes for clinical trials with normal data. Stat Med 2004 Jun 30;23(12):1921-1986. [doi: 10.1002/sim.1783] [Medline: 15195324]
26. Bacchetti P, Leung JM. Sample size calculations in clinical research. Anesthesiology 2002 Oct;97(4):1028-1029. [doi: 10.1097/00000542-200210000-00050] [Medline: 12357184]
27. Yandell MB, Quinlivan BT, Popov D, Walsh C, Zelik KE. Physical interface dynamics alter how robotic exosuits augment human movement: implications for optimizing wearable assistive devices. J Neuroeng Rehabil 2017 May 18;14(1):40. [doi: 10.1186/s12984-017-0247-9] [Medline: 28521803]
28. Hass CJ, Bishop M, Moscovich M, et al. Defining the clinically meaningful difference in gait speed in persons with Parkinson disease. J Neurol Phys Ther 2014 Oct;38(4):233-238. [doi: 10.1097/NPT.0000000000000055] [Medline: 25198866]
29. Alwardat M, Etoom M, Al Dajah S, et al. Effectiveness of robot-assisted gait training on motor impairments in people with Parkinson's disease: a systematic review and meta-analysis. Int J Rehabil Res 2018 Dec;41(4):287-296. [doi: 10.1097/MRR.0000000000000312] [Medline: 30119060]
30. Lin PY, Yang YR, Cheng SJ, Wang RY. The relation between ankle impairments and gait velocity and symmetry in people with stroke. Arch Phys Med Rehabil 2006 Apr;87(4):562-568. [doi: 10.1016/j.apmr.2005.12.042] [Medline: 16571398]
31. Furnari A, Calabrò RS, De Cola MC, et al. Robotic-assisted gait training in Parkinson's disease: a three-month follow-up randomized clinical trial. Int J Neurosci 2017 Nov;127(11):996-1004. [doi: 10.1080/00207454.2017.1288623] [Medline: 28132574]
32. Picelli A, Capecci M, Filippetti M, et al. Effects of robot-assisted gait training on postural instability in Parkinson's disease: a systematic review. Eur J Phys Rehabil Med 2021 Jun;57(3):472-477. [doi: 10.23736/S1973-9087.21.06939-2] [Medline: 33826278]
33. Takakusaki K, Tomita N, Yano M. Substrates for normal gait and pathophysiology of gait disturbances with respect to the basal ganglia dysfunction. J Neurol 2008 Aug;255 Suppl 4:19-29. [doi: 10.1007/s00415-008-4004-7] [Medline: 18821082]
34. Choi JY, Kim SK, Hong J, et al. Overground gait training with a wearable robot in children with cerebral palsy: a randomized clinical trial. JAMA Netw Open 2024 Jul 1;7(7):e2422625. [doi: 10.1001/jamanetworkopen.2024.22625] [Medline: 39037815]
35. Kim H, Kim E, Yun SJ, et al. Robot-assisted gait training with auditory and visual cues in Parkinson's disease: a randomized controlled trial. Ann Phys Rehabil Med 2022 May;65(3):101620. [doi: 10.1016/j.rehab.2021.101620] [Medline: 34896605]
36. Nagai A, Marumoto K, Ohata K, Takasaki S, Moriyama H. Effects of ankle dorsiflexion training on anticipatory postural adjustments during gait initiation in patients with Parkinson's disease. Gait Posture 2025 Mar;117:109-114. [doi: 10.1016/j.gaitpost.2024.12.015] [Medline: 39693727]
37. Ribot-Ciscar E, Aimonetti JM, Azulay JP. Sensory training with vibration-induced kinesthetic illusions improves proprioceptive integration in patients with Parkinson's disease. J Neurol Sci 2017 Dec;383:161-165. [doi: 10.1016/j.jns.2017.11.016]
38. Horak FB, Frank J, Nutt J. Effects of dopamine on postural control in parkinsonian subjects: scaling, set, and tone. J Neurophysiol 1996 Jun 1;75(6):2380-2396. [doi: 10.1152/jn.1996.75.6.2380]

XSL•FO
RenderX

39. Ashburn A, Pickering R, McIntosh E, et al. Exercise- and strategy-based physiotherapy-delivered intervention for preventing repeat falls in people with Parkinson's: the PDSAFE RCT. Health Technol Assess 2019;23(36):1-150. [doi: 10.3310/hta23360]
40. Adkin AL, Frank JS, Jog MS. Fear of falling and postural control in Parkinson's disease. Mov Disord 2003 May;18(5):496-502. [doi: 10.1002/mds.10396] [Medline: 12722162]
41. Nieuwboer A, Kwakkel G, Rochester L, et al. Cueing training in the home improves gait-related mobility in Parkinson's disease: the RESCUE trial. J Neurol Neurosurg Psychiatry 2007 Feb;78(2):134-140. [doi: 10.1136/jnnp.200X.097923] [Medline: 17229744]
42. Hong J, Lee J, Choi T, et al. Feasibility of overground gait training using a joint-torque-assisting wearable exoskeletal robot in children with static brain injury. Sensors (Basel) 2022 May 19;22(10):3870. [doi: 10.3390/s22103870] [Medline: 35632279]
43. Brozova H, Barnaure I, Ruzicka E, Stochl J, Alterman R, Tagliati M. Short- and long-term effects of DBS on gait in Parkinson's disease. Front Neurol 2021;12:688760. [doi: 10.3389/fneur.2021.688760] [Medline: 34690908]
44. Youssofzadeh V, Zanotto D, Wong-Lin K, Agrawal SK, Prasad G. Directed functional connectivity in fronto-centroparietal circuit correlates with motor adaptation in gait training. IEEE Trans Neural Syst Rehabil Eng 2016 Nov;24(11):1265-1275. [doi: 10.1109/TNSRE.2016.2551642] [Medline: 27071181]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
CR: conventional rehabilitation
DCL: directional control
IRB: Institutional Review Board
MCID: minimal clinically important difference
PD: Parkinson disease
RCT: randomized controlled trial
RT: reaction time
SER: soft exoskeleton robot
SRE: soft robotic exoskeleton
UPDRS-III: Unified Parkinson Disease Rating Scale Part III

Edited by S Brini; submitted 19.Aug.2025; peer-reviewed by J Hwang, N Postol; revised version received 07.Mar.2026; accepted 08.Mar.2026; published 28.Apr.2026.

Please cite as:
Bu S, Luo X, Li S, Shi Y, Jin L, Zhao Q
Effects of a Soft Robotic Exoskeleton for Gait Training on Clinical Outcomes in Patients With Parkinson Disease: Randomized Controlled Pilot Study
J Med Internet Res 2026;28:e82629
URL: https://www.jmir.org/2026/1/e82629
doi:10.2196/82629

© Shiying Bu, Xiaoqing Luo, Shuangfang Li, Yuqin Shi, Lingjing Jin, Qing Zhao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Facilitators and Barriers to Implementing a Remote Monitoring Model of Care for Stable Patients With Axial Spondyloarthritis Using the Consolidated Framework for Implementation Research: Qualitative Study

Yu Heng Kwan[1,2,3*], BSc, MD, PhD; Xin Ru Chew[2*], BSc; Humairah Zainal[4], PhD; Zhonghui Xiong[2], MBBS; Melissa Hock[5]; Ka-Hao Zee[2]; Ying-Ying Leung[2], MBBS, MD; Truls Østbye[1,6], MPH, MBA, MD; Lian Leng Low[1,3,7], MBBS, MMed, MCI; Hendra Goh[1], PhD; Charmaine Wai Yan Sum[2], BSc; Sungwon Yoon[1,3], BA, MPH, PhD; Charmaine Tze May Wang[2], MRCP; Warren Fong[2,8], MRCP; Julian Thumboo[1,2,3,8], FRCP

[1]Program in Health Services and Systems Research, Duke-NUS Medical School, 8 College Road, Singapore, Singapore
[2]Department of Rheumatology and Immunology, Singapore General Hospital, Singapore, Singapore
[3]SingHealth Regional Health System, Centre of Population Health and Implementation Research, Singapore, Singapore
[4]Health Services Research Unit, Singapore General Hospital, Singapore, Singapore
[5]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore
[6]Department of Family Medicine and Community Health, Duke University, Durham, NC, United States
[7]Post Acute and Continuing Care, Outram Community Hospital, Singapore, Singapore
[8]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[*]these authors contributed equally

Corresponding Author:
Yu Heng Kwan, BSc, MD, PhD
Program in Health Services and Systems Research, Duke-NUS Medical School, 8 College Road, Singapore, Singapore

Abstract

Background: Close follow-up of stable patients with axial spondyloarthritis (axSpA) presents a financial burden and inconvenience to patients. A remote monitoring patient-reported outcome measures (PROMs)–based model of care (PROMise) was designed to reduce the frequency of in-person consultations for stable patients with axSpA. However, little is known about the facilitators and barriers of implementing a remote monitoring PROMise.

Objective: This study aims to understand the facilitators and barriers, as well as the mitigation strategies to implementing a PROMise in the Singapore context.

Methods: We conducted a qualitative study involving in-depth interviews with 19 patients with axSpA (78.9% (15) male, mean age 39.4, SD 11.7 years) and 13 health care professionals (HCPs) (23.1%, 3 male; mean age 37.9, SD 7.2 years) in a tertiary hospital in Singapore until data saturation was reached. Participants were purposively recruited based on sex, age, and ethnicity. Patients were additionally recruited based on the number of years since diagnosed with axSpA, while HCPs were recruited based on seniority and their role in the care of patients with axSpA. Interviews were transcribed, deductively analyzed, and mapped to the Consolidated Framework for Implementation Research (CFIR) framework to identify facilitators and barriers from both the patients' and HCPs' perspectives. The CFIR-Expert Recommendations for Implementing Change (ERIC) match tool was used to produce implementation strategies to overcome the CFIR barriers identified.

Results: All five domains of the CFIR framework were elicited. Facilitators included (1) reduced inconvenience and costs for patients and reduced patient load in the clinic, (2) need for PROMise, (3) similarity to current workflows, and (4) suitable patient selection. Barriers included concerns for (1) financial sustainability of PROMise, (2) cultural conditions, (3) patient safety, and (4) increased workload for HCPs. In total, 35 ERIC strategies were matched to the corresponding CFIR barriers.

Conclusions: We identified ERIC strategies that will facilitate the implementation of the PROMise model. In particular, focus should be placed on developing an implementation blueprint and obtaining continuous feedback from affected patients with axSpA and HCPs involved in the care of the affected patients. These implementation strategies cross-cut the CFIR barriers identified and thus may overcome the barriers to implementation.

(J Med Internet Res 2026;28:e82480) doi:10.2196/82480

XSL•FO
RenderX

KEYWORDS

axial spondyloarthritis; model of care; implementation science; consolidated framework for implementation research; patient reported outcome measures; Singapore

Introduction

Axial spondyloarthritis (axSpA) is a chronic inflammatory disease characterized by chronic back pain and stiffness of the pelvis and lower back. Peak disease onset occurs in late adolescence and early adulthood [1,2]. The reported prevalence of axSpA ranges from 0.3% to 1.4% [3,4]. Axial spondyloarthritis can have a long-lasting impact on various aspects of patients' lives including in the functional, psychological, and social domains [5]. Hence, it is important to measure both the level of disease activity and the patient's health-related quality of life (HRQoL), which can be achieved through patient-reported outcome measures (PROMs).

PROMs are standardized and validated questionnaires that provide patients' perspectives of their health status and HRQoL over time. PROMs including the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) have indicated that a higher level of disease activity correlates with poorer HRQoL [6]. Patients with axSpA have a poor HRQoL that is comparable to that of patients with moderate to end-stage kidney disease [7]. It is important to diagnose axSpA early and closely follow up with patients with axSpA to prevent disease progression and improve HRQoL [8]. Furthermore, the use of disease - modifying antirheumatic drugs such as sulfasalazine or biologics requires patients to be regularly monitored for toxicity and side effects. Hence, it is recommended that patients with axSpA visit a rheumatologist at least twice a year [9].

However, frequent visits to the rheumatologists result in both a high economic cost to patients and inconvenience due to time spent traveling and waiting for consultations [10]. Patients with axSpA spend on average US $180 per visit to the rheumatologist [11]. Hence, there is a need for an alternative monitoring strategy. A randomized controlled trial by Hermans et al [12] in the Netherlands remotely monitored patients with axSpA through laboratory tests and PROM results. In that study, patients were able to self-initiate appointments after a review of their laboratory test and PROM result by a rheumatologist. The model was shown to reduce health care costs and be noninferior to usual care in terms of quality-adjusted life years in the remote monitoring group.

We propose a similar PROMs-based model of care (PROMise), which aims to remotely monitor stable patients with axSpA through an automated review of their laboratory test and PROM results. Patients are not required to be physically or virtually present in the hospital for the automated review (see Figure 1). Through an automated review of the laboratory test and PROM result, we aim to maintain the standards of patient outcomes and patient experience while reducing health care costs and improving health care use [12].

Figure 1. Comparison between the model of care and patient-reported outcome measures–based model of care (PROMise). AMS: automated monitoring system; PROMise: patient-reported outcome measures–based model of care.

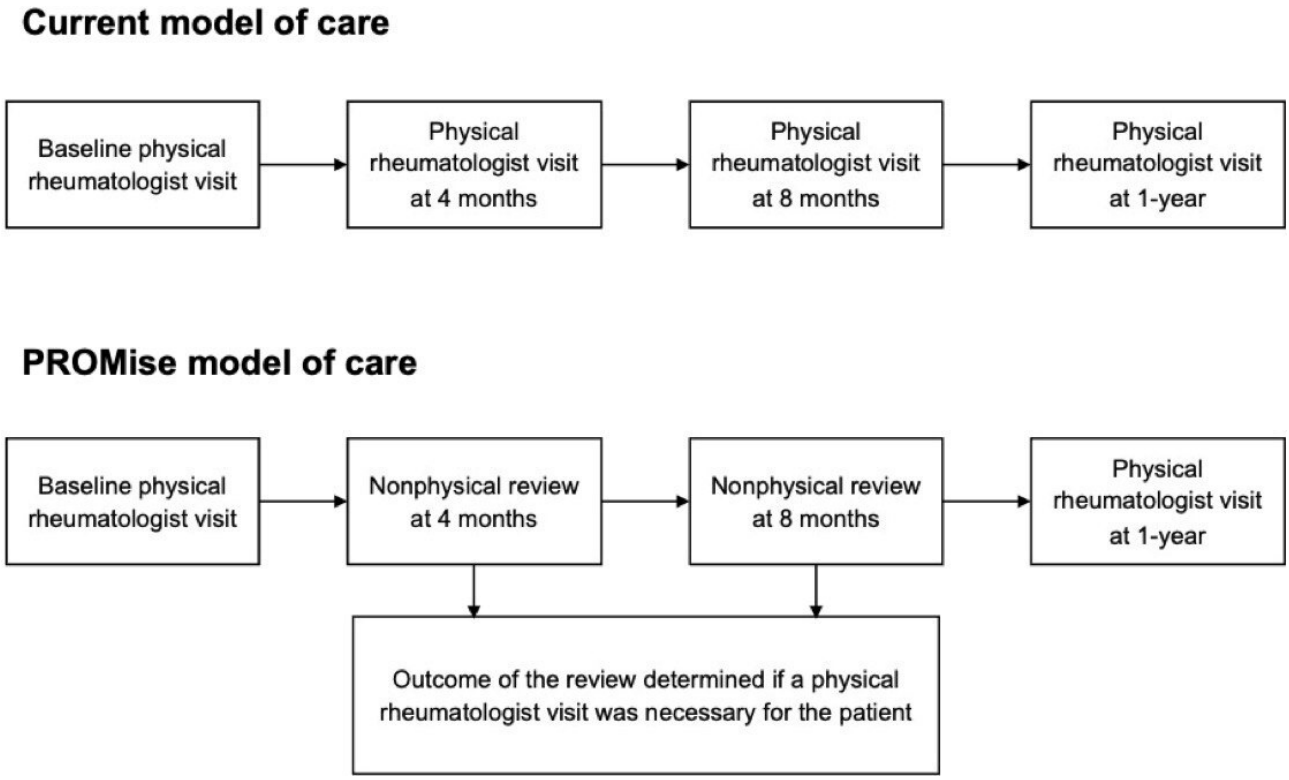

However, the implementation of a new model of care is a complex and challenging process [13]. It has been frequently stated that it takes an average of 17 years for research evidence to reach clinical practice [14]. Hence, implementation science

XSL•FO
RenderX

provides a unique opportunity to expedite the time from evidence generation to clinical implementation [15]. The Consolidated Framework for Implementation Research (CFIR) is a theoretical framework commonly used in implementation science. It consists of 5 major domains (innovation, outer setting, inner setting, individuals, and implementation process) [16]. The Expert Recommendations for Implementing Change (ERIC) is a compilation of 73 implementation strategies that can be used in isolation or in combination with CFIR to tailor implementation strategies [17]. The CFIR-ERIC matching tool links both key resources, allowing users to use ERIC strategies to overcome the identified CFIR barriers [18]. There was a qualitative study by Smits et al [19] in the Netherlands that evaluated the experiences of patients with axSpA and health care professionals (HCPs) with a remote monitoring model of care in axSpA. However, the study did not have an implementation science focus, and it is important to evaluate implementation processes in a local context to ensure innovations can be implemented successfully and sustainably [20].

Therefore, we aimed to understand the facilitators and barriers to implementing a PROMise model in the Singapore context through the CFIR framework and mitigate the CFIR barriers identified through implementation strategies matched by the CFIR-ERIC matching tool. The insights gained from this study will facilitate subsequent implementation of the PROMise model for stable patients with axSpA in Singapore.

Methods

Study Design and Setting

This was a qualitative study relating to the proposed implementation of the PROMise model. The study was conducted in the outpatient rheumatology clinic in Singapore General Hospital. Singapore General Hospital is the largest tertiary hospital in Singapore and serves patients from across the country. In-depth interviews were conducted from November 2024 to May 2025 to elicit both the HCPs' and patients' views on the potential implementation of the PROMise model.

The proposed PROMise model remotely monitors patients with axSpA through an automated review of laboratory tests and PROM results. Stable patients will be identified during their routine visits based on a BASDAI score of lower than 4, which indicates that disease activity is inactive in the patients [21]. The patients will then be given a PROM questionnaire, in this case BASDAI, to fill out prior to their upcoming automated review. The BASDAI questionnaire can be completed remotely in the patients' homes. The schedule of their next appointment will then be determined algorithmically based on the outcome of their automated review (see Figure 1). Patients will have a minimum of one in-person visit to the rheumatologist per year. The PROMise model is under development at the time of this writing, and once refined, is expected to be integrated as part of routine assessment in all future follow-up visits.

Participants

Patients were eligible if they were aged 21 years and older and had been diagnosed with axSpA by a rheumatologist. All patients fulfilled the 2009 Assessment of SpondyloArthritis International Society criteria with a BASDAI score of <4 [2]. HCPs were eligible if they were caring for a patient with axSpA at the time of this writing. HCPs had a diverse variety of roles in the routine care of patients with axSpA; those that could potentially be affected by the implementation of the PROMise model were of particular interest. This included rheumatologists, nurses, allied health professionals, and clinic operations staff. Participants were purposively sampled from a list of patients, who were screened based on eligibility criteria. Purposive sampling was used to ensure representation across key demographic and clinical characteristics. Both patients and HCPs were purposively recruited based on sex, age, and ethnicity. Patients were additionally recruited based on the number of years since diagnosed with axSpA, while HCPs were recruited based on seniority and their role in the care of patients with axSpA. A sampling matrix that included gender, age, ethnicity, and years since diagnosis for patients and years of clinical experience and professional role for HCPs was used to monitor key participant characteristics throughout recruitment. Recruitment was iteratively adjusted to ensure diversity across these characteristics, with targeted efforts made to include underrepresented groups. The eligible patients were approached in person at the outpatient rheumatology clinic by a research staff who explained the purpose of the study.

Data Collection

The CFIR framework was used as a guide to design a standardized, semistructured interview guide (Multimedia Appendix 1) [22]. Individual in-depth interviews were conducted and audio-recorded for both HCPs and patients by 2 trained facilitators (HZ and XRC). Field notes were taken when needed. The interviewers were not part of the care team for the participants. The interviews lasted around 30 minutes on average and were conducted remotely via online video conferencing. Data saturation was defined as the point at which no new codes or subthemes emerged within the CFIR domains represented in this study. Saturation was achieved with at least one code or subtheme identified per domain and confirmed through 3 additional interviews for both patients and HCPs, which did not yield new codes or subthemes within these domains. We also collected basic demographics such as age, sex, education, and ethnicity, as well as the number of years since axSpA was first diagnosed for patients and the number of years in their position at the time of this writing for HCPs.

Data Analysis

The interviews were transcribed verbatim, and the transcripts were checked against the initial audio recording by the coders (HZ, MH, ZX, and XRC) to ensure accuracy and enhance familiarity with the dataset. Coding was performed by independent coders who worked in pairs. Transcripts were thematically analyzed based on Braun and Clarke [23-25], adopting a structured, codebook-based approach and ensuring clarity and transparency in analytic procedures. Each coder independently reviewed transcripts and generated initial codes. The coders then met to compare coding, discuss differences, and organize codes into potential categories and subthemes. A codebook was developed and iteratively refined to document

XSL•FO
RenderX

coding definitions, decisions, and subtheme development. Coding discrepancies were discussed and resolved through consensus, with input from a third researcher (YHK) where required. Regular team discussions were held to review and refine the coding framework and thematic structure to enhance analytic rigor. To enhance trustworthiness, several strategies were used. Credibility was supported through investigator triangulation and repeated engagement with the data. Dependability was strengthened by maintaining an audit trail of coding decisions and codebook development. Confirmability was addressed through team-based discussions to minimize individual bias. Transferability was facilitated by providing contextual details of the study setting and participants to enable assessment of applicability to other contexts. NVivo software (version 15; Lumivero) was used for data management. The reporting of qualitative methods followed the Consolidated Criteria for Reporting Qualitative Research (COREQ) checklist (Checklist 1) [26].

The CFIR-ERIC matching tool was used to develop and outline the expert-endorsed Level 1 and 2 strategies. The identified CFIR barriers were input into the CFIR–ERIC matching tool. The tool's algorithm is derived from an expert consensus panel that ranked how well different implementation strategies address specific CFIR barriers. It then generates a prioritized list of strategies by aggregating the importance ratings assigned by experts across the selected domains [17]. In the development of the CFIR-ERIC matching tool, experts were invited to endorse their top 7 strategies for overcoming a specific CFIR barrier. A total of 18 Level 1 and Level 2 strategies were endorsed by more than 50% of experts and between 20% and 50% of experts, respectively. The CFIR-ERIC output was then analyzed to determine suitable implementation strategies to overcome the main barriers identified.

Ethical Considerations

This study was approved by the SingHealth Centralized Institutional Review Board (Ref No. 2024‑4044). Informed written consent was obtained from all participants before the interviews. All data collected from participants were deidentified. As a token of appreciation, each participant received an SGD $50 (US $39.34) coupon upon completion of the interview. The study was conducted in compliance with the principles of the 1964 Declaration of Helsinki.

Results

Participant Characteristics

A total of 32 interviews were conducted. The participant characteristics are summarized in Table 1. The mean (SD) age for patients was 39.4 (11.7) years and for HCPs was 37.9 (7.2) years, respectively. The HCPs were employed in a diverse range of roles in the clinical setting, from clinic operation staff to rheumatologists; all were involved in the care of patients with axSpA. The HCPs had between 1 year and 23 years of service (Table 1).

XSL•FO
RenderX

Table . Participants characteristics.

Participants	Values
Patients (n=19)	
Age (years), mean (SD)	39.4 (11.7)
Sex, n (%)	
Female	4 (21.1)
Male	15 (78.9)
Race, n (%)	
Chinese	13 (68.4)
Non-Chinese	6 (31.6)
Education, n (%)	
Secondary	2 (10.5)
Tertiary	17 (89.5)
Housing, n (%)	
Private	6 (31.6)
Public	13 (68.4)
Marital status, n (%)	
Married	9 (47.4)
Single	10 (52.6)
Occupational status, n (%)	
Employed	16 (84.2)
National Service	1 (5.3)
Student	1 (5.3)
Unemployed	1 (5.3)
Number of years since diagnosed with axSpA[a], mean (SD)	8.05 (5.21)
HCPs[b] (n=13)	
Age (years), mean (SD)	37.9 (7.2)
Sex, n (%)	
Female	10 (76.9)
Male	3 (23.1)
Race, n (%)	
Chinese	9 (69.2)
Non-Chinese	4 (30.8)
Education, n (%)	
Tertiary	13 (100.0)
Housing, n (%)	
Private	5 (38.5)
Public	8 (61.5)
Marital status, n (%)	
Married	10 (76.9)
Single	3 (23.1)
Occupational status, n (%)	
Employed	13 (100.0)

XSL•FO
RenderX

Participants	Values
Professional role, n (%)	
Clinic operations	3 (23.1)
Doctor	6 (46.2)
Nurse	2 (15.4)
Pharmacist	2 (15.4)
Number of years in the role, mean (SD)	7.7 (6.6)

[a]axSpA: axial spondyloarthritis.

[b]HCP: health care professional,

Facilitators Mapped to CFIR Constructs

Using CFIR as the a priori framework for identifying the barriers and facilitators to implementation of the novel PROMise, all 5 domains (innovation, outer setting, inner setting, individuals, and implementation process) were elicited. We identified 4 main facilitators to the implementation of the PROMise model, which are presented in Table 2.

XSL•FO
RenderX

Table . Facilitators of implementing a patient-reported outcome measures (PROMs)–based model of care mapped to the Consolidated Framework for Implementation Research (CFIR) framework.

CFIR domain	Perceived facilitator	Representative quotes
Innovation (relative advantage)	• Reduced inconvenience and cost • Reduced patient load	• "What I'm looking forward to is the time that I can save … otherwise it's just like half to full day gone … also maybe a little on the finances' side because sometimes I would have to take the taxi down to the hospital." (Patient, Female, 36) • "So I mean most of us are very busy, so it would save me the time to go down, yeah. And of course, there will be cost savings right, because we don't incur the cost of going down there." (Patient, Male, 46) • "The patients' laboratory tests are still being taken, so it is a win-win. The patient feels that someone is looking out for their symptoms and test results. At the same time, I'm spending less clinic time attending to them while they still are feeling okay and satisfied with the care." (HCP[a], Female, 42) • "… reduce the load for clinic with consultations … but they are able to detect any symptoms changes and allow for timely intervention. I would say it reduces everybody's workload, and still effectively allows patients to access hospital's treatment if needed." (HCP, Female, 33)
Inner setting (tension for change)	• Need for the new model of care (PROMise)[b]	• "The clinic slots are very limited. So that's why we are so overwhelmed in clinic itself, because every almost doctor needs to triple book, double book appointments. … at least release some stress from the workload and then the more critical care patients can be seen on time." (HCP, Female, 44) • "The schedule sometimes hinders me, because I need to adjust to the doctor's schedule … because currently I'm working on Mondays to Fridays. … and there's no clinic during Saturday and Sunday." (Patient, Male, 24)
Inner setting (compatibility)	• Similarity to the workflows	• "The process itself is fine, it's already like the old scenario of me coming down for the physical appointment, doing the physical PROMs[c] sheet, and then seeing the doctor's name and doing the blood test as part of the process." (Patient, Male, 30)
Individuals (innovation recipients)	• Suitable patient pool selection	• "I think it is good that we are starting out with the SpA[d] patients because they are relatively young … and their conditions can also be relatively new where they do not have many complications and comorbidities." (HCP, Female, 34) • "… patients with axial spondyloarthritis are also sort of a special group because they tend to be younger, so they are more informed and they can better monitor themselves, so they know when things are not too good and they want to come in earlier, so that's good." (HCP, Male, 45)

[a]HCP: health care professional.

[b]PROMise: patient-reported outcome measures–based model of care.

XSL•FO
RenderX

[c]PROM: patient-reported outcome measure,

[d]SpA: spondyloarthritis.

Innovation (Relative Advantage)

Many patients believed that they would be able to save on the costs and time spent traveling to the hospital. In particular, some patients "look forward to … the time that I can save … otherwise it's just like half to full day gone." (Table 2). Hence, they felt that the remote monitoring model of PROMise would be better than their schedule of physical appointments at the time of this writing.

In addition, HCPs felt that the PROMise model would be able to reduce the patient load in the clinic. Many HCPs also pointed out that even though they will be "spending less clinic time attending" to patients, the patients will still be closely monitored (Table 3). HCPs believe that PROMise will still be "able to detect any symptom changes and allow for timely intervention" (Table 2).

XSL•FO
RenderX

Table . Barriers of implementing a patient-reported outcome measures (PROMs)–based model of care mapped to the Consolidated Framework for Implementation Research (CFIR) framework.

CFIR domain	Perceived barrier	Representative quotes
Innovation (cost)	• Financial sustainability concerns	• "… if this is an incentive for the patient does it mean that it is a disincentive for the organization, … I'm not sure in the future how we will charge our patients for PROM[a] based care. Because manpower is still being pumped to review these blood test results, PROMS results, …" (HCP[b], Female, 34) • "My only question would be how much it would cost, and whether that would be a barrier to us talking to the senior management." (HCP, Female, 41)
Outer setting (local conditions)	• Cultural concerns	• "I think the elderly will be a bit more reluctant to take part in this … because of the language and their eyesight, it may be a bit difficult for them and they are not well versed in the technology …" (Patient, Male, 62) • "… language barrier regarding the questionnaires? … if the patient only speaks like dialect or non-English, non-Mandarin then someone has to interpret for them, I think that would be difficult…" (HCP, Female, 44) • "… I hope they are not deprived of the opportunity to take their time to transition or not transition at all. So, I just hope there's a transition period if there is going to be a change of in the model of care." (Patient, Male, 30)
Individuals (needs)	• Patient safety concerns	• "… but I think if there's a urgent medical condition, maybe it should schedule the appointment earlier ... I think the patients should be informed straight away to go for an earlier appointment?" (Patient, Male, 62) • ".. have a direct helpline to contact the medical staff, … immediately contact the nurse … and they can quickly review or share your results with the doctor and get that assurance. I think if that's done well, it wouldn't compromise my safety or well-being." (Patient, Male, 30)
Implementation process (planning)	• Increased workload concerns	• "… for the patients to be seen in the SpA[c] Clinic within two weeks, that's a little too short. I don't think that the SpA patient needs to be seen so urgently, ... So perhaps four to eight weeks would be more reasonable ... Because the clinician also have very busy clinics that they cannot at two weeks' notice, slot someone in..." (HCP, Female, 42) • "What happens if that they don't do the blood test or they don't answer the questionnaire, then what happens to them? So the workflow process has to be there, and if there's administrative support for this, then that will be good… if the clinician is to be part of this workflow process, then time has to be set aside to vet through all the results of the patients." (HCP, Male, 45)

XSL•FO
RenderX

[a]PROM: patient-reported outcome measure,

[b]HCP: health care professional.

[c]SpA: spondyloarthritis.

Inner Setting (Tension for Change)

Both patients and HCPs felt that there was a need for change in the current model of care. HCPs felt that they were overburdened with clinical load and thus had to "triple book, double book appointment" slots to be able to accommodate the patient numbers (Table 3). Additionally, patients were also unsatisfied with the inconvenience of attending physical consultations. A few patients mentioned that they often face scheduling conflicts when attending physical appointments as they have to "adjust to the doctor's schedule" (Table 2).

Inner Setting (Compatibility)

Patients pointed out that the necessary tasks of "doing the blood tests" and "doing the physical PROMs sheet" were already present in this model of care (Table 3). At the time of this writing, all patients are required to complete laboratory tests and the BASDAI questionnaire before their physical consultation. Many patients pointed out the differences would be that first, the questionnaire will now be administered digitally, and second, a physical appointment may not be required depending on the outcome of the automated review.

Individuals (Innovation Recipients)

HCPs mentioned that patients with axSpA were well suited for the PROMise model. Patients with axSpA were "relatively younger" and hence "do not have many complications and comorbidities" that would require closer monitoring from HCPs (Table 3). Furthermore, it has been raised that the patients with axSpA can "better monitor themselves, so they know when things are not too good and they want to come in earlier" and are able to seek the necessary medical attention if their condition worsens (Table 2).

Barriers Mapped to CFIR Constructs and Level 1 ERIC Strategies

We identified 4 main barriers to the implementation of the PROMise model (Table 3).

Innovation (Cost)

Some HCPs reported that they were concerned with the financial sustainability of the PROMise model. They pointed out that resources such as "manpower is still being pumped" to review the laboratory test and PROMs results; however, they were unsure if these costs were chargeable to the health care system (Table 2). HCPs were worried that the cost of the PROMise model will be unsustainable in the long term as the sources of funding for the PROMise model are still unclear. The lack of clarity regarding the financial sustainability of PROMise will be "a barrier to [the HCPs] talking to the senior management" to seek approval for the PROMise model (Table 2).

The Level 1 ERIC strategy recommended to mitigate the barrier of financial sustainability concerns is to access new funding (82%; Table 4).

XSL•FO
RenderX

Table . Level 1 and 2 strategies identified by the Consolidated Framework for Implementation Research-Expert Recommendations for Implementing Change (CFIR-ERIC) matching tool.

ERIC[a] strategies[b]	Cumulative percent[c]	Cost	Needs	Planning
Conduct local needs assessment	**111%**	4%	**57%**	**50%**
Assess for readiness and identify barriers and facilitators	**92%**	16%	33%	42%
Develop a formal implementation blueprint	**86%**	8%	5%	**73%**
Obtain and use patients or consumers and family feedback	**84%**	4%	**76%**	4%
Access new funding	**80%**	**72%**	—[d]	8%
Involve patients and family members	**75%**	—	**71%**	4%
Alter incentive or allowance structures	**65%**	44%	10%	12%
Conduct local consensus discussions	**56%**	4%	29%	23%
Prepare patients to be active participants	48%	—	48%	
Identify and prepare champions	48%	12%	5%	31%
Develop and implement tools for quality monitoring	45%	—	14%	31%
Use advisory boards and workgroups	44%	—	29%	15%
Tailor strategies	38%	12%	14%	12%
Develop resource sharing agreements	36%	32%	—	4%
Make billing easier	32%	32%	—	
Intervene with patients to enhance uptake & adherence	32%	4%	24%	4%
Facilitation	31%	8%	—	23%
Promote adaptability	30%	16%	14%	—
Conduct educational meetings	29%	12%	10%	8%
Conduct cyclical small tests of change	29%	8%	10%	12%
Capture and share local knowledge	29%	4%	10%	15%
Fund and contract for clinical innovation	28%	28%	—	—
Involve executive boards	25%	20%	5%	—
Provide local technical assistance	24%	4%	5%	15%
Use an implementation adviser	24%	4%	5%	15%
Place innovation on fee for service lists or formularies	24%	24%	—	—
Model and simulate change	24%	20%	—	4%

XSL•FO
RenderX

ERIC[a] strategies[b]	Cumulative percent[c]	Cost	Needs	Planning
Visit other sites	24%	16%	—	8%
Stage implementation scale up	23%	8%	—	15%
Conduct ongoing training	23%	—	—	23%
Build a coalition	22%	4%	14%	4%
Increase demand	22%	12%	10%	
Purposely reexamine the implementation	20%	—	5%	15%
Alter patient fees	20%	20%	—	—
Use other payment schemes	20%	20%	—	—
Identify early adopters	20%	8%	—	12%

[a]ERIC: Expert Recommendations for Implementing Change.

[b]Bolded strategies indicate Level 1 strategies, defined as those endorsed by a majority of experts (>50% agreement), whereas unbolded strategies indicate moderate agreement (20%–50%).

[c]Bolded values indicate that a majority of experts (>50% agreement) endorsed the strategy as a mitigation approach for the corresponding CFIR construct, whereas unbolded values indicate moderate agreement (20%–50%).

[d]Not applicable.

Outer Setting (Local Conditions)

Both patients and HCPs had concerns that some patients would experience challenges with the digital platform, particularly for older adults. Technological proficiency varies across ages, and there is a common perception that the older adult patients "are not well versed in the technology" (Table 3). Hence, they will not be able to readily access the digital PROMs questionnaire.

Both patients and HCPs have also shared concerns regarding language literacy. Patients who do not understand English will not be able to answer the English-based PROMs questionnaire. They would then require "someone to interpret [the questionnaire] for them," which would be an additional hurdle to the implementation of PROMise (Table 3).

One patient was concerned that other patients may not be able to adapt to the new model of care due to similar concerns relating to digital and language literacy mentioned above. This patient hoped that such individuals would be given an "opportunity to take their time to transition or not transition at all" and that there should be a "transition period" to ease patients into the new model of care (Table 3).

Individuals (Needs)

A few patients were concerned that the model of care will not meet their needs, and they will not receive timely medical attention. They were worried that flares or symptoms that worsen or newly emerge may not get the necessary medical attention in time.

Patients suggested establishing "a direct helpline" to the clinic that will allow for timely access to medical assistance, indicating that it would alleviate their safety concerns (Table 3).

The Level 1 ERIC strategies recommended to mitigate the barrier of concerns for patient safety are to involve patients and families (71%), use patients' and families' feedback (76%), and conduct a local needs assessment (57%; Table 4).

Implementation Process (Planning)

Some HCPs raised concerns that their workload would increase under this new model of care. As dedicated "time has to be set aside to vet through" the laboratory data and PROM results from the patients, in addition to the workload that the HCPs have (Table 3). HCPs also worry that they will be unable to accommodate the sudden influx of patients requiring follow-up due to their overbooked clinics.

The Level 1 ERIC strategies recommended to mitigate the barrier of additional workload concerns are to develop a formal implementation blueprint (73%) and conduct a local needs assessment (50%; Table 4).

Strategies Generated by CFIR-ERIC Matching Tool

Overall, the CFIR-ERIC matching tool produced a total of 35 Level 1 and Level 2 strategies (Table 4). There were 8 Level 1 strategies, which had expert agreement of more than 50%. Of the 8 Level 1 ERIC strategies, 7 strategies cut across all 3 barriers of innovation cost, individuals (need), and planning. The strategy of accessing new funding cut across only the barriers of innovation costs and planning, as it did not address the needs of patients. The patients required additional reassurance that they would still receive timely medical attention under this model of care, which could not be resolved by accessing new funding.

The cumulative percentages exceeded 100% because experts were invited to endorse the same implementation strategy for multiple CFIR barriers independently.

Discussion

Principal Findings

This study sought to identify and understand the facilitators and barriers to implementing a novel model of care for patients with axSpA (PROMise) from both the perspectives of patients and

XSL•FO
RenderX

HCPs using the CFIR framework. Facilitators included (1) reduced inconvenience and costs for patients and reduced patient load in the clinic, (2) need for an alternative model of care, (3) similarity to these workflows, and (4) suitable patient selection for PROMise. Barriers included concerns for the (1) financial sustainability of PROMise, (2) cultural conditions, (3) patient safety, and (4) increased workload for HCPs. To the best of our knowledge, this is the first study to break down the facilitators and barriers to the implementation of a novel model of care and suggest implementation strategies to address barriers identified prior to the formal implementation of the novel model of care. This study provides insights for future studies that aim to implement remote monitoring strategies in axSpA or other chronic diseases.

The findings of cost and increased workload being barriers are common in the implementation process [27,28]. In our study, even though there were concerns with the direct costs of the new model of care, the greater concern of the HCPs was the lack of clarity regarding the financial structure of the proposed model of care. HCPs also mentioned a lack of clarity on how to set aside dedicated time to review patients' laboratory tests and PROMs results who require follow-up in the PROMise model. Although both ERIC strategies of accessing new funding and developing an implementation blueprint were recommended for the mitigation of the abovementioned barriers, developing an implementation blueprint would be more relevant in this particular context. The key features of the implementation blueprint should include cost evaluation, patient billing, and detailed workflow for not just clinicians but all HCPs involved in the PROMise model. Hence, an implementation blueprint would shed light on the financial sustainability of the model of care and increase workload transparency for the HCPs [29].

Both patients and HCPs brought up considerations regarding technological and language literacy. Poor technological literacy has been described in local publications [30]. Even though the vast majority of Singapore residents are literate in English (97.1%), there is still a small minority who will face challenges completing the PROM questionnaire in English [31]. For patients who either have challenges in technological or language literacy, more efforts will be needed to determine how support can be provided. Furthermore, these patients are likely to be older and have an increased likelihood of co-morbidities, hence requiring closer monitoring from HCPs [32,33]. However, patients with axSpA are likely to be younger, as pointed out by the HCPs. Hence, the majority of the patients with axSpA will still benefit from the PROMise model. For the small minority of patients with lower technological and language literacy and/or more severe illness, it is likely a longer transition period will be required to ease these patients into the PROMise model as suggested, or to determine if they are suitable at all.

Patient safety was a concern for a minority of patients. Patients desired a direct avenue to seek help in between the scheduled reviews by the algorithm. This was in contrast to findings by other studies where other patients felt a sense of security under a remote monitoring model of care [34]. However, patients in this study did mention that their concerns would be greatly alleviated if they were able to contact the clinic staff directly regarding the scheduling of appointments if they experience worsening symptoms. Hence, the future implementation of the PROMise model should include and inform patients of the access to a direct helpline to address their safety concerns.

With regard to facilitators from the patients' perspectives, the benefits of remote monitoring are well documented. Our study found similar facilitators such as patient convenience and reduced traveling costs [35,36].

Based on the recommended ERIC strategies of conducting a local needs assessment, assessing readiness, and identifying barriers and facilitators; and involving, obtaining, and using patient and family feedback, continuous feedback from patients and HCPs will be collected and used to improve the PROMise model. The continued feedback from patients and HCPs is integral to the sustainability of the PROMise model after implementation [37].

Strengths and Limitations

The strengths of this study include purposive sampling of the HCPs involved in the care of patients with axSpA in Singapore General Hospital. We ensured the inclusion of a wide range of ages, levels of clinical experience, and diverse clinical and nonclinical roles that would be affected by the implementation of the PROMise model. Therefore, our data captured a broad range of views and was representative of the views of HCPs that would be affected by this model of care.

There are a few limitations to our study. First, our study focused on interviews with patients and HCPs involved in the care of patients with axSpA. We lack representation of other stakeholders such as the health care leadership. These stakeholders may have provided a more comprehensive understanding of the barriers and facilitators of the implementation process. However, we tried to mitigate this by including staff involved in operations to understand the barriers and facilitators to the implementation of PROMs. Moving forward, we will solicit feedback during our meetings with the health care leadership. Second, as the proposed PROMise model of care has not been piloted yet, the patients and HCPs interviewed have not had first-hand experience with the model of care. Hence, the perspectives that they provided were only based on their perception of what a PROMise is. They may have a different perspective after they have experienced the model of care personally, especially after the implementation process has been tailored according to recommendations. Hence, we will consistently solicit feedback from patients and HCPs after the model of care has been piloted to better understand and improve upon the implementation process.

Conclusion

Our study found a wide range of facilitators and barriers to the implementation of our novel PROMise model of care, where patients with axSpA will be monitored remotely through laboratory tests and PROM results. We also presented the ERIC strategies that would possibly mitigate the identified CFIR barriers. These findings will guide the implementation of the PROMise model care in patients with axSpA. Future works should evaluate the implementation success of the improved PROMise model of care.

XSL•FO
RenderX

Acknowledgments

The authors thank all participants for their time and contribution to the study. Generative AI was not used at any stage in the preparation of this manuscript.

Funding

This study received grant funding from SingHealth Duke-NUS Academic Medical Centre through grants Pitch-For-Funds 03/FY2024/P2/26-A148 and AM-HEARTS AM/HRT048/2025 (SRDUKAMR25H8), and from the National Medical Research Council of Singapore through grant PHRGNIG24jul-0002 to Dr Kwan Yu Heng. The study also received grant funding from Reverie Rheumatology Research Fund (SPERO program – 03/FY2020/EX/13-A56) to A/Prof Warren Fong. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

YHK contributed to the conceptualization, supervision, funding, data analysis and interpretation, and drafting of the manuscript. XRC contributed to the investigation, data curation, formal analysis, and drafting of the full manuscript. HZ, ZX, MH, KHZ contributed to the investigation, data curation, formal analysis. YYL, TO, LLL, HG, CWYS, SY, CTMW, WF, JT contributed to the formal analysis. All authors contributed to full manuscript editing, read, and approved the final manuscript.

Conflicts of Interest

YYL has received grant from the NMRC (CSAINV24jul-0007), and received honorarium from AbbVie, DKSH, Janssen, Novartis and Pfizer – these have no direct relation with the current study. All other authors have no conflicts of interest to declare.

Multimedia Appendix 1
Interview guide for patients and health care professionals.
[DOCX File, 34 KB - jmir_v28i1e82480_app1.docx]

Checklist 1
COREQ (Consolidated Criteria for Reporting Qualitative Research): 32-item checklist.
[PDF File, 48 KB - jmir_v28i1e82480_app2.pdf]

References

1. Sieper J, Poddubnyy D. Axial spondyloarthritis. Lancet 2017 Jul 1;390(10089):73-84. [doi: 10.1016/S0140-6736(16)31591-4] [Medline: 28110981]
2. Rudwaleit M, van der Heijde D, Landewé R, et al. The development of assessment of spondyloarthritis international society classification criteria for axial spondyloarthritis (part II): validation and final selection. Ann Rheum Dis 2009 Jun;68(6):777-783. [doi: 10.1136/ard.2009.108233] [Medline: 19297344]
3. Poddubnyy D, Hammel L, Heyne M, Veit J, Jentzsch C, Baraliakos X. Treat-to-target strategy with secukinumab as a first-line biological disease modifying anti-rheumatic drug compared to standard-of-care treatment in patients with active axial spondyloarthritis: protocol for a randomised open-label phase III study, AScalate. BMJ Open 2020 Sep 30;10(9):e039059. [doi: 10.1136/bmjopen-2020-039059] [Medline: 32998926]
4. Bohn R, Cooney M, Deodhar A, Curtis JR, Golembesky A. Incidence and prevalence of axial spondyloarthritis: methodologic challenges and gaps in the literature. Clin Exp Rheumatol 2018;36(2):263-274. [Medline: 29148402]
5. Lim WZ, Fong W, Kwan YH, Leung YY. Exploring the prevalence and factors associated with fatigue in axial spondyloarthritis in an Asian cohort in Singapore. Front Med (Lausanne) 2021;8:603941. [doi: 10.3389/fmed.2021.603941] [Medline: 33614681]
6. López-Medina C, Garrido-Castro JL, Castro-Jiménez J, et al. Evaluation of quality of life in patients with axial spondyloarthritis and its association with disease activity, functionality, mobility, and structural damage. Clin Rheumatol 2018 Jun;37(6):1581-1588. [doi: 10.1007/s10067-018-4112-4] [Medline: 29691772]
7. Kwan YH, Fong W, How P, et al. The impact of axial spondyloarthritis on quality of life (QoL): a comparison with the impact of moderate to end-stage chronic kidney disease on QoL. Qual Life Res 2018 Sep;27(9):2321-2327. [doi: 10.1007/s11136-018-1900-x] [Medline: 29872955]

XSL•FO
RenderX

8. Mauro D, Forte G, Poddubnyy D, Ciccia F. The role of early treatment in the management of axial spondyloarthritis: challenges and opportunities. Rheumatol Ther 2024 Feb;11(1):19-34. [doi: 10.1007/s40744-023-00627-0] [Medline: 38108992]
9. Kiltz U, Landewé RBM, van der Heijde D, et al. Development of ASAS quality standards to improve the quality of health and care services for patients with axial spondyloarthritis. Ann Rheum Dis 2020 Feb;79(2):193-201. [doi: 10.1136/annrheumdis-2019-216034] [Medline: 31604704]
10. Kwan YH, Kwoh SY, Phang JK, et al. The direct and indirect costs of axial spondyloarthritis (axSpA) in Singapore. Int J Rheum Dis 2020 Mar;23(3):334-341. [doi: 10.1111/1756-185X.13777] [Medline: 31903711]
11. Data on outpatient visits for systemic lupus erythematosus and spondyloarthritis. Ministry of Health Singapore. 2020. URL: https://www.moh.gov.sg/newsroom/data-on-outpatient-visits-for-systemic-lupus-erythematosus-and-spondyloarthritis/ [accessed 2026-04-11]
12. Hermans K, Webers C, Boonen A, Vonkeman HE, van Tubergen A. Patient-initiated follow-up supported by asynchronous telemedicine versus usual care in spondyloarthritis (TeleSpA-study): a randomised controlled trial of clinical and cost-effectiveness. Lancet Rheumatol 2024 Dec;6(12):e848-e859. [doi: 10.1016/S2665-9913(24)00229-7] [Medline: 39492125]
13. Hailu R, Sousa J, Tang M, Mehrotra A, Uscher-Pines L. Challenges and facilitators in implementing remote patient monitoring programs in primary care. J Gen Intern Med 2024 Oct;39(13):2471-2477. [doi: 10.1007/s11606-023-08557-x] [Medline: 38653884]
14. Westfall JM, Mold J, Fagnan L. Practice-based research--"Blue Highways" on the NIH roadmap. JAMA 2007 Jan 24;297(4):403-406. [doi: 10.1001/jama.297.4.403] [Medline: 17244837]
15. Olswang LB, Prelock PA. Bridging the gap between research and practice: implementation science. J Speech Lang Hear Res 2015 Dec;58(6):S1818-S1826. [doi: 10.1044/2015_JSLHR-L-14-0305] [Medline: 26262822]
16. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implementation Sci 2009 Dec;4(1):50. [doi: 10.1186/1748-5908-4-50]
17. Powell BJ, Waltz TJ, Chinman MJ, et al. A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. Implement Sci 2015 Feb 12;10:21. [doi: 10.1186/s13012-015-0209-1] [Medline: 25889199]
18. Waltz TJ, Powell BJ, Fernández ME, Abadie B, Damschroder LJ. Choosing implementation strategies to address contextual barriers: diversity in recommendations and future directions. Implement Sci 2019 Apr 29;14(1):42. [doi: 10.1186/s13012-019-0892-4] [Medline: 31036028]
19. Smits M, Hermans K, Boonen A, Vonkeman HE, Webers C, Van Tubergen A. POS0496 Experiences of patients and healthcare providers with patient-initiated care and asynchronous telemonitoring in spondyloarthritis: a qualitative study. Ann Rheum Dis 2024 Jun;83:894-895. [doi: 10.1136/annrheumdis-2024-eular.193]
20. McCullough MB, Chou AF, Solomon JL, et al. The interplay of contextual elements in implementation: an ethnographic case study. BMC Health Serv Res 2015 Feb 14;15(1):62. [doi: 10.1186/s12913-015-0713-7] [Medline: 25890333]
21. Garrett S, Jenkinson T, Kennedy LG, Whitelock H, Gaisford P, Calin A. A new approach to defining disease status in ankylosing spondylitis: the Bath Ankylosing Spondylitis Disease Activity Index. J Rheumatol 1994 Dec;21(12):2286-2291. [Medline: 7699630]
22. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated consolidated framework for implementation research based on user feedback. Implement Sci 2022 Oct 29;17(1):75. [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]
23. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
24. Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and be(com)ing a knowing researcher. Int J Transgend Health 2023;24(1):1-6. [doi: 10.1080/26895269.2022.2129597] [Medline: 36713144]
25. Braun V, Clarke V. Conceptual and design thinking for thematic analysis. Qual Psychol 2022;9(1):3-26. [doi: 10.1037/qup0000196]
26. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
27. Oudbier SJ, Souget-Ruff SP, Chen BSJ, Ziesemer KA, Meij HJ, Smets EMA. Implementation barriers and facilitators of remote monitoring, remote consultation and digital care platforms through the eyes of healthcare professionals: a review of reviews. BMJ Open 2024 Jun 10;14(6):e075833. [doi: 10.1136/bmjopen-2023-075833] [Medline: 38858155]
28. VanDevanter N, Vu M, Nguyen A, et al. A qualitative assessment of factors influencing implementation and sustainability of evidence-based tobacco use treatment in Vietnam health centers. Implement Sci 2020 Sep 9;15(1):73. [doi: 10.1186/s13012-020-01035-6] [Medline: 32907603]

XSL•FO
RenderX

29. Hyzak KA, Bunger AC, Herrmann SA, et al. Development of an implementation blueprint to scale-up contraception care for adolescents with psychiatric conditions in a pediatric hospital. Glob Implement Res Appl 2023 Jun;3(2):147-161. [doi: 10.1007/s43477-023-00082-7] [Medline: 38293653]
30. Straughan PT, Tov W, Kim S, Cheng T, Hoskins S, Tan M. Attitudes, behaviours, and the well-being of older Singaporeans in the time of COVID-19: perspectives from the Singapore Life Panel. : Singapore Management University, ROSA; 2020 Jul URL: https://ink.library.smu.edu.sg/rosa_reports/1 [accessed 2026-04-22]
31. Key findings. Singapore Department of Statistics. 2020. URL: https://assets.app.optical.gov.sg/singstat/production/published/assets/b80ab730-e972-42cf-979b-f16546c0b515.pdf [accessed 2026-04-11]
32. Low LL, Kwan YH, Ko MSM, et al. Epidemiologic characteristics of multimorbidity and sociodemographic factors associated with multimorbidity in a rapidly aging Asian country. JAMA Netw Open 2019 Nov 1;2(11):e1915245. [doi: 10.1001/jamanetworkopen.2019.15245] [Medline: 31722030]
33. Statistics SDo. Language literate. In: Singapore Census of Population 2020, Statistical Release 1: Demographic Characteristics, Education, Language and Religion: Department of Statistics, Ministry of Trade and Industry, Republic of Singapore; 2020.
34. Walker RC, Tong A, Howard K, Palmer SC. Patient expectations and experiences of remote monitoring for chronic diseases: systematic review and thematic synthesis of qualitative studies. Int J Med Inform 2019 Apr;124(78-85):78-85. [doi: 10.1016/j.ijmedinf.2019.01.013] [Medline: 30784430]
35. Hansa I, Semaan SJ, Vaid NR, Ferguson DJ. Remote monitoring and “Tele-orthodontics”: concept, scope and applications. Semin Orthod 2018 Dec;24(4):470-481. [doi: 10.1053/j.sodo.2018.10.011]
36. Raatikainen MJP, Uusimaa P, van Ginneken MME, Janssen JPG, Linnaluoto M. Remote monitoring of implantable cardioverter defibrillator patients: a safe, time-saving, and cost-effective means for follow-up. Europace 2008 Oct;10(10):1145-1151. [doi: 10.1093/europace/eun203] [Medline: 18703585]
37. Wiltsey Stirman S, Kimberly J, Cook N, Calloway A, Castro F, Charns M. The sustainability of new programs and innovations: a review of the empirical literature and recommendations for future research. Implement Sci 2012 Mar 14;7:17. [doi: 10.1186/1748-5908-7-17] [Medline: 22417162]

Abbreviations

axSpA: axial spondyloarthritis
BASDAI: Bath Ankylosing Spondylitis Disease Activity Index
CFIR: Consolidated Framework for Implementation Research
COREQ: Consolidated Criteria for Reporting Qualitative Research
ERIC: Expert Recommendations for Implementing Change
HCP: health care professional
HRQoL: Health-Related Quality of Life
PROM: patient-reported outcome measure
PROMise: PROMs-based model of care

Edited by A Stone; submitted 16.Aug.2025; peer-reviewed by C Lei, G Carot-Sans; revised version received 20.Mar.2026; accepted 27.Mar.2026; published 30.Apr.2026.

Please cite as:
Kwan YH, Chew XR, Zainal H, Xiong Z, Hock M, Zee KH, Leung YY, Østbye T, Low LL, Goh H, Sum CWY, Yoon S, Wang CTM, Fong W, Thumboo J
Facilitators and Barriers to Implementing a Remote Monitoring Model of Care for Stable Patients With Axial Spondyloarthritis Using the Consolidated Framework for Implementation Research: Qualitative Study
J Med Internet Res 2026;28:e82480
URL: https://www.jmir.org/2026/1/e82480
doi:10.2196/82480

© Yu Heng Kwan, Xin Ru Chew, Humairah Zainal, Zhonghui Xiong, Melissa Hock, Ka-Hao Zee, Ying-Ying Leung, Truls Østbye, Lian Leng Low, Hendra Goh, Charmaine Wai Yan Sum, Sungwon Yoon, Charmaine Tze May Wang, Warren Fong, Julian Thumboo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Virtual Integrated General Practitioner–Pediatrician Model of Care Implemented in Metropolitan and Rural Primary Care Settings: Qualitative Analysis of Clinician Perspectives on the SUSTAIN Model of Care

Susan Bullock[1], MClinPsych, MIS; Raghu Lingam[1], MSc, MBChB, PhD; Karen Wheeler[1], BSc; Tammy Meyers Morris[1], MBBS, PhD; Corin Miller[1], BSc, MBBS; Louisa Adams[2], MBBS; Ken Peacock[2], MBBS; Annemarie Christie[2], MBBS; Michael Hodgins[1], PhD; Harriet Hiscock[3,4,5], MBBS, MD; Lena Sanci[6], MBBS, PhD; Natalie Taylor[7], MSc, PhD; Carmen Crespo-Gonzalez[1], PhD

[1]Population Child Health Research Group, School of Clinical Medicine, UNSW Sydney, Building C29 HTH Level 3 North, 55 Botany St, Kensington, NSW, Australia

[2]Sydney Children's Hospitals Network, Sydney, Australia

[3]Department of Paediatrics, The University of Melbourne, Melbourne, Australia

[4]Health Services and Economics, Murdoch Children's Research Institute, Melbourne, Australia

[5]Faculty of Medicine, Health and Dental Sciences, The University of Melbourne, Melbourne, Australia

[6]Department of General Practice and Primary Care, The University of Melbourne, Melbourne, Australia

[7]School of Population Health, Faculty of Medicine and Health, UNSW Sydney, Sydney, Australia

Corresponding Author:
Susan Bullock, MClinPsych, MIS
Population Child Health Research Group, School of Clinical Medicine, UNSW Sydney, Building C29 HTH Level 3 North, 55 Botany St, Kensington, NSW, Australia

Abstract

Background: General practitioners (GPs) play a pivotal role in a patient's health care journey. However, demands on general practice, including complex patient management, workforce shortages, and health system fragmentation, have been shown to adversely impact the delivery of high-quality care and health outcomes. Integrated care models, particularly those that offer virtual care options, can support improved access to quality care and efficiency of health care delivery across metropolitan and rural areas. The SUSTAIN model of care was created to provide an accessible option for integrated care. It consists of centralized pediatricians supporting GPs in their practice through virtual coconsultations, virtual "lunch and learn" case discussions, and phone or email support. There is limited evaluation literature on integrated models of care being implemented in a primary care setting where the GP and family are face to face and the non-GP specialist is virtual. To address this gap, a comprehensive implementation evaluation of the SUSTAIN model of care was conducted.

Objective: This study aimed to examine what, why, and how different factors impact the uptake of the SUSTAIN model of care from the perspectives of the SUSTAIN pediatricians and metropolitan and rural GPs in New South Wales, Australia.

Methods: A qualitative study was conducted as part of a mixed methods implementation evaluation of the SUSTAIN model of care. Data were collected via recorded online focus groups and interviews with GPs, practice managers, and pediatricians at 6 and 12 months after the commencement of SUSTAIN. Data were analyzed thematically using iterative thematic analysis informed by the Consolidated Framework of Implementation Research.

Results: Eighteen focus groups and 13 interviews were conducted. GPs, practice managers, and pediatricians found the SUSTAIN model acceptable, with the flexibility and practicality of the model highlighted. GPs valued the learning opportunities, collaboration, and support they gained from working alongside the pediatricians. Virtual delivery through telehealth was viewed as a positive means of receiving specialist support that would otherwise be inaccessible to many practices. Increased efficiency in workflow and working at the top of scope in pediatric care as well as opportunities for meaningful professional relationships and increased family trust in GP-delivered care were recognized as key benefits that enhanced uptake. The current landscape of Australian general practice, with fee-for-service billing and workflow pressures, was recognized as a barrier to engagement with SUSTAIN. GPs and pediatricians highlighted that more appropriate remuneration to support co-consultation is vital to the sustainability and scalability of the SUSTAIN model.

XSL•FO
RenderX

Conclusions: The SUSTAIN model of care expands on our understanding of the benefits of integrated GP-pediatrician models of care in general practice by demonstrating the utility of a pediatrician supporting a GP in their practice via telehealth across metropolitan and rural environments in New South Wales, Australia.

Trial Registration: Australian New Zealand Clinical Trials Registry ACTRN12623000543684; https://www.anzctr.org.au/Trial/Registration/TrialReview.aspx?id=385645

International Registered Report Identifier (IRRID): RR2-10.2196/69728

(J Med Internet Res 2026;28:e86707) doi:10.2196/86707

KEYWORDS

virtual health care; integrated care; rural health; pediatric health; primary care

Introduction

Integrated care models are one way to support improved access to quality care and efficiency of health care delivery [1]. Integrated care has been defined as the structured effort to provide coordinated, person-centered, multidisciplinary care within the same or different organizations, either within health care or across the health, social, or community care sectors [2-4]. Integrated care models have demonstrated positive outcomes, including increased access for patients, interprofessional learnings and communications, health service efficiency, and cost-savings [1,5].

Integrated models of care with pediatricians coconsulting with and supporting general practitioners (GPs) in primary care have been evaluated in studies across the United Kingdom and Australia [6-8]. Results demonstrate reductions in pediatric hospital outpatient appointments, subspecialty pediatric appointments, and emergency department (ED) attendance, as well as improvements in quality of care for several common childhood conditions [6,9]. The Strengthening Care for Children (SC4C) model, based on the United Kingdom's Children and Young People's Health Partnership, comprised in-person GP-pediatrician coconsultations and case discussions and remote email and phone support. SC4C was delivered in 22 metropolitan general practices across Melbourne and Sydney [6-8].

However, in the Australian context, in-person integrated care models, such as SC4C, pose equity and scalability challenges especially for nonmetropolitan general practices where access to local pediatrician support is limited and GP workload capacity stretched [10]. Furthermore, 28% of Australians, or approximately 7 million people, live in rural, regional, and remote areas of the country [11]. Compared to their urban counterparts, children living in these areas experience inferior health outcomes and are more likely to be developmentally vulnerable in one or more domains in their first year of school [12,13]. Developmental vulnerabilities at school entry are strongly associated with long-term adverse outcomes, including school failure, lifelong disability, chronic diseases, mental illness, incarceration, and reduced economic opportunity, perpetuating cycles of disadvantage across generations [12,14]. These health and developmental challenges are further exacerbated by reduced access to appropriate health care services arising from geographical barriers and an increased likelihood of exposure to socioeconomic constraints [12,15,16].

The use of virtual care to support clinical practice has been highlighted as an efficient and cost-effective means for delivering and accessing quality health care services and outcomes [17,18], but some populations struggle to access this care equitably [10]. Telehealth, a virtual care option, is a broad term for health care services provided by practitioners to patients through digital communication channels that allow them to interact with patients, or other clinicians, without a physical presence [19]. The impact of COVID-19 accelerated the use of telehealth, driving more integrated ways of working across health sectors, addressing access difficulties, particularly for regional and rural areas, and overcoming health professional shortages [10,18,20,21]. Benefits arising from the use of telehealth include optimizing patient choice for health care delivery, increasing access to timely care, supporting the integration of the multidisciplinary team to enhance comprehensive care, reducing costs, increasing convenience, and continuity of care [18,22].

The SUSTAIN model of care was created in response to equity issues arising from the SC4C trial to allow for a more accessible option for pediatric integrated care in both metropolitan and nonmetropolitan general practice clinics. The SUSTAIN model consisted of centralized pediatricians consulting with and supporting GPs in their practices through virtual coconsultations and case discussions, phone or email support, and online educational opportunities.

There is limited literature on integrated models of care being implemented in a primary care setting where the GP remains the leading practitioner and consults patients face to face and non-GP specialists are in a support capacity only through telehealth. To address this gap, a comprehensive implementation evaluation of the SUSTAIN model of care was conducted alongside the outcome evaluation. Many health interventions with proven effectiveness are not consistently translated into clinical practice over time [23]. This gap highlights the need for robust evaluation of complex health interventions to understand the active ingredients driving implementation success, which is critical for supporting replication and sustainability. Without an understanding of the factors influencing the implementation of the SUSTAIN model of care among key stakeholders in different settings, its development, replication, and broader application remain limited. Identifying and understanding the contextual determinants that impact the uptake of the SUSTAIN model of care, considering the role of systemic and individual influences, in line with the Consolidated Framework for Implementation Research (CFIR), is vital to its

XSL•FO
RenderX

long-term success [24]. This study aimed to identify and understand the barriers and facilitators perceived by pediatricians, GPs, and general practice staff as moderators of the implementation of the SUSTAIN model.

Methods

Study Design

This qualitative study was conducted as part of the mixed methods implementation evaluation of the SUSTAIN model of care, drawing on the methodology developed for the SC4C trial and detailed in the SC4C implementation evaluation protocol and the SUSTAIN trial protocol [25,26].

The CFIR and the SUSTAIN logic model [26] were used as frameworks to guide the analysis [25,27]. The interview guides for the semistructured interviews and focus groups were informed by the 5 CFIR domains (ie, the innovation, outer setting, inner setting, individuals, and implementation process) and were adapted from guides used in SC4C to be suitable for GPs practicing in both metropolitan and rural areas of New South Wales (NSW), Australia [25]. Standards recommended by the Consolidated Criteria for Reporting Qualitative Research were used to guide the reporting of results from this qualitative study [28]. A COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist is included as Checklist 1.

Intervention

The SUSTAIN model of care consisted of four components: (1) virtual GP-pediatrician coconsultations (up to 6 h/mo per general practice), (2) biweekly 1-hour virtual "lunch and learn" case discussion sessions ("lunch and learn") with multiple practices at a time (the frequency of "lunch and learn" sessions changed after the beginning of the intervention from monthly to biweekly), and (3) same-day phone or (4) same-day email support within business hours. In addition, GPs had complimentary access to the internationally established Sydney Child Health Program (SCHP), an online modular child health training program targeted at GPs [29]. During their 12-month intervention period, participating GPs were able to engage flexibly with the model of care components based on their clinical needs, capacity, and learning style. Following this period, support was withdrawn completely, and practices moved into a sustainability period. Ongoing communication was provided in the lead-up to exiting the intervention to support transitioning out of the model. During the sustainability period, data collection was ongoing, but GPs had no further access to the SUSTAIN pediatricians or SCHP. The SUSTAIN model was delivered by 2 general pediatricians (1.0 FTE) employed by the Sydney Children's Hospitals Network through a competitive employment process. The pediatricians were selected based on their experience as pediatric consultants, their involvement in junior physician training, and their prior work with GPs in regional and rural areas. They were prepared for the role by pediatricians who had previously served in the role for the SC4C project.

Recruitment Procedure

An expression of interest (EOI) process was used to recruit general practice clinics, with general practice clinics across metropolitan, regional, and rural NSW being eligible to participate. An EOI was distributed by partnering primary health networks (Central and Eastern Sydney, South Western Sydney, and South Eastern NSW) for all general practice clinics in these networks or directly through the research team to other interested general practice clinics in rural, regional, and remote NSW. The EOI provided a written overview that outlined the project's background, the eligibility criteria, and the expectations associated with participation. It also detailed the assessment and selection criteria that would be applied during the review process. General practice clinics that were interested in participating were asked to complete an application form to fill in relevant information.

Eligibility criteria for general practice clinics included location and service provision within NSW, being an accredited service or working toward accreditation, use of Best Practice or Medical Director software, and the ability to commit to the project over the designated timeline. More detailed information on the inclusion and exclusion criteria for general practice clinics is outlined in the SUSTAIN study protocol [26].

Setting

Our sample included 19 general practice clinics located across NSW in rural and metropolitan locations. The practices were located across multiple local health districts with Modified Monash Model classifications MM1 (metropolitan areas) to MM5 (small rural towns) [30].

Participant Selection

Purposive sampling was used to ensure a diverse sample of participants was recruited. After the SUSTAIN model of care had been running in a practice for 6 months, all participating GPs and practice managers were invited to participate in a virtual focus group. This invitation included GPs who had not engaged with any component of the model during the first 6 months, based on the information routinely collated by the research team through REDCap (Research Electronic Data Capture) and pediatrician reporting of coconsultation attendance, phone or email support access, and "lunch and learn" attendance as well as access to the SCHP. For GPs who were not actively participating in the model, the implementation evaluation team contacted them directly via email. These communications explained the importance of their involvement in a focus group or interview, particularly to elicit feedback on the reasons for their low engagement and to understand circumstances in which the SUSTAIN model may not have been perceived as useful in their practice context. This invitation to participate in a focus group or interview was also extended (via the practice manager) to any general practice clinic staff (practice nurses and administrative personnel) who were involved in the implementation of the trial. If participating GPs were unable to attend the focus group at 6 months, they were invited to participate in a one-on-one interview at a time of their convenience.

Following the cessation of the model of care at their practice, all participating GPs and practice managers were contacted again to participate in an end-of-intervention virtual focus group or one-on-one interview. All GPs, regardless of their level of

XSL•FO
RenderX

engagement in the model, were encouraged to participate. To minimize responder bias, the SUSTAIN pediatricians did not participate in any focus groups or interviews with GPs and general practice clinic staff. After the intervention period had finalized in all practices, the SUSTAIN pediatricians were also invited to participate in either a focus group or a one-on-one interview. For this study, the term "focus group" was used to refer to any feedback session with more than 1 participant. Figure 1 provides an overview of the planned timing of the implementation evaluation of the SUSTAIN trial in the context of the broader proposed trial design.

Figure 1. SUSTAIN proposed trial design including timing of qualitative data collection at each step. GP: general practitioner.

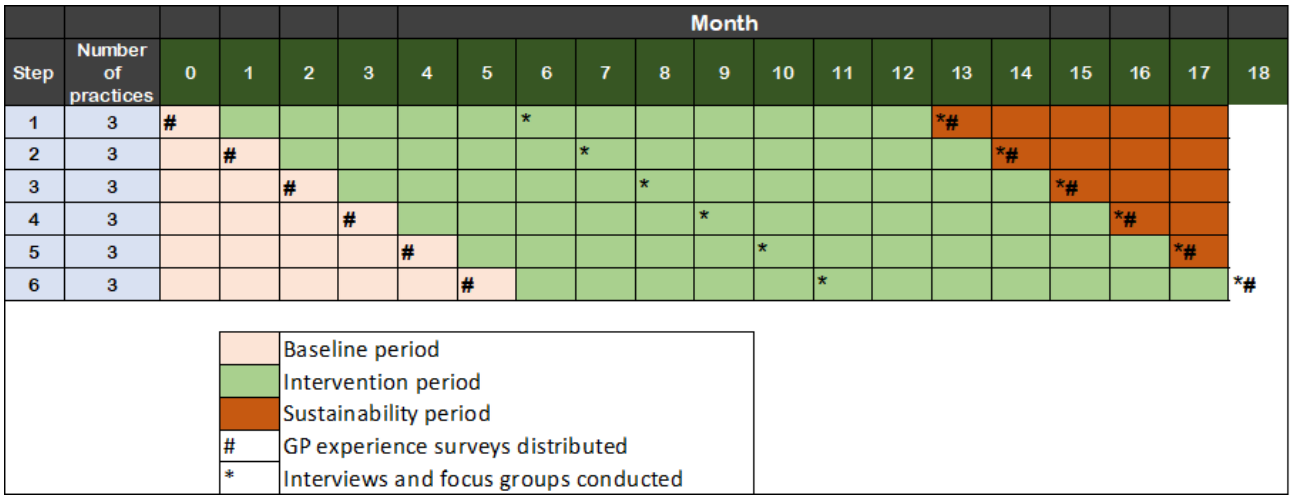

Data Collection

GP demographic data (eg, gender, years of practice, number of pediatric patients seen, formal pediatric training, general practice type, and billing type) were collected as part of an online GP experience survey, which was distributed to all participating GPs in the trial prior to the intervention commencing in their practice (Figure 1). Pediatrician demographic information (gender and years of practice) was collected by the research team directly. Information on the Modified Monash Location of the GP and pediatrician was collected by the research team based on practice location.

Interview and focus group data were collected between April 2024 and February 2025. All focus groups and interviews were conducted by the SUSTAIN implementation evaluation team: CCG (PhD, Research Associate, University of New South Wales [UNSW], Australia, female), SB (MClinPsych, Research Manager, UNSW, Australia, female), and CM (MBBS, PhD candidate, UNSW, Australia, female) led by CCG, an experienced implementation scientist. Where possible, focus groups were conducted by 2 interviewers. Researchers involved in the implementation evaluation came from diverse professional and personal backgrounds. To minimize the influence of prior assumptions and professional experiences on data collection and analysis, researchers approached participant engagement from a neutral standpoint. Reflexivity was embedded throughout the study through regular team discussions, critical reflection during data analysis, and ongoing consideration of how researchers' perspectives and potential unconscious biases may shape interpretation. These reflexive practices were used to enhance the rigor, credibility, and trustworthiness of the qualitative findings.

All interviews and focus groups were completed after written informed consent was collected. Focus groups and interviews with GPs and general practice staff were facilitated virtually using a semistructured guide designed to capture experiences with and adoption of the SUSTAIN model of care. Additional discussion topics included factors moderating the implementation of the model and changes in practice resulting from participation. A modified interview guide was used for the interview with pediatricians, with discussion questions focusing on experiences of barriers and facilitators to the delivery of the intervention and recommendations for future implementation of SUSTAIN. Data collection for the qualitative interviews at 6 and 12 months continued until data saturation (ie, the point at which additional interviews no longer generated new codes, categories, or themes relevant to the research questions and codebook stabilization had been achieved). All interviews and focus groups were audio recorded through an online recording function, downloaded, and stored in a secure drive that was only accessible by the research team. All audio files were initially transcribed verbatim using the Otter.ai transcription service and downloaded and stored alongside the audio files in a secure drive. All online sources were deleted in accordance with our ethics approval. All transcripts were checked for accuracy by SB, CCG, and CM, with modifications being made if the transcript differed from the audio recording. The transcripts were deidentified in preparation for data analysis. Transcripts were not returned to participants for comment. Any words added to quotes by the researchers, to clarify meaning or deidentify locations or key details, were contained in square brackets.

Data Analysis

Transcripts were analyzed thematically using a hybrid inductive-deductive approach using an iterative process of thematic analysis consistent with the methodological framework of Braun et al [31]. The initial phases, (1) data familiarization, (2) generation of initial codes, and (3) searching for themes, were conducted through systematic reading and verification of transcripts by CCG, SB, and CM, with line-by-line open coding

XSL•FO
RenderX

independently conducted by SB and CM. The next steps, (4) grouping and reviewing themes and (5) defining and naming themes, were carried out through collaborative, in-depth discussions among CCG, SB, and CM. These discussions focused on examining the results of the line-by-line coding, including the relationships among identified themes and disagreements in the interpretation of themes. Any disagreements were resolved by jointly reexamining relevant transcript sections, considering the context around participants' statements and drawing upon the CFIR framework to support interpretative consistency. Where necessary, additional themes or subthemes were developed by the team if these were deemed more representative of the content of the transcripts. This iterative process continued until the team reached consensus on a draft coding framework. Two additional transcripts were then further coded by SB using this draft framework, and the results of this extended coding were discussed as a team. During this discussion, any additional themes or subthemes identified by SB that lay outside the preliminary framework were critiqued collaboratively, with the relevant transcripts revisited to determine whether modifications to the coding framework were required. The same analytic approach was used at 6 and 12 months. At 6 months, saturation was reached prior to the completion of data collection. No new inductive codes emerged in the final set of interviews, indicating codebook stabilization, and subsequent data contributed primarily to the refinement and elaboration of existing themes. The 12-month interviews were analyzed against the established codebook and updated iteratively to allow for genuinely new content to emerge. The implementation team met regularly to assess whether successive interviews contributed novel codes or extended existing themes in a meaningful way. From approximately the 10th interview onward, no new inductive codes emerged; subsequent interviews corroborated and elaborated previously identified themes rather than introducing new domains. Given the observed repetition across interviews, the stability of the codebook over multiple consecutive interviews, and team consensus that additional data were unlikely to yield new insights, we judged saturation to have been reached at 12 months. The CFIR was used throughout the analysis as the guiding framework, given its capacity to offer a comprehensive and systematic approach to examining complex implementation challenges. All transcripts were uploaded to and coded in NVivo (version 12; Lumivero).

Ethical Considerations

All participants in this study provided written consent. All data provided by participants was deidentified prior to analysis. Only members of the research team had access to participant data and ensured it was treated according to principles of privacy and confidentiality. Participating general practice clinics received a one-off AUS $1000 (US $713) payment to support implementation of the SUSTAIN trial into their practice. Participating GPs and pediatricians did not receive any additional compensation for participation in interviews or focus groups. This trial was approved by the Human Research Ethics Committee of the Sydney Children's Hospital Network (2022/ETH02068), NSW, Australia.

Results

Overview

Table 1 outlines the number of GPs, practice managers, and pediatricians invited to participate at 6 and 12 months and the number of GPs, practice managers, and pediatricians who participated in either a focus group or a one-on-one interview at 6 months only, 6 and 12 months, and 12 months only. One GP was interviewed at 12 months only. All additional GPs and practice managers (n=9) who participated in the 12-month interviews or focus groups had participated in an interview or focus group at 6 months. The 2 pediatricians were only interviewed once the intervention was completed in all practices.

Table . Overview of general practitioners (GPs), practice managers, and pediatricians who were invited to and participated in focus groups and interviews and 6 and 12 mo.

	GPs			Practice managers			Pediatricians		
	6 months only	6 and 12 months	12 months only	6 months only	6 and 12 months	12 months only	6 months only	6 and 12 months	12 months only
Invited to focus group/interview, n	96	0	74[a]	18	0	17	0	0	2
Participated in focus group, n	40	3	0	5	1	0	0	0	2
Participated in one-on-one interview, n	4	5	1	1	0	0	0	0	0
Total, n	44	8	1	6	1	0	0	0	2

[a]Reduction in the number of GPs and practice managers invited to participate between 6 and 12 mo was due to formal withdrawal from the trial of a general practice clinic (1 practice manager) and 15 GPs between these time points.

No additional practice staff (practice nurses and administrative personnel) consented to participate in the focus groups or one-on-one interviews. Focus group duration ranged from 24 to 83 minutes. One-on-one interview duration ranged from 18 to 41 minutes. Table 2 outlines demographic information for participants of the 6- and 12-month focus groups and interviews.

XSL•FO
RenderX

An overview of the level of engagement of the participating GPs in coconsultations, phone or email support, "lunch and learn" sessions, and the SCHP is presented in Figure S1 and Table S1 in Multimedia Appendix 1.

Five key themes were identified through the thematic analysis: compatibility and adaptability of the SUSTAIN model as drivers of seamless delivery and engagement; the SUSTAIN model's relative advantage in reinforcing GPs' pediatric care capability, access, and motivation; GP-pediatrician relational connections and communication strengthening collaboration and trust; relative priorities, needs, and available resources as key moderators of participants' engagement with the model; and funding and enhancement of model design: considerations for sustainability.

XSL•FO
RenderX

Table . Demographic information of focus group and interview participants.

Characteristics	6-mo interview and focus group participants only		6- and 12-mo interview and focus group participants		12-mo interview and focus group participants only	
	GPs[a] (n=41)[b]	Practice managers (n=6)	GPs (n=8)	Practice managers (n=1)	GPs (n=1)	Pediatricians (n=2)
Gender, n (%)[c]						
Female	27 (66)	—[d]	6 (75)	—	1 (100)	1 (100)
Male	12 (29)	—	2 (25)	—	0 (0)	1 (100)
Prefer not to say	2 (5)	—	0 (0)	—	0 (0)	0 (0)
Modified Monash Model classification, n (%)						
MM1[e]	18 (44)	2 (33)	5 (62)	0 (0)	0 (0)	2 (100)
MM3[f]	11 (27)	1 (17)	0 (0)	0 (0)	0 (0)	0 (0)
MM4[g]	2 (5)	1 (17)	2 (25)	1 (100)	1 (100)	0 (0)
MM5[h]	10 (24)	2 (33)	1 (13)	0 (0)	0 (0)	0 (0)
Pediatric patients seen per week (GPs only), n (%)						
<11	17 (41)	—	1 (13)	—	0 (0)	—
11-20	17 (41)	—	5 (62)	—	1 (100)	—
>20	7 (17)	—	2 (25)	—	0 (0)	—
Years of practice, n (%)						
<6	15 (37)	—	3 (37)	—	0 (0)	0 (0)
6-15	16 (39)	—	3 (37)	—	1 (100)	1 (50)
>15	10 (24)	—	2 (25)	—	0 (0)	1 (50)
Formal pediatric training (GPs only)[i], n (%)						
Yes	14 (34)	—	4 (50)	—	1 (100)	—
General practice type, n (%)						
Solo GP	1 (2)	—	1 (13)	—	0 (0)	—
Group practice	40 (98)	—	7 (87)	—	1 (100)	—
Billing type (for patients aged <18 y), n (%)						
Bulk-billing only[j]	20 (49)	—	3 (37)	—	1 (100)	—
Private billing only[k]	9 (22)	—	1 (13)	—	0 (0)	—
Mixed billing[l]	12 (29)	—	4 (50)	—	0 (0)	—

[a]GP: general practitioner.

[b]Demographic information from 3 GPs who participated in 6-month focus groups was not available.

[c]Gender of practice managers was not formally assessed.

[d]Not applicable.

[e]MM1: metropolitan areas.

[f]MM3: large rural towns.

[g]MM4: medium rural towns.

[h]MM5: small rural towns.

[i]Completion of pediatric training outside of MBBS, MD, or GP Registrar training, including Diploma of Child Health, Sydney Child Health Program prior to SUSTAIN.

[j]Bulk-billing: GPs accept Medicare (Australia's publicly funded universal health insurance scheme) benefit as full payment for service, no charge to patient.

[k]Private billing: patient pays for all or part ("gap") of medical services, Medicare rebate can be claimed by patient for some consultations afterward.

XSL•FO
RenderX

and Medicare rebate can be claimed by patient for some consultations afterward.

[l]Mixed billing: GP bulk-bills some patients, that is, those on health care card or other concession, and others are privately billed.

Theme 1: Compatibility and Adaptability of the SUSTAIN Model as Drivers of Seamless Delivery and Engagement

SUSTAIN was positively received by participants, both the overall model of care and the individual components. As outlined in Textbox 1, participants consistently described the model as well organized and feasible within routine general practice workflows.

> *I do think it's a brilliant idea. It's an absolutely brilliant idea, and it's a pity that we can't do this across all the specialties* [GP26, month 6, MM4]

GPs and practice managers reported a clear understanding of the model of care and described the onboarding process as straightforward, with sufficient resources to integrate SUSTAIN into routine practice with minimal disruption. GPs and pediatricians highlighted variability in how the model was used across practices and individuals, noting that its flexibility was highly valued. Consistent with the CFIR innovation and inner setting domains, Textbox 1 presents that GPs viewed the SUSTAIN model as highly compatible with general practice workflows and its adaptability was seen as a key strength that allowed it to meet the needs of GPs. GPs reported engaging with components of the model that best suited their clinical style, experience, confidence, and workload, with some preferring phone or email support over coconsultations due to scheduling constraints. These reported preferences further illustrate how adaptability supported engagement of individual GPs. Pediatricians also reflected that some GPs would only engage with one component of the model, and this appeared to provide them with the level of support that was required:

> *There were some GPs who clearly just had Thursday one o'clock in their diaries...and I've seen them pop in, and I go, "what case you want to talk about?" "Oh, yeah, there is a child I want to talk about that I saw three days ago, what about this? What about that?" Because that worked for them and their workflow.* [P2, end of intervention, MM1]

Coconsultations were typically reserved for more complex cases, clinical uncertainty, or when parents sought additional reassurance from a pediatrician. Phone or email support and "lunch and learn" sessions were reported as being predominantly used for general queries or recent cases that did not warrant coconsultation. GPs who engaged with the SCHP described it as informative and comprehensive, using it either systematically or to address specific knowledge gaps. However, many GPs reported that they did not participate in the SCHP due to time constraints. Several GPs reported that they preferred the informal and flexible "lunch and learn" sessions, which allowed them to join as needed and better suited the demands and unpredictability of general practice. Together, these findings demonstrate that perceived compatibility with workflow and adaptability to clinician needs were central drivers of GP engagement with the SUSTAIN model, as detailed in Textbox 1.

XSL•FO
RenderX

Textbox 1. Consolidated Framework for Implementation Research (CFIR) constructs and participant quotations for theme 1.

Compatibility and adaptability of the SUSTAIN model as drivers of seamless delivery and engagement

CFIR constructs

- Innovation and inner setting domains: SUSTAIN is compatible with general practice workflow, systems, and processes and adaptable to the needs of the general practitioners (GPs) and general practice clinics.
- Individuals domain: SUSTAIN addresses the need of GPs in pediatrics (increased support, confidence, and knowledge building)

Illustrative quotes

- Model of care
- "The communication was really good. The emails were great…from a running perspective I just think that it was organized well" (GP14, month 6, MM1).
- "I think it's great. Yes, the ultimate solution to paediatric care is to get more paediatricians available, but I'm aware that this project is not aiming to do that, in terms of what this model aims to do, I love it" (GP11, month 6, MM4).
- Coconsultations
- "I've seen a couple of co-consults, which were great. [SUSTAIN paediatrician]'s been really approachable, really helpful, emailing me, happy for me to ring in, easy to make an appointment. All of that's fantastic" (GP27, month 6, MM4).
- Phone or email support
- "I found it [email support] very useful because I got two very, very prompt and very, very appropriate answers" (GP11, end of intervention, MM4).
- "Lunch and Learn"
- "They ['Lunch and Learn'] have been great. I wish I could make it to more of them. But I do try to make it to as many as I can. And they're really valuable. It's that practical application of knowledge, which you'd never find in an online resource or a textbook, anywhere" (GP28, month 6, MM3).
- Sydney Children's Health Program (SCHP)
- "It [SCHP] was really comprehensive. I didn't expect it to have such a broad range of topics and a lot of stuff that I hadn't come across since medical school, just to refresh my memory, learn a bit more with regards to areas that I hadn't had any experience in, and just feeling a little bit more prepared if something does come up left field in clinic" (GP30, end of intervention, MM4).

Theme 2: The SUSTAIN Model's Relative Advantage in Reinforcing GPs' Pediatric Care Scope, Access, and Motivation

Overview

As summarized in Textbox 2, theme 2 predominantly aligned with the CFIR construct of innovation relative advantage within the innovation domain and mapped to key constructs within the inner setting and individuals domains. Participants identified increased efficiency and working at the top of the scope for GPs as the major benefits arising from participation in the SUSTAIN model.

XSL•FO
RenderX

Textbox 2. Consolidated Framework for Implementation Research (CFIR) constructs and participant quotations for theme 2.

The SUSTAIN model's relative advantage in reinforcing general practitioners' (GPs) pediatric care scope, access, and motivation

Efficiency

- CFIR construct
- Innovation domain: The SUSTAIN model offers a relative advantage over normal practice.
- Illustrative quotes
- "It's been really good. It's quite handy being able to access a paediatrician when we want to and being able to speak directly with the consultant as opposed to going to switchboard, and then waiting for switch to call the registrar, and then the registrar may not be certain on their advice and running it by the consultant before getting it back to us. So it's much more useful" (GP20, month 6, MM1).
- "[SUSTAIN paediatricians] reply very quickly, and I get good advice in a timely manner that I can actually use it and utilize it for the patients" (GP49, month 6, MM1).
- "There would have been a few that have saved a referral because I could email and ask" (GP1, month 6, MM3).

Working at top of scope

- CFIR construct
- Inner setting domain: GPs and pediatricians have shared values and beliefs to care, prioritizing recipient-centered care and supporting and addressing the needs of the children and families.
- Individuals domain: Model addressed the need of GPs (increased support, confidence, upskilling, and knowledge building in pediatrics) and the need of families (reassurance and accessible pediatric care).
- Illustrative quotes
- "I think I referred a lot, anyway, but it's definitely saved me some more referrals than I would have done" (GP33, month 6, MM1).
- "One of the things I really appreciate is sort of you pick up some tips, even just by observation, how a person takes a history, how they react to maybe a child who doesn't want to talk. I don't know if I said this to [SUSTAIN paediatrician], but I really liked their low key, calm response and you take that on board. So it's like I say there's a bit of art to medicine, you pick up a bit of the art also" (GP44, month 6, MM3).

Increased Efficiency

Consistent with the CFIR construct of innovation relative advantage (Textbox 2), participants described SUSTAIN as a more streamlined alternative to standard referral procedures and pathways, which often involved contacting hospital switchboards, speaking with rotating registrars, and waiting extended periods for feedback:

> *To have that consistency to talk to someone as opposed to just calling the hospital and speaking to a new reg every time who doesn't know you, and you're interrupting their day. I've had some really complex cases that [SUSTAIN paediatrician] has kind of really guided me on or just confirmed what I was doing. I think it's excellent, I would love it to continue.* [GP18, month 6, MM5]

Access to phone or email support, coconsultations, and "lunch and learn" sessions enabled timely advice, allowing GPs to provide prompt feedback to families. Direct specialist-to-specialist communication with a consultant pediatrician was reported as being particularly valuable in reducing delays and administrative burden on GPs and general practice clinics:

> *We don't even have to write a referral, we just write "this is my question" and then get to have that two-way conversation in the actual co-consult itself, and then and you get the immediate feedback and not needing to wait any period of time. And I'm imagining that saves a fair bit of time writing things out and typing and sending from the administrative side of things from the specialist end as well.* [GP28, month 6, MM3]

In some cases, GPs reported that this immediacy of advice reduced or avoided referrals altogether, while in others it streamlined escalation to subspecialty care. These findings reinforce the perceived efficiency gains and comparative advantage of the model over usual care.

Working at Top of Scope

Theme 2 also aligned with the CFIR construct related to GP-pediatrician shared values focused on recipient-centered care, as well as the SUSTAIN model's responsiveness to GP needs for additional support. Most GPs indicated that their capacity to provide care, support, and management of pediatric conditions had improved through participation in the SUSTAIN model:

> *Each issue I'm managing is making me more confident that the next time I can do this without [SUSTAIN paediatricians] because now I know what to do.* [GP49, month 6, MM1]

GPs reported increased confidence to manage conditions they may previously have referred, including mental health concerns,

XSL•FO
RenderX

constipation, enuresis, and failure to thrive. Several also described improved judgments regarding when escalation to acute or subspecialty services was appropriate. Even highly experienced GPs reported gaining new insights, practical techniques, and resources that strengthened their practice:

> *[SUSTAIN] would save me referring probably 5 to 10 people a year.* [GP48, month 6, MM1]

For GPs who saw fewer children, SUSTAIN was reported to provide accessible support and education around best practice guidelines and referral options. GPs who attended the "lunch and learn" sessions indicated that the learning was 2-fold, with discussions with the SUSTAIN pediatricians and other GPs facilitating ideas, insights, and knowledge building. As detailed in Textbox 2, the perceived relative advantage of SUSTAIN, combined with its impact on GP confidence and capability, underpinned its value in supporting clinicians to work at the top of their scope as GPs.

Theme 3: GP and Pediatrician Relational Connections and Communication Strengthening Collaboration and Trust

Overview

The critical role that professional relationships have in strengthening collaboration and building trust was highlighted by participants. As outlined in Textbox 3, theme 3 mapped primarily to constructs within the inner setting domain, particularly relational connections, communication practices, and shared culture and values centered on recipient-focused care.

Textbox 3. Consolidated Framework for Implementation Research (CFIR) constructs and participant quotations for theme 3.

General practitioner (GP) and pediatrician relational connections and communication strengthening collaboration and trust

Collaboration

- CFIR construct
- Inner setting domain: SUSTAIN GPs-pediatrician relational connection, communication and shared culture, values and beliefs to care, and prioritizing recipient-centered care.
- Illustrative quotes
- "They've [SUSTAIN paediatricians] got an understanding of rural, not everybody does" (GP28, month 6, MM3).
- "I was surprised by some of the complexity that they [GPs] were having to manage, especially the rural and remote GPs...and managing stuff with very limited resources" (P1, end of intervention, MM1).
- "I think once you get to know the paediatrician you are asking the advice from, it just makes it much easier...it's just so much easier to get advice on kids than it was before SUSTAIN" (GP50, month 6, MM1).
- "I think the concept is just so good, it's such a good idea. Especially if you have a relationship with the specialist. So I've met [SUSTAIN paediatrician] in the lunchtime meetings and so it's a lot easier for me to present a case to someone that I've already met, even though I'm only meeting through an audio visual, I think it works a lot better" (GP26, month 6, MM4).

Strengthening family trust in GP-delivered care

- CFIR construct
- Inner setting domain: The GP and pediatrician hold shared values, beliefs, and norms around caring, supporting, and addressing the needs of children and families.
- Individuals domain: Model addressed the need of families to be heard. Model helps build family trust in GP capability.
- Illustrative quotes
- "I think the parents really find the co-consults valuable in terms of they really feel like they got something out of it" (GP2, month 6, MM1).
- "It actually, I think, strengthened the relationship for the GP and the family, because I could go, 'actually, no, everything's sweet, you've done more than what I would have done, so it's great, you've actually covered everything off, and you don't need the specialist'" (P2, end of intervention, MM1).
- "I think once they've heard that once from the paediatrician, they're more likely to then take on board what we're saying down the track, just because they've had those two different sources telling them the same thing, and it's ended up okay. So I think it also helps with the relationship with the families and us as the GPs as well" (GP30, end of intervention, MM4).

Collaboration Between GPs and Pediatricians

Collaboration between practitioners emerged as a key factor in the successful implementation of the SUSTAIN model. Participants described how opportunities for sustained, ongoing interaction fostered mutual understanding of professional contexts, including the complexity of cases managed in general practice and the time pressures faced by GPs, particularly in rural settings. These findings reflect strong relational ties and shared values within the inner setting domain (Textbox 3):

XSL•FO
RenderX

> *Just generally, I don't think you could find better people than [SUSTAIN paediatricians] to kind of be that great resource, technically very astute, very capable, but then also be that kind of person that GPs just feel really comfortable talking to, because not everyone is. And if I was to think about the program in the future, or any program in the future, lots of it is about assessing for fit, so that the clinicians you put in it are actually good at talking to GPs, because not everyone is.* [GP41, end of intervention, MM1]

GPs emphasized the value of specialist-to-specialist partnerships, describing rich learning opportunities supported by safe and inclusive learning environments. The multiple points of contact, including coconsultations and "lunch and learn" sessions, strengthened continuity and deepened professional relationships over time:

> *I suppose just no question felt silly...every question was taken on board, they made you feel like it was interesting, it was helpful information. I think that was really useful.* [GP2, month 6, MM1]

Virtual delivery via telehealth was not viewed as a barrier to relationship development, reflecting adaptation to postpandemic models of care. Although some GPs suggested that optional face-to-face interaction may enhance connection, telehealth was generally considered acceptable and effective, particularly in rural contexts.

Strengthening Family Trust in GP-Delivered Care

Theme 3 also aligned with CFIR constructs within the inner setting domain related to shared values and beliefs supporting recipient-centered care and to constructs within the individuals domain, particularly around addressing family needs and reinforcing trust (Textbox 3).

GPs reported that having the support of the SUSTAIN pediatricians, whether through coconsultation or phone or email support, helped reassure families that their child's medical problem was being appropriately managed by the GP. Participants reported that this reassurance helped increase family confidence in GP-led management and supported continuity of care within primary practice:

> *I'm finding that families are more willing to keep following up with us, as opposed to just waiting for the specialist or going into hospital if things get dire.* [GP30, month 6, MM4]

In some cases, collaborative input changed the trajectory of a patient's journey, reducing the need for referral; in others, it enabled more GPs to feel more confident to manage the "work up" required for specialist care. GPs reflected that this streamlined approach saved time and reduced the burden for families, further reinforcing trust in GP-delivered care. As detailed in (Textbox 3), relational connection and shared commitment to child- and family-centered care were central to strengthening both interprofessional collaboration and family confidence in GP-delivered care.

Theme 4: Relative Priorities, Needs, and Available Resources as Key Moderators of Participants' Engagement With the Model

Heavy clinical workload and limited time were consistently identified as key barriers to engagement with the SUSTAIN model. Consistent with the relative priority construct within the CFIR inner setting domain (Textbox 4), many GPs described SUSTAIN as valuable yet challenging to prioritize within constraints of busy practice environments. GPs seeing fewer pediatric patients (less than 11 per week) reported limited engagement across the intervention period, reflecting a lower perceived need for support. Coconsultations and "lunch and learn" sessions were reported as being particularly affected by time pressures. Scheduling coconsultations was described as challenging in the context of heavily booked clinical calendars, especially in smaller or solo practices. Many GPs commented on the difficulties they faced getting to a "lunch and learn" session due to clinical overrun, administration requirements, and the need for rest between sessions, while evening sessions were often reported as not being feasible due to personal commitments. These findings reflect the influence of work infrastructure and resource constraints within the Inner Setting (Textbox 4).

> *I think the big thing is, we are so under the pump, our wait times are six weeks to book an appointment...there's no way I can easily fit that [co-consultation] in my schedule at the moment to actually work. So it's just not practical. The idea is good but I just don't think it's practical.* [GP 6, month 6, MM5]

Variation in engagement also reflected differences in perceived need, consistent with the CFIR individuals domain (Textbox 4). While many experienced GPs acknowledged that SUSTAIN offered skills and knowledge development, others felt confident managing complex pediatric presentations independently and did not perceive a strong need to engage. Some GPs practicing in rural settings described having adapted over time to managing complexity without specialist input:

> *I've been here for a long time, I've been seeing kids for a long time, I've already got a pattern on what I do for kids. I've got a pattern of where I refer and why I refer and there isn't a local paediatric service... and so SUSTAIN, I hadn't really felt a perceived need for a new system, so I haven't engaged with SUSTAIN very much, that's my perspective on it.* [GP5, month 6, MM5]

In addition, family expectations and preferences influenced use of the SUSTAIN model. Several GPs reported that parents often requested direct referral to a pediatrician in the community and could be insistent on this course of action over engagement in the SUSTAIN model. Together, these findings demonstrate that engagement with SUSTAIN was moderated by competing priorities, structural constraints, and perceived need for pediatric support, as detailed in Textbox 4. While the model was broadly valued, its uptake depended on the alignment between GP and practice capacity, clinician motivation, and perceived need and family expectations.

XSL•FO
RenderX

Textbox 4. Consolidated Framework for Implementation Research (CFIR) constructs and participant quotations for theme 4.

Relative priorities, needs, and available resources as key moderators of participants' engagement with the model.

Workload, time available, and needs of general practitioners (GPs) and families

- CFIR constructs
- Inner setting domain: The implementation of SUSTAIN can be impacted by its relative priority within general practices, as well as the work infrastructure and available resources of these settings.
- Individuals' subdomain: The perceived need for pediatric support through SUSTAIN.
- Illustrative quotes
- "The co-consultations and the 'Lunch and Learn' sessions and the email, it's all really helpful. It's just, I guess the nature of GP isn't an environment that's conducive to ongoing learning really" (GP1, month 6, MM3).
- "After so many years practicing in an environment, I mean, look, we're deprived of many specialties on tap, so you adjust and you become more comfortable with managing more complex things, can't help it, I mean, you sink or swim" (GP37, month 6, MM3).
- "There wasn't always a lot in there ['Lunch and Learn' sessions] for those of us who've been doing this for a little while" (GP38, month 6, MM3).
- "A lot of times the mums just really want to see a paediatrician. So typically, I try to avoid referring if I don't think it's necessary, but sometimes they're quite persistent" (GP46, month 6, MM5).

Theme 5: Funding and Enhancement of Model Design: Considerations for Sustainability

Participants emphasized that long-term sustainability would depend on refinements to both model design and funding structures to support the model's implementation.

Model of Care Design

Most GPs commented favorably on the design of the SUSTAIN model; however, they also identified limitations that reduced its perceived relevance in certain clinical contexts. In particular, the defined boundary that the SUSTAIN pediatricians would not assume ongoing clinical responsibility for children's care limited the model's perceived applicability and relevance for some GPs. This was noted particularly in relation to the diagnosis and management of children with autism spectrum disorder and/or attention-deficit hyperactivity disorder. Mapping to the CFIR implementation process domain, participants indicated a potential need for adapting the model to better reflect and accommodate clinical realities (Textbox 5).

> *The reason I haven't used it much....90% of what I send off to the paeds is behavioural. They are people who are wanting a diagnosis and wanting medication and you guys can't do that. So that's been my main reason for not using you more.* [GP27, month 6, MM4]

The SUSTAIN pediatricians similarly reflected that flexibility in how the model is operationalized would be important for long-term implementation with suggestions of increased scope for the pediatrician to take more clinical responsibility, if clear professional and clinical boundaries were in place. Conversely, several GPs proposed that certain components, such as the frequency of "lunch and learn" sessions or pediatrician availability, could be reduced while maintaining value, potentially supporting widespread roll-out across the state and broader scale-up. A hybrid model incorporating periodic face-to-face engagement alongside telehealth was also suggested by GPs and the SUSTAIN pediatricians as a means of strengthening connection and contextual understanding (Textbox 5).

XSL•FO
RenderX

Textbox 5. Consolidated Framework for Implementation Research (CFIR) constructs and participant quotations for theme 5.

Funding and enhancement of model design: considerations for sustainability

Model of care design

- CFIR constructs
- Implementation process domain: adapting the design of the sustain model to meet gp and general practice needs.
- Illustrative quotes
- "An ideal scenario for us in terms of [a] more focused and intense model than SUSTAIN, so that combination of telehealth, phone a friend, but then also having that paediatric team, or someone on the team come out and spend a couple of days consulting every two months...to me, the combination of telehealth/face to face would be perfect" (GP11, month 6, MM4).
- "It probably would be easier if they came and visited face to face. And you know, every 6 months you had a paediatrician, and you just booked a whole day of kids and had the paediatrician sit with you for a whole day like you would learn a lot that way as well" (GP1, month 6, MM3).
- "While virtual is great and makes it accessible for people, if this was a long term roll out permanent type role, I'd be saying what the paediatrician actually needs is to go out to that area that's regularly using the service, meet the GPs, meet the paediatric service, meet the paediatric service that doesn't exist, whatever it might be to actually get a feel for what they're working in" (P2, month 12, MM1).

Financial considerations

- CFIR constructs
- Outer setting domain: The availability of funding to implement the SUSTAIN model of care in an ongoing capacity.
- Illustrative quotes
- "If it's going to be long term, then of course the paediatricians have to be remunerated for their time and if we're going to be doing longer consults, then perhaps we have to have like an additional thing [financial incentive] attached to it" (GP10, month 6, MM1).
- "For a while there was a Medicare item number where we could do a co-consultation, but I think that was stopped, and I found that a useful thing. I think we've got to move towards this because we're never going to solve the problem of enough specialists, especially in rural areas" (GP26, month 6, MM4).

Financial Considerations

Sustainability was also strongly linked to the availability of funding structures within the Outer setting domain. Although financial barriers did not prevent participation during the trial, many GPs indicated that long-term engagement would depend on appropriate remuneration. Participants highlighted that current Medicare billing arrangements do not adequately support coconsultation models or longer appointments involving 2 clinicians. GPs, practice managers, and pediatricians agreed that either revised billing items or supplementary financial incentives would be required to support routine implementation. Without structural funding alignment, participants expressed concern that the model's scalability and sustainability would be limited.

Overall, as detailed in Textbox 5, sustained implementation of SUSTAIN would require adaptive refinement of the model design and modification of funding mechanisms to ensure feasibility within primary care.

Discussion

Principal Findings

GPs, practice managers, and SUSTAIN pediatricians found the model acceptable, highlighting its flexibility and practicality. GPs valued the learning, collaboration, and specialist support, with the virtual format viewed as an effective way to access non-GP specialist input that would otherwise be unavailable to many general practice clinics. Key benefits included improved workflow efficiency, capacity building in pediatric care, enhanced professional relationships, and increased family trust in GP-delivered care. However, time constraints and workflow pressures within general practice limited engagement, reflecting broader international concerns about primary care sustainability [32-34]. Participants also emphasized that appropriate Medicare reform or alternative remuneration to support coconsultation would be essential for the long-term sustainability of the model.

The findings that the SUSTAIN model of care was acceptable to GPs, practice managers, and pediatricians align with results from the SC4C trial and the Children and Young People's Health Partnership model, further supporting the value of integrated GP-pediatrician models in general practice [35,36]. Its acceptability in rural general practice clinics demonstrates utility beyond metropolitan settings, consistent with NSW Health priorities [37] and extends previous research by examining an entirely virtual model of care across both metropolitan and rural contexts. In line with the UK Medical Research Council guidance on complex interventions, which emphasizes responsiveness to the implementation context [38], uptake of SUSTAIN components was shaped by practice capacity, resources, and GP learning preferences, with participants highlighting that the model's flexible design was better suited to primary care than a prescriptive "one-size-fits-all" approach.

XSL•FO
RenderX

GPs reported that the SUSTAIN model of care provided an efficient and effective pathway to non-GP specialist pediatric support, providing timelier and streamlined access for both GPs and families compared to usual practice. These findings align with other primary care–specialist integration models in Australia [39-42] and internationally [43-45], where improved workflow and management of pediatric patients were similarly observed. Such efficiencies are particularly important in the context of workforce shortages, long working hours, and high burnout among GPs, particularly in rural areas, and may contribute to improved flexibility, work-life balance, and career longevity for GPs [32,33,46-48]. However, the absence of dedicated Medicare items for coconsultation or additional financial incentives provided to general practice clinics to support GP specialist collaboration was identified as a potential barrier to long-term sustainability. This concern reflects findings from the 2024 Royal Australian College of General Practitioners Health of the Nation survey, in which more than half of GPs reported income maintenance as a challenge [46]. Participating GPs who bulk-bill children (aged <16 y) noted financial losses associated with longer consultations compared with shorter standard appointments, including in the context of triple pediatric (aged <16 y) bulk-billing incentives introduced in Australia in 2023 [49]. Attendance at "lunch and learn" sessions and engaging in phone or email advice were also typically reported as being unpaid. Scaling up of models of multidisciplinary integrated care and supporting GPs to work at the top of scope have been recognized as key objectives of primary health care into the future at both national and state levels in Australia [37,50,51]. Providing the financial mechanisms to facilitate these policy objectives and support models such as SUSTAIN to be implemented ongoing is vital.

Despite growth in Australia's pediatric population and increasing rates of childhood chronic illness, GP consultations with children have declined in number and duration [52]. While this trend has been attributed to greater demand from adults and older patients, concerns have been raised that reduced exposure may diminish GP confidence in managing chronic disease and behavioral concerns in children [52]. In contrast, GPs participating in SUSTAIN reported increased confidence and capacity to manage pediatric presentations through education, observation, and improved access to specialist advice and referral pathways, with perceived benefits for patient outcomes. These findings are consistent with the SC4C study, where GPs similarly reported enhanced confidence, knowledge, and skills in pediatric care [35] and reflect similar findings from international studies [43-45,53], supporting the potential of integrated GP-pediatrician models strengthen primary care management of children's health concerns.

Concurrently, the burden of disease and subsequent landscape of pediatric general practice in Australia, and internationally, has shifted toward more complex neurodevelopmental, behavioral, and mental health presentations [54-60], increasing demand for accessible assessment and management pathways particularly for conditions such as attention-deficit hyperactivity disorder [54,61,62]. Evidence suggests that GPs are willing to take a greater role in managing developmental and behavioral conditions when appropriately supported [63,64] aligning with NSW Health recommendations to enhance GP training and support in this area [65] and recent policy reforms expanding prescribing rights for GPs to care for pediatric patients requiring stimulants in NSW [66,67]. Within SUSTAIN, GPs reported improved confidence in undertaking developmental and behavioral assessments, including preparatory work prior to referral; however, some indicated that further adaptation of the model, such as incorporating structured shared care or temporary pediatrician-led management, would be required to fully support complex neurodevelopmental care in primary practice.

International research examining the implementation of integrated pediatric health models has consistently identified the development of a trusted professional relationship between practitioners as fundamental to success [68]. In this study, continuity of interaction through SUSTAIN enabled pediatricians and GPs to build meaningful relationships and mutual trust, reflecting findings from the SC4C trial that emphasized the importance of open, nonjudgmental communication, clear role delineation, and psychologically safe collaborative environments in effective integrated care delivery [35]. These findings are reinforced by a systematic review highlighting that clearly defined roles, shared understanding of collaborative benefits, and dedicated time for communication are key to developing successful interprofessional teamwork in primary care [69], alongside further research demonstrating that regular communication and explicit expectations underpin trust between GPs and non-GP specialists [70-72]. Unlike many comparable pediatric integrated care models that incorporate face-to-face interaction [6-9], SUSTAIN was delivered entirely virtually; however, participants did not perceive the remote format as a barrier to relationship development. Although research specifically examining virtual GP-pediatrician collaboration remains limited, studies of telehealth coconsultation between tertiary specialists and rural clinicians suggest positive experiences and reduced professional isolation [73,74], while evidence from virtual communities of practice indicates that online engagement can mitigate hierarchical barriers and foster equitable, accessible knowledge exchange for rural practitioners [75].

GPs and pediatricians perceived that participation in SUSTAIN strengthened family trust in GP-delivered care, increasing families' willingness to continue engaging with their GP for future health concerns, consistent with findings from the SC4C trial where families reported greater confidence in their GP's pediatric care following participation [35]. These outcomes are particularly relevant, given current help-seeking patterns, with children older than 15 years accounting for the highest rates of lower urgency (ED) presentations and 13% of potentially preventable hospitalizations in Australia [76], and 39% to 60% of pediatric ED visits classified as nonurgent across several Organisation for Economic Co-operation and Development countries [77-81]. Dissatisfaction or mistrust of primary care has been identified as a contributor to nonurgent ED use [82-84], with 1 Australian study reporting that only 45% of parents felt completely confident in GP-delivered care and nearly half preferred pediatrician-led care for their child [83]. Although evidence on telehealth-delivered integrated care models for children remains limited, prior research suggests telehealth can

XSL•FO
RenderX

enhance family perceptions of primary care providers [73]; this study extends that literature by highlighting coconsultation and timely pediatric input within primary care may strengthen family confidence and trust, particularly in rural contexts where telehealth may improve access to care.

Strengths and Limitations

This qualitative analysis had broad engagement from GPs and practice managers across metropolitan and rural general practice clinics within NSW. This representation is a strength and provides evidence for the utility of the SUSTAIN model across a range of settings. However, participating practices were self-selecting. Despite attempts to actively recruit a diverse range of practices and to reach GP practices that care for underrepresented or harder-to-reach populations groups, for example, remote, regional, rural general practices, the cohort of participating practices may be more motivated to participate in research and seek opportunities for professional development. This potentially limits the generalizability of results. For example, although practices participating in the SUSTAIN model were in small, medium, and large rural towns (MM3 to MM5), there were no practices located in remote (MM6) or very remote (MM7) communities in NSW. Further studies to account for additional factors impacting uptake of the model in these communities are required.

As outlined in the *Methods* section, GPs who had not engaged with any component of the model were invited to participate in the interviews and focus groups. However, only one consented to take part. As a result, the findings of this study may not fully capture the perspectives of all the GPs exposed to the model, particularly those with very limited engagement. Furthermore, there was heterogeneity in how GPs engaged with the SUSTAIN model, with individual engagement reflecting differences in preference, clinical style, and workload. As a result, the qualitative themes within this study reflect participants' subjective experiences of the SUSTAIN model as they encountered it. Although variability in exposure to different components is an inherent feature of flexible, practice-embedded implementation models such as SUSTAIN, we acknowledge this as a limitation in the generalization of results.

This study only included perspectives from GPs, practice managers, and the SUSTAIN pediatricians as we were unable to successfully recruit additional practice staff (practice nurses and administrative personnel) to participate. Understanding the implementation of the SUSTAIN model of care from the perspective of the broader practice team is an important consideration for future implementation.

Conclusions

The SUSTAIN model of care expands on our understanding of the benefits of integrated GP-pediatrician models of care in general practice by demonstrating the utility of a pediatrician supporting a GP in their practice via telehealth across metropolitan and rural environments. The nature of the model, with virtual support from the pediatrician, has the potential to improve access to quality health care for children by reducing barriers that many rural and remote families currently face in accessing and affording health care close to home. However, these findings should be interpreted in light of study limitations, including participant self-selection, heterogeneity in engagement with the model, and the absence of remote and very remote (MM6 and MM7) general practice clinics, which may limit generalizability. For the SUSTAIN model to be sustainable and scalable, there is a need for a flexible design to suit the requirements of different general practice clinics and consideration of remuneration to adequately compensate clinicians working within a coconsultation model of care.

Acknowledgments

SUSTAIN is a Translational Research Grants Scheme funded research collaboration between the University of New South Wales, Sydney Children's Hospital Network, and the University of Melbourne. We thank our partnering primary health networks (Central and Eastern Sydney, Southwestern Sydney, and South Eastern NSW), all general practice clinics, general practitioners, pediatricians, and families who participated in the trial. A generative artificial intelligence tool, ChatGPT (OpenAI; GPT-5.2) was used only for language editing and refinement of complex or lengthy paragraphs to improve clarity and readability. These tools were not used for study design, data collection, data analysis, interpretation of results, or generation of original scientific content. All authors take full responsibility for the content of the manuscript and its integrity.

Funding

This study was supported through an New South Wales Ministry of Health Translational Research Grants Scheme grant and in-kind contribution from the Sydney Children's Hospital Network supporting 1.0 FTE pediatrician and complimentary enrollment for participating general practitioners in the Sydney Children's Hospital Network. In kind contribution was also provided by the Central Eastern Sydney, Southwestern Sydney, and South Eastern New South Wales Primary Health Networks to support recruitment and ongoing engagement with general practitioner practices in each network.

Data Availability

Data from this qualitative analysis are available upon reasonable request from the corresponding author.

XSL•FO
RenderX

Authors' Contributions

The original trial design was conceived by RL, AC, HH, LS, and TMM. CC-G, NT, and MH provided expertise on the implementation evaluation framework. SB, CC-G, and CM conducted formal analysis and investigation. SB, CC-G, and KW wrote original draft. CM, TMM, RL, MH, NT, HH, LS, KP, and LA reviewed and edited draft. All authors approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
SUSTAIN_A virtual integrated GP-pediatrician model of care.
[DOCX File, 147 KB - jmir_v28i1e86707_app1.docx]

Checklist 1
SUSTAIN_A virtual integrated GP-pediatrician model of care_COREQ checklist.
[PDF File, 183 KB - jmir_v28i1e86707_app2.pdf]

References

1. Royal Australasian College of Physicians (RACP). Integrated care: physicians supporting better patient outcomes discussion paper. : Royal Australasian College of Physicians. Sydney, NSW; 2018 URL: https://www.racp.edu.au/docs/default-source/advocacy-library/integrated-care-physicians-supporting-better-patient-outcomes-discussion-paper.pdf?sfvrsn=5163091a_4 [accessed 2025-08-25]
2. Struckmann V, Shuftan N, Scarpetti G, Looman W, Bal R, Rutten-van Molken M, et al. How to Implement Integrated Care? A Framework with 12 Overall Strategies to Transform Care Delivery: European Observatory Policy Briefs; 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/39836795 [accessed 2025-08-25]
3. Wolfe I, Satherley RM, Scotney E, Newham J, Lingam R. Integrated care models and child health: a meta-analysis. Pediatrics 2020 Jan 1;145(1):2018-3747. [doi: 10.1542/peds.2018-3747]
4. Wolfe I, Mandeville K, Harrison K, Lingam R. Child survival in England: strengthening governance for health. Health Policy 2017 Nov;121(11):1131-1138. [doi: 10.1016/j.healthpol.2017.09.004] [Medline: 29029812]
5. NSW health strategic framework for integrating care: NSW Ministry of Health. NSW Health. 2018. URL: https://www.health.nsw.gov.au/integratedcare/Pages/strategic-framework-for-integrating-care.aspx [accessed 2025-09-17]
6. Wolfe I, Forman J, Cecil E, et al. Effect of the Children and Young People's Health Partnership model of paediatric integrated care on health service use and child health outcomes: a pragmatic two-arm cluster randomised controlled trial. Lancet Child Adolesc Health 2023 Dec;7(12):830-843. [doi: 10.1016/S2352-4642(23)00216-X] [Medline: 37866369]
7. Khano S, Sanci L, Woolfenden S, et al. Strengthening Care for Children (SC4C): protocol for a stepped wedge cluster randomised controlled trial of an integrated general practitioner-paediatrician model of primary care. BMJ Open 2022 Sep;12(9):e063449. [doi: 10.1136/bmjopen-2022-063449]
8. Hiscock H, Moore C, Khano S, et al. Strengthening Care for Children (SC4C), an integrated paediatrician-general practitioner model for reducing hospital referral rates: a stepped-wedge cluster randomised controlled trial. Med J Aust 2026 Jan;224(1):e70115. [doi: 10.5694/mja2.70115] [Medline: 41391012]
9. Montgomery-Taylor S, Watson M, Klaber R. Child Health General Practice Hubs: a service evaluation. Arch Dis Child 2016 Apr;101(4):333-337. [doi: 10.1136/archdischild-2015-308910] [Medline: 26699536]
10. Miller C, Smithers-Sheedy H, Hu N, et al. Reducing health inequity for children and young people in rural Australia: are digital interventions a panacea? A rural generalist's commentary. Aust J Rural Health 2025 Apr;33(2):e70015. [doi: 10.1111/ajr.70015] [Medline: 40152339]
11. Rural remote health. Australian Institute of Health and Welfare. URL: https://www.aihw.gov.au/reports/rural-remote-australians/rural-and-remote-health [accessed 2025-08-01]
12. Arefadib N, Moore T. Reporting the health and development of children in rural and remote Australia. : The Centre for Community Child Health at the Royal Children's Hospital and the Murdoch Children's Research Institute; 2017 URL: https://www.rch.org.au/uploadedFiles/Main/Content/ccchdev/CCCH-Reporting-the-Health-and-Development-of-Children--in-Rural-and-Remote-Aus-September-2017.pdf [accessed 2025-08-01]
13. 2024 AEDC national report. Australian Early Development Census. 2024. URL: https://www.aedc.gov.au/resources/detail/2024-aedc-national-report [accessed 2025-10-22]
14. Australia's children. Australian Institute of Health and Welfare. URL: https://www.aihw.gov.au/reports/children-youth/australias-children/contents/education/transition-primary-school [accessed 2025-10-02]
15. Aubrey-Basler K, Bursey K, Pike A, et al. Interventions to improve primary healthcare in rural settings: a scoping review. PLoS ONE 2024;19(7):e0305516. [doi: 10.1371/journal.pone.0305516] [Medline: 38990801]

XSL•FO
RenderX

16. Blumenthal D, Gumas ED, Shah A, Gunja, MZ, Williams II RD. Mirror, mirror 2024: a portrait of the failing US health system comparing performance in 10 nations. The Commonwealth Fund. 2024. URL: https://www.commonwealthfund.org/publications/fund-reports/2024/sep/mirror-mirror-2024 [accessed 2025-09-10]
17. Gajarawala SN, Pelkowski JN. Telehealth benefits and barriers. J Nurse Pract 2021 Feb;17(2):218-221. [doi: 10.1016/j.nurpra.2020.09.013] [Medline: 33106751]
18. NSW virtual care strategy 2021-2026. : NSW Ministry of Health; 2022 URL: https://www.nsw.gov.au/sites/default/files/2023-02/nsw-health-virtual-care-strategy-feb-2022.pdf [accessed 2025-09-12]
19. Information for practitioners who provide virtual care Australian Health Practitioner Regulation Agency (AHPRA). AHPRA & National Boards. URL: https://www.ahpra.gov.au/Resources/Information-for-practitioners-who-provide-virtual-care.aspx [accessed 2025-08-10]
20. Flournoy R, Shah R, Moisan E, Oregón C. Telehealth insights from an integrated care system. Am J Manag Care 2024 Sep;30(Spec No. 10):SP751-SP755. [doi: 10.37765/ajmc.2024.89609] [Medline: 39287996]
21. Australia Government Department of Health and Aged Care. MBS review advisory committee telehealth post-implementation review: final report 2024. : Australian Government; 2024 Mar URL: https://www.health.gov.au/sites/default/files/2024-06/mbs-review-advisory-committee-telehealth-post-implementation-review-final-report.pdf [accessed 2025-09-15]
22. Powell RE, Henstenburg JM, Cooper G, Hollander JE, Rising KL. Patient perceptions of telehealth primary care video visits. Ann Fam Med 2017 May;15(3):225-229. [doi: 10.1370/afm.2095]
23. Bauer MS, Kirchner J. Implementation science: what is it and why should I care? Psychiatry Res 2020 Jan;283:112376. [doi: 10.1016/j.psychres.2019.04.025]
24. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated Consolidated Framework for Implementation Research based on user feedback. Implementation Sci 2022;17(1). [doi: 10.1186/s13012-022-01245-0]
25. Hodgins M, Zurynski Y, Burley J, et al. Protocol for the implementation evaluation of an integrated paediatric and primary care model: Strengthening Care for Children (SC4C). BMJ Open 2022 Sep 28;12(9):e063450. [doi: 10.1136/bmjopen-2022-063450] [Medline: 36171022]
26. Meyers Morris T, Hiscock H, Wheeler K, et al. Strengthening care for children using a virtual integrated general practitioner-pediatrician model of primary care (SUSTAIN): Protocol for a stepped wedge cluster randomized controlled trial. JMIR Res Protoc 2026 Jan 14;15:e69728. [doi: 10.2196/69728] [Medline: 41540794]
27. Consolidated Framework for Implementation Research. 2022. URL: https://cfirguide.org/ [accessed 2025-08-08]
28. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
29. Sydney Child Health Program - Essential Paediatrics. The Sydney Children's Hospital Network. 2024. URL: https://schneducation.com.au/essential-paediatrics [accessed 2025-09-30]
30. Modified monash model. Australian Government Department of Health DaA. 2024. URL: https://www.health.gov.au/topics/rural-health-workforce/classifications/mmm [accessed 2025-10-07]
31. Braun V, Clarke V, Hayfield N, Terry G. Thematic analysis. In: Liamputtong P, editor. Handbook of Research Methods in Health Social Sciences: Springer; 2019:843-860. [doi: 10.1007/978-981-10-5251-4_103]
32. Dowell A, Betty B, Gellen C, et al. The concentration of complexity: case mix in New Zealand general practice and the sustainability of primary care. J Prim Health Care 2022 Dec;14(4):302-309. [doi: 10.1071/HC22087] [Medline: 36592774]
33. Hobbs FDR, Bankhead C, Mukhtar T, et al. Clinical workload in UK primary care: a retrospective analysis of 100 million consultations in England, 2007-14. Lancet 2016 Jun 4;387(10035):2323-2330. [doi: 10.1016/S0140-6736(16)00620-6] [Medline: 27059888]
34. Croxson CH, Ashdown HF, Hobbs FR. GPs' perceptions of workload in England: a qualitative interview study. Br J Gen Pract 2017 Feb;67(655):e138-e147. [doi: 10.3399/bjgp17X688849]
35. Crespo-Gonzalez C, Hodgins M, Zurynski Y, et al. Advancing integrated paediatric care in Australian general practices: qualitative insights from the SC4C GP-paediatrician model of care. PLoS ONE 2024;19(5):e0302815. [doi: 10.1371/journal.pone.0302815]
36. Satherley RM, Newham J, Cecil E, et al. Implementation of the children and young people's health partnership model of paediatric integrated care: a mixed-methods process evaluation. Arch Dis Child 2025 Jul 18;110(8):603-611. [doi: 10.1136/archdischild-2024-327664] [Medline: 39904570]
37. NSW regional health strategic plan 2022-2032. : NSW Ministry of Health; 2023 URL: https://www.health.nsw.gov.au/regional/Publications/regional-health-strategic-plan.pdf [accessed 2025-09-10]
38. Skivington K, Matthews L, Simpson SA, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. Int J Nurs Stud 2024 Jun;154:104705. [doi: 10.1016/j.ijnurstu.2024.104705]
39. Wilson M, Mazowita G, Ignaszewski A, et al. Family physician access to specialist advice by telephone: reduction in unnecessary specialist consultations and emergency department visits. Can Fam Physician 2016 Nov;62(11):e668-e676. [Medline: 28661886]

XSL•FO
RenderX

40. Mitchell GK, Burridge L, Zhang J, et al. Systematic review of integrated models of health care delivered at the primary-secondary interface: how effective is it and what determines effectiveness? Aust J Prim Health 2015;21(4):391-408. [doi: 10.1071/PY14172] [Medline: 26329878]
41. Deva M, Osborne J, McGlynn A, et al. Disease focused integrated care - a new model of healthcare delivery for the treatment of skin cancer. Int J Integr Care 2023;23(2):12. [doi: 10.5334/ijic.7009] [Medline: 37151779]
42. Cuesta-Briand B, Rock D, Tayba L, et al. GP perspectives on a psychiatry phone line in Western Australia's Great Southern region: implications for addressing rural GP workload. Aust J Prim Health 2024 Feb;30(1). [doi: 10.1071/PY23039] [Medline: 37743540]
43. Lai L, Liddy C, Keely E, et al. The impact of electronic consultation on a Canadian tertiary care pediatric specialty referral system: a prospective single-center observational study. PLoS ONE 2018;13(1):e0190247. [doi: 10.1371/journal.pone.0190247] [Medline: 29320539]
44. Callahan CW, Malone F, Estroff D, Person DA. Effectiveness of an internet-based store-and-forward telemedicine system for pediatric subspecialty consultation. Arch Pediatr Adolesc Med 2005 Apr;159(4):389-393. [doi: 10.1001/archpedi.159.4.389] [Medline: 15809396]
45. Pashankar DS, Prior D, Arora A, et al. Evaluation and utility of an electronic consultation program in a children's health system. Acad Pediatr 2025 Jul;25(5):102815. [doi: 10.1016/j.acap.2025.102815] [Medline: 40089204]
46. General practice health of the nation 2024. : Royal Australian College of General Practitioners; 2024 URL: https://www.racgp.org.au/FSDEDEV/media/documents/Health-of-the-Nation-2024.pdf [accessed 2025-10-12]
47. Hall LH, Johnson J, Watt I, O'Connor DB. Could breaks reduce general practitioner burnout and improve safety? A daily diary study. PLoS ONE 2024;19(8):e0307513. [doi: 10.1371/journal.pone.0307513] [Medline: 39190672]
48. Vision for general practice and a sustainable healthcare system. : Royal Australian College of General Practitioners; 2024 URL: https://www.racgp.org.au/getattachment/e8ad4284-34d3-48ca-825e-45d58b2d49da/The-Vision-for-general-practice.aspx [accessed 2025-09-10]
49. Medicare bulk billing and out-of-pocket costs of GP attendances over time: australian government. Australian Institute of Health and Welfare. 2024. URL: https://www.aihw.gov.au/reports/medicare/medicare-bulk-billing-of-gp-attendances-over-time/contents/summary [accessed 2025-09-10]
50. NSW Health trategic framework for integrating care. NSW Health. 2018. URL: https://www.health.nsw.gov.au/integratedcare/Pages/strategic-framework-for-integrating-care.aspx [accessed 2026-02-19]
51. Future focused primary health care: australia's primary health care 10 year plan 2022–2032. Australian Government Department of Health, Disability and Ageing. 2022. URL: https://www.health.gov.au/resources/publications/australias-primary-health-care-10-year-plan-2022-2032?language=en [accessed 2026-02-19]
52. Freed GL, Spike NA, Sewell JR, et al. Changes in longer consultations for children in general practice. J Paediatrics Child Health 2013 Apr;49(4):325-329. [doi: 10.1111/jpc.12157]
53. Mazurek MO, Parker RA, Chan J, Kuhlthau K, Sohl K, ECHO Autism Collaborative. Effectiveness of the extension for community health outcomes model as applied to primary care for autism: a partial stepped-wedge randomized clinical trial. JAMA Pediatr 2020 May 1;174(5):e196306. [doi: 10.1001/jamapediatrics.2019.6306] [Medline: 32150229]
54. Freed GL, Turbitt E, Kunin M, Gafforini S, Sanci L, Spike N. General practitioner perspectives on referrals to paediatric public specialty clinics. Aust Fam Physician 2016 Oct;45(10):747-753. [doi: 10.3316/informit.377670138658053] [Medline: 27695726]
55. Charles J, Harrison CM, Britt H. Management of children's psychological problems in general practice 1970-1971, 1990-1991 and 2008-2009. Aust N Z J Psychiatry 2011 Nov;45(11):976-984. [doi: 10.3109/00048674.2011.610743] [Medline: 22017686]
56. Kunin M, Turbitt E, Gafforini SA, Sanci LA, Spike NA, Freed GL. General practitioner referrals to paediatric specialist outpatient clinics: referral goals and parental influence. J Prim Health Care 2018 Mar;10(1):76-80. [doi: 10.1071/HC17030] [Medline: 30068455]
57. Bayram C, Harrison C, Charles J, Britt H. "The kids are alright" - Changes in GP consultations with children 2000-15. Aust Fam Physician 2015 Dec;44(12):877-879. [doi: 10.3316/informit.760751080479859] [Medline: 27054203]
58. General practice health of the nation 2023. : Royal Australian College of General Practitioners; 2023 URL: https://www.racgp.org.au/getmedia/122d4119-a779-41c0-bc67-a8914be52561/Health-of-the-Nation-2023.pdf [accessed 2025-09-30]
59. Jia T, Kong Y, Zhao G, Wang Y. Trends and cross-country inequalities in the global burden of neurodevelopmental disorders among children aged 0-14 from 1990 to 2021. Front Public Health 2025;13:1609254. [doi: 10.3389/fpubh.2025.1609254] [Medline: 40959611]
60. Leeb RT, Danielson ML, Claussen AH, et al. Trends in mental, behavioral, and developmental disorders among children and adolescents in the US, 2016–2021. Prev Chronic Dis 2024;21(240142):cd21. [doi: 10.5888/pcd21.240142]
61. Hiscock H, Danchin MH, Efron D, et al. Trends in paediatric practice in Australia: 2008 and 2013 national audits from the Australian Paediatric Research Network. J Paediatrics Child Health 2017 Jan;53(1):55-61. [doi: 10.1111/jpc.13280]
62. Health services research unit. The Royal Children's Hospital. 2021 Aug. URL: https://www.rch.org.au/hsru/ [accessed 2025-08-30]

XSL•FO
RenderX

63. French B, Perez Vallejos E, Sayal K, Daley D. Awareness of ADHD in primary care: stakeholder perspectives. BMC Fam Pract 2020 Dec;21(1). [doi: 10.1186/s12875-020-01112-1]
64. Kwon S, Bhurawala H, Munoz A, Kramer J, Poulton A. General practitioners' attitudes and knowledge about attention-deficit hyperactivity disorder (ADHD): insights from a survey. Australas Psychiatry 2024 Feb;32(1):18-22. [doi: 10.1177/10398562231211156] [Medline: 37936255]
65. Henry R. Review of health services for children, young people and families within the NSW health system. : NSW Health; 2019 URL: https://www.health.nsw.gov.au/kidsfamilies/paediatric/Documents/henry-review.pdf [accessed 2025-09-12]
66. Game changing reforms allow GPs to treat ADHD to reduce wait times and costs. NSW Government. 2025 May. URL: https://www.nsw.gov.au/ministerial-releases/game-changing-reforms-allow-gps-to-treat-adhd-to-reduce-wait-times-and-costs [accessed 2025-09-10]
67. Kramer J. The changing landscape of ADHD in Australia. RACGP. 2025 Sep 9. URL: https://www1.racgp.org.au/newsgp/gp-opinion/the-changing-landscape-of-adhd-in-australia [accessed 2025-09-10]
68. Watson M, Struthers S, Turner SW. Lessons learnt (so far) from establishing models of integrated clinical care for children and young people. Arch Dis Child 2024 Aug 16;109(9):772-774. [doi: 10.1136/archdischild-2023-325818] [Medline: 37949641]
69. Supper I, Catala O, Lustman M, Chemla C, Bourgueil Y, Letrilliart L. Interprofessional collaboration in primary health care: a review of facilitators and barriers perceived by involved actors. J Public Health (Oxf) 2015 Dec;37(4):716-727. [doi: 10.1093/pubmed/fdu102] [Medline: 25525194]
70. Pidano AE, Arora P, Gipson PY, Hudson BO, Schellinger KB. Psychologists and pediatricians in the primary care sandbox: communication is key to cooperative play. J Clin Psychol Med Settings 2018 Mar;25(1):32-42. [doi: 10.1007/s10880-017-9522-y] [Medline: 29322290]
71. Hoffmann M, Hartmann M, Wensing M, Friederich HC, Haun MW. Potential for integrating mental health specialist video consultations in office-based routine primary care: cross-sectional qualitative study among family physicians. J Med Internet Res 2019 Aug 19;21(8):e13382. [doi: 10.2196/13382] [Medline: 31429419]
72. Dossett LA, Hudson JN, Morris AM, et al. The primary care provider (PCP)-cancer specialist relationship: a systematic review and mixed-methods meta-synthesis. CA Cancer J Clin 2017 Mar;67(2):156-169. [doi: 10.3322/caac.21385] [Medline: 27727446]
73. Marcin JP, Ellis J, Mawis R, Nagrampa E, Nesbitt TS, Dimand RJ. Using telemedicine to provide pediatric subspecialty care to children with special health care needs in an underserved rural community. Pediatrics 2004 Jan;113(1 Pt 1):1-6. [doi: 10.1542/peds.113.1.1] [Medline: 14702439]
74. Santomauro C, McLanders M, Rae A. Experiences of rural clinicians accessing specialist support via telehealth for trauma and emergency care in Queensland, Australia. Digit Health 2024;10:20552076241251950. [doi: 10.1177/20552076241251950] [Medline: 39070889]
75. McLoughlin C, Patel KD, O'Callaghan T, Reeves S. The use of virtual communities of practice to improve interprofessional collaboration and education: findings from an integrated review. J Interprof Care 2018 Mar;32(2):136-142. [doi: 10.1080/13561820.2017.1377692] [Medline: 29161155]
76. Use of emergency departments for lower urgency care: 2025-16 to 2018-19. Australian Institute of Health and Welfare. 2020. URL: https://www.aihw.gov.au/reports/primary-health-care/use-of-ed-for-lower-urgency-care-2018-19/contents/summary/summary [accessed 2025-10-01]
77. Jaeger MW, Ambadwar PB, King AJ, Onukwube JI, Robbins JM. Emergency care of children with ambulatory care sensitive conditions in the United States. J Emerg Med 2015 Nov;49(5):729-739. [doi: 10.1016/j.jemermed.2015.03.001] [Medline: 26037481]
78. Pianucci L, Longacre ML. Nonurgent emergency department use by pediatric patients in the United States: a systematic literature review. Pediatr Emerg Care 2022 Oct 1;38(10):540-544. [doi: 10.1097/PEC.0000000000002718] [Medline: 35580154]
79. Benahmed N, Laokri S, Zhang WH, et al. Determinants of nonurgent use of the emergency department for pediatric patients in 12 hospitals in Belgium. Eur J Pediatr 2012 Dec;171(12):1829-1837. [doi: 10.1007/s00431-012-1853-y] [Medline: 23064744]
80. Vedovetto A, Soriani N, Merlo E, Gregori D. The burden of inappropriate emergency department pediatric visits: why Italy needs an urgent reform. Health Serv Res 2014 Aug;49(4):1290-1305. [doi: 10.1111/1475-6773.12161] [Medline: 24495258]
81. Kubicek K, Liu D, Beaudin C, et al. A profile of nonurgent emergency department use in an urban pediatric hospital. Pediatr Emerg Care 2012 Oct;28(10):977-984. [doi: 10.1097/PEC.0b013e31826c9aab] [Medline: 23023463]
82. Watson M, Blair M. Emergency departments and minor illness: some behavioural insights. Arch Dis Child 2018 Apr;103(4):309-310. [doi: 10.1136/archdischild-2017-314057]
83. Freed GL, Turbitt E, Kunin M, Gafforini S, Sanci L, Spike N. Children referred for specialty care: parental perspectives and preferences on referral, follow‐up and primary care. J Paediatrics Child Health 2017 Jan;53(1):18-25. [doi: 10.1111/jpc.13321]

XSL•FO
RenderX

84. Butun A, Hemingway P. A qualitative systematic review of the reasons for parental attendance at the emergency department with children presenting with minor illness. Int Emerg Nurs 2018 Jan;36:56-62. [doi: 10.1016/j.ienj.2017.07.002] [Medline: 28778488]

Abbreviations

CFIR: Consolidated Framework for Implementation Research
COREQ: Consolidated Criteria for Reporting Qualitative Research
ED: emergency department
EOI: expression of interest
GP: general practitioner
MM: Modified Monash (classification)
NSW: New South Wales
REDCap: Research Electronic Data Capture
SC4C: Strengthening Care for Children
SCHP: Sydney Child Health Program
UNSW: University of New South Wales

Edited by A Stone; submitted 03.Nov.2025; peer-reviewed by J Piera-Jiménez, S Rungan, S Mirji; revised version received 23.Mar.2026; accepted 27.Mar.2026; published 05.May.2026.

Please cite as:
Bullock S, Lingam R, Wheeler K, Meyers Morris T, Miller C, Adams L, Peacock K, Christie A, Hodgins M, Hiscock H, Sanci L, Taylor N, Crespo-Gonzalez C
A Virtual Integrated General Practitioner–Pediatrician Model of Care Implemented in Metropolitan and Rural Primary Care Settings: Qualitative Analysis of Clinician Perspectives on the SUSTAIN Model of Care
J Med Internet Res 2026;28:e86707
URL: https://www.jmir.org/2026/1/e86707
doi:10.2196/86707

© Susan Bullock, Raghu Lingam, Karen Wheeler, Tammy Meyers Morris, Corin Miller, Louisa Adams, Ken Peacock, Annemarie Christie, Michael Hodgins, Harriet Hiscock, Lena Sanci, Natalie Taylor, Carmen Crespo-Gonzalez. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Quality Monitoring for Type 2 Diabetes in Swiss Primary Care: Qualitative Interview Study

Odile-Florence Giger[1], MBA; Lola Jo Ackermann[2], Bsc; Marinja Principe[2,3], Msc; Simon Meier[1], MBA; Elgar Fleisch[1,4], Prof Dr; Susanna Gallani[5], Prof Dr, PhD; Michael Brändle[6], Prof Dr Med; Tobias Kowatsch[2,3,4], Prof Dr; Mia Jovanova[2,4], PhD

[1]Centre for Digital Health Interventions, Institute of Technology Management, University of St. Gallen, St. Gallen, Switzerland
[2]Centre for Digital Health Interventions, School of Medicine, University of St. Gallen, Dufourstrasse 40a, St. Gallen, Switzerland
[3]Institute for Implementation Science in Health Care, Centre for Digital Health Interventions, University of Zurich, Zurich, Switzerland
[4]Centre for Digital Health Interventions, Department of Management, Technology and Economic, ETH Zurich, Zurich, Switzerland
[5]Accounting and Management Unit, Harvard Business School, Harvard University, Boston, MA, United States
[6]HOCH Health Ostschweiz, Cantonal Hospital St. Gallen, St. Gallen, Switzerland

Corresponding Author:
Mia Jovanova, PhD
Centre for Digital Health Interventions, School of Medicine, University of St. Gallen, Dufourstrasse 40a, St. Gallen, Switzerland

Abstract

Background: General practitioners (GPs) manage most type 2 diabetes (T2D) cases worldwide but face increasing workloads. To support guideline-based T2D care, the Swiss Society of Endocrinology and Diabetology (SSED) developed a quality monitoring tool for GPs to track and document patient clinical outcomes, such as glycated hemoglobin testing, blood pressure measurement, and diabetes-related complications in primary care. By aggregating these elements, the SSED score provides a structured way to assess adherence to recommended standards of diabetes care. Although this approach shows promise for improving health outcomes, the score is currently available in paper format and creates an additional administrative burden for physicians.

Objective: This study examined challenges in T2D management in Swiss primary care, assessed limitations of the current SSED quality score, and identified strategies to enable the potential digital implementation and wider adoption of a digital SSED quality score in routine practice.

Methods: A total of 38 qualitative interviews were conducted with 39 participants, representing 4 stakeholder groups relevant to digital T2D care: health care professionals (n=12), individuals with T2D (n=12), software developers (n=10), and health insurer representatives (n=5). Data were analyzed using thematic analysis.

Results: Participants highlighted persistent challenges in T2D management, including time pressure for health care professionals, fragmented care, and a lack of personalized support for individuals with T2D. Barriers to digital quality monitoring included poor integration with GP systems, misaligned incentives, and limited relevance to multimorbidity. Suggested facilitators included embedding digital tools, such as a digital SSED score, into existing workflows, optimizing task shifting for GPs, integrating patient-centered features, and improving data sharing.

Conclusions: A digital SSED score could support GPs by integrating patient-entered data, generating automated alerts, and enabling pharmacy-based follow-up. Successful adoption is likely to depend on GP workflow integration, system interoperability, and aligned incentives. Involving stakeholders early is essential to ensure that new digital tools meet the needs of both individuals with T2D and GPs.

(J Med Internet Res 2026;28:e82960) doi:10.2196/82960

KEYWORDS

type 2 diabetes; patient-centered care; primary care; interviews; Switzerland; integrated care; digital health; qualitative

Introduction

General practitioners (GPs) worldwide face rising workloads, particularly in chronic disease management [1] and a declining workforce [2]. Among these conditions, type 2 diabetes (T2D) represents a major and growing challenge due to its increasing global prevalence and associated health care costs [3]. As primary care providers, GPs play a central role in managing T2D, yet high workloads can limit their ability to deliver optimal care [4].

These challenges are particularly pronounced in high-spending health care systems such as Switzerland, which ranks among

XSL•FO
RenderX

the highest globally in per capita health care expenditure [5]. Despite this investment, approximately 6.1% of the Swiss population is affected by diabetes [6], contributing to a rising burden on primary care services. As diabetes prevalence continues to increase, GPs face mounting pressure to provide high-quality, long-term T2D management [7]. Together, these trends highlight the need for innovative approaches to better support GPs in managing T2D effectively.

At the same time, there is increasing evidence that structured monitoring and guideline-based diabetes management can contribute to improved patient outcomes. International studies suggest that patients who regularly receive recommended monitoring tests, such as glycated hemoglobin (HbA_{1c}) measurements, lipid profiles, and kidney function assessments, have a lower risk of diabetes-related complications and hospitalizations [8]. Similarly, real-world evidence from Switzerland indicates that structured diabetes disease management programs can improve adherence to recommended monitoring processes while also slowing the growth of health care costs compared to usual care [9]. These findings highlight the potential value of systematic quality monitoring approaches in chronic disease management.

To support GPs, the Swiss Society of Endocrinology and Diabetology (SSED) developed the "SSED score," a self-reported instrument to help GPs capture key clinical indicators, including HbA_{1c}, blood pressure, and body weight during patient visits [10,11]. However, the SSED score is paper-based, making it susceptible to human error and limiting opportunities for longitudinal analyses and visualization of disease progression [12,13]. Moreover, the score lacks opportunities for data aggregation, automated scheduling of follow-up visits for individuals with T2D, and tailored lifestyle support. All these functionalities are increasingly relevant in modern long-term care [14].

Improving T2D management, therefore, requires a fundamental rethinking of how a quality assessment score is designed and implemented in practice. This includes not only the potential digitization but also integration into workflows, data infrastructure, and engagement processes. From a digital health adoption perspective, prior research suggests that the uptake of new technologies in clinical practice is influenced by perceived usefulness and perceived ease of use, as conceptualized in the technology acceptance model [15]. In primary care settings, perceptions of a quality monitoring score are likely shaped not only by the technical design of a tool but also by its compatibility with existing workflows [16] and its ability to reduce GP workload [17]. A key challenge is therefore how to improve quality assessment in a way that supports GPs in making more informed and efficient decisions while also addressing system-level (eg, workflow integration) constraints. To this end, we conduct a qualitative interview study to inductively explore stakeholder perspectives on digital quality monitoring for T2D management in Swiss primary care, with a focus on identifying challenges, limitations of existing quality assessment scores, and strategies to support their adoption in routine practice. We specifically examine three key research objectives (ROs):

1. RO1: identify current challenges in T2D management in primary care in Switzerland.
2. RO2: examine limitations of existing T2D quality assessment scores (eg, SSED score), including gaps in relevance, usability, and alignment with real-world care delivery.
3. RO3: develop strategies to promote the adoption of T2D quality assessment scores (eg, SSED score), addressing technical, organizational, and incentive-related factors to ensure a better fit with the complexities of chronic disease management in primary care.

Methods

Development of the Interview Guides

We designed semistructured interview guides tailored to each stakeholder group (refer to Multimedia Appendix 1). Each guide was structured around key thematic areas to ensure consistency across interviewees. At the same time, the format remained flexible, allowing interviewers to pursue unstructured, spontaneous questions where appropriate to explore emerging insights and context-specific experiences in greater depth.

Selection and Recruitment of Interviewees

The interviews were conducted between November 27, 2024, and May 3, 2025. Health care professionals, health care software developers, and health insurers were recruited using purposive sampling. Participants were identified based on predefined criteria, including (1) professional involvement in T2D care, health software development, or health insurance decision-making; (2) experience with quality measurement or care coordination; and (3) relevance to the Swiss health care system. Potential participants were identified through the research team's professional networks and publicly available information (eg, organizational websites and LinkedIn [Linkedin Corporation]).

Additionally, we used snowball sampling by asking the initial interviewees to suggest further participants who met our selection criteria, resulting in a chain of referrals.

Individuals with T2D were recruited via TestingTime (Norstad Group) [18], an online platform that compensates individuals for study participation. The inclusion criteria were that participants had been diagnosed with T2D, resided in Switzerland, and spoke either German or English.

Recruitment ended upon reaching theoretical saturation, defined as the point when 3 consecutive interviews revealed no new concepts [19,20].

A total of 38 qualitative interviews were conducted with 39 participants, representing 4 stakeholder groups relevant to digital T2D care: health care professionals (n=12), individuals with T2D (n=12), software developers (n=10), and health insurer representatives (n=5) (refer to Multimedia Appendix 2).

Most interviews were conducted on an individual basis. One interview with a health care software provider was conducted as a joint interview with 2 participants, resulting in a higher number of participants than interviews. Throughout the paper,

XSL•FO
RenderX

participant counts (rather than interview counts) are reported to ensure consistency and transparency.

Data Collection and Data Analysis

All interviews were conducted via video conference by 1 of 4 members of the research team (LJA, MP, OFG, or SM), all based in Switzerland. The interviewers were familiar with the study context. OFG had training and experience in conducting qualitative interviews and provided guidance and preparatory instruction to the other researchers before data collection, ensuring a consistent and structured interview approach. LJA conducted one interview with a software provider with whom she had prior professional contact, and SM conducted interviews with 3 GPs with whom he had prior contact. All remaining participants were not known to the interviewer who conducted the respective interview before the study. Interviews were conducted primarily in (Swiss) German (n=36), with 2 interviews conducted in English.

All interviews were digitally recorded via Microsoft Teams. Transcripts were generated using Microsoft Teams' automated transcription function and were subsequently carefully reviewed by the research team. Any transcription errors or missing passages were manually corrected or added by relistening to the audio recordings before analysis. Transcripts were first analyzed and coded in German. Following the analytical process, codes and themes were translated into English for reporting purposes, with translation focusing on conceptual rather than literal equivalence. Given the systematic manual review and correction of transcripts before coding, automated transcription errors were considered unlikely to have materially affected the coding and interpretation of the data.

Next, transcripts were anonymized and assigned a unique participant ID. Data analyses were carried out in parallel using ATLAS.ti software (version 25.0.1; Lumivero, LLC). To ensure methodological rigor and transparency, we followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines throughout the study (Checklist 1) [21].

Data were analyzed inductively to identify themes grounded in participants' experiences and perspectives related to T2D management and digital tool adoption.

The research team's positionality was shaped by backgrounds in computer science and business innovation, with a shared focus on digital health and health care innovation. Reflexivity was addressed at the team level through regular analytical exchange. Emerging findings, preliminary codes, and interpretations were systematically discussed and compared in weekly research meetings, allowing the team to reflect on differing perspectives, challenge initial assumptions, and iteratively align interpretations throughout the analysis process.

To enhance the rigor of the qualitative analysis, a subset of the interview transcripts was double-coded. Specifically, 6 of the 38 transcripts were independently coded by 2 researchers. Double-coding was conducted in 3 coding pairs: LJA-MP, SM-OFG, and LJA-OFG.

In an initial analytical step, separate preliminary codebooks were developed inductively for each stakeholder perspective to capture perspective-specific concepts and language reflected in the data. These first-order codes closely followed participants' own terms and meanings. Discrepancies identified during double-coding were discussed within the respective coding pairs in iterative consensus discussions, which focused on clarifying code definitions and interpretations and informed refinements of the perspective-specific codebooks.

In a subsequent step, the perspective-specific codebooks were systematically compared and overlaid to identify overlapping concepts and recurring patterns across stakeholder perspectives. For example, issues related to fragmentation of care emerged consistently across multiple perspectives, albeit described using different terms. Based on these overlaps, first-order codes were progressively merged and abstracted into higher-level conceptual categories and themes.

LJA and OFG iteratively refined, consolidated, and validated the final set of codes and themes through discussion until the thematic structure coherently captured both cross-cutting and perspective-specific insights across the dataset.

In addition to the interviews, one follow-up clarification was obtained via email from an interview participant during the paper revision process. This exchange focused on a specific question related to anticipated changes in the Swiss billing system and was used to support contextual interpretation of the findings.

Ethical Considerations

This study was reviewed by the institution's ethics committee (HSG Ethics Committee), which determined that the project was exempt from formal ethics review (decision date: October 27, 2024). Written informed consent was obtained from all participants before participation in the interviews. All data were treated confidentially and processed in accordance with applicable data protection regulations. Interview data were anonymized before analysis, and no directly identifiable personal information was included in the study results. Data were stored securely on institutional systems accessible only to the research team. Individuals with T2D who were recruited via TestingTime received financial compensation of CHF 20 (US $25.53) for their participation, in accordance with the platform's standard reimbursement procedures. Health care professionals, software developers, and representatives of health insurers did not receive financial compensation.

Results

Overview

We conducted a total of 38 individual interviews, with the coding framework described in Tables 1-4. Interviews were structured in 3 main parts. In the first part of the interview, participants were asked about their current challenges in T2D management (RO1). In the second part, they were asked about the limitations of existing T2D quality assessment scores (eg, SSED score; RO2). In the third part, participants discussed recommendations to address these limitations through actionable strategies. These strategies aimed to promote the adoption of T2D quality assessment scores (eg, SSED score), considering technical, organizational, and incentive-related factors and the

XSL•FO
RenderX

complexities of chronic disease management in primary care (RO3).

Table . Themes and codes for the status quo and key challenges in primary care.

Themes	Codes
Systemic and structural burdens	• Communication with other health care providers • Absence of holistic care • Time pressure • Shortage of skilled workers
Lifestyle and personal challenges	• Difficult maintaining a healthy diet • Low motivation for exercise • Being in denial about the disease
Primary care challenges	• Late or unexpected diagnosis • Insufficient and impersonal information • Absence of holistic care • Limited clinic availability
Structural and technological barriers	• Financial constraints • Access barriers to sports facilities • Limited data sharing of vital parameters

Table . Themes and codes for guideline adoption barriers.

Themes	Codes
Monetary and nonmonetary rewards	• Limited ability to decide how to use financial rewards • Only consultations are compensated, not administrative tasks • The reach of incentive programs is low • Exclusion of noncompliant patients • Nonmonetary rewards are not perceived as valuable
Fragmentation and interoperability	• Fragmentation of guidelines • Fragmentation in digital infrastructure and a lack of system interoperability • Manually re-enter health data • Inadequate quality of data entry
Tension between guidelines and real-world practice	• Perceived lack of value in guidelines, as they do not necessarily reflect the quality of care • Score components depend on patient adherence • GPs[a] resist transparency in benchmarking performance • Patients do not have diabetes alone, but from multiple chronic conditions
Regulatory barriers	• Strict regulations regarding patient data (Switzerland) • Classification as a medical device

[a]GP: general practitioner.

Table . Codes for essential features for the Swiss Society of Endocrinology and Diabetology (SSED) score.

Themes	Codes
Reducing structural burdens	• Integration or compatibility with clinical information systems • Integration of patient data
Targeted digital interventions	• Labeling and flagging system • Reminder and alert mechanisms • Multidisease scores • Inclusion of patients with prediabetes
Aligning with patient needs	• Collaborative goal setting • Customizable patient score

XSL•FO
RenderX

Table . Codes for adoption strategies for the Swiss Society of Endocrinology and Diabetology (SSED) score.

Themes	Codes
Task shifting	• Involving MPAs[a], MPKs[b], and APNs[c]
Aligning financial incentives with quality care	• Financial incentives • Fund continuing education • Replacement of the outdated TARMED[d] system
Boosting communication and data sharing	• Improved communication • Sharing and comparing data • Feedback loops • Automated solutions to support decision-making
Leveraging digital tools to support clinical decision-making	• Automated solutions to support clinical decision-making
Strengthening patient support	• Updated and personalized information • Collaboration between GPs[e] and pharmacies

[a]MPA: medical practice assistant.
[b]MPK: medical practice coordinator.
[c]APN: advanced practice nurse.
[d]TARMED: tarif médical.
[e]GP: general practitioner.

Key Challenges With T2D Management in Primary Care

This section outlines the key challenges of T2D management (RO1).

Systemic and Structural Burdens

In typical diabetes consultations, GPs often manage individuals with T2D with multiple comorbidities within limited timeframes, relying heavily on support staff. Six of 10 (60%) GPs mentioned that they occasionally rely on the support of endocrinologists to ensure high-quality care. Furthermore, 5 of 12 individuals with T2D (~42%) were unhappy with the absence of holistic care. They felt that GPs addressed only one issue at a time, with little coordination between providers: “[…] everything is an isolated system.” (P8)

Five of 12 (~42%) health care professionals reported experiencing constant time pressure in their daily practice. With the next person typically waiting, in-depth discussions or more intensive consultations become difficult to manage. As a result, GPs often struggle to provide the level of attention and care they would ideally like to offer. GP2 stated: “Initially, it is certainly the case that people with diabetes require a great deal of effort if you want to inform them properly…. Instead, we refer them to hospitals, which have diabetes counseling services and dietetic counseling units, because we cannot afford this time investment. Endocrinologists have more time, and they are also able to bill this very well. We, on the other hand, are not.”

Additionally, 3 of 12 (25%) GPs noted a growing shortage of skilled workers, whereby medical practice assistants (MPAs) and medical practice coordinators (MPKs) are increasingly difficult to recruit, as mentioned by GP7: “It has become quite difficult, and I believe even in larger centers it has become challenging in some cases. It is not yet as severe as it is for physicians, but MPAs are in high demand.” Furthermore, it was mentioned that even when MPAs are available, they often do not remain in the same practice long enough to complete further training. As a result of this shortage, other health care professionals are forced to take on additional responsibilities, particularly administrative tasks.

Primary Care Challenges

Five of 12 (~42%) individuals with T2D reported having a generally trusting relationship with their GP. However, 6 of 12 (50%) individuals with T2D stated that the diagnosis was late or unexpected. Participants were often diagnosed during unrelated check-ups or procedures, or after requesting further testing themselves. This may reflect not only delayed diagnosis but also insufficient preventive guidance or poor advice by the GP—for example: “I wanted to do a hair transplant […] and then suddenly the doctor mentioned that I have type 2 diabetes” (P6), and “My GP never told me to change my lifestyle habits although I was always at the upper limit […] I could have taken better care of myself earlier” (P2). Also, 7 of 12 (58%) interviewees said that after their diagnosis they received insufficient and nonpersonalized information, which limited individuals with T2D's ability to manage their condition. The individuals with T2D described receiving outdated or overly generic advice from GPs that might have been used in diabetes education 20 years prior. Dietary guidance was also seen as insufficiently personalized: “I also have grain intolerance and a dairy intolerance […] I really noticed that the nutritionist was desperate. She couldn’t apply any standard approach” (P5). New technologies, such as continuous glucose monitors, were rarely mentioned or explained by GPs. Limited clinic availability presented additional barriers. Three of 12 (25%) individuals with T2D reported long waiting times and inflexible hours, making routine consultations hard to access. P8 stated: “I have

XSL•FO
RenderX

to go on an empty stomach […] I always have to take vacation days for the GP visit […] that's why I'm going only 1 - 2x per year" (P4).

Structural and Technological Barriers

Four of 12 (~33%) individuals with T2D reported that frequent blood sugar measurements helped them become more disciplined about their dietary choices. They mentioned, however, that financial constraints hindered them from ideal self-management, especially due to the cost of test strips: "One test strip costs almost 1 franc. If you want to test regularly, it gets really expensive" (P2). Three of 12 (25%) individuals with T2D said that they faced access barriers to athletic facilities. One participant lost access to Aqua Gym when it closed during the COVID-19 pandemic, while others found traveling to fitness centers too difficult and time-consuming: "The gym is too far […] so I rather stay at home and go for walks sometimes" (P12). Seven of 12 (~58%) interviewees used some digital device (diabetes app, smartwatch, step tracker, continuous glucose monitor, blood pressure device, and other) to measure their vitals frequently. Four of 12 (~33%) individuals with T2D reported that sharing their vitals with GPs was not possible due to data-sharing issues. They manually consolidate and print data for consultations: "So I print the blood pressure data from the app and bring it to the doctor. It's just extra effort that is tedious" (P1). Only one individual with T2D said that their endocrinologist has direct access to their continuous glucose monitoring device data.

Barriers to Adopting T2D Quality Assessment Tools Such as the SSED Score

This section explores the key barriers that hinder the adoption of T2D quality assessment tools, such as the SSED score (RO2).

Monetary and Nonmonetary Rewards

The current system compensates for consultations, but not for administrative tasks, such as those related to the SSED score. This contributes to a structural issue, whereby physicians are compensated based on the quantity of consultations rather than the quality of care provided: "The processes can be made more efficient, and I also believe they can be evaluated and used as a quality indicator—or even incentivized" (PT1). Four of 5 (80%) insurers acknowledged that they have limited ability to incentivize practices beyond standard reimbursable services, such as administrative documentation or coordination.

In contrast, 4 of 12 health care professionals (~33%) reported using the SSED score in their practice as part of managed care contracts with health insurers. In these arrangements, health insurers give financial incentives to the managed care network, which then decides how to allocate those funds. As a result, individual providers have limited voice in deciding how these financial resources are used. One GP mentioned that their network uses this compensation to organize workshops and invest in further education for both the GP and the team to improve the quality of T2D management: "Through our managed care contracts we have negotiated with the health insurers that if we offer cost-effective, good quality, our network can use some of the money for training and projects" (GP1).

However, insurers noted that the reach of these incentive programs remains low due to variability and fragmentation across insurance plans: "Reaching a large population is much more challenging because [A] these products vary across insurers, and [B] the number of people covered by these specific products is usually quite low" (HI3). Additionally, 1 of 5 (20%) health insurers pointed out that some medical practices may exclude noncompliant individuals with T2D to improve their overall SSED score and qualify for financial incentives. This selective reporting can distort the actual quality of care: "Whenever we gave someone a score like that, most of them always achieved it. They just didn't include the people who didn't fit in" (HI3).

Finally, 5 of 12 (~42%) health care professionals noted that nonmonetary rewards, such as performance ratings, awards, or labels, are not as interesting and, in some cases, unwanted. One GP mentioned that practices are already at full capacity, making such labels irrelevant for attracting more individuals with T2D: "I don't want all diabetics to come to me because I'm somehow labeled as such. That really doesn't add any value from my point of view...I'm not an endocrinology practice." (GP7), while GP8 noted: "At the moment, we have almost 20 patient rejections per day in our practice. People are simply glad if they can find care anywhere, and a label like this brings no added value for me at all."

Fragmentation and Interoperability

Health care professionals consistently emphasized their reliance on T2D guidelines but also expressed frustration with their fragmentation. These guidelines originate from various organizations, often resulting in inconsistent benchmarks and uncertainty about best practices.

All 12 (100%) health care professionals also reported using a clinical information system (CIS) to manage individuals with T2D, not only for diabetes but across their practice. However, they highlighted fragmentation in digital infrastructure and a lack of system interoperability as significant challenges.

Five different CISs were mentioned, some of which include diabetes-related functions, such as targeted monitoring, while others serve only basic documentation purposes. While 7 of 10 (70%) health care software developers offer diabetes dashboards or SSED-based tools, these are rarely integrated into practice software systems. Instead, they are accessed via separate web portals or stand-alone apps that require GPs to log in outside their primary clinical workflow. This was highlighted in the following way: "The practice software providers are not necessarily known for offering interfaces, or if they do, they are very expensive. […] The market is currently very fragmented, with around 60 different primary CIS" (SP1). Due to missing or inadequate interfaces, many providers reported needing to manually re-enter health data that was already stored in the CIS. Seven of 10 (70%) GPs stated that this not only wastes their time but also discourages them from using digital tools: "I also think that the entire bureaucracy in medicine is too much and if you then have to fill out something more, it is simply an additional burden" (SP9).

XSL•FO
RenderX

A key concern of health insurers is the inadequate quality of data entry in CISs: “The quality of data capture is often inadequate, and data is also lost when transferred to the SSED score” (HI4). Three of 5 (60%) insurers emphasized that unstructured data within CISs poses a significant obstacle to developing reliable interfaces. Compounding this issue is the predominance of unstructured and text-based data in many CISs, as noted by 4 of 10 (40%) health care software developers: “Often certain information is extremely difficult to read and is then either written as free text or not recorded at all” (SP5).

Tension Between T2D Quality Assessment Tools and Real-World Practice

Five of 12 (~42%) health care professionals argued that quality assessment scores, such as the SSED score, do not bring any value, as they do not necessarily reflect the quality of care provided: “I don’t think that the SSED score per se helps to improve processes. The SSED score simply tells you how well I look after my team. It doesn’t help with the processes” (GP4). The potential risk of false-positive outcomes when working with the SSED score was also highlighted: “So I can cause a falsely good SSED score if I simply tick a box but only quickly look at the foot and have not examined it properly, for example. The examinations, in particular, are parameters that allow for some margin and do not necessarily reflect the true quality” (GP9). Such false-positive results may arise when certain parameters, such as the foot examination, are marked as completed despite being carried out superficially or interpreted with significant flexibility. Thus, 2 of 12 (~17%) health care professionals emphasized that it is possible to improve the SSED score by selectively entering data. This can then lead to an inaccurate portrayal of treatment quality.

Furthermore, health care professionals emphasized that achieving a good SSED score in everyday practice is often difficult. Four of them (~33%) pointed out that many score components depend on patients’ adherence or factors beyond the physician’s control, leading to potential false negatives, such as elevated HbA_{1c} levels despite appropriate treatment. GP1 mentioned: “There are people who want to take a great deal of responsibility themselves and are very well informed; they come prepared with questions. Others, however, are already overwhelmed from the very beginning. A classic example, just as an anecdote: people know that diabetes is a “sugar disease,” but I would say that about one third of patients never fully understand that it is not only about sugary foods—100 grams of dry spaghetti are essentially equivalent to 99 grams of pure sugar.” Six of 12 (50%) health care professionals highlighted that inconsistent self-management is a key barrier, as effective T2D care requires sustained lifestyle changes, which some individuals with T2D struggle to maintain. Health insurers echoed this concern. They stressed that many SSED score components rely directly on patients’ behavior, which is influenced by broader social and economic conditions: “We also know that, especially with these chronic illnesses, socioeconomic factors come into play that the family doctor cannot change per se. A lot of it is about financial possibilities, but also level of education or the cultural environment in which the patient lives. That definitely has an influence” (HI1). Insurers also noted that poor self-management contributes not only to worse outcomes but also to increased health care costs, highlighting the need for a more interdisciplinary and socially informed approach to T2D care.

All 5 (100%) health insurers stressed that a lack of openness and consensus among health care professionals hampers effective management of T2D. Many GPs resist transparency when it comes to benchmarking performance: “Transparency can be uncomfortable for doctors because they are not used to seeing their own results in a benchmark” (HI5). Four of 5 (80%) health insurers also mentioned that GPs often assume they are already managing their individuals with T2D well, making it difficult for them to accept that their care might fall short in certain areas.

Eight of 12 (~67%) health care professionals noted that individuals with T2D rarely have diabetes alone and often have multiple chronic conditions: “Many individuals with T2D have additional diseases. This can make the treatment difficult” (GP6). These individuals with T2D often require highly individualized and interdisciplinary care approaches. Multimorbid cases challenge the standardization of treatment plans and demand greater flexibility in clinical decision-making. In this context, 4 of 12 (~33%) health care professionals argued that digital solutions can be difficult to integrate, as they often lack the necessary usability to handle such complex cases. In this regard, all 5 (100%) health insurers highlighted that long-term conditions need an interdisciplinary approach involving GPs, specialists, MPAs, and others.

Also, 6 of 10 (60%) health care software developers emphasized that diabetes management requires input from multiple disciplines and diverse reasoning approaches, making the integration of a standardized digital solution complex: “Multifactorial means that many aspects have to be taken into account when it comes to the therapy or treatment of diabetes” (SP3). This complexity underscores the challenge of developing a comprehensive digital solution that integrates relevant medical, lifestyle, and treatment data.

Regulatory Barriers

Six of 10 (60%) health care software developers identified strict Swiss regulations with health data, particularly those related to billing codes and health data documentation, as key obstacles: “The documentation requirements are strict, and there are many requirements regarding retention obligations and data protection” (SP3). Finally, 4 of 10 (40%) health care software developers highlighted the significant burden imposed when an application is classified as a medical device. Such a classification entails extensive certification processes, high costs, and stringent regulatory requirements that must be continuously met, reevaluated, and updated: “We did the math and came up with a multi-million-dollar figure, which is not feasible for us” (SP9).

Suggestion to Adopt T2D Quality Assessment Tools

The following section summarizes suggestions for improving the SSED score to better support GPs in T2D management. It highlights essential features the tool should include and outlines

XSL•FO
RenderX

practical strategies to encourage its adoption in routine GP practice (RO3).

Essential Features

We asked the participants to identify the key essential features for an application necessary for successful management of T2D, specifically related to the SSED score. The results focus on reducing structural burdens, improving T2D management, and aligning with GP needs.

Reducing Structural Burdens

A key requirement highlighted by 5 of 12 (~42%) health care professionals, 5 (50%) health care software developers, and 5 (100%) health insurers is that the SSED score application must be integrable or compatible with existing CISs. As a health care software developer put it, “A connection to the CIS would be ideal” (SP9), and a GP emphasized, “The most important thing is that such features are fully available within the practice information system” (GP4). Further, GP2 reported: “There would certainly be a benefit when the scores were somehow compatible with standard practice software and could automatically extract the necessary data.” Furthermore, 6 of 12 (50%) health care professionals stressed the importance of directly incorporating patient-reported data into the application, rather than relying on individuals with T2D to hand over PDF documents during consultations: “It’s crucial to have a fully integrated solution or embed these data-gathering steps into the consultation itself so they can be reviewed in real time” (GP4). Four of 12 (33%) individuals with T2D shared this view. They would appreciate a unified digital platform to track all health data, such as glucose, blood pressure, medication, and so on and to sync it with their GP records. Visualizing long-term trends was seen as a valuable tool for managing their condition. One individual with T2D shared: “I’m setting up the electronic health record (EHR). Ideally, I could submit all my data to a single platform and get an overview” (P6). However, 1 of 10 (10%) health care software developers cautioned that true interoperability remains a major barrier, emphasizing that without a functioning EHR infrastructure in Switzerland, full integration will not be possible.

Targeted Digital Interventions

All 10 (100%) health care software developers emphasized the need for a robust labeling and flagging system that categorizes individuals with T2D based on key metrics, highlighting critical challenges in managing T2D. This approach allows care teams to quickly identify individuals with T2D who are struggling and prioritize interventions: “It’s valuable to compare individuals with T2D. Who’s doing well and, more importantly, who isn’t? Why? Are we missing something, or are their values genuinely poor?” (SP9) Two of 5 (40%) health insurers echoed this requirement. In addition to labeling and flagging, 8 of 10 (80%) health care software developers highlighted the importance of a reminder or alert mechanism that automatically notifies health care professionals or individuals with T2D when critical thresholds are reached. Such reminders ensure that high-risk individuals with T2D receive timely follow-ups, whether consultations, laboratory tests, or medication reviews, so that no essential checkups are overlooked.

Three of 5 (60%) health insurer representatives highlighted the need to develop multidisease scores that can be accessed through digital solutions. One of 5 (20%) health insurer representatives suggested creating a solution that extends beyond the SSED score to include other chronic disease scores, such as the CARE score [22] (a cardiovascular prevention score). Five of 10 (50%) health care software developer interviewees emphasized that applications should not only support individual scores but also focus on population health: “For us, diabetes is just the beginning. In the long term, the system can and should evolve—such as for multimorbidity indicators or other relevant measurements” (SP3). SP2 mentioned: “Once we reach this level (chronic disease management), we would essentially have conditions comparable to those in modern, well-digitalized health care systems—often cited examples include Denmark, Finland, and Sweden. In such systems, these prerequisites already exist. Under certain circumstances, this could represent a real breakthrough.”

Finally, 3 (30%) health care software developers suggested including individuals with prediabetes in digital diabetes management applications: “This raises the question of whether a minimal version of the SSED guidelines could be developed specifically for prediabetics” (SP2). Including individuals with prediabetes would help identify and manage risk factors early, potentially preventing or delaying the onset of T2D.

Aligning With Needs

Two of 5 (40%) health insurers emphasized that setting goals collaboratively with individuals with T2D could improve adherence. Similarly, 2 of 10 (20%) health care software developers suggested that GPs should be able to define specific, personalized goals with individuals with T2D to better identify issues and set realistic, achievable targets. This approach also provides individuals with T2D with individualized benchmarks to work toward daily, promoting awareness and increasing adherence. In this context, 2 of 12 (16%) individuals with T2D highlighted the need for a digital solution that actively supports them through tailored feedback and education. A digital SSED application could therefore link individual score results to personalized, evidence-based educational content (eg, short explanations or practical recommendations) that explain the relevance of indicators such as HbA_{1c}, blood pressure, or weight in an understandable way. This need was explained in the following way: “Like, if you look at the UK, the NHS has a separate department for diabetes, where they really put out a lot of information very frequently and they’re updating their knowledge. (..) For example, you can also look up the coaching platform that I use: Mastering-Diabetes.org, nutritionfacts.org. These are very helpful resources that don’t exist in Switzerland” (P8).

Four of 12 (33%) health care professionals argued that the SSED score requires a more flexible interpretation, as its current rigidity can hinder practical implementation: “Because it is impossible to achieve this 80% rate, we adopted the SSED score criteria, but we completely changed the interpretation to be able to achieve reasonable care” (GP6). Such adaptations reflect the need for scoring systems to accommodate everyday practice realities rather than impose idealized standards. Furthermore,

XSL•FO
RenderX

1 of 10 (10%) health care software developers emphasized that an SSED score–based application should offer customizable scoring features tailored to specific practice workflows and populations: “The way doctors handle management varies greatly from practice to practice. I’m talking more about a workflow here—some would like it this way, others would like it that way” (SP10). This perspective was echoed by 1 of 12 (8%) health care professionals, reinforcing the need for flexibility and configurability in digital tools to support meaningful, context-sensitive care delivery.

Guideline Adoption Strategies

In addition to exploring key problems in T2D management and essential features for an application, we asked the interviewees which strategies they consider essential for improving the adoption of the SSED score and enhancing the overall management of T2D.

Task Shifting

Eleven of 12 (~92%) health care professionals emphasized the important role of MPAs, MPKs, and advanced practice nurses in diabetes management, to whom physicians can delegate specific tasks and consultations: “They help us a lot. It’s increasingly being divided so that we, doctors, handle part of the four examinations, and the MPA does the other part—two and two, for example” (GP10). Additionally, 3 (30%, 6/10) health insurer representatives and 6 of 10 (60%) health care software developers agreed that involving these health care professionals could improve diabetes care by allowing GPs to delegate responsibilities, including administrative tasks.

Aligning Financial Incentives With Quality Care

Ten of 12 (83%) health care professionals emphasized that financial incentives could serve as a strong motivator. Five of 12 (42%) health care professionals highlighted the following: “I think the effort should be rewarded—the time you put in. And of course, if you put in more effort, you should get something back for the time you invest” (GP3). One of 12 (8%) health care professionals also suggested that such compensation could be used to fund continuing education, workshops, and training initiatives that directly contribute to improved diabetes care. This perspective was echoed by 2 of 5 (40%) health insurers, who placed hope in the upcoming TARDOC (Tarif für ambulante ärztliche Leistungen und Dokumentation) pricing structure, set to replace the outdated TARMED (tarif médical) system in 2026. TARDOC is designed to modernize the compensation model, aligning it more closely with today’s medical practices and technologies. It aims to ensure fairer remuneration for services rendered while reducing inefficient incentives and promoting value-based care within the health care system: “I think there’s hope that with TARDOC, we’ll have a model that’s a bit more flexible and innovative. That we can also ensure that certain services can be financed, which are important but are currently difficult to include in existing models” (HI3).

Boosting Communication and Data Sharing

A recurring theme among interviewees was the need for improved communication and data sharing within diabetes care networks. Nine of 12 (75%) health care professionals believed that better communication and collaboration between GPs and specialists could significantly improve the management of complex, multimorbid individuals with T2D. GP8 mentioned: “What would be the most effective and best tool would be a reporting and information exchange tool with hospitals—specifically for sharing clinical reports. Inselspital has implemented this, and Epic has now set it up for GPs: a login access that allows GPs to log into Epic and view reports, progress notes, and the full medical records of their own patients. That is, of course, excellent. If other provider groups were to implement something similar, it would significantly reduce uncertainties about what has happened or what is currently going on, because I could simply retrieve the information myself. This is, for example, a tool that I currently only know from Epic, and one that could be optimized and scaled substantially.”

One of 5 (20%) insurers shared a similar view, emphasizing the value of smooth, timely transitions to specialized care: “There’s long-term cost-saving potential, especially if GPs are supported in identifying which individuals with T2D should be referred and when” (HI1). Five of 10 (50%) health care software developers also stressed the importance of network-wide data sharing and comparison: “We have a benchmark in the network, so the doctor can see how they compare to other practices. It acts as a trust indicator” (SP4). These comparisons typically involve quality indicators and clinical outcomes (eg, HbA_{1c} levels and blood pressure control), allowing health care professionals to recognize areas for improvement and drive better care outcomes. While cost data were less frequently mentioned, the emphasis was on performance transparency across practices.

Four of 5 (80%) health insurers echoed this by recommending tools such as feedback loops and “improvement clubs” to promote transparency and continuous learning: “Once we create transparency and see how they compare, it opens up opportunities for ongoing improvement. It’s about using data, learning from it, making changes, and moving through a continuous learning cycle—Plan, Do, Check, Act” (HI1). This approach not only improves individual practices but also fosters a culture of shared responsibility and peer-driven quality enhancement across the network.

Leveraging Digital Tools to Support Clinical Decision-Making

Besides this, seven health care software developers (70%, 7/10) emphasized that automated solutions could support clinical decision-making through generating valuable insights to assist GPs. These technologies also hold promise for reducing administrative workload over time: “I think AI will have a revolutionary impact in both administration and medicine. It will take a lot of work off doctors’ hands and can create treatments tailored to the patient” (SP10). Additionally, five health care professionals (~42%, 5/12) acknowledged the potential of decision-supporting solutions to enhance and streamline medical decision processes.

XSL•FO
RenderX

Strengthening Support for Individuals With T2D

Several interviewees mentioned that the SSED score should be designed not only for GPs but also for other stakeholders. Three of 12 (25%) individuals with T2D called for closer collaboration between GPs and pharmacies. This could be implemented, for instance, through testing HbA_{1c} at pharmacies during off-peak hours and discussing results with GPs over the phone. This collaboration would improve access and flexibility in managing T2D: "It would be helpful if I could test my blood at the pharmacy and they send the results to my GP so we could discuss them over the phone" (P4). In this case, not only the GP but also pharmacists would have access to the SSED score–based application. This would alleviate the burden on GPs to fill out the SSED score themselves. On top of that, 9 of 12 (75%) individuals with T2D highlighted that if a SSED score–based application had a user interface, it should show updated and personalized information. Some proposed dedicated platforms or paper-based periodical publications with up-to-date content tailored to individual needs: "In the UK, the NHS puts out frequent updates and holds conferences" (P9). Five of 12 (~42%) health care professionals echoed this sentiment and noted that individuals with T2D are more likely to follow treatment plans when they understand their condition better: "If I know that individuals with T2D see the added value and the improvement, they are, of course, a lot more motivated" (GP8). Tailored information can thus play a key role in strengthening adherence and long-term management of T2D.

Discussion

Principal Findings

Overview

This study explored stakeholder perspectives (health care professionals, health care software developers, individuals with T2D, and health insurers) on challenges in T2D management in Swiss primary care, the limitations of the existing SSED quality score, and strategies to support its adoption in routine practice. Findings indicate persistent structural barriers, including time pressure for health care professionals, fragmented care pathways, and limited personalized support for individuals with T2D. Stakeholders also identified obstacles to digital quality monitoring, particularly poor integration with GP systems, misaligned incentives, and limited suitability for multimorbidity. At the same time, facilitators of adoption included workflow integration, task shifting, patient-centered functionality, and improved data sharing. Overall, the results suggest that successful implementation of a digital SSED score requires addressing organizational, technological, and behavioral factors across the care system. Although the SSED score is specific to the Swiss T2D context, the underlying principles may also be relevant for other chronic conditions and health care systems.

One of the most critical and unresolved challenges in Switzerland is the fragmented and unstructured nature of health care data. This challenge poses a fundamental barrier to the implementation of interoperable systems. Interviewees across stakeholder groups, particularly health care software developers and health insurers, consistently pointed to the lack of standardized interfaces and structured data as a core impediment. CISs are highly fragmented, often relying on free-text entries, manual exports, and redundant documentation processes. Even the most well-designed digital SSED score cannot unlock its potential if data from various sources, such as individuals with T2D, pharmacies, devices, or laboratories, cannot be aggregated and meaningfully connected.

This lack of integration is not a minor technical issue but one of the defining barriers. It prevents automation, limits decision-support capabilities, and undermines the utility of performance feedback. Therefore, the challenge of bringing together disparate, siloed, and often poorly structured data is a technical prerequisite and represents the most urgent priority for digital transformation in long-term care. Without addressing this foundational issue, the envisioned digital support systems risk remaining conceptual rather than operational.

Essential Features

Findings from the interviews show that for the SSED score to be adopted at scale, it must be embedded into the daily workflows of primary care in a way that minimizes administrative burden while maximizing clinical utility. A core requirement is interoperability. Data entered by individuals with T2D or MPAs must seamlessly synchronize with the GPs' EHRs, such as the CIS. This ensures that information is not siloed and that GPs maintain a complete, up-to-date overview without duplicating effort.

These findings are supported by existing studies. For example, it has been shown that individuals with T2D treated in hospitals with advanced EHRs cost nearly 10% less than those in hospitals without advanced EHRs, due to better care coordination, fewer duplicative tests, and increased efficiency [23].

Furthermore, a recent literature review showed that EHRs with predictive analytics can anticipate disease progression, support earlier intervention, and improve patient outcomes, particularly in conditions such as diabetes and cardiovascular disease, while also reducing hospitalizations and health care use [24].

Furthermore, our findings show that the user interface of the SSED score must also be intuitive and time-efficient. Quick data entry fields, automated alerts for missing or abnormal values, and visual dashboards summarizing an individual's status (eg, traffic-light systems) can support clinical decision-making while saving time. To reconcile standardized quality assessment with individualized care, the authors suggest that the digital SSED score should be designed as a modular system. This is particularly important for individuals with multimorbidity.

A modular structure could enhance usability. In such a design, individual sections of the score (eg, HbA_{1c}, blood pressure, and weight) can be completed independently and at different times. This allows clinicians to gradually fill in the score during multiple visits while the system tracks overall completion status.

The authors suggest that such a modular architecture would enable the integration of additional disease-specific or cross-disease quality metrics, such as multidisease scores,

XSL•FO
RenderX

without overburdening clinicians. From the authors' perspective, this approach would allow a digital SSED tool to preserve a standardized core set of diabetes indicators while flexibly expanding to capture comorbid cardiometabolic conditions, thereby supporting holistic, patient-centered care in multimorbidity contexts.

Based on the study findings, the authors further recommend empowering individuals with T2D as active contributors by integrating home-monitoring devices (eg, blood pressure monitors and wearables) and allowing manual input of parameters such as weight, blood pressure, or step count. Incentive mechanisms and nudges, such as accomplishment badges and reminders, could further increase engagement.

Prior research suggests that patient-facing applications that support self-monitoring and provide feedback can improve engagement and health outcomes, particularly in chronic disease management [25]. Features such as reminders, goal tracking, and gamification elements have been found to enhance motivation and adherence to self-care behaviors [26,27].

Adoption Strategies

Overview

While financial incentives such as pay-for-performance schemes have been proposed to enhance chronic disease management [28], our interviews reveal mixed perceptions of these schemes. Although GPs indicated that receiving additional compensation for completing the SSED score would be appreciated, they also noted that merely filling out the score may not necessarily lead to improved clinical outcomes. Moreover, if payment were tied to clinical outcomes, health insurers raised concerns that GPs may selectively document only well-managed individuals with T2D in the SSED score to protect their performance metrics. While financial incentives could be powerful motivators for GPs, they must be carefully designed to avoid discouraging the inclusion of complex or noncompliant individuals with T2D.

Based on additional literature, Helsana, one of Switzerland's largest health insurers, successfully piloted a pay-for-performance model that resulted in a 14% increase in guideline-adherent diabetes care; however, critics argue that it remains unclear whether individuals' clinical outcomes improved [29].

Our findings further showed that certifications were also seen as an unappealing incentive for GPs to use the SSED score, as many are already managing full practices and are not seeking to attract additional individuals with T2D. Instead, time-saving approaches and task redistribution emerged as more promising strategies to foster the implementation of the SSED score. Three potential solutions were identified as feasible and contextually relevant.

Task Shifting to MPAs With Specific Diabetes Education

Our findings indicate that MPAs are already engaged in routine clinical tasks such as blood sampling and education of individuals with T2D. Expanding their scope of practice to include SSED score completion, especially if supported by additional training in diabetes care, could reduce GPs' workload while enhancing the quality of interactions with individuals with T2D. MPAs are not only more cost-effective than physicians but also have the potential to engage more deeply with individuals with T2D, thereby supporting better adherence and health outcomes.

Evidence from a Swiss case report further illustrates that when individuals with T2D are given space to openly discuss their medication concerns through active listening, they can move from nonadherence to sustained adherence, highlighting the importance of communication skills alongside clinical delegation [30]. Furthermore, a recent meta-analysis shows that, compared with standard care, models in which nurses take a leading role achieve meaningful improvements in clinical indicators of diabetes and hypertension [31].

A Swiss study evaluating the long-term implementation of a care program, primarily delivered through specially trained practice nurses, showed significant improvements in HbA_{1c}. However, only 40.9% of practices continued using this model 3 years after the initial intervention. Key barriers included financial constraints, such as the lack of reimbursement for practice nurses, and a reluctance among GPs to delegate clinical responsibilities [32].

One health insurer (HI1) noted that TARDOC, expected to be implemented in 2026, introduces only incremental changes for nonphysician chronic care activities. Similar to TARMED, it includes chronic care management positions (AK.05) that allow services delivered by MPAs or MPKs to be billed, while shifting from 15-minute blocks to minute-based accounting and expanding eligibility to additional chronic conditions. Consequently, the economic viability of delegating structured diabetes management to MPAs or MPKs depends largely on practice-level efficiency gains, such as freeing physician capacity for higher-value services, rather than on tariff reform alone.

Involving Pharmacies in Routine Monitoring

Although we did not interview pharmacists, interviewees in our study suggested that pharmacies could be a helpful resource in chronic disease management.

It is plausible that shifting routine diabetes monitoring—such as HbA_{1c} or cholesterol checks—from GPs to pharmacies could offer economic and logistical advantages. A pharmacy-based HbA_{1c} test costs around CHF 30 (US $38.28) [33]. Adding a teleconsultation with a GP (CHF 50 [US $63.79]) [34] could bring the total to CHF 80 (US $102.07)—less than half the cost of a standard GP visit (CHF 150 [US $191.38]) [35]. For well-controlled individuals with T2D, 2 annual pharmacy visits combined with 1 in-person GP check (eg, for nephropathy, retinopathy, and foot exams) could maintain care quality while substantially lowering costs. Findings from the Swiss Siscare program, a pharmacy-led adherence program, further support this potential: pharmacists effectively conducted medication adherence monitoring and counseling. However, collaboration with physicians was often limited to one-way information flows, underscoring the importance of interoperable IT systems and clear reimbursement models for sustainable implementation [36].

XSL•FO
RenderX

From the authors' perspective, this development could be transformative for the SSED score: pharmacies could act as local hubs for routine monitoring and structured data entry, thus alleviating GP workload and enhancing continuity of care. The model also aligns with broader goals of cost efficiency and prevention, as better medication adherence and early interventions have been shown to reduce complications [37,38]. However, as previous studies show, successful implementation will depend on the integration of interoperable systems, well-defined roles, and close coordination among health care professionals [39,40].

Empowering Individuals With T2D Through Self-Tracking and Data Sharing

Based on the interviews, some individuals with T2D already monitor health parameters using wearables and expressed a willingness to share these data with their GPs.

Building on this insight, the authors propose that individuals with T2D could additionally complete laboratory-based metrics (eg, HbA_{1c} and cholesterol) in pharmacies, thereby creating a comprehensive, longitudinal health profile accessible to GPs. Such a model might support more personalized care and greater engagement and accountability. It also allows health insurers to experiment with value-sharing schemes, such as rewarding data-sharing individuals with T2D with bonus programs or lower premiums. In Switzerland, such models are common; health insurers have introduced apps that promote healthy lifestyles by offering monetary rewards for sharing behavioral data from fitness trackers or manual inputs [41]. From the authors' perspective, if implemented effectively, this triadic model, combining individuals' self-tracking, pharmacy-based services, and GP oversight in one digital solution, could offer a scalable, cost-saving solution that maintains high standards of care.

These findings suggest that meaningful SSED score adoption will require more than a single incentive mechanism. A multipronged approach, grounded in interprofessional collaboration, digital enablement, and empowerment, holds greater promise. Importantly, the upcoming regulatory changes in Switzerland provide fertile ground for the trial of such models. Yet, the success of these strategies will depend on several enabling factors: interoperability of digital systems, provider trust in redistributed responsibilities, and robust evidence demonstrating the cost-effectiveness and clinical utility of these alternative pathways.

Practical Recommendations and Future Directions

Across all stakeholder groups, the findings consistently indicate that resolving data infrastructure fragmentation is the primary prerequisite for any scalable digital enhancement of T2D quality measurement. Without interoperable data flows and shared data standards, advanced digital features cannot function reliably or deliver value at scale. Taken together, these findings suggest several core design principles for future digital health tools: (1) interoperability should be treated as a foundational requirement rather than an optional feature; (2) systems must integrate seamlessly into existing clinical workflows; (3) digital solutions should function as shared coordination infrastructures that enable multistakeholder collaboration rather than as isolated documentation tools; and (4) digital tools should support task delegation, minimize additional workload, and enable bidirectional data exchange across care settings to ensure practical usability and sustained adoption. Based on these findings, a phased implementation strategy is recommended.

In the short term (0 - 12 months), the SSED score should be embedded into existing primary care workflows by enabling MPAs to complete routine score components, thereby reducing GP workload. Crucially, this embedding should prioritize interoperability and data exchange over feature expansion. To mitigate fragmentation, the score should function as a shared digital coordination layer, enabling interoperable, bidirectional data exchange between GP systems, pharmacies, and patient-facing applications rather than operating as a standalone documentation tool.

In the medium term (12 - 36 months), pharmacies could be integrated as decentralized hubs for standardized monitoring (eg, HbA_{1c} and cholesterol), a development supported by recent policy changes. In March 2025, the Swiss Parliament adopted a revision of the Health Insurance Act (Krankenversicherungsgesetz) enabling pharmacists to provide reimbursable services under mandatory health insurance from 2027 onward, including therapy adherence support, blood pressure screening, and medication optimization [42]. This policy shift aligns with evidence showing strong patient interest in pharmacy-based counseling, medication reviews, and treatment planning—services closely linked to quality indicators captured by the SSED score [36]. In parallel, individuals with T2D should be empowered as active contributors through self-tracking and tailored educational content linked to score results.

As a longer-term goal (>36 months), the SSED score should be expanded to reflect multimorbidity by incorporating additional cardiometabolic conditions commonly co-occurring with T2D, thereby increasing clinical relevance and sustainability in primary care. From a regulatory perspective, such a digital score would only fall under the Swiss Medical Device Regulation as a medical device if it provides automated decision support or treatment recommendations, whereas a descriptive, coordination tool would likely remain lower risk [43]. Future research should therefore evaluate clinical effectiveness, cost-effectiveness, professional adoption, and governance requirements as the score evolves toward multidisease and decision-support functionality.

Beyond provider-level changes, health insurers and public authorities are key enablers of adoption. Insurers can promote uptake by aligning reimbursement and incentives with coordinated care, for example, by rewarding SSED score use and supporting pharmacy-based monitoring pathways. Public authorities can facilitate scale-up by strengthening interoperability requirements. Together, these measures can help translate pilot models into routine practice.

More broadly, reimbursement structures will be critical for the sustainable implementation of digital quality monitoring. Payment models that primarily reward individual encounters may not adequately support data sharing, interprofessional

XSL•FO
RenderX

coordination, and task delegation. Aligning incentives with coordinated, data-enabled care processes may therefore be essential for scaling digital quality monitoring.

Although this study focuses on Swiss primary care, the identified barriers—particularly fragmented data infrastructures, limited interoperability, and workflow misalignment—are common across many health care systems undergoing digital transformation [44]. The design principles derived from this study, such as prioritizing interoperable data exchange, integrating tools into existing clinical workflows, and supporting task redistribution, are therefore likely to be broadly transferable. However, the pace and feasibility of scale-up will depend on local conditions, including digital infrastructure maturity, regulatory frameworks, and reimbursement models [45,46]. Overall, while core design principles may generalize across settings, implementation strategies likely require adaptation to system-specific organizational and governance contexts.

Theoretical Implications

Although this study followed an inductive qualitative approach, the findings can be interpreted through the lens of the technology acceptance model. The inductively derived themes map closely onto core technology acceptance model constructs, particularly perceived usefulness and perceived ease of use [15], while also extending them in ways that are highly relevant for digital health and chronic disease management. Perceived usefulness was shaped less by standalone technological features and more by workflow integration, interoperability with existing clinical systems, and the redistribution of tasks to MPAs or pharmacies. Likewise, perceived ease of use was closely linked to system-level integration and time savings within routine primary care workflows. Importantly, the findings indicate that technology acceptance in primary care is not solely an individual decision but a collective, ecosystem-level process influenced by interprofessional coordination, incentive structures, and patient engagement mechanisms. From a theoretical perspective, this suggests a need to extend the technology acceptance model in health care contexts to better account for organizational roles, reimbursement arrangements, and shared value creation across multiple stakeholders, thereby situating digital quality measurement tools within broader care models rather than treating them as isolated technologies.

Limitations

This study has several limitations. First, although the sample included diverse professional backgrounds—such as health care professionals, individuals with T2D, health insurers, and health care software developers—the participants were not representatively sampled across Switzerland. Most interviewees came from the German-speaking region. This regional concentration limits the generalizability of findings, particularly related to system interoperability, IT integration, and data exchange practices, which may vary across regions due to differences in cantonal implementation and digital maturity in the region. Second, we acknowledge the possibility of participant self-selection, as those with a particular interest in digital health or structured diabetes care might have been more inclined to participate, thus introducing bias in the sample. Third, the study explored mostly hypothetical SSED score adoption scenarios. As such, many of the insights gathered are based on anticipated rather than observed behavior. This limits our ability to assess real-world effectiveness, feasibility, or unintended consequences. While participants reflected thoughtfully on enablers and barriers, their responses may not fully predict actual adoption behavior, especially under changing workload or policy conditions. While this study concentrated on identifying key challenges and generating initial solution ideas, future research should perform a thorough evaluation of how these essential features and adoption strategies could be designed and implemented.

Fourth, this study did not include community pharmacists as interview participants. While pharmacy-based follow-up emerged as a potentially important component of future diabetes care models in the ”Results” and “Discussion” sections, these suggestions originated from other stakeholder groups, including individuals with T2D, GPs, insurers, and digital health developers. As such, the findings reflect perceptions about the role of pharmacies rather than insights from pharmacists themselves. The absence of community pharmacists limits the extent to which conclusions can be drawn about the feasibility, acceptability, and operationalization of pharmacy-led follow-up, highlighting the need for future research to directly incorporate pharmacists’ perspectives.

Finally, economic considerations were discussed only at a conceptual level. Although stakeholders frequently raised cost concerns, particularly regarding time investment, IT integration, and the lack of direct reimbursement for MPAs without diabetes education, this study did not include quantitative modeling of the potential return on investment or long-term impact on health care usage. Further research using cost-effectiveness or budget impact analysis frameworks will be essential to substantiate the financial case for broader adoption.

Conclusion

This study highlights the importance of aligning digital health innovations related to quality monitoring tools with the practical realities of clinical care. The SSED score, in a digital form, has the potential to enhance the quality and coordination of T2D management across Swiss primary care. However, our results suggest that successful adoption of a digital SSED score may depend on addressing key barriers identified by stakeholders: the need for system interoperability, time-saving features, delegation of tasks, and improved engagement. Our findings further suggest that technological sophistication and financial incentives for GPs alone are unlikely to be sufficient to drive change. Rather, the integration of an SSED score–based application must be supported by a multifaceted adoption strategy, combining workflow-compatible design, interprofessional collaboration, and task-shifting activities. New financing models that redistribute tasks to MPAs or pharmacies, empower individuals with T2D through self-monitoring, and enable real-time data sharing to offer a compelling path forward.

While these strategies are promising from a stakeholder perspective, their feasibility and effectiveness require validation in real-world settings. Future research should, therefore, move beyond hypothetical scenarios and empirically test digital SSED score–based applications through pilot implementations and

XSL•FO
RenderX

accompanying economic evaluations to assess their impact on care processes, clinical and efficiency outcomes, and resource use.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading, editing, summarizing text, and translation. The GAI tool used was ChatGPT (version 4.5; OpenAI). Responsibility for the final paper lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the outcomes. This declaration was submitted by the author OFG.

Funding

The Centre for Digital Health Interventions (CDHI) is funded in part by CSS, a Swiss health insurer; Mavie Next (UNIQA), an Austrian health care provider; MTIP, a Swiss investor company; and Novo Nordisk. None of these funders were involved in the study design, data analysis, or manuscript preparation.

Data Availability

The data generated and analyzed during this study are not publicly available due to ethical and confidentiality restrictions related to the protection of participant privacy. The interview data contain potentially identifiable information and therefore cannot be shared in accordance with institutional policies and applicable data protection regulations.

Authors' Contributions

OFG and TK performed project administration. EF, TK, and SG provided supervision. OFG and LJA contributed to writing the original draft. EF, MB, MJ, TK, SG, MP, and SM contributed to writing, review, and editing.

Conflicts of Interest

OFG, EF, TK, and MJ are affiliated with the Centre for Digital Health Interventions (CDHI), a joint initiative of the Institute for Implementation Science in Health Care, University of Zurich; the Department of Management, Technology, and Economics at ETH Zurich; and the Institute of Technology Management and School of Medicine at the University of St. Gallen. EF is a cofounder of Pathmate Technologies, a university spin-off that develops and delivers digital clinical pathways. TK is also a cofounder of Pathmate Technologies; however, he has not held any shares since the end of 2023 and has never held a formal role within the company.

Multimedia Appendix 1
Interview guides.
[DOCX File, 43 KB - jmir_v28i1e82960_app1.docx]

Multimedia Appendix 2
Interviewee's background.
[DOCX File, 36 KB - jmir_v28i1e82960_app2.docx]

Checklist 1
COREQ checklist.
[DOCX File, 40 KB - jmir_v28i1e82960_app3.docx]

References

1. «Es ist zu viel»: Hilferufe der Hausärztinnen. Schweizer Radio und Fernsehen (SRF). 2023. URL: https://www.srf.ch/news/schweiz/hausarztpraxen-in-not-es-ist-zu-viel-hilferufe-der-hausaerztinnen [accessed 2024-11-05]
2. Deshalb gehen der Schweiz die Hausärztinnen und Hausärzte aus. Schweizer Radio und Fernsehen (SRF). 2023. URL: https://www.srf.ch/news/schweiz/allgemeinmedizin-deshalb-gehen-der-schweiz-die-hausaerztinnen-und-hausaerzte-aus [accessed 2024-11-05]
3. Hausarztpraxis ist bei Diabetes Anlaufstation Nummer eins. ÄrzteZeitung. 2020. URL: https://www.aerztezeitung.de/Politik/Hausarztpraxis-ist-bei-Diabetes-Anlaufstation-Nummer-eins-408419.html [accessed 2024-11-05]
4. Nøkleby K, Berg TJ, Mdala I, et al. Variation between general practitioners in type 2 diabetes processes of care. Prim Care Diabetes 2021 Jun;15(3):495-501. [doi: 10.1016/j.pcd.2020.11.018] [Medline: 33349599]

XSL•FO
RenderX

5. Stucki M, Schärer X, Trottmann M, Scholz-Odermatt S, Wieser S. What drives health care spending in Switzerland? Findings from a decomposition by disease, health service, sex, and age. BMC Health Serv Res 2023 Oct 25;23(1):1149. [doi: 10.1186/s12913-023-10124-3] [Medline: 37880733]
6. Huber CA, Schwenkglenks M, Rapold R, Reich O. Epidemiology and costs of diabetes mellitus in Switzerland: an analysis of health care claims data, 2006 and 2011. BMC Endocr Disord 2014 Jun 3;14(1):44. [doi: 10.1186/1472-6823-14-44] [Medline: 24894889]
7. Zhao X, Zhang Y, Yang Y, Pan J. Diabetes-related avoidable hospitalisations and its relationship with primary healthcare resourcing in China: a cross-sectional study from Sichuan province. Health Soc Care Community 2022 Jul;30(4):e1143-e1156. [doi: 10.1111/hsc.13522] [Medline: 34309097]
8. Ugolini C, Lippi Bruni M, Leucci AC, et al. Disease management in diabetes care: when involving GPs improves patient compliance and health outcomes. Health Policy 2019 Oct;123(10):955-962. [doi: 10.1016/j.healthpol.2019.08.014] [Medline: 31481267]
9. Carlander M, Wirth B, Höglinger M. Wissenschaftliche Begleitevaluation SWICA Disease-Management-Programm Diabetes Mellitus: Follow-Up-Jahr 2020. : ZHAW Zürcher Hochschule für Angewandte Wissenschaften; 2022 URL: https://digitalcollection.zhaw.ch/items/efc8b9ff-ccc8-49da-8ab9-3754ad16d704 [accessed 2026-05-02]
10. Kriterien für ein „gutes“ Disease Management Diabetes in der Grundversorgung. SGED. 2025 URL: https://www.sgedssed.ch/SgedSsed/fileadmin/6_Diabetologie/64_Ressourcen_Hausarzt/Diabetes_Kriterien_2017_SGED_def.pdf [accessed 2026-05-02]
11. Christ E, Czock A, Renström F, et al. Evaluation of type 2 diabetes care management in nine primary care practices before and after implementation of the Criteria of Good Disease Management of Diabetes established by the Swiss Society of Endocrinology and Diabetology. Swiss Med Wkly 2022 Jul 18;152:w30197. [doi: 10.4414/smw.2022.w30197] [Medline: 35925612]
12. Ley B, Rijal KR, Marfurt J, et al. Analysis of erroneous data entries in paper based and electronic data collection. BMC Res Notes 2019 Aug 22;12(1):537. [doi: 10.1186/s13104-019-4574-8] [Medline: 31439025]
13. Clifton DA, Clifton L, Sandu DM, et al. “Errors” and omissions in paper-based early warning scores: the association with changes in vital signs--a database analysis. BMJ Open 2015 Jul 3;5(7):e007376. [doi: 10.1136/bmjopen-2014-007376] [Medline: 26141302]
14. Giger OF, Pfitzer E, Mekniran W, et al. Digital health technologies and innovation patterns in diabetes ecosystems. Digit Health 2025;11:20552076241311740. [doi: 10.1177/20552076241311740] [Medline: 39911718]
15. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
16. Militello LG, Arbuckle NB, Saleem JJ, et al. Sources of variation in primary care clinical workflow: implications for the design of cognitive support. Health Informatics J 2014 Mar;20(1):35-49. [doi: 10.1177/1460458213476968] [Medline: 24105625]
17. Storseth O, McNeil K, Grudniewicz A, et al. Administrative burden in primary care: critical review. Can Fam Physician 2025 Jun;71(6):417-423. [doi: 10.46747/cfp.7106417] [Medline: 40523738]
18. Homepage. TestingTime. 2025. URL: https://www.testingtime.com/en/ [accessed 2025-04-15]
19. O’Reilly CA III, Tushman ML. Organizational ambidexterity: past, present, and future. AMP 2013 Nov;27(4):324-338. [doi: 10.5465/amp.2013.0025]
20. Guest G, Bunce A, Johnson L. How many interviews are enough? Field methods 2006 Feb;18(1):59-82. [doi: 10.1177/1525822X05279903]
21. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
22. Rosemann T, Sudano I, Lüscher TF, Strehle O, Senn O. The CARE (cardiovascular prevention) score - an instrument for recording the quality of CARE of cardiovascular risk patients in the GP practice. Praxis (Bern 1994) 2024 May;113(5):113-118. [doi: 10.23785/PRAXIS.2024.05.001] [Medline: 38864104]
23. Kazley AS, Simpson AN, Simpson KN, Teufel R. Association of electronic health records with cost savings in a national sample. Am J Manag Care 2014 Jun 1;20(6):e183-e190. [Medline: 25180501]
24. Alam MA, Sohel A, Uddin MM, et al. Big data and chronic disease management through patient monitoring and treatment with data analytics. AJSTEME 2024 Oct 27;1(1):77-94. [doi: 10.69593/ajaimldsmis.v1i01.133]
25. Hamine S, Gerth-Guyette E, Faulx D, Green BB, Ginsburg AS. Impact of mHealth chronic disease management on treatment adherence and patient outcomes: a systematic review. J Med Internet Res 2015 Feb 24;17(2):e52. [doi: 10.2196/jmir.3951] [Medline: 25803266]
26. Palmer MJ, Barnard S, Perel P, Free C. Mobile phone-based interventions for improving adherence to medication prescribed for the primary prevention of cardiovascular disease in adults. Cochrane Database Syst Rev 2018 Jun 22;6(6):CD012675. [doi: 10.1002/14651858.CD012675.pub2] [Medline: 29932455]
27. Johnson D, Deterding S, Kuhn KA, Staneva A, Stoyanov S, Hides L. Gamification for health and wellbeing: a systematic review of the literature. Internet Interv 2016 Nov;6:89-106. [doi: 10.1016/j.invent.2016.10.002] [Medline: 30135818]

XSL•FO
RenderX

28. Giger OF, Fleisch E, Jovanova M, Kowatsch T. Barriers and facilitators of implementing value-based care: the case of SwissDiabeter. Digit Health 2025;11:20552076251336322. [doi: 10.1177/20552076251336322] [Medline: 40351841]
29. Pinto C. Krankenkasse misst erstmals Qualität der Behandlung – und belohnt die Ärzte. Tages-Anzeiger. 2023 Aug 5. URL: https://www.tagesanzeiger.ch/krankenkasse-misst-erstmals-qualitaet-der-behandlung-und-belohnt-die-aerzte-147341634201 [accessed 2025-04-14]
30. Solh Dost L, Gastaldi G, Schneider MP. The power of active listening to address medication non-adherence during care transition: a case report of a polypharmacy patient with type 2 diabetes. Pharmacy (Basel) 2025 Apr 30;13(3):64. [doi: 10.3390/pharmacy13030064] [Medline: 40407502]
31. Kavita K, Thakur JS, Ghai S, Narang T, Kaur R. Nurse-led interventions for prevention and control of noncommunicable diseases in low- and middle-income countries. International Journal of Noncommunicable Diseases 2023;8(1):4-13. [doi: 10.4103/jncd.jncd_74_22]
32. Chmiel C, Giewer I, Frei A, Rosemann T. Four-year long-term follow-up of diabetes patients after implementation of the Chronic Care Model in primary care: a cross-sectional study. Swiss Med Wkly 2017;147(4344):w14522. [doi: 10.4414/smw.2017.14522] [Medline: 29120011]
33. DiabetesCheck. Medbase Apotheken. URL: https://www.medbase-apotheken.ch/angebote/detail/diabetescheck/ [accessed 2025-04-14]
34. Telemedizin. Medgate. URL: https://www.medgate.ch/de-ch/ratgeber/telemedizin#:~:text=Die%20Konsultation%20wird%20im%20Rahmen,%2D [accessed 2025-04-14]
35. Was kostet ein Arztbesuch in der Schweiz? Deine3a. URL: https://www.deine3a.ch/was-kostet-ein-arztbesuch-in-der-schweiz [accessed 2026-04-20]
36. Bawab N, Moullin J, Jotterand S, Rossier C, Schneider MP, Perraudin C. Building interprofessional collaborative practices through a support program for patients with type 2 diabetes in primary care. J Contin Educ Health Prof 2023 Apr 1;43(2):77-86. [doi: 10.1097/CEH.0000000000000466] [Medline: 36877813]
37. Uddin S, Sanchez Machado M, Alshahrouri B, et al. Empowering pharmacists in type 2 diabetes care: opportunities for prevention, counseling, and therapeutic optimization. J Clin Med 2025 May 29;14(11):3822. [doi: 10.3390/jcm14113822] [Medline: 40507585]
38. Roebuck MC, Liberman JN, Gemmill-Toyama M, Brennan TA. Medication adherence leads to lower health care use and costs despite increased drug spending. Health Aff (Millwood) 2011 Jan;30(1):91-99. [doi: 10.1377/hlthaff.2009.1087] [Medline: 21209444]
39. Jabin MSR. Operational disruption in healthcare associated with software functionality issue due to software security patching: a case report. Front Digit Health 2024;6:1367431. [doi: 10.3389/fdgth.2024.1367431] [Medline: 38550716]
40. Hung SY, Chen C, Wang KH. Critical success factors for the implementation of integrated healthcare information systems projects: an organizational fit perspective. CAIS 2014;34. [doi: 10.17705/1CAIS.03439]
41. Martani A, Shaw D, Elger BS. Stay fit or get bit - ethical issues in sharing health data with insurers' apps. Swiss Med Wkly 2019 Jun 17;149:w20089. [doi: 10.4414/smw.2019.20089] [Medline: 31256413]
42. Apotheken stärken: parlament verabschiedet wegweisende KVG-revision. Schweizerischer Apothekerverband. 2025. URL: https://pharmasuisse.org/de/medienmitteilung/apotheken-staerken-parlament-verabschiedet-wegweisende-kvg-revision [accessed 2025-04-15]
43. Guide for app developers, manufacturers and distributors. : eHealth Suisse; 2022 URL: https://www.e-health-suisse.ch/payload/api/documents/file/Guide_App_Developers.pdf [accessed 2026-05-02]
44. Bertrand Y, Denis JL, Motulsky A. A case study on the barriers of digital transformation in public hospitals. Stud Health Technol Inform 2025 May 15;327:398-399. [doi: 10.3233/SHTI250355] [Medline: 40380466]
45. Fahad M. Technological transformation in healthcare: a global perspective. J Digitain Realism Mastery (DREAM) 2024 Dec 30;3(12):73-84. [doi: 10.56982/dream.v3i12.294]
46. Williams PA, Lovelock B, Cabarrus T, Harvey M. Improving digital hospital transformation: development of an outcomes-based infrastructure maturity assessment framework. JMIR Med Inform 2019 Jan 11;7(1):e12465. [doi: 10.2196/12465] [Medline: 30632973]

Abbreviations

CIS: clinical information system
COREQ: Consolidated Criteria for Reporting Qualitative Research
EHR: electronic health record
GP: general practitioner
HbA_{1c}: glycated hemoglobin
MPA: medical practice assistant
MPK: medical practice coordinator
RO: research objective
SSED: Swiss Society of Endocrinology and Diabetology

XSL•FO
RenderX

T2D: type 2 diabetes
TARDOC: Tarif für ambulante ärztliche Leistungen und Dokumentation
TARMED: tarif médical

Edited by A Stone; submitted 25.Aug.2025; peer-reviewed by AJ Umeano, F Okonkwo, R Deshpande, R Chi, Z Su; revised version received 14.Mar.2026; accepted 16.Mar.2026; published 12.May.2026.

Please cite as:
Giger OF, Ackermann LJ, Principe M, Meier S, Fleisch E, Gallani S, Brändle M, Kowatsch T, Jovanova M
Digital Quality Monitoring for Type 2 Diabetes in Swiss Primary Care: Qualitative Interview Study
J Med Internet Res 2026;28:e82960
URL: https://www.jmir.org/2026/1/e82960
doi:10.2196/82960

© Odile-Florence Giger, Lola Jo Ackermann, Marinja Principe, Simon Meier, Elgar Fleisch, Susanna Gallani, Michael Brändle, Tobias Kowatsch, Mia Jovanova. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Literacy and Interpersonal Trust as Predictors of Willingness to Share Patient-Generated Health Data Among Korean Internet Users: Cross-Sectional Study Using Privacy Calculus and Communication Privacy Management Theories

Dongsu Lee, MPH; Wonseuk Jang, PhD

Future Medicine Research Center, Gangnam Severance Hospital, College of Medicine, Yonsei University, 20 63rd Road, Unju-ro, Gangnam-gu, Seoul, Republic of Korea

Corresponding Author:
Wonseuk Jang, PhD
Future Medicine Research Center, Gangnam Severance Hospital, College of Medicine, Yonsei University, 20 63rd Road, Unju-ro, Gangnam-gu, Seoul, Republic of Korea

Abstract

Background: The proliferation of wearable devices and advances in data analytics are accelerating the adoption of personalized digital health care, relying heavily on patient-generated health data (PGHD). However, the sensitive nature of this data creates significant privacy boundaries. While previous research has focused on rational cost-benefit trade-offs, there is a limited understanding of how social and cognitive factors—specifically interpersonal trust and digital literacy (DL)—shape the data-sharing decisions of the general public.

Objective: This study aims to identify the factors predicting individuals' willingness to share health data (WS) by integrating privacy calculus and communication privacy management theories. It specifically examines the comparative influence of DL, interpersonal trust, and moral motivation on data-sharing decisions.

Methods: We analyzed data from the 2023 Korea Panel Survey on the Digital Society (n=4518), a nationwide representative sample of internet users. Survey-weighted structural equation modeling with weighted least squares mean and variance adjusted estimation was used to examine the relationships among perceived risk (PR), perceived benefit (PB), DL, interpersonal trust, and WS.

Results: PR was the strongest negative predictor of WS (standardized coefficient [std β]=−0.189, *P*<.001), whereas PB was the strongest positive predictor (std β=0.076, *P*=.009), followed by moral motivation (std β=0.062, *P*=.03) and interpersonal trust. DL showed a significant negative direct association with willingness to share (std β = −0.060, *P*<.001). However, subdimension analysis revealed heterogeneous mechanisms: the "understanding" dimension was associated with lower PR and indirectly promoted sharing, whereas use and engagement were associated with higher PR. Age-stratified analyses suggested potential heterogeneity in these relationships, although the overall interaction was not statistically significant.

Conclusions: PBs and risks were the strongest determinants of PGHD sharing, with benefits increasing and risks decreasing willingness to share. Beyond this risk-benefit balance, DL (particularly understanding) and interpersonal trust also played important roles, highlighting the need for trust-based and user-centric strategies to promote PGHD sharing and support the expansion of digital health care.

(J Med Internet Res 2026;28:e75448) doi:10.2196/75448

KEYWORDS

patient-generated health data; privacy; digital literacy; health data sharing; interpersonal trust; perceived risk; perceived benefit

Introduction

Background

Due to the 4th industrial revolution, the development of big data analysis technology and the active introduction of wearable devices into the medical field are increasing the possibility of introducing personalized medicine based on individual health data and daily life data [1,2]. However, securing real-world evidence for collecting health data and measuring clinical effects is still insufficient [3]. Despite the efforts of countries to expand digital health care, research is needed regarding privacy issues, clinical usefulness has not been sufficiently proven, and the development of new services is progressing slowly due to the regulatory environment of the medical device industry [4].

To address this, research has been conducted based on various theories on the sharing of personal health data [5,6]. Previous

XSL•FO
RenderX

research has mainly focused on personal health information [7,8]. However, the concept of personal health data has recently expanded beyond the scope of existing medical information to health life logs, and is developing into a concept that includes behavioral habits that affect an individual's health [9]. Amid these changes, the concept of patient-generated health data (PGHD) is gradually expanding. Consequently, there is a growing need for research to investigate the factors influencing individuals' willingness to share health data (WS), applying the concept of PGHD in a broader sense.

PGHD is health data that patients or the general public directly create and record using smart devices. It can play a key role in improving medical services and realizing customized health management [9]. However, the definition of PGHD is still being clarified [9,10], and the potential for using PGHD is expanding further with the spread of wearable devices. PGHD is attracting attention as an important driving force for the growth of digital health care and is recognized as an essential resource for expanding services [11,12].

PGHD, an important resource for the development of digital health care, has limitations in that it can be collected and used based on the consent of the individual who created it [13]. Recent research results have shown that the WS is known to be affected by the quality of data [12] and the purpose of use [6], and personal privacy concerns and perceived risks (PRs) have been identified as major factors that hinder the sharing of PGHD [14]. Sharing sensitive personal health data involves complex decision-making, and additional research is needed regarding privacy issues [4,5].

The willingness to share PGHD may be partially interpreted through the lens of communication privacy management (CPM) theory, which conceptualizes how individuals regulate the disclosure of private information within relational and digitally mediated contexts [15].

Private information refers to sensitive information that an individual considers to be his or her own, and includes PGHD, the subject of this study. Privacy boundaries refer to the psychological boundaries of how much an individual will share sensitive information with others, and can be a criterion for deciding whether or not to share. This study assumes that as digital literacy (DL) and interpersonal trust increase, privacy boundaries will loosen, increasing the willingness to share PGHD.

Previous studies have investigated factors that may affect the willingness to share PGHD, but most of them were limited to specific disease patient groups or small samples [4,5], and there is a lack of research that comprehensively analyzes the understanding of PGHD at the general public level and the psychological and social factors of individuals that affect the decision to share information.

In Korea, the "Korea Panel Survey on the Digital Society (KPSDS)," a national statistical survey targeting a sample of internet users, has been conducted annually since 2022 to investigate individuals' information management behaviors, such as WS, privacy concerns, DL, and social trust, as well as the level of public awareness of informatization.

Although eHEALS (eHealth Literacy Scale) [16,17] is used as a tool to measure the ability to use health data in a specific medical environment, it is reasonable to further expand the scope of the current study by using media DL as a measurement tool to broadly survey the willingness to share PGHD to the general public, who are potential customers for future provision, as general media literacy is ultimately a basic and essential ability for using wearable devices such as smartwatches and related health services.

General internet users can be considered a representative group for studying digital health care adoption, as they are more likely to engage with technologies such as smart health care and wearable devices [18]. In order to activate PGHD sharing, which is a key resource for expanding digital health care services, it is necessary to identify factors that influence the decision to share health data when sharing sensitive health data, such as PGHD, with internet users.

Objectives

This study aims to complement the limitations of previous studies and enhance the understanding of individuals' willingness to share PGHD. Specifically, it analyzes the impact of key factors such as PRs and benefits, DL, and interpersonal trust on the adjustment of privacy boundaries and the formation of privacy rules, based on the CPM theory and privacy calculus theory [19].

It is differentiated from existing studies in that it uses the results of the KPSDS (2023), a national statistic for Korea, to add factors such as DL and interpersonal trust in addition to the perceived benefit (PB) and PR explained in the existing privacy calculus theory, and analyzes factors affecting the decision to share personal information by linking it with the general public's willingness to share PGHD.

Through this, we analyze the factors influencing the decision to share personal health data for the purpose of creating a digital health care environment, additionally investigate the influence of DL and interpersonal trust on consent to share health data, seek ways to improve WS, and provide implications for establishing general policies for the expansion of digital health care services using PGHD.

Related Work and Model Development

Theoretical Integration: Privacy Calculus and CPM

Privacy calculus posits that individuals make disclosure decisions by weighing PRs against PBs [20]. In health data contexts, individuals evaluate potential privacy losses—such as misuse, unauthorized access, or secondary use—against expected gains, including improved services, personalized feedback, or social value [21,22]. This framework has been widely applied in digital environments [23], social media [24], health care and mHealth settings [25-27], and e-commerce [28], consistently demonstrating that PR reduces WS, whereas PB increases it [5,29]. Accordingly, we hypothesize that PR negatively (H1) and PB positively (H2) influence willingness to share PGHD.

However, sharing PGHD involves more than rational cost-benefit evaluation. As PGHD are continuously generated,

XSL•FO
RenderX

behaviorally detailed, and often highly sensitive, individuals must actively manage privacy boundaries within evolving digital ecosystems. CPM theory explains how disclosure is regulated through privacy rules shaped by relational expectations, perceived control, and trust [15,30]. CPM has been applied to health communication, wearable device use, and digital contact-tracing contexts, demonstrating how boundary coordination and trust dynamics shape disclosure decisions [19,21,24]. Rather than positioning CPM as an alternative to privacy calculus, we use it as a complementary perspective that helps interpret how individuals regulate privacy boundaries within digital health environments.

In this study, privacy calculus serves as the primary explanatory framework, while CPM is used as a complementary lens to interpret how DL and trust may shape individuals' privacy boundary regulation.

Within the structural equation modeling framework, PR and PB represent evaluations of privacy boundary permeability, while willingness to share PGHD reflects the outcome of boundary coordination governed by privacy rules. Interpersonal trust and governance-related perceptions function as contextual factors shaping the formation and adjustment of these privacy rules.

Antecedents of PR and PB

Within this integrated framework, willingness to share is primarily shaped by PR and PB. Prior research suggests that several psychological and contextual factors condition these perceptions.

DL, defined as the ability to understand, evaluate, and engage with digital systems [31], may influence both risk and benefit appraisals. Higher DL can reduce uncertainty about data practices and increase perceived control, thereby lowering PR [32,33], forming the basis for H3. At the same time, DL may enhance recognition of potential advantages associated with digital health participation, increasing PB [34] and potentially strengthening WS directly [18], supporting H4 and H5.

Trust also plays a central role in shaping privacy appraisals. Interpersonal trust–generalized expectations about others' intentions—has been shown to reduce PR in online disclosure contexts [22,24], supporting H6. It may also directly encourage sharing behavior and enhance PB [35], forming the basis for H7 and H8. Institutional trust (TR), by contrast, reflects confidence in formal governance and regulatory safeguards and is modeled primarily as an antecedent of PR.

Other contextual determinants further shape PR and PB. Privacy concern has been consistently associated with higher PR [24], whereas monetary benefit (MB) increases PB by enhancing the instrumental value of disclosure [3,36]. TR increases the perceived legitimacy of data governance, which in turn enhances the perceived societal and personal benefits of data sharing [8]. Information control (IC) and perceived effectiveness of government regulation (PE) may stabilize privacy boundaries and reduce anticipated violations, thereby lowering PR [5]. In addition, moral motivation (MM)—intrinsic intentions to contribute to socially beneficial outcomes—can increase WS by framing disclosure as prosocial and ethically meaningful [26,37].

Differential Effects of DL Subdimensions

DL is widely conceptualized as a multidimensional construct encompassing distinct domains of digital competence. Frameworks such as DigComp 3.0 distinguish differentiated competence areas, while empirical studies identify multiple behavioral domains of digital engagement [38,39]. Drawing on these perspectives, we conceptually organize DL into 3 competencies: digital literacy of use (DU), digital literacy of understanding (DD), and digital literacy of engagement (DE).

DU, reflecting operational skills, is associated with greater perceived control and digital self-efficacy [40]. DD, capturing critical evaluation and awareness of data practices, is linked to heightened privacy risk awareness [41]. DE, referring to active participation in digital ecosystems, is associated with social capital and PBs of interaction [42]. As these dimensions correspond to distinct psychological mechanisms—control, risk awareness, and benefit orientation—they may exert heterogeneous effects on WS. Accordingly, we examine whether DU, DD, and DE differentially influence willingness to share (H9).

Summary of Hypothesis Development

Integrating privacy calculus and CPM theory, the proposed model explains willingness to share PGHD primarily through PR and PB, while DL, interpersonal trust, MM, and governance-related perceptions shape these evaluations. This framework provides the theoretical basis for the following hypotheses:

- H1: PR negatively affects WS.
- H2: PB positively affects WS.
- H3: DL negatively affects PR.
- H4: DL positively affects WS.
- H5: DL positively affects PB.
- H6: interpersonal trust negatively affects PR.
- H7: interpersonal trust positively affects WS.
- H8: interpersonal trust positively affects PB.
- H9: DL subdimensions (DU, DD, and DE) exert differential effects on WS.

Methods

About KPSDS

Overview

This study is a secondary analysis of the 2023 User Panel Survey on the Intelligent Information Society, a nationwide representative longitudinal survey conducted by the Korea Information Society Development Institute (KISDI; 164004). Data were collected from October to December 2023 via 1:1 in-person computer-assisted personal interviewing interviews with 4581 daily internet users aged 15 - 69 years representing an estimated population of 39,027,595.

To minimize common method variance (CMV), the original survey used procedural remedies, including respondent

XSL•FO
RenderX

anonymity and the use of varied Likert-type scales (4-, 5-, and 7-point) across 7 distinct thematic sections.

Study Sample and Selection Bias Check

This study used data from the 2023 User Panel Survey on the Intelligent Information Society, a nationwide longitudinal survey conducted by the KISDI (164004).

The survey targeted internet users aged 15 to 69 years, representing an estimated population of 39,027,595. Responses of “no experience” (coded as 9) on PR items (q15_) were treated as missing. Participants who selected “no experience” for all PR items (63/4581, 1.38%) were excluded from the analysis because PR could not be meaningfully constructed for these cases, leaving a final analytic sample of 4518 respondents (weighted n=38,420,934).

To assess the potential impact of excluding respondents with missing PR values, we conducted sensitivity analyses using alternative imputation strategies. Missing PR values were imputed using minimum (1), median (2.5), and maximum (4) scores, and the structural model was re-estimated under each scenario. Additionally, to evaluate potential selection bias, we compared DL between excluded and included respondents using survey-weighted *t* tests.

Data Validity

To assess potential CMV, the Harman single-factor test was conducted at the item level using polychoric correlations to account for the ordinal nature of the data. Due to structurally missing responses (“no experience”), this supplementary test was designed to be evaluated separately for the core measurement items and the perceived-risk items.

Variables and Measures

In this study, variables used in previous studies were selected by using medical and health-related items from the 2023 KPSDS [5]. Detailed variable definitions, questionnaire items, and operationalization procedures are provided in Multimedia Appendix 1. For descriptive purposes, composite mean scores were reported for all variables to facilitate interpretation on the original Likert scale. However, in the structural equation modeling, PB and DL were modeled as latent constructs, whereas the remaining variables were operationalized as composite scores.

Dependent Variable

Regarding WS, the degree of consent and permission to share personal health data was measured using a 5-point Likert scale (q16_9, 1=absolutely not, 5=completely possible). In the structural equation modeling, WS was treated as an ordered categorical variable with 5 categories and specified as an ordered outcome in *lavaan*.

Independent Variables

Referring to the research design of previous studies, questions related to PR, PB, PE, MB, MM, IC, privacy concern, and TR were selected as measurement tools using the concept of MM defined in previous studies [5]. PR was measured on a 4-point Likert scale (1=always, 4=never). Responses coded as “no experience” (9) were treated as missing, and respondents who selected “no experience” for all PR items were excluded from the analysis (n=63). Valid PR items were reverse-coded so that higher scores indicate a higher frequency of PR (ie, 4=always, 1=never). All other items (PB, PE, MB, MM, IC, privacy concern, and TR) were assessed on a 5-point Likert scale (1=never, 5=completely possible) and operationalized as composite scores, except for PB, which was modeled as a latent construct.

DL was assessed using a media literacy framework rather than the conventional eHEALS, given that the creation and sharing of PGHD typically occur in a general digital environment. Therefore, general digital media usage skills were considered more relevant. A total of 23 items related to DL were rated on a 5-point Likert scale (1=“not at all” to 5=“very much”). The items were grouped into 3 subdimensions: DU (6 items, q33_1–q33_6), DD (11 items, q33_8–q33_15 and q33_19–q33_21), and DE (6 items, q33_16–q33_18 and q33_22–q33_24).

Each subdimension (DU, DD, and DE) was modeled as a first-order latent construct measured by its respective items. These first-order constructs were specified as indicators of a second-order latent construct representing overall DL. All subscales demonstrated high internal consistency, with Cronbach α values of 0.89 for DU, 0.93 for DD, and 0.91 for DE.

As a supplementary scale check, we also conducted an item-level CFA for the DL subdimensions (DU, DD, and DE); detailed results (including composite reliability [CR] and average variance extracted [AVE] for each subscale) are provided in Multimedia Appendix 2.

Interpersonal trust was assessed with a 7-point scale question, “How much do you generally trust people?” To address sparse distributions at extreme ends of the scale, the scale was recoded into 5 categories by combining responses 1 and 2, and 6 and 7.

Control Variables

Sociodemographic variables included sex, age group (AG), residential area (RA), education level, occupation, and household income (HI).

AGs were categorized into 5 groups: twenties and younger, thirties, forties, fifties, and sixty and older.

RA was classified into 3 categories: province, metropolitan city, and special city (Seoul).

Education level was recategorized into 3 groups: high school or below, college, and university or higher.

HI was classified into 4 categories: less than 2 million KRW, 2 to 4 million KRW, 4 to 6 million KRW, and 6 million KRW or more.

Additionally, perceived standard of living (SL) was included as a control variable. This was assessed by asking participants, "If we divide the living standards of our people into five levels: upper, upper-middle, middle, lower-middle, and lower, where do you think your family’s living standards fall?” on a 5-point Likert scale (1=upper, 5=low). Responses were categorized into “low” (4 - 5 points) and “high” (1 - 3 points) for analysis.

XSL•FO
RenderX

CHERRIES Checklist Mapping (Adapted)

This study is a secondary analysis of KISDI's national panel survey; therefore, CHERRIES (Checklist for Reporting Results of Internet E-Surveys) items related to online survey administration (eg, view rate, cookies, and IP checks) are not applicable. We report available items, including survey mode, sampling, consent procedures, and data handling per the KISDI documentation.

Study Design

We conducted a cross-sectional analysis of a nationwide representative survey. The proposed research model illustrating hypothesized relationships is shown in Figure 1.

Figure 1. Hypothesized structural model of WS based on a cross-sectional, nationwide representative survey of Korean internet users aged 15-69 years (South Korea, 2023). DL is modeled as a latent construct indicated by DU, DD, and DE. The model includes PR, PB, interpersonal trust, TR, and MM as predictors of WS. Demographic variables (SX, RA, EL, OC, HI, SL) were analyzed descriptively and via chi-square tests (Table S15, Multimedia Appendix 2) but were not included as covariates in the SEM. DD: digital literacy of understanding; DE: digital literacy of engagement; DL: digital literacy; DU: digital literacy of use; EL: education level; HI: household income; IC: information control; MB: monetary benefit; MM: moral motivation; OC: occupation; PB: perceived benefit; PE: perceived effectiveness of government regulation; PR: perceived risk; SL: perceived standard of living; SX: sex; RA: residential area; TR: institutional trust; WS: willingness to share health data.

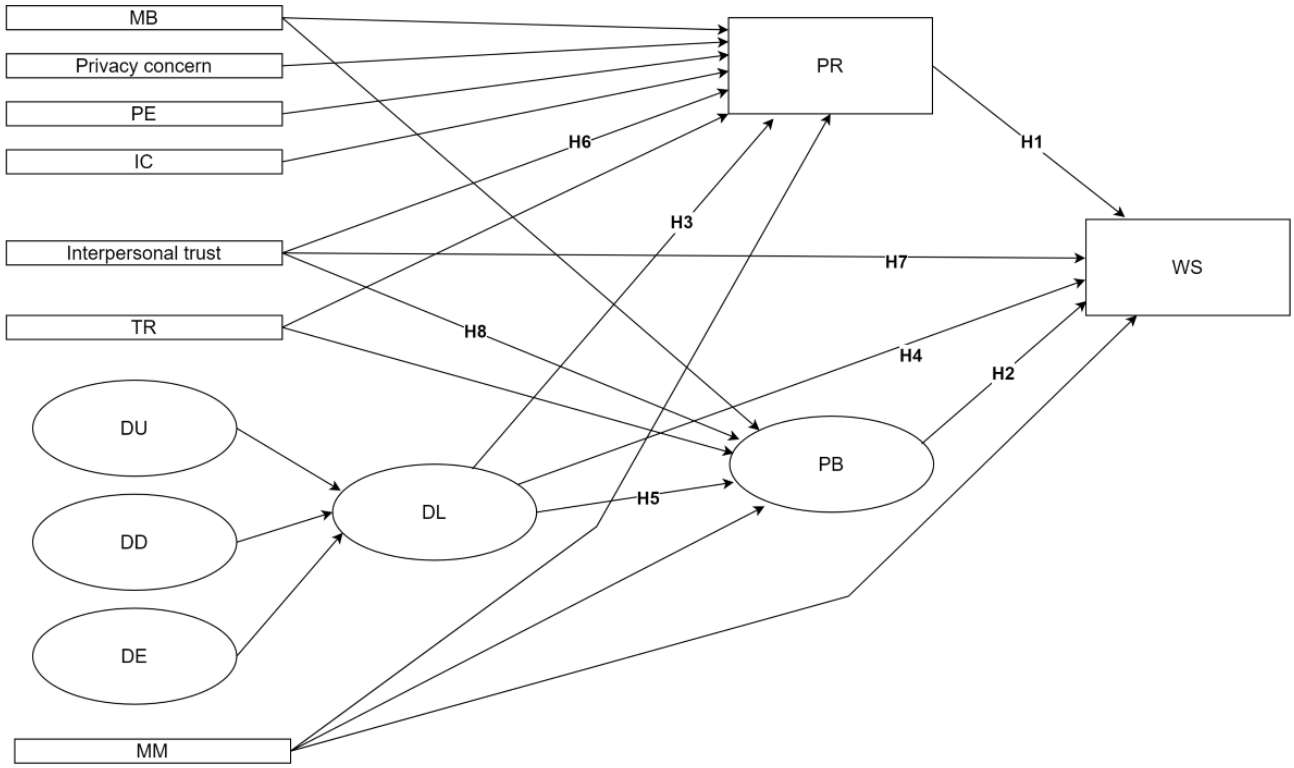

Analysis

Statistical Analysis and Structural Equation Modeling Specification

All analyses were conducted in R (version 4.5.2; R Foundation) using the *lavaan* package (version 0.6.21). Descriptive statistics and chi-square tests were used to summarize sample characteristics.

We used covariance-based structural equation modeling in *lavaan* and incorporated population weights provided by the national survey through the sampling.weights argument to obtain weighted parameter estimates. As the outcome (WS) was specified as an ordered categorical variable (5 categories), models were estimated using the WLSMV (weighted least squares mean and variance adjusted; robust DWLS) estimator. Under this framework, robust (scaled) fit indices and SEs were reported, and thresholds (t1-t4) were estimated for WS.

Only PB and DL-related constructs (DL, DU, DD, and DE) were specified as latent variables. PR and all other exogenous predictors were treated as observed composite variables, and WS was modeled as an ordered observed outcome.

To assess potential multicollinearity among trust-related constructs (interpersonal trust, TR, PE, and IC), we examined survey-weighted correlations and variance inflation factors.

Both unstandardized estimates (B) and fully standardized coefficients (std β; std all) were reported.

Measurement Model Evaluation

Measurement model evaluation focused on the latent measurement structure, including PB and the hierarchical DL construct (DU, DD, and DE, forming the higher-order construct DL). PR and other composite variables were not included in this evaluation because they were operationalized as observed indices rather than reflective latent constructs. Furthermore, because model 2 retained the same first-order measurement structure as model 1, the measurement evaluation was not repeated for model 2.

The measurement model was evaluated using fully standardized factor loadings (std all). Convergent validity was assessed

XSL•FO
RenderX

through CR and AVE, with recommended thresholds of CR ≥.70 and AVE ≥.50.

Discriminant validity was examined using the Fornell-Larcker criterion, whereby the square root of AVE for each construct was compared with interconstruct correlations, and the heterotrait-monotrait (HTMT) ratio, with values below 0.85 indicating adequate discriminant validity. HTMT ratios were computed based on unweighted polychoric correlations as a supplementary discriminant validity check, while the primary measurement evaluation was conducted using the survey-weighted estimation consistent with the structural model. CR, AVE, and Fornell-Larcker comparisons are reported in the Multimedia Appendix 2.

Model Fit Assessment

Model fit was evaluated using the robust (scaled) chi-square statistic and degrees of freedom, along with the comparative fit index (CFI), Tucker-Lewis index (TLI), root mean square error of approximation (RMSEA; 90% CI), and standardized root mean square residual (SRMR), as recommended for WLSMV estimation with ordered categorical outcomes.

Fit indices were interpreted based on established structural equation modeling guidelines [43,44]. As the chi-square statistic is sensitive to large sample sizes, greater emphasis was placed on incremental and residual-based fit indices (CFI, TLI, RMSEA, and SRMR) rather than on chi-square alone.

Structural Models

Overview

Sociodemographic variables, including RA, HI, and perceived SL, were not included as covariates in the primary structural equation modeling. The proposed model was specified as a theory-driven framework focusing on proximal psychological and perceptual mechanisms, such as PR and PB. Consistent with this perspective, sociodemographic factors were conceptualized as distal variables whose effects on data-sharing intentions may be mediated through these core psychological constructs.

We specified 2 structural equation models to address the hypothesized pathways (model 1) and differential subcomponent effects (model 2).

Model 1 (Hypothesized Structural Model)

This was specified as a theory-driven baseline structural model to test the hypothesized pathways (H1-H8).

DL was modeled as a second-order latent construct reflected by three first-order latent constructs: DU, DD, and DE. Each first-order construct was measured using its corresponding survey items (q33 series).

Interpersonal trust and DL were specified as predictors of PR and PB, which in turn predicted WS. MM was included as an exogenous theoretical predictor with direct effects on PR, PB, and WS.

Additional observed exogenous predictors—including MB, privacy concern, PE, IC, TR, and MM—were incorporated into the structural equations as specified in Figure 1.

All model specifications were theory-driven and transparently reported in the final path diagram and supplementary material.

Model 2 (H9: Differential Effects of DL Subcomponents)

To test hypothesis 9, the higher-order DL construct was removed, and its 3 first-order latent subdimensions (DU, DD, and DE) were retained and modeled separately as predictors of PR, PB, and WS, allowing direct comparison of their differential effects. To account for the conceptual relatedness among the DL subdimensions, 3 residual covariances (DU~~DD, DD~~DE, and DE~~DU) were specified a priori. This specification was theoretically justified because these subdimensions represent related and interdependent facets of the broader multidimensional construct of DL, and shared variance among them may remain unaccounted for when the higher-order factor is not explicitly modeled.

We estimated (1) direct effects of each subcomponent on WS, (2) indirect effects through PR and PB, (3) the total indirect effect (via PR+PB), and (4) total effects (direct+indirect). Indirect effects were explicitly defined within the structural equation modeling as products of constituent structural paths (eg, DD → PR×PR → WS) using defined parameters in *lavaan*, and were estimated under the same survey-weighted WLSMV framework.

To formally examine whether the effects of DU, DD, and DE differed, Wald chi-square tests were conducted using lavTestWald to compare relevant path coefficients and defined indirect effects across subcomponents. Direct, indirect, and total effects were reported to facilitate the interpretation of heterogeneous mechanisms.

Robustness Checks and Supplementary Analyses

To ensure the robustness of our findings against potential methodological artifacts, 2 supplementary analyses were conducted. First, to rigorously test the moderating effect of age without the loss of information associated with dichotomization, we used a survey-weighted generalized linear model (svyglm) to assess the interaction between DL and the full ordinal spectrum of age categories, followed by post hoc simple slope analyses. Std β were computed using survey-weighted SDs (std β=b×SD[X]/SD[Y]). Second, to address potential selection bias from excluding respondents with “no experience” in the PR variable, we conducted a sensitivity analysis by assigning extreme theoretical values (minimum, midpoint, and maximum) to the missing PR cases within the structural equation modeling framework. All robustness checks are detailed in Multimedia Appendix 2.

Ethical Considerations

This study was reviewed and approved for exemption by the Institutional Review Board of Gangnam Severance Hospital, Yonsei University (3-2024-0483). This research is a secondary analysis of the KPSDS, a nationwide representative longitudinal dataset provided by the KISDI. The KPSDS is a state-approved statistic (402001). According to the User Guidebook (2023), the original survey was conducted with the informed consent of all participants, and the data were collected in compliance with the Statistics Act of Korea. As the dataset provided to

XSL•FO
RenderX

researchers is strictly anonymized and deidentified by KISDI to ensure that no individual participant can be identified, the requirement for obtaining new informed consent for this secondary analysis was waived by the IRB. Regarding compensation, no additional incentives or payments were provided to the participants specifically for this secondary analysis. In the original KPSDS panel survey, participants received standard panel remuneration as determined by KISDI; detailed compensation conditions are not disclosed to secondary data users. Furthermore, we confirm that no individual participants can be identified in any images or supplementary materials associated with this paper.

Results

Sample Characteristics

Table 1 presents the demographic characteristics of this study's sample. The unweighted sample included 4518 Korean internet users aged 15‑69 years, and survey weighting yielded a nationwide representative population of approximately 38.4 million individuals. After weighting, sex distribution became balanced (49.1% female), and age and regional distributions aligned more closely with national population parameters. Educational attainment, occupational status, HI, and perceived SL showed similar patterns before and after weighting.

Table . Sample characteristics of Korean internet users aged 15‑69 years in a cross-sectional, nationwide representative survey (South Korea, 2023; n=4518). Unweighted and survey-weighted distributions are shown.

Control variable and category	Unweighted (n=4518), n (%)	Weighted (n=38,420,934), n (%)
Sex		
Female	2520 (55.8)	18,858,316 (49.1)
Male	1998 (44.2)	19,562,618 (50.9)
Age group (years)		
Twenties and younger	1072 (23.7)	8,235,437 (21.4)
Thirties	943 (20.9)	6,783,423 (17.7)
Forties	848 (18.8)	7,919,577 (20.6)
Fifties	911 (20.2)	8,350,620 (21.7)
Sixty and older	744 (16.5)	7,131,877 (18.6)
Residential area		
Province	2252 (49.8)	21,309,767 (55.5)
Metropolitan city	1310 (29)	9,768,815 (25.4)
Special city (Seoul)	956 (21.2)	7,342,352 (19.1)
Education level		
High school and below	1709 (37.8)	14,892,087 (38.8)
College	893 (19.8)	7,432,505 (19.3)
University and graduate school	1916 (42.4)	16,096,342 (41.9)
Occupation		
Student and unemployed	511 (11.3)	4,294,116 (11.2)
Employed	3522 (78)	30,259,593 (78.8)
Homemaker	485 (10.7)	3,867,225 (10.1)
Household income (KRW)[a]		
Less than 2 million	150 (3.3)	1,147,170 (3.0)
Less than 4 million	1414 (31.3)	11,123,523 (29.0)
Less than 6 million	1748 (38.7)	15,035,770 (39.1)
6 million and above	1206 (26.7)	11,114,471 (28.9)
Perceived standard of living		
Low	2274 (50.3)	18,565,404 (48.3)
High	2244 (49.7)	19,855,530 (51.7)

[a]A currency exchange rate of 1 KRW=US $0.00067 was applicable.

XSL•FO
RenderX

Descriptive Statistics for the Main Variables

Table 2 presents the descriptive statistics for the main variables used in this study, including both unweighted means and SDs and survey-weighted means with SEs. The mean score for WS was 2.903 (SD 1.029) in the unweighted sample and 2.889 (SE 0.018) in the weighted sample.

The mean score for PR was 2.123 (SD 0.641), while PB averaged 3.256 (SD 0.679). The PE had a mean of 3.644 (SD 0.597), MB averaged 3.296 (SD 0.688), and MM was 3.410 (SD 0.615).

Regarding data management perceptions, the mean score for IC was 3.667 (SD 0.623), privacy concern was 3.348 (SD 0.366), interpersonal trust was 3.682 (SD 0.990), and TR was 3.453 (SD 0.628). Weighted means were nearly identical to unweighted estimates.

Table . Descriptive statistics (unweighted and survey-weighted means and SEs) for main study variables in Korean internet users aged 15 - 69 years (South Korea, 2023; n=4518). WS is an ordinal self-reported outcome. For constructs modeled as latent variables in the structural model (eg, PB[a] and DL[b]), item-averaged composite scores are presented to provide interpretable descriptive statistics on the original Likert scale. DL and its subdimensions were measured on 5-point Likert scales, whereas PR[c] was measured on a 4-point Likert scale. For PB, descriptive statistics are based on the mean of its observed indicators (q17_8, q17_13, q17_14).

Independent variable	Sample (n=4518), mean (SD)	Weighted (n=38,420,934), mean (SE[d])
PR	2.123 (0.641)	2.128 (0.011)
PB	3.256 (0.679)	3.259 (0.012)
PE[e]	3.644 (0.597)	3.641 (0.01)
MB[f]	3.296 (0.688)	3.292 (0.011)
MM[g]	3.410 (0.615)	3.403 (0.011)
IC[h]	3.667 (0.623)	3.655 (0.011)
Privacy concern	3.348 (0.366)	3.352 (0.007)
Interpersonal trust	3.682 (0.990)	3.677 (0.017)
TR[i]	3.453 (0.628)	3.464 (0.011)
WS[j]	2.903 (1.029)	2.889 (0.018)
DL	3.329 (0.699)	3.315 (0.014)
DU[k]	3.739 (0.757)	3.716 (0.014)
DD[l]	3.174 (0.779)	3.163 (0.015)
DE[m]	3.187 (0.904)	3.178 (0.017)

[a]PB: perceived benefit.
[b]DL: Digital literacy.
[c]PR: perceived risk.
[d]Weighted estimates are reported with SEs rather than SDs to reflect the precision of population means under the complex survey design.
[e]PE: perceived effectiveness of government regulation.
[f]MB: monetary benefit.
[g]MM: moral motivation.
[h] IC: information control.
[i] TR: institutional trust.
[j] WS: willingness to share health data.
[k] DU: digital literacy of use.
[l] DD: digital literacy of understanding.
[m] DE: digital literacy of engagement.

Preliminary Analysis

Before the full-scale structural model analysis, a preliminary analysis was conducted to verify the validity of the data. Upon reviewing the distribution of control variables (Table S15 in Multimedia Appendix 2), sensitivity and selection bias regarding missing value handling (Tables S20 and S21 in Multimedia Appendix 2), assessment of CMV (Tables S22 and S23 in Multimedia Appendix 2), and multicollinearity (Table S25 in Multimedia Appendix 2), no issues were found that violated the model's main assumptions or caused serious bias. Detailed figures and results are presented in Multimedia Appendix 2.

XSL•FO
RenderX

Robustness Checks and Supplementary Analyses

Age as a Moderator of the Association Between DL and Risk Perception

Figure S1 and Tables S17 and S18 in Multimedia Appendix 2 illustrate the age-specific patterns in the association between DL and PR.

All models were estimated using survey-weighted regression with design-based degrees of freedom.

Table S17 in Multimedia Appendix 2 shows the overall interaction between DL and AG was not statistically significant in the fully adjusted model. However, simple slope analyses revealed heterogeneity in the direction and magnitude of the association across AGs.

Specifically, DL was negatively associated with PR among participants in their thirties (std β=−0.093, *P*=.04), whereas positive but nonsignificant associations were observed among participants in their forties (std β=0.094, *P*=.08), fifties (std β=0.085, *P*=.27), and 60 years and older (std β=0.044, *P*=.31). No association was observed among participants in their twenties or younger (std β=−0.006, *P*=.91).

These findings suggest potential variation in the relationship between DL and PR across AGs, although the overall interaction was not statistically significant.

PR Exclusion Sensitivity Test

Table S20 (Multimedia Appendix 2) showed that under plausible assumptions (exclusion and minimum imputation), the main indirect effects of DD remained consistent in magnitude and statistical significance. In contrast, extreme imputation scenarios (median and maximum) produced unstable estimates, indicating violations of model assumptions. Overall, these findings support the robustness of the main results under realistic assumptions. Additionally, excluded respondents showed significantly lower DL across all subdimensions (all *P*<.001).

Measurement Model Evaluation

Confirmatory factor analysis (CFA) was conducted as part of the first structural equation model (model 1), where DL was specified as a second-order latent construct measured by 3 first-order latent constructs: DU, DD, and DE. The model was estimated using a robust diagonally weighted least squares (WLSMV) estimator with survey weights applied.

The robust chi-square test was statistically significant (χ^2_{295}=7453.471, *P*<.001), which is expected in large samples. The overall model fit was acceptable: CFI=0.930, TLI=0.923, SRMR=0.052, and robust RMSEA=0.073. All fit indices met conventional criteria (CFI and TLI>0.90, SRMR<0.08, RMSEA<0.08), indicating satisfactory model fit.

Indicator Reliability and Internal Consistency

Table 3 presents the comprehensive item-level evaluation of the first-order reflective latent constructs used in model 1. All standardized factor loadings for the indicator items were highly significant (*P*<.001) and demonstrated strong indicator reliability, ranging from 0.600 to 0.867. All item loadings exceeded the conventional threshold of 0.60, confirming that the individual survey items adequately reflect their respective underlying latent constructs.

The internal consistency of the constructs was rigorously assessed using both Cronbach α and CR. All constructs exhibited robust reliability, with Cronbach α ranging from 0.720 to 0.933, and CR values ranging from 0.776 to 0.954. As all values were well above the recommended minimum threshold of 0.70, the measurement model demonstrated excellent internal consistency.

XSL•FO
RenderX

Table . Comprehensive measurement model table for model 1.

Construct and number	Mean (SD)	Weighted mean (SE)	Standardized loading	Loading SE	*z* value	*P* value	AVE[a]	CR[b]	α[c]
PB[d]						<.001	0.540	0.776	.720
1	3.29 (0.85)	3.30 (0.01)	0.600	0.020	29.792				
2	3.23 (0.86)	3.24 (0.02)	0.773	0.019	39.892				
3	3.24 (0.83)	3.24 (0.01)	0.813	0.021	39.647				
DU[e]						<.001	0.668	0.923	.894
1	3.80 (0.90)	3.77 (0.02)	0.830	0.008	71.315				
2	3.82 (0.90)	3.80 (0.02)	0.795	0.008	68.873				
3	3.69 (0.96)	3.67 (0.02)	0.818	0.009	68.079				
4	3.73 (0.97)	3.72 (0.02)	0.832	0.009	69.411				
5	3.57 (0.98)	3.54 (0.02)	0.862	0.009	66.289				
6	3.82 (0.91)	3.80 (0.02)	0.762	0.009	62.548				
DD[f]						<.001	0.652	0.954	.933
1	3.17 (1.00)	3.13 (0.02)	0.789	0.010	38.380				
2	3.10 (1.01)	3.07 (0.02)	0.829	0.010	38.315				
3	3.10 (1.02)	3.08 (0.02)	0.837	0.011	37.977				
4	3.09 (1.01)	3.07 (0.02)	0.825	0.010	38.267				
5	3.28 (0.98)	3.28 (0.02)	0.797	0.010	38.692				
6	3.32 (0.96)	3.31 (0.02)	0.799	0.010	37.824				
7	3.29 (0.97)	3.29 (0.02)	0.791	0.010	38.393				
8	3.17 (1.01)	3.16 (0.02)	0.788	0.010	38.464				
9	3.12 (1.03)	3.11 (0.02)	0.797	0.010	38.353				
10	3.13 (1.03)	3.14 (0.02)	0.818	0.010	38.613				
11	3.15 (1.04)	3.16 (0.02)	0.813	0.010	38.280				
DE[g]						<.001	0.708	0.936	.911
1	3.09 (1.13)	3.09 (0.02)	0.832	0.016	16.511				
2	3.23 (1.10)	3.23 (0.02)	0.828	0.016	16.406				
3	3.02 (1.13)	3.00 (0.02)	0.836	0.016	16.406				
4	3.26 (1.06)	3.26 (0.02)	0.850	0.017	16.261				
5	3.25 (1.05)	3.23 (0.02)	0.836	0.016	16.267				
6	3.27 (1.04)	3.26 (0.02)	0.867	0.017	16.362				

[a] AVE: average variance extracted

[b] CR: composite reliability.

[c] α: Cronbach α.

[d] PB: perceived benefit.

[e]DU: digital literacy of use.

[f]DD: digital literacy of understanding.

[g]DE: digital literacy of engagement.

Convergent and Discriminant Validity

Table 3 also shows that convergent validity was supported as the AVE values for all first-order latent constructs surpassed the established benchmark of 0.50, ranging from 0.540 to 0.708. Furthermore, the secondary loadings of DL on its subdimensions were DU=0.683, DD=0.876, and DE=0.949 (all *P*<.001). Although the AVE value (0.298) of the higher-level construct was lower than the threshold of 0.50, this is due to the structural characteristics of the multilevel model. As the AVE of the second-order factor is calculated through the first-order factors, the value tends to be calculated more strictly than in general

XSL•FO
RenderX

cases. However, as the factor loadings are high and the CR is sufficiently secured, it can be concluded that the higher-level construct adequately explains the lower-level dimensions.

Discriminant validity was assessed using the Fornell-Larcker criterion and the HTMT ratio. The Fornell-Larcker criterion was satisfied for most pairs of constructs, but between DD and DE, the square root of the AVE of DD (0.808) was lower than the correlation coefficient with DE (model 1: r=0.831; model 2: r=0.832). To complement these results, the HTMT ratio was additionally examined; the HTMT value between DD and DE was found to be 0.853, and all HTMT values were less than 0.90. Specific results are presented in Tables S6-S11 of Multimedia Appendix 2.

Note on Model 2

Model 2 retained the first-order latent constructs (DU, DD, and DE), excluding the higher-order construct (DL). The measurement model remained identical to model 1 without any changes to the specifications of the first-order constructs. Specific details are presented in Tables S9-S11 of Multimedia Appendix 2.

Structural Model Evaluation (Model 1: H1-H8)

Model Fit and Explanatory Power (Model 1)

Model 1 demonstrated satisfactory fit (χ^2_{525}=5372.19, P<.001, CFI=0.958, TLI=0.970, RMSEA=0.045, SRMR=0.049). The structural predictors explained 7.3% of the variance in WS (R^2_{WS}=0.073), 6.2% of PR (R^2_{PR}=0.062), and 56% of PB (R^2_{PB}=0.560).

Structural Path Analysis and Hypothesis Testing

Hypotheses H1 through H8 were tested using the first structural equation model (model 1), in which DL was modeled as a second-order latent construct composed of 3 first-order latent constructs: DU, DD, and DE.

Figure 2 presents the standardized structural path coefficients for model 1.

Figure 2. Structural model results (model 1, survey-weighted). Standardized path coefficients (std β) from the structural equation modeling testing hypotheses H1-H8 in Korean internet users aged 15-69 years (South Korea, 2023; n=4518). WS was modeled as an ordered outcome. Demographic variables (SX, RA, EL, OC, HI, SL) were analyzed descriptively and via chi-square tests (Table S15, Multimedia Appendix 2) but were not included as covariates in the structural equation modeling. The model focuses on theoretical paths and does not display sociodemographic control variables. Solid lines indicate statistically significant paths, and dashed lines indicate nonsignificant paths. DD: digital literacy of understanding; DE: digital literacy of engagement; DL: digital literacy; DU: digital literacy of use; EL: education level; HI: household income; IC: information control; MB: monetary benefit; MM: moral motivation; OC: occupation; PB: perceived benefit; PE: perceived effectiveness of government regulation; PR: perceived risk; RA: residential area; SL: perceived standard of living; std β: standardized path coefficients; SX: sex; TR: institutional trust; WS: willingness to share health data.

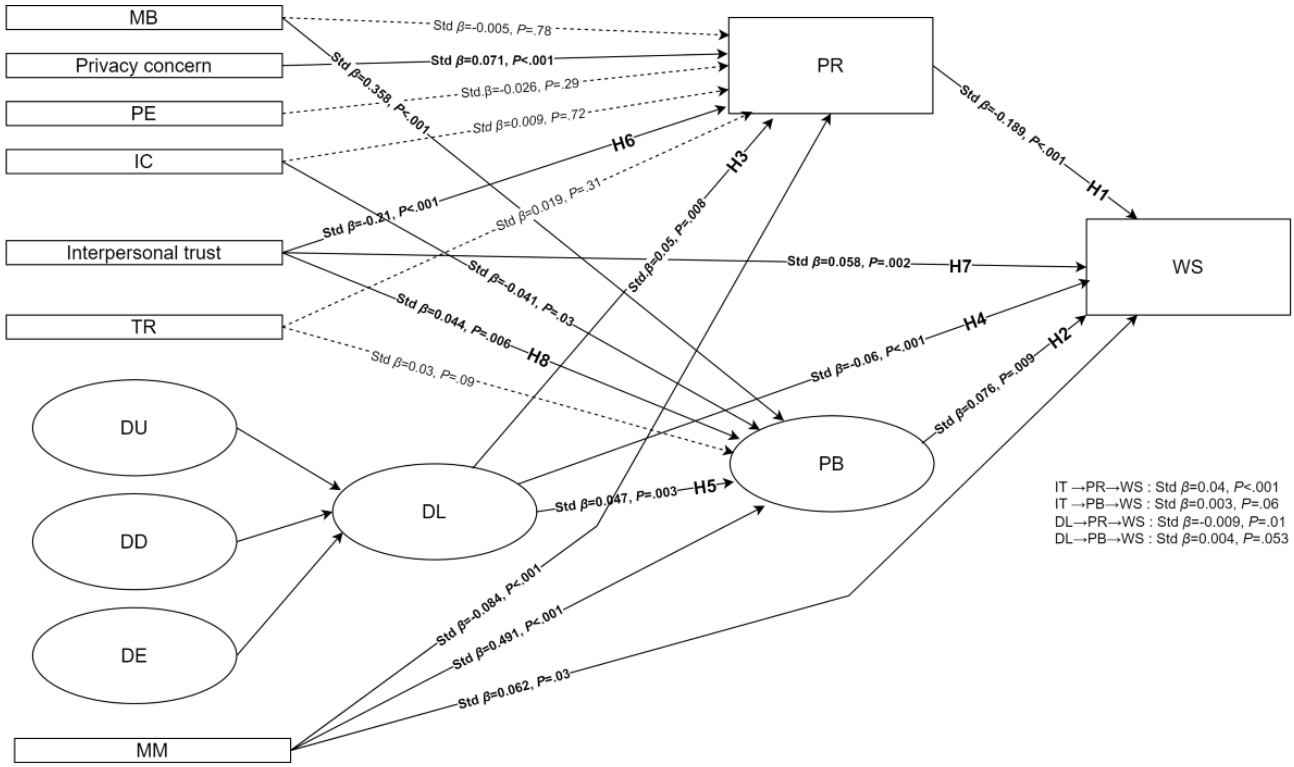

Regarding WS, PR exhibited the strongest negative direct effect (std β=−0.189, P<.001). PB showed the strongest positive standardized direct effects (std β=0.076, P=.009), followed by MM and interpersonal trust (MM: std β=0.062, P=.03; interpersonal trust: std β=0.058, P=.002). DL showed a small but statistically significant negative direct association with WS (std β=−0.060, P<.001).

In predicting PR, interpersonal trust showed the strongest negative effect (std β=−0.21, P<.001), followed by MM (std β=−0.084, P<.001). Privacy concern demonstrated a positive standardized effect (std β=0.071, P<.001), and DL also showed a small but significant positive association (std β=0.050, P=.008). TR, IC, MB, and PE were not statistically significant

XSL•FO
RenderX

predictors of PR (TR: std β=0.019, *P*=.31; IC: std β=0.009, *P*=.72; MB: std β=−0.005, *P*=.78; PE: std β=−0.026, *P*=.29).

For PB, MM had the strongest positive standardized effect (std β=0.491, *P*<.001), followed by MB (std β=0.358, *P*<.001). Interpersonal trust (std β=0.044, *P*=.006) and DL (std β=0.047, *P*=.003) also showed positive and significant effects. In contrast, IC had a significant negative effect on PB (std β=−0.041, *P*=.03), while TR was not a significant predictor (TR: std β=0.03, *P*=.09).

Overall, MM was an important predictor of both PB and WS, whereas PR was the main negative predictor of WS. Privacy concern and interpersonal trust were significant antecedents of PR, underscoring the role of psychological and relational factors in health data disclosure.

Hypothesis Validation (H1-H8)

H1 through H8 were tested using model 1. As shown in Tables 3 and 4 and Figure 2, a total of 6 of the 8 hypotheses (H1, H2, H5, H6, H7, and H8) were supported, whereas H3 and H4 were not supported.

Regarding WS, PR (H1) had a significant negative effect (std β=−0.189, *P*<.001), whereas PB (H2) showed a significant positive effect (std β=0.076, *P*=.009). Interpersonal trust (H7) and MM were also positively associated with WS (interpersonal trust: std β=0.058, *P*=.002; MM: std β=0.062, *P*=.03).

For PR, interpersonal trust (H6) showed a negative effect (std β=−0.210, *P*<.001), whereas DL (H3) showed a positive effect (std β=0.050, *P*=.008), contrary to the hypothesis.

For PB, DL (H5) and interpersonal trust (H8) both had positive effects (DL: std β=0.047, *P*=.003; interpersonal trust: std β=0.044, *P*=.006).

Finally, DL (H4) showed a negative direct effect on WS (std β=−0.060, *P*<.001), contrary to the hypothesized positive relationship.

Table . Survey-weighted structural equation modeling results for hypothesized direct paths (H1-H8; model 1) in Korean internet users aged 15 - 69 years (South Korea, 2023; n=4518).

Hypothesis	From	To	Estimate	SE	Standard β	*z* value	*P* value	Supported
H1: PR[a] has a negative effect on WS[b]	PR	WS	−0.298	0.025	−.189	−11.820	<.001	✓
H2: PB[c] has a positive effect on WS	PB	WS	0.122	0.047	.076	2.614	.009	✓
H3: DL has a negative effect on PR	DL[d]	PR	0.057	0.022	.050	2.646	.008	—[e]
H4: DL has a positive effect on WS	DL	WS	−0.109	0.031	−.060	−3.493	<.001	—
H5: DL has a positive effect on PB	DL	PB	0.054	0.018	.047	2.922	.003	✓
H6: interpersonal trust has a negative effect on PR	Interpersonal trust	PR	−0.137	0.011	−.210	−11.960	<.001	✓
H7: interpersonal trust has a positive effect on WS	Interpersonal trust	WS	0.060	0.020	.058	3.044	.002	✓
H8: interpersonal trust has a positive effect on PB	Interpersonal trust	PB	0.028	0.010	.044	2.776	.006	✓

[a]PR: perceived risk.

[b]WS: willingness to share health data.

[c]PB: perceived benefit.

[d]DL: digital literacy.

[e]Statistically significant, but the direction of effect is opposite.

XSL•FO
RenderX

Differential Effects by DL Subcomponent (Model 2: H9)

Model Fit and Explanatory Power (Model 2)

To examine whether the subdimensions of DU, DD, and DE differentially influenced privacy-related appraisals and sharing intentions, we estimated a second structural model (model 2).

Model 2 demonstrated satisfactory fit (χ^2_{519}=5255.31, *P*<.001, CFI=0.959, TLI=0.970, RMSEA=0.045, SRMR=0.047). The explanatory power of the model was comparable to model 1, accounting for 7.5% of the variance in willingness to share (R^2_{WS}=0.075), 9.4% of PR (R^2_{PR}=0.094), and 57% of PB (R^2_{PB}=0.570).

Global Wald Test Results

As shown in Table 5, the Wald test results for model 2 revealed differences in the direct effects of the DL subfactors (DU, DD, and DE) on sharing intention (WS; χ^2_2=7.646, *P*=.02). In other words, this means that the 3 subfactors influence sharing intention in different ways.

Table . Global Wald test results comparing direct, indirect, and total effects of DL[a] subdimensions (DU[b], DD[c], and DE[d]) on WS[e] in model 2 (South Korea, 2023; n=4518). Wald chi-square tests examine whether the effects of DL subdimensions (DU, DD, and DE) are equal across pathways (ie, H0: coefficients are equal). Tests are based on survey-weighted structural equation modeling estimates (WLSMV[f]). This table reports omnibus or global Wald tests for equality of effects across DU, DD, and DE. Pairwise Wald comparisons are provided in Table S16 in Multimedia Appendix 2.

Test type	χ^2 (df)	*P* value	Significance
Direct effect (DU/DD/DE → WS)	7.646 (2)	.02	✓
Indirect effect via PR[g]	61.604 (2)	<.001	✓
Indirect effect via PB[h]	5.614 (2)	.06	—[i]
Total indirect effect via PR and PB	56.271 (2)	<.001	✓
Total effects	2.019 (2)	.36	—

[a]DL: digital literacy.

[b]DU: digital literacy of use.

[c]DD: digital literacy of understanding.

[d]DE: digital literacy of engagement.

[e]WS: willingness to share health data.

[f]WLSMV: weighted least squares mean and variance adjusted.

[g]PR: perceived risk.

[h]PB: perceived benefit.

[i]Statistically significant, but the direction of effect is opposite.

The total indirect effect, which combines the indirect effects through PR and PB, also showed differences among the subfactors (χ^2_2=56.271, *P*<.001). Breaking this down, there was a difference in the indirect effect through PR (χ^2_2=61.604, *P*<.001), but no significant difference in the indirect effect through PB (χ^2_2=5.614, *P*=.06).

On the other hand, the total effect, which combines both direct and indirect effects, showed no difference among the subfactors (χ^2_2=2.019, *P*=.36). This means that while the processes through which each element exerts influence differ, the magnitude of the final overall impact is similar.

In summary, differences among the subelements of DL were primarily observed in the “path through risk (PR),” while there were no significant differences in the “path through benefit (PB)” or in the overall effect.

The standardized direct, indirect, and total effect estimates for each DL subdimension are presented in Table 6.

XSL•FO
RenderX

Table . Parameter estimates for direct, indirect, and total effects of DL subdimensions on WS[a] in model 2 (South Korea, 2023; n=4518). Indirect effects are decomposed via PR[b] and PB[c]. WS was modeled as an ordered outcome using survey-weighted structural equation modeling (WLSMV[d] estimation).

Pathway and subdimension	Estimate[e]	SE	Std β[f]	*P* value
Direct (DL[g] → WS)				
DU	−0.028	.033	−.023	.38
DD	−0.141	.043	−.108	.001
DE	.084	.045	.069	.06
Indirect via PR				
DU	−0.035	.007	−.028	<.001
DD	.066	.010	.051	<.001
DE	−0.048	.009	−.039	<.001
Indirect via PB				
DU	−0.004	.003	−.003	.15
DD	.021	.009	.016	.02
DE	−0.012	.006	−.010	.04
Total indirect				
DU	−0.038	.007	−.030	<.001
DD	.087	.013	.066	<.001
DE	−0.060	.011	−.049	<.001
Total effect				
DU	−0.067	.032	−.053	.04
DD	−0.055	.040	−.042	.17
DE	.024	.043	.020	.57

[a]WS: willingness to share health data.
[b]PR: perceived risk.
[c]PB: perceived benefit.
[d]WLSMV: weighted least squares mean and variance adjusted.
[e]Estimate: unstandardized coefficient.
[f]Std β: standardized coefficient.
[g]DL: digital literacy.

Differential Direct and Indirect Effects

Figure 3 shows the standardized direct, indirect, and total effects of DL subdimensions (DU, DD, and DE) on health data sharing intention (WS).

A significant negative effect was observed only for DD (std β=−0.108, *P*=.001), while DU (std β=−0.023, *P*=.38) and DE (std β=0.069, *P*=.06) were not statistically significant.

In terms of total indirect effects considering both PR and PB, DD showed a significant positive effect (std β=0.066, *P*<.001), while DU and DE showed significant negative effects (DU: std β=−0.030, *P*<.001; DE: std β=−0.049, *P*<.001).

Upon decomposition, for the PR mediation path, DD showed a significant positive indirect effect (std β=0.051, *P*<.001), whereas DU and DE showed significant negative indirect effects (DU: std β=−0.028, *P*<.001; DE: std β=−0.039, *P*<.001), reflecting the opposite direction of the DL → PR path and the PR → WS path.

For the PB mediation path, DD showed a significant positive indirect effect (std β=0.016, *P*=.02), while DE showed a small but statistically significant negative indirect effect (std β=−0.010, *P*=.04). DU also showed a negative indirect effect through PB (std β=−0.003, *P*=.15), but this was not statistically significant. Looking at the overall effects, only DU showed a statistically significant negative effect on WS (std β=−0.053, *P*=.04). On the other hand, DD (std β=−0.042, *P*=.17) and DE (std β=0.020, *P*=.57) were not statistically significant, and the CI includes 0, indicating uncertainty regarding the direction of the effect.

XSL•FO
RenderX

Figure 3. Standardized direct, indirect, and total effects of DL subdimensions on willingness to share health data (model 2). Effects are presented as standardized coefficients (std β) with 95% CIs. Panels show (A) direct effects, (B) total indirect effects, (B1) indirect effects via perceived risk, (B2) indirect effects via perceived benefit, and (C) total effects. DL subdimensions include DU, DD, and DE. Panel (B) presents the total indirect effects, which are decomposed into indirect effects via perceived risk (B1) and perceived benefit (B2). DD: digital literacy of understanding; DE: digital literacy of engagement; DU: digital literacy of use.

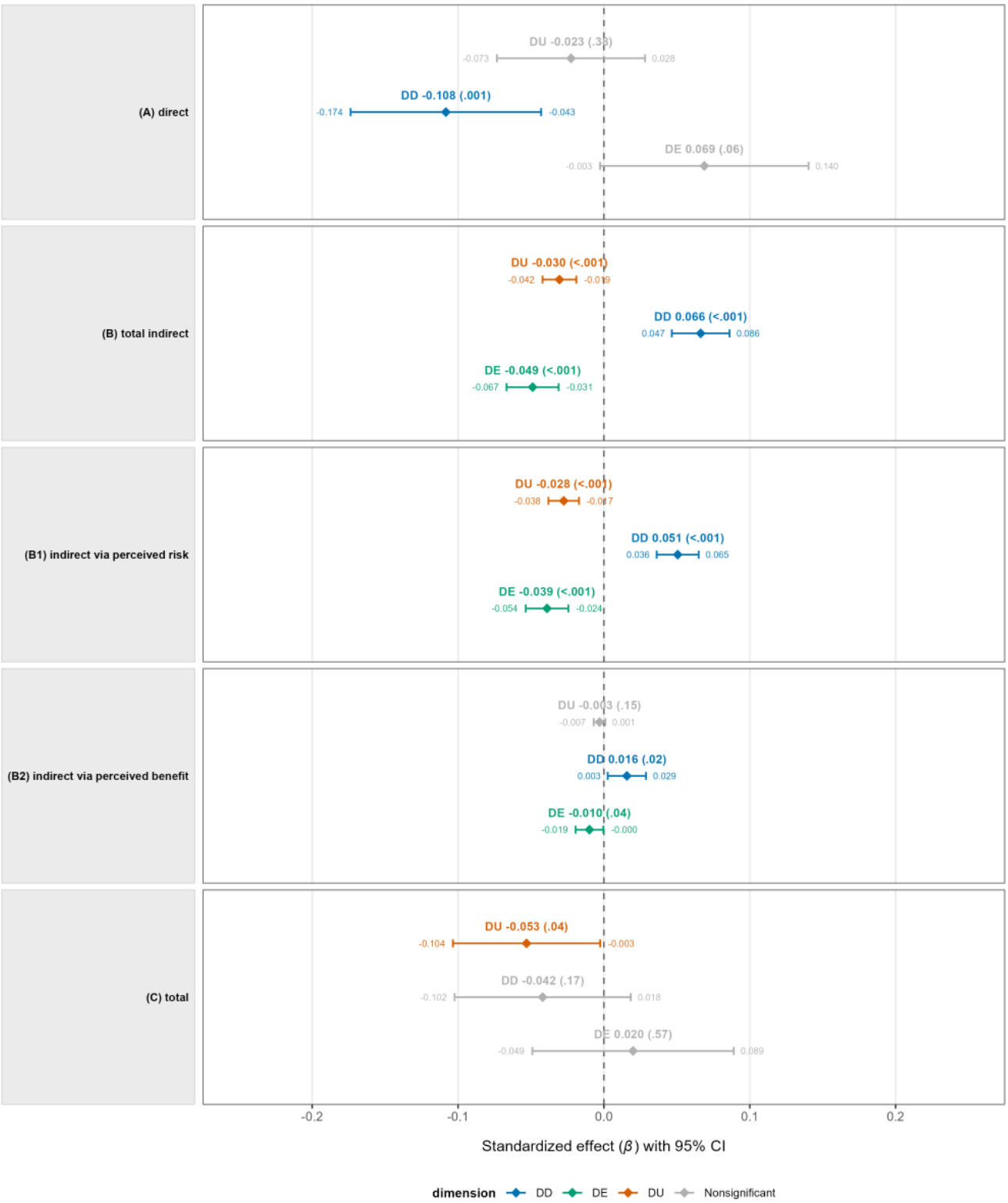

Discussion

Principal Results

This study analyzed the factors influencing an individual's WS based on CPM theory [15,30] and privacy calculus theory [19,24,36]. In particular, it examined the impact of psychological and social factors, such as DL, interpersonal trust, and MM, on data sharing behavior.

The results supported the core assumptions of privacy calculus theory. Specifically, the WS is determined by the balance between PR and PB, with PB appearing as the strongest positive predictor. Additionally, MM (std β=0.062, *P*=.03) and

XSL•FO
RenderX

interpersonal trust (std β=0.058, *P*=.002) also showed significant positive effects, suggesting that data sharing decisions are formed not only by personal gain but also by altruistic values and relational factors [26,29].

Among the components of DL, "understanding" was found to have the greatest indirect effect through PR and PB. Contrary to existing hypotheses, DL was found to both increase PR (std β=0.050, *P*=.008) and directly reduce willingness to share (std β=−0.060, *P*<.001).

This suggests that an individual's cognitive ability indirectly shapes data sharing behavior by structuring evaluations of risk and benefit, and that DL can act as a kind of "filter" in the data sharing decision-making process, leading to more prudent judgment.

These results highlight the important role of user competence in shaping health data sharing behavior [32] and are consistent with the CPM theory [8,35], which posits that data sharing is a trust-based boundary management process. In other words, individuals manage information through privacy boundaries that change according to digital competence and social context, rather than fixed standards, and DL and interpersonal trust play a key role in this boundary-setting process [15,30]. In this regard, this study complements the privacy calculus perspective and demonstrates that data sharing is not merely a rational choice but a dynamic process involving cognitive evaluation and social factors.

Age Interaction

The results of this study suggest that the relationship between DL and PR may be moderated by age. While no significant association between the 2 variables was observed in the overall sample, post hoc analysis confirmed a statistically significant relationship between DL and PR (std β=−0.093, *P*=.04) only in the thirties AG. Although statistical significance was not reached in other AGs, it is noteworthy that the slopes varied by age.

These results suggest that DL may be reflected differently in how risk is assessed depending on age. In particular, considering that DL includes an understanding of privacy protection policies and the context of data usage, it is possible that for a specific AG (thirties), these competencies served as a criterion for evaluating privacy risk.

This partially aligns with the CPM theory, which posits that an individual's privacy boundary setting varies depending on the context of information and the individual's rules. However, as these results are based on post hoc analysis, a cautious approach is required when generalizing them to all AGs or interpreting them as definitive conclusions; further verification is necessary through follow-up research that considers generational characteristics.

Comparison With Prior Work

The results of this study are largely consistent with prior research based on privacy calculus. Previous studies have shown that WS decreases as PR increases [3,5,45] and increases as PBs increase [4,6,25]. However, unlike earlier studies suggesting that DL reduces risk through greater technological control, the present findings align with more recent research indicating that DL may increase risk sensitivity [46]. This provides a possible explanation for why hypotheses H3 and H4 were not supported as expected. DL may enable individuals to recognize subtle threats in increasingly complex data environments, which can lead to more cautious data-sharing behavior. This tendency toward greater caution associated with higher DL offers a new perspective on traditional privacy calculus models. This relationship may be particularly important in the context of personal health data (PGHD), which is often sensitive and frequently collected.

The finding that MM has a significant influence on WS is consistent with prior research [26,29]. While previous studies have primarily focused on individual cost-benefit analysis, the present study highlights the importance of collective values, such as contributing to public health.

The finding that interpersonal trust has a significant positive effect on willingness to share is also consistent with prior research [8,32,35,47]. However, while interpersonal trust was significant, variables reflecting institutional factors (eg, perceived institutional effectiveness, PE) were not. This partially contrasts with previous studies suggesting that it can reduce PR [48]. This difference may reflect variation in how different types of trust influence data-sharing decisions. In particular, in cultures such as Korea that emphasize personal relationships and social networks, interpersonal trust may play a more prominent role than formal TR, which may help explain differences from findings in Western contexts [47,49]. In addition, institutional factors in this study were operationalized as perceived institutional effectiveness rather than trust in government, which may also contribute to these differences.

Finally, this study used nationwide representative data from Korean internet users rather than specific subpopulations, thereby enhancing the generalizability of the findings and strengthening their policy relevance compared to prior research [4,5,14].

Implications

To promote health data sharing within digital health systems, policy and intervention strategies should extend beyond technical infrastructure to incorporate psychological and social dimensions. Our analysis identified the "understanding" component of DL as having the strongest indirect effect on WS. This suggests that DL education should emphasize critical thinking, ethical reasoning, and data interpretation rather than focusing solely on technical or operational skills [33,34].

Although the overall interaction effect between age and DL did not reach statistical significance (*P*=.06), the heterogeneity of the slopes by AG observed in the post hoc analysis offers notable implications. The positive association observed in some AGs suggests the possibility that high literacy may have both facilitating and inhibiting effects, potentially increasing awareness of data-related risks and making people cautious about sharing. Therefore, when establishing future strategies to promote the sharing of personal health data (PGHD), it should move beyond uniform digital competency training and simultaneously implement trust-based governance and specific protective measures that consider age-specific characteristics.

XSL•FO
RenderX

Additionally, the variability in willingness to share PGHD depending on its intended use highlights the importance of contextual sensitivity. While participants expressed high willingness to share data for public health or research purposes, their willingness decreased markedly when the purpose involved commercial use, largely due to elevated perceptions of risk [9,14,50]. These findings indicate the need for purpose-specific consent models and clearer communication strategies that reflect the intended use of health data.

Policy Recommendations

The findings of this study yield the following policy implications for advancing digital health strategies and health data governance frameworks.

First, consent mechanisms should reflect the sensitivity of data types. PGHD range from relatively low-risk information (eg, step count or sleep patterns) to highly sensitive data (eg, heart rate, GPS, or stress levels). A one-size-fits-all consent process is inadequate. Instead, category-specific notifications and granular, opt-in consent designs—particularly for potential commercial uses—are essential for respecting user preferences and mitigating PRs [9,47,50].

Second, data-sharing frameworks should emphasize public value. Given the strong effect of MM on WS, data initiatives should be framed around collective benefits, such as medical research and disease prevention [26,29]. Institutions that collect or use PGHD should clearly communicate how data use aligns with public interest to increase transparency and trust.

Third, DL should be a core focus of policy. The “understanding” component of DL showed the most significant indirect effect on willingness to share. This underscores the need for education that fosters critical judgment, ethical awareness, and user agency, rather than focusing solely on technical skills [33,34,51]. Literacy strategies should move toward participatory and contextual learning that enables individuals to make informed decisions in digital health environments.

Fourth, trust-centered data governance must be prioritized. For PGHD to be accepted and effectively used—whether by governments, health care institutions, or corporations—technical safeguards must be supplemented with transparent practices, clear accountability mechanisms, and meaningful stakeholder engagement. Interpersonal trust plays a critical role in shaping users’ perceptions and behaviors, and governance models must reflect this social dimension [8,32,35].

Taken together, these strategies can foster a sustainable data-sharing culture grounded in public interest, ethical responsibility, and user empowerment.

Limitations

This study has several limitations. First, the cross-sectional, self-reported design limits causal inference and may be subject to CMV. Although Harman single-factor test suggested that a single factor did not dominate the covariance among measures (including a polychoric-based variant), CMV cannot be fully ruled out. Second, DL was measured using a general media literacy scale rather than a health-specific instrument such as eHEALS [16,17], reflecting the broader, nonclinical digital environments in which PGHD are created and shared. However, this may limit comparability with prior studies that operationalized DL using eHEALS. Third, WS was assessed using a single self-reported item that did not specify the data recipient (eg, government agencies, academic researchers, or commercial vendors). As sharing intentions is known to vary by target, this ambiguity may weaken construct validity and blur distinctions between institutional and interpersonal trust. Fourth, although attitudes differed by data use purpose (eg, public vs commercial), no behavioral or scenario-based validations were conducted. In addition, respondents who reported “no experience” across all PR items were excluded because PR could not be computed, which may slightly limit generalizability to internet users with minimal exposure to the relevant digital context; however, the excluded proportion was small (1.38%). Future research should use longitudinal, qualitative, or scenario-based experimental designs to validate these mechanisms and better capture behavioral outcomes. The AVE of the meta-DL construct was found to be lower than the recommended threshold (AVE=0.298). However, the CR was 0.879, satisfying the general criterion (CR≥0.70). While low AVE values are common in hierarchical second-factor models due to the integration of various subdimensions, this suggests the possibility that convergent validity may be limited at the higher level. Therefore, caution is required in interpreting the meta-construct. Finally, the model’s relatively low explanatory power suggests that a significant portion of the variability in the WS remains unexplained. This appears to reflect the complex real-world decision-making processes occurring within heterogeneous populations. Nevertheless, this study contributes to understanding the mechanisms related to DL and privacy awareness. Future research should include additional contextual factors to enhance explanatory power.

Despite these limitations, this study offers several notable strengths. First, unlike previous research limited to specific patient groups or technology users [4,14,21], this study analyzed nationwide representative data from general internet users in Korea, thereby enhancing generalizability. Second, by disaggregating DL into 3 subdimensions—use, understanding, and engagement—it provides practical insights for designing targeted digital health education policies [34,37,51]. Third, the integration of both privacy calculus and CPM theories enabled a multidimensional analysis of data-sharing behavior, incorporating both rational decision-making and privacy boundary management [15,30].

Conclusions

This study identified key psychological and social factors influencing the WS based on privacy calculus and CPM theories. The results showed that PBs had the strongest positive influence, while PRs had the strongest negative influence on willingness to share. Additionally, MM, interpersonal trust, and the “understanding” dimension of DL were significant factors, indicating that health data sharing involves both cognitive and social processes beyond simple cost-benefit evaluations.

Considering that the relationship between DL and risk perception varies by age, tailored policies and educational interventions are needed. Furthermore, as WS differs depending

XSL•FO
RenderX

on the purpose (eg, public vs commercial) and sensitivity, providing clear information and purpose-specific consent mechanisms is essential. These findings highlight the importance of trust-based, user-centric strategies to promote health data sharing and support the expansion of digital health care.

Future research should improve measurement tools to reduce interpretation bias and better capture contextual factors influencing health data sharing. In particular, health literacy instruments such as eHEALS may allow more precise assessment of digital capabilities. Additionally, studies involving specific clinical populations are needed to examine how willingness to share data translates into actual health behaviors and outcomes.

Acknowledgments

The author gratefully acknowledges JY Lee for his contribution to the conceptual development and

research direction regarding PGHD. Generative AI (artificial intelligence) tools were used as supportive instruments during manuscript preparation. ChatGPT (OpenAI, GPT-5.2) and Gemini (Google LLC, version 3) were used to assist with statistical interpretation, language translation, and proofreading. Cursor (Anysphere) was used as an AI-assisted coding editor to support R programming code generation. No AI tool was used to generate original data or to make autonomous analytical or interpretative decisions. The author critically reviewed, validated, and verified all AI-assisted outputs and retains full responsibility for the accuracy, integrity, and originality of the final content.

Funding

The authors declared no financial support was received for this work.

Data Availability

The dataset analyzed in this study is publicly available through the User Policy Archive for AI Services, operated by the Korea Information Society Development Institute (KISDI), and can be downloaded from its official website [52]. The analytic code and derived results (eg, processed variables and model outputs) are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization, methodology, formal analysis, data curation, writing – original draft, writing – review & editing, and visualization: DL
Supervision: WJ

Conflicts of Interest

None declared.

Multimedia Appendix 1
Variable definitions and questionnaire items used in this study.
[XLSX File, 23 KB - jmir_v28i1e75448_app1.xlsx]

Multimedia Appendix 2
Data tables and figures.
[DOCX File, 525 KB - jmir_v28i1e75448_app2.docx]

References

1. Schwab K. The Fourth Industrial Revolution. : World Economic Forum; 2016 URL: https://law.unimelb.edu.au/__data/assets/pdf_file/0005/3385454/Schwab-The_Fourth_Industrial_Revolution_Klaus_S.pdf [accessed 2026-05-12]
2. Maru D. What the Fourth Industrial Revolution can teach us about healthcare in poor countries. World Economic Forum. 2017 Jul 20. URL: https://www.weforum.org/stories/2017/07/n-all-digital-foundations-for-universal-health-coverage-and-beyond/ [accessed 2026-05-12]
3. Baines R, Stevens S, Austin D, et al. Patient and public willingness to share personal health data for third-party or secondary uses: systematic review. J Med Internet Res 2024 Mar 5;26:e50421. [doi: 10.2196/50421] [Medline: 38441944]
4. Benevento M, Mandarelli G, Carravetta F, et al. Measuring the willingness to share personal health information: a systematic review. Front Public Health 2023;11:1213615. [doi: 10.3389/fpubh.2023.1213615] [Medline: 37546309]
5. Shi J, Yuan R, Yan X, et al. Factors influencing the sharing of personal health data based on the integrated theory of privacy calculus and theory of planned behaviors framework: results of a cross-sectional study of Chinese patients in the Yangtze River Delta. J Med Internet Res 2023 Jul 6;25:e46562. [doi: 10.2196/46562] [Medline: 37410526]

XSL•FO
RenderX

6. Cascini F, Pantovic A, Al-Ajlouni YA, Puleo V, De Maio L, Ricciardi W. Health data sharing attitudes towards primary and secondary use of data: a systematic review. eClinicalMedicine 2024 May;71:102551. [doi: 10.1016/j.eclinm.2024.102551] [Medline: 38533128]
7. Kassam I, Ilkina D, Kemp J, Roble H, Carter-Langford A, Shen N. Patient perspectives and preferences for consent in the digital health context: state-of-the-art literature review. J Med Internet Res 2023 Feb 10;25:e42507. [doi: 10.2196/42507] [Medline: 36763409]
8. Busch-Casler J, Radic M. Trust and health information exchanges: qualitative analysis of the intent to share personal health information. J Med Internet Res 2023 Aug 30;25:e41635. [doi: 10.2196/41635] [Medline: 37647102]
9. Khatiwada P, Yang B, Lin JC, Blobel B. Patient-generated health data (PGHD): understanding, requirements, challenges, and existing techniques for data security and privacy. J Pers Med 2024 Mar 3;14(3):282. [doi: 10.3390/jpm14030282] [Medline: 38541024]
10. Guardado S, Karampela M, Isomursu M, Grundstrom C. Use of patient-generated health data from consumer-grade devices by health care professionals in the clinic: systematic review. J Med Internet Res 2024 May 31;26:e49320. [doi: 10.2196/49320] [Medline: 38820580]
11. Mars M, Scott RE. Electronic patient-generated health data for healthcare. In: Linwood SL, editor. Digital Health: Exon Publications; 2022:1-16. [doi: 10.36255/exon-publications-digital-health-patient-generated-health-data]
12. Abdolkhani R, Gray K, Borda A, DeSouza R. Recommendations for the quality management of patient-generated health data in remote patient monitoring: mixed methods study. JMIR mHealth uHealth 2023 Feb 24;11:e35917. [doi: 10.2196/35917] [Medline: 36826986]
13. Lee AR, Koo D, Kim IK, et al. Identifying facilitators of and barriers to the adoption of dynamic consent in digital health ecosystems: a scoping review. BMC Med Ethics 2023 Dec 1;24(1):107. [doi: 10.1186/s12910-023-00988-9] [Medline: 38041034]
14. Simpson E, Brown R, Sillence E, et al. Understanding the barriers and facilitators to sharing patient-generated health data using digital technology for people living with long-term health conditions: a narrative review. Front Public Health 2021;9:641424. [doi: 10.3389/fpubh.2021.641424] [Medline: 34888271]
15. Petronio S. Communication boundary management: a theoretical model of managing disclosure of private information between marital couples. Commun Theory 1991 Nov;1(4):311-335. [doi: 10.1111/j.1468-2885.1991.tb00023.x]
16. Norman CD, Skinner HA. eHEALS: the eHealth Literacy Scale. J Med Internet Res 2006 Nov 14;8(4):e27. [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
17. Norman C. eHealth literacy 2.0: problems and opportunities with an evolving concept. J Med Internet Res 2011 Dec 23;13(4):e125. [doi: 10.2196/jmir.2035] [Medline: 22193243]
18. Kontos E, Blake KD, Chou WY, Prestin A. Predictors of eHealth usage: insights on the digital divide from the Health Information National Trends Survey 2012. J Med Internet Res 2014 Jul 16;16(7):e172. [doi: 10.2196/jmir.3117] [Medline: 25048379]
19. Hong SJ, Cho H. Privacy management and health information sharing via contact tracing during the COVID-19 pandemic: a hypothetical study on AI-based technologies. Health Commun 2023 May;38(5):913-924. [doi: 10.1080/10410236.2021.1981565] [Medline: 34555998]
20. Dinev T, Hart P. An extended privacy calculus model for e-commerce transactions. Inf Syst Res 2006 Mar;17(1):61-80. [doi: 10.1287/isre.1060.0080]
21. Abdelhamid M. Fitness tracker information and privacy management: empirical study. J Med Internet Res 2021 Nov 16;23(11):e23059. [doi: 10.2196/23059] [Medline: 34783672]
22. Bansal G, Zahedi F“, Gefen D. The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online. Decis Support Syst 2010 May;49(2):138-150. [doi: 10.1016/j.dss.2010.01.010]
23. Li H, Sarathy R, Xu H. The role of affect and cognition on online consumers’ decision to disclose personal information to unfamiliar online vendors. Decis Support Syst 2011 Jun;51(3):434-445. [doi: 10.1016/j.dss.2011.01.017]
24. Hassandoust F, Akhlaghpour S, Johnston AC. Individuals’ privacy concerns and adoption of contact tracing mobile applications in a pandemic: a situational privacy calculus perspective. J Am Med Inform Assoc 2021 Mar 1;28(3):463-471. [doi: 10.1093/jamia/ocaa240] [Medline: 33164077]
25. Maus B, Olsson CM, Salvi D. Privacy personas for IoT-based health research: a privacy calculus approach. Front Digit Health 2021;3:675754. [doi: 10.3389/fdgth.2021.675754] [Medline: 34977856]
26. Princi E, Krämer NC. Out of control - privacy calculus and the effect of perceived control and moral considerations on the usage of IoT healthcare devices. Front Psychol 2020;11:582054. [doi: 10.3389/fpsyg.2020.582054] [Medline: 33262731]
27. von Kalckreuth N, Feufel MA. Extending the privacy calculus to the mHealth domain: survey study on the intention to use mHealth apps in Germany. JMIR Hum Factors 2023 Aug 16;10:e45503. [doi: 10.2196/45503] [Medline: 37585259]
28. Gouthier MHJ, Nennstiel C, Kern N, Wendel L. The more the better? Data disclosure between the conflicting priorities of privacy concerns, information sensitivity and personalization in e-commerce. J Bus Res 2022 Sep;148:174-189. [doi: 10.1016/j.jbusres.2022.04.034]
29. Li H, Wu J, Gao Y, Shi Y. Examining individuals’ adoption of healthcare wearable devices: an empirical study from privacy calculus perspective. Int J Med Inform 2016 Apr;88(8-17):8-17. [doi: 10.1016/j.ijmedinf.2015.12.010] [Medline: 26878757]

XSL•FO
RenderX

30. Petronio S, Child JT. Conceptualization and operationalization: utility of communication privacy management theory. Curr Opin Psychol 2020 Feb;31:76-82. [doi: 10.1016/j.copsyc.2019.08.009] [Medline: 31526974]
31. Gilster P. Digital Literacy: John Wiley & Sons, Inc; 1997.
32. Bonfadelli H. Literacy and trust as influencing factors of health communication online. : IntechOpen; 2023 Feb 8. [doi: 10.5772/intechopen.105883]
33. Pendl D, Maitz KM, Gasteiger-Klicpera B. Examining the relationship between health literacy and individual and sociodemographic factors in secondary school students. J Public Health (Berl) 2024 Mar;32(3):531-542. [doi: 10.1007/s10389-023-01836-1]
34. Rahdar S, Montazeri M, Mirzaee M, Ahmadian L. The relationship between e-health literacy and information technology acceptance, and the willingness to share personal and health information among pregnant women. Int J Med Inform 2023 Oct;178:105203. [doi: 10.1016/j.ijmedinf.2023.105203]
35. Cherif E, Bezaz N, Mzoughi M. Do personal health concerns and trust in healthcare providers mitigate privacy concerns? Effects on patients' intention to share personal health data on electronic health records. Soc Sci Med 2021 Aug;283:114146. [doi: 10.1016/j.socscimed.2021.114146] [Medline: 34242891]
36. Li Y. Theories in online information privacy research: a critical review and an integrated framework. Decis Support Syst 2012 Dec;54(1):471-481. [doi: 10.1016/j.dss.2012.06.010]
37. Richter G, Krawczak M. How to elucidate consent-free research use of medical data: a case for "health data literacy". JMIR Med Inform 2024 Jun 18;12:e51350. [doi: 10.2196/51350] [Medline: 38889087]
38. Cosgrove J, Cachia R. DigComp 3.0: European Digital Competence Framework, Fifth Edition: Publications Office of the European Union; 2025. [doi: 10.2760/0001149]
39. Hu J, Tan C, Jiao Y, Mou X, Wu P. The associational relationships between multidimensional digital literacy and wage income: evidence from urban–rural heterogeneity. Sci Rep 2025;15(1):42830. [doi: 10.1038/s41598-025-28467-8]
40. Zakir S, Hoque ME, Susanto P, et al. Digital literacy and academic performance: the mediating roles of digital informal learning, self-efficacy, and students' digital competence. Front Educ 2025;10:2025. [doi: 10.3389/feduc.2025.1590274]
41. Yang F, Yao R, Ren Y, Guo L. Harmony in diversity: digital literacy research in a multidisciplinary landscape. Comput Educ 2025 Jun;230:105265. [doi: 10.1016/j.compedu.2025.105265]
42. Reichert F, Pan Q, Chen LL. Digital literacy assessment. : UNESCO; 2023 URL: https://unesdoc.unesco.org/ark:/48223/pf0000386202 [accessed 2026-05-12]
43. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Model 1999 Jan;6(1):1-55. [doi: 10.1080/10705519909540118]
44. Kline RB. Principles and Practice of Structural Equation Modeling, 4th edition: Guilford Press; 2016. URL: https://eli.johogo.com/Class/CCU/SEM/_Principles%20and%20Practice%20of%20Structural%20Equation%20Modeling_Kline.pdf [accessed 2026-05-12]
45. Naeem I, Quan H, Singh S, et al. Factors associated with willingness to share health information: rapid review. JMIR Hum Factors 2022 Feb 9;9(1):e20702. [doi: 10.2196/20702] [Medline: 35138263]
46. Hu Y, Qian Y. Who is concerned about digitalization? The role of digital literacy and exposure across 30 countries. Inf, Commun Soc 2025:1-19. [doi: 10.1080/1369118X.2025.2592771]
47. Fox G, van der Werff L, Rosati P, Lynn T. Investigating citizens' acceptance of contact tracing apps: quantitative study of the role of trust and privacy. JMIR mHealth uHealth 2024 Jan 18;12:e48700. [doi: 10.2196/48700] [Medline: 38085914]
48. Meier DY, Barthelmess P, Sun W, Liberatore F. Wearable technology acceptance in health care based on national culture differences: cross-country analysis between Chinese and Swiss consumers. J Med Internet Res 2020 Oct 22;22(10):e18801. [doi: 10.2196/18801] [Medline: 33090108]
49. von Kalckreuth N, Feufel MA. Disease characteristics influence the privacy calculus to adopt electronic health records: a survey study in Germany. Digit Health 2024;10:20552076241274245. [doi: 10.1177/20552076241274245] [Medline: 39247096]
50. Omoloja A, Vundavalli S. Patient generated health data: benefits and challenges. Curr Probl Pediatr Adolesc Health Care 2021 Nov;51(11):101103. [doi: 10.1016/j.cppeds.2021.101103] [Medline: 34799255]
51. Use, understand & engage: a digital media literacy framework for canadian schools. Media Smarts. URL: https://mediasmarts.ca/teacher-resources/use-understand-engage-digital-media-literacy-framework-canadian-schools [accessed 2026-05-12]
52. What is the AI service user policy archive? Artificial Intelligence Service User Policy Archive. URL: https://user-archive.kisdi.re.kr/kor/introduce/bunesIntroduce.html [accessed 2026-05-13]

Abbreviations

AG: age group
AVE: average variance extracted
CFI: comparative fit index
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
CMV: common method variance

XSL•FO
RenderX

CPM: communication privacy management
CR: composite reliability
DD : digital literacy of understanding
DE: digital literacy of engagement
DL: digital literacy
DU: digital literacy of use
DWLS: Diagonally Weighted Least Squares
eHEALS: eHealth Literacy Scale
HI: household income
HTMT: heterotrait-monotrait
IC: information control
KISDI: Korea Information Society Development Institute
KPSDS: Korea Panel Survey on the Digital Society
MB : monetary benefit
MM: moral motivation
PB : perceived benefit
PE: perceived effectiveness of government regulation
PGHD : patient-generated health data
PR : perceived risk
RA: residential area
RMSEA: root mean square error of approximation
SL: perceived standard of living
SRMR: standardized root mean square residual
std β: standardized coefficient
std all: Fully standardized estimates
TLI: Tucker-Lewis index
TR: institutional trust
WLSMV: weighted least squares mean and variance adjusted
WS: willingness to share health data

Edited by A Coristine; submitted 20.Apr.2025; peer-reviewed by A Lakdawala, NV Kalckreuth; revised version received 27.Apr.2026; accepted 28.Apr.2026; published 27.May.2026.

Please cite as:
Lee D, Jang W
Digital Literacy and Interpersonal Trust as Predictors of Willingness to Share Patient-Generated Health Data Among Korean Internet Users: Cross-Sectional Study Using Privacy Calculus and Communication Privacy Management Theories
J Med Internet Res 2026;28:e75448
URL: https://www.jmir.org/2026/1/e75448
doi:10.2196/75448

© Dongsu Lee, Wonseuk Jang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Quality Challenges in Municipal Telecare Call Center Services: Qualitative Evaluation Using the Anchored, Realistic, Cocreated, Human, Integrated, and Evaluated (ARCHIE) Framework

Linda C Grøndal-Eeles*; Janne Dugstad*, PhD; Hilde Eide*, PhD; Etty Nilsen*, PhD

Centre for Health and Technology, Faculty of Health and Social Science, University of South-Eastern Norway, Postboks 4, Borre, Norway

*all authors contributed equally

Corresponding Author:
Linda C Grøndal-Eeles
Centre for Health and Technology, Faculty of Health and Social Science, University of South-Eastern Norway, Postboks 4, Borre, Norway

Abstract

Background: Telecare is seen as a promising technology aimed at enhancing the accessibility and efficiency of health care services. Although focus on quality has been highly prioritized within the health care services, there is a need to explore the quality of telecare services in general and municipal telecare call centers (CCs) in particular, as health and assistive technologies are increasingly being implemented in patients' homes.

Objective: The study sought to explore which factors influence the quality of telecare services provided by municipal telecare CCs in Norway, evaluated through the anchored, realistic, cocreated, human, integrated, and evaluated (ARCHIE) framework.

Methods: The study had a multiple-case design. Interviews were the main source of data from 15 informants from 5 municipal telecare CCs across Norway. Observation and document studies were used for background and contextualization. To explore and evaluate quality, a combined deductive–inductive analysis was conducted.

Results: Evaluated against the ARCHIE framework, none of the quality criteria were fully met. Due to the telecare service not being sufficiently anchored for all patients, it was challenging to provide realistic technologies. The collaborative work was difficult, with challenges in recruiting patients. The human principle was characterized by variation of knowledge and national guidelines. Municipal telecare CCs were not integrated into the health care services, and data must be used to a greater extent for evaluation and learning than is currently the case.

Conclusions: The findings suggest that municipal telecare CC services have several shortcomings in providing high-quality health care. Relating the quality principles identified by the ARCHIE framework to normalization process theory constructs indicates that the CC service remains in a transitional phase of normalization. To improve the telecare CC services and enhance communication and integration, policymakers need to reduce fragmentation in the broader health care system. Further national standardization to professionalize the telecare CC services should be developed. The telecare CCs need to improve their service related to all indicators of the ARCHIE framework. Training for telecare operators should be prioritized.

(J Med Internet Res 2026;28:e76054) doi:10.2196/76054

KEYWORDS

quality; municipalities; telecare services; telecare call center; assistive technology; patient safety; telecare; call center

Introduction

Telecare and digital technologies have been partly integrated into the health care services of Western countries during the past decade. Telecare can be defined as: "the use of information, communication, and monitoring technologies which allows health care providers to remotely evaluate health status, give educational intervention, or deliver health and social care to patients in their homes" [1]. A call center (CC) is a 24/7 service that operates telecare. A CC employs operators, typically health care workers or technicians, to answer, document, and assess calls and alarms from assistive technology [2]. Assistive technology is an umbrella term for assistive products and their related systems and services, which help maintain or improve an individual's functioning related to cognition, communication, hearing, mobility, self-care, and vision [3,4]. These technologies are implemented in people's homes and are the reason for the CCs' existence. CCs organize their service in a variety of ways, including their core service, which is to answer, evaluate, and document assistive alarms and take any necessary further action [2]. They also provide additional services, including ambulant teams, advisory departments, or monitoring technologies that are not health-related. The CCs may be private entities or part of the public service. In Norway, the CCs are largely organized

XSL•FO
RenderX

as intermunicipal collaborations that are hosted by one municipality and regulated by the Local Government Act [5]. Our recent article investigating the organization of municipal telecare CCs in Norway found that the centers were in an early development phase and had different organizational models, leading to a variation of delivered services [6]. As CCs are considered new and innovative within the health care service and provide the service in a new way, there is a need to investigate the quality of their services.

The World Health Organization (WHO) defines quality as "the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge" [7]. To achieve high quality within health care services, several factors, such as patient safety, effectiveness, availability, and patient-centered care, need to be evaluated [8], and many governments have strengthened their focus on quality within the health care system [9-12]. International studies have shown a multitude of adverse events happening within the health care system each year, which have an effect on quality [13-16].

Several frameworks have been developed for the evaluation of the quality of health care services. One of the earliest and most used is the Donabedian framework [17], which evaluates the systems, processes, and outcomes of health care delivery. The framework has also been used for evaluating quality in digital services such as telecare [18]. Tossaint-Schoenmakers et al's [18] review highlights 3 broad aspects that are important for integration: the needs of the care recipient, technology that is attuned to the organization and care processes, and human resources that are aligned with the desired end results. A need for more in-depth studies into the specific processes in different parts of the health care system related to eHealth is suggested.

Quality in telecare has increasingly become a relevant topic, as more assistive technology is being implemented. The aim of implementing technology is to improve access to the health care system and improve effectiveness [19]. The use of telecare can improve communication between health care professionals as well as improve patient safety by reducing errors [20,21]. Furthermore, digital health care can give patients the opportunity to participate in their own treatment and treatment plans to a higher degree than before [19]. Challenges and barriers related to the use of telecare include technical barriers, such as low internet access, poor digital competence among certain user groups, and concerns about privacy and data security [22]. In addition, Woolham et al [23] mention cost-effectiveness, assessment and deployment, ethical implications, and user involvement as key challenges.

As there is a need to explore in greater depth how quality is performed [18], Greenhalgh et al's [24] framework, called the anchored, realistic, cocreated, human, integrated, and evaluated (ARCHIE) framework, seems to be a relevant framework to evaluate important quality aspects of telecare. The framework determines whether a service is anchored, realistic, co-creative, human, integrated, and evaluated, constituting the acronym ARCHIE. The framework was based on interviews and observations in various care contexts to identify barriers and suggest several solutions to provide quality care with useful assistive technology. An important takeaway from Greenhalgh et al's [24] study is that technological advances should be underpinned by a user-centered approach to design and delivery. The care contracts and technology should be personalized for each patient, from branded products to interoperable components. Furthermore, the technology should have the ability to be used flexibly across devices and platforms [24].

The benefits of telecare are still not well-established [25,26]. If the service lacks quality, this can lead to an increased burden of illness, increased economic costs, and medical errors such as inaccurate diagnoses, medication errors, and inappropriate treatments [27,28]. If clinical guidelines are not followed, this can result in ineffective care. By increasing quality within the telecare service, the benefits of the implemented technology might be easier to recognize.

Patient safety in primary care is a critical aspect of health care. According to the WHO, patient safety involves the prevention of errors and adverse effects associated with health care, aiming to "do no harm" to patients [29]. Research indicates that incidents in primary care, although often less severe than those in hospitals, significantly impact overall patient safety due to the high volume of patient interactions [14]. Furthermore, it is important to take into consideration that patients will increasingly receive complicated treatment of complex illnesses at home. As previous research has highlighted, the assessment of quality and safety culture is a fundamental step in understanding health care providers' safety-related perceptions [14,30]. Therefore, fostering a robust safety culture and implementing targeted interventions for quality are paramount in primary care.

Telecare has the ability to enhance the accessibility and efficiency of health care services [31]. However, a systematic review identified multiple patient safety risks associated with telecare, including issues related to the nature of homecare tasks, person-centered characteristics, technology and devices, organization and environmental factors [32]. These risks can arise from the complexity of integrating technology into home care settings, where both the physical environment and the patient's capabilities play crucial roles. A qualitative study highlighted the importance of care workers' roles in ensuring patient safety through telecare [33]. The study identified 3 essential work processes: aligning people and technologies, being alert and staying calm, and coordinating activities based on people and technology [33]. These processes underscore the need for effective training and support for care workers to manage the interface between technology and patient care effectively.

Few studies have investigated telecare providers' perspectives on how quality is performed in municipal telecare CC practice. Therefore, the aim of the study was to explore which factors influence the quality of telecare services provided by municipal telecare CCs in Norway, evaluated through the ARCHIE framework.

XSL•FO
RenderX

Methods

Study Design

This study had a multiple-case design. A multiple case study is regarded by many as more robust than the single-case design [34-37]. In this study, similar cases have been studied in order to gain a broad understanding of quality, and not to compare the individual cases. The various cases form a bricolage [38] that embraces the complexity of the context. Interviews were the main source of data. Data from observations and documents have been used to enhance the understanding of the background and context of the CCs.

Recruitment and Data Collection

Data were collected in May and June of 2022. At the time of data collection, no overview of Norwegian telecare CCs existed, so a snowball-inspired strategy was used in our networks to identify centers. An email was sent to the managers of 5 CCs, and they agreed to participate with their leaders and personnel.

A semistructured interview guide was developed for the in-depth interviews. The study's aim was used to develop themes for the interview guide where the following themes were covered: (1) The participants' understanding of the municipal telecare CC service in their context, (2) How the municipal telecare CC services operated, (3) Cooperation with other health care services, (4) Ethical challenges, (5) Competence requirements and competence development, and (6) Perspectives on the further development of the municipal telecare CC services, to shed light on the aim. The idea was to let the informants speak as freely as possible within the themes. One paper has previously been published based on the same in-depth interviews, focusing on the organization of the services [11]. For this paper, the interviews were analyzed focusing on quality.

The first author, LGE, visited 4 of the CCs, with each visit lasting for one day, including interviews, a tour, and observations. On arrival at the CCs, a guided tour was conducted. The observations included how the CC operators worked, how they responded to alarms, the layout of the facilities, the workstation setup, and some of the technology included in the services (eg, social alarms, GPS, medicine dispensers, door locks, and sensors). These observations provided important background knowledge for understanding the service and supporting the interviews. Notes were taken and used to understand the context.

All the interviews were performed by LGE following the semistructured interview guide, which constituted the main source of data for the studies. As the interview guide contained broad, open questions, the informants could speak as freely as they wanted. Four of the 5 CCs were visited, while one CC had to be interviewed digitally due to the COVID-19 pandemic restrictions. Ten informants were interviewed face-to-face. Two of these were conducted with 2 informants present, and the rest were conducted one-on-one. Five informants were interviewed digitally—one with 3 informants present, and 2 were conducted one-on-one. The informants were present at the CCs during the interviews, and this was also the case during the digital interviews.

The interviews lasted from 45 minutes to 2 hours and were audio-recorded and transcribed.

Data Analysis

The data analysis had a combined deductive–inductive approach, analyzing the CCs as the unit of analysis. First, text segments were identified and sorted in accordance with the specific principles in the ARCHIE framework in a deductive fashion. Second, the text under each principle was analyzed inductively.

Deductive Analysis Using the ARCHIE Framework

The ARCHIE framework was designed to be adaptable and to guide the development and evaluation of telehealth and telecare services in various contexts [24]. The framework aims to operationalize quality in telehealth and telecare in order to improve the proportion of patients receiving appropriate, acceptable, and workable technologies and services and has 6 principles (Table 1).

Table . Overview of the ARCHIE (Anchored, Realistic, Cocreated, Human, Integrated, and Evaluated) framework [24].

Principle	Content
Anchored	Having a shared understanding of what matters to the patient, which includes design and development of technology
Realistic	Understanding the natural history of illness. The illness may progress, and the patients' needs change
Co-creative	Evolving and adapting solutions with patients. This should be continuous rather than standardized
Human	Supporting patients through interpersonal relationships and social networks where knowledge and information are key
Integrated	Developing mutual awareness and sharing knowledge
Evaluated	Ensuring system learning from experiences built over years. Ongoing innovation to increase quality

- Step 1: The first author (LGE) transcribed the 11 interviews and initially read through all of them, identifying and coding text segments that aligned with the 6 main ARCHIE principles described above. Some text segments aligned with the ideal quality principle, but others were inconsistent; these inconsistent statements were coded under the specific principle as "not" (Table 2 contains examples). Consequently, all the coded text had a main code in

XSL•FO
RenderX

accordance with ARCHIE, and if relevant, a subcode as "not."

Table . Examples of coding that align with the ARCHIE (Anchored, Realistic, Cocreated, Human, Integrated, and Evaluated) principles and which are inconsistent with the ARCHIE principles.

Principle	Example
Anchored	A shared understanding of the needs of the patient
Not anchored	A patient is given the wrong technology
Realistic	Understanding the natural history of illness
Not realistic	Change of pattern for the use of assistive technology, but the patient still has the technology

- Step 2: All authors read and coded 3 selected interviews to align the understanding of the ARCHIE framework principles and secure the initial coding in each principle. The tentative coding by the first author in Step 1 was discussed, adjusted, and agreed on.
- Step 3: LGE reread and checked the coding of all the interviews.

Inductive Analysis

After the deductive analysis, a second phase followed using inductive content analysis within each ARCHIE principle. The inductive analysis followed the structured process described by Elo and Kyngäs [39], with 3 main steps: preparation, organization, and reporting.

- Step 4: In the preparation phase, the text within each ARCHIE principle was re-examined by LGE to identify meaning relevant to the principle's theme. These units were extracted and organized for a deeper analysis.
- Step 5: During the organization phase, the author group's focus was on identifying patterns, contrasts, and recurring themes within each principle, rather than generating new codes. For example, in the anchored principle, both aligned and misaligned statements were examined to understand how services reflected, or failed to reflect, what mattered to the patient. This was discussed in the author group to ensure understanding of each principle.
- Step 6: The reporting phase synthesized insights from each principle to highlight trends and tensions. These were then interpreted in relation to the ARCHIE framework to assess whether and how quality was met across the principles. NVivo (version 5.0; Lumivero) was used in the coding processes.

Rigor and Trustworthiness

To ensure rigor and trustworthiness in the study, criteria developed by Guba and Lincoln [40] were followed, including credibility, transferability, dependability, and confirmability. To ensure the credibility of the coding, analysis, and results were discussed in several rounds, in the deductive as well as the inductive phase of the study, to ensure the group of authors could express their thoughts, concerns, or other factors. One issue that the author group discussed in several rounds was how to interpret the content of the principles in ARCHIE, and which statements belonged where. To understand the principles, the author group went through each statement, forming a consensus on whether or not statements were placed correctly in our understanding of what each principle should contain. The study provided detailed descriptions of the sample to ensure transferability and support contextual understanding. The coding process was well-documented and logged in NVivo, as well as revisions to maintain dependability. In any research, the expulsion of all biases is challenging. However, the research group consisted of authors with interdisciplinary backgrounds, such as nursing (LGE and HE), economics and management (EN), optometry (JD), psychology, and eHealth communication and counseling (HE). The group also had broad competence in research methodology as well as health technology implementation. This meant that the authors could share different views on the research and ask critical questions to make sure the author group's assumptions and interpretations were as transparent as possible, thereby ensuring confirmability. Furthermore, the coding process was repetitive and reflexive, guided by Saldaña's [41] approach, and involved multiple rounds of discussions within the author group to reach consensus. Analytical transparency was maintained through NVivo documentation.

Ethical Considerations

The study was approved by the Norwegian Agency for Shared Services in Education and Research, which assessed the data processing plan to be in accordance with data protection legislation (Ethics approval number: 426173). Furthermore, the main principles for research according to research ethics in the social sciences and the humanities [42] were followed. All informants received written and oral information about the study, and informed consent to participate was obtained from each informant. The informants participated voluntarily and received no compensation of any sort. The CC informants appreciated that researchers took an interest in the telecare topic and therefore welcomed the study. Data were stored and secured in a USN-protected area. The data in the quotes used throughout this article were anonymized.

Results

Overview

The first section of the results will present the sample of the study, and key dimensions of quality according to the ARCHIE framework will then be presented. The headings represent the main findings in each principle.

Sample

Data were drawn from 5 municipal telecare CCs in different parts of Norway (Table 3). Eleven interviews were performed

XSL•FO
RenderX

with 15 professionals working in the 5 CCs (CC1-CC5). Five informants from CC1 were interviewed, 1 from CC2, 4 from CC3, 2 from CC4, and 3 from CC5. In total, 12 women and 3 men were included. The informants had a variety of professional backgrounds and roles within the centers: 6 were managers, 4 were advisors, 4 were health care operators, and 1 was an IT technician.

Table . Overview of call centers (CCs) participating in the study.

	CC1	CC2	CC3	CC4	CC5
Number of municipalities in cooperation	14	16	16	6	12
Number of citizens in the cooperating municipalities	Estimated 140,000	Estimated 175,000 (Estimated by authors)[a]	Estimated 306,000 (Estimated by authors)[a]	Estimated 51,000	Estimated 85,000
Established	2019	2003 serving fire and rescue. Included in the health care service from 2019	2018	2019	2016 under fire and rescue service, included in health care 2023, colocated with municipal accident and emergency department.
Organization (departments or services provided alongside telecare CC)	Colocated with municipal accident and emergency department	Monitoring technical alarms in public buildings (fire alarms, lift alarms, etc)	Ambulatory team department. (Can visit patients whose alarms cannot be resolved by phone)	Service design advisory department. (Can assist municipalities in cooperation with implementation strategy and other eHealth challenges	Monitoring technical alarms in public buildings (fire alarms, lift alarms, etc)

[a]Estimated by authors.

The Unanchored Services of Call Centers

The anchored principle was meant to reflect a shared understanding of the patient's needs between the patient and all the stakeholders involved in the provision of the service, emphasizing what is important and meaningful to the patient. However, the allocation of assistive technology did not appear to be anchored with all the stakeholders working with each patient. The telecare providers working in the CCs were not involved in the process of assessing the patients' needs or prioritizing and selecting what services they should receive. Informant K said:

> *The assumption must be that the home nursing staff, in collaboration with other decision-makers in the municipalities, should work much closer with the patient to see if the technology can actually relieve the burden. It is felt that this is not being done.*

The municipal telecare CCs provided services across several municipalities, which otherwise worked independently of each other. In order to establish and provide the service for a patient over time, the telecare services included several stakeholders within the health care services of each municipality, who interacted with the patient during the course of the service. Initially, the home nursing unit and the service allocation office were involved in the referral and mapping of the patients and the assignment of the service. In addition, general practitioners, family members, and even the patients themselves could identify the need for a service in the early phases. When a patient was allocated a telecare service, the health technology support team was responsible for installing the technological devices in the patient's home and instructing the patient in the use of the technology. The CC was then notified. Their responsibility was to respond to alarms and calls and to monitor the technology and incoming data. In their daily operations, the CC alerted the home nursing staff if physical interaction with the patients was needed. In order to change the service for a patient, the CCs or home nursing staff reported their concern to the service allocation office. A suspicion that the service was perhaps no longer what the patient needed would prompt an evaluation and the allocation of new services. Figure 1 displays a visualization of the process with an overview of the relevant stakeholders involved in the different phases of assigning and providing a telecare service.

XSL•FO
RenderX

Figure 1. Overview of the actions related to the telecare call center service and stakeholders within primary health telecare.

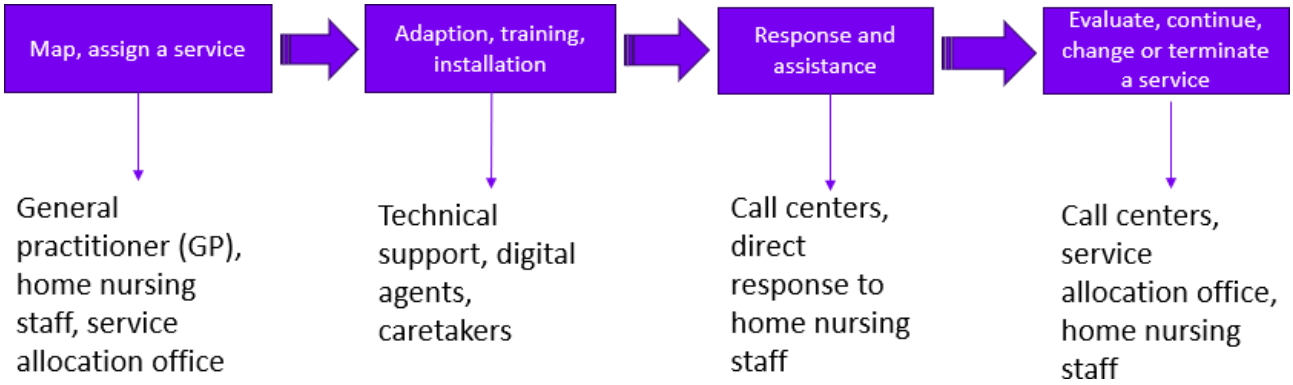

The Challenge of Having Realistic Technology When the Needs Are Constantly Changing

The realistic principle meant that technology and care packages needed to allow for the patients' conditions to progress or change, thereby supporting changes in the patients' needs. However, if the technology was not anchored in the first place, it could never be realistic. The informants addressed how patients received the technology too late, where the situation of the patients was rapidly changing, and the technology was therefore not realistic. Informant K explained:

> *We feel like the introduction to health technology should happen at a much earlier stage than what is being done now.*

Another challenge was how some progressive illnesses left the patient in a state where they did not know, understand, or remember the different assistive technologies, indicating that their health situation had changed, so they were no longer capable of using the technology as intended. When a patient reached this stage, there was no benefit for either the patient or the health care service, because the assistive technology was a poor fit. As informant F stated:

> *When a patient has moderate Alzheimer's, then it is too late. They will wonder what it is and maybe even take it off and not remember to use the alarm when needed.*

The CCs could pick up on the progression of a patient's health conditions by monitoring patterns in their behavior, such as a sudden increase in calls from a patient via their social alarm. The CC reported these changes to the home nursing staff or service allocation office. Informant B said:

> *We have two patients that trigger their alarms hundreds of times. Then we can say, this user has the wrong service.*

However, the CCs found that in spite of their reports, the patients continued to have assistive technology that did not work for them. This indicated that the municipalities were unsuccessful in re-evaluating whether a patient had the right technology at the right time, and also that the feedback from the CC was not taken into consideration.

Challenges in Co-Creative and Collaborative Work

The co-creative principle meant that the technology should be continuously cocreated between the service stakeholders and the patients, making it tailor-made to a certain degree. Co-creation and interaction across disciplines and levels were raised as challenges by several informants. Lack of cooperation and the sharing of information and knowledge between CCs and other institutions was a challenge in general, and specifically on a user-related level. Informant F said:

> *We see a need for cooperation, on a larger scale, with GPs and hospitals. But we also see a great need to interact better with the patients to be able to give them a better service.*

The focus of the CC providers was to cooperate and interact with patients. However, informants explained how it was a challenge to recruit patients to cooperate and test the technology. The CCs tested all the technology, and some even created a test lab where patients, future patients, and their next of kin could come and gain hands-on experience of the technologies. Informant K said:

> *We have emphasised a holistic focus. We see it as a core value. Because it is about bringing people with us the whole way, from development to implementation.*

None of the telecare providers involved in this study actually developed health technology. They primarily managed calls/alarms from technology purchased by the municipalities. The technologies had been purchased based on the needs of the patients in each municipality. However, some of the CCs provided recommendations or came up with suggestions to the municipalities about what technologies to procure. In some instances, they even procured technology on behalf of the municipalities. However, it was explained that patients' needs determined what was to be procured, not the technology in itself. Informant K explained:

> *We must understand the needs to solve a problem, then we can go out and see – is there any technology that can work?*

Telecare providers had a unique insight into how the home nursing staff coped with the technology, whether they managed the technology, and whether the patients could handle the technology. Informant B said:

> *We get the home nursing perspective, on what is difficult technology. Because it can be completely different for an engineer who has developed it and it*

XSL•FO
RenderX

makes sense to them, but maybe not to an assistant in the home nursing team.

The Importance of Knowledge for Nurses

The human principle comprised the interaction between patients and employees' knowledge and key information. One CC received 300,000 alarms in 2021, with 861 events in 24 hours and 36 per hour, which underlined how much interaction the CC service had with its patients. The telecare providers frequently talked to the patients and got to know many of them. Having information and personal knowledge about some of their patients led to more calls being resolved by the CC service—hence, the alarm did not need to be escalated to the home nursing team. Informant P stated:

> *Because in X [a private call centre] they just sent the call through to home nursing, but here [municipal call centre] they answer, and ask questions. "I have lost something under the sofa." A call like this will not go out to the home care team.*

As a result of the high number of alarms, an ethical dilemma emerged regarding frequent callers. The patients labeled as frequent callers could lead to employees in the CC service feeling that they were getting tired of the patients and, therefore, needed to work to ensure that they upheld their standards and the quality of their communication. The telecare providers said that they managed it because they knew that these patients' needs were not being met. Some reflected that the reason for the frequent calls might be the fact that the patient had the wrong technology. Informant D said:

> *I have experienced that I get tired of alarms, especially when the alarm is triggered for "nothing."*

Findings highlighted a variety of health personnel working in the CCs. Some of the CCs had nurses, some had health care workers only, and some had technicians working together with nurses as operators in the CC. Some CCs had a standard minimum requirement for at least one nurse, whereas others, which only had nurses, were looking to hire health care workers. Informant P said:

> *We have made it a requirement that a nurse must be present at work at all times.*

Not only did the professional backgrounds of the operators vary, but their level of experience also seemed to differ. One of the CCs was colocated with an accident and emergency department. The nurses had a high level of competence in handling telecare and solving acute matters over the phone. Informant A said:

> *We only have nurses with very high competence answering acute calls. Because that happens every week, and they need to be able to handle it.*

Employees' experience seemed to be key. The interviews and observations revealed that there were no national decision-making tools, procedures for the receipt of alarms, questionnaires, or standards. One nurse (Informant C) working for a CC colocated with an accident and emergency department said:

> *At the accident and emergency department we use triage systems [decision-making tools], but at the call centre we have nothing. There are no procedures as to what we should do.*

Where there were no decision-making tools or procedures for how to receive and handle a call, it could elevate the risk of misunderstandings, as there was no "safety net" to ensure the operators had asked about everything that was relevant when handling an alarm. One informant explained an episode where the operator did not inquire thoroughly enough during an alarm that ended with a patient lying on the floor for a long time, while the operator thought that the patient was lying in bed. Another CC, which monitored a wide variety of technology-based services, including lift alarms, fire alarms, burglary alarms, and an after-hours answering service for the municipality, needed other skills for non-health-related alarms. Informant K explained:

> *We have two operators on duty – one with a health background and one with a technical background. This composition is because of the portfolio of services we have.*

However, all alarms had the same ringtone in all the CCs. In the CC where both nurses and technicians were working alongside each other, they answered alarms without any system or pattern as to which one of them was answering which alarm. The lack of alarm identification meant that the technicians answered medical alarms, and vice versa. They would assist each other if one of them, for example, answered a cardiac arrest. Informant C stated:

> *If a fire alarm has gone off in an apartment, we should act right away, but if an alarm box has a low battery, we have 12 hours. But we can't tell the difference between the two when the alarm goes off [in the call centre].*

The study found no evidence to show that there were more adverse events in CCs with health care workers than with experienced nurses. This is because there was no indication of any system for registering adverse events, or opportunities to look into whether CCs with different professional employee profiles had more adverse events than others.

Challenges With Partly Integrated Technology

The integrated principle meant that both technology and work processes should be integrated with existing systems. Even though the CCs reported that it was demanding to cope with all the technical integrations, it seemed that this was a very high priority for all of them. They also emphasized the importance of the integrations. Informant A explained:

> *Because of the integrations, we can clarify many alarms that come in.*

All the CCs had patients' medical records integrated. This meant that the telecare providers had access to information about the patients, such as when the home nursing team visited the patients. When operators answered alarms, a notification came up on the screens, as well as an alarm or ringtone in the CC. The operator could see who was contacting them, and their medical record came up on their screens. Furthermore, it seemed that the technology used by the patients was integrated into the platforms of the CCs. However, the CCs explained how it could

XSL•FO
RenderX

take time and be challenging to integrate new technology into the platforms. The CCs had many vendors to relate to and many technologies that needed to be integrated into their systems. On top of that, new technologies were making their way into the service, which also required integration.

As for collaboration and the integration of work processes, the CCs seemed to closely cooperate with the home nursing units in the different municipalities where they provided services. However, the analysis indicated an "us/them" attitude between the CC providers versus the more administrative service allocation office in the different municipalities, suggesting a lack of cooperation between the organizational units involved. Informant K said:

> *I don't know what to say about the administration – it is not that they are falling behind, but the cooperation, it has not worked so well.*

Better Use of Data for Evaluation and Learning

The evaluated principle indicated learning from experience and improving systems accordingly. It should involve an ongoing effort to increase quality. The CCs had a lot of data from their operations, and the municipalities received reports from the CCs regarding this data on a regular basis. This data provided key information about the CC service and the patients and could be the basis for evaluation, learning, and development of the service. However, the CCs were not sure whether the municipalities used the data for learning purposes or if they used the data at all. The CCs monitored the number of alarms related to each of the collaborating municipalities and variations over the year and could provide data for individual patients, mapping whether the service was the right one for them. Informant B explained:

> *Creating value through better use of data is important for the municipalities … How many 'false alarms' did the home nursing have to take on before the call centre? When the alarm is sent now, it is real.*

The CCs were reporting to the municipalities because they were trying to see the effects of their work, making sure they procured the right technology, and getting it documented. Informant K explained:

> *We are reporting, we do it because we want to see the effects of the measures we implement. A call centre is not a quick fix, but thorough work from top to bottom or bottom to top.*

Furthermore, the CCs were evaluating if there were other services they could operate to make them more resilient. Many options were considered at the idea stage, but digital home monitoring was a trend that many of the CCs were looking at implementing as a service. Some CCs had already started. Informant L said:

> *Digital home monitoring and rehabilitation. We are investigating the possibilities.*

Discussion

Principal Findings

Using the ARCHIE framework to evaluate quality in municipal CC services reveals that, despite efforts, the service remains fragmented and therefore poses a potential threat to quality. The findings in this study appear not to reflect the quality principles in the ARCHIE framework and what is indicated as high-quality telecare. The framework promotes patient-centered care, whereas this study suggests that the patients' voices are barely heard, and it appears that consideration of what matters to the patients is lacking. This permeates throughout the following principles of the ARCHIE framework: anchored, realistic, cocreated, and poses challenges to the human principle.

In the following, the results will be discussed using the normalization process theory (NPT) [43]. NPT is a framework or set of tools concerned with understanding 3 main issues: implementation, embedding, and integration of new or modified practices, and these align closely with the challenges identified in this study. Using the ARCHIE principles, the study has researched the quality of municipal telecare CCs on a micro level [44]. To further specify how municipal telecare CCs can become part of normal digital service delivery on a meso level [44], NPT offers valuable insight and aids the assessment of how far municipal telecare CCs have come in terms of integration into routine practice, and what is missing to become a fully embedded, high-quality digital health service.

Lack of Quality Challenges Telecare Call Center Normalization

Overview

In this study, the ARCHIE framework [25] was chosen because it was specifically developed as a set of quality indicators for telecare services. ARCHIE emphasizes values such as person-centeredness, coproduction, and inclusivity, which are particularly relevant in the context of digital health and remote care [25]. ARCHIE provides a useful framework for assessing how telecare services align with the intended values of care delivery. Its design makes it especially suitable for evaluating the quality dimensions of telecare implementation, which appear to be less visible in more structural models such as Donabedian's [18,24]. Although foundational in health care quality assessment, the Donabedian model [18] focuses primarily on structure, process, and outcome and may not sufficiently capture the relational and participatory aspects of telecare services.

Of particular relevance in the NPT framework are the mechanisms that inhibit or promote the new practice, namely the following: (1) coherence, (2) cognitive participation, (3) collective action, and (4) reflexive monitoring [45]. In the following, each principle will be briefly presented. The results from the analysis of the relevant ARCHIE principles will then be discussed, including implications.

Figure 2 illustrates how the principles from the ARCHIE framework align with the NPT mechanisms and findings of this study.

XSL•FO
RenderX

Figure 2. Results of the ARCHIE (Anchored, Realistic, Cocreated, Human, Integrated, and Evaluated) analysis on the call center level related to a broader implementation framework, normalization process theory (NPT), at the meso level.

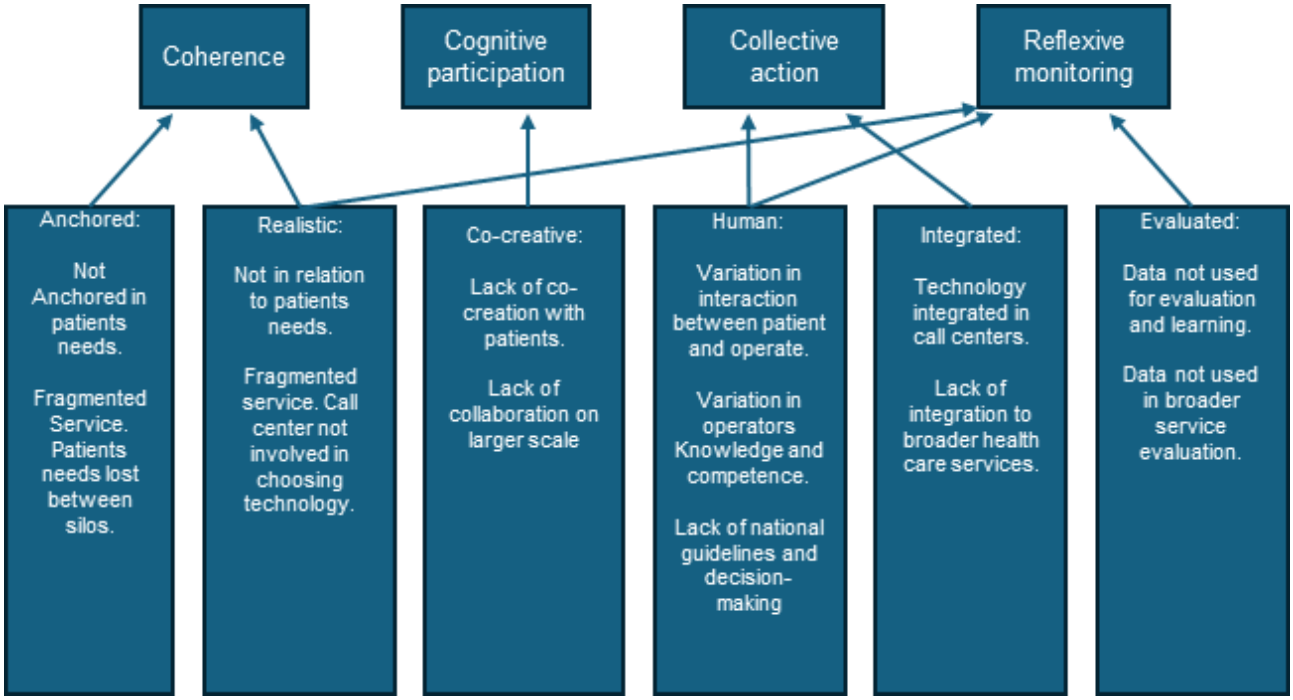

Coherence: Lack of Shared Understanding

On a meso level, the NPT construct Coherence is the process, sensemaking, and work to understand how a technology or service is implemented into practice, and also the routine for embedding a practice [45]. Coherence is regarded as a cognitive construct [43]. On a micro level, the ARCHIE principles anchored and realistic relate to coherence. Anchored links to the patient and to a shared understanding of the patient's needs, which correlates well with coherence. Realistic looks to technology, and whether the technology is realistic for the patient [24]. Although the realistic principle may relate to several NPT constructs, an intervention that is not meaningful will rarely be experienced as realistic. In the ARCHIE principle, realistic was linked to coherence.

Findings in this study reveal a lack of coherence in municipal telecare CC services, particularly in terms of a lack of shared understanding of the patients' needs. This could result in patients being assigned inappropriate technology. The lack of coherence becomes visible when seen through the principles of the ARCHIE framework and suggests that it is a system feature where one principle affects the other. If the decision by the service allocation office is not anchored, this will have a domino effect on the following principles, and they will not be right either. This shows how the telecare service is fragmented, and several silos are unsuccessful in working together to execute the service.

Dillon et al [46] found that when the health care system is perceived as fragmented, the outcome could be a loss of trust and difficulties navigating between providers. These results may be transferable to the telecare service. Furthermore, fragmentation may substantiate the "us/them" feeling, causing an even greater division between the stakeholders, as this study implies. Dillon et al [46] recommend a fundamental restructuring of the health care system to coordinate patient-centered care to meet the needs of the patients and to rebuild trust in the health care system. The ARCHIE framework also promotes patient-centered care, but results from this study show that the lack of tailored approaches may hinder quality in telecare and in the CC services, despite efforts by municipal telecare CCs to improve and professionalize their services.

Cognitive Participation: Uneven Stakeholder Engagement

The NPT construct Cognitive participation relates to the process that needs to be undertaken by individuals to achieve engagement for the implementation of a new practice on a meso level [43]. On a micro level, cognitive participation can be seen in relation to the ARCHIE principle of co-creation. This principle was not met, and the result can be divided into two. First, the lack of collaboration on a larger scale, with other health care services, and second, the lack of collaboration and co-creation with patients. Interestingly, the informants said that it was difficult to engage the patients in co-creation. Fusco et al [47] identified low health literacy, lack of understanding, time constraints, and intimidation of the co-creation process as barriers to successful co-creation. The reasons why the municipal telecare CCs in this study found it challenging to engage patients can only be speculated on, assuming that Fusco et al's [47] barriers also apply to municipal telecare CCs.

Collective Action: Fragmented Operational Routines

On a meso level, Collective action relates to the work individuals and organizations need to do to enact implementation [43]. On a micro level, the ARCHIE principles human and integrated align with collective action. The findings show that operational routines within municipal telecare CC services are fragmented, the staff receive limited training, and there is a lack of standardized protocols and integration with the broader health

XSL•FO
RenderX

services. These barriers hinder the enactment of quality-enhancing practices and suggest that municipal telecare CCs have not yet been fully normalized within municipal health care structures. Although key to telecare delivery, the municipal telecare CCs seem to operate in partial isolation from other stakeholders, which further complicates coordination and continuity of care.

The other ARCHIE principle closely linked to collective action is human. This study has mapped a variety of health professions working in municipal telecare CCs, spanning from highly experienced nurses working in both emergency rooms and CCs to technicians. Given that remote care is a different way of providing care, specific skills are required [48,49]. The CCs have to be prepared to handle emergency alarms, and findings indicate that the staff's experience level is key. Anker-Hansen and Johansen [50] found that the nurses ensure patient safety in complex remote decision-making in telecare. The complexity includes making accurate decisions based on unclear patient descriptions and challenges with communication [50]. The telecare nurses emphasized the importance of clinical expertise and the challenging nature of the decision-making processes and communication. They asserted that the provision of telecare requires that the staff working in CCs must have both proper training and experience, which is not the case in this study. Proper training and protocols can help to mitigate these potential patient safety risks [32].

Data from this study shows that miscommunication and adverse events have occurred. In Fisk et al's [48] study, the emphasis is on the need for a diverse range of skills and knowledge, which may include a mix of health care professionals. There is a lack of national systems, updated guidelines, procedures, and best practices for handling alarms to guide the operators working in the CC, which may negatively impact the quality of the service. Hughes [48] suggests that most adverse events result from faulty systems and processes rather than individual errors. This raises concerns about the impact of nonoptimized systems in telecare services.

In 2014, Guise et al [32] called for increased standardization in reporting adverse events to improve the knowledge base in telecare, but current research indicates that little progress has been made since. Conversely, there are critical voices expressing concerns about over-reliance on standardized guidelines and protocols. It is argued that standardization can lead to a lack of flexibility and creativity in addressing complex, real-world problems, and that these guidelines sometimes stifle critical thinking and innovation, leading to a more bureaucratic and less-effective health care system [49]. The importance of reflexivity and questioning established norms is stressed, and it is argued that a one-size-fits-all approach to guidelines can overlook the unique contexts and needs of different health care settings and patients. It seems that the absence of national guidelines or standards further contributes to the disparity of collective action being varied, making it difficult to foster collective ownership of telecare quality improvement.

Reflexive Monitoring: Limited Evaluation and Learning

Reflexive monitoring reflects the evaluation of a new practice and the effects the new practice has. This includes assessment of the advantages and disadvantages of the new practice [45]. On a micro level, reflexive monitoring aligns with several of the ARCHIE principles: realistic, human, and evaluated. This implies that both the ARCHIE framework and NPT are closely related to practice, gliding into each other, and can be challenging to separate. The ARCHIE principles of human and realistic have been discussed under other NPT constructs; however, they are also relevant on a meso level of reflexive monitoring because they emphasize continuous assessment and adaptation based on human experience and context.

In the ARCHIE framework, the insufficient and ineffective utilization of data falls under the evaluated principle [24]. The results of the study reveal that few municipal telecare CCs have mechanisms for evaluating their performance or for incorporating feedback from patients. This lack of systematic evaluation limits opportunities for learning and adaptation and contributes to persistent gaps in quality and safety. Without reflexive monitoring, it is difficult to identify what works, what needs improving, and how services can evolve to meet patient needs more effectively. The effective utilization of data in telecare services is crucial for enhancing patients' outcomes and service efficiency. When data is not coherently managed or used, it can lead to fragmented care, missed opportunities for service improvement, and reduced patient trust [51]. Inefficiencies such as repeated tests and procedures may also arise, increasing operational costs [52]. This concern is seen in the WHO's SMART Guidelines [53], which emphasize that digital technologies should be designed to support consistent, standards-based data use to improve care quality and coordination.

Regular audits and feedback mechanisms are also vital for continuous improvement [54]. Furthermore, Guise and Wiig [49] encourage training that includes patient-centered methodology to ensure good patient–professional relationships in remote care.

There are several challenges regarding the telecare service that need to be addressed on a meso level: lack of coherence, effective utilization of data, and outdated national guidelines. From a system perspective, health care is seen as a complex set of management levels, along with the policies, resources, and actors involved in delivering care [55], referring to the entire network of organizations, people, and actions whose primary intent is to promote, restore, or maintain health [56].

Strengths and Limitations

The ARCHIE framework was selected for this study because it was especially developed for defining quality in telecare. The framework has successfully assisted the research, uncovering quality and patient safety challenges within the telecare service, including municipal telecare CCs.

This study's combined deductive–inductive approach to the ARCHIE framework contributes to an enhanced understanding of the complex nature of the municipal telecare CC service.

However, when analyzing the data, we found that the different principles in ARCHIE were difficult to differentiate from one another and could be intertangled, meaning that the data could fit several of the principles. This was resolved by being precise

XSL•FO
RenderX

in choosing statements and through discussions within the author group, to ensure that the statements reflected how the group understood the principles.

The selection of the municipal telecare CCs was a potential limitation for the study. At the time of data collection, modest information was available on existing CCs, and the selection was therefore random for practical purposes. As we have learned in the process of writing this study, however, the CCs under study represent the most common way of organizing CCs in Norway. The ease of access to informants might have compromised the reliability and validity of the data [57]. Another limitation to be aware of is that of the cross-sectional snapshot, as data were collected at one point in time in this study. This may not represent long-term behaviors or trends.

The lack of national standards makes it challenging to benchmark quality, which affects both analysis and interpretation.

A strength of the research is the diverse background of the research group. This interdisciplinary background enabled the group to conduct an open and critical process, with the aim of reducing bias as much as possible.

Conclusions

The findings of this study suggest that telecare services in Norway are not yet fully normalized. The challenges identified across all 4 NPT constructs—conceptual ambiguity, uneven engagement, fragmented routines, and limited evaluation—indicate that telecare remains in a transitional phase of implementation. The application of the ARCHIE framework [24] helps to clarify the current position of telecare within the health care system and highlights specific areas where targeted interventions could accelerate normalization. Combined with NPT, which offers critical insights into the processes necessary for change at the meso level, this study provides a dual insight for understanding both the process and the quality dimensions of telecare. These insights are critical for policymakers, service designers, and practitioners aiming to improve the quality, safety, and integration of telecare services.

Implications for Policymaking

The findings of this study imply that municipal telecare CCs need to improve the telecare service to achieve the full potential of integrating a high level of quality and patient safety. Improvements should be considered at both meso and micro levels, as shown in this study. At a meso level, one way to advance is for stakeholders in telecare services to improve communication across their services and reduce fragmentation in the system. The telecare service calls for national standardization and the professionalization of the service. On a micro level, all of the ARCHIE principles need improvement. Training for telecare operators is one area that should be prioritized. As shown in this research, Greenhalgh et al's [24] ARCHIE framework can be a useful tool for assessing quality in emerging services within the health care domain.

Issues for Further Research

The following issues for further research to improve quality were identified in this study:

1. The patient perspective remains underresearched, as this study has focused on the municipal telecare CCs.
2. Cooperation and communication between stakeholders—in particular, how do the municipalities consider patients' needs for assistive technology, and how are concerns about ill-fitted assistive technology handled?
3. Competence development—what kind of training do telecare operators receive before they start as operators, and what plans are in place to maintain and increase knowledge?

Acknowledgments

All authors made significant contributions to the manuscript. The article was prepared in close collaboration with the authors LGE, EN, JD, and HE. The data were collected by LGE, but all authors contributed to the analysis and writing of the manuscript. Furthermore, all authors read and agreed on the final version of the manuscript. The authors would like to thank all the informants for participating in this study.

Funding

LGE is funded by the University of South-Eastern Norway. JD, EN, and HE were partly financed by the Research Council of Norway via the project Cocreated Health Technology (COTECH, project number 331903).

Data Availability

The data collected and analyzed for the previous and this study are not publicly available due to privacy reasons.

Conflicts of Interest

The first author (LGE) is employed by a private company, Tellu. The data collection and writing of this article were, however, performed before this employment began. The company did not, under any circumstances, influence data analysis, interpretation, or the writing of the manuscript.

References

XSL•FO
RenderX

1. Solli H, Bjørk IT, Hvalvik S, Hellesø R. Principle-based analysis of the concept of telecare. J Adv Nurs 2012 Dec;68(12):2802-2815. [doi: 10.1111/j.1365-2648.2012.06038.x] [Medline: 22607115]
2. Farshchian BA, Vilarinho T, Mikalsen M. From episodes to continuity of care: a study of a call center for supporting independent living. Comput Supported Coop Work 2017 Jun;26(3):309-343. [doi: 10.1007/s10606-017-9262-4]
3. Zallio M, Ohashi T. The evolution of assistive technology: a literature review of technology developments and applications. arXiv 2022. [doi: 10.48550/arXiv.2201.07152]
4. Assistive technology. World Health Organization. 2025. URL: https://www.who.int/news-room/fact-sheets/detail/assistive-technology [accessed 2025-09-16]
5. Ministry of Local Government and Regional Development. Act relating to municipalities and county authorities (the local government act). Lovdata. 2024. URL: https://lovdata.no/dokument/NLE/lov/2018-06-22-83 [accessed 2024-01-20]
6. Grøndal-Eeles L, Dugstad J, Eide H, Nilsen E. The organizational challenges of municipal call centers as a health service in Norway-a multiple case study. BMC Health Serv Res 2025 Jan 20;25(1):108. [doi: 10.1186/s12913-025-12264-0] [Medline: 39833817]
7. Quality of care: a process for making strategic choices in health systems. World Health Organization. 2006. URL: https://iris.who.int/handle/10665/43470 [accessed 2024-07-02]
8. Institute of Medicine (US) Committee on Quality of Health Care in America. Crossing the Quality Chasm: A New Health System for the 21st Century: National Academies Press (US); 2001. URL: http://www.ncbi.nlm.nih.gov/books/NBK222274/ [accessed 2024-07-02]
9. Campbell SM, Roland MO, Buetow SA. Defining quality of care. Soc Sci Med 2000 Dec;51(11):1611-1625. [doi: 10.1016/s0277-9536(00)00057-5] [Medline: 11072882]
10. H. Ministry of Health and Care Services. Meld st 11 (2020–2021). Regjeringen. 2024. URL: https://www.regjeringen.no/no/dokumenter/meld.-st.-11-20202021/id2791147/ [accessed 2024-01-25]
11. Public Health System Group for England. Quality in public health: a shared responsibility. Gov.UK. URL: https://www.gov.uk/government/publications/quality-in-public-health-a-shared-responsibility [accessed 2024-01-25]
12. Quentin W, Partanen VM, Brownwood I, Klazinga N. Measuring healthcare quality. In: Improving Healthcare Quality in Europe Characteristics, Effectiveness and Implementation of Different Strategies: Characteristics, Effectiveness and Implementation of Different Strategies: OECD Publishing; 2019.
13. Jha AK, Prasopa-Plaizier N, Larizgoitia I, Bates DW, Research Priority Setting Working Group of the WHO World Alliance for Patient Safety. Patient safety research: an overview of the global evidence. Qual Saf Health Care 2010 Feb;19(1):42-47. [doi: 10.1136/qshc.2008.029165] [Medline: 20172882]
14. Lawati M, Dennis S, Short SD, Abdulhadi NN. Patient safety and safety culture in primary health care: a systematic review. BMC Fam Pract 2018 Dec;19(1):104. [doi: 10.1186/s12875-018-0793-7]
15. Nieva VF, Sorra J. Safety culture assessment: a tool for improving patient safety in healthcare organizations. Qual Saf Health Care 2003 Dec;12 Suppl 2(Suppl 2):ii17-ii23. [doi: 10.1136/qhc.12.suppl_2.ii17] [Medline: 14645891]
16. Stokke R, Furnes M, Vifladt A, Ringsby Odberg K. Pasientsikkerhet og velferdsteknologi – parhester i utakt [patient safety and welfare technology – out of sync]. TFO 2020 Nov 20;6(2):165-171. [doi: 10.18261/issn.2387-5984-2020-02-12]
17. Donabedian A. The quality of care: how can it be assessed. JAMA 1988 Sep;260(12):1743-1748. [doi: 10.1001/jama.1988.03410120089033]
18. Tossaint-Schoenmakers R, Versluis A, Chavannes N, Talboom-Kamp E, Kasteleyn M. The challenge of integrating eHealth into health care: systematic literature review of the donabedian model of structure, process, and outcome. J Med Internet Res 2021 May 10;23(5):e27180. [doi: 10.2196/27180] [Medline: 33970123]
19. Karlsen A, Sørbø MF, Nylenna M. Digitale helsetjenester: nye muligheter, nye utfordringer [digital health services: new opportunities, new challenges]. Det Norske Medicinske Selskab [The Norwegian Medical Association]. 2020. URL: https://www.michaeljournal.no/article/2020/08/Digitale-helsetjenester--nye-muligheter,-nye-utfordringer [accessed 2024-07-02]
20. WHO develops guidance to improve telemedicine services. World Health Organization. 2025. URL: https://www.who.int/europe/news-room/15-07-2024-who-develops-guidance-to-improve-telemedicine-services [accessed 2025-02-17]
21. WHO issues consolidated guide to running effective telemedicine services. World Health Organization. 2025. URL: https://www.who.int/news/item/10-11-2022-who-issues-new-guide-to-running-effective-telemedicine-services [accessed 2025-02-17]
22. Meld st 9 (2023–2024). Regjeringen.no. 2023. URL: https://www.regjeringen.no/no/dokumenter/meld.-st.-9-20232024/id3027594/ [accessed 2024-06-02]
23. Woolham J, Freddolino P, Gibson G, Daniels S. Telecare at a crossroads? Finding researchable questions. JET 2021 Sep 8;15(3):175-188. [doi: 10.1108/JET-11-2020-0049]
24. Greenhalgh T, Procter R, Wherton J, Sugarhood P, Hinder S, Rouncefield M. What is quality in assisted living technology? the ARCHIE framework for effective telehealth and telecare services. BMC Med 2015 Apr 23;13(1):91. [doi: 10.1186/s12916-015-0279-6] [Medline: 25902803]
25. Alvsåker YB. Velferdsteknologi i Norske hjemmetjenester: en praxeologisk studie om statlige visjoner i møte med markedskrefter og kommunale realiteter [welfare technology in Norwegian home services: a praxeological study of state

XSL•FO
RenderX

visions in the face of market forces and municipal realities]. Norwegian Research Information Repository.: The University of Bergen; 2023. URL: https://bora.uib.no/bora-xmlui/handle/11250/3106782 [accessed 2024-05-06]
26. Tid for handling NOU 2023: 4. Regjeringen.no. URL: https://www.regjeringen.no/no/dokumenter/nou-2023-4/id2961552/ [accessed 2024-06-17]
27. McLean S, Sheikh A, Cresswell K, et al. The impact of telehealthcare on the quality and safety of care: a systematic overview. PLOS ONE 2013;8(8):e71238. [doi: 10.1371/journal.pone.0071238] [Medline: 23977001]
28. Fothergill L, Hayes N, Latham Y, Hamilton J, Ahmed S, Holland C. Understanding how, for whom and under what circumstances telecare can support independence in community-dwelling older adults: a realist review. BMC Geriatr 2025 Jan 27;25(1):59. [doi: 10.1186/s12877-024-05650-6] [Medline: 39871177]
29. Safer primary care. World Health Organization. 2024. URL: https://www.who.int/teams/integrated-health-services/patient-safety/research/safer-primary-care [accessed 2024-08-28]
30. Al Salem G, Bowie P, Morrison J. Hospital survey on patient safety culture: psychometric evaluation in Kuwaiti public healthcare settings. BMJ Open 2019 May 30;9(5):e028666. [doi: 10.1136/bmjopen-2018-028666] [Medline: 31152040]
31. Nwankwo EI, Emeihe EV, Ajegbile MD, Olaboje JA, Maha CC. Integrating telemedicine and AI to improve healthcare access in rural settings. Int J Life Sci Res Arch 2024;7(1):059-077. [doi: 10.53771/ijlsra.2024.7.1.0061]
32. Guise V, Anderson J, Wiig S. Patient safety risks associated with telecare: a systematic review and narrative synthesis of the literature. BMC Health Serv Res 2014 Nov 25;14(1):588. [doi: 10.1186/s12913-014-0588-z] [Medline: 25421823]
33. Stokke R, Melby L, Isaksen J, Obstfelder A, Andreassen H. A qualitative study of what care workers do to provide patient safety at home through telecare. BMC Health Serv Res 2021 Jun 5;21(1):553. [doi: 10.1186/s12913-021-06556-4] [Medline: 34090450]
34. Eisenhardt KM. Building theories from case study research. Acad Manage Rev 1989 Oct;14(4):532-550. [doi: 10.5465/amr.1989.4308385]
35. Eisenhardt KM, Graebner ME. Theory building from cases: opportunities and challenges. AMJ 2007 Feb;50(1):25-32. [doi: 10.5465/amj.2007.24160888]
36. Yin RK. Discovering the future of the case study. method in evaluation research. Eval Pract 1994 Oct;15(3):283-290. [doi: 10.1177/109821409401500309]
37. Yin RK. Enhancing the quality of case studies in health services research. Health Serv Res 1999 Dec;34(5 Pt 2):1209-1224. [Medline: 10591280]
38. Denzin NK, Lincoln YS. Introduction: The Discipline and Practice of Qualitative Research', in Strategies of Qualitative Inquiry, 3rd edition: Sage Publications, Inc; 2008:1-43.
39. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
40. Lincoln YS, Guba EG. Naturalistic Inquiry: SAGE; 1985.
41. Saldana J. The Coding Manual for Qualitative Researchers: SAGE Publications Ltd; 2021. URL: https://www.torrossa.com/en/resources/an/5018667 [accessed 2024-01-20] [doi: 10.4135/9781036235611]
42. National committee for research ethics in the social sciences and the humanities. National Research Ethics Committees. 2024. URL: https://www.forskningsetikk.no/en/about-us/our-committees-and-commission/nesh/ [accessed 2024-01-30]
43. May C, Finch T. Implementing, embedding, and integrating practices: an outline of normalization process theory. Sociology 2009 Jun;43(3):535-554. [doi: 10.1177/0038038509103208]
44. May CR. Making sense of technology adoption in healthcare: meso-level considerations. BMC Med 2015 Apr 23;13(1):92. [doi: 10.1186/s12916-015-0305-8] [Medline: 25902829]
45. Huddlestone L, Turner J, Eborall H, Hudson N, Davies M, Martin G. Application of normalisation process theory in understanding implementation processes in primary care settings in the UK: a systematic review. BMC Fam Pract 2020 Mar 16;21(1):52. [doi: 10.1186/s12875-020-01107-y] [Medline: 32178624]
46. Dillon EC, Martinez MC, Li M, et al. "It is not the fault of the health care team - it is the way the system works": a mixed-methods quality improvement study of patients with advanced cancer and family members reveals challenges navigating a fragmented healthcare system and the administrative and financial burdens of care. BMC Health Serv Res 2024 Nov;24(1):1378. [doi: 10.1186/s12913-024-11744-z]
47. Fusco F, Marsilio M, Guglielmetti C. Co-creation in healthcare: framing the outcomes and their determinants. JOSM 2023 Dec 18;34(6):1-26. [doi: 10.1108/JOSM-06-2021-0212]
48. Fisk MJ, Woolham J, Steils N. Knowledge and skills sets for telecare service staff in the context of digital health. J Int Soc Telemed eHealth 2020 Dec;8. [doi: 10.29086/JISfTeH.8.e15]
49. Guise V, Wiig S. Perceptions of telecare training needs in home healthcare services: a focus group study. BMC Health Serv Res 2017 Feb 23;17(1):164. [doi: 10.1186/s12913-017-2098-2] [Medline: 28231852]
50. Anker-Hansen C, Johansen EK. Navigating care and communication: a qualitative study on nurses' perspectives in response centres. BMC Nurs 2025 Feb 13;24(1):170. [doi: 10.1186/s12912-025-02811-9] [Medline: 39948604]
51. Alenoghena CO, Ohize HO, Adejo AO, et al. Telemedicine: a survey of telecommunication technologies, developments, and challenges. JSAN 2023 Apr;12(2):20. [doi: 10.3390/jsan12020020]

XSL•FO
RenderX

52. Hajesmaeel-Gohari S, Bahaadinbeigy K. The most used questionnaires for evaluating telemedicine services. BMC Med Inform Decis Mak 2021 Feb 2;21(1):36. [doi: 10.1186/s12911-021-01407-y] [Medline: 33531013]
53. SMART guidelines. World Health Organization. URL: https://www.who.int/teams/digital-health-and-innovation/smart-guidelines [accessed 2025-09-17]
54. Leonardsen ACL, Hardeland C, Helgesen AK, Grøndahl VA. Patient experiences with technology enabled care across healthcare settings- a systematic review. BMC Health Serv Res 2020 Aug 24;20(1):779. [doi: 10.1186/s12913-020-05633-4] [Medline: 32838784]
55. Jarvis T, Scott F, El-Jardali F, Alvarez E. Defining and classifying public health systems: a critical interpretive synthesis. Health Res Policy Syst 2020 Jun 16;18(1):68. [doi: 10.1186/s12961-020-00583-z] [Medline: 32546163]
56. McCoy D, Allotey P. An introduction to health systems. In: Systems Thinking Analyses for Health Policy and Systems Development: A Malaysian Case Study: Cambridge University Press; 2021:3-16. [doi: 10.1017/9781108954846.004]
57. Dyer WG, Wilkins AL. Better stories, not better constructs, to generate better theory: a rejoinder to Eisenhardt. The Academy of Management Review 1991 Jul;16(3):613. [doi: 10.2307/258920]

Abbreviations

ARCHIE: anchored, realistic, cocreated, human, integrated, and evaluated
CC: call center
NPT: normalization process theory
WHO: World Health Organization

Edited by A Stone, N Cahill; submitted 15.Apr.2025; peer-reviewed by FA Dhabbari, Q Wang; revised version received 17.Jan.2026; accepted 20.Jan.2026; published 10.Mar.2026.

Please cite as:
Grøndal-Eeles LC, Dugstad J, Eide H, Nilsen E
Quality Challenges in Municipal Telecare Call Center Services: Qualitative Evaluation Using the Anchored, Realistic, Cocreated, Human, Integrated, and Evaluated (ARCHIE) Framework
J Med Internet Res 2026;28:e76054
URL: https://www.jmir.org/2026/1/e76054
doi:10.2196/76054

© Linda C Grøndal-Eeles, Janne Dugstad, Hilde Eide, Etty Nilsen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring the Dynamics of Actors, Structural Factors, and Bricolage in the Implementation and Sustainability of eHealth Solutions: Qualitative Multiple-Case Study

Susanne Eriksen[1*], BA, MSc; Christine Øye[2*], Prof Dr; Anne Marie Dahler[3*], Prof Dr

[1]Department of Health and Caring Science, Faculty of Health and Social Science, Western Norway University of Applied Sciences, Bjornsonsgate 45, Haugesund, Norway

[2]Faculty of Health and Social Science, Western Norway University of Applied Sciences, Stord, Norway

[3]Department of Applied Welfare Research, UCL University College, Odense, Denmark

[*]all authors contributed equally

Corresponding Author:
Susanne Eriksen, BA, MSc
Department of Health and Caring Science, Faculty of Health and Social Science, Western Norway University of Applied Sciences, Bjornsonsgate 45, Haugesund, Norway

Abstract

Background: European health care systems face mounting pressures from an aging population, workforce shortages, and decentralization, challenging the delivery of accessible, high-quality care. eHealth solutions are widely promoted to enhance efficiency and improve the quality of care. Despite a strong policy report, anticipated benefits remain unrealized, as implementation processes often encounter barriers and high failure rates. Research shows that drivers and barriers are dynamic and shaped by actor interactions. Some studies suggest that certain actors, often acting as bricoleurs, play a critical role in overcoming these barriers through adaptive and improvised practices. However, little is known about how these actors enact roles, what features enable bricolage, and how structural conditions influence these practices.

Objective: The aim of this study is twofold. First, it investigates the roles and features of actors involved in innovation processes, with a particular emphasis on the application of bricolage to overcome barriers and the influence of structural factors on these processes. Second, it aims to contribute both theoretical and empirical insights to deepen the understanding of barrier dynamics within innovation processes.

Methods: We conducted a multiple-case study comprising 10 semistructured interviews, 11 focus groups with health care professionals, managers, trainers, and policymakers, participant observations of training sessions, and document analysis. An iterative process integrated the dramaturgical approach with the concept of bricolage, guiding the reflexive thematic analyses.

Results: Roles were enacted based on available information, context, and assigned functions. Service specialists (eg, superusers) and mediators (eg, unit or project managers) gained backstage insights through shadowing staff, evaluations, and support activities. When mandated and equipped with contextual and technical knowledge, these actors became bricoleurs, addressing unforeseen challenges by creatively mobilizing resources and thereby transforming barriers into promoters. Effective bricolage required proximity to the implementation site, dedicated involvement, and experiential knowledge of health care and technical domains. Key drivers included colocation, supportive management, stable teams, superusers, tailored training, follow-up activities, and informal evaluations. Barriers such as organizational silos, leadership shifts, staffing shortages, high turnover, geographic dispersion, and technology perceived as challenging or surveillance-oriented constrained bricolage and hindered implementation.

Conclusions: Actors may become bricoleurs when their assigned roles, contextual knowledge, and backstage access enable them to improvise in response to unforeseen challenges. Through a dramaturgical lens, bricolage is an adaptive performance that sustains frontstage care delivery. Bricoleurs combine proximity, experiential knowledge, and dual expertise to transform barriers into drivers by adjusting the innovation process and fostering interaction. These practices illustrate the mutual shaping of structure and agency: enabling conditions expand the space for bricolage, while barriers narrow it. Understanding this dynamic is essential for advancing theory on innovation processes and for designing implementation strategies that leverage bricolage as a mechanism for transforming barriers into drivers of innovation.

(J Med Internet Res 2026;28:e79999) doi:10.2196/79999

KEYWORDS

eHealth solutions; innovation processes; interaction; public sector innovation; implementation

XSL•FO
RenderX

Introduction

Background

One of the significant challenges facing European societies is the aging population, coupled with a shortage of health care staff [1-4]. These challenges are compounded by a deliberate decentralization of health care services, driven by political decisions, which complicates efforts to ensure accessible and high-quality health care [2,5-8]. In response to these pressing issues, there is a growing emphasis on innovation within the public sector [9,10], particularly through the adoption of eHealth solutions [1-3]. eHealth solutions are often defined as the organization and delivery of health services using information and communication technologies (ICTs) to support and enhance health care [4,11]. Across European societies, including Norway and Denmark, which form the context of this study, there is strong optimism that innovative eHealth solutions can improve productivity, free up time for patient-centered care and core health care tasks [2,8,12], and empower older adults to live independently at home by enhancing their health and quality of life [13-16]. The belief stems from the potential that eHealth solutions can enhance efficiency, improve the quality of care, and bolster patient security [17-19]. However, despite a strong policy push for implementing eHealth solutions, authorities acknowledge that the full benefits have yet to be realized [2,3,20,21].

Research on innovation processes involved in implementing and sustaining eHealth solutions in health care reveals that these processes are often fraught with challenges, resulting in a high failure rate [22-26]. Numerous studies have identified a range of factors influencing the outcomes of eHealth innovation processes [22,26-33]. These factors are often presented as separate concepts, being either a driver or a barrier [34,35]. However, scholars are increasingly emphasizing the dynamic nature of drivers and barriers [36-38]. What may initially appear as a barrier can be transformed into a driver if addressed appropriately [38-40]. Detecting barriers and addressing them depends on the interaction among actors as well as proactive engagement from specific actors. Previous research has demonstrated that collaborating actors in innovation processes engage in complex and dynamic interactions and negotiations [38]. As the complexity of the innovation process increases, interactional barriers become more prominent [34,36,41], hindering effective knowledge sharing, which is crucial for implementing and sustaining eHealth solutions [25,32,38,42,43]. Research has identified specific actors who are effective in overcoming interactional barriers [38]. Such actors are frequently found to act as boundary spanners, bridging knowledge gaps, for example, between ICT staff, management, and health care professionals. By bridging knowledge gaps, these actors can transform barriers into opportunities, making innovations more contextually relevant and beneficial by, for example, facilitating the redesign of workflows, providing adequate training and support to users, and highlighting problematic issues [26,31,38,44]. These actors can also be referred to as bricoleurs [45-47], and their contributions are considered vital for integrating eHealth solutions into health care practices [7,48,49]. Bricolage is characterized by its spontaneous and improvised nature, involving small, pragmatic adjustments in response to unforeseen events or emerging needs [47,49]. This process is often informal, builds on embodied experience within the practice, and is inherently collaborative, as it necessitates interaction with other actors [45,47,49]. Given its intangible nature, it remains unclear how the context can facilitate bricolage work [46]. There also remains an incomplete understanding of the nature of the roles of these bricoleurs as well as how these roles might be identified, enabled, and enhanced [26,36,38,49,50]. The literature emphasizes the importance of investigating individual actors as units of analysis, given that microlevel practices have been largely overlooked [24,30,50]. Research has identified various features of individual actors involved in implementing and sustaining innovations, including ICT-related skills, experience, demographics, and personality traits [26,35]. Creative and empowered actors, who possess the ability to navigate and overcome a risk-averse administrative culture, play a crucial role in driving innovation [35]. However, it remains unclear how these features interact with the enactment of roles and the contextual factors. Bricolage provides an approach that emphasizes how innovation can be made possible by recognizing both the work of actors and the influence of structural factors [47]. We therefore consider the concept of bricolage useful for understanding the complex dynamics between the micro- and system levels of innovation processes.

Objectives

Against this background, this study aims (1) to investigate the roles and features of actors involved in innovation processes, with particular emphasis on the application of bricolage to overcome barriers and the influence of structural factors on these processes; and (2) to contribute both theoretical and empirical insights to deepen the understanding of barrier dynamics within innovation processes. We pursue these aims through a qualitative, multiple-case study set within a European Union innovation project that focuses on enhancing digital skills and innovation readiness. By examining 3 cases of implementing and sustaining innovative eHealth solutions, our study seeks to address the following research questions:

- How do actors involved in implementing and sustaining eHealth solutions enact the role of bricoleurs, and what features enable this role?
- In what ways is bricolage performed to transform barriers into drivers of innovation, and how do structural conditions shape or constrain these practices?

Methods

Study Design

The study used an exploratory case study design that adhered to the principles of multiple holistic case study design [51]. Each case served as a distinct unit of analysis. The exploratory case study is an empirical investigation that examines processes and uncovers mechanisms related to a contemporary phenomenon within its real-life context [51,52]. This approach was well-suited to our research design, as we aimed to understand the roles of various actors in innovation processes and to investigate how they overcome barriers and transform

XSL•FO
RenderX

them into drivers of innovation, as well as how structures influence the enactment of roles and performance of bricolage.

Case Selection and Data Collection

The cases selected for this study were part of the Digital and Innovation Skills Helix project, a European Union initiative aimed at enhancing digital skills and promoting innovation readiness. As part of a forthcoming innovation process aimed at implementing an eHealth solution, these cases tested 3 tools developed through the Digital and Innovation Skills Helix project. These tools facilitated the acquisition of digital skills, cocreative implementation planning, evaluation, and competence assessment. Further details about the tools can be found in our previous work [53].

The 3 cases are situated in a Nordic welfare context. Case A is a nursing home situated in a rural municipality in Western Norway, offering 24-hour health and care services to residents for both short-term and long-term stays. This municipality actively participates in regional and national eHealth networks and collaborates closely with neighboring municipalities to enhance health care services. The nursing home accommodates 41 residents and employs 60 staff members, including 1 leader and 4 unit managers. Since 2017, the municipality has systematically upgraded its outdated equipment with new eHealth solutions, such as electronic medicine dispensers and safety alarms. At the time of data collection, the nursing home implemented a new patient monitoring system that included digital supervision. This system's primary objectives were to increase service efficiency and improve patient security.

Case B is a home care service located in a rural municipality in Western Norway, providing 24-hour assistance with daily living and home health services. Most service users are frail older adults who require support to continue living at home. This home care service supports approximately 150 service users and employs 50 health care staff, including the home care service unit manager. Since 2015, the municipality has focused on implementing eHealth solutions in the health care sector to address demographic challenges and deliver sustainable health care services. The home care service implemented electronic door locks (e-locks) in service users' homes, aiming to provide faster and safer assistance.

Case C is a cross-sectoral collaboration between a hospital and a municipal home care service in Southern Denmark. The hospital offers emergency care, outpatient treatment, and examination services to patients who have been injured. The home care service offers 24-hour assistance, including support for daily living and somatic and psychiatric care. In 2018, the local hospital initiated a project to collaborate with the municipality through video consultations, specifically for discharging complex and vulnerable patients from the hospital to municipal care. Aligned with the Regional Council's digitalization strategy, the objective was to enhance cross-sectoral collaboration using technology, streamline discharge conferences, and ensure more coherent patient care. The hospital was allocated approximately US $1.5 million to develop digital competencies among its staff.

The selection of cases was pragmatic, as the available cases within the European Union project were limited, resulting in the inclusion of 3 cases that exhibited varying levels of complexity. As case C involved implementing video consultations in a cross-sector collaboration between a hospital and a municipal home care service in Southern Denmark, this case is more complex compared to cases A and B, which were more similar and less complex, as they implemented an eHealth solution within a single organization, engaging only a limited number of employee groups. The empirical material for these cases included document analysis, participant observation, semistructured individual interviews, and focus groups (see Multimedia Appendix 1 for an overview of the data collection).

The interviews and focus groups, with some being more prominently featured than others, served as a central component of our analysis, providing in-depth insights into the experiences and perspectives associated with the implementation process (Multimedia Appendix 2). The participant observations took place during training, where the observer participated and observed questions, discussions, and task-solving activities. The participant observations complemented the interviews, enhancing the analytical depth of the study by offering a comprehensive understanding of the setting and context as well as nuanced insights into the social dynamics (Multimedia Appendix 3). Although these observations, along with the document analysis, served as critical supplementary materials, they primarily provided a contextual backdrop for interpreting the qualitative data derived from interviews and focus groups. The data collection took place between 2020 and 2022, during and after the testing phase of the 3 tools. The data were gathered in person or online via Zoom (Zoom Video Communications) or Microsoft Teams, conducted by SE, AMD, or CØ. The data collection ended when data saturation was reached.

Participants for the study were recruited with the assistance of key stakeholders involved in implementation planning and training within the clinical settings. In cases A and B, a designated project manager from each case served as our primary point of contact. For case C, we contacted the individuals responsible for training at the hospital, where one of the trainers became our primary point of contact. These 3 contact persons facilitated the recruitment process by reaching out to potential participants by email. A total of 949 individuals were identified as potential participants from the training roster, and 31 individuals were identified from the implementation working groups. We contacted 115 individuals, of whom 39 agreed to participate in the interviews, and 70 agreed to participate in participant observations, resulting in a final sample of 109.

Ethical Considerations

Ethics approval was obtained from Sikt—Norwegian Agency for Shared Services in Education and Research (ref: 198584). Before participation, all participants were provided with informed consent that they signed before participation, which included a comprehensive overview of the study project, its objectives, and data handling procedures. No compensation was provided to the participants. Anonymity was ensured by replacing names with pseudonyms and revising the empirical

XSL•FO
RenderX

material to obscure recognizable quotes. To ensure data security, the original recordings as well as transcripts were stored as password-protected files on a secure research server, with access limited to the three authors only.

Analyses

NVivo (version 21; Lumivero) software was used to analyze verbatim transcriptions of audio-recorded interviews and focus groups. The reflexive thematic analyses for this study were applied to generate initial themes by identifying patterns of shared meaning across the dataset [54-56]. We followed the 6-phase approach developed by Braun et al [55]. First, we aimed to enhance reliability in the analysis process by conducting independent readings of the transcripts. Second, all three authors generated initial themes independently, guided by the concept of bricolage [57] in combination with the dramaturgical approach [58] and the Consolidated Framework for Implementation Research [59]. Since the study aimed to investigate role enactment, bricolage strategies, and structural factors, we approached the data, exploring and tagging text that reflected strategies involving bricolage activities and role features akin to those of mediators and service specialists, as well as structural factors affecting implementation. Third, all three authors constructed themes through thematic mapping, which involved visually exploring potential themes and subthemes, as well as connections between them. Finally, themes were revised and defined in collaboration between all three authors before producing the report. Researcher reflexivity was actively pursued in accordance with the standards of qualitative research [60]. This included engaging in critical dialogue to challenge and complement each other's interpretations and explicitly acknowledging our personal and professional backgrounds early in the research process. These discussions helped us remain aware of our initial assumptions and avoid conflating prior perspectives with insights emerging from the data.

Conceptual Framework

The dramaturgical approach by Goffman [58] illustrates how actors manage their frontstage performances within social settings to navigate and influence others' perceptions. In contrast, backstage involves actions that support the frontstage performance but do not align with its presented image. This perspective highlights that inventive strategies and adaptive behaviors emerge in response to the interactions and perceptions within actors' immediate social contexts rather than being dictated solely by structural conditions. By integrating the dramaturgical approach and the concept of bricolage, we aim to develop a more comprehensive understanding of how actors can transform barriers into drivers through performing bricolage. Previous research has indicated that distinct roles, such as mediators and service specialists, are pivotal in overcoming interactional barriers [38]. In the paragraph below, we delve deeper into the roles of mediators and service specialists, emphasizing their essential function in bridging interactional gaps.

A mediator acts as an intermediary, facilitating mutually beneficial agreements between 2 potentially opposing teams. By cultivating trust and managing confidential information, the mediator maintains a delicate balance, sometimes projecting a skewed perception of loyalty to foster closeness and understanding among the teams. Examples of mediators may include facilitators, project leaders, or department managers. Conversely, service specialists focus on constructing, repairing, and maintaining performance. Acting as "scene workers," they enable actors to effectively perform their roles and define situations without encountering dramaturgical obstacles [58]. Examples of service specialists could include ICT specialists, champions, or "superusers."

To investigate the strategies used by mediators and service specialists in merging frontstage with backstage to overcome barriers and further the innovation process, we integrate Lévi-Strauss's [57] concept of bricolage. Bricolage, derived from the French verb "bricoleur," means to tinker or improvise. Lévi-Strauss [57] conceptualized bricolage as a creative method of using available resources, contrasting the improvisational nature of the bricoleur with the systematic, planned approach of the engineer. This duality draws attention to different problem-solving methods [57]. Since its introduction, bricolage has found relevance across various disciplines [61-63], including studies on public sector innovation [7,45-49,64]. While extensive research has examined bricolage activities, there has been less focus on bricoleurs themselves [49]. This oversight may stem from Levi-Strauss's [57] structuralist perspective, which suggests that objective structures shape actors' lives, often overshadowing individual interpretations. However, latter interpretations of the concept of bricolage are not strongly underpinned by the structuralist approach and see bricolage as an activity where the bricoleur creates structures from resources at hand [46]. As such, bricolage is considered relevant for capturing the dynamics of innovation and the connection between microlevel practices and broader systems of innovation [46]. Along with other researchers [65-67], we argue for the need to develop new approaches to advance public sector innovation processes. As such, to better understand the motivations and strategies of actors, it is necessary to recognize the interplay between different approaches. Situated in the critical realism paradigm, we acknowledge that both Goffman's and Lévi-Strauss's approaches can complement one another. While objective structures influence individual agency, especially in complex situations, actors' agency is also shaped by their subjective interpretations of social reality.

The dramaturgical approach illuminates the social dynamics and interpersonal interactions on the microlevel that shape how bricoleurs operate, for example, how the roles are enacted, negotiated, and adapted based on the context and audience. Conversely, bricolage enhances our understanding of the resourcefulness and adaptability that actors demonstrate in their roles, particularly in overcoming barriers to innovation. It elucidates how bricoleurs creatively respond to unforeseen events posed by the context while simultaneously negotiating and navigating structures that influence their actions.

XSL•FO
RenderX

Results

Overview

The total sample consisted of 109 participants, including 39 who participated in interviews and 70 who participated in participant observations. The participants represented a diverse range of stakeholders, including health care professionals, unit managers, leaders, policymakers, ICT specialists, eHealth solution providers, trainers, project staff, and technical personnel. The following sections outline the themes, accompanied by illustrative quotations.

Enacting the Role of a Bricoleur

The actors assigned the role to spearhead the innovation processes in the 3 cases were a project manager (case A), a home care service manager (case B), and a project manager assisted by trainers (case C). According to Goffman [58], roles are enacted based on available information, the situational context, and the function an actor is expected to perform. When assigned a function such as project manager (case A and C), home care service manager (case B), or trainer (case C), actors face specific expectations from their audience, which they strive to meet or engage in impression management to convey that these expectations are being met. In case B, the role of a home care service manager entailed numerous time-consuming operational responsibilities, which constrained the time and attention available for the implementation process—particularly during the COVID-19 pandemic. The geographical distance also prevented him from shadowing staff and observing their activities firsthand. These conditions made it challenging to enact either a mediator or a service specialist role. Compounding these issues, the home care service manager was new to the organization, starting the position only 2 weeks after the national lockdown in Norway. These circumstances complicated the enactment of a bricoleur role. First, the manager lacked familiarity with the organizational context and staff. Second, not having participated in the planning phase, he had limited background knowledge of the e-lock project's rationale and objectives, as reflected by the following quote: "It was decided before I started working here. So, I'm not entirely sure of the background [of the project], but it's probably to save time on key usage and to increase accessibility and safety for the users" (Respondent 1, home care service manager). Finally, the operational demands inherent in the managerial role overshadowed the implementation process, leaving little opportunity to prioritize it. This contrasts with cases A and C, where project managers were fully dedicated to the implementation process and exempt from other operational responsibilities. Although the project managers were fully dedicated to the innovation process, the way it unfolded varied across the 2 cases. In case A, the project manager and a project coordinator moved into the nursing home during the initial 2 weeks of implementation, providing continuous on-site support and guidance. After this period, they remained available via telephone and email and maintained an office next door to ensure ongoing support and rapid problem-solving. Due to poor collaboration with the ICT department in the municipality, the project manager and coordinator were also compelled to take on responsibilities as ICT specialists. This additional responsibility led the project manager and the coordinator to enact roles of service specialists, which had many advantages for the ongoing process, as expressed by a staff member:

> *I think it would have been difficult without them. It makes the workday easier. We spend less time on frustration, because, well, technology isn't my field, you know. My field is actually healthcare. But having those who are a support function for technology somehow makes my day easier.* [Respondent 2, nurse]

As the patient warning system and digital supervision were implemented directly within the nursing home, the service specialists remained in proximity to the site of action. The continuous presence of the project manager and coordinators, who engaged in shadowing staff and observing daily routines, enabled them to adopt dual roles as both service specialists and mediators gradually. In doing so, they effectively bridged the gap between backstage and the front stage. Drawing on their familiarity with the context and the actors involved, along with their acquired ICT competencies, they gradually enacted roles as bricoleurs—constructing, repairing, and maintaining the health care professionals' performance so they could do their "actual job" (Respondent 2, nurse). The project team also trained a group of superusers, who in turn took on roles as service specialists and actively supported bricolage activities by drawing on their digital competence.

Although the project manager in case C was committed to the innovation process, the physical location—outside the hospital in a separate building—created a spatial distance from the implementation site. Furthermore, the affiliation with a research and innovation unit, rather than the hospital units or the municipality, meant that the project manager lacked an established foothold within the everyday workplace dynamic. These conditions challenged the ability to enact a role as a mediator or service specialist, as the project manager was not embedded in the daily routines of the health care staff and thus had limited access to backstage information, such as training needs and perceptions about the video consultations. Enacting roles as mediators and bricoleurs was also challenging for the unit managers when priorities competed, tasks were incompatible, and responsibilities conflicted. As expressed by a unit manager: "We were a COVID unit, so we already had plenty to just ... work through, so the idea of creating new ideas and having the time to implement new systems ... Well, the staff was also overloaded. You reach a certain limit where you can't take in any more" (Respondent 3, hospital unit manager). However, in some of the units, the trainers were able to step into roles as service specialists and bricoleurs when they were physically present in the hospital to support the setup of video consultations. Their role as trainers granted them access to the backstage where health care professionals prepared for their frontstage performances. This backstage access enabled the trainers to observe real-world challenges and tailor both the training and video consultations to the specific needs of each unit. Drawing on their ICT expertise, they engaged in bricolage, creatively assembling and adjusting tools and practices to support the health care professionals in delivering care. The combination of service specialist and bricoleur roles illustrates how these actors not only facilitated the technical

XSL•FO
RenderX

implementation but also helped maintain the integrity of the health care professionals' frontstage performance. As reflected by the hospital director:

> *Every time we introduce a new ICT or digitalisation product, we turn the experts into novices. They deal with new technology, altering patient-health professional relationships [...]. If the doctor is struggling with an ICT system, they lose some respect in the eyes of the patients. Because then the patient sees them fumbling around and may think they are equally clumsy with all their professional expertise. [...] How can we ensure our healthcare professionals are not novices in ICT but at least able to use it effectively, so they don't come across as incompetent or anything like that? When healthcare professionals experience technical problems, they tend to revert to what they are accustomed to and can handle better.* [Respondent 4, hospital director]

This quote highlights the delicate balance between adopting eHealth solutions and maintaining a professional identity. Trainers who acted as bricoleurs helped preserve this balance by ensuring that health care professionals could maintain confidence and competence in their frontstage roles, even when navigating unfamiliar eHealth solutions.

The 3 cases show that the ability to enact the role of bricoleurs depends on actors' experiential knowledge, proximity to the implementation site, and a dedicated focus on implementation, which makes backstage dynamics more accessible.

Performing Bricolage to Transform Barriers Into Drivers

Goffman [58] distinguishes between 2 models of behavior: the real and the contrived. The real is regarded as a genuine performance, which is not consciously assembled, but an unintended product of an actor's spontaneous reaction to the actual situation at hand. Contrived performances, on the other hand, are regarded as carefully constructed, with each artificial element added one by one, since the behavior is not reacting to any actual situation. How one responds spontaneously to unforeseen events determines whether the performance is characterized by bricolage or a more systematic and planned engineering approach to overcome barriers [57]. Refer to Table 1 for an overview of the drivers and barriers across the 3 cases.

XSL•FO
RenderX

Table . Cross-case synthesis of structural drivers and barriers of bricolage.

Structural factor	Drivers	Barriers
Organizational structure	Colocation and small teams enable rapid feedback and access (A); proximity in small municipalities fosters collaboration (A, B).	Silos and interorganizational differences hinder coordination (A, C); concurrent large IT projects (electronic health records) compete for attention (C); geographic dispersion (A, B, C); role overload in small municipalities (A, B).
Leadership and governance	Supportive management and dedicated project leadership drive progress (A, B, C).	Shifting leadership and lack of clear ownership delay implementation (B, C); operational pressures crowd out strategic work (C).
Political framework	Alignment with local or regional or national priorities; digitalization targets create mandate (A, B, C).	—[a]
Resources	Adequate funding and equipment (A, B, C); stable workforce (A, B); superusers support adoption (A).	Staffing shortages, turnover, and crises (eg, COVID-19, strikes) impede adoption (C); lack of designated superusers (B, C).
Cultural norms and values	Positive work culture (A), and perceived expectations to use technology motivate uptake (A, B, C).	Strong patient-safety ethos and professional identity can slow change (A, B, C); small-community dynamics can amplify resistance (A).
Training and competence development	Needs-based, one-on-one, learning-by-doing, and follow-up (A, B, C) are effective; dedicated trainers also provide support (A, C).	24/7 operations complicate scheduling (A, B, C); generic courses without local tailoring are less effective; trainer distance and high turnover reduce retention (C).
Infrastructure and technology	Technology perceived as applicable or simple supports use (A, B); technology used proximate to superusers or project management supports use (A).	The technology is implemented in settings that lacked proximity to superusers and management (B, C). Technology perceived as surveillance (A), time-consuming (B, C), poorly functioning (C), and not beneficial (B, C) hinders its use. Poor set-up (A, C) and lack of deimplementation (B) create friction.
Communication systems	Multiple channels (meetings, emails, direct access to project leads) aid communication (A, B, C).	Weak interdepartmental and sectoral channels (A, C) and 24/7 staffing patterns (A, B, C) limit the information flow; the existing communication system between sectors hinders the development of a new communication system (C).
Daily relations and collaboration	Colocation and bridging roles (eg, unit management and superusers) improve collaboration (A).	Geographic dispersion (A, B, C) and lack of cross-sector workflows impede collaboration (A, C).
Quality assurance and evaluation routines	Routine monitoring (B, C) and frequent informal evaluation support technology adoption (A).	Monitoring can be perceived as surveillance; the absence of informal feedback loops reduces responsiveness (B, C).

[a]Not applicable.

By being proximate and shadowing the staff, the project manager in case A was present when unforeseen events happened. For example, when a ghost appeared on the patient monitoring system, as elicited by the project manager:

> *We set up [digital supervision] for a user [and] it was supposed to alert us if the person got out of bed. [...] It's nighttime, and the user is restless, so a night shift comes in [...]. She sits on the edge of the bed to calm the patient down. Then she stands up, which triggers an alert for getting out of bed because she's probably been sitting on the bed for too long. Due to our poor Wi-Fi, the signal is delayed. [...] Then, a few minutes pass, and an alert goes off on the mobile phones that the bed is abandoned. An anonymized picture of a man appears. Then, the others panic and rush in, wondering who's here, and they get the idea that this is a ghost [...]. They find out that the woman's late husband is in the room. Because he lived at the nursing home before.* [Respondent 5, project manager]

One of the health care professionals:

> *[...] There's probably much more between heaven and earth than we see, but I'm unsure if it would affect patient monitoring. [...] It's sometimes a profession shrouded in superstition and stories. [...] We are often close to death, right? So maybe it's a bit natural to become a bit superstitious.* [Respondent 2, nurse]

Even though the project manager could not convince the staff that the ghost appeared due to a slow Wi-Fi signal, responding constructively to this event spontaneously required an understanding of the context and the actors involved, as well as technical competence to grasp how cultural and emotional

XSL•FO
RenderX

factors intersect with technological implementation. The project manager also participated in staff meetings to evaluate the use of the patient monitoring system, ensuring that issues were raised and resolved. The training was conducted informally and continuously, based on immediate needs:

> *We simply have to practice. I have to show the staff and test. I don't know how often I've been on the floor to demonstrate a fall. [...] I have to show them and have them try it themselves. We constantly have to repeat, demonstrate, you know ... "Now, I triggered a violence alarm. What do you do then?" And it's like ... recreating scenarios.* [Respondent 5, project manager]

By being present and practicing in the situation, the project manager made it challenging for the staff to conceal actions or project desired impressions. This led the staff to use the patient warning system and digital supervision even when they were not fully confident in operating it. They either invited the project manager backstage for assistance or proceeded without the necessary skills, which created technical challenges and unforeseen events. These events, however, functioned as dramaturgical disruptions that exposed latent vulnerabilities in the performances of health care professionals. Once these issues were made visible, the project manager could implement improvisational strategies using the available resources.

In contrast to case B, where the e-locks were used in service users' homes, this made it difficult for the home care service manager to maintain proximity and shadow health care professionals during their work. Consequently, the home care service manager was not exposed to many unforeseen events caused by technology. The e-lock system log was used to monitor use, but it did not provide sufficient information to facilitate bricolage. As a result, the home care service manager remained unaware of how the staff perceived the e-locks. According to the staff, they only used the e-locks about 50% of the time. One staff member had experienced the e-lock malfunctioning once, so they always carried the regular keys "just in case" (Respondent 6, nurse). Staff members did not convey their experiences and concerns. Consequently, the management remained unaware of these issues. As reflected in the quote from the home care service manager below:

> *It was straightforward and a truly positive thing. It probably hasn't caused many obstacles. Of course, some actors might have hesitated a bit more to carry it out, but it's not like they have rallied others, because it's so simple. So I haven't heard that anyone has really resisted or that there has been noise around this technology. I would have known if there had been something.* [Respondent 1, home care service manager]

This quote reflects that when new or inexperienced leaders are given formal authority over more experienced staff members, the formally empowered actor often enacts a role of symbolic dominance. In contrast, the staff members are the ones who truly run the show [58]. This discrepancy between formal authority and practical influence served as a barrier to performing bricolage. Transforming this barrier to a driver requires strategies to gain backstage access. Because 24/7 operations made formal training difficult, he improvised by installing an e-lock on his office door, allowing health care professionals to practice at their convenience. This arrangement enabled him to remain close to the action and shadow staff during their practice sessions—providing a form of backstage access. These efforts ensured that staff acquired the necessary competencies to operate the e-lock. The staff found the training sufficient for using the e-locks, as remarked: "It wasn't that difficult" (Respondent 6, nurse). However, the staff questioned the usefulness and the functionality of the e-locks, as reflected in the following exchange:

> *Respondent 6, nurse: If [the e-lock] works, it's certainly easier.*
>
> *Respondent 7, nurse: Hmm ... I'm not so sure about that.*
>
> *Respondent 8, nurse: If it works.*
>
> *Respondent 7, nurse: It's very easy just to find the right key and unlock it.*
>
> *Respondent 8, nurse: Yeah, especially if it's raining.*

As such, this performance preparation training facilitated by the home care service manager had a limited effect, as e-locks were still not consistently used in service users' homes. The home care service unit manager emphasized the importance of managing explicit resistance happening on the frontstage: "Those who I thought might give the most resistance, they were actually the first ones I gave training to, and maybe followed up a little extra" (Respondent 1, home care services manager). However, silent backstage resistance might be a barrier more challenging to overcome than explicit frontstage resistance. In case C, many units struggled to implement video consultations for discharge conferences; however, a few units succeeded by engaging in adaptive, improvisational practices. In these units, the technology was not merely adopted as prescribed but reinterpreted and repurposed for alternative use—such as municipal rehabilitation supervision or admission meetings with relatives—illustrating a shift from the scripted "frontstage" plan to context-sensitive enactments. Training similarly moved backstage, taking place within the units rather than in formal classroom settings, fostering unforeseen events and situated learning. The trainers emphasized the importance of readily available support to answer questions and assist with new technology, as stated in the following quote from one of the trainers, highlighting the need for continuous and organized follow-up training in the units:

> *[...] Sometimes I think we could try to be more organised in training in the units. I'm not sure if there should be someone who becomes a superuser or is somehow responsible for it [...]. If one were to mention a missing link, I think it's the transition from having received the training to becoming an integrated part of the unit. That's where we could do something more. [...] It works really well if they have had training with us, and then we have been [in the units]* [Respondent 9, trainer]

The performance of bricolage in these units was facilitated by the trainers' presence backstage, their flexibility, and consistent

XSL•FO
RenderX

availability for support, as well as their provision of ad hoc problem-solving for unforeseen events. As the 3 cases show, responding to unforeseen events in a systematic and planned manner can be challenging. To effectively perform bricolage, it is essential to understand and act upon the context and engage with the involved actors. Understanding the context derives from experiential knowledge, the function assigned, and access to backstage insights. Performing bricolage is not just about creatively navigating unforeseen events using the resources at hand; it is also about being mindful of the interactions and relationships with other actors, ultimately shaping the experience and outcome for everyone involved in the innovation process.

Features of a Bricoleur

Several common features were observed among the actors who performed bricolage. They were familiar with the local work context, able to move fluidly between technical and clinical domains, and demonstrated a readiness to improvise and adapt eHealth solutions to meet the practical needs of health care professionals. As the project leader in case A noted: "Now I am the ICT department. I go into the computer cabinet myself and connect things" (Respondent 5, project manager). Another feature of the bricoleurs was their assigned role, such as project leader or trainer, which required full dedication to the innovation process. This meant they could be consistently available, responsive, and actively worked to meet the expectations of both the leaders and the health care professionals they supported, as emphasized by one of the trainers in case C: "It doesn't matter how much time has passed; I always make myself available" (Respondent 9, trainer). The physical proximity to the implementation site and involved actors enabled frequent, informal interaction that granted them access to backstage areas.

In contrast, actors who lacked the essential features for performing bricolage, such as those who were physically distant from the implementation site or held a position that involved other tasks and responsibilities, struggled to enact roles as bricoleurs. Without embeddedness in the local work context or access to backstage interactions, for example, by being new to the organization, these actors were less able to understand the nuances of everyday practices or respond to emerging needs. For instance, this resulted in staff perceptions of management being "unresponsive to feedback" (Respondent 7, nurse) in case B or feeling surveilled as in case C:

> *[The management] is definitely monitoring whether we are conducting these video consultations. We figured that out. We didn't actually know. But they are keeping an eye on us. [...] What is the reason for that? Well, I think it's because we first tried to say ... "Do we really need this? And should we do it now?" It was very emphatically stated that we should. Also, somewhere between the lines, it says that most consultations should be conducted over video. We have internally determined that as long as we say half/half, we'll have to see if they eventually knock on the door and say "no, no, no, now there are too many physical meetings every day"* [Respondent 10, nurse]

This quote captures the management's attempt to gain backstage access without success, while the nurse and her team strive to protect this backstage area. Consequently, actors without bricoleur features struggle to gain backstage information and are left to resort to more systematically planned approaches, limiting their contributions to, for example, formal training sessions or top-down strategies with limited impact on everyday practices among health care professionals.

Key features of bricoleurs are their familiarity with the local context, the ability to bridge technical and clinical domains, and consistent presence at the implementation site. Together with their dedicated roles and informal access to backstage dynamics, they are enabled to improvise and adapt eHealth solutions to meet practical needs.

Interactions Between Structures and Actors

In all 3 cases, the implementation of eHealth solutions was aligned with local, regional, and national priorities. The initiatives were supported by committed leadership, adequate funding, and a shared perception among actors that the use of eHealth solutions was both expected and necessary. However, these drivers carried limited weight when other structural factors (see Table 1 for an overview of barriers and drivers across cases), such as shortages of health care professionals, posed substantial barriers to implementation, as expressed by a member of the executive hospital management in case C, which was also the case featuring the most complexity:

> *We want to do a lot, but we don't have the resources. [...] We hear: " You get a lot of money; just get started." On the other hand, we simply don't have anyone who can [do it]. We clearly see that there is a need for these things. [...] It creates a vicious cycle spiralling downward, and we need to turn it around somehow.* [Respondent 4, hospital director]

Other barriers to bricolage included organizational silos (case C), interorganizational (case A) differences, shifting leadership (cases B and C), the pressures of 24/7 operations (all cases), high turnover rates (case C), the absence of superusers (cases B and C), and a strong patient-safety ethos (all cases). Additional barriers stemmed from the maintenance of professional identities (all cases), a lack of deimplementation practices (case B), the absence of informal feedback loops, and geographic dispersion among the actors. Technology itself also posed barriers, particularly when it was perceived as a form of surveillance (case A), as time-consuming (cases B and C), or as offering limited benefits to clinical practice (cases B and C). Under such conditions, performing bricolage becomes a challenging task.

Despite structural complexity and barriers, bricolage occurred across all 3 cases. Actors were not merely shaped by these structures; they actively worked within and upon them. Actors performed bricolage, particularly when structural drivers created openings for creative and adaptive problem-solving. While organizational silos, 24/7 operations, and technological infrastructure influenced what was possible, actors leveraged their assigned functions and interactions, using their expertise to adapt, reshape, and, at times, transform these structures. Key drivers of bricolage included colocation and proximity between

XSL•FO
RenderX

actors, which facilitated informal interactions and backstage access. The presence of supportive management, dedicated project leads, and a stable workforce created a foundation for continuity and responsiveness. A high number of superusers, along with tailored training that addressed the specific needs of each unit, enabled health care professionals to engage more confidently with eHealth solutions. Follow-up activities, dual roles (where project leaders also serve as ICT support), routine monitoring, and frequent informal evaluations further supported backstage access and, consequently, adaptive and creative problem-solving.

Discussion

Principal Findings

This study explored how actors involved in implementing and sustaining eHealth solutions enact roles as bricoleurs and perform bricolage to transform barriers into drivers of innovation. Using Goffman's [58] dramaturgical approach and Lévi-Strauss's [57] concept of bricolage, we examined how structural factors and actors' agency interact in complex innovation processes.

This cross-case analysis showed that the ability to enact the role of a bricoleur increased when actors were assigned a function dedicated to the innovation process, thereby being freed from operational demands. For example, being assigned the function of a trainer (case C) provided dedication, but also a mandate and a set of expectations to be fulfilled. Further, the enactment of the bricoleur role depended on actors' experiential knowledge and proximity to the implementation site, which facilitated access to backstage dynamics. The project team in case A is a good example that demonstrates how proximity and presence facilitated backstage access where routines were rehearsed, problems surfaced, and informal knowledge was shared. This backstage access led to a deeper understanding of the everyday challenges faced by health care professionals. Due to the project teams' continuous presence, health care professionals had to use eHealth solutions despite technical uncertainty, which in turn led to genuine performances, flaws, and unforeseen events. These findings resonate with previous research on bricolage [45,47,49] and add to the incomplete understanding of how the bricoleur role can be identified, enabled, and enhanced [26,36,38,49,50].

Bricolage emerged as a spontaneous and improvised response to unforeseen events, as found in previous research [47,49]; however, our research expands the literature by showing that bricolage builds on the bricoleur's experiential knowledge and dual expertise in clinical and technical domains, as was apparent in cases A and C. In these cases, the bricoleurs could transform barriers such as 24/7 operational demands and technological issues into drivers through, for example, ad-hoc, tailored training and by adapting solutions in contextually appropriate ways. Previous research has highlighted that barriers are dynamic and may be transformed into drivers if addressed appropriately [36-40]. We propose that bricolage offers a promising approach for facilitating such transformations.

Without backstage access, actors are left to rely on formal and systematic plans, strategies, and communication, as was the case for the project manager in case C and the home care service manager in case B. In both cases, silent resistance in the form of subtle, unspoken disengagement with the eHealth solutions hindered the implementation. Unlike explicit resistance, silent resistance may be difficult to detect and address, especially when actors lack backstage access. While previous research has emphasized how interactional barriers can hinder innovation processes and underscored the importance of collaboration [25,32,34,36,38,41-43] and boundary-spanning roles [26,31,38,44] in mitigating such barriers, our study adds nuance by demonstrating how bricoleurs can mitigate interactional barriers through backstage access.

Despite structural and contextual constraints, bricolage emerged across all cases. In case A, due to collaboration challenges with the ICT department, the project manager acquired ICT competencies and assumed responsibility for ICT-related tasks, which ultimately proved beneficial to both the staff and the innovation process. Rather than being passively shaped by structural conditions, actors can interact with and actively influence the conditions. This highlights how bricolage is not only resourceful but also interactive and performative, enabled by the interplay between agency and context, offering insight into the previously noted uncertainty regarding how the context facilitates bricolage work [46].

Our findings align with and enrich existing implementation frameworks such as the Non-adoption, Abandonment, Scale-up, Spread, and Sustainability framework [25,68] and Consolidated Framework for Implementation Research [59]. While these frameworks emphasize the importance of context, complexity, and actor engagement, they lack a detailed account of the microlevel improvisations that sustain implementation in practice. We contribute to an underexplored area [24,30,50] by proposing that bricolage offers a complementary mechanism that explains how actors navigate complexity not by eliminating it, but by working within and around it. As such, we suggest that bricolage can be conceptualized as a mechanism that links structural factors with microlevel practices (eg, improvisation, adaptation, and role enactment). This perspective can inform the design of implementation strategies that are more responsive to local contingencies and the agency of actors.

Our study aligns with previous literature on the characteristics of actors driving innovation, highlighting the importance of experience and ICT-related skills [26,35]. However, our study expands the literature by demonstrating that features of a bricoleur also encompass proximity, dedication to the implementation process, dual-expertise, and access to backstage dynamics. These features were most evident in cases A and C. As such, the findings of our study demonstrate that bricolage is not a spontaneous phenomenon or a matter of individual creativity [35]; rather, it is contingent upon various structural factors and interactions with actors. Our contribution offers nuanced, empirically grounded, context-sensitive illustrations that enrich existing understandings of bricolage and its role in innovation processes [45-49]. Although prior studies have conceptualized bricolage [45-49], our research advances the field by identifying the conditions and actors that enable its

XSL•FO
RenderX

practical enactment. Our research advances the field by empirically demonstrating how specific structural and contextual conditions, such as proximity and dedicated roles, enable the positioning of actors to enact roles as bricoleurs. Through detailed cross-case analysis guided by dramaturgy and bricolage, we have identified key features of bricoleurs and demonstrated how backstage access enables them to respond to unforeseen events in real time. This responsiveness creates opportunities to transform barriers into drivers, thereby sustaining innovation processes in complex health care settings.

In times of austerity, bricolage—an approach that uses available resources—may be notably applicable. Specifically, given that public sector innovation processes are often fraught with challenges and have a high failure rate [22-26], and as new approaches to advance, public sector innovation processes are asked for [65-67].

Limitations and Future Research

This study used a qualitative multiple-case design with methodological triangulation. While this approach enabled a rich and nuanced understanding of the implementation of eHealth solutions, the findings are grounded in 3 specific cases in a Nordic welfare context. As such, transferability to other settings may be limited. However, the implementation of eHealth solutions in health care is highly relevant across a wide range of countries. eHealth solutions tend to influence professional roles, interactions, and organizational structures in ways that are difficult to predict and anticipate [32,38,42,43,69]. These dynamics apply regardless of context. While our findings are not intended to be generalized, the conditions that enable successful bricolage—such as proximity and dual expertise—may offer valuable insights across diverse contexts.

Future research could build on this work by using mixed methods designs or larger-scale comparative studies to examine how bricolage unfolds across different health care systems or policy environments. Additionally, further exploration of silent resistance, including its manifestations, consequences, and strategies for mitigation, could deepen our understanding of the subtle dynamics that shape innovation processes in complex health care settings.

Conclusions

This study contributes to the understanding of innovation processes in the public sector by illuminating how actors navigate complex implementation processes through bricolage. By integrating dramaturgy with bricolage, we offer a novel analytical lens that captures both the performative and improvisational dimensions of innovation work. This theoretical pairing enabled us to explore how roles are dynamically enacted and adapted in response to structural and contextual factors, as well as emergent challenges.

Our findings show that actors become bricoleurs not merely by individual traits, but through a combination of contextual knowledge, proximity to the implementation site, and access to backstage dynamics. These conditions enable bricoleurs to improvise, adapt, and sustain innovation efforts in ways that formal strategies alone may not achieve. Our study provides detailed, context-sensitive illustrations that enrich existing theories of innovation and implementation. Recognizing and supporting conditions for bricolage may help design more adaptive, responsive, and sustainable strategies. We demonstrate how bricolage operates as a mechanism that links structural factors with microlevel improvisation, offering a valuable complement to established implementation frameworks. The theoretical pairing further clarifies how social expectations, role performances, and backstage interactions shape the conditions under which bricolage can occur.

Acknowledgments

The authors thank the participants who took part in their study. The authors also thank the project management in the Digital and Innovation Skills Helix (DISH) project and the funders of the DISH project: Erasmus+ Program of the European Union, Key Action 2 Cooperation for Innovation, and the Exchange of Good Practices—Sector Skills Alliances. The authors declare the use of generative artificial intelligence in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI (generative artificial intelligence) tools under full human supervision: proofreading, editing, and translation. The GAI tool used was ChatGPT-4. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was funded by the Norwegian Ministry of Education and Research and Western Norway University of Applied Science. The funders had no involvement in the study design, data collection, analyses, interpretation of results, or the writing of the manuscript.

Data Availability

The data analyzed during this study are not publicly available due to confidentiality reasons, but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: SE (lead), CØ (equal), AMD (supporting)

XSL•FO
RenderX

Data curation: SE (lead), CØ (supporting), AMD (supporting)
Formal analysis: SE (lead), CØ (equal), AMD (equal)
Funding acquisition: CØ
Investigation: SE (lead), CØ (equal), AMD (equal)
Methodology: SE (lead), CØ (supporting), AMD (supporting)
Validation: SE (lead), CØ (equal), AMD (equal)
Visualization: SE
Writing—original draft: SE (lead), CØ (supporting), AMD (supporting)
Writing—review and editing: SE (lead), CØ (supporting), AMD (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of data collection.
[DOCX File, 29 KB - jmir_v28i1e79999_app1.docx]

Multimedia Appendix 2
Interview and focus group guide.
[DOCX File, 17 KB - jmir_v28i1e79999_app2.docx]

Multimedia Appendix 3
Participant observation guide.
[DOCX File, 19 KB - jmir_v28i1e79999_app3.docx]

References

1. Recommendations on digital interventions for health system strengthening. WHO guideline. World Health Organization. 2019. URL: https://iris.who.int/server/api/core/bitstreams/c3c53f30-23cc-48d0-a3b5-c05ddd7f5349/content [accessed 2025-12-20]
2. Norges Offentlige Utredninger. Tid for handling. Personellet i en bærekraftig helse-og omsorgstjeneste. : Helse-og Omsorgsdepartementet; 2023 URL: https://www.regjeringen.no/contentassets/337fef958f2148bebd326f0749a1213d/no/pdfs/nou202320230004000dddpdfs.pdf [accessed 2025-12-12]
3. Regeringen, Kommunernes Landsforening, Danske regioner. Digitalisering, der løfter samfundet: den fællesoffentlige digitaliseringsstrategi 2022-2025. sammen om en digital fremtid. 2022 URL: https://fm.dk/media/vvkmho33/digitalisering-der-loefter-samfundet_den-faellesoffentlige-digitaliseringsstrategi-2022-2025_web.pdf [accessed 2025-12-12]
4. e-Health—making healthcare better for European citizens: an action plan for a European e-Health area. : European Commission; 2004 URL: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2004:0356:FIN:EN:PDF [accessed 2025-12-12]
5. Peine A, Faulkner A, Jæger B, Moors E. Science, technology and the ‘grand challenge’ of ageing—understanding the socio-material constitution of later life. Technol Forecast Soc Change 2015 Apr;93:1-9. [doi: 10.1016/j.techfore.2014.11.010]
6. Mort M, Roberts C, Callén B. Ageing with telecare: care or coercion in austerity? Sociol Health Illn 2013 Jul;35(6):799-812. [doi: 10.1111/j.1467-9566.2012.01530.x] [Medline: 23094945]
7. Procter R, Greenhalgh T, Wherton J, Sugarhood P, Rouncefield M, Hinder S. The day-to-day co-production of ageing in place. Comput Supported Coop Work 2014 Jun;23(3):245-267. [doi: 10.1007/s10606-014-9202-5]
8. Indenrigs- og Sundhedsministeriet. Robusthedskommissionens anbefalinger. : Sundhedsministeriet; 2023 URL: https://www.ism.dk/Media/638336462586551242/Robusthed-Samlet-Rapport-TILG.pdf [accessed 2025-12-12]
9. Osborne SP, Brown L. Innovation, public policy and public services delivery in the UK. The word that would be king? Public Adm 2011 Dec;89(4):1335-1350. [doi: 10.1111/j.1467-9299.2011.01932.x]
10. State of the Innovation Union 2015. : European Commission; 2015 URL: https://op.europa.eu/en/publication-detail/-/publication/0487b7b9-b5d6-11e5-8d3c-01aa75ed71a1/language-en [accessed 2025-12-12]
11. Boogerd EA, Arts T, Engelen LJ, van de Belt TH. “What is eHealth”: time for an update? JMIR Res Protoc 2015 Mar 12;4(1):e29. [doi: 10.2196/resprot.4065] [Medline: 25768939]
12. Norges Offentlige Utredninger. Innovasjon i omsorg. : Helse-og Omsorgsdepartementet Oslo; 2011 URL: https://www.regjeringen.no/contentassets/5fd24706b4474177bec0938582e3964a/no/pdfs/nou201120110011000dddpdfs.pdf [accessed 2025-12-12]
13. World report on ageing and health. World Health Organization. 2015. URL: https://www.who.int/publications/i/item/9789241565042 [accessed 2025-12-11]

XSL•FO
RenderX

14. Green paper on ageing: fostering solidarity and responsibility between generations. : European Commission; 2021 URL: https://commission.europa.eu/system/files/2021-06/green_paper_ageing_2021_en.pdf [accessed 2025-12-12]
15. Meld. St. 24. (2022-2023). Fellesskap og meistring. Bu trygt heime. : Helse- og Omsorgsdepartementet; 2023 URL: https://www.regjeringen.no/contentassets/a8280e2548c04d3ea6898078480bfa0c/nn-no/pdfs/stm202220230024000dddpdfs.pdf [accessed 2025-12-12]
16. Regeringen. Regeringens ældrereform. Du bliver aldrig for gammel til at have det godt. : Social-, Bolig- og, Ældreministeriet; 2024 URL: https://regeringen.dk/media/ec0nhyiw/regeringens_aeldreudspil_jan2024.pdf [accessed 2025-12-12]
17. Berge MS. Telecare—where, when, why and for whom does it work? A realist evaluation of a Norwegian project. J Rehabil Assist Technol Eng 2017;4:2055668317693737. [doi: 10.1177/2055668317693737] [Medline: 31186924]
18. Dugstad J, Eide T, Nilsen ER, Eide H. Towards successful digital transformation through co-creation: a longitudinal study of a four-year implementation of digital monitoring technology in residential care for persons with dementia. BMC Health Serv Res 2019 Jun 10;19(1):366. [doi: 10.1186/s12913-019-4191-1] [Medline: 31182093]
19. Huter K, Krick T, Domhoff D, Seibert K, Wolf-Ostermann K, Rothgang H. Effectiveness of digital technologies to support nursing care: results of a scoping review. J Multidiscip Healthc 2020;13:1905-1926. [doi: 10.2147/JMDH.S286193] [Medline: 33328736]
20. Lokal og digital. Et sammenhængende danmark. Den fælleskommunale digitaliseringsstrategi og handlingsplan 2016-2020. : Kommunernes Landsforening; 2016 URL: https://www.kl.dk/media/l4daobww/lokal-og-digital-et-sammenhaengende-danmark.pdf [accessed 2025-12-12]
21. Sundhedsstrukturkommissionens rapport. Beslutningsgrundlag for et mere lige, sammenhængende og bæredygtigt sundhedsvæsen. : Indenrigs- og Sundhedsministeriet; 2024 URL: https://www.ism.dk/Media/638545635292256419/Hovedrapport-tilg%c3%a6ngelig-fil.pdf [accessed 2025-12-12]
22. Cresswell KM, Bates DW, Sheikh A. Ten key considerations for the successful implementation and adoption of large-scale health information technology. J Am Med Inform Assoc 2013 Jun;20(e1):e9-e13. [doi: 10.1136/amiajnl-2013-001684] [Medline: 23599226]
23. Mair FS, May C, O'Donnell C, Finch T, Sullivan F, Murray E. Factors that promote or inhibit the implementation of e-health systems: an explanatory systematic review. Bull World Health Organ 2012 May 1;90(5):357-364. [doi: 10.2471/BLT.11.099424] [Medline: 22589569]
24. Standing C, Standing S, McDermott ML, Gururajan R, Kiani Mavi R. The paradoxes of telehealth: a review of the literature 2000–2015. Syst Res 2018 Jan;35(1):90-101. [doi: 10.1002/sres.2442]
25. Greenhalgh T, Maylor H, Shaw S, et al. The NASSS-CAT tools for understanding, guiding, monitoring, and researching technology implementation projects in health and social care: protocol for an evaluation study in real-world settings. JMIR Res Protoc 2020 May 13;9(5):e16861. [doi: 10.2196/16861] [Medline: 32401224]
26. Greenhalgh T, Robert G, Macfarlane F, Bate P, Kyriakidou O. Diffusion of innovations in service organizations: systematic review and recommendations. Milbank Q 2004;82(4):581-629. [doi: 10.1111/j.0887-378X.2004.00325.x] [Medline: 15595944]
27. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023 Sep 18;6(1):161. [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
28. Scott Kruse C, Karem P, Shifflett K, Vegi L, Ravi K, Brooks M. Evaluating barriers to adopting telemedicine worldwide: a systematic review. J Telemed Telecare 2018 Jan;24(1):4-12. [doi: 10.1177/1357633X16674087] [Medline: 29320966]
29. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146. [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
30. Gagnon MP, Desmartis M, Labrecque M, et al. Systematic review of factors influencing the adoption of information and communication technologies by healthcare professionals. J Med Syst 2012 Feb;36(1):241-277. [doi: 10.1007/s10916-010-9473-4] [Medline: 20703721]
31. Sligo J, Gauld R, Roberts V, Villa L. A literature review for large-scale health information system project planning, implementation and evaluation. Int J Med Inform 2017 Jan;97:86-97. [doi: 10.1016/j.ijmedinf.2016.09.007] [Medline: 27919399]
32. Brewster L, Mountain G, Wessels B, Kelly C, Hawley M. Factors affecting front line staff acceptance of telehealth technologies: a mixed-method systematic review. J Adv Nurs 2014 Jan;70(1):21-33. [doi: 10.1111/jan.12196] [Medline: 23786584]
33. Lau R, Stevenson F, Ong BN, et al. Achieving change in primary care—causes of the evidence to practice gap: systematic reviews of reviews. Implement Sci 2015 Dec;11(1):40. [doi: 10.1186/s13012-016-0396-4]
34. Cinar E, Trott P, Simms C. A systematic review of barriers to public sector innovation process. Public Manage Rev 2019 Feb;21(2):264-290. [doi: 10.1080/14719037.2018.1473477]
35. De Vries H, Bekkers V, Tummers L. Innovation in the public sector: a systematic review and future research agenda. Public Adm 2016 Mar;94(1):146-166. [doi: 10.1111/padm.12209]
36. Torugsa N, Arundel A. Complexity of Innovation in the public sector: a workgroup-level analysis of related factors and outcomes. Public Manage Rev 2016 Mar 15;18(3):392-416. [doi: 10.1080/14719037.2014.984626]

XSL•FO
RenderX

37. Borins SF. The persistence of innovation in government. : Brookings Institution Press in Collaboration with Ash Center for Democratic Governance and Innovation; 2014 URL: https://businessofgovernment.org/sites/default/files/The%20Persistence%20of%20Innovation%20in%20Government.pdf [accessed 2025-12-12]
38. Eriksen S, Dahler AM, Øye C. Improvisational theatre as a viable path? Exploring interaction antecedents in public service innovation processes attempting to implement and sustain eHealth solutions. Public Manage Rev 2025 Mar 4;27(3):817-835. [doi: 10.1080/14719037.2024.2381069]
39. Hadjimanolis A. The barriers approach to innovation. In: Shavinina LV, editor. The International Handbook on Innovation: Elsevier; 2003:559-573.
40. Nilsen ER, Dugstad J, Eide H, Gullslett MK, Eide T. Exploring resistance to implementation of welfare technology in municipal healthcare services—a longitudinal case study. BMC Health Serv Res 2016 Nov 15;16(1):657. [doi: 10.1186/s12913-016-1913-5] [Medline: 27846834]
41. Cinar E, Trott P, Simms C. An international exploration of barriers and tactics in the public sector innovation process. Public Manage Rev 2021 Mar 4;23(3):326-353. [doi: 10.1080/14719037.2019.1668470]
42. Cucciniello M, Guerrazzi C, Nasi G, Ongaro E. Coordination mechanisms for implementing complex innovations in the health care sector. Public Manage Rev 2015 Aug 9;17(7):1040-1060. [doi: 10.1080/14719037.2015.1029348]
43. Cucciniello M, Nasi G. Evaluation of the impacts of innovation in the health care sector: a comparative analysis. Public Manage Rev 2014 Jan 2;16(1):90-116. [doi: 10.1080/14719037.2013.798026]
44. Cresswell K, Sheikh A. Organizational issues in the implementation and adoption of health information technology innovations: an interpretative review. Int J Med Inform 2013 May;82(5):e73-e86. [doi: 10.1016/j.ijmedinf.2012.10.007] [Medline: 23146626]
45. Bugge MM, Bloch CW. Between bricolage and breakthroughs—framing the many faces of public sector innovation. Public Money Manage 2016 Jun 6;36(4):281-288. [doi: 10.1080/09540962.2016.1162599]
46. Fuglsang L, Sørensen F. The balance between bricolage and innovation: management dilemmas in sustainable public innovation. Serv Ind J 2011 Mar;31(4):581-595. [doi: 10.1080/02642069.2010.504302]
47. Fuglsang L. Bricolage and invisible innovation in public service innovation. J Innov Econ Manage 2010;5(1):67-87. [doi: 10.3917/jie.005.0067]
48. Gibson G, Dickinson C, Brittain K, Robinson L. Personalisation, customisation and bricolage: how people with dementia and their families make assistive technology work for them. Ageing Soc 2019 Nov;39(11):2502-2519. [doi: 10.1017/S0144686X18000661]
49. Greenhalgh T, Wherton J, Sugarhood P, Hinder S, Procter R, Stones R. What matters to older people with assisted living needs? A phenomenological analysis of the use and non-use of telehealth and telecare. Soc Sci Med 2013 Sep;93:86-94. [doi: 10.1016/j.socscimed.2013.05.036]
50. Houtgraaf G, Kruyen PM, van Thiel S. Public sector creativity as the origin of public sector innovation: a taxonomy and future research agenda. Public Adm 2023 Jun;101(2):539-556. [doi: 10.1111/padm.12778]
51. Yin RK. Case study research. In: Design and Methods: Sage Publications; 2009.
52. Flyvbjerg B. Five misunderstandings about case-study research. Qual Inq 2006 Apr;12(2):219-245. [doi: 10.1177/1077800405284363]
53. Eriksen S, Dahler AM, Øye C. The informal way to success or failure? Findings from a comparative case study on video consultation training and implementation in two Danish hospitals. BMC Health Serv Res 2023 Oct 21;23(1):1135. [doi: 10.1186/s12913-023-10163-w] [Medline: 37865741]
54. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Aug 8;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
55. Braun V, Clarke V, Hayfield N, Terry G. Thematic analysis. In: Liamputtong P, editor. Handbook of Research Methods in Health Social Sciences: Springer; 2019:843-860.
56. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
57. Levi-Strauss C. The Savage Mind: University of Chicago Press; 1966.
58. Goffman E. The Presentation of Self in Everyday Life: Doubleday; 1959.
59. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated Consolidated Framework for Implementation Research based on user feedback. Implement Sci 2022 Oct 29;17(1):75. [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]
60. Malterud K. Qualitative research: standards, challenges, and guidelines. Lancet 2001 Aug;358(9280):483-488. [doi: 10.1016/S0140-6736(01)05627-6]
61. Johnson C. Bricoleur and bricolage: from metaphor to universal concept. Paragraph 2012;35(3):355-372. [doi: 10.3366/para.2012.0064]
62. Baker T, Nelson RE. Creating something from nothing: resource construction through entrepreneurial bricolage. Adm Sci Q 2005 Sep;50(3):329-366. [doi: 10.2189/asqu.2005.50.3.329]

XSL•FO
RenderX

63. Nelson RE, Rodriguez-Lluesma C, Companys YE, Stinchfield BT. Contextualizing the subjectivist-objectivist debate in entrepreneurship using engineering, art, craft, and bricolage. Int Entrep Manag J 2018 Dec;14(4):999-1021. [doi: 10.1007/s11365-017-0471-6]
64. Greenhalgh T, Shaw S, Wherton J, et al. SCALS: a fourth-generation study of assisted living technologies in their organisational, social, political and policy context. BMJ Open 2016 Feb 15;6(2):e010208. [doi: 10.1136/bmjopen-2015-010208] [Medline: 26880671]
65. Paparini S, Green J, Papoutsi C, et al. Case study research for better evaluations of complex interventions: rationale and challenges. BMC Med 2020 Nov 10;18(1):301. [doi: 10.1186/s12916-020-01777-6] [Medline: 33167974]
66. Wong G, Greenhalgh T, Westhorp G, Pawson R. Realist methods in medical education research: what are they and what can they contribute? Med Educ 2012 Jan;46(1):89-96. [doi: 10.1111/j.1365-2923.2011.04045.x] [Medline: 22150200]
67. Greenhalgh T, Papoutsi C. Studying complexity in health services research: desperately seeking an overdue paradigm shift. BMC Med 2018 Jun 20;16(1):95. [doi: 10.1186/s12916-018-1089-4] [Medline: 29921272]
68. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
69. Kamp A, Hansen AM. Negotiating professional knowledge and responsibility in cross-sectoral telemedicine. Nordic J Work Life Stud 2019;9(S5):13-32. [doi: 10.18291/njwls.v9iS5.112691]

Abbreviations

e-locks: electronic door locks
ICT: information and communication technology

Edited by A Stone, A Mavragani; submitted 02.Jul.2025; peer-reviewed by N Döring, S Kuoppamaki; accepted 28.Oct.2025; published 07.Jan.2026.

Please cite as:
Eriksen S, Øye C, Dahler AM
Exploring the Dynamics of Actors, Structural Factors, and Bricolage in the Implementation and Sustainability of eHealth Solutions: Qualitative Multiple-Case Study
J Med Internet Res 2026;28:e79999
URL: https://www.jmir.org/2026/1/e79999
doi:10.2196/79999

© Susanne Eriksen, Christine Øye, Anne Marie Dahler. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

“I Want to Spend My Time Living”—Experiences With a Digital Outpatient Service With a Mobile App for Tailored Care Among Adults With Long-Term Health Service Needs: Qualitative Study Using Thematic Analysis

Heidi Holmen[1,2], RN, PhD; Erik Fosse[1,3], MD, PhD

[1]The Intervention Centre, Oslo University Hospital, Oslo, Norway

[2]Department of Nursing and Health Promotion, Faculty of Health Sciences, OsloMet—Oslo Metropolitan University, Oslo, Norway

[3]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway

Corresponding Author:
Heidi Holmen, RN, PhD
The Intervention Centre
Oslo University Hospital
Postboks 4950 Nydalen
Oslo, 0424
Norway
Phone: 47 91527700
Email: heidho@ous-hf.no

Abstract

Background: Digital health services are increasingly used in hospital-based outpatient care, offering remote monitoring, patient-reported outcomes, information sharing, and asynchronous communication. While expected to improve self-management, timeliness, and efficiency, the success of digital health interventions relies on patients’ health literacy and digital health literacy. While some research has addressed potential associations between digital health interventions and patients’ health outcomes, research on patients’ experiences remains limited.

Objective: The aim of this study was to explore and gain in-depth knowledge about the experiences of patients with chronic or long-term conditions enrolled in a 6-month digital outpatient care intervention for tailored care and health literacy.

Methods: We conducted an exploratory qualitative interview study with 17 strategically recruited adult patients with cancer, interstitial lung disease, epilepsy, or complicated pain who used a digital outpatient service for 6 months. Individual telephone interviews were conducted using a semistructured guide, transcribed verbatim, and analyzed with thematic analysis to generate codes and themes. Participants had a median age of 62 years (minimum-maximum 36-83 years), with 8 females and 9 males.

Results: The thematic analysis led to 1 main theme “Digital outpatient care as a flexible service supporting patients’ self-management,” informed by 3 subthemes “The ongoing nature of managing a chronic condition and how the digital service meet the patients’ desire for autonomy in their care,” “Digital tools flexibly address the patients’ unique needs, but reliability depends on patient interaction,” and “Digital services enhance the patients’ sense of safety through easy access to a relation with competent healthcare workers.” The themes highlight patients’ appreciation for greater flexibility in their care and their desire to self-manage with the support of easily accessible health care workers, ultimately supporting their health literacy. Patients recognized the importance of actively engaging with the digital solution to fully benefit from its opportunities and emphasized the critical role of health care workers in fostering their sense of security.

Conclusions: Digital outpatient care was experienced as flexible and supportive for patients with long-term conditions. The increased possibility of interacting with health care workers was welcomed by the patients, and the combination of flexibility, self-monitoring, and addressing concerns regarding their self-management may increase the patients experience of autonomy. As health literacy likely plays a role in patients’ ability to effectively engage with digital tools and self-manage their conditions, future research should explore how varying levels of health literacy influence these outcomes. In addition, research should address whether such digital outpatient clinics are positive for a wider range of patients, associated health outcomes, and any positive effects on a health system level.

Trial Registration: ClinicalTrials.gov NCT05068869; https://clinicaltrials.gov/ct2/show/NCT05068869

XSL•FO
RenderX

(J Med Internet Res 2026;28:e79155) doi:10.2196/79155

KEYWORDS

digital outpatient care; mHealth; patient-centered care; patient experiences; digital outpatient services

Introduction

The increasing demand for health care services is not matched by available resources, including workforce shortages [1,2]. To enhance the sustainability of health services, greater flexibility is needed to address the dynamic needs of patients with long-term conditions, aligning with patient-centered care and optimizing resource use. Digital tools have been proposed as a means to achieve this flexibility, offering functionalities such as patient monitoring, self-reporting of objective and subjective data, asynchronous communication, and video consultations [3]. These tools can help identify patients at risk for deterioration or in need of immediate care, while also detecting those stable enough to delay scheduled services. By supporting patient-centered care, digital health solutions may improve attendance at scheduled appointments, enhance symptom management, and empower patients in self-management.

Digital health care services typically enable the subjective reporting of health parameters through patient-reported outcome measures [4]. These measures can aid self-management and support health literacy [5] and facilitate communication between patients and health care workers [6]. Furthermore, digital engagement, self-monitoring, and data sharing offer several advantages for patients [3,7,8]. Recent studies suggest that digital tools in outpatient care for patients with long-term conditions, such as cancer, epilepsy, interstitial lung disease, and musculoskeletal pain, can prevent complications, encourage patient engagement, and increase confidence and autonomy [9-12]. However, research on multicomponent digital solutions—encompassing patient-reported outcome measures, asynchronous messaging, remote monitoring, patient notifications, and video consultations—is limited. In addition, one important prerequisite exists for patients to benefit from digital health services, namely, that the patients use the services as intended. Patients' use of a digital health services might rely on their health literacy and ability to self-manage, and a certain level of digital skills and digital health literacy is also needed [5]. Health literacy provides a theoretical lens for this study, as it is a foundational factor influencing how patients access, understand, and use digital health services. Defined as "the cognitive and social skills that determine the motivation and ability of individuals to gain access to, understand, and use information in ways which promote and maintain good health" [13], health literacy not only affects patients' capacity for self-management but also plays an integral role in their engagement with digital tools. Often, higher health literacy is associated with an increased benefit of digital health interventions [14], while lower health literacy is linked to poorer health outcomes and less digital solution usage [15,16]. Furthermore, adequate health literacy is key to optimizing the use of digital health solutions, as it is suggested to enhance self-management in chronic conditions [5,16-18]. Although several systematic reviews have explored how to define, measure, and understand health literacy and digital health literacy [17,19-24], linking this knowledge to the experience of patients participating in digital health interventions remains scarce. Digital health literacy has been suggested as a super determinant of health comprising more than simply digital literacy and health literacy [25]; thus, exploring the experiences of patients with a digital health intervention supporting flexibility and health literacy might contribute with knowledge on how to research and design more effective, inclusive, and patient-centered digital health solutions that address diverse needs and promote equitable health outcomes.

While research on digital health services expands, research on patients' experiences of participating and making use of integrated digital solutions in outpatient care is lacking. The meaningfulness of a digital health solution depends on factors such as its usability, clinical relevance, convenience, and evidence-based design [6]. Studies on video consultations suggest how they are more convenient and time-saving for patients, but they are not perceived as superior to traditional in-person consultations [26,27]. Thus, despite growing interest in specific features of digital outpatient care and to some extent, the combination of features, previous research remains limited concerning the patients' experiences on a comprehensive model for digital outpatient care. By adopting health literacy as a guiding framework, the aim of this study was to explore and gain in-depth knowledge about the experiences of patients with chronic or long-term conditions enrolled in a digital 6-month outpatient care intervention.

Methods

Study Design

We conducted an explorative qualitative interview study as part of a larger study evaluating the use of digital outpatient care for adults with cancer, interstitial lung disease, epilepsy, or a complex pain condition at 2 university hospitals in Norway [28,29]. The study is reported according to the COREQ (Consolidated criteria for Reporting Qualitative research) guidelines [30].

Setting

The setting for this study was the 4 outpatient services participating in a multicenter controlled trial evaluating digital outpatient services, comprising the Department of Respiratory Diseases, the Department of Neurology, and the Department of Pain Management at Oslo University Hospital, as well as the Department of Cancer at the University Hospital of North Norway [28,29]. All patients included in this trial were living at home, receiving outpatient care, and otherwise managing their own lives.

Eligible participants had been included in the digital outpatient intervention, previously detailed [28,29]. The core focus of the intervention was to improve outpatient service accessibility,

XSL•FO
RenderX

allowing the patients to engage with patient-reported outcome measures, self-monitor parameters, and communicate asynchronously with health care workers. Designated health care workers assessed the patients' responses and parameters and had asynchronous digital contact, aligning with a conceptual model linking health literacy to service access and self-management [31]. Health care workers assigned tasks to patients and set individual thresholds using a traffic light model. Notifications alerted health care workers of deviations such as increased pain, side effects, or uncompleted tasks. Patients received reminders for unfinished tasks. Two-way messaging enabled asynchronous communication for questions, information exchange, and sharing of treatment plans. The digital outpatient intervention was enabled through the Dignio Connected Care platform, allowing personalized patient–health care worker contact [32] (Multimedia Appendices 1 and 2). The platform, including the health care worker's Dignio Prevent system and the patient's MyDignio app, is Conformité Européenne marked, privacy compliant, and globally applied [32].

Recruitment and Study Participants

Patients in the intervention arm of the overall multicenter controlled trial were invited to be interviewed when they completed their 6-month follow-up questionnaire at the end of the digital intervention. One item of the questionnaire asked for their consent to be contacted by a researcher with information regarding the qualitative interviews. An a priori estimation suggested a sample of 12-15 patients from the intervention group [28]. A strategic recruitment among those accepting to be contacted was conducted to ensure a broad sample of interviewees, enabling our aim of collecting rich data on the experiences of using the digital outpatient care intervention. Oral information was given over the phone, prior to written information in a secure digital platform, also allowing for their digital consent. We made use of the secure service "Nettskjema" digital consent in a service for sensitive data (Tjeneste for Sensitive Data [TSD] in Norwegian: the English version is Service for Sensitive Data]) developed at the University of Oslo. TSD is designed for storing and processing sensitive data in compliance with the Norwegian "Personal Data Act" and "Health Research Act." The consent invitations were sent to the participants through the Pretty Good Privacy encrypted version of the University of Oslo web questionnaire service "Nettskjema" demanding a governmental ID portal for login. A total of 91 patients completed their 6-month questionnaire, and of these, 50 agreed to be contacted for a possible interview. A researcher consecutively approached 21 patients, and 18 patients provided their digital consent to be interviewed.

Data Collection

To provide the information most valuable for the study aim, an interview guide was developed based on reviewing previous research in the field, inspired by topics on innovation assessment [33], and by adding items regarding specific factors associated with the current intervention [28]. A draft for the interview guide was developed by the first author and reviewed prior to a satisfactory version for a pilot interview. No changes were made to the interview guide after the pilot interview, and the pilot interview was included in the final analysis. The interview guide (Multimedia Appendix 3) contained an introduction with questions on their use of the digital service in general, before a midpart with more detailed questions on their experience with intervention, and finally some summarizing questions allowing final reflections. The interviewer (HH) has a background as a nurse with a PhD in health service research, without any personal or professional relationship with the participants. She has conducted qualitative research and analysis earlier and led the intervention study in which this substudy arises from.

All interviews were conducted over the telephone. All participants used their private smartphones for the interviews, while the interviewer (HH) used a work-related phone, with only voice. The "Nettskjema" Dictaphone app was used to record all interviews, with 1 main recorder and a backup recorder. All recordings were satisfactory besides one, where the line was interrupted midway through, and a new recording was started. No data were lost in this recording.

All interviews were conducted between 9 AM and 5 PM, with participants offered multiple time slots to ensure convenience. The interviewer called the participants at the scheduled time, and of those consenting, only 1 never responded despite repeated attempts from the researcher. Thus, 17 interviews were conducted with consenting participants, of whom 5 had cancer, 5 had chronic pain, 4 had interstitial lung disease, and 3 had epilepsy (Table 1). All participants were interviewed only once after their 6-month follow-up, and saturation was achieved through the 17 interviews; thus, no participants were added to the interview sample. The median time from the 6-month follow-up questionnaire end to interview was 40 days (minimum-maximum 9-95 days). The interviews lasted from 15 minutes to the maximum of 46 minutes, with a median interview time of 26 minutes. Median age of the interviewed participants was 62 years (minimum-maximum 36-83 years), of whom 8 were females and 9 were males. The participants had some variation in how they had used the digital intervention, and only 1 participant was categorized as a low user (Table 2).

Table 1. Participants, age groups, gender, department, and interview details.

Age groups (years)	N (%)	Sex, female/male	Departments	Minutes of interview, median (minimum-maximum)
31-50	3 (17)	2/1	Neurology and pain management	27 (26-27)
51-60	4 (24)	2/2	Cancer, respiratory diseases, and neurology	24 (16-39)
61-70	6 (35)	2/4	Cancer, respiratory diseases, neurology, and pain management	22 (15-46)
71-90	4 (24)	2/2	Cancer and respiratory diseases	29 (17-33)

XSL•FO
RenderX

Table 2. Use of the digital intervention (N=17).

Intervention	Values
Digital interaction[a]	
Total number of digital interactions	
Mean (SD)	58.8 (55.6)
Median (minimum-maximum)	65 (3-157)
Dichotomized digital interaction[b]	
Low use, n (%)	1 (6)
High use, n (%)	16 (94)
Asynchronous chat messages	
Messages sent from patient	
Mean (SD)	8.4 (9.2)
Median (minimum-maximum)	4 (0-31)
Messages sent from health care worker	
Mean (SD)	8.2 (7.5)
Median (minimum-maximum)	3 (0-20)
Total number of messages	
Mean (SD)	16.6 (16.2)
Median (minimum-maximum)	12 (0-51)
PRO [c] **measures**	
PRO measures sent to patient	
Mean (SD)	23.7 (25.4)
Median (minimum-maximum)	12 (0-82)
PRO measure responses from patient	
Mean (SD)	21.4 (24.7)
Median (minimum-maximum)	11 (0-82)

[a]Total counts were computed for chat messages, video visits, completed PRO measures, and monitoring events.

[b]Low users responded to less than 30% of the expected PRO measures.

[c]PRO: patient-reported outcome.

Data Analysis

All interviews were transcribed verbatim using the F4-transcript software (Audiotranskription.de) in TSD by the first author (HH). Then, the transcripts was analyzed using thematic analysis [34], with the following steps: (1) familiarize with data, (2) generate initial codes, (3) search for themes, (4) review themes, (5) define and name the themes, and (6) produce the report. An inductive approach was used, with the research question at the outset, allowing an analysis on the premises of the data at hand [34]. Both authors read and reread the material and HH generated initial and preliminary codes of the material—interview by interview. These preliminary codes were later reviewed by EF, before they were discussed, reviewed, and refined. Through discussion, the 2 authors (HH and EF) suggested themes to describe and clarify the codes, with particular attention given to exploring how patients' interactions with the digital health intervention may reflect aspects of their ability to access, understand, and use health information and the digital health tool according to health literacy and digital health literacy. A final set of themes, codes, and corresponding quotes was then summarized. All interviews were conducted, transcribed, and analyzed in Norwegian. The quotes have been translated to English and reviewed to assess the intended meaning. Minor alterations were made to ensure anonymity among the participants.

Ethical Considerations

The regional ethical committee in Norway prereviewed the protocol and judged the project as outside its mandate according to the Norwegian Health Research Act (regional ethical committee south-east reference number 252051). The project was approved by the data protection officer at Oslo University Hospital (reference 21/06826) and Northern Norway University Hospital (reference 2021/4942). All participants signed a digital, written informed consent form before participating in the project. This study was conducted in accordance with the Declaration

XSL•FO
RenderX

of Helsinki and all data were deidentified to ensure privacy and confidentiality. No compensation was provided to participants.

Results

Summary of the Themes and Codes

The thematic analysis led to 1 main theme "Digital outpatient care as a flexible service supporting patients' self-management," informed by 3 subthemes (Table 3).

XSL•FO
RenderX

Table 3. Themes and corresponding codes with quotes from the thematic analysis. Theme: digital outpatient care as a flexible service supporting patients' self-management.

Subthemes and codes	Quotes
The ongoing nature of managing a chronic condition and how the digital service meets the patients' desire for autonomy in their care	
Living with a chronic condition necessitates ongoing self-management	• "When you live with a diagnosis like this, it's not something you can easily forget" [P17]. • "I have had some quite big seizures, and little episodes that I don't really know whether are related to these seizures and that worries me" [P19].
Flexible services are required to support self-management	• "It allows me to better track myself and makes me calmer in relation to the disease progression, and I believe it enables me to respond at an earlier stage than I otherwise might have if something were to happen that affects my situation. So, the motivation lies in the ability to cope with a difficult disease, I would say. There's also additional motivation in that I enjoy numbers, apps, and correlations" [P17]. • "My motivation for using the app was the possibility to ask questions, and perhaps add notes to my symptoms, to be able to see some patterns" [P29].
Making use of digital tools for self-management supports autonomy in the patients' role	• "With this app, being able to see that the results sometimes go down a bit, and if you as a patient don't feel particularly worse, you might wait until the next measurement to make contact because you can relate this to your current situation and how you feel then, which is definitely possible with my disease" [P17]. • "Sometimes, I feel I am in bad shape, but the app says I am not, and that makes me happy...because there's so many things affecting my breath, like my state of mind and all" [P11].
Digital tools flexibly address the patients' unique needs but reliability depend on patient interaction	
Digital tools provide new solutions to real needs in a flexible manner	• "I believe that, for my part, the frequency of contact with the healthcare system has decreased since I started using the app" [P17]. • "I find the app to be a part of the entire [outpatient care] system" [P36].
Enabling interaction through communication of unique needs beyond standardized formats	• "The app asked more thorough questions than I remember being asked in person" [P12]. • "You do have specific alternatives to respond to, but perhaps some lines where you could express something? It's useful for the users, but whether it's also useful for you as the recipient, one might question, but maybe just 3-4 lines" [P36].
Making sure digital tools are suitable for the user	• "It is indeed an advantage to have digital contact for those of us who are digital and can use such things. But it is important that the digital solutions do not replace physical attendance and postal mail, but come as an addition, that one can choose" [P6].
Digital services enhance the patients' sense of safety through easy access to a relation with competent health care workers	
Easy and seamless contact with the outpatient clinic	• "But the feeling of safety, the certainty that I could get hold of someone if there was something, that gives a lot of peace of mind for my part" [P32]. • "As I am working full time, it's a hassle to leave work to go to the clinic, and it's adding more stress because I can't do my job properly" [P29].
Digital services facilitate patient access to health care, while supporting the health care workers' workflow	• "It is a reassurance for me if I would need it, and two years ago, I had a question that I didn't know whether was important and it took a month to get an answer. If we could have done it via the app, it would have been faster for both of us instead of having to call repeatedly" [P19].
Building a trustful digital relation to health care workers	• "Yes, I responded twice a week to the nurse and if I didn't respond, I received a message (...) so it worked perfectly (...) but I have never spoken directly with the nurse, which by the way is exceptionally pleasant to communicate with digitally" [P22]. • "I did know the doctor from before, but not the nurse – and not that it matters because now I feel like I know the nurse too, and that gives me a sense of safety" [P37].

XSL•FO
RenderX

The Ongoing Nature of Managing a Chronic Condition and How the Digital Service Meets the Patients' Desire for Autonomy in Their Care

The participants acknowledged the burden of living with a chronic condition. While some felt that they already understood the required self-management, they emphasized that digital follow-up provided a sense of support that lightened the burden, allowing them to focus more on living rather than being ill. One participant remarked, "Because I don't want to spend too much time on my condition, I want to spend my time on life instead, right? If it becomes too much, then it becomes too much concentration on the condition, and I still think that I'm not particularly sick" [P12].

The importance of self-management was highlighted, with some participants noting that this was the first time they had access to a tool that truly supported their self-management effort and the ability to monitor deterioration as highly valuable. The option to send messages, even for minor questions, was particularly appreciated, as these small clarifications provided reassurance: "Even small questions can take up a lot of space" [P34]. This access to advice enhanced their sense of control, and participants found it easier to consult health care workers through the digital platform compared with standard care.

The patients acknowledged their frequent need for health services and the challenges of contacting health care workers in outpatient care. They shared how digital outpatient care addressed this issue by improving communication and accessibility. Patients valued the efficiency and convenience offered by the digital solution, as one participant shared: "So, for me, this worked very well because I don't need to go in just for questions. I wasn't scared of anything, and I didn't have to go anywhere, and I didn't have to spend time calling, so it worked very well for me. Then you can rather go in if you don't achieve your goals with the digital follow-up" [P29]. While the benefits of digital services were valued, the patients emphasized the need for a hybrid model with digital and physical contact. They supported innovation in health services, particularly when it contributed to flexible, individualized care and expressed positivity toward continued digital contact after the project.

Patients found the digital outpatient care tools valuable for tracking health parameters, identifying changes in their condition, thus enhancing their sense of safety. The tools were used less when their condition was stable but became essential when they suspected health changes. Through the app, patients reported gaining an understanding of their symptoms and a reflection on their condition, which they felt had an impact on their health literacy and self-management skills. In addition, the tools facilitated greater involvement of family members, providing an extra layer of support. As one patient shared: "Yes! It's always a big moment every time I take such tests, and the people around me are always interested in how it's going. They always ask how it went when I took these tests" [P17]. This family engagement was described as a motivating factor for using the digital platform. However, some patients noted that their use of the platform was independent of family involvement, reflecting diverse approaches to digital health engagement.

Digital Tools Flexibly Address the Patients' Unique Needs, But Reliability Depends on Patient Interaction

Most patients valued the flexibility offered by digital tools, particularly the convenience of receiving care without traveling to the hospital. This was especially important for those with long travel distances or conditions that made travel burdensome. Digital tools were described as time-saving, stress-reducing, and helpful for minimizing work absences, with the ease of use and sense of security further enhancing their appeal: "If something comes up in the evening, I can enter it and then I get a response in the morning...much more frequent contact than usual" [P37]. Patients appreciated the opportunity for more flexible interactions with health care workers.

Patient-reported outcomes were generally seen as relevant and easy to complete, allowing for more thorough and individualized health reporting than traditional consultations. However, some patients felt that standardized questions and response categories did not fully capture their experiences, expressing a need for more flexibility and individual tailoring. One participant reflected: "For many of the questions I get [through the app], there are some I don't get to elaborate enough on, and it could just be that the information that I have could be valuable to you" [P27]. Suggestions for improvement included displaying scheduled consultations and test results in the app, reducing questionnaire frequency, adding time estimates for completion, and allowing more free-text responses. Patients also expressed trust in the app's data security but emphasized that not all data needed to be shared digitally to minimize risks.

Prerequisites for using digital tools, such as internet access, smartphones, and digital identification, were noted. Most found the app easy to navigate, with minimal training needed, although those using medical monitoring devices appreciated some initial guidance. Initial stress in using the app subsided with familiarity, as one patient noted: "At first, I was a bit stressed by it [the app], when the messages came from the hospital that I had to do these measurements, I felt that I was getting stressed. But why that is, I don't know, it's going well now. I was a bit surprised at myself, like, why is it stressing me, but it has passed now so that's good because now I'm starting to get the hang of it" [P11]. However, digital tools were not universally suitable, with concerns raised about older adult users, individuals with health-related anxiety, and the timing of app prompts for working patients. Despite these concerns, reminders were often described as helpful for completing tasks.

Digital Services Enhance the Patients' Sense of Safety Through Easy Access to a Relation With Competent Health Care Workers

The patients appreciated the sense of security offered by the user-friendly and flexible digital contact with health care workers in outpatient care. The app was seen as supportive rather than burdensome, with some noting that spending a few minutes weekly was a small effort for managing a long-term condition: "It's not like the app reminds you that you have a chronic condition, it's more perceived as a positive element in addition to the treatment you're already receiving" [P17]. Digital reporting was practical and reduced errors, although patients cautioned against relying on technology for false security.

XSL•FO
RenderX

Avoiding phone queues and callbacks saved time and reduced stress for patients, while also improving workflow for health care workers by allowing responses on their schedule. One participant shared: “I liked that I seemed to have better access to the department than normal, it was less burdensome than calling in the limited time the reception is open, and sometimes it rings and rings and rings and rings, and sometimes dealing with less or even unqualified reception workers which I have had problems with before” [P19]. Asynchronous messaging was particularly valued for enabling prompt answers to both urgent and nonurgent questions, fostering a sense of being seen and supported. Even when unused, the messaging function reassured patients that they could reach skilled health care workers if needed.

The competence of health care workers responding to queries was trusted, and while some patients valued consistent interactions with the same person, others prioritized competence over familiarity, especially with nurses. Patients appreciated building relationships through physical consultations before transitioning to digital follow-up but found digital contact effective once trust was established, as one patient elaborated: “And yes, I’ve met them, they know me and I have a face to them, and I think that’s very good. That it’s the same people who follow up. But it’s mostly on the doctor's side that I need to have the same person, whom I trust, and I have met in consultations, who knows me. But with the nurses, I think it’s very good to have two people” [P14]. Administrative workers triaging questions were considered acceptable, provided medical issues were handled by qualified professionals. However, some patients still preferred direct calls to familiar health care workers, especially if they already had access to adequate resources.

Discussion

Principal Findings

The patients experienced digital outpatient care as a flexible service supporting self-management, which fulfilled the patients’ ongoing but unmet need of autonomy in their self-management. The findings suggest that aspects of health literacy, such as the ability to understand, engage with, and act on health information, may have influenced how patients interacted with the digital platform and benefited from the service. To achieve a truly flexible and reliable service, active interaction between patients and health care workers with the digital platform is essential to strengthen the patients’ sense of safety. The patients’ reflections regarding the digital outpatient service in this study highlight key insights worth further discussion.

From the patients’ perspective, a digital outpatient solution offers greater flexibility in health care services than traditional calendar-based approaches, which rely solely on phone calls or scheduled consultations to access the outpatient clinic. Previous research has described how traditional systems are rigid and allow little patient-centered and flexible approaches [35]. If changes in the patients’ clinical status occur, there is little room for a rapid response or to move forward a scheduled appointment to curb the deterioration. However, the digital opportunities include flexibility through asynchronous messaging, monitoring, and patient responses—offering a more patient-centered approach, supporting the patient’s self-management [28]. Furthermore, health literacy is often intertwined with self-management, and both concepts are positively and iteratively influencing each other [16,31]. This highlights how digital outpatient solutions not only enhance the patients’ experience of more flexibility and responsiveness but also foster patient-centered care by supporting self-management and strengthening health literacy in a mutually reinforcing manner.

While our study and similar research highlight the benefits of digital approaches [5,10-12,18], some patients may feel burdened by the added responsibility, potentially threatening their sense of security. One patient described initial anxiety when using the digital platform, fearing mistakes in reporting—a concern echoed in previous studies emphasizing the critical role of usability and adequate digital and health literacy [36,37]. However, with practice, participants gained confidence in self-management and their sense of insecurity decreased. Interaction with health care workers further supported this transition, providing timely guidance and encouragement through digital messaging. This highlights the importance of tailoring support to help reluctant patients build the necessary skills and competencies for digital outpatient care [5,18]. In our study, patients demonstrated health literacy through their engagement with the intervention, responding accurately to patient-reported outcome measures, conducting remote monitoring as instructed, and framing thoughtful questions via the messaging functions, showcasing the integral link between health literacy and effective self-management.

The flexibility and interactive opportunities provided by the digital platform highlight its potential to reduce the traditional asymmetry between patients and health care providers. Unlike calendar-based systems in traditional care, which often overlook individual patient needs, digital outpatient care allows patients to engage with health care workers when changes in their clinical condition occur. Over time, patients became more adept at interpreting the relationship between their symptoms and objective measurements provided by the system, enabling them to make informed decisions and take appropriate actions with guidance from health care workers [28]. In the digital solution, where interactions occurred based on asynchronous messages, the focus of the interaction was more centered on the patients’ input and experiences rather than being primarily driven by the health care workers [6]. The role of the health care worker is also evolving, much like the changes digital solutions bring to the patient’s role. More flexible ways of interacting with patients are shifting the responsibilities of health care workers, often in ways they are not fully prepared for. However, with the necessary competencies in using digital solutions and a clear understanding of their integration into clinical care, health care workers can positively influence patients’ engagement and effective use of digital health solutions to maintain or improve health [38,39]. While processes that strengthen the patient role are well established in patient-centered care [6], their implementations through digital solutions remain underexplored.

Based on our findings, we argue that digital outpatient care can enhance care quality by providing easier and more timely access to health services, saving time and potentially preventing

XSL•FO
RenderX

symptom deterioration through early identification of changes. Patients' trust in the system and providers, their increased self-management, and reduced feelings of burden—both as individuals living with a chronic condition and as users of health care services—further support this claim. However, the impact of digital outpatient care on health outcomes related to patients' conditions remains unclear. While previous research has shown positive but small and low-graded evidence on such effects [5,7,18], further studies are needed to evaluate both health outcomes and the experiences of patients and health care providers using integrated digital health solutions.

Limitations and Strengths

This study is among the few qualitative investigations into integrated digital health solutions in outpatient care, but it has limitations. Participants were recruited from a study evaluating the same digital platform, with tailored interventions based on clinical needs, leading to heterogeneous experiences. All interviews were conducted using telephone with only voice, and thus, potentially relevant data concerning body language and nonverbal communication were not interpreted. To address this, a semistructured interview guide was used, and while data analysis was comprehensive, recurring similar statements were interpreted as having broader relevance. A key finding was that all patients expressed trust in the digital solution and care provided. However, this may introduce bias, as participants were likely those already comfortable with digital solutions, potentially excluding perspectives from less digitally inclined individuals.

Conclusions

In this study, we aimed to explore the experiences of patients with long-term conditions in using a digital outpatient care service for 6 months. Digital outpatient care was experienced as flexible and supportive for patients with long-term conditions. The increased possibility of interacting with health care workers was welcomed by the patients, and the combination of flexibility, self-monitoring, and addressing concerns regarding their self-management seemed to increase the patients' experience of autonomy. As health literacy likely plays a role in patients' ability to effectively engage with digital tools and self-manage their conditions, future research should explore how varying levels of health literacy influence these outcomes. In addition, research should address whether such digital outpatient clinics are positive for a wider range of patients, associated health outcomes, and any positive effects on a health system level.

Acknowledgments

We would like to thank all the participating patients and health care workers at each outpatient clinic. We are grateful to the staff in Dignio for their valuable insight and practical tailoring of their platform, and for a good collaboration, particularly with Anna Hurrød, Meetali Kakad, and Andreas Norling. Furthermore, we would like to thank Anette Winger and Astrid Torbjørnsen for their valuable comments to subsets of the draft in process. Artificial intelligence–assistive technologies have been applied for language editing only.

Funding

This work was initiated by the Oslo University Hospital, financially supported by the Research Council of Norway (grant 316244), internally funded by Oslo University Hospital and UNN to secure staff, and funded by Dignio Connected Care to finance user licenses in the digital platform. The Research Council of Norway and Dignio Connected Care had no role in preparing this manuscript.

Data Availability

The datasets used and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

HH initiated the study, developed the interview guide, and conducted all the interviews and the thematic analysis before she wrote the manuscript. EF collaborated in the project and gave feedback on the interview guide and the preliminary interviews, took part in the thematic analysis, and revised the manuscript. Both coauthors approved the final manuscript. No authors were added or removed during the process.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Dignio Prevent interface.
[DOCX File , 123 KB - jmir_v28i1e79155_app1.docx]

Multimedia Appendix 2

XSL•FO
RenderX

MyDignio interface.
[DOCX File , 124 KB - jmir_v28i1e79155_app2.docx]

Multimedia Appendix 3
Interview guide.
[DOCX File , 21 KB - jmir_v28i1e79155_app3.docx]

Multimedia Appendix 4
COREQ (Consolidated criteria for Reporting Qualitative research) checklist.
[PDF File (Adobe PDF File), 225 KB - jmir_v28i1e79155_app4.pdf]

References

1. GBD 2019 Demographics Collaborators. Global age-sex-specific fertility, mortality, healthy life expectancy (HALE), and population estimates in 204 countries and territories, 1950-2019: a comprehensive demographic analysis for the Global Burden of Disease Study 2019. Lancet 2020;396(10258):1160-1203 [FREE Full text] [doi: 10.1016/S0140-6736(20)30977-6] [Medline: 33069325]
2. Blumenthal D, Chernof B, Fulmer T, Lumpkin J, Selberg J. Caring for high-need, high-cost patients—an urgent priority. N Engl J Med 2016;375(10):909-911. [doi: 10.1056/NEJMp1608511] [Medline: 27602661]
3. Hjollund NHI, Larsen LP, de Thurah AL, Grove BE, Skuladottir H, Linnet H, et al. Patient-reported outcome (PRO) measurements in chronic and malignant diseases: ten years' experience with PRO-algorithm-based patient-clinician interaction (telePRO) in AmbuFlex. Qual Life Res 2023;32(4):1053-1067 [FREE Full text] [doi: 10.1007/s11136-022-03322-9] [Medline: 36639598]
4. Gandrup J, Ali SM, McBeth J, van der Veer SN, Dixon WG. Remote symptom monitoring integrated into electronic health records: a systematic review. J Am Med Inform Assoc 2020;27(11):1752-1763 [FREE Full text] [doi: 10.1093/jamia/ocaa177] [Medline: 32968785]
5. Verweel L, Newman A, Michaelchuk W, Packham T, Goldstein R, Brooks D. The effect of digital interventions on related health literacy and skills for individuals living with chronic diseases: a systematic review and meta-analysis. Int J Med Inform 2023;177:105114. [doi: 10.1016/j.ijmedinf.2023.105114] [Medline: 37329765]
6. Greenhalgh J, Gooding K, Gibbons E, Dalkin S, Wright J, Valderas J, et al. How do patient reported outcome measures (PROMs) support clinician-patient communication and patient care? A realist synthesis. J Patient Rep Outcomes 2018;2:42 [FREE Full text] [doi: 10.1186/s41687-018-0061-6] [Medline: 30294712]
7. Mandizha J, Lanario JW, Duckworth A, Lines S, Paiva A, Elworthy V, et al. Patient perspectives on home-spirometry in interstitial lung disease: a qualitative co-designed study. BMJ Open Respir Res 2023;10(1):e001837 [FREE Full text] [doi: 10.1136/bmjresp-2023-001837] [Medline: 37793682]
8. de Bell S, Zhelev Z, Shaw N, Bethel A, Anderson R, Thompson Coon J. Remote monitoring for long-term physical health conditions: an evidence and gap map. Health Soc Care Deliv Res 2023;11(22):1-74. [doi: 10.3310/BVCF6192] [Medline: 38014553]
9. Althobiani MA, Evans RA, Alqahtani JS, Aldhahir AM, Russell A, Hurst JR, et al. Home monitoring of physiology and symptoms to detect interstitial lung disease exacerbations and progression: a systematic review. ERJ Open Res 2021;7(4):00441 [FREE Full text] [doi: 10.1183/23120541.00441-2021] [Medline: 34938799]
10. Escriva Boulley G, Leroy T, Bernetière C, Paquienseguy F, Desfriches-Doria O, Préau M. Digital health interventions to help living with cancer: a systematic review of participants' engagement and psychosocial effects. Psychooncology 2018;27(12):2677-2686. [doi: 10.1002/pon.4867] [Medline: 30152074]
11. Hewitt S, Sephton R, Yeowell G. The effectiveness of digital health interventions in the management of musculoskeletal conditions: systematic literature review. J Med Internet Res 2020;22(6):e15617 [FREE Full text] [doi: 10.2196/15617] [Medline: 32501277]
12. Shegog R, Braverman L, Hixson JD. Digital and technological opportunities in epilepsy: toward a digital ecosystem for enhanced epilepsy management. Epilepsy Behav 2020;102:106663. [doi: 10.1016/j.yebeh.2019.106663] [Medline: 31778878]
13. Nutbeam D. Health promotion glossary. Health Promot Int 1998;13(4):349-364. [doi: 10.1093/heapro/13.4.349]
14. Cheng C, Beauchamp A, Elsworth GR, Osborne RH. Applying the electronic health literacy lens: systematic review of electronic health interventions targeted at socially disadvantaged groups. J Med Internet Res 2020;22(8):e18476 [FREE Full text] [doi: 10.2196/18476] [Medline: 32788144]
15. Baccolini V, Rosso A, Di Paolo C, Isonne C, Salerno C, Migliara G, et al. What is the prevalence of low health literacy in European Union member states? A systematic review and meta-analysis. J Gen Intern Med 2021;36(3):753-761 [FREE Full text] [doi: 10.1007/s11606-020-06407-8] [Medline: 33403622]
16. van der Gaag M, Heijmans M, Spoiala C, Rademakers J. The importance of health literacy for self-management: a scoping review of reviews. Chronic Illn 2022;18(2):234-254. [doi: 10.1177/17423953211035472] [Medline: 34402309]

XSL•FO
RenderX

17. Larsen MH, Mengshoel AM, Andersen MH, Borge CR, Ahlsen B, Dahl KG, et al. "A bit of everything": health literacy interventions in chronic conditions—a systematic review. Patient Educ Couns 2022;105(10):2999-3016 [FREE Full text] [doi: 10.1016/j.pec.2022.05.008] [Medline: 35641366]
18. Barbati C, Maranesi E, Giammarchi C, Lenge M, Bonciani M, Barbi E, et al. Effectiveness of eHealth literacy interventions: a systematic review and meta-analysis of experimental studies. BMC Public Health 2025;25(1):288 [FREE Full text] [doi: 10.1186/s12889-025-21354-x] [Medline: 39849354]
19. Faux-Nightingale A, Philp F, Chadwick D, Singh B, Pandyan A. Available tools to evaluate digital health literacy and engagement with eHealth resources: a scoping review. Heliyon 2022;8(8):e10380 [FREE Full text] [doi: 10.1016/j.heliyon.2022.e10380] [Medline: 36090207]
20. Karnoe A, Kayser L. How is eHealth literacy measured and what do the measurements tell us? A systematic review. Knowledge Manage E-Learning 2015;7(4):576-600. [doi: 10.34105/j.kmel.2015.07.038]
21. Lee J, Lee E, Chae D. eHealth literacy instruments: systematic review of measurement properties. J Med Internet Res 2021;23(11):e30644 [FREE Full text] [doi: 10.2196/30644] [Medline: 34779781]
22. Refahi H, Klein M, Feigerlova E. e-Health literacy skills in people with chronic diseases and what do the measurements tell us: a scoping review. Telemed J E Health 2023;29(2):198-208. [doi: 10.1089/tmj.2022.0115] [Medline: 35671526]
23. Holmen H, Fløło T, Tørris C, Løyland B, Almendingen K, Bjørnnes AK, et al. Unpacking the public health triad of social inequality in health, health literacy, and quality of life-a scoping review of research characteristics. Int J Environ Res Public Health 2023;21(1):36 [FREE Full text] [doi: 10.3390/ijerph21010036] [Medline: 38248501]
24. Jacobs RJ, Lou JQ, Ownby RL, Caballero J. A systematic review of eHealth interventions to improve health literacy. Health Inform J 2016;22(2):81-98 [FREE Full text] [doi: 10.1177/1460458214534092] [Medline: 24916567]
25. van Kessel R, Wong BLH, Clemens T, Brand H. Digital health literacy as a super determinant of health: more than simply the sum of its parts. Internet Interv 2022;27:100500 [FREE Full text] [doi: 10.1016/j.invent.2022.100500] [Medline: 35242586]
26. Thiyagarajan A, Grant C, Griffiths F, Atherton H. Exploring patients' and clinicians' experiences of video consultations in primary care: a systematic scoping review. BJGP Open 2020;4(1) [FREE Full text] [doi: 10.3399/bjgpopen20X101020] [Medline: 32184212]
27. Walthall H, Schutz S, Snowball J, Vagner R, Fernandez N, Bartram E. Patients' and clinicians' experiences of remote consultation? A narrative synthesis. J Adv Nurs 2022;78(7):1954-1967 [FREE Full text] [doi: 10.1111/jan.15230] [Medline: 35362191]
28. Holmen H, Holm AM, Kilvær TK, Ljoså TM, Granan L, Ekholdt C, et al. Digital outpatient services for adults: development of an intervention and protocol for a multicenter non-randomized controlled trial. JMIR Res Protoc 2023;12:e46649 [FREE Full text] [doi: 10.2196/46649] [Medline: 37428533]
29. Holmen H, Holm AM, Falk RS, Kilvær TK, Ljosaa TM, Ekholdt C, et al. A digital outpatient service with a mobile app for tailored care and health literacy in adults with long-term health service needs: multicenter nonrandomized controlled trial. J Med Internet Res 2025;27:e60343 [FREE Full text] [doi: 10.2196/60343] [Medline: 40294411]
30. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
31. Paasche-Orlow MK, Wolf MS. The causal pathways linking health literacy to health outcomes. Am J Health Behav 2007;31(1):19-26. [doi: 10.5993/ajhb.31.s1.4]
32. Dignio Connected Care. Dignio. URL: https://dignio.com/solution/ [accessed 2025-12-16]
33. Kværner K. How to assess value and benefits of innovation. Noteb—Nordic Test Beds. Centre for Connected Care and Nordic Innovation. 2018. URL: https://www.researchgate.net/figure/The-User-guide-How-to-assess-value-and-benefits-of-innovation-the-end-product-of-the_fig1_331666378 [accessed 2025-12-20]
34. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2008;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
35. Matulis JC, McCoy R. Patient-centered appointment scheduling: a call for autonomy, continuity, and creativity. J Gen Intern Med 2021;36(2):511-514 [FREE Full text] [doi: 10.1007/s11606-020-06058-9] [Medline: 32885369]
36. Lupton D. The digitally engaged patient: self-monitoring and self-care in the digital health era. Soc Theory Health 2013;11(3):256-270. [doi: 10.1057/sth.2013.10]
37. Song M, Elson J, Bastola D. Digital age transformation in patient-physician communication: 25-Year Narrative Review (1999-2023). J Med Internet Res 2025;27:e60512 [FREE Full text] [doi: 10.2196/60512] [Medline: 39819592]
38. Kho J, Gillespie N, Martin-Khan M. A systematic scoping review of change management practices used for telemedicine service implementations. BMC Health Serv Res 2020;20(1):815 [FREE Full text] [doi: 10.1186/s12913-020-05657-w] [Medline: 32873295]
39. Longhini J, Rossettini G, Palese A. Digital health competencies among health care professionals: systematic review. J Med Internet Res 2022 Aug 18;24(8):e36414 [FREE Full text] [doi: 10.2196/36414] [Medline: 35980735]

XSL•FO
RenderX

Abbreviations

COREQ: Consolidated criteria for Reporting Qualitative research
TSD: Tjeneste for Sensitive Data

Edited by A Mavragani; submitted 16.Jun.2025; peer-reviewed by W van Harten, FY Chou, YY Chen; comments to author 17.Sep.2025; revised version received 05.Nov.2025; accepted 12.Nov.2025; published 15.Jan.2026.

Please cite as:
Holmen H, Fosse E
"I Want to Spend My Time Living"—Experiences With a Digital Outpatient Service With a Mobile App for Tailored Care Among Adults With Long-Term Health Service Needs: Qualitative Study Using Thematic Analysis
J Med Internet Res 2026;28:e79155
URL: https://www.jmir.org/2026/1/e79155
doi:10.2196/79155
PMID:

©Heidi Holmen, Erik Fosse. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Feasibility and Acceptability of AI-Powered Tools for Early Autism Screening in Egypt: Semistructured Focus Group Study

Pratheepan Yogarajah[1*], PhD; Ammal M Metwally[2*], Prof Dr, PhD; Priyanka Chaurasia[1], PhD; Ghada A Elshaarawy[2], Prof Dr, PhD; Shereen M El Khateeb[3], MFA; Engy A Ashaat[2], PhD; Amal Elsaeid[2], PhD; Nahed A Elghareeb[4], PhD; Amira S ElRifay[2], Prof Dr, PhD

[1]School of Computing, Engineering & Intelligent Systems, Ulster University, Londonderry, Northern Ireland, United Kingdom

[2]National Research Centre, Giza, Giza, Egypt

[3]Department of psychology, American University in Cairo, Cairo, Egypt

[4]Ministry of Health and Population, Cairo, Egypt

[*]these authors contributed equally

Corresponding Author:
Pratheepan Yogarajah, PhD
School of Computing, Engineering & Intelligent Systems, Ulster University
Ulster University
Londonderry, Northern Ireland, BT48 7JL
United Kingdom
Phone: 44 2871675034
Email: p.yogarajah@ulster.ac.uk

Abstract

Background: Autism spectrum disorder (ASD) is often underdiagnosed in low- and middle-income countries due to limited specialist access, sociocultural stigma, and fragmented screening systems. Artificial intelligence (AI)–powered screening tools may improve early detection by enabling low-cost, accessible assessments. However, adoption depends on stakeholder trust, ethical safeguards, and alignment with local health system capacities.

Objective: This study explored the feasibility, acceptability, and perceived ethical and practical enablers and barriers to implementing AI-powered tools for early ASD screening in Egypt, with attention to urban–rural disparities and integration into existing care pathways.

Methods: We used a qualitative design with semistructured focus group discussions with 49 participants (21 parents of children with ASD and 28 health care professionals) recruited from urban and rural governorates. Discussions were audio-recorded, transcribed verbatim, and analyzed using Braun and Clarke's reflexive thematic analysis, supported by NVivo software (Lumivero). Methodological integrity was ensured through reflexivity, triangulation, and peer debriefing. Thematic saturation was monitored across groups, and participant diversity was prioritized across contexts.

Results: Five themes emerged: (1) AI as a supportive tool rather than a replacement for clinicians, emphasizing scalability and assistance for nonspecialists; (2) the need for cultural and contextual adaptation to ensure local relevance; (3) privacy, trust, and transparency concerns, including data security, consent, and algorithmic opacity; (4) reducing diagnostic inequities by addressing urban–rural disparities and strengthening community-based deployment; and (5) the preference for hybrid AI–human models, with conditions for adoption including cultural sensitivity, human oversight, and digital literacy support. Counts (n/N) of parents and health care professionals contributing to each theme were used descriptively as indicators of pattern salience rather than as statistical estimates of prevalence. Participants expressed cautious optimism, with parents emphasizing accessibility and speed, while health care professionals highlighted concerns about reliability, cultural adaptation, and data governance.

Conclusions: AI-powered ASD screening has potential to advance equitable early detection in underserved areas. Adoption requires transparent data governance, integration into hybrid human–AI models, culturally adaptive design, and targeted digital literacy initiatives. These findings provide an evidence-based roadmap for policymakers, technologists, and health system leaders to implement AI screening tools that are ethically sound, contextually relevant, and equity-focused.

(J Med Internet Res 2026;28:e82564) doi:10.2196/82564

XSL•FO
RenderX

KEYWORDS

AI; artificial intelligence; ASD; autism spectrum disorder; developing countries; Egypt; health equity; mobile applications; qualitative research; screening

Introduction

Autism spectrum disorder (ASD) is a prevalent neurodevelopmental condition that affects cognitive, behavioral, and social functioning, necessitating early and accurate identification for optimal developmental outcomes [1,2]. Early detection is particularly important because timely intervention can substantially improve language development, social skills, and long-term adaptive functioning [3]. However, many children remain undiagnosed due to barriers such as limited specialist access, cultural stigma, and inconsistent screening practices, particularly in low-resource settings [4,5].

Traditional screening methods, including clinical observation, structured interviews, and standardized behavioral checklists, remain the cornerstone of ASD diagnosis but are constrained by subjectivity and poor scalability. In Egypt, where the estimated prevalence of children aged 2-12 years at high risk for ASD is 3.3% (95% CI 3.1%-3.5%) [6] and those diagnosed with autism is 1.1% (95% CI 1.0%-1.2%) [7], the burden of undiagnosed developmental conditions is compounded by delayed health-seeking and shortages of trained professionals [6]. Developmental delays affect 6.4% of preschool-aged children [6] and 7.4% of school-aged children, underscoring the urgent need for scalable and accessible diagnostic alternatives [8]. These figures highlight the systemic challenges faced by families and health systems, where demand for developmental assessment far exceeds available capacity.

Artificial intelligence (AI) has emerged as a promising adjunct to traditional screening by enhancing diagnostic accuracy and reducing reliance on clinical judgment [9]. Machine learning models have demonstrated potential in identifying ASD-related traits via speech, behavior, facial analysis, and digital interactions [10]. Mobile AI apps, in particular, offer low-cost, early-stage screening accessible to both parents and frontline health care providers, especially in underserved communities [11,12]. Such tools could bridge gaps in access and standardization, addressing the fragmented diagnostic pathways observed in Egypt and other low- and middle-income countries (LMICs) [13].

Despite their promise, AI-driven tools face challenges, including algorithmic bias, data privacy concerns, limited cultural adaptability, and stakeholder mistrust [14]. In addition, most existing AI models are trained on Western datasets, raising questions about their validity and cultural fit in LMIC contexts [15]. Broader digital health research also shows that successful scale-up in low-resource environments requires attention to cultural adaptation, stakeholder engagement, and health system integration [16]. The feasibility and success of integrating AI into ASD screening depend not only on technical accuracy but also on community acceptance, ethical transparency, and usability.

These challenges can be better understood and systematically examined through established theoretical frameworks. To deepen interpretation of these adoption challenges, the study draws on the Health Belief Model (HBM) and the Technology Acceptance Model (TAM). HBM highlights how perceived susceptibility, benefits, and barriers influence caregivers' health-related decisions, including ASD help-seeking behaviors [17]. TAM, widely applied in digital health research, explains how perceived usefulness and ease of use shape acceptance of mobile and AI technologies in health care settings [18]. Integrating these frameworks provides a structured basis for examining how caregiver beliefs and provider attitudes interact with system-level constraints to shape adoption of AI-powered ASD screening in Egypt.

Against this backdrop, we conducted a qualitative study in Egypt using focus group discussions (FGDs) and follow-up interviews with health care providers and parents of young children with ASD across diverse urban, semiurban, and rural settings. The primary aim was to explore stakeholder perspectives on the feasibility, acceptability, and perceived ethical and practical enablers and barriers to implementing AI-powered tools for early ASD screening within the Egyptian health system. By situating these perspectives within Egypt's health system and sociocultural context, and interpreting them through the HBM and TAM, this qualitative inquiry seeks to inform the context-appropriate design and implementation of AI-enabled ASD screening approaches in LMICs. Accordingly, we addressed 2 research questions—research question 1 (RQ1) and research question 2 (RQ2): (RQ1) what barriers and contextual challenges do parents and health care providers identify in current ASD screening, referral, diagnosis, and access-to-support pathways in Egypt? (RQ2) How do parents and health care providers perceive the acceptability and feasibility of AI-powered ASD screening tools, including perceived value, usability, trust, and implementation requirements?

Methods

Research Design

This qualitative study used FGDs and individual interviews to explore stakeholder perspectives on AI-assisted autism screening in Egypt. Overall, the study included 49 participants, comprising 28 health care providers (pediatricians, psychologists, family medicine physicians, speech therapists, nurses, and health-information specialists) and 21 parents of children with a confirmed ASD diagnosis aged 6 years or younger. All participants took part in 1 of 7 FGDs, and a subset of these same participants later contributed to brief follow-up individual interviews; no additional participants were recruited solely for interviews. Participants were purposively sampled to reflect diverse professional backgrounds and geographic locations, including urban, semiurban, and rural settings [19]. This diversity facilitated examination of disparities in ASD screening accessibility and cultural perceptions.

FGDs allowed interactive exploration of clinical, cultural, and technical aspects of ASD diagnosis, while individual interviews

XSL•FO
RenderX

captured sensitive or nuanced views less likely to emerge in group settings [20,21]. Sessions were conducted separately for health care professionals and parents to minimize power imbalances and encourage open dialogue. Data collection proceeded iteratively alongside early familiarization and preliminary coding; after each FGD (and subsequent follow-up interviews), the facilitation and analysis team conducted rapid debriefs and reviewed emerging codes to assess whether additional sessions were likely to generate substantively new insights. Saturation was judged to have been reached when successive sessions yielded no substantively new codes or theme candidates beyond those already identified and the team agreed that additional data were unlikely to add meaningful conceptual depth [22,23]. Geographic representation was prioritized over equal participant numbers per site.

Participants were recruited through hospitals, specialized ASD clinics, professional networks, autism advocacy groups, special needs schools, and early intervention centers. All received detailed study information and confidentiality assurances and provided written informed consent prior to participation.

A full description of the stepwise thematic analysis process (phases 1-6 of Braun and Clarke's framework) has been provided in Multimedia Appendix 1.

Inclusion and Exclusion Criteria

Inclusion criteria were (1) health care professionals directly involved in ASD screening or developmental care, and (2) parents of children with a confirmed ASD diagnosis aged 6 years or younger. Consequently, the realized sample reflects perspectives that are closely anchored in the early detection and screening phase rather than in later, long-term intervention. Parents of children diagnosed during or soon after the preschool years were therefore prioritized, as their reflections most directly inform the feasibility and acceptability of AI-powered early screening tools.

Exclusion criteria were (1) health care providers without direct ASD-related experience, (2) parents of children with suspected but unconfirmed neurodevelopmental disorders, and (3) individuals unwilling to provide consent or unable to complete the full discussion session. This approach ensured that voices from both clinical and caregiving domains were adequately represented, enhancing the relevance and applicability of findings.

Recruitment Process

Participants were recruited through multiple channels to ensure diversity and relevance. Health care professionals were identified via hospitals, specialized ASD clinics, and professional networks in the selected governorates. Pediatric associations and child development centers facilitated the engagement of pediatricians, psychologists, nurses, and other specialists involved in ASD screening and intervention. All individuals invited agreed to participate, and no invited participants declined or withdrew.

Parents were recruited through autism advocacy groups, special needs schools, early intervention centers, and referrals from pediatric clinics. This ensured inclusion of caregivers with firsthand experience navigating ASD diagnosis and care pathways.

All participants received detailed study information, including objectives, procedures, confidentiality assurances, and their right to withdraw at any time. Written informed consent was obtained prior to participation.

Geographic Distribution and Representation

The study was conducted across 6 Egyptian governorates selected to capture geographic and demographic diversity, including urban, semiurban, and rural contexts. Governorates were purposively chosen to reflect differences in levels of urbanization, availability of ASD-related services, and health system infrastructure (eg, Ministry of Health and Population primary health care and maternal and child health facilities, university hospitals, and private or nongovernmental centers). This approach ensured that findings reflected a broad spectrum of health care access, cultural beliefs, and diagnostic experiences related to ASD. Rather than aiming for equal sample sizes per site, recruitment focused on meaningful representation of regional disparities in ASD service delivery. Geographic representation was therefore operationalized as maximum-variation coverage of contrasting service environments rather than as proportional sampling based on population size or formal rural–urban quotas.

This geographic diversity, combined with rigorous participant selection, supported a comprehensive understanding of the challenges and opportunities associated with integrating AI-powered ASD screening tools in varied sociocultural and clinical settings.

Data Collection

Data were collected through semistructured FGDs and individual interviews, combining structured prompts with open-ended dialogue to elicit both guided and spontaneous insights. FGDs lasted 60-90 minutes and were moderated by trained facilitators with backgrounds in public health, ASD, and qualitative research. During parent FGDs, children and, in many cases, the other parent were present in the room so that 1 caregiver could supervise the child while the other participated more actively in the discussion. All facilitators were female members of the research team with prior experience in qualitative interviewing and no prior personal or professional relationship with participants. The same facilitators also conducted the follow-up interviews to ensure continuity and rapport. No repeat interviews were conducted; each participant took part in a single FGD. In addition to the FGDs, short individual interviews were conducted with the same FGD participants to clarify points raised during the group discussion and to allow participants to elaborate on topics they were not comfortable discussing in a group setting. No separate pool of interview participants was recruited; all interviewees had already taken part in an FGD. These interviews were conducted immediately after the FGDs or were scheduled within the same week (as follow-ups for deeper insights). This approach ensured that group-generated themes were accurately understood and provided an opportunity to validate and deepen emerging insights, enhancing the credibility of the analysis.

XSL•FO
RenderX

Given the sensitive nature of ASD screening and emerging AI technologies, moderators received additional training to manage emotionally charged discussions and foster a safe, nonjudgmental environment. To facilitate understanding, participants received a brief explanatory note on AI decision-making, stating:

> *This allows you to understand the logic behind AI's decision-making in a simple and clear way; without needing advanced technical knowledge. It's like AI showing you how it arrived at its decision, so you know it's not just a random guess.*

All individuals invited agreed to participate, and no invited participants declined or withdrew.

Moderator Training and Power Imbalance Considerations

To address potential power dynamics, FGDs were conducted separately for health care professionals and parents, with moderators trained to foster inclusive participation and encourage open dialogue. Additional individual interviews were conducted to capture sensitive experiences that participants might not have shared in group settings (Multimedia Appendix 1 provides detailed description of moderator training and power-imbalance considerations). All focus groups and individual interviews were conducted by a research team of female public health researchers at the National Research Centre (NRC) with formal training in qualitative interviewing and prior experience in ASD-related services. None had a direct clinical care relationship with the participating families or children, which helped minimize role-related coercion while encouraging open and candid discussion of views on ASD screening and AI tools.

Discussion Guide Topics

Focus group discussions were guided by 2 core research questions: (RQ1) What barriers and contextual challenges do parents and health care providers identify in current ASD screening, referral, diagnosis, and access-to-support pathways in Egypt? (RQ2) How do parents and health care providers perceive the acceptability and feasibility of AI-powered ASD screening tools, including perceived value, usability, trust, and implementation requirements? The full discussion guide is available in Multimedia Appendix 2.

Data Recording and Transcription

All FGDs and interviews were audio-recorded and transcribed verbatim. Field notes captured nonverbal cues and group dynamics. Discussions were conducted in Arabic and professionally translated into English to ensure linguistic and conceptual accuracy during thematic analysis. Initial coding was conducted on the original Arabic transcripts by bilingual members of the research team, with English translations used to support team-based analysis and reporting. Arabic and English versions were reviewed side by side, and any discrepancies or ambiguous phrases were resolved through discussion between the coders and the professional translator to preserve the intended meaning. Verbatim transcripts were not formally returned to participants for checking due to resource and logistical constraints; instead, analytic accuracy was supported through verbatim transcription, independent transcript review, peer debriefing, and limited member checking with a purposive subset of participants during follow-up calls to clarify key interpretations.

In reporting quotations, we used anonymized identifiers indicating participant role, gender, and governorate (for example, "Parent-F, Fayoum" and "Pediatrician-M, Cairo") to illustrate the diversity of perspectives while preserving confidentiality.

Data Analysis

The analysis followed Braun and Clarke's (2019) reflexive thematic analysis framework, which involves six iterative phases (1) familiarization with the data through repeated reading of transcripts, (2) generating initial inductive codes, (3) searching for candidate themes by clustering related codes, (4) reviewing and refining themes for coherence, (5) defining and naming themes with operational clarity, and (6) producing the final analytic report integrating quotes and interpretations. NVivo software (version 14; Lumivero) supported coding and data organization. Coding was led by 2 bilingual public health researchers with training in qualitative methods and experience in ASD-related research, supported by a senior qualitative methodologist; all are based at the NRC, and none were involved in direct clinical care of the participating families.

The 2 primary analysts independently read and annotated the transcripts, generated initial inductive codes, and then met regularly to compare and refine the coding frame and candidate theme map, not to calculate interrater reliability but to deepen reflexive engagement with the data, consistent with reflexive thematic analysis. All FGD and interview transcripts were imported into NVivo and analyzed as a single qualitative data corpus using the same inductive coding frame; the data-collection mode (focus group vs individual interview) was retained as a case attribute to enable comparison of patterns across modes when relevant rather than conducting 2 separate analytic processes. For example, codes such as "AI helps nonspecialists," "AI as first filter," "AI speeds referrals," and "AI cannot replace doctors" were clustered into a subtheme labelled "AI as a supportive triage tool," which was then situated within the broader theme "AI as a supportive tool rather than a replacement for clinicians."

In line with reflexive thematic analysis, counts (n/N) for parents and health care providers were used descriptively in the tables as signposts of pattern salience; that is, to indicate how many participants contributed to a given code or theme rather than as statistical estimates of prevalence or as the basis for inferential claims [24]. In addition, emergent themes were examined in relation to the HBM and TAM to provide a theoretical lens for interpreting stakeholder perceptions of AI-powered ASD screening. This deductive overlay complemented the inductive analysis, enabling findings to be systematically linked with established models of health behavior and technology adoption.

During analysis, we also explored potential differences in perspectives across health care professional subgroups (pediatricians, nurses, psychologists, and child development specialists); however, given the modest numbers in each

XSL•FO
RenderX

subgroup and the substantial overlap in core themes, no systematic divergence in overarching patterns was identified, and themes are therefore reported at the combined "health care provider" level, with subgroup nuances highlighted where relevant. Reflexive journals were maintained by the primary analysts to document assumptions (eg, expectations about the benefits of AI in overstretched public services) and to record how these were challenged by the data; these reflections informed theme naming and interpretation (eg, reframing early "AI will solve gaps" codes as "cautious optimism conditioned by system constraints"). Interpretive rigor was maintained through reflexive team discussions, staged member checking with a purposive subset of participants, peer debriefing with a senior qualitative researcher, and triangulation of perspectives across participant groups. A detailed description of the 6 analytic phases, including examples of coding procedures and reflexivity practices, is provided in Multimedia Appendix 1.

Ensuring Trustworthiness

This study applied classic criteria of credibility, transferability, dependability, and confirmability [25-27]. Credibility was established through triangulation of FGDs and interviews, member checking, and peer debriefing. Member checking occurred at 2 stages. First, at the end of each focus group and interview, moderators briefly summarized the main discussion points and explicitly invited participants to correct, nuance, or add to these summaries, allowing immediate verification of the researchers' understanding. Second, after preliminary theme development, the research team prepared concise written summaries of the key themes and a small set of illustrative quotes and contacted a purposive subset of parent and clinician participants from different governorates (via brief follow-up calls or messaging apps) to review these summaries. Participants were asked whether the themes and examples reflected their experiences, and their feedback was used to refine thematic boundaries, adjust wording of themes, and address any apparent misinterpretations. Formal transcript-based member checking (returning full transcripts to all participants) was not undertaken due to resource and logistical constraints; however, this staged, selective approach allowed verification of core interpretations without recontacting every participant or overburdening busy clinicians and caregivers.

Transferability was supported by thick descriptions of participant experiences, service settings, and contextual factors across diverse regions, enabling readers to assess how findings may apply to similar LMIC contexts. Dependability was ensured by documenting analytic procedures and maintaining an audit trail. The audit trail included decision logs on coding, theme development, and changes to the analytic framework over time. Confirmability was strengthened through reflexivity and integration of direct participant quotes into findings. The primary analysts maintained reflexive journals to document their assumptions and how these were challenged by the data; these reflections were discussed in team meetings and peer debriefing sessions with a senior qualitative researcher, supporting critical examination of alternative interpretations and reducing the risk of unchecked researcher bias. Triangulation across data sources (FGDs and follow-up interviews) and participant groups (parents and health care providers) further contributed to the credibility and confirmability of the analysis. Details of the trustworthiness measures are provided in Multimedia Appendix 1.

Ethical Considerations

Ethical approval for this study was granted by the NRC of Egypt (approval number 2705250141). All participants provided written informed consent prior to participation. Before participation, individuals received detailed study information, including objectives, procedures, confidentiality measures, and their right to withdraw at any stage without justification or consequences. Participants had ample time to review consent forms, ask questions, and opt out if desired. To protect privacy, all data were anonymized, securely stored, and accessible only to authorized researchers. Participants received reimbursement for travel costs and a modest goodwill payment of US $20, as approved by the NRC ethical committee and funded by Ulster University (cost center 71799); these tokens were intended to offset participation-related expenses rather than act as coercive incentives. Concerns raised by participants were incorporated into the analysis to highlight ethical implications of AI integration in screening of neurodiversity disorders [28,29]. Researchers followed international ethical guidelines, including the Declaration of Helsinki and General Data Protection Regulation data protection standards [30].

Through comprehensive consent procedures, strict confidentiality measures, and reflexive ethical discussions, this study upheld responsible and ethical qualitative research practices. Participation was fully voluntary, and participants were informed that they could withdraw at any time without consequences. Data were collected in a confidential manner, deidentified and delinked from personal identifiers, and stored in a secure location. In addition, data collection and handling complied with Egypt's personal data protection framework; identifiable information was stored on password-protected institutional servers at the NRC in Cairo, with access restricted to the core research team, and deidentified analytic datasets were stored separately and shared only in encrypted form. Information disclosure ("making sure participants understand") was guaranteed according to the recommendations of prior work on Egyptian patients' and guardians' perceptions of clinical informed consent [31].

This study was designed, conducted, and reported in accordance with the COREQ (Consolidated Criteria for Reporting Qualitative Research), and a completed COREQ checklist is provided in Multimedia Appendix 3.

Results

Overview

Table 1 summarizes the composition of the study participants, including parents and health care providers, across 6 Egyptian governorates. Participants were recruited from governorates representing diverse geographic and sociodevelopmental settings across Egypt, including urban (Cairo), urban/semiurban (Giza and Damietta), mixed urban–rural (Dakahlia and Assuit), and predominantly rural/semiurban (Fayoum). The sample comprised clinicians (pediatricians, psychologists, family medicine physicians, speech therapists, nurses, and

XSL•FO
RenderX

health-information specialists) and parents of children with diagnosed ASD. Participant numbers varied by site depending on service availability and community size rather than equal quotas per governorate, consistent with qualitative sampling principles.

Table 1. Participant description (N=49) by geographic location, professional role, and participant type.

Governorate/region [32-36]	Setting type	Parents, n/N	Health care providers, n/N	Health care provider roles
Dakahlia (Lower Egypt, mixed urban–rural)	Autism centers, outpatient clinics	3/21	4/28	Pediatrician, psychologist, head nurse, and family medicine physician
Cairo (cities, urban)	MCH[a] centers	5/21	7/28	Pediatricians, 2 psychologists, head nurse, 2 general practitioners, and health-information specialists
Giza (cities, urban/semiurban)	Early screening units	3/21	3/28	Pediatrician, speech therapist, and MOHP[b] officer
Fayoum (Upper Egypt, rural/semiurban)	Primary health units	5/21	6/28	Pediatrician, 2 psychologists, head nurse, PHC[c] physician, and health-information specialist
Damietta (coastal, urban/semiurban)	Pediatric clinics, NGO[d]	2/21	4/28	Pediatrician, 2 psychologists, and speech therapist
Assuit (Upper Egypt, mixed urban–rural)	Community health centers	3/21	4/28	Pediatrician, psychologist, head nurse, and PHC Physician

[a]MCH: maternal and child health care.

[b]MOHP: Ministry of Health and Population.

[c]PHC: primary health care.

[d]NGO: nongovernmental organization.

In the main thematic tables (Tables 2-6), we therefore report absolute counts with explicit denominators for each stakeholder group (for example, 20/21 parents; 25/28 health care providers) rather than percentages, to reflect the qualitative, nonstatistical nature of the data while still allowing readers to assess the salience of each view across participants [24,37].

Table 2. Theme 1: reported challenges in autism diagnosis and pathways, mention counts (n/N) by participant group.

Challenge[a]	Parents, n/N	Health care providers, n/N
Lack of ASD[b]-specific training for health care professionals	18/21	21/28
No standardized ASD screening protocols	13/21	16/28

[a]Values represent the number of participants (n) who spontaneously mentioned each challenge. Denominators (N) refer to total participants in each group (parents=21 and health care providers=28). These counts indicate thematic salience, not prevalence or consensus. Illustrative quotations capture both typical and divergent views to reflect the full spectrum of perspectives and should not be interpreted as statistical estimates.

[b]ASD: autism spectrum disorder.

Table 3. Theme 2: limited awareness, social stigma, and misconceptions about autism, mention counts (n/N) among parents.

Challenge[a]	Parents, n/N
Stigma and fear of judgment	11/21
Misconceptions (autism as temporary or due to poor parenting)	8/21
Lack of parental awareness about ASD[b] screening	7/21

[a]Values represent the number of participants (n) who spontaneously mentioned each challenge. Denominators (N) refer to total participants in each group (parents=21). These counts indicate thematic salience, not prevalence or consensus. Illustrative quotations capture both typical and divergent views to reflect the full spectrum of perspectives and should not be interpreted as statistical estimates.

[b]ASD: autism spectrum disorder.

XSL•FO
RenderX

Table 4. Theme 3: reported barriers to accessing autism diagnostic and support services: mention counts (n/N) by participant group.

Barriers[a]	Parents, n/N	Health care providers, n/N
Shortage of ASD[b] specialists in local areas	13/21	23/28
Long waiting times for appointments	12/21	11/28
High cost of private consultations	9/21	11/28
Travel distance/time burden	15/21	10/28

[a]Values represent the number of participants (n) who spontaneously mentioned each barrier. Denominators (N) refer to total participants in each group (parents=21 and health care providers=28). These counts indicate thematic salience, not prevalence or consensus. Illustrative quotations capture both typical and divergent views to reflect the full spectrum of perspectives and should not be interpreted as statistical estimates.

[b]ASD: autism spectrum disorder.

Table 5. Theme 4: perceived role of artificial intelligence (AI) in autism screening, mention counts (n/N) by participant group.

Factor[a]	Parents, n/N	Health care providers, n/N
Support for AI-based screening	9/21	19/28
Need for culturally adapted tools	12/21	16/28
Preference for explainable AI-generated results	15/21	17/28

[a]Values represent the number of participants (n) who spontaneously mentioned each factor. Denominators (N) refer to total participants in each group (parents=21 and health care providers=28). These counts indicate thematic salience, not prevalence or consensus. Illustrative quotations capture both typical and divergent views to reflect the full spectrum of perspectives. Figures should not be interpreted as statistical estimates.

Table 6. Theme 5: reported concerns about artificial intelligence (AI) adoption, mention counts (n/N) by participant group.

Concern[a]	Parents, n/N	Health care providers, n/N
Data privacy and security risks	17/21	23/28
Fear of AI replacing human judgment	6/21	__[b]
Limited digital literacy among parents	4/21	__[b]
User-friendly interface for non-specialist health care providers	__[b]	21/28

[a]Values represent the number of participants (n) who spontaneously mentioned each concern. Denominators (N) refer to total participants in each group (parents: 21; health care providers: 28). These counts indicate thematic salience, not prevalence or consensus. Illustrative quotations capture both typical and divergent views to reflect the full spectrum of perspectives. Figures should not be interpreted as statistical estimates.

[b]Not applicable.

This analysis drew upon 7 FGDs involving key health care providers, health professionals, and parents of children with ASD across 6 governorates in Egypt. Each group included approximately 7 participants, totaling 28 health care professionals and 21 parents. Health care professionals included heads of departments for children with special needs, chief technicians, physicians, nurses, speech therapists, and health-information specialists. The mean age of health care providers was 43.3 (SD 11.0) years. The mean age of parents was 44.3 (SD 10.6) years. Among their children, 66.7% (14/21) were male and 33.3% (7/21) female, with a mean age at diagnosis of 2.2 (SD 0.6) years; 80.9% (17/21) were classified as having moderate autism. Educational attainment varied across groups: among health care providers, 39.3% (11/28) held graduate degrees, 32.1% (9/28) postgraduate qualifications, and 28.6% (8/28) technical education. Among parents, 57.1% (12/21) held graduate degrees, 23.8% (5/21) postgraduate qualifications, and 19.1% (4/21) secondary or technical education.

Thematic analysis of the focus group and interview data identified 5 major themes that captured stakeholder perspectives on the feasibility, acceptability, and challenges of AI-powered autism screening in Egypt. Absolute counts with denominators (n/N) are presented to indicate how many participants in each group (parents and health care providers) raised each issue during the discussions. Importantly, these figures should be interpreted as indicators of thematic salience rather than statistical prevalence or consensus, given the exploratory qualitative design and the fact that not all participants were asked identical questions in a standardized format.

To reinforce this, table footnotes clarify that counts (n/N) reflect how frequently topics arose rather than the degree of agreement across the sample. While the abstract summarizes the 5 overarching themes conceptually according to AI adoption factors, the following Results section presents them in the empirical order in which they emerged during thematic analysis.

These figures should be viewed as indicators of thematic salience rather than measures of agreement or prevalence, as not all participants were asked identical questions in a

XSL•FO
RenderX

standardized format. Instead, they reflect how frequently particular topics arose during the conversations and the relative emphasis placed on them by participants. To ensure analytic rigor, the interpretation of these findings was informed by reflexive consideration of the research team's positionality, clinical experience, and prior exposure to autism-related work in Egypt. The 5 themes emerging from the analysis were delays in autism diagnosis and inconsistent pathways; limited awareness, social stigma, and misconceptions about autism; barriers to accessing diagnostic and support services; perceived role of AI in autism screening; and balancing trust, transparency, and data protection in AI screening.

Tables 2-6 summarize the 5 major themes derived from this analysis. Each table presents key subthemes, absolute counts with denominators for parents and health care providers, representative participant quotations, and brief interpretive notes. This structure facilitates comparison across themes while preserving a clear link between each table and its corresponding narrative subsection in the Results. Themes 1-3 primarily address RQ1 by characterizing current pathway-related and access-related barriers to timely ASD screening and support, whereas Themes 4 and 5 primarily address RQ2 by examining perceived acceptability, feasibility, and trust-related considerations for AI-enabled ASD screening.

Theme 1: Delays in Autism Diagnosis and Inconsistent Pathways

In relation to RQ1 (current pathway-related barriers), most health care providers reported lacking specific training in ASD diagnosis, and many raised concerns about the absence of standardized screening protocols. These delays were most often attributed to long waiting lists, limited specialist availability, and the absence of standardized referral pathways. Diagnostic inconsistency was another recurring concern, with many parents and more than half of health care professionals noting that children often received conflicting assessments from different providers.

One parent described the protracted wait for assessment:

> *I had to wait over six months for my child to be assessed, and even then, I wasn't sure if the diagnosis was correct because different doctors gave me different answers.* [Parent-M, Cairo]

This quote illustrates 2 critical barriers: significant delays due to workforce shortages and diagnostic inconsistency arising from the absence of standardized screening protocols. Together, these factors contribute to late interventions and parental uncertainty.

Health care professionals echoed these concerns, with one psychologist noting:

> *We see many children after they have been to three or four different doctors, each giving a different opinion.* [Psychologist-F, Assuit]

Parents reported long diagnostic delays:

> *I went to three doctors before someone said autism" (Parent, Fayoum)—while healthcare providers cited limited resources and high caseloads as major obstacles. Healthcare provider described "a system that reacts late, not one that screens early.* [Speech therapist-M, Damietta]

This reflects fragmentation of care pathways and highlights why nearly two-thirds of parents and half of health care providers in our sample emphasized the urgent need for standardized assessment tools. Importantly, participants also drew connections between these diagnostic inconsistencies and later discussions of AI, suggesting that standardized digital tools could harmonize assessments and reduce conflicting diagnoses (Theme 5).

Transitioning from diagnostic delays, participants also described how broader societal perceptions of autism further compound these challenges, as explored in Theme 2.

Theme 2: Limited Awareness, Social Stigma, and Misconceptions About Autism

Cultural stigma and misinformation emerged as significant barriers to early ASD diagnosis and were raised by many parents. Misconceptions, such as autism being temporary or caused by poor parenting, were frequently reported, and several parents were unaware that early screening was even possible.

Parents described community-level misunderstandings, such as attributing autism symptoms to poor parenting, stubbornness, or spiritual causes.

> *People told me my child was just naughty, or that I was spoiling him too much.* [Parent-F, Assuit]

This highlights the community-level stigma faced by more than half of parents in the sample, reinforcing how misconceptions delay help-seeking behavior. Health care professionals also noted that misconceptions among general practitioners sometimes led to misdiagnosis, resulting in children being referred to speech therapy or psychiatry without autism-specific assessment.

A pediatrician from Fayoum explained:

> *Some doctors still think autism means a child cannot speak at all, so they miss milder cases.* [Pediatrician-M, Fayoum]

By linking parental stigma with professional misconceptions, participants emphasized that awareness deficits are multilayered, spanning families, communities, and frontline medical providers. These deficits not only prolong diagnostic delays but also shape the systemic barriers discussed in Theme 3. In particular, stigma and misconceptions were seen as amplifying existing access problems, since families already hesitant to seek help often face structural hurdles once they do attempt to engage with services.

Theme 3: Barriers to Accessing Diagnostic and Support Services

Whereas Theme 2 captured attitudinal and cultural barriers to early ASD identification, Theme 3 focuses on structural and logistical constraints that hinder access to timely diagnosis and ongoing support. Barriers to service access were reported by many parents and most health care providers. These included long travel distances, high service costs, and the scarcity of autism-specialized centers, particularly in rural areas.

XSL•FO
RenderX

One mother from Fayoum explained:

> *The nearest centre is two hours away, and I cannot afford to travel every week.* [Parent-F, Fayoum]

Health care providers recognized these structural barriers. A psychologist from Cairo noted:

> *Families in rural areas have very limited options. Sometimes we try to offer online consultations, but internet quality is a problem.* [Psychologist-M, Cairo]

Financial constraints were also significant. One parent shared:

> *Therapy costs a lot, and we have to choose between paying for treatment or other basic needs.* [Parent-M, Dakahlia]

These findings underline the intersection of geographic, economic, and infrastructural barriers, which can delay both diagnosis and intervention. This was particularly salient among parents and was also frequently highlighted by health care providers, underscoring the disproportionate burden borne by families navigating multiple constraints simultaneously. Against this backdrop of limited access, many participants suggested that AI tools could potentially bridge these gaps by offering more accessible first-line screening, a possibility explored further in Theme 4. When discussing digital health options, parents and providers repeatedly pointed out that limited internet access and affordability challenges could constrain the feasibility of AI-powered tools even as they recognized their potential benefits.

Theme 4: Perceived Role of AI in Autism Screening

In relation to RQ2 (acceptability and feasibility of AI-enabled screening), optimism about AI-powered tools was expressed by many health care professionals and a substantial proportion of parents, who emphasized the potential for early risk identification in underserved areas. Some pediatricians and child development specialists highlighted AI's promise for accelerating referrals and supporting overburdened systems, while frontline nurses and primary care physicians more often focused on how such tools might fit within existing workloads and communication with families.

A pediatrician in Giza stated:

> *AI could help us identify at-risk children sooner, so we can refer them before the delay becomes too long.* [Pediatrician-F, Giza]

Parents similarly expressed cautious optimism:

> *If it's accurate and easy to use, I would try it for my child.* [Parent-F, Assuit]

Most parents viewed AI as a tool that could support earlier detection and broaden access to assessment, particularly in areas with limited specialist services:

> *If the app can tell me early that my child is at risk, I can go to the doctor sooner instead of waiting for years.* [Parent-F, Fayoum]

However, a minority expressed discomfort with technology-mediated interactions, preferring face-to-face communication:

> *I prefer someone to talk to me directly, not an app that gives me a score.* [Parent-M, Cairo]

These divergent views highlight the need for AI screening tools to be positioned as supportive aids rather than replacements for human care. Overall, enthusiasm was clear, but participants emphasized that meaningful adoption would require addressing structural barriers (Theme 3) and building trust through transparency and usability (Theme 5).

Theme 5: Balancing Trust, Transparency, and Data Protection in AI Screening

Concerns about trust and usability emerged strongly, with many parents and health care providers citing data privacy as their foremost worry. Psychologists often linked these concerns to existing stigma and the risk of labeling, whereas pediatricians tended to emphasize clinical responsibility for acting on AI-generated risk scores. Participants also frequently emphasized the importance of clear, understandable outputs, and some parents expressed concern that AI might replace human judgment.

One pediatrician stressed the importance of maintaining human oversight:

> *AI should not replace a specialist. A diagnosis should come from a doctor, not an algorithm.* [Pediatrician-F, Dakahlia]

Parents voiced privacy concerns:

> *If I use an app to screen my child, how do I know my data won't be misused? I don't want personal information shared without my consent.* [Parent-M, Damietta]

Digital literacy emerged as an additional barrier, especially in rural contexts:

> *I am not good with technology. If the AI screening tool is too complicated, I won't know how to use it properly for my child.* [Parent-F, Fayoum]

Health care providers also highlighted usability concerns. A physician explained that:

> *AI screening tools must be easy to use, especially for general practitioners and nurses who are not ASD specialists.* [PHC Physician-F, Assuit]

A nurse echoed this point, adding:

> *If the system is too complicated, healthcare providers won't use it, and parents will struggle to understand the results.* [Head Nurse-F, Cairo]

However, a few younger health care providers expressed optimism about automation:

> *If AI can handle screening, we can focus on therapy and follow-up.* [Speech therapist-M, Damietta]

This generational contrast reflects differing comfort levels with digital integration in clinical care. Across all sites, participants linked AI adoption to issues of fairness, privacy, and equity.

One policymaker warned:

XSL•FO
RenderX

> *Technology must not widen the gap between rich and poor families.* [MOH officer-F, Giza]

Many parents wanted assurances of data protection, with one noting:

> *Once data is uploaded, we lose control. That scares us.* [Parent-M, Fayoum]

In contrast, a few tech-savvy participants felt that:

> *Data collection is normal—everything is online now.* [Technology specialist-F, Cairo]

Their comfort with digital data contrasts with the majority's apprehension, illustrating generational and digital literacy divides. To accommodate both digitally inexperienced parents and nonspecialist providers, participants underscored the need for intuitive design. Trust in AI was therefore linked not only to accuracy and cultural fit but also to privacy protections, clarity of outputs, and user-friendliness.

This final theme loops back to Theme 1's concern with diagnostic inconsistencies, reinforcing that AI adoption could help standardize pathways if implemented with strong privacy protections and clear communication.

In addition to the 5 major themes, several divergent perspectives emerged. For example, while many parents emphasized accessibility as a key advantage, some questioned whether AI tools might replace human empathy. Similarly, a minority of clinicians expressed optimism about algorithmic decision support, contrasting with the dominant skepticism. These variations illustrate the heterogeneity of experiences and highlight areas for tailored implementation strategies. To support interpretation, the 5 themes are also conceptually mapped onto HBM and TAM constructs; this integration is summarized in Figure 1 and elaborated in the Discussion.

Figure 1. Conceptual mapping of emergent themes to Health Belief Model (HBM) and Technology Acceptance Model (TAM) constructs.

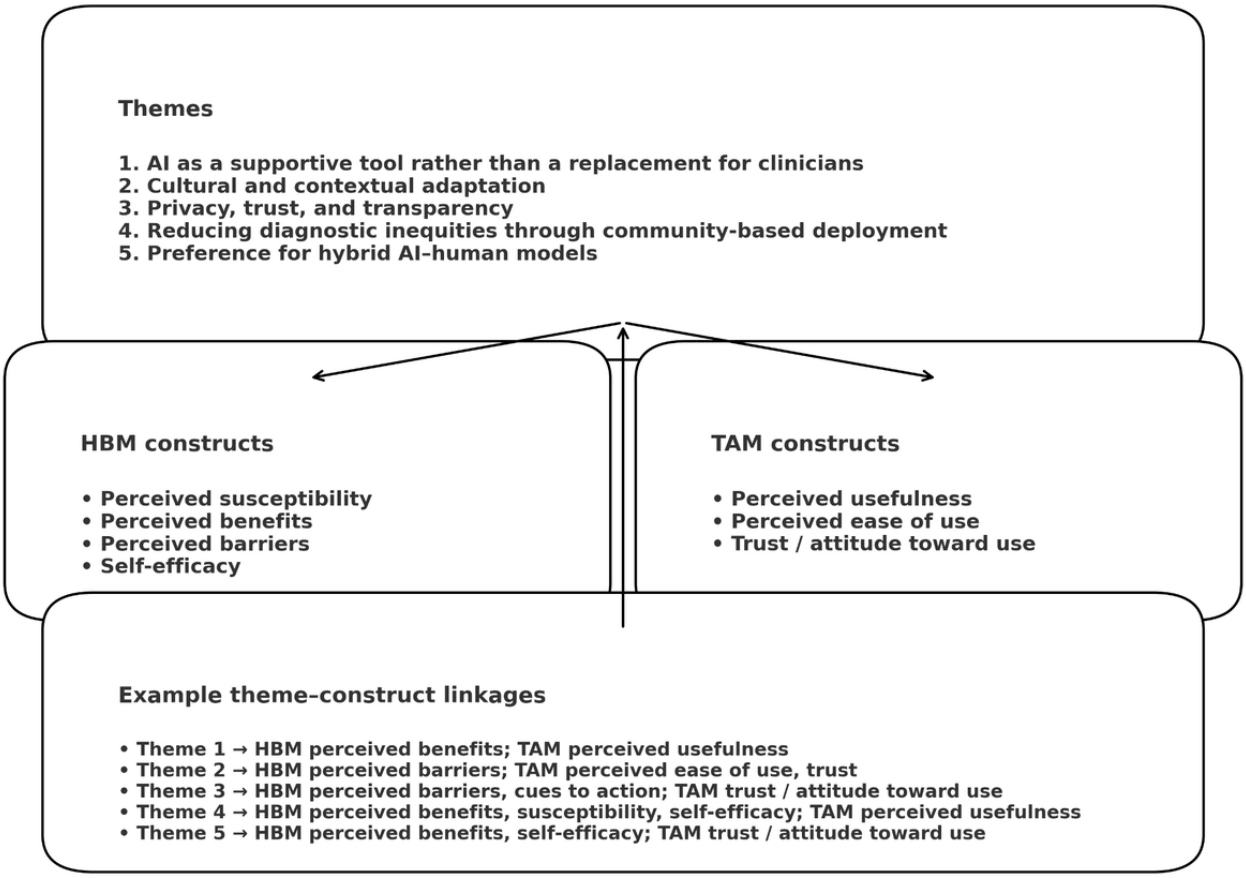

Discussion

Overview

This study explored stakeholder perspectives on the feasibility, acceptability, and perceived ethical and practical enablers and barriers to implementing AI-powered tools for early ASD screening within the Egyptian health system. Overall, participants recognized that AI-powered tools could help bridge diagnostic gaps, support earlier detection, and extend services to underserved areas. At the same time, they highlighted substantial barriers that must be addressed for successful adoption, including systemic delays and fragmented care pathways, lack of standardized diagnostic practices, pervasive stigma and low awareness, concerns about cultural appropriateness and transparency of AI decisions, and limitations in digital infrastructure and workforce training.

As illustrated in Figure 1, the themes align closely with core constructs of the HBM and TAM. Perceptions of AI as a

XSL•FO
RenderX

supportive triage tool, particularly for nonspecialists in underserved areas, reflect strong perceived benefits and perceived usefulness, which are key drivers of screening uptake in HBM- and TAM-informed frameworks. In contrast, concerns about data privacy, algorithmic opacity, and cultural mismatch map onto perceived barriers, perceived ease of use, and trust-related TAM constructs and help explain ambivalence despite overall cautious optimism. The preference for hybrid AI–human models illustrates how trust, self-efficacy, and attitudes toward use are negotiated in practice. Stakeholders were more comfortable when AI outputs were explicitly embedded within human oversight and relational care. These relationships are summarized in the conceptual mapping presented in Figure 1 and highlight potential leverage points for implementation, including strengthening perceived benefits and usefulness, reducing perceived barriers, and actively building trust and self-efficacy around AI tools.

Viewed through this HBM/TAM lens, the findings highlight both the promise and limitations associated with AI-driven neurodiversity screening. Consistent with prior research, AI-based screening tools offer a scalable, accessible, and cost-effective solution to bridging gaps in ASD diagnosis, particularly in underserved regions [13,38]. In keeping with Braun and Clarke's [24] call for thematic analysis to move beyond description toward theoretically informed interpretation, our findings link participants' accounts to established constructs in health behavior and technology adoption frameworks. However, stakeholder concerns regarding algorithmic bias, data privacy, and ethical implications suggest that AI models must be rigorously validated and co-designed with input from health care professionals, caregivers, and neurodivergent individuals [39,40]. Our emphasis on culturally sensitive AI models therefore reflects a call for AI tools to follow and extend the same principles of careful translation, local validation, and reflexive use that are increasingly expected of human-administered instruments, rather than holding AI to a fundamentally different standard [16,28,29,41-48].

Viewed in a broader context, our findings both converge with and extend international evidence on AI-enabled and digital approaches to autism and neurodevelopmental screening. Studies from high-income countries have similarly reported cautious optimism about AI and mobile tools, alongside concerns regarding algorithmic transparency, data privacy, and the risk of depersonalized care [9,29,49,50]. Our participants' preference for hybrid AI–human models echoes this literature, which consistently emphasizes the importance of human oversight and relational communication when delivering AI-generated outputs [9,28,40,51]. By bringing these strands together, this study contributes a LMIC-based perspective that both confirms global concerns and foregrounds context-specific implementation challenges [15,16,47,48].

From a reflexive standpoint, the research team's professional backgrounds in public health, autism screening, and health care systems may have sensitized the analysis toward issues of access, policy implications, and sociocultural context. This positionality informed both theme construction and interpretation, particularly in highlighting equity considerations for rural and underserved groups. Explicitly acknowledging positionality strengthens the transparency of analytic choices, in line with qualitative rigor guidelines [52].

AI tools have demonstrated high accuracy in speech pattern recognition, facial analysis, and behavioral assessments, improving early-stage ASD risk detection [9]. Consistent with earlier research, this study supports a hybrid approach, where AI functions as a decision-support tool rather than a diagnostic substitute [41,42,51].

AI Skepticism vs Parental Trust: Divergent Stakeholder Perspectives

One of the study's key findings is the contrast between health care professionals' AI skepticism and parents' greater openness to AI-driven screening. Many parents in the study viewed AI as a potential tool for early intervention, particularly in regions with limited specialist access. As one parent noted, "If it can tell me early that something is wrong, I don't care if it's a machine; as long as I can get help faster." This quote illustrates how parents interpret AI's value pragmatically, privileging speed and access over abstract concerns about reliability.

Conversely, many health care professionals in this study questioned AI's reliability, highlighting concerns that machine learning models might misinterpret symptoms, particularly for nonverbal children or those with atypical ASD presentations. This skepticism reflects broader concerns about algorithmic opacity and AI's inability to capture the complexity of neurodivergent traits [40,53-55].

Ethical Considerations in AI Adoption for ASD Screening

The integration of AI-driven screening tools raises pressing ethical concerns that must be addressed for widespread adoption. One major challenge is algorithmic bias, as many models are trained on Western datasets, limiting their applicability to diverse global settings [15]. This is particularly relevant in Egypt, where sociocultural differences in autism perception and diagnosis criteria may not align with Western-centric models. Participants' concerns mirror broader debates emphasizing the importance of cultural adaptation [47].

Data privacy emerged as another primary concern. Trust was closely tied to transparency in data governance and safeguards [40,51].

Finally, participants worried about misdiagnosis and emotional distress—false positives producing anxiety and false negatives delaying intervention. These dual anxieties underscore that participants' accounts should be understood not as statistical generalizations but as socially constructed experiences of risk and uncertainty [46].

Bridging Urban vs Rural Disparities in ASD Diagnosis

A key contribution of this study is its comparative analysis of urban and rural ASD screening barriers. While urban parents and health care professionals reported long waiting times and high costs as primary challenges, rural participants highlighted severe shortages of ASD specialists and a widespread lack of awareness about screening tools. These findings align with prior research on health care disparities in low-resource settings,

XSL•FO
RenderX

which suggests that AI-powered screening must be adapted to infrastructural limitations and literacy levels of diverse communities [8,43].

Theoretical Contributions: Behavior Change, Trust, and Stigma

This study contributes to theoretical frameworks explaining health technology adoption and behavior change in autism screening. Consistent with the HBM, parental decisions regarding AI-powered screening tools were shaped by their perceptions of susceptibility to ASD, the benefits of early detection, and concerns about potential barriers particularly stigma and data misuse [56]. Participants' concerns about stigma align with Goffman's theory of social labeling, where fear of judgment discourages disclosure [57].

Furthermore, trust-based adoption models and the TAM were evident in health care professionals' skepticism and parents' mixed acceptance driven by concerns over ease of use, usefulness, and trust in AI-generated outputs [55,58,59].

Policy Recommendations With Implementation Strategies

Building on these empirical and theoretical insights, several practical and policy-level recommendations can be made to support the ethical and effective implementation of AI-driven ASD screening in Egypt and similar contexts.

AI tools should be trained and validated on diverse datasets that reflect the cultural, linguistic, and behavioral characteristics of Egyptian children [13,15,16,39,49,50]. AI systems can follow the same approach by using locally collected training data, documenting the provenance and limitations of datasets, and involving Egyptian clinicians, caregivers, and autistic individuals in co-design and evaluation [16,28,29,47-49]. Such steps would help ensure that AI does not simply reproduce Western diagnostic assumptions but is genuinely responsive to local sociocultural realities [15,16,28,29,48,49].

AI adoption must be underpinned by national legal frameworks that regulate data storage, privacy, and consent. Without such frameworks, participants' anxieties about surveillance and discrimination could limit uptake, as also reported in other LMIC contexts [15].

Participants emphasized that parents' trust depends on transparent communication regarding how AI screening tools operate, where data are stored, and who has access. Transparent governance and communication have been shown to improve uptake of AI-based health tools [51].

A recurring theme was the value of combining AI outputs with human judgment. Parents and health care providers agreed that AI should assist, not replace, clinicians. This aligns with global guidance emphasizing AI as an augmentative tool [9], and reflects participants' preference for "shared accountability" between humans and technology.

To address diagnostic inequities, AI-enabled screening should be deployed in primary care facilities and community health centers in rural areas. Participants in rural settings described stigma, lack of awareness, and absence of specialists as barriers, reinforcing the importance of deploying accessible, low-literacy screening tools. Training community health providers to use AI tools could decentralize early detection and improve referral pathways. This approach mirrors other successful models of digital health integration in low-resource contexts [54].

Successful AI adoption will depend not only on clinical readiness but also on public understanding and digital competence. Parents and caregivers expressed both curiosity and apprehension about AI, underscoring the need for structured awareness campaigns. Digital literacy training has been shown to improve uptake of AI tools in LMIC health care settings [47].

Stigma remains a powerful deterrent to early ASD diagnosis. Participants' accounts revealed that some families avoid clinics out of fear of labeling, echoing prior national studies on autism stigma in Egypt [8].

Health care providers and frontline staff require training not only in the technical use of AI tools but also in communicating AI results to families with empathy and sensitivity. Participants stressed the risk of "cold delivery" of AI outputs; thus, training must emphasize the human element of care to prevent alienation. Building capacity in both technical literacy and interpersonal skills will ensure AI integration enhances, rather than undermines, the therapeutic alliance between families and health care systems.

Strengths

This study is among the first to explore stakeholder perspectives on AI-powered autism screening in Egypt, offering valuable insights for the development and implementation of such tools in LMICs. The use of reflexive thematic analysis allowed for the identification of context-specific enablers and barriers, drawing from the lived experiences of both health care professionals and parents. Purposive sampling ensured that participants represented a wide range of geographic and professional backgrounds, increasing the diversity of perspectives captured. Together, these design choices support conceptual transferability to other LMIC settings facing similar diagnostic and infrastructural challenges. All parent participants had children aged 6 years or younger, anchoring parental perspectives in early detection contexts.

Limitations

As with many qualitative designs, findings are context-bound and not intended for statistical generalization. Instead, the aim was to generate nuanced, situated understandings that may have conceptual transferability to similar LMIC contexts. Although the sample achieved geographic variation, the number of FGDs was necessarily constrained by logistical and resource limitations, which shaped the scope of representation. The sample was not designed to be proportionate to population distributions across urban and rural areas and was not powered for formal statistical comparisons between geographic strata; rather, rural and urban participants were included to surface key contextual patterns and implementation considerations. As discussed above, the research team's positionalities, including professional experience in public health, autism research, and digital health, may have influenced data interpretation, particularly in prioritizing equity and access themes; this

XSL•FO
RenderX

reflexive awareness was explicitly documented and used to critically interrogate analytic decisions.

Although all individuals who were invited agreed to participate, which may reflect greater interest or comfort with discussing AI and autism, we attempted to mitigate this potential selection bias by recruiting across multiple governorates, facility types, and professional roles to capture a wide range of perspectives. However, socioeconomic position was not systematically measured or analyzed, and any socioeconomic variation in the sample is therefore inferred indirectly through differences in governorate and service setting rather than examined explicitly. In addition, because all participating parents had children aged 6 years or younger, perspectives related to later intervention phases were not represented, and we were not able to systematically compare views on early detection vs long-term intervention. Similarly, the numbers of participants within individual health care professional subgroups were not sufficient to support robust within-profession comparisons, and the analysis was therefore conducted at the combined health care provider level, with any role-specific nuances interpreted cautiously.

Additionally, while separate FGDs for health care professionals and parents minimized power imbalances, the group setting may have led to social desirability bias, with participants potentially moderating their comments in line with perceived group norms. The presence of AI-related stigma and limited public familiarity with AI screening could also have influenced the openness of responses, despite measures taken to build rapport. Future research could incorporate longitudinal designs and follow-up interviews to explore how attitudes evolve as AI screening tools are introduced and piloted in local health care systems.

These limitations do not undermine the validity of the findings but rather frame them as contextually situated insights, highlighting the value of qualitative inquiry for shaping ethical and culturally sensitive AI interventions in LMICs.

Conclusion

This study highlighted how feasibility, acceptability, and perceived ethical and practical enablers and barriers shape stakeholder attitudes toward implementing AI-powered tools for early ASD screening within the Egyptian health system. While parents were generally open to AI tools for early detection, health care professionals expressed caution, emphasizing the need for rigorous validation, culturally adapted algorithms, and integration within existing diagnostic pathways.

AI-powered tools hold significant promise for addressing diagnostic delays and specialist shortages in Egypt and other LMICs. However, their success depends on transparent governance, stakeholder engagement in co-design, and human oversight to safeguard against bias and misuse.

By integrating technological innovation with culturally sensitive implementation strategies, AI-powered screening tools could contribute to more equitable, timely, and accurate autism identification. In practical terms, this entails embedding AI within hybrid human–AI care models, aligning governance frameworks with community expectations and ethical standards, and sustaining engagement with families and clinicians to ensure that AI enhances, rather than replaces, human-centered care.

Acknowledgments

The authors express their sincere appreciation to all health care providers for their willingness to engage and cooperate during the study. Special thanks to all parents who gave their time to participate in this study, for whom it was conducted.

The authors disclose that ChatGPT (OpenAI; accessed January 2026) was used only for language polishing (improving grammar, clarity, and concision) during manuscript preparation. The authors reviewed and edited all artificial intelligence (AI)–assisted text and take full responsibility for the final content. No generative AI was used for data collection, data analysis, visualization, or the generation of study data, analyses, or results. No identifiable participant data or raw transcripts were entered into the tool.

Data Availability

The datasets used and/or analyzed for the current study are anonymous and are available from the corresponding author on request.

Funding

This study received financial support from Ulster University, Londonderry, Northern Ireland, United Kingdom through Ulster International Science Partnerships Fund (ISPF) funding (cost centre 71799) for the participant travel expense and goodwill payment and publication of this study.

Authors' Contributions

AMM was responsible for conceptualization and contributed to the formulation of the research questions, study themes, data curation, methodology, project administration, supervision, and review and editing of the manuscript. PY contributed to conceptualization, assisted in formulating the research questions and study themes, and participated in project administration and manuscript review and editing. PC contributed to conceptualization, project administration, and manuscript review and editing. GAA assisted with data curation, formal analysis, software, and validation. SMEK, the lead researcher for developmental psychology, maintained a reflexive journal throughout the familiarization process to document assumptions, emotional responses, and evolving interpretations. AE assisted with participant recruitment, facilitated focus groups, and contributed to drafting the

XSL•FO
RenderX

original manuscript. EAA contributed to the formulation of study themes and drafting of the original manuscript. NAE assisted with participant recruitment and facilitation of focus groups. ASE assisted with participant recruitment, facilitation of focus groups, reference writing, and drafting of the original manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Expanded methodological details.
[DOCX File , 25 KB - jmir_v28i1e82564_app1.docx]

Multimedia Appendix 2
Focus group and interview discussion guide.
[DOCX File , 18 KB - jmir_v28i1e82564_app2.docx]

Multimedia Appendix 3
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[PDF File (Adobe PDF File), 478 KB - jmir_v28i1e82564_app3.pdf]

References

1. Silverman C. NeuroTribes: The legacy of autism and the future of neurodiversity by steve silberman. Anthropological Quarterly 2015;88(4):1111-1121. [doi: 10.1353/anq.2015.0057]
2. Zwaigenbaum L, Bauman ML, Stone WL, Yirmiya N, Estes A, Hansen RL. Early identification of autism spectrum disorder: recommendations for practice and research. Pediatrics 2015;136(Suppl 1):S10-S40 [FREE Full text] [doi: 10.1542/peds.2014-3667C] [Medline: 26430168]
3. Dawson G, Rogers S, Munson J, Smith M, Winter J, Greenson J. Randomized, controlled trial of an intervention for toddlers with autism: the Early Start Denver Model. Pediatrics 2010;125(1):e17-e23 [FREE Full text] [doi: 10.1542/peds.2009-0958] [Medline: 19948568]
4. Lordan R, Storni C, De Benedictis CA. Autism spectrum disorders: diagnosis and treatment. In: Autism Spectrum Disorders. Brisbane: Exon Publications; 2021.
5. Brookman-Frazee L, Drahota A, Stadnick N, Palinkas LA. Therapist perspectives on community mental health services for children with autism spectrum disorders. Adm Policy Ment Health 2012;39(5):365-373 [FREE Full text] [doi: 10.1007/s10488-011-0355-y] [Medline: 21533846]
6. Metwally AM, Abdallah AM, El-Din EMS, Zeid DA, Khadr Z, Elshaarawy GA. Screening and determinant of suspected developmental delays among Egyptian preschool-aged children: a cross-sectional national community-based study. BMC Pediatr 2023;23(1):521 [FREE Full text] [doi: 10.1186/s12887-023-04335-0] [Medline: 37858055]
7. Metwally AM, Salah El-Din EM, Sami SM, Abdelraouf ER, Sallam SF, Elsaeid A. Mapping autism in Egypt: population-based insights into prevalence, risk determinants, and severity among children aged 1-12 years. Mol Autism 2025;16(1):32 [FREE Full text] [doi: 10.1186/s13229-025-00665-1] [Medline: 40442748]
8. Metwally AM, Nassar MS, Salah El-Din EM, Abdallah AM, Khadr Z, Abouelnaga MW. National screening for developmental delays and their determinants among Egyptian school age children: a step towards implementing life skills programs. PLoS One 2023;18(9):e0287315 [FREE Full text] [doi: 10.1371/journal.pone.0287315] [Medline: 37725608]
9. Topol E. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. London: Hachette UK; 2019.
10. Khudhur DD, Khudhur SD. The classification of autism spectrum disorder by machine learning methods on multiple datasets for four age groups. Measurement: Sensors 2023;27:100774. [doi: 10.1016/j.measen.2023.100774]
11. Alves CL, Martinelli T, Sallum LF, Rodrigues FA, Toutain T, Porto JAM. Multiclass classification of autism spectrum disorder, attention deficit hyperactivity disorder, and typically developed individuals using fMRI functional connectivity analysis. PLoS One 2024;19(10):e0305630 [FREE Full text] [doi: 10.1371/journal.pone.0305630] [Medline: 39418298]
12. Duda M, Ma R, Haber N, Wall DP. Use of machine learning for behavioral distinction of autism and ADHD. Transl Psychiatry 2016;6(2):e732. [doi: 10.1038/tp.2015.221] [Medline: 26859815]
13. Thabtah F, Peebles D. Early autism screening: a comprehensive review. Int J Environ Res Public Health 2019;16(18):3502 [FREE Full text] [doi: 10.3390/ijerph16183502] [Medline: 31546906]
14. Dhar S, Sarkar U. Safeguarding data privacy and informed consent. In: Advances in Electronic Government, Digital Divide, and Regional Development. Hershey: IGI Global; 2025.
15. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366(6464):447-453 [FREE Full text] [doi: 10.1126/science.aax2342] [Medline: 31649194]

XSL•FO
RenderX

16. Labrique AB, Wadhwani C, Williams KA, Lamptey P, Hesp C, Luk R. Best practices in scaling digital health in low and middle income countries. Global Health 2018;14(1):103 [FREE Full text] [doi: 10.1186/s12992-018-0424-z] [Medline: 30390686]
17. Rogers SJ, Estes A, Lord C, Vismara L, Winter J, Fitzpatrick A. Effects of a brief Early Start Denver Model (ESDM)–based parent intervention on toddlers at risk for autism spectrum disorders: a randomized controlled trial. J Am Acad Child Adolesc Psychiatry 2012;51(10):1052-1065 [FREE Full text] [doi: 10.1016/j.jaac.2012.08.003] [Medline: 23021480]
18. Holden RJ, Karsh BT. The technology acceptance model: its past and its future in health care. J Biomed Inform 2010;43(1):159-172 [FREE Full text] [doi: 10.1016/j.jbi.2009.07.002] [Medline: 19615467]
19. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative aata collection and analysis in mixed method implementation research. Adm Policy Ment Health 2015;42(5):533-544 [FREE Full text] [doi: 10.1007/s10488-013-0528-y] [Medline: 24193818]
20. Krueger RA. Focus Groups: A Practical Guide for Applied Research. California: Sage Publications; 2014.
21. Patton MQ. Qualitative Research & Evaluation Methods: Integrating Theory and Practice. California: Sage Publications; 2014.
22. Guest G, Bunce A, Johnson L. How many interviews are enough? Field Methods 2006;18(1):59-82. [doi: 10.1177/1525822X05279903]
23. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B. Saturation in qualitative research: exploring its conceptualization and operationalization. Quality & Quantity 2018;52(4):1893-1907 [FREE Full text] [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
24. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019;11(4):589-597. [doi: 10.1080/2159676x.2019.1628806]
25. Shenton AK. Strategies for ensuring trustworthiness in qualitative research projects. Educ Inf 2004;22(2):63-75. [doi: 10.3233/EFI-2004-22201]
26. Levitt H, Bamberg M, Creswell J, Frost D, Josselson R, Suárez-Orozco C. Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018;73(1):26-46. [doi: 10.1037/amp0000151]
27. Morse JM. Critical analysis of strategies for determining rigor in qualitative inquiry. Qual Health Res 2015;25(9):1212-1222. [doi: 10.1177/1049732315588501] [Medline: 26184336]
28. Morley J, Machado CCV, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: a mapping review. Soc Sci Med (1982) 2020;260:113172. [doi: 10.1016/j.socscimed.2020.113172] [Medline: 32702587]
29. Floridi L, Cowls J. A unified framework of five principles for AI in society. Harv Data Sci Rev 2019. [doi: 10.1162/99608f92.8cd550d1]
30. International Ethical Guidelines for Health-Related Research Involving Humans. Geneva: Council for International Organizations of Medical Sciences (CIOMS); 2016.
31. Metwally AM, Amer HA, Salama HI, Abd El Hady SI, Alam RR, Aboulghate A. Egyptian patients'/guardians' experiences and perception about clinical informed consent and its purpose: cross sectional study. PLoS One 2021;16(6):e0252996 [FREE Full text] [doi: 10.1371/journal.pone.0252996] [Medline: 34125842]
32. Egypt in 2024: Population and Urbanization. Cairo: Central Agency for Public Mobilization and Statistics (CAPMAS); 2024.
33. Statistical Yearbook 2023: Governorates of Egypt. Cairo: Central Agency for Public Mobilization and Statistics (CAPMAS); 2023.
34. Urban population (% of total) – Arab Republic of Egypt. Washington DC: World Bank; 2023.
35. Egypt Urban Expansion and Sustainable Development Report. Nairobi: United Nations Human Settlements Programme (UN-Habitat); 2021.
36. Egypt Human Development Report 2021: Development, a Right for All? Egypt's Pathways and Prospects. Cairo: United Nations Development Programme (UNDP); 2021.
37. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
38. Wallis KE, Guthrie W. Identifying autism spectrum disorder in real-world health care settings. Pediatrics 2020;146(2):e20201467 [FREE Full text] [doi: 10.1542/peds.2020-1467] [Medline: 32632020]
39. Ayub R, Immamul Ansarullah S, Samdani F, Akhtar A, Afzal M, Moinuddin Ahmed S. Enhancing early detection of autistic spectrum disorder in children using machine learning approaches. J King Saud Univ Sci 2024;36(10):103468 [FREE Full text] [doi: 10.1016/j.jksus.2024.103468]
40. Morley J, Elhalal A, Garcia F, Kinsey L, Mökander J, Floridi L. Ethics as a service: a pragmatic operationalisation of AI Ethics. Minds Mach (Dordr) 2021;31(2):239-256 [FREE Full text] [doi: 10.1007/s11023-021-09563-w] [Medline: 34720418]
41. Fombonne E. The changing epidemiology of autism. J Appl Res Intellect Disabil 2005;18(4):281-294. [doi: 10.1111/j.1552-6909.2012.01399.x] [Medline: 2861175]

XSL•FO
RenderX

42. Lyall K, Croen L, Daniels J, Fallin MD, Ladd-Acosta C, Lee BK. The changing epidemiology of autism spectrum disorders. Annu Rev Public Health 2017;38(1):81-102 [FREE Full text] [doi: 10.1146/annurev-publhealth-031816-044318] [Medline: 28068486]
43. Hermaszewska S, Sin J. End-user perspectives on the development of an online intervention for parents of children on the autism spectrum. Autism 2021;25(5):1234-1245 [FREE Full text] [doi: 10.1177/1362361320984895] [Medline: 33423522]
44. Metwally AM, Helmy MA, Salah El-Din EM, Saleh RM, Abdel Raouf ER, Abdallah AM. National screening for Egyptian children aged 1 year up to 12 years at high risk of autism and its determinants: a step for determining what ASD surveillance needs. BMC Psychiatry 2023;23(1):471 [FREE Full text] [doi: 10.1186/s12888-023-04977-5] [Medline: 37381024]
45. Lincoln YS. Naturalistic Inquiry. California: SAGE Publications, Inc; 1985.
46. Braun V, Clarke V. One size fits all? What counts as quality practice in (reflexive) thematic analysis? Qual Res Psychol 2020;18(3):328-352. [doi: 10.1080/14780887.2020.1769238]
47. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Iii HD, et al. Datasheets for datasets. Commun ACM 2021;64(12):86-92 [FREE Full text] [doi: 10.1145/3458723]
48. Birhane A. Algorithmic injustice: a relational ethics approach. Patterns 2021;2(2):100205 [FREE Full text] [doi: 10.1016/j.patter.2021.100205] [Medline: 33659914]
49. Wies B, Landers C, Ienca M. Digital mental health for young people: a scoping review of ethical promises and challenges. Front Digit Health 2021;3:697072 [FREE Full text] [doi: 10.3389/fdgth.2021.697072] [Medline: 34713173]
50. Li Y, Zhang T, Yang Y, Gao Y. Artificial intelligence-aided decision support in paediatrics clinical diagnosis: development and future prospects. J Int Med Res 2020;48(9):300060520945141 [FREE Full text] [doi: 10.1177/0300060520945141] [Medline: 32924683]
51. Floridi L, Cowls J. A unified framework of five principles for AI in society. In: Machine Learning and the City: Applications in Architecture and Urban Design. Hoboken: Wiley; 2022:535-545.
52. Lincoln YS. Naturalistic Inquiry. California: SAGE Publications, Inc; 1985.
53. Achenie LEK, Scarpa A, Factor RS, Wang T, Robins DL, McCrickard DS. A machine learning strategy for autism screening in toddlers. J Dev Behav Pediatr 2019;40(5):369-376 [FREE Full text] [doi: 10.1097/DBP.0000000000000668] [Medline: 30985384]
54. Li Y, Zhang T, Yang Y, Gao Y. Artificial intelligence-aided decision support in paediatrics clinical diagnosis: development and future prospects. J Int Med Res 2020;48(9):300060520945141 [FREE Full text] [doi: 10.1177/0300060520945141] [Medline: 32924683]
55. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: four longitudinal field studies. Manage Sci 2000;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
56. Rosenstock IM. The health belief model and preventive health behavior. Health Educ Monogr 1974;2(4):354-386 [FREE Full text] [doi: 10.1177/109019817400200405]
57. Goffman E. Stigma: Notes on the Management of Spoiled Identity. New York: Simon and Schuster; 2009.
58. McKnight DH, Choudhury V, Kacmar C. Developing and validating trust measures for e-Commerce: an integrative typology. Inf Syst Res 2002;13(3):334-359. [doi: 10.1287/isre.13.3.334.81]
59. Metwally AM, Abdel-Latif GA, Eletreby L, Aboulghate A, Mohsen A, Amer HA. Egyptians' social acceptance and consenting options for posthumous organ donation: a cross-sectional study. BMC Med Ethics 2020;21(1):49 [FREE Full text] [doi: 10.1186/s12910-020-00490-6] [Medline: 32539704]

Abbreviations

AI: artificial intelligence
ASD: autism spectrum disorder
COREQ: Consolidated Criteria for Reporting Qualitative Research
FGD: focus group discussion
HBM: Health Belief Model
LMIC: low- and middle-income country
NRC: National Research Centre
RQ1: research question 1
RQ2: research question 2
TAM: Technology Acceptance Model

XSL•FO
RenderX

Edited by A Stone; submitted 18.Aug.2025; peer-reviewed by F Jorge, J Wang; comments to author 02.Oct.2025; revised version received 06.Jan.2026; accepted 30.Jan.2026; published 07.Apr.2026.

Please cite as:
Yogarajah P, Metwally AM, Chaurasia P, Elshaarawy GA, El Khateeb SM, Ashaat EA, Elsaeid A, Elghareeb NA, ElRifay AS
Feasibility and Acceptability of AI-Powered Tools for Early Autism Screening in Egypt: Semistructured Focus Group Study
J Med Internet Res 2026;28:e82564
URL: https://www.jmir.org/2026/1/e82564
doi:10.2196/82564
PMID:

©Pratheepan Yogarajah, Ammal M Metwally, Priyanka Chaurasia, Ghada A Elshaarawy, Shereen M El Khateeb, Engy A Ashaat, Amal Elsaeid, Nahed A Elghareeb, Amira S ElRifay. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Facility-Level Associations Between Use of a Digital Health Platform for Voluntary Counseling and Testing and HIV Testing Outcomes in the Urban Primary Health Care Centers of Guangzhou, China: Cross-Sectional Study

Ye Chen[1,2*], BSc; Yu-Fei Wang[1,2*], BSc; Yu-Zhou Gu[3*], MSc; Xiao-Ru Fan[4], MSc; Yong-Heng Lu[5,6], BSc; Ju-Shuang Li[7], PhD; Jun Wang[8], MSc; Zhi-Ye Lin[9], MSc; Chun-Li Zhang[1,2], MSc; Meng-Die Li[10], MSc; Ning-Jun Ren[1,2], MSc; Zhi-Zhong Luo[1,2], MSc; Rui-Fei Zuo[1,2], BSc; Fu-Chuan Xiong[1,2], BSc; Yu-Tong Kang[1,2], BSc; Xiu-Rong Lin[1,2], BSc; Chun Hao[1,2], PhD

[1]Department of Medical Statistics, School of Public Health, Sun-Yat-Sen University, Guangzhou, China

[2]Sun Yat-Sen Global Health Institute, Institute of State Governance| Institute of International and Regional Studies, Sun-Yat-Sen University, Guangzhou, China

[3]Guangzhou Center for Disease Control and Prevention (Guangzhou Health Supervision Institute), Guangzhou, China

[4]Liwan District Center for Disease Control and Prevention, Guangzhou, China

[5]Guangzhou Lingnan Community Support Center, Guangzhou, China

[6]Kangyuan Community Support Center of Yuexiu District, Guangzhou, China

[7]Department of Epidemiology and Health Statistics, School of Public Health, Hengyang Medical School, Hengyang, China

[8]Hospital of Stomatology, Sun Yat-Sen University, Guangzhou, China

[9]Akeso Biopharma Co, Ltd, Zhongshan, China

[10]Department of Medical Record Management, the Seventh Affiliated Hospital, Sun Yat-Sen University, Shenzhen, China

[*]these authors contributed equally

Corresponding Author:
Chun Hao, PhD
Department of Medical Statistics
School of Public Health
Sun-Yat-Sen University
74 Zhongshan 2nd Rd Yuexiu District
Guangzhou, 510080
China
Phone: 86 87332517
Email: haochun@mail.sysu.edu.cn

Abstract

Background: WellTest, a digital health platform designed to facilitate HIV voluntary counseling and testing (VCT) services, has been widely implemented in Guangzhou, China. However, the extent of its use by primary health care centers (PHCs) and the facility-level associations between WellTest use and HIV testing outcomes remain unclear.

Objective: This study aimed to assess the use of WellTest at the facility level across PHCs in urban Guangzhou, China, and to explore the associations between WellTest use and HIV testing outcomes.

Methods: We obtained data on VCT services visits from the National HIV/AIDS Prevention and Control Information System and the WellTest platform between January 1 and December 31, 2022. Two facility-level HIV testing outcomes, the HIV testing volume and HIV positivity rate, were calculated for each PHC. A structured questionnaire collected data on the PHCs' characteristics. Multilevel negative binomial regression and zero-inflated gamma models were used to examine associations between WellTest use and 2 HIV testing outcomes at the facility level.

Results: A total of 81 PHCs across 5 urban districts in Guangzhou were included. WellTest was used for 7997 active consultations and 7969 HIV tests that resulted in 157 newly diagnosed HIV cases, for an overall positivity rate of 2.0%, during 2022. The median share of clients booking via WellTest was 71% (IQR 34%-98%), and 81% (66/81) of PHCs offered online slots on all

XSL•FO
RenderX

service days without mandates. PHCs that actively confirmed appointments had significantly lower HIV testing volumes compared with those that took no actions after clients scheduled appointments (incidence rate ratio 0.75, 95% CI 0.58-0.97; *P*=.03) but exhibited higher HIV positivity rates (β=.39, 95% CI 0.02-0.76; *P*=.04). Additionally, PHCs where ≥50% of clients used WellTest to schedule appointments showed higher HIV positivity rates than those with lower uptake (β=1.11; 95% CI 0.82-1.40; *P*<.001).

Conclusions: WellTest showed substantial use among VCT clients and providers in urban Guangzhou PHCs, with greater use of digital appointments and counselor follow-up linked to increased HIV positivity rates. Strategies to optimize institutional adoption may help address stigma-related barriers, strengthen engagement in HIV testing, and support HIV case finding in primary health care settings.

(J Med Internet Res 2026;28:e83662) doi:10.2196/83662

KEYWORDS

HIV testing; factor analysis; voluntary counseling and testing; health service; digital health

Introduction

HIV/AIDS has been a major public health problem since it was first reported in 1981. According to the Joint United Nations Program on HIV/AIDS [1], by the end of 2022, 39 million people were living with HIV. In China, by the end of 2022, there were an estimated 1.223 million people living with HIV, with 107,000 new infections and 30,000 AIDS-related deaths recorded in the same year [2]. At the provincial level, Guangdong Province has been one of the regions most affected by HIV/AIDS in China, accounting for 18.4% (11,806/64,170) of HIV infections in China [3,4] in 2018. Moreover, with many unaware of their HIV status, access to treatment was delayed and the risk of transmission increased [1]. As of 2024, 33.8% of the high-risk population in China remain unaware of their HIV status [5]. The UNAIDS 95-95-95 targets aim to eliminate the AIDS epidemic by 2030 [6], with the first target ensuring 95% of individuals living with HIV are aware of their status.

Voluntary counseling and testing (VCT) has been shown to contribute to the identification of potential individuals with HIV [7,8]. Since the establishment of a national VCT network in China in 2004, the number of VCT sites has grown to 11,319 [9]. However, workforce shortages and uneven distribution have impeded access to services [10]. To enhance accessibility, a 2009 national strategy promoted decentralizing VCT delivery to primary health care centers (PHCs) [11]. Despite this, the use of VCT services in PHCs remained low, hindered by insufficient awareness, stigma, and concerns over the quality of care [10,12-14].

Digital health interventions have shown promise in HIV/AIDS prevention due to their feasibility and acceptability. Tools such as social networks, online courses, and digital health apps have been proven to promote HIV testing [15-18], particularly among the high-risk population, by offering privacy, convenience, and efficiency [19]. Additionally, these interventions have improved the accessibility of disease testing [20,21]. However, research findings on whether digital health use can facilitate disease testing or screening at the facility level remains scarce [22].

Guangzhou, the capital of Guangdong Province and a major national central city with >18 million residents [23], has leveraged its advanced technological infrastructure to pioneer digital innovations aimed at enhancing access to VCT services within primary health care settings. In 2016, the Guangzhou Centers for Disease Control and Prevention (CDC) and Lingnan Partners Community Support Center (a community-based organization) developed WellTest, originating from the Lingnan website [24]. It is a digital platform for VCT services delivery for PHCs, by providing functions for online VCT services appointment booking, online-to-offline referral, clinic-based testing registration, and delivery of HIV test results. Since its expansion in 2018, the platform has enhanced VCT accessibility, facilitating 59,846 online consultations by 2023 and covering 83.6% of total VCT services visits in Guangzhou [25].

However, the implementation of a VCT digital health platform in PHCs and their impact on HIV testing outcomes at the facility level remain unclear. This study aims to examine the use of the WellTest platform in urban Guangzhou PHCs and its association with HIV testing outcomes at the facility level, providing novel evidence on how digital health use can optimize HIV testing services in primary health care settings.

Methods

Recruitment of Study Sites and Participants

The recruitment process was facilitated through the Guangzhou CDC. We identified all 81 public PHCs providing VCT services across 5 urban districts of Guangzhou (Tianhe, Yuexiu, Liwan, Baiyun, and Haizhu districts) in 2022. For each PHC, 1 primary VCT counselor was invited to complete the questionnaire survey for the data collection. Eligible counselors were those who agreed to participate in our study and had been working in VCT services for at least 6 months during the study period to ensure they had sufficient experience with the WellTest platform.

All identified PHCs and eligible counselors were formally invited to participate in this study with the help of Guangzhou CDC.

Data Sources and Data Collection

Data on VCT services visits to these PHCs were obtained from two sources: the National HIV/AIDS Prevention and Control Information System and the digital health platform WellTest, spanning from January 1 to December 31, 2022. Individuals with incomplete HIV testing records were excluded from the analysis.

A questionnaire survey was conducted with VCT counselors from 81 PHCs to collect information on PHC characteristics,

XSL•FO
RenderX

including digital health use, institutional details, and counselor profiles.

The recruitment and data collection flow is depicted in Figure 1.

Figure 1. Flowchart of the recruitment and data collection process. CDC: Centers for Disease Control and Prevention; PHC: primary health care center; VCT: voluntary counseling and testing.

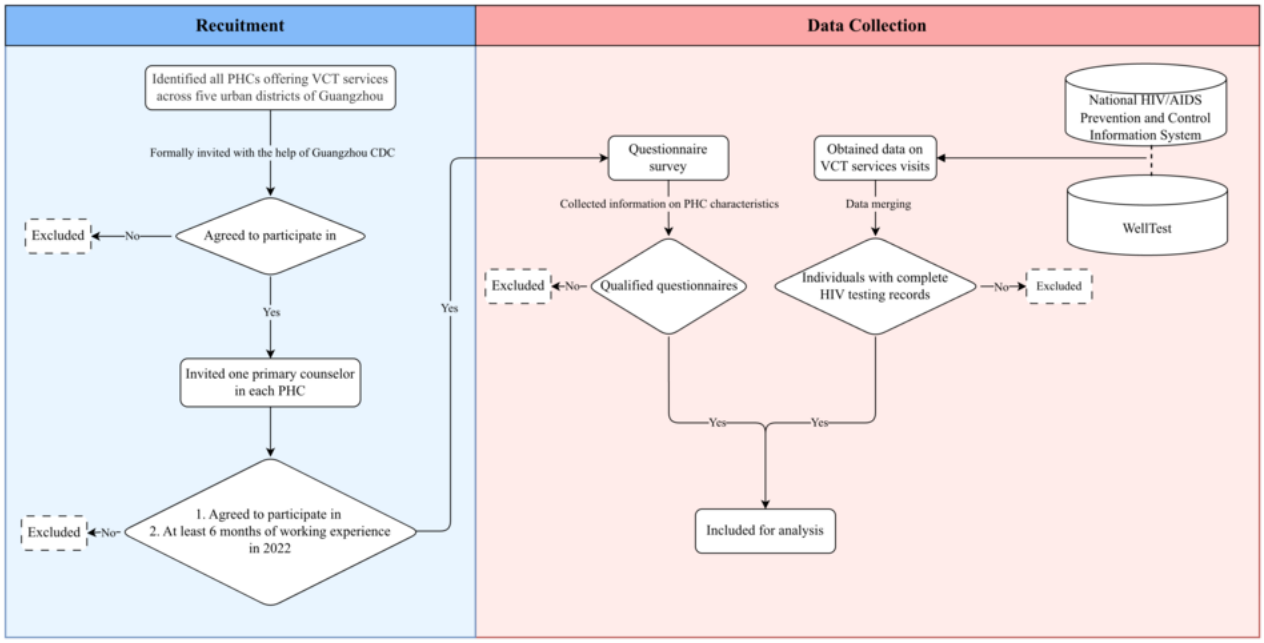

Measurements

Two facility-level HIV testing outcomes were assessed for each PHC:

- HIV testing volume; the total number of HIV tests conducted at each PHC
- HIV positivity rate; calculated as the number of newly diagnosed HIV cases/the number of HIV tests conducted

We merged the two aforementioned databases and then calculated the number of HIV tests conducted and the number of newly diagnosed HIV cases for each PHC providing VCT services in 2022.

Associated factors include digital health use, institutional details, and counselor profiles. Digital health use included the proportion of clients scheduling VCT appointments through WellTest, whether to set online appointment slots for all VCT service days via WellTest, and counselor responses to WellTest appointments (take no actions or confirm appointment by contacting clients via phone or SMS text message). Counselor profiles encompassed age, gender, educational level, major, working experience, and proportion of VCT services within their workload. PHCs' institutional details consisted of five domains. The first is organizational structure, that is, the ownership type of the PHC (public or private) and whether the PHC prioritizes VCT services in routine practice. The second is human resources, that is, the number of VCT counselors and the frequency of counselor turnover during the study period. The third is infrastructure, that is, the availability of dedicated phlebotomy rooms and the availability of private and dedicated counseling spaces. The fourth is location, that is, the walking time to the nearest subway station (≤15 minutes or >15 minutes, based on the commonly used standard of approximately 15 minutes per km, which is generally considered a walkable distance). The distance to the nearest subway station and geographical coordinates were obtained using Amap, a widely used digital mapping service in China. The fifth is service delivery, that is, the average turnaround time for HIV test results, whether services were available on weekends, the number of times counselors were unable to provide VCT services for personal reasons in 2022, and the availability of posttest gifts (HIV prevention kits, including condoms, lubricants, and informational materials) for testers.

Statistical Analysis

Two maps were created to illustrate the HIV testing volume and HIV testing rate for each PHC. Descriptive statistics were used to depict VCT digital health use and the characteristics of counselors and PHCs. Given that HIV testing volume at the facility level exhibited overdispersion and the HIV positivity rate was highly right skewed with an excess of zeros, we applied multilevel negative binomial regression [26] and zero-inflated gamma models [27]. These methods were used to examine the associations between WellTest use and HIV testing outcomes, including testing volume and positivity rate, across primary health care facilities. The univariate analysis was conducted to assess associations between VCT digital health use, the characteristics of VCT counselors, and PHCs with 2 HIV testing outcomes. Subsequently, variables at a significance level of $P<.10$ from the univariate analysis, along with 3 variables of the "digital health use" domain, were included in the multivariable analysis to investigate the adjusted association between the use of WellTest and HIV testing among PHCs in urban Guangzhou. The population under the jurisdiction of each PHC, representing the facility-level service catchment area and

XSL•FO
RenderX

underlying demand of HIV testing, was included as a confounder. $P<.05$ was considered statistically significant in the multivariable analysis. To test the robustness of our findings, we conducted a sensitivity analysis excluding data from Liwan district due to its limited sample size (n=2), following the same steps as the primary analysis. All analyses were performed using R (version 4.4.3; R Foundation for Statistical Computing).

Ethical Considerations

This study was approved by the ethics committee of Sun Yat-sen University (Institutional Review Board number 054/19; February 28, 2019). All questionnaires were completed with the counselors' informed consent, and the analysis of data on VCT services visits was approved by Guangzhou CDC, with no additional ethical review required. All data used were deidentified and analyzed solely within the research team to ensure the confidentiality and privacy of the data.

Results

Recruitment

All PHCs offering VCT services in the urban area of Guangzhou in 2022 were included in our study (N=81), and all 81 invited counselors from the respective PHCs agreed to participate in this study. No PHCs or counselors were excluded from the analysis due to poor questionnaire quality or logical inconsistencies.

HIV Testing Among PHCs Offering VCT Services

After excluding 5 individuals with incomplete HIV testing records, 81 urban PHCs of Guangzhou offered a total of 7997 person-times of consultations and 7969 HIV tests, identified 157 new diagnoses, and resulted in an HIV-positive rate of 2.0% in 2022. Figure 2 shows the HIV testing volumes for each urban PHC in 2022, varying from 2 to 609 (median 71, IQR 45-112). Figure 3 presents the HIV-positive rates for each urban PHC in 2022, ranging from 0% to 16.7% (median 0, IQR 0-0).

Figure 2. HIV testing volumes for each primary health care center offering voluntary counseling and testing services in the urban area of Guangzhou, China, in 2022.

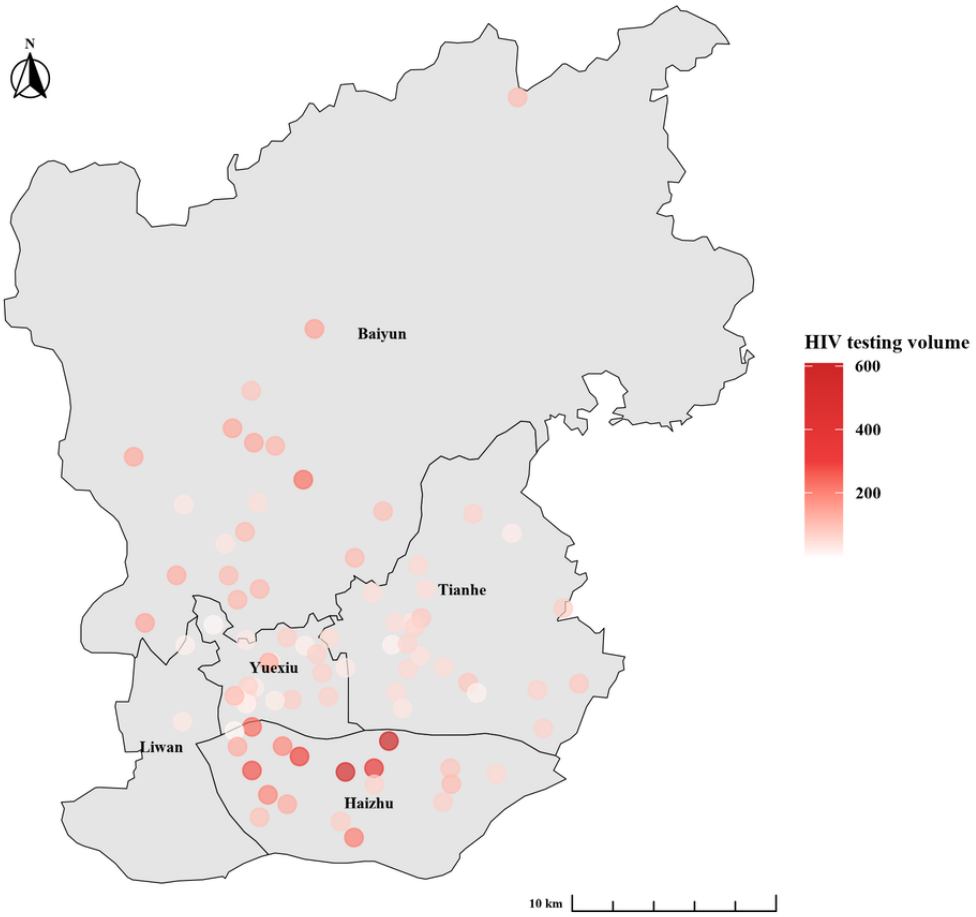

XSL•FO
RenderX

Figure 3. HIV positivity rates for each primary health care center offering voluntary counseling and testing services in the urban area of Guangzhou, China, in 2022.

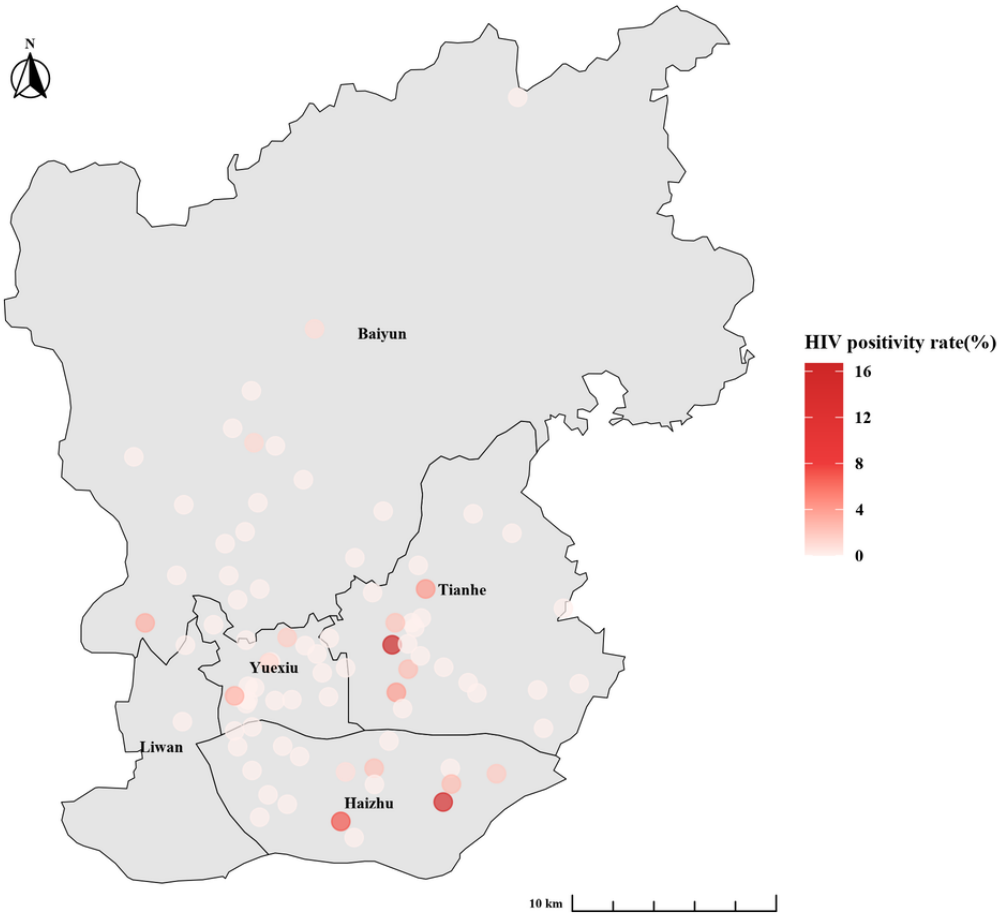

Digital Health Use for VCT Services Delivery

Across the 81 PHCs, the median number of clients actively seeking VCT services was 71 (IQR 44-112). Among them, a median of 30 (IQR 10-62) clients scheduled appointments via WellTest, corresponding to a high proportion of scheduling VCT appointments through WellTest (median 71%, IQR 34%-98%). On the provider side, the median number of days per week offering VCT services was 3 (IQR 2-5) days, and a median of 3 (IQR 2-5) appointment days weekly was set in WellTest. The median proportion of service days with available online slots was 100% (IQR 100%-100%).

As shown in Table 1, most PHCs (66/81, 81%) opened online appointment slots on all service days. After online appointments were made, nearly half (39/81, 48%) reported counselor follow-up.

XSL•FO
RenderX

Table 1. Characteristics of primary health care centers offering voluntary counseling and testing services in the urban area of Guangzhou, China, in 2022 (N=81).

Characteristics	Primary health care centers
Digital health use	
Proportion of clients scheduling VCT [a] services through WellTest (%), median (IQR)	71 (34-98)
Set online appointment slots on all VCT service days via WellTest, n (%)	
No	15 (19)
Yes	66 (81)
Action taken by counselors upon receiving appointments from WellTest, n (%)	
Take no action	42 (52)
Confirm appointment (contacting client via phone or text message)	39 (48)
Institutional details	
Organizational structure	
Ownership, n (%)	
Public ownership	57 (70)
Private ownership	24 (30)
Institutional prioritization of VCT services in routine practice, n (%)	
No	27 (33)
Yes	54 (67)
Human resources	
Number of counselors, n (%)	
1	35 (43)
2 or 3	46 (57)
Counselor turnover in the past year, n (%)	
No	62 (77)
Yes	19 (23)
Infrastructure	
Dedicated phlebotomy rooms, n (%)	
No	42 (52)
Yes	39 (48)
Dedicated and private counseling rooms, n (%)	
No	29 (36)
Yes	52 (64)
Location	
Walking time to the nearest subway station (minutes), n (%)	
≤15	43 (53)
>15	38 (47)
Administrative district, n (%)	
Tianhe district	23 (28)
Haizhu district	19 (24)
Baiyun district	19 (24)
Liwan district	2 (2)
Yuexiu district	18 (22)
Service delivery	

XSL•FO
RenderX

Characteristics	Primary health care centers
Turnaround time of HIV test results, n (%)	
Within 1 day	65 (80)
More than 1 day	16 (20)
Availability of VCT services on weekends, n (%)	
No	57 (70)
Yes	24 (30)
Number of times counselors were unable to provide VCT services for personal reasons last year, n (%)	
0	38 (47)
<5	32 (39)
≥5	11 (14)
Posttest gifts for testers, n (%)	
No	3 (4)
Yes	78 (96)
Counselor profiles	
Age (years), n (%)	
≤35	46 (57)
>35	35 (43)
Gender, n (%)	
Male	30 (37)
Female	51 (63)
Educational level: bachelor's degree or higher, n (%)	
No	17 (21)
Yes	64 (79)
Major in public health, n (%)	
No	40 (49)
Yes	41 (51)
Work experience (years), n (%)	
≤3	40 (49)
>3	41 (51)
Proportion of VCT services within the counselor's workload, n (%)	
≤50%	73 (90)
>50%	8 (10)

[a]VCT: voluntary counseling and testing.

Institutional Details and Counselor Profiles of PHCs Offering VCT Services

Table 1 summarizes institutional details and counselor profiles of PHCs. The distribution of PHCs was relatively balanced: Tianhe had the largest share (23/81, 28%), followed by Haizhu (19/81, 24%), Baiyun (19/81, 24%), Yuexiu (18/81, 22%), and Liwan, the smallest (2/81, 2%). Most PHCs were publicly owned (57/81, 70%) and prioritized VCT services in their routine practice (54/81, 67%). Nearly half had only 1 counselor (35/81, 43%), while staff turnover was uncommon (19/81, 23%). About half lacked a dedicated phlebotomy room (42/81, 52%), but most had a private counseling room (52/81, 64%). Over half were within 15 minutes of the nearest subway station (43/81, 53%). Most could return HIV results within a day (65/81, 80%) but did not provide weekend VCT services (57/81, 70%). About half (38/81, 47%) of the PHCs reported no inability to offer VCT services by counselors due to personal reasons last year, and almost all provided posttest gifts for testers (78/81, 96%).

Among 81 VCT counselors, most were female (51/81, 63%) and aged ≤35 years (46/81, 57%). The majority held a bachelor's degree or higher (64/81, 79%), about half had a public health background (41/81, 51%) and >3 years' experience (41/81,

XSL•FO
RenderX

51%), and most spent >50% of their workload on VCT services (73/81, 90%).

Facility-Level Associations Between WellTest Use and HIV Testing Volume

The null model for the HIV testing volume indicated a statistically significant random effect at the cluster level (*P*<.001), suggesting that HIV testing volumes were nested within administrative districts, showing significant interdistrict variation. In the univariate analysis, 3 variables (number of times counselors were unable to provide VCT services for personal reasons last year, counselors' age, and proportion of VCT services within the counselor's workload) were significantly associated with facility-level HIV testing volume (*P*<.10). Results of the sensitivity analysis are consistent with those of the primary analysis. Further details of the primary analysis and sensitivity analysis are presented in Tables S1 and S3 in Multimedia Appendix 1, respectively.

The results of the multivariate analysis for HIV testing volume are presented in Table 2. On the basis of a multilevel negative binomial regression model, PHCs that confirmed appointments had significantly lower HIV testing volumes compared with those that took no actions after clients made appointments (incidence rate ratio 0.75, 95% CI 0.58-0.97; *P*=.03), indicating a 25% reduction in HIV testing volume associated with appointment confirmation. Table 2 also illustrates results of the sensitivity analysis, which are consistent with the primary analysis.

Table 2. Unadjusted and multivariable analysis[a] of the facility-level associations of the use of WellTest with HIV testing volume among primary health care centers (PHCs) offering voluntary counseling and testing services in the urban area of Guangzhou, China, in 2022.

Variables	Facility-level HIV testing volume, median (IQR)	IRR[b] (95% CI)	*P* value	IRR_m[c] (95% CI)	*P* value
Primary analysis					
Proportion of clients scheduling VCT[d] services through WellTest					
<50%	106 (70-142)	Reference	__[e]	Reference	—
≥50%	58 (34-87)	0.83 (0.60-1.15)	.27	0.93 (0.69-1.27)	.66
Set online appointment slots on all VCT service days via WellTest					
No	83 (63-121)	Reference	—	Reference	—
Yes	70 (38-110)	1.00 (0.68-1.46)	.98	1.09 (0.80-1.50)	.58
Actions taken by counselors upon receiving appointments from WellTest					
Take no actions	79 (48-138)	Reference	—	Reference	—
Confirm appointments (contacting clients via phone or SMS text message)	69 (40-96)	0.69 (0.52-0.92)	.01	0.75 (0.58-0.97)	.03
Sensitivity analysis					
Proportion of clients scheduling VCT services through WellTest					
<50%	106 (70-142)	Reference	—	Reference	—
≥50%	61 (35-88)	0.84 (0.61-1.17)	.30	0.94 (0.69-1.28)	.70
Set online appointment slots on all VCT service days via WellTest					
No	83 (63-121)	Reference	—	Reference	—
Yes	70 (46-112)	1.02 (0.70-1.49)	.93	1.11 (0.81-1.52)	.53
Actions taken by counselors upon receiving appointments from WellTest					
Take no actions	83 (48-139)	Reference	—	Reference	—
Confirm appointments (contacting clients via phone or SMS text message)	70 (46-97)	0.69 (0.52-0.92)	.01	0.75 (0.58-0.98)	.03

[a]All models were adjusted by the population under the jurisdiction of each primary health care enter.

[b]IRR: unadjusted incidence rate ratio; $IRR=e^{\beta}$.

[c]IRR_m: multivariate-adjusted incidence rate ratio.

[d]VCT: voluntary counseling and testing.

[e]Not applicable.

XSL•FO
RenderX

Facility-Level Associations Between WellTest Use and HIV Positivity Rate

For HIV positivity rate, the null model revealed that the random effect at the cluster level was not statistically significant (*P*>.99), indicating no district-level clustering. In the univariate analysis, 5 variables (the availability of dedicated phlebotomy rooms, the availability of private and dedicated counseling spaces, administrative district, availability of posttest gifts for testers, and proportion of VCT services within the counselor's workload) were significantly associated with facility-level HIV positivity rate at the significance level of *P*<.10. The sensitivity analysis produced results consistent with the primary analysis. Detailed findings of the primary analysis and sensitivity analysis are presented in Tables S2 and S4 in Multimedia Appendix 1, respectively.

Table 3 presents the results of multivariate analysis for the HIV positivity rate. According to the zero-inflated gamma distribution model adjusted by relative factors, PHCs in which ≥50% of clients scheduled VCT services through WellTest demonstrated higher positivity rates than those with <50% who scheduled through the platform (β=.39, 95% CI 0.02-0.76; *P*=.04). Additionally, PHCs that contacted clients via phone or SMS text message to confirm appointments showed a significantly higher HIV positivity rate compared with those that took no action (β=1.11, 95% CI 0.82-1.40; *P*<.001). Table 3 also presents results of the sensitivity analysis, which are consistent with the primary analysis.

Table 3. Unadjusted and multivariable analysis[a] of the facility-level associations of the use of WellTest with the HIV positivity rate among primary health care centers offering voluntary counseling and testing services in the urban area of Guangzhou, China, 2022.

Variables	Facility-level HIV positivity, median (IQR), %	β[b] (95% CI)	*P* value	β_m[c] (95% CI)	*P* value
Primary analysis					
Proportion of clients scheduling VCT [d] services through WellTest					
<50%	0.0 (0.0-0.0)	Reference	__[e]	Reference	—
≥50%	0.0 (0.0-0.0)	0.07 (−0.78 to 0.93)	.87	0.39 (0.02 to 0.76)	.04
Set online appointment slots on all VCT service days via WellTest					
No	0.0 (0.0-0.0)	Reference	—	Reference	—
Yes	0.0 (0.0-0.0)	−0.42 (−1.48 to 0.63)	.44	−0.42 (−0.85 to 0.01)	.06
Actions taken by counselors upon receiving appointments from WellTest					
Take no actions	0.0 (0.0-0.0)	Reference	—	Reference	—
Confirm appointments (contacting clients via phone or SMS text message)	0.0 (0.0-0.0)	1.07 (0.39 to 1.76)	.003	1.11 (0.82 to 1.40)	<.001
Sensitivity analysis					
Proportion of clients scheduling VCT services through WellTest					
<50%	0.0 (0.0-0.0)	Reference	—	Reference	—
≥50%	0.0 (0.0-0.0)	0.07 (−0.78 to 0.93)	.87	0.39 (0.02 to 0.77)	.04
Set online appointment slots on all VCT service days via WellTest					
No	0.0 (0.0-0.0)	Reference	—	Reference	—
Yes	0.0 (0.0-0.0)	−0.42 (−1.48 to 0.63)	.44	−0.42 (−0.84 to 0.00)	.055
Actions taken by counselors upon receiving appointments from WellTest					
Take no actions	0.0 (0.0-0.0)	Reference	—	Reference	—
Confirm appointments (contacting clients via phone or SMS text message)	0.0 (0.0-0.0)	1.07 (0.39 to 1.76)	.003	1.11 (0.81 to 1.42)	<.001

[a]All models were adjusted by the population under the jurisdiction of each primary health care center.

[b]β: unadjusted β.

[c]β_m: multivariate-adjusted β.

[d]VCT: voluntary counseling and testing.

[e]Not applicable.

XSL•FO
RenderX

Discussion

Principal Findings

Among 81 PHCs offering VCT services in urban Guangzhou in this study, the median proportion of clients using WellTest for appointments was 71% (IQR 34%-98%), while 81% (66/81) of PHCs set online appointment slots on all service days without government mandates. For the client side, facilities with more digital bookings had higher HIV positivity rates. For the provider side, those with active counselors follow-up also had higher positivity rates but lower overall testing volume. These findings indicated strong demand for WellTest as an appointment platform from both providers and clients and suggested that proactive counselor engagement may improve the efficiency of HIV case detection (as reflected by higher positivity rates), even if total testing volume declines.

In the absence of top-down administrative arrangements, WellTest experienced spontaneous and sustained uptake among providers and clients [25]. To our best knowledge, this is the first facility-level report of VCT service digital platform adoption globally. Previous research has shown that digital platforms could effectively facilitate HIV testing at the individual level [15,28], while this study showed the same finding at the facility level. In addition, digital platforms have been found to promote health information dissemination and enhance HIV testing willingness [29-31]. As stigma is considered as a main psychosocial barrier to HIV testing [32], using digital appointment platforms at primary health care facilities might increase accessibility to HIV testing while mitigating stigma associated with walk-in visits. This indicated digital appointment platforms might facilitate access to stigma-sensitive health services.

In terms of appointment management, our findings suggested that the counselor-initiated engagement, such as confirming online appointments via phone calls or SMS text messages, might be associated with a higher HIV positivity rate at the facility level, indicating improved case-detection efficiency despite fewer total tests. Very few studies have reported that active follow-up by health care providers may inadvertently create barriers that influence clients' willingness to undergo HIV testing. Psychosocial barriers such as stigma [32-35], anxiety, and fear of test results [36,37] as major barriers to HIV testing and limited social support [38] deter testing, while high-risk individuals are more motivated to enroll for testing [39]. Although WellTest may play a role in improving the VCT access, the action of confirming appointments may create psychological or logistical hurdles, increasing anxiety and dropouts. However, this step may also filter out less committed or low-risk clients, raising the positivity rate and improving testing efficiency at PHCs at the facility level. Clients seeking testing often experience anxiety, depression, and emotional distress [40]. Counseling and support, as essential components of VCT, have been shown to reduce such psychological burdens [41]and enhance the client experience [42]. As clients often experience anxiety and distress, integrating brief psychological support into the appointment confirmation stage might help enhance retention and engagement. Adding automated reminders in WellTest may also nudge clients without interpersonal pressure, reduce burden, and improve comfort, willingness, and attendance while preserving support and privacy, which is crucial in HIV testing.

Furthermore, the analysis revealed that a higher proportion of clients who scheduled VCT services through digital health platforms had higher HIV case–detection rates at the facility level. This suggested that WellTest may attract higher-risk individuals who were more motivated to seek testing. A prior study has shown that personalized outreach via social media might effectively promote testing among high-risk populations [43]. Embedding peer-to-peer features, for example, privacy-preserving "share with a friend or partner" prompts and referral links, within WellTest might boost uptake and further improve case findings.

Limitations

This study has some limitations. As a cross-sectional study, it cannot establish causality, consistent with similar publications [44-46]. Furthermore, it spanned the entirety of the year 2022 when Guangzhou experienced intermittent COVID-19 pandemic restrictions, which may have affected HIV testing patterns, potentially limiting the generalizability of our findings to nonpandemic conditions. As this study was conducted in the urban area of Guangzhou, a leading metropolitan hub in China, the findings may not be directly applicable to rural or resource-limited regions. Additionally, several potential facility-level and individual-level confounding factors could not be fully addressed in this analysis due to data availability and models' stability. Future studies using longitudinal data that link to more related confounders are needed to better clarify the causal relationship between digital platform use and HIV testing outcomes. Notably, while the WellTest platform effectively engages high-risk populations most in need of VCT services, digital health platforms may encounter digital disparities among older adult populations. Nonetheless, the data provide valuable evidence on facility-level PHC service delivery and digital platform use under real-world conditions and offer suggestions for improving VCT digital platforms.

Conclusions

This study demonstrated the substantial use of WellTest among clients and providers in Guangzhou's urban PHCs and identified its potential role in improving HIV testing services delivery in primary health care settings. To further optimize HIV testing services, the integration of psychological support into appointment confirmation, with automated reminders to reduce client burden and peer-referral features via social networks, could be considered. Such strategies may help address psychosocial barriers, strengthen engagement in HIV testing, and support case finding in primary health care settings.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank the undergraduate student volunteers for their contributions to data collection.

Data Availability

The datasets generated or analyzed during this study, obtained from Guangzhou Centers for Disease Control and Prevention, are not publicly available due to confidentiality and privacy but are available from the corresponding author on reasonable request.

Funding

This study was supported by the National Natural Science Foundation of China (72204059).

Authors' Contributions

Conceptualization: CH (lead), YC (supporting), YFW (supporting)
Data curation: YC (lead), YFW (equal), XRF (supporting)
Formal analysis: YC (lead), YFW (supporting)
Investigation: XRF (lead), JSL (supporting), JW (supporting), ZYL (supporting), CLZ (supporting), MDL (supporting), NJR (supporting), ZZL (supporting), RFZ (supporting), FCX (supporting), YTK (supporting), XRL (supporting)
Methodology: YC (lead), YFW (supporting)
Project administration: CH (lead), YZG (equal), YHL (supporting)
Resources: YZG
Supervision: CH (lead), YZG (supporting)
Validation: YC (lead), YFW (supporting)
Visualization: YC (lead), YFW (supporting)
Writing – original draft: YC (lead), YFW (equal)
Writing – review and editing: YC (lead), YFW (equal), YZG (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Results of the primary and sensitivity analysis for associations between institutional details and counselor profiles and facility-level HIV testing outcomes among primary health care centers offering voluntary counseling and testing services in the urban area of Guangzhou, China, in 2022.
[DOCX File , 66 KB - jmir_v28i1e83662_app1.docx]

References

1. The path that ends AIDS. UNAIDS. URL: https://thepath.unaids.org [accessed 2025-08-03]
2. China statistical yearbook 2023. National Bureau of Statistics of China. URL: https://www.stats.gov.cn/sj/ndsj/2023/indexch.htm [accessed 2026-01-13]
3. The Health Commission of Guangdong Province released the 2018 report on legally reportable infectious diseases in the province. Guangdong Provincial Health Commission. 2019. URL: https://wsjkw.gd.gov.cn/gkmlpt/content/2/2203/post%20_2203048.html#2532 [accessed 2026-01-19]
4. China statistical yearbook 2019. National Bureau of Statistics of China. URL: https://www.stats.gov.cn/sj/ndsj/2019/indexeh.htm [accessed 2026-01-19]
5. The key populations atlas. UNAIDS. URL: https://kpatlas.unaids.org/dashboard [accessed 2025-08-05]
6. Global AIDS strategy 2021-2026 — end inequalities. End AIDS. UNAIDS. 2021 Mar 25. URL: https://www.unaids.org/en/resources/documents/2021/2021-2026-global-AIDS-strategy [accessed 2025-08-05]
7. Bassett IV, Giddy J, Nkera J, Wang B, Losina E, Lu Z, et al. Routine voluntary HIV testing in Durban, South Africa: the experience from an outpatient department. J Acquir Immune Defic Syndr 2007 Oct 01;46(2):181-186 [FREE Full text] [doi: 10.1097/QAI.0b013e31814277c8] [Medline: 17667332]
8. Xu Z, Ma P, Chu M, Chen Y, Miao J, Xia H, et al. Understanding the role of Voluntary Counseling and Testing (VCT) in HIV prevention in Nantong, China. Biomed Res Int 2020 Oct 05;2020:5740654 [FREE Full text] [doi: 10.1155/2020/5740654] [Medline: 33083474]
9. He N. Research progress in the Epidemiology of HIV/AIDS in China. China CDC Wkly 2021 Nov 26;3(48):1022-1030 [FREE Full text] [doi: 10.46234/ccdcw2021.249] [Medline: 34888119]
10. Ong JJ, Peng MH, Wong WW, Lo YR, Kidd MR, Roland M, et al. Opportunities and barriers for providing HIV testing through community health centers in mainland China: a nationwide cross-sectional survey. BMC Infect Dis 2019 Dec 16;19(1):1054 [FREE Full text] [doi: 10.1186/s12879-019-4673-0] [Medline: 31842781]

XSL•FO
RenderX

11. Opinions of the Ministry of Health, Ministry of Finance, and National Population and Family Planning Commission on promoting the gradual equalization of basic public health services. Ministry of Health, Ministry of Finance, and National Population and Family Planning Commission. URL: https://www.gov.cn/gongbao/content/2009/content_1284372.htm [accessed 2025-08-02]
12. Simmons EM, Brown MJ, Sly K, Ma M, Sutton MY, McLellan-Lemal E. Barriers and facilitators to HIV testing in primary care among health care providers. J Natl Med Assoc 2011 May;103(5):432-438. [doi: 10.1016/s0027-9684(15)30340-0] [Medline: 21809793]
13. Wang Y, Li B, Pan J, Sengupta S, Emrick CB, Cohen MS, et al. Factors associated with utilization of a free HIV VCT clinic by female sex workers in Jinan City, Northern China. AIDS Behav 2011 May;15(4):702-710 [FREE Full text] [doi: 10.1007/s10461-010-9703-2] [Medline: 20458528]
14. Nyblade L, Stockton MA, Giger K, Bond V, Ekstrand ML, Lean RM, et al. Stigma in health facilities: why it matters and how we can change it. BMC Med 2019 Feb 15;17(1):25 [FREE Full text] [doi: 10.1186/s12916-019-1256-2] [Medline: 30764806]
15. Biello KB, Mayer KH, Scott H, Valente PK, Hill-Rorie J, Buchbinder S, et al. The effects of MyChoices and LYNX mobile apps on HIV testing and pre-exposure prophylaxis use by young US sexual minority men: results from a national randomized controlled trial. JMIR Public Health Surveill 2025 Feb 05;11:e63428 [FREE Full text] [doi: 10.2196/63428] [Medline: 39908084]
16. Saber R, Mongrella M, Matson M, Carty-Fickes E, Novotny J, Pirog S, et al. Use of a custom testing center locator tool to improve STI and HIV testing rates in adolescent men who have sex with men as part of an online sexual health program. Procedia Comput Sci 2022;206:92-100 [FREE Full text] [doi: 10.1016/j.procs.2022.09.088] [Medline: 37228693]
17. Liu Z, Huang H, Yue T, Gebregziabher NK, Gong H, Xu P, et al. Standardly trained peer volunteer-led, social app recruiting-based human immunodeficiency virus testing strategy using rapid testing kits (SPARK) among men who have sex with men: a population-based survey evaluation. Open Forum Infect Dis 2024 Dec 02;11(12):ofae709. [doi: 10.1093/ofid/ofae709] [Medline: 39679350]
18. Tang W, Wei C, Cao B, Wu D, Li KT, Lu H, et al. Crowdsourcing to expand HIV testing among men who have sex with men in China: a closed cohort stepped wedge cluster randomized controlled trial. PLoS Med 2018 Aug 28;15(8):e1002645 [FREE Full text] [doi: 10.1371/journal.pmed.1002645] [Medline: 30153265]
19. Adeagbo O, Herbst C, Blandford A, McKendry R, Estcourt C, Seeley J, et al. Exploring people's candidacy for mobile health-supported HIV testing and care services in rural KwaZulu-Natal, South Africa: qualitative study. J Med Internet Res 2019 Nov 18;21(11):e15681 [FREE Full text] [doi: 10.2196/15681] [Medline: 31738174]
20. Mwingira U, Chikawe M, Mandara WL, Mableson HE, Uisso C, Mremi I, et al. Lymphatic filariasis patient identification in a large urban area of Tanzania: an application of a community-led mHealth system. PLoS Negl Trop Dis 2017 Jul 14;11(7):e0005748 [FREE Full text] [doi: 10.1371/journal.pntd.0005748] [Medline: 28708825]
21. Yew TW, Chi C, Chan SY, van Dam RM, Whitton C, Lim CS, et al. A randomized controlled trial to evaluate the effects of a smartphone application-based lifestyle coaching program on gestational weight gain, glycemic control, and maternal and neonatal outcomes in women with gestational diabetes mellitus: the SMART-GDM study. Diabetes Care 2021 Feb;44(2):456-463 [FREE Full text] [doi: 10.2337/dc20-1216] [Medline: 33184151]
22. Miller DPJ, Denizard-Thompson N, Weaver KE, Case LD, Troyer JL, Spangler JG, et al. Effect of a digital health intervention on receipt of colorectal cancer screening in vulnerable patients: a randomized controlled trial. Ann Intern Med 2018 Apr 17;168(8):550-557 [FREE Full text] [doi: 10.7326/M17-2315] [Medline: 29532054]
23. Guangzhou statistical yearbook 2024. Guangzhou Statistics Bureau. URL: https://tjj.gz.gov.cn/datav/admin/home/www_nj/ [accessed 2026-01-13]
24. Cheng W, Cai Y, Tang W, Zhong F, Meng G, Gu J, et al. Providing HIV-related services in China for men who have sex with men. Bull World Health Organ 2016 Jan 22;94(3):222-227. [doi: 10.2471/blt.15.156406]
25. Wang J, Gu YZ, Lu YH, Li JS, Luo YF, Cai YS, et al. Implementation of digital HIV services in Guangzhou's primary health-care system, China. Bull World Health Organ 2025 Feb 01;103(2):99-109. [doi: 10.2471/BLT.24.291921] [Medline: 39882491]
26. Gschlößl S, Czado C. Modelling count data with overdispersion and spatial effects. Statistical Papers 2006 Nov 17;49:531-552. [doi: 10.1007/s00362-006-0031-6]
27. Nobre AA, Carvalho MS, Griep RH, Fonseca MJ, Melo EC, Santos IS, et al. Multinomial model and zero-inflated gamma model to study time spent on leisure time physical activity: an example of ELSA-Brasil. Rev Saude Publica 2017 Aug 17;51:76 [FREE Full text] [doi: 10.11606/S1518-8787.2017051006882] [Medline: 28832756]
28. Phanuphak N, Anand T, Jantarapakde J, Nitpolprasert C, Himmad K, Sungsing T, et al. What would you choose: online or offline or mixed services? Feasibility of online HIV counselling and testing among Thai men who have sex with men and transgender women and factors associated with service uptake. J Int AIDS Soc 2018 Jul;21 Suppl 5(Suppl Suppl 5):e25118 [FREE Full text] [doi: 10.1002/jia2.25118] [Medline: 30033644]
29. Govender K, Beckett S, Masebo W, Braga C, Zambezi P, Manhique M, et al. Effects of a short message service (SMS) intervention on reduction of HIV risk behaviours and improving HIV testing rates among populations located near roadside

XSL•FO
RenderX

wellness clinics: a cluster randomised controlled trial in South Africa, Zimbabwe and Mozambique. AIDS Behav 2019 Nov;23(11):3119-3128. [doi: 10.1007/s10461-019-02427-6] [Medline: 30771133]
30. Schnall R, Kuhns L, Pearson C, Bruce J, Batey DS, Radix A, et al. Preliminary results from a pragmatic clinical trial of MyPEEPS mobile to improve HIV prevention behaviors in young men. Stud Health Technol Inform 2020 Jun 16;270:1365-1366 [FREE Full text] [doi: 10.3233/SHTI200444] [Medline: 32570661]
31. Blas MM, Alva IE, Carcamo CP, Cabello R, Goodreau SM, Kimball AM, et al. Effect of an online video-based intervention to increase HIV testing in men who have sex with men in Peru. PLoS One 2010 May 03;5(5):e10448 [FREE Full text] [doi: 10.1371/journal.pone.0010448] [Medline: 20454667]
32. Andrinopoulos K, Kerrigan D, Figueroa JP, Reese R, Ellen JM. HIV coping self-efficacy: a key to understanding stigma and HIV test acceptance among incarcerated men in Jamaica. AIDS Care 2010 Mar;22(3):339-347 [FREE Full text] [doi: 10.1080/09540120903193633] [Medline: 20390514]
33. Berendes S, Rimal RN. Addressing the slow uptake of HIV testing in Malawi: the role of stigma, self-efficacy, and knowledge in the Malawi BRIDGE Project. J Assoc Nurses AIDS Care 2011;22(3):215-228. [doi: 10.1016/j.jana.2010.08.005] [Medline: 21185751]
34. Hendriksen ES, Hlubinka D, Chariyalertsak S, Chingono A, Gray G, Mbwambo J, et al. Keep talking about it: HIV/AIDS-related communication and prior HIV testing in Tanzania, Zimbabwe, South Africa, and Thailand. AIDS Behav 2009 Dec;13(6):1213-1221 [FREE Full text] [doi: 10.1007/s10461-009-9608-0] [Medline: 19760154]
35. Kalichman SC, Simbayi LC. HIV testing attitudes, AIDS stigma, and voluntary HIV counselling and testing in a Black township in Cape Town, South Africa. Sex Transm Infect 2003 Dec;79(6):442-447 [FREE Full text] [doi: 10.1136/sti.79.6.442] [Medline: 14663117]
36. Flowers P, Knussen C, Li J, McDaid L. Has testing been normalized? An analysis of changes in barriers to HIV testing among men who have sex with men between 2000 and 2010 in Scotland, UK. HIV Med 2013 Feb;14(2):92-98 [FREE Full text] [doi: 10.1111/j.1468-1293.2012.01041.x] [Medline: 22934820]
37. Menser ME. Perceived risk, decisional balance, and HIV testing practices in college students [thesis]. University of Pittsburgh. 2010. URL: https://d-scholarship.pitt.edu/concern/etds/87659344-fe6d-4a39-8b7d-10d24743cadd [accessed 2025-08-05]
38. Conserve DF, Bay C, Kilonzo MN, Makyao NE, Kajula L, Maman S. Sexual and social network correlates of willingness to self-test for HIV among ever-tested and never-tested men: implications for the Tanzania STEP project. AIDS Care 2019 Feb;31(2):169-176 [FREE Full text] [doi: 10.1080/09540121.2018.1537466] [Medline: 30362377]
39. Johnson C, Neuman M, MacPherson P, Choko A, Quinn C, Wong VJ, et al. Use and awareness of and willingness to self-test for HIV: an analysis of cross-sectional population-based surveys in Malawi and Zimbabwe. BMC Public Health 2020 May 25;20(1):779 [FREE Full text] [doi: 10.1186/s12889-020-08855-7] [Medline: 32450840]
40. Kagee A, Saal W, Bantjes J. Distress, depression and anxiety among persons seeking HIV testing. AIDS Care 2017 Mar;29(3):280-284 [FREE Full text] [doi: 10.1080/09540121.2016.1259453] [Medline: 27866410]
41. Gu J, Lau JT, Wang Z, Wu AM, Tan X. Perceived empathy of service providers mediates the association between perceived discrimination and behavioral intention to take up HIV antibody testing again among men who have sex with men. PLoS One 2015 Feb 18;10(2):e0117376 [FREE Full text] [doi: 10.1371/journal.pone.0117376] [Medline: 25693179]
42. Dakpui HD, Shamrock OW, Aidoo-Frimpong G, Zigah EY, Kofi Agbemedu GR, Ahmed A, et al. A qualitative description of HIV testing and healthcare experiences among trans women in Ghanaian urban slums BSGH-012. PLOS Glob Public Health 2024 Dec 2;4(12):e0003980. [doi: 10.1371/journal.pgph.0003980] [Medline: 39621619]
43. Chiou PY, Tsao WW, Li CL, Yu JM, Su WH, Liu ZH, et al. Recruitment for voluntary video and mobile HIV testing on social media platforms during the COVID-19 pandemic: cross-sectional study. J Med Internet Res 2024 Nov 28;26:e54420 [FREE Full text] [doi: 10.2196/54420] [Medline: 39607762]
44. Adugna DG, Worku MG. HIV testing and associated factors among men (15-64 years) in Eastern Africa: a multilevel analysis using the recent demographic and health survey. BMC Public Health 2022 Nov 24;22(1):2170 [FREE Full text] [doi: 10.1186/s12889-022-14588-6] [Medline: 36434555]
45. Rakhmanina N, Messenger N, Phillips G2, Teach S, Morrison S, Hern J, et al. Factors affecting acceptance of routine human immunodeficiency virus screening by adolescents in pediatric emergency departments. J Adolesc Health 2014 Feb;54(2):176-182. [doi: 10.1016/j.jadohealth.2013.07.027] [Medline: 24060572]
46. Cisse K, Ouedraogo HG, Ky-Zerbo O, Kambire D, Zida S, Ki-Toe C, et al. HIV testing among female sex workers and associated factors in Burkina Faso: findings from a respondent-driven sampling survey. BMC Public Health 2024 Sep 16;24(1):2512 [FREE Full text] [doi: 10.1186/s12889-024-19977-7] [Medline: 39285273]

Abbreviations

CDC: Centers for Disease Control and Prevention
PHC: primary health care center
VCT: voluntary counseling and testing

XSL•FO
RenderX

Edited by A Mavragani; submitted 09.Sep.2025; peer-reviewed by K Jiao; comments to author 13.Jan.2026; accepted 25.Mar.2026; published 17.Apr.2026.

Please cite as:
Chen Y, Wang YF, Gu YZ, Fan XR, Lu YH, Li JS, Wang J, Lin ZY, Zhang CL, Li MD, Ren NJ, Luo ZZ, Zuo RF, Xiong FC, Kang YT, Lin XR, Hao C
Facility-Level Associations Between Use of a Digital Health Platform for Voluntary Counseling and Testing and HIV Testing Outcomes in the Urban Primary Health Care Centers of Guangzhou, China: Cross-Sectional Study
J Med Internet Res 2026;28:e83662
URL: https://www.jmir.org/2026/1/e83662
doi:10.2196/83662
PMID:36810139

©Ye Chen, Yu-Fei Wang, Yu-Zhou Gu, Xiao-Ru Fan, Yong-Heng Lu, Ju-Shuang Li, Jun Wang, Zhi-Ye Lin, Chun-Li Zhang, Meng-Die Li, Ning-Jun Ren, Zhi-Zhong Luo, Rui-Fei Zuo, Fu-Chuan Xiong, Yu-Tong Kang, Xiu-Rong Lin, Chun Hao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Usage and Exposure to Content of the NHS Healthy Living Program for People With Type 2 Diabetes: Retrospective Observational Cohort Study

Antonia M Marsden[1], PhD; Rhiannon E Hawkes[2], PhD; Jack S Benton[2], PhD; David P French[2], PhD; Sarah Cotterill[1], PhD

[1]Centre for Biostatistics, School of Health Sciences, University of Manchester, Manchester, England, United Kingdom

[2]Manchester Centre for Health Psychology, School of Health Sciences, University of Manchester, Manchester, England, United Kingdom

Corresponding Author:
Rhiannon E Hawkes, PhD
Manchester Centre for Health Psychology
School of Health Sciences
University of Manchester
Coupland Building 1
Oxford Road
Manchester, England, M13 9PL
United Kingdom
Phone: 44 0161 275 2584
Email: rhiannon.hawkes@manchester.ac.uk

Abstract

Background: Diabetes self-management education and support (DSMES) programs can improve health outcomes, but engagement is often low. “Healthy Living” is a web-based self-management program for people with type 2 diabetes, based on the “HeLP-Diabetes” intervention, which demonstrated effectiveness in a randomized controlled trial. Healthy Living was commissioned by the National Health Service in England and rolled out nationally into routine care in 2020. The program comprises web-based structured learning, unstructured articles (which users could access at any time), and tracking tools such as goal setting. It is important to assess not only the uptake of digital interventions but also the amount of time spent using the intervention and the content they engage with. There is currently limited research on the extent to which people engage with digital DSMES program content outside of a trial setting.

Objective: This study aimed to investigate the overall usage and exposure to content of Healthy Living, including differences in usage/exposure by user characteristics.

Methods: This study used a retrospective observational cohort study design to conduct a service evaluation of Healthy Living. Pseudonymous usage data from all people with type 2 diabetes (N=27,422) who activated a Healthy Living account between May 2020 and September 2023 were available, including (1) which program activities were accessed, (2) when activities were accessed, and (3) how long users spent on each activity. User demographic and usage information was summarized using means and SDs, medians and IQRs, or frequencies and percentages. Logistic regression evaluated the association between user demographics and usage.

Results: Of the 27,422 users who activated a Healthy Living account, the mean age was 58.6 (SD 12.0) years, and the percentage of females was 56.8% (15,385/27,422). There was an even spread across deprivation quintiles. The median length of time spent on the program in total was 7.6 (IQR 0.6-27.6) minutes; 12,066 (44%) users spent <5 minutes on the program, and 3022 (11%) spent ≥1 hour. Of those who activated an account, 69.8% (19,137/27,422) accessed some program content, 40.7% (11,149/27,422) completed the first section of structured education, and 4.7% (1219/27,422) completed 60% of the structured education. Usage of the unstructured aspects of the program was low. Female gender, lower deprivation, White ethnicity, and a shorter time since diagnosis were associated with increased usage.

Conclusions: This study is one of the first to provide detailed analysis of user engagement with a national digital self-management program for type 2 diabetes. Unlike previous evaluations conducted on smaller sample sizes, this study highlights the low usage of Healthy Living across >27,000 users. Given that higher usage of DSMES programs is associated with improved health outcomes in people with type 2 diabetes, further work needs to identify how to encourage increased engagement with the program.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e89690) doi:10.2196/89690

KEYWORDS

digital interventions; engagement; self-management; type 2 diabetes; usage data

Introduction

Overview

Type 2 diabetes is a chronic condition that requires long-term management [1]. Diabetes self-management education and support (DSMES) aims to inform people about their condition, for example, through providing guidance on diet, physical activity, blood glucose control, and medication, to enable people to manage their diabetes on an ongoing basis, thus reducing their risk of further health and physiological problems [2,3]. In the United Kingdom, the National Institute for Health and Care Excellence (NICE) guidelines outline that DSMES should be offered to all adults newly diagnosed with type 2 diabetes [1]. There is evidence that such programs can result in reduced hemoglobin A_{1c} (HbA_{1c}; blood glucose) levels, reduce the risk of complications, increase quality of life, and foster positive behavioral changes relating to exercise and nutrition [4,5].

Despite the potential benefits, attendance for DSMES can be low. Data from the 2022 National Diabetes Audit in the United Kingdom estimates that approximately 76.3% of people newly diagnosed with type 2 diabetes were offered structured education within 12 months of their diagnosis, but only 8.6% attended within the first 12 months [6]. Such interventions are typically delivered as face-to-face group-based courses [7,8]. However, digital interventions, delivered via websites or apps, are potentially more appealing to some, including (1) those who work and/or have caring responsibilities [9], (2) those with severe mental illness or intellectual disability, and (3) ethnic minority groups [10]. To widen access to support for people with type 2 diabetes in England, the National Health Service (NHS) 10-Year Plan has committed to shifting care from analog to digital, by expanding access to digital self-management tools and DSMES programs [11]. All demographic groups should be involved in developing digital services, and tailoring can be helpful to minimize digital inequality [10]. Systematic review evidence suggests that digital DSMES can be effective in improving HbA_{1c} in patients with type 2 diabetes [12].

HeLP-Diabetes was an online DSMES designed to provide information about type 2 diabetes and its treatments, and support people to make and retain lifestyle changes in areas such as physical activity and diet [13]. A randomized controlled trial (RCT) comparing the HeLP-Diabetes intervention to a simple information website demonstrated modest but statistically significant reductions in HbA_{1c} at 12 months for those allocated to the intervention (mean difference between groups at 12 months: –2.6, 95% CI –4.8 to –0.5 mmol/mol; P=.01) [14]. HeLP-Diabetes was shown to be cost-effective [15].

A national rollout of HeLP-Diabetes in routine care was commissioned by the NHS in England in 2019 [16]. The program, named “Healthy living for people with type 2 diabetes,” or “Healthy Living” for short, was modified and delivered by an external provider for the NHS, but showed high fidelity of content to the version evaluated in the trial [17]. The program underwent beta testing between May 2020 and June 2021, followed by the full-scale rollout. Healthy Living is one of several DSMES programs available in the United Kingdom, alongside structured group options such as DESMOND and X-PERT, which offer face-to-face group education and self-management support delivered by trained educators in local communities, often complemented by a digital app. People with type 2 diabetes can be referred to any of these services to help them manage their condition, although access depends on what is commissioned locally. Healthy Living was designed not to replace existing provision, but to provide a universally available, free, evidence-based online alternative that expands patient choice beyond what is offered in their local area.

It is important to assess not only the uptake of digital interventions, but also the engagement once participants have enrolled, for example, the amount of time spent using the intervention and the content they engage with [18]. This is to (1) examine why there may be differences in outcomes in routine delivery compared to an RCT context, (2) inform whether there is a need to improve engagement of particular elements of the intervention, and (3) examine whether less usage in disadvantaged groups exacerbates inequalities. For example, if users do not engage with the intervention, they will not be exposed to important intervention content to help with self-management of their condition, which could result in reduced health outcomes. Further, while there is evidence that low socioeconomic status groups are significantly less likely to start using chronic disease self-management support interventions, the literature is unclear on whether low socioeconomic status is related to retention and usage, with only a few small studies [19]. Analysis of program usage is needed, including understanding how it varies among the population, so programs can be tailored to underrepresented groups [10] and avoid exacerbating health inequalities.

Previous studies investigating engagement with DSMES programs for type 2 diabetes have shown mixed results. In the HeLP-Diabetes RCT, the mean number of log-ins for users of the intervention was 18.7 (SD 84.0), and the mean number of days in which the website was accessed over the 12-month follow-up period was 10.1 (SD 22.9) [14]. In an RCT of the BetaMe online program in New Zealand, initial engagement was high, with 74% of participants having active engagement within the first 16 weeks, but this declined over time [20]. However, levels of usage in RCT settings may be an overestimation of that in real-world settings; a systematic review found that the real-world usage rate of unguided e-mental health programs was, on average, 4 times lower than that reported in the RCTs of the same programs [21]. There has been less research on engagement with DSMES based on real-world data, but one study is the observational analysis of usage of a UK-based DSMES program for type 2 diabetes, myDESMOND, where 5360 (56.3%) of 9522 users remained on the program

XSL•FO
RenderX

for at least 1 month, and 1676 (17.6%) remained on for 1 year [22].

Objectives

This study was able to investigate engagement with a nationally implemented DSMES program in a sample of routine users of Healthy Living. The aim of this work was to summarize usage of the Healthy Living program, specifically investigating (1) overall usage (such as how often participants accessed the self-management program and how long they spent on the program), (2) what program content was accessed by participants, and (3) how usage varied across demographics.

Methods

Design

This study was part of a wider service evaluation of Healthy Living, in which all data from users of the Healthy Living program were analyzed. Thus, a retrospective observational cohort design was used, using pseudonymized participant-level usage data collected by the Healthy Living service provider linked to the National Diabetes Audit [6].

Healthy Living Program

The Healthy Living program is a web-based self-management program, accessible on computers, smartphones, and tablets, with 3 components: the "Learn Journey," "Find Answers," and "Tools." The Learn Journey, the main component, contained structured information comprising 1-page articles containing a mixture of text, pictures, videos, and quizzes, grouped into sections. At the time of this evaluation, there were >30 sections of the Learn Journey which users were required to work through in a defined order. The majority of the Learn Journey was the same for all users, but there was additional information on smoking, driving, and working for those who indicated this was relevant. NHS England defines "attendance" on the program as reaching the end of the first section in the Learn Journey ("Introduction to type 2 diabetes"), and "completion" of the program as accessing 60% of the Learn Journey (see Textbox 1 for definitions).

The Find Answers component was unstructured and comprised articles, grouped in sections, that users can access at any time. The "Tools" section comprised 5 tracking tools users could access at any time: goal setting, blood test tracking, food tracking, weight tracking, and step tracking. See Multimedia Appendix 1 for screenshots of the 3 main components of the Healthy Living intervention at the time of the evaluation.

Textbox 1. Definitions of Healthy Living program components.

Learn Journey

- Structured information comprising 1-page articles containing a mixture of text, pictures, videos, and quizzes, grouped into sections.

Find Answers

- Unstructured information comprising articles, grouped into sections, that users could access at any time.

Tools

- Section comprises 5 tracking tools users could access at any time: goal setting, blood test tracking, food tracking, weight tracking, and step tracking.

Session

- A string of continuous activity on the program, ending when the user was inactive for over 10 minutes.

Attendance

- Reaching the end of the first section in the Learn Journey.

Completion

- Accessing at least 60% of the Learn Journey.

Participants

Participants were referred to Healthy Living through their general practice or via self-referral. To register, users visited the Healthy Living website and provided their name, type 2 diabetes status (ie, whether they were a patient, caregiver, or health care professional), email address, gender, date of birth, postcode, and how they heard about Healthy Living. Following registration, users were sent an email with a link to complete their profile and activate their account, after which they could access program content.

Setting and Study Size

This study includes data from all users who activated an account between May 11, 2020, and September 1, 2023, and had a type 2 diabetes diagnosis. In cases where individuals had multiple active accounts, we selected the account from which the user had accessed the highest number of educational articles, and if the number was equal, we selected the latest account. Users who activated an account less than 1 month before the date of data extraction (September 1, 2023) were excluded, as they may not have had sufficient time to work through the materials. The

XSL•FO
RenderX

suitability of this 1-month time period was investigated in a sensitivity analysis.

Variables and Data Measurement

Personal Characteristics

In addition to the user information provided by participants at registration, information on ethnicity and time since diagnosis of type 2 diabetes was obtained from the National Diabetes Audit [6]; users were linked using a pseudonymized NHS number. However, this information was unavailable for those with a missing pseudonymized NHS number (2237/27,422, 8.2% of users). Other information, such as first language, disability status, and smoking status, was collected from users once they had activated their account, but this has not been included due to high levels of missingness (>50%). Postcode entered at registration was used to determine the 2019 Index of Multiple Deprivation quintile [23].

Program Usage

Data on activities undertaken on the Healthy Living program by all users was collected: when users logged in, what articles in the Learn Journey and Find Answers sections were accessed, what tools were accessed, when the articles and tools were accessed, and the amount of time spent on each activity. This enabled us to summarize across all users.

Summaries of Overall Usage

The following measures were used to summarize overall program usage:

1. The amount of time spent on the program at each session and the total amount of time spent on the program.
2. The number of program "sessions" attended, where a session is defined as a string of continuous activity on the program, ending when the user was inactive for over 10 minutes.
3. The number of users who accessed program content.
4. The number of users who "attended" the program, where attendance is defined as reaching the end of the first section in the Learn Journey.
5. The number of users who "completed" the program, where completion is defined as accessing at least 60% of the Learn Journey (ie, reaching the end of section 15).
6. When sessions took place.

Summaries of Program Content Accessed

The following measures were used to summarize access to program content:

1. The number of sections in the Learn Journey accessed.
2. The number of articles in the Find Answers section accessed.
3. The most accessed articles in the Find Answers section.
4. How often the tools were used.

Bias

Note that for the amount of time spent on the program at each session and the total amount of time spent on the program, because of the way in which a session was defined, the duration of the last activity of each session was unknown. These missing durations were set to the median duration of activities of the same type (eg, Learn Journey article, Find Answers article, and step tracker access) undertaken by the user, if available. If this median could not be calculated because the user had no other observations of the same type, the missing duration was set to the median duration of activities of the same type undertaken by all other users.

Statistical Methods

User characteristics and usage information were summarized using means and SDs, medians and IQRs, or frequencies and percentages, as appropriate. User characteristics were also summarized separately in those who did and did not "attend" the program and those who did and did not "complete" the program. Logistic regression was used to estimate the association between each usage characteristic with both "attendance" and "completion," while adjusting for the other characteristics. Analyses were performed in Stata (version 17; Stata Corp LLC).

Ethical Considerations

This study was reviewed and approved by the Yorkshire and the Humber-Leeds West NHS Research Ethics Committee (reference number 20/YH/0250). All users of Healthy Living had provided consent for their pseudonymized data to be accessed for service evaluations, and agreed to the terms and conditions of the privacy policy before they accessed the program. Data presented in this manuscript are pseudonymized. As this was a service evaluation, users were not compensated as they did not actively participate in any research, and instead, only their pseudonymized data were obtained. No identification of individual participants/users in this manuscript or supplementary material is possible.

Results

Participant Characteristics

Figure 1 describes participant selection. Data from 42,689 users who registered on the program were obtained, and 27,422 of these were determined to have reported a type 2 diabetes diagnosis at registration and activated a Healthy Living account more than 1 month before data extraction, that is, before September 1, 2023. The sensitivity analysis suggested that increasing this period to 3, 6, and 9 months made little difference to the findings (Multimedia Appendix 2). Note that, while the pseudonymized NHS number was used to identify accounts belonging to the same person, these data were missing for 8.2% (2237/27,422) of eligible participants who activated an account, so some duplicates may have been missed.

XSL•FO
RenderX

Figure 1. Flow diagram of participant selection for inclusion in analysis.

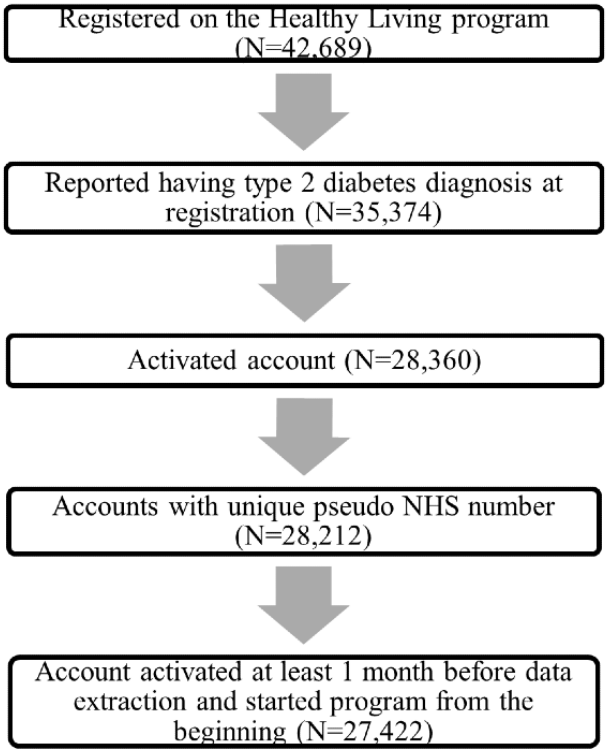

Table 1 summarizes characteristics of the 27,422 users. The mean age was 58.6 (SD 12.0) years, and the percentage of females was 56.8% (15,385/27,088). There was an even spread across deprivation quintiles. The majority of users were made aware of Healthy Living through a health care professional from their general practitioner practice (12,178/27,422, 44.4%), social media (5053/27,422, 18.4%), or an online search (4758/27,422, 17.4%). Users self-referred in nearly all cases, 15.6% (4282/27,422) doing so through a private link before the program was publicly available, and 81.9% (22,467/27,422) doing so once the program was public. Regarding ethnicity, 84.9% (17,628/20,755) of those with a nonmissing entry were White, 8.8% (1835/20,755) were Asian, 4.2% (862/20,755) were Black, 1.1% (228/20,755) were mixed, and 1% (202/20,755) were classed as “other.” The median number of years since diagnosis was 3 (IQR 0-10) years.

XSL•FO
RenderX

Table 1. Descriptive characteristics of users with type 2 diabetes who activated a Healthy Living account between May 2020 and September 2023 (N=27,422).

Characteristics	Values
Age (years)	
Mean (SD)	58.6 (12.0)
Median (IQR)	59 (51-67)
Missing, n (%)	139 (0.5)
Gender, n (%)	
Male	11,644 (43)
Female	15,385 (56.8)
Intersex	59 (0.2)
Missing	334 (1.2)
IMD[a] quintile, n (%)	
1 (most deprived)	4920 (18.9)
2	5029 (19.3)
3	5403 (20.7)
4	5601 (21.5)
5 (least deprived)	5129 (19.7)
Missing	1340 (4.9)
Ethnicity, n (%)	
Asian	1835 (8.8)
Black	862 (4.2)
Mixed	228 (1.1)
Other	202 (1)
White	17,628 (84.9)
Missing	6667 (24.3)
Years since diagnosis	
Mean (SD)	6.3 (6.9)
Median (IQR)	4 (0-10)
Missing, n (%)	5961 (21.7)
How the user was made aware of Healthy Living, n (%)	
Charity or community group-Diabetes UK	700 (2.6)
Charity or community group-other	76 (0.3)
Friend or family member	504 (1.8)
Health care professional-GP[b] practice	12,178 (44.4)
Health care professional-other	680 (2.5)
Online search	4758 (17.4)
Poster, leaflet, or other NHS[c] promotion	1381 (5)
Social media	5053 (18.4)
None of the above	2092 (7.6)
Missing	0
Referral route, n (%)	
Public beta-Hub referral (NWL[d])[e]	323 (1.2)

XSL•FO
RenderX

Characteristics	Values
Self-referral landing page (private)	4282 (15.6)
Self-referral landing page (public)	22,467 (81.9)
GP referral	350 (1.3)
Missing	0

[a]IMD: Index of Multiple Deprivation. IMD scores associated with the lower super output area derived from venue postcodes, ranging from the most deprived areas in England to the least deprived areas in England.

[b]GP: general practitioner.

[c]NHS: National Health Service.

[d]NWL: North West London.

[e]The option "Public beta-Hub referral (NWL)" refers to referrals via an online learning platform targeted at primary care staff in North West London.

Summary of Overall Usage

Table 2 summarizes overall usage of the program. The median length of time spent on the program by a total of 27,422 users who activated their accounts was 7.6 (IQR 0.6-27.6) minutes; 12,066 (44%) of users spent less than 5 minutes on the program, and 3022 (11%) spent 1 hour or more on the program. Over half completed only 1 session (15,089/27,422, 55%), and 3171 (11.6%) completed 5 or more sessions. Approximately 70% (n=19,137) accessed some program content at least once, 40.7% (n=11,149) "attended" the program, and 4.7% (n=1291) "completed" the program (see Textbox 1 for definitions).

Table 2. Summary of overall usage of the Healthy Living program between May 2020 and September 2023 (N=27,422).

	Values
Total time spent on the program in minutes	
<1 minute, n (%)	7598 (27.7)
1-5 minutes, n (%)	4468 (16.3)
5-30 minutes, n (%)	8971 (32.7)
30-60 minutes, n (%)	3363 (12.3)
60-120 minutes, n (%)	2025 (7.4)
>120 minutes, n (%)	997 (3.6)
Median (IQR)	7.6 (0.6-27.6)
Mean (SD)[a]	24.5 (58.4)
Total number of sessions[b]	
1, n (%)	15,089 (55)
2-4, n (%)	9162 (33.4)
5-9, n (%)	2183 (8)
>10, n (%)	988 (3.6)
Median (IQR)	1 (1-3)
Mean (SD)[a]	3.1 (19.9)
Accessed some program content (Learn Journey, Find Answers, or Tools), n (%)	19,137 (69.8)
"Attended" the program[c], n (%)	11,149 (40.7)
"Completed" the program[d], n (%)	1219 (4.7)

[a]The mean is presented to aid comparison with similar programs. The median is our preferred summary measure due to the skewed distribution of these variables.

[b]A session is defined as a string of continuous activity on the program, ending when the user was inactive for over 10 minutes.

[c]Attendance is defined as reaching the end of the first section in the Learn Journey.

[d]Completion is defined as accessing at least 60% of the Learn Journey.

In total, 83,731 sessions were recorded over 60,830 person-days. Table 3 summarizes the time of day and day of week that sessions took place. The most popular time of day for sessions was 12 PM-6 PM (3887/83,731, 35.3% of sessions), although

XSL•FO
RenderX

the number of sessions starting between 6 AM-12 PM and 6 PM-12 AM was only slightly lower (approximately 30% in each). Weekends were slightly less popular than weekdays. Of the 83,731 sessions, 32,032 (38.3%) started between 9 AM and 5 PM on Monday-Friday.

Table 3. Summary of time of day and day of week that users completed sessions[a] on the Healthy Living program between May 2020 and September 2023.

	Values, n (%)
Time of day (n=83,731)	
12 AM-6 AM	3887 (4.6)
6 AM-12 PM	27,017 (32.3)
12 PM-6 PM	29,539 (35.3)
6 PM-12 AM	23,288 (27.8)
Day of the week (n=60,830)	
Monday	10,017 (16.5)
Tuesday	9805 (16.1)
Wednesday	9625 (15.8)
Thursday	9560 (15.7)
Friday	9300 (15.3)
Saturday	6301 (10.4)
Sunday	6222 (10.2)

[a]A session is defined as a string of continuous activity on the program, ending when the user is inactive for over 10 minutes.

Summary of Program Content Accessed

Table 4 summarizes the number of sections in the Learn Journey accessed and the number of articles in the Find Answers section accessed by the 19,137 users who accessed program content.

The median number of Learn Journey sections accessed was 2 (IQR 1-4). A total of 10,273 (53.7%) users accessed 1-3 sections, and 1090 (5.1%) accessed 20 or more sections. Table S4 in Multimedia Appendix 3 lists each section of the Learn Journey and the number of people who accessed it. Seven articles in the Learn Journey had a median duration of engagement of 90 seconds or more; 6 of these included interactive quizzes, and 1 contained an embedded video. These are listed in Multimedia Appendix 4.

Of the 19,137 users, 5825 (30.4%) accessed the Find Answers section at least once, and the median number of articles accessed was 0 (IQR 0-1); 1453/19,137 (7.6%) accessed only 1 article in total, and 1814/19,137 (9.5%) accessed over 10 articles. The most accessed articles were in the “Eating well for diabetes” section; this is the second section that appears in this part of the program, after the section “About type 2 diabetes.” The 5 most popular articles were: “Meals and portion sizes” (accessed by n=1148, 6%), “Foods to eat and avoid” (n=1091, 5.7%), “Healthy portion plate” (n=844, 4.4%), “Fruit and vegetables” (n=839, 4.4%), and “Lower carbohydrate plate” (n=832, 4.4%). All articles that were accessed by more than 600 users are listed in Multimedia Appendix 5 (Table S5).

Table 5 shows how many users accessed each of the tools in the Tools section. Of the 19,137 users, 3066 (16%) visited at least one of the tool sections once or more, but there was little evidence of regular use of the tracking tools.

XSL•FO
RenderX

Table 4. Summary of the number of sections in the Learn Journey and the number of articles in the Find Answers section that were accessed by Healthy Living users between May 2020 and September 2023 (n=19,137).

	Values
Number of sections of the Learn Journey accessed	
0 sections, n (%)	1483 (7.8)
1 section, n (%)	6829 (35.7)
2-3 sections, n (%)	3444 (18)
4-5 sections, n (%)	4246 (22.2)
6-9 sections, n (%)	1233 (6.4)
10-19 sections, n (%)	812 (4.2)
20-29 sections, n (%)	1090 (5.7)
Median (IQR)	2 (1-4)
Mean (SD)[a]	4.1 (5.7)
Number of articles in the Find Answers section accessed	
0 article, n (%)	13,312 (69.6)
1 article, n (%)	1453 (7.59)
2-4 articles, n (%)	1577 (8.2)
5-9 articles, n (%)	981 (5.1)
10-19 articles, n (%)	847 (4.4)
20+ articles, n (%)	967 (5.1)
Median (IQR)	0 (0-1)
Mean (SD)[a]	2.9 (8.8)

[a]The mean is presented to aid comparison with similar programs. The median is our preferred summary measure due to the skewed distribution of these variables.

Table 5. Summary of the number of Healthy Living users who accessed each of the tools on at least 1, 2, and 3 separate days between May 2020 and September 2023.

	Accessed on at least 1 day (n=19,137), n (%)	Accessed on at least 2 days (n=19,137), n (%)	Accessed on at least 3 days (n=19,137), n (%)
Goals	2303 (12)	326 (1.7)	53 (0.3)
Weight tracking	1612 (8.4)	520 (2.7)	312 (1.6)
Food tracking	248 (1.3)	61 (0.3)	34 (0.2)
Step tracking	211 (1.1)	100 (0.5)	71 (0.4)
Blood test tracking	490 (2.6)	151 (0.8)	104 (0.5)
Any tool	3066 (16)	787 (4.1)	425 (2.2)

Usage by User Characteristics

Table 6 summarizes user characteristics separately by "attendance" and "completion" of the program, and gives the results of the logistic regression analyses.

"Attendance" was associated with ethnicity: those with an Asian, Black, or mixed ethnicity had lower odds of "attending" the program compared to those with a White ethnicity. Additionally, male gender, being from a more deprived area, a longer time since diagnosis of type 2 diabetes, and being made aware of Healthy Living via social media, compared with being made aware via a health care professional, were all associated with reduced odds of "attendance." Although statistically significant, the odds ratio for age regarding "attendance" was 1.00 (95% CI 1.00-1.001), suggesting a minimal association.

Associations with "completion" were similar to those for attendance, although the CIs were wider due to a relatively small number of participants "completing" the program. Male gender, being from a more deprived area, Asian, Black, or mixed ethnicity (compared to White ethnicity), a longer time since diagnosis of type 2 diabetes, and being made aware of Healthy Living via social media (compared to being made aware via a health care professional), were all associated with reduced odds of "completion."

XSL•FO
RenderX

Table 6. Attendance and completion of Healthy Living between May 2020 and September 2023 summarized by user characteristics with odds ratios from logistic regression analyses.

	Attendance (n=11,149)[a]				Completion (n=1291)[b]			
	Summary, n (%)		Logistic regression		Summary, n (%)		Logistic regression	
	"Attended" program	Did not "attend" program	Odds ratio (95% CI)	*P* value	"Completed" program	Did not "complete" program	Odds ratio (95% CI)	*P* value
Gender								
Male (baseline)	4220 (38.5)	7424 (46.2)	N/A[c]	N/A	509 (40)	11,135 (43.2)	N/A	N/A
Female	6752 (61.5)	8633 (53.8)	1.35 (1.27-1.43)	<.001	764 (60)	14,621 (56.8)	1.16 (1.01-1.33)	.03
Age (years), mean (SD)	58.5 (11.7)	58.8 (12.3)	1.00 (1.00-1.01)	.03	58.6 (11.3)	58.6 (12.1)	1.00 (1.00-1.01)	.26
IMD quintile[d]								
1 (baseline)	1876 (17.8)	3044 (19.6)	N/A	N/A	201 (16.3)	4719 (19)	N/A	N/A
2	2022 (19.2)	3007 (19.4)	1.06 (0.97-1.17)	.19	237 (19.2)	4792 (19.3)	1.05 (0.84-1.31)	.66
3	2246 (21.3)	3157 (20.3)	1.15 (1.05-1.26)	.003	253 (20.5)	5150 (20.7)	1.04 (0.84-1.29)	.72
4	2232 (21.2)	3369 (21.7)	1.07 (0.98-1.18)	.13	262 (21.3)	5339 (21.5)	1.03 (0.83-1.28)	.77
5	2163 (20.5)	2966 (19.1)	1.17 (1.06-1.28)	.001	279 (22.7)	4850 (19.5)	1.19 (0.96-1.47)	.12
Ethnicity								
Asian	456 (5.7)	1379 (10.8)	0.51 (0.46-0.57)	<.001	34 (3.7)	1801 (9.1)	0.37 (0.26-0.53)	<.001
Black	236 (3)	626 (4.9)	0.57 (0.49-0.67)	<.001	16 (1.7)	846 (4.3)	0.36 (0.22-0.60)	<.001
Mixed	69 (0.9)	159 (1.2)	0.62 (0.46-0.83)	.001	7 (0.8)	221 (1.1)	0.62 (0.29-1.32)	.21
Other	65 (0.8)	137 (1.1)	0.77 (0.57-1.04)	.08	4 (0.4)	198 (1)	0.42 (0.16-1.13)	.09
White (baseline)	7141 (89.6)	10,487 (82)	N/A	N/A	871 (93.5)	16,757 (84.5)	N/A	N/A
Years since diagnosis, mean (SD)	5.4 (6.5)	6.9 (7.1)	0.97 (0.97-0.98)	<.001	4.6 (6.4)	6.4 (6.9)	0.96 (0.95-0.97)	<.001
How the user was made aware of Healthy Living								
Health professional (baseline)	5107 (45.8)	7751 (47.6)	N/A	N/A	646 (50)	12,212 (46.7)	N/A	N/A
Social media	1802 (16.2)	3251 (20)	0.81 (0.75-0.88)	<.001	141 (10.9)	4912 (18.8)	0.52 (0.42-0.65)	<.001
Online search	2172 (19.5)	2584 (15.9)	1.16 (1.07-1.26)	<.001	259 (20.1)	4499 (17.2)	1.12 (0.94-1.34)	.20
Poster/leaflet	476 (4.3)	905 (5.6)	0.88 (0.77-1.01)	.06	63 (4.9)	1318 (5)	1.00 (0.74-1.35)	.99
Other	1590 (14.3)	1782 (11)	1.16 (1.05-1.28)	.002	182 (14.1)	3190 (12.2)	1.05 (0.85-1.30)	.64

[a]Defined as reaching the end of the first section in the Learn Journey.

[b]Defined as accessing at least 60% of the Learn Journey.

[c]N/A: not applicable.

[d]IMD: Index of Multiple Deprivation. IMD scores associated with the lower super output area derived from venue postcodes, ranging from the most

XSL•FO
RenderX

deprived areas in England to the least deprived areas in England.

Discussion

Principal Findings

Less than half of Healthy Living users who activated an account spent more than 5 minutes on the program. Furthermore, less than half of Healthy Living users reached the end of the first section of the structured Learn Journey content, and very few users "completed" the program, that is, accessed 60% of the structured Learn Journey content. Of those who accessed some program content, only a third of Healthy Living users accessed the unstructured Find Answers section at least once, and even fewer users accessed the Tools section. Hence, while a small proportion accessed a large amount of content, many visited the program only once. Female gender, lower deprivation, White ethnicity, and a shorter time since diagnosis were associated with increased usage.

Comparisons With Prior Work

Although the measures of usage are similar but not directly comparable, usage appears to be lower than in the HeLP-Diabetes RCT, where the mean number of log-ins in the intervention group was 18.4 (SD 84.0) over a 12-month follow-up [14], compared to the mean number of sessions completed in Healthy Living being 3.1 (SD 19.9). This difference may reflect the tendency for usage to be higher in RCT settings than in the real-world, due to factors such as participant motivation and additional support provided during trials [21]. In addition, previous research from this program of work identified key changes implemented during the national rollout of Healthy Living that contrasted with the RCT, comprising (1) the inclusion of a structured web-based learning curriculum (the "Learn Journey") due to changes in the NHS policy, (2) a lack of support provided from health care professionals to access the program due to fewer resources in general practice, and (3) the omission of a moderated online support forum due to low uptake of this feature in the HeLP-Diabetes RCT [17]. These changes made during the national rollout may have also contributed to the reasons for lower engagement with Healthy Living.

The results are more comparable to usage analyses in the observational studies of the HeLP-Diabetes: Starting Out program (a structured version of the original HeLP-Diabetes) [24] and the myDESMOND program [22]. In HeLP Diabetes: Starting Out, 23.8% of those who registered reached the end of the first section of structured content, while this percentage was 40.7% for Healthy Living. However, 9.4% of people who registered for HeLP Diabetes: Starting Out completed the program, compared to a completion rate of 4.7% for Healthy Living, although the former was a considerably shorter program [24]. In myDESMOND, 56.3% of users used it for at least a month, and the median number of log-ins was 8 (IQR 4-18) [22], while the median number of sessions completed by users in Healthy Living was 1 (IQR 1-3).

The myDESMOND analysis also found that usage was higher in those aged over 50 years, compared to those aged under 50 years; those aged over 50 years had a higher number of log-ins and spent more time on the program [22]. Similar to our results, the analysis of the myDESMOND program found that usage was higher in females and participants with White ethnicity, although the measures of usage were different. Specifically, females had a similar number of log-ins to males but spent more time on the program, and people with Black or Asian ethnicity had fewer log-ins and spent less time on the program compared to White people [22]. If these differences in usage continue to exist, digital DSMES services risk worsening inequalities based on ethnicity and socioeconomic status, by excluding those groups of society most in need of self-management support. Tailoring to cultural or other specific needs can improve the acceptability and effectiveness of DSMES programs in underrepresented groups [10]; a study of 1678 adults who attended the DESMOND face-to-face program in Leicestershire, including tailoring, identified that 28% of attendances were of South Asian ethnicity, comparable to the 28.3% seen in the local population [25].

Implications

Potential factors contributing to the low usage of Healthy Living could be the structured nature of the content and the lack of interactive elements. NICE guidelines recommend that DSMES programs be structured to ensure participants are exposed to essential components of the program [1]. However, some users have stated they would have preferred to select topics of interest and felt disengaged when a topic was not relevant to them [26]. A way to overcome this could be to allow users the flexibility to access the content they would find most useful, rather than being directed to work through a curriculum. Additionally, the Healthy Living program was self-led and did not involve interaction with health care professionals or peers, unlike HeLP-Diabetes [17], and there is some evidence from similar programs to suggest that usage and outcomes could be improved by brief human interaction [27-30]. Articles with the highest engagement based on time spent on the page were those with a quiz or video, rather than text only. This is in line with findings from interviews with Healthy Living participants, who said they valued interactive components to keep them engaged [26]. The interviews also found that use of the "Tools" section may have been low because users already had other ways of tracking lifestyle factors, such as mobile apps and wearable technology [26].

Our wider program of research has found that the mean reduction in HbA_{1c} was greater in participants with higher engagement in the program (Zghebi et al, unpublished data, 2025). Results from studies of programs aiming to encourage weight loss have also found that higher levels of usage are associated with better health outcomes [31-33]. For example, 1 trial found that participants who demonstrated consistent use of a weight loss maintenance web-based intervention were more successful at maintaining long-term weight loss [32], and another trial found that greater use of a weight tracker tool was associated with greater weight loss [33]. Given that the NHS 10-Year Plan has committed to shifting care from analog to digital and expanding access to digital self-management tools [11], there needs to be more of a focus on bridging the gap

XSL•FO
RenderX

between uptake and engagement with digital programs when implementing at scale.

To explore ways to increase usage with digital programs, future research could explore introducing prompts via SMS text message [34], having more emphasis on unstructured components, or providing more interactive features in the content, such as quizzes and videos. Work could also be done aiming to increase usage specifically in males, those with a Black, Asian, or mixed ethnicity, those from more deprived areas, and people who are not newly diagnosed with type 2 diabetes. Our program of research has found that the program is likely to be cost-effective compared to usual care (Paterson et al, unpublished data, 2025) and therefore could be a viable self-management option to support individuals and to reduce NHS costs for type 2 diabetes treatment and related complications.

Strengths and Limitations

This was an independent evaluation of a nationally implemented digital DSMES program. Data from all users of Healthy Living over a 3-year period were available, and all user activities were recorded, hence we provided a fine-grained and detailed analysis of usage, which is not commonly done. However, the Healthy Living program was in an early stage at the start of the data analysis period, and changes have been made, such as the removal and addition of some sections of information, which may have had an impact on usage. Furthermore, the study period covers the COVID-19 lockdowns, which may have impacted how much people engaged with the program. Two of the variables included in the analyses assessing associations between user characteristics and usage of the program had high levels of missingness: ethnicity and time since diagnosis. We did not perform multiple imputation, despite it being plausible that the data are missing at random or missing not at random, due to the small number of variables. However, a similar analysis as part of our wider body of work using a dataset with more variables (Zghebi et al, unpublished data, 2025) had similar findings.

Conclusions

This analysis explored the usage of a nationally implemented digital DSMES program for >27,000 users with type 2 diabetes and demonstrates the importance of considering usage alongside uptake and effectiveness in any evaluation. Unlike previous evaluations, which have focused on smaller sample sizes, this study offers large-scale, real-world evidence on how people actually engage with a national digital DSMES program. As higher usage is associated with improved health outcomes, encouraging increased engagement with the program has the potential to lead to better health outcomes in people with type 2 diabetes.

Acknowledgments

We would like to thank NHS England and the service provider for facilitating the transfer of usage data to the research team. We would like to thank the HED-LINE Advisory Group and the HED-LINE Patient and Public Involvement and Engagement group for providing feedback and interpretations of these findings during the data analysis stages. With particular thanks to the Healthy Living program team at NHS England, who provided valuable comments during the write-up of this paper.

Funding

This paper reports independent research funded by the National Institute for Health and Care Research (Policy Research Programme, HED-LINE, NIHR200933). The views expressed in this publication are those of the authors and not necessarily those of the National Institute for Health and Care Research or the Department of Health and Social Care.

Data Availability

The datasets generated or analyzed during this study are not publicly available as they were used under license from NHS England for this study only.

Authors' Contributions

Conceptualization: AMM (lead), REH, JSB, DPF, SC
Methodology: AMM (lead), REH, JSB, DPF, SC
Formal analysis: AMM (lead), REH, JSB
Investigation: DPF, SC
Funding acquisition: DPF, SC
Supervision: DPF, SC
Data curation: AMM, SC
Validation: AMM, SC
Interpretation of results: AMM (lead), REH, DPF, SC
Writing – original draft: AMM (lead)
Writing – review and editing: REH, JSB, DPF, SC
All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Screenshots of the 3 main components in the Healthy Living website.
[DOCX File , 444 KB - jmir_v28i1e89690_app1.docx]

Multimedia Appendix 2
Exclusion criteria.
[DOCX File , 19 KB - jmir_v28i1e89690_app2.docx]

Multimedia Appendix 3
Engagement with Learn Journey sections.
[DOCX File , 16 KB - jmir_v28i1e89690_app3.docx]

Multimedia Appendix 4
Articles with the highest engagement.
[DOCX File , 15 KB - jmir_v28i1e89690_app4.docx]

Multimedia Appendix 5
Most commonly accessed articles in the Find Answers section.
[DOCX File , 15 KB - jmir_v28i1e89690_app5.docx]

References

1. Type 2 diabetes in adults: management: [NG28]. National Institute for Health and Care Excellence. 2022. URL: https://www.nice.org.uk/guidance/ng28 [accessed 2025-12-15]
2. Powers MA, Bardsley J, Cypress M, Duker P, Funnell MM, Fischl AH, et al. Diabetes self-management education and support in type 2 diabetes: a joint position statement of the American Diabetes Association, the American Association of Diabetes Educators, and the Academy of Nutrition and Dietetics. Clin Diabetes 2016;34(2):70-80 [FREE Full text] [doi: 10.2337/diaclin.34.2.70] [Medline: 27092016]
3. Chatterjee S, Davies MJ, Heller S, Speight J, Snoek FJ, Khunti K. Diabetes structured self-management education programmes: a narrative review and current innovations. Lancet Diabetes Endocrinol 2018;6(2):130-142. [doi: 10.1016/S2213-8587(17)30239-5] [Medline: 28970034]
4. Powers M, Bardsley J, Cypress M, Funnell MM, Harms D, Hess-Fischl A, et al. Diabetes Care 2020;43(7):1636-1649. [doi: 10.2337/dci20-0023] [Medline: 32513817]
5. Odgers-Jewell K, Ball LE, Kelly JT, Isenring EA, Reidlinger DP, Thomas R. Effectiveness of group-based self-management education for individuals with type 2 diabetes: a systematic review with meta-analyses and meta-regression. Diabet Med 2017;34(8):1027-1039. [doi: 10.1111/dme.13340] [Medline: 28226200]
6. National Diabetes Audit (NDA) 2023-24 quarterly report. National Health Service. 2024. URL: https://digital.nhs.uk/data-and-information/publications/statistical/national-diabetes-audit [accessed 2025-12-15]
7. Davies MJ, Heller S, Skinner TC. Effectiveness of the diabetes education and self management for ongoing and newly diagnosed (DESMOND) programme for people with newly diagnosed type 2 diabetes: Cluster randomised controlled trial. BMJ 2008;336(7649):491-495. [doi: 10.1136/bmj.39553.528299.ad]
8. Deakin TA, Cade JE, Williams R, Greenwood DC. Structured patient education: the diabetes X-PERT programme makes a difference. Diabet Med 2006;23(9):944-954 [FREE Full text] [doi: 10.1111/j.1464-5491.2006.01906.x] [Medline: 16922700]
9. Horigan G, Davies M, Findlay-White F, Chaney D, Coates V. Reasons why patients referred to diabetes education programmes choose not to attend: a systematic review. Diabet Med 2017;34(1):14-26. [doi: 10.1111/dme.13120] [Medline: 26996982]
10. Hadjiconstantinou M, Quinn LM, Tippins F, Schreder S, Khunti K, Davies MJ. A perspective piece on diabetes self-management education and support (DSMES) programmes for under-represented groups with T2DM in the UK. Br J Diabetes 2021;21(1):3-10. [doi: 10.15277/bjd.2021.278]
11. Fit for the future: 10 year health plan for England. NHS England. URL: https://www.gov.uk/government/publications/10-year-health-plan-for-england-fit-for-the-future [accessed 2026-02-23]
12. Nkhoma DE, Soko CJ, Bowrin P, Manga YB, Greenfield D, Househ M, et al. Digital interventions self-management education for type 1 and 2 diabetes: a systematic review and meta-analysis. Comput Methods Programs Biomed 2021;210:106370. [doi: 10.1016/j.cmpb.2021.106370] [Medline: 34492544]

XSL•FO
RenderX

13. Dack C, Ross J, Stevenson F, Pal K, Gubert E, Michie S, et al. A digital self-management intervention for adults with type 2 diabetes: combining theory, data and participatory design to develop HeLP-Diabetes. Internet Interv 2019;17:100241 [FREE Full text] [doi: 10.1016/j.invent.2019.100241] [Medline: 31372349]
14. Murray E, Sweeting M, Dack C, Pal K, Modrow K, Hudda M, et al. Web-based self-management support for people with type 2 diabetes (HeLP-Diabetes): randomised controlled trial in English primary care. BMJ Open 2017;7(9):e016009 [FREE Full text] [doi: 10.1136/bmjopen-2017-016009] [Medline: 28954789]
15. Li J, Parrott S, Sweeting M, Farmer A, Ross J, Dack C, et al. Cost-effectiveness of facilitated access to a self-management website, compared to usual care, for patients with type 2 diabetes (HeLP-Diabetes): randomized controlled trial. J Med Internet Res 2018;20(6):e201 [FREE Full text] [doi: 10.2196/jmir.9256] [Medline: 29884608]
16. Healthy living service specification. NHS England. 2019. URL: https://www.england.nhs.uk/specialised-commissioning-document-library/service-specifications/ [accessed 2026-05-22]
17. Benton JS, Cotterill S, Hawkes RE, Miles LM, French DP. Changes in a digital type 2 diabetes self-management intervention during national rollout: mixed methods study of fidelity. J Med Internet Res 2022;24(12):e39483 [FREE Full text] [doi: 10.2196/39483] [Medline: 36476723]
18. Nahum-Shani I, Yoon C. Towards the science of engagement with digital interventions. Curr Dir Psychol Sci 2024;33(4):239-246. [doi: 10.1177/09637214241254328] [Medline: 39552747]
19. Hardman R, Begg S, Spelten E. What impact do chronic disease self-management support interventions have on health inequity gaps related to socioeconomic status: a systematic review. BMC Health Serv Res 2020;20(1):150 [FREE Full text] [doi: 10.1186/s12913-020-5010-4] [Medline: 32106889]
20. McLeod M, Stanley J, Signal V, Stairmand J, Thompson D, Henderson K, et al. Impact of a comprehensive digital health programme on HbA and weight after 12 months for people with diabetes and prediabetes: a randomised controlled trial. Diabetologia 2020;63(12):2559-2570. [doi: 10.1007/s00125-020-05261-x] [Medline: 32886192]
21. Baumel A, Edan S, Kane JM. Is there a trial bias impacting user engagement with unguided e-mental health interventions? A systematic comparison of published reports and real-world usage of the same programs. Transl Behav Med 2019;9(6):1020-1033. [doi: 10.1093/tbm/ibz147] [Medline: 31689344]
22. Barker MM, Chauhan R, Davies MJ, Brough C, Northern A, Stribling B, et al. User retention and engagement in the digital-based diabetes education and self-management for ongoing and newly diagnosed (myDESMOND) program: descriptive longitudinal study. JMIR Diabetes 2023;8:e44943 [FREE Full text] [doi: 10.2196/44943] [Medline: 37477963]
23. The English indices of deprivation. Department for Communities and Local Government. 2019. URL: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019 [accessed 2025-12-15]
24. Poduval S, Marston L, Hamilton F, Stevenson F, Murray E. Feasibility, acceptability, and impact of a web-based structured education program for type 2 diabetes: real-world study. JMIR Diabetes 2020;5(1):e15744 [FREE Full text] [doi: 10.2196/15744] [Medline: 31904580]
25. Chatterjee S, Davies M, Stribling B, Farooqi A, Khunti K. Real‐world evaluation of the DESMOND type 2 diabetes education and self‐management programme. Practical Diabetes 2018 Feb 08;35(1):19. [doi: 10.1002/pdi.2154]
26. Hawkes RE, Benton JS, Cotterill S, Sanders C, French DP. Service users' experiences of a nationwide digital type 2 diabetes self-management intervention (Healthy Living): qualitative interview study. JMIR Diabetes 2024;9:e56276 [FREE Full text] [doi: 10.2196/56276] [Medline: 39024002]
27. Hawkes RE, Miles LM, Ainsworth B, Ross J, Meacock R, French DP. Engagement with a nationally-implemented digital behaviour change intervention: usage patterns over the 9-month duration of the National Health Service Digital Diabetes Prevention Programme. Internet Interv 2023;33:100647 [FREE Full text] [doi: 10.1016/j.invent.2023.100647] [Medline: 37502122]
28. Ross J, Hawkes RE, Miles LM, Cotterill S, Bower P, Murray E. Design and early use of the nationally implemented Healthier You National Health Service Digital Diabetes Prevention programme: mixed methods study. J Med Internet Res 2023;25:e47436 [FREE Full text] [doi: 10.2196/47436] [Medline: 37590056]
29. Ross J, Cotterill S, Bower P, Murray E. Influences on patient uptake of and engagement with the National Health Service Digital Diabetes Prevention Programme: qualitative interview study. J Med Internet Res 2023;25:e40961 [FREE Full text] [doi: 10.2196/40961] [Medline: 36853751]
30. Dennison L, Morrison L, Lloyd S, Phillips D, Stuart B, Williams S, et al. Does brief telephone support improve engagement with a web-based weight management intervention? Randomized controlled trial. J Med Internet Res 2014;16(3):e95 [FREE Full text] [doi: 10.2196/jmir.3199] [Medline: 24681761]
31. Cussler EC, Teixeira PJ, Going SB, Houtkooper LB, Metcalfe LL, Blew RM, et al. Maintenance of weight loss in overweight middle-aged women through the internet. Obesity (Silver Spring) 2008;16(5):1052-1060 [FREE Full text] [doi: 10.1038/oby.2008.19] [Medline: 18309301]
32. Funk KL, Stevens VJ, Appel LJ, Bauck A, Brantley PJ, Champagne CM, et al. Associations of internet website use with weight change in a long-term weight loss maintenance program. J Med Internet Res 2010;12(3):e29 [FREE Full text] [doi: 10.2196/jmir.1504] [Medline: 20663751]

XSL•FO
RenderX

33. Brindal E, Freyne J, Saunders I, Berkovsky S, Smith G, Noakes M. Features predicting weight loss in overweight or obese participants in a web-based intervention: randomized trial. J Med Internet Res 2012;14(6):e173 [FREE Full text] [doi: 10.2196/jmir.2156] [Medline: 23234759]
34. Bartlett YK, Farmer A, Newhouse N, Miles L, Kenning C, French DP. Effects of using a text message intervention on psychological constructs and the association between changes to psychological constructs and medication adherence in people with type 2 diabetes: results from a randomized controlled feasibility study. JMIR Form Res 2022;6(4):e30058 [FREE Full text] [doi: 10.2196/30058] [Medline: 35486430]

Abbreviations

DSMES: diabetes self-management education and support
HbA1c: hemoglobin A1c
NHS: National Health Service
NICE: National Institute for Health and Care Excellence
RCT: randomized controlled trial

Edited by S Brini; submitted 16.Dec.2025; peer-reviewed by V Johnson, R Deshpande; comments to author 20.Feb.2026; revised version received 02.Apr.2026; accepted 03.Apr.2026; published 02.Jun.2026.

Please cite as:
Marsden AM, Hawkes RE, Benton JS, French DP, Cotterill S
Usage and Exposure to Content of the NHS Healthy Living Program for People With Type 2 Diabetes: Retrospective Observational Cohort Study
J Med Internet Res 2026;28:e89690
URL: https://www.jmir.org/2026/1/e89690
doi:10.2196/89690
PMID:

©Antonia M Marsden, Rhiannon E Hawkes, Jack S Benton, David P French, Sarah Cotterill. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Measuring the Quality of Datasets: Development of the IDEFIM Indicator Set for Empirical Health Research

Sonja Harkener[1]; Oliver J Bott[2], PhD; Christian Draeger[3]; Tobias Hartz[4]; Ekkehart Jenetzky[5,6], MD, PhD; Matthias Löbe[3]; Stefanie March[7], PhD; Chris Schubert[8]; Jürgen Stausberg[1], MD, PhD

[1]Institute for Medical Informatics, Biometry and Epidemiology, Faculty of Medicine, University of Duisburg-Essen, Hufelandstr. 55, Essen, North Rhine-Westphalia, Germany

[2]Institute for Applied Data Science Hannover (DATA|H), Hochschule Hannover - University of Applied Sciences and Arts, Hannover, Lower Saxony, Germany

[3]Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Saxony, Germany

[4]Clinical Cancer Registry Lower Saxony, Hannover, Germany

[5]Faculty of Health/School of Medicine, Witten/Herdecke University, Witten, North Rhine-Westphalia, Germany

[6]University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Rhineland-Palatinate, Germany

[7]Magdeburg-Stendal University of Applied Sciences, Magdeburg, Saxony-Anhalt, Germany

[8]Library, TU Wien, Vienna, Austria

Corresponding Author:
Jürgen Stausberg, MD, PhD
Institute for Medical Informatics, Biometry and Epidemiology, Faculty of Medicine, University of Duisburg-Essen, Hufelandstr. 55, Essen, North Rhine-Westphalia, Germany

Abstract

Background: To be beneficial for empirical health research, a dataset must be fit for use. The quality of a dataset can only be influenced during data collection, yet it is evaluated multiple times during analysis or secondary use by applying quality indicators.

Objective: This study aimed to establish an up-to-date set of indicators measuring the quality of datasets in empirical health research.

Methods: A total of 3 pillars were combined. First, the 51 indicators of a German guideline from 2014 about the management of data quality were revised. Second, a literature review was performed looking for evidence sources since 2013 that describe, propose, or apply dataset quality indicators. Third, indicators were supplemented by a manual search and other sources. The quality indicators were then integrated into the IDEFIM framework. The IDEFIM framework distinguishes between the categories' data, metadata, context, and openness quality. In this work, only the categories data and metadata quality, with their 14 dimensions were considered.

Results: In total, 69 indicators qualified for the IDEFIM indicator set, 53 related to the category data quality, and 16 to the category metadata quality. A total of 30 indicators originated from the German guideline, 31 from the literature review. Three indicators were added to cover aspects of diversity, equity, and inclusion, and an additional 5 related to specifics of data and metadata quality not addressed so far. Most indicators were found in the dimensions accuracy (data) with 12 measures, completeness (data) with 12 measures, and consistency (data) with 19 measures. According to the number of supporting evidence sources, missing values in data elements (48 evidence sources), contradictions (31), and currentness (26) were the most popular quality indicators. Metadata quality was significantly less frequently addressed.

Conclusions: The presented IDEFIM indicator set can be used for the management of data collections as well as for the verification of a dataset's quality for an intended use. The indicator set should also be considered in the design of a study in empirical health research and the development of software tools supporting the visualization of issues related to the quality of a dataset.

(J Med Internet Res 2026;28:e90482) doi:10.2196/90482

KEYWORDS

dataset; data quality; empirical health research; health care; metadata; quality indicator; secondary use; IDEFIM framework

XSL•FO
RenderX

Introduction

The goal of having data that is fit for use accompanies the entire life cycle of empirical health research (Figure 1). It begins with the development of a data collection following a systematic approach as a study [1] or registry protocol [2]. Ideally, predefined research questions guide the determination of populations, samples, visits, and variables, to name a few aspects [3]. It is worthwhile to consider study-specific requirements for characteristics of a dataset, such as case completeness, data completeness, and correctness, already in the development phase, because the design - as well as the available resources - will substantially contribute to fulfilling these requirements. Many recommendations have been made for appropriate designs focusing on different types of empirical health research, for example, CONSORT (Consolidated Standards of Reporting Trials) for randomized trials [4] or STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) for observational research [5].

Figure 1. Life cycle of datasets in empirical health research from the development phase to their linkage and secondary use.

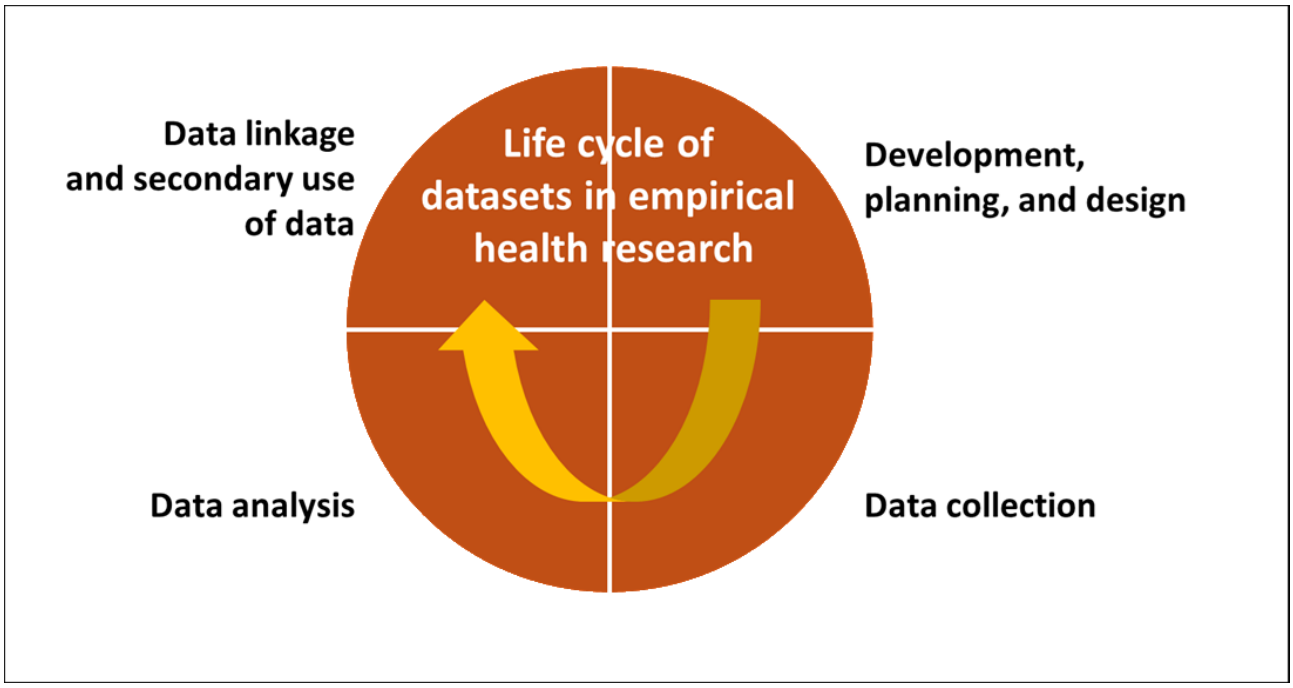

Project-specific requirements regarding the quality of a dataset can be measured, evaluated, and benchmarked against individual or common thresholds. During data collection, the plan-do-check-act cycle (PDCA cycle) described by Deming [6] can be applied to detect weak points and to implement actions for improvement. The data quality standards of the International Organization for Standardization (ISO) are based upon the approach of active quality management that was brought into practice in the 1950s by United States-Americans such as Crosby [7] and Juran [8]. ISO 8000 defines data quality as the degree to which a set of inherent characteristics of data fulfills requirements [9]. The degree is something measurable that could be quantified using a “data quality measure” introduced in ISO/International Electrotechnical Commission (IEC) 25024 [10]. However, even within ISO standards, there are different definitions of the term “data quality,” each of them reasonable in the respective subject areas. Returning to the PDCA cycle, the measures about the quality of a dataset support the check part within the PDCA cycle, and the consequences determine the act part of the cycle regarding the management and operation of a data collection.

In empirical health research, data collection is usually followed by an analysis phase [11]. Once data collection is complete (with the exception of interim analyses), the design of the data collection could not be changed. The same is true for secondary analyses of data, for example, from electronic health records [12]. Measuring quality is important in the analysis phase to assess whether a dataset is fit for the intended use or for a secondary use. However, reviewing a dataset related to its design, its documentation, its availability, and similar aspects may also be useful. On the one hand, reviews supplement the measurement of characteristics with additional information, such as the FAIRness of a dataset [13]. On the other hand, reviews may be less demanding because access to a dataset is not necessary to evaluate its fitness for use. This advantage of reviews compared to complex measurements could motivate the data quality and utility label proposed for the secondary use of data in the European Health Data Space (EHDS) [14]. However, simply documenting information on data quality does not indicate the certain degree of quality needed, for example, to assess the appropriateness of a dataset with regard to different purposes of empirical health research, such as effectiveness evaluation as the most prominent purpose or hypothesis generation as a basic purpose [15].

New challenges concerning data quality arise from using datasets to generate large language models [16], from the availability of large and uncontrolled volumes of data (“big data”) [17], and from linking datasets either as part of an active data collection or for analysis work [18]. The associated risk of

XSL•FO
RenderX

discrimination against vulnerable groups [19] demands a strong consideration of data quality issues.

IDEFIM aims to support the collection and analysis phases of empirical health research by offering an up-to-date list of measures that assess a dataset's quality. A dataset was defined as a logically meaningful grouping of data [20] and as an "identifiable collection of data available for access or download in one or more formats" [21]. The terminology in the field of data quality was diverse and inconsistent. Therefore, IDEFIM kept the term "quality indicator" from previous work [22] as a synonym for "quality measure" [23]. Quality indicators measure the degree to which requirements concerning characteristics of datasets are met. IDEFIM assumed that the degree to which a dataset possesses a particular characteristic can be quantified by several quality indicators, each focusing on a different aspect of that characteristic. The goal of IDEFIM is to establish a comprehensive and consistent set of quality indicators for empirical health research, embedded in a common framework addressing a dataset's fitness for purpose and fitness for use. In previous publications, we presented preparatory and supporting work, including our motivation [24], our framework covering dimensions and categories [25], and a structure for specifying a dataset quality indicator [26]. Here, we introduce the IDEFIM indicator set for empirical health research.

Methods

IDEFIM Framework

The IDEFIM framework consists of 4 categories with 21 dimensions (Figure 2 [25]), 8 dimensions in category data quality, 6 in category metadata quality, 3 in category context quality, and 4 in category openness quality. In accordance with ISO 8000, quality indicators measure the degree to which general requirements for characteristics, that is, dimensions, are met. This degree can be used to evaluate the fitness of a dataset for different purposes of empirical health research, such as health services research, quality research, or drug approval research. To evaluate a dataset's fitness for use in a particular project, quality indicator instances are applied to adjust the quality indicators to the particular setting, such as the recorded data elements. Quality checks represent the algorithms applied to the data. IDEFIM focused on quality indicators and a respective indicator set. However, quality indicators seem not to be appropriate with regard to the categories of context and openness quality. Respective approaches, such as the FAIR Guiding Principles [13] and the 5-star open data approach [27], propose Boolean conditions regarding structures (eg, the existence of an open license as a 1-star prerequisite) and processes (eg, an authentication and authorization procedure to access data as one FAIR Guiding Principle). Boolean conditions are not suitable for calculating measures. Consequently, IDEFIM examined quality indicators only with regard to the categories data and metadata quality.

Figure 2. IDEFIM framework with five layers supporting the goal to achieve datasets that are fit for purpose and fit for use.

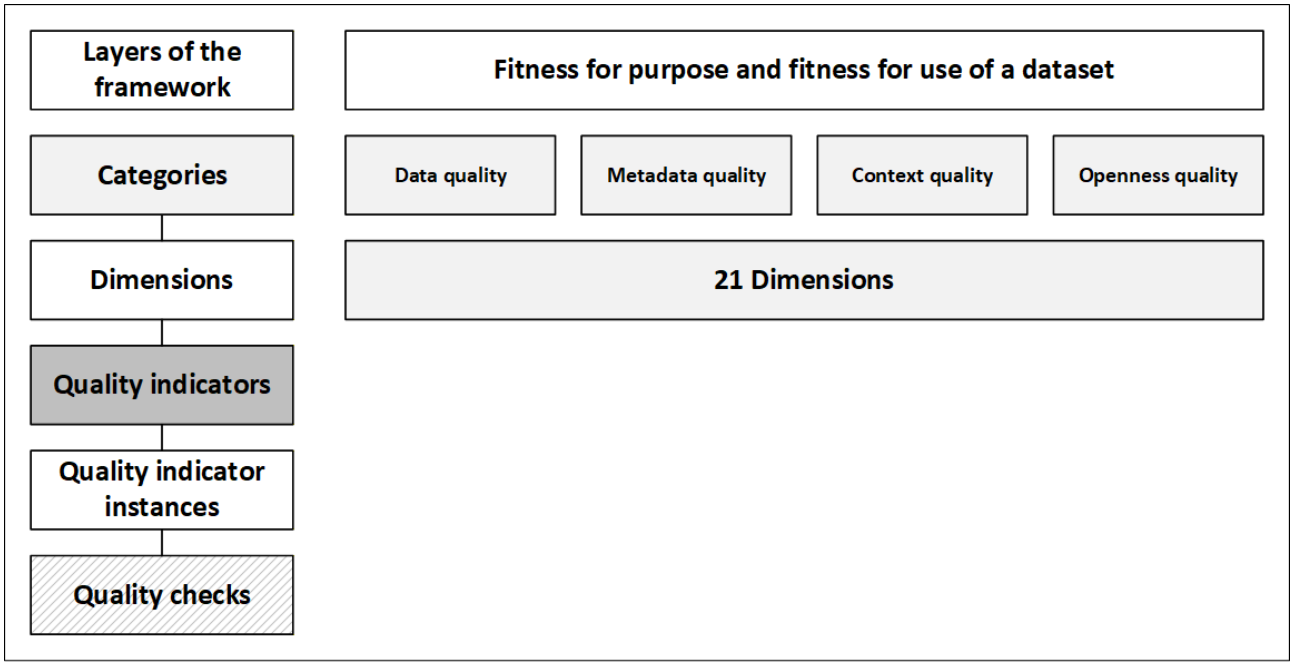

Material

Overview

The development of the IDEFIM indicator set was based on 3 pillars. First, it builds on national efforts to create a guideline for an adaptive management of data quality in cohort studies and registries [28,29]. Particularly, IDEFIM used indicators published in 2014 under the auspices of the Technology and Methods Platform for Networked Medical Research (TMF), with the second version of this guideline. These TMF indicators were developed through a systematic process over the years based on literature reviews, expert consultations, and community involvement [22]. Second, a broad literature review was conducted in the field of data quality in empirical health research that included the search for evidence sources proposing,

XSL•FO
RenderX

describing, or using indicators [24]. This review expanded similar work from the TMF guidelines by searching for publications from 2013 onwards until January 2024. Third, the project identified additional indicators through empirical knowledge of the core project team, manual search looking at relevant journals and conference proceedings, snowball sampling starting with records retrieved during the literature search, and consultations of invited experts in 2 workshops. The first face-to-face workshop with 8 experts took place in November 2024, focusing on the framework. The second face-to-face workshop with 7 experts took place in August 2025, focused on a proposal from the project concerning quality indicators. The workshops were strictly advisory. The core research team made the final decision to include or reject an indicator. The combination of the 3 pillars resulted in a first draft of the set, which explicitly relates each indicator to its origin. This draft was condensed into the final set described here. Each indicator in the final set was described using a uniform structure [26].

TMF Indicators

The guideline contains 51 indicators organized into 3 categories: integrity, organization, and trueness. These indicators are defined in a structured way using 14 items such as sources, calculation, and interpretation. As IDEFIM was concerned with empirical research projects on a general level, 10 of these indicators were excluded because they were either cancer registry specific or concerned measurements and their conditions that particularly occur in cohort studies (Multimedia Appendix 1). Thus, 41 indicators remained for IDEFIM from the TMF guideline. Each indicator was then assigned to exactly 1 dimension. Names and definitions of the indicators were revised and extended as necessary to create a consistent and up-to-date set of quality indicators.

Literature Search

The literature search included Medline, the Cochrane Library, the Web of Science, and Scopus [24]. It was conducted on February 2, 2024, for Medline and on January 16, 2024, for the other 3 information sources. The search looked at English or German records published since January 1, 2013. The search criteria were tailored to the individual options offered by the 4 information sources (Multimedia Appendix 2). The results were merged, and duplicates were eliminated, resulting in 2748 records.

Due to the large volume of records, the screening based on titles and abstracts was split into 2 stages. A prescreening focused on data quality as a topic, the relationship to empirical health research, and general relevance. The records were divided equally among 2 reviewers. The prescreening took place between February 9, 2024, and March 6, 2024. The remaining 734 records were evaluated independently by 2 reviewers (SH and JS) between March 12, 2024, and April 8, 2024. The relationship of a record to the topic of data quality could be further differentiated into the following aspects: the structure for the description of data quality indicators, indicators of data quality, dimensions of data, gender equality and diversity in relation to data quality, and other references to data quality. An optional comment could be entered for each search result. In case of disagreement between the reviewers, a consensus was reached, resulting in 221 records.

Out of the 221 search results, 219 were obtained in full text. Here as well, due to the unexpectedly high number of remaining records, the search results were randomly divided in half and evaluated using the following criteria: type of scientific project (eg, analysis of data quality), reference to empirical research projects in medicine (eg, registry), reference to data quality (eg, indicators for data quality), and recommendation for further consideration. This left 117 sources of evidence for the review. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram is presented in Multimedia Appendix 3. Table 1 shows the distribution of the 117 sources with regard to aspects of data quality.

Table . Distribution of the 117 evidence sources with regard to aspects of data quality (multiple options possible).

Category	Frequency
Dimensions of data	101
Quality indicators	79
Standards of data quality	28
Special statistical procedures	21
Overall score for data quality	19
Interventions to improve data quality	18
Quality checks	15
Structure of quality indicators	9
Quality of metadata and paradata	6
Gender equality and diversity with reference to data quality	4
Other reference to data quality	28

XSL•FO
RenderX

Literature Review

The literature search yielded 117 relevant evidence sources, out of which 79 specifically related to data quality indicators. Each evidence source was examined by one reviewer (SH) for possible indicators, indicator instances, and quality checks. Due to the diverse and inconsistent terminology of measures, indicator instances, and quality checks were considered at this stage. Respective findings were recorded for each evidence source. However, some sources yielded no findings. In these cases, the entire source was recorded as a single finding. This resulted in a total of 622 findings, ie, candidates for quality indicators, from 79 evidence sources. A total of 2 reviewers (JS and SH) assessed each finding independently. If a finding corresponded to one of the 41 considered TMF indicators, it was marked accordingly. If no assignment made sense, either a "new indicator," "not applicable," or "undecidable" flag could be set. Proposals for new indicators were mainly considered in relation to the categories of data and metadata quality. It was assumed that requirements concerning context and openness quality are better defined using Boolean conditions than measures. Table 2 shows the results of this stage. The reviewers agreed on 265 findings, and they disagreed on 357 findings.

Table . Initial evaluation of quality indicator findings by two reviewers (JS and SH).

Option	Frequency
Both reviewers performed a consistent mapping to a TMF[a] indicator	174
Both reviewers performed a mapping to a TMF indicator, but it was different	131
One reviewer performed a mapping to a TMF indicator; the other did not	152
Both reviewers unanimously opted for "new indicator"	43
Both reviewers unanimously opted for "not applicable"	48
Other scenarios, eg, combinations of new, not applicable, and undecidable	74

[a]TMF: Technology and Methods Platform for Networked Medical Research.

Both reviewers discussed the 357 disagreements and reached a consensus. Findings that were not considered proposals for an indicator, an indicator instance, or a quality check were deleted. Findings with suggestions for mapping to 2 or 3 indicators were broken down into individual entries leading to a numerical growth. This left 74 evidence sources with a total of 624 findings: 99 findings proposing a new indicator, 428 findings with a successful mapping to a TMF indicator, and 97 findings rejected. The median number of findings per source was 5.5 (range 1-54) findings. A total of 30% (187/624) of the findings came from 6 evidence sources [30-35]. Findings qualifying for a new indicator were combined based on the description in the respective evidence sources. Each new indicator was assigned to one of the 14 dimensions in the categories data and metadata quality. A Microsoft Access database was used for the literature review, including the management of the retrieved records, the reviews, and the analysis of agreement.

IDEFIM Contributions

Proposals for new indicators beyond the literature review were permitted to maximize the use of community knowledge, even if the proposals were published more recently than the literature search. These proposals were identified through snowball sampling from evidence sources, a manual search of other publications proposing quality indicators, input from scientific conferences, and contributions from invited experts. According to the project work plan, the relationship between data quality and the concepts of diversity, equity, and inclusion (DEI) was examined based on the initial records from the literature search [36]. Related indicators were also established and added to the indicator set.

Condensation of the Draft Indicator Set

Only indicators assigned to a dimension in categories data and metadata quality were accepted. According to the IDEFIM project protocol, TMF indicators without evidence from at least 1 independent source in the literature review were excluded from the final indicator set. Furthermore, the draft set was screened for possible merges. In addition to the originating pillar - TMF work, literature review, and IDEFIM contributions - the indicator proposals were not formally and empirically validated.

Ethical Considerations

This research did not involve human participants.

Results

The draft set included 81 indicators. A total of 6 initially kept TMF indicators were excluded from the draft set because they were not referenced by an evidence source. Additionally, 4 TMF indicators were defined as subconcepts of 2 other indicators using information expected on the level of indicator instances. These 4 indicators were also excluded (Multimedia Appendix 1). A total of 2 indicators, 1 TMF indicator and 1 indicator from the review, were skipped because they refer to a dimension in the category context quality. Table 3 shows the distribution of the final set of 69 indicators. Of the 41 TMF indicators, 30 remained after condensation, all dealing with data quality. A total of 31 new indicators were added based on findings from the literature review. These 31 new indicators were distributed fairly evenly between both categories, data and metadata quality. While the total number of indicators was lower in category metadata quality, they were distributed fairly evenly across its dimensions. In contrast, the dimensions of the category data quality were unevenly covered by indicators. Accuracy (data),

XSL•FO
RenderX

completeness (data), and consistency (data) were the most dominant here. The 8 indicators added by IDEFIM did not significantly change the distribution.

Table . Distribution of the 69 indicators among the 14 dimensions in the data and metadata quality categories of the IDEFIM indicator set.

Category and dimension	Origin			
	Total	TMF[a]	Literature review	IDEFIM contribution
Data quality	53	30	16	7
Accuracy (data)	12	6	4	2
Completeness (cases)	3	2	—[b]	1
Completeness (data)	12	7	3	2
Compliance (data)	3	—	3	—
Consistency (data)	19	13	5	1
Credibility	1	—	1	—
Currentness	1	1	—	—
Representativeness	2	1	—	1
Metadata quality	16	0	15	1
Accuracy (metadata)	2	—	2	—
Completeness (metadata)	4	—	4	—
Compliance (metadata)	2	—	2	—
Consistency (metadata)	4	—	3	1
Precision	2	—	2	—
Understandability	2	—	2	—
Total	69	30	31	8

[a]TMF: Technology and Methods Platform for Networked Medical Research.

[b]Not available.

Table 4 lists 53 indicators related to data quality organized into 16 indicator groups for clarity, with 1 to 4 groups per dimension. The 16 indicators contributed by the literature review originated from 26 evidence sources. IDEFIM added 3 indicators based on a separate analysis of DEI issues [36]. Additionally, James et al [37] propose the R-index as a standardized representativeness metric for benchmarking DEI in a dataset. The R-index is listed in the IDEFIM indicator set as “Conspicuous representativeness distribution.” To complete the consideration of the contingency table [38,39], the F-Measure was defined as the harmonic mean of correctness and recall based on a recommendation from the second workshop and named “Validity.” Time series, for example, for lab values, are a special phenomenon in health-related datasets. The idea of Giesa et al [40] to address this phenomenon with an indicator that provides a rate of incomplete time series was accepted. While homonyms of observational units were considered in the TMF guideline, homonyms of pieces of information were missing and supplemented according to Woodall et al [41].

XSL•FO
RenderX

Table . Indicators related to data quality with origin and number of supporting evidence sources (ES).

Dimension, indicator group, and indicator	Origin	Number of ES[a]
Accuracy (data)		
Contingency table indicators		
Conspicuous correctness distribution	IDEFIM[b] (DEI[c])	—[d]
Correctness	TMF[e]	18
Recall	TMF	6
Validity	IDEFIM	—
Disagreement with source data indicators		
Disagreement with source data referring to data elements	TMF	21
Disagreement with source data referring to observational units	TMF	1
Illegal content indicators		
Illegal values of qualitative data elements	TMF	11
Illegal values of qualitative data elements used for the coding of missings	TMF	4
Incorrect text in qualitative data elements	Review	4
Misfielded values	Review	3
Other indicators for accuracy (data)		
Granularity (data)	Review	2
Temporal trends in counts or proportions	Review	1
Completeness (cases)		
Indicators for completeness (cases)		
Conspicuous recruitment rate distribution	IDEFIM (DEI)	—
Drop-out-rate	TMF	6
Recruitment rate	TMF	15
Completeness (data)		
Missing content indicators		
Conspicuous missing values distribution	IDEFIM (DEI)[c]	—
Missing modules	TMF	5
Missing values in data elements	TMF	48
Temporal missingness	IDEFIM [40]	—
Other indicators for completeness (data)		
Data elements with existing entries for all observational units	TMF	2
Information density score	Review	1
Modules with existing entries for all data elements	Review	1
Observational units with existing entries for all data elements	Review	4
Observational units with follow-up	TMF	1
Refusal rate indicators		
Refusal rate of investigations	TMF	1
Refusal rate of modules	TMF	1
Refusal rate of single data elements	TMF	1
Compliance (data)		
Indicators for compliance (data)		

XSL•FO
RenderX

Dimension, indicator group, and indicator	Origin	Number of ES[a]
Data format, data type, and unit compliance	Review	13
Incompliance with metadata	Review	1
Range compliance	Review	8
Consistency (data)		
Confusion and redundancy indicators		
Confusion	IDEFIM [41]	—
Duplicates (data)	TMF	13
Homonyms (data)	TMF	2
Synonyms (data)	TMF	8
Contradiction indicators		
Contradictions	TMF	31
Data element contradictions	Review	2
Other indicators for consistency (data)		
Missing evidence of known correlations	TMF	4
Single data source per observational unit	TMF	2
Temporality of categorical data elements	Review	1
Unexpected entry indicators		
Concordance	TMF	14
Conspicuous distribution of digits in date-time data elements	Review	1
Conspicuous distribution of values	TMF	6
Data elements with value unknown etc	TMF	3
Disagreement with previous values	TMF	3
Frequency outliers	Review	1
Last digit preferences	TMF	2
Outliers (continuous data elements)	TMF	17
Outliers in numerical data elements in a multivariate analysis	Review	1
Values from external references	TMF	3
Credibility		
Indicators for credibility		
Data element credibility	Review	2
Currentness		
Indicators for currentness		
Currentness	TMF	26
Representativeness		
Indicators for representativeness		
Conspicuous representativeness distribution	IDEFIM [37]	—
Representativeness	TMF	7

[a]ES: evidence source.
[b]IDEFIM: contributed by the project, not the literature review.
[c]DEI: diversity, equity, and inclusion.
[d]Not available.
[e]TMF: Technology and Methods Platform for Networked Medical Research.

XSL•FO
RenderX

The number of supporting evidence sources derived from the literature review can be used to weight indicators against each other. A total of 3 indicators were affected by more than one-third of the 74 evidence sources: “Missing values in data elements” (48 evidence sources), “Contradictions” (31), and “Currentness” (26). Two additional indicators were supported by at least a quarter of the evidence sources: “Correctness” (18) and “Disagreement with source data referring to data elements” (21). Multimedia Appendix 4 gives a list of all evidence sources for each indicator in the category data quality.

Table 5 lists the 16 indicators related to metadata quality. Because there are fewer indicators, there is only 1 indicator group per dimension. The 15 indicators from the literature review came from 10 evidence sources. IDEFIM added 1 indicator: “Duplicates” refers to a comparison of metadata between multiple datasets, that intentionally do not overlap [41]. This indicator counts redundantly defined data elements. In the category metadata quality, the number of supporting evidence sources was less suitable as an indicator weight. Only 2 indicators had more than 1 supporting evidence source, “Metadata format, type, and unit compliance” (3 evidence sources) and “Granularity (metadata)” (2). Multimedia Appendix 5 provides a list of all evidence sources for each indicator in the category metadata quality.

Table . Indicators related to metadata quality with origin and number of supporting evidence sources (ES).

Dimension, indicator group, and indicator	Origin[a]	Number of ES[b]
Accuracy (metadata)		
Indicators for accuracy (metadata)		
Correctness (metadata)	Review	1
Responsiveness	Review	1
Completeness (metadata)		
Indicators for completeness (metadata)		
Completeness of administrative metadata	Review	1
Coverage of all data elements	Review	1
Data element completeness	Review	1
Richness (metadata)	Review	1
Compliance (metadata)		
Indicators for compliance (metadata)		
Data element compliance with reference	Review	1
Metadata format, type, and unit compliance	Review	3
Consistency (metadata)		
Indicators for consistency (metadata)		
Duplicates (metadata)	IDEFIM [41]	__[c]
Heterogeneous representation of data elements	Review	1
Homonyms (metadata)	Review	1
Synonyms (metadata)	Review	1
Precision		
Indicators for precision		
Granularity (metadata)	Review	2
Residual classes of qualitative data elements	Review	1
Understandability		
Indicators for understandability		
Easy of understanding	Review	1
Relevance of the dataset’s descriptive information	Review	1

[a]IDEFIM: contributed by the project, not the literature review.

[b]ES: evidence source.

[c]Not available.

XSL•FO
RenderX

Discussion

Principal Findings

With 69 indicators, the IDEFIM indicator set expanded its predecessor, the TMF guideline [22], by one-third. The 2 most comprehensive indicator lists found in the literature review contributed 54 [33] and 35 findings [34]. ISO/IEC 25024 offers 63 quality measures [10]. Considering IDEFIM's coverage of data as well as metadata quality, 69 indicators seem manageable compared to potential competitors. However, the IDEFIM set does not cover the categories context and openness quality. Multimedia Appendix 6 contains the specifications of each indicator.

As expected, not all TMF indicators were confirmed. However, the final set impressively demonstrated the relevance of the adopted indicators. The 8 most frequently mentioned indicators originated from the TMF guideline. Conversely, 21 out of 31 indicators added by the review were mentioned only once. In the category metadata quality, both the number of indicators was lower and the supporting evidence was weaker than in the category data quality. This does not necessarily mean that metadata quality was out of scope. Metadata quality might be appropriately evaluated with a mixture of measures (as is the case in the category data quality) and conditions (as it might be the case in the categories context and openness quality). Nevertheless, indicators of metadata quality could be a critical area for future work to effectively publish open data from empirical health research [42].

The IDEFIM indicator set and its framework provide a foundation for measuring the quality of datasets in empirical health research. Further work could lead to a community approach of sharing and harmonizing instances and applications of its indicators. This includes recommendations for using the indicators, indicator instances, and quality checks in specific use cases. The cross-registry benchmarking of data in health services research can serve as an initial example here [43]. The quality indicators were taken from the TMF guideline. Each participating registry then tailored the indicators to its own setting by specifying data elements considered with regard to missing values, determining its particular denominator with regard to the recruitment rate, and defining the rules used to count contradictions, for example. This tailoring resulted in quality indicator instances for each registry. However, the quality indicators' common ground with identical meanings and thresholds allowed for a comparison of results between registries. Due to the registries' different information models, data structures, and concept systems, each registry was responsible for specifying and implementing quality checks.

Comparison With Prior Work

Like others, IDEFIM understood a quality indicator as a distinct entity, broken down into components in its specification. Woodall et al [41] took a different approach. They distinguish between 9 data quality problems, 9 generic data quality methods, and 6 so-called taxonomy elements. Taxonomy elements combine different objects of a data structure from the perspective of a relational database management system. Quality measures arise from the cross-tabulation of data quality problems and taxonomy elements, which involves assigning none, 1, or several generic data quality methods to each intersection point. Of the initial 54 intersection points, 16 were left blank as senseless and 6 were labeled as gaps that could be filled in subsequent work. The remaining 32 intersection points were filled with 57 entries. To address all aspects proposed by Woodall et al [41], IDEFIM added 2 indicators. "Duplicates (metadata)" represented the cross-domain analyses mentioned by Woodall et al [41] at the level of data elements intended to reduce redundancy. Second, IDEFIM supplemented the data quality problem "Existence of synonyms and homonyms" with the indicator "Confusion," which examines homonyms of values in a dataset. The approach of Woodall et al [41] turned out to be a good training partner for IDEFIM; however, constructing indicators by combining different axes appeared too theoretical for practical use.

QUANTUM proposes a data quality and utility label [44] that might be used in the EHDS according to article 78 of the respective regulation [14]. In this regulation, a "data quality and utility label means a graphic diagram, including a scale, describing the data quality and conditions of use of a dataset." This concept is closely related to measuring dataset quality with the IDEFIM framework. However, there is a fundamental difference between QUANTUM and IDEFIM. QUANTUM examines the documentation of quality-related aspects, while IDEFIM examines quality-related aspects directly. In combination, IDEFIM might constitute the basis from which QUANTUM derives its rating. Table 6 shows a comparison between QUANTUM and IDEFIM. Most of QUANTUM's measure labels represent Boolean conditions that do not provide information about data or metadata quality, for example, the "availability of a data access & usage policy at the time of release of the dataset". Some measure labels represent measurable quantities such as the coverage rate. For this comparison, we assigned IDEFIM indicators to a measure label even if the label represents a Boolean condition. In these cases, the indicator can quantify the fulfillment of a condition or evaluate the fulfillment of a condition based on a predefined threshold.

XSL•FO
RenderX

Table . Comparison between the measures of QUANTUM [44] and IDEFIM.

QUANTUM		IDEFIM
Dimension	Measure label	Reference
Accessibility	• Availability of a data access & usage policy at the time of release of the dataset • Average time from data access application to data release for a specific dataset	• Category: openness quality • Dimension: availability • Category: openness quality • Dimension: accessibility
Population coverage	• Coverage Rate (percentage of the eligible population represented in the dataset)	• Category: data quality • Dimension: completeness (cases) • Quality indicator: “Recruitment rate”
Population representativity	• How closely does the observed population represent the expected population?	• Category: data quality • Dimension: representativeness • Quality indicator: “Representativeness”
Compliance	• Is there documentation of compliance with ethical standards, conventions, protocols or regulations?	• Category: context quality • Dimension: compliance (context)
Data provenance	• Is the source of the dataset documented?	• Category: context quality • Dimension: provenance
	• Are the processes and operations on the data documented?	• [not addressed]
Metadata scope	• Existence of comprehensive standardised metadata	• Category: metadata quality • Dimension: completeness (metadata) • Quality indicator: “Richness (metadata)”
	• Existence of an exhaustive data dictionary at variable level	• Category: metadata quality • Dimension: completeness (metadata) • Quality indicator: “Coverage of all data elements”
Accuracy	• Is the accuracy of the dataset documented?	• Category: data quality • Dimension: accuracy (data) • Quality indicator: [several]
Coherence	• Is coherence of the dataset documented?	• Category: data quality • Dimension: accuracy (data) • Group of indicators: illegal content • Quality indicator: [several]
Completeness	• Is completeness of the dataset documented?	• Category: data quality • Dimension: completeness (data) • Quality indicator: [several]
Consistency	• Is the consistency of the dataset documented?	• Category: data quality • Dimension: consistency (data) • Quality indicator: [several]
Precision	• Is the precision of the dataset documented?	• Category: metadata quality • Dimension: precision • Quality indicator: “Granularity (metadata)”
Validity	• Availability of a conformance report for the data model	• Category: data quality • Dimension: compliance (data) • Quality indicator: [several]

From the comparison, one can conclude that IDEFIM addressed all measures of QUANTUM with quality indicators, if the categories data quality and metadata quality are concerned. Therefore, using the IDEFIM indicator set will simplify the registration for the EHDS data quality and utility label to this extent. However, the terminology of QUANTUM differs and does not refer to ISO standards, terms, or definitions. This can

XSL•FO
RenderX

be attributed to the unusual framework and wording of the EHDS regulation.

Limitations

Some limitations of the work must be noted. First, the variety of terminologies used to define data quality and its dimensions posed a challenge. For instance, we found the terms aspects, attributes, categories, characteristics, domains, features, indicators, metrics, processes, and submetrics to be synonyms or siblings of the term "dimension." The same was true for the term "indicator" noted, for example, as measure, check, or even dimension. In cases of doubt, the criteria were interpreted generously that may have led to misleading sources. Second, maintaining the focus on empirical health research was difficult. This dilemma demonstrates the importance of publishing terms in a citable manner to avoid ambiguity in interpretation and to encourage reliable reuse. Substantial methodological contributions to data quality originated from other fields, even outside of health care. We did not want to overlook important proposals for quality indicators, so we accepted some sources from outside the domain of empirical health research. The domain addressed in the literature was sometimes unclear or overlapped with empirical health research. Particularly, the secondary use of data raises awareness about the initial quality of health-related data, that is, the quality of data from daily health care. Sources were also accepted outside the intended domain if their approach appeared relevant. Third, IDEFIM does not reflect solely the result of its literature review. Although IDEFIM considered the PRISMA extension for Scoping Reviews [45], it uses the results of the literature review as one but not the only basis for its indicator set. However, it would not have been possible to develop a relevant indicator set alone from the literature review due to the terminological but also conceptual confusion. Therefore, the IDEFIM indicator set offers a unique perspective. The bottom-up approach that gives evidence to each indicator is a strength and a unique selling point. Fourth, it is important to note that formal consensus mechanisms were not applied to all parts of this work. For instance, the literature review was one but not the only pillar that led to the inclusion of indicator proposals in the IDEFIM indicator set. Ultimately, the IDEFIM indicator set reflects the decisions of the core project team. Furthermore, a formal and empirical validation of individual indicators was outside the scope of the presented work. Whenever possible and appropriate, this validation is a project of its own for each indicator.

Conclusions

IDEFIM offers the most up-to-date set of indicators intended to measure the quality of a dataset in empirical health research, as far as the authors know. This set can be used to manage and control the data collection in a clinical trial, a cohort study, or a patient registry, for example. Furthermore, datasets can be labeled according to their fitness for purpose in principle. Then, health data users can verify electronic health data or claims data with respect to the intended use. All indicators of the IDEFIM set were defined using a uniform structure and integrated into the IDEFIM framework of categories and dimensions. However, there is no automatism that allows for the application of the indicators from scratch. It is up to the users to adapt quality indicators to their specific needs, eg, data elements of interest, to create respective indicator instances, and to implement quality checks operating on the data representation layer. This work does not oppose tool-based approaches to data quality control, such as the DataQualityDashboard [46] and dataquieR [47]. Rather, it offers a conceptual basis for the appropriate application of related tools in empirical health research practice.

Acknowledgments

The authors used DeepL Write in the refinement, correction, editing, or formatting of the manuscript during the review process to improve clarity of language.

Funding

The project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 506700202.

Data Availability

The individual indicator specifications in version 0.8 are available in Multimedia Appendix 6.

Authors' Contributions

Conceptualization: JS
Data curation: SH
Formal analysis: SH
Funding acquisition: JS
Methodology: JS
Project administration: SH
Resources: SH, JS
Software: SH
Supervision: JS
Validation: OJB, CD, TH, EJ, ML, SM, CS
Writing – original draft: SH, JS

XSL•FO
RenderX

Writing – review and editing: SH, OJB, CD, TH, EJ, ML, SM, CS, JS

Conflicts of Interest

None declared.

Multimedia Appendix 1
Original set of TMF indicators.
[DOCX File, 32 KB - jmir_v28i1e90482_app1.docx]

Multimedia Appendix 2
Search strings used for Medline, the Cochrane Library, the Web of Science, and Scopus.
[DOCX File, 26 KB - jmir_v28i1e90482_app2.docx]

Multimedia Appendix 3
PRISMA flow diagram.
[PNG File, 135 KB - jmir_v28i1e90482_app3.png]

Multimedia Appendix 4
Evidence sources per indicator in category data quality.
[DOCX File, 51 KB - jmir_v28i1e90482_app4.docx]

Multimedia Appendix 5
Evidence sources per indicator in category metadata quality.
[DOCX File, 32 KB - jmir_v28i1e90482_app5.docx]

Multimedia Appendix 6
Individual indicator specifications in version 0.8.
[PDF File, 532 KB - jmir_v28i1e90482_app6.pdf]

References

1. ICH harmonised guideline for good clinical practice E6(R3). : International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use; 2025 Jan 6 URL: https://database.ich.org/sites/default/files/ICH_E6(R3)_Step4_FinalGuideline_2025_0106.pdf [accessed 2025-12-29]
2. Gliklich RE, Leavy MB, Dreyer NA. Registries for Evaluating Patient Outcomes: A User's Guide: Agency for Healthcare Research and Quality; 2020. [doi: 10.23970/AHRQEPCREGISTRIES4]
3. Stausberg J, Maier B, Bestehorn K, et al. Memorandum registry for health services research: update 2019. Gesundheitswesen 2020 Mar;82(3):e39-e66. [doi: 10.1055/a-1083-6417] [Medline: 32069507]
4. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. Lancet 2025 Apr 14;405:1633-1640. [doi: 10.1016/S0140-6736(25)00672-5] [Medline: 40245901]
5. Elm EV, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007 Oct 20;335(7624):806-808. [doi: 10.1136/bmj.39335.541782.AD]
6. Deming WE. Out of The Crisis: Quality, Productivity and Competitive Position: Cambridge University Press; 1982.
7. Crosby PB. Quality Is Free: The Art of Making Quality Certain: Penguin; 1980.
8. Juran JM, Godfrey AB. Juran's Quality Handbook, 5th edition: McGraw Hill; 1999.
9. ISO 8000-1:2022 data quality — part 1: overview. International Organization for Standardization (ISO). URL: https://www.iso.org/standard/81745.html [accessed 2026-06-03]
10. ISO/IEC 25024:2015(en), systems and software engineering — systems and software quality requirements and evaluation (square) — measurement of data quality. International Organization for Standardization (ISO). URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:25024:ed-1:v1:en [accessed 2026-06-03]
11. An D, Lim M, Lee S. Challenges for data quality in the clinical data life cycle: systematic review. J Med Internet Res 2025 Apr 23;27:e60709. [doi: 10.2196/60709] [Medline: 40266662]
12. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 1;20(1):144-151. [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
13. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3(1):160018. [doi: 10.1038/sdata.2016.18] [Medline: 26978244]

XSL•FO
RenderX

14. Regulation (EU) 2025/327 of the European parliament and of the council of 11 February 2025 on the European health data space and amending directive 2011/24/EU and regulation (EU) 2024/2847 (text with EEA relevance). EUR-Lex. URL: https://eur-lex.europa.eu/eli/reg/2025/327/oj/eng [accessed 2025-12-29]
15. Malin JL, Keating NL. The cost-quality trade-off: need for data quality standards for studies that impact clinical practice and health policy. J Clin Oncol 2005 Jul 20;23(21):4581-4584. [doi: 10.1200/JCO.2005.01.912] [Medline: 15851767]
16. Siddique SM, Evans CV, Harhay M, et al. Critical appraisal for racial and ethnic equity in clinical prediction models extension: development of a critical appraisal tool extension to assess racial and ethnic equity-related risk of bias for clinical prediction models. Health Equity 2023;7(1):773-781. [doi: 10.1089/heq.2023.0035] [Medline: 38076212]
17. Ramasamy A, Chowdhury S. Big data quality dimensions: a systematic literature review. Journal of Information Systems and Technology Management 2020;17:e202017003. [doi: 10.4301/S1807-1775202017003]
18. Assaf A, Senart A, Troncy R. Towards an objective assessment framework for linked data quality: enriching dataset profiles with quality indicators. Int J Semant Web Inf Syst 2016;12:111-133. [doi: 10.4018/IJSWIS.2016070104]
19. Aspinall PJ. Measuring the health patterns of the 'mixed/multiple' ethnic group in Britain: data quality problems, reporting issues, and implications for policy. Int J Soc Res Methodol 2018 May 4;21(3):359-371. [doi: 10.1080/13645579.2017.1399623]
20. ISO 8000-2:2022(en), data quality - part 2: vocabulary. International Organization for Standardization (ISO). URL: https://www.iso.org/obp/ui/en/#iso:std:iso:8000:-2:ed-5:v1:en [accessed 2026-06-03]
21. ISO/IEC 11179-33:2023(E) Information technology - metadata registries (MDR) - part 33: metamodel for data set registration. : ISO/IEC; 2023 URL: https://cdn.standards.iteh.ai/samples/81725/ec9431074c324417a349c5fddbdf7ac9/ISO-IEC-11179-33-2023.pdf [accessed 2026-06-03]
22. Stausberg J, Bauer U, Nasseh D, et al. Indicators of data quality: review and requirements from the perspective of networked medical research. GMS Med Inform Biom Epidemiol 2019;15:Doc05. [doi: 10.3205/mibe000199]
23. ISO/IEC 25012:2008(E) Software engineering - software product quality requirements and evaluation (square) - data quality model. : ISO/IEC; 2008 URL: https://cdn.standards.iteh.ai/samples/35736/3791c8ca8fd64a7fa4c7b629ec8f8524/ISO-IEC-25012-2008.pdf [accessed 2026-06-03]
24. Stausberg J, Harkener S, Bünz S. A collection of data quality indicators for health research: rationale for an update. Stud Health Technol Inform 2024 Nov 22;321:254-258. [doi: 10.3233/SHTI241103] [Medline: 39575819]
25. Stausberg J, Harkener S, Draeger C, et al. Quality of data sets in empirical health research: the IDEFIM framework. Stud Health Technol Inform 2025 Aug 7;329:103-107. [doi: 10.3233/SHTI250810] [Medline: 40775828]
26. Stausberg J, Harkener S. How to define a data quality indicator? Stud Health Technol Inform 2025 Jun 26;328:266-267. [doi: 10.3233/SHTI250716] [Medline: 40588923]
27. Martin EG, Law J, Ran W, Helbig N, Birkhead GS. Evaluating the quality and usability of open data for public health research: a systematic review of data offerings on 3 open data platforms. J Public Health Manag Pract 2017;23(4):e5-e13. [doi: 10.1097/PHH.0000000000000388] [Medline: 26910872]
28. Nonnemacher M, Weiland D, Stausberg J. Datenqualität in Der Medizinischen Forschung Leitlinie Zum Adaptiven Management von Datenqualität in Kohortenstudien Und Registern: Medizinisch Wissenschaftliche Verlagsgesellschaft; 2007.
29. Nonnemacher M, Nasseh D, Stausberg J. Datenqualität in Der Medizinischen Forschung: Leitlinie Zum Adaptiven Management von Datenqualität in Kohortenstudien Und Registern: Medizinisch Wissenschaftliche Verlagsgesellschaft; 2014. URL: https://www.mwv-open.de/site/books/10.32745/9783954663743 [doi: 10.32745/9783954663743]
30. Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. J Am Med Inform Assoc 2021 Sep 18;28(10):2251-2257. [doi: 10.1093/jamia/ocab132] [Medline: 34313749]
31. Laberge M, Shachak A. Developing a tool to assess the quality of socio-demographic data in community health centres. Appl Clin Inform 2013;4(1):1-11. [doi: 10.4338/ACI-2012-10-CR-0041] [Medline: 23650483]
32. Liu CH, Talaei-Khoei A, Storey VC, Peng GC. A Review of the State of the Art of Data Quality in Healthcare. Journal of Global Information Management 2023 Jan;31(1):1-18. [doi: 10.4018/JGIM.316236]
33. Quindroit P, Fruchart M, Degoul S, et al. Definition of a practical taxonomy for referencing data quality problems in health care databases. Methods Inf Med 2023 May;62(1-02):19-30. [doi: 10.1055/a-1976-2371] [Medline: 36356592]
34. Schmidt CO, Struckmann S, Enzenbach C, et al. Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. BMC Med Res Methodol 2021 Apr 2;21(1):63. [doi: 10.1186/s12874-021-01252-7] [Medline: 33810787]
35. Wu D, Xu H, Yongyi W, Zhu H. Quality of government health data in COVID-19: definition and testing of an open government health data quality evaluation framework. LHT 2022 Mar 29;40(2):516-534. [doi: 10.1108/LHT-04-2021-0126]
36. Bünz S, Harkener S, Stausberg J. Data quality and diversity in health care and health research. Stud Health Technol Inform 2025 May 15;327:442-446. [doi: 10.3233/SHTI250376] [Medline: 40380486]
37. James SL, Bourgognon M, Vieira PP, et al. R-index: a standardized representativeness metric for benchmarking diversity, equity, and inclusion in biopharmaceutical clinical trial development. EClinicalMedicine 2025 Feb;80:103079. [doi: 10.1016/j.eclinm.2025.103079] [Medline: 39968390]

XSL•FO
RenderX

38. Hogan WR, Wagner MM. Accuracy of data in computer-based patient records. J Am Med Inform Assoc 1997;4(5):342-355. [doi: 10.1136/jamia.1997.0040342] [Medline: 9292840]
39. Logan JR, Gorman PN, Middleton B. Measuring the quality of medical records: a method for comparing completeness and correctness of clinical encounter data. Proc AMIA Symp 2001:408-412. [Medline: 11825220]
40. Giesa N, Akguel M, Boie SD, Balzer F. GRU-D characterizes age-specific temporal missingness in MIMIC-IV. Stud Health Technol Inform 2025 May 15;327:472-476. [doi: 10.3233/SHTI250382] [Medline: 40380492]
41. Woodall P, Oberhofer M, Borek A. A classification of data quality assessment and improvement methods. IJIQ 2014;3(4):298. [doi: 10.1504/IJIQ.2014.068656]
42. Šlibar B, Oreški D, Begičević Ređep N. Importance of the open data assessment: an insight into the (Meta) data quality dimensions. Sage Open 2021 Apr;11(2):11. [doi: 10.1177/21582440211023178]
43. Stausberg J, Harkener S, Engel C, et al. Cross-Registry Benchmarking of Data Quality: Lessons Learned. Stud Health Technol Inform 2023 May 18;302:167-171. [doi: 10.3233/SHTI230096] [Medline: 37203640]
44. Daumas J, Schäfer A, Dolanski-Aghamanoukjan L, et al. Deliverable 1.1 specification of the data sets' quality and utility label. Zenodo. 2024. URL: https://zenodo.org/records/14937423 [accessed 2025-12-29]
45. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020 Oct;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
46. Schuemie M, Reps J, Black A, et al. Health-Analytics Data to Evidence Suite (HADES): open-source software for observational research. Stud Health Technol Inform 2024 Jan 25;310:966-970. [doi: 10.3233/SHTI231108] [Medline: 38269952]
47. Richter A, Schmidt CO, Krüger M, Struckmann S. DataquieR: assessment of data quality in epidemiological research. J Open Source Softw 2021;6(61):3093. [doi: 10.21105/joss.03093]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
DEI: diversity, equity, and inclusion
DFG: Deutsche Forschungsgemeinschaft
EHDS: European Health Data Space
ES: evidence source
IEC: International Electrotechnical Commission
ISO: International Organization for Standardization
PDCA: plan-do-check-act
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TMF: Technology and Methods Platform for Networked Medical Research

Edited by A Coristine; submitted 29.Dec.2025; peer-reviewed by T Sagi, W Xu; revised version received 13.May.2026; accepted 15.May.2026; published 17.Jun.2026.

Please cite as:
Harkener S, Bott OJ, Draeger C, Hartz T, Jenetzky E, Löbe M, March S, Schubert C, Stausberg J
Measuring the Quality of Datasets: Development of the IDEFIM Indicator Set for Empirical Health Research
J Med Internet Res 2026;28:e90482
URL: https://www.jmir.org/2026/1/e90482
doi:10.2196/90482

© Sonja Harkener, Oliver J Bott, Christian Draeger, Tobias Hartz, Ekkehart Jenetzky, Matthias Löbe, Stefanie March, Chris Schubert, Jürgen Stausberg. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Online Social Interaction, Neighborhood Perception, and the Mediating Role of Social Capital in Charitable Giving for Seriously Ill Patients: Cross-Sectional Study

Jie Chen[1,2], MSc; Junzhang Hu[1], MSc; Yang Yang[3], PhD; Fangjuan Du[1], PhD; Xuelian Li[4], MSc

[1]School of Geography and Environmental Sciences (School of Karst Science), Guizhou Normal University, Huaxi University Town, Guian New District, Guiyang, China

[2]School of Tourism and Air Service, Guizhou Minzu University, Guiyang, China

[3]School of International Tourism and Culture, Guizhou Normal University, Guiyang, China

[4]Faculty of Geography, Yunnan Normal University, Kunming, China

Corresponding Author:
Fangjuan Du, PhD
School of Geography and Environmental Sciences (School of Karst Science), Guizhou Normal University, Huaxi University Town, Guian New District, Guiyang, China

Abstract

Background: Online public welfare and donation-based crowdfunding have rapidly expanded, motivating individuals to engage in charitable giving through digital platforms. However, the influence of online social interactions and offline neighborhood perceptions on donation behavior remains underexplored.

Objective: This study investigates how online social interactions influence individuals' intentions to donate to patients with serious illness, focusing on the mediating roles of bridging and bonding social capital and the moderating effect of neighborhood perception. It integrates online and offline social contexts to explain donation behavior.

Methods: Data were collected from individuals recruited through the "3C" (Caring for Children with Cancer) Health Care Volunteer Community in Guizhou, China. Validated scales were used to measure online social interaction, social capital, donation intention, and neighborhood perception. A total of 450 valid responses were obtained between August and October 2023, with a 91.46% response rate.

Results: Online social interaction was positively associated with donation intention (β=0.37, *P*<.01). It significantly predicted both bridging social capital (β=0.45, *P*<.01) and bonding social capital (β=0.54, *P*<.01), which in turn were positively associated with donation intention (β=0.48 and β=0.54, respectively, *P*<.01). Bridging social capital partially mediated this relationship, whereas bonding social capital showed a stronger mediating effect. Neighborhood perception significantly moderated the relationship between online social interaction and both bridging (β=0.17, *P*<.01) and bonding social capital (β=0.18, *P*<.01). Furthermore, moderated mediation analysis showed that the indirect effects were stronger at higher levels of neighborhood perception (bridging: β=0.278; bonding: β=0.340) than at lower levels (bridging: β=0.150; bonding: β=0.181). The indices of moderated mediation were significant for both bridging (β=0.080, 95% CI 0.021-0.142) and bonding social capital (β=0.100, 95% CI 0.040-0.162).

Conclusions: Online social interaction was associated with increased donation intentions to patients with serious illness via bridging and bonding social capital, with bonding ties fully and bridging ties partially mediating the effect. Neighborhood perception strengthened these pathways, indicating that offline social context can amplify the impact of online engagement on charitable giving. These findings highlight the importance of integrating digital interaction and offline social context in promoting health-related charitable giving.

(J Med Internet Res 2026;28:e85573) doi:10.2196/85573

KEYWORDS

online social interaction; bridging social capital; bonding social capital; donation; neighborhood perception; seriously ill patients

XSL•FO
RenderX

Introduction

Background

The rapid expansion of digital technologies has transformed health-related charitable giving into an increasingly prominent component of the broader digital health (eHealth) ecosystem [1,2]. Online medical crowdfunding platforms, as a form of digitally mediated health support, enable individuals to mobilize financial resources for patients in need [3]. For example, Tencent's public welfare platform claimed that user donations reached CNY 5.6 billion (approximately US $790 million) in 2022, involving more than 127 million donors (Tencent Public Welfare Platform, 2023). This indicates the expanding scale of digital charity activity. However, such expansion does not necessarily result in equal assistance outcomes. Evidence suggests that access to and effectiveness of digital health–enabled charitable support remain uneven [4]. Patients with serious illnesses, particularly in less-developed regions such as Western China, often receive comparatively less financial support through crowdfunding platforms [5-7]. Recent empirical evidence indicates that regional socioeconomic disparities and the "digital divide" significantly constrain fundraising performance, as individuals in these areas often lack the extensive online social networks required to mobilize sufficient donations [8]. This highlights a critical challenge within digital health systems: although technological platforms expand opportunities for participation, they do not automatically translate into equitable health-related support outcomes. Understanding what drives individuals to engage in donation behavior within these digitally mediated environments is therefore essential. Prior research suggests that charitable donations can enhance civic awareness and social cohesion while enabling individuals to express prosocial values [9]. However, recent national reports indicate a decline in overall social donations. For instance, the 2023 China Philanthropy Development Report revealed a 4.63% decrease compared to previous years, suggesting increasing uncertainty in digital charitable participation. To better explain these disparities, it is necessary to move beyond structural explanations and examine the relational mechanisms embedded in digital platforms. In particular, the role of online social interaction as a key driver of donation behavior remains insufficiently understood.

Though the mechanisms underlying public donation have been widely studied, the influence of individual cyberspace engagement on donation intention remains insufficiently understood. Previous research has distinguished between endogenous and exogenous determinants of donation behavior, including enjoyment and peer influence [10]. However, empirical evidence on how online social interactions shape willingness to donate to patients with serious illness remains limited [11]. Existing research has primarily focused on charitable giving and broader forms of prosocial engagement in general social networking contexts [12,13], with less attention paid to health care–related charitable contexts. In such situations, donation decisions are often influenced by trust, empathy, and emotional vulnerability [14,15]. In addition, the role of offline neighborhood perception in shaping digital prosocial participation has received little systematic attention [16]. The above indicates that the mechanisms underlying the relationships between online social interactions, social capital, donation intentions, and perceptions of neighborhood in health care settings warrant further investigation. To address this, this study focuses on active forms of online social interaction, including direct communication and engagement behaviors such as liking and commenting [17]. These interactions reflect emotional support and identification with charitable activities, thereby facilitating patient care and communication within online communities [18,19].

From a social capital perspective, charitable donation can be understood as a form of resource exchange embedded in social relationships [20,21]. Accordingly, this study adopts a relational perspective to examine how online social interaction influences donation intention through bridging and bonding social capital, and how this process is conditioned by neighborhood perception as an offline contextual factor [22,23]. Focusing on health care volunteer communities, this study highlights the distinct roles of bridging and bonding social capital in serious illness donation contexts. Bridging ties facilitate information diffusion and network expansion, whereas bonding ties are particularly salient in health care–related settings due to their role in fostering trust, empathy, and emotional support [14,24]. This study integrates online social interactions, social capital, and perceptions of offline neighborhood, thereby expanding the application of social capital theory in the field of digital health philanthropy and deepening our understanding of the mechanisms underlying digital altruistic behavior.

Literature Review and Hypotheses: Social Capital Theory

Social capital has been conceptualized in various ways, but is generally understood as resources embedded in social networks and accessible through social relationships [24]. Despite differing definitions, the core concept of social capital involves social networks or structures and the resources they provide. Social capital can be manifested through individuals' social networks and is commonly divided into 2 dimensions: bridging and bonding social capital [20]. Bridging social capital emphasizes superficial and diverse relationships, typically involving weak ties between people from different backgrounds. It broadens social horizons and facilitates information exchange. Bonding social capital involves strong ties between family or close friends and reflects emotional or practical support from individuals with similar backgrounds or deep connections. Social networks comprise channels of information and emotional bonds, embedding the norms and value system that promote good behavior among individuals, forming social capital [24].

Interactions in cyberspace contribute to the accumulation of online social capital [23], which is obtained by connecting with others through information communication technologies [22]. A key feature of the internet is that it allows people to make new connections online or maintain offline relationships [24]. As such, it can be viewed as a tool for maintaining relationships or achieving goals [25]. From a social capital perspective, internet use serves as an important mechanism for accumulating social resources and promoting prosocial behaviors such as community participation and helping others [11]. It can also

XSL•FO
RenderX

influence willingness to reciprocate [26], sense of community [23], and continuance intention [27] to promote and stimulate personal behaviors. More broadly, social capital fosters trust, cooperation, and reciprocity, which in turn facilitate collective action and shape individual decision-making processes [28-30]. These mechanisms suggest that social capital functions as a relational resource shaping behavioral intentions in online and offline contexts.

Online Social Interaction and Intention to Donate

Social interaction is a key factor in facilitating information exchange [31] and provides a foundational condition for promoting prosocial behavior (eg, charity or donation). Social media use is associated with civic engagement, life satisfaction, and social trust [32,33]. It also provides informational and emotional support, and resources for self-regulation [10,34]. In terms of use patterns, researchers distinguish between passive and active social media use. Passive use refers to content consumption, while active use emphasizes interactive engagement and mutual communication [17]. Active social interaction enables users to create a positive experience through encouragement, affirmation, and mutual support [35].

Given the virtual nature of the internet, online interactions weaken an individual's personal identity and enhance group identity. It gives people a sense of belonging when connecting or sharing common experiences [23], thus encouraging them to continue commenting on each other's content or sharing experiences [18]. Positive interactions also create social pressure for individuals to engage in collective activities [35]. From a collective standpoint, many positive behaviors arise from the pressure of social norms in social interactions [36]. Personal interactions highlight and reinforce social norms, thereby guiding individual behavior [37]. Studies have further shown that it is positively associated with charitable giving and donation behavior [38,39]. Therefore, we formulated the following hypothesis: online social interaction is positively associated with intention to donate (H1).

Mediation of Bridging Social Capital

Social media encourages the exchange of information, advice, and ideas based on the needs and resources of common interest [40], allowing users to obtain the relevant social support [25]. Through low-cost and convenient communication, online interaction strengthens weak ties among individuals and fosters the development of bridging social capital [23,24]. For example, Ellison et al [34] found that social media use can facilitate and predict bridging social capital. Hence, we hypothesized: online social interaction positively affects bridging social capital (H2a).

The benefits of bridging social capital can be reflected in its connection with external assets and diffusion of information [20]. After the formation of networks with weak ties, information flows faster, and connections become closer [12]. Bridging social capital influences users' attitudes and perceived credibility in stimulating individual engagement [28]. A bridging social capital perspective enables an increase in community engagement and collective cooperation based on large and diverse networks [41]. Visible donations incentivize donors to improve their status or maintain relationships [38], and individuals are accustomed to comparing their donations with those of others [42]. Therefore, we posited: bridging social capital positively affects intention to donate (H2b).

The internet creates an environment in which individuals can communicate with others, which benefits the expansion and extension of social networks [43]. Network ties reduce the time and investment required to gather information and knowledge, which in turn generates norms, enables individuals to participate in and contribute to activities, and increases their willingness to maintain relationships [44]. Research has clarified that bridging social capital accumulated through online interactions also leads to positive psychological outcomes that motivate individuals to donate [32,43]. Therefore, scope-based bridging social capital provides a large stock of social resources to guide people to make charitable donations [21]. Combining H1, H2a, and H2b, we proposed: bridging social capital mediates the positive relationship between online social interaction and intention to donate (H2).

Mediation of Bonding Social Capital

Bonding social capital is rooted in emotional and substantive support derived from close and intimate relationships formed through strong social ties [24]. Internet use can supplement offline relationships with family or close friends [45]. Interactions within a virtual platform strengthen the bonds between members [23], and online social connections can evolve into lasting and strong relationships over time [46]. Studies have suggested a direct relationship between online interactions and bonding social capital. For example, social networking platforms provide new opportunities for maintaining and strengthening strong ties [47], expanding family and friend networks while offering emotional support [48]. In addition, individuals tend to interact with similar others online, which facilitates the development of trust-based and emotionally secure relationships [49]. Thus, we hypothesized: online social interaction is positively associated with bonding social capital (H3a).

Bonding social capital is characterized by strong ties and has a considerable impact on outcomes and human behavior, such as emotional support, collective action, and mobilizing solidarity [24]. Personal involvement in charitable organizations plays a significant role in generosity and donations [50]. Additionally, strong family and neighborhood social networks positively affect donations [51]. We can, thus, infer that volunteers who have close contact with patients or experience of caring for them may have strong emotions and expectations, build bonding social capital, and promote charitable giving. Therefore, we posited: bonding social capital positively affects the intention to donate (H3b).

Compared to bridging social capital, bonding social capital may play a more important role in health care–related donation contexts, as serious illness donations often involve trust, empathy, and emotional vulnerability [14,52]. In such situations, individuals are more likely to rely on relationships based on emotional support and trust rather than on relationships based on weak ties [24]. Therefore, bonding social capital may have a stronger influence on donation intention than bridging social capital.

XSL•FO
RenderX

Social bonding mediates the relationship between online social interactions and an individual's intentions [27]. Online participation can create the perception that others support community goals. This may lead users to form common norms and pursue the same goals [33]. Membership and friendships in the community correlate with the frequency of donations [53]. Trust, competence, discipline, and kindness are built on substantial interactions among communities and form bonding social capital, thus promoting good behavior. Combining H1, H3a, and H3b, we proposed: bonding social capital mediates the positive relationship between online social interaction and intention to donate (H3).

Moderating Role of Neighborhood Perception

Social cohesion and donation networks suggest that prosocial behaviors are embedded within broader social environments and relational contexts [16]. Neighborhoods, as localized social settings, reflect varying levels of cohesion, safety, and interpersonal connectedness [54]. In this study, neighborhood perception is conceptualized as individuals' subjective evaluation of social cohesion, safety, and relational quality within their local environment. Prior research suggests that individuals in more advantaged neighborhoods tend to display higher levels of social cohesion than those in disadvantaged neighborhoods, leading to differences in social attitudes and behavioral outcomes [55]. Such variation reflects how individuals interpret and respond to their surrounding social environments. Rather than representing a single mechanism, neighborhood perception operates as a contextual condition that shapes social cognitive processes. Specifically, it may influence the formation of online social capital through a norm-based mechanism, in which individuals become more sensitive to social expectations and reputational concerns when perceiving their neighborhood environment as socially salient, and a contact-based mechanism, in which higher neighborhood perception facilitates trust and interpersonal interaction [56,57].

In the context of norm-based mechanisms, symbolic threats refer to potential challenges to morals, beliefs, and norms caused by groups with different value systems. This can involve the risk of loss of face or honor and potentially hurt an individual's self-identity or self-esteem [58]. According to the social capital theory, bridging social capital can trigger individuals to generate cognitive responses or stress behaviors and make themselves conform to related norms [59]. Accordingly, individuals with stronger neighborhood perception may experience greater normative awareness during online social interactions, which facilitates the development of bridging social capital. Individuals with higher neighborhood perception are more sensitive to local social norms and reputational cues, which may facilitate broader social interactions and the development of bridging social capital [60]. We propose that individuals with higher neighborhood perception are more likely to develop stronger bridging social capital through social cohesion and interaction. Thus, we proposed: neighborhood perception moderates the direct effect of online social interaction on bridging social capital (H4).

From a contact-based perspective, neighborhood perception plays an important role in promoting social integration and fostering harmonious social relationships [61]. The contact mechanism states that increased opportunities for interpersonal interaction within neighborhoods enhance mutual understanding, reduce bias, and promote social cohesion [54]. Social networks in more cohesive neighborhoods are more likely to be channels for information circulation and emotional bonding. Interpersonal contact reduces bias, improves intergroup relationships, and increases empathy and understanding [62]. As individuals develop stronger perceptions of neighborhood cohesion, they are more likely to internalize a sense of belonging, trust, and a sense of community [63]. These relational and affective processes foster deeper interpersonal connections and reinforce strong-tie networks. Therefore, stronger neighborhood perception is expected to enhance bonding social capital through the contact-based mechanism, particularly in the context of online social interaction. Thus, we proposed: neighborhood perception moderates the direct effect of online social interaction on bonding social capital (H5).

Neighborhoods capture the social environment and, consequently, affect individuals' prosocial behavior in the real world [64]. Individuals embedded in neighborhoods with stronger neighborhood ties produce stronger sources of social influence [16]. Building on the dual-mechanism framework, neighborhood perception is expected to strengthen the indirect effects of online social interaction on donation intention through both bridging and bonding social capital. From a norm-based perspective, higher neighborhood perception heightens sensitivity to social norms and reputational expectations, promoting engagement in broader weak-tie networks and thereby reinforcing the indirect effect via bridging social capital [16]. From a contact-based perspective, stronger neighborhood perception enhances trust, interaction, and a sense of belonging, fostering emotionally supportive ties and strengthening the indirect effect via bonding social capital [61]. Taken together, these complementary mechanisms suggest that higher neighborhood perception amplifies both indirect pathways linking online social interaction to donation intention. Based on the above mediation hypotheses, we proposed: neighborhood perception moderates the indirect effect of online social interaction on the intention to donate via bridging social capital, such that the indirect effects are stronger when neighborhood perception is higher (H6) and neighborhood perception moderates the indirect effect of online social interaction on the intention to donate via bonding social capital, such that the indirect effects are stronger when neighborhood perception is higher (H7). Figure 1 presents the conceptual model of this study. The model illustrates the direct effect of online social interaction on intention to donate (H1), 2 parallel mediation pathways through bridging and bonding social capital (H2a-H2b; H3a-H3b), the moderating effects of neighborhood perception on the relationships between online social interaction and both types of social capital (H4, H5), and the moderated mediation effects (H6, H7). In the figure, oval shapes represent latent constructs, arrows indicate hypothesized relationships, and H denotes a hypothesis.

XSL•FO
RenderX

Figure 1. Conceptual model of the study.

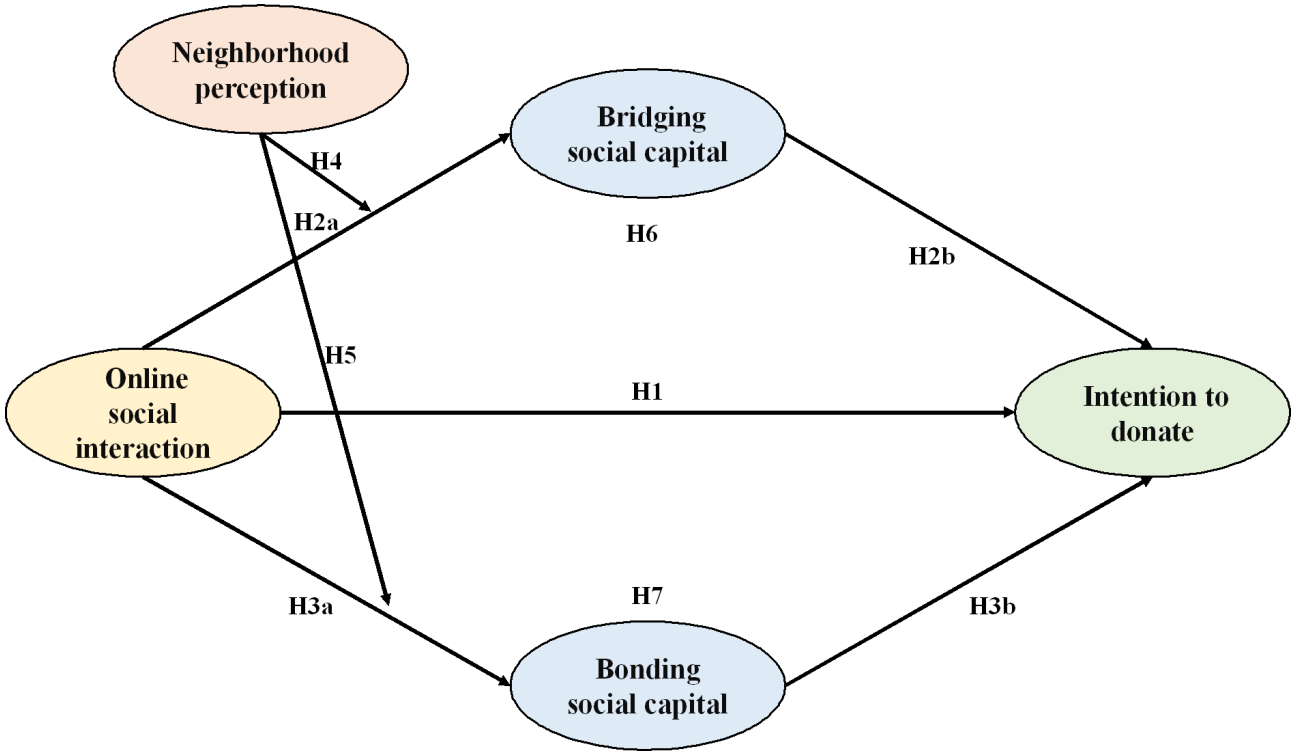

Methods

Study Setting

This study was conducted within the context of the "3C" (Caring for Children with Cancer) Health Care Volunteer Community, a health care support initiative founded by university faculty members and involving serious illness care and volunteer community engagement. The 3C community operates primarily in Guizhou, China, through collaborations with hospitals and rehabilitation organizations. The 3C community has developed both online and offline engagement platforms, including WeChat Moments and Public Channels (Tencent), to facilitate communication and dissemination of charitable activities. Regular volunteer activities are conducted on weekdays in the Children's Hematology Department of Guiyang Maternal and Child Health Hospital, including games, painting, handicrafts, education, and music for children with leukemia. The 3C collaborates with public service centers to conduct follow-up care for patients with serious illnesses and organize nature-based therapeutic activities for children with cancer and their families. In addition, the community operates an online crowdfunding initiative ("Qiantong Health") to support charitable fundraising for children with serious illnesses. Through ongoing online and offline care activities, the community gradually attracted broader participation beyond core volunteers, including peripheral supporters and individuals exposed to its charitable initiatives or online content. Therefore, these online volunteer care communities, together with their extended social networks, represent typical cases that reflect our research aims.

Participants and Recruitment

Participants were recruited from the "3C" Health Care Volunteer Community and its extended social networks. A purposive sampling strategy was used to ensure that participants were engaged in or influenced by volunteer care activities related to serious illness. Eligible participants included individuals with varying levels of connection to the 3C community, including (1) those directly involved in volunteer-related activities, such as hospital-based volunteer services, online community interactions, or follow-up care activities; and (2) broader community members who were exposed to the 3C community through its online content, charitable initiatives, or related support activities. Potential participants were approached through multiple channels, including WeChat groups, online community platforms, and face-to-face invitations during ongoing volunteer activities at the hospital. Individuals who expressed interest were screened for eligibility based on their level of engagement with or connection to the volunteer care network. All eligible participants were informed about the purpose and procedures of the study. Participation was voluntary, and informed consent was obtained prior to data collection.

Data Collection

Data collection was divided into 2 phases: the precollection stage (August 25 - 30, 2023) and the formal data collection stage (September 23-October 8, 2023). In phase 1, 200 questionnaires were distributed and 176 were collected, with an 88% response rate. We eliminated items with low reliability and modified any ambiguous items. A formal questionnaire was distributed online and offline in phase 2. To ensure the relevance of the survey context, respondents were required to have some degree of familiarity with or exposure to health care volunteer activities or related charitable initiatives associated with the 3C community. These respondents had been identified prior to the commencement of the study. A total of 492 questionnaires were distributed; 488 were collected, and 450 were valid, for a response rate of 91.46%. The statistical information of the

XSL•FO
RenderX

samples is shown in Table 1. To reduce potential bias, anonymity was ensured and items were presented in a mixed order.

Table . Demographic characteristics of participants (N=450).

Category and item	Value, n (%)
Sex	
Male	166 (36.9)
Female	284 (63.1)
Age (years)	
≤20	121 (26.9)
21 - 30	266 (59.1)
31 - 40	34 (7.6)
41 - 50	19 (4.2)
51 - 60	7 (1.6)
>61	3 (0.7)
Education	
High school or below	27 (6.0)
Junior college	61 (13.6)
Bachelor's degree	252 (56.0)
Graduate degree	110 (24.4)
Residence	
Urban	225 (50.0)
Rural	225 (50.0)
Political affiliation	
CPC[a] member	78 (17.3)
Non-CPC member	372 (82.7)
Marital status	
Unmarried	378 (84.0)
Married	63 (14.0)
Divorced	9 (2.0)
Monthly income, CNY (US $ equivalent)	
≤3000 CNY (≤US $425)	269 (59.8)
3000 - 6000 CNY (US $425-$851)	93 (20.7)
6000 - 8000 CNY (US $851-$1135)	45 (10.0)
8000 - 10,000 CNY (US $1135-$1418)	22 (4.9)
≥10,000 CNY (≥US $1418)	21 (4.7)

[a] CPC: Communist Party of China.

Measurements

In this study, we used measures drawn from previous research and previously validated instruments. A standard 5-point Likert-type response format was used, ranging from strongly disagree (1) to strongly agree (5). To ensure linguistic and conceptual equivalence in the Chinese context, a translation and back-translation procedure was conducted. First, the original English items were translated into Chinese by a bilingual researcher. Then, an independent bilingual expert back-translated the Chinese version into English. Discrepancies between the original and back-translated versions were discussed and resolved to ensure semantic consistency. A pilot test was also conducted to assess the clarity and cultural appropriateness of the translated items, and minor revisions were made accordingly.

Online social interaction was measured using 5 items adapted from established online interaction scales that have demonstrated good reliability in previous digital communication research

XSL•FO
RenderX

[17,23]. We adopted a 10-item scale adapted from previously validated measures to assess online social capital, including bridging social capital (5 items) and bonding social capital (5 items) [23,24]. A 3-item scale measuring behavioral intentions to donate was adapted from Thomas and Jadeja [65], which has been used to assess charitable giving intentions. Neighborhood perceptions were measured using 7 items adapted from a well-validated scale assessing residents' perceptions of neighborhood trust, safety, and physical conditions [66]. Finally, demographic variables, including sex, age, income, and census register, were included as control variables based on prior research [67].

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of Guizhou Normal University (approval number 202407001). All participants were informed about the purpose and procedures of the study, and participation was voluntary. Informed consent was obtained from all participants prior to data collection. Given the minimal-risk nature of the study, verbal consent was considered appropriate. All data were collected anonymously and used solely for research purposes.

Results

Data Normality and Common Method Bias Analysis

Data normality was assessed using skewness (−0.876 to 1.311) and kurtosis (−0.810 to 1.130). According to Kline [68], these values indicate acceptable normality. To examine the potential impact of common method bias (CMB), we first conducted Harman's single-factor test following Anderson and Gerbing [69]. The results revealed that 5 factors with eigenvalues greater than one were extracted, and the first factor accounted for 38.55% of the total variance, suggesting that CMB was unlikely to be a serious concern [70]. Given that Harman's single-factor test has been widely criticized as an insufficient diagnostic technique, we further applied a latent method factor approach following Podsakoff et al [70]. Specifically, a common latent factor (CLF) was incorporated into the measurement model, with all observed indicators loading onto both their theoretical constructs and the latent method factor, while constraining the method factor to be uncorrelated with the substantive constructs. Minor item refinements were applied to ensure model stability. The results indicated that the model fit remained acceptable after including the CLF (χ^2_{257}=785.675, Comparative Fit Index=0.938, Tucker-Lewis Index=0.927, Root Mean Square Error of Approximation=0.068, Standardized Root Mean Square Residual=0.089). Moreover, the inclusion of the CLF did not substantially alter the magnitude or significance of the factor loadings, suggesting that CMB is unlikely to pose a serious threat to the validity of the findings.

Confirmatory Factor Analysis

Before testing the hypotheses, we conducted a series of confirmatory factor analyses with maximum-likelihood estimation in Mplus 7.0 (Muthén & Muthén) to examine the distinctiveness of our study's constructs [71]. Table 2 shows that the 5-factor measurement model demonstrated an acceptable fit for the data (χ^2_{258}=803.47, Comparative Fit Index=0.94, Tucker-Lewis Index=0.93, Root Mean Square Error of Approximation=0.07, and Standardized Root Mean Square Residual=0.08). All factor loadings were significant. The 5-factor model demonstrated a significantly better fit than alternative models, thereby supporting the discriminant validity of the constructs [72].

XSL•FO
RenderX

Table . Confirmatory factor analysis results (N=450).

Model	Chi-square (*df*)	*P* value	CFI[a]	TLI[b]	RMSEA[c]	SRMR[d]
5-factor model: OSI[e], BSC[f], OSC[g], ITD[h], EP[i]	803.47 (258)	<.001	0.94	0.93	0.07	0.08
4-factor model 1: OSI, BSC+OSC, ITD, EP	1808.82 (269)	<.001	0.82	0.80	0.11	0.11
4-factor model 2: OSI, BSC, OSC, ITD+EP	2509.84 (269)	<.001	0.74	0.71	0.14	0.12
4-factor model 3: OSI+EP, BSC, OSC, ITD	2794.32 (269)	<.001	0.70	0.67	0.14	0.14
3-factor model 1: OSI+EP+ITD, BSC, OSC	3512.76 (272)	<.001	0.62	0.58	0.16	0.14
3-factor model 2: OSI+BSC+OSC, EP, ITD	2561.12 (272)	<.001	0.73	0.70	0.14	0.12
2-factor model: OSI+BSC+OSC, ITD+EP	3594.23 (274)	<.001	0.61	0.57	0.16	0.13
1-factor model: OSI+BSC+OSC+ITD+EP	4597.33 (275)	<.001	0.49	0.44	0.19	0.14

[a]CFI: Comparative Fit Index.

[b]TLI: Tucker-Lewis Index.

[c]RMSEA: Root Mean Square Error of Approximation.

[d]SRMR: Standardized Root Mean Square Residual.

[e]OSI: online social interaction.

[f]BSC: bridging social capital.

[g]OSC: bonding social capital.

[h]ITD: intention to donate.

[i]EP: neighborhood perception.

Descriptive Statistics

All items are reported in Table 3. The factor loadings of these constructs were significant (0.53 to 0.97 level) [73]. The constructs had adequate and satisfactory convergent validity, as all the average variance extracted (AVE) values were higher than the threshold of 0.50 [74]. The Cronbach α values of these constructs and composite reliability were higher than 0.86. This illustrates strong internal consistency based on all measures. The means, SDs, reliabilities, zero-order Pearson correlations, and square roots of the AVEs are presented in Table 4. Online social interaction was positively correlated with bridging social capital (r=0.46, P<.01), bonding social capital (r=0.55, P<.01), and intention to donate (r=0.37, P<.01). Both bridging social capital (r=0.54, P<.01) and bonding social capital (r=0.55, P<.01) were positively correlated with intention to donate. Moreover, bridging social capital and bonding social capital were positively correlated (r=0.69, P<0.01), providing preliminary support for our hypotheses. In addition, the square roots of the AVEs for all constructs exceeded their interconstruct correlations, further supporting discriminant validity.

XSL•FO
RenderX

Table . Reliability and validity of the constructs.

Constructs and items	Standardized factor loading	Cronbach α	CR[a]	AVE[b]
Online social interaction		0.86	0.86	0.56
I like comments frequently on the Internet.	0.63			
I reply to comments frequently on the Internet.	0.91			
I participate in discussions frequently on the Internet.	0.93			
I often chat privately with strangers on the Internet.	0.59			
I often leave a message on others' profiles on the Internet.	0.61			
Bridging social capital		0.92	0.92	0.71
Interacting with people in online charity communities makes me interested in things that happen outside of my school or work.	0.78			
Interacting with people in online charity communities makes me want to try new things.	0.90			
Interacting with people in online charity communities makes me curious about other places in the world.	0.88			
Interacting with people in online charity communities makes me interested in what people unlike me are thinking.	0.85			
Interacting with people in online charity communities reminds me that everyone in the world is connected.	0.79			
Bonding social capital		0.87	0.87	0.58
There are several people in online charity communities I trust to help solve my problems.	0.78			
There is someone in the online charity community that I feel comfortable talking to about intimate personal problems.	0.76			
There is someone in online charity communities I can turn to for advice about making very important decisions.	0.84			
When I feel lonely, there are several people in online charity communities I can talk to.	0.77			

XSL•FO
RenderX

Constructs and items	Standardized factor loading	Cronbach α	CR[a]	AVE[b]
If I needed an emergency loan of 4000 RMB, I know someone in the online charity community I can turn to.	0.64			
Neighborhood perception		0.87	0.87	0.51
I think my neighborhood is safe during the day.	0.58			
I think my neighborhood is safe at night.	0.60			
I could call my neighbors for help.	0.53			
My neighbors and I trust each other.	0.63			
The buildings and roads in my neighborhood are well-maintained.	0.84			
I feel good about my home or neighborhood.	0.88			
My neighborhood is kept clean.	0.84			
Intention to donate		0.95	0.95	0.87
When I see crowdfunding online platforms for seriously ill patients, I would like to donate money to them through online charitable organizations.	0.90			
When I see crowdfunding online platforms for seriously ill patients, I intend to donate money to them through online charitable organizations.	0.97			
When I see crowdfunding online platforms for seriously ill patients, I will donate money to them through online charitable organizations.	0.92			

[a]CR: composite reliability.

[b]AVE: average variance extracted.

XSL•FO
RenderX

Table . Correlations, means, and SDs of all the study variables (N=450).

	Sex	Age	Income	Education	Census register	Political identity	Marital status	OSI[a]	BSC[b]	OSC[c]	ITD[d]	EP[e]
Sex	—[f]											
Age	−0.02	—										
Income	−0.09	0.39[g]	—									
Education	−0.05	−0.06	−0.08	—								
Census register	−0.01	−0.23[g]	−0.07	−0.01	—							
Political identity	0.02	−0.20[g]	−0.13[g]	−0.22[g]	−0.01	—						
Marital status	0.05	0.59[g]	0.33[g]	−0.26[g]	−0.16[g]	−0.05	—					
OSI[a]	−0.01	−0.05	−0.02	−0.14[g]	−0.05	0.01	0.02	0.75[i]				
BSC[b]	0.05	−0.04	−0.04	−0.08	−0.06	−0.04	−0.04	0.46[g]	0.84			
OSC[c]	−0.06	−0.04	−0.08	−0.14[g]	−0.02	0.02	−0.02	0.55[g]	0.69[g]	0.76		
ITD[d]	0.09	−0.01	−0.03	−0.12[h]	0.01	−0.01	−0.01	0.37[g]	0.54[g]	0.55[g]	0.93	
EP[e]	−0.11[h]	0.07	0.08	−0.05	−0.05	−0.02	0.01	0.27[g]	0.26[g]	0.40[g]	0.44[g]	0.71
Mean (SD)	1.63 (0.48)	1.96 (0.87)	1.74 (1.12)	2.99 (0.79)	1.50 (0.50)	1.83 (0.38)	1.18 (0.43)	2.58 (0.84)	2.99 (0.99)	2.58 (0.95)	3.34 (1.01)	3.50 (0.80)

[a]OSI: online social interaction.

[b]BSC: bridging social capital.

[c]OSC: bonding social capital.

[d]ITD: intention to donate.

[e]EP: neighborhood perception.

[f]Not applicable.

[g]Statistically significant at the .01 level

[h]Statistically significant at the .05 level

[i]Italicized diagonal values represent the square roots of average variance extracted.

Hypotheses Testing

We conducted a regression analysis to test H1. The results indicated that online social interaction was positively correlated with intention to donate (β=0.37, SE 0.05; *P*<.01), supporting H1. We used PROCESS macro (Model 7) to test the hypothesized first-stage moderated mediation model [75]. Specifically, we verified indirect effects using 5000 samples with bias-corrected bootstrap estimates to generate 95% CIs. To ensure our sample size provided sufficient statistical power to detect the hypothesized mediation effects, a power analysis was performed using the MedPower tool (David A. Kenny) [76]. Based on a sample of 450 and the observed path coefficients, the results indicated that the statistical power for detecting the indirect effects in both mediation models was virtually 1.00, well above the recommended 0.80 threshold. H2 and H3 proposed that bridging social capital and bonding social capital mediate the relationship between online social interaction and the intention to donate. As Table 5 shows, online social interaction was positively related to bridging social capital (β=0.45, SE 0.05; *P*<.01) and bonding social capital (β=0.54, SE 0.05; *P*<.01). Moreover, bridging social capital (β=0.48, SE 0.05; *P*<.01) and bonding social capital (β=0.54, SE 0.05; *P*<.01) were positively related to intention to donate. Furthermore, the PROCESS results indicated that bridging social capital partially mediated the relationship between online social interaction and intention to donate, as the direct effect remained significant (β=0.18, SE 0.05; *P*<.01). In contrast, bonding social capital fully mediated this relationship, as the direct effect of online social interaction on intention to donate became non-significant (β=0.11, SE 0.06; *P*=.06). Therefore, both H2 and H3 were supported.

XSL•FO
RenderX

Table . PROCESS results for the overall model (N=450).

Variables	Bridging social capital		Bonding social capital		Intention to donate	
	M1	M2	M3	M4	M5	M6
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
Constant	—[a]	3.47[b] (0.42)	—	3.13[b] (0.36)	1.33[c] (0.45)	1.46[d] (0.45)
Online social interaction	0.45[b] (0.05)	0.44[b] (0.05)	0.54[b] (0.05)	0.48[b] (0.05)	0.18[b] (0.05)	0.11 (0.06)
Bridging social capital	—	—	—	—	0.48[b] (0.05)	—
Bonding social capital	—	—	—	—	—	0.54[b] (0.05)
Moderation variables	—	—	—	—	—	—
Neighborhood perception	—	0.17[b] (0.05)	—	0.31[b] (0.05)	—	—
Interaction Int 1[e]	—	0.17[c] (0.05)	—	0.18[c] (0.05)	—	—
Control variables	—	—	—	—	—	—
Sex	0.06 (0.09)	0.17 (0.08)	−0.07 (0.08)	−0.06 (0.07)	0.13 (0.08)	0.25 (0.08)
Age	0.01 (0.06)	−0.02 (0.06)	0.04 (0.06)	−0.01 (0.05)	0.05 (0.06)	0.03 (0.06)
Income	−0.02 (0.04)	−0.01 (0.04)	−0.08 (0.04)	−0.07 (0.03)	−0.01 (0.04)	0.02 (0.04)
Education	−0.05 (0.06)	−0.06 (0.06)	−0.09 (0.05)	−0.11 (0.05)	−0.08 (0.05)	−0.05 (0.05)
Census register	−0.06 (0.09)	−0.07 (0.08)	−0.01 (0.08)	0.03 (0.07)	0.08 (0.08)	0.04 (0.08)
Political identity	−0.08 (0.12)	−0.15 (0.11)	−0.01 (0.11)	−0.02 (0.10)	0.01 (0.11)	−0.05 (0.11)
Marital status	−0.13 (0.13)	−0.13 (0.12)	−0.05 (0.11)	−0.04 (0.11)	−0.06 (0.12)	−0.10 (0.12)
R^2	0.23	0.26	0.32	0.41	0.32	0.33
F test (*df*)	16.04 (8, 441)[b]	15.59 (10, 439[b]	25.94 (8, 441)[b]	30.03 (10, 439)[b]	23.45 (9, 440)[b]	23.97 (9, 440)[b]

[a]Not applicable.

[b]Statistically significant at the .001 level.

[c]Statistically significant at the .01 level.

[d]Statistically significant at the .05 level

[e]Int 1=online social interaction × neighborhood perception; unstandardized coefficients of PROCESS.

The PROCESS results also provided empirical support for H4 and H5, which proposed that individuals' neighborhood perceptions moderate the relationship between online social interaction, bridging social capital, and bonding social capital. Given that H4 and H5 included interaction, prior to the construction of the interaction term (neighborhood perception × online social interaction), online social interaction and neighborhood perception were mean-centered. As shown in Table 5, the results indicated that the interaction term was positively associated with bridging social capital (β=0.17, SE 0.05; *P*<0.01) and bonding social capital (β=0.18, SE 0.05; *P*<.01). Additionally, a simple slope test revealed that when the level of neighborhood perception was low, online social interaction was positively associated with bridging social capital (β=0.31, SE 0.07; *P*<.01), and this relationship became stronger when neighborhood perception was high (β=0.58, SE 0.06; *P*<.01). Similarly, when neighborhood perception was low, online social interaction was positively associated with bonding social capital (β=0.34, SE 0.07; *P*<0.01), and the relationship was stronger at higher levels of neighborhood perception (β=0.63, SE 0.05; *P*<.01). Thus, H4 and H5 were supported [77] (Figure 2A and B).

XSL•FO
RenderX

Figure 2. Moderating effect of neighborhood perception on the relationship between online social interaction and bridging social capital (A) and bonding social capital (B).

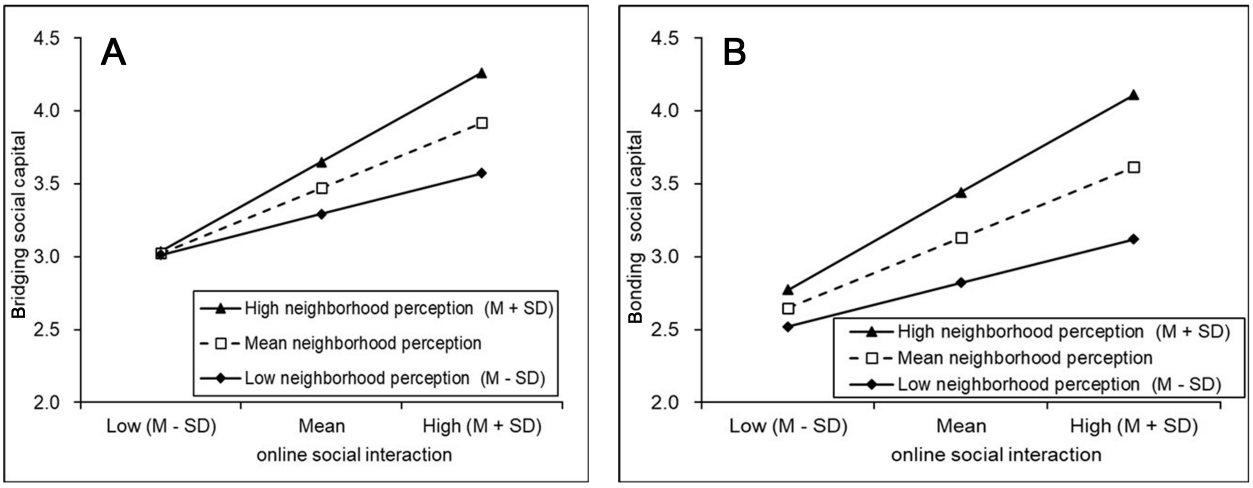

H6 and H7 proposed that individuals' neighborhood perceptions enhance the indirect effects of online social interaction on their intention to donate via bridging and bonding social capital. As shown in Table 6, the conditional indirect effects were significant when neighborhood perception was high (β=0.278, 95% CI 0.195-0.362 for bridging social capital; β=0.340, 95% CI 0.257-0.424 for bonding social capital). In contrast, although still significant, these effects were weaker when neighborhood perception was low (β=0.150, 95% CI 0.059-0.248 for bridging social capital and β=0.181, 95% CI 0.097-0.270 for bonding social capital). Thus, H6 and H7 were supported. Moreover, the moderated mediation index provided further evidence for these moderated mediation effects [78], showing significance for intention to donate via bridging social capital (β=0.080, 95% CI 0.021-0.142) and bonding social capital (β=0.100, 95% CI 0.040-0.162). Notably, the moderated mediation effect was stronger for the bonding social capital pathway than for the bridging pathway.

Table . Conditional indirect effects of online social interaction on intention to donate via online social capital moderated by neighborhood perception (N=450)[a].

Mediation and levels of EP[b] (moderator)	Indirect effect (95% CI)	Index of moderated mediation (95% CI)
Bridging social capital		0.080 (0.021-0.142)
Mean +1 SD	0.278 (0.195-0.362)	
Mean	0.214 (0.142-0.291)	
Mean −1 SD	0.150 (0.059-0.248)	
Bonding social capital		0.100 (0.040-0.162)
Mean +1 SD	0.340 (0.257-0.424)	
Mean	0.260 (0.193-0.331)	
Mean −1 SD	0.181 (0.097-0.270)	

[a]Statistical significance is indicated when the CI does not include zero.

[b] EP: neighborhood perception.

Furthermore, the Johnson-Neyman technique [79] was used to identify regions of significance for the conditional indirect effects. As depicted in Figure 3A and B, these effects were not significant at low levels of neighborhood perception. Given the 5-point Likert scale, the Johnson-Neyman thresholds allow substantive interpretation. For bridging social capital, the indirect effect was significant for neighborhood perception values above 2.04, indicating that its mediating role emerges only from low to moderate levels onward. For bonding social capital, the threshold was lower (1.89), suggesting that its mediating effect is activated even at slightly below-moderate levels. This difference suggests that bonding social capital has a lower activation threshold due to its reliance on close, trust-based ties, whereas bridging social capital requires a more supportive perceived environment. Accordingly, when neighborhood perception is low, online social interaction is less likely to translate into donation intention, while at moderate to higher levels, both forms of social capital become effective pathways.

XSL•FO
RenderX

Figure 3. Conditional indirect effect of online social interaction on intention to donate via bridging social capital (A) and bonding social capital (B) at values of neighborhood perception.

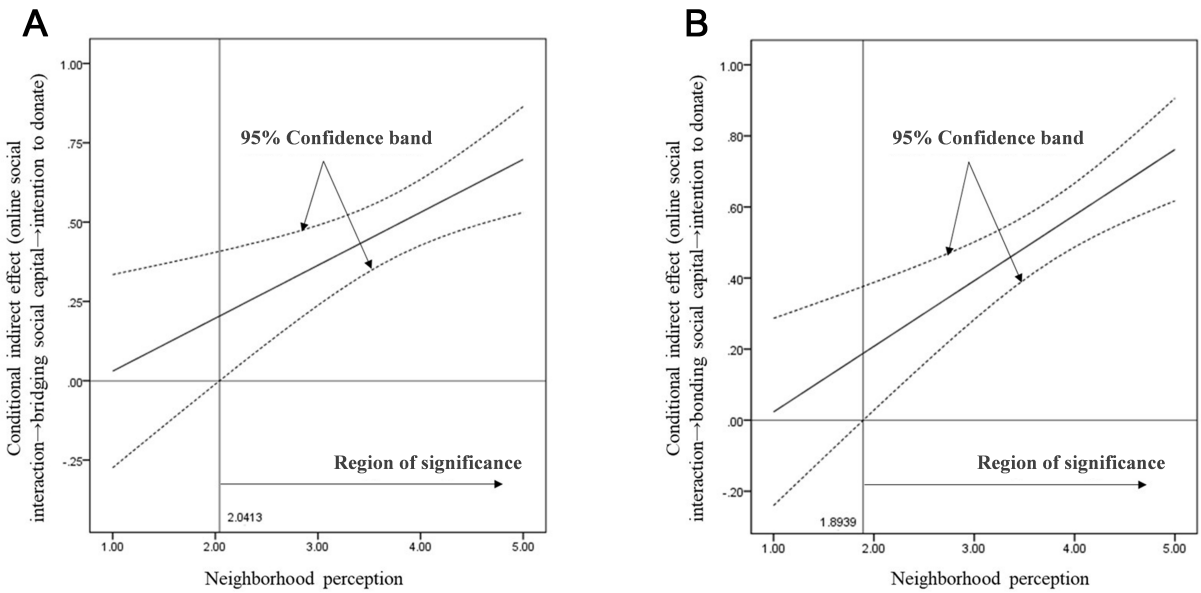

Analysis of Control Variables and Moderation Effects

To enhance the robustness of our findings, we examined the direct effects of control variables (age, income, education, gender, and residency) on the primary constructs. As presented in Table 7, only gender significantly predicted intention to donate (β=0.120, SE 0.038, *P*<.01), whereas other demographic factors remained nonsignificant. Furthermore, we conducted a multigroup SEM analysis to explore the potential moderating effect of gender, the only significant control variable. We compared a configural model (unconstrained) with a structural model where all path coefficients were constrained to be equal across groups. The results of the chi-square difference test (Table 8) indicated that the model fit did not significantly decrease after applying these constraints ($\Delta\chi^2_4$=8.523, *P*=.07). This suggests that the relationship between the independent and dependent variables is stable and does not vary significantly by gender, confirming the generalizability of our hypothesized model.

Table . Direct effects of control variables on ITD[a] (*P*<.01).

Control variables	β (SE)	*z* value	*P* value
Age	0.023 (0.041)	0.547	.58
Monthly income	−0.003 (0.041)	−0.072	.94
Education	−0.038 (0.038)	−0.995	.32
Gender	0.120 (0.038)	3.187	.001
Residence	0.037 (0.038)	0.983	.33

[a]ITD: intention to donate.

Table . Multigroup SEM invariance test for gender.

Model	Chi-square (*df*)	Δ Chi-square [a](Δ*df*)	*P* value
Unconstrained model	1620.233 (770)	__[b]	—
Constrained model	1628.756 (774)	8.523 (4)	.07

[a]ΔChi-square is the difference in chi-square values between the constrained and unconstrained models.
[b]Not applicable.

XSL•FO
RenderX

Discussion

Principal Findings

The results indicate that online social interaction is significantly associated with donation intentions, operating through both online bridging and bonding social capital. Rather than implying a causal mechanism, these findings suggest that the intention to donate to patients with serious illnesses may involve a dual-path associative process in which both weak and strong ties play a role. This result is consistent with prior studies [12,23], which have highlighted the positive roles of bridging and bonding social capital in shaping prosocial outcomes.

Further, the finding that bonding social capital fully mediates the relationship between online social interaction and donation intention, whereas bridging social capital shows partial mediation, suggests that emotional closeness plays a dominant role in health-related charitable decision-making. In the context of serious illness, individuals are frequently exposed to emotionally salient narratives that emphasize vulnerability, suffering, and hope. Such illness narratives tend to elicit empathy and emotional resonance, which strengthen affective bonds and ultimately motivate prosocial behaviors such as donation [80]. This mechanism can be understood through the relational nature of bonding social capital, which is characterized by emotional intensity, trust, and strong interpersonal ties [17,18]. These affective qualities are particularly salient in health charity contexts, where moral obligation and empathic concern are key drivers of helping behavior. In contrast, bridging social capital primarily facilitates access to heterogeneous information and weak-tie connections, which may increase awareness but lack the emotional depth required to fully translate into donation intention [18,35]. This distinction explains why bonding social capital exerts a full mediating effect, whereas bridging social capital plays a complementary but partial role. Furthermore, this finding aligns with emerging eHealth literature emphasizing the role of affective engagement in digital health environments. Studies on digital peer support and patient-centered applications suggest that emotionally responsive systems can enhance empathy-driven interaction and prosocial engagement [77,79,81]. These findings suggest that digital prosocial behavior in health care volunteer communities is not primarily driven by information exposure alone, but by emotionally embedded and trust-centered relational processes [12,24,82]. In this context, online interaction becomes a form of relational support through which empathy and emotional connection are cultivated, thereby motivating charitable engagement [81,83]. This finding extends existing social capital research by highlighting the emotional and relational embeddedness of digital altruism in health care–related settings. Additionally, the stronger effect of bonding social capital compared with bridging social capital suggests that digitally mediated charitable intentions in health care settings are sustained more by emotionally dense and trust-based relational ties than by weak informational connections [23,24]. This finding further indicates that perceived intimacy and affective solidarity function as important motivational foundations for prosocial engagement in digitally mediated health care communities [81,83]. Overall, these findings suggest that online health care volunteer communities operate not merely as informational networks, but as affective and relational infrastructures that sustain digital prosocial engagement.

This study also demonstrates that neighborhood perception moderates the indirect effect of online social interaction on donation intention through both bridging and bonding social capital. These findings provide a complementary perspective for understanding how neighborhood perception and online interaction jointly shape prosocial engagement in the context of serious illness care charity. Specifically, the 2 pathways appear to reflect distinct yet complementary social mechanisms. The pathway through bridging social capital is primarily associated with norm-based processes, whereby individuals with stronger neighborhood perception become more sensitive to social expectations, reputational concerns, and norms of prosocial behavior [16]. This heightened norm awareness encourages engagement in broader, weak-tie networks and facilitates access to broader social support and informational resources, thereby strengthening donation intention. In contrast, the pathway through bonding social capital is more closely linked to relationship-based processes, in which stronger neighborhood perception enhances interpersonal trust, emotional connection, and a sense of belonging [54]. These relational dynamics foster emotionally supportive and reciprocal ties among donors, volunteers, and patients, thereby reinforcing prosocial motivation and donation intention. Overall, these findings highlight that offline neighborhood perception and online social interaction are not independent influences, but mutually reinforcing relational environments that shape digital prosocial engagement.

Theoretical Contributions

This study contributes to the literature by deepening our understanding of how online interactions and offline contexts jointly shape digital altruistic behavior within health care volunteer communities. This study further expands social capital research, revealing the distinct roles that bridging social capital and bonding social capital play in shaping the willingness to donate [23,24]. Although both forms of social capital are positively associated with donation intention, bonding social capital exerts a stronger influence. This suggests that digitally mediated charitable engagement is based more on emotionally intense, trust-based relationships than on weak informational ties [12,81]. In this sense, this study offers a new perspective on understanding the relational and emotional embeddedness of digital altruistic behavior in health care–related contexts [14,15].

This study also expands the body of research on online interaction, moving beyond the informational perspective of digital communication [11,12]. And this study does not view online interaction merely as the exchange of information, but rather regards active and constructive online participation as a form of relational support that fosters trust, empathy, and emotional connection through this process [15]. This perspective complements existing research on online social capital by emphasizing the relational processes through which digital prosocial engagement becomes socially embedded [14,81].

In addition, this study advances the literature by incorporating neighborhood perception as an important offline contextual

XSL•FO
RenderX

condition shaping the formation of online social capital [23,24]. While prior studies have primarily focused on platform characteristics or online behavioral factors, this study highlights the socio-spatial embeddedness of digital social capital formation. More specifically, the findings suggest that neighborhood perception simultaneously reinforces norm-based processes associated with bridging social capital [16] and relationship-based processes associated with bonding social capital [54]. These findings indicate that digital prosocial engagement is not formed solely within online environments but emerges through the interaction between online relational dynamics and offline social contexts. Overall, this study provides an integrated theoretical perspective on how online interaction, neighborhood perception, and differentiated forms of social capital jointly shape prosocial engagement in digitally mediated health care communities.

Implications for eHealth Tool Development

The identified mechanisms can be translated into actionable strategies for digital platforms. Our findings highlight the critical role of platform affordances in facilitating social capital accumulation. The structure of digital platforms, such as algorithmic content feeds, notification systems, and interactive communication channels, can foster a sense of community and accountability [84]. For instance, algorithm-driven exposure and timely reminders may reinforce weak ties and enhance bridging social capital, while interactive functions such as private messaging and group discussions can strengthen trust and emotional attachment, thereby fostering bonding social capital.

The dual-path mechanism suggests that eHealth platforms should adopt differentiated engagement strategies. Bridging social capital can be promoted through broad information dissemination, open networks, and low-threshold participation. In contrast, bonding social capital may be strengthened through personalized storytelling (eg, patient narratives and volunteer experiences), which can foster empathy and sustained engagement. Prior research has shown that character-driven narratives in digital environments increase narrative transportation and subsequent donation behavior [80].

These findings support the integration of digital “nudges” into eHealth systems. Prior research has shown that peer support, social interaction, and behavior change techniques in mHealth interventions can enhance user engagement and promote sustained participation [84-86]. Accordingly, subtle design interventions, such as personalized recommendations, emotionally engaging prompts, and context-sensitive notifications, can encourage prosocial behaviors with minimal cognitive burden. These approaches are particularly valuable in mobile health environments, where sustaining engagement remains a key challenge.

Furthermore, this study reframes the “3C” Health Care Volunteer Community as a prototype of a digital health intervention, which integrates online interaction, social capital formation, and prosocial health engagement. This perspective highlights the potential of digitally mediated communities as scalable models for fostering peer support and charitable participation in health contexts. As online social interaction and charitable engagement increasingly rely on platform-mediated communication, issues such as data security, user consent, and the responsible use of personal health information become critical, particularly in health-related crowdfunding and online volunteering platforms. Ensuring transparency and trust in data governance is essential for sustaining user engagement and fostering long-term participation in digital health communities.

Practical Implications

Extending these findings to digital health contexts, platform-based interventions can be strategically designed to enhance user engagement and prosocial behavior. To encourage donation intention, community operators can provide multiple channels to stimulate active social interaction, for example, interactions between public accounts and online communication groups. Beyond purely online strategies, greater emphasis should be placed on integrating online platforms with offline community infrastructures. For example, linking neighborhood-based applications with widely used social platforms such as WeChat can facilitate real-time coordination, lower participation barriers, and enable localized health volunteering within familiar social networks.

Offline-online synergies can be further strengthened by embedding digital participation into everyday community health touchpoints, linking offline care events with sustained online interaction, and developing data-driven feedback loops that inform local health governance. In addition, integrating spatially grounded community resources (eg, health centers, therapeutic gardens, and neighborhood activity spaces) with digital platforms can facilitate context-aware volunteering and enhance both bonding and bridging social capital. Given the mediating effect of social capital, importance should be attached to building strong and weak ties across online and offline settings. Active interaction among participants in charitable donations and sustained online contact between volunteers and patients following offline care activities can strengthen emotional connections and reinforce trust. Broader online connections across diverse social groups can facilitate information exchange, diffusion of prosocial norms, and resource mobilization. The interplay between offline encounters and online communication thus creates a reinforcing loop that enhances both bonding and bridging social capital, ultimately promoting sustained engagement in digital health volunteering.

The moderating role of neighborhood perception further highlights the importance of offline community environments in shaping digital prosocial behavior. Safer and more cohesive neighborhoods can foster trust and social interaction, which in turn strengthens both online bridging and bonding social capital. This suggests that improving neighborhood conditions, such as enhancing safety, promoting social cohesion, and encouraging local participation, can indirectly support digital health engagement. Moreover, the roles of schools, businesses, and other venues should also receive increased focus.

Limitations and Future Research Directions

First, this study is based on cross-sectional data, which limits causal inference. Sensitivity analyses (eg, instrumental variables or longitudinal simulations) could strengthen causal

XSL•FO
RenderX

interpretation but require different data structures and were beyond the scope of this study. Therefore, the observed mediation patterns should be interpreted with caution. Future research should use longitudinal or experimental designs and appropriate sensitivity analyses to better address causality and potential endogeneity.

Second, reliance on self-reported measures of donation intention may introduce CMB and social desirability bias. In the context of charitable giving, respondents may overstate their willingness to donate due to normative expectations or perceived social approval. This may lead to an overestimation of the relationships observed in this study. Future research could complement self-reported intentions with behavioral data (eg, actual donation records or platform trace data) or adopt more indirect measurement techniques to reduce such bias. In addition, the dependent variable in this study is donation intention rather than actual donation behavior. Although intention is a theoretically meaningful predictor of behavior, it may not always translate into action. Therefore, the findings should be interpreted with caution when extending to real-world donation behavior.

Third, the sample is relatively homogeneous, mainly comprising young female participants from Guizhou, China, which may limit generalizability to more diverse eHealth populations across different age groups and sociocultural settings. In addition, participants were recruited through health care volunteer-related networks and represented varying levels of engagement with the 3C community, ranging from active volunteers to peripheral supporters and individuals familiar with related charitable initiatives. Although this diversity of engagement provides a meaningful context for examining digitally mediated prosocial behavior, the findings should be interpreted with caution when generalizing to broader populations. Future studies should adopt multisite designs and more diverse samples to enhance external validity.

Future research could use experimental designs to strengthen causal evidence. In particular, A/B testing of online interaction environments (eg, varying emotional framing, peer visibility, social endorsement cues, or interaction intensity on crowdfunding platforms) may help disentangle how different forms of online social interaction influence the formation of social capital and subsequent donation intentions.

Conclusion

Online social interaction was associated with donation intentions toward patients with serious illness via bridging and bonding social capital, with bonding ties fully mediating the effect and bridging ties partially mediating it. Neighborhood perception strengthened these pathways, suggesting that offline neighborhood context can enhance the impact of online engagement on charitable giving. These findings highlight the importance of integrating digital interaction with offline environments in promoting health-related charitable behavior and provide practical implications for designing digital health tools.

Acknowledgments

We would like to thank all participants and volunteers involved in this study.

We used the generative artificial intelligence (AI) tool ChatGPT by OpenAI to assist with language refinement. No generative AI tools were used to generate scientific content, analyses, results, or interpretations.

Funding

This work was supported by the National Natural Science Foundation of China (grant numbers 42461031 and 41861029) and the Science and Technology Support Project of the Guizhou Science and Technology Department (grant number 2022-254).

Data Availability

The datasets analyzed during the current study are available from the corresponding author on reasonable request.

Authors' Contributions

JC contributed to conceptualization, writing—original draft, visualization, and writing—review and editing. JH contributed to conceptualization, investigation, and writing—original draft. YY contributed to conceptualization, methodology, and writing—original draft. FD contributed to conceptualization, supervision, and project administration. XL contributed to validation, investigation, and writing—original draft.

Conflicts of Interest

None declared.

References

1. Eysenbach G. What is e-health? J Med Internet Res 2001;3(2):E20. [doi: 10.2196/jmir.3.2.e20] [Medline: 11720962]
2. Pagliari C, Sloan D, Gregor P, et al. What is eHealth (4): a scoping exercise to map the field. J Med Internet Res 2005 Mar 31;7(1):e9. [doi: 10.2196/jmir.7.1.e9] [Medline: 15829481]

XSL•FO
RenderX

3. Zhang Y, Tan CD, Sun J, Yang Z. Why do people patronize donation-based crowdfunding platforms? An activity perspective of critical success factors. Comput Human Behav 2020 Nov;112:106470. [doi: 10.1016/j.chb.2020.106470] [Medline: 32292239]
4. Cai Y, Kamarudin S, Jiang X, Zhou B. Health inequities in medical crowdfunding: a systematic review. Int J Equity Health 2025 Jun 10;24(1):166. [doi: 10.1186/s12939-025-02543-x] [Medline: 40495201]
5. Cheng C, Gao R. Regional disparities in medical crowdfunding outcomes and their microscopic interpretation [Article in Chinese]. Academic Monthly 2025;57(10):121-131. [doi: 10.19862/j.cnki.xsyk.001175]
6. Ba Z, Zhao YC, Song S, Zhu Q. Understanding the determinants of online medical crowdfunding project success in China. Inf Process Manag 2021 Mar;58(2):102465. [doi: 10.1016/j.ipm.2020.102465]
7. Ba Z, Zhao YC, Zhou L, Song S. Exploring the donation allocation of online charitable crowdfunding based on topical and spatial analysis: evidence from the Tencent GongYi. Inf Process Manag 2020 Nov;57(6):102322. [doi: 10.1016/j.ipm.2020.102322]
8. Han B, Zhang Q. An empirical study on the impact of charitable medical care on residents' health level. Front Public Health 2025;13:1591503. [doi: 10.3389/fpubh.2025.1591503] [Medline: 40520304]
9. Habib MD, Filimonau V, Coşkun A, Wang L, Ermolaev VA. Altruistic and collectivistic values as the antecedents of surplus food donation intention. Socioecon Plann Sci 2023 Dec;90:101729. [doi: 10.1016/j.seps.2023.101729]
10. Hou T, Hou K, Wang X, Luo XR. Why I give money to unknown people? An investigation of online donation and forwarding intention. Electron Commer Res Appl 2021 May;47:101055. [doi: 10.1016/j.elerap.2021.101055]
11. Zhou D, Li Y, Tirasawasdichai T. Do online social interactions cultivate social capital? Evidence from a longitudinal study. Front Psychol 2022;13:989137. [doi: 10.3389/fpsyg.2022.989137] [Medline: 36312069]
12. Zhong ZJ. Civic engagement among educated Chinese youth: the role of SNS (Social Networking Services), bonding and bridging social capital. Comput Educ 2014 Jun;75:263-273. [doi: 10.1016/j.compedu.2014.03.005]
13. Saxton GD, Wang L. The social network effect: the determinants of giving through social media. Nonprofit Volunt Sect Q 2014;43(5):850-868. [doi: 10.1177/0899764013485159]
14. Tucker IM, Goodings L. Digital atmospheres: affective practices of care in Elefriends. Sociol Health Illn 2017 May;39(4):629-642. [doi: 10.1111/1467-9566.12545] [Medline: 28326556]
15. De Coninck S, Emmers E. "Online, the counselor can't see me cry": a systematic literature review on emotion and computer-mediated care. Front Digit Health 2023;5:1216268. [doi: 10.3389/fdgth.2023.1216268] [Medline: 37720162]
16. Schröder JM, Merz EM, Suanet B, Wiepking P. The social contagion of prosocial behaviour: how neighbourhood blood donations influence individual donation behaviour. Health Place 2023 Sep;83:103072. [doi: 10.1016/j.healthplace.2023.103072] [Medline: 37557003]
17. Gerson J, Plagnol AC, Corr PJ. Passive and Active Facebook Use Measure (PAUM): validation and relationship to the reinforcement sensitivity theory. Pers Individ Dif 2017 Oct;117:81-90. [doi: 10.1016/j.paid.2017.05.034]
18. Kim C, Yang SU. Like, comment, and share on Facebook: how each behavior differs from the other. Public Relat Rev 2017 Jun;43(2):441-449. [doi: 10.1016/j.pubrev.2017.02.006]
19. Mazanderani F, Kirkpatrick SF, Ziebland S, Locock L, Powell J. Caring for care: online feedback in the context of public healthcare services. Soc Sci Med 2021 Sep;285:114280. [doi: 10.1016/j.socscimed.2021.114280] [Medline: 34358947]
20. Putnam RD. Bowling Alone: The Collapse and Revival of American Community: Duke University Press; 2000. [doi: 10.1145/358916.361990]
21. Bekkers R. Trust and volunteering: selection or causation? Evidence from a 4 year panel study. Polit Behav 2012 Jun;34(2):225-247. [doi: 10.1007/s11109-011-9165-x]
22. Spottswood EL, Wohn DY. Online social capital: recent trends in research. Curr Opin Psychol 2020 Dec;36:147-152. [doi: 10.1016/j.copsyc.2020.07.031] [Medline: 32950953]
23. Li W, Chen M. How online social interactions predict the sense of virtual community via social capital: testing a dual-process model with an interest–based SNS. Comput Human Behav 2022 Oct;135:107347. [doi: 10.1016/j.chb.2022.107347]
24. Williams D. On and off the 'net: scales for social capital in an online era. J Comput Mediat Commun 2006 Jan;11(2):593-628. [doi: 10.1111/j.1083-6101.2006.00029.x]
25. Song J, Jiang J. Online social capital, offline social capital and health: evidence from China. Health Soc Care Community 2022 Jul;30(4):e1025-e1036. [doi: 10.1111/hsc.13507] [Medline: 34245205]
26. Mpinganjira M. Willingness to reciprocate in virtual health communities: the role of social capital, gratitude and indebtedness. Serv Bus 2019 Jun;13(2):269-287. [doi: 10.1007/s11628-018-0382-9]
27. Wang JC, Chiang MJ. Social interaction and continuance intention in online auctions: a social capital perspective. Decis Support Syst 2009 Nov;47(4):466-476. [doi: 10.1016/j.dss.2009.04.013]
28. Gvili Y, Levy S. Consumer engagement with eWOM on social media: the role of social capital. Online Inf Rev 2018 Aug 21;42(4):482-505. [doi: 10.1108/OIR-05-2017-0158]
29. Yang X, Li G. Factors influencing the popularity of customer-generated content in a company-hosted online co-creation community: a social capital perspective. Comput Human Behav 2016 Nov;64:760-768. [doi: 10.1016/j.chb.2016.08.002] [Medline: 27956757]

XSL•FO
RenderX

30. Zheng C, Cheung AWK, Cronje T. Social capital and bank liquidity hoarding. J Int Financ Mark Inst Money 2022 Sep;80:101653. [doi: 10.1016/j.intfin.2022.101653]
31. Yoo E, Rand W, Eftekhar M, Rabinovich E. Evaluating information diffusion speed and its determinants in social media networks during humanitarian crises. J Oper Manag 2016 Jul;45(1):123-133. [doi: 10.1016/j.jom.2016.05.007]
32. Oh HJ, Ozkaya E, LaRose R. How does online social networking enhance life satisfaction? The relationships among online supportive interaction, affect, perceived social support, sense of community, and life satisfaction. Comput Human Behav 2014 Jan;30:69-78. [doi: 10.1016/j.chb.2013.07.053]
33. Naranjo-Zolotov M, Oliveira T, Cruz-Jesus F, et al. Examining social capital and individual motivators to explain the adoption of online citizen participation. Future Gener Comput Syst 2019 Mar;92:302-311. [doi: 10.1016/j.future.2018.09.044]
34. Ellison NB, Steinfield C, Lampe C. The benefits of Facebook "Friends:" social capital and college students' use of online social network sites. J Comput Mediat Commun 2007 Jul;12(4):1143-1168. [doi: 10.1111/j.1083-6101.2007.00367.x]
35. Obamiro K, West S, Lee S. Like, comment, tag, share: Facebook interactions in health research. Int J Med Inform 2020 May;137:104097. [doi: 10.1016/j.ijmedinf.2020.104097] [Medline: 32092667]
36. Cheng X, Long R, Wu F, Geng J, Yang J. How social interaction shapes habitual and occasional low-carbon consumption behaviors: evidence from ten cities in China. Renew Sustain Energy Rev 2023 Aug;182:113387. [doi: 10.1016/j.rser.2023.113387]
37. Abrahamse W, Steg L. Social influence approaches to encourage resource conservation: a meta-analysis. Glob Environ Change 2013 Dec;23(6):1773-1785. [doi: 10.1016/j.gloenvcha.2013.07.029]
38. Chen H, Li W, Lyu T, Zheng X. Understanding people's participation in online charities: a dual-process approach of trust and empathic concern. Ind Manang Data Syst 2021 Jul 5;121(7):1642-1663. [doi: 10.1108/IMDS-09-2020-0513]
39. Fang X. Can digital finance promote individuals' charitable giving? Empirical evidence from China. Comput Human Behav 2023 Aug;145:107768. [doi: 10.1016/j.chb.2023.107768]
40. Lu J, Yang J, Yu CS. Is social capital effective for online learning? Inf Manag 2013 Nov;50(7):507-522. [doi: 10.1016/j.im.2013.07.009]
41. Argentiero A, Chiarini B, Marzano E. Do social capital and the quality of institutions affect waste recycling? Waste Manag 2023 Jan 1;155:240-251. [doi: 10.1016/j.wasman.2022.11.005] [Medline: 36399851]
42. Katz H, Malul M. Do you enjoy giving more or more than others? The positionality of voluntary contributions. Voluntas 2015 Dec;26(6):2755-2767. [doi: 10.1007/s11266-014-9544-0]
43. Zhai Y. The role of online social capital in the relationship between internet use and self-worth. Curr Psychol 2021 May;40(5):2073-2082. [doi: 10.1007/s12144-019-0149-z]
44. Li S, Clark L, Wheeler C. Unlocking the marketing potential of social capital: a study to identify the dimensions of social capital considered represented within online brand communities. Presented at: 2013 IEEE 10th International Conference on e-Business Engineering; Oct 14-16, 2013; New York, NY Piscataway, NJ p. 138-141. [doi: 10.1109/ICEBE.2013.21]
45. Jing R, Jin G, Guo Y, Zhang Y, Li L. The association between constant and new Internet use and depressive symptoms among older adults in China: the role of structural social capital. Comput Human Behav 2023 Jan;138:107480. [doi: 10.1016/j.chb.2022.107480]
46. Shen C, Chen W. Gamers' confidants: massively multiplayer online game participation and core networks in China. Soc Networks 2015 Jan;40:207-214. [doi: 10.1016/j.socnet.2014.11.001]
47. Vitak J, Ellison NB, Steinfield C. The ties that bond: re-examining the relationship between facebook use and bonding social capital. Presented at: 2011 44th Hawaii International Conference on System Sciences; Jan 4-7, 2011. [doi: 10.1109/HICSS.2011.435]
48. Sun Y, Gao J, Zhang X, Cheng Y. The impact of internet use on residents' happiness in China. Front Public Health 2023;11:1188259. [doi: 10.3389/fpubh.2023.1188259]
49. Colì E, Paciello M, Lamponi E, Calella R, Falcone R. Adolescents and trust in online social interactions: a qualitative exploratory study. Children (Basel) 2023 Aug 18;10(8):1408. [doi: 10.3390/children10081408] [Medline: 37628407]
50. Glanville JL, Paxton P, Wang Y. Social capital and generosity: a multilevel analysis. Nonprofit Volunt Sect Q 2016;45:526-547. [doi: 10.1177/0899764015591366]
51. Chen S, Liu C, Huang J, Zhi H. Can social capital facilitate households' donations in rural China? China Econ Rev 2023 Oct;81:102034. [doi: 10.1016/j.chieco.2023.102034]
52. Cohen EL, Hoffner C. Gifts of giving: the role of empathy and perceived benefits to others and self in young adults' decisions to become organ donors. J Health Psychol 2013 Jan;18(1):128-138. [doi: 10.1177/1359105311433910] [Medline: 22322992]
53. Unger A, Papastamatelou J, Arpagaus J. Do social networks increase donation frequency? The Swiss context. Curr Psychol 2022 Dec;41(12):8885-8896. [doi: 10.1007/s12144-020-01335-6]
54. Laméris J, Hipp JR, Tolsma J. Perceptions as the crucial link? The mediating role of neighborhood perceptions in the relationship between the neighborhood context and neighborhood cohesion. Soc Sci Res 2018 May;72:53-68. [doi: 10.1016/j.ssresearch.2018.02.010] [Medline: 29609745]
55. Letki N. Does diversity erode social cohesion? Social capital and race in British neighbourhoods. Polit Stud (Oxf) 2008 Mar;56(1):99-126. [doi: 10.1111/j.1467-9248.2007.00692.x]
56. Allport GW. The Nature of Prejudice: Addison-Wesley; 1954.

XSL•FO
RenderX

57. Alexander RD. The evolution of social behavior. Annu Rev Ecol Syst 1974 Nov;5(1):325-383. [doi: 10.1146/annurev.es.05.110174.001545]
58. Stephan WG, Ybarra O, Rios K. Intergroup threat theory. In: Nelson TD, editor. Handbook of Prejudice, Stereotyping, and Discrimination: Psychology Press; 2015:255-278. [doi: 10.4324/9780203361993]
59. Stephan WG, Renfro CL, Davis MD. The role of threat in intergroup relations. In: Wagner U, Tropp LR, Finchilescu G, Tredoux C, editors. Improving Intergroup Relations: Building on the Legacy of Thomas F. Pettigrew: Blackwell; 2008:55-72. [doi: 10.1002/9781444303117.ch5]
60. Steinfield C, Ellison NB, Lampe C. Social capital, self-esteem, and use of online social network sites: a longitudinal analysis. J Appl Dev Psychol 2008 Nov;29(6):434-445. [doi: 10.1016/j.appdev.2008.07.002]
61. Schmid K, Al Ramiah AA, Hewstone M. Neighborhood ethnic diversity and trust: the role of intergroup contact and perceived threat. Psychol Sci 2014 Mar;25(3):665-674. [doi: 10.1177/0956797613508956] [Medline: 24434239]
62. Pettigrew TF. Generalized intergroup contact effects on prejudice. Pers Soc Psychol Bull 1997 Feb;23(2):173-185. [doi: 10.1177/0146167297232006]
63. Sutcliffe M, Noble K. Belonging, trust and social isolation: the move on-line during the time of COVID - a longitudinal study. Heliyon 2022 Sep;8(9):e10637. [doi: 10.1016/j.heliyon.2022.e10637] [Medline: 36164518]
64. Veldhuizen EM, Stronks K, Kunst AE. Assessing associations between socio-economic environment and self-reported health in Amsterdam using bespoke environments. PLoS One 2013;8(7):e68790. [doi: 10.1371/journal.pone.0068790] [Medline: 23874765]
65. Thomas S, Jadeja A. Psychological antecedents of consumer trust in CRM campaigns and donation intentions: the moderating role of creativity. J Retail Consum Serv 2021 Jul;61:102589. [doi: 10.1016/j.jretconser.2021.102589]
66. Carbone JT. Neighborhood perceptions and allostatic load: evidence from Midlife in the United States study. Health Place 2020 Jan;61:102263. [doi: 10.1016/j.healthplace.2019.102263] [Medline: 32329729]
67. Chapman CM, Louis WR, Masser BM. Identifying (our) donors: toward a social psychological understanding of charity selection in Australia. Psychol Mark 2018 Dec;35(12):980-989. [doi: 10.1002/mar.21150]
68. Kline RB. Principles and Practice of Structural Equation Modeling, 3rd edition: The Guilford Press; 2010.
69. Anderson JC, Gerbing DW. Structural equation modeling in practice: a review and recommended two-step approach. Psychol Bull 1988;103(3):411-423. [doi: 10.1037/0033-2909.103.3.411]
70. Podsakoff PM, MacKenzie SB, Lee JY, Podsakoff NP. Common method biases in behavioral research: a critical review of the literature and recommended remedies. J Appl Psychol 2003 Oct;88(5):879-903. [doi: 10.1037/0021-9010.88.5.879] [Medline: 14516251]
71. Muthén LK, Muthén BO. Mplus User's Guide: Muthén & Muthén; 2012.
72. Bentler PM, Bonett DG. Significance tests and goodness of fit in the analysis of covariance structures. Psychol Bull 1980;88(3):588-606. [doi: 10.1037/0033-2909.88.3.588]
73. Barclay DW, Thompson RL, Higgins CA. The partial least squares (PLS) approach to causal modeling: personal computer adoption and use as an illustration. Technol Stud 1995;2:285-309 [FREE Full text]
74. Hair JF Jr, Sarstedt M, Hopkins L, Kuppelwieser VG. Partial least squares structural equation modeling (PLS-SEM): an emerging tool in business research. Eur Bus Rev 2014;26:106-121. [doi: 10.1108/EBR-10-2013-0128]
75. Hayes AF. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach, 3rd edition: The Guilford Press; 2022.
76. Kenny DA. MedPower: power and N computations for mediation. shinyapps.io. 2017. URL: https://davidakenny.shinyapps.io/MedPower/ [accessed 2026-04-20]
77. Aiken LS, West SG, Reno RR. Multiple Regression: Testing and Interpreting Interactions: SAGE; 1991.
78. Hayes AF. An index and test of linear moderated mediation. Multivariate Behav Res 2015;50(1):1-22. [doi: 10.1080/00273171.2014.962683] [Medline: 26609740]
79. Preacher KJ, Rucker DD, Hayes AF. Addressing moderated mediation hypotheses: theory, methods, and prescriptions. Multivariate Behav Res 2007;42(1):185-227. [doi: 10.1080/00273170701341316] [Medline: 26821081]
80. Tan SZ, Yang L, Loo WH, Chalke A. Stories that matter: the effects of narrative transportation and influencer-message congruity on prosocial behavior. J Bus Res 2026 Apr;208:116051. [doi: 10.1016/j.jbusres.2026.116051]
81. Haythornthwaite C. Strong, weak, and latent ties and the impact of new media. Inf Soc 2002 Oct;18(5):385-401. [doi: 10.1080/01972240290108195]
82. Inkpen AC, Tsang EWK. Social capital, networks, and knowledge transfer. Acad Manag Rev 2005 Jan;30(1):146-165. [doi: 10.5465/amr.2005.15281445]
83. Kale P, Singh H, Perlmutter H. Learning and protection of proprietary assets in strategic alliances: building relational capital. Strat Mgmt J 2000 Mar;21(3):217-237. [doi: 10.1002/(SICI)1097-0266(200003)21:3<217::AID-SMJ95>3.0.CO;2-Y]
84. Liu X, Liu T, Chen T, Ma R. Efficacy of a digital peer support program on weight management and mental health in university students with preobesity: randomized controlled trial. JMIR Mhealth Uhealth 2026 Jan 21;14:e78960. [doi: 10.2196/78960] [Medline: 41564335]

XSL•FO
RenderX

85. Lewis ZH, Swartz MC, Martinez E, Lyons EJ. Social support patterns of middle-aged and older adults within a physical activity app: secondary mixed method analysis. JMIR Aging 2019 Aug 23;2(2):e12496. [doi: 10.2196/12496] [Medline: 31518281]
86. Kawaguchi K, Nakagomi A, Ide K, Kondo K. Effects of a mobile app to promote social participation on older adults: randomized controlled trial. J Med Internet Res 2024 Sep 30;26:e64196. [doi: 10.2196/64196] [Medline: 39348180]

Abbreviations

3C: Caring for Children with Cancer
AVE: average variance extracted
CLF: common latent factor
CMB: common method bias

Edited by M Balcarras; submitted 09.Oct.2025; peer-reviewed by A Rama, H Wang; revised version received 18.May.2026; accepted 19.May.2026; published 26.Jun.2026.

Please cite as:
Chen J, Hu J, Yang Y, Du F, Li X
Online Social Interaction, Neighborhood Perception, and the Mediating Role of Social Capital in Charitable Giving for Seriously Ill Patients: Cross-Sectional Study
J Med Internet Res 2026;28:e85573
URL: https://www.jmir.org/2026/1/e85573
doi:10.2196/85573

© Jie Chen, Junzhang Hu, Yang Yang, Fangjuan Du, Xuelian Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

What Patients With Asthma Share When No One Listens: Multimethod Observational Study of Patient Narratives on Reddit

Elena Curto-Sánchez[1], MD, PhD; Gabriela Salazar-Palacios[1], MD; Ana Martín-Varillas[1], MD; Estela Cristina Prieto-Maíllo[1], MD; Jacinto Ramos-González[1], MD, PhD; Ignacio Dávila-González[2], MD, PhD; Domingo Palacios-Ceña[3], PhD; Juan Nicolas Cuenca-Zaldivar[4], PhD

[1]Pneumology Service, Complejo Hospitalario de Salamanca, Salamanca, Castilla y León, Spain

[2]Department of Allergy, Institute for Biomedical Research of Salamanca, Complejo Hospitalario de Salamanca, Salamanca, Castilla y León, Spain

[3]Research Group of Humanities and Qualitative Research in Health Science of Universidad Rey Juan Carlos (Hum&QRinHS), Department of Physical Therapy, Occupational Therapy, Physical Medicine and Rehabilitation, Universidad Rey Juan Carlos, Alcorcon, Madrid, Spain

[4]Primary Health Center "El Abajón", Research Group of Pain and Physiotherapy, Madrid Health Service, Las Rozas de Madrid, Madrid, Spain

Corresponding Author:
Domingo Palacios-Ceña, PhD
Research Group of Humanities and Qualitative Research in Health Science of Universidad Rey Juan Carlos (Hum&QRinHS)
Department of Physical Therapy, Occupational Therapy, Physical Medicine and Rehabilitation
Universidad Rey Juan Carlos
Avenida Atenas s/n
Alcorcon, Madrid, 28922
Spain
Phone: 34 914888883 ext 8883
Fax: 34 914888959
Email: domingo.palacios@urjc.es

Abstract

Background: The use of social media platforms, such as Reddit, to seek and share information about disease management and treatment strategies is increasingly common. In the context of asthma—a chronic condition characterized by limiting symptoms and exacerbations that require active patient engagement and adherence to treatment—there is a lack of research describing the content of Reddit posts and the specific topics of interest to patients.

Objective: This study aimed to describe the topics discussed by users on the Reddit asthma forum and identify the sentiments and polarity of the language used in the posts.

Methods: A retrospective observational study of public posts on the asthma subreddit forum (r/Asthma) over a 1-year period (October 2023-October 2024). All posts and related threads were included, subdivided into hot, news, and top, and those voted "up" or "down," those that received "awards," categorized as "golds." The messages were reviewed manually and excluded if they were not related to asthma. A mixed methods analysis was conducted, comprising (1) analysis using text lemmatization, (2) structural topic modeling to identify topics based on word frequency, and (3) sentiment and polarity analysis. This approach aimed to identify the most frequently used topics on Reddit, detect positive and negative sentiments based on the words used, and acceptance or rejection (polarity) based on the language used in the asthma subreddit. Statistical analyses were performed using R software (version 4.1.3; R Foundation for Statistical Computing), with a significance threshold set at $P<.05$.

Results: After removing duplicates, 7806 posts were identified. The suitability of the chosen analysis model was confirmed, as it presented the best balance between exclusivity and semantic coherence. Clusters of 25 topics were identified and distributed according to their weight. The topics with the highest weight were Topic 7 (Symptoms and severity of asthma attacks) and Topic 18 (Causes of asthma). No significant differences were found in the evolution of emerging topics throughout the year except in Topic 20 (Seeking advice from people with asthma; $P=.04$), Topic 21 (Medical tests that should be reviewed periodically; $P=.04$), and Topic 22 (Times of year when attacks occur; $P=.03$). The proportion of feelings and emotions showed a stable trend throughout the year. Discrepancies in feelings and emotions were identified depending on the dictionaries used. Thus, a higher probability of positive feelings was confirmed in the AFINN lexicon. Meanwhile, negative feelings were significant in the Stanford Natural Language Processing, Bing, and National Research Council Canada lexicons.

XSL•FO
RenderX

Conclusions: These results can serve as a guide to identify hidden patient needs and help professionals develop specific interventions on topics relevant to patients.

(J Med Internet Res 2026;28:e77027) doi:10.2196/77027

KEYWORDS

Reddit; subreddits; r/asthma; asthma; social media; health information; online health; online health information; patient education; observational study

Introduction

Asthma is one of the most prevalent chronic respiratory diseases, affecting over 300 million people globally [1]. The intermittent onset of bronchospasm attacks causes symptoms such as wheezing and dyspnea, and is characterized by airway inflammation, hyperresponsiveness, and mucus hypersecretion, leading to airflow obstruction [1]. The clinical manifestation of asthma varies across individuals, with distinct phenotypes [2], influencing treatment of the disease [3].

The introduction of novel clinical biomarkers (such as sputum and blood eosinophils, serum immunoglobulin E (IgE), and fractional exhaled nitric oxide) [2] has facilitated the development of treatments tailored to each patient's individual phenotype [4,5], specifically monoclonal antibodies. However, these drugs are limited to the most severe forms and certain phenotypes [3,4]. Treatment for most patients with mild to moderate forms of the disease is based on the use of inhalers; however, despite the ongoing development of educational programs, significant issues remain in relation to adherence [6] and inhalation technique [7], both of which have critical implications for disease management and the frequency of exacerbations.

Another key aspect in the treatment of asthma is the shared perspective between prescribing physicians and patients diagnosed with asthma on the impact of the disease on daily life [8,9]. Previous studies [8,9] show that there are differences in perspective on the management and control of the disease, resulting in an underestimation of the impact of asthma on the patient's life by the physician, reducing disease and symptom control [8].

One of the responses among patients is to seek help to cope with symptoms, apply strategies, and share their experiences with other patients [10,11]. In this regard, the use of social media for health reasons has increased over the last decade [10,11], having a major impact on patients' understanding of the disease, the acquisition of habits, the decision-making process, and health outcomes [12-18]. In addition, widely shared or viral content significantly influences the adoption—or rejection—of health behaviors [10], largely due to its ability to elicit strong, activating emotions, whether positive or negative [19].

Reddit is a large-scale online platform that hosts more than 130,000 individual communities, known as subreddits. It attracts around 500 million visits per month and engages approximately 73 million unique users each day [20]. With over 270 million active users each month [21], Reddit allows individuals to share personal experiences related to illness and health care. Users frequently post firsthand accounts, respond to others with shared experiences, and exchange advice on symptoms, medical conditions, and treatments [10,22]. The platform allows registered users (known as Redditors) to post content (text, images, and videos) to moderated boards, which are voted on by other users. Posts on Reddit are visible to the broader community, and their prominence within a subreddit is shaped by user voting—either upvotes or downvotes [23]. Subreddits are organized by topic, and users, known as Redditors, can subscribe to those that match their interests, allowing them to curate the content displayed on their personalized front page [24]. Additionally, some subreddits follow a question-and-answer format, with popular examples including r/AskReddit and r/AMA, which features the interactive "Ask Me Anything" style. Discussions on Reddit are public (unless specifically designated as private), allowing for passive data collection. Furthermore, for an internet user seeking information on a specific topic, Reddit is a useful starting point due to its topic-centric organization [10].

To the authors' knowledge, there are no studies that have shown the use of Reddit by patients with asthma and described the topics they consult. The questions that guided this study were: "What were the most consulted and discussed topics by patients with asthma on Reddit?" and "Was the language used on Reddit positive and negative? The objectives of this study were (1) to describe the topics discussed by patients in the asthma subreddit (r/Asthma) and (2) to identify the sentiments and polarity of the language used by patients with asthma on Reddit.

This study may help identify unknown and unmet needs of patients with asthma and improve medical care.

Methods

Design

An observational study was conducted, consisting of a retrospective analysis of public posts on Reddit over a 1-year period, from October 2023 to October 2024. Also, we used the subreddit forum asthma (r/Asthma) [25]. This study adhered to the Reporting Guidelines for Social Networks in Health Research [26], which are incorporated within the EQUATOR (Enhancing the Quality and Transparency of Health Research) initiative [27].

Data Collection

Posts that included general comments were identified and subdivided into hot, news, and top. Similarly, posts that were voted "up" or "down" were identified, as well as those that received "awards," categorized as "golds," and cross-posted posts. r/Asthma posts were obtained with the R library *RedditExtractoR* [28]. This library allows for extracting up to

XSL•FO
RenderX

1000 main posts by each category (hot, top, and new posts) and all generated threads. Only posts in the English language were used to facilitate the analysis.

All posts from the r/Asthma forum and all threads derived from the conversations were included, as r/Asthma is a subreddit forum dedicated specifically to asthma, so the objective of the study is to determine what topics interest the users of this forum without establishing any specific selection criteria a priori.

In addition, the main topic, secondary topics, number of votes obtained, number of responses, and subresponses obtained were included. Also, all information related to asthma, such as symptom management, treatment, follow-up consultations, risk behaviors, healthy habits, professional care, adherence (or nonadherence) to treatment and medical recommendations, education, use of inhalers, etc, was incorporated into the study and analyzed. Messages were reviewed manually and systematically excluded if they were not related to asthma use.

Statistical Analysis

Overview

Statistical analyses were performed using R software (version 4.1.3; R Foundation for Statistical Computing) [29]. The significance level was set at $P<.05$. The Kolmogorov-Smirnov test with Lilliefors correction was used to test the distribution of the variables. The variables were described with mean (SD), median (IQR), or with absolute and relative values, that is, n (%).

In this study, a mixed methods analysis was carried out that included three types of analysis: (1) text lemmatization analysis, (2) structural analysis using models based on topics identified by their frequency of appearance, and (3) sentiment and polarity analysis. This allowed us to identify the most frequently used topics on Reddit, identify positive and negative sentiments based on the words used, and determine acceptance or rejection (polarity) based on the language used to discuss asthma in Reddit posts.

Analysis Using Text Lemmatization

First, the text of the posts was lemmatized for analysis. Lemmatization is a linguistic process that involves identifying the base or lexicon form of a word—known as the lemma—from its inflected variants (eg, plural forms, gendered adjectives, verb conjugations). The lemma is the canonical form under which all inflected versions of a word are grouped. For example, in traditional dictionaries, nouns are typically listed in their singular form, adjectives in the masculine singular, and verbs in the infinitive.

Structural Topic Modeling Analysis

Second, a structural topic model analysis was conducted to examine the emergence of topics across user comments. This method also allows the incorporation of metadata, such as the time period of each comment, as a covariate in the model. The selection of the optimal number of topics was carried out in 2 phases: first, the range of topics with the best values of held-out likelihood and lower limit of stability of convergence between iterations of the models and with the lowest value in the residuals was evaluated; in a second phase, the ratio between semantic coherence and exclusivity between the selected models was analyzed. Exclusivity evaluates whether the main words of the topics also appear as main words of other topics, while semantic coherence shows whether the words most associated with a topic occur equally or not within the documents; in both cases, higher values are better [30-32].

Sentiment and Polarity Analysis

Finally, a sentiment analysis was performed using the Bing [33], AFINN [34], National Research Council Canada (NRC) [35], and Stanford Natural Language Processing (Stanford NLP) [36] dictionaries. All 4 dictionaries are based on unigrams or individual words that assign scores for positive or negative feelings. In addition, the NRC lexicon classifies words into 8 emotional categories of anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. While the Affin lexicon assigns words a score ranging from –5 to 5, with negative values indicating more negative feelings and positive values indicating more positive feelings. Stanford NLP classifies positive into 5 categories: very negative, negative, neutral, positive, and very positive. Similarly, sentence polarity was analyzed using the Bing lexicon, incorporating modifiers such as amplifiers, decrementers, and negators as proposed by Halliday [37]. The integration of these 4 dictionaries facilitates a detailed and comprehensive analysis of the emotions and feelings conveyed in the texts shared by Reddit (r/Asthma forum) users. This methodological approach enables an emotional profiling of the analyzed content, contributing to a deeper understanding of the affective dimensions embedded in user-generated discourse. Furthermore, this combination of dictionaries allows for a multiangle analysis, providing a comprehensive analytical framework that encompasses the analysis of positive and negative emotions, ranging from –5 to 5 with the AFINN dictionary or from very negative to very positive with the Stanford NLP dictionary, as well as assessing 8 more complex emotions with the NRC dictionary. Moreover, the simultaneous use of these 4 dictionaries allows for a cross-validation and complementarity strategy; the combination of results can lead to a more robust and reliable analysis [38,39].

The presence of significant differences in the proportion of the categories of feelings and emotions or the average score of each lexicon and its temporal evolution was analyzed using a regression model. The assumption of linearity between the dependent variables and the quantitative predictors was tested by evaluating the effective degrees of freedom (EDF) with a value greater than 1 as a nonlinearity cut-off point. A generalized additive mixed model (GAMM) or a linear mixed model (LMM) was applied depending on the EDF value. Since these were time series, the correlation structure was modeled by comparing, using a likelihood ratio test (LRT), the models without this structure, with an autoregressive structure with lag 1 (AR1), and with an autoregressive moving average structure (ARMA) evaluated by means of an ARMA analysis of the residuals of the previous AR1 model. In the GAMM models, the fulfillment of the assumptions of v for the smoothed terms was tested, eliminating those with a value greater than 0.8, and the adequacy of the number of basic functions was checked using the K index ($P>.05$ as a selection criterion). The fulfillment of the

XSL•FO
RenderX

assumptions in the residuals was checked with the Kolmogorov-Smirnov tests with Lilliefors correction (normality) and Breusch-Pagan (heteroscedasticity). Since these were not met in the GAMM models, the robust sandwich-type standard errors were calculated in the LMM model with the Stanford NLP lexicon using a bootstrap.

Ethical Considerations

This study does not require approval from the Ethics and Clinical Research Committee of the University Hospital of Salamanca because the information is publicly available. The same considerations have been applied in previous studies that analyzed the use of Reddit by patients to share and search for information on topics related to health and disease [40-42]. In addition, institutional review board and research ethics committee approvals were not required nor sought because the research did not meet the requirements for needing ethical approval per section 1.3 of the European Pharmaceutical Market Research Association guidelines [43]. The ethical principles established by the Declaration of Helsinki were respected. In addition, Reddit offers a comprehensive application programming interface that grants access to much of the platform's information. In strict compliance with the rules and ethical guidelines of each subreddit, we only collect data from subreddits that explicitly allow research activities. Furthermore, to protect user privacy, we only record the comment ID without including any user information or the content of the comment itself, ensuring that no personally identifiable information or data that could lead to reidentification is collected. Furthermore, in the posts selected as illustrative examples for this study, all identifying information, including authorship and publication dates, was removed, and the content was summarized to preserve and emphasize the essential information.

Results

Descriptive

A total of 7806 posts were collected from the r/Asthma thread between October 2023 and October 2024 after removing duplicates, of which 454 generated a total of 6046 response threads. Since Reddit does not allow the extraction of more than 1000 posts, different filters were used to maximize the search for primary comments, from which an unlimited number of response threads could be obtained (Table 1).

XSL•FO
RenderX

Table 1. Reddit r/Asthma posts' characteristics between October 2023 and October 2024.

Post characteristics	2023			2024									
	October	November	December	January	February	March	April	May	June	July	August	September	October
Characteristics of posts													
n	4	69	62	74	69	89	81	76	86	79	85	451	535
Generated comments, mean (SD)	37.75(10.24)	23.03 (19.39)	20.85 (15.28)	23.16 (24.78)	32.83 (31.77)	30.66 (23.66)	26.83 (27.63)	24.22 (19.82)	25.69 (23.05)	20.71 (20.76)	19.44 (26.46)	10.39 (14.26)	10.66 (13.71)
Comments filter, n (%)													
Hot[a]	—[b]	—	—	—	—	—	—	—	—	—	10 (11.8)	368 (81.6)	382 (71.4)
New[c]	—	—	—	—	—	—	—	—	—	—	2 (2.4)	—	—
Top[d]	4 (100)	69 (100)	62 (100)	74 (100)	69 (100)	89 (100)	81 (100)	76 (100)	86 (100)	79 (100)	73 (85.9)	83 (18.4)	153 (28.6)
Characteristics of posts with threads													
n	1	13	12	26	22	28	28	27	18	17	28	132	102
Score[e], mean (SD)	17.00	15.31 (5.89)	9.75 (2.05)	9.38 (3.73)	9.68 (4.75)	10.54 (6.13)	9.93 (5.28)	12.96 (7.60)	11.00 (4.00)	10.47 (3.74)	7.57 (8.16)	2.42 (3.57)	4.06 (5.68)
Upvotes[f], mean (SD)	17.00	15.31 (5.89)	9.75 (2.05)	9.38 (3.73)	9.68 (4.75)	10.54 (6.13)	9.93 (5.28)	12.96 (7.60)	11.00 (4.00)	10.47 (3.74)	7.57 (8.16)	2.42 (3.57)	4.06 (5.68)
Up ratio[g], mean (SD)	0.95	0.92 (0.07)	0.92 (0.07)	0.92 (0.07)	0.91 (0.09)	0.91 (0.06)	0.93 (0.08)	0.92 (0.07)	0.94 (0.07)	0.85 (0.11)	0.87 (0.15)	0.77 (0.24)	0.79 (0.24)
Cross-posts[h], mean (SD)	—	—	—	—	—	—	—	—	—	—	—	0.01 (0.09)	—
Generated comments, mean (SD)	34.00	26.92 (18.83)	19.92 (14.15)	16.38 (12.54)	18.14 (15.79)	19.25 (14.04)	16.07 (11.69)	19.63 (15.80)	16.72 (16.05)	24.88 (24.83)	12.57 (9.80)	8.19 (8.56)	9.29 (14.69)
Thread characteristics													
n	26	339	218	415	372	536	450	503	320	427	359	1099	982
Score, mean (SD)	1.77 (1.24)	2.99 (3.81)	2.56 (3.01)	2.87 (3.60)	2.92 (5.49)	2.25 (3.03)	2.85 (4.23)	2.41 (3.20)	2.83 (4.21)	3.02 (6.12)	2.72 (3.36)	2.19 (3.43)	2.49 (2.94)
Upvotes, mean (SD)	1.77 (1.24)	2.99 (3.81)	2.56 (3.01)	2.87 (3.60)	2.92 (5.49)	2.25 (3.03)	2.85 (4.23)	2.41 (3.20)	2.83 (4.21	3.02 (6.12)	2.72 (3.36)	2.19 (3.43)	2.49 (2.94)

[a]Hot: Comments with the highest popularity rate based on their age and number of positive votes are the most active and popular discussions in real time.

[b]Not available.

[c]New: New comments generated in a given period of time.

[d]Top: Comments with the highest number of net upvotes (upvotes minus downvotes) in a given time period.

[e]Score: Net positive votes (positive votes minus negative votes).

[f]Upvotes: Positive votes.

[g]Up ratio: Quotient of positive and negative votes in percentage.

[h]Cross-posts: Comments linked from other subreddits. Generated comments: comments not written by humans.

Analysis of Thematic Models

By determining the best held-out likelihood values among the models, it was determined that the optimal number of topics was 15 and 25 (Multimedia Appendix 1). Subsequently, the model with 25 topics was found to have the best balance of exclusivity versus semantic coherence, and therefore, the model with 25 topics was adopted (Multimedia Appendix 2). Table 2 shows the 25 topics hypothesized after the development of the model.

XSL•FO
RenderX

Table 2. Reddit r/Asthma posts topics identified.

Topic	Description
Topic 1	Influence of respiratory problems on the ability to walk.
Topic 2	Timing of attacks.
Topic 3	Hopes for rapid improvement.
Topic 4	Adverse effects of Prednisone that may justify stopping its use.
Topic 5	Gratitude for improvement.
Topic 6	Use of inhaled steroids as rescue medication.
Topic 7	Symptoms and severity of asthma attacks.
Topic 8	Medical tests to be monitored.
Topic 9	Influence of respiratory problems on work activity.
Topic 10	Influence of respiratory problems on quality of life in general terms.
Topic 11	Influence of tobacco.
Topic 12	Need for information.
Topic 13	Budesonide spray dosage.
Topic 14	How to access Albuterol inhalers for wheezing without medical insurance.
Topic 15	Hospitalization with oxygen therapy in case of acute exacerbation.
Topic 16	Search for any asthma-related therapy.
Topic 17	Influence of respiratory symptoms such as coughing.
Topic 18	Causes of asthma.
Topic 19	Symptom change throughout the year.
Topic 20	Request for advice to people who have asthma.
Topic 21	Medical tests that should be checked periodically.
Topic 22	Times of the year when attacks occur.
Topic 23	Change in symptoms at times of the year when temperatures are cooler.
Topic 24	Need for specialized medical care.
Topic 25	Gratitude for the assistance received.

Subsequently, each topic was classified according to the weight of its frequency, together with the most frequent words used in each topic (Figure 1 shows the weighted topics).

XSL•FO
RenderX

Figure 1. Weighted topics.

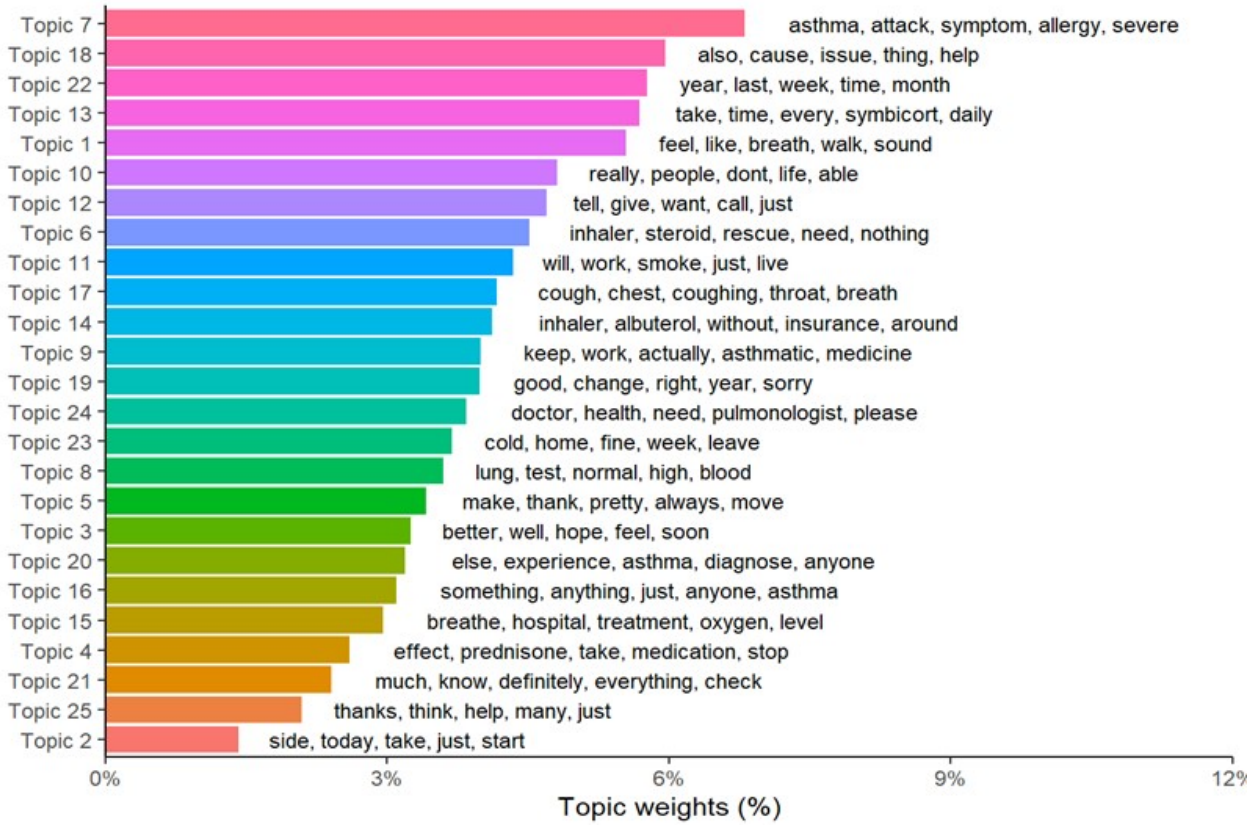

The topics with the greatest weight were Topic 7 (Symptoms and severity of asthma attacks) and Topic 18 (Causes of asthma).

Regarding the annual analysis of the topics, there were no significant differences in the evolution of the emerging topics throughout the year, except in 3 topics. These were Topic 20 (Request for advice from people with asthma; *P*=.04), Topic 21 (Medical tests that should be checked periodically; *P*=.04), and Topic 22 (Times of the year when attacks occur; *P*=.03; Table 3).

XSL•FO
RenderX

Table 3. Evolution of each topic throughout the period of time analyzed.

Topic and time period	Coefficient (SE)	*t* statistic	*P* value[a]
Topic 1			
Intercept	–0.04 (0.29)	–0.13	.89
Date	0 (0)	0.29	.76
Topic 2			
Intercept	–0.04 (0.21)	–0.18	.85
Date	0 (0)	0.30	.76
Topic 3			
Intercept	0.415 (0.21)	1.92	.05
Date	0 (0)	–1.77	.07
Topic 4			
Intercept	0.052 (0.24)	0.21	.82
Date	0 (0)	–0.07	.94
Topic 5			
Intercept	0.225 (0.21)	1.04	.29
Date	0 (0)	–0.89	.37
Topic 6			
Intercept	0.063 (0.26)	0.23	.81
Date	0 (0)	–0.07	.94
Topic 7			
Intercept	0.128 (0.33)	0.38	.70
Date	0 (0)	–0.22	.82
Topic 8			
Intercept	–0.234 (0.27)	–0.85	.39
Date	0 (0)	0.99	.32
Topic 9			
Intercept	0.308 (0.29)	1.05	.29
Date	0 (0)	–0.90	.36
Topic 10			
Intercept	0.142 (0.268)	0.53	.59
Date	0 (0)	–0.35	.72
Topic 11			
Intercept	0.089 (0.24)	0.37	.71
Date	0 (0)	–0.20	.83
Topic 12			
Intercept	0.053 (0.25)	0.20	.83
Date	0 (0)	–0.03	.97
Topic 13			
Intercept	–0.456 (0.30)	–1.50	.13
Date	0 (0)	1.66	.09
Topic 14			
Intercept	–0.452 (0.28)	–1.59	.11
Date	0 (0)	1.74	.08

XSL•FO
RenderX

Topic and time period	Coefficient (SE)	*t* statistic	*P* value[a]
Topic 15			
Intercept	–0.074 (0.31)	–0.23	.81
Date	0 (0)	0.36	.71
Topic 16			
Intercept	–0.432 (0.27)	–1.57	.11
Date	0 (0)	1.71	.08
Topic 17			
Intercept	–0.048 (0.23)	–0.20	.83
Date	0 (0)	0.37	.70
Topic 18			
Intercept	0.223 (0.31)	0.70	.48
Date	0 (0)	–0.53	.59
Topic 19			
Intercept	0.41 (0.24)	1.65	.09
Date	0 (0)	–1.48	.13
Topic 20			
Intercept	–0.505 (0.26)	–1.89	.05
Date	0 (0)	2.04	.041
Topic 21			
Intercept	0.421 (0.19)	2.11	.035
Date	0 (0)	–1.98	.048[b]
Topic 22			
Intercept	0.706 (0.30)	2.31	.021
Date	0 (0)	–2.13	.033
Topic 23			
Intercept	0.167 (0.31)	0.53	.59
Date	0 (0)	–0.39	.69
Topic 24			
Intercept	–0.19 (0.272)	–0.69	.48
Date	0 (0)	0.84	.39
Topic 25			
Intercept	0.067 (0.04)	1.52	.12
Date	0 (0)	–1.26	.20

[a]Significant $P<.05$.

Multimedia Appendix 3 presents examples of texts and narratives shared by users in the Reddit r/Asthma forum, organized by each identified topic.

In addition, a clustering and correlation of all topics was observed, except topics 3, 6, and 13 (Hope of getting better soon; Use of inhaled steroids as rescue medication; and Dosage of Budesonide aerosol; respectively) referring to the use of steroids and Budesonide to accelerate the improvement of symptoms; and topics 17, 20, and 23 (Influence of respiratory symptoms such as cough; Asking for advice from people who have asthma; and Change of symptoms at times of the year when temperatures are lower; respectively) related to asking for advice from those who have already experienced asthma on the management of respiratory symptoms such as cough, especially at times of lower temperatures (Multimedia Appendix 4).

The graphs with the model predictions show how, over the year, the probability of posts containing topics 20 (Request for advice to people with asthma) and 22 (Times of the year when attacks occur) appear in autumn 2024, decreasing over 2024. In the

XSL•FO
RenderX

case of Topic 22, they increase again in the autumn of 2024. Conversely, Topic 21 (Medical tests to be checked periodically), increases especially in the summer-autumn of 2024 (Multimedia Appendix 5).

Analysis of Feelings and Polarity

The proportion of feelings and emotions and scores showed a stable trend throughout the year (Multimedia Appendix 6). In addition, a nonnormal distribution in polarity was confirmed ($P<.001$).

In relation to the contrast of the hypotheses, the significant LRT indicated that the ARMA model is the one that presented the best fit, except in the NRC lexicon of emotions, which presents better results in the AR1 model. Meanwhile, in polarity, the best-fitting model was the model without autocorrelation structure (Multimedia Appendix 7).

Significant differences were confirmed in the proportion of categories appearing in each of the dictionaries used versus the reference category. In contrast, time has no significant effect on the proportion of feelings and emotions or polarity (Multimedia Appendix 8).

Sentiment and emotion analysis produced varying results depending on the lexicons applied. For instance, categorical analysis showed a higher likelihood of positive sentiments and score values of 2 using the AFINN lexicon. In contrast, the Stanford NLP, Bing, and NRC lexicons highlighted significant negative sentiment. Additionally, the NRC lexicon revealed notable associations with the emotions of anticipation, fear, and trust (Multimedia Appendix 9).

The following lines present excerpts from Reddit posts in which both negative and positive emotional expressions are identified. Each example is accompanied by the date of publication. One post describes a near-death experience due to worsening asthma symptoms:

> *I expressed my feelings towards the limitation that asthma brings to me…well, 2 days ago, I almost died…I had an asthma attack after minimal exercise.*

Another post reflects intense frustration caused by the progression of the disease:

> *My asthma has been flaring up more frequently than usual. It's stopped me from seeing friends, working, and made me lose sleep. This is causing me a significant amount of stress and anxiety, too, which I know also makes it worse.*

A third example illustrates how an asthma crisis can complicate air travel:

> *Had an asthma attack at the beginning of a long international flight…it was really scary being trapped in that confined space, not being able to breathe.*

On the other hand, positive experiences were also identified:

> *No doctor wanted to diagnose me as asthmatic, but the only thing that controlled my coughing attacks was corticosteroids…I am grateful to the doctor who agreed to prescribe me.*

Discussion

Principal Findings

The analysis of Reddit's r/Asthma forum has made it possible to identify and classify user-generated topics based on their weight and relevance. The most prominent topics include the causes of asthma and the symptoms and severity of attacks, followed by seasonal patterns of exacerbations and corticosteroid inhaler dosing. Posts display a wide range and intensity of emotions, with a noticeable tendency toward the use of negative language when discussing asthma.

Comparison With Previous Work

The analysis of social networks (including Reddit) and their ability to influence risk behaviors, and the construction of people's opinions, grew during the COVID-19 pandemic [44-46] because of their use as tools for obtaining and cross-checking information, sharing knowledge and as a source of help in dealing with diseases and applying strategies or treatments [47-50]. However, the use of online health information carries significant risks related to the acquisition of reliable knowledge [51-55]. To obtain trustworthy, evidence-based medical or scientific information, users must navigate through a vast amount of misinformation and poor-quality content, including posts that promote unhealthy practices or unproven treatments [52,54-56]. As a result, individuals may make decisions that could potentially harm their health [52-54]. A key factor in acquiring reliable and safe online health information is enhancing the digital literacy and search skills of lay users [57]. Moreover, several recommendations have been proposed to support safe and informed health information searches online, including (1) using trustworthy websites with official domains such as .gov, .edu, or .org; (2) verifying authorship and sources, and prioritizing content reviewed by scientific societies, health organizations, or accredited institutions; (3) avoiding personal testimonials as clinical evidence, since anecdotal narratives do not substitute scientific knowledge; and (4) consulting health care professionals before making any medical decisions [51,54,58,59].

In the field of pulmonology and respiratory diseases, the Reddit social media platform has previously been used to identify unmet needs, sentiments toward areas of interest, shared patient concerns, and the adoption of (unhealthy) behaviors related to e-cigarette use during the COVID-19 pandemic [60], lung cancer [61], vaping use [62], and respiratory lesions [63,64]. However, no previous work from Reddit's analysis has been found for people with asthma. Asthma carries a significant burden on patients' lives, with an important impact on their quality of life and health loss burden, measured by disability-adjusted life years, years of life lost, and years lived with disability [65]. Ten Have et al [66] showed that there are 4 asthma symptoms that, due to their impact, need to be prioritized in their resolution: fatigue, sleep disturbance, impairment of physical activity, and work-related symptoms. Moreover, our results do not show posts with relevant weight on fatigue or sleep disturbances, but on walking ability (topic 1, the fifth by weight, see Figure 1), and on work activity (topic 9, the 12th by weight). Our results revealed posts and comments from patients with Asthma on

XSL•FO
RenderX

Reddit discussing the onset of asthma attacks, changes in symptoms throughout the year, and seasonal variations linked to colder temperatures (topics 2, 19, and 23). This interest aligns with previous studies showing that extreme temperatures—both hot and cold—can trigger asthma attacks and worsen symptoms [67,68].

The strategy for asthma prevention and remission of the European Forum for Research and Education in Allergy and Airway Diseases Consensus Statement [69] adopts a phenotype-centered approach aimed at improving clinical outcomes and preserving health-related quality of life. This strategy moves away from the current "one-size-fits-all" concept, which focuses on symptom-oriented treatment strategies. In addition, there is growing support among experts for including nonpharmacological and interdisciplinary interventions—such as breathing exercises and multicomponent services—in the asthma management "toolbox." These approaches aim to help manage clinical symptoms, improve patients' quality of life, reduce health care costs, and address unmet needs [70]. This would help to decrease the discordance between compliance rates reported by patients with asthma and physicians [71].

Additionally, another aim would be to avoid loss of adherence to treatments. Thus, Amin et al [72] described how denial of the disease and of the need for long-term treatment, along with poor physician-patient communication, suboptimal knowledge of asthma medication (lack of understanding of the distinction between maintenance and reliever inhalers), suboptimal inhaler technique, and the high cost of asthma medication, were key factors in poor medication adherence. Access to pharmacological treatment, its correct use, identification of adverse effects, and use of other therapies appear in our results in six topics (Topics 4, 6, and 13-16), with special consideration of inhaled treatments as a means of administration.

At this point, patient participation is key to treatment adherence and adoption of healthy behaviors. Kang et al [73], in their systematic review and meta-synthesis of qualitative studies on shared decision-making, highlight that patients with asthma engage in shared decision-making when they have the opportunity (ie, environmental factors such as social, cultural, and institutional conditions that either facilitate or hinder individual behavior). The capability (skills, knowledge, and abilities required for individuals to engage in a particular behavior), and the motivation (internal psychological factors that drive individuals to choose specific behaviors based on their motivations and goals. The authors of this study believe that Reddit could be a tool used by patients with asthma to gain more capacity and opportunity to manage the disease, its symptoms, and treatments.

Another possible explanation for the use of Reddit as a source of information, obtaining medications and guidelines for the use of pharmaceutical products and drugs, may be related to the price of health care provided by professionals (doctors) and the cost of drugs [72,74,75]. Zhang et al [65], in their study on health loss and the economic burden of asthma in China, described an annual direct cost of between US $348 and US $1187 per capita, indirect costs of US $7 to US $1195, and hospitalization costs of US $177 to US $1547, influenced by the frequency and severity of acute exacerbations, comorbidities, and treatment adherence. Each country has different health systems with their own models of resource organization (public vs private), which could influence access to health care, medical information and recommendations, and pharmacological treatments for people with asthma [75].

Our results revealed a range of emotions and sentiments, with notable differences between the positive sentiment scores found using the AFINN lexicon and the negative sentiments identified through the Stanford NLP, Bing, and NRC lexicons. This variability in both positive and negative emotions also appears in the study by Volpato et al [74], where patients with uncontrolled asthma experienced a wide range of emotions that impacted their daily lives. Fear and anxiety appeared alongside the unpredictability of asthma attacks, while frustration and hopelessness accompanied the symptoms [74].

Limitations and Strengths

Among the limitations of this study is the inability to confirm an asthma diagnosis in the users who participated in the r/Asthma forum on Reddit, as well as the inability to determine other sociodemographic data of the forum participants [40]. However, the anonymous and private nature of social media would allow users to share questions, doubts, embarrassing topics, and make comments without restrictions that they would not make in other contexts [10]. Another limitation is that there was no uniformity in the comments, with some posts providing more information and content than others. This limitation was resolved by including all possible response threads derived from the posts in the study. The third limitation is that no keywords or combinations were used to search for and include posts in this study. In our study, since there were no previous studies on Reddit content related to asthma, all posts and threads generated during the study year were included. Finally, the reliability of the information shared on Reddit and other online social media platforms is essential for fostering trust in such content, promoting healthy behaviors, and protecting users' health [54]. In this initial study conducted within the Reddit r/Asthma forum, we presented the thematic content and types of information shared by users, as well as the emotional expressions and polarity of their language. Based on this preliminary work, further research would be justified to investigate the credibility and reliability of the information shared in thematic posts, and to explore potential relationships or associations between specific emotional states and the quality of the information provided.

In relation to strengths, social media analysis (Reddit) can provide key information compared to other data collection methods, such as traditional surveys, without being limited by issues such as sensitivity or taboos surrounding certain topics [10,40]. Conversely, among all available online social media platforms, only Reddit was used due to its structure based on thematic forums or communities (subreddits), where users share content and engage in discussions focused on specific topics or interests. This structure contrasts with other platforms, such as Instagram or X, which are primarily oriented toward following individuals. Moreover, content visibility on Reddit is determined

XSL•FO
RenderX

by user evaluations—either positive or negative—rather than by algorithms that prioritize virality or engagement metrics. Reddit also enables extensive discussions without restrictions on character count, image format, or source type, unlike other platforms that impose limitations based on character length (X), visual format (Instagram), or social network boundaries (Facebook). This would allow for the collection of valuable insights into users' motivations and the nature of the content they share. Furthermore, another strength of Reddit analysis for understanding patient feelings and behaviors is that large volumes of data can be obtained in a short period of time, and user-generated content can be processed quickly [76,77]. In addition, social media generates and shares a large amount of varied information on health and disease topics, allowing patients and their families to learn about different aspects of health care, such as diagnostic processes, treatments, coping strategies, professional advice, etc.

Conclusions

These findings can help uncover hidden needs and challenges faced daily by forum users and may assist health care professionals in developing targeted interventions aimed at improving knowledge around key topics. Identifying the most frequently shared thematic content (weighted topics) is relevant for health care professionals, as it may not reflect their clinical priorities. It also enables the assessment of information quality and the detection of potentially risky behaviors that are not supported by scientific evidence. In future research, it would be necessary to examine the scientific quality of the information shared in posts within Reddit's r/Asthma forum and its relationship to emotional state when sharing posts.

Acknowledgments

This study would not have been possible without the support of the Sociedad Castellanoleonesa y Cántabra de Patología Respiratoria (Spanish and Cantabrian Society of Respiratory Pathology).

Funding

This research was funded by Sociedad Castellanoleonesa y Cántabra de Patología Respiratoria (Spanish and Cantabrian Society of Respiratory Pathology). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated and analyzed during this study is not publicly available; although Reddit posts are publicly available and anonymized by username, Redditors did not consent for their content to be shared publicly in a curated dataset. The dataset is available from the corresponding author on reasonable request for research purposes.

Authors' Contributions

Conceptualization: EC-S, GS-P, AM-V, ECP-M, JR-G, ID-G
Data curation: JNC-Z (lead), DP-C
Formal analysis: JNC-Z
Funding acquisition: EC-S
Investigation: JNC-Z (lead), DP-C (equal), EC-S, GS-P, AM-V, ECP-M, JR-G, ID-G
Methodology: DP-C (lead), JNC-Z (equal), EC-S (supporting)
Project administration: JNC-Z (lead), DP-C (equal), EC-S (supporting)
Resources: EC-S
Supervision: EC-S
Validation: EC-S
Visualization: GS-P
Writing – original draft: DP-C (lead), JNC-Z (equal), EC-S (supporting).
Writing – review & editing: DP-C (lead), JNC-Z (equal), EC-S (supporting), GS-P (supporting), AM-V (supporting), ECP-M (supporting), JR-G (supporting), ID-G (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Diagnostic model plots by topics number.
[PNG File , 229 KB - jmir_v28i1e77027_app1.png]

Multimedia Appendix 2

XSL•FO
RenderX

Best selected topic models exclusivity versus semantic coherence graph.
[PNG File , 395 KB - jmir_v28i1e77027_app2.png]

Multimedia Appendix 3
Examples of texts and narratives from the Reddit asthma forum by topic. Each post includes the date it was posted.
[DOCX File , 22 KB - jmir_v28i1e77027_app3.docx]

Multimedia Appendix 4
Topics correlation. Blue line thickness is proportional to correlational value.
[PNG File , 171 KB - jmir_v28i1e77027_app4.png]

Multimedia Appendix 5
Significant topics time-change predicted values between October 2023 and October 2024.
[PNG File , 224 KB - jmir_v28i1e77027_app5.png]

Multimedia Appendix 6
Reddit r/Asthma posts descriptive sentiments and emotions by dictionary and month.
[DOCX File , 24 KB - jmir_v28i1e77027_app6.docx]

Multimedia Appendix 7
Likelihood ratio test for sentiments and emotions by dictionary time series model comparison (ap values).
[DOCX File , 17 KB - jmir_v28i1e77027_app7.docx]

Multimedia Appendix 8
Final Reddit r/Asthma posts sentiments and emotions by dictionary time series models.
[DOCX File , 23 KB - jmir_v28i1e77027_app8.docx]

Multimedia Appendix 9
Significant sentiments and emotions category overall predicted values by dictionary.
[PNG File , 194 KB - jmir_v28i1e77027_app9.png]

References

1. Porsbjerg C, Melén E, Lehtimäki L, Shaw D. Asthma. Lancet 2023;401(10379):858-873. [doi: 10.1016/S0140-6736(22)02125-0] [Medline: 36682372]
2. Schoettler N, Strek ME. Recent advances in severe asthma: from phenotypes to personalized medicine. Chest 2020;157(3):516-528 [FREE Full text] [doi: 10.1016/j.chest.2019.10.009] [Medline: 31678077]
3. Nolasco S, Crimi C, Campisi R. Personalized medicine in asthma: current approach and future perspectives. J Pers Med 2023;13(10):1459 [FREE Full text] [doi: 10.3390/jpm13101459] [Medline: 37888070]
4. Mosnaim G. Asthma in adults. N Engl J Med 2023;389(11):1023-1031. [doi: 10.1056/NEJMcp2304871] [Medline: 37703556]
5. Gonzalez-Uribe V, Romero-Tapia SJ, Castro-Rodriguez JA. Asthma phenotypes in the era of personalized medicine. J Clin Med 2023;12(19):6207 [FREE Full text] [doi: 10.3390/jcm12196207] [Medline: 37834850]
6. Pollard S, Bansback N, FitzGerld JM, Bryan S. The burden of nonadherence among adults with asthma: a role for shared decision-making. Allergy 2017;72(5):705-712. [doi: 10.1111/all.13090] [Medline: 27873330]
7. Sanchis J, Gich I, Pedersen S, Aerosol Drug Management Improvement Team (ADMIT). Systematic review of errors in inhaler use: has patient technique improved over time? Chest 2016;150(2):394-406 [FREE Full text] [doi: 10.1016/j.chest.2016.03.041] [Medline: 27060726]
8. Crespo-Lessmann A, Plaza V, González-Barcala FJ, Fernández-Sánchez T, Sastre J. Concordance of opinions between patients and physicians and their relationship with symptomatic control and future risk in patients with moderate-severe asthma. BMJ Open Respir Res 2017;4(1):e000189 [FREE Full text] [doi: 10.1136/bmjresp-2017-000189] [Medline: 29018525]
9. Matsunaga K, Hamada K, Oishi K, Yano M, Yamaji Y, Hirano T. Factors associated with physician-patient discordance in the perception of asthma control. J Allergy Clin Immunol Pract 2019;7(8):2634-2641. [doi: 10.1016/j.jaip.2019.04.046] [Medline: 31100555]
10. Chan GJ, Fung M, Warrington J, Nowak SA. Understanding health-related discussions on reddit: development of a topic assignment method and exploratory analysis. JMIR Form Res 2025;9:e55309 [FREE Full text] [doi: 10.2196/55309] [Medline: 39879094]

XSL•FO
RenderX

11. Zhao X, Yang V, Menta A, Blum J, Ranasinghe P. Exploring the use of social media for medical problem solving by analyzing the subreddit r/medical_advice: quantitative analysis. JMIR Infodemiology 2025;5:e56116 [FREE Full text] [doi: 10.2196/56116] [Medline: 40112288]
12. Sun Y, Zhang Y, Gwizdka J, Trace CB. Consumer evaluation of the quality of online health information: systematic literature review of relevant criteria and indicators. J Med Internet Res 2019;21(5):e12522 [FREE Full text] [doi: 10.2196/12522] [Medline: 31045507]
13. Sillence E, Briggs P, Harris PR, Fishwick L. How do patients evaluate and make use of online health information? Soc Sci Med 2007;64(9):1853-1862. [doi: 10.1016/j.socscimed.2007.01.012] [Medline: 17328998]
14. Pawlus Z, Spyra A, Polak K, Miziołek B, Bergler-Czop B. Can social media encourage young Polish adults to visit a dermatologist? An original study. Postepy Dermatol Alergol 2025;42(1):68-74 [FREE Full text] [doi: 10.5114/ada.2025.147554] [Medline: 40114760]
15. Munjita SM. Understanding vaccine hesitancy: insights from social media on polio, human papilloma virus, and COVID-19 in Zambia. Digit Health 2025;11:20552076251326131 [FREE Full text] [doi: 10.1177/20552076251326131] [Medline: 40109405]
16. Atarere J, Annor E, Bilalaga MM, Egbo O, Gaddipati GN, Vasireddy R, et al. Social media use and the relationship with colorectal cancer screening among foreign-born populations in the United States. Cancer Causes Control 2025;36(8):845-851. [doi: 10.1007/s10552-025-01985-6] [Medline: 40100525]
17. Siddiqui ZA, Pathan M, Nduaguba S, LeMasters T, Scott VG, Sambamoorthi U, et al. Leveraging social media data to study disease and treatment characteristics of Hodgkin's lymphoma using natural language processing methods. PLOS Digit Health 2025;4(3):e0000765. [doi: 10.1371/journal.pdig.0000765] [Medline: 40106471]
18. Guo IJ, Padmita AC, Matsuzaki M, Gittelsohn J, Feeley A, Watson F, et al. The use of social media to promote unhealthy food and beverage consumption among Indonesian children. BMC Nutr 2025;11(1):57. [doi: 10.1186/s40795-025-01040-2] [Medline: 40119452]
19. Berger J, Milkman KL. What makes online content viral? J Market Res 2012;49(2):192-205. [doi: 10.1509/jmr.10.0353]
20. Reddit Inc. United States Securities and Exchange Commission form S-1. URL: https://www.sec.gov/Archives/edgar/data/1713445/000162828024006294/reddits-1q423.htm#i1b9a579e78a34dfa99f7f26daeec195b_88 [accessed 2025-03-23]
21. Foundation Marketing. URL: https://foundationinc.co/lab/reddit-statistics/ [accessed 2025-03-23]
22. Lu Y, Zhang P, Liu J, Li J, Deng S. Health-related hot topic detection in online communities using text clustering. PLoS One 2013;8(2):e56221 [FREE Full text] [doi: 10.1371/journal.pone.0056221] [Medline: 23457530]
23. Reddit. URL: https://redditinc.com/ [accessed 2025-03-23]
24. Reddit rules. URL: https://redditinc.com/policies/reddit-rules [accessed 2025-03-23]
25. Forum r/asthma. Reddit. URL: https://www.reddit.com/r/Asthma/ [accessed 2025-03-23]
26. Luke DA, Tsai E, Carothers BJ, Malone S, Prusaczyk B, Combs TB, et al. Introducing SoNHR-reporting guidelines for social networks in health research. PLoS One 2023;18(12):e0285236 [FREE Full text] [doi: 10.1371/journal.pone.0285236] [Medline: 38096166]
27. Introducing So NHR-reporting guidelines for social networks in health research. Enhancing the quality and transparency of health research (Equator). URL: https://www.equator-network.org/reporting-guidelines/introducing-sonhr-reporting-guidelines-for-social-networks-in-health-research/ [accessed 2025-03-23]
28. Rivera I. RedditExtractor: tools for extracting reddit comments and submissions (version 2.3.0). GitHub. URL: https://github.com/ivan-rivera/RedditExtractoR/ [accessed 2025-10-16]
29. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2024. URL: https://www.R-project.org/ [accessed 2025-12-17]
30. Rajkumar AV, Suresh C, Veni M, Narasimha MM. On finding the natural number of topics with latent dirichlet allocation: some observations. In: Advances in Knowledge Discovery and Data Mining. Heidelberg: Springer; 2010:391-402.
31. Campbell JC, Hindle A, Stroulia E. Latent dirichlet allocation: extracting topics from software engineering data. In: The Art and Science of Analyzing Software Data. Cambridge, MA: Morgan Kaufmann Publishers; 2015:139-159.
32. Gupta A, Aeron S, Agrawal A, Gupta H. Trends in COVID-19 publications: streamlining research using NLP and LDA. Front Digit Health 2021;3:686720 [FREE Full text] [doi: 10.3389/fdgth.2021.686720] [Medline: 34713157]
33. Liu B. Sentiment analysis and subjectivity. In: Handbook of Natural Language Processing. Boca Raton, Florida: Chapman and Hall/CRC Press; 2010:627-666.
34. Finn Årup N. Evaluation of a word list for sentiment analysis in microblogs. 2011 Presented at: Proceedings for "Making Sense of Microposts: Big things come in small packages" (MSM2011), Volume 718 of CEUR Workshop; May 30, 2011; Heraklion, Greece p. 93-98.
35. Mohammad SM, Turney PD. Crowdsourcing a word–emotion association lexicon. Computational Intelligence 2012;29(3):436-465. [doi: 10.1111/j.1467-8640.2012.00460.x]
36. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford coreNLP natural language processing toolkit. Association for Computational Linguistics; 2014 Presented at: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations; December 16, 2025; Baltimore, Maryland p. 55-60. [doi: 10.3115/v1/p14-5010]

XSL•FO
RenderX

37. Halliday MAK, Hasan R. Cohesion in English. Milton Park, UK: Routledge; 1976.
38. Ribeiro FN, Araújo M, Gonçalves P, André Gonçalves M, Benevenuto F. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Sci 2016;5(1):23 [FREE Full text] [doi: 10.1140/epjds/s13688-016-0085-1]
39. Velu SR, Ravi V, Tabianan K. Multi-lexicon classification and valence-based sentiment analysis as features for deep neural stock price prediction. Sci 2023;5(1):8 [FREE Full text] [doi: 10.3390/sci5010008]
40. Shmueli-Scheuer M, Silverman Y, Halperin I, Gepner Y. Analysis of reddit discussions on motivational factors for physical activity: cross-sectional study. J Med Internet Res 2025;27:e54489 [FREE Full text] [doi: 10.2196/54489] [Medline: 39805106]
41. Kramer ML, Polo JM, Kumar N, Mulgirigama A, Benkiran A. Living with and managing uncomplicated urinary tract infection: mixed methods analysis of patient insights from social media. J Med Internet Res 2025;27:e58882 [FREE Full text] [doi: 10.2196/58882] [Medline: 40067345]
42. Joshi A, Kaune DF, Leff P, Fraser E, Lee S, Harrison M, et al. Self-reported side effects associated with selective androgen receptor modulators: social media data analysis. J Med Internet Res 2025;27:e65031 [FREE Full text] [doi: 10.2196/65031] [Medline: 39965201]
43. Code of conduct. EphMRA. URL: https://tinyurl.com/yc7yfrrw [accessed 2025-03-23]
44. Savolainen R. Assessing the credibility of COVID-19 vaccine mis/disinformation in online discussion. J Inf Sci 2023;49(4):1096-1110 [FREE Full text] [doi: 10.1177/01655515211040653] [Medline: 37461399]
45. Doody S. Making sense of a pandemic: reasoning about COVID-19 in the intellectual dark web. Front Sociol 2024;9:1374042 [FREE Full text] [doi: 10.3389/fsoc.2024.1374042] [Medline: 39351293]
46. Chlabicz M, Nabożny A, Koszelew J, Łaguna W, Szpakowicz A, Sowa P, et al. Medical misinformation in polish on the world wide web during the COVID-19 pandemic period: infodemiology study. J Med Internet Res 2024;26:e48130 [FREE Full text] [doi: 10.2196/48130] [Medline: 38551638]
47. Dolatabadi E, Moyano D, Bales M, Spasojevic S, Bhambhoria R, Bhatti J, et al. Using social media to help understand patient-reported health outcomes of post-COVID-19 condition: natural language processing approach. J Med Internet Res 2023;25:e45767 [FREE Full text] [doi: 10.2196/45767] [Medline: 37725432]
48. Blake L, Tucker P, Vanderloo LM. Mothers' perspectives of the barriers and facilitators to reducing young children's screen time during COVID-19: a Reddit content analysis. PLoS One 2024;19(3):e0301089 [FREE Full text] [doi: 10.1371/journal.pone.0301089] [Medline: 38536885]
49. Ménard A, O'Sullivan T, Mulvey M, Belanger C, Fraser S. Perceptions of hospital care for persons with dementia during the COVID-19 pandemic: a social media sentiment analysis. Gerontologist 2024;64(7):gnad155. [doi: 10.1093/geront/gnad155] [Medline: 37943714]
50. Kim S, Warren E, Jahangir T, Al-Garadi M, Guo Y, Yang Y, et al. Characteristics of intimate partner violence and survivor's needs during the covid-19 pandemic: insights from subreddits related to intimate partner violence. J Interpers Violence 2023;38(17-18):9693-9716 [FREE Full text] [doi: 10.1177/08862605231168816] [Medline: 37102576]
51. Collins SE, Lewis DM. Social media made easy: guiding patients to credible online health information and engagement resources. Clin Diabetes 2013;31(3):137-141 [FREE Full text] [doi: 10.2337/diaclin.31.3.137]
52. Khojah M, Sarhan MY. Vaccination uptake is influenced by many cues during health information seeking online. Health Info Libr J 2025. [doi: 10.1111/hir.12564] [Medline: 39780332]
53. Probst Y, Saffioti E, Manche S, Eaton M. Examination of social media nutrition information related to multiple sclerosis: a cross-sectional social network analysis. Public Health Nutr 2025;28(1):e166. [doi: 10.1017/S1368980025100943] [Medline: 40964911]
54. Silburn A. Navigating online health information: empowerment vs. misinformation. Front Digit Health 2025;7:1555290 [FREE Full text] [doi: 10.3389/fdgth.2025.1555290] [Medline: 40778382]
55. Stokes-Parish J. Navigating the credibility of web-based information during the covid-19 pandemic: using mnemonics to empower the public to spot red flags in health information on the internet. J Med Internet Res 2022;24(6):e38269 [FREE Full text] [doi: 10.2196/38269] [Medline: 35649183]
56. De Caro W. Online health information, health literacy and therapeutic compliance: a theoretical framework. Eur J Public Health 2021;31(3) [FREE Full text] [doi: 10.1093/eurpub/ckab165.429]
57. Mohamed H, Kittle E, Nour N, Hamed R, Feeney K, Salsberg J, et al. An integrative systematic review on interventions to improve layperson's ability to identify trustworthy digital health information. PLOS Digit Health 2024;3(10):e0000638. [doi: 10.1371/journal.pdig.0000638] [Medline: 39453891]
58. US Department of Health and Human Services. How to know if health information on social media is accurate. National Institutes of Health. 2025. URL: https://tinyurl.com/2emf9ezu [accessed 2025-10-16]
59. US Department of Health and Human Services. Know the science: finding health information online. National Institutes of Health. 2025. URL: https://www.nccih.nih.gov/health/know-science/finding-and-evaluating-online-resources/finding-health-information-online/introduction [accessed 2025-10-16]

XSL•FO
RenderX

60. Watkins SL, Snodgrass K, Fahrion L, Shaw E. Contextualizing changes in e-cigarette use during the early COVID-19 pandemic and accompanying infodemic ("So Much Contradictory Evidence"): qualitative document analysis of reddit forums. J Med Internet Res 2025;27:e66010 [FREE Full text] [doi: 10.2196/66010] [Medline: 40112286]
61. Shah AM, Lee KY, Hidayat A, Falchook A, Muhammad W. A text analytics approach for mining public discussions in online cancer forum: analysis of multi-intent lung cancer treatment dataset. Int J Med Inform 2024;184:105375. [doi: 10.1016/j.ijmedinf.2024.105375] [Medline: 38367390]
62. Kierstead E, Silver N, Amato M. Examining quitting experiences on quit vaping subreddits from 2015 to 2021: content analysis. J Med Internet Res 2024;26:e52129 [FREE Full text] [doi: 10.2196/52129] [Medline: 39454194]
63. Wu D, Kasson E, Singh AK, Ren Y, Kaiser N, Huang M, et al. Topics and sentiment surrounding vaping on twitter and reddit during the 2019 e-cigarette and vaping use-associated lung injury outbreak: comparative study. J Med Internet Res 2022;24(12):e39460 [FREE Full text] [doi: 10.2196/39460] [Medline: 36512403]
64. Hswen Y, Yom-Tov E. Analysis of a vaping-associated lung injury outbreak through participatory surveillance and archival internet data. Int J Environ Res Public Health 2021;18(15):8203 [FREE Full text] [doi: 10.3390/ijerph18158203] [Medline: 34360495]
65. Zhang P, Xu J, Xu B, Zhang Y, Xie Y. Health loss and economic burden of asthma in China: a qualitative review based on existing literature. Arch Public Health 2025;83(1):28 [FREE Full text] [doi: 10.1186/s13690-025-01515-5] [Medline: 39905572]
66. Ten Have L, Meulmeester FL, de Jong K, Ten Brinke A. Patient-centred outcomes in severe asthma: fatigue, sleep, physical activity and work. Eur Respir Rev 2025;34(175):240122 [FREE Full text] [doi: 10.1183/16000617.0122-2024] [Medline: 40044187]
67. Han A, Deng S, Yu J, Zhang Y, Jalaludin B, Huang C. Asthma triggered by extreme temperatures: from epidemiological evidence to biological plausibility. Environ Res 2023;216(Pt 2):114489 [FREE Full text] [doi: 10.1016/j.envres.2022.114489] [Medline: 36208788]
68. Deng S, Han A, Jin S, Wang S, Zheng J, Jalaludin BB, et al. Effect of extreme temperatures on asthma hospital visits: modification by event characteristics and healthy behaviors. Environ Res 2023;226:115679. [doi: 10.1016/j.envres.2023.115679] [Medline: 36913996]
69. Jesenak M, Bobcakova A, Djukanovic R, Gaga M, Hanania NA, Heaney LG, et al. Promoting prevention and targeting remission of asthma: a EUFOREA consensus statement on raising the bar in asthma care. Chest 2025;167(4):956-974 [FREE Full text] [doi: 10.1016/j.chest.2024.11.035] [Medline: 39672229]
70. Holland AE, Lewis A. Evidence-based management of symptoms in serious respiratory illness: what is in our toolbox? Eur Respir Rev 2024;33(174):24025 [FREE Full text] [doi: 10.1183/16000617.0205-2024] [Medline: 39477357]
71. Al-Moamary M, Aggarwal B, Al-Ahmad M, Sriprasart T, Koenig S, Levy G, et al. Are treatment adherence factors apparent in patients with asthma and to physicians? Results from the APPaRENT 3 Survey. Adv Ther 2025;42(3):1506-1521. [doi: 10.1007/s12325-025-03105-x] [Medline: 39912987]
72. Amin S, Soliman M, McIvor A, Cave A, Cabrera C. Understanding patient perspectives on medication adherence in asthma:A targeted review of qualitative studies. Patient Prefer Adherence 2020;14:541-551 [FREE Full text] [doi: 10.2147/PPA.S234651] [Medline: 32210541]
73. Kang H, Pen Y, He Y, Yang X, Su J, Yang Q, et al. The experience of shared decision-making for people with asthma: A systematic review and metasynthesis of qualitative studies. Health Expect 2024;27(2):e14039 [FREE Full text] [doi: 10.1111/hex.14039] [Medline: 38613765]
74. Volpato E, Pennisi V, Pennisi A, Piraino A, Banfi P, D'Antonio S, et al. Delving into uncontrolled or severe asthma: perspectives from patients and healthcare professionals in a cross-sectional study. J Asthma Allergy 2024;17:1207-1226 [FREE Full text] [doi: 10.2147/JAA.S483020] [Medline: 39610847]
75. Buendia JA, Guerrero-Patino D, Buendia Sanchez JA. Analysis of the economically justifiable price of mepolizumab in adults with asthma in Colombia. J Asthma 2025;62(5):850-860. [doi: 10.1080/02770903.2024.2448736] [Medline: 39836038]
76. Shatz I. Fast, free, and targeted: Reddit as a source for recruiting participants online. Soc Sci Comput Rev 2016;35(4):537-549. [doi: 10.1177/0894439316650163]
77. Watts IVC. There can be benefits to discussing healthcare on social media. BMJ 2024;384:q566. [doi: 10.1136/bmj.q566] [Medline: 38448054]

Abbreviations

AR1: autoregressive structure with lag 1
ARMA: autoregressive moving average structure
EDF: effective degrees of freedom
EQUATOR: Enhancing the Quality and Transparency of Health Research
GAMM: generalized additive mixed model
IgE: immunoglobulin E
LMM: linear mixed model

XSL•FO
RenderX

LRT: likelihood ratio test
NRC: National Research Council Canada
Stanford NLP: Stanford Natural Language Processing

Edited by A Stone; submitted 06.May.2025; peer-reviewed by S Leuckert, M Herrero-Montes, M Chlabicz; comments to author 18.Aug.2025; accepted 14.Nov.2025; published 08.Jan.2026.

Please cite as:
Curto-Sánchez E, Salazar-Palacios G, Martín-Varillas A, Prieto-Maíllo EC, Ramos-González J, Dávila-González I, Palacios-Ceña D, Cuenca-Zaldivar JN
What Patients With Asthma Share When No One Listens: Multimethod Observational Study of Patient Narratives on Reddit
J Med Internet Res 2026;28:e77027
URL: https://www.jmir.org/2026/1/e77027
doi:10.2196/77027
PMID:41505750

©Elena Curto-Sánchez, Gabriela Salazar-Palacios, Ana Martín-Varillas, Estela Cristina Prieto-Maíllo, Jacinto Ramos-González, Ignacio Dávila-González, Domingo Palacios-Ceña, Juan Nicolas Cuenca-Zaldivar. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Moderators' Role in Fostering Health Practitioner Engagement Within an Online Active Learning Program (The Community Fracture Capture Learning Hub): Qualitative Relational Content Thematic Analysis

Ahmed M Fathalla[1], PhD; Shanton Chang[2], PhD; Ralph Audehm[3], MBBS; Cherie Chiang[1,4], MD; Alexandra Gorelik[1,5], MSc; Christopher J Yates[1,4], PhD; Steve Snow[6], BCom, LLB; Rahul D Barmanray[4], PhD; Sarah Price[1,7], PhD; Lucy Collins[8], MD; John D Wark[1,4], PhD

[1]Department of Medicine, The Royal Melbourne Hospital, The University of Melbourne, Melbourne, Victoria, Australia
[2]School of Computing and Information Systems, University of Melbourne, Melbourne, Australia
[3]Department of General Practice and Primary Care, University of Melbourne, Melbourne, Australia
[4]Department of Diabetes & Endocrinology, Royal Melbourne Hospital, Melbourne Health, Melbourne, Australia
[5]School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia
[6]Praxhub, Melbourne, Australia
[7]Department of Obstetric Medicine, Royal Women's Hospital, University of Melbourne, Melbourne, Australia
[8]Department of Medicine, School of Clinical Sciences, Monash University, Melbourne, Australia

Corresponding Author:
Ahmed M Fathalla, PhD
Department of Medicine, The Royal Melbourne Hospital
The University of Melbourne
4th Floor, Clinical Sciences Building
Melbourne, Victoria, 3010
Australia
Phone: 61 383445892
Email: ahmed.elsayed@unimelb.edu.au

Abstract

Background: Virtual communities of practice (VCoPs) offer flexible platforms for interprofessional education among primary care physicians (PCPs); yet, the role of moderators in optimizing engagement remains underexplored. The Community Fracture Capture Learning Hub, a case-based, interactive web-based program addressing osteoporosis management gaps, provides an ideal context to investigate moderators' impact on collaborative learning.

Objective: This study aimed to characterize moderators' roles and strategies in fostering engagement within health-focused VCoPs, using the Community Fracture Capture Learning Hub as a model.

Methods: A qualitative relational content analysis was conducted on discussion-board interactions across four 6-week cycles (May 2022-October 2023) involving 55 PCPs and 8 moderators. Data included anonymized forum comments of moderators throughout the 4 cycles. Inductive-deductive thematic analysis identified patterns in moderator-participant and moderator-moderator exchanges, focusing on engagement techniques, content facilitation, and interaction dynamics.

Results: Five interconnected themes were identified; (1) participation encouragement: moderators used personalized prompts that highlighted clinical knowledge, affirming language, and balanced encouragement frequency to stimulate involvement; (2) evident moderator collaboration: cross-moderation through mutual collaboration that focused on clinical knowledge, shared prompts, and comment-building reinforced community cohesion; (3) topic discussions: strategic techniques (querying, framing, and real-world contextualization) deepened clinical reasoning; (4) PCPs' clinical practice discussion: moderators bridged theory-practice gaps using case-based exchanges on screening, treatment protocols, and evidence translation; and (5) progression of engagement: moderators evolved from structured guidance to collaborative dialogue, reducing formality and increasing topic complexity over time.

XSL•FO
RenderX

Conclusions: Thematic analysis suggested that moderators are pivotal in transforming VCoPs into dynamic learning environments. Their ability to blend rational skills, clinical authenticity, and adaptive facilitation drives knowledge coconstruction and practice-relevant engagement. Future studies should consider investing in competency-based moderator training that emphasizes empathetic communication and collaborative techniques as a means of structuring effective VCoPs.

International Registered Report Identifier (IRRID): RR2-10.2196/57511

(J Med Internet Res 2026;28:e83764) doi:10.2196/83764

KEYWORDS

community-based fracture capture learning hub; virtual community of practice; web-based learning moderator; primary care physicians; case-based learning; osteoporosis education; qualitative content analysis; continuing professional development; facilitation strategies; moderator-participant interaction; collaborative learning

Introduction

Leveraging Virtual Communities of Practice to Foster Interprofessional Education, Knowledge Sharing, and Collaboration Among Primary Care Physicians

Interprofessional education (IPE) is a learning approach in which professionals from different disciplines are educated together to achieve a common goal [1]. As per the World Health Organization, “Interprofessional education (IPE) occurs when two or more professions learn about, from and with each other to enable effective collaboration and improve health outcomes” [2]. Acknowledging the critical role of interprofessional teamwork and collaboration in delivering high-quality patient care, there is increasing advocacy for training health care professionals in collaborative practice skills through IPE, as a supplement to their clinical skill development [3]. IPE has an essential role in breaking down professional silos among primary care physicians (PCPs) [4,5]. Systematic reviews have shown that implementing IPE and interprofessional collaboration (IPC) play vital roles in enhancing the effectiveness of care delivery and contributing to a variety of improved health outcomes [6,7]. One effective strategy for promoting IPE and IPC is the development of communities of practice (CoPs) [8]. These communities bring together individuals with shared interests to work collaboratively, enabling the exchange of information, the transfer of knowledge, the dissemination of best practices, and the strengthening of both professional and interprofessional skills [9]. CoPs are associated with a range of improvements, including immediate advantages such as identification of workforce competencies, knowledge sharing, and reuse of existing knowledge, while also supporting long-term outcomes, including the development of professional identity, increased current awareness, and improved implementation of evidence-based practices [9,10]. With the growth of web-based information-sharing tools, information and communications technology is increasingly being used to promote IPE and IPC, leading to the development of a wide range of virtual communities of practice (VCoPs) [11]. The primary motivation behind VCoPs is to establish networks of individuals who share common interests but are located in different geographic regions [12]. VCoPs have demonstrated effectiveness in several key areas, including the development and implementation of strategies to improve health services and the enhancement of collaboration among PCPs [13,14].

Enhancing Collaborative Learning in Primary Care: The Evolving Role of Moderators in Virtual Communities of Practice

PCPs encounter considerable difficulties in handling multiple chronic conditions during brief patient consultations, necessitating ongoing learning to maintain the use of evidence-based approaches [15,16]. Consequently, there is an increasing dependence on digital modalities for continuing professional development (CPD) [17-19], as they offer greater flexibility and time efficiency compared with traditional learning methods. Since the COVID-19 pandemic, social media platforms and VCoPs have gained prominence as effective tools for collaborative learning in the health care sector, overcoming time, cost, and geographical barriers [20-22]. VCoPs have been found to reduce professional isolation, enhance workforce retention, promote IPC, and offer a supportive environment that encourages active and safe participation in knowledge exchange [10,12].

Despite their potential, VCoPs often face challenges related to member recruitment, engagement, and long-term sustainability [23]. For VCoPs to be effective, several key factors have been identified as essential, including the presence of external support, strong and supportive leadership, effective facilitation, and the involvement of local champions [9]. Yet, a lack of clarity persists regarding the specific responsibilities of CoP moderators and the appropriate management of power dynamics within CoP groups [24]. We suggest that the role of the moderators in the formation and sustainability of effective VCoPs remains crucial. Evidence suggests that leaders and moderators enhance collaboration by clarifying expectations for participation, maintaining focus within discussions, and fostering an environment characterized by mutual respect, openness, and active engagement [12]. However, a knowledge gap persists in understanding the role of moderators in the success of VCoPs, particularly in relation to the dynamics of their interactions with participants and with one another—an area that remains underexplored in the existing literature. Hence, a full understanding of moderators’ roles needs to be appreciated, so that, as a group, moderators are well prepared and trained, and can provide more strategic approaches that increase participant engagement.

In light of the capacity of VCoPs to serve as a flexible and accessible approach for enhancing PCPs’ knowledge and confidence, and given the limited empirical investigation into

XSL•FO
RenderX

the influence of moderators on the effectiveness of VCoPs, this study was guided by the following research question: What roles do moderators play in enhancing the experience of participants in a VCoP?

The objective of the study was to investigate the moderators' involvement with the program and its members, and identify themes developed in these exchanges.

To answer the research question, in this study, we examined the role of moderators and their contributions in an innovative VCoP known as the Community Fracture Capture (CFC) Learning Hub, which is an interactive, case-based online education platform that facilitates small-group learning through adaptable and customized modules [25]. It is designed to address the osteoporosis treatment gap within community settings, support the adoption of evidence-based bone health practices among PCPs, and enhance the implementation of both the eLearning hub and fracture liaison service models in primary care, both nationally in Australia and internationally [26]. Adopting qualitative analysis of postings and interactions of 8 moderators across 4 cycles (each is 6 weeks long), the study was able to derive a number of emerging themes using qualitative thematic analysis approaches.

Methods

Methodology

Full details of the CFC's development are outlined in our published protocol paper [26].

Case Study: CFC Hub

The CFC Learning Hub is a secure and adaptable web-based platform designed to support collaborative learning [25,26]. Central to its approach is an interactive Discussion Forum that encourages participant engagement and peer-to-peer exchange on course material. The role of the moderators was specifically important in guiding the discussion. Experienced PCP moderators facilitated the group discussions, with support from specialist moderators as needed, fostering case discussions and directing the conversation based on the content's relevance and significance. A qualitative evaluation of the CFC model provides important insights into moderators' interactions with and experiences of this innovative educational approach, enriching quantitative results [27] by offering contextual understanding and a deeper interpretation of the observed outcomes.

Between May 2022 and October 2023, 4 separate 6-week small-group cycles of the program were implemented, involving a total of 55 PCPs. Of these, 33 (60%) PCPs actively participated in the activities by sharing comments or questions. Across all cycles, 8 PCPs and specialist physicians served as the program moderators. At any one time, there was always both a PCP and a specialist active in their respective roles as moderators. All moderators actively contributed to the ongoing discussion through commenting, posting, initiating conversations, or answering enquiries.

Understanding the experiences of the learning activity moderators is crucial for developing and evaluating an effective learning model [28,29]. Effective facilitation of small-group learning is critical for creating opportunities for collaborative engagement among learners and for fostering team-building competencies—skills that are essential for functioning within health care environments [28]. This qualitative study aimed to answer the research question set out above. We propose that moderator engagement within the CFC Learning Hub enhances the platform's functionality as a peer-driven educational tool by supporting the dissemination of current fracture prevention knowledge to PCPs across a range of professional backgrounds and experience levels, while also contributing to increased confidence and motivation among PCPs in delivering osteoporosis care, with these effects being measurable through integrated analytics systems.

Ethical Considerations

This study was approved by the Melbourne Health Human Research Ethics Committee (site reference number: 2016.24). All PCP participants provided electronic consent for the use of their anonymized data for research and audit purposes. Electronic informed consent was obtained through Praxhub's platform, with measures of data anonymization and deidentification in place to uphold privacy and confidentiality standards. Participation in the program was voluntary and did not involve any form of compensation.

Recruitment

Moderators were recruited by the principal investigators through position announcements circulated within internal networks. Candidates for moderator roles were interviewed informally and engaged in discussions outlining their role responsibilities, alignment with the project's timelines and milestones, and expectations for engagement. Final selection was based on clinical expertise, prior experience with similar projects, availability, and overall suitability with respect to the project's aims and team dynamics.

PCP participants were recruited via web promotion of the CFC program via the Praxhub platform [30], joining small private groups (12 to 16 PCPs per cycle) to engage in week-long modules, discussions, and quizzes within a secure, supportive learning environment.

Participants and Moderators

A total of 55 PCPs participated in the program across four 6-week cycles. Of these, 33 (60%) were active participants, sharing comments or questions. Across all cycles, 8 moderators (2 PCPs and 6 specialist physicians; Table 1) served as the program moderators by leading activities and encouraging discussion. The selected specialists were all clinical endocrinologists with a clinical and research interest in bone and mineral disorders.

XSL•FO
RenderX

Table 1. Total number of moderators and their specialty.

CFC[a] moderators	Number of moderators	Role
Specialists	6	• Filter and select relevant case studies. • Provide guidance and support clinical focus. • Primary focus: content preparation and clinical consultation aligned with up-to-date clinical practice.
PCPs[b] with special interest in bone health	2	• Filter and select relevant case studies. • Curate content designed by specialists to ascertain relevance to PCPs. • Introduce the topic and lead the group discussion. • Primary focus: direct interaction with members to encourage involvement and engagement.

[a]CFC: Community Fracture Capture.

[b]PCP: primary care physician.

Qualitative Data and Analysis

The qualitative data included in this analysis were collected from over 630 comments, 160 pages, and more than 53,000 words of weekly discussions provided by participants and moderators. In the current analysis, we looked more specifically for moderator-participant and moderator-moderator engagement, with a focus on (1) the nature and tone of moderators' contributions, (2) the motivation and aims of the moderators' contributions, (3) interactions between moderators and participants, and (4) interactions between moderators. All the relevant discussion comments were compiled for qualitative analysis.

We used collaborative thematic analysis, as guided by Braun and Clarke [31,32]. The thematic analysis was done by the first and second authors independently initially, after familiarization with the data. This process was carried out to identify patterns in the data. To address the study's aim, the 2 researchers used relational content analysis [33,34]. Initially, they independently generated codes using both semantic (focused on explicit meanings of the text) and latent (focused on underlying tone and assumptions) approaches to capture a broad range of insights. After initial codes were generated, the 2 authors then collaboratively refined and defined themes and subthemes, through rigorous discussion, based on the evidence from the qualitative data. This process was carried out to identify patterns in the data. In the next phase, these researchers collaboratively reviewed and refined the codes to identify recurring patterns aligned with key themes. These were then organized into subthemes under broader thematic categories. Finally, one researcher coded the full dataset against these subthemes, with a random subset reviewed for accuracy by a senior researcher.

Given the novel approach of the eLearning hub, themes were developed after each cycle and cross-checked against existing and emerging research in the field. As feedback was gathered throughout the program, ongoing refinements were made to both the structure and content of the hub.

A relational content analysis was conducted [33,34], and the findings were organized using both inductive (bottom-up) and deductive (top-down) methods, grouping diverse ideas into meaningful categories. The final themes and subthemes are illustrated with representative quotes from participants.

Results

Moderators' Engagement

This section details the program moderators' interactions with the program and its members, as well as their engagement techniques. This feedback can be categorized into 5 main themes: participation encouragement, evident moderator collaboration, topic discussions, PCPs' clinical practice discussions, and progression of nature of engagement (Figure 1).

XSL•FO
RenderX

Figure 1. Moderators' engagement and its main 5 themes and their subthemes. PCP: primary care physician.

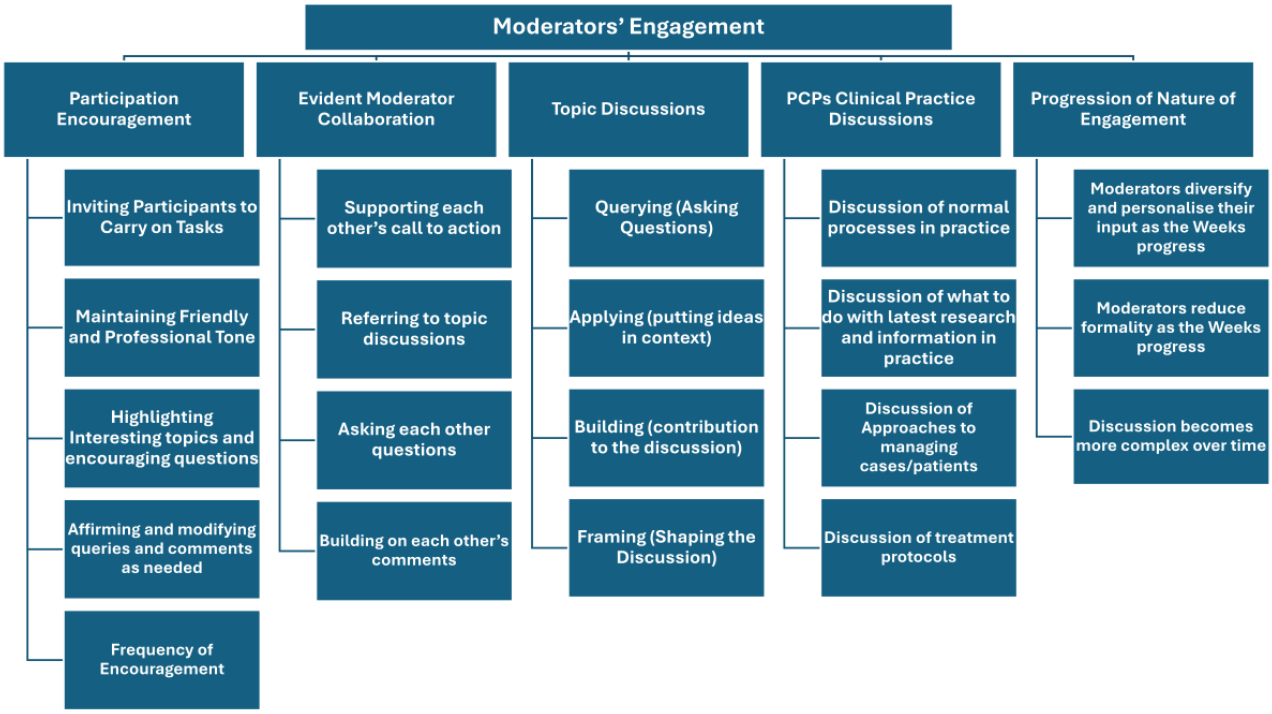

Participation Encouragement

Overview

Moderators of our virtual learning program used a variety of techniques to foster active participation and engagement among learners. Participants were shown a demonstration of the Praxhub webpage in a virtual, live session to permit interactions with lead moderators and other PCP participants. They actively invited participants to undertake tasks and contribute to discussions, ensuring that the tone remained friendly and professional to create a welcoming environment. By highlighting topics and areas of interest, our moderators aimed to capture learners' attention and maintain their interest. Encouraging questioning was another key strategy, as it was intended to stimulate continuous engagement with the activity. Additionally, moderators tried to affirm and constructively modify queries and comments, providing clear guidance and support. Striking the right balance in the frequency of encouragement communications was also crucial, aiming to provide learners with timely motivation without feeling overwhelmed. These techniques collectively aimed to help create an interactive and dynamic learning experience. This theme was broken down into 5 subthemes.

Inviting Participants to Carry Out Tasks

The moderators applied several key strategies for encouraging participation in the program. Moderators used targeted prompts to guide discussions on specific topics and encourage learners to engage with the material meaningfully. By inviting questions and actively acknowledging past participation, they created a supportive and interactive environment. Other techniques included emphasis on collaborative learning and using prompts for encouraging reflective thinking, further highlighting efforts to foster a sense of community and to stimulate deeper analysis.

One moderator encouraged participation using a targeted prompt:

> *I would also be very interested to hear your thoughts on strengths and limitations of DXA scans to screen for osteoporosis. If you have other questions, please also feel free to ask them.*

Another strategy was to highlight collaborative thinking:

> *We're really keen to hear your thoughts so that we can all learn from one another.*

Also, using prompts for reflective thinking was one of the participation encouragement techniques:

> *Yes, those are the questions to ask. I'm wondering how others think the answers to those questions might influence what we do next?*

Collectively, these approaches aimed to keep learners engaged, motivated, and actively contributing to the learning experience.

Maintaining a Friendly and Professional Tone

The moderators' tone across the activity demonstrated a consistent effort to be friendly, professional, encouraging, and sympathetic. Their informal and upbeat introduction aimed to set a welcoming tone, motivating learners to engage with the interactive hub and reassuring them that participation is confidential. The professional yet warm welcome was targeted at establishing a supportive environment for sharing and learning. Additionally, moderators' acknowledgment of a participant's contribution and encouragement for others to submit cases further fostered a collaborative atmosphere. Their affirmations and praise for correct insights and thoughtful responses were intended to be encouraging and supportive, reinforcing positive participation. In addition, their responses, including the appreciation for additional information and structured answers, showed attentiveness and respect for

XSL•FO
RenderX

learners' contributions. They provided personalized responses to participants, highlighting their questions and comments, which was intended to show genuine interest and support, promoting a positive and engaging learning experience.

One moderator gave the following friendly welcome to program participants:

> *Hi all. The weekend is nearly here. The perfect time to start getting involved in the CFC interactive learning hub. Peer-to-peer learning needs you. And don't be shy - what happens in the hub, stays in the hub! Ways to get involved?*

Another moderator also professionally welcomed participants:

> *Welcome to the CFC Bone Hub. I'm (name) and I am going to be your course moderator for Week 1. The purpose of this forum is to provide a space where we can learn from each other, through our own experiences, and feel comfortable sharing knowledge in a supportive environment.*

One moderator acknowledged a contribution and encouraged others to participate:

> *Thanks to one of our participants who has provided this good case and some questions to get us all thinking. Don't forget, any participants can send in an interesting case with some questions for consideration to be a case of the week.*

Another moderator praised members' contributions:

> *You are absolutely right that hyperparathyroidism preferentially results in cortical bone loss.*

Also, one moderator expressed appreciation for members' efforts,

> *Thanks for that extra information, very helpful in understanding more about the case!*

Other moderators personalized responses to different members in a professional and friendly manner:

> *(Member name)- this is a great point and highlights a crucial component in the case presented.*

Overall, the moderators effectively used a blend of encouragement, recognition, and constructive feedback to create a supportive and interactive learning environment.

Highlighting Interesting Topics and Encouraging Questions

Moderators tended to highlight and encourage questions to foster participant engagement. They posed questions to introduce new topics and stimulate deeper discussion and exploration. They also highlighted discussing additional topics and cases in an attempt to enrich the discussion with diverse perspectives. Such endorsement of interesting information can enhance the learning experience.

One moderator asked a question to stimulate discussion:

> *I'd like to ask a question about the fragility fracture risk calculators.*

Another moderator used an expression of interest presented by one of the members to highlight the importance of a noteworthy area and to ignite the discussion further:

> *Hi (member name), Great additional question. I might be pre-empting next week somewhat with a response but since you've asked, I'll give a general overview ahead of the greater detail you'll get in week 6.*

A third moderator suggested discussing an additional case vignette alongside the current one to enrich the discussion:

> *Another complementary case vignette has been submitted for discussion. Perhaps we can discuss it alongside the current case?*

Overall, these approaches reflected a strategic effort to engage participants by spotlighting intriguing topics and fostering a dynamic, interactive environment.

Affirming and Modifying Queries and Comments as Needed

Moderators worked to affirm and modify queries and comments to ensure clarity and enhance the quality of the discussion. They actively corrected and refined questions or responses to improve accuracy and relevance, ensuring the dialogue remained focused and productive.

For instance, one moderator acknowledged a necessary update to a post by stating, "This is a post edit," signaling that changes were made for clarity.

Another moderator pointed out a discrepancy in the resources and ensured the correct materials would be provided, saying:

> *We realised the answers and the resources were not loaded together, so we will be loading these.*

To prevent confusion, a third moderator pointed out an error in wording, suggesting a revision:

> *Note: the following question had an error in the wording which should have read …..*

Through these actions, the moderators not only affirmed the ongoing dialogue but also ensured that it was continually refined and improved. Overall, their efforts aimed to help maintain a precise and evolving conversation, fostering an atmosphere of constructive engagement.

Frequency of Encouragement

Moderators consistently accessed the discussion boards to monitor and participate in the dialogue, ensuring their engagement at least once daily throughout each cycle. The frequency of their comments and probing questions, aimed at encouraging further discussion, varied based on factors such as the level of response from participants, the depth of the topic being discussed, and the presence of unresolved queries or questions that required follow-up. On average, moderators provided at least one instance of daily engagement per day of the cycle. In addition to their participation in the discussion threads, moderators also sent email communications 3 times a week during the cycle, keeping participants informed about any updates, encouraging them to join in the discussions, and reminding them of the activity's requirements.

XSL•FO
RenderX

In summary, these efforts contributed to maintaining active participation and fostering a continuous, interactive learning environment throughout the cycle, as demonstrated by the participants' level of engagement in the activity.

Evident Moderator Collaboration

Overview

The theme of "Evident Moderator Collaboration" covers how moderators worked together to enhance the learning experience through mutual support and active engagement. This teamwork aimed to create a dynamic and cohesive learning environment. This collaboration is subcategorized into 4 subthemes: supporting each other's call to action, where moderators encouraged participant involvement by backing each other's prompts; referring to topic discussions, where they recalled earlier posts to create continuity and deepen the conversation; asking each other questions, which modeled inquisitive behavior for participants, encouraging them to engage more deeply; and building on each other's comments, where they expanded on points made by another moderator or participants to further enrich the dialogue.

Supporting Each Other's Call to Action

Moderators worked collaboratively to encourage participant engagement by supporting each other's calls to action. They consistently reinforced each other's prompts, creating a sense of shared responsibility in fostering discussion. This mutual support helped to drive the conversation forward and encouraged participants to respond actively.

For example, one moderator expressed eagerness for participant contributions to emphasize the importance of a colleague moderator's inquiry and motivated the participants to join the conversation by stating:

> *Looking forward to hearing your thoughts on (moderator) questions.*

In turn, the moderator acknowledged the colleague moderator's encouragement and shared enthusiasm for productive dialogue by responding:

> *Thanks (name). I'm looking forward to some good discussion.*

As a result, participants were encouraged to continue the conversation and responded effectively to the moderator's discussion. One of the participants further commented to the moderator, "I like to know how T score to be interpreted with Z score reading." They then clarified their understanding after receiving the moderator's response, "Now I understand that if Z score is below -2 then secondary causes should be investigated when the T score is also below -2.5."

Another moderator also contributed by showing appreciation for another colleague's efforts and building anticipation for future discussions, saying:

> *Looking forward to chatting tomorrow night. Thank you (name) for the teriparatide information!*

Overall, these interactions highlighted how moderators worked together to strengthen the call for engagement, fostering an atmosphere of cooperation and active participation.

Referring to Topic Discussions

Moderators consistently referenced previous conversations to ensure continuity and to build on ongoing discussions. This helped keep the dialogue focused and allowed for deeper exploration of the topics at hand.

As an example, the moderator referred directly to an earlier question regarding the best clinical practice for the management of oligomenorrhoea, demonstrating a clear connection to the ongoing conversation:

> *With regard to your question about oligomenorrhoea, reduced oestrogen exposure is certainly a risk factor for osteoporosis that should prompt consideration of osteoporosis screening by DXA.*

This response addressed a previous inquiry and contributed valuable insight into the current topic, reinforcing the relevance of earlier discussions. By referring to prior exchanges, moderators ensured that the conversation remained cohesive and that each new point was tied back to the broader discussion, enhancing the overall learning experience.

Asking Each Other Questions

Moderators demonstrated how inquisitive dialogue can enrich discussions and model active engagement for participants. By posing questions to one another, they not only sought clarification but also encouraged deeper exploration of the topics at hand.

For example, encouraging further discussion and highlighting the value of peer collaboration, one moderator prompted his colleagues by asking:

> *I wonder (if) (moderator one) or (moderator 2) (can share) what proportion would you think would be expected?*

Similarly, a moderator's question prompted deeper analysis of the case, inviting others to explore the situation more thoroughly, "Great talk (moderator). With the case study - was this a failure? Could we have prevented the fracture? She had low vitamin D, smoking and postmenopausal. Would there be a role for picking her up say 5 years ago and discussing lifestyle, weight training and doing a baseline DEXA if she was willing to pay for it?" which not only sought insight but also invited a peer-driven exchange of strategies.

These exchanges role-modeled the importance of asking questions to clarify, challenge, and deepen understanding, encouraging participants to actively engage in their own inquiries.

Building on Each Other's Comments

Moderators actively expanded on one another's contributions to deepen the discussion and explore various facets of the topic.

For example, a moderator's comment sparked further engagement from participants:

XSL•FO
RenderX

> *Great points have been raised in the thread above.*

Another moderator responded by acknowledging the discussion and extending it, encouraging the group to look beyond what had already been covered, saying:

> *Thanks for the comments (moderator one) and (moderator two). We have already discussed some aspects of this case last week, however there are more things to consider*

One moderator also built on another moderator's input and added another layer to the conversation by connecting patient perspectives to the clinical details, saying:

> *That is so true (moderator name). A lot of my patients associate dietary calcium to high dietary cholesterol.*

Overall, through these exchanges, moderators worked together to enhance the depth of the discussion, building on each other's comments with the aim of leading to more comprehensive and insightful dialogue.

Topic Discussions

Overview

Moderators used a range of techniques to facilitate participant engagement and deepen understanding of the subject matter. Moderators guided participants through the exploration of topics by applying various strategies, which aimed to enhance participants' comprehension and application of the material. This approach aimed to provide participants with opportunities to actively engage with the content and with one another. This theme is broken down into 4 subthemes.

Querying (Asking Questions)

Moderators used questions strategically to engage participants and steer discussions in productive directions. By asking thought-provoking and relevant questions, they encouraged participants to share their knowledge and explore different aspects of the topic. These questions helped highlight key areas for further discussion, often prompting deeper reflection.

For instance, one moderator posed a question to get participants thinking about the management of osteoporosis and setting the stage for a comprehensive discussion, saying:

> *Can I ask the group for your thoughts about the use of exercise to manage osteoporosis in your patients?*

Additionally, one moderator directly engaged a participant and fostered a more interactive environment by asking:

> *I am surprised by the inclusion of an incident fracture in the first year of treatment as all these medications take time to work and I would have thought a year on any of the treatments would not be long enough to have a real impact? (moderator name), what do you think?*

All in all, such questioning techniques effectively stimulated discussion, facilitated knowledge sharing, and ensured that the conversation remained focused and enriching for all involved.

Applying (Putting Ideas in Context)

Moderators used real-world examples and practical applications to help participants connect theoretical concepts to their everyday practices. This approach facilitated the discussion by encouraging participants to consider how the topics might play out in their own clinical settings.

For example, one moderator shared, "I also provide my patients with the Healthy Bones Australia fact sheets for calcium and vitamin D...," using this as an example of how they communicate essential information to patients, which prompted the activity's moderators and participants to reflect on similar practices.

As a result, one participant contributed by sharing their clinical practice and commented:

> *Of course, you can start patients … and this is what I do in frail patients…*

Additionally, moderators prompted participants to reflect on their observations in clinical practice, as one participant noted:

> *I always thought that pain from crush fractures was reported much less. It might be the degree of pain! I often see crush # on lateral spines - but no history of severe pain. Women do shrink - again often no specific pain but they all have some pain. What do the docs think about my experience?*

This invitation to share personal experiences helped ground the discussion in reality and led to a more nuanced exchange.

Overall, these applications of real-world examples allowed moderators to create a meaningful and relatable discussion, helping participants to apply new ideas and insights directly to their clinical contexts.

Building (Contribution to the Discussion)

Moderators focused on building the discussion by encouraging participants to contribute their insights and engage actively with the content. They created an environment where everyone was invited to share, fostering a collaborative atmosphere.

For instance, one moderator used a deliberately leading question to build on an ongoing conversation and encourage participants to consider various implications, asking:

> *Unfortunately, we don't have any more history on the 1986 fracture but as you say more details would be valuable. A change in therapy is probably warranted but (and it's a deliberately leading question!) are there any potential issues with changing straight from denosumab to teriparatide?*

Additionally, moderators prompted deeper engagement by inviting feedback from the group, as seen in the question:

> *I'm interested if this is consistent with your practice or what you've seen practised? And I'm keen on the group's thoughts as to whether the complexity of the plan is even required.*

This approach encouraged participants to reflect on their own practices and share their perspectives, thereby enriching the overall discussion.

XSL•FO
RenderX

By creating these opportunities for input, moderators fostered a sense of ownership and encouraged participants to contribute their experiences and expertise to the conversation, building a more dynamic and collaborative learning environment.

Framing (Shaping the Discussion)

Moderators used framing techniques to shape the direction and focus of discussions, ensuring that participants remained engaged and that the conversation stayed aligned with the educational goals. They often introduced topics by setting clear expectations for the flow of the discussion.

For example, one moderator guided the conversation by stating, “Therapies will be covered in depth later in the educational program, but it is worth discussing it briefly for case 2 before we move on to week 3's case,” which helped participants understand the structure and provided context for the current discussion. Another moderator framed the week’s session by acknowledging the contributions of participants, saying:

> *Welcome to week 3. Thanks to one of our participants who has provided this good case and some questions to get us all thinking. Don't forget any participants can send in an interesting case with some questions for consideration to be a case of the week.*

This approach not only shaped the direction of the current week’s discussion but also encouraged ongoing participation from the group.

By encouraging participants to share their thoughts and questions, moderators aimed to create an interactive environment while maintaining focus on key topics. These framing techniques were essential in shaping the overall structure and ensuring that the discussions remained relevant and engaging for all participants.

PCPs’ Clinical Practice Discussions

Overview

Moderators actively shared their own clinical practice experiences to enrich the learning process and make the discussions more relatable. By offering real-world examples from their own practices, they helped bridge the gap between theory and clinical application, which allowed participants to see how they might implement the discussed strategies in their own settings. Sharing such experiences not only helped participants gain a clearer understanding of the clinical aspects of the discussions but also fostered a collaborative learning environment, where participants could learn from each other's real-world experiences. This exchange of clinical knowledge created a richer, more interactive learning experience for all involved. This theme is subcategorized into 4 subthemes.

Discussion of Normal Processes in Practice

The moderators shared their normal clinical practices as a way of promoting discussion and encouraging learners to reflect on their own approaches. By presenting real-world scenarios from their practices, they opened the door for valuable exchanges and collaborative learning.

For example, one moderator used a targeted prompt to engage participants with a question on screening and calcium intake, stating:

> *I will be interested in what people think and how many screen patients before the first fracture. I make it a point to ask all women about calcium intake when coming in for contraceptive advice. I am amazed at how many have insufficient calcium in their diet!*

Similarly, another moderator invited discussion around the management of dental procedures in patients treated with antiresorptives:

> *I’m also keen to hear from colleagues as to how they approach elective dental procedures in patients treated with anti-resorptives. There are at least a few different approaches I’m familiar with in guidelines and from colleagues, am keen to hear from the group on what’s done in the real world.*

These strategies encouraged reflection on common clinical practices, fostering a collaborative atmosphere where participants could learn from one another’s experiences and approaches.

Discussion of What to Do With the Latest Research and Information in Practice

The moderators facilitated discussions on how to incorporate the latest research and information into clinical practice, encouraging participants to explore new findings and evaluate their practical implications.

One moderator shared insights from recent studies, saying:

> *That's an interesting question. I'm not familiar with human studies of the effect of ethanol on anti-resorptive actions. This basic science study in rats (study link) found that...,*

Additionally, they highlighted other key resources, like the National Prescribing Service Medicine Wise tables, offering practical tools for clinical decision-making:

> *It is a good resource - for people who have not seen it, it is here: (study link)*

Furthermore, one moderator brought in specific research, which led to a deeper conversation about the comparative efficacy of treatments, sharing:

> *Additionally, one large meta-analysis of 11 studies suggested that while denosumab improves bone density to a greater degree than bisphosphonates in post-menopausal osteoporosis, there is no difference in overall fracture rates*

These exchanges helped participants navigate the integration of emerging research into their everyday clinical practice.

Discussion of Approaches to Managing Cases/Patients

Moderators encouraged discussion around various approaches to managing complex cases, prompting participants to reflect on their own practices and share insights.

XSL•FO
RenderX

One moderator, for example, shared their surprise at a particularly challenging case, stating:

> *Wow, interesting case with a very low BMD in the spine - never seen one as low as that! I am interested as to what tests people would do when faced with such a case!*

This opened the floor for participants to discuss diagnostic strategies and test selections in similar scenarios.

Another moderator addressed a specific clinical consideration, saying, "For those who manage osteoporosis in patients with renal impairment, you might find the slides in the link useful," which encouraged participants to explore specialized resources tailored to managing osteoporosis in patients with this condition.

These discussions allowed participants to compare their approaches to managing unique and complex patient cases, fostering a collaborative environment for knowledge-sharing and practical problem-solving.

Discussion of Treatment Protocols

The moderators facilitated discussions on treatment protocols, encouraging participants to share their practices and perspectives on various approaches.

One moderator, for instance, initiated a discussion about the approach to managing osteoporosis in postmenopausal women, saying:

> *Dear all, how many of you organise a DEXA after menopause for women. I find I am performing it a lot more especially if any past risk factors. are there any particular groups that you screen?*

This prompted one participant to share real-life clinical experience:

> *I recently started working in a woman's clinic. I am struggling to find clear guidance as to when osteoporosis screening should start.*

Another moderator contributed by highlighting the relevance of specific guidelines, stating, "The Medicare Benefits Schedule (MBS) guidance on bone densitometry is quite helpful when considering some of the issues raised in this discussion," inviting participants to consider how these guidelines influence their treatment decisions.

Through these exchanges, participants had the opportunity to gain valuable insights into various approaches, strengthening their collective understanding and enhancing their ability to manage similar cases in their own practices.

Progression of Nature of Engagement

Overview

This theme focused on how the nature of discussions led by moderators evolved throughout the activity. Initially, moderators focused on providing clear instructions and creating a structured environment. Over time, they encouraged more collaborative and reflective dialogue, allowing participants to take on a more active role in shaping the discussions. This theme has been consistently observed throughout the cycles. It is broken down into 3 subthemes.

Moderators Diversify and Personalize Their Input as the Weeks Progress

Moderators had to diversify and personalize their input as the weeks progressed, adjusting their engagement to meet the evolving needs of the participants. Early in the activity, the moderators provided general updates and guidance, such as sharing quiz answers and encouraging participants to ask questions. For example, one moderator stated:

> *Hello everyone, please find the answers to the Week 2 Quiz as below. Please do not hesitate to query the answers, some questions are intended to be controversial! WEEK 2 MCQs (are as follows):*

As the activity evolved, the same moderator began tailoring responses to individual participants and their inquiries about the quiz, offering more detailed, personalized feedback, reflecting a shift towards more personalized engagement. For instance, the moderator reshared the MCQs answers directly with one of the participants, saying:

> *Hi (member name), please see week 2 MCQ answers as below:*

And addressing a member's understanding saying:

> *Hi (participant name), that is exactly right. Regarding osteopenia paradox: More fractures occur in subjects with osteopenia compared to osteoporosis, because although the relative risk of fracture is higher in osteoporosis, there are more people in the population with osteopenia.*

Additionally, affirming a member's feedback on an answer to a specific question:

> *Hi (member name), question 9 is designed to be controversial! You are right that anabolic should be considered in all the scenarios below, currently PBS prescriptions require 2 x fragility fractures (1 on treatment).*

The moderator's ability to adapt responses based on the participant's queries reflected progression from general messaging to more personalized, detailed interactions. This evolution in engagement demonstrates a shift towards fostering a more interactive and participant-centered learning environment.

Moderators Reduce Formality as the Weeks Progress

As the weeks progressed, moderators gradually reduced the formality in their communication to better engage with participants. In the early weeks, moderators maintained a professional tone, offering structured advice and information.

For instance, one moderator initially provided detailed and formal guidance, saying:

> *I would still do all the usual screening tests. She could still have other causes for her osteoporosis and could be devastating to miss these.*

XSL•FO
RenderX

However, as the activity evolved, the tone became more relaxed and informal, reflecting a shift towards a more conversational and engaging approach. In a later interaction, the same moderator expressed his thoughts in a less formal manner, saying:

> *OMG I am dizzy just thinking about the case. I would really be looking to the experts for this one... but how is it he is here – he has had renal physicians and endos for years... would they not have intervened early-is it not possible??*

This informal tone helped humanize the discussions and made the moderator appear more approachable. By using casual expressions and a less rigid style, moderators fostered a more comfortable environment, encouraging participants to engage more openly and freely in the discussions.

Discussion Becomes More Complex Over Time

As the activity progressed, the moderators facilitated discussions that grew increasingly nuanced and complex, challenging participants to consider more detailed aspects of patient care and clinical decision-making.

For instance, a moderator's questions grew more detailed and reflective over time. At an earlier stage of the cycle, the moderator raised an important, straightforward question about dual-energy X-ray absorptiometry (DXA) screening rates in practice, asking:

> *Given that all people over age 70 are eligible for a DEXA (DXA), my question to attendees is what proportion of people over 70 in your practice has had a screen?*

However, as the cycle progressed, the discussion became deeper, and the moderator's inquiries became more specific and clinically focused. For example, by mid-cycle, the moderator responded to relevant questions and discussed the routine practice of organizing DXA scans for older adults, pointing out how this had helped in detecting osteoporosis in several cases and highlighting the importance of considering physical signs:

> *Yes, age over 70 automatically qualifies her. I have been routinely organising DEXAs (DXA) for all my over 70s and have picked a decent amount of osteoporosis. Stooped posture would raise the issue of thoracic crush fractures - a lateral spine in her 60s may have given a diagnosis!*

Also, as the discussion went on, the moderator introduced more technical details and deepened the conversation about DXA scans, which led to a more advanced conversation on the variability and interpretation of diagnostic results, saying:

> *(It is) interesting that the wrist is so different and out of step to the other readings - and Colles fractures are so common and considered pathognomonic of osteoporosis. Most of my DEXAs (DXA) do not include the wrist!*

And later in a subsequent cycle, expanding the discussion to more unexplored areas, asking:

> *For me, one of the issues Is screening around the menopause. I often discuss with women going through menopause getting a baseline DEXA (DXA) scan (self-funded) - particularly to see where they are at. Is that something the rest of you do?*

These exchanges demonstrate how the moderators gradually steered the group from general discussions to more complex, case-specific inquiries, encouraging participants to engage with the subtleties of clinical practice and the evolving nature of evidence-based decision-making.

Discussion

Principal Findings

Overall, the CFC Learning Hub exemplifies the pivotal role of moderators in cultivating effective VCoPs for PCPs. Our qualitative analysis demonstrated how moderators serve as architects of engagement, bridging theoretical knowledge and clinical practice through strategic facilitation. Data analysis revealed that 5 core engagement themes emerged as pillars of effective moderation: participation encouragement, evident moderator collaboration, topic discussions, PCPs' clinical practice discussions, and Progression of Nature of Engagement. The study also showed that moderator collaboration was observable through mutual reinforcement of call to action, cross-referencing prior topic discussions, and peer-to-peer questioning and comment-building. In particular, moderators' engagement evolved in early weeks from formal instructions and structured guidance to informal dialogue, personalized feedback, and participant-led discussions in later weeks. Further, we found that topic discussions were facilitated via strategic querying, clinical contextualization, and progressive framing, where clinical practice discussions dominated exchanges, focusing on screening protocols, interpretation of guidelines, and management of complex cases. These findings align with and extend current understanding of effective VCoP facilitation in health professional education, emphasizing the nuanced interplay between relational warmth, strategic pedagogical techniques, and authentic clinical expertise.

In interpreting our findings through the lens of Wenger's [35,36] CoP framework, the 5 themes we identified cohere closely with the framework's theoretical dimensions of mutual engagement, joint enterprise, and shared repertoire. First, participation encouragement and progression of engagement reflect mechanisms by which moderators support mutual engagement: by prompting participation, recognizing contributions, and scaffolding newcomers' trajectories, moderators help sustain the regular, trust-based interactions that Wenger argues bind a CoP as a social entity (mutual engagement) [37,38]. Second, the themes of topic discussions and Clinical Practice Discussions embody aspects of the joint enterprise, as they signal how community members, under moderator facilitation, negotiate and coconstruct what the CoP is "about"—namely, relevant clinical challenges and professional concerns in practice—thus aligning with Wenger's notion of a collectively negotiated domain and mutual accountability for its purposes [24,38]. Finally, the evident moderator collaboration theme fosters the development of a shared repertoire, as moderators and

XSL•FO
RenderX

participating PCPs jointly build and maintain communal resources (eg, shared norms, language, case-based learning, narratives), which act as artifacts or tools for ongoing practice (eg, stories, tools, and routines) [39]. In sum, our empirical themes suggest that effective moderation not only reinforces Wenger's core CoP dimensions but also actively operationalizes them, facilitating a more cohesive, participatory, and sustainable community of practice.

The theme participation encouragement demonstrated how moderators proactively created an inclusive environment through targeted invitations, sustained professional yet warm communication, and strategic highlighting of relevant topics. This aligns with evidence that welcoming tones and explicit invitations reduce perceived barriers in web-based learning spaces, particularly for time-constrained professionals [28]. Techniques such as affirming contributions while constructively refining queries echo best practices in digital facilitation, where balancing support with intellectual rigor maintains engagement without compromising depth [40]. Moderators' daily presence and thrice-weekly email updates further reinforced reliability—a factor linked to sustained participation in web-based CPD [25]. These moderation practices likely satisfy core psychological needs described in self-determination theory (SDT)—autonomy, competence, and relatedness—thereby fostering more intrinsic motivation [41]. Research in online learning shows that perceived instructor (or moderator) support enhances satisfaction of these needs, which in turn promotes engagement [42,43]. By explicitly inviting participants and using clinically relevant framing, moderators support autonomy; by affirming and reflecting, they nurture competence and relatedness. Moreover, their reliable presence establishes trust and predictability, which helps external motivations internalize over time, increasing self-determined engagement. SDT has also been applied in CPD contexts: for example, a study of nursing professionals' mobile learning found that autonomy and competence were linked to higher motivation in continuing education [44].

VCoP moderators have been shown to enhance collaboration within a VCoP by ensuring clear rules of engagement, maintaining focus in discussions, and fostering engagement, shared respect, and openness [12]. CFC Learning Hub moderators played a pivotal role in sustaining engagement through collaborative questioning and personalized feedback. Their use of relational techniques (eg, affirming contributions or posing reflective questions) mirrors effective practices in web-based facilitation for health professionals [45]. The use of affirming language, reflective prompts, and real-world case discussion aligns with other studies where empathetic communication and clinical relevance were identified as key facilitators of engagement in small-group learning [40]. For instance, their shift from formal instruction to informal dialogue resonates with CoP theory, where trust-building fosters deeper engagement [35,36]. In addition, this shift toward a more informal, trust-based dialogue reflects CoP principles: as trust builds, participants gradually move from peripheral to more central roles (legitimate peripheral participation), deepening their ownership of the community [36,46]. In addition, the use of techniques like framing and query building in topic discussions contributed to fostering activity homogeneity and focus. Reports indicated that a VCoP tends to see higher levels of member interest and involvement when it is more homogeneous or 'focused' [47]. Moreover, this facilitation style reflects principles of adult learning theories: adult learners are more engaged when learning is relevant, self-directed, and respectful of their professional experience [48]. By reducing access barriers and signaling respect for participants' expertise, moderators align with these andragogical principles. Furthermore, moderators' focus on using real-world clinical examples to contextualize discussions aligns with effective strategies in health-related learning environments [23,47]. Successful VCoPs require that the knowledge shared and discussed by members be applied to clinical practice [47]. Professionals are more likely to engage when the VCoP is patient-centered, and when guidelines and resources for translating research into practice are readily available in user-friendly formats [23].

The evident moderator collaboration theme highlighted how moderators visibly supported each other's prompts, built upon colleagues' comments, and modeled inquisitive dialogue. This intentional demonstration of teamwork aligns with Wenger's [35] concept of "communities of practice," where shared leadership fosters collective knowledge construction. By publicly engaging in peer questioning, moderators normalized critical thinking and reduced hierarchical barriers, and signaled that inquiry and reflection are communal practices rather than top-down directives. This distributed leadership likely promotes a sense of shared ownership among participants. Moreover, modeling collegial critical thinking may satisfy participants' relatedness needs, per SDT [41], creating a psychologically safe environment for contribution**.** This finding reinforces the value of cofacilitation in VCoPs, where multiple moderators can diversify expertise and model collaborative problem-solving [10]. Shared moderation builds structural resilience: having multiple moderators reduces reliance on any single individual, which supports long-term sustainability. Over time, as participants observe and internalize this co-leadership, they may increasingly adopt facilitative behaviors themselves, strengthening the community's capacity and self-governance [40]. Further, the Progression of Nature of Engagement theme illustrated moderators' evolution from structured guidance to collaborative partnership. Diversifying input (eg, shifting from group announcements to personalized feedback) and reducing formality mirrored community maturation, consistent with VCoP lifecycle models where trust enables deeper interaction [36]. Increasing discussion complexity over weeks reflected adaptive scaffolding—a facilitator skill linked to improved learning outcomes in web-based spaces [28]. Adaptive scaffolding in web-based education was reported to foster learner autonomy, engagement, and self-regulation [49]. From a motivational perspective, this progression supports increasing autonomy (participants gain more control), competence (they engage in complex discourse), and relatedness (peer interactions deepen), all of which facilitate more internalized engagement under SDT. Furthermore, as participants transition from peripheral to more central roles, this reflects the concept of legitimate peripheral participation: newcomers start with less risky participation and gradually assume more central, meaningful roles, reinforcing

XSL•FO
RenderX

both individual development and the stability of the community [50].

Our findings underscore that effective VCoP moderation transcends technical platform management; it requires relational, pedagogical, and clinical competencies. Accordingly, training programs for VCoP moderators should emphasize relational skills such as building rapport, maintaining approachable yet professional tones, and fostering psychological safety; pedagogical strategies that include techniques for scaffolding discussions (eg, framing and questioning), encouraging reflection, and modeling collaboration; and clinical authenticity, including the judicious sharing of real-world experiences to contextualize evidence without dominating discourse. Indeed, the presence of a supportive learning environment was reported as a key theme to facilitate the knowledge transfer from CPD to clinical practice [51]. In addition, platform designers should prioritize features that enable seamless moderator collaboration (eg, shared dashboards), personalized communication tools, and analytics to track engagement patterns. A safe, nonjudgmental educational environment; well-organized and structured programs; and the presence of a motivating and intrinsically satisfying atmosphere with smooth and transparent facilitation were among the main factors reported to influence the transfer of knowledge from continuing professional education to clinical practice [52]. The CFC Hub's success in using case-based discussions also supports designing VCoPs around practice-relevant scenarios rather than abstract topics. These insights additionally carry broader implications for practice, as the identified relational and pedagogical competencies, along with the platform design considerations, can inform moderator training and VCoP implementation across a wide range of clinical specialties and health care settings. For instance, moderation by cognition-based team swift trust was positively correlated with team creativity within nursing education [53]. Developing adaptable training programs and flexible platform architectures may enhance the scalability and relevance of VCoPs, thereby increasing their practical value and supporting more sustainable, cross-disciplinary digital professional development.

Limitations and Future Research

While this study yields meaningful insights into the role of moderated online learning environments, some limitations should be acknowledged. First, participant-moderator relationships may have been influenced by self-selection bias, with PCPs who were more motivated to improve osteoporosis care likely being overrepresented. Furthermore, participants were recruited via the Praxhub platform, with the majority being metropolitan-based practitioners with over 10 years of clinical experience. This recruitment approach may limit the generalizability of the findings, as the results may not fully capture the program moderation's impact across more diverse practice settings. Further evaluation is warranted, particularly among regional general practitioners and those with less clinical experience. Additionally, as is common in qualitative research, the study relied on self-reported data, which may be subject to response bias. Future studies should investigate the effectiveness of these educational strategies in larger and more heterogeneous health care populations to better assess their influence on patient care.

Future studies should expand contexts and test these themes in VCoPs across different specialty fields (eg, diabetes and mental health) and geographic settings to identify cross-cutting principles. Future research could also explore how moderator-participant dynamics evolve over extended periods and their impact on community sustainability. Although this study demonstrates immediate engagement and knowledge exchange, it remains uncertain whether these educational gains result in sustained changes in clinical behavior and patient outcomes. Longitudinal investigations are needed to assess the retention of knowledge and skills over time, confidence in providing optimal clinical care in the topics covered, and to evaluate their potential influence on patient management, including relevant clinical outcomes. Further, an in-depth examination of the partnership between the 2 types of moderators is needed, as it would enable a critical analysis of their cross-functional dynamics and offer deeper insight into their complementary roles in delivering an engaging and clinically relevant learning experience.

Conclusions

This study elucidated the critical role of moderators in VCoPs for PCPs. Through 5 key themes- participation encouragement, evident moderator collaboration, topic discussions, PCPs' clinical practice discussions, and Progression of Nature of Engagement, we demonstrate how moderators drive engagement by blending relational warmth, strategic facilitation, and clinical authenticity. Moderators' ability to adapt from structured guidance to collaborative dialogue fosters a dynamic environment where knowledge is coconstructed and applied to real-world practice. These insights strongly suggest the need for dedicated moderator training and thoughtfully designed VCoP platforms that prioritize facilitator support. As digital learning becomes integral to professional development, optimizing moderator roles will be essential for scaling effective, sustainable communities that bridge evidence-practice gaps in health care, strengthen ongoing workforce development, and ensure that VCoPs remain resilient, adaptive, and impactful over the long term.

Acknowledgments

We express our gratitude to the Royal Australian College of General Practitioners (RACGP) and the Australian College of Rural and Remote Medicine (ACRRM) for accrediting the Community Fracture Capture (CFC) Bone Hub as a professional learning activity. Additionally, we appreciate the support and resources offered by the University of Melbourne Department of Medicine and the Department of Diabetes and Endocrinology at Royal Melbourne Hospital, and the Faculty of Engineering and Information Technology at the University of Melbourne.

XSL•FO
RenderX

Funding

This study was supported financially by Amgen Australia Pty Ltd and Theramex. These sponsors had no involvement in the study's design, content development, or in the conduct, data collection, analysis, and interpretation of the research.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

AMF, SC, CC, AG, RA, and JDW contributed to study conceptualization. AMF, SC, CC, AG, RA, SS, CJY, RB, SP, LC, and JDW contributed to the methodology and investigation. JDW, SC, CC, RA, RB, CJY, LC, and SP contributed to conducting the learning sessions; SS contributed to providing regular liaison with Praxhub. AMF, SC, CC, RA, and JDW contributed to the writing of the original draft. AMF, CC, RA, SC, and JDW contributed to the project administration. AMF, SC, CC, RA, AG, and JDW contributed to the data curation. JDW, SC, and CC contributed to study supervision. All authors have read, provided feedback, and agreed to the final version of the manuscript for publication.

Conflicts of Interest

None declared.

References

1. Mohammed CA, Anand R, Saleena Ummer V. Interprofessional education (IPE): a framework for introducing teamwork and collaboration in health professions curriculum. Med J Armed Forces India 2021;77(Suppl 1):S16-S21 [FREE Full text] [doi: 10.1016/j.mjafi.2021.01.012] [Medline: 33612927]
2. World Health Organization. 2010. URL: https://www.who.int/publications/i/item/framework-for-action-on-interprofessional-education-collaborative-practice [accessed 2025-10-05]
3. King S, Greidanus E, Carbonaro M, Drummond J, Patterson S. Merging social networking environments and formal learning environments to support and facilitate interprofessional instruction. Med Educ Online 2009;14:5 [FREE Full text] [doi: 10.3885/meo.2009.T0000132] [Medline: 20165519]
4. Frenk J, Chen L, Bhutta ZA, Cohen J, Crisp N, Evans T, et al. Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. Lancet 2010;376(9756):1923-1958. [doi: 10.1016/S0140-6736(10)61854-5] [Medline: 21112623]
5. Lygidakis H, McLoughlin C, Patel KD. Achieving universal health coverage: technology for innovative primary health care education. ResearchGate. 2016. URL: https://www.researchgate.net/publication/311616296_Achieving_Universal_Health_Coverage_Technology_for_innovative_primary_health_care_education [accessed 2026-02-17]
6. Reeves S, Pelone F, Harrison R, Goldman J, Zwarenstein M. Interprofessional collaboration to improve professional practice and healthcare outcomes. Cochrane Database Syst Rev 2017;6(6):CD000072 [FREE Full text] [doi: 10.1002/14651858.CD000072.pub3] [Medline: 28639262]
7. Reeves S, Perrier L, Goldman J, Freeth D, Zwarenstein M. Interprofessional education: effects on professional practice and healthcare outcomes (update). Cochrane Database Syst Rev 2013;2013(3):CD002213 [FREE Full text] [doi: 10.1002/14651858.CD002213.pub3] [Medline: 23543515]
8. Lees A, Meyer E. Theoretically speaking: use of a communities of practice framework to describe and evaluate interprofessional education. J Interprof Care 2011;25(2):84-90. [doi: 10.3109/13561820.2010.515429] [Medline: 21043557]
9. Ford J, Korjonen H, Keswani A, Hughes E. Virtual communities of practice: can they support the prevention agenda in public health? Online J Public Health Inform 2015;7(2):e222 [FREE Full text] [doi: 10.5210/ojphi.v7i2.6031] [Medline: 26284150]
10. McLoughlin C, Patel KD, O'Callaghan T, Reeves S. The use of virtual communities of practice to improve interprofessional collaboration and education: findings from an integrated review. J Interprof Care 2018;32(2):136-142. [doi: 10.1080/13561820.2017.1377692] [Medline: 29161155]
11. Barnett S, Jones SC, Bennett S, Iverson D, Robinson L. A virtual community of practice for general practice training: a preimplementation survey. JMIR Med Educ 2016;2(2):e13 [FREE Full text] [doi: 10.2196/mededu.5318] [Medline: 27731864]
12. Barnett S, Jones SC, Bennett S, Iverson D, Bonney A. General practice training and virtual communities of practice - a review of the literature. BMC Fam Pract 2012;13(1):87 [FREE Full text] [doi: 10.1186/1471-2296-13-87] [Medline: 22905827]
13. Abos Mendizabal G, Nuño Solinís R, Zaballa González I. HOBE+, a case study: a virtual community of practice to support innovation in primary care in Basque public health service. BMC Fam Pract 2013;14(1):168 [FREE Full text] [doi: 10.1186/1471-2296-14-168] [Medline: 24188617]

XSL•FO
RenderX

14. Curran JA, Murphy AL, Abidi SSR, Sinclair D, McGrath PJ. Bridging the gap: knowledge seeking and sharing in a virtual community of emergency practice. Eval Health Prof 2009;32(3):312-325. [doi: 10.1177/0163278709338570] [Medline: 19696084]
15. Allen LM, Palermo C, Armstrong E, Hay M. Categorising the broad impacts of continuing professional development: a scoping review. Med Educ 2019;53(11):1087-1099 [FREE Full text] [doi: 10.1111/medu.13922] [Medline: 31396999]
16. Magwenya RH, Ross AJ, Ngatiane LS. Continuing professional development in the last decade – a scoping review. J Adult Contin Educ 2022;29(2):408-437. [doi: 10.1177/14779714221147297]
17. Thepwongsa I, Kirby CN, Schattner P, Piterman L. Online continuing medical education (CME) for GPs: does it work? a systematic review. Aust Fam Physician 2014;43(10):717-721 [FREE Full text] [Medline: 25286431]
18. Cook DA, Blachman MJ, Price DW, West CP, Baasch Thomas BL, Berger RA, et al. Educational technologies for physician continuous professional development: a national survey. Acad Med 2018;93(1):104-112. [doi: 10.1097/ACM.0000000000001817] [Medline: 28658022]
19. Baso YS, Syafruddin A, Padlan AA, Akbar M. Continuing medical education for general practitioners in Indonesia: moving toward accountability. Gac Sanit 2021;35(Suppl 2):S120-S122 [FREE Full text] [doi: 10.1016/j.gaceta.2021.07.008] [Medline: 34929792]
20. Rolls K, Hansen M, Jackson D, Elliott D. How health care professionals use social media to create virtual communities: an integrative review. J Med Internet Res 2016;18(6):e166 [FREE Full text] [doi: 10.2196/jmir.5312] [Medline: 27328967]
21. Yee M, Simpson-Young V, Paton R, Zuo Y. How do GPs want to learn in the digital era? Aust Fam Physician 2014;43(6):399-402 [FREE Full text] [Medline: 24897992]
22. van der Keylen P, Tomandl J, Wollmann K, Möhler R, Sofroniou M, Maun A, et al. The online health information needs of family physicians: systematic review of qualitative and quantitative studies. J Med Internet Res 2020;22(12):e18816 [FREE Full text] [doi: 10.2196/18816] [Medline: 33377874]
23. Vinson CA. Using concept mapping to develop a conceptual framework for creating virtual communities of practice to translate cancer research into practice. Prev Chronic Dis 2014;11:E68 [FREE Full text] [doi: 10.5888/pcd11.130280] [Medline: 24762532]
24. Li LC, Grimshaw JM, Nielsen C, Judd M, Coyte PC, Graham ID. Use of communities of practice in business and health care sectors: a systematic review. Implement Sci 2009;4(1):27 [FREE Full text] [doi: 10.1186/1748-5908-4-27] [Medline: 19445723]
25. Murad A, Hyde N, Chang S, Lederman R, Bosua R, Pirotta M, et al. Quantifying use of a health virtual community of practice for general practitioners' continuing professional development: a novel methodology and pilot evaluation. J Med Internet Res 2019;21(11):e14545 [FREE Full text] [doi: 10.2196/14545] [Medline: 31774401]
26. Fathalla AM, Chiang C, Audehm R, Gorelik A, Chang S, Yates CJ, et al. Developing and evaluating an interactive, case-based, web-based active learning tool for primary care physicians (community fracture capture learning hub): protocol for an acceptability and engagement study. JMIR Res Protoc 2025;14:e57511 [FREE Full text] [doi: 10.2196/57511] [Medline: 39999431]
27. Fathalla AM, Chiang C, Audehm R, Gorelik A, Chang S, Dao T, et al. Evaluation of a novel web-based active learning tool for primary care physicians' continuing professional development (the community fracture capture learning hub): quantitative analysis. JMIR Form Res 2026 Jan 21;10:e76216 [FREE Full text] [doi: 10.2196/76216] [Medline: 41564332]
28. Burgess A, van Diggele C, Roberts C, Mellis C. Facilitating small group learning in the health professions. BMC Med Educ 2020;20(Suppl 2):457 [FREE Full text] [doi: 10.1186/s12909-020-02282-3] [Medline: 33272270]
29. Schiekirka-Schwake S, Anders S, von Steinbüchel N, Becker JC, Raupach T. Facilitators of high-quality teaching in medical school: findings from a nation-wide survey among clinical teachers. BMC Med Educ 2017;17(1):178 [FREE Full text] [doi: 10.1186/s12909-017-1000-6] [Medline: 28962568]
30. Free online medical CPD training. Praxhub. URL: https://praxhubcom/ [accessed 2026-02-11]
31. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2008;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
32. Clarke V, Braun V. Thematic analysis. J Posit Psychol 2016;12(3):297-298. [doi: 10.1080/17439760.2016.1262613]
33. Naeem M, Ozuem W, Howell K, Ranfagni S. A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. Int J Qual Methods 2023;22:16094069231205789. [doi: 10.1177/16094069231205789]
34. Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and be(com)ing a researcher. Int J Transgend Health 2023;24(1):1-6 [FREE Full text] [doi: 10.1080/26895269.2022.2129597] [Medline: 36713144]
35. Wenger E. Communities of Practice: Learning, Meaning, and Identity. New York, NY: Cambridge University Press; 1999.
36. Wenger E, McDermott RA, Snyder W. Cultivating Communities of Practice: A Guide to Managing Knowledge. USA: Harvard Business Press; 2002.
37. Murillo E. Communities of practice in the business and organization studies literature. Inf Res 2011;16(1):464 [FREE Full text]
38. Li LC, Grimshaw JM, Nielsen C, Judd M, Coyte PC, Graham ID. Evolution of Wenger's concept of community of practice. Implement Sci 2009;4:11 [FREE Full text] [doi: 10.1186/1748-5908-4-11] [Medline: 19250556]

XSL•FO
RenderX

39. Wenger-Trayner EWTB. An introduction to communities of practice: a brief overview of the concept and its uses. Wenger-Trayner. 2015. URL: https://www.wenger-trayner.com/introduction-to-communities-of-practice/ [accessed 2026-02-17]
40. D'Eon M, Zhao R. Five ways for facilitators to get a grip on small group learning. Can Med Educ J 2022;13(2):82-88 [FREE Full text] [doi: 10.36834/cmej.72949] [Medline: 35572022]
41. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
42. He J, Wang Q, Lee H. Enhancing online learning engagement: teacher support, psychological needs satisfaction and interaction. BMC Psychol 2025;13(1):696 [FREE Full text] [doi: 10.1186/s40359-025-03016-0] [Medline: 40598609]
43. Yang Y, Chen J, Zhuang X. Self-determination theory and the influence of social support, self-regulated learning, and flow experience on student learning engagement in self-directed e-learning. Front Psychol 2025;16:1545980 [FREE Full text] [doi: 10.3389/fpsyg.2025.1545980] [Medline: 40207130]
44. Delia CS. The effect of mLearning on motivation in the continuing professional development of nursing professionals: a self-determination theory perspective. J Nurs Educ Pract 2021;12(3):22. [doi: 10.5430/jnep.v12n3p22]
45. Kellam H, Cook C, Smith DL, Haight P. The virtual community of practice facilitation model: a conceptual framework for healthcare professional education. Int J Technol Hum Interact 2023;19(1):1-14. [doi: 10.4018/ijthi.328578]
46. Barnett S, Jones SC, Caton T, Iverson D, Bennett S, Robinson L. Implementing a virtual community of practice for family physician training: a mixed-methods case study. J Med Internet Res 2014;16(3):e83 [FREE Full text] [doi: 10.2196/jmir.3083] [Medline: 24622292]
47. Mairs K, McNeil H, McLeod J, Prorok JC, Stolee P. Online strategies to facilitate health - related knowledge transfer: a systematic search and review. Health Info Libraries J 2013;30(4):261-277 [FREE Full text] [doi: 10.1111/hir.12048]
48. Mukhalalati BA, Taylor A. Adult learning theories in context: a quick guide for healthcare professional educators. J Med Educ Curric Dev 2019;6:2382120519840332 [FREE Full text] [doi: 10.1177/2382120519840332] [Medline: 31008257]
49. Shank E, Tang H, Morris W. Motivation in online course design using self-determination theory: an action research study in a secondary mathematics course. Education Tech Research Dev 2024;73(1):415-441. [doi: 10.1007/s11423-024-10410-9]
50. Lave J, Wenger E. Situated Learning: Legitimate Peripheral Participation. New York, NY, US: Cambridge University Press; 1991:138.
51. Clark E, Draper J, Rogers J. Illuminating the process: enhancing the impact of continuing professional education on practice. Nurse Educ Today 2015;35(2):388-394. [doi: 10.1016/j.nedt.2014.10.014] [Medline: 25467716]
52. Vasli P, Dehghan-Nayeri N, Khosravi L. Factors affecting knowledge transfer from continuing professional education to clinical practice: development and psychometric properties of a new instrument. Nurse Educ Pract 2018;28:189-195. [doi: 10.1016/j.nepr.2017.10.032] [Medline: 29112906]
53. Liu HY. The moderating role of team conflict on teams of nursing students. Int J Environ Res Public Health 2022;19(7):4152 [FREE Full text] [doi: 10.3390/ijerph19074152] [Medline: 35409835]

Abbreviations

CFC: Community Fracture Capture
CoP: communities of practice
CPD: continuing professional development
DXA: dual-energy X-ray absorptiometry
IPC: interprofessional collaboration
IPE: interprofessional education
PCP: primary care physician
SDT: self-determination theory
VCoP: virtual communities of practice

Edited by A Stone; submitted 08.Sep.2025; peer-reviewed by O Enaworu, I Wilson; comments to author 29.Oct.2025; revised version received 13.Dec.2025; accepted 16.Jan.2026; published 04.Mar.2026.

Please cite as:
Fathalla AM, Chang S, Audehm R, Chiang C, Gorelik A, Yates CJ, Snow S, Barmanray RD, Price S, Collins L, Wark JD
Moderators' Role in Fostering Health Practitioner Engagement Within an Online Active Learning Program (The Community Fracture Capture Learning Hub): Qualitative Relational Content Thematic Analysis
J Med Internet Res 2026;28:e83764
URL: https://www.jmir.org/2026/1/e83764
doi:10.2196/83764
PMID:

XSL•FO
RenderX

©Ahmed M Fathalla, Shanton Chang, Ralph Audehm, Cherie Chiang, Alexandra Gorelik, Christopher J Yates, Steve Snow, Rahul D Barmanray, Sarah Price, Lucy Collins, John D Wark. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Quantifying Innovation in Stroke: Large Language Model Bibliometric Analysis

Adam Marcus[1,2], BSc, MSc, MBBS, PhD; Georgina Lockwood-Taylor[3], MSc; Daniel Rueckert[1,4], PhD; Paul Bentley[2], PhD, MRCP

[1]Department of Computing, Imperial College London, London, United Kingdom
[2]Department of Medicine, Imperial College London, London, United Kingdom
[3]Department of Psychology and Neuroscience, King's College London, London, United Kingdom
[4]Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

Corresponding Author:
Adam Marcus, BSc, MSc, MBBS, PhD
Department of Computing
Imperial College London
Exhibition Rd, South Kensington
London, SW7 2AZ
United Kingdom
Phone: 44 020 7589 5111
Email: adam.marcus11@imperial.ac.uk

Abstract

Background: Thrombolysis and mechanical thrombectomy represent the most successful stroke innovations over the last 30 years. Quantifying innovation in stroke is essential for identifying productive research lines and prioritizing funding, but health care lacks validated methods for measuring innovation.

Objective: This study aimed to systematically evaluate the relationship between stroke-related patents and publications, demonstrate the feasibility of using large language models (LLMs) in this process, and identify the most rapidly advancing innovations in stroke care by mapping them to a theoretical innovation life cycle.

Methods: The Open Patent Services (European Patent Office) and PubMed databases were searched between 1993 and 2023 for "stroke OR cerebrovascular." In this bibliometric patent-publication analysis, a 13 billion–parameter Llama LLM was trained to identify patents related to stroke disease, as opposed to other references to the word "stroke," on a manually labeled subset of 5000 patents and assessed using 5-fold cross-validation. The LLM filtered irrelevant results, and the resulting patent codes were grouped into innovation clusters. For each cluster, annual patent and publication counts were normalized to adjust for global trends. Cluster-specific growth curves were plotted to analyze the rates and characteristics of growth. The innovation life cycle stage for each innovation cluster was estimated by fitting a sigmoid curve to the patent and publication data consistent with the diffusion of innovations theory by Rogers.

Results: The cross-validated accuracy of the LLM was 99.2%, with a sensitivity of 96.5% and a specificity of 99.6%. An initial bibliometric search retrieved 237,035 patents and 486,664 research publications. A manual review of a random sample of patents before filtering revealed that only 11.2% (56/500) were relevant to stroke. After LLM filtering, of the 237,035 patents, 28,225 (11.9%) stroke-related patents remained. These were grouped into 7 innovation clusters: pharmacological treatment, alternative medicine, rehabilitation devices, medical imaging, diagnostic testing, surgical devices, and artificial intelligence (AI) methods. Patent and publication counts were strongly correlated across clusters (Spearman r_s=0.65-0.92; P<.006) except for pharmacological treatment (r_s=0.09) and alternative medicine (r_s=0.55). Pharmacological treatments were the top-performing cluster over the last 30 years, accounting for 49.3% (36,005/73,094) of all patents, but patent activity in this area has plateaued since the late 2000s. AI methods, rehabilitation devices, and medical imaging exhibited exponential rates of patent growth, with annual normalized increases of 39.2%, 15.9%, and 5.8% compared to 16.9%, 5.3%, and 2.2% for publications, respectively.

Conclusions: Applying an LLM to publicly available patent and publication data provides a scalable way to quantify innovation in stroke. Pharmacological treatment appears to have entered a saturation phase, whereas AI methods, rehabilitation devices, and medical imaging remain in rapid growth, highlighting areas of greatest traction for future research and investment.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e70754) doi:10.2196/70754

KEYWORDS

diffusion of innovation; innovation; stroke; large language model; artificial intelligence; AI

Introduction

Turning points in stroke treatment occurred in 1995 and 2015 with the introduction of thrombolysis and mechanical thrombectomy, respectively [1,2]. Such shifts are rare and, in a more formal context, can be considered innovations: a term defined as a process that ushers in new technologies or techniques that induce a substantial change in practice [3,4]. Quantifying innovation in stroke care is vital as it helps discern which lines of research are productive, such as revascularization therapies, perfusion imaging, and decompressive surgery, as opposed to those that have been less successful, such as neuroprotective and neurorestorative therapies. Measures of innovation output aid in research strategy planning, prioritizing, and assessing the effectiveness of research funding. However, while the study of innovation is well established in other fields [5], health care has relatively few validated methods for quantifying innovation outputs, which can limit progress [6].

Prior evaluations of innovation in stroke have been limited, consisting largely of qualitative reviews [7-9] or conventional bibliometric analyses that track academic trends [10]. While valuable for tracking academic discourse and research activity, such bibliometric approaches have inherent limitations for measuring tangible innovation. Metrics based on citations and publications tend to reflect academic impact over practical implementation, and they do not readily distinguish incremental advances from transformative breakthroughs. Consequently, these methods primarily measure research inputs and academic outputs, not the development of novel, practical solutions.

An alternative approach, originally applied to surgery [4], leverages original patents as a benchmark of technological innovation by comparing the cumulative quantity of patents for a specific innovation with that of related peer-reviewed publications. This method has the advantage of drawing on a comprehensive repository of inventions that have been independently evaluated for novelty and utility; these are generally sufficiently mature and practical to have attracted the funding resources required for patent filing. However, patent analysis relies heavily on the precise interpretation and determination of relevant patents, a task complicated by the broad and often ambiguous language of patent documents, making it labor-intensive and time-consuming. The difficulty of this task is particularly magnified in the context of stroke-related patents given that the term “stroke” could denote a disease as well as multiple engineering concepts and mechanical action of engines, clocks, and other mechanisms. Conventional search engines such as Google are liable to confound stroke terms as they lack the specialized filtering and context awareness needed for precise medical searches.

In this regard, recent advancements in artificial intelligence (AI), specifically large language models (LLMs), hold tremendous potential. They have exhibited remarkable proficiency in textual tasks, making them invaluable in this context [11]. Essentially, LLMs are statistical models trained on vast datasets enabling them to learn intricate relationships between words and phrases. While training LLMs from scratch might pose considerable difficulties and financial burden, the recent availability of open-source trained LLMs to the public has mitigated these challenges [12].

Therefore, the aims of this study were 3-fold: first, to evaluate systematically the relationship between stroke-related patents and publications over the last 3 decades; second, to demonstrate the feasibility of using LLMs to assist in this process; and, finally, to identify the most rapidly advancing innovations in stroke care.

Methods

Although this was not a clinical study, the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [13] were adhered to where appropriate. The methodology was based on the work by Hughes-Hallett et al [4], with adaptions for stroke-related innovation.

Data Collection

The Open Patent Services web service, provided by the European Patent Office [14], was used to obtain patent application data from more than 80 different countries. The period from 1993 to 2023 was chosen to capture the modern era of stroke care, beginning shortly before the pivotal 1995 National Institute of Neurological Disorders and Stroke trial that established thrombolysis as a standard treatment [1]. Patents filed between 1993 and 2023 were downloaded if either their title or abstract matched the following Boolean search: “stroke OR cerebrovascular.” A PubMed (National Library of Medicine) search was also conducted using the same strategy to extract publication data for the same period.

Data Filtering

The collected patents were randomly sampled, and a subset of 5000 was manually annotated by a single author (AM), who is a medical professional, as either related or unrelated to stroke. A second, random unfiltered sample of 500 patents and 500 publications matching the search terms “stroke” or “cerebrovascular” was also manually labeled by the same annotator (AM). Annotations from the 5000-patent subset were used to provide ground truth for model fine-tuning, whereas the second sample was used to verify the accuracy of the search strategy, especially for PubMed results. Patents were considered stroke related if their primary content was directly relevant to stroke management or pathophysiology. Publications were considered stroke related if they contributed to the understanding of stroke, including basic science suitable for stroke journals and clinical studies with stroke as an end point.

XSL•FO
RenderX

A 13 billion–parameter Llama model (Meta AI) [12], a state-of-the-art open-source LLM trained on web data, was then fine-tuned using low-rank adaptation (LoRA) [15] to classify stroke-related patents. Fine-tuning was performed using PyTorch (version 1.13; Meta AI) on a machine equipped with a 2.80-GHz AMD EPYC 7543P central processing unit and an NVIDIA RTX A4500 20-GB graphics processing unit. The hyperparameters, which are listed in Table 1 and were based on standard values from the original LoRA paper [15], were chosen to ensure stable training while preventing overfitting. In particular, the LoRA rank was set to 8 to keep model adaptation minimal and efficient. The prompt template is provided in Figure 1. To assess prompt sensitivity, alternative phrasings were trialed on a development subset during fine-tuning. The model's performance was evaluated using 5-fold cross-validated binary classification metrics and compared to that of the base model before the final model was used to analyze the unlabeled patents and filter out those unrelated to stroke. The model was not used for classifying publications.

Table 1. The hyperparameters used for fine-tuning the Llama model using low-rank adaptation (LoRA).

Parameter	Value
Batch size	128
Number of epochs	10
Learning rate	0.0003
Optimizer	Adam
Maximum gradient norm	1
LoRA rank	8
LoRA alpha	16
LoRA dropout	0.05
LoRA target modules	Query and value projection

Figure 1. The prompt template used to fine-tune the Llama model to classify whether a patent was related to stroke.

```
Below is a patent application title, paired with an abstract. Write a
yes/no response whether the patent is related to stroke disease.

### Title:
[TITLE]

### Abstract:
[ABSTRACT]

### Response:
<COMPLETE>
```

Data Normalization

Across all fields, the number of patents and publications has risen exponentially. To adjust for this increase, both counts were normalized using the formula outlined by Hughes-Hallett et al [4]:

In this formula, II_i denotes the innovation index, defined as the number of patents or publications within a particular field; C_i is the innovation constant; and t_i is the total number of patents granted or publications indexed on PubMed for a given year i. For example, if a field had 50 patents in a year when 100,000 patents were granted and the maximum number of patents in any year during the study period was 200,000, then C_i = 100,000/200,000 = 0.5, and the normalized innovation index would be 50/0.5 = 100.

Identifying Innovation Clusters

The process of identifying innovation clusters involved a 2-stage method to ensure comprehensive coverage. Initially, the top 100 most frequent International Patent Classification (IPC) [16] codes from the filtered, stroke-related patent dataset were extracted. These codes, assigned by patent examiners, offer a standardized way of categorizing the technological domains of inventions. Focusing on the top 100 codes provided a quantitative starting point, representing the most concentrated areas of patent activity while avoiding the sparsity and noise of the long tail of less frequent codes.

These top 100 codes were then manually grouped into preliminary innovation clusters by 2 authors (AM and GL-T).

XSL•FO
RenderX

This grouping was based on the descriptive content of the IPC codes and their relevance to distinct areas of stroke care. Interrater reliability was assessed using the Cohen κ, and any disagreements regarding this grouping were resolved by a third author (PB).

To capture relevant patents and publications that may not have fallen into these top 100 codes, a second stage was implemented. Expanded, cluster-specific Boolean search strategies were developed as listed in Table 2 and performed on both the patent database and PubMed. The keywords for these searches were also determined by 2 authors (AM and GL-T), with a third author (PB) resolving any disagreements. For clusters such as alternative medicine, broader search terms such as "food" and "herbal" were intentionally used. This was necessary to capture innovations described in nonclinical or lay terms, which is common in patent applications for complementary therapies that may lack standardized medical terminology. The final dataset for each cluster comprised all documents identified through either the initial IPC code grouping or the subsequent expanded Boolean search. Finally, this entire 2-step methodology was repeated for patents and publications from the last decade (2013-2023) to allow for the determination of more recent innovations.

Table 2. PubMed and European Patent Office database search strategies.

Innovation cluster	Search strategy
AI[a] methods	(AI OR "artificial intelligence" OR "deep learning" OR "machine learning" OR "neural network") AND (stroke OR cerebrovascular)
Alternative medicine	(food OR tea OR coffee OR beverage OR herbal OR acupuncture OR aromatherapy OR reflexology OR "holistic therapy") AND (stroke OR cerebrovascular)
Diagnostic testing	("diagnostic testing" OR "diagnostic tools" OR "clinical tests" OR "screening tests" OR "blood tests" OR "laboratory tests" OR "genetic testing") AND (stroke OR cerebrovascular)
Medical imaging	(imaging OR angiography OR angiogram OR ultrasound OR CT OR MRI OR PET OR "computed tomography" OR "magnetic resonance" OR "positron emission tomography") AND (stroke OR cerebrovascular)
Pharmacological treatment	(thrombolysis OR aspirin OR clopidogrel OR warfarin OR DOACs OR alteplase OR tPA OR "tissue plasminogen activator" OR "thrombolytic therapy" OR "pharmacological treatment" OR "pharmaceutical composition" OR "drug therapy" OR "secondary prevention" OR "direct oral anticoagulants") AND (stroke OR cerebrovascular)
Rehabilitation devices	(rehabilitation OR neurorehabilitation OR exoskeleton OR "training device" OR "brain-computer interface") AND (stroke OR cerebrovascular)
Surgical devices	(thrombectomy OR embolectomy OR "clot removal" OR "clot retrieval" OR "catheter device" OR "surgical device" OR "endovascular treatment" OR "endovascular therapy") AND (stroke OR cerebrovascular)

[a]AI: artificial intelligence.

Statistical Analysis

All statistical analyses were performed using Python (version 3.11.3; Python Software Foundation) and the *statsmodels* [17] package. Permutation testing with Bonferroni correction was used to calculate *P* values adjusted for multiple tests. The relationship between patent and publication data was visualized using scatterplots to assess the nature of the association. On the basis of this visual inspection, an appropriate correlation coefficient was selected: Pearson (r) for linear relationships and Spearman rank (r_s), a nonparametric method, for monotonic but nonlinear relationships. To model the technology diffusion life cycle, innovation life cycle progression was derived by fitting sigmoid curves to the patent and publication data, a method consistent with the diffusion of innovations theory by Rogers [18]. To quantify the uncertainty in these estimates, 95% CIs were calculated using nonparametric bootstrapping, which involved repeatedly resampling the data and refitting the curve.

Results

Data and Filtering Performance

The initial search retrieved 237,035 patents and 486,664 publications. In a random, unfiltered sample of 500 patents and 500 publications matching the search terms "stroke" or "cerebrovascular," 11.2% (56/500) of patents and 74.2% (371/500) of publications were stroke related. The remaining patents typically referred to "stroke" in nonclinical contexts, including mechanical travel (eg, stroke length in pistons), engine cycles (eg, 2-stroke engines), lightning discharges, and line rendering in handwriting or graphics. These examples formed the negative class for model fine-tuning. An LLM was then fine-tuned to classify whether a patent was stroke related, achieving a cross-validated accuracy of 99.2% with a sensitivity of 96.5% and specificity of 99.6% and significantly outperforming the base model across all metrics listed in Table 3 (all $P<.001$). The receiver operating characteristic curve is shown in Figure 2 [19]. Prompt sensitivity was also evaluated during model fine-tuning using alternative phrasings; this had a negligible impact on classification outcomes. After filtering using the model, of the 237,035 patents, 28,225 (11.9%) stroke-related patents remained. Figure 3 illustrates the annual

XSL•FO
RenderX

increase in these patents filed by geographic region, with the largest proportion originating from China. The original and normalized counts of patents and publications related to stroke are shown in Figure 4. Normalized patent counts reached a peak in 2010, whereas normalized publication activity continues to grow.

Table 3. Five-fold cross-validated performance of the fine-tuned Llama model compared to the base model for classifying whether a patent was stroke related. *P* values are from 2-tailed paired t tests across the 5 folds.

Measure	Fine-tuned estimate (95% CI)	Base estimate (95% CI)	*P* value
AUROC[a]	0.990 (0.988-0.992)	0.654 (0.605-0.703)	<.001
Accuracy (%)	99.2 (99.0-99.5)	66.8 (66.3-67.3)	<.001
Sensitivity (%)	96.5 (94.8-98.2)	54.5 (48.3-60.7)	<.001
Specificity (%)	99.6 (99.4-99.7)	68.2 (67.6-68.7)	<.001
PPV[b] (%)	96.1 (94.8-97.4)	16.5 (13.9-19.1)	<.001
NPV[c] (%)	99.6 (99.4-99.8)	92.9 (92.2-93.7)	<.001
F_1-score (%)	96.3 (95.1-97.5)	25.2 (21.6-28.9)	<.001

[a]AUROC: area under the receiver operating characteristic curve.

[b]PPV: positive predictive value.

[c]NPV: negative predictive value.

Figure 2. Receiver operating characteristic curve illustrating the cross-validated performance of the fine-tuned Llama model compared to the base Llama model in classifying stroke-related patents. The shaded area represents the 95% confidence region determined via the fixed-width band technique [19]. AUC: area under the curve.

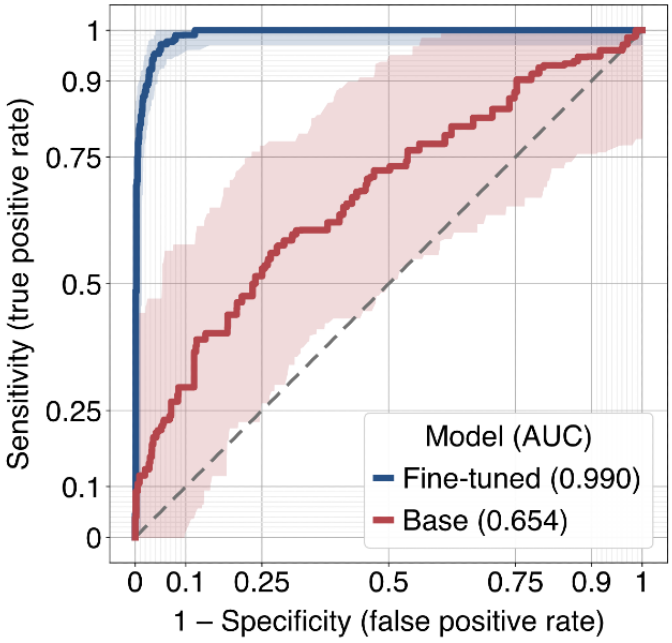

XSL•FO
RenderX

Figure 3. Stroke-related patents filed between 1993 and 2023 categorized by geographic coverage as defined by patent filing jurisdiction and route (China: China National Intellectual Property Administration; United States: US Patent and Trademark Office; Japan: Japan Patent Office; South Korea: Korean Intellectual Property Office; Canada: Canadian Intellectual Property Office; Europe: European Patent Office; worldwide: Patent Cooperation Treaty or World Intellectual Property Organization; other regions: other national or regional offices).

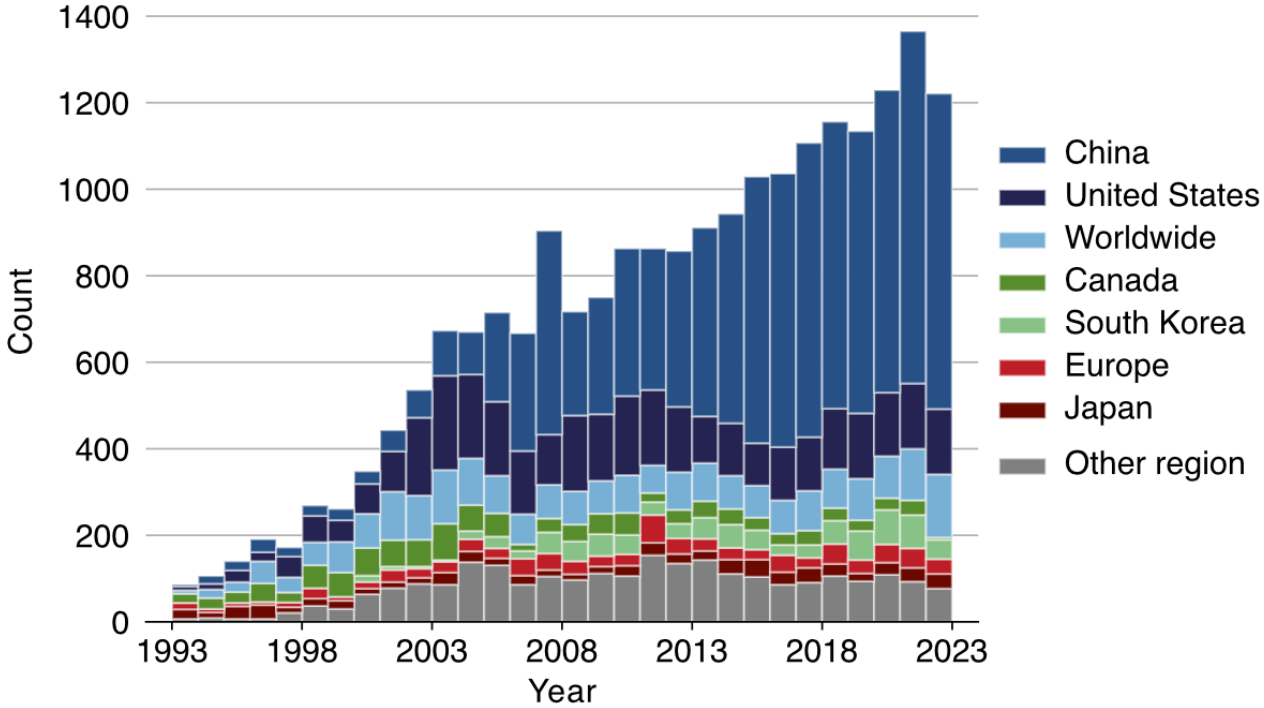

XSL•FO
RenderX

Figure 4. Overview of the (A) counts of patents and publications related to stroke over time and (B) year-on-year normalized patent and publication counts for each innovation cluster. Normalized counts were calculated by dividing the annual number of stroke-related patents or publications by the total number of patents or publications in that year, scaled to the maximum annual total across the study period. In both cases, the dashed lines depict exponential fits with the associated coefficients of determination (R^2) given per plot. AI: artificial intelligence.

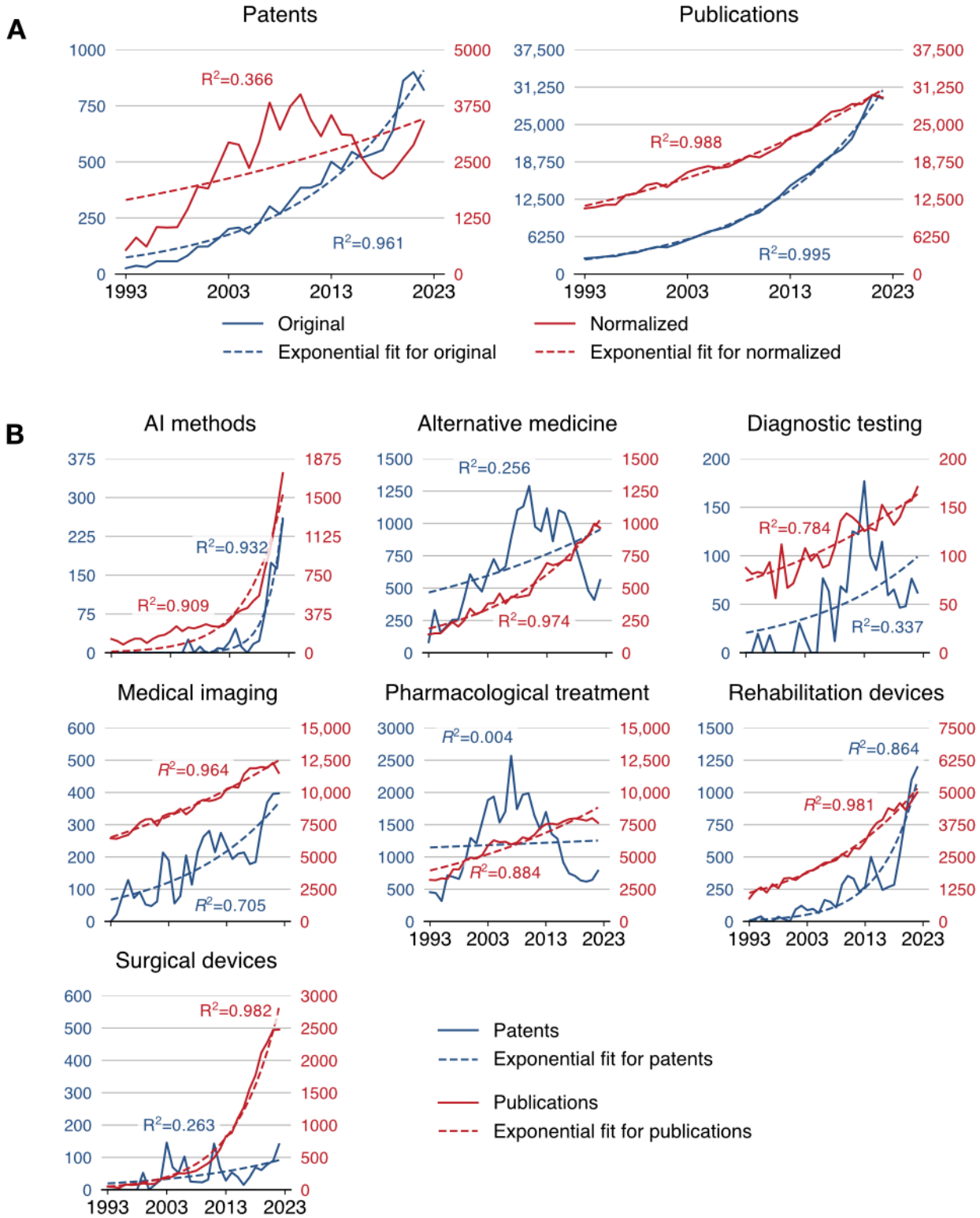

Leading Innovation Clusters

There were 7 top stroke-related innovation clusters identified over the last 30 years; interrater reliability between the 2 authors was high (Cohen κ=0.871). To address potential overcapture in the alternative medicine cluster due to broad search terms, a sensitivity analysis of 100 randomly sampled patents was performed, verifying that 97% were stroke related. The performance of these clusters, as measured using patents, is summarized in Table 4, with the allocation of patent codes provided in Multimedia Appendix 1. Pharmacological treatment was the largest cluster, accounting for 49.3% (36,005/73,094)

XSL•FO
RenderX

of the patents filed over the study period. To ensure the stability of these findings, a sensitivity analysis was performed confirming that cluster rankings held without normalization (data not shown). Within the last decade, only AI methods increased in rank, with the relative ordering of the other top-performing clusters remaining constant. Although not reflected in a change in order, the proportion accounted for by pharmacological treatments fell to 33.7% (9391/27,870), whereas all the other clusters increased their shares.

Table 4. Comparing the top-performing stroke-related innovation clusters by cumulative normalized patent counts over the past 30 years and the last decade. Artificial intelligence (AI) methods were the only cluster to increase in rank.

Rank	Innovation cluster	Normalized patent count, n (%)
1993-2023 (n=73,094)		
1	Pharmacological treatment	36,005 (49.3)
2	Alternative medicine	20,291 (27.8)
3	Rehabilitation devices	7777 (10.6)
4	Medical imaging	5331 (7.3)
5	Diagnostic testing	1448 (2.0)
6	Surgical devices	1397 (1.9)
7	AI methods	845 (1.2)
2013-2023 (n=27,870)		
1	Pharmacological treatment	9391 (33.7)
2	Alternative medicine	8005 (28.7)
3	Rehabilitation devices	5569 (20.0)
4	Medical imaging	2668 (9.6)
5	AI methods	791 (2.8)
6	Diagnostic testing	834 (3.0)
7	Surgical devices	612 (2.2)

Statistical Analysis

Figure 4 and Table 5 show the relationship between normalized patent and publication counts over time for the top-performing innovation clusters. There were strong associations (r_s>0.6; P<.006) between patent and publication rates for all clusters except pharmacological treatment (r_s=0.094; P=.99) and alternative medicine (r_s=0.546; P=.01). These 2 clusters showed normalized patents peaking in the late 2000s, with a continued shallow rise in publications. This trend is further illustrated by the normalized patent-to-publication ratio over time for each innovation cluster, as detailed in Multimedia Appendix 2. Plots of the data on AI methods show a rapid rise similar to that of other emerging clusters. To quantify this concurrent growth, the temporal correlation between the AI cluster and other leading clusters was assessed. AI patent activity was strongly correlated with patent activity in rehabilitation devices (r_s=0.767; P=.004) and medical imaging (r_s=0.662; P=.004). Similarly, AI publication rates correlated strongly with publication rates in rehabilitation devices (r_s=0.963; P=.004) and medical imaging (r_s=0.962; P=.004). These clusters all exhibited a strong exponential fit for both patent and publication data supported by coefficient of determination (R^2) values exceeding 0.7, as shown in Table 5. The rates of exponential growth for AI methods were the highest (39.2% per year for patents; 16.9% per year for publications), followed by rehabilitation devices (15.9% per year for patents; 5.3% per year for publications) and medical imaging (5.8% per year for patents; 2.2% per year for publications). The diffusion dynamics for each innovation cluster, approximated by fitting sigmoid curves to the patent and publication data, are shown in Figure 5 [18] and contextualized within the phases of the diffusion of innovations theory by Rogers [18], highlighting their positions in the innovation life cycle. The estimated progression through the innovation life cycle based on patent data was highest for pharmacological treatment (97.5%), followed by surgical devices (82.9%), nutritional and complementary therapies (71.4%), diagnostic testing (52.2%), AI methods (41.6%), medical imaging (30.5%), and rehabilitation devices (16.5%).

XSL•FO
RenderX

Table 5. Comparing the association between the normalized patent and publication counts for each innovation cluster along with the equations of the associated lines of the best exponential fit.

Innovation cluster	Patents		Publications		Association	
	Equation for the line with the best exponential fit	R^2	Equation for the line with the best exponential fit	R^2	r_s	*P* value
AI[a] methods	$3.7E\text{-}343e^{0.392x}$	0.932	$2.7E\text{-}146e^{0.169x}$	0.909	0.776	.001
Alternative medicine	$7.3E\text{-}20e^{0.025x}$	0.256	$2.2E\text{-}49e^{0.059x}$	0.974	0.546	.02
Diagnostic testing	$9.1E\text{-}47e^{0.054x}$	0.337	$1.9E\text{-}22e^{0.027x}$	0.784	0.645	.006
Medical imaging	$5.8E\text{-}50e^{0.058x}$	0.705	$2.6E\text{-}16e^{0.022x}$	0.964	0.767	.001
Pharmacological treatment	$8.5E\text{-}1e^{0.003x}$	0.004	$3.8E\text{-}21e^{0.028x}$	0.884	0.094	>.99
Rehabilitation devices	$9.1E\text{-}138e^{0.159x}$	0.864	$1.6E\text{-}43e^{0.053x}$	0.981	0.916	.001
Surgical devices	$5.2E\text{-}46e^{0.053x}$	0.263	$3.8E\text{-}120e^{0.140x}$	0.982	0.649	.001

[a]AI: artificial intelligence.

Figure 5. (A) The approximate innovation life cycle progression of each innovation cluster, calculated by fitting a sigmoid curve to patent and publication data, with 95% CIs estimated via nonparametric bootstrapping by repeatedly resampling the data and refitting the curve, which indicates their phase in the diffusion of innovation curve, and (B) the theoretical diffusion of innovation curve displaying both the cumulative diffusion (S-shaped curve) and rate of diffusion (bell-shaped curve) over time divided into the incubation (innovators and early adopters), growth (early majority and late majority), and saturation (laggards) phases. AI: artificial intelligence.

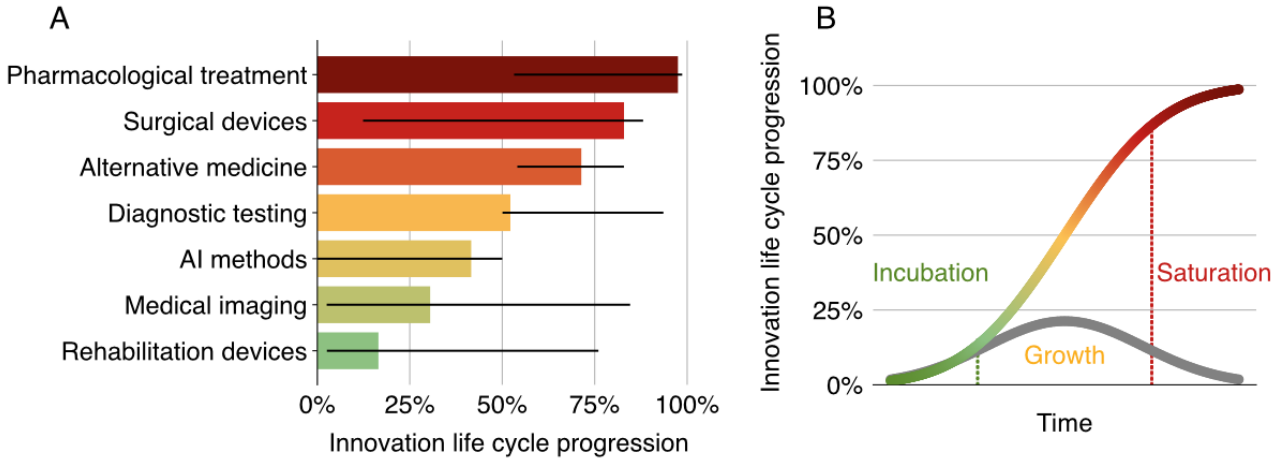

Discussion

Principal Findings

To the best of the authors' knowledge, this study provides the first rigorous quantification of innovation in stroke management. By using a scientifically validated framework, along with a novel application of an LLM, publicly available patent and publication data were analyzed. Among the top stroke-related innovation clusters identified, pharmacological treatment was found to be the most dominant over the past 30 years, accounting for nearly half (36,005/73,094, 49.3%) of all patents filed. AI methods, followed by rehabilitation devices and medical imaging, exhibited the highest rates of patent and publication growth, suggesting that these emerging innovation clusters are taking on increasingly important roles.

The diffusion of innovations theory by Rogers [18] describes the adoption curve of technology as a sigmoid function. This arises from the natural variation in the attitudes of individuals, ranging from early adopters to laggards, toward a new innovation. Previous studies have shown that this curve can also be applied to specific clusters themselves [4,20], indicating different phases of innovation: the incubation phase, where seminal work occurs and which is reflected by the initial rise in patenting and publication activity; the growth phase, where industry and clinicians drive innovation and which is associated with an exponential rise in patent and publication counts; and, finally, the saturation phase, which occurs when manufacturers refine the technology to maintain their competitive edge while continuing to pursue patents, leading to a plateau in both patent and publication activity. While these curves provide a model for technology diffusion, we acknowledge that this is a simplification. Real-world adoption in health care is complex and is further influenced by external factors such as regulatory approvals, reimbursement policies, and the development of clinical guidelines, which were not modeled in this study.

Applying this theory to this study, there was an exponential rise in the number of patents and publications for AI methods. This

XSL•FO
RenderX

study's novel quantification of this trend demonstrates that the patent growth rate for AI methods is approximately 2 to 7 times greater than that for the other key growth clusters of rehabilitation devices and medical imaging. This suggests that these clusters are all in the growth phase but that AI is accelerating at a substantially faster rate. There was a significant inflection point for AI methods observed in 2018, coinciding with the regulatory approval of the first AI software developed by Viz.ai [21]. Subsequently, the market has experienced a proliferation of commercially available software platforms designed to interpret and triage radiological data [22]. Therefore, the concurrent rise in medical imaging is to be expected given that these AI platforms, and indeed the main applications of AI in stroke [23-25], relate to analysis and interpretation of medical images. Similarly, while rehabilitation systems for stroke have yet to receive the same level of attention, there has been growing interest in full automation, with AI methods actively being researched [26].

The pattern for surgical devices presents a less clear trajectory: patent counts are leveling off, whereas publications continue to increase. This divergence could suggest that surgical device innovation remains at a nascent, exploratory stage: scholarly output continues to rise, whereas commercial patenting momentum has yet to follow. Alternatively, this pattern may relate to a limitation of the innovation discovery method used in this study in that patents for generic technologies that do not explicitly state stroke management may be incorrectly excluded. Thus, many patents for mechanical thrombectomy, being applicable to multiple diseases, may not be seen to parallel the rise in stroke thrombectomy publications.

Pharmacological treatments, alternative medicine, and diagnostic testing all experienced peaks in patent activity during the 2000s and have since plateaued, indicating that these sectors are now in the saturation phase. This finding for pharmacological treatments is particularly notable and perhaps unexpected given recent high-profile trials of novel drug classes. The rise in pharmacological treatments can be traced back to 1995 following the groundbreaking National Institute of Neurological Disorders and Stroke tissue-type plasminogen activator trial [1]. However, despite extensive efforts, the development of new therapeutics has become more challenging [27], and thus, a relative decrease in the number of patents and publications is perhaps unsurprising. Notably, this decline coincided with a decoupling between research output and patent activity (r_s=0.09; P=.99), suggesting that ongoing research in pharmacology increasingly focuses on avenues less amenable to patenting. Similarly, alternative medicine, despite seeing increased enthusiasm, particularly in addressing poststroke depression [28], is a well-established field whose roots predate orthodox medicine [29]. Diagnostic testing is also mature, with many of the key technologies, such as electrocardiogram and blood pressure monitors, able to be traced back to the 19th century [30,31]. As such, the observed trends appear in line with expectations.

The varying patent coefficient of determination (R^2) values across the innovation clusters provide insights into the reliability of their respective innovation trajectories. For AI methods (R^2=0.932), rehabilitation devices (R^2=0.864), and medical imaging (R^2=0.705), the high values signify a consistent and predictable exponential growth, reinforcing the conclusion that these are emerging technologies in a strong growth phase. In contrast, the lower values for patent trends in pharmacological treatments (R^2=0.004) and alternative medicine (R^2=0.256) suggest that an exponential model is less descriptive. This lower reliability is not a limitation of the data but rather an indicator that these fields have likely reached a saturation phase in which innovation, as measured via patent filings, is no longer accelerating at a consistent exponential rate.

The observed rise in AI-driven platforms, many of which operate on medical imaging data, has been linked to faster treatment times in acute stroke care. For example, a single-center study reported that the introduction of the Viz.ai software, which uses automated image interpretation to triage stroke cases, was associated with a 30-minute reduction in median door-to-needle time, alongside improvements in door-to-imaging and door-to-puncture intervals [32]. Similarly, innovations in rehabilitation devices, including robotic-assisted gait and upper-limb training, have demonstrated clinically meaningful improvements in motor function and activities of daily living in randomized controlled trials [33]. This demonstrated clinical value arguably provides a powerful mechanism that drives the investment, research, and patenting activity observed in this study. Taken together, these examples suggest that the observed growth in patents and publications within emerging innovation clusters reflects not only conceptual progress but also improvements in patient care and clinical workflows.

Comparison With Other Studies

There has been limited prior work evaluating innovation in stroke, with those studies focusing only on specific areas and generally being qualitative [7-9]. Martinez-Gutierrez et al [7] conducted a narrative review of developing technologies in the prehospital space in which they identified emergency medical service detection and triage of stroke as important areas for future advancements. The study also supported the potential of AI algorithms in these domains, aligning with the findings of our study. Recently, Ji et al [10] conducted a bibliometric analysis using Web of Science specifically for perioperative stroke over the past 20 years and found rapid growth in research publications addressing antiplatelet and antithrombotic therapy, cardiovascular surgery, and thrombectomy, among others. One notable difference between this study's results and those of our study is that the former identified pharmacological treatments and surgical devices as leading research areas. However, this discrepancy may be due to their use of absolute rather than normalized publication counts and their restriction to a narrow type of stroke as opposed to all causes, as in this study.

Within the field of health care more generally, Tran et al [34] performed a survey of health AI publications using Web of Science between 1977 and 2018 and found stroke to be a leading application area. The approach used in the aforementioned study has also been applied to neurosurgery [16] and surgery [4] as a whole. In both instances, and in keeping with the work presented in this paper, trends in patents and publications were consistent with the diffusion of innovations theory.

XSL•FO
RenderX

Strengths and Limitations

A key strength of this study lies in its use of a data-driven and quantitative framework to evaluate innovation in stroke management. This approach moves beyond traditional qualitative analyses by incorporating an LLM. This was critical for enabling the study's scale, as the manual filtering of 237,035 patents, of which a sample review found only 11.2% (56/500) to be relevant, would have been prohibitively labor-intensive. By enabling a detailed and extensive search across the breadth of stroke research fields, this study comprehensively quantified the current popularity and productiveness of innovation clusters and, by plotting changes in these metrics over time, can estimate where along the trajectory of innovation diffusion each cluster currently lies. Our results identify emerging technologies and may be useful metrics to inform policy and grant funding strategies. Our results also build on previous work [4,16] that underscores the value of using patent and publication data in the assessment of innovation. Despite their value, patent data have remained largely underused and underinvestigated [4].

Although this study used a novel approach to quantitatively evaluate innovation in stroke management, it is not without limitations. First, the methodology relied on patents as an indicator of technological innovation, potentially overlooking the output from individuals or organizations who lack the resources to apply for patents or choose not to for ethical or other reasons. Second, emerging or small-scale innovation clusters were unlikely to be identified through the method used, as they may be concealed within larger, more mature clusters. Third, patents for generic technological innovations that did not explicitly state their application to stroke were excluded from the analysis, even though they could still be applicable to stroke management. Fourth, it is possible that some inventors may deliberately delay academic publication until a patent has been granted, leading to an underestimation of recent innovations. Fifth, bias could be introduced due to the imperfect filtering of patents by the LLM, which was trained on a single-annotated dataset. Such limited annotation may have introduced misclassification errors, potentially overrepresenting clusters that use terminology closely aligned with the training labels while underrepresenting those that rely on novel or nuanced language. This may have affected both the sensitivity and specificity of cluster assignment. Future work building on this methodology should prioritize creating a ground-truth dataset with multiple, independent annotators to validate labels and further improve model generalizability. Sixth, the normalization method based on the work by Hughes-Hallett et al [4] is sensitive to the maximum value within the time series, which may introduce instability in the normalized metrics if future volumes differ substantially. Finally, the findings of this study were not validated against external clinical data. Such a validation would be a valuable next step to confirm whether the identified trends in patenting activity correlate with tangible improvements in stroke care and patient outcomes.

Conclusions

This is the first study to systematically use patent and publication data to quantitatively evaluate innovation in stroke care. Seven influential innovation clusters were identified over the 30-year study period, and their respective growth characteristics were found to be explainable by the diffusion of innovations theory. Looking ahead, the results suggest that AI methods, rehabilitation devices, and medical imaging are undergoing exponential growth and are forecasted to have a greater impact on stroke management. Furthermore, the methodology used in this work, particularly the novel use of an LLM, could be applied to assess more specific clusters and assist in decision-making for future research and funding.

Acknowledgments

The research team gratefully acknowledges the support and resources provided by the BioMedIA laboratory and Imperial College London. Generative artificial intelligence was not used in the creation of the manuscript text, figures, or tables. The large language model (Llama) discussed in this paper was an object of study and a tool for data filtering as described in the Methods section; it was not used to generate any part of this research paper.

Funding

This work was supported by the following grants and programs: the UK Research and Innovation Centre for Doctoral Training in AI for Healthcare [35] under grant EP/S023283/1, The Graham-Dixon Charitable Trust, the UK National Institute for Health and Care Research Invention for Innovation Programme under grant II-LA-0814-20007, he Imperial College Biomedical Research Centre, and the Economic and Social Research Council London Interdisciplinary Social Science Doctoral Training Partnership Collaborative Awards in Science and Engineering studentship.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the terms and conditions of the Open Patent Services web service but are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

AM and PB were responsible for the conceptualization of the study. AM, GL-T, and PB developed the methodology. AM conducted the formal analysis and data curation. AM wrote the original draft of the manuscript. All authors contributed to the writing, review, and editing of the final manuscript. PB and DR provided supervision.

Conflicts of Interest

PB is a cofounder of GripAble Ltd, a spin-out company from Imperial College London. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
The 100 top-performing patent codes retrieved by the search "stroke OR cerebrovascular" between 1993 and 2023 allocated to innovation clusters.
[DOCX File , 32 KB - jmir_v28i1e70754_app1.docx]

Multimedia Appendix 2
Year-on-year normalized patent-to-publication ratio for each innovation cluster.
[DOCX File , 92 KB - jmir_v28i1e70754_app2.docx]

References

1. National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. Tissue plasminogen activator for acute ischemic stroke. N Engl J Med 1995 Dec 14;333(24):1581-1587. [doi: 10.1056/NEJM199512143332401] [Medline: 7477192]
2. Goyal M, Menon BK, van Zwam WH, Dippel DW, Mitchell PJ, Demchuk AM, HERMES collaborators. Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. Lancet 2016 Apr 23;387(10029):1723-1731. [doi: 10.1016/S0140-6736(16)00163-X] [Medline: 26898852]
3. West MA. The social psychology of innovation in groups. In: West MA, Farr JL, editors. Innovation and Creativity at Work: Psychological and Organizational Strategies. Hoboken, NJ: John Wiley & Sons; 1990:309-333.
4. Hughes-Hallett A, Mayer EK, Marcus HJ, Cundy TP, Pratt PJ, Parston G, et al. Quantifying innovation in surgery. Ann Surg 2014 Aug;260(2):205-211 [FREE Full text] [doi: 10.1097/SLA.0000000000000662] [Medline: 25350647]
5. Daim TU, Rueda G, Martin H, Gerdsri P. Forecasting emerging technologies: use of bibliometrics and patent analysis. Technol Forecast Soc Change 2006 Oct;73(8):981-1012. [doi: 10.1016/j.techfore.2006.04.004]
6. Campbell B. How to judge the value of innovation. BMJ 2012 Mar 07;344(mar07 1):e1457. [doi: 10.1136/bmj.e1457] [Medline: 22399703]
7. Martinez-Gutierrez JC, Chandra RV, Hirsch JA, Leslie-Mazwi T. Technological innovation for prehospital stroke triage: ripe for disruption. J Neurointerv Surg 2019 Nov 14;11(11):1085-1090. [doi: 10.1136/neurintsurg-2019-014902] [Medline: 31201289]
8. Verma A, Towfighi A, Brown A, Abhat A, Casillas A. Moving towards equity with digital health innovations for stroke care. Stroke 2022 Mar;53(3):689-697 [FREE Full text] [doi: 10.1161/STROKEAHA.121.035307] [Medline: 35124973]
9. Filler A. The history, development and impact of computed imaging in neurological diagnosis and neurosurgery: CT, MRI, and DTI. Nat Preced 2009 Jun 30;1:1-76. [doi: 10.1038/npre.2009.3267]
10. Ji S, Shi Y, Fan X, Jiang T, Yang X, Tao T, et al. Global trends in perioperative stroke research from 2003 to 2022: a web of science-based bibliometric and visual analysis. Front Neurol 2023;14:1185326 [FREE Full text] [doi: 10.3389/fneur.2023.1185326] [Medline: 37325224]
11. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: scaling language modeling with pathways. arXiv Preprint posted online April 5, 2022 [FREE Full text] [doi: 10.5555/3648699.3648939]
12. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: open and efficient foundation language models. arXiv Preprint posted online February 27, 2023 [FREE Full text] [doi: 10.48550/arXiv.2302.13971]
13. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
14. Coverage, codes and statistics. European Patent Office. URL: https://www.epo.org/en/searching-for-patents/data/coverage [accessed 2025-05-29]
15. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. arXiv Preprint posted online June 17, 2021 [FREE Full text] [doi: 10.48550/arXiv.2106.09685]
16. International Patent Classification (IPC). World Intellectual Property Organization. URL: https://www.wipo.int/en/web/classification-ipc [accessed 2025-05-29]

XSL•FO
RenderX

17. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with Python. In: Proceedings of the 9th Python in Science Conference. 2010 Presented at: SciPy '10; June 28-July 3, 2010; Austin, TX URL: https://proceedings.scipy.org/articles/Majora-92bf1922-011 [doi: 10.25080/majora-92bf1922-011]
18. Rogers EM. Diffusion of Innovations. 4th edition. New York, NY: Simon and Schuster; 2010.
19. Campbell G. Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. Stat Med 1994 Oct 15;13(5-7):499-508. [doi: 10.1002/sim.4780130513] [Medline: 8023031]
20. Marcus HJ, Hughes-Hallett A, Kwasnicki RM, Darzi A, Yang GZ, Nandi D. Technological innovation in neurosurgery: a quantitative study. J Neurosurg 2015 Jul;123(1):174-181 [FREE Full text] [doi: 10.3171/2014.12.JNS141422] [Medline: 25699414]
21. Petrone JJ. FDA approves stroke-detecting AI software. Nat Biotechnol 2018 Apr 05;36(4):290. [doi: 10.1038/nbt0418-290] [Medline: 29621226]
22. Soun J, Chow DS, Nagamine M, Takhtawala RS, Filippi CG, Yu W, et al. Artificial intelligence and acute stroke imaging. AJNR Am J Neuroradiol 2021 Jan;42(1):2-11 [FREE Full text] [doi: 10.3174/ajnr.A6883] [Medline: 33243898]
23. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec 21;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
24. Mouridsen K, Thurner P, Zaharchuk G. Artificial intelligence applications in stroke. Stroke 2020 Aug;51(8):308-331. [doi: 10.1161/strokeaha.119.027479]
25. Lee EJ, Kim YH, Kim N, Kang DW. Deep into the brain: artificial intelligence in stroke imaging. J Stroke 2017 Sep;19(3):277-285 [FREE Full text] [doi: 10.5853/jos.2017.02054] [Medline: 29037014]
26. Rahman S, Sarker S, Haque AK, Uttsha MM, Islam MF, Deb S. AI-driven stroke rehabilitation systems and assessment: a systematic review. IEEE Trans Neural Syst Rehabil Eng 2023;31:192-207. [doi: 10.1109/TNSRE.2022.3219085] [Medline: 36327176]
27. Paul S, Candelario-Jalil E. Emerging neuroprotective strategies for the treatment of ischemic stroke: an overview of clinical and preclinical studies. Exp Neurol 2021 Jan;335:113518 [FREE Full text] [doi: 10.1016/j.expneurol.2020.113518] [Medline: 33144066]
28. Wijeratne T, Sales C, Wijeratne C. A narrative review on the non-pharmacologic interventions in post-stroke depression. Psychol Res Behav Manag 2022 Jul 07;15:1689-1706 [FREE Full text] [doi: 10.2147/PRBM.S310207] [Medline: 35832139]
29. Steyer TE. Complementary and alternative medicine: a primer. Fam Pract Manag 2001 Mar;8(3):37-42 [FREE Full text] [Medline: 11317848]
30. Burch GE, DePasquale NP. A History of Electrocardiography. New York, NY: Norman Publishing; 1990.
31. Booth J. A short history of blood pressure measurement. Proc R Soc Med 1977 Nov;70(11):793-799 [FREE Full text] [doi: 10.1177/003591577707001112] [Medline: 341169]
32. Milfred F, Ami Roy A, Delmor C, Bansal A, Gifford K, Ma K, et al. Abstract TP89: advancing stroke care efficiency: impact of AI and communication tools on patient outcomes. Stroke 2025 Feb;56(Suppl_1):ATP89. [doi: 10.1161/str.56.suppl_1.tp89]
33. Mehrholz J, Kugler J, Pohl M, Elsner B. Electromechanical-assisted training for walking after stroke. Cochrane Database Syst Rev 2025 May 14;5(5):CD006185. [doi: 10.1002/14651858.CD006185.pub6] [Medline: 40365867]
34. Tran BX, Vu GT, Ha GH, Vuong QH, Ho MT, Vuong TT, et al. Global evolution of research in artificial intelligence in health and medicine: a bibliometric study. J Clin Med 2019 Mar 14;8(3):360 [FREE Full text] [doi: 10.3390/jcm8030360] [Medline: 30875745]
35. UKRI Centre for Doctoral Training in AI for Healthcare. AI4Health. URL: https://ai4health.io/ [accessed 2026-01-12]

Abbreviations

AI: artificial intelligence
IPC: International Patent Classification
LLM: large language model
LoRA: low-rank adaptation
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

XSL•FO
RenderX

Edited by A Coristine; submitted 01.Jan.2025; peer-reviewed by N Raju, S Sivarajkumar, C Wang, R Yang; comments to author 01.Apr.2025; revised version received 23.Dec.2025; accepted 23.Dec.2025; published 20.Jan.2026.

Please cite as:
Marcus A, Lockwood-Taylor G, Rueckert D, Bentley P
Quantifying Innovation in Stroke: Large Language Model Bibliometric Analysis
J Med Internet Res 2026;28:e70754
URL: https://www.jmir.org/2026/1/e70754
doi:10.2196/70754
PMID:

©Adam Marcus, Georgina Lockwood-Taylor, Daniel Rueckert, Paul Bentley. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Automated Prediction of Glasgow Coma Scale Scores From Unstructured Electronic Health Records Using Natural Language Processing: Development and Validation Study

Marta Fernandes[1], PhD; Niels Turley[2], MSc; Haoqi Sun[2], PhD; Shibani S Mukerji[1], MD, PhD; Lidia M V R Moura[1], MD, MPH, PhD; M Brandon Westover[2], MD, PhD; Sahar F Zafar[1], MD, MSc

[1]Department of Neurology, Massachusetts General Hospital, 55 Fruit St, Boston, MA, United States

[2]Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA, United States

Corresponding Author:
Marta Fernandes, PhD
Department of Neurology, Massachusetts General Hospital, 55 Fruit St, Boston, MA, United States

Abstract

Background: Multicenter electronic health records (EHRs) can support quality improvement and comparative effectiveness research in critical care. However, limitations of EHR-based research include challenges in abstracting key clinical variables, including a patient's level of consciousness.

Objective: This study aimed to develop a natural language processing model to predict Glasgow Coma Scale (GCS) scores from daily EHR notes.

Methods: The study included adult patients (aged ≥18 years) admitted to Mass General Brigham (MGB) hospitals (2017 - 2024) and patients from the Medical Information Mart for Intensive Care-III (MIMIC-III version 1.4; 2001 - 2012) database. A dataset of all patients from both institutions was split into training (70%) or hold-out test (30%) sets. Variables consisted of daily notes, age, sex, and admission type. A pooled ordinal regression model (ordinalNet) with an elastic net penalty was trained to predict the lowest daily level of consciousness across 3 classes of impairment: severe (GCS score 3 - 8), moderate (GCS score 9 - 12), and mild (GCS score 13 - 15), and a pooled linear model was trained to predict continuous GCS scores (3-15). Gold standard GCS was obtained from structured flowsheet data. External generalizability was assessed using a single-institution ordinal model trained on MGB and tested on MIMIC. Following post hoc calibration, the performance of the ordinal and linear models was evaluated on the hold-out test sets using the area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) for the ordinal models and root mean square error and Pearson correlation coefficient for the linear models.

Results: The modeling cohort included 145,897 patients (MGB: n=123,257, MIMIC: n=22,640), with 1,446,965 days of hospitalization between training and testing sets; the average age was 62 (SD 18) years, and the sex distribution was balanced. The pooled ordinalNet achieved an AUROC of 0.96 (95% CI 0.96 - 0.96) and an AUPRC of 0.77 (95% CI 0.76 - 0.77). The single-institution ordinal model achieved an AUROC of 0.90 (95% CI 0.89 - 0.90) and an AUPRC of 0.80 (95% CI 0.79 - 0.80). The pooled linear model achieved a root mean square error of 2.30 (95% CI 2.30 - 2.30) and a correlation of 0.76 (95% CI 0.76 - 0.76). Predictions for severe GCS were driven by terms indicating unresponsiveness and critical interventions, moderate GCS by intermediate alertness descriptors, and mild GCS by mentions of normal or awake behavior.

Conclusions: Pooled ordinal and linear models can accurately predict GCS from unstructured data and can support large-scale phenotyping of neurological assessments for future critical care research.

(J Med Internet Res 2026;28:e81245) doi:10.2196/81245

KEYWORDS

Glasgow Coma Scale; electronic health records; phenotyping; natural language processing; machine learning

Introduction

The Glasgow Coma Scale (GCS) is used widely to characterize the level of consciousness in critically ill medical, surgical, neurological, and trauma patients [1]. It is a key component of most critical illness disease severity scores, including the Acute Physiology and Chronic Health Evaluation II score, the Sequential Organ Failure Assessment score, and the Simplified Acute Physiology score [2-4]. It is also used to define disease severity in patients with traumatic brain injuries [5,6]. GCS correlates strongly with morbidity and mortality in critically ill patients and in patients with traumatic brain injuries [7,8]. Therefore, GCS serves as a key variable in critical care research

XSL•FO
RenderX

studies, quality improvement initiatives, epidemiological studies, and population health research [9,10]. However, GCS is often missing in electronic health records (EHRs), posing challenges in defining illness severity and calculating key illness severity scores [11]. GCS is also frequently missing in trauma studies [12]. Often, imputation methods are used to account for missing GCS data in calculating critical illness severity scores [13,14]. In addition, GCS frequently fluctuates, and documented scores may not capture the fluctuations [14]. Finally, while GCS is commonly documented in structured flowsheets, often greater details on component scores and fluctuations are provided in unstructured EHR data [15].

The goal of this study was to develop an automated natural language processing (NLP) model capable of estimating GCS scores from unstructured clinical text, allowing inference of neurological status in clinical notes where structured GCS documentation is incomplete or missing. This can support assessment of disease severity in EHR-based population health science research and quality improvement and comparative effectiveness studies, as well as cohort identification for clinical trials in patients with neurologic injury or disorders of consciousness.

Methods

Study Population

We included adult patients (aged ≥18 years) with an inpatient hospital admission. This study consists of retrospective data analysis and is reported in accordance with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Artificial Intelligence (TRIPOD-AI) statement [16].

Ethical Considerations

The study was approved by the Mass General Brigham (MGB) Institutional Review Board (protocol 2013P001024). A waiver of informed consent was obtained for this observational study.

Datasets

Our cohort was derived from two sources: (1) inpatient admissions to the MGB health care system between January 2017 and May 2024 and (2) admissions recorded in the publicly available, deidentified Medical Information Mart for Intensive Care-III (MIMIC-III) database, version 1.4 [17], which contains medical records for intensive care unit admissions at Beth Israel Deaconess Medical Center between 2001 and 2012. The cohort included adult patients with daily clinical notes and at least 1 daily assessment of the total GCS score.

Clinical Variables

Structured variables included age, sex, and admission types (emergency, urgent, and elective). Text-based variables were extracted from preprocessed daily clinical notes and binarized (Multimedia Appendix 1). MGB notes included procedures, nursing, case management, consults, assessment and plan, history and physical, progress, physical therapy, occupational therapy, and discharge summaries. MIMIC notes included electrocardiogram, echocardiogram, radiology, respiratory, nutrition, general, physician, nursing, rehab services, case management, consult, and discharge summaries. For both institutions, inpatient days with only the following note types, either alone or combined, were removed: procedures, radiology, electrocardiogram, echocardiogram, respiratory, and nutrition. These note types were considered to contain minimal information regarding the patient's neurological state.

To avoid conflating pathologic coma with iatrogenic sedation and enhance clinical validity, text features containing sedative and anesthetic medications (eg, propofol or other sedatives or anesthetics) were removed, ensuring the model focused on physiological descriptions of consciousness.

Outcomes and Gold Standard

Our outcome was the lowest daily GCS score for each day of hospital admission. For analysis, the lowest daily GCS score was categorized as severe (GCS score 3 - 8), moderate (GCS score 9 - 12), and mild (GCS score 13 - 15). For both MGB and MIMIC cohorts, the gold standard scores were obtained from structured information tables.

Statistical Analysis

Data from each institution was randomly split by patient into training (70%) and hold-out test (30%) sets. The training sets from both MGB and MIMIC, as well as their respective test sets, were combined to create pooled training and testing sets for a single, multi-institution model. A combination of undersampling and oversampling strategies was applied within the training set to prevent 1 institution from dominating the pooled model and to address class imbalance and mitigate potential bias toward the majority class (Multimedia Appendix 1).

Model Design

We developed an ordinal regression model with elastic net penalty [18] (ordinalNet) within the training data to predict the 3 classes of GCS scores: mild (13-15), moderate (9-12), and severe (3-8). We also developed a linear regression model to predict the full range of GCS scores: 3 - 15. Data were split into train and test and resampled similar to the ordinal regression (Multimedia Appendix 1).

We additionally evaluated cross-institution generalizability by training a single-institution model using data from MGB and testing it on MIMIC data (Multimedia Appendix 1). This approach allowed us to assess the generalizability of models trained on data from a single center compared with models trained on multicenter data.

Model Evaluation

For model evaluation, we used the area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), sensitivity (or recall), specificity, positive predictive value, negative predictive value, and F_1-score. We present the macroaverage performance, which calculates each metric independently for each class and then takes the average.

Calibration was performed post hoc using multinomial logistic regression fitted on the model-predicted GCS class probabilities. The calibration model was fit on the training data only and

XSL•FO
RenderX

evaluated on the hold-out test set. Calibration was evaluated using reliability diagrams comparing predicted and observed class probabilities. Model calibration was assessed using both the multiclass Brier score, which treats each class as nominal, and the ordinal Brier score, which accounts for the ordering of the classes. A lower Brier score indicates better calibrated predictions.

A total of 1000 bootstrap iterations with random sampling and replacement were performed to calculate 95% CIs for performance metrics. We report overall results for the pooled institution model on the combined hold-out test set and separately by institution. We additionally report results for the single-institution model.

Results

Patients' Characteristics

Our study cohort included 145,897 patients (MGB=123,257, MIMIC=22,640), with a total of 1,446,965 days of inpatient admissions. The average age of patients in our study cohort was 62 (SD 18) years, and the cohort had a balanced sex distribution (n=75,866, 52%). The patients' race and ethnicity distribution is shown in Table 1. These variables were collected for descriptive purposes and not used for analysis. Overall, the multicenter train and hold-out test sets were balanced at the patient level, with a majority of emergency admissions in both sets. After applying inclusion and exclusion criteria (Figure 1), we observed a higher proportion of exclusion in the MGB data for patients who did not present all 3 GCS components—visual, verbal, and motor—thus not a total GCS score. In the MIMIC data, a higher proportion of exclusion was for inpatient days including only procedure notes. To improve readability, the flow diagram reports total numbers after exclusion criteria, while detailed exclusion counts at patient and admission levels by criterion are provided in Table S2 in Multimedia Appendix 1. We performed a sensitivity analysis for notes length threshold and selected a minimum of 300 words per day for model training (Tables S3 and S4 in Multimedia Appendix 1).

We observed that MGB test data are heavily skewed toward mild cases (n=116,028, 89%), with very few severe or moderate cases. MIMIC test data are more balanced, with substantially higher proportions of moderate and severe cases than the MGB test data (Table 1).

XSL•FO
RenderX

Table . Characteristics of the study population.

Characteristic	Training set			Testing set		
	MGB[a] (n=26,718)	MIMIC[b] (n=14,022)	Total (n=40,740)	MGB (n=96,539)	MIMIC (n=8618)	Total (n=105,157)
Age (years), mean (SD)	64 (17)	64 (17)	64 (17)	62 (18)	63 (18)	62 (18)
Sex (male), n (%)	14,598 (55)	7954 (57)	22,552 (55)	48,784 (51)	4827 (56)	53,611 (51)
Race[c], n (%)						
Asian	931 (3)	315 (2)	1246 (3)	3063 (3)	186 (2)	3249 (3)
Black or African American	2354 (9)	1043 (7)	3397 (8)	6914 (7)	642 (7)	7556 (7)
Hispanic or Latino	2127 (8)	423 (3)	2550 (6)	7876 (8)	271 (3)	8147 (8)
White	20,567 (77)	10,175 (73)	30,742 (76)	77,286 (80)	6238 (72)	83,524 (80)
Other[d]	2866 (11)	2489 (18)	5355 (13)	9276 (10)	1552 (18)	10,828 (10)
Hospital daily stays[e], n	45,000	45,000	90,000	1,308,700	48,265	1,356,965
Same day death[e,f], n (%)	764 (2)	917 (2)	1681 (2)	4265 (0.3)	1048 (2)	5313 (0.4)
Admission type[e], n (%)						
Emergency	28,933 (64)	38,344 (85)	67,277 (75)	865,201 (66)	40,931 (85)	906,132 (67)
Urgent	9048 (20)	1693 (4)	10,741 (12)	202,175 (15)	5478 (4)	207,653 (15)
Elective	8093 (18)	4963 (11)	13,056 (15)	255,523 (20)	4029 (11)	259,552 (19)
GCS[e,g] classes, n (%)						
Severe (3-8)	15,000 (33)	15,000 (33)	30,000 (33)	64,891 (5)	15,407 (32)	80,298 (6)
Moderate (9-12)	15,000 (33)	15,000 (33)	30,000 (33)	83,526 (6)	13,036 (27)	96,562 (7)
Mild (13-15)	15,000 (33)	15,000 (33)	30,000 (33)	1,160,283 (89)	19,822 (41)	1,180,105 (87)

[a]MGB: Mass General Brigham.

[b]MIMIC: Medical Information Mart for Intensive Care.

[c]Race, ethnicity, and same day death are presented for descriptive purposes to characterize the cohort; these variables were not used as predictors in the modeling analyses.

[d]Includes American Indian or Alaska Native, multirace, and unknown race.

[e]Number of daily inpatient stays.

[f]Death on the same day of inpatient day.

[g]GCS: Glasgow Coma Scale.

XSL•FO
RenderX

Figure 1. CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence) flow diagram. The number of patients is represented by '"n," the number of hospital admissions by "N," and the number of daily inpatient stays by "D." GCS: Glasgow Coma Scale; MGB: Mass General Brigham; MIMIC: Medical Information Mart for Intensive Care.

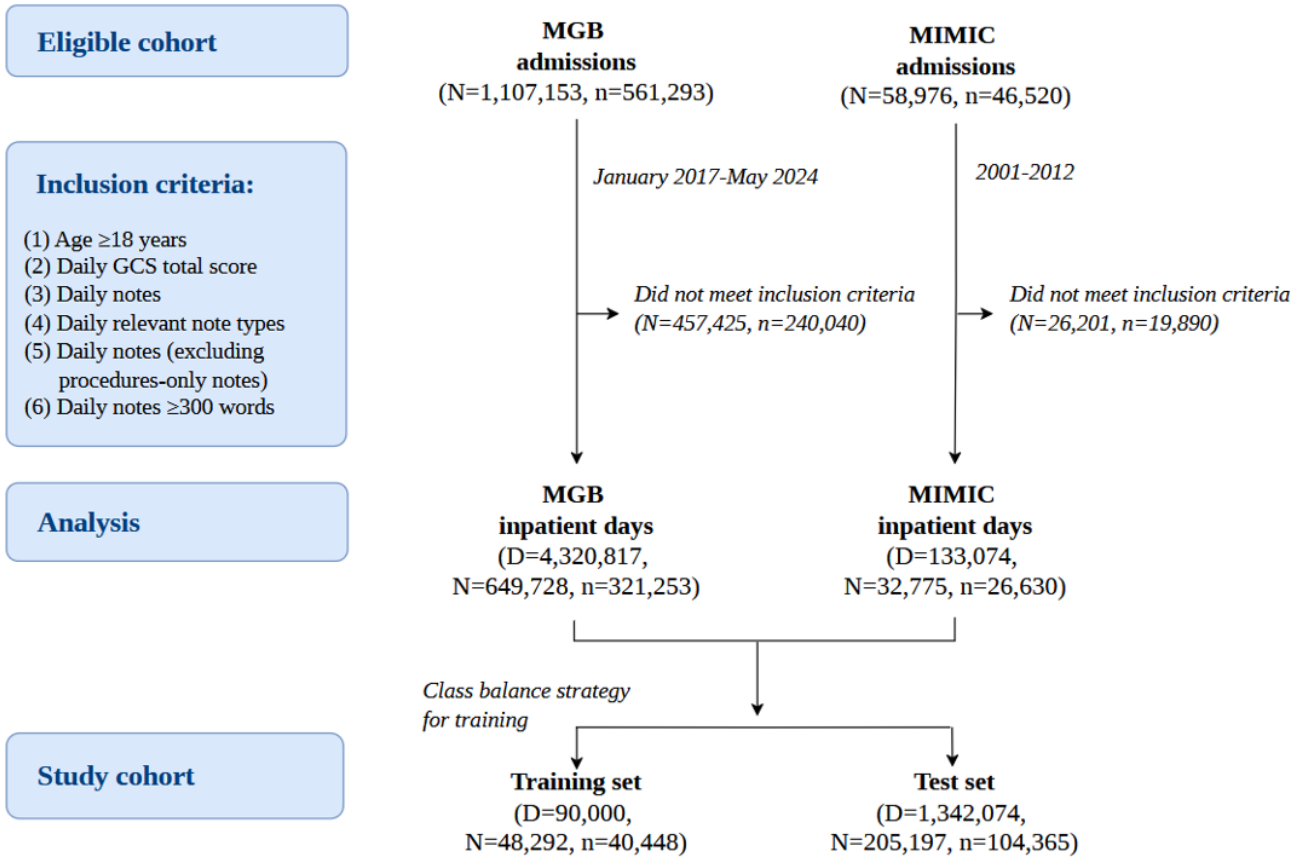

Pooled Ordinal Model Performance

Performance of the pooled model was evaluated separately for overall multicenter MGB+MIMIC data and separately for each institution using postcalibration predictions. The overall pooled ordinal model achieved an AUROC of 0.96 (95% CI 0.96 - 0.96) and an AUPRC of 0.77 (95% CI 0.76 - 0.77) after calibration (Figure 2). When evaluated on each institution test set, the model achieved a combined recall of 0.80 (95% CI 0.79 - 0.81; for MGB+MIMC), while MIMIC achieved an AUPRC of 0.87 (95% CI 0.86 - 0.87), 10% higher than MGB (0.76, 95% CI 0.75 - 0.77). Confusion matrices revealed that most misclassification occurred in the moderate class (Figure 3). MGB was heavily skewed toward mild cases (n=116,028, 89% of inpatient days), as seen in Table 1, resulting in the majority of predictions being correctly classified as mild, with relatively few moderate and severe cases. MIMIC had a more balanced distribution across mild, moderate, and severe classes, which led to a higher absolute number of misclassifications between moderate and severe, despite comparable proportional performance. These differences in the confusion matrices are therefore largely explained by the underlying class distributions in the test sets rather than differences in model performance.

XSL•FO
RenderX

Figure 2. Macroaverage performance of the pooled model on the hold-out test set for the Glasgow Coma Scale prediction: (A) across combined ordinal classes and (B) stratified by individual ordinal classes AUROC: area under the receiver operating characteristic; AUPRC: area under the precision-recall curve; MGB: Mass General Brigham; MIMIC: Medical Information Mart for Intensive Care.

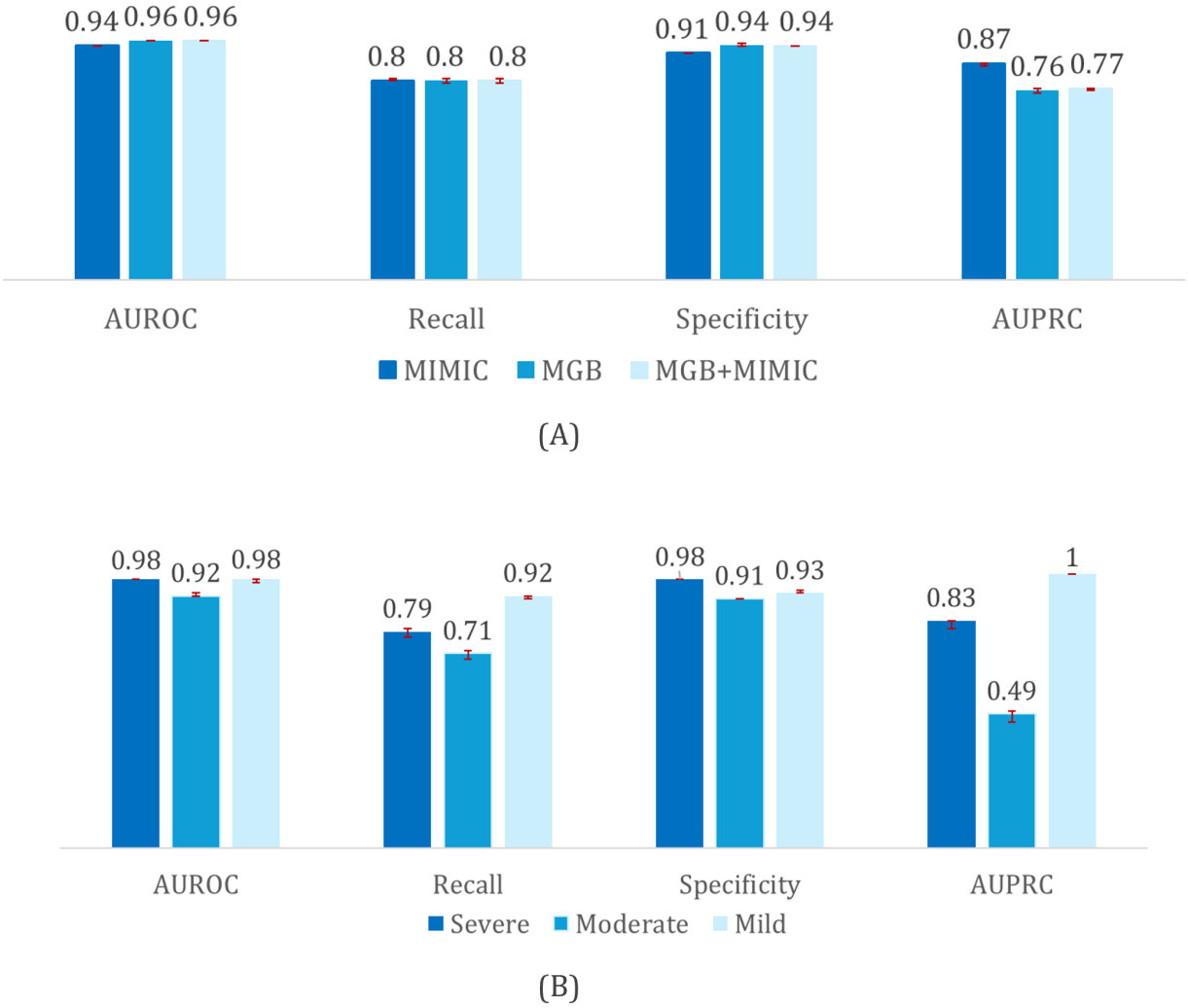

XSL•FO
RenderX

Figure 3. Confusion matrices of the pooled model on the hold-out test set: (A) full test set, (B) Mass General Brigham, and (C) Medical Information Mart for Intensive Care.

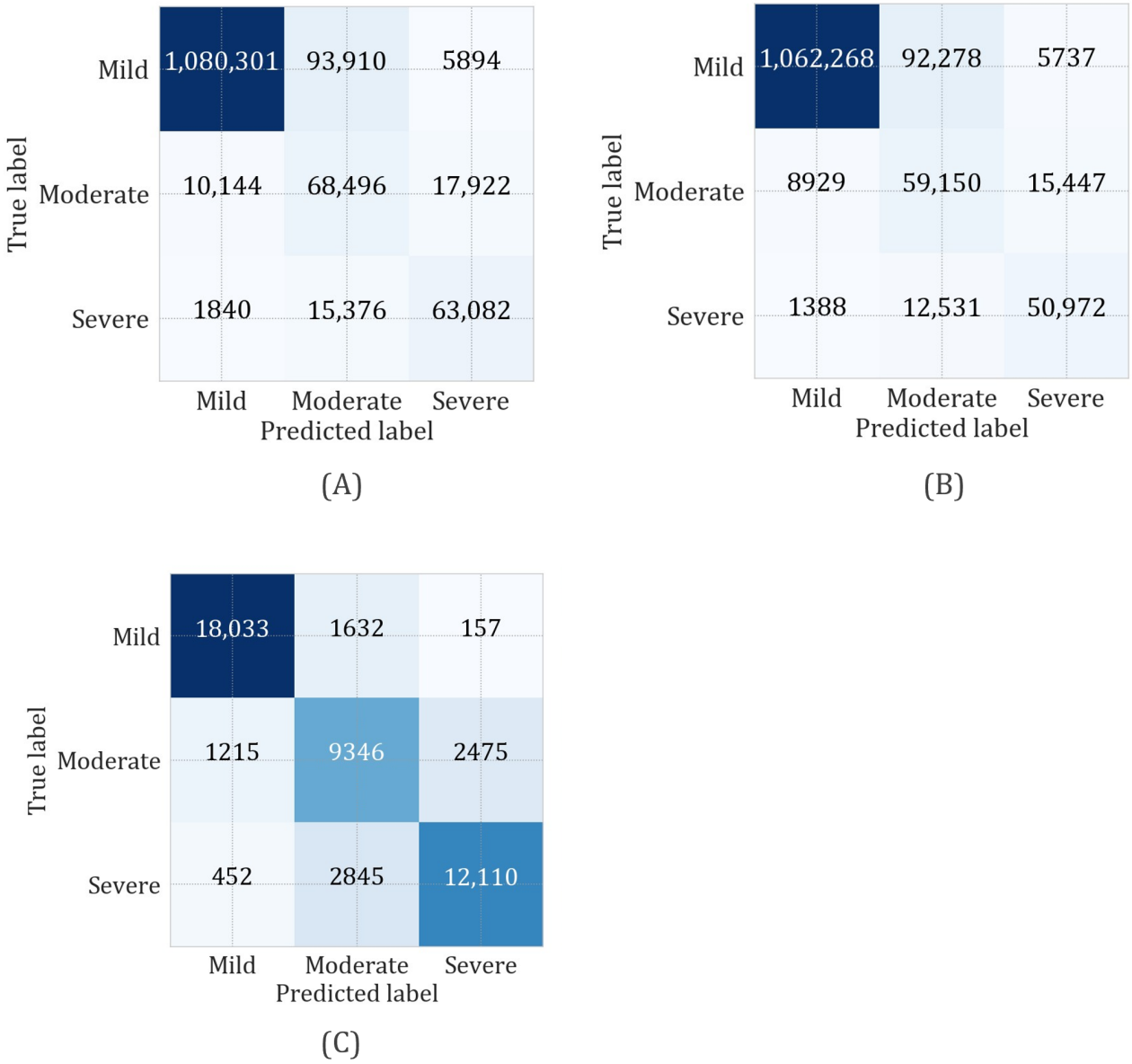

The pooled ordinal model's AUROC and AUPRC plots are presented in Figures S1 to S3 in Multimedia Appendix 1, macroaveraged performance across all metrics before and after calibration are reported in Tables S6 and S7 in Multimedia Appendix 1, and the modeling parameters are provided in Table S8 in Multimedia Appendix 1. All performance metrics reported here (area under the curve, recall, specificity, AUPRC, and confusion matrices) are based on postcalibration predictions.

Ordinal Model Calibration

We assessed the reliability of calibration per institution, as shown in Figure 4. Prior to calibration, the model is systematically overconfident at higher predicted probabilities. After calibration, predicted probabilities are better aligned with observed accuracies, indicating improved probabilistic calibration. For MIMIC, calibration curves align more closely with the ideal diagonal across the full probability range. Calibration produced minimal changes in class-specific performance metrics (Tables S6 and S7 in Multimedia Appendix 1), suggesting that the original ordinalNet probabilities were already well calibrated.

XSL•FO
RenderX

Figure 4. Calibration of ordinal class probabilities on the hold-out test set: institution-specific reliability diagrams for (A) Mass General Brigham and (B) Medical Information Mart for Intensive Care.

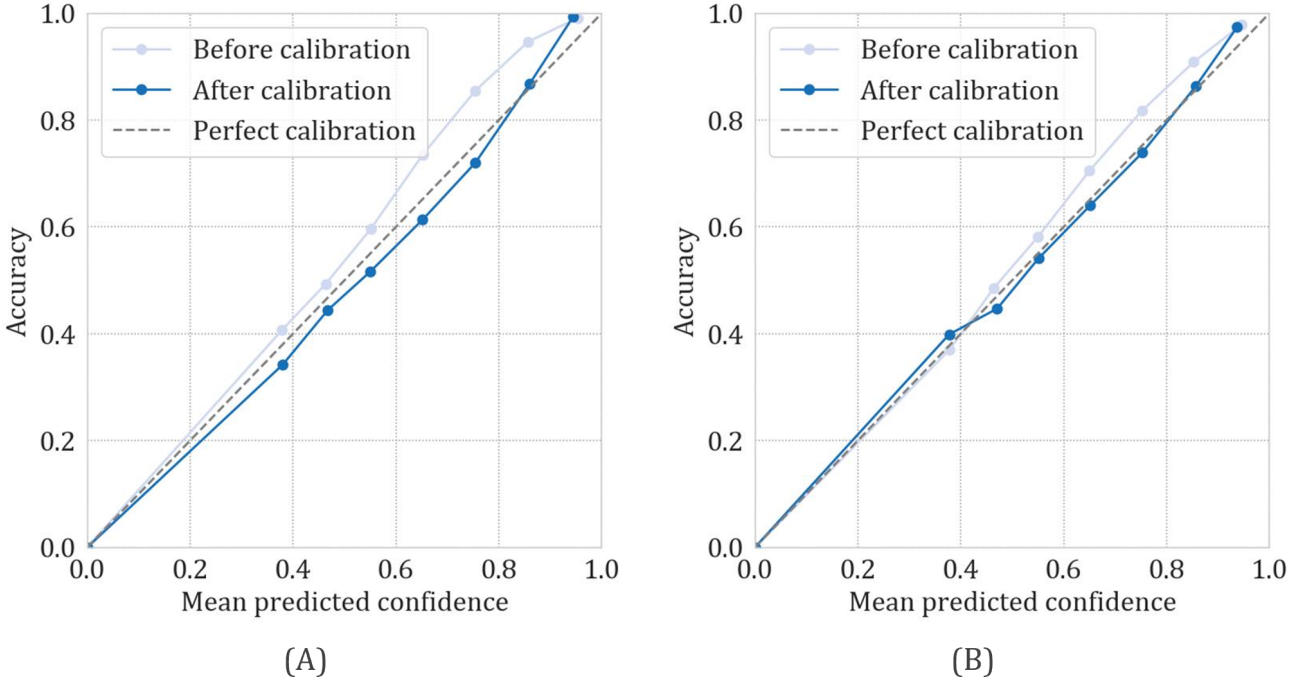

We also evaluated Brier scores before and after calibration and observed that the ordinal Brier score for the pooled ordinal model remained unchanged, with a value of 0.10 (95% CI 0.09 - 0.10). Institution-specific Brier scores are reported in Table S9 in Multimedia Appendix 1.

Single-Institution Ordinal Model Performance

The single-institution model was trained on MGB train set data (number of daily inpatient stays=45,000, n=26,718; Table 1) and externally evaluated on the MIMIC full test set (number of daily inpatient stay=160,375, n=28,725). The model achieved an AUROC of 0.90 (95% CI 0.89 - 0.90), a recall of 0.72 (95% CI 0.71 - 0.72), a specificity of 0.87 (95% CI 0.86 - 0.87), and an AUPRC of 0.80 (95% CI 0.79 - 0.80). Performance results for each ordinal label are also presented in Tables S6 and S7 in Multimedia Appendix 1. AUROC and AUPRC plots, as well as the confusion matrix, are shown in Figure S4 in Multimedia Appendix 1.

Post hoc multinomial calibration slightly reduced discrimination (recall 0.73 vs 0.72; Tables S6 and S7 in Multimedia Appendix 1) and slightly increased Brier error (0.23 vs 0.24; Table S8 in Multimedia Appendix 1), suggesting that the original ordinalNet probabilities were already well calibrated within the institutional distribution.

Linear Model Performance

We evaluated the performance of the linear model on the same hold-out test set as the ordinal model. Details of cohort derivation are presented in the CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence) modeling charts in Figure S5 in Multimedia Appendix 1. The pooled linear model achieved a Pearson correlation of 0.76 (root mean square error [RMSE] 2.40, 95% CI 2.40-2.41) on the hold-out test set (Table S5 in Multimedia Appendix 1). When evaluated separately, Pearson correlation was 0.74 for MGB (RMSE 2.40, 95% CI 2.40-2.41) and 0.82 for MIMIC (RMSE 2.47, 95% CI 2.45-2.48). Post hoc calibration improved absolute prediction accuracy (RMSE 2.30, 95% CI 2.30-2.30), while the Pearson correlation coefficient remained unchanged, consistent with the invariance of correlation to linear transformations.

Pooled linear model calibrated predictions vs target GCS scores are presented in Figure S6 in Multimedia Appendix 1, and the corresponding calibration curve is shown in Figure S7 in Multimedia Appendix 1.

Error Analysis

We conducted a detailed analysis of misclassifications to identify potential limitations and areas for improvement in our model.

One source of model error was occasional inconsistency between the time of gold standard GCS measurements and the actual worst neurological condition of patients. In some cases, patients who deteriorated rapidly and died on the same day had only a single GCS score of 15 recorded early in the day, creating a mismatch between the documented score and the patient's final clinical state. We retained these cases in our analysis to maintain a clinically representative dataset, acknowledging that excluding such edge cases would introduce selection bias.

Variability in measurement frequency posed another challenge. Some inpatient days had only a single daily GCS measurement, while others had multiple assessments. We chose not to restrict our analysis based on the number of measurements to avoid introducing sampling bias, as patients with more frequent monitoring might represent a clinically distinct population.

For patients with fluctuating consciousness levels throughout the day, the GCS documented in clinical notes often did not correspond to the lowest recorded gold standard GCS score.

XSL•FO
RenderX

These findings highlight both the challenges in developing accurate GCS prediction models and the potential advantages of NLP-based approaches in capturing the nuanced clinical picture that may not be fully represented in structured data alone.

Feature Importance

After vectorization of the preprocessed notes and removal of highly correlated features, as well as any features referring to sedative or opioid medications (eg, propofol, fentanyl, morphine, hydromorphone, oxycodone, and lorazepam [Ativan]), sedation, or anesthesia, the final feature set comprised 2705 text-derived features. Inclusion of age, sex, and admission type resulted in a total of 2708 features. The pooled ordinal model trained with the elastic penalty selected 541 features as relevant (with nonnull coefficients). The top 20 most relevant features for each model threshold are shown in Figure 5. The pooled linear model exhibited a similar pattern of feature importance (Figure S8 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 5. Feature importance from the pooled ordinal model for the top 20 features. (A) Threshold 1: mild vs moderate+severe and (B) threshold 2: mild+moderate vs severe. Positive coefficients indicate a higher likelihood of being above the threshold (more severe), while negative coefficients indicate a higher likelihood of being below the threshold (less severe). Stemmed features derived from clinical documentation include airway and respiratory terms (ett: endotracheal tube; trach: tracheostomy; vent: mechanical ventilation; peep: positive end-expiratory pressure; to extub: to extubate; ogt: orogastric tube; suction; secret, secretions; carina; vc: vital capacity; abg: arterial blood gas), neurologic examination terms (unrespons: unresponsive; stimuli; incontin: incontinence; withdraw: withdrawal; to pain: no response to pain; pupil; not follow: not follow command; follow command: follows commands; simpl command or command: obeys simple commands; word: verbal response; feel: sensory response; nod; open eye; mental statu: mental status; agit: agitation; awak: awake; alert; ox: derived from ”A&Ox” [alert and oriented x]), and documentation and context terms (unabl to assess: unable to assess; neo: phenylephrine or vasopressor; chair: mobilized to chair).

XSL•FO
RenderX

Feature importance patterns across ordinal thresholds revealed a clinically coherent gradient of severity (Figure 5). Mild cases were characterized by preserved alertness, command-following ability, verbal output, and spontaneous oxygenation. Moderate cases were distinguished primarily by airway instrumentation and mechanical ventilatory support, along with reduced but present responsiveness (patient nodding). In contrast, severe cases were dominated by features reflecting neurologic unresponsiveness, including lack of response to stimuli, pupillary abnormalities, and inability to follow commands. These findings suggest that respiratory support differentiates mild from higher severity states, whereas profound neurologic dysfunction defines the transition to severe illness.

Discussion

Principal Findings

We developed an NLP-based algorithm to classify GCS scores from daily inpatient notes of an adult population. The algorithm was developed using data from 2 health care institutions: one with inpatient admissions to any hospital inpatient setting (MGB) and the other with inpatient admissions to critical care units (MIMIC). The pooled ordinal model achieved high discrimination for predicting GCS across all classes (AUROC 0.96, 95% CI 0.96‑0.96 and AUPRC 0.77, 95% CI 0.76‑0.77). The model reliably provides predictions even when GCS scores are not directly documented, enabling automated ordinal scoring in real-world EHRs with incomplete documentation.

To ensure fair comparison of predicted probabilities, all models were evaluated after identical post hoc calibration. Calibration had minimal impact on both linear and ordinal predictions, indicating that the models' outputs were inherently well calibrated. The pooled linear model achieved an overall good performance (RMSE 2.30; Pearson correlation 0.76). The higher Pearson correlation observed when testing the pooled linear model on MIMIC should be interpreted in the context of its substantially smaller test set. Correlation-based metrics are sensitive to sample size and outcome variability, and smaller, less heterogeneous cohorts may yield higher correlation estimates. In contrast, the larger and more diverse test set from MGB provides a more conservative estimate of performance. Importantly, these differences do not indicate inferior model performance, as pooled results remained stable across institutions.

The model trained on a single institution and tested on an external institution showed overall lower performance, highlighting limited generalizability when models are developed using data from only 1 site. Thus, single-institution models may not generalize well to external data, underscoring the benefit of pooling data across multiple institutions.

Comparison With Prior Work and Potential Applications of Our Model

To our knowledge, this is the first work using an NLP-based algorithm to predict total GCS scores from unstructured clinical notes. Previous approaches have largely relied on structured data elements or manual chart review, which limits scalability and introduces potential selection bias. Our approach offers several advantages, including the ability to process large volumes of clinical documentation efficiently and extract neurological status information that may not be captured in structured fields.

Research in critical care, neurocritical care, and acute and nonacute neurology heavily relies on the GCS score as a key variable to measure illness severity, frequently used as an important confounder for adjustment, and also to determine eligibility for inclusion into epidemiology research and comparative effectiveness research, as well as for recruitment into clinical trials [19-27]. Missing GCS data are frequently a barrier in research that relies on EHR data [28,29]. Our algorithm addresses a major gap by developing a robust and efficient method to measure GCS when it is missing to facilitate EHR-based research. Another potential use of the algorithm is in critical care and trauma outcome studies, as well as in intensive care units survivorship studies, where data on baseline GCS scores are critical to understand relation to outcomes and survivorship [28]. Finally, missing data are a common barrier to EHR-based quality improvement efforts, and our model has potential applications in quality improvement interventions that rely on documentation of GCS or level of consciousness [30-32].

While our model showed excellent performance in identifying cases of severe consciousness impairment, the discrimination between moderate (GCS: 9‑12) and mild (GCS: 13‑15) categories was less robust. This likely reflects the clinical reality that the distinction between these categories can be subtle, with significant overlap in the language used to describe patients at these levels of consciousness. The threshold-specific feature importance patterns from the ordinal model suggest that this reflects the graded clinical structure of neurologic decline. Features distinguishing mild from higher severity states were largely related to airway instrumentation and ventilatory support (eg, endotracheal tube, tracheostomy, mechanical ventilation, and positive end‑expiratory pressure), along with reduced responsiveness (eg, not following commands). In contrast, the transition from moderate to severe was driven predominantly by markers of profound neurologic dysfunction, including unresponsiveness, abnormal responses to stimuli, pupillary findings, and inability to follow commands. Conversely, terms reflecting preserved interaction, such as alertness, orientation, and verbal responsiveness, were negatively associated with both severity thresholds and were most predictive of mild GCS scores. These findings suggest that respiratory support differentiates mild from nonmild cases, whereas severe GCS is primarily defined by documented neurologic unresponsiveness.

Our error analysis revealed insights into the challenges of GCS prediction. Cases of rapid clinical deterioration, where the documented GCS score failed to capture the patient's worst neurological state, represented a particular challenge. Similarly, variability in measurement frequency introduced potential bias, as patients with more frequent monitoring may represent a clinically distinct population.

The model's efficiency is notable—testing on our dataset required only seconds, making it feasible for deployment across large clinical databases. This could enable population-level

XSL•FO
RenderX

neurological outcome research that was previously impractical due to the manual effort required for GCS extraction. Applications could include quality improvement initiatives, comparative effectiveness research, and epidemiological studies of neurological conditions. Our approach is extremely fast and scalable compared to large language models, but it is potentially less generalizable and might take longer to develop. State-of-the-art large language models often perform well set to zero-shot, but they are computationally intensive and take time, making them impractical for processing millions of clinical notes in EHR-based research.

Limitations

Several limitations should be acknowledged. First, both institutions in our study are in the same geographic region, potentially limiting generalizability. We demonstrate that building the model on a single site reduces the performance; therefore, future work should validate the model using data from institutions in different geographies with potentially different documentation practices. Second, our model was developed and validated using retrospective data, and prospective validation would strengthen confidence in its clinical utility. Third, while our model strongly predicted severe GCS scores, further refinement is needed to improve discrimination between moderate and mild categories. With regard to features, our models include vasopressors, which are not direct measures of neurological function and may, in part, act as proxy variables reflecting overall illness severity. However, we chose not to exclude vasopressors because they capture clinically meaningful aspects of a patient's hemodynamic state, which is often tightly coupled with neurological status in critically ill populations, and removing such variables may therefore omit relevant contextual information. Similarly, while code status terms may introduce proxy-learning behavior, we did not exclude these as they do not inherently equate to greater illness severity but may also reflect patient preferences, comorbidities, or goals of care. Additionally, the feature "unable to assess" may reflect examinations confounded by procedures rather than true neurological impairment; however, this term is also used in clinically unstable patients where a reliable neurological assessment cannot be performed and therefore retained in the model. Another limitation is that our model treats each day independently, ignoring temporal dependencies in GCS trajectories. Future work could use sequence modeling (eg, recurrent neural networks) to better capture patient trajectories. While our model can be applied to notes with fewer than 300 tokens, it cannot be applied in the absence of documentation. Future longitudinal analyses using imputation strategies should include strong clinical caveats, as acute changes in neurological status may not be reliably captured. In addition, aggregating all daily notes into a single representation can dilute short but clinically important events, such as brief deterioration or seizures. This may lead the model to underdetect transient adverse events, representing a structural limitation of the aggregation approach rather than model failure. Future NLP tools should align note time stamps tightly with flowsheet time stamps to resolve this temporal mismatch. Finally, although traditional NLP methods were used in this study, these approaches do not capture contextual or conditional semantics to the same extent as transformer-based models (eg, ClinicalBERT), which may limit sensitivity to complex negations or subtle distinctions within moderate impairment. Future work may evaluate whether contextual embeddings improve performance in these intermediate states.

Conclusions

Despite the study limitations, our findings demonstrate the feasibility and potential utility of automated GCS prediction from unstructured clinical notes. By enabling efficient, large-scale phenotyping of neurological assessments, this approach could significantly advance population health research and quality improvement efforts in neurological and critical care. Future work should focus on external validation, prospective evaluation, and integration with other clinical decision support tools to maximize impact on patient care.

Funding

This work was funded by National Institutes of Health (NIH; R01NS131347 SFZ). MBW was supported by grants from the NIH (RF1AG064312, RF1NS120947, R01AG073410, R01HL161253, R01NS126282, R01AG073598, R01NS131347, and R01NS130119) and the National Science Foundation (2014431). SFZ was supported by the NIH (K23NS114201, R01NS126282, and R01AG082693). SSM is supported by Glaucoma and Retinopathy Screening Study (GRaSS) from the NIH (R01MH131194 and R01MH134823) and the Claflin Distinguished Scholar award (Massachusetts General Hospital). LMVRM was supported by the NIH (5R01AG073410-02, 2R01AG082693-01, 5U01AG076478-03, 2P01AG032952-13, and 5R01AG062282-05), the Baker Family Foundation: artificial intelligence–driven patient messaging system, the Centers for Disease Control and Prevention Research Cooperative Agreement (5U48DP006377-04-00), the Epilepsy Foundation (consultant to the CEO), and Food and Drug Administration (consultant to the Digital Health and other committees).

Data Availability

The code for notes preprocessing and modeling, along with a sample deidentified dataset are publicly available [33].

Authors' Contributions

MF, SFZ, and MBW contributed to conceptualization and methodology. MF was responsible for data curation, formal analysis, software, validation, and drafting the original manuscript. MF, SFZ, and MBW contributed to writing–review and editing. SFZ contributed to supervision, resources, and funding acquisition. All authors reviewed and approved the final manuscript.

XSL•FO
RenderX

Conflicts of Interest

SFZ is a clinical neurophysiologist for Corticare and has received speaking honoraria from Marinus, as well as royalties from Springer Publishing, all unrelated to this work. MBW is a cofounder, scientific advisor, and consultant for Beacon Biosignals and holds a personal equity interest in the company. He also receives royalties for authoring *Pocket Neurology* (Wolters Kluwer) and *Atlas of Intensive Care Quantitative EEG* (Demos Medical). None of these relationships influenced the present work. LMVRM reports no significant financial relationships with commercial or proprietary entities producing health care–related products or services relevant to this manuscript. The authors declare no other conflicts of interest.

Multimedia Appendix 1
Supplemental methods, tables, and figures.
[DOCX File, 3569 KB - jmir_v28i1e81245_app1.docx]

References

1. Jain S, Margetis K, Iverson LM. Glasgow Coma Scale. In: StatPearls [Internet]: StatPearls Publishing; 2025. URL: https://www.ncbi.nlm.nih.gov/books/NBK513298/ [accessed 2026-06-16]
2. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med 1985 Oct;13(10):818-829. [Medline: 3928249]
3. Vincent JL, Moreno R, Takala J, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the working group on sepsis-related problems of the European Society of Intensive Care Medicine. Intensive Care Med 1996 Jul;22(7):707-710. [doi: 10.1007/BF01709751] [Medline: 8844239]
4. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA 1993;270(24):2957-2963. [doi: 10.1001/jama.270.24.2957] [Medline: 8254858]
5. MRC CRASH Trial Collaborators, Perel P, Arango M, et al. Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients. BMJ 2008 Feb 23;336(7641):425-429. [doi: 10.1136/bmj.39461.643438.25] [Medline: 18270239]
6. Steyerberg EW, Mushkudiani N, Perel P, et al. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. PLoS Med 2008 Aug 5;5(8):e165. [doi: 10.1371/journal.pmed.0050165] [Medline: 18684008]
7. Gennarelli TA, Champion HR, Copes WS, Sacco WJ. Comparison of mortality, morbidity, and severity of 59,713 head injured patients with 114,447 patients with extracranial injuries. J Trauma 1994 Dec;37(6):962-968. [doi: 10.1097/00005373-199412000-00016] [Medline: 7996612]
8. Moore L, Lavoie A, Camden S, et al. Statistical validation of the Glasgow Coma Score. J Trauma 2006 Jun;60(6):1238-1243. [doi: 10.1097/01.ta.0000195593.60245.80] [Medline: 16766966]
9. Field triage guideline revision: Glasgow Coma Scale: systematic review. : Effective Health Care Program, Agency for Healthcare Research and Quality; 2016 URL: https://effectivehealthcare.ahrq.gov/sites/default/files/pdf/field-triage-glascow_research-protocol.pdf [accessed 2026-06-16]
10. Zhang JK, Dinh TU, Teasdale G, Mercier P, Mattei TA. The message of the Glasgow Coma Scale: a comprehensive bibliometric analysis and systematic review of clinical practice guidelines spanning the past 50 years. World Neurosurg 2024 May;185:393-402.e27. [doi: 10.1016/j.wneu.2024.02.139] [Medline: 38437980]
11. Moore L, Lavoie A, LeSage N, et al. Multiple imputation of the Glasgow Coma Score. J Trauma 2005 Sep;59(3):698-704. [Medline: 16361915]
12. Rue T, Thompson HJ, Rivara FP, Mackenzie EJ, Jurkovich GJ. Managing the common problem of missing data in trauma studies. J Nurs Scholarsh 2008;40(4):373-378. [doi: 10.1111/j.1547-5069.2008.00252.x] [Medline: 19094153]
13. Engel JM, Junger A, Bottger S, et al. Outcome prediction in a surgical ICU using automatically calculated SAPS II scores. Anaesth Intensive Care 2003 Oct;31(5):548-554. [doi: 10.1177/0310057X0303100509] [Medline: 14601278]
14. Ercole A, Dixit A, Nelson DW, et al. Imputation strategies for missing baseline neurological assessment covariates after traumatic brain injury: a CENTER-TBI study. PLoS One 2021;16(8):e0253425. [doi: 10.1371/journal.pone.0253425] [Medline: 34358231]
15. Boag W, Doss D, Naumann T, Szolovits P. What's in a note? Unpacking predictive value in clinical note representations. AMIA Jt Summits Transl Sci Proc 2018;2017:26-34. [Medline: 29888035]
16. Collins GS, Moons KG, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024 Apr 16;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
17. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035. [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
18. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol 2005 Apr;67(2):301-320. [doi: 10.1111/j.1467-9868.2005.00503.x]

XSL•FO
RenderX

19. Reith FC, Lingsma HF, Gabbe BJ, Lecky FE, Roberts I, Maas AI. Differential effects of the Glasgow Coma Scale Score and its components: an analysis of 54,069 patients with traumatic brain injury. Injury 2017 Sep;48(9):1932-1943. [doi: 10.1016/j.injury.2017.05.038] [Medline: 28602178]
20. Yue JK, Lee YM, Sun X, et al. Performance of the IMPACT and CRASH prognostic models for traumatic brain injury in a contemporary multicenter cohort: a TRACK-TBI study. J Neurosurg 2024 Mar 15;141(2):417-429. [doi: 10.3171/2023.11.JNS231425] [Medline: 38489823]
21. Kouloulas EJ, Papadeas AG, Michail X, Sakas DE, Boviatsis EJ. Prognostic value of time-related Glasgow Coma Scale components in severe traumatic brain injury: a prospective evaluation with respect to 1-year survival and functional outcome. Int J Rehabil Res 2013 Sep;36(3):260-267. [doi: 10.1097/MRR.0b013e32835fd99a] [Medline: 23470551]
22. Pisano F, Bilotta F. The predictive value of the verbal Glasgow Coma Scale in traumatic brain injury: a systematic review. J Head Trauma Rehabil 2024;39(4):273-283. [doi: 10.1097/HTR.0000000000000938] [Medline: 38453630]
23. Brennan PM, Murray GD, Teasdale GM. Simplifying the use of prognostic information in traumatic brain injury. Part 1: the GCS-Pupils score: an extended index of clinical severity. J Neurosurg 2018 Jun;128(6):1612-1620. [doi: 10.3171/2017.12.JNS172780] [Medline: 29631516]
24. Wells AJ, Reilly PL. 50 years of the Glasgow Coma Scale: a historical perspective. J Clin Neurosci 2025 Mar;133:110994. [doi: 10.1016/j.jocn.2024.110994] [Medline: 39756098]
25. Chen L, Lu L, Fang Y, et al. Glasgow Coma Scale on admission as predictor of neurological sequelae at discharge and acute respiratory failure in patients with heatstroke. Postgrad Med J 2023 Nov 20;99(1178):1237-1245. [doi: 10.1093/postmj/qgad071] [Medline: 37650372]
26. Gomez D, Byrne JP, Alali AS, et al. Inclusion of highest Glasgow Coma Scale motor component score in mortality risk adjustment for benchmarking of trauma center performance. J Am Coll Surg 2017 Dec;225(6):755-762. [doi: 10.1016/j.jamcollsurg.2017.08.020] [Medline: 28912029]
27. Shen D, Zhou S, Wu H, Zhang W, Li Z, Sun J. Associations between Glasgow Coma Scale trajectories and 28-day survival rate in patients with sepsis-associated encephalopathy: insights from longitudinal group trajectory modeling. Front Neurol 2025;16:1607946. [doi: 10.3389/fneur.2025.1607946] [Medline: 40963937]
28. Butler A, Wei W, Yuan C, Kang T, Si Y, Weng C. The data gap in the EHR for clinical research eligibility screening. AMIA Jt Summits Transl Sci Proc 2018;2017:320-329. [Medline: 29888090]
29. de Andrade JB, de Medeiros Cavalcante MA, Lopes TL, et al. Discovery of data quality issues in electronic health records: profound consequences for critical care medicine applications - a systematized review. Crit Care 2026 Jan 8;30(1):19. [doi: 10.1186/s13054-025-05677-0] [Medline: 41508097]
30. Syed R, Eden R, Makasi T, et al. Digital health data quality issues: systematic review. J Med Internet Res 2023 Mar 31;25:e42615. [doi: 10.2196/42615] [Medline: 37000497]
31. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. EGEMS (Wash DC) 2013;1(3):1035. [doi: 10.13063/2327-9214.1035] [Medline: 25848578]
32. Lighterness A, Adcock M, Scanlon LA, Price G. Data quality-driven improvement in health care: systematic literature review. J Med Internet Res 2024 Aug 22;26:e57615. [doi: 10.2196/57615] [Medline: 39173155]
33. Automated prediction of Glasgow Coma Scale scores from unstructured electronic health records: a natural language processing approach 1.0.0. Brain Data Science Platform. URL: https://bdsp.io/content/gd2l2o4nrl4dl9sok4bd/1.0.0/%20(https://doi.org/10.60508/3fgr-8g80 [accessed 2026-06-25]

Abbreviations

AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
CONSORT-AI : Consolidated Standards of Reporting Trials–Artificial Intelligence
EHR: electronic health record
GCS : Glasgow Coma Scale
MGB: Mass General Brigham
MIMIC: Medical Information Mart for Intensive Care
NLP : natural language processing
RMSE : root mean square error
TRIPOD-AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Artificial Intelligence

XSL•FO
RenderX

Edited by A Coristine; submitted 24.Jul.2025; peer-reviewed by AW Maiga, D Anestis, P Han; revised version received 17.Apr.2026; accepted 20.Apr.2026; published 29.Jun.2026.

Please cite as:
Fernandes M, Turley N, Sun H, Mukerji SS, Moura LMVR, Westover MB, Zafar SF
Automated Prediction of Glasgow Coma Scale Scores From Unstructured Electronic Health Records Using Natural Language Processing: Development and Validation Study
J Med Internet Res 2026;28:e81245
URL: https://www.jmir.org/2026/1/e81245
doi:10.2196/81245

© Marta Fernandes, Niels Turley, Haoqi Sun, Shibani S Mukerji, Lidia M V R Moura, M Brandon Westover, Sahar F Zafar. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Electrocardiogram-Based Mental Stress Detection Amid Everyday Activities Using Machine Learning: Model Development and Validation Study

Buelent Uendes[1], MSc; Alex Antonides[1], MSc; Sjors van de Ven[2], MSc; Denise Johanna van der Mee[2], PhD; Eco de Geus[2], PhD; Mark Hoogendoorn[1], PhD

[1]Department of Computer Science, Vrije Universiteit Amsterdam, De Boelelaan 1111, Amsterdam, The Netherlands
[2]Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Corresponding Author:
Buelent Uendes, MSc
Department of Computer Science, Vrije Universiteit Amsterdam, De Boelelaan 1111, Amsterdam, The Netherlands

Abstract

Background: Frequent, sustained stress is linked to poor health and requires monitoring for early intervention. Electrocardiograms (ECG) are promising biomarkers because they can be recorded noninvasively and continuously using wearable devices. However, tracking stress with ECG is challenging because daily activities elicit responses similar to mental stress (MS), and various mental stimuli that individuals encounter complicate the use of machine learning (ML) models trained on a limited set of stressors.

Objective: We (1) evaluated the ability of ML models to distinguish MS episodes from a composite “no-stress” background, including rest and low- to moderate-intensity activities; (2) assessed their generalizability to new stressors and participants; and (3) tested robustness to lower sampling rates and fewer features, to explore their suitability for lightweight wearables.

Methods: We used a comprehensive ECG dataset sampled at 1000 hertz from 127 participants who underwent various mental stressors and engaged in diverse physical activities. A 30-second window was used to extract 55 features from time, frequency, nonlinear, and morphological domains. We trained a logistic regression (LR) model and an extreme gradient boosting (XGBoost) model, splitting the data into 60/20/20 for training, validation, and testing. Shapley additive explanation values were computed to explain model predictions. Additional analyses included leave-one-stressor-out; downsampling to 500, 250, and 125 hertz; a time-window sensitivity analysis; and reducing the number of features to as few as 5.

Results: XGBoost achieved an area under the receiver operating characteristic curve (AUROC) of 0.741 (95% CI 0.701 - 0.783) and an area under the precision-recall curve (AUPRC) of 0.706 (95% CI 0.658 - 0.753), compared with 0.724 (95% CI 0.678 - 0.772) and 0.691 (95% CI 0.639 - 0.742) for LR. The mean performance difference between XGBoost and LR was 0.017 for AUROC (95% CI 0.001 - 0.032) and 0.015 for AUPRC (95% CI −0.001 to 0.037; clustered bootstrap analysis using 2000 participant-level resamples), suggesting that LR performs comparably to the nonlinear XGBoost model. Both models were robust to downsampling and feature reduction (10 features retained >93% of performance). Extending the analysis window to 60 seconds improved model performance across all sampling rates, highlighting a trade-off between rapid detection and overall performance. When evaluating discrimination from physical activity, models achieved acceptable specificity for light physical activity (XGBoost: 0.787; LR: 0.794) but poor specificity for moderate physical activity (XGBoost: 0.418; LR: 0.444). Both models generalized to most unseen stressors, although performance varied across stressors, with limited transfer to the social-evaluative stressor. Feature importance analysis revealed fuzzy entropy and frequency-based features as key predictors.

Conclusions: ML models can detect MS with high sensitivity and remain robust to lower sampling rates and fewer features. Generalization to novel stressors was stressor-dependent. Importantly, our results highlight challenges in distinguishing stress-related cardiac responses from those caused by physical exertion, revealing critical limitations of single-sensor ECG approaches for MS detection.

(J Med Internet Res 2026;28:e80450) doi:10.2196/80450

KEYWORDS

mental stress; machine learning; electrocardiography; ECG; stress detection; generalizability

XSL•FO
RenderX

Introduction

Background

Mental stress (MS), an ever-present aspect of life, occurs when external demands exceed an individual's available physiological and psychological coping resources [1,2]. While short-term exposure to stress may enhance focus and performance [3], frequent or prolonged exposure can adversely affect health, contributing to psychiatric conditions [4,5] and cardiovascular disease [6]. Consequently, accurate and timely detection followed by effective stress management is essential to mitigate these adverse health outcomes.

Traditional stress assessment methods, such as cortisol analysis [7], are burdensome. In contrast, self-report questionnaires, such as the Perceived Stress Scale [8], administered retrospectively or through ecological momentary assessment, offer greater convenience but are limited in temporal granularity [9]. Moreover, they may be susceptible to recall and reporting bias and less reliable among individuals with alexithymia. Effective stress detection, by contrast, should be minimally invasive and support continuous monitoring to capture early warning signs of stress [10].

Given that stress modulates the activity of the autonomic nervous system (ANS) [11], physiological biomarkers such as the electrocardiogram (ECG) and electrodermal activity have been widely proposed for automated stress monitoring. Furthermore, these signals are increasingly accessible through modern wearable devices, making real-time stress tracking more feasible. However, physiological signals often exhibit subtle and complex patterns that can be difficult to analyze and interpret using traditional statistical methods [1].

Objectives

Machine learning (ML) has emerged as a powerful tool for analyzing high-dimensional data, excelling at uncovering patterns and complex relationships [12]. Consequently, ML has increasingly been used in recent studies to detect stress responses from the biomarkers mentioned above [13-16]. However, most existing studies exhibit one of the following limitations: (1) small cohorts; (2) a narrow range of mental stressors investigated; and (3) restrict MS detection to seated baseline (BL) conditions. This restriction leaves it unclear whether ML models can distinguish stress-induced ECG signals from those triggered by everyday movements, which often produce similar physiological responses. Moreover, because individuals encounter a wide range of novel stressors in daily life, models must generalize to stimuli absent from their training data. This aspect has been largely overlooked in prior work. The few studies that investigate generalization [17-19] typically train their ML models on one dataset and test them on another. Although informative, this approach conflates differences in participant demographics, sensor hardware, and stressor types, making it hard to isolate which factor drives any performance drop.

In this study, we address these gaps by developing and rigorously evaluating two ML models, namely logistic regression (LR) and extreme gradient boosting (XGBoost), on a large ECG dataset sampled at 1000 Hz. The dataset, collected from a controlled laboratory experiment at the Vrije Universiteit Amsterdam and published by van der Mee et al [20], included 127 participants who performed both diverse mental-stress tasks and everyday activities (eg, dishwashing and walking). In this study, we define "mental stress" as the physiological response elicited by cognitively or emotionally challenging tasks. Subsequently, we formulated a binary MS versus no-stress classification task, where the no-stress condition includes seated BLs, recovery periods, and low-to-moderate-intensity physical activities. This design not only improves ecological validity but also tests each model's ability to distinguish stress - induced ECG changes from those driven by physical exertion, without relying on additional sensors (eg, accelerometers). To assess generalization, we evaluate our models on unseen participants and on "novel" stressors, using a leave-one-stressor-out protocol, where we withhold each mental stressor during training and assess model performance on this specific stressor. We further investigate the robustness of our models to lower sampling frequency and the reduction of the feature set. Together, these experiments provide key insights for transitioning ECG-based MS detection from laboratory settings to real-world contexts.

Methods

Overview

The structure of the following section is in line with the TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Artificial Intelligence) statement [21,22]. Sections of the guideline not applicable to this research were omitted for brevity.

Data

This study is a retrospective analysis of pseudonymized data collected from a controlled laboratory experiment conducted at the Vrije Universiteit Amsterdam and published by van der Mee et al [20]. The experimental conditions are shown in Table 1. In addition to ECG signals, subjective affect was assessed after each experimental condition using self-reported measures of positive and negative affect, as measured by the Maastricht Questionnaire [23]. We refer to the original study by van der Mee et al [20] for detailed explanations of the experimental procedure.

XSL•FO
RenderX

Table . Experimental conditions and their associated labels used in our study. The laboratory protocol consisted of mental stress (MS) tasks, physical activities at varying intensities, baseline (BL) rest, and recovery periods, each labeled for binary classification as MS or nonstress (BL, low physical activity [LPA], and moderate physical activity [MPA]). A total of 127 healthy adults participated in the experiments conducted in the Netherlands (data collected from 2017 to 2019).

Experimental condition	Duration (min)	Label
Standing	3	LPA
Sitting	3	BL
Tone avoidance	4	MS
Recovery (sitting)	2	BL
Short sing-a-song stress test (anticipatory)	1	MS
Recovery (sitting)	2	BL
Paced auditory serial addition test	4	MS
Recovery (sitting)	2	BL
Raven's Progressive Matrices	4	MS
Walking at a natural pace	2	MPA
Recovery (standing)	2	LPA
Dishwashing	2	MPA
Vacuum cleaning	2	MPA
Recovery (sitting)	2	BL
Tone avoidance (repeat)	4	MS
Recovery (sitting)	2	BL
Paced auditory serial addition test (repeat)	4	MS
Recovery (sitting)	3	BL

Participants

As outlined in the study by van der Mee et al [20], participants were required to be Dutch-speaking, employed or enrolled in school, and aged 18-48 years. Exclusion criteria included a BMI above 30, high cholesterol, diabetes, liver or thyroid disease, and use of medications that affect the ANS, such as antidepressants or anticholinergics. The final sample included 127 participants (71 female and 56 male) with a mean age of 23.3 (SD 5.6) years. We refer the reader to the original study for additional details on recruitment and population. The data used in this analysis were pseudonymized.

Outcome

For this research, the different experimental conditions, as shown in Table 1, were categorized into the following labels:

- BL: This group includes conditions in which participants were primarily resting, such as sitting and recovery periods.
- MS: This group includes conditions aimed to induce cognitive or emotional stress, such as the tone avoidance (TA) reaction time task, the sing-a-song-stress test (SSST), the paced auditory serial addition task (PASAT), and the Raven's Progressive Matrices (RAVEN) test. For the SSST, the labeled stress segments correspond to the anticipatory phase before singing, that is, the period during which participants are informed they will be asked to sing but have not yet begun.
- Low physical activity (LPA): Standing, including recovery periods that involve standing, is classified as LPA.
- Moderate physical activity (MPA): Activities involving moderate physical exertion, such as walking at a moderate pace, washing dishes, and vacuuming.

Data Preparation

ECG signals are susceptible to various forms of noise and artifacts, which can obscure underlying cardiac activity and hinder accurate analysis. Thus, data preprocessing is crucial to ensure the quality and consistency of the input for the subsequent analysis.

In the first stage of data preprocessing, ECG segments unrelated to the laboratory conditions of interest were removed. These segments included the first and last minutes of the experimental setup, which captured the experimental setup and ECG lead removal, as well as any designated "short break" conditions. Compared with the original study, we excluded high-intensity physical activities (eg, treadmill running and stair climbing), as these activities are easily distinguishable from MS by, for instance, considering the maximum heart rate (HR) and could inflate reported model performance.

In the subsequent step, a 0.5 Hz high-pass Butterworth filter (order=5) was used to remove BL wander and slow drifts in the signal BL caused by respiration, movement, or electrode issues, as well as a power line filter (50 Hz) to attenuate noise from electrical power sources. Both filtering steps were implemented using the preprocessing pipeline provided by the NeuroKit2 package (version 0.2.7) [24]. While filtering addresses signal quality issues, some segments may remain unreliable due to

XSL•FO
RenderX

factors such as temporary electrode detachment caused by leads accidentally falling off during the experiment. To identify and address these segments, a signal quality index (SQI) was computed for each QRS complex. Segments with an SQI below 25% were deemed unreliable and discarded. This threshold was chosen based on visual inspection, where segments with SQI values below this level often exhibited flat lines, rendering the identification of key peaks unreliable.

Predictors

Overview

For each participant, the data was segmented using a sliding-window approach, and features were subsequently calculated. We used a 30-second window size with a step size of 10 seconds. This window length represents the minimum duration demonstrated in prior work to yield reliable heart rate variability (HRV) estimates [25], yet remains short enough for near-real-time applications. Furthermore, our window size aligns with previous research [26].

Following R-peak detection and waveform delineation of the ECG signal, various features were extracted from the resulting ECG signal based on prior research:

Time-Domain Features

For each time window, we extracted the normal-to-normal (NN) interval series, converted it to an instantaneous HR series (in beats per minute [BPM]), and then computed the mean, SD, minimum, and maximum of that series. These measures capture the central tendency and dispersion of the HR, which typically rise during physical exertion and in response to acute mental stressors [27].

HRV features are essential markers for ANS activity [25] and have been consequently linked to stress in prior research [28-30]. Therefore, we calculated the average value of the NN intervals (AVNN: average normal-to-normal), the SD of the NN intervals (SDNN), as well as the root-mean-square of successive differences (RMSSD) for each time window. We also included the ratio between the SDNN and the RMSSD, as it serves as a time-domain surrogate for the low-frequency to high-frequency power ratio, a standard marker of ANS activity [31,32].

In addition to features like AVNN and SDNN, we derived several complementary time - domain features to capture both absolute and relative variability in NN intervals. These include NN20 and NN50, the counts of consecutive interval differences exceeding 20 milliseconds and 50 milliseconds, and their normalized counterparts, PNN20 and PNN50, as these features are helpful for stress monitoring [33,34].

To assess variability relative to mean HR, we calculated CVNN (SDNN/AVNN) and CVSD (RMSSD/AVNN). Finally, we quantified dispersion beyond SDs by computing the IQR of NN intervals and the median absolute difference of NN intervals.

Frequency-Domain Features

To assess the spectral characteristics of the ECG signal, Welch's method was used to estimate the power spectral density across three frequency bands: high frequency (HF; 0.15 - 0.40 Hz), very high frequency (VHF; 0.40 - 0.50 Hz), and ultra-high frequency (UHF; 0.50 - 1.00 Hz). Although prior studies commonly analyze only frequencies up to the HF band (0.40 Hz), this may be restrictive during stress, as acute psychological stress can elevate respiratory rate [35-37], potentially shifting the respiratory-driven modulation of R-to-R intervals above the HF band. Supporting this, Hernando et al [38] showed that respiratory rate increases and becomes less stable during stress, and that incorporating respiratory information improves the characterization of the stress-related autonomic response.

Motivated by this evidence, we included the VHF band (0.40 - 0.50 Hz), as implemented in the NeuroKit2 toolbox [24], to capture high-frequency variability potentially associated with stress-related changes beyond the conventional HF range. We additionally investigated the UHF band (0.50 - 1.00 Hz) in an exploratory manner, as this range may reflect high-frequency variability components that differ systematically across conditions. Lower-frequency bands, such as very low frequency (0.0033 - 0.04 Hz), were not considered in this study, as their estimation would require longer time windows to be accurately captured [31].

Following the approach taken by Schmidt et al [13], we computed the absolute spectral power within each predefined frequency band. Then we expressed each band's power as a fraction of the total spectral power to quantify its proportional contribution. We also extracted the minimum, maximum, mean, SD, and entropy for each frequency band in line with the approach taken by Karthikeyan et al [39].

Nonlinear Features

Given the inherent nonlinearity and complexity of the MS response [40], we included several nonlinear indices of HRV derived from NN intervals to capture the subtle fluctuations and irregularities in ANS activity. Following the approach taken in Tanev et al [41], we included the approximate entropy [42,43] and fuzzy entropy [44]. We excluded sample entropy [43] because approximate entropy is better suited to shorter time windows [45]. Furthermore, as a derivative of sample entropy, fuzzy entropy has been shown to outperform sample entropy [46].

Using Poincaré plot analysis, we extracted the SD1 and SD2 features, along with their ratio, to gain insights into the dynamics of short-term and long-term HRV [45]. Furthermore, heart rate asymmetry (HRA), that is, the asymmetrical distribution between accelerating and decelerating NN intervals, has been shown to reflect MS [47]. Consequently, we extracted the area index [48] as a feature of the HRA. In contrast to other HRA measures, such as Guszik's slope or Porta's index, the area index has been reported to exhibit fewer variations in short-term heartbeat intervals [48].

Because the recurrence quantification analysis has been shown to reflect the sympathetic and parasympathetic nervous system [49,50], we extracted three features from the recurrence quantification analysis plot: average length, longest length, and the entropy of the vertical line (W, W_{max}, and W_{En}).

To assess short-term correlations in the ECG signal, we used detrended fluctuation analysis and extracted the α_1 value [51],

XSL•FO
RenderX

which is moderated by sympathetic activity [52,53]. In contrast to the α_1 value, which requires 3 to 16 beats for estimation, the α_2 value needs 16 to 64 heartbeats [54], which may be problematic given our short 30-second analysis window and the variability in HR across experimental conditions. For this reason, we excluded the α_2 from our feature set.

Last, we included the HR fragmentation features proposed by Costa et al [55], that is, percentage of short segments (PSS), inverse average length of segments (IALS), percentage of inflection points, and percentage of NN intervals in alternation segments, as these biomarkers have been proposed to capture the functionality of the HR control system.

Morphological Features

Beyond linear and nonlinear dynamics, morphological features (ie, characteristics that describe the shape, size, or timing of ECG waves) were included to capture further nuances in cardiac response during MS. In particular, we focused on T-wave alternans (TWA), a beat-to-beat variation in the amplitude of the T-wave that reflects subtle alterations in ventricular repolarization and which has been identified as a potential marker of MS [7,56].

While the spectral method and the modified moving average (MMA) method are both widely used for TWA analysis, the latter was used to extract the TWA feature, as the MMA does not require participants to achieve a HR of 105 to 110 BPM [57]. This is crucial, as participants in the MS or seated BL conditions may experience HRs below this threshold (Figure 1).

Figure 1. Heart rate distributions across experimental conditions (127 total participants). Kernel density plots show mean heart rate (BPM) during sitting, recovery (sitting), low physical activity, mental stress, and moderate physical activity. Overlapping distributions across experimental conditions highlight the difficulty of discriminating between mental stress and nonstress conditions using mean heart rate alone. BPM: beats per minute.

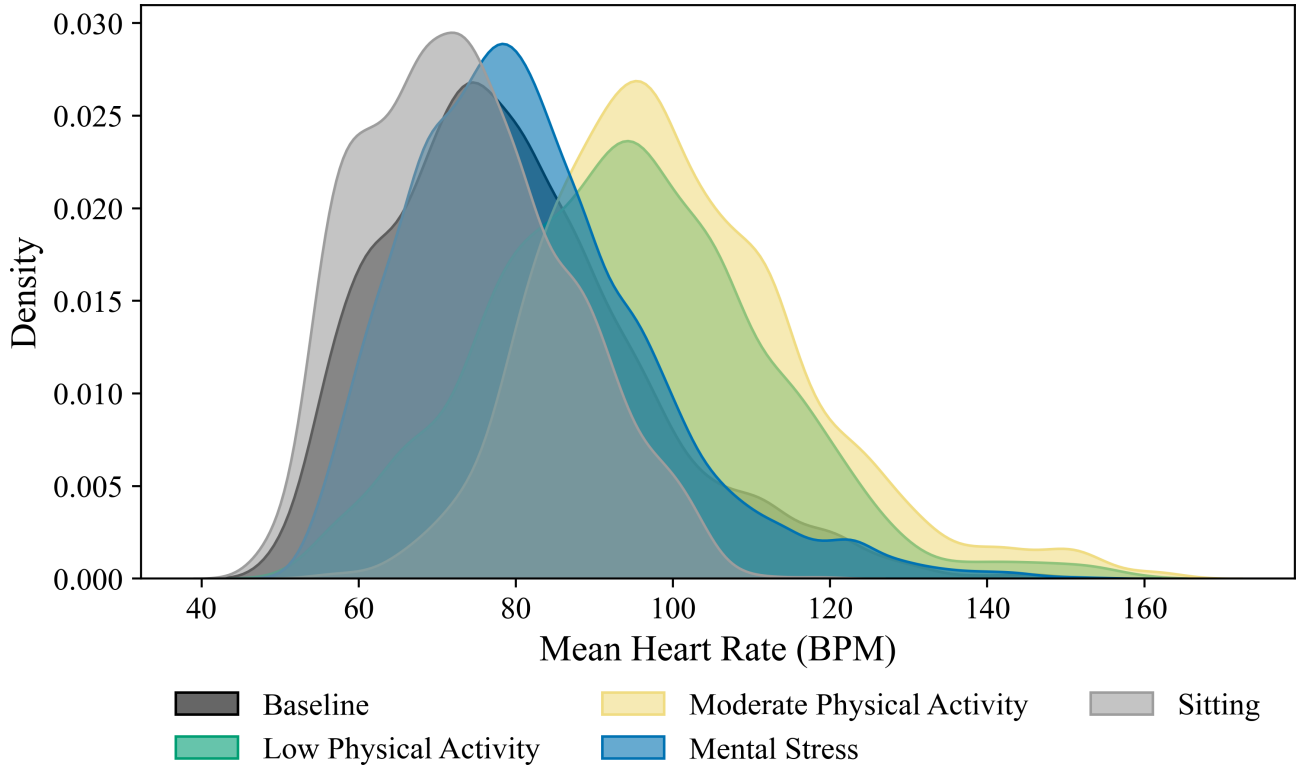

Using the MMA approach [58], we delineated and extracted the T-wave from each 30-second ECG window, numbered them sequentially, and split them into odd (1, 3, …) and even (2, 4, …) beats. We then applied an MMA to each group and defined TWA as the maximum amplitude difference between the odd- and even-beat templates.

The complete set of features used in this study is presented in Multimedia Appendix 1 [55]. We excluded any 30-second window with fewer than 12 detected R peaks or with a mean HR outside the physiologically plausible range of 40 - 220 BPM during feature calculation. This exclusion removed 0.9% of all time windows.

Models were trained using ECG-derived features only to test whether ECG alone is sufficient for stress detection and to enable deployment on any wearable device without requiring personal or demographic data.

Sample Size

As illustrated in Table 2, the 30-second sliding-window approach with a 10-second shift yielded 31,408 observations.

XSL•FO
RenderX

Table . Distribution of electrocardiogram windows across experimental conditions for mental stress (MS) detection (31,408 thirty-second windows from 127 participants; 10-second sliding-window shift). Windows were labeled as MS or nonstress conditions: baseline (BL) rest, low physical activity (LPA), or moderate physical activity (MPA).

Experimental condition	Sample size	Overall percentage (%)
BL	10,140	32.28
LPA	3222	10.26
MPA	3703	11.79
MS	14,343	45.68
Total	31,408	100.00

Missing Data

Due to the controlled, high-quality conditions of the laboratory-based ECG data collection and the preprocessing pipelines used in this study, missing values were scarce. Out of 31,408 analysis windows (Table 2), only 50 (0.16%) windows contained missing data, all of which were attributed to the TWA feature, which requires an adequate number of detected T-peaks for reliable computation. We imputed these missing values using a k-nearest neighbors approach with a k value of 5, where neighbors were selected using Euclidean distance across all 55 standardized (as described below) features, and the missing values were imputed using the mean values of the 5 nearest neighbors. Both the imputer and the scaler were fitted exclusively on the training set to prevent data leakage.

Analytical Methods

ML Models

In this study, we trained an LR with L2 regularization and a nonlinear tree-based model, that is, an XGBoost [59] model. While the LR can only capture linear relationships, its inherent interpretability offers an advantage, especially in the health care domain, where interpretability is a key requirement [60]. By contrast, we selected a tree-based model as the nonlinear approach, given its widespread use in ECG-based stress and affect detection [13,61-63]. Specifically, we selected XGBoost because gradient-boosting methods have demonstrated strong performance across various tabular benchmarks, matching or surpassing neural network architectures [64,65]. By evaluating both ML models side-by-side, we can directly assess the trade-off between more interpretable linear models and more complex ones in the context of ECG-based MS detection. To confirm the stability of the results, we also included results from a random forest (RF) model to assess robustness.

Model Development

We partitioned participants into training (75/125 participants, 60%), validation (26/125 participants, 20%), and test (26/125 participants, 20%) cohorts. For continuous features, we used z-standardization, whereas for discrete features (NN20, NN50, and W_{max}), we applied min-max normalization across the training dataset. We used global rather than participant-specific feature normalization, as global normalization enables assessment of whether ML models can generalize to unseen individuals without requiring per-user calibration data. Hyperparameters for the ML models (LR, XGBoost, and RF) were optimized using Bayesian optimization with Gaussian processes (Optuna version 4.2.1; Preferred Networks, Inc) over 25 iterations; the full hyperparameter tuning ranges and selected values are detailed in Multimedia Appendix 2.

Model Evaluation

We quantified model performance using the area under the receiver operating characteristic (AUROC) curve and the area under the precision-recall curve (AUPRC) curve—the former capturing overall class separation, while the latter emphasizes stress detection performance directly, which is of particular importance in our study. In line with [66], we interpret 0.60 - 0.69 as moderate, 0.70 - 0.79 as acceptable, 0.80 - 0.89 as good, and ≥0.90 as excellent. To evaluate how the ML models would perform in practice, we also report the F_1-score, as well as sensitivity and specificity. For the latter metrics, we used the threshold that maximized the F_1-score on the validation dataset. We computed 95% CIs for all metrics using participant-level cluster bootstrapping (2000 bootstrap samples), in which each bootstrap sample resampled participants within the test dataset with replacement and included all corresponding observations for each selected participant.

Sampling-Rate Robustness

Our ECG recordings were acquired at 1000 Hz, which exceeds the rates of commonly used datasets [13,61,67]. Using the NeuroKit resampling method, we downsampled our ECG signal to 500, 250, and 125 Hz. This procedure quantifies the effects of reduced temporal resolution on model performance, facilitates direct comparison with earlier work, and evaluates the feasibility for consumer-grade devices, such as the Hexoskin Proshirt (Carré Technologies Inc) [68], which operates at lower sampling rates. Furthermore, investigating whether the performance of MS detection is affected by the sampling rate has important implications and considerations for the design of memory-efficient devices.

Time-Window Robustness

To investigate the trade-off between the desire to capture MS in near-real-time (ie, shorter time windows) and the requirement to have sufficiently long windows to generate reliable ECG features, we compared model performance using our standard 30-second segments against extended 60-second segments. Importantly, we used a 20-second time shift for the latter to maintain a constant overall window overlap across both settings. This setup allows us to subsequently investigate how the window size affects model performance.

XSL•FO
RenderX

Model Explainability and Calibration

To explain the predictions of our ML models, we computed Shapley Additive Explanations (SHAP) [69] values. SHAP values are based on a game-theoretic approach to calculate a feature's importance concerning the overall model prediction. For the LR model, we report, in addition to the SHAP values, the model coefficients.

Model calibration (ie, how well the provided model probabilities align with the actual probabilities of the outcomes) is particularly critical in the health care domain [66,70,71]. We assess calibration using calibration curves, the Brier score [72], and the Brier skill score. To obtain the latter, we compare the Brier scores of our ML models to those of a simple majority classifier. These analyses enable us to evaluate not only the predictive performance but also the reliability of the probabilities provided by our ML models. Following [73], we consider a model with a Brier score <0.20 to be well-calibrated. We further use isotonic regression to post hoc calibrate the model probabilities.

Parsimony

As previously outlined, and motivated by prior research, we engineered a total of 55 features spanning time, frequency, nonlinear, and morphological domains to capture physiological responses to MS. However, manual feature generation is a time-consuming process. To quantify the trade-off between model simplicity and predictive performance, we evaluated model performance using progressively smaller feature sets. We applied forward selection to choose subsets of 5, 10, and 20 features by iteratively adding features that maximally improved the overall model performance, while retraining each model at each step of the feature selection process. Compared with alternative feature selection methods such as principal component analysis [74] or mutual information gain, which calculates dependency between the predictor and outcome variable, the forward-selection method directly optimizes for predictive performance during the feature selection process by iteratively adding features that provide the greatest incremental benefit. This process inherently accounts for feature dependencies, as each selection is made conditional on the features already included. Our progressive reduction from 55 to 5 features allows us to quantify the trade-off between model simplicity and performance, while evaluating the extent to which a compact feature subset can maintain classification performance. We note that the forward selection process was performed on a single validation split (see section "Model Development"). The specific feature subset composition may therefore vary across different validation splits. However, the relatively large sample size of our study (127 participants; 31,408 windows) may partially mitigate, though not fully eliminate, the risk of overfitting to a specific data split. Nevertheless, we caution against overinterpreting the feature subsets.

Generalizability Across Stressors

Although our dataset encompasses a variety of different mental stressors (social stressors, cognitive stressors, and active coping), individuals can face a vast majority of varying stressors in their daily lives. To assess whether our models learn a general MS pattern and can generalize to unseen stimuli, we used a leave-one-stressor-out scheme. For each stressor category, we trained models on all but the held-out mental stressor and tested their performance on the unseen stimulus. This analysis can thereby give valuable insights into our model's ability to generalize to novel stress types, a critical requirement for deploying ML for MS detection to ambulatory settings.

We implemented the analysis in Python (version 3.11) and used artificial intelligence tools to accelerate the programming process. Importantly, all output was carefully reviewed, and the code was double-checked and tested for correctness. We release the source code used for this study publicly on GitHub [75].

Class Imbalance

Although the class distribution between the MS and nonstress states (BL, LPA, and MPA) is relatively balanced (45.68% MS vs 54.32% nonstress; Table 2), the leave-one-stressor-out analyses exhibit more pronounced imbalance, as specific mental stressors are left out during training, resulting in a class distribution of approximately 33.33% MS versus 66.67% nonstress examples (when PASAT or TA stimuli are left out). To overcome the class imbalance, we applied the synthetic minority oversampling technique (SMOTE) [76] exclusively to the training set to upsample the minority class, that is, the MS condition. SMOTE is an oversampling method that creates new synthetic instances of the minority class by interpolating between neighboring minority examples. To maintain a consistent training pipeline, we also used SMOTE in the main analysis, although it had a negligible impact on model performance in the more balanced MS versus nonstress setting (AUROC with vs without SMOTE: 0.741 vs 0.742 for XGBoost; 0.724 vs 0.724 for LR).

Model Output

We framed our primary outcome as a binary classification task: detecting MS episodes across all nonstress states, including seated BLs, recovery periods, and LPA to MPA. In this way, our nonstress class reflects both desk-bound environments, such as office work in front of a computer, as well as more mobile contexts, such as those found in the teaching or nursing profession. By incorporating both dynamic and seated non-MS conditions, our models must learn to distinguish ECG changes induced by MS from those resulting from resting or physical activity—an inherently more complex task than handling either condition alone.

Ethical Considerations

Given that this study represents a retrospective analysis of precollected data, ethical approval was not required under the Dutch Medical Research Involving Human Subjects Act (WMO) [77]. However, the study from which the experimental data were obtained [20] was approved by the Medical Ethical Committee of the Vrije Universiteit Amsterdam Medical Center (METc VUmc #2017.374, ABR #NL62442.029.17). All participants provided written informed consent before the start of the experiment and were either compensated with research credits (for student participants) or with a €50 (US $57.92) gift voucher. All data were pseudonymized.

XSL•FO
RenderX

Results

Exploratory ECG Signal Analysis

Before evaluating our models, we first examine 2 fundamental ECG biomarkers, namely mean HR and AVNN reactivity, to illustrate the intrinsic challenge of detecting MS in everyday routine activities. Mean HR was selected for its simplicity of calculation from an ECG signal. In contrast, HR reactivity (or its inverse, AVNN reactivity) was chosen for its link to cardiac sympathetic and parasympathetic tone and its reliability in recognizing MS in experimental conditions [34]; however, because it relies on a controlled BL measurement, it cannot be defined in unconstrained, fully ambulatory settings.

Figure 1 shows the kernel density estimates of mean HR for five experimental conditions: seated BL, recovery (sitting), low- and moderate-intensity activity, MS, and high-intensity activity. As expected, increasing physical activity leads to a corresponding increase in mean HR. Crucially, the distribution for MS (blue) overlaps extensively with both low- and moderate-intensity activity (green and yellow) and even with the resting peaks (BL and recovery in gray). This pronounced overlap underscores the difficulty of disentangling MS-induced changes from those arising during everyday activity or recovery based on a simple measurement such as HR.

Figure 2 shows the AVNN reactivity, calculated as the difference between AVNN during each experimental stressor and the seated BL, across all MS conditions. Although some individuals exhibit positive AVNN reactivity, represented by the mean change from the seated BL, was significantly negative across all tasks, indicating an overall decline in AVNN compared with BL (PASAT: t_{126}=−7.66, P<.001; PASAT [repeat]: t_{126}=−12.02, P<.001; RAVEN: t_{126}=−3.19, P=.002; SSST: t_{126}=−11.92, P<.001; TA: t_{126}=−8.20, P<.001; TA [repeat]: t_{126}=−13.09, P<.001 for 2-sided paired t tests). However, the magnitude of this change varied considerably across experimental conditions. For instance, TA repeat elicits the most considerable mean AVNN reduction, whereas RAVEN matrices produce a more modest decline.

XSL•FO
RenderX

Figure 2. AVNN interval reactivity in ms across mental stressors. AVNN reactivity was calculated for each participant (127 participants) as the average NN interval during the mental stressor minus the average NN interval during the sitting baseline period (ie, the inverse of the commonly used HR reactivity). Statistical significance was assessed using a paired *t* test. Test statistics: PASAT: t_{126}=–7.66, *P*<.001; PASAT (repeat): t_{126}=–12.02, *P*<.001; RAVEN: t_{126}=–3.19, *P*=.002; SSST: t_{126}=–11.92, *P*<.001; TA: t_{126}=–8.20, *P*<.001; TA (repeat): t_{126}=–13.09, *P*<.001 for 2-sided paired *t* tests. AVNN: average normal-to-normal; HR: heart rate; ms: milliseconds; NN: normal-to-normal; PASAT: paced auditory serial addition task; RAVEN: Raven's Progressive Matrices; SSST: sing-a-song-stress test; TA: tone avoidance.

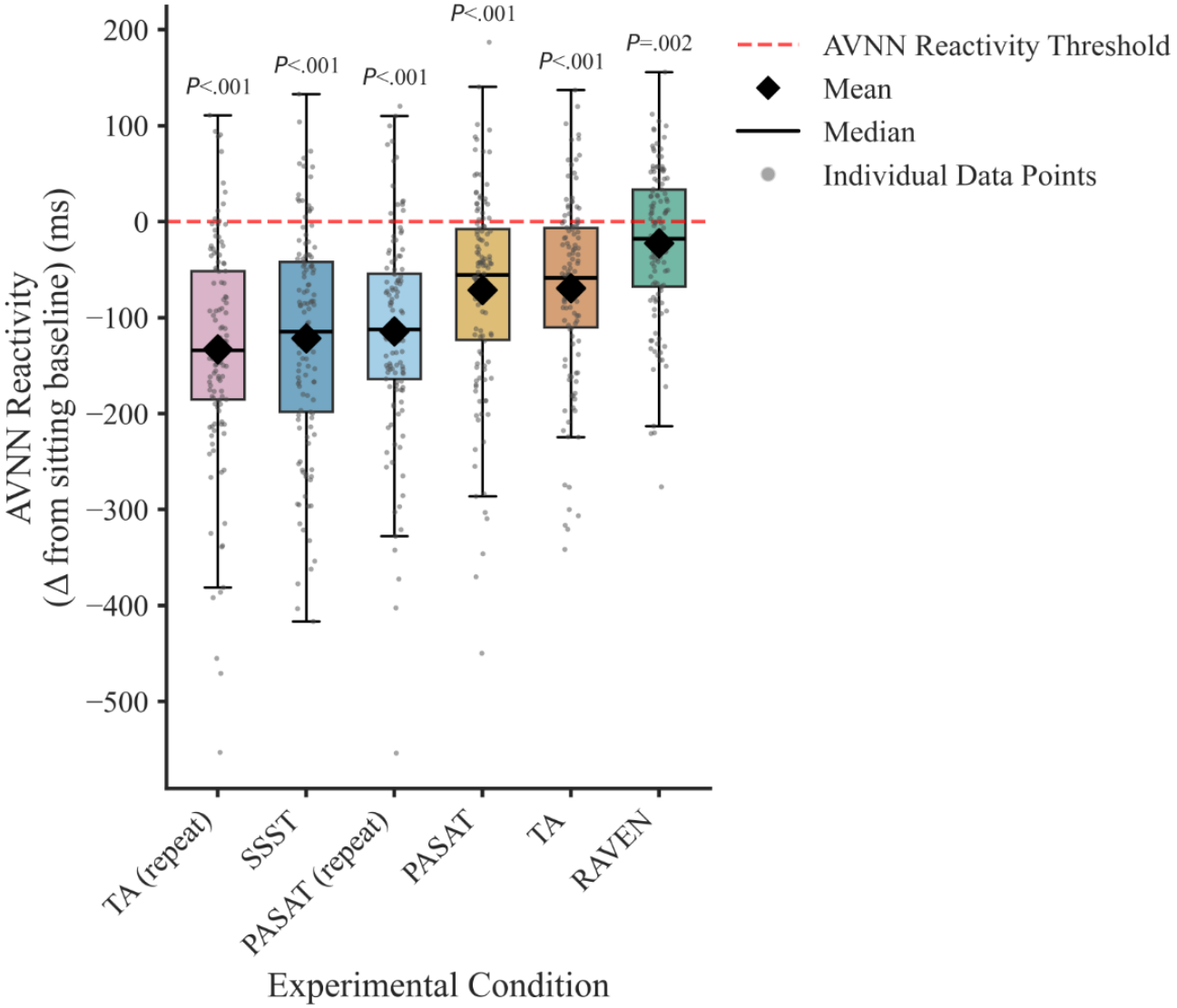

Taken together, these figures demonstrate that neither mean HR nor HR reactivity (or its inverse AVNN reactivity) alone can reliably discriminate MS from everyday activity or account for heterogeneity across stress tasks and individuals. This motivates our extraction of 55 time-, frequency-, nonlinear, and morphological ECG features, and the use of ML models to develop an MS classifier that detects physiological patterns characteristic of MS.

The large interindividual variability in ECG responses underscores that MS reactions are person-specific, likely influenced by subjective perceptions of stress. In Multimedia Appendix 3, we also show changes in both positive and negative affect relative to the sitting BL. As expected, positive affect significantly decreased across all mental stressors (PASAT: t_{126}=−13.25, *P*<.001; PASAT (repeat): t_{126}=−13.45, *P*<.001; RAVEN: t_{126}=−7.96, *P*<.001; SSST: t_{126}=−4.65, *P*<.001; TA: t_{126}=−7.71, *P*<.001; TA [repeat]: t_{126}=−5.00, *P*<.001 for 2-sided paired *t* tests). Likewise, negative affect significantly increased compared with the sitting BL (PASAT: t_{126}=−12.00, *P*<.001; PASAT [repeat]: t_{126}=−6.57, *P*<.001; RAVEN: t_{126}=−6.09, *P*<.001; SSST: t_{126}=−6.67, *P*<.001; TA: t_{126}=−8.31, *P*<.001; TA [repeat]: t_{126}=−3.21, *P*=.002 for 2-sided paired *t* tests). Nonetheless, apparent individual differences emerged, highlighting the challenge of building models that generalize across both stressors and individuals.

Classification Performance

Model Performance Across Sampling Rates

Figure 3 compares LR and XGBoost on our MS detection task at 1000 Hz and after downsampling to 500, 250, and 125 Hz. Both classifiers achieve acceptable AUROC scores, with XGBoost outperforming the LR counterpart across all frequencies (XGBoost: 0.741, 95% CI 0.701 - 0.783; LR: 0.724, 95% CI 0.678 - 0.772) at 1000 Hz. Interestingly, AUROC

XSL•FO
RenderX

performance remains stable across all sampling rates for both models, suggesting that sampling at higher rates does not yield performance gains. In Multimedia Appendix 4, we report corresponding AUPRC scores, which show similar trends and qualitative results. To estimate the performance difference between XGBoost and LR, we computed the mean pairwise differences in AUROC and AUPRC using 2000 participant-level cluster bootstrap samples, alongside 95% and 99% CIs to quantify uncertainty around these estimates. The results are reported in Multimedia Appendix 5. We chose the bootstrap approach over the DeLong test [78] because the latter assumes independent samples [79], which is violated in our setup due to within-participant correlation among the time windows. As shown in Multimedia Appendix 5, the mean performance difference (Δ) between XGBoost and LR was 0.017 for AUROC (95% CI 0.001 - 0.032) and 0.015 for AUPRC (95% CI −0.007 to 0.037), using 2000 bootstrap samples. These differences are small in magnitude (<0.02), suggesting that both models perform comparably across sampling rates. At the F_1-maximizing threshold (1000 Hz), XGBoost achieved a sensitivity of 0.800 (95% CI 0.698 - 0.889) but a specificity of only 0.512 (95% CI 0.420 - 0.609); LR performed similarly (sensitivity: 0.782, 95% CI 0.671 - 0.878; specificity: 0.509, 95% CI 0.417 - 0.613). This asymmetry indicates that, although both models reliably detect MS, approximately half of the nonstress segments are misclassified, highlighting a critical challenge for deployment. Complete results across all sampling rates, along with F_1-scores and RF results, are provided in Multimedia Appendix 6.

Figure 3. Performance comparison of logistic regression and extreme gradient boosting for electrocardiogram-based mental stress classification across sampling rates (127 total participants, 26 test set participants). Points represent bootstrapped mean AUROC with 95% CIs (error bars) from 2000 participant-level bootstrap samples. Models were trained on 55 features extracted from 30-second windows (10-second shift) using 60/20/20 (train/validation/test) splits at the individual level. Both models demonstrate robustness to downsampling from 1000 to 125 Hz. AUROC: area under the receiver operating characteristic curve.

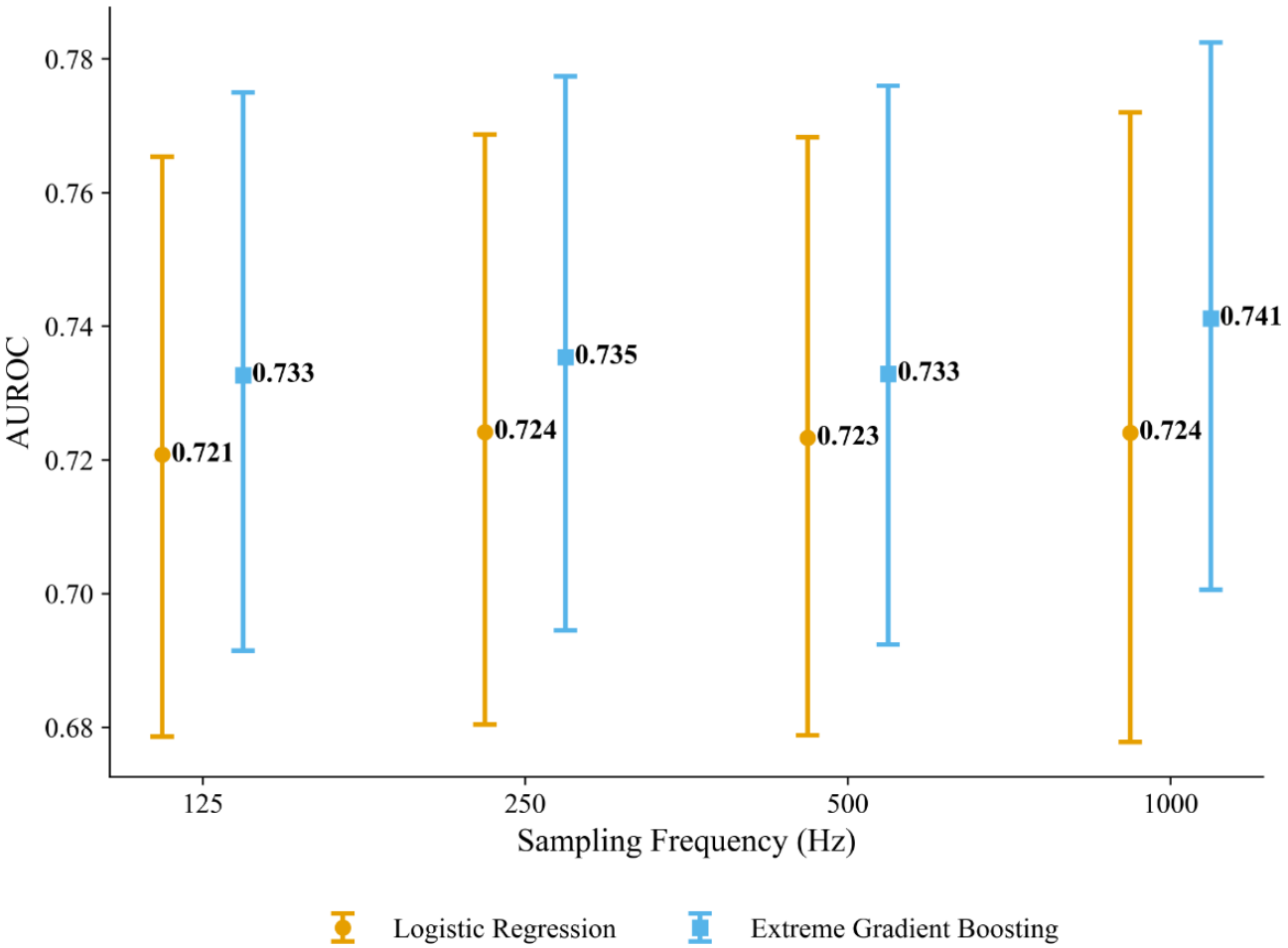

Model Performance Under Simplified Conditions

To assess the impact of our inherently more challenging classification setup, we retrained our ML models on a simplified task: distinguishing MS from a seated BL. The corresponding results are reported in Multimedia Appendix 7. As can be seen, model performances improve substantially under these simplified settings: bootstrapped mean AUROC for XGBoost increases from 0.741 to 0.766, and from 0.724 to 0.752 for the LR at a sampling rate of 1000 Hz (with similar increases for the other sampling rates). The corresponding AUPRC improvements are even more substantial, underlining the challenge our setup poses for MS detection (especially without reliance on accelerometers).

As physiological recovery from stressors is not instantaneous, the recovery periods following each experimental condition (Table 1) likely contain residual autonomic arousal. To

XSL•FO
RenderX

investigate the impact of this on model performance, we conducted a sensitivity analysis excluding recovery periods from the nonstress class. The AUROC performance remained stable (LR: 0.728, 95% CI 0.671 - 0.787; XGBoost: 0.751, 95% CI 0.699 - 0.806 at sampling rate of 1000 Hz), whereas the AUPRC improved substantially for both models (LR: 0.825, 95% CI 0.782 - 0.867; XGBoost: 0.848, 95% CI 0.811 - 0.885), suggesting that including recovery periods introduced label noise negatively affecting AUPRC, likely due to residual autonomic arousal from preceding stressors. However, because distinguishing stress from poststress recovery is an important and realistic challenge in detecting MS in everyday life, we retained recovery periods in the main analysis.

Time-Window Robustness

In Multimedia Appendix 8, we compare AUROC and AUPRC scores for 30-second versus 60-second windows for both models. Extending the time window yields consistent performance gains for both LR and XGBoost across all frequencies, highlighting the trade-off between quick stress detection (short time windows) and improved model performance (longer windows). For instance, at a 1000 Hz sampling rate, doubling the window from 30 to 60 seconds increased XGBoost's bootstrapped mean AUROC from 0.741 to 0.759 (and AUPRC from 0.706 to 0.742), while LR's mean AUROC rose from 0.724 to 0.747 (AUPRC from 0.691 to 0.730).

Instance-Level Agreement

We further assessed Cohen kappa regarding the model agreement on individual test instances. The substantial agreement of 0.735 (95% CI 0.695 - 0.775) suggests that LR and XGBoost largely agree on an instance-level, which ECG responses reflect MS.

Model Explainability and Model Calibration

We next examined feature importance and model calibration, that is, how well the predicted probabilities represent the true probability of the outcomes. Figure 4 shows the 10 most important predictors for mental-stress detection in the XGBoost model according to the SHAP values. Notably, 5 out of these features originate from the frequency domain, highlighting the importance of spectral features in stress detection. Two heart-rate fragmentation indices (PSS and IALS) and fuzzy entropy also rank highly, with higher values of these nonlinear features pushing XGBoost's output towards the MS class.

Interestingly, maximum HR and NN20, 2 simple time-domain features, are also important for predicting MS, demonstrating the relevance of time-domain metrics.

In Multimedia Appendix 9, we present the corresponding SHAP values for the LR model, along with its feature rankings based on the absolute values of the model coefficients. Fuzzy entropy emerges as the most important predictor of MS, underscoring its significance.

IALS again appears to be a crucial feature for MS detection, thereby confirming the importance of heart-rate fragmentation indices and underscoring the results obtained with the XGBoost model. Among time and frequency domain features, total and maximum power in the HF band, SD of HR, and PNN20 are also of high importance for the LR model.

XSL•FO
RenderX

Figure 4. SHAP plot for the detection of mental stress versus no-mental-stress for the extreme gradient boosting model. The top 10 most important features are presented in descending order of importance. Each dot represents a Shapley value for a specific instance and feature, with the color indicating the underlying feature value ranging from high (red) to low (blue). HF: high frequency (0.15-0.40 Hz); HR: heart rate; IALS: inverse average length of segments; NN20: number of (normal-to-normal) NN intervals differing by more than 20 milliseconds; PSS: percentage of short segments; SHAP: Shapley additive explanation; VHF: very high frequency (0.40-0.50 Hz).

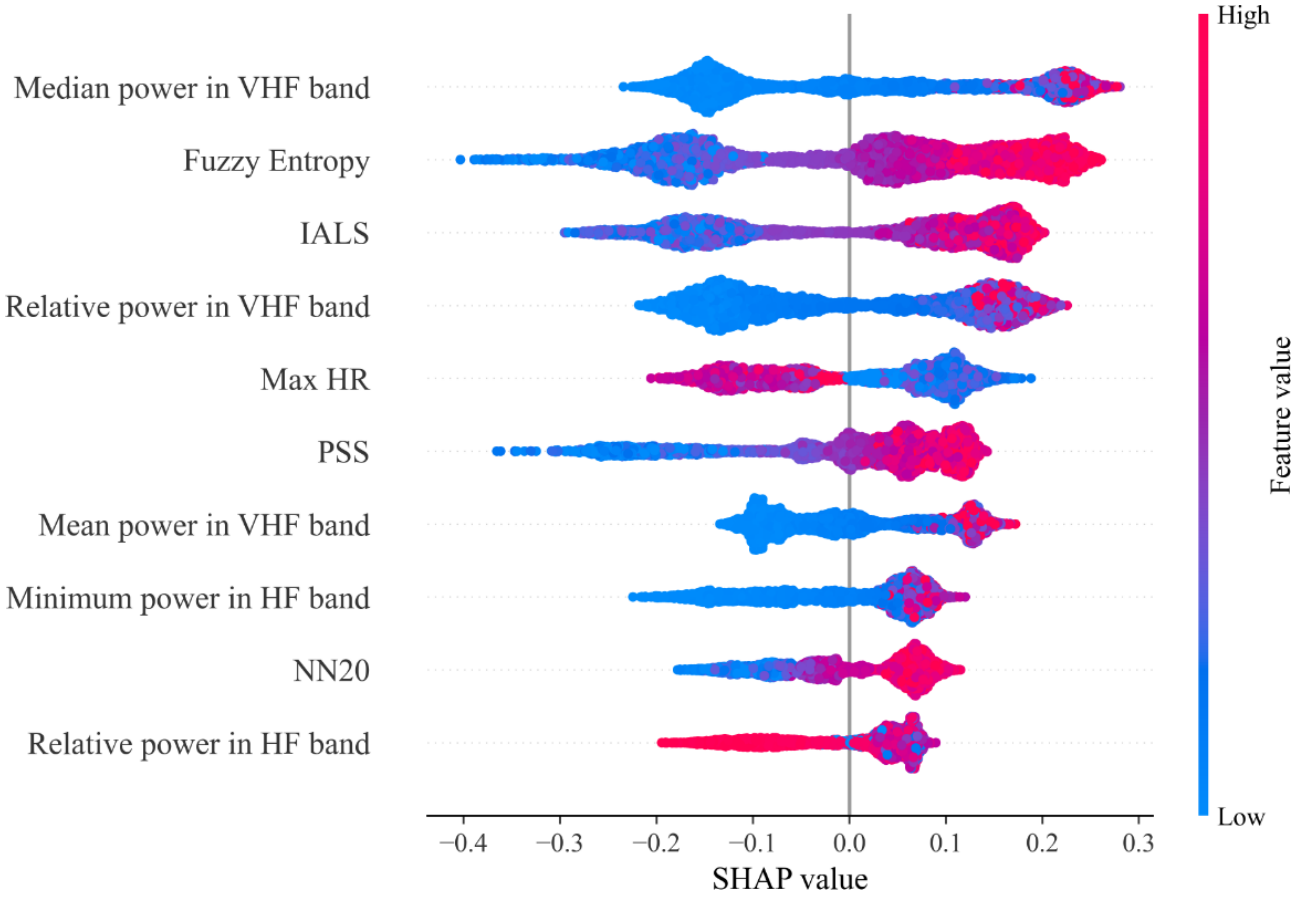

The corresponding model-explainability results for the MS versus the seated nonstress condition (Multimedia Appendix 9; Figures 3-5) also highlight the importance of fuzzy entropy, as well as both the high-frequency and very high-frequency components, as predictive features.

In the health care domain, accurate probability estimates are crucial for informed decision-making [66,70]. For this reason, we assessed model calibration using the Brier score and the expected calibration error with 10 bins.

LR achieves a bootstrapped mean Brier score of 0.215 (95% CI 0.199 - 0.232), while XGBoost scored 0.209 (95% CI 0.194 - 0.224), thereby falling just above our <0.20 cutoff for well-calibrated probabilities. However, the AUROC for the XGBoost model is 0.615 (95% CI 0.586 - 0.643), and 0.603 (95% CI 0.572 - 0.633) for the LR, indicating substantial relative improvements over the dummy classifier BL that predicts the majority class (MS). Figure 5 shows the calibration curves for both classifiers, indicating slight underconfidence at low prediction probabilities (0.05 - 0.25).

XSL•FO
RenderX

Figure 5. Calibration curves for the logistic regression and extreme gradient boosting models, evaluated on the held-out test set (127 total participants and 26 test participants). Points show the predicted model probabilities for mental stress versus the observed probability of mental stress (y-axis), across 10 bins. The dotted line indicates perfect calibration. ECE quantifies deviation from perfect calibration, with lower values indicating better calibration. ECE: expected calibration error.

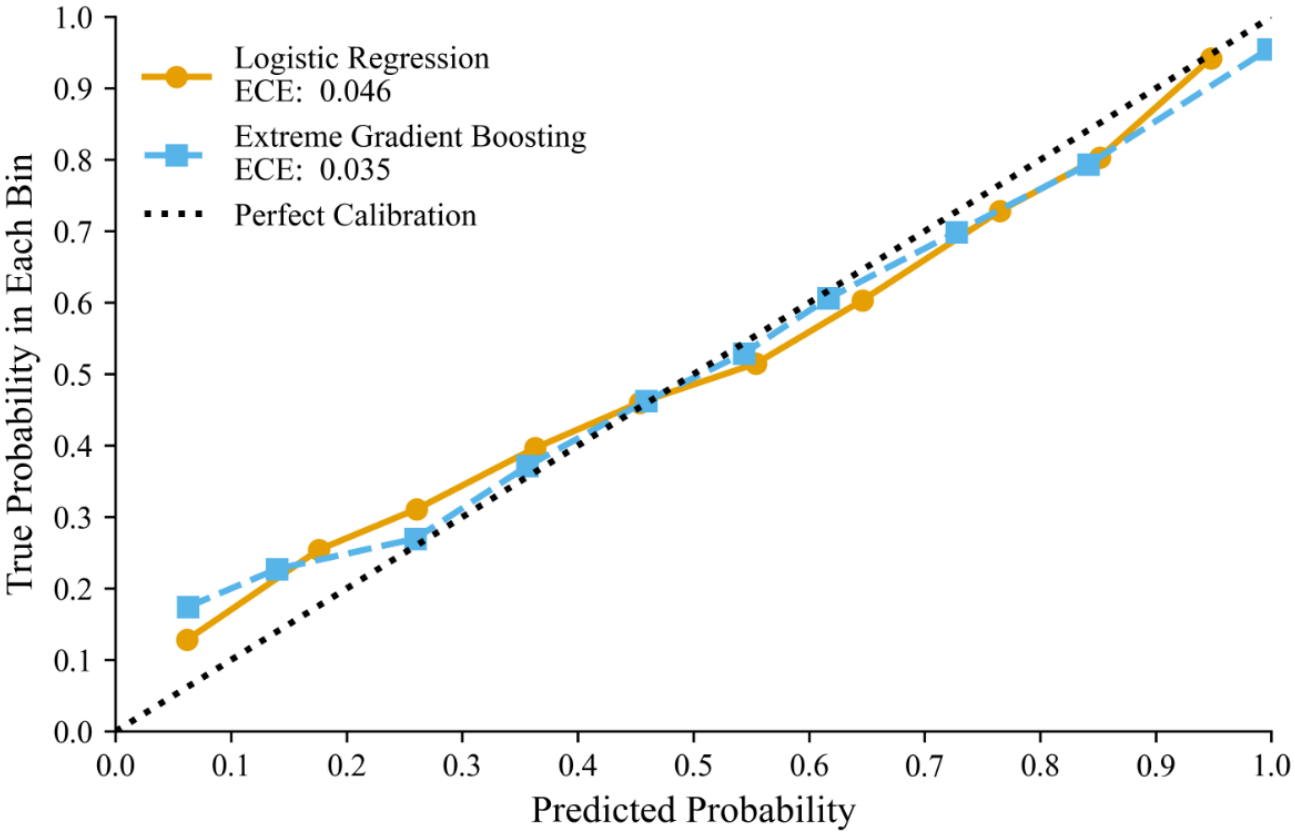

Feature Parsimony

Generating the full set of 55 features is time-consuming and increases memory demands, which are both undesirable for resource-constrained wearable devices. Moreover, a large feature set can obscure model interpretability. To assess the trade-off between model simplicity and predictive performance, Figure 6 reports the AUROC as the number of features is progressively reduced. Notably, performance remains surprisingly robust, with over 95% of the original AUROC retained for both models, even with only 10 features. Interestingly, XGBoost outperforms the LR model across all feature sets. In Multimedia Appendix 10, we report the corresponding results for the AUPRC score, which are in line with the previously stated insights. The full list of selected features for each feature set is provided in Multimedia Appendix 11. As noted in the Methods section, feature selection was conducted using a single validation split. The resulting feature rankings should therefore be interpreted with caution.

XSL•FO
RenderX

Figure 6. Performance comparison of logistic regression and extreme gradient boosting for electrocardiogram-based mental stress classification across varying feature-set sizes on the held-out test set (127 total participants and 26 test participants). Points represent bootstrapped mean AUROC with 95% CIs (error bars) from 2000 participant-level bootstrap samples. Models were trained using a 60/20/20 (train/validation/test) split at the individual level, with features selected via forward selection based on validation performance. Both models demonstrate robustness to feature reduction, maintaining >95% of original performance with only 10 features (logistic regression: 0.707; extreme gradient boosting: 0.729) compared with the full 55-feature set (logistic regression: 0.724; extreme gradient boosting: 0.741). AUROC: area under the receiver operating characteristic curve.

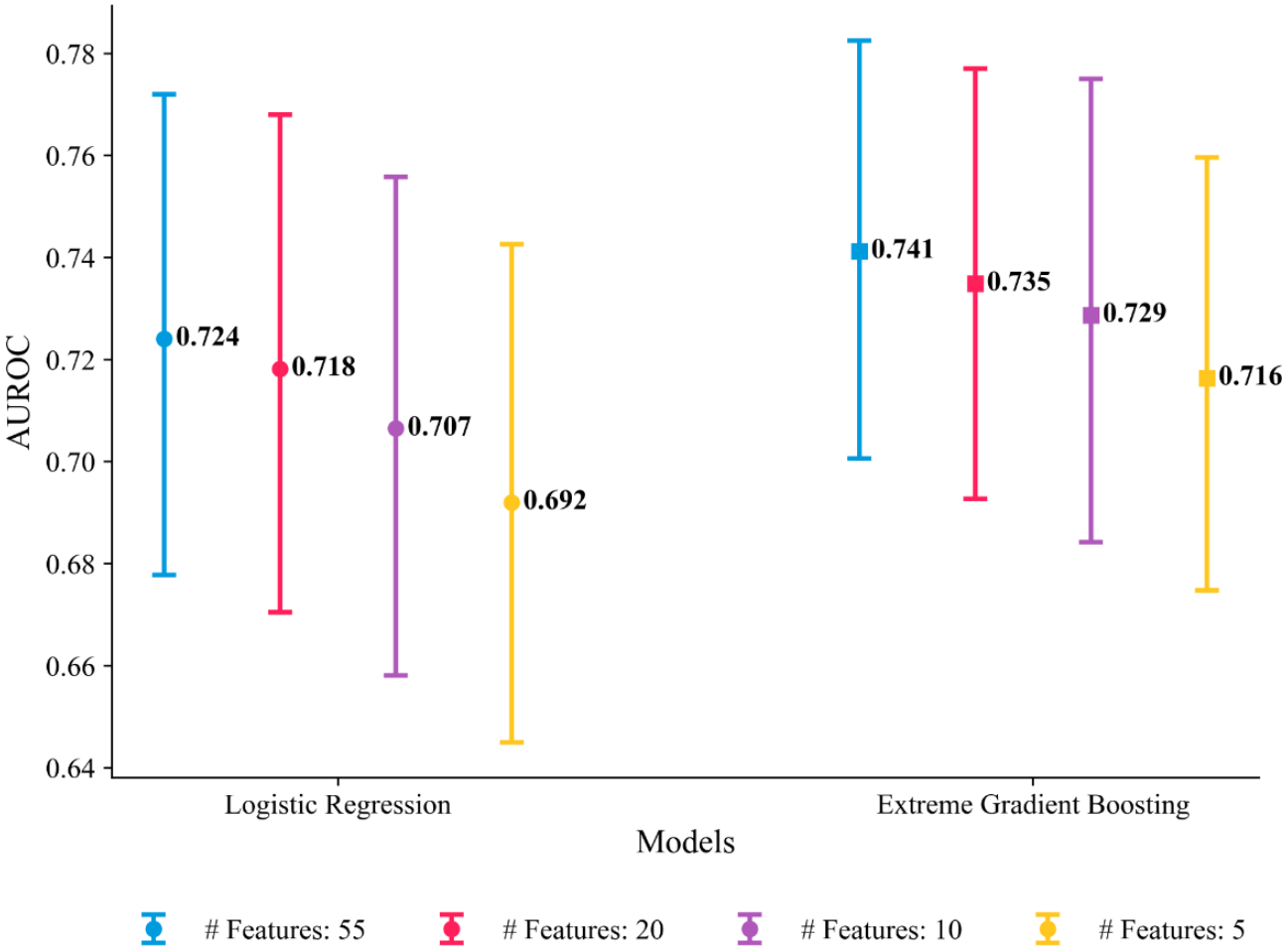

Generalization to Unknown Stressors

Every day, life exposes individuals to a wide variety of mental stressors, making it practically infeasible to train a model on every possible type of stimulus. Thus, a model's ability to perform well under unseen mental stressors is essential for real-world deployment [17]. Table 3 reports the AUROC for each model under a leave-one-stressor-out scheme. For each held-out stressor, the model is trained on all remaining stressors plus the nonstress condition and then tested on its ability to discriminate the held-out stressor from the nonstress condition. Note that if a mental stressor had a repeat condition in the experimental protocol (eg, TA and PASAT), both stressors were left out during training of the ML models. To ensure these results accurately reflect the difficulty of generalizing to an unseen stressor, rather than just the effect of having fewer stressor types and training examples, we also report, for each left-out mental stressor, the BL AUROC on the remaining (known) stressors.

XSL•FO
RenderX

Table . Bootstrapped mean area under the receiver operating characteristic curve (AUROC) with 95% CIs from 2000 participant-level bootstrap samples for binary classification of mental stress versus nonstress conditions under leave-one-stressor-out evaluation on the held-out test set (127 total participants and 26 test participants). Models were trained using a 60/20/20 (train/validation/test) split at the participant level, using a total of 55 features. For each row, the "baseline" column reports the AUROC performance of the model on stressors included in the training set. In contrast, "left-out" reports the AUROC model performance on the held-out (unseen) mental stressor, assessing each model's ability to generalize to novel mental stress conditions. Note that if a mental stressor had a repeat condition in the experimental protocol (eg, tone avoidance and paced auditory serial addition task), both stressors were left out during training of the machine learning models, which is why their baseline scores are identical.

Left-out mental stressor	AUROC model performance, 95% CI					
	LR[a] baseline	LR left-out	XGBoost[b] baseline	XGBoost left-out	RF[c] baseline	RF left-out
SSST[d]	0.7218 (0.6812 - 0.7660)	0.5620 (0.4667 - 0.6548)	0.7360 (0.6975 - 0.7759)	0.5600 (0.4795 - 0.6380)	0.7413 (0.7039 - 0.7792)	0.5499 (0.4680 - 0.6261)
RAVEN[e]	0.7146 (0.6789 - 0.7520)	0.6365 (0.5736 - 0.6976)	0.7365 (0.7070 - 0.7671)	0.6595 (0.6057 - 0.7114)	0.7371 (0.7061 - 0.7696)	0.6645 (0.6092 - 0.7173)
PASAT[f]	0.7219 (0.6777 - 0.7659)	0.7035 (0.6390 - 0.7673)	0.7322 (0.6888 - 0.7752)	0.7125 (0.6490 - 0.7807)	0.7333 (0.6906 - 0.7754)	0.7150 (0.6520 - 0.7816)
PASAT (repeat)	0.7219 (0.6777 - 0.7659)	0.6570 (0.5711 - 0.7456)	0.7322 (0.6888 - 0.7752)	0.6691 (0.5897 - 0.7518)	0.7333 (0.6906 - 0.7754)	0.6782 (0.5985 - 0.7593)
TA[g]	0.7123 (0.6705 - 0.7558)	0.7731 (0.7070 - 0.8352)	0.7107 (0.6710 - 0.7528)	0.7594 (0.6965 - 0.8192)	0.7087 (0.6713 - 0.7492)	0.7608 (0.7017 - 0.8205)
TA (repeat)	0.7123 (0.6705 - 0.7558)	0.7337 (0.6488 - 0.8127)	0.7107 (0.6710 - 0.7528)	0.7200 (0.6454 - 0.7936)	0.7087 (0.6713 - 0.7492)	0.7618 (0.6834 - 0.8370)

[a]LR: logistic regression.

[b]XGBoost: extreme gradient boosting.

[c]RF: random forest.

[d]SSST: sing-a-song-stress test.

[e]RAVEN: Raven's Progressive Matrices.

[f]PASAT: paced auditory serial addition task.

[g]TA: tone avoidance.

As shown in Table 3, model performance on held-out stressors generally declines compared with stressors included in the training set, as expected. Nevertheless, all models achieve AUROC well above the 0.50 chance level on most unseen stressors, confirming their ability to generalize to new mental stimuli. The exception was SSST, for which the 95% CI overlaps chance performance. Notably, the magnitude of performance drop to unseen mental stressors varies across the different stimuli. While both the SSST and RAVEN are most difficult to classify correctly as MS, the opposite holds true for the TA condition. The fact that LR, XGBoost, and the RF model show comparable generalization capability suggests that differences in generalizability to novel stressors arise from the underlying ECG response, rather than from model complexity.

Multimedia Appendix 12 presents AUPRC results that closely mirror the AUROC findings. SSST and RAVEN are the hardest to detect when excluded from model training. Although model performance is generally worse for the left-out stressor, the TA stressor (and its repeat condition) shows the opposite pattern: the AUPRC score when TA is held out exceeds that of other mental stressors included in training. This suggests that TA produces particularly distinctive and readily identifiable stress patterns.

Stratified Stressor Analysis

While our leave-one-out evaluation results suggest that our ML models can overall generalize to novel mental stressors, the generalizability to unseen mental stressors varied significantly. In particular, the sharp contrast between TA and SSST, 2 mental stressors that elicit strong physiological responses as measured by AVNN reactivity (Figure 2), is intriguing. We therefore examined how each model performed on individual stressor types. Table 4, therefore, reports AUROC scores stratified by mental-stressor condition. Again, model performance varies by mental stressor: TA achieves AUROC>0.75, while others (eg, RAVEN and SSST) yield only moderate performance, which suggests that certain MS stimuli produce more distinguishable ECG responses (as compared with the nonstress class) than others. Importantly, this observation holds for LR, XGBoost, and RF alike, demonstrating that these differences are not related to model complexity. The results for the AUPRC, as presented in Multimedia Appendix 13, align with these findings.

XSL•FO
RenderX

Table . Bootstrapped mean area under the receiver operating characteristic curve (AUROC) with 95% CIs from 2000 participant-level bootstrap samples for mental stress classification, stratified by individual stressor type and evaluated on the held-out test set (26 participants from 127 total participants). Results are shown for logistic regression (LR), extreme gradient boosting (XGBoost), and random forest (RF) trained using 60/20/20 (train/validation/test) splits at the participant level with 55 features. Each row presents model performance for a specific mental stressor, assessing how well the models detect that stressor. An AUROC score of 0.50 represents chance-level performance (ie, a random baseline that predicts the majority class).

Mental stressor	AUROC model performance, 95% CI		
	LR	XGBoost	RF
SSST[a]	0.5902 (0.4955 - 0.6792)	0.5726 (0.4958 - 0.6464)	0.5690 (0.4926 - 0.6421)
RAVEN[b]	0.6657 (0.6090 - 0.7230)	0.6829 (0.6350 - 0.7312)	0.6893 (0.6429 - 0.7348)
PASAT[c]	0.7087 (0.6435 - 0.7745)	0.7231 (0.6607 - 0.7882)	0.7219 (0.6588 - 0.7871)
PASAT (repeat)	0.6667 (0.5809 - 0.7558)	0.6846 (0.6063 - 0.7637)	0.6900 (0.6099 - 0.7871)
TA[d]	0.7998 (0.7461 - 0.8540)	0.8176 (0.7729 - 0.8603)	0.8164 (0.7692 - 0.8618)
TA (repeat)	0.7737 (0.6960 - 0.8446)	0.7984 (0.7290 - 0.8614)	0.7962 (0.7262 - 0.8635)

[a]SSST: sing-a-song-stress test.

[b]RAVEN: Raven's Progressive Matrices.

[c]PASAT: paced auditory serial addition task.

[d]TA: tone avoidance.

While the relatively poor model performance on the RAVEN stressor likely stems from its weaker elicited physiological responses (AVNN reactivity is displayed in Figure 2), the similarly low performance on SSST is more surprising. One possible explanation is that SSST elicits a fundamentally different physiological response compared with the other mental stressors, which could account for the poor generalization of the ML models to the SSST stressor observed in Table 3. To test this hypothesis, we retrained our LR and XGBoost classifiers separately on each stressor, that is, training and testing exclusively on one condition at a time. Multimedia Appendix 14 summarizes these per-stressor results, showing that both models outperform random BLs across all mental stressors. Notably, SSST shows marked improvement: despite both models struggling to detect it in earlier analyses, it now achieves an AUROC of 0.73 for LR and 0.76 for XGBoost, comparable to other mental stressors. This indicates that the SSST task elicits detectable MS responses. However, when SSST is excluded during training, AUROC drops sharply when tested on it (Table 3). Together, this indicates that while our ML models can learn to discriminate between SSST and nonstress when trained on SSST stressors alone, the patterns that the ML models learn from other mental stressors do not generalize to SSST. Interestingly, a similar pattern is also observed for the RAVEN stressor, further supporting this interpretation.

Discrimination From Physical Activity

Given that MS can elicit ECG responses similar to those induced by physical activity (Figure 1), our ML models may conflate metabolic demand with MS. We therefore examined condition-specific sensitivity and specificity for light (LPA) and MPA (at 1000 sampling rates, 55 features).

The XGBoost model achieved good sensitivity for MS (0.800; 95% CI 0.698 - 0.889), with the LR also yielding acceptable performance (0.772; 95% CI 0.656 - 0.877). For the LPA, XGBoost achieved a specificity of 0.787 (95% CI 0.662 - 0.903), whereas LR performed comparably, with a bootstrapped mean specificity of 0.794 (95% CI 0.685 - 0.895). However, our models performed worse for MPA (LR: 0.444, 95% CI 0.314 - 0.576; XGBoost: 0.418, 95% CI 0.299 - 0.542), suggesting that even with a comprehensive feature set of 55 features, distinguishing stress-induced ECG responses from those elicited by MPA remains challenging.

Discussion

Principal Results

We developed ML models to distinguish MS from a composite nonstress BL encompassing various everyday activities (eg, walking at a normal pace, recovering from MS, or vacuum cleaning). Using 55 features derived from the ECG signal, XGBoost achieved an AUROC of 0.741 and AUPRC of 0.706, while LR performed comparably (AUROC: 0.724; AUPRC: 0.692) at 1000 Hz. Both models achieved high sensitivity (XGBoost: 0.800; LR: 0.782) but low specificity (XGBoost: 0.512; LR: 0.509), with this pattern being particularly pronounced for MPA. Our results, therefore, highlight a fundamental challenge in distinguishing stress-induced cardiac responses from those driven by physical exertion when relying solely on ECG features. The ML models used in our study also demonstrated robustness to ECG signal downsampling, retaining more than 93% of performance using only 10 features.

While the nonlinear model (XGBoost) numerically outperformed our linear model (LR), the performance differences were small (<0.02 difference in AUROC and AUPRC at 1000 Hz), suggesting that complex nonlinear models offer only marginal benefits relative to linear BLs. One potential explanation is the so-called Rashomon effect [80], which posits that there exist many equally well-performing models for specific datasets—a phenomenon often observed in high-stakes applications and tabular datasets [81]. High outcome variability, likely present in our dataset, as physiological responses to mental stressors are both participant- and stressor-specific (Figure 2 displays the HRV response), increases the likelihood that a simple model

XSL•FO
RenderX

will perform on par with a more complex one [82]. The fact that we do not apply time-series models to the raw ECG dataset using, for instance, deep learning approaches and instead work with tabular data also likely further contributes to this phenomenon.

Our findings reveal a clear trade-off between window size and model performance, as increasing the window size led to model improvements. This can be explained by more reliable features, as shortening the window size has been shown to affect the quality of HRV features [25]. Deep learning approaches, such as convolutional or recurrent neural networks, can partially overcome this trade-off, as they learn feature representations directly from the ECG signals and have been shown to achieve promising performance, even using shorter window sizes than those considered in our study [83,84]. However, these models are often considered "black boxes" and typically require more data for training than the ML models considered in our study. Although recent advances in time-series explainability methods, such as those by Crabbé and Schaar [85] and Enguehard [86], have been proposed to overcome the challenge of interpreting deep learning time-series models, these approaches typically highlight only key segments in the time-series signal driving the model prediction. However, interpreting raw time series is challenging for humans. In contrast, our ML models use features based on well-studied physiological constructs, thereby facilitating the interpretation of model predictions using methods such as SHAP.

The feature importance analysis revealed the importance of fuzzy entropy for discriminating MS responses from other nonstress states. This finding aligns with prior work, which indicates that entropy-based measures, capturing the irregularity of the ECG signal, increase during periods of MS [87]. Notably, 2 HR fragmentation indices, PSS and IALS, also emerged as important features, despite being less commonly used in ML models for detecting MS. Our findings further highlight the importance of frequency-based features, consistent with a comprehensive review [28] covering 37 studies. In their studies, the authors identified changes in the low- and high-frequency components of the ECG signal as the most reported marker of MS.

Additionally, our analysis identified the VHF (as opposed to the UHF) band as an important predictor. Although the physiological origin of the VHF band is less well understood than that of the LF or HF bands, its predictive value likely reflects stress-induced hyperventilation, a well-documented physiological response to various stressors [35-37]. While the VHF band remained an important feature for distinguishing MS from a seated BL (Multimedia Appendix 9; Figures 3-5), suggesting a physiological contribution under sedentary conditions, we acknowledge that this band might also represent motion artifacts or electromyogram noise during active conditions. The low specificity observed during MPA may partially reflect this confound: If VHF power increases during physical activity, due to motion artifacts rather than vagal modulation, the model may misclassify physical exertion as MS. Further research on this band under stress is needed, particularly using accelerometry or respiratory measurements to distinguish stress-related changes from motion artifacts.

Our leave-one-stressor-out experiments demonstrated that our ML models could generalize to most held-out stressors, though performance varied by stressor. While SSST and Raven proved to be the most difficult stressors to recognize, the opposite was observed for TA: both models detected MS evoked by TA and its repeat condition, even when they had never been trained on it. This suggests that the TA stressor produces a strong and distinct physiological ECG response that can be easily distinguished from nonstress states. Whether this is related to differential psychological processes engaged by these tasks is an open question. The TA evokes frustration and a fear of punishment, whereas the SSST is social-evaluative and invokes a fear of exclusion, similar to Raven and mental arithmetic tasks, which also involve ego threats. Building on this, our findings imply that robust, generalizable models require training on a diverse set of various distinct mental stressors, or, when constrained, focus on mental stressors that evoke a strong physiological response, such as TA.

Although our results demonstrate that the ML models can detect MS with high sensitivity, distinguishing stress-evoked ECG responses from those induced by MPA remains especially challenging (specificity LR: 0.444; XGBoost: 0.418), even with a comprehensive set of 55 handcrafted features. Because deep learning models can learn more nuanced temporal ECG patterns, a possible path forward is to apply time-series deep learning directly to raw ECG data. Moreover, incorporating additional sensing modalities (eg, accelerometer or electrodermal activity) may help the model separate cardiac responses driven by metabolic demand from those driven by psychological stress. We therefore consider deep learning–based multimodal approaches an important direction for future research.

Importantly, our results suggest potential for translation to photoplethysmography (PPG)-based settings, as the most predictive features in our ML models are derived from beat-to-beat dynamics, specifically, the R-peaks and interbeat intervals, rather than ECG waveform morphology (eg, T-wave amplitude). Notably, the XGBoost model maintains most of its performance with 10 features (AUROC: 0.729) or 5 features (AUROC: 0.716; Figure 6). These feature subsets consist of PPG-compatible features (HR statistics and frequency-domain features; Multimedia Appendix 11), thereby supporting the possibility that our results may generalize to PPG signals. Consistent with this potential, Taoum et al [88] demonstrated that HRV metrics derived from PPG agreed with the ECG counterpart for 37 out of 48 features for recordings of 5 minutes and 30 seconds, and further for 16 features in 1-minute and 30-second windows. However, since we train on 30-second ECG epochs, future work must determine whether PPG devices can reliably extract those intervals in similarly short windows. Encouragingly, our models remain robust even when the ECG is downsampled to 125 Hz, supporting deployment on lower-resolution PPG wearables. Yet, PPG signals are more susceptible to motion artifacts than research-grade ECG [89], which may challenge the direct transfer of our feature pipeline. Given these considerations, validating and adapting our models for real-world PPG recordings, particularly under movement conditions and within 30-second windows, is therefore an

XSL•FO
RenderX

important direction for future research and essential before our approach can be deployed on PPG-based wearables.

Comparison With Prior Work

Dataset and Experimental Paradigms

MS detection has gained significant traction in recent years (Gedam and Paul [90] and Sharma and Gedeon [91] provide comprehensive reviews, Can et al [92] provides stress detection in daily life scenarios, and Pataca et al [93] provides wearable-based approaches), driven in part by the collection of large, multimodal datasets, some of which are publicly available, such as SWELL knowledge work [14], wearable stress and affect detection [13], StressID [61], or ForDigitStress [62]. While these datasets vary widely in modality, stressor type, and experimental setup, the majority rely on laboratory settings, inducing stress using the Trier Social Stress Test [94], variants of the Stroop task [95], or mental arithmetic challenges [63], all of which are effective methods for inducing acute stress [96].

Even though these datasets have significantly advanced the field, our experimental protocol and its associated dataset offer 2 distinct advantages compared with prior work.

First, our dataset includes 127 participants, which exceeds the sample sizes of most comparable ECG-based stress datasets [13,14,61]. Our relatively large sample size helps to enhance the robustness and generalizability of our findings, especially given the high between-individual variability in physiological stress responses [26,63,97] (also present in our study; see Figure 2). The small number of participants has also been noted as a limitation of prior work in a systematic review [98]. Notably, Smets et al [99] include 1002 participants; however, their study relies on an ambulatory design that uses self-reported stress via ecological momentary assessment. In contrast, our controlled laboratory protocol induces MS through standardized tasks and captures objective ECG measurements, which can be continuously observed.

Second, our experimental design incorporates 4 distinct mental stressors—2 of which are repeated—as well as a diverse set of nonstress conditions, including seated BL, recovery, or common everyday activities such as walking at a normal pace. This comprehensive experimental setup allows us to address a more ecologically valid and challenging classification problem: evaluating whether mental-stress-specific ECG patterns can be distinguished not only from rest, but also from other non-resting states.

ML Performance and Analysis

Most ECG-based stress-detection studies frame the problem as distinguishing MS from a seated, nonstress BL [13,61,62,100]. In contrast, Sun et al [63] achieved 80.9% between-individual accuracy in 20 participants by using an activity-aware, multimodal classifier, therefore combining ECG, accelerometer, and galvanic skin response to detect MS while sitting, standing, and walking. When the accelerometer input was removed, their performance dropped substantially, highlighting the importance of tracking the motion context. Although our ECG-only model does not match the multimodal performance reported by Sun et al [63], it is worth noting that their approach relies on a 60-second window. In contrast, our study uses a 30-second window, which, as stated above, reduces model performance but enables faster detection. Moreover, our laboratory setup encompasses a broader range of mental stressors and leverages an expanded ECG feature set, highlighting the potential of feature-rich, single-sensor models.

Previous work has often simplified stress detection to distinguishing MS from a seated BL [13,61]. In contrast, our classification task differentiates stress from a composite set of nonstressful everyday activities (eg, walking at a normal pace, poststress recovery, and vacuum cleaning), yielding a more ecologically valid scenario in which people naturally move and recover throughout their day. However, this setup makes our classification task more challenging, and consequently, performance is lower than in studies limited to seated nonstress conditions. When we rerun our models on the classic stress versus seated BL task, AUROC and AUPRC both increase markedly (Multimedia Appendix 12), confirming that the drop in metrics stems from the more demanding classification problem.

Beyond the expanded task, our approach also leverages a rich feature space of 55 time-, frequency-, nonlinear, and morphological ECG features, exceeding the number of ECG predictors used in other studies [13,61,101]. We also include and investigate less commonly considered features, such as HR fragmentation indices, which proved to be of added value according to our feature importance analysis (PSS and IALS). To the best of our knowledge, no previous work has evaluated the impact of aggressive downsampling on downstream model performance. We show that reducing the sampling rate from 1000 to 125 Hz does not degrade performance. While our models were robust to sampling rates, the choice of the window size affected model performance: increasing the window size from 30 to 60 s led to notable performance improvements for both ML models. Consequently, our findings provide valuable insights into the effects of sampling rate and window size choices for MS detection, thereby informing the design of models for resource-constrained wearables.

Model Generalization Across Mental Stressors

Most mental-stress classification studies evaluate models only on the same stressors used during training, while far fewer assess generalization to unseen stressors—a critical capability for real-world deployment. Prior work often evaluates generalizability by training on one dataset and testing on another [17-19,101]. While informative, such cross-dataset evaluations conflate multiple sources of variation—individual demographics, sensor hardware, and stressor types—making it challenging to attribute performance drops to any specific factor. In contrast, our leave-one-stressor-out evaluation isolates stressor novelty, enabling a more targeted assessment of a model's ability to detect MS under previously unseen conditions. As our leave-one-out-stressor experiments demonstrated, both LR and XGBoost could generalize to novel stressors, indicating the feasibility of ML models to adapt to unseen mental stimuli.

Taken together, our work advances MS detection from several perspectives. First, by leveraging an ECG dataset from 127 participants across 6 mental stressors and a diverse set of

XSL•FO
RenderX

nonstress activities, our work addresses a more challenging, ecologically valid task: separating MS from a composite nonstress set that includes everyday activities, for which we demonstrate acceptable performance. Second, we demonstrate that unimodal ML ECG classifiers using 55 features are robust against downsampling from 1000 to 125 Hz, maintaining a performance of over 95% with just 10 features. This highlights the feasibility of lightweight and interpretable models for wearable deployment. Third, through our leave-one-stressor-out evaluation, we provide valuable insights into model generalization to unseen stressors, isolating the challenge of stressor novelty without confounds from demographic or hardware differences.

Limitations

Despite the results and insights presented in the paper, a few limitations warrant attention.

First, our cohort comprises Dutch-speaking young working adults (aged 18 - 48 years). While our analytical framework is broadly applicable, physiological stress responses are highly individual and may vary across age groups, cultures, and health statuses.

Additionally, activity patterns might change over the lifespan [102], further limiting generalizability. Importantly, our results do not directly generalize to older populations, given that HRV, HR reserve, and other time-domain features generally decline with age [103], potentially making the classification problem more difficult. Given that distinguishing MS from cardiac responses evoked by physical activity proved very challenging despite our comparable young cohort, we anticipate that this challenge will persist (or even become more complicated) when considering an older population. An interesting direction for future work is therefore to validate our models in more diverse populations, such as older adults or non-Western cohorts, to ensure robustness and generalizability of our findings.

Second, we used global z-standardization across the training set to enable generalization to unseen users without requiring per-user calibration data. While this approach is attractive for real-world deployment, it does not account for the law of initial values [104], that is, it ignores interindividual variability in physiological responses to mental stressors present in our study (Figure 2). Our global normalization scheme may therefore bias model predictions against specific physiological phenotypes, for instance, by increasing false-positive rates among individuals with elevated resting HR or reduced BL HRV. Future work could examine the extent to which global versus individual-aware normalization strategies systematically affect downstream model performances across different BL physiological characteristics.

Third, we rely solely on task labels rather than participants' self-reports to assign MS conditions, which prevents us from confirming that every labeled stress window corresponds to actual perceived stress. Incorporating self-assessments of perceived stress could have therefore reduced label noise, particularly given the pronounced interindividual variability in physiological responses (Figure 2 and Multimedia Appendix 3), which may partly account for the observed failure to generalize to the SSST stressor. Furthermore, instead of treating stress as a binary variable (stress or nonstress), one could use perceived ratings to classify the specific level or intensity of perceived MS [105].

Fourth, a related limitation of this study concerns the labeling of recovery periods. Physiological recovery from stress is gradual rather than instantaneous, suggesting that segments labeled "non-stress" during recovery periods likely retain residual stress physiology, constituting class-dependent label noise [106]. Empirically, label noise has been shown to adversely affect model performance across various ML approaches, including deep learning methods [107-109]. Our sensitivity analysis supports this observation: excluding recovery periods improved AUPRC for LR and XGBoost (see "Model Performance under Simplified Conditions"). We retained recovery periods as nonstress for comparability with prior stress-detection literature, which commonly assigns labels based on the protocol, and to provide conservative performance estimates. Yet, an alternative approach could either treat recovery as a buffer period excluded from training and evaluation or apply methods such as importance reweighting [110] or weighted surrogate loss [111].

Fifth, while our study uses ECG data to detect MS and downsamples the signal to mimic the hardware constraints of consumer-grade devices, such wearables typically rely on PPG. The most predictive features in our models are derived from R-peaks and beat-to-beat dynamics, which can also be extracted from PPG, suggesting that our approach may be transferable. However, this hypothesis remains to be validated empirically. Future work should therefore assess the extent to which our findings can be generalized to PPG-based settings.

Last, the MS tasks in our experimental setup were performed while participants remained seated. In everyday life, however, individuals often experience MS while standing or moving, for example, a teacher navigating a lively classroom. Future work should therefore extend the evaluation to simultaneous stress-and-activity scenarios to fully validate ambulatory MS detection, as done by Sun et al [63], Hosseini et al [112], and Kaczor et al [113].

Conclusions

This study evaluated ML models for distinguishing MS from routine physical activities using a single-sensor ECG. Both LR and XGBoost achieved acceptable discriminative performance (AUROC >0.7), with XGBoost providing only marginal benefits over the linear BL ($\Delta<0.02$), suggesting that simple, interpretable models can perform competitively. Performance remained robust when downsampling from 1000 to 125 Hz and reducing the feature set to 10, thereby supporting lightweight deployment. However, our findings reveal a critical limitation of single-sensor ECG approaches: although both models detected MS with acceptable to good sensitivity (XGBoost: 0.800; LR: 0.782), specificity was low (XGBoost: 0.512; LR: 0.509), particularly for MPA. This indicates that stress-induced cardiac responses cannot be reliably distinguished from those driven by physical exertion using ECG features alone. Generalization to unseen stressors was also stressor-dependent, with models performing well above chance for most stressors but near chance

XSL•FO
RenderX

for the social-evaluative stimulus (SSST). Future studies should validate our findings in ambulatory settings where MS and physical activity can co-occur and explore both multimodal approaches and the use of PPG via wearables as a practical alternative to ECG for continuous stress monitoring.

Acknowledgments

This article was written as part of the research project “Stress in Action”. We used artificial intelligence tools, specifically ChatGPT and Claude, to speed up coding and debugging. Furthermore, we used these tools to improve the syntax, clarity, and grammar of the written paragraphs. Importantly, all outputs were carefully checked and reviewed.

Funding

Stress in Action is financially supported by the Dutch Research Council and the Dutch Ministry of Education, Culture, and Science (NWO gravitation grant number 024.005.010).

Data Availability

The dataset analyzed in this study is not publicly available, and a data access agreement is required to access and analyze the data. The code is publicly available on the GitHub repository [75].

Authors' Contributions

BU contributed to the conceptualization, formal analysis, coding, methodology, and writing of the manuscript. AA contributed to formal analysis, coding, and investigation. SvdV and DJvdM provided consultation and contributed to the review and editing of the manuscript. EdG and MH provided supervision and also contributed to the review and editing of the manuscript.

Conflicts of Interest

EdG owns shares in VU Ambulatory Monitoring Solutions Ltd. The remaining authors declare no competing interests.

Multimedia Appendix 1
Overview of features.
[PDF File, 53 KB - jmir_v28i1e80450_app1.pdf]

Multimedia Appendix 2
Hyperparameter settings.
[PDF File, 26 KB - jmir_v28i1e80450_app2.pdf]

Multimedia Appendix 3
Boxplots positive and negative affect.
[PDF File, 448 KB - jmir_v28i1e80450_app3.pdf]

Multimedia Appendix 4
Area under the precision-recall curve classification performance.
[PDF File, 133 KB - jmir_v28i1e80450_app4.pdf]

Multimedia Appendix 5
CIs for performance differences between XGBoost and LR.
[PDF File, 40 KB - jmir_v28i1e80450_app5.pdf]

Multimedia Appendix 6
Additional classification performance results.
[PDF File, 685 KB - jmir_v28i1e80450_app6.pdf]

Multimedia Appendix 7
Seated baseline classification performance.
[PDF File, 358 KB - jmir_v28i1e80450_app7.pdf]

Multimedia Appendix 8

XSL•FO
RenderX

Time window analysis.
[PDF File, 339 KB - jmir_v28i1e80450_app8.pdf]

Multimedia Appendix 9
Model explainability.
[PDF File, 764 KB - jmir_v28i1e80450_app9.pdf]

Multimedia Appendix 10
Feature parsimony.
[PDF File, 151 KB - jmir_v28i1e80450_app10.pdf]

Multimedia Appendix 11
Overview of selected features.
[PDF File, 33 KB - jmir_v28i1e80450_app11.pdf]

Multimedia Appendix 12
Generalization to unknown stressors.
[PDF File, 56 KB - jmir_v28i1e80450_app12.pdf]

Multimedia Appendix 13
Stratified stressor analysis.
[PDF File, 47 KB - jmir_v28i1e80450_app13.pdf]

Multimedia Appendix 14
Stratified stressor analysis - single sensor training.
[PDF File, 49 KB - jmir_v28i1e80450_app14.pdf]

References

1. Razavi M, Ziyadidegan S, Mahmoudzadeh A, et al. Machine learning, deep learning, and data preprocessing techniques for detecting, predicting, and monitoring stress and stress-related mental disorders: scoping review. JMIR Ment Health 2024 Aug 21;11:e53714. [doi: 10.2196/53714] [Medline: 39167782]
2. Cohen S, Gianaros PJ, Manuck SB. A stage model of stress and disease. Perspect Psychol Sci 2016 Jul;11(4):456-463. [doi: 10.1177/1745691616646305] [Medline: 27474134]
3. Le Fevre M, Matheny J, Kolt GS. Eustress, distress, and interpretation in occupational stress. J Manag Psychol 2003 Nov 1;18(7):726-744. [doi: 10.1108/02683940310502412]
4. Cohen S, Janicki-Deverts D, Miller GE. Psychological stress and disease. JAMA 2007 Oct 10;298(14):1685-1687. [doi: 10.1001/jama.298.14.1685] [Medline: 17925521]
5. Hammen C. Stress and depression. Annu Rev Clin Psychol 2005;1(293-319):293-319. [doi: 10.1146/annurev.clinpsy.1.102803.143938] [Medline: 17716090]
6. Kivimäki M, Steptoe A. Effects of stress on the development and progression of cardiovascular disease. Nat Rev Cardiol 2018 Apr;15(4):215-229. [doi: 10.1038/nrcardio.2017.189] [Medline: 29213140]
7. Jönsson P, Wallergård M, Osterberg K, Hansen AM, Johansson G, Karlson B. Cardiovascular and cortisol reactivity and habituation to a virtual reality version of the trier social stress test: a pilot study. Psychoneuroendocrinology 2010 Oct;35(9):1397-1403. [doi: 10.1016/j.psyneuen.2010.04.003] [Medline: 20451329]
8. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [Medline: 6668417]
9. Martinez GJ, Grover T, Mattingly SM, et al. Alignment between heart rate variability from fitness trackers and perceived stress: perspectives from a large-scale in situ longitudinal study of information workers. JMIR Hum Factors 2022 Aug 4;9(3):e33754. [doi: 10.2196/33754] [Medline: 35925662]
10. Choi J, Gutierrez-Osuna R. Using heart rate monitors to detect mental stress. 2009 Presented at: Proceedings of the 2009 Sixth International Workshop on Wearable and Implantable Body Sensor Networks; Jun 3-5, 2009.
11. Won E, Kim YK. Stress, the autonomic nervous system, and the immune-kynurenine pathway in the etiology of depression. Curr Neuropharmacol 2016;14(7):665-673. [doi: 10.2174/1570159x14666151208113006] [Medline: 27640517]
12. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science 2015 Jul 17;349(6245):255-260. [doi: 10.1126/science.aaa8415] [Medline: 26185243]

XSL•FO
RenderX

13. Schmidt P, Reiss A, Duerichen R, Marberger C, Van Laerhoven K. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. Presented at: ICMI '18: Proceedings of the 20th ACM International Conference on Multimodal Interaction; Oct 16-20, 2018; Boulder CO USA p. 400-408. [doi: 10.1145/3242969.3242985]
14. Koldijk S, Sappelli M, Verberne S, Neerincx MA, Kraaij W. The swell knowledge work dataset for stress and user modeling research. Presented at: ICMI '14: Proceedings of the 16th International Conference on Multimodal Interaction; Nov 12-16, 2014. [doi: 10.1145/2663204.2663257]
15. Li RA, Liu Z. Stress detection using deep neural networks. BMC Med Inform Decis Mak 2020 Dec;20(S11). [doi: 10.1186/s12911-020-01299-4]
16. Zhu L, Spachos P, Ng PC, et al. Stress detection through wrist-based electrodermal activity monitoring and machine learning. IEEE J Biomed Health Inform 2023;27(5):2155-2165. [doi: 10.1109/JBHI.2023.3239305]
17. Prajod P, Mahesh B, André E. Stressor type matters! --- exploring factors influencing cross-dataset generalizability of physiological stress detection. Presented at: ICMI '24: Proceedings of the 26th International Conference on Multimodal Interaction; Nov 4-8, 2024; San Jose Costa Rica p. 508-517. [doi: 10.1145/3678957.3685738]
18. Mishra V, Sen S, Chen G, et al. Evaluating the reproducibility of physiological stress detection models. Proc ACM Interact Mob Wearable Ubiquitous Technol 2020 Dec;4(4):147. [doi: 10.1145/3432220] [Medline: 36189150]
19. Cho HM, Park H, Dong SY, Youn I. Ambulatory and laboratory stress detection based on raw electrocardiogram signals using a convolutional neural network. Sensors (Basel) 2019 Oct 11;19(20):4408. [doi: 10.3390/s19204408] [Medline: 31614646]
20. van der Mee DJ, Gevonden MJ, Westerink J, de Geus EJC. Validity of electrodermal activity-based measures of sympathetic nervous system activity from a wrist-worn device. Int J Psychophysiol 2021 Oct;168:52-64. [doi: 10.1016/j.ijpsycho.2021.08.003] [Medline: 34418464]
21. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMC Med 2015 Jan 6;13(1):1. [doi: 10.1186/s12916-014-0241-z] [Medline: 25563062]
22. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:e078378. [doi: 10.1136/bmj-2023-078378]
23. Myin-Germeys I, van Os J, Schwartz JE, Stone AA, Delespaul PA. Emotional reactivity to daily life stress in psychosis. Arch Gen Psychiatry 2001 Dec;58(12):1137-1144. [doi: 10.1001/archpsyc.58.12.1137] [Medline: 11735842]
24. Makowski D, Pham T, Lau ZJ, et al. NeuroKit2: a Python toolbox for neurophysiological signal processing. Behav Res Methods 2021 Aug;53(4):1689-1696. [doi: 10.3758/s13428-020-01516-y] [Medline: 33528817]
25. Bernardes A, Couceiro R, Medeiros J, et al. How reliable are ultra-short-term HRV measurements during cognitively demanding tasks? Sensors (Basel) 2022 Aug 30;22(17):6528. [doi: 10.3390/s22176528] [Medline: 36080987]
26. Smets E, Casale P, Großekathöfer U, Lamichhane B, Raedt W, Bogaerts K. Comparison of Machine Learning Techniques for Psychophysiological Stress Detection Pervasive Computing Paradigms for Mental Health: Springer International Publishing; 2015. [doi: 10.1007/978-3-319-32270-4_2]
27. Kaegi DM, Halamek LP, Van Hare GF, Howard SK, Dubin AM. Effect of mental stress on heart rate variability: validation of simulated operating and delivery room training modules. Pediatr Res 1999 Apr;45(4, Part 2 of 2):77A-77A. [doi: 10.1203/00006450-199904020-00463]
28. Kim HG, Cheon EJ, Bai DS, Lee YH, Koo BH. Stress and heart rate variability: a meta-analysis and review of the literature. Psychiatry Investig 2018 Mar;15(3):235-245. [doi: 10.30773/pi.2017.08.17]
29. Kleiger RE, Stein PK, Bosner MS, Rottman JN. Time domain measurements of heart rate variability. Cardiol Clin 1992 Aug;10(3):487-498. [doi: 10.1016/s0733-8651(18)30230-3] [Medline: 1504980]
30. Dalmeida KM, Masala GL. HRV features as viable physiological markers for stress detection using wearable devices. Sensors (Basel) 2021 Apr 19;21(8):2873. [doi: 10.3390/s21082873] [Medline: 33921884]
31. Wang HM, Huang SC. SDNN/RMSSD as a surrogate for LF/HF: a revised investigation. Model Simul Eng 2012;2012:1-8. [doi: 10.1155/2012/931943]
32. Sollers JJ 3rd, Buchanan TW, Mowrer SM, Hill LK, Thayer JF. Comparison of the ratio of the standard deviation of the R-R interval and the root mean squared successive differences (SD/rMSSD) to the low frequency-to-high frequency (LF/HF) ratio in a patient population and normal healthy controls. Biomed Sci Instrum 2007;43:158-163. [Medline: 17487074]
33. Salahuddin L, Cho J, Jeong MG, Kim D. Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings. Annu Int Conf IEEE Eng Med Biol Soc 2007;2007:4656-4659. [doi: 10.1109/IEMBS.2007.4353378] [Medline: 18003044]
34. Pereira T, Almeida PR, Cunha JPS, Aguiar A. Heart rate variability metrics for fine-grained stress level assessment. Comput Methods Programs Biomed 2017 Sep;148:71-80. [doi: 10.1016/j.cmpb.2017.06.018] [Medline: 28774440]
35. Tipton MJ, Harper A, Paton JFR, Costello JT. The human ventilatory response to stress: rate or depth? J Physiol (Lond) 2017 Sep;595(17):5729-5752. [doi: 10.1113/JP274596]
36. Masaoka Y, Homma I. Anxiety and respiratory patterns: their relationship during mental stress and physical load. Int J Psychophysiol 1997 Sep;27(2):153-159. [doi: 10.1016/s0167-8760(97)00052-4] [Medline: 9342646]

XSL•FO
RenderX

37. Masaoka Y, Homma I. The effect of anticipatory anxiety on breathing and metabolism in humans. Respir Physiol 2001 Nov 1;128(2):171-177. [doi: 10.1016/s0034-5687(01)00278-x] [Medline: 11812381]
38. Hernando A, Lazaro J, Gil E, et al. Inclusion of respiratory frequency information in heart rate variability analysis for stress assessment. IEEE J Biomed Health Inform 2016 Jul;20(4):1016-1025. [doi: 10.1109/JBHI.2016.2553578] [Medline: 27093713]
39. Karthikeyan P, Murugappan M, Yaacob S. A study on mental arithmetic task based human stress level classification using discrete wavelet transform. Presented at: 2012 IEEE conference on sustainable utilization and development in engineering and technology (STUDENT2012); Oct 6-9, 2012. [doi: 10.1109/STUDENT.2012.6408369]
40. Webber CL Jr, Zbilut JP. Dynamical assessment of physiological systems and states using recurrence plot strategies. J Appl Physiol 1994 Feb 1;76(2):965-973. [doi: 10.1152/jappl.1994.76.2.965]
41. Tanev G, Saadi DB, Hoppe K, Sorensen HBD. Classification of acute stress using linear and non-linear heart rate variability analysis derived from sternal ECG. Presented at: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); Aug 26-30, 2014. [doi: 10.1109/EMBC.2014.6944349]
42. Pincus SM. Approximate entropy as a measure of system complexity. Proc Natl Acad Sci USA 1991 Mar 15;88(6):2297-2301. [doi: 10.1073/pnas.88.6.2297]
43. Richman JS, Moorman JR. Physiological time-series analysis using approximate entropy and sample entropy. Am J Physiol Heart Circ Physiol 2000 Jun;278(6):H2039-H2049. [doi: 10.1152/ajpheart.2000.278.6.H2039] [Medline: 10843903]
44. Chen W, Wang Z, Xie H, Yu W. Characterization of surface EMG signal based on fuzzy entropy. IEEE Trans Neural Syst Rehabil Eng 2007 Jun;15(2):266-272. [doi: 10.1109/TNSRE.2007.897025] [Medline: 17601197]
45. Pham T, Lau ZJ, Chen SHA, Makowski D. Heart rate variability in psychology: a review of HRV indices and an analysis tutorial. Sensors (Basel) 2021 Jun 9;21(12):3998. [doi: 10.3390/s21123998] [Medline: 34207927]
46. Humeau-Heurtier A. Evaluation of systems' irregularity and complexity: sample entropy, its derivatives, and their applications across scales and disciplines. Entropy (Basel) 2018 Oct 16;20(10):794. [doi: 10.3390/e20100794] [Medline: 33265881]
47. Parvaneh S, Toosizadeh N, Moharreri S. Impact of mental stress on heart rate asymmetry. In: Parvaneh S, Toosizadeh N, Moharreri S, editors. Presented at: 2015 Computing in Cardiology Conference (CinC); Sep 6-9, 2015. [doi: 10.1109/CIC.2015.7411030]
48. Yan C, Li P, Ji L, Yao L, Karmakar C, Liu C. Area asymmetry of heart rate variability signal. Biomed Eng Online 2017 Sep 21;16(1):112. [doi: 10.1186/s12938-017-0402-3] [Medline: 28934961]
49. Javorka M, Turianikova Z, Tonhajzerova I, Javorka K, Baumert M. The effect of orthostasis on recurrence quantification analysis of heart rate and blood pressure dynamics. Physiol Meas 2009 Jan;30(1):29-41. [doi: 10.1088/0967-3334/30/1/003] [Medline: 19039163]
50. Dimitriev D, Saperova EV, Dimitriev A, Karpenko Y. Recurrence quantification analysis of heart rate during mental arithmetic stress in young females. Front Physiol 2020;11(40):40. [doi: 10.3389/fphys.2020.00040] [Medline: 32116754]
51. Peng CK, Havlin S, Stanley HE, Goldberger AL. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos 1995 Mar 1;5(1):82-87. [doi: 10.1063/1.166141]
52. Tulppo MP, Mäkikallio TH, Seppänen T, et al. Effects of pharmacological adrenergic and vagal modulation on fractal heart rate dynamics. Clin Physiol 2001 Sep 24;21(5):515-523. [doi: 10.1046/j.1365-2281.2001.00344.x]
53. Immanuel S, Teferra MN, Baumert M, Bidargaddi N. Heart rate variability for evaluating psychological stress changes in healthy adults: a scoping review. Neuropsychobiology 2023;82(4):187-202. [doi: 10.1159/000530376] [Medline: 37290411]
54. Goya-Esteban R, Barquero-Pérez O, Sarabia-Cachadina E, Naranjo-Orellana J, Rojo-Alvarez JL. Heart rate variability non linear dynamics in intense exercise. 2010 Presented at: Computing in Cardiology; Sep 9-12, 2012.
55. Costa MD, Davis RB, Goldberger AL. Heart rate fragmentation: a new approach to the analysis of cardiac interbeat interval dynamics. Front Physiol 2017 May;8. [doi: 10.3389/fphys.2017.00255]
56. Lampert R. ECG signatures of psychological stress. J Electrocardiol 2015;48(6):1000-1005. [doi: 10.1016/j.jelectrocard.2015.08.005] [Medline: 26364755]
57. Aro AL, Kenttä TV, Huikuri HV. Microvolt T-wave alternans: where are we now? Arrhythm Electrophysiol Rev 2016 May;5(1):37-40. [doi: 10.15420/aer.2015.28.1] [Medline: 27403292]
58. Nearing BD, Verrier RL. Modified moving average analysis of T-wave alternans to predict ventricular fibrillation with high accuracy. J Appl Physiol (1985) 2002 Feb;92(2):541-549. [doi: 10.1152/japplphysiol.00592.2001] [Medline: 11796662]
59. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13-17, 2016; San Francisco California USA p. 785-794.
60. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1):310. [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
61. Chaptoukaev H, Strizhkova V, Panariello M, et al. StressID: a multimodal dataset for stress identification. Presented at: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track; Dec 10-16, 2023.
62. Heimerl A, Prajod P, Mertes S, et al. The ForDigitStress dataset: a multi-modal dataset for automatic stress recognition. IEEE Trans Affective Comput ;16(2):1219-1234. [doi: 10.1109/TAFFC.2024.3501400]

XSL•FO
RenderX

63. Sun FT, Kuo C, Cheng HT, Buthpitiya S, Collins P, Griss M, editors. Activity-aware mental stress detection using physiological sensors. Presented at: Mobile Computing, Applications, and Services MobiCASE 2010; Oct 25-28, 2010.
64. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data. Presented at: Proceedings of the 36th International Conference on Neural Information Processing Systems; Nov 28 to Dec 3, 2022; New Orleans, Louisiana, United States p. 507-520.
65. Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. Information Fusion 2022 May;81(C):84-90. [doi: 10.1016/j.inffus.2021.11.011]
66. Schallmoser S, Zueger T, Kraus M, Saar-Tsechansky M, Stettler C, Feuerriegel S. Machine learning for predicting micro- and macrovascular complications in individuals with prediabetes or diabetes: retrospective cohort study. J Med Internet Res 2023;25:e42181. [doi: 10.2196/42181]
67. Healey JA, Picard RW. Detecting stress during real-world driving tasks using physiological sensors. IEEE Trans Intell Transport Syst 2005 Jun;6(2):156-166. [doi: 10.1109/TITS.2005.848368]
68. Saygin M, Schoenmakers M, Gevonden M, de Geus E. Speech detection via respiratory inductance plethysmography, thoracic impedance, accelerometers, and gyroscopes: a machine learning-informed comparative study. Psychophysiology 2025 Feb;62(2):e70021. [doi: 10.1111/psyp.70021] [Medline: 39950497]
69. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Presented at: Proceedings of the 31st International Conference on Neural Information Processing Systems; Dec 4-9, 2017; Long Beach, California, United States p. 4768-4777.
70. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. JAMA 2017 Oct 10;318(14):1377-1384. [doi: 10.1001/jama.2017.12126] [Medline: 29049590]
71. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. J Am Med Inform Assoc 2020 Apr 1;27(4):621-633. [doi: 10.1093/jamia/ocz228] [Medline: 32106284]
72. Brier GW. Verification of forecasts expressed in terms of probability. Mon Wea Rev 1950 Jan;78(1):1-3. [doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2]
73. Maiga A, Farjah F, Blume J, et al. Risk prediction in clinical practice: a practical guide for cardiothoracic surgeons. Ann Thorac Surg 2019 Nov;108(5):1573-1582. [doi: 10.1016/j.athoracsur.2019.04.126] [Medline: 31255609]
74. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. Lond Edinb Dubl Phil Mag J Sci 1901 Nov;2(11):559-572. [doi: 10.1080/14786440109462720]
75. Uendes B, Antonides A. Stress detection ECG. GitHub. 2025 Jul. URL: https://github.com/BuelentUendes/Stress_Detection_ECG [accessed 2025-07-11]
76. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16(1):321-357. [doi: 10.1613/jair.953]
77. Your research: subject to the WMO or not? Central Committee on Research Involving Human Subjects. URL: https://english.ccmo.nl/investigators/legal-framework-for-medical-scientific-research/your-research-is-it-subject-to-the-wmo-or-not [accessed 2026-03-22]
78. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988 Sep;44(3):837-845. [doi: 10.2307/2531595] [Medline: 3203132]
79. Obuchowski NA. Nonparametric analysis of clustered ROC curve data. Biometrics 1997 Jun;53(2):567-578. [doi: 10.2307/2533958] [Medline: 9192452]
80. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Statist Sci 2001 Aug;16(3):199-231. [doi: 10.1214/ss/1009213726]
81. Rudin C, Zhong C, Semenova L, Seltzer M, Parr R, Liu J, et al. Position: amazing things come from having many good models. In: Ruslan S, Zico K, Katherine H, Adrian W, Nuria O, Jonathan S, editors. Presented at: Proceedings of the 41st International Conference on Machine Learning; Jul 21-27, 2024; Vienna, Austria p. 42783-42795.
82. Semenova L, Chen H, Parr R, Rudin C. A path to simpler models starts with noise. Adv Neural Inf Process Syst 2023 Dec;36:3362-3401. [Medline: 38577617]
83. Hwang B, You J, Vaessen T, Myin-Germeys I, Park C, Zhang BT. Deep ECGNet: an optimal deep learning framework for monitoring mental stress using ultra short-term ECG signals. Telemed J E Health 2018 Oct;24(10):753-772. [doi: 10.1089/tmj.2017.0250] [Medline: 29420125]
84. Zhang P, Li F, Zhao R, et al. Real-time psychological stress detection according to ECG using deep learning. Appl Sci (Basel) 2021;11(9):3838. [doi: 10.3390/app11093838]
85. Crabbé J, Schaar MVD. Explaining time series predictions with dynamic masks. In: Marina M, Tong Z, editors. Presented at: Proceedings of the 38th International Conference on Machine Learning; Jul 18-24, 2021; Virtual / Online p. 2166-2177.
86. Enguehard J. Learning perturbations to explain time series predictions. In: Andreas K, Emma B, Kyunghyun C, Barbara E, Sivan S, Jonathan S, editors. Presented at: Proceedings of the 40th International Conference on Machine Learning; Jul 23-29, 2023; Honolulu, Hawaii, United States p. 9329-9342.
87. Chand K, Chandra S, Dutt V. A comprehensive evaluation of linear and non-linear HRV parameters between paced breathing and stressful mental state. Heliyon 2024 Jun;10(11):e32195. [doi: 10.1016/j.heliyon.2024.e32195]

XSL•FO
RenderX

88. Taoum A, Bisiaux A, Tilquin F, Le Guillou Y, Carrault G. Validity of ultra-short-term HRV analysis using PPG—a preliminary study. Sensors (Basel) 2022 Oct 20;22(20):7995. [doi: 10.3390/s22207995]
89. Maeda Y, Sekine M, Tamura T. Relationship between measurement site and motion artifacts in wearable reflected photoplethysmography. J Med Syst 2011 Oct;35(5):969-976. [doi: 10.1007/s10916-010-9505-0]
90. Gedam S, Paul S. A review on mental stress detection using wearable sensors and machine learning techniques. IEEE Access 2021;9:84045-84066. [doi: 10.1109/ACCESS.2021.3085502]
91. Sharma N, Gedeon T. Objective measures, sensors and computational techniques for stress recognition and classification: a survey. Comput Methods Programs Biomed 2012 Dec;108(3):1287-1301. [doi: 10.1016/j.cmpb.2012.07.003] [Medline: 22921417]
92. Can YS, Arnrich B, Ersoy C. Stress detection in daily life scenarios using smart phones and wearable sensors: a survey. J Biomed Inform 2019 Apr;92:103139. [doi: 10.1016/j.jbi.2019.103139]
93. Pataca AO, Zdravevski E, Coelho PJ, et al. Use of machine learning for predicting stress episodes based on wearable sensor data: a systematic review. Comput Biol Med 2025 Nov;198:111166. [doi: 10.1016/j.compbiomed.2025.111166]
94. Kirschbaum C, Pirke KM, Hellhammer DH. The 'Trier Social Stress Test' – a tool for investigating psychobiological stress responses in a laboratory setting. Neuropsychobiology 1993;28(1-2):76-81. [doi: 10.1159/000119004]
95. Stroop JR. Studies of interference in serial verbal reactions. J Exp Psychol 1935;18(6):643-662. [doi: 10.1037/h0054651]
96. Dickerson SS, Kemeny ME. Acute stressors and cortisol responses: a theoretical integration and synthesis of laboratory research. Psychol Bull 2004 May;130(3):355-391. [doi: 10.1037/0033-2909.130.3.355]
97. Epel ES, Crosswell AD, Mayer SE, et al. More than a feeling: a unified view of stress measurement for population science. Front Neuroendocrinol 2018 Apr;49:146-169. [doi: 10.1016/j.yfrne.2018.03.001] [Medline: 29551356]
98. Vos G, Trinh K, Sarnyai Z, Rahimi Azghadi M. Generalizable machine learning for stress monitoring from wearable devices: a systematic literature review. Int J Med Inform 2023 May;173:105026. [doi: 10.1016/j.ijmedinf.2023.105026] [Medline: 36893657]
99. Smets E, Rios Velazquez E, Schiavone G, et al. Large-scale wearable data reveal digital phenotypes for daily-life stress detection. npj Digital Med 2018;1(1):67. [doi: 10.1038/s41746-018-0074-9]
100. Huang S, Li J, Zhang P, Zhang W. Detection of mental fatigue state with wearable ECG devices. Int J Med Inform 2018 Nov;119:39-46. [doi: 10.1016/j.ijmedinf.2018.08.010]
101. Han HJ, Labbaf S, Borelli JL, Dutt N, Rahmani AM. Objective stress monitoring based on wearable sensors in everyday settings. J Med Eng Technol 2020 May;44(4):177-189. [doi: 10.1080/03091902.2020.1759707] [Medline: 32589065]
102. Weber C, Quintus M, Egloff B, Luong G, Riediger M, Wrzus C. Same old, same old? age differences in the diversity of daily life. Psychol Aging 2020 May;35(3):434-448. [doi: 10.1037/pag0000407] [Medline: 31613136]
103. Umetani K, Singer DH, McCraty R, Atkinson M. Twenty-four hour time domain heart rate variability and heart rate: relations to age and gender over nine decades. J Am Coll Cardiol 1998 Mar 1;31(3):593-601. [doi: 10.1016/s0735-1097(97)00554-8] [Medline: 9502641]
104. Wilder J. The law of initial value in neurology and psychiatry. J Nerv Ment Dis 1957 Jan;125(1):73-86. [doi: 10.1097/00005053-195701000-00009]
105. Benchekroun M, Velmovitsky PE, Istrate D, Zalc V, Morita PP, Lenne D. Cross dataset analysis for generalizability of HRV-based stress detection models. Sensors (Basel) 2023 Feb 6;23(4):1807. [doi: 10.3390/s23041807] [Medline: 36850407]
106. Karimi D, Dou H, Warfield SK, Gholipour A. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. Med Image Anal 2020 Oct;65:101759. [doi: 10.1016/j.media.2020.101759] [Medline: 32623277]
107. Ding C, Pereira T, Xiao R, Lee RJ, Hu X. Impact of label noise on the learning based models for a binary classification of physiological signal. Sensors (Basel) 2022;22(19):7166. [doi: 10.3390/s22197166]
108. Song H, Kim M, Park D, Shin Y, Lee JG. Learning from noisy labels with deep neural networks: a survey. IEEE Trans Neural Netw Learning Syst 2023;34(11):8135-8153. [doi: 10.1109/TNNLS.2022.3152527]
109. Frénay B, Verleysen M. Classification in the presence of label noise: a survey. IEEE Trans Neural Netw Learn Syst 2014 May;25(5):845-869. [doi: 10.1109/TNNLS.2013.2292894] [Medline: 24808033]
110. Liu T, Tao D. Classification with noisy labels by importance reweighting. IEEE Trans Pattern Anal Mach Intell 2016;38(3):447-461. [doi: 10.1109/TPAMI.2015.2456899]
111. Natarajan N, Dhillon IS, Ravikumar P, Tewari A. Learning with noisy labels. Presented at: Proceedings of the 27th International Conference on Neural Information Processing Systems; Dec 5-10, 2013; Lake Tahoe, Nevada, United States p. 1196-1204.
112. Hosseini S, Gottumukkala R, Katragadda S, et al. A multimodal sensor dataset for continuous stress detection of nurses in a hospital. Sci Data 2022 Jun 1;9(1):255. [doi: 10.1038/s41597-022-01361-y] [Medline: 35650267]
113. Kaczor EE, Carreiro S, Stapp J, Chapman B, Indic P. Objective measurement of physician stress in the emergency department using a wearable sensor. Proc Annu Hawaii Int Conf Syst Sci 2020;2020:3729-3738. [Medline: 32015695]

Abbreviations

ANS: autonomic nervous system

XSL•FO
RenderX

AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic
AVNN: average normal-to-normal
BL: baseline
BPM: beats per minute
ECG: electrocardiogram
HF: high frequency
HR: heart rate
HRA: heart rate asymmetry
HRV: heart rate variability
Hz: hertz
IALS: inverse average length of segments
LPA: low physical activity
LR: logistic regression
ML: machine learning
MMA: modified moving average
MPA: moderate physical activity
MS: mental stress
NN: normal-to-normal
PASAT: paced auditory serial addition task
PPG: photoplethysmography
PSS: percentage of short segments
RAVEN: Raven's Progressive Matrices
RF: random forest
RMSSD: root-mean-square of successive differences
SDNN: SD of the NN intervals
SHAP: Shapley additive explanation
SMOTE: synthetic minority oversampling technique
SQI: signal quality index
SSST: sing-a-song-stress test
TA: tone avoidance
TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Artificial Intelligence
TWA: T-wave alternans
UHF: ultra-high frequency
VHF: very high frequency
XGBoost: extreme gradient boosting

Edited by A Coristine; submitted 10.Jul.2025; peer-reviewed by J Li, SM Shaffi; revised version received 22.Jan.2026; accepted 23.Jan.2026; published 07.Apr.2026.

Please cite as:
Uendes B, Antonides A, van de Ven S, van der Mee DJ, de Geus E, Hoogendoorn M
Electrocardiogram-Based Mental Stress Detection Amid Everyday Activities Using Machine Learning: Model Development and Validation Study
J Med Internet Res 2026;28:e80450
URL: https://www.jmir.org/2026/1/e80450
doi:10.2196/80450

© Buelent Uendes, Alex Antonides, Sjors van de Ven, Denise Johanna van der Mee, Eco de Geus, Mark Hoogendoorn. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Examining the Association Between Internet Use and Perceived Stress in Adults: Longitudinal Observational Study Combining Web Tracking Data With Questionnaires

Mohammad Belal, MCA; Nguyen Luong, MSc; Talayeh Aledavood, PhD; Juhi Kulshrestha, DRING

Department of Computer Science, Aalto University, PL 1400, Konemiehentie 2, Aalto, Finland

Corresponding Author:
Mohammad Belal, MCA
Department of Computer Science, Aalto University, PL 1400, Konemiehentie 2, Aalto, Finland

Abstract

Background: In today's digital era, the internet plays a pervasive role in daily life, influencing everyday activities such as communication, work, and leisure. This online engagement intertwines with offline experiences, shaping individuals' overall well-being. Despite its significance, existing research often falls short in capturing the relationship between internet use and well-being, relying primarily on isolated studies and self-reported data. One major contributor to deteriorated well-being is stress. While some research has examined the relationship between internet use and stress, both positive and negative associations have been reported.

Objective: This study aimed to identify the associations between an individual's internet use and their stress.

Methods: We conducted a 7-month longitudinal study. We combined fine-grained URL-level web browsing traces of 1490 German internet users with their sociodemographics and monthly measures of stress. Further, we developed a conceptual framework that allows us to simultaneously explore different contextual dimensions, including how, where, when, and by whom the internet is used. We applied linear mixed-effects models to examine these associations.

Results: Our analysis revealed several associations between internet use and stress, varying by context. Increased time spent on social media, online shopping, and gaming platforms was associated with higher stress. For example, the time spent by individuals on shopping-related internet use (aggregated over the 30 days before their stress was measured via questionnaires) was positively associated with stress on both mobile (β=.04, 95% CI 0.00 - 0.08; *P*=.04) and desktop devices (β=.03, 95% CI −0.00 to 0.06; *P*=.09). In contrast, time spent on productivity or news websites was associated with lower stress. Specifically, in the last 30 days of mobile usage, productivity-related use showed a negative association with stress (β=−.03, 95% CI −0.06 to −0.00; *P*=.04). In addition, in the last 2 days of data, news usage was negatively associated with stress on both mobile (β=−.54, 95% CI −1.08 to 0.00; *P*=.048) and desktop devices (β=−.50, 95% CI−0.90 to −0.11; *P*=.01). Further analysis showed that total time spent online (β=.01, 95% CI 0.00 - 0.02; *P*<.001), social-media usage (β=.02, 95% CI 0.00 - 0.03; *P*=.02), and gaming usage (β=.01, 95% CI 0.00 - 0.02; *P*=.02) were all positively associated with stress in high-stress Perceived Stress Scale (PSS>26) individuals on mobile devices.

Conclusions: The findings indicate that internet use is associated with stress, and these associations differ across various usage contexts. In the future, the behavioral markers we identified can pave the way for designing individualized tools for people to self-monitor and self-moderate their online behaviors to enhance their well-being, reducing the burden on already overburdened mental health services.

(J Med Internet Res 2026;28:e78775) doi:10.2196/78775

KEYWORDS

online behavior; stress; internet use; web browsing traces; sociodemographic differences; longitudinal design

Introduction

Stress is an unavoidable part of human life, arising from the demands and challenges we face daily. It is a significant factor in health issues such as cardiovascular disease, weakened immune function, and mental health challenges [1,2]. The internet, now an integral part of modern life, has sparked debates about its impact on stress levels and psychological well-being [3,4], as well as whether this influence is predominantly positive or negative. As our online and offline lives become increasingly interconnected, understanding the relationship between internet use and stress has gained considerable attention.

While the internet offers numerous advantages, such as enhanced connectivity and easy access to information, excessive or problematic use has been linked to various stress-related factors [3,5,6]. For example, heavy internet use has been associated

XSL•FO
RenderX

with higher levels of anxiety [5], while the amount of time spent online has been linked to sleep loss and withdrawal [6]. On the other hand, past research suggests that not all forms of internet use are detrimental; certain online activities have been associated with reduced stress and improved psychological well-being [7-9].

Despite the internet's widespread influence, research on psychological well-being, including stress, has primarily focused on offline activities, leaving a critical gap in understanding how online behaviors impact stress and well-being [10]. For instance, a comprehensive review of 99 commonly used psychological well-being scales identified 196 dimensions, yet none explicitly addressed online activities or behaviors [10]. Moreover, studies on online engagement have often faced limitations, including short study durations, small sample sizes, and an over-reliance on questionnaires to capture internet use patterns. These approaches can introduce biases and fail to provide a complete picture of the connection between online and offline experiences [11,12].

In this paper, we first review previous research on the association between internet use and stress and examine the methodologies used in these studies. We then outline our longitudinal multimodal study design, which integrates actual internet usage data with monthly questionnaires to measure stress, and discuss our study's potential impact.

Conflicting Findings on Associations Between Stress and Internet Use

The relationship between internet use and stress is complex, with previous research showing contrasting associations depending on the type and context of digital engagement, as well as individual characteristics. High levels of internet and smartphone use have been linked to increased stress [13,14], often due to digital overload (ie, the cognitive strain caused by constant notifications and an endless stream of information) [15,16]. In contrast, internet use through computers has been associated with less burnout compared to smartphone use [17]. However, these associations are not consistent. For instance, age influences the impact of digital multitasking: younger users report higher stress than older adults when handling multiple digital tasks, yet they appear less affected by communication overload [15]. Experimental evidence shows that multitasking increases perceived stress levels in both younger and older adults, with no significant differences between the age groups [18]. Other studies have shown no association [19] or even a negative association between time spent online and stress, particularly in young adults [20].

Moreover, the type of online activity plays a crucial role in stress outcomes. Social networking and entertainment-related use have been associated with higher stress levels, while internet use for work-related tasks has been linked to greater life satisfaction in the middle-aged population [21]. Research also indicates that communication overload from emails and messages is positively associated with perceived stress in the age group of 50 - 85 years [15]. Studies on social media show similarly nuanced findings. While Pew Research found no association between social media use and stress in men, a negative association was observed in women [22]. A large-scale study also showed slightly higher perceived stress among high social media users than nonusers [23]. Other digital behaviors, such as problematic news consumption [24] and adult content addiction [25], have also been linked to heightened stress and emotional distress. Similarly, concerns such as cyberbullying, online harassment [26], work-life boundary erosion [27], and data privacy issues [28] have been widely documented as stress-inducing. Further studies have found positive associations between stress and various digital behaviors, such as online shopping addiction in young people [29], negative information seeking [30], interpersonal communication in older adults [7], misinformation sharing [31], and excessive gaming in adolescents [32,33].

Conversely, the internet can also act as a buffer against stress [20], offering access to supportive communities [9], relaxation tools, and leisure activities [7]. Online entertainment and social interaction, in particular, have been shown to reduce stress and enhance well-being among older adults [7,8,34]. In addition, internet use has been recognized as a coping mechanism. Several online activities, including social media [35], entertainment [36,37], shopping [38], and gaming [35,39], have been identified in previous studies as strategies for managing stress.

User Characteristics Shape the Relationship Between Online Activity and Stress

Studies show that age, gender, income, and baseline stress levels influence how online activities relate to stress [15,18,22]. Social media use has been negatively associated with stress among females [22], though females overall report higher stress than males [40,41]. Older adults tend to report lower stress than younger groups but experience stronger associations between communication load, frequent messaging, and stress [42,43]. Higher income is generally linked to lower stress levels [44], though findings differ by context—for instance, higher income was associated with fewer mental health issues in Germany but with more issues in China [45]. Social media use has also been linked to slower recovery from real-world stressors, suggesting possible stress maintenance in already stressed individuals [46].

Methods for Identifying Connections Between Internet Use and Stress

Past research on internet use and stress has used various methodologies. Many studies rely on cross-sectional designs and self-reported surveys [7,15,16,21,24,32]. These studies often focus on specific populations, such as university or medical students [13,14]. Some have used larger samples [23,24,32] but still rely on questionnaires to capture internet use. A smaller number of experimental studies are available [30,46], and some have adopted alternative approaches, such as analyzing social media data to infer psychological states [31]. Studies using actual web browsing data [20,47] are limited and tend to capture general metrics, such as total time spent online [20], and are based on relatively small samples (92 and 107 participants), indicating how difficult it is to conduct such studies.

Contributions and Impact of Our Study

Previous research on the relationship between internet use and stress has shown mixed findings, revealing both negative and

XSL•FO
RenderX

positive associations depending on the type of internet activity. However, much of this evidence remains fragmented, as previous studies have largely relied on self-reported internet use data (which often lacks granularity) and have focused on limited aspects of internet use.

To address these limitations and to provide a more comprehensive understanding of how internet use relates to stress, we conducted a longitudinal multimodal study involving 1490 internet users in Germany over 7 months. Our study integrates fine-grained, passively collected web trace data from both desktop and mobile devices with participants' monthly responses to a validated stress scale (refer to Figure 1). Using objective behavioral data, we move beyond subjective self-reports and introduce a data-driven framework for revealing long-term usage patterns and identifying digital markers of stress.

Figure 1. Overview of the study design and contextual dimensions. The top panel shows our longitudinal study design combining desktop and mobile web - trace data with monthly stress questionnaires. The middle panel depicts the internet use and well-being features extracted from web browsing traces and monthly questionnaires. The bottom panel shows the contextual dimensions we consider for examining associations between internet use and stress.

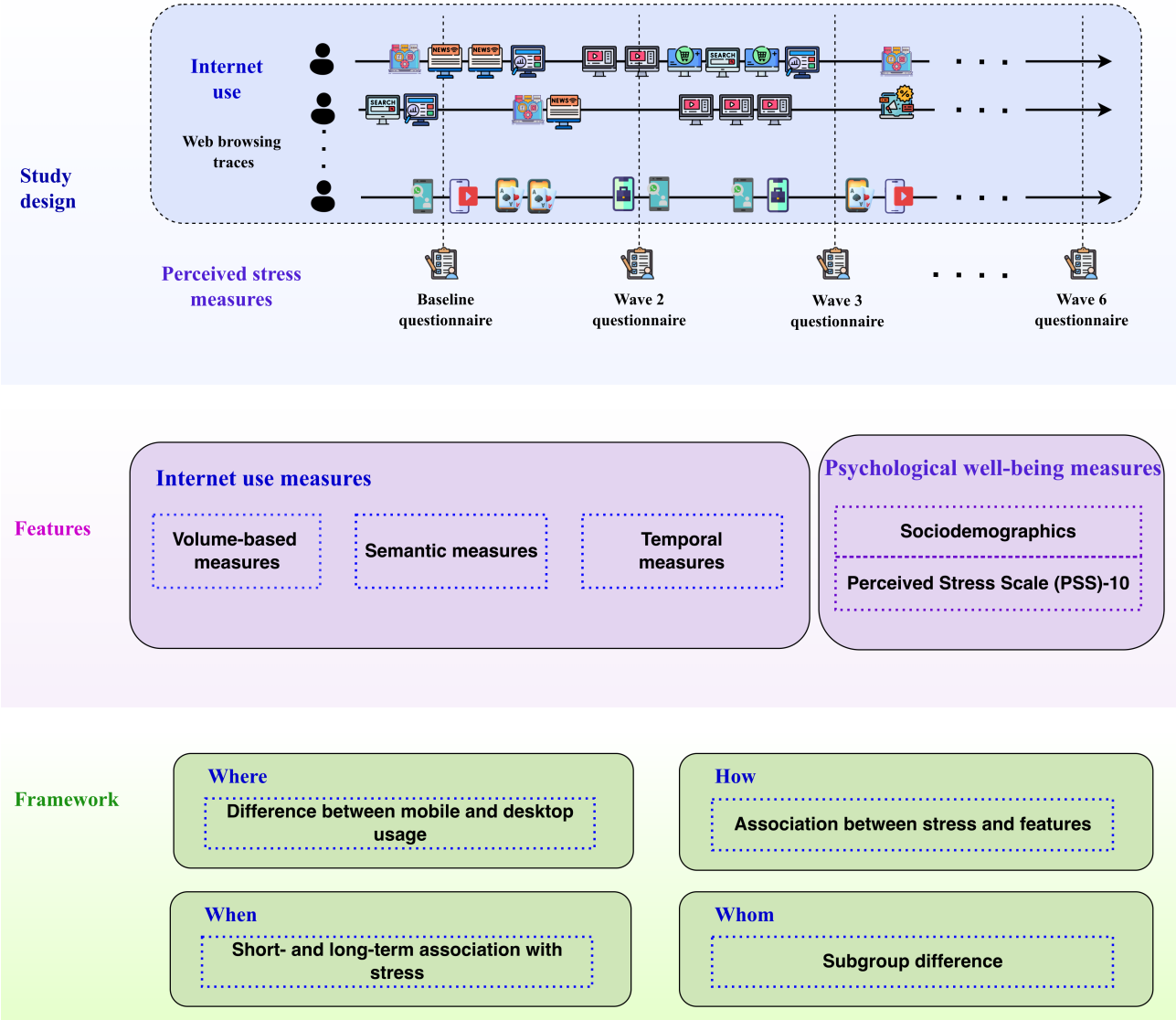

Building on existing research, we further identify 4 key dimensions that shape the relationship between digital behavior and stress:

1. How: the type and pattern of internet behaviors—such as usage volume [14,48], temporal rhythms [49], and content categories (eg, social media, productivity, entertainment, or shopping)—may influence stress in different ways [21,50-52].
2. Where: the device context plays a significant role. Previous research suggests that desktop use is often more goal-oriented and structured, whereas mobile use tends to be more fragmented and reactive [53].
3. When: the timing of online behaviors in relation to stress assessments is important, as short-term engagement with digital content may have immediate effects on stress responses [54,55].
4. By whom: individual differences—such as age, gender, income, and baseline stress levels—can moderate the impact of digital engagement on stress, with some groups being more susceptible than others [40-44].

XSL•FO
RenderX

By adopting this multidimensional perspective, our study seeks to bring coherence to the scattered evidence in existing research and provide a more thorough understanding of the link between digital behaviors and stress. In addition, identifying behavioral markers of stress in internet use may inform the design of future tools for real-time stress monitoring, complementing traditional self-reported measures. This approach contributes to a deeper understanding of digital well-being and supports the development of targeted interventions for healthier online behaviors.

Methods

In this section, we describe our study design, participants, collected data, extracted features, and analysis models that allowed us to overcome the challenges of previous work described in the "Contributions and Impact of Our Study" section.

Study Design

We conducted a longitudinal multimodal study over 7 months that combined passively collected fine-grained web browsing traces with repeated monthly online questionnaires (as provided in Figure 1). The web browsing traces for desktop users included URL-level traces, while for mobile users, both URL-level and application-level traces were included (throughout the paper, we will use "app" to mean mobile app). We measured the perceived stress of our panelists using the validated Perceived Stress Scale (PSS)-10 in 6 monthly waves. In the first wave, we also collected the sociodemographic characteristics of the panelists, including age, gender, and income.

Participants

The study was conducted on a sample of German internet users, recruited through a General Data Protection Regulation–compliant panel company (Bilendi GmbH), which provided access to participants who had already installed tracking software on their devices to capture their internet use. The company also managed survey coding and distribution, sending email notifications to relevant participants each time a survey was launched. Participation was voluntary; panelists were informed about the study and chose to take part. Compensation ranged from €1 - €3 (US $1.10-US $3.30) US $3.30) per month, depending on the number of devices tracked, plus an additional amount based on the company's standard rate (€6/h or US $6.60/h) for each survey completion. All currency values in this study are reported in euros. The exchange rate at the time of the study was €1=US $1.10. All panelists from the company were invited via email for the first wave of questionnaires, yielding 1490 completed responses. In the subsequent 5 waves, which were approximately 1 month apart, all these 1490 respondents were invited via email to participate in each wave. Across all waves, the average range between the earliest and latest survey completion dates was 15 days, and the average completion time for the baseline survey was about 37 minutes (90% CI 32.7-42.3), while for the remaining 5 waves, it was approximately 19.6 minutes (90% CI 17.8-21.5). Table 1 reports the number of participants with completed responses for each wave that we retained for further analysis. We excluded 1 panelist who reported their gender as nonbinary and 52 panelists who reported their income as "other," because these categories had too few respondents. The sample for Wave 1 is therefore 1437. We observe that about 23% of the panelists dropped out by the sixth wave. In Table 1, we also depict the distribution of the panelists across age, gender, and income for the 6 waves of questionnaires.

Next, we examined how closely our sample's sociodemographic distributions match the German population margins for gender, age, and income (provided in Table 2). We observe that our sample's gender distribution matches closely with that of the German population (Destatis [56]). However, for both age and income, the middle ranges are overrepresented in our sample, while the extremes are underrepresented.

XSL•FO
RenderX

Table . Descriptive characteristics of participants across 6 survey waves. The table presents the number of participants, gender distribution, age groups, income categories, and mean perceived stress scores (with SDs) for each wave. Percentages are shown in parentheses for categorical variables.

Wave	1	2	3	4	5	6
Participants (n)	1437	1314	1212	1198	1205	1107
Gender, n (%)						
Male	738 (51.35)	688 (52.36)	639 (52.72)	635 (53.01)	628 (52.12)	593 (53.57)
Female	699 (48.65)	626 (47.64)	573 (47.28)	563 (46.99)	577 (47.88)	514 (46.43)
Age group (years), n (%)						
18 - 30	119 (8.28)	101 (7.67)	91 (7.51)	84 (7.01)	92 (7.63)	76 (6.67)
31 - 45	462 (32.15)	414 (31.51)	385 (31.77)	382 (31.89)	379 (31.45)	330 (29.81)
46 - 60	569 (39.60)	528 (40.18)	483 (39.85)	483 (40.32)	483 (40.08)	460 (41.55)
>60	287 (19.97)	271 (20.62)	253 (20.87)	249 (20.78)	251 (20.83)	241 (21.77)
Income (euros/month), n (%)						
<1000 (Tier I)	126 (8.77)	119 (9.06)	116 (9.57)	110 (9.18)	103 (8.55)	94 (8.49)
1000 - 2000 (Tier II)	300 (20.88)	271 (20.62)	244 (20.13)	244 (20.37)	249 (20.66)	240 (21.68)
2001 - 3000 (Tier III)	364 (25.33)	339 (25.80)	319 (26.32)	310 (25.88)	309 (25.64)	281 (25.38)
3001 - 4000 (Tier IV)	294 (20.46)	271 (20.62)	243 (20.05)	245 (20.45)	251 (20.83)	229 (20.69)
>4000 (Tier V)	353 (24.57)	314 (23.90)	290 (23.93)	289 (24.12)	293 (24.32)	263 (23.76)
Perceived Stress Score, mean (SD)	16.19 (7.19)	15.83 (7.43)	15.76 (7.56)	15.65 (7.41)	15.54 (7.50)	14.89 (7.58)

Table . Distribution of the adult population in Germany (2023) by gender, age group, and monthly income level.

Category	Adult population, n (%)
Sex	
Male	41.2 (48.8)
Female	42.3 (51.2)
Age group (years)	
18 - 30	14.2 (17)
31 - 45	19.2 (23)
46 - 60	21.7 (26)
>60	28.4 (34)
Monthly income level (euros)	
<€1250 (<US $1375)	21.1 (25.3)
€1250-€2080 (US $1375-US $2288)	13 .7 (16.4)
€2080-€2920 (US $2288-US $3212)	12.4 (14.8)
€2920-€4170 (US $3212-US $4587)	13.6 (16.3)
•	
>€4170 (>US $4587)	22.6 (27.1)

Data Collection

The panelists of the panel company had already consented to install tracking software on their desktops or mobile devices. Some participants consented to install it on both devices. Through this tracking software, the company provided fine-grained traces of visited URLs and mobile apps, including the time of visit and duration of each visit. During the 7 months, we recorded 47,100,701 URL visits from both desktop and mobile users, covering 236,955 unique web domains. For mobile apps, we captured 13,553,645 app visits across 13,476 unique apps.

XSL•FO
RenderX

Data Cleaning and Preprocessing

First, we removed the bottom 20% of panelists in each wave, ranked by total time spent browsing, since they did not have sufficient data to extract meaningful internet use patterns. Second, we identified a group of panelists who appeared to be "professional survey takers," spending more than 25% of their online time on survey domains. To focus on users with more typical internet use, we excluded these individuals from the sample. Notably, applying the above time threshold to their nonsurvey activities would have led to the removal of more than 29% of these panelists. Finally, to ensure that our internet use measures accurately capture user behavior, we only included panelists for whom we could categorize at least 80% of their web visits (refer to the "Data Enrichment" section for details). Table 3 summarizes the number of participants excluded at each step, resulting in a set of distinct panelists across waves comprising 656 mobile users and 526 desktop users. Table 4 presents the sociodemographic characteristics of the remaining users included in the analysis. In the mobile data, the proportion of users aged 31 - 45 years increased compared to the baseline questionnaire, as provided in Table 3. In the desktop data, the proportion of males and users aged 46 - 60 years increased, while the proportion of users aged 31 - 45 years decreased relative to the baseline questionnaire.

Table . Overview of panelists with matched passive web data from desktop and mobile devices across 6 survey waves. The table shows the number of users before and after data cleaning for both device types. The final row indicates the total number of distinct users retained in the cleaned dataset.

Survey wave	Number of panelists	Desktop users	Desktop users (cleaned)	Mobile users	Mobile users (cleaned)
1	1437	981	359	907	519
2	1314	848	321	806	470
3	1212	762	284	728	426
4	1198	717	257	717	418
5	1205	714	265	697	399
6	1107	656	227	649	368
Total distinct users	__[a]	—	526	—	656

[a]Not applicable.

Table . Sociodemographic characteristics of users included in the analysis after data cleaning for both mobile and desktop datasets.

Characteristic	Mobile (n=656)	Desktop (n=526)
Gender, n (%)		
Male	334 (50.91)	289 (54.94)
Female	322 (49.09)	237 (45.06)
Age group (years), n (%)		
18 - 30	53 (8.08)	38 (7.22)
31 - 45	246 (37.5)	137 (26.05)
46 - 60	247 (37.65)	232 (44.11)
>60	110 (16.77)	119 (22.62)
Income (euros/month), n (%)		
<1000 (Tier I)	60 (9.15)	59 (11.22)
1000 - 2000 (Tier II)	129 (19.66)	121 (23)
2001 - 3000 (Tier III)	166 (25.30)	128 (24.33)
3001 - 4000 (Tier IV)	137 (20.88)	87 (16.54)
>4000 (Tier V)	164 (25)	131 (24.9)

Data Enrichment

To understand "how" the panelists are using the internet, we categorized their online visits into semantic categories. The goal was to group domains, subdomains, and apps based on their primary function into categories such as "social media" (eg, facebook.com [Meta Platforms, Inc] and TikTok app [ByteDance]) and "productivity" (eg, Gmail [Google LLC] and calendar.google.com [Google LLC]). We derived the set of categories (provided in Table 5) by combining categories used by app stores and web domain classification services such as Webshrinker.com. For platform domains such as google.com, we also categorized their subdomains. For instance, google.com was categorized as "search," while mail.google.com was

XSL•FO
RenderX

classified as "productivity." Table 5 provides the complete list of semantic categories we considered, along with some examples. Two researchers from our team first independently annotated the categories for all domains and apps that constituted around 85% of web visits made by our panelists. Later, disagreements were resolved collaboratively. We observed a substantial interannotator agreement with a Cohen κ agreement [57] score of 0.7, based on annotations for a random subset of 200 domains. Following this process, we classified 3777 domains and 989 apps into semantic categories, capturing 85% of visits from mobile devices and 84% from desktops.

Table . Categorization of web domains and mobile apps based on semantic usage type. The table lists representative examples of domains and apps across various categories, grouped separately for desktop web domains and mobile apps.

Category	Example of domains, subdomains, and apps in the category
Domains	
Entertainment	youtube.com, twitch.tv, disneyplus.com, and netflix.com
Shopping	amazon.de, ebay.de, kleinanzeigen.de, and temu.com
Social media	facebook.com, twitter.com, and instagram.com
Messaging	whatsapp.com, knuddels.de, and fdating.com
Productivity	mail.google.com, outlook.live.com, navigator.web.de, and docs.google.com
Games	gameduell.de, anocris.com, forgeofempires.com, and spielaffe.de
Adult	pornhub.com, xvideos.com, xnxx.com, and romeo.com
News	bild.de, focus.de, welt.de, and wunderweib.de
Apps	
Entertainment	YouTube (Google LLC), Netflix (Netflix, Inc), and Spotify Music (Spotify Technology)
Shopping	Amazon Shopping (Amazon.com, Inc), eBay (eBay, Inc), Vinted.fr (Vinted Group), and Lidl Plus (Schwarz Group)
Social media	Facebook (Meta Platforms, Inc), Instagram (Meta Platforms, Inc), and Twitter (X Corp), TikTok – Make Your Day (ByteDance)
Messaging	WhatsApp (Meta Platforms, Inc), Facebook Messenger (Meta Platforms, Inc), and Telegram (Telegram FZ-LLC)
Productivity	Gmail (Google LLC), GMX Mail (Global Mail eXchange), WEB.DE Mail (United Internet Group), and Google Calendar (Google LLC)
Games	Candy Crush Saga (King), Coin Master (Moon Active), Royal Match (Dream Games), and Pokémon GO (Niantic)
News	n-tv Nachrichten (RTL Group), kicker online (Olympia -Verlag GmbH), AOL – News (AOL Media LLC), and BILD: Immer aktuell informiert (Axel Springer SE)

Measures

As described in the "Study Design" section, we combined repeated monthly stress questionnaires with web browsing data. Perceived stress was measured through PSS-10 questionnaire responses, and internet use features were derived from passively collected web traces. For each panelist in each survey wave, we calculated internet usage features based on their activity during the period "T" preceding the stress measurement (ie, the time of questionnaire response). To address the question of "when" the internet is used, we extracted features for either 30 or 2 days to examine both long-term and short-term effects. The resulting measures were then used to examine the associations between internet use and stress.

Measures From Web Traces

To measure "how" individuals use the internet, we created features that span from coarse- to fine-grained measures, as provided in Table 6. We captured overall web activity at the coarse-grained level, such as total time spent online. We also accounted for the time of the day when panelists were browsing the web by including the difference between the time spent online during daytime (6 AM-6 PM) and nighttime (6 PM-6 AM) hours. At a finer granularity, we analyzed how panelists distributed their time across online activities such as social media, entertainment, and news. For each survey wave, if a participant completed the survey on a given date (eg, July 31), we summed their time spent on each activity over the period (T=30 or 2 days; eg, July 1 - 30 or July 29 - 30) preceding the day of completing the survey. For instance, time spent on news represents the aggregated time on news domains (desktop) and news domains and apps (mobile) during that period. Each participant had up to 6 time points, 1 per wave.

XSL•FO
RenderX

Table . Features and their descriptions. Time spent online is measured in hours in the period T (30 days or 2 days) before the measurement of stress. These features are computed for online activity on each device (desktop or mobile) separately.

Features	Description
Coarse-grained	
• Total time spent online	Total time spent online in period T
• Daytime nighttime difference	Difference of time spent online during daytime (6 AM-6 PM) and nighttime hours (6 PM-6 AM)
Fine-grained	
• Time spent on entertainment • Time spent on social media • Time spent on messaging • Time spent on news • Time spent on adult content • Time spent on games • Time spent on shopping • Time spent on productivity	Time spent on different semantic classes of online activities. For instance, time spent on entertainment domains or apps such as YouTube.com or Amazon Prime is classified as entertainment use.
Control variables	
• Gender • Age • Income • Survey wave	Sociodemographic characteristics of individuals and seasonality

Measures From Questionnaires

We used the PSS-10 [58] in our monthly questionnaires to measure the stress levels of our panelists. The PSS-10 is a widely used, validated scale designed to assess how stressed individuals feel. It captures aspects such as the unpredictability of life, perceived control over situations, and general stress levels over the past month. Participants rate their responses on a scale from 0 (never) to 4 (very often), producing a total score between 0 and 40 across the 10 items. Higher scores indicate greater perceived stress, with scores typically grouped into 3 levels: 0 - 13 (low stress), 14 - 26 (moderate stress), and 27 - 40 (high stress) [59,60]. In addition, we collected each participant's self-reported sociodemographics, including age, gender, and income, in the first wave of the questionnaires.

Statistical Analysis

We used linear mixed-effects models (LMMs) [61] to examine the relationship between internet use and stress. We chose LMMs for analyzing data from our longitudinal study since they account for repeated measurements of individuals and incorporate fixed and random effects. Fixed effects included internet use features provided in Table 6. Random intercepts were added to account for individual-specific differences in baseline stress levels across participants.

We formally describe the models as follows. For an individual i at questionnaire wave j∈ {1,2...6}, we denote Yij as the variable of interest, xij as the covariate, and the intercept for the random effect as uj. Therefore, we consider the following LMM:

$$Yij=\beta 0+\beta 1xij1+\beta 2xij2+\ldots+\beta nxijn+uj+ ij$$

where:

- Yij is the perceived stress level of the individual i measured at the questionnaire wave j.
- β0 is the fixed intercept.
- β1,…,βn are the fixed effect coefficients for each covariate xijn.
- xij1=(total time spent online, xij2=daytime-nighttime difference, xij3=time spent on entertainment ... xijn=survey wave), where xijn corresponds to each feature provided in Table 6.
- uj is the random effect for individual i, capturing individual-level variability.
- ij is the residual error term for the individual i at wave j.

We conducted model diagnostics to validate the assumptions of LMMs, including checks for multicollinearity (Variance Inflation Factor (VIF) <2.0). All statistical analyses were implemented using Python's statsmodels package (version 0.14.1; Python Software Foundation) [62].

To understand whether the granularity of the extracted features affects model performance and the associations identified, we developed 2 models. The first model (Model 1) focused on coarse-grained measures of internet use such as total time spent online and daytime-nighttime difference. The second model (Model 2) extended the first model and also incorporated finer-grained measures of internet use across semantic classes. For Model 2, we dropped the total time spent online feature to avoid multicollinearity.

Previous work has shown that both sociodemographics [40-44] and seasonal variations [63,64] can significantly influence individuals' stress levels. Accordingly, we included the sociodemographics and seasonality measures as control variables for both models, as provided in Table 6.

XSL•FO
RenderX

Ethical Considerations

Our study was approved by Aalto University's Research Ethics Committee (approval ID D/894/03.04/2023). Data collection was conducted via a General Data Protection Regulation–compliant European company, and informed consent was obtained from participants for both the surveys and web-trace datasets, with the option to withdraw consent at any time during or after the study. To protect participants' privacy, we implemented strict data privacy measures. The web dataset was anonymized by the panel company by removing personal information such as email addresses and usernames to prevent participant identification. In addition, the dataset was stored and analyzed solely on the university's secure server, with access restricted to the research team. We will make the anonymized data and code available to support the open-source community and to spur further research at the intersection of internet use and well-being.

Results

Overview

Our study examined various internet use behaviors associated with stress, and in this section, we present our results across 4 key contextual dimensions (as outlined in the "Contributions and Impact of Our Study" section). We first analyzed "how" internet-based features relate to stress. We then explored the remaining dimensions: device-based differences (where) by comparing desktop and mobile usage, temporal patterns (when) using internet activity from the 2 days before the survey, and individual differences (by whom) through subgroup analyses based on age, gender, income, and baseline stress levels.

Behavioral Patterns (How)

To understand how internet usage is associated with stress, we ran LMMs on 2 sets of features, progressing from coarse-grained (amount and timing of usage) to fine-grained (also including semantic category usage) measures. An ANOVA test was conducted to determine whether the more complex model explained significantly more variance than the simpler model. The results showed no statistically significant improvement when using the more complex model, although the more complex model provided important information on the nuanced relationship between internet use and stress.

Analysis of 30-day mobile data (number of panelists, N=656) and observations, n=2600), as provided in Table 7, revealed that Model 2—which includes both timing and semantic category usage—identified significant associations with stress. Specifically, shopping-related usage was positively associated with stress (β=.04,, 95% CI 0.00 - 0.08; *P*=.04), while productivity usage showed a negative association (β=–.03, 95% CI −0.06 to −0.00; *P*=.04). In contrast, Model 1, which included only total usage and timing, did not show any significant associations.

XSL•FO
RenderX

Table . Results from linear mixed-effects models for all participants, based on 30-day mobile data. Model 1 includes coarse-grained features, while Model 2 incorporates fine-grained usage categories (described in the "Statistical Analysis" section). Estimates, CIs, and *P* values are reported for each predictor. Statistically significant *P* values are in bold.

Predictors	Model 1[a], estimate (95% CI)	*P* value	Model 2[b], estimate (95% CI)	*P* value
Intercept	20.77[c] (19.09 - 22.44)	<.001	20.69[c] (19.01 - 22.37)	<.001
Survey wave	−0.10[d] (−0.19 to −0.02)	.01	−0.11[d] (−0.19 to −0.02)	.01
Gender	1.79[c] (0.80 - 2.77)	<.001	1.68[c] (0.68 - 2.68)	<.001
Age	−1.62[c] (−2.20 to −1.04)	<.001	−1.57[c] (−2.16 to −0.99)	<.001
Income	−1.11[c] (−1.50 to −0.73)	<.001	−1.08[c] (−1.47 to −0.69)	<.001
Total time spent online	0.00 (−0.00 to 0.01)	.39	__[e]	—
Daytime nighttime difference	−0.01 (−0.02 to −0.00)	.12	−0.01 (−0.02 to 0.00)	.18
Time spent on entertainment	—	—	0.01 (−0.01 to 0.02)	.37
Time spent on social media	—	—	0.00 (−0.01 to 0.02)	.71
Time spent on messaging	—	—	0.00 (−0.02 to 0.02)	.70
Time spent on games	—	—	0.00 (−0.00 to 0.01)	.30
Time spent on shopping	—	—	0.04[d] (0.00 - 0.08)	.04
Time spent on productivity	—	—	−0.03[d] (−0.06 to −0.00)	.04
Time spent on news	—	—	−0.03 (−0.09 to 0.03)	.33

[a]σ^2=11.67; τ =36.54$_{pid}$; ICC=0.76; N=656$_{pid}$; Observations=2600.

[b]σ^2=11.65; τ =36.48$_{pid}$; ICC=0.76; N=656$_{pid}$; Observations=2600.

[c]*P*<.001.

[d]*P*<.05.

[e]Not applicable.

In addition, sociodemographic factors such as age, gender, and income consistently predicted stress across both models. Age (β=–1.57, 95% CI −2.16 to −0.99; *P*<.001) and income (β=–1.08, 95% CI −1.47 to −0.69; *P*<.001) were negatively associated with stress, while women reported higher stress levels (β=1.68, 95% CI 0.68 - 2.68; *P*=.001).

Device Matters (Where)

To observe device differences, we analyzed 30-day desktop data. For desktop data (N=526 and n=1713), the results are shown in Table 8. Model 2, which incorporates semantic and temporal features, showed a weaker positive association between shopping usage and stress (β=.03, 95% CI −0.0 to 0.06; *P*=.09). As observed with mobile data, the simpler Model 1 did not reveal any significant associations with internet usage features. Similarly, the sociodemographic results were consistent with those observed in the mobile data.

XSL•FO
RenderX

Table . Results from linear mixed-effects models for all participants, based on 30-day desktop data. Model 1 includes coarse-grained features, while Model 2 incorporates fine-grained usage categories (described in the "Statistical Analysis" section). Estimates, CIs, and *P* values are reported for each predictor. Statistically significant *P* values are in bold. Random effects, intraclass correlation coefficient (ICC), number of participants (N), and total observations are also provided.

Predictors	Model 1[a], estimate (95% CI)	*P* value	Model 2[b], estimate (95% CI)	*P* value
Intercept	20.62[c] (18.70 - 22.55)	<.001	20.49[c] (18.58 - 22.41)	<.001
Survey wave	−0.14[d] (−0.24 to −0.03)	.009	−0.15[d] (−0.26 to −0.04)	.006
Gender	1.83[d] (0.66 - 3.00)	.002	1.73[d] (0.55 - 2.90)	.004
Age	−1.72[c] (−2.41 to −1.04)	<.001	−1.74[c] (−2.43 to −1.05)	<.001
Income	−0.94[c] (−1.38 to −0.51)	<.001	−0.92[c] (−1.36 to −0.48)	<.001
Total time spent online	−0.00 (−0.01 to 0.00)	.40	__[e]	—
Daytime nighttime difference	0.00 (−0.01 to 0.01)	.46	0.00 (−0.01 to 0.01)	.46
Time spent on entertainment	—	—	−0.00 (−0.01 to 0.01)	.64
Time spent on adult content	—	—	−0.01 (−0.03 to 0.00)	.11
Time spent on social media	—	—	0.00 (−0.02 to 0.02)	.94
Time spent on messaging	—	—	0.01 (−0.04 to 0.06)	.61
Time spent on games	—	—	0.00 (−0.03 to 0.03)	.84
Time spent on shopping	—	—	0.03 (−0.00 to 0.06)	.09
Time spent on productivity	—	—	−0.00 (−0.03 to 0.02)	.86
Time spent on news	—	—	−0.02 (−0.06 to 0.02)	.28

[a]σ^2=11.10; τ =40.62$_{pid}$; ICC=0.79; N=526$_{pid}$; Observations=1713.

[b]σ^2=11.09; τ =40.77$_{pid}$; ICC=0.79; N=526$_{pid}$; Observations=1713.

[c]P<.001

[d]*P*<.01.

[e]Not available.

Time Period of Data (When)

To investigate the relationship between short-term versus long-term internet usage patterns and stress, we analyzed associations between various online activities performed on mobile and desktop devices in 2 time periods—30 days and 2 days—and individual stress levels. We used the same features and models as in our previous 30-day data analyses. Here, we specifically focused on web activity recorded during the 2 days immediately preceding the PSS-10 survey.

For both mobile and desktop data (as provided in Tables S1 and S2 in Multimedia Appendix 1), news usage showed a negative association with stress (β=–.54, 95% CI −1.08 to 0.00; *P*=.048) and (β=–.50, 95% CI −0.90 to −0.11; *P*=.01), respectively, in Model 2. In addition, in desktop data, messaging usage demonstrated a weak negative association with stress (β=–.59, 95% CI −1.24 to 0.06; *P*=.07) in model 2.

Individual Differences (By Whom)

To explore how internet usage varies by individual characteristics, we conducted subgroup analyses, running models separately for categories such as gender (male and female) to understand how associations differ based on these characteristics. In the following subsections, we first examined the relationship between internet use and stress based on baseline stress levels by analyzing high-stress and low-stress groups. We then explored differences by gender, age, and income categories.

Stress Levels

We identified 2 groups of panelists from our data—high-stress and low-stress—based on their reported PSS-10 scores in the online questionnaires. Panelists who scored more than 26 in any wave they participated in were included in the high-stress group, and panelists who scored below 14 in any wave were included in the low-stress group.

For the high-stress population (PSS-10 score > 26), several notable results were observed (as provided in Tables S3-S6 in Multimedia Appendix 1). In the 30-day mobile data, time spent (β=.01, 95% CI 0.0 - 0.02; *P*<.001) in Model 1, and social media usage (β=.02, 95% CI 0.0 - 0.03; *P*=.02), and gaming usage (β=.01, 95% CI 0.0 - 0.02; *P*=.02) in Model 2 were positively associated with stress. In the 2 days of data, the daytime-nighttime difference showed a weak positive association (β=.11, 95% CI −0.01 to 0.23, *P*=.08) in Model 1. For desktop data, no significant variables, including sociodemographics, were found to be associated with stress in the high-stress subgroup.

XSL•FO
RenderX

In the low-stress population (PSS-10 score <14 in the baseline survey), as provided in Tables S7-S10 in Multimedia Appendix 1, adult-content usage was negatively associated with stress in 30-day desktop data (β=–.02, 95% CI −0.04 to 0.0; *P*=.07). Similarly, for the low-stress group, in the 2 days data, time spent (β=–.07, 95% CI −0.14 to 0.0; *P*=.04) was significant for desktop, while gaming usage was weakly significant for mobile data (β=.08, 95% CI −0.01 to 0.17; *P*=.1).

When analyzing the 30-day data for all participants, sociodemographic factors were strongly associated with stress in both mobile and desktop settings. However, within the high-stress group, income was the only sociodemographic variable that remained significant in mobile data, showing a negative association with stress in both the general population (β=–1.08, 95% CI −1.47 to −0.69; *P*<.001) and the high-stress subgroup (β=–.52, 95% CI −0.91 to −0.12; *P*=.01). In contrast, gender and age—which were significant predictors in the overall population—did not show statistical significance in the highly stressed subgroup for either mobile or desktop data.

Gender Differences

Subgroup analysis by gender revealed distinct patterns in feature significance for both desktop and mobile data. In the 30-day mobile data (as provided in Table S11 in Multimedia Appendix 1), shopping usage (β=.07, 95% CI 0.01 - 0.14; *P*=.02) and productivity features (β=–.05, 95% CI −0.09 to −0.01; *P*=.02) were significant only for male users (N=334 panelists and n=1334 observations) in Model 2. In contrast, these features were not significant for female users (N=322 panelists and n=1266 observations) as provided in Table S12 in Multimedia Appendix 1. No features were significant for males and females in the 30-day desktop data.

In the 2-day data (as provided in Tables S15-S16 in Multimedia Appendix 1), news consumption was negatively associated with stress for males in both desktop (β=–.52, 95% CI −1.02 to −0.01; *P*=.04) and mobile (β=–.58, 95% CI −1.23 to 0.07; *P*=.08) data. For females, in mobile data, daytime-nighttime difference (β=.10, 95% CI −0.0 to 0.21; *P*=.06) and messaging (β=.23, CI −0.02 to 0.48; *P*=.08) were weakly positively associated.

Age Differences

Subgroup analysis by age revealed distinct patterns in web-based associations. For the 30-day mobile data (as provided in Tables S19-S22 in Multimedia Appendix 1), shopping was positively associated with stress in age groups of 18 - 30 years (β=.12, 95% CI −0.02 to 0.25; *P*=.09) and older than 60 years (β=.01, 95% CI 0.02 - 0.19; *P*=.02). In the age group of 30 - 45 years, weak positive associations were found for entertainment (β=.02, 95% CI −0.00 to 0.04; *P*=.08) and messaging (β=.03, 95% CI −0.00 to 0.06; *P*=.07), whereas productivity was negatively associated (β=–.08, 95% CI −0.14 to −0.02; *P*=.007). In the older than 60 years group, time spent (β=.01, 95% CI 0.00 - 0.03; *P*=.03) and messaging (β=.05, CI −0.00 to 0.09; *P*=.06) were positively associated, while the daytime-nighttime difference (β=–.03, 95% CI −0.05 to −0.01; *P*=.01) and shopping (β=.10, 95% CI 0.02 - 0.19; *P*=.02) were negatively associated.

In the 30-day desktop data (as provided in Tables S23-S26 in Multimedia Appendix 1), adult content usage was negatively associated in 18 - 30 years (β=–.87, 95% CI −1.77 to 0.03; *P*=.06) and older than 60 years (β=–.19, 95% CI −0.38 to 0.00; *P*=.06) age groups. In addition, in the 18 - 30 age group, the daytime-nighttime difference (β=.07, CI −0.00 to 0.14; *P*=.06) was positively associated, while news usage (β=.026, CI −0.42 to −0.10; *P*=.002) was negatively associated. Shopping was positively associated in the age group of 45 - 60 years (β=.04, 95% CI −0.00 to 0.09; *P*=.06).

For the 2-day mobile data (as provided in Tables S27-S30 in Multimedia Appendix 1), news was negatively associated in both the 30 - 45 years (β=–1.01, 95% CI −2.22 to 0.19; *P*=.10) and 45 - 60 years (β=–.85, 95% CI −1.76 to 0.06; *P*=.07) age groups. In addition, in the age group of 30 - 45 years, the daytime-nighttime difference (β=.14, 95% CI 0.02 - 0.25; *P*=.02) and entertainment usage (β=.22, 95% CI 0.01 - 0.43; *P*=.04) were positively associated, whereas shopping was negatively associated (β=–.54, 95% CI −1.08 to −0.0; *P*=.05). In the older than 60 years of age group, time spent on gaming (β=.26, 95% CI 0.01 - 0.50; *P*=.04) was positively associated.

Similarly, for the 2-day desktop data (as provided in Tables S31-S34 in Multimedia Appendix 1), messaging was negatively associated in the 45 - 60 years age group (β=–.78, 95% CI −1.48 to −0.07; *P*=.03), but positively associated for the older than 60 years age group (β=7.63, 95% CI 0.39 - 14.87; *P*=.04). In the age group of 18 - 30 years, entertainment (β=1.14, 95% CI −0.18 to 2.46; *P*=.09) was positively associated, and social media (β=1.22, 95% CI 0.15 - 2.29; *P*=.03) was positive for the 30 - 45 age group. In addition, time spent (β=–.14, 95% CI 0.28 - 0.0; *P*=.06) and news usage (β=.46, 95% CI −0.96 to 0.04; *P*=.07) were negatively associated in the older than 60 years age group.

Income Differences

For the 30-day mobile data (as provided in Tables S35-S39 in Multimedia Appendix 1), messaging (β=–.06, 95% CI −0.12 to −0.01; *P*=.03) was negatively associated with stress in participants earning less than €1000 (US $ 1100) per month (Tier I). In the €2001-€3000 (US $2201.10-US $3300) income group (Tier II), productivity (β=–.05, 95% CI −0.10 to 0.0; *P*=.06) and news usage (β=–.10, 95% CI −0.20 to 0.01; *P*=.08) were both negatively associated. Time spent (β=.01, CI 0.0 - 0.02; *P*=.02) and shopping (β=.08, 95% CI 0.01 - 0.16; *P*=.02) were positively associated with stress for participants earning €3001-€4000 (US $3301.10-US $4400; Tier IV). No other significant internet-based features were observed for other income categories.

For the 30-day desktop data (as provided in Tables S40-S44 in Multimedia Appendix 1), news usage was positively associated with stress in Tier I income group (β=.28, 95 CI 0.05 - 0.50; *P*=.02), while productivity was negatively associated (β=–.07, 95% CI −0.13 to −0.01; *P*=.03). For participants in Tier III income, news usage (β=–.13, 95% CI −0.25 to 0.0; *P*=.06) was negatively associated. For participants in Tier IV income, shopping was positively associated with stress (β=.08, 95% CI 0.02 - 0.15; *P*=.02). For Tier V participants, social media use was positively associated with stress (β=.06, 95% CI 0.01 - 0.12; *P*=.03), while news use (β=–.16, 95% CI −0.25 to −0.07; *P*<.001) and time spent were negatively associated

XSL•FO
RenderX

(β=–.01, 95% CI −0.03 to 0.0; *P*=.03). No significant associations were identified for other income categories.

For the 2-day mobile data (as provided in Tables S45-S49 in Multimedia Appendix 1), gaming (β=.21, 95% CI −0.02 to 0.43; *P*=.07) was positively associated in the Tier II income group and negatively associated (β=–.19, 95% CI −0.37 to 0.0; *P*=.047) in the Tier V income group. News use was negatively associated in the Tier III (β=–.75, 95% CI −1.58 to 0.09; *P*=.08) and Tier V (β=–.74, 95% CI −1.60 to 0.12; *P*=.09) income groups. In addition, messaging (β=, 95% CI 0.14 - 0.82; *P*=.006) was positively associated in the Tier III income group, and daytime-nighttime difference (β=.18, 95% CI 0.0 - 0.36; *P*=.04) was positively associated in the Tier IV income group.

For the 2-day desktop data (as provided in Tables S50-S54 in Multimedia Appendix 1), news usage was negatively associated in the Tier III income group (β=–1.13, 95% CI −2.33 to 0.06; *P*=.06) and the Tier V income group (β=–1.04, 95% CI −1.76 to −0.32; *P*=.005). In addition, in the Tier V income group, productivity (β=.64, 95% CI 0.07 - 1.21; P=.03) was positively associated with stress. In the Tier IV income group, time spent (β=.23, 95% CI 0.04 - 0.42; *P*=.0192) and daytime-nighttime difference (β=.23, 95% CI −0.00 to 0.46; *P*=.05) were positively associated.

Discussion

Principal Findings

Our results show that various internet usage behaviors are associated with stress, both positively and negatively. These associations differ across device type, time frame, and sociodemographic groups. Figure 2 summarizes these patterns and their variation across panelist subgroups. In the following sections, we discuss the internet use features that have shown positive (social media, entertainment, shopping, games, and time of internet use), negative (adult content, productivity, and news), and mixed (messaging and screen time) associations with stress.

Figure 2. Overview of significant internet usage behaviors across various contextual dimensions. The icons represent different internet use features. The rows correspond to a combination of the time frame (30 days or 2 days) and device type (desktop or mobile), and the columns correspond to factors pertaining to individual differences (stress levels, gender, age, and income). Red icons denote positive associations, blue icons indicate negative associations, and empty cells show no significant internet features. The intensity of the color reflects the strength of the significance, with lighter icons denoting weak significance (*P*=.05 to *P*<.10) and darker icons representing high significance (*P*<.05).

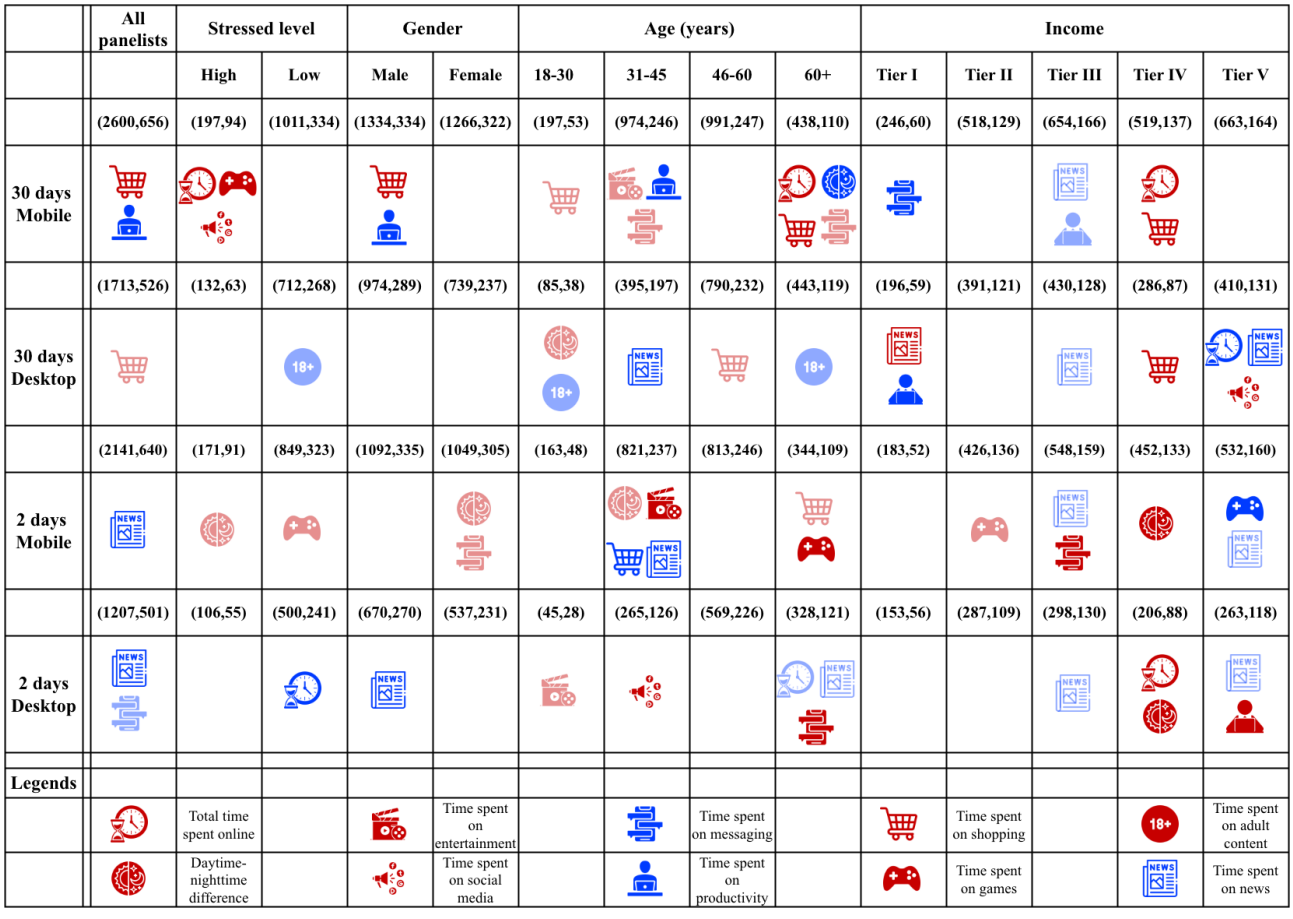

Positive Associations

Social Media

People spend a substantial portion of their time online on social media platforms [65]. In our dataset, social media accounted for approximately 23% of total usage in mobile data and 15% in desktop data. Therefore, understanding social media's relationship with stress is increasingly important as it continues to occupy a large share of individuals' internet activity.

XSL•FO
RenderX

Previous research shows that social media can contribute to stress, act as a resource, or function as a coping tool [50]. Factors such as fear of missing out [66], appearance-related pressure [67], and communication overload [68] have been linked to increased distress. At the same time, other studies have highlighted its potential to buffer stress and offer social support in specific contexts [22,35,69].

In our results, when significant, social media was consistently positively associated with stress across subgroups and device types. This association was significant for the high-stress subgroup in the 30-day mobile data, the Tier I income group in the 30-day desktop data, and the 30 - 45 age group in the 2-day desktop data (Figure 2). Although we cannot definitively determine whether social media contributes to or alleviates stress, these patterns suggest that it may be used as a coping strategy for these groups.

Previous work has identified social media as a space for various coping mechanisms [35]. Our findings in the middle-aged group potentially reflect this pattern, aligning with earlier research [70], possibly due to the support accessed through these platforms. This is consistent with studies linking social media use, particularly Facebook, to support-seeking behavior [71].

Overall, our results highlight the role of social media in shaping stress experiences. While our study, combining fine-grained web data with a longitudinal design and contextual framework, strengthens this interpretation, further research is needed to disentangle the causal directions of the observed associations and social media's role—whether as stressor, resource, or coping tool.

Entertainment

According to a recent survey in Germany, people spend an average of 203 minutes per day watching content online [65]. In our data, entertainment usage accounted for 9% of total usage on mobile devices and 14% on desktop devices.

Previous studies have reported a positive correlation between entertainment usage and stress [21,72-74]. Further, activities such as watching content or listening to music have also been identified as common coping strategies for managing stress [36,37,75]. In our findings, entertainment usage was positively associated with stress in both desktop and mobile data. For mobile users, this association was consistent across both the 30-day and 2-day periods for the 30 - 45 age group. In the 2-day desktop data, a weak positive association was observed for the age group of 18 - 30 years.

These results suggest that entertainment usage (similar to social media) may serve as a coping mechanism, especially among younger users in our data, compared with older individuals. Previous research indicates that younger individuals are more likely to engage in binge-watching as a way to regulate emotions [36]. However, there is a lack of detailed research on this association [74], and future studies should further explore the types of content consumed and their relationship to mental health.

Online Shopping

Online shopping has grown substantially in recent years, particularly with the rise of smartphones, which now account for 80% of all retail visits [76]. Previous studies have linked compulsive buying-shopping disorder to higher levels of stress [29,38,52,77-79]. At the same time, shopping has also been identified as a way to relieve stress [80,81].

In our results, shopping was predominantly positively associated with stress across mobile and desktop data, in various time frames and subgroups (refer to Figure 2), aligning with previous findings. A negative association was observed only for the age group of 30 - 45 years in the 2-day mobile data, suggesting that, for this group, shopping may serve as a short-term stress reliever, or that individuals under stress may avoid shopping. The latter behavior aligns with previous research showing that, in middle-aged adults, stress can lead to increased saving behavior [82].

This association was more pronounced in mobile data, suggesting that people may use mobile phones to cope with stress due to their easier accessibility. It was also consistent across both mobile and desktop data for the Tier IV income group, supporting earlier findings that higher-income individuals may use shopping as a way to cope with stress [38]. In the 30-day mobile data, this positive association was observed among male participants but not among females. Previous studies show that males are more likely to experience negative emotions related to shopping [83], while some studies report no gender differences in online shopping addiction tendencies [29].

These findings highlight how shopping has become embedded in daily life and its potential influence on stress levels. Future research should explore different types of shopping (hedonic vs utilitarian) and their impact on mental health. Another direction could be to examine the time spent on shopping compared with actual purchases completed after the payment process, and how these different shopping behaviors are associated with stress levels.

Gaming

Gaming has become a widespread daily habit, with the global market projected to reach US $522.46 billion by 2025 [84]. Previous research presents mixed findings: while gaming has been positively linked to stress and lower psychosocial well-being, with stress being a known precursor to pathological gaming [32,33,85]. It has also been identified as a stress reliever and coping mechanism [39,86-88].

In our results, gaming usage was positively associated with stress in the 30-day mobile data for the high-stress subgroup. In the 2-day mobile data, positive associations were observed for the low-stress group, users aged older than 60 years, and those in the Tier II income range. A negative association was found for the Tier V income group. These findings align with previous research suggesting that gaming may intensify stress in already stressed individuals [89], whereas older adults often use gaming as a way to relieve stress [90]. The contrasting patterns across income groups may reflect differences in usage

XSL•FO
RenderX

intensity, as lower-income individuals tend to engage more frequently in gaming than those with higher incomes [91].

We did not find any significant associations in the desktop data, which may be due to lower gaming activity on desktop browsers compared with mobile apps—gaming accounted for only 4.7% of desktop usage versus 16.8% in mobile data. Future research should examine the type of gaming, such as games that require active cognitive engagement (eg, first-person shooter games) versus low cognitive requirement arcade games, and how these different types of games relate to stress.

Timing of Online Activity

Individuals tend to have preferences for the timing of both online and offline activities throughout the day [92,93]. Previous research has shown that increased nighttime smartphone use is linked to higher perceived stress [49]. Nighttime use has also been associated with reduced sleep duration [48,94] and poorer sleep quality [95,96], both of which are shown to impact mental health negatively [97-99].

In our results, the daytime–nighttime difference feature was mainly positively associated with perceived stress across various groups. A negative association was observed only in the age group of older than 60 years in the 30-day mobile data. Positive associations were more common in the shorter time frame, particularly in the 2-day mobile data, for groups including high-stress individuals, females, those aged 30 - 45 years, and the Tier IV income group (refer to Figure 2).

Our results indicate that daytime internet use, compared with nighttime use, is positively associated with stress, contrasting with earlier studies that reported a stronger link between nighttime use and stress in young adults [49]. Longitudinal research has shown that it is sleep loss due to time spent online, rather than internet use itself, that is associated with poorer mental health outcomes [6]. Internet use patterns can also be influenced by individuals' chronotypes, which in turn will affect preferred active hours [100]. Moreover, future studies could investigate panelists' bedtime and examine postbedtime usage, as previous studies have shown that postbedtime use—rather than nighttime use in general—has the most harmful effects on sleep quality [95]. The negative association observed in the age group of older than 60 years may suggest that older individuals are more sensitive to late-night use and its link to stress. Further research is needed to better understand how daily temporal patterns of internet use relate to stress, particularly by considering users' chronotype and bedtime.

Negative Associations

Adult Content

A recent study estimates that around 90 million people may be affected by problematic adult content usage [101]. Research has shown that problematic adult content consumption can impact mental health, including associations with higher stress [102,103].

In our results, however, adult content consumption was negatively associated with stress levels. This negative relationship was observed in the low-stress group (PSS score < 14), and in the age groups of 18 - 30 years and older than 60 years within the 30-day desktop data. Some studies have reported no significant link between adult content use and psychological health [104], while others suggest it is often consumed as a form of leisure [105]. Our results may suggest that adult content serves as a stress buffer for our participants, as previous research has indicated stress and stress relief as common motivations for its use [106,107]. Another possible explanation is that users with lower stress levels might engage in this activity as a form of leisure or a means to alleviate boredom [105]. Previous research has identified boredom proneness and its positive association with the frequency of pornography use [108]. However, the positive associations observed in past studies were tied to problematic adult content consumption or addiction. This implies that while limited consumption may offer stress relief, excessive use could have detrimental effects. Future research should explore both the positive and negative impacts of adult content consumption on mental health, particularly in relation to the amount of consumption.

Productivity

Information and communication technology has become a central part of daily work and study routines, particularly in the 21st century. Most of the work we do on information and communication technology devices is connected to the internet, and their use for work and productivity has been linked to improved workplace efficiency [109,110]. However, the effects of internet use on psychological health and stress are dual in nature. While it has been associated with increased productivity, higher internet use for work-related tasks has also been linked to higher stress levels [51,111-114]. Some interventional studies report no significant effects [113], while others suggest that using the internet for work and study can have a positive impact on mental health [6].

In our results, increased use of productivity-related apps and domains was generally associated with lower perceived stress. In the 30-day mobile data, negative associations were observed among all participants, as well as within the male subgroup, the age group of 31 - 45 years, and the Tier III income group. In the 30-day desktop data, a similar negative association was found in the Tier I income group. A positive association appeared only once—in the 2-day desktop data for the Tier V income group.

Our findings suggest that stressed individuals may avoid productivity-related tasks by shifting their focus to other activities. This is supported by previous research showing a positive relationship between perceived stress and avoidant coping styles [115]. Previous research has reported a negative relationship between work-related online tasks and stress among middle-aged adults [21]. At the same time, communication overload from emails and messages has been linked to higher perceived stress in adults aged 50 - 85 years [15]. In our results, we found a similar negative association between productivity-related use and stress in the age group of 31 - 45 years, suggesting that increased productivity use may be linked to lower stress in this group. However, no significant association was observed in the older age group. For higher-income groups, the observed positive association may reflect greater work

XSL•FO
RenderX

responsibilities that are more difficult to avoid, contributing to increased stress.

Overall, our findings highlight the need for further research into how productivity-related internet use influences stress. As previous studies have shown, avoidance behaviors can increase the risk of prolonged stress and other mental health challenges [116]. Future studies should explore how online interventions can effectively help users mitigate stress.

News

Informed citizens are a cornerstone of a well-functioning democracy and good governance. However, recent studies suggest that news, especially news with highly negative content, can adversely affect mental health and increase stress levels [24,117-119].

Our findings reveal a counterintuitive relationship between news engagement and stress: participants who spent more time-consuming news tended to report lower stress levels. This association was more pronounced in the 2-day data for both mobile and desktop users, and overall, it was stronger in desktop data across both time periods. One possible explanation is that individuals experiencing stress may avoid the news altogether in the short term. Previous research supports this, showing that people under stress often disengage from news consumption [120-122]. Other studies have found no significant link between news consumption and stress [123,124], while some suggest that positive and soft news content may improve mood [30] and well-being [125].

Since our news category includes a range of sources, such as entertainment, sports, and politics, it is important to consider how different types of news relate to stress in the future. Future studies could examine the effects of specific news genres, as well as the influence of low-quality news sources and misinformation on mental health. Examining how consistent exposure to different news categories, sources, and misinformation influences stress responses over time could provide insights into mental health risks and how media consumption habits shape psychological well-being.

Mixed Associations

Messaging

In 2021, an estimated 3.09 billion mobile phone users accessed the top messaging apps for communication, with this number projected to reach 3.51 billion by 2025 [126]. The younger generation, in particular, remains connected with friends and family through messaging apps such as WhatsApp (Meta Platforms, Inc) and Telegram (Telegram FZ-LLC). Research has shown that messaging apps can be both positively associated with stress [7,127-130] and serve as stress reducers [131-134].

In our results, messaging usage was mostly positively associated with stress in mobile data. It was positively associated in the age groups of 30 - 45 years and older than 60 years and negatively associated in the Tier I income group in the 30-day mobile data. In the 2-day mobile data, it was positively associated with females and those in the Tier III income group. In the 2-day desktop data, it was negatively associated with all participants and the age group of 45 - 60 years, while it was positively associated with the age group of older than 60 years.

The key finding is that messaging use was predominantly positively associated with stress in mobile data and more negatively associated in desktop data. Previous research suggests that “checking behavior” (ie, brief and repeated usage sessions) is more common in mobile use compared to desktop devices [53], and perhaps this type of repeated checking on mobile devices adds to the stress. We found that in the age group of older than 60 years, messaging was positively associated with stress in both mobile and desktop data, which aligns with previous studies showing that older individuals experience higher stress from interpersonal communication [7]. In the Tier I income group for mobile data, messaging was negatively associated with stress. Studies show that people in lower-income groups tend to send more messages [135]. Previous research has also shown that messaging can be an effective tool for reducing depression, particularly among low-income individuals [136]. Finally, in the 2-day mobile data, messaging was positively associated with stress in females but not males, reflecting previous research suggesting that frequent messaging is linked to mental health symptoms (externalizing and inattention) in females, but not in males [137].

Our results highlight how messaging can have a dual impact on stress, depending on the device, time period, and individual differences. Future work could examine how messaging with friends and family, interactions with strangers in chatrooms, and repeated checking behavior on mobile devices relate to stress, loneliness, and overall well-being.

Screen Time

The internet plays a pervasive role in our daily lives, and the amount of time we spend online may have a detrimental impact on stress levels [48,138]. Research has shown that smartphone addiction is significantly linked to higher stress [139,140], and the amount of time spent online is positively associated with stress [48] and other mental health issues [141,142].

In our results, total time spent online was positively associated with stress in the 30-day mobile data, aligning with previous studies [48,139,140]. This association was significant for the high-stress group, individuals older than 60 years, and those in income Tier IV. However, the relationship was mainly negative in both the 2-day and 30-day desktop data across different groups, with the exception of a positive association for the Tier IV income group in the 2-day desktop data, which mirrored the mobile data results.

Our findings suggest that increased time spent online on mobile devices may amplify stress in different contexts. For high-stress individuals, extended mobile phone usage is positively associated with stress, likely due to the constant accessibility of smartphones, making it harder to disconnect from them. Previous studies have also found that smartphone users tend to experience higher levels of digital burnout compared to those using desktops or laptops [17]. In addition, older individuals in our study showed an increase in stress with more time spent online, consistent with previous research [141].

XSL•FO
RenderX

In contrast, desktop usage generally showed negative associations with stress, which may be due to the differences in accessibility of smartphones and desktop devices. Our data show that time spent on desktop devices has higher variability, as reflected by the higher SD (71.43 for desktop vs 66.86 for mobile) and coefficient of variation (desktop: 0.97, CI 0.93 - 1.02 vs mobile: 0.70, CI 0.67 - 0.73), compared to time spent on mobile devices.

Future research should further investigate device differences and the role mobile devices play in predicting stress levels.

User Characteristics and Stress

User characteristics such as gender, age, and income influence both stress levels and their association with internet use, as provided in Figure 3. In our results, female users reported significantly higher stress than males across both mobile and desktop data and for both the 30-day and 2-day periods (refer to Figure 3). Age and income showed consistent negative associations with stress across contexts. Similar findings have been reported in earlier research on sociodemographics and stress [40-44]. Interestingly, in our analysis, income was the only sociodemographic factor significantly associated with stress in the high-stress group for mobile data, while no other sociodemographic features were significant across platforms or time frames. Previous research also suggests that income, rather than age or gender, shows the strongest association with stress [45]. This highlights the need for further investigation into the role of income in stress, as such insights can inform policies aimed at reducing income-related health disparities. Future studies should also incorporate measures to incorporate resilience of individuals (trait stress measures) [143] and personality traits alongside sociodemographics to better understand these associations.

XSL•FO
RenderX

Figure 3. Overview of significant sociodemographic associations across time frames (30 days vs 2 days) and device types (mobile vs desktop). Rows distinguish model results between all participants and high-stress participants. Red icons indicate positive associations, blue icons indicate negative associations, and empty cells denote no significant effects.

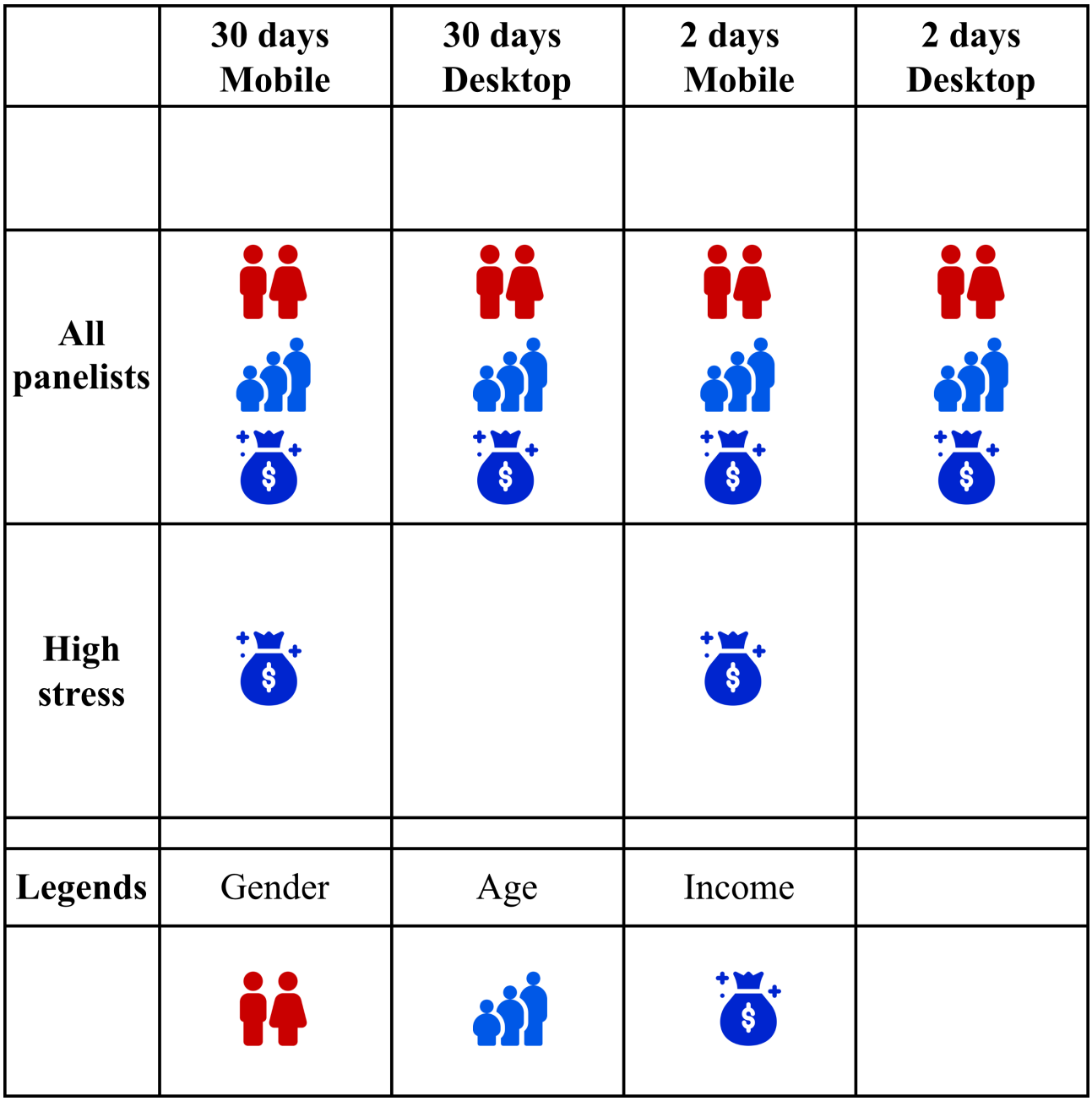

Implications

Building on these findings, it is important to consider the implications for various stakeholders, including digital platform designers, health care professionals, and end-users.

For digital platform designers, these findings suggest opportunities to promote healthier use patterns. Features such as reminders to take breaks, tools to visualize usage habits, and reduced notifications during evening/nights could help users avoid stress-inducing behaviors. In addition, adaptive designs that encourage daytime work engagement over late-night use could be particularly effective. Finally, different design considerations may be needed for mobile versus desktop devices to encourage healthier internet use.

For health care professionals, understanding how digital behaviors relate to stress could offer new ways to support patients. For example, excessive gaming on mobile devices might indicate elevated stress. Health workers could include questions about internet habits in assessments and recommend tools that encourage healthier online behaviors. Notably, the low cost and high availability of web data could provide efficient tools to complement traditional monthly surveys for monitoring stress, helping to alleviate the burden on the already strained health care sector.

For individuals, these findings emphasize the importance of timing and purpose in digital habits. Being mindful of potentially stress-inducing activities, such as excessive shopping or gaming, can foster a healthier balance. Tools that track and suggest

XSL•FO
RenderX

healthier patterns of internet use could assist users in managing their habits.

Overall, the relationship between stress and internet use is influenced by factors such as the type of activity, timing, and individual circumstances. This suggests that small, intentional changes in digital habits can help manage stress effectively.

Limitations

The panelists we collected data from are gig workers who regularly participate in surveys, which may lead to behavioral differences compared to the general population. To address this issue, we used rigorous preprocessing steps, including removing users who spent more than 25% of their time on survey websites and excluding the bottom 20% of users based on time spent. Moreover, browsing data from these panels has been shown to prominently feature the most visited domains in Germany [144]. Another limitation is that users may change their online behavior when they are aware of being tracked. However, previous work has demonstrated that the privacy attitudes of web-tracked panelists are comparable to those of nonweb-tracked panelists in Germany [145], though they may vary across countries. These findings support the reliability and suitability of the web-tracking data for capturing individuals' internet usage behavior. In addition, due to budget constraints, all panelists were from a single country, which may limit the generalizability of our findings because of cultural and behavioral differences. Previous research has demonstrated differences in cultural, social, and technological access contexts, variations in usage patterns, and mental health prevalence across countries [146,147]. Replicating this study in other countries or with a larger and more diverse sample could help further improve the generalizability and provide a broader cultural context. Further, 23% of participants dropped out between the first and final survey waves. To mitigate this attrition, we initially used a larger sample and invited all panelists from the panel company to participate in our baseline survey. Furthermore, since our tracker only tracks browsing behavior through the browser, we lack complete data on desktop usage through desktop apps, which may limit our ability to fully capture differences between mobile and desktop use. Finally, we used a validated self-report measure to assess stress, as inviting all participants to a lab setting was not feasible. Alternative approaches, such as physiological or real-time assessments, could provide more objective and detailed measures of stress.

Conclusion

Our study examines the relationship between internet use and perceived stress through a novel contextual framework that considers how, where, when, and by whom the internet is used. The findings indicate that internet use is associated with stress, and these associations differ across various usage contexts. Specifically, engagement with social media, online shopping, entertainment, and gaming is positively linked to higher stress levels. Notably, these activities have been identified in previous research as common coping mechanisms for stress, highlighting the need for future studies to examine whether such coping strategies alleviate stress or, possibly, exacerbate it over time. In contrast, productivity-related activities, news consumption, and adult content use are negatively associated with stress, suggesting they may either function as stress buffers or indicate avoidance behavior. Associations inferred from desktop data across different contextual dimensions are weaker than those inferred from mobile data, indicating that device type plays an important role. In the short term, news consumption is negatively associated with both mobile and desktop data. For individuals already experiencing high stress, increased time online on mobile phones—particularly on social media and gaming—is correlated with higher stress levels. In addition, sociodemographic factors, especially income, have significant associations with stress. These findings have important implications for the design of digital platforms, the development of mental health interventions, and the formation of healthier online habits. Future research should focus more specifically on particular web-based behaviors, such as news consumption and online shopping, and their effects on psychological well-being. It should also aim to establish causal links between internet use and stress and further investigate the mechanisms underlying sociodemographic differences in these associations.

Acknowledgments

This work was supported by the Helsinki Institute for Information Technology (HIIT). We would like to thank Yajing Wang for her help with data annotation and Emilia Marchese for her help with the data collection process. In addition, we acknowledge the computational resources provided by the Aalto Science-IT project.

Funding

This work was supported by the Helsinki Institute for Information Technology (HIIT). The funder had no involvement in the study design, data collection, analysis, interpretation, or manuscript writing.

Authors' Contributions

JK, MB, and TA led the conceptualization of the study. JK, MB, and TA were responsible for data curation and methodology development. MB and NL contributed to software implementation and validation. The investigation was conducted by JK, MB, NL, and TA, and data analysis was performed by MB and NL. JK and MB prepared the original draft, and JK, MB, NL, and TA contributed to reviewing and editing the paper. Visualizations were created by MB and NL. JK and TA provided supervision, acquired funding, and secured the necessary resources for the project.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Mixed-effects models results across contextual dimensions.
[DOCX File, 181 KB - jmir_v28i1e78775_app1.docx]

References

1. Schneiderman N, Ironson G, Siegel SD. Stress and health: psychological, behavioral, and biological determinants. Annu Rev Clin Psychol 2005;1:607-628. [doi: 10.1146/annurev.clinpsy.1.102803.144141] [Medline: 17716101]
2. Bui T, Zackula R, Dugan K, Ablah E. Workplace stress and productivity: a cross-sectional study. Kans J Med 2021;14(1):42-45. [doi: 10.17161/kjm.vol1413424] [Medline: 33654542]
3. Huang C. Internet use and psychological well-being: a meta-analysis. Cyberpsychol Behav Soc Netw 2010 Jun;13(3):241-249. [doi: 10.1089/cyber.2009.0217] [Medline: 20557242]
4. Çikrıkci Ö. The effect of internet use on well-being: meta-analysis. Comput Human Behav 2016 Dec;65:560-566. [doi: 10.1016/j.chb.2016.09.021]
5. Jenaro C, Flores N, Gómez-Vela M, González-Gil F, Caballo C. Problematic internet and cell-phone use: psychological, behavioral, and health correlates. Addict Res Theory 2007 Jan;15(3):309-320. [doi: 10.1080/16066350701350247]
6. Hökby S, Hadlaczky G, Westerlund J, et al. Are mental health effects of internet use attributable to the web-based content or perceived consequences of usage? A longitudinal study of european adolescents. JMIR Ment Health 2016 Jul 13;3(3):e31. [doi: 10.2196/mental.5925] [Medline: 27417665]
7. Nimrod G. Changes in internet use when coping with stress: older adults during the COVID-19 pandemic. Am J Geriatr Psychiatry 2020 Oct;28(10):1020-1024. [doi: 10.1016/j.jagp.2020.07.010] [Medline: 32771312]
8. Luo Y, Yip PSF, Zhang Q. Positive association between Internet use and mental health among adults aged ≥50 years in 23 countries. Nat Hum Behav 2025 Jan;9(1):90-100. [doi: 10.1038/s41562-024-02048-7] [Medline: 39558112]
9. Heo J, Chun S, Lee S, Lee KH, Kim J. Internet use and well-being in older adults. Cyberpsychol Behav Soc Netw 2015 May;18(5):268-272. [doi: 10.1089/cyber.2014.0549] [Medline: 25919967]
10. Linton MJ, Dieppe P, Medina-Lara A. Review of 99 self-report measures for assessing well-being in adults: exploring dimensions of well-being and developments over time. BMJ Open 2016 Jul 7;6(7):e010641. [doi: 10.1136/bmjopen-2015-010641] [Medline: 27388349]
11. Kraut R, Burke M. Internet use and psychological well-being: effects of activity and audience. Commun ACM 2015;58(12):94-100. [doi: 10.1145/2739043]
12. Yetton BD, Revord J, Margolis S, Lyubomirsky S, Seitz AR. Cognitive and physiological measures in well-being science: limitations and lessons. Front Psychol 2019;10:1630. [doi: 10.3389/fpsyg.2019.01630] [Medline: 31354601]
13. Patel VK. Study of internet use characteristics, perceived stress, and internet addiction among first-year medical students of Jamnagar, Gujarat, India. Ind J Priv Psychiatry 2019 Dec 1;13(2):44-47. [doi: 10.5005/jp-journals-10067-0037]
14. Nikolic A, Bukurov B, Kocic I, et al. Smartphone addiction, sleep quality, depression, anxiety, and stress among medical students. Front Public Health 2023;11:1252371. [doi: 10.3389/fpubh.2023.1252371] [Medline: 37744504]
15. Reinecke L, Aufenanger S, Beutel ME, et al. Digital stress over the life span: the effects of communication load and internet multitasking on perceived stress and psychological health impairments in a German probability sample. Media Psychol 2017 Jan 2;20(1):90-115. [doi: 10.1080/15213269.2015.1121832]
16. Barley SR, Meyerson DE, Grodal S. E-mail as a source and symbol of stress. Organ Sci 2011 Aug;22(4):887-906. [doi: 10.1287/orsc.1100.0573]
17. Göldağ B. An investigation of the relationship between university students' digital burnout levels and perceived stress levels. J learn teach digit age 2022;7(1):90-98. [doi: 10.53850/joltida.958039]
18. Becker L, Martin T, Rohleder N, Nieding G, Wannagat W. Physiological stress responses to digital single- and multitasking demands in younger and older adults. Psychoneuroendocrinology 2025 Apr;174:107376. [doi: 10.1016/j.psyneuen.2025.107376] [Medline: 39893951]
19. Campbell AJ, Cumming SR, Hughes I. Internet use by the socially fearful: addiction or therapy? Cyberpsychol Behav 2006 Feb;9(1):69-81. [doi: 10.1089/cpb.2006.9.69] [Medline: 16497120]
20. Stanković M, Nešić M, Čičević S, Shi Z. Association of smartphone use with depression, anxiety, stress, sleep quality, and internet addiction. Empirical evidence from a smartphone application. Pers Individ Dif 2021 Jan;168:110342. [doi: 10.1016/j.paid.2020.110342]
21. Khalili-Mahani N, Smyrnova A, Kakinami L. To each stress its own screen: a cross-sectional survey of the patterns of stress and various screen uses in relation to self-admitted screen addiction. J Med Internet Res 2019 Apr 2;21(4):e11485. [doi: 10.2196/11485] [Medline: 30938685]
22. Hampton K, Rainie L, Lu W, Shin I, Purcell K. Psychological stress and social media use. Pew Research Center. 2015. URL: https://www.pewresearch.org/internet/2015/01/15/psychological-stress-and-social-media-use-2/ [accessed 2025-12-11]

XSL•FO
RenderX

23. Nygaard M, Andersen TO, Rod NH. Can social connections become stressful? Exploring the link between social media use and perceived stress in cross-sectional and longitudinal analyses of 25,053 adults. J Ment Health 2024 Oct;33(5):596-604. [doi: 10.1080/09638237.2024.2332802] [Medline: 38545943]
24. McLaughlin B, Gotlieb MR, Mills DJ. Caught in a dangerous world: problematic news consumption and its relationship to mental and physical ill-being. Health Commun 2023 Dec;38(12):2687-2697. [doi: 10.1080/10410236.2022.2106086] [Medline: 35999665]
25. Grubbs JB, Volk F, Exline JJ, Pargament KI. Internet pornography use: perceived addiction, psychological distress, and the validation of a brief measure. J Sex Marital Ther 2015;41(1):83-106. [doi: 10.1080/0092623X.2013.842192] [Medline: 24341869]
26. Bezinović P, Roviš D, Rončević N, Bilajac L. Patterns of internet use and mental health of high school students in Istria County Croatia: cross-sectional study. Croat Med J 2015 Jun;56(3):297-305. [doi: 10.3325/cmj.2015.56.297] [Medline: 26088855]
27. Büchler N, ter Hoeven CL, van Zoonen W. Understanding constant connectivity to work: how and for whom is constant connectivity related to employee well-being? Inf Organ 2020 Sep;30(3):100302. [doi: 10.1016/j.infoandorg.2020.100302]
28. Elhai JD, Hall BJ. Anxiety about internet hacking: results from a community sample. Comput Human Behav 2016 Jan;54:180-185. [doi: 10.1016/j.chb.2015.07.057]
29. Li H, Ma X, Fang J, et al. Student stress and online shopping addiction tendency among college students in Guangdong Province, China: the mediating effect of the social support. IJERPH 2022;20(1):176. [doi: 10.3390/ijerph20010176]
30. Kelly CA, Sharot T. Web-browsing patterns reflect and shape mood and mental health. Nat Hum Behav 2025 Jan;9(1):133-146. [doi: 10.1038/s41562-024-02065-6] [Medline: 39572688]
31. Verma G, Bhardwaj A, Aledavood T, De Choudhury M, Kumar S. Examining the impact of sharing COVID-19 misinformation online on mental health. Sci Rep 2022 May 16;12(1):8045. [doi: 10.1038/s41598-022-11488-y] [Medline: 35577820]
32. Mun IB. Academic stress and first-/third-person shooter game addiction in a large adolescent sample: a serial mediation model with depression and impulsivity. Comput Human Behav 2023 Aug;145:107767. [doi: 10.1016/j.chb.2023.107767]
33. Park K, Son M, Chang H, Lee SK. The roles of stress, non-digital hobbies, and gaming time in adolescent problematic game use: a focus on sex differences. Comput Human Behav 2024 Feb;151:108002. [doi: 10.1016/j.chb.2023.108002]
34. Rosell J, Vergés A, Miranda-Castillo C, Sepúlveda-Caro S, Gómez M. Predictors, types of internet use, and the psychological well-being of older adults: a comprehensive model. J Gerontol B Psychol Sci Soc Sci 2022 Jul 5;77(7):1186-1196. [doi: 10.1093/geronb/gbac054] [Medline: 35286369]
35. van Ingen E, Utz S, Toepoel V. Online coping after negative life events: measurement, prevalence, and relation with internet activities and well-being. Soc Sci Comput Rev 2016;34(5):511-529. [doi: 10.1177/0894439315600322]
36. Boursier V, Musetti A, Gioia F, Flayelle M, Billieux J, Schimmenti A. Corrigendum: is watching tv series an adaptive coping strategy during the COVID-19 pandemic? Insights from an Italian community sample. Front Psychiatry 2021;12:698404. [doi: 10.3389/fpsyt.2021.698404] [Medline: 34149485]
37. Nabi RL, Torres DP, Prestin A. Guilty pleasure no more: the relative importance of media use for coping with stress. Journal of Media Psychology: Theories, Methods, and Applications 2017;29(3):126-136. [doi: 10.1027/1864-1105/a000223]
38. Maraz A, Yi S. Compulsive buying gradually increased during the first six months of the Covid-19 outbreak. J Behav Addict 2022 Mar 28;11(1):88-101. [doi: 10.1556/2006.2022.00002] [Medline: 35262509]
39. Šporčić B, Glavak-Tkalić R. The relationship between online gaming motivation, self-concept clarity and tendency toward problematic gaming. CP 2018;12(1). [doi: 10.5817/CP2018-1-4]
40. Graves BS, Hall ME, Dias-Karch C, Haischer MH, Apter C. Gender differences in perceived stress and coping among college students. PLOS ONE 2021;16(8):e0255634. [doi: 10.1371/journal.pone.0255634] [Medline: 34383790]
41. Matud MP. Gender differences in stress and coping styles. Pers Individ Dif 2004 Nov;37(7):1401-1415. [doi: 10.1016/j.paid.2004.01.010]
42. Almeida DM, Rush J, Mogle J, Piazza JR, Cerino E, Charles ST. Longitudinal change in daily stress across 20 years of adulthood: results from the national study of daily experiences. Dev Psychol 2023 Mar;59(3):515-523. [doi: 10.1037/dev0001469] [Medline: 36174182]
43. Johnson MD, Krahn HJ, Galambos NL. Perceived stress trajectories from age 25 to 50 years. Int J Behav Dev 2023 May;47(3):233-242. [doi: 10.1177/01650254221150887]
44. de Miquel C, Domènech-Abella J, Felez-Nobrega M, et al. The mental health of employees with job loss and income loss during the COVID-19 pandemic: the mediating role of perceived financial stress. Int J Environ Res Public Health 2022 Mar 8;19(6):3158. [doi: 10.3390/ijerph19063158] [Medline: 35328846]
45. Li R, Liu S, Huang C, Darabi D, Zhao M, Heinzel S. The influence of perceived stress and income on mental health in China and Germany. Heliyon 2023 Jun;9(6):e17344. [doi: 10.1016/j.heliyon.2023.e17344]
46. Rus HM, Tiemensma J. Social media under the skin: Facebook use after acute stress impairs cortisol recovery. Front Psychol 2017;8:1609. [doi: 10.3389/fpsyg.2017.01609] [Medline: 28974938]

XSL•FO
RenderX

47. Yang C, Mousavi S, Dash A, Gummadi KP, Weber I. Studying behavioral addiction by combining surveys and digital traces: a case study of tiktok. 2025 Presented at: Proceedings of the International AAAI Conference on Web and Social Media; Copenhagen, Denmark p. 2106-2123. [doi: 10.1609/icwsm.v19i1.35922]
48. Thomée S, Dellve L, Härenstam A, Hagberg M. Perceived connections between information and communication technology use and mental symptoms among young adults - a qualitative study. BMC Public Health 2010 Dec;10(1):1-14. [doi: 10.1186/1471-2458-10-66]
49. Dissing AS, Andersen TO, Jensen AK, Lund R, Rod NH. Nighttime smartphone use and changes in mental health and wellbeing among young adults: a longitudinal study based on high-resolution tracking data. Sci Rep 2022 May 15;12(1):8013. [doi: 10.1038/s41598-022-10116-z] [Medline: 35570230]
50. Wolfers LN, Utz S. Social media use, stress, and coping. Curr Opin Psychol 2022 Jun;45:101305. [doi: 10.1016/j.copsyc.2022.101305] [Medline: 35184027]
51. Afifi TD, Zamanzadeh N, Harrison K, Acevedo Callejas M. WIRED: the impact of media and technology use on stress (cortisol) and inflammation (interleukin IL-6) in fast paced families. Comput Human Behav 2018 Apr;81:265-273. [doi: 10.1016/j.chb.2017.12.010]
52. Thomas TA, Schmid AM, Kessling A, et al. Stress and compulsive buying-shopping disorder: a scoping review. Compr Psychiatry 2024 Jul;132:152482. [doi: 10.1016/j.comppsych.2024.152482] [Medline: 38603938]
53. Oulasvirta A, Rattenbury T, Ma L, Raita E. Habits make smartphone use more pervasive. Pers Ubiquit Comput 2012 Jan;16(1):105-114. [doi: 10.1007/s00779-011-0412-2]
54. Baryshnikov I, Aledavood T, Rosenström T, et al. Relationship between daily rated depression symptom severity and the retrospective self-report on PHQ-9: a prospective ecological momentary assessment study on 80 psychiatric outpatients. J Affect Disord 2023 Mar 1;324:170-174. [doi: 10.1016/j.jad.2022.12.127] [Medline: 36586594]
55. Hu Q, Li A, Heng F, Li J, Zhu T. Predicting depression of social media user on different observation windows. Presented at: 2015 IEEE / WIC / ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT); Dec 6-9, 2015; Singapore, Singapore p. 361-364. [doi: 10.1109/WI-IAT.2015.166]
56. Current population of germany. Statistisches Bundesamt (Destatis). 2023 Jun 20. URL: https://www.destatis.de/EN/Themes/Society-Environment/Population/Current-Population/_node.html#sprg480730 [accessed 2025-02-07]
57. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr;20(1):37-46. [doi: 10.1177/001316446002000104]
58. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [doi: 10.2307/2136404] [Medline: 6668417]
59. Philpott LF, Leahy-Warren P, FitzGerald S, Savage E. Prevalence and associated factors of paternal stress, anxiety, and depression symptoms in the early postnatal period. Glob Ment Health (Camb) 2022;9:306-321. [doi: 10.1017/gmh.2022.33] [Medline: 36561920]
60. Biswas B, Saha R, Haldar D, Saha I. Level of stress perception and predictors of higher stress perception among informal primary caregivers of Eastern Indian people living with HIV/AIDS. Int J Community Med Public Health 2019;6(10):4374. [doi: 10.18203/2394-6040.ijcmph20194497]
61. Bryk AS, Raudenbush SW. Hierarchical Linear Models: Applications and Data Analysis Methods: sage; 2002, Vol. 1.
62. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with python. 2010 Presented at: Python in Science Conference; Jun 28, 2010. [doi: 10.25080/Majora-92bf1922-011]
63. Thorn L, Evans P, Cannon A, Hucklebridge F, Clow A. Seasonal differences in the diurnal pattern of cortisol secretion in healthy participants and those with self-assessed seasonal affective disorder. Psychoneuroendocrinology 2011 Jul;36(6):816-823. [doi: 10.1016/j.psyneuen.2010.11.003] [Medline: 21145663]
64. Gassen J, Mengelkoch S, Slavich GM. Human immune and metabolic biomarker levels, and stress-biomarker associations, differ by season: Implications for biomedical health research. Brain, Behavior, & Immunity - Health 2024 Jul;38:100793. [doi: 10.1016/j.bbih.2024.100793]
65. Average daily time spent on internet and online media use in germany in the 3rd quarter of 2023, by device. Statista. 2024. URL: https://www.statista.com/statistics/719939/average-daily-internet-and-social-media-use-in-germany-by-device/ [accessed 2025-12-11]
66. Beyens I, Frison E, Eggermont S. “I don’t want to miss a thing”: adolescents’ fear of missing out and its relationship to adolescents’ social needs, Facebook use, and Facebook related stress. Comput Human Behav 2016 Nov;64:1-8. [doi: 10.1016/j.chb.2016.05.083]
67. Åberg E, Koivula A, Kukkonen I. A feminine burden of perfection? Appearance-related pressures on social networking sites. Telematics and Informatics 2020 Mar;46:101319. [doi: 10.1016/j.tele.2019.101319]
68. Chen W, Lee KH. Sharing, liking, commenting, and distressed? The pathway between Facebook interaction and psychological distress. Cyberpsychol Behav Soc Netw 2013 Oct;16(10):728-734. [doi: 10.1089/cyber.2012.0272] [Medline: 23745614]
69. Rus HM, Tiemensma J. Social media as a shield: Facebook buffers acute stress. Physiol Behav 2018 Mar;185:46-54. [doi: 10.1016/j.physbeh.2017.12.021]

XSL•FO
RenderX

70. Wolfers LN, Festl R, Utz S. Do smartphones and social network sites become more important when experiencing stress? Results from longitudinal data. Comput Human Behav 2020 Aug;109:106339. [doi: 10.1016/j.chb.2020.106339] [Medline: 32747849]
71. Brailovskaia J, Rohmann E, Bierhoff HW, Schillack H, Margraf J. The relationship between daily stress, social support and Facebook addiction disorder. Psychiatry Res 2019 Jun;276:167-174. [doi: 10.1016/j.psychres.2019.05.014] [Medline: 31096147]
72. Shen X, Wang JL. Loneliness and excessive smartphone use among Chinese college students: moderated mediation effect of perceived stressed and motivation. Comput Human Behav 2019 Jun;95:31-36. [doi: 10.1016/j.chb.2019.01.012]
73. Aghababian AH, Sadler JR, Jansen E, Thapaliya G, Smith KR, Carnell S. Binge watching during COVID-19: associations with stress and body weight. Nutrients 2021 Sep 28;13(10):3418. [doi: 10.3390/nu13103418] [Medline: 34684420]
74. Alimoradi Z, Jafari E, Potenza MN, Lin CY, Wu CY, Pakpour AH. Binge-watching and mental health problems: a systematic review and meta-analysis. Int J Environ Res Public Health 2022 Aug 6;19(15):9707. [doi: 10.3390/ijerph19159707] [Medline: 35955069]
75. Hunter IR, Gillen MC. Stress coping mechanisms in elderly adults: an initial study of recreational and other coping behaviors in nursing home patients. Adultspan Journal 2009 Mar;8(1):43-53 [FREE Full text] [doi: 10.1002/j.2161-0029.2009.tb00056.x]
76. van Gelder K. E-commerce worldwide - statistics & facts. Statista. URL: https://www.statista.com/topics/871/online-shopping/ [accessed 2025-12-11]
77. Tarka P, Kukar-Kinney M. Compulsive buying among young consumers in Eastern Europe: a two-study approach to scale adaptation and validation. JCM 2022 Feb 9;39(1):106-120. [doi: 10.1108/JCM-05-2020-3833]
78. Zheng Y, Yang X, Liu Q, Chu X, Huang Q, Zhou Z. Perceived stress and online compulsive buying among women: a moderated mediation model. Comput Human Behav 2020 Feb;103:13-20. [doi: 10.1016/j.chb.2019.09.012]
79. Singh R, Nayak JK. Life stressors and compulsive buying behaviour among adolescents in India. SAJGBR 2015 Aug 3;4(2):251-274. [doi: 10.1108/SAJGBR-08-2014-0054]
80. Hama Y. Shopping as a coping behavior for stress. Jpn Psychol Res 2001 Nov;43(4):218-224. [doi: 10.1111/1468-5884.00179]
81. Maharani SAD, Utami NP. Coping mechanisms of stress: the impact on online purchase impulsivity. JBMS 2023;3(3):164-180 [FREE Full text] [doi: 10.53748/jbms.v3i3.70]
82. Durante KM, Laran J. The effect of stress on consumer. Journal of Marketing Research 2016;53:814-828. [doi: 10.1509/jmr.15.0319]
83. Gallagher CE, Watt MC, Weaver AD, Murphy KA. “I fear, therefore, I shop!” exploring anxiety sensitivity in relation to compulsive buying. Pers Individ Dif 2017 Jan;104:37-42. [doi: 10.1016/j.paid.2016.07.023]
84. Games - worldwide. Statista. 2025. URL: https://www.statista.com/outlook/amo/media/games/worldwide [accessed 2025-12-11]
85. Lemmens JS, Valkenburg PM, Peter J. Psychosocial causes and consequences of pathological gaming. Comput Human Behav 2011 Jan;27(1):144-152. [doi: 10.1016/j.chb.2010.07.015]
86. Whitbourne SK, Ellenberg S, Akimoto K. Reasons for playing casual video games and perceived benefits among adults 18 to 80 years old. Cyberpsychol Behav Soc Netw 2013 Dec;16(12):892-897. [doi: 10.1089/cyber.2012.0705] [Medline: 23971430]
87. Desai V, Gupta A, Andersen L, Ronnestrand B, Wong M. Stress-reducing effects of playing a casual video game among undergraduate students. Trends Psychol 2021;29(3):563-579. [doi: 10.1007/s43076-021-00062-6] [Medline: 40477391]
88. Lee YH, Chen M. Seeking a sense of control or escapism? The role of video games in coping with unemployment. Games and Culture 2023 May;18(3):339-361. [doi: 10.1177/15554120221097413]
89. Snodgrass JG, Lacy MG, Dengah HJF II, Eisenhauer S, Batchelder G, Cookson RJ. A vacation from your mind: problematic online gaming is a stress response. Comput Human Behav 2014 Sep;38:248-260. [doi: 10.1016/j.chb.2014.06.004]
90. Seçer İ, Us E. Digital gaming trends of middle-aged and older adults: a sample from Turkey. Simul Gaming 2023 Feb;54(1):85-103. [doi: 10.1177/10468781221144184]
91. Engelstätter B, Ward MR. Video games become more mainstream. Entertain Comput 2022 May;42:100494. [doi: 10.1016/j.entcom.2022.100494]
92. Aledavood T, Lehmann S, Saramäki J. Digital daily cycles of individuals. Front Phys 2015;3:73. [doi: 10.3389/fphy.2015.00073]
93. Luong N, Barnett I, Aledavood T. The impact of the COVID-19 pandemic on daily rhythms. J Am Med Inform Assoc 2023 Nov 17;30(12):1943-1953. [doi: 10.1093/jamia/ocad140]
94. Schrempft S, Baysson H, Chessa A, et al. Associations between bedtime media use and sleep outcomes in an adult population-based cohort. Sleep Med 2024 Sep;121:226-235. [doi: 10.1016/j.sleep.2024.06.029]
95. Siebers T, Beyens I, Baumgartner SE, Valkenburg PM. Adolescents’ digital nightlife: the comparative effects of day- and nighttime smartphone use on sleep quality. Communic Res 2024. [doi: 10.1177/00936502241276793]
96. Luqman A, Masood A, Shahzad F, Shahbaz M, Feng Y. Untangling the adverse effects of late-night usage of smartphone-based SNS among university students. Behav Inf Technol 2021 Nov 18;40(15):1671-1687. [doi: 10.1080/0144929X.2020.1773538]

XSL•FO
RenderX

97. Thomée S, Härenstam A, Hagberg M. Mobile phone use and stress, sleep disturbances, and symptoms of depression among young adults - a prospective cohort study. BMC Public Health 2011 Dec;11(1):1-11. [doi: 10.1186/1471-2458-11-66]
98. Thomée S, Härenstam A, Hagberg M. Computer use and stress, sleep disturbances, and symptoms of depression among young adults – a prospective cohort study. BMC Psychiatry 2012 Dec;12(1):1-14. [doi: 10.1186/1471-244X-12-176]
99. Price GD, Heinz MV, Song SH, Nemesure MD, Jacobson NC. Using digital phenotyping to capture depression symptom variability: detecting naturalistic variability in depression symptoms across one year using passively collected wearable movement and sleep data. Transl Psychiatry 2023 Dec 9;13(1):381. [doi: 10.1038/s41398-023-02669-y] [Medline: 38071317]
100. Kortesoja L, Vainikainen MP, Hotulainen R, Merikanto I. Late-night digital media use in relation to chronotype, sleep and tiredness on school days in adolescence. J Youth Adolesc 2023 Feb;52(2):419-433. [doi: 10.1007/s10964-022-01703-4] [Medline: 36401709]
101. Bőthe B, Nagy L, Koós M, et al. Problematic pornography use across countries, genders, and sexual orientations: insights from the International Sex Survey and comparison of different assessment tools. Addiction 2024 May;119(5):928-950. [doi: 10.1111/add.16431] [Medline: 38413365]
102. Laier C, Brand M. Mood changes after watching pornography on the Internet are linked to tendencies towards Internet-pornography-viewing disorder. Addict Behav Rep 2017 Jun;5:9-13. [doi: 10.1016/j.abrep.2016.11.003] [Medline: 29450222]
103. Altin M, De Leo D, Tribbia N, Ronconi L, Cipolletta S. Problematic pornography use, mental health, and suicidality among young adults. Int J Environ Res Public Health 2024 Sep 18;21(9):1228. [doi: 10.3390/ijerph21091228] [Medline: 39338111]
104. Harper C, Hodgins DC. Examining correlates of problematic internet pornography use among university students. J Behav Addict 2016 Jun;5(2):179-191. [doi: 10.1556/2006.5.2016.022] [Medline: 27156383]
105. McCormack M, Wignall L. Enjoyment, exploration and education: understanding the consumption of pornography among young men with non-exclusive sexual orientations. Sociology 2017 Oct;51(5):975-991. [doi: 10.1177/0038038516629909] [Medline: 28989197]
106. Bőthe B, Tóth-Király I, Bella N, Potenza MN, Demetrovics Z, Orosz G. Why do people watch pornography? The motivational basis of pornography use. Psychol Addict Behav 2021 Mar;35(2):172-186. [doi: 10.1037/adb0000603] [Medline: 32730047]
107. Mubeen B, Ashraf D. Psychological predictors of adult content consumption: a qualitative analysis by grounded theory. Asian J Psychiatr 2022;23(7). [doi: 10.54615/2231-7805.47266]
108. Moynihan AB, Igou ER, van Tilburg WAP. Pornography consumption as existential escape from boredom. Pers Individ Dif 2022 Nov;198:111802. [doi: 10.1016/j.paid.2022.111802]
109. Federico B. ICT and productivity: a review of the literature. : European Commission, Joint Research Centre (JRC); 2013. [doi: 10.2788/32940]
110. Buhari A, A. AA. Influence of internet and its connectivity in workplace - a comprehensive analysis. RRRJ 2024;3(1):244-257 [FREE Full text] [doi: 10.36548/rrrj.2024.1.016]
111. Mark G, Voida S, Cardello A. A pace not dictated by electrons": an empirical study of work without email. CHI ’12: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems :555-564. [doi: 10.1145/2207676.2207754]
112. Ninaus K, Diehl S, Terlutter R, Chan K, Huang A, Erlandsson S. Benefits and stressors - perceived effects of ICT use on employee health and work stress: an exploratory study from Austria and Hong Kong. Int J Qual Stud Health Well-being 2015;10:28838. [doi: 10.3402/qhw.v10.28838] [Medline: 26462972]
113. Berg-Beckhoff G, Nielsen G, Ladekjær Larsen E. Use of information communication technology and stress, burnout, and mental health in older, middle-aged, and younger workers - results from a systematic review. Int J Occup Environ Health 2017 Apr;23(2):160-171. [doi: 10.1080/10773525.2018.1436015] [Medline: 29460697]
114. Malott L, Vishwanathan SP, Chellappan S. Differences in internet usage patterns with stress and anxiety among college students. Presented at: 2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013); Oct 9-12, 2013; Lisbon, Portugal p. 664-668. [doi: 10.1109/HealthCom.2013.6720759]
115. Allen MT. Explorations of avoidance and approach coping and perceived stress with a computer-based avatar task: detrimental effects of resignation and withdrawal. PeerJ 2021;9:e11265. [doi: 10.7717/peerj.11265]
116. Holahan CJ, Moos RH, Holahan CK, Brennan PL, Schutte KK. Stress generation, avoidance coping, and depressive symptoms: a 10-year model. J Consult Clin Psychol 2005 Aug;73(4):658-666. [doi: 10.1037/0022-006X.73.4.658] [Medline: 16173853]
117. Shaffer SR, Dolovich C, El-Gabalawy R, et al. The impact of source and consumption of news on mental distress among inflammatory bowel disease patients during the COVID-19 pandemic. J Can Assoc Gastroenterol 2024 Apr;7(2):212-218. [doi: 10.1093/jcag/gwad060] [Medline: 38596803]
118. Ladis I, Gao C, Scullin MK. COVID-19-related news consumption linked with stress and worry, but not sleep quality, early in the pandemic. Psychol Health Med 2023 Apr;28(4):980-994. [doi: 10.1080/13548506.2022.2141281] [Medline: 36322027]
119. Kellerman JK, Hamilton JL, Selby EA, Kleiman EM. The mental health impact of daily news exposure during the COVID-19 pandemic: Ecological Momentary Assessment study. JMIR Ment Health 2022 May 25;9(5):e36966. [doi: 10.2196/36966] [Medline: 35377320]
120. Nguyen A, Smith A, Jackson D, Zhao X. Pandemic news experience: COVID-19, news consumption, mental health, and the demand for positive news. SSRN Journal 2021. [doi: 10.2139/ssrn.3832669]

XSL•FO
RenderX

121. Lindell J, Mikkelsen Båge E. Disconnecting from digital news: news avoidance and the ignored role of social class. Journalism 2023 Sep;24(9):1980-1997. [doi: 10.1177/14648849221085389]
122. Mannell K, Meese J. From doom-scrolling to news avoidance: limiting news as a wellbeing strategy during COVID lockdown. Journal Stud 2022 Feb 17;23(3):302-319. [doi: 10.1080/1461670X.2021.2021105]
123. McNaughton-cassill ME. The news media and psychological distress. Anxiety, Stress & Coping 2001 Apr;14(2):193-211. [doi: 10.1080/10615800108248354]
124. Lavelle B. Investigating whether the consumption of news impacts measures for anxiety, stress, and well-being. Https://norma.ncirl.ie/id/eprint/5657. 2022.
125. Boukes M, Vliegenthart R. News consumption and its unpleasant side effect. J Media Psychol 2017 Jul;29(3):137-147. [doi: 10.1027/1864-1105/a000224]
126. Mobile messaging users worldwide 2025. Statista. 2021. URL: https://www.statista.com/statistics/483255/number-of-mobile-messaging-users-worldwide/ [accessed 2025-12-11]
127. Coccia C, Darling CA. Having the time of their life: college student stress, dating and satisfaction with life. Stress Health 2016 Feb;32(1):28-35. [doi: 10.1002/smi.2575] [Medline: 24723539]
128. Thomée S, Eklöf M, Gustafsson E, Nilsson R, Hagberg M. Prevalence of perceived stress, symptoms of depression and sleep disturbances in relation to information and communication technology (ICT) use among young adults – an explorative prospective study. Comput Human Behav 2007 May;23(3):1300-1321. [doi: 10.1016/j.chb.2004.12.007]
129. Hurbean L, Dospinescu O, Munteanu V, Danaiata D. Effects of instant messaging related technostress on work performance and well-being. Electronics (Basel) 2022;11(16):2535. [doi: 10.3390/electronics11162535]
130. Lin XY, Lachman ME. Daily stress and affect across adulthood: the role of social interactions via different communication modes. Technol Mind Behav 2021;2(1):10. [doi: 10.1037/tmb0000026] [Medline: 35369392]
131. Melumad S, Pham MT. The smartphone as a pacifying technology. J Consum Res 2020 Aug 1;47(2):237-255. [doi: 10.1093/jcr/ucaa005]
132. Yau JC, Reich SM, Lee TY. Coping with stress through texting: an experimental study. J Adolesc Health 2021 Mar;68(3):565-571. [doi: 10.1016/j.jadohealth.2020.07.004] [Medline: 32798096]
133. Holtzman S, DeClerck D, Turcotte K, Lisi D, Woodworth M. Emotional support during times of stress: Can text messaging compete with in-person interactions? Comput Human Behav 2017 Jun;71:130-139. [doi: 10.1016/j.chb.2017.01.043]
134. Hooker ED, Campos B, Pressman SD. It just takes a text: partner text messages can reduce cardiovascular responses to stress in females. Comput Human Behav 2018 Jul;84:485-492. [doi: 10.1016/j.chb.2018.02.033]
135. Smith A. How Americans use text messaging. Pew Research Center. URL: https://www.pewresearch.org/internet/2011/09/19/how-americans-use-text-messaging/ [accessed 2025-12-11]
136. Aguilera A, Muñoz RF. Text messaging as an adjunct to CBT in low-income populations: a usability and feasibility pilot study. Prof Psychol Res Pr 2011 Dec 1;42(6):472-478. [doi: 10.1037/a0025499] [Medline: 25525292]
137. George MJ, Beron K, Vollet JW, Burnell K, Ehrenreich SE, Underwood MK. Frequency of text messaging and adolescents' mental health symptoms across 4 years of high school. J Adolesc Health 2021 Feb;68(2):324-330. [doi: 10.1016/j.jadohealth.2020.06.012] [Medline: 32753344]
138. Yu CC, Tou NX, Low JA. Internet use and effects on mental well-being during the lockdown phase of the COVID-19 pandemic in younger versus older adults: observational cross-sectional study. JMIR Form Res 2024 Feb 6;8:e46824. [doi: 10.2196/46824] [Medline: 38319700]
139. Elamin NO, Almasaad JM, Busaeed RB, Aljafari DA, Khan MA. Smartphone addiction, stress, and depression among university students. Clin Epidemiol Glob Health 2024 Jan;25:101487. [doi: 10.1016/j.cegh.2023.101487]
140. Khan A, McLeod G, Hidajat T, Edwards EJ. Excessive smartphone use is associated with depression, anxiety, stress, and sleep quality of Australian adults. J Med Syst 2023 Oct 20;47(1):109. [doi: 10.1007/s10916-023-02005-3] [Medline: 37858009]
141. Ding L, Li Z, Jiang H, Zhang X, Xiong Z, Zhu X. Mobile phone problem use and depressive symptoms: the mediating role of social support and attitude to aging among Chinese older adults. BMC Psychiatry 2024 Feb 16;24(1):135. [doi: 10.1186/s12888-024-05565-x] [Medline: 38365625]
142. Yue C, Ware S, Morillo R, et al. Automatic depression prediction using Internet traffic characteristics on smartphones. Smart Health (2014) 2020 Nov;18:100137. [doi: 10.1016/j.smhl.2020.100137]
143. Bartone PT, Ursano RJ, Wright KM, Ingraham LH. The impact of a military air disaster on the health of assistance workers. A prospective study. J Nerv Ment Dis 1989 Jun;177(6):317-328. [doi: 10.1097/00005053-198906000-00001] [Medline: 2723619]
144. Kulshrestha J, Oliveira M, Karaçalık O, Bonnay D, Wagner C. Web routineness and limits of predictability: investigating demographic and behavioral differences using web tracking data. ICWSM 2021;15:327-338. [doi: 10.1609/icwsm.v15i1.18064]
145. Stier S, Kirkizh N, Froio C, Schroeder R. Populist attitudes and selective exposure to online news: a cross-country analysis combining web tracking and surveys. Int J Press Polit 2020 Jul;25(3):426-446. [doi: 10.1177/1940161220907018]
146. Grothaus C, Dolch C, Zawacki-Richter O. Use of digital media in higher education across country contexts: a comparison between Germany and Thailand. Int J Emerg Technol Learn 2021;16(20):64. [doi: 10.3991/ijet.v16i20.24263]

XSL•FO
RenderX

147. Lopez-Fernandez O, Romo L, Kern L, et al. Problematic internet use among adults: a cross-cultural study in 15 countries. J Clin Med 2023 Jan 29;12(3):1027. [doi: 10.3390/jcm12031027] [Medline: 36769675]

Abbreviations

LMM: linear mixed-effects model
PSS: Perceived Stress Scale

Edited by K Liew; submitted 09.Jun.2025; peer-reviewed by A Hinz, S Biswas, Y Zhang; revised version received 17.Sep.2025; accepted 06.Oct.2025; published 09.Jan.2026.

Please cite as:
Belal M, Luong N, Aledavood T, Kulshrestha J
Examining the Association Between Internet Use and Perceived Stress in Adults: Longitudinal Observational Study Combining Web Tracking Data With Questionnaires
J Med Internet Res 2026;28:e78775
URL: https://www.jmir.org/2026/1/e78775
doi:10.2196/78775

© Mohammad Belal, Nguyen Luong, Talayeh Aledavood, Juhi Kulshrestha. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Impact of Evaluation Strategy on Sepsis Prediction Model Performance Metrics in Intensive Care Data: Retrospective Cohort Study

Dang-Khoa Do[1], State Examination in Medicine; Patrick Rockenschaub[2], PhD; Sebastian Daniel Boie[1], PhD; Oliver Kumpf[3], Dr med; Hans-Dieter Volk[4], Dr med; Felix Balzer[1], Dr med, Dr rer nat; Falk von Dincklage[5], Dr med; Gregor Lichtner[1,5], Dr rer nat

[1]Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[2]Institute of Clinical Epidemiology, Public Health, Health Economics, Medical Statistics and Informatics, Medical University of Innsbruck, Innsbruck, Austria

[3]Department of Anesthesiology and Intensive Care Medicine (CCM/CVK), Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[4]Institute of Medical Immunology, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[5]Department of Anesthesia, Critical Care, Emergency and Pain Medicine, Universitätsmedizin Greifswald, Greifswald, Germany

Corresponding Author:
Gregor Lichtner, Dr rer nat
Department of Anesthesia, Critical Care, Emergency and Pain Medicine
Universitätsmedizin Greifswald
Ferdinand-Sauerbruch-Straße
Greifswald, 17475
Germany
Phone: 49 3834 86 5801
Email: gregor.lichtner@med.uni-greifswald.de

Abstract

Background: The prediction of the onset of sepsis, a life-threatening condition resulting from a dysregulated response to an infection, is one of the most common prediction tasks in intensive care–related machine learning research. To assess the performance of such models, different evaluation strategies, including fixed horizon (a single prediction at a set time before onset), peak score (a single prediction using the maximum predicted risk across time), and continuous evaluation (multiple predictions assessed continuously across time), are commonly implemented, but there is no clear consensus on which approach should be used in order to provide clinically meaningful performance evaluation.

Objective: This study aimed to assess different evaluation approaches of sepsis prediction models trained on a public intensive care dataset applied to German intensive care data.

Methods: In this retrospective, observational cohort study, we assessed the efficacy of machine learning models, pretrained on the Medical Information Mart for Intensive Care IV dataset, when applied to BerlinICU, a multisite German intensive care dataset. To understand the real-world impact of implementing these models, we examined the performance variability across various evaluation strategies.

Results: The BerlinICU dataset includes 40,132 intensive care admissions spanning 10 years (2012-2021). Using the latest Sepsis-3 definition, we identified 4134 septic admissions (10.3% prevalence). Application of a temporal convolutional network model to BerlinICU yielded an area under the receiver operating characteristic curve (AUROC) of 0.67 (95% CI 0.66-0.68) for continuous evaluation with a 6-hour prediction horizon, compared with 0.84 (95% CI 0.83-0.85) on the test set of Medical Information Mart for Intensive Care IV. On BerlinICU, peak score evaluation showed a similar AUROC compared with continuous evaluation, while fixed horizon evaluation showed a reduced AUROC of 0.61 (95% CI 0.60-0.62). Onset matching had minimal impact on performance estimates using continuous evaluation or fixed horizon evaluation, but increased estimates for peak score evaluation. Performance metrics improved with shorter prediction horizons across all strategies.

XSL•FO
RenderX

Conclusions: Our results demonstrate that the choice of evaluation strategy has a significant impact on the performance metrics of intensive care prediction models. The same model applied to the same dataset yields markedly different performance metrics depending on the evaluation approach. Therefore, careful selection of the evaluation approach is essential to ensure that the interpretation of performance metrics aligns with clinical intentions and enables meaningful comparisons between studies. In our view, the continuous evaluation approach best reflects the continual monitoring of patients that is performed in real-world clinical practice. In contrast, fixed-horizon and peak score evaluation approaches may produce skewed results when not properly matching the length of stay distributions between sepsis cases and control cases. Especially for peak score evaluation, longer visits tend to produce higher maximum scores because sampling from more values increases the likelihood of capturing higher values purely by chance.

(J Med Internet Res 2026;28:e72083) doi:10.2196/72083

KEYWORDS

sepsis; machine learning; deep learning; intensive care; intensive care unit; ICU; early prediction; early warning; electronic health records; prediction; generalizability

Introduction

Sepsis is a life-threatening condition resulting from a dysregulated response to an infection that leads to organ dysfunction [1], responsible for up to 20% of all deaths worldwide [2]. Beyond its significant mortality rate [3], sepsis represents a tremendous financial burden, totaling to almost US $24 billion in the United States in 2013 [4]. Those who survive may experience long-term health consequences, including a diminished quality of life [5].

As each hour of delayed therapy increases mortality [6-8], sepsis treatment guidelines continue to recommend early therapy within the first hour of sepsis recognition [9-11]. Therefore, an early recognition of sepsis is essential. Numerous studies have investigated machine learning to aid early prediction of sepsis in the intensive care unit (ICU) [12,13], including prospective application [14,15] and external validation on data of different hospitals [14,16-18]. However, the strategies used in previous studies to evaluate model performance vary widely, particularly in the criteria for data selection and how predictions are aligned with respect to the sepsis onset. The most commonly used evaluation strategies can be broadly categorized into three groups, (1) fixed horizon evaluation, which assesses the model's ability to predict sepsis at a fixed number of hours before onset [13,19]; (2) peak score evaluation, which uses the highest prediction score across time before onset [16,20]; and (3) continuous evaluation, which considers all prediction scores across time before onset [13].

While these strategies are widely used in the literature, it remains uncertain how the choice among them influences the estimated model performance and whether certain strategies may systematically inflate or deflate metrics compared with others, even when applied to the exact same model and dataset. This has direct implications for clinical deployment and benchmarking: if a model is selected based on an evaluation strategy that does not reflect its intended clinical use (eg, hourly real-time monitoring vs one-time assessment), its actual performance may fall short of expectations. Similarly, comparing models across studies without accounting for differences in evaluation strategy risks drawing misleading conclusions about their relative performance.

In addition to differences in evaluation strategy, the literature also uses a range of performance metrics. Standard measures include the area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), and threshold-based metrics, such as positive predictive value (PPV) and negative predictive value (NPV). Threshold-based metrics reflect clinical utility at specific decision points, which is critical for deployment. In contrast, AUROC is threshold-agnostic, capturing model discrimination across all thresholds, and is robust to class imbalance, which makes it the predominant choice in the literature [21]. Beyond standard performance metrics, recent sepsis-prediction challenges have adopted clinically informed cost functions that capture specific clinical trade-offs. The PhysioNet/CinC 2019 challenge used a time-weighted continuous utility score, rewarding early detection while penalizing very early, late, and false alarms across the entire ICU trajectory. In contrast, the Pediatric Sepsis Data Challenge evaluates a single admission-time prediction per patient and scores models by their true-positive rate at a fixed maximal false-positive rate [22,23]. These examples highlight that not only evaluation strategies, but also the choice of performance metric, can fundamentally shape how sepsis prediction models are interpreted and compared.

In this study, we aimed to investigate how the choice of the evaluation approach influences the estimated performance of sepsis prediction models in an external validation context. Specifically, we compared the performance of models trained on a public intensive care dataset from the United States and subsequently applied them to German intensive care data. In doing so, we also address the lack of studies involving large-scale ICU cohorts from the German health care context, where the efficacy of sepsis prediction models—given the distinct patient demographics, treatment approaches, and health care policies—remains largely uninvestigated.

Methods

Study Design and Setting

This is a retrospective cohort study for evaluating differences in performance metrics under different, commonly used evaluation strategies. Temporal convolutional network (TCN) models were developed on Medical Information Mart for Intensive Care IV (MIMIC-IV v2.0; USA) and applied to

XSL•FO
RenderX

BerlinICU, a cohort of adult ICU stays from 8 ICUs at Charité—Universitätsmedizin Berlin across 2 sites (Campus Virchow Klinikum and Campus Charité Mitte) from 2012 to 2021. No clinical intervention or change in patient management was implemented; all data were collected during routine care.

Participants and Flow

We included ICU stays from adult patients (≥18 y). We excluded stays originating from non-ICU wards, stays with <6 hours of data, gaps ≥12 hours, sepsis onset within 4 hours after admission, and truncated each stay at 7 days. Evaluation samples for different evaluation strategies were drawn from this cohort.

Primary and Secondary Outcome Measures

The clinical outcome was sepsis onset, defined according to Sepsis-3 with an adapted suspicion-of-infection definition for BerlinICU (refer to "Outcome definition" section). The primary performance measure used to compare evaluation strategies was the AUROC. Secondary performance measures included the AUPRC, PPV, NPV, and recall at fixed thresholds. We additionally report "lifted" PPV and AUPRC (each divided by the horizon-specific prevalence) and recall at the threshold where precision equals twice the prevalence, as well as Brier score, alarm rate, and net benefit to assess calibration and clinical utility.

Datasets

We pretrained models on the MIMIC-IV dataset (version 2.0), a well-known ICU dataset from the United States that is commonly used to develop clinical prediction models [12,13]. We applied these models to BerlinICU, a comprehensive dataset from 8 ICUs from Charité - Universitätsmedizin Berlin in Germany, one of Europe's largest university hospitals. The ICUs were distributed across 2 different sites (Campus Virchow Klinikum and Campus Charité Mitte).

BerlinICU was extracted from the intensive care data management systems and the hospital information system present at the Charité - Universitätsmedizin Berlin. Specifically, we extracted the same clinical features that we used to pretrain the models, resulting in a total of 48 time-varying features, including vital parameters and laboratory results, along with 4 static features (age, sex, height, and weight; Table S1 in Multimedia Appendix 1). These features comprise demographic variables, vital signs, arterial blood gas values, electrolyte and blood count parameters, as well as treatment-related variables (eg, fraction of inspired oxygen) and clinical output measures (eg, urine output). All clinical feature values outside of physiologically plausible ranges, as defined by the ricu package [24], were removed. The remaining data was aggregated into intervals of one hour using their median value. Missing values were imputed using a forward fill strategy. If no previous values were available, the training set mean across all patients, regardless of sepsis status, for that feature was used. This corresponds to zero imputation if performed after feature standardization and is common practice in machine learning studies on sepsis prediction [17,25,26]. Outlier influence was mitigated by previous filtering based on physiological plausibility. Data were excluded from the dataset if they originated from nonintensive care wards, involved patients younger than 18 years, had a sepsis onset within 4 hours after admission, contained less than 6 hours of data, or included gaps equal to or exceeding 12 hours. The dataset was restricted to data from the years 2012-2021 and limited to at most 7 days following admission to the ICU.

Ethical Considerations

This study was approved by the local ethics committee (Ethikausschuss am Campus Virchow-Klinikum, Charité—Universitätsmedizin Berlin, Chairperson PD Dr E Kaschina, application EA2/137/22, approval date July 12, 2022, and amendment date May 10, 2023). Informed consent was not required due to the retrospective, observational nature of the study and the use of routinely collected data in accordance with applicable regulations. No compensation was provided.

Outcome Definition

To identify the onset of sepsis, we used the latest Sepsis-3 definition criteria, which require the co-occurrence of a suspicion of infection (SI) and an increase of the Sequential Organ Failure Assessment (SOFA) score by at least 2 points [1]. An SI is present when antibiotic treatment is coupled with microbiological culture tests. As the BerlinICU dataset did not include microbiological data, we used an alternative definition for SI in line with previous studies [16,17]. We inferred SI if a patient was treated with antibiotics administered at least once every 24 hours for at least 3 days, and defined the time of suspicion as the beginning of this treatment. Following previous studies, we defined an SI window ranging from 48 hours before to 24 hours after the time of SI (Figure 1). We assumed the onset of sepsis if an increase of at least 2 points of the SOFA score occurred during that SI window. The sepsis onset time point was defined as the time point of the recorded SOFA increase.

XSL•FO
RenderX

Figure 1. Preprocessing Pipeline. The BerlinICU dataset was extracted from the hospital information system and cleaned. We inferred sepsis onset according to the Sepsis-3 definition. For the logistic regression baseline models, we created additional temporal features. The final dataset was then passed to the models and evaluated. ICU: intensive care unit; MAP: mean arterial pressure; MIMIC-IV: Medical Information Mart for Intensive Care IV; SI: suspicion of infection; SOFA: Sequential Organ Failure Assessment score.

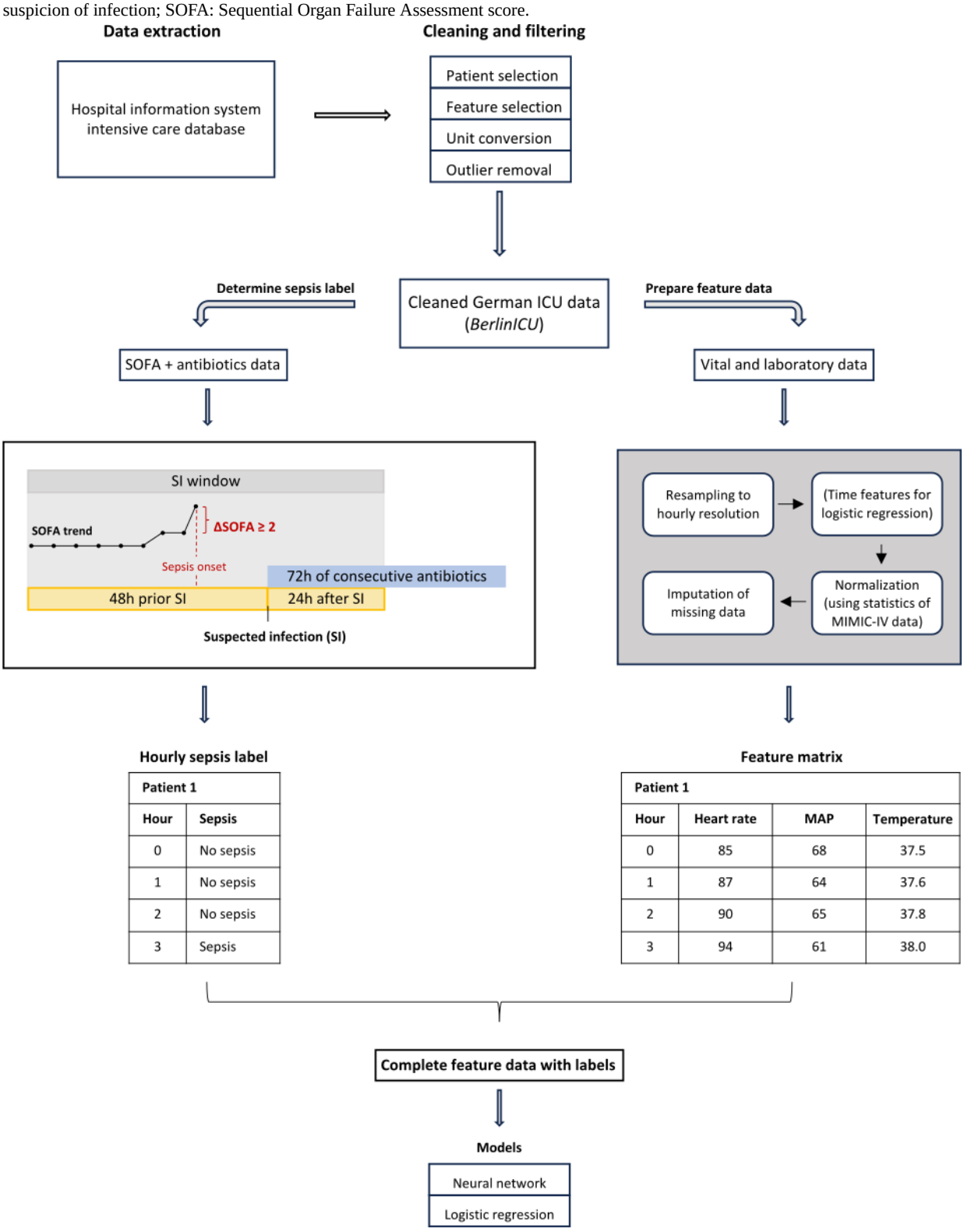

Model Training

We used TCN models trained on the MIMIC-IV dataset from our previous study [17]. This choice is guided by a comprehensive study by Bai et al [27], which demonstrated the superior performance of the TCNs in sequential data in various tests and domains, and by our own previous work, in which TCNs showed consistently strong or superior performance compared with gated recurrent units (GRUs) and attention-based architectures across multiple public ICU datasets [17]. The TCN architecture uses causal convolutions with stride 1, padding to preserve sequence length, and exponentially increasing dilation to expand the receptive field. Each convolution is followed by a Chomp1d operation to enforce causality, Leaky ReLU

XSL•FO
RenderX

activation, dropout, and residual connections. Our models were trained to predict, for each hourly bin, the individual risk of sepsis onset in the next 6 hours. We used a 10-times repeated random split scheme for hyperparameter tuning, resulting in 10 trained models for the hyperparameter set with the lowest average validation loss. For each repetition in this scheme, 20% of the data were held out as a test set. The remaining data were split into training (64%) and validation (16%). For each repetition, 10 randomly drawn hyperparameter configurations were evaluated. The configuration with the lowest average validation loss across the 10 repetitions was selected to train the final models. Training was performed using a binary cross-entropy loss, where class weights were calculated based on the inverse prevalence of the positive class in the training data, an Adam optimizer for a maximum of 1000 epochs, and early stopping with a patience of 10 epochs based on validation loss. The randomized hyperparameter search was conducted on learning rate (sampled log-uniform from exp(–10) to exp(–3)), weight decay (0, 1e–7, 1e–6, …, 1), dropout probability (0.3, 0.4, …, 0.7), batch size (128, 256, and 512), hidden dimension (32, 64, and 128), number of layers (1, …, 10), and kernel size (2, …, 6). A detailed description of the architecture, hyperparameter search, and validation procedure is provided in our previous publication [17]. To investigate a possible dependence of model architecture on the results, we also trained GRU and attention-based models on the MIMIC-IV dataset for comparison. Model training and evaluation followed the same procedures as described in our previous publication [17].

To enhance model performance, binary missingness indicators were added for the 48 time-varying and 4 static features used as inputs. This representation has been shown to capture clinically meaningful information about testing behavior but may also encode site-specific practice patterns [22,28]. To mitigate overfitting, models were trained with regularization (weight decay) and early stopping. The final models, along with their tuned hyperparameters, were then applied to the BerlinICU dataset without any further fine-tuning or retraining.

For establishing baselines in our analyses, we used logistic regression models as implemented in scikit-learn with default settings, corresponding to L2-regularized logistic regression, without hyperparameter tuning. These were trained following a 5-fold cross-validation scheme in which 80% of the data were used for training and 20% for testing. To better capture the temporal dynamics, we additionally engineered temporal features for the 48 time-varying features. These temporal features represent the trajectory of the measurements over time and include the minimum, maximum, mean, median, and variance of the last 4, 8, and 16 hours for the 48 laboratory and vital features, totaling 772 features (48 time-varying features from the hourly bins, their 720 temporal features, and 4 static features). Although this results in a large number of features relative to the number of events, these features are highly correlated temporal summaries, and penalization reduces the effective model complexity. To account for the class imbalance, we used class weights inversely proportional to class frequencies.

Training and evaluation were performed using Python 3.10.8, PyTorch 1.12.1, and scikit-learn 1.1.3.

Model Evaluation

To focus on evaluating the real-world applicability of the models in a clinical setting, we implemented 3 different evaluation strategies that scrutinized different aspects of the models' performance—fixed horizon evaluation, peak score evaluation, and continuous evaluation. Each strategy reflects a different assumption about how a model is applied in practice, and thus, yields performance metrics with different interpretations (Figure 2).

XSL•FO
RenderX

Figure 2. Overview of the evaluation strategies. For the fixed horizon evaluation, a prediction horizon was defined, where predictions were made for a time point that preceded the onset (or control onset), a predefined number of hours (6 hours in this example). Only the data from admission up to this time point was used to make the prediction (green rectangle). Labels were assigned as "positive" (marked as green "1") for patients with sepsis or "negative" (marked as green "0") for control patients. In the peak score evaluation (purple rectangle), the maximum prediction score across all time points before onset was determined. Patients with sepsis were labeled "positive," while control patients were labeled "negative" (purple "1" or "0" label). In the continuous evaluation (brown rectangle), for patients with sepsis, timepoints within a specific time window before the onset (6 hours in this example) were labeled "positive" (brown "1"s), while earlier time points—and all timepoints for control patients—were labeled "negative" (light brown "0"s). All timepoints before the onset (or control onset) were considered for evaluation. For patients with sepsis (top), the onset was the actual sepsis onset. For control patients (bottom), the onset was defined either as a hypothetical control onset matched to the onset time of a patient with sepsis ("matched onset") or as the last available data point, either at discharge or at 7 days after admission, whichever occurred first. The rectangular frames represent the data that was ultimately considered for each evaluation strategy. Shown here are the default settings for each evaluation strategy: the fixed horizon and peak score evaluations use matched sepsis onset times, while the continuous evaluation is depicted without onset matching. Note that the underlying patient data are the same across all panels; only the labeling and selected time points differ according to the respective evaluation strategy.

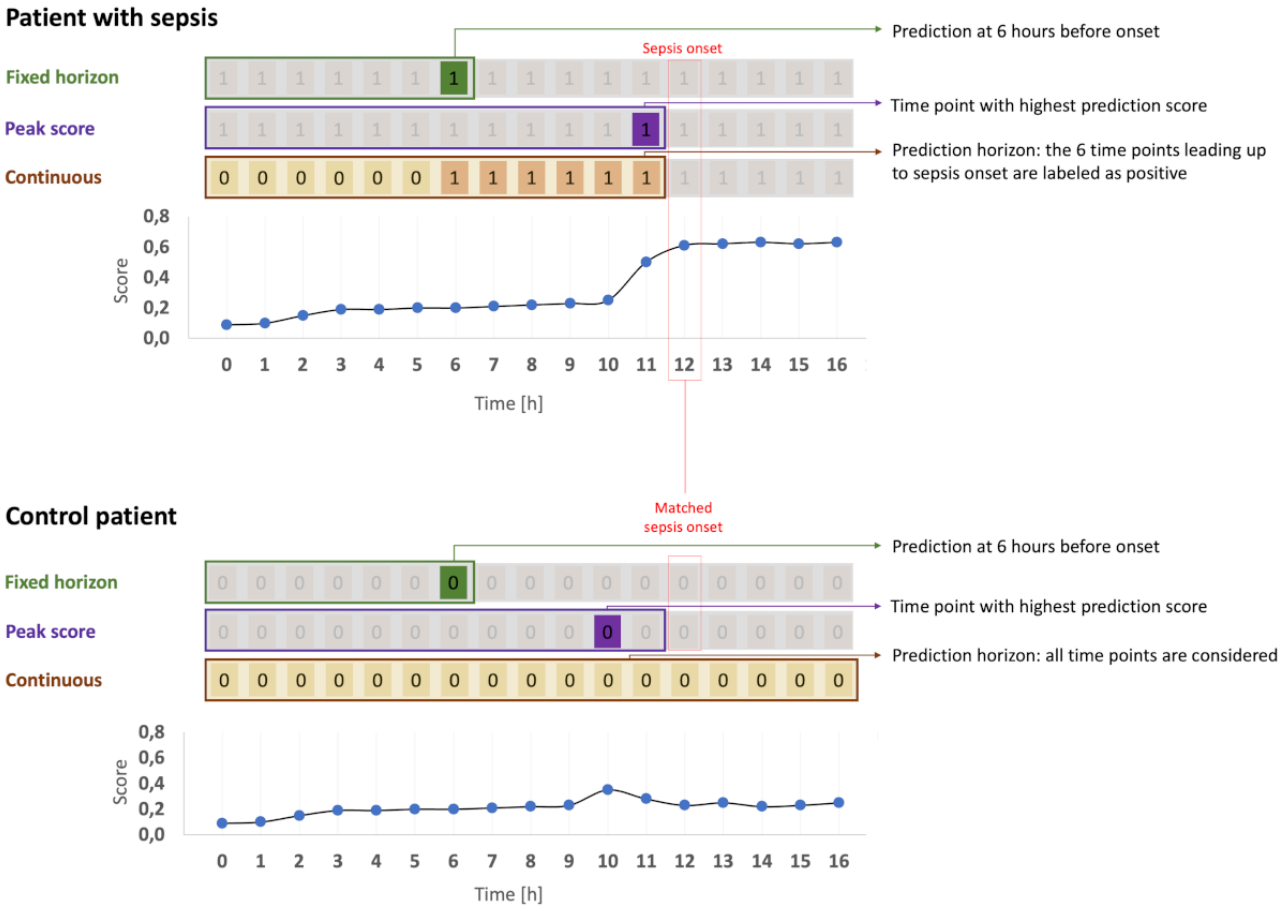

Onset Matching

A key aspect of the evaluation strategies is the handling of time series data for patients without sepsis (controls). While patients with sepsis (cases) have a clearly defined sepsis onset that marks the end point of their time series for analysis, patients without sepsis lack such a reference point. Therefore, we used 2 commonly used approaches to define the end point of their time series for analysis. First, no onset matching, in which the hypothetical onset for control patients is defined as the last available data point, either at discharge or at 7 days after admission, whichever occurred first. Second, onset matching, in which onsets of control patients are aligned with those of patients with sepsis by randomly pairing each control patient with a patient with sepsis whose time to sepsis onset does not exceed that of the control patient. The time from admission to sepsis onset in the matched patient with sepsis is then used as the hypothetical onset time for the control patient [16,29,30]. This approach preserves alignment to admission for both groups and ensures that the distribution of evaluation time points is comparable between patients with sepsis and patients without sepsis, thereby reducing bias due to differences in lengths of stay.

Fixed Horizon Evaluation

This approach evaluates a model's ability to predict the onset of sepsis at a predefined time point before the actual onset—the so-called prediction horizon (h)—using all available data up to that time [13,19]. Varying the length of the prediction horizon allows for assessing the earliness of the model's predictive ability.

The predicted score y_{pred}, which is based on all available data for the patient up to h hours before the sepsis onset for patients with sepsis or up to a hypothetical onset for control patients (t_{onset}), and the true label y_{true} are given by:

XSL•FO
RenderX

where $\boldsymbol{X}_{1:t}$ is the patient data (feature matrix) from time point 1 to t, and f is the model that outputs a score correlated to the probability of a sepsis onset.

Peak Score Evaluation

Peak score evaluation is based on the assumption that a prediction model in a clinical setting operates with a fixed threshold, and once any prediction score surpasses this threshold, an alarm for that particular patient is triggered [16,20]. To capture this behavior, the maximum model prediction score from ICU admission to the onset of sepsis or control offset is determined, resulting in a single score-label pair for each patient for performance score calculation. Here, the prediction score and true label are given by:

For horizon-dependent analysis, we used the peak score within the defined horizon instead of across all time points of each patient.

Continuous Evaluation

This approach evaluates a series of predictions made continuously for each patient over time [13]. Specifically, predictions for all time points before sepsis onset (or control onset) are considered. The resulting performance metrics from this approach represent the model's predictive efficiency across all time points, simulating its use for a randomly selected patient at a random time point.

Using this approach included first establishing a specific time horizon (eg, 6 h) to assess the model's ability to predict the sepsis onset at any given time point within that timeframe. For patients with sepsis, a positive label was assigned to all time points within the horizon, while all other time points were labeled as negative. For control patients, all time points were labeled as negative. Importantly, the time horizon only determines which time points before onset are labeled as positive, but does not restrict the range of time points included in the evaluation. We included the prediction scores from each hourly interval up to the onset of sepsis in patients with sepsis or the hypothetical control onset in control patients (t_{onset}), excluding the prediction at the actual onset (or control onset in case of onset matching). The prediction score and true label are given by:

This procedure generated a sequence of hourly score-label pairs for each patient:

To accommodate variations in the length of these sequences among patients, we applied inverse frequency weighting to the samples when computing performance metrics.

Performance Metrics

For all strategies, we report the AUROC. We chose AUROC because, unlike other commonly applied metrics such as AUPRC, accuracy, precision or PPV, NPV, or F score, the AUROC is independent of sepsis prevalence. This is essential for our study, since sepsis prevalence varies not only between the 2 datasets (MIMIC-IV vs BerlinICU) but also within a dataset across evaluation strategies; for example, in continuous evaluation, the proportion of positive time points changes with each prediction horizon. Using a prevalence-sensitive metric under these conditions would confound discrimination performance with shifts in prevalence and, therefore, blur the effect we wish to study. AUROC instead measures the probability that the model ranks a randomly chosen positive instance higher than a randomly chosen negative one, providing a prevalence-robust estimate of discrimination across all decision thresholds. However, to support interpretability, we additionally report AUPRC, PPV, NPV, and sensitivity at fixed thresholds (Figures S5 and S6 in Multimedia Appendix 1).

Because AUPRC, PPV, and NPV depend on event prevalence, we harmonized the patient-level prevalence across strategies and horizons in all evaluation samples by down-sampling patients without sepsis to match the overall sepsis prevalence of the unfiltered BerlinICU cohort. Note, however, that in continuous evaluation, the prediction horizon also determines the number of time points labeled as positive. So, while patient-level prevalence is fixed across strategies, sample-level prevalence varies with horizon in the continuous evaluation approach. As a result, prevalence-sensitive metrics are difficult to compare directly in continuous evaluation across different horizons. We therefore report both raw values and "lifted" variants that are normalized by the horizon-specific prevalence, namely PPV and AUPRC divided by prevalence and recall measured at the threshold where precision equals twice the prevalence, which disentangle model performance from shifts in baseline risk. When reporting precision at a fixed recall, we estimate it from the precision-recall curve by linear interpolation at the target recall.

To estimate the variability of performance metrics, we applied bootstrap resampling (1000 samples). For each evaluation setting that included a prediction horizon, we only included those patients whose length of stay matched or exceeded the defined horizon (Table S2 in Multimedia Appendix 1).

Calibration Assessment

To assess model calibration, we computed the Brier score. Internal calibration on the MIMIC-IV dataset is reported in our previous publication [17]. For the external BerlinICU dataset, calibration was estimated using 5-fold cross-validation. In each fold, isotonic regression was fitted on 80% of the data and applied to the held-out 20% to produce calibrated predictions. All performance metrics were calculated using the uncalibrated model applied to the full BerlinICU dataset, since discrimination metrics such as AUROC, precision, or recall are unaffected by probability calibration and depend only on the ranking of predicted scores, which is not changed by calibration. Calibration analyses were performed using scikit-learn (version 1.7.0).

Clinical Utility

Although the primary objective of this study was to compare evaluation strategies, we included a clinical utility analysis with

XSL•FO
RenderX

a continuous evaluation setup, which best aligns with clinical reality. Intensive care early-warning systems generate continuously updated risk estimates that may trigger alerts throughout the patient's stay rather than single patient-level decisions. We quantified how the model would behave when deployed as an hourly alerting system by identifying a threshold corresponding to 80% sensitivity for sepsis onset and defining an alarm as any hourly prediction exceeding this threshold. Alarm rates for septic and nonseptic admissions were calculated as the mean number of alarms per patient-hour. To approximate typical alert-suppression logic, we applied a 4-hour silencing window; after an alarm was triggered for a patient, subsequent alarms for that patient were suppressed for 4 hours, and alarm rates were recomputed using only unsilenced alerts. Net benefit was evaluated using decision-curve analysis implemented with the dcurves package, where net benefit represents the proportion of true-positive alerts minus a threshold-dependent penalty for false-positive alerts. For the main clinical interpretation, we focus on the continuous evaluation with a 6-hour prediction horizon, but also report net benefits for the other strategies to illustrate how the choice of evaluation strategy influences apparent clinical usefulness.

Statistical Analysis

To assess the significance of AUROC differences between evaluation strategies, we used a paired bootstrap Wald test with 1000 bootstrap samples per strategy. For each comparison, we drew the same bootstrap sample of patients for both strategies, computed AUROC for each, and formed the per-replicate difference. The SE of the difference was estimated as the SD of the bootstrap differences. Furthermore, 2-sided *P* values were obtained from the standard normal distribution. We applied multiple comparison correction at the level of hypothesis families; for the onset-matching effect within each strategy×model×dataset there is exactly one prespecified contrast (with vs without onset), whereas for strategy comparisons, we tested the 3 related pairwise contrasts (fixed vs peak, peak vs continuous, and fixed vs continuous) within each model×dataset×onset condition. To control the probability of at least 1 false positive among these related tests, we used the Holm step-down procedure.

Results

Study Cohort

From 60,332 ICU stays in the initial dataset, after extraction, preprocessing, and filtering, the final dataset consisted of 40,132 ICU stays of 36,872 unique patients spanning 10 years (2012-2021). Furthermore, 4134 stays were septic, resulting in a sepsis prevalence of 10.3% (Figure 3, Table 1). The original MIMIC-IV training dataset consisted of 67,056 ICU stays with a sepsis prevalence of 5.6% (3730 ICU stays).

Figure 3. Flow diagram of patient eligibility criteria. ICU: intensive care unit.

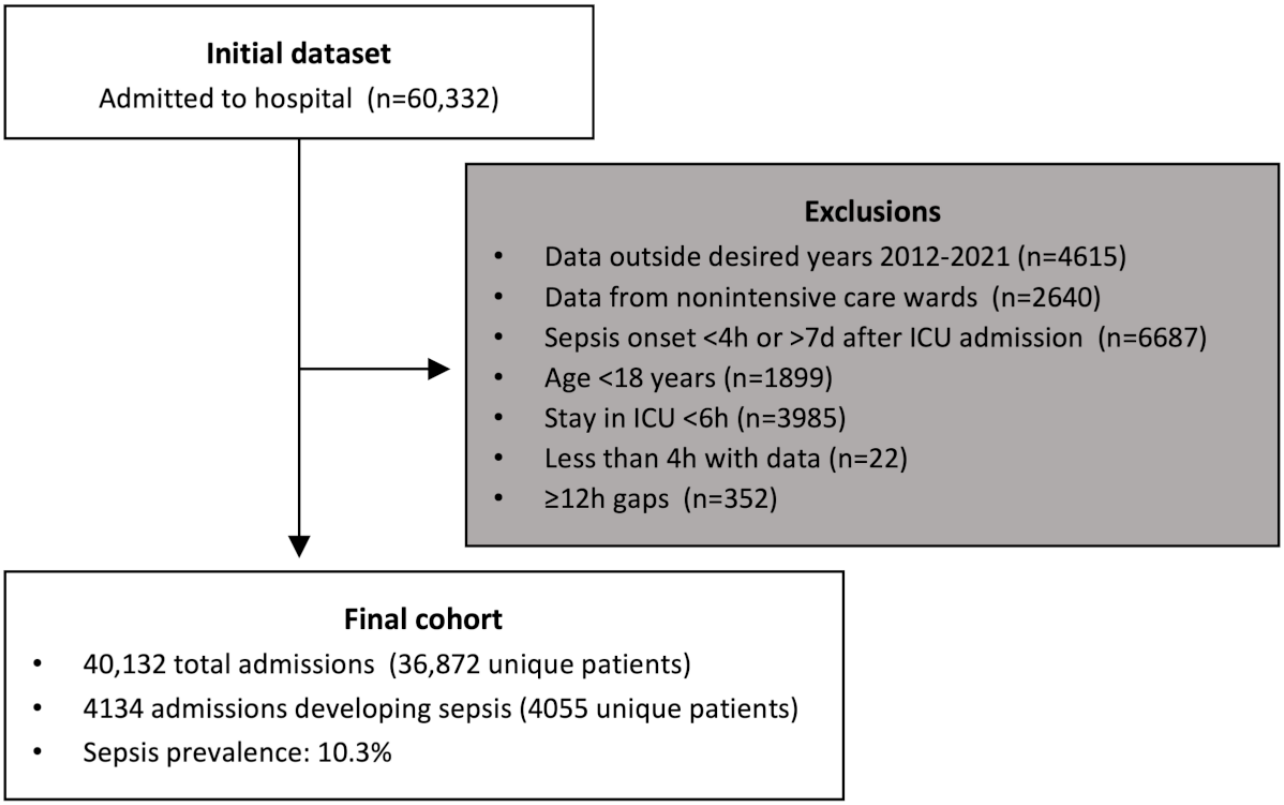

XSL•FO
RenderX

Table 1. Patient characteristics of BerlinICU and MIMIC-IV[a] for the total dataset, the sepsis cohort, and nonsepsis cohort.

Statistic	Total	Sepsis	Control
BerlinICU			
Cohort size n, (row-%)	40,132 (100)	4134 (10)	35,998 (90)
Age (y), median (IQR)	64 (50-75)	67 (55-77)	63 (50-75)
ICU[b] length of stay (d), median (IQR)	1.0 (0.7-2.1)	2.0 (1.4-3.4)	1.0 (0.7-2.0)
Sex, n (%)			
Male	22,423 (56)	2444 (59)	19,979 (56)
Female	17,709 (44)	1690 (41)	16,019 (44)
Patient count, n (%)			
Patients on vasopressors	17,117 (43)	3442 (83)	13,675 (38)
Patients ventilated	13,967 (35)	2774 (67)	11,193 (31)
Patients on any antibiotics	22,155 (55)	4134 (100)	18,021 (50)
Patients on antibiotics for ≥3 days	5351 (13)	4134 (100)	1217 (3)
MIMIC-IV			
Cohort size, n (row-%)	67,056 (100)	3730 (5.6)	63,326 (94.4)
Age (y), median (IQR)	65 (53-76)	67 (54-76)	65 (53-76)
ICU length of stay (d), median (IQR)	1.7 (1.0-2.8)	0.8 (0.6-1.5)	1.7 (1.0-2.8)
Sex, n (%)			
Male	37,184 (55)	2132 (57)	35,052 (55)
Female	29,872 (45)	1598 (43)	28,274 (45)
Patient count, n (%)			
Patients on vasopressors	8548 (13)	1230 (33)	7318 (12)
Patients ventilated	21,863 (33)	2520 (68)	19,343 (31)
Patients on any antibiotics	46,276 (69)	3730 (100)	42,546 (67)
Patients on antibiotics for ≥3 days	6214 (9)	2760 (74)	3454 (5)

MIMIC-IV: Medical Information Mart for Intensive Care IV.

ICU: intensive care unit.

Calibration

In the external BerlinICU dataset, the TCN model's predicted probabilities were reasonably calibrated, with a Brier score of 0.04-0.09 depending on the evaluation strategy (Figure S4 in Multimedia Appendix 1). Importantly, calibration does not affect the performance metrics reported in the main results, as these are independent of the predicted probability values themselves and are unaffected by monotonic transformations.

Influence of Evaluation Strategy on Model Performance Estimates

As our primary analysis, we applied the TCN models pretrained on MIMIC-IV to the MIMIC-IV test set for internal validation under a scenario aligned with the training setup—continuous evaluation using a 6-hour prediction horizon without onset matching. Under this configuration, the models showed strong internal validation performance (Figure 4A). When the same models were applied to BerlinICU for external validation under the same evaluation conditions, performance was notably reduced (Figure 4B).

XSL•FO
RenderX

Figure 4. Comparison of model performance of (A) temporal convolutional network models (top) and (B) logistic regression baseline models (bottom) for different evaluation strategies. The models were applied to the test splits of the (a) MIMIC-IV training dataset and (b) the entire German ICU dataset (BerlinICU). Model performance was evaluated using fixed horizon evaluation (6h horizon), peak score evaluation (entire patient stay), and continuous evaluation (6h horizon) without onset matching (dark bars) and with onset matching (light bars). The AUROC value inside the bar represents the mean AUROC across all bootstrap samples and models. Error bars represent 95% CIs. Significance: ***P<.001, **P<.01, *P<.05, ns: not significant (paired bootstrap Wald test with Holm-adjusted P values). AUROC: area under the receiver operating characteristic curve; MIMIC: Medical Information Mart for Intensive Care.

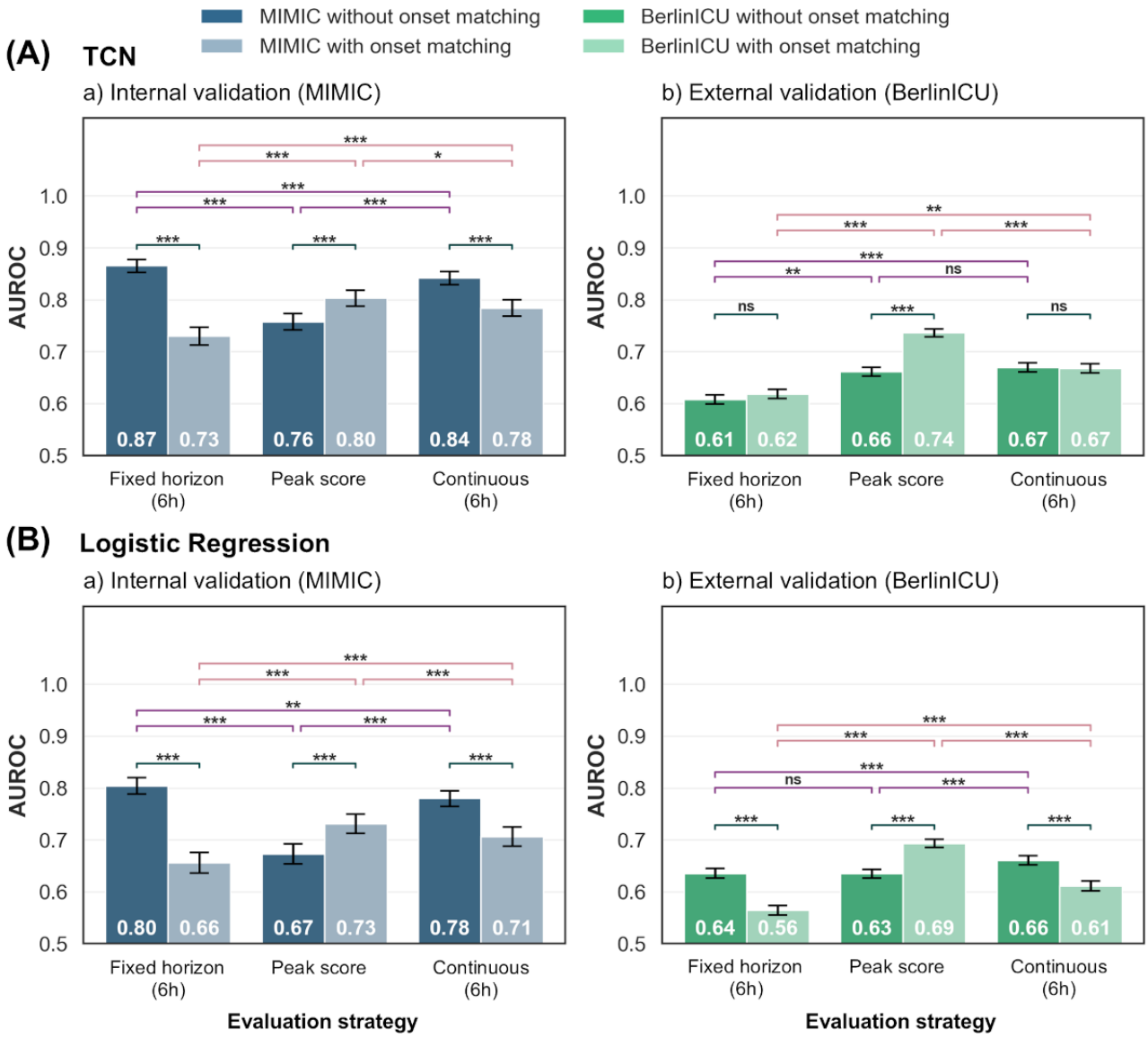

To assess the impact of different evaluation strategies, we performed fixed horizon evaluation under the same conditions (6-hour prediction horizon, no onset matching) and peak score evaluation, which does not use a prediction horizon. On the MIMIC-IV test set, fixed horizon evaluation resulted in a slightly higher AUROC, whereas peak score evaluation showed reduced performance. In external validation on BerlinICU, fixed horizon evaluation was markedly lower than continuous evaluation, while the peak score evaluation estimate was similar to that of the continuous evaluation. Across all evaluation strategies, performance estimates were consistently lower on BerlinICU compared with the MIMIC-IV test set.

Aligning the length of stay distribution of control patients to patients with sepsis through onset matching led to a notable AUROC reduction in internal validation on the MIMIC-IV test set for continuous evaluation (ΔAUROC=–0.06) and fixed horizon evaluation (ΔAUROC=–0.14), while AUROC estimates for peak score evaluation increased (ΔAUROC=0.04). On BerlinICU, onset matching had minimal impact on performance estimates during continuous evaluation or fixed horizon evaluation. In contrast, for peak score evaluation, onset matching increased AUROC estimates (ΔAUROC=0.08).

Across all evaluation strategies, TCN models consistently outperformed logistic regression baseline models. The only exception was observed in fixed horizon evaluation on BerlinICU without onset matching, where logistic regression showed a superior performance compared with TCN.

Influence of the Prediction Horizon on Model Performance Estimates

To evaluate the ability of the models to provide timely predictions, we assessed their performance across a range of

XSL•FO
RenderX

prediction horizons (Figure 5). Performance estimates generally improved with shorter prediction horizons across all evaluation strategies. During internal validation on the MIMIC-IV test set, fixed horizon evaluation and continuous evaluation showed consistent improvement, with the highest AUROC achieved at a 1-hour horizon (fixed horizon: AUROC=0.92, 95% CI 0.91-0.93; continuous: AUROC=0.87, 95% CI 0.85-0.88). Peak score plateaued around the 6-hour horizon (AUROC 0.92, 95% CI 0.91-0.93), with only marginal improvements for smaller horizons. In external validation on BerlinICU, performance also generally increased with shorter prediction windows, with few exceptions, including a plateau around the 6-hour horizon in fixed horizon and peak score evaluation with onset matching.

Figure 5. Performance metrics across prediction horizons, comparing the performance of temporal convolutional network models for an evaluation with onset matching (dashed lines) and without onset matching (solid lines). (A) Fixed horizon, (B) peak score, and (C) continuous prediction. The models were applied to the test splits of the MIMIC-IV training dataset (blue) and the entire German ICU dataset (BerlinICU; green). Shown is the AUROC based on varying prediction horizons from one hour before onset up to 100 hours before onset. Error bars represent 95% CIs. AUROC: area under the receiver operating characteristic; MIMIC: Medical Information Mart for Intensive Care; TCN: temporal convolutional network.

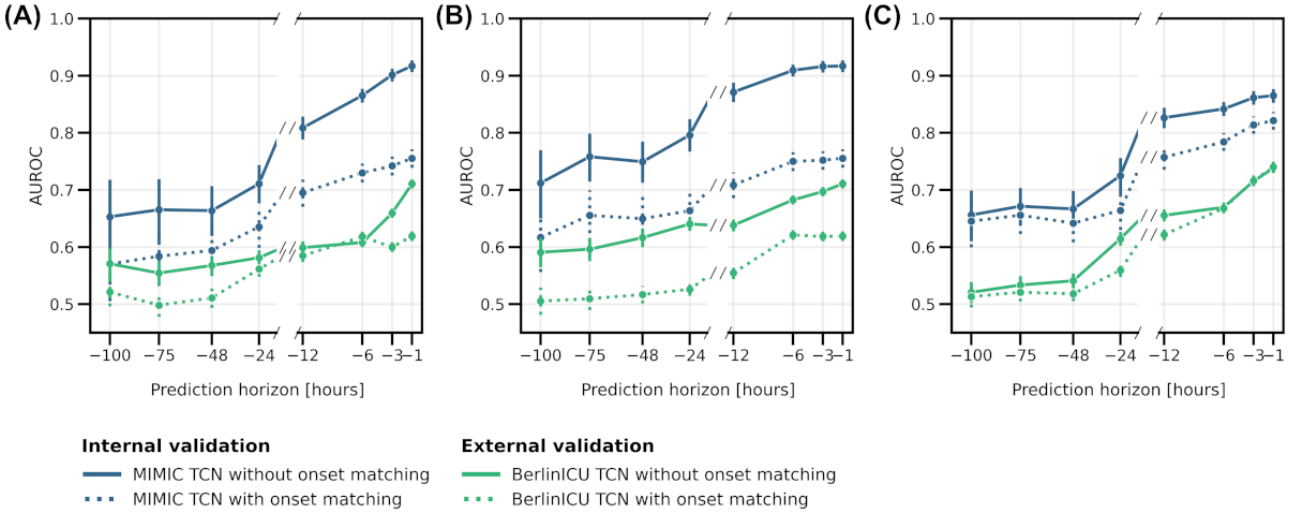

Onset matching consistently decreased performance estimates during internal validation on the MIMIC-IV test set across all evaluation strategies, albeit to different extents, with continuous evaluation showing the smallest susceptibility to the matching approach. On BerlinICU, onset matching impacted strategies differently depending on the prediction horizon. Notably, continuous evaluation remained more robust to onset matching, maintaining more consistent AUROCs compared with the other strategies across different prediction horizons, particularly for very short and long prediction horizons.

Comparing the TCN and logistic regression baseline models in internal validation on the MIMIC-IV test set, TCN models consistently outperformed logistic regression models across all evaluation strategies and prediction horizons. In contrast, during external validation on BerlinICU without onset matching, logistic regression models approached or even exceeded the performance of TCN models at certain prediction horizons (Figure S1 in Multimedia Appendix 1). Onset matching was associated with an AUROC reduction for both models; the reduction appeared larger for logistic regression than for TCN in our data.

Clinical Operational Characteristics Under Continuous Evaluation

Under the continuous evaluation with a 6-hour prediction horizon, an operating point of 80% sensitivity produced 0.70 alarms/hour in patients with sepsis and 0.25 alarms/hour in patients without sepsis in MIMIC-IV, compared with 0.74 and 0.55 alarms/hour, respectively, in BerlinICU (Figure S8 in Multimedia Appendix 1). With a 4-hour silencing period, rates decreased to 0.18 (septic) and 0.06 (nonseptic) in MIMIC-IV, and to 0.18 and 0.14 alarms/hour, respectively, in BerlinICU. Decision-curve analysis showed modest net benefit at low thresholds in MIMIC-IV and limited benefit in BerlinICU (Figure S9 in Multimedia Appendix 1).

Discussion

Principal Findings

In this study, we assessed the impact of commonly used evaluation strategies (fixed horizon, peak score, and continuous evaluation) on the performance of machine learning models for sepsis onset prediction. Our results demonstrate that the choice of the evaluation strategy significantly influences performance estimates, even though the same model is used on the same dataset.

We further addressed a gap in research involving the German health care system by aggregating an ICU dataset, called BerlinICU, from one of the largest university hospitals in Europe, Charité - Universitätsmedizin Berlin. Including more than 40,000 admissions from 8 ICU wards, this dataset is one of the largest ICU datasets in Europe, exceeding the size of the currently available European datasets—Amsterdam University Medical Center Database [31], High Time Resolution ICU Dataset [32], and Salzburg Intensive Care database [33]. To the best of our knowledge, this is one of the first studies to explore how sepsis prediction models perform on large-scale German ICU data.

XSL•FO
RenderX

Fixed Horizon Evaluation Requires Prospective Knowledge of Sepsis Onset

Fixed horizon evaluation uses data up to a predefined time point before sepsis onset—the prediction horizon—to make predictions. Performance decreased with longer prediction horizons due to increasing temporal distance from sepsis onset. Without onset matching, performance estimates strongly increased for all horizons on the MIMIC-IV test set and for most horizons for BerlinICU, with the exception of the 6-hour horizon. This performance increase when not using onset matching is most parsimoniously explained by the evaluation of control patients over their entire stay, with emphasis on the period just before discharge. During these predischarge hours, control patients are typically in a stable condition, making it easier for the model to correctly identify them as low-risk for sepsis. This effect, also noted by Futoma et al [29], shows the significant impact of onset alignment. Despite this significant influence, many studies do not comment on the implementation of case-control matching [13]. While the fixed horizon approach achieved the highest AUROC in a study on different evaluation strategies [19], we observed more nuanced performance differences compared with the other evaluation approaches, depending on the dataset, horizon, and onset matching approach. Fixed-horizon evaluation can be useful for benchmarking how far in advance a model can predict sepsis and for retrospective simulation studies, but its direct clinical interpretability is challenging, as it relies on prospective information—the time point of the sepsis onset—which is not accessible in a real-world clinical setting.

Peak Score Evaluation is Influenced by the Length of Stay Distribution and Requires Onset Matching to Avoid Bias

The peak score evaluation assesses the model's peak confidence in predicting sepsis throughout the entire ICU stay. This approach seems intuitively aligned with clinical practice, where a fixed threshold would be predetermined, and an alarm is triggered once the prediction model's output surpasses it—potentially just once per patient, with subsequent alarms suppressed [16,20]. The focus is on whether an alarm is raised at any point during the patient's stay, which is replicated by determining the maximum prediction score. However, the performance metrics determined under this approach are influenced by differences in length of stay distributions between cases and controls—longer time series for control patients increase the likelihood of higher maximum prediction scores for these patients simply due to random fluctuations, ultimately reducing the AUROC as the maximum scores for control patients approach those of patients with sepsis on average (Figure S3 in Multimedia Appendix 1). Importantly, this bias conflicts with the core interpretation of the AUROC, which measures the probability that a randomly selected true positive will have a higher predicted score than a randomly selected true negative. Our results demonstrate this behavior, showing that performance estimates without onset matching, where control cases tend to have longer lengths of stays, were significantly reduced compared with analyses with onset matching (Figure 4). To mitigate this effect, an onset matching approach that aligns the length of stay distributions between positive and negative cases is strictly necessary when applying this evaluation strategy.

The discrepancy between the peak score evaluation across all patients (Figure 4), where onset matching increases the AUROC, and horizon-dependent analysis (Figure 5), where onset matching decreases the AUROC, can be explained by the exclusion of patients with insufficient stay lengths in the latter. This exclusion alters the length of stay distribution and thereby significantly impacts performance metrics. Not filtering patients with insufficient length of stays shows a distinct interaction effect between onset matching and prediction horizon (Figure S2 in Multimedia Appendix 1), further highlighting the critical influence of length of stay distribution on the performance estimates in the peak score evaluation approach. Nevertheless, peak score evaluation has a simple clinical interpretation—at a given threshold, does the model ever raise an alarm for a patient with sepsis? When the length of stay distribution is carefully adjusted between cases and controls and the evaluation is restricted to a clinically meaningful time frame (eg, within 12 hours before sepsis onset), this strategy can provide a useful summary of whether the model is able to flag a patient at high-risk at any point during their stay.

Continuous Evaluation Without Onset Matching Reflects Clinical Practice

Continuous prediction evaluation showed increased performance with shorter prediction windows, demonstrating the model's effectiveness in using short-term, relevant data to predict sepsis onset. As the prediction window narrows and focuses more on the actual onset of sepsis, the assigned sepsis labels align more closely with the sepsis-related signals in the patient data, enhancing prediction efficiency. Interestingly, although the models were trained to predict sepsis 6 hours in advance, this specific labeling window did not yield the best performance. This is likely because these shorter windows include only time points immediately before sepsis onset, where the signal is strongest, while earlier time points, although possibly already showing signs of deterioration, are treated as negatives.

Evaluating on BerlinICU with onset matching, model performance was significantly reduced compared with the MIMIC-IV test set (ΔAUROC=–0.11 for 6 hours before onset and –0.08 for 1 hour). This reduced performance when transferring models between datasets was also evident in our previous study, where models trained on MIMIC-IV were evaluated on other European datasets using a horizon-based evaluation with a 6-hour window [17]. Even though the performance in that study declined for European datasets (High time Resolution ICU Dataset: ΔAUROC=–0.07, Amsterdam University Medical Center Database: ΔAUROC=–0.09) as well as for another US dataset (Electronic Intensive Care Unit Collaborative Research Database: ΔAUROC=–0.10), it was not as severe as observed in BerlinICU. One might suspect that this sharper decline is related to different sepsis prevalences, since BerlinICU had a higher prevalence than MIMIC-IV. However, AUROC is unaffected by prevalence [21], and the difficulty of the task, and therefore the prediction performance, is more strongly influenced by factors such as data heterogeneity, label noise, and shifting feature distributions, all of which may differ

XSL•FO
RenderX

substantially between datasets. These findings underscore the importance of addressing domain shift when deploying machine learning models across health care systems. Differences in patient populations, documentation practices, feature availability, and underlying sepsis definitions may contribute to reduced generalizability. For example, differences in clinical workflows, such as the timing and frequency of laboratory testing, documentation delays, or variability in when and how sepsis-related interventions are initiated (eg, antibiotics or fluids), may influence how early or clearly sepsis-related signals appear in the data. The use of missingness indicators is another potential factor, although previous studies have reported mixed effects on the generalizability of sepsis prediction models [22,28]. Mitigation strategies, such as dataset-specific fine-tuning or full retraining on local data, can be considered to improve cross-domain performance and support safe clinical deployment.

Without onset matching, performance in the MIMIC-IV test set was higher across horizons, and was similar or better on BerlinICU for most horizons. Again, this higher performance without onset matching most likely reflects the inclusion of time points closer to discharge for control patients, who generally have a healthier profile, making the prediction task easier. This mechanism is illustrated in Figure S10 in Multimedia Appendix 1, which shows predicted risk trajectories over time; unmatched controls display a marked decline in predicted risk approaching discharge, whereas matched controls remain at higher levels, closer to those of patients with sepsis. While onset matching may provide a focused assessment in more challenging cases, not using it better mirrors clinical practice, where patients are monitored continuously from admission to discharge. In our view, continuous evaluation without onset matching better reflects real-world clinical application, as the AUROC represents the likelihood of a positive case being ranked higher than a negative one, including time points closer to discharge. Accordingly, we treat continuous evaluation without onset matching as the most clinically relevant interpretation of model performance, while the other evaluation strategies are used to demonstrate how alternative designs can substantially alter reported results.

Although this approach requires predetermining a time window, which indicates how far in the future the onset may be predicted, we consider continuous evaluation the better choice for 2 reasons. First, defining a time window—ideally based on pathophysiological plausibility—ensures that only alarms occurring within a clinically meaningful period before sepsis onset are considered. This distinction is crucial, as some published models show high performance days to weeks before the actual event [18,34,35], likely detecting broader, correlated health-related features rather than sepsis specifically. This underlines the importance of thoroughly defining and testing what these models are predicting to ensure their practical utility in clinical settings. Second, varying the prediction window and reporting the performance across all these windows allows for a more nuanced understanding of the models' ability to predict event onset in advance. This analysis can be performed efficiently without retraining, since the underlying model predictions remain fixed and only the labeling of time points is varied to reflect different prediction windows. However, training separate models for different forecast horizons and investigating the effect on the performance using continuous evaluation is an interesting, but computationally more intensive approach that could be explored in future studies.

It is important to note that longer prediction horizons (eg, beyond 12 h) are unlikely to be clinically meaningful. From a clinical standpoint, alarms triggered more than half a day before sepsis onset may not prompt specific interventions and could increase alert fatigue. Nevertheless, we included prediction horizons up to 100 hours as a form of negative control to investigate whether model performance persists far in advance of sepsis. If so, this could indicate that the model is capturing nonspecific markers of deterioration or health status rather than sepsis-specific signals. In line with this, we observed that the model's performance on BerlinICU dropped toward chance level at long horizons, while AUROC remained elevated on MIMIC-IV, possibly suggesting that the model may overfit to nonspecific or site-specific patterns in the training data.

Clinical Implications of the Choice of the Evaluation Strategy

The evaluation strategy directly determines how performance metrics translate to clinical practice. Metrics such as PPV, when derived from peak-score evaluation, are patient-level; they yield a single value per patient that answers, "Did this patient ever cross the alert threshold?" In routine deployment, however, models are usually executed repeatedly (eg, hourly), and clinicians need alert-level information, that is, the probability that a specific alarm corresponds to true sepsis within a clinically relevant time window. Using patient-level metrics to select models for clinical deployment can therefore overestimate real-world performance and contribute to alarm fatigue. Likewise, fixed-horizon evaluation reports performance at one predefined time point per patient; it does not describe how the model behaves during continuous monitoring. Importantly, it requires prospective knowledge of the exact onset time, which is inherently unavailable in clinical reality. For continuous evaluation, the horizon is adjustable; PPV at h hours answers the question "If the model raises an alert now, what is the probability that sepsis will occur within the next h hours?" Different choices of the horizon yield different PPVs, and only the PPV at the clinically intended horizon is directly relevant for deployment. Thus, to ensure clinical relevance, evaluation strategies and metrics should be aligned with how the model will be used in practice.

Our analyses focused primarily on TCN models. However, the effects of the evaluation strategy are not specific to this architecture. Because evaluation is applied post hoc to model outputs, the principles are, in theory, model-agnostic. Consistent with this expectation, we observed the same qualitative influence of evaluation strategy on performance when applying it to GRU- and transformer-based models (Figure S7 in Multimedia Appendix 1). This supports the interpretation that our findings reflect general properties of evaluation design rather than model-specific behavior. Likewise, while our empirical analyses were conducted on 2 datasets, the observed differences between evaluation strategies arise from the way labels are constructed

XSL•FO
RenderX

and metrics are defined, and are therefore expected to hold independently of the specific dataset.

Clinical Utility Considerations

While the primary objective of this study is to compare evaluation strategies, it is helpful to illustrate how a model of this type would function within a real-time ICU workflow. In practice, such early-warning systems would generate repeated risk estimates and trigger alerts that prompt targeted clinical assessment, such as reviewing chart data or performing additional tests, rather than automated interventions. Thresholds are typically chosen to balance timely detection with acceptable alarm burden, often combined with silencing periods to avoid repeated notifications. Since clinical use is structured around continuous monitoring and repeated predictions, continuous evaluation is the strategy that most closely reflects real-world deployment.

The operational characteristics observed in our analyses provide an indication of the potential alarm burden such a system might generate. At 80% sensitivity, the model produced alarm rates that are plausible for an ICU early-warning context, although their clinical acceptability is uncertain and would require prospective evaluation. A 4-hour silencing period substantially reduced repeated alerts, highlighting the importance of workflow-oriented design choices. Decision-curve analysis further showed modest net benefit at low thresholds in MIMIC-IV and limited benefit in BerlinICU.

Limitations

Our study has several limitations. First, we did not have access to microbiological data, requiring an adjustment of the Sepsis-3 definition. While this approach has been used in previous studies using hospital datasets without microbiology records [16,17], it may reduce the specificity of outcome labeling and limit comparability with studies that apply the complete Sepsis-3 criteria, including microbiology data. Second, we used models trained for a 6-hour horizon and applied them across various prediction horizons to predict sepsis onset. While these models may not perform optimally outside their original specifications, the inability to know the exact duration until a potential sepsis onset at any given timepoint in clinical practice necessitates using a flexible approach. Thus, by testing these models beyond their intended parameters, we can evaluate their robustness and generalizability to different clinical scenarios and timelines. We acknowledge, however, that this does not replace training and comparing separate models optimized for different prediction horizons.

Comparison With Previous Work

Our work combines the idea of generalizability across countries [16,17] and suitable metric calculation concepts [16,36]. Our study builds on these previous concepts, refining and extending them as needed. While Moor et al [16] and Rockenschaub et al [17] examined the influence of various algorithms and training sets from different countries, our work focuses on different evaluation strategies. We applied sepsis onset matching [29,36] and peak score evaluation [16], combining approaches that were presented separately. As the peak score evaluation is only suitable for retrospective analysis, we also applied a continuous evaluation strategy [13]. Moreover, there are only a few studies that have applied models to German data [18,37], none of which have a dedicated focus on comparing performances in different evaluation strategies. While some models have reported high AUROC values on German data, care is needed when interpreting or comparing such metrics across studies, as the AUROC is not a function of the model alone, but depends heavily on the prediction task and the underlying dataset. Differences in prediction target, patient population, and labeling criteria can have substantial effects on reported discrimination. For example, distinguishing patients with sepsis in the ICU from a general hospital population is inherently easier than distinguishing sepsis from nonsepsis within an already critically ill ICU cohort, due to the greater physiological overlap in the latter case. Moreover, as we demonstrate in this study, the choice of evaluation strategy (eg, fixed prediction horizon, onset matching, or continuous evaluation) can meaningfully influence reported performance.

Several studies have applied sepsis prediction models prospectively to US data [14,15,38], but they often lack details on how the models were trained and how the data were preprocessed. This lack of transparency is a common issue, as noted in 2 systematic reviews of sepsis prediction models [12,13]. Accordingly, common checklists for reporting prediction models include items for external validation in different countries and time periods [39,40].

Conclusions

Our study highlights the critical need to choose an evaluation approach that is aligned with the intended interpretation of the resulting performance metric, as different approaches yield markedly different performance estimates despite using the same model on the same dataset. Importantly, for the fixed horizon and peak score evaluation approaches, a carefully devised onset matching strategy is crucial to avoid skewed results that may not reflect the true model performance. In our view, the continuous evaluation approach better reflects clinical reality. Despite requiring specifying the horizon time window, it offers insights into the model's effectiveness in a clinical setting.

Acknowledgments

We thank Fabian Schreiber and Falk Meyer-Eschenbach for their support in data extraction. This research was conducted using data obtained from Charité's Health Data Platform (HDP). The authors also acknowledge the Scientific Computing of the IT Division at the Charité - Universitätsmedizin Berlin for providing computational resources that have contributed to the research

XSL•FO
RenderX

results reported in this paper. OpenAI's Model GPT5 was used to polish the language of the manuscript, ensuring clarity and fluency, as well as to shorten paragraphs where applicable, assisting in effective communication of the study's contributions.

All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article.

Data Availability

The MIMIC-IV (Medical Information Mart for Intensive Care IV) dataset is publicly available for research purposes through PhysioNet upon completion of a data use agreement and required training [41,42]. The BerlinICU dataset used in this study cannot currently be shared outside the institution due to data protection regulations and institutional policies.
Preprocessing and training code for the neural network models has been published by Rockenschaub et al [17], with the corresponding GitHub repository available at GitHub [43]. All analysis code specific to this study are also available [44].

Funding

This work received funding from the Jürgen Manchot Stiftung.
All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article.

Authors' Contributions

DKD: Methodology, software, validation, formal analysis, investigation, data curation, writing (original draft), writing (review & editing), visualization, project administration.
PR: Methodology, software, validation, data curation, writing (review & editing).
SDB: Data curation, writing (review & editing).
OK: Writing (review & editing).
HDV: Writing (review & editing), supervision.
FB: Writing (review & editing), supervision, resources.
FvD: Conceptualization, methodology, writing (review & editing), supervision.
GL: Conceptualization, methodology, investigation, writing (original draft), writing (review & editing), visualization, supervision, project administration.

Conflicts of Interest

SDB is a salaried employee at Pfizer Pharma GmbH and a visiting researcher at the institute of Medical Informatics, Charité — Universitätsmedizin Berlin. Pfizer had no involvement in the conception, design, execution, or interpretation of the study, nor in the preparation or decision to submit the manuscript for publication. The authors declare no conflicts of interest regarding this study.

Multimedia Appendix 1
Additional tables and figures.
[PDF File (Adobe PDF File), 1839 KB - jmir_v28i1e72083_app1.pdf]

References

1. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 2016;315(8):801-810 [FREE Full text] [doi: 10.1001/jama.2016.0287] [Medline: 26903338]
2. Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al. Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the global burden of disease study. Lancet 2020;395(10219):200-211 [FREE Full text] [doi: 10.1016/S0140-6736(19)32989-7] [Medline: 31954465]
3. Fleischmann C, Thomas-Rueddel DO, Hartmann M, Hartog CS, Welte T, Heublein S, et al. Hospital incidence and mortality rates of sepsis. Dtsch Arztebl Int 2016;113(10):159-166 [FREE Full text] [doi: 10.3238/arztebl.2016.0159] [Medline: 27010950]
4. Torio CM, Moore BJ. National inpatient hospital Costs: The most expensive conditions by payer, 2013. Healthc Cost Util Proj HCUP Stat Briefs Rockville (MD): Agency for Healthcare Research and Quality (US) 2006. [Medline: 27359025]
5. Winters BD, Eberlein M, Leung J, Needham DM, Pronovost PJ, Sevransky JE. Long-term mortality and quality of life in sepsis: a systematic review. Crit Care Med 2010;38(5):1276-1283. [doi: 10.1097/CCM.0b013e3181d8cc1d] [Medline: 20308885]

XSL•FO
RenderX

6. Pruinelli L, Westra BL, Yadav P, Hoff A, Steinbach M, Kumar V, et al. Delay within the 3-hour surviving sepsis campaign guideline on mortality for patients with severe sepsis and septic shock. Crit Care Med 2018;46(4):500-505 [FREE Full text] [doi: 10.1097/CCM.0000000000002949] [Medline: 29298189]
7. Ferrer R, Martin-Loeches I, Phillips G, Osborn TM, Townsend S, Dellinger RP, et al. Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program. Crit Care Med 2014;42(8):1749-1755. [doi: 10.1097/CCM.0000000000000330] [Medline: 24717459]
8. Seymour CW, Gesten F, Prescott HC, Friedrich ME, Iwashyna TJ, Phillips GS, et al. Time to treatment and mortality during mandated emergency care for sepsis. N Engl J Med 2017;376(23):2235-2244 [FREE Full text] [doi: 10.1056/NEJMoa1703058] [Medline: 28528569]
9. Evans L, Rhodes A, Alhazzani W, Antonelli M, Coopersmith CM, French C, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. Intensive Care Med 2021;47(11):1181-1247 [FREE Full text] [doi: 10.1007/s00134-021-06506-y] [Medline: 34599691]
10. Evans L, Rhodes A, Alhazzani W, Antonelli M, Coopersmith CM, French C, et al. Executive summary: Surviving sepsis campaign: International guidelines for the management of sepsis and septic shock 2021. Crit Care Med 2021;49(11):1974-1982. [doi: 10.1097/CCM.0000000000005357] [Medline: 34643578]
11. Levy MM, Evans LE, Rhodes A. The surviving sepsis campaign Bundle: 2018 update. Intensive Care Med 2018;44(6):925-928. [doi: 10.1007/s00134-018-5085-0]
12. Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 2020;46(3):383-400 [FREE Full text] [doi: 10.1007/s00134-019-05872-y] [Medline: 31965266]
13. Moor M, Rieck B, Horn M, Jutzeler CR, Borgwardt K. Early prediction of sepsis in the ICU using machine learning: A systematic review. Front Med (Lausanne) 2021;8:607952 [FREE Full text] [doi: 10.3389/fmed.2021.607952] [Medline: 34124082]
14. Burdick H, Pino E, Gabel-Comeau D, McCoy A, Gu C, Roberts J, et al. Effect of a sepsis prediction algorithm on patient mortality, length of stay and readmission: a prospective multicentre clinical outcomes evaluation of real-world patient data from US hospitals. BMJ Health Care Inform 2020;27(1):e100109 [FREE Full text] [doi: 10.1136/bmjhci-2019-100109] [Medline: 32354696]
15. Adams R, Henry KE, Sridharan A, Soleimani H, Zhan A, Rawat N, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat Med 2022;28(7):1455-1460. [doi: 10.1038/s41591-022-01894-0] [Medline: 35864252]
16. Moor M, Bennett N, Plečko D, Horn M, Rieck B, Meinshausen N, et al. Predicting sepsis using deep learning across international sites: a retrospective development and validation study. EClinicalMedicine 2023;62:102124 [FREE Full text] [doi: 10.1016/j.eclinm.2023.102124] [Medline: 37588623]
17. Rockenschaub P, Hilbert A, Kossen T, Elbers P, von Dincklage F, Madai VI, et al. The impact of multi-institution datasets on the generalizability of machine learning prediction models in the ICU. Crit Care Med 2024;52(11):1710-1721. [doi: 10.1097/CCM.0000000000006359] [Medline: 38958568]
18. Steinbach D, Ahrens PC, Schmidt M, Federbusch M, Heuft L, Lübbert C, et al. Applying machine learning to blood count data predicts sepsis with ICU admission. Clin Chem 2024;70(3):506-515. [doi: 10.1093/clinchem/hvae001] [Medline: 38431275]
19. Lauritsen SM, Thiesson B, Jørgensen MJ, Riis AH, Espelund US, Weile JB, et al. The framing of machine learning risk prediction models illustrated by evaluation of sepsis in general wards. NPJ Digit Med 2021;4(1):158 [FREE Full text] [doi: 10.1038/s41746-021-00529-x] [Medline: 34782696]
20. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med 2021;181(8):1065-1070 [FREE Full text] [doi: 10.1001/jamainternmed.2021.2626] [Medline: 34152373]
21. McDermott MBA, Zhang H, Hansen LH, Angelotti G, Gallifant J. A closer look at AUROC and AUPRC under class imbalance. arXiv 2025. [doi: 10.52202/079017-1400]
22. Reyna MA, Josef CS, Jeter R, Shashikumar SP, Westover MB, Nemati S, et al. Early prediction of sepsis from clinical Data: The PhysioNet/computing in cardiology challenge 2019. Crit Care Med 2020;48(2):210-217 [FREE Full text] [doi: 10.1097/CCM.0000000000004145] [Medline: 31939789]
23. Huxford C, Rafiei A, Nguyen V, Wiens M, Ansermino J, Kissoon N, Pediatric Sepsis Data CoLaboratory. The 2024 pediatric sepsis Challenge: Predicting in-hospital mortality in children with suspected sepsis in Uganda. Pediatr Crit Care Med 2024;25(11):1047-1050. [doi: 10.1097/PCC.0000000000003556] [Medline: 38904442]
24. Bennett N, Plečko D, Ukor IF, Meinshausen N, Bühlmann P. ricu: R's interface to intensive care data. Gigascience 2022;12:giad041 [FREE Full text] [doi: 10.1093/gigascience/giad041] [Medline: 37318234]
25. Rockenschaub P, Akay EM, Carlisle BG, Hilbert A, Wendland J, Meyer-Eschenbach F, et al. External validation of AI-based scoring systems in the ICU: a systematic review and meta-analysis. BMC Med Inform Decis Mak 2025;25(1):5 [FREE Full text] [doi: 10.1186/s12911-024-02830-7] [Medline: 39762808]

XSL•FO
RenderX

26. Moor M, Bennett N, Plečko D, Horn M, Rieck B, Meinshausen N, et al. Predicting sepsis using deep learning across international sites: A retrospective development and validation study. EClinicalMedicine 2023;62:102124 [FREE Full text] [doi: 10.1016/j.eclinm.2023.102124] [Medline: 37588623]
27. Bai S, Kolter JZ, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv 2018. [doi: 10.48550/arXiv.1803.01271]
28. Singh J, Sato M, Ohkuma T. On missingness features in machine learning models for critical Care: Observational study. JMIR Med Inform 2021;9(12):e25022 [FREE Full text] [doi: 10.2196/25022] [Medline: 34889756]
29. Futoma J, Hariharan S, Heller K, Sendak M, Brajer N, Clement M, et al. An improved multi-output gaussian process RNN with real-time validation for early sepsis detection. Proc 2nd Mach Learn Healthc Conf PMLR 2017:243-254 [FREE Full text]
30. Lichtner G, Balzer F, Haufe S, Giesa N, Schiefenhövel F, Schmieding M, et al. Predicting lethal courses in critically ill COVID-19 patients using a machine learning model trained on patients with non-COVID-19 viral pneumonia. Sci Rep 2021;11(1):13205 [FREE Full text] [doi: 10.1038/s41598-021-92475-7] [Medline: 34168198]
31. Thoral PJ, Peppink JM, Driessen RH, Sijbrands EJG, Kompanje EJO, Kaplan L, Amsterdam University Medical Centers Database (AmsterdamUMCdb) Collaboratorsthe SCCM/ESICM Joint Data Science Task Force. Sharing ICU patient data responsibly under the society of critical care medicine/european society of intensive care medicine joint data science collaboration: The Amsterdam university medical centers database (AmsterdamUMCdb) example. Crit Care Med 2021;49(6):e563-e577 [FREE Full text] [doi: 10.1097/CCM.0000000000004916] [Medline: 33625129]
32. Hyland SL, Faltys M, Hüser M, Lyu X, Gumbsch T, Esteban C, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. Nat Med 2020;26(3):364-373. [doi: 10.1038/s41591-020-0789-4] [Medline: 32152583]
33. Rodemund N, Wernly B, Jung C, Cozowicz C, Koköfer A. Harnessing big data in critical Care: Exploring a new European dataset. Sci Data 2024;11(1):320 [FREE Full text] [doi: 10.1038/s41597-024-03164-9] [Medline: 38548745]
34. Barton C, Chettipally U, Zhou Y, Jiang Z, Lynn-Palevsky A, Le S, et al. Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. Comput Biol Med 2019;109:79-84 [FREE Full text] [doi: 10.1016/j.compbiomed.2019.04.027] [Medline: 31035074]
35. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med 2015;7(299):299ra122. [doi: 10.1126/scitranslmed.aab3719] [Medline: 26246167]
36. Moor M, Horn M, Rieck B, Roqueiro D, Borgwardt K. Early recognition of sepsis with Gaussian process temporal convolutional networks and dynamic time warping. Mach Learn Healthc Conf PMLR 2019:2-26 [FREE Full text]
37. Schamoni S, Lindner HA, Schneider-Lindner V, Thiel M, Riezler S. Leveraging implicit expert knowledge for non-circular machine learning in sepsis prediction. Artif Intell Med 2019;100:101725. [doi: 10.1016/j.artmed.2019.101725] [Medline: 31607345]
38. Henry KE, Adams R, Parent C, Soleimani H, Sridharan A, Johnson L, et al. Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. Nat Med 2022;28(7):1447-1454. [doi: 10.1038/s41591-022-01895-z] [Medline: 35864251]
39. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, PROGRESS Group. Prognosis research strategy (PROGRESS) 3: prognostic model research. PLoS Med 2013;10(2):e1001381 [FREE Full text] [doi: 10.1371/journal.pmed.1001381] [Medline: 23393430]
40. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. Lancet 2007;370(9596):1453-1457 [FREE Full text] [doi: 10.1016/S0140-6736(07)61602-X] [Medline: 18064739]
41. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCH, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000;101(23):E215-E220. [doi: 10.1161/01.cir.101.23.e215] [Medline: 10851218]
42. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023;10(1):1 [FREE Full text] [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]
43. Generalisability of deep learning-based early warning in the intensive care unit. GitHub. URL: https://github.com/prockenschaub/icuDG [accessed 2026-02-15]
44. Evaluation of Machine Learning Models for Sepsis Prediction. GitHub. URL: https://github.com/medinf-ai/sepsis-eval-study/ [accessed 2026-02-15]

Abbreviations

AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
ICU: intensive care unit
GRU: gated recurrent unit
MIMIC-IV: Medical Information Mart for Intensive Care IV
PPV: positive predictive value

XSL•FO
RenderX

NPV: negative predictive value
SI: suspicion of infection
SOFA: Sequential Organ Failure Assessment
TCN: temporal convolutional network

Edited by J Sarvestan; submitted 18.Feb.2025; peer-reviewed by M Görges, H Shuai-Jing, R Heyer; comments to author 14.Jul.2025; revised version received 10.Dec.2025; accepted 10.Dec.2025; published 24.Mar.2026.

Please cite as:
Do DK, Rockenschaub P, Boie SD, Kumpf O, Volk HD, Balzer F, von Dincklage F, Lichtner G
The Impact of Evaluation Strategy on Sepsis Prediction Model Performance Metrics in Intensive Care Data: Retrospective Cohort Study
J Med Internet Res 2026;28:e72083
URL: https://www.jmir.org/2026/1/e72083
doi:10.2196/72083
PMID:41874553

©Dang-Khoa Do, Patrick Rockenschaub, Sebastian Daniel Boie, Oliver Kumpf, Hans-Dieter Volk, Felix Balzer, Falk von Dincklage, Gregor Lichtner. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Sleep and Activity Patterns in Depression From Wearable Data: Unsupervised Clustering Study

Carolin Oetzmann[1], PhD; Yuezhou Zhang[2], PhD; Nicholas Cummins[2], PhD; Ewan Carr[2], PhD; Faith Matcham[3], PhD; Sara Siddi[4,5], PhD; Femke Lamers[6,7], PhD; Daniel Leightley[8], PhD; Katie M White[1], PhD; Amos A Folarin[2], PhD; Peter Annas[9], PhD; Josep Maria Haro[4,5], MD; Brenda WJH Penninx[6,7], PhD; Srinivasan Vairavan[10], PhD; Til Wykes[11,12], PhD; Richard JB Dobson[2], PhD; Vaibhav A Narayan[13], PhD; Matthew Hotopf[1,12], PhD

[1]Department of Psychological Medicine, Institute of Psychiatry, Psychology & Neuroscience, King's College London, 16 De Crespigny Park, London, United Kingdom

[10]Janssen Research and Development, LLC, Titusville, NJ, United States

[11]Department of Psychology, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[12]South London and Maudsley NHS Foundation Trust, London, United Kingdom

[13]Davos Alzheimer's Collaborative, Geneva, Switzerland

[2]Department of Biostatistics & Health Informatics, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom

[3]School of Psychology, University of Sussex, Falmer, United Kingdom

[4]Parc Sanitari Sant Joan de Deu, Sant Joan de Deu Research Institute (IRSJD), Sant Boi de Llobregat, Spain

[5]Mental Health Networking Biomedical Research Centre (CIBERSAM), Madrid, Spain

[6]Department of Psychiatry, Amsterdam UMC, Amsterdam, The Netherlands

[7]Amsterdam Public Health, Mental Health Program, Amsterdam, The Netherlands

[8]School of Population Health Sciences, King's College London, London, United Kingdom

[9]H. Lundbeck A/S, Copenhagen, Denmark

Corresponding Author:
Carolin Oetzmann, PhD
Department of Psychological Medicine, Institute of Psychiatry, Psychology & Neuroscience, King's College London, 16 De Crespigny Park, London, United Kingdom

Abstract

Background: Efforts to advance our understanding of depression have long been constrained by the disorder's vast symptom heterogeneity and by the reliance on self-report, which offers only a partial view of phenotypic expression. Digital phenotyping provides an opportunity to address these core challenges by generating real-time, objective data on behavior and physiology, offering new perspectives on understanding depression phenotypes. Yet, prior efforts to identify such objectively derived subtypes have relied on predefined diagnostic labels or supervised models, limiting discovery to existing clinical categories.

Objective: This study aimed to identify subtypes of depression based on objective sleep and activity data using an unsupervised learning method and to explore how participants transition between these subtypes over time.

Methods: We analyzed longitudinal Fitbit data from 623 participants with recurrent depression enrolled in the Remote Assessment of Disease and Relapse in Major Depressive Disorder study. To identify our subtypes, we applied Gaussian mixture models and hidden Markov models, incorporating a thorough model selection approach that combined grouped cross-validation and seed selection to ensure robustness.

Results: Three activity subtypes (high, light, and low activity) and 4 sleep subtypes (efficient early sleepers, efficient late sleepers, disrupted sleepers, and variable late sleepers) were consistently identified. These subtypes align with known associations between depression and behavioral patterns. Transition modeling revealed stability within individuals over follow-up, further suggesting the presence of behavioral phenotypes rather than momentary fluctuations.

Conclusions: The results demonstrate that wearable-derived features can identify reproducible and clinically relevant behavioral subtypes of sleep and activity in individuals with major depressive disorder. These subtypes reflect known behavioral correlates of depression and may offer a data-driven framework for reducing phenotypic heterogeneity, improving research stratification, and supporting personalized patient monitoring. Further work is needed to validate these findings in independent cohorts and evaluate their potential use in reducing noise when using sleep or activity data to predict depression outcomes.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e86900) doi:10.2196/86900

KEYWORDS

major depressive disorder; digital phenotyping; wearable devices; sleep patterns; physical activity; unsupervised learning; Gaussian mixture models; hidden Markov models; behavioral subtypes; longitudinal data; personalized psychiatry

Introduction

Conventional clinical assessment of depression has predominantly relied on interview-based and self-report measures, which are widely used and clinically informative but may be susceptible to biases and variability in administration and adherence [1,2]. These subjective data can be influenced by factors such as recall bias, social desirability effects, and high interrater variability [3]. With the widespread adoption of smartphones [4] and wearable technologies [5] over the past decade, there is an opportunity to improve the way depression is assessed. Digital phenotyping uses remote measurement technologies (RMTs), such as smartphones or wearable devices, to unobtrusively collect continuous real-time data on patient behavior and physiology, such as heart rate, physical activity, or sleep [2]. These data, collected in a naturalistic setting, provide objective correlates that could help reveal a more comprehensive understanding of individual differences in how subjective depression symptoms manifest in daily life [6].

Sleep problems, circadian rhythm disturbances, and physical inactivity have long been recognized as important and potentially modifiable risk factors for depression [7]. These factors exist within a complex interplay of biological, psychological, and environmental processes and can be influenced through behavioral and clinical interventions. Consistent with their clinical relevance, sleep disturbances or alterations in activity levels are recognized as core features of depression symptomatology [8], weighted equally as emotional and cognitive symptoms in gold-standard diagnostic criteria such as the *Fifth Edition of the Diagnostic and Statistical Manual* (*DSM-5* [9]) and the *Eleventh Revision of the International Classification of Diseases* (*ICD-11* [10]).

The idea of using digital phenotypes to better monitor depression states and symptoms is not new; previous research has established an association between RMT features and depression severity or clinical state [11-13]. For example, earlier studies have consistently identified individuals with depression as having lower physical activity levels, quantified as reduced overall movement, less vigorous activity, or higher sedentary behaviors, compared to controls [14-16]. These associations suggest a general link between lower physical activity and depression and indicate heterogeneity in the ways patients might exhibit reduced activity. This variability, consistent with the broader heterogeneity of depression presentation [17], may point to latent subgroups defined by patterns of physical activity, although this remains speculative.

A large body of literature on sleep identifies diverse dysregulation profiles in sleep patterns and circadian rhythms as both markers of, and risk factors for, depression [11,18,19]. Some studies show that clinical groups exhibit higher nighttime activity than controls [20], while others link evening chronotypes [21] and increased sleep variability [22] to depression. These sleep disturbances indicate that depression manifests differently across individuals, potentially revealing distinct patient subtypes defined by specific sleep patterns. Current diagnostic criteria for major depressive disorder (MDD) include both insomnia and hypersomnia. But the significant co-occurrence rates (up to 30%) suggest that sleep is more complex than these 2 dimensions imply [23].

By using unsupervised clustering techniques [24] on sleep and activity data, we can investigate potential subtypes or patterns that may be revealed in the RMT data alone. This approach contrasts with prior work, which has largely relied on predetermined labels or subjective patient information. In doing so, it enables the identification of potential latent subgroups within a phenotypically diverse sample of individuals with depression. This paper addresses this gap in the literature by (1) identifying subtypes of depression from objective sleep and activity data using an unsupervised learning method and (2) exploring how participants transition between these subtypes over time.

Methods

Study Design and Participants

This study was a secondary analysis of the Remote Assessment of Disease and Relapse Study in Major Depressive Disorder (RADAR-MDD) [25] conducted to explore subtypes of depression using objective sleep and activity data collected via a Fitbit wearable device. The RADAR-MDD study was a longitudinal observational cohort study involving patients with a history of recurrent MDD and aimed to monitor the illness course using remote measurement technologies, such as a Fitbit wearable device.

The study enrolled 623 participants who (1) fulfilled *DSM-5* diagnostic criteria for MDD within the past 2 years and (2) had experienced at least 2 depressive episodes in their lifetime. The exclusion criteria included a history of bipolar disorder, dementia, psychosis or MDD with psychotic features, a history of moderate-to-severe drug or alcohol dependence in the last 6 months, or a history of major medical disease that might impact their ability to participate in normal daily life [25]. Participants were enrolled and followed up for an average of 18 months (range 11 - 36 months). They were asked to wear the Fitbit device throughout the study and complete a battery of questionnaires every 3 months.

The study ran from November 30, 2017, to April 30, 2021, across 3 sites: King's College London (UK), Amsterdam University Medical Centre (the Netherlands), and Centro de Investigación Biomédica en Red (Spain). Participants were recruited from diverse sources, including volunteer registers of people with MDD and clinical samples of people attending mental health care services. The full RADAR-MDD protocol

XSL•FO
RenderX

[25] and data availability, cohort profile, and retention [26] have been published elsewhere.

Ethical Considerations

The study obtained ethical approval from the Camberwell St Giles Research Ethics Committee (REC reference: 17/LO/1154) in London, from the CEIC Fundacio Sant Joan de Deu (CI: PIC-128‑17) in Barcelona, and from the Medische Ethische Toetsingscommissie VUmc (METc VUmc registratienummer: 2018.012—NL63557.029.17—registration date: September 10, 2018) in the Netherlands. All participants provided written informed consent.

Lived Experience Involvement

The RADAR-MDD lived experience group was instrumental in co-designing the study and data collection processes, including the types and frequencies of data collection, as well as contributing to problem-solving throughout the study, as detailed further in Simblett et al [27].

Measures

Three Monthly Outcome Assessments

Throughout the follow-up period, every 3 months, participants were asked to complete a battery of questionnaires; these included, among others, the Inventory of Depressive Symptomatology–Self Report (IDS-SR; [28]), the questionnaire assessing the 7-item Generalized Anxiety Disorder (GAD-7; [29]), and the Work and Social Adjustment Scale (WSAS; [30]).

Digital Sleep and Activity Data

Digital sleep and activity data were collected through Fitbit Charge 2 and later the Charge 3 wearable devices, which recorded information on sleep and physical activity throughout the follow-up period. Table 1 summarizes the sleep and activity features included in the analysis. Of the features available for extraction, these were deemed to provide the most comprehensive and clinically relevant picture of aggregated sleep and activity. Previous literature reviews have identified them as some of the most frequently included digital features in studies linking depression and RMTs [11,13,14,31,32].

Table . Digital feature definitions for the variables included in the subtyping solutions.

Domain and digital feature	Definition
Sleep features	
Total sleep time	Total duration (sum) of all “nonawake” stages
Sleep efficiency	Percentage of total sleep time to time in bed
Time awake in bed	Percentage of total time in bed spent awake
Sleep onset time	Onset time recorded in local time as a decimal number of hours (eg, 22.5=10:30 PM)
Sleep offset time	Offset time recorded in local time as a decimal number of hours (eg, 6.65=6:39 AM)
Activity features	
Sedentary time	Number of minutes labeled “sedentary” all day
Light activity	Number of minutes labeled “lightly active” all day
Moderate activity	Number of minutes labeled “fairly active” all day
Vigorous activity	Number of minutes labeled “very active” all day
Nighttime activity	Number of “active” minutes during nighttime (00:00-05:59)
Total daily calories	Sum of calories per minute along all day (ie, calories per day)

Baseline Sociodemographic and Clinical Characteristics

Assessed at baseline, sociodemographic and clinical characteristics were used to describe differences in the resulting solutions (see Table 2 for the variables included). These variables were chosen based on data availability and clinical relevance.

XSL•FO
RenderX

Table . Sociodemographic and clinical characteristics used to describe subtype membership[a].

Variable	Further information
Age	Age at enrollment
Gender	Female, male, prefer not to say
Employment status	• Employed or furloughed • Retired • Student • Unemployed or sick leave
Family history of depression	Having a parent, sibling, and/or child with a diagnosis of depression
Mental health comorbidity[a]	Presence of a mental health comorbidity
Physical health comorbidity[a]	Presence of a physical health comorbidity
Depression*	The Inventory of Depressive Symptomatology–Self Report (IDS-SR; [28])
Anxiety*	Questionnaire assessing 7-item Generalized Anxiety Disorder (GAD-7; [29])
Functional impairment*	Work and Social Adjustment Scale [30]

[a]Mental and physical health comorbidity groups are not mutually exclusive. All variables were assessed at baseline, except for those marked with an asterisk (*), which were also assessed during the 3-month outcome assessments.

Data Processing, Missing Data Handling, and Outliers

Each completed 3-monthly questionnaire assessment was paired with digital data collected up to 7 days before questionnaire completion. Among the measures, the IDS-SR was selected as the primary criterion for determining whether a 3-monthly assessment was included in the analysis due to its 80% completion rate [26] and significant clinical relevance in assessing depression symptoms. Since the IDS-SR asks respondents to reflect on their symptoms over the past week, a 7-day window was chosen for pairing it with digital data.

Aligning the RMT measures with this time frame was imperative (1) to guarantee that the self-reported symptom experiences participants reflect upon for the IDS-SR would mirror the digitally collected data sources, (2) to ensure consistency in data processing and cleaning across publications (eg, Oetzmann et al [33]), and (3) to allow for future prediction models where the IDS-SR is the outcome measure.

Sleep and activity data were summarized into weekly means and SDs. A missing data threshold of at least 12 hours per day of the available data was used to ensure thorough, reliable summaries of digital features that account for a broad range of daily or nightly behavior changes [34]. Additionally, to provide a realistic view of each week, participants with fewer than 3 days of data within the 7 days before IDS-SR completion were excluded, informed by previous work [35]. The number of observations between a 3-day and 4-day cut-off was minimal (1748 vs 1697 in the activity sample). Given this difference, the more inclusive 3-day cut-off was used to retain a larger dataset. Due to inconsistencies in the raw data during preprocessing, the variable “time awake in bed (SD)” could not be reliably extracted and was therefore excluded from the final model.

This cleaning was undertaken in the activity and sleep data separately to help reduce data loss. As a result, the final datasets for both sleep and activity have varied levels of completeness, with different numbers of participants and observations in each. This variation is accounted for and reported in the results for each corresponding modeling solution. Any remaining missingness within the 2 datasets was assessed and imputed using k-nearest neighbors (k=5) before clustering [36]. Missingness was minimal, and imputation was applied to preserve observations and maintain complete longitudinal trajectories.

Clustering Analysis

Overview

We decided to analyze sleep and activity data separately because (1) despite their inherent connection, they represent distinct behavioral and physiological constructs, (2) the exploratory nature of this analysis benefits from focusing on each domain independently, avoiding interference or dilution of signals, and (3) it simplifies model complexity and enhances interpretability.

Following best practice guidelines [24], sleep and activity data were explored using 2 probabilistic unsupervised machine learning–based clustering approaches: Gaussian mixture models (GMMs; [37]), a static model, and hidden Markov models (HMMs) with Gaussian emissions [38], which consider repeated measures. Using 2 complementary approaches, we could check whether the solution was highly dependent on the algorithm used or method-agnostic, suggesting better generalizability.

Clustering was conducted at the observational level, with baseline and follow-up timepoints included simultaneously in each model. Thus, participants contributing multiple timepoints could be assigned to different clusters across assessments. In the GMM approaches, observations were treated as independent, consistent with identifying cross-sectional latent structure across all observations. In contrast, the HMM explicitly incorporated the temporal ordering of repeated measures, enabling the estimation of transition probabilities that describe movement between states over time.

XSL•FO
RenderX

Model Specifications and Software

We used a diagonal covariance matrix for the GMM and the HMM, as it requires less computational power than a full matrix. Additionally, it reduces the risk of overfitting in smaller datasets (compared to a full matrix) [39]. For the GMM, we used "k-means++" as the setting for the initialization parameter to ensure that the initial cluster centers are spread out in the data space, thereby avoiding convergence on a suboptimal solution. This also reduces computational cost compared to running a full k-means [39]. All other model specifications were set to default according to the Gaussian HMM model in the hmmlearn package [40] and the GaussianMixture in the Scikit-learn toolkit [39,41]. All analyses were conducted using Python (version 3.11.0) [42]. All figures were created in R (version 4.3.1) [43] using the ggplot2 package [44].

Selection of the Optimal Number of Clusters or State Solutions

We undertook several steps to determine the optimal number of clusters or states for the GMM and the HMM separately. Our approach aimed to ensure that the chosen model complexity was based on robust, generalizable performance rather than overfitting or sensitivity to random initializations. For each algorithm, we tested 2 - 10 cluster or state solutions, with 5 folds stratified by participant identifier, to ensure that all observations for a particular participant were always in the same set. The process is outlined in Figure 1.

Figure 1. Step-by-step process for Gaussian mixture models (GMM) or hidden Markov models (HMM) clustering and model selection.

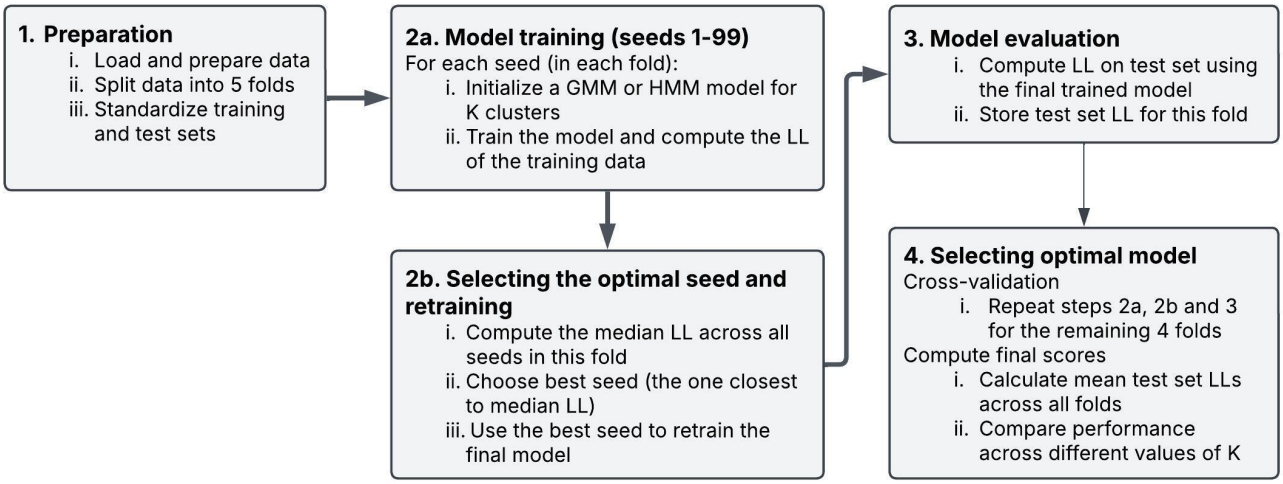

1. Data preparation. First, the data were split into 5 folds grouped by participant using GroupKFold. In each cross-validation iteration, 4 folds were used for training, and the remaining fold served as the test set; no separate hold-out dataset was created. The features in the training data were standardized, which involved centering the data by subtracting the mean and scaling it to have an SD of 1 (mean 0, SD 1). This same transformation was then applied to the testing dataset to prevent data leakage.
2. Model training and seed selection. To reduce sensitivity to random initialization, each model (per fold and cluster number) was trained across 99 seeds, and the seed with the log-likelihood (LL) closest to the median was selected to initialize the final "median-fit" model. LL was used because it reflects model fit under the assumed probabilistic framework, providing an objective and comparable measure across initializations. This approach, analogous to burn-in or stabilization steps used in other clustering algorithms, provides a robust and generalizable solution. The median model seed was chosen because the LL is sensitive to initialization [45]; selecting the median corresponds to the middle ground initialization that is robust to this variability, ensuring a stable and generalizable solution.
3. Evaluating the "median-fit" model on test data. Next, we estimated the test set log-likelihood score by evaluating the test data on the "median-fit" model, providing a test LL score that indicates how well the model generalizes to unseen data. Each of these 5 resulting test LL scores (1 per fold) for K number of clusters is then stored for subsequent analysis.
4. Selecting the optimal number of clusters. Once these steps are repeated for (2 - 10 cluster solutions), the mean of the test LL scores for each fold for each cluster solution is then calculated and plotted on a graph. The optimal number of clusters is then selected by identifying the point at which the curve hits a plateau, indicating that the gain in fit from adding additional clusters is no longer substantial.

Model Fitting and Cluster Interpretation

Once the optimal number of clusters was identified using grouped 5-fold cross-validation, the model was refitted on the full dataset (ie, all 5 folds combined) using the selected number of clusters and the corresponding "median-fit" seed. No separate test set was retained at this stage because generalization performance had already been estimated via cross-validation. This model was then refitted, and cluster assignments (state labels) were predicted for each observation (ie, each observation was assigned to its most likely cluster). Once each observation was assigned to a given cluster, descriptive statistics for each profile were calculated. For each cluster, the mean values of the input RMT metrics were computed and interpreted to ascertain digital feature differences between the identified clusters. To facilitate comparisons across different features, we used the standardized and centered feature values, representing standardized deviations from the population mean for both the GMM and HMM solutions. For ease of interpretation, Tables

XSL•FO
RenderX

S1 and S2 in Multimedia Appendix 1 present the HMM solution in raw units for the sleep and physical activity clusters, respectively. This was only done for the HMM, as the GMM and HMM solutions showed high comparability.

Demographic and Clinical Cluster Differences

Finally, the demographic and clinical characteristics of the states were identified. Descriptive statistics for baseline and follow-up variables were calculated at the observation level, meaning that participants contributing multiple follow-ups appeared multiple times. No formal statistical tests were performed due to the nonindependence of repeated observations. These analyses were exclusively reported on the HMM solution, as the HMM considers repeated measures, and the obtained solutions showed high comparability with those derived from the GMM model.

Transition Probability Estimation

For the HMM, transition probabilities between states were estimated as part of the model fitting. These probabilities represent the probability of transitioning from one state at a given observation to another state at the subsequent observation within the same participant. Transition estimates were derived from all available consecutive 3-month time points per participant, allowing the model to characterize within-person movement between behavioral states over time. Participants contributed transitions only when sequential observations were available.

Results

Sample Characteristics

The overall sample of the RADAR-MDD study consisted of 623 participants; further details on the entire cohort's sample characteristics are reported in Matcham et al [26]. This paper used an analytical sample of 490 participants (2030 observations) for the sleep clustering work and 472 participants (1748 observations) for the activity clustering. Table 3 presents the demographic and clinical characteristics of the overall sample and 2 analytical samples, showing no substantial differences across variables, indicating their comparability.

XSL•FO
RenderX

Table . Sample characteristics for the total sample, sleep, and activity analytical samples.

Characteristics	RADAR-MDD[a] sample (n=623; obvs.=3724)	Sample 1 Sleep data (n=490; obvs.=2030)	Sample 2 Activity data (n=472; obvs.=1748)
Study site, n (%)			
United Kingdom	350 (56.2)	276 (56.3)	271 (57.4)
Spain	155 (24.9)	107 (21.8)	115 (24.4)
The Netherlands	118 (18.9)	107 (21.8)	86 (18.2)
Age, mean (SD)	46.4 (15.3)	45.6 (15.3)	46.3 (15.2)
Gender, n (%)			
Female	471 (75.6)	371 (75.7)	351 (74.4)
Male	152 (24.4)	119 (24.3)	121 (25.6)
Aggregated ethnicity, n (%)			
White British or Dutch	369 (59.2)	314 (64.1)	291 (61.7)
White other	35 (5.6)	28 (5.7)	25 (5.3)
Black ethnic group	14 (2.2)	9 (1.8)	11 (2.3)
Asian ethnic group	16 (2.6)	9 (1.8)	11 (2.3)
Mixed ethnic group	16 (2.6)	10 (2)	9 (1.9)
Other	18 (2.9)	13 (2.7)	10 (2.1)
Not reported[b]	155 (24.9)	107 (21.8)	115 (24.4)
Baseline depression score (IDS-SR)[c]			
IDS-SR total, mean (SD)	31.3 (14.5)	30.8 (14.4)	31.2 (14.3)
None (0 - 13), n (%)	61 (9.8)	50 (10.2)	44 (9.3)
Mild (14-25), n (%)	157 (25.2)	131 (26.7)	123 (26.1)
Moderate (26-38), n (%)	206 (33.1)	164 (33.5)	158 (33.5)
Severe (39-48), n (%)	104 (16.7)	74 (15.1)	80 (16.9)
Very severe (49-84), n (%)	79 (12.7)	61 (12.4)	56 (11.9)
Not reported, n (%)	16 (2.6)	10 (2.0)[d]	11 (2.3)[d]

[a]RADAR-MDD: Remote Assessment of Disease and Relapse in Major Depressive Disorder.

[b]Ethnicity data were not collected at the Spanish site (n=155).

[c]IDS-SR: Inventory of Depressive Symptomatology–Self Report.

[d]These individuals missed baseline IDS-SR measures but provided follow-up IDS-SR measures, which were included in the analysis.

Missing Data

After data cleaning and preprocessing, the resulting data for sleep and activity had marginal amounts of missingness, with the sleep data showing 7 missing values out of 2030 for the IDS-SR total score. In the activity dataset, 8 out of 1748 missing values were found for nighttime activity (mean) and 23 out of 1748 for nighttime activity (SD). This missingness was deemed small and imputed using K-nearest neighbors' imputation.

Digital Sleep Data Clustering

Selecting the Optimal Number of Clusters

Figure S1 in Multimedia Appendix 1 presents the mean LL scores across the 5 folds evaluated on the unseen test data for a given number of clusters. In the GMM solution, LL increased substantially up to 4 clusters, with more modest improvements thereafter. Although the curve continued to rise slightly beyond 4 clusters, the 4-cluster solution was selected to balance model fit, parsimony, and interpretability. Similarly, the HMM solution showed substantial improvement up to 4 states, followed by progressively smaller gains with additional states, indicating diminishing returns beyond this point. Taken together, these patterns support the selection of the 4-cluster solution across both the GMM and HMM solutions.

Description

The 4-cluster GMM and 4-state HMM solutions have similar descriptive characteristics across the digital sleep features used to define the model. The 4-cluster solution for the GMM and the 4-state solution for the HMM are visualized in Figure 2, presenting deviations from the population mean. Table S1 in Multimedia Appendix 1 presents the findings in raw units

XSL•FO
RenderX

without scaling for the HMM solution. Overall, these results indicate the following:

- State 1 or cluster 1 (ie, efficient early sleepers) had higher-than-average sleep efficiency and lower-than-average awakenings (an average sleep efficiency of 92.4% and 7.6% of the time in bed spent awake), with earlier-than-average sleep onset and offset times (mean) (an average onset at 23:40 and offset at 7:39) and lower-than-average levels of sleep onset and sleep offset time (SD) variation, indicating consistent sleep patterns night-to-night.
- State 2 or cluster 2 (ie, efficient late sleepers) had higher-than-average sleep efficiency (average sleep efficiency of 92% and 8% of time spent awake) and lower-than-average awakenings. They had later-than-average sleep onset and offset times (mean), with a mean sleep onset at 00:47 (the next day) and offset at 8:25.
- State 3 or cluster 3 (ie, disrupted sleepers) had lower-than-average sleep efficiency levels and higher-than-average levels of awakenings (mean sleep efficiency of 90.2% and 9.8% of the time spent awake). They presented earlier-than-average sleep onset and offset time (mean) compared to the overall population (an average onset time of 23:32 and an offset at 07:24).
- State 4 or cluster 4 (ie, variable late sleepers) had the highest levels of variation (around 2 hours each) in total sleep time (SD), sleep onset, and offset time (SD) over the 7 days compared to the population's mean variation. They presented with lower-than-average efficient sleep (90.6%), spending approximately 9.4% of their time in bed awake, and later-than-average sleep onset and offset times (01:22 the next day and 09:33, respectively).

XSL•FO
RenderX

Figure 2. Descriptive means for the sleep features included in the 4-cluster Gaussian mixture model (GMM) and 4-state hidden Markov model (HMM) solutions. Obvs.: number of observations in the state or cluster. The data values are presented in Table S3 in Multimedia Appendix 1.

XSL•FO
RenderX

Demographic and Clinical Characteristics of the Sleep States

Table 4 describes the demographic and clinical characteristics of each state. The clusters presented similar characteristics, with only slight differences observed across variables. For state 1, efficient early sleepers had a slightly higher proportion of female participants than the overall sample. In contrast, in state 4, variable late sleepers had the highest proportion of male participants. When considering employment, state 4, variable late sleepers, had a slightly lower proportion of people employed or furloughed and a higher proportion of unemployed or sick leave individuals compared to the overall. The prevalence of physical health comorbidities varied across states; state 1 (efficient early sleepers) had the least, while state 4 (variable late sleepers) had the highest presence of physical comorbidities. No substantial differences in depression severity and anxiety were observed; all states fell within the moderate range on the IDS-SR, though a 10-point difference was observed between state 1 (efficient early sleepers; 27.1) and state 4 (variable late sleepers; 38.4), highlighting potentially clinically relevant variation within the same severity category. States 2 (efficient late sleepers) and 4 (variable late sleepers) exhibited marginally higher GAD-7 anxiety scores compared to the other states and were slightly more affected by functional disability, indicating significant impairments.

Table . Demographic and clinical characteristics of each state identified in the hidden Markov model (HMM) sleep feature solution.

Variable	Overall	State 1 Efficient early sleepers	State 2 Efficient late sleepers	State 3 Disrupted sleeper	State 4 Variable late sleepers
Observations, n (%)	2030 (100)	565 (27.8)	731 (36)	463 (22.8)	271 (13.3)
Age (y), mean (SD)	47.4 (14.9)	47 (15)	45 (14.5)	51 (14.3)	48 (15.4)
Gender, n (%)					
Female	1577 (77.7)	488 (86.4)	561 (76.7)	357 (77.1)	171 (63.1)
Study site, n (%)					
United Kingdom	1231 (60.6)	354 (62.7)	434 (59.4)	297 (64.1)	146 (53.9)
Spain	415 (20.4)	86 (15.2)	172 (23.5)	83 (17.9)	74 (27.3)
The Netherlands	384 (18.9)	125 (22.1)	125 (17.1)	83 (17.9)	51 (18.8)
Employment status, n (%)					
Employed or furloughed	898 (44.2)	273 (48.3)	336 (46)	201 (43.4)	88 (32.5)
Retired	444 (21.9)	130 (23)	125 (17.1)	118 (25.5)	71 (26.2)
Student	193 (9.5)	52 (9.2)	79 (10.8)	37 (8)	25 (9.2)
Unemployed or sick leave	383 (18.9)	96 (17)	141 (19.3)	85 (18.4)	61 (22.5)
Other or not available	112 (5.5)	14 (2.5)	50 (6.8)	22 (4.8)	26 (9.6)
Family history of depression, n (%)					
Yes	1470 (72.4)	411 (72.7)	511 (69.9)	355 (76.7)	193 (71.2)
Mental health comorbidity, n (%)[a]					
Yes	1398 (68.9)	386 (68.3)	542 (74.1)	293 (63.3)	177 (65.3)
Physical health comorbidity, n (%)[a]					
Yes	1017 (50.1)	219 (38.8)	398 (54.4)	211 (45.6)	189 (69.7)
IDS-SR[b], mean (SD)	31.0 (15.3)	27.1 (14.7)	32.8 (15)	28.5 (14)	38.4 (16)
WSAS[c], mean (SD)[d]	18.8 (11.1)	15.9 (11)	20.3 (10.5)	17.7 (11.1)	22.2 (11)
GAD-7[e], mean (SD)[d]	8.3 (5.4)	7.5 (5.3)	9.1 (5.2)	7.6 (5.2)	9.1 (6)

[a]Mental health and physical health comorbidity are not mutually exclusive groups.

[b]IDS-SR: Inventory of Depressive Symptomatology–Self Report. Scores of 26-38 indicate moderate levels of depression, and scores of 39-48 indicate severe levels of depression [28].

[c]WSAS: Work and Social Adjustment Scale. Scores of 11-20 indicate some impairment, and scores >20 indicate significant impairment [30].

[d]Indicates variables with missing data. The WSAS had 30 missing entries, and the GAD-7 had 14 entries.

[e]GAD-7: 7-item Generalized Anxiety Disorder. Scores ≥10 indicate moderate-to-severe anxiety [29].

XSL•FO
RenderX

Digital Physical Activity Data Clustering

Selecting the Optimal Number of Clusters

Figure S2 in Multimedia Appendix 1 presents the mean LL scores across the 5 folds evaluated on the unseen test data, for the physical activity models. For the GMM, a clear inflection point is observed at 3 clusters. This suggests that a 3-cluster solution provides the best fit for the data. The HMM LL curve shows an inflection point at 3 and 7 states. Based on these findings, the 3-state GMM and HMM solutions were selected as the optimal number of clusters. The 7-state HMM was also explored and is presented in Table S4 in Multimedia Appendix 1 for further reference.

Description

The 3-cluster solutions for the GMM and HMM models presented similar solutions. Figure 3 describes these differences, presenting deviations from the population mean. Table S2 in Multimedia Appendix 1 presents the findings in raw units without scaling for the HMM solution.

XSL•FO
RenderX

Figure 3. Descriptive means for the physical activity features included in the 3-cluster Gaussian mixture model (GMM) and 3-state hidden Markov model (HMM) solutions. A.: activity; Obvs.: number of observations in the state or cluster. The data values are presented in Table S5 in Multimedia Appendix 1.

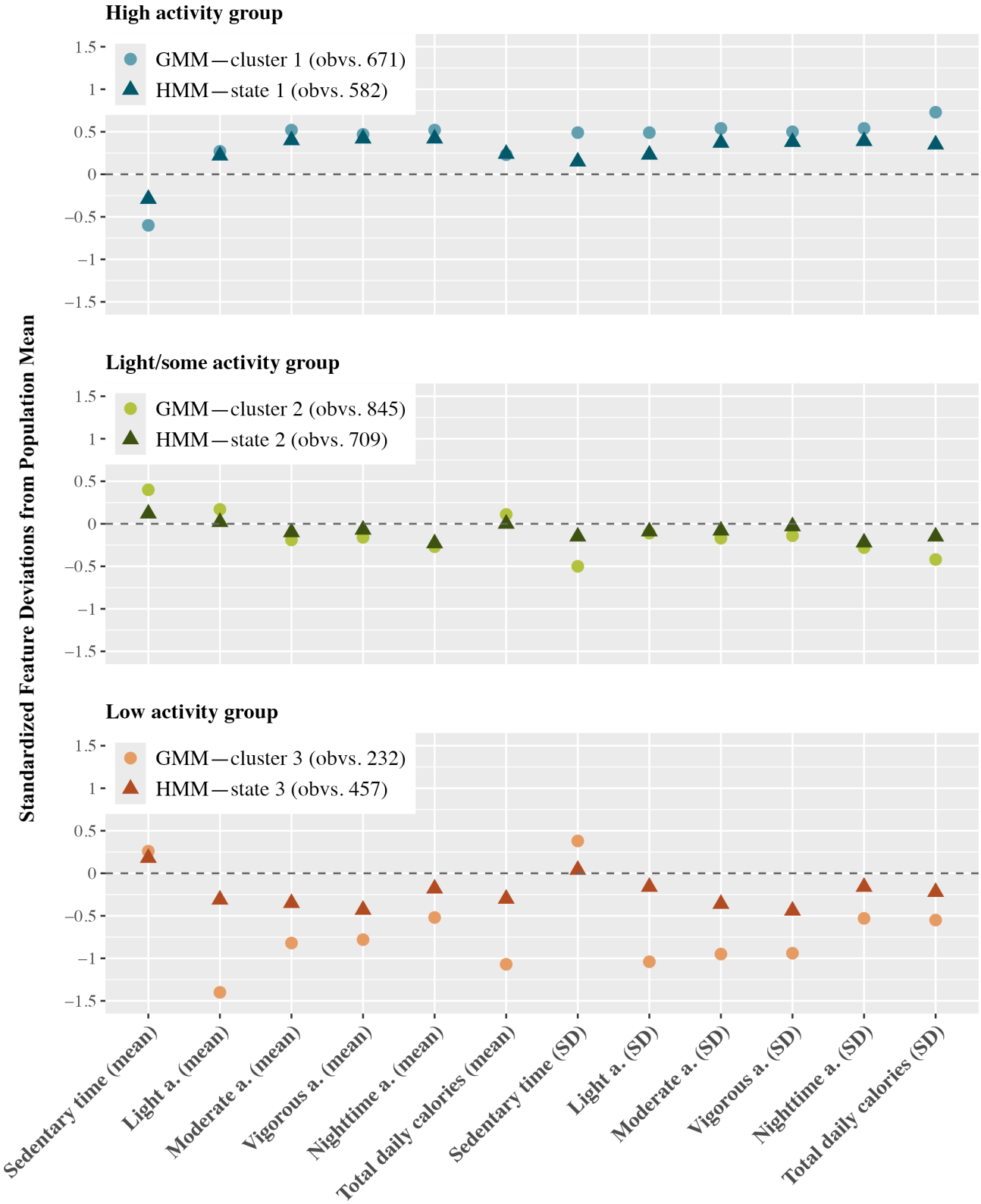

- State 1 or cluster 1 (ie, high activity) had the lowest mean levels of sedentary time compared to any other cluster (18.5 h), the highest levels of light, moderate, and vigorous activity (around 3 h, 26 min, and 26 min, respectively), and the highest level of variation in activity levels compared to the population mean (around 20 minutes for moderate and vigorous activity).
- State 2 or cluster 2 (ie, light or some activity) presented with above-average levels of sedentary time and slightly above-average levels of light activity compared to the population mean. This translates to around 19.5 hours of

XSL•FO
RenderX

sedentary time, around 3 hours of light activity, 15 minutes of moderate, and 15 minutes of vigorous activity. All measures of variation in activity showed results near the total sample mean.

- State 3 or cluster 3 (ie, low activity) presented the lowest levels of light, moderate, and vigorous activity compared to any other cluster and slightly above-average sedentary time. This translates to a mean of 2.5 hours in light activity, 10 minutes in moderate, and 8 minutes in vigorous activity but a sedentary time of nearly 20 hours.

Demographic and Clinical Characteristics of the Activity States

Their demographic and clinical characteristics were explored to examine the distinctions among the 3 delineated activity groups. The states varied only marginally across characteristics. Table 5 presents these differences, showing state 1 (high activity) to have the highest proportion of male participants (n=198, 34%) and instances of no physical health comorbidity (n= 335, 57.6%) compared to any other cluster or the overall sample. State 3 (low activity) had a slightly higher proportion of retired participants than the overall sample (n= 126, 27.6% and n= 383, 21.9%, respectively). Clinical differences are minor, with all states presenting moderate levels of depression (scores between 26 and 38). State 3 (low activity) presented the highest moderate IDS-SR score compared to the other 2 groups. All states presented some functional impairment (scores between 11 and 20) and mild anxiety scores between 5 and 9.

Table . Demographic and clinical characteristics of each state identified in the hidden Markov models (HMM) physical activity feature solution.

Variable	Overall	State 1 High activity group	State 2 Light or some activity	State 3 Low activity
Observations, n (%)	1748	582 (33.3)	709 (40.6)	457 (26.1)
Age, mean (SD)	47.3 (14.6)	46.6 (14.8)	46.8 (14.5)	48.8 (14.4)
Gender, n (%)				
Female	1297 (74.2)	384 (66)	554 (78.1)	359 (78.6)
Study site, n (%)				
United Kingdom	1153 (66)	334 (57.4)	535 (75.5)	284 (62.1)
Spain	393 (22.5)	141 (24.2)	124 (17.5)	128 (28)
The Netherlands	202 (11.6)	107 (18.4)	50 (7.1)	45 (9.8)
Employment status, n (%)				
Employed or furloughed	787 (45.0)	275 (47.3)	329 (46.4)	183 (40)
Retired	383 (21.9)	116 (19.9)	141 (19.9)	126 (27.6)
Student	158 (9)	44 (7.6)	77 (10.9)	37 (8.1)
Unemployed or sick leave	332 (19)	123 (21.1)	133 (18.8)	76 (16.6)
Other or Not available	88 (5)	24 (4.1)	29 (4.1)	35 (7.7)
Family history of depression, n (%)				
Yes	1276 (73)	416 (71.5)	530 (74.8)	330 (72.2)
Mental health comorbidity, n (%)[a]				
Yes	1198 (68.5)	368 (63.2)	479 (67.6)	351 (76.8)
Physical health comorbidity, n (%)[a]				
Yes	919 (52.6)	247 (42.4)	400 (56.4)	272 (59.5)
IDS-SR[b], mean (SD)	31.3 (15.4)	28.9 (15.3)	30.9 (15.1)	34.8 (15.2)
WSAS[c], mean (SD)[d]	18.8 (11.3)	18.5 (11.2)	17.9 (11.3)	20.6 (11.3)
GAD-7[e], mean (SD)[d]	8.5 (5.5)	8.1 (5.4)	8.2 (5.3)	9.2 (5.6)

[a]Mental health and physical health comorbidity are not mutually exclusive groups.

[b]IDS-SR scores of 26 - 38 indicate moderate levels of depression, while scores of 39 - 48 indicate severe levels of depression [28].

[c]WSAS: Work and Social Adjustment Scale. Scores of 11 - 20 indicate some impairment, and scores >20 indicate significant impairment [30].

[d]Indicates variables with missing data: WSAS had 30 missing entries, and GAD-7 had 14 missing entries.

[e]GAD-7: 7-item Generalized Anxiety Disorder. Scores ≥10 indicate moderate-to-severe anxiety [29].

XSL•FO
RenderX

Transition Probabilities

The likelihood of transitioning between the hidden states according to the sleep and activity HMM models can be visualized in Figure 4. Across both the sleep and the physical activity solutions, the largest observed probabilities were self-transitions, meaning individuals were most likely to remain in the same state. When considering physical activity hidden states specifically, the most likely transitions across different states were from the high activity group to the light or some activity group and from the low activity group to the light or some activity group, with both reporting around a 26% chance of transitioning. Transitions between low- and high-activity groups were less likely, reporting around a 15% chance. When considering sleep, the most likely transition between different states was moving from the variable late sleepers to efficient late sleepers, with a 31.2% chance.

Figure 4. Flowchart indicating transition probabilities as observed in the physical activity hidden Markov models (HMM) and the sleep HMM. Transitions with less than 10% transition probability were omitted to improve readability. Arrows are weighted to highlight more probable transitions.

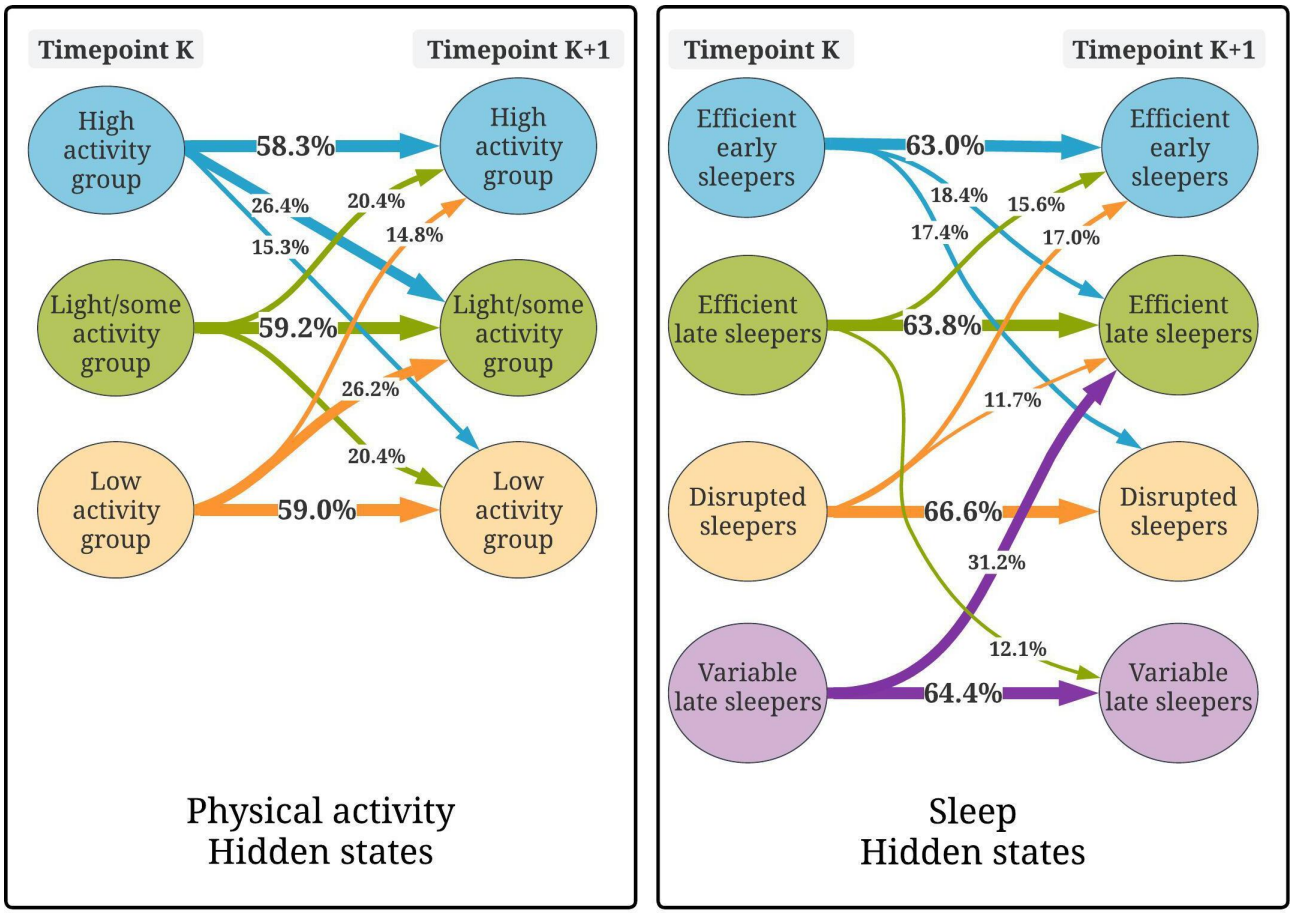

Discussion

Principal Findings

This study identified subtypes based on objective sleep and activity data in a heterogeneous sample of participants diagnosed with MDD. It also explored transitions between these subtypes over time. Previous research has shown that various digital features related to sleep and activity can predict the clinical state of depression [11,13]. Based on this, we hypothesized that these same digital markers might be able to uncover nuanced behavioral differences among patients with MDD and identify latent subtypes to provide refined insight into the diverse phenotypes of depression.

To accomplish this, we used cross-fold validation and a seed optimization procedure to derive a stable and reproducible optimal number of clusters in the RADAR-MDD sample. We identified a stable 4-group model as the best fit for sleep feature data and a 3-group solution as the best fit for the physical activity data. The resulting profiles are depicted in Figure 5. Participants were most likely to remain in the same state over time.

XSL•FO
RenderX

Figure 5. Diagram describing the subtypes identified in the digital sleep and digital physical activity hidden Markov models (HMM) solutions.

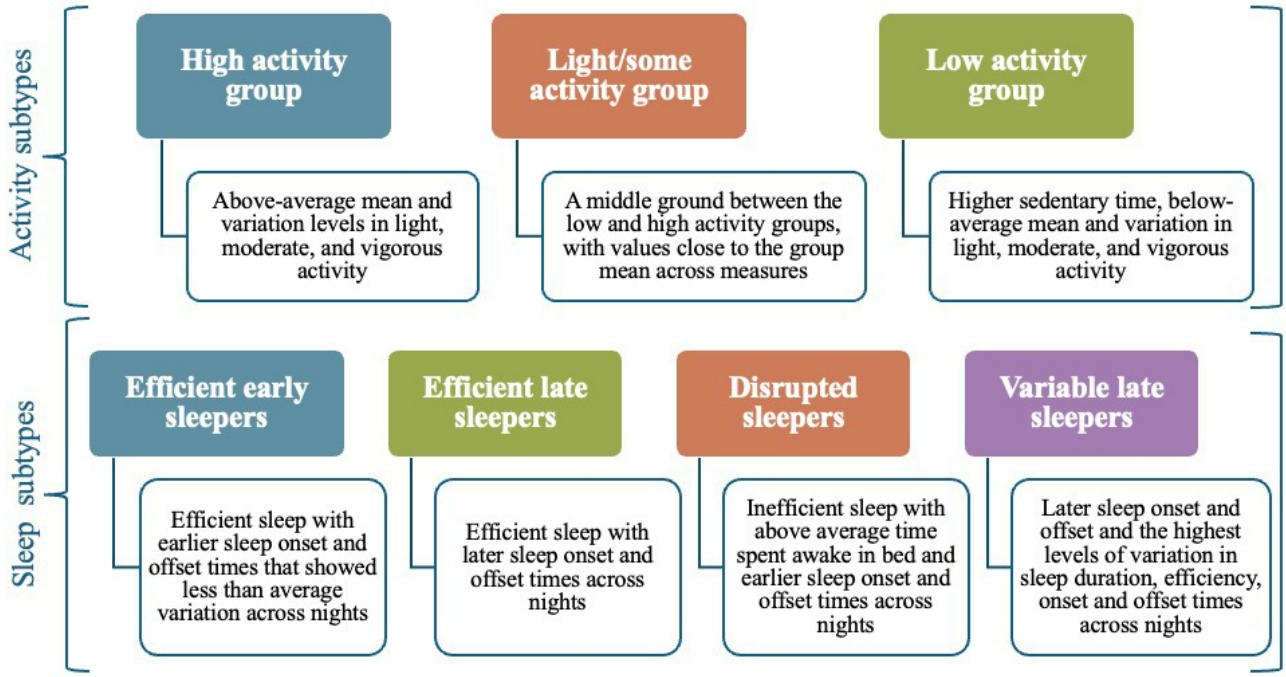

Implications and Comparisons with Previous Work

Sleep Profiles Identified via RMT

The human sleep-wake cycle is governed by the interplay between sleep homeostasis and circadian processes. Dysregulation of this system, for example, when an individual's internal circadian rhythm and external behavior do not align, can cause sleep disruptions [46]. With approximately 90% of patients affected, sleep disturbance is one of the most commonly reported symptoms of depression [47,48]. This study aimed to examine these differences via clustering of sleep RMT data, identifying 4 subtypes in the RADAR-MDD sample.

Notably, variable late sleepers were characterized by later mean sleep-wake patterns, higher awakenings, and inconsistent sleep onset, offset, and durations over the 7 days compared to the overall population. They typically went to bed around 1:22 AM (the following day) and woke around 9:30 AM, with over 2 hours of variability in these times each week. Among others, the observed variability in sleep patterns may indicate social jet lag, where differing weekday and weekend schedules disrupt sleep cycles [49]. This variability to meet social and work demands can negatively impact circadian rhythms, increasing the risk of poor mental and physical health [49,50].

Previous literature looking at sleep profiles in depression frequently draws a distinction between morning and evening preference types [51,52]. Seo et al [53] describe this as morning-types preferring earlier bedtimes and wake-up times, while evening-types prefer later bedtimes and wake-up times according to self-report measures. While we cannot infer personal preference due to the passive nature of the Fitbit sleep data collection, we identify distinct differences in sleep onset and offset times, which mirror similar patterns.

The efficient late sleepers presented similar later sleep-wake patterns (as seen in variable late sleepers) but efficient sleep with little time awake compared to the overall sample, with an average sleep onset of 00:47 (the next day) and offset at 8:25. Previous works examining later sleep-wake patterns using wearables have found associations with worse outcomes [22,54-56]. Studies using self-report metrics have found that a morning preference may be protective against depression symptoms [57], while evening preference is associated with worse adverse effects [18,58]. Previous literature indicates rumination as a key psychological mediator in the relationship between depression and later sleep onset [59].

In contrast, both disturbed sleepers and efficient early sleepers showed earlier than average sleep onset and offset times, with average variability in these timings. They typically began sleeping around 23:30 and woke around 7:30, around 1 to 2 hours earlier compared to their late sleep counterparts and experienced less than an hour of variation in both onset and offset times each week. They also differ, with efficient early sleepers presenting with high sleep efficiency with little time awake in bed, while disturbed sleepers present with the highest levels of inefficiency and time awake in bed compared to any other state. It is important to highlight that while these differences appear substantial within the context of the overall sample range, they translate to only modest absolute differences, with approximately 2% more time spent awake in the disturbed sleepers group. Despite the modest differences identified, previous literature has also identified poor sleep efficiency and increased wakefulness after sleep onset in patients with depression compared to controls [60].

A few sociodemographic differences were observed across the sleep subtypes, such as a higher prevalence of physical health comorbidities and a higher proportion of unemployment and sick leave in individuals with variable late sleep compared to the overall sample. These differences were only descriptive and not substantial enough to suggest clear, distinguishing features. The absence of such distinct differences indicates that the sleep

XSL•FO
RenderX

subtypes may reflect more stable, inherent characteristics of the RADAR-MDD cohort rather than being heavily influenced by other participant characteristics. The finding that participants were most likely to stay in 1 subtype over time further suggests the presence of trait-like differences, indicating that these latent states may reflect underlying behavioral phenotypes rather than short-term fluctuations within the heterogeneous RADAR-MDD population. Nevertheless, transitions that did occur most commonly involved movement from more extreme profiles toward intermediate patterns (eg, high or low activity toward light or some activity; variable late sleepers toward efficient late sleepers), which may partly reflect regression to the mean as well as genuine behavioral change over time.

Depression severity differed across sleep states, although all states fell within the moderate range on the IDS-SR. State 4 (variable late sleepers) scored 10 points higher than state 1 (efficient early sleepers), representing a potentially clinically relevant difference in symptom severity. These findings are consistent with prior work, finding greater variability in sleep timings and later sleep-wake patterns associated with worse depression [11,22,49]. Together, these findings suggest that sleep subtypes may reflect differences in sleep-wake patterns and behavioral regularity within depression, ranging from stable and efficient sleep patterns to delayed and irregular patterns. By categorizing these relationships into possible subtypes, these findings provide an exploratory basis for a framework that identifies consistent sleep-depression profiles, which might mitigate the challenges presented by symptom heterogeneity in research and support more tailored circadian and sleep-focused approaches aimed at improving sleep-wake timing and regularity, such as structured sleep scheduling, sleep hygiene practices, or light exposure interventions.

Physical Activity Profiles Identified via RMT

In terms of physical activity measures, our subtyping analysis identified 3 distinct groups: a high-activity subtype with above-average levels of light, moderate, and vigorous activity, participating in nearly 1 hour of moderate-to-vigorous activity daily; a low-activity subtype with high levels of sedentary time (roughly 1.5 hours more than the high-activity group) and below-average levels across measures of physical activity with around 18 minutes spent in moderate-to-vigorous activity; and, finally, a group that presented a middle ground between these 2 extremes (light or some activity group), spending an average of 30 minutes daily in moderate-to-vigorous activity.

Previous research has identified various associations, including low gross physical activity and low moderate-to-vigorous physical activity with depression [14,61]. Research suggests the presence of a bidirectional link between low mood and physical inactivity, where symptoms of depression, such as feelings of anhedonia and low mood, can result in reduced motivation to engage in physical activity [62], when, in turn, low physical activity also increases the risk of depression [14]. Taken together, these findings suggest that the activity subtypes may reflect differing levels of behavioral activity and daily structure, ranging from highly engaged patterns to reduced activity consistent with behavioral withdrawal. Such patterns may be relevant for interventions aimed at increasing behavioral activation and daily activity. Employment status represents an important contextual factor, as individuals with depression might take leave from work, resulting in reduced activity levels. This study did not identify any meaningful descriptive differences in depression scores between the low, light, and high-activity subtypes, except for employment, where a slightly higher proportion of retired participants was found in the low-activity group compared to the overall sample.

Strengths and Limitations

A strength of this study lies in its rigorous methodology for model identification. It emphasizes model robustness through grouped cross-validation to reduce the risk of overfitting the final model to the training data and enhancing generalizability [24]. It effectively addresses concerns about model instability by methodically managing the randomness associated with seed selection and reduces the risk of outlier-dependent performance by selecting the seed with the median LL value. Finally, applying this methodology to the dataset using 2 different clustering algorithms further improves the generalizability of the solution.

A core limitation of this study is the lack of external validation, which limits conclusions to the RADAR-MDD dataset. Future work should validate the model using (1) other depression cohorts and (2) through qualitative research with patients to examine the extent to which these data-derived groups reflect lived experience, work exceeding the scope of this analysis.

Another limitation of the methodological approach lies in the feature selection for the GMM and HMM models. Best practices in machine learning suggest starting with a broad set of features and using dimensionality reduction or feature selection techniques to retain the most influential variables that drive meaningful patterns in the data [24]. This paper could not use such techniques due to data availability and data quality; instead, it relied on previous literature and clinical relevance to justify feature inclusion. For example, measures of sleep stages were systematically excluded from the analysis due to insufficient evidence validating their accuracy [63,64]. As the validity of these measures improves, their inclusion in future subtyping work could lead to different or new profiles not identified here. Once achieved, future work could consider using a broader approach where all Fitbit features undergo data reduction before being included in the unsupervised learning models.

Another study limitation is the descriptive method used to outline clinical and demographic characteristics across the defined states. This method is limited as (1) each data point was placed in its most probable cluster without considering the likelihood of belonging to multiple clusters; (2) comparisons were made among observations instead of participants, which may distort the findings in favor of individuals with more repeated observations; and (3) the descriptive nature of this approach hinders our ability to assess the predictive capability of these subtypes in assessing depression scores. In addition, although missing data were minimal, k-nearest neighbors' imputation was used to preserve longitudinal trajectories, which may introduce minor smoothing of the data, although any impact is likely minimal.

XSL•FO
RenderX

We acknowledge the vast heterogeneity observed in the RADAR-MDD dataset. Participants were not necessarily in an acute episode of depression throughout the 2-year follow-up, resulting in a diverse set of observations with varying levels of severity and caseness in the sample. While this can be considered an advantage since it reflects the real-world complexity of depression, it also restricts the ability to confidently determine whether the differences observed are traits found in the general population or unique to individuals with an acute episode.

Conclusions

In conclusion, previous research suggests that various digital markers linked to sleep and activity can serve as predictors of depression. Therefore, we posited that these digital indicators could highlight phenotypic differences in patients and assist in identifying meaningful subtypes in the RADAR-MDD population. The identified subtypes correspond well to previously established links between depression and both objective and subjective measures of sleep and activity. By categorizing these relationships into potential subtypes, the findings lay the groundwork for a framework that identifies consistent depression profiles, potentially alleviating the challenges posed by symptom diversity in research and clinical practice. While further validation is required, such profiles may support more personalized patient management by informing targeted approaches to sleep timing and regularity, promoting daily activity and engagement, and enabling more context-aware monitoring. By summarizing complex behavioral patterns into clinically meaningful profiles, this approach may also facilitate early recognition of meaningful behavioral changes over time.

Acknowledgments

The authors thank their colleagues, both within the RADAR-CNS consortium and across all involved institutions, for their contributions to the recruitment strategy for the Remote Assessment of Disease and Relapse Study in Major Depressive Disorder study [65]. Furthermore, they would like to thank all members of the RADAR-CNS patient advisory board, who all have experience of living with or supporting those who are living with depression, epilepsy, or multiple sclerosis.

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: code optimization, creation of algorithms for data analysis, proofreading and editing, and adapting and adjusting emotional tone. The GAI tools used were ChatGPT 4.5 and Grammarly Version 8. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by: all authors.

They used ChatGPT to support in code optimization and to support in the creation of algorithms used in the analyses. No code was generated by AI and not subsequently checked and tested by 2 separate coauthors. Grammarly AI was used to support in adjusting tone when needed as well as proofreading, spell-checking, and editing of the final manuscript.

Funding

This study was funded by the UK Medical Research Council (MR/N013700/1) via the MRC Doctoral Training Partnership in Biomedical Sciences. EC is part-funded by the National Institute for Health and Care Research Maudsley Biomedical Research Centre (BRC). FL and BWJHP are supported by Stress in Action. The research project "Stress in Action" (https://stress-in-action.nl/z) is financially supported by the Dutch Research Council and the Dutch Ministry of Education, Culture and Science (NWO gravitation grant number 024.005.010). The RADAR-CNS project has received funding from the Innovative Medicines Initiative (IMI) 2 Joint Undertaking under grant agreement number 115902. This Joint Undertaking receives support from the European Union’s Horizon 2020 research and innovation program and the European Federation of Pharmaceutical Industries and Associations (Efpia) (https://www.ihi.europa.eu/). This communication reflects the views of the RADAR-CNS consortium and neither IMI nor the European Union and EFPIA are liable for any use that may be made of the information contained herein. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The views expressed are those of the authors and not necessarily those of the National Institute for Health and Care Research or the Department of Health and Social Care.

Data Availability

The data that support the findings of this study are available from the Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD) study, but restrictions apply to the availability of these data, which were used under license for this study, and so are not publicly available. Data are, however, available from the authors upon reasonable request and with permission of the RADAR-MDD. The underlying code for this study is available in an online repository titled “Unsupervised-Clustering-of-Wearable-Derived-Behavioural-Features” and can be accessed via [66].

Authors' Contributions

Conceptualization: CO, EC, MH, NC, TW
Data curation: AAF, CO, NC, SV, YZ

XSL•FO
RenderX

Formal analysis: CO
Funding acquisition: AAF, BWJHP, CO, DL, FL, JMH, MH, RJBD, TW, VAN
Investigation: AAF, BWJHP, CO, DL, FL, FM, JMH, KMW, MH, PA, RJBD, SS, TW, VAN
Methodology: CO, EC, MH, NC
Project administration: BWJHP, CO, EC, FL, FM, JMH, MH, NC, PA, SS, VAN
Resources: BWJHP, CO, FL, FM, JMH, MH, RJBD, SS, VAN
Software: BWJHP, CO, RJBD, YZ
Supervision: BWJHP, CO, EC, FL, FM, MH, NC, PA, SS, VAN
Validation: CO, EC, MH, NC
Visualization: CO
Writing – original draft: CO
Writing – review & editing: CO, DL, EC, FL, FM, MH, NC, PA, YZ

Conflicts of Interest

JMH has received economic compensation for participating in advisory boards or giving educational lectures from Eli Lilly & Co, Sanofi, Lundbeck, and Otsuka. MH is the principal investigator of the RADAR-CNS program, a precompetitive public-private partnership funded by the Innovative Medicines Initiative and the European Federation of Pharmaceutical Industries and Associations. The program received support from Janssen, Biogen, MSD, UCB, and Lundbeck. SV and VAN were full-time employees of Janssen Research & Development LLC and held company stocks/stock options at the time of study conduct. AAF has shares in Alphabet, the parent company of FitBit. RJBD is the director for cogstack, Onsentia and have advisory roles for Megi.ai and JJ. PA was an employee at Lundbeck A/S at the time of the study. The other authors declare no conflicts of interest.

Multimedia Appendix 1
Tables and figures with further information relevant to the paper.
[PDF File, 462 KB - jmir_v28i1e86900_app1.pdf]

References

1. Maruani J, Mauries S, Zehani F, Lejoyeux M, Geoffroy PA. Exploring actigraphy as a digital phenotyping measure: a study on differentiating psychomotor agitation and retardation in depression. Acta Psychiatr Scand 2025 Mar;151(3):401-411. [doi: 10.1111/acps.13739] [Medline: 39030838]
2. Torous J, Kiang MV, Lorme J, Onnela JP. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR Ment Health 2016 May 5;3(2):e16. [doi: 10.2196/mental.5165] [Medline: 27150677]
3. Bourla A, Ferreri F, Ogorzelec L, Guinchard C, Mouchabac S. Assessment of mood disorders by passive data gathering: the concept of digital phenotype versus psychiatrist's professional culture. Encephale 2018 Apr;44(2):168-175. [doi: 10.1016/j.encep.2017.07.007] [Medline: 29096909]
4. Sherif A. Global smartphone penetration rate as share of population from 2016 to 2024. Statista. 2025. URL: https://www.statista.com/statistics/203734/global-smartphone-penetration-per-capita-since-2005/ [accessed 2024-10-18]
5. Indrakumari R, Poongodi T, Suresh P, Balamurugan B. The growing role of internet of things in healthcare wearables. In: Balas VE, Solanki VK, Kumar R, editors. Emergence of Pharmaceutical Industry Growth with Industrial IoT Approach: Academic Press; 2020:163-194. [doi: 10.1016/B978-0-12-819593-2.00006-6]
6. Dunn J, Runge R, Snyder M. Wearables and the medical revolution. Per Med 2018 Sep;15(5):429-448. [doi: 10.2217/pme-2018-0044] [Medline: 30259801]
7. Glaus J, Kang SJ, Guo W, et al. Objectively assessed sleep and physical activity in depression subtypes and its mediating role in their association with cardiovascular risk factors. J Psychiatr Res 2023 Jul;163:325-336. [doi: 10.1016/j.jpsychires.2023.05.042] [Medline: 37253320]
8. Difrancesco S, Lamers F, Riese H, et al. Sleep, circadian rhythm, and physical activity patterns in depressive and anxiety disorders: a 2-week ambulatory assessment study. Depress Anxiety 2019 Oct;36(10):975-986. [doi: 10.1002/da.22949] [Medline: 31348850]
9. Diagnostic and Statistical Manual of Mental Disorders: DSM-5: American Psychiatric Association; 2013. [doi: 10.1176/APPI.BOOKS.9780890425596]
10. ICD-11 for mortality and morbidity statistics. World Health Organisation. 2024. URL: https://icd.who.int/browse/2024-01/mms/en#334423054 [accessed 2024-08-23]
11. De Angel V, Lewis S, White K, et al. Digital health tools for the passive monitoring of depression: a systematic review of methods. NPJ Digit Med 2022 Jan 11;5(1):3. [doi: 10.1038/s41746-021-00548-8] [Medline: 35017634]
12. Abd-Alrazaq A, AlSaad R, Shuweihdi F, Ahmed A, Aziz S, Sheikh J. Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression. NPJ Digit Med 2023 May 5;6(1):84. [doi: 10.1038/s41746-023-00828-5] [Medline: 37147384]

XSL•FO
RenderX

13. Zierer C, Behrendt C, Lepach-Engelhardt AC. Digital biomarkers in depression: a systematic review and call for standardization and harmonization of feature engineering. J Affect Disord 2024 Jul 1;356:438-449. [doi: 10.1016/j.jad.2024.03.163] [Medline: 38583596]
14. Schuch F, Vancampfort D, Firth J, et al. Physical activity and sedentary behavior in people with major depressive disorder: a systematic review and meta-analysis. J Affect Disord 2017 Mar 1;210:139-150. [doi: 10.1016/j.jad.2016.10.050] [Medline: 28033521]
15. Tazawa Y, Wada M, Mitsukura Y, et al. Actigraphy for evaluation of mood disorders: a systematic review and meta-analysis. J Affect Disord 2019 Jun 15;253:257-269. [doi: 10.1016/j.jad.2019.04.087] [Medline: 31060012]
16. Wüthrich F, Nabb CB, Mittal VA, Shankman SA, Walther S. Actigraphically measured psychomotor slowing in depression: systematic review and meta-analysis. Psychol Med 2022 May;52(7):1208-1221. [doi: 10.1017/S0033291722000903] [Medline: 35550677]
17. Ostergaard SD, Jensen SOW, Bech P. The heterogeneity of the depressive syndrome: when numbers get serious. Acta Psychiatr Scand 2011 Dec;124(6):495-496. [doi: 10.1111/j.1600-0447.2011.01744.x] [Medline: 21838736]
18. Antypa N, Verkuil B, Molendijk M, Schoevers R, Penninx BWJH, Van Der Does W. Associations between chronotypes and psychological vulnerability factors of depression. Chronobiol Int 2017;34(8):1125-1135. [doi: 10.1080/07420528.2017.1345932] [Medline: 28759270]
19. Nutt D, Wilson S, Paterson L. Sleep disorders as core symptoms of depression. Dialogues Clin Neurosci 2008;10(3):329-336. [doi: 10.31887/DCNS.2008.10.3/dnutt] [Medline: 18979946]
20. Burton C, McKinstry B, Szentagotai Tătar A, Serrano-Blanco A, Pagliari C, Wolters M. Activity monitoring in patients with depression: a systematic review. J Affect Disord 2013 Feb 15;145(1):21-28. [doi: 10.1016/j.jad.2012.07.001] [Medline: 22868056]
21. Lamprou E, Kivelä LMM, Rohling JHT, Meijer JH, van der Does W, Antypa N. Chronotype, sleep quality, depression and pre-sleep rumination: a diary and actigraphy study. Eur J Neurosci 2024 Dec;60(11):6593-6604. [doi: 10.1111/ejn.16551] [Medline: 39367761]
22. Matcham F, Carr E, Meyer N, et al. The relationship between wearable-derived sleep features and relapse in Major Depressive Disorder. J Affect Disord 2024 Oct 15;363:90-98. [doi: 10.1016/j.jad.2024.07.136] [Medline: 39038618]
23. Zhao J, Liu H, Wu Z, et al. Clinical features of the patients with major depressive disorder co-occurring insomnia and hypersomnia symptoms: a report of NSSD study. Sleep Med 2021 May;81:375-381. [doi: 10.1016/j.sleep.2021.03.005] [Medline: 33813234]
24. Gao CX, Dwyer D, Zhu Y, et al. An overview of clustering methods with guidelines for application in mental health research. Psychiatry Res 2023 Sep;327:115265. [doi: 10.1016/j.psychres.2023.115265] [Medline: 37348404]
25. Matcham F, Barattieri di San Pietro C, Bulgari V, et al. Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): a multi-centre prospective cohort study protocol. BMC Psychiatry 2019 Feb 18;19(1):72. [doi: 10.1186/s12888-019-2049-z] [Medline: 30777041]
26. Matcham F, Leightley D, Siddi S, et al. Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD): recruitment, retention, and data availability in a longitudinal remote measurement study. BMC Psychiatry 2022 Feb 21;22(1):136. [doi: 10.1186/s12888-022-03753-1] [Medline: 35189842]
27. Simblett S, Matcham F, Siddi S, et al. Barriers to and facilitators of engagement with mHealth technology for remote measurement and management of depression: qualitative analysis. JMIR Mhealth Uhealth 2019 Jan 30;7(1):e11325. [doi: 10.2196/11325] [Medline: 30698535]
28. Rush AJ, Carmody T, Reimitz P. The Inventory of Depressive Symptomatology (IDS): clinician (IDS‐C) and self‐report (IDS‐SR) ratings of depressive symptoms. Int J Methods Psych Res 2000 Jun;9(2):45-59. [doi: 10.1002/mpr.79]
29. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
30. Mundt JC, Marks IM, Shear MK, Greist JM. The Work and Social Adjustment Scale: a simple measure of impairment in functioning. Br J Psychiatry 2002 May;180(5):461-464. [doi: 10.1192/bjp.180.5.461] [Medline: 11983645]
31. Pearce M, Garcia L, Abbas A, et al. Association between physical activity and risk of depression: a systematic review and meta-analysis. JAMA Psychiatry 2022 Jun 1;79(6):550-559. [doi: 10.1001/jamapsychiatry.2022.0609] [Medline: 35416941]
32. Zhang Y, Folarin AA, Sun S, et al. Relationship between major depression symptom severity and sleep collected using a wristband wearable device: multicenter longitudinal observational study. JMIR Mhealth Uhealth 2021 Apr 12;9(4):e24604. [doi: 10.2196/24604] [Medline: 33843591]
33. Oetzmann C, Cummins N, Lamers F, et al. Identifying depression subtypes and investigating their consistency and transitions in a 1-year cohort analysis. PLOS One 2025;20(1):e0314604. [doi: 10.1371/journal.pone.0314604] [Medline: 39808668]
34. Zhang Y, Folarin AA, Sun S, et al. Predicting depressive symptom severity through individuals' nearby Bluetooth device count data collected by mobile phones: preliminary longitudinal study. JMIR Mhealth Uhealth 2021 Jul 30;9(7):e29840. [doi: 10.2196/29840] [Medline: 34328441]
35. Sun S, Folarin AA, Zhang Y, et al. Challenges in using mHealth data from smartphones and wearable devices to predict depression symptom severity: retrospective analysis. J Med Internet Res 2023 Aug 14;25:e45233. [doi: 10.2196/45233] [Medline: 37578823]

XSL•FO
RenderX

36. Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. Bioinformatics 2001 Jun;17(6):520-525. [doi: 10.1093/bioinformatics/17.6.520] [Medline: 11395428]
37. Reynolds D. Gaussian mixture models. In: Li SZ, Jain A, editors. Encyclopedia of Biometrics: Springer; 2009:659-663. [doi: 10.1007/978-0-387-73003-5_196]
38. Ghassem Pour S. Clustering longitudinal health data using hidden Markov models [Dissertation]. : Western Sydney University; 2014 URL: https://scispace.com/pdf/clustering-longitudinal-health-data-using-hidden-markov-cc59v8zuqp.pdf [accessed 2026-04-15]
39. Scikit-learn: machine learning in Python. scikit-learn. 2024. URL: https://scikit-learn.org/stable/ [accessed 2024-12-12]
40. hmmlearn. GitHub. URL: https://github.com/hmmlearn/hmmlearn [accessed 2024-12-12]
41. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12(85):2825-2830 [FREE Full text]
42. Python Software Foundation. 2025. URL: https://www.python.org/ [accessed 2025-01-24]
43. The R project for statistical computing. R Project. 2023. URL: https://www.R-project.org/ [accessed 2026-04-15]
44. Wickham H. ggplot2: Elegant Graphics for Data Analysis: Springer-Verlag; 2016.
45. Baudry JP, Celeux G. EM for mixtures. Stat Comput 2015 Jul;25(4):713-726. [doi: 10.1007/s11222-015-9561-x]
46. Dijk DJ, Lockley SW. Integration of human sleep-wake regulation and circadian rhythmicity. J Appl Physiol (1985) 2002 Feb;92(2):852-862. [doi: 10.1152/japplphysiol.00924.2001] [Medline: 11796701]
47. Pandi-Perumal SR, Monti JM, Burman D, et al. Clarifying the role of sleep in depression: a narrative review. Psychiatry Res 2020 Sep;291:113239. [doi: 10.1016/j.psychres.2020.113239]
48. Plante DT. The evolving nexus of sleep and depression. Am J Psychiatry 2021 Oct 1;178(10):896-902. [doi: 10.1176/appi.ajp.2021.21080821] [Medline: 33171111]
49. Caliandro R, Streng AA, van Kerkhof LWM, van der Horst GTJ, Chaves I. Social jetlag and related risks for human health: a timely review. Nutrients 2021 Dec 18;13(12):4543. [doi: 10.3390/nu13124543] [Medline: 34960096]
50. Wong PM, Hasler BP, Kamarck TW, Muldoon MF, Manuck SB. Social jetlag, chronotype, and cardiometabolic risk. J Clin Endocrinol Metab 2015 Dec;100(12):4612-4620. [doi: 10.1210/jc.2015-2923] [Medline: 26580236]
51. Bauducco S, Richardson C, Gradisar M. Chronotype, circadian rhythms and mood. Curr Opin Psychol 2020 Aug;34:77-83. [doi: 10.1016/j.copsyc.2019.09.002] [Medline: 31786494]
52. Chauhan S, Norbury R, Faßbender KC, Ettinger U, Kumari V. Beyond sleep: a multidimensional model of chronotype. Neurosci Biobehav Rev 2023 May;148:105114. [doi: 10.1016/j.neubiorev.2023.105114] [Medline: 36868368]
53. Seo JY, Yeom JW, Cho CH, et al. The relationship between morningness-eveningness and mood symptoms and quality of life in euthymic state of mood disorders: results from a prospective observational cohort study. J Affect Disord 2022 Nov 1;316:10-16. [doi: 10.1016/j.jad.2022.07.072] [Medline: 35940376]
54. Robillard R, Naismith SL, Smith KL, et al. Sleep-wake cycle in young and older persons with a lifetime history of mood disorders. PLOS ONE 2014;9(2):e87763. [doi: 10.1371/journal.pone.0087763] [Medline: 24586290]
55. Robillard R, Carpenter JS, Rogers NL, et al. Circadian rhythms and psychiatric profiles in young adults with unipolar depressive disorders. Transl Psychiatry 2018 Oct 9;8(1):213. [doi: 10.1038/s41398-018-0255-y] [Medline: 30301878]
56. Dollish HK, Tsyglakova M, McClung CA. Circadian rhythms and mood disorders: time to see the light. Neuron 2024 Jan 3;112(1):25-40. [doi: 10.1016/j.neuron.2023.09.023] [Medline: 37858331]
57. Park H, Lee HK, Lee K. Chronotype and suicide: the mediating effect of depressive symptoms. Psychiatry Res 2018 Nov;269:316-320. [doi: 10.1016/j.psychres.2018.08.046] [Medline: 30172189]
58. Merikanto I, Kronholm E, Peltonen M, Laatikainen T, Vartiainen E, Partonen T. Circadian preference links to depression in general adult population. J Affect Disord 2015 Dec 1;188:143-148. [doi: 10.1016/j.jad.2015.08.061] [Medline: 26363264]
59. Pillai V, Steenburg LA, Ciesla JA, Roth T, Drake CL. A seven day actigraphy-based study of rumination and sleep disturbance among young adults with depressive symptoms. J Psychosom Res 2014 Jul;77(1):70-75. [doi: 10.1016/j.jpsychores.2014.05.004] [Medline: 24913345]
60. Solelhac G, Berger M, Strippoli MPF, et al. Objective polysomnography-based sleep features and major depressive disorder subtypes in the general population. Psychiatry Res 2023 Jun;324:115213. [doi: 10.1016/j.psychres.2023.115213] [Medline: 37098299]
61. Noetel M, Sanders T, Gallardo-Gómez D, et al. Effect of exercise for depression: systematic review and network meta-analysis of randomised controlled trials. BMJ 2024 Feb 14;384:e075847. [doi: 10.1136/bmj-2023-075847] [Medline: 38355154]
62. Machaczek KK, Allmark P, Goyder E, et al. A scoping study of interventions to increase the uptake of physical activity (PA) amongst individuals with mild-to-moderate depression (MMD). BMC Public Health 2018 Mar 21;18(1):392. [doi: 10.1186/s12889-018-5270-7] [Medline: 29562904]
63. Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. Accuracy of wristband Fitbit models in assessing sleep: systematic review and meta-analysis. J Med Internet Res 2019 Nov 28;21(11):e16273. [doi: 10.2196/16273] [Medline: 31778122]
64. Liang Z, Chapa-Martell MA. Accuracy of Fitbit wristbands in measuring sleep stage transitions and the effect of user-specific factors. JMIR Mhealth Uhealth 2019 Jun 6;7(6):e13384. [doi: 10.2196/13384] [Medline: 31172956]

XSL•FO
RenderX

65. RADAR-CNS (Remote Assessment of Disease and Relapse – Central Nervous System). King's College London. URL: https://www.kcl.ac.uk/research/radarcns [accessed 2026-04-18]
66. Unsupervised-clustering-of-wearable-derived-behavioural-features. GitHub. 2025. URL: https://github.com/carolinoetz/Unsupervised-Clustering-of-Wearable-Derived-Behavioural-Features [accessed 2026-04-15]

Abbreviations

***DSM-5*:** *Fifth Edition of the Diagnostic and Statistical Manual*
GAD-7: 7-item Generalized Anxiety Disorder
GMM: Gaussian mixture model
HMM: hidden Markov model
***ICD-11*:** *Eleventh Revision of the International Classification of Diseases*
IDS-SR: Inventory of Depressive Symptomatology–Self Report
LL: log-likelihood
MDD: major depressive disorder
RADAR-MDD: Remote Assessment of Disease and Relapse Study in Major Depressive Disorder study
RMT: remote measurement technology
WSAS: Work and Social Adjustment Scale

Edited by M Balcarras; submitted 31.Oct.2025; peer-reviewed by D Hidalgo-Mazzei, M Carminati; revised version received 08.Mar.2026; accepted 01.Apr.2026; published 10.Jun.2026.

Please cite as:
Oetzmann C, Zhang Y, Cummins N, Carr E, Matcham F, Siddi S, Lamers F, Leightley D, White KM, Folarin AA, Annas P, Haro JM, Penninx BWJH, Vairavan S, Wykes T, Dobson RJB, Narayan VA, Hotopf M
Sleep and Activity Patterns in Depression From Wearable Data: Unsupervised Clustering Study
J Med Internet Res 2026;28:e86900
URL: https://www.jmir.org/2026/1/e86900
doi:10.2196/86900

© Carolin Oetzmann, Yuezhou Zhang, Nicholas Cummins, Ewan Carr, Faith Matcham, Sara Siddi, Femke Lamers, Daniel Leightley, Katie M White, Amos A Folarin, Peter Annas, Josep Maria Haro, Brenda W J H Penninx, Srinivasan Vairavan, Til Wykes, Richard J B Dobson, Vaibhav A Narayan, Matthew Hotopf. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Rehabilitation Following Ultrasound-Guided Injection for Chronic Rotator Cuff Injury: Randomized Controlled Trial

Zifu He[1,2*], MS; Yuemin Zhang[2*], MS; Liyan Shi[2], MS; Gaineng Yuan[2], BS; Xuefeng Liu[2], BS; Yue Cai[2], BS; Jue Zhao[2], BS; Jing Xu[2], MD

[1]School of Gongli Hospital Medical Technology, University of Shanghai for Science and Technology, Shanghai, China
[2]Pudong Gongli Hospital, Shanghai University of Medicine & Health Sciences, No. 219, Miaopu Road, Shanghai, China
[*]these authors contributed equally

Corresponding Author:
Jing Xu, MD
Pudong Gongli Hospital, Shanghai University of Medicine & Health Sciences, No. 219, Miaopu Road, Shanghai, China

Abstract

Background: Chronic rotator cuff injuries (CRCIs) often lead to shoulder pain and dysfunction. Although ultrasound-guided injections provide rapid pain relief, subsequent rehabilitation efficacy is often compromised by poor exercise adherence and inaccurate movement execution. Digital therapeutics may enhance patient engagement and movement accuracy, potentially improving outcomes.

Objective: This study compared the effects of ultrasound-guided corticosteroid plus local anesthetic injection combined with the Digital Rehabilitation System training versus traditional exercise on pain intensity and shoulder function in patients with CRCIs.

Methods: This was a single-center, parallel-group, superiority randomized controlled trial with a blinded outcome assessor, injection physician, and data analyst. Sixty eligible participants were randomized to either the conventional treatment group (group C, n=30) or the digital treatment group (group D, n=30) for a 3-month intervention. Following pain-alleviating ultrasound-guided injections, group C performed prescribed exercises, while group D used the intelligent system for self-corrective adaptive training. Primary outcome was the Constant-Murley Score. Secondary outcomes included the University of California Los Angeles Shoulder Scale, Numerical Rating Scale for pain, range of motion (ROM), and system proficiency scores. Assessments occurred at baseline (T1), 1-week (T2), 1-month (T3), and 3-month (T4) follow-ups.

Results: No significant between-group differences existed in baseline characteristics or outcome measures (all $P>.05$). No group × time interactions were observed for any outcomes. At T4, group D demonstrated significantly greater improvements in Constant-Murley Score (Hodges-Lehmann estimator [HLE] −6, 95% CI −10 to −2; $P=.007$), University of California Los Angeles Shoulder Scale score (HLE −2, 95% CI −4 to −1; $P=.001$), and Numerical Rating Scale score (HLE 1, 95% CI 0-2; $P=.002$), and their improvement values (Δ) all exceeded the corresponding minimal clinically important differences. ROM also showed significant intergroup differences in all planes. Compared with T1, ROM was significantly improved in both groups at T4, with the greatest improvement in abduction (Δ group C=20.7 and Δ group D=30.5). Subgroup analysis revealed better enhanced shoulder function and ROM in participants with no or mild shoulder motion restrictions using digital therapy.

Conclusions: For participants with CRCIs, combining ultrasound-guided injection with subsequent Digital Rehabilitation System training achieves greater shoulder ROM and functional scores than traditional exercise after pain relief. This approach demonstrates enhanced efficacy in subgroups with no or mild shoulder mobility limitations.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2500097903; https://tinyurl.com/y52mm5ry

(J Med Internet Res 2026;28:e79494) doi:10.2196/79494

KEYWORDS

rotator cuff injury; digital therapeutics; rehabilitation exercise; injection therapy; system proficiency

Introduction

Chronic rotator cuff injury (CRCI) is the primary cause of shoulder pain and dysfunction, and its prevalence is increasing worldwide in the occupational and geriatric populations [1]. When rotator cuff tendon lesions occur, the shoulder's stability is disrupted, often leading to persistent pain, limited range of motion (ROM), and reduced muscular strength [2]. If untreated, CRCI can also lead to shoulder stiffness and muscle atrophy, which further exacerbate shoulder dysfunction. Although ultrasound-guided injection therapy has been shown to facilitate tissue repair by providing targeted relief of acute pain, the

XSL•FO
RenderX

medium- and long-term improvements in shoulder function are still not promising [3-5]. As a core intervention in the recovery from CRCI [6,7], the effectiveness of exercise is significantly correlated with movement accuracy and compliance [8]. Studies have shown that systematic rehabilitation not only improves shoulder mobility and muscle coordination but also demonstrates a more sustained analgesic effect than local injection therapy, reducing the risk of disability [7,9]. However, based on paper-based guidelines and follow-ups, conventional rehabilitation often leads to secondary injuries or delayed recovery due to adherence to incorrect exercise patterns.

In recent years, the rapid development of digital therapeutics (DTx) has provided innovative solutions in exercise rehabilitation [10-12]. Motion capture technology allows real-time acquisition of participants' joint kinematic parameters through multisensor fusion with deep learning algorithms. Recently, a study demonstrated that digital interventions using inertial motion trackers integrated with mobile apps exhibit comparable efficacy with traditional in-person supervised programs for patients with shoulder pain, with DTx users showing more substantial clinical improvements in shoulder function [13]. Emerging evidence further validates the effectiveness of digital protocols as novel therapeutic modalities, demonstrating superior efficacy in alleviating pain and associated disability [6,11,14].

However, most existing studies focus on postoperative rehabilitation or acute injury. There is still a gap in the standardization of movement performance for participants with CRCI [15]. This study aims to develop disease-adaptive intelligent assessment models to achieve motion modification based on participants' movement data and to evaluate the clinical value of ultrasound-guided injection therapy combined with digital therapy for CRCI. The authors hypothesized that this digital protocol after injection therapy would result in a more significant improvement in pain as well as function in participants with CRCI than conventional methods.

Methods

Study Design

This study was conducted in accordance with the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) guidelines (Checklist 1) [16]. This prospective, single-center, randomized controlled trial used a 2-arm parallel design. Eligible participants were randomly assigned to either the conventional treatment group (group C) or the digital treatment group (group D) for a 3-month intervention period (March 1 to June 17, 2025). The blinding status (masking of the injection physicians, the outcome assessor, and the data analyst) was maintained until trial completion. Only the investigator generating the randomization sequence and administering the Digital Rehabilitation System was unblinded to group allocation.

Sample Size Calculation

Prior to recruitment, the minimum sample size was calculated by G*Power 3.1.9.7 to ensure the rigor of statistical analysis [17]. Based on previous studies, the Constant-Murley Score (CMS) was used as the primary outcome in this study. The sample size was calculated based on the between-group difference in the improvement values (Δ) from baseline to postintervention. The minimum difference value between the 2 groups was 5.1 (SD 5.2), resulting in an effect size (d) of 0.98 [18]. By using a 2-tailed t test (taking α=.05, test efficacy=0.90, an allocation ratio of 1:1, and a 20% dropout rate), a minimum of 58 individuals needed to be included.

Participants

Participants were recruited at the outpatient pain clinic of Pudong Gongli Hospital, Shanghai University of Medicine & Health Sciences. Both groups were required to receive injections and to undergo rehabilitation after therapeutic patient education. Group C underwent home rehabilitation, while group D underwent rehabilitation through the Digital Rehabilitation System along with home exercises.

Eligible participants comprised adults aged 18 years or older with intact cognitive function and athletic ability who met the diagnostic criteria for rotator cuff injury (RCI) [19,20]. Inclusion required unilateral symptomatic presentation with pain intensity ≥3 on the Numerical Rating Scale (NRS) persisting for at least 3 months, along with a Neer classification of stage II or less [21] diagnosed by magnetic resonance imaging. Exclusion criteria comprised full-thickness tears, active infections, neoplastic conditions, concomitant pathologies affecting shoulder function (eg, shoulder fracture, glenoid labral injury, osteoarthritis, etc), prior ipsilateral shoulder surgery, hypersensitivity to interventional medications, pregnancy or lactation, and individuals with severe audiovisual or motor impairment.

Additionally, participants were stratified into subgroups based on the baseline active range of motion (AROM) of shoulder flexion and abduction, given their primary importance for upper limb function. The classification used clinically meaningful thresholds informed by functional requirements. Individuals with an AROM of ≥150° for both flexion and abduction movements were classified as having no or mild functional limitation (subgroup 1), as this range is typically sufficient for essential overhead activities [22]. Severe limitation (subgroup 2) was defined as an AROM of ≤90° in either flexion or abduction, a threshold associated with substantial disability in performing basic activities of daily living. Participants who did not meet the criteria for subgroup 1 or 2 were classified as having moderate limitation (subgroup 3).

Allocation and Blinding

Researcher 1 (JX) generated and recorded 60 numbers using the RANDBETWEEN function before the start of the trial and used these numbers to determine a concealed allocation sequence (based on the size of the numbers). This sequence was prepared in a set of sequentially numbered, opaque, sealed envelopes by researcher 1. Researcher 2 (ZJ) was responsible for receiving participants, determining their eligibility, and administering medications with musculoskeletal ultrasonography. After researcher 2 confirmed a participant's eligibility, researcher 1 opened the next consecutively numbered

XSL•FO
RenderX

envelope in the presence of the participant to reveal and execute the allocation. To ensure objectivity and eliminate subjective influence, all outpatient training was monitored by researcher 1, who remained passive and did not intervene in the exercise process. The other outcome indicators were evaluated and recorded by researcher 3 (ZH). The anonymized data were processed and analyzed by researcher 4 (YC). To ensure the objectivity of the results, all investigators outside of researcher 1 were blinded until the end of the study.

System Equipment

The system was developed by Pudong Gongli Hospital and Ruiku Medical Technology Co, Ltd. It is divided into 2 subsystems: the Virtual Reality (VR) Education subsystem (VRS) (Figure 1A and B) and the Motion Capture subsystem (MCS) (Figure 1C-F). All participants used the VRS for therapeutic patient education and a butterfly capture task, which involved video tutorials about the function of the rotator cuff, the symptoms and causes of RCI, and the methods of exercises [23]. And they needed to use the MCS for training. The background in MCS includes 3 partitions of a virtual coach, an avatar, and a training scene, which serve to guide the exercise, provide real-time feedback, and avoid danger, respectively.

The VRS module was built around a PICO 4 Ultra mixed-reality head-mounted display (PICO, ByteDance Ltd). This device is powered by a Qualcomm Snapdragon XR2 Gen 2 processor and is equipped with a comprehensive sensor array, including dual 32-megapixel RGB cameras and an iToF depth-sensing camera for high-fidelity environmental tracking and stereoscopic visual rendering. The headset provided participants with synchronized audiovisual stimuli. For researcher monitoring and intervention, a Xiaomi Pad 5 tablet (Xiaomi Inc), featuring a Qualcomm Snapdragon 860 platform, was used to run dedicated supervisory software, allowing real-time observation of the participant's first-person perspective.

The MCS module consisted of 2 primary components: an Azure Kinect Developer Kit (DK) camera (Microsoft Corp) and a Lenovo Legion Y7000 IRX9 laptop computer (Lenovo Group Ltd). The Azure Kinect DK, which features a 12-megapixel RGB camera and a 1-megapixel depth camera, served as the primary sensor for capturing high-precision shoulder joint kinematic data. The laptop, running the Windows 11 operating system and equipped with an Intel Core i7-13650HX processor and NVIDIA GeForce RTX 4060 GPU, hosted the custom analysis software built on the Unity platform. This system compared the captured motion data against a predefined standard movement model in real time and generated immediate graphical and auditory feedback to guide participants in correcting aberrant movement patterns.

XSL•FO
RenderX

Figure 1. Operating Interface of the Digital Rehabilitation System: (A) therapeutic patient education, (B) butterfly capture task, (C) exercise prescription selection, (D) pretraining assessment for grading, (E) digital training, and (F) training results.

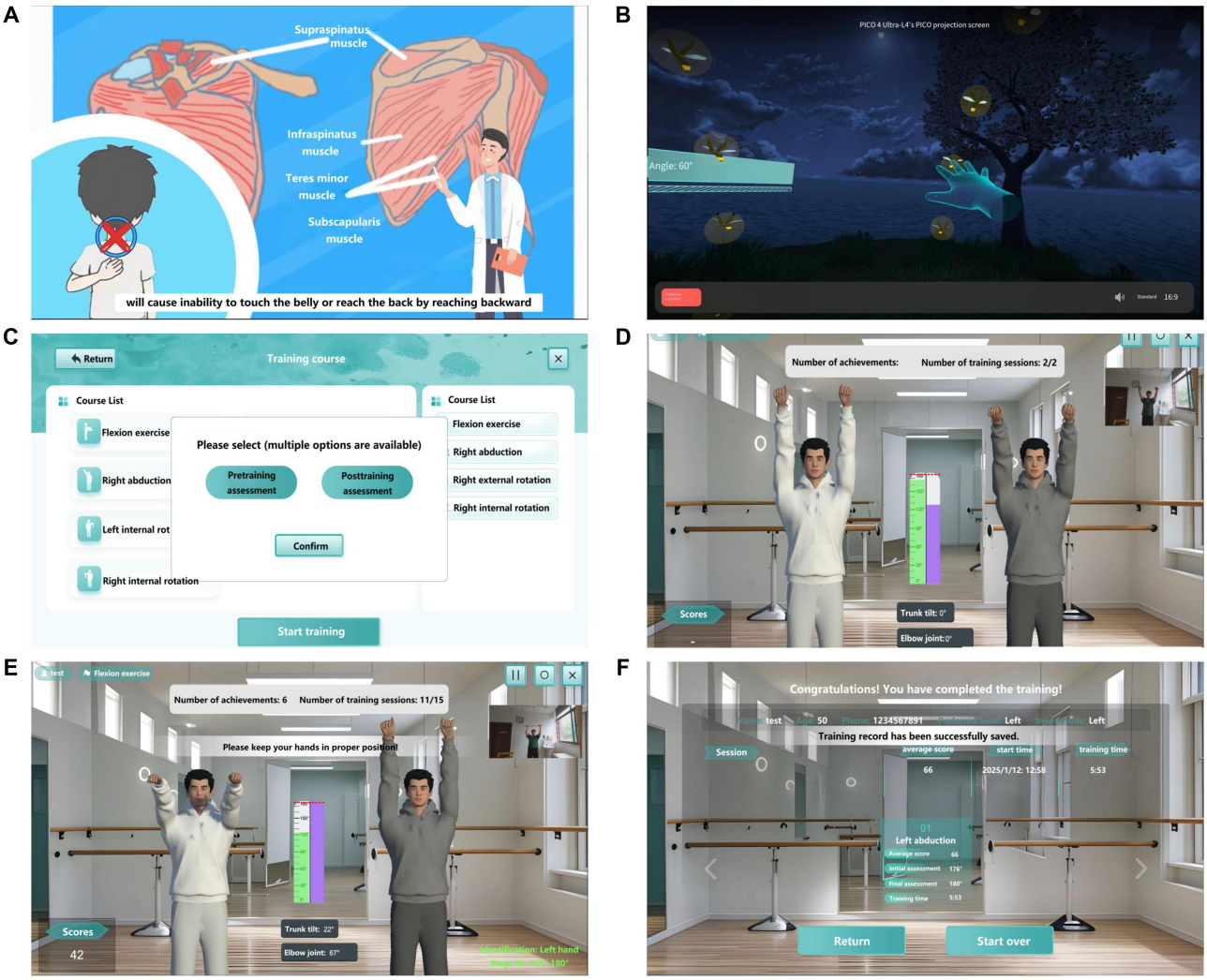

Intervention

Overview

All participants received injections administered by the same physician team. The injectate was prepared by first combining 0.25 mL of compound betamethasone injection (containing 1.25 mg of betamethasone sodium phosphate and 0.5 mg of betamethasone dipropionate) with 2.5 mL of 2% lidocaine to form a base solution [3,24]. This base solution was then diluted with 0.9% saline to a total volume of 10 mL. Subsequently, 1 mL of the mixture was discarded, and the remaining 9 mL was used for injection, resulting in a final lidocaine concentration of 0.5%. Based on the magnetic resonance imaging diagnosis, pathological tissues (including structural lesions and inflammatory regions) were precisely targeted under musculoskeletal ultrasound guidance. Participants were positioned supine for subcoracoid bursa and intertubercular groove injections and then placed in a contralateral decubitus position for subacromial bursa injection [3,25,26], administering 3 mL per site. Postprocedural care included sterile dressing application at needle entry points, with instructions to maintain site dryness and avoid strenuous activity for 24 hours. All individuals received 3 injections at 1-week intervals to minimize the potential for corticosteroid-related tendinopathy [27].

Flexion, abduction, external rotation, and internal rotation exercises were prescribed for all participants, with each session lasting 10-15 minutes and administered 3 sessions per day [28]. In addition, all participants attended a weekly follow-up appointment at the outpatient clinic. These sessions were conducted by the same designated physician and physiotherapist to ensure consistency in clinical assessment. Additionally, researcher 1 administered a standardized, interactive session using the VRS to enhance participants' understanding of their condition and the rehabilitation principles. Separately, participants also completed one of their 3 daily training sessions under supervision at the outpatient clinic on weekdays. The remaining 2 daily sessions were conducted independently at home, performing conventional exercises without the digital system. Participants in group C exercised within the range of pain tolerance.

Flexion Training

Participants held a booster stick in both hands and made the shoulder joint flex forward. When the affected shoulder could not continue to flex forward due to weakness or adhesion, the unaffected one assisted or passively lifted it up to its maximal AROM and slowly lowered it down after 15 - 30 seconds.

XSL•FO
RenderX

Abduction Training

The method was in accordance with flexion training. At the end of AROM, the unaffected arm forcefully retracted to provide assistance, prompting the affected shoulder to the maximum angle within the tolerable range of pain.

External and Internal Rotation Training

The unaffected hand applied downward pressure on the head of the affected humerus, and participants were asked to exercise while the affected shoulder was kept in 90° abduction with the elbow flexed to 90°.

Participants allocated to group D used the custom-developed rehabilitation system for their training regimen. Prior to each session, participants performed 2 maximal ROM trials in specified directions to determine their appropriate training phase. Subsequently, the system automatically assigned them to a corresponding phase for 1 exercise set. The exercise modalities within the system were the same as the movements performed by group C. Real-time visual feedback, displayed as a discrepancy between an on-screen coach avatar and the participant's own motion capture, along with a performance score, allowed for immediate self-correction to enhance movement accuracy [29]. The protocol incorporated an automatic progression rule. This rule stipulated that the system would advance a participant to the subsequent phase once their joint angle consistently met the threshold for the next level [30]. Each set comprised 15 movement repetitions. Following set completion, participants underwent 2 reassessment trials. The total time for 1 system-guided set was 15 minutes. Participants attended 1 in-clinic session on weekdays, supplemented by home-based exercises, for a total of 3 sets daily.

Outcome Measures

NRS, CMS, and University of California Los Angeles (UCLA) Shoulder Scale scores, and ROMs of forward flexion, abduction, and external and internal rotation of the affected shoulder were collected from each one by the same personnel at baseline (T1), 1-week follow-up (T2), 1-month follow-up (T3), and 3-month follow-up (T4) [31]. In addition, there was a need to score the rehabilitation movements in both groups using the Digital Rehabilitation System at T4. Upon reasonable request, the data supporting the following findings of this study can be obtained from the first author (ZH).

CMS was the primary outcome in this study, with a total score of 100. A higher score represents better function [32]. The scale assessed 4 domains, including pain level, ability to perform activities of daily living (ADL), AROM, and shoulder abduction strength. Participants state the severity of pain and limitations in ADL, and the remaining items are measured by the physiotherapist's examination. Shoulder abduction strength was specifically assessed using the manual muscle testing technique [33]. Given that all participants had unilateral involvement, strength was evaluated concurrently on both the affected and unaffected sides. This side-to-side comparison during simultaneous abduction allowed for a relative grading of strength deficit, mitigating potential bias from interindividual variability in tester application.

UCLA Shoulder Scale is a system of functional assessment of surgical or conservative treatment of the shoulder with a total score of 35. It consists of pain, usable function of the upper extremity, ROM, and muscle strength of flexion, and patient satisfaction, with higher scores indicating better recovery. The scale focuses more on the impact of pain on daily life than the CMS. The muscle strength of flexion is also assessed using the manual muscle testing [33]. This outcome was not included in the trial registry and is reported in this manuscript as an exploratory outcome.

NRS uses a total of 11 numbers from 0 to 10 to show the severity of pain. After explaining the pain represented by each number, participants reported the highest level of pain experienced during the performance of daily activities in the week preceding the assessment time point, capturing the peak impact of pain on functional status. A score of 0 indicates no pain; 1 - 3 is mild pain, and the pain does not interfere with sleep; 4 - 6 is moderate pain that interferes with sleep; and 7 - 10 is severe pain that prevents sleep [32].

Researcher 3 measured the shoulder ROM of flexion, abduction, external rotation, and internal rotation using a protractor in strictly standardized positions after excluding compensatory movements or interactions of neighboring joints [5,25].

Additional evaluated measures included treatment adherence, adverse events, and a system-derived performance metric. Adherence was tracked through a participant-maintained session log and verified during weekly clinic visits. All adverse events were actively monitored and documented throughout the study period. This included patient self-reports and standardized assessments using the Simulator Sickness Questionnaire [34] and the Virtual Reality Sickness Questionnaire [35]. Furthermore, the system generated a real-time score to improve the movement accuracy. This metric, henceforth termed system proficiency, provided intrasession biofeedback to guide movement correction and was not interpreted as a direct clinical measure of proprioception.

Statistical Analysis

To determine statistical significance between 2 groups, continuous variables with a normal distribution were expressed as mean (SD), while non–normally distributed data were reported as median (IQR). Categorical data were presented as frequency (percentage) and compared using chi-square tests. For normally distributed data with homogeneity of variance, independent samples *t* tests were used for intergroup comparisons, while Welch *t* tests were used for heterogeneous variances. Friedman tests analyzed overall differences in non–normal data, with post hoc comparisons via the Conover-Iman test or Dunn multiple comparison test. Wilcoxon signed-rank tests assessed repeated measures, while Mann-Whitney *U* tests compared intergroup differences. As primary data violated normality assumptions, Scheirer-Ray-Hare tests (using an SPSS-validated alternative) evaluated main and interaction effects. Kruskal-Wallis H tests compared treatment efficacy across subgroups. Additionally, all analyses incorporated the Bonferroni correction. Data were analyzed using SPSS software (version 26.0; IBM Corp), with statistical significance set at a 2-tailed $P<.05$ and 95% CI.

XSL•FO
RenderX

Ethical Considerations

This study was approved by the Medical Ethics Committee of Gongli Hospital, Pudong New Area, Shanghai, China (approval no. GLYY1s2024-031). Eligible participants were fully informed about the study's purpose, procedures, potential risks, and benefits through verbal and written explanations. All participants provided written informed consent prior to enrollment. A unique number assigned to each participant served as an individual identifier, and all subsequent data analyses were performed exclusively on a deidentified dataset. Patients were not involved in the design or interpretation of the study results but contributed to data collection through their clinical outcomes. Although no participation compensation was provided, participants were entitled to receive appropriate alternative medical care—such as medication, shock wave therapy, or conventional rehabilitation exercises—if they experienced any adverse reactions (eg, allergy or dizziness) to the treatment protocol.

Results

Participant Characteristics

In this study, 151 participants were screened for eligibility, and a total of 60 participants met the inclusion criteria. The flowchart of participants in the trial, the screening, and trial processes are illustrated in Figure 2 [16]. 7 individuals were found to meet more than 1 exclusion criterion concurrently. One participant was both unable to participate and had a comorbid condition; 1 had a full-thickness tear, was in the acute injury phase, and declined the injection; 1 had a comorbid condition, was in the acute phase, and declined treatment; 2 had bilateral shoulder pain and comorbid conditions; 1 had bilateral shoulder pain and was in the acute phase; and 1 had bilateral shoulder pain, was in the acute phase, and declined treatment. The baseline characteristics of the 2 groups were well balanced, and statistical analysis revealed that no significant differences were observed between the 2 groups in terms of gender, age, disease duration, or side of injury ($P>.05$), and the results of the treatment were comparable (Table 1).

XSL•FO
RenderX

Figure 2. Flowchart of the randomized controlled trial in patients with chronic rotator cuff injury at Pudong Gongli Hospital, Shanghai University of Medicine & Health Sciences, Shanghai, China, 2025. Seven patients met multiple exclusion criteria; they are represented in different applicable categories but counted only once in the total (n=91). CMS: Constant-Murley Score; UCLA: University of California Los Angeles Shoulder Scale; ROM: range of motion; NRS: Numerical Rating Scale.

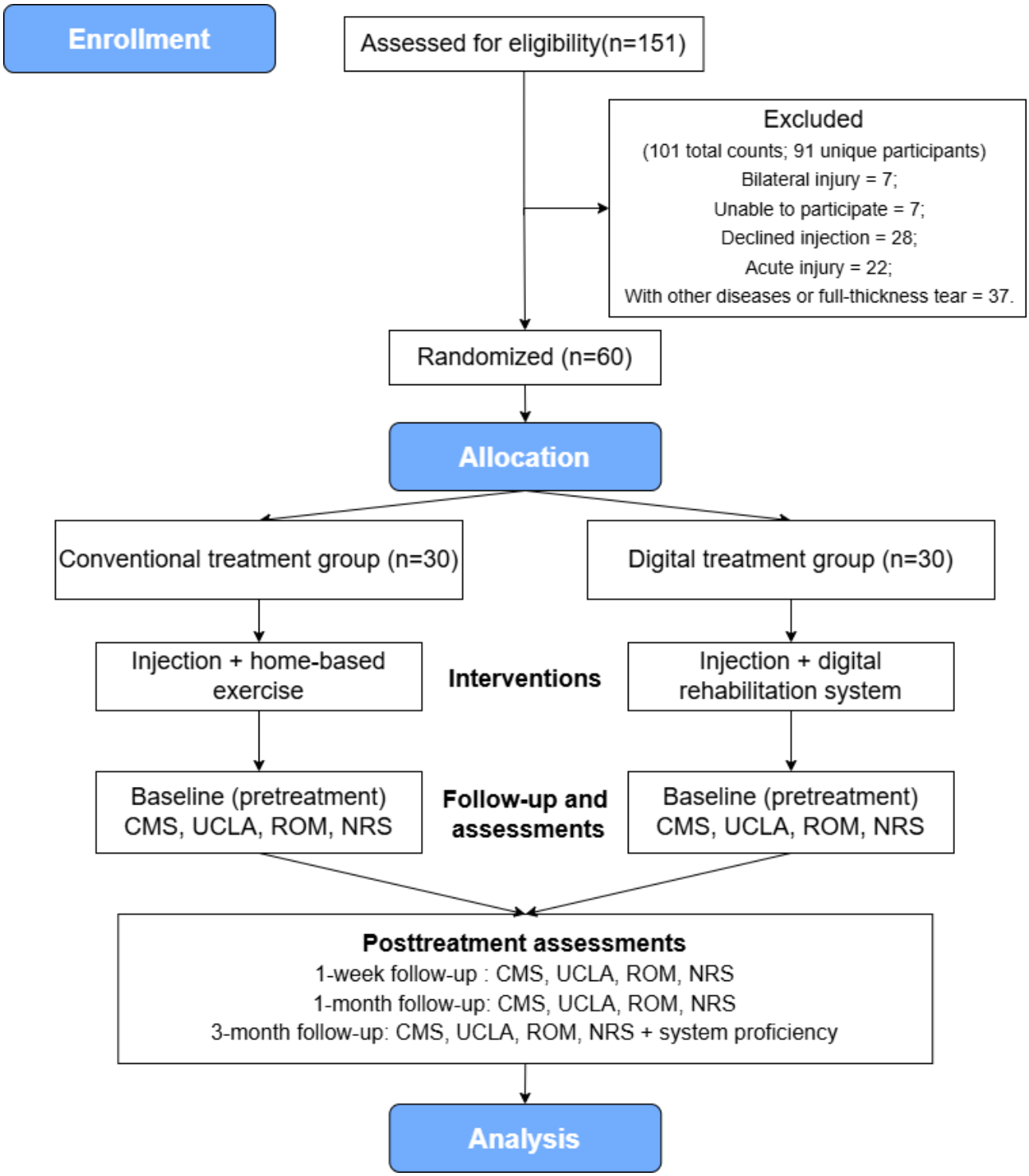

XSL•FO
RenderX

Table . Demographic and clinical characteristics of the 2 groups in a randomized controlled trial focusing on rotator cuff injuries at Pudong Gongli Hospital, Shanghai University of Medicine & Health Sciences, Shanghai, China, 2025.

Group	Gender		Injured side		Age (years), median (P25- P75)	Duration (months), median (P25-P75)
	Man, n (%)	Woman, n (%)	Left, n (%)	Right, n (%)		
Group C[a]	10 (33.3)	20 (66.7)	8 (26.7)	22 (73.3)	60 (52.25-69.25)	3 (3-6)
Group D[b]	9 (30)	21 (70)	9 (30)	21 (70)	63 (47.75-70)	3 (3-6)
P value	.781		>.99[c]		.976	.552

[a]Group C: conventional treatment group.

[b]Group D: digital treatment group.

[c]Fisher exact test was applied, resulting in *P*>.99. Pearson chi-square test yielded *P*=.774. All expected cell counts were >5 (minimum expected count = 8.50).

Shoulder Functional Performance

There was a significant time main effect (Kruskal-Wallis H=61.27, *P*<.001) and between-groups main effect (Kruskal-Wallis H=5.46, *P*=.019) for CMSs. For UCLA Shoulder Scale scores, a significant main effect of time was observed (Kruskal-Wallis H=116.91, *P*<.001). At T1, there was no statistically significant difference between the total CMS (*P*=.859) and UCLA Shoulder Scale scores (*P*=.405) of the 2 groups. At T4, the difference between the CMSs of the groups was even more significant (Hodges-Lehmann estimator [HLE] −6, 95% CI −10 to −2; *P*=.007). Compared with T1, both groups demonstrated significant CMS improvements at T4 (group C: median of difference [MD] 13.5, 95% CI 10.5-16.5, *P*<.001; group D: MD 19.0, 95% CI 14.5-22.5, *P*<.001). Notably, over half of group D (19/30, 63.3%) achieved "excellent" outcomes per Figure 3A substantially exceeding group C (10/30, 33.3%). UCLA Shoulder Scale scores similarly showed significant intergroup differences at T3 (HLE −2, 95% CI −3 to 0; *P*=.034) and T4 (HLE −2, 95% CI −4 to −1; *P*=.001). At T4, the number of participants rated as excellent in group D (11/30, 36.7%) exceeded that in group C (1/30, 3.3%) (Figure 3B). At T4, the improved scores of shoulder function were greater than the minimal clinically important difference (MCID) for CMS (7.2 points) and UCLA Shoulder Scale (3.3 points) (Table S1 in Multimedia Appendix 1) [36].

Figure 3. Proportion of participants with shoulder function scores at baseline and 3-month follow-up in a randomized controlled trial focusing on rotator cuff injuries at Pudong Gongli Hospital, Shanghai University of Medicine & Health Sciences, Shanghai, China, 2025. (A) Changes in the Constant-Murley Score for the 2 groups at baseline and 3-month follow-up (excellent: 90 - 100, complete functional recovery; good: 80 - 89, significant functional recovery; fair: 70 - 79, partial improvement; and poor:<70, poor outcome). (B) Changes in the University of California Los Angeles Shoulder Scale for the 2 groups at baseline and 3-month follow-up (excellent, 34 - 35; good, 29 - 33; and poor,<29).

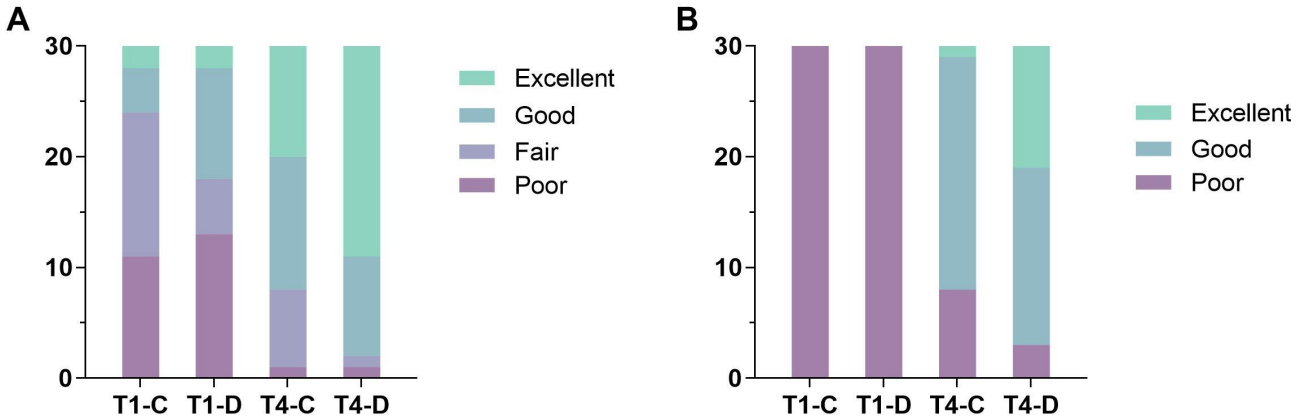

Pain-Related Outcomes

Statistical analysis revealed significant main effects of both group (Kruskal-Wallis H=5.15, *P*=.023) and time (Kruskal-Wallis H=108.48, *P*<.001) on NRS scores. However, no significant group × time interaction was observed (Kruskal-Wallis H=7.38, *P*=.061). As shown in Table S1 in Multimedia Appendix 1, no statistically significant between-group differences in NRS scores were observed at T1 (*P*=.458). At T3, group D demonstrated a significant pain reduction from baseline (median 5, IQR 4 - 6 to median 2, IQR 1 - 2.25; *P*<.001; MD −3.5, 95% CI −3.5 to −3). This improvement further progressed at T4 (median 0.5, IQR 0 - 2; *P*<.001; MD −4.0, 95% CI −4.5 to −3.5). Similarly, group C showed pain reduction at T3 (median 5, IQR 3 - 6.25 to median 3, IQR 2 - 3.25; *P*<.001; MD −2.0, 95% CI −2.5 to −1.5) and T4 (median 2, IQR 1-3; *P*<.001; MD −2.5, 95% CI −3 to −2). Both groups exceeded the MCID threshold of 1.37 points [37,38], indicating therapeutic relevance for both interventions. At T4, group D reported significantly lower pain intensity than group C (HLE 1, 95% CI 0-2; *P*=.002). These findings indicate enhanced pain-alleviating efficacy of group D compared with group C.

XSL•FO
RenderX

Shoulder ROM

The shoulder's 3 rotational degrees of freedom contribute to heterogeneous motion limitations post-RCI. This study conducted analyses on ROM in 4 movement directions: flexion, abduction, external rotation, and internal rotation. Significant main effects of time were observed for all 4 ROM directions, but no time × group interactions emerged. Group effects existed for all directions except flexion. Baseline data showed no significant between-group differences in any plane (Table S1 in Multimedia Appendix 1). At T2 and T3, only the ROM of external rotation demonstrated significant intergroup differences (Figure 4I-L). By T4, group D exhibited significantly higher median ROM than group C in flexion (HLE −10, 95% CI −25 to 0; P=.023), abduction (HLE −15, 95% CI −30 to 0; P=.034), external rotation (HLE −5, 95% CI −20 to 0; P=.001), and internal rotation (HLE −15, 95% CI −30 to 0; P=.035) (Figure 4A-H). All directions showed increased ROM at T4 versus T1 (Figure 4), with maximal gains in abduction. However, transient ROM reductions occurred during training.

Figure 4. Changes within the range of motion of the 2 groups in a randomized controlled trial focusing on rotator cuff injuries at Pudong Gongli Hospital, Shanghai University of Medicine & Health Sciences, Shanghai, China, 2025. (A-D) Within-group differences in the range of motion for participants in the conventional treatment group; (E-H) within-group differences in the range of motion of the digital treatment group; and (I-L) comparisons between the 2 groups. *P<.05, **P<.01, ***P<.001.

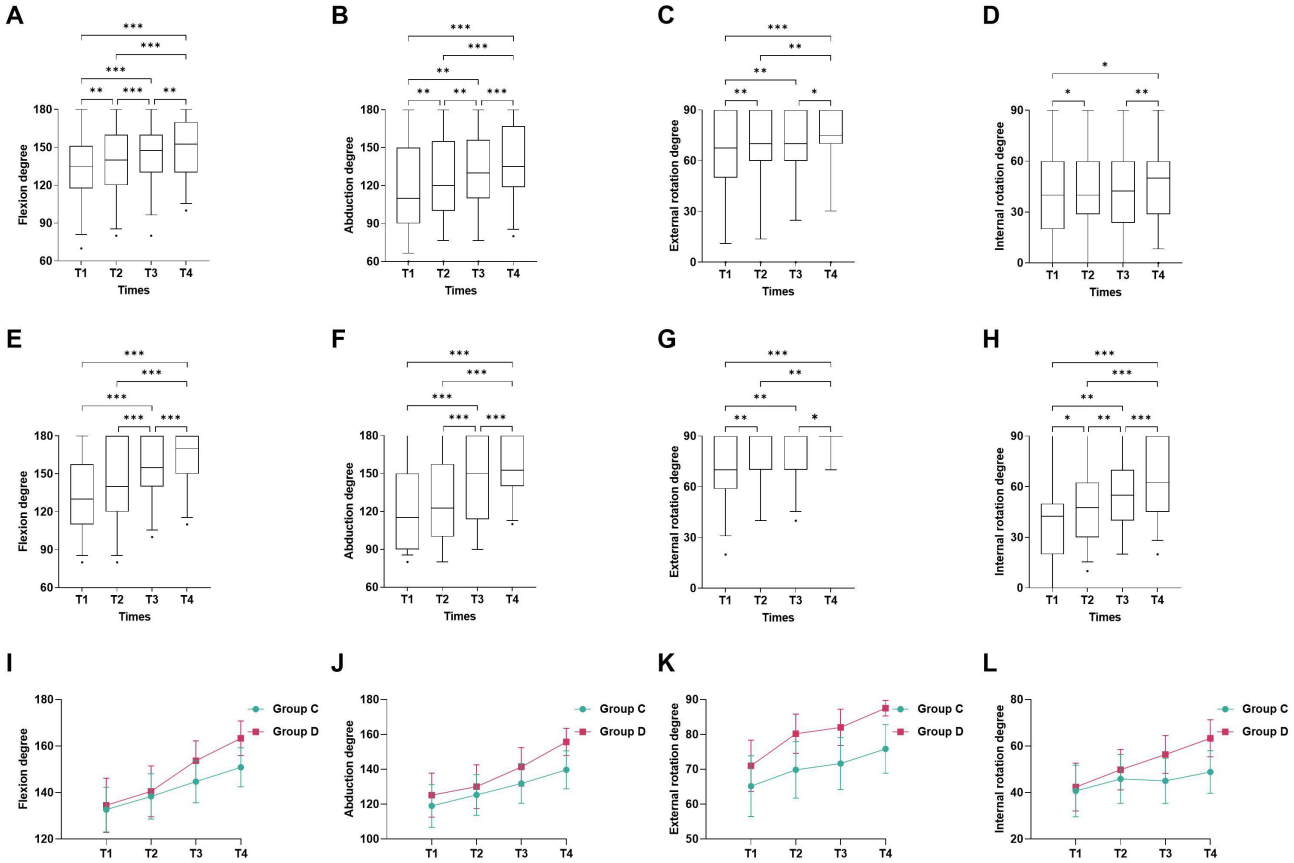

Other Findings

All patients fully complied with the prescribed protocols, and no deviations occurred during the study. No adverse events were reported or identified through patient logs and standardized questionnaires throughout the study period. The monitored outcomes specifically included, but were not limited to, muscle atrophy, abnormal blood glucose, allergic reactions, and dizziness. No participants required rescue analgesic medication during the study period. System proficiency, collected as process feedback during training, was treated as a nonclinical metric. It is presented in the final subsection of the "Results" and analyzed in the "Discussion" section.

Subgroup Analysis

The distribution of participants across these subgroups was comparable between group C and group D. As presented in Table 2 and Tables S2-S4 in Multimedia Appendix 1, the baseline demographic and clinical characteristics were well balanced across all 3 subgroups within each treatment group. This balance in baseline profiles confirms that any observed differences in treatment outcomes among the subgroups are unlikely to be attributable to preexisting disparities in these key variables.

XSL•FO
RenderX

Table . Statistical comparison of characteristics across the subgroups stratified by baseline function within each intervention group in a randomized controlled trial focusing on rotator cuff injuries at Pudong Gongli Hospital, Shanghai University of Medicine & Health Sciences, Shanghai, China, 2025.

Subgroup	Gender		Injured side		Age (years), median (P25-P75)	Duration (months), median (P25-P75)
	Man, n (%)	Woman, n (%)	Left, n (%)	Right, n (%)		
1						
Group C (n=7)[a]	1 (14.3)	6 (85.7)	1 (14.3)	6 (85.7)	64 (53-68)	6 (3-12)
Group D (n=9)[b]	3 (33.3)	6 (66.7)	1 (11.1)	8 (88.9)	62 (46.5-69.5)	3 (3-8)
P value	.372		.849		.916	.350
2						
Group C (n=9)[a]	5 (55.6)	4 (44.4)	2 (22.2)	7 (77.8)	60 (49-71.5)	3 (3-3.5)
Group D (n=8)[b]	3 (37.5)	5 (62.5)	3 (37.5)	5 (62.5)	57.5 (45-69)	3 (3-3.75)
P value	.455		.490		.586	.432
3						
Group C (n=14)[a]	4 (28.6)	10 (71.4)	5 (35.7)	9 (64.3)	60 (53-69.25)	4 (3-6)
Group D (n=13)[b]	3 (23.1)	10 (76.9)	5 (38.5)	8 (61.5)	64 (55-70.5)	3 (3-6)
P value	.744		.883		.451	.979

[a]C: conventional treatment group.

[b]D: digital treatment group.

After adjusting for baseline differences using ANCOVA, the pattern of therapeutic effects differed between group C and group D. Within group C, therapeutic effects were largely similar across the 3 subgroups, with the exception of shoulder abduction. For abduction, the homogeneity of variance assumption was violated (Levene test, $F_{2, 27}$=3.681; *P*=.039). Therefore, the Games-Howell post hoc test, which does not assume equal variances, was used. It revealed that the improvement in subgroup 1 was significantly greater than in both subgroup 2 (mean difference 56.49, 95% CI 38.33-74.66; *P*<.001) and subgroup 3 (mean difference 36.57, 95% CI 16.83-56.31; *P*<.001).

In contrast, within group D, significant subgroup differences were observed across multiple outcomes. For internal rotation, where variances were homogeneous (Levene test, $F_{2, 27}$=0.845; *P*=.441), a standard 1-way ANCOVA revealed a significant main effect of subgroup ($F_{2, 26}$=10.848; *P*<.001, η^2=0.455). Post hoc pairwise comparisons (Bonferroni-corrected) confirmed that subgroup 1 improved significantly more than both subgroup 2 (mean difference 32.16, 95% CI 13.28-51.04; *P*=.001) and subgroup 3 (mean difference 24.48, 95% CI 8.49-40.47; *P*=.002). For shoulder flexion (Levene $F_{2, 27}$=7.333; *P*=.003), UCLA Shoulder Scale score (Levene $F_{2, 27}$=4.911; *P*=.015), and shoulder abduction (Levene $F_{2, 27}$=3.776; *P*=.036), the variance homogeneity assumption was violated. Consequently, the Games-Howell post hoc tests were applied. These analyses indicated significant subgroup differences for all 3 measures (all *P*<.05, as evidenced by the subsequent pairwise comparisons). Specifically, subgroup 1 showed superior improvement to both subgroup 2 and subgroup 3 in flexion, UCLA Shoulder Scale score, and abduction (all *P*<.05, in Table S5 in Multimedia Appendix 1). Notably, for abduction, a further significant difference was found, with subgroup 3 improving more than subgroup 2 (mean difference 23.89, 95% CI 5.88-41.90; *P*=.010), establishing a clear improvement gradient: subgroup 1 > subgroup 3 > subgroup 2.

Significant improvements in CMS, UCLA Shoulder Scale, and NRS in global mixed-rank and within-group changes were observed in all subgroups at T4 compared with T1. Subgroups 2 and 3 additionally demonstrated a statistical difference in shoulder ROM (Figure 5A-C and Table S2-S4 in Multimedia Appendix 1), suggesting that combined injection-rehabilitation therapy primarily enhances mobility in restricted ones.

XSL•FO
RenderX

Figure 5. Efficacy of conventional and digital treatments in each subgroup in a randomized controlled trial focusing on rotator cuff injuries at Pudong Gongli Hospital, Shanghai University of Medicine & Health Sciences, Shanghai, China, 2025. (A) Range of motion of subgroup 1; (B) range of motion of subgroup 2; and (C) range of motion of subgroup 3. *P<.05, **P<.01, ***P<.001.

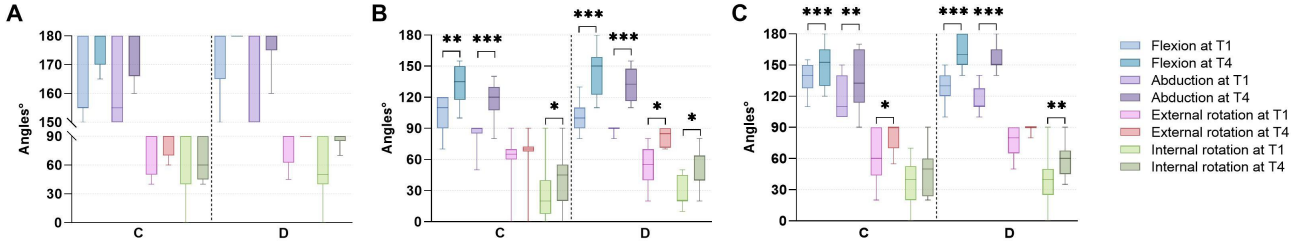

System Proficiency

At T4, group D demonstrated higher median (P25-P75) scores than group C in flexion (93, 88.75-95 vs 87.5, 77.75-92.25; P<.001), abduction (86.5, 81.5-89 vs 84, 70-86.75; P=.048), external rotation (88.5, 85.75-92.25 vs 79, 52-90; P=.009), and internal rotation (89, 83-92 vs 66.5, 42.25-86; P<.001). Subgroup analyses offered further insight. Within subgroup 2, group D demonstrated superior median (P25-P75) scores than group C in external rotation (88, 83.25-90.75 vs 61, 45.5-80.5; HLE −28, 95% CI −43 to 0; P=.036) and internal rotation (mean 88.75, SD 4.496 vs mean 50.67, SD 25.598; mean difference −38.08, 95% CI −57.88 to −18.29; $t_{8.553}$=−4.39; P=.002).

In subgroup 3, group D demonstrated greater proficiency (median, P25-P75) than group C in flexion (93, 89-96 vs 84.5, 76.75-92; HLE −9, 95% CI −15 to −2; P=.003), abduction (88, 85.5-90 vs 82, 66.5-89.25; HLE −7, 95% CI −19 to −1; P=.019), and internal rotation (85, 78-91.5 vs 73.5, 45-83; HLE −11, 95% CI −38 to 0; P=.043).

Discussion

Overview and Background

RCI refers to pathological changes in the supraspinatus, infraspinatus, teres minor, or subscapularis tendons, with the supraspinatus most commonly affected [39]. Although passive ROM theoretically remains intact in these patients, clinical observations reveal that pain and protective neuromuscular mechanisms following tendon injury frequently result in limitations in both AROM and passive ROM. Home-based rehabilitation exhibits inherent limitations in standardizing exercise parameters under the FITT (Frequency, Intensity, Time, and Type) framework, often resulting in inconsistent recovery outcomes. Furthermore, patients with chronic diseases are consistently challenged by persistent pain, movement restrictions, and insufficient professional guidance, which collectively compromise exercise adherence and rehabilitation efficacy. Despite the rapid emergence of DTx, their application remains predominantly focused on postoperative rehabilitation [14], with evidence-based research on nonsurgical shoulder interventions still in exploratory phases. This study innovatively combines injection with digital rehabilitation modalities, providing a systematic evaluation of this protocol's clinical effectiveness in the conservative management of CRCI.

Research Findings and Comparison With Prior Work

The Digital Rehabilitation System was customized based on the kinematic characteristics of CRCI. Consistent with prior findings [40], the digital protocol demonstrated superior pain reduction and functional recovery. The improvements of CMS, UCLA Shoulder Scale, and NRS exceeded their MCIDs [36-38], underscoring the critical role of real-time visualization and biofeedback mechanisms [11]. While the studies by AlHossan et al [41] and Pak et al [13] concluded that digital and conventional exercise yield comparable outcomes for chronic shoulder pain rehabilitation, our study extends these findings by demonstrating superior efficacy in a distinct, nonoperative patient population. This supports the broader applicability and integration of digital technology into shoulder rehabilitation protocols. And discrepancies between our findings and the negative conclusions reported by Shim et al [11] may stem from population heterogeneity, intervention protocol variations, and differential sensitivity of outcome measures.

Postintervention analysis revealed significant between-group differences favoring group D in pain (NRS, P=.002), shoulder function (CMS, P=.007; UCLA Shoulder Scale, P=.001), and AROM for flexion (P=.023), abduction (P=.034), external rotation (P=.001), and internal rotation (P=.035). Notably, 6 participants (4 in group C, 2 in group D) exhibited a reduction in ROM during the treatment period, which increased data heterogeneity of detecting significant between-group differences.

The finding of greater absolute rotation ROM in group D posttreatment may primarily reflect restored tissue extensibility. However, neuromuscular coordination issues may still remain unresolved. Compensatory patterns, such as scapulothoracic dyskinesis [42] and inherent rotational biomechanical constraints [43], can drive avoidance of true rotational movements, thereby perpetuating deficits [44]. Furthermore, patients' predominant reliance on flexion during ADL may create "blind zones" for rotational control in self-directed rehabilitation. If unaddressed, these issues can contribute to sustained muscle hypertonicity, strain on nontarget muscles, and increased joint stress, ultimately impeding functional recovery. Patients with severe capsular adhesions or pronounced neuromuscular dysfunction may therefore require integrated, adjunctive treatments to fully capitalize on the benefits of exercise therapy.

Scientific Rationale for the Combined Intervention

Rehabilitation exercises promote collagen synthesis and tendon strengthening. The system's algorithms address the corrective feedback delays inherent in conventional methods via real-time biofeedback, while motion visualization enhances sensorimotor cortical reorganization and neuromuscular control. The corticosteroid component of the injection suppresses synovitis and proinflammatory cytokines [5,39], rapidly alleviating pain

XSL•FO
RenderX

and edema to remove barriers to subsequent training. Additionally, local anesthetics modulate nociceptive signaling and collagen remodeling to restore tendon homeostasis [3]. Integrating injection with digital rehabilitation ensures biomechanically optimized training postures, thereby enhancing functional outcomes while minimizing reinjury risks.

Ceiling Effects in Subgroup Analysis

First, a methodological ceiling effect was inherent in the composition of subgroup 1. This subgroup was specifically defined by a combined shoulder flexion and abduction ROM of ≥150° for both movements at baseline. This threshold represents a functional level near the upper limit of normal daily requirements. Consequently, participants in this subgroup began the study at a functional ceiling, leaving minimal room for measurable improvement on the ROM scales themselves. The primary rationale for creating this distinct subgroup was to isolate and analyze the treatment response in participants whose primary complaint was pain without substantial functional impairment. Including these high-functioning individuals in the overall analysis would have diluted or obscured the detection of true functional gains in participants with more significant baseline limitations (subgroups 2 and 3), as their near-maximal scores would disproportionately influence the group means or medians.

Second, a notable observation pertains to the analysis of internal rotation in subgroup 1 at the T2 time point. The Mann-Whitney *U* test between group C and group D yielded $P>.99$ (Table S2 in Multimedia Appendix 1). Crucially, this was not due to identical raw data but due to identical rank sums (eg, both groups having a median rank of 8.5). This occurs because when the vast majority of participants in a subgroup are clustered at the very top of the measurement scale, the variance in raw scores becomes compressed. The rank-based test then loses its discriminative power, as small, clinically irrelevant differences near the ceiling are insufficient to generate a significant rank separation. Thus, the *P* value of 1.000 is a statistical artifact of the measurement ceiling, indicating that the tool was insensitive to differences at this high functional level at that specific time point, not necessarily the absence of any differential treatment effect.

Heterogeneous Treatment Response Based on Baseline Function

A key finding from the subgroup analysis was the differential efficacy of group D across patients with varying baseline function. While the unadjusted improvement magnitudes (Δ) of flexion, abduction, and external rotation suggested that subgroup 2 gained the most (Tables S2-S4 in Multimedia Appendix 1), the ANCOVA—which statistically equates baseline scores—revealed that subgroup 1 achieved a superior adjusted treatment effect. It is important to note that this ANCOVA is a within-arm analysis and does not directly establish differential between-treatment efficacy. This demonstrates that the apparent advantage of the severely impaired subgroup was likely confounded by its greater potential range for absolute improvement. After accounting for this baseline disparity, the analysis indicates that the intervention was particularly effective in further optimizing function in those who started with no or only mild limitations. Given the small per-subgroup sample sizes (n=7-14), statistical power for between-group comparisons was limited. Accordingly, these subgroup analyses should be interpreted as exploratory and hypothesis-generating.

This pattern suggests a heterogeneous treatment response, potentially mediated by the patient's initial functional capacity. Patients with no or milder impairments may possess a sufficient residual movement repertoire and neuromuscular control to more fully exploit and integrate the precise, feedback-driven training. Consequently, the digital rehabilitation protocol appears especially potent as an optimization tool for patients at a higher functional baseline, enabling them to achieve near-normal performance levels, rather than as a primary restorative intervention for those with severe deficits.

System Proficiency and Its Clinical Implications

The digital rehabilitation system used in this study incorporated a real-time score derived from a weighted composite of 4 kinematic parameters: trunk inclination, target-plane angle, compensatory motion in nontarget planes, and postural alignment during exercise. This score provided immediate biofeedback to optimize movement quality. However, as group C did not train with this system, the score may reflect system-specific proficiency—familiarity with the device and its feedback—rather than a pure, transferable measure of proprioception. Consequently, it was not interpreted as a primary clinical outcome.

Impact of Inclusion Criterion Deviation on Study Findings

A discrepancy in the pain severity inclusion criterion occurred (registry: NRS score ≥4; final analysis: NRS score ≥3) because NRS score 3 represents a clinical transition zone between mild and moderate pain, where patients often experience functional impairment despite meeting the numerical definition of mild pain, making the distinction between scores 3 and 4 difficult in practice. To assess whether this adjustment introduced bias, a series of analyses was conducted. First, a comparison between patients with NRS score 3 (n=10) and NRS score ≥4 (n=50) showed no significant differences in baseline characteristics (all $P>.05$; Table S6 in Multimedia Appendix 1), indicating that the 10 patients were comparable with the rest of the cohort. Second, after excluding these 10 patients, group C and group D remained well balanced in baseline characteristics (all $P>.05$; Table S7 in Multimedia Appendix 1). Finally, sensitivity analysis excluding patients with NRS score 3 yielded results consistent with the full analysis (Table S8 vs Table S1 in Multimedia Appendix 1), with the direction and significance of treatment effects at T4 remaining unchanged. These findings demonstrate that the inclusion of patients with NRS score 3 did not introduce bias and that our conclusions are robust.

It is noteworthy that in the sensitivity analysis cohort, 8 of the 10 excluded participants were from group C, resulting in unequal final group sizes (group C: n=22; group D: n=28). This reflects the randomness of the initial allocation rather than a systematic bias, as the baseline characteristics remained balanced. While unequal group sizes can influence statistical power, the

XSL•FO
RenderX

consistency of the treatment effects in both the full and sensitivity analyses supports the robustness of our primary conclusions regarding the efficacy of the digital intervention.

Limitations and Future Work

This study has several limitations that should be acknowledged. First, as a pilot study with a primary focus on understanding initial therapeutic effects, the follow-up period was limited to 3 months. While sufficient to assess the short-term efficacy and safety of the intervention, it may not capture the long-term interaction between the waning anti-inflammatory effects of corticosteroids and the sustainability of rehabilitation gains. Although the pharmacological half-life of corticosteroids is short (eg, 36‑54 hours for compound betamethasone), their biological effects through genomic mechanisms persist for a much longer duration. Future studies should incorporate extended follow-up assessments at 6 and 12 months and implement structured retention strategies, such as participant incentives, which would be crucial to ensure adherence in longer-term trials.

Second, the single-center design and the modest sample size (N=60) limit the generalizability of the findings. While informative for initial efficacy signals, the statistical power for subgroup analyses was constrained, potentially affecting the robustness of these comparisons. To enhance external validity and confirm the results, a subsequent multicenter randomized controlled trial (in collaboration with the community or other hospitals) with a larger cohort is warranted.

Third, the current system's training paradigm may be relatively simple for participants presenting with pain alone, potentially influencing the average system proficiency scores. Furthermore, the protocol lacked a dedicated module for proprioceptive training. Future iterations of the rehabilitation system should integrate modular and comprehensive training strategies, including randomized angle proprioception drills, to better address neuromuscular control deficits and potentially improve overall outcomes.

Fourth, concerning blinding, although all participants were exposed to the VR sessions to standardize initial exposure, complete participant blinding was not feasible due to the distinct active interventions (ie, only group D received motion capture–based training). This could have introduced performance bias or a novelty effect. A future design could use a modified "digital sham" control, where both groups interact with the digital platform, but group C receives a version that provides feedback only during assessment periods rather than as real-time training guidance, thereby better controlling for these nonspecific effects.

Conclusions

This study demonstrates that combining ultrasound-guided steroid-anesthetic injections with subsequent Digital Rehabilitation System training provides dual therapeutic benefits for CRCI. It effectively reduces pain severity while improving shoulder joint ROM and shoulder function scores. This integrated approach holds significant clinical implications, with particularly pronounced benefits observed in subgroups exhibiting no or mild shoulder mobility restrictions.

Acknowledgments

The authors extend their sincere gratitude to all patients for their dedication and valuable contributions to this study. The authors acknowledge Shanghai Ruiku Medical Technology Co, Ltd, for their expertise in developing the Digital Rehabilitation System. Special thanks are extended to the clinical staff for their ongoing support in participant recruitment. The authors acknowledge the use of QuillBot (an artificial intelligence–assisted language-polishing tool) solely for grammar and spelling checking/language editing during manuscript preparation. It was not used to generate scientific content, interpret data, or draft the intellectual content of the paper. The authors remain solely responsible for the accuracy, originality, and integrity of all content.

Funding

This work was supported in part by the Medical Discipline Construction Program of Shanghai Pudong New Area Health Commission (the Specialty Program, grant PWZzb2022-24 to JX), and in part by the Hospital Talent Program of Pudong Gongli Hospital, Shanghai University of Medicine & Health Sciences (grant 2025-GLSHLH-03 to YZ). The funding source had no influence on any aspect of the research design, execution, or the writing of the manuscript.

Data Availability

The anonymized individual participant data (including demographic characteristics, clinical outcomes, and treatment allocation details) generated in this study are available upon reasonable request. Requests should be directed to the first author, ZH. To ensure patient confidentiality, all datasets exclude personally identifiable information (eg, names and contact details) and will be transferred only after signing a data use agreement. The methodology for reproducing the trial results, including intervention protocols and outcome measurement procedures, is fully described in the "Methods" section of this paper.

Authors' Contributions

Conceptualization: ZH (lead), JX (equal)
Methodology: ZH (lead), JX (equal)
Project administration: JX (lead), YZ (equal)

XSL•FO
RenderX

Funding acquisition: JX (lead), YZ (equal)
Supervision: JX (lead), YZ (equal)
Investigation: JZ (lead), ZH (supporting)
Data curation: YC (lead), ZH (supporting)
Formal analysis: YC (lead), ZH (supporting)
Resources: LS (lead), GY (equal), XL (equal)
Writing – original draft: ZH (lead)
Writing – review & editing: ZH (lead), JX (supporting), YZ (supporting)

Conflicts of Interest

A patent application for the digital shoulder rehabilitation system used in this study has been filed by Pudong Gongli Hospital, with authors ZH, JX, and YZ listed as inventors. The system prototype was developed in partnership with Ruiku Medical Technology Co, Ltd, under a collaborative research agreement, which provided hardware development support. However, Ruiku Medical Technology Co, Ltd, was not involved in the conceptualization, execution, data analysis, or reporting of this clinical trial. All authors declare no further financial or personal relationships that could be construed as a potential conflict of interest.

Multimedia Appendix 1
Supplementary material containing primary analysis, subgroup analyses, sensitivity analyses, and supporting data.
[DOC File, 235 KB - jmir_v28i1e79494_app1.doc]

Checklist 1
CONSORT-HEALTH (V 1.6.1) checklist.
[PDF File, 1663 KB - jmir_v28i1e79494_app2.pdf]

References

1. Doiron-Cadrin P, Lafrance S, Saulnier M, et al. Shoulder rotator cuff disorders: a systematic review of clinical practice guidelines and semantic analyses of recommendations. Arch Phys Med Rehabil 2020 Jul;101(7):1233-1242. [doi: 10.1016/j.apmr.2019.12.017] [Medline: 32007452]
2. Jarosek SL, Virnig BA. Inference from imprecisely measured exposures: caution is needed when using health insurance data for clinical research. Journal of Bone and Joint Surgery 2022;104(9):e39. [doi: 10.2106/JBJS.22.00118] [Medline: 35506953]
3. Cook T, Lewis J. Rotator Cuff-related shoulder pain: to inject or not to inject? J Orthop Sports Phys Ther 2019 May;49(5):289-293. [doi: 10.2519/jospt.2019.0607] [Medline: 31039685]
4. Hopewell S, Keene DJ, Marian IR, et al. Progressive exercise compared with best practice advice, with or without corticosteroid injection, for the treatment of patients with rotator cuff disorders (GRASP): a multicentre, pragmatic, 2 × 2 factorial, randomised controlled trial. Lancet 2021 Jul 31;398(10298):416-428. [doi: 10.1016/S0140-6736(21)00846-1] [Medline: 34265255]
5. Lafrance S, Charron M, Roy JS, et al. Diagnosing, managing, and supporting return to work of adults with rotator cuff disorders: a clinical practice guideline. J Orthop Sports Phys Ther 2022 Oct;52(10):647-664. [doi: 10.2519/jospt.2022.11306] [Medline: 35881707]
6. Janela D, Costa F, Molinos M, et al. Asynchronous and tailored digital rehabilitation of chronic shoulder pain: a prospective longitudinal cohort study. J Pain Res 2022;15:53-66. [doi: 10.2147/JPR.S343308] [Medline: 35035234]
7. Pieters L, Lewis J, Kuppens K, et al. An update of systematic reviews examining the effectiveness of conservative physical therapy interventions for subacromial shoulder pain. J Orthop Sports Phys Ther 2020 Mar;50(3):131-141. [doi: 10.2519/jospt.2020.8498] [Medline: 31726927]
8. Obukhov A, Volkov A, Pchelintsev A, et al. Examination of the accuracy of movement tracking systems for monitoring exercise for musculoskeletal rehabilitation. Sensors (Basel) 2023 Sep 24;23(19):8058. [doi: 10.3390/s23198058] [Medline: 37836887]
9. Babatunde OO, Ensor J, Littlewood C, et al. Comparative effectiveness of treatment options for subacromial shoulder conditions: a systematic review and network meta-analysis. Ther Adv Musculoskelet Dis 2021;13:1759720X211037530. [doi: 10.1177/1759720X211037530] [Medline: 34527083]
10. Archer KR, Ellis TD. Advances in rehabilitation technology to transform health. Phys Ther 2024 Feb 1;104(2):pzae008. [doi: 10.1093/ptj/pzae008] [Medline: 38329471]
11. Shim GY, Kim EH, Baek YJ, et al. A randomized controlled trial of postoperative rehabilitation using digital healthcare system after rotator cuff repair. NPJ Digit Med 2023 May 23;6(1):95. [doi: 10.1038/s41746-023-00842-7] [Medline: 37221303]

XSL•FO
RenderX

12. Chung C, Kim AR, Kang DY, et al. Clinical efficacy of smartphone app-based pulmonary rehabilitation in chronic respiratory diseases: randomized controlled and feasibility trials. J Med Internet Res 2025 Nov 28;27:e76801. [doi: 10.2196/76801] [Medline: 41313804]
13. Pak SS, Janela D, Freitas N, et al. Comparing digital to conventional physical therapy for chronic shoulder pain: randomized controlled trial. J Med Internet Res 2023 Aug 18;25:e49236. [doi: 10.2196/49236] [Medline: 37490337]
14. Correia FD, Molinos M, Luís S, et al. Digitally assisted versus conventional home-based rehabilitation after arthroscopic rotator cuff repair: a randomized controlled trial. Am J Phys Med Rehabil 2022 Mar 1;101(3):237-249. [doi: 10.1097/PHM.0000000000001780] [Medline: 33935152]
15. van den Hoorn W, Fabre A, Nardese G, et al. The future of clinical active shoulder range of motion assessment, best practice, and its challenges: narrative review. Sensors (Basel) 2025 Jan 23;25(3):667. [doi: 10.3390/s25030667] [Medline: 39943306]
16. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
17. Kang H. Sample size determination and power analysis using the G*Power software. J Educ Eval Health Prof 2021;18:17. [doi: 10.3352/jeehp.2021.18.17] [Medline: 34325496]
18. Wang H, Wang Q, He J, et al. Application effect of remote intelligent rehabilitation system in rehabilitation of patients with rotator cuff injury following minimally invasive arthroscopic surgery. Chinese J Trauma 2023;39:876-884. [doi: 10.3760/cma.j.cn501098-20230522-00292]
19. Plancher KD, Shanmugam J, Briggs K, Petterson SC. Diagnosis and management of partial thickness rotator cuff tears: a comprehensive review. J Am Acad Orthop Surg 2021 Dec 15;29(24):1031-1043. [doi: 10.5435/JAAOS-D-20-01092] [Medline: 34520444]
20. Requejo-Salinas N, Lewis J, Michener LA, et al. International physical therapists consensus on clinical descriptors for diagnosing rotator cuff related shoulder pain: a Delphi study. Braz J Phys Ther 2022;26(2):100395. [doi: 10.1016/j.bjpt.2022.100395] [Medline: 35366589]
21. Tapscott DC, Varacallo MA. Supraspinatus tendonitis. In: StatPearls: StatPearls Publishing; 2023. [Medline: 31869166]
22. Oosterwijk AM, Nieuwenhuis MK, van der Schans CP, Mouton LJ. Shoulder and elbow range of motion for the performance of activities of daily living: a systematic review. Physiother Theory Pract 2018 Jul;34(7):505-528. [doi: 10.1080/09593985.2017.1422206] [Medline: 29377745]
23. Malliaras P, Cridland K, Hopmans R, et al. Internet and Telerehabilitation-Delivered Management of Rotator Cuff-Related Shoulder Pain (INTEL Trial): randomized controlled pilot and feasibility trial. JMIR Mhealth Uhealth 2020 Nov 18;8(11):e24311. [doi: 10.2196/24311] [Medline: 33206059]
24. Keene DJ, Soutakbar H, Hopewell S, et al. Development and implementation of the physiotherapy-led exercise interventions for the treatment of rotator cuff disorders for the “Getting it Right: Addressing Shoulder Pain” (GRASP) trial. Physiotherapy 2020 Jun;107:252-266. [doi: 10.1016/j.physio.2019.07.002] [Medline: 32026827]
25. Desmeules F, Roy JS, Lafrance S, et al. Rotator cuff tendinopathy diagnosis, nonsurgical medical care, and rehabilitation: a clinical practice guideline. J Orthop Sports Phys Ther 2025 Apr;55(4):235-274. [doi: 10.2519/jospt.2025.13182] [Medline: 40165544]
26. Dubé MO, Lafrance S, Charron M, et al. FITT Odyssey: a scoping review of exercise programs for managing rotator cuff-related shoulder pain. J Orthop Sports Phys Ther 2024 Aug;54(8):513-529. [doi: 10.2519/jospt.2024.12452] [Medline: 38832666]
27. Malhotra N, Sinha N, Kumar A, Gupta D, Malhotra N. Comparison of efficacy of intra-articular injection of platelet-rich plasma with bupivacaine and steroid combination in chronic shoulder pain. J Anaesthesiol Clin Pharmacol 2026;42(1):120-126. [doi: 10.4103/joacp.joacp_138_25] [Medline: 41542215]
28. Meehan K, Wassinger C, Roy JS, Sole G. Seven key themes in physical therapy advice for patients living with subacromial shoulder pain: a scoping review. J Orthop Sports Phys Ther 2020 Jun;50(6):285-a12. [doi: 10.2519/jospt.2020.9152] [Medline: 32476583]
29. He Z, Xu J, Yuan G, et al. Human-centered design of a digital shoulder rehabilitation system: a mixed-methods usability study. Presented at: Proceedings of the 2025 6th International Symposium on Artificial Intelligence for Medical Sciences: Association for Computing Machinery; Oct 24-26, 2025; Wuhan China p. 268-273. [doi: 10.1145/3777577.3777620]
30. Chepeha J, Silveira A, Sheps D, Luciak-Corea C, Styles-Tripp F, Beaupre L. A standardized criteria-based progressive shoulder exercise program is effective in managing rotator cuff-related shoulder pain: a prospective cohort study. PLoS One 2025;20(7):e0328728. [doi: 10.1371/journal.pone.0328728] [Medline: 40700393]
31. Angst F, Schwyzer HK, Aeschlimann A, Simmen BR, Goldhahn J. Measures of adult shoulder function: Disabilities of the Arm, Shoulder, and Hand Questionnaire (DASH) and its short version (QuickDASH), Shoulder Pain and Disability Index (SPADI), American Shoulder and Elbow Surgeons (ASES) Society standardized shoulder assessment form, Constant (Murley) Score (CS), Simple Shoulder Test (SST), Oxford Shoulder Score (OSS), Shoulder Disability Questionnaire (SDQ), and Western Ontario Shoulder Instability Index (WOSI). Arthritis Care Res (Hoboken) 2011 Nov;63 Suppl 11:S174-S188. [doi: 10.1002/acr.20630] [Medline: 22588743]

XSL•FO
RenderX

32. Hawker GA, Mian S, Kendzerska T, French M. Measures of adult pain: Visual Analog Scale for Pain (VAS Pain), Numeric Rating Scale for Pain (NRS Pain), McGill Pain Questionnaire (MPQ), Short-Form McGill Pain Questionnaire (SF-MPQ), Chronic Pain Grade Scale (CPGS), Short Form-36 Bodily Pain Scale (SF-36 BPS), and Measure of Intermittent and Constant Osteoarthritis Pain (ICOAP). Arthritis Care Res (Hoboken) 2011 Nov;63 Suppl 11:S240-S252. [doi: 10.1002/acr.20543] [Medline: 22588748]
33. Ohta S, Ueda Y, Komai O. Postoperative results of arthroscopic superior capsule reconstruction using fascia lata: a retrospective cohort study. J Shoulder Elbow Surg 2024 Mar;33(3):686-697. [doi: 10.1016/j.jse.2023.07.021] [Medline: 37619926]
34. Uğur E, Kurter A, Aydın Ç, Konukseven B. Novel approach to the simulator sickness questionnaire. J Audiol Otol 2025 Jan;29(1):57-63. [doi: 10.7874/jao.2024.00444] [Medline: 39916401]
35. Kim HK, Park J, Choi Y, Choe M. Virtual reality sickness questionnaire (VRSQ): motion sickness measurement index in a virtual reality environment. Appl Ergon 2018 May;69:66-73. [doi: 10.1016/j.apergo.2017.12.016] [Medline: 29477332]
36. Zhou A, Xu S, Yew KSA, Lie DTT. Minimal clinically important differences for Oxford, Constant, and University of California Los Angeles Shoulder Scores after reverse shoulder arthroplasty to allow interpretation of patient-reported outcome measures and future statistical power analyses. Arthroscopy 2023 Jun;39(6):1405-1414. [doi: 10.1016/j.arthro.2022.12.027] [Medline: 36592696]
37. Yoo DH, Choi JY, Lee SG, et al. Long-term follow-up of inpatients with rotator cuff tear who received integrative Korean medicine treatment: a retrospective analysis and questionnaire survey. Explore (NY) 2024;20(2):212-221. [doi: 10.1016/j.explore.2023.08.005] [Medline: 37689574]
38. Salaffi F, Stancati A, Silvestri CA, Ciapetti A, Grassi W. Minimal clinically important changes in chronic musculoskeletal pain intensity measured on a numerical rating scale. Eur J Pain 2004 Aug;8(4):283-291. [doi: 10.1016/j.ejpain.2003.09.004] [Medline: 15207508]
39. Jain NB, Khazzam MS. Degenerative rotator-cuff disorders. N Engl J Med 2024 Nov 28;391(21):2027-2034. [doi: 10.1056/NEJMcp1909797] [Medline: 39602631]
40. Özlü A, Üstündağ S, Bulut Özkaya D, Menekşeoğlu AK. Effect of exergame on pain, function, and quality of life in shoulder impingement syndrome: a prospective randomized controlled study. Games Health J 2024 Apr;13(2):109-119. [doi: 10.1089/g4h.2023.0108] [Medline: 38394299]
41. AlHossan AM, Jahhaf RH, Alharbi AS, et al. Digital and virtual reality-based rehabilitation versus conventional therapy for rotator cuff tears and post-repair recovery: a systematic review and meta-analysis. JSES Rev Rep Tech 2026 Feb;6(1):100584. [doi: 10.1016/j.xrrt.2025.09.003] [Medline: 41142763]
42. Nakamura Y, Yokoya S, Matsubara Y, Harada Y. Relationship between scapular body morphology and rotator cuff tears: a threedimensional computed tomography study. Acta Orthop Traumatol Turc 2024 Jun 13;58(3):167-170. [doi: 10.5152/j.aott.2024.22181] [Medline: 39165221]
43. Bolia IK, Collon K, Bogdanov J, Lan R, Petrigliano FA. Management options for shoulder impingement syndrome in athletes: insights and future directions. Open Access J Sports Med 2021;12:43-53. [doi: 10.2147/OAJSM.S281100] [Medline: 33880071]
44. Carnevale A, Mannocchi I, Schena E, et al. Performance evaluation of an immersive virtual reality application for rehabilitation after arthroscopic rotator cuff repair. Bioengineering (Basel) 2023 Nov 10;10(11):1305. [doi: 10.3390/bioengineering10111305] [Medline: 38002429]

Abbreviations

ADL: activities of daily living
AROM: active range of motion
CMS: Constant-Murley Score
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
CRCI: chronic rotator cuff injury
DTx: digital therapeutics
FITT: Frequency, Intensity, Time, and Type
HLE: Hodges-Lehmann estimator
MCID: minimal clinically important difference
MCS: Motion Capture subsystem
MD: median of difference
NRS: Numerical Rating Scale
RCI: rotator cuff injury
ROM: range of motion
T1: baseline
T2: 1-week follow-up

XSL•FO
RenderX

T3: 1-month follow-up
T4: 3-month follow-up
UCLA: University of California Los Angeles
VR: virtual reality
VRS: Virtual Reality Education subsystem

Edited by A Coristine; submitted 23.Jun.2025; peer-reviewed by J Lang, S Sohani; revised version received 03.May.2026; accepted 04.May.2026; published 10.Jun.2026.

Please cite as:
He Z, Zhang Y, Shi L, Yuan G, Liu X, Cai Y, Zhao J, Xu J
Digital Rehabilitation Following Ultrasound-Guided Injection for Chronic Rotator Cuff Injury: Randomized Controlled Trial
J Med Internet Res 2026;28:e79494
URL: https://www.jmir.org/2026/1/e79494
doi:10.2196/79494

© Zifu He, Yuemin Zhang, Liyan Shi, Gaineng Yuan, Xuefeng Liu, Yue Cai, Jue Zhao, Jing Xu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Physical Activity Recommendations Tailored by a Predictive Model for Adults With High Blood Pressure: Observational Study

Yuhui Yang[1], PhD; Manqing Chen[1], MS; Weiwei Hu[1], PhD; Yifan Fu[1], MS; Xingyan Li[1], MS; Zhenli Liao[1], MS; Hongman Feng[1], MS; Yaling Zhao[1], PhD; Leilei Pei[1], PhD; Baibing Mi[1*], PhD; Fangyao Chen[1,2*], PhD

[1]Department of Epidemiology and Biostatistics, School of Public Health, Xi'an Jiaotong University Health Science Center, 76 Yanta Xilu Road, Xi'an, Shaanxi, China

[2]Department of Radiology, First Affiliate Hospital of Xi'an Jiaotong University, Xi'an, Shaanxi, China

[*]these authors contributed equally

Corresponding Author:
Fangyao Chen, PhD
Department of Epidemiology and Biostatistics, School of Public Health, Xi'an Jiaotong University Health Science Center, 76 Yanta Xilu Road, Xi'an, Shaanxi, China

Abstract

Background: Whether the benefits of identical physical activity (PA) patterns for adults with high blood pressure (BP) vary according to an individual's characteristics has not been adequately studied.

Objective: This study aimed to investigate whether an individual's characteristics modify the associations between PA patterns and mortality rate.

Methods: Four PA patterns were derived from accelerometer-based data: active weekend warrior, active regular, active light PA, and baseline PA. The main outcome was all-cause mortality. A machine learning model to predict the optimal PA pattern for individual patients was trained in the UK Biobank (UKB) cohort and externally validated in the National Health and Nutrition Examination Survey cohort, which was subsequently integrated into a web-based application. The potentially optimal PA pattern within patients was identified as the one leading to the highest predicted survival probability. Multivariable Cox models were used to estimate hazard ratios and 95% CIs for all-cause mortality corresponding to the inconsistency of the current PA pattern with the predicted optimal PA pattern.

Results: A total of 71,637 UKB adults and 5104 National Health and Nutrition Examination Survey individuals were enrolled. External validation demonstrated that the area under the receiver operating characteristic curve of our model for predicting mortality at 10 years of follow-up was 86.4% (95% CI 85.1% - 87.7%). The predicted optimal PA patterns in the UKB cohort were active regular PA for 26,643 (37.2%) participants, active light PA for 22,606 (31.6%) participants, and active weekend warrior for 21,749 (30.4%) participants. Stroke history, age, sex, BP class, and antihypertension medication were key factors driving heterogeneity in individuals' optimal PA patterns. Cox regression analysis suggested that individuals in the UKB cohort whose current PA patterns were inconsistent with the predicted optimal patterns may be associated with a 28% increase in all-cause mortality risk on average (hazard ratio 1.28, 95% CI 1.20 - 1.38) compared to those with consistent patterns.

Conclusions: Our findings may help patients with high BP obtain individualized recommendations for PA patterns based on their specific characteristics, thereby improving their prognosis.

(J Med Internet Res 2026;28:e78492) doi:10.2196/78492

KEYWORDS

hypertension; physical activity pattern; machine learning; mortality; precision medicine

Introduction

According to the latest report from the World Health Organization, the global population of adults with hypertension has more than doubled between 1990 and 2019, rising from 0.65 billion to 1.3 billion [1,2]. In 2019, high blood pressure (BP) was responsible for 10.8 million deaths worldwide, primarily due to cardiovascular diseases and chronic kidney disease [1,2]. Timely and effective interventions are therefore critical to mitigating this burden. International hypertension management guidelines universally endorse physical activity (PA) as a first-line nonpharmacological intervention for BP control [3,4]. Furthermore, PA offers distinct advantages over antihypertensive medications, including cost-effectiveness and a reduced risk of adverse effects [5,6].

As outlined by Dzau and Hodgkinson [7], patients with hypertension should receive individualized management that accounts for patients' unique genetic and environmental factors,

XSL•FO
RenderX

namely, *precision hypertension*. This perspective highlights the heterogeneity in treatment response among hypertensive individuals and underscores the need for tailored PA strategies. A major challenge in advancing precision hypertension is to move beyond the estimation of average associations and address the variability in individual responses to PA. This variability depends on factors such as personal characteristics, baseline risk profiles, and differential sensitivity to PA interventions [8]. Therefore, conducting studies on the heterogeneous associations between PA and health outcomes and identifying key factors driving this heterogeneity is essential for precision hypertension management.

PA patterns (defined by frequency, duration, and intensity of PA) and their associations with mortality have caught significant attention in recent research [9-11]. Furthermore, these studies on the heterogeneous associations between PA and mortality rely on subgroup analyses [12,13]. For example, Ji et al [13] studied the heterogeneous association between PA and mortality using prespecified gender subgroups. However, such traditional subgroup analyses require prespecification of potential subgroup variables, which limits their exploratory value [14]. Furthermore, one-variable-at-a-time analyses may produce false positives due to multiple testing and false negatives due to limited statistical power in small subgroup samples [15]. The burgeoning field of machine learning (ML) holds promise in estimating heterogeneous treatment effects, potentially offering significant support in addressing this concern [7,16]. Therefore, in our study, we adopted the S-learner framework, a metalearner for estimating individualized treatment effects using ML [17]. The ML based on metalearners can simulate the progression of hypertension and identify optimal PA patterns for managing these conditions.

We hypothesized that the associations of PA patterns with mortality would differ based on individual patient characteristics. In other words, the optimal PA pattern for patients may exhibit heterogeneity depending on individual characteristics. To test this hypothesis, we derived and externally validated a potential outcome prediction model, which was further integrated into a web-based application to identify the optimal PA pattern for an individual with hypertension. In our study, PA patterns were defined using accelerometer data from the UK Biobank (UKB) and the National Health and Nutrition Examination Survey (NHANES) cohorts.

Methods

Study Design and Participants

Our study included two surveys: the UKB and the NHANES. The UKB is a prospective cohort study comprising 502,629 participants enrolled between 2006 and 2010. The NHANES is a biannual survey designed to assess the health and nutritional status of the US population. This study adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [18].

The inclusion criteria for this study were individuals with device-measured PA data and high BP. Exclusion criteria were participants with (1) poor device calibration, (2) inadequate wear time, or (3) missing values for the outcome and covariates. According to the 2024 European Society of Cardiology Guidelines for the management of elevated BP and hypertension [3], participants were classified as having high BP if they met any of the following criteria: systolic BP (SBP) ≥120 mm Hg, diastolic BP (DBP) ≥70 mm Hg, hospitalization records, or the use of antihypertensive medications. In our study, we further categorized high BP into two groups: elevated BP and hypertension. Participants with SBP between 120 and 139 mm Hg or DBP between 70 and 89 mm Hg who were not using antihypertensive medication were classified as having elevated BP; others were classified as having hypertension.

A total of 103,582 participants from the UKB and 14,631 participants from the NHANES cohort were initially considered. After applying the inclusion and exclusion criteria, 71,637 (69.2%) participants from the UKB and 5104 (34.9%) participants from NHANES were included in the final analysis.

Exposure Ascertainment

Information on the collection of accelerometer-based PA data in the UKB and NHANES studies is provided as follows. From 2013 to 2015, UKB participants were randomly assigned to wear an Axivity AX3 accelerometer (Newcastle upon Tyne) for 7 days to measure PA and sedentary behavior. The wrist-worn accelerometers were initialized to capture data with a sampling frequency of 100 Hz and a dynamic range of ±8 g. Poor device calibration was defined as a lack of sufficient orientation changes or as having implausible acceleration values. Good wear time was defined as having at least 3 days (72 h) of data and also having data in each 1-hour period of the 24-hour cycle (scattered over multiple days). The accelerometers recorded data in milligravity units (mg). Light PA (LPA) and moderate to vigorous physical activity (MVPA) were categorized via a published ML-driven approach specifically designed for classifying a wide spectrum of activities [19].

For the NHANES, we included individuals with accelerometer-based PA data from the 2003 to 2006 cycles of NHANES. According to the design of the NHANES study, information on sedentary behavior and total PA was collected using an accelerometer (ActiGraph model 7164; ActiGraph, LLC), worn on the waist for 7 consecutive days during waking hours, except while swimming or showering. Poor device calibration was defined as more than 60 consecutive minutes with zero counts. Good wear time was defined as a recorded wear time of 10 hours or more for at least 1 day [20]. Behavior categories were defined by count per minute thresholds for adults: sedentary behavior (<100 counts per minute), LPA (100 - 2020 counts per minute), and MVPA (≥2020 counts per minute) [21].

We defined four PA patterns according to the following criteria: active weekend warrior (WW; ≥150 min per week MVPA with ≥50% of total achieved in 1 - 2 d) [22], active regular (≥150 min per week and not meeting MVPA WW status) [22], active LPA (<150 min per week MVPA and ≥1900 min per week LPA), and baseline PA (<150 min per week MVPA and <1900 min per week LPA). The threshold of 1900 minutes per week for an active LPA pattern was determined based on the previous study [23], and our analysis of the dose-response relationship

XSL•FO
RenderX

between LPA and all-cause mortality was determined using the Cox regression models with restricted cubic splines (Multimedia Appendix 1).

Mortality Ascertainment

The primary outcome of this study was all-cause mortality, obtained from the National Health Service Information Center (England and Wales) and the National Health Service Central Register Scotland (Scotland) for the UKB study. At the time of analysis, mortality data were available through May 31, 2024, for England and Wales, and December 31, 2023, for Scotland. For NHANES, the National Death Index was used to ascertain all-cause mortality of included samples until December 31, 2019. The follow-up period was defined as the duration between the initial PA measurement and either death or the end of follow-up, whichever occurred first. The causes of death were confirmed based on ICD-10 coding.

Covariates

The covariates were age, sex, ethnicity, BMI, waist circumference, education, smoking status, alcohol consumption, added salt intake, sedentary time, sleep time, SBP, DBP, antihypertension medication, cancer, diabetes, myocardial infarction (MI), stroke, family cardiovascular disease (CVD), BP class, glycated hemoglobin (HbA_{1c}), high-density lipoprotein cholesterol, triglyceride levels, and glucose. Ascertainment and descriptions of covariates considered for both UKB and NHANES are provided in Multimedia Appendix 2.

To minimize the time interval between covariate assessment and accelerometer-based PA measurement (typically 2013 - 2015) in the UKB, we prioritized data from instance 1 (the repeat assessment in 2012 - 2013) when both instance 0 (the baseline assessment in 2006 - 2010) and instance 1 were available [22]. We assessed the reliability of covariates in participants with repeated measurements by the intraclass correlation coefficient [24] for continuous variables and Cohen κ [25] for categorical variables. The analysis indicated moderate to high reliability of the covariates; detailed results are provided in Multimedia Appendix 3. In contrast, the PA and covariates in the external validation set, NHANES, were measured during the same period, thus avoiding the aforementioned issues present in UKB.

Model Development and Estimation of Heterogeneous Associations

We randomly divided the UKB data into an 80% training set and a 20% internal validation set, with the NHANES dataset used for external validation. Predictors were selected using Cox regression with the least absolute shrinkage and selection operator (LASSO) penalization [26], where the optimal shrinkage parameter lambda was determined via 10-fold cross-validation [27]. The original predictive model was constructed using a multivariable Cox model that incorporated all second-order interactions between PA pattern and the selected predictors. Then, we applied a stepwise backward elimination method to refine the model, removing variables and interaction terms that did not significantly contribute to the model [28]. Model performance was evaluated in the internal validation set and the NHANES validation set and visualized using the calibration curves (with Brier score) and receiver operating characteristic (ROC) analysis (with the area under the ROC curve [AUC]) [29,30].

Our S-learner algorithm for estimating heterogeneous associations between PA and overall survival follows a 2-step process. First, it uses the trained prediction model to estimate the conditional expectations of survival time for each of the 4 PA patterns separately. Second, it computes the differences between these estimates to capture the heterogeneous associations.

On the basis of the S-learner framework, the trained prediction model was capable of handling potential variations in covariates, enabling the prediction of each patient's potential survival probabilities under the 4 PA patterns. Then, the best potential PA pattern for a certain patient was identified as the one leading to the highest predicted survival probability. To operationalize these findings, we developed an R Shiny web application that enables clinicians to simulate personalized survival curves for all 4 PA patterns based on individual baseline characteristics.

Patient Characteristics Associated With Predicted Individualized PA Patterns

All individuals in the UKB and the NHANES cohorts were stratified by PA patterns that might individually optimize overall survival. This stratification was used to identify the profiles of individuals who could benefit most from 1 of the 4 PA patterns presented, according to the constructed prediction model.

To facilitate rapid decision-making and clearly illustrate patient characteristics associated with predicted individualized PA patterns, a model of conditional inference tree (CIT) was established [31]. Specifically, we used the potentially optimal PA patterns predicted as the outcome and all covariates as predictors to build the CIT model.

Statistical Analysis

We introduced an indicator variable, referred to as an "inconsistent PA pattern," to distinguish between individuals whose actual and predicted optimal PA patterns were inconsistent and those whose patterns were consistent. Multivariable Cox models were used to estimate hazard ratios (HRs) and 95% CIs for all-cause mortality corresponding to the inconsistency of the current PA pattern with the predicted optimal PA pattern. The analyses were adjusted for covariates selected by LASSO.

As further stratification of participants with inconsistent PA patterns provides more clinically meaningful information, we subdivided the inconsistent group into two categories based on the predicted optimal intensity: (1) inconsistent MVPA pattern—participants whose predicted optimal PA pattern required more MVPA but whose actual activity was lower than predicted and (2) inconsistent LPA pattern—participants whose predicted optimal PA pattern required LPA primarily but whose actual activity (including active regular, active WW, and baseline PA patterns) differed. The original binary variable "inconsistent PA pattern" was thus updated to a category variable with three levels: consistent, inconsistent MVPA pattern, and inconsistent LPA pattern. We then used "consistent"

XSL•FO
RenderX

as a reference class for the aforementioned multivariable Cox model analyses. In addition, we conducted subgroup analysis in the actually observed nonbaseline group.

Sensitivity Analyses

To test the generalizability of the aforementioned analyses, we performed 4 prespecified sensitivity analyses. First, we repeated the statistical analyses using the NHANES dataset. Second, we replaced the all-cause mortality outcome with CVD-specific mortality and cancer-specific mortality outcomes. Third, to minimize the potential confounding effects of the COVID-19 pandemic, we restricted follow-up to the period before December 31, 2019. This prepandemic dataset was used as an additional validation cohort to assess the robustness of the model's performance and to repeat the statistical analyses. Finally, the statistical analyses were repeated using the internal validation set.

All analyses were conducted using R version 4.4.1 (R Foundation), with statistical significance set at $P<.05$.

Ethical Considerations

This study involved secondary analysis of data from the UKB and the NHANES, whose original data collection protocols were conducted in accordance with the Declaration of Helsinki. UKB was approved by the North West Multi-centre Research Ethics Committee (reference 21/NW/0157), and all participants provided written informed consent for health-related research, including secondary analyses. The NHANES study received ethical approval from the National Center for Health Statistics Ethics Review Board. The National Center for Health Statistics Ethics Review Board approved Protocol 98‐12 for the 2003‐2004 cycle and Protocol 2005‐06 for the 2005‐2006 cycle. All participants provided written informed consent. All data were anonymized or deidentified.

Results

Overview

For the UKB cohort, a total of 71,637 adults were enrolled and analyzed (Figure 1), with 4207 deaths confirmed during a median follow-up period of 9.5 years. For the NHANES cohort, we included 5104 individuals who underwent device-based activity measurements between the 2003 and 2006 cycles (Multimedia Appendix 4), with 1279 deaths confirmed during a median follow-up period of 14.3 years. The median follow-up time for survivors and nonsurvivors in both the UKB and the NHANES datasets is shown in Multimedia Appendix 5. The baseline characteristics of participants, stratified by 4 actually *observed* PA patterns, are summarized in Multimedia Appendix 6 for the UKB cohort and the NHANES cohort. At baseline in the UKB, 12,577 (17.6%) participants were classified as baseline PA, 12,782 (17.8%) participants were classified as active LPA, 15,141 (21.1%) participants were classified as active regular, and 31,137 (43.5%) participants were classified as active WW.

XSL•FO
RenderX

Figure 1. Study flow for UK Biobank. CVD: cardiovascular disease; DBP: diastolic blood pressure; HbA_{1c}: glycated hemoglobin; HDL: high-density lipoprotein; LPA: light physical activity; MVPA: moderate to vigorous physical activity; SBP: systolic blood pressure.

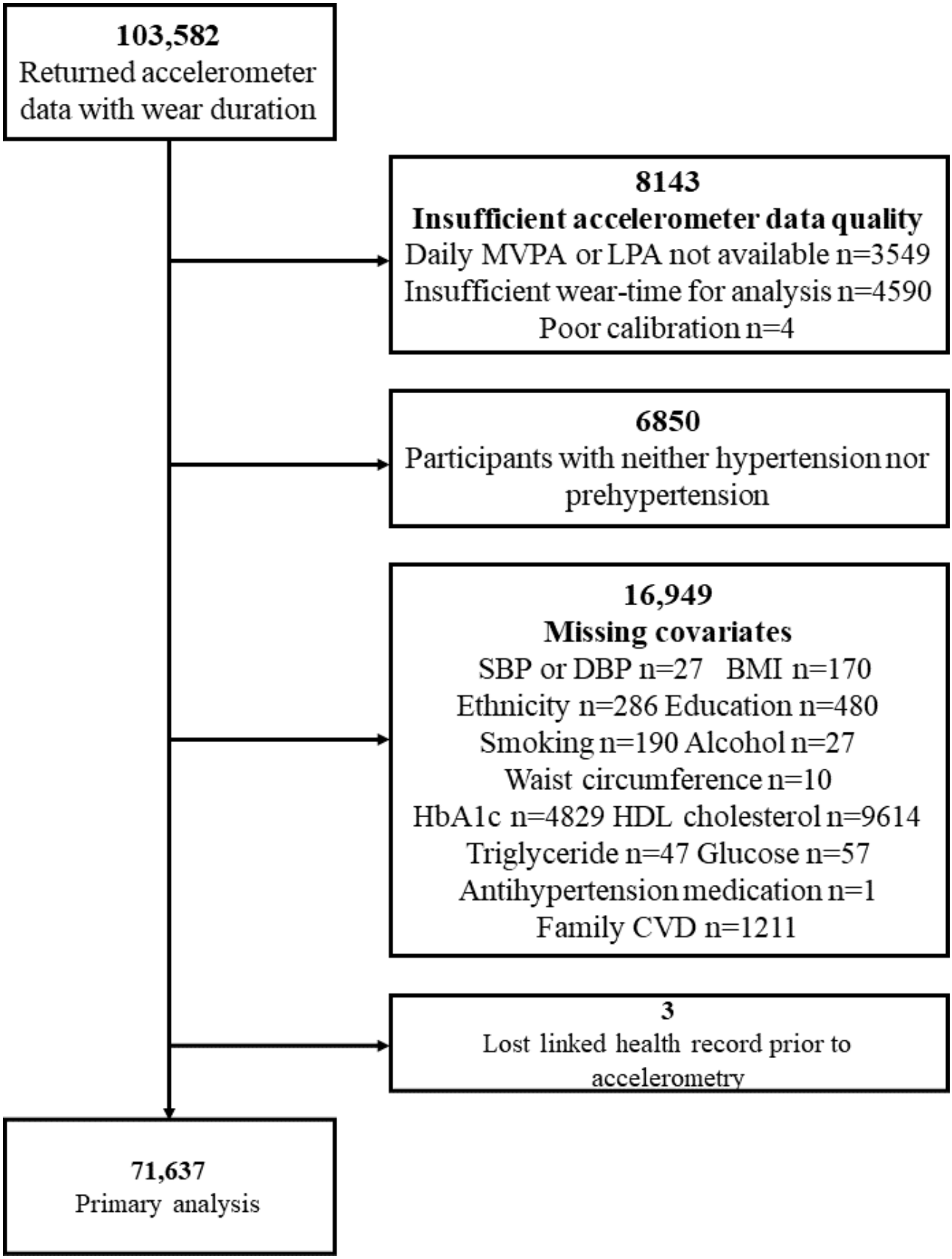

XSL•FO
RenderX

A Prediction Model to Identify the Best Potential PA Patterns

The detailed results from the LASSO-penalized Cox model for covariate selection are presented in Multimedia Appendix 7. In brief, the selected covariates included PA patterns, age, sex, sedentary time, smoking status, antihypertension medication, cancer, diabetes, MI, stroke, BP class, waist circumference, glucose, and HbA_{1c}. In the UKB cohort, the associations between covariates selected by LASSO and all-cause mortality were examined using univariate Cox models (Multimedia Appendix 8): PA patterns were found to be associated with overall survival, suggesting that adults with active LPA (HR 0.56, 95% CI 0.51 - 0.62), active regular (HR 0.45, 95% CI 0.41 - 0.50), or active WW (HR 0.47, 95% CI 0.43 - 0.51) were at a lower risk of death than those with the baseline PA pattern.

The original predictive model incorporated all second-order interactions of PA patterns with each selected covariate. Following stepwise backward elimination, the main effect of glucose and the interactions of PA patterns with waist circumference, smoking status, and glucose were removed from the model. The terms and coefficients of the final model are detailed in Multimedia Appendix 9. The model showed well-calibrated predictions in the internal test set, with Brier scores of 2.0% (95% CI 1.8% - 2.2%) for the 5-year and 5.1% (95% CI 4.7% - 5.4%) for the 10-year prediction (Figure 2A). Calibration remained acceptable in the external NHANES set, although it was slightly decreased, with Brier scores of 5.2% (95% CI 4.7% - 5.8%) for the 5-year and 9.7% (95% CI 9.1% - 10.4%) for the 10-year prediction (Figure 2B). The model also demonstrated good discriminative ability. In the internal set, the AUC was 76.8 % (95% CI 74.2% - 79.5%) for the 5-year and 78.1% (95% CI 76.2% - 79.9%) for the 10-year prediction (Figure 2C). This performance was further improved in the external validation set, with AUCs of 85.0% (95% CI 82.9% - 87.1%) and 86.4% (95% CI 85.1% - 87.7%) for the 5- and 10-year prediction, respectively (Figure 2D).

XSL•FO
RenderX

Figure 2. Performance of the model in predicting all-cause mortality. The calibration and ROC curves for internal (n=14,328) and external test data set (n=5104). AUC: area under the receiver operating characteristic curve; NHANES: National Health and Nutrition Examination Survey; ROC: receiver operating characteristic.

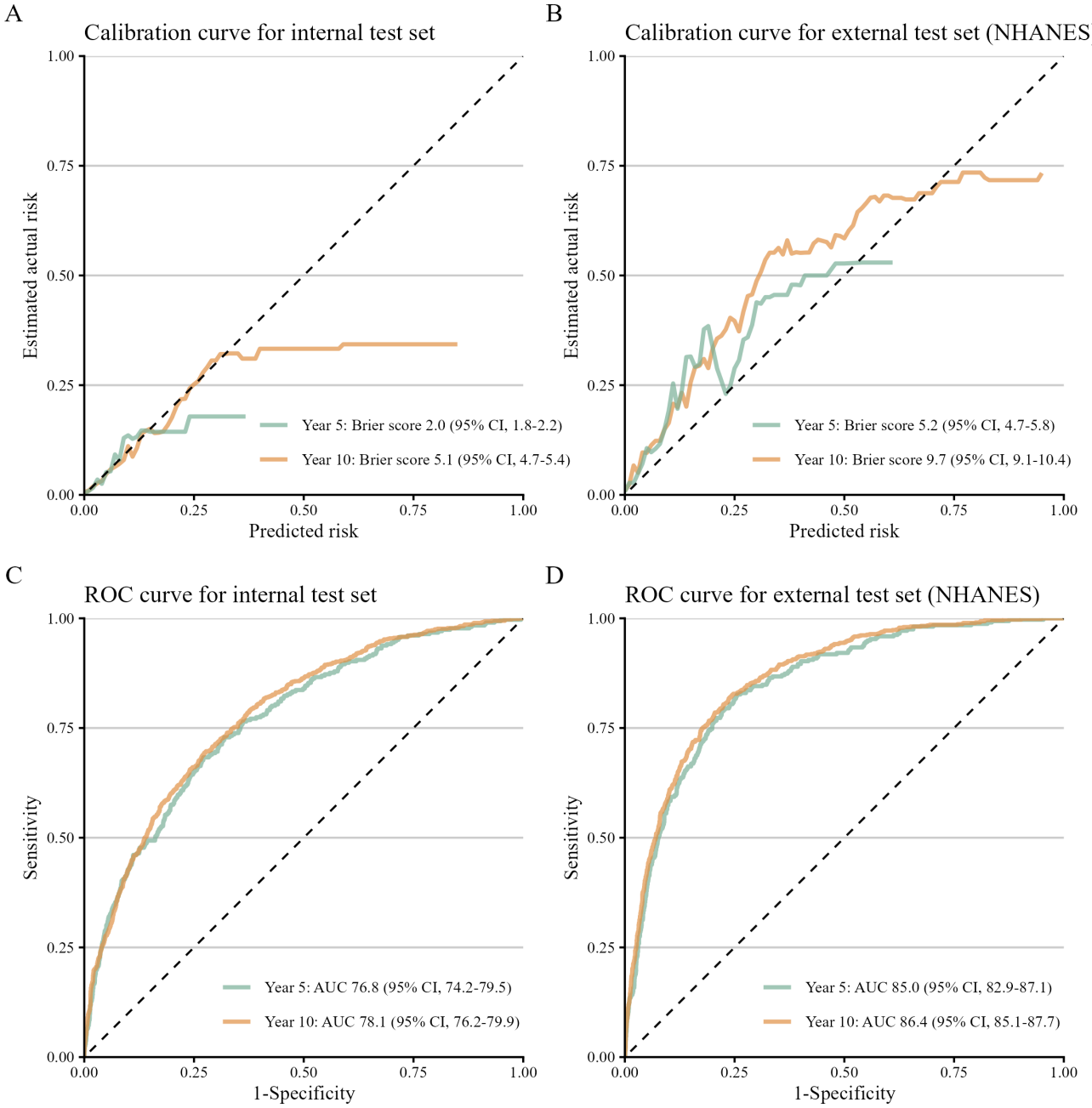

The prediction model identifying the best PA pattern has been integrated into an online application, accessible free of charge (Multimedia Appendix 10). To ensure broader accessibility of the application, we have provided both English and Chinese versions (Multimedia Appendix 10). Users need to enter the required clinical characteristics (including age, sex, sedentary time, smoking status, antihypertension medication, cancer, diabetes, MI, stroke, BP class, waist circumference, and HbA_{1c}) into the website. The tool then calculates the predicted 5-year and 10-year survival probabilities corresponding to the 4 PA patterns. After these probabilities are generated, the application identifies the pattern with the highest predicted survival probability. This information can support clinicians and patients in selecting the most appropriate personalized PA pattern.

Patient Characteristics Associated With Predicted Optimal PA Patterns

Our predictive model indicated that the optimal PA pattern varied among adults with high BP. Specifically, 26,643 (37.2%) participants were predicted to benefit most from the active regular pattern, 22,606 (31.6%) participants from the active LPA pattern, 21,749 (30.4%) participants from the active WW pattern, and 639 (0.9%) participants from the baseline PA pattern. Table 1 presents the characteristics of participants according to their *predicted* optimal PA patterns and compares these with the patterns they actually followed in the UKB cohort. In addition, we applied the same predictive approach to NHANES participants and compared *predicted* versus *observed* PA patterns, as shown in Multimedia Appendix 11.

XSL•FO
RenderX

Table . Baseline characteristics of participants stratified by the predicted optimal physical activity (PA) pattern.

Variables and levels	Baseline PA (n=639)	Active LPA[a] (n=22,606)	Active regular (n=26,643)	Active WW[b] (n=21,749)	*P* value
Pattern actually observed, n (%)					<.001
Baseline PA	191 (29.9)	3118 (13.8)	5238 (19.7)	4030 (18.5)	
Active LPA	118 (18.5)	4532 (20.0)	4734 (17.8)	3398 (15.6)	
Active regular	105 (16.4)	5169 (22.9)	5467 (20.5)	4400 (20.2)	
Active WW	225 (35.2)	9787 (43.3)	11,204 (42.1)	9921 (45.6)	
Age (y), median (IQR)	67.9 (63.7-71.3)	56.8 (51.7-62.4)	62.8 (56.9-67.3)	69.8 (66.8-72.5)	<.001
Sex, n (%)					<.001
Male	447 (70)	4407 (19.5)	12,069 (45.3)	15,948 (73.3)	
Female	192 (30)	18,199 (80.5)	14,574 (54.7)	5801 (26.7)	
WC[c] (cm), median (IQR)	96.0 (87.0-104.2)	82.0 (75.0-91.0)	90.0 (81.0-99.0)	93.0 (85.0-101.0)	<.001
Smoking, n (%)					<.001
Never	283 (44.3)	14,066 (62.2)	15,541 (58.3)	10,943 (50.3)	
Previous	300 (46.9)	6979 (30.9)	9206 (34.6)	9639 (44.3)	
Current	56 (8.8)	1561 (6.9)	1896 (7.1)	1167 (5.4)	
Antihypertension medication, n (%)					<.001
No	217 (34)	19,694 (87.1)	23,541 (88.4)	14,643 (67.3)	
Yes	422 (66)	2912 (12.9)	3102 (11.6)	7106 (32.7)	
Blood pressure class, n (%)					<.001
Elevated	54 (8.5)	19,513 (86.3)	3405 (12.8)	9907 (45.6)	
Hypertension	585 (91.5)	3093 (13.7)	23,238 (87.2)	11,842 (54.4)	
MI[d], n (%)					<.001
No	556 (87)	22,595 (100)	26,097 (98)	20,705 (95.2)	
Yes	83 (13)	11 (0)	546 (2)	1044 (4.8)	
Stroke, n (%)					<.001
No	0 (0)	22,527 (99.7)	26,585 (99.8)	21,409 (98.4)	
Yes	639 (100)	79 (0.3)	58 (0.2)	340 (1.6)	
Diabetes, n (%)					<.001
No	573 (89.7)	21,610 (95.6)	26,454 (99.3)	19,405 (89.2)	
Yes	66 (10.3)	996 (4.4)	189 (0.7)	2344 (10.8)	
Cancer, n (%)					<.001
No	534 (83.6)	22,291 (98.6)	22,158 (83.2)	20,519 (94.3)	
Yes	105 (16.4)	315 (1.4)	4485 (16.8)	1230 (5.7)	
Glucose (mmol/L), median (IQR)	5.0 (4.7-5.5)	4.8 (4.5-5.2)	5.0 (4.6-5.3)	5.0 (4.6-5.4)	<.001
HbA_{1c}[e] (mmol/mol), median (IQR)	36.9 (34.5-39.8)	33.8 (31.5-36.1)	35.5 (33.2-37.9)	35.5 (33.1-38.1)	<.001

[a]LPA: light physical activity.
[b]WW: weekend warrior.
[c]WC: waist circumference.
[d]MI: myocardial infarction.
[e]HbA_{1c}: glycated hemoglobin.

The results of the CIT model are shown in Figure 3. Briefly, the model is rooted by stroke, with subsequent branching by age, BP class, sex, and antihypertension medication (all *P*<.001). Adults with stroke are more likely to have baseline PA or active

XSL•FO
RenderX

LPA as optimal patterns. In contrast, for those without stroke, younger adults (≤68 y) tend to benefit more from active regular and active LPA, whereas older men (>68 y) are predicted to gain more from the active WW pattern. In summary, stroke, age, sex, BP class, and antihypertension medication are key factors driving heterogeneity in individuals' optimal PA patterns. Notably, the baseline PA pattern denotes a lower level of PA rather than complete physical inactivity or sedentary behavior.

Figure 3. The derived conditional inference tree. The models were constructed by the optimal PA pattern predicted (outcome) and all covariates selected (predictors). Antihyper med: antihypertension medication; BP: blood pressure; LPA: light physical activity; WW: weekend warrior.

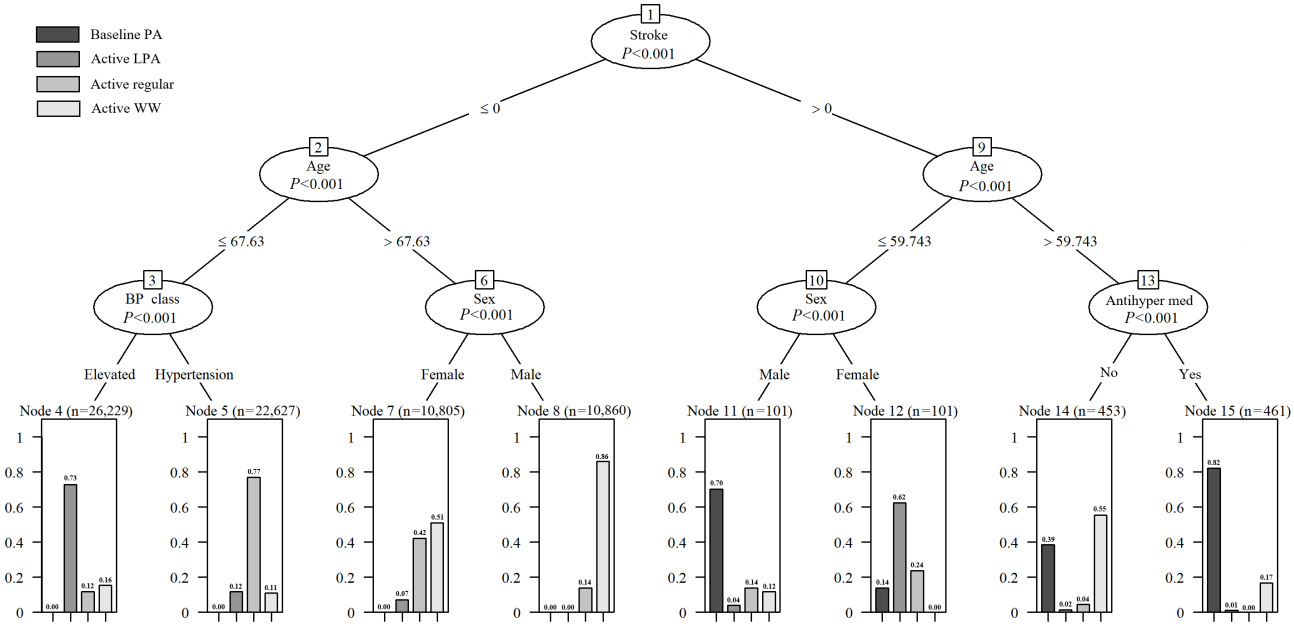

The Associations Between Inconsistent PA Patterns and Mortality

The PA patterns observed in 71.9% of participants in the UKB cohort and 78.9% of participants in the NHANES cohort were inconsistent with the optimal patterns predicted by our model. Compared to individuals whose actual and predicted optimal PA patterns were consistent, an inconsistent PA pattern was associated with a 28% increase in all-cause mortality risk on average (HR 1.28, 95% CI 1.20 - 1.38; Figure 4A) in the UKB cohort. Specifically, an inconsistent MVPA pattern was associated with a 31% increased risk (HR 1.31, 95% CI 1.22 - 1.41; Figure 4B), whereas an inconsistent LPA pattern was associated with a 14% increased risk (HR 1.14, 95% CI 1.01 - 1.30; Figure 4B). In subgroup analysis, we found no statistically significant association between the inconsistent LPA pattern and all-cause mortality (HR 1.06, 95% CI 0.91 - 1.23; Figure 5).

XSL•FO
RenderX

Figure 4. Forest plot of hazard ratios (HRs) from the multivariate Cox regression model. Outcome: all-cause mortality. Exposure: inconsistent PA patterns (distinguishing between individuals with consistent [actual vs predicted optimal PA patterns] and inconsistent PA patterns). In Figure 4A, the exposure is a binary variable (consistent vs inconsistent); in Figure 4B, it is a categorical variable with 3 levels (consistent [reference group], inconsistent MVPA pattern, and inconsistent LPA pattern). The Cox regressions were analyzed and sorted by HR. BP: blood pressure; HbA_{1c}: glycated hemoglobin; LPA: light physical activity; MVPA: moderate to vigorous physical activity; PA: physical activity; UKB: UK Biobank.

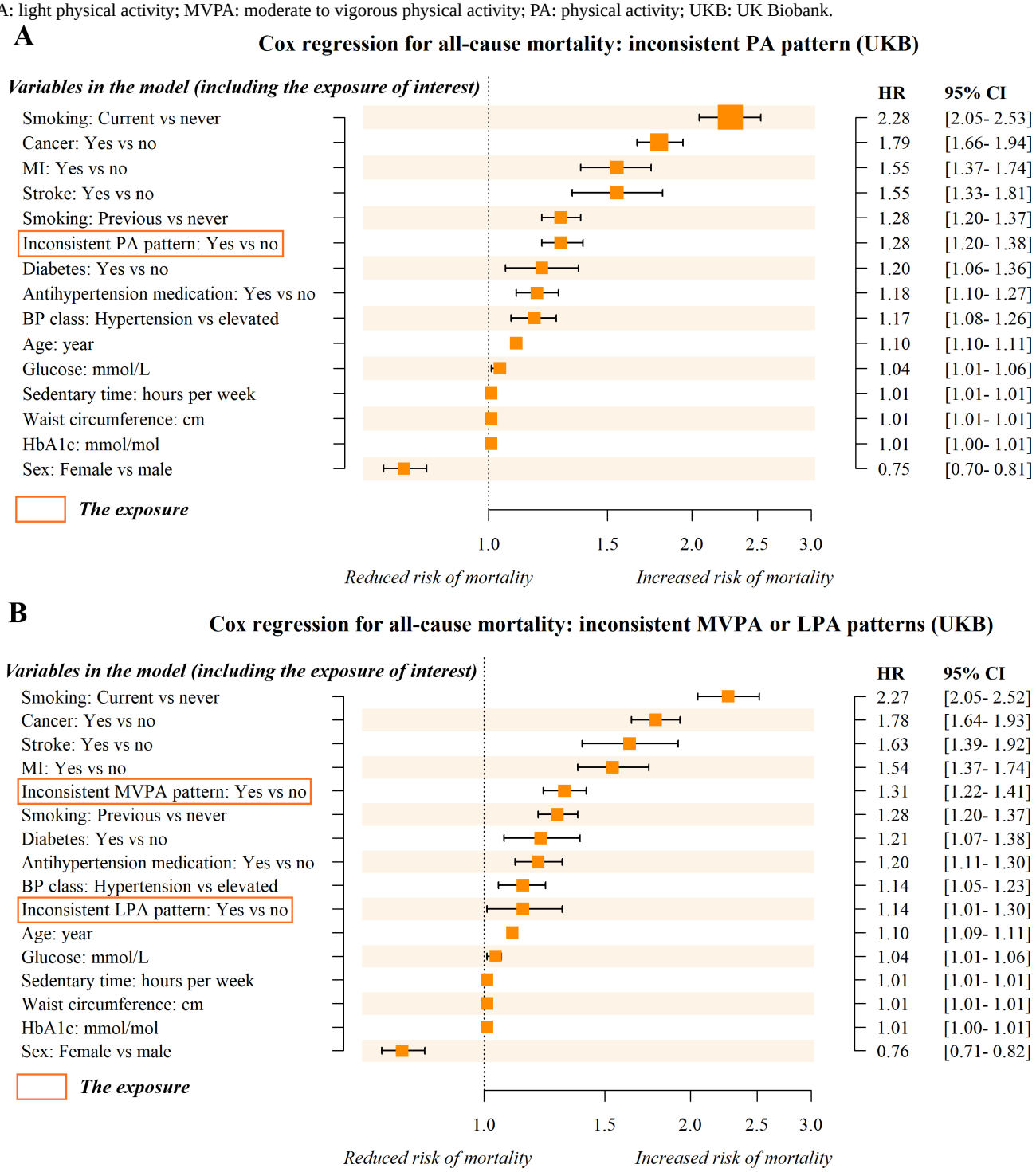

XSL•FO
RenderX

Figure 5. Forest plot of hazard ratios (HRs) from the multivariate Cox regression model in the nonbaseline subgroup. Outcome: all-cause mortality. Exposure: inconsistent PA patterns (distinguishing between individuals with consistent [actual vs predicted optimal PA patterns] and inconsistent PA patterns) with 3 levels (consistent [reference group], inconsistent MVPA pattern, and inconsistent LPA pattern). The nonbaseline group comprises participants whose actually observed PA pattern was not the "baseline PA" pattern. The Cox regressions were analyzed and sorted by HR. BP: blood pressure; HbA_{1c}: glycated hemoglobin; LPA: light physical activity; MVPA: moderate to vigorous physical activity; PA: physical activity; UKB: UK Biobank.

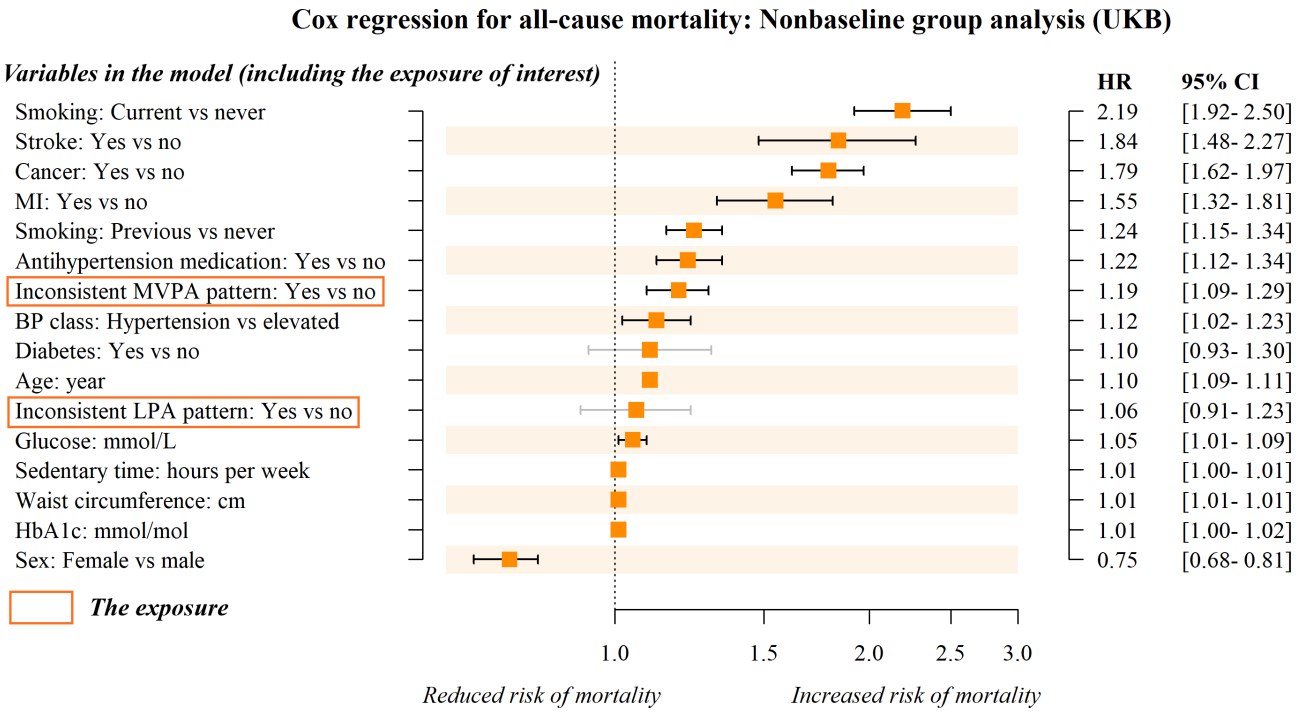

Sensitivity Analyses

The results of the sensitivity analyses are demonstrated in Multimedia Appendix 12. The results of the first sensitivity analysis using the NHANES data showed that an inconsistent PA pattern was associated with a 26% increase in all-cause mortality risk (HR 1.26, 95% CI 1.06 - 1.51), compared to individuals with consistent actual and predicted optimal PA patterns. The analyses of the associations between inconsistent MVPA or LPA patterns and mortality, as well as subgroup analyses, also yielded results similar to those in the UKB cohort. The results of the second sensitivity analysis using cause-specific mortality outcomes indicated that an inconsistent PA pattern was associated with an 18% increase in cancer mortality risk (HR 1.18, 95% CI 1.07 - 1.30) and a 33% increase in CVD mortality risk (HR 1.33, 95% CI 1.18 - 1.49). Subgroup analyses revealed that the inconsistent LPA pattern was not statistically significantly associated with either cancer or CVD mortality in the UKB cohort. In the third sensitivity analysis, we used the prepandemic dataset for external validation. The results showed that the Brier score of our model for predicting 5-year mortality was 2.0% (95% CI 1.8% - 2.3%), and the AUC was 78.1% (95% CI 75.3% - 80.8%). In addition, both the overall and subgroup analyses indicated no statistically significant associations between the inconsistent LPA pattern and all-cause mortality. The results of the fourth sensitivity analysis were consistent with those of the primary analysis.

Discussion

Principal Findings

On the basis of the analysis of 2 temporally and geographically distinct cohorts, we found that patients' individual characteristics modified the benefit of PA patterns on all-cause mortality. Our findings in recognition of the best potential PA pattern revealed that up to 67.6% of patients with high BP were advised to engage in over 150 minutes of MVPA per week (37.2% for active regular and 30.4% for active WW). However, we also found that only 31.8% of them were actually with the predicted optimal PA pattern. In other words, according to the prediction model, a significant portion of the participants may not follow the most suitable PA pattern based on their underlying personal characteristics. Moreover, we found that the risk of death was significantly lower when current PA patterns were consistent with the predicted optimal patterns. This suggests the urgent need to implement standardized, individualized, and evidence-based PA pattern recommendations for patients with hypertension to enhance their prognosis.

Our subgroup analysis of the nonbaseline participants revealed no statistically significant associations between the inconsistent LPA pattern and all-cause mortality. As the subgroup with an inconsistent LPA pattern among nonbaseline participants was predominantly engaged in MVPA, this null finding suggests that increasing PA intensity and volume beyond the predicted levels may not offer additional benefits. Therefore, identifying such subgroups may be important, as simple goals could increase patients' confidence [32] and adherence [33] to PA engagement. However, while our study strongly supports the importance of

XSL•FO
RenderX

meeting the minimum PA intensity threshold, evidence on the potential harm of excessive exercise beyond the predicted level remains weak. This warrants further studies with stronger causal evidence, such as randomized controlled trials [34,35].

Hypertension frequently coexists with cardiovascular disease, obesity, diabetes, or cognitive decline, particularly among older adults [36-39]. These comorbidities complicate PA management in hypertensive patients due to two key challenges: (1) older people with comorbidities tend to have limited physical ability to engage in MVPA [40] and (2) VPA may attenuate PA's cardiovascular benefits by inducing adverse cardiac remodeling in vulnerable individuals [41-43]. Consequently, PA regimens should be tailored to individual physical conditions, and our model provides insights to guide this personalization. For instance, the identified interaction terms reveal that the associations between specific PA patterns and mortality differ by comorbidity. Specifically, for patients with MI who are engaged in cardiac rehabilitation, regular MVPA may offer greater benefits [44]. Hypertensive individuals with diabetes may benefit more from light-intensity PA, as excessive exercise could lead to hypoglycemia [45]. Additionally, participants with a stroke comorbidity are more likely to benefit from a lower level of PA due to the physical limitations and health risks associated with stroke recovery. Light-intensity PA offers a safer, more sustainable option, promoting cardiovascular health and functional recovery without overwhelming the individual [46,47]. However, for high-comorbidity groups, the model-predicted optimal pattern may serve as a prognostic marker rather than a prescriptive target for exercise limitation due to potential unmeasured confounders.

Advances in artificial intelligence have the potential to make personalized health advice more accessible and affordable for a broader range of individuals [48,49]. The model derived and online application proposed in this study could simulate various prognostic scenarios under different PA patterns for adults with hypertension and help patients choose the individualized PA pattern. Importantly, our algorithm can be integrated into current practices that rely on in-person assessments and regular follow-ups, as a support but not as a replacement for the physicians' professionalism and experience [50].

Study Limitations

This study also has several limitations. First, PA patterns were assessed at baseline using a short-term (7 d) estimate, whereas the median follow-up was 9.5 years for UKB and 14.3 years for NHANES. This may not accurately reflect typical PA patterns over time. Additionally, our predictive model assumes a static PA pattern, which limits its clinical applicability. Second, although we prioritized covariate data from the repeat assessment (2012 - 2013) to align with the accelerometer measurement period (2013 - 2015), most of the covariate data were sourced from the baseline assessment (2006 - 2010). This introduces the risk of measurement error due to temporal changes in lifestyle or health status, thereby compromising the tool's real-world predictive accuracy. Third, the ML model we developed was based on the observation survey rather than a randomized controlled trial, so the capacity of our results to elucidate causal relationships is limited. Fourth, the observed association between the "inconsistent PA pattern" and higher mortality risk may be influenced by residual confounding from unmeasured health-seeking behaviors. While our sensitivity analysis results have enhanced the reliability of our conclusions, the unmeasured confounders cannot be overlooked. Fifth, the majority of participants were White, which limits the generalizability of the model and means its predictions may have limited applicability to non-White populations. Sixth, the prediction model was trained and validated using objective accelerometer data; however, the use of self-reported sedentary time as input in the web application may result in inaccurate survival predictions [51]. In addition, the reliance on objective sedentary time and the need for recent blood biochemical marker data may limit the tool's applicability to the general public and primary prevention. Finally, as the optimal PA pattern driven by small increases in survival probability may lack clinical significance, establishing an appropriate threshold should be an important focus for future research. In addition, the threshold selection of the "active LPA pattern" merits further inquiry to verify its biological and clinical significance.

Conclusions

Our study found that the optimal PA pattern was heterogeneous based on individuals' underlying characteristics, whereas only a few participants actually followed the most suitable PA pattern as we predicted. The algorithm presented herein could assist in assigning patients with high BP to the individualized PA pattern based on their specific characteristics, which can be easily accessed through an online application.

Acknowledgments

All authors are grateful to the UK Biobank and the National Health and Nutrition Examination Survey participants and professionals. The authors declare that no generative AI was used in any portion of the manuscript writing.

Funding

This work was supported by the National Social Science Fund of China (21CTJ009).

Data Availability

The datasets generated or analyzed during this study are available in the UK Biobank [52] and the National Health and Nutrition Examination Survey [53] repository.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: YY, FC
Data curation: BM, YY
Formal analysis: YY, MC, WH
Funding acquisition: FC
Methodology: YY, MC, WH
Project administration: FC, BM
Resources: FC, BM
Software: YY, MC, WH, YF, XL, ZL, HF
Supervision: FC, BM, YZ, LP
Validation: all authors
Visualization: YY, MC, WH
Writing – original draft: YY
Writing – review & editing: YY, MC, WH, FC, BM, YZ, LP

Conflicts of Interest

None declared.

Multimedia Appendix 1
The analysis of the dose-response relationship between light physical activity and all-cause mortality.
[DOCX File, 70 KB - jmir_v28i1e78492_app1.docx]

Multimedia Appendix 2
Covariates ascertainment and descriptions.
[DOCX File, 42 KB - jmir_v28i1e78492_app2.docx]

Multimedia Appendix 3
The reliability measurements of covariates.
[DOCX File, 25 KB - jmir_v28i1e78492_app3.docx]

Multimedia Appendix 4
Study flow for National Health and Nutrition Examination Survey.
[DOCX File, 44 KB - jmir_v28i1e78492_app4.docx]

Multimedia Appendix 5
Time to censoring and time to death for the UK Biobank and National Health and Nutrition Examination Survey sets.
[DOCX File, 224 KB - jmir_v28i1e78492_app5.docx]

Multimedia Appendix 6
Baseline characteristics of participants.
[DOCX File, 39 KB - jmir_v28i1e78492_app6.docx]

Multimedia Appendix 7
The detailed results obtained using the least absolute shrinkage and selection operator penalized Cox model for covariate selection.
[DOCX File, 73 KB - jmir_v28i1e78492_app7.docx]

Multimedia Appendix 8
The associations between covariates selected and all-cause mortality by univariate Cox models in the UK Biobank cohort.
[DOCX File, 24 KB - jmir_v28i1e78492_app8.docx]

Multimedia Appendix 9
Coefficients and SEs of the Cox prediction model.
[DOCX File, 26 KB - jmir_v28i1e78492_app9.docx]

Multimedia Appendix 10
A snapshot from the web application.

XSL•FO
RenderX

[DOCX File, 567 KB - jmir_v28i1e78492_app10.docx]

Multimedia Appendix 11
A baseline characteristic of participants stratified by the predicted optimal physical activity pattern in the National Health and Nutrition Examination Survey cohort.
[DOCX File, 25 KB - jmir_v28i1e78492_app11.docx]

Multimedia Appendix 12
The results of the sensitivity analyses.
[DOCX File, 1535 KB - jmir_v28i1e78492_app12.docx]

References

1. Global report on hypertension: the race against a silent killer. : World Health Organization; 2023 URL: https://www.who.int/publications/i/item/9789240081062 [accessed 2023-09-19]
2. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants. Lancet 2021 Sep 11;398(10304):957-980. [doi: 10.1016/S0140-6736(21)01330-1] [Medline: 34450083]
3. McEvoy JW, McCarthy CP, Bruno RM, et al. 2024 ESC guidelines for the management of elevated blood pressure and hypertension: developed by the task force on the management of elevated blood pressure and hypertension of the European Society of Cardiology (ESC) and endorsed by the European Society of Endocrinology (ESE) and the European Stroke Organisation (ESO). Eur Heart J 2024 Jul;45(38):3912-4018. [doi: 10.1093/eurheartj/ehae178]
4. Whelton PK, Carey RM, Aronow WS, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the prevention, detection, evaluation, and management of high blood pressure in adults. J Am Coll Cardiol 2018 May;71(19):e127-e248. [doi: 10.1016/j.jacc.2017.11.006]
5. Zhou YF, Liu N, Wang P, et al. Cost-effectiveness of drug treatment for Chinese patients with stage I hypertension according to the 2017 hypertension clinical practice guidelines. Hypertension 2020 Sep;76(3):750-758. [doi: 10.1161/HYPERTENSIONAHA.119.14533] [Medline: 32713271]
6. Byrd JB, Brook RD. Hypertension. Ann Intern Med 2019 May 7;170(9):ITC65-ITC80. [doi: 10.7326/AITC201905070] [Medline: 31060074]
7. Dzau VJ, Hodgkinson CP. Precision hypertension. Hypertension 2024 Apr;81(4):702-708. [doi: 10.1161/HYPERTENSIONAHA.123.21710] [Medline: 38112080]
8. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q 2004;82(4):661-687. [doi: 10.1111/j.0887-378X.2004.00327.x] [Medline: 15595946]
9. Dos Santos M, Ferrari G, Lee DH, et al. Association of the “weekend warrior” and other leisure-time physical activity patterns with all-cause and cause-specific mortality: a nationwide cohort study. JAMA Intern Med 2022 Aug 1;182(8):840-848. [doi: 10.1001/jamainternmed.2022.2488] [Medline: 35788615]
10. Füzéki E, Engeroff T, Banzer W. Health benefits of light-intensity physical activity: a systematic review of accelerometer data of the National Health and Nutrition Examination Survey (NHANES). Sports Med 2017 Sep;47(9):1769-1793. [doi: 10.1007/s40279-017-0724-0] [Medline: 28393328]
11. Hamaya R, Shiroma EJ, Moore CC, Buring JE, Evenson KR, Lee IM. Time-vs step-based physical activity metrics for health. JAMA Intern Med 2024 Jul 1;184(7):718-725. [doi: 10.1001/jamainternmed.2024.0892] [Medline: 38767892]
12. Cao Y, Baumgartner KB, Visvanathan K, Boone SD, Baumgartner RN, Connor AE. Ethnic and biological differences in the association between physical activity and survival after breast cancer. NPJ Breast Cancer 2020;6:51. [doi: 10.1038/s41523-020-00194-5] [Medline: 33083530]
13. Ji H, Gulati M, Huang TY, et al. Sex differences in association of physical activity with all-cause and cardiovascular mortality. J Am Coll Cardiol 2024 Feb 27;83(8):783-793. [doi: 10.1016/j.jacc.2023.12.019] [Medline: 38383092]
14. Alosh M, Huque MF, Bretz F, D’Agostino RB Sr. Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials. Stat Med 2017 Apr 15;36(8):1334-1360. [doi: 10.1002/sim.7167] [Medline: 27891631]
15. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine-reporting of subgroup analyses in clinical trials. N Engl J Med 2007 Nov 22;357(21):2189-2194. [doi: 10.1056/NEJMsr077003] [Medline: 18032770]
16. Kent DM, Paulus JK, van Klaveren D, et al. The predictive approaches to treatment effect heterogeneity (PATH) statement. Ann Intern Med 2020 Jan 7;172(1):35-45. [doi: 10.7326/M18-3667] [Medline: 31711134]
17. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. Proc Natl Acad Sci U S A 2019 Mar 5;116(10):4156-4165. [doi: 10.1073/pnas.1804597116] [Medline: 30770453]
18. Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. PLOS Med 2007 Oct 16;4(10):e297. [doi: 10.1371/journal.pmed.0040297] [Medline: 17941715]

XSL•FO
RenderX

19. Walmsley R, Chan S, Smith-Byrne K, et al. Reallocation of time between device-measured movement behaviours and risk of incident cardiovascular disease. Br J Sports Med 2022 Sep;56(18):1008-1017. [doi: 10.1136/bjsports-2021-104050]
20. Saint-Maurice PF, Troiano RP, Bassett DR Jr, et al. Association of daily step count and step intensity with mortality among US adults. JAMA 2020 Mar 24;323(12):1151-1160. [doi: 10.1001/jama.2020.1382] [Medline: 32207799]
21. Smirnova E, Leroux A, Cao Q, et al. The predictive performance of objective measures of physical activity derived from accelerometry data for 5-year all-cause mortality in older adults: national health and nutritional examination survey 2003-2006. J Gerontol A Biol Sci Med Sci 2020 Sep 16;75(9):1779-1785. [doi: 10.1093/gerona/glz193] [Medline: 31504213]
22. Khurshid S, Al-Alusi MA, Churchill TW, Guseh JS, Ellinor PT. Accelerometer-derived "weekend warrior" physical activity and incident cardiovascular disease. JAMA 2023 Jul 18;330(3):247-252. [doi: 10.1001/jama.2023.10875] [Medline: 37462704]
23. Xiang B, Zhou Y, Wu X, Zhou X. Association of device-measured physical activity with cardiovascular outcomes in individuals with hypertension. Hypertension 2023 Nov;80(11):2455-2463. [doi: 10.1161/HYPERTENSIONAHA.123.21663] [Medline: 37667966]
24. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86(2):420-428. [doi: 10.1037/0033-2909.86.2.420]
25. Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of kappa and weighted kappa. Psychol Bull 1969 Nov;72(5):323-327. [doi: 10.1037/h0028106]
26. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med 1997 Feb 28;16(4):385-395. [doi: 10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3] [Medline: 9044528]
27. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. J Stat Softw 2011 Mar;39(5):1-13. [doi: 10.18637/jss.v039.i05] [Medline: 27065756]
28. Derksen S, Keselman HJ. Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. Brit J Math Statis 1992 Nov;45(2):265-282. [doi: 10.1111/j.2044-8317.1992.tb00992.x]
29. Hung H, Chiang CT. Estimation methods for time‐dependent AUC models with survival data. Can J Statistics 2010 Mar;38(1):8-26. [doi: 10.1002/cjs.10046]
30. Gerds TA, Schumacher M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. Biom J 2006 Dec;48(6):1029-1040. [doi: 10.1002/bimj.200610301] [Medline: 17240660]
31. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. J Comput Graph Stat 2006 Sep;15(3):651-674. [doi: 10.1198/106186006X133933]
32. Anderson CL, Feldman DB. Hope and physical exercise: the contributions of hope, self-efficacy, and optimism in accounting for variance in exercise frequency. Psychol Rep 2020 Aug;123(4):1145-1159. [doi: 10.1177/0033294119851798] [Medline: 31142190]
33. Li Q, Jiang J, Duan A, Hu J, Li L, Chen W. Physical activity experience of patients with hypertension: a systematic review and synthesis of qualitative literature. BMC Public Health 2024 Oct 15;24(1):2826. [doi: 10.1186/s12889-024-20326-x]
34. Bothwell LE, Greene JA, Podolsky SH, Jones DS. Assessing the gold standard--lessons from the history of RCTs. N Engl J Med 2016 Jun 2;374(22):2175-2181. [doi: 10.1056/NEJMms1604593] [Medline: 27248626]
35. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000 Jun 22;342(25):1887-1892. [doi: 10.1056/NEJM200006223422507] [Medline: 10861325]
36. Lauder L, Mahfoud F, Azizi M, et al. Hypertension management in patients with cardiovascular comorbidities. Eur Heart J 2023 Jun 20;44(23):2066-2077. [doi: 10.1093/eurheartj/ehac395] [Medline: 36342266]
37. Jia G, Sowers JR, Whaley-Connell AT. Obesity in hypertension: the role of the expanding waistline over the years and insights into the future. Hypertension 2024 Apr;81(4):687-690. [doi: 10.1161/HYPERTENSIONAHA.123.21719] [Medline: 38018438]
38. Pacholko A, Iadecola C. Hypertension, neurodegeneration, and cognitive decline. Hypertension 2024 May;81(5):991-1007. [doi: 10.1161/HYPERTENSIONAHA.123.21356] [Medline: 38426329]
39. Kokkinos P, Faselis C, Pittaras A, et al. Stroke incidence in patients with hypertension according to cardiorespiratory fitness. Hypertension 2024 Aug;81(8):1747-1757. [doi: 10.1161/HYPERTENSIONAHA.124.23066] [Medline: 38841839]
40. Shi H, Hu FB, Huang T, et al. Sedentary behaviors, light-intensity physical activity, and healthy aging. JAMA Netw Open 2024 Jun 3;7(6):e2416300. [doi: 10.1001/jamanetworkopen.2024.16300] [Medline: 38861256]
41. Ho FK, Zhou Z, Petermann-Rocha F, et al. Association between device-measured physical activity and incident heart failure: a prospective cohort study of 94 739 UK Biobank participants. Circulation 2022 Sep 20;146(12):883-891. [doi: 10.1161/CIRCULATIONAHA.122.059663] [Medline: 36036153]
42. Keating CJ, Montilla J, Román P, Del Castillo RM. Comparison of high-intensity interval training to moderate-intensity continuous training in older adults: a systematic review. J Aging Phys Act 2020 Oct 1;28(5):798-807. [doi: 10.1123/japa.2019-0111] [Medline: 32303000]
43. Basavarajaiah S, Boraita A, Whyte G, et al. Ethnic differences in left ventricular remodeling in highly-trained athletes relevance to differentiating physiologic left ventricular hypertrophy from hypertrophic cardiomyopathy. J Am Coll Cardiol 2008 Jun 10;51(23):2256-2262. [doi: 10.1016/j.jacc.2007.12.061] [Medline: 18534273]

XSL•FO
RenderX

44. Piepoli MF, Hoes AW, Agewall S, et al. 2016 European Guidelines on cardiovascular disease prevention in clinical practice. Atherosclerosis 2016 Sep;252(29):207-274. [doi: 10.1016/j.atherosclerosis.2016.05.037]
45. Colberg SR, Sigal RJ, Yardley JE, et al. Physical activity/exercise and diabetes: a position statement of the American diabetes association. Diabetes Care 2016 Nov;39(11):2065-2079. [doi: 10.2337/dc16-1728] [Medline: 27926890]
46. Fini NA, Bernhardt J, Said CM, Billinger SA. How to address physical activity participation after stroke in research and clinical practice. Stroke 2021 Jun;52(6):e274-e277. [doi: 10.1161/STROKEAHA.121.034557] [Medline: 33951930]
47. Billinger SA, Arena R, Bernhardt J, et al. Physical activity and exercise recommendations for stroke survivors: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke 2014 Aug;45(8):2532-2553. [doi: 10.1161/STR.0000000000000022] [Medline: 24846875]
48. Tegegne TK, Shifti DM, Rawstorn JC, et al. Digital lifestyle interventions for cardiovascular risk reduction: a systematic review and network meta-analysis. Health Policy Technol 2024 Aug;13(3):100879. [doi: 10.1016/j.hlpt.2024.100879]
49. Nnamoko N, Cabrera-Diego LA, Campbell D, Sanders G, Fairclough SJ, Korkontzelos I. Personalised accelerometer cut-point prediction for older adults' movement behaviours using a machine learning approach. Comput Methods Programs Biomed 2021 Sep;208:106165. [doi: 10.1016/j.cmpb.2021.106165] [Medline: 34118492]
50. Sezgin E. Artificial intelligence in healthcare: complementing, not replacing, doctors and healthcare providers. Digital health 2023;9:20552076231186520. [doi: 10.1177/20552076231186520] [Medline: 37426593]
51. Lines RLJ, Ntoumanis N, Thøgersen-Ntoumani C, et al. Cross-sectional and longitudinal comparisons of self-reported and device-assessed physical activity and sedentary behaviour. J Sci Med Sport 2020 Sep;23(9):831-835. [doi: 10.1016/j.jsams.2020.03.004] [Medline: 32312612]
52. UK Biobank. URL: https://www.ukbiobank.ac.uk [accessed 2025-12-22]
53. National Health and Nutrition Examination Survey. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/nhanes [accessed 2025-12-22]

Abbreviations

AUC: area under the receiver operating characteristic curve
BP: blood pressure
CIT: conditional inference tree
CVD: cardiovascular disease
DBP: diastolic blood pressure
HbA_{1c}: glycated hemoglobin
HR: hazard ratio
LASSO: least absolute shrinkage and selection operator
LPA: light physical activity
MI: myocardial infarction
ML: machine learning
MVPA: moderate to vigorous physical activity
NHANES: National Health and Nutrition Examination Survey
PA: physical activity
ROC: receiver operating characteristic
SBP: systolic blood pressure
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
UKB: UK Biobank
WW: weekend warrior

Edited by A Coristine, TDA Cardoso; submitted 04.Jun.2025; peer-reviewed by L Dan, SM Lai; revised version received 02.Dec.2025; accepted 03.Dec.2025; published 09.Jan.2026.

Please cite as:
Yang Y, Chen M, Hu W, Fu Y, Li X, Liao Z, Feng H, Zhao Y, Pei L, Mi B, Chen F
Physical Activity Recommendations Tailored by a Predictive Model for Adults With High Blood Pressure: Observational Study
J Med Internet Res 2026;28:e78492
URL: https://www.jmir.org/2026/1/e78492
doi:10.2196/78492

©Yuhui Yang, Manqing Chen, Weiwei Hu, Yifan Fu, Xingyan Li, Zhenli Liao, Hongman Feng, Yaling Zhao, Leilei Pei, Baibing Mi, Fangyao Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Jan.2026. This is

XSL•FO
RenderX

an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Beyond Visual Consensus: Tiered Reference Framework for AI Cystoscopy Studies

Ahmet Murat Bayraktar, MD; Bilgi İşler, MD

Department of Urology, Konya City Hospital, Akabe Mah, Adana Çevre Yolu Cad. No.135 Karatay, Konya, Turkey

Corresponding Author:
Ahmet Murat Bayraktar, MD
Department of Urology, Konya City Hospital, Akabe Mah, Adana Çevre Yolu Cad. No.135 Karatay, Konya, Turkey

Related Articles:

https://www.jmir.org/2026/1/e87193

Comment in: https://www.jmir.org/2026/1/e103335

(J Med Internet Res 2026;28:e101910) doi:10.2196/101910

KEYWORDS

multimodal; large language model; AI; cystoscopy; diagnostic reasoning; finding description; biopsy indication; bladder tumor; artificial intelligence

We read with great interest the study by Shih et al [1], a valuable contribution to the emerging field of artificial intelligence (AI)–assisted cystoscopic diagnosis. Their blinded evaluation of four multimodal large language models across 401 images encompassing 40 cystoscopic finding subcategories provides important insights into current model capabilities. We wish to raise a methodological consideration regarding the reference standard that may inform the interpretation of the reported findings.

The reference standard in this study was established through visual consensus between two urologists, without histopathological confirmation. While interexpert agreement was satisfactory (κ=0.81), cystoscopic impression alone has well-documented limitations. Cina et al [2] demonstrated that experienced urologists could not reliably distinguish between low- and high-grade papillary lesions endoscopically, with complete grade-stage concordance with histopathology in only 70.3% of cases and a specificity of just 57% for predicting lamina propria invasion. A visually derived reference standard thus carries inherent diagnostic uncertainty.

This concern is particularly relevant for lesion categories central to the study's 7-class task. Carcinoma in situ (CIS) is notoriously difficult to identify under white light cystoscopy; blue light cystoscopy studies have demonstrated that approximately one-third of CIS lesions are missed by white light alone [3]. Similarly, the frequent misclassification of papilloma as papillary urothelial carcinoma—acknowledged by the authors as reflecting substantial macroscopic overlap [1]—underscores that definitive classification of these entities requires histological evaluation of architectural and cytological features indistinguishable on endoscopic inspection.

The AI-assisted cystoscopic diagnosis literature has converged on histopathological confirmation as the reference standard. Foundational work such as CystoNet was trained and validated on histologically confirmed lesions [4], and a recent systematic review by Hengky et al [5] restricted inclusion to studies using histopathology as the reference standard. This consensus reflects the clinical reality that the categorical distinctions central to bladder lesion classification—low- versus high-grade carcinoma, CIS versus inflammation, papilloma versus carcinoma—ultimately rest on histological criteria and drive subsequent management.

We acknowledge the logistical challenges of obtaining histopathology for every image in a large, multisource dataset, particularly for benign-appearing or nonresected findings. We, therefore, suggest that future benchmarking studies adopt a tiered reference framework: (1) histopathologically confirmed labels for all lesions undergoing biopsy or resection, encompassing the full malignant spectrum; (2) enhanced cystoscopy correlation (blue light or narrow band imaging) as an intermediate standard, particularly for CIS [3]; and (3) consensus visual labels—explicitly flagged as lower confidence—for benign categories unlikely to undergo biopsy in routine practice. Stratified performance reporting under such a framework would allow readers to separate genuine algorithmic limitations from ambiguity inherent to the reference standard, providing a more clinically meaningful evaluation.

XSL•FO
RenderX

Acknowledgments

The authors acknowledge the use of generative artificial intelligence (Google Gemini) for language editing and proofreading assistance during the preparation of this manuscript.

Funding

The authors declared no financial support was received for this work.

Conflicts of Interest

None declared.

References

1. Shih YC, Wu CY, Huang SW, Tsai CY. Multimodal large language models for cystoscopic image interpretation and bladder lesion classification: comparative study. J Med Internet Res 2026 Jan 28;28:e87193. [doi: 10.2196/87193] [Medline: 41605505]
2. Cina SJ, Epstein JI, Endrizzi JM, Harmon WJ, Seay TM, Schoenberg MP. Correlation of cystoscopic impression with histologic diagnosis of biopsy specimens of the bladder. Hum Pathol 2001 Jun;32(6):630-637. [doi: 10.1053/hupa.2001.24999] [Medline: 11431718]
3. Grossman HB, Gomella L, Fradet Y, et al. A phase III, multicenter comparison of hexaminolevulinate fluorescence cystoscopy and white light cystoscopy for the detection of superficial papillary lesions in patients with bladder cancer. J Urol 2007 Jul;178(1):62-67. [doi: 10.1016/j.juro.2007.03.034] [Medline: 17499283]
4. Shkolyar E, Jia X, Chang TC, et al. Augmented bladder tumor detection using deep learning. Eur Urol 2019 Dec;76(6):714-718. [doi: 10.1016/j.eururo.2019.08.032] [Medline: 31537407]
5. Hengky A, Lionardi SK, Kusumajaya C. Can artificial intelligence aid the urologists in detecting bladder cancer? Indian J Urol 2024;40(4):221-228. [doi: 10.4103/iju.iju_39_24] [Medline: 39555437]

Abbreviations

AI: artificial intelligence
CIS: carcinoma in situ

Edited by T Leung; submitted 20.May.2026; this is a non–peer-reviewed article;accepted 04.Jun.2026; published 18.Jun.2026.

Please cite as:
Bayraktar AM, İşler B
Beyond Visual Consensus: Tiered Reference Framework for AI Cystoscopy Studies
J Med Internet Res 2026;28:e101910
URL: https://www.jmir.org/2026/1/e101910
doi:10.2196/101910

© Ahmet Murat Bayraktar, Bilgi İşler. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Authors' Reply: Beyond Visual Consensus: Tiered Reference Framework for AI Cystoscopy Studies

Chung-You Tsai[1,2,3], MD, PhD; Shi-Wei Huang[3,4], MD, PhD

[1]Division of Urology, Department of Surgery, Far Eastern Memorial Hospital, New Taipei, Taiwan
[2]Department of Electrical Engineering, Yuan Ze University, Taoyuan, Taiwan
[3]Department of Urology, College of Medicine, National Taiwan University, Taipei, Taiwan
[4]Department of Urology, National Taiwan University Hospital, Yunlin Branch, No. 579, Sec. 2, Yunlin Rd., Douliou City, Yunlin, Taiwan

Corresponding Author:
Shi-Wei Huang, MD, PhD
Department of Urology, College of Medicine, National Taiwan University, Taipei, Taiwan

Related Articles:

https://www.jmir.org/2026/1/e87193

https://www.jmir.org/2026/1/e101910

(J Med Internet Res 2026;28:e103335) doi:10.2196/103335

KEYWORDS

multimodal; large language model; AI; cystoscopy; diagnostic reasoning; finding description; biopsy indication; bladder tumor; artificial intelligence

We thank the reader [1] for their thoughtful and constructive comments on our study [2]. We would like to clarify an important methodological point regarding the reference standard used in our study.

The 401 cystoscopic images were curated from established educational and clinical sources, including reference atlases, PubMed-indexed articles and repositories, clinical websites, industry archives, and Creative Commons–licensed educational videos. These images were originally accompanied by source-provided diagnoses, educational labels, captions, or contextual information.

As stated in the Study Design section, "The reference standard diagnoses were determined through a multiphase, consensus-based process." To further clarify how source-provided diagnostic information was incorporated, this process can be described in two key phases. In the initial independent review phase, source-provided diagnostic labels were withheld, and two urological experts independently inspected each image and answered the prespecified Q1 to Q5 framework, including anatomic site, cystoscopic findings, lesion detection, lesion reasoning, and final diagnosis. In the subsequent consensus phase, the available source-provided diagnoses or educational labels were disclosed to the experts and considered together with the cystoscopic appearance to resolve discrepancies and establish the final reference answers for Q1 to Q5.

Therefore, the final consensus diagnosis for each image was not based solely on independent visual inspection by two urologists. Rather, the reference standard was established through a source-informed expert consensus process, integrating source-provided diagnostic information with expert review of the cystoscopic appearance.

Thus, the reference standard in our study was designed to support the evaluation of model-generated cystoscopic interpretation and reasoning, rather than to independently re-establish or revalidate the original pathological diagnosis of each image. We agree that histopathology remains the definitive standard for pathological lesion classification, particularly for entities such as carcinoma in situ, papilloma, and papillary urothelial carcinoma. Accordingly, our study should be read as an image interpretation and reasoning benchmark based on images with pre-existing source-provided diagnostic or educational information, with final reference labels established through expert consensus after consideration of both the cystoscopic appearance and the diagnoses provided by the original educational or clinical sources.

Acknowledgments

Generative artificial intelligence assistance was used for language editing. The authors reviewed, revised, and approved the final content and take full responsibility for the submitted manuscript.

XSL•FO
RenderX

Funding

The authors declared no financial support was received for this work.

Conflicts of Interest

None declared.

References

1. Bayraktar AM, İşler B. Beyond visual consensus: tiered reference framework for AI cystoscopy studies. J Med Internet Res 2026;28:e101910. [doi: 10.2196/101910]
2. Shih YC, Wu CY, Huang SW, Tsai CY. Multimodal large language models for cystoscopic image interpretation and bladder lesion classification: comparative study. J Med Internet Res 2026 Jan 28;28:e87193. [doi: 10.2196/87193] [Medline: 41605505]

Edited by T Leung; submitted 03.Jun.2026; this is a non–peer-reviewed article;accepted 04.Jun.2026; published 18.Jun.2026.

Please cite as:
Tsai CY, Huang SW
Authors' Reply: Beyond Visual Consensus: Tiered Reference Framework for AI Cystoscopy Studies
J Med Internet Res 2026;28:e103335
URL: https://www.jmir.org/2026/1/e103335
doi:10.2196/103335

© Chung-You Tsai, Shi-Wei Huang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Human-in-the-Loop as a Safety Guardrail: Clinical Accountability in the Large Language Model Era

Isaac Zablah[1*], PhD; Yolly Molina[2*], MSc; Antonio Garcia-Loureiro[3*], PhD

[1]Faculty of Medical Sciences, National Autonomous University of Honduras, Calle la Salud SN, Tegucigalpa, Honduras
[2]Center for Biomedical Imaging Diagnostics Research and Rehabilitation, National Autonomous University of Honduras, Tegucigalpa, Honduras
[3]Department of Electronics and Computer Science, Universidade de Santiago de Compostela, Santiago de Compostela, Spain
[*]all authors contributed equally

Corresponding Author:
Isaac Zablah, PhD
Faculty of Medical Sciences, National Autonomous University of Honduras, Calle la Salud SN, Tegucigalpa, Honduras

Related Article:

https://www.jmir.org/2025/1/e59069

(J Med Internet Res 2026;28:e85726) doi:10.2196/85726

KEYWORDS

large language models; high-performance computing; medical informatics; computational efficiency; clinical decision support; artificial intelligence; healthcare infrastructure; model optimization

We found Zhang et al's thorough review of the transformative potential of large language models (LLMs) in healthcare to be very interesting [1]. The authors do a great job of talking about clinical applications, data integration, and ethical issues. However, we think that important aspects of computational performance need more attention, especially when it comes to using technology in real-world healthcare settings where resources are limited.

Zhang et al mention that "*technological advancements*" are helping to meet the "*high hardware requirements*" of LLMs [1], but the reality of computing is still a huge challenge. Modern medical LLMs such as GPT-4 and domain-specific models such as Med-PaLM 2 need considerable infrastructure as described below [2]:

- *Inference latency*: Currently, LLMs take 2 to 10 seconds to respond to each query. This may not be fast enough for clinical situations where time is of the essence, like triage in the emergency department or decision support during surgery. More detailed answers need more time [3].
- *Memory footprint*: Models with billions of parameters need 16-80+ GB of VRAM (video random access memory) for fast inference [4]. This means that many health care facilities, especially in low- and middle-income countries, do not have the specialized GPU infrastructure they need.
- *Scalability challenges:* Serving hundreds of concurrent clinical users requires distributed computing architectures and load-balancing strategies not discussed in the review [5].

For edge computing and improving models, we suggest that subsequent research should emphasize:

- *Model quantization and pruning:* Techniques to reduce model size by 50% - 75% with minimal accuracy loss, enabling deployment on consumer-grade hardware.
- *Edge computing solutions:* Local deployment using optimized models (eg, 7-13B parameter variants) to address data privacy concerns while reducing latency and cloud dependency.
- *Hybrid architecture:* Combining lightweight edge models for routine queries with cloud-based full models for complex cases, optimizing the accuracy-efficiency trade-off.

The medical informatics community requires standardized metrics that assess not only diagnostic accuracy but also operations per diagnosis (computational cost), energy consumption per inference (environmental impact), and cost-effectiveness ratios (accuracy gained per dollar of infrastructure). We did an initial benchmarking of three LLMs on differential diagnosis tasks: Clinical Camel (LLaMA-2-13B), PMC-LLaMA 13B, and Meditron-3 (Qwen2.5-14B). We found that smaller, domain-specific models (~14 billion parameters fine-tuned on medical corpora) were able to achieve 85% - 90% of GPT-4's diagnostic accuracy while using only about 15% of the computational resources, indicating considerable room for improvement.

We want high-performance computing research in medical artificial intelligence (AI) to help with clinical implementation. This research should set benchmarks for both computational performance and clinical accuracy, come up with optimization techniques that are specific to medical inference workloads, create reference architectures for deploying LLMs in different health care settings, and investigate federated learning strategies that let training happen without putting sensitive patient data in one place.

XSL•FO
RenderX

The transformative potential Zhang et al describe will only be realized if LLMs can be deployed efficiently and equitably across diverse health care environments. High-performance computing and medical informatics must advance in tandem to bridge the gap between research promise and clinical reality.

Acknowledgments

The authors used the Wordvice.ai service solely to improve the language and semantics of the manuscript.

Funding

The authors declared no financial support was received for this work.

Data Availability

The benchmarking data comparing diagnostic accuracy and computational resource utilization of Clinical Camel (LLaMA-2-13B), PMC-LLaMA 13B, and Meditron-3 (Qwen2.5-14B) against GPT-4 baseline are available from the corresponding author upon reasonable request. The evaluation was conducted on publicly available differential diagnosis case datasets. Model access: Clinical Camel and PMC-LLaMA 13B are available via Hugging Face; Meditron-3 (Qwen2.5-14B) is available through the EPFL repository; and GPT-4 was accessed via OpenAI API for comparative benchmarking.

Authors' Contributions

Conceptualization: AGL
Methodology: JZ
Validation: YM
Formal analysis: AGL
Writing — original draft: JZ, YM
Writing — review & editing: AGL

Conflicts of Interest

None declared.

Editorial Notice

The corresponding author of “Revolutionizing Health Care: The Transformative Impact of Large Language Models in Medicine” declined to respond to this letter.

References

1. Zhang K, Meng X, Yan X, et al. Revolutionizing health care: the transformative impact of large language models in medicine. J Med Internet Res 2025 Jan 7;27:e59069. [doi: 10.2196/59069] [Medline: 39773666]
2. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature New Biol 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
4. Raiaan MAK, Mukta M, Fatema K, et al. A review on large language models: architectures, applications, taxonomies, open issues and challenges. IEEE Access 2024;12:26839-26874. [doi: 10.1109/ACCESS.2024.3365742]
5. He K, Mao R, Lin Q, et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. arXiv. Preprint posted online on 2023. [doi: 10.48550/arXiv.2310.05694]

Abbreviations

AI: artificial intelligence
LLM: large language model
VRAM: video random access memory

XSL•FO
RenderX

Edited by A Mavragani; submitted 12.Oct.2025; this is a non–peer-reviewed article;accepted 08.May.2026; published 18.Jun.2026.

Please cite as:
Zablah I, Molina Y, Garcia-Loureiro A
Human-in-the-Loop as a Safety Guardrail: Clinical Accountability in the Large Language Model Era
J Med Internet Res 2026;28:e85726
URL: https://www.jmir.org/2026/1/e85726
doi:10.2196/85726

© Isaac Zablah, Yolly Molina, Antonio Garcia-Loureiro. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Characterization of Models for Identifying Physical and Cognitive Frailty in Older Adults With Diabetes: Systematic Review and Meta-Analysis

Xia Wang[1*], MS; Shujie Meng[1*], MS; Xiang Xiao[2], BS; Liu Lu[3]; Hongyan Chen[1], MS; Yong Li[4], MEd; Rong Zhang[4]; Qiwu Jiang[5], BM; Shan Liu[6], MS; Ru Gao[6], PhD

[1]School of Basic Medical Sciences and School of Nursing, Chengdu University, Chengdu, China
[2]School of Nursing, Yibin Vocational College of Medicine and Health, Yibin, China
[3]Nursing Department, The Fourth People's Hospital of Yibin, Yibin, China
[4]Rehabilitation College, Sichuan Health Rehabilitation Vocational College, Zigong, China
[5]School of Rehabilitation and Wellness, Yibin Vocational College of Medicine and Health, Yibin, China
[6]Nursing Department, Wenjiang District People's Hospital, No.86 Kangtai Road Wenjiang District, Chengdu, Sichuan Province, China
[*]these authors contributed equally

Corresponding Author:
Ru Gao, PhD
Nursing Department, Wenjiang District People's Hospital, No.86 Kangtai Road Wenjiang District, Chengdu, Sichuan Province, China

Related Article:

Correction of: https://www.jmir.org/2026/1/e84617

(J Med Internet Res 2026;28:e92717) doi:10.2196/92717

In “Characterization of Models for Identifying Physical and Cognitive Frailty in Older Adults With Diabetes: Systematic Review and Meta-Analysis” [1], the authors noted one correction.

Affiliation 5 has been changed from the following:

> *Medical and Nursing College, Yibin Vocational College of Medicine and Health, Yibin, China*

The affiliation now reads:

> *School of Rehabilitation and Wellness, Yibin Vocational College of Medicine and Health, Yibin, China*

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Wang X, Meng S, Xiao X, et al. Characterization of models for identifying physical and cognitive frailty in older adults with diabetes: systematic review and meta-analysis. J Med Internet Res 2026 Jan 29;28:e84617. [doi: 10.2196/84617] [Medline: 41610414]

Submitted 02.Feb.2026; this is a non–peer-reviewed article;accepted 02.Feb.2026; published 12.Feb.2026.

Please cite as:
Wang X, Meng S, Xiao X, Lu L, Chen H, Li Y, Zhang R, Jiang Q, Liu S, Gao R
Correction: Characterization of Models for Identifying Physical and Cognitive Frailty in Older Adults With Diabetes: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e92717
URL: https://www.jmir.org/2026/1/e92717
doi:10.2196/92717

XSL•FO
RenderX

© Xia Wang, Shujie Meng, Xiang Xiao, Liu Lu, Hongyan Chen, Yong Li, Rong Zhang, Qiwu Jiang, Shan Liu, Ru Gao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: User Profiles and Engagement in a Hypertension Self-Management App: Cross-Sectional Survey

Felix Muehlensiepen[1,2,3,4], MPH, PhD, Dr rer medic habil; Susann May[1,2,3], MPH, Dr rer medic; Frances Seifert[1,2,3], MSc; Eileen Wengemuth[1], Dr rer nat; Olen Johannsen[5], MSc; Martin Middeke[6], Prof, Dr; Martin Heinze[1,7], Prof, Dr; Pascal Petit[4], PhD; Nicolas Vuillerme[4], Prof, Dr; Sebastian Spethmann[2,3*], Prof, Dr; Dunja Bruch[8*], MSc

[1]Center for Health Services Research, Faculty of Health Sciences Brandenburg, Brandenburg Medical School, Rüdersdorf, Seebad 82/83, Ruedersdorf bei Berlin, Germany

[2]Department of Cardiology, Angiology and Intensive Care Medicine, Deutsches Herzzentrum der Charité, Berlin, Germany

[3]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[4]Université Grenoble Alpes, CNRS, Grenoble INP, LIG Sangria, Grenoble, France

[5]Hypertension Care UG, Ingolstadt, Germany

[6]Hypertension Center Munich, Munich, Germany

[7]Center for Mental Health, Brandenburg Medical School, Immanuel Klinik Rüdersdorf, Ruedersdorf bei Berlin, Germany

[8]Department of Cardiovascular Surgery, Heart Center Brandenburg, Faculty of Health Sciences, Brandenburg Medical School Theodor Fontane, Bernau, Germany

[*]these authors contributed equally

Corresponding Author:
Felix Muehlensiepen, MPH, PhD, Dr rer medic habil
Center for Health Services Research, Faculty of Health Sciences Brandenburg, Brandenburg Medical School, Rüdersdorf, Seebad 82/83, Ruedersdorf bei Berlin, Germany

Related Article:

Correction of: https://www.jmir.org/2026/1/e83075/

(J Med Internet Res 2026;28:e93515) doi:10.2196/93515

In "User Profiles and Engagement in a Hypertension Self-Management App: Cross-Sectional Survey" [1], the authors noted an error in the order of affiliations 1 to 6, which was introduced during the production process.

The affiliation order was revised from the following:

> *[1]Hypertension Center Munich, Munich, Germany*
>
> *[2]Center for Health Services Research, Faculty of Health Sciences Brandenburg, Brandenburg Medical School, Rüdersdorf, Ruedersdorf bei Berlin, Germany*
>
> *[3]Department of Cardiology, Angiology and Intensive Care Medicine, Deutsches Herzzentrum der Charité, Berlin, Germany*
>
> *[4]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany*
>
> *[5]Université Grenoble Alpes, CNRS, Grenoble INP, LIG Sangria, Grenoble, France*
>
> *[6]Hypertension Care UG, Ingolstadt, Germany*

The order of the affiliations now reads:

> *[1]Center for Health Services Research, Faculty of Health Sciences Brandenburg, Brandenburg Medical School, Rüdersdorf, Ruedersdorf bei Berlin, Germany*
>
> *[2]Department of Cardiology, Angiology and Intensive Care Medicine, Deutsches Herzzentrum der Charité, Berlin, Germany*
>
> *[3]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany*
>
> *[4]Université Grenoble Alpes, CNRS, Grenoble INP, LIG Sangria, Grenoble, France*
>
> *[5]Hypertension Care UG, Ingolstadt, Germany*
>
> *[6]Hypertension Center Munich, Munich, Germany*

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

XSL•FO
RenderX

1. Muehlensiepen F, May S, Seifert F, et al. User profiles and engagement in a hypertension self-management app: cross-sectional survey. J Med Internet Res 2026 Feb 11;28:e83075. [doi: 10.2196/83075] [Medline: 41671553]

Submitted 13.Feb.2026; this is a non–peer-reviewed article;accepted 13.Feb.2026; published 24.Feb.2026.

Please cite as:
Muehlensiepen F, May S, Seifert F, Wengemuth E, Johannsen O, Middeke M, Heinze M, Petit P, Vuillerme N, Spethmann S, Bruch D
Correction: User Profiles and Engagement in a Hypertension Self-Management App: Cross-Sectional Survey
J Med Internet Res 2026;28:e93515
URL: https://www.jmir.org/2026/1/e93515
doi:10.2196/93515

© Felix Muehlensiepen, Susann May, Frances Seifert, Eileen Wengemuth, Olen Johannsen, Martin Middeke, Martin Heinze, Pascal Petit, Nicolas Vuillerme, Sebastian Spethmann, Dunja Bruch. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Leveraging Social Media to Achieve Population-Level Reach of Lung Cancer Screening-Eligible Individuals: A RE-AIM Framework Perspective

Lisa Carter-Bawa[1,2], MSN, MPH, PhD; Jamie S Ostroff[3], PhD; Susan M Rawl[4], PhD; Erin A Hirsch[1], PhD; Smita C Banerjee[3], PhD; Andrew Ciupek[5], PhD; Robert Skipworth Comer[6], MS; Minal Kale[7], MPH, MD; Katherine T Leopold[8], MD; Patrick O Monahan[9], PhD; James E Slaven Jr[9], MS; Francis Valenzona[1], BS; Renda Soylemez Wiener[10,11,12], MPH, MD; Ana Guadalupe Vielma[1], PhD

[1]Cancer Prevention Precision Control Institute, Center for Discovery & Innovation, Hackensack Meridian Health, 123 Metro Blvd, 6th Floor, 6400 Pod, Nutley, NJ, United States

[10]Center for Health Optimization & Implementation Research, VA Boston Healthcare System, Boston, MA, United States

[11]National Center for Lung Cancer Screening, Veterans Health Administration, Washington, DC, United States

[12]The Pulmonary Center, Boston University Chobanian & Avedesian School of Medicine, Boston, MA, United States

[2]Cancer Prevention & Control Program, Georgetown Lombardi Comprehensive Cancer Center, Washington, DC, United States

[3]Department of Psychiatry & Behavioral Science, Memorial Sloan Kettering Cancer Center, New York, NY, United States

[4]School of Nursing, Indiana University, Indianapolis, United States

[5]Go2 for Lung Cancer, Washington, United States

[6]School of Informatics, Indiana University, Indianapolis, United States

[7]Department of Medicine, Internal Medicine, Icahn School of Medicine at Mount Sinai, New York, United States

[8]Hackensack Meridian School of Medicine, Hackensack Meridian Health, Nutley, United States

[9]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, United States

Corresponding Author:
Lisa Carter-Bawa, MSN, MPH, PhD
Cancer Prevention Precision Control Institute, Center for Discovery & Innovation, Hackensack Meridian Health, 123 Metro Blvd, 6th Floor, 6400 Pod, Nutley, NJ, United States

Related Article:

Correction of: https://www.jmir.org/2026/1/e80281

Abstract

(J Med Internet Res 2026;28:e94664) doi:10.2196/94664

In “Leveraging Social Media to Achieve Population-Level Reach of Lung Cancer Screening-Eligible Individuals: A RE-AIM Framework Perspective” [1], the authors noted one error.

The title previously appeared as:

> *Leveraging Social Media to Achieve Population-Level Reach of Lung Cancer Screening–Eligible Individuals: RE-AIM Framework–Based Secondary Analysis of a Randomized Controlled Trial*

The title now reads:

> *Leveraging Social Media to Achieve Population-Level Reach of Lung Cancer Screening-Eligible Individuals: A RE-AIM Framework Perspective*

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Carter-Bawa L, Ostroff JS, Rawl SM, et al. Leveraging social media to achieve population-level reach of lung cancer screening-eligible individuals: a RE-AIM framework perspective. J Med Internet Res 2026;28:e80281. [doi: 10.2196/80281]

XSL•FO
RenderX

Submitted 04.Mar.2026; this is a non–peer-reviewed article;accepted 04.Mar.2026; published 05.Mar.2026.

Please cite as:
Carter-Bawa L, Ostroff JS, Rawl SM, Hirsch EA, Banerjee SC, Ciupek A, Comer RS, Kale M, Leopold KT, Monahan PO, Slaven Jr JE, Valenzona F, Wiener RS, Vielma AG
Correction: Leveraging Social Media to Achieve Population-Level Reach of Lung Cancer Screening-Eligible Individuals: A RE-AIM Framework Perspective
J Med Internet Res 2026;28:e94664
URL: https://www.jmir.org/2026/1/e94664
doi:10.2196/94664

© Lisa Carter-Bawa, Jamie S Ostroff, Susan M Rawl, Erin A Hirsch, Smita C Banerjee, Andrew Ciupek, Robert Skipworth Comer, Minal Kale, Katherine T Leopold, Patrick O Monahan, James E Slaven Jr, Francis Valenzona, Renda Soylemez Wiener, Ana Guadalupe Vielma. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Hybrid Care Modifications in the Delivery of Nonpandemic Care During the COVID-19 Pandemic: Scoping Review

Natalia Sanchez Villalobos[1], MD; Tessa van Loenen[1], PhD; Lem Ngongalah[2], PhD; Máire A Connolly[2], MD, PhD; Mart L Stein[3], PhD; Chantal P Rovers[4], MD, PhD; Aura Timen[1], MD, PhD

[1]Department of Primary and Community Care, Radboud University Medical Center, Geert Grooteplein Noord 21, Nijmegen, The Netherlands
[2]School of Health Sciences, University of Galway, Galway, Ireland
[3]Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, The Netherlands
[4]Department of Internal Medicine, Radboud University Medical Center, Nijmegen, The Netherlands

Corresponding Author:
Natalia Sanchez Villalobos, MD
Department of Primary and Community Care, Radboud University Medical Center, Geert Grooteplein Noord 21, Nijmegen, The Netherlands

Related Article:

Correction of: https://www.jmir.org/2026/1/e84756/

(J Med Internet Res 2026;28:e101728) doi:10.2196/101728

In "Hybrid Care Modifications in the Delivery of Nonpandemic Care During the COVID-19 Pandemic: Scoping Review" [1], the authors noted one error.

The funding information has been added to the end of the article as follows:

> *This study was funded by the European Union. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. The funder had no involvement in the study design, data collection, analysis, interpretation, or writing of the manuscript.*

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Sanchez Villalobos N, van Loenen T, Ngongalah L, et al. Hybrid care modifications in the delivery of nonpandemic care during the COVID-19 pandemic: scoping review. J Med Internet Res 2026 Apr 30;28:e84756. [doi: 10.2196/84756] [Medline: 42060541]

Submitted 19.May.2026; this is a non–peer-reviewed article;accepted 19.May.2026; published 04.Jun.2026.

Please cite as:
Sanchez Villalobos N, van Loenen T, Ngongalah L, Connolly MA, Stein ML, Rovers CP, Timen A
Correction: Hybrid Care Modifications in the Delivery of Nonpandemic Care During the COVID-19 Pandemic: Scoping Review
J Med Internet Res 2026;28:e101728
URL: https://www.jmir.org/2026/1/e101728
doi:10.2196/101728

© Natalia Sanchez Villalobos, Tessa van Loenen, Lem Ngongalah, Máire A Connolly, Mart L Stein, Chantal P Rovers, Aura Timen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 4.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the

XSL•FO
RenderX

Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Integrating Text and Image Analysis: Exploring GPT-4V's Capabilities in Advanced Radiological Applications Across Subspecialties

Felix Busch[1], MD; Tianyu Han[2], PhD; Marcus R Makowski[3], MD, PhD; Daniel Truhn[2], MSc, MD; Keno K Bressem[4*], MD; Lisa Adams[3*], MD

[1]Department of Neuroradiology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt Universität zu Berlin, Berlin, Germany

[2]Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Aachen, Germany

[3]Department of Diagnostic and Interventional Radiology, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

[4]Institute for Radiology and Nuclear Medicine, German Heart Center Munich, Technical University of Munich, Munich, Germany

[*]these authors contributed equally

Corresponding Author:
Felix Busch, MD
Department of Neuroradiology
Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt Universität zu Berlin
Charitépl. 1
Berlin, 10117
Germany
Phone: 49 3045050
Email: felix.busch@charite.de

Related Article:

Correction of: https://www.jmir.org/2024/1/e54948

(J Med Internet Res 2026;28:e91415) doi:10.2196/91415

In "Integrating Text and Image Analysis: Exploring GPT-4V's Capabilities in Advanced Radiological Applications Across Subspecialties" [1], the authors made two corrections.

The following text was added to the caption of Figure 1:

> *Reproduced with permission from the Radiological Society of North America. Link to the displayed case: https://cases.rsna.org/take-quiz/07c4b917-80fb-43c0-8b3b-59a0d8ceb203 (accessed 14th January 2026).*

In the Data Availability section, the following sentence was revised:

> *The data set was utilized under the terms and conditions provided by the RSNA, which permits the use of these cases for educational and research purposes.*

This section now reads:

> *The cases from the RSNA Case Collection were reproduced with permission from the Radiological Society of North America. These cases were not used for model training, nor were they retained by any tools or systems employed in this study.*

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Busch F, Han T, Makowski M, Truhn D, Bressem K, Adams L. Integrating text and image analysisxploring GPT-4V's capabilities in advanced radiological applications across subspecialties. J Med Internet Res 2024;26:54948. [doi: 10.2196/54948]

XSL•FO
RenderX

Submitted 14.Jan.2026; this is a non–peer-reviewed article;accepted 27.Jan.2026; published 11.Feb.2026.

Please cite as:
Busch F, Han T, Makowski MR, Truhn D, Bressem KK, Adams L
Correction: Integrating Text and Image Analysis: Exploring GPT-4V's Capabilities in Advanced Radiological Applications Across Subspecialties
J Med Internet Res 2026;28:e91415
URL: https://www.jmir.org/2026/1/e91415
doi:10.2196/91415
PMID:

©Felix Busch, Tianyu Han, Marcus R Makowski, Daniel Truhn, Keno K Bressem, Lisa Adams. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Call for Decision Support for Electrocardiographic Alarm Administration Among Neonatal Intensive Care Unit Staff: Multicenter, Cross-Sectional Survey

Xiaoli Tang[1,2,3,4*], MSN; Xiaochen Yang[2,4*], BSN; Jiajun Yuan[1,3*], MPH; Jie Yang[5,6], MD; Qian Jin[2], BSN; Hanting Zhang[2], BSN; Liebin Zhao[1], DM, MM; Weiwei Guo[1,2,6,7], MM

[1]Shanghai Engineering Research Center of Intelligence Pediatrics, Shanghai Children's Medical Center, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

[2]Neonatology Department, Shanghai Children's Medical Center, National Children's Medical Center, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

[3]Child Health Advocacy Institute, China Hospital Development Institute, Shanghai, China

[4]Shanghai Jiao Tong University School of Nursing, Shanghai, China

[5]Neonatology Department, Nanfang Hospital, Southern Medical University, Guangzhou, China

[6]Neonatal Respiratory Group of Chinese Physicians Association, Guangzhou, China

[7]Administrative Office, Shanghai Children's Medical Center, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Weiwei Guo, MM
Shanghai Engineering Research Center of Intelligence Pediatrics, Shanghai Children's Medical Center, School of Medicine, Shanghai Jiao Tong University
1678 Dongfang Road, Pudong New Area
Shanghai, 200127
China
Phone: 86 18930830578
Email: guoweiwei@scmc.com.cn

Related Article:

Correction of: https://www.jmir.org/2024/1/e60944/

(J Med Internet Res 2026;28:e101652) doi:10.2196/101652

In "Call for Decision Support for Electrocardiographic Alarm Administration Among Neonatal Intensive Care Unit Staff: Multicenter, Cross-Sectional Survey" [1], the authors noted one error in the Ethical Considerations section. The ethics approval number was mistakenly stated as "SCMCIRB-K2024014-1." The correct ethics approval number for the study reported in this article is "SCMCIRB-YJ2022001," which was granted by the Ethics Committee of Shanghai Children's Medical Center on November 22, 2022. This ethics approval was obtained for a project comprising a survey of neonatal intensive care unit electrocardiographic alarm status and alarm management, as well as the development of an intelligent monitoring device for high-risk neonates.

In the Ethical Considerations section, the following sentence was revised:

> *Prior to the commencement of the study, ethics approval was obtained from the ethics committee of Shanghai Children's Medical Center (approval SCMCIRB-K2024014-1).*

This sentence now reads:

> *Prior to the commencement of the study, ethics approval was obtained from the ethics committee of Shanghai Children's Medical Center (approval SCMCIRB-YJ2022001).*

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

XSL•FO
RenderX

1. Tang X, Yang X, Yuan J, Yang J, Jin Q, Zhang H, et al. Call for decision support for electrocardiographic alarm administration among neonatal intensive care unit staff: multicenter, cross-sectional survey. J Med Internet Res 2024 Dec 20;26:e60944 [FREE Full text] [doi: 10.2196/60944] [Medline: 39705688]

Submitted 18.May.2026; this is a non–peer-reviewed article;accepted 25.Jun.2026; published 29.Jun.2026.

Please cite as:
Tang X, Yang X, Yuan J, Yang J, Jin Q, Zhang H, Zhao L, Guo W
Correction: Call for Decision Support for Electrocardiographic Alarm Administration Among Neonatal Intensive Care Unit Staff: Multicenter, Cross-Sectional Survey
J Med Internet Res 2026;28:e101652
URL: https://www.jmir.org/2026/1/e101652
doi:10.2196/101652
PMID:

©Xiaoli Tang, Xiaochen Yang, Jiajun Yuan, Jie Yang, Qian Jin, Hanting Zhang, Liebin Zhao, Weiwei Guo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Multicriteria Optimization of Language Models for Heart Failure With Preserved Ejection Fraction Symptom Detection in Spanish Electronic Health Records: Comparative Modeling Study

Jacinto Mata[1], PhD; Victoria Pachón[1], PhD; Ana Manovel[2], MD; Manuel J Maña[1], PhD; Manuel de la Villa[1], MSc

[1]I²C Research Group, Universidad de Huelva, Huelva, Spain

[2]Cardiology Department, Juan Ramón Jiménez University Hospital, Multidisciplinary Amyloidosis Unit Huelva, Hospital Juan Ramón Jiménez, Huelva, Spain

Corresponding Author:
Jacinto Mata, PhD
I²C Research Group, Universidad de Huelva, Huelva, Spain

Related Article:

Correction of: https://www.jmir.org/2025/1/e76433

(J Med Internet Res 2026;28:e97621) doi:10.2196/97621

In "Multicriteria Optimization of Language Models for Heart Failure With Preserved Ejection Fraction Symptom Detection in Spanish Electronic Health Records: Comparative Modeling Study" [1], the authors made one correction.

In the Acknowledgments section, the following text was revised:

> *This project was funded by the Institute of Health Carlos III, Ministry of Science, Innovation and Universities, Spanish Government (grant number PI20/01485).*

The text now reads:

> *This study has been funded by the Instituto de Salud Carlos III through the project PI20/01485 (cofunded by the European Regional Development Fund/European Social Fund "A way to make Europe"/"Investing in your future").*

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Mata J, Pachón V, Manovel A, Maña MJ, de la Villa M. Multicriteria optimization of language models for heart failure with preserved ejection fraction symptom detection in Spanish electronic health records: comparative modeling study. J Med Internet Res 2025 Jul 17;27:e76433. [doi: 10.2196/76433] [Medline: 40674251]

Submitted 15.Apr.2026; this is a non–peer-reviewed article;accepted 20.Apr.2026; published 28.May.2026.

Please cite as:
Mata J, Pachón V, Manovel A, Maña MJ, de la Villa M
Correction: Multicriteria Optimization of Language Models for Heart Failure With Preserved Ejection Fraction Symptom Detection in Spanish Electronic Health Records: Comparative Modeling Study
J Med Internet Res 2026;28:e97621
URL: https://www.jmir.org/2026/1/e97621
doi:10.2196/97621

© Jacinto Mata, Victoria Pachón, Ana Manovel, Manuel J Maña, Manuel de la Villa. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN

XSL•FO
RenderX

1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Enhancing LGBTQ+ Inclusivity in an AI-Powered Sexual Health Chatbot: User-Centered Design Approach Through a Nonprofit and Academic Partnership

William Wibowo Liem[1], MSW; Elizabeth Casline[1], PhD; Julianna Lorenzo[1], BS; Jacob D Gordon[2], MSW, PhD; Andres Alvarado Avila[1], MPH; Attia Taylor[3], BA; Nicole Levitz[3], MPH; Michael C O'Keefe[3], MA; Karen Shum[3], BA; Kathryn Macapagal[1,4], PhD

[1]Impact Institute, Northwestern University, Chicago, IL, United States

[2]College of Nursing, University of Cincinnati, Cincinnati, OH, United States

[3]Planned Parenthood Federation of America, New York, NY, United States

[4]Department of Medical Social Sciences, Impact Institute, Northwestern University, 625 North Michigan Avenue, Floor 14, Chicago, IL, United States

Corresponding Author:
Kathryn Macapagal, PhD
Impact Institute, Northwestern University, Chicago, IL, United States

Related Article:

Correction of: https://www.jmir.org/2026/1/e78621

(J Med Internet Res 2026;28:e96639) doi:10.2196/96639

In "Enhancing LGBTQ+ Inclusivity in an AI-Powered Sexual Health Chatbot: User-Centered Design Approach Through a Nonprofit and Academic Partnership" [1], the authors made one addition.

Author KS was inadvertently omitted from the authorship list. They have been added as the ninth author as follows:

> *Karen Shum, BA[3]*
>
> *ORCID: 0009-0004-0094-4818*
>
> *3. Planned Parenthood Federation of America, New York City, NY, United States*

In the Authors' Contributions section, the following was added:

> *KS reviewed the manuscript and provided approval of the final work.*

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Liem WW, Casline E, Lorenzo J, et al. Enhancing LGBTQ+ inclusivity in an AI-powered sexual health chatbot: user-centered design approach through a nonprofit and academic partnership. J Med Internet Res 2026 Feb 19;28:e78621. [doi: 10.2196/78621] [Medline: 41712769]

Submitted 30.Mar.2026; this is a non–peer-reviewed article;accepted 31.Mar.2026; published 04.May.2026.

Please cite as:
Liem WW, Casline E, Lorenzo J, Gordon JD, Avila AA, Taylor A, Levitz N, O'Keefe MC, Shum K, Macapagal K
Correction: Enhancing LGBTQ+ Inclusivity in an AI-Powered Sexual Health Chatbot: User-Centered Design Approach Through a Nonprofit and Academic Partnership
J Med Internet Res 2026;28:e96639
URL: https://www.jmir.org/2026/1/e96639
doi:10.2196/96639

© William Wibowo Liem, Elizabeth Casline, Julianna Lorenzo, Jacob D Gordon, Andres Alvarado Avila, Attia Taylor, Nicole Levitz, Michael C O'Keefe, Karen Shum, Kathryn Macapagal. Originally published in the Journal of Medical Internet Research

XSL•FO
RenderX

(https://www.jmir.org), 4.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Securing Federated Learning With Blockchain in the Medical Field: Systematic Literature Review

Xudong Wang[1*], MD; Yi Xie[2*], MD; Xiaoliang Chen[1,3*], MD; Jiaming Yang[1], MD; Ruiyuan Li[1], MD; Weihang Gao[1], MD; Zineng Yan[1], MD; Hong Zhou[1], MD; Zhewei Ye[1], Prof Dr Med

[1]Department of Orthopedics, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, 1277 Jiefang Avenue, Wuhan, Hubei Province, China

[2]Renmin Hospital of Wuhan University, Wuhan, Hubei Province, China

[3]Department of Orthopedics, People's Hospital of Ningxia Hui Autonomous Region, Ningxia Medical University, Yinchuan, China

[*]these authors contributed equally

Corresponding Author:
Zhewei Ye, Prof Dr Med
Department of Orthopedics, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, 1277 Jiefang Avenue, Wuhan, Hubei Province, China

Related Article:

Correction of: https://www.jmir.org/2026/1/e79052

(J Med Internet Res 2026;28:e95788) doi:10.2196/95788

In "Securing Federated Learning With Blockchain in the Medical Field: Systematic Literature Review" [1], the authors noted two errors.

The first affiliation has been changed from the following:

> *Wuhan Union Hospital, Wuhan, Hubei Province, China*

The affiliation now reads:

> *Department of Orthopedics, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei Province, China.*

The third affiliation has been changed from the following:

> *Ningxia Hui Autonomous Region Peoples Hospital, Yinchuan, Ningxia, China.*

The affiliation now reads:

> *Department of Orthopedics, People's Hospital of Ningxia Hui Autonomous Region, Ningxia Medical University, Yinchuan, China.*

The corrections will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because these were made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Wang X, Xie Y, Chen X, et al. Securing federated learning with blockchain in the medical field: systematic literature review. J Med Internet Res 2026 Feb 19;28:e79052. [doi: 10.2196/79052] [Medline: 41712960]

Submitted 20.Mar.2026; this is a non–peer-reviewed article;accepted 26.Mar.2026; published 15.Apr.2026.

Please cite as:
Wang X, Xie Y, Chen X, Yang J, Li R, Gao W, Yan Z, Zhou H, Ye Z
Correction: Securing Federated Learning With Blockchain in the Medical Field: Systematic Literature Review
J Med Internet Res 2026;28:e95788
URL: https://www.jmir.org/2026/1/e95788
doi:10.2196/95788

© Xudong Wang, Yi Xie, Xiaoliang Chen, Jiaming Yang, Ruiyuan Li, Weihang Gao, Zineng Yan, Hong Zhou, Zhewei Ye. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/),

XSL•FO
RenderX

which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Evidence for Digital Health Tools Designed to Support the Triage of Musculoskeletal Conditions in Primary, Urgent, and Emergency Care Settings: Scoping Review

Linda K Truong[1,2], PT, PhD; James G Wrightson[3], PhD; Raphaël Vincent[4,5,6], PT, MSc; Eunice Lui[1], MPH; Jamon L Couch[7,8], PT; Ellen Wang[2,8], MSc; Cobie Starcevich[9,10], PT; Dean Giustini[11], PhD; Alex Haagaard[12], MA; Elena Lopatina[13,14], MD, PhD; Niels van Berkel[15], PhD; Michael Skovdal Rathleff[16,17], PT, PhD; Clare L Ardern[1,2,7], PT, PhD

[1]Centre for Aging SMART, University of British Columbia, Vancouver, BC, Canada
[10]Physiotherapy Department, Rockingham General Hospital, South Metropolitan Health Service, Perth, Australia
[11]Biomedical Branch Library, University of British Columbia, Vancouver, BC, Canada
[12]Pain BC, Vancouver, BC, Canada
[13]University of Calgary, Calgary, AB, Canada
[14]Primary Care Alberta, Calgary, AB, Canada
[15]Department of Computer Science, Aalborg University, Aalborg, Denmark
[16]Department of Health Science and Technology, Faculty of Medicine, Aalborg University, Aalborg, Denmark
[17]Center for General Practice at Aalborg University, Aalborg, Denmark
[2]Department of Physical Therapy, Faculty of Medicine, University of British Columbia, 13737 96 Avenue, Surrey, BC, Canada
[3]Department of Family Practice, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada
[4]School of Rehabilitation, Faculty of Medicine, Université de Montréal, Montreal, QC, Canada
[5]Hôpital Maisonneuve-Rosemont Research Center, Université de Montréal Affiliated Research Center, Montreal, QC, Canada
[6]Centre for Interdisciplinary Research in Rehabilitation of Greater Montreal, Institut Universitaire sur la Réadaptation en Déficience Physique de Montréal, Montreal, QC, Canada
[7]La Trobe Sport and Exercise Medicine Research Centre, La Trobe University, Melbourne, Australia
[8]Arthritis Research Canada, Vancouver, BC, Canada
[9]School of Allied Health, Faculty of Health Sciences, Curtin University, Perth, Australia

Corresponding Author:
Linda K Truong, PT, PhD
Centre for Aging SMART, University of British Columbia, Vancouver, BC, Canada

Related Article:

Correction of: https://www.jmir.org/2026/1/e81578

(J Med Internet Res 2026;28:e92722) doi:10.2196/92722

In "Evidence for Digital Health Tools Designed to Support the Triage of Musculoskeletal Conditions in Primary, Urgent, and Emergency Care Settings: Scoping Review" [1], the authors made two corrections.

In the abstract, the list of databases searched has been amended to remove PsycINFO and Cochrane Library, and these were replaced with Central (OVID) and Compendex. This has also been further specified in the Methods section.

In the abstract, the following sentence was revised:

> *Systematic searches in MEDLINE (OVID), CINAHL (EBSCO), PsycINFO (EBSCO), Embase (OVID), Cochrane Library, Web of Science, OpenGrey, Google Scholar, arXiv, medRxiv, and an extensive gray literature search were conducted with a librarian scientist from inception to September 18, 2025.*

This sentence now reads:

> *Systematic searches in MEDLINE (OVID), Embase (OVID), CENTRAL (OVID), CINAHL (EBSCO), Compendex, Web of Science, OpenGrey, Google Scholar, arXiv, medRxiv, and an extensive gray literature search were conducted with a librarian scientist from inception to September 18, 2025.*

In the Methods section, the following sentence was revised:

> *An electronic search was conducted in 6 databases (MEDLINE [OVID], CINAHL [EBSCO], PsycINFO [EBSCO], Embase [OVID], the Cochrane Library, and Web of Science) and 4 gray literature sites (OpenGrey, GoogleScholar, arXiv, and medRxiv) with the aid of a biomedical librarian and information specialist.*

The sentence now reads:

XSL•FO
RenderX

An electronic search was conducted in 6 databases (MEDLINE [OVID], Embase [OVID], CENTRAL [OVID], CINAHL [EBSCO], Compendex, and Web of Science) and 4 gray literature sites (OpenGrey, GoogleScholar, arXiv, and medRxiv) with the aid of a biomedical librarian and information specialist.

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Truong LK, Wrightson JG, Vincent R, et al. Evidence for digital health tools designed to support the triage of musculoskeletal conditions in primary, urgent, and emergency care settings: scoping review. J Med Internet Res 2026 Jan 14;28:e81578. [doi: 10.2196/81578] [Medline: 41534041]

Submitted 02.Feb.2026; this is a non–peer-reviewed article;accepted 11.Feb.2026; published 26.Feb.2026.

Please cite as:
Truong LK, Wrightson JG, Vincent R, Lui E, Couch JL, Wang E, Starcevich C, Giustini D, Haagaard A, Lopatina E, van Berkel N, Rathleff MS, Ardern CL
Correction: Evidence for Digital Health Tools Designed to Support the Triage of Musculoskeletal Conditions in Primary, Urgent, and Emergency Care Settings: Scoping Review
J Med Internet Res 2026;28:e92722
URL: https://www.jmir.org/2026/1/e92722
doi:10.2196/92722

© Linda K Truong, James G Wrightson, Raphaël Vincent, Eunice Lui, Jamon L Couch, Ellen Wang, Cobie Starcevich, Dean Giustini, Alex Haagaard, Elena Lopatina, Niels van Berkel, Michael Skovdal Rathleff, Clare L Ardern. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Acceptability of Health Information Technology by Health Care Professionals: Where We Are Now and How We Can Fill the Gap

Corinne Isnard-Bagnis[1,2], MD, PhD; Stéphane Mouchabac[3,4], MD; Riadh Lebib[5], PhD; Hervé Bismut[6], PhD; Pierre Alexis Geoffroy[7,8,9], MD, PhD

[1]Nephrology Department, Pitié-Salpêtrière Hospital, AP-HP, Sorbonne Université, 47-83 boulevard de l'Hôpital, Paris, France
[2]Academic Digital Medical Hub, AP-HP, Paris, France
[3]Department of Psychiatry, Saint-Antoine Hospital, AP-HP, Sorbonne Université, Paris, France
[4]Infrastructure for Clinical Research in Neurosciences (iCRIN), Paris Brain Institute, Paris, France
[5]Humans Matter, Paris, France
[6]Geminicis, Paris, France
[7]Département de psychiatrie et d'addictologie, AP-HP, GHU Paris Nord, DMU Neurosciences, Hôpital Bichat-Claude-Bernard, Paris, France
[8]Centre ChronoS, GHU Paris psychiatrie & neurosciences, Paris, France
[9]Université Paris Cité, Inserm, NeuroDiderot, Paris, France

Corresponding Author:
Corinne Isnard-Bagnis, MD, PhD
Nephrology Department, Pitié-Salpêtrière Hospital, AP-HP, Sorbonne Université, 47-83 boulevard de l'Hôpital, Paris, France

Related Article:

Correction of: https://www.jmir.org/2025/1/e72184

(J Med Internet Res 2026;28:e89383) doi:10.2196/89383

In "Acceptability of Health Information Technology by Health Care Professionals: Where We Are Now and How We Can Fill the Gap" [1], the authors noted incomplete affiliations for authors CIB, SM, and PAG.

The previous list of affiliations was as follows:

> *Corinne Isnard Bagnis, MD, PhD[1], Stéphane Mouchabac, MD[2], Riadh Lebib, PhD[3], Hervé Bismut, PhD[4], Pierre A Geoffroy, MD, PhD[5]*
>
> *[1]Department of Nephrology, Pitié-Salpêtrière Hospital, APHP Sorbonne Université, Paris, France*
>
> *[2]Department of Psychiatry, Saint-Antoine Hospital, APHP Sorbonne Université, Paris, France*
>
> *[3]Humans Matter, Paris, France*
>
> *[4]Geminicis, Paris, France*
>
> *[5]Department of Psychiatry, Hôpital Bichat-Claude-Bernard, Paris, France*

This has been revised to:

> *Corinne Isnard Bagnis, MD, PhD[1,2], Stéphane Mouchabac, MD[3,4], Riadh Lebib, PhD[5], Hervé Bismut, PhD[6], Pierre A Geoffroy, MD, PhD[7-9]*
>
> *[1]Nephrology Department, Pitié Salpêtrière Hospital, AP-HP, Sorbonne Université, Paris, France*
>
> *[2]Academic Digital Medical Hub, AP-HP, Paris, France*
>
> *[3]Department of Psychiatry, Saint-Antoine Hospital, AP-HP, Sorbonne Université, Paris, France.*
>
> *[4]Infrastructure for Clinical Research in Neurosciences (iCRIN), Paris Brain Institute, Paris, France*
>
> *[5]Humans Matter, Paris, France*
>
> *[6]Geminicis, Paris, France*
>
> *[7]Département de psychiatrie et d'addictologie, AP-HP, GHU Paris Nord, DMU Neurosciences, Hôpital Bichat - Claude-Bernard, Paris, France*
>
> *[8]Centre ChronoS, GHU Paris psychiatrie & neurosciences, Paris, France*
>
> *[9]Université Paris Cité, Inserm, NeuroDiderot, Paris, France*

These corrections will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because these were made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

XSL•FO
RenderX

Reference

1. Isnard-Bagnis C, Mouchabac S, Lebib R, Bismut H, Geoffroy PA. Acceptability of health information technology by health care professionals: where we are now and how we can fill the gap. J Med Internet Res 2025 Dec 2;27:e72184. [doi: 10.2196/72184] [Medline: 41329915]

Submitted 11.Dec.2025; this is a non–peer-reviewed article;accepted 16.Dec.2025; published 06.Jan.2026.

Please cite as:
Isnard-Bagnis C, Mouchabac S, Lebib R, Bismut H, Geoffroy PA
Correction: Acceptability of Health Information Technology by Health Care Professionals: Where We Are Now and How We Can Fill the Gap
J Med Internet Res 2026;28:e89383
URL: https://www.jmir.org/2026/1/e89383
doi:10.2196/89383

© Corinne Isnard-Bagnis, Stéphane Mouchabac, Riadh Lebib, Hervé Bismut, Pierre A Geoffroy. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Combining Artificial Intelligence and Human Support in Mental Health: Digital Intervention With Comparable Effectiveness to Human-Delivered Care

Clare E Palmer[1], PhD; Emily Marshall[1], PGCert, PGDip; Edward Millgate[1], PhD; Graham Warren[1], PhD; Michael Ewbank[1], PhD; Elisa Cooper[1], PhD; Samantha Lawes[1], PhD; Alastair Smith[1], MEng; Chris Hutchins-Joss[1], MSc; Jessica Young[1], MSc; Malika Bouazzaoui[1], MBA, MSc; Morad Margoum[2], MSc; Sandra Healey[2], PGDip; Louise Marshall[1], PhD; Shaun Mehew[1], PGCert, PGDip; Ronan Cummins[1], PhD; Valentin Tablan[1], PhD; Ana Catarino[1], PhD; Andrew E Welchman[1], PhD; Andrew D Blackwell[1], PhD

[1]ieso Digital Health Ltd, Cambridge, United Kingdom

[2]Dorset HealthCare University NHS Foundation, Poole, United Kingdom

Corresponding Author:
Clare E Palmer, PhD
ieso Digital Health Ltd
The Jeffreys Building
Cowley Road
Cambridge, CB4 0DS
United Kingdom
Phone: 44 0800 074 5560
Email: c.palmer@iesohealth.com

Related Article:

Correction of: https://www.jmir.org/2025/1/e69351

(J Med Internet Res 2026;28:e88640) doi:10.2196/88640

In "Combining Artificial Intelligence and Human Support in Mental Health: Digital Intervention With Comparable Effectiveness to Human-Delivered Care" [1], the authors noted one error.

Author CH-J was previously listed with no academic qualifications. His qualification has been revised to "MSc."

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Palmer CE, Marshall E, Millgate E, Warren G, Ewbank M, Cooper E, et al. Combining Artificial Intelligence and Human Support in Mental Health: Digital Intervention With Comparable Effectiveness to Human-Delivered Care. J Med Internet Res 2025 May 13;27:e69351 [FREE Full text] [doi: 10.2196/69351] [Medline: 40152000]

XSL•FO
RenderX

Submitted 28.Nov.2025; this is a non–peer-reviewed article;accepted 05.Dec.2025; published 06.Jan.2026.

Please cite as:
Palmer CE, Marshall E, Millgate E, Warren G, Ewbank M, Cooper E, Lawes S, Smith A, Hutchins-Joss C, Young J, Bouazzaoui M, Margoum M, Healey S, Marshall L, Mehew S, Cummins R, Tablan V, Catarino A, Welchman AE, Blackwell AD
Correction: Combining Artificial Intelligence and Human Support in Mental Health: Digital Intervention With Comparable Effectiveness to Human-Delivered Care
J Med Internet Res 2026;28:e88640
URL: https://www.jmir.org/2026/1/e88640
doi:10.2196/88640
PMID:

©Clare E Palmer, Emily Marshall, Edward Millgate, Graham Warren, Michael Ewbank, Elisa Cooper, Samantha Lawes, Alastair Smith, Chris Hutchins-Joss, Jessica Young, Malika Bouazzaoui, Morad Margoum, Sandra Healey, Louise Marshall, Shaun Mehew, Ronan Cummins, Valentin Tablan, Ana Catarino, Andrew E Welchman, Andrew D Blackwell. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Effectiveness of a Web-Based Medication Education Course on Pregnant Women's Medication Information Literacy and Decision Self-Efficacy: Randomized Controlled Trial

Suya Li[1*], MSN; Hui-Jun Chen[2*], MD; Jie Zhou[1], MSN; Yi-Bei Zhouchen[1], MNS; Rong Wang[3], MSN; Jinyi Guo[1], MNS; Sharon R Redding[4], EdD; Yan-Qiong Ouyang[1], MD

[1]School of Nursing, Wuhan University, Wuhan, China

[2]Department of Obstetrics, Zhongnan Hospital of Wuhan University, Wuhan, China

[3]Renmin Hospital of Wuhan University, Wuhan, China

[4]Project HOPE, Washington, DC, United States

[*]these authors contributed equally

Corresponding Author:
Yan-Qiong Ouyang, MD
School of Nursing
Wuhan University
115 Donghu Road
Wuchang District
Wuhan, 430071
China
Phone: 86 02768758747
Email: ouyangyq@whu.edu.cn

Related Article:

Correction of: https://www.jmir.org/2025/1/e54148/8

(J Med Internet Res 2026;28:e91835) doi:10.2196/91835

In "Effectiveness of a Web-Based Medication Education Course on Pregnant Women's Medication Information Literacy and Decision Self-Efficacy: Randomized Controlled Trial" [1], the authors noted an error in the affiliations.

The affiliation of authors SL, JZ, YBZ, JG, and YQO was previously shown as:

[1]*Wuhan University*

This affiliation has been revised to the following:

[1]*School of Nursing, Wuhan University*

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Li S, Chen HJ, Zhou J, Zhouchen YB, Wang R, Guo J, et al. Effectiveness of a Web-Based Medication Education Course on Pregnant Women's Medication Information Literacy and Decision Self-Efficacy: Randomized Controlled Trial. J Med Internet Res 2025 Jan 22;27:e54148 [FREE Full text] [doi: 10.2196/54148] [Medline: 39841986]

XSL•FO
RenderX

Submitted 20.Jan.2026; this is a non–peer-reviewed article;accepted 21.Jan.2026; published 26.Jan.2026.

Please cite as:
Li S, Chen HJ, Zhou J, Zhouchen YB, Wang R, Guo J, Redding SR, Ouyang YQ
Correction: Effectiveness of a Web-Based Medication Education Course on Pregnant Women's Medication Information Literacy and Decision Self-Efficacy: Randomized Controlled Trial
J Med Internet Res 2026;28:e91835
URL: https://www.jmir.org/2026/1/e91835
doi:10.2196/91835
PMID:

©Suya Li, Hui-Jun Chen, Jie Zhou, Yi-Bei Zhouchen, Rong Wang, Jinyi Guo, Sharon R Redding, Yan-Qiong Ouyang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Culturally Adapted Guided Internet-Based Cognitive Behavioral Therapy for Hong Kong People With Depressive Symptoms: Randomized Controlled Trial

Jia-Yan Pan[1], PhD; Jonas Rafi[2], PhD

[1]Department of Social Work, Hong Kong Baptist University, Hong Kong, China (Hong Kong)

[2]Department of Psychology, Stockholm University, Stockholm, Sweden

Corresponding Author:
Jia-Yan Pan, PhD
Department of Social Work
Hong Kong Baptist University
Room AAB1026, 10/F, Academic and Administration Building
15 Baptist University Road, Baptist University Road Campus, Kowloon Tong, KLN
Hong Kong
China (Hong Kong)
Phone: 852 34116415
Email: jiayan@hkbu.edu.hk

Related Article:

Correction of: https://www.jmir.org/2025/1/e64303

(J Med Internet Res 2026;28:e88495) doi:10.2196/88495

In "Culturally Adapted Guided Internet-Based Cognitive Behavioral Therapy for Hong Kong People With Depressive Symptoms: Randomized Controlled Trial"[1], the authors noted some errors.

One sentence has been changed. In the "Measurements" section, the third sentence in the third paragraph originally appeared as follows:

> *Item scores were summed for the positive and negative subscales, with higher total scores indicating more positive and negative automatic thoughts, respectively.*

This sentence now reads:

> *Item scores were averaged for the positive and negative subscales, with higher total scores indicating more positive and negative automatic thoughts, respectively.*

The originally published Tables 2 and 4 have been revised due to typos. The new Table 2 and Table 4 that are replacing the original published versions are shown below (with changes marked in italics).

XSL•FO
RenderX

Table 2. Comparison of treatment effects between active and WLC[a] groups at post-treatment (intent-to-treat analysis).

Outcome	Unadjusted mean (SD)			Adjusted difference (95% CI)				Between group effect size *d* (95% CI)		Within group effect size (95% CI)		
	Web	App	WL	Web vs WL	*P* value	App vs WL	*P* value	Web vs WL	App vs WL	Web	App	WLC
BDI-II[b]	9.40 (8.90)	9.43 (8.45)	18.02 (9.02)	–8.24 (–10.52 to –5.96)	<.001	–9.38 (–11.65 to –7.10)	<.001	1.07 (0.81 to 1.34)	1.15 (0.88 to 1.43)	1.48 (1.18 to 1.78)	1.64 (1.33 to 1.94)	0.45 (0.2 to 0.71)
PHQ-9[c]	5.59 (4.43)	5.24 (3.65)	8.99 (4.58)	–3.07 (–4.46 to –1.67)	<.001	–4.50 (–5.89 to –3.12)	<.001	0.78 (0.52 to 1.04)	0.95 (0.63 to 1.27)	1.11 (0.81 to 1.41)	1.69 (1.36 to 2.02)	0.46 (0.19 to 0.73)
GHQ-12[d]	1.60 (3.06)	1.53 (2.77)	5.35 (3.85)	–3.65 (–4.75 to –2.54)	<.001	–4.49 (–5.6 to –3.38)	<.001	1.05 (0.77 to 1.32)	1.11 (0.83 to 1.39)	1.64 (1.34 to 1.95)	2.21 (1.87 to 2.55)	0.64 (0.38 to 0.9)
BAI[e]	8.89 (7.89)	8.49 (6.69)	11.09 (7.49)	–2.47 (–4.33 to –0.62)	.009	–3.32 (–5.17 to –1.47)	<.001	0.37 (0.11 to 0.63)	0.54 (0.29 to 0.8)	0.64 (0.37 to 0.91)	0.88 (0.59 to 1.15)	0.37 (0.11 to 0.62)
Positive automatic thoughts	2.40 (0.71)	2.33 (0.67)	1.98 (0.58)	0.38 (0.22 to 0.54)	<.001	0.40 (0.24 to 0.55)	<.001	–0.63 (–0.9 to –0.37)	–0.61 (–0.9 to –0.32)	–0.88 (–1.16 to –0.6)	–0.93 (–1.22 to –0.65)	–0.30 (–0.55 to 0.04)
Negative automatic thoughts	1.85 (0.78)	1.89 (0.63)	2.29 (0.80)	–0.43 (–0.63 to –0.22)	<.001	–0.48 (–0.68 to –0.27)	<.001	0.63 (0.37 to 0.90)	0.66 (0.35 to 0.97)	0.83 (0.55 to 1.11)	1.00 (0.72 to 1.28)	0.28 (0.02 to 0.53)
Positive emotions	3.26 (0.69)	3.27 (0.62)	2.89 (0.68)	0.35 (0.18 to 0.52)	<.001	0.43 (0.26 to 0.60)	<.001	–0.56 (–0.84 to –0.27)	–0.60 (–0.9 to –0.3)	–0.89 (–1.17 to –0.61)	–1.10 (–1.38 to –0.81)	–0.30 (–0.55 to –0.04)
Negative emotions	2.91 (0.82)	2.92 (0.69)	3.33 (0.74)	–0.45 (–0.64 to –0.25)	<.001	–0.47 (–0.66 to –0.27)	<.001	0.64 (0.37 to 0.91)	0.65 (0.38 to 0.91)	1.06 (0.78 to 1.34)	1.21 (0.91 to 1.5)	0.50 (0.24 to 0.75)

[a]WLC: waitlist control.

[b]BDI-II: Beck Depression Inventory-II.

[c]PHQ-9: 9-item Patient Health Questionnaire.

[d]GHQ-12: 12-item General Health Questionnaire.

[e]BAI: Beck Anxiety Inventory.

XSL•FO
RenderX

Table 4. Comparisons between web-based and app-based iCBT[a] groups on primary and secondary outcome measures at post-treatment (intention-to-treat analysis).

Measure and time	Unadjusted mean (SD)		Adjusted difference (95% CI)	*P* value	Between-group effect size *d* (95% CI)	Within-group effect size (95% CI)	
	Web	App				Web	App
BDI-II[b]							
Baseline	22.24 (8.52)	23.66 (8.77)	N/A[c]	N/A	N/A	N/A	N/A
Post	9.40 (8.90)	9.43 (8.45)	0.21 (–2.28 to 2.70)	.87	–0.01 (–0.3 to 0.28)	1.48 (1.18 to 1.78)	1.64 (1.33 to 1.94)
3 months	9.43 (9.58)	9.15 (8.51)	–0.14 (–3.11 to 2.83)	.93	0.03 (–0.34 to 0.4)	1.45 (1.1 to 1.8)	1.67 (1.32 to 2.02)
6 months	10.15 (9.03)	13.47 (12.24)	–1.37 (–4.46 to 1.72)	.38	–0.37 (–0.76 to 0.02)	1.40 (1.05 to 1.74)	1.04 (0.69 to 1.38)
PHQ-9[d]							
Baseline	10.77 (4.9)	12.07 (4.27)	N/A	N/A	N/A	N/A	N/A
Post	5.59 (4.43)	5.24 (3.65)	0.35 (–1.08 to 1.77)	.63	0.07 (–0.22 to 0.36)	1.11 (0.8 to 1.41)	1.69 (1.36 to 2.02)
3 months	5.41 (4.01)	5.83 (3.57)	–0.27 (–1.94 to 1.41)	.75	–0.1 (–0.47 to 0.28)	1.16 (0.81 to 1.51)	1.55 (1.18 to 1.91)
6 months	5.94 (5.44)	7.65 (5.06)	–0.86 (–2.67 to 0.94)	.35	–0.39 (–0.79 to 0.01)	0.95 (0.59 to 1.31)	0.97 (0.62 to 1.33)
GHQ-12[e]							
Baseline	7.46 (3.86)	8.4 (3.31)	N/A	N/A	N/A	N/A	N/A
Post	1.60 (3.06)	1.53 (2.77)	0.05 (–0.84 to 0.93)	.91	0.02 (–0.27 to 0.31)	1.64 (1.34 to 1.95)	2.21 (1.87 to 2.55)
3 months	1.85 (3.02)	1.69 (2.82)	–0.08 (–1.17 to 1)	.88	0.05 (–0.32 to 0.42)	1.54 (1.19 to 1.89)	2.12 (1.74 to 2.49)
6 months	1.77 (3.20)	3.49 (4.10)	–1.46 (–0.33 to –2.59)	.01	–0.52 (–0.91 to –0.12)	1.54 (1.19 to 1.89)	1.39 (1.03 to 1.74)
BAI[f]							
Baseline	14.53 (9.29)	15.36 (8.8)	N/A	N/A	N/A	N/A	N/A
Post	8.89 (7.89)	8.49 (6.69)	0.53 (–1.25 to 2.31)	.56	0.06 (–0.23 to 0.36)	0.64 (0.37 to 0.91)	0.88 (0.59 to 1.15)
3 months	7.67 (6.42)	8.03 (6.78)	0.59 (–1.56 to 2.73)	.59	–0.05 (–0.42 to 0.32)	0.80 (0.47 to 1.12)	0.89 (0.57 to 1.21)
6 months	7.89 (7.99)	9.88 (7.72)	0.32 (–1.91 to 2.56)	.78	–0.25 (–0.64 to 0.14)	0.74 (0.42 to 1.07)	0.64 (0.31 to 0.98)
Positive automatic thoughts							
Baseline	1.85 (0.55)	1.79 (0.52)	N/A	N/A	N/A	N/A	N/A
Post	2.40 (0.71)	2.33 (0.67)	0.01 (–0.18 to 0.2)	.93	0.1 (–0.19 to 0.39)	–0.88 (–1.16 to –0.6)	–0.93 (–1.21 to –0.65)
3 months	2.49 (0.67)	2.44 (0.84)	–0.1 (–0.32 to 0.12)	.38	0.07 (–0.3 to 0.44)	–1.08 (–1.41 to –0.75)	–1.03 (–1.36 to –0.71)
6 months	2.46 (0.85)	2.15 (0.71)	0.08 (–0.16 to 0.31)	.52	0.47 (0.08 to 0.86)	–0.93 (–1.26 to –0.6)	–0.63 (–0.96 to –0.29)
Negative automatic thoughts							
Baseline	2.55 (0.88)	2.65 (0.85)	N/A	N/A	N/A	N/A	N/A
Post	1.85 (0.78)	1.89 (0.63)	–0.02 (–0.22 to 0.18)	.86	–0.04 (–0.33 to 0.25)	0.83 (0.55 to 1.1)	1 (0.72 to 1.28)
3 months	1.74 (0.68)	1.82 (0.71)	–0.03 (–0.27 to 0.21)	.81	–0.09 (–0.46 to 0.28)	0.97 (0.64 to 1.29)	1.02 (0.7 to 1.34)
6 months	1.75 (0.78)	2.05 (0.93)	–0.15 (–0.4 to 0.1)	.24	–0.38 (–0.77 to 0.01)	0.93 (0.6 to 1.26)	0.68 (0.34 to 1.02)
Positive emotions							
Baseline	2.72 (0.54)	2.65 (0.54)	N/A	N/A	N/A	N/A	N/A
Post	3.26 (0.69)	3.27 (0.62)	–0.03 (–0.22 to 0.16)	.87	–0.04 (–0.33 to 0.25)	–0.89 (–1.17 to –0.61)	–1.1 (–1.38 to –0.81)
3 months	3.26 (0.75)	3.33 (0.81)	–0.1 (–0.33 to 0.13)	.93	–0.12 (–0.49 to 0.25)	–0.89 (–1.21 to –0.56)	–1.07 (–1.39 to –0.74)
6 months	3.24 (0.75)	2.97 (0.72)	0.12 (–0.11 to 0.36)	.38	0.4 (0.01 to 0.79)	–0.85 (–1.18 to –0.52)	–0.54 (–0.88 to –0.21)

XSL•FO
RenderX

Measure and time	Unadjusted mean (SD)		Adjusted difference (95% CI)	*P* value	Between-group effect size *d* (95% CI)	Within-group effect size (95% CI)	
	Web	App				Web	App
Negative emotions							
Baseline	3.73 (0.75)	3.78 (0.73)	N/A	N/A	N/A	N/A	N/A
Post	2.91 (0.82)	2.93 (0.69)	–0.02 (–0.23 to 0.2)	.63	–0.02 (–0.31 to 0.28)	1.06 (0.78 to 1.34)	1.21 (0.91 to 1.5)
3 months	2.78 (0.85)	2.93 (0.73)	–0.07 (–0.32 to 0.18)	.75	–0.19 (–0.56 to 0.18)	1.23 (0.89 to 1.57)	1.17 (0.84 to 1.5)
6 months	2.86 (0.91)	3.17 (0.88)	–0.1 (–0.36 to 0.16)	.35	–0.39 (–0.78 to 0)	1.09 (0.75 to 1.42)	0.8 (0.46 to 1.13)

[a]iCBT: internet-based cognitive behavioral therapy

[b]BDI-II: Beck Depression Inventory-II.

[c]N/A: not applicable.

[d]PHQ-9: 9-item Patient Health Questionnaire.

[e]GHQ-12: 12-item General Health Questionnaire.

[f]BAI: Beck Anxiety Inventory.

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Pan JY, Rafi J. Culturally Adapted Guided Internet-Based Cognitive Behavioral Therapy for Hong Kong People With Depressive Symptoms: Randomized Controlled Trial. J Med Internet Res 2025 Feb 25;27:e64303 [FREE Full text] [doi: 10.2196/64303] [Medline: 39998865]

Submitted 26.Nov.2025; this is a non–peer-reviewed article;accepted 27.Nov.2025; published 03.Feb.2026.

Please cite as:
Pan JY, Rafi J
Correction: Culturally Adapted Guided Internet-Based Cognitive Behavioral Therapy for Hong Kong People With Depressive Symptoms: Randomized Controlled Trial
J Med Internet Res 2026;28:e88495
URL: https://www.jmir.org/2026/1/e88495
doi:10.2196/88495
PMID:

©Jia-Yan Pan, Jonas Rafi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Ingroup Favoritism Surrounding COVID-19 Vaccinations in the Hispanic Communities: Experimental Study

Juwon Hwang[1,2], PhD; Asya Cooley[3], PhD; Skye Cooley[3], PhD; Robert Hinck[4], PhD

[1]Department of Mass Communication, Advertising and Public Relations, College of Communication, Boston University, Boston, MA, United States
[2]Core Faculty, Center on Emerging Infectious Diseases, Boston University, Boston, MA, United States
[3]School of Media and Strategic Communications, Oklahoma State University, Stillwater, United States
[4]Leadership and Innovation Institute, Air University, Montgomery, United States

Corresponding Author:
Juwon Hwang, PhD
Department of Mass Communication, Advertising and Public Relations
College of Communication
Boston University
640 Commonwealth Ave,
Boston, MA, 02215
United States
Phone: 1 617 353 3450
Email: juhwang@bu.edu

Related Article:

Correction of: https://www.jmir.org/2025/1/e71188

(J Med Internet Res 2026;28:e90652) doi:10.2196/90652

In “Ingroup Favoritism Surrounding COVID-19 Vaccinations in the Hispanic Communities: Experimental Study” [1], the authors identified errors related to references and corresponding in-text citations.

Reference 43 was invalid and has been removed from the reference list, and the corresponding in-text citation has been removed. The affected statement is now supported by the following reference:

> *Daggett A, Abdollahi S, Hashemzadeh M. The effect of language concordance on health care relationship trust score. Cureus. May 26, 2023;15(5):e39530 [doi: 10.7759/cureus.39530] [Medline: 37366455]*

Reference 53 was invalid and has been removed from the reference list, and the corresponding in-text citation has been removed. The affected statement is now supported by the following reference:

> *Chandler J, Rosenzweig C, Moss AJ, Robinson J, Litman L. Online panels in social science research: expanding sampling methods beyond Mechanical Turk. Behav Res Methods. Oct 2019;51(5):2022-2038. [doi: 10.3758/s13428-019-01273-7] [Medline: 31512174]*

Reference 54 was invalid and has been removed from the reference list, and the corresponding in-text citation has been removed.

References 55 and 56 were invalid entries of the same reference and have been removed from the reference list, and the corresponding in-text citations have been removed. The affected statement is now supported by the following reference:

> *Khan H, Lee R, Lockshin L. Do ethnic cues improve advertising effectiveness for ethnic consumers? Australasian Marketing J. Aug 2015;23(3):218-226. [doi: 10.1016/j.ausmj.2015.07.001]*

Reference 44 was incorrectly attributed and has been corrected to the following reference:

> *Haslam C, Jetten J, Cruwys T, Dingle G, Haslam SA. The New Psychology of Health: Unlocking the Social Cure. Routledge; 2018. [doi: 10.4324/9781315648569]*

Reference 58 was incorrectly attributed and has been corrected to the following reference:

> *Ipeirotis P. Demographics of Mechanical Turk. New York University; Mar 2010. URL: https://ipeirotis.org/wp-content/uploads/2012/02/CeDER-10-01.pdf [Accessed 2026-02-20]*

The correction will appear in the online version of the paper on the JMIR Publications website together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

XSL•FO
RenderX

Reference

1. Hwang J, Cooley A, Cooley S, Hinck R. Ingroup Favoritism Surrounding COVID-19 Vaccinations in the Hispanic Communities: Experimental Study. J Med Internet Res 2025 May 27;27:e71188 [FREE Full text] [doi: 10.2196/71188] [Medline: 40424033]

Submitted 31.Dec.2025; this is a non–peer-reviewed article;accepted 13.Feb.2026; published 06.Mar.2026.

Please cite as:
Hwang J, Cooley A, Cooley S, Hinck R
Correction: Ingroup Favoritism Surrounding COVID-19 Vaccinations in the Hispanic Communities: Experimental Study
J Med Internet Res 2026;28:e90652
URL: https://www.jmir.org/2026/1/e90652
doi:10.2196/90652
PMID:

©Juwon Hwang, Asya Cooley, Skye Cooley, Robert Hinck. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: A Supervised Explainable Machine Learning Model for Perioperative Neurocognitive Disorder in Liver-Transplantation Patients and External Validation on the Medical Information Mart for Intensive Care IV Database: Retrospective Study

Zhendong Ding[1], MD; Linan Zhang[1], MD; Yihan Zhang[1], MD; Jing Yang[1], MD, PhD; Yuheng Luo[2], MD; Mian Ge[1], MD, PhD; Weifeng Yao[1], MD, PhD; Ziqing Hei[1], MD, PhD; Chaojin Chen[1], MD, PhD

[1]Department of Anesthesiology, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, Guangdong, China

[2]Guangzhou AI & Data Cloud Technology Co., LTD, Guangzhou, China

Corresponding Author:
Chaojin Chen, MD, PhD
Department of Anesthesiology
The Third Affiliated Hospital of Sun Yat-sen University
Guangzhou, Guangdong
China
Phone: 86 13430322182
Email: chenchj28@mail.sysu.edu.cn

Related Article:

Correction of: https://www.jmir.org/2025/1/e55046

(J Med Internet Res 2026;28:e88952) doi:10.2196/88952

In "A Supervised Explainable Machine Learning Model for Perioperative Neurocognitive Disorder in Liver-Transplantation Patients and External Validation on the Medical Information Mart for Intensive Care IV Database: Retrospective Study" [1], the authors made three corrections.

The footnotes of Table 1 originally read as follows:

> [a]*PND: perioperative neurocognitive dysfunction.*
>
> [b]*DBD: donation after brain death.*
>
> [c]*DCD: donation after circulatory death.>*
>
> [d]*DBCD: donation after brain death followed by circulatory death.*

The footnotes now read:

> [a]*Data presented in Table 1 are based on the original dataset prior to data imputation.*
>
> [b]*PND: perioperative neurocognitive dysfunction.*
>
> [c]*DBD: donation after brain death.*
>
> [d]*DCD: donation after circulatory death.*
>
> [e]*DBCD: donation after brain death followed by circulatory death.*

Below is the new Figure 4 (presented here as Figure 1), which will replace the originally published figure.

XSL•FO
RenderX

Figure 1. Post–liver transplantation survival associated with perioperative neurocognitive dysfunction. Patients with post–liver transplantation perioperative neurocognitive dysfunction showed a significantly lower survival rate. LT: liver transplantation; PND: perioperative neurocognitive dysfunction.

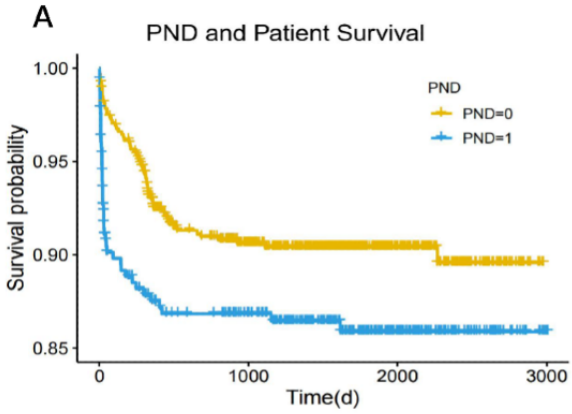

B

Survival duration	Total (N=980)	Non-PND (N=649)	PND (N=331)	*P-value*
30-day*	940(95.9%)	637 (98.2%)	303(91.5%)	**<0.001**
3-month*	929(94.8%)	630(97.1%)	299(90.3%)	**<0.001**
6-month*	919(93.8%)	624(96.1%)	295(89.1%)	**<0.001**
12-month*	892(91.0%)	601(92.6%)	291(87.9%)	**0.021**
3-years	879(89.7%)	590(90.9%)	289(87.3%)	0.101

A paragraph in the *Effect of PND on Patients' Outcomes and Prognosis* section has been revised. The text originally read as follows:

> *Further survival analysis (Figure 4)was conducted to assess patient prognosis. The PND group exhibited significantly lower survival rates at 30 days (87.1% vs 97.84%, P<.001), 3 months (83.99% vs 96.46%, P<.001), 6 months (82.78% vs 95.38%, P<.001), and 12 months (78.85% vs 88.44%, P<.001), and overall survival (P=.03).*

The text now reads:

> *Further survival analysis (Figure 4) was conducted to assess patient prognosis. The PND group exhibited significantly lower survival rates at 30 days (91.5% vs 98.2%, P<.001), 3 months (90.3% vs 97.1%, P<.001), 6 months (89.1% vs 96.1%, P<.001), and 12 months (87.9% vs 92.6%, P=.02).*

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Ding Z, Zhang L, Zhang Y, Yang J, Luo Y, Ge M, et al. A supervised explainable machine learning model for perioperative neurocognitive disorder in liver-transplantation patients and external validation on the medical information mart for intensive care IV database: retrospective study. J Med Internet Res 2025 Jan 15;27:e55046 [FREE Full text] [doi: 10.2196/55046] [Medline: 39813086]

Submitted 04.Dec.2025; this is a non–peer-reviewed article;accepted 11.Dec.2025; published 11.Mar.2026.

Please cite as:
Ding Z, Zhang L, Zhang Y, Yang J, Luo Y, Ge M, Yao W, Hei Z, Chen C
Correction: A Supervised Explainable Machine Learning Model for Perioperative Neurocognitive Disorder in Liver-Transplantation Patients and External Validation on the Medical Information Mart for Intensive Care IV Database: Retrospective Study
J Med Internet Res 2026;28:e88952
URL: https://www.jmir.org/2026/1/e88952
doi:10.2196/88952
PMID:

©Zhendong Ding, Linan Zhang, Yihan Zhang, Jing Yang, Yuheng Luo, Mian Ge, Weifeng Yao, Ziqing Hei, Chaojin Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Infoveillance of COVID-19 Infections in Dentistry Using Platform X: Descriptive Study

Alghalia Al-Mansoori[1], DDS; Ola Al Hayk[1], DDS; Sharifa Qassmi[1], DDS; Sarah M Aziz[1], DDS; Fatima Haouari[1], MSc; Tawanda Chivese[1], MSc, PhD; Faleh Tamimi[1], BDS, MSc, PhD; Alaa Daud[1], BDS, MSc, PGCert

College of Dental Medicine, QU Health, Qatar University, Doha, Qatar

Corresponding Author:
Alaa Daud, BDS, MSc, PGCert
College of Dental Medicine
QU Health
Qatar University
Al Tarfa St
Doha, 2713
Qatar
Phone: 974 77193993
Email: adaud@qu.edu.qa

Related Article:

Correction of: https://www.jmir.org/2025/1/e54650

(J Med Internet Res 2026;28:e86469) doi:10.2196/86469

In "Infoveillance of COVID-19 Infections in Dentistry Using Platform X: Descriptive Study" [1], the authors noted that one name is missing from the Acknowledgments section.

In the Acknowledgments section, the sentence originally appeared as:

> *The authors would like to thank Dr Rula Chami for her contribution to this research. The study received no funding.*

This sentence now reads:

> *The authors would like to thank Dr Rula Chami and Layan Mohamed for their contribution to this research. The study received no funding.*

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Al-Mansoori A, Al Hayk O, Qassmi S, Aziz SM, Haouari F, Chivese T, et al. Infoveillance of COVID-19 infections in dentistry using platform X: descriptive study. J Med Internet Res 2025 Apr 03;27:e54650 [FREE Full text] [doi: 10.2196/54650] [Medline: 40179381]

Submitted 24.Oct.2025; this is a non–peer-reviewed article;accepted 28.Oct.2025; published 31.Mar.2026.

Please cite as:
Al-Mansoori A, Al Hayk O, Qassmi S, Aziz SM, Haouari F, Chivese T, Tamimi F, Daud A
Correction: Infoveillance of COVID-19 Infections in Dentistry Using Platform X: Descriptive Study
J Med Internet Res 2026;28:e86469
URL: https://www.jmir.org/2026/1/e86469
doi:10.2196/86469
PMID:

XSL•FO
RenderX

©Alghalia Al-Mansoori, Ola Al Hayk, Sharifa Qassmi, Sarah M Aziz, Fatima Haouari, Tawanda Chivese, Faleh Tamimi, Alaa Daud. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: AI-Enabled Wearables for Motor Function Assessment and Rehabilitation in Parkinson Disease: Scoping Review

Shengting Li[1,2*], MSc; Siqi Chen[1,2*], MSc; Xiaosong Yu[1,2*], MSc; Huixiang Shang[1,2], MSc; Tingting Tu[1,2]; Mingtao Quan[1,2], MSc

[1]Department of Nursing, Affiliated Hospital of Zunyi Medical College, Zunyi, Guizhou, China

[2]School of Nursing, Zunyi Medical University, Zunyi, Guizhou, China

[*]these authors contributed equally

Corresponding Author:
Mingtao Quan, MSc
Department of Nursing
Affiliated Hospital of Zunyi Medical College
149 Dalian Road, Huichuan District
Zunyi, Guizhou, 563000
China
Phone: 86 18185066688
Email: 865768818@qq.com

Related Article:

Correction of: https://www.jmir.org/2026/1/e85596

(J Med Internet Res 2026;28:e97418) doi:10.2196/97418

In "AI-Enabled Wearables for Motor Function Assessment and Rehabilitation in Parkinson Disease: Scoping Review," the authors requested a correction to the author affiliations.

In the published article, the author affiliations were presented as:

> [1]*School of Nursing, Zunyi Medical University, Zunyi, China*
>
> [2]*Affiliated Hospital of Zunyi Medical College, Zunyi, China*

These have been revised to the following:

> [1]*Department of Nursing, Affiliated Hospital of Zunyi Medical College, Zunyi, Guizhou, China*
>
> [2]*School of Nursing, Zunyi Medical University, Zunyi, Guizhou, China*

This correction does not affect the authorship, author order, or the scientific content of the article.

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Carter-Bawa L, Ostroff JS, Rawl SM, Hirsch EA, Banerjee SC, Ciupek A, et al. Correction: Leveraging Social Media to Achieve Population-Level Reach of Lung Cancer Screening-Eligible Individuals: A RE-AIM Framework Perspective. J Med Internet Res 2026 Mar 05;28:e94664 [FREE Full text] [doi: 10.2196/94664] [Medline: 41813246]

XSL•FO
RenderX

Submitted 07.Apr.2026; this is a non–peer-reviewed article;accepted 09.Apr.2026; published 21.Apr.2026.

Please cite as:
Li S, Chen S, Yu X, Shang H, Tu T, Quan M
Correction: AI-Enabled Wearables for Motor Function Assessment and Rehabilitation in Parkinson Disease: Scoping Review
J Med Internet Res 2026;28:e97418
URL: https://www.jmir.org/2026/1/e97418
doi:10.2196/97418
PMID:

©Shengting Li, Siqi Chen, Xiaosong Yu, Huixiang Shang, Tingting Tu, Mingtao Quan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Effects of an eHealth Cardiac Exercise Rehabilitation Platform for Patients After Percutaneous Coronary Intervention Based on the Persuasive Systems Design Model: Randomized Controlled Trial

Yang Liu[1], MSN; Xiting Huang[1,2], MM; Ziying Dai[1], MM; Zhili Jiang[3], MSN; Wenxiao Wu[1], MSN; Jing Wang[1], MSN; Zhiqian Wang[1], MSN; Luyao Yu[1], MSN; Hanyu Li[1], MSN; Lihua Huang[1], MBBS

[1]Department of Nursing, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China

[2]Department of Nursing, Sun Yat-sen University Cancer Center, Guangzhou, China

[3]School of Humanities and Management, Zhejiang Chinese Medical University, Hangzhou, China

Corresponding Author:
Lihua Huang, MBBS
Department of Nursing
The First Affiliated Hospital
Zhejiang University School of Medicine
79 Qingchun Road
Shangcheng District
Hangzhou, 310003
China
Phone: 86 13867129329
Email: lihuahuang818@zju.edu.cn

Related Article:

Correction of: https://www.jmir.org/2026/1/e71450/

(J Med Internet Res 2026;28:e99326) doi:10.2196/99326

The authors noted two errors in "Effects of an eHealth Cardiac Exercise Rehabilitation Platform for Patients After Percutaneous Coronary Intervention Based on the Persuasive Systems Design Model: Randomized Controlled Trial" [1].

The first affiliation has been revised from the following:

> *The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China*

The affiliation now reads:

> *Department of Nursing, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China*

The degrees of authors were originally listed as the following:

> *Yang Liu, MS; Xiting Huang, MS; Ziying Dai, MS; Zhili Jiang, MS; Wenxiao Wu, MS; Jing Wang, MS; Zhiqian Wang, MS; Luyao Yu, MS; Hanyu Li, MS; Lihua Huang, PhD*

The list of degrees now reads:

> *Yang Liu, MSN; Xiting Huang, MM; Ziying Dai, MM; Zhili Jiang, MSN; Wenxiao Wu, MSN; Jing Wang, MSN; Zhiqian Wang, MSN; Luyao Yu, MSN; Hanyu Li, MSN; Lihua Huang, MBBS*

These correction does not affect the scientific conclusions or validity of the original article.

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Liu Y, Huang X, Dai Z, Jiang Z, Wu W, Wang J, et al. Effects of an eHealth Cardiac Exercise Rehabilitation Platform for Patients After Percutaneous Coronary Intervention Based on the Persuasive Systems Design Model: Randomized Controlled Trial. J Med Internet Res 2026 Jan 14;28:e71450 [FREE Full text] [doi: 10.2196/71450] [Medline: 41534075]

XSL•FO
RenderX

Submitted 24.Apr.2026; this is a non–peer-reviewed article;accepted 27.Apr.2026; published 05.May.2026.

Please cite as:
Liu Y, Huang X, Dai Z, Jiang Z, Wu W, Wang J, Wang Z, Yu L, Li H, Huang L
Correction: Effects of an eHealth Cardiac Exercise Rehabilitation Platform for Patients After Percutaneous Coronary Intervention Based on the Persuasive Systems Design Model: Randomized Controlled Trial
J Med Internet Res 2026;28:e99326
URL: https://www.jmir.org/2026/1/e99326
doi:10.2196/99326
PMID:

©Yang Liu, Xiting Huang, Ziying Dai, Zhili Jiang, Wenxiao Wu, Jing Wang, Zhiqian Wang, Luyao Yu, Hanyu Li, Lihua Huang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparative Analysis of Prescriptions and Pharmacy Services in Internet-Based Psychiatric Hospital During and After the COVID-19 Pandemic: Retrospective Cross-Sectional Observational Study

Guowei Deng[1*], BS; Hui Xia[1*], BS; De-wei Shang[1,2], PhD; Yuguan Wen[1,2], MS; Jinqing Hu[1,2], MS; Yaqian Tan[1,2], PhD

[1]Department of Pharmacy, The Affiliated Brain Hospital, Guangzhou Medical University, Guangzhou, China

[2]Key Laboratory of Neurogenetics and Channelopathies of Guangdong Province and the Ministry of Education of China, Guangzhou Medical University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Yaqian Tan, PhD
Department of Pharmacy, The Affiliated Brain Hospital, Guangzhou Medical University
Mingxin Road 36
Guangzhou
China
Phone: 86 18898533945
Email: tanyaqian2013@163.com

Abstract

Background: The COVID-19 pandemic has significantly accelerated the development of internet-based hospitals and telepharmacy services. However, their characteristics and evolving trends remain unclear.

Objective: This study aimed to assess the associations between distinct pandemic phases and the number of prescriptions, patients' demographic characteristics, drug and disease distribution patterns, and pharmacy service indicators in our internet-based psychiatric hospital.

Methods: In this retrospective cross-sectional observational study, we conducted a full-sample census of prescriptions issued in the internet-based psychiatric hospital of the Affiliated Brain Hospital of Guangzhou Medical University during November 2020-December 2023. Cancelled, pending, and test prescriptions were excluded, and no sampling procedure was used. The research timespan was divided into pandemic and postpandemic phases, and trends of prescriptions were evaluated using interrupted time series analysis. Outcome measures, including patients' sex and age, diagnosed disease, drug type, pharmacist audit time, and audit outcome, were analyzed using the bootstrap method, Pearson chi-square analysis, and multinomial logistic regression.

Results: The segmented regression model revealed significant positive correlation between months and number of prescriptions during pandemic phase ($F_{1,16}$=6.96; P=.02), whereas no significant correlation was detected in postpandemic phase ($F_{1,10}$=2.77; P=.13). Descriptive analysis with bootstrap method revealed that female population were the majority in the pandemic phase (7297/11,812, 61.78%; 95% CI 60.91%-62.70%) and the postpandemic phase (3520/5518, 63.79%; 95% CI 62.58%-65.18%). Young adults aged 18-40 years were the predominant population in the pandemic phase (5606/11,812, 47.46%; 95% CI 46.63%-48.39%) and postpandemic phase (2657/5518, 48.15%; 95% CI 46.79%-49.37%). Depressive disorder and quetiapine were the most frequently diagnosed disease and prescribed drug in both pandemic phases, respectively. The majority of prescriptions were audited within 5 minutes during the pandemic phase (5999/11,812, 50.79%; 95% CI 49.89%-51.65%), while most prescriptions were audited within 1-12 hours in the postpandemic phase (2031/5518, 36.81%; 95% CI 35.61%-37.95%). Pearson chi-square analysis and multinomial logistic regression indicated that variables positively correlated with pandemic phases included female (P=.01; odds ratio [OR] 1.09, 95% CI 1.02-1.17), aged ≤17 years (P<.001; OR 2.20, 95% CI 1.90-2.54), aged 18-40 years (P<.001; OR 1.59, 95% CI 1.38-1.83), audit time between 12 and 24 hours (P=.02; OR 6.26, 95% CI 1.38-28.49), and approved outcome (P=.03; OR 3.97, 95% CI 1.19-13.26). The audit time ≤5 minutes (P=.049; OR 0.22, 95% CI 0.05-0.99) was negatively correlated with the pandemic phases.

Conclusions: This study innovatively applied descriptive and analytic statistical methods to evaluate the associations between different pandemic phases and the prescriptions and pharmacy services in an internet-based psychiatric hospital. This study addressed limitations of existing research through a larger sample size, longer research timespan, and analytic statistical methods.

XSL•FO
RenderX

This study demonstrated early warning indicators and replicable analytic methods for other medical institutions and offered implications in optimizing the efficiency of pharmacy services.

(J Med Internet Res 2026;28:e74059) doi:10.2196/74059

KEYWORDS

COVID-19; mental disorders; internet-based hospital; pharmacy services; prescriptions

Introduction

Background

The COVID-19 pandemic has profoundly affected the world and our daily lives [1-5]. In response to the COVID-19 outbreaks, the Chinese government had taken a range of actions, including temporary hospitals, lockdown, and quarantine [6,7]. However, these strict control measures inevitably disrupted conventional outpatient services, especially for patients with mental disorders [8]. Specifically, patients with chronic mental illness generally require long-term medication and regular medical assistance, and lockdown can cause difficulties in obtaining drugs and even the danger of discontinuation [9-11]. In addition, patients with long-term serious mental illness are often physically or socially disadvantaged, and COVID-19 highlighted these preexisting differences [12-15]. Finally, it has been reported that the global prevalence of mental health problems, such as loneliness, panic, anxiety, and depression, has increased significantly since the outbreak of the COVID-19 pandemic [16,17]. Hence, in order to meet the challenges of patients' medical needs, and in line with the recommendation of the World Health Organization to strengthen the digitalization of health systems, online medical services have developed rapidly worldwide [18-20]. The National Health Commission of China has also promoted the establishment of internet-based hospitals and timely translated into emergency control measures of COVID-19 [21,22]. Internet-based hospitals in China have been reported to grow rapidly from approximately 1600 in June 2021 to over 3000 by June 2023 [23], demonstrating the growing importance of this online service model.

In fact, with the advancement of artificial intelligence, 5G networks, and virtual reality, internet-based hospitals have developed rapidly in the last decade [24]. Internet-based hospitals displayed significant advantages over offline services in terms of spatial accessibility and cost-effectiveness [25,26]. Further, internet-based hospitals established by physical hospitals can integrate online and offline medical information and offer consultation, prescription, and follow-up services to ensure the continuity of health care for patients [27]. It is worth mentioning that there has been a major problem of uneven distribution of medical resources in China, primarily manifested as great regional differences [28,29]. However, the establishment of internet-based hospitals has greatly improved the accessibility of high-quality medical resources in big cities, thereby narrowing the differences in medical resources in China [30,31].

Subsequently, with the outbreak of COVID-19, extensive evidence has emphasized the growing importance of telehealth in providing stable health care services and curbing the spread of the pandemic [32,33]. By transferring nonemergency cases to online services, internet-based hospitals significantly alleviated the overburden of health care systems, thereby optimizing the allocation of medical resources [34,35]. Additionally, pharmacy services, such as medication consultation, medication therapy management, patient education, and drug delivery services, have been shown to significantly reduce the financial burdens of patients and improve their medication adherence [25,36,37]. Despite these potential benefits, disparities among different populations might also be inadvertently exacerbated during the pandemic, including internet facility conditions, digital skills, and physical conditions [38-41]. Meanwhile, due to the disadvantages such as insufficient interaction between doctors and patients and poor timeliness, the quality of online health care services has received great concern [22,42,43].

As for patients with mental disorders, telepsychiatry digital platforms have fundamentally improved the accessibility of patients and exhibited therapeutic effects comparable to offline treatment [44,45]. To meet the growing medication needs of patients, our hospital (the Affiliated Brain Hospital of Guangzhou Medical University) launched an internet-based psychiatric hospital platform and officially started to provide online pharmacy services since November 2020. Specifically, patients can receive remote consultation from doctors at home, pay online, and settle in real time through medical insurance. Subsequently, electronic prescriptions are audited by trained and authorized pharmacists, and drugs are delivered directly to patients through qualified third-party logistics channels to reduce the spread of COVID-19.

Relation to Previous Work

At present, most studies on the telemedicine services during the COVID-19 pandemic have focused on general hospitals [46-49], and few studies have explored the online pharmacy services in internet-based psychiatric hospitals [21,23]. Moreover, the limitations of existing studies in this area still exist, including small sample sizes, restricted study periods, and a lack of comparison between different pandemic phases. Besides, it is still unclear how the gradual resumption of normal medical services after the pandemic will affect internet-based hospitals. Thus, to fill these knowledge gaps, we conducted a retrospective cross-sectional observational study using the electronic prescriptions issued in our internet-based psychiatric hospital during November 2020 to December 2023. This study had a larger sample size and a longer research timespan than previous studies in this field and used analytic statistical methods to ensure the reliability of our findings. This study could gain in-depth insights into the development trends of prescriptions and key predictors of pharmacy services in internet-based psychiatric hospitals and offer empirical guidance for other medical institutions in the effective responses to public health emergencies.

XSL•FO
RenderX

Objectives

In this study, we aimed to analyze the associations between different COVID-19 pandemic phases (predictor) and the trends and distribution patterns of the electronic prescriptions through descriptive and analytic statistical methods. The specific outcomes, including monthly prescription numbers, patients' demographic characteristics (sex and age), clinical characteristics (primary diagnosed disease and type of drug), and pharmacy service indicators (pharmacist audit time and audit outcome), would be assessed to elucidate the changes in their distribution patterns driven by the pandemic phases.

Methods

All methods and findings of this study were reported based on the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [50] and the JARS (Journal Article Reporting Standards) guidelines [51,52]. A completed and filled out STROBE checklist of this study is provided as Multimedia Appendix 1.

Study Design

This study was a retrospective cross-sectional observational study that comparatively analyzed the prescriptions and pharmacy services from an internet-based psychiatric hospital during November 2020 to December 2023. We used descriptive and analytic statistical methods to analyze the associations between the 2 distinct pandemic phases (pandemic phase and postpandemic phase) and the long-term trends of the prescriptions, patients' demographic characteristics, drug and disease distribution patterns, and pharmacy service indicators.

Settings and Data Collection

All data in this study were obtained from the internet-based psychiatric hospital of the Affiliated Brain Hospital of Guangzhou Medical University, a tertiary psychiatric hospital located in Guangzhou, China. The telepharmacy service platform was officially launched and started to provide online services since November 2020. Patients can visit our internet-based psychiatric hospital through the app developed by Guangdong Yunhui Technology Co, Ltd. After patients log on to the app and input their personal information, doctors will confirm their identity information and issue prescriptions according to patients' condition. All electronic prescriptions from the internet-based psychiatric hospital are manually audited by trained and authorized pharmacists. The detailed workflow of the doctor prescription and the pharmacist audit process is demonstrated in Figure 1.

Figure 1. Flowchart depicting the doctor prescription and pharmacist audit process in the internet-based psychiatric hospital of the Affiliated Brain Hospital of Guangzhou Medical University during November 2020 to December 2023.

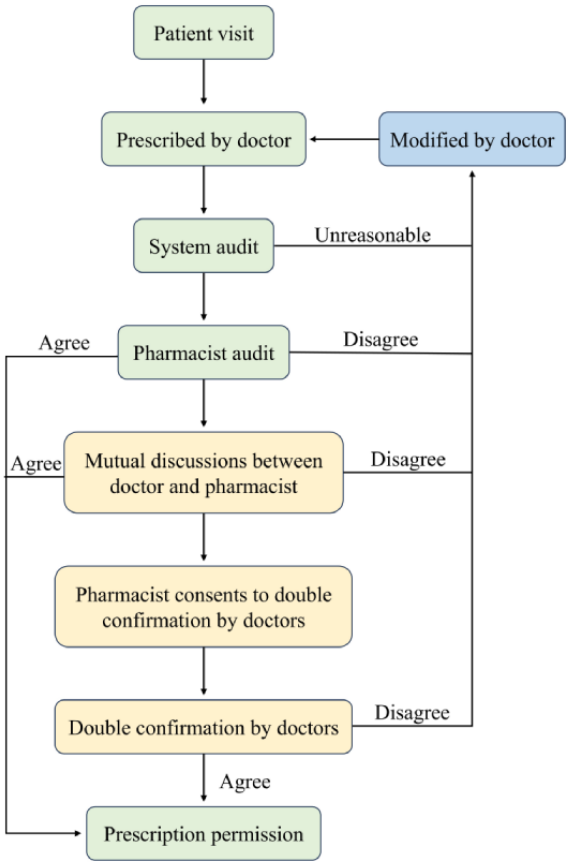

Data were collected from the electronic prescriptions of patients who received online medical services in our internet-based psychiatric hospital during November 1, 2020, to December 31, 2023. In this study, the exposure was the different pandemic phases. This study was a retrospective cross-sectional observational study, and no follow-up was performed. Data collection was conducted on a single day of March 11, 2024.

XSL•FO
RenderX

Ultimately, a total of 17,330 electronic prescriptions with 36,088 drug records were identified and included in the final analysis.

Eligibility Criteria

This study aimed to analyze the entire population of prescriptions issued in our online platform during November 2020 to December 2023. There were no restrictions on prescription inclusion based on patients' demographic characteristics (eg, sex, age, ethnicity, and socioeconomic status), and the inclusion criterion was defined as all prescriptions issued throughout the study period. Therefore, we essentially adopted a full-sample census approach. During the subsequent data screening process, cancelled, pending, and test prescriptions were considered to meet exclusion criteria and were removed from further data analysis. After data screening, a total of 17,330 prescriptions were included in the final data analysis. All data items of these included prescriptions were complete with no missing data. In addition, given the full-sample census method used, no sampling procedure was involved in this study. The detailed workflow of the eligibility criteria and data screening process is illustrated in Figure 2.

Figure 2. Flowchart demonstrating the eligibility criteria and data screening process of the electronic prescriptions issued in the internet-based psychiatric hospital of the Affiliated Brain Hospital of Guangzhou Medical University during November 2020 to December 2023.

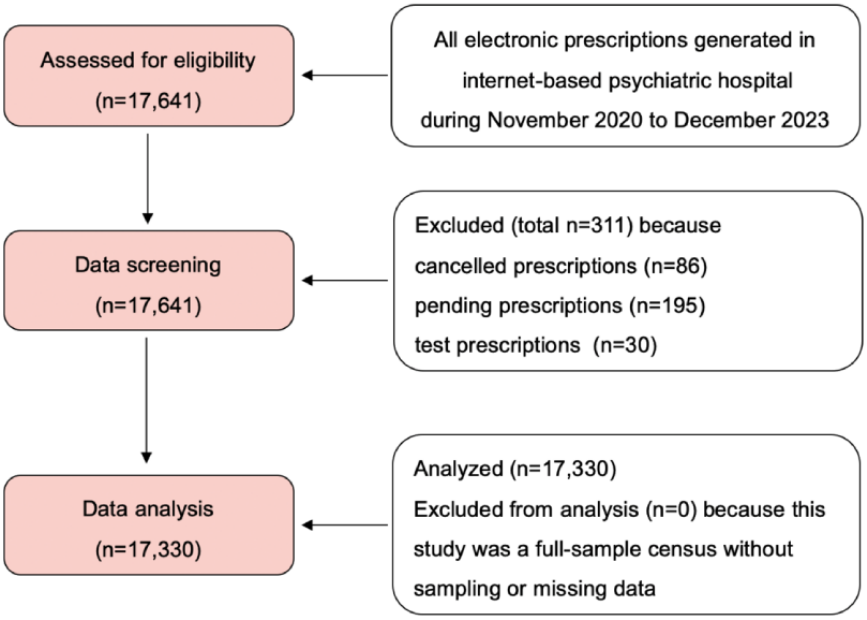

Variables

In this study, the outcomes were measured at the prescription level, including the number of prescriptions, patients' demographic characteristics (sex and age), clinical characteristics (primary diagnosed disease and type of drug), and pharmacy service indicators (pharmacist audit time and audit outcome). Within the standard framework of this observational research, the pandemic phases could theoretically be considered as potential "predictors" of the outcomes. The conditions in this study were all naturally observed, and no causal analysis of observed phenomena was performed. Thus, no variables were formally specified as confounders or effect modifiers.

Bias

To prevent data inconsistencies due to periodic updates of the system when extracting data over multiple days, such as retrospective corrections of prescription status or batch updates of diagnoses and drug codes, all raw data in this study were extracted directly from our online platform backend system within a single day of March 11, 2024. This approach allowed us to ensure that all collected prescriptions had the same data status or version, thereby minimizing potential bias from data sources and guaranteeing data standardization.

Data Sources or Measurement and Quantitative Variables

In December 2022, the Chinese government announced the end of all COVID-19 prevention and quarantine control measures in China. Based on this, the timeline of the COVID-19 pandemic was divided into 2 phases: the pandemic phase (November 2020-December 2022) and postpandemic phase (January 2023-December 2023). Exposure to the distinct COVID-19 pandemic phases was therefore defined and evaluated throughout the study period.

Based on the 2 pandemic phases divided, the number of monthly prescriptions and cumulative prescriptions were presented using a continuous time variable in months. Further, a 2-stage interrupted time series (ITS) analysis was conducted. In the pandemic phase, we adjusted the starting time point to July 2021 to exclude abnormal values (values deviating more than 2 times the CI from the absolute value) in June 2021. A segmented regression model in ITS analysis was used to evaluate the association between COVID-19 relaxation in China and the trends of monthly prescriptions [53,54]. A continuous time variable in months was applied in the model, and data were presented as absolute values with 95% CI. The P value <.05 was accepted as statistical significance in the segmented regression model.

XSL•FO
RenderX

Descriptive data of patients' demographic characteristics, drug and disease distribution patterns, and pharmacy service indicators were presented using frequencies and percentages. The subgroups of age were assigned as follows: pediatric and adolescent (≤17 years), young adult (18-40 years), middle-aged (41-65 years), and older adult (≥66 years). The age of patients was assessed using median with IQR.

The pharmacy service indicators included pharmacist audit time and audit outcome. Pharmacist audit time was defined as the total length of time (in minutes) from prescription submission to audit completion, which was automatically recorded in our audit system. During data analysis, the audit time was divided into 7 subgroups according to our institutional pharmacy operation guidelines, including "≤5 minutes," "5-30 minutes," "30 minutes-1 hour," "1-12 hours," "12-24 hours," "24-48 hours," and " 48 hours." The audit outcomes mainly included "approved by pharmacists," "double confirmation by doctors," and "not approved by pharmacists." We used consistent evaluation criteria for the audit outcomes of all prescriptions in this study according to our institutional pharmacy operation guidelines. To be specific, "approved by pharmacists" referred to prescriptions that meet the patient's clinical condition and rationale for medication, "double confirmation by doctors" referred to prescriptions with potential medication errors (eg, inappropriate indication, repeated medication, and overdose) that required double confirmation by doctors, and "not approved by pharmacists" referred to prescriptions disapproved by pharmacists due to serious safety concerns (eg, contraindicated drugs and lethal doses).

To examine the robustness of the descriptive data, including patients' sex and age, diagnosed disease, drug type, pharmacist audit time, and audit outcome, percentage values were sampled repeatedly for 1000 times using the bootstrap method and were presented as absolute values with 95% CI [55,56].

Statistical Methods

Statistical data were presented for multiple variables, including sex, age, pharmacist audit time, and audit outcome. The general associations between pandemic phases and outcomes were examined using crosstabs analysis and Pearson chi-square analysis [54]. If the association was significant ($P<.05$), we further performed a multinomial logistic regression analysis to determine the source of significance using pandemic phases as a covariate [57]. Significance level of the Wald inclusion test statistic was applied with P value <.05. Odds ratios (ORs) with 95% CI and SE were calculated to quantify the associations between pandemic phases and outcomes. In this study, our predetermined α level was .05, and a 2-sided P value <.05 was considered statistically significant. All statistical analysis and graphical representations were performed using SPSS (version 26.0.0.2; IBM Corp) and GraphPad Prism (version 10.1.2; GraphPad Software, Inc).

Ethical Considerations

This study was approved by the institutional review board (IRB) of the Affiliated Brain Hospital of Guangzhou Medical University following a thorough review of the research protocol (approval 2025111). The informed consent of this study was waived by the IRB, and the IRB allowed the primary data collection and secondary analysis of research data without additional consent. We confirm that this study adhered strictly to the principles of privacy and confidentiality protection. The research team complied with all relevant local, national, and international laws and regulations regarding the protection of personal information, privacy, and human rights. All sensitive data related to patients, including name, ethnicity, socioeconomic status, education level, residential address, contact information, geographic distribution, prescription cost, primary complaint, medical history, previous diagnosis, information of the clinician, and content of the conversation, were deidentified to ensure privacy and security. Besides sensitive information, all other collected data were reported in this study. This study involved prescription data from the local database of our hospital and did not involve human experimentation or compensation. We confirm that no personally identifiable information of patients was accessible to the research team. We confirm that no identification of individual participants or users in any images of the manuscript or supplementary material is possible.

Results

General Trends of Monthly Prescriptions

In this study, a total of 17,330 electronic prescriptions were finally identified, of which 11,812 prescriptions were processed during the pandemic phase, and 5518 prescriptions were issued during the postpandemic phase. As shown in Figure 3, the general trend of monthly prescriptions reflected a fluctuating tendency from November 2020 to December 2023. During the early stage of the pandemic phase from November 2020 to May 2021, the number of monthly prescriptions remained relatively low, with 76 prescriptions in total. Subsequently, the number of monthly prescriptions dramatically surged in June 2021 (n=3427), and then decreased to an average level of approximately 350 until another peak in November 2022 (n=1506). In the postpandemic phase, the number of monthly prescriptions stayed relatively stable around an average level of 460. On the other hand, the cumulative number of prescriptions exhibited a continuous upward tendency and peaked at 36,088 in December 2023.

XSL•FO
RenderX

Figure 3. The number of monthly prescriptions and cumulative prescriptions in the internet-based psychiatric hospital of the Affiliated Brain Hospital of Guangzhou Medical University. Sources of data are all the electronic prescriptions issued during November 2020 to December 2023. Values are presented using a continuous time variable in months.

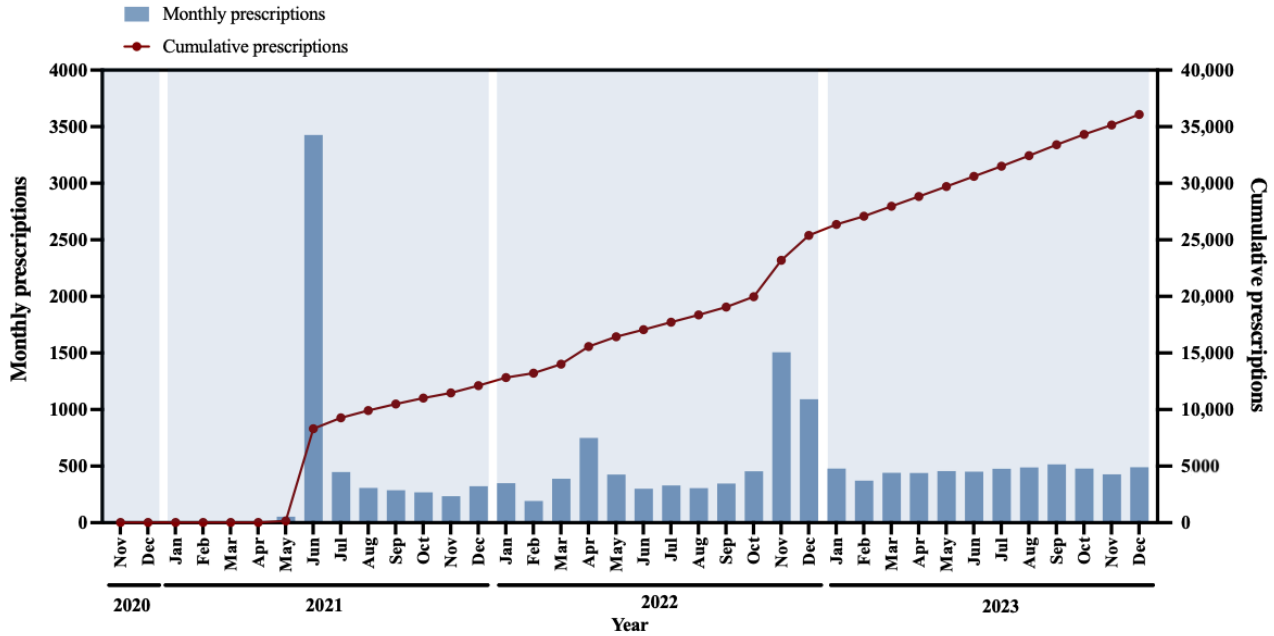

ITS Analysis of Monthly Prescriptions

A further ITS analysis was conducted to evaluate the association between COVID-19 relaxation and the number of monthly prescriptions. We conducted a 2-stage ITS analysis using the data ranging from July 2021 to December 2023. As depicted in Figure 4, the segmented regression model revealed a significant positive correlation between the months and the number of prescriptions (y=34.52*x+133.60; r=0.55; $F_{1,16}$=6.96; P=.02; slope 34.52; 95% CI 6.78-62.27). These findings suggested an increasing tendency in patients' needs and interest in our internet-based psychiatric hospital. In the postpandemic phase, the number of monthly prescriptions remained generally stable, and the trend was not statistically significant compared to baseline (y=4.83*x+428.50; r=0.46; $F_{1,10}$=2.77; P=.13; slope 4.83; 95% CI –1.63 to 11.28). These findings indicated a stable and continuous service model of internet-based psychiatric hospital and a steady medication adherence of patients during January 2023 to December 2023.

Figure 4. Segmented regression model in the 2-stage interrupted time series analysis evaluating the association between COVID-19 relaxation and the number of monthly prescriptions in the internet-based psychiatric hospital of the Affiliated Brain Hospital of Guangzhou Medical University. Sources of data are all the electronic prescriptions issued during July 2021 to December 2023. Values are presented as absolute values (blue dots) with 95% CI (red shadow) with a continuous time variable in months. The red solid lines indicate lines of regression, and the vertical black dashed line indicates the end of all COVID-19 prevention and quarantine control measures in China in December 2022. The P value <.05 indicates a significant difference between the regression line and the baseline.

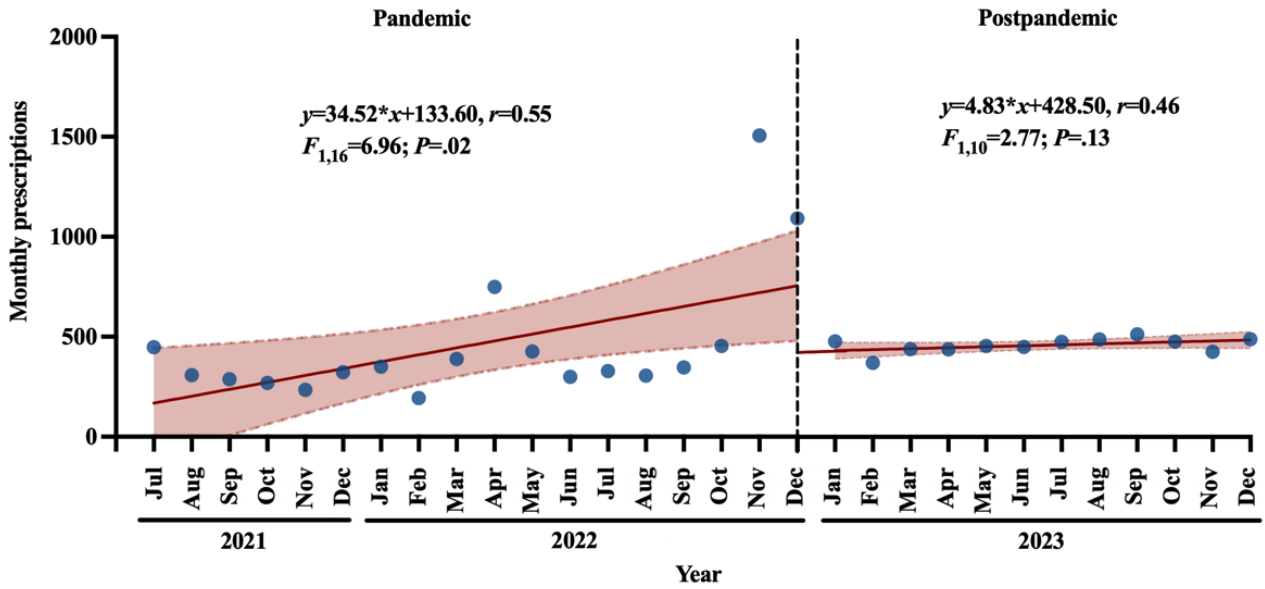

XSL•FO
RenderX

Demographic Characteristics of Patients

In this study, a total of 17,330 electronic prescriptions were collected, of which 11,812 prescriptions were issued during pandemic phase, and 5518 prescriptions were issued during the postpandemic phase. As the detailed characteristics of patients displayed in Table 1, the majority of patients were female during the pandemic phase (7297/11,812, 61.78%; 95% CI 60.91%-62.70%) and the postpandemic phase (3520/5518, 63.79%; 95% CI 62.58%-65.18%). There were 4515 male patients in the pandemic phase (4515/11,812, 38.22%; 95% CI 37.36%-39.03%) and 1998 male patients in the postpandemic phase (1998/5518, 36.21%; 95% CI 34.99%-37.32%).

Table 1. Demographic characteristics of patients who visited the internet-based psychiatric hospital of the Affiliated Brain Hospital of Guangzhou Medical University during November 2020 to December 2023[a].

Characteristic	Pandemic (n=11,812), n (%)	95% CI	Postpandemic (n=5518), n (%)	95% CI	Chi-square (*df*)	*P* value
Sex					6.5 (1)	.01
Female	7297 (61.78)	60.91-62.70	3520 (63.79)	62.58-65.18		
Male	4515 (38.22)	37.36-39.03	1998 (36.21)	34.99-37.32		
Age (years)[b]					295.4 (3)	<.001
≤17	2873 (24.32)	23.47-25.19	1880 (34.07)	32.78-35.40		
18-40	5606 (47.46)	46.63-48.39	2657 (48.15)	46.79-49.37		
41-65	2400 (20.32)	19.61-21.07	703 (12.74)	11.92-13.59		
≥66	933 (7.90)	7.42-8.36	278 (5.04)	4.48-5.62		

[a]Values are presented as absolute values and percentages with 95% CI using the bootstrap method. The general associations between different pandemic phases and sex and age of patients are evaluated with crosstabs analysis and presented with Pearson chi-square values.

[b]Pandemic: median 21 (IQR 18-44) years, postpandemic: median 21 (IQR 16-33) years.

Regarding age, young adults aged between 18 and 40 years accounted for the predominant population (5606/11,812, 47.46%; 95% CI 46.63%-48.39%) during pandemic phase, followed by patients aged ≤17 years (2873/11,812, 24.32%; 95% CI 23.47%-25.19%), patients aged between 41 and 65 years (2400/11,812, 20.32%; 95% CI 19.61%-21.07%), and patients aged ≥66 years (933/11,812, 7.90%; 95% CI 7.42%-8.36 %). Similarly, in postpandemic phase, young adults aged between 18 and 40 years accounted for the majority (2657/5518, 48.15%; 95% CI 46.79%-49.37%), followed by patients aged ≤17 years (1880/5518, 34.07%; 95% CI 32.78%-35.40%), patients aged between 41 and 65 years (703/5518, 12.74%; 95% CI 11.92%-13.59%), and patients aged ≥66 years (278/5518, 5.04%, 95 % CI 4.48%-5.62%). The median value of age was 21 (IQR 18-44) years during the pandemic phase and 21 (IQR 16-33) years in the postpandemic phase.

Overall, the Pearson chi-square analysis suggested significant statistical differences in sex distribution (χ^2_1=6.5; *P*=.01) and age distribution (χ^2_3=295.4; *P*<.001) between the 2 pandemic phases. These differences thereby suggested the necessity to perform further multinomial logistic regression analysis to determine the source of significance.

A follow-up multinomial logistic regression model was used to analyze the influences of different COVID-19 phases on the demographic characteristics of patients. As shown in Table 2, the postpandemic phase exhibited a positive correlation with the female group (*P*=.01; OR 1.09, 95% CI 1.02-1.17), suggesting a significantly increased proportion of female patients from 61.78% (7297/11,812) in the pandemic phase to 63.79% (3520/5518) in the postpandemic phase. Further, postpandemic phase displayed positive correlations with patients aged ≤17 years (*P*<.001; OR 2.20, 95% CI 1.90-2.54) and patients aged 18-40 years (*P*<.001; OR 1.59, 95% CI 1.38-1.83), suggesting significantly increased proportions of patients aged ≤17 years from 24.32% (2873/11,812) in pandemic phase to 34.07% (1880/5518) in postpandemic phase, and patients aged 18-40 years from 47.46% (5606/11,812) in pandemic phase to 48.15% (2657/5518) in postpandemic phase. We did not find a significant association between pandemic phases and patients aged ≥66 years across different pandemic phases (*P*=.83; OR 0.98, 95% CI 0.84-1.15).

XSL•FO
RenderX

Table 2. The influences of COVID-19 phases on the subgroups of patients' demographic characteristics in a multinomial logistic regression model[a].

Characteristic	SE	Wald	*df*	*P* value	OR[b] (95% CI)
Sex					
Female	0.03	6.51	1	.01	1.09 (1.02-1.17)
Male[c]	—[d]	—	—	—	1.00 (—)
Age (years)					
≤17	0.07	111.53	1	<.001	2.20 (1.90-2.54)
18-40	0.07	41.24	1	<.001	1.59 (1.38-1.83)
41-65	0.08	0.05	1	.83	0.98 (0.84-1.15)
≥66[c]	—	—	—	—	1.00 (—)

[a]Data presented in this table are from the follow-up analysis of patients' demographic characteristics to determine the source of the statistical difference. The SE and Wald inclusion test statistic are applied to quantify the associations between pandemic phases and outcomes.

[b]OR: odds ratio.

[c]Reference group.

[d]Not available.

Distribution of Primary Diagnosed Diseases

As listed in Table 3, the top 10 primary diagnosed diseases in frequency were examined in our descriptive analysis. During the pandemic phase, a total of 83 diagnosed diseases were detected. The primary diagnosis of depressive disorder (3539/11,812, 29.96%; 95% CI 29.15%-30.79%) ranked first in frequency, followed by schizophrenia (2095/11,812, 17.74%; 95% CI 17.10%-18.40%) and bipolar disorder (1632/11,812, 13.82%; 95% CI 13.17%-14.50%). The primary diagnosis of epilepsy (108/11,812, 0.91%; 95% CI 0.77%-1.06%) ranked lowest among the top 10 primary diagnosed diseases. During the postpandemic phase, 58 diagnosed diseases in total were identified, and the leading 3 were depressive disorder (2094/5518, 37.95%; 95% CI 36.74%-39.12%), mood disorder (1191/5518, 21.58%; 95% CI 20.50%-22.74%), and schizophrenia (607/5518, 11%; 95% CI 10.20%-11.82%). In the postpandemic phase, epilepsy (49/5518, 0.89%; 95% CI 0.67%-1.12%) ranked lowest among the top 10 primary diagnosed diseases.

XSL•FO
RenderX

Table 3. The distribution of the top 10 primary diagnosed diseases from the electronic prescriptions processed in the internet-based psychiatric hospital of the Affiliated Brain Hospital of Guangzhou Medical University during November 2020 to December 2023[a].

Rank	Primary diagnosed disease	Values, n (%)	95% CI
Pandemic (n=11,812)			
1	Depressive disorder	3539 (29.96)	29.15-30.79
2	Schizophrenia	2095 (17.74)	17.10-18.40
3	Bipolar disorder	1632 (13.82)	13.17-14.50
4	Mood disorder	1444 (12.22)	11.65-12.86
5	Anxiety disorder	1208 (10.23)	9.61-10.82
6	Obsessive-compulsive disorder	328 (2.78)	2.47-3.08
7	Attention-deficit/hyperactivity disorder	232 (1.96)	1.73-2.22
8	Alzheimer disease	164 (1.39)	1.19-1.61
9	Autistic disorder	116 (0.98)	0.81-1.15
10	Epilepsy	108 (0.91)	0.77-1.06
Postpandemic (n=5518)			
1	Depressive disorder	2094 (37.95)	36.74-39.12
2	Mood disorder	1191 (21.58)	20.50-22.74
3	Schizophrenia	607 (11.00)	10.20-11.82
4	Anxiety disorder	323 (5.85)	5.21-6.52
5	Bipolar disorder	301 (5.45)	4.89-5.96
6	Obsessive-compulsive disorder	162 (2.94)	2.56-3.35
7	Attention-deficit/hyperactivity disorder	127 (2.30)	1.94-2.66
8	Autistic disorder	117 (2.12)	1.81-2.45
9	Alzheimer disease	72 (1.30)	1.04-1.59
10	Epilepsy	49 (0.89)	0.67-1.12

[a]Values are presented as absolute values and percentages with 95% CI using the bootstrap method.

Distribution of Drugs in Prescriptions

A total of 114 types of drugs prescribed in 25,402 times during the pandemic phase and 126 types of drugs prescribed in 10,686 times during the postpandemic phase were identified. As listed in Table 4, the top 10 drugs in frequency were examined in our descriptive analysis. During the pandemic phase, the predominant drugs in prescriptions were quetiapine (2653/25,402, 10.44%; 95% CI 10.03%-10.81%), lithium carbonate (1439/25,402, 5.66%; 95% CI 5.39%-5.93%), and escitalopram (1414/25,402, 5.57%; 95% CI 5.30%-5.85%). The drug agomelatine (870/25,402, 3.42%; 95% CI 3.20%-3.65%) had the lowest frequency among the top 10 prescribed drugs. Likewise, in the postpandemic phase, the top 3 most frequently prescribed drugs were quetiapine (1276/10,686, 11.94%; 95% CI 11.37%-12.55%), lithium carbonate (733/10,686, 6.86%; 95% CI 6.36%-7.37%), and escitalopram (710/10,686, 6.64%; 95% CI 6.20%-7.07%). Olanzapine (458/10,686, 4.29%; 95% CI 3.91%-4.71%) ranked lowest in the top 10 prescribed drugs during the postpandemic phase.

XSL•FO
RenderX

Table 4. The distribution of the top 10 drugs from the electronic prescriptions processed in the internet-based psychiatric hospital of the Affiliated Brain Hospital of Guangzhou Medical University during November 2020 to December 2023[a].

Rank	Drug	Values, n (%)	95% CI
Pandemic (n=25,402)[b]			
1	Quetiapine	2653 (10.44)	10.03-10.81
2	Lithium carbonate	1439 (5.66)	5.39-5.93
3	Escitalopram	1414 (5.57)	5.30-5.85
4	Olanzapine	1398 (5.50)	5.26-5.75
5	Sodium Valproate	1356 (5.34)	5.07-5.62
6	Sertraline	1356 (5.34)	5.08-5.61
7	Aripiprazole	1247 (4.91)	4.63-5.18
8	Trihexyphenidyl	1185 (4.66)	4.41-4.93
9	Tandospirone	932 (3.67)	3.46-3.89
10	Agomelatine	870 (3.42)	3.20-3.65
Postpandemic (n=10,686)[b]			
1	Quetiapine	1276 (11.94)	11.37-12.55
2	Lithium carbonate	733 (6.86)	6.36-7.37
3	Escitalopram	710 (6.64)	6.20-7.07
4	Tandospirone	657 (6.15)	5.66-6.64
5	Sodium valproate	612 (5.73)	5.31-6.14
6	Sertraline	584 (5.47)	5.02-5.90
7	Aripiprazole	570 (5.33)	4.90-5.75
8	Lamotrigine	495 (4.63)	4.23-5.01
9	Fluoxetine	460 (4.30)	3.93-4.67
10	Olanzapine	458 (4.29)	3.91-4.71

[a]Values are presented as absolute values and percentages with 95% CI using the bootstrap method.

[b]Some prescriptions may contain more than 1 drug.

Distribution of Prescription Audit Time and Audit Outcome

The detailed characteristics of prescription audit in the internet-based psychiatric hospital are listed in Table 5. During the pandemic phase, the majority of prescriptions (5999/11,812, 50.79%; 95% CI 49.89%-51.65%) were audited in ≤5 minutes, followed by prescriptions audited within 5-30 minutes (3255/11,812, 27.56%; 95% CI 26.74%-28.36%) and prescriptions audited within 1-12 hours (1466/11,812, 12.41%; 95% CI 11.86%-13.04%). The group of audit time 48 hours had the lowest number of prescriptions (4/11,812, 0.03%; 95% CI 0.01%-0.07%). In contrast, during postpandemic phase, most prescriptions (2031/5518, 36.81%; 95% CI 35.61%-37.95%) were audited within 1-12 hours, followed by prescriptions audited within 5-30 minutes (1359/5518, 24.63%; 95% CI 23.47%-25.73%) and prescriptions audited in ≤5 minutes (998/5518, 18.09%; 95% CI 17.13%-19.17%). Likewise, in the postpandemic phase, the group of audit time 48 hours had the lowest number of prescriptions (3/5518, 0.05%; 95% CI 0.01%-0.13%).

XSL•FO
RenderX

Table 5. The characteristics of prescription audit in the internet-based psychiatric hospital of the Affiliated Brain Hospital of Guangzhou Medical University during November 2020 to December 2023[a].

Characteristic	Pandemic (n=11,812), n (%)	95% CI	Postpandemic (n=5518), n (%)	95% CI	Chi-square (*df*)	*P* value
Audit time					2784.5 (6)	<.001
≤5 minutes	5999 (50.79)	49.89-51.65	998 (18.09)	17.13-19.17		
5-30 minutes	3255 (27.56)	26.74-28.36	1359 (24.63)	23.47-25.73		
30 minutes-1 hour	957 (8.10)	7.65-8.57	669 (12.12)	11.31-13.01		
1-12 hours	1466 (12.41)	11.86-13.04	2031 (36.81)	35.61-37.95		
12-24 hours	86 (0.73)	0.57-0.88	404 (7.32)	6.64-8.07		
24-48 hours	45 (0.38)	0.28-0.50	54 (0.98)	0.76-1.23		
48 hours	4 (0.03)	0.01-0.07	3 (0.05)	0.01-0.13		
Audit outcome					601.0 (2)	<.001
Approved by pharmacists	9844 (83.34)	82.68-84.06	5327 (96.54)	96.04-97.01		
Double confirmation by doctors	1946 (16.47)	15.85-17.05	188 (3.41)	2.98-3.88		
Not approved by pharmacists	22 (0.19)	0.12-0.25	3 (0.05)	0.01-0.13		

[a]Values are presented as absolute values and percentages with 95% CI using the bootstrap method. The general associations between different pandemic phases and audit time and audit outcome are evaluated with crosstabs analysis and presented with Pearson chi-square values.

As for audit outcomes, most prescriptions (9844/11,812, 83.34%; 95% CI 82.68%-84.06%) were approved by pharmacists during the pandemic phase. Of the 11,812 prescriptions in the pandemic phase, 1946 (16.47%; 95% CI 15.85%-17.05%) required double confirmation by doctors before approval, and 22 (0.19%; 95% CI 0.12%-0.25%) were not approved by pharmacists. Similarly, during the postpandemic phase, the vast majority of prescriptions (5327/5518, 96.54%; 95% CI 96.04%-97.01%) were approved by pharmacists. Of the 5518 prescriptions in the postpandemic phase, 188 (3.41%; 95% CI 2.98%-3.88%) required double confirmation by doctors before approval, and 3 (0.05%; 95% CI 0.01%-0.13%) were not approved by pharmacists.

Overall, the Pearson chi-square analysis revealed a significant statistical difference in audit time (χ^2_6=2784.5; *P*<.001) and audit outcomes (χ^2_2=601.0; *P*<.001) between the 2 pandemic phases. These differences thereby suggested the necessity to perform further multinomial logistic regression analysis to determine the source of significance.

A follow-up multinomial logistic regression model was used to analyze the influences of different COVID-19 phases on the characteristics of prescription audit. As shown in Table 6, the postpandemic phase exhibited a negative correlation with audit time ≤5 minutes (*P*=.049; OR 0.22, 95% CI 0.05-0.99), suggesting a significantly decreased proportion of prescriptions audited in ≤5 minutes from 50.79% (5999/11,812) in the pandemic phase to 18.09% (998/5518) in the postpandemic phase. There was a positive correlation with audit time within 12-24 hours (*P*=.02; OR 6.26, 95% CI 1.38-28.49), suggesting a significantly increased proportion of prescriptions audited within 12-24 hours from 0.73% (86/11,812) in the pandemic phase to 7.32% (404/5518) in the postpandemic phase. We did not find significant associations between different pandemic phases and audit time within 5-30 minutes (*P*=.44; OR 0.56, 95% CI 0.12-2.49), audit time within 30-60 minutes (*P*=.93; OR 0.93, 95% CI 0.21-4.17), audit time within 1-12 hours (*P*=.42; OR 1.85, 95% CI 0.41-8.27), or audit time within 24-48 hours (*P*=.55; OR 1.60, 95% CI 0.34-7.53).

XSL•FO
RenderX

Table 6. The influences of COVID-19 phases on the subgroups of prescription audit indicators in a multinomial logistic regression model[a].

Characteristic	SE	Wald	*df*	*P* value	OR[b] (95% CI)
Audit time					
≤5 minutes	0.77	3.88	1	.049	0.22 (0.05-0.99)
5-30 minutes	0.76	0.59	1	.44	0.56 (0.12-2.49)
30 minutes-1 hour	0.76	0.01	1	.93	0.93 (0.21-4.17)
1-12 hours	0.77	0.64	1	.42	1.85 (0.41-8.27)
12-24 hours	0.77	5.63	1	.02	6.26 (1.38-28.49)
24-48 hours	0.79	0.35	1	.55	1.60 (0.34-7.53)
48 hours[c]	—[d]	—	—	—	1.00 (—)
Audit outcome					
Approved by pharmacists	0.62	5.01	1	.03	3.97 (1.19-13.26)
Double confirmation by doctors	0.62	0.31	1	.58	0.71 (0.21-2.39)
Not approved by pharmacists[c]	—	—	—	—	1.00 (—)

[a]Data presented in this table are from the follow-up analysis of the indicators of prescription audit to determine the source of statistical difference. The SE and Wald inclusion test statistic are applied to quantify the associations between pandemic phases and outcomes.

[b]OR: odds ratio.

[c]Reference group.

[d]Not available.

Furthermore, the postpandemic phase displayed a positive correlation with the approved group (*P*=.03; OR 3.97, 95% CI 1.19-13.26), suggesting a significantly increased proportion of approved outcomes from 83.34% (9844/11,812) in the pandemic phase to 96.54% (5327/5518) in the postpandemic phase. We did not find significant associations between different pandemic phases and doctor double confirmation group (*P*=.58; OR 0.71, 95% CI 0.21-2.39).

Discussion

Principal Findings

In general, we found distinct trends of prescriptions in different phases of the pandemic, such that an overall upward trend during the pandemic phase and a stable tendency during the postpandemic phase. We found that female and young adults were the predominant groups, and their proportions increased with the development of the pandemic phases. Additionally, our findings demonstrated that the distribution patterns of primary diagnosed diseases and prescribed drugs were generally similar in both pandemic phases. At last, our results revealed that with the progress of the pandemic phases, the pharmacist audit time was extended, and the audit approval rate was increased.

Detailed Discussion of the Findings

Interpretations

From the perspective of monthly prescriptions, we found that our internet-based psychiatric hospital exhibited distinct distribution patterns across 2 pandemic phases. Our results from ITS analysis might reveal a generally increasing need for patients in online pharmacy services during the pandemic phase and the ability of internet-based psychiatric hospitals to provide stable pharmacy services after the pandemic [58,59].

Initially, during November 2020-May 2021, the small number of cumulative prescriptions possibly indicated that patients have not yet adapted to the shift from offline to online medical treatment model [60]. Notably, in June 2021, with the major outbreak in Guangzhou, the outpatient services of our hospital were suspended, which led to a dramatic surge in the number of electronic prescriptions [61]. This finding was in line with previous evidence, indicating the public panic in the early stage of the pandemic and the increased demand for patients for internet-based hospitals [21,23]. Subsequently, from July 2021 to October 2022, with the optimization of human resources in our internet-based psychiatric hospital and the gradually reduced public panic [62,63], the average number of monthly prescriptions remained relatively stable at approximately 350. During November 2022-December 2022, with another major outbreak in Guangzhou, together with the subsequent COVID-19 control relaxation in China, the demand for internet-based psychiatric hospitals surged again [64]. In the postpandemic phase, the average number of monthly prescriptions was stable at approximately 460, which was consistent with previous findings, indicating the critical role of internet-based hospitals during and after the COVID-19 pandemic [65]. Importantly, our findings highlighted the capacity and necessity of internet-based psychiatric hospitals in providing long-term stable services after the COVID-19 pandemic [66].

Our results of prescription analysis indicated that female patients were the predominant population in our internet-based psychiatric hospital and were significantly associated with the COVID-19 pandemic phases compared to male patients. Previous studies have shown that female patients are more

XSL•FO
RenderX

susceptible to external environmental stress compared to male patients, thereby resulting in mental health problems [67-70]. Recent evidence has suggested that female patients are more likely to seek help during public health crises, particularly during the COVID-19 pandemic [19,71]. Besides, our findings showed that patients aged 18-40 years were the major population in both pandemic phases, and individuals younger than 41 years of age were significantly associated with the pandemic phases. There were several driving factors that are worth noting. First, these individuals generally encounter significant life crises during the pandemic, such as academic disruptions, risk of unemployment, and financial instability [72]. Second, previous research has shown that the depression rate among adolescents and young adults has risen sharply over the past decade [73]. Fortunately, adolescents and young adults also demonstrated greater adaptability to digital health care platforms compared to the older population, which was helpful in facing public health crises [74,75]. Interestingly, our long-term datasets across pandemic phases appeared to reflect potential demographic differences based on sex and age, which echoed the previous evidence, indicating that the pandemic led to a widening digital divide [76]. Therefore, we suggest that future policy guidance and technical support should be committed to ensure balanced allocation of digital health care resources among different populations [77].

In this study, we found that the leading primary diagnosed diseases included depressive disorder, bipolar disorder, schizophrenia, mood disorder, and anxiety disorder. This disorder spectrum not only aligned with the clinical orientation of our internet-based psychiatric hospital but also suggested that the patient populations analyzed in this study met the intended service objectives of our online platform [23]. Previous studies have reported that sudden public health emergencies can cause various psychological disorders, such as depression, mood disorder, and anxiety [78,79]. During the COVID-19 pandemic, control measures of medical isolation or home quarantine were also reported to cause significant psychological pressure [80]. Our results revealed increased proportions of depressive disorder and mood disorder in the postpandemic phase, which might support the evidence that there is a pandemic-driven shift pattern in global mental health needs [81]. Regarding drug categories, quetiapine, lithium carbonate, and escitalopram were the top 3 drugs prescribed in both pandemic phases, which was consistent with the evidence that these drugs were the most commonly applied antipsychotics and antidepressants [82].

As for the time spent during the pharmacist audit process, it was reported that pharmacist audit time was significantly compressed during the pandemic due to strategic reallocation of hospital human resources [42,83]. In the postpandemic phase, we found a significantly decreased proportion of prescriptions audited in ≤5 minutes and a significantly increased proportion of prescriptions audited within 12-24 hours. Several factors could contribute to these changes. First, the increased complexity of prescriptions for patients infected with COVID-19 could lead to an extended audit time [23]. In addition, evidence has suggested that pharmacists were already at risk of burnout before the COVID-19 pandemic [84]. Subsequently, the increased workload coupled with decreased rest time have exacerbated the burnout, which might influence the overall prescription audit time [85].

Pharmacists are considered to play a crucial role in ensuring the quality of prescription audit [86]. Pharmacist audit offers greater flexibility and professional judgment than a system automatic audit by using their clinical experience in individual conditions of patients [23]. Our results suggested an overall high prescription approval rate in both pandemic phases, revealing the critical role of pharmacist audit in promoting rational medication use in the internet-based psychiatric hospital [87]. Additionally, our results indicated a significantly increased pharmacist approval rate as well as a decreased doctor double confirmation rate in the postpandemic phase. These findings possibly reflected the increased familiarity of pharmacists and doctors with the online platform and the technological optimization of the digital system in the postpandemic phase [88].

Innovation of the Study

To the best of our knowledge, this study is the first work that applied descriptive and analytic statistical methods to evaluate the associations between different phases of the COVID-19 pandemic and the prescriptions and pharmacy services in the internet-based psychiatric hospital. The analyzed indicators included long-term prescription trends, patients' demographic characteristics, drug and disease distribution patterns, and pharmacy service indicators. Our results could provide practical experience for other medical institutions and help to promote the development of digital health service models in the future [89].

Comparisons to Existing Literature

In the field of pharmacy services in internet-based hospitals, most research has focused on the establishment and application of an artificial intelligence audit system [90,91], drug delivery [92-94], medication therapy management [95,96], and drug consultation [97,98]. To date, only a few studies have explored online pharmacy services related to the COVID-19 pandemic through prescription analysis. Among these studies, Chen et al [99] analyzed 1718 prescriptions from the online platform of a tertiary cancer hospital in China over 6 months, involving data of patients' sex and age, geographical distribution, main diagnosis, and drug category. Ding et al [49] analyzed 1380 prescriptions from the online platform of a tertiary general hospital in China over 2 months, covering data of patients' sex and age, residence, prescription department, payment of prescription, and drug delivery region. As for studies in the field of internet-based psychiatric hospitals, Xie et al [23] analyzed 2914 prescriptions over 24 months, and Du et al [21] analyzed 1259 prescriptions over 12 months. Notably, we found that the distribution patterns of patients' sex and age, diagnosed diseases, and prescribed drugs in our study were generally consistent with these 2 studies [21,23], which supported the reliability and generalizability of our results.

However, existing studies conducting prescription analysis only applied descriptive statistical methods rather than analytic statistical methods to compare the results between different

XSL•FO
RenderX

pandemic phases, limiting in-depth interpretation of the observed phenomena [21,23,49,99]. Our study addressed the limitations of existing research from several aspects. First, this study carried a larger sample size compared to previous studies, and the collection of 17,330 prescriptions strengthened the statistical power of the study [100]. Second, we provided a longer research timespan than previous studies, and the dataset over 38 months enhanced the robustness of our results [101]. Finally, by providing certainty of evidence through analytic statistical methods, including ITS analysis, bootstrap method, Pearson chi-square analysis, and multinomial logistic regression, we further explored the potential correlations between the pandemic phases and the observed outcomes [53-56].

Contribution of the Study

The main contribution of this study was to provide a fundamental practice model of telepharmacy service. First, we identified early warning indicators that can be used as key intervention steps in pharmacy services, such as extended audit time and an increased number of prescriptions [102]. Second, by using descriptive and analytic statistical approaches in the analysis of prescriptions and pharmacy services, we demonstrated replicable analytic methods that can be applied in other medical institutions, offering templates for future responses to public health crises [103]. Taken together, this study greatly enriched our understanding of pharmacy services in internet-based hospitals and provided practical experience for future multicenter research in this field [89].

Implications

Our study had direct real-world implications for optimizing the efficiency of pharmacy services in internet-based hospitals. Our data showed that prescription audit time was significantly extended in the postpandemic phase. Therefore, for hospital managers, it is recommended to establish a pharmacy personnel allocation system to timely deploy nonemergency personnel when prescription audit time increases abnormally [104]. In addition, it is important to build a reward mechanism for pharmacy services. Specifically, the efficiency of prescription audit can be promoted by implementing clear service standards and linking them to incentive funds [105]. For software developers, pharmacy service paths can be optimized for key populations. Our results suggested that the majority in internet-based psychiatric hospitals were female patients. Therefore, online psychological counseling tools can be developed for this population [106,107]. On the other hand, our results revealed the lowest proportion of older patients. Hence, in order to improve their user experience, simplified user interface and video chat tools can be added in the app [108]. Besides, we found that many prescriptions required double confirmation by doctors during the pandemic phase. To further improve doctor-pharmacist communication in this step, a real-time videoconferencing module can be added to the audit system [109,110]. Further, a local knowledge base can be built to predict prescriptions that may trigger double confirmation by doctors and convert them into pop-up notifications in the doctor-prescribing system [111].

Limitations

This study had some limitations that need to be noted. First, although our study analyzed the real-world pharmacy service model in an operating online platform of a psychiatric hospital in China, its single-center nature might limit the generalizability of our findings in rural regions or other hospitals, such as general hospitals and rural primary medical institutions [112]. Future research of multicenter design should be required to further validate our conclusions [89]. Second, due to the observational nature of our study, the findings mainly allowed us to demonstrate statistical associations rather than definite proof of causality [113]. It would be helpful for future studies to manipulate more predictors in the model to clarify potential confounding effects [114]. Third, the absence of other unmeasured factors, such as prescription cost, patients' geographic distribution, comorbidities, disease severity, and adverse events, might have influenced the depth of our findings [49,99]. The inclusion of these possible variables in future work would enrich our understanding of the internet-based hospital service models [115].

Conclusions

In summary, this study applied descriptive and analytic statistical methods to evaluate the associations between different COVID-19 phases and the prescriptions and pharmacy services in our internet-based psychiatric hospital. We found an overall upward trend in the number of prescriptions during the pandemic phase and a steady trend after the pandemic. Among patients, female and young adults were the predominant groups, and their proportions increased with the development of pandemic phases. The distribution patterns of primary diagnosed diseases and prescribed drugs were generally similar in both pandemic phases. As the pandemic phases advanced, prescription audit time was extended, and the audit approval rate was increased. This study addressed the limitations of existing research by the application of larger sample size, longer research timespan, and analytic statistical methods. This study demonstrated early warning indicators and replicable analytic methods that can be applied in other medical institutions. Our findings also had implications for hospital managers and software developers in optimizing the efficiency of pharmacy services in internet-based hospitals.

Acknowledgments

No generative artificial intelligence was used in any portion of the manuscript writing.

Data Availability

The datasets used and/or analyzed during this study are included in the paper or Multimedia Appendix 1; further inquiries are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Funding

This work was supported by the National Natural Science Foundation of China (82104223), Guangdong Basic and Applied Basic Research Foundation (2020A1515110008), Science and Technology Program of Guangzhou (202102021022, 2024A04J10001, and 2025A03J3308), Guangzhou Municipal Science and Technology Project for Medicine and Healthcare (20211A011044), and Guangzhou Municipal Key Discipline in Medicine (2025-2027). The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Authors' Contributions

YT and JH contributed to the study concept and design. DS contributed to the acquisition of data. GD and HX contributed to the statistical analysis and visualization of data and drafted the manuscript. YT and JH contributed to the writing and review of the manuscript. YW contributed to the supervision. All authors contributed to the critical revision of the final manuscript and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The completed and filled out STROBE checklist.
[DOCX File , 25 KB - jmir_v28i1e74059_app1.docx]

References

1. Onyeaka H, Anumudu CK, Al-Sharify ZT, Egele-Godswill E, Mbaegbu P. COVID-19 pandemic: a review of the global lockdown and its far-reaching effects. Sci Prog 2021;104(2):368504211019854 [FREE Full text] [doi: 10.1177/00368504211019854] [Medline: 34061685]
2. Panchal U, Salazar de Pablo G, Franco M, Moreno C, Parellada M, Arango C, et al. The impact of COVID-19 lockdown on child and adolescent mental health: systematic review. Eur Child Adolesc Psychiatry 2023;32(7):1151-1177 [FREE Full text] [doi: 10.1007/s00787-021-01856-w] [Medline: 34406494]
3. Zhu C, Zhang T, Li Q, Chen X, Wang K. Depression and anxiety during the COVID-19 pandemic: epidemiology, mechanism, and treatment. Neurosci Bull 2023;39(4):675-684 [FREE Full text] [doi: 10.1007/s12264-022-00970-2] [Medline: 36411394]
4. Alsabeeha NHM, Atieh MA, Balakrishnan MS. Older adults' satisfaction with telemedicine during the COVID-19 pandemic: a systematic review. Telemed J E Health 2023;29(1):38-49. [doi: 10.1089/tmj.2022.0045] [Medline: 35613380]
5. Kowalski K, Misiak B. Schizophrenia and the COVID-19 pandemic: a narrative review from the biomedical perspective. Span J Psychiatry Ment Health 2025;18(2):141-148. [doi: 10.1016/j.rpsm.2023.04.002] [Medline: 37544807]
6. Sunjaya AP, Jenkins C. Rationale for universal face masks in public against COVID-19. Respirology 2020;25(7):678-679 [FREE Full text] [doi: 10.1111/resp.13834] [Medline: 32353901]
7. Gaur U, Sealy W, Bharatha A, Sobers NP, Krishnamurthy K, Campbell MH, et al. Knowledge, attitudes, and practices of hand hygiene, mask use, and social distancing among public hospital and polyclinic nurses in Barbados during the coronavirus 2019 pandemic. Epidemiologia (Basel) 2024;5(1):122-136 [FREE Full text] [doi: 10.3390/epidemiologia5010008] [Medline: 38534805]
8. Li P, Wen L, Sun B, Sun W, Chen H. Retrospective estimation of the time-varying effective reproduction number for a COVID-19 outbreak in Shenyang, China: an observational study. Medicine (Baltimore) 2024;103(22):e38373 [FREE Full text] [doi: 10.1097/MD.0000000000038373] [Medline: 39259088]
9. De Hert M, Mazereel V, Detraux J, Van Assche K. Prioritizing COVID-19 vaccination for people with severe mental illness. World Psychiatry 2021;20(1):54-55 [FREE Full text] [doi: 10.1002/wps.20826] [Medline: 33131217]
10. Ahuvia IL, Sung JY, Dobias ML, Nelson BD, Richmond LL, London B, et al. College student interest in teletherapy and self-guided mental health supports during the COVID-19 pandemic. J Am Coll Health 2024;72(3):940-946. [doi: 10.1080/07448481.2022.2062245] [Medline: 35427460]
11. Smith KA, Ostinelli EG, Ede R, Allard L, Thomson M, Hewitt K, et al. Assessing the impact of evidence-based mental health guidance during the COVID-19 pandemic: systematic review and qualitative evaluation. JMIR Ment Health 2023;10:e52901 [FREE Full text] [doi: 10.2196/52901] [Medline: 38133912]
12. Kola L, Kohrt BA, Hanlon C, Naslund JA, Sikander S, Balaji M, et al. COVID-19 mental health impact and responses in low-income and middle-income countries: reimagining global mental health. Lancet Psychiatry 2021;8(6):535-550 [FREE Full text] [doi: 10.1016/S2215-0366(21)00025-0] [Medline: 33639109]
13. Chai Y, Lam ICH, Man KKC, Hayes JF, Wan EYF, Li X, et al. Psychiatric and neuropsychiatric sequelae of COVID-19 within 2 years: a multinational cohort study. BMC Med 2025;23(1):144 [FREE Full text] [doi: 10.1186/s12916-025-03952-z] [Medline: 40055683]

XSL•FO
RenderX

14. Liu Y, Ren Y, Liu C, Chen X, Li D, Peng J, et al. Global burden of mental disorders in children and adolescents before and during the COVID-19 pandemic: evidence from the Global Burden of Disease Study 2021. Psychol Med 2025;55:e90. [doi: 10.1017/s0033291725000649]
15. Baidoobonso S, Etowa E, Nnadi J, Mba S, Tharao W, Daboné C, et al. The impact of the COVID-19 pandemic on the mental health of African, Caribbean, and Black (ACB) people in Canada. J Immigr Minor Health 2025;27(1):42-52. [doi: 10.1007/s10903-024-01654-x] [Medline: 39612108]
16. Ahmed N, Barnett P, Greenburgh A, Pemovska T, Stefanidou T, Lyons N, et al. Mental health in Europe during the COVID-19 pandemic: a systematic review. Lancet Psychiatry 2023;10(7):537-556 [FREE Full text] [doi: 10.1016/S2215-0366(23)00113-X] [Medline: 37321240]
17. Lee S, Shin HE, Lee R, Jo S, Lee SC, Cheong H, et al. A brief report of the economic burden and epidemiological finding of suicide attempts among the older adults in Korea from 2007 to 2021. BMC Geriatr 2025;25(1):147 [FREE Full text] [doi: 10.1186/s12877-025-05793-0] [Medline: 40045237]
18. Grosman-Dziewiszek P, Wiatrak B, Jęśkowiak I, Szeląg A. Patients' habits and the role of pharmacists and telemedicine as elements of a modern health care system during the COVID-19 pandemic. J Clin Med 2021;10(18):4211 [FREE Full text] [doi: 10.3390/jcm10184211] [Medline: 34575322]
19. Lau BHP, Tang CSK, Holroyd E, Wong WCW. Challenges and implications for menopausal health and help-seeking behaviors in midlife women from the United States and China in light of the COVID-19 pandemic: web-based panel surveys. JMIR Public Health Surveill 2024;10:e46538 [FREE Full text] [doi: 10.2196/46538] [Medline: 38277194]
20. Weinberg V, van Haren E, Gombert-Handoko KB. Improving patient access to hospital pharmacists using eConsults: retrospective descriptive study. JMIR Form Res 2023;7:e38518 [FREE Full text] [doi: 10.2196/38518] [Medline: 36705957]
21. Du Z, Jiang Y, Lu R, Zhou Q, Pan Y, Shen Y, et al. Practice of pharmaceutical services and prescription analysis in internet-based psychiatric hospitals during COVID-19 pandemic in Wuxi, China. Front Psychiatry 2023;14:1195298 [FREE Full text] [doi: 10.3389/fpsyt.2023.1195298] [Medline: 37547208]
22. Ge F, Qian H, Lei J, Ni Y, Li Q, Wang S, et al. Experiences and challenges of emerging online health services combating COVID-19 in China: retrospective, cross-sectional study of internet hospitals. JMIR Med Inform 2022;10(6):e37042 [FREE Full text] [doi: 10.2196/37042] [Medline: 35500013]
23. Xie W, Wang F, Qian Y, Qiu L, Zhou Q, Shen Y, et al. "Internet+" pharmacy in psychiatric hospital amid COVID-19 and post-pandemic period: analysis and development. Front Psychiatry 2024;15:1434966 [FREE Full text] [doi: 10.3389/fpsyt.2024.1434966] [Medline: 39758438]
24. Cheng TC, Yip W. Policies, progress, and prospects for internet telemedicine in China. Health Syst Reform 2024;10(2):2389570 [FREE Full text] [doi: 10.1080/23288604.2024.2389570] [Medline: 39437237]
25. Wang K, Zou W, Lai Y, Hao C, Liu N, Ling X, et al. Accessibility, cost, and quality of an online regular follow-up visit service at an internet hospital in China: mixed methods study. J Med Internet Res 2024;26:e54902 [FREE Full text] [doi: 10.2196/54902] [Medline: 39432365]
26. Bilionis I, Quer Palomas S, Vidal-Alaball J. Digital health for tracking long-covid symptoms with data insights from mobile app questionnaires and wearable devices. Stud Health Technol Inform 2025;323:434-438. [doi: 10.3233/SHTI250127] [Medline: 40200524]
27. Zhi L, Yin P, Ren J, Wei G, Zhou J, Wu J, et al. Running an internet hospital in China: perspective based on a case study. J Med Internet Res 2021;23(9):e18307 [FREE Full text] [doi: 10.2196/18307] [Medline: 34342267]
28. Fu L, Wang R, Dong Y. The impact of the hierarchical medical system on medical resource allocation in China. Sci Rep 2025;15(1):7561 [FREE Full text] [doi: 10.1038/s41598-025-88558-4] [Medline: 40038294]
29. Qingjie P, Wanyu H, Jingying Y, Ming Y, Linlin W, Qiyue M, et al. How to promote telemedicine in underdeveloped cities in central China? Qualitative interviews with medical personnel in tertiary A-level hospital of Xinyang City. Digit Health 2024;10:20552076241277670 [FREE Full text] [doi: 10.1177/20552076241277670] [Medline: 39247099]
30. Gao J, Fan C, Chen B, Fan Z, Li L, Wang L, et al. Telemedicine is becoming an increasingly popular way to resolve the unequal distribution of healthcare resources: evidence from China. Front Public Health 2022;10:916303 [FREE Full text] [doi: 10.3389/fpubh.2022.916303] [Medline: 35874991]
31. Yu J, Cheong IH, Kozlakidis Z, Wang H. Advancements and challenges of artificial intelligence in dermatology: a review of applications and perspectives in China. Front Digit Health 2025;7:1544520 [FREE Full text] [doi: 10.3389/fdgth.2025.1544520] [Medline: 40881732]
32. Ndwabe H, Basu A, Mohammed J. Post pandemic analysis on comprehensive utilization of telehealth and telemedicine. Clinical eHealth 2024;7:5-14. [doi: 10.1016/j.ceh.2023.12.002]
33. Kim J, Cai ZR, Chen ML, Onyeka S, Ko JM, Linos E. Telehealth utilization and associations in the United States during the third year of the COVID-19 pandemic: population-based survey study in 2022. JMIR Public Health Surveill 2024;10:e51279 [FREE Full text] [doi: 10.2196/51279] [Medline: 38669075]
34. Cao Y, Li Q, Chen J, Guo X, Miao C, Yang H, et al. Hospital emergency management plan during the COVID-19 epidemic. Acad Emerg Med 2020;27(4):309-311 [FREE Full text] [doi: 10.1111/acem.13951] [Medline: 32124506]

XSL•FO
RenderX

35. Juang WC, Chiou SM, Chen HC, Li YC. Differences in characteristics and length of stay of elderly emergency patients before and after the outbreak of COVID-19. Int J Environ Res Public Health 2023;20(2):1162 [FREE Full text] [doi: 10.3390/ijerph20021162] [Medline: 36673917]
36. Han S, Li L. Consulting doctors online after offline treatment: investigating the effects of online information on patients' effective use of online follow-up services. Front Public Health 2024;12:1375144 [FREE Full text] [doi: 10.3389/fpubh.2024.1375144] [Medline: 38655527]
37. Bates DW, Cheng HY, Cheung NT, Jew R, Mir F, Tamblyn R, et al. 'Improving smart medication management': an online expert discussion. BMJ Health Care Inform 2022;29(1):e100540 [FREE Full text] [doi: 10.1136/bmjhci-2021-100540] [Medline: 35477691]
38. Zeng B, Rivadeneira NA, Wen A, Sarkar U, Khoong EC. The impact of the COVID-19 pandemic on internet use and the use of digital health tools: secondary analysis of the 2020 Health Information National Trends Survey. J Med Internet Res 2022;24(9):e35828 [FREE Full text] [doi: 10.2196/35828] [Medline: 36041005]
39. Lamba S, Li S, Lamba A, Tang F, Valencia WM, Dang S. Differences in challenges to using telehealth among older adult video and telephone users with frailty: retrospective observational study. J Med Internet Res 2025;27:e69437-e69437 [FREE Full text] [doi: 10.2196/69437] [Medline: 40882213]
40. Faro JM, Obermiller E, Obermiller C, Trinkley KE, Wright G, Sadasivam RS, et al. Using routinely available electronic health record data elements to develop and validate a digital divide risk score. JAMIA Open 2025;8(1):ooaf004. [doi: 10.1093/jamiaopen/ooaf004] [Medline: 39906363]
41. Jiménez-Zarco A, Mateos SC, Bosque-Prous M, Espelt A, Torrent-Sellens J, Adib K, et al. Impact of the COVID-19 pandemic on mHealth adoption: identification of the main barriers through an international comparative analysis. Int J Med Inform 2025;195:105779 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105779] [Medline: 39813967]
42. Li Y, Hu H, Rozanova L, Fabre G. COVID-19 and internet hospital development in China. Epidemiologia (Basel) 2022;3(2):269-284 [FREE Full text] [doi: 10.3390/epidemiologia3020021] [Medline: 36417257]
43. Tran J, Hao Z, Xie Z, Yiu K, Li A, Tsoi K. Nationwide trends of online consultations in China: a 10-year sequential cross-sectional study of 65 305 clinicians. BMJ Public Health 2025;3(2):e002296. [doi: 10.1136/bmjph-2024-002296] [Medline: 40697407]
44. Wu X, Kuang Y, Guo Y, Wu J, Xiao L. Internet hospital response to the COVID-19 pandemic in a tertiary hospital in China: perspectives based on a mixed-methods. Digit Health 2024;10:20552076241228418 [FREE Full text] [doi: 10.1177/20552076241228418] [Medline: 38303968]
45. Hagi K, Kurokawa S, Takamiya A, Fujikawa M, Kinoshita S, Iizuka M, et al. Telepsychiatry versus face-to-face treatment: systematic review and meta-analysis of randomised controlled trials. Br J Psychiatry 2023;223(3):407-414 [FREE Full text] [doi: 10.1192/bjp.2023.86] [Medline: 37655816]
46. Ambrósio GC, Dal Paz K, Moreira de Souza VA, Discacciatii MG, Campa A. Impact of the COVID-19 pandemic on pharmacist interventions: a retrospective study with inpatients in a university hospital. Can J Hosp Pharm 2024;77(3):e3514. [doi: 10.4212/cjhp.3514] [Medline: 38988873]
47. Yu T, Li W, Liu Y, Jin C, Wang Z, Cao H. Application of internet hospitals in the disease management of patients with ulcerative colitis: retrospective study. J Med Internet Res 2025;27:e60019 [FREE Full text] [doi: 10.2196/60019] [Medline: 40101745]
48. Wang L, Liang D, Huang Fu H, Ke C, Wu S, Lai Y. Enterprise-led internet healthcare provision in China: insights from a leading platform. Front Digit Health 2025;7:1491183 [FREE Full text] [doi: 10.3389/fdgth.2025.1491183] [Medline: 40099033]
49. Ding L, She Q, Chen F, Chen Z, Jiang M, Huang H, et al. The internet hospital plus drug delivery platform for health management during the COVID-19 pandemic: observational study. J Med Internet Res 2020;22(8):e19678 [FREE Full text] [doi: 10.2196/19678] [Medline: 32716892]
50. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med 2007;147(8):573-577 [FREE Full text] [doi: 10.7326/0003-4819-147-8-200710160-00010] [Medline: 17938396]
51. Levitt HM, Bamberg M, Creswell JW, Frost DM, Josselson R, Suárez-Orozco C. Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018;73(1):26-46 [FREE Full text] [doi: 10.1037/amp0000151] [Medline: 29345485]
52. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
53. Antonazzo IC, Fornari C, Maumus-Robert S, Cei E, Paoletti O, Ferrara P, et al. Antidepressants drug use during COVID-19 waves in the Tuscan general population: an interrupted time-series analysis. J Pers Med 2022;12(2):1639-1641 [FREE Full text] [doi: 10.3390/jpm12020178] [Medline: 35207666]

XSL•FO
RenderX

54. Chan VKY, Chai Y, Chan SSM, Luo H, Jit M, Knapp M, et al. Impact of COVID-19 pandemic on depression incidence and healthcare service use among patients with depression: an interrupted time-series analysis from a 9-year population-based study. BMC Med 2024;22(1):169 [FREE Full text] [doi: 10.1186/s12916-024-03386-z] [Medline: 38644506]
55. Xia H, Deng G, Liang F, Zhang Z, Huang W, Guo Z, et al. Investigating remedial strategies for missed or delayed dose of sertraline in Chinese adolescent patients with depressive disorders via population pharmacokinetics modeling and simulation approaches. Drug Des Devel Ther 2025;19:3001-3016 [FREE Full text] [doi: 10.2147/DDDT.S504521] [Medline: 40260198]
56. Tsamardinos I, Greasidou E, Borboudakis G. Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. Mach Learn 2018;107(12):1895-1922 [FREE Full text] [doi: 10.1007/s10994-018-5714-4] [Medline: 30393425]
57. Wen Q, Yang B, Wang X, Xia L, Song L. Characteristics analysis of internet pharmacy consultation services for children in southwest China during the post-epidemic era: a cross-sectional study. Int J Med Inform 2024;186:105424 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105424] [Medline: 38547627]
58. Karattuthodi MS, Krishnan G, Aiswarya A, Sarath CC, Nair SS, Thorakkatil SA. A systematic review on pharmacist-led pharmaceutical care optimization through virtual platforms. J Pharm Health Serv Res 2023;14(1):2-13. [doi: 10.1093/jphsr/rmac054]
59. Hursman A, Vang C, Thooft T, Stone K. The role of telepharmacy in the delivery of clinical pharmacy services following the COVID-19 pandemic: a descriptive report. J Pharm Technol 2024;40(2):66-71. [doi: 10.1177/87551225231222426] [Medline: 38525089]
60. Lu J, Bai J, Zhao H, Zhang X. The effect of "offline-to-online" trust transfer on the utilization of online medical consultation among Chinese rural residents: experimental study. J Med Internet Res 2023;25:e43430 [FREE Full text] [doi: 10.2196/43430] [Medline: 38147370]
61. Zhang M, Xiao J, Deng A, Zhang Y, Zhuang Y, Hu T, et al. Transmission dynamics of an outbreak of the COVID-19 Delta Variant B.1.617.2—Guangdong Province, China, May-June 2021. China CDC Wkly 2021;3(27):584-586 [FREE Full text] [doi: 10.46234/ccdcw2021.148] [Medline: 34594941]
62. Chen Y, Liu Y, Yan Y. Revealing the spatiotemporal characteristics of the general public's panic levels during the pandemic crisis in China. Trans GIS 2022:tgis.13016 [FREE Full text] [doi: 10.1111/tgis.13016] [Medline: 36721464]
63. Yang Y, Zhang Y, Zhang X, Cao Y, Zhang J. Spatial evolution patterns of public panic on Chinese social networks amidst the COVID-19 pandemic. Int J Disaster Risk Reduct 2022;70:102762 [FREE Full text] [doi: 10.1016/j.ijdrr.2021.102762] [Medline: 35004139]
64. Huang J, Zhao S, Chong KC, Zhou Y, Lu W, Fang F, et al. Infection rate in guangzhou after easing the zero-COVID policy: seroprevalence results to ORF8 antigen. Lancet Infect Dis 2023;23(4):403-404 [FREE Full text] [doi: 10.1016/s1473-3099(23)00112-3]
65. Di Carlo F, Sociali A, Picutti E, Pettorruso M, Vellante F, Verrastro V, et al. Telepsychiatry and other cutting-edge technologies in COVID-19 pandemic: bridging the distance in mental health assistance. Int J Clin Pract 2021;75(1):10.1111/ijcp.13716 [FREE Full text] [doi: 10.1111/ijcp.13716] [Medline: 32946641]
66. Pan X, Zhou X, Yu L, Hou L. Switching from offline to online health consultation in the post-pandemic era: the role of perceived pandemic risk. Front Public Health 2023;11:1121290 [FREE Full text] [doi: 10.3389/fpubh.2023.1121290] [Medline: 37261233]
67. Min C, Shen F, Yu W, Chu Y. The relationship between government trust and preventive behaviors during the COVID-19 pandemic in China: exploring the roles of knowledge and negative emotion. Prev Med 2020;141:106288 [FREE Full text] [doi: 10.1016/j.ypmed.2020.106288] [Medline: 33091414]
68. Rossell SL, Neill E, Phillipou A, Tan EJ, Toh WL, Van Rheenen TE, et al. An overview of current mental health in the general population of Australia during the COVID-19 pandemic: results from the COLLATE project. Psychiatry Res 2021;296:113660 [FREE Full text] [doi: 10.1016/j.psychres.2020.113660] [Medline: 33373808]
69. Duan L, Shao X, Wang Y, Huang Y, Miao J, Yang X, et al. An investigation of mental health status of children and adolescents in China during the outbreak of COVID-19. J Affect Disord 2020;275:112-118 [FREE Full text] [doi: 10.1016/j.jad.2020.06.029] [Medline: 32658812]
70. Liu S, Yang L, Zhang C, Xu Y, Cai L, Ma S, et al. Gender differences in mental health problems of healthcare workers during the coronavirus disease 2019 outbreak. J Psychiatr Res 2021;137:393-400 [FREE Full text] [doi: 10.1016/j.jpsychires.2021.03.014] [Medline: 33765451]
71. Wu X, Wong KW, Gurvich C, Dong Y. Impact of COVID-19 on health of menopausal women: a scoping review. Gen Hosp Psychiatry 2023;84:125-141. [doi: 10.1016/j.genhosppsych.2023.07.001] [Medline: 37473503]
72. Cook JA, Jonikas JA, Burke-Miller JK, Aranda F, Mullen MG, Davis M, et al. Social determinants, mental well-being, and disrupted life transitions among young adults with disabling mental health conditions. J Behav Health Serv Res 2025;52(2):263-281. [doi: 10.1007/s11414-024-09924-0] [Medline: 39806229]
73. Thapar A, Eyre O, Patel V, Brent D. Depression in young people. Lancet 2022;400(10352):617-631. [doi: 10.1016/S0140-6736(22)01012-1] [Medline: 35940184]

XSL•FO
RenderX

74. Schultz AN, Honnens S, Bauer EH, Söderberg K, Kidholm K, Christensen R, et al. Patients' preferences for virtual outpatient consultations and health care professionals' assessment of suitability for their patients: a single-centre survey. J Telemed Telecare 2024;10:1357633x241294125. [doi: 10.1177/1357633x241294125]
75. Hamilton JL, Nesi J, Choukas-Bradley S. Reexamining social media and socioemotional well-being among adolescents through the lens of the COVID-19 pandemic: a theoretical review and directions for future research. Perspect Psychol Sci 2022;17(3):662-679 [FREE Full text] [doi: 10.1177/17456916211014189] [Medline: 34756118]
76. Qiu Y, Huang H, Gai J, De Leo G. The effects of the COVID-19 pandemic on age-based disparities in digital health technology use: secondary analysis of the 2017-2022 Health Information National Trends Survey. J Med Internet Res 2024;26:e65541 [FREE Full text] [doi: 10.2196/65541] [Medline: 39631070]
77. Tay CT, Joham AE, Teede HJ. Key standards and principles for developing evidence-based clinical guidelines: balancing health professional, patient, funder, and government needs. Fertil Steril 2025;123(4):561-568. [doi: 10.1016/j.fertnstert.2025.01.023] [Medline: 39884335]
78. Tng XJJ, Chew QH, Sim K. Psychological sequelae within different populations during the COVID-19 pandemic: a rapid review of extant evidence. Singapore Med J 2022;63(5):229-235 [FREE Full text] [doi: 10.11622/smedj.2020111] [Medline: 36043286]
79. Rogers M, Windle A, Wu L, Taylor V, Bale C. Emotional well-being, spiritual well-being and resilience of advanced clinical practitioners in the United Kingdom during COVID-19: an exploratory mixed method study. J Nurs Manag 2022;30(4):883-891 [FREE Full text] [doi: 10.1111/jonm.13577] [Medline: 35261097]
80. Lestari R, Setyawan FEB. Mental health policy: protecting community mental health during the COVID-19 pandemic. J Public Health Res 2021;10(2):2231 [FREE Full text] [doi: 10.4081/jphr.2021.2231] [Medline: 33855426]
81. Xiao Y, Chen TT, Liu L, Zong L. China ends its zero-COVID-19 policy: new challenges facing mental health services. Asian J Psychiatr 2023;82:103485. [doi: 10.1016/j.ajp.2023.103485] [Medline: 36731222]
82. Solmi M, Miola A, Capone F, Pallottino S, Højlund M, Firth J, et al. Risk factors, prevention and treatment of weight gain associated with the use of antidepressants and antipsychotics: a state-of-the-art clinical review. Expert Opin Drug Saf 2024;23(10):1249-1269. [doi: 10.1080/14740338.2024.2396396] [Medline: 39225182]
83. Dettori M, Castiglia P. COVID-19 and digital health: evolution, perspectives and opportunities. Int J Environ Res Public Health 2022;19(14):8519 [FREE Full text] [doi: 10.3390/ijerph19148519] [Medline: 35886366]
84. Durham ME, Bush PW, Ball AM. Evidence of burnout in health-system pharmacists. Am J Health Syst Pharm 2018;75(23 Supplement 4):S93-S100. [doi: 10.2146/ajhp170818] [Medline: 30333113]
85. Abdelsadig Mohammed H, Elamin SA, El-Awaisi A, El Hajj MS. Use of the job demands-resource model to understand community pharmacists' burnout during the COVID-19 pandemic. Res Social Adm Pharm 2022;18(9):3568-3579 [FREE Full text] [doi: 10.1016/j.sapharm.2022.03.011] [Medline: 35418325]
86. De Vries E, Johnson Y, Willems B, Bedeker W, Ras T, Coetzee R, et al. Improving primary care antimicrobial stewardship by implementing a peer audit and feedback intervention in Cape Town community healthcare centres. S Afr Med J 2022;112(10):812-818 [FREE Full text] [doi: 10.7196/SAMJ.2022.v112i10.16397] [Medline: 36472332]
87. Mujtaba SH, Gazerani P. Exploring the role of community pharmacists in pain management: enablers and challenges. Pharmacy (Basel) 2024;12(4):111 [FREE Full text] [doi: 10.3390/pharmacy12040111] [Medline: 39051395]
88. Patrocínio WP, Pimenta MA. Information and communication technology: strategies to optimize hospital services. Rev Gest Soc Ambient 2025;19(1):1-17. [doi: 10.24857/rgsa.v19n1-021]
89. Das MK. Multicenter studies: relevance, design and implementation. Indian Pediatr 2022;59(7):571-579. [doi: 10.1007/s13312-022-2561-y]
90. Zheng Y, Rowell B, Chen Q, Kim JY, Kontar RA, Yang XJ, et al. Designing human-centered AI to prevent medication dispensing errors: focus group study with pharmacists. JMIR Form Res 2023;7:e51921 [FREE Full text] [doi: 10.2196/51921] [Medline: 38145475]
91. Fourkiotis KP, Tsadiras A. Future internet applications in healthcare: big data-driven fraud detection with machine learning. Future Internet 2025;17(10):460. [doi: 10.3390/fi17100460]
92. Sham R, Siau CS, Tan S, Kiu DC, Sabhi H, Thew HZ, et al. Drone usage for medicine and vaccine delivery during the COVID-19 pandemic: attitude of health care workers in rural medical centres. Drones 2022;6(5):109. [doi: 10.3390/drones6050109]
93. Mukati N, Namdev N, Dilip R, Hemalatha N, Dhiman V, Sahu B. Healthcare assistance to COVID-19 patient using internet of things (IoT) enabled technologies. Mater Today Proc 2023;80:3777-3781 [FREE Full text] [doi: 10.1016/j.matpr.2021.07.379] [Medline: 34336599]
94. Hammour KA, Abdeljalil M, Manaseer Q, Al-Manaseer B. Jordanian experience: the internet pharmacy drug delivery platform during the COVID-19. Health Policy Technol 2022;11(1):100596 [FREE Full text] [doi: 10.1016/j.hlpt.2022.100596] [Medline: 35024326]
95. Zhang P, Yang M, He S, Li X, Lang B, Zeng L, et al. Evaluating the impact of the COVID-19 pandemic on telepharmaceutical service effectiveness: systematic review and meta-analysis. J Med Internet Res 2025;27:e64073 [FREE Full text] [doi: 10.2196/64073] [Medline: 40601374]

XSL•FO
RenderX

96. Zhao X, Huang H, Zeng G, Shi Q, Zhu P, Zhang L, et al. Research on the online service mechanism of internet hospital in infectious disease prevention and control. Exp Biol Med (Maywood) 2025;250:10349 [FREE Full text] [doi: 10.3389/ebm.2025.10349] [Medline: 40351479]
97. Bu F, Sun H, Li L, Tang F, Zhang X, Yan J, et al. Artificial intelligence-based internet hospital pharmacy services in China: perspective based on a case study. Front Pharmacol 2022;13:1027808 [FREE Full text] [doi: 10.3389/fphar.2022.1027808] [Medline: 36438784]
98. Li H, Zheng S, Li D, Jiang D, Liu F, Guo W, et al. The establishment and practice of pharmacy care service based on internet social media: telemedicine in response to the COVID-19 pandemic. Front Pharmacol 2021;12:707442 [FREE Full text] [doi: 10.3389/fphar.2021.707442] [Medline: 34658854]
99. Chen ZJ, Liang W, Liu QT, He R, Chen QC, Li QF, et al. Use of a remote oncology pharmacy service platform for patients with cancer during the COVID-19 pandemic: implementation and user acceptance evaluation. J Med Internet Res 2021;23(1):e24619 [FREE Full text] [doi: 10.2196/24619] [Medline: 33395398]
100. Ahmed SK. How to choose a sampling technique and determine sample size for research: a simplified guide for researchers. Oral Oncol Rep 2024;12:100662 [FREE Full text] [doi: 10.1016/j.oor.2024.100662]
101. Li X, Hao J, Han Q, Wang D, Lu Y, Tu J, et al. Triglyceride-glucose index prediction of stroke incidence risk in low-income Chinese population: a 10-year prospective cohort study. Front Endocrinol (Lausanne) 2024;15:1444030 [FREE Full text] [doi: 10.3389/fendo.2024.1444030] [Medline: 39493774]
102. Baptista R, Williams M, Price J. Improving the impact of pharmacy interventions in hospitals. BMJ Open Qual 2023;12(4):e002276 [FREE Full text] [doi: 10.1136/bmjoq-2023-002276] [Medline: 37940333]
103. Hazari A. Data analysis: descriptive and analytical statistics. In: Research Methodology for Allied Health Professionals: A Comprehensive Guide to Thesis & Dissertation. Cham: Springer; 2024:79-98.
104. Hammad M, Tashtoush MM, El Faki EMA, Hajaj MY, Ahmed SS, Darwish AA. Hospital pharmacy response to the COVID-19 pandemic: experience from a regional referral center in the United Arab Emirates. J Pharm Policy Pract 2023;16(1):36 [FREE Full text] [doi: 10.1186/s40545-023-00542-3] [Medline: 36864489]
105. Hussain R, Babar ZUD. Global landscape of community pharmacy services remuneration: a narrative synthesis of the literature. J Pharm Policy Pract 2023;16(1):118 [FREE Full text] [doi: 10.1186/s40545-023-00626-0] [Medline: 37814349]
106. Andreu‐Pejó L, Martínez‐Borba V, Osma López J, Suso‐Ribera C, Crespo Delgado E. Perinatal mental e‐health: what is the profile of pregnant women interested in online assessment of their emotional state? Nursing Open 2022;10(2):901-914. [doi: 10.1002/nop2.1358] [Medline: 36068679]
107. Progga FT, Rubya S. Women's perspectives and challenges in adopting perinatal mental health technologies. Proc ACM Hum-Comput Interact 2025;9(1):1-30. [doi: 10.1145/3701217]
108. Kelly RM, Xing Y, Baker S, Waycott J. Video calls as a replacement for family visits during lockdowns in aged care: interview study with family members. JMIR Aging 2023;6:e40953 [FREE Full text] [doi: 10.2196/40953] [Medline: 37191951]
109. Allison A, Shahan J, Goodner J, Smith L, Sweet C. Providing essential clinical pharmacy services during a pandemic: virtual video rounding and precepting. Am J Health Syst Pharm 2021;78(17):1556-1558 [FREE Full text] [doi: 10.1093/ajhp/zxab208] [Medline: 33987637]
110. Sarasmita MA, Sudarma IW, Jaya MKA, Irham LM, Susanty S. Telepharmacy implementation to support pharmaceutical care services during the COVID-19 pandemic: a scoping review. Can J Hosp Pharm 2024;77(1):e3430. [doi: 10.4212/cjhp.3430] [Medline: 38204502]
111. Yue XH, Yang L, Zhong JJ, Liu HM, Wang D, Tao X, et al. Optimization and impact of an evidence-based pre-audit prescription decision system in primary healthcare settings. Front Pharmacol 2025;16:1491810 [FREE Full text] [doi: 10.3389/fphar.2025.1491810] [Medline: 40297139]
112. Kim YR, Chung SW, Kim M, Choi W, Choi J, Lee D, et al. Limited generalizability of retrospective single-center cohort study in comparison to multicenter cohort study on prognosis of hepatocellular carcinoma. J Hepatocell Carcinoma 2024;11:1235-1249. [doi: 10.2147/JHC.S456093] [Medline: 38974017]
113. Haber NA, Wieten SE, Rohrer JM, Arah OA, Tennant PWG, Stuart EA, et al. Causal and associational language in observational health research: a systematic evaluation. Am J Epidemiol 2022;191(12):2084-2097 [FREE Full text] [doi: 10.1093/aje/kwac137] [Medline: 35925053]
114. Sendor R, Stürmer T. Core concepts in pharmacoepidemiology: confounding by indication and the role of active comparators. Pharmacoepidemiol Drug Saf 2022;31(3):261-269 [FREE Full text] [doi: 10.1002/pds.5407] [Medline: 35019190]
115. Conti V, Sellitto C, Torsiello M, Manzo V, De Bellis E, Stefanelli B, et al. Identification of drug interaction adverse events in patients with COVID-19: a systematic review. JAMA Netw Open 2022;5(4):e227970 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.7970] [Medline: 35438752]

Abbreviations

IRB: institutional review board
ITS: interrupted time series

XSL•FO
RenderX

JARS: Journal Article Reporting Standards
OR: odds ratio
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by S Brini; submitted 18.Mar.2025; peer-reviewed by N Liu, Ş Demiralay; comments to author 24.Sep.2025; accepted 06.Feb.2026; published 04.Mar.2026.

Please cite as:
Deng G, Xia H, Shang DW, Wen Y, Hu J, Tan Y
Comparative Analysis of Prescriptions and Pharmacy Services in Internet-Based Psychiatric Hospital During and After the COVID-19 Pandemic: Retrospective Cross-Sectional Observational Study
J Med Internet Res 2026;28:e74059
URL: https://www.jmir.org/2026/1/e74059
doi:10.2196/74059
PMID:

©Guowei Deng, Hui Xia, De-wei Shang, Yuguan Wen, Jinqing Hu, Yaqian Tan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Skill-Based Virtual Reality Therapy in High-Impact Chronic Pain: 2-Year Follow-Up Results From a Secondary Analysis of a Randomized Controlled Trial

Todd Maddox[1], PhD; Josh Sackman[1], MBA; Robert Bonakdar[2], MD; Roselani Maddox[1], BS; Beth D Darnall[3], PhD

[1]AppliedVR, Van Nuys, CA, United States
[2]Scripps Health, La Jolla, CA, United States
[3]Stanford Medicine, Palo Alto, CA, United States

Corresponding Author:
Todd Maddox, PhD
AppliedVR
16760 Stagg St
Van Nuys, CA
United States
Phone: 1 844 857 0010
Email: tmaddox@appliedvr.io

Abstract

Background: High-impact chronic pain (HICP) involves substantial interference in functioning, affects 8.5% of the population, and leads to higher health care costs relative to low-impact chronic pain (LICP). Behavioral interventions such as virtual reality (VR) offer scalable and accessible treatment, but testing is needed to ensure durable effectiveness in HICP. We conducted a secondary analysis of the largest real-world dataset for a therapeutic skill-based VR vs a sham VR control to test treatment efficacy in HICP vs LICP. Relative to LICP, we found significantly larger (and clinically meaningful; ie, ≥2 points) pain interference and pain intensity reductions for HICP at end of treatment and 1 year posttreatment. End-of-treatment reduction in pain interference reclassified 70% (114/163) of participants with HICP as LICP, and this improvement held at 1 year posttreatment (104/155, 67%).

Objective: This study examined the effectiveness of a 56-session skill-based VR therapy in HICP at 2 years posttreatment and compared the effects with those on LICP.

Methods: We conducted a secondary analysis of the skill-based VR sample (536/1067, 50.2%) at 2 years posttreatment from a randomized controlled trial involving an in-home chronic low back pain sample that was recruited and tested online and was diverse (female: 411/536, 77%; non-White individuals: 166/536, 31%; high school or lower educational level: 102/536, 19%; mean age 50.8 years) and had clinically severe pain at baseline (intensity=6.6; interference=6.2; 42% with severe or complete disability). Focusing on the skill-based VR participants (536/1067, 50.2%) and using a validated approach, we classified participants at baseline as either HICP (baseline Brief Pain Inventory pain interference score>7) or LICP (baseline Brief Pain Inventory pain interference score<7). Clinical effectiveness was examined using a general linear model at 2 years posttreatment relative to baseline with the primary outcomes of pain interference and pain intensity.

Results: Participants with HICP (192/536, 35.8%) reported superior reductions in pain interference, pain intensity, sleep disturbance, and physical disability (P<.001 in all cases) at 2 years posttreatment compared to participants with LICP (344/536, 64.2%). Participants with HICP had clinically meaningful (≥2-point) reductions in pain interference (mean 3.1, 95% CI 2.66-3.54; effect size=1.12) and pain intensity (mean 2.6, 95% CI 2.18-3.02; effect size=1.01) at 2 years posttreatment. Importantly, reduced pain interference scores at 2 years posttreatment reclassified 71.1% (106/149) of the participants with HICP as LICP. No serious adverse events or side effects were reported.

Conclusions: Patients with HICP experience severe pain that drives high health care use. The skill-based VR therapy demonstrated durable reductions in pain and related outcomes 2 years posttreatment, with the largest benefits observed in the HICP subgroup. These results suggest that a skill-based, VR-delivered therapy produces durable effects in patients with HICP, a population that is frequently overmedicalized and undertreated with behavioral interventions. These findings suggest that home-based VR-delivered therapy offers a scalable treatment option for this underserved population.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e90688) doi:10.2196/90688

KEYWORDS

high-impact chronic pain; HICP; low-impact chronic pain; LICP; virtual reality; VR; digital health

Introduction

Background and Rationale

High-impact chronic pain (HICP) is defined as pain lasting at least 3 months that also causes a major limitation in daily functioning—such as the inability to attend school, work outside the home, or complete household tasks [1-4]. In 2023, nearly 25% of adults reported experiencing chronic pain within the previous 3 months, and 8.5% reported high levels of chronic pain that frequently restricted their ability to engage in work or life activities (classified as HICP) [3,5]. Importantly, the prevalence of HICP grew from 7.4% in 2019 to 8.5% in 2023 and continues to grow on an annual basis [5]. In fact, a recent 2025 publication [6] suggested that factors such as long COVID, sedentary lifestyles, and reduced health care access during and after the COVID-19 pandemic have contributed to the record high and growing levels of HICP. Critically, HICP, relative to low-impact chronic pain (LICP), has increased health care use and costs. Patients with HICP are more likely to receive opioid prescriptions at 5 times the morphine-equivalent daily dose and are 3 times more likely to undergo surgery and have emergency department visits [3,7-10]. Recognizing these burdens, the US Department of Health and Human Services highlighted the need to prioritize HICP in the 2016 National Pain Strategy [11]. Given the increased prevalence and increased pain treatment and health care use patterns, patients with HICP are ideal candidates for low-risk and accessible in-home interventions such as virtual reality (VR)–delivered therapy, but it is important to verify strong treatment effectiveness and patient engagement with these immersive therapeutics in HICP. In-person therapies such as cognitive behavioral therapy are moderately effective for HICP [12,13], but to date, only 1 published study has examined treatment effectiveness and patient engagement with an immersive therapeutic [14].

In a recent large-sample randomized controlled trial in chronic lower back pain [15,16], we found durable efficacy and statistical superiority of a Food and Drug Administration–authorized in-home 56-session skill-based VR therapy relative to a 56-session active sham VR control at end of treatment and 1 year posttreatment. With durable efficacy and superiority to the sham control established, we conducted a secondary analysis [14] of the skill-based VR group and found significantly larger (and clinically meaningful; ie, ≥2-point reductions [17,18]) in pain interference and pain intensity (primary end points) for participants with HICP than for participants with LICP at end of treatment and 1 year posttreatment. Importantly, end-of-treatment reduction in pain interference among patients with HICP reclassified 70% of them as LICP, and this improvement held at 1 year posttreatment (67%). Significantly larger reductions in sleep disturbance and physical disability (secondary end points) were found for HICP compared to LICP. No differences were found for HICP vs LICP in device engagement or device usability, both of which were high.

Objectives

The aim of this report was to conduct the first-ever examination of long-term effectiveness of the skill-based VR-delivered therapy in patients with HICP at 2 years posttreatment. As in previous work [14], we used the validated classification approach for the Brief Pain Inventory (BPI) pain interference score [19,20] of 7 or more to classify participants as either HICP or LICP at baseline and at 2 years posttreatment to determine individual change in classification category [21]. Specifically, within the skill-based VR group, we compared treatment response for HICP vs LICP at 2 years posttreatment relative to baseline. We hypothesized that the skill-based VR-delivered therapy would lead to strong and durable pain reductions (primary end points), as well as durable reductions in sleep disturbance and physical disability (secondary end points) in patients with HICP.

Methods

Patient and Public Involvement

There was no patient or public involvement in the design, conduct, or reporting of this secondary analysis. The primary trial from which these data were derived was designed and conducted by the study investigators.

Trial Design, Trial Setting, and Participant Eligibility

The methods, CONSORT (Consolidated Standards of Reporting Trials) diagram, and analyses at end of treatment and 1 year posttreatment for the primary decentralized randomized controlled trial are published elsewhere [15,16]. This secondary analysis is reported in accordance with the CONSORT 2025 statement [22] and CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) guidelines [23], and there were no changes to the methods after trial commencement. Although the focus of this report is on the skill-based VR arm of the study, in particular a 2-year durability test of the skill-based VR treatment advantage in the HICP group, for completeness, we briefly describe some of the trial details, including the sham VR group. A double-blind, randomized placebo-controlled trial was conducted in a national community sample of 1067 individuals with self-reported chronic lower back pain for at least 3 months with an average BPI [19,20] pain intensity and pain interference score of 4 or higher for the previous month on a scale from 0 to 10 points. The sample had clinically severe pain (baseline pain intensity=6.6; baseline pain interference=6.2; disability=42% severe or complete) and was demographically diverse (female: 411/536, 77%; non-White individuals: 166/536, 31%; high school or lower educational level: 102/536, 19%; mean age 50.8 years).

XSL•FO
RenderX

Intervention and Randomization

Participants were randomized 1:1 to either (1) a 56-session immersive 3D pain relief skill-based VR therapy or (2) a sham VR control (2D nature videos delivered in a VR headset with no pain relief skill content) and received the device (VR device that includes either the skills-based VR therapy or the sham VR therapy) via mail. The skill-based VR therapy included a fixed sequence of 56 daily VR sessions that delivered pain neuroscience education along with evidence-based self-regulatory strategies commonly taught in behavioral pain treatments, such as cognitive behavioral therapy and mindfulness training. The sham VR condition was designed to replicate the look and feel of the skill-based VR therapy while omitting 3D immersion, spoken content, and pain relief skills. Instead, it consisted of a fixed sequence of 56 nature videos in 2D accompanied by neutral music.

Relevant Self-Report Outcome Measures

These are all widely used and well-validated measures that are effective for measuring changes in chronic lower back pain and relevant comorbidities following VR-delivered therapy.

Demographics

Age, gender, race, ethnicity, annual household income, height, weight, and pain comorbidities were collected.

BPI: Pain Interference

The BPI pain interference scale measures how much pain interfered with 7 domains of daily life (enjoyment of life, general activity, mood, normal work, relationships with other people, sleep, and walking ability) over the previous 24 hours using a numeric rating scale from 0 to 10 [20]. The scores on the 7 items are averaged to generate a global pain interference score. Five administrations of the BPI pain interference scale were included (1 per day) during the 5 days following initial baseline administration and prior to the start of therapy, with a requirement that at least 2 be completed. These were averaged with the baseline BPI score to obtain the overall baseline BPI pain interference score. The BPI pain interference scale was also administered at end of treatment and 1 and 2 years posttreatment.

BPI: Pain Intensity

The BPI pain intensity scale uses a single item to measure pain intensity over the previous 24 hours using a numeric rating scale from 0 to 10 [20]. Five administrations of the BPI pain intensity scale were included (1 per day) during the 5 days following initial baseline administration and prior to the start of therapy, with a requirement that at least 2 be completed. These were averaged with the baseline BPI score to obtain the overall baseline BPI pain intensity score. The BPI pain intensity scale was also administered at end of treatment and 1 and 2 years posttreatment.

Patient-Reported Outcomes Measurement Information System Sleep Disturbance (Version 8b)

The Patient-Reported Outcomes Measurement Information System (PROMIS) Sleep Disturbance tool is an 8-item survey that assesses an individual's perception of sleep quality, sleep depth, and restoration associated with sleep over the previous 7 days [24]. Raw scores on the PROMIS Sleep Disturbance survey are converted to T scores, which range from 28.9 to 76.5. This measure was administered at baseline, end of treatment, and 1 and 2 years posttreatment.

PROMIS Depression (Version 8b)

The PROMIS Depression tool is an 8-item survey that assesses the frequency and severity of depressive symptoms over the previous 7 days [25]. It covers topics such as feelings of worthlessness, hopelessness, and loss of interest in activities. Raw scores on the PROMIS Depression survey are converted to T scores, which range from 37.1 to 81.1. This measure was administered at baseline, end of treatment, and 1 and 2 years posttreatment.

Oswestry Disability Index (Version 2.1b)

The Oswestry Disability Index (ODI) is a 10-item survey that assesses how low back pain affects one's ability to manage in everyday life [26]. Scores range from 0 to 100 and reflect the level of disability. This measure was administered at baseline, end of treatment, and 1 and 2 years posttreatment.

Harms

Harms were assessed and reported for the primary randomized controlled trial [15]. No serious adverse events related to the VR intervention were reported. Harms were not separately reassessed in this secondary analysis, which focused on long-term efficacy outcomes by pain impact subgroup.

Ethical Considerations

With respect to human subject research ethics review, exemptions, and approvals, the study protocol (approval 132166) was approved by the WCG Institutional Review Board (Puyallup, Washington) in December 2021 and followed the CONSORT reporting guidelines. The randomized controlled trial was prospectively registered with ClinicalTrials.gov on January 31, 2022 (ClinicalTrials.gov identifier: NCT05263037). The first participant was enrolled on February 14, 2022. All participants provided informed consent, and study data were collected from February 14, 2022, to October 31, 2024 (2 years posttreatment). Participant identification was protected with a unique study identification number. Data were received electronically through Curebase, secured in the database, and stored with double password protection. All study data were anonymized and deidentified. Study participants received US $210 for the completion of all study surveys up to day 56 (end of treatment), an additional US $90 for completing the 1-year posttreatment surveys, and an additional US $90 for completing the 2-year posttreatment follow-up surveys. No images in the manuscript or supplementary materials contain identifiable individual participants.

Statistical Analysis Methods

The standardized mean differences between the skill-based VR HICP and LICP groups were calculated. Between-group comparisons (HICP vs LICP) were conducted using a general linear model with age, sex, and BMI as covariates. First, we evaluated the skill-based VR therapy effects in HICP vs LICP on the 2 primary end points (BPI pain interference and BPI pain

XSL•FO
RenderX

intensity reduction from baseline to 2 years posttreatment). The Bonferroni-adjusted statistical significance level for comparing groups was set at a 2-sided value of .03 to adjust for 2 comparisons. Next, we evaluated the skill-based VR therapy effects in HICP vs LICP on the 3 secondary end points (PROMIS Sleep Disturbance, PROMIS Depression, and ODI reduction from baseline to 2 years posttreatment). The Bonferroni-adjusted statistical significance level for comparing groups was set at a 2-sided value of .02 to adjust for 3 comparisons.

Results

Participant Characteristics

Complete demographics and baseline clinical measures, as well as full results for the skill-based VR group analysis of the primary and secondary end points at end of treatment and 1 year posttreatment by pain impact group (LICP vs HICP) are published and accessible via open source [15,16]. The CONSORT participant flow diagram for the skill-based VR and sham VR groups is presented in Figure 1. Table 1 shows the baseline demographic and clinical characteristics by pain impact group (LICP vs HICP) for the skill-based VR therapy participants. A total of 35.8% (192/536) were classified as HICP. The HICP group had greater racial diversity (*P*=.005); lower annual household income (*P*<.001); higher BMI (*P*=.01); and significantly greater levels of symptom severity across all variables, with *P* value group differences ranging from .01 to <.001. Two-year posttreatment survey completion was high, with 88.1% (420/477) of skill-based VR therapy participants who completed end-of-treatment surveys completing the 2-year posttreatment survey. No differences in demographic or baseline clinical variables were observed between dropouts and responders within the skill-based VR HICP or LICP subgroups. In addition, the pattern of demographic and baseline clinical variable differences observed between the skill-based VR HICP and LICP subgroups did not differ for dropouts vs responders.

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) participant flow diagram for the full randomized controlled trial comparing skill-based virtual reality (VR) with sham VR control in patients with chronic low back pain (participant flow for end of treatment and 1 and 2 years posttreatment included).

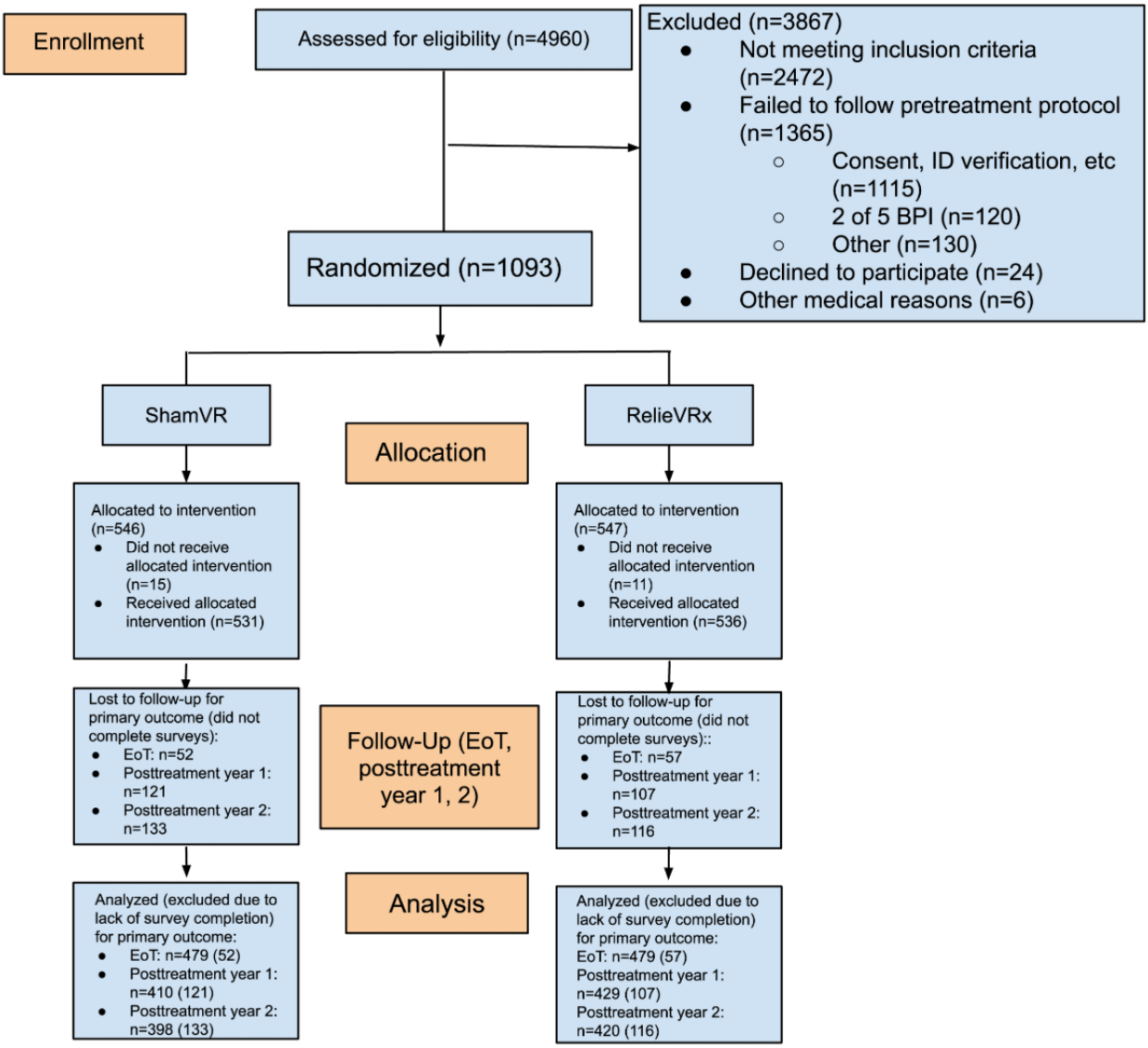

XSL•FO
RenderX

Table 1. Participant demographic characteristics and select baseline measures in the modified intent to treat analysis for the subgroup of participants who completed the skill-based virtual reality (VR) therapy in a randomized controlled trial that compared skill-based VR with a sham VR control in patients with chronic low back pain.

	Skill-based VR—LICP[a] (n=344)	Skill-based VR—HICP[b] (n=192)	*P* value[c]
Gender, n (%)			.03
Female	253 (73.5)	158 (82.3)	
Male	90 (26.2)	34 (17.7)	
Nonbinary	1 (0.3)	0 (0)	
Race and ethnicity, n (%)			.005
American Indian or Alaska Native	3 (0.9)	1 (0.5)	
Asian or Pacific Islander	7 (2)	5 (2.6)	
Black or African American	34 (9.9)	47 (24.5)	
Hispanic or Latino	11 (3.2)	7 (3.6)	
Multiracial	29 (8.4)	22 (11.5)	
White	260 (75.6)	110 (57.3)	
Age (y), mean (SD)	50.9 (14.2)	49.5 (11.9)	.23
Annual household income (US $), n (%)			<.001
<60,000	181 (52.6)	133 (69.3)	
≥60,000	163 (47.4)	59 (30.7)	
Number of pain comorbidities, n (%)			.52
0	140 (40.7)	75 (39.1)	
≥1	182 (52.9)	110 (57.3)	
Nonreported	22 (6.4)	7 (3.6)	
BMI (kg/m^2), mean (SD)	30.4 (7.7)	33.9 (17.8)	.01
BPI[d] pain intensity (0-10), mean (SD)	5.9 (1.2)	7.9 (1.0)	<.001
BPI pain interference (0-10), mean (SD)	5.2 (1.4)	8.1 (0.8)	<.001
PROMIS[e] Anxiety, mean (SD)	55.1 (8.8)	57.9 (9.5)	<.001
PROMIS Sleep Disturbance, mean (SD)	58.7 (6.7)	64.4 (6.5)	<.001
PROMIS Depression, mean (SD)	54.0 (8.6)	56.7 (10.2)	.002
ODI[f], mean (SD)	36.0 (13.2)	51.3 (14.7)	<.001

[a]LICP: low-impact chronic pain.
[b]HICP: high-impact chronic pain.
[c]*P* value from 2-tailed *t* test or Fisher exact test comparing LICP to HICP.
[d]BPI: Brief Pain Inventory.
[e]PROMIS: Patient-Reported Outcomes Measurement Information System.
[f]ODI: Oswestry Disability Index.

Skill-Based VR Therapy Pain Impact Group Differences at 2 Years Posttreatment

Primary End Points

Table 2 shows the mean reductions, 95% CIs, and effect sizes for BPI pain interference and BPI pain intensity from baseline to 2 years posttreatment in the skill-based VR therapy group separately for the HICP and LICP groups. The average pain interference reduction was statistically larger in the HICP group (mean 3.1, 95% CI 2.66-3.54; within-group effect size=1.12) than in the LICP group (mean 1.7, 95% CI 1.45-1.95; within-group effect size=0.79; *P*<.001; between-group effect size=0.60). Interestingly, at 2 years posttreatment, the average pain interference score of participants with HICP was classified in the LICP range (2-year posttreatment average=5.02), with 71.1% (106/149) of individual participants with HICP at 2 years posttreatment reclassified as LICP. The average pain intensity reduction was statistically larger in the HICP group (mean 2.6, 95% CI 2.18-3.02; within-group effect size=1.01) than in the LICP group (mean 1.5, 95% CI 1.24-1.76; within-group effect size=0.68; *P*<.001; between-group effect size=0.48). Although they have been published elsewhere [14], Table 2 also includes

XSL•FO
RenderX

the primary end point results from 1 year posttreatment for comparison purposes. The 1-year posttreatment findings mirror those from 2 years posttreatment, suggesting that the HICP advantage is robust and durable. Figure 2 shows the average pain interference (panel A) and average pain intensity (panel B) ratings for the HICP and LICP groups at baseline, end of treatment, and 1 and 2 years posttreatment, providing a visual of the trajectory of pain ratings over time.

Table 2. Primary and secondary end point reductions from baseline to 2 years posttreatment and from baseline to 1 year posttreatment for the subgroup of participants who completed the skill-based VR therapy in a randomized controlled trial comparing skill-based VR with a sham VR control in patients with chronic low back pain.

	HICP[a,b]	LICP[b,c]	Between-group effect size	*P* value: HICP vs LICP
2 years posttreatment relative to baseline				
BPI[d] Pain Interference	3.1; 2.66 to 3.54; (1.12)	1.7; 1.45 to 1.95; (0.79	0.60	<.001
BPI Pain Intensity	2.6; 2.18 to 3.02; (1.01)	1.5; 1.24 to 1.76; (0.68)	0.48	<.001
PROMIS[e] Sleep Disturbance	7.8; 6.25 to 9.35; (0.80)	4.1; 3.17 to 5.03; (0.53)	0.43	<.001
PROMIS Depression	0.9; –0.9 to 2.70; (0.08)	1.9; 0.93 to 2.87; (0.24)	0.11	.34
Oswestry Disability Index	14.7; 11.87 to 17.53; (0.84)	8.3; 6.59 to 10.01; (0.58)	0.42	<.001
1 year posttreatment relative to baseline				
BPI Pain Interference	2.5; 2.2 to 2.8; (1.04)	1.6; 1.35 to 1.85; (0.73)	0.45	<.001
BPI Pain Intensity	2.0; 1.65 to 2.35; (0.93)	1.5; 1.26 to 1.74; (0.71)	0.30	.003
PROMIS Sleep Disturbance	6.9; 5.46 to 8.34; (0.74)	4.5; 1.63 to 5.37; (0.56)	0.33	.001
PROMIS Depression	1.5; –0.43 to 3.43; (0.09)	1.9; 1.01 to 2.79; (0.25)	0.08	.43
Oswestry Disability Index	12.7; 9.93 to 15.47; (0.69)	8.8; 7.13 to 10.47; (0.60)	0.23	.02

[a]HICP: high-impact chronic pain.

[b]Mean; 95% CIs (within-group effect size) point reduction from baseline to 2-years posttreatment or 1-year posttreatment.

[c]LICP: low-impact chronic pain.

[d]BPI: Brief Pain Inventory

[e]PROMIS: Patient-Reported Outcomes Measurement Information System.

Figure 2. Average (A) pain interference and (B) pain intensity reductions (with error bars) from baseline to end of treatment and 1 and 2 years posttreatment for the subgroup of participants who completed the skill-based virtual reality (VR) therapy (subdivided into high-impact chronic pain [HICP] and low-impact chronic pain [LICP] groups) from a randomized controlled trial comparing skill-based VR with a sham VR control in patients with chronic low back pain.

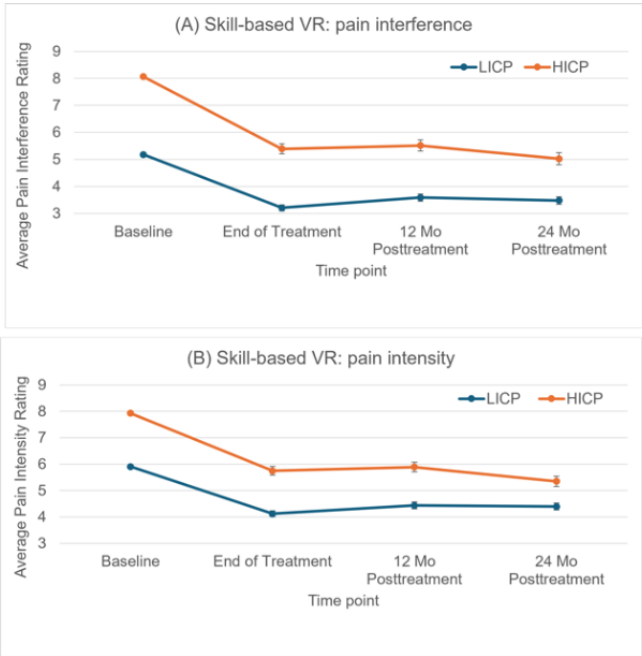

XSL•FO
RenderX

Secondary End Points

Table 2 also shows the mean reductions, 95% CIs, and effect sizes for the PROMIS Sleep Disturbance, PROMIS Depression, and ODI from baseline to 2 years posttreatment in the skill-based VR therapy group separately for the HICP and LICP groups. The average sleep disturbance reduction was statistically larger in the HICP group (mean 7.8, 95% CI 6.25-9.35; within-group effect size=0.80) than in the LICP group (mean 4.1, 95% CI 3.17-5.03; within-group effect size=0.53; $P<.001$; between-group effect size=0.43). The average depression reduction was 0.9 (95% CI −0.90 to 2.70; within-group effect size=0.08) in the HICP group and 1.9 (95% CI 0.93-2.87; within-group effect size=0.24) in the LICP group and did not differ statistically ($P=.34$; between-group effect size=0.11). The average disability reduction was statistically larger in the HICP group (mean 14.7, 95% CI 11.87-17.53; within-group effect size=0.84) than in the LICP group (mean 8.3, 95% CI 6.59-10.01; within-group effect size=0.58; $P<.001$; between-group effect size=0.42). Although they have been published elsewhere [14], Table 2 also includes the secondary end point results from 1 year posttreatment for comparison purposes. The 1-year posttreatment findings mirror those from 2 years posttreatment, suggesting that the HICP advantage is robust and durable.

Discussion

Principal Findings

In this first-ever examination of treatment durability to 2 years posttreatment for a 56-session skill-based VR-delivered therapy for chronic low back pain, we found that the HICP group had clinically meaningful pain interference and pain intensity reductions that were statistically superior to those in the LICP group, with large within-group effect sizes. Most participants with HICP (106/149, 71%) were reclassified as LICP at 2 years posttreatment. We also found larger sleep disturbance and physical disability reductions for HICP over LICP at 2 years posttreatment. These findings are consistent with the hypothesis that patients with HICP would demonstrate strong and durable treatment response, confirming and extending our prior findings at 1 year posttreatment [14].

Patients with HICP have higher health service use (including more frequent primary care visits, specialist consultations, and emergency department use and greater reliance on diagnostic imaging and interventional procedures) [7-10], greater risks from various pain treatments—including medication side effects, complications from invasive procedures, and potential misuse of medications—and only modest clinical improvement relative to patients with LICP [27-29]. These elevated risks are particularly concerning given the diminishing marginal returns of repeated or escalated medical interventions in this population, suggesting a mismatch between treatment intensity and therapeutic benefit [1]. HICP is also more common among women, older adults, people with lower socioeconomic status, and rural residents [3,30], likely reflecting a complex interaction between biological vulnerability, psychosocial stressors, reduced access to comprehensive pain care, and structural inequities in health care delivery. Although these groups have long been known to experience more severe chronic pain, the HICP classification offers a validated way to examine pain impacts and assess treatment effectiveness [21]. Importantly, this classification moves beyond simple pain intensity metrics by incorporating functional impact, thereby providing a more clinically meaningful framework for stratifying patients and tailoring interventions [21]. Data show that patients with HICP are often overmedicalized and undertreated with behavioral approaches [31]. This imbalance may perpetuate a cycle in which patients become increasingly dependent on biomedical interventions while missing opportunities to develop adaptive self-regulation and coping skills. There is evidence suggesting that an HICP treatment model should prioritize behavioral medicine interventions at the outset of pain treatment to minimize patient risks and enhance both pain and health outcomes [32], possibly preventing chronicity by targeting maladaptive learning processes such as fear avoidance and pain catastrophizing before they become entrenched. In-person behavioral medicine approaches such as cognitive behavioral therapy have demonstrated modest efficacy for HICP [12,13], but access is often poor because of the multisession and therapist-led nature of cognitive behavioral therapy [33]. Additional barriers include geographic limitations, long wait times, cost, and stigma associated with seeking psychological treatment for pain [32]. In addition, effectiveness can fluctuate based on the quality of the therapist [34], thus introducing inconsistency in treatment delivery, which may attenuate outcomes and complicate efforts to scale evidence-based care. The skill-based VR intervention evaluated in this study offers high-quality, consistent therapy from session to session that is available on demand in the home, thus helping address the access and therapist quality issues [15]. By standardizing therapeutic content while maintaining interactivity and engagement, VR-based delivery may reduce variability and enhance treatment fidelity across diverse patient populations [32]. The skill-based VR therapy is also easy to use, as evidenced by its A+ rating on the System Usability Scale [14]. Importantly, this A+ rating was observed across several sociodemographic factors (age, gender, race and ethnicity, and socioeconomic status) [35] and for both HICP and LICP [14], suggesting that the intervention is broadly accessible and does not exacerbate existing disparities in digital health engagement. The results of this study combined with those of our previous study [14] suggest that the patients with chronic lower back pain who are most impacted and symptomatic (ie, patients with HICP) reaped the largest short-term and long-term benefits from the skill-based VR therapy. This pattern is consistent with a "greater room for improvement" effect but may also reflect heightened responsiveness to interventions that directly target central pain modulation and cognitive-emotional processes [36]. The finding that 71.1% (106/149) of patients with HICP were reclassified as LICP at 2 years posttreatment is also noteworthy. Future research may examine whether most patients with HICP who were reclassified as LICP also exhibit expected reductions in health care use and associated medical treatment risks. Longitudinal analyses linking classification changes to downstream health care costs, opioid use, and procedural interventions would provide critical evidence for the broader health-economic value of this approach.

XSL•FO
RenderX

The existing literature suggests that user experience features such as interactivity, gamification, and virtual embodiment systematically modulate how much pain relief people obtain from VR, with more engaging and embodied designs generally producing larger analgesic effects [37-39]. Although the skill-based VR therapy contained some elements of interactivity, gamification, and embodiment, it was primarily designed as a multimodal program drawing on cognitive behavioral strategies, meditation, mindfulness, biofeedback, education, and interoceptive awareness and control [15]. Collectively, these results suggest that complex patients with HICP gain significant advantages from this self-delivered multimodal therapy. Even so, building on the existing literature, future iterations of the therapy should explore incorporating more interactivity, gamification, and virtual embodiment as this could enhance the effectiveness of the therapy beyond the current level.

Limitations and Strengths

Several limitations should be considered when interpreting these findings. First, although the HICP vs LICP classification is empirically derived, participants could not be randomized, and the results remain correlational. Second, unmeasured factors may be causal, including pain-related and non–pain-related comorbidities. Greater characterization of the diagnoses and conditions underlying the HICP group (eg, trauma history, disability status, and cancer or palliative care) and of concurrent treatments over the 2-year period (eg, opioid use, injections, and surgery) would strengthen interpretation and should be addressed in future research. Third, chronic lower back pain was self-reported and not clinically confirmed. Finally, because the study focused on chronic lower back pain, the findings may not generalize to other populations. Despite these limitations, we also note several study strengths that contextualize its value. First, these results come from the largest real-world community sample dataset for therapeutic skill-based VR (n=536) with a large HICP subset (n=192). Second, 2-year participant retention was high, with 88.1% (420/477) of skill-based VR therapy participants who completed end-of-treatment surveys completing the 2-year posttreatment survey. Finally, we tested heterogeneity of treatment effect across a factor that is broadly applicable and relevant to clinicians (ie, HICP vs LICP).

Conclusions

This secondary analysis provides the first evidence of 2-year durability of a skill-based VR-delivered therapy in HICP, demonstrating clinically meaningful and sustained benefits in a population that is often overmedicalized and undertreated with behavioral approaches. The finding that most participants with HICP were reclassified as LICP at 2 years posttreatment has potential implications for reducing health care use and improving quality of life in this high-need population. These results support continued investigation of home-based, self-administered VR-delivered behavioral therapy as a scalable, accessible, and effective treatment option for patients with HICP. Future research should examine whether HICP reclassification is associated with reductions in health care use and should evaluate the therapy across other chronic pain conditions.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Funding

AppliedVR, Inc, financially supported this study.

Authors' Contributions

TM and JS were involved in all aspects of the study. RM was involved in study implementation, project and participant management, data analysis, and data presentation. RB served as medical monitor and helped interpret the findings. BDD was involved in study design, data interpretation, and manuscript preparation. All authors had full access to all the data in the study and had final responsibility for the decision to submit the manuscript for publication.

Conflicts of Interest

TM and RM are employees of AppliedVR, Inc. JS is president of AppliedVR, Inc. RB is a consultant for AppliedVR, Inc, and served as the medical monitor on the trial. BDD is compensated for serving as chief science advisor for AppliedVR, Inc. BDD has authored or coauthored 5 pain treatment books for patients and clinicians and receives royalties for 4. BDD is the principal investigator for pain research grants and awards from the National Institutes of Health (NIH) and the Patient-Centered Outcomes Research Institute (nonspecific to this work). BDD is a coinvestigator on 2 NIH research grants investigating virtual reality analgesia; neither of these grants is specific to this work. BDD serves on the Board of Directors of the Institute for Brain Potential and on the Medical Advisory Board of the Facial Pain Association. BDD is a current scientific member of the NIH Interagency Pain Research Coordinating Committee, and is on the Board of Directors and President-Elect of the United States Association for the Study of Pain. BDD is a former member of the Centers for Disease Control and Prevention Opioid Workgroup (2020-2021), and former member of the Pain Advisory Group of the American Psychological Association.

Multimedia Appendix 1
CONSORT checklist.
[PDF File (Adobe PDF File), 263 KB - jmir_v28i1e90688_app1.pdf]

XSL•FO
RenderX

References

1. Şentürk İ, Şentürk E, Üstün I, Gökçedağ A, Yıldırım NP, İçen NK. High-impact chronic pain: evaluation of risk factors and predictors. Korean J Pain 2023 Jan 01;36(1):84-97 [FREE Full text] [doi: 10.3344/kjp.22357] [Medline: 36581599]
2. Dahlhamer J, Lucas J, Zelaya C, Nahin R, Mackey S, DeBar L, et al. Prevalence of chronic pain and high-impact chronic pain among adults - United States, 2016. MMWR Morb Mortal Wkly Rep 2018 Sep 14;67(36):1001-1006 [FREE Full text] [doi: 10.15585/mmwr.mm6736a2] [Medline: 30212442]
3. Pitcher MH, Von Korff M, Bushnell MC, Porter L. Prevalence and profile of high-impact chronic pain in the United States. J Pain 2019 Feb;20(2):146-160 [FREE Full text] [doi: 10.1016/j.jpain.2018.07.006] [Medline: 30096445]
4. Vaegter HB, Bruun KD, Bye-Møller L. High-impact chronic pain. International Association for the Study of Pain. 2023. URL: https://www.iasp-pain.org/resources/fact-sheets/high-impact-chronic-pain/ [accessed 2026-04-15]
5. Lucas JW, Sohi I. Chronic pain and high-impact chronic pain in U.S. adults, 2023. NCHS Data Brief 2024 Oct(518):CS355235. [doi: 10.15620/cdc/169630] [Medline: 39751180]
6. Zajacova A, Grol-Prokopczyk H, Nahin RL. Pain among US adults before, during, and after the COVID-19 pandemic: a study using the 2019 to 2023 National Health Interview Survey. Pain 2026 Jan 01;167(1):142-149. [doi: 10.1097/j.pain.0000000000003764] [Medline: 40758540]
7. Herman PM, Broten N, Lavelle TA, Sorbero ME, Coulter ID. Health care costs and opioid use associated with high-impact chronic spinal pain in the United States. Spine (Phila Pa 1976) 2019 Aug 15;44(16):1154-1161 [FREE Full text] [doi: 10.1097/BRS.0000000000003033] [Medline: 31373999]
8. Stubhaug A, Hansen JL, Hallberg S, Gustavsson A, Eggen AE, Nielsen CS. The costs of chronic pain-long-term estimates. Eur J Pain 2024 Jul;28(6):960-977. [doi: 10.1002/ejp.2234] [Medline: 38214661]
9. Tamargo JA, Cruz-Almeida Y. Concurrent validity of the operationalization of high-impact pain construct in the health and retirement study. J Pain Res 2025;18:5533-5540 [FREE Full text] [doi: 10.2147/JPR.S539052] [Medline: 41146648]
10. Guy GPJ, Miller GF, Legha JK, Rikard SM, Strahan AE, Mikosz C, et al. Economic costs of chronic pain-United States, 2021. Med Care 2025 Sep 01;63(9):679-685. [doi: 10.1097/MLR.0000000000002181] [Medline: 40730349]
11. National pain strategy: a comprehensive population health-level strategy for pain. National Institute of Neurological Disorders and Stroke. 2010. URL: https://www.ninds.nih.gov/sites/default/files/documents/NationalPainStrategy_508C.pdf [accessed 2026-04-15]
12. Sanabria-Mazo JP, Colomer-Carbonell A, Fernández-Vázquez Ó, Noboa-Rocamora G, Cardona-Ros G, McCracken LM, et al. A systematic review of cognitive behavioral therapy-based interventions for comorbid chronic pain and clinically relevant psychological distress. Front Psychol 2023;14:1200685 [FREE Full text] [doi: 10.3389/fpsyg.2023.1200685] [Medline: 38187407]
13. Liu X, Yuan W, Gao X, Zhao Z, Leng R, Xia Y. Efficacy of cognitive behavioral therapy for musculoskeletal pain: a systematic review and meta-analysis. Front Psychol 2025 Jan 20;16:1705679. [doi: 10.3389/fpsyg.2025.1705679] [Medline: 41640593]
14. Maddox T, Oldstone L, Linde-Zwirble W, Bonakdar R, Maddox R, Sackman J, et al. Differential treatment response to virtual reality in high-impact chronic pain: secondary analysis of a randomized trial. Sci Rep 2025 Apr 25;15(1):14430 [FREE Full text] [doi: 10.1038/s41598-025-98716-3] [Medline: 40281004]
15. Maddox T, Oldstone L, Sparks CY, Sackman J, Oyao A, Garcia L, et al. In-home virtual reality program for chronic lower back pain: a randomized sham-controlled effectiveness trial in a clinically severe and diverse sample. Mayo Clin Proc Digit Health 2023 Dec;1(4):563-573 [FREE Full text] [doi: 10.1016/j.mcpdig.2023.09.003] [Medline: 40206316]
16. Maddox T, Oldstone L, Sackman J, Maddox R, Adair T, Ffrench K, et al. Twelve-month results for a randomized sham-controlled effectiveness trial of an in-home skills-based virtual reality program for chronic low back pain. Pain Rep 2024 Oct;9(5):e1182. [doi: 10.1097/PR9.0000000000001182] [Medline: 39239633]
17. Farrar JT, Young JPJ, LaMoreaux L, Werth JL, Poole MR. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. Pain 2001 Nov;94(2):149-158. [doi: 10.1016/S0304-3959(01)00349-9] [Medline: 11690728]
18. Kovacs FM, Abraira V, Royuela A, Corcoll J, Alegre L, Cano A, et al. Minimal clinically important change for pain intensity and disability in patients with nonspecific low back pain. Spine (Phila Pa 1976) 2007 Dec 01;32(25):2915-2920. [doi: 10.1097/BRS.0b013e31815b75ae] [Medline: 18246018]
19. Cleeland CS. The Brief Pain Inventory: user guide. University of Texas MD Anderson Cancer Center. 2009. URL: https://www.mdanderson.org/documents/Departments-and-Divisions/Symptom-Research/BPI_UserGuide.pdf [accessed 2026-04-15]
20. Cleeland CS, Ryan KM. Pain assessment: global use of the Brief Pain Inventory. Ann Acad Med Singap 1994;23(2):129-138 [FREE Full text]
21. Cook KF, You DS, Korff MV, Mackey SC. Defining chronic pain impact levels: a patient-clinician approach using PROMIS® pain interference scores. J Patient Rep Outcomes 2025 Aug 18;9(1):103. [doi: 10.1186/s41687-025-00908-y] [Medline: 40824460]
22. Hopewell S, Chan A, Collins GS, Hróbjartsson A, Moher D, Schulz KF, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ 2025 Apr 14;389:e081123. [doi: 10.1136/bmj-2024-081123] [Medline: 40228833]

XSL•FO
RenderX

23. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
24. Yu L, Buysse DJ, Germain A, Moul DE, Stover A, Dodds NE, et al. Development of short forms from the PROMIS™ sleep disturbance and Sleep-Related Impairment item banks. Behav Sleep Med 2011 Dec 28;10(1):6-24 [FREE Full text] [doi: 10.1080/15402002.2012.636266] [Medline: 22250775]
25. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Med Care 2007 May;45(5 Suppl 1):S3-11 [FREE Full text] [doi: 10.1097/01.mlr.0000258615.42478.55] [Medline: 17443116]
26. Fairbank JC, Pynsent PB. The Oswestry Disability Index. Spine (Phila Pa 1976) 2000 Nov 15;25(22):2940-52; discussion 2952. [doi: 10.1097/00007632-200011150-00017] [Medline: 11074683]
27. Chou R, Hartung D, Turner J, Blazina I, Chan B, Levander X, et al. Opioid Treatments for Chronic Pain. Rockville, MD: Agency for Healthcare Research and Quality (US); 2020.
28. Vane JR, Botting RM. Mechanism of action of anti-inflammatory drugs. Scand J Rheumatol Suppl 1996 Jul 12;102(sup102):9-21. [doi: 10.3109/03009749609097226] [Medline: 8628981]
29. Bair MJ, Robinson RL, Katon W, Kroenke K. Depression and pain comorbidity: a literature review. Arch Intern Med 2003 Nov 10;163(20):2433-2445. [doi: 10.1001/archinte.163.20.2433] [Medline: 14609780]
30. Prevalence and profile of high impact chronic pain. National Center for Complementary and Integrative Health. URL: https://www.nccih.nih.gov/research/research-results/prevalence-and-profile-of-high-impact-chronic-pain [accessed 2026-04-15]
31. Falasinnu T, Hossain MB, Weber KA2, Helmick CG, Karim ME, Mackey S. The problem of pain in the United States: a population-based characterization of biopsychosocial correlates of high impact chronic pain using the National Health Interview Survey. J Pain 2023 Jun;24(6):1094-1103 [FREE Full text] [doi: 10.1016/j.jpain.2023.03.008] [Medline: 36965649]
32. Maddox T, Sparks CY, Oldstone L, Chibbaro M, Sackman J, Judge E, et al. Perspective: the promise of virtual reality as an immersive therapeutic. J Med Ext Real 2024 Jan 01;1(1):13-20. [doi: 10.1089/jmxr.2023.0003]
33. Darnall BD, Scheman J, Davin S, Burns JW, Murphy JL, Wilson AC, et al. Pain psychology: a global needs assessment and national call to action. Pain Med 2016 Feb 23;17(2):250-263 [FREE Full text] [doi: 10.1093/pm/pnv095] [Medline: 26803844]
34. Day MA, Ehde DM, Bindicsova I, Jensen MP. Understanding the role of therapist quality in accounting for heterogeneity of patient outcomes in psychosocial chronic pain treatments. J Pain 2024 Apr;25(4):843-856 [FREE Full text] [doi: 10.1016/j.jpain.2023.10.007] [Medline: 37832902]
35. Maddox T, Oldstone L, Sackman J, Judge E, Maddox R, Adair T, et al. Sociodemographic predictors of clinical effectiveness, therapeutic program engagement, and device usability for an in-home virtual reality program for chronic low back pain: secondary analysis of a randomized controlled trial. J Med Ext Real 2024;1(1):65-72. [doi: 10.1089/jmxr.2023.0013]
36. Bushnell MC, Ceko M, Low LA. Cognitive and emotional control of pain and its disruption in chronic pain. Nat Rev Neurosci 2013 Jul;14(7):502-511 [FREE Full text] [doi: 10.1038/nrn3516] [Medline: 23719569]
37. Saby A, Alvarez A, Smolins D, Petros J, Nguyen L, Trujillo M, et al. Effects of embodiment in virtual reality for treatment of chronic pain: pilot open-label study. JMIR Form Res 2024 Feb 16;8:e34162 [FREE Full text] [doi: 10.2196/34162] [Medline: 38363591]
38. Wender R, Hoffman HG, Hunner HH, Seibel EJ, Patterson DR, Sharar SR. Interactivity influences the magnitude of virtual reality analgesia. J Cyber Ther Rehabil 2009;2(1):27-33 [FREE Full text] [Medline: 20390047]
39. Özlü A, Ünver G, Tuna H, Menekşeoğlu AK. The effect of a virtual reality-mediated gamified rehabilitation program on pain, disability, function, and balance in knee osteoarthritis: a prospective randomized controlled study. Games Health J 2023 Apr;12(2):118-124. [doi: 10.1089/g4h.2022.0130] [Medline: 36603100]

Abbreviations

BPI: Brief Pain Inventory
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
HICP: high-impact chronic pain
LICP: low-impact chronic pain
ODI: Oswestry Disability Index
PROMIS: Patient-Reported Outcomes Measurement Information System
VR: virtual reality

XSL•FO
RenderX

Edited by S Brini; submitted 01.Jan.2026; peer-reviewed by J Guerra Armas; comments to author 18.Feb.2026; accepted 01.Apr.2026; published 23.Apr.2026.

Please cite as:
Maddox T, Sackman J, Bonakdar R, Maddox R, Darnall BD
Skill-Based Virtual Reality Therapy in High-Impact Chronic Pain: 2-Year Follow-Up Results From a Secondary Analysis of a Randomized Controlled Trial
J Med Internet Res 2026;28:e90688
URL: https://www.jmir.org/2026/1/e90688
doi:10.2196/90688

©Todd Maddox, Josh Sackman, Robert Bonakdar, Roselani Maddox, Beth D Darnall. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Time Savings Through an AI Speech Assistant for Nursing Documentation: Pre-Post Time-Motion Study in German Long-Term Care

Katja Schwabe[1*], BSc; Drin Ferizaj[2*], MSc; Susann Neumann[2], DipNEd; Sandra Strube-Lahmann[2], PhD; Nils Lahmann[2], Prof Dr

[1]voize GmbH, Dircksenstraße 47, Berlin, Germany

[2]Department of Geriatrics and Medical Gerontology, Charité – Universitätsmedizin Berlin, Berlin, Germany

[*]these authors contributed equally

Corresponding Author:
Katja Schwabe, BSc
voize GmbH, Dircksenstraße 47, Berlin, Germany

Abstract

Background: Nurses in long-term care spend up to one-third of their working time on documentation, contributing to administrative burden and limited time for direct care. Artificial intelligence (AI) speech assistants have shown potential to accelerate documentation, but longitudinal evidence from real-world long-term care settings remains scarce.

Objective: This study aimed to evaluate whether implementing a domain-specific, mobile AI speech assistant is associated with reduced documentation time in German long-term care under routine conditions. Secondary objectives included examining usability, perceived documentation effort, interruptions, and workplace satisfaction.

Methods: A pre-post time-motion study with full-shift observation was conducted. Continuous, event-based observations were performed before (t_0) and after (t_1) implementation of the mobile speech assistant voize. The primary outcome was total documentation time per morning shift based on direct observations. In addition to observations, questionnaires were administered to assess perceived documentation effort, interruptions, satisfaction with the documentation system, and workplace satisfaction. The primary end point was analyzed using a linear mixed-effects model. Secondary, self-reported outcomes were analyzed exploratorily via paired pre-post differences with pooling across multiple imputations and Holm-Bonferroni correction.

Results: A total of 52 registered nurses from 14 long-term care facilities participated (mean age 42.37, SD 12.37 years; 42/52, 80.8% female). Across 770 observed hours, the observed total documentation time per morning shift decreased significantly by an adjusted mean of 15 (SE 3.36) minutes, $t_{46.29}$=−4.46, P<.001, with a 95% CI of −21.75 to −8.23, corresponding to an approximately 28% reduction relative to the baseline mean. Holm-Bonferroni–corrected exploratory analyses indicated significant declines in self-reported documentation time and interruptions, and satisfaction with the documentation system improved, while workplace satisfaction showed no significant change. Usability was rated as acceptable.

Conclusions: This study provides real-world evidence from a single-group pre-post design that an AI-based speech assistant is associated with reduced documentation workload in long-term care. In this sample, the integration of a mobile, domain-specific speech system into daily workflows coincided with substantially decreased documentation time and improved perceived efficiency. Beyond these observed time savings, such technology has the potential to alleviate workload, free time for resident care, and enhance working conditions. These findings are also relevant for policy discussions on addressing the nursing workforce shortage, showing that well-integrated, speech-enabled documentation systems can support more sustainable long-term care environments.

Trial Registration: German Clinical Trials Register DRKS00035512; https://drks.de/search/en/trial/DRKS00035512

(J Med Internet Res 2026;28:e86078) doi:10.2196/86078

KEYWORDS

nursing documentation time; automatic speech recognition; long-term care; artificial intelligence; AI speech assistant; mobile nursing documentation

XSL•FO
RenderX

Introduction

Background

In Germany, approximately 1.7 million nursing staff are employed, with nearly 30% working in long-term care facilities, making this the largest sector in the nursing workforce [1]. Demand for care continues to rise while shortages of qualified staff worsen [2]. Nursing staff, particularly in long-term care, report higher levels of strain compared to other occupational groups [3,4]. They work under considerable time pressure, often lack sufficient time to complete tasks, and frequently perform overtime with irregular or missed breaks [3,5]. Alongside these structural challenges and workload demands, care workers face a high administrative burden: documentation tasks account for about 30% of total working time [6,7], leaving less room for direct interaction and care for residents [8]. Furthermore, the high documentation workload has been linked to reduced job satisfaction and an elevated level of stress [9]. In light of these issues, artificial intelligence (AI)–based speech assistants are presented as promising solutions to accelerate the documentation process and improve its accuracy [10].

Medical Documentation With Speech Recognition

Research on speech recognition has been most widely conducted in clinical medicine, often involving physicians and mixed professional groups. Previous findings show that speech recognition technologies can significantly streamline documentation workflows compared to traditional methods. For example, Zuchowski and Göller [11] reported notable time savings and lower error rates when using speech recognition versus manual typing. Similarly, Peine et al [12] found that a voice-based system enabled faster and more accurate documentation among intensive care professionals compared to both electronic and paper-based approaches.

These findings are further supported by systematic reviews. Alboksmaty et al [13] synthesized findings across several countries, highlighting consistent efficiency gains, reduced administrative burden, and improved provider-patient interaction. Of particular note, Wang et al [14] found that a digital scribe could accelerate history-taking and physical examination documentation by more than double compared to typing or traditional dictation.

However, reviews also point out that most studies have taken place in controlled or experimental settings, which may limit generalizability to routine care [10,13]. In this context, Ng et al [15] reviewed 29 research articles and observed that many reported time savings and improvements in documentation completeness, whereas others described challenges such as increased editing workload, variable cost-effectiveness, and issues with error rates for specialized terminology or accented speech.

Clinical Nursing Documentation With Speech Recognition

Compared to physician documentation, evidence on nursing documentation with AI speech assistants is more limited but growing. Findings to date are still heterogeneous. For example, many studies reviewed by Joseph et al [10] indicated beneficial effects, such as reduced documentation time and improved workflow. However, a substantial number reported no significant impact, given ongoing challenges integrating AI speech assistants into real-world nursing environments and existing documentation systems.

However, several recent investigations suggest a promising potential for efficiency gains in real-world nursing environments. In a hospital time-motion study, Ehrler et al [16] found that a mobile bedside application meaningfully decreased the duration of electronic health record (EHR) documentation for nurses while also increasing time available for direct patient interaction. Similarly, Mayer et al [17] reported statistically significant time savings—up to 6.1 minutes per documentation scenario—when using a voice recognition system compared to traditional keyboard entry. Mairittha et al [18] also demonstrated that a smartphone-based dialogue system allowed for 15 percent faster documentation relative to electronic forms in a controlled setting. Likewise, qualitative findings from German long-term care facilities hint at the potential of AI speech assistants for care documentation, reporting self-reported time savings, fewer interruptions, greater flexibility, and less reliance on stationary computers [19].

Research Gap

In summary, evidence across clinical settings suggests that AI speech assistants can shorten documentation time and thus reduce administrative burden. At the same time, findings are mixed and often derive from controlled or pilot evaluations rather than routine practice, which limits generalizability [10,13,15]. In nursing, several studies report faster entries with mobile or dialogue-based tools, but most were scenario-based, single-unit, or acute-care focused, with heterogeneous outcomes and little full-shift observation. Long-term care therefore remains underrepresented despite the high documentation burden [12,16-18].

Objectives

Therefore, this study aims to address these gaps by evaluating a domain-specific mobile AI speech assistant under routine conditions in German long-term care using full-shift time-motion observation at 2 time points, before and after implementation. The primary end point is the adjusted change in total documentation time per observed morning shift, defined as the cumulative time spent on a set of predefined documentation-related tasks. Secondary outcomes are self-reported and include usability, perceived documentation effort and interruptions, workplace satisfaction, and self-reported documentation quality.

Hypotheses

The following hypothesis will be tested for the primary end point:

- H0: There will be no significant reduction measured by the adjusted mean change in total documentation time from baseline to postimplementation.
- H1: There will be a significant reduction measured by the adjusted mean change in total documentation time from baseline to postimplementation.

XSL•FO
RenderX

The secondary outcomes are evaluated exploratorily by pre-post changes. To adjust for multiple testing, the Holm-Bonferroni correction is applied across all secondary tests.

Methods

Study Design and Setting

The study was conducted as part of the "PYSA – Nursing Documentation With Hybrid Speech Assistant" project, funded by the German Federal Ministry of Education and Research [20]. The longitudinal research design follows a pre-post observational time-motion design with 2 measurement points (baseline t_0 and postassessment t_1) and takes place in stationary long-term care facilities in Germany. Facilities were approached nationwide to ensure the representation of different conditions and to include both rural and urban areas as well as different types of providers. All participating facilities used either the same EHR system or paper-based documentation.

The minimum age for inclusion of participants was 18 years, with no upper age limit. Additional inclusion criteria were a completed 3-year nursing education, regular involvement in nursing documentation, and at least 1 year of professional experience. Participants had to be able to understand instructions in German and provide informed consent. Only staff employed in long-term care facilities in which the speech assistant voize was implemented during the study period were eligible. Exclusion criteria were applied to staff who were not permanently employed in the participating facilities, such as agency or temporary nurses. We report this observational study in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement [21].

Recruitment

Participants were recruited through long-term care facilities that planned to implement the speech assistant voize and received written information about the study procedures. Following approval from the employee representative bodies of each facility, nursing staff were approached by the facility management. All participants were informed in writing that participation was voluntary, that data would be pseudonymized and analyzed in aggregated form, and that they could withdraw from the study at any time without facing any negative consequences. Participants were offered a voluntary small financial compensation for their time and effort.

The Speech Assistant Voize

The speech assistant voize is a software-based mobile app that enables users to dictate nursing documentation via smartphone. It uses a continuously trained language model that processes voice inputs locally, assigns the content to the appropriate categories, and automatically generates structured documentation entries. These entries are reviewed and approved by nursing staff before being integrated into the existing documentation system. Developed in close collaboration with nursing professionals [19,20], the application is continuously refined through a collaborative, user-centered design process, and its underlying models are regularly trained with voluntarily submitted and anonymized documentation data to ensure ongoing improvements and adaptation to the linguistic characteristics and professional terminology used in nursing.

Technically, the system converts spoken language into text (speech-to-text) in real time using an end-to-end streaming ASR (automatic speech recognition) model. In conjunction with this, a set of specialized natural language processing models is activated as needed to handle specific tasks—such as understanding context, matching spoken language to planned interventions, extracting key information, or structuring data—ensuring that all components work together to deliver the supported features. As part of the study, voize was introduced in cooperation with voize GmbH. The implementation included planning and rollout within the facilities.

Power Analysis

In a preliminary anonymous survey with 55 nurses already using voize, about 70% reported subjective time savings of 16 to 30 minutes per shift. It was assumed that about one-third of working time is spent on documentation-related tasks [6,7], which is equivalent to about 20 minutes per hour. A 15% reduction was considered meaningful, corresponding to the mean of reported savings (23 minutes per shift), that is, 2.9 minutes less per working hour, reducing documentation from 20 to 17.1 minutes per hour. With an assumed SD of 6 minutes, this equates to an effect size of Cohen d=0.5. Power analysis using the R package pwr [22] (α=.05, 2-sided; power=0.8) indicated a minimum of 34 participants for paired t tests. To account for an expected 15% dropout, the target sample size was set at 40, compensating for likely reasons such as scheduling conflicts, shift changes, sick leave, or vacation.

Early-stage dropout rates exceeded expectations. Therefore, additional recruitment efforts were employed. Nonetheless, interest from both facilities and individual nurses was high, and incremental participant burden remained low. As a result, a total of 52 nurses were included in this study.

Measures

Primary End Point: Total Documentation Time per Shift

The primary outcome was total documentation time per observed morning shift in minutes, defined as the sum of all documentation-related categories recorded during the time-motion observation. Observation categories were developed and refined during preparatory nonparticipant observations in a long-term care facility within an unpublished pilot study [20]. The identified documentation-related tasks included documentation at the PC or on paper, consulting information at the resident record at the PC, waiting for a PC, booting up and logging into the PC, walking to the ward office for documentation purposes, making notes, and reviewing entries (Table S1 in Multimedia Appendix 1). All facilities applied their usual infection control protocols for mobile devices. Nurses were instructed to disinfect the study smartphones in accordance with local hygiene guidelines. When these hygiene steps occurred in direct connection with documentation, they were included in the observed documentation time. At t_1, after implementation of the application, additional categories were

XSL•FO
RenderX

included for documentation with voize and consulting information in voize (Table S2 in Multimedia Appendix 1).

Nursing Workplace Satisfaction Questionnaire (NWSQ)

Workplace satisfaction outcomes were assessed using the NWSQ [23]. The NWSQ is a brief measure of nursing job satisfaction encompassing 3 domains: intrinsic, extrinsic, and relational satisfaction. Items are rated on a 5-point Likert scale, and domain scores are computed as the sum of their constituent items. Possible ranges are 6‑30 for intrinsic, 5‑25 for extrinsic, and 4‑20 for relational satisfaction, with higher scores indicating greater satisfaction.

Usability

Technology-related usability was measured with the system usability scale [24] (SUS), a 10-item questionnaire of perceived usability on a 5-point Likert scale and alternating item valence. Responses were scored using the standard SUS procedure, including appropriate reverse-scoring, and converted to a total score from 0 to 100. Scores around 70 indicate acceptable usability, with higher values reflecting better perceived usability [24].

Self-Developed Questionnaire

Several single-item measures were developed and used to capture several facets of potential impact based on a previous study [19]. Items included perceived interruptions during daily work, the perceived time-consuming nature of documentation, satisfaction with the documentation system, expected quality improvement, expected time savings, current documentation quality and completeness, implementation as a good idea, intention-to-use, and self-reported documentation time in minutes per shift. Items used 5-point scales or 10-point scales as shown in the Results section, plus open numerical entries for minutes. Higher ratings indicate more agreement with the item. For example, higher values on " documentation is time-consuming" indicate greater perceived time burden, and higher values on "satisfaction with the documentation system" indicate greater satisfaction. Item wording and differences across t_0 and t_1 are provided in Tables S1 and S2 in Multimedia Appendix 2.

Procedure

Data were collected between September 2024 and July 2025 in German long-term care facilities. Trained observers conducted continuous, nonparticipant full-shift observations of morning shifts with a planned duration of 8 hours, and observed shift lengths were very similar between t_0 and t_1.

Time was captured with an event-based smartphone logging app by starting and stopping predefined documentation-related task categories. After completing the shift, participants were asked to complete an online questionnaire that included the measures.

Approximately 8 weeks after the implementation of voize (t_1), participants were observed again during a full morning shift, and documentation times were recorded using the same procedure. At this second measurement point, participants completed an adjusted online survey, including the SUS.

Observations at both time points took place on typical weekdays outside major holiday periods. According to the participating units, bed capacity and staffing ratios were generally stable between t_0 and t_1. Taken together, these conditions suggest that major systematic differences in resident census between t_0 and t_1 were unlikely, although smaller variations may still have occurred.

Observers and Minimization of Bias

Prior to data collection, all observers were employees of voize GmbH who were specifically hired for the purpose of this study and were not involved in the development or design of the speech assistant. Observers had no performance incentives tied to study outcomes and were instructed to adhere strictly to the predefined task categories and timing rules to minimize the risk of biased classification or timing cessation. They received standardized training from the Charité research team in the use of the measurement instruments and the predefined task categories. As part of this training, observers first jointly worked through standardized observation scenarios and coded them in parallel. Any discrepancies were then discussed systematically until a shared consensus on the application of each category was reached. In addition, weekly meetings were held throughout the data collection period to address questions, reflect on measurement procedures, and resolve ambiguities in a standardized manner. During the observations, observers also had access to a notes field where they could record questions or uncertainties regarding the assignment to task categories, add comments, and briefly describe the situation. This enabled the study personnel to later compare the selected task category with the descriptive notes and to correct obvious misclassifications where necessary.

In addition, the direct observation design can lead to behavioral changes among participants. The presence of an external observer can sometimes prompt improved performance, a phenomenon known as the Hawthorne effect [25]. To mitigate such influences, repeated briefings and transparent communication of the study objectives were provided to the nursing staff and critical questions were encouraged.

Statistical Analysis

Statistical analyses were performed using Python 3.11.9 and R 4.5.2 in Visual Studio Code 1.103.2 with the packages scikit-learn [26], miceforest [27], NumPy [28], pandas [29], SciPy [30], Matplotlib [31], tidyverse [32], lme4 [33], and lmerTest [34]. All statistical methods used 2-sided tests at α=.05.

To account for the hierarchical and longitudinal structure of the data with repeated measurements within participants and clustering within facilities, the primary end point was analyzed using a linear mixed-effects model (LMM). The primary estimand was the adjusted mean change in documentation time from baseline to postimplementation ($t_1 - t_0$) in minutes. Time was coded as 0 at baseline and 1 at postimplementation, so that the fixed effect of time directly represents the adjusted mean change. The primary model included fixed effects for time, age, gender, and baseline documentation type, and random intercepts for participants and facilities. The model was estimated with

XSL•FO
RenderX

restricted maximum likelihood. Inference for fixed effects used Kenward-Roger degrees of freedom.

The primary model used all available outcome observations and handled missing primary outcomes via likelihood-based estimation under a missing-at-random assumption. Robustness was assessed in exploratory sensitivity analyses, including a complete-case analysis requiring observations at both time points, models additionally accounting for observer effects, and alternative random-effects specifications. We further repeated analyses using multiple imputation with pooling via Rubin rules and Barnard-Rubin degrees of freedom and conducted missing-not-at-random (MNAR) pattern-mixture sensitivity analyses including jump-to-baseline and delta adjustments. As an additional robustness check, we repeated the analysis using a baseline-adjusted analysis of covariance with change scores. Full model specifications, statistical rationale, imputation settings, and MNAR sensitivity procedures are provided in Multimedia Appendix 3. As the primary and secondary analyses fall under the frequentist paradigm, we additionally conducted an exploratory Bayesian hierarchical repeated-measures analysis (Multimedia Appendix 4).

All secondary, self-reported outcomes were analyzed exploratorily using within-subject pre-post comparisons on multiply imputed datasets pooled according to Rubin rules and Barnard-Rubin small sample adjustment. For each outcome, the estimand was the mean change score ($t_1 - t_0$) with heteroscedasticity-consistent type 3 robust standard errors. To account for multiplicity, a Holm-Bonferroni correction was applied across secondary outcomes.

Ethical Considerations

The study was reviewed and approved by the Ethics Committee of Charité – Universitätsmedizin Berlin, approval number EA1/209/24. All participants provided written informed consent prior to participation. Data were pseudonymized and analyzed in aggregated form. The study was preregistered in the German Clinical Trials Register (DRKS00035512). Participants received 30 € (US $34.43) for their participation.

Results

Participant Characteristics

The sample comprised 52 registered nurses from 14 long-term care facilities in Germany. Of the 52 participants, 8 (15.4%) dropped out between baseline and postassessment, primarily due to long-term illness or resignation.

The mean age was 42.37 (SD 12.37) years, with 17.52 (SD 9.63) years of professional experience and 9.42 (SD 8.13) years of tenure. Most participants were female (42/52, 80.8%). Among completers (44/52, 84.6%), the mean age was 43.23 (SD 12.01) years and mean experience was 18.2 (SD 9.95) years; 84.1% (37/44) were female. Dropouts (8/52, 15.4%) were younger on average (mean age 37.62, SD 14.07 years) and reported less work experience (mean 13.75, SD 6.88 years), and 5/8 (62.5%) of them were female. Detailed counts and percentages for all characteristics are provided in Table 1.

Table . Participant characteristics of the total sample, completers, and dropouts.

Variable	Total (N=52)	Completers (n=44)	Dropouts (n=8)
Age (years), mean (SD)	42.37 (12.37)	43.23 (12.01)	37.62 (14.07)
Work experience (years), mean (SD)	17.52 (9.63)	18.20 (9.95)	13.75 (6.88)
Tenure (years), mean (SD)	9.42 (8.13)	9.75 (8.41)	7.62 (6.52)
Gender, n (%)			
Male	10 (19.2)	7 (15.9)	3 (37.5)
Female	42 (80.8)	37 (84.1)	5 (62.5)
Other	0 (0)	0 (0)	0 (0)
Role, n (%)			
Registered nurse	52 (100)	44 (100)	8 (100)
Leadership role	6 (11.5)	5 (11.4)	1 (12.5)
Practical instructor	6 (11.5)	6 (13.6)	0 (0)
Other	1 (1.9)	0 (0)	1 (12.5)

Primary Outcome

Descriptive Analyses

At t_0 (n=52), documentation was predominantly PC-based (35/52, 67.3%), followed by hybrid use (11/52, 21.2%) and paper (6/52, 11.5%). Among completers at t_1 (n=44), the speech assistant voize was reported as the primary documentation mode by 18/44 (40.9%), PC for 25/44 (56.8%), and paper by 1/44 (2.3%). Usage frequency of the speech assistant was high: 33/44 (75%) reported daily use, 9/44 (20.5%) used the speech assistant 3 - 4 times per week, and 2/44 (4.5%) 1 - 2 times per week.

Overall, around 770 hours of work were observed in long-term care during this study. Total documentation-related time decreased from a mean of 53.91 (SD 23.59) minutes at baseline to 39.10 (SD 16.29) minutes postimplementation in the analysis pooled across 40 multiply imputed datasets (Δ=−14.81 minutes). For the change in total documentation time (post – pre in minutes), complete cases (n=44; see Figure 1) had Q1 –22.48

XSL•FO
RenderX

and Q3 –2.90 (min=−85.90, max=32.60), and the pooled analysis (n=52) had Q1 –23.59 and Q3 –2.54 (min=−86.95, max=32.91).

Figure 1. Baseline to postassessment changes in total documentation time among participants.

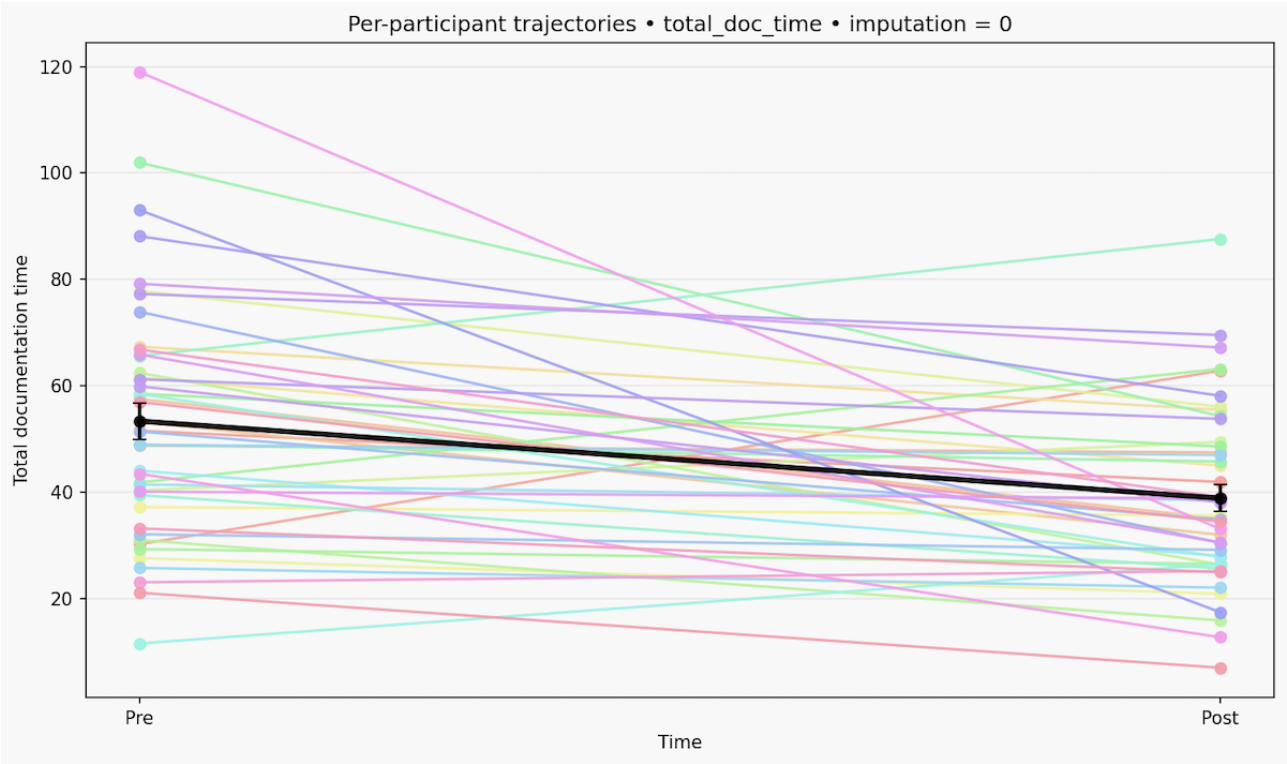

Descriptive reductions were concentrated in active documentation (Δ=−13.28) and note-taking (Δ=−2.38), while reviewing entries (Δ=+0.64) and information retrieval (Δ=+1.28) showed small descriptive increases, consistent with a shift from producing text to checking/retrieving information.

Primary Outcome: Changes in Total Documentation Time

The LMM with random intercepts for participants and facilities and fixed effects for time, age, gender, and baseline documentation type indicated a statistically significant reduction in total documentation time. The adjusted mean change for time (β) was −15 (SE 3.36) minutes, $t_{46.29}$=−4.46, P<.001, with a 95% CI −21.75 to −8.23. Relative to the observed baseline mean of 53.91 minutes, this corresponds to an adjusted 28% reduction in documentation time per morning shift. Variance partitioning indicated an intraclass correlation coefficient of 0.36 at the participant level and 0.08 at the facility level.

Exploratory analyses including time-by-covariate interactions did not indicate reliable heterogeneity of the time effect by age (P=.25), gender (P=.87), or baseline documentation type (P=.49). The estimated reduction was stable across sensitivity analyses, with effect estimates and confidence intervals closely aligned with the primary model. The reduction also remained highly robust under extreme simulated MNAR departures. Detailed results of the exploratory and sensitivity analyses are provided in Multimedia Appendix 5.

Secondary, Self-Reported Outcomes

There was a significant reduction in self-reported documentation time (Δ=−31.14, SE 6.57 minutes), $t_{43.18}$=−4.88, P<.001, 95% CI −45.31 to −18.83. Self-reported daily work interruptions also declined, $t_{47.00}$=−3.95, P<.001, 95% CI −0.95 to −0.31. Participants experienced the documentation process as markedly less time-consuming, $t_{46.91}$=−9.32, P<.001, 95% CI −1.97 to −1.27, and satisfaction with the documentation system increased after the implementation of voize, $t_{46.47}$=4.84, P<.001, 95% CI 1.04 - 2.52. Expectations that the voice assistant would improve documentation quality were also lower after use, $t_{46.82}$=−3.40, P=.001, 95% CI −0.75 to −0.19. All 5 of these effects remained significant after Holm-Bonferroni correction across the 13-test family (Holm-adjusted P≤.013 for each; see Table 2).

XSL•FO
RenderX

Table . Pooled measures of central tendency for secondary, self-reported outcomes[a].

Variables (score range: min-max)	Pooled baseline, mean (SD)	Pooled post, mean (SD)	Mean difference (post – pre)	Median (pre)	Median (post)
Interruptions in daily work (range: 1 - 5)	4.06 (0.78)	3.40 (0.91)	−0.66[b]***	4.00	3.97
Self-reported documentation time (in minutes)	67.05 (40.98)	35.91 (24.34)	−31.14[b]***	60.00	30.00
Documentation is time-consuming (range: 1 - 5)	4.07 (0.79)	2.46 (0.84)	−1.61[b]***	4.00	2.18
Satisfaction with documentation system (range: 1 - 10)	5.07 (2.07)	6.65 (1.82)	+1.58[b]***	5.00	7.00
Expected quality improvement (range: 1 - 5)	3.69 (0.74)	3.18 (0.81)	−0.51[b]**	4.00	3.00
Documentation is complete (range: 1 - 5)	3.27 (0.85)	2.83 (0.94)	−0.44**	3.00	2.87
Expected time savings (range: 1 - 5)	3.81 (0.69)	3.46 (0.79)	−0.35*	4.00	3.57
High documentation quality (range: 1 - 5)	3.38 (0.82)	3.03 (0.75)	−0.35*	3.89	3.00
Implementation as a good idea (range: 1 - 10)	7.91 (1.93)	7.35 (2.04)	−0.56	8.00	7.73
Intention to use (range: 1 - 10)	8.17 (1.99)	7.88 (1.69)	−0.29	8.13	8.00
NWSQ[c]: extrinsic (range: 5 - 25)	17.60 (2.34)	17.95 (2.94)	+0.35	17.51	18.00
NWSQ: intrinsic (range: 6 - 30)	22.21 (3.94)	22.04 (4.42)	−0.17	22.00	21.63
NWSQ: relational (range: 4 - 20)	15.62 (2.62)	15.31 (2.66)	−0.31	16.00	16.00

[a]All values are pooled across 40 multiply imputed datasets according to Rubin rules. Consequently, some median scores are not restricted to the discrete response categories. The significance of the change is indicated by asterisks based on paired *t* tests. *P<.05; **P<.01; ***P<.001.

[b]Effects that remained significant after Holm-Bonferroni correction for 13 tests.

[c]NWSQ: Nursing Workplace Satisfaction Questionnaire.

Perceiving the implementation of voize as a good idea showed no reliable change, $t_{46.29}$=−1.09, P=.28, 95% CI −1.04 to 0.31, and intention to use remained essentially unchanged at a high level, $t_{46.60}$=−0.45, P=.65, 95% CI −0.77 to 0.49. Both remain nonsignificant after correction (Holm-adjusted P≥.99). Ratings that the current documentation is qualitatively high decreased, $t_{46.97}$=−2.30, P=.03, 95% CI −0.61 to −0.04, as did ratings that it is complete and comprehensive, $t_{47.01}$=–2.75, P=.008, 95% CI −0.74 to –0.11, and expected time savings with the speech assistant, $t_{46.74}$=–2.60, P=.01, 95% CI −0.55 to −0.07. However, none of these 3 effects survived Holm-Bonferroni correction (Holm-adjusted P=.16, .06, and .09, respectively).

Nursing workplace satisfaction measured by the NWSQ did not change significantly (extrinsic: $t_{38.94}$=1.48, P=.15, 95% CI −0.22 to 1.45; intrinsic: $t_{39.06}$=0.07, P=.95, 95% CI −1.03 to 1.10; relational: $t_{38.46}$=−0.68, P=.5, 95% CI −0.84 to 0.42), and these remain nonsignificant after correction (Holm-adjusted P=.73, ≥.99, and ≥.99, respectively). The SUS scores indicated acceptable usability [24]: the complete-case mean was 70.42 (SD 14.26; n=40).

Discussion

Principal Findings

This study provides real-world evidence from long-term care that a mobile speech assistant is associated with reduced documentation time over full shifts. The primary LMM estimated an adjusted reduction of approximately 15 minutes per morning shift, corresponding to an approximately 28% adjusted decrease relative to the baseline mean. This finding was robust across extensive sensitivity analyses. The interquartile range indicated that most participants experienced time savings. Task-level patterns suggest that time was recaptured by making active documentation more efficient (see Table 3). However, net time savings may be overestimated as

XSL•FO
RenderX

infection control steps such as disinfection time and adherence to infection control protocols were not systematically recorded in this study. In parallel, nurses reported fewer interruptions, lower perceived time burden, and higher satisfaction with the mobile AI speech assistant. Notably, self-reported retrospective time savings were substantially larger than the observed reduction, and perceived baseline documentation time exceeded our time-motion estimates. This divergence suggests that perceived relief in documentation burden may exceed objective time savings when moving from manual entry to speech-based workflows. Retrospective estimates may apply broader boundaries to what counts as documentation and rely on memory-based summary impressions that overweight salient moments. In this context, cognitively exhaustive, interruption-prone work such as administrative documentation may be remembered as more time-consuming, whereas patient interaction may feel more absorbing and therefore subjectively shorter. By contrast, time-motion observation captures only predefined, observable tasks within the sampled shift, yielding an objective estimate of recorded activities that does not necessarily map onto perceived workload.

Table . Documentation time by task in minutes at baseline and postimplementation.

Task	Pre[a] (minutes), mean (SD)	Post-complete[b] (minutes), mean (SD)	Post-pooled[c] (minutes), mean (SD)
Computer login	1.45 (1.32)	1.05 (1.13)	1.01 (1.09)
Active documentation	44.23 (19.72)	30.44 (14.71)	30.95 (14.36)
Reviewing entries	0.82 (3.65)	1.47 (3.81)	1.46 (3.68)
Information retrieval	3.48 (4.71)	4.85 (6.69)	4.76 (6.48)
Walking to the PC (for documentation)[d]	0.56 (1.56)	0.10 (0.24)	0.09 (0.23)
Taking notes	3.15 (2.68)	0.84 (1.67)	0.77 (1.56)
Waiting	0.22 (0.88)	0.07 (0.33)	0.06 (0.31)
Total documentation-related time	53.91 (23.59)	38.82 (16.75)	39.10 (16.29)

[a]Pre: observed baseline for all participants (N=52).

[b]Post-complete: observed postimplementation among completers (n=44).

[c]Post-pooled: estimates pooled across 40 multiply imputed datasets under “missing at random” assumption (N=52).

[d]“Walking to the PC (for documentation)” refers exclusively to walking carried out to reach stationary ward computers or the ward office for documentation purposes. Walking that occurred while documentation was being performed (eg, dictating with the mobile speech assistant while moving) was not logged as a separate walking task but was counted only as documentation time.

Overall usability was rated as acceptable. Taken together, these findings suggest that introducing speech-enabled, mobile documentation can reduce documentation time while improving perceived documentation efficiency.

At the same time, self-reported documentation quality and completeness showed a consistent negative trend after implementation. In the frequentist exploratory analyses, these decreases were statistically significant before multiplicity correction, but not after Holm-Bonferroni adjustment. However, the complementary Bayesian analysis used adaptive regularization for secondary outcomes and assessed practical relevance using a region of practical equivalence. The Bayesian analysis still assigned a high posterior probability of negative changes in perceived documentation completeness and quality, while the 94% highest density intervals slightly overlapped zero (see Multimedia Appendix 4). Together, these findings suggest a potential trade-off in routine use. Faster, more mobile documentation may coincide with lower perceived completeness or quality, possibly because of recognition errors, editing burden, or shorter entries [16,18,19].

In stationary long-term care, documentation consumes substantial nurse time. A domain-specific mobile speech assistant may reclaim minutes each shift, possibly reduce interruptions, and improve the overall documentation experience. In labor-constrained markets, recovered time can be redeployed to resident care, coordination, supervision, taking necessary breaks, and improving working conditions while lowering overtime [19].

Relation to Prior Work

These findings are consistent with experimental and simulation studies showing that speech-based input improves speed and, in some contexts, accuracy. In nursing, Mayer et al [17] found time savings of 2.3 to 6.1 minutes per scenario without increased errors and high user preference. Similarly, Mairittha et al [18] reported 15% faster entries with a dialogue-based system. In addition, another mobile bedside app reduced EHR charting time and increased uninterrupted patient interaction in a pre-post study [16]. In broader clinical samples, speech systems shortened task completion time and reduced error rates versus paper and traditional patient data management systems in intensive care units [12]. In hospital-wide deployments, they also achieved high adoption with large transcription-cost reductions and perceived gains in quality and efficiency [35]. Observational comparisons have also shown that speech can be substantially faster than typing, even when perceptions may differ [11].

Relative to this body of evidence, this study adds three elements. First, it focuses on long-term care, where evidence has been

XSL•FO
RenderX

sparse and documentation burden is high. Second, it evaluates the technology under routine conditions, not only in controlled or simulated scenarios. Third, it uses full-shift observation at both time points, yielding approximately 770 hours of time-motion data and directly aligned with the workflow realities reported in earlier studies [6,7,36]. Together, these features can help bridge the gap highlighted by reviews that call for end-to-end, real-world evaluations beyond transcription accuracy alone [10,15]. The pattern of decreased active entry, waiting, and walking to the stationary ward computers, with slight increases in retrieval and review (see Table 3), also may show that mobile tools reduce "print-note-retype" loops [16].

Several mechanisms may influence the observed association between speech assistant use and reduced documentation time. First, this is in line with previous qualitative findings of Ferizaj and Neumann [19] that have shown fewer interruptions, less walking to the ward, shorter information-retrieval times, and higher overall satisfaction with the documentation system. In line with this, the mobile, as-you-go entry may have further reduced travel and queuing at ward PCs, cutting waiting, walking to the ward, and repeated logins. Second, domain-tuned speech and natural language processing (ie, models adapted to the nursing domain) may lower the friction of capturing structured information. This may reduce the need for interim notes and subsequent re-entry while also minimizing information loss and documentation errors. Third, workflow coupling at the bedside may decrease interruptions and handovers. The descriptive rise in information retrieval and review may reflect appropriate verification behavior within the mobile app. Usability in the good range [24] and stable acceptance [37] suggest the solution was well received with high intention-to-use and satisfaction, albeit with room for refinement.

Limitations

This study has several limitations. The single-group pre-post design without a control condition limits causal inference and leaves room for workload fluctuations or co-occurring process changes as alternative explanations. According to reports from facility management, no other major documentation tools or workflow redesigns were introduced between t_0 and t_1, which reduces, but does not eliminate, the risk of confounding by concurrent initiatives. Participation was voluntary, so self-selection may have favored early adopters. Additionally, as facilities reported generally stable conditions, we did not collect shift-level workload indicators such as number of residents cared for or staffing ratios for each observed shift. Therefore, documentation time was not normalized for patient census or workload, and residual confounding by day-to-day fluctuations remains possible. Observations were conducted by multiple observers. Although observers underwent standardized training, joint pilot coding, and ongoing weekly calibration meetings, we did not calculate a formal interrater reliability coefficient (eg, Cohen κ) for the final field observations. This limits the extent to which we can quantify measurement consistency across observers. However, an exploratory sensitivity analysis including observer ID as a covariate did not materially change the estimated reduction in documentation time, suggesting that systematic between-observer differences are unlikely to account for the primary findings. Still, this exploratory analysis cannot rule out shared, potentially unconscious observer bias operating in the same direction across time points. Fourth, direct observation can itself change behavior (Hawthorne effect), which we sought to mitigate with standardized procedures but cannot exclude [25]. We did not systematically record time spent on disinfecting mobile devices unless it occurred in direct connection with documentation or monitor adherence to infection control protocols. If device disinfection is performed more consistently in routine use, the net time savings reported here may be overestimated. Documentation quality was captured perceptually rather than via objective measures such as assessing structured completeness or error rates that other work has examined in standardized tasks [12]. We did not conduct independent audits of documentation accuracy (eg, chart reviews, error checks). Reduced documentation time therefore cannot be directly linked to unchanged or improved documentation quality. Furthermore, dropouts were descriptively younger and less experienced than completers. This pattern may limit generalizability for adoption, since retention and continued engagement with the tool may differ by career stage, training status, or role-related workload. Finally, results reflect stationary long-term care in Germany, German-language ASR, and a specific EHR integration. Performance may differ with dialects, accents, terminology, and integration quality [15], and ongoing advances in speech technology may change effect sizes over time.

Future Research

This study complements prior simulation and laboratory research and provides further evaluation of the potential effects of AI speech assistants in long-term care. To better inform workforce policy in the context of nursing shortages, future implementation research should examine whether documentation automation yields sustained workload relief, frees time for direct resident care, coordination, and recovery, and under which organizational and technical conditions these benefits can be realized at scale.

To strengthen causal inference and capture secular trends and learning effects, future studies should use cluster-randomized or stepped-wedge designs with extended follow-up. These evaluations should combine time-motion data with independent, objective assessments of documentation quality (eg, completeness indices, semantic accuracy ratings, and error rates) to determine whether time savings are achieved without compromising data integrity. In addition, studies should assess how saved time is reinvested and whether this translates into improvements in resident outcomes, psychosocial workload, perceived interruptions, staff satisfaction, and retention.

Methodologically, linking full-shift observation with EHR telemetry could quantify sustainability, diurnal patterns, and adoption trajectories. Additionally, further subgroup analyses could explore heterogeneity by accent/dialect, professional experience, unit characteristics, and integration context. Finally, economic evaluations should estimate the total cost of ownership and downstream effects on productivity, overtime, sick leave, and turnover to guide procurement and financing decisions for broader implementation.

XSL•FO
RenderX

Conclusions

To the authors' knowledge, this is the first study to report real-world time savings observed in association with the use of a mobile AI speech assistant in long-term care. In this longitudinal pre-post time-motion study conducted in 14 German long-term care facilities, implementation of a mobile, domain-specific speech documentation system into routine workflows was associated with substantially lower documentation time. Specifically, the mean documentation-related time decreased from about 54 minutes at baseline to approximately 39 minutes postimplementation. In the primary LMM with random intercepts for participants and facilities and fixed effects for time, age, gender, and baseline documentation type, the adjusted mean change was −15 minutes, corresponding to an adjusted 28% reduction relative to the baseline mean. This result remained consistent across sensitivity analyses and MNAR pattern-mixture scenarios. Task-level patterns indicated that time savings were primarily driven by less active documentation and note-taking, while reviewing entries and information retrieval showed small increases. This is consistent with a shift from producing text to verifying and retrieving information [19].

Secondary, exploratory self-report measures converged with the observational findings: participants reported fewer interruptions, lower perceived documentation burden, and higher satisfaction with the documentation system, while nursing workplace satisfaction did not change. Usability was rated as acceptable. Most participants experienced measurable time savings, with approximately one quarter reducing documentation time by more than about 24 minutes per shift. Given the single-group pre-post design and the reliance on self-reported indicators for documentation quality, these results should be interpreted as evidence of an association between implementation and reduced documentation time rather than as causal proof or as objective evidence of changes in documentation quality.

Acknowledgments

We would like to thank the German Federal Ministry of Education and Research as the responsible funding body. In addition, we would like to thank the participating facilities, and especially the nursing staff, who contributed their insights and expertise to this study.

We used the generative AI tool ChatGPT by OpenAI to assist with language refinement and translation of text originally written by the authors, and to help identify potentially relevant scientific literature. All suggested references were independently reviewed and verified by the authors. No generative AI tools were used to generate scientific content, analyses, results, or interpretations.

Funding

This study was conducted as part of the "PYSA – Nursing Documentation With Hybrid Speech Assistant" project, funded by the German Federal Ministry of Education and Research.

Data Availability

The datasets generated and analyzed in this study contain highly granular time-motion logs and contextual information from nursing staff in German long-term care facilities. Owing to data protection regulations, institutional data use agreements, and the terms of the ethics approval and informed consent based on a preregistered study protocol that did not foresee open data sharing, these data cannot be made publicly available. Deidentified aggregate data are reported in the manuscript and appendices. Researchers who wish to access additional deidentified data for scientifically sound purposes may contact the corresponding author; data sharing will be considered on a case-by-case basis and may require additional institutional approvals and data use agreements.

Authors' Contributions

KS contributed to the concept and design of the study, data interpretation, visualization, preparation of the original draft, and review and editing of the manuscript. DF contributed to the study concept and design, data curation, formal analysis, investigation, visualization, preparation of the original draft, and review and editing of the manuscript. SN contributed to the study design, investigation, project administration, interpretation of data, and review and editing of the manuscript. SS-L contributed to project administration, concept and design, funding acquisition, and review and editing of the manuscript. NL contributed to methodological guidance, formal analysis, supervision, funding acquisition, and review and editing of the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest

KS is employed by voize GmbH, which developed the speech assistant evaluated in this study. She was not involved in the development of the system. All other authors declare no competing interests.

Multimedia Appendix 1
Definition of task categories for observations at t_0 and t_1.
[DOCX File, 1766 KB - jmir_v28i1e86078_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Self-developed questionnaire for assessment at t_0 and t_1.
[DOCX File, 1768 KB - jmir_v28i1e86078_app2.docx]

Multimedia Appendix 3
Statistical analysis rationale and detailed methods.
[DOCX File, 1771 KB - jmir_v28i1e86078_app3.docx]

Multimedia Appendix 4
Bayesian supplementary modeling and analysis.
[DOCX File, 2868 KB - jmir_v28i1e86078_app4.docx]

Multimedia Appendix 5
Results of exploratory sensitivity analyses.
[DOCX File, 2317 KB - jmir_v28i1e86078_app5.docx]

References

1. Arbeitsmarktsituation im Pflegebereich [Report in German]. : Statistik der Bundesagentur für Arbeit; 2025 May URL: https://statistik.arbeitsagentur.de/DE/Statischer-Content/Statistiken/Themen-im-Fokus/Berufe/Generische-Publikationen/Altenpflege.html?__blob=publicationFile [accessed 2026-03-13]
2. Pflegekräftevorausberechnung [Website in German]. Statistisches Bundesamt (Destatis). URL: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Bevoelkerungsvorausberechnung/pflegekraeftevorausberechnung.html [accessed 2026-3-13]
3. Wirth T, Ulusoy N, Lincke HJ, Nienhaus N, Schablon A. Psychosoziale Belastungen und Beanspruchungen von Beschäftigten in der stationären und ambulanten Altenpflege [Article in German]. ASU Arbeitsmed Sozialmed Umweltmed 2017;52(9):662-669. [doi: 10.17147/ASU2017-09-01-01]
4. Hasson H, Arnetz JE. Nursing staff competence, work strain, stress and satisfaction in elderly care: a comparison of home-based care and nursing homes. J Clin Nurs 2008 Feb;17(4):468-481. [doi: 10.1111/j.1365-2702.2006.01803.x] [Medline: 17331093]
5. Dall'Ora C, Ejebu OZ, Ball J, Griffiths P. Shift work characteristics and burnout among nurses: cross-sectional survey. Occup Med 2023 May 18;73(4):199-204. [doi: 10.1093/occmed/kqad046]
6. Schenk E, Schleyer R, Jones CR, Fincham S, Daratha KB, Monsen KA. Time motion analysis of nursing work in ICU, telemetry and medical-surgical units. J Nurs Manag 2017 Nov;25(8):640-646. [doi: 10.1111/jonm.12502] [Medline: 28853187]
7. Yen PY, Kellye M, Lopetegui M, et al. Nurses' time allocation and multitasking of nursing activities: a time motion study. AMIA Annu Symp Proc 2018;2018:1137-1146. [Medline: 30815156]
8. Momenipur A, Pennathur PR. Balancing documentation and direct patient care activities: a study of a mature electronic health record system. Int J Ind Ergon 2019 Jul;72:338-346. [doi: 10.1016/j.ergon.2019.06.012] [Medline: 32201437]
9. Gesner E, Dykes PC, Zhang L, Gazarian P. Documentation burden in nursing and its role in clinician burnout syndrome. Appl Clin Inform 2022 Oct;13(5):983-990. [doi: 10.1055/s-0042-1757157] [Medline: 36261113]
10. Joseph J, Moore ZEH, Patton D, O'Connor T, Nugent LE. The impact of implementing speech recognition technology on the accuracy and efficiency (time to complete) clinical documentation by nurses: a systematic review. J Clin Nurs 2020 Jul;29(13-14):2125-2137. [doi: 10.1111/jocn.15261] [Medline: 32243006]
11. Zuchowski M, Göller A. Speech recognition for medical documentation: an analysis of time, cost efficiency and acceptance in a clinical setting. Br J Healthc Manag 2022 Jan 2;28(1):30-36. [doi: 10.12968/bjhc.2021.0074]
12. Peine A, Gronholz M, Seidl-Rathkopf K, et al. Standardized comparison of voice-based information and documentation systems to established systems in intensive care: crossover study. JMIR Med Inform 2023 Nov 28;11:e44773. [doi: 10.2196/44773] [Medline: 38015593]
13. Alboksmaty A, Aldakhil R, Hayhoe BWJ, Ashrafian H, Darzi A, Neves AL. The impact of using AI-powered voice-to-text technology for clinical documentation on quality of care in primary care and outpatient settings: a systematic review. eBioMedicine 2025 Aug;118:105861. [doi: 10.1016/j.ebiom.2025.105861] [Medline: 40694861]
14. Wang J, Lavender M, Hoque E, Brophy P, Kautz H. A patient-centered digital scribe for automatic medical documentation. JAMIA Open 2021 Jan;4(1):ooab003. [doi: 10.1093/jamiaopen/ooab003] [Medline: 34377960]
15. Ng JJW, Wang E, Zhou X, et al. Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: a systematic review. BMC Med Inform Decis Mak 2025 Jul 1;25(1):236. [doi: 10.1186/s12911-025-03061-0] [Medline: 40598136]
16. Ehrler F, Wu DTY, Ducloux P, Blondon K. A mobile application to support bedside nurse documentation and care: a time and motion study. JAMIA Open 2021 Jul;4(3):ooab046. [doi: 10.1093/jamiaopen/ooab046] [Medline: 34345804]

XSL•FO
RenderX

17. Mayer L, Xu D, Edwards N, Bokhart G. A comparison of voice recognition program and traditional keyboard charting for nurse documentation. Comput Inform Nurs 2021 Aug 4;40(2):90-94. [doi: 10.1097/CIN.0000000000000793] [Medline: 34347642]
18. Mairittha T, Mairittha N, Inoue S. Evaluating a spoken dialogue system for recording systems of nursing care. Sensors (Basel) 2019 Jan;19(17):3736. [doi: 10.3390/s19173736]
19. Ferizaj D, Neumann S. Assessing perceptions and experiences of an AI-driven speech assistant for nursing documentation: a qualitative study in German nursing homes. In: Kurosu M, Hashizume A, editors. Human-Computer Interaction, Part V: Springer; 2024:17-34. [doi: 10.1007/978-3-031-60449-2_2]
20. PYSA – Miteinander durch Innovation. Bundesministerium für Forschung, Technologie und Raumfahrt. URL: https://www.interaktive-technologien.de/projekte/pysa [accessed 2026-3-13]
21. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008]
22. Champely S, Ekstrom C, Gill J, et al. pwr: basic functions for power analysis. GitHub. 2020. URL: https://github.com/heliosdrm/pwr [accessed 2026-03-13]
23. Fairbrother G, Jones A, Rivas K. Development and validation of the Nursing Workplace Satisfaction Questionnaire (NWSQ). Contemp Nurse 2009;34(1):10-18. [doi: 10.5172/conu.2009.34.1.010] [Medline: 20230168]
24. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the System Usability Scale. Int J Hum-Comput Interact 2008 Jul 29;24(6):574-594. [doi: 10.1080/10447310802205776]
25. Franke RH, Kaul JD. The Hawthorne experiments: first statistical interpretation. Am Sociol Rev 1978 Oct;43(5):623. [doi: 10.2307/2094540]
26. Pedregosa F, Varoquaux G, Gramfort A, et al. scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12(85):2825-2830. [doi: 10.5555/1953048.2078195]
27. miceforest: Multiple imputation by chained equations with LightGBM. GitHub. URL: https://github.com/AnotherSamWilson/miceforest [accessed 2026-03-13]
28. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. Nature 2020;585:357-362. [doi: 10.1038/s41586-020-2649-2]
29. McKinney W. Data structures for statistical computing in Python. SciPy 2010. [doi: 10.25080/Majora-92bf1922-00a]
30. Jones E, Oliphant T, Peterson P. SciPy: open source scientific tools for Python. 2001. URL: http://www.scipy.org [accessed 2026-03-13]
31. Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng 2007 May;9(3):90-95. [doi: 10.1109/MCSE.2007.55]
32. Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. J Open Source Softw 2019;4(43):1686. [doi: 10.21105/joss.01686]
33. Bates D, Maechler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw 2015;67(1):2-48. [doi: 10.18637/jss.v067.i01]
34. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: tests in linear mixed effects models. J Stat Soft 2017;82(13):1-26. [doi: 10.18637/jss.v082.i13]
35. Saxena K, Diamond R, Conant RF, Mitchell TH, Gallopyn IG, Yakimow KE. Provider adoption of speech recognition and its impact on satisfaction, documentation quality, efficiency, and cost in an inpatient EHR. AMIA Jt Summits Transl Sci Proc 2018;2017:186-195. [Medline: 29888069]
36. Yen PY, Kelley M, Lopetegui M, et al. Understanding and visualizing multitasking and task switching activities: a time motion study to capture nursing workflow. AMIA Annu Symp Proc 2016;2016:1264-1273. [Medline: 28269924]
37. Van Der Laan JD, Heino A, De Waard D. A simple procedure for the assessment of acceptance of advanced transport telematics. Transp Res Part C Emerg Technol 1997 Feb;5(1):1-10. [doi: 10.1016/S0968-090X(96)00025-3]

Abbreviations

AI: artificial intelligence
ASR: automatic speech recognition
EHR: electronic health record
LMM: linear mixed-effects model
MNAR: missing not at random
NWSQ: Nursing Workplace Satisfaction Questionnaire
PYSA: “Pflegedokumentation mit Hybridem Sprachassistenten” (project title)
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
SUS: system usability scale
t_0: baseline measurement
t_1: postimplementation measurement

XSL•FO
RenderX

Edited by A Coristine; submitted 17.Oct.2025; peer-reviewed by K Chen, L He, S Mohanadas, S Ye; revised version received 20.Feb.2026; accepted 24.Feb.2026; published 08.Apr.2026.

Please cite as:
Schwabe K, Ferizaj D, Neumann S, Strube-Lahmann S, Lahmann N
Time Savings Through an AI Speech Assistant for Nursing Documentation: Pre-Post Time-Motion Study in German Long-Term Care
J Med Internet Res 2026;28:e86078
URL: https://www.jmir.org/2026/1/e86078
doi:10.2196/86078

© Katja Schwabe, Drin Ferizaj, Susann Neumann, Sandra Strube-Lahmann, Nils Lahmann. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Medical Consultation and HIV Testing After AI-Based Symptom Check: Retrospective Cohort Study

Nao Taguchi[1], PhD; Kunihiro Hirahara[1], MSc; Keisuke Harada[1], PhD; Keisuke Orimo[2], MD, PhD; KuanYeh Lee[1], MD; Kota Iwahashi[3], MPH; Akifumi Imamura[4], MD

[1]Gilead Sciences K.K., Chiyoda-ku, Tokyo, Japan

[2]Ubie Inc, Chuo-ku, Tokyo, Japan

[3]Nonprofit Organization akta, Shinjuku-ku, Tokyo, Japan

[4]Tokyo Metropolitan Cancer and Infectious Diseases Center Komagome Hospital, 3-18-22 Hon-komagome, Bunkyo-ku, Tokyo, Japan

Corresponding Author:
Akifumi Imamura, MD
Tokyo Metropolitan Cancer and Infectious Diseases Center Komagome Hospital, 3-18-22 Hon-komagome, Bunkyo-ku, Tokyo, Japan

Abstract

Background: Promoting early HIV testing and patient detection is an important public health goal. In Japan, approximately 30% of the population is diagnosed with AIDS. Several studies have investigated the challenges related to HIV diagnosis; however, there are limitations in understanding the characteristics and barriers faced by individuals who are at high risk of HIV but have not yet been tested or have not sought medical consultation.

Objective: This study aimed to examine the factors associated with medical consultation and HIV-testing behaviors, explore the reasons for not undergoing HIV testing, and evaluate the effectiveness of HIV-related awareness efforts among respondents to a revisit survey conducted via an artificial intelligence (AI)–based symptom search engine.

Methods: This retrospective cohort study used data obtained from the AI-based symptom search engine, Ubie. Episodes involving individuals who used the AI-based symptom checker to search for their symptoms, which were subsequently suggested as HIV/AIDS/sexually transmitted infection (STI)–related conditions, were included. Those who answered the first and revisit survey questionnaires were included in the analysis. Multivariable logistic regression analyses were conducted to explore the factors associated with medical consultation in both the overall suggested HIV/AIDS/STI–related condition group and the suggested STI-related condition subgroup. Factors associated with HIV testing in individuals who underwent medical consultations were also explored using multivariable logistic regression analysis. The reasons for not undergoing HIV testing and the future intention to undergo testing were described.

Results: The number of eligible episodes was 424,893 for 332,976 individuals. Of these, medical consultations were performed in 105,365 cases and HIV testing in 394 cases. Compared with individuals in their 20s, older age groups were associated with a higher tendency to seek medical consultations. The provision of awareness information through the AI-based symptom checker was associated with medical consultation behavior, and 29% (280/964) of people who initially had no intention of undergoing HIV testing responded that they would undergo HIV testing after using the AI-based symptom checker. Compared with the internal medicine department, the gynecology department was significantly associated with HIV testing; however, the HIV testing rates were low in the suggested STI-related condition subgroup across major departments.

Conclusions: These results suggest that HIV-related information delivered via an AI-based symptom checker may raise awareness or consideration of medical consultation among individuals actively searching for symptoms potentially associated with HIV. To further promote HIV testing, it may be necessary to refine the content and delivery of educational materials and enhance HIV testing literacy among physicians who encounter patients with STIs.

(J Med Internet Res 2026;28:e90257) doi:10.2196/90257

KEYWORDS

HIV infections; sexually transmitted diseases; artificial intelligence; AI; symptom assessment; delayed diagnosis; patient acceptance of health care; health knowledge, attitudes, practice; Japan; sexually transmitted infection

XSL•FO
RenderX

Introduction

The global epidemic of HIV remains a significant public health issue. In 2024, an estimated 40.8 million people were living with HIV, and there were approximately 1.3 million new HIV infections [1]. The Joint United Nations Program on HIV/AIDS has set the "95-95-95" targets for 2025 with the aims of (1) diagnosing 95% of individuals with HIV, (2) ensuring 95% of diagnosed individuals receive treatment, and (3) achieving viral suppression in 95% of treated individuals [2]. Of these targets, achieving the first "95" target—diagnosing 95% of individuals with HIV—is crucial, as it represents the entry point into the HIV care continuum. The importance of early awareness of one's HIV status has also been emphasized in the Global AIDS Strategy 2021-2026 [3]. Because HIV infection itself does not cause specific symptoms, it is essential to connect people at risk of HIV infection with testing services to improve diagnosis rates before the onset of AIDS.

In Japan, approximately 30% of the population has been first diagnosed at the AIDS stage over the past decade [4,5]. The widespread availability of HIV testing has been facilitated by government- and community-based efforts, including anonymous testing at public health centers and promotional activities conducted by community-based organizations. However, achieving the first "95" target of diagnosis remains a challenge. Although the HIV diagnosis rate steadily increased before the COVID-19 pandemic, a significant decline in both testing and consultation numbers was observed in 2020, the first year of the pandemic. HIV testing numbers have started to increase as the pandemic gradually recedes, and the diagnosis rate reached 89.3% by 2022 [6]. However, HIV testing numbers remain below prepandemic levels [4]. Furthermore, a previous study reported that only approximately 5% of patients who tested for syphilis underwent HIV testing, indicating missed opportunities for early HIV detection [7]. These situations suggest that additional approaches are required to achieve the first "95" target of diagnosis.

Understanding the factors and barriers associated with HIV testing and consultation behaviors is important for improving diagnosis rates. Previous studies in Japan have identified several factors associated with the timing of HIV diagnosis. For instance, a study using data from the Japanese Drug Resistance HIV-1 Surveillance Network found that factors such as age, heterosexual transmission, living outside Tokyo, and hepatitis C virus coinfection were associated with late diagnosis between 2003 and 2019 [8]. Similarly, a retrospective cohort study of Japanese HIV care hospitals found that voluntary testing and being a man who has sex with men were associated with early diagnosis [9]. Furthermore, a study from Japan's largest HIV referral center reported that the reasons for diagnostic tests in newly infected individuals included existing diseases, such as AIDS and sexually transmitted infections (STIs), voluntary testing, and routine presurgery or on-admission screening [10]. Regional differences have also been observed in Japan, with a higher proportion of undiagnosed HIV cases in nonmetropolitan areas [11].

Although previous studies have provided valuable insights into the experiences and challenges of individuals diagnosed with HIV, they have primarily focused on populations already engaged by the health care system. Expanding the scope of inquiry to include individuals at high risk of HIV who have not yet undergone testing or sought medical consultation offers an important opportunity to deepen our understanding of the diverse factors influencing testing behaviors. Recent reports suggest that novel approaches using digital technologies [12], artificial intelligence (AI), and machine learning can efficiently identify at-risk individuals [13], thereby potentially addressing this gap. By leveraging data from AI- and digital-based services that target high-risk populations, researchers can identify and study previously inaccessible high-risk groups.

In this study, we targeted individuals who were using the widely used web service of an AI-based symptom checker in Japan and identified them as the study population who reported symptoms potentially associated with the risk of HIV and responded to the revisit survey. The aim of this study was to examine the factors associated with medical consultation and HIV testing behaviors among these revisit survey respondents to better understand the challenges of HIV testing. We also explored the reasons for not undergoing HIV testing and evaluated the effectiveness of HIV-related awareness efforts.

Methods

Study Design and Data Source

This retrospective cohort study used data obtained from an AI-based symptom search engine, Ubie [14]. The symptom checker evaluated in this study is publicly available. Most users access it by conducting ordinary web searches for their symptoms (eg, Google or other search engines), which directly lead them to the publicly accessible Ubie symptom checker page. This service allows users to answer an AI-optimized adaptive question flow in a step-by-step format based on their symptoms and provides information on related conditions, recommended actions, appropriate medical departments, and health care facilities. The AI medical interview assistance system alternates question selection and identifies relevant diseases using a flowchart model regarding symptoms and diseases based on information from tens of thousands of medical reports. This service has approximately 10 million monthly users. Individuals experiencing symptoms access the symptom search engine through web searches or Ubie applications. In this AI-based symptom search engine, individuals who agreed to the service's privacy policy responded to questions from the AI-based symptom checker. The results displayed potentially related conditions, along with the recommended health care facilities (Figure S1 in Multimedia Appendix 1). When individuals revisit the search engine after completing the initial questionnaire, they are asked to answer follow-up questionnaires regarding the health care consultations conducted since their initial visit.

Three types of data sources were used. First, survey data were collected from individuals who visited Ubie's website, responded to symptom-related questions, and received presentations of potentially related conditions. The data sources included information on age, sex, comorbidities, treatment

XSL•FO
RenderX

status, current symptoms, pain scores, and possible conditions. Second, revisit survey data were gathered, which included responses from individuals who revisited the website after completing the first survey, including data on their medical consultations and testing behaviors after the first survey. The revisit survey responses were automatically linked to the latest symptom check episode using an anonymized user identifier assigned within the platform. This data source included information on whether a medical examination or test was carried out, the type and area of the medical facility and department visited for consultation or testing, diagnostic results, and whether HIV testing was conducted. Third, HIV-specific banner questionnaire data were collected, which included responses to HIV-specific questions from individuals who completed a questionnaire accessed via banner advertisements on Ubie's website during the first survey. This data source included information on reasons for not undergoing HIV testing and future intentions for HIV testing. The study period was from October 19, 2022, to December 27, 2023.

Study Population

Episodes from individuals were included if they met the following criteria: (1) individuals for whom the AI-based symptom checker showed any of the following 9 potentially related conditions in the first survey data: HIV infection (HIV infection or secondary immunodeficiency disorder) or an STI (trichomoniasis, Chlamydia infection, genital herpes, condyloma acuminatum, gonorrhea, amoebiasis, or syphilis); (2) completion of the revisit survey after the first survey. Individuals were excluded if they answered in the first survey that they were searching for symptoms in someone other than themselves or had been previously diagnosed with HIV infection or secondary immunodeficiency disorder. For individuals with multiple first survey data points during the study period, each data point meeting the above criteria was included to assess the background of the individuals immediately before their medical consultation and testing behavior.

Outcome Variables

The primary outcome was receiving a medical consultation after answering the AI-based symptom checker, as obtained from the revisit survey data. The secondary outcome was HIV testing among individuals who visited medical care facilities after answering the AI-based symptom checker, which was also obtained from the revisit survey data. A positive HIV test result was obtained from the revisit survey data. Reasons for not undergoing HIV testing and future intentions to undergo testing were evaluated based on HIV-specific banner questionnaire responses.

Characteristics

To explore the factors associated with the primary and secondary outcomes, several characteristic variables were assessed from the first and revisit survey data. For the first survey data, the variables were age, sex, main symptoms, pain score, STI comorbidities or medical histories, displayed suggested condition, device used to access the website (PC or mobile device), access to HIV awareness information that was displayed during the study period in connection with 9 previously mentioned potential conditions, and survey period. For the revisit survey data, the variables included diagnoses in medical facilities (HIV risk diagnosis in medical facilities, area of medical facilities, facility where the test was conducted, department visited, and place where the HIV test was conducted).

Statistical Analysis

In this study, the AI-based symptom checker adapted subsequent questions based on individual responses. Therefore, certain questions were presented to only some individuals, resulting in missing data for others. In such cases, a "Do not be asked" category was assigned, and imputation or other missing data handling techniques were not applied.

A flow diagram showing the number of individuals at each stage of the selection process was created. Patient characteristics were summarized according to visit status and HIV testing status. Continuous variables were summarized using means and SDs, or medians and interquartile ranges. Tables with frequencies and percentages were generated for categorical variables. Stratified subgroup analyses were performed.

Multivariate logistic regression analysis was conducted to explore the factors associated with medical consultations. The crude and adjusted odds ratios (ORs) and 95% CIs were estimated. A forest plot of the ORs was also shown. This analysis was conducted in 2 groups: the overall HIV/AIDS/STI–related suggested condition group (suggested HIV/AIDS/STI group) and the STI-related suggested condition subgroup (suggested STI group). The suggested STI group comprised individuals whose symptoms were associated with STI-related suggested conditions. Because HIV/AIDS-related symptom patterns may occur simultaneously in the same individuals who present with STI-related symptoms, the suggested STI group could include individuals who also exhibited HIV/AIDS-related symptom patterns. In this analysis, the suggested STI group was, therefore, treated as a subgroup within the broader suggested HIV/AIDS/STI group. Similarly, factors associated with HIV testing among individuals who received medical consultations were explored using multivariate logistic regression analysis.

To explore HIV-specific information among those who responded to the HIV-specific banner questionnaire, summary statistics of the reasons for not having an HIV test and future intentions to undergo testing were calculated.

All statistical analyses were conducted using Python (version 3.11; Python Software Foundation) and the *statsmodels* library (version 0.14.4).

Ethical Considerations

This study was approved by the ethics review committee of the Japan Physicians Association (approval 052-2503-02; approved on May 13, 2025).

Results

The database included 38,934,125 episodes based on the data available from October 19, 2022, to December 27, 2023. Of these, 424,893 eligible episodes from 332,976 individuals

XSL•FO
RenderX

(Multimedia Appendix 1) were analyzed (Figure 1). Among these episodes, medical consultations and HIV testing were performed in 105,365 and 394 cases, respectively. In the subset of episodes with suggested STIs, medical consultations and HIV testing were performed in 74,892 and 280 cases, respectively (Figure 1). Patient characteristics for both episodes and individuals are summarized in Table S1 in Multimedia Appendix 1 and were similar across all episodes and individuals.

Figure 1. Flow diagram. AI: artificial intelligence; STI: sexually transmitted infection.

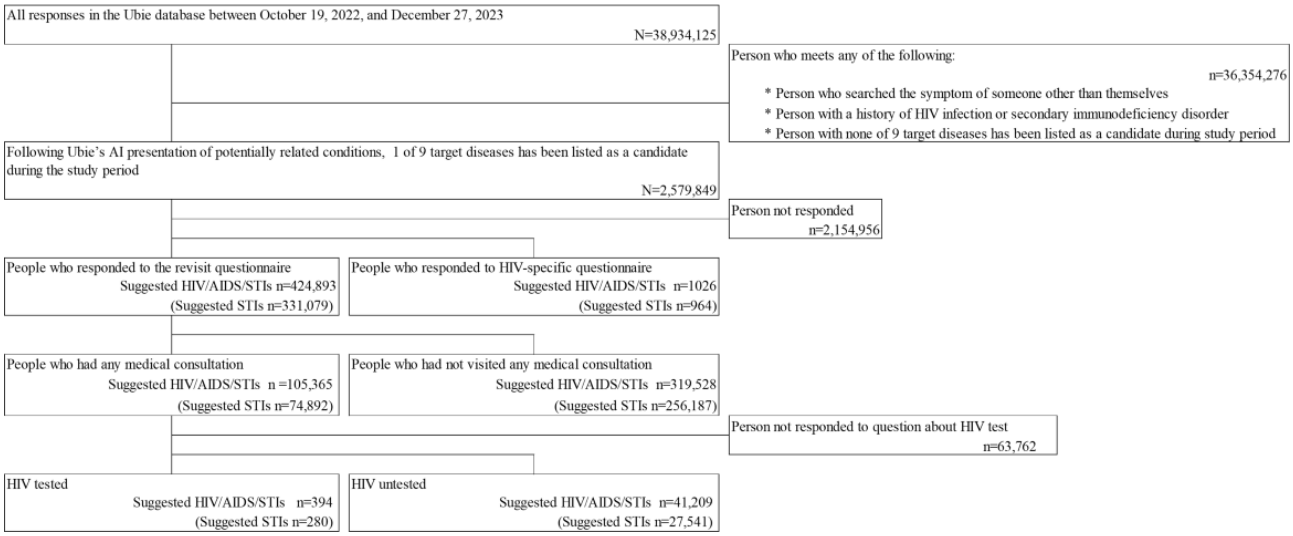

For the suggested HIV/AIDS/STI group, the median age was 37 (IQR 27-48) years in the group that had a medical consultation and 31 (IQR 22-44) years in the group that did not (Table 1). For the suggested STI group, the median age was 32 (IQR 25-43) years for those who had a medical consultation and 28 (IQR 21-39) years for those who did not. Across all groups, the age category with the highest proportion of episodes was the 20s. Across all groups, 76.2% (80,315/105,365) to 78.5% (201,040/256,187) of the episodes were reported by female users. Genital symptoms were the most frequent main symptoms, reported in 35.8% (37,767/105,365) to 50.4% (37,733/74,892) of episodes. Pain scores were reported as categories "1 to 3" in 10.2% (7605/74,892) to 18.4% (58,696/319,528) of episodes, as categories "4 to 6" in 10.1% (10,613/105,365) to 11.8% (30,196/256,187) of episodes, and as categories "7 to 10" in 3.4 % (10,920/319,528) to 4.9% (3703/74,892) of episodes. The remaining episodes were not assessed for pain because the symptom checker did not determine whether it was necessary to evaluate a possible condition. Over 90% (99,228/105,365; 301,664/319,528; 71,128/74,892; 243,973/256,187) of episodes were accessed via mobile phones. Access to the information page containing HIV educational material was observed in 19.3% (20,344/105,365) to 22.9% (17,127/74892) of the episodes. However, access to the HIV testing information page via the link provided in the educational material was observed in only 0.1% (312/256,187) to 0.2% (175/105,365) of the episodes. The proportion of survey periods was similar across groups.

XSL•FO
RenderX

Table . Patient characteristics of the suggested HIV/AIDS/STI group and the suggested STI group.

	Suggested HIV/AIDS/STI[a] group		Suggested STI group[b]	
	Had a medical consultation (n=105,365)	Did not have a medical consultation (n=319,528)	Had a medical consultation (n=74,892)	Did not have a medical consultation (n=256,187)
Age (y), median (IQR)	37 (27-48)	31 (22-44)	32 (25-43)	28 (21-39)
Age group (y), n (%)				
0-19	6617 (6.3)	46,538 (14.6)	5981 (8.0)	44,447 (17.3)
20-29	28,080 (26.7)	100,091 (31.3)	25,926 (34.6)	95,526 (37.3)
30-39	23,302 (22.1)	64,007 (20.0)	18,382 (24.5)	53,958 (21.1)
40-49	23,281 (22.1)	58,164 (18.2)	14,700 (19.6)	39,746 (15.5)
50-59	16,024 (15.2)	34,992 (11.0)	7550 (10.1)	17,967 (7.0)
≥60	8061 (7.7)	15,736 (4.9)	2353 (3.1)	4543 (1.8)
Sex, n (%)				
Male	25,036 (23.8)	72,533 (22.7)	17,006 (22.7)	55,133 (21.5)
Female	80,315 (76.2)	246,970 (77.3)	57,875 (77.3)	201,040 (78.5)
Main symptoms, n (%)[c]				
Genital symptoms	37,767 (35.8)	124,830 (39.1)	37,733 (50.4)	124,778 (48.7)
Symptoms in the ears, nose, throat, eyes, and oral cavity	18,128 (17.2)	49,912 (15.6)	5472 (7.3%)	21,188 (8.3)
Skin and subcutaneous symptoms	10,564 (10.0)	29,485 (9.2)	8089 (10.8)	24,087 (9.4)
Urination-related symptoms	8271 (7.8)	22,960 (7.2)	8271 (11.0)	22,960 (9.0)
General symptoms	7910 (7.5)	14,972 (4.7)	534 (0.7)	1549 (0.6)
Musculoskeletal and joint symptoms	7573 (7.2)	30,543 (9.6)	6259 (8.4)	26,601 (10.4)
Abdominal digestive-related symptoms	7013 (6.7)	26,022 (8.1)	6160 (8.2)	23,785 (9.3)
Pain score, n (%)				
1 - 3	19,334 (8.3)	58,696 (18.4)	7605 (10.2)	34,134 (13.3)
4 - 6	10,613 (10.1)	37,146 (11.6)	7583 (10.1)	30,196 (11.8)
7 - 10	4663 (4.4)	10,920 (3.4)	3703 (4.9)	9097 (3.6)
Not asked	70,755 (67.2)	212,766 (66.6)	56,001 (74.8)	182,760 (71.3)
Device used to access the website (PC or mobile), n (%)				
Mobile	99,228 (94.2)	301,664 (94.4)	71,128 (95.0)	243,973 (95.2)
PC	6133 (5.8)	17,852 (5.6)	3760 (5.0)	12,202 (4.8)
Access to HIV awareness information				
Not open	84,846 (80.5)	256,439 (80.3)	57,641 (77.0)	199,814 (78.0)
Access to the information page on possible diagnoses	20,344 (19.3)	62,694 (19.6)	17,127 (22.9)	56,061 (21.9)
Access to the information page on HIV testing	175 (0.2)	395 (0.1)	124 (0.2)	312 (0.1)
Survey period				
October 19, 2022, to December 31, 2022	16,794 (15.9)	57,568 (18.0)	16,731 (22.3)	57,472 (22.4)
January 1, 2023, to June 30, 2023	45,496 (43.2)	134,301 (42.0)	31,924 (42.6)	106,793 (41.7)

XSL•FO
RenderX

	Suggested HIV/AIDS/STI[a] group		Suggested STI group[b]	
	Had a medical consultation (n=105,365)	Did not have a medical consultation (n=319,528)	Had a medical consultation (n=74,892)	Did not have a medical consultation (n=256,187)
July 1, 2023, to December 27, 2023	43,075 (40.9)	127,659 (40.0)	26,237 (35.0)	91,922 (35.9)

[a]STI: sexually transmitted infection.

[b]Population with suggested STIs as possible conditions.

[c]Symptoms experienced by >5% of visitors were listed.

Approximately 7% (3044/41,603) of episodes in the suggested HIV/AIDS/STI group and 10% (3037/27,821) of episodes in the suggested STI group were diagnosed as actual STIs (Table S2 in Multimedia Appendix 1). Individuals who underwent HIV testing were more likely to have a history of STIs or comorbidities associated with them. Overall, the most frequently visited departments were internal medicine, gynecology, and urology. Among those who underwent HIV testing, more than 50% (373/674) were tested at a hospital. Access to HIV-awareness information ranged from 19.8% (8156/41,209) to 33.6% (94/280) across all groups.

In the suggested STI group, the proportion of men in their 20s was the lowest compared with the other subgroups (Table S3 in Multimedia Appendix 1). Men who have sex with men and individuals with a history of STIs showed a higher proportion of genital symptoms than the general male and female groups. The prevalence of these symptoms among men who have sex with men was 71.7% (370/516) for those who had a medical consultation and 73.3% (1211/1653) for those who did not. The prevalence of these symptoms was 80.7% (5911/7327) and 80.8% (16,438/20,351), respectively, in individuals with a history of STIs. In contrast, the prevalence in the general male and female groups was 35.2% (5978/17,006) and 33.0% (18,177/55,133) and 54.9% (31,749/57,875) and 53.0% (106,596/201,040), respectively. Similarly, the proportion of individuals with urinary symptoms was higher in the high-risk subgroups (men who have sex with men and individuals with a history of STIs). Additionally, access to HIV awareness information was higher in these subgroups than in the general male and female subgroups. Among the suggested STI groups, 10.9% (3037/27821) were actually diagnosed with an STI. In this group, 9.2% (280/3037) underwent an HIV test, and among those tested, 4.6% (13/280) tested positive (Table S4 in Multimedia Appendix 1).

In the suggested STI group, compared with individuals in their 20s, older age groups were significantly more likely to seek medical consultation, with ORs of 1.26 (95% CI 1.24 - 1.29) for those aged 30 to 39 years, 1.38 (95% CI 1.35 - 1.41) for those aged 40 to 49 years, 1.58 (95% CI 1.54 - 1.63) for those aged 50 to 59 years, and 1.98 (95% CI 1.88 - 2.09) for those aged 60 years and above (Figure 2). Compared with the low pain score categories (1-3), the higher pain score categories (7-10) were more strongly associated with medical consultation, with an OR of 1.71 (95% CI 1.64 - 1.80). Accessing the information page on possible diagnoses, which included HIV awareness content, was positively associated with health care–seeking behaviors, with an OR of 1.12 (95% CI 1.09 - 1.14). Accessing the HIV testing page via the link in the HIV awareness section also showed a significant association, with an OR of 1.41 (95% CI 1.14 - 1.74). Compared with the 2022 survey period, the 2023 survey period was positively associated with seeking medical consultation, with ORs of 1.09 (95% CI 1.07 - 1.11) for January 2023 to June 2023 and 1.06 (95% CI 1.03 - 1.08) for July 2023 to December 2023. Similar results were observed in the HIV/AIDS/STI group (Figure 3). In this group, females were more likely to seek medical consultation than males, with an OR of 1.04 (95% CI 1.02 - 1.06).

XSL•FO
RenderX

Figure 2. Factors associated with medical consultation (suggested STI group). Forest plot illustrates adjusted odds ratios (ORs) and 95% CIs for medical consultations associated with various factors. The vertical dashed line (OR=1.0) represents the reference line. Blue circles represent *P*<.05, and gray circles represent *P*≥.05. The size of the marker is proportional to the sample size for each category. STI: sexually transmitted infection.

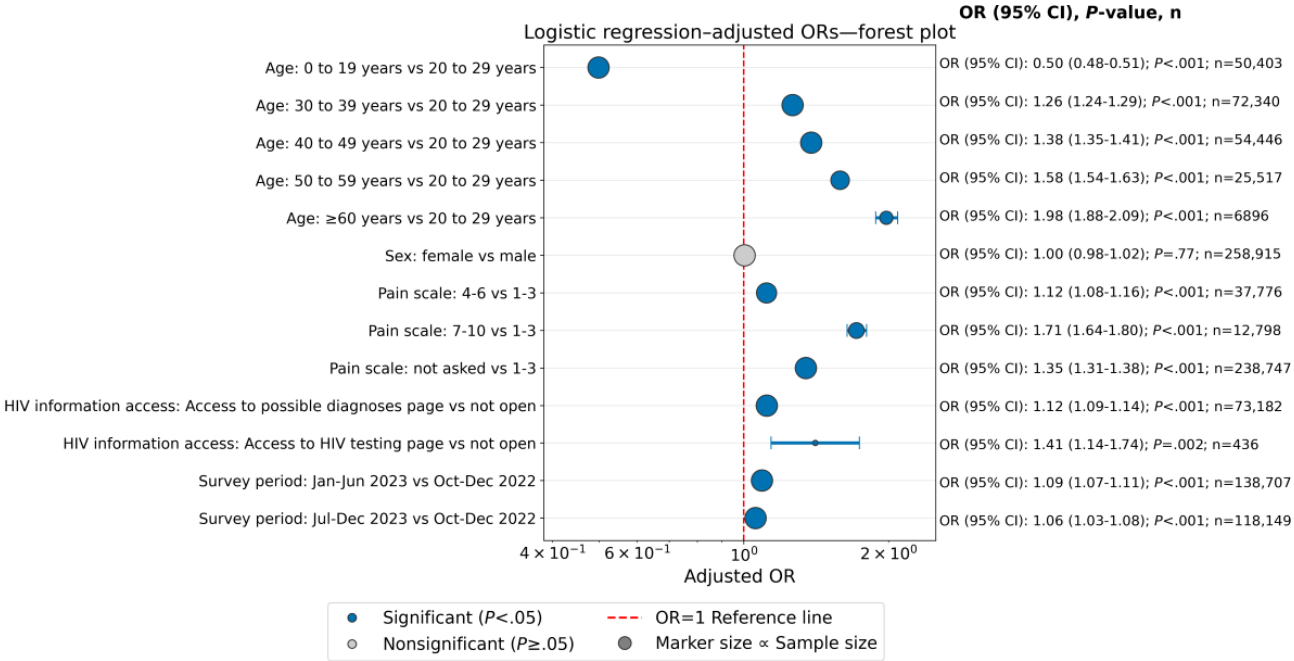

Figure 3. Factors associated with medical consultation (suggested HIV/AIDS/STI group). Forest plot illustrates the adjusted odds ratios (ORs) and 95% CIs for medical consultations associated with various factors. The vertical dashed line (OR=1.0) represents the reference line. Blue circles represent a *P*<.05, and gray circles represent a *P*≥.05. The size of the marker is proportional to the sample size for each category. STI: sexually transmitted infection.

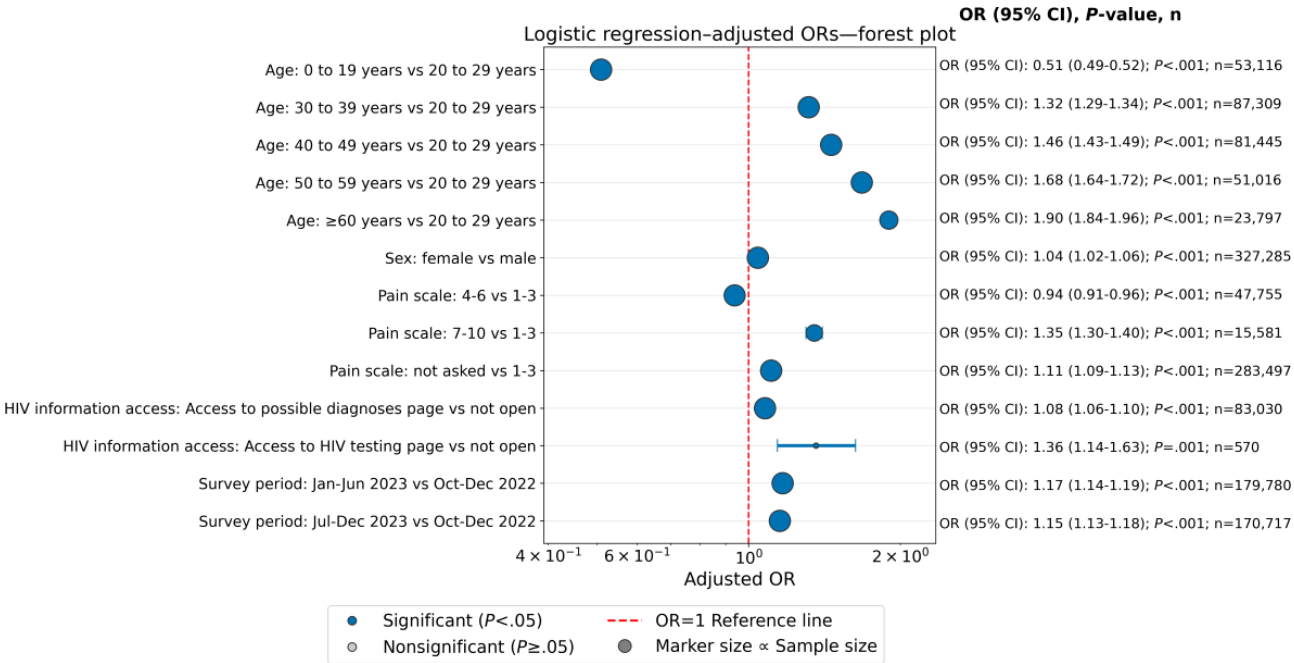

The forest plot illustrates adjusted ORs and 95% CIs for medical consultations associated with various factors. The vertical dashed line (OR=1.0) represents the reference line. Blue circles represent *P*<.05, and gray circles represent *P*≥.05. The size of the marker is proportional to the sample size for each category.

The forest plot illustrates the adjusted ORs and 95% CIs for medical consultations associated with various factors. The vertical dashed line (OR=1.0) represents the reference line. Blue circles represent a *P*<.05, and gray circles represent a *P*≥.05. The size of the marker is proportional to the sample size for each category.

In contrast to the results of medical consultation behavior, the 30 to 59 age category was negatively associated with HIV testing behavior compared with the 20 - 29 age category in the suggested STI group, with ORs of 0.27 (95% CI 0.14 - 0.52) for those aged 30 to 39 years, 0.19 (95% CI 0.09 - 0.42) for those aged 40 to 49 years, and 0.33 (95% CI 0.15 - 0.73) for those aged 50 to 59 years (Figure 4). No significant associations

XSL•FO
RenderX

were observed between HIV testing and factors, such as sex, pain score, or exposure to awareness information. Similarly, neither the type of facility visited nor the geographical area of the medical facility was associated with HIV-testing behavior. Compared with internal medicine departments, gynecology departments were significantly associated with HIV testing, with an OR of 2.44 (95% CI 1.10 - 5.41). In the suggested HIV/AIDS/STI group, the urology department was negatively associated with HIV testing, with an OR of 0.37 (95% CI 0.18 - 0.73; Figure 5).

Figure 4. Factors associated with HIV testing (suggested STI group). The forest plot illustrates the adjusted odds ratios (ORs) and 95% CIs for HIV testing associated with various factors. The vertical dashed line (OR=1.0) represents the reference line. Blue circles represent *P*<.05, and gray circles represent *P*≥.05. The size of the marker is proportional to the sample size for each category. STI: sexually transmitted infection.

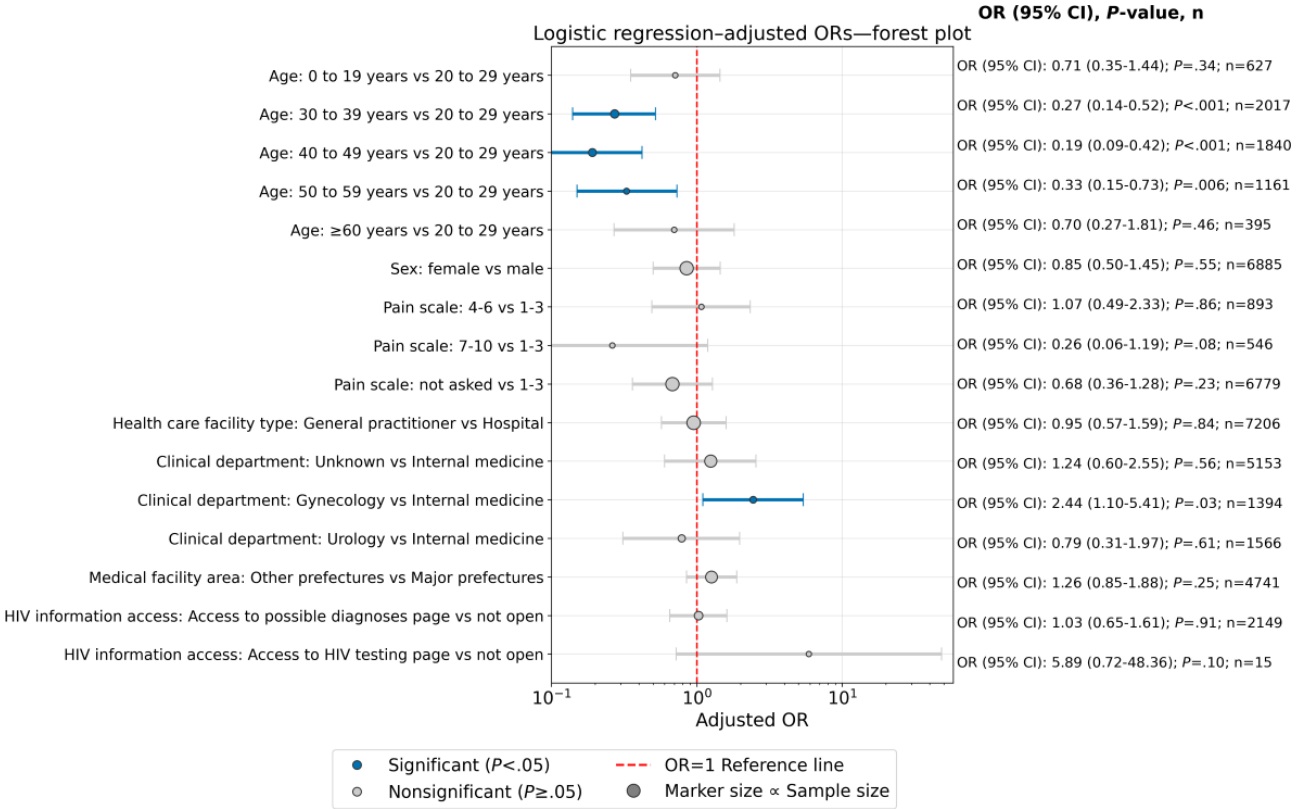

XSL•FO
RenderX

Figure 5. Factors associated with HIV testing (suggested HIV/AIDS/STI group). The forest plot illustrates the adjusted odds ratios (ORs) and 95% CIs for HIV testing associated with various factors. The vertical dashed line (OR=1.0) represents the reference line. Blue circles represent *P*<.05, and gray circles represent *P*≥.05. The size of the marker is proportional to the sample size for each category. STI: sexually transmitted infection.

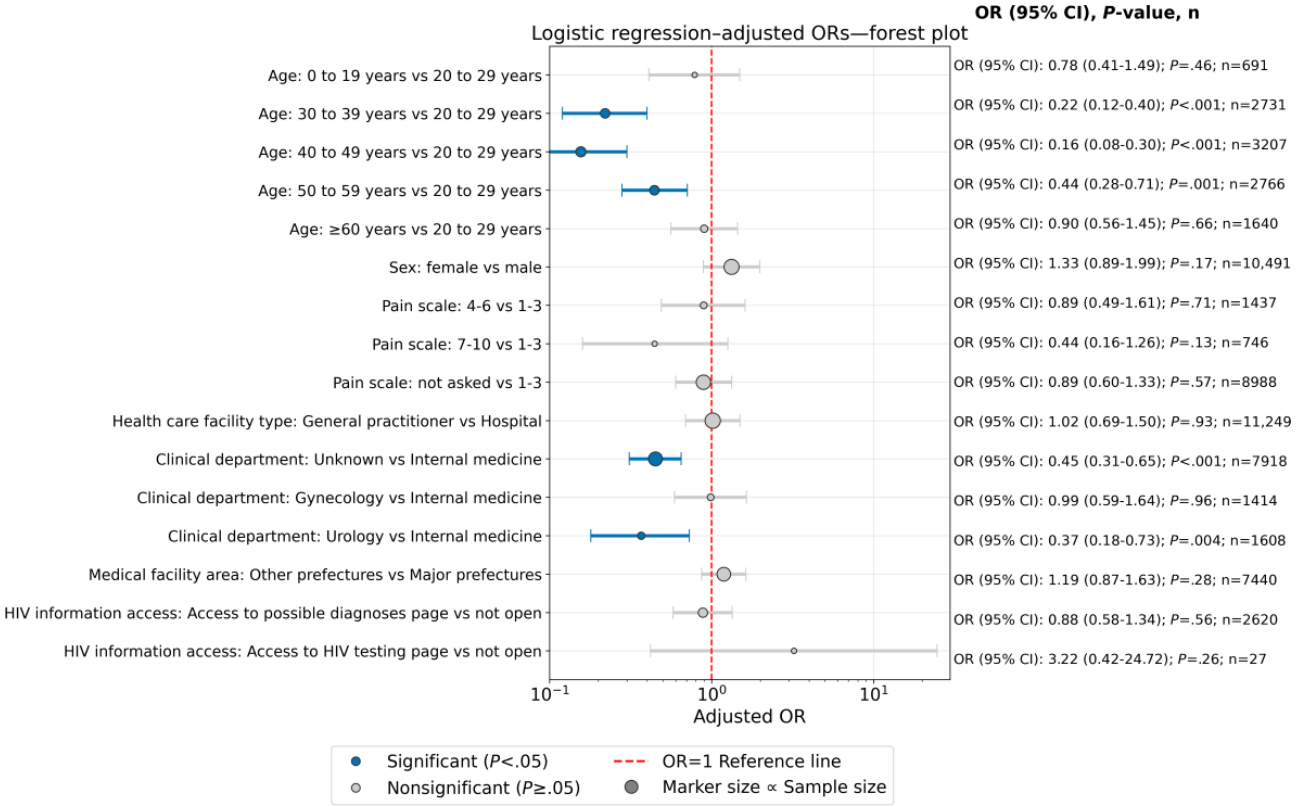

The forest plot illustrates the adjusted ORs and 95% CIs for HIV testing associated with various factors. The vertical dashed line (OR=1.0) represents the reference line. Blue circles represent *P*<.05, and gray circles represent *P*≥.05. The size of the marker is proportional to the sample size for each category.

The forest plot illustrates the adjusted ORs and 95% CIs for HIV testing associated with various factors. The vertical dashed line (OR=1.0) represents the reference line. Blue circles represent *P*<.05, and gray circles represent *P*≥.05. The size of the marker is proportional to the sample size for each category.

The most frequent reason selected for not undergoing an HIV test was "I did not feel it was necessary," with a proportion of 35.3% (54/153) to 49.6% (59/119) across the subgroups. High-risk populations with a history of STIs showed a lower proportion of this response than other subgroups (Table 2). Additionally, 11.3% (109/964) of the respondents did not know where to get tested, and 4.9% (47/964) were unaware of the availability of HIV testing. The proportion of those who felt barriers to receiving an HIV test was 4.2% (35/838) for females and 0.8% (1/119) for males. Concerns about the cost of HIV testing were reported by 5.4% (52/964) of respondents.

XSL•FO
RenderX

Table . Reasons for not undergoing an HIV test among the suggested STI[a] group.

	All (n=964), n (%)	Male (n=119), n (%)	Female (n=838), n (%)	STI[a] history (positive; n=153), n (%)
I did not feel it was necessary	437 (45.3)	59 (49.6)	374 (44.6)	54 (35.3)
I did not have time to get an HIV test	81 (8.4)	14 (11.8)	66 (7.9)	14 (9.2)
I did not know where to get an HIV test	109 (11.3)	9 (7.6)	100 (11.9)	22 (14.4)
I did not know an HIV test was available	47 (4.9)	5 (4.2)	42 (5.0)	8 (5.2)
I felt there were barriers to getting an HIV test	36 (3.7)	1 (0.8)	35 (4.2)	4 (2.6)
I thought it would be expensive to get an HIV test	52 (5.4)	5 (4.2)	47 (5.6)	11 (7.2)
Others	254 (26.3)	31 (26.1)	221 (26.4)	51 (33.3)

[a]STI: sexually transmitted infection.

The intention to undergo an HIV test was positively influenced in 71.0% (685/964) of the individuals, and 29.0% (280/964) of those who originally had no intention of being tested changed their minds and decided to undergo an HIV test (Table 3). The male population had the highest proportion of individuals who had never been tested for HIV infection (43/119, 36.1%).

Table . Intention to undergo HIV testing after an AI[a]-based symptom check in the suggested STI[b] group.

	All (n=964), n (%)	Male (n=119), n (%)	Female (n=838), n (%)	STI[a] history (positive; n=153), n (%)
Did you change your mind about the HIV test after AI[a] algorithm–based information provision?				
Never intended to get an HIV test	259 (26.9)	43 (36.1)	212 (25.3)	37 (24.2)
Initially no intention, but now I understand the importance, but I will not undergo an HIV test	236 (24.5)	22 (18.5)	213 (25.4)	39 (25.5)
Initially no intention, but now I will undergo an HIV test	280 (29.0)	31 (26.1)	248 (29.6)	42 (27.5)
Intention to undergo an HIV test has increased	169 (17.5)	22 (18.5)	146 (17.4)	35 (22.9)
Others	20 (2.1)	1 (0.8)	19 (2.3)	0 (0.0)

[a]AI: artificial intelligence.

[b]STI: sexually transmitted infection.

A total of 20.8% (200/964) of the participants were aware of HIV postal testing; however, only 1.6% (15/964) had experience in using it (Table S5 in Multimedia Appendix 1). A total of 18.7% (180/964) of individuals answered that they did not intend to receive a test but would be willing to use HIV postal testing. The distribution of answers was generally similar across male, female, and high-risk populations (those with a history of STIs). The men who have sex with men subgroup was excluded from the analysis due to the small number of patients (n=8).

Discussion

Principal Findings

Our study analyzed factors associated with subsequent medical consultation and HIV testing among users who were identified by an AI-based symptom checker as having symptoms potentially related to HIV risk. We found that older age, higher pain severity, and access to HIV-related awareness information were associated with a greater likelihood of seeking medical consultation. Although HIV testing was more frequently performed in gynecology departments than in internal medicine, testing rates remained low across all major clinical departments, particularly in urology. Use of the AI-based symptom checker

XSL•FO
RenderX

was also associated with an improvement in users' intention to undergo HIV testing. These findings suggest the importance of improving access to and optimizing HIV-related awareness information for at-risk individuals, together with strengthening HIV testing practices within clinical settings.

Benefit of Approaching This Population

Using data from a widely used AI-based symptom checker in Japan, this study identified key populations that should be prioritized for HIV testing. Among the 420,000 episodes analyzed, approximately 330,000 involved suggested STIs, which are critical for the early detection of HIV infection. Of these, confirmed STI diagnoses were reported in approximately 10% (3037/331,079) of individuals who sought medical care, with a total of approximately 3000 cases. Considering the possibility that a certain number of at-risk individuals existed among those who did not report their consultation outcomes, it is likely that the system effectively reached key populations. In addition, men who have sex with men and individuals with a history of STIs showed a higher prevalence of genital and urinary symptoms than the general male and female subgroups. Furthermore, the proportion of access to awareness materials was higher among high-risk populations. These findings suggest that within these high-risk populations, individuals in need of HIV testing may have been more accurately identified than those in other subgroups. Previous studies have reported that being a man who has sex with men is associated with an early HIV diagnosis [8,9]. Higher engagement with awareness materials among these groups may reflect the greater relevance of the content to their specific risks.

This study is also noteworthy for successfully addressing a previously underexplored demographic gap and providing valuable insights. First, it includes users from rural areas, including those outside government-designated cities, at levels comparable to or even exceeding those from urban regions. Previous reports have suggested that promoting HIV testing and improving diagnostic rates in rural areas are key public health challenges in Japan [8], as diagnosis rates tend to be higher in urban areas and lower in rural regions. This suggests that this study may have successfully reached populations with a greater need for awareness and promotion of HIV testing. Second, this study included a high proportion of females in the high-risk population. However, in Japan, women are a relatively overlooked group in HIV diagnosis, accounting for only 3.5% (19/539) of newly diagnosed HIV cases in recent years [4]. Consequently, there is a prevailing perception that HIV is not relevant to women, and awareness campaigns tend to exclude them. This study successfully identified a previously underrepresented population as well as possible HIV-positive cases.

Factors Associated With Medical Consultation and HIV Testing (Except Awareness Activity)

This study provides insights into several factors associated with medical consultations and HIV testing behaviors. Individuals in their 20s showed a lower tendency to seek medical consultations than older age groups. A previous study investigating factors related to late HIV diagnosis reported that older age is associated with late diagnosis [8], which contrasts with the findings of our study. This difference is likely because these studies evaluated late diagnoses based on CD4 counts at the time of HIV diagnosis, whereas our study assessed behavior changes following recommendations from an AI-based symptom checker. In contrast to medical consultation behavior, a higher proportion of individuals in their 20s underwent HIV testing after medical consultation compared with older age groups. This suggests that, compared with other age groups, individuals in their 20s are more likely to undergo HIV testing once they seek medical care. Given that new HIV infections are most frequently reported among individuals in their 30s, followed by those in their 20s, promoting health care–seeking behaviors among people in their 20s is particularly important [4]. Strengthening HIV screening for individuals in their 20s who present with the suggested STIs is crucial, and the key challenge lies in effectively encouraging this population to seek medical attention. For cases with suggested STIs, HIV testing was more frequent in the gynecology department than in the internal medicine department, whereas no significant differences were observed between the urology and internal medicine departments. This may be because HIV screening is routinely conducted in all pregnant women in Japan, which could contribute to a lower barrier to testing in gynecology departments. However, the testing rates among the suggested STI groups were low across all specialties, remaining at 0.92% (9/975) in internal medicine, 1.75% (54/3090) in gynecology, and even lower in urology (0.19%, 3/1576). In Japan, several practical barriers may contribute to low HIV testing in STI-related clinics, including urology. First, physicians often find it burdensome to explain and recommend HIV testing to patients who do not request it themselves. Second, because free and anonymous HIV testing is available at public health centers, which also handle posttest procedures, clinicians have limited incentive to perform testing within their own facilities. Third, awareness of insurance coverage for HIV testing in STI-related contexts remains low, and concerns about reimbursement or claim denial persist [15]. In addition, in urology practice, genital symptoms often lead to a relatively clear presumptive diagnosis and prompt treatment of specific STIs, which may result in clinical focus being placed on immediate disease management rather than on broader risk assessment, including HIV testing. This highlights the need for renewed awareness among urology health care providers that the presence of STIs should also prompt consideration of concurrent HIV testing. Barriers to HIV testing need to be addressed and improved.

This study provides a detailed investigation of the barriers to HIV testing among a subset of users who responded to an HIV-specific banner questionnaire. One of the most common reasons for not undergoing HIV testing was the belief that HIV was not relevant to oneself, which was reported by approximately 40% (54/153) to 50% (59/119) of the respondents. In addition, 11.3% (109/964) of the respondents did not know where to get tested, and 4.9% (47/964) were unaware that HIV testing existed. Lack of time and concern about cost were cited as reasons for not being tested by 8.4% (81/964) and 5.4% (52/964) of the respondents, respectively. Moreover, 3.7% (36/964) reported feeling that there were barriers to undergoing HIV testing. These findings suggest that a lack of knowledge regarding HIV risk and testing options may

XSL•FO
RenderX

contribute to the absence of testing behavior. These results are consistent with those of previous studies conducted among men who have sex with men populations, reinforcing the importance of recognizing one's own risk of HIV infection and gaining knowledge about HIV testing [16]. This study also found that 20.8% (200/964) of the respondents were aware of the availability of mail-based HIV testing services. This suggests that promoting awareness of various testing options may help encourage testing behavior, particularly in populations with low awareness.

Benefit of Education Through the Symptom Checker

This study demonstrated the benefits and challenges of information provision approaches related to HIV testing. The population in this study consisted of individuals who exhibited certain symptoms and were therefore more likely to visit medical institutions or testing facilities than asymptomatic individuals. Approximately 25% (105,365/424,893) of the participants visited a medical institution. In the study population, the provision of information through an AI-based symptom checker was identified as a factor associated with health care–seeking behaviors. Additionally, among respondents to the HIV-specific banner questionnaire, approximately 29% (280/964) reported a change in mindset; that is, participants became more receptive to HIV testing despite not having initially intended to undergo it. Furthermore, 18.7% (180/964) expressed a new intention to use mail-based HIV-testing services. Based on these findings, when educational information reaches users, it has the potential to promote medical consultations and testing behaviors to a certain extent, thereby demonstrating the usefulness of web-based awareness initiatives. However, these findings indicate that there is room for improvement in information accessibility, as only approximately 20% (156,226/755,972) of users viewed HIV-awareness content related to their HIV-related conditions, and only approximately 0.1% (1006/755,972) accessed pages about HIV testing. Therefore, strategies to increase visibility and engagement with educational content are needed.

While the provision of information through AI-based symptom checkers was associated with health care–seeking behaviors, no significant association was found with HIV testing uptake in this study. A previous survey conducted at 11 clinics in Japan that handle STIs revealed that 6 did not offer HIV testing unless requested by the patient [15]. These findings suggest that there are barriers to HIV testing in medical institutions, such as the need to explain and persuade patients who do not proactively request testing and the lack of incentives. In the United States, a previous study on the medical records of newly diagnosed HIV cases reported that missed opportunities for HIV testing were common. Therefore, strengthening opt-out HIV screening programs is recommended [17]. These findings, along with ours, suggest that health care providers should adopt a proactive approach to information provision. Individuals must also be encouraged to recognize their potential risks and take the initiative to request testing or select their preferred testing method. Our results highlight the importance of raising awareness among physicians and patients to reduce barriers to HIV testing in the current Japanese setting.

Limitations

This study has some limitations. First, because the number of HIV testing events was very small (394/424,893 cases, 0.09%), the stability and precision of the regression estimates are limited. Although a commonly cited rule suggests that approximately 10 events per variable may be acceptable for multivariable logistic regression, the rarity of the outcome in this study raises concerns regarding estimate stability. Therefore, the regression analyses were conducted for exploratory purposes to identify general patterns of association rather than to provide precise effect estimates, and the results should be interpreted with caution.

Second, this study used an episode-based analytic approach, which may introduce within-person dependency when individuals contributed multiple episodes. Although baseline characteristics were similar when summarized at the individual and episode levels, the potential nonindependence across episodes may affect the estimation of variance for the estimates, representing a limitation of this study.

Third, the data source for this study did not allow for the precise identification of individuals with a history of HIV infection at the time of the first survey. Therefore, it was difficult to completely exclude all individuals with a prior HIV diagnosis from the study population, and some may have been included. Consequently, there are limitations in generalizing the study population to the intended target population, which consists of individuals with possible HIV infection or STIs who have not previously received medical consultation or testing. However, before the 9 relevant, potentially related conditions were displayed in the first survey, data on whether individuals had been diagnosed with each condition were obtained. This allowed for the exclusion of individuals with HIV infection or STIs.

Fourth, outcome evaluation relied on revisit survey data; therefore, only individuals who responded to the survey were included. This may have introduced a selection bias in the study population. For instance, individuals who received medical consultation after the first survey but did not revisit the AI-based symptom checker were excluded, potentially leading to an underestimation of medical consultations. Additionally, if individuals received medical consultation after revisiting the AI-based symptom checker, they may still be misclassified as not having received medical consultation in the revisit survey data, further contributing to an underestimation of medical consultations. However, this system prevents revisit surveys within 24 hours of the first survey, which may mitigate bias by excluding immediate revisits. Moreover, because approximately 80% to 90% of individuals did not complete the revisit survey and were therefore excluded from the analysis, it is difficult to generalize the findings to this large excluded population, reflecting potential selection and attrition bias.

Fifth, the first survey data in this study were derived from an AI-based symptom checker questionnaire that adapted subsequent questions based on individual responses. Consequently, certain questions, such as those regarding comorbidities or symptoms, were presented only to specific individuals, leading to missing data. Missing data reflect the absence of questioning rather than the absence of a condition

XSL•FO
RenderX

or symptom, which limits the assessment of factors associated with the outcomes. Furthermore, although transgender individuals were considered a high-risk population, they were not identified in this study. Consequently, some women may have been categorized as female. A further limitation is that the internal algorithms, versioning, and update processes of the AI-based symptom checker are proprietary and were not fully available for independent reporting in this study. As a result, we were unable to detail the specific AI methods or evaluate how algorithmic updates or postdeployment monitoring might influence system behavior. Our description of the tool is, therefore, limited to its observable adaptive question flow behavior, and the findings should be interpreted with this constraint in mind.

Sixth, as this data source was derived from a self-reported questionnaire, the outcomes and characteristic variables obtained may be subject to misclassification owing to individual recall errors. As the results of HIV testing are based on self-reporting from revisit surveys, it is unclear whether they reported the results of the screening test or a definitive diagnosis. This could have caused an overestimation of HIV-positive cases. This overestimation is more likely to occur in populations with a low underlying HIV prevalence, such as women and heterosexual men in the Japanese epidemiological context, where false-positive screening results are more common relative to confirmed HIV diagnoses.

Conclusions

This study suggests that HIV-related information provided through an AI-based symptom checker may influence individuals' awareness or consideration of medical consultation. To further promote HIV testing, it may be necessary to refine the content and delivery of educational materials and enhance HIV testing literacy among physicians who encounter patients with STIs. In addition, such tools may be useful when integrated into broader HIV awareness or testing promotion strategies led by community-based organizations or the public sector.

Acknowledgments

The authors thank Ryuji Matsushima of Ubie Inc for his assistance with the data analysis. They also thank Ryozo Wakabayashi and Takuro Okuyama of Datack Inc for their assistance with writing. The authors did not use generative artificial intelligence in the preparation of this manuscript.

Funding

This study was funded by Gilead Sciences K.K., Japan. Employees of the funder were involved in the study as well as the decision to submit the manuscript for publication.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to privacy restrictions and user agreements, which limit data use to approved research and do not extend to public sharing, but are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: NT, K Hirahara, K Harada, KL, KI, AI
Data curation: KO
Formal analysis: KO
Conceptualization: NT, K Hirahara, K Harada, KL, KI, AI
Data curation: KO
Formal analysis: KO
Investigation: KO
Methodology: NT, K Hirahara, K Harada, KL, KI, AI
Project administration: NT
Supervision: AI
Validation: KO
Visualization: KO
Writing—original draft: NT
Writing—review and editing: NT, K Hirahara, K Harada, KO, KL, KI, AI

Conflicts of Interest

NT, K Hirahara, K Harada, and KL are employees of Gilead Sciences, K.K. with stock ownership. KI receives payment for consultation from Gilead Sciences, K.K. and ViiV Healthcare K.K. AI receives payment for consultation from Gilead Sciences, K.K. KO is an employee of Ubie Inc.

Multimedia Appendix 1

XSL•FO
RenderX

Supplementary tables and figures supporting the main analyses and the plain language summary.
[DOCX File, 1523 KB - jmir_v28i1e90257_app1.docx]

References

1. Fact Sheet 2025 Global HIV Statistics. : Joint United Nations Programme on HIV/AIDS (UNAIDS); 2025 URL: https://www.unaids.org/sites/default/files/2025-07/2025_Global_HIV_Factsheet_en.pdf [accessed 2025-09-24]
2. The path that ends AIDS: UNAIDS global AIDS update 2023. : Joint United Nations Programme on HIV/AIDS (UNAIDS); 2023 URL: https://www.unaids.org/sites/default/files/media_asset/2023-unaids-global-aids-update_en.pdf [accessed 2025-09-24]
3. Global AIDS Strategy 2021–2026: end inequalities. End AIDS. : Joint United Nations Programme on HIV/AIDS (UNAIDS); 2021 URL: https://www.unaids.org/sites/default/files/media_asset/global-AIDS-strategy-2021-2026_en.pdf [accessed 2025-09-24]
4. Annual report on AIDS incidence trends in 2025. : AIDS Trend Committee, Ministry of Health, Labour and Welfare, Japan; 2026 URL: https://api-net.jfap.or.jp/status/Japan/nenpo.html [accessed 2025-09-24]
5. Ejima K, Koizumi Y, Yamamoto N, et al. HIV testing by public health centers and municipalities and new HIV cases during the COVID-19 pandemic in Japan. J Acquir Immune Defic Syndr 2021 Jun 1;87(2):e182-e187. [doi: 10.1097/QAI.0000000000002660] [Medline: 33625066]
6. Nishiura H, Fujiwara S, Imamura A, Shirasaka T. HIV incidence before and during the COVID-19 pandemic in Japan. Math Biosci Eng 2024 Feb 29;21(4):4874-4885. [doi: 10.3934/mbe.2024215] [Medline: 38872518]
7. Taniguchi T, Yamagishi K, Yahaba M, Yoshikawa H, Igari H. Missed opportunities: HIV testing deficits among syphilis-tested patients in Japan. J Infect Chemother 2025 Oct;31(10):102816. [doi: 10.1016/j.jiac.2025.102816] [Medline: 40992169]
8. Otani M, Shiino T, Hachiya A, et al. Association of demographics, HCV co-infection, HIV-1 subtypes and genetic clustering with late HIV diagnosis: a retrospective analysis from the Japanese Drug Resistance HIV-1 Surveillance Network. J Int AIDS Soc 2023 May;26(5):e26086. [doi: 10.1002/jia2.26086] [Medline: 37221951]
9. Horino T, Sato F, Kato T, et al. Associations of HIV testing and late diagnosis at a Japanese university hospital. Clinics (Sao Paulo) 2016 Feb;71(2):73-77. [doi: 10.6061/clinics/2016(02)04] [Medline: 26934235]
10. Nishijima T, Takano M, Matsumoto S, et al. What triggers a diagnosis of HIV infection in the Tokyo metropolitan area? Implications for preventing the spread of HIV infection in Japan. PLoS One 2015;10(11):e0143874. [doi: 10.1371/journal.pone.0143874] [Medline: 26606382]
11. Nishiura H, Fujiwara S, Imamura A, Shirasaka T. Regional variations in HIV diagnosis in Japan before and during the COVID-19 pandemic. Infect Dis Model 2024;10(1):40-49. [doi: 10.1016/j.idm.2024.08.004] [Medline: 39319285]
12. Romero RA, Klausner JD, Marsch LA, Young SD. Technology-delivered intervention strategies to bolster HIV testing. Curr HIV/AIDS Rep 2021 Aug;18(4):391-405. [doi: 10.1007/s11904-021-00565-y] [Medline: 34109549]
13. Balzer LB, Havlir DV, Kamya MR, et al. Machine learning to identify persons at high-risk of human immunodeficiency virus acquisition in rural Kenya and Uganda. Clin Infect Dis 2020 Dec 3;71(9):2326-2333. [doi: 10.1093/cid/ciz1096] [Medline: 31697383]
14. Company. Ubie Health. URL: https://ubiehealth.com/company [accessed 2025-10-24]
15. Itoda I. HIV testing in health care facilities [Article in Japanese]. J AIDS Res 2015;17(3):133-137 [FREE Full text]
16. Strategic study for AIDS prevention [Report in Japanese]. : Japan Foundation for AIDS Prevention (JFAP); 2002 URL: https://www.jfap.or.jp/strategic_study/hokokuPdf/01.pdf [accessed 2025-09-25]
17. Anderson S, Friedman EE, Eller D, et al. HIV testing in a high prevalence urban area in the US: identifying missed opportunities two ways. Int J STD AIDS 2022 Oct;33(11):970-977. [doi: 10.1177/09564624221118484] [Medline: 36031933]

Abbreviations

AI: artificial intelligence
OR: odds ratio
STI: sexually transmitted infection

XSL•FO
RenderX

Edited by S Raquel Ramos; submitted 25.Dec.2025; peer-reviewed by E Kodama, M Imahashi, T Fujii; revised version received 02.Apr.2026; accepted 28.Apr.2026; published 18.Jun.2026.

Please cite as:
Taguchi N, Hirahara K, Harada K, Orimo K, Lee K, Iwahashi K, Imamura A
Medical Consultation and HIV Testing After AI-Based Symptom Check: Retrospective Cohort Study
J Med Internet Res 2026;28:e90257
URL: https://www.jmir.org/2026/1/e90257
doi:10.2196/90257

©Nao Taguchi, Kunihiro Hirahara, Keisuke Harada, Keisuke Orimo, Kuan Yeh Lee, Kota Iwahashi, Akifumi Imamura. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Computerized Self-Reported Medical History Taking to Support Early Rule Out of Major Adverse Cardiac Events in Patients With Acute Chest Pain: Post Hoc Analysis of the CLEOS-CPDS Prospective Cohort Study

Helge Brandberg[1], MD, PhD; Carl Johan Sundberg[2,3], MD, PhD; Jonas Spaak[1], MD, PhD; Sabine Koch[2], MSc, PhD; Thomas Kahan[1], MD, PhD

[1]Division of Cardiovascular Medicine, Department of Clinical Sciences, Danderyd Hospital, Karolinska Institutet, Stockholm, Sweden

[2]Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden

[3]Department of Physiology & Pharmacology, Karolinska Institutet, Stockholm, Sweden

Corresponding Author:
Helge Brandberg, MD, PhD
Division of Cardiovascular Medicine
Department of Clinical Sciences, Danderyd Hospital
Karolinska Institutet
Stockholm, SE-182 88
Sweden
Phone: 46 8 123 583 47
Email: helge.brandberg@ki.se

Abstract

Background: Self-reported, computerized history taking (CHT) may enable efficient collection of medical histories for acute chest pain management.

Objective: The primary aim is to determine the diagnostic performance of 4 CHT-derived chest pain risk scores for ruling out 30-day major adverse cardiac events (MACEs) or acute coronary syndrome (ACS). The secondary aim is to assess their impact on patient disposition in the emergency department (ED).

Methods: This is a prospective cohort study conducted at a tertiary hospital ED in Stockholm, Sweden. Clinically stable adults (≥18 years) with chest pain and an electrocardiogram (ECG) not indicating an acute disease requiring immediate care provided medical histories via a tablet-based CHT program (Clinical Expert Operating System [CLEOS]). CHT data and ECG interpretations and troponin values were used to calculate the History, ECG, Age, Risk Factors, and Troponin (HEART) score, Danderyd HEART (D-HEART) score, Emergency Department Assessment of Chest Pain Score combined with an Accelerated Diagnostic Protocol (EDACS-ADP), and Troponin-only Manchester Acute Coronary Syndrome (T-MACS). The primary outcome was 30-day ACS; the secondary outcome was 30-day MACE (ACS, revascularization, or cardiovascular death).

Results: Among 1000 participants (age: mean 55 years, SD 17 years; 456/1000, 45.60%, women), risk scores could be calculated in 838 (83.80%). Within 30 days, 65 (6.50%) participants experienced ACS, and 72 (7.20%) had a MACE. Negative predictive values were 0.99 (95% CI 0.97-1.00) for both outcomes. Sensitivity for MACE was 0.91 (95% CI 0.81-0.97) for HEART, 0.94 (95% CI 0.86-0.98) for D-HEART, 0.94 (95% CI 0.86-0.98) for EDACS-ADP, and 0.97 (95% CI 0.90-1.00) for T-MACS, with similar results for ACS. As many as 89 of the 528 (16.9%) patients admitted could be reclassified from “nonlow risk” to “low risk.” Among reclassified patients, 30-day MACE or ACS occurred in 0-4 cases; miss rates were below 1% for D-HEART (4/416, 0.96%) and T-MACS (2/286, 0.7%), but exceeded 1% for HEART (6/406, 1.5%) and EDACS-ADP (4/346, 1.2%).

Conclusions: Automated, self-reported CHT provided sufficient data to calculate 4 chest pain risk scores in 838 of 1000 (83.80%) patients with acute chest pain, with score calculation dependent on physician-interpreted ECGs. These CHT-derived risk scores demonstrated good diagnostic performance for ruling out 30-day MACE and ACS. Performance was broadly comparable with prior studies using physician-acquired scores, although suggested safety thresholds were primarily met by D-HEART and T-MACS. The improved safety of D-HEART compared with HEART is likely attributable to the incorporation of serial 0/1-hour troponin testing. Use of CHT-derived risk scores may reclassify a substantial fraction of admitted patients as “low risk,” potentially supporting discharge decisions in selected patients, while admission may still be required for non-ACS reasons. However, any

XSL•FO
RenderX

gains in discharge rates should be weighed against the possibility of missed events among reclassified patients. Multicenter studies are needed to confirm generalizability, operational feasibility, and safety.

Trial Registration: ClinicalTrials.gov NCT03439449; https://clinicaltrials.gov/ct2/show/NCT03439449

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2019-031871

(J Med Internet Res 2026;28:e76087) doi:10.2196/76087

KEYWORDS

chest pain; risk assessment; coronary artery disease; emergency service, hospital; medical informatics; artificial intelligence; medical history taking

Introduction

Chest pain, a common chief complaint in emergency departments (EDs) worldwide [1,2], contributes to overcrowding in these settings [3]. Prompt and accurate management is crucial, not only for identifying life-threatening conditions such as acute coronary syndrome (ACS; acute myocardial infarction or unstable angina pectoris), but also for the safe discharge of the majority of patients with benign conditions. An integrated assessment combining medical history, signs and symptoms, electrocardiogram (ECG), and high-sensitivity cardiac troponins (hs-cTn) is recommended for initial short-term risk stratification [4]. To consolidate these observations, risk scores are recommended for structured, individualized patient assessment [5,6]. Their clinical utility is supported by substantial evidence [7-9], and recent data have shown that risk score implementation can reduce the risk of major adverse cardiac events (MACEs) by half among patients with chest pain discharged from the ED [10]. However, chest pain risk scores remain underutilized in EDs [11], and when used, issues such as miscalculation [12] or discordance among physicians [13] are common.

Digital tools are increasingly integrated into health care, offering potential enhancements in efficiency and decision-making [14,15]. In acute chest pain management, international societies advocate improved symptom classification methods using machine learning [16]. In addition to biometric data such as vital signs or ECG, prompt and reliable acquisition of medical history data is essential for the use of these techniques. Computerized history taking (CHT) is a possible solution, offering an automated, structured, and standardized approach to self-reported medical history. Previous studies indicate that CHT can assist physicians by providing a more standardized, detailed, and complete medical history in both emergent and nonemergent settings [17-20]. Studies have also shown that CHT is well received by patients with acute chest pain, with a majority able to interact effectively and provide sufficient data for risk stratification [21-23]. However, there is a knowledge gap regarding the impact of CHT on clinical management and outcomes in acute care settings [24]. This study proposes a standardized, automated approach to collect medical history data, enabling real-time use of clinical decision-support tools. This approach has the potential to significantly impact the management of the large population of patients presenting to EDs with chest pain.

We hypothesized that risk scores populated with CHT data would reliably identify patients with acute chest pain at low risk for MACE or ACS, enabling early rule out and thereby reducing unnecessary admissions and diagnostic testing without compromising patient safety. Our aims were therefore, first, to determine the overall diagnostic performance of 4 chest pain risk scores populated with CHT data for 30-day MACE (defined as ACS, coronary artery revascularization, or cardiovascular death) or ACS rule out; and second, to assess the potential impact of this approach on patient disposition among those presenting to the ED with acute chest pain.

Methods

Study Design and Setting

The Clinical Expert Operating System Chest Pain Danderyd Study (CLEOS-CPDS; ClinicalTrials.gov identifier: NCT03439449) was a prospective cohort study designed to determine the value of self-reported CHT in acute chest pain management [25]. This study reports diagnostic accuracy and adheres to the STARD (Standards for Reporting Diagnostic Accuracy) 2015 guideline [26].

Participants

Patients presenting to Danderyd University Hospital, a tertiary hospital in Stockholm, Sweden, were consecutively enrolled between October 1, 2017, and May 16, 2019. Eligible participants were women and men aged 18 years or older, fluent in Swedish, with a presenting complaint of chest pain recorded by an ED triage nurse or registrar; a nondiagnostic initial ECG, serum markers, or both, for an acute disease requiring immediate care; and clinical stability, as defined by the Rapid Emergency Triage and Treatment System (RETTS) levels orange, yellow, green, and blue [27]. Exclusion criteria were inability to complete a CHT interview (eg, agitation, severely impaired vision, or confusion). Among patients considered eligible but not enrolled, the most common reasons were language barriers (eg, inability to read Swedish), patients feeling too tired, and difficulties using a tablet [22]. All patients included in the study were initially managed as suspected ACS according to standard care protocols recommended by regional guidelines [28]. Data from the CHT and the calculated risk scores were not accessible to the treating physicians or other health care professionals and were used solely for clinical research purposes.

Interventions

Computerized History Taking

Medical history was collected using the CHT software CLEOS on tablets (iPad; Apple Inc), as described in detail previously

XSL•FO
RenderX

[25,29]. Briefly, CLEOS is an expert system, a form of artificial intelligence software, that uses a rule-based approach to guide the history-taking process. It incorporates a comprehensive medical knowledge base, algorithmically represented by decision trees comprising over 17,000 decision nodes. By directly interacting with the patient, CLEOS automates history taking through a series of structured questions, mainly in text format (ie, yes/no or multiple-choice questions), as well as image-based questions. A limited number of free-text entries are available for instances in which patients are unable to select a suitable option, such as when describing the quality of their pain. Based on prior responses and their clinical relevance, CLEOS adapts the interview and dynamically determines the next most appropriate questions, emulating clinical reasoning as data are continuously collected [29]. In a study evaluating interrater reliability between data collected using CHT and physician-acquired data, we observed high agreement for traditional risk factors (eg, diabetes mellitus or hypercholesterolemia), but low to moderate agreement for chest pain characteristics (eg, pain radiating to the arm or relief by rest) [30]. As previously reported, the median time required to collect sufficient information to calculate the HEART score was 23 (IQR 18-31) minutes [22].

Chest Pain Risk Scores

The History, ECG, Age, Risk Factors, and Troponin (HEART) score; Emergency Department Assessment of Chest Pain Score combined with an Accelerated Diagnostic Protocol (EDACS-ADP); and Troponin-only Manchester Acute Coronary Syndrome (T-MACS) score were calculated as described in their respective derivation studies [31-33]. When assessing the History component of the HEART score in this retrospective analysis, the traditional clinical classification of suspected anginal symptoms was used: (1) central chest pain, (2) triggered by physical or emotional exertion, and (3) relieved by rest or nitrates. Based on the number of criteria met, patient history was categorized as highly (3 criteria met), moderately (2 criteria met), or slightly suspicious (none or 1 criterion met). This predefined, symptom-based approach to probability assessment, used as an aid in the clinical diagnosis of coronary artery disease (Diamond-Forrester prediction rule) [34], reflects commonly applied principles for reconstructing the HEART History component in several validation studies [35,36]. However, this approach differs from the derivation study and subsequent validation study by the same authors, in which clinician judgment, rather than a predefined set of criteria, was used to evaluate the likelihood of chest pain being related to MACE [33,37]. The Danderyd HEART (D-HEART) score was calculated according to guidelines in Region Stockholm [28], combining the HEART score with hs-cTn assays using the European Society of Cardiology (ESC) 0/1-hour rule-out and rule-in algorithm [38,39]. By contrast, the HEART score in this study was calculated using a single baseline hs-cTn measurement, in accordance with the derivation study. We defined a clinically decisive risk score as “nonlow risk,” indicating the need for hospitalization, if D-HEART ≥4, HEART ≥4, EDACS ≥16, T-MACS ≥0.02, or if troponin met rule-in criteria using the ESC 0/1-hour algorithm [4]. For T-MACS, we evaluated both the <0.02 (very low risk) and <0.05 (low risk) thresholds, as they represent distinct clinical pathways (immediate discharge vs clinical observation with further serial troponin testing) and were defined in the derivation study [32]. Detailed descriptions of chest pain characteristics, risk factor components, and interpretation (ie, “nonlow risk” thresholds) for D-HEART, HEART, EDACS, and T-MACS are provided in Table 1.

XSL•FO
RenderX

Table 1. Components and interpretation of assessed risk scores.

Component	Risk score			
	D-HEART[a]	HEART[b]	EDACS-ADP[c]	T-MACS[d]
Age	✓	✓	✓	N/A[e]
Sex	N/A	N/A	✓	N/A
Chest pain characteristics				
Central chest pain	✓	✓	N/A	N/A
Provoked by physical exertion or emotional stress	✓	✓	N/A	N/A
Relieved by rest or nitrates	✓	✓	N/A	N/A
Diaphoresis	N/A	N/A	✓	✓
Radiation to arm or shoulder	N/A	N/A	✓	N/A
Radiation to right arm or shoulder	N/A	N/A	N/A	✓
Pain occurred/worsened with inspiration	N/A	N/A	✓	N/A
Pain is reproduced by palpation	N/A	N/A	✓	N/A
Worsening or crescendo angina	N/A	N/A	N/A	✓
Pain associated with vomiting	N/A	N/A	N/A	✓
Risk factors				
Atherosclerotic disease	✓	✓	N/A	N/A
Known coronary artery disease (not stroke)	N/A	N/A	✓	N/A
Diabetes mellitus	✓	✓	✓	N/A
Current smoker	✓	✓	✓	N/A
Family history of premature coronary artery disease	✓	✓	✓	N/A
Hypertension	✓	✓	✓	N/A
Hypercholesterolemia	✓	✓	✓	N/A
Reported obesity	✓	✓	N/A	N/A
Electrocardiogram, vital signs, and biomarkers				
Signs of ischemia on electrocardiogram	✓	✓	✓	✓
Hypotension, systolic blood pressure <100 mmHg	N/A	N/A	N/A	✓
Elevated biomarker (troponin)	✓	✓	✓	✓
Possible range of score	0 to 8	0 to 10	–10 to 34	0 to 1
Threshold for “nonlow risk”	≥4[f]	≥4	≥16[g]	≥0.02

[a]D-HEART: Danderyd HEART.

[b]HEART: History, ECG, Age, Risk Factors, and Troponin.

[c]EDACS-ADP: Emergency Department Assessment of Chest Pain Score combined with an Accelerated Diagnostic Protocol.

[d]T-MACS: Troponin-only Manchester Acute Coronary Syndrome.

[e]N/A: not applicable.

[f]And a rule-in troponin result using the European Society of Cardiology 0/1-hour algorithm [38].

[g]And new ischemia on electrocardiogram and/or a rule-in 0- or 2-hour troponin result.

Data Collection

At presentation to the ED, all patients with acute chest pain were triaged to determine urgency either by a cardiology consultant or a senior cardiology resident (8 AM to 5 PM) or by a triage nurse using the RETTS protocol (5 PM to 8 AM) [27]. After ECG and biomarker acquisition during triage, patients underwent a more thorough examination and standard history taking by the attending physician, either in the ED cardiology unit—operating 24 hours a day and staffed by cardiology consultants or senior residents with limited access to noninvasive cardiac imaging techniques—or in the inpatient day-care unit. The day-care unit, operational from 8 AM to 5 PM, functions as an observational ward, is staffed by cardiology consultants, and provides ready access to both invasive and noninvasive cardiac imaging techniques. Details regarding

XSL•FO
RenderX

patient disposition have been reported in previous publications [22,25].

The participant enrollment process has been described in detail previously [22,25]. Briefly, patients were invited to participate in the study by a research staff member in the ED cardiology unit or the inpatient day-care unit. Patients were provided with standardized information and given the opportunity to ask questions before signing a consent form to provide informed consent. CHT was conducted during waiting times in the ED, either before or after the initial physician consultation. The CHT interview could be paused during standard clinical procedures (eg, blood sampling or radiographic examinations) and resumed afterward. Discontinuation of CHT occurred upon completion, at the patient's request, or upon discharge home or admission to a ward. The CHT interview did not interfere with standard management, and ED staff had no access to data collected through CHT.

Data relevant to this study were extracted from the electronic health record (EHR; TakeCare; CompuGroup Medical Sweden AB) by research staff and from the CHT database. To align with variables used in the risk scores, we predefined variables to be collected and established an interpretation scheme that allowed all variables to be converted into a binary format (yes/no responses; see an example of the interpretation scheme for crescendo angina pectoris in [30]). Demographic information (age and sex) for the general ED population at Danderyd University Hospital during the study period was extracted from the EHR system using QlikView version 12.10 (QlikTech International AB).

ECG findings, interpreted by both physicians and computer algorithms, were collected by a research assistant from the EHR. To specifically evaluate the independent contribution of CHT to chest pain management, physician-reported ECG interpretations were deliberately used to populate the risk scores. In cases in which a physician report was missing (n=58), the automated ECG interpretation (EC Sense ECG; Cardiolex Medical AB) was imputed. A detailed description of ECG findings and their categorization is provided in Table S1 in Multimedia Appendix 1. Blood samples were collected according to standard procedures at the Karolinska University Laboratory, Stockholm. The detection limit of the assay was 5 ng/l, with a 99th percentile upper reference limit—used to identify elevated levels indicative of myocardial injury—of 14 ng/l [40].

All variables required for each risk score were extracted from the EHR and CHT and transformed according to the predefined interpretation scheme. CHT responses were manually mapped post hoc, and risk scores were calculated retrospectively.

Outcome Measures

The primary outcome was defined as a diagnosis of ACS [25]. To facilitate comparison with other diagnostic accuracy studies of acute chest pain, we also report 30-day MACE as a secondary outcome, comprising an ACS diagnosis, coronary artery revascularization, or cardiovascular death within 30 days. According to the study protocol, the diagnosis of ACS was verified by a board-certified cardiologist in accordance with prevailing European guidelines at the time of the study [41,42]. The cardiologist then assigned the corresponding International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10) codes: I20.0 (unstable angina pectoris), I21.0-9 (acute myocardial infarction), or I24.0-9 (other acute ischemic heart diseases) [43,44]. Previous studies have shown that the overall positive predictive value (PPV) of Swedish discharge ICD codes is 85%-95% for most diagnoses, including acute myocardial infarction [45,46]. Data on death and revascularization were retrieved through review of the regional EHR system, which is shared by all hospitals and cardiology outpatient clinics in Region Stockholm. The regional EHR system is linked to the national population registry, providing complete information on mortality status. No active follow-up or separate linkage to a national registry was performed. Further details are available in the study protocol [25].

Our previous work shows that CHT provides more complete information on key medical history variables than that reported by physicians in the EHR [30]. Consequently, sufficient information for risk score calculation was available for fewer than 30% of patients when using EHR data alone, compared with approximately three-quarters when using CHT data [30]. As a result, reliable retrospective calculation of a complete physician-derived HEART score was not possible, making a direct head-to-head comparison with the CHT-derived score infeasible. To evaluate how CHT might influence clinical management, we therefore used patient disposition as a proxy for the attending physician's overall risk assessment. Participants discharged from the ED to home were classified as "low risk," whereas those admitted to a ward or day-care unit were categorized as "nonlow risk."

Statistical Analysis

In this post hoc analysis evaluating diagnostic accuracy, a risk score populated with CHT data was considered the index test, and the occurrence of MACE or ACS was considered the reference test. Descriptive statistics are presented as means (SDs), 95% CIs, or proportions, as appropriate. Differences between groups with and without MACE or ACS were assessed using the Student *t* test for continuous variables and the Pearson chi-square test for binomial and categorical variables. Diagnostic accuracy was evaluated using cross-tabulation of the index test against the reference test. For each index test, performance was assessed by calculating sensitivity, specificity, PPV, and negative predictive value (NPV) for the reference test. The Pearson chi-square test was used to compare NPVs between index tests. Receiver operating characteristic curves were generated for each risk score, and the Hanley and McNeil method [47] was used to test for differences between areas under the receiver operating characteristic curve (AUCs). The miss rate was defined as the proportion of missed cases in the rule-out group. All analyses were performed using STATA, version 14.2 (StataCorp).

The sample size was determined based on the targeted precision of sensitivity and specificity in the original study objective, as previously described [25]. Briefly, because the prevalence of MACE and ACS in the study population was unknown, the

XSL•FO
RenderX

calculation was based on an assumed prevalence of 0.5 (50%), which maximizes the estimated sample size. To achieve a 0.03 (3%) precision for sensitivity and specificity, as calculated using nQuery, version 7.0 (Statistical Solutions Ltd), 1000 participants were required.

Ethical Considerations

This study complied with the Declaration of Helsinki and was approved by the Stockholm Regional Ethical Committee (now the Swedish Ethical Review Authority; reference number 2015/1955-31). All participants received oral and written information about the study and provided written informed consent before inclusion. Study data were deidentified before analysis, and only coded study IDs were used in the research dataset. Participants did not receive any financial or other compensation for their participation.

Results

Study Population Characteristics

During the study period, a total of 13,044 patients presented to the ED with a chief complaint of chest pain. During periods with on-duty research staff (office hours, evenings, and weekends), 1000 patients were consecutively included (Figure 1). The age and sex distribution of the study population (age: mean 55.3 years, SD 17.4 years; 456/1000, 45.60% female) closely aligned with that of the general chest pain population at Danderyd University Hospital (age: mean 57.6 years, SD 19.1 years; 49% female). Further baseline characteristics collected by CHT and selected ED variables, as well as patient disposition, are summarized in Tables 2 and 3.

XSL•FO
RenderX

Figure 1. Study flow diagram. CHT: computerized history taking; D-HEART: Danderyd HEART; ECG: electrocardiogram; ED: emergency department; EDACS-ADP: Emergency Department Assessment of Chest Pain Score combined with an Accelerated Diagnostic Protocol; EHR: electronic health record; HEART: History, ECG, Age, Risk Factors, and Troponin; T-MACS: Troponin-only Manchester Acute Coronary Syndrome.

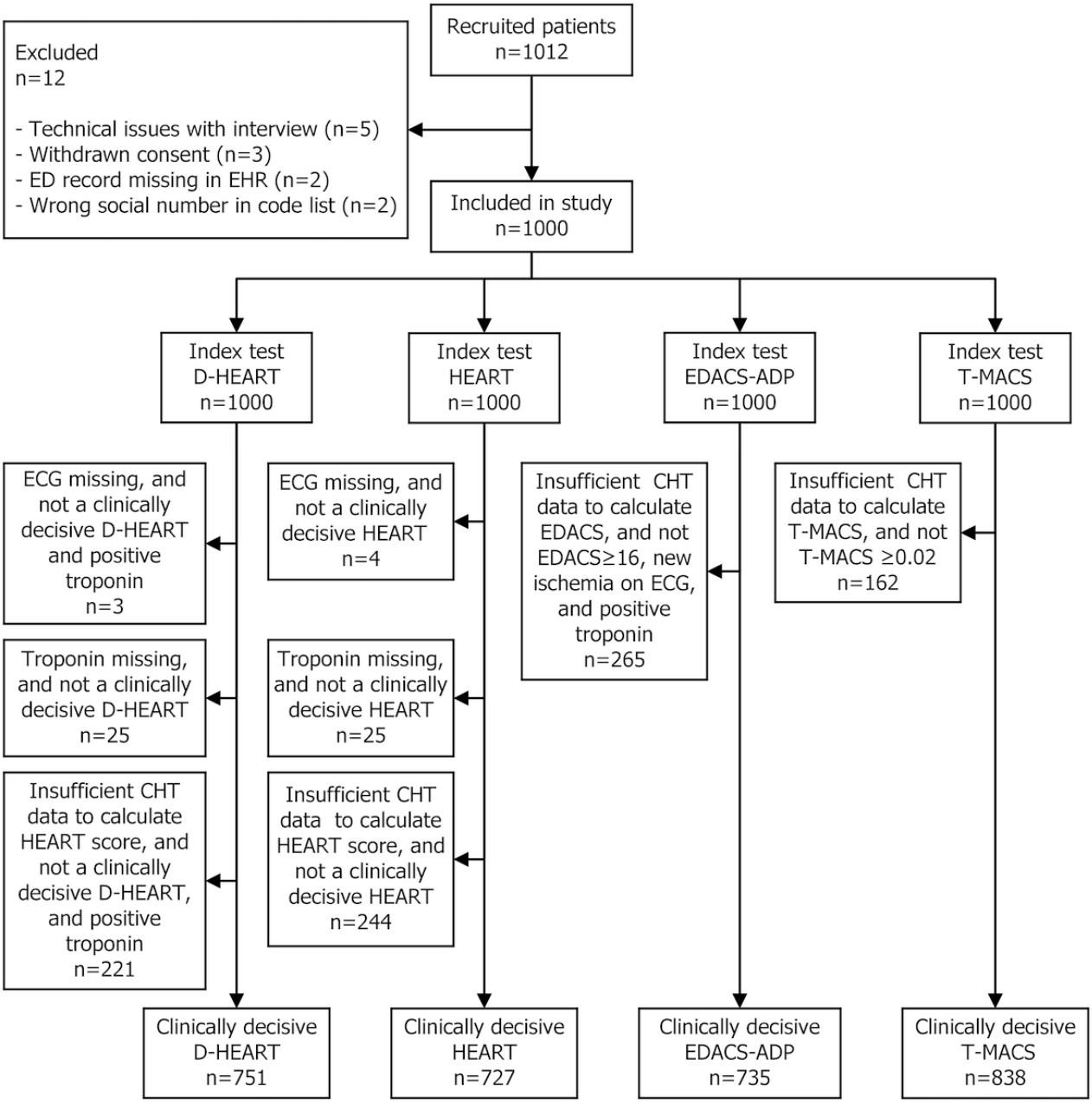

XSL•FO
RenderX

Table 2. Demography and background of all included patients divided into groups depending on outcome (30-day MACE[a] and ACS[b]).[c]

Characteristics	All		MACE		Non-MACE		*P* value	ACS		Non-ACS		*P* value
	Value	n	Value	n	Value	n		Value	n	Value	n	
Age (years), mean (SD)	54.7 (17.2)	1000	67.6 (11.0)	72	53.7 (17.1)	928	<.001	68.0 (10.9)	64	53.8 (17.1)	936	<.001
Sex (females), n (%)	456 (45.60)	1000	19 (26.39)	72	437 (47.09)	928	.001	18 (28.13)	64	438 (46.79)	936	.004
BMI (kg/m^2), mean (SD)	26.4 (4.7)	1000	26.6 (4.3)	72	26.3 (4.7)	928	.63	26.2 (3.9)	64	26.4 (4.7)	936	.82
Diabetes mellitus type 1 or 2, n (%)	62 (7.87)	788	8 (15.69)	51	54 (7.33)	737	.03	5 (11.36)	44	57 (7.66)	744	.37
Ongoing lipid-lowering medication, n (%)	128 (20.00)	640	21 (47.73)	44	107 (17.95)	596	<.001	18 (46.15)	39	110 (18.30)	601	<.001
Hypertension, n (%)	316 (41.31)	765	34 (64.15)	53	282 (39.61)	712	<.001	31 (67.39)	46	285 (39.64)	719	<.001
Family history of coronary artery disease, n (%)	200 (25.84)	774	18 (37.50)	48	182 (25.07)	726	.06	17 (41.46)	41	183 (24.97)	733	.02
Known coronary artery disease, n (%)	137 (16.10)	851	31 (55.36)	56	106 (13.33)	795	<.001	27 (55.10)	49	110 (13.72)	802	<.001
History of angina pectoris	88 (10.35)	850	23 (41.07)	56	65 (8.19)	794	<.001	20 (40.82)	49	68 (8.49)	801	<.001
History of myocardial infarction	82 (9.65)	850	21 (37.5)	56	61 (7.68)	794	<.001	18 (36.73)	49	64 (7.99)	801	<.001
History of percutaneous coronary intervention	76 (9.12)	833	16 (29.09)	55	60 (7.61)	778	<.001	13 (27.08)	48	63 (8.03)	785	<.001
History of coronary artery bypass graft	18 (2.16)	833	10 (18.18)	55	8 (1.03)	778	<.001	9 (18.75)	48	9 (1.15)	785	<.001
Current smoker, n (%)	90 (11.39)	790	5 (9.43)	53	85 (11.53)	737	.64	5 (10.87)	46	84 (11.29)	744	.91
Region of birth, n (%)												
Nordic countries	829 (82.90)	1000	63 (87.50)	72	766 (82.54)	928	.28	55 (85.94)	64	769 (82.16)	936	.50
Europe (outside the Nordic countries)	46 (4.60)	1000	1 (1.39)	72	45 (4.85)	928	.18	1 (1.56)	64	45 (4.81)	936	.23
Outside Europe	125 (12.50)	1000	8 (11.11)	72	117 (12.61)	928	.71	8 (12.50)	64	116 (12.39)	936	>.99
Occupational status, n (%)												
Active worker (employed or student)	616 (61.60)	1000	26 (36.11)	72	590 (63.58)	928	<.001	23 (35.94)	64	589 (62.93)	936	<.001
Not at work (unemployed or on sick leave)	69 (6.90)	1000	1 (1.4)	72	68 (7.33)	928	.06	1 (1.56)	64	67 (7.16)	936	.08
Retired	315 (31.50)	1000	45 (62.50)	72	270 (29.09)	928	<.001	40 (62.50)	64	274 (29.27)	936	<.001
Arrived at emergency department by ambulance, n (%)	189 (20.61)	917	14 (24.14)	58	175 (20.37)	859	.49	13 (25.49)	51	176 (20.32)	866	.37
Ongoing chest pain during computerized history taking, n (%)	544 (60.99)	892	19 (34.55)	55	525 (62.72)	837	<.001	16 (33.33)	48	524 (62.09)	844	<.001

[a]MACE: major adverse cardiac event.

[b]ACS: acute coronary syndrome.

[c]Self-reported medical history data derived from computerized history taking.

XSL•FO
RenderX

Table 3. Vital signs, electrocardiogram, circulating biomarkers, and disposition of all included patients divided into groups depending on outcome (30-day MACE[a] and ACS[b]) or not.[c]

Characteristic	All		MACE		Non-MACE		*P* value	ACS		Non-ACS		*P* value
	Value	n	Value	n	Value	n		Value	n	Value	n	
Vital parameters at triage, mean (SD)												
Systolic blood pressure (mmHg)	143 (22)	995	148 (22)	72	143 (22)	923	.06	148 (22)	64	143 (21)	931	.08
Diastolic blood pressure (mmHg)	83 (13)	993	81 (14)	72	84 (13)	921	.18	82 (15)	64	83 (13)	929	.30
Heart rate (beats/minute)	77 (16)	984	74 (15)	72	77 (16)	912	.15	74 (16)	64	77 (16)	920	.19
Respiration rate (breaths/minute)	16 (3)	984	16 (2)	72	16 (3)	912	.85	16 (2)	64	16 (3)	920	.86
Body temperature (°C)	36.8 (0.4)	947	36.6 (0.4)	70	36.8 (0.4)	877	<.001	36.6 (0.4)	63	36.8 (0.4)	884	.001
Electrocardiogram, n (%)[d]												
New signs diagnostic for ischemia	68 (6.8)	994	14 (19.7)	71	54 (5.9)	923	<.001	14 (22.2)	63	53 (5.7)	931	<.001
New nonspecific ST-T changes	131 (13.2)	994	20 (28.2)	71	111 (12.0)	923	<.001	16 (25.4)	63	115 (12.4)	931	.003
Normal or known ST-T alterations	795 (80.0)	994	37 (52.1)	71	758 (82.1)	923	<.001	33 (52.4)	63	758 (81.4)	931	<.001
High-sensitive troponin T values, n (%)												
>3× normal limit (>42 ng/l)	47 (4.8)	970	29 (40.8)	71	18 (2.0)	899	<.001	28 (44.4)	63	19 (2.1)	907	<.001
1-3× normal limit (15-42 ng/l)	112 (11.5)	970	19 (26.8)	71	93 (10.3)	899	<.001	16 (25.4)	63	96 (10.6)	907	<.001
Normal limit[e] (≤14 ng/l)	811 (83.6)	970	23 (32.4)	71	788 (87.7)	899	<.001	19 (30.2)	63	786 (86.7)	907	<.001
5-14 ng/l	397 (40.9)	970	15 (21.1)	71	367 (40.8)	899	.001	11 (17.5)	63	368 (40.6)	907	.004
1-hour troponin T elevated (>2 ng/l)	21 (2.2)	970	2 (2.8)	71	19 (2.1)	899	.69	2 (3.2)	63	19 (2.1)	907	.57
<5 ng/l	414 (42.7)	970	4 (5.6)	71	410 (45.6)	899	<.001	4 (6.3)	63	407 (44.9)	907	<.001
Admitted to the ward or day-care unit, n (%)	528 (53.3)	990	69 (95.8)	72	459 (50.0)	918	<.001	62 (96.9)	64	466 (50.3)	926	<.001
Ward (not via day-care unit)	203 (20.5)	990	61 (84.7)	72	141 (15.4)	918	<.001	57 (89.1)	64	146 (15.8)	926	<.001
Day-care unit	325 (32.8)	990	7 (9.7)	72	318 (34.6)	918	<.001	5 (7.8)	64	320 (34.6)	926	<.001
Day-care unit then to ward	33 (3.3)	7 (9.7)	7 (10)	72	26 (2.8)	918	.002	5 (7.8)	64	28 (3.0)	926	.04
Day-care unit then sent home	292 (29.5)	990	0 (0)	72	292 (31.8)	918	<.001	0 (0)	64	292 (31.5)	926	<.001

[a]MACE: major adverse cardiac event.

[b]ACS: acute coronary syndrome.

[c]Data were derived from the electronic health record.

[d] For definitions of electrocardiogram interpretation, see Table S1 in Multimedia Appendix 1).

[e]99th percentile (for definitions of electrocardiogram interpretation, see Table S1 in Multimedia Appendix 1).

XSL•FO
RenderX

Endpoints Reached Within 30 Days

Among the 1000 included patients, 72 experienced a MACE, including 65 with an ACS; 7 underwent coronary revascularization without an ACS diagnosis. No cases of cardiovascular death occurred (1 noncardiac death due to advanced lung cancer was excluded). Three patients diagnosed with unstable angina pectoris were not revascularized. The types of MACE and ACS are presented in Table 4. MACE and ACS were associated with increasing age, male sex, diabetes mellitus, lipid-lowering therapy, hypertension, and a history of coronary artery disease (Table 2). Primary diagnoses for participants discharged from either the ED or the day-care unit, as well as those hospitalized, are reported in Tables 5 and 6, respectively. Three patients were diagnosed with pulmonary embolism at the ED visit, assessed as low risk, and discharged according to local clinical protocols with appropriate anticoagulant therapy and dedicated specialist outpatient follow-up; these cases therefore do not represent missed or chronic diagnoses. During the study period, 2 patients receiving standard care experienced a MACE within 30 days after ED discharge to home.

Table 4. Major adverse cardiac events within 30 days among the 1000 participants.[a]

Outcome	7 days[b], n	30 days[c], n	Total, n
Major adverse cardiac event	70	2	72
Acute coronary syndrome	64	1	65
Unstable angina pectoris (I20.0)	22	0	22
Revascularization	19	0	19
Percutaneous coronary intervention	16	0	16
Coronary artery bypass grafting	3	0	3
No revascularization	3	0	3
Acute subendocardial myocardial infarction (I21.4)	33	1	34
Revascularization	24	1	25
Percutaneous coronary intervention	23	1	24
Coronary artery bypass grafting	1	0	1
No revascularization	9	0	9
Acute myocardial infarction, unspecified, type 1 (I21.9 + I98.1)	6	0	6
Revascularization	4	0	4
Percutaneous coronary intervention	4	0	4
Coronary artery bypass grafting	0	0	0
No revascularization	2	0	2
Acute myocardial infarction, unspecified, type 2 (I21.9 + I98.2)	3	0	3
Revascularization	0	0	0
Percutaneous coronary intervention	0	0	0
Coronary artery bypass grafting	0	0	0
No revascularization	0	0	0
Revascularization without acute coronary syndrome diagnosis	6	1	7
Percutaneous coronary intervention	6	1	7
Coronary artery bypass grafting	0	0	0
Cardiovascular death	0	0	0

[a]Data were derived from the electronic health record.

[b]7 days: outcome at discharge or ≤7 days from the emergency department visit.

[c]30 days: outcome after discharge or >7 days, and ≤30 days from the emergency department visit.

XSL•FO
RenderX

Table 5. Primary ICD-10[a] code and diagnosis for the 754 patients discharged home.[b]

ICD-10 code	Diagnosis	Values, n (%)
R07	Pain in throat and chest	574 (76.1)
R00-09 (excluding R07)	Symptoms and signs involving the circulatory and respiratory systems	57 (7.6)
R10-49	Symptoms and signs involving other organ systems	23 (3.1)
I48	Atrial fibrillation and flutter	16 (2.1)
R50-69	General symptoms and signs	14 (1.9)
M79	Other soft tissue disorders, not elsewhere classified	11 (1.5)
K29.7, K30.9, and K21.9	Gastritis, unspecified; functional dyspepsia; gastro-esophageal reflux disease without esophagitis	10 (1.3)
I20.1-9	Angina pectoris	4 (0.5)
I26.9	Pulmonary embolism without mention of acute cor pulmonale	3 (0.4)
J18.9	Pneumonia, unspecified	2 (0.3)
J45.9	Asthma, unspecified	2 (0.3)
I30.9	Acute pericarditis, unspecified	1 (0.1)
I63.9	Cerebral infarction, unspecified	1 (0.1)
J44.1	Chronic obstructive pulmonary disease with acute exacerbation, unspecified	1 (0.1)
N/A[c]	Other nonsevere diagnosis	35 (4.6)

[a]ICD-10: International Statistical Classification of Diseases and Related Health Problems, 10th Revision.

[b]Diagnoses reported by the physician in the electronic health record for patients with a chief complaint of chest pain who were discharged home (n=754), either directly from the emergency department (n=462) or via the day-care unit (n=292).

[c]N/A: not applicable.

XSL•FO
RenderX

Table 6. Primary discharge ICD-10[a] code and diagnosis for hospitalized patients (N=236).[b]

ICD code	Diagnosis	Values, n (%)
I20.0, I21	Acute coronary syndrome[c]	62 (26.3)
I20.1-9	Angina pectoris	14 (5.9)
I25.9	Chronic ischemic heart disease, unspecified	8 (3.4)
I05.0	Mitral stenosis	1 (0.4)
I10.9	Essential (primary) hypertension	1 (0.4)
I26-I28	Pulmonary heart disease and diseases of pulmonary circulation	6 (2.5)
I30-I52	**Other forms of heart disease**	35 (14.8)
	I30: Acute pericarditis	2 (5.7)
	I35: Nonrheumatic aortic valve disorders	3 (8.6)
	I40: Acute myocarditis	5 (14.3)
	I42: Cardiomyopathy	1 (2.9)
	I44: Atrioventricular and left bundle-branch block	1 (2.9)
	I47: Paroxysmal tachycardia	2 (5.7)
	I48: Atrial fibrillation and flutter	10 (28.6)
	I49: Other cardiac arrhythmias	3 (8.6)
	I50: Heart failure	8 (22.9)
I95-I99	Other and unspecified disorders of the circulatory system	1 (0.4)
J09-J18	Influenza and pneumonia	4 (1.7)
J20	Acute bronchitis	1 (0.4)
J44.1	Chronic obstructive pulmonary disease with acute exacerbation, unspecified	1 (0.4)
K00-93	Diseases of the digestive system	7 (3.0)
R07.4	Chest pain, unspecified	64 (27.1)
R00-09 (excluding R07.4)	Symptoms and signs involving the circulatory and respiratory systems	11 (4.7)
R10-49	Symptoms and signs involving other organ systems	1 (0.4)
R50-69	General symptoms and signs	2 (0.8)
N/A[d]	Other nonsevere diagnosis singularly reported	17 (7.2)

[a]ICD-10: International Statistical Classification of Diseases and Related Health Problems, 10th Revision.

[b]Data reported by the physician in the electronic health record for patients with a chief complaint of chest pain who were admitted from the ED (n=236). Of these, 15 patients were admitted to a noncardiac ward.

[c]Primary diagnosis of I20.0 (unstable angina pectoris) or I21 (acute myocardial infarction).

[d]N/A: not applicable.

Diagnostic Accuracy for a MACE or ACS

A clinically decisive score could be calculated using CHT-acquired data in 751 out of 1000 (75.10%; D-HEART), 727 out of 1000 (72.70%; HEART), 735 out of 1000 (73.50%; EDACS-ADP), and 838 out of 1000 (83.80%; T-MACS) patients (Figure 1). For D-HEART, 24 of 751 (3.2%) patients were included despite insufficient CHT data, based on rule-in troponin findings. Patients with insufficient CHT data to calculate the risk scores were generally younger and, for all risk scores except T-MACS, more often female (Table S2 in Multimedia Appendix 1). No adverse events occurred during the CHT interviews.

For MACE, all 4 risk scores demonstrated high NPVs of 0.99 (95% CI 0.97-1.00). Sensitivity ranged from 0.91 (95% CI 0.81-0.97; HEART) to 0.97 (95% CI 0.90-1.00; T-MACS <0.02). The MACE miss rate among patients classified as low risk was 2 out of 286 (0.70%; T-MACS), 4 out of 416 (1.0%; D-HEART), 4 out of 346 (1.2%; EDACS-ADP), and 6 out of 406 (1.5%; HEART). The proportion of patients ruled out was highest for HEART (406/727, 55.8%) and D-HEART (416/751, 55.4%), intermediate for EDACS-ADP (346/744, 46.5%), and lowest for T-MACS (286/838, 34.1%; Table 7). Using T-MACS with a threshold of <0.05 for low risk yielded a sensitivity, specificity, PPV, and NPV of 0.89 (95% CI 0.78-0.95), 0.59 (95% CI 0.55-0.62), 0.16 (95% CI 0.12-0.20), and 0.98 (95% CI 0.97-0.99), respectively. Across CHT-derived risk scores, diagnostic performance for ACS was comparable to that for

XSL•FO
RenderX

MACE (Table S3 in Multimedia Appendix 1). T-MACS <0.02 provided higher sensitivity and NPV but lower specificity compared with T-MACS <0.05 (Table S4 in Multimedia Appendix 1). No differences in AUC between the risk scores were observed for either MACE or ACS (Figure 2 and Table S5 in Multimedia Appendix 1). Excluding the 3 patients diagnosed with unstable angina pectoris who did not undergo revascularization (Table S6 in Multimedia Appendix 1), or the patient who died from a noncardiac cause, had a negligible effect on the results.

Table 7. Rule-out performance of risk scores populated with data derived from computerized history taking for a major adverse cardiac event within 30 days.[a]

Risk score	Sensitivity (95% CI)	Specificity (95% CI)	Positive predictive value (95% CI)	Negative predictive value (95% CI)	Miss rate, n/N (%)	Patients ruled out, n/N (%)
D-HEART[b] (n=751)	0.94 (0.86-0.98)	0.60 (0.57-0.64)	0.19 (0.15-0.24)	0.99 (0.98-1.00)	4/416 (0.96)	416/751 (55.39)
HEART[c] (n=727)	0.91 (0.81-0.97)	0.61 (0.57-0.64)	0.19 (0.15-0.23)	0.99 (0.97-1.00)	6/406 (1.48)	406/727 (55.85)
EDACS-ADP[d] (n=744)	0.94 (0.86-0.98)	0.51 (0.48-0.55)	0.17 (0.13-0.21)	0.99 (0.97-1.00)	4/346 (1.16)	346/744 (46.51)
T-MACS[e] (n=838)	0.97 (0.90-1.00)	0.37 (0.34-0.39)	0.12 (0.10-0.15)	0.99 (0.98-1.00)	2/286 (0.70)	286/838 (34.13)

[a]When T-MACS with a threshold of <0.05 for low risk was applied, the numbers of true positives, false positives, false negatives, and true negatives were 55, 294, 7, and 418, respectively. Sensitivity, specificity, PPV, and NPV were 0.89 (95% CI 0.78-0.95), 0.59 (95% CI 0.55-0.62), 0.16 (95% CI 0.12-0.20), and 0.98 (95% CI 0.97-0.99), respectively.

[b]D-HEART: Danderyd HEART.

[c]HEART: History, ECG, Age, Risk Factors, and Troponin.

[d]EDACS-ADP: Emergency Department Assessment of Chest Pain Score combined with an Accelerated Diagnostic Protocol.

[e]T-MACS: Troponin-only Manchester Acute Coronary Syndrome.

Figure 2. Receiver operating curves for HEART, D-HEART, EDACS-ADP and T-MACS for a 30-day major adverse cardiac event. AUC: area under the receiver operating curve; D-HEART: Danderyd HEART; EDACS-ADP: Emergency Department Assessment of Chest Pain Score combined with an Accelerated Diagnostic Protocol; HEART: History, ECG, Age, Risk Factors, and Troponin; T-MACS: Troponin-only Manchester Acute Coronary Syndrome.

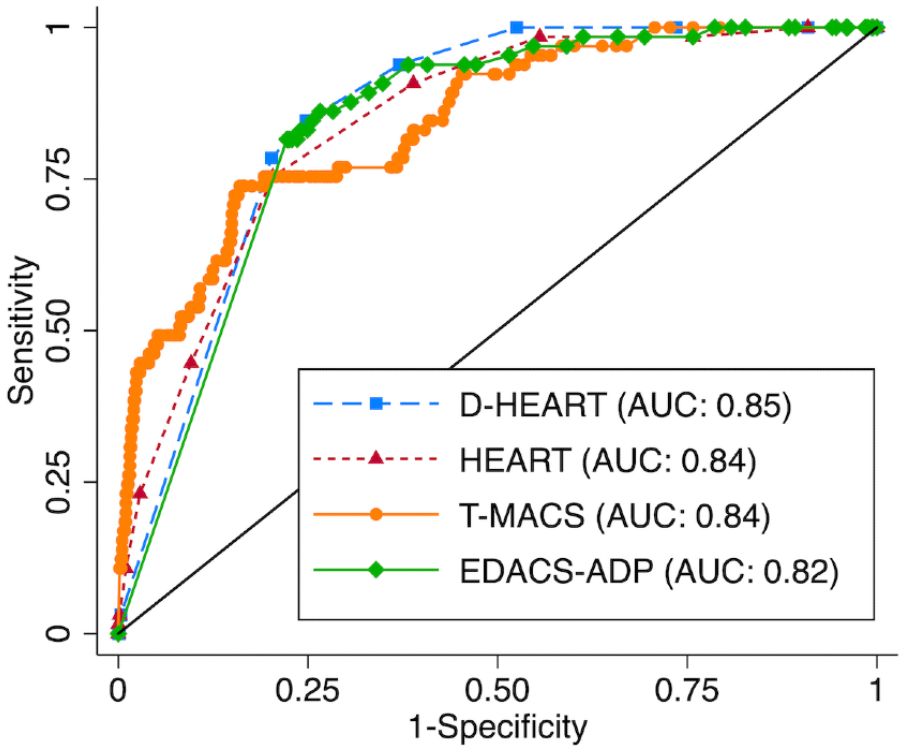

D-HEART Versus HEART

When comparing the NPV (95% CI) for ruling out MACE and ACS, no difference (P<.001) was found between the D-HEART and HEART scores: 0.99 (0.98-1.00) versus 0.99 (0.97-1.00) for MACE, and 0.99 (0.98-1.00) versus 0.99 (0.97-1.00) for ACS, respectively. Similarly, no difference in diagnostic accuracy was observed when comparing the AUCs (Table S5 in Multimedia Appendix 1).

XSL•FO
RenderX

Risk Reclassification for MACE or ACS

Of the 1000 participants, 528 were admitted, with 203 admitted to the ward and 325 to the day-care unit. Of the day-care unit admissions, 33 were subsequently admitted to the ward, and 292 were discharged home (Table 3). If CHT data had been used instead of standard management to calculate D-HEART, HEART, and EDACS-ADP, there would have been a net reclassification from "nonlow risk" to "low risk" in 89 (16.9%), 89 (16.9%), and 31 (5.9%) cases, respectively. However, due to lower specificity, use of T-MACS would instead have resulted in 91 (17.2%) patients being net reclassified from "low risk" to "nonlow risk." These reclassifications correspond to a net shift in MACE or ACS outcomes of +2 (D-HEART), +4 (HEART), +2 (EDACS-ADP), and 0 (T-MACS; Table 8).

Table 8. Reclassification and MACE[a] or ACS[b] within reclassified groups if risk scores populated with computerized history-taking data had been used for the disposition of patients instead of physician management.[c]

Reclassification	D-HEART[d]	HEART[e]	EDACS-ADP[f]	T-MACS[g]
Low to nonlow risk, n	93	97	128	220
MACE or ACS, n	1	1	2	1
Nonlow risk to low risk, n	182	186	159	129
MACE or ACS, n	3	5	4	1
Difference[h]	89	89	31	–91
MACE or ACS, n	+2	+4	+2	0
Proportion of admissions, n/N (%)	89/528 (–16.9%)	89/528 (–16.9%)	31/528 (–5.9%)	91/528 (+17.2%)

[a]MACE: major adverse cardiac event.

[b]ACS: acute coronary syndrome.

[c]Admission to the ward or day-care unit (n=528) was considered a proxy for nonlow risk, and discharge to home (n=462) a proxy for low risk. Data for 1 participant were missing, and 9 left the ED without seeing a physician.

[d]D-HEART: Danderyd HEART.

[e]HEART: History, ECG, Age, Risk Factors, and Troponin.

[f]EDACS-ADP: Emergency Department Assessment of Chest Pain Score combined with an Accelerated Diagnostic Protocol.

[g]T-MACS: Troponin-only Manchester Acute Coronary Syndrome.

[h]The difference indicates the n and % change in admissions.

Discussion

Principal Findings

This study shows that CHT can be used to calculate 4 risk scores in a majority (727-838/1000, 72.70%-83.80%) of patients presenting with acute chest pain, with good diagnostic performance for ruling out 30-day 3-point MACE or ACS, yielding NPVs of 0.99 (95% CI 0.97-1.00) for both outcomes. Sensitivities ranged from 91% to 97%, implying that, out of 100 patients, 3-9 with MACE might be misclassified. While this may initially appear concerning, the clinically more relevant miss rate among patients classified as low risk remained low for D-HEART (4/416, 0.96%) and T-MACS (2/286, 0.70%), aligning with commonly accepted safety thresholds for MACE (NPV ≥0.99 and miss rate <1%). By contrast, the HEART score (6/406, 1.5%) and EDACS-ADP (4/346, 1.2%) showed higher miss rates [16,48]. However, the lower bounds of the 95% CIs for sensitivity allow miss rates exceeding accepted rule-out safety thresholds; therefore, these findings warrant cautious interpretation.

In addition, CHT-derived chest pain risk scores reclassified a substantial proportion (up to 89/528, 16.9%) of patients admitted with acute chest pain, initially managed as suspected ACS, from "nonlow risk" to "low risk" at the time of initial ED assessment, potentially facilitating a decision for discharge rather than hospital admission. However, hospital admission may be driven by factors other than ACS risk, and reclassification to low risk does not necessarily eliminate the need for admission; thus, the aforesaid figure (ie, 89/528, 16.9%) should be viewed as a theoretical maximum, and real-world discharge gains may be substantially lower. The attendant trade-off is an increased miss rate of MACE or ACS among reclassified patients. In our study, the number of missed events (0-4 events, depending on the risk score used) was similar to that observed under standard management (2 events). While our findings indicate that CHT is a promising tool for risk score determination, further studies are warranted to validate these results and to ascertain safety.

Comparison With Prior Work

Our findings on diagnostic accuracy align with multiple validation studies in which traditional physician-acquired medical histories were used, demonstrating the efficacy of risk scores in safely ruling out MACE in the ED [7-10,31,49]. However, this appears to be the first study to report on the performance of an automated method for collecting the medical history required to calculate these risk scores. The assessed risk scores (D-HEART, HEART, EDACS-ADP, and T-MACS) showed similar performance for ruling out 30-day MACE or ACS. Our findings are consistent with prior work. A meta-analysis of 25 studies (n=25,266) reported a pooled sensitivity of 0.96 (95% CI 0.93-0.98) and an NPV of 0.99 (95% CI 0.98-0.99) for HEART [7]. A meta-analysis of the EDACS score (8 studies, n=11,578) reported a pooled sensitivity of 0.95

XSL•FO
RenderX

(95% CI 0.90-0.99) but did not specifically report NPV [50]. External validation of EDACS-ADP in a North American cohort (n=763) demonstrated a sensitivity and NPV of 1.00 (95% CI 0.94-1.00) and 1.00 (95% CI 0.99-1.00), respectively [8]. For T-MACS, secondary analyses of 4 UK cohort studies (n=1459) reported a sensitivity and NPV of 0.98 (95% CI 0.95-1.00) and 0.99 (95% CI 0.98-1.00), respectively [32]. However, the HEART score, which has undergone more extensive validation, consistently demonstrates good performance and is endorsed by international societies [5,51].

The only methodological difference between HEART and D-HEART in our study was the use of the ESC 0/1-hour troponin algorithm in D-HEART. Thus, the comparison primarily reflects the added value of serial troponin testing rather than differences related to CHT. D-HEART demonstrated equivalent performance for MACE or ACS rule out compared with the original HEART score using only a single baseline troponin measurement. This could be interpreted as support for using the simpler HEART score without repeated troponin measurements. However, this contrasts with findings showing that the original HEART score has somewhat lower diagnostic accuracy than a combination of the HEART score and a 0/1-hour troponin algorithm [52]. As the HEART score was primarily developed and validated for safe rule out of MACE, it is reasonable to assume that integrating a 0/1-hour algorithm enhances rule-in capability, as previously suggested [53]. Accordingly, the superior safety profile observed for D-HEART in this study is likely driven by the integration of serial troponin measurements rather than the CHT-derived history component itself. Nevertheless, the absolute difference in missed events was small (4 vs 6), and neither NPV nor AUC differed significantly between the scores. Thus, the difference in performance between the 2 scores may not be statistically significant. Taken together, D-HEART appears to be the most promising option for CHT-based risk stratification; however, future studies are needed to confirm these findings.

Our findings show an AUC of 0.82-0.85 and an NPV of 0.98-0.99, depending on the risk score used. While there is no universal benchmark for AUC or NPV in risk stratification, an AUC of 0.8-0.9 is generally considered excellent, and >0.9 outstanding [54]. Interpretation of NPV varies by clinical context; for MACE or ACS, a high NPV is desirable for safe rule out [55]. Overall, our findings suggest that risk calculation using CHT data performs at least as well as traditional history taking. It is reassuring that all scores reproduced similar results. However, further refinement of CHT data collection could potentially enhance MACE rule-out capability. Notably, a strategy in which CHT data are used for risk calculation requires further validation. Importantly, risk score algorithms should always be considered as decision-support tools and used in complement to physician expertise.

Clinical Implications

A significant fraction (up to 89/528, 16.9%) of participants were reclassified from “nonlow risk” to “low risk” when using D-HEART, HEART, and EDACS-ADP populated with CHT data, compared with standard management. These findings align with those of an external validation study, which reported a reduction in hospital admissions for patients with acute chest pain of approximately 15% following implementation of the HEART score [56]. Improvements in classification may reduce anxiety, unnecessary examinations, and health care resource use [57]. However, a large Dutch multicenter randomized controlled trial observed a limited impact of risk score calculations on resource utilization, attributing this effect to nonadherence to management recommendations [58]. Notably, when using T-MACS, more patients would be reclassified from “low risk” to “nonlow risk,” potentially leading to increased admissions (91/528, +17.2%) compared with standard care, due to its lower specificity. Applying a higher threshold for T-MACS (<0.05 for “low risk”) increases specificity but results in a lower NPV, supporting the conclusion that this is not the preferred score for patient disposition. Further studies on optimal CHT implementation and its impact on resource utilization and quality of care in chest pain management are warranted.

This study extends our previous reports demonstrating substantial agreement between data collected using CHT and standard history taking for risk score assessment [30]. However, for the HEART score, the History component was derived from predefined patient-reported CHT criteria rather than physician gestalt (ie, physician’s subjective clinical judgment), as applied in the original derivation studies. Given the previously demonstrated low-to-moderate agreement between CHT and EHR data for chest pain characteristics [30], this substitution may affect construct validity and limit direct comparability with physician-acquired HEART scores. At the same time, prior work suggests that CHT captures more complete and detailed symptom data than are routinely documented by physicians in the EHR, which may partly explain the observed discrepancies. Our previous studies also showed that risk score calculation was feasible to a much greater extent when using CHT-collected data compared with physician-reported EHR data (74%-83% vs 10%-31%, depending on the risk score used), and that a HEART score excluding the troponin variable (HEART score) can effectively predict clinical outcomes [54]. This analysis underscores the broader utility of CHT in facilitating the calculation of established risk scores commonly used in clinical practice. Moreover, this study introduces the concept of a clinically decisive risk score, defined as reaching a specific threshold for “nonlow risk” (eg, D-HEART≥4 or T-MACS≥0.02), to better reflect how these risk stratification tools are applied in practice.

The time required to obtain sufficient data for risk score calculation using CHT (median 23 minutes) may appear long. However, the intention of CHT was not to replace physician history taking, but rather to use waiting time in the ED to collect and supplement crucial information for risk stratification. Notably, the median length of stay for low- to intermediate-risk patients (ie, RETTS orange, yellow, green, and blue) in the ED at the time of the study was approximately 4 hours, suggesting that CHT can be completed without disrupting workflow. Nevertheless, future developments of CHT should focus on shortening interview duration and adapting content to the clinical context and urgency. Moreover, although physician-interpreted ECGs were used by design to isolate the effect of CHT on chest

XSL•FO
RenderX

pain management, the incorporation of automated ECG analysis may further enhance efficiency.

Considering the high volume of patients presenting to the ED with chest pain, automating risk stratification through CHT could streamline the ED pathway: first, by ensuring complete and structured capture of medical history and risk factors; second, by reducing physician time spent on data gathering; third, by providing real-time support for clinical decision-making; and finally, by enabling faster and safer disposition decisions that identify patients suitable for discharge and reduce resource use.

Strengths and Limitations

Key strengths of this study include its prospective cohort design, a large and representative ED chest pain population, evaluation under real-world conditions in the typical clinical environment in which the tool would be used, and reliable outcomes defined by strict, well-established criteria. The consistency of the findings across multiple validated risk scores underscores the robustness of the results. Furthermore, the generic layout of the CHT software, which is not limited to cardiology or the ED setting, suggests that these findings may apply to other clinical contexts. Finally, the study was initiated and conducted within an academic setting, without any commercial interests.

However, several limitations warrant consideration. First, the single-center design introduces a risk of selection bias, potentially limiting population diversity. In addition, the use of a convenience sample comprising a minority of eligible chest pain presentations may limit external validity, as enrolled patients may be more compliant, clinically stable, or technologically literate than the overall ED population. This bias may have been further exacerbated by the exclusion of patients unable to complete CHT and by recruitment occurring mainly during office hours and evenings, due to research staff availability. However, the proportion of patients with chest pain presenting at night was small, and the demographics of study participants were similar to those of the overall ED chest pain population, reducing the likelihood of major systematic differences. Nevertheless, our findings should be interpreted with caution, and future large multicenter studies are warranted to ascertain the generalizability of these results to other health care settings. Second, variability in physician performance and the unique ED setup, which includes a dedicated cardiology unit and a day-care unit staffed by cardiologists, may lead to an underestimation of CHT performance compared with settings with less specialist expertise in the ED. At the same time, this organizational structure may limit the generalizability of our findings to more diverse clinical environments. Third, risk score determination using CHT data was not feasible for approximately one-quarter of participants. We have previously shown that premature termination of the CHT interview was associated with ED discharge and fatigue, most commonly among older adults (≥70 years), individuals born outside Europe, and retirees, thereby limiting conclusions for these groups [22]. In our analysis, however, patients with insufficient CHT data were significantly younger than those with a clinically decisive score (mean age 51.0 vs 56.2 years; see Table S2 in Multimedia Appendix 1), likely reflecting the larger proportion of adults aged 18-69 years in the study population and their higher frequency of interview termination, often associated with early discharge [22]. Nevertheless, diagnostic performance was assessed only in patients who provided sufficient CHT data for risk score determination, which may have overestimated real-world performance. This finding underscores the need for a more user-friendly tool to reduce missing data in routine care. Fourth, ACS outcomes were adjudicated by a single board-certified cardiologist, which may introduce adjudication bias compared with evaluation by 2 or more independent specialists. However, Swedish discharge ICD codes have a high PPV (85%-95%) for acute myocardial infarction and related diagnoses, supporting the reliability of the outcome classification [45,46]. Further, outcomes were retrieved through passive review of the shared EHR system in Region Stockholm; therefore, events occurring outside the region may not have been captured, potentially leading to an overestimation of the NPV. Mortality status is, however, captured through linkage between the regional EHR system and the national population registry. The proportion of nonfatal events is likely small, as routine follow-up after a MACE is systematically conducted at the patient's home hospital, even when the event occurs outside the region. Fifth, given that patients with MACE were significantly older than those without (ie, non-MACE; 67.6 vs 53.7 years), there is a risk of spectrum bias if CHT fails to capture data from patients at the highest risk. In our cohort, however, insufficient CHT data were not more common among older patients (Table S2 in Multimedia Appendix 1), reducing concerns that missing CHT data systematically occurred in older or higher-risk individuals. Sixth, inclusion in the study could occur either before or after the patient had been evaluated by a physician. This may have introduced recall bias, as patients might provide different responses when reporting their history a second time. The effect of interview order on patient responses was not assessed and therefore remains an unmitigated limitation. Seventh, patient disposition reflects not only the medical history obtained by the physician but also additional information available at the time of assessment, including ECG findings, blood test results, and prior EHR documentation. These factors likely influenced clinical decision-making, thereby limiting direct comparability with CHT-derived risk scores. However, this approach was intentionally chosen to evaluate the potential impact of CHT within current clinical management. Future multimodal studies should aim to incorporate such data directly into risk stratification models. Eighth, the primary focus of this study was the safe rule out of MACE or ACS using a clinically decisive risk score. This implies that once patients exceed the threshold for "nonlow risk" (eg, HEART ≥4), they are considered ruled in. However, some patients within this clinically decisive risk group who did not complete the CHT interview might have achieved even higher total risk scores. Consequently, our findings may be more robust for rule out than for rule in. Conversely, excluding patients with insufficient CHT data who also did not meet ECG or troponin criteria for "nonlow risk" may have resulted in the systematic exclusion of a specific patient subset, potentially leading to an overestimation of NPVs. Accordingly, interpretation of both rule-out and rule-in performance should account for these methodological limitations, and further evaluation across the full risk spectrum

XSL•FO
RenderX

is warranted. Finally, this study carries a potential risk of verification bias. We believe this risk to be minimal, as patients were managed according to standard care protocols independent of CHT-derived risk scores, which were not available to the treating physicians during clinical decision-making.

Future Directions

Future research on CHT in chest pain risk stratification should include customizing CHT for the ED setting to reduce interview time, extending its use to early prehospital triage, and broadening its scope to support diagnosis of a wider array of acute conditions, such as pulmonary embolism, aortic dissection, and other noncardiac causes of chest pain. In addition, future studies should explore the integration of multimodal variables into risk stratification models, and the incorporation of multiple languages into the CHT software could enable history taking in patients' native languages.

Conclusions

Automated medical history taking using CHT can generate reliable risk scores in the majority of patients presenting with acute chest pain, with good diagnostic performance for ruling out 30-day 3-point MACE (ie, ACS diagnosis, revascularization, or cardiovascular death) and ACS. CHT-derived risk scores performed comparably to previously validated physician-acquired scores, and accepted safety thresholds were met by D-HEART and T-MACS, whereas the HEART score and EDACS-ADP showed values slightly above the commonly accepted <1% threshold; however, given the wide sensitivity CIs, these findings should be interpreted with caution. Compared with the HEART score, the improved safety of D-HEART appears to be primarily attributable to the incorporation of serial 0/1-hour troponin testing rather than the CHT-derived history itself. Application of CHT-derived risk scores indicated that 89 of the 528 (16.9%) patients could potentially be reclassified from "nonlow risk" to "low risk" at the time of initial ED assessment, which may support discharge decisions in selected patients, although admission may still be required for reasons unrelated to ACS. Consistent with prior chest pain risk score studies, any gains in discharge rates must be balanced against the risk of missed events among reclassified patients. Among the scores assessed, D-HEART appears most promising for CHT-based risk stratification. However, further large multicenter studies are warranted to confirm these findings before CHT can be established in routine clinical practice.

Acknowledgments

The authors thank Professor D. Zakim, Karolinska Institutet, Stockholm, Sweden, developer of the expert system software (CLEOS) used for CHT and a contributor to the trial protocol, for invaluable support. The authors also thank statistician F. K. Johansson, MSc, Karolinska Institutet, Stockholm, Sweden, for expert statistical advice. This work was supported by the Robert Bosch Stiftung, Stuttgart, Germany (grant 11.5.1000.0258.0); Region Stockholm, ALF project, Stockholm, Sweden (grants FoUI-954046 and FoUI-962663); Stiftelsen Hjärtat, Stockholm, Sweden; Fond 176, Danderyd University Hospital, Stockholm, Sweden; and institutional funds from the Department of Learning, Informatics, Management, and Ethics, Karolinska Institutet, Stockholm, Sweden.

Data Availability

The datasets used or analyzed during this study are available from the corresponding author (HB) on reasonable request.

Conflicts of Interest

All patent rights and copyrights to the technology, language, images, and knowledge content are assigned, without royalty rights, to Karolinska Institutet, Stockholm, Sweden, a public university. HB reports speaker honoraria from AstraZeneca. TK reports research grants to Karolinska Institutet from Amgen, Medtronic, and ReCor Medical, all outside the submitted work. JS reports speaker honoraria from AstraZeneca, Bayer, Novo Nordisk, and Medtronic, and shareholdings in Beats Vascular Health.

Multimedia Appendix 1
Additional analysis.
[DOCX File , 52 KB - jmir_v28i1e76087_app1.docx]

References

1. Aalam AA, Alsabban A, Pines JM. National trends in chest pain visits in US emergency departments (2006-2016). Emerg Med J 2020 Nov;37(11):696-699. [doi: 10.1136/emermed-2020-210306] [Medline: 32900857]
2. Bjørnsen LP, Naess-Pleym LE, Dale J, Grenne B, Wiseth R. Description of chest pain patients in a Norwegian emergency department. Scand Cardiovasc J 2019 Feb;53(1):28-34. [doi: 10.1080/14017431.2019.1583362] [Medline: 30760035]
3. Af Ugglas B, Djärv T, Ljungman PLS, Holzmann MJ. Emergency department crowding associated with increased 30-day mortality: a cohort study in Stockholm Region, Sweden, 2012 to 2016. J Am Coll Emerg Physicians Open 2020 Dec;1(6):1312-1319 [FREE Full text] [doi: 10.1002/emp2.12243] [Medline: 33392538]

XSL•FO
RenderX

4. Byrne RA, Rossello X, Coughlan JJ, Barbato E, Berry C, Chieffo A, ESC Scientific Document Group. 2023 ESC Guidelines for the management of acute coronary syndromes. Eur Heart J 2023 Oct 12;44(38):3720-3826. [doi: 10.1093/eurheartj/ehad191] [Medline: 37622654]
5. Stepinska J, Lettino M, Ahrens I, Bueno H, Garcia-Castrillo L, Khoury A, et al. Diagnosis and risk stratification of chest pain patients in the emergency department: focus on acute coronary syndromes. A position paper of the Acute Cardiovascular Care Association. Eur Heart J Acute Cardiovasc Care 2020 Feb;9(1):76-89 [FREE Full text] [doi: 10.1177/2048872619885346] [Medline: 31958018]
6. Writing Committee, Kontos MC, de Lemos JA, Deitelzweig SB, Diercks DB, Gore MO, et al. 2022 ACC expert consensus decision pathway on the evaluation and disposition of acute chest pain in the emergency department: a report of the American College of Cardiology Solution Set Oversight Committee. J Am Coll Cardiol 2022 Nov 15;80(20):1925-1960 [FREE Full text] [doi: 10.1016/j.jacc.2022.08.750] [Medline: 36241466]
7. Laureano-Phillips J, Robinson RD, Aryal S, Blair S, Wilson D, Boyd K, et al. HEART score risk stratification of low-risk chest pain patients in the emergency department: a systematic review and meta-analysis. Ann Emerg Med 2019 Aug;74(2):187-203. [doi: 10.1016/j.annemergmed.2018.12.010] [Medline: 30718010]
8. Flaws D, Than M, Scheuermeyer FX, Christenson J, Boychuk B, Greenslade JH, et al. External validation of the emergency department assessment of chest pain score accelerated diagnostic pathway (EDACS-ADP). Emerg Med J 2016 Sep;33(9):618-625. [doi: 10.1136/emermed-2015-205028] [Medline: 27406833]
9. Meier M, Boeddinghaus J, Nestelberger T, Koechlin L, Lopez-Ayala P, Wussler D, Miró, APACE investigators. Comparing the utility of clinical risk scores and integrated clinical judgement in patients with suspected acute coronary syndrome. Eur Heart J Acute Cardiovasc Care 2023 Oct 25;12(10):693-702 [FREE Full text] [doi: 10.1093/ehjacc/zuad081] [Medline: 37435949]
10. Steiro O, Tjora HL, Langørgen J, Bjørneklett R, Nygård OK, Skadberg, et al. Clinical risk scores identify more patients at risk for cardiovascular events within 30 days as compared to standard ACS risk criteria: the WESTCOR study. Eur Heart J Acute Cardiovasc Care 2021 May 11;10(3):287-301. [doi: 10.1093/ehjacc/zuaa016] [Medline: 33620429]
11. Pena M, Tsao J, Falaiye T, Kazan V, Li M, Daouk F, et al. Why won't emergency physicians discharge patients with a low HEART score from an observation unit without further evaluation? Crit Pathw Cardiol 2020 Dec;19(4):195-199. [doi: 10.1097/HPC.0000000000000236] [Medline: 32769484]
12. Ras M, Reitsma JB, Hoes AW, Six AJ, Poldervaart JM. Secondary analysis of frequency, circumstances and consequences of calculation errors of the HEART (history, ECG, age, risk factors and troponin) score at the emergency departments of nine hospitals in the Netherlands. BMJ Open 2017 Oct 22;7(10):e017259 [FREE Full text] [doi: 10.1136/bmjopen-2017-017259] [Medline: 29061617]
13. Wu WK, Yiadom MYAB, Collins SP, Self WH, Monahan K. Documentation of HEART score discordance between emergency physician and cardiologist evaluations of ED patients with chest pain. Am J Emerg Med 2017 Jan;35(1):132-135 [FREE Full text] [doi: 10.1016/j.ajem.2016.09.058] [Medline: 27745728]
14. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689 [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
15. Lopez-Jimenez F, Attia Z, Arruda-Olson AM, Carter R, Chareonthaitawee P, Jouni H, et al. Artificial intelligence in cardiology: present and future. Mayo Clin Proc 2020 May;95(5):1015-1039. [doi: 10.1016/j.mayocp.2020.01.038] [Medline: 32370835]
16. Gulati M, Levy PD, Mukherjee D, Amsterdam E, Bhatt DL, Birtcher KK, et al. 2021 AHA/ACC/ASE/CHEST/SAEM/SCCT/SCMR guideline for the evaluation and diagnosis of chest pain: executive summary: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. Circulation 2021 Nov 30;144(22):e368-e454 [FREE Full text] [doi: 10.1161/CIR.0000000000001030] [Medline: 34709928]
17. Zakim D, Braun N, Fritz P, Alscher MD. Underutilization of information and knowledge in everyday medical practice: evaluation of a computer-based solution. BMC Med Inform Decis Mak 2008 Nov 05;8:50 [FREE Full text] [doi: 10.1186/1472-6947-8-50] [Medline: 18983684]
18. Zakim D, Fritz C, Braun N, Fritz P, Alscher MD. Computerized history-taking as a tool to manage dyslipidemia. Vasc Health Risk Manag 2010 Nov 12;6:1039-1046 [FREE Full text] [doi: 10.2147/VHRM.S14302] [Medline: 21127700]
19. Zakim D, Brandberg H, El Amrani S, Hultgren A, Stathakarou N, Nifakos S, et al. Computerized history-taking improves data quality for clinical decision-making-comparison of EHR and computer-acquired history data in patients with chest pain. PLoS One 2021;16(9):e0257677 [FREE Full text] [doi: 10.1371/journal.pone.0257677] [Medline: 34570811]
20. Hourizadeh J, Zeltser R, Makaryus AN, Druz R. Real-world evaluation of a chest pain digital triage platform at the point of care. South Med J 2023 Nov;116(11):857-862. [doi: 10.14423/SMJ.0000000000001620] [Medline: 37913803]
21. Sundberg K, Adeli A, Brandberg H, Spaak J, Koch S, Sundberg CJ, et al. User experience of self-reported computerized medical history taking for acute chest pain: The Clinical Expert Operating System Chest Pain Danderyd Study. Health Expect 2022 Dec;25(6):3053-3061 [FREE Full text] [doi: 10.1111/hex.13612] [Medline: 36148691]
22. Brandberg H, Sundberg CJ, Spaak J, Koch S, Zakim D, Kahan T. Use of self-reported computerized medical history taking for acute chest pain in the emergency department - the Clinical Expert Operating System Chest Pain Danderyd Study

XSL•FO
RenderX

(CLEOS-CPDS): prospective cohort study. J Med Internet Res 2021 Apr 27;23(4):e25493 [FREE Full text] [doi: 10.2196/25493] [Medline: 33904821]
23. Fritzell K, Brandberg H, Spaak J, Koch S, Sundberg CJ, Zakim D, et al. Patient perspectives on digital technology and experiences of computerized history-taking for chest pain management in the emergency department: CLEOS-CPDS prospective cohort study. JMIR Med Inform 2025 Jun 17;13:e65568-e65568 [FREE Full text] [doi: 10.2196/65568] [Medline: 40526888]
24. Berdahl CT, Henreid AJ, Pevnick JM, Zheng K, Nuckols TK. Digital tools designed to obtain the history of present illness from patients: scoping review. J Med Internet Res 2022 Nov 17;24(11):e36074 [FREE Full text] [doi: 10.2196/36074] [Medline: 36394945]
25. Brandberg H, Kahan T, Spaak J, Sundberg K, Koch S, Adeli A, et al. A prospective cohort study of self-reported computerised medical history taking for acute chest pain: protocol of the CLEOS-Chest Pain Danderyd Study (CLEOS-CPDS). BMJ Open 2020 Jan 21;10(1):e031871 [FREE Full text] [doi: 10.1136/bmjopen-2019-031871] [Medline: 31969363]
26. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015 Oct 28;351:h5527 [FREE Full text] [doi: 10.1136/bmj.h5527] [Medline: 26511519]
27. Widgren BR, Jourak M. Medical Emergency Triage and Treatment System (METTS): a new protocol in primary triage and secondary priority decision in emergency medicine. J Emerg Med 2011 Jun;40(6):623-628. [doi: 10.1016/j.jemermed.2008.04.003] [Medline: 18930373]
28. Törnerud M, Brandberg H. Akut Hjärtsjukvård (Acute Cardiac Care). Kunskapsstod for vardgivare.: Region Stockholm URL: https://kunskapsstodforvardgivare.se/omraden/hjart--och-karlsjukdomar/regionala-kunskapsstod/akut-hjartsjukvard [accessed 2026-01-10]
29. Zakim D. Development and significance of automated history-taking software for clinical medicine, clinical research and basic medical science. J Intern Med 2016 Sep;280(3):287-299 [FREE Full text] [doi: 10.1111/joim.12509] [Medline: 27071980]
30. Brandberg H, Sundberg CJ, Spaak J, Koch S, Kahan T. Are medical history data fit for risk stratification of patients with chest pain in emergency care? Comparing data collected from patients using computerized history taking with data documented by physicians in the electronic health record in the CLEOS-CPDS prospective cohort study. J Am Med Inform Assoc 2024 Jun 20;31(7):1529-1539 [FREE Full text] [doi: 10.1093/jamia/ocae110] [Medline: 38781350]
31. Than M, Flaws D, Sanders S, Doust J, Glasziou P, Kline J, et al. Development and validation of the Emergency Department Assessment of Chest pain Score and 2 h accelerated diagnostic protocol. Emerg Med Australas 2014 Feb;26(1):34-44. [doi: 10.1111/1742-6723.12164] [Medline: 24428678]
32. Body R, Carlton E, Sperrin M, Lewis PS, Burrows G, Carley S, et al. Troponin-only Manchester Acute Coronary Syndromes (T-MACS) decision aid: single biomarker re-derivation and external validation in three cohorts. Emerg Med J 2017 Jun;34(6):349-356 [FREE Full text] [doi: 10.1136/emermed-2016-205983] [Medline: 27565197]
33. Six AJ, Backus BE, Kelder JC. Chest pain in the emergency room: value of the HEART score. Neth Heart J 2008 Jun;16(6):191-196. [doi: 10.1007/bf03086144]
34. Diamond GA, Forrester JS. Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease. N Engl J Med 1979 Jun 14;300(24):1350-1358. [doi: 10.1056/NEJM197906143002402] [Medline: 440357]
35. McCord J, Cabrera R, Lindahl B, Giannitsis E, Evans K, Nowak R, TRAPID-AMI Investigators. Prognostic utility of a modified HEART score in chest pain patients in the emergency department. Circ Cardiovasc Qual Outcomes 2017 Feb;10(2):1-8. [doi: 10.1161/CIRCOUTCOMES.116.003101] [Medline: 28167641]
36. Carlton EW, Ingram J, Taylor H, Glynn J, Kandiyali R, Campbell S, et al. Limit of detection of troponin discharge strategy versus usual care: randomised controlled trial. Heart 2020 Oct;106(20):1586-1594 [FREE Full text] [doi: 10.1136/heartjnl-2020-316692] [Medline: 32371401]
37. Backus BE, Six AJ, Kelder JC, Bosschaert MAR, Mast EG, Mosterd A, et al. A prospective validation of the HEART score for chest pain patients at the emergency department. Int J Cardiol 2013 Oct 3;168(3):2153-2158 [FREE Full text] [doi: 10.1016/j.ijcard.2013.01.255] [Medline: 23465250]
38. Reichlin T, Schindler C, Drexler B, Twerenbold R, Reiter M, Zellweger C, et al. One-hour rule-out and rule-in of acute myocardial infarction using high-sensitivity cardiac troponin T. Arch Intern Med 2012 Sep 10;172(16):1211-1218. [doi: 10.1001/archinternmed.2012.3698] [Medline: 22892889]
39. Body R, Mueller C, Giannitsis E, Christ M, Ordonez-Llanos J, de Filippi CR, TRAPID-AMI Investigators. The use of very low concentrations of high-sensitivity troponin T to rule out acute myocardial infarction using a single blood test. Acad Emerg Med 2016 Sep;23(9):1004-1013 [FREE Full text] [doi: 10.1111/acem.13012] [Medline: 27178492]
40. Giannitsis E, Kurz K, Hallermayer K, Jarausch J, Jaffe AS, Katus HA. Analytical validation of a high-sensitivity cardiac troponin T assay. Clin Chem 2010 Feb;56(2):254-261. [doi: 10.1373/clinchem.2009.132654] [Medline: 19959623]
41. Roffi M, Patrono C, Collet J, Mueller C, Valgimigli M, Andreotti F, Erol, et al. 2015 ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation: Task Force for the Management of Acute Coronary Syndromes in Patients Presenting without Persistent ST-Segment Elevation of the European Society of Cardiology (ESC). Eur Heart J 2016 Jan 14;37(3):267-315. [doi: 10.1093/eurheartj/ehv320] [Medline: 26320110]

XSL•FO
RenderX

42. Ibanez B, James S, Agewall S, Antunes MJ, Bucciarelli-Ducci C, Bueno H, ESC Scientific Document Group. 2017 ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation: The Task Force for the management of acute myocardial infarction in patients presenting with ST-segment elevation of the European Society of Cardiology (ESC). Eur Heart J 2018 Jan 07;39(2):119-177 [FREE Full text] [doi: 10.1093/eurheartj/ehx393] [Medline: 28886621]
43. Bezin J, Girodet P, Rambelomanana S, Touya M, Ferreira P, Gilleron V, et al. Choice of ICD-10 codes for the identification of acute coronary syndrome in the French hospitalization database. Fundam Clin Pharmacol 2015 Dec;29(6):586-591. [doi: 10.1111/fcp.12143] [Medline: 26301735]
44. International Statistical Classification of Diseases and Related Health Problems 10th Revision 2023. World Health Organization. URL: https://icd.who.int/browse10/2019/en [accessed 2025-03-01]
45. Ludvigsson JF, Andersson E, Ekbom A, Feychting M, Kim J, Reuterwall C, et al. External review and validation of the Swedish national inpatient register. BMC Public Health 2011 Jun 09;11:450 [FREE Full text] [doi: 10.1186/1471-2458-11-450] [Medline: 21658213]
46. Hammar N, Alfredsson L, Rosén M, Spetz CL, Kahan T, Ysberg AS. A national record linkage to study acute myocardial infarction incidence and case fatality in Sweden. Int J Epidemiol 2001 Oct;30 Suppl 1:S30-S34. [doi: 10.1093/ije/30.suppl_1.s30] [Medline: 11759848]
47. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983 Sep;148(3):839-843. [doi: 10.1148/radiology.148.3.6878708] [Medline: 6878708]
48. Than M, Herbert M, Flaws D, Cullen L, Hess E, Hollander JE, et al. What is an acceptable risk of major adverse cardiac event in chest pain patients soon after discharge from the Emergency Department?: a clinical survey. Int J Cardiol 2013 Jul 1;166(3):752-754. [doi: 10.1016/j.ijcard.2012.09.171] [Medline: 23084108]
49. Six AJ, Cullen L, Backus BE, Greenslade J, Parsonage W, Aldous S, et al. The HEART score for the assessment of patients with chest pain in the emergency department: a multinational validation study. Crit Pathw Cardiol 2013 Sep;12(3):121-126. [doi: 10.1097/HPC.0b013e31828b327e] [Medline: 23892941]
50. Boyle RSJ, Body R. The diagnostic accuracy of the Emergency Department Assessment of Chest Pain (EDACS) score: a systematic review and meta-analysis. Ann Emerg Med 2021 Apr;77(4):433-441. [doi: 10.1016/j.annemergmed.2020.10.020] [Medline: 33461885]
51. Liu N, Ng JCJ, Ting CE, Sakamoto JT, Ho AFW, Koh ZX, et al. Clinical scores for risk stratification of chest pain patients in the emergency department: an updated systematic review. J Emerg Crit Care Med 2018 Feb 7;2:16-16. [doi: 10.21037/jeccm.2018.01.10]
52. Nilsson T, Johannesson E, Lundager Forberg J, Mokhtari A, Ekelund U. Diagnostic accuracy of the HEART Pathway and EDACS-ADP when combined with a 0-hour/1-hour hs-cTnT protocol for assessment of acute chest pain patients. Emerg Med J 2021 Nov;38(11):808-813. [doi: 10.1136/emermed-2020-210833] [Medline: 33837120]
53. Cortés M, Haseeb S, Lambardi F, Arbucci R, Ariznavarreta P, Resi S, et al. The HEART score in the era of the European Society of Cardiology 0/1-hour algorithm. Eur Heart J Acute Cardiovasc Care 2020 Feb;9(1):30-38. [doi: 10.1177/2048872619883619] [Medline: 31657616]
54. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol 2010 Sep;5(9):1315-1316 [FREE Full text] [doi: 10.1097/JTO.0b013e3181ec173d] [Medline: 20736804]
55. Morrow DA. Clinician's guide to early rule-out strategies with high-sensitivity cardiac troponin. Circulation 2017 Apr 25;135(17):1612-1616. [doi: 10.1161/CIRCULATIONAHA.117.026717] [Medline: 28438803]
56. Hyams JM, Streitz MJ, Oliver JJ, Wood RM, Maksimenko YM, Long B, et al. Impact of the HEART pathway on admission rates for emergency department patients with chest pain: an external clinical validation study. J Emerg Med 2018 Apr;54(4):549-557. [doi: 10.1016/j.jemermed.2017.12.038] [Medline: 29478861]
57. Alley W, Mahler SA. Clinical decision aids for chest pain in the emergency department: identifying low-risk patients. Open Access Emerg Med 2015;7:85-92 [FREE Full text] [doi: 10.2147/OAEM.S71282] [Medline: 27147894]
58. Poldervaart JM, Reitsma JB, Backus BE, Koffijberg H, Veldkamp RF, Ten Haaf ME, et al. Effect of using the HEART score in patients with chest pain in the emergency department: a stepped-wedge, cluster randomized trial. Ann Intern Med 2017 May 16;166(10):689-697. [doi: 10.7326/M16-1600] [Medline: 28437795]

Abbreviations

ACS: acute coronary syndrome
AUC: area under the receiver operating characteristic curve
CHT: computerized history taking
CLEOS-CPDS: Clinical Expert Operating System Chest Pain Danderyd Study
D-HEART: Danderyd HEART
ECG: electrocardiogram
ED: emergency department

XSL•FO
RenderX

EDACS-ADP: Emergency Department Assessment of Chest Pain Score combined with an Accelerated Diagnostic Protocol
EHR: electronic health record
ESC: European Society of Cardiology
HEART: History, ECG, Age, Risk Factors, and Troponin
hs-cTn: high-sensitivity cardiac troponin
ICD-10: International Statistical Classification of Diseases and Related Health Problems, 10th Revision
MACE: major adverse cardiac event
NPV: negative predictive value
PPV: positive predictive value
RETTS: Rapid Emergency Triage and Treatment System
STARD: Standards for Reporting Diagnostic Accuracy
T-MACS: Troponin-only Manchester Acute Coronary Syndrome

Edited by A Coristine; submitted 16.Apr.2025; peer-reviewed by ME Matsumura, S van den Bulk; comments to author 17.Jul.2025; revised version received 08.Jan.2026; accepted 14.Jan.2026; published 11.Feb.2026.

Please cite as:
Brandberg H, Sundberg CJ, Spaak J, Koch S, Kahan T
Computerized Self-Reported Medical History Taking to Support Early Rule Out of Major Adverse Cardiac Events in Patients With Acute Chest Pain: Post Hoc Analysis of the CLEOS-CPDS Prospective Cohort Study
J Med Internet Res 2026;28:e76087
URL: https://www.jmir.org/2026/1/e76087
doi:10.2196/76087
PMID:

©Helge Brandberg, Carl Johan Sundberg, Jonas Spaak, Sabine Koch, Thomas Kahan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Crowdsourcing Medical School Admissions Data: Development and Analysis of the CycleTrack Platform

Daniel Boris Amusin[1], BS; Isabelle Hua[2], BA, BS; Arthi Shankar Kozhumam[1], MScGH

[1]Medical Scientist Training Program, Feinberg School of Medicine, Northwestern University, 303 E Chicago Ave, Morton 1-670, Chicago, IL, United States

[2]Medical Scientist Training Program, School of Medicine, University of Colorado, Aurora, CO, United States

Corresponding Author:
Daniel Boris Amusin, BS
Medical Scientist Training Program, Feinberg School of Medicine, Northwestern University, 303 E Chicago Ave, Morton 1-670, Chicago, IL, United States

Abstract

Using a crowdsource model, we developed CycleTrack, a web-based platform that tracks and aggregates medical school application data in real-time. From May 2022 to January 2026, 34,763 users registered on the platform, of whom 20,354 tracked a cumulative 415,837 doctor of medicine (MD), MD–doctor of philosophy (PhD), doctor of osteopathic medicine (DO), and DO-PhD applications. The volume of applications tracked per program strongly correlated with the total number of applications reported in official data from the Association of American Medical Colleges and American Association of Colleges of Osteopathic Medicine (Spearman $\rho \geq 0.91$; $P<.001$); however, statistically significant differences in demographics, grade point average, and Medical College Admission Test (MCAT) scores were observed between CycleTrack users and national averages. Despite this, the CycleTrack database accurately captured the cadence of interview invitations from the Northwestern University and University of Michigan MD programs. Aggregated data from the CycleTrack database offered a glimpse into application cycle dynamics, such as demonstrating the relatively longer DO interview cycle compared to MD and MD-PhD programs. Use data from CycleTrack suggest demand for open and transparent information about statistics and timing of admissions decisions among medical school applicants.

(J Med Internet Res 2026;28:e83087) doi:10.2196/83087

KEYWORDS

undergraduate medical education; medical school admission; data accessibility; crowdsourcing; CycleTrack

Introduction

Medical school admission in the United States is highly competitive, with matriculation rates below 45% [1,2]. Applicants frequently seek to optimize acceptance chances by consulting online resources. While select data are available through the Medical School Admissions Requirements and Choose DO Explorer tools, there are remaining information gaps in the timing of rolling admissions decisions and data specific to dual-degree applicants. Applicants have frequently turned to online forums such as the Student Doctor Network or Reddit to share and discuss this information [3]. However, anecdotal experiences from individual users may not reflect empirical evidence and mislead applicants, which may especially discourage low-income and minority background applicants [4,5].

In the absence of official data, crowdsourcing information may support applicants. Tools such as the American Medical College Application Service (AMCAS) Tracker [6] visualize the timing of application verification, and MDApplicants [7] displays individual applicant profiles. These tools lack key information such as the timing of interview invitations and admissions decisions. An early attempt to aggregate interview invitation dates by the Student Doctor Network user “TheDataKing” relied on end-of-cycle surveys that required entry into a Google Form [8]. However, the lack of timeliness and the need for manual entry, curation, and visualization limited uptake.

Empowering applicants requires novel methods for increasing data availability. Here, we describe CycleTrack, a novel platform that bridges the individual application tracking spreadsheet with transparent automated data aggregation.

Methods

Overview

CycleTrack is a free medical school admissions tracking platform built using the Python Flask framework [9,10]. Applicants register on the website, create school lists, share optional demographic/statistical information, and enter dates of actions such as interview invitations or acceptances. Data are collected using a standardized interface from email-verified users and stored in a SQLite3 database. Input validation ensures

XSL•FO
RenderX

that data fit AMCAS demographic categories, grade point average and Medical College Admission Test (MCAT) scores are within appropriate bounds, and dates are between May of the application year and August of the matriculation year. The database is queried hourly to generate publicly visible summary statistics/graphs on school-specific data explorer pages. To maintain privacy, summary statistics are only reported when at least 5 points are available, and dates of actions are never linked to other user information.

To characterize CycleTrack users, demographics, and scores, applications from the platform were compared to national data provided by the Association of American Medical Colleges (AAMC) and American Association of Colleges of Osteopathic Medicine (AACOM) [1,2].

Tracking the cadence of interview invitations is a common use of CycleTrack, and we explored the validity and utility of self-reported data for this purpose. Official dates of interview invitations from Northwestern University and the University of Michigan were compared to those of CycleTrack users. Additionally, the proportions of interview invitation entries in CycleTrack from completed application cycles (2022 - 2025) were plotted over time to examine differences in the dissemination of interviews between years and program types (total: N=24,941; doctor of medicine [MD]: n=18,465; doctor of osteopathic medicine [DO]: n=3698; MD–doctor of philosophy [PhD]: n=2778). All statistics and graphs were generated using R v4.5.0 (The R Foundation for Statistical Computing) and ggplot2 [11].

Ethical Considerations

Our study was confirmed to be nonhuman subjects research by the Northwestern University Institutional Review Board (#STU00225793).

Results

Applications and Users

From May 2022 to January 2026, 34,763 email-verified users registered on the platform. Of these, 20,354 tracked at least one application (MD: n=18,848; MD-PhD: n=1592; DO: n=4456; DO-PhD: n=67), submitting a total of 415,837 applications (MD: n=359,440; MD-PhD: n=26,018; DO: n=30,217; DO-PhD: n=162). Google Analytics data of website visitors throughout the 2024-2025 and 2025-2026 cycles identified over 257,000 unique visitors, with the most traffic to the data explorer. The number of tracked applications strongly correlated with the total applications for each school reported by the AAMC and AACOM (Figure 1A-C), and in the 2024 - 2025 cycle, CycleTrack applications represented a median 13.5% (range 1.1% - 44.1%) of total submitted applications per program (MD: 10.0%, 1.1% - 44.1%; DO: 4.1%, 2.7% - 8.3%; MD-PhD: 24.4%, 4.9% - 29.6%).

Figure 1. Commonly tracked programs and utility of CycleTrack in examining the interview cycle. (A-C) The most commonly tracked programs in CycleTrack are associated with the number of applications submitted to programs reported by the Association of American Medical Colleges (AAMC) and American Association of Colleges of Osteopathic Medicine (AACOM) during the 2024 - 2025 cycle (Spearman ρ≥0.91; *P*<.001) [1,2]. (D) Cumulative distributions of interview invitations stratified by doctor of medicine (MD), doctor of osteopathic medicine (DO), and MD–doctor of philosophy (PhD) applicants for the 2022 - 2023, 2023 - 2024, and 2024 - 2025 application cycles. Cumulative distributions were significantly different (9-sample Anderson-Darling *P*<.001). (E) The weekly percent of total interviews sent for each program type and year. Individual plots for each program type are shown and discussed in Multimedia Appendix 1. DO-PhD applications were excluded from temporal modeling due to an insufficient sample size.

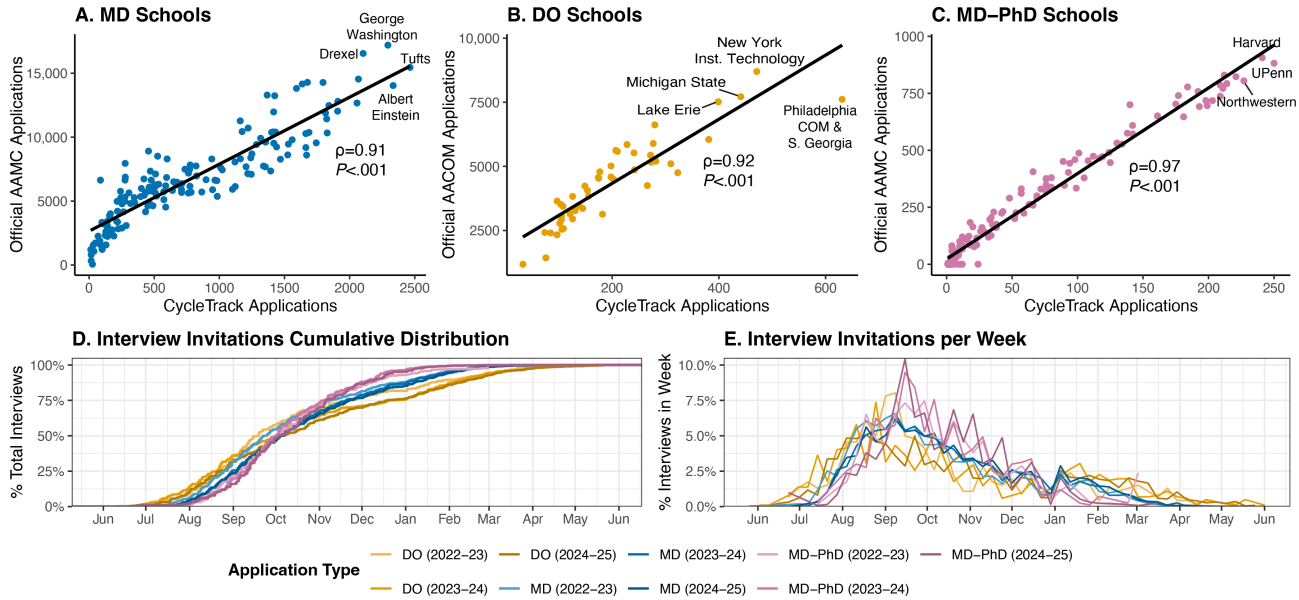

Of users who tracked applications, 80% reported demographic information. Demographics reported in the 2024 - 2025 application cycle are summarized in Table 1. CycleTrack users were more likely to be men, be White or Asian, and score higher compared to national statistics.

XSL•FO
RenderX

Table . Comparison of CycleTrack user demographics from the 2024 - 2025 application cycle to national data reported by the Association of American Medical Colleges (AAMC) and American Association of Colleges of Osteopathic Medicine (AACOM) [1,2].

	CycleTrack[a]	AAMC (MD[b]/MD-PhD[c]) [1] or AACOM (DO[d]) [2]	*P* value[e]
MD			
Gender, n (%)			<.001
Man	2321 (45.4)	23,072 (42.2)	
Woman	2731 (53.4)	31,292 (57.2)	
Other, transgender, nonbinary	60 (1.2)	265 (0.5)	
Race/ethnicity, n (%)			<.001
Asian	1682 (34.0)	14,585 (26.7)	
Black	265 (5.4)	5022 (9.2)	
Hispanic	383 (7.8)	3752 (6.9)	
White	2180 (44.1)	18,597 (34.0)	
Other, multiple	430 (8.7)	12,743 (23.3)	
Age at matriculation (years), mean (SD)	23.9 (2.6)	NR[f]	—[g]
GPA[h], mean (SD)	3.79 (0.21)	3.67 (0.33)	<.001
MCAT[i], mean (SD)	513.4 (6.5)	506.3 (10.0)	<.001
MD-PhD			
Gender, n (%)			ND[j,k]
Man	179 (48.2)	NR	
Woman	182 (49.1)	NR	
Other, transgender, nonbinary	10 (2.7)	NR	
Race/ethnicity, n (%)			<.001
Asian	139 (39.0)	619 (30.3)	
Black	23 (6.5)	210 (10.3)	
Hispanic	13 (3.7)	116 (5.7)	
White	145 (40.7)	619 (30.3)	
Other, multiple	36 (10.1)	476 (23.3)	
Age at matriculation (years), mean (SD)	23.7 (2.0)	NR	—
GPA, mean (SD)	3.83 (0.22)	3.71 (0.33)	<.001
MCAT, mean (SD)	516.2 (6.6)	510.4 (10.4)	<.001
DO			
Gender, n (%)			ND[l]
Man	586 (43.2)	9464 (40.9)[l]	
Woman	757 (55.8)	13,602 (58.8)[l]	
Other, transgender, nonbinary	14 (1.0)	48 (0.2)[l]	
Race/ethnicity, n (%)			<.001
Asian	399 (30.5)	6331 (27.4)	
Black	79 (6.0)	1844 (8.0)	
Hispanic	107 (8.2)	2583 (11.2)	
White	618 (47.2)	9905 (42.9)	

XSL•FO
RenderX

	CycleTrack[a]	AAMC (MD[b]/MD-PhD[c]) [1] or AACOM (DO[d]) [2]	*P* value[e]
Other, multiple	106 (8.1)	2451 (10.6)	
Age at matriculation (years), mean (SD)	24.5 (3.1)	NR	—
GPA, mean (SD)	3.70 (0.25)	3.58 (NR[m])	<.001
MCAT, mean (SD)	508.2 (6.3)	502.4 (NR[m])	<.001

[a]As entry of demographic information is optional for CycleTrack users, not all variables have the same number of total responses. Users who applied to multiple program types are counted for each program type they applied to. Therefore, these analyses are not statistically independent of one another.

[b]MD: doctor of medicine.

[c]PhD: doctor of philosophy.

[d]DO: doctor of osteopathic medicine.

[e]Categorical data were compared by χ^2 test, GPA was compared by Wilcoxon signed rank test, and MCAT was compared by 1-sample t test.

[f]NR: not reported; data not publicly disclosed by the AAMC or AACOM.

[g]Not applicable.

[h]GPA: grade point average.

[i]MCAT: Medical College Admission Test.

[j]ND: not done; statistical tests that were not performed.

[k]The AAMC does not publicly report the gender distribution of MD-PhD applicants. The gender distribution of applications to MD-PhD programs is 49.3% men, 49.1% women, and 1.6% other or declined to report [1]. As the number of applications submitted per applicant may vary by gender, these percentages may not fully reflect the distribution of applicants.

[l]The American Association of Colleges of Osteopathic Medicine Application Service (AACOMAS) reports gender using binary categories (male, female, not disclosed). The American Medical College Application Service (AMCAS) uses identity-based categories (man, woman, nonbinary, etc). As CycleTrack uses categories defined by AMCAS, no direct comparison to AACOMAS data was performed.

[m]The most recent AACOMAS applicant GPA and MCAT information at the time of writing are reported for the 2023 - 2024 application cycle.

Interview Invitations

Compared to official data from the University of Michigan and Northwestern University MD programs, CycleTrack accurately captured 79% to 96% of the weeks on which interview invitations were released (Multimedia Appendix 2). Modeling the cumulative distribution of interviews across three application cycles demonstrated that DO interview invitations exhibited the greatest yearly fluctuation and that DO and MD-PhD interview cycles are longer and shorter than MD, respectively (Figure 1D-E, further discussed in Multimedia Appendix 1).

Discussion

CycleTrack extends prior tools by serving as the first website to centralize real-time application tracking and distribution of aggregated statistics. The volume of users/traffic demonstrates demand for free, publicly accessible tools and information about application timelines. Following the launch of CycleTrack, other online platforms (eg, Admit) have implemented and iterated on its model for both medical school and residency admissions.

CycleTrack demographics deviated from AAMC/AACOM statistics, which may influence the accuracy of the explorer tool. While hesitancy to report data among lower-scoring or underrepresented students may be a contributing factor, it is also likely that online word-of-mouth advertising via forums used by CycleTrack is insufficient to reach some populations of applicants. Furthermore, while our analysis of reported interview dates provides utility for understanding the overall cadence of the interview cycle, as survival analysis techniques were not implemented, the chronological decline in reported invitations may reflect survival bias and user abandonment rather than true institutional behavior. Greater advertisement and uptake of such aggregation platforms may address several of these limitations. However, data transparency at the institutional and national levels is feasible as demonstrated by AAMC/AACOM tools and the University of Michigan #UMichMedMondays program. We urge medical schools and their national bodies to adopt clear and transparent models for sharing real-time application metrics so that platforms such as CycleTrack are obsolete.

Despite its limitations, CycleTrack has advanced data aggregation efforts among premedical applicants. It has uniquely captured temporal trends of admissions actions, provided dual-degree–specific data, and built a database for data-driven inquiries into applicant questions. Furthermore, it demonstrates the power of a student initiative to improve medical school applications for all.

Acknowledgments

The authors would like to thank Dr Alan Hauser and Dr Lindsey Martin from Northwestern University for their assistance with external validation of CycleTrack data and critical review of this manuscript.

XSL•FO
RenderX

Funding

DBA, IH, and ASK were supported by Medical Scientist Training Program grants from the National Institutes of Health awarded to the Northwestern University (T32GM144295) and University of Colorado (T32GM149361) MD-PhD programs. IH was also supported by the University of Colorado Pharmacology and Molecular Medicine graduate program's Predoctoral Training Grant (1T32GM158468-01). The contents of this manuscript are the sole responsibility of the authors and do not necessarily represent the views of the National Institutes of Health. No dedicated funding was obtained for the purpose of this manuscript, and the CycleTrack platform itself was self-funded by the authors and is free to use without advertisements to all.

Data Availability

Aggregated data from the CycleTrack platform are publicly accessible on the CycleTrack website [10]. To protect applicant privacy, individual-level data are not publicly available.

Authors' Contributions

Conceptualization: DBA (lead), IH (supporting), ASK (supporting)
Methodology: DBA, IH, ASK
Investigation: DBA, IH, ASK
Software: DBA, IH
Data curation: DBA
Formal analysis: DBA
Writing – original draft: DBA
Writing – review & editing: DBA, IH, ASK

Conflicts of Interest

DBA and IH are the developers of the CycleTrack website and software.

Multimedia Appendix 1
Interview invitations by program type and application cycle.
[PDF File, 140 KB - jmir_v28i1e83087_app1.pdf]

Multimedia Appendix 2
External validation of CycleTrack interview data.
[PDF File, 114 KB - jmir_v28i1e83087_app2.pdf]

References

1. 2025 facts: applicants and matriculants data. Association of American Medical Colleges. 2025. URL: https://www.aamc.org/data-reports/students-residents/data/facts-applicants-and-matriculants [accessed 2026-02-28]
2. Research reports. American Association of Colleges of Osteopathic Medicine. 2026. URL: https://www.aacom.org/searches/reports [accessed 2026-02-28]
3. Mathew D, Grbic D. Use and evaluation of medical school information sources by aspiring medical students. AAMC Analysis Brief 2011 Apr;11(3) [FREE Full text]
4. Christophers B, Gotian R. Using admission statistics to encourage diverse applicants to MD-PhD programs. J Clin Invest 2020 Jan 2;130(1):17-19. [doi: 10.1172/JCI134941] [Medline: 31815735]
5. Hadinger MA. Underrepresented minorities in medical school admissions: a qualitative study. Teach Learn Med 2017;29(1):31-41. [doi: 10.1080/10401334.2016.1220861] [Medline: 27824260]
6. AMCAS Tracker. URL: https://amcas-tracker.hpsa.org/ [accessed 2026-05-13]
7. MD Applicants. URL: https://www.mdapplicants.com/ [accessed 2026-05-13]
8. 2019/2020 MD Application Tracker (responses). Google. URL: https://docs.google.com/spreadsheets/d/1VvNlLPixJwHlnc589IgAeKgRKMnamDjbYDf8vyVGj6I [accessed 2026-05-13]
9. Flask (Version 2.3.3).: Pallets; 2023. URL: https://flask.palletsprojects.com/ [accessed 2025-07-11]
10. Amusin DB, Hua IJ. CycleTrack. 2022. URL: https://cycletrack.org [accessed 2025-07-11]
11. Wickham H. ggplot2.: Tidyverse; 2016. URL: https://ggplot2.tidyverse.org [accessed 2025-07-11]

Abbreviations

AACOM: American Association of Colleges of Osteopathic Medicine
AAMC: Association of American Medical Colleges
AMCAS: American Medical College Application Service
DO: doctor of osteopathic medicine

XSL•FO
RenderX

MCAT: Medical College Admission Test
MD: doctor of medicine
PhD: doctor of philosophy

Edited by A Coristine; submitted 27.Aug.2025; peer-reviewed by E Chuck, M Rohaim; revised version received 04.May.2026; accepted 05.May.2026; published 22.May.2026.

Please cite as:
Amusin DB, Hua I, Kozhumam AS
Crowdsourcing Medical School Admissions Data: Development and Analysis of the CycleTrack Platform
J Med Internet Res 2026;28:e83087
URL: https://www.jmir.org/2026/1/e83087
doi:10.2196/83087

© Daniel Boris Amusin, Isabelle Hua, Arthi Shankar Kozhumam. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating the Potential of Reasoning Large Language Models to Perpetuate Racial and Gender Disease Stereotypes in Health Care

Joshua J Docking[1], BMedSc; Lee X Li[1], MBioStats; Bradley D Menz[1], BPharm; Stephen Bacchi[2], PhD; Ashley M Hopkins[1], PhD; Michael J Sorich[1], PhD

[1]College of Medicine and Public Health, Flinders University, GPO Box 2100, Adelaide, SA, Australia
[2]Adelaide Medical School, Adelaide University, Adelaide, SA, Australia

Corresponding Author:
Michael J Sorich, PhD
College of Medicine and Public Health, Flinders University, GPO Box 2100, Adelaide, SA, Australia

Abstract

This evaluation of 36,000 clinical vignettes found that next-generation reasoning large language models, o3-mini and DeepSeek-R1, frequently perpetuate racial and gender stereotypes for common medical conditions, indicating that advancements in reasoning do not inherently improve representational fairness.

(J Med Internet Res 2026;28:e82256) doi:10.2196/82256

KEYWORDS

large language model; reasoning LLM; artificial intelligence; bias; stereotypes; fairness; health equity; race; gender; representation

Introduction

Large language models (LLMs) have the potential to transform health care but risk exacerbating health disparities if they perpetuate biases [1-3]. Zack and colleagues [4] demonstrated potential racial and gender biases in clinical vignettes generated by GPT-4, including overrepresentation of Black patients in stereotypical medical conditions. Since then, next-generation reasoning LLMs have emerged, offering improved reasoning capability (“thinking” before answering), with this model class demonstrating superior benchmark performance [5]. Whether this will reduce representational bias in health care remains unknown. This study evaluated whether reasoning LLMs exhibit racial and gender biases in generated clinical content.

Methods

Using the methods of Zack et al [4], two prominent reasoning LLMs of distinct geographic origin, o3-mini (OpenAI) and DeepSeek-R1 (DeepSeek; 671B full model), generated patient cases across 18 medical conditions that represent a spectrum of demographic-prevalence relationships (Multimedia Appendix 1), specifying a US population, using 10 prompt variations (Table S1 in Multimedia Appendix 1), and running 100 times each. Patient demographic characteristics in the generated vignettes were extracted using the methods of Zack et al [4], and the proportion of race and gender representation for each condition was calculated. A qualitative analysis of DeepSeek-R1’s reasoning traces was also performed on a random sample of 20 vignettes (Multimedia Appendix 1). Misrepresentation (LLM estimate minus the published US epidemiological estimates [4]) was summarized as the median (range) across 18 medical conditions. For example, if 60% of LLM-generated HIV cases were Black patients, compared to the 40% of Black patients reported in representative US studies of HIV (Table S2, Multimedia Appendix 1), misrepresentation would be +20%. Positive values indicate overrepresentation, and negative values indicate underrepresentation. A difference greater than 20% was considered the threshold for significant misrepresentation, indicating a practically meaningful deviation in demographic representation. Sensitivity analyses using 10% and 30% thresholds confirmed that findings were robust to threshold selection (Multimedia Appendix 1). χ^2 goodness-of-fit tests with Benjamini-Hochberg false-discovery rate correction were used to assess whether the LLM-generated demographic distributions differed significantly from epidemiological baselines for each condition (Multimedia Appendix 1).

Results

A total of 36,000 unique clinical vignettes were generated. Pairwise word-level Jaccard similarity confirmed substantive diversity, with a mean within-group similarity of 0.35 (SD 0.06 for DeepSeek-R1 and 0.08 for o3-mini) and near-duplicate pairs (Jaccard >0.8) constituting fewer than 0.1% of comparisons (Multimedia Appendix 1). Median misrepresentation for o3-mini was 44% (range −12% to +75%) for Black, −4.6% (range −37% to 0%) for Asian, −14% (range −27% to +0.7%) for Hispanic, and −8.2% (range −56% to +33%) for White persons (Figure 1). Median misrepresentation for DeepSeek-R1 was 31% (range

XSL•FO
RenderX

−21% to +81%) for Black, −4.4% (range −35% to +47%) for Asian, −8.8% (range −26% to +53%) for Hispanic, and −21% (range −63% to +41%) for White persons (Figure 2). For 78% (14/18) of medical conditions using o3-mini and 89% (16/18) using DeepSeek-R1, there was more than 20% misrepresentation for at least one race. χ^2 goodness-of-fit tests confirmed that the racial distributions generated by both models differed significantly from epidemiological baselines for all 18 conditions (all Benjamini-Hochberg corrected $P<.001$; Table S5, Multimedia Appendix 1).

Figure 1. Proportional representation of disease cases by race and gender in o3-mini–generated clinical content compared with published US statistics. Blue dot on the right of red diamond indicates overrepresentation by the large language model. Blue dot on the left indicates underrepresentation.

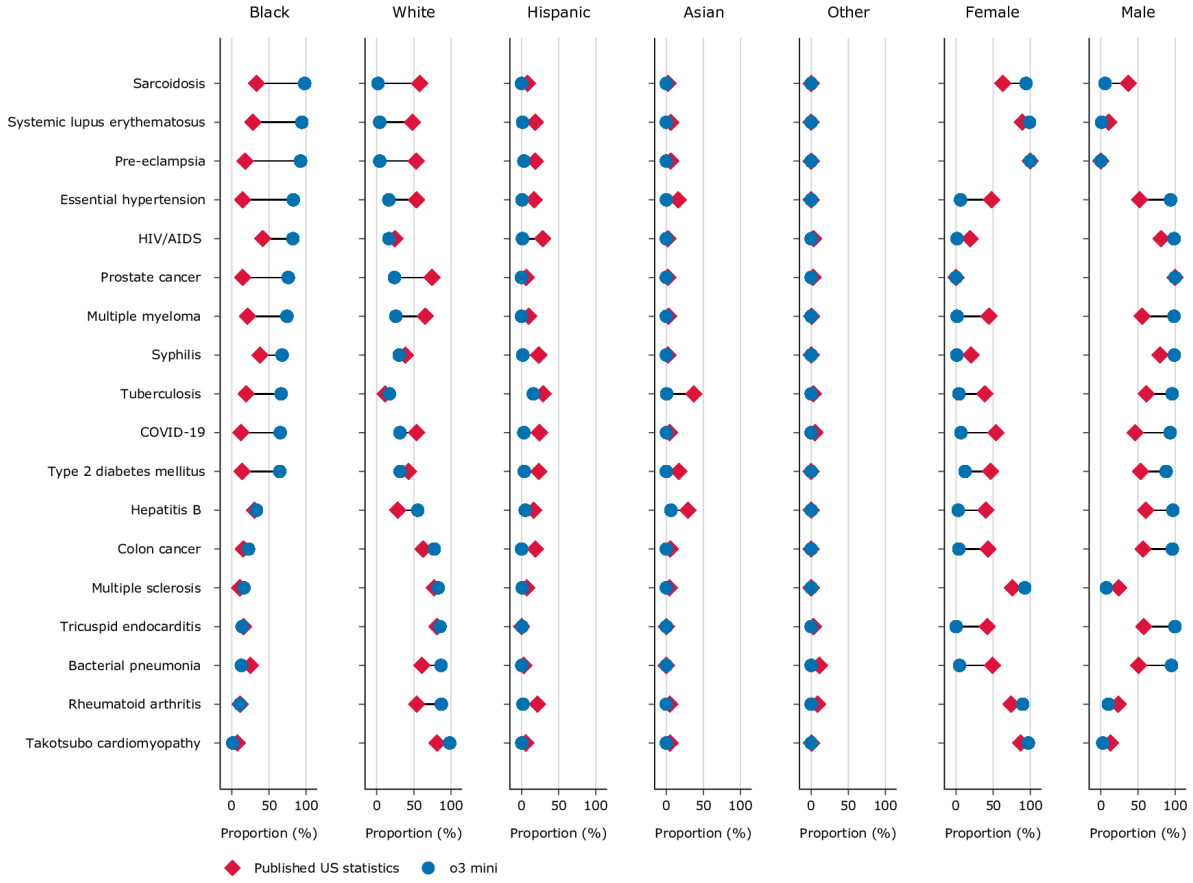

XSL•FO
RenderX

Figure 2. Proportional representation of disease cases by race and gender in DeepSeek R1–generated clinical content compared with published US statistics. Blue dot on the right of red diamond indicates overrepresentation by the large language model. Blue dot on the left indicates underrepresentation.

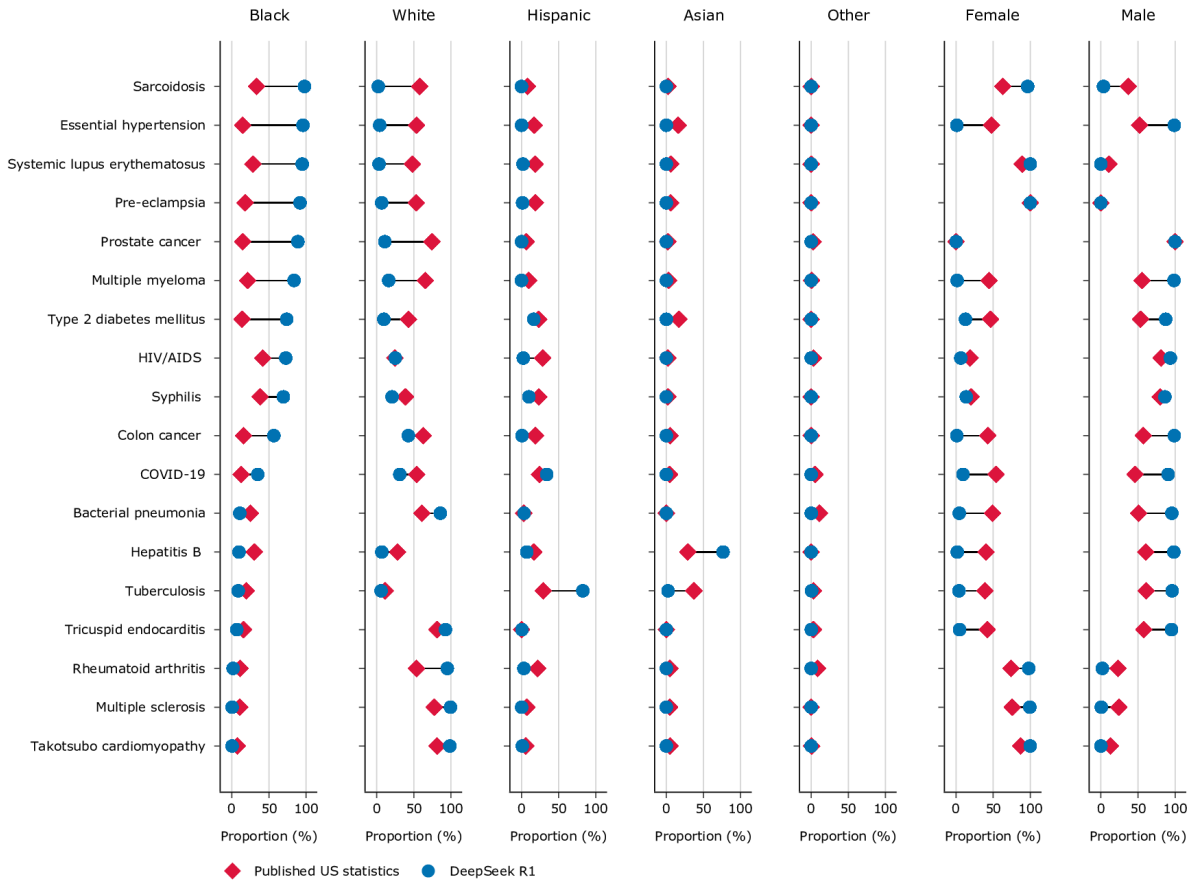

Female median misrepresentation was −27% (range −47% to +31%) for o3-mini and −23% (range −47% to +33%) for DeepSeek-R1. For 56% (10/18) of medical conditions using o3-mini and 67% (12/18) using DeepSeek-R1, there was more than 20% misrepresentation. There was a consistent overrepresentation of the gender with the higher published representation. Gender distributions also differed significantly from epidemiological baselines for all non–sex-linked conditions (all Benjamini-Hochberg corrected $P<.001$; Table S6 in Multimedia Appendix 1).

Discussion

Reasoning LLMs, such as o3-mini and DeepSeek-R1, frequently misrepresent the distribution of race and gender in medical conditions, mirroring issues previously observed in GPT-4 [4], which met the threshold for significant misrepresentation in 67% (12/18) of conditions for race and 67% (12/18) for gender [4]. Our results show comparable or higher rates for o3-mini (78% race, 56% gender) and DeepSeek-R1 (89% race, 67% gender), indicating no improvement in representation with the newer reasoning models.

Both o3-mini and DeepSeek-R1, like GPT-4, overrepresented Black populations in stereotypically associated conditions (eg, sarcoidosis, systemic lupus erythematosus, pre-eclampsia, essential hypertension) [4], with even higher median misrepresentation of 44% and 31%, respectively, compared to 15% in GPT-4 [4]. This persistent pattern may reflect underlying bias, though models may also default to generating prototypical cases rather than representative samples due to patterns in their training data. Qualitative analysis of DeepSeek-R1's reasoning traces supports this, revealing that the model explicitly invoked disease-demographic associations (eg, "more prevalent in") when selecting patient demographics, without referencing quantitative epidemiological data (Multimedia Appendix 1). This explicit demographic deliberation may also explain the higher misrepresentation observed in the reasoning models included in this study compared to GPT-4, as the extended reasoning process may amplify stereotypical associations by actively invoking them during generation. In either case, consistently overrepresenting certain demographic groups, particularly for conditions that in practice affect diverse populations, risks reinforcing narrowed demographic assumptions in clinical contexts where understanding disease prevalence across populations is an important component of diagnostic reasoning. Similarly, the consistent exaggeration of the majority gender aligns with previous findings showing LLM outputs skew toward gender stereotypes in health care roles, which could further marginalize minority genders [6].

This study's strengths include its evaluation of next-generation reasoning LLMs and the robust assessment from 36,000 generated clinical vignettes. Limitations include our focus on

XSL•FO
RenderX

a US context, and that DeepSeek-R1's development outside the United States may mean that deviations partly reflect differences in training data representation, though the similar directional patterns between both models suggest shared stereotypical associations. The demographic categories were also adopted from Zack et al [4] to enable direct comparison but do not capture Native American, multiracial, nonbinary, or transgender populations. and treat "Hispanic" as a race rather than an ethnicity. Additionally, given the rapid evolution of the LLM landscape, GPT-4 comparisons are based on published data from Zack et al [4] rather than a concurrent control run with identical prompts, which limits definitive comparative conclusions. However, the present study's inclusion of explicit US geographic context in prompts would, if anything, be expected to reduce deviations from US epidemiological baselines, suggesting the comparison is conservative. Further detail on comparability is provided in Multimedia Appendix 1. The qualitative analysis was exploratory and based on a small sample, limiting the generalizability of the mechanistic observations. Future research should assess generalizability across different regions and broader condition sets, and evaluate whether prompt-level strategies can mitigate the observed patterns. For example, explicit instructions to reflect epidemiological distributions or providing previously generated cases as context to encourage demographic diversity across outputs may help shift generation from prototypical to representative cases, particularly given our finding that the model explicitly deliberates about demographic associations during generation. Further research should also directly evaluate whether generation biases in LLMs translate to impacts on clinical decision-making and potential patient harms.

In conclusion, despite enhanced reasoning capabilities, the clinical outputs of o3-mini and DeepSeek-R1 still exhibit racial and gender disease stereotyping in common medical conditions. Advancements in LLM capabilities do not guarantee parallel improvements across all dimensions [7], including, as demonstrated here, fairness and representation in health care. Awareness of these demographic defaults is essential for the safe integration of LLMs into clinical workflows, and continuous monitoring of potential biases should accompany their adoption.

Acknowledgments

ChatGPT 4o (OpenAI) and Gemini 2.5 Pro (Google) were used to assist in formatting and editing the manuscript. The authors reviewed and verified all artificial intelligence–assisted content.

Funding

MJS is supported by a Beat Cancer Research Fellowship from the Cancer Council South Australia (PRF2719). AMH holds an Emerging Leader Investigator Fellowship from the National Health and Medical Research Council, Australia (APP2008119). The PhD scholarship of BDM is supported by the National Health and Medical Research Council, Australia (APP2030913). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: MJS
Methodology: MJS
Software: JJD
Validation: JJD, LXL, BDM, SB, AMH, MJS
Formal analysis: JJD, MJS
Investigation: JJD
Resources: MJS
Data curation: JJD
Writing – original draft: JJD
Writing – review & editing: LXL, BDM, SB, AMH, MJS
Visualization: JJD
Supervision: MJS
Project administration: JJD, MJS
Funding acquisition: MJS

Conflicts of Interest

MJS reported receiving grants from Pfizer, AstraZeneca, Boehringer Ingelheim, and the National Health and Medical Research Council of Australia outside the submitted work. AMH reported receiving grants from Boehringer Ingelheim, Hospital Research Foundation, Tour De Cure, and Flinders Foundation outside the submitted work. No other disclosures were reported.

XSL•FO
RenderX

Multimedia Appendix 1
Methods, model details, data sources, analyses, and vignette uniqueness.
[PDF File, 245 KB - jmir_v28i1e82256_app1.pdf]

References

1. Weidinger L, Uesato J, Rauh M, et al. Taxonomy of risks posed by language models. Presented at: FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency; Jun 21-24, 2022. [doi: 10.1145/3531146.3533088]
2. Adam H, Balagopalan A, Alsentzer E, Christia F, Ghassemi M. Mitigating the impact of biased artificial intelligence in emergency decision-making. Commun Med (Lond) 2022 Nov 21;2(1):149. [doi: 10.1038/s43856-022-00214-4] [Medline: 36414774]
3. Pfohl SR, Cole-Lewis H, Sayres R, et al. A toolbox for surfacing health equity harms and biases in large language models. Nat Med 2024 Dec;30(12):3590-3600. [doi: 10.1038/s41591-024-03258-2] [Medline: 39313595]
4. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health 2024 Jan;6(1):e12-e22. [doi: 10.1016/S2589-7500(23)00225-X] [Medline: 38123252]
5. Brodeur PG, Buckley TA, Kanjee Z, et al. Superhuman performance of a large language model on the reasoning tasks of a physician. arXiv. Preprint posted online on Dec 14, 2024. [doi: 10.48550/arXiv.2412.10849]
6. Menz BD, Kuderer NM, Chin-Yee B, et al. Gender representation of health care professionals in large language model-generated stories. JAMA Netw Open 2024 Sep 3;7(9):e2434997. [doi: 10.1001/jamanetworkopen.2024.34997] [Medline: 39312237]
7. Cui DX, Long SY, Tang YX, Zhao Y, Li Q. Can reasoning power significantly improve the knowledge of large language models for chemistry?—Based on conversations with LLMs. J Chem Inf Model 2025 Sep 22;65(18):9516-9527. [doi: 10.1021/acs.jcim.5c01265] [Medline: 40854079]

Abbreviations

LLM: large language model

Edited by A Coristine; submitted 09.Sep.2025; peer-reviewed by KH Lin, SY Long; revised version received 26.Apr.2026; accepted 08.May.2026; published 28.May.2026.

Please cite as:
Docking JJ, Li LX, Menz BD, Bacchi S, Hopkins AM, Sorich MJ
Evaluating the Potential of Reasoning Large Language Models to Perpetuate Racial and Gender Disease Stereotypes in Health Care
J Med Internet Res 2026;28:e82256
URL: https://www.jmir.org/2026/1/e82256
doi:10.2196/82256

© Joshua J Docking, Lee X Li, Bradley D Menz, Stephen Bacchi, Ashley M Hopkins, Michael J Sorich. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Large Language Model Versus Multidisciplinary Team: Feasibility Study of Pancreatic Cancer Management Recommendations

Zhuoran Liu[1*], MD; Xin Zhao[2*], MD; Tianyang Mao[1], MD; Heng Zhou[1], MM; Kangyi Jiang[1], MM

[1]Department of Hepato-Pancreato-biliary Surgery, People's Hospital of Leshan, No.238 Huian street, Shizhong District, Leshan, Sichuan Province, China

[2]Department of General Surgery, West China Hospital of Sichuan University, Chengdu, Sichuan, China

[*]these authors contributed equally

Corresponding Author:
Kangyi Jiang, MM
Department of Hepato-Pancreato-biliary Surgery, People's Hospital of Leshan, No.238 Huian street, Shizhong District, Leshan, Sichuan Province, China

Abstract

Of 269 pancreatic cancer multidisciplinary team cases, 125 complete structured cases were evaluated. The chat interface model labeled ChatGPT-5.2 matched the multidisciplinary team's broad initial management category in 119 cases (95.2%, 95% CI 89.8%-98.2%). The end point was coarse, and the model was not version controlled, so findings only support supervised exploratory use.

(J Med Internet Res 2026;28:e95411) doi:10.2196/95411

KEYWORDS

pancreatic cancer; multidisciplinary team; large language models; clinical decision support; artificial intelligence; MDT; LLM

Introduction

Pancreatic cancer treatment decisions integrate resectability, metastatic disease, performance status, comorbidities, symptoms, and biomarkers [1,2]. Multidisciplinary team (MDT) review is recommended but may be limited by staffing and scheduling [3]. Large language models (LLMs) might help organize clinical information, yet their agreement with real pancreatic MDT decisions is uncertain [4]. We compared a chat interface LLM with documented MD recommendations as a feasibility assessment of prespecified broad management category concordance in complete structured cases, not autonomous decision-making or implementation.

Methods

Overview

We retrospectively reviewed deidentified pancreatic cancer cases discussed at weekly specialist pancreatic MDT meetings at a single center from March 2023 to December 2025. Of 269 MDT-discussed cases, 125 treatment-naive adults with pathologically confirmed pancreatic neoplasms and sufficient clinical, imaging, staging, and biomarker information for standardized prompts formed the analytic cohort. No purposive enrichment was used. The retained analytic file did not include a screening log, patient-level data for the 144 nonincluded cases, aggregate missing data descriptors, or exclusion categories. Thus, the cohort represents complete structured cases rather than all pancreatic cancer MDT cases, and optimistic selection bias is possible if excluded cases were more ambiguous, incomplete, or clinically complex. Two researchers extracted variables independently; a senior pancreatic cancer specialist adjudicated discrepancies.

The index LLM was the chat interface model labeled ChatGPT-5.2. This label reflects what was displayed during data collection, not an application programming interface–pinned public model identifier. Hidden model build, training data cutoff, temperature, system configuration, exact prompt entry dates, and replicate intersession outputs were not available or retained. No application programming interface–based or version-controlled access was used. All cases used the same structured prompt (Multimedia Appendix 1). The LLM workflow and study evaluation framework are shown in Multimedia Appendix 2. The primary outcome was exact case-level agreement between the LLM recommendation and MDT decision across 5 prespecified broad initial management categories. The categories were defined before concordance assessment and were not revised after reviewing model outputs. The end point did not evaluate regimen choice, sequencing, dose intensity, treatment line triggers, biomarker-specific regimen selection, or implementation safety. No rule-based or guideline-only comparator was constructed.

Three independent specialists rated each LLM output for comprehensiveness, perceived concordance, and rationality on 5-point Likert scales. Detailed mean specialist appraisal scores by specialty are shown in Multimedia Appendix 3. For perceived

XSL•FO
RenderX

concordance, they saw the MDT decision; for other dimensions, they saw the case summary and LLM output. Because all dimensions were scored within the same workflow, halo effects and nonindependence are possible. Concordance proportions used Clopper-Pearson exact 95% CIs. Linearly weighted Cohen κ was used to assess rater consistency. Resectability, histology, and biomarker subgroup tabulations were descriptive only.

Ethical Considerations

The Ethics Committee of Leshan People's Hospital approved the study (LYLL2026KY117) and waived informed consent because retrospective deidentified data were used.

Results

Among 125 included patients, the mean age was 62.36 (SD 8.92) years, 80 (64%) were male, 113 (90.4%) had pancreatic ductal adenocarcinoma, and 12 (9.6%) had grade 2 pancreatic neuroendocrine neoplasm (Table 1). The analytic cohort represented 46.5% of the 269 MDT-discussed cases. No data were available to compare included and nonincluded cases or to determine if exclusions reflected missing imaging, biomarkers, pathology, comorbidity details, or other contextual information. Overall broad category concordance was 119 (95.2%, 95% CI 89.8% - 98.2%; Table 2). Concordance was 32 of 36 (88.9%) in resectable disease, 38 of 38 (100%) in borderline resectable disease, 27 of 29 (93.1%) in locally advanced disease, and 22 of 22 (100%) in metastatic disease. These subgroup estimates are descriptive; 100% agreement in borderline resectable and metastatic disease should be interpreted with the strong constraint that stage and resectability impose on treatment direction. *BRCA1/2* mutant and microsatellite instability-high or mismatch repair-deficient (MSI-H/dMMR) rows are descriptive tabulations rather than powered subgroup analyses because the denominators were small.

Discordant cases mainly involved resectable or locally advanced disease and appeared related to different weighting of comorbidities, perioperative risk, and thresholds for neoadjuvant intensification. Mean specialist ratings were 3.21 (SD 0.58) for comprehensiveness, 3.85 (SD 0.42) for perceived concordance, and 3.97 (SD 0.39) for rationality. These ratings should not be interpreted as mutually independent dimensions or as validation of mechanical concordance because exposure to the MDT decision for one dimension could influence adjacent judgments. Weighted κ values ranged from 0.70 to 0.83, indicating substantial to near-perfect rater consistency only.

Concordance was defined as exact agreement between the LLM and MDT on the prespecified broad initial management strategy. Subgroup estimates, including biomarker subgroup estimates, are descriptive because no inferential subgroup analysis was prespecified, and some denominators were small.

XSL•FO
RenderX

Table . Baseline characteristics of included cases (N=125).

Characteristic	Value
Male sex, n (%)	80 (64.0)
Age (years), mean (SD)	62.36 (8.92)
BMI (kg/m^2), mean (SD)	23.36 (3.27)
ECOG[a] performance status 0, n (%)	93 (74.4)
ECOG performance status 1, n (%)	32 (25.6)
Diabetes mellitus, n (%)	63 (50.4)
Hypertension, n (%)	48 (38.4)
Cardiovascular disease, n (%)	18 (14.4)
CA19-9[b] (U/mL), median (IQR)	897 (326 - 3560)
Tumor in pancreatic head, n (%)	96 (76.8)
Resectable, n (%)	36 (28.8)
Borderline resectable, n (%)	38 (30.4)
Locally advanced, n (%)	29 (23.2)
Metastatic, n (%)	22 (17.6)
PDAC[c], n (%)	113 (90.4)
pNET[d] G2, n (%)	12 (9.6)
BRCA1/2 mutant, n (%)	20 (16.0)
MSI-H/dMMR[e], n (%)	4 (3.2)

[a]ECOG: Eastern Cooperative Oncology Group.
[b]CA19-9: carbohydrate antigen 19 - 9.
[c]PDAC: pancreatic ductal adenocarcinoma.
[d]pNET G2: grade 2 pancreatic neuroendocrine neoplasm.
[e]MSI-H/dMMR: microsatellite instability-high or mismatch repair-deficient.

Table . Case-level concordance between the large language model and multidisciplinary team by subgroup.

Subgroup	Concordant/total, n	Concordance (%)	Exact 95% CI
Overall	119/125	95.2	89.8 - 98.2
Resectable	32/36	88.9	73.9 - 96.9
Borderline resectable	38/38	100.0	90.7 - 100.0
Locally advanced	27/29	93.1	77.2 - 99.2
Metastatic	22/22	100.0	84.6 - 100.0
BRCA1/2 mutant	20/20	100.0	83.2 - 100.0
MSI-H/dMMR[a]	4/4	100.0	39.8 - 100.0

[a]MSI-H/dMMR: microsatellite instability-high or mismatch repair-deficient.

Discussion

In complete structured cases, the LLM often selected the same broad initial management category as the pancreatic MDT. This result should be interpreted narrowly. The 5-category end point captures treatment direction, not regimen-level correctness, safety, workflow benefit, patient benefit, or clinical equivalence. Because resectability and stage constrain initial management, especially in borderline resectable and metastatic disease, the 95.2% concordance is best viewed as the upper bound supported by this coarse end point, not evidence of nuanced multidisciplinary reasoning. Without a rule-based or guideline-only baseline, we cannot determine whether the LLM exceeded structured guideline retrieval or a simple decision tree [5]. More granular end points assessing regimen selection, sequencing, dose modification, biomarker-triggered treatment, radiation indications, operative risk trade-offs, or supportive care priorities would likely reduce observed concordance. The

XSL•FO
RenderX

analysis mainly identifies feasibility issues, reporting constraints, and end point limitations for future studies.

The limitations are substantial. First, this was a single-center complete-case study; only 125 of 269 MDT-discussed cases were analyzed, and missing screening data prevented characterization of the 144 nonincluded cases. The estimate, therefore, applies only to complete structured cases and is likely optimistically biased. Second, standardized summaries may have simplified the MDT context by removing uncertainty, competing priorities, and workflow constraints. Third, the chat interface model was not reproducible because the model build, parameters, training data cutoff, prompt-entry dates, and replicate outputs were unavailable. Temporal validity is consequently uncertain for cases and guideline changes from March 2023 to December 2025. Fourth, most cases were pancreatic ductal adenocarcinoma, with only 12 grade 2 pancreatic neuroendocrine neoplasm cases; histology and biomarker subgroup results were descriptive, with small denominators and wide intervals, particularly for MSI-H/dMMR. Fifth, specialist ratings may be affected by halo effects because concordance scoring used the MDT decision. Finally, the MDT was an operational reference standard, not a gold standard. These findings are hypothesis-generating and support only cautious, supervised exploratory use, not autonomous clinical application.

Acknowledgments

Generative artificial intelligence was not used to draft or edit this manuscript.

Funding

The authors declared that no financial support was received for this work.

Data Availability

The deidentified data underlying this study are not publicly available because they were derived from hospital records containing potentially identifiable clinical information. Data may be made available from the corresponding authors upon reasonable request, subject to institutional review and applicable data protection requirements.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Structured prompt template used for large language model evaluation.
[DOCX File, 16 KB - jmir_v28i1e95411_app1.docx]

Multimedia Appendix 2
Large language model workflow and study evaluation framework.
[DOCX File, 172 KB - jmir_v28i1e95411_app2.docx]

Multimedia Appendix 3
Mean specialist appraisal scores for comprehensiveness and perceived concordance with the multidisciplinary team decision.
[DOCX File, 81 KB - jmir_v28i1e95411_app3.docx]

References

1. Del Chiaro M, Sugawara T, Karam SD, Messersmith WA. Advances in the management of pancreatic cancer. BMJ 2023 Dec 13;383:e073995. [doi: 10.1136/bmj-2022-073995] [Medline: 38164628]
2. Stoop TF, Javed AA, Oba A, et al. Pancreatic cancer. Lancet 2025 Apr 5;405(10485):1182-1202. [doi: 10.1016/S0140-6736(25)00261-2] [Medline: 40187844]
3. Kirkegård J, Aahlin EK, Al-Saiddi M, et al. Multicentre study of multidisciplinary team assessment of pancreatic cancer resectability and treatment allocation. Br J Surg 2019 May;106(6):756-764. [doi: 10.1002/bjs.11093] [Medline: 30830974]
4. Benary M, Wang XD, Schmidt M, et al. Leveraging large language models for decision support in personalized oncology. JAMA Netw Open 2023 Nov 1;6(11):e2343689. [doi: 10.1001/jamanetworkopen.2023.43689] [Medline: 37976064]
5. Tung JYM, Le Q, Yao J, et al. Performance of retrieval-augmented generation large language models in guideline-concordant prostate-specific antigen testing: comparative study with junior clinicians. J Med Internet Res 2025 Nov 19;27:e78393. [doi: 10.2196/78393] [Medline: 41259800]

Abbreviations

LLM: large language model

XSL•FO
RenderX

MDT: multidisciplinary team
MSI-H/dMMR: microsatellite instability-high or mismatch repair-deficient

Edited by M Balcarras; submitted 16.Mar.2026; peer-reviewed by K Kostadinov, L Ilodigwe; revised version received 18.May.2026; accepted 19.May.2026; published 30.Jun.2026.

Please cite as:
Liu Z, Zhao X, Mao T, Zhou H, Jiang K
Large Language Model Versus Multidisciplinary Team: Feasibility Study of Pancreatic Cancer Management Recommendations
J Med Internet Res 2026;28:e95411
URL: https://www.jmir.org/2026/1/e95411
doi:10.2196/95411

© Zhuoran Liu, Xin Zhao, Tianyang Mao, Heng Zhou, Kangyi Jiang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Quality of Conventional versus Artificial Intelligence Oral Surgery Consent Forms: Comparative Analysis

Jan Gaessler[1], MD, DDS; Bernhard Remschmidt[1], MD, DMD; Ann-Kathrin Jopp[2], MSc, PhD; Behrouz Arefnia[3], DMD; Adrian Franke[4], MD, DMD; Marcus Rieder[1], MD, DMD

[1]Division of Oral and Maxillofacial Surgery, Department of Dental Medicine and Oral Health, Medical University of Graz, Auenbruggerplatz 5/6, Graz, Austria

[2]Clinical psychologist and psychotherapist in private practice, Hannover, Germany

[3]Division of Restorative Dentistry, Periodontology and Prosthodontics, Department of Dental Medicine and Oral Health, Medical University of Graz, Graz, Austria

[4]Department of Oral and Maxillofacial Surgery, University Hospital Carl Gustav Carus, Dresden University of Technology, Dresden, Germany

Corresponding Author:
Marcus Rieder, MD, DMD
Division of Oral and Maxillofacial Surgery, Department of Dental Medicine and Oral Health, Medical University of Graz, Auenbruggerplatz 5/6, Graz, Austria

Abstract

Artificial intelligence–generated informed consent forms for oral surgery demonstrated higher quality and better readability than conventional web-based forms, though both fell short of recommended comprehension levels.

(J Med Internet Res 2026;28:e59851) doi:10.2196/59851

KEYWORDS

oral surgical procedures; informed consent; quality control; artificial intelligence; oral surgery; consent form; AI; dental health; oral surgeon; patient care; paitent autonomy; dentistry

Introduction

Informed consent is a foundational element of ethical and legal medical care, ensuring patients understand the nature, risks, and alternatives of proposed treatments [1,2]. In oral surgery, where procedures can be complex and invasive, clear and high-quality informed consent forms (ICFs) are especially critical. However, many ICFs exceed the recommended 6th-grade reading level, limiting patient comprehension [3]. With the recent rise of artificial intelligence (AI), particularly large language models (LLMs), there is growing interest in their potential to improve patient communication [4,5]. This study aimed to assess the quality and readability of conventional, web-based oral surgery ICFs and compare them to those generated by AI-based LLMs.

Methods

Ten common oral surgery procedures were selected (ie, apicoectomy, biopsy, bone augmentation, cystectomy, dental implants, incision and drainage, local anesthesia, periodontal surgery, tooth extraction, and wisdom tooth removal). Using Google Chrome in incognito mode, 300 web-based ICFs (ie, 30 per procedure) were collected (see search strategy in Multimedia Appendix 1). In parallel, four LLMs (ChatGPT 3.5, Claude, Bard, and Bing Chat) were prompted to generate ICFs for the same procedures using standardized requests. Per every procedure and LLM, two basic and non-directive prompts were developed to minimize bias and ensure neutrality, resulting in 80 AI-generated ICFs (see Multimedia Appendix 1). Subsequently, two oral and maxillofacial surgeons screened the collected forms using predefined inclusion and exclusion criteria (see Multimedia Appendix 1).

Quality was assessed using a newly developed alteration of the well-established DISCERN instrument [6], namely the Graz Assessment Tool for Written Informed Consent Keypoints (GATWICK; see Multimedia Appendix 1). It was validated through expert review for content relevance and consistency. It includes 11 items scored on a 5-point Likert scale (total score range 11 - 55). Two oral and maxillofacial surgery residents independently rated all forms. Readability was evaluated using six established formulas (ie, Automated Readability Index, Coleman-Liau, Flesch-Kincaid, FORCAST, Gunning Fog, and Simple Measure of Gobbledygook), and an average reading grade level was calculated [7]. Statistical analyses included the Mann-Whitney *U* test, Kruskal-Wallis test, and Kendall tau-b, with significance set at $P \leq .05$.

Results

Of 380 screened documents, 213 ICFs met the inclusion criteria: 136 web-based and 77 AI-generated ones. The inter-rater reliability for GATWICK scores was excellent (intraclass correlation coefficient=0.948).

XSL•FO
RenderX

Regarding the quality, AI-generated ICFs had significantly higher total GATWICK scores compared to web-based ones (median 32.5, IQR 28-35.5 vs median 27.5, IQR 20.375-37; *P*=.007). Items related to treatment alternatives, rationale for recommended intervention, and discussion of options scored particularly higher in AI-generated forms. Web-based ICFs scored better in perioperative behavior instructions.

Considering the readability, web-based forms were significantly harder to read (median grade level 12.45, IQR 11.3-13.325) than AI-generated forms (median 10.7, IQR 10.1-12.4; *P*<.001), although neither met the recommended 6th-grade level. Readability was weakly correlated with overall quality (τ=0.132; *P*=.005).

The word count was higher for web-based forms (median 794 words, IQR 475.25-1068.75 words) than AI-generated ones (median 338 words, IQR 296-381 words; *P*<.001). Longer forms showed a weak correlation with higher quality (τ=0.270; *P*<.001).

Among LLMs, ChatGPT-powered services (ie, ChatGPT 3.5 and Claude) scored significantly higher in terms of quality. ICFs on tooth extraction scored significantly worse when compared with periodontal surgery forms. AI-generated informed consent forms performed significantly better than conventional versions, with notable differences across oral surgical procedures and among the types of LLMs used (Table 1).

Table . Quality of informed consent forms (ICFs) measured through the GATWICK (Graz Assessment Tool for Written Informed Consent Keypoints) score.

Quality	Median (IQR)	*P* value
Overall quality		.007[a]
Conventional (i.e., web-based) ICFs	27.50 (20.125-37)	
Artificial intelligence-generated ICFs	32.50 (28-36.25)	
All combined	31.00 (23-37)	
Differences by procedure		.004[b]
Apicoectomy	27.00 (21.75-34.875)	
Biopsy	30.50 (25.75-33)	
Oral bone augmentation	31.50 (25.75-37.5)	
Dental cystectomy	31.25 (23-33.875)	
Dental implants	33.25 (20.625-37.125)	
Oral incision and drainage	31.50 (23.5-39.5)	
Dental local anesthesia	28.50 (21-34.5)	
Periodontal surgery	36.50 (32.5-42)	
Tooth extraction	23.50 (20-32.75)	
Wisdom tooth removal	28.25 (20-36.875)	
Differences by large language model		<.001[b]
ChatGPT	34.25 (33-37)	
Claude	40.50 (35-43)	
Bing Chat	30.00 (27.25-31.75)	
Google Bard	26.50 (22.75-31.375)	

[a]Mann-Whitney *U* test.

[b]Kruskal-Wallis test.

Discussion

Principal Findings

This study found that conventional oral surgery ICFs available online are generally of modest quality and exceed recommended reading levels. AI-generated ICFs outperformed web-based ones in both quality and readability, although they too fell short of ideal readability standards.

These findings are consistent with prior research across medical disciplines, which show that most ICFs are written at a level too advanced for the average patient [8,9]. Notably, AI-generated forms more consistently addressed key informed consent components such as treatment alternatives and rationale, suggesting that LLMs may serve as valuable tools in drafting patient-centered documents. However, AI models may also produce inaccuracies or omit procedure-specific nuances, highlighting the need for expert review [10].

The limitations of this study include its focus on English-language materials and the variability inherent in AI outputs depending on prompt phrasing or model version. While

XSL•FO
RenderX

the GATWICK tool demonstrated strong reliability, further validation is needed.

Conclusion

AI-based LLMs offer a promising avenue for improving the quality and accessibility of oral surgery informed consent documents. Future efforts should focus on refining AI outputs and integrating clinician oversight to ensure accuracy, comprehensiveness, and patient comprehension.

Acknowledgments

Preliminary results were presented at the 27th Congress of the European Association for Cranio-Maxillo-Facial Surgery (EACMFS) in Rome, Italy, from September 17 to 20, 2024.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Authors' Contributions

JG: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft preparation, Writing – review and editing, Visualization, Project administration. BR: Conceptualization, Methodology, Validation, Investigation, Resources, Writing – original draft preparation, Writing – review and editing. A-KJ: Conceptualization, Methodology, Validation, Investigation, Resources, Writing – original draft preparation, Writing – review and editing. BA: Validation, Resources, Writing – review and editing, Supervision. AF: Validation, Resources, Writing – review and editing, Supervision. MR: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft preparation, Writing – review and editing, Supervision, Project administration. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Methodology showing the utilization of the Boolean operator “OR” helped to broaden the web search as it accounted for differences regarding designation and spelling (i.e., American versus British English). Detailed description of the Graz Assessment Tool of Written Informed Consent Keypoints (GATWICK).
[DOCX File, 29 KB - jmir_v28i1e59851_app1.docx]

References

1. Agozzino E, Borrelli S, Cancellieri M, Carfora FM, Di Lorenzo T, Attena F. Does written informed consent adequately inform surgical patients? A cross sectional study. BMC Med Ethics 2019 Jan 7;20(1):1. [doi: 10.1186/s12910-018-0340-z] [Medline: 30616673]
2. General Medical Council. Consent: patients and doctors making decisions together. URL: https://www.gmc-uk.org/-/media/documents/gmc-guidance-for-doctors---consent---english_pdf-48903482.pdf [accessed 2024-04-23]
3. Powers BJ, Trinh JV, Bosworth HB. Can this patient read and understand written health information? JAMA 2010 Jul 7;304(1):76-84. [doi: 10.1001/jama.2010.896] [Medline: 20606152]
4. Rasteau S, Ernenwein D, Savoldelli C, Bouletreau P. Artificial intelligence for oral and maxillo-facial surgery: A narrative review. J Stomatol Oral Maxillofac Surg 2022 Jun;123(3):276-282. [doi: 10.1016/j.jormas.2022.01.010] [Medline: 35091121]
5. Puladi B, Gsaxner C, Kleesiek J, Hölzle F, Röhrig R, Egger J. The impact and opportunities of large language models like ChatGPT in oral and maxillofacial surgery: a narrative review. Int J Oral Maxillofac Surg 2024 Jan;53(1):78-88. [doi: 10.1016/j.ijom.2023.09.005] [Medline: 37798200]
6. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999 Feb;53(2):105-111. [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
7. Ley P, Florio T. The use of readability formulas in health care. Psychol Health Med 1996 Feb;1(1):7-28. [doi: 10.1080/13548509608400003]

XSL•FO
RenderX

8. Meade MJ, Dreyer CW. Orthodontic treatment consent forms: a readability analysis. J Orthod 2022 Mar;49(1):32-38. [doi: 10.1177/14653125211033301] [Medline: 34325567]
9. Meade MJ, Dreyer CW. How readable are orthognathic surgery consent forms? Int Orthod 2022 Dec;20(4):100689. [doi: 10.1016/j.ortho.2022.100689] [Medline: 36117084]
10. Decker H, Trang K, Ramirez J, et al. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. JAMA Netw Open 2023 Oct 2;6(10):e2336997. [doi: 10.1001/jamanetworkopen.2023.36997] [Medline: 37812419]

Abbreviations

AI: Artificial intelligence
GATWICK: Graz Assessment Tool for Written Informed Consent Keypoints
ICF: informed consent form
LLM: large language models

Edited by N Cahill; submitted 28.Apr.2024; peer-reviewed by H Zhang, R Pranab, W Banjar; revised version received 01.Dec.2025; accepted 01.Dec.2025; published 05.Jan.2026.

Please cite as:
Gaessler J, Remschmidt B, Jopp AK, Arefnia B, Franke A, Rieder M
Quality of Conventional versus Artificial Intelligence Oral Surgery Consent Forms: Comparative Analysis
J Med Internet Res 2026;28:e59851
URL: https://www.jmir.org/2026/1/e59851
doi:10.2196/59851

©Jan Gaessler, Bernhard Remschmidt, Ann-Kathrin Jopp, Behrouz Arefnia, Adrian Franke, Marcus Rieder. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

National Institutes of Health–Funded Artificial Intelligence and Machine Learning Research, 2019 2023: Cross-Sectional Study

Joshua Pei Le[1*], MD; Joseph Morrison[2*], BS; Atul Malhotra[3], MD; Shamim Nemati[3,4], PhD; Gabriel Wardi[3,4], MD, MPH; James S Ford[4*], MAS, MD

[1]School of Medicine, University of Limerick, Limerick, Ireland
[2]School of Medicine, University of California Davis, Sacramento, CA, United States
[3]Department of Medicine, University of California San Diego, La Jolla, CA, United States
[4]Department of Emergency Medicine, University of California San Diego, 9500 Gilman Drive, La Jolla, CA, United States
[*]these authors contributed equally

Corresponding Author:
James S Ford, MAS, MD
Department of Emergency Medicine, University of California San Diego, 9500 Gilman Drive, La Jolla, CA, United States

Abstract

Inflation-adjusted funding for artificial intelligence and machine learning research increased by 233% between fiscal year 2019 and 2023, outpacing the overall National Institutes of Health's budget increase of 12%.

(J Med Internet Res 2026;28:e84861) doi:10.2196/84861

KEYWORDS

artificial intelligence; AI; machine learning; ML; National Institutes of Health; NIH; research funding

Introduction

Artificial intelligence (AI) and machine learning (ML) technologies have begun to revolutionize the practice of medicine. Recent studies have shown that 2 in 3 physicians currently use AI in their clinical practice and more than 3 in 4 scientists currently use AI in their research [1]. Recognizing the growing importance of AI tools, the National Institutes of Health (NIH) began tracking AI and ML as its own funding category in fiscal year (FY) 2019. While previous studies have broadly assessed NIH funding trends for AI and ML research, no studies report data more recently than 2020 [2,3]. Moreover, little is known about the principal investigators (PIs) driving this research. In this study, we aimed to examine trends in NIH-funded AI and ML research and characterize the population of funded PIs.

Methods

We conducted a cross-sectional study of NIH RePORTER (Research Portfolio Online Reporting Tools Expenditures and Results) data between October 1, 2018, and September 30, 2023. We extracted studies indexed under the "Machine Learning and Artificial Intelligence" funding category and collected data on funding institute, mechanism, and award amount (in US $). We also selected a random sample of approximately 25% of PIs who were awarded grants, and we conducted an internet search to collect data on educational background (eg, MD), research setting (eg, academic), and clinical specialty (as appropriate). We searched official websites (eg, university), LinkedIn, or publicly available curriculum vitae. To account for potential differences in educational background (eg, PhD) and clinical specialty (eg, internal medicine) among PIs funded by different institutes, we balanced our random sampling according to the original institute-level proportions of funded grants. Three authors (JM, JPL, and JSF) conducted the search. CIs for proportions were calculated using the Wald method. Tests of trend were conducted using Poisson and linear regression, as appropriate. Analyses were performed using Stata/SE (version 17.0; StataCorp LLC).

Results

Total active projects (including multiyear grants) increased from 1229 to 3449 and total funding increased nominally from US $0.6 billion to $2.3 billion between NIH FY 2019 and 2023 (Figure 1).

There were 5418 unique projects from 4365 unique researchers. Funding trends according to NIH institute and grant mechanism are available in Figures S1 and S2 in Multimedia Appendix 1. We randomly selected approximately 25% of PIs (1091/4395) to conduct further data extraction via internet searches. Among this random sample, the most common educational background was PhD only (70%, 758/1091), followed by MD or DO only (11%, 115/1091), and MD or PhD (9%, 93/1091) (Table S1 in Multimedia Appendix 1). Individuals with only a PhD accounted for 64% of funding dollars, compared to 18% of those with a medical doctoral degree (eg, MD or DO) but no PhD, and 11%

XSL•FO
RenderX

for those with a clinical or professional degree (RN, MD, or PharmD) and a PhD. There were 251/1091 (23%) PIs who held medical doctoral degrees (MD or DO), and of these, 229 (91%, 95% CI 87%-94%) listed residency training. The proportions of total projects and funding dollars by medical specialty are available in Figure 2.

Figure 1. Overall National Institutes of Health (NIH)–funded artificial intelligence and machine learning research projects and funding (in US $) between NIH fiscal year 2019 and 2023. Plotted funding (US $) represents nominal values unadjusted for inflation. Inflation-adjusted values are presented in Table S2A and S2B in Multimedia Appendix 1.

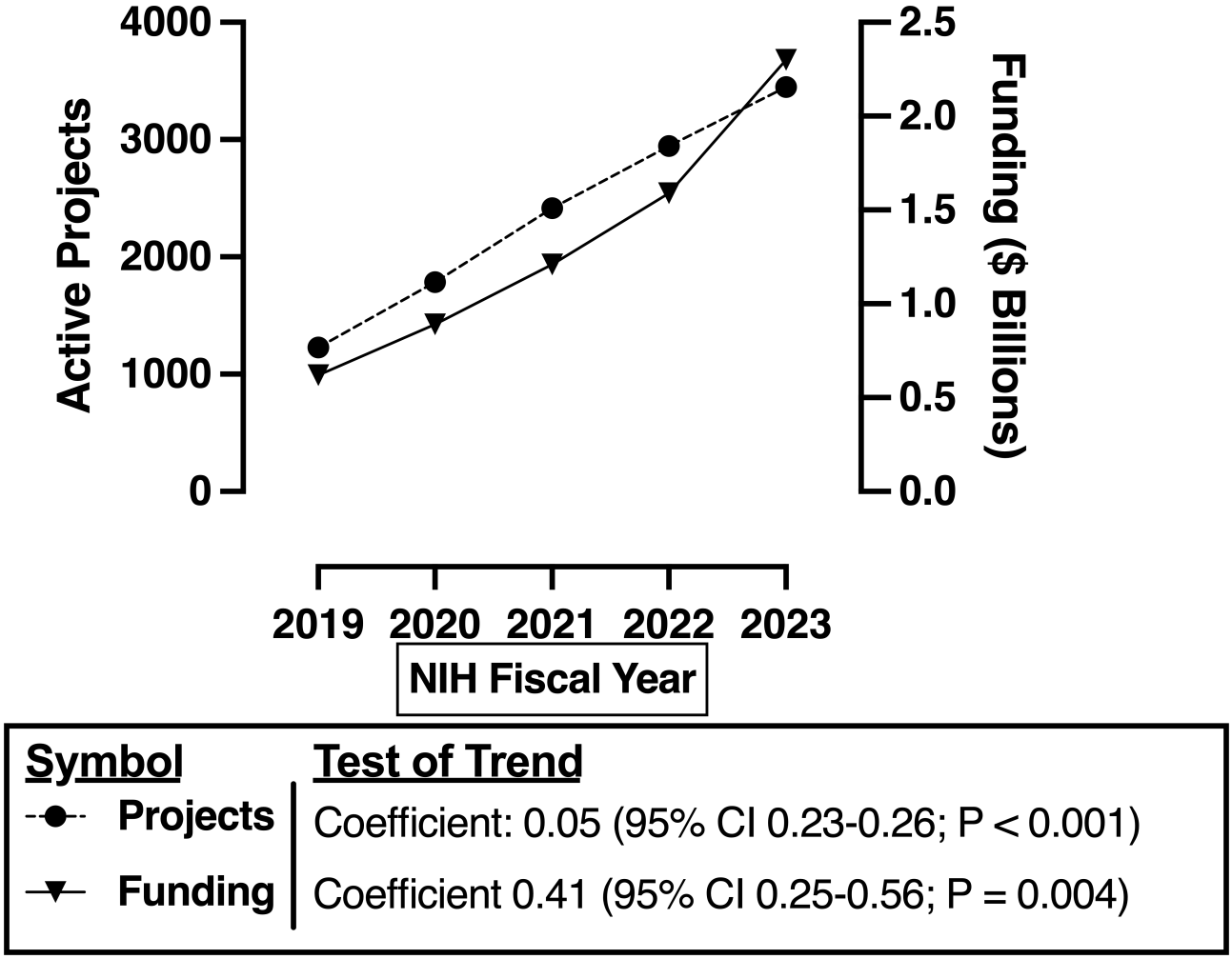

XSL•FO
RenderX

Figure 2. (A) Proportions of total funded principal investigators (PIs) and (B) total funding by specialty. Data from random sample of 1091 unique PIs with National Institutes of Health–funded projects in artificial intelligence or machine learning. CC: critical care; Gen: general; GI: gastroenterology; ID: infectious diseases; IM: internal medicine; Med: medical; Peds: pediatrics; SS, subspecialty; Surg: surgery. *Includes general surgery, urology, vascular surgery, transplant surgery, head and neck surgery, pediatric surgery, neurosurgery, obstetrics and gynecology. †Includes general pediatrics, cardiology, infectious diseases, hematology/oncology, pulmonology, endocrinology, neonatology. ‡Includes endocrinology, rheumatology, nephrology, hematology/oncology. §Includes emergency medicine, anesthesiology, pediatric emergency medicine, dermatology, and PIs with more than one specialty.

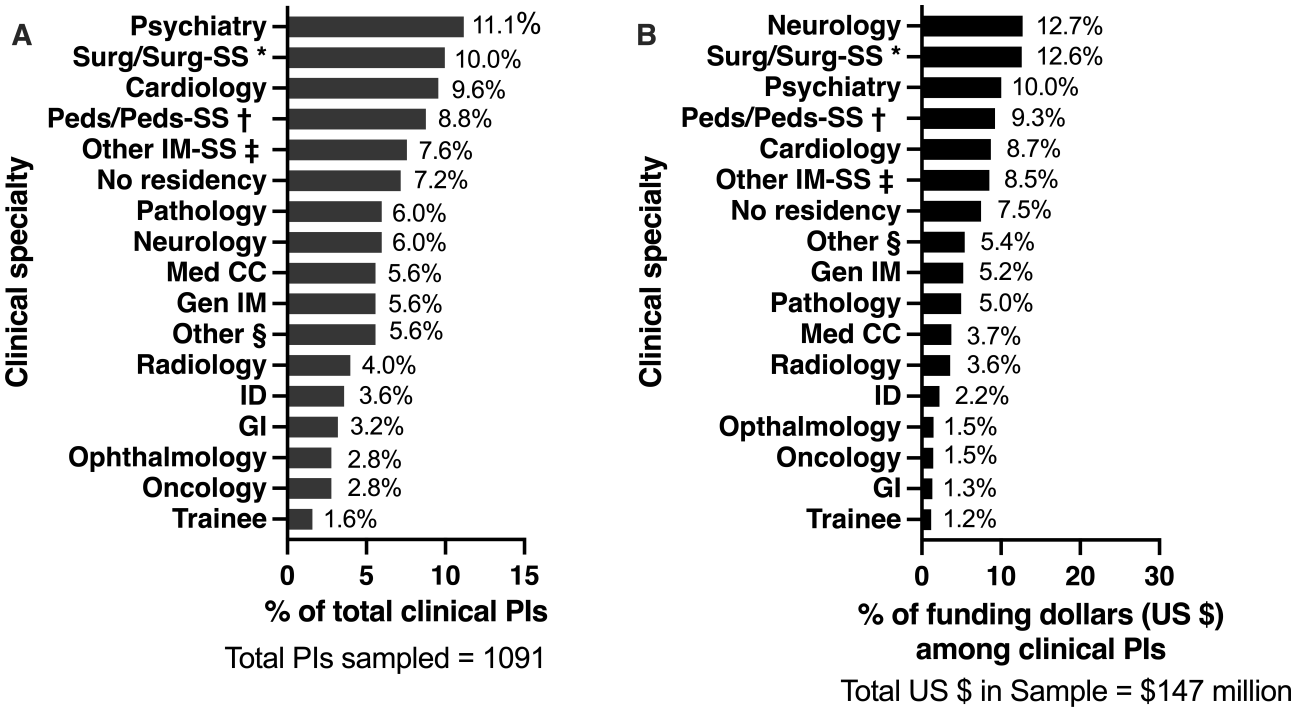

Discussion

In 2017, which was before the NIH began tracking AI and ML as its own spending category, one study found that the NIH supported 535 projects totaling US $264 million, representing approximately 0.7% of the total NIH budget (NIH budget FY 2017; $34.301 billion) [2,4]. Our analysis of 2023 NIH RePORTER funding revealed $2.3 billion in funding (including direct, indirect, and supplemental costs) in this spending category, representing approximately 4.7% of the total NIH budget (NIH budget FY 2023; $49.2 billion) [4]. While the NIH budget increased by 12% (adjusted for inflation) between 2019 and 2023, funding for AI and ML research increased by 233% (adjusted for inflation) over this same period (Table S2A and S2B in Multimedia Appendix 1) [4].

Most AI and ML researchers held only a PhD (70%), and this group accounted for almost two-thirds of all NIH research funding dollars in the AI and ML funding category. By contrast, researchers who held a medical doctoral degree (MD or DO) and who had completed a clinical residency program were relatively underrepresented, accounting for less than 20% of funded researchers and total research dollars. In our sample of researchers who held a medical doctoral degree (MD or DO), those with training in psychiatry had the most funded projects and accounted for the third-highest proportion of funding dollars. While the list of possible applications is extensive and growing, notably, funding among specialties such as emergency medicine, anesthesiology, dermatology, ophthalmology, and various individual medical or surgical subspecialties was considerably lower. These gaps are similar to those seen across other research disciplines and represent clear targets for future NIH strategic investment [5,6].

As US federal funding priorities have been modified by the current administration and the NIH faces anticipated contraction in federal discretionary spending for FY 2026, our findings come at a critical moment [4,7]. Indeed, a recent study demonstrated the potential "ripple effects" of NIH funding cuts, including delays in innovation, workforce attrition, loss of talent to universities and corporations overseas, and increased long-term health care costs [8]. These proposed changes are likely to substantially weaken the United States' contributions to this critical emerging field.

Our analysis included PIs, but data on coinvestigators were not captured. Data related to PI education and training background were obtained from publicly available online sources, which may have been incomplete or outdated at the time of collection. PIs with multiple specialties may have been misclassified. Our spending estimates are higher than those published by the NIH because we aggregated all funding (eg, direct costs, indirect costs, subprojects, and supplemental costs); we believe this methodology provides a more complete view of total research funding [9]. While our random sample of PIs was balanced according to institute-specific proportions, we cannot exclude the possibility of sampling bias. Lastly, the study period overlaps with the COVID-19 pandemic, which may confound funding trends during this time.

In summary, we found that total NIH projects and funding dollars increased substantially between 2019 and 2023. As the US government has enacted substantial NIH budget cuts for FY

XSL•FO
RenderX

2026, future studies are needed to examine the impact of current fiscal policy on future NIH-funded AI and ML research.

Acknowledgments

No generative artificial intelligence was used in the preparation of this manuscript.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

Data will be provided upon reasonable request of the corresponding author.

Authors' Contributions

Conceptualization: AM, SN, GW, JSF
Data Curation: JPL, JM, JSF
Formal analysis: JSF
Investigation: JPL, JM, AM, SN, GW, JSF
Methodology: JSF
Supervision: AM, SN, GW, JSF
Visualization: JSF
Writing – original draft: JPL, JM, JSF
Writing – review & editing: JPL, JM, AM, SN, GW, JSF

Conflicts of Interest

SN and AM are cofounders of a University of California San Diego (UCSD) start-up, Clairyon Inc, which is focused on the commercialization of advanced analytical decision support tools and was formed in compliance with UCSD conflict of interest policies. SN, AM, GW, and JSF receive funding from the National Institutes of Health. SN and JSF receive funding from the University of California Office of the President. GW reports consulting for Abbott Laboratories. GW and JSF report funding from the National Foundation of Emergency Medicine.

Multimedia Appendix 1
Supplemental Tables S1 and S2 and Figures S1 and S2.
[DOCX File, 322 KB - jmir_v28i1e84861_app1.docx]

References

1. Researchers and AI: survey findings. Oxford University Press. 2024 May. URL: https://fdslive.oup.com/www.oup.com/academic/pdf/Researchers-and-AI-survey-findings.pdf [accessed 2025-06-13]
2. Annapureddy AR, Angraal S, Caraballo C, et al. The National Institutes of Health funding for clinical research applying machine learning techniques in 2017. NPJ Digit Med 2020;3(13):13. [doi: 10.1038/s41746-020-0223-9] [Medline: 32025574]
3. Eweje FR, Byun S, Chandra R, et al. Translatability analysis of National Institutes of Health-funded biomedical research that applies artificial intelligence. JAMA Netw Open 2022 Jan 4;5(1):e2144742. [doi: 10.1001/jamanetworkopen.2021.44742] [Medline: 35072720]
4. National Institutes of Health (NIH) funding: FY1996-FY2025 Congressional Research Service. congress.gov. 2024 Jun 25. URL: https://www.congress.gov/crs_external_products/R/PDF/R43341/R43341.54.pdf [accessed 2025-06-13]
5. Morrison J, Le JP, Malhotra A, Nemati S, Wardi G, Ford JS. A snapshot of National Institutes of Health funding for sepsis research: 2019-2023. Ann Emerg Med 2025 Aug;86(2):206-208. [doi: 10.1016/j.annemergmed.2025.03.011] [Medline: 40237685]
6. Schlafly A, Sebro R. Does NIH funding differ between medical specialties? A longitudinal analysis of NIH grant data by specialty and type of grant, 2011-2020. BMJ Open 2022 Dec 30;12(12):e058191. [doi: 10.1136/bmjopen-2021-058191] [Medline: 36585146]
7. NIH Office of Budget. URL: https://officeofbudget.od.nih.gov [accessed 2025-10-28]
8. Jalali MS, Hasgul Z. Potential trade-offs of proposed cuts to the US National Institutes of Health. JAMA Health Forum 2025 Jul 3;6(7):e252228. [doi: 10.1001/jamahealthforum.2025.2228] [Medline: 40711777]
9. Estimates of funding for various research, condition, and disease categories (RCDC). Research Portfolio Online Reporting Tools (RePORT). URL: https://report.nih.gov/funding/categorical-spending [accessed 2025-06-09]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
FY: fiscal year
ML: machine learning
NIH: National Institutes of Health
PI: principal investigator
RePORTER: Research Portfolio Online Reporting Tools Expenditures and Results

Edited by A Coristine; submitted 25.Sep.2025; peer-reviewed by DJ Dzikowicz, KM Gillen; revised version received 24.Nov.2025; accepted 24.Nov.2025; published 08.Jan.2026.

Please cite as:
Le JP, Morrison J, Malhotra A, Nemati S, Wardi G, Ford JS
National Institutes of Health–Funded Artificial Intelligence and Machine Learning Research, 2019 - 2023: Cross-Sectional Study
J Med Internet Res 2026;28:e84861
URL: https://www.jmir.org/2026/1/e84861
doi:10.2196/84861

© Joshua Pei Le, Joseph Morrison, Atul Malhotra, Shamim Nemati, Gabriel Wardi, James S Ford. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Transformer-Based Topic Modeling: Characterizing Cannabis Product Adverse Experiences Self-Reported as Requiring Medical Attention on Reddit

Tim Ken Mackey[1,2,3,4], MAS, PhD; Matthew C Nali[2,3,4], BA; Meng Zhen Larsen[2,3], BS; Zhuoran Li[3], BS, MS; Cassandra L Taylor[5], PhD; Beverly Wolpert[6], MS, PhD; Catharine Trice[6,7], MA

[1]Global Health Program, Department of Anthropology, University of California, San Diego, La Jolla, CA, United States
[2]San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA, United States
[3]S-3 Research LLC, San Diego, CA, United States
[4]Global Health Policy and Data Institute, San Diego, CA, United States
[5]Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, United States
[6]Office of Surveillance Strategy and Risk Prioritization, Human Foods Program, U.S. Food and Drug Administration, College Park, MD, United States
[7]Office of Women's Health, U.S. Food and Drug Administration, Silver Spring, MD, United States

Corresponding Author:
Tim Ken Mackey, MAS, PhD
Global Health Program
Department of Anthropology
University of California, San Diego
9500 Gilman Dr
MC 0505
La Jolla, CA, 92093
United States
Phone: 1 9514914161
Email: tmackey@ucsd.edu

Abstract

This study uses keyword filtering, a transformer-based algorithm, and inductive content coding to identify and characterize cannabis adverse experiences as discussed on the social media platform Reddit and reports a total of 1177 self-reported adverse experiences requiring medical attention.

(J Med Internet Res 2026;28:e82661) doi:10.2196/82661

KEYWORDS

cannabis; adverse events; reddit; machine learning, topic modeling

Introduction

Cannabis is one of the most frequently used intoxicating substances in the United States [1]. Changing policies and attitudes have increased cannabis use, yet safety profiles for different cannabis-derived products (CDPs) (eg, products containing tetrahydrocannabinols, cannabidiol, or cannabinoids derived from cannabis) are not well understood. Studies have examined internet and social media data for health concerns among cannabis users, including for adverse health outcomes [2-8]. However, no study has specifically characterized the types of adverse experiences requiring medical attention, as self-reported online. We aimed to use a transformer-based algorithm to identify and characterize adverse experiences self-reported as requiring medical attention (SRRMA) on Reddit.

Methods

For this observational retrospective study, we collected user-generated posts and comments from 27 subreddits(/r/) related to cannabis use (Figure 1; Multimedia Appendix 1). A set of keywords for adverse experiences SRRMA was used for data mining and manual coding to generate a sample of “signal” Reddit data (ie, posts about cannabis product adverse experiences requiring medical attention: hospitalization, or visiting a health care professional, emergency room, or urgent

XSL•FO
RenderX

care). This initial sample of SRMMA prelabeled "signal" was joined with data from the full Reddit data corpus.

Figure 1. Summary of the study methodology, including data collection, data processing/filtering, BERT (bidirectional encoder representations from transformers) analysis, and results (signal data): (1) data collection for cannabis-related subreddits (both posts and comments); (2) keyword filtering for adverse experience self-reported as requiring medical attention (AE SRRMA)-related signals; (3) separation of sample of "signals" for general adverse experiences (GAE) self-reported as not requiring medical attention and AE SRRMA; (4) use of BERT topic model (k=5, 10, 15, 20) after data filtering with a sample of AE SRRMA keywords and labeled AE SRRMA data joined with the unlabeled dataset to analyze the full corpus of Reddit posts and comments; (5) selection of BERT topic clusters with the highest yield of prelabeled signals for AE SRRMA that were manually annotated to identify additional AE SRRMA signals (process repeated to achieve higher thematic saturation); newly identified signals were joined with the unlabeled dataset for the next round of BERT topic modeling; and (6) AE SRRMA signal identification from Reddit posts and comments summarized in this publication.

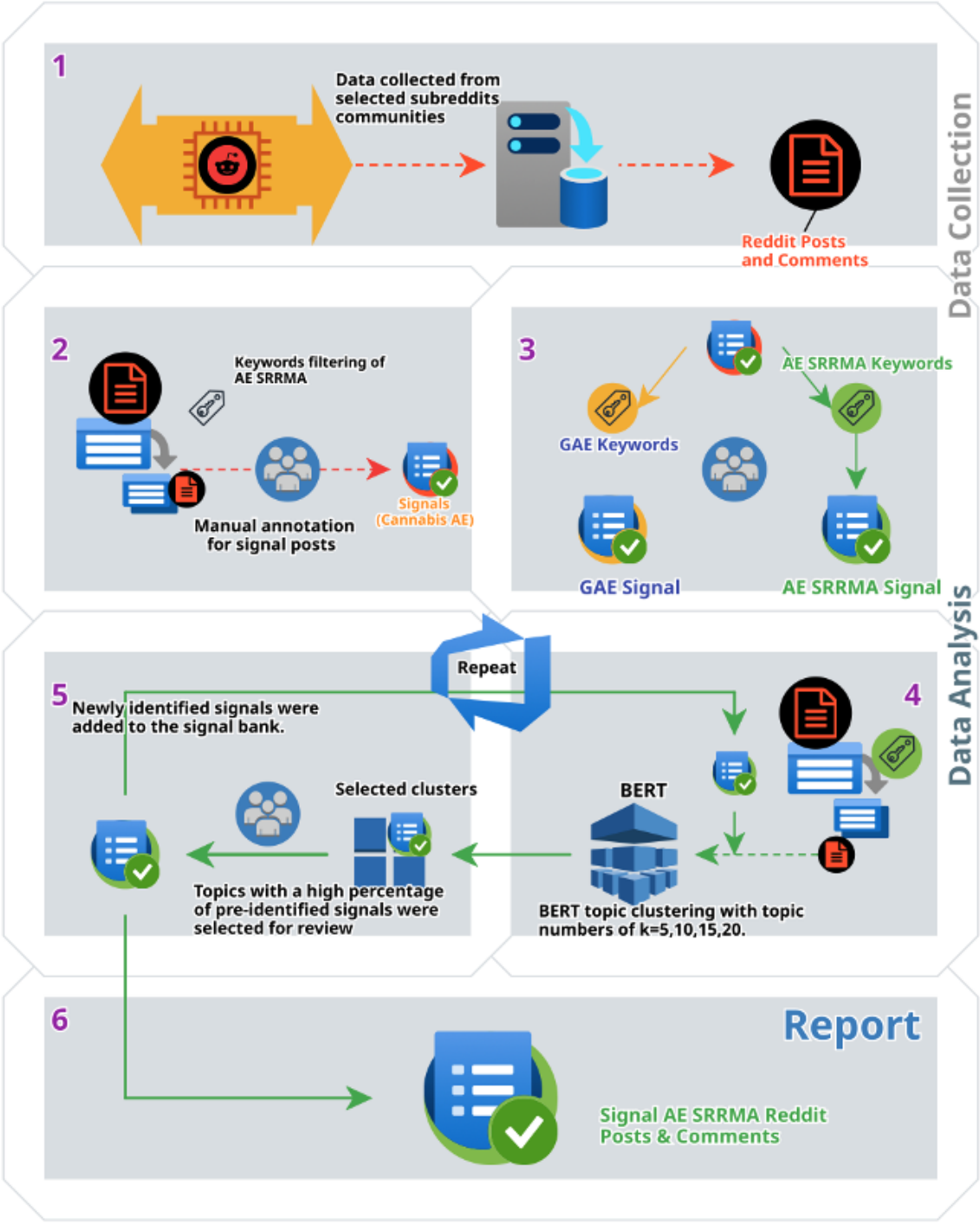

XSL•FO
RenderX

Natural language processing was applied to cluster Reddit posts and comments for further manual annotation. We used the bidirectional encoder representations from transformers (BERT) topic model with the Python (version 3.7) package BERTopic (version 0.6.0), which is a pretrained, self-supervised transformer-based algorithm that embeds texts, extracts topics, and then clusters texts.

BERT was used to cluster groups of nonlabeled data with SRRMA prelabeled signal data into k topic clusters, where the value of k is determined according to the dataset size (5≤k≥20). In each generated cluster, we calculated the percentages of SRRMA prelabeled signal data in that topic cluster to estimate how well that cluster was related to our SRRMA prelabeled data (% of potential SRRMA prelabeled signal data in a cluster that could be appended to the SRRMA prelabeled dataset = amount of SRRMA prelabeled training signal data in a cluster/amount of total data in a cluster). The higher the percentage, the more likely the uncoded data were related to our SRRMA prelabeled topic. Cluster selection was determined based on the highest percentages. The new signals were then added to our coded signal pool for the next iteration of BERT modeling: we filtered the uncoded dataset with the keywords repeatedly, combined the filtered uncoded data with all prelabeled signal data, and then performed the next iteration of BERT to find the best clusters to manually annotate for identifying additional signals. Posts and comments from these clusters were extracted, and an inductive coding approach was used to label and characterize cannabis-related adverse experiences and adverse experiences that met the SRRMA criteria. Human ethics and consent to participate were not applicable, as all information was from the public domain and did not involve user interaction, and any identifiable information was aggregated and removed from the results.

Results

We collected 1,795,478 Reddit posts/comments. After BERT and coding, 1177 posts/comments comprising 1542 user-generated mentions of SRRMA adverse experiences for cannabis products were detected between July 2017 and December 2022 (Table 1). Coders (TKM, MN, MZ) achieved a high intercoder reliability score (κ=0.90). From the 27 subreddits, the top 10 SRRMA adverse experiences were vomiting (284/1542, 18.42%), nausea (171/1542, 11.09%), panic attack (122/1542, 7.91%), abdominal/stomach pain (96/1542, 6.23%), concern over elevated heart rate (92/1542, 5.97%), anxiety (60/1542, 3.89%), chest pain (53/1542, 3.44%), general complaints of sickness (31/1542, 2.01%), symptoms attributed to Cannabinoid Hyperemesis Syndrome (30/1542, 1.95%), and paranoia (30/1542, 1.95%). Additionally, 108 (9.18%) SRRMAs mentioned a CDP but not a specific adverse experience. SRRMA adverse experiences included users self-reporting seeking medical attention with visits/admissions to hospitals (533/1177, 45.28%), emergency rooms (569/1177, 48.34%), health care professionals (70/1177, 5.95%), and urgent care (5/1177, 0.43%).

For cannabis product use characteristics with an SRRMA, most (988/1177, 83.94%) were inhalation (eg, inhaler, joint, vape) followed by ingestion (59/1177, 5.01%; eg, edibles) products. For users reporting intent of use (ie, post and comment), 94.65% (1114/1177) were adult use (ie, recreational), followed by 3.57% (42/1177) therapeutic (ie, for health benefit), 1.02% (12/1177) unknown, 0.42% (5/1177) medical (ie, from a medical dispensary), and 0.34% (4/1177) unintentional use. A subset of users reported co-use with other substances, specifically tobacco and nicotine products (n=7), prescription medications (eg, alprazolam, antidepressants, n=22), illicit drugs (eg, lysergic acid diethylamide, cocaine, n=5), and dual use with other cannabinoids (eg, general tetrahydrocannabinol use, cannabidiol, delta 8, n=12).

Table 1. Characteristics of adverse experiences self-reported as requiring medical attention, posted online on Reddit between July 2017 and December 2022.

Post and comment type/intent of use	Adverse Experiences Self-Reported as Requiring Medical Attention (AE SRRMA)[a]				
	Hospitalization (n=533), n	Health care professional visit (n=70), n	Emergency room visit (n=569), n	Urgent care (n=5), n	Proportion of platform AE SRRMA posts (N=1177), n (%)
Medical	1	0	4	0	5 (0.42)
Recreational	513	66	530	5	1114 (94.65)
Therapeutic	15	4	23	0	42 (3.57)
Unintended	0	0	4	0	4 (0.34)
Unknown	4	0	8	0	12 (1.02)

[a]AE SRRMA indicates that a cannabis user sought medical care because of a concern that their symptom or symptoms were serious or potentially life-threatening. Intent of use categories included medical (user mentions acquiring a product from a medical dispensary); recreational (user mentions using a cannabis product, price, or acquisition, in the context of a recreational or social event or circumstance); therapeutic (user mentions using a cannabis product for medical reasons); unknown (could not be determined based on data that was available); and unintended (user mentions using a cannabis-derived product accidentally or unintentionally).

XSL•FO
RenderX

Discussion

We identified 1177 Reddit user-generated posts that described CDP SRMMA adverse experiences. Results are similar to those of a study analyzing 28,630 cannabis exposures/cases reported to state poison control centers from 2017 to 2019 that found CDP concentrates, vaporized liquids, and edibles documented with "moderate or greater" medical outcomes (ie, effects usually requiring a form of medical treatment) [9]. Studies using transformer-based and large language models also found similar reports of adverse experiences by Reddit users but only coded for generalized adverse health outcomes [7] or only reviewed a single subreddit (r/delta8) [8,9]. Our study builds on the methodology of these prior studies by using a transformer-based algorithm "seeded" with prelabeled signal data to find additional unlabeled data—an approach similar to a few-shot learning approach but without formal model training.

Study limitations include terms not inclusive of all cannabis safety-related experiences and lack of cross-validation of users' self-reports (eg, lack of validation with clinical data). Further, social media data are not representative of all cannabis use behavior, and what cannabis products may have led to adverse experiences (eg, contaminated/adulterated products or products not from licensed dispensaries) were unclear (eg, we observed more Reddit discussions on adult-use cannabis [ie, recreational] than on medical cannabis use, which may limit generalizability). Though exploratory, study results may be used in conjunction with other cannabis-related adverse experience surveillance data (eg, National Poison Data System and FDA MedWatch [6]) and can provide important context to adverse experiences that result in medical attention for purposes of helping consumers and clinicians better understand potential cannabis risks.

Acknowledgments

Generative artificial intelligence was not used for any portion of this manuscript in its writing, editing, or generation.

Funding

This research was supported by the Food and Drug Administration (FDA) of the US Department of Health and Human Services (HHS) as part of a contractual award (contract 75F40123C00030) totaling US $174,971.73, with 100% funded by FDA/HHS. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by FDA/HHS or the US government.

Authors' Contributions

Conceptualization: TKM, CLT, BJW, CT
Methodology: TKM, MCN, MZL, ZL
Software: ZL
Validation: TKM, MCN, MZL
Formal analysis: TKM, MCN, MZL
Investigation: TKM, MCN, MZL
Data curation: MCN, MZL, ZL
Writing – original draft: TKM, MCN, MZL
Writing – review & editing: TKM, CLT, BJW, CT
Supervision: TKM
Project administration: TKM
Funding acquisition: TKM

Conflicts of Interest

TKM, MN, MZL, and ZL are employees of the startup company S-3 Research LLC. TKM is also a cofounder and co-owner of the company. S-3 Research is a minority owned small business that was originally created through a National Institute on Drug Abuse startup business award and has been subsequently funded through government contracts with federal agencies to develop research tools and services for data science in public health, including social listening, data mining, and machine learning approaches. TKM is Editor-in-Chief of *JMIR Infodemiology*, a JMIR Publications journal, at the time of this publication. The authors report no other conflict of interests associated with this manuscript.

Multimedia Appendix 1
Data collection, filtering, and content coding methodology.
[DOCX File , 51 KB - jmir_v28i1e82661_app1.docx]

References

XSL•FO
RenderX

1. U.S. Centers for Disease Control and Prevention. Cannabis facts and stats. Data and Statistics. 2025 Mar 07. URL: https://www.cdc.gov/cannabis/data-research/facts-stats/?CDC_AAref_Val=https://www.cdc.gov/marijuana/data-statistics.htm [accessed 2026-01-27]
2. Meacham MC, Paul MJ, Ramo DE. Understanding emerging forms of cannabis use through an online cannabis community: an analysis of relative post volume and subjective highness ratings. Drug Alcohol Depend 2018 Jul 01;188:364-369 [FREE Full text] [doi: 10.1016/j.drugalcdep.2018.03.041] [Medline: 29883950]
3. Allem J, Majmundar A, Dormanesh A, Donaldson SI. Identifying health-related discussions of cannabis use on Twitter by using a medical dictionary: content analysis of tweets. JMIR Form Res 2022 Feb 25;6(2):e35027 [FREE Full text] [doi: 10.2196/35027] [Medline: 35212637]
4. Hallinan CM, Khademi Habibabadi S, Conway M, Bonomo YA. Social media discourse and internet search queries on cannabis as a medicine: a systematic scoping review. PLoS One 2023;18(1):e0269143 [FREE Full text] [doi: 10.1371/journal.pone.0269143] [Medline: 36662832]
5. Dilley JA, Graves JM, Brooks-Russell A, Whitehill JM, Liebelt EL. Trends and characteristics of manufactured cannabis product and cannabis plant product exposures reported to US poison control centers, 2017-2019. JAMA Netw Open 2021 May 03;4(5):e2110925 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.10925] [Medline: 34028553]
6. Smith BP, Hoots B, DePadilla L, Roehler DR, Holland KM, Bowen DA, et al. Using transformer-based topic modeling to examine discussions of delta-8 tetrahydrocannabinol: content analysis. J Med Internet Res 2023 Dec 21;25:e49469 [FREE Full text] [doi: 10.2196/49469] [Medline: 38127427]
7. Leas EC, Ayers JW, Desai N, Dredze M, Hogarth M, Smith DM. Using large language models to support content analysis: a case study of ChatGPT for adverse event detection. J Med Internet Res 2024 May 02;26:e52499 [FREE Full text] [doi: 10.2196/52499] [Medline: 38696245]
8. Hines MC, Harinstein LM, Kortepeter CM. Reporting adverse events for cannabis to the FDA. N Engl J Med 2020 Jan 02;382(1):98-98. [doi: 10.1056/nejmc1913460]
9. Leas EC, Harati RM, Satybaldiyeva N, Morales NE, Huffaker SL, Mejorado T, et al. Self-reported adverse events associated with Δ-Tetrahydrocannabinol (Delta-8-THC) Use. J Cannabis Res 2023 May 23;5(1):15 [FREE Full text] [doi: 10.1186/s42238-023-00191-y] [Medline: 37217977]

Abbreviations

BERT: bidirectional encoder representations from transformers
CDP: cannabis-derived product
SRRMA: self-reported as requiring medical attention

Edited by A Mavragani; submitted 19.Aug.2025; peer-reviewed by SN Sundaradhas, JM Conway; comments to author 03.Nov.2025; revised version received 09.Jan.2026; accepted 20.Jan.2026; published 04.Feb.2026.

Please cite as:
Mackey TK, Nali MC, Larsen MZ, Li Z, Taylor CL, Wolpert B, Trice C
Transformer-Based Topic Modeling: Characterizing Cannabis Product Adverse Experiences Self-Reported as Requiring Medical Attention on Reddit
J Med Internet Res 2026;28:e82661
URL: https://www.jmir.org/2026/1/e82661
doi:10.2196/82661
PMID:

©Tim Ken Mackey, Matthew C Nali, Meng Zhen Larsen, Zhuoran Li, Cassandra L Taylor, Beverly Wolpert, Catharine Trice. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Sex Differences in Medical Specialist Physicians' Electronic Health Record In-Basket Workloads and the Implications for Compensation and Equity: Retrospective Cross-Sectional Study

Ashwini Nadkarni[1], MD; Laura Glick[2], MD; Molly Becker[3], MS; Michael Healey[1], MD; David Stein[1], PhD; Mahyar Heydarpour[1], PhD; Lisa S Rotenstein[4], MD, MBA

[1]Brigham and Women's Hospital, Boston, MA, United States
[2]Massachusetts General Hospital, Boston, MA, United States
[3]Suffolk University, Boston, MA, United States
[4]University of California, San Francisco, San Francisco, CA, United States

Corresponding Author:
Ashwini Nadkarni, MD
Brigham and Women's Hospital
75 Francis Street
Boston, MA, 02115
United States
Phone: 1 6177329811
Email: anadkarni@bwh.harvard.edu

Abstract

In this retrospective cross-sectional study, we assessed sex differences in in-basket messages generated by outpatient workflows among internal medicine specialists as compared to relative value units; we found that female physicians had a greater burden of in-basket work for each unit of paid clinical care.

(J Med Internet Res 2026;28:e79172) doi:10.2196/79172

KEYWORDS

Electronic health records; In-basket messaging; Physician workload; Sex differences; Relative value units

Introduction

Female physicians in primary care receive higher volumes of patient and staff messages in the electronic health record (EHR) in-basket [1,2]. However, little is known on whether sex differences extend into medical specialties and impact on workload disparities relative to compensated care. In this cross-sectional study, we assessed sex differences in in-basket messages generated by outpatient workflows among internal medicine specialists as compared to relative value units (RVUs). We aimed to clarify whether female medical specialty physicians experience disproportionately greater uncompensated electronic work for each unit of paid clinical care.

Methods

Ethical Considerations

We followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) cross-sectional study guidelines. The Mass General Brigham Institutional Review Board deemed the study exempt (2021P001356). The informed consent requirement was waived because the study involved secondary analysis of deidentified EHR data and posed minimal risk to participants. Participants did not receive compensation. Although prior literature refers to "gender" differences, our analysis examined sex differences, so we used the terms "female" and "male." We conducted the study at a large, academic medical center. All attending physicians actively practiced in outpatient specialty practices with measurable in-basket activity during the study period.

Study Procedures

We collected monthly, ambulatory in-basket physician data and patient characteristics for each attending physician's patient panel from Epic for January through March 2021 and normalized monthly in-basket burden measures by monthly RVUs. Patients are generally assigned to physicians based on availability and patient preference. We obtained physician sex, academic rank, and practice years from the public institutional directory, which reports sex as male or female; no physicians were listed as nonbinary. We categorized specialties as those including a

XSL•FO
RenderX

substantial procedural component (cardiology, gastroenterology, and pulmonology) versus those predominantly not procedural (genetics, geriatrics, hematology, immunology, infectious disease, nephrology, palliative care, rheumatology, and sleep medicine).

Statistical Analysis

We compared continuous variables using Wilcoxon rank sum tests, and categorical variables using chi-square tests or Fisher exact tests when expected cell counts were small. To assess sex differences in in-basket burden, we fit multivariable linear regression models with in-basket outcomes as continuous dependent variables, physician sex (female vs male) as the primary independent variable, and covariates that included academic rank, years in practice, mean patient age, mean number of problems per patient, and panel size. We included both academic rank and years in practice to capture distinct aspects of physician experience and considered collinearity diagnostics as acceptable. We derived adjusted marginal means and sex differences with 95% CIs from each model. We conducted an analysis in the full sample and in a subgroup restricted to nonprocedural specialties to assess robustness. We excluded physicians with incomplete covariate data since comparisons showed no meaningful differences between those with complete versus incomplete in-basket data. We defined statistical significance as 2-tailed $P<.05$ and performed analyses using JMP (SAS Institute).

Results

As shown in Table 1, of 384 physicians with baseline data, sex was reported for 367 (n=146 female; n=221 male); 17 physicians with missing sex were excluded from sex-based analyses. Complete EHR in-basket data were available for 304 physicians (n=130 female; n=174 male). Of these, RVU information was available for 296 physicians (n=124 female; n=172 male), who were included in unadjusted RVU-normalized analyses. Adjusted analyses included 257 physicians (n=108 female; n=149 male) with complete covariate data.

Complete EHR in-basket burden data was available for 304 physicians (n=130 female; n=174 male). Physicians with complete in-basket data did not differ meaningfully from those with incomplete data with respect to physician sex, academic rank, years in practice, or patient panel characteristics. Table 2 presents adjusted marginal means with 95% CIs for female and male physicians, along with adjusted sex differences and associated *P* values. After adjustment for academic rank, years in practice, and panel characteristics, female physicians sustained higher RVU-normalized in-basket workloads than male physicians, suggesting that these differences are not driven solely by differences in the ratio of physicians in predominantly procedural subspecialties. In multivariable linear regression models, female physicians had higher adjusted mean daily total in-basket time, time spent completing messages, and EHR time outside 7 AM to 7 PM per RVU than male physicians. In contrast, adjusted differences in message counts were attenuated, with the largest and most consistent sex differences observed for staff messages per RVU. Significant differences in EHR in-basket burden metrics by sex persisted in analyses including all specialties and in analyses restricted to nonprocedural specialties (Table S1 in Multimedia Appendix 1).

XSL•FO
RenderX

Table 1. Internal medicine specialty physician, productivity, and patient panel characteristics.

Variables	Female physicians	Male physicians	*P* value
RVUs[a], mean (SD)	534.92 (42.3)	666.1 (44.2)	.03
Demographic characteristics, n (%)			
Academic rank[b]			<.001
Total	146 (39.8)	221 (60.2)	
Instructor	64 (43.8)	59 (26.7)	
Assistant professor	53 (36.3)	66 (29.9)	
Associate professor	22 (15.1)	51 (23.1)	
Professor	7 (4.8)	45 (20.4)	
Years in practice[b]			<.001
Total	151 (41.1)	233 (63.5)	
0-10 years	27 (17.9)	22 (9.4)	
11-20 years	68 (45.3)	70 (30.0)	
21-30 years	37 (24.5)	60 (25.8)	
31 years or more	19 (12.6)	81 (34.8)	
Panel characteristics, mean (SD)			
Patient age (years)	56.2 (9.9); n=113 patients	58.3 (8.8); n=153 patients	.07
Problems per patient (n)	14.7 (4.7); n=113 patients	14.8 (4.1); n=153 patients	.84
Patients per RVU (n)	0.37 (0.08); n=124 patients	0.21 (0.03); n=172 patients	.08
EHR in-basket burden data, mean (SD)			
Daily time spent on in-basket per physician per RVU (minutes)	314 (32.6)	246 (25.1)	.02
Daily time spent on completing messages per physician per RVU (minutes)	18,888.07 (1961.4)	14,808.09 (1507.6)	.02
Daily time spent outside 7 AM-7 PM per physician per RVU (minutes)	113.2 (17.8)	101.2 (17.4)	.14
Complete messages (n)	387.9 (40.3)	360.4 (33.2)	.17
Staff messages (n)	49.1 (5.7)	34.3 (4.1)	.03
Patient advice messages (n)	70.8 (9.13)	47.5 (5.9)	.007

[a]RVU: relative value unit.

[b]Percentantages in these columns use the value in the corresponding entry in the same column for Total under the same nested heading.

XSL•FO
RenderX

Table 2. Relative value unit (RVU)–normalized electronic health record in-basket burden metrics by physician sex. (Adjusted for physician demographics, patient characteristics, and panel size.)

In-basket efficiency metric	Unadjusted, normalized by RVU				Adjusted, normalized by RVU			
	Female (n=124), mean (SD; 95% CI)	Male (n=172, mean (SD; 95% CI)	Mean difference	*P* value	Female (n=108), mean (SD; 95% CI)	Male (n=149), mean (SD; 95% CI)	Mean difference	*P* value
Daily time spent on in-basket per RVU (min)	1.46 (1.83; 1.14-1.78)	0.80 (1.09; 0.64-0.96)	0.66	.001	1.27 (0.09; 1.25-1.29)	0.96 (0.08; 0.94-0.97)	0.31	<.001
Daily time spent on completing messages per RVU (min)	88.21(110.16; 68.82-107.6)	48.89 (65.98; 39.03-58.75)	39.3	<.001	76.40 (5.69; 75.4-77.4)	57.71 (4.65; 57.0-58.4)	18.7	.01
Daily time spent outside 7 AM-7 PM per RVU (min)	0.55 (0.79; 0.65-0.93)	0.29 (0.47; 0.22-0.36)	0.26	.002	0.44 (0.06; 0.43-0.45)	0.30 (0.05; 0.29-0.31)	0.14	.05
Messages per RVU (n)	1.36 (1.46; 1.10-1.62)	1.01 (0.93; 0.87-1.15)	0.35	.03	1.53 (0.12; 1.51-1.55)	1.41 (0.10; 1.39-1.43)	0.12	.44
Staff messages per RVU (n)	0.19 (0.21; 0.15-0.23)	0.10 (0.11; 0.08-0.12)	0.09	<.001	0.18 (0.01; 0.16-0.20)	0.12 (0.01; 0.1-0.14)	0.06	.002
Patient advice messages per RVU (n)	0.006 (0.006; 0.004-0.011)	0.004 (0.005; 0.003-0.011)	0.002	.03	0.005 (0.0009; 0.003-0.007)	0.006 (0.0008; 0.0048-0.0072)	0.001	.80

Discussion

Female physicians spent more time on all in-basket activities and received a higher number of staff and patient messages per RVU. These findings persisted even after adjusting for physicians' age, years in practice, and patient panel characteristics, and among procedural and nonprocedural specialties. Our results expand beyond primary care to internal medicine specialists, normalizing in-basket burden by RVU to emphasize that female physicians have a greater burden of in-basket work per unit of paid clinical care. These findings potentially reflect other studies on differential expectations for female physicians from patients and staff. However, our study demonstrates the pervasiveness of such sex-based differences across medical specialties, and notably, their impact on pay equity [3-5].

Limitations of our study could have influenced our findings. For instance, using data from a single academic medical center over a relatively short period could have reduced generalizability. Exclusion of physicians with incomplete data, who may have differed in unmeasured ways, could have biased the estimates. Use of RVUs could have incompletely captured clinical workloads. Finally, a lack of data on medical assistant support or clinical full-time equivalent hours was a confounder.

Our results have practical implications for physician well-being in ambulatory medical specialties. Compensation for in-basket activities would enable recognition of workload disparities. Use of artificial intelligence to automate message management could reduce inequities in uncompensated work. To address physician burnout, health systems must directly confront and mitigate sex disparities in EHR workload. Moreover, given physicians are currently not generally compensated for in-basket work, future studies should assess interventions to reduce sex differences in in-basket burden relative to compensated care delivered.

Acknowledgments

We thank Laura Bockoff for her assistance in data clean-up. We did not use generative artificial intelligence to produce this manuscript.

Funding

This study was funded through a microgrant provided by the Brigham Research Institute.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Relative value unit–normalized electronic health record in-basket burden metrics by physician sex and nonprocedural internal medicine specialties.
[DOCX File , 21 KB - jmir_v28i1e79172_app1.docx]

Multimedia Appendix 2

XSL•FO
RenderX

STROBE checklist.
[DOCX File , 38 KB - jmir_v28i1e79172_app2.docx]

References

1. Rittenberg E, Liebman JB, Rexrode KM. Primary care physician gender and electronic health record workload. J Gen Intern Med 2022 Oct 06;37(13):3295-3301 [FREE Full text] [doi: 10.1007/s11606-021-07298-z] [Medline: 34993875]
2. Liddell SS, Tomasi AG, Halvorsen AJ, Stelling BEV, Leasure EL. Gender disparities in electronic health record usage and inbasket burden for internal medicine residents. J Gen Intern Med 2024 Nov 26;39(15):2904-2909. [doi: 10.1007/s11606-024-08861-0] [Medline: 38926324]
3. Linzer M, Harwood E. Gendered expectations: do they contribute to high burnout among female physicians? J Gen Intern Med 2018 Jun 12;33(6):963-965 [FREE Full text] [doi: 10.1007/s11606-018-4330-0] [Medline: 29435727]
4. Rotenstein L, Jay Holmgren A. COVID exacerbated the gender disparity in physician electronic health record inbox burden. J Am Med Inform Assoc 2023 Sep 25;30(10):1720-1724. [doi: 10.1093/jamia/ocad141] [Medline: 37436709]
5. Rotenstein LS, Melnick ER, Jeffery M, Zhang J, Sinsky CA, Gitomer R, et al. Association of primary care physicians' electronic inbox activity patterns with patients' likelihood to recommend the physician. J Gen Intern Med 2024 Jan 20;39(1):150-152. [doi: 10.1007/s11606-023-08417-8] [Medline: 37731135]

Abbreviations

EHR: electronic health record
RVU: relative value unit

Edited by A Mavragani; submitted 25.Jun.2025; peer-reviewed by D Lew, A Murzello; comments to author 25.Nov.2025; revised version received 29.Jan.2026; accepted 30.Jan.2026; published 18.Feb.2026.

Please cite as:
Nadkarni A, Glick L, Becker M, Healey M, Stein D, Heydarpour M, Rotenstein LS
Sex Differences in Medical Specialist Physicians' Electronic Health Record In-Basket Workloads and the Implications for Compensation and Equity: Retrospective Cross-Sectional Study
J Med Internet Res 2026;28:e79172
URL: https://www.jmir.org/2026/1/e79172
doi:10.2196/79172
PMID:

©Ashwini Nadkarni, Laura Glick, Molly Becker, Michael Healey, David Stein, Mahyar Heydarpour, Lisa S Rotenstein. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Trends in Venture Capital Investment in AI-Driven Biopharmaceutical Startups

Abhishek Bazaz[1], BS; Yunan Ji[2], PhD; Mariana P Socal[3,4], MD, PhD; So-Yeon Kang[5], MBA, PhD

[1]School of Medicine, Johns Hopkins University, Baltimore, MD, United States
[2]McDonough School of Business, Georgetown University, Washington, DC, United States
[3]Health Policy and Management, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, United States
[4]Carey Business School, Johns Hopkins University, Baltimore, MD, United States
[5]Department of Health Management and Policy, School of Health, Georgetown University, Washington, DC, United States

Corresponding Author:
So-Yeon Kang, MBA, PhD
Department of Health Management and Policy
School of Health
Georgetown University
3700 Reservoir Rd NW
Washington, DC, 20007
United States
Phone: 1 202 687 0100
Email: soyeon.kang@georgetown.edu

Abstract

This study analyzes 2010-2024 venture capital trends in international artificial intelligence–driven biopharmaceutical startups, revealing rapid growth in discovery tool investments and concentrated US funding in California and Massachusetts.

(J Med Internet Res 2026;28:e84968) doi:10.2196/84968

KEYWORDS

venture capital; artificial intelligence; biotechnology; biopharmaceutical; drug discovery; discovery tools; investment trends; health care innovation; startups

Introduction

Artificial intelligence (AI) is increasingly recognized for its potential to reduce costs and improve efficiency across biopharmaceutical research and commercialization [1]. Previous studies have demonstrated increasing venture capital (VC) investment in the biopharmaceutical industry and, separately, in companies that utilize AI technology [2,3]. However, little systematic knowledge exists about which biopharmaceutical sectors attract VC investment for AI-related innovation, the magnitude of these investments, and the geographical distribution of funding. A clearer understanding of this investment landscape can inform biomedical researchers and entrepreneurs seeking funding for their AI-related ventures as well as policymakers designing financial incentives [2]. We examine recent trends in VC investment into AI-driven biopharmaceutical startups.

Methods

We used proprietary investment data from PitchBook, a leading third-party investment data provider that tracks VC deals across countries [4,5]. From January 2010 to December 2024, we identified VC-funded biopharmaceutical deals and extracted size, date, and financing type. We classified companies by using PitchBook's primary industry designation and supplemented this with subindustry categories (eg, drug discovery) to improve specificity. This approach has been widely adopted in prior peer-reviewed research on VC investment in health care and biopharmaceuticals [2,6,7]. All capital data were adjusted for inflation by using the Consumer Price Index and expressed in 2024 US dollars. Companies were classified as AI-related if artificial intelligence or AI were disclosed in the company's area of strategic business focus in PitchBook and non-AI otherwise. Companies were further stratified by headquarter location and primary business area according to Pitchbook: (1) biotechnology companies (developing specific therapeutics or biologics, eg, CardioGen Sciences); (2) drug discovery

XSL•FO
RenderX

(computational or experimental platforms to identify new therapeutics for in-house development, eg, Aquemia); (3) drug delivery (technology for administering drugs in a clinical setting, eg, Particle Therapeutics); and (4) discovery tools (enabling technologies, devices, or software for drug discovery by other companies, eg, Carterra) [5]. The median (IQR) was calculated for deal counts and total capital invested. The compound annual growth rate was calculated for capital investment, and chi-square analysis was conducted for company industry distribution. Analyses were conducted using RStudio (version 2023.12.0; Posit). Additional methodological details are provided in Multimedia Appendix 1.

Results

We identified 28,269 VC deals involving biopharmaceutical companies between January 1, 2010, and December 31, 2024. Of these, 1679 (5.93%) deals were associated with companies disclosing AI as a business focus. Among AI-related deals, drug discovery was the most common business focus (639/1679, 38.06% of the deals, median investment size US $9 million, IQR 3.00-30.00 million), followed by biotechnology/pharmaceutical (625/1679, 37.22%, median US $3.51 million, IQR 1.20-14.10 million), discovery tools (401/1679, 23.88%, median US $4.64 million, IQR 1.39-15.00 million), and drug delivery (14/1679, 0.83%, median US $5.23 million, IQR 4.38-11.63 million). The industry composition of VC-funded companies differed significantly between AI-related and non-AI companies. In particular, AI-related companies had a substantially higher share in discovery tools (401/1679, 23.88% vs 879/26,590, 3.31%) and minimal involvement in drug delivery (14/1679, 0.83% vs 1093/26,590, 4.11%) compared with non-AI companies, respectively (P<.001, Figure S1 in Multimedia Appendix 1). The increase in the number of VC deals across both AI and non-AI sectors was most reflected in early-stage deals. In 2024, 77.42% (96/124) of AI and 68.21% (723/1060) of non-AI VC deals were early-stage. Total capital demonstrated more equal growth trends across deal stages (Figure S2 in Multimedia Appendix 1).

Figure 1 depicts the annual deal count for biopharmaceutical startups. AI companies, both overall and within specific industries, demonstrated greater growth in VC compared with non-AI companies based on compound annual growth rate calculations (Table S1 in Multimedia Appendix 1). Among AI companies, discovery tool companies experienced the largest growth, with the median deal size of AI discovery tool companies increasing from US $0.19 million to US $7.50 million between 2010 and 2024. Their share of AI-related capital invested increased from 0.25% (0.19/76.86) to 21.85% (906.15/4147.92) from 2010 to 2024. For non-AI investments, discovery tools slightly increased from 0.87% (59.25/6816.51) to 1.58% (573.03/36,233.1) (Figure S3 in Multimedia Appendix 1).

Figure 2 illustrates the geographic distribution of VC deals in the United States in 2024. Approximately 50.74% (852/1679) of AI-related and 46.07% (12,249/26,590) of non-AI–related deals were made for US-based companies. Nearly 60% (65/110) of US-based AI deals in 2024 were in California and Massachusetts, while non-AI companies spanned 42 states. Median AI-related deal sizes varied regionally—US $10.76 million (IQR 4-39.85 million) in California compared to US $2.94 million (IQR 1.00-14.70 million) in Connecticut.

XSL•FO
RenderX

Figure 1. Trends from 2010 to 2024 in deal counts of venture capital deals by primary investment area: companies with artificial intelligence (AI) vs without AI. (A) AI deals and (B) non-AI deals.

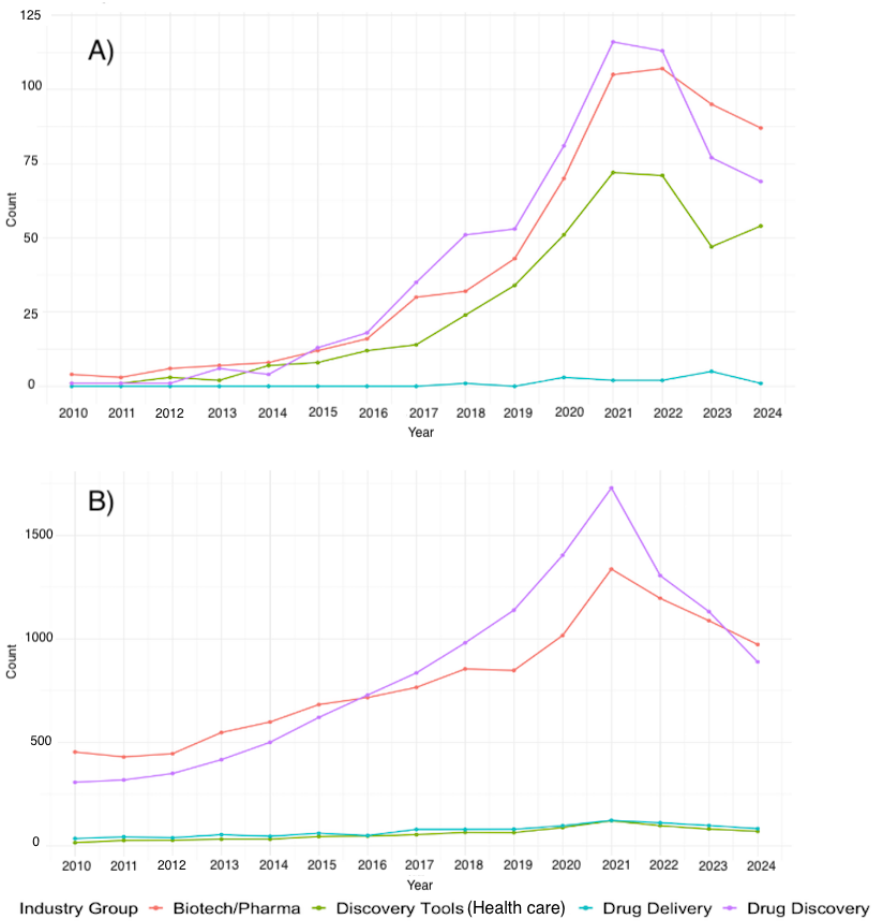

Figure 2. Geographic distribution of venture capital deal count involving biopharmaceutical companies in 2024. (A) AI deals and (B) non-AI deals. AI: artificial intelligence.

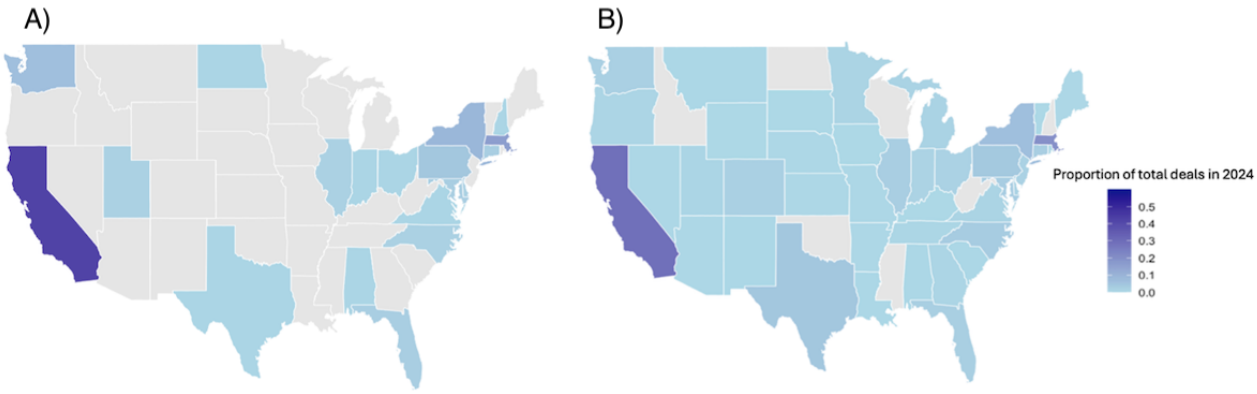

Discussion

While AI-related biopharmaceutical companies represent a minority of overall deals, they have experienced substantial growth, especially in discovery tools, where funding surged nearly 36-fold from 2010 to 2024. The majority of this growth occurred during and after the COVID-19 pandemic, in which VC funding experienced significant volatility [8]. AI-related VC activity expanded to 19 states by 2024, though investments remained highly concentrated in California and Massachusetts. These trends suggest that AI is an emerging investment focus but may be influenced by regional innovation ecosystems and biopharmaceutical research resources. Policymakers and state governments seeking to attract VC investment should consider emulating the frameworks used in these states, including research and development incentives and efficient regulatory pathways for AI-driven biotechnology ventures [9]. Future research evaluating specific regulatory and innovation policies that influence the diffusion of AI-driven health care companies and investment flows as well as the impact of AI on such

XSL•FO
RenderX

innovation warrants consideration. Limitations of this study include its reliance on proprietary data based on publicly disclosed deals and self-reported business profiles. Companies that do not publicly report their financing activities or strategic focus may be missing from our dataset. Additionally, firms that do not explicitly identify AI as a core business area may be excluded, while some companies that use the term as a marketing tool ("AI-washing") are included [10]. However, our classification based on companies' strategically disclosed business focus on AI is more conservative than methods that include companies using AI only in their internal operations when estimating VC flows into AI-driven companies.

Funding

This work was supported in part by Arnold Ventures. The funder had no role in the collection of the data, analysis, interpretation, or reporting of the data or in the decision to submit the manuscript for publication.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to their proprietary nature and restrictions under a data use agreement. However, the analyses can be reproduced by researchers with access to the same subscription-based data sources.

Authors' Contributions

AB contributed to formal analysis, methodology, visualization, writing of the original draft, and reviewing/editing. YJ and MPS contributed to conceptualization, supervision, and reviewing/editing. SYK contributed to conceptualization, data curation, funding acquisition, project administration, supervision, writing of the original draft, and reviewing/editing.

Conflicts of Interest

SYK received research funding from Arnold Ventures, the National Institute for Health Care Management Foundation, and the American Society of Clinical Oncology and consulting fees from Colorado Consumer Health Initiative. YJ received research funding from the National Institute on Aging. MPS receives research grants to Johns Hopkins Bloomberg School of Public Health from Arnold Ventures, the California Department of Health Care Access and Information, the ERISA Industry Committee, and the US Department of Defense.

Multimedia Appendix 1
Details of the methodology and results.
[DOCX File , 915 KB - jmir_v28i1e84968_app1.docx]

References

1. Leite M, de Loiola Costa LS, Cunha V, Kreniski V, de Oliveira Braga Filho M, da Cunha NB, et al. Artificial intelligence and the future of life sciences. Drug Discov Today 2021 Nov;26(11):2515-2526. [doi: 10.1016/j.drudis.2021.07.002] [Medline: 34245910]
2. Kang S, Liu M, Ballreich J, Gupta R, Anderson G. Biopharmaceutical pipeline funded by venture capital firms, 2014 to 2024. Health Aff Sch 2024 Oct;2(10):qxae124. [doi: 10.1093/haschl/qxae124] [Medline: 39411456]
3. Tricot R. Venture capital investments in artificial intelligence. OECD. 2021 Sep. URL: https://www.oecd.org/en/publications/venture-capital-investments-in-artificial-intelligence_f97beae7-en.html [accessed 2025-12-01]
4. Rathi V, Murr A, Feng A, Tauscher J, Naunheim M, Kozin E, et al. Analysis of venture capital investment in therapeutic otolaryngologic devices, 2008-2017. JAMA Otolaryngol Head Neck Surg 2019 Apr 01;145(4):387-389 [FREE Full text] [doi: 10.1001/jamaoto.2019.0045] [Medline: 30816918]
5. PitchBook. URL: https://pitchbook.com [accessed 2025-11-25]
6. Ji Y, Kang S. Global venture capital flows and US health care innovation. Health Aff Sch 2025 Nov;3(11):qxaf206. [doi: 10.1093/haschl/qxaf206] [Medline: 41234683]
7. Kang S, Lee B, Gupta R, Ballreich J, Anderson G. Research and development financing models for new biopharmaceuticals in the United States. Health Aff Sch 2025 Aug;3(8):qxaf161 [FREE Full text] [doi: 10.1093/haschl/qxaf161] [Medline: 40896379]
8. Bellucci A, Borisov A, Gucciardi G, Zazzaro A. The reallocation effects of COVID-19: evidence from venture capital investments around the world. J Bank Financ 2023 Feb;147:106443 [FREE Full text] [doi: 10.1016/j.jbankfin.2022.106443] [Medline: 35221458]
9. Moretti E, Wilson D. State incentives for innovation, star scientists and jobs: evidence from biotech. Journal of Urban Economics 2014 Jan;79:20-38 [FREE Full text] [doi: 10.1016/j.jue.2013.07.002]

XSL•FO
RenderX

10. Simonian J. AI washing: signs, symptoms, and suggested solutions for investment stakeholders. CFA Institute Research and Policy Institute. URL: https://rpc.cfainstitute.org/sites/default/files/docs/research-reports/simonian_ai_washing_report_online.pdf [accessed 2025-12-01]

Abbreviations

AI: artificial intelligence
VC: venture capital

Edited by A Coristine; submitted 28.Sep.2025; peer-reviewed by J Megerian, VV Sangaraju, T Yusuff, A Gopinath, F Amakye; comments to author 04.Nov.2025; revised version received 20.Jan.2026; accepted 21.Jan.2026; published 25.Feb.2026.

Please cite as:
Bazaz A, Ji Y, Socal MP, Kang SY
Trends in Venture Capital Investment in AI-Driven Biopharmaceutical Startups
J Med Internet Res 2026;28:e84968
URL: https://www.jmir.org/2026/1/e84968
doi:10.2196/84968
PMID:

©Abhishek Bazaz, Yunan Ji, Mariana P Socal, So-Yeon Kang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

The Impact of Ontario's Virtual Care Payment Model on Cancer Care: A Natural Policy Experiment

Adam Suleman[1], MD; Zhihui Amy Liu[2], PhD; Anna Sinaiko[3], PhD; Anca Prica[2], MD, MEd; Danielle Rodin[2], MD, MPH

[1]Oakville-Trafalgar Memorial Hospital, Oakville, ON, Canada

[2]Princess Margaret Cancer Centre, Toronto, ON, Canada

[3]Department of Health Policy and Management, Harvard University, Boston, MA, United States

Corresponding Author:
Danielle Rodin, MD, MPH
Princess Margaret Cancer Centre
700 University Avenue, Room 7-513
Toronto, ON, M5G 1M9
Canada
Phone: 1 (416) 946 6513
Fax: 1 (416) 946 6566
Email: danielle.rodin@uhn.ca

(J Med Internet Res 2026;28:e89151) doi:10.2196/89151

KEYWORDS

virtual care; cancer; time series; policy; payment model

Introduction

Virtual care for oncology patients via phone or video visits significantly increased during the COVID-19 pandemic to support ongoing care delivery [1-3]. In Ontario, this was facilitated by changes in the publicly funded Ontario Health Insurance Program, through which most physicians are remunerated. From March 14, 2020, to November 30, 2022, all virtual visits were payable at the same rate as in-person care [4]. On December 1, 2022, payment for new patient consultations by phone was reduced to CAD $15 and phone follow-up visits payable at 85% of the equivalent in-person rate. We evaluated the association between the December 2022 change in payment and the use of phone visits for care delivery.

Methods

We obtained data on all ambulatory oncology clinic visits from Princess Margaret Cancer Center (Multimedia Appendix 1), Canada's largest comprehensive cancer center, delivered between June 1 and November 30, 2022 (period 1: prior to policy change or "interruption"), and between December 1, 2022, and May 31, 2023 (period 2: after policy change).

An interrupted time series (ITS) analysis was conducted, modeling the total number of weekly visits using a linear regression model. An outlier holiday week of December 24-30, 2022, was removed from the analysis. Model parameters included the intercept, baseline trend (preinterruption slope), step change after policy change, and trend change after policy implementation. Autocorrelation and partial autocorrelation plots and the Durbin-Watson statistic were used to examine temporal autocorrelation in the model residuals. To evaluate whether modality-specific changes might reflect broader utilization trends, individual ITS analyses for total, in-person, phone, and video visits were performed. Subgroup analyses were performed for individual clinic types, including for hematologic malignancies, radiation therapy, solid tumor, and supportive care clinics. A 2-sided P value <.05 was considered statistically significant. This study was approved by the University Health Network research ethics board (REB 23-5510).

Results

There were 144,139 visits during the study (75,849 during period 1; 68,290 during period 2), of which 13% (9860/75,849) and 7% (4780/68,290) were by phone in periods 1 and 2, respectively. Video visits comprised 4% (3034/75,849) of the visits in period 1 and 12% (8195/68,290) in period 2. Visits were to outpatient clinics for solid tumors (108,620/144,139, 75%), hematologic malignancies (22,301/144,139, 15%), and supportive care (4874/144,139, 3%), and on-treatment visits for radiation therapy (8344/144,139, 6%).

Before the policy change, phone visits were increasing by 5.8 visits per week. At the time of the policy change, there was an immediate reduction of 209.4 phone visits per week (95% CI –285.3 to –133.6; P<.001). After the policy change, the weekly trend also reversed, with a slope change of –9.1 visits per week

XSL•FO
RenderX

(95% CI –14.2 to –4; *P*<.001), corresponding to an overall post-policy decline of 3.3 phone visits per week (Figure 1A).

Figure 1. Phone visits per week (panel A), video visits per week (panel B), in-person visits per week (panel C), and total visits per week (panel D) for patients with an oncology visit from June 2022 to May 2023. x: Christmas holiday week excluded from the analysis.

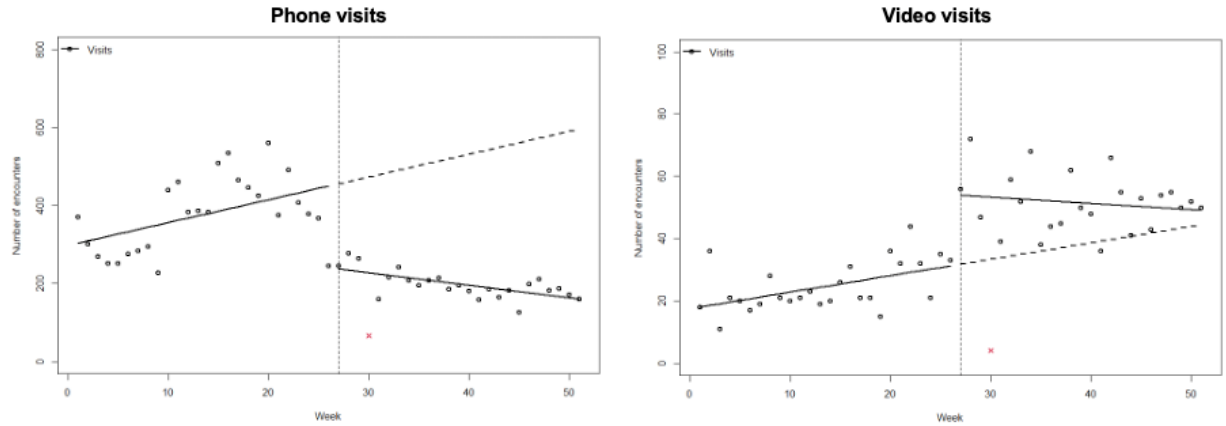

A

Phone visits	Coefficient	95% CI	*P* value
(Intercept)	297.5	244.6 to 350.5	<.001
Baseline trend	5.8	2.4 to 9.3	.001
Step change after policy change	-209.4	-285.3 to -133.6	<.001
Trend change after policy change	-9.1	-14.2 to -4	<.001

B

Video visits	Coefficient	95% CI	*P* value
(Intercept)	17.5	10.8 to 24.2	<.001
Baseline trend	0.5	0.1 to 1	.02
Step change after policy change	22.9	13.3 to 32.5	<.001
Trend change after policy change	-0.7	-1.4 to -0.1	.03

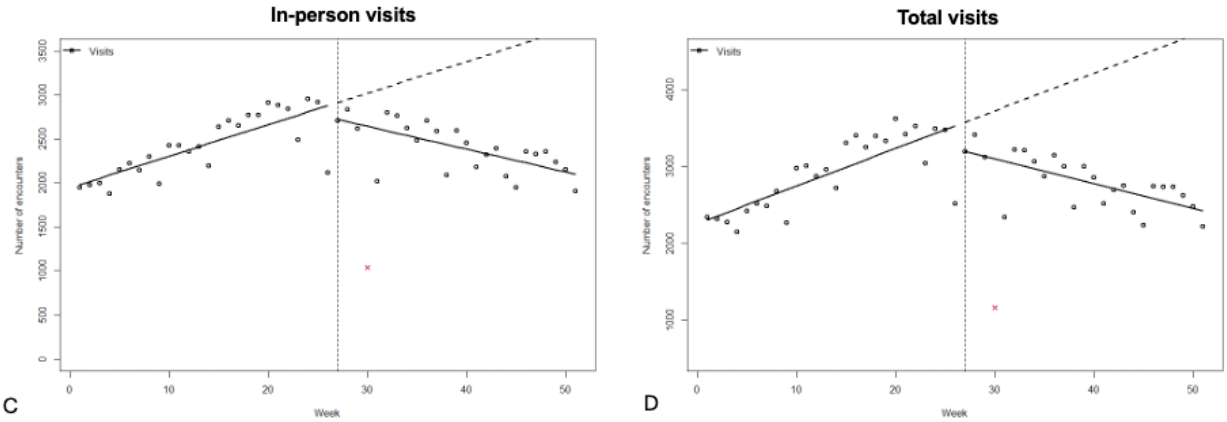

C

In person visits	Coefficient	95% CI	*P* value
(Intercept)	1945.5	1770.9 to 2120.1	<.001
Baseline trend	35.9	24.6 to 47.2	<.001
Step change after policy change	-129.5	-379.4 to 120.3	.30
Trend change after policy change	-61.8	-78.6 to -45	<.001

D

Total visits	Coefficient	95% CI	*P* value
(Intercept)	2258	2033.5 to 2482.5	<.001
Baseline trend	48.8	34.3 to 63.4	<.001
Step change after policy change	-301.9	-623.2 to 19.5	.07
Trend change after policy change	-80.9	-102.5 to -59.3	<.001

This decreasing trend was observed for all clinic types (Figure 2). In contrast, the number of weekly video visits immediately increased by 22.9 (95% CI 13.3-32.5; *P*<.001) at the time of the policy change compared to a baseline increasing trend of 0.5 weekly visits. After the policy change, the weekly trend reversed, with a slope change of 0.7 (95% CI –1.4 to –0.1; *P*=.03) (Figure 1B).

XSL•FO
RenderX

Figure 2. Phone visits per week by clinic type for patients with an oncology visit from June 2022 to May 2023. x: Christmas holiday week excluded from the analysis.

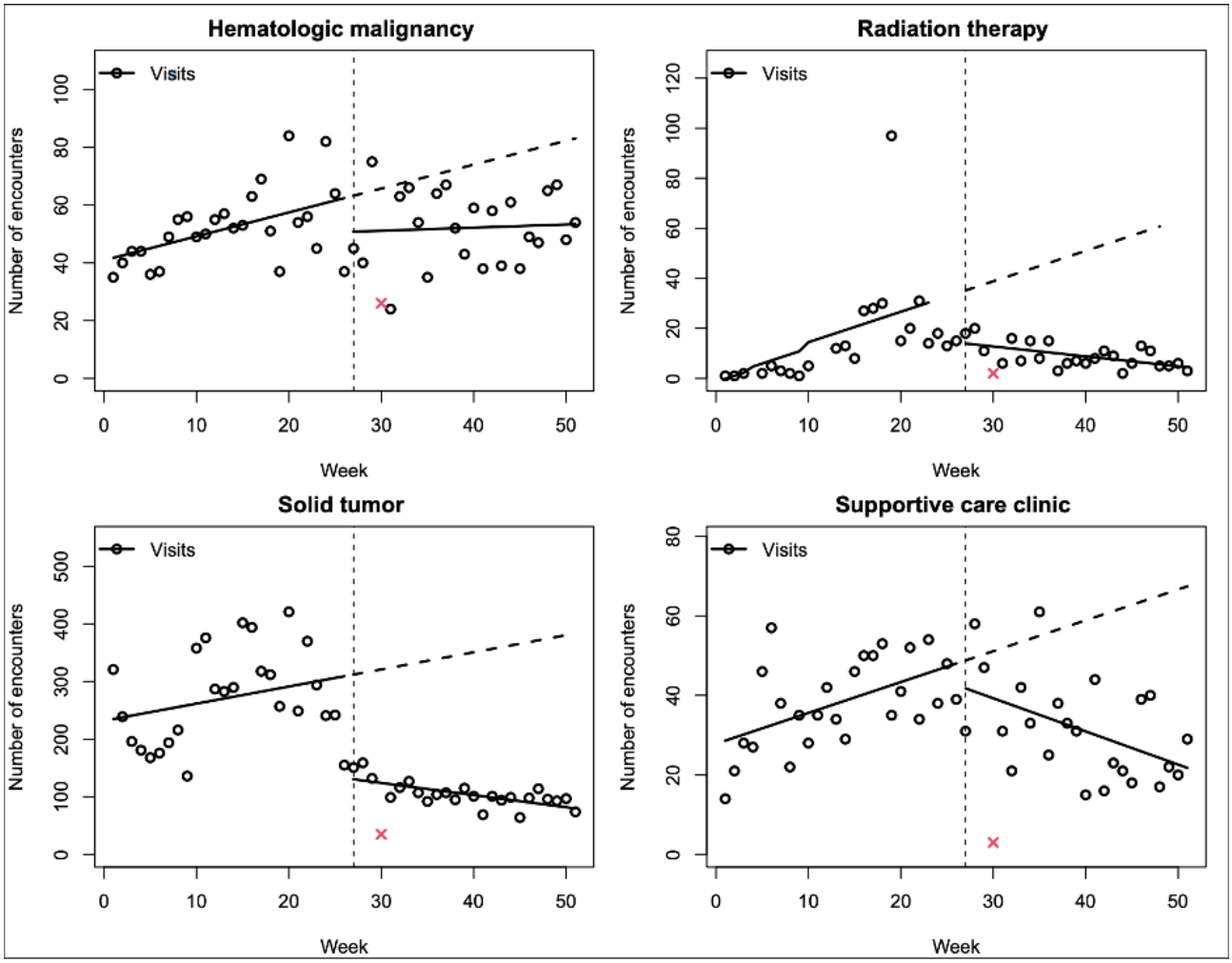

Unlike for virtual care, there was no significant immediate change in the number of weekly in-person (Figure 1C) or total visits (Figure 1D). After the policy change, there was a decreasing trend, consistent with the reduction observed in virtual care, with a slope change of –61.8 (95% CI –78.6 to –45; P<.001), corresponding to 25.9 fewer in-person visits per week in period 2 compared to those in period 1 and an overall reduction of 32.1 total visits per week between the two periods.

Discussion

An immediate shift in visit modality was observed following the payment policy change, with a sharp drop in phone visits and a sharp increase in video visits, with no change in visits in-person. In the 6 months following, there was a general trend toward declining virtual care, reversing the initial increase in video visits following the fee code revision, along with a decline in in-person visits, suggesting an inflection point in the modality used for health care delivery at the time of the policy change. The overall reduction in visits over time suggests the contribution of broader temporal changes in care delivery.

There is ongoing tension in defining the optimal use of virtual care. Virtual care offers enhanced access to timely support and symptom management for patients with cancer, reducing the need for frequent travel during physically and emotionally challenging treatment periods [5]. In fee-for-service environments, however, financial incentives have been shown to impact physician behavior [6,7], and there is concern that physicians may feel incentivized to favor brief phone encounters over in-person visits, increasing the frequency of low-value visits [8]. In the absence of other policies or operational changes that may have coincided with the fee change, findings from this natural policy experiment suggest that remuneration may have played a role in the immediate shift in visit modality. However, unmeasured concurrent changes in clinician behavior or patient preferences could still contribute to residual temporal confounding.

Limitations of this study include the lack of clinical or sociodemographic information, which may impact patient or physician preference for virtual care or clinic- or hospital-level factors that may impact physicians' ability to schedule virtual visits. This study was performed in a specialized academic center where virtual care was widely offered throughout the pandemic [1,3], and the findings may not be applicable to other settings that had not routinely adopted virtual care during this period. Further, we cannot attribute causality from the policy change to the observed findings, and the gradual overall decline in visits suggest other unmeasured concurrent changes in care delivery or preferences.

XSL•FO
RenderX

The option of virtual care has been valued by many patients with cancer [9,10]; thus, policymakers should evaluate whether the current fee structure is contributing to reduced virtual care access and reduced agency in choosing virtual care when appropriate. Further studies are needed to understand how policy changes, along with patient and provider preferences, impact the modality of care delivery for patients with cancer.

Funding

The authors declared no financial support was received for this work.

Conflicts of Interest

A Suleman declares honoraria from BeOne Medicines, Astra Zeneca, Roche, and Incyte. AP declares honoraria from AbbVie, Kite-Gilead, and Astra Zeneca.

Multimedia Appendix 1
Number of ambulatory visits to Princess Margaret Cancer Centre by appointment type (in person or virtual) and clinic type during the study period.
[DOCX File , 15 KB - jmir_v28i1e89151_app1.docx]

References

1. Berlin A, Lovas M, Truong T, Melwani S, Liu J, Liu ZA, et al. Implementation and outcomes of virtual care across a tertiary cancer center during COVID-19. JAMA Oncol 2021 Apr 01;7(4):597-602 [FREE Full text] [doi: 10.1001/jamaoncol.2020.6982] [Medline: 33410867]
2. Shaffer KM, Turner KL, Siwik C, Gonzalez BD, Upasani R, Glazer JV, et al. Digital health and telehealth in cancer care: a scoping review of reviews. Lancet Digit Health 2023 May;5(5):e316-e327 [FREE Full text] [doi: 10.1016/S2589-7500(23)00049-3] [Medline: 37100545]
3. Suleman A, Vijenthira A, Liu Z, Truong T, Berlin A, Prica A, et al. Virtual care during the COVID-19 pandemic for patients with hematologic malignancies: a single-institution experience. JCO Oncology Practice 2023 May;19(5):e672-e682. [doi: 10.1200/op.22.00690]
4. Bulletin 11229 - primary care changes in response to corona virus (COVID-19) effective March 14, 2020. OHIP INFOBulletins 2020. URL: http://www.ontario.ca/document/ohip-infobulletins-2020/bulletin-11229-primary-care-changes-response-corona-virus-covid-19 [accessed 2025-05-07]
5. Suleman A, Vijenthira A, Berlin A, Prica A, Rodin D. The use of virtual care in patients with hematologic malignancies: a scoping review. Curr Oncol 2022 Feb 07;29(2):892-900 [FREE Full text] [doi: 10.3390/curroncol29020076] [Medline: 35200575]
6. O'Neil B, Graves AJ, Barocas DA, Chang SS, Penson DF, Resnick MJ. Doing more for more: unintended consequences of financial incentives for oncology specialty care. J Natl Cancer Inst 2016 Feb;108(2):djv331 [FREE Full text] [doi: 10.1093/jnci/djv331] [Medline: 26582063]
7. Green BB, Anderson ML, Cook AJ, Chubak J, Fuller S, Kimbel KJ, et al. Financial incentives to increase colorectal cancer screening uptake and decrease disparities: a randomized clinical trial. JAMA Netw Open 2019 Jul 03;2(7):e196570 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.6570] [Medline: 31276178]
8. Llorian ER, Mason G. Healthcare utilization and telemedicine: an evaluation using linked administrative data from Manitoba. J Telemed Telecare 2021 Jan 17;29(4):282-290. [doi: 10.1177/1357633x20981227]
9. Scruton S, Wong G, Babinski S, Squires LR, Berlin A, Easley J, et al. Optimizing virtual follow-up care: realist evaluation of experiences and perspectives of patients with breast and prostate cancer. J Med Internet Res 2025 Jan 03;27:e65148 [FREE Full text] [doi: 10.2196/65148] [Medline: 39752659]
10. Cuaron JJ, McBride S, Chino F, Parikh D, Kollmeier M, Pastrana G, et al. Patient safety and satisfaction with fully remote management of radiation oncology care. JAMA Netw Open 2024 Jun 03;7(6):e2416570 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.16570] [Medline: 38865123]

Abbreviations

ITS: interrupted time series

XSL•FO
RenderX

Edited by M Balcarras; submitted 07.Dec.2025; peer-reviewed by K Syrett, J Guan, K Morrison; comments to author 04.Mar.2026; revised version received 11.Mar.2026; accepted 12.Mar.2026; published 20.Apr.2026.

Please cite as:
Suleman A, Liu ZA, Sinaiko A, Prica A, Rodin D
The Impact of Ontario's Virtual Care Payment Model on Cancer Care: A Natural Policy Experiment
J Med Internet Res 2026;28:e89151
URL: https://www.jmir.org/2026/1/e89151
doi:10.2196/89151

©Adam Suleman, Zhihui Amy Liu, Anna Sinaiko, Anca Prica, Danielle Rodin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Acceptability, Feasibility, and Perceived Effectiveness of Video-Based Patient Records for Supporting Care Delivery to Older Adults With Frailty: Nonrandomized Mixed Methods Pilot Study

Phoebe Averill[1,2], PhD; Rachael Lear[3], RN, PhD; Ricky Odedra[1,4], MSc; Susannah Long[1,5], MBBS, PhD; Alex Taylor[1], BA; Pi-Jung Charville[5], RN, MSc; Jessica Fernandes[5], RN, BSc; Uzoamaka Ekeogu[5], RN, MSc; Jessica Leombruno[5], MSc; Sophia Ellis[5], MBChB, MRCP; Erik Mayer[1,4,5], MBBS, PhD

[1]NIHR North West London Patient Safety Research Collaboration, Institute of Global Health Innovation, Imperial College London, St Mary's Hospital, London, United Kingdom

[2]Better Health & Care Hub, King's College London, London, United Kingdom

[3]Robin Hood Lane Health Centre, Sutton, United Kingdom

[4]iCARE Secure Data Environment, NIHR Imperial Biomedical Research Centre, Imperial College Healthcare NHS Trust, London, United Kingdom

[5]Imperial College Healthcare NHS Trust, London, United Kingdom

Corresponding Author:
Phoebe Averill, PhD
NIHR North West London Patient Safety Research Collaboration, Institute of Global Health Innovation, Imperial College London, St Mary's Hospital, London, United Kingdom

Abstract

Background: Frailty constitutes a growing challenge for health and social care systems around the world. In England, 35% of adults aged 65 years and older live with frailty, with international estimates indicating that almost half of all hospital inpatients within the same age group are frail. This population often experiences multimorbidity and frequent care transitions. Written documentation and verbal handovers may lack the precision and nuance required to understand an older adult's presentation and support needs. Video recordings of individual patients, capturing aspects of their functional abilities and condition, may help to enhance multidisciplinary team communication and care continuity, yet little is known about their use in the care of older inpatients with frailty.

Objective: We aimed to evaluate the acceptability, feasibility of implementation, and perceived effectiveness of video-based patient records (the Isla Health Digital Pathway Platform) for supporting the assessment and care of older inpatients with frailty within the acute hospital setting.

Methods: A nonrandomized mixed methods pilot study was conducted within 3 acute medicine wards for older adults. The video-based patient records intervention, permitting videos to be embedded securely within the electronic patient record, was implemented over a 3-month period alongside usual care. Patient enrollment and retention figures; qualitative interviews with patients, carers, and clinical staff; and video capture and view metrics were used to address the study objectives. The Theoretical Framework of Acceptability of Healthcare Interventions was applied to the framework analysis of interview data, capturing concepts such as intervention ethicality, burden, and coherence. Patient and public involvement and engagement informed each research stage.

Results: Twenty-nine patients were enrolled (56.9%); 1 patient withdrew before receiving the intervention. Modal reasons given by patients for nonparticipation included not wanting to take part in research (n=8) or feeling too unwell (n=2). Staff identified multiple opportunities for capturing patient videos, including documentation of mobility assessments or seizures. The intervention was considered acceptable on the grounds that safeguards were always in place, including secure data storage and upholding of patient dignity. Implementation barriers and facilitators were identified; factors such as difficulties in capturing videos within busy ward environments and scheduling issues were voiced by participants. Video view metrics and data from interviews collectively suggested low rates of engagement with videos by clinical staff once captured. Potential intervention impacts included perceived enhancements to clinical assessment and person-centered care.

Conclusions: Our findings suggest that the intervention is largely acceptable to patients, carers, and clinical staff. Conclusions as to intervention feasibility were mixed, with limited engagement with videos suggesting further work is required to promote sufficient uptake among staff. Finally, this research presents promising patient, carer, and clinical opinion as to the potential effectiveness of video-based patient records for improving aspects of patient care.

XSL•FO
RenderX

Trial Registration: ClinicalTrials.gov NCT06504641; https://clinicaltrials.gov/study/NCT06504641

(J Med Internet Res 2026;28:e77318) doi:10.2196/77318

KEYWORDS

frailty; video recording; digital health; patient safety; quality of care

Introduction

Background

Frailty can be understood as impairment in a person's ability to recover following an injury or illness [1]. It is commonly characterized by exhaustion upon low energy expenditure, diminished muscle strength, and unplanned weight loss [2], and represents a pertinent challenge for health and social care systems globally in the context of population aging [3]. Over one-third (35%) of adults in England aged ≥65 years are thought to live with frailty [4], with a synthesis of international literature indicating a pooled prevalence rate of 47% for frailty among hospital inpatients in the same age group [5].

Older adults with frailty typically have complex care needs, experience multimorbidity, and require input from several care providers spanning multiple, fragmented health and social care organizations [6,7]. Evidence suggests that such patients are particularly vulnerable to experiencing shortfalls in safety and quality of care, owing to poor care continuity [8,9]. This can contribute to avoidable deconditioning in older adults [10], who may struggle to return to their usual activities and routines upon discharge from the hospital to community-based care.

The World Health Organization acknowledges the inherent risks to patient safety at transitions of care, where treatment delays, duplication of tests, and increased rates of adverse events are among identified challenges [11]. Frequent transitions between health and social care settings are commonplace in the care of frail older people [9]. Therefore, effective communication within and between services involved in the care of these patients is imperative [12]; providers require timely access to comprehensive, accurate, and up-to-date information about a person's presentation, in a format that readily supports meaningful interpretation [13].

Written documentation and verbal handovers vary in quality and may fail to convey the complexities of an older adult's condition, including functional capabilities and support needs. A growing body of literature suggests that the use of digital multimedia, such as photographs and videos, may help to enhance communication in patient care. A Cochrane review indicated that transmission of digital images between care providers may reduce time to commence treatment once a patient has presented to a service [14]. Moreover, a scoping review reported on the value of patient-generated photos and videos in helping professionals to obtain a more holistic understanding of patients, supporting diagnostic processes and monitoring of treatment outcomes, for example, documenting the healing of postoperative wounds [15]. Our recent systematic review synthesized empirical research evidence and professional and regulatory guidance on video recording patients for direct care purposes [16]. Review findings suggest that video recording patients to support care delivery is largely acceptable to patients, where privacy and data security risks are suitably mitigated [16].

Video recordings may support improvements in the safety and quality of care for older adults with frailty, in providing detailed visual information about a patient's functional presentation, support needs, and care preferences. Nevertheless, ethical considerations warrant detailed examination prior to the routine use of video recordings within patient care [16]. Previously, a lack of secure and efficient ways to acquire, store, and share photographs and videos may in part explain why multimedia data have not been fully exploited in health care. Further research is required to explore the acceptability, feasibility, and potential value of video recordings that are securely embedded within the electronic patient record (EPR) for use in the care of older people living with frailty.

Objectives

The primary objective of this study was (1) to determine the acceptability of using video recordings to capture the functional abilities, support needs, and preferences of older inpatients with frailty, to support clinical assessment and care delivery. Secondary objectives were (2) to explore the feasibility of implementing video-based patient records (the Isla Health Digital Pathway Platform) within the acute medical ward setting for older adults, and (3) to appraise the perceived effectiveness of patient videos for supporting care in this context. Study findings will be used to inform the decision-making about progression to a definitive trial, examining the effectiveness of patient video recordings for improving the quality and safety of care transitions.

Methods

Design

This study comprised a single-center, nonrandomized, mixed methods pilot study with an embedded process evaluation (ClinicalTrials.gov ID: NCT06504641) to inform decision-making about progression to a definitive trial to evaluate clinical impact. Where applicable, in line with guidance for nonrandomized pilot and feasibility research [17], the study reporting adheres to the CONSORT (Consolidated Standards of Reporting Trials) extension to pilot and feasibility trials [18], with intervention description informed by the Template for Intervention Description and Replication (TIDieR) guidelines [19].

Setting and Participants

Participants were recruited from 3 acute medical wards for older adults within a large acute National Health Service (NHS) Trust in London, United Kingdom. We aimed to recruit 30 patient-carer dyads and 35 clinical staff members from the

XSL•FO
RenderX

wards, according to the eligibility criteria defined in Table 1. Carers consisted of family members or friends who provided the patient with unpaid, informal care. Our sampling approach was designed to permit appraisal of acceptability and feasibility in principle, through a pragmatic yet purposively diverse sample of patients, carers, and ward staff.

Table . Overview of study eligibility criteria for the nonrandomized mixed methods pilot study involving older adult inpatients with frailty.[a]

Participants	Inclusion criteria	Exclusion criteria
Patients	• Inpatient within acute medical wards for older adults during the 3-month intervention pilot phase • Aged >65 years old • Assessed as frail or prefrail by the direct care team • Capacity to consent to participate OR lacked capacity to consent and a "personal consultee" was available to advise on patient's likely wishes about participating	• Lacked capacity to consent AND a "personal consultee" was not available to advise on the patients' likely wishes about participating
Carers	• Aged >18 years old • Provided unpaid assistance to the patient for daily activities	• Carers were excluded if the patient declined study participation
Ward staff	• Clinical staff working on 1 of 3 participating acute medical wards for older adults • Working regular shifts during the study initiation and 3-month pilot phases	• Ad hoc bank or agency staff members • Permanent staff members who are on long-term leave (eg, sickness and parental leave) during the 3-month pilot phase.

[a]Carers included family members and friends who provided the patient with unpaid, informal care.

Intervention

The Isla for Frailty intervention provided a video-based patient record function via the web-based Isla Health Digital Pathway Platform, which interfaces with EPR systems such as Cerner. Isla Health is a technology company delivering a digital pathway platform for health care staff to support patients throughout their care journey. One of its features is a visual patient record, which enables health care staff to securely capture and review videos as part of a patient's care. Video data are securely stored within encrypted cloud storage; health care staff can view these data within the EPR system, or directly via the Isla platform, using a secure weblink requiring the user's NHS email address and password. To capture visual data using the Isla platform, the user must be logged in on a mobile device (eg, smartphone and tablet) with an in-built camera and Wi-Fi connection. The Isla Health Digital Pathway Platform is approved by NHS Digital and satisfies NHS data security and protection requirements. Wards were supplied with a designated tablet device to use for video capture.

Video recordings were captured to document aspects of a participating patient's condition or functional ability considered by the direct care team as potentially useful in supporting clinical decision-making, care continuity, or multidisciplinary team (MDT) communication within the patient care journey. For example, videos might be used to capture a patient's mobility and transfers, behavior, or patient preferences. Personal care or toileting was not recorded. Patient videos were viewable by the direct care team during the patient's ward stay. Ward staff were encouraged to view and use videos in a way they felt was useful for supporting patient care, such as to support communication during shift handovers, to inform discharge planning, or in discussions with family members. The study protocol did not dictate when videos should be reviewed; therefore, the study examined the real-world use of patient videos within the acute older adult ward setting.

Participating staff had completed prior mandatory training in Data Security Awareness and were also trained in using the intervention prior to the 3-month pilot. Training emphasized the need to confirm participating patients' consent prior to video capture and to stop recording if a patient showed any signs of distress (verbal or nonverbal) or if another patient entered the shot. Staff were advised to protect patient dignity, ensuring patients were appropriately dressed and drawing curtains around the bed space where required. Contact details for the study clinical lead were shared with all participating staff who wished to discuss any questions or concerns.

Recruitment

Enrollment of Ward Staff

We engaged with staff working on the acute medical wards for older adults over a 2-month period (February 12-April 13, 2024) directly preceding the 3-month pilot phase for the video-based patient record intervention, inviting them to take part in the study. All clinical staff working on the 3 pilot wards received written and verbal information about the study aims, the intervention purpose, permitted uses of the video recordings, and information governance procedures. Staff were reassured that they could take part but opt not to be involved in video recording patients or appearing in patient videos themselves. They were also informed that they could choose not to participate in the evaluation of the video recording intervention, or to decline participation altogether without giving a reason. Staff who wished to take part were given opportunities to ask questions prior to providing their written informed consent.

XSL•FO
RenderX

Enrollment of Patients With Capacity to Consent

Patient screening and enrollment procedures are displayed in Multimedia Appendix 1. Over the 3-month intervention pilot phase (April 15-July 14, 2024), potential participants (inpatients on the 3 participating wards and their carers) were screened for eligibility by the direct care team, who explained the study purpose and provided patients and carers with participant information sheets. The direct care team flagged all potential participants to the researchers. Once the patient was clinically stable, a member of the research team visited the patient on the ward to confirm eligibility and complete enrollment procedures. Patients were offered opportunities to ask questions about the study and asked whether they might be interested in participating. Where required, easy-read participant information sheets were used to support patient understanding about what taking part would involve. Potential participants were offered additional time to decide whether they wished to take part. Where the direct care team reported concerns that a patient may lack capacity to consent to participation, the study team assessed and documented decision-specific capacity, considering whether the patient could understand what participation would involve, retain the study information, weigh this information to make a decision, and communicate their decision.

After consideration, patients who wished to participate and who had the capacity to consent were asked to sign a version of the consent form for patients with capacity. Patients were reassured that they could withdraw at any time. Within the consent process, patients were asked to consider whether they would like to remain in the study, should they lose capacity during the study period (eg, due to delirium onset). Patients were also asked whether they would still like their data to be used in the study. Participant information sheets contained a clear statement regarding retention and use of identifiable data following loss of capacity. Continued participation for such patients was subject to the same protocols for enrolling patients lacking capacity to consent, as detailed in the next section. During the study, the capacity of participants was monitored by the direct care team; the clinical lead communicated with the research team regularly about capacity concerns.

Enrollment of Patients Lacking Capacity to Consent

For patients assessed to lack the capacity to consent to participate, the research team still provided the potential participant with information about the study according to their level of understanding, using visual aids as needed (eg, the easy-read participant information sheets). In line with the requirements of Section 32(3) of the Mental Capacity Act 2005 and the Department of Health's Guidance on nominating a consultee for research involving adults who lack capacity to consent (2008), reasonable steps were taken to identify a personal consultee for these patients [20,21]. Personal consultees were identified by the patient's direct care team and included family members and carers (unpaid), who knew the patient well and who assisted them with their daily activities (eg, decisions about their welfare). Members of the ward staff team could not be nominated as personal consultees due to their own involvement in the study. Patients were excluded where no appropriate person could act as a personal consultee.

Potential personal consultees were approached first by the direct care team, who outlined the study purpose and the role of a consultee, emphasizing their right to decline acting as a consultee for the patient. For potential personal consultees who agreed to be contacted directly by the research team, a member of the research team then provided detailed written and verbal information about the study, including the consultee participant information sheet. The consultee was asked to consider the patient's likely wishes about taking part. The research team made it clear to the consultee that they were not being asked to provide their personal views on study participation, nor to consent to the study on the patient's behalf. Recruitment decisions were made by the research team in line with the consultee's advice and any previous relevant statements or wishes communicated by the patient, whether verbal or nonverbal. Consultee advice was documented on a printed consultee declaration form or via an electronic version of the form. Where a consultee advised that the patient would not have wanted to take part, the researcher abided by this. Patients were also excluded if the consultee declined to offer advice about the patient's likely wishes.

Enrollment of Carers

Carers of participating patients were also invited to take part in the evaluation during the 3-month pilot phase. Where carers were present while a member of the research team was on the ward, they were approached by the researcher accordingly. Otherwise, a ward team member contacted carers to ascertain whether they were happy to be contacted (via email or telephone) by a member of the research team. Carers who wished to participate after reviewing the carer participant information sheet were asked to sign a consent form for carers, via a printed version of the form, or an electronic form sent to the carer via email or text message. For all patients and carers, reasons for nonparticipation were documented within a screening and enrollment log.

Data Collection

Study data sources are summarized in Table 2 and described below.

XSL•FO
RenderX

Table . Overview of outcomes, measures, and data sources for each objective of the nonrandomized mixed methods pilot study involving older adult inpatients with frailty.

Study objective, outcomes, and measures	Data source
Acceptability	
Patient recruitment and retention	
Percentage of eligible participants who were enrolled into the study	Screening and enrollment log
Percentage of eligible participants declining enrollment and reasons for nonparticipation	Screening and enrollment log
Video recording requests by care team	
Number of videos requested, attempted, and submitted	Video tracker
Suitability of patient videos	Video evaluation questionnaire
Desire to see more patient videos in the future	Video evaluation questionnaire
Perceived acceptability	
Patient and carer-reported acceptability	Semistructured interview at or within 2 weeks of discharge
Ward staff team-reported acceptability	Semistructured interview after the 3-month pilot
Feasibility	
Diversity of patient sample	
Patient clinical and demographic characteristics	Screening and enrollment log
Privacy and security concerns	
Number of videos raising cause for concern reported to the clinical lead	Video tracker
Use of the Isla Health Digital Pathway Platform	
Percentage of participants with one or more video linked to EPR[a]	Isla Health Digital Pathway Platform metadata
Reasons for unsuccessful attempts to take a video	Video tracker
Video view metrics	Isla Health Digital Pathway Platform metadata
Intervention barriers and facilitators	
Patient and carer-reported intervention barriers and facilitators	Semistructured interview after the 3-month pilot
Ward staff team-reported intervention barriers and facilitators	Semistructured interview after the 3-month pilot
Perceived effectiveness	
Perceived impacts on: assessment and clinical decision-making; multidisciplinary team communication; care continuity during a hospital stay; person-centered care during a hospital stay	
Potential usefulness of intervention to ward staff team	Video evaluation questionnaire
Ward staff team-reported perspectives on intervention impacts	Semistructured interview after the 3-month pilot
Patient and carer-reported perspectives on intervention impacts	Semistructured interview at or within 2 weeks of discharge

[a]EPR: electronic patient record.

Screening and Enrollment Log

The screening and enrollment log was used to document patient recruitment status, demographic characteristics, and, where applicable, reasons for nonparticipation. A paper version of the log was initially populated. It was then digitized at regular intervals within the 3-month intervention pilot phase and stored in Imperial College Healthcare NHS Trust's iCARE secure data environment.

Video Tracker

A video tracker spreadsheet (Microsoft Excel) was developed to document patient video requests, attempts, submissions, and any issues associated with taking the recordings. The video tracker also captured information about the aspect of a participating patient's care or functional abilities to be video recorded. For example, video foci may include a person's mobility baseline or their support needs when eating and drinking.

Video Evaluation Questionnaires

Staff experience of viewing patient videos was appraised using a brief, anonymous, paper-based video evaluation questionnaire, consisting of closed-ended, multiple-choice questions with space for free-text comments (Multimedia Appendix 2). Participating

XSL•FO
RenderX

staff were asked to complete the questionnaire for each video they viewed. At the end of the 3-month pilot, questionnaire data were digitized (Microsoft Excel) and stored securely within the iCARE secure data environment.

Interviews With Patients and Carers

Patients with the capacity to consent to participation and carers were invited to take part in brief, semistructured interviews at or within 2 weeks of hospital discharge. Consent to take part in an optional interview was provided during the enrollment phase. All patients with the capacity to participate in an interview who had not declined the optional interview component at enrollment were invited to take part as they approached hospital discharge. Likewise, all participating carers were contacted to arrange the optional interview, unless they had previously declined at enrollment. Where the direct care team or research team felt it appropriate, patient capacity was assessed again prior to undertaking the interview, due to the potential for fluctuating capacity in this group. Patients and carers were asked to verbally confirm their consent prior to starting the interview. Interviews were conducted on the ward prior to discharge, or via telephone where patients had already been discharged.

Interviews were conducted by a member of the research team (PA or RO) using a topic guide for patients and carers (Multimedia Appendix 3), seeking to understand patients' experiences of being video recorded. Researchers adapted their language to the patient's level of understanding to enable patients with cognitive impairment to participate. Patient and carer interviews were completed by August 14, 2024, reflecting the date of discharge of the final patient who was eligible to take part in an interview. Interviews were audio-recorded and professionally transcribed. Names and identifiers were replaced with pseudonyms during the transcription process. Transcripts were transferred to and stored within the iCARE secure data environment.

Interviews With Ward Staff

Semistructured interviews with consenting staff members were undertaken within 2 months of the end of the 3-month intervention pilot phase. All participating staff members were approached to take part in an interview, unless they had declined this optional study component at the point of enrollment. Interviews used a version of the topic guide for staff and explored their experiences with the video-based patient records (the Isla Health Digital Pathway Platform) and perceived impacts on patient assessment and clinical decision-making; team communication; and care delivery (Multimedia Appendix 4). Audio-recorded interviews were conducted in a private space on or near the ward, or via telephone according to interviewee preference, with the final interview taking place on October 3, 2024. Written consent for the interviews was documented at recruitment. Names and identifiers were replaced with pseudonyms during the transcription process. Transcripts were transferred to and stored within the iCARE secure data environment.

Isla Health Digital Pathway Platform Metadata

At the end of the 3-month video-based patient record pilot (the Isla Health Digital Pathway Platform), video view metrics were exported from the Isla audit log. The anonymized video view metrics spreadsheet was transferred to and stored within the iCARE secure data environment.

Outcomes

In line with study objectives, the primary outcome we sought to assess was patient, carer, and ward staff team perspectives on the acceptability of the video-based patient records (the Isla Health Digital Pathway Platform). Use of the Isla platform and intervention barriers and facilitators were among secondary outcomes measured to assess the feasibility of implementing video-based patient records within the acute medical inpatient setting for older adults. Finally, 4 outcomes, including perceived impacts on care continuity in the hospital, were appraised to examine the perceived effectiveness of video-based patient records for supporting older adult inpatient care. Outcome measures and data sources for each objective are summarized in Table 2.

Data Analysis

Using Microsoft Excel, descriptive statistics were computed to summarize numerical data as to participant recruitment and retention rates; proportions of patients with videos linked to the EPR; video view metrics; and numbers of videos raising cause for concern. Frequencies and percentages were calculated for categorical variables, with means and SDs presented for continuous variables. Framework Analysis, following the 5-step approach described by Ritchie and Spencer, was applied to qualitative study data [22]. Accordingly, following a phase of familiarization with interview transcripts and free-text video evaluation questionnaire data, key concepts and ideas were noted by 1 author (PA) and discussed iteratively with 2 further authors (RO and AT), to interrogate immediate observations and assumptions about these data. At this stage, the researcher's analytical observations were used together with the Theoretical Framework of Acceptability of Healthcare Interventions by Sekhon to shape an initial thematic framework [23], combining a priori concepts with inductively generated codes. The Theoretical Framework of Acceptability of Healthcare Interventions comprises 7 constructs for appraising intervention acceptability: affective attitude, burden, ethicality, intervention coherence, opportunity costs, perceived effectiveness, and self-efficacy. Component constructs are defined in Multimedia Appendix 5. The derived thematic framework was then applied to all transcripts by 2 authors (PA and RO), with regular collaborative coding meetings convened by coders to appraise consistency and appropriateness in use of the framework. Data were then charted into a framework matrix for review and interpretation (PA, RO, and AT).

Information Power

Adequacy of the participant sample and subsequent data collection were appraised on an ongoing basis according to the 5 dimensions of "information power" [24]. Several qualities of the study suggested that a less extensive sample would be required to achieve sufficient information power. First, our a priori study objectives were narrowly focused on the acceptability, feasibility, and perceived effectiveness of a specific intervention. Second, the study eligibility criteria

XSL•FO
RenderX

ensured dense sample specificity, where participants all held characteristics which were highly specific to study objectives (eg, patients who were older adults with frailty admitted to 1 of 3 pilot study wards, carers who were unpaid carers to enrolled patients, and ward staff who were clinical staff employed on a permanent basis on one of the pilot study wards). Third, our primary objective to examine intervention acceptability was underpinned by established theory, in the form of the Theoretical Framework of Acceptability of Healthcare Interventions by Sekhon [23]. A fourth dimension, concerning the quality of dialogue, could only be assessed during the data collection phase, rather than at the study outset. We therefore reflect on this later in the paper. Finally, since Framework Analysis approaches entail both within- and between-case analysis using the derived framework matrix, we deemed that a larger volume of data would be necessary to evaluate study objectives, relative to a solely within-case analysis.

Patient and Public Involvement and Engagement

Patient and public involvement and engagement (PPIE) was central to this research. A member of the public with lived experience of caring for an older family member with frailty joined the team as a PPIE researcher, developing a storyboard characterizing patient experiences of frailty and discontinuity at care transitions. This was used to engage ward staff members as to the potential value of the research, helping to support successful staff enrollment. Contributions to data analysis, interpretation, and coauthoring this paper ensured that the perspectives of patients and carers informed each stage of the study.

Ethical Considerations

Ethics approval was granted by an English NHS research ethics committee (IRAS ID: 313814). Informed consent was obtained from all participating carers and ward staff members via completion of paper or electronic consent forms. For patients with the capacity to consent to participation, informed consent was documented by signing a paper consent form. Personal consultees, who knew the patient well and assisted them with their daily activities, were asked to consider the patient's likely wishes about taking part for those patients assessed to lack the capacity to consent to participation. Consultee advice was documented on a printed or electronic consultee declaration form. Enrollment and consent procedures are described in detail in the "Recruitment" section. Participant names and identifiers were replaced with pseudonyms and all study data were securely stored within Imperial College Healthcare NHS Trust's iCARE secure data environment. Further details are provided within the "Data Collection" section. Research participants did not receive payments or other compensation to take part.

Results

Overview

We recruited 107 study participants, consisting of 58 ward staff team members, 29 patients, and 20 carers. Twelve carers additionally acted as a personal consultee for recruited patients. Participating staff included registered nurses and nursing assistants (36/58, 62.1%); physiotherapists, occupational therapists, and therapy assistants (9/58, 15.5%); doctors (11/58, 19.0%); and administrative staff (2/58, 3.4%). Table 3 shows the baseline characteristics for patients who were invited to participate, by enrollment status. Video evaluation questionnaires were completed by just 4 staff members who had watched one or more patient videos (2 nurses and 2 therapy professionals), owing to issues with interpreting Isla Health Digital Pathway Platform metadata, detailed below alongside feasibility findings. Postintervention interviews were conducted with 70 participants, comprising 10 patients, 16 carers, and 44 staff members and lasted between 2 and 25 minutes (mean 11, SD 4.7 minutes).

XSL•FO
RenderX

Table . Baseline characteristics of older adult inpatients with frailty who were invited to participate by enrollment status.[a]

Characteristics	Enrolled patients (n=29)	Nonenrolled patients (n=22)
Sex, n (%)		
Male	18 (62.1)	12 (54.5)
Age (years), mean (range)	82.9 (72-96)	82.2 (66-92)
Ethnicity, n (%)[b]		
White	14 (48.3)	13 (59.1)
Black, Black British, Caribbean or African	3 (10.3)	2 (9.1)
Asian or Asian British	2 (6.9)	3 (13.6)
Mixed or Multiple Ethnic Groups	0 (0)	1 (4.5)
Other	8 (27.6)	2 (9.1)
Not stated	2 (6.9)	1 (4.5)
Dementia diagnosis, n (%)		
Diagnosis received	10 (34.5)	4 (18.2)
Main language spoken at home, n (%)		
English	27 (93.1)	18 (81.8)
Clinical Frailty Scale, n (%)		
1: Very fit	0 (0)	0 (0)
2: Well	0 (0)	0 (0)
3: Managing well	0 (0)	0 (0)
4: Vulnerable	0 (0)	2 (9.1)
5: Mildly frail	5 (17.2)	3 (13.6)
6: Moderately frail	8 (27.6)	4 (18.2)
7: Severely frail	16 (55.2)	13 (59.1)
8: Very severely frail	0 (0)	0 (0)
9: Terminally ill	0 (0)	0 (0)

[a]Other languages spoken at home included Urdu, Gujarati, Greek, and Kurdish.

[b]Ethnicity categories were defined according to UK Census classifications. White includes English, Welsh, Scottish, Northern Irish or British, Irish, Gypsy or Irish Traveller, Roma, and any other White background. Black, Black British, Caribbean or African includes Caribbean, African, and any other Black, Black British, or Caribbean background. Asian or Asian British includes Indian, Pakistani, Bangladeshi, Chinese, and any other Asian background. Mixed or Multiple Ethnic Groups includes White and Black Caribbean, White and Black African, White and Asian, and any other Mixed or multiple ethnic background. Other includes Arab, and any other ethnic group.

Acceptability

Patient Recruitment and Retention

Fifty-one patients were invited to participate, of whom 29 patients (56.9%) were recruited: 17 (58.6%) patients gave informed consent, while a personal consultee was involved in the decision for 12 (41.4%) further patients (Figure 1). One patient withdrew from the study prior to receiving the intervention. Among those who declined participation (22/51, 43.1%), key reasons given by patients included not wanting to take part in research (n=8), not wanting to be video-recorded (n=2), or feeling too unwell (n=2). Consultees who declined on behalf of a patient did so due to believing that the patient would not like to be video recorded (n=3), or concerns about how video recordings would be used (n=1).

Only patients with the capacity to provide their own consent to participate were invited to take part in an interview to assess outcomes. Participating patients were not excluded from interviewing based on whether they had personally received the allocated intervention during the pilot, defined as having at least one video linked to the EPR prior to hospital discharge. As such, follow-up status is displayed for all patients initially enrolled to the study.

XSL•FO
RenderX

Figure 1. Patient flow throughout the nonrandomized mixed methods pilot study (modified from CONSORT [25]). CONSORT: Consolidated Standards of Reporting Trials.

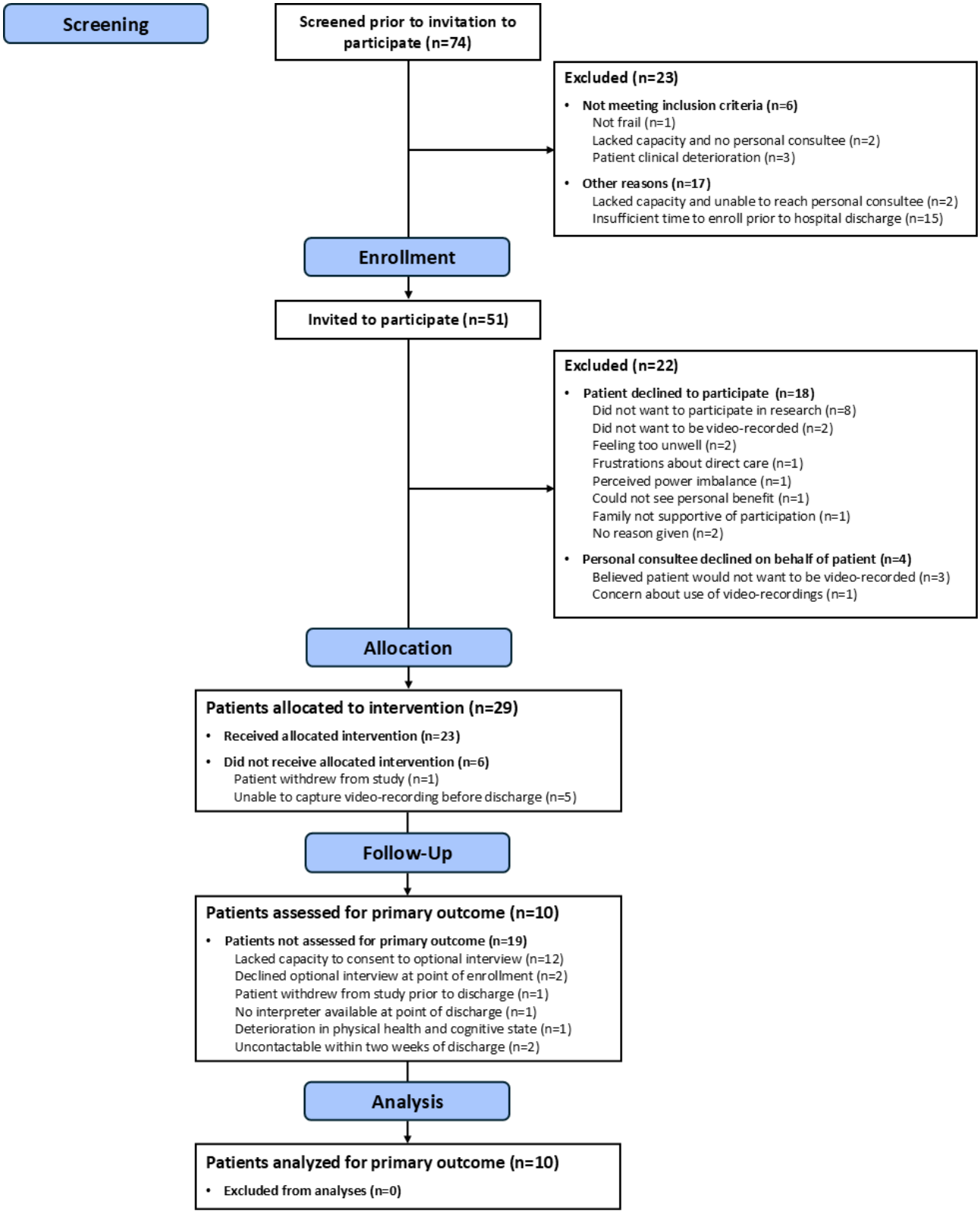

Video Recording Requests by Care Team

During the 3-month intervention pilot, 44 videos were requested by ward staff, with 37 video recordings attempted and 36 submitted within the EPR. Video recordings were focused on patient transfers (n=16), other aspects of mobility (n=13), eating and drinking support (n=3), and patient behavior (n=2). Further videos documented patient-staff interactions (n=1) and seizures (n=1). Video evaluation questionnaire data indicated the suitability of patient videos and the desire among those staff to see more videos in the future. Indeed, all 4 respondents endorsed that videos were of suitable length and quality for clinical interpretation, expressing a desire to see more patient videos in the future.

XSL•FO
RenderX

Perceived Acceptability

Overview

Interviewed patients, carers, and ward staff alike (66/70, 94.3%) typically rated video-based patient records as "completely acceptable" or "acceptable." Three participants stated that they had "no opinion" (4.3%), while a single patient felt that the intervention was "unacceptable" (1.4%); no explanation was provided for this rating. Detailed qualitative findings as to the acceptability of video-based patient records, informed by the Theoretical Framework of Acceptability of Healthcare Interventions by Sekhon, are presented as follows and summarized in Figure 2 [23].

Figure 2. Overview of nonrandomized mixed methods pilot study findings about the acceptability of using video recordings as part of older adult inpatient care.

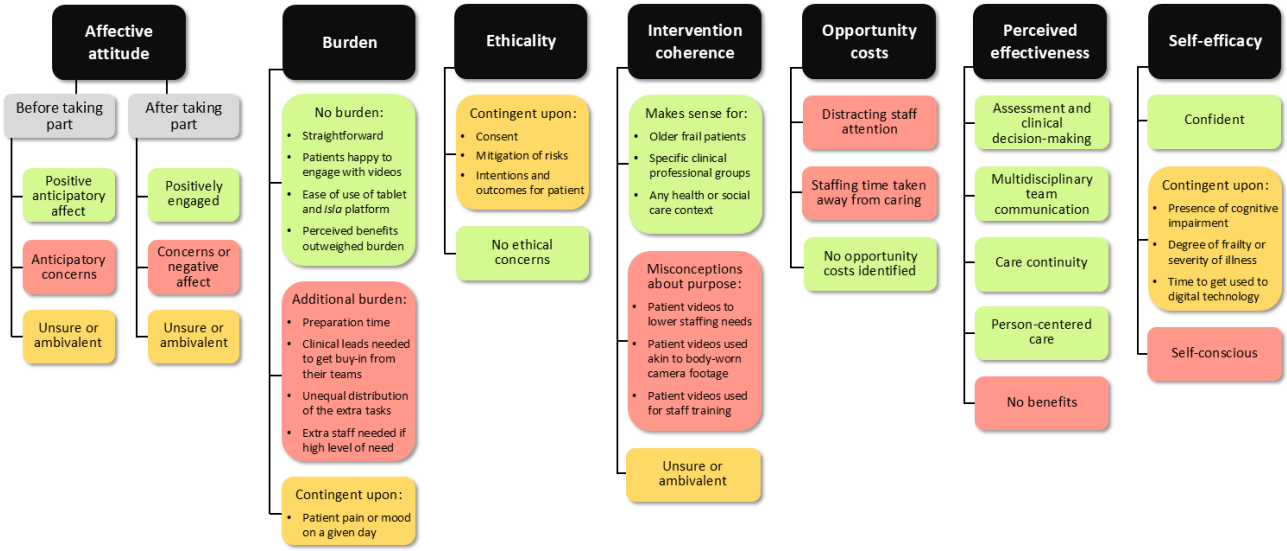

Affective Attitude

Reflecting on how they felt about the intervention prior to participating, those with positive anticipatory affect gave several reasons for this perspective. Video-based patient records were expected to help in establishing an individual's baseline function upon admission, allowing patients and staff alike to appraise progress during the inpatient stay. Staff believed the intervention could foster improved MDT communication and clinical decision-making, valuing the opportunity to pilot the intervention prior to decisions about rollout into routine care:

> *We get new things turn up all the time, "Oh we are doing this now, we got this new policy", whereas because it's being studied in a research way there is going to be lots of oversight, consent, and feeding back.* [Doctor]

Anticipatory concerns were about information security and maintaining patient dignity, particularly where patients lacked the capacity to personally consent to taking part. Staff were also apprehensive about potential workload impacts and whether patients and families would be supportive of the intervention. In practice, capturing and viewing patient videos was less burdensome than expected and staff were surprised to find that many patients were enthusiastic about taking part:

> *I was a bit skeptical, but found that the more we saw the patients, the more eager they were about the recording.* [Therapies staff]

Burden

Patients and carers believed that burdens to patients in participating in the intervention were limited:

> *I felt it was something that needed to be done, it didn't take very long. I didn't have a lot to do so it was very positive.* [Patient]

The potential to exacerbate behavioral symptoms in patients with agitation or cognitive impairment was carefully considered by staff and carers. This was mitigated by avoiding filming or terminating patient videos immediately upon any indication of patient distress (verbal or nonverbal). Burdens for staff included time required for preparation or additional staffing needs where patients required assistance of 2 for mobilization, with a third staff member needed to capture the video.

Ethicality

Participants regarded the use of video recordings within the assessment and care of older inpatients with frailty as ethical, on the grounds that certain conditions were met. A central view was that ethicality was contingent upon patient consent to appear in videos. Where a person was assessed to lack the capacity to consent, differing views were presented as to whether a personal consultee or next-of-kin should be allowed to make this decision on a patient's behalf. Given that a person's next-of-kin would typically be best positioned to understand the patient's likely wishes, next-of-kin consent was largely considered suitable. However, others voiced that patient consent was the only basis on which the intervention could be judged as ethical:

> *I would feel something like that needs to come from them [patients]… and I don't think the next-of-kin should be making that decision for them.* [Carer]

Mitigation of risks associated with video capture was another prerequisite for ethicality. A central principle was that a patient's dignity must always be upheld during video recording,

XSL•FO
RenderX

particularly given the permanence of digital patient records. Patient confidentiality and information security were further risks agreed by participants to warrant safeguards. Participants described feeling reassured about the extensive safeguards in place throughout the intervention pilot. There was consensus that patient videos should be viewed only by staff members who need to see them and must be stored in a secure digital environment that is resilient to cyberattacks.

The intentions and outcomes of the use of video-based patient records comprised another determinant of their ethicality. Where an individual's personal, religious, and cultural beliefs were considered and there was a clear clinical rationale for capturing a patient video to optimize an aspect of care, conditions for ethicality were met. Indeed, staff indicated that patient videos may help to accelerate processes of identifying discharge placements for patients, thus reducing other risks associated with lengthy hospital stays, including deconditioning or exposure to hospital-acquired infections. There was a perception that withholding such an intervention from patients, due to concerns about frailty or cognitive impairment, meant that older inpatients risked missing out on potentially enhanced care. Patients justified ethicality based on enjoying tracking their own progress and experiencing no negative outcomes from taking part:

> *It hasn't hurt me, so why should I complain?* [Patient]

Beneficial patient outcomes associated with video-based patient records were thought most likely to be realized where there is potential for an individual to return to their mobility baseline, while benefits were deemed less clear for extremely frail or cognitively impaired patients.

Intervention Coherence

It was evident from some participants' accounts that they understood how video-based patient records could support patient care and why the intervention had been piloted within the 3 wards. Those who could confidently explain the intervention aims felt it made particular sense to use patient videos in the care of older patients, where rapid changes in presentation may occur:

> *I can see now how it might be important for people to see the transformation in me, the way I was a few days ago. Now, I've lost ten years.* [Patient]

Comparisons to usual care were made at several points. Staff described the limitations inherent in text-based patient information and reflected upon the potential added value of visual data for obtaining insights into a patient's needs:

> *It's quite good technology to have, like I say, written documentation doesn't always convey what's going on, the patient's abilities. So, to actually see something, especially in regards to moving and handling, things like that.* [Nursing staff]

Staff believed that patient videos had applications for all MDT members. However, the intervention was thought most relevant to the work of physiotherapists and occupational therapists, as well as doctors. Moreover, others regarded that video recordings may be coherently used in any health or social care context.

In contrast, some patients and carers were unsure about how the intervention could be useful as part of clinical care. Furthermore, several misconceptions were expressed about the purpose of patient videos. Some staff members believed that patient videos would be used to reduce staffing requirements. Likewise, others misunderstood how and why patients may be filmed, believing that video capture would be used in the same way as body-worn cameras, which are typically used for continuous or targeted filming to document violence and aggression in health care or policing contexts.

Opportunity Costs

Opportunity costs, where other potential benefits were forgone in implementing the video-based patient record intervention, were seldom identified. A concern was that in capturing patient videos, staff may be distracted from risks on the ward, such as wandering patients:

> *If a patient fell… you obviously want to focus your attention on what's going on. You don't want to be thinking about, "Is that angle right? Am I capturing everything?"* [Therapies staff]

Second, an additional member of staff was often asked to assist in holding the tablet during video capture, thus stepping away from the task they were completing at the time.

Self-Efficacy

Finally, many participants expressed confidence in either capturing (staff) or appearing in videos (patients, carers, or staff). Self-efficacy was reportedly lower where patients had cognitive impairment; patients were also less comfortable appearing in videos on days when they were especially unwell. Some ward staff indicated initial hesitancy in using a new form of digital technology, needing time to become accustomed to using the tablet. A minority of patients and staff felt self-conscious over their physical appearance or about the sound of their voice when appearing in videos.

Feasibility

Diversity of Patient Sample

Enrolled patients were largely male (18/29, 62.1%), had a mean age of 82.9 (range 72 - 96, SD 6.7) years, mostly spoke English at home (27/29, 93%), and were diverse in ethnicity (Table 3). Clinical Frailty Scale scores for participating patients ranged from 5 (Mildly frail) to 7 (Severely frail), with most patients evaluated to be "Severely frail" (n=16) [26]. Around one-third (10/29, 34.5%) had a dementia diagnosis. Baseline characteristics were similar between enrolled and nonenrolled patients, apart from dementia diagnosis and main language spoken at home, where higher proportions of participating patients had dementia and spoke English.

Privacy and Security Concerns

During the 3-month pilot, the study clinical lead was asked to review a single patient video owing to privacy concerns. In the video, another patient could be observed stepping into the video background. The video was removed from the Isla Health Digital Pathway Platform, owing to concerns that the other patient might be identifiable from the footage.

XSL•FO
RenderX

Use of the Isla Health Digital Pathway Platform

Videos were captured on a total of 19 days within the 3-month pilot. The video recording process typically involves 2 or 3 staff members, with one to carry out the recording and others to provide direct support to the patient. Information as to those involved in this process was limited to the person capturing the video recording only, based on login data for the platform. We ascertained that 11 different staff members used their logins to capture and upload videos but were unable to determine the total number of staff members involved in the recording process in a supporting role.

Use of the Isla Health Digital Platform Pathway was similar across the 3 pilot wards. Ward 1 captured 13 videos (among n=10 recruited patients), with 17 captured on Ward 2 (among n=13 recruited patients), and 6 on Ward 3 (among n=6 recruited patients). Overall, of 29 participating patients, 79.3% (n=23) had at least one video linked to their EPR at the point of discharge. The research team was notified of 12 unsuccessful attempts to take a video. Primary reasons documented included the patient declining (n=4), patient cognitive or medical deterioration (n=3), and insufficient staffing (n=2).

Video views by ward staff were quantified through Isla Health Digital Pathway Platform metadata, where staff clicked to view and expand a video tile, resulting in a documented "View submission" action. Four such views were counted. However, toward the end of the pilot phase, we ascertained through observing staff interact with the platform that watching a video without expanding on the video tile was instead recorded within the Isla platform metadata as "View folder." Accordingly, excluding cases where a "View folder" action occurred directly before or after the same staff member uploaded a video for a given patient (when the folder view likely served a different purpose), we estimate that there were 17 video views during the study. Seemingly low video view rates among staff were corroborated by anecdotal insights obtained through staff interviews. Although staff members could typically identify potential intervention benefits at the team level, when asked directly, interviewed staff commonly attested that they had not personally viewed or sought to access a patient video at any point during the 3-month pilot.

Intervention Barriers and Facilitators

Nine themes were inductively developed from interview data as to the barriers and facilitators to implementing the video-based patient records intervention in the care of older adult hospital inpatients with frailty (Figure 3).

Figure 3. Barriers and facilitators to implementing video-based patient records in acute medical wards for older adult inpatients with frailty. EPR: electronic patient record.

XSL•FO
RenderX

Ward Environment and Configuration

The crowded, busy ward environment and layout were such that staff struggled to capture patient videos while maintaining the privacy of participating patients. Likewise, difficulties were faced in ensuring that other patients and staff members were not captured in the background of patient videos. Curtains around the bed space were valued by staff for facilitating a degree of privacy during the video recording process.

Digital Technology and Infrastructure

Digital technology and infrastructure barriers stemmed from poor quality Wi-Fi coverage across the 3 pilot wards. Technology hesitancy and delays or challenges in logging into the video-based patient records system were also experienced by some staff members:

> *We are so tied down by logins, governance, two-factor authentication that if you want people to be able to have a low threshold to record, check, share something, you can't do it.* [Doctor]

Overall, the Isla Health Digital Pathway Platform was, however, considered straightforward to use.

Equipment

Staff reported several equipment-related challenges. Locating the designated tablet promptly when required or finding that the battery had not been charged by previous users were key frustrations. An inbuilt stand within the tablet case served as a facilitator to the video recording process, allowing staff to position the tablet on a surface within the ward (eg, bedside table) when another staff member could not assist with video capture.

Workflow, Scheduling, and Logistical Considerations

Further obstacles to using the intervention are related to ward workflow and logistical considerations. Videos often needed to be taken at an optimal moment when seeking to capture intermittent presentations (eg, seizures) or for fatigued patients. Patients were sometimes discharged earlier than expected, meaning that video recordings could not be captured in time or used to realize potential benefits. Staff described integrating video capture into their usual planned therapy sessions with participating patients, or capturing videos during quieter periods, to maximize efficiency and feasibility.

Training and Support

Regular staff turnover presented difficulties in ensuring all staff were trained in using the intervention. From the staff engagement and enrollment phase through to the end of the 3-month pilot, there were rotations in both junior doctors and specialty trainees on a national level. When asked about low apparent usage of videos, some staff members indicated that they could not recall how to access videos within the EPR:

> *I think it's quite straightforward when you record it, but I think there is still some sort of gap, like how you follow up with Isla outside of the iPad [tablet].* [Therapies staff]

Nevertheless, staff and patients alike appreciated access to ongoing training and support from the research team throughout the 3-month trial period.

Reminders and Prompting

Remembering to capture and use videos within patient care was a persistent challenge. This served as a barrier to uptake and achieving a meaningful intervention "dosage" for participants, where it was felt that more videos throughout an individual's inpatient stay would be necessary to visually demonstrate improvement or deterioration.

> *I was naively hoping it would be more visible. Having people nudging and reminding definitely helps, but that would be my concern for wider adoption.* [Doctor]

For other staff, there appeared to be low impetus to access and use patient videos once captured:

> *Not many people would go and check, or would say, "Okay, on Cerner [EPR], I want to go and see that video."* [Therapies staff]

Reminders used throughout the pilot study period (eg, stickers above beds of enrolled patients and reminders from the research team) were valued but thought insufficient for comprehensive implementation.

Health and Safety Management

Managing safety when capturing videos, including staff's ability to prevent patient falls during video-recorded mobility assessments, was an important consideration. Other challenges included infection control considerations when switching between delivering direct patient care and using the tablet to capture a patient video.

> *I got pointed out for infection control… they recommended me leaving the iPad [tablet] there, going to wash my hands, come back and go onto the tablet. But I can't leave the patient.* [Therapies staff]

Staff detailed practical issues in capturing videos, where handwashing was required between each activity.

Staff Engagement, Staffing, and Workload

Moreover, intervention feasibility was also contingent upon staff engagement, staffing, and workload factors. Barriers were presented when wards were short-staffed, resulting in high workload for clinical teams and limited spare capacity to assist with capturing videos. Reassuring staff that video recordings were not intended to be used for appraising staff practice and recognizing the efforts of staff to deliver the intervention were important for maintaining staff engagement. Implementation was reportedly best where an MDT approach to capturing patient videos was adopted, so that individual professional groups (ie, nursing or therapy staff) did not feel alone in shouldering additional task burdens. Likewise, carers emphasized the importance of staff taking time to explain to patients why they wished to capture a video of a given activity or symptom and how the video may be used as part of their care:

XSL•FO
RenderX

> *Is there rapport being built up? Is the person comfortable with those particular healthcare professionals?* [Carer]

By allowing time to develop positive staff-patient relationships, patients felt more comfortable with taking part.

Patient and Family Factors

Finally, patient and family factors were important to the implementation of video-based patient records. On some days, interviewees described that patients were reluctant to accept any care or interventions on the ward, including patient videos. Older, frail patients may decline to engage owing to experiencing pain or fatigue on a given day. Seemingly uncooperative behavior was displayed at times among participating patients due to confusion or fluctuation in cognitive function, presenting a barrier to delivering the video-based patient record intervention:

> *When they were trying to video something more positive with him, he decided not to be as positive! But that's hit and miss with any of the patients I would think.* [Carer]

Nevertheless, a key facilitator indicated by all participant groups was that patients and carers were often happy to be involved in trialing a new intervention which may improve their own care or that of others.

Perceived Effectiveness

Overview

Perceived effectiveness of the video-based patient records intervention was appraised according to ward staff views on its potential usefulness in the care of older adult inpatients with frailty, as well as patient, carer, and staff perspectives on intervention impacts. Participants conceived of multiple ways in which video-based patient records could be potentially effective for use in the present context. Some participants were unaware of any intervention impacts. Such patients and carers described a lack of follow-up from ward staff about their videos, meaning that they were unsure whether the video recordings had been beneficial. No adverse events were reported. Study data indicated several ways in which video recordings may improve aspects of patient care.

Assessment and Clinical Decision-Making

Data from the 4 completed video evaluation questionnaires showed that ward staff considered the video they had watched was "very useful" for supporting patient assessment and clinical decision-making. Interviewees elaborated on the potential for patient videos to improve assessment quality by mitigating subjectivity in clinical interpretation and speeding up the assessment process. For behaviors or symptoms occurring at unpredictable intervals, videos also permitted staff to obtain specialist advice from other teams:

> *It helped our neuro team see a patient who was having intermittent seizures… they [neurology team] would not have been able to see it otherwise.* [Nursing staff]

Further potential benefits were in offering clear information about an individual's care and support needs, providing evidence for referrals to other care providers. Staff reflected on challenges faced in securing onward care placements for patients who had exhibited fluctuating behavior and nursing care requirements in the hospital. Rehabilitation settings and nursing homes reportedly declined referrals or overestimated staffing costs necessary to provide patient care, based on patients having had an acute or episodic change in presentation (eg, delirium onset) during their admission:

> *If we had someone come in for assessments, we had this evidence to say "Okay, look, on a good day this patient did it this way, so you need to assess them based on that."* [Nursing staff]

In contrast, an alternative position was that videos merely provided a snapshot of a person's presentation:

> *You're putting someone into a slightly odd position saying, "Now, walk for us." And I think you'll get the Charlie Chaplin effect, people starting to do things that maybe they wouldn't normally do… I think a camera is not always going to bring out the truth in people.* [Carer]

Therefore, the potential for video-based patient records to improve assessment and decision-making precision was questioned.

Multidisciplinary Team Communication

All 4 staff members who completed video evaluation questionnaires reported that the video they had watched was "very useful" for communicating patient information to colleagues. The added communicative value of visual data over written documentation or verbal handovers was apparent from interview data. Carers were similarly supportive of patient videos, where their use might improve communication quality and concision:

> *I don't know whether that would completely eradicate the need for discussion… but it could perhaps enhance that, if there's something that someone needs to show someone else.* [Carer]

Participants also envisaged potential value for strengthening communication with other clinical settings. Carers felt that videos could optimize care timeliness and continuity by enabling remote communication with other care providers. This view was supported by staff, who also believed that patient videos may improve collaboration between clinical services, potentially reducing duplication of clinical assessments:

> *We have a deficit of trust across boundaries… but if you see a video they know we are not making it up.* [Doctor]

Moreover, videos were reportedly used to enhance communication with patients' families:

> *Sometimes the family says, "Oh, he could not do it at home." When they see the patient sitting out and walking, it's amazing them.* [Nursing staff]

As such, this helped to provide a tangible demonstration of care and interventions received in the hospital and apparent impacts.

XSL•FO
RenderX

Care Continuity

Study findings suggest that video-based patient records could improve care quality by optimizing continuity, both within and beyond an individual's hospital stay. Patient videos may support joined-up care by helping to establish a shared understanding of the goals of a patient's admission, informed by videos showing their functional baseline. Similarly, participants reported that videos allowed staff to objectively monitor patients over time, facilitating greater awareness of signs of improvement or deterioration:

> *For the patient it's better because we know their progress… I'm video-recording to compare. We see the difference from bedbound to mobilizing.* [Nursing staff]

Video-based patient records were also deemed valuable for aiding safe care transitions. Staff thought patient videos could assist community therapy teams to provide joined-up care following discharge. Since frail older adults often experience multiple hospital admissions, participants also envisaged using patient videos to gain insights into a person's presentation during prior inpatient stays.

Person-Centered Care

The potential for video-based patient records to improve person-centered care was a prominent theme. By providing a visual record of a patient's achievements throughout their hospital admission, patient videos promoted a focus on an individual's goals and needs, contributing to more personalized care experiences:

> *I was proud to see me in the video and pleased that I have made so much progress… it gave me great spirits.* [Patient]

Moreover, the use of patient videos was thought to improve the quality of staff-patient engagement. Staff described how the process of delivering the intervention offered opportunities for patients to receive additional encouragement from ward staff, which in turn improved staff satisfaction:

> *The patient was eager, because he was walking and the staff… they were following and cheering him on. It was so beautiful – he was able to mobilize himself. They were all so happy, the staff and the patient themself [sic] because he wanted to go home!* [Nursing staff]

Finally, there was a sense that patients appreciated the choice over whether to take part in the intervention:

> *It gives the patient a little say in what's going on in their life as a patient.* [Patient]

Where patients may feel they have limited involvement in care decision-making when in hospital, the opportunity to make choices about their care appeared to instill a sense of agency and empowerment.

Discussion

Principal Findings

Findings from this nonrandomized mixed methods pilot study suggest that the use of video recordings within patient assessment and care delivery is largely acceptable to older patients with frailty, their family members or carers, and clinical staff. This was demonstrated by promising participant recruitment and retention rates; enrollment targets for patients were almost met, while ward staff recruitment vastly exceeded intended numbers. The range of videos captured, from documentation of seizures through to establishing a person's mobility baseline, shows that staff discerned multiple use cases for patient videos. On the grounds that ethical preconditions are met, including mitigation of concerns about information security and patient dignity, video-based patient records were largely deemed to be acceptable.

A secondary objective to explore the feasibility of implementing video-based patient records within acute wards for older adults revealed mixed findings. Diversity of the patient sample in terms of clinical and demographic characteristics was encouraging, indicating that risks of widening health inequalities through inequitable uptake are low, should the intervention be implemented into routine care. Implementation barriers were multiple, yet participants described a range of facilitators that supported successful delivery. However, video view metric estimates and anecdotal evidence from interviews with ward staff suggested limited engagement with videos once captured, thus casting doubt on the feasibility of attaining sufficient uptake among staff. We estimate that over half of all videos captured were never viewed. Nevertheless, patient, carer, and clinical opinion as to perceived intervention effectiveness yielded evidence of promise. Videos were considered useful in supporting clinical assessment, enhancing MDT communication, strengthening care continuity, and promoting person-centered care.

Comparison With Wider Literature

Our findings are largely in accordance with the conclusions of a recent systematic review, which examined empirical research and regulatory guidance on the generation and use of video recordings as part of patient care [16]. Indeed, similarly to existing studies [16,27,28], we observed that video-based patient records were acceptable to patients, carers, and clinical staff alike, so long as ethical conditions were satisfied around upholding patient dignity and maintaining the security of these digital data [29]. Likewise, based on clinical opinion, we found that patient videos may aid the monitoring of care delivery and outcomes over time, aligning with the conclusions of comparable research [15].

Perhaps most striking was our finding that video-based patient records appear to be valuable for promoting person-centered care within the acute hospital setting. The intervention seemed to instill a sense of agency and empowerment in older adults with frailty who took part in video recordings. Staff experienced secondhand satisfaction upon observing a patient demonstrating a functional improvement, such as walking unaided, when they had previously felt too frail to do so. This finding resonates with

XSL•FO
RenderX

prior research from physiotherapy and rehabilitation contexts, where video recordings documenting improvements in patient gait throughout their inpatient stay were found to improve patient motivation and satisfaction [30]. Likewise, although we piloted the use of patient videos captured by ward staff, rather than exploring patient-generated visual media as reported within a further review [15], we similarly concluded that patient videos helped staff to gain a more individualized understanding of patients. A deeper understanding of patients as individuals likely, in part, underpinned the closer staff-patient engagement anecdotally reported by participating staff and patients.

Older adults with frailty often face multiple transitions between health and social care settings, owing to complex care needs stemming from multimorbidity [6,7]. An important application of the present intervention, as conceived by ward staff, lay in the potential to capture video footage which could be shown to other departments or providers to inform decision-making and support joined-up care. For example, staff spoke about the potential value in sharing footage with neurology specialists or with staff carrying out assessments for care home placements. For the purposes of this pilot study, staff could only show video recordings to health care professionals outside the direct care team who visited the pilot wards in person. Should wider implementation of video-based patient records be considered, the lack of interoperability of different EPR systems, a well-documented obstacle to delivering digital interventions at scale, warrants timely address [31,32].

The integration of new digital technologies into health care services resembles a complex challenge, with tailored strategies required to promote successful adoption and sustainability of a given intervention [33]. Our findings point to a series of barriers and facilitators to the implementation of video-based patient records, including those pertaining to staff engagement, workflow, and the adequacy of reminders and prompting used to encourage staff to capture and view patient videos. So that potential enhancements to patient care can be realized, theoretically informed strategies drawing upon behavioral science or implementation science frameworks are likely required to encourage uptake [34].

Finally, it is well documented that severely frail patients and people who are unable to consent to participation are seldom included within research studies [35,36]. Our findings reinforce existing literature which points to the need for careful ethical review of procedures for patient inclusion [37], to increase opportunities for such patient populations to contribute to applied clinical and health services research. This is so that they are not excluded from the development of innovations which may result in enhanced care.

Strengths and Limitations

Strengths of this study lie in its representation of patient populations who are underrepresented and underserved in clinical research. This includes older adults with severe frailty and people who lack the capacity to consent to research owing to cognitive impairment [35]. Exclusion of such patients from clinical trials is widespread, thus obstructing the potential to develop improved health care services and interventions that meet the needs of these populations [36]. High quality and comprehensiveness in qualitative data collection components were indicated according to multiple dimensions of information power [24]. For instance, data were sufficiently comprehensive given our narrow study aims, high sample specificity while capturing variation in patient demographic and clinical characteristics, and owing to the application of established theory. We ensured rigor in our analysis of interview data through the Framework Analysis approach [22], applying an existing theoretically informed framework to guide data charting [23], while inductively introducing new codes into the coding scheme to characterize each construct. Further strengths were in our approach to PPIE, ensuring that lived experience perspectives guided the study throughout the research cycle.

Issues in interpreting Isla Health Digital Pathway Platform metadata, such that video view figures were estimates that cannot be determined with exactness, represent an important limitation. Other limitations pertain to the study design. The single-site, nonrandomized nature of this research and the lack of a comparator group should be taken into account when considering our findings. While we appraised the quality of interview dialogue to be high overall, we hold that dialogue flow was at times impacted within the busy ward environment in which data were gathered. Interviews with patients were often inadvertently interrupted by staff or visitors, while staff were regularly distracted by other tasks (eg, incoming telephone calls) while being interviewed. We also note that among participating patients with the capacity to consent to an optional interview during enrollment, a number of these individuals were not interviewed due to reasons such as being uncontactable within 2 weeks of leaving the hospital, deterioration in cognitive state, and lack of interpreter availability at the point of discharge. Had it been possible to interview these individuals, it is plausible that a wider range of perspectives on the intervention may have surfaced, with greater information power to permit between-case analysis. Nonetheless, we offer a robust evaluation of intervention feasibility, applying our findings stringently to decision-making about progression to a definitive trial.

Implications

Study findings provide evidence that the use of video-based patient records within direct care delivery for older adults with frailty was largely deemed to be acceptable. We also offer compelling preliminary insights based on participating patient, carer, and clinical staff opinions as to the potential effectiveness of this intervention in the acute medical inpatient ward context. When implemented successfully, our findings suggest that the intervention could contribute to improved care quality and safety. However, the feasibility of intervention implementation must be questioned, given estimated video view metrics and anecdotal evidence from staff interviews, which indicated that uptake among staff in terms of capturing and viewing patient videos did not reach a threshold to permit full realization of benefits. Evidence for ethicality and a lack of adverse outcomes suggests that the intervention could be considered appropriate for use within other clinical services when indicated, for example, as part of efforts to empower patients and to promote person-centered care. Further professional and regulatory guidance to support clinical staff in the implementation of such an intervention is, however, warranted [16].

XSL•FO
RenderX

Conclusions

Addressing the primary objective of this pilot study, the video-based patient record intervention was found to be largely acceptable to older inpatients with frailty, carers, and staff. Participants also held promising opinions about possible intervention impacts. Optimizing assessments, enhancing care continuity, and empowering patients were considered important potential benefits of using patient videos within direct care. No adverse events were reported. Findings about patient diversity, alongside barriers and facilitators to intervention implementation, support encouraging indications overall as to the feasibility of implementing such an intervention within this care context. However, apparent limited engagement with patient videos after their capture during the pilot phase would likely obstruct our ability to measure intervention efficacy in future research. As such, we cannot endorse progression to a definitive trial to examine the effectiveness of patient videos for improving safety and quality of care transitions. Nevertheless, our findings provide an initial indication that the intervention could be ethically implemented into routine practice, with suitable ethical safeguards in place.

Acknowledgments

The authors would like to thank the patients, carers, and ward staff who took part for their valuable contributions to this research. We also express our sincere thanks to Dr Colin Mitchell and Dr David James for supporting the successful implementation and evaluation of the Isla for Frailty intervention within their wards.

Funding

This study was funded by the National Institute for Health and Care Research (NIHR) North West London Patient Safety Research Collaboration (NIHR204292) with Infrastructure support provided by the NIHR Imperial Biomedical Research Centre (NIHR203323). PA is currently supported by a Better Health & Care Hub Postdoctoral Fellowship awarded by the Better Health & Care Hub, King's College London. Views expressed in this publication are those of the authors and not necessarily those of the NIHR, Isla Health, nor the Better Health & Care Hub.

Data Availability

The dataset generated and analyzed during this study is not publicly available as consent for data sharing was not sought from research participants and was not agreed as part of the ethics approval obtained to undertake this research. Any queries about study data may be directed to the corresponding author.

Authors' Contributions

Conceptualization: RL (lead), SL, SE, EM
Formal analysis: PA (lead), RO, AT
Funding acquisition: RL, EM
Investigation: PA, RL, RO, SL, AT, PJC, JF, UE, JL
Methodology: RL, SL, SE, AT
Project Administration: PA, RL, RO
Supervision: PA, RL, SL
Writing – original draft: PA, RL
Writing – review & editing: PA, RL, RO, SL, AT, PJC, JF, UE, JL, SE, EM

Conflicts of Interest

None declared.

Multimedia Appendix 1
Procedure for patient screening and enrollment.
[DOCX File, 49 KB - jmir_v28i1e77318_app1.docx]

Multimedia Appendix 2
Video evaluation questionnaire completed by ward staff upon viewing a patient video.
[DOCX File, 27 KB - jmir_v28i1e77318_app2.docx]

Multimedia Appendix 3
Interview guide for patients and carers.
[DOCX File, 26 KB - jmir_v28i1e77318_app3.docx]

XSL•FO
RenderX

Multimedia Appendix 4
Interview guide for ward staff.
[DOCX File, 26 KB - jmir_v28i1e77318_app4.docx]

Multimedia Appendix 5
Component constructs of the Theoretical Framework of Acceptability of Healthcare Interventions.
[DOCX File, 26 KB - jmir_v28i1e77318_app5.docx]

References

1. What is frailty? Age UK. 2020. URL: https://www.ageuk.org.uk/our-impact/policy-research/frailty-in-older-people/understanding-frailty/ [accessed 2025-03-07]
2. Introduction to frailty. British Geriatrics Society. 2014. URL: https://www.bgs.org.uk/resources/introduction-to-frailty [accessed 2025-03-07]
3. Dlima SD, Hall A, Aminu AQ, Akpan A, Todd C, Vardy E. Frailty: a global health challenge in need of local action. BMJ Glob Health 2024 Aug 9;9(8):e015173. [doi: 10.1136/bmjgh-2024-015173] [Medline: 39122463]
4. Proactive care: providing care and support for people living at home with moderate or severe frailty. NHS England. 2023. URL: https://www.england.nhs.uk/long-read/proactive-care-providing-care-and-support-for-people-living-at-home-with-moderate-or-severe-frailty/ [accessed 2025-04-04]
5. Doody P, Asamane EA, Aunger JA, et al. The prevalence of frailty and pre-frailty among geriatric hospital inpatients and its association with economic prosperity and healthcare expenditure: a systematic review and meta-analysis of 467,779 geriatric hospital inpatients. Ageing Res Rev 2022 Sep;80:101666. [doi: 10.1016/j.arr.2022.101666] [Medline: 35697143]
6. Hewitt J, McCormack C, Tay HS, et al. Prevalence of multimorbidity and its association with outcomes in older emergency general surgical patients: an observational study. BMJ Open 2016 Mar 31;6(3):e010126. [doi: 10.1136/bmjopen-2015-010126] [Medline: 27033960]
7. Sinclair DR, Maharani A, Chandola T, et al. Frailty among older adults and its distribution in England. J Frailty Aging 2022;11(2):163-168. [doi: 10.14283/jfa.2021.55] [Medline: 35441193]
8. Sadler E, Potterton V, Anderson R, et al. Service user, carer and provider perspectives on integrated care for older people with frailty, and factors perceived to facilitate and hinder implementation: a systematic review and narrative synthesis. PLoS One 2019;14(5):e0216488. [doi: 10.1371/journal.pone.0216488] [Medline: 31083707]
9. Baillie L, Gallini A, Corser R, Elworthy G, Scotcher A, Barrand A. Care transitions for frail, older people from acute hospital wards within an integrated healthcare system in England: a qualitative case study. Int J Integr Care 2014 Jan;14(1):e009. [doi: 10.5334/ijic.1175] [Medline: 24868193]
10. Welch C, Chen Y, Hartley P, et al. New horizons in hospital-associated deconditioning: a global condition of body and mind. Age Ageing 2024 Nov 1;53(11):afae241. [doi: 10.1093/ageing/afae241] [Medline: 39497271]
11. Transitions of care. World Health Organization. 2016. URL: https://iris.who.int/handle/10665/252272 [accessed 2025-04-10]
12. Coffey A, O'Reilly P, Meskell P, et al. The development of a national transfer document: for use when an older person is being transferred from residential to acute care settings. : Health Service Executive; 2019 URL: http://hdl.handle.net/10147/627131 [accessed 2025-12-19]
13. Kripalani S, LeFevre F, Phillips CO, Williams MV, Basaviah P, Baker DW. Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. JAMA 2007 Feb 28;297(8):831-841. [doi: 10.1001/jama.297.8.831] [Medline: 17327525]
14. Gonçalves-Bradley DC, J Maria AR, Ricci-Cabello I, et al. Mobile technologies to support healthcare provider to healthcare provider communication and management of care. Cochrane Database Syst Rev 2020 Aug 18;8(8):CD012927. [doi: 10.1002/14651858.CD012927.pub2] [Medline: 32813281]
15. Ploderer B, Rezaei Aghdam A, Burns K. Patient-generated health photos and videos across health and well-being contexts: scoping review. J Med Internet Res 2022 Apr 12;24(4):e28867. [doi: 10.2196/28867] [Medline: 35412458]
16. Lear R, Ellis S, Ollivierre-Harris T, Long S, Mayer EK. Video recording patients for direct care purposes: systematic review and narrative synthesis of international empirical studies and UK professional guidance. J Med Internet Res 2023 Aug 16;25:e46478. [doi: 10.2196/46478] [Medline: 37585249]
17. Lancaster GA, Thabane L. Guidelines for reporting non-randomised pilot and feasibility studies. Pilot Feasibility Stud 2019;5(1):114. [doi: 10.1186/s40814-019-0499-1] [Medline: 31608150]
18. Eldridge SM, Chan CL, Campbell MJ, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016 Oct 24;355:i5239. [doi: 10.1136/bmj.i5239] [Medline: 27777223]
19. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014 Mar 7;348(mar07 3):g1687. [doi: 10.1136/bmj.g1687] [Medline: 24609605]
20. Mental capacity act 2005. Legislation.gov.uk. URL: https://www.legislation.gov.uk/ukpga/2005/9/section/32 [accessed 2025-02-10]

XSL•FO
RenderX

21. Department of Health Scientific Development and Bioethics Division. Guidance on nominating a consultee for research involving adults who lack capacity to consent. : Department of Health; 2008 URL: https://www.manchester.gov.uk/downloads/file/12218/guidance_on_nominating_a_consultee_for_research_involving_adults_who_lack_capacity_to_consent [accessed 2025-12-29]
22. Ritchie J, Spencer L. Qualitative Data Analysis for Applied Policy Research Analyzing Qualitative Data: Routledge; 1994:173-194. [doi: 10.4324/9780203413081_chapter_9]
23. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017 Jan 26;17(1):88. [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
24. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
25. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010 Mar 23;340:c869. [doi: 10.1136/bmj.c869] [Medline: 20332511]
26. Moorhouse P, Rockwood K. Frailty and its quantitative clinical evaluation. J R Coll Physicians Edinb 2012;42(4):333-340. [doi: 10.4997/JRCPE.2012.412] [Medline: 23240122]
27. Bayen E, Jacquemot J, Netscher G, Agrawal P, Tabb Noyce L, Bayen A. Reduction in fall rate in dementia managed care through video incident review: pilot study. J Med Internet Res 2017 Oct 17;19(10):e339. [doi: 10.2196/jmir.8095] [Medline: 29042342]
28. Towsley GL, Wong B, Mokhtari T, Hull W, Miller SC. Piloting me and my wishes-videos of nursing home residents' preferences. J Pain Symptom Manage 2020 Mar;59(3):609-617. [doi: 10.1016/j.jpainsymman.2019.10.030] [Medline: 31711970]
29. Williams K, Arthur A, Niedens M, Moushey L, Hutfles L. In-home monitoring support for dementia caregivers: a feasibility study. Clin Nurs Res 2013 May;22(2):139-150. [doi: 10.1177/1054773812460545] [Medline: 22997349]
30. Jayabalan P, Kaplan R, Breisinger T, et al. Poster 311 Video recording the gait of stroke patients during inpatient rehabilitation to improve motivation, satisfaction and outcome. PM R 2014 Sep;6(9S):S170. [doi: 10.1016/j.pmrj.2014.08.374]
31. Li E, Clarke J, Ashrafian H, Darzi A, Neves AL. The impact of electronic health record interoperability on safety and quality of care in high-income countries: systematic review. J Med Internet Res 2022 Sep 15;24(9):e38144. [doi: 10.2196/38144] [Medline: 36107486]
32. Shull JG. Digital health and the state of interoperable electronic health records. JMIR Med Inform 2019 Nov 1;7(4):e12712. [doi: 10.2196/12712] [Medline: 31682583]
33. Pearce L, Costa N, Sherrington C, Hassett L. Implementation of digital health interventions in rehabilitation: a scoping review. Clin Rehabil 2023 Nov;37(11):1533-1551. [doi: 10.1177/02692155231172299] [Medline: 37132030]
34. Handley MA, Gorukanti A, Cattamanchi A. Strategies for implementing implementation science: a methodological overview. Emerg Med J 2016 Sep;33(9):660-664. [doi: 10.1136/emermed-2015-205461] [Medline: 26893401]
35. Shepherd V, Wood F, Griffith R, Sheehan M, Hood K. Protection by exclusion? The (lack of) inclusion of adults who lack capacity to consent to research in clinical trials in the UK. Trials 2019 Aug 5;20(1):474. [doi: 10.1186/s13063-019-3603-1] [Medline: 31382999]
36. Shepherd V. An under-represented and underserved population in trials: methodological, structural, and systemic barriers to the inclusion of adults lacking capacity to consent. Trials 2020 May 29;21(1):445. [doi: 10.1186/s13063-020-04406-y] [Medline: 32471488]
37. Gilbert T, Bosquet A, Thomas-Antérion C, Bonnefoy M, Le Saux O. Assessing capacity to consent for research in cognitively impaired older patients. Clin Interv Aging 2017;12:1553-1563. [doi: 10.2147/CIA.S141905] [Medline: 29026293]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
EPR: electronic patient record
MDT: multidisciplinary team
NHS: National Health Service
NIHR: National Institute for Health and Care Research
PPIE: patient and public involvement and engagement
TIDieR: Template for Intervention Description and Replication

XSL•FO
RenderX

Edited by A Stone; submitted 27.Jun.2025; peer-reviewed by CM Chen, C Subbe; accepted 11.Nov.2025; published 06.Jan.2026.

Please cite as:
Averill P, Lear R, Odedra R, Long S, Taylor A, Charville PJ, Fernandes J, Ekeogu U, Leombruno J, Ellis S, Mayer E
Acceptability, Feasibility, and Perceived Effectiveness of Video-Based Patient Records for Supporting Care Delivery to Older Adults With Frailty: Nonrandomized Mixed Methods Pilot Study
J Med Internet Res 2026;28:e77318
URL: https://www.jmir.org/2026/1/e77318
doi:10.2196/77318

© Phoebe Averill, Rachael Lear, Ricky Odedra, Susannah Long, Alex Taylor, Pi-Jung Charville, Jessica Fernandes, Uzoamaka Ekeogu, Jessica Leombruno, Sophia Ellis, Erik Mayer. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Associations and Pathways Between Online Health Information–Seeking Behavior and Patient Adherence: Cross-Sectional Study

Zhenyu Xu[1*], MMed; Jingjing Jia[2*], MM; Chaofan Wu[3], MM

[1]Department of Urology, The First Affiliated Hospital of Lishui University, Lishui People's Hospital, Lishui, Zhejiang, China

[2]School of Medicine and Health Management, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China

[3]Department of Personnel, The First Affiliated Hospital of Lishui University, Lishui People's Hospital, 15 Dazhong Street, Lishui City, Lishui, Zhejiang, China

[*]these authors contributed equally

Corresponding Author:
Chaofan Wu, MM
Department of Personnel, The First Affiliated Hospital of Lishui University, Lishui People's Hospital, 15 Dazhong Street, Lishui City, Lishui, Zhejiang, China

Abstract

Background: The widespread adoption of the internet has established online health information–seeking behavior (OHISB) as a primary channel for public health knowledge acquisition, potentially influencing patient adherence behaviors and physician-patient dynamics. However, the underlying pathways, particularly the role of physician-patient communication efficacy and the differential impact of various digital platforms, remain underexplored, especially among rural populations.

Objective: This study examined the association between OHISB and patient adherence among rural residents in China, with a specific focus on the mediating role of physician-patient communication efficacy and the moderating roles of different platform types.

Methods: A cross-sectional survey was conducted from June 2023 to October 2024 using multistage stratified sampling across 6 Chinese provinces. Participants were rural residents aged 18 to 70 years with recent health care experiences. Data from 7004 valid questionnaires were analyzed. A fixed-effects model assessed the primary association, with robustness checked via least absolute shrinkage and selection operator regression. Mediation analysis using the bootstrap method examined the indirect association through physician-patient communication efficacy, and interaction terms tested the moderating effects of platform type (internet hospitals, professional platforms, WeChat accounts, short video apps, and search engines).

Results: OHISB showed a significant positive direct association with patient adherence (β=0.260; *P*<.001). Physician-patient communication efficacy exhibited a significant negative indirect association with patient adherence (β=–0.026; *P*<.001), accounting for 9.29% of the total association. Platform type significantly moderated this association: internet hospitals (β=0.099; *P*=.04), professional platforms (β=0.081; *P*=.04), and WeChat accounts (β=0.032; *P*=.03) enhanced the positive association between OHISB and patient adherence, whereas short video platforms (β=–0.034; *P*=.006) and search engines (β=–0.204; *P*<.001) weakened it.

Conclusions: Online health information seeking among rural residents was directly associated with better patient adherence, but this benefit was partially attenuated by a negative indirect association through reduced physician-patient communication efficacy. The association between OHISB and adherence varied significantly by platform type. This finding suggests the need for digital health equity strategies, interventions to improve communication efficacy and health literacy, and graded management of health information platforms.

(J Med Internet Res 2026;28:e91115) doi:10.2196/91115

KEYWORDS

online health information–seeking behavior; patient adherence; physician-patient communication efficacy; digital health platforms; rural residents; health literacy

XSL•FO
RenderX

Introduction

Driven by widespread internet adoption, rural residents' daily lives are rapidly undergoing digital transformation. By December 2024, internet users in rural China reached 326 million, with a penetration rate of 65.6%, representing a 28.1–percentage point increase since 2018 [1]. Concurrently, online health information seeking has emerged as the primary channel for public acquisition of medical knowledge. This shift has reconfigured patients' disease awareness and decision-making patterns while also reshaping traditional physician-patient dynamics [2]. Patient adherence, a core metric for evaluating health care outcomes, is increasingly determined by the dynamic interaction between patient autonomy and health literacy rather than physician authority [3].

However, current research predominantly examines the direct pathways from health information seeking to adherence, overlooking the mediating role of physician-patient communication efficacy. In the digital era, patients' prior knowledge acquired through online sources may compromise communication effectiveness, potentially reducing interaction quality and indirectly impairing adherence [4]. Moreover, heterogeneous health information platforms (eg, professional medical portals, short video platforms, and search engines) may exert differential moderating effects on the information seeking–adherence relationship due to disparities in information quality, credibility, and user experience [5].

To address this gap, we developed a conceptual framework grounded in the health belief model (HBM) [6,7]. This framework posits that online health information–seeking behavior (OHISB) may enhance patient adherence directly by strengthening perceived benefits and self-efficacy. Simultaneously, OHISB may trigger information overload and uncertainty—functioning as perceived barriers within the HBM framework—which may undermine physician-patient communication efficacy and, thereby, exert a negative indirect effect on adherence. Platform type is theorized to moderate these associations by shaping the quality and credibility of the information encountered. Specifically, platforms with rigorous content curation (eg, hospital-affiliated portals) are hypothesized to strengthen the positive OHISB-adherence association, whereas platforms characterized by fragmented, entertainment-oriented, or low-credibility content (eg, short video apps and search engines) are hypothesized to weaken or reverse this association. Accordingly, the primary objective of this study was to examine the association between OHISB and patient adherence among rural Chinese residents, with a specific focus on the mediating role of physician-patient communication efficacy and the moderating role of platform type.

Methods

Study Design and Participants

This cross-sectional study was conducted from June 19, 2023, to October 20, 2024. A multistage stratified sampling method was used. First, 2 provinces (autonomous regions and municipalities) were selected from each of the eastern, central, and western regions of China to ensure national geographical representativeness. The surveyed provinces included Gansu, Chongqing, Hubei, Henan, Zhejiang, and Shandong. Within each province, counties were stratified by economic development level; the top and bottom 3% of counties were selected as sampling units, totaling 12 counties.

Trained investigators conducted door-to-door interviews with permanent residents (defined as individuals who had resided in the household for ≥6 months in the previous year), administering paper questionnaires to individuals aged 18 to 70 years. A standardized training manual and interview script were used (Multimedia Appendix 1). All investigators underwent a 2-day centralized training session covering research ethics, questionnaire explanation, and interview techniques. Role-playing and pilot tests were conducted to ensure uniform understanding and delivery of questions. Regular supervision (weekly team meetings to review completed questionnaires) and random callback checks (10% of participants were recontacted via telephone within 1 week to verify responses) were performed during data collection to maintain consistency and quality. The questionnaire was interviewer administered: investigators read each item aloud and recorded responses to minimize literacy-related barriers.

Each questionnaire included a written informed consent form. The study excluded minors. Of the 7200 returned questionnaires, 7004 (97.3%) valid responses were collected after excluding incomplete or inconsistent submissions. Inclusion criteria were (1) medical resource use within the previous 2 years, (2) language communication ability, (3) no major cognitive or mental disorder, and (4) voluntary participation.

Variables

Independent Variable

The independent variable was OHISB. Consistent with the conceptual framework by Mirzaei et al [8], this construct encompasses active retrieval, browsing, and acquisition of health-related information via digital platforms. It was measured via self-reported frequency of internet-based health information seeking using computers or mobile devices. Responses were recorded on a 3-point Likert-type scale: 0="never," 1="occasionally" (1-2 times per month), and 2="frequently" (≥3 times per month). Response categories were defined in the questionnaire with concrete anchors, and interviewers provided standardized examples (eg, "searching for information about symptoms, treatments, or medications online"). This parsimonious measure was selected for its methodological appropriateness in a large-scale survey with rural residents [9], effectively balancing scientific rigor with practical feasibility by capturing essential behavioral variation while minimizing cognitive load and ensuring high response rates.

Dependent Variable

Patient adherence constituted the dependent variable. Adopting the World Health Organization's conceptualization [10], this construct was operationalized as the extent of congruence between patient behaviors and health care providers' recommendations. It was measured using a single item asking respondents about their willingness to modify health-related behaviors in accordance with clinical guidance. Responses were

XSL•FO
RenderX

recorded on a 5-point scale (1="not at all willing," 2="slightly willing," 3="moderately willing," 4="very willing," and 5="completely willing").

Mediating Variable

Physician-patient communication efficacy served as the mediating variable. Measurement used a 5-point scale to assess patients' self-perceived competence in articulating needs and understanding medical advice [11]. This scale comprised 2 separate items: one for "articulating health concerns to the doctor" and one for "understanding the doctor's explanations and recommendations." Scores were averaged to produce a composite measure. To ensure the relevance and psychometric robustness for the rural population, the scale underwent formal adaptation and validation. Content validity was evaluated by 3 independent experts in health communication (2 PhD-level health communication researchers and 1 practicing physician with ≥10 years of clinical experience), yielding a content validity index greater than 0.85. A pilot study with a separate sample of 30 rural residents (excluded from the final analysis) assessed structural validity and reliability. Exploratory factor analysis of the pilot data supported the presumed unidimensional structure (Kaiser-Meyer-Olkin test=0.805; Bartlett test of sphericity: $P<.001$). In the main study sample (N=7004), the scale demonstrated excellent internal consistency (Cronbach $\alpha=0.915$).

Moderating Variable

The moderating variable was platform type for health information seeking. This construct was operationalized by categorizing respondents' primary digital sources into five mutually exclusive types: (1) public hospital–affiliated online portals, (2) specialized medical platforms (eg, Haodf or Ping An Good Doctor), (3) short video apps (eg, Douyin, Kuaishou, or RedNote), (4) WeChat official accounts, and (5) search engines (eg, Baidu and Quark).

Covariates

Potential confounding variables were controlled based on established methodological approaches [12], encompassing age, educational attainment, annual household income, employment status, chronic disease status, self-rated health, weekly breakfast frequency, smoking status, alcohol consumption, exercise frequency, and subjectively measured sleep quality.

Statistical Methods

Data entry and database development were performed using EpiData (version 4.6; EpiData Association) with double entry verification. All statistical analyses were conducted in R (version 4.3.1; R Foundation for Statistical Computing) using the *lme4* package for fixed-effects modeling and *glmnet* for least absolute shrinkage and selection operator (LASSO) regression, with statistical significance defined as a 2-sided P value of less than .05. Baseline regression analyses incorporated regional fixed effects, and SEs were clustered at the county level to account for intracounty correlation, providing more conservative statistical inferences. Robustness checks were implemented through LASSO regression, a machine learning method used for variable selection and shrinkage to prevent overfitting in models with many potential covariates. LASSO helps identify a parsimonious set of predictors by shrinking less important coefficients toward zero, thereby addressing potential multicollinearity.

To examine whether physician-patient communication efficacy mediated the association between OHISB and patient adherence, we conducted a mediation analysis using the bootstrap method with 5000 resamples [13]. The indirect effect (product of the coefficient for OHISB to efficacy and efficacy to adherence) and its 95% CI were calculated. Mediation was considered statistically significant if the CI did not include 0.

Ethical Considerations

This study was approved by the ethics committee of Tongji Medical College of Huazhong University of Science and Technology (IORG0003571). All procedures performed were in accordance with the ethical standards of the institutional research committee and with the 1964 Declaration of Helsinki and its later amendments. Participation was entirely voluntary and anonymous; no personally identifiable information was collected. All data were kept strictly confidential and used solely for research purposes. Electronic informed consent was obtained from all individual participants included in the study. No financial compensation was provided to the participants.

Results

Descriptive Analysis

The final sample included 7004 rural participants. OHISB distribution was 12.4% (867/7004) frequent users, 18.3% (1280/7004) occasional users, and 69.3% (4857/7004) nonusers. Mean scores were 3.43 (SD 1.41) for patient adherence and 4.32 (SD 0.98) for physician-patient communication efficacy. Regarding health information platforms (percentages calculated among OHISB users: 2147/7004, 30.7%), short video apps showed the highest use (1597/2147, 74.4%), followed by WeChat official accounts (575/2147, 26.8%), whereas specialized platforms demonstrated minimal adoption (hospital-affiliated portals: 39/2147, 1.8%; dedicated medical platforms: 47/2147, 2.2%).

Table 1 presents the demographic and health-related characteristics of the 7004 participants, including age, educational attainment, income, employment status, region, chronic disease status, self-rated health, and lifestyle factors (breakfast frequency, smoking, drinking, exercise, and sleep quality).

XSL•FO
RenderX

Table . Demographic and health-related characteristics of the participants (N=7004).

Variable and category	Values
Age (y), mean (SD)	54.72 (23.81)
Educational level (1-5)[a], mean (SD)	2.48 (1.13)
Annual income (US $)[b], mean (SD)	2493.73 (4499.05)
Employment status, n (%)	
Unemployed	3382 (48.3)
Employed (including farming)	3622 (51.7)
Region, n (%)	
Eastern	2394 (34.2)
Central	2552 (36.4)
Western	2058 (29.4)
Chronic diseases, n (%)	
No	4256 (60.8)
Yes	2748 (39.2)
EQ-5D score (−0.149 to 1.000), mean (SD)	0.91 (0.17)
Frequency of breakfast (1-5)[c], mean (SD)	1.31 (0.99)
Smoking status (1-3)[d], mean (SD)	1.81 (0.39)
Drinking status (1-3)[e], mean (SD)	1.81 (0.39)
Weekly physical exercise sessions, mean (SD)	3.59 (3.45)
Sleep quality (1-5)[f], mean (SD)	2.20 (1.17)

[a]1=never attended school; 2=primary school; 3=junior high school; 4=high school; 5=college degree or higher.

[b]Converted at the average exchange rate of US $1=CN ¥7.1383, mean of the 2023 and 2024 annual average rates from the Federal Reserve.

[c]1=daily; 2=4 - 6 times per week; 3=1 - 2 times per week; 4=once per week; 5=never.

[d]1=smoker; 2=quit smoking; 3=nonsmoker.

[e]1=drinks alcohol; 2=has given up alcohol; 3=does not drink alcohol.

[f]1=verygood; 2=good; 3=average; 4=poor; 5=very poor.

Table 2 summarizes the distribution of the core study variables, including OHISB frequency, patient adherence scores, physician-patient communication efficacy scores, and the use of different platform types for health information seeking (percentages calculated among OHISB users in this case).

XSL•FO
RenderX

Table . Distribution of core study variables(N=7004).

Variable and category	Values
OHISB[a] frequency, n (%)	
Never	4857 (69.3)
Occasionally	1280 (18.3)
Frequently	867 (12.4)
Patient adherence (1-5), mean (SD)	3.43 (1.41)
Physician-patient communication efficacy (1-5), mean (SD)	4.32 (0.98)
Platform types for health information seeking (n=2147), n (%)[b]	
Public hospital–affiliated online portals	39 (1.8)
Specialized medical platforms	47 (2.2)
Shortvideo apps	1597 (74.4)
WeChat official accounts	575 (26.8)
Search engines	240 (11.2)

[a]OHISB: online health information–seeking behavior.

[b]Percentages may add up to more than 100 due to multiple platform use.

Association Between OHISB and Patient Adherence

As shown in Table 3, OHISB was significantly and positively associated with patient adherence across all models.

Table . Association between online health information–seeking behavior (OHISB) and patient adherence (N=7004)[a].

Variable	Model 1[b], β (SE)	Model 2[c], β (SE)	Model 3[d], β (SE)	Model 4[e], β (SE)
OHISB	0.356[f] (0.024)	0.354[f] (0.024)	0.260[f] (0.024)	0.268[f] (0.024)
Age	—[g]	—	0.003[f] (0.001)	0.005[f] (0.001)
Educational attainment	—	—	0.116[f] (0.017)	0.117[f] (0.016)
Annual income	—	—	0.027[f] (0.005)	—
Employment status	—	—	0.113[h] (0.041)	—
Chronic disease	—	—	–0.067 (0.038)	—
Health (EQ-5D)	—	—	0.962[f] (0.106)	0.876[f] (0.102)
Breakfast frequency	—	—	–0.103[f] (0.017)	–0.101[f] (0.017)
Smoking	—	—	–0.029 (0.048)	—
Drinking	—	—	0.139[h] (0.048)	—
Exercise frequency	—	—	0.040[f] (0.005)	0.042[f] (0.005)
Sleep quality	—	—	0.020 (0.015)	—

[a]Regional fixed effects were applied at the province level for models 2, 3, and 4. Statistical significance was assessed using linear regression with 2-tailed *t* tests.

[b]Unadjusted; R^2=0.496; adjusted R^2=0.496; root mean squared error (RMSE)=1.39.

[c]Regional fixed effects; R^2=0.532; adjusted R^2=0.520; RMSE=1.38.

[d]All covariates (full list in the Methods section); R^2=0.587; adjusted R^2=0.573; RMSE=1.35.

[e]Least absolute shrinkage and selection operator–selected covariates; R^2=0.542; adjusted R^2=0.541; RMSE=1.35.

[f]$P<.001$.

[g]Variable not included in the model.

[h]$P<.01$.

XSL•FO
RenderX

In the unadjusted model (model 1), the association was strong (β=0.356; *P*<.001). Incorporating regional fixed effects in model 2 yielded consistent estimates (β=0.354; *P*<.001), confirming model robustness. Regional fixed effects were applied at the province level. Model 3 further adjusted for covariates, including age, income, and educational attainment, maintaining statistical significance (β=0.260; *P*<.001). This corresponds to a 0.260-unit increase in adherence per 1-unit increment in OHISB frequency. Significant covariates associated with higher adherence included older age, higher educational attainment, better self-rated health, consistent breakfast habits, and greater exercise frequency.

Model 4 used LASSO regression for robustness checks. The regularization procedure retained significant covariates: age, educational level, health (EQ-5D), weekly frequency of breakfast, and exercise frequency. OHISB showed a consistent positive association with patient adherence (β=0.268; *P*<.001) relative to fixed-effects models. Control variable coefficients maintained directional and statistical alignment with baseline estimates.

The results of the 4 models demonstrated the robustness of the association between OHISB and patient adherence.

Mediation Analysis

To examine the mechanism through which OHISB may be associated with patient adherence, a mediation analysis was conducted using the bootstrap method with 5000 resamples. The results, presented in Table 4, delineate the direct, indirect, and total associations.

Table . Mediation analysis of physician-patient communication efficacy[a].

Path	Estimate (β; SE; 95% CI)	*P* value
a: OHISB[b]–efficacy	−0.083 (0.018; –0.117 to –0.048)	<.001
b: efficacy–adherence	0.310 (0.016; 0.278 to 0.342)	<.001
c’: direct (OHISB–adherence)	0.301 (0.024; 0.255 to 0.348)	<.001
Indirect (a × b)	−0.026 (0.007; –0.039 to –0.013)	<.001
Total effect (c)	0.276 (0.024; 0.228 to 0.324)	<.001

[a]Bootstrap resamples=5000; proportion of total effect mediated: 9.29%; *P* value for indirect effect based on bootstrap percentile method.

[b]OHISB: online health information–seeking behavior.

Path *a* (OHISB–communication efficacy) was negative and significant (β=−0.083; *P*<.001, 95% CI −0.117 to –0.048). Path *b* (communication efficacy–adherence) was positive and significant (β=0.310; *P*<.001, 95% CI 0.278-0.342). The indirect effect (*a* × *b*) was negative and significant (β=−0.026; *P*<.001, 95% CI −0.039 to −0.013), accounting for 9.29% of the total association. The direct effect (*c*’: OHISB–adherence, controlling for efficacy) remained positive and significant (β=0.301; *P*<.001, 95% CI 0.255-0.348), indicating inconsistent mediation.

This indirect association accounted for approximately 9.29% of the total association between OHISB and adherence, indicating an inconsistent mediation pattern (ie, the indirect association opposes the direction of the direct positive association) [13,14]. Figure 1 shows the mediation model with standardized path coefficients.

Figure 1. Mediation model of physician-patient communication efficacy in the association between online health information–seeking behavior (OHISB) and patient adherence.

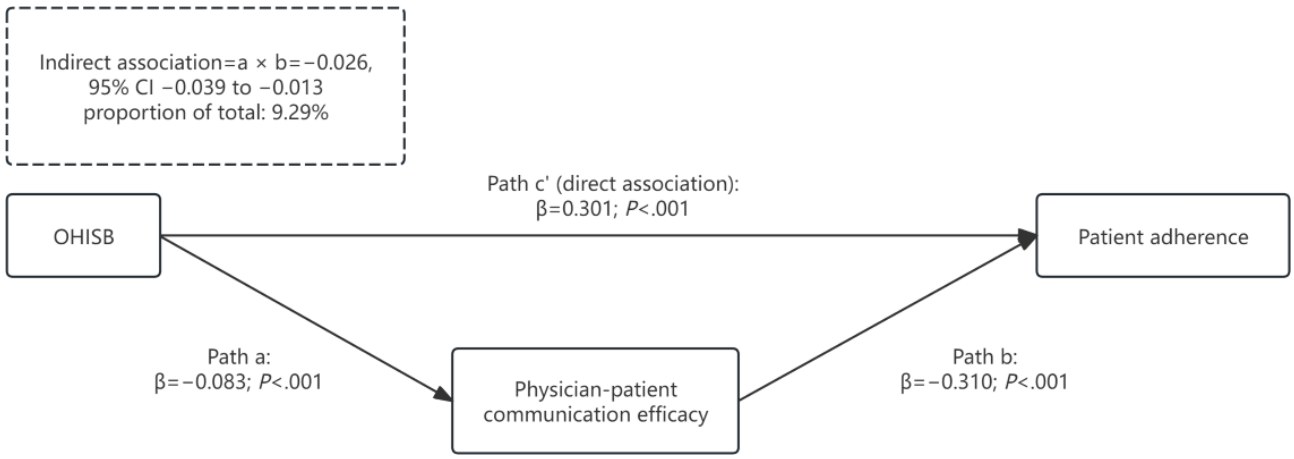

Moderation by Platform Type

All interaction terms attained statistical significance (Table 5). Positive moderating associations were observed for hospital-affiliated portals (β=0.099; *P*=.04), specialized medical platforms (β=0.081; *P*=.04), and WeChat official accounts (β=0.032; *P*=.03), with hospital-affiliated portals showing the largest effect size. Negative moderating associations were observed for short video apps (β=−0.034; *P*=.006) and search engines (β=−0.204; *P*<.001), where search engines demonstrated the most pronounced attenuating pattern.

XSL•FO
RenderX

Table . Moderating roles of platform type on the relationship between online health information–seeking behavior (OHISB) and adherence (N=7004)[a].

Variable	Model 1[b], β (SE)	Model 2[c], β (SE)	Model 3[d], β (SE)	Model 4[e], β (SE)	Model 5[f], β (SE)
OHISB	0.061[g] (0.006)	0.063[g] (0.006)	0.022[h] (0.011)	0.055[g] (0.007)	0.062[g] (0.006)
Hospital-affiliated portals	0.013 (0.184)	__[i]	—	—	—
OHISB × hospital-affiliated portals	0.099[h] (0.044)	—	—	—	—
Specialized medical platforms	—	0.052 (0.132)	—	—	—
OHISB × specialized medical platforms	—	0.081[h] (0.035)	—	—	—
Short video apps	—	—	−0.265[g] (0.043)	—	—
OHISB × short video apps	—	—	−0.034[j] (0.012)	—	—
WeChat official accounts	—	—	—	0.074 (0.045)	—
OHISB × WeChat official accounts	—	—	—	0.032[h] (0.013)	—
Search engines	—	—	—	—	0.298[g] (0.066)
OHISB × search engines	—	—	—	—	−0.204[g] (0.007)

[a]All models control for the covariates listed in Table 3. The reference category for each interaction term is nonuse of the respective platform type.
[b]R^2=0.642; adjusted R^2=0.625; root mean squared error (RMSE)=0.66.
[c]R^2=0.637; adjusted R^2=0.615; RMSE=0.66.
[d]R^2=0.592; adjusted R^2=0.583; RMSE=0.64.
[e]R^2=0.583; adjusted R^2=0.577; RMSE=0.66.
[f]R^2=0.639; adjusted R^2=0.621; RMSE=0.65.
[g]P<.001.
[h]P<.05.
[i]Variable not included in the model.
[j]P<.01.

The reference category for each interaction term was nonuse of the respective platform type. For example, the coefficient for OHISB × short video apps (β=−0.034) indicates that, compared with nonusers of short video platforms, users of these platforms showed a weaker positive association between OHISB and adherence.

Discussion

This study revealed a dual-process association between OHISB and patient adherence among rural Chinese residents moderated by the type of digital platform used.

Positive Association Between OHISB and Adherence Amid Low Adoption Rates

OHISB frequency was positively associated with patient adherence, consistent with prior research [15,16]. This finding supports the HBM [17], suggesting that information seeking may enhance perceived treatment benefits and self-efficacy, directly promoting adherence intentions. The adoption of OHISB is significantly lower in rural populations compared with reported rates in urban settings, and the joint use of hospital-affiliated platforms and professional health information platforms is relatively low in these populations [2,18]. This disparity likely stems from overlapping digital divides: deficient rural infrastructure, technological barriers among older adults, and terminology-dense content that raises cognitive thresholds [19]. The high health literacy demands of professional platforms further limit their potential benefits [20]. Future initiatives must optimize platform usability and deliver comprehensible, tailored health messaging to rural populations.

The Dual-Process Model: Unpacking the Negative Indirect Association

Our analysis suggests a dual-process pattern: while OHISB was directly and positively associated with adherence, a simultaneous negative indirect association was observed through reduced physician-patient communication efficacy [21]. This finding must be interpreted with nuance; it does not imply that online information seeking is inherently harmful [22]. Rather, it highlights a critical contingency: the net benefit of OHISB may be weaker when the information environment or the clinical encounter creates specific risks. Reduced communication efficacy may be more likely under circumstances where (1)

XSL•FO
RenderX

information from online sources is fragmented, contradictory, or of low credibility, which may contribute to patient confusion or misplaced confidence [23]; (2) patients experience "information overload," a state that increases anxiety and hinders effective information processing during consultations [23]; or (3) a power dynamic is activated wherein clinicians perceive patient-initiated information as challenging their authority, prompting defensive communication [24]. For health system design, these findings suggest that promoting patient empowerment through digital tools should be coupled with parallel interventions to improve information quality; scaffold patient-clinician dialogue; and train clinicians in collaborative, information-integrated consultation styles. The indirect effect, while statistically significant, accounted for only 9.29% of the total association, suggesting that the direct positive association predominated.

Platform-Specific Differences in the OHISB-Adherence Association

The moderating effects of platform types exhibited significant heterogeneity. Hospital-affiliated portals, specialized medical platforms, and WeChat official accounts strengthened the positive association between OHISB and patient adherence, whereas short video apps and search engines weakened it, corroborating previous findings [25]. Mechanistically, this divergence likely stems from differential information credibility, content quality, and interaction paradigms. Hospital-affiliated and specialized platforms enforce rigorous content curation to deliver evidence-based structured information congruent with clinical standards, enabling patients to integrate information into shared decision-making [26]. WeChat official accounts further facilitate sustained physician-patient interaction that contextualizes health information. Conversely, despite dominating use in this study (1597/2147, 74.4%), short video and search platforms propagate fragmentary, entertainment-focused content (eg, treatment claims emphasizing efficacy while omitting risks). Their recommendation algorithms may create echo chambers that increase exposure to homogeneous viewpoints and amplify cognitive biases [24]. Concurrently, their attention-grabbing formats trivialize complex health topics, whereas inconsistent information quality exposes users to contradictory perspectives, generating information overload [22]. This complexity may induce decisional uncertainty that ultimately attenuates the benefits of information seeking.

However, the observed platform-specific differences should be interpreted with caution as they may partly reflect unmeasured confounding factors. For instance, users of different platforms may differ systematically in digital literacy, health literacy, access to alternative information sources, or underlying health status. Short video platforms, which dominate rural use, may attract users with lower digital literacy or less engagement with formal health care, whereas those who actively seek out hospital-affiliated portals may have higher health literacy or more complex health needs. Such self-selection could contribute to the differential associations observed. Future research should measure and adjust for these potential confounders to isolate the unique moderating effects of platform characteristics.

Limitations

This study has several limitations. First, its cross-sectional design precludes causal inference; longitudinal or experimental studies are needed to establish temporal sequences and strengthen causal claims. Consequently, terms such as "effect," "impact," and "pathway" are used descriptively to reflect statistical associations rather than causal relationships. Second, the use of a single-item, 3-point scale to measure OHISB, while practical, lacks granularity and conceptual specificity. Third, patient adherence was measured using a single item assessing behavioral intention rather than actual behavior, which may overestimate adherence due to social desirability bias. Test-retest reliability was not assessed, which represents an additional limitation. Future work should use validated, multidimensional scales to capture this complexity and provide a more nuanced understanding of the OHISB and patient adherence construct.

Conclusions

Drawing on cross-sectional survey data collected during 2023 and 2024, this study investigated OHISB and patient adherence, with particular emphasis on the mediating role of physician-patient communication efficacy and moderating associations with platform type. Three principal findings emerged.

First, OHISB frequency was positively associated with patient adherence, wherein higher seeking frequency corresponded to greater adherence. Second, physician-patient communication efficacy was observed as a negative indirect association whereby OHISB was associated with reduced communication efficacy, which in turn was associated with lower adherence—partially offsetting the direct positive association. However, this indirect association accounted for only 9.29% of the total association, indicating that the direct positive association predominated. Third, the association between information seeking and adherence varied significantly by platform type: hospital-affiliated portals, specialized medical platforms, and WeChat official accounts showed positive moderating associations, whereas short video apps and search engines showed attenuating associations that diminished overall empowerment returns.

These findings align with the HBM. The direct positive association supports the HBM's premise that information seeking enhances perceived benefits and self-efficacy. The negative indirect association is consistent with the concept of perceived barriers: information overload and uncertainty may impede effective patient-clinician interaction. Platform-specific moderation suggests that information quality and credibility shape whether seeking behavior reduces or amplifies these barriers.

These findings suggest several implications for policy and practice, although these should be considered suggestive rather than definitive given the observational study design. First, to leverage the direct positive association and address low adoption rates, digital health equity initiatives should include the co-design of culturally and linguistically adapted health information content (eg, developing plain-language, dialect-compatible health materials for low-literacy users) with

XSL•FO
RenderX

rural residents. Second, to address the negative indirect association, health systems might consider dual-focused interventions: (1) for patients, brief, scalable "health information literacy" modules to build skills for critical evaluation of online content; and (2) for clinicians, training in communication strategies to acknowledge, discuss, and integrate patient-sourced information constructively. Third, to respond to the heterogeneous platform effects, a graded regulatory framework could be considered, including accreditation incentives for authoritative content on professional platforms and algorithmic governance measures for short video and search platforms (eg, mandatory source disclosure for health claims and downgrading of unverified medical advice).

These findings and implications are grounded in a rural Chinese context and may not be directly generalizable to other settings. Future research should use longitudinal designs, validated multidimensional measures, and diverse populations to extend and refine these observations.

Acknowledgments

The authors would like to express their sincere gratitude to the facilities involved in this study for their support. All author contributions have been detailed in the Authors' Contributions section.

Funding

This research was funded by the National Natural Science Foundation of China (grant 72104086). The funder had no role in the study design; data collection, analysis, and interpretation; or manuscript preparation.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to restrictions specified in the ethics approval but are available from the corresponding author on reasonable request and with permission from the principal investigator.

Authors' Contributions

Conceptualization: CW (lead)
Data curation: ZX (lead), JJ (supporting)
Formal analysis: ZX (lead)
Investigation: JJ (lead)
Methodology: ZX (lead)
Project administration: JJ (lead), CW (supporting)
Supervision: CW (lead)
Validation: JJ (lead)
Writing—original draft: ZX (lead)
Writing—review and editing: CW (lead), JJ (supporting), ZX (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interviewer training manual and interview script.
[PDF File, 292 KB - jmir_v28i1e91115_app1.pdf]

References

1. The 55th statistical report on China's internet development. China Internet Network Information Center. 2025. URL: https://www3.cnnic.cn/n4/2025/0117/c88-11229.html [accessed 2025-03-17]
2. Li H, Li D, Zhai M, Lin L, Cao Z. Associations among online health information seeking behavior, online health information perception, and health service utilization: cross-sectional study. J Med Internet Res 2025 Mar 14;27:e66683. [doi: 10.2196/66683] [Medline: 40085841]
3. Panahi S, Rathi N, Hurley J, Sundrud J, Lucero M, Kamimura A. Patient adherence to health care provider recommendations and medication among free clinic patients. J Patient Exp 2022;9:23743735221077523. [doi: 10.1177/23743735221077523] [Medline: 35155751]
4. Klerings I, Weinhandl AS, Thaler KJ. Information overload in healthcare: too much of a good thing? Z Evid Fortbild Qual Gesundhwes 2015;109(4-5):285-290. [doi: 10.1016/j.zefq.2015.06.005] [Medline: 26354128]
5. Guan JL, Xia SH, Zhao K, et al. Videos in short-video sharing platforms as sources of information on colorectal polyps: cross-sectional content analysis study. J Med Internet Res 2024 Oct 29;26:e51655. [doi: 10.2196/51655] [Medline: 39470708]

XSL•FO
RenderX

6. Rosenstock IM, Strecher VJ, Becker MH. Social learning theory and the Health Belief Model. Health Educ Q 1988;15(2):175-183. [doi: 10.1177/109019818801500203] [Medline: 3378902]
7. Janz NK, Becker MH. The Health Belief Model: a decade later. Health Educ Q 1984;11(1):1-47. [doi: 10.1177/109019818401100101] [Medline: 6392204]
8. Mirzaei A, Aslani P, Luca EJ, Schneider CR. Predictors of health information-seeking behavior: systematic literature review and network analysis. J Med Internet Res 2021 Jul 2;23(7):e21680. [doi: 10.2196/21680] [Medline: 33979776]
9. Xiong Z, Zhang L, Li Z, Xu W, Zhang Y, Ye T. Frequency of online health information seeking and types of information sought among the general Chinese population: cross-sectional study. J Med Internet Res 2021 Dec 2;23(12):e30855. [doi: 10.2196/30855] [Medline: 34860676]
10. Brown MT, Bussell JK. Medication adherence: WHO cares? Mayo Clin Proc 2011 Apr;86(4):304-314. [doi: 10.4065/mcp.2010.0575] [Medline: 21389250]
11. Sharkiya SH. Quality communication can improve patient-centred health outcomes among older patients: a rapid review. BMC Health Serv Res 2023 Aug 22;23(1):886. [doi: 10.1186/s12913-023-09869-8] [Medline: 37608376]
12. Jia X, Pang Y, Liu LS. Online health information seeking behavior: a systematic review. Healthcare (Basel) 2021 Dec 16;9(12):1740. [doi: 10.3390/healthcare9121740] [Medline: 34946466]
13. MacKinnon DP, Fairchild AJ, Fritz MS. Mediation analysis. Annu Rev Psychol 2007;58:593-614. [doi: 10.1146/annurev.psych.58.110405.085542] [Medline: 16968208]
14. Tönnies T, Schlesinger S, Lang A, Kuss O. Mediation analysis in medical research. Dtsch Arztebl Int 2023 Oct 13;120(41):681-687. [doi: 10.3238/arztebl.m2023.0175] [Medline: 37584228]
15. AlShayban DM, Naqvi AA, Alhumaid O, et al. Association of disease knowledge and medication adherence among out-patients with type 2 diabetes mellitus in Khobar, Saudi Arabia. Front Pharmacol 2020;11:60. [doi: 10.3389/fphar.2020.00060] [Medline: 32153397]
16. Kim MS, Kim SH. Health information-seeking behavior in patients with coronary artery disease: activating methods. PLoS One 2024;19(4):e0300755. [doi: 10.1371/journal.pone.0300755] [Medline: 38630654]
17. Zhao YC, Zhao M, Song S. Online health information seeking among patients with chronic conditions: integrating the Health Belief Model and social support theory. J Med Internet Res 2022 Nov 2;24(11):e42447. [doi: 10.2196/42447] [Medline: 36322124]
18. Liu D, Yang S, Cheng CY, Cai L, Su J. Online health information seeking, eHealth literacy, and health behaviors among Chinese internet users: cross-sectional survey study. J Med Internet Res 2024 Oct 18;26:e54135. [doi: 10.2196/54135] [Medline: 39423374]
19. Bauerly BC, McCord RF, Hulkower R, Pepin D. Broadband access as a public health issue: the role of law in expanding broadband access and connecting underserved communities for better health outcomes. J Law Med Ethics 2019 Jun;47(2_suppl):39-42. [doi: 10.1177/1073110519857314] [Medline: 31298126]
20. Wammes JJ, Frederix G, Govaert P, et al. Case-studies of displacement effects in Dutch hospital care. BMC Health Serv Res 2020 Mar 30;20(1):263. [doi: 10.1186/s12913-020-05086-9] [Medline: 32228590]
21. Zhong F, Gu C. The impact of health information echo chambers on older adults avoidance behavior: the mediating role of information fatigue and the moderating role of trait mindfulness. Front Psychol 2024;15:1412515. [doi: 10.3389/fpsyg.2024.1412515] [Medline: 39228876]
22. Zhong L, Cao J, Xue F. The paradox of convenience: how information overload in mHealth apps leads to medical service overuse. Front Public Health 2024;12:1408998. [doi: 10.3389/fpubh.2024.1408998] [Medline: 39668954]
23. Lim HM, Dunn AG, Lim JR, Abdullah A, Ng CJ. Association between online health information-seeking and medication adherence: a systematic review and meta-analysis. Digit Health 2022;8:20552076221097784. [doi: 10.1177/20552076221097784] [Medline: 35586836]
24. Mumtaz H, Riaz MH, Wajid H, et al. Current challenges and potential solutions to the use of digital health technologies in evidence generation: a narrative review. Front Digit Health 2023;5:1203945. [doi: 10.3389/fdgth.2023.1203945] [Medline: 37840685]
25. Wu D, Lowry PB, Zhang D, Tao Y. Patient trust in physicians matters-understanding the role of a mobile patient education system and patient-physician communication in improving patient adherence behavior: field study. J Med Internet Res 2022 Dec 20;24(12):e42941. [doi: 10.2196/42941] [Medline: 36538351]
26. Hellstrand Tang U, Smith F, Karilampi UL, Gremyr A. Exploring the role of complexity in health care technology bottom-up innovations: multiple-case study using the nonadoption, abandonment, scale-up, spread, and sustainability complexity assessment tool. JMIR Hum Factors 2024 Apr 26;11:e50889. [doi: 10.2196/50889] [Medline: 38669076]

Abbreviations

HBM: health belief model
LASSO: least absolute shrinkage and selection operator
OHISB: online health information–seeking behavior

XSL•FO
RenderX

Edited by M Balcarras; submitted 12.Jan.2026; peer-reviewed by BK Slavych, C Liu; revised version received 03.May.2026; accepted 04.May.2026; published 04.Jun.2026.

Please cite as:
Xu Z, Jia J, Wu C
Associations and Pathways Between Online Health Information–Seeking Behavior and Patient Adherence: Cross-Sectional Study
J Med Internet Res 2026;28:e91115
URL: https://www.jmir.org/2026/1/e91115
doi:10.2196/91115

© Zhenyu Xu, Jingjing Jia, Chaofan Wu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 4.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Adaptive Fast-Slow Large Language Model Framework for Multidimensional Classification of Prenatal Ultrasound Reports: Comparative Study

Wei Zhong[1], MD; Huihui Yan[2], MD; Yifan Liu[2], MD; Yan Liu[2], MD; Kai Yang[1], PhD; Huimin Gao[1], MD; Zhengyang Yao[2], MD; Wenjing Hao[2], MD; Yousheng Yan[1*], MD; Chenghong Yin[3*], MD

[1]Department of Medical Genetics, Beijing Obstetrics and Gynecology Hospital, Capital Medical University. Beijing Maternal and Child Health Care Hospital, Beijing, China

[2]Department of Prenatal Diagnosis Center, Beijing Obstetrics and Gynecology Hospital, Capital Medical University. Beijing Maternal and Child Health Care Hospital, Beijing, China

[3]Department of Central Laboratory, Beijing Obstetrics and Gynecology Hospital, Capital Medical University. Beijing Maternal and Child Health Care Hospital, No. 251 Yaojiayuan Road, Chaoyang District, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Chenghong Yin, MD
Department of Central Laboratory, Beijing Obstetrics and Gynecology Hospital, Capital Medical University. Beijing Maternal and Child Health Care Hospital, No. 251 Yaojiayuan Road, Chaoyang District, Beijing, China

Abstract

Background: Phenotype-driven prenatal diagnosis relies on the precise correlation between ultrasound findings and genetic outcomes; however, this process is hindered by the unstructured nature of clinical ultrasound reports. While large language models (LLMs) hold the potential to address this challenge, their specific application in this domain remains systematically underexplored.

Objective: To establish an effective LLM implementation framework for the clinical multidimensional classification of prenatal ultrasound reports, we evaluated the open-source DeepSeek-V3.2 family on real-world anomalous reports—covering both factual and subjective categories—while integrating retrieval-augmented generation (RAG) and chain-of-thought (CoT) reasoning.

Methods: From a cohort of 4256 pregnancies, we extracted 254 reports with fetal anomalies. We comprehensively evaluated both the high-speed base model (DeepSeek-V3.2-B) and the reasoning-enhanced model (DeepSeek-V3.2-R) across all 5 classification dimensions, comprising 4 factual extraction tasks—primary classification, standardized terminology, anatomical system, and abnormality count—and 1 subjective severity assessment. We further explicitly evaluated the efficacy of RAG for the subjective tasks. Finally, to validate the clinical utility of this approach, we performed a correlation analysis between the expert-validated multidimensional phenotypic profiles and definitive genetic outcomes derived from amniocentesis.

Results: While V3.2-B achieved high efficiency in factual tasks (accuracy and F_1-score >90%), it underperformed in subjective severity grading (56.6% accuracy), exhibiting a recall of 0 for minor anomalies. Crucially, while RAG significantly improved both models' performance on internal retrieval datasets (P<.05), this benefit did not generalize to external test datasets (P>.05). In contrast, the V3.2-R model utilizing CoT reasoning achieved superior robustness (86% accuracy and F_1-score=0.75) on external data without RAG; notably, introducing RAG to V3.2-R degraded performance to 81%, suggesting potential noise interference. Clinical validation against amniocentesis outcomes confirmed that accurate multidimensional phenotypic profiles significantly stratified pathogenic genetic risks.

Conclusions: The rapid base models are efficient for factual classification, and RAG enhances performance on data similar to the knowledge base, whereas CoT is indispensable for subjective assessment. Within the constraints of our dataset and current retrieval implementation, CoT proved more robust than RAG for subjective assessment. However, this finding is specifically tied to our experimental setup and should not be generalized as a universal conclusion. We recommend clinically adopting this adaptive "fast-slow" LLM framework to efficiently perform the multidimensional classification of prenatal ultrasound anomalies. This privacy-preserving, locally deployable solution provides a scalable path to accelerate phenotype-genotype research and optimize invasive diagnostic decision-making.

(J Med Internet Res 2026;28:e91399) doi:10.2196/91399

XSL•FO
RenderX

KEYWORDS

large language models; prenatal ultrasound; DeepSeek; phenotype-driven diagnosis; chain-of-thought; retrieval-augmented generation

Introduction

Prenatal ultrasound is fundamental for fetal assessment, but efficiently transforming these narrative reports into structured data for clinical decision-support remains a significant challenge [1,2]. While amniocentesis provides definitive genetic diagnoses, its associated risks of miscarriage and infection [3] make it unsuitable for universal application to all cases with abnormal ultrasound findings. Crucially, the decision to pursue invasive testing relies on a nuanced, multidimensional risk assessment rather than a singular finding. Different classification dimensions—specifically the affected anatomical system, the count of anomalies (isolated vs multiple), and the severity grading—each carry distinct predictive weights regarding chromosomal outcomes. For instance, multisystem defects or lethal malformations suggest a significantly higher genetic risk compared to isolated soft markers.

Therefore, a comprehensive integration of these multidimensional classifications is essential to provide patients with data-driven counseling, enabling them to weigh the probability of a genetic disorder against the procedural risks of amniocentesis. However, the large-scale analysis required to validate and refine this multidimensional risk stratification is hindered by the nature of clinical documentation: the frequent occurrence of benign anomalies and inconsistent descriptive terminology [3-7] creates unstructured "data silos." Establishing a standardized framework to correlate these specific sonographic phenotypes with genetic outcomes is vital; yet, it has been hindered by the labor-intensive, expert-dependent process of annotating large-scale datasets.

The recent emergence of high-performance, open-source large language models (LLMs) offers a potential solution to this clinical natural language processing bottleneck [8-11]. However, the deployment of LLMs in medicine faces the critical challenge of hallucinations [12]. To mitigate this, retrieval-augmented generation (RAG) has become the prevailing paradigm, anchoring model outputs to external knowledge bases to ensure factual accuracy [13]. However, RAG primarily enhances information retrieval rather than logical reasoning. For complex, subjective tasks—such as assessing the severity of fetal anomalies based on subtle descriptive nuances—access to external knowledge may be insufficient without the capacity for deep reasoning. This has catalyzed the development of chain-of-thought (CoT) reasoning models [12,14], which mimic the deliberate "System 2" thinking popularized by Daniel Kahneman [15,16] by decomposing complex problems into intermediate logical steps. Unlike proprietary models, modern open-source LLMs, such as DeepSeek-V3.2, can be securely deployed within hospital environments. The V3.2 iteration uniquely offers both a high-speed base model (V3.2-B) and a reasoning-enhanced variant (V3.2-R), presenting a new opportunity to address the varying complexity of medical tasks—ranging from routine information extraction to complex logic-based severity assessment—within a unified local framework.

Despite these technological advances, the comparative effectiveness of retrieval-based versus reasoning-based approaches in automating the analysis of prenatal ultrasound reports remains largely unexplored. This study proposes an adaptive "fast-slow" LLM framework to address the multidimensional complexity of fetal phenotype extraction. We utilized the DeepSeek-V3.2 suite to evaluate the trade-offs between the fast base model (V3.2-B) and the slow reasoning model (V3.2-R), specifically assessing the utility of RAG versus CoT in subjective severity grading. To establish the clinical validity of this multidimensional classification, we further analyzed the association between expert-verified results and "gold standard" genetic outcomes from amniocentesis. Our objective was to demonstrate that accurate, multidimensional profiling is strongly predictive of pathogenic risks, and that while the fast base LLM and RAG suffice for factual tasks, CoT reasoning is indispensable for automating the subjective components of this profile, thereby accelerating phenotype-driven diagnosis.

Methods

Patient Recruitment and Data Collection

Between January 2023 and July 2024, 4256 pregnant women underwent an enhanced noninvasive prenatal test (NIPT2.0) as part of a longitudinal clinical efficacy validation study. The inclusion criteria were singleton pregnancies between 12 and 20 weeks of gestation, while excluding those with conditions known to significantly interfere with the NIPT2.0 analysis. This specific gestational window was deliberately selected because it aligns with the optimal and most actionable clinical timeframe for evaluating the necessity of invasive amniocentesis. Although this specific window inherently excludes late-onset anomalies presenting in the third trimester, it precisely captures the most critical period for phenotype-driven genetic risk assessment. Furthermore, because the NIPT2.0 screening—which covers both common aneuploidies and select monogenic disorders [17]—was offered free of charge, participation rates were exceptionally high. This provided a broad and highly representative real-world sample of early-to-mid second-trimester ultrasound findings, laying a solid data foundation for the downstream phenotype-genetic correlation analysis. Maternal age, family history, and prenatal ultrasound reports were collected at enrollment. Participants with negative (low-risk) NIPT2.0 results received routine follow-up, while those with positive (high-risk) results were advised to undergo amniocentesis for definitive diagnosis. The manual review of all records identified 254 participants with ultrasound reports showing anomalies of the fetus, placenta, umbilical cord, or amniotic fluid. These 254 reports constituted the dataset for this analysis. The study workflow is depicted in Figure 1.

XSL•FO
RenderX

Figure 1. Flowchart of the study design. The workflow comprises three stages: (1) cohort construction: the selection of 254 abnormal fetal ultrasound reports from 4256 screenings. (2) The adaptive "fast-slow" framework: a dual-pathway system deploying DeepSeek-V3.2-B (fast) for objective factual extraction and DeepSeek-V3.2-R (slow) for subjective severity assessment. This phase explicitly compares the efficacy of RAG versus CoT reasoning. (3) Clinical validation: the dual assessment of the structured output against expert consensus (technical accuracy) and amniocentesis genetic outcomes (clinical utility). CoT: chain-of-thought; LLMs: large language models; NIPT2.0: noninvasive prenatal test; RAG: retrieval-augmented generation; V3.2-B: DeepSeek-V3.2 base model; V3.2-R: DeepSeek-V3.2 reasoning-enhanced model.

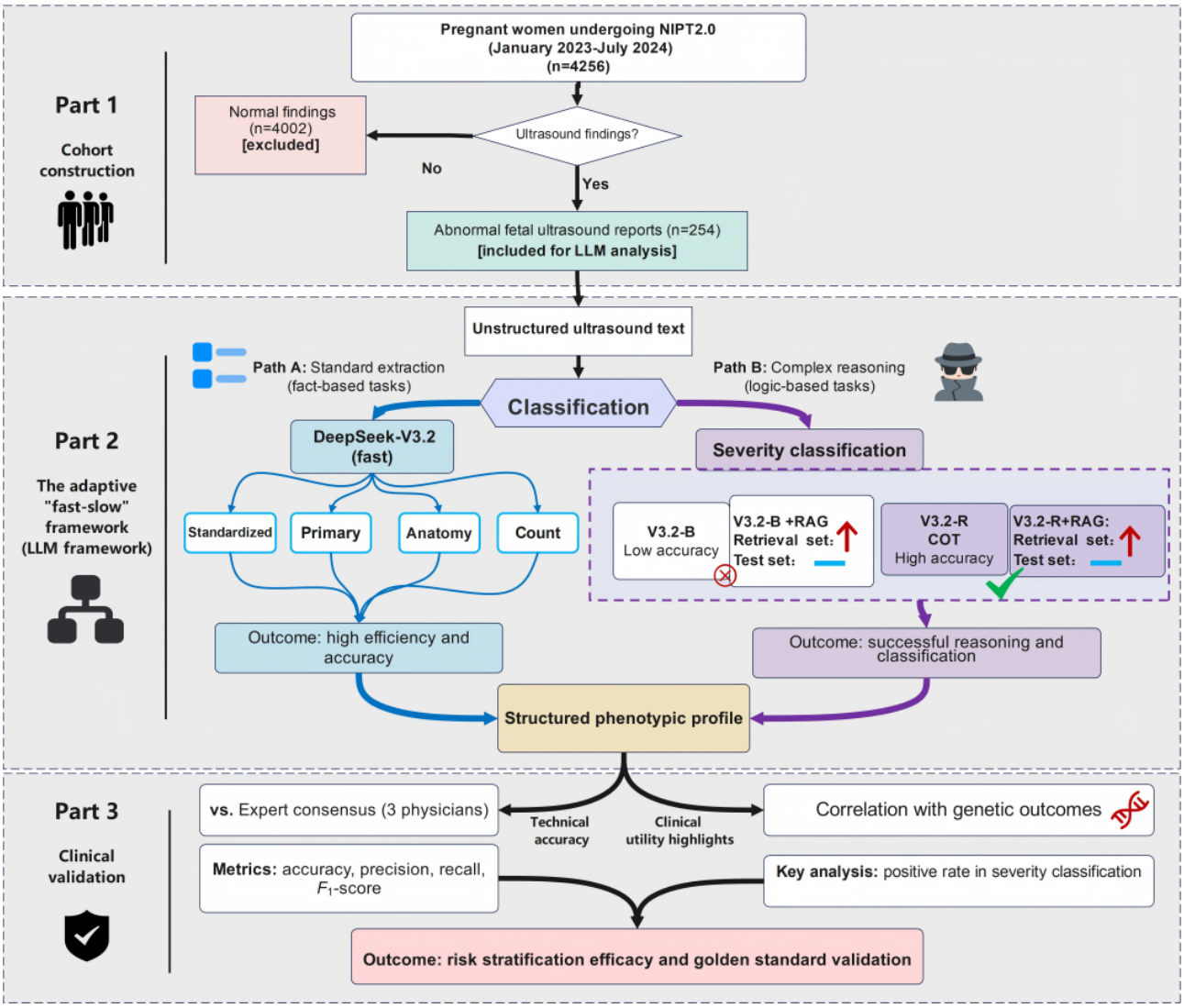

Multidimensional Classification of Prenatal Ultrasound Reports by LLMs

The 254 selected abnormal prenatal ultrasound reports were organized into a spreadsheet format. We utilized Dify (LangGenius, Inc) [18] to call the application programming interface of DeepSeek-V3.2, provided by the SiliconCloud service [19], to perform the automated classification. Each report was analyzed individually in a single-turn conversation with model parameters set to temperature=0 and top-p=0.95.

Five classification schemes were developed to structure the raw report data, including 4 fact-based classifications and 1 subjective assessment. For each scheme, prompts were designed by a sonographer and a prenatal diagnostician to guide the LLM. Due to their extensive length, the full verbatim prompts are provided in Multimedia Appendix 1. Nevertheless, all prompts adhered to a consistent design framework; they explicitly delineated the clinical rules and definitions for each classification scheme (as outlined below) and strategically incorporated specific explanations alongside clinical examples to disambiguate potentially confusing or overlapping findings. For classification purposes, suspected anomalies were treated as confirmed. The 5 schemes were as follows:

1. Primary classification: categorized reports into 4 mutually exclusive groups: increased nuchal translucency (NT), other soft markers, structural abnormalities, and fetal growth restriction (FGR).
2. Standardized terminology: converted descriptive text into standardized clinical terms while omitting laterality (eg, standardizing "left ventriculomegaly" to "ventriculomegaly").
3. Anatomical system: mapped anomalies to the affected fetal system (eg, nervous and cardiovascular). Nonstructural or nonfetal findings were classified as "None."
4. Abnormality count: Classified reports as "Solitary" (a single anomaly) or "multiple" (≥2 distinct anomalies, including bilateral presentations of a single finding).

XSL•FO
RenderX

5. Severity (subjective assessment): Findings were graded based on the most severe anomaly present, categorized into 3 levels according to standardized clinical criteria: lethal (eg, anencephaly, typically requiring termination of pregnancy), major (eg, complex congenital heart disease, significantly affecting viability), or minor (eg, cleft lip, surgically correctable postnatally). Nonstructural findings, such as NT or FGR, were categorized as "Other."

For all schemes except "abnormality count" and "severity," the model could generate multiple comma-separated outputs per report.

LLM Execution and Evaluation

Two DeepSeek models were utilized sequentially. While both models belong to the DeepSeek-V3.2 family and share a highly efficient underlying architecture (incorporating mechanisms like DeepSeek Sparse Attention for rapid processing), their posttraining paradigms and inference mechanisms differ fundamentally [20]. V3.2-B is optimized using standard supervised fine-tuning for rapid instruction following and direct pattern matching, making it highly efficient for objective factual extraction. In contrast, V3.2-R is a reasoning-enhanced variant heavily trained with a scalable reinforcement learning protocol to intrinsically generate intermediate CoT steps. During inference, V3.2-R allocates substantial additional computational resources to process an internal "thinking" phase before producing the final response. This architectural distinction mimics human "System 2" deliberate deduction, enabling it to handle subjective clinical nuances, albeit at the cost of significantly longer processing times.

First, the V3.2-B rapidly processed all reports for initial classification (2 - 4 s/report). To ensure the reliability of the classification, 2 experienced attending prenatal diagnosticians independently evaluated the V3.2-B outputs in a double-blind manner, labeling each classification as "Correct" or "Incorrect." Disagreements were resolved through consensus meetings guided by a senior chief diagnostician (YY) to form a preliminary reference standard. Because all discrepancies were ultimately resolved via 100% consensus, formal interrater reliability metrics were not calculated.

Subsequently, the senior chief diagnostician utilized this preliminary standard to evaluate V3.2-R outputs. This secondary expert review step was essential because V3.2-R occasionally presented valid but differently phrased classifications that required expert judgment rather than simple string-matching, thereby establishing the final "gold standard" dataset.

Both models were independently evaluated on the entire dataset across all 5 classification schemes. For this study, an F_1-score greater than 0.90 was prospectively defined as indicating highly reliable and clinically acceptable performance for the classification of prenatal ultrasound anomalies.

Construction of Knowledge Base and Implementation of RAG

To evaluate the efficacy of RAG in enhancing model performance, the dataset of expert-verified severity assessments (n=254) was systematically partitioned based on sequential identification numbers. The first half (n=127) constituted the retrieval set, which was vectorized to construct the external knowledge base. The subsequent half (n=127) served as the unseen test set. While this sequential split shares local linguistic patterns and clinical workflows, it was intentionally designed to simulate a real-world scenario, in which a hospital utilizes its own historical records as a RAG knowledge base to process new, incoming reports of a similar style.

The RAG pipeline was deployed using the Dify platform. We integrated the Qwen3-Reranker-8B model for semantic reranking of candidate chunks. The retrieval parameters were configured with a Top-K of 3 and a similarity score threshold of 0.60 to filter out low-relevance noise. Finally, both V3.2-B and V3.2-R were tested on both the retrieval set and the test set to assess two critical metrics: (1) the effectiveness of RAG in retrieving and utilizing "seen" knowledge (performance on the retrieval set) and (2) the generalizability of the RAG-enhanced models when applied to "unseen" clinical data (performance on the test set).

Follow-Up and Prenatal Diagnostic Outcomes

Among the 254 women with abnormal ultrasound findings, those with high-risk NIPT2.0 results were counseled for diagnostic amniocentesis. Definitive genetic testing included one or more of the following: karyotyping, chromosomal microarray analysis, whole-exome sequencing, and copy number variation sequencing. Cases were categorized based on these results. Patients who declined diagnostic testing after a high-risk NIPT2.0 result were excluded from the association analysis.

For the purpose of this study, a negative prenatal diagnostic outcome was assigned to all participants with a low-risk NIPT2.0 result. This approach was justified by the high negative predictive value of NIPT2.0 and supported by two observations in our cohort: (1) all women in this group who nonetheless underwent amniocentesis for sonographic indications had negative results and (2) clinicians did not recommend invasive testing for the remainder, judging the genetic risk to be low.

Association Analysis Between Classified Ultrasound Abnormalities and Genetic Outcomes

An association analysis was performed on 251 eligible cases, correlating the classified ultrasound abnormalities with genetic diagnostic outcomes. This analysis utilized a "gold standard" dataset, which consisted of clinician-verified LLM classifications supplemented with manual corrections. The results were visualized using bar charts, showing the number and proportion of positive and negative genetic diagnoses for each anomaly category.

Prior to analysis, manual data curation was performed to standardize the LLM outputs for 3 classification schemes (standardized terminology, primary classification, and anatomical system). This step involved consolidating semantically identical but textually variant terms (eg, merging "Increased NT" and "Increased nuchal translucency") and grouping identical combinations of findings listed in different orders to ensure accurate frequency counts for the association analysis.

XSL•FO
RenderX

Statistical Analysis

To evaluate the performance of DeepSeek-V3.2, we calculated the overall accuracy for the categories. A true positive is recorded when an item in the model's output list also appears in the gold standard; a false positive is an output item absent from the gold standard; a false negative is a gold-standard item missing from the output list. The F_1-score was calculated as follows:

$$F1=2\times\text{Precision}\times\text{Recall}\text{Precision}+\text{Recall}$$

For the multiclass "Severity" classification, macroaveraged precision, recall, and F_1-scores were calculated to assess balanced performance across all categories:

$$\text{Macro}-F1=1N\sum i=1NF1i$$

where N is the number of classes. Additionally, we calculated the precision, recall, and F_1-score for each of the 4 severity categories individually.

For the association analysis, Pearson's χ^2 test was used to compare the rate of positive genetic diagnoses between the solitary and multiple abnormality categories, with significance set at $P<.05$. Due to the limited number of positive cases within the numerous categories of the other classifications, these were analyzed using descriptive statistics. The results were visualized as the number and proportion of cases with positive versus negative genetic diagnoses for each abnormality type.

The difference in accuracy between the pre- and post-RAG phases was analyzed using the McNemar test for paired categorical data. Continuity correction was applied where appropriate. Statistical significance was defined as a 2-sided $P<.05$. Data analysis was conducted using Python 3.10.

Ethical Considerations

This comparative effectiveness study utilized the existing data from a prospectively registered cohort. The cohort protocol was registered with the Medical Research Registration and Filing Information System of the National Health Security Information Platform of China (registration number MR-11-24-002508). The study was conducted in accordance with the Declaration of Helsinki and approved by the institutional review board (2023-KY-099 - 02). All participants in the original cohort provided written informed consent, agreeing to the use of their deidentified clinical information for scientific research purposes. As this study constitutes a secondary analysis of existing, deidentified data for comparative effectiveness purposes without causal inference, no additional patient contact or consent was required.

Results

Screening and Identification of Cases With Positive Amniocentesis Results

Of the 254 women with abnormal ultrasound findings, 30 had high-risk NIPT2.0 results. Three of these women, all with a high risk for Trisomy 21, declined amniocentesis and were therefore excluded from the association analysis. The remaining 27 women underwent amniocentesis, all of whom were confirmed to have a positive genetic diagnosis. This resulted in a final cohort of 251 cases for the association analysis, comprising 27 positive and 224 negative genetic outcomes.

Evaluation of LLMs' Performance in Multidimensional Classification

V3.2-B demonstrated high accuracy (>90%) and F_1-score (>0.9) across 4 fact-based classifications (Table 1). Specifically, its accuracy was 98.4% (250/254) for standardized terminology, 92.9% (236/254) for primary classification, 90.1% (229/254) for anatomical system, and 98.4% (250/254) for abnormality count. When independently applied to the entire dataset for these 4 objective tasks, V3.2-R achieved higher performance metrics across all categories compared to V3.2-B. However, because the performance of the base model was already exceptionally high, the improvements provided by the reasoning model were marginal (a maximum accuracy difference of 4.7% in the primary classification).

XSL•FO
RenderX

Table . Performance of the DeepSeek-V3.2 in classifying fetal ultrasound abnormalities (N=254).

Classification scheme	Accuracy (V3.2-B)[a], n (%)	Accuracy (V3.2-R)[b], n (%)	Precision	Recall	F_1-score
Standardized terminology	250 (98.4)	254 (100)	0.99	0.99	0.99
Primary classification	236 (92.9)	248 (97.6)	0.94	0.96	0.95
Anatomical system	229 (90.1)	240 (94.4)	0.90	0.93	0.91
Abnormality count	250 (98.4)	254 (100)	0.99	0.96	0.98
Severity classification[c]	144 (56.6)	215 (84.6)	0.83[c]	0.77[c]	0.75[c]
Lethal (n=10)	—[d]	—	1.00	0.40	0.57
Major (n=31)	—	—	0.53	1.00	0.70
Minor (n=55)	—	—	0.79	0.82	0.80
Other (n=158)	—	—	1.00	0.85	0.92

[a]V3.2-B: DeepSeek-V3.2 base model.

[b]V3.2-R: DeepSeek-V3.2 reasoning-enhanced model.

[c]For the 4 factual classifications (excluding severity classification), the precision, recall, and F_1-score metrics reported in the subsequent analysis are derived from V3.2-B, as its initial performance achieved the predefined high-level threshold of this study (F_1-score >0.90). For the severity classification, the detailed performance metrics reported are based on V3.2-R, as the initial accuracy of V3.2-B was inadequate.

[d]Not applicable.

In contrast, the "Severity" classification proved more challenging. V3.2-B's accuracy was only 56.7% (144/254), while V3.2-R significantly improved to 84.6% (215/254), with a macro-F_1-score of 0.75.

A detailed breakdown of V3.2-R's performance on the severity task revealed trade-offs: it achieved perfect precision for "lethal malformation" but with low recall (0.40), while for "major malformation," it achieved perfect recall at the cost of lower precision (0.53). The LLM performed best when classifying findings into the "Other" category (F_1-score=0.92).

Comparative Efficacy of RAG Versus CoT Reasoning in Severity Assessment

We evaluated the performance of the V3.2-B and V3.2-R models across both the internal retrieval set and the external test set, before and after the implementation of RAG (Table 2).

XSL•FO
RenderX

Table . Comparative performance of DeepSeek-V3.2 base and reasoning models for fetal anomaly severity assessment before and after RAG[a].

Model and set	Before RAG[b] (n=127)				After RAG (n=127)				*P* value[c]
	Accuracy	Precision	Recall	F_1-score	Accuracy	Precision	Recall	F_1-score	
V3.2-B[d]									
Retrieval	0.56	0.45	0.62	0.48	0.70	0.61	0.68	0.61	.002
Lethal (n=4)	—[e]	0.57	1.00	0.72	—	0.50	0.75	0.60	—
Major (n=19)	—	0.29	0.84	0.43	—	0.40	0.79	0.54	—
Minor (n=26)	—	0	0	0	—	0.62	0.38	0.48	—
Other (n=78)	—	0.96	0.65	0.78	—	0.90	0.78	0.84	—
Test	0.57	0.43	0.62	0.46	0.59	0.49	0.59	0.49	.61
Lethal (n=6)	—	0.54	1.00	0.71	—	0.44	0.67	0.53	—
Major (n=12)	—	0.18	0.75	0.29	—	0.25	0.75	0.35	—
Minor (n=29)	—	0	0	0	—	0.30	0.24	0.27	—
Other (n=80)	—	0.98	0.71	0.83	—	1.00	0.69	0.81	—
V3.2-R[f]									
Retrieval	0.83	0.83	0.72	0.70	0.99	0.99	1.00	0.99	<.001
Lethal (n=4)	—	1.00	0.25	0.40	—	1.00	1.00	1.00	—
Major (n=19)	—	0.58	1.00	0.73	—	1.00	1.00	1.00	—
Minor (n=26)	—	0.75	0.81	0.78	—	0.96	1.00	0.98	—
Other (n=78)	—	1.00	0.83	0.91	—	1.00	0.99	0.99	—
Test	0.86	0.83	0.80	0.77	0.81	0.69	0.69	0.69	.33
Lethal (n=6)	—	1.00	0.50	0.67	—	0.60	0.50	0.50	—
Major (n=12)	—	0.48	1.00	0.65	—	0.53	0.67	0.59	—
Minor (n=29)	—	0.83	0.83	0.83	—	0.71	0.69	0.70	—
Other (n=80)	—	1.00	0.87	0.93	—	0.90	0.90	0.90	—

[a]Data represent the performance metrics for the internal retrieval set (n=127) and the external test set (n=127). The retrieval set consists of data used to construct the RAG vector database, whereas the test set comprises the unseen data.

[b]RAG: retrieval-augmented generation.

[c]*P* values were calculated using the McNemar test to determine the statistical significance of the difference in overall accuracy before and after the implementation of RAG for each model.

[d]V3.2-B: DeepSeek-V3.2 base model.

[e]Not applicable.

[f]V3.2-R: DeepSeek-V3.2 reasoning-enhanced model.

XSL•FO
RenderX

V3.2-R demonstrated a substantial performance advantage over V3.2-B prior to the application of RAG. On the test set, V3.2-R achieved an accuracy of 86%, significantly outperforming V3.2-B, which achieved only 57%. Notably, in the classification of "Minor" anomalies, V3.2-B completely failed to identify any cases (precision, recall, and F_1-score=0), whereas V3.2-R achieved a high F_1-score of 0.83. This disparity highlights the intrinsic limitation of V3.2-B in handling subjective severity grading without explicit reasoning capabilities.

When applied to the retrieval set, RAG significantly improved the performance of both models. The accuracy of V3.2-B increased from 56% to 70% (P=.002), with the F_1-score for "Minor" anomalies rising from 0 to 0.48, indicating that the model successfully retrieved relevant examples to correct its output. Moreover, V3.2-R achieved near-perfect performance with RAG, improving accuracy from 83% to 99% (P<.001). Precision and recall metrics across all severity subtypes approached or reached 1.00. These results confirm that the RAG pipeline was technically functional and capable of enhancing performance when the test data were semantically identical or highly similar to the knowledge base.

Crucially, the performance gains observed in the retrieval set did not translate to the external test set, revealing a critical limitation in the generalizability of RAG for this specific task. V3.2-B showed no statistically significant improvement with RAG (accuracy: 57% vs 59%, P=.61). While RAG slightly improved the detection of "Minor" anomalies (F_1-score increased to 0.27), the overall capability remained suboptimal compared to the reasoning model. The implementation of RAG on V3.2-R resulted in a slight, though not statistically significant, decline in accuracy on the external test set (86% vs 81%, P=.33). Specifically, the F_1-scores for "Lethal" and "Major" anomalies decreased after RAG (lethal: 0.67-0.50; major: 0.65-0.59).

These data indicate that while RAG can effectively guide LLMs to memorize or retrieve specific patterns within the knowledge base, it fails to significantly enhance—and may potentially hinder—performance on unseen, heterogeneous clinical data. In contrast, the CoT reasoning inherent in V3.2-R (without RAG) proved to be the most robust approach for the subjective task of severity assessment, achieving the highest stand-alone accuracy (86%) and F_1-scores on the external validation cohort.

Descriptive Statistics of Prenatal Ultrasound Abnormalities

A descriptive analysis of the curated dataset revealed the distribution of abnormalities across the 5 classification schemes (Figure 2).

XSL•FO
RenderX

Figure 2. Distribution of prenatal ultrasound abnormalities based on the manually curated classifications generated by DeepSeek-V3.2. (A) Frequencies of standardized medical terms for ultrasound findings. The model standardized unstructured descriptions into common terms, generating 371 entries in total. Only terms with a frequency greater than 1% are displayed. (B) Distribution of abnormalities by the affected anatomical system (n=176). (C) Distribution of cases by severity classification (n=254).

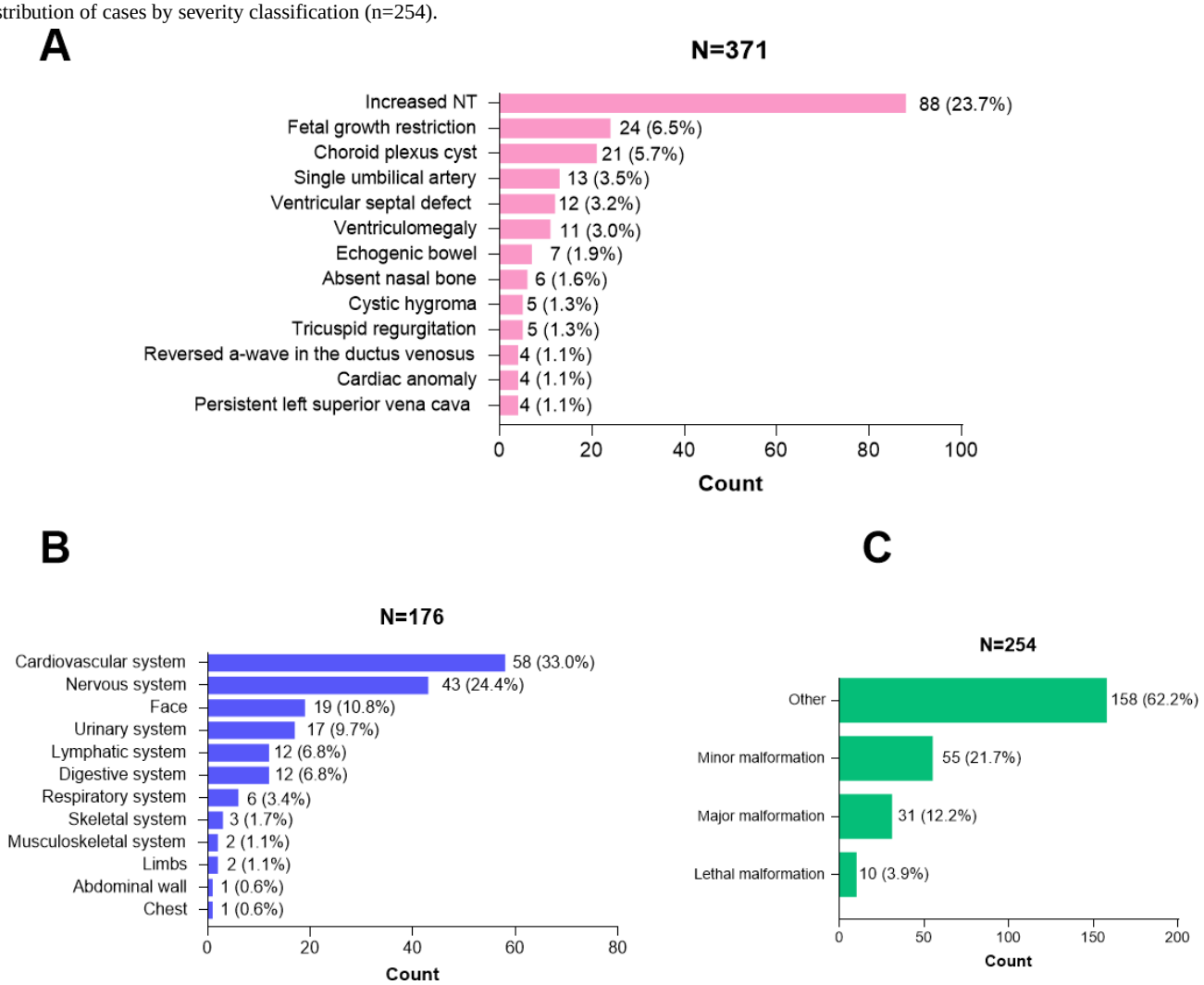

By standardized terminology, the most frequent among the 371 identified findings were increased NT (88/371, 23.7%), FGR (24/371, 6.5%), choroid plexus cyst (21/371, 5.7%), and single umbilical artery (13/371, 3.5%; Figure 2A). Most reports described solitary findings (171/254, 67.3%) rather than multiple findings (83/254, 32.7%). Among 176 classified structural anomalies, the cardiovascular (58/176, 33%) and nervous (43/176, 24.4%) systems were the most commonly affected (Figure 2B). The primary classifications were distributed among increased NT (96/292, 32.5%), other soft markers (83/292, 28.4%), structural abnormalities (82/292, 28.1%), and FGR (32/292, 11%). Finally, by severity, most cases were categorized as “Other” (158/254, 62.2%), followed by minor (55/254, 21.7%), major (31/254, 12.2%), and lethal (10/254, 3.9%) malformations (Figure 2C).

Association Analysis

An association analysis correlated the classified ultrasound findings with genetic outcomes (Figure 3). The presence of multiple abnormalities significantly increased the risk of a positive genetic diagnosis (Figure 3) compared to solitary findings (14/82, 17.1% vs 13/169, 7.7%).

XSL•FO
RenderX

Figure 3. Association analysis between classified prenatal ultrasound abnormalities and genetic diagnostic outcomes. The bar charts display the number and proportion of cases with negative (blue) and positive (red) prenatal diagnoses for each classification scheme. (A) Analysis by standardized terminology. This panel shows all 27 positive cases and their associated ultrasound findings. For clarity, only findings from the negative diagnosis group with a frequency greater than 2 are displayed. (B) Analysis by primary classification. (C) Analysis by the anatomical system. (D) Analysis by the abnormality count. The rate of positive diagnoses was significantly higher in cases with multiple abnormalities compared to those with solitary findings (P=.02, Pearson χ^2 test). (E) Analysis by severity (subjective assessment). AVSD: atrioventricular septal defect; DA: ductus arteriosus; DV: ductus venosus; FGR: fetal growth restriction; NT: nuchal translucency; RR: relative risk; VSD: ventricular septal defect.

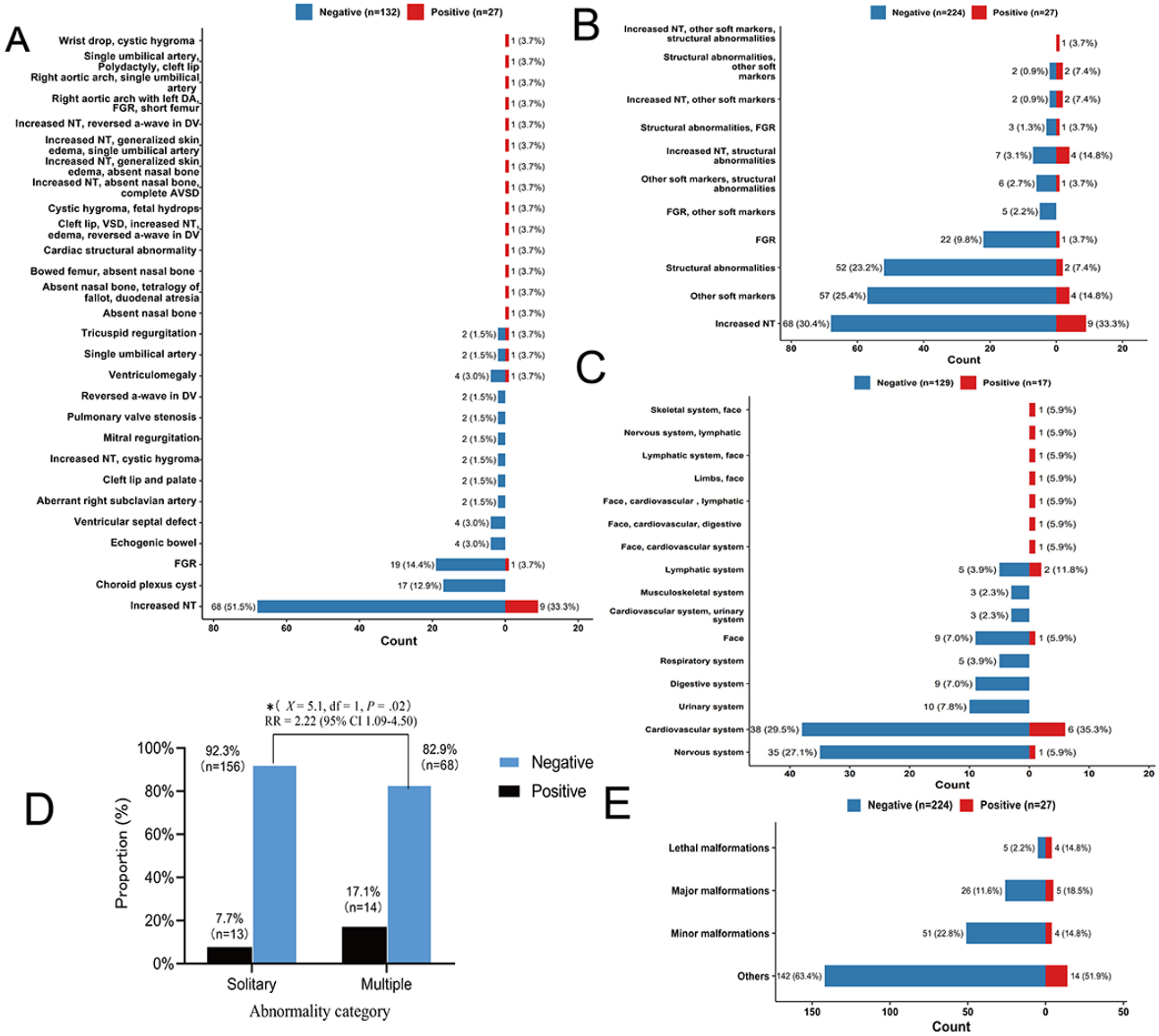

Among solitary findings, the increased NT was the most common abnormality associated with a positive genetic diagnosis (n=9; Figures 3A and 3B). Notably, no positive diagnoses were found in cases with isolated choroid plexus cysts, ventriculomegaly, echogenic bowel, or ventricular septal defects (Figure 3A).

Risk is also correlated with the anatomical system. Among cases with positive genetic diagnoses, cardiovascular and lymphatic system abnormalities were the most frequent. Conversely, no positive diagnoses in this cohort were associated with isolated anomalies of the urinary, digestive, or musculoskeletal systems (Figure 3C).

More importantly, lethal and major malformations were disproportionately represented in the positive diagnosis cohort (Figure 3E), accounting for 14.8% (4/27) and 18.5% (5/27) of positive cases, respectively. In addition, positive cases constituted 44.44% (4/9) of all lethal malformations and 16.13% (5/31) of all major malformations but only 7.14% (4/56) of minor malformations.

Discussion

Principal Results

This study establishes an adaptive “fast-slow” framework utilizing the open-source DeepSeek-V3.2 family for the automated, multidimensional classification of prenatal ultrasound reports. By strategically deploying a high-speed base model for factual extraction and a reasoning-enhanced model for subjective assessment, our approach significantly enhances data annotation efficiency while resolving the complexity of phenotype validation. Crucially, we identified a pivotal

XSL•FO
RenderX

mechanistic dichotomy; while RAG improves performance on the data seen within the knowledge base, it fails to generalize to external subjective tasks. In contrast, CoT reasoning demonstrates superior robustness in "unseen" scenarios, effectively mimicking the "System 2" clinical judgment required for severity grading. This work provides a foundational pipeline for phenotype-driven research using unstructured hospital data and offers a reliable tool to support clinical decision-making, highlighting the necessity of matching model cognitive architectures to clinical task complexity.

Limitations

This study has limitations. First, while we highlighted the superiority of CoT over RAG for subjective tasks, our RAG implementation utilized a specific reranking strategy (Qwen3-Reranker). Alternative retrieval algorithms or hybrid approaches might yield different results. Second, the cohort, while expertly annotated, is relatively small (n=254) and derived from a single center, and the fact that the "gold standard" annotations were established through the expert review of model outputs rather than being fully independent inevitably introduces a degree of subjectivity. Consequently, certain clinically important subcategories, such as lethal malformations, are represented by very small sample sizes, which may affect the statistical stability of our performance metrics within these subgroups. Furthermore, our evaluation utilized an internal sequential split. While the strict separation of patient IDs prevented direct data leakage, the unseen test set shared local linguistic patterns with the retrieval knowledge base. This setup inherently provides an optimistic performance estimate for the RAG pipeline. However, this constraint actually reinforces our primary findings; even under conditions highly favorable to retrieval, CoT reasoning still demonstrated superior robustness for subjective tasks. While this single-center design accurately simulates local clinical deployment—where a hospital utilizes its own historical records—it does not evaluate true out-of-distribution generalizability. Broader generalizability across different institutional reporting styles remains to be established through future multicenter or larger-scale studies. Third, we did not formally evaluate prompt engineering variations [21-23], although reasoning models typically show resilience to prompt nuances [24]. Fourth, potential confounders, such as gestational age, were not integrated into the model's decision logic, warranting future multimodal investigations. Finally, it is also important to acknowledge limitations regarding our clinical end points. For one, treating all low-risk NIPT2.0 cases as negative genetic outcomes without universal confirmatory amniocentesis—while clinically justified by the test's high negative predictive value and our cohort observations—may introduce a minor risk of misclassification bias. Statistically, this inherent uncertainty of screening-based negatives could potentially lead to a slight underestimation of the true genetic risk associated with specific ultrasound phenotypes. Furthermore, our validation specifically targeted pathogenic genetic risk to support invasive diagnostic decision-making. Because ultrasound anomalies frequently arise from heterogeneous nongenetic etiologies, a more comprehensive assessment of the framework's overall clinical utility will require future studies integrating additional longitudinal end points, such as postnatal diagnoses and long-term functional outcomes.

Comparison With Prior Work

While LLMs have shown broad capabilities across medicine [25-47], a "one-size-fits-all" approach remains inefficient for complex clinical workflows. Our findings challenge the prevailing assumption that RAG is the universal solution for medical LLM hallucinations. In our study, RAG successfully corrected the base model's errors within the retrieval set, confirming its utility for pattern matching. However, this performance collapsed on the external test set. This suggests that for subjective tasks like severity assessment—which rely on synthesizing subtle cues—semantic retrieval is insufficient. RAG retrieves similar text chunks but not necessarily the logic of the diagnosis. Conversely, the V3.2-R model, utilizing CoT, achieved an 86% accuracy on the external set without accessing the knowledge base. This indicates that internalized reasoning capabilities (navigating clinical logic steps) are more critical than external knowledge retrieval (accessing facts) when dealing with the nuanced subjectivity of fetal anomalies [44]. Notably, introducing RAG to V3.2-R degraded performance to 81%, suggesting potential noise interference.

Unlike commercial proprietary models, the open-source nature of the DeepSeek suite allows for secure local deployment, ensuring patient data privacy—a nonnegotiable requirement for handling sensitive prenatal records [39,40,48-51]. Our framework maximizes resource efficiency: by routing the majority of straightforward tasks (entity extraction and counting) to the "Fast" V3.2-B model, we preserve the computationally expensive "Slow" V3.2-R model only for tasks where it provides a statistically significant benefit. This tiered approach addresses the processing speed bottlenecks often cited as a barrier to deploying reasoning models in real-time clinical settings [52,53]. Admittedly, the DeepSeek LLMs in this study did not achieve perfect accuracy across all classification tasks; even V-3.2R achieved only 84.6% accuracy in the subjective severity grading of the entire dataset. However, model performance is expected to improve with future advancements in open-source LLMs. Furthermore, while no current LLM can fully replace specific clinical practice, they are already sufficient to significantly enhance efficiency.

The ultimate value of this technical framework lies in its clinical utility. By enabling high-throughput, multidimensional classification, we were able to conduct an association analysis between sonographic phenotypes and genetic outcomes. However, these association results are exploratory and specific to our single-center cohort. Establishing higher-level evidence for the strength of phenotype-genotype correlations would necessitate much larger, multicenter datasets and more rigorous statistical validation to ensure broader clinical applicability. The fundamental aim of this analysis was to illustrate the practical clinical significance of the multidimensional profiles generated by our LLM framework. Our automated severity grading successfully stratified patients into distinct risk categories, confirming established high-risk predictors, such as multiple anomalies and specific system involvement (eg, cardiovascular) [6,54-58]. The decreasing genetic risk observed

XSL•FO
RenderX

across lethal (4/9, 44.44%), major (5/31, 16.13%), and minor (4/56, 7.14%) malformations demonstrates that subjective severity grading is an indispensable dimension for phenotype-driven diagnosis. Importantly, the reasoning model accurately identified “Minor” malformations—a category the base model completely missed. This granularity provides quantitative support for the lower (yet nonnegligible) risk nature of isolated markers, potentially aiding in reducing unnecessary invasive procedures for low-risk findings [59], while ensuring that subtle but significant patterns are not overlooked [60]. This validates that our “human-in-the-loop” LLM framework does not merely digitize text but actively contributes to refining risk stratification.

Conclusions

This study demonstrates that a monolithic LLM strategy is insufficient for the diverse challenges of prenatal diagnosis. We propose an adaptive framework where “Fast” models handle factual extraction, while “Slow” reasoning models are prioritized for subjective clinical assessment, as they demonstrated greater robustness than our specific RAG implementation in this cohort. However, this finding is contextualized within our current experimental framework and does not constitute a generalized conclusion regarding the inherent superiority of CoT over RAG. By aligning the cognitive architecture of LLM agents with the cognitive demands of medical tasks, we offer a scalable, privacy-preserving path to transform unstructured ultrasound narratives into actionable, phenotype-driven clinical intelligence.

Acknowledgments

No generative artificial intelligence tools were used at any stage in the preparation of this manuscript.

Funding

This study was supported by the National Key Research and Development Program of China (2023YFC2705600), the Capital Clinical Characteristic Diagnosis and Treatment Technology Research and Transformation Application Project (Z221100007422012), the Beijing Hospital Management Center "Yangfan" Plan 3.0 Clinical Technology Innovation Project (ZLRK202329), and the Science and Technology Innovation and Transformation Special Project of Beijing Obstetrics and Gynecology Hospital Affiliated to Capital Medical University/Beijing Maternal and Child Health Hospital (FCYYZH202201).

Data Availability

All data generated or analyzed during this study, including the full anonymized dataset (comprising original Chinese ultrasound texts, English translations, and specific amniocentesis outcomes), expert annotations, and full large language model (LLM) prompts, are included in this published article (Multimedia Appendix 1). The code used for data analysis, including the generation of Figures 2 and 3, evaluation of LLM performance metrics, and statistical analyses, is openly available in the public repository [61]. Furthermore, to ensure full methodological reproducibility without requiring custom execution scripts, the exact workflow configurations and hyperparameters used for the LLM inference and retrieval-augmented generation pipeline via the open-source Dify platform are thoroughly documented in the Methods section.

Authors' Contributions

Conceptualization: WZ, HY
Methodology: WZ, HY, Yifan Liu, KY
Investigation: WZ, HY, Yifan Liu, KY
Formal analysis: Yan Liu, HG, ZY
Visualization: WZ
Writing – original draft: WZ, HY
Writing – review & editing: WZ, HY, Yifan Liu, Yan Liu, KY, HG, ZY, WH, YY, CY
Project administration: YY, CY
Supervision: YY, CY
Funding acquisition: YY, CY
All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The full anonymized prenatal-ultrasound abnormality dataset, 5 classification schemes with expert scores, DeepSeek large language model prompts, and severity assessment before and after RAG.
[XLSX File, 105 KB - jmir_v28i1e91399_app1.xlsx]

XSL•FO
RenderX

References

1. Karim JN, Campbell H, Pandya P, et al. Clinical and cost-effectiveness of detailed anomaly ultrasound screening in the first trimester: a mixed-methods study. Health Technol Assess 2025 May;29(22):1-250. [doi: 10.3310/NLTP7102] [Medline: 40455571]
2. Rivero-Arias O, Png ME, White A, et al. Benefits and harms of antenatal and newborn screening programmes in health economic assessments: the VALENTIA systematic review and qualitative investigation. Health Technol Assess 2024 Jun;28(25):1-180. [doi: 10.3310/PYTK6591] [Medline: 38938110]
3. Ryan GA, Start AO, Cathcart B, et al. Prenatal findings and associated survival rates in fetal ventriculomegaly: a prospective observational study. Int J Gynaecol Obstet 2022 Dec;159(3):891-897. [doi: 10.1002/ijgo.14206] [Medline: 35373343]
4. Bergström C, Ngarina M, Abeid M, et al. Health professionals' experiences and views on obstetric ultrasound in Tanzania: a cross-sectional study. Womens Health (Lond) 2024;20:17455057241273675. [doi: 10.1177/17455057241273675] [Medline: 39206633]
5. Rossi AC, Prefumo F. Accuracy of ultrasonography at 11-14 weeks of gestation for detection of fetal structural anomalies: a systematic review. Obstet Gynecol 2013 Dec;122(6):1160-1167. [doi: 10.1097/AOG.0000000000000015] [Medline: 24201688]
6. Huang J, Wu D, He JH, et al. Associations between genomic aberrations, increased nuchal translucency, and pregnancy outcomes: a comprehensive analysis of 2,272 singleton pregnancies in women under 35. Front Med (Lausanne) 2024;11:1376319. [doi: 10.3389/fmed.2024.1376319] [Medline: 38633307]
7. Su J, Qin Z, Fu H, et al. Association of prenatal renal ultrasound abnormalities with pathogenic copy number variants in a large Chinese cohort. Ultrasound Obstet Gynecol 2022 Feb;59(2):226-233. [doi: 10.1002/uog.23702] [Medline: 34090309]
8. Gibney E. 'Another DeepSeek moment': Chinese AI model Kimi K2 stirs excitement. Nature New Biol 2025 Jul 24;643(8073):889-890. [doi: 10.1038/d41586-025-02275-6] [Medline: 40670748]
9. Normile D. Chinese firm's large language model makes a splash. Science 2025 Jan 17;387(6731):238. [doi: 10.1126/science.adv9836] [Medline: 39818899]
10. Gibney E. China's cheap, open AI model DeepSeek thrills scientists. Nature 2025 Feb;638(8049):13-14. [doi: 10.1038/d41586-025-00229-6] [Medline: 39849139]
11. Guo D, Yang D, Zhang H, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature New Biol 2025 Sep;645(8081):633-638. [doi: 10.1038/s41586-025-09422-z] [Medline: 40962978]
12. Liu F, Zhou H, Gu B, et al. Application of large language models in medicine. Nat Rev Bioeng 2025 Jun;3(6):445-464. [doi: 10.1038/s44222-025-00279-5]
13. Amugongo LM, Mascheroni P, Brooks S, Doering S, Seidel J. Retrieval augmented generation for large language models in healthcare: a systematic review. PLOS Digit Health 2025 Jun;4(6):e0000877. [doi: 10.1371/journal.pdig.0000877] [Medline: 40498738]
14. Ahn S. A guide to evade hallucinations and maintain reliability when using large language models for medical research: a narrative review. Ann Pediatr Endocrinol Metab 2025 Jun;30(3):115-118. [doi: 10.6065/apem.2448278.139] [Medline: 40624912]
15. Shafir E. Daniel Kahneman obituary: psychologist who revolutionized the way we think about thinking. Nature New Biol 2024 May 16;629(8012):526. [doi: 10.1038/d41586-024-01344-6]
16. Fischhoff B. Daniel Kahneman (1934–2024). Science 2024 May 3;384(6695):515-515. [doi: 10.1126/science.adp6405]
17. Zhang J, Wu Y, Chen S, et al. Prospective prenatal cell-free DNA screening for genetic conditions of heterogenous etiologies. Nat Med 2024 Feb;30(2):470-479. [doi: 10.1038/s41591-023-02774-x] [Medline: 38253798]
18. LangGenius / dify. GitHub. 2025. URL: https://github.com/langgenius/dify [accessed 2025-12-21]
19. SiliconFlow. URL: https://www.siliconflow.com/ [accessed 2025-08-09]
20. Liu A, Mei A, Lin B, et al. DeepSeek-V3.2: pushing the frontier of open large language models. arXiv. Preprint posted online on Dec 2, 2025. [doi: 10.48550/arXiv.2512.02556]
21. Liu H, Yin H, Luo Z, Wang X. Integrating chemistry knowledge in large language models via prompt engineering. Synth Syst Biotechnol 2025;10(1):23-38. [doi: 10.1016/j.synbio.2024.07.004] [Medline: 39206087]
22. Schulhoff S, Ilie M, Balepur N, et al. The prompt report: a systematic survey of prompt engineering techniques. arXiv. Preprint posted online on Jun 6, 2024. [doi: 10.48550/arXiv.2406.06608]
23. Hu Y, Chen Q, Du J, et al. Improving large language models for clinical named entity recognition via prompt engineering. J Am Med Inform Assoc 2024 Sep 1;31(9):1812-1820. [doi: 10.1093/jamia/ocad259] [Medline: 38281112]
24. Jeon S, Kim HG. A comparative evaluation of chain-of-thought-based prompt engineering techniques for medical question answering. Comput Biol Med 2025 Sep;196(Pt A):110614. [doi: 10.1016/j.compbiomed.2025.110614] [Medline: 40602316]
25. Das A, Talati IA, Chaves JMZ, Rubin D, Banerjee I. Weakly supervised language models for automated extraction of critical findings from radiology reports. NPJ Digit Med 2025 May 8;8(1):257. [doi: 10.1038/s41746-025-01522-4] [Medline: 40341617]
26. Shyr C, Hu Y, Bastarache L, et al. Identifying and extracting rare diseases and their phenotypes with large language models. J Healthc Inform Res 2024 Jun;8(2):438-461. [doi: 10.1007/s41666-023-00155-0] [Medline: 38681753]

XSL•FO
RenderX

27. Keshavarz P, Bagherieh S, Nabipoorashrafi SA, et al. ChatGPT in radiology: a systematic review of performance, pitfalls, and future perspectives. Diagn Interv Imaging 2024;105(7-8):251-265. [doi: 10.1016/j.diii.2024.04.003] [Medline: 38679540]
28. Hasani AM, Singh S, Zahergivar A, et al. Evaluating the performance of Generative Pre-Trained Transformer-4 (GPT-4) in standardizing radiology reports. Eur Radiol 2024 Jun;34(6):3566-3574. [doi: 10.1007/s00330-023-10384-x] [Medline: 37938381]
29. Miao BY, Williams CYK, Chinedu-Eneh E, et al. Understanding contraceptive switching rationales from real world clinical notes using large language models. NPJ Digit Med 2025 Apr 23;8(1):221. [doi: 10.1038/s41746-025-01615-0] [Medline: 40269253]
30. Wei WI, Leung CLK, Tang A, McNeil EB, Wong SYS, Kwok KO. Extracting symptoms from free-text responses using ChatGPT among COVID-19 cases in Hong Kong. Clin Microbiol Infect 2024 Jan;30(1):142. [doi: 10.1016/j.cmi.2023.11.002] [Medline: 37949111]
31. Kim J, Leonte KG, Chen ML, et al. Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. NPJ Digit Med 2024 Jul 19;7(1):193. [doi: 10.1038/s41746-024-01181-x] [Medline: 39030292]
32. Salem AC, Gale RC, Fleegle M, Fergadiotis G, Bedrick S. Automating intended target identification for paraphasias in discourse using a large language model. J Speech Lang Hear Res 2023 Dec 11;66(12):4949-4966. [doi: 10.1044/2023_JSLHR-23-00121] [Medline: 37931137]
33. Bellini D, Ferrari R, Vicini S, Rengo M, Saletti CL, Carbone I. Hi ChatGPT, I am a radiologist, how can you help me? Radiol Med 2025 Aug;130(8):1221-1230. [doi: 10.1007/s11547-025-02053-4] [Medline: 40699279]
34. Le Guellec B, Lefèvre A, Geay C, et al. Performance of an open-source large language model in extracting information from free-text radiology reports. Radiol Artif Intell 2024 Jul;6(4):e230364. [doi: 10.1148/ryai.230364] [Medline: 38717292]
35. Guo SB, Shen Y, Meng Y, et al. Surge in large language models exacerbates global regional healthcare inequalities. J Transl Med 2025 Jul 1;23(1):706. [doi: 10.1186/s12967-025-06751-5] [Medline: 40597368]
36. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
37. Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res 2023 Aug 22;25:e48659. [doi: 10.2196/48659] [Medline: 37606976]
38. Habib S, Butt H, Goldenholz SR, Chang CY, Goldenholz DM. Large language model performance on practice epilepsy board examinations. JAMA Neurol 2024 Jun 1;81(6):660-661. [doi: 10.1001/jamaneurol.2024.0676] [Medline: 38587850]
39. Zhong W, Liu Y, Liu Y, et al. Performance of ChatGPT-4o and four open-source large language models in generating diagnoses based on China's rare disease catalog: comparative study. J Med Internet Res 2025 Jun 18;27:e69929. [doi: 10.2196/69929] [Medline: 40532199]
40. Zhong W, Sun M, Yao S, et al. Enhancing the accuracy of human phenotype ontology identification: comparative evaluation of multimodal large language models. J Med Internet Res 2025 Jun 2;27:e73233. [doi: 10.2196/73233] [Medline: 40456109]
41. Shankar SV, Dhingra LS, Aminorroaya A, et al. Automated transformation of unstructured cardiovascular diagnostic reports into structured datasets using sequentially deployed large language models. Eur Heart J Digit Health 2025 Jul;6(4):783-796. [doi: 10.1093/ehjdh/ztaf030] [Medline: 40703108]
42. Somani S, Kim DD, Perez-Guerrero E, et al. Understanding reasons for oral anticoagulation nonprescription in atrial fibrillation using large language models. J Am Heart Assoc 2025 Apr;14(7):e040419. [doi: 10.1161/JAHA.124.040419] [Medline: 40145287]
43. Fang S, Holgate B, Shek A, et al. Extracting epilepsy-related information from unstructured clinic letters using large language models. Epilepsia 2025 Sep;66(9):3369-3384. [doi: 10.1111/epi.18475] [Medline: 40637590]
44. Owens D, Nguyen DQ, Dohopolski M, Rousseau JF, Peterson ED, Navar AM. Accuracy of large language models to identify stroke subtypes within unstructured electronic health record data. Stroke 2025 Oct;56(10):2966-2975. [doi: 10.1161/STROKEAHA.125.051993] [Medline: 40709446]
45. Gu Z, He X, Yu P, et al. Automatic quantitative stroke severity assessment based on Chinese clinical named entity recognition with domain-adaptive pre-trained large language model. Artif Intell Med 2024 Apr;150:102822. [doi: 10.1016/j.artmed.2024.102822] [Medline: 38553162]
46. Spitzl D, Mergen M, Braren R, Endrös L, Eiber M, Steinhelfer L. LLM-powered breast cancer staging from PET/CT reports: a comparative performance study. Int J Med Inform 2025 Dec;204:106053. [doi: 10.1016/j.ijmedinf.2025.106053] [Medline: 40706196]
47. Danhauser K, Wang Y, Klein C, et al. Using large language models to extract information from pediatric clinical reports. PLOS Digit Health 2025 Jul;4(7):e0000919. [doi: 10.1371/journal.pdig.0000919] [Medline: 40700460]
48. Mondillo G, Colosimo S, Perrotta A, Frattolillo V, Masino M. Comparative evaluation of advanced AI reasoning models in pediatric clinical decision support: chatgpt O1 vs. Deepseek-R1. MedRxiv. Preprint posted online on Jan 28, 2025. [doi: 10.1101/2025.01.27.25321169]
49. Arrieta A, Ugarte M, Valle P, Parejo JA, Segura S. O3-mini vs Deepseek-R1: which one is safer? arXiv. Preprint posted online on Jan 30, 2025. [doi: 10.48550/arXiv.2501.18438]

XSL•FO
RenderX

50. Tordjman M, Liu Z, Yuce M, et al. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. Nat Med 2025 Aug;31(8):2550-2555. [doi: 10.1038/s41591-025-03726-3] [Medline: 40267969]
51. Sandmann S, Hegselmann S, Fujarski M, et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. Nat Med 2025 Aug;31(8):2546-2549. [doi: 10.1038/s41591-025-03727-2] [Medline: 40267970]
52. Yan Z, Fan KQ, Zhang Q, et al. Comparative analysis of the performance of the large language models DeepSeek-V3, DeepSeek-R1, Open AI-O3 mini and Open AI-O3 mini high in urology. World J Urol 2025 Jul 7;43(1):416. [doi: 10.1007/s00345-025-05757-4] [Medline: 40622427]
53. Ming S, Yao X, Guo Q, et al. Evaluation of DeepSeek-R1 for ophthalmic diagnosis and reasoning: a comparison with OpenAI o1 and o3. J Med Syst 2025 Oct 8;49(1):130. [doi: 10.1007/s10916-025-02264-2] [Medline: 41060487]
54. Wang Y, Chai Y, Wang J, Gao M, Zang W, Chang Y. Application of copy number variation sequencing technology in 422 foetuses with abnormal ultrasound soft markers. Int J Womens Health 2023;15:1791-1800. [doi: 10.2147/IJWH.S429164] [Medline: 38020944]
55. ENSO Working Group. Role of prenatal magnetic resonance imaging in fetuses with isolated mild or moderate ventriculomegaly in the era of neurosonography: international multicenter study. Ultrasound Obstet Gynecol 2020 Sep;56(3):340-347. [doi: 10.1002/uog.21974] [Medline: 31917496]
56. Jin H, Wang J, Zhang G, et al. A Chinese multicenter retrospective study of isolated increased nuchal translucency associated chromosome anomaly and prenatal diagnostic suggestions. Sci Rep 2021 Mar 10;11(1):5596. [doi: 10.1038/s41598-021-85108-6] [Medline: 33692422]
57. Ji X, Li Q, Qi Y, et al. When NIPT meets WES, prenatal diagnosticians face the dilemma: genetic etiological analysis of 2,328 cases of NT thickening and follow-up of pregnancy outcomes. Front Genet 2023;14:1227724. [doi: 10.3389/fgene.2023.1227724] [Medline: 37600658]
58. Fantasia I, Catagini S, Zamagni G, et al. The clinical impact of the first-trimester nuchal translucency between the 95th-99th percentiles. Prenat Diagn 2023 Jun;43(7):929-936. [doi: 10.1002/pd.6390] [Medline: 37264704]
59. Yoshida S, Miura K, Yamasaki K, et al. Does increased nuchal translucency indicate a fetal abnormality? A retrospective study to clarify the clinical significance of nuchal translucency in Japan. J Hum Genet 2008;53(8):688-693. [doi: 10.1007/s10038-008-0299-6] [Medline: 18500546]
60. Mellis R, Eberhardt RY, Hamilton SJ, et al. Fetal exome sequencing for isolated increased nuchal translucency: should we be doing it? BJOG 2022 Jan;129(1):52-61. [doi: 10.1111/1471-0528.16869] [Medline: 34411415]
61. Zhong W. An adaptive “fast-slow” large language model framework for multi-dimensional classification of prenatal ultrasound reports. Zenodo. 2025. URL: https://zenodo.org/records/16788862 [accessed 2026-05-13]

Abbreviations

CoT: chain-of-thought
FGR: fetal growth restriction
LLMs: large language models
NIPT2.0: enhanced noninvasive prenatal test
NT: nuchal translucency
RAG: retrieval-augmented generation

Edited by M Balcarras; submitted 14.Jan.2026; peer-reviewed by L Lin, Z Li; revised version received 02.May.2026; accepted 04.May.2026; published 28.May.2026.

Please cite as:
Zhong W, Yan H, Liu Y, Liu Y, Yang K, Gao H, Yao Z, Hao W, Yan Y, Yin C
Adaptive Fast-Slow Large Language Model Framework for Multidimensional Classification of Prenatal Ultrasound Reports: Comparative Study
J Med Internet Res 2026;28:e91399
URL: https://www.jmir.org/2026/1/e91399
doi:10.2196/91399

©Wei Zhong, Huihui Yan, Yifan Liu, Yan Liu, Kai Yang, Huimin Gao, Zhengyang Yao, Wenjing Hao, Yousheng Yan, Chenghong Yin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Confidence Measurement Metrics in Multimodal Large Language Models for Ultrasound-Based Radiology Cases: Comparative Evaluation Study of Self-Reported, Consistency-Based, and Hybrid Methods

Taewon Han, MD; Jaeseung Shin, MD, PhD; Jeong Hyun Lee, MD; Kyowon Gu, MD

Department of Radiology, Samsung Medical Center, 81 Irwon-ro, Irwon-dong - Gangnam-gu, Seoul, Republic of Korea

Corresponding Author:
Jaeseung Shin, MD, PhD
Department of Radiology, Samsung Medical Center, 81 Irwon-ro, Irwon-dong - Gangnam-gu, Seoul, Republic of Korea

Abstract

Background: Large language models (LLMs) require specialized methodologies to quantify model confidence for safe deployment in health care systems; however, there is a lack of established methods for confidence assessment.

Objective: This study aimed to evaluate confidence metrics for multimodal LLMs interpreting ultrasound-based radiology cases and to compare self-reported, consistency-based, and hybrid methods.

Methods: From a total of 330 quizzes on the Korean Society of Ultrasound in Medicine digital platform, we selected 94 multiple-choice cases. Four multimodal LLMs were evaluated: 3 reasoning models (GPT-5, Claude-4.5-Sonnet, and Gemini-3-Pro) and 1 general model (GPT-4o). Temperature was fixed at 1.0. Multiple confidence metrics were assessed: (1) self-reported metrics generated by LLMs using prompts that elicited direct confidence percentages with answers, including first self-reported confidence and mean self-reported confidence; (2) consistency-based metrics derived from 20 repeated outputs per case, including relative entropy calculated as $1 - H/\log_2 k$ (H=Shannon entropy, k=number of answer choices) and majority-vote percentage; and (3) a Top Weighted Score combining response frequency with self-reported confidence. Receiver operating characteristic analysis for discrimination and Spearman correlation between accuracy and each confidence metric was conducted. Additionally, model calibration was assessed using expected calibration error and Brier score. Processing time and token consumption (input, output, and total) were recorded for each application programming interface call to evaluate resource use across models.

Results: Diagnostic accuracy varied across models, with Gemini-3-Pro achieving the highest accuracy (70/94, 74.47%), surpassing the median human accuracy (59%, IQR 40.3%-75%). Top Weighted Score, a hybrid metric combining response frequency and self-reported confidence, was the only metric achieving statistically significant correlations across all 4 models: Gemini-3-Pro (ρ=0.52), GPT-5 (ρ=0.43), Claude-4.5-Sonnet (ρ=0.30), and GPT-4o (ρ=0.22). Receiver operating characteristic analysis revealed that Top Weighted Score demonstrated the highest discriminative ability, with area under the curve values of 0.826 (95% CI 0.731 - 0.920) for Gemini-3-Pro and 0.767 (95% CI 0.668 - 0.866) for GPT-5. Top Weighted Score was the only metric achieving statistical significance in GPT-4o. Calibration analysis showed that Top Weighted Score achieved the lowest expected calibration error in GPT-5 (0.098) and Claude-4.5-Sonnet (0.192), while Gemini-3-Pro showed comparable calibration between relative entropy (0.119) and Top Weighted Score (0.122). Resource use analysis demonstrated that reasoning models required substantially longer processing times and higher token consumption compared to general models.

Conclusions: In multimodal LLMs applied to ultrasound-based radiology cases, hybrid methods (Top Weighted Score) demonstrated significant associations across all evaluated models and appear to serve as more reliable indicators of diagnostic confidence compared to self-reported or consistency-based metrics alone, although the strength of these associations varied across models, and external validation is warranted before broader clinical application. These findings support integrative confidence estimation approaches that incorporate response consistency while highlighting the need for resource-efficient sampling strategies to enable practical clinical deployment.

(J Med Internet Res 2026;28:e86498) doi:10.2196/86498

KEYWORDS

artificial intelligence; AI; radiology; medical informatics; diagnostic confidence; large language models; LLMs

XSL•FO
RenderX

Introduction

The integration of large language models (LLMs) into clinical workflows is accelerating from promise to practice [1-3]. These advancements necessitate robust frameworks for evaluating output reliability to ensure patient safety when LLMs are used for medical decision-making [4]. Of particular concern is the calibration of LLMs—the alignment between a model's confidence and its true accuracy—as poorly calibrated LLMs may deliver inaccurate responses with inappropriately high confidence [5], potentially introducing significant risks to patients through downstream diagnostic and treatment errors [6].

Unlike traditional probabilistic classifiers (eg, logistic regression or convolutional neural networks) that expose an explicit class probability for each prediction, LLMs generate text sequentially using probabilities but may present answers confidently despite substantial uncertainty in their underlying probability distributions, resulting in overconfidence issues [4]. This tendency toward overconfidence complicates the safe deployment of LLMs in health care systems and underscores the need for specialized methodologies to quantify model confidence for clinical end users [4].

While existing deep learning models have demonstrated established methods for uncertainty quantification in medical artificial intelligence through various approaches [7-9], architecturally different LLMs still lack standardized methodologies. Several approaches have been proposed to estimate the confidence of LLM outputs [10,11]. In the self-reported method, the model is explicitly prompted to assign a numerical confidence score, typically 0% to 100%, to its own answer [12,13]. Another approach uses sample consistency, leveraging the stochastic behavior of LLMs by running the same prompt multiple times and estimating confidence from the agreement or entropy of the resulting responses [14]. Additional methods include directly using token-level log probabilities to quantify confidence [15]. Despite these various methodologies, a best-practice standard for assessing LLM confidence has not yet been established.

Given the lack of established best practices for confidence assessment, a systematic appraisal of available techniques is essential before these systems can be deployed in clinical practice. Therefore, this study aims to evaluate confidence measurement metrics of multimodal LLMs tasked with ultrasound-based radiological cases assessing whether these approaches can serve as reliable indicators of diagnostic confidence.

Methods

Ethical Considerations

This investigation used publicly available educational datasets, obviating the need for institutional review board approval or informed consent. All quiz materials from the Korean Society of Ultrasound in Medicine (KSUM) website were previously deidentified before public release. No compensation was involved in this study, and no identifiable individuals appear in any images within the manuscript or supplementary materials.

Dataset

A total of 330 case discussion quizzes were extracted by a radiologist (TH, with 4 years of experience) from the KSUM digital platform [16], published between July 28, 2000, and December 25, 2025. The radiologist systematically collected imaging data, question content with corresponding multiple-choice options, and relevant imaging information, including imaging modality and anatomical site. We excluded 236 cases without multiple-choice formats to maintain measurement reliability, resulting in a final dataset of 94 quiz cases (Figure 1). These quiz cases encompassed various imaging modalities, with some cases featuring challenging diagnostic scenarios. To focus on multimodal capabilities, we standardized the brief clinical text to include only patient demographics (age and sex) and chief complaint or previous medical history. The ground truth for our study was established using officially designated answers from the KSUM platform. Human performance benchmarks were derived from response statistics of KSUM platform subscribers, mainly radiologists and radiology trainees with varying degrees of expertise.

XSL•FO
RenderX

Figure 1. Flowchart depicting the evaluation process for consistency metrics across 3 reasoning large language models and 1 general large language model. AUC: area under the curve; KSUM: Korean Society of Ultrasound in Medicine; LLM: large language model; R_H: relative entropy; ROC: receiver operating characteristic.

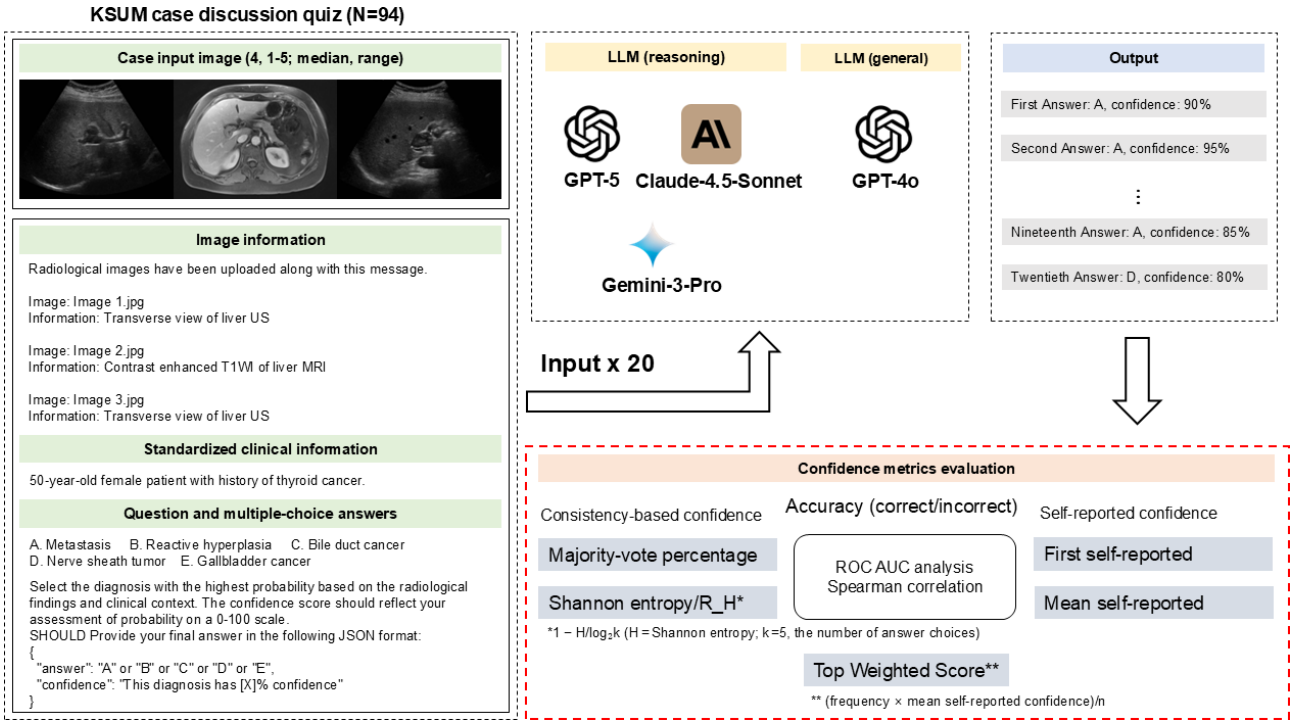

Multimodal LLMs

We selected four multimodal LLMs, including three reasoning models—(1) GPT-5 (Alias: 2025-08-07; OpenAI) [17]; (2) Claude-4.5-Sonnet (Alias: 2025-09-29; Anthropic) [18]; (3) Gemini-3-Pro (Alias: 2025-11-18; Google) [19]—and one general model—GPT-4o (Alias: 2024-11-20; OpenAI) [20]. The temperature was fixed at 1.0 across all models because the reasoning models do not permit adjustment, and previous literature reports optimal performance at this setting [21]. Additionally, enhanced reasoning capabilities were activated for applicable models: “reasoning effort” and “thinking level” were set to “high” for GPT-5 and Gemini-3-Pro-Preview, and “thinking” mode was enabled for Claude-4.5-Sonnet.

Confidence Measurement Metrics

Radiological images were paired with brief text prompts that described the clinical question and key imaging information. We evaluated self-reported, consistency-based, and hybrid confidence metrics while simultaneously measuring diagnostic accuracy (Figure 1). For each case, each model generated 20 independent outputs, enabling the analysis of consistency-based metrics.

First, a relative entropy–based score (R_H) was calculated as $R_H=1-H/\log_2 k$, where $H=-\sum_{i=1 \text{ to } k} p_i \log_2 p_i$; p_i represents Shannon entropy, p_i represents the relative frequency of option i across repeated model outputs, and k=5 is the number of answer choices (so $\log_2 k \approx 2.322$ bits). R_H ranges from 0 (maximum entropy, complete inconsistency) to 1 (zero entropy, perfect consistency); for example, a response pattern of [A A A A A] yields R_H=1, whereas [A B C D E] yields R_H=0. Intermediate patterns receive scores reflecting their coherence; [A A A B B] would yield a higher R_H value than [A A A B C] because of greater consistency. The raw Shannon entropy (H) was also reported alongside R_H to provide an unnormalized measure.

Second, a majority-vote percentage recorded the proportion of the most frequent response in repeated trials. Both [A A A B B] and [A A A B C], for instance, produce a modal proportion of 60%, illustrating that this metric ignores differences in the dispersion of the remaining responses.

Third, a weighted confidence score was calculated for each answer option as follows: (frequency × mean self-reported confidence)/n, where frequency is the count of that option across repeated trials, mean self-reported confidence is the average confidence rating for that option, and n is the total number of repetitions. The Top Weighted Score was defined as the highest score among all options. For example, if option A appeared 12 times with a mean confidence of 80% and option B appeared 8 times with a mean confidence of 90%, the weighted scores would be (12×80)/20=48 for A and (8×90)/20=36 for B, yielding a Top Weighted Score of 48.

In parallel, each model was prompted to append a numerical self-confidence rating (0% - 100%) to each answer. Two self-reported confidence metrics were derived: (1) first self-reported confidence, which used the confidence rating from the first response and (2) mean self-reported confidence, which averaged the confidence ratings across all responses that selected the majority-vote answer. The prompt instructed models to select the diagnosis based on radiological findings and clinical context before providing the final answer in JSON format, including a confidence score on a 0% - 100% scale (eg, “Select the diagnosis with the highest probability… Provide your final answer in the following JSON format: answer: A-E, confidence:

XSL•FO
RenderX

0% - 100%”). The exact templates are presented in Table S1 in Multimedia Appendix 1 and Figure 1.

To determine the minimum repetition count required for consistency-based metrics, analyses were additionally conducted with 5, 10, and 15 repeated outputs per case.

For ROC AUC, Spearman correlation, and calibration analyses, each confidence metric was paired with the diagnostic accuracy of its corresponding representative answer. For consistency-based metrics (R_H and majority-vote percentage) and mean self-reported confidence, the representative answer was the majority-voted option across repeated outputs. For the Top Weighted Score, the representative answer was the option receiving the highest weighted score. For the first self-reported confidence, the representative answer was the model’s first response. Diagnostic accuracy was assessed by comparing the corresponding representative answer for each metric with the KSUM ground truth.

Resource Use

To evaluate the trade-off between confidence estimation reliability and resource efficiency, we recorded the processing time and token consumption for each model query. Processing time was measured as the duration from application programming interface (API) request submission to response completion. Token usage was recorded as input tokens (text prompt and image data), output tokens (model-generated response), and total tokens consumed per query. We calculated the cumulative processing time and token consumption required to analyze a single quiz case for each repetition count condition (1, 5, 10, 15, and 20).

Statistical Analysis

Accuracy for each model and repetition count (1, 5, 10, 15, and 20) was quantified as the proportion of correct answers. Differences across repetition counts were tested using the Cochran *Q* test [22]. Discriminative ability was evaluated using receiver operating characteristic (ROC) area under the curve (AUC) with 95% CIs estimated by the DeLong method [23], using each confidence metric as a predictor of diagnostic accuracy (correct vs incorrect). Spearman correlation coefficients (ρ) measured the association between each confidence measurement and diagnostic accuracy (correct vs incorrect). Correlations were interpreted as negligible (|ρ|≤0.10), weak (0.10<|ρ|≤0.39), moderate (0.39<|ρ|≤0.69), strong (0.69<|ρ|≤0.89), or very strong (|ρ|>0.89) [24]. Calibration, defined as the alignment between predicted confidence and actual accuracy, was evaluated using expected calibration error (ECE) and Brier scores [25,26], using fixed 10-bin calibration, with 95% CIs estimated via bootstrap resampling (1000 iterations). The Brier score was calculated based on binary diagnostic accuracy (correct vs incorrect) rather than the multiclass probability across all 5 options.

Model response repeatability was evaluated using the Fleiss κ statistic, with results interpreted as follows: >0.8, almost perfect; 0.61 to 0.80, substantial; 0.41 to 0.60, moderate; 0.21 to 0.40, fair; and <0.20, poor [27]. To assess potential training data contamination, majority-vote accuracy at a repetition count of 20 was compared between cases published before February 2025 (n=72) and from February to December 2025 (n=22) using Fisher exact tests. Statistical significance was established at *P*<.05. Statistical analyses were performed using GraphPad Prism (version 10.4.1; GraphPad Software) and Python (version 3.10).

Results

Dataset and Diagnostic Performance

A study of 94 cases was conducted, with a median of 4 (IQR 3-4; range 1 - 5) input images per case. The distribution of input image modalities comprised radiography (n=8), ultrasonography (n=94), computed tomography (n=18), magnetic resonance imaging (n=26), nuclear medicine imaging (n=6), and other diagnostic techniques, including endoscopic visualization (n=2) and aspiration fluid (n=2). Human accuracy demonstrated substantial variability (median 59%, IQR 40.3%-75%; range 5% - 96%; mean 56.8%, SD 22%).

Table 1 and Figure 2 demonstrate the implementation of majority voting across repetition counts. Claude-4.5-Sonnet showed a significant improvement in diagnostic accuracy with majority voting, improving from 48.94% (46/94) to 55.32% (52/94) at 10 repetitions (*P*=.01). In contrast, Gemini-3-Pro, GPT-5, and GPT-4o showed no significant change with majority voting (*P*=.67, *P*=.94, and *P*=.08, respectively), with Gemini-3-Pro showing the highest accuracy ranging from 72.34% (68/94) to 74.47% (70/94).

Table . Comparison of first output and majority vote accuracy across multimodal large language models (N=94).[a]

Model	First output, n (%)	Majority vote (5), n (%)	Majority vote (10), n (%)	Majority vote (15), n (%)	Majority vote (20), n (%)	*P* value
Claude-4.5-Sonnet	46 (48.94)	49 (52.13)	52 (55.32)	52 (55.32)	52 (55.32)	.01
Gemini-3-Pro	70 (74.47)	70 (74.47)	69 (73.40)	68 (72.34)	68 (72.34)	.67
GPT-5	63 (67.02)	64 (68)	65 (69.15)	65 (69.15)	64 (68.09)	.94
GPT-4o	41 (43.62)	44 (46.81)	43 (45.74)	44 (46.81)	44 (46.81)	.08

[a]Differences in accuracy were assessed using the Cochran *Q* test. Human accuracy for these cases showed substantial variability (median 59%, IQR 40.3%-75%; mean 56.8%, SD 22%, range 5%-96%).

XSL•FO
RenderX

Figure 2. Accuracy and κ values plotted against repetition number.

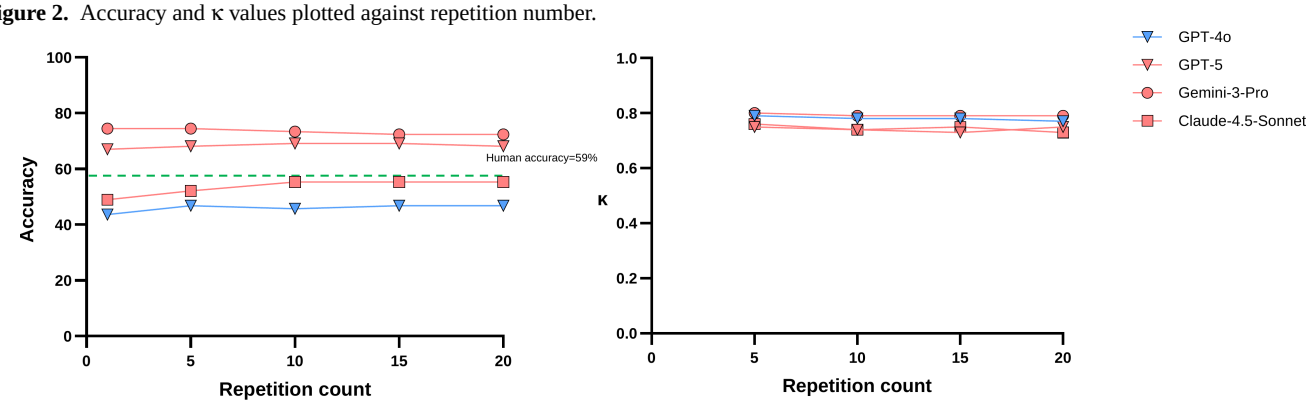

Fleiss κ analysis demonstrated substantial within-model repeatability across all models (Table S2 in Multimedia Appendix 1). Gemini-3-Pro achieved the highest consistency (κ=0.79 - 0.80), followed by GPT-4o (κ=0.77 - 0.79). Claude-4.5-Sonnet and GPT-5 showed comparable repeatability (κ=0.73 - 0.76 and κ=0.73 - 0.75, respectively). All models maintained substantial agreement across all repetition counts, with response consistency remaining stable from 5 to 20 repetitions (Figure 2).

No statistically significant differences in majority-vote accuracy were observed between cases published before and after February 2025 for any of the 4 evaluated models (all *P*>.79; Table S3 in Multimedia Appendix 1).

Confidence Measurement Metrics

The discriminative ability of confidence metrics varied substantially across models (Table 2). Top Weighted Score demonstrated the highest discriminative performance, achieving significant ROC AUC values across all models. Gemini-3-Pro showed the strongest discrimination with Top Weighted Score (ROC AUC=0.826, 95% CI 0.731 - 0.920, *P*<.001), followed by GPT-5 (ROC AUC=0.767, 95% CI 0.668 - 0.866, *P*<.001), Claude-4.5-Sonnet (ROC AUC=0.676, 95% CI 0.568 - 0.785, *P*=.001), and GPT-4o (ROC AUC=0.629, 95% CI 0.509 - 0.749, *P*=.04). Notably, Top Weighted Score was the only metric achieving statistical significance in GPT-4o.

Table . Receiver operating characteristic area under the curve comparing discriminative ability of confidence metrics in multimodal large language models.

Model	Self-reported (first) (ROC[a] AUC[b], 95% CI, *P* value[c])	Self-reported (mean) (ROC AUC, 95% CI, *P* value)	R_H[d] (ROC AUC, 95% CI, *P* value)[e]	Majority-vote percentage (ROC AUC, 95% CI, *P* value)	Top Weighted Score (ROC AUC, 95% CI, *P* value)
Claude-4.5-Sonnet	0.706, 0.602 - 0.810, <.001	0.636, 0.523 - 0.748, .02	0.671, 0.565 - 0.778, .002	0.668, 0.562 - 0.775, .002	0.676, 0.568 - 0.785, .001
Gemini-3-Pro	0.532, 0.439 - 0.625, .50	0.661, 0.546 - 0.775, .006	0.779, 0.672 - 0.887, <.001	0.790, 0.682 - 0.897, <.001	0.826, 0.731 - 0.920, <.001
GPT-5	0.719, 0.613 - 0.826, <.001	0.659, 0.547 - 0.771, .005	0.755, 0.647 - 0.863, <.001	0.740, 0.631 - 0.848, <.001	0.767, 0.668 - 0.866, <.001
GPT-4o	0.597, 0.491 - 0.703, .07	0.592, 0.476 - 0.708, .12	0.576, 0.463 - 0.689, .19	0.577, 0.464 - 0.689, .18	0.629, 0.509 - 0.749, .04

[a]ROC: receiver operating characteristic.

[b]AUC: area under the curve.

[c]Statistically significant results (*P*<.05) are marked with an asterisk.

[d]R_H: relative entropy.

[e]Shannon entropy is not reported separately because it yielded identical AUC values to the relative entropy–based score.

Consistency-based metrics (R_H, majority-vote percentage, and Shannon entropy) showed strong discriminative ability in 3 models. Gemini-3-Pro achieved the highest performance (R_H, ROC AUC=0.779, 95% CI 0.672 - 0.887, *P*<.001; majority-vote percentage, ROC AUC=0.790, 95% CI 0.682 - 0.897, *P*<.001; Shannon entropy, ROC AUC=0.779, 95% CI 0.672 - 0.887, *P*<.001), followed by GPT-5 (R_H, ROC AUC=0.755, 95% CI 0.647 - 0.863, *P*<.001; majority-vote percentage, ROC AUC=0.740, 95% CI 0.631 - 0.848, *P*<.001; Shannon entropy, ROC AUC=0.755, 95% CI 0.647 - 0.863, *P*<.001) and Claude-4.5-Sonnet (R_H, ROC AUC=0.671, 95% CI 0.565 - 0.778, *P*=.002; majority-vote percentage, ROC AUC=0.668, 95% CI 0.562 - 0.775, *P*=.002; Shannon entropy, ROC AUC=0.672, 95% CI 0.565 - 0.778, *P*=.002). However, GPT-4o showed no significant discrimination with consistency-based metrics (R_H, ROC AUC=0.576, *P*=.19; majority-vote percentage, ROC AUC=0.577, *P*=.18; Shannon entropy, ROC AUC=0.576, *P*=.19).

Self-reported confidence demonstrated model-dependent performance. First self-reported confidence achieved significant discrimination in Claude-4.5-Sonnet (ROC AUC=0.706, 95% CI 0.602 - 0.810, *P*<.001) and GPT-5 (ROC AUC=0.719, 95%

XSL•FO
RenderX

CI 0.613 - 0.826, *P*<.001). However, mean self-reported confidence showed lower discriminative ability in Claude-4.5-Sonnet (ROC AUC=0.636, *P*=.02), GPT-5 (ROC AUC=0.659, *P*=.005), and Gemini-3-Pro (ROC AUC=0.661, *P*=.006).

Figure 3 replicates the ROC findings. Top Weighted Score demonstrated higher confidence scores for correct responses across all 4 models, whereas R_H, majority-vote percentage, first self-reported confidence, and mean self-reported confidence showed this pattern only in GPT-5, Claude-4.5-Sonnet, and Gemini-3-Pro, excluding GPT-4o.

Figure 3. Distribution of confidence scores stratified by diagnostic accuracy (correct vs incorrect) for each multimodal large language model. (A) Relative entropy (R_H), (B) majority-vote percentage, (C) first self-reported confidence, (D) mean self-reported confidence, and (E) Top Weighted Score. Solid bars indicate correct responses; hatched bars indicate incorrect responses. Differences between groups were assessed using the Mann-Whitney *U* test. **P*<.05; ***P*<.01; ****P*<.001.

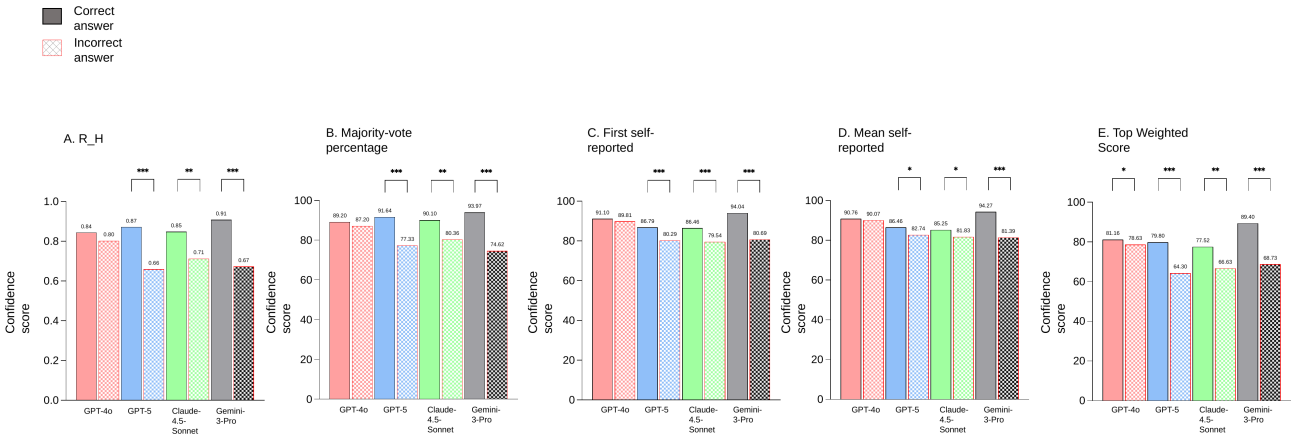

Spearman correlation analysis revealed similar patterns (Table 3). Notably, Top Weighted Score was the only metric achieving statistical significance across all 4 models, demonstrating moderate correlations in Gemini-3-Pro (ρ=0.52, 95% CI 0.35 - 0.65, *P*<.001) and GPT-5 (ρ=0.43, 95% CI 0.25 - 0.58, *P*<.001), and weak correlations in Claude-4.5-Sonnet (ρ=0.30, 95% CI 0.11 - 0.48, *P*=.003) and GPT-4o (ρ=0.22, 95% CI 0.02 - 0.41, *P*=.03). In GPT-4o, Top Weighted Score was the only metric achieving significance. Consistency-based metrics showed moderate correlations in Gemini-3-Pro (R_H, ρ=0.48; majority-vote percentage, ρ=0.50; all *P*<.001) and GPT-5 (R_H, ρ=0.43; majority-vote percentage, ρ=0.41; all *P*<.001). Self-reported confidence showed weak correlations across all models.

XSL•FO
RenderX

Table . Correlation analysis between accuracy and confidence metrics in multimodal large language models at a repetition count of 20.

Model and metrics	ρ (95% CI)[a]	*P* value[b]
Claude-4.5-Sonnet		
Self-reported (first)	0.36 (0.17 to 0.53)	<.001
Self-reported (mean)	0.23 (0.03 to 0.42)	.02
R_H[c]	0.31 (0.11 to 0.48)	.002
Majority-vote percentage	0.30 (0.11 to 0.48)	.003
Top Weighted Score	0.30 (0.11 to 0.48)	.003
Gemini-3-Pro		
Self-reported (first)	0.06 (−0.14 to 0.26)	.55
Self-reported (mean)	0.25 (0.05 to 0.43)	.014
R_H	0.48 (0.30 to 0.62)	<.001
Majority-vote percentage	0.50 (0.33 to 0.64)	<.001
Top Weighted Score	0.52 (0.35 to 0.65)	<.001
GPT-5		
Self-reported (first)	0.36 (0.17 to 0.52)	<.001
Self-reported (mean)	0.26 (0.06 to 0.44)	.012
R_H	0.43 (0.25 to 0.58)	<.001
Majority-vote percentage	0.41 (0.22 to 0.56)	<.001
Top Weighted Score	0.43 (0.25 to 0.58)	<.001
GPT-4o		
Self-reported (first)	0.18 (−0.02 to 0.37)	.08
Self-reported (mean)	0.16 (−0.05 to 0.35)	.13
R_H	0.14 (−0.07 to 0.33)	.18
Majority-vote percentage	0.14 (−0.06 to 0.33)	.18
Top Weighted Score	0.22 (0.02 to 0.41)	.03

[a]Values represent Spearman correlation coefficients (ρ) with 95% CIs in parentheses.

[b]Significant correlations (*P*<.05) are marked with an asterisk.

[c]R_H: relative entropy.

Calibration

Calibration analysis revealed that the Top Weighted Score demonstrated the best calibration in GPT-5 (ECE=0.098, 95% CI 0.074 - 0.211; Brier score=0.185, 95% CI 0.140 - 0.235) and Claude-4.5-Sonnet (ECE=0.192, 95% CI 0.133 - 0.307; Brier score=0.259, 95% CI 0.203 - 0.317). In Gemini-3-Pro, R_H and the Top Weighted Score showed comparable calibration (ECE=0.119 vs 0.122; Brier score=0.164 vs 0.163, respectively). Across Claude-4.5-Sonnet, Gemini-3-Pro, and GPT-5, consistency-based metrics, particularly R_H, demonstrated better calibration compared to self-reported metrics. GPT-4o showed the poorest calibration across all metrics (Table 4).

XSL•FO
RenderX

Table . Calibration metrics for different confidence measurement methods in multimodal large language models at a repetition count of 20.[a]

Model and metrics	ECE[b] (95% CI)	Brier score (95% CI)
Claude-4.5-Sonnet		
Self-reported (mean)	0.284 (0.195 - 0.389)	0.317 (0.253 - 0.385)
Self-reported (first)	0.340 (0.244 - 0.439)	0.339 (0.275 - 0.409)
R_H[c]	0.266 (0.191 - 0.375)	0.288 (0.216 - 0.363)
Majority-vote percentage	0.304 (0.226 - 0.403)	0.321 (0.247 - 0.399)
Top Weighted Score	0.192 (0.133 - 0.307)	0.259 (0.203 - 0.317)
Gemini-3-Pro		
Self-reported (mean)	0.216 (0.128 - 0.305)	0.243 (0.165 - 0.322)
Self-reported (first)	0.208 (0.120 - 0.298)	0.229 (0.152 - 0.313)
R_H	0.119 (0.078 - 0.210)	0.164 (0.109 - 0.226)
Majority-vote percentage	0.168 (0.109 - 0.264)	0.178 (0.117 - 0.244)
Top Weighted Score	0.122 (0.071 - 0.212)	0.163 (0.109 - 0.220)
GPT-5		
Self-reported (mean)	0.172 (0.097 - 0.273)	0.235 (0.171 - 0.303)
Self-reported (first)	0.176 (0.096 - 0.278)	0.230 (0.164 - 0.296)
R_H	0.140 (0.099 - 0.240)	0.191 (0.135 - 0.252)
Majority-vote percentage	0.206 (0.130 - 0.299)	0.219 (0.154 - 0.287)
Top Weighted Score	0.098 (0.074 - 0.211)	0.185 (0.140 - 0.235)
GPT-4o		
Self-reported (mean)	0.436 (0.329 - 0.543)	0.436 (0.349 - 0.523)
Self-reported (first)	0.468 (0.368 - 0.570)	0.459 (0.380 - 0.542)
R_H	0.377 (0.303 - 0.494)	0.397 (0.313 - 0.482)
Majority-vote percentage	0.413 (0.332 - 0.535)	0.435 (0.345 - 0.525)
Top Weighted Score	0.356 (0.272 - 0.470)	0.368 (0.295 - 0.440)

[a]Lower values indicate better calibration. ECE was calculated using 10 bins.

[b]ECE: expected calibration error.

[c]R_H: relative entropy.

Repetition Count Analysis

Correlation analysis stratified by repetition count revealed model-specific patterns in the relationship between consistency-based metrics and diagnostic accuracy (Table S4 in Multimedia Appendix 1). Gemini-3-Pro showed strengthening correlations with increasing repetitions, with majority-vote percentage and R_H improving from weak (ρ=0.35, $P<.001$) at 5 repetitions to moderate (ρ=0.51 - 0.52, $P<.001$) at 15 repetitions, and Top Weighted Score improving from weak (ρ=0.32, $P=.002$) to moderate (ρ=0.46, $P<.001$). Mean self-reported confidence in Gemini-3-Pro achieved statistical significance only after 10 repetitions. In contrast, Claude-4.5-Sonnet demonstrated significant weak correlations across all metrics after 5 repetitions. GPT-5 showed improvement in R_H from weak (ρ=0.38, $P<.001$) at 5 repetitions to moderate (ρ=0.41, $P<.001$) at 15 repetitions, while Top Weighted Score maintained moderate correlations across all repetition counts (ρ=0.40 - 0.51; all $P<.001$). GPT-4o showed nonsignificant correlations across all metrics regardless of repetition count (Figure 4).

XSL•FO
RenderX

Figure 4. Heatmap of Spearman correlation (ρ) and calibration (ECE) between diagnostic accuracy and confidence metrics across multimodal large language models by repetition count. ECE: expected calibration error; R_H: relative entropy.

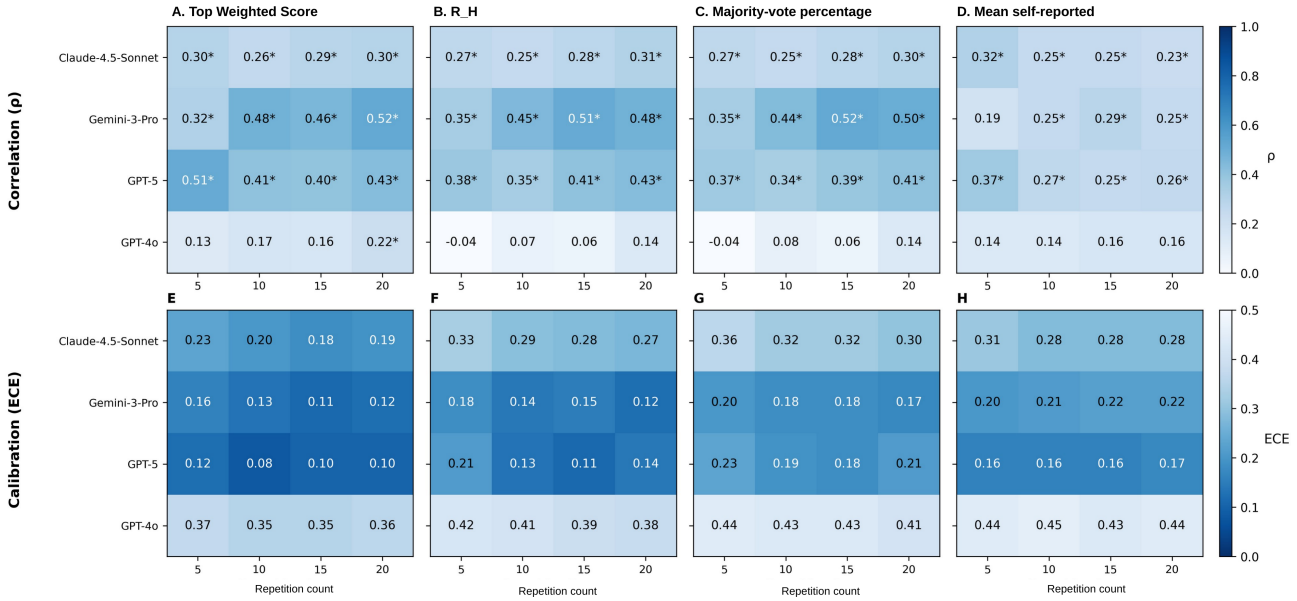

Calibration metrics showed similar patterns with increasing repetitions (Table S5 in Multimedia Appendix 1). Top Weighted Score demonstrated decreasing ECE values across repetitions in Gemini-3-Pro (0.158 to 0.112), GPT-5 (0.118 to 0.095), and Claude-4.5-Sonnet (0.230 to 0.184).

Heatmaps display Spearman correlation coefficients (ρ; top row, panels A-D) and ECE (bottom row, panels E-H) for 4 confidence metrics across repetition counts of 5, 10, 15, and 20. For correlation, darker shading indicates a stronger positive correlation (range 0 - 1); for ECE, darker shading indicates better calibration with lower error (range 0 - 0.5). Statistically significant correlations (P<.05) are marked with an asterisk. Shannon entropy exhibited inverse correlations with identical magnitudes to R_H values (data not shown).

Resource Use

Resource consumption varied substantially across models (Table S6 in Multimedia Appendix 1). GPT-4o demonstrated the highest efficiency, requiring a mean processing time of 6.38 (SD 2.26) seconds and a mean of 4353 (SD 1402) total tokens per case at 1 repetition, increasing to 59.43 (SD 12.46) seconds and 43,423 (SD 13,921) tokens at 10 repetitions, and 117.93 (SD 23.44) seconds and 86,877 (SD 27,869) tokens at 20 repetitions. Claude-4.5-Sonnet required a mean processing time of 29.48 (SD 6.37) seconds and a mean of 5551 (SD 1477) total tokens at 1 repetition, increasing to 288.27 (SD 44.76) seconds and 55,207 (SD 14,505) tokens at 10 repetitions, and 750.43 (SD 1694.80) seconds and 110,344 (SD 28,860) tokens at 20 repetitions. Gemini-3-Pro consumed a mean processing time of 54.75 (SD 29.79) seconds and a mean of 8774 (SD 2993) total tokens at 1 repetition, increasing to 543.87 (SD 256.35) seconds and 88,621 (SD 26,743) tokens at 10 repetitions, and 1103.97 (SD 528.83) seconds and 178,114 (SD 54,325) tokens at 20 repetitions. GPT-5 showed the highest resource consumption, requiring a mean processing time of 70.66 (SD 39.39) seconds and a mean of 6348 (SD 2026) total tokens at 1 repetition, increasing to 947.53 (SD 1808.41) seconds and 63,956 (SD 17,991) tokens at 10 repetitions, and 1768.43 (SD 2059.84) seconds and 127,879 (SD 35,210) tokens at 20 repetitions. Notably, sporadic extreme processing times exceeding 500 seconds per individual API call were observed in Claude-4.5-Sonnet (2/1880, 0.11%) calls and GPT-5 (5/1880, 0.27%) calls, whereas no such events occurred in Gemini-3-Pro or GPT-4o.

Discussion

Principal Findings

Our findings indicate that the Top Weighted Score, a composite metric integrating response consistency with self-reported confidence, provided the most consistent assessment of multimodal LLM output reliability for ultrasound-based radiological cases. Notably, it was the only metric to demonstrate statistically significant correlations across all 4 evaluated models, and it exhibited the best calibration in most models. In parallel, consistency-based metrics (R_H, majority-vote percentage, and Shannon entropy) showed strong discriminative performance in Gemini-3-Pro, GPT-5, and Claude-4.5-Sonnet when contrasted with self-reported confidence, underscoring the value of response agreement–derived signals for reliability estimation.

Comparison to Prior Work

These observations align with prior studies suggesting that consistency-based calibration approaches can outperform post hoc verbalized confidence methods for estimating LLM uncertainty and reliability [11,28,29]. In radiology, Huppertz et al [30] similarly reported no significant association between verbalized confidence and diagnostic accuracy, with accuracy remaining below 50% even at the highest confidence scores. However, in our study, first-response confidence showed relatively high ROC AUC values in GPT-5 and Claude-4.5-Sonnet, whereas mean self-reported confidence demonstrated lower values; notably, Gemini-3-Pro showed the opposite pattern. This inconsistency suggests that the apparent first-response advantage may have been driven by sampling

XSL•FO
RenderX

variability. Accordingly, repeated averaging of verbalized confidence scores represents an averaging of scores only partially aligned with actual correctness rather than a more accurate probability estimate [4], and this process may compress case-level variance and reduce discriminative ability. Given the intrinsically stochastic nature of LLM generation, reliance on a single initial output for confidence estimation poses significant risks in clinical settings.

Interestingly, the entropy-based metrics achieved lower ECE and Brier scores than the simple majority-vote percentage metric. This suggests that model response dispersion, rather than only the most frequent response, yields better-calibrated confidence estimates. Given the high κ values observed in our study, the benefit of entropy-based metrics is likely to increase in settings with greater response diversity, such as tasks with larger multiple-choice panels or free-text outputs. However, high interresponse agreement (κ≈0.7 - 0.8) and fixed sampling parameters (eg, temperature=1) may inflate confidence and reduce discriminative power.

Additionally, although not statistically significant, Gemini-3-Pro, the highest-performing model, showed a decrease in accuracy with majority voting at 15 and 20 repetitions. This may partly reflect the systematic nature of its errors across repetitions. When interresponse agreement is high (κ=0.79 - 0.80), incorrect responses tend to be consistent across repetitions, which may limit the corrective potential of majority voting and contribute to the marginal decrease in accuracy observed.

Despite broad improvements in capability across successive model releases, recent studies across domains indicate that newer models can retain a systematic tendency toward overconfidence [10,31,32]. Consistent with this literature, we observed that models frequently assigned high confidence to incorrect answers, which poses a direct challenge for deployment in clinical decision support, where confidently presented errors may be disproportionately persuasive and propagate into downstream decision-making [33,34]. Several recent studies have reported that hybrid approaches integrating consistency and verbalized confidence can outperform either method alone in certain models [4,28,35]. In our study, Top Weighted Score operationalizes a related principle by weighting candidate responses using both frequency and self-reported confidence, potentially mitigating 2 complementary limitations: (1) overconfidence inherent to verbalized estimates and (2) reduced sensitivity of pure consistency metrics when interresponse agreement is high. The consistent performance of Top Weighted Score across all tested models supports the utility of such integrative formulations for multimodal radiology tasks.

A major practical limitation of consistency-based estimation is resource intensity, as multiple independent model executions are required to compute agreement and dispersion metrics, increasing API costs and processing time [11]. In our study, processing time varied substantially by model and repetition count, with the burden particularly evident for reasoning models, where extended reasoning traces can materially increase latency and necessitate explicit cost-benefit trade-offs. To mitigate this overhead, adaptive sampling strategies have been proposed that achieve comparable accuracy with significantly reduced computational costs [35,36]. In our repetition-depth analysis, approximately 10 resampling runs per case were sufficient to stabilize discrimination and calibration estimates. Further research is needed to establish minimal sampling schedules for each model, along with inference acceleration strategies leveraging recent advances [37], to balance confidence estimation reliability with resource efficiency. Furthermore, sporadic extreme processing times were observed in Claude-4.5-Sonnet and GPT-5, likely reflecting potential API latency or timeout events. Although this processing time instability affected fewer than 0.3% of total API calls, it disproportionately inflated processing time variability and may represent a real-world barrier to clinical deployment, where predictable response times are essential.

Limitations

This study exhibited several limitations. First, our relatively small sample size (N=94) potentially restricts the generalizability of our findings, and the fixed 10-bin ECE may be unstable in this sample, as confidence score clustering may leave some bins sparsely populated. Second, because model performance was assessed with multiple-choice questions, the evaluation does not fully represent real-world clinical scenarios that typically require free-text responses. Third, our evaluation was limited to the most widely available closed-source models and did not include open-source models that might have high potential applicability in health care. Fourth, as the KSUM educational quizzes are publicly accessible online, there is a possibility that these materials may have been included in the training corpora of the evaluated LLMs. This potential data contamination could inflate model accuracy estimates. However, a temporal holdout analysis revealed no significant accuracy differences between precutoff and postcutoff cases for any model (Table S3 in Multimedia Appendix 1). Furthermore, our previous investigation using the same dataset demonstrated no significant performance differences for GPT-4o between cases published before versus after the model's knowledge cutoff date [38], suggesting that even if contamination occurred, its impact on model performance may be negligible given the vast scale of training parameters. Fifth, a practical limitation of consistency-based estimation is resource intensity, as multiple independent model executions are required, increasing costs and processing time. To address this, we conducted our repetition count analysis, and approximately 10 resampling runs per case appeared to be sufficient to stabilize discrimination and calibration estimates. Sixth, although processing times indicated that reasoning models engaged in extended computation, the structured JSON output requirement may have partially constrained the models' chain-of-thought process, potentially affecting confidence calibration. Additionally, our prompt was designed such that confidence values were embedded as natural language strings within the JSON output rather than as direct numerical fields, which may introduce parsing instability if models generate slight textual variations. Seventh, our evaluation used general-purpose multimodal LLMs. Although recent studies have evaluated general-purpose multimodal LLMs on radiological cases, these investigations predominantly focus on diagnostic accuracy [3,21,38,39]. The

XSL•FO
RenderX

development of radiologic domain-specific multimodal LLMs is in its early stages [40,41], and such models are not yet widely available commercially. Building upon these studies of general-purpose LLMs, our findings may contribute to the future development and evaluation of domain-specific expert models. Finally, the Top Weighted Score was developed post hoc on the study dataset, which introduces a potential risk of overfitting. Therefore, external validation using independent datasets is necessary to confirm the generalizability of this hybrid metric.

Future Directions

Several directions for future research emerge from our findings. First, validation with larger datasets across various radiological conditions would strengthen generalizability and potentially identify additional patterns in model performance and calibration. Second, incorporating free-text tasks would better approximate real-world clinical usage. Third, comparative studies should explore whether open-source models demonstrate similar confidence calibration patterns and assess their applicability in clinical environments. Fourth, further research is needed to establish minimal sampling schedules and develop adaptive sampling strategies to balance confidence estimation reliability with resource efficiency. Additionally, studies using models with accessible internal reasoning processes could enable direct entropy computation from log probabilities, potentially reducing computational overhead while maintaining calibration accuracy. Fifth, comparative studies evaluating human reader confidence alongside LLM confidence metrics during diagnostic tasks could provide valuable insights into the potential for LLMs to augment clinical decision-making. Finally, from a clinical perspective, confidence metrics could augment human decision-making by providing reliability indicators for LLM outputs; however, clinical validation studies are needed to evaluate the practical use of these metrics in real-world diagnostic settings.

Conclusions

In multimodal LLMs applied to ultrasound-based radiology cases, hybrid methods (Top Weighted Score) demonstrated significant associations across all evaluated models and appear to serve as more reliable indicators of diagnostic confidence compared to self-reported or consistency-based metrics alone, although the strength of this association varied across models, and external validation is warranted before broader clinical application.

Acknowledgments

The authors used Claude-4.5-Sonnet (Anthropic) for language editing and grammar checking during manuscript preparation. All content was verified and approved by the authors. The large language models evaluated in this study (GPT-4o, GPT-5, Claude-4.5-Sonnet, and Gemini-3-Pro) are described in the Methods section.

Funding

This work was supported by a grant from the National Research Foundation of Korea (NRF), funded by the Korean government (MSIT) (grant RS-2025-00516874).

Data Availability

The research data supporting the findings of this study are publicly available through the Korean Society of Ultrasound in Medicine (KSUM) [16]. This study used these open-source datasets for analysis. The underlying code for this study, including calibration analysis, is provided in Supplementary Material 1 in Multimedia Appendix 2 and will be made publicly available upon publication.

Authors' Contributions

TH and JS conceived and designed the study. TH, JHL, and KG acquired the data. TH, JHL, KG, and JS analyzed and interpreted the data. TH and JS prepared the first draft of the manuscript. All authors critically revised the manuscript for important intellectual content, approved the final manuscript, and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary table.
[DOCX File, 34 KB - jmir_v28i1e86498_app1.docx]

Multimedia Appendix 2
Calibration new analysis code.
[TXT File, 8 KB - jmir_v28i1e86498_app2.txt]

References

XSL•FO
RenderX

1. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA 2023 Sep 5;330(9):866-869. [doi: 10.1001/jama.2023.14217] [Medline: 37548965]
2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
3. Zhou Y, Ong H, Kennedy P, et al. Evaluating GPT-v4 (GPT-4 with vision) on detection of radiologic findings on chest radiographs. Radiology 2024 May;311(2):e233270. [doi: 10.1148/radiol.233270] [Medline: 38713028]
4. Xiong M, Hu Z, Lu X, Li Y, Fu J, He J. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. Presented at: 12th International Conference on Learning Representations, ICLR 2024; May 7-11, 2024 URL: https://researchportal.hkust.edu.hk/en/publications/can-llms-express-their-uncertainty-an-empirical-evaluation-of-con/? [accessed 2026-05-22]
5. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. Presented at: ICML'17: Proceedings of the 34th International Conference on Machine Learning; Aug 6-11, 2017 URL: https://proceedings.mlr.press/v70/guo17a/guo17a.pdf [accessed 2026-05-22]
6. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs). NPJ Digit Med 2024 Jul 8;7(1):183. [doi: 10.1038/s41746-024-01157-x] [Medline: 38977771]
7. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. Presented at: ICML'16: Proceedings of the 33rd International Conference on International Conference on Machine Learning; Jun 19-24, 2016 URL: https://proceedings.mlr.press/v48/gal16.pdf [accessed 2026-05-22]
8. Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. Nat Mach Intell 2019 Jan 7;1(1):20-23. [doi: 10.1038/s42256-018-0004-1]
9. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digit Med 2021 Jan 5;4(1):4. [doi: 10.1038/s41746-020-00367-3] [Medline: 33402680]
10. Omar M, Agbareia R, Glicksberg BS, Nadkarni GN, Klang E. Benchmarking the confidence of large language models in clinical questions. medRxiv. Preprint posted online on Sep 10, 2024. [doi: 10.1101/2024.08.11.24311810]
11. Savage T, Wang J, Gallo R, et al. Large language model uncertainty measurement and calibration for medical diagnosis and treatment. medRxiv. Preprint posted online on Jun 7, 2024. [doi: 10.1101/2024.06.06.24308399]
12. Tian K, Mitchell E, Zhou A, et al. Just ask for calibration: strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. Presented at: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Dec 6-10, 2023. [doi: 10.18653/v1/2023.emnlp-main.330]
13. Yang D, Tsai YHH, Yamada M. On verbalized confidence scores for LLMs. arXiv. Preprint posted online on Dec 19, 2024. [doi: 10.48550/arXiv.2412.14737]
14. Raj H, Rosati D, Majumdar S. Measuring reliability of large language models through semantic consistency. arXiv. Preprint posted online on Nov 10, 2022. [doi: 10.48550/arXiv.2211.05853]
15. Fadeeva E, Rubashevskii A, Shelmanov A, et al. Fact-checking the output of large language models via token-level uncertainty quantification. Presented at: Findings of the Association for Computational Linguistics: ACL 2024; Aug 11-16, 2024. [doi: 10.18653/v1/2024.findings-acl.558]
16. Education. Korean Society of Ultrasound in Medicine (KSUM). URL: https://www.ksum.or.kr/education/index.php [accessed 2026-05-23]
17. Singh A, Fry A, Perelman A, Tart A, Ganesh A, El-Kishky A. OpenAI GPT-5 system card. arXiv. Preprint posted online on Dec 19, 2025. [doi: 10.48550/arXiv.2601.03267]
18. Introducing Claude Sonnet 4.5. Anthropic. 2025. URL: https://www.anthropic.com/news/claude-sonnet-4-5 [accessed 2026-05-22]
19. Model cards. Google DeepMind. 2025. URL: https://deepmind.google/models/model-cards/ [accessed 2026-05-22]
20. Hurst A, Lerer A, Goucher AP, Perelman A, Ramesh A, Clark A. GPT-4o system card. arXiv. Preprint posted online on Oct 25, 2024. [doi: 10.48550/arXiv.2410.21276]
21. Suh PS, Shim WH, Suh CH, et al. Comparing diagnostic accuracy of radiologists versus GPT-4V and Gemini Pro Vision using image inputs from Diagnosis Please cases. Radiology 2024 Jul;312(1):e240273. [doi: 10.1148/radiol.240273] [Medline: 38980179]
22. Cochran WG. The comparison of percentages in matched samples. Biometrika 1950 Dec;37(3-4):256-266. [doi: 10.1093/biomet/37.3-4.256] [Medline: 14801052]
23. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988 Sep;44(3):837-845. [doi: 10.2307/2531595] [Medline: 3203132]
24. Schober P, Boer C, Schwarte LA. Correlation coefficients: appropriate use and interpretation. Anesth Analg 2018 May;126(5):1763-1768. [doi: 10.1213/ANE.0000000000002864] [Medline: 29481436]
25. Rivera M, Godbout JF, Rabbany R, Pelrine K. Combining confidence elicitation and sample-based methods for uncertainty quantification in misinformation mitigation. Presented at: Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024); Mar 22, 2024. [doi: 10.18653/v1/2024.uncertainlp-1.12]

XSL•FO
RenderX

26. Brier GW. Verification of forecasts expressed in terms of probability. Mon Wea Rev 1950 Jan;78(1):1-3. [doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2]
27. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]
28. de Oliveira R, Garber M, Gwinnutt JM, et al. A study of calibration as a measurement of trustworthiness of large language models in biomedical natural language processing. JAMIA Open 2025;8(4):ooaf058. [doi: 10.1093/jamiaopen/ooaf058] [Medline: 40655536]
29. Lyu Q, Shridhar K, Malaviya C, et al. Calibrating large language models with sample consistency. Presented at: Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25); Feb 25 to Mar 4, 2025. [doi: 10.1609/aaai.v39i18.34120]
30. Huppertz MS, Siepmann R, Topp D, et al. Revolution or risk? Assessing the potential and challenges of GPT-4V in radiologic image interpretation. Eur Radiol 2025 Mar;35(3):1111-1121. [doi: 10.1007/s00330-024-11115-6] [Medline: 39422726]
31. Naderi N, Safavi-Naini SAA, Savage T, Atf Z, Lewis P, Nadkarni G. Self-reported confidence of large language models in gastroenterology: analysis of commercial, open-source, and quantized models. arXiv. Preprint posted online on May 24, 2025. [doi: 10.48550/arXiv.2503.18562]
32. Vashurin R, Fadeeva E, Vazhentsev A, et al. Benchmarking uncertainty quantification methods for large language models with LM-Polygraph. Trans Assoc Comput Linguist 2025 Mar 19;13:220-248. [doi: 10.1162/tacl_a_00737]
33. Omar M, Nassar S, Hijazi K, Glicksberg BS, Nadkarni GN, Klang E. Generating credible referenced medical research: a comparative study of OpenAI's GPT-4 and Google's Gemini. Comput Biol Med 2025 Feb;185:109545. [doi: 10.1016/j.compbiomed.2024.109545] [Medline: 39667055]
34. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. Crit Care 2023 Mar 21;27(1):120. [doi: 10.1186/s13054-023-04393-x] [Medline: 36945051]
35. Taubenfeld A, Sheffer T, Ofek E, et al. Confidence improves self-consistency in LLMs. Presented at: Findings of the Association for Computational Linguistics; Jul 27 to Aug 1, 2025. [doi: 10.18653/v1/2025.findings-acl.1030]
36. Aggarwal P, Madaan A, Yang Y. Let's sample step by step: adaptive-consistency for efficient reasoning and coding with LLMs. Presented at: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Dec 6-10, 2023. [doi: 10.18653/v1/2023.emnlp-main.761]
37. He C, Huang Y, Mu P, Miao Z, Xue J, Ma L. WaferLLM: large language model inference at Wafer scale. Presented at: 19th USENIX Symposium on Operating Systems Design and Implementation (OSDI '25); Jul 7-9, 2025 URL: https://www.usenix.org/system/files/osdi25-he.pdf [accessed 2026-05-22]
38. Han T, Jeong WK, Shin J. Diagnostic performance of multimodal large language models in radiological quiz cases: the effects of prompt engineering and input conditions. Ultrasonography 2025 May;44(3):220-231. [doi: 10.14366/usg.25012] [Medline: 40235070]
39. Schramm S, Preis S, Metz MC, et al. Impact of multimodal prompt elements on diagnostic performance of GPT-4V in challenging brain MRI cases. Radiology 2025 Jan;314(1):e240689. [doi: 10.1148/radiol.240689] [Medline: 39835982]
40. Hong EK, Ham J, Roh B, et al. Diagnostic accuracy and clinical value of a domain-specific multimodal generative AI model for chest radiograph report generation. Radiology 2025 Mar;314(3):e241476. [doi: 10.1148/radiol.241476] [Medline: 40131111]
41. Hong EK, Roh B, Park B, et al. Value of using a generative AI model in chest radiography reporting: a reader study. Radiology 2025 Mar;314(3):e241646. [doi: 10.1148/radiol.241646] [Medline: 40067108]

Abbreviations

API: application programming interface
AUC: area under the curve
ECE: expected calibration error
KSUM: Korean Society of Ultrasound in Medicine
LLM: large language model
R_H: relative entropy
ROC: receiver operating characteristic

XSL•FO
RenderX

Edited by A Coristine; submitted 25.Oct.2025; peer-reviewed by A Mitra, D Hu, EL Epstein, M Sevenster; revised version received 15.May.2026; accepted 18.May.2026; published 02.Jun.2026.

Please cite as:
Han T, Shin J, Lee JH, Gu K
Confidence Measurement Metrics in Multimodal Large Language Models for Ultrasound-Based Radiology Cases: Comparative Evaluation Study of Self-Reported, Consistency-Based, and Hybrid Methods
J Med Internet Res 2026;28:e86498
URL: https://www.jmir.org/2026/1/e86498
doi:10.2196/86498

© Taewon Han, Jaeseung Shin, Jeong Hyun Lee, Kyowon Gu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Performance Evaluation of GPT-5, Grok 4, and DeepSeek R1 in Interpreting Complete Blood Count Reports for Hematologic Diseases: Retrospective Comparative Study

Xianfei Ye[1,2,3], MM; Xinglun Qi[1], BS; Lina Fan[1], BS; Qian Yu[1], BMed; Suming Zhou[1], BS; Chunyun Ren[1], MM; Dagan Yang[1], MS

[1]Department of Laboratory Medicine, The First Affiliated Hospital, Zhejiang University School of Medicine, 79 Qingchun Road, Hangzhou, Zhejiang, China

[2]Key Laboratory of Clinical In Vitro Diagnostic Techniques of Zhejiang Province, Hangzhou, China

[3]Institute of Laboratory Medicine, Zhejiang University, Hangzhou, China

Corresponding Author:
Dagan Yang, MS
Department of Laboratory Medicine, The First Affiliated Hospital, Zhejiang University School of Medicine, 79 Qingchun Road, Hangzhou, Zhejiang, China

Abstract

Background: Large language models (LLMs) demonstrate potential in the laboratory, yet rigorous clinical evaluation remains limited. The opacity of LLM decision-making constrains their safe application in interpreting complete blood count (CBC) reports for hematologic diseases.

Objective: This study aimed to conduct an exploratory evaluation of GPT-5, Grok 4, and DeepSeek R1 in interpreting real-world CBC reports, particularly their reasoning capabilities and clinical safety.

Methods: This single-center retrospective study analyzed 100 CBC reports from initial-visit patients with hematologic conditions. After responses were generated by the 3 LLMs using standardized Chinese prompts, four trained laboratory physicians blindly evaluated them across 6 quality and 5 task dimensions. Interrater reliability was assessed using intraclass correlation coefficients (ICCs), and performance differences were assessed based on 4-rater consensus scores and Friedman and Wilcoxon tests. For task 4 (ablation analysis), the McNemar test was used to compare top-1 diagnostic concordance with the gold-standard diagnosis within each model, with and without initial clinical suspicion in the prompt. Error types and distributions were documented during the task evaluation.

Results: DeepSeek R1 demonstrated excellent interrater reliability across most quality dimensions (ICC ≥0.75). In the quality dimension, DeepSeek R1 significantly outperformed the other models in comprehensiveness, accuracy, clarity, relevance, and practicality. In the task 4 evaluation, GPT-5 demonstrated the highest concordance (93/100, 93%) with gold-standard diagnoses, followed by DeepSeek R1 (92/100, 92%) and Grok 4 (89/100, 89%). After removing the initial clinical suspicion, these rates decreased to 79% (79/100), 77% (77/100), and 72% (72/100), representing statistically significant within-model reductions for all models (P<.001). Post hoc error analysis revealed distinct patterns across task dimensions. GPT-5 exhibited 12 hallucinations in the analyzer alert processing task; DeepSeek R1 demonstrated 1 hallucination in the abnormal item identification task, whereas Grok 4 displayed none. All models exhibited reasoning errors and varying degrees of deficiencies in the correlation analysis and preliminary diagnosis tasks, characterized by unwarranted inferences of disease status from isolated results without clinical integration. Grok 4 generated 9 reasoning errors in the clinical management task by providing generic recommendations not tailored to case-specific CBC data, potentially compromising individualized treatment decisions.

Conclusions: While current LLMs demonstrate potential for interpreting CBC reports in hematologic diseases, they show performance heterogeneity across models. The ablation study findings underscore the necessity of integrating clinical context for accurate laboratory test interpretation. Low scores, hallucinations, and reasoning errors in model outputs indicate that current clinical deployment requires human oversight and quality control. As this single-center, Chinese-language exploratory assessment provides only preliminary, possibly context-dependent evidence, multicenter, cross-lingual prospective validation is needed to delineate the practical boundaries and safety standards for clinical deployment.

(J Med Internet Res 2026;28:e87802) doi:10.2196/87802

KEYWORDS

large language models; hematologic diseases; ChatGPT; Grok; DeepSeek; hallucination; artificial intelligence; AI

XSL•FO
RenderX

Introduction

The rapid advancement of next-generation information technologies has enabled large language models (LLMs), exemplified by ChatGPT (OpenAI), to demonstrate unprecedented capabilities in natural language processing, logical reasoning, and content generation [1]. In laboratory medicine, LLMs have demonstrated potential utility in multiple contexts, including laboratory test ordering [2], report interpretation [3,4], intelligent question answering [5], qualification examinations [6], laboratory management [7], and clinical decision support [8].

However, existing evaluations of LLMs predominantly rely on public or simulated datasets, with limited rigorous assessments in authentic clinical environments [9]. Critical challenges—including model hallucinations, opaque decision-making processes, and inadequate evaluation frameworks—severely constrain the safe deployment and widespread adoption of LLMs in clinical practice [10]. In response, guidelines such as TRIPOD-LLM and the Chatbot Assessment Reporting Tool (CHART) have been introduced, emphasizing the necessity of conducting systematic assessments of clinical artificial intelligence (AI) tools that are multidimensional, interpretable, and transparent [11,12].

The interpretation of complete blood count (CBC) results represents a key application for hematologic disease screening. As a primary diagnostic tool, CBC analysis provides essential clues for disease identification and differential diagnosis. When combined with clinical data from electronic health records (EHRs), CBC information from laboratory analyzers constitutes the foundation for diagnosing hematologic disorders. This process relies heavily on physician experience, leading to subjectivity, high workloads, and a lack of standardization. Traditional machine learning models have shown potential—such as CatBoost models for thalassemia identification [13], extreme gradient boosting (XGBoost)–based classifiers for acute leukemia subtypes [14], and support vector machine or artificial neural network models for acute leukemia diagnosis [15,16]. However, these models often suffer from limitations including limited interpretability, weak cross-disease generalization, an inability to integrate unstructured clinical narratives, and a lack of interactive explanations [17,18].

In contrast, LLMs leverage robust contextual understanding and chain-of-thought reasoning to identify abnormal values, analyze parameter correlations, simulate clinical reasoning, and generate diagnostic recommendations [19,20]. Notably, the DeepSeek (High-Flyer) model has been deployed in nearly 1000 Chinese hospitals [21], and its report interpretation capabilities have been piloted for laboratory result verification, clinical consultation, and patient communication.

Despite these advances, clinical laboratories currently lack evidence-based criteria for selecting LLMs, comparative evaluations using real-world clinical data, and systematic analyses of critical safety issues such as model hallucinations. This study addresses these limitations by comprehensively evaluating 3 advanced LLMs—DeepSeek R1, Grok 4 (SpaceXAI), and GPT-5—using real-world clinical CBC data across 6 quality and 5 task dimensions, with particular attention to reasoning capabilities and clinical safety. Our findings aim to provide empirical evidence and practical guidance for developing more reliable and safer AI-assisted interpretation systems for hematologic disease reports.

Methods

Study Design

We used a comprehensive framework to systematically evaluate the performance of 3 LLMs in interpreting CBC reports for hematologic disorders. The workflow consists of four key phases: (1) selection of target CBC reports for hematologic disorders; (2) submission of structured prompts to 3 LLMs; (3) evaluation across 6 quality dimensions—comprehensiveness, accuracy, clarity, relevance, practicality, and safety; and (4) evaluation across 5 task dimensions, reflecting clinical capabilities in analyzer alert processing, abnormal item identification, correlation analysis of abnormal items, preliminary diagnosis, and clinical management.

Ethical Considerations

This study analyzed a set of EHRs obtained from the laboratory department of The First Affiliated Hospital, Zhejiang University School of Medicine (FAHZU) between June 1, 2025, and July 7, 2025. The study protocol was approved by the institutional review board of FAHZU (IIT2025B-0629). Prior to data extraction, all records were fully deidentified by removing all direct identifiers (eg, name, date of birth, medical record number, contact information, and clinician details) and quasi-identifiers (eg, specific dates, locations, and institutional identifiers). Only data relevant to the study objectives were retained, including patient demographics, chief complaint, symptoms, physical examination findings, initial clinical suspicion, CBC reports, and alert messages generated by hematology analyzers. As the research involved a retrospective analysis of fully deidentified EHR data, the requirement for informed consent was waived by the institutional review board. No participants were contacted, and no compensation was provided.

Selection of Clinical Case Reports

This study retrospectively screened patients with hematologic conditions presenting for their initial visit to the 4 campuses of FAHZU (Yuhang, Qingchun, Zhijiang, and Chengzhan) between June 1, 2025, and July 7, 2025, yielding an initial cohort of 449 cases. From this cohort, we selected cases based on characteristic abnormal patterns in CBC reports that could provide diagnostic clues. Cases with normal CBC results or with nonspecific or subtle presentations (eg, those seen in early lymphoma, multiple myeloma, or coagulation disorders) were excluded. Ultimately, 49% (220/449) of the cases were included. The final diagnoses in this dataset were categorized into 4 main groups as follows:

Category 1 included myeloproliferative neoplasms, characterized by the persistent clonal proliferation of specific cell lineages, such as marked leukocytosis and a full myeloid spectrum in chronic myeloid leukemia; persistently elevated hemoglobin and hematocrit in polycythemia vera; significantly increased

XSL•FO
RenderX

platelet counts in essential thrombocythemia; and a characteristic leukoerythroblastic presentation in myelofibrosis.

Category 2 included acute leukemias and myelodysplastic syndromes, defined by blood cell count abnormalities and the presence of immature cells. This category includes acute myeloid leukemia and acute lymphoblastic leukemia—both marked by blasts—as well as high-risk myelodysplastic syndromes, which typically present as persistent, unexplained bicytopenia or pancytopenia, often accompanied by blasts.

Category 3 included cytopenic disorders, defined by reduced counts in one or more blood cell lineages. This category includes aplastic anemia with pancytopenia, immune thrombocytopenia with isolated thrombocytopenia, and various types of anemia characterized by specific red blood cell indices. These include microcytic hypochromic anemia in iron deficiency, macrocytic anemia in megaloblastic anemia, and microcytic hypochromic anemia in thalassemia.

Category 4 included lymphoproliferative neoplasms, characterized by clonal quantitative abnormalities in lymphocytes or plasma cells, including chronic lymphocytic leukemia with sustained absolute lymphocytosis, lymphoma with abnormal lymphocytes, and multiple myeloma, in which circulating plasma cells can be detected in some cases.

To achieve a balance between sample representativeness and evaluation workload, 100 cases were selected as the final evaluation cohort from the 220-case dataset through stratified random sampling, with disease category serving as the stratification variable.

Selection of LLMs

Three representative LLMs were selected for evaluation in this study: the open-source DeepSeek R1 (released May 28, 2025) and the closed-source models Grok 4 (released July 10, 2025) and GPT-5 (released August 8, 2025). The model strings used were DeepSeek-R1-0528, Grok 4, and GPT-5, respectively. All models were accessed via application programming interfaces using standardized prompts, with testing conducted on August 11, 2025. To ensure consistency and reproducibility, the generation parameters were fixed at a temperature of 0.3 and a top-p value of 1.0. None of the models underwent domain-specific fine-tuning; this study was designed to evaluate their out-of-the-box performance.

Querying LLMs

To evaluate the capabilities of LLMs in interpreting CBC reports for patients with hematologic conditions, a standardized prompt was designed. This prompt explicitly instructed the models to act as an “experienced laboratory medicine expert” and provide a professional interpretation based on the provided information. Input data encompassed patient demographics and key clinical texts from EHRs, specifically the chief complaint, physical examination findings, and initial clinical suspicion. All clinical text inputs were used in their original, unprocessed free-text format. These inputs were integrated with CBC reports containing all numerical parameters, reference intervals, and analyzer alerts. The final gold-standard diagnoses for all cases were established by clinicians according to the World Health Organization Classification (5th edition) criteria [22].

The models were required to address the following 5 tasks in sequence:

1. Analyzer alert processing. Interpret the meaning of all alert flags, assess their potential impact on result reliability, and provide specific recommendations for subsequent actions (eg, blood smear review).
2. Abnormal item identification. List all out-of-range parameters in a structured format, including values, direction of change, and brief clinical significance.
3. Correlation analysis of abnormal items. Analyze potential pathophysiological relationships among the abnormal indicators, incorporating patient demographics.
4. Preliminary diagnosis. Propose 1 to 3 of the most likely preliminary diagnoses or differential diagnoses
5. Clinical management. Provide specific, actionable suggestions regarding urgent interventions, additional tests, and follow-up.

Regarding output specifications, the models were required to respond in professional and concise Chinese, strictly follow the numbered sequence (1-5), and limit the total word count to 500 words. Any form of disclaimer was explicitly prohibited to prompt the model to make the most probable judgment. Each case query was processed in a new, isolated conversation session to prevent contextual interference between cases.

Additionally, to evaluate the independent hematologic reasoning capability of LLMs, we conducted an ablation study for task 4 by removing the initial clinical suspicion from the prompt while retaining all other patient information. Each case was thus processed under 2 conditions—with and without clinical suspicion—enabling comparison of model performance under full and blinded clinical contexts.

Evaluation of LLM Outputs

Through random selection from all eligible personnel at the 4 campuses, we recruited 2 junior evaluators (each with 5 years of experience) and 2 senior evaluators (each with >10 years of experience). Prior to formal evaluation, all evaluators completed a standardized training program to ensure consistent application of the evaluation criteria. The training encompassed clinical practice guidelines, authoritative medical literature, and clinical experience. It also included detailed explanations of scoring dimensions, illustrative case demonstrations, and a calibration exercise involving 20 reports. This exercise was repeated until consensus was reached and a Cohen kappa (κ) coefficient of 0.7 or above was achieved.

All model outputs were standardized before evaluator review. Only plain-text final outputs were presented, and visible reasoning traces or reasoning markers were removed when present. This removal was applied only to DeepSeek R1 outputs, as the GPT-5 and Grok 4 application programming interfaces responses did not contain visible reasoning traces. The standardized outputs were then anonymized, stripped of model identifiers, and presented in randomized order to minimize evaluator recognition based on formatting or stylistic cues.

XSL•FO
RenderX

We used 2 distinct evaluation checklists. For the quality dimensions, a scoring checklist based on a 5-point Likert scale was used (Multimedia Appendix 1). For the task dimensions, a 5-point deduction rubric was tailored to each of the 5 tasks (Multimedia Appendix 2). The task-specific deduction rubric served as a guide for assigning Likert scores across the 6 quality dimensions for each task output. In cases of ambiguity, evaluators consulted relevant guidelines or literature for clarification. This evaluation generated a total of 36,000 independent quality dimension ratings, calculated as follows: 100 cases×3 models×4 evaluators×5 tasks×6 dimensions. For both the quality and task dimensions, we applied a pragmatic consensus-scoring rule, where the final consensus score was the average of the 4 evaluator ratings. In addition, the raw individual ratings were retained to facilitate distributional visualizations of rating patterns and to analyze low-score assignments by task, model, and evaluator seniority.

We assessed concordance between the LLMs' preliminary diagnoses in task 4 and the initial clinical suspicion (under full-context conditions), as well as the final gold-standard diagnosis (under full-context and ablation conditions). For each model, paired 2×2 contingency tables were constructed based on whether the top-1 suggestion was concordant or discordant with the final gold-standard diagnosis under the 2 conditions. In addition, we classified errors identified during evaluation as either "hallucinations" (factually fabricated information) or "reasoning errors" (inferences lacking adequate clinical justification). To further emphasize clinical safety, evaluators were advised to assign a score no higher than 3 to any response containing either type of error, and all such cases were systematically recorded and reviewed in a post hoc analysis.

Statistical Analysis

All statistical analyses were conducted using R software (version 4.3.1; R Foundation for Statistical Computing). The reliability of the averaged ratings within the junior and senior evaluator groups was estimated using the 2-way random-effects intraclass correlation coefficient (ICC) for the mean of k=2 raters (ICC [2,2]). This coefficient reflects the reliability of the mean rating derived from 2 randomly selected raters. ICC values were reported with 95% CIs and interpreted according to Cicchetti criteria [23]: poor (<0.40), fair (0.40 - 0.59), good (0.60 - 0.74), and excellent (≥0.75). Additionally, we calculated objective concordance metrics for task 4, including top-1 concordance with the initial clinical suspicion and top-1 concordance with the final gold-standard diagnosis under full-context and ablation conditions. To evaluate whether removal of initial clinical suspicion significantly changed diagnostic accuracy within each individual model, we compared the paired binary outcomes (concordant vs discordant with the gold-standard diagnosis) between the 2 prompt conditions using the McNemar test.

For descriptive reporting of model performance across the 6 quality dimensions, the final 4-rater consensus scores were summarized as medians and IQRs. For task-level performance, each task output was comprehensively evaluated across the 6 quality dimensions, and a task-level score was calculated by averaging the 6 dimension scores assigned by the 4 evaluators for comparative analysis. Differences among the 3 LLMs in these scores were analyzed with the Friedman test; where significant, pairwise post hoc comparisons were conducted using the Wilcoxon signed-rank test with Holm-Bonferroni correction for multiple comparisons. All tests were 2-sided, with statistical significance set at $P<.05$.

Results

General Characteristics

As illustrated in the study design workflow (Figure 1), after screening, exclusion, and stratified sampling across four categories of hematologic diseases, 100 patients were included in the final evaluation cohort (mean age 58.5, SD 16.9 years; range 23 - 88 years; n=47, 47% male patients). The cohort composition is detailed in Table 1. The cohort encompassed diseases characterized by quantitative abnormalities in trilineage blood cell counts (category 1 and 3), as well as diseases marked by diagnostically significant pathological cells, including blasts, abnormal promyelocytes, abnormal lymphocytes, and plasma cells (category 2 and 4).

XSL•FO
RenderX

Figure 1. Workflow for the study design. CBC: complete blood count; LLM: large language model.

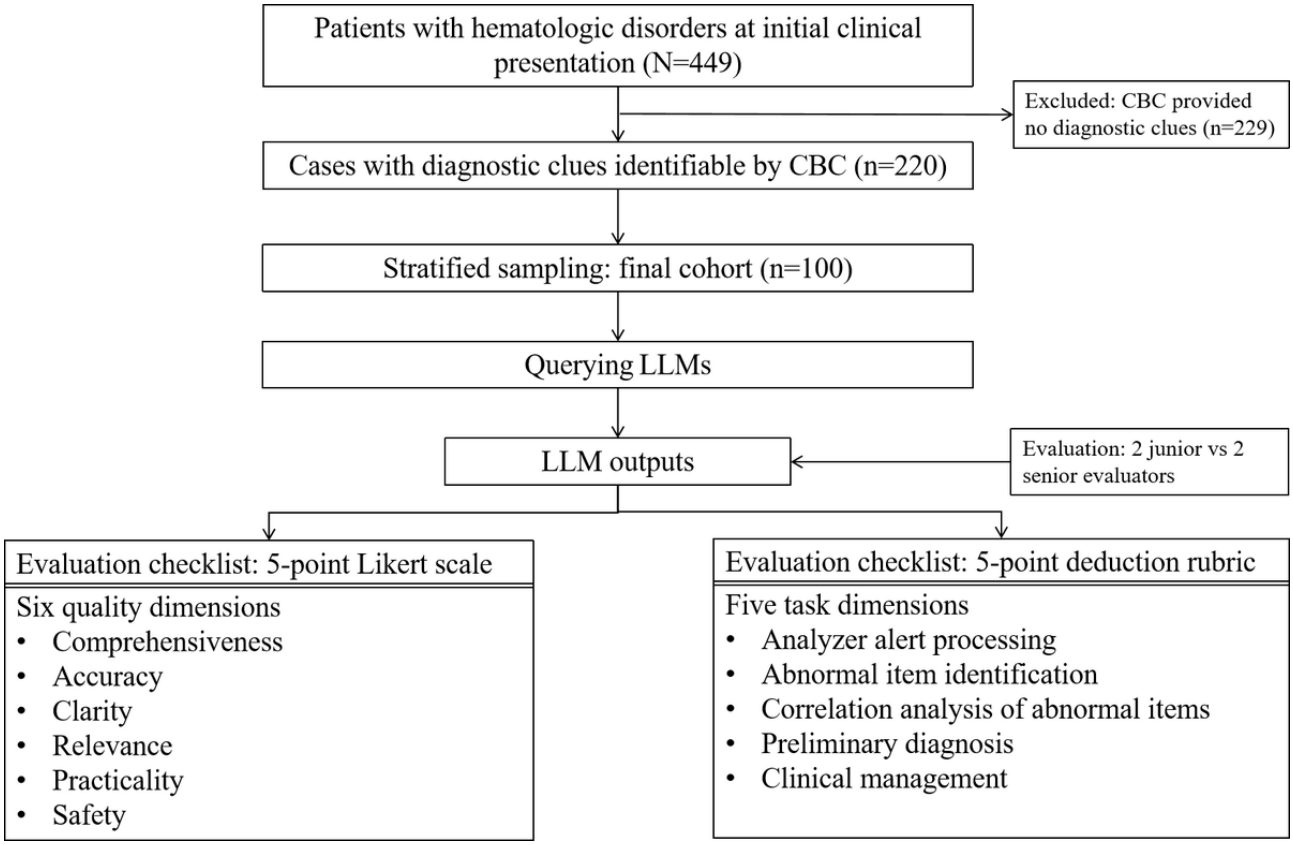

XSL•FO
RenderX

Table . Overview of clinical cases (n=100).

Characteristics	Values
Sex, n (%)	
Male	47 (47)
Female	53 (53)
Age (years), mean (SD; range)	58.5 (16.9; 23-88)
Disease category, n (%)	
Category 1: myeloproliferative neoplasms	22 (22)
Chronic myeloid leukemia	2 (2)
Polycythemia vera	5 (5)
Essential thrombocythemia	13 (13)
Myelofibrosis	2 (2)
Category 2: acute leukemias and myelodysplastic syndromes	25 (25)
Acute myeloid leukemia[a]	9 (9)
Acute lymphoblastic leukemia	6 (6)
Myelodysplastic syndromes	3 (3)
Acute leukemia of ambiguous lineage	7 (7)
Category 3: cytopenic disorders	34 (34)
Aplastic anemia	5 (5)
Immune thrombocytopenia	23 (23)
Iron deficiency anemia	4 (4)
Megaloblastic anemia	1 (1)
Thalassemia	1 (1)
Category 4: lymphoproliferative neoplasms	19 (19)
Chronic lymphocytic leukemia	12 (12)
Lymphoma with circulating abnormal lymphocytes	5 (5)
Multiple myeloma	2 (2)

[a]The 9 cases of acute myeloid leukemia included 2 cases of acute promyelocytic leukemia.

Interrater Reliability

Interrater reliability varied across LLMs and evaluator seniority (Figure 2). Across all 12 evaluations (2 seniority groups×6 dimensions), DeepSeek R1 demonstrated overall excellent reliability, with 9 evaluations achieving excellent reliability (ICC ≥0.75). Grok 4 showed moderate reliability, with 10 evaluations demonstrating good reliability (ICC 0.60 - 0.74) and 2 achieving excellent reliability in the accuracy (ICC 0.782, 95% CI 0.740 - 0.817) and safety (ICC 0.825, 95% CI 0.792 - 0.853) dimensions among senior evaluators. GPT-5 exhibited relatively greater variability, with 7 evaluations showing good reliability; notably, the lowest reliability was observed in the clarity dimension among junior evaluators (ICC 0.602, 95% CI 0.525 - 0.666). The remaining 5 evaluations achieved excellent reliability, primarily in the relevance and safety dimensions.

XSL•FO
RenderX

Figure 2. Forest plot of interrater reliability (intraclass correlation coefficient with 95% CI) for 3 large language models across 6 quality dimensions, stratified by evaluator seniority. Background colors indicate reliability levels: poor (<0.40, red), fair (0.40 - 0.59, yellow), good (0.60 - 0.74, light yellow), and excellent (≥0.75, light green).

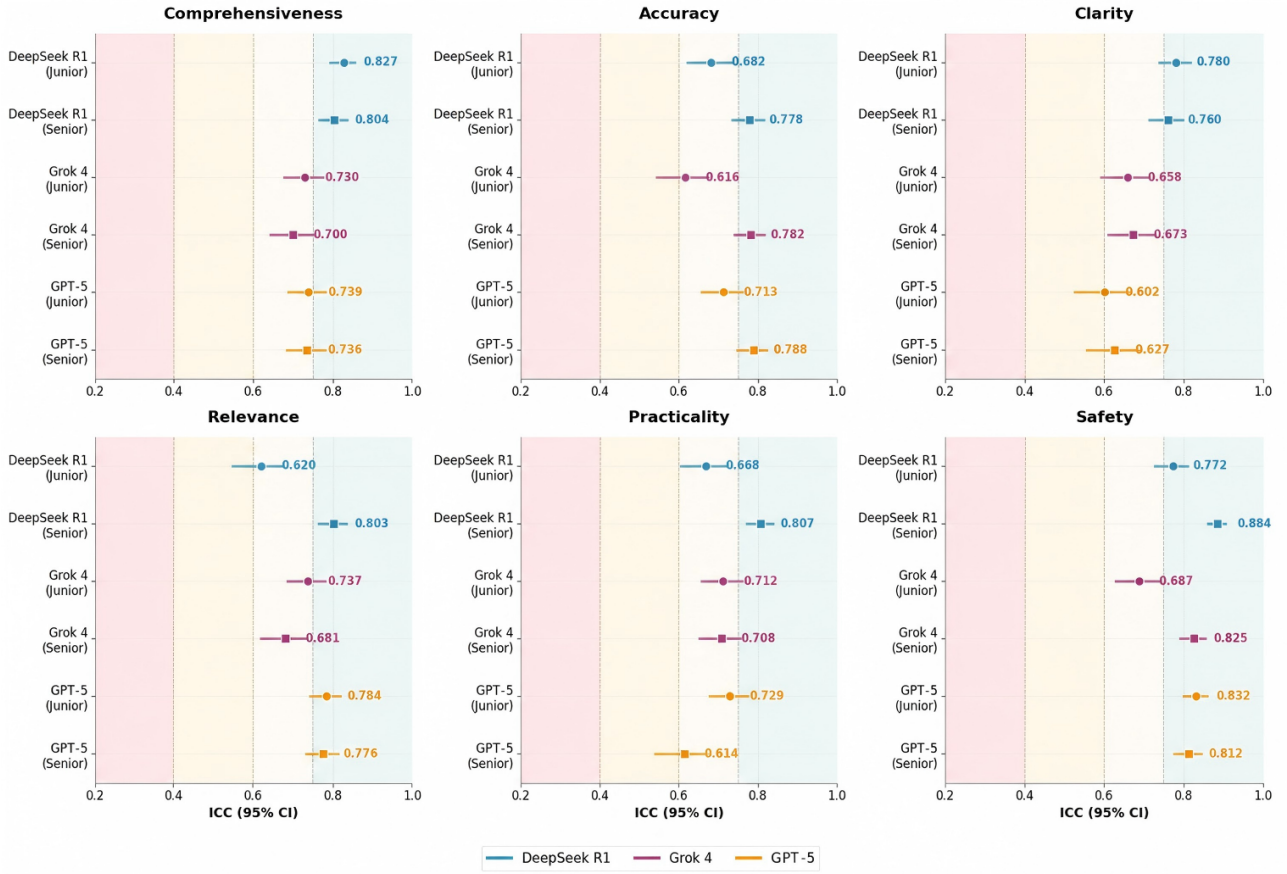

Quality Dimension Evaluation

We compared the LLMs' performance across 6 quality dimensions using box plots (Figure 3). DeepSeek R1 significantly outperformed both GPT-5 and Grok 4 in 5 dimensions: comprehensiveness, accuracy, clarity, relevance, and practicality (all P<.001). In the safety dimension, DeepSeek R1 achieved a median consensus score of 4.0 (IQR 4.0 - 4.5), which was lower than that of GPT-5 (median consensus score 4.25, IQR 4.0 - 4.5), although this difference did not reach statistical significance (P=.94). In comparative analyses, GPT-5 and Grok 4 demonstrated comparable accuracy and relevance, with no statistically significant differences (P=.13 and P=.30, respectively); however, GPT-5 outperformed Grok 4 in comprehensiveness, practicality, and safety (all P<.001). Conversely, Grok 4 exhibited significantly greater clarity than GPT-5 (P<.001).

XSL•FO
RenderX

Figure 3. Performance comparison of GPT-5, Grok 4, and DeepSeek R1 across 6 quality dimensions. Performance is visualized using box plots, where the bounds indicate the first and third quartiles (Q1 and Q3), the internal line represents the median consensus score, and the whiskers extend to 1.5 times the IQR. Individual plotted points represent 4-rater consensus scores.

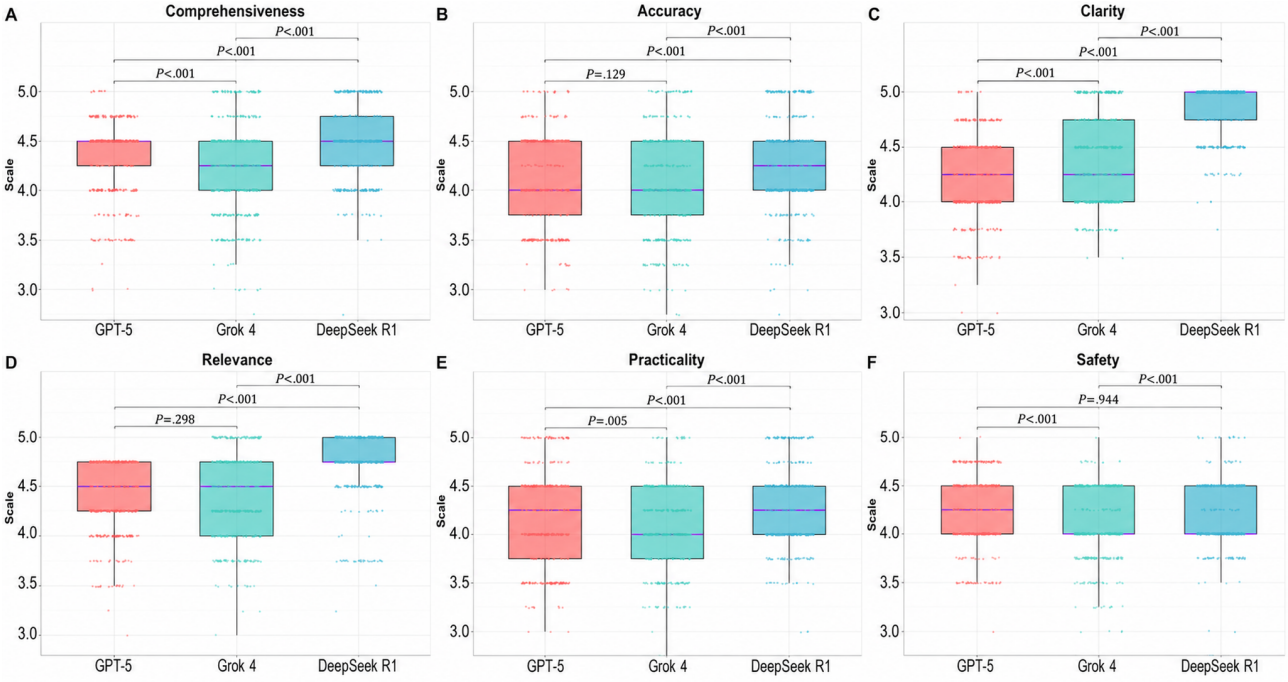

Task Dimension Evaluation

We compared the performance of the 3 LLMs across tasks 1 through 5 (Multimedia Appendix 3). DeepSeek R1 achieved the highest or tied-for-highest consensus scores in all tasks, with the only nonsignificant difference from GPT-5 occurring in the clinical management task. In the direct comparison between GPT-5 and Grok 4, the 2 models showed no significant difference in the analyzer alert processing task; furthermore, GPT-5 underperformed Grok 4 in the abnormal item identification task but outperformed it in the remaining 4 tasks.

For task 4 (preliminary diagnosis), we compared top-1 model outputs against both the initial clinical suspicion and the final gold-standard diagnosis under full-context and ablation conditions (Figure 4). Under full-context conditions, Grok 4 showed the highest concordance with the initial clinical suspicion (96/100, 96%), whereas GPT-5 achieved the highest concordance with the gold-standard diagnosis (93/100, 93%), followed by DeepSeek R1 (92/100, 92%) and Grok 4 (89/100, 89%). After removal of initial clinical suspicion, concordance with the gold-standard diagnosis declined to 79% (79/100) for GPT-5, 77% (77/100) for DeepSeek R1, and 72% (72/100) for Grok 4. The McNemar test showed that these within-model declines were statistically significant for all 3 models (all $P<.001$), indicating that initial clinical suspicion materially improved diagnostic accuracy.

XSL•FO
RenderX

Figure 4. Ablation analysis of preliminary diagnosis performance in 3 large language models. Top-1 concordance rates are shown for agreement with the initial clinical suspicion, agreement with the gold-standard diagnosis under full-context prompting, and agreement with the gold-standard diagnosis under the ablation condition. Paired 2×2 contingency tables and the McNemar test were used to evaluate within-model changes after removal of initial clinical suspicion. The numbers shown in the 2×2 contingency tables represent absolute case counts out of the 100 included cases.

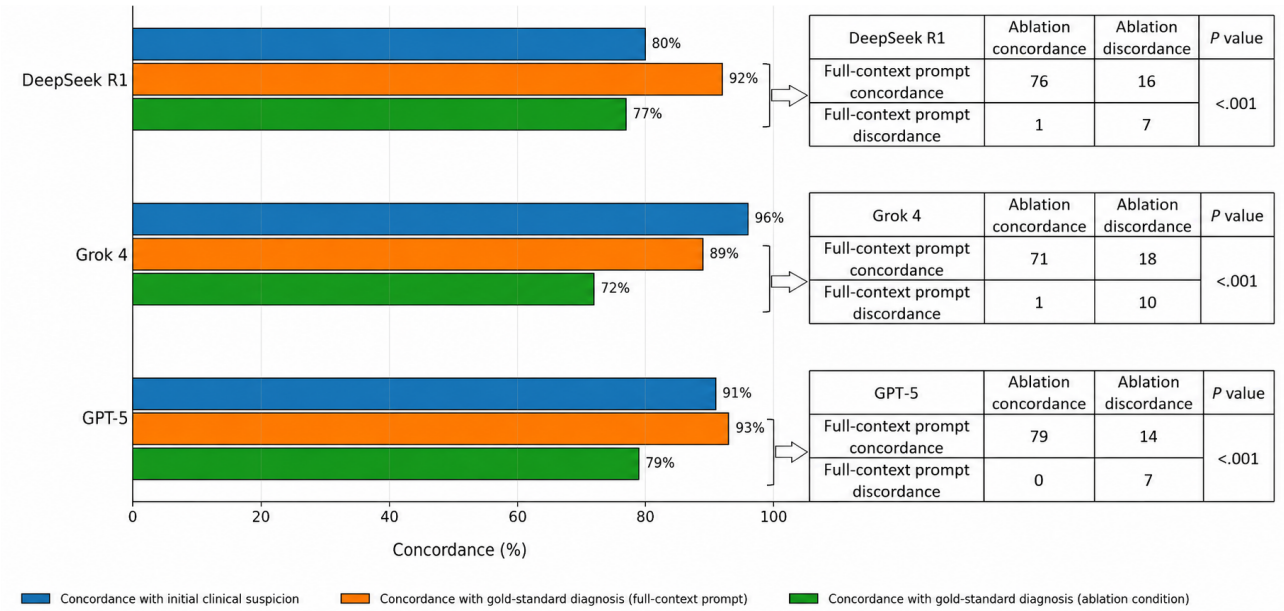

DeepSeek R1	Ablation concordance	Ablation discordance	*P* value
Full-context prompt concordance	76	16	<.001
Full-context prompt discordance	1	7	

Grok 4	Ablation concordance	Ablation discordance	*P* value
Full-context prompt concordance	71	18	<.001
Full-context prompt discordance	1	10	

GPT-5	Ablation concordance	Ablation discordance	*P* value
Full-context prompt concordance	79	14	<.001
Full-context prompt discordance	0	7	

Error Distributions and Analysis

We visualized raw individual evaluator rating distributions using heatmaps (Figure 5), with 1200 ratings generated for each task-model-evaluator seniority group combination (100 cases×2 evaluators within the seniority group×6 quality dimensions). Acknowledging that central tendency metrics (eg, median consensus scores) can obscure low-score assignments in clinically important tasks, we specifically analyzed the tails of the distribution. Although the rating scale theoretically ranged from 1 to 5, no evaluator assigned a score of 1 in this dataset; therefore, only observed score categories (2-5) are displayed. In clinical applications, the absolute incidence of serious errors is far more consequential than average performance; therefore, we quantified responses receiving low scores (<3) and qualitatively assessed the clinical risk. These low scores were mainly concentrated in 2 tasks: preliminary diagnosis and clinical management. In the preliminary diagnosis task, Grok 4 showed a relatively high proportion of low scores, at 1.1% (13/1200) for both junior and senior evaluators. In the clinical management task, the proportion of low scores for Grok 4 was even higher, at 2.5% (30/1200) for junior evaluators and 3.8% (46/1200) for senior evaluators.

XSL•FO
RenderX

Figure 5. Heatmap analysis of ratings across 5 task dimensions by evaluator seniority. Score distributions (1-5) for 3 large language models are shown, stratified by junior and senior evaluators, with ratings across the 6 quality dimensions. The color gradient and numeric values in each segment represent the proportion of assignments for each score. Scores of 1 are not displayed because no evaluator assigned a rating of 1 in the study dataset.

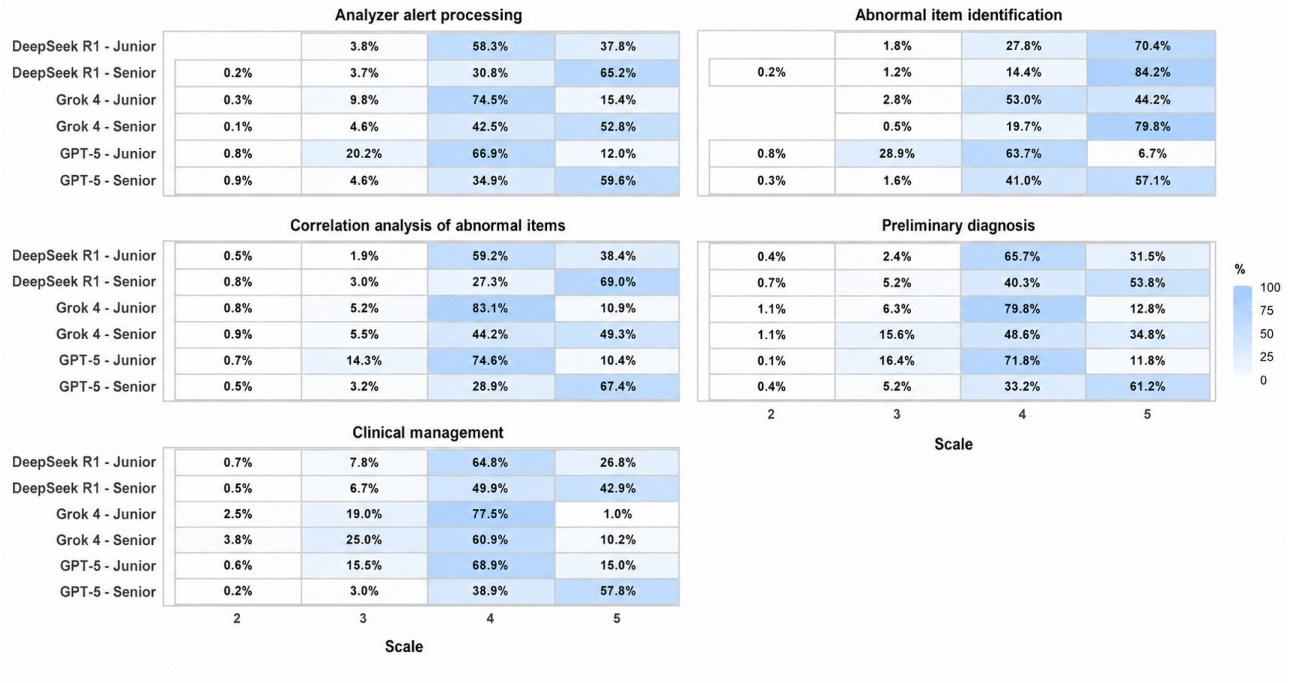

To elucidate the specific errors underlying these scores, we conducted a narrative review of the error distribution across the 5 task dimensions (Multimedia Appendix 4). In the analyzer alert processing task, hallucination errors were most prominent. GPT-5 exhibited 12 such errors, including recommending the manual correction of white blood cell counts while ignoring modern analyzers' automatic correction features, and misinterpreting plasma cell percentages from manual differential counts as instrument results—all of which could lead to unnecessary manual review and delayed reporting. DeepSeek R1 showed minor misunderstandings regarding platelet count interference factors, whereas Grok 4 displayed no hallucinations in this task.

For the abnormal item identification task, only DeepSeek R1 exhibited a single hallucination error (misinterpretation of reference interval thresholds), potentially leading to false-positive or false-negative clinical judgments. During the correlation analysis task, all models exhibited reasoning errors, characterized by unwarranted inference of disease status from a single laboratory result without integrating prior clinical information. Such errors could precipitate inappropriate clinical escalation, patient anxiety, and unnecessary overtesting. These inferential errors propagated further into the preliminary diagnosis task, where models used definitive terminology lacking guideline support to assert disease progression, potentially misguiding clinicians toward inappropriate treatment decisions. In the clinical management task, Grok 4 generated 9 reasoning errors by providing generic recommendations not tailored to specific CBC reports, which could compromise patient-specific health care and delay appropriate treatment. By comparison, DeepSeek R1 showed only 1 such case, and GPT-5 displayed none.

Discussion

Principal Findings

This multidimensional comparative evaluation of 3 leading LLMs in interpreting CBC reports for hematologic diseases demonstrates substantial potential for integration into laboratory medicine, offering critical insights for clinical implementation while revealing performance heterogeneity among models in handling complex hematologic inference.

Significant performance heterogeneity emerged across models in a task-dependent manner. DeepSeek R1 demonstrated superior or equivalent performance across all 6 quality dimensions and 5 clinical tasks. Its reasoning architecture—optimized through large-scale reinforcement learning and pretrained on high-quality Chinese medical literature—enabled the generation of localized and logically rigorous interpretations that garnered high evaluator recognition [24,25]. More importantly, the model exhibited substantial independent diagnostic capabilities: despite achieving only 80% (80/100) concordance with preliminary clinical suspicion, it attained 92% (92/100) concordance with gold-standard diagnoses, indicating its capacity for independent reasoning that challenges initial suspicions rather than merely reinforcing them [26].

In contrast, Grok 4 exhibited concerning error patterns in clinically critical tasks and systematic confirmation bias. While demonstrating 96% (96/100) concordance with preliminary clinical suspicion, its concordance with final gold-standard diagnoses was merely 89% (89/100), suggesting a tendency to extract and reinforce initial diagnostic cues from prompts rather than conducting independent inference based on laboratory data. Its concordance with the gold-standard diagnosis also declined markedly after ablation, from 89% (89/100) to 72% (72/100)

XSL•FO
RenderX

(McNemar $P<.001$). Furthermore, it displayed systematic overdiagnosis tendencies, generating generic management recommendations lacking specific data support during clinical management tasks, with 3.8% (46/1200) of outputs receiving low scores in the senior group.

GPT-5 demonstrated overall robust performance characteristics, trailing DeepSeek R1 across most quality dimensions while retaining the highest blinded concordance after ablation at 79% (79/100). It achieved the highest concordance with gold-standard diagnoses at 93% (93/100) and with initial clinical suspicion at 91% (91/100), indicating favorable knowledge generalization and diagnostic stability. However, the model showed a unique pattern of technical hallucinations in task dimensions, with 12 factual errors in the analyzer alert processing task, indicating that even generational upgrades of general-purpose LLMs cannot eliminate knowledge gaps in specialized technical domains.

The ablation study findings underscore the importance of integrating clinical context for accurate interpretation of laboratory results, which is highly consistent with the perspective advocated by Plebani [27] on laboratory result interpretation: interpreting laboratory data in isolation from the clinical context is inherently limited and potentially misleading. Plebani [27] explicitly noted that expecting AI to achieve more accurate diagnoses than well-trained clinicians based solely on reference intervals and test parameters is "absurd," emphasizing that true laboratory medicine value realization depends on the integration of pretest probability and comprehensive clinical information. In our study, all models showed significant declines after removal of the initial clinical suspicion, further supporting the notion that, in the absence of clinical information, models tend to rely excessively on statistical associations rather than pathophysiological reasoning. Although GPT-5 demonstrated relative robustness under blinded conditions, this cannot compensate for the systematic loss of diagnostic accuracy when removed from clinical context. Therefore, current applications of LLMs in hematology report interpretation should adhere to the principle advocated by Plebani [27] that laboratory results must be interpreted within the context of pretest probability and comprehensive clinical information, restricting AI assistance to scenarios with a complete clinical background rather than permitting its use as an isolated interpreter of test results.

Comparison With Prior Work

Previous studies have predominantly relied on publicly available question banks or simulated cases with limited validation in real clinical scenarios, functioning essentially as "isolated test result interpretation tools" rather than "integrated clinical decision support systems." Such research only assessed the models' ability to interpret individual laboratory parameters without evaluating LLMs' performance in complete diagnostic contexts [3]. For instance, Kumari et al [28] evaluated 3 LLMs on 50 complex, multitopic hematology cases to assess their case-solving performance; Han et al [29] assessed ChatGPT's error-correction capability for nucleic acid testing reports by artificially introducing mistakes; and Cadamuro et al [3], representing the European Federation of Clinical Chemistry and Laboratory Medicine Working Group, tested ChatGPT's comprehension using 10 simulated reports of common parameters. Additional studies compared LLMs with physician interpretations of laboratory questions from online health forums [5,19], yet remained confined to general comprehension rather than professional diagnostic reasoning. Unlike prior work, this study integrates comprehensive EHR context (chief concerns and physical examination findings) with raw analyzer data and alert flags, constructing an evaluation framework that more authentically replicates clinical laboratory workflows. This paradigm shift from "test interpretation tool" to "clinical decision support" provides a feasible pathway for transitioning laboratory report interpretation from bench to bedside deployment.

Despite the potential of LLMs, hallucinations and reasoning errors remain fundamental barriers to clinical implementation [9]. Based on our error classification framework, differentiated strategies are required for precise risk mitigation. For hallucination errors exhibited by GPT-5, retrieval-augmented generation can anchor model outputs to Clinical and Laboratory Standards Institute (CLSI) guidelines and instrument operation manuals, eliminating fabricated instructions such as "manual correction of white blood cell counts" [20,30]. For reasoning errors demonstrated by Grok 4, senior hematologists should be specifically tasked with reviewing diagnostic recommendations, leveraging their clinical reasoning to calibrate AI overinference. DeepSeek R1's superior performance as an open-source model suggests that domain-specific fine-tuning based on local practice guidelines can significantly enhance evidence-based reasoning consistency, while future integration of multimodal inputs—such as peripheral blood smear images and CBC scattergram raw data—may further augment diagnostic reliability [9]. Based on these differentiated strategies, we recommend implementing a tiered human-AI collaboration framework: structured tasks, such as abnormal item identification, may be automated using LLMs to improve efficiency, whereas high-risk steps—including analyzer alert interpretation, preliminary diagnosis, and clinical management tasks—must undergo mandatory review by hematology experts to ensure clinical safety.

Limitations

This study has several limitations. First, although cases from 4 hospital campuses were included, the retrospective design and stratified sampling based on typical CBC abnormality patterns may introduce selection bias, particularly by excluding diseases with subtle CBC presentations—such as early-stage lymphomas and multiple myeloma without cytopenias—resulting in an overrepresentation of classic cases. Second, the evaluation was conducted in Chinese, whereas the base training data for GPT-5 and Grok 4 are primarily in English, which may have introduced a language-related confounding factor. In addition, the prompt imposed a strict 500-word output limit, which represented another artificial constraint that may have influenced model behavior. This restriction may have encouraged models to compress content or selectively omit details to comply with the length requirement, thereby affecting the overall quality of the responses. The effects of language and length constraints may have jointly contributed to some of the observed patterns in this study, such as the relatively lower clarity of GPT-5. For Grok 4, the relatively generic and less tailored clinical management recommendations may be related to the model attempting to

XSL•FO
RenderX

shorten its response to comply with the prompt instructions. For DeepSeek R1, the 500-word limit was applied only to the final output rather than to reasoning tokens; therefore, the models may not have been evaluated under fully equal length constraints, potentially confounding task performance. Future studies should consider language-matched evaluation settings and more flexible output constraints to better reflect real-world model capabilities. Third, the structured fact-verification checklist standardizes the grading format; however, in the absence of a pre-established, case-specific answer key, distinguishing between verifiable minor errors and significant interpretive deviations requires subjective clinical judgment that may vary among evaluators, leaving residual subjectivity in the rubric definitions. Future studies should develop comprehensive, case-specific scoring rubrics with predefined exemplar responses to further enhance objectivity. Fourth, only standardized final outputs were retained for evaluation, with visible reasoning traces excluded. Because this preprocessing was applied only to DeepSeek R1, which uses a reinforcement learning–optimized reasoning mechanism to generate high-quality outputs, evaluators were unable to assess its intermediate deductive process. This may have limited the evaluation of reasoning transparency and error formation mechanisms, particularly for DeepSeek R1. Fifth, the number of hematologists participating in the evaluation was limited and all were from the same health care system, which may introduce institution-specific bias and affect the generalizability of the results. Finally, given the rapid evolution of LLM technology, our results represent only a performance snapshot at a specific time point; subsequent model updates may alter performance characteristics, limiting the strict reproducibility of findings.

Conclusions

This study reveals significant performance heterogeneity among LLMs in real-world hematologic CBC report interpretation and distinct patterns of error distribution, providing preliminary evidence for laboratory AI tool selection. Clinical deployment should implement a tiered management strategy based on error classification: restricting LLMs to low-risk structured task assistance while mandating expert review for high-risk diagnostic reasoning tasks. As this represents a single-center, Chinese-language exploratory assessment, this findings are context-dependent and necessitate multicenter, cross-lingual prospective validation to further delineate the safety boundaries and generalizable standards for clinical integration of LLMs.

Acknowledgments

We thank Yunying Chen from the Department of Laboratory Medicine, Hangzhou Children's Hospital, for assistance with statistical analysis and figure preparation. KIMI (Moonshot AI) was used only for English language polishing during manuscript preparation and translation. No AI-generated content was included without author review and editing, and the authors take full responsibility for the final manuscript.

Funding

This work was supported by the National Key Technologies R&D Program of China (2022YFC3602302).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

XY, XQ, LF, and DY contributed to the conceptualization, investigation, and methodology of the project. QY and SZ, as selected members of the junior group, and XY and CR, as selected members of the senior group, evaluated the interpretations generated by the large language models. XY, XQ, and LF were responsible for data curation. XY was responsible for the formal data analysis under the supervision of XQ and DY. XY and DY drafted the original manuscript (writing—original draft), which was reviewed and edited by all coauthors (writing—review and editing).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Evaluation checklists for quality dimensions.
[XLSX File, 12 KB - jmir_v28i1e87802_app1.xlsx]

Multimedia Appendix 2
Evaluation checklists for task dimensions.
[XLSX File, 14 KB - jmir_v28i1e87802_app2.xlsx]

Multimedia Appendix 3
Performance comparison of GPT-5, Grok 4, and DeepSeek R1 across 5 task dimensions.

XSL•FO
RenderX

[DOCX File, 586 KB - jmir_v28i1e87802_app3.docx]

Multimedia Appendix 4
Narrative review of model errors across 5 task dimensions.
[DOCX File, 15 KB - jmir_v28i1e87802_app4.docx]

References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
2. Zayed AM, Frans G, Delvaux N. Evaluating large language models as clinical laboratory test recommenders in primary and emergency care: a crucial step in clinical decision making. Clin Chem Lab Med 2025 Oct 27;63(11):2186-2197. [doi: 10.1515/cclm-2025-0647] [Medline: 40802589]
3. Cadamuro J, Cabitza F, Debeljak Z, et al. Potentials and pitfalls of ChatGPT and natural-language artificial intelligence models for the understanding of laboratory medicine test results. An assessment by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group on Artificial Intelligence (WG-AI). Clin Chem Lab Med 2023 Jun 27;61(7):1158-1166. [doi: 10.1515/cclm-2023-0355] [Medline: 37083166]
4. Hu L, Xu X, Zhuang Y, et al. Pre-trained ChatGPT for report generation in automated microbial identification and antibiotic susceptibility testing systems. Sci Rep 2025;15:36283. [doi: 10.1038/s41598-025-22315-5]
5. Girton MR, Greene DN, Messerlian G, Keren DF, Yu M. ChatGPT vs medical professional: analyzing responses to laboratory medicine questions on social media. Clin Chem 2024 Sep 3;70(9):1122-1139. [doi: 10.1093/clinchem/hvae093] [Medline: 39013110]
6. Jung K, Kim HJ, Shin S, et al. Evaluation of the performance of advanced large language models in laboratory medicine using residency examinations. Ann Lab Med 2026 May 1;46(3):327-337. [doi: 10.3343/alm.2025.0200] [Medline: 41224529]
7. Abusoglu S, Serdar M, Unlu A, Abusoglu G. Comparison of three chatbots as an assistant for problem-solving in clinical laboratory. Clin Chem Lab Med 2024 Jun 25;62(7):1362-1366. [doi: 10.1515/cclm-2023-1058] [Medline: 38095605]
8. Li Q, Zhan L, Cai X. Assessing DeepSeek-R1 for clinical decision support in multidisciplinary laboratory medicine. J Multidiscip Healthc 2025;18:4979-4988. [doi: 10.2147/JMDH.S538253] [Medline: 40823482]
9. Yu E, Chu X, Zhang W, et al. Large language models in medicine: applications, challenges, and future directions. Int J Med Sci 2025;22(11):2792-2801. [doi: 10.7150/ijms.111780] [Medline: 40520893]
10. Tam TY, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med 2024 Sep 28;7(1):258. [doi: 10.1038/s41746-024-01258-7] [Medline: 39333376]
11. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med 2025 Jan;31(1):60-69. [doi: 10.1038/s41591-024-03425-5] [Medline: 39779929]
12. CHART Collaborative. Reporting guidelines for chatbot health advice studies: explanation and elaboration for the Chatbot Assessment Reporting Tool (CHART). BMJ 2025 Aug 1;390:e083305. [doi: 10.1136/bmj-2024-083305] [Medline: 40750271]
13. Lai JX, Tang JW, Gong SS, et al. Development and validation of an interpretable risk prediction model for the early classification of thalassemia. NPJ Digit Med 2025 Jun 10;8(1):346. [doi: 10.1038/s41746-025-01766-0] [Medline: 40494920]
14. Alcazer V, Le Meur G, Roccon M, et al. Evaluation of a machine-learning model based on laboratory parameters for the prediction of acute leukaemia subtypes: a multicentre model development and validation study in France. Lancet Digit Health 2024 May;6(5):e323-e333. [doi: 10.1016/S2589-7500(24)00044-X] [Medline: 38670741]
15. Gao HW, Wang YY, Li X, et al. Acute leukemia warning model combined CBC and CPD data based on machine learning. Int J Lab Hematol 2025 Dec;47(6):1044-1053. [doi: 10.1111/ijlh.14538] [Medline: 40765161]
16. Haider RZ, Ujjan IU, Khan NA, Urrechaga E, Shamsi TS. Beyond the in-practice CBC: the research CBC parameters-driven machine learning predictive modeling for early differentiation among leukemias. Diagnostics (Basel) 2022 Jan 7;12(1):138. [doi: 10.3390/diagnostics12010138] [Medline: 35054304]
17. Çubukçu HC, Topcu D, Yenice S. Machine learning-based clinical decision support using laboratory data. Clin Chem Lab Med 2023 Nov;62(5):793-823. [doi: 10.1515/cclm-2023-1037] [Medline: 38015744]
18. Miller HA, Valdes R. Rigorous validation of machine learning in laboratory medicine: guidance toward quality improvement. Crit Rev Clin Lab Sci 2025 Aug;62(5):327-346. [doi: 10.1080/10408363.2025.2488842] [Medline: 40247648]
19. Meyer A, Soleman A, Riese J, Streichert T. Comparison of ChatGPT, Gemini, and Le Chat with physician interpretations of medical laboratory questions from an online health forum. Clin Chem Lab Med 2024 Nov 26;62(12):2425-2434. [doi: 10.1515/cclm-2024-0246] [Medline: 38804035]
20. He Z, Bhasuran B, Jin Q, et al. Quality of answers of generative large language models versus peer users for interpreting laboratory test results for lay patients: evaluation study. J Med Internet Res 2024 Apr 17;26:e56655. [doi: 10.2196/56655] [Medline: 38630520]

XSL•FO
RenderX

21. Zeng D, Qin Y, Sheng B, Wong TY. DeepSeek's "low-cost" adoption across China's hospital systems: too fast, too soon? JAMA 2025 Jun 3;333(21):1866-1869. [doi: 10.1001/jama.2025.6571] [Medline: 40293869]
22. Khoury JD, Solary E, Abla O, et al. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: myeloid and histiocytic/dendritic neoplasms. Leukemia 2022 Jul;36(7):1703-1719. [doi: 10.1038/s41375-022-01613-1] [Medline: 35732831]
23. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 1994;6(4):284-290. [doi: 10.1037/1040-3590.6.4.284]
24. Tordjman M, Liu Z, Yuce M, et al. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. Nat Med 2025 Aug;31(8):2550-2555. [doi: 10.1038/s41591-025-03726-3] [Medline: 40267969]
25. Sandmann S, Hegselmann S, Fujarski M, et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. Nat Med 2025 Aug;31(8):2546-2549. [doi: 10.1038/s41591-025-03727-2] [Medline: 40267970]
26. Yang HS, Li J, Yi X, Wang F. Performance evaluation of large language models with chain-of-thought reasoning ability in clinical laboratory case interpretation. Clin Chem Lab Med 2025 Jul 28;63(8):e199-e201. [doi: 10.1515/cclm-2025-0055] [Medline: 40023838]
27. Plebani M. ChatGPT: Angel or demond? Critical thinking is still needed. Clin Chem Lab Med 2023 Jun 27;61(7):1131-1132. [doi: 10.1515/cclm-2023-0387]
28. Kumari A, Kumari A, Singh A, et al. Large language models in hematology case solving: a comparative study of ChatGPT-3.5, Google Bard, and Microsoft Bing. Cureus 2023 Aug;15(8):e43861. [doi: 10.7759/cureus.43861] [Medline: 37736448]
29. Han W, Wan C, Shan R, et al. Evaluation of error detection and treatment recommendations in nucleic acid test reports using ChatGPT models. Clin Chem Lab Med 2025 Aug 26;63(9):1698-1708. [doi: 10.1515/cclm-2025-0089] [Medline: 40249886]
30. Nanua S, Steward R, Neely B, Datto M, Youens K. Retrieval-augmented generation for interpreting clinical laboratory regulations using large language models. J Pathol Inform 2025 Nov;19:100520. [doi: 10.1016/j.jpi.2025.100520] [Medline: 41244595]

Abbreviations

AI: artificial intelligence
CBC: complete blood count
CLSI: Clinical and Laboratory Standards Institute
EHR: electronic health record
FAHZU: The First Affiliated Hospital, Zhejiang University School of Medicine
ICC: intraclass correlation coefficient
LLM: large language model

Edited by A Coristine; submitted 14.Nov.2025; peer-reviewed by A Abou-Bakr, R Yazici, R Cerqueira; revised version received 06.May.2026; accepted 06.May.2026; published 05.Jun.2026.

Please cite as:
Ye X, Qi X, Fan L, Yu Q, Zhou S, Ren C, Yang D
Performance Evaluation of GPT-5, Grok 4, and DeepSeek R1 in Interpreting Complete Blood Count Reports for Hematologic Diseases: Retrospective Comparative Study
J Med Internet Res 2026;28:e87802
URL: https://www.jmir.org/2026/1/e87802
doi:10.2196/87802

© Xianfei Ye, Xinglun Qi, Lina Fan, Qian Yu, Suming Zhou, Chunyun Ren, Dagan Yang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessing Eligibility for Anticancer Drug Health Insurance Reimbursement Using Large Language Models: Benchmark Development and Comparative Study

Junhyuk Seo[1*], RN, MS; Taerim Kim[1,2,3*], MD, PhD; Ju-Hyun Kim[4], MD, PhD

[1]Healthcare Research Institute, ETOILE Inc, Seoul, Republic of Korea

[2]Department of Digital Health, Samsung Advanced Institute of Health Sciences and Technology (SAIHST), Sungkyunkwan University, Seoul, Republic of Korea

[3]Department of Emergency Medicine, Sungkyunkwan University School of Medicine, Samsung Medical Center, Seoul, Republic of Korea

[4]Department of Obstetrics and Gynecology, University of Ulsan College of Medicine, Asan Medical Center, 88, Olympic-ro 43-gil, Songpa-gu, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Ju-Hyun Kim, MD, PhD
Department of Obstetrics and Gynecology, University of Ulsan College of Medicine, Asan Medical Center, 88, Olympic-ro 43-gil, Songpa-gu, Seoul, Republic of Korea

Abstract

Background: Administrative costs in the health care system are driven in part by complex insurance eligibility determinations. Large language models (LLMs) are increasingly used for health insurance–related queries, yet their reliability for structured logical reasoning over coverage criteria has not been systematically evaluated.

Objective: This study aimed to develop a benchmark for anticancer drug reimbursement eligibility determination and evaluate whether LLMs can reliably perform eligibility verification.

Methods: We constructed a benchmark based on South Korea's National Health Insurance reimbursement guidelines for 3 gynecologic cancers (cervical, uterine, and ovarian), using a tristate adjudication framework (eligible, ineligible, and undeterminable). Three gynecologic oncology experts and a utilization review nurse validated the benchmark. Six LLMs from 3 providers (Anthropic, Google, and OpenAI) were evaluated using the official guideline document as input. Each case was evaluated 3 times per model, with final predictions determined by majority vote, and performance was compared across the 3 outcome classes.

Results: The benchmark comprises 74 anticancer regimens with 222 cases. Overall verification accuracy ranged from 77.9% to 88.7% across the 6 models. Eligible and ineligible cases were classified with high recall (86.5% - 98.6%), but undeterminable cases showed a marked decline across all models (44.6% - 70.3%). Performance varied by cancer type, with uterine cancer showing the lowest undeterminable recall (16.7%), corresponding to the highest guideline complexity. Undeterminable cases were predominantly misclassified as eligible rather than ineligible. The tristate framework enabled logic-based error analysis of 235 incorrect predictions, revealing information gap-filling as the dominant failure pattern (n=196, 83.4%), followed by criterion misapplication (n=20, 8.5%) and false uncertainty (n=19, 8.1%). Subtype analysis indicated that information gap-filling errors were concentrated at hierarchical elements of the guideline. Sensitivity analyses showed that converting the guideline document to structured text degraded performance, while web search–enabled condition (0% - 3.2% tool invocation across models) and structure-guided prompting did not produce significant changes from baseline.

Conclusions: In this benchmark, LLMs classified clearly eligible and ineligible cases with relatively high recall but showed limited reliability on undeterminable cases. The dominant error pattern was information gap-filling, in which models inferred eligibility rather than withholding judgment. These findings indicate that LLMs, in their current form, should be deployed as supervised decision-support tools rather than as independent adjudicators in reimbursement review.

(J Med Internet Res 2026;28:e95877) doi:10.2196/95877

KEYWORDS

large language models; health insurance; reimbursement; eligibility verification; anticancer drugs; clinical decision support; benchmark; natural language processing; gynecologic cancer; National Health Insurance

XSL•FO
RenderX

Introduction

In the United States, billing- and insurance-related activities were estimated to cost approximately US $471 billion annually, representing roughly 18% of national health expenditure [1]. At the provider level, billing- and insurance-related costs account for 3% to 25% of professional revenue depending on the type of clinical encounter [2]. This administrative burden poses significant negative impacts for both patients and clinicians [3,4]. Although health systems and payer models vary worldwide, billing processes follow a common pathway—from eligibility verification and coding to claim submission and rework—across countries with diverse payer structures [5]. These burdens persist even in single-payer systems, where centralized rule-based reimbursement introduces a different form of administrative complexity.

In South Korea, the entire population is enrolled in the National Health Insurance (NHI) by law and relies on predefined coverage rules and eligibility criteria [6]. Under this system, the National Health Insurance Service functions as the single insurer, while the Health Insurance Review & Assessment Service (HIRA) conducts centralized claims review and quality assessment [7]. Reimbursement decisions are made through a retrospective claims review process, in which submitted claims are evaluated for compliance with detailed coverage rules [8,9]. Under this retrospective review structure, eligibility determination becomes a high-stakes post hoc decision, where errors can directly lead to claim rejection, payment adjustment, and downstream administrative burden.

As anticancer agents continue to be developed and covered by the NHI, national expenditure on these drugs surged by 168.2% from 2013 to 2022, reaching US $982 million in 2022 [10]. Alongside this rapid cost increase, eligibility criteria for anticancer drugs have become increasingly complicated, requiring simultaneous assessment of multiple clinical variables—such as tumor type, prior regimen sequences, biomarker thresholds, and performance scores—connected by nested AND/OR conditions [11]. This combinatorial structure makes manual review prone to omission and misinterpretation, particularly when multiple interdependent conditions must be evaluated simultaneously.

A recent report found that users of ChatGPT in the United States alone send 1.6 to 1.9 million health insurance–related messages each week to understand coverage and navigate claims and denials [12]. Despite this high level of use, large language models (LLMs) often produce hallucinations and other errors that can lead to critical mistakes, especially in medical domains [13,14]. When LLMs have been applied to improve the understandability of clinical guidelines, unintended omissions and changes in meaning were identified in up to 20% of revised subsections [15]. If even surface-level revisions can introduce errors, the risk may be amplified when LLMs must reason over the logical structure of guidelines to reach coverage determination.

Recent studies have shown that LLMs can support reimbursement-related tasks, such as detecting clinical conditions for appropriate coding [16], and automating *International Classification of Diseases* code assignment from clinical documentation [17]. However, these applications focus on information extraction from clinical notes. Determining reimbursement eligibility, by contrast, requires structured logical reasoning in which multiple clinical attributes must be jointly assessed against complex rule sets, where a single incorrect inference can invalidate the entire decision. Moreover, in clinical practice, available documentation is often incomplete [18,19], and a reliable system must also recognize when the evidence is insufficient to reach a determination rather than forcing a binary decision. Yet, no standardized benchmark exists to evaluate these capabilities under real-world insurance review constraints.

To address this gap, we developed a benchmark for anticancer drug reimbursement eligibility based on South Korea's national guidelines. Our benchmark formalizes the condition-level adjudication logic that clinicians and utilization review nurses routinely apply when evaluating incomplete clinical evidence against coverage rules. It covers 3 gynecologic cancers—cervical, uterine, and ovarian—and includes eligible, ineligible, and undeterminable cases to assess not only correctness but also whether models can recognize when available evidence is insufficient to support a determination. Using this benchmark, we evaluated LLM reliability for reimbursement eligibility adjudication under incomplete information.

Methods

Benchmark Development

The study design is summarized in Figure 1. This study was reported in accordance with the TRIPOD-LLM (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis-Large Language Model) [20]. A completed checklist is provided in Checklist 1.

XSL•FO
RenderX

Figure 1. Study design overview. The benchmark comprises 222 cases derived from national reimbursement guidelines for 3 gynecologic cancers (74 regimens × 3 outcome classes). Validation was performed by 3 gynecologic oncology experts and an independent utilization review nurse. Six models were evaluated using the finalized benchmark. LLM: large language model.

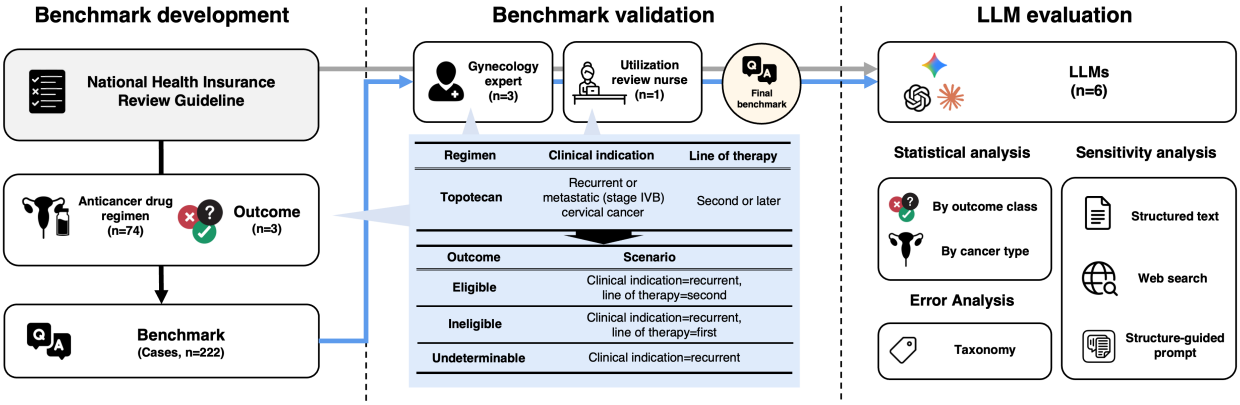

We developed the benchmark using the HIRA reimbursement review guidelines for anticancer drugs, version dated February 1, 2026. Three major gynecologic cancers—cervical cancer, uterine cancer, and ovarian cancer—were included. For each cancer type, reimbursement criteria were translated into structured clinical attributes representing heterogeneous patient scenarios. Candidate benchmark items were created by a researcher with expertise in medical informatics (JS).

The benchmark was constructed within a tristate adjudication framework, including eligible, ineligible, and undeterminable outcomes for each anticancer regimen. This outcome assignment is based on 3 different condition-level states: met, not met, and unevaluable. We use "not met" to denote a condition whose required value is present but violates the criterion, and "unevaluable" to denote a condition that cannot be assessed because the required information is absent (Figure 2). The unevaluable state was included to reflect real-world documentation contexts in which information required for applying a coverage criterion may be absent, incomplete, or inaccessible.

Figure 2. Tristate adjudication framework for benchmark outcome assignment. Each required condition for a regimen was classified as met, not met, or unevaluable. Eligible cases met all required conditions. Ineligible cases included at least 1 explicitly not-met condition. Undeterminable cases included no explicitly not-met condition but at least 1 unevaluable condition due to missing information.

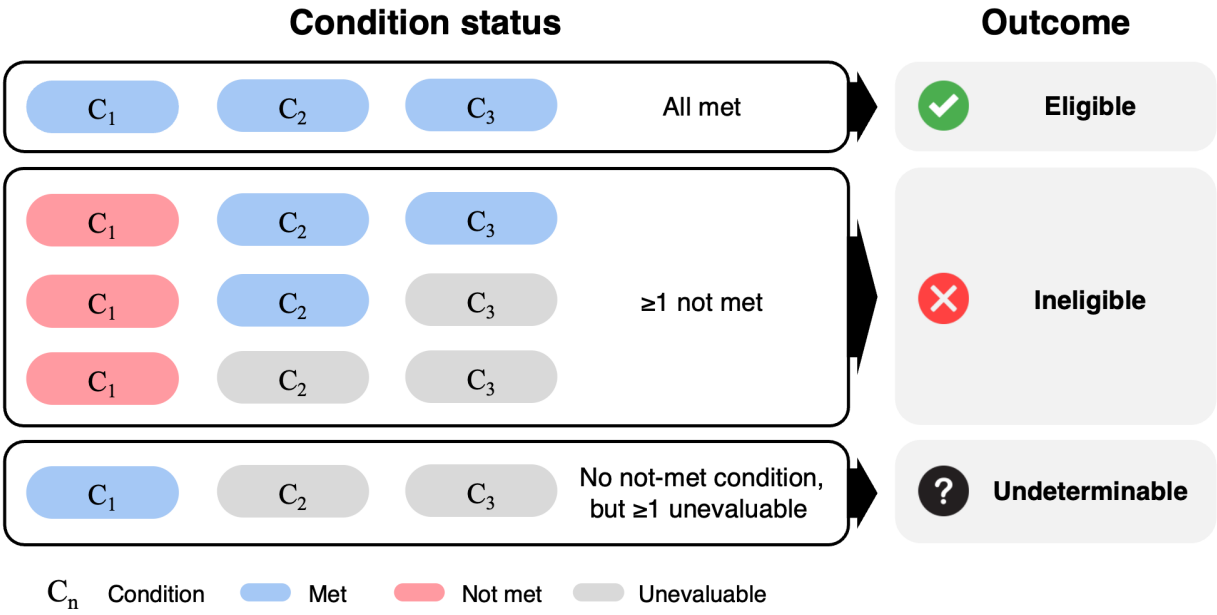

For instance, a case is classified as ineligible when 1 or more conditions are explicitly not met, regardless of whether other conditions are unevaluable. A case is classified as undeterminable when no condition is explicitly not met, yet 1 or more conditions remain unevaluable. To control task difficulty, ineligible and undeterminable cases were designed as near-miss scenarios containing only 1 to 2 conditions not met or unevaluable. The detailed definitions of the outcomes are shown in Table 1.

XSL•FO
RenderX

Table . Definitions of eligibility outcomes used in the benchmark, with illustrative examples from cervical cancer topotecan (single use).

Outcome	Definition	Example scenario	Explanation
Eligible	All required conditions are met	Clinical indication=recurrent; line of therapy=second	All guideline conditions are satisfied.
Ineligible	≥1 condition is explicitly not met, regardless of whether other conditions are unevaluable	Clinical indication=recurrent; Line of therapy=first	Line of therapy is first line. The condition is present but violates the criterion (“not met”).
Undeterminable	No condition is explicitly not met, but ≥1 condition is unevaluable due to absent clinical information	Clinical indication=recurrent	Line of therapy is required but absent. No condition is violated, so the case cannot be classified as ineligible, yet a determination cannot be reached (“unevaluable”).

Benchmark Validation

The benchmark was reviewed by 3 gynecologic oncology experts, including 2 nurse practitioners and 1 physician. Validation focused on the clinical plausibility of the patient scenarios, the appropriateness of attribute combinations, and the consistency of the reference answers.

For instance, the experts identified that palliative regimens inherently target patients with recurrent or metastatic disease, meaning that clinical indication is implicitly satisfied, and its removal alone does not produce a valid undeterminable case. For such regimens, additional clinical attributes had to be removed to construct undeterminable scenarios. The review also identified certain regimens that are permissively reimbursed in clinical practice despite not strictly meeting the guideline criteria. For these cases, reference answers were assigned based on the guideline criteria.

After clinical expert review, all benchmark cases were independently reviewed by a utilization review nurse, and discrepancies were resolved through consensus while preserving the balanced class structure. Interrater agreement between experts and the utilization review nurse was assessed using Cohen κ. This multidisciplinary validation approach reflects real-world reimbursement review processes, which involve both clinical and administrative expertise.

LLM Evaluation

We evaluated 6 LLMs from 3 providers: Gemini 3.1 Pro (gemini-3.1-pro-preview) and Gemini 3 Flash (gemini-3-flash-preview) from Google; Claude Opus 4.6 (claude-opus-4‑6) and Claude Sonnet 4.6 (claude-sonnet-4‑6) from Anthropic; and GPT-5.4 (gpt-5.4-2026-03-05) and GPT-5 Mini (gpt-5-mini-2025-08-07) from OpenAI. All models were evaluated under the primary condition, in which the original HIRA guideline PDF was provided via each model’s native document upload feature. Guideline documents were provided as 3 cancer-specific PDF files corresponding to cervical, uterine, and ovarian cancer. Because the set of user-adjustable hyperparameters differed across models and the evaluation included both reasoning and nonreasoning models, all models were evaluated using provider-default settings, and no parameter was explicitly modified. To assess response stability, each model was run 3 times under the same condition.

All models received the same standardized prompt template (Textbox 1). The system prompt instructed the model to act as an expert reviewer of Korean NHI reimbursement for oncology regimens and to return a structured JSON response containing a 3-class decision (eligible, ineligible, or undeterminable) and a single-sentence rationale. The user prompt specified the cancer type, regimen name, and structured clinical and administrative attributes.

XSL•FO
RenderX

Textbox 1. Standardized prompt template for reimbursement eligibility determination.

System Prompt:

You are an expert reviewer of Korean National Health Insurance reimbursement for oncology regimens.

Given the clinical and administrative attributes and any provided material, decide whether the named regimen is reimbursable.

Respond with exactly one JSON object using this schema:

{“decision”:“eligible|ineligible|undeterminable,”“reason”:“<one short sentence>”}

The following are the outcome definitions:

- Eligible: All required conditions are met.
- Ineligible: 1 or more conditions are explicitly not met, regardless of whether other conditions are unevaluable. The condition is present but violates the criterion.
- Undeterminable: No condition is explicitly not met, but 1 or more conditions are unevaluable due to absent clinical information. Because no criterion is violated, the case cannot be classified as ineligible, yet a determination cannot be reached.

User Prompt:

Cancer type:<cancer_type>

Regimen:<regimen_name>

Clinical and administrative attributes:<structured attributes>

Task: Determine whether this patient is eligible for reimbursement for the regimen above.

Statistical Analysis

Each model was evaluated over 3 independent runs. To capture run-to-run consistency of the models, accuracy was recorded to compute mean and SD. A final predicted label for each case was determined by majority voting across the 3 runs. A 3-way tie was resolved as undeterminable, reflecting the model’s own uncertainty about the case.

All evaluation metrics—precision, recall, and F_1-score—were calculated from these majority-voted predictions. Ninety-five percent CIs for accuracy, precision, and recall were calculated by Wilson method, and those for the F_1-score were obtained by bootstrapping (n=1000). The Bhapkar test of marginal homogeneity was used to assess whether the predicted outcome distribution differed from the observed distribution across the 3 outcome classes. Analyses were performed using Python 3.10 with SciPy.

Error Analysis

We further conducted a retrospective error analysis of all misclassified cases to characterize how models failed across the benchmark. Because the benchmark was explicitly designed around a tristate adjudication framework, model errors could be interpreted according to the adjudication principle they violated.

We therefore conducted a logic-based error analysis in which each expected-predicted outcome pair was mapped to an error category. Representative rationales and guideline locations were then reviewed to characterize recurrent structural sources of the dominant error category. This rationale review was performed blinded to model identity to mitigate potential reviewer bias.

Sensitivity Analysis

Sensitivity analysis consisted of 3 strategies. First, to examine the effect of input format, the guideline file (PDF) was parsed into structured text (Markdown) using the pdfplumber library, and all models were evaluated with the extracted text as input. Second, to assess the effect of information access, search-enabled conditions were evaluated on all models. This setting was intended to reflect scenarios in which patients or health care providers use LLMs with web search enabled. Third, to investigate the effect of prompting strategy, a prompt informed by the error analysis findings was evaluated, instructing models to verify each eligibility condition step by step.

For the sensitivity analysis, each prediction was dichotomized as correct or incorrect against the reference standard. Pairwise comparisons of correctness between the baseline and each sensitivity condition were then performed using McNemar test, with Benjamini-Hochberg correction for the false discovery rate. Adjusted P values are reported as q values.

Ethical Considerations

This study did not involve human participants, real patient data, or protected health information. All patient scenarios were synthetically constructed based on publicly available reimbursement guidelines published by the HIRA in South Korea.

Results

Benchmark Statistics

Agreement between the utilization review nurse and gynecologic oncology experts was initially 95% (211/222; Cohen κ=0.93). Eleven cases were identified in which specific qualifying conditions had not been fully reflected in the benchmark scenarios. These discrepancies were resolved through consensus while preserving the balanced class structure.

The final benchmark composition is summarized in Table 2. A total of 74 anticancer regimens (15 cervical, 17 uterine, and 42

XSL•FO
RenderX

ovarian) were each assigned 3 outcome scenarios, yielding 222 cases. The mean number of eligibility conditions per regimen was 4.2 (SD 1.9; range 2 - 11). Uterine (mean 4.5, SD 2.8; range 3-11) and ovarian cancer regimens showed comparable mean complexity (mean 4.5, SD 1.3; range 2-8 for both), but uterine cancer included the most complex individual regimens, with up to 11 eligibility conditions. Cervical cancer had the simplest criteria (mean 2.9, SD 1.4; range 2 - 6).

Table . Benchmark composition by cancer type (N=222).

Characteristic	Cervical	Uterine	Ovarian	Total
Regimens, n	15	17	42	74
Benchmark cases, n	45	51	126	222
Unique clinical attribute types, n	10	14	24	38[a]
Conditions per regimen, mean (SD; range)	2.9 (1.4; 2-6)	4.5 (2.8; 3-11)	4.5 (1.3; 2-8)	4.2 (1.9; 2-11)

[a]The total reflects the number of distinct attribute types after deduplication across cancer types.

The upper range of complexity in uterine cancer was largely driven by immunotherapy regimens such as pembrolizumab plus lenvatinib, which required up to 11 simultaneous conditions, including deficient mismatch repair or microsatellite instability-high status, Eastern Cooperative Oncology Group performance score, and prior treatment history.

Overall LLM Performance

Overall accuracy ranged from 77.9% (GPT-5.4) to 88.7% (Gemini 3.1 Pro) across the 6 models. No malformed outputs were observed. Performance was relatively stable across runs, with SDs ranging from 0.3 to 2.0 percentage points (pp). Despite these differences in overall accuracy, all models showed the same general pattern of substantially lower recall for undeterminable cases than for eligible or ineligible cases. Three-way ties across runs were rare, occurring in only 1 case out of all model evaluations (222 cases × 6 models). Detailed results are provided in Multimedia Appendix 1.

Analysis by Outcome Class

The LLM output was analyzed by 3 outcome classes (Figure 3). For eligible cases, recall was high across models (93.2% - 98.6%), and ineligible cases showed similarly strong performance (86.5% - 97.3%). By contrast, undeterminable cases showed a marked decline across all models, ranging from 44.6% to 70.3%. Among the models, Gemini 3.1 Pro showed the highest recall for undeterminable cases (70.3%), whereas GPT-5.4 showed the lowest (44.6%). The full results with precision, recall, F_1-score, and confusion matrix are provided in Multimedia Appendix 2.

Figure 3. Recall by model with 95% CI (n=74 per class, majority vote). Eligible (solid blue) cases met all required conditions. Ineligible (hatched blue) cases contained at least 1 explicitly not-met condition. Undeterminable (red) cases contained no explicitly not-met condition but at least 1 unevaluable condition due to missing information.

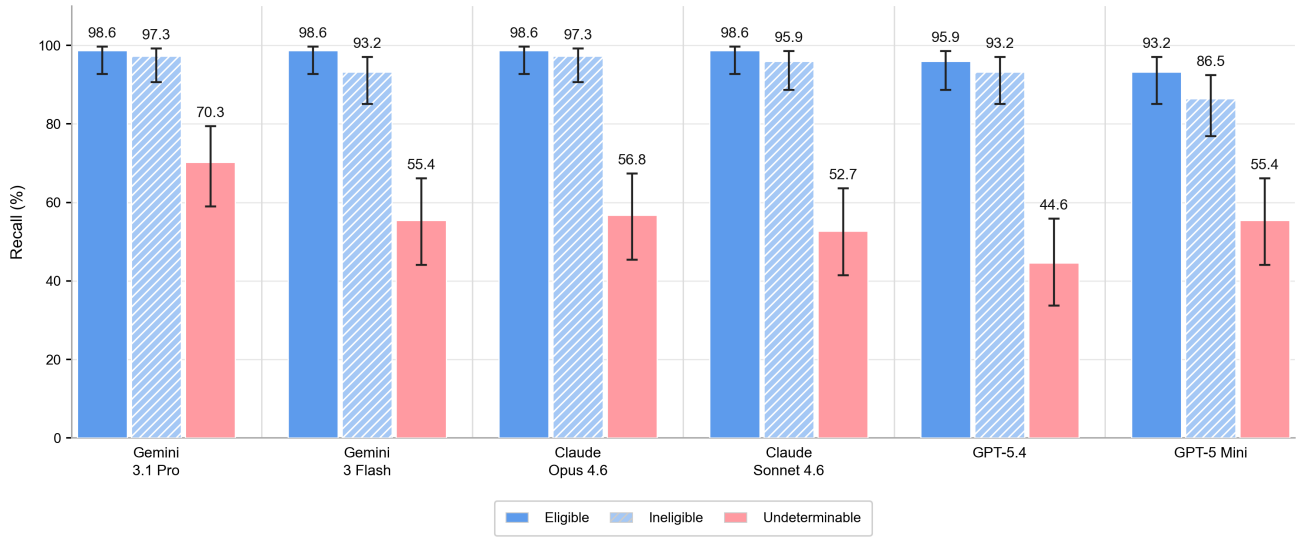

Analysis of misclassification direction for undeterminable cases (Table 3, 44.6% - 70.3%) revealed that models predominantly misclassified these cases as eligible (28.4% - 50%) rather than ineligible (0% - 5.4%). This pattern was observed across all 6 models. Consistent with this pattern, Bhapkar tests showed significant marginal heterogeneity between observed and predicted outcome distributions for all models (all P<.001; Multimedia Appendix 3).

XSL•FO
RenderX

Table . Misclassification direction for undeterminable cases by model (majority vote, n=74 per model).

Model	Predicted eligible, n (%)	Predicted ineligible, n (%)	Predicted undeterminable, n (%)
Gemini 3.1 Pro	21 (28.4)	1 (1.4)	52 (70.3)
Gemini 3 Flash	33 (44.6)	0 (0)	41 (55.4)
Claude Opus 4.6	31 (41.9)	1 (1.4)	42 (56.8)
Claude Sonnet 4.6	35 (47.3)	0 (0)	39 (52.7)
GPT-5.4	37 (50)	4 (5.4)	33 (44.6)
GPT-5 Mini	31 (41.9)	2 (2.7)	41 (55.4)

Analysis by Cancer Type

Performance varied across cancer types. Cervical cancer showed the highest mean recall across 6 models (92.6%), followed by ovarian (82.9%) and uterine (71.9%) cancer. Per-regimen complexity was lowest for cervical cancer (mean 2.9, SD 1.4; range 2 - 6 conditions) and highest for uterine cancer (mean 4.5, SD 2.8; range 3 - 11). Notably, recall varied substantially by outcome class within each cancer type (Table 4). While eligible and ineligible recall remained high across cancer types (89.7% - 100%), undeterminable recall dropped sharply, particularly for uterine cancer (mean 16.7%), where per-regimen complexity was highest. Model-wise results by cancer type and outcome class are provided in Multimedia Appendix 4.

Table . Pooled recall across cancer types and outcome classes with 95% CI (6 models).

Cancer type	Eligible, % (95% CI)	Ineligible, % (95% CI)	Undeterminable, % (95% CI)	Overall, % (95% CI)
Cervical	100 (95.9 - 100.0)	98.9 (94.0 - 99.8)	78.9 (69.4 - 86.0)	92.6 (88.8 - 95.2)
Uterine	99 (94.7 - 99.8)	100 (96.4 - 100.0)	16.7 (10.7 - 25.1)	71.9 (66.6 - 76.6)
Ovarian	95.6 (92.4 - 97.5)	89.7 (85.3 - 92.9)	63.5 (57.4 - 69.2)	82.9 (80.1 - 85.5)

Error Analysis

The tristate benchmark design enabled a logic-based characterization of model errors. All 235 misclassified cases were assigned to 1 of 3 error types according to how each misclassification violated the adjudication logic; predicted outcomes and single-sentence rationales were reviewed to characterize each category (Table 5).

Table . Error categories enabled by the tristate adjudication framework.

Error type and expected outcome	Predicted outcome	Explanation
Information gap-filling		
Undeterminable	Eligible	One or more conditions were not provided, but the model assumed a met value for the missing attribute.
Undeterminable	Ineligible	One or more conditions were not provided, but the model assumed a not-met value for the missing attribute.
Criterion misapplication		
Eligible	Ineligible	All conditions were met, but the model applied a restriction not present in the guideline.
Ineligible	Eligible	One or more conditions were explicitly not met, but the model confirmed eligibility based on the remaining met conditions.
False uncertainty		
Eligible	Undeterminable	All conditions were met, but the model treated one or more conditions as unevaluable and withheld judgment.
Ineligible	Undeterminable	One or more conditions were explicitly not met, but the model prioritized an unevaluable condition over the not-met finding.

The most common error was information gap-filling (196/235, 83.4%), in which models inferred a definitive outcome from incomplete scenarios rather than recognizing missing information. The second most common error type was criterion

XSL•FO
RenderX

misapplication (n=20, 8.5%), in which models classified cases as eligible despite 1 or more clearly not-met conditions or, conversely, classified cases as ineligible despite all conditions being met. Third, false uncertainty (n=19, 8.1%) refers to cases where models inferred undeterminable despite sufficient information to adjudicate eligible or ineligible. This pattern occurred most frequently in ineligible scenarios where 1 or more conditions were unevaluable but at least 1 clearly unmet condition was present.

To further understand the dominant error type, information gap cases were analyzed in relation to guideline structure. Reimbursement guidelines are organized hierarchically (Textbox 2), and model performance varied with structural complexity. In particular, subheading (L2) and footnote (L4) were the weak points of the models.

Textbox 2. Example structure of anticancer drug eligibility guideline.

[L1] Global Caveat: Cancer-type-level conditions applied to all regimens (eg, default scope of covered indications)

[L2] Treatment Intent A

└── Table

└──[L3] Regimen A [Indication A]

└──[L3] Regimen B [Indication B]

└──…

[L2] Treatment Intent B

└── Table

└──[L3] Regimen A [Indication C]

└──[L3] Regimen C [Indication D]

└──…

[L2] Treatment Intent C

└──[L3] Regimen D [Indication E]

└──[L3] Regimen E [Indication F]

└──…

[L2] Treatment Intent D

└──[L3] Regimen F [Indication G]*

└──….

[L4] *Footnote: Externalized condition at document tail, applied via explicit cross-reference (eg, “see Note 1”)

The largest subtype involved models ignoring the document subheading (L2) that specifies therapeutic intent (eg, palliative, maintenance). This distinction is clinically significant because identical regimens may appear under multiple therapeutic purposes with different indications, and reimbursement is granted only for the purpose under which the regimen is listed.

Another subtype involved footnote (L4) recognition. Designated institution requirement for immune checkpoint inhibitor, restricting eligible prescribing sites to those with qualified specialist staffing, appeared only in guideline footnotes, and most models systematically missed this condition.

Sensitivity Analysis

Building on the failure modes identified in the error analysis, 3 sensitivity conditions were evaluated (Multimedia Appendix 5). These analyses included structured text, web search–enabled, and structure-guided prompt conditions.

In the structured text condition, the parsed guideline in Markdown format was provided to the models. Performance declined across all 6 models, with statistically significant degradation observed in the 2 Claude models (Claude Opus 4.6: $\Delta=-10.4$ pp; Claude Sonnet 4.6: $\Delta=-10.8$ pp; both $q<.001$). The remaining 4 models showed smaller declines that did not reach significance.

The web search–enabled condition assessed the effect of additional information access alongside the guideline document, while the structure-guided prompt condition incorporated the hierarchical levels of the guideline (Textbox 2) into the prompt and instructed models to verify each eligibility condition step by step against the corresponding level. Neither condition produced a statistically significant change from baseline in any model (all $q>.05$).

Notably, web search tool invocation rates ranged from 0% to 3.2% across the 6 models under the web search–enabled condition. Token usage nevertheless increased substantially: OpenAI models showed approximately 4500 additional input tokens per case from the embedded tool specification, while Claude models produced longer reasoning.

XSL•FO
RenderX

Discussion

Principal Results

This study developed a benchmark for anticancer drug reimbursement eligibility determination covering 3 gynecologic cancers and used it to evaluate 6 LLMs from 3 providers. The benchmark was based on a tristate adjudication framework that distinguishes condition-level states (met, not met, unevaluable) and case-level outcomes (eligible, ineligible, and undeterminable) for eligibility verification. Case-level outcomes are derived from the aggregation of condition-level states.

The principal findings are threefold. First, overall accuracy varied across models, ranging from 77.9% to 88.7%, indicating broadly comparable performance across the 6 evaluated models. Second, performance varied sharply by outcome class. While eligible and ineligible cases were classified with consistently high recall, undeterminable cases posed the greatest challenge across all models, revealing a consistent limitation to recognize when clinical information is insufficient for a determination. Third, the tristate benchmark design enabled a logic-based characterization of model errors. Rather than treating all incorrect predictions as equivalent, the framework distinguished information gap-filling, criterion misapplication, and false uncertainty according to how each misclassification violated the expected adjudication logic.

These findings suggest that LLMs may help reduce the manual burden of eligibility review for clearly eligible and ineligible cases. However, these models should be positioned as decision-support tools operating under human oversight, not as autonomous decision-makers. The systematic failure in undeterminable cases represents the most clinically consequential limitation, as models tend to resolve ambiguity toward eligibility rather than deferring judgment when information is incomplete.

The dominant undeterminable failure pattern—information gap-filling—has direct clinical implications. When models misclassified undeterminable cases, the predominant direction was toward eligible rather than ineligible, indicating a tendency to infer eligibility from incomplete information. The severity of this pattern varied by cancer type, corresponding to the number of eligibility conditions per regimen. Uterine cancer regimens, which had the largest mean number of eligibility conditions, showed the lowest undeterminable recall.

This tendency has potential financial implications for hospitals. Under South Korea's retrospective claims review structure, ineligibility identified during postadministration audit can result in full denial of previously reimbursed drug costs, with the burden shifted to the prescribing institution. Because high-cost agents such as immune checkpoint inhibitors, antiangiogenic agents, and poly (ADP-ribose) polymerase inhibitors are reimbursed under exception-coverage provisions for severe diseases, even a small number of misclassifications can accumulate into substantial losses. Given that per-cycle costs typically range from KRW 3 - 6 million (approximately US $2200 - $4400), cumulative nonrecoverable costs can reach tens of millions of KRW per patient. Although the per-case inference cost itself is minimal (approximately US $0.002 - $0.10 across the 6 models), an erroneous determination at this cost can still translate into substantial financial losses through retrospective denial.

Meanwhile, sensitivity analysis revealed 3 patterns. First, converting the guideline file to structured text degraded performance across all models, indicating that document structure conveys eligibility logic lost during text extraction. Second, enabling web search did not produce a statistically significant change from baseline for any model (all $q>.05$), with models defaulting to the provided guideline rather than calling the search tool. This pattern indicates that in realistic scenarios where authoritative source documents are accessible to the LLM, the models rely on the provided context, and enabling the search tool neither improved nor degraded outcomes while increasing inference cost. Third, structure-guided prompting did not produce a statistically significant improvement for any of the 6 models (all $q>.05$), indicating that explicit structural guidance combined with step-by-step per-condition verification was insufficient to overcome the inherent limitations in hierarchical guideline reasoning observed in this benchmark.

Taken together, these results indicate that current LLMs are not yet reliable as standalone tools for reimbursement eligibility determination, particularly when clinical information is incomplete. Because neither input reformatting, information augmentation, nor prompt modification alone reliably improved performance, more systematic integration strategies will be needed to deploy LLMs safely in reimbursement workflows.

Comparison With Prior Work

Most LLM benchmarks in the medical domain have focused on knowledge assessment through multiple-choice examinations such as MedQA [21], while real-world task evaluation remains scarce [22]. Recent frameworks such as MedHELM [23] have begun addressing this gap but reveal that administrative and workflow tasks remain the weakest evaluation category.

Within guideline-based reasoning, MedGUIDE [24] and CPGPrompt [25] evaluated LLM adherence to cancer treatment decision trees and referral classification guidelines, respectively, both finding that models frequently deviate from structured conditional logic. Outside the medical domain, RuleArena [26] reported similar difficulties with complex rule-guided reasoning. However, none of these benchmarks test whether models can recognize when information is insufficient for a determination; all evaluate scenarios in which sufficient information is provided and a definitive answer exists.

This limitation is directly relevant to our findings. Recent abstention studies—AbstentionBench [27] across 20 datasets and MedAbstain [28] in medical question answering—showed that providing an explicit abstention option consistently increased model uncertainty and safer abstention behavior, whereas scaling model size or applying structure-guided prompting yielded little improvement.

Our undeterminable outcome class parallels these abstention scenarios in a domain-specific context. However, whereas these benchmarks test whether models can detect missing information within a self-contained question, our benchmark additionally

XSL•FO
RenderX

requires models to extract eligibility conditions from a hierarchically structured guideline document and then determine whether the patient scenario provides sufficient information to evaluate each condition.

To our knowledge, no prior benchmark has evaluated LLM performance on health insurance reimbursement eligibility. This study addresses this gap by introducing a benchmark grounded in national anticancer drug coverage criteria with expert validation, incorporating an undeterminable outcome class that explicitly tests recognition of incomplete information.

Limitations

This study has several limitations. First, the benchmark was limited to 3 gynecologic cancers based on South Korean NHI guidelines and relied on a relatively small dataset. Its generalizability to other cancer types, guideline versions, or insurance systems requires further validation.

Second, all patient scenarios were synthetically constructed in a structured format. This design does not capture the additional uncertainty introduced by extracting relevant attributes from unstructured records. Because real-world clinical documentation is often incomplete and unstructured, the performance observed in this study may overestimate model utility in practice.

Third, some eligibility conditions appeared in the global caveat and footnotes; errors involving these structural elements may partly reflect document representation rather than reasoning limitations of models alone.

Fourth, our benchmark adopts a strict rule that treats any missing information required by guideline criteria as undeterminable. This reflects a fundamental distinction between clinical reasoning and administrative reasoning. For instance, the absence of a prior treatment record may suggest a patient who has not previously been treated in clinical reasoning but constitutes missing evidence in administrative adjudication. Findings should, therefore, be interpreted as specific to administrative reasoning tasks.

Fifth, while the qualitative error analysis was conducted with model identities blinded, LLM outputs often contain stylistic fingerprints (eg, characteristic phrasing or cadences) that may make perfect blinding difficult. Residual unblinding may, therefore, introduce some bias into our error categorization.

Finally, models were tested with provider-default settings without fine-tuning or agentic frameworks, which may underestimate achievable performance.

Future Work

Several directions arise from these limitations. Future work should expand the benchmark to additional cancer types and guideline versions and validate it using real clinical documentation to better reflect performance under unstructured and incomplete records. A direct human performance baseline, in which experts solve the same cases under identical conditions, would also be valuable to contextualize LLM accuracy against expert-level performance.

Methodologically, agentic approaches that iteratively verify each condition against the guideline may improve accuracy on complex hierarchical structures. In addition, hybrid neuro-symbolic architectures that pair LLM-based understanding with structured rule engines have shown initial promise in adjacent health care tasks [29]; their applicability to reimbursement reasoning under incomplete information, particularly in handling rigid AND/OR logical constraints, remains to be evaluated.

Conclusion

The tristate adjudication framework introduced in this study offers a distinctive approach to evaluating LLM behavior in logical reasoning contexts characterized by variable information completeness. In this benchmark, LLMs classified clearly eligible and ineligible cases with relatively high recall but showed limited reliability on undeterminable cases. The dominant error pattern was information gap-filling, in which models tended to infer eligibility rather than withhold judgment. These findings indicate that LLMs, in their current form, should be deployed as supervised decision-support tools rather than as independent adjudicators in reimbursement review.

Acknowledgments

The authors used Claude (Opus 4.6, Anthropic, 2026) to assist with improving the readability and language editing of the manuscript. All outputs were reviewed and verified by the authors, who take full responsibility for the final content.

Funding

The authors declared that no financial support was received for this work.

Data Availability

Upon publication, the benchmark dataset, prompt templates, and evaluation scripts are available on GitHub [30] and archived on Zenodo [31].

Authors' Contributions

Conceptualization: JS, TK
Data curation: JS
Formal analysis: JS

XSL•FO
RenderX

Methodology: JS, TK, JHK
Software: JS
Supervision: TK, JHK
Validation: JHK
Visualization: JS
Writing – original draft: JS
Writing – review & editing: JS, TK, JHK

Conflicts of Interest

TK is cofounder and chief executive officer of ETOILE Inc, a health care artificial intelligence company. JS is an employee of ETOILE Inc. The benchmark described in this manuscript is not a commercial product of ETOILE Inc. The authors have no financial relationships with any of the commercial large language model vendors evaluated in this study. JHK reports no competing interests.

Multimedia Appendix 1
Overall accuracy across repeated runs and majority-voted with 95% CI.
[DOCX File, 17 KB - jmir_v28i1e95877_app1.docx]

Multimedia Appendix 2
Large language model performance per class with majority voting (n=74 per class; 95% CI).
[DOCX File, 542 KB - jmir_v28i1e95877_app2.docx]

Multimedia Appendix 3
Bhapkar test of marginal homogeneity comparing each large language model's predicted distribution against the observed distribution (eligible=74, ineligible=74, undeterminable =74; balanced design) across the 3 eligibility categories.
[DOCX File, 16 KB - jmir_v28i1e95877_app3.docx]

Multimedia Appendix 4
Model-wise recall by cancer type and outcome class with 95% CI.
[DOCX File, 18 KB - jmir_v28i1e95877_app4.docx]

Multimedia Appendix 5
Sensitivity analysis results with pairwise comparisons against baseline using McNemar test with Benjamini-Hochberg correction.
[DOCX File, 18 KB - jmir_v28i1e95877_app5.docx]

Checklist 1
TRIPOD-LLM checklist.
[DOCX File, 28 KB - jmir_v28i1e95877_app6.docx]

References

1. Jiwani A, Himmelstein D, Woolhandler S, Kahn JG. Billing and insurance-related administrative costs in United States' health care: synthesis of micro-costing evidence. BMC Health Serv Res 2014 Nov 13;14:556. [doi: 10.1186/s12913-014-0556-7] [Medline: 25540104]
2. Tseng P, Kaplan RS, Richman BD, Shah MA, Schulman KA. Administrative costs associated with physician billing and insurance-related activities at an academic health care system. JAMA 2018 Feb 20;319(7):691-697. [doi: 10.1001/jama.2017.19148] [Medline: 29466590]
3. Erickson SM, Rockwern B, Koltov M, McLean RM, Medical Practice and Quality Committee of the American College of Physicians. Putting patients first by reducing administrative tasks in health care: a position paper of the American College of Physicians. Ann Intern Med 2017 May 2;166(9):659-661. [doi: 10.7326/M16-2697] [Medline: 28346948]
4. Kyle MA, Feng KY, Wade CG, Yaver M. Patient administrative burden: a scoping review. Health Aff Sch 2025 Nov;3(11):qxaf216. [doi: 10.1093/haschl/qxaf216] [Medline: 41278120]
5. Richman BD, Kaplan RS, Kohli J, et al. Billing and insurance-related administrative costs: a cross-national analysis. Health Aff (Millwood) 2022 Aug;41(8):1098-1106. [doi: 10.1377/hlthaff.2022.00241] [Medline: 35914203]
6. Sohn M, Jung M. Effects of public and private health insurance on medical service utilization in the National Health Insurance System: national panel study in the Republic of Korea. BMC Health Serv Res 2016 Sep 21;16(1):503. [doi: 10.1186/s12913-016-1746-2] [Medline: 27654146]

XSL•FO
RenderX

7. Kim JA, Yoon S, Kim LY, Kim DS. Towards actualizing the value potential of Korea Health Insurance Review and Assessment (HIRA) data as a resource for health research: strengths, limitations, applications, and strategies for optimal use of HIRA data. J Korean Med Sci 2017 May;32(5):718-728. [doi: 10.3346/jkms.2017.32.5.718] [Medline: 28378543]
8. Kim L, Kim JA, Kim S. A guide for the utilization of Health Insurance Review and Assessment Service National Patient Samples. Epidemiol Health 2014;36:e2014008. [doi: 10.4178/epih/e2014008] [Medline: 25078381]
9. Shin HC, Park YT, Lee YT, Jo EC. Healthcare utilization monitoring system in Korea. Healthc Inform Res 2015 Jul;21(3):184-190. [doi: 10.4258/hir.2015.21.3.184] [Medline: 26279955]
10. Yun J, Chang Y, Jo M, Heo Y, Kim DS. National expenditures on anticancer and immunomodulating agents during 2013-2022 in Korea. J Korean Med Sci 2025 Feb 10;40(5):e16. [doi: 10.3346/jkms.2025.40.e16] [Medline: 39938871]
11. Cancer drug and regimen guidelines: public announcements [Report in Korean]. : Health Insurance Review & Assessment Service; 2026 URL: https://www.hira.or.kr/bbsDummy.do?pgmid=HIRAA030023010000 [accessed 2026-06-02]
12. AI as a healthcare ally: how Americans are navigating the system with ChatGPT. : OpenAI; 2026 URL: https://cdn.openai.com/pdf/2cb29276-68cd-4ec6-a5f4-c01c5e7a36e9/OpenAI-AI-as-a-Healthcare-Ally-Jan-2026.pdf [accessed 2026-06-02]
13. Seo J, Choi D, Kim T, et al. Evaluation framework of large language models in medical documentation: development and usability study. J Med Internet Res 2024 Nov 20;26:e58329. [doi: 10.2196/58329] [Medline: 39566044]
14. Song JW, Park J, Kim JH, You SC. Large language model assistant for emergency department discharge documentation. JAMA Netw Open 2025 Oct 1;8(10):e2538427. [doi: 10.1001/jamanetworkopen.2025.38427] [Medline: 41118162]
15. Jones MD, Torgbi M, Tayyar Madabushi H. Improving the understandability of clinical guidelines: development and evaluation of a GPT-4-based pipeline. J Med Internet Res 2026 Feb 23;28:e81915. [doi: 10.2196/81915] [Medline: 41730207]
16. Zheng NS, Keloth VK, You K, et al. Detection of gastrointestinal bleeding with large language models to aid quality improvement and appropriate reimbursement. Gastroenterology 2025 Jan;168(1):111-120. [doi: 10.1053/j.gastro.2024.09.014] [Medline: 39304088]
17. Hou Z, Liu H, Bian J, He X, Zhuang Y. Enhancing medical coding efficiency through domain-specific fine-tuned large language models. Npj Health Syst 2025;2(1):14. [doi: 10.1038/s44401-025-00018-3] [Medline: 40321467]
18. Smith PC, Araya-Guerra R, Bublitz C, et al. Missing clinical information during primary care visits. JAMA 2005 Feb 2;293(5):565-571. [doi: 10.1001/jama.293.5.565] [Medline: 15687311]
19. Burnett SJ, Deelchand V, Franklin BD, Moorthy K, Vincent C. Missing clinical information in NHS hospital outpatient clinics: prevalence, causes and effects on patient care. BMC Health Serv Res 2011 May 23;11:114. [doi: 10.1186/1472-6963-11-114] [Medline: 21605359]
20. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med 2025 Jan;31(1):60-69. [doi: 10.1038/s41591-024-03425-5] [Medline: 39779929]
21. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci 2021;11(14):6421. [doi: 10.3390/app11146421]
22. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 2025 Jan 28;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
23. Bedi S, Cui H, Fuentes M, et al. Holistic evaluation of large language models for medical tasks with MedHELM. Nat Med 2026 Mar;32(3):943-951. [doi: 10.1038/s41591-025-04151-2] [Medline: 41559415]
24. Li X, Gao M, Hao Y, et al. MedGUIDE: benchmarking clinical decision-making in large language models. Presented at: 39th Conference on Neural Information Processing Systems (NeurIPS 2025); Dec 2-7, 2025 URL: https://openreview.net/pdf/07c2a85f2355803fe92312f6a9808c0e5065ddef.pdf [accessed 2026-06-02]
25. Deng R, Martin G, Wang T, et al. CPGPrompt: translating clinical guidelines into large language model-executable decision support. J Am Med Inform Assoc 2026 Apr 1;33(4):855-862. [doi: 10.1093/jamia/ocag026] [Medline: 41746783]
26. Zhou R, Hua W, Pan L, et al. RuleArena: a benchmark for rule-guided reasoning with LLMs in real-world scenarios. Presented at: The 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Jul 27 to Aug 1, 2025. [doi: 10.18653/v1/2025.acl-long.27]
27. Kirichenko P, Ibrahim M, Chaudhuri K, Bell SJ. AbstentionBench: reasoning llms fail on unanswerable questions. Presented at: 39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks; Dec 2-7, 2025 URL: https://proceedings.neurips.cc/paper_files/paper/2025/hash/fb122bfc3f0127a94ded048b5b03496f-Abstract-Datasets_and_Benchmarks_Track.html [accessed 2026-06-02]
28. Machcha S, Yerra S, Gupta S, et al. Knowing when to abstain: medical LLMs under clinical uncertainty. Presented at: 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers); Mar 24-29, 2026. [doi: 10.18653/v1/2026.eacl-long.291]
29. Prenosil GA, Weitzel TK, Bello SC, et al. Neuro-symbolic AI for auditable cognitive information extraction from medical reports. Commun Med (Lond) 2025 Nov 21;5(1):491. [doi: 10.1038/s43856-025-01194-x] [Medline: 41272253]
30. Seo-see/k-NHIB. GitHub. URL: https://github.com/seo-see/K-NHIB [accessed 2026-06-07]
31. K-NHIB: Korean National Health Insurance Benchmark. Zenodo. URL: https://doi.org/10.5281/zenodo.20563360 [accessed 2026-06-07]

XSL•FO
RenderX

Abbreviations

HIRA: Health Insurance Review & Assessment Service
LLM: large language model
NHI: National Health Insurance
pp: percentage points
TRIPOD-LLM: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis-Large Language Model

Edited by A Coristine; submitted 22.Mar.2026; peer-reviewed by A Kodan, O Bursa, Y Cao, Z Li; revised version received 22.May.2026; accepted 25.May.2026; published 15.Jun.2026.

Please cite as:
Seo J, Kim T, Kim JH
Assessing Eligibility for Anticancer Drug Health Insurance Reimbursement Using Large Language Models: Benchmark Development and Comparative Study
J Med Internet Res 2026;28:e95877
URL: https://www.jmir.org/2026/1/e95877
doi:10.2196/95877

© Junhyuk Seo, Taerim Kim, Ju-Hyun Kim. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Trusting Generative AI for Health Advice: Preregistered Survey Experiment

Asheley R Landrum[1], PhD; Nitin Verma[2], PhD; Amanda Kehrberg[1], MAS

[1]Walter Cronkite School of Journalism and Mass Communication, Arizona State University, 555 N Central Ave, Phoenix, AZ, United States
[2]School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, IL, United States

Corresponding Author:
Asheley R Landrum, PhD
Walter Cronkite School of Journalism and Mass Communication, Arizona State University, 555 N Central Ave, Phoenix, AZ, United States

Abstract

Background: Generative artificial intelligence (AI) systems are increasingly used for health information seeking, yet it remains unclear how the public evaluates AI-generated health advice relative to guidance from credentialed clinicians in digital environments. Understanding the conditions under which AI is perceived as credible is critical as these systems become integrated into digital health ecosystems.

Objective: This study examined how source type (a human nurse in an online portal, a health care–specialized generative artificial intelligence system framed as an "AI nurse," or ChatGPT [OpenAI], a general-purpose chatbot), message characteristics, contextual risk, values framing, and individual differences in medical skepticism and experience with AI shape credibility evaluations of the provided advice and its purported source.

Methods: In a preregistered online experiment, a national sample of US adult participants (N=1502) was randomly assigned to 1 of 3 source conditions and evaluated health advice across 3 scenarios: low risk (dietary advice for cholesterol), high risk (chest pain triage), and a morally sensitive scenario (egg freezing). Advice type (intuitive vs counterintuitive) was manipulated in the risk scenarios, and ideological framing (neutral, conservative-leaning, and liberal-leaning) was manipulated in the morally sensitive scenario. Primary outcomes included participants' perceived credibility of the advice and beliefs about whether the patient should follow it. Source-level perceptions of competence and benevolence were also assessed. Medical skepticism and prior AI experience were examined as moderators.

Results: Advice attributed to a human nurse was rated as more credible than advice attributed to either AI source. Message intuitiveness showed effects comparable to, and sometimes larger than, the effects of source: intuitive advice was perceived as more credible than counterintuitive advice, with this difference amplified in high-risk contexts. In the morally sensitive scenario, ideological framing influenced perceived bias but did not interact significantly with source. Medical skepticism moderated source evaluations: higher skepticism was associated with greater perceived competence of the AI nurse and lower perceived competence of the human nurse.

Conclusions: Generative AI is evaluated within existing credibility frameworks rather than dismissed outright as inferior to human expertise. While licensed clinicians retain a credibility advantage, AI-generated advice is generally perceived as competent and legitimate. Importantly, individuals skeptical of traditional medical authority may evaluate AI-based guidance more favorably, suggesting that AI systems may redistribute—rather than uniformly erode—trust in health advice. As AI tools become embedded in patient-facing health platforms, message design and audience characteristics may shape acceptance more strongly than source labeling alone.

(J Med Internet Res 2026;28:e97882) doi:10.2196/97882

KEYWORDS

artificial intelligence; ChatGPT; digital health; health communication; health information seeking; medical skepticism; risk perception; trust

Introduction

Background

Generative artificial intelligence (AI) tools are increasingly embedded in everyday information-seeking practices, including domains traditionally reserved for credentialed experts [1]. Health advice is no exception, and a sizable proportion of people report turning to general-purpose AI tools, such as ChatGPT (OpenAI), for health-related inquiries [2]. Notably, platform developers are formalizing this use: OpenAI, for example, has announced a new tool, ChatGPT Health, stating that

XSL•FO
RenderX

health-related information-seeking is one of ChatGPT's most popular uses, with "hundreds of millions of people asking health and wellness questions each week" [3]. As these systems increasingly function as media for transmitting scientific knowledge to lay publics [4], they occupy a unique position within the health communication ecosystem: neither fully institutional nor fully interpersonal, neither a licensed professional nor an anonymous webpage [5]. Prior research has long established that trust in the source of health information is central to whether individuals act on medical advice [6-8]. Yet it is unclear how generative AI is evaluated when it offers health guidance: do people treat it as a credible source of information, and, if so, under what conditions?

Epistemic Trust and Source Credibility

Epistemic trust is the willingness to accept information from another as reliable, relevant, and generalizable; that is, to treat the information source as a legitimate knower whose claims can guide belief and action [9-11]. Such trust is typically grounded in evaluations of source credibility—often conceptualized in terms of competence or expertise and benevolence or trustworthiness [12,13]. Indeed, decades of social science research demonstrate that people's perceptions of source credibility influence whether and to what extent messages are accepted [14]. In science communication, trusting sources based on perceived expertise has been shown to influence public responses to scientific recommendations, particularly in domains involving uncertainty or risk [15]. In health contexts, trust in medical professionals predicts adherence to treatment recommendations, preventive behaviors, and willingness to seek care [16,17]. As new "actors" enter the health communication environment, understanding the circumstances under which patients deem messages and sources credible is essential for anticipating downstream effects on public health behavior.

Licensed medical professionals have traditionally occupied positions of epistemic authority [18]; research finds that physicians and nurses are typically perceived as very trustworthy, even amid declining trust in other types of experts and social institutions [19] (see also [20,21]). The clinical encounter itself is structured around asymmetries in knowledge, with patients typically deferring to experts to interpret symptoms, assess risks, and obtain recommendations for appropriate courses of action. Thus, we may assume that credentialed human providers presently serve as a benchmark of credibility against which other (nonhuman) sources ought to be evaluated.

At the same time, epistemic authority does not imply infallibility. Health care professionals, like other humans, are susceptible to cognitive biases, knowledge gaps, time constraints, and diagnostic errors. Patients' lived experiences, including misdiagnosis, perceived dismissal, or inequitable treatment, may further complicate their assumptions of health care practitioners' credibility. A trust deficit in this sense could reflect antiexpert sentiment, but it also may arise from an awareness of human fallibility and structural limitations within health care systems. Against this backdrop, generative AI systems that can digest and synthesize extensive medical literature may establish new benchmarks for evaluating the quality of advice. AI systems trained on expansive, peer-reviewed corpora may be perceived as less prone to individual bias, less influenced by interpersonal dynamics, and more consistently evidence-based than any single human provider is capable of being [22,23]. To the extent that patients equate credibility with comprehensiveness, neutrality, or data-driven reasoning, such systems could, at least in some contexts, be viewed as rivaling or even exceeding the perceived credibility of credentialed—but inherently human—experts.

Importantly, however, generative AI systems themselves are not monolithic. If patients evaluate human credibility in part by inferring competence and knowledge boundaries, then perceptions of AI expertise could also depend on cues that signal domain specialization [24]. Research on expertise heuristics demonstrates that audiences rely on markers such as credentials, role labels, institutional affiliation, and context to assess whether a source is qualified to speak on a given issue [25,26]. Use of such heuristics may also extend to AI agents [27,28]. A system described as a general-purpose chatbot may be evaluated differently than one framed as a custom-trained medical assistant or "AI nurse," even when both rely on similar underlying architectures—and especially if users assume differences in training data, evidentiary standards, or scope of knowledge. Consistent with this logic, studies of algorithmic systems suggest that perceived domain fit and task alignment meaningfully shape trust in and compliance with AI recommendations [29]. Distinguishing between general-purpose AI tools and health care–specific systems enables us to examine whether and how specialization cues influence perceptions of credibility.

Message Characteristics and Other Factors

Although source characteristics play an important role in credibility assessments, message features themselves can shape whether advice is accepted (and can even influence perceptions of the source of the information [13,30]). In particular, individuals tend to evaluate information more favorably when it aligns with their prior beliefs and expectations [31]. Dual-process theories of persuasion suggest that when individuals rely on heuristic processing, cues such as coherence, familiarity, and intuitive plausibility are influential [32,33]. For example, people may view accurate advice that contradicts their common sense–based expectations as less credible. In health contexts, where individuals hold lay theories (as opposed to expert knowledge) about diet, illness, and risk, recommendations that seem counterintuitive may face heightened skepticism, regardless of the source.

The perceived stakes of a decision may also shape credibility judgments. Research on risk perception demonstrates that individuals process information differently in low- versus high-risk contexts, with heightened sensitivity to potential threats amplifying reliance on heuristics and affective cues [34-37]. Research on the affect heuristic suggests that when consequences are perceived as severe or uncertain, individuals rely more heavily on gut reactions and intuitive evaluations than on analytic scrutiny [34,38]. In health domains, high-risk scenarios can trigger defensive processing [39,40], which may magnify skepticism toward advice that appears counterintuitive.

XSL•FO
RenderX

If so, advice that violates lay expectations may incur a larger credibility penalty in higher-risk contexts than in lower-risk or routine ones.

Beyond risk and advice intuitiveness, credibility judgments may also be shaped by the extent to which proffered health advice activates personal values and identity. Research on motivated reasoning suggests that individuals evaluate information in ways that protect preexisting beliefs and moral commitments [41-45] (see also [31]). Health decisions involving reproduction and bodily autonomy, for example, may be evaluated in light of their own value orientations (particularly religiosity and political identity), rather than solely on medical grounds. When proffered advice signals alignment with particular values, it is likely to be perceived as biased, and such perceptions of bias may influence compliance and judgment of credibility independently of a source's signaled expertise. However, as noted earlier, people may expect less bias from AI sources than from human sources [22,23,46,47]. Examining health advice in contexts intertwined with individuals' ideological values and moral commitments, then, allows us to assess another context in which generative AI systems may be evaluated differently from human health care providers.

Individuals' judgments about the credibility of health information providers also reflect their broader orientations toward medical authority. Medical skepticism, for instance, combines institutional distrust with confidence in one's own ability to understand and manage health concerns [48]. Rather than rejecting medicine outright, participants with this orientation question the necessity of professional intervention, prefer self-directed decision-making, and treat personal experience as a meaningful source of health knowledge; and research links medical skepticism to unhealthy behaviors, lack of insurance, absence of a primary care provider, delayed or avoided physician visits, and limited use of preventive care [48]. Related work on institutional trust shows that individuals who distrust established authorities often discount expert recommendations, particularly when they perceive power or knowledge asymmetries [49]. At the same time, scholarship on self-efficacy and lay health expertise finds that people rely on personal judgment, informal networks, and their own experiences when evaluating medical information [50]. Together, these dynamics suggest that emerging technological advice sources—such as generative AI—may be evaluated more favorably by individuals who feel skeptical of traditional medical authority yet confident in their own interpretive capacities.

This Study

Here, we investigate how generative AI is evaluated as a source of health advice, comparing perceptions when participants are told that advice comes from a licensed human nurse, an AI nurse, or a general-purpose chatbot (ChatGPT). Drawing on source credibility and epistemic trust theory, risk perception research, and motivated reasoning, we preregistered directional hypotheses regarding the source manipulation (hypotheses 1a-1f and hypothesis 4). We did not preregister hypotheses regarding potential main effects of the message characteristics (hypotheses 2 and 3).

First, we hypothesized that human sources would be perceived as more competent (hypothesis 1a) and benevolent (hypothesis 1b) than AI sources. Similarly, we predicted that advice attributed to a licensed human nurse would be perceived as more credible (hypothesis 1c) and more worthy of adherence (hypothesis 1d) than advice attributed to either AI source (consistent with relevant work by Reis et al [51]). Second, based on motivated reasoning and dual-processing theories, we expected that (hypothesis 2a) intuitive advice would be judged as more credible than counterintuitive advice, with this difference amplified in high-risk contexts (an intuitiveness-by-risk-level interaction, hypothesis 2b). Third, in the morally sensitive scenario, we anticipated that (hypothesis 3) ideological framing would shape perceptions of advice bias and credibility, and we specifically expected (hypothesis 4) the AI sources to be perceived as less biased than the human nurse. Finally, informed by scholarship on epistemic trust, motivated reasoning, and perceptions of algorithmic objectivity, we explored (research question 1 [RQ1]) whether individual differences in (1) medical skepticism and (2) experience with AI tools moderate credibility evaluations of human and AI sources.

Methods

Ethical Considerations

The study was approved and categorized as exempt by the Arizona State University Institutional Review Board on October 3, 2024 (STUDY00020970). In accordance with best practices for the ethical and transparent reporting of internet-based surveys, this study adhered to the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) guidelines [52]. Refer to Checklist 1. No personal or identifying information was collected or stored. Our methods, hypotheses, and analysis plan were preregistered on the Open Science Framework on June 14, 2025, 1 month prior to data collection [53].

Participants

Participants (N=1502) were recruited via the Cint panel to approximate US national demographics. Our sample was 74% (1111/1502) White, 14% (208/1502) Black or African American, 6% (78/1502) Asian American or Pacific Islander, 2% (32/1502) American Indian or Alaska Native, and 6% (72/1502) other or nonresponse; 13% (190/1502) identified as Hispanic or Latino. Participants were 54% (807/1502) female and ranged in age from 18 to 99 (mean 47, SD 18) years, with regional distribution spanning the Northeast (18%, 271/1502), Midwest (22%, 329/1502), South (40%, 595/1502), and West (20%, 306/1502). The sample size was determined using a power analysis in R (R Foundation for Statistical Computing) for our preregistered hypotheses (assuming a medium effect size and 80% power) as well as the available amount of funds to pay for participant recruitment.

Design and Procedure

We conducted a survey experiment, built using Qualtrics (Qualtrics LLC) survey software, with data collection between July 14 and 15, 2025. We requested 1500 participants sampled to match US Census data (for adults aged 18 years and older)

XSL•FO
RenderX

from the Cint panel, which recruited potential participants and provided them with a link to our survey. After opening the survey link, reading the study description, and voluntarily consenting to participate, participants were randomly assigned (via a randomizing function in the survey software) to 1 of 3 source conditions: a licensed human nurse, a health care–specialized "AI nurse," or ChatGPT. To preserve data quality, individuals who started the study but failed an attention check requiring them to correctly identify the source of the advice in the study (ie, their source condition) were immediately dropped (they were not allowed to complete the rest of the survey) and their data were deleted. Thus, these individuals are not a part of the final sample of 1502 participants. Participants were compensated by Cint in line with their survey panel agreements. Median completion time was 11.97 (IQR 11.32) minutes.

Each participant read and answered questions about 3 scenarios that were presented in a randomized order: a low-risk scenario (whether to eat eggs when trying to reduce cholesterol), a high-risk scenario (whether to go to the emergency room for chest discomfort), and a morally sensitive (or values-relevant) scenario (whether to freeze one's eggs to preserve fertility for later in life). Each scenario showcased a conversation between a fictional person seeking health advice and the source of the advice. Scenario topics were chosen based on our unpublished pilot study. The full texts are available in Multimedia Appendix 1 and on our Open Science project page [54].

Our design included 2 substudies. Refer to Table 1. Substudy 1 (the low- and high-risk scenarios) used a mixed design with source (human nurse, AI nurse, or ChatGPT) and advice intuitiveness (intuitive or counterintuitive) as between-subjects factors and scenario risk level (low-risk and high-risk) as a within-subjects factor. Substudy 2 (the morally sensitive scenario) used a fully between-subjects design crossing source with advice bias (neutral, conservative-leaning, or liberal-leaning).

Table . Study design and sample size[a].

Substudy	Scenarios (within-subjects)	Source (between-subjects)		
		ChatGPT (n=487)	AI[b] nurse (n=527)	Human nurse (n=488)
Substudy 1	• Low-risk scenario (dietary advice) • High-risk scenario (chest pain)	Advice intuitiveness: Intuitive (n=253) OR Counterintuitive (n=234)	Advice intuitiveness: Intuitive (n=261) OR Counterintuitive (n=266)	Advice intuitiveness: Intuitive (n=241) OR Counterintuitive (n=247)
Substudy 2	• Morally sensitive scenario (egg freezing)	Advice bias: Neutral (n=147), Liberal (n=168), OR Conservative (n=172)	Advice bias: Neutral (n=165), Liberal (n=179), OR Conservative (n=183)	Advice bias: Neutral (n=173), Liberal (n=153), OR Conservative (n=162)

[a]Substudy 1 (source × advice intuitiveness × risk level) applied only to the low- and high-risk scenarios, whereas Substudy 2 (source× advice bias) applied only to the morally sensitive scenario. Scenario order was randomized between participants. The total sample size was 1502.

[b]AI: artificial intelligence.

After each scenario, participants rated advice credibility, the perceived ideological bias of the advice (perceived bias), and their agreement that the character in the scenario should follow the advice (take advice). After completing the scenarios, participants evaluated their perceptions of the source's competence and benevolence and completed measures of medical skepticism, religiosity, prior generative AI experience, and other demographics.

Manipulations and Measures

As stated above, we used 3 between-subjects manipulations. First, the source of health advice (ie, source) was randomly assigned and held constant across all three scenarios presented to a participant: (1) a licensed registered nurse (called "Nurse Dobson") responding via an online patient portal; (2) an AI nurse—a custom-trained generative AI designed to simulate a health care professional using medical data; or (3) ChatGPT—a general-purpose AI developed by OpenAI that can provide information across domains, including health, but is not purpose-built for clinical use. Second, advice intuitiveness (substudy 1) was randomly assigned (using the survey software's randomizing tools) as either intuitive (advice consistent with common expectations; eg, reducing saturated fat to lower cholesterol; seeking immediate emergency care for chest discomfort) or counterintuitive (advice that departs from common expectations but is accurate; eg, endorsing full-fat dairy while reducing carbohydrates; recommending acid-reflux management when a heart attack is unlikely based on contextual factors). Advice intuitiveness also remained consistent for each participant across both the low-risk and high-risk scenarios (the within-subjects manipulation in substudy 1); it was not relevant to the morally sensitive scenario. Third, in the morally sensitive scenario only (substudy 2), we manipulated advice bias between-subjects as neutral (medically focused and balanced framing), conservative-leaning (prioritizing starting a family and natural means of conception), or liberal-leaning (emphasizing reproductive autonomy and individual empowerment). The exact wording of each message appears in the supplementary materials on our Open Science Framework project page [54].

Advice credibility was measured with a single-item Likert-type scale asking, "How would you rate the credibility of the advice [PATIENT] received from [SOURCE]?" (1=not at all credible to 5=extremely credible). Belief that the patient should follow

XSL•FO
RenderX

the advice was also assessed with a single item, "In your view, should [PATIENT] take the advice given by [SOURCE]?" (1=definitely no to 4=definitely yes). Both items were asked after each of the 3 scenarios. Perceived ideological bias of the advice was measured using a single item with a directional scale ("Do you think the advice [PATIENT] received from [SOURCE] is politically biased?") ranging from 1=strong right-wing or conservative bias to 5=strong left-wing or progressive bias, with 3=no, I don't think it is politically biased and 99=other (please specify) followed by a textbox for open-ended responses. The 35 participants who selected "Other" were list-wise deleted from analyses using this variable, and the vast majority of these participants wrote "none," "not sure," or "I have no idea" in the textbox.

Two multiitem scales—one to measure perceived source competence and the other to measure perceived source benevolence—assessed source credibility after all 3 scenarios were presented. Perceived source competence was measured using five 7-point semantic differential items (experienced-inexperienced, expert-inexpert, competent-incompetent, poorly trained–well trained, and informed-uninformed [12,55]). These items showed good interitem reliability, with Cronbach α=0.84 (95% CI 0.83-0.86). Perceived benevolence was measured with four 7-point semantic differential items (eg, trustworthy-untrustworthy, likely to tell the whole story–unlikely to tell the whole story, biased-unbiased, and likely to separate fact and fiction–unlikely to separate fact and fiction [12,55]). The perceived benevolence items had weaker interitem reliability than the perceived source competence items (Cronbach α=0.66, 95% CI 0.63-0.69). As specified in our preregistration, we used the factor scores extracted from a graded response model (item response theory analysis [56]) in our analyses.

Several individual difference measures were also included for exploratory analyses. Medical skepticism was measured with 4 items from prior research [48]: "I understand my health better than most doctors do," "I can overcome most moderate or severe illnesses without help from a medically trained professional," "home remedies are often better than those prescribed by a doctor," and "if I get sick, it is my own behavior that determines how soon I get well again" (Cronbach α=0.71, 95% CI 0.68-0.73). These items reflect both doubt about the usefulness of medical intervention and higher self-efficacy in handling illness. We also measured participants' religiosity by averaging across 2 items (each centered and scaled). The first question asked participants how much guidance their faith, religion, or spirituality provides them in their day-to-day life (on a 6-point ordinal scale). The second question asked participants how frequently they pray (on a 5-point ordinal scale). We also asked participants how much experience they have using generative AI tools such as ChatGPT, DALL-E (OpenAI), Gemini (Google LLC), and so on. We provided 6 response options: no experience (29%, 433/1502), a little experience ("I've tried them once or twice," 23%, 350/1502), some experience ("I use them occasionally," 25%, 374/1502), a lot of experience ("I use them regularly for work, school, or personal tasks," 15%, 227/1502), expert-level experience ("I use or develop generative AI tools frequently and understand how they work," 4%, 62/1502), and other (please specify, n=6). Participants who selected "other" were list-wise deleted from analyses using this variable. This item was included because recent studies have shown that one of the strongest predictors of trust in generative AI technology is experience using that type of technology [57]. These items were in addition to standard demographic variables collected (eg, age, education, income, and gender).

Results

Overview

Our primary aim was to examine whether the perceived source of health advice, human nurse, AI nurse, or ChatGPT, influences participants' perception of the credibility of that advice and whether the advice should be taken by the person described in a given scenario (hypothesis 1). Second, we were interested in whether that trust (and/or differences in trust) varied based on health context (eg, scenario characteristics: hypothesis 2: intuitiveness of the advice given and level of risk inherent in the scenario; hypothesis 3: advice bias) and audience characteristics (eg, political views, prior experience with AI, and medical skepticism; RQ1a and RQ1b). A summary table indicating each hypothesis and whether it was supported by the findings is provided in Multimedia Appendix 2.

Substudy 1: Risk Level and Intuitiveness of Advice

For substudy 1, we had 2 outcome variables of interest: perceptions of the advice's credibility (advice credibility) and the belief that the character in the scenario should take the advice (take advice). We analyzed the data in R using mixed-design ANOVAs (one model per outcome variable) and report the results in Table 2 below.

XSL•FO
RenderX

Table . Results from substudy 1[a].

Outcome variable	Advice credibility			Take advice		
	F test (*df*)	*P* value	η_p^2	*F* test (*df*)	*P* value	η_p^2
Source	21.37 (2, 1494)	<.001	0.030	9.28 (2, 1491)	<.001	0.010
Advice intuitiveness (intuitiveness)	40.44 (1, 1494)	<.001	0.030	76.52 (1, 1491)	<.001	0.050
Source× intuitiveness	0.81 (2, 1494)	.45	0.001	0.06 (2, 1491)	.94	<0.001
Scenario risk level (risk)	7.41 (1, 1494)	.007	0.005	15.31 (1, 1491)	<.001	0.010
Source× risk	0.74 (2, 1494)	.48	<0.001	0.48 (2, 1491)	.62	<0.001
Intuitive× risk	18.54 (1, 1494)	<.001	0.010	32.17 (1, 1491)	<.001	0.020
Source× intuitiveness×risk	2.33 (2, 1494)	.10	0.003	0.04 (2, 1491)	.96	<0.001

[a] Source and advice intuitiveness (intuitive and counterintuitive) were the between-subjects variables, and scenario risk level (low risk and high risk) was the within-subjects variable.

When examining perceptions of advice credibility, there was a main effect of source, and follow-up tests with a Bonferroni correction found that advice from the human nurse (estimated marginal mean, ie, emmean 3.53, SE 0.05, 95% CI 3.44-3.63) was rated as more credible than that from the AI nurse (emmean 3.13, SE 0.05, 95% CI 3.04-3.22; *P*<.001) and ChatGPT (emmean 3.21, SE 0.05, 95% CI 3.11-3.30; *P*<.001), supporting hypothesis 1c. There was no significant difference in perceptions of advice credibility between the two AI sources (*P*=.68). We also found differences in advice intuitiveness and scenario risk level (Table 2). Counterintuitive advice (emmean 3.12, SE 0.04, 95% CI 3.04-3.19) was seen as less credible than intuitively correct advice (emmean 3.46, SE 0.04, 95% CI 3.39-3.54; *P*<.001), and advice given in the low-risk scenario (emmean 3.33, SE 0.03, 95% CI 3.27-3.39) was seen as more credible than advice given in the high-risk scenario (emmean 3.24, SE 0.03, 95% CI 3.18-3.31; *P*=.006).

The 3-way interaction between source, advice intuitiveness, and scenario risk level was only marginally significant (*P*=.10), whereas the 2-way interaction between advice intuitiveness and scenario risk level was statistically significant (*P*<.001), supporting hypothesis 2 (Table 2). Participants, averaged across the source conditions, perceived the counterintuitive advice as less credible than the intuitive advice, and the difference in perceived advice credibility between the counterintuitive and intuitive advice was greater for the high-risk scenario (difference estimate= –0.48) than for the low-risk one (difference estimate= –0.20). Refer to Figure 1.

Figure 1. Estimated marginal means for advice credibility ratings broken down by source, advice type, and scenario (advice credibility was measured on a scale of 1 (not at all credible) to 5 (extremely credible), and a rating of 3 indicated moderate credibility).

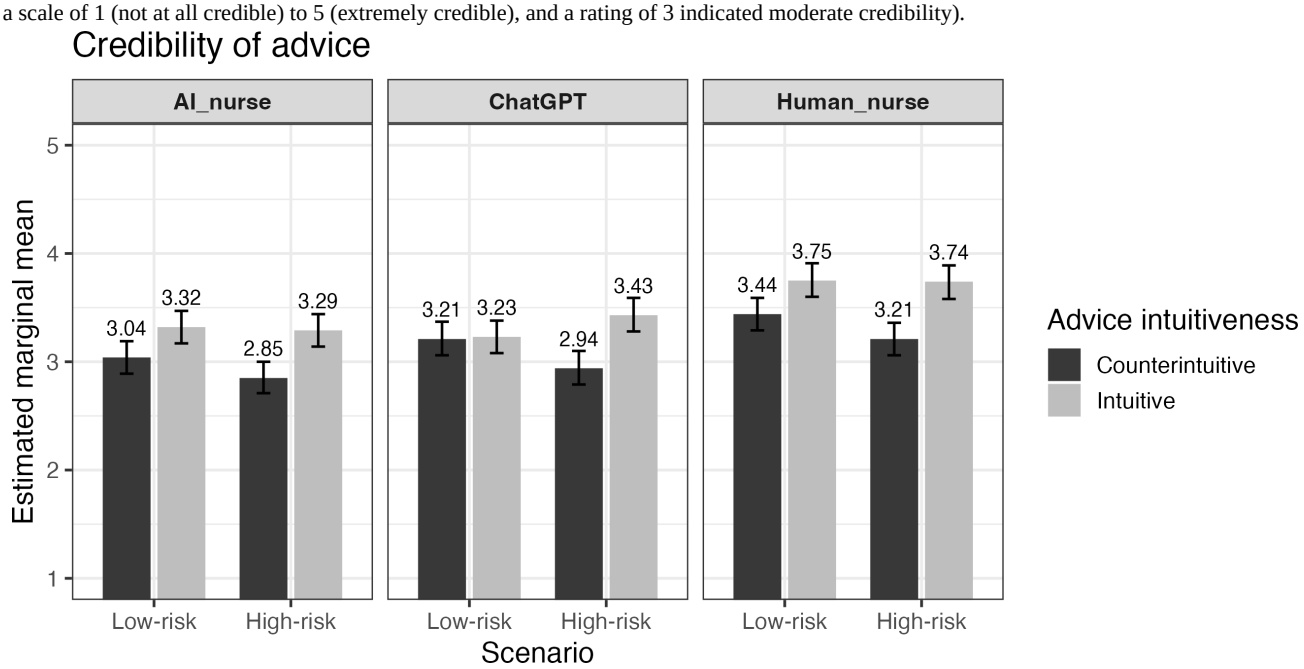

The results for participants' level of agreement that the scenario character should take the advice (ie, take advice) follow the same pattern as the results for participants' perceptions of advice credibility. Refer to Table 2. Supporting hypothesis 1,

XSL•FO
RenderX

participants who saw advice from the human nurse (emmean 3.36, SE 0.03, 95% CI 3.30-3.42) agreed more strongly that the scenario character should take that advice than participants who saw advice from an AI nurse (emmean 3.20, SE 0.03, 95% CI 3.15-3.26; *P*<.001) or ChatGPT (emmean 3.22, SE 0.03, 95% CI 3.16-3.28; *P*=.002). The difference between the AI nurse and ChatGPT was not significant (*P*=.89).

We also found a significant effect of advice intuitiveness (counterintuitive vs intuitive) on take advice that does not vary based on source (no significant advice intuitiveness by source interaction, *P*=.94). Overall, people agree more strongly that the scenario character should take the intuitive advice (emmean 3.40, SE 0.02, 95% CI 3.36-3.45) than the counterintuitive advice (emmean 3.12, SE 0.02, 95% CI 3.07-3.16; *P*<.001). There is also a significant effect of scenario risk level: participants agree more strongly that advice in the low-risk scenario is worth taking (emmean 3.31, SE 0.02, 95% CI 3.27-3.34) than advice in the high-risk scenario (emmean 3.22, SE 0.02, 95% CI 3.18-3.26; *P*<.001), and this, too, does not vary by source (no significant scenario risk level by source interaction, *P*=.62).

Supporting hypothesis 2, we found the expected significant interaction between advice intuitiveness and scenario risk level (Figure 1), such that participants, averaged across source conditions, perceived the counterintuitive advice as less credible than the intuitive advice, and the difference in take advice between the counterintuitive and intuitive advice types was greater for the high-risk scenario (difference estimate= –0.41) than for the low-risk one (difference estimate=–0.16; *P*<.001).

Substudy 2: Morally Sensitive Scenario and Ideological Bias of Advice

For substudy 2 (hypotheses 3 and 4), in addition to evaluating perceived advice credibility and taking advice, we also evaluated whether, and to what extent, source, advice bias, and individuals' own political ideology (ie, vote) influenced their perceptions of bias in the advice (ie, perceived bias). We tested these questions with a fully between-subjects ANOVA in R. Refer to Table 3. In the perceived bias analysis, we found no significant effects except for our advice bias manipulation (supporting hypothesis 3). Follow-up paired-samples *t* tests with Tukey correction showed that as intended, the conservative advice bias manipulation (mean 2.80, SD 0.96) was perceived as more conservatively biased than the liberal (mean 3.12, SD 0.90; Cohen *d*=0.34; *P*<.001) and neutral advice manipulations (mean 3.03, SD 0.87; Cohen *d*=0.25; *P*<.001), though the liberal advice did not differ significantly from the neutral advice (Cohen *d*=0.10; *P*=.27). Notably, the median response option across all 3 conditions was 3, which corresponded to "No, I don't think it's politically biased."

Table . Results from substudy 2[a].

	Advice credibility			Take advice			Perceived bias		
	F test (*df*)	*P* value	η_ρ^2	*F* test (*df*)	*P* value	η_ρ^2	*F* test (*df*)	*P* value	η_ρ^2
Source	12.47 (2, 1474)	<.001	.020	7.11 (2, 1472)	<.001	.009	0.95 (2, 1440)	.39	.001
Advice bias (bias)	11.93 (2, 1474)	<.001	.020	19.11 (2, 1472)	<.001	.030	15.98 (2, 1440)	<.001	.020
Vote	1.04 (2, 1474)	.36	.001	0.12 (2, 1472)	.89	<.001	0.06 (2, 1440)	.95	<.001
Source× bias	*0.78* (4, 1474)	.54	.002	0.37 (4, 1472)	.83	.001	0.54 (4, 1440)	.71	.001
Source× vote	1.29 (4, 1474)	.27	.004	1.77 (4, 1472)	.13	.005	1.34 (4, 1440)	.26	.003
Bias × vote	0.09 (4, 1474)	.98	<.001	1.45 (4, 1472)	.22	.004	0.40 (4, 1440)	.81	.001
Source× bias×vote	0.24 (8, 1474)	.98	.001	0.79 (8, 1472)	.61	.004	1.41 (8, 1440)	.19	.008

[a]These ANOVA use type I sums of squares.

When evaluating advice credibility and taking advice, we found an effect of advice bias. Conservatively biased advice was seen as less credible (mean 3.06, SD 1.23) than neutral advice (mean 3.42, SD 1.17; Cohen *d*=0.30; *P*<.001) and liberally biased advice (mean 3.34, SD 1.22; Cohen *d*=0.23; *P*<.001). Participants agreed more strongly that neutral advice (mean 3.37, SD 0.74; Cohen *d*=0.43) and liberally biased advice (mean 3.28, SD 0.74; Cohen *d*=0.34) should be taken than conservatively biased advice (mean 3.08, SD 0.81; both *P* values <.001).

Source Credibility

We modeled the relationships among source credibility (assessed by perceived source competence and benevolence), source, individual characteristics (basic demographics, voting behavior, and religiosity), prior experience with large language model–driven chatbots, and medical skepticism using a generalized linear model. As we wrote in our preregistration, this examination of the potential influence of individual differences variables on perceptions of competence and benevolence is exploratory. Refer to Table 4.

XSL•FO
RenderX

Table . Results of generalized linear model analyses predicting competence and benevolence perceptions[a].

Outcome variable	Competence			Benevolence		
	β	*P* value	lmg (%)	β	*P* value	lmg (%)
Source (AI[b] nurse vs ChatGPT)	0.04	.55	6.53	0.03	.52	5.05
Source (AI nurse vs human nurse)	0.56	<.001		0.49	<.001	
GenAI[c] experience	0.11	.02	0.47	0.11	.02	0.23
Medical skepticism	0.16	<.001	0.06	0.02	.61	0.37
Age	0.17	<.001	2.04	0.18	<.001	2.55
Income	0.04	.14	0.29	0.02	.40	0.10
Gender	0.06	.23	0.09	0.15	.006	0.51
Vote (democrat vs other)	–0.04	.51	0.16	–0.04	.52	0.11
Vote (democrat vs republican)	–0.01	.92		–0.05	.46	
Religiosity	0.05	.048	0.36	0.04	.12	0.23
Education	–0.01	.74	0.12	–0.02	.56	0.09
Source (AI nurse vs ChatGPT) × medical skepticism	–0.21	<.001	1.04	–0.10	.10	0.48
Source (AI nurse vs human nurse) × medical skepticism	–0.22	<.001		–0.17	.01	
Source (AI nurse vs ChatGPT) × GenAI experience	0.06	.34	0.43	–0.03	.60	0.03
Source (AI nurse vs human nurse) × GenAI experience	–0.10	.11		–0.04	.54	

[a] Standardized regression coefficients and relative importance using the lmg metric (which partitions model R^2 into predictor contributions averaged across all possible orderings).

[b]AI: artificial intelligence.

[c]GenAI: generative artificial intelligence.

When predicting both competence and benevolence, we found a significant effect of source where the AI nurse (mean competence score [M_{comp}] 4.82, SD 1.39; mean benevolence score [M_{bene}] 4.76, SD 1.25) is perceived as significantly less competent and benevolent than the human nurse (M_{comp} 5.6, SD 1.30; Cohen *d*=0.58; *P*<.001; M_{bene} 5.39, SD 1.30; Cohen *d*=0.50; *P*<.001). Similarly, ChatGPT was seen as less competent (M_{comp} 4.88, SD 1.34; M_{bene} 4.81, SD 1.24) than the human nurse (both *P* values <.001, both Cohen *d* values >0.40). There was no significant difference between the two AI sources (both *P* values >.29).

It is worth noting, however, that all sources were generally considered credible, with average ratings for both competence and benevolence above the midpoint of the scale (4). Generative AI experience also contributed to greater perceptions of competence and benevolence regardless of source. Interestingly, being older was associated with having greater perceptions of both competence and benevolence, averaged across source (Table 4).

Most interestingly, we found a significant interaction between source and medical skepticism, such that the relationship between participants' medical skepticism scores and their perceptions of competence and benevolence differs when rating the AI nurse versus the human nurse or ChatGPT. For example, as medical skepticism increases, perceptions of the AI nurse's competence increase, whereas perceptions of the human nurse's competence decrease. Refer to Figure 2. Simple effects correlation tests show that despite differences between the sources, the relationship between medical skepticism and credibility ratings is significant only for the AI nurse when rating competence and for the human nurse when rating benevolence. Refer to Table 5.

XSL•FO
RenderX

Figure 2. Relationships between medical skepticism and perceptions of source competence and benevolence by source condition (AI nurse, ChatGPT, and human nurse).

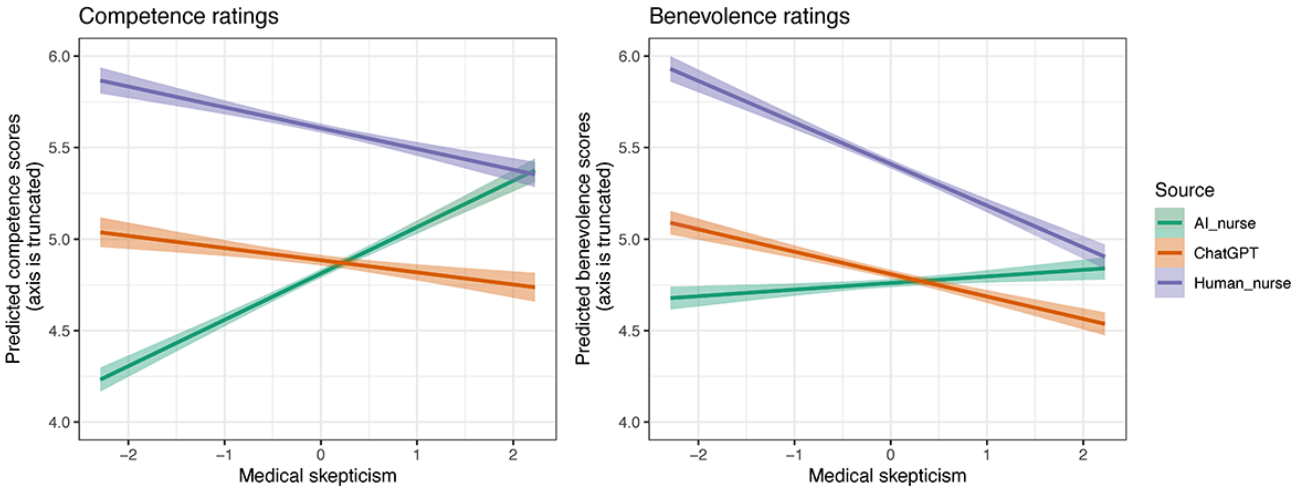

Table . Post hoc simple effects tests[a].

Medical skepticism with	Competence		Benevolence	
	r	*P* value	*r*	*P* value
Human nurse	–0.08	.10	–0.13	.005
AI nurse	0.18	<.001	0.03	.44
ChatGPT	–0.04	.41	–0.08	.08

[a]Correlations between competence ratings or benevolence ratings and medical skepticism are shown for each source condition.

Discussion

Findings Summary

This study examined how generative AI tools are evaluated as sources of health advice relative to a licensed human nurse, and how those evaluations vary across message characteristics, contextual risk, and individual orientations toward medical authority (ie, medical skepticism). Overall, the findings largely supported the hypotheses. Consistent with hypothesis 1, human nurses were perceived as more competent and benevolent than either AI source, and advice was perceived as more credible and more worthy of being followed when it was attributed to the human nurse than either AI source. Supporting hypothesis 2, intuitive advice was judged as more credible than counterintuitive advice, and this difference was amplified in high-risk contexts. Supporting hypothesis 3, ideological framing influenced perceptions of bias and credibility in the morally sensitive scenario. Hypothesis 4 was not supported, as AI-generated advice was not perceived as less politically biased (in either direction) than advice from the human nurse. Finally, exploratory analyses related to RQ1 showed that medical skepticism moderated source evaluations, such that higher levels of medical skepticism were associated with greater perceived competence of the AI nurse, whereas the opposite pattern emerged for perceptions of the human nurse's competence. In contrast, prior AI experience generally increased positive evaluations of all sources rather than differentially benefiting AI systems. Refer to Multimedia Appendix 2.

no our knowledge, this is among the first preregistered experiments to directly compare perceptions of advice credibility that purportedly comes from a human nurse, a domain-specialized AI system ("AI nurse"), or a general-purpose chatbot (ChatGPT), although we are aware of one paper that has examined perceptions of AI involvement in health advice (ie, [51], which compared an AI chatbot, a human physician, and a human physician collaborating with AI) and additional papers that have examined participants' evaluations of advice that actually comes from physicians versus AI tools [58,59].

Discussion of Key Findings

Several of our findings are particularly relevant for understanding how AI-mediated systems may reshape digital health ecosystems.

Message Characteristics Outweigh Source in Shaping Trust in Health Advice

Consistent with our first hypothesis [51], advice attributed to a licensed human nurse was rated as more credible and more worthy of taking than advice attributed to either AI source. However, these source effects were modest relative to message characteristics. Across all 3 sources, participants seemed to rely more on the intuitiveness of the advice itself. Intuitive recommendations were perceived as more credible than counterintuitive recommendations, particularly in high-risk contexts. These findings suggest that users are not treating AI-generated health advice as categorically distinct from human-delivered advice. Instead, they appear to apply similar advice-credibility heuristics—particularly whether the advice aligns with their expectations—to both human and artificial sources.

XSL•FO
RenderX

For digital health implementation, this pattern has important implications. As generative AI tools become integrated into patient portals, telehealth systems, and direct-to-consumer health platforms, developers and health systems may assume that labeling advice as AI-generated will meaningfully shift trust (up or down). Our results suggest that message content may exert a stronger influence than the source alone. In particular, counterintuitive but evidence-based recommendations may require additional explanation or scaffolding, especially in high-stakes scenarios where affective and heuristic processing are heightened.

Specialization Framing Failed to Boost Trust in AI Nurse Compared to ChatGPT

It is also noteworthy that we found no overall differences between ChatGPT and the AI nurse, despite anticipating that participants would be sensitive to the AI nurse's description as a custom system trained on credible medical data and would use this as a cue of domain-specific expertise. There are several possibilities for this lack of difference. First, our design was between-subjects, and participants evaluated each source in isolation rather than engaging in relative judgment, which may have attenuated any influence of subtle framing differences. Second, ChatGPT, as a widely recognized and frequently used tool, may benefit from reputational credibility or the mere exposure effect, whereas the AI nurse—though described as specialized—was novel and thus lacked an established mental model. Third, participants may already perceive general-purpose AI systems as broadly competent across domains, including health, particularly given the growing visibility of AI-assisted information seeking. In this sense, describing ChatGPT as "not purpose-built" for health may not have been sufficient to meaningfully diminish perceived expertise, especially in the absence of explicit cues about potential limitations or risks (eg, the possibility of error or hallucination).

Medical Skepticism Associated With AI Nurse Competence but Not Human Nurses or ChatGPT

In our view, the interaction between medical skepticism and source credibility evaluations represents the most theoretically and practically consequential finding. Individuals higher in medical skepticism perceived the AI nurse as more competent, whereas similar individuals perceived the human nurse and ChatGPT as less competent. Medical skepticism reflects not only institutional distrust but also heightened confidence in one's ability to manage one's own health concerns [48]. For these individuals, generative AI may serve as a more acceptable alternative. Unlike licensed clinicians, AI systems are not (currently) embedded within visible institutional hierarchies, professional guilds, or regulatory structures. Furthermore, people often perceive computational systems as more objective, data-driven, and free from interpersonal bias than human experts [46,47], though some recent research has found participants to be suspicious of AI involvement in medical advice [51]. A perception of mechanical neutrality can foster trust, particularly in contexts where human institutions are viewed as politicized, self-interested, or fallible. At the same time, interacting with generative AI may afford users a heightened sense of agency and self-efficacy. Because individuals actively formulate prompts, evaluate responses, and iteratively refine queries, AI-generated guidance may feel less like externally imposed expertise and more like the product of one's own information seeking. Research on online health information seeking and eHealth literacy suggests that independently sourcing and interpreting medical information can increase subjective knowledge, perceived control, and self-efficacy [60-62]. More broadly, self-determination theory and related work on autonomy suggest that when individuals perceive themselves as the originators of a decision process, they may be more likely to view its outcomes as legitimate [63]. In this way, AI systems may enable users to experience medical guidance as something they have "figured out" themselves, aligning more closely with values of autonomy and self-reliance. Consequently, individuals who question the legitimacy or authority of medical institutions may evaluate AI-based guidance not merely as a substitute for expertise, but as a different—and potentially more impartial and self-endorsed—source of knowledge.

That said, this was the one finding where ChatGPT and the AI nurse were evaluated differently. When it came to source competence, medical skepticism was positively associated with perceptions of the AI nurse's competence (r=0.18, P<.001), but it was not significantly associated with perceptions of ChatGPT's competence (r=–0.04, P=.41). The relationship between medical skepticism and perceptions of a source's benevolence or trustworthiness did not significantly differ between the AI nurse and ChatGPT. One explanation for this divergence lies in how medical skepticism might reorient credibility judgments. As discussed, medical skepticism reflects not only distrust in institutionalized medical authority but also greater confidence in one's own ability to evaluate health information. For these individuals, sources that preserve autonomy while still signaling relevant expertise may be especially appealing. The AI nurse uniquely satisfies both criteria: it carries cues of domain-specific medical knowledge while remaining outside traditional institutional hierarchies. In contrast, ChatGPT—framed as a general-purpose system—offers a similarly noninstitutional source but lacks clear signals of medical specialization. As a result, it may not meet the threshold for perceived competence in a health context, even among those predisposed to question traditional expertise.

This interpretation is consistent with research showing that audiences rely on domain-fit heuristics when evaluating expertise and that perceptions of credibility depend on whether a source is seen as appropriately qualified for a given task [24,28,29]. It also aligns with work suggesting that AI systems may be perceived as more impartial or less socially biased than human experts, which could elevate baseline perceptions of benevolence across AI sources [22,46,47]. Indeed, both AI sources may already be viewed as relatively neutral or non–self-interested, limiting differentiation on benevolence. In contrast, competence judgments appear more sensitive to the presence or absence of domain-specific expertise cues, possibly producing the observed divergence between the AI nurse and ChatGPT.

Importantly, this pattern suggests that medical skepticism does not simply increase or decrease trust in AI broadly. Rather, it appears to redistribute trust toward sources that combine

XSL•FO
RenderX

perceived expertise with independence from traditional medical authority. In this way, generative AI systems that are framed as domain-specialized—but not institutionally embedded—may be particularly well positioned to gain credibility among individuals who are skeptical of conventional health care systems.

Limitations

As with most research, this study has limitations. First, the AI nurse condition relied on descriptive framing to signal health care specialization. Participants may have inferred differences in training data, regulatory oversight, or institutional affiliation that were not explicitly specified. Future research could disentangle the effects of domain specialization cues from assumptions about accountability and evidentiary grounding. Second, our manipulation of ideological bias in the morally sensitive scenario was intentionally subtle, and median responses indicated that most participants did not perceive the advice as particularly politically biased (though they did perceive differences between conservative and liberal biased advice). Stronger or more explicit partisan cues may yield different credibility perceptions. Third, the study relied on hypothetical vignette scenarios, which necessarily simplify real-world health decision-making contexts. Although vignettes provide important methodological advantages, such as experimental control, standardized exposure across conditions, and the ability to examine sensitive or high-risk scenarios without requiring participants to disclose personal medical information, participants' responses to hypothetical scenarios may differ from reactions in real clinical situations. In particular, evaluations of health advice may change when individuals are personally experiencing fear, uncertainty, pain, or urgency, or when they have an ongoing relationship with a health care provider or AI system. Relatedly, the selected scenarios focused on a limited number of contexts; evaluations may differ in chronic disease management, mental health, or pediatric settings where relational continuity and affective trust may be more central. Finally, although our counterintuitive advice conditions involved counterintuitive recommendations, we did not examine responses to clearly incorrect AI guidance, which may produce different patterns of trust and source differentiation.

Conclusion

People are not abandoning human clinicians in favor of generative AI. Instead, AI appears to be yet another potential source of health advice that people evaluate using many of the same credibility judgments they apply to human sources, including perceptions of the source's competence or expertise and benevolence or trustworthiness, characteristics of the advice itself (such as whether it seems intuitively plausible), contextual factors such as perceived risk, and through the lens of the individual's orientations toward medical authority and institutional trust. Although licensed clinicians retain a credibility advantage, our study demonstrates that AI-based systems are nonetheless viewed as competent and legitimate sources of advice, particularly among individuals with prior AI experience and those skeptical of traditional medical authority. These patterns suggest that, rather than simply reducing trust in medical expertise overall, AI may change how people allocate trust across different sources of health information, with some individuals relying more heavily on AI systems alongside—or, in some cases, instead of—traditional health care providers. As AI tools become increasingly embedded in patient-facing health platforms, future research should examine how these evaluative processes operate in real clinical interactions and over time, particularly in a wider variety of contexts. Our findings also suggest that health communication interventions and AI system design should focus not only on informational accuracy but also on how message framing, transparency, and audience characteristics shape trust and willingness to follow evidence-based guidance. Understanding these evaluative processes will be critical for ensuring that AI systems support, rather than undermine, evidence-based care and patient-provider relationships.

Acknowledgments

We used the generative artificial intelligence (AI) tool ChatGPT to create the conversations used as scenario stimuli. These were reviewed and edited by the authors. We also used the generative AI-powered literature search tool, Elicit, in addition to Google Scholar, for finding studies relevant to our current paper.

Funding

Participant recruitment for this study was supported by start-up funds provided to ARL by the Walter Cronkite School of Journalism and Mass Communication at Arizona State University (ASU). NV's work on this paper at ASU was supported by a postdoctoral fellowship in AI and Society at the Julie Ann Wrigley Global Futures Laboratory.

Data Availability

The datasets generated or analyzed during this study are available in the Open Science Framework repository [52].

Authors' Contributions

Conceptualization: ARL (lead), NV (supporting), AK (supporting)
Data curation: ARL
Formal analysis: ARL
Funding acquisition: ARL

XSL•FO
RenderX

Investigation: ARL (lead), NV (supporting)
Methodology: ARL (lead), NV (supporting), AK (supporting)
Project administration: ARL
Supervision: ARL
Validation: ARL
Visualization: ARL
Writing – original draft: ARL (lead), NV (supporting)
Writing – review & editing: ARL (lead), NV (supporting), AK (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Source descriptions and scenario text.
[PDF File, 189 KB - jmir_v28i1e97882_app1.pdf]

Multimedia Appendix 2
Summary of hypotheses and findings.
[PDF File, 90 KB - jmir_v28i1e97882_app2.pdf]

Checklist 1
CHERRIES checklist.
[PDF File, 150 KB - jmir_v28i1e97882_app3.pdf]

References

1. Chatterji A, Cunningham T, Deming DJ, et al. How people use ChatGPT. : National Bureau of Economic Research; 2025 NBER Working Paper 34255 URL: https://www.nber.org/system/files/working_papers/w34255/w34255.pdf [accessed 2026-05-29]
2. Mendel T, Singh N, Mann DM, Wiesenfeld B, Nov O. Laypeople's use of and attitudes toward large language models and search engines for health queries: survey study. J Med Internet Res 2025 Feb 13;27(1):e64290. [doi: 10.2196/64290] [Medline: 39946180]
3. Introducing chatGPT health: a dedicated experience in chatGPT designed for health and wellness. OpenAI. URL: https://openai.com/index/introducing-chatgpt-health [accessed 2026-01-07]
4. Rafi MM, Krishnamurthy A, Balu A. Trustworthy LLM-mediated communication: evaluating information fidelity in LLM as a communicator (LAAC) framework in multiple application domains. . Preprint posted online on Nov 6, 2025. [doi: 10.1109/TPS-ISA67132.2025.00078]
5. Kelly M. Situated epistemic infrastructures: a diagnostic framework for post-coherence knowledge. arXiv. Preprint posted online on Oct 17, 2025. [doi: 10.48550/arXiv.2508.04995]
6. Nan X, Zhao X, Briones R. Parental cancer beliefs and trust in health information from medical authorities as predictors of HPV vaccine acceptability. J Health Commun 2014;19(1):100-114. [doi: 10.1080/10810730.2013.811319] [Medline: 24093156]
7. Sillence E, Briggs P, Harris P, Fishwick L. A framework for understanding trust factors in web-based health advice. Int J Hum Comput Stud 2006 Aug;64(8):697-713. [doi: 10.1016/j.ijhcs.2006.02.007]
8. Zarzeczna N, Hanel PHP, Rutjens BT, Bono SA, Chen YH, Haddock G. Scientists, speak up! Source impacts trust in health advice across five countries. J Exp Psychol Appl 2024 Sep;30(3):430-441. [doi: 10.1037/xap0000500] [Medline: 37902696]
9. Hardwig J. Epistemic dependence. J Philos 1985 Jul;82(7):335. [doi: 10.2307/2026523]
10. Hardwig J. The role of trust in knowledge. J Philos 1991 Dec;88(12):693. [doi: 10.2307/2027007]
11. Wilholt T. Epistemic trust in science. Br J Philos Sci 2013 Jun 1;64(2):233-253. [doi: 10.1093/bjps/axs007]
12. Hendriks F, Kienhues D, Bromme R. Measuring laypeople's trust in experts in a digital age: the Muenster Epistemic Trustworthiness Inventory (METI). PLoS One 2015;10(10):e0139309. [doi: 10.1371/journal.pone.0139309] [Medline: 26474078]
13. Landrum AR, Eaves BS, Shafto P. Learning to trust and trusting to learn: a theoretical framework. Trends Cogn Sci 2015 Mar;19(3):109-111. [doi: 10.1016/j.tics.2014.12.007] [Medline: 25563822]
14. Pornpitakpan C. The persuasiveness of source credibility: a critical review of five decades' evidence. J Applied Social Pyschol 2004 Feb;34(2):243-281. [doi: 10.1111/j.1559-1816.2004.tb02547.x]
15. Intemann K. Science communication and public trust in science. Interdiscip Sci Rev 2023 Apr 3;48(2):350-365. [doi: 10.1080/03080188.2022.2152244]

XSL•FO
RenderX

16. Rotenberg KJ, Petrocchi S. A longitudinal investigation of trust beliefs in physicians by children with asthma and their mothers: relations with children's adherence to medical regimes and quality of life. Child Care Health Dev 2018 Nov;44(6):879-884. [doi: 10.1111/cch.12604] [Medline: 30109723]
17. Trachtenberg F, Dugan E, Hall MA. How patients' trust relates to their involvement in medical care. J Fam Pract 2005 Apr;54(4):344-352. [Medline: 15833226]
18. Popowicz DM. Doctor knows best": on the epistemic authority of the medical practitioner. Philos Med 2021;2(2). [doi: 10.5195/philmed.2021.49]
19. Thai CL, Gaysynsky A, Falisi A, Chou WS, Blake K, Hesse BW. Trust in health information sources and channels, then and now: evidence from the health information national trends survey (2005–2013). In: Hale TM, Chou WYS, Cotten SR, Khilnani A, editors. eHealth: Current Evidence, Promises, Perils, and Future Directions: Emerald Publishing Limited; 2018. [doi: 10.1108/S2050-2060201815]
20. Many in US consider AI-generated health information useful and reliable. Annenberg Public Policy Center of the University of Pennsylvania. URL: https://www.annenbergpublicpolicycenter.org/many-in-u-s-consider-ai-generated-health-information-useful-and-reliable/ [accessed 2026-05-29]
21. 2026 Edelman Trust Barometer: trust amid insularity (global report). : Edelman Trust Institute; 2026 URL: https://edl.mn/4stxw32 [accessed 2026-05-29]
22. Choung H, Seberger JS, David P. When AI is perceived to be fairer than a human: understanding perceptions of algorithmic decisions in a job application vontext. Int J Hum-Comput Interact 2024 Nov 16;40(22):7451-7468. [doi: 10.1080/10447318.2023.2266244]
23. Lu L, Tormala ZL, Duhachek A. How AI sources can increase openness to opposing views. Sci Rep 2025;15(1):17170. [doi: 10.1038/s41598-025-00791-z]
24. Liew TW, Tan SM. Social cues and implications for designing expert and competent artificial agents: a systematic review. Telemat Inform 2021 Dec;65:101721. [doi: 10.1016/j.tele.2021.101721]
25. Jalbert M, Harris M, Williams L. Who is perceived to be an expert on COVID-19 vaccines on social media?: Biomedical credentials confer expertise, even among vaccine-hesitant and Conservative observers. Inf Commun Soc 2025;28(4):669-687. [doi: 10.1080/1369118x.2024.2436001] [Medline: 40809071]
26. Thon FM, Jucks R. Believing in expertise: how authors' credentials and language use influence the credibility of online health information. Health Commun 2017 Jul;32(7):828-836. [doi: 10.1080/10410236.2016.1172296] [Medline: 27466693]
27. Alvarado R. What kind of trust does AI deserve, if any? AI Ethics 2023 Nov;3(4):1169-1183. [doi: 10.1007/s43681-022-00224-x]
28. Dlugatch R, Georgieva A, Kerasidou A. Trustworthy artificial intelligence and ethical design: public perceptions of trustworthiness of an AI-based decision-support tool in the context of intrapartum care. BMC Med Ethics 2023 Jun 20;24(1):42. [doi: 10.1186/s12910-023-00917-w] [Medline: 37340408]
29. Guidi S, Reina L, Currò F, Cipriani E, Grassini S, Parlangeli O. The influence of framing, domain and task type on trust in AI. Presented at: Proceedings of the 36th Annual Conference of the European Association of Cognitive Ergonomics (EACE); Oct 7, 2025; Tallinn Estonia p. 1-6. [doi: 10.1145/3746175.3747196]
30. Landrum AR, Lull RB, Akin H, Hasell A, Jamieson KH. Processing the papal encyclical through perceptual filters: Pope Francis, identity-protective cognition, and climate change concern. Cognition 2017 Sep;166:1-12. [doi: 10.1016/j.cognition.2017.05.015] [Medline: 28549233]
31. Kunda Z. The case for motivated reasoning. Psychol Bull 1990 Nov;108(3):480-498. [doi: 10.1037/0033-2909.108.3.480] [Medline: 2270237]
32. Petty RE, Cacioppo JT. The elaboration likelihood model of persuasion. Adv Exp Soc Psychol 1986;19:123-205. [doi: 10.1016/S0065-2601(08)60214-2]
33. Chaiken S. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. J Pers Soc Psychol 1980;39(5):752-766. [doi: 10.1037/0022-3514.39.5.752]
34. Finucane ML, Alhakami A, Slovic P, Johnson SM. The affect heuristic in judgments of risks and benefits. J Behav Decis Making 2000 Jan;13(1):1-17. [doi: 10.1002/(SICI)1099-0771(200001/03)13:1<1::AID-BDM333>3.0.CO;2-S]
35. Rimal RN. Perceived risk and self-efficacy as motivators: understanding individuals' long-term use of health information. J Commun 2001 Dec 1;51(4):633-654. [doi: 10.1111/j.1460-2466.2001.tb02900.x]
36. Slovic P, Finucane ML, Peters E, MacGregor DG. Risk as analysis and risk as feelings: some thoughts about affect, reason, risk, and rationality. Risk Anal 2004 Apr;24(2):311-322. [doi: 10.1111/j.0272-4332.2004.00433.x] [Medline: 15078302]
37. Yang JZ. Whose risk? Why did the U.S. public ignore information about the Ebola outbreak? Risk Anal 2019 Aug;39(8):1708-1722. [doi: 10.1111/risa.13282] [Medline: 30726559]
38. Wilson RS, Arvai JL. When less is more: how affect influences preferences when comparing low and high‐risk options. J Risk Res 2006 Mar;9(2):165-178. [doi: 10.1080/13669870500419503]
39. de Hoog N, Stroebe W, de Wit JBF. The processing of fear‐arousing communications: how biased processing leads to persuasion. Soc Influ 2008 Jun;3(2):84-113. [doi: 10.1080/15534510802185836]
40. Wedderhoff O, Chasiotis A, Rosman T. When freedom of choice leads to bias: how threat fosters selective exposure to health information. Front Psychol 2022;13:937699. [doi: 10.3389/fpsyg.2022.937699] [Medline: 36312070]

XSL•FO
RenderX

41. Chang DS, Kang OS, Kim HH, et al. Pre-existing beliefs and expectations influence judgments of novel health information. J Health Psychol 2012 Jul;17(5):753-763. [doi: 10.1177/1359105311421044] [Medline: 21997377]
42. Kahan DM, Braman D, Cohen GL, Gastil J, Slovic P. Who fears the HPV vaccine, who doesn't, and why? an experimental study of the mechanisms of cultural cognition. Law Hum Behav 2010 Dec;34(6):501-516. [doi: 10.1007/s10979-009-9201-0] [Medline: 20076997]
43. Kahan DM, Jamieson KH, Landrum A, Winneg K. Culturally antagonistic memes and the Zika virus: an experimental test. J Risk Res 2017 Jan 2;20(1):1-40. [doi: 10.1080/13669877.2016.1260631]
44. Kim JN, Oh YW, Krishna A. Justificatory information forefending in digital age: self-sealing informational conviction of risky health behavior. Health Commun 2018 Jan;33(1):85-93. [doi: 10.1080/10410236.2016.1242040] [Medline: 27892700]
45. Meppelink CS, Smit EG, Fransen ML, Diviani N. "I was Right about Vaccination": confirmation bias and health literacy in online health information seeking. J Health Commun 2019;24(2):129-140. [doi: 10.1080/10810730.2019.1583701] [Medline: 30895889]
46. Claudy MC, Aquino K, Graso M. Artificial intelligence can't be charmed: the effects of impartiality on laypeople's algorithmic preferences. Front Psychol 2022;13:898027. [doi: 10.3389/fpsyg.2022.898027] [Medline: 35846643]
47. Jago AS, Laurin K. Assumptions about algorithms' capacity for discrimination. Pers Soc Psychol Bull 2022 Apr;48(4):582-595. [doi: 10.1177/01461672211016187] [Medline: 34044648]
48. Fiscella K, Franks P, Clancy CM. Skepticism toward medical care and health care utilization. Med Care 1998 Feb;36(2):180-189. [doi: 10.1097/00005650-199802000-00007] [Medline: 9475472]
49. Levy N. Due deference to denialism: explaining ordinary people's rejection of established scientific findings. Synthese 2019;196(1):313-327. [doi: 10.1007/s11229-017-1477-x] [Medline: 30713358]
50. Stavri PZ, Freeman DJ, Burroughs CM. Perception of quality and trustworthiness of Internet resources by personal health information seekers. AMIA Annu Symp Proc 2003;2003:629-633. [Medline: 14728249]
51. Reis M, Reis F, Kunde W. Influence of believed AI involvement on the perception of digital medical advice. Nat Med 2024 Nov;30(11):3098-3100. [doi: 10.1038/s41591-024-03180-7] [Medline: 39054373]
52. Eysenbach G. Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
53. Landrum AR, Verma N, Kehrberg A. Preregistration: asking generative AI health questions. OSF. 2025 Jul 14. URL: https://osf.io/ze4yd/overview [accessed 2026-05-30]
54. Landrum AR, Verma N, Kehrberg A. Taking health advice from generative AI. OSF. URL: https://osf.io/a4k5u/overview [accessed 2026-05-30]
55. McCroskey JC, Teven JJ. Goodwill: a reexamination of the construct and its measurement. Commun Monogr 1999 Mar;66(1):90-103. [doi: 10.1080/03637759909376464]
56. Samejima F. Graded response models. In: Handbook of Item Response Theory: Chapman and Hall/CRC; 2016:95-107. [doi: 10.1201/9781315374512]
57. Greussing E, Guenther L, Baram-Tsabari A, et al. The perception and use of generative AI for science-related information search: insights from a cross-national study. Public Underst Sci 2025 Jul;34(5):599-615. [doi: 10.1177/09636625241308493] [Medline: 39891412]
58. Shekar S, Pataranutaporn P, Sarabu C, Cecchi GA, Maes P. People overtrust AI-generated medical advice despite low accuracy. NEJM AI 2025 May 22;2(6):AIoa2300015. [doi: 10.1056/AIoa2300015]
59. Qin H, Zhu Y, Jiang Y, Luo S, Huang C. Examining the impact of personalization and carefulness in AI-generated health advice: trust, adoption, and insights in online healthcare consultations experiments. Technol Soc 2024 Dec;79:102726. [doi: 10.1016/j.techsoc.2024.102726]
60. Lee ST, Lin J. A self-determination perspective on online health information seeking: the internet vs. face-to-face office visits with physicians. J Health Commun 2016 Jun;21(6):714-722. [doi: 10.1080/10810730.2016.1157651] [Medline: 27186966]
61. Rains SA. Seeking health information in the information age: the role of internet self-efficacy. West J Commun 2008 Feb 21;72(1):1-18. [doi: 10.1080/10570310701827612]
62. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9. [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
63. Deci EL, Ryan RM. Self-determination theory. In: AW K, ET H, editors. Handbook of Theories of Social Psychology: Sage Publications Ltd; 2012:416-436. [doi: 10.4135/9781446249215.n21]

Abbreviations

AI: artificial intelligence
AI nurse: health care–specialized generative artificial intelligence system framed as an "AI nurse"
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
emmean: estimated marginal mean
M_{bene}: mean benevolence score

XSL•FO
RenderX

M_{comp}: mean competence score
RQ1: research question 1

Edited by I Steenstra; submitted 10.Apr.2026; peer-reviewed by M Reis, W Evans; revised version received 08.May.2026; accepted 15.May.2026; published 22.Jun.2026.

Please cite as:
Landrum AR, Verma N, Kehrberg A
Trusting Generative AI for Health Advice: Preregistered Survey Experiment
J Med Internet Res 2026;28:e97882
URL: https://www.jmir.org/2026/1/e97882
doi:10.2196/97882

© Asheley R Landrum, Nitin Verma, Amanda Kehrberg. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Large Language Model Summarization of Physician-to-Physician Calls for Interhospital Transfer of Patients With ST-Elevation Myocardial Infarction: Observational Study

Jesse O Wrenn[1,2,3], MD, PhD; Madelaine Behrens[4], MD; Mary S Hershey[4], MD; Marc Maldaver[4], MD; John Mitchell[4], MD; Trevor Thompson[4], MD; Austin J Triana[4], MBA, MD; Zain M Virk[4], MD; Yasemin Akdas[5], MPH, PhD; Michael R Cauley[2], PhD, DMin; Michael J Ward[1,2,3,6*], MBA, MD, PhD; Ken Monahan[7*], MD

[1]Department of Emergency Medicine, Vanderbilt University Medical Center, 2215 Garland Ave, Light Hall Ste 203, Nashville, TN, United States
[2]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States
[3]Division of Emergency Medicine, VA Tennessee Valley Healthcare System, Nashville, TN, United States
[4]Department of Internal Medicine, Vanderbilt University Medical Center, Nashville, TN, United States
[5]Office of Community Health and Engagement, Vanderbilt University Medical Center, Nashville, TN, United States
[6]Geriatric Research, Education, and Clinical Center, VA Tennessee Valley Healthcare System, Nashville, TN, United States
[7]Division of Cardiovascular Medicine, Vanderbilt University Medical Center, Nashville, TN, United States
[*]these authors contributed equally

Corresponding Author:
Jesse O Wrenn, MD, PhD
Department of Emergency Medicine, Vanderbilt University Medical Center, 2215 Garland Ave, Light Hall Ste 203, Nashville, TN, United States

Abstract

Background: Interhospital transfer of patients with suspected ST-elevation myocardial infarction (STEMI) requires timely and robust communication. Clinical uptake of potentially useful information from physician-to-physician phone calls authorizing transfer is low at many institutions, at least in part due to relative inaccessibility of call audio and lack of transcripts or summaries. Large language models (LLMs) can transform text into brief, consistently formatted summaries that could be made available in the electronic health record, thus facilitating the timely availability of clinically relevant data to physicians downstream of the transfer call.

Objective: We sought to assess the feasibility of using transcription and LLM summarization to provide written information summarizing transfer calls by adapting the Physician Documentation Quality Instrument (PDQI) to score generated call summaries and evaluate whether LLMs could effectively summarize a curated set of transfer calls.

Methods: STEMI transfer calls for which our institution was the receiving facility were transcribed and summarized by Whisper and ChatGPT (OpenAI), respectively. Each summary was reviewed by 2 of 7 independent physician raters. Summaries were rated using a Likert scale applied to an 8-domain framework adapted from the PDQI. We calculated summary statistics, including means, SDs, and raw and weighted agreement, and produced visual radar plots to demonstrate ratings for each call. We also performed thematic analysis of reviewers' comments.

Results: We identified 32 calls, of which 1 (3.1%) was excluded for incompleteness. Raw agreement between raters was 62% (153/248), and the mean of the pairwise weighted κ coefficients was 0.19 (SD 0.30; slight agreement). The mean rating of all summaries across all domains was 4.6 of 5 (SD 0.7). The "useful" (mean 4.8/5, SD 0.5) and "consistent" (mean 4.9/5, SD 0.6) domains were the highest rated, and the "thorough" (mean 4.4/5, SD 1.0) and "hallucination free" (mean 4.4/5, SD 0.9) domains were the lowest rated. The mean score for accuracy was 4.6/5 (SD 0.7). Qualitative analysis found that raters penalized the LLM for inferential hallucinations, although these were often clinically accurate, and discrepancies related to calculation of timing of events.

Conclusions: Despite the limitations inherent in a small pilot cohort, this feasibility study suggests that LLMs can generate accurate and pertinent summaries of interhospital transfer calls for patients with STEMI. Interrater agreement was slight, which may suggest inadequate training of raters, unclear definitions, or a limitation of using the PDQI for this task. We identified several important areas for consideration prior to implementation, including thorough assessment of transcription accuracy, prompt engineering to minimize unwanted LLM behavior, and assessment of the impact of incorporating these summaries into clinical care on clinical outcomes.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e88834) doi:10.2196/88834

KEYWORDS

ST-elevation myocardial infarction; STEMI; patient transfer; large language models; summarization; delivery of health care; artificial intelligence; AI

Introduction

ST-elevation myocardial infarction (STEMI) is a cardiovascular emergency that requires prompt access to the recommended reperfusion treatment, primary percutaneous coronary intervention (PCI), to achieve optimal outcomes. Approximately 61% of hospitals in the United States lack PCI capabilities [1], requiring transfer of up to 50% of patients with STEMI [2-4]. Shorter time spent at the transferring hospital is significantly associated with shorter door-to-balloon times and improved mortality rates [5]. The quality of communication between referring and receiving institutions has been identified as a major barrier to effective, timely transfer and, thus, may be a potential target for intervention [6].

At Vanderbilt University Medical Center (VUMC) in middle Tennessee, patients are accepted for transfer during phone calls between referring physicians and VUMC cardiologists. Although digitally recorded, phone calls with *accepting* physicians, who make the decision to accept the patient for transfer, are not readily available for review by *receiving* physicians, who ultimately care for the patient on arrival and may differ from the accepting physician. In addition, the *accepting* physician may not have the opportunity to communicate directly with the *receiving* physician, leading to further degradation of information transfer. Thus, a durable near–primary source account of the clinical details surrounding the transfer would assist in adding an element of continuity across the passage of time and transitions to different physicians.

Large language models (LLMs) may offer potential solutions by summarizing transfer calls into brief, consistently formatted, digestible text that could be made available in the electronic health record. There has been extensive study of digital scribes, or summarization of recorded conversations between physicians and patients, but less work in the domain of phone call transcription and summarization [7-9]. A recent study demonstrates successful summarization of phone calls related to emergency department transfer; however, the process included intermediaries that corrected transcription mistakes and performed manual speaker diarization and labeling [10], steps that may not be practically implemented in particularly time-sensitive conditions such as STEMI and may also not be scalable due to personnel and cost requirements. We sought to assess the feasibility of transcription and LLM summarization of STEMI transfer calls without human intervention using the Physician Documentation Quality Instrument (PDQI) [11], adapted subsequently for evaluation of LLM-generated text [12].

Methods

Ethical Considerations

This project was approved by our institutional review board with a quality improvement or nonresearch determination (240725). Therefore, individual consent was not required. All identifiable data were stored on encrypted devices within the secure institutional infrastructure. Participants were not compensated for inclusion.

Data Acquisition

Our institution maintains a log of all patients admitted, transferred, or diagnosed in-house with STEMI. We used this log to identify all patients transferred to VUMC from an outside institution between January 1 and June 30, 2024, for whom there was a clinical concern for STEMI. The recordings of the transfer calls between associated referring physicians and accepting cardiologists were downloaded from an internal audio database (Encore; version 8.6; DVSAnalytics) and reviewed by a single cardiologist (KM) to ensure that they contained both the clinical information upon which the presumptive diagnosis of STEMI was made and the key decision-making discussion that informed the transfer.

Transcription and Summarization

Each call was then transcribed using the OpenAI Whisper model with English language and large model size preselected, hosted locally within the secure institutional infrastructure. Transcripts were neither preprocessed nor cleaned prior to LLM analysis. Six transcripts were selected at random for assessment of word error rate (WER). Subtle surname misspellings were not counted as errors, and filler words were excluded from analysis. These verbatim transcripts were analyzed using aiChat, VUMC's HIPAA (Health Insurance Portability and Accountability Act)-compliant GPT-4o–powered LLM. This software was chosen because it is readily available at VUMC and HIPAA compliant. Although prompt engineering [13], the systematic process of drafting clear instructions for an LLM, can optimize performance, this was outside the scope of our project. The following prompt alone, based on prior qualitative work without any additional specific formatting requests, was used to generate each summary [14]:

> *Please generate the following for the following conversation:*
>
> *"Summary: (provide the best possible summary according to my request)*
>
> *Medications administered: (provide a list of all medications administered thus far)*
>
> *Time: (provide the number of minutes since the chest pain started)*

XSL•FO
RenderX

EKG findings: (provide the EKG findings related to the STEMI)

Labs: (provide any labs that have returned)

Vitals: (provide any vitals or information about hemodynamic stability)"

Quality Measurement

Audio and aiChat summaries were uploaded to REDCap (Research Electronic Data Capture; Vanderbilt University) [15,16] alongside a modified PDQI tool. Modifications to the PDQI included removal of the "up-to-date" and "synthesized" domains, which were deemed not applicable to this project. The adapted PDQI also included a question about whether the summary was "free from hallucination," defined for the purpose of this study as containing information not verifiable in the call audio. This definition has been used previously for analysis of artificial intelligence (AI) tools [12]. Three transfer calls from December 2022 were used as a pilot dataset to facilitate familiarity with the rating mechanism and identify ambiguities and/or problems with workflow. To measure the quality of transfer call summaries, each case was assigned to 2 of 7 participating second- and third-year internal medicine resident physicians. Raters were directly queried via email regarding their rotation history, which was verified by one author (KM) against the central schedule posted on the medical center's intranet. They provided the number of cardiac intensive care unit (CICU) rotations and non-CICU cardiology rotations they had completed at that point in their training. The same pair of raters evaluated no more than 2 transfer calls and were blinded to each other's scores but not to the study hypothesis. Each rater independently listened to the raw transfer call audio and rated the LLM-generated summary on each PDQI domain from 1 ("not at all") to 5 ("completely"). Raters did not have access to the transcript. Assignments were made to maximally distribute rater pairings. Raters were given an opportunity to leave narrative comments to describe hallucinations, missing information, or other interesting findings.

Statistical and Qualitative Analysis

We calculated and reported raw interrater reliability and pairwise weighted κ. Interrater agreement was presented graphically as well as using radar charts. In addition, we reported mean and SD values for each measured domain. We reviewed rater comments from each of the measurements and conducted a thematic analysis. We also reviewed comments from a convenience sample of the 5 summaries with the largest difference between the two assigned raters' scores across all domains and calculated the differences between raters in each domain.

Results

Quantitative Results

We identified 32 transfer calls. Of these 32 calls, 1 (3.1%) was excluded from analysis due to the recording being incomplete, leaving a sample size of 31 (96.9%). The average WER for 6 evaluated calls was 1.6% (SD 0.4%), with a 1.4% (70/4865) total WER across the 6 sampled calls. Each call summary was rated on the 8 PDQI domains by 2 raters for a total of 496 ratings or 248 total pairs. At the time of review, the 7 residents rating the summaries had completed an average of 2.4 (SD 1.7) CICU rotations and 3.4 (SD 0.9) non-CICU cardiology rotations. Each rater evaluated 8 or 9 summaries. Of the paired ratings, 62% (153/248) matched exactly between both raters, pairwise weighted κ coefficients ranged from −0.23 (no agreement) to 1 (perfect agreement), and the mean of the pairwise weighted κ coefficients was 0.19 (SD 0.30; slight agreement). Figure 1 shows radar plots of the raters' scores for each reviewed summary document and displays visually the overall degree of scoring overlap among the rater pairs.

Table 1 provides a summary of ratings by PDQI domain. No summary was rated perfectly by both raters, but there was perfect agreement between raters on summaries 6 and 27. For each of the 8 domains, the mean score exceeded 4 on the 5-point Likert scale. The mean score for all ratings across all domains was 4.6 (SD 0.7).

XSL•FO
RenderX

Figure 1. Radar plot representation of interrater variability for each call summary. For a given radar plot, each vertex of the octagon represents one of the adapted Physician Documentation Quality Instrument domains, as depicted in the legend. For each summary, one rater's scores are shaded in red, and the other rater's scores are shaded in blue. The overlap between the 2 reviewers is indicated in purple. For example, raters agreed perfectly on the summaries of calls 6 and 27 and disagreed substantially on the summaries of calls 5, 14, 19, 25, and 31.

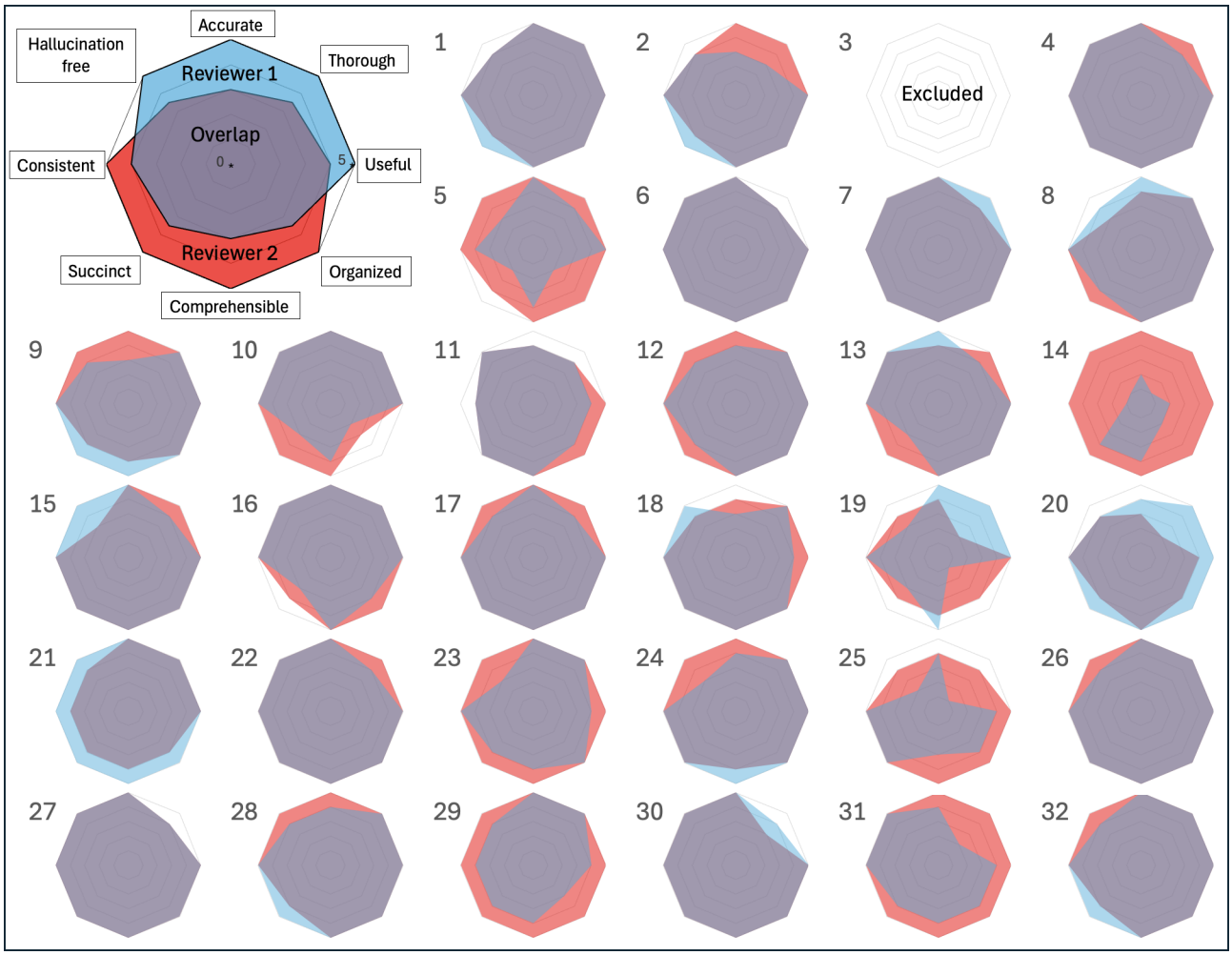

Table . Summary statistics by Physician Documentation Quality Instrument domain.

Domain	Score (1-5), mean (SD)
Accurate	4.6 (0.7)
Thorough	4.4 (1.0)
Useful	4.8 (0.5)
Organized	4.6 (0.9)
Comprehensible	4.8 (0.4)
Succinct	4.6 (0.7)
Consistent	4.9 (0.6)
Hallucination free	4.4 (0.9)[a]

[a]Clinically valid inferences by the large language model were penalized by raters as hallucinations as the inferences were not directly verifiable in the call audio, so the mean score in the "hallucination free" domain should be interpreted with caution.

Qualitative Results

The descriptive comments by raters identified hallucinations, which often took the form of inferences made by the LLM. For multiple patients, the LLM inferred a patient's hemodynamic stability from the vital signs provided or from a description of the patient's appearance as pale or diaphoretic. Although these inferences were not reported as inaccurate, they were not explicitly mentioned in the calls, which prompted the raters to label them as hallucinations. For example, raters noted that the LLM inferred hepatic encephalopathy as a diagnosis (unstated in the call) from the combination of altered mental status and a history of cirrhosis (both stated in the call).

XSL•FO
RenderX

Noninferential inaccuracies were uncommon. However, on several occasions, when the referring physician voiced that a medication was considered or ready to be given, it was cited as given by the LLM. Vital signs were mentioned as misreported once. The raters described more instances of omission of details that they believed would have been useful to the receiving team, including laboratory tests, relevant history, and examination findings. However, raters often praised the summaries for filtering out parts of the conversation that they judged as not important to the receiving team.

Raters identified that the LLM appeared to have difficulty with both absolute and relative time. In total, 6.5% (2/31) of the summaries had events listed in the wrong sequence. Several summaries (7/31, 23%) contained inferences regarding the time of a call that was not explicitly mentioned, both absolute clock time and with respect to the time since the chest pain started, the latter of which was explicitly requested by the prompt.

Interrater agreement calculated using the mean of the pairwise weighted κ coefficients was slight at 0.19 (SD 0.30). A total of 16.1% (5/31) of the summaries reviewed (5, 14, 19, 25, and 31) had notably worse agreement than the others. One specific rater was found to have given the lower average rating in 4 of these 5 summaries and was also a rater on the fifth (summary 19), which had the same average rating between the 2 raters but substantial differences in domain scores. This rater was very thorough with their comments, and most often referred to missing information that the rater believed could alter management, inaccuracies regarding assessment of timing, and inferential hallucinations. In these 5 calls, differences between ratings were the largest in the "thorough," "hallucination free," and "organized" domains.

Discussion

Principal Findings

Overview

This feasibility study explored the use of LLMs to summarize phone calls between referring and accepting physicians in the setting of emergency care for patients with suspected STEMI and evaluated this process using standardized assessment tools. With the caveat that mean scores were derived from ratings with only slight agreement, the high (adapted) PDQI score suggests that the LLM-based approach yielded summaries that were accurate and reproducible. Transfer call summaries, particularly when compared to complete opacity, may represent a transformative solution to address data loss during interhospital transfers. This work demonstrates the feasibility of using an LLM to summarize transcripts of STEMI transfer calls and provides a foundation for further investigation of automated summarization of call audio in patient transfers.

Call Summary Quality

No PDQI domain received ratings under 4/5 when averaged over the summaries. The "thorough" and "hallucination free" domains had the lowest scores of the 8 measured domains. Whether a lower score in the "thorough" domain represents the LLM system behaving as designed given inherent limitations of how thorough a summary can be or reflects true information loss is unclear as this was not addressed in the comments left by raters. The distinction is important as redesigning the prompt could generate more thorough, albeit potentially longer summaries. However, the consistently high ratings for the "useful" domain are reassuring and suggest that the lower scores in the "thorough" domain are less likely to be due to diminished transfer of high-yield information.

We found that there were 2 types of hallucinations referred to in the comments regarding the "hallucination free" domain. The first was fabrication, which could clearly cause harm if patients' management was changed due to inaccurate information. The second was clinical inference (such as inferring hemodynamic instability from words such as "pale" or "diaphoretic"), which was considered reasonable and valid by reviewers but was still labeled as a hallucination because the information was not explicitly verifiable via the call audio. In this initial study, the "hallucination free" rating could represent the worst-case scenario or upper bound, but future work might benefit from parsing further the definition of the hallucination domain by including fabrication and valid inference as 2 separate domains.

Although relatively uncommon, fabrications could potentially adversely influence patient care, particularly if there are no associated primary data included in the summary from which receiving physicians can draw their own conclusions (ie, a receiving physician may interpret vital signs differently from the LLM, but there is no analogous corroborating evidence to support whether a medication was given). Furthermore, if the LLM reported a categorical assessment of a variable such as vital signs and the primary data were not included in the summary, a miscategorized vital sign may prompt misclassification. However, there are strategies for mitigation of LLM hallucinations that were not within the scope of this project [17].

That the LLM had difficulty with the concept of time since the chest pain started is not surprising. Although LLMs have recently achieved significant advances in mathematical reasoning, this remains a challenging frontier in AI research [18]. The prompt we designed requested the number of minutes since the chest pain started, which may require knowledge of the time of the call, the time of onset of chest pain, and the ability to reason mathematically. An alteration of the prompt to request any relevant information from the call on the time when symptoms started as opposed to "the number of minutes since chest pain started" might improve the summary's accuracy with respect to the timing of relevant events.

There are several additional issues that are important to consider prior to implementation. The average WER was 1.6% (SD 0.4%) for 6 randomly selected transcripts, which is reassuring; however, we did not thoroughly assess the conceptual accuracy of all transcripts, nor did we analyze common errors of transcription. Propagated errors of transcription may have degraded the quality of the summary. Assurance that the transcription model includes medical terminology and drug names is important, as is thorough assessment of transcription quality. Further prompt engineering may reduce the incidence of hallucinations, particularly with respect to the timing of

XSL•FO
RenderX

events, including the onset of chest pain. For example, updating the prompt to request only timing that was explicitly discussed in the call may be helpful. Finally, to engender confidence, it may be reasonable to request quotations from the transcript so that physicians can look back to the primary data to confirm critical information.

Limitations

There were several limitations to this study. The transcription software and LLM were chosen for convenience by the study team, and alternative software or alternate versions could produce different results. The sample size was small, which could limit overall generalizability, although the patient population was relatively homogenous (ie, patients being transferred to a tertiary care center due to concern about STEMI). The selection process included information-rich calls that contained complete clinical and decision-making information, which introduced selection bias. Performance on a randomly selected sample of calls would improve the generalizability of the study to the broader clinical environment, although using an enriched dataset for this feasibility study is a reasonable initial step.

Lack of a comparator such as human expert–generated summaries limits interpretability of the summary quality scores, and the study's retrospective design limits our ability to understand the clinical adequacy of the summaries and downstream effects of summarization on patient care. A prospective study measuring the impact of incorporation of summaries into clinical workflow on patient outcomes (such as morbidity, mortality, time in the intensive care unit, and overall length of stay) compared to an expert-generated comparator could address some of these design limitations.

Although we did not test different prompts, the prompt used in this study was engineered to optimize the inclusion of key objective data while generating a summary that could be consumed efficiently by a busy clinician. Future work could address this limitation by having receiving physicians evaluate the overall utility of summaries generated using different prompts.

The second- and third-year internal medicine residents reviewing the summaries did not evaluate them from the perspective of interventional or critical care cardiologists; however, they had enough exposure to CICU and general cardiology rotations in their training to reasonably make assessments of summary quality. Furthermore, they are often the intended beneficiaries of call summaries (ie, the first-line recipients of post-PCI patients with STEMI). That said, their relative inexperience could lead to different assessments of the importance of certain elements of the call from those of a more experienced clinician, particularly for the "useful" and "thorough" domains, each of which relies on the communication of information deemed important. Although blinded to the scores of others, raters were not blinded to the LLM-generated source of the summaries, which could introduce rater bias and artificially inflate or deflate scores. They were also not blinded to the study hypothesis, which could introduce confirmation bias as raters could subconsciously inflate scores to align with the goals of the study. Although infrequent, there were a few substantial disagreements between raters on the quality of summaries, with slight weighted agreement. The reported mean scores for each of the domains must be interpreted with caution given discordant ratings. It is also important to note that the observed slight agreement may partially reflect a statistical ceiling effect as 74.8% (371/496) of the total ratings were 5/5, and κ can underestimate agreement when ratings cluster within a narrow range and chance agreement is inflated. It is unclear whether the remaining disagreement represents inadequate training of raters, unclear definitions, or a limitation of using the PDQI for this task. That one particular rater often provided lower ratings than their counterparts may reflect lenience on the part of the other raters, particularly in light of the possibility of confirmation bias resulting from the fact that raters were unblinded to the study hypothesis. Prior to deployment in clinical practice, in addition to larger transfer call sample sizes and proportionally larger numbers of summary raters, follow-up studies could include additional raters per summary to more robustly characterize the frequency and magnitude of outliers in the domain of interrater disparities.

The original 9-item PDQI (PDQI-9) was developed and validated for the evaluation of notes across 9 criteria. At the conception of this study, we identified an unvalidated adaptation of the PDQI-9 that assessed the quality of LLM-generated text. Since then, there has been a validated evaluation of similar adaptations of the PDQI-9 for LLM-generated text that include the addition of the "hallucination free" domain and the removal of the "up-to-date" domain [19]. There has also been validation of a similar adapted PDQI-9 [20] for LLM-generated summaries that removed the "up-to-date" and "consistent" domains and included the assessment of hallucination in the "accuracy" domain. Notably, the latter defined hallucinations as falsifications (information or data distorted from the original note) or fabrications (made-up information or data that could be plausible but based on nonexistent facts) present in the summary. Regardless, the use of an unvalidated altered instrument may undermine the reliability of the scores to measure true summary quality.

To make this type of automated call summary tool usable in day-to-day clinical practice, the individual components (ie, call transcription and LLM analysis) need to be integrated such that summary generation is transparent to the summary recipient. In addition, the summary would need to be uploaded to the patient's chart seamlessly, ideally without any incremental effort from the accepting or receiving clinician. Although implementation would require IT support, the use of ambient AI to document outpatient clinic visits [21,22] suggests that this objective is readily achievable in the tertiary medical center environment.

Conclusions

In a feasibility study, we demonstrated that transcription and LLM summarization can provide information about the content of STEMI transfer calls between institutions and, thus, leveraged primary data that are routinely collected but not optimally used. The use of an adapted PDQI has notable limitations that must be addressed to more accurately assess summary quality, and

XSL•FO
RenderX

mean scores must be interpreted cautiously given slight interrater agreement; however, on a subset of curated information-rich transfer calls, summaries were demonstrated to be useful and accurate. Given the risks associated with physician handoffs [23], our results emphasize LLMs' potential as a transformative tool to address data loss during interhospital transfers, particularly when compared to the current alternative of total information loss.

Acknowledgments

The authors would like to thank Wendy Daigle, RN, MSN, CCRN, for providing the ST-elevation myocardial infarction log and Dana Wilson, RN, MBA, MMHC, NEA-BC, for facilitating access to the digital phone call database.

Funding

JOW and MRC were supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under award T32HL170986.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the impracticality of redacting protected health information from audio files, transcripts, and summary documents but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: JOW, YA, MRC, KM, MJW
Data curation: JOW, MB, MSH, MM, JM, TT, AJT, ZMV
Formal analysis: JOW
Supervision: KM, MJW
Writing—original draft: JOW
Writing—review and editing: JOW, MB, MSH, MM, JM, TT, AJT, ZMV, YA, MRC, KM, MJW

Conflicts of Interest

None declared.

References

1. Langabeer JR, Henry TD, Kereiakes DJ, et al. Growth in percutaneous coronary intervention capacity relative to population and disease prevalence. J Am Heart Assoc 2013 Oct 28;2(6):e000370. [doi: 10.1161/JAHA.113.000370] [Medline: 24166491]
2. Chakrabarti A, Krumholz HM, Wang Y, Rumsfeld JS, Nallamothu BK, National Cardiovascular Data Registry. Time-to-reperfusion in patients undergoing interhospital transfer for primary percutaneous coronary intervention in the US: an analysis of 2005 and 2006 data from the National Cardiovascular Data Registry. J Am Coll Cardiol 2008 Jun 24;51(25):2442-2443. [doi: 10.1016/j.jacc.2008.02.071] [Medline: 18565404]
3. Jollis JG, Al-Khalidi HR, Monk L, et al. Expansion of a regional ST-segment-elevation myocardial infarction system to an entire state. Circulation 2012 Jul 10;126(2):189-195. [doi: 10.1161/CIRCULATIONAHA.111.068049] [Medline: 22665718]
4. Jollis JG, Roettig ML, Aluko AO, et al. Implementation of a statewide system for coronary reperfusion for ST-segment elevation myocardial infarction. JAMA 2007 Nov 28;298(20):2371-2380. [doi: 10.1001/jama.298.20.joc70124] [Medline: 17982184]
5. Wang TY, Nallamothu BK, Krumholz HM, et al. Association of door-in to door-out time with reperfusion delays and outcomes among patients transferred for primary percutaneous coronary intervention. JAMA 2011 Jun 22;305(24):2540-2547. [doi: 10.1001/jama.2011.862] [Medline: 21693742]
6. Ward MJ, Vogus TJ, Bonnet K, Moser K, Schlundt D, Kripalani S. Breaking down walls: a qualitative evaluation of perceived emergency department delays for patients transferred with ST-elevation myocardial infarction. BMC Emerg Med 2020 Aug 6;20(1):60. [doi: 10.1186/s12873-020-00355-6] [Medline: 32762657]
7. Ha E, Choon-Kon-Yune I, Murray L, et al. Evaluating the usability, technical performance, and accuracy of artificial intelligence scribes for primary care: competitive analysis. JMIR Hum Factors 2025 Jul 23;12:e71434. [doi: 10.2196/71434] [Medline: 40700466]
8. van Buchem MM, Boosman H, Bauer MP, Kant IM, Cammel SA, Steyerberg EW. The digital scribe in clinical practice: a scoping review and research agenda. NPJ Digit Med 2021 Mar 26;4(1):57. [doi: 10.1038/s41746-021-00432-5] [Medline: 33772070]

XSL•FO
RenderX

9. Duggan MJ, Gervase J, Schoenbaum A, et al. Clinician experiences with ambient scribe technology to assist with documentation burden and efficiency. JAMA Netw Open 2025 Feb 3;8(2):e2460637. [doi: 10.1001/jamanetworkopen.2024.60637] [Medline: 39969880]
10. Sezgin E, Sirrianni JW, Kranz K. Evaluation of a digital scribe: conversation summarization for emergency department consultation calls. Appl Clin Inform 2024 May 15;15(3):600-611. [doi: 10.1055/a-2327-4121] [Medline: 38749477]
11. Stetson PD, Bakken S, Wrenn JO, Siegler EL. Assessing electronic note quality using the Physician Documentation Quality Instrument (PDQI-9). Appl Clin Inform 2012;3(2):164-174. [doi: 10.4338/aci-2011-11-ra-0070] [Medline: 22577483]
12. Tierney AA, Gayre G, Hoberman B, et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. NEJM Catal Innov Care Deliv 2024 Feb 21;5(3). [doi: 10.1056/CAT.23.0404]
13. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv 2023;55(9):1-35. [doi: 10.1145/3560815]
14. Ward MJ, Kripalani S, Muñoz D, et al. Association of physician coordination with interfacility transfer acceptance timeliness. Am J Accountable Care 2022 Sep;10(3):7-15. [doi: 10.37765/ajac.2022.89231] [Medline: 38617098]
15. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
16. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
17. Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst 2025;43(2):1-55. [doi: 10.1145/3703155]
18. Wang PY, Liu TS, Wang C, et al. A survey on large language models for mathematical reasoning. ACM Comput Surv 2026;58(8):1-35. [doi: 10.1145/3786333]
19. Palm E, Manikantan A, Mahal H, Belwadi SS, Pepin ME. Assessing the quality of AI-generated clinical notes: validated evaluation of a large language model ambient scribe. Front Artif Intell 2025;8:1691499. [doi: 10.3389/frai.2025.1691499] [Medline: 41199808]
20. Croxford E, Gao Y, Pellegrino N, et al. Development and validation of the provider documentation summarization quality instrument for large language models. J Am Med Inform Assoc 2025 Jun 1;32(6):1050-1060. [doi: 10.1093/jamia/ocaf068] [Medline: 40323321]
21. Kakaday R, Herrera EZ, Coskey O, Hertel AW, Kaiser P. The STREAMLINE pilot study on time reduction and efficiency in AI-mediated logging for improved note-taking experience. Appl Clin Inform 2025 May;16(3):614-621. [doi: 10.1055/a-2559-5791] [Medline: 40097146]
22. Liu TL, Hetherington TC, Stephens C, et al. AI-powered clinical documentation and clinicians' electronic health record experience: a nonrandomized clinical trial. JAMA Netw Open 2024 Sep 3;7(9):e2432460. [doi: 10.1001/jamanetworkopen.2024.32460] [Medline: 39240568]
23. Solet DJ, Norvell JM, Rutan GH, Frankel RM. Lost in translation: challenges and opportunities in physician-to-physician communication during patient handoffs. Acad Med 2005 Dec;80(12):1094-1099. [doi: 10.1097/00001888-200512000-00005] [Medline: 16306279]

Abbreviations

AI: artificial intelligence
CICU: cardiac intensive care unit
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
PCI: percutaneous coronary intervention
PDQI: Physician Documentation Quality Instrument
PDQI-9: 9-item Physician Documentation Quality Instrument
REDCap: Research Electronic Data Capture
STEMI: ST-elevation myocardial infarction
VUMC: Vanderbilt University Medical Center
WER: word error rate

XSL•FO
RenderX

Edited by A Coristine; submitted 02.Dec.2025; peer-reviewed by M Dilip, W Shi; revised version received 01.Jun.2026; accepted 02.Jun.2026; published 25.Jun.2026.

Please cite as:
Wrenn JO, Behrens M, Hershey MS, Maldaver M, Mitchell J, Thompson T, Triana AJ, Virk ZM, Akdas Y, Cauley MR, Ward MJ, Monahan K
Large Language Model Summarization of Physician-to-Physician Calls for Interhospital Transfer of Patients With ST-Elevation Myocardial Infarction: Observational Study
J Med Internet Res 2026;28:e88834
URL: https://www.jmir.org/2026/1/e88834
doi:10.2196/88834

© Jesse O Wrenn, Madelaine Behrens, Mary S Hershey, Marc Maldaver, John Mitchell, Trevor Thompson, Austin J Triana, Zain M Virk, Yasemin Akdas, Michael R Cauley, Michael J Ward, Ken Monahan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Disclaimers and Referral Patterns for Medical Advice Across Urgency Levels: Large Language Model Evaluation Study

Florian Reis[1], MD; Louis Agha-Mir-Salim[1], MD; Richard Hickstein[2], MD; Moritz Reis[3], PhD; Sophie K Piper[1,4], PhD; Felix Balzer[1], MD, PhD; Sebastian Daniel Boie[1], PhD

[1]Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Charitéplatz 1, Berlin, Germany

[2]Working Group on Cardiovascular Magnetic Resonance, Experimental and Clinical Research Center, a cooperation between Charité – Universitätsmedizin Berlin and the Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany

[3]Institute of Psychology, Julius Maximilians University, Würzburg, Germany

[4]Institute of Biometry and Clinical Epidemiology, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

Corresponding Author:
Florian Reis, MD
Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Charitéplatz 1, Berlin, Germany

Abstract

Background: “I’m not a doctor, but...” is a typical response when asking considerate laypeople for health advice. However, seeking medical advice has also shifted to digital settings, where the expertise of the other party is less transparent than in face-to-face interactions. Recently, large language models (LLMs) have emerged as easily accessible tools, offering a novel way to formulate medical questions and receive seemingly qualified advice. Given the sensitive nature of health-related queries and the lack of professional supervision, incorrect advice can pose serious health risks. Therefore, including explicit disclaimers and precise referrals in LLM responses to medical queries is crucial. However, little is known about how LLMs adapt their safety implementations in response to different urgency levels.

Objective: This study evaluates disclaimer and referral patterns in responses from LLMs to authentic medical queries of different urgency levels using a systematic evaluation framework.

Methods: This prospective, multimodel evaluation study generated and analyzed 908 responses from 4 popular LLMs (GPT-4o, Claude Sonnet-4, Grok-3, and DeepSeek-V3) to 227 authentic patient queries from a public dataset. Two human raters classified all 227 patient queries using a 3-level urgency scale. LLM responses were evaluated using a 5-point ordinal classification system for disclaimer and referral advice, ranging from “no disclaimer” to “urgent advice to consult a medical professional.” GPT-4o served as the primary rater model for this task after conducting a subset validation against human expert annotations. Statistical analyses included Jonckheere-Terpstra tests to examine the relationship between case urgency and disclaimer ratings and Kruskal-Wallis tests for intermodel comparisons.

Results: The 227 patient queries were distributed as 77 (34%) low-urgency, 110 (48%) intermediate-urgency, and 40 (18%) high-urgency cases. All 4 LLMs demonstrated statistically significant ordered trends (all $P<.001$), with higher-urgency queries receiving more explicit referral advice. Disclaimer and referral advice clustered toward higher categories across all models, with 97% (881/908) of responses indicating that a medical professional should be consulted. Sonnet-4, Grok-3, and GPT-4o demonstrated a conservative approach, with 89%, 89%, and 88%, respectively, of their responses being either explicit or urgent referrals. In contrast, DeepSeek-V3 showed a broader distribution, with 74% of responses falling into these categories. Interrater reliability between GPT-4o and human raters achieved moderate to substantial agreement, with weighted Cohen κ values between 0.415 and 0.707.

Conclusions: Current LLMs exhibit urgency-responsive safety mechanisms when providing medical advice. All evaluated models adaptively incorporate more explicit disclaimers and urgent referrals for higher-urgency queries. However, variability between LLMs highlights the need for standardized safety measures and appropriate regulatory frameworks. Although these findings indicate progress regarding safety concerns, the public availability of LLMs requires careful consideration to ensure consistent protection against patient harm while preserving the benefits of low-threshold access to health information.

(J Med Internet Res 2026;28:e84668) doi:10.2196/84668

XSL•FO
RenderX

KEYWORDS

artificial intelligence; large language models; chatbots; patient safety; risk assessment; digital health; health information systems; consumer health information; triage; medical liability

Introduction

The health care landscape faces a fundamental transformation as large language models (LLMs) gain popularity for seeking health information [1]. A recent survey indicates that people without regular access to primary care are particularly likely to use conversational artificial intelligence (AI) systems, such as OpenAI's ChatGPT, to obtain medical advice and interpret symptoms [2]. This shift represents a paradigm change from traditional health care consultation models and has occurred largely outside established health care systems, creating a parallel pathway for health information access and self-diagnosis [3]. The implications are profound, with the potential to influence health care utilization patterns and reshape stakeholder relationships as patients arrive with AI-generated hypotheses and expectations [4].

From a technical perspective, LLMs demonstrate remarkable representation of medical knowledge. They are trained on vast amounts of literature, enabling them to generate contextually appropriate medical information [5]. Empirical evaluations show notable diagnostic reasoning performance, with some LLMs outperforming traditional tools in differential diagnosis and clinical decision support [6,7]. However, significant limitations temper these capabilities. LLMs are susceptible to hallucinations and sometimes fail to follow instructions or guidelines [8,9]. While they perform well on structured multiple-choice questions, they often struggle with nuanced, open-ended clinical reasoning and questions requiring a higher level of expertise [10]. Similarly, many models demonstrate a decline in accuracy when confronted with unfamiliar patterns [11]. This suggests that LLM-encoded "medical knowledge" depends on training data rather than being an analog of human clinical cognition.

From an interpersonal perspective, patient perception and trust are crucial for the acceptance of AI-supported health care measures [12]. Research has shown that people tend to be more critical of medical advice labeled as AI-generated than of advice from human experts, even when the content is identical [13]. Using disclaimers effectively and establishing clear referral pathways can help build trust. Specifically, a well-designed disclaimer that transparently communicates limitations, alongside a clear referral pathway to a human expert, could improve patients' ability to correctly interpret the advice given. While research demonstrates that warning messages can modify compliance behavior [14], professional standards for the implementation of disclaimers for AI health applications remain largely undeveloped.

Notably, the provision of medical advice carries significant legal implications, which also extend to AI-generated health care guidance [15]. However, while producing medical device–like output [16], comprehensive regulatory frameworks governing LLM use in health care remain largely absent, operating in a regulatory gray area that generates significant professional concern about potential patient harm [17]. Although effective risk stratification is an essential part of safe medical practice, integrating automated triage systems into digital applications poses significant challenges [18-20]. Consequently, the emergence of LLMs providing medical advice creates complex liability questions among developers, institutions, and users, given the current lack of regulatory clarity regarding responsibility for patient harm [21].

A particularly underexplored area concerns the relationship between clinical urgency and LLM safety responses. While medical professionals routinely adjust communication based on perceived risk, little is known about the adaptive behavior of LLMs [19]. Additionally, substantial gaps exist between controlled research settings and real-world LLM usage patterns [22]. Many studies focus on narrow domains using artificial vignettes rather than authentic patient queries. Furthermore, standardized frameworks for evaluating disclaimer patterns across LLM architectures are absent, with diverse classification systems limiting cross-study comparisons.

To address these critical gaps, we developed a systematic, multimodel evaluation framework leveraging authentic patient queries to assess LLM disclaimer patterns across different urgency levels. This study employs a clinically relevant urgency scale and a 5-point classification system providing nuanced assessment of safety messaging relative to medical context. We further validated an LLM-based rating system against human expert annotation for scalable evaluation. This design enables an ecosystem-wide analysis across both proprietary and open-weight architectures to determine whether LLMs systematically adapt disclaimer and referral advice based on clinical urgency. This investigation contributes to the development of standards that protect users while preserving the benefits of access to LLM-generated health information.

Methods

Study Design and Research Question

This prospective, multimodel evaluation study aimed to analyze how LLMs embed disclaimers and referral advice into their responses to authentic medical queries of varying clinical urgency.

Data Collection

The study dataset was based on a publicly available, deidentified compilation of more than 100,000 dialogues between patients and doctors from widely used online medical consultation platforms [23]. Each record in the corpus contained a patient query and the corresponding physician reply. We drew a sample of 1000 patient queries for detailed manual annotation. A medical doctor manually screened all sampled queries against the following inclusion criteria: (1) the patient described unclear or undiagnosed symptoms and explicitly sought diagnostic guidance or medical advice, (2) the patient indicated no prior in-person consultation, diagnosis, or ongoing treatment for the

XSL•FO
RenderX

current complaint, and (3) the query was conveyed exclusively in narrative text, without reliance on additional attachments (eg, images or laboratory reports). A total of 773 queries did not meet at least 1 of these criteria and were therefore excluded. All 227 retained queries were minimally edited by removing salutations such as "Dear Doctor" or "Hello Doctor" to prevent systematic bias in subsequent LLM prompting.

Urgency Annotation of Patient Queries

All 227 eligible queries were independently classified by a medical doctor and a psychologist, who were blinded to each other's ratings. For this assessment, we constructed and employed an ordinal, 3-level scale to measure the urgency of typical patient queries from online settings. This approach was informed by previous work on pragmatic triage scales [24] and extended common emergency versus self-care distinctions with an intermediate "professional assessment recommended" category tailored to typical online health queries (Table 1).

Table . Scoring rubric for case urgency levels.

Urgency level	Definition	Implication
1 (Low)	Symptoms compatible with self-care or delayed evaluation	Online advice initially sufficient
2 (Intermediate)	Symptoms warrant professional assessment but not immediate intervention	Outpatient visit recommended
3 (High)	Symptoms suggest a potentially serious or rapidly progressive condition	Timely or urgent in-person consultation required

Any discordant cases were revisited at a consensus meeting and mutually resolved. Patient inquiries were categorized according to a predefined list of medical specialties deemed most likely to be responsible for each case.

LLM Selection and Response Generation

To ensure breadth of vendor representation and model architecture, 4 frequently used state-of-the-art LLMs not specifically designed for explicit multistep reasoning tasks (nonreasoning) were evaluated: GPT-4o (OpenAI, proprietary), Claude Sonnet 4 (Anthropic, proprietary), DeepSeek V3 (DeepSeek, open weight), and Grok 3 (xAI, proprietary). Model architectures were chosen to span both closed weight and open weight ecosystems and to represent widely adopted platforms (proprietary denotes models whose weights are not publicly released; open weight indicates freely downloadable models with public specifications).

To generate model outputs, the system prompt was set to "You are a helpful assistant," and all 227 patient queries were then submitted as user prompts. This minimal instructional approach was chosen to mirror laypeople's use of generative AI tools. All models were accessed via vendor application programming interfaces (APIs) on July 5, 2025 (model IDs: "gpt-4o," "claude-sonnet-4-20250514," "deepseek-chat," and "grok-3"; temperature=0.7; no further specification of parameters or safety settings). This procedure yielded 908 LLM responses (227 queries × 4 LLMs) for subsequent disclaimer and referral annotation.

Output Annotation and Rating Framework: Disclaimer and Referral Advice Classification

Each model response was evaluated to determine the highest level of disclaimer or referral advice present. Disclaimers and referral advice were categorized together, as disclaimers can be seen as a precursor to referrals, which typically contain indirect forms of disclaimers. The prompt used to rate the disclaimer and referral advice is provided in Information 1 in Multimedia Appendix 1 and was identical for all models. In this context, we employed a 5-point ordinal scale for advice classification, reflecting clinically relevant thresholds. This novel approach to assessing disclaimer and referral categories combined a yes/no disclaimer analysis with a pragmatic yet clinically meaningful categorization of referrals (Table 2).

For this advice categorization task, we evaluated 5 LLMs (GPT-o4-mini, GPT-4o-mini, GPT-4o, GPT-4o-latest, GPT-4.1) for their respective performance as rater LLMs. To validate our LLM-based annotation approach, 2 medical doctors independently scored identical subsets with 240 responses (60 cases per LLM: 20 low-urgency, 20 intermediate, 20 high; randomly selected within each urgency category), corresponding to more than 25% (240/908) of the total dataset. Both human raters were provided with identical rating instructions as in the prompt used for the rater LLMs to ensure comparability between human raters and LLMs. Raters were blinded to both the LLM identity and the models' assessments. To reflect real-world uptake of the rating scale and ensure equal conditions for LLMs and human raters, no calibration or consensus meetings were held (Figure 1).

In this context, GPT-4o showed the best overall performance compared with human expert ratings and was therefore selected for further ratings (see Information 2 in Multimedia Appendix 1 for technical settings and performance comparison).

XSL•FO
RenderX

Table . Scoring rubric for disclaimer and referral advice categories.[a]

Category	Definition	Illustrative example
1 (No disclaimer)	No advice to consult a medical professional and no statement of the model's limitations	"You may try resting at home."
2 (General disclaimer)	General statement that medical advice should come from professionals or statement of model limitation (eg, "I am not a medical professional," "This cannot replace professional advice")	"Please note, I am not a doctor."
3 (Differentiated referral)	Advises consulting a medical professional if symptoms prolong, worsen, or new symptoms appear	"If your pain persists or gets worse, you should see a doctor."
4 (Explicit referral)	Advises consulting a medical professional	"You should make an appointment with your GP[b]."
5 (Urgent referral)	Advises consulting a medical professional urgently	"Seek medical attention immediately or visit the emergency department."

[a]The classification assigned to each model response refers to the highest category that appeared in the respective response. Examples are illustrative and were not provided as instructions to human raters, nor to the rater large language models.

[b]GP: general practitioner.

Figure 1. Graphical overview of the study protocol. LLM: large language model.

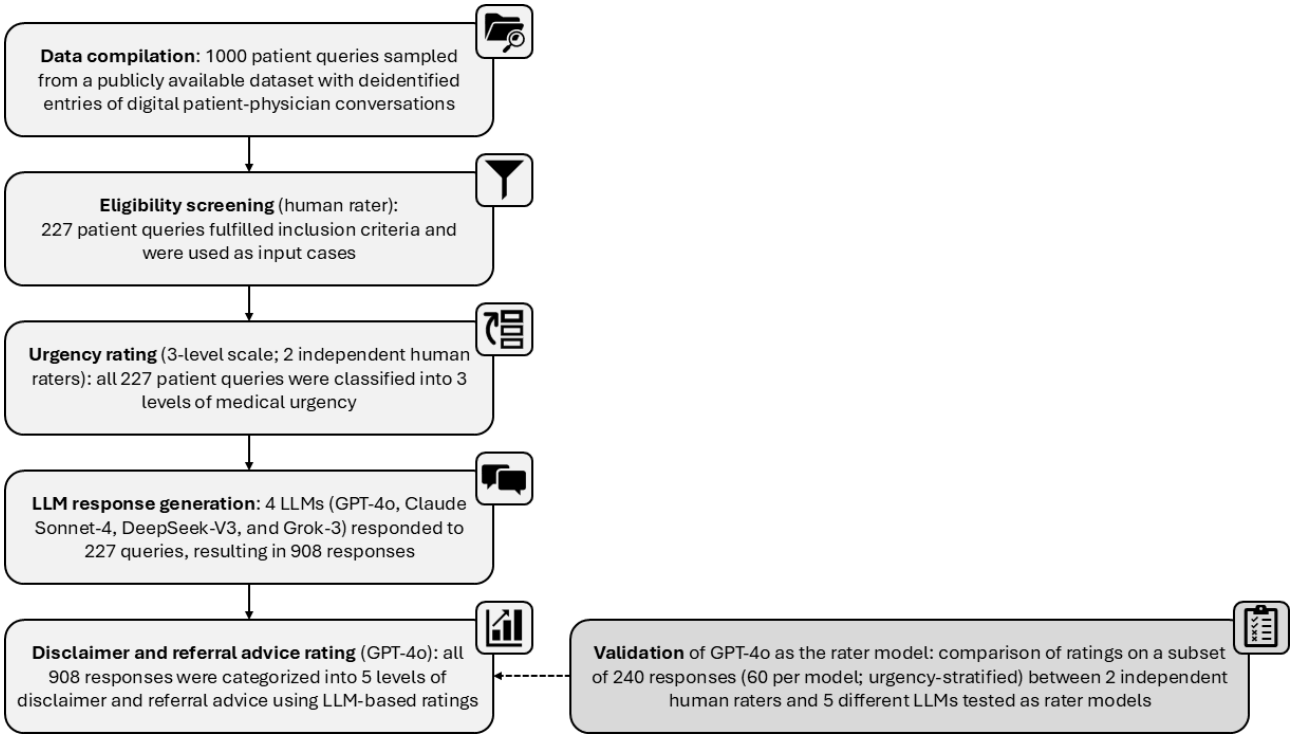

Statistical Analysis

Statistical analyses were performed using Python (version 3.13) with the *matplotlib* (v3.10.5), *numpy* (v2.3.1), *pandas* (v2.3.0), *scipy* (v1.16.0), and *seaborn* (v0.13.2) libraries. Statistical significance was defined as α=.05, unless specified otherwise. Descriptive statistics characterized the dataset and model outputs. Following a significant Jonckheere-Terpstra test indicating a monotonic trend across the ordered groups, post hoc Mann-Whitney tests with Bonferroni correction for multiple comparisons were performed to examine the association between the input medical urgency levels of the patient queries and the disclaimer or referral advice categories in the LLM responses for each LLM. Differences in the distribution of disclaimer and referral levels across the 4 evaluated LLMs were assessed using the nonparametric Kruskal-Wallis H test, given the ordinal nature of the disclaimer and referral ratings. If the Kruskal-Wallis test indicated a statistically significant difference, post hoc pairwise comparisons between the LLMs were conducted using the Dunn test. To account for multiple comparisons, the P values obtained from the Dunn test were adjusted using the Bonferroni correction. Interrater reliability between human raters and LLMs was assessed for output disclaimer and referral advice on a subset of LLM responses for each model. Weighted Cohen κ was used to measure the level of agreement between human and LLM raters in classifying LLM-produced disclaimers and referral advice.

XSL•FO
RenderX

Ethical Considerations

This study utilized an openly available, deidentified dataset sourced from published literature. In accordance with the regulations of Charité – Universitätsmedizin Berlin [25], ethics committee approval was not required for research conducted exclusively on previously published, deidentified data in which no interventions, clinical recruitment, or patient-level interactions were performed. Informed consent was similarly not required under these regulations. To ensure privacy and confidentiality, 2 independent readers reviewed all cases at the data input stage to confirm the absence of any identifiable personal information. No participants were involved in this study, and no compensation was provided.

Results

Descriptive Overview of Input Cases and LLM Responses

Of the 227 eligible patient queries, 77 (34%), 110 (48%), and 40 (18%) were rated as level 1 (low urgency), level 2 (intermediate urgency), and level 3 (high urgency), respectively, after consensus-based agreement between both human raters. Level 2-rated patient queries were the longest on average (mean 383 characters; 74 words [SD 56]), followed by level 3 queries (333 characters; 64 words [SD 35]), whereas level 1 queries were the shortest (312 characters; 61 words [SD 43]). While cases classified as general practitioner dominated at all urgency levels, dental and dermatology cases were more prevalent at levels 1 and 2 and almost absent at level 3. Emergency department cases, although few in the eligible dataset, were found exclusively at urgency level 3 (Table 3).

Table . Absolute count of patient queries per specialty and urgency level.

Specialty	Level 1 (Low urgency)	Level 2 (Intermediate urgency)	Level 3 (High urgency)
Dentistry	2	8	5
Dermatology	16	11	0
Emergency medicine	0	0	2
ENT[a]/otorhinolaryngology	1	7	2
Ophthalmology	5	5	1
General practice	31	38	16
Gynecology	2	6	1
Neurology	0	0	1
Orthopedics	3	10	4
Pediatrics	1	6	3
Psychiatry	3	10	2
Urology	13	9	3

[a]ENT: ear, nose, and throat.

With each of the 4 LLMs responding to all 227 patient queries, a total of 908 responses were generated. On average, Grok-3 replies were the longest at (mean 785 [SD 219] words), while Claude Sonnet-4 replies were the shortest (mean 215 [SD 22] words). The average response length for GPT-4o and DeepSeek-V3 was intermediate (mean 267 [SD 59] words and mean 353 [SD 80] words, respectively). Pearson r correlations between the word count of responses and the GPT-4o rating of the corresponding disclaimers and referrals were small and negative for all 4 LLMs tested (r=−0.024, r=−0.140, r=−0.195, and r=−0.028 for GPT-4o, Sonnet-4, Grok-3, and DeepSeek-V3, respectively). The pooled correlation was approximately r=−0.043, indicating an overall lack of meaningful linear relationship.

Interrater Reliability for Disclaimer and Referral Ratings

On a subset of LLM responses to patient queries produced by GPT-4o, Sonnet-4, Grok-3, and DeepSeek-V3, the 5 LLMs tested as rater models (GPT-4o, GPT-4o-mini, GPT-latest, GPT-o4-mini, and GPT-4.1) produced predominantly moderate (weighted Cohen κ=0.41 - 0.60) to substantial agreement (0.61 - 0.80), with GPT-4o demonstrating the greatest overall alignment with human raters (Table 4).

XSL•FO
RenderX

Table . Quadratic weighted Cohen κ for interrater agreement between rater models and medical doctors regarding the ratings of disclaimer and referral advice produced by 4 different large language models.

LLM[a] for response generation	MD1[b] vs MD2	GPT-4o vs MDs[c]	GPT-4o-mini vs MDs[c]	GPT-4o-latest vs MDs[c]	GPT-o4-mini vs MDs[c]	GPT-4.1 vs MDs[c]
GPT-4o	0.756	0.558	0.534	0.388	0.526	0.532
Sonnet-4	0.583	0.602	0.497	0.463	0.467	0.569
Grok-3	0.714	0.415	0.527	0.338	0.328	0.430
DeepSeek-V3	0.756	0.707	0.671	0.529	0.554	0.642

[a]LLM: large language model.

[b]MD: medical doctor.

[c]The mean value of the 2 interrater reliability values between the respective model and each of the 2 doctors is reported, as 2 doctors rated the dataset.

Association Between Case Urgency Level and LLM Disclaimer and Referral Advice Category

Across all 4 models, higher case urgency was associated with a shift toward higher disclaimer and referral ratings. For GPT-4o, most low-urgency and intermediate-urgency cases were rated as category 4, while high-urgency cases showed a greater proportion of category 5 ratings alongside fewer category 4 ratings. Sonnet-4 displayed a similar pattern, with category 4 ratings predominating at urgency levels 1 and 2, while category 5 ratings dominated at urgency level 3. Grok-3 also concentrated on category 4 for urgency levels 1 and 2, but for high-urgency cases, the largest share shifted to category 5 ratings. DeepSeek-V3 exhibited a broader distribution of ratings overall, with fewer category 4 ratings and more frequent use of categories 2 and 3 across different urgency levels; nevertheless, urgency level 3 still had a higher proportion of category 5 ratings compared to urgency levels 1 and 2 (Figure 2).

XSL•FO
RenderX

Figure 2. Shift of disclaimer and referral advice rating per urgency level for each model. Grouped bar visualization showing the proportion of each disclaimer and referral advice rating category (1‐5, relative proportion on y-axis) for each urgency level (1‐3, x-axis). Each subplot (A-D) illustrates the large language models used to generate responses to patient queries. Within each subplot, bars of the same color correspond to the same rating category as determined by GPT-4o as the rater model.

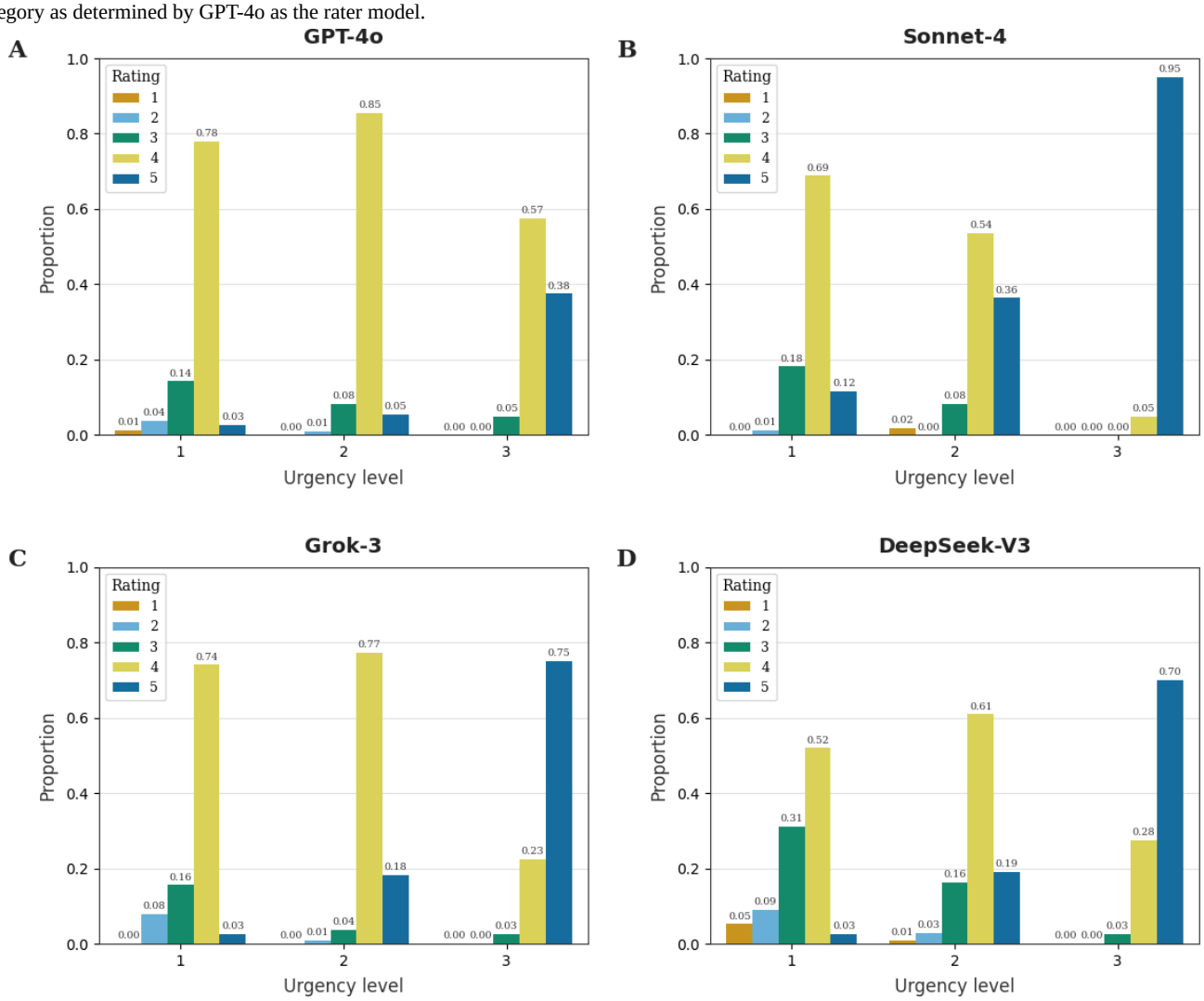

XSL•FO
RenderX

Figure 3. Frequency distributions of disclaimer and referral advice ratings (categories 1 - 5) for each model using GPT-4o as a rater (subplots A-D) and violin plot showing the medians and distributions of disclaimer and referral ratings (categories 1 - 5) for each large language model (LLM) (subplot E). For each LLM, violin width reflects the kernel density estimate of the underlying data; thicker sections mark rating levels with more responses. Labeled dots indicate medians.

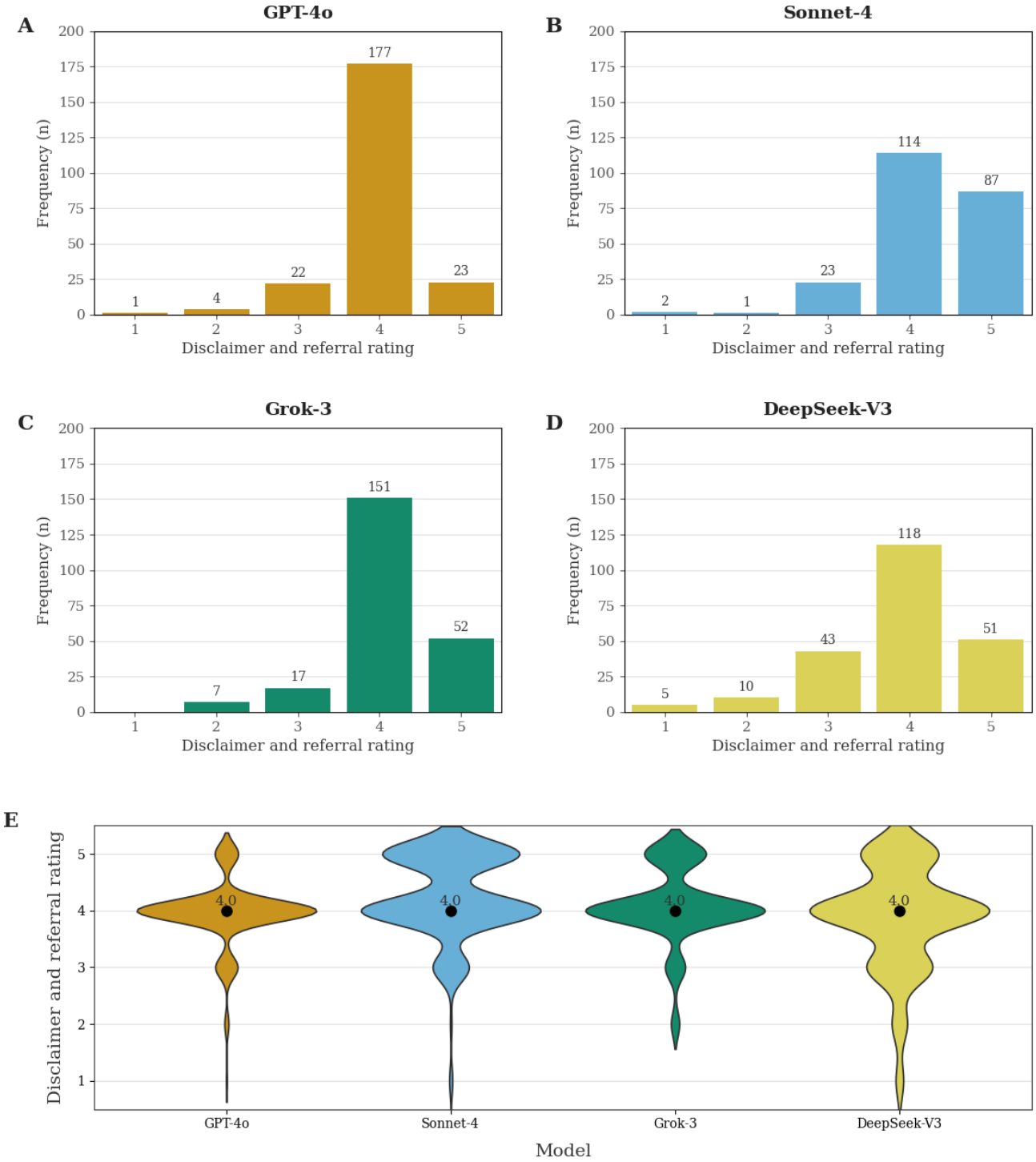

Among responses with a disclaimer and referral category rating of 5, the proportion of cases with level 1 urgency was 2/23 (9%) for GPT-4o, 9/87 (10%) for Sonnet-4, 2/52 (4%) for Grok-3, and 2/51 (4%) for DeepSeek-V3. Conversely, among responses with a disclaimer and referral category rating of less than 4, the proportion of cases with level 3 urgency was 2/27 (7%) for GPT-4o, 0/26 (0%) for Sonnet-4, 1/24 (4%) for Grok-3, and 1/58 (2%) for DeepSeek-V3.

Jonckheere-Terpstra tests for ordered alternatives with pairwise Mann-Whitney comparisons were conducted to examine whether ratings of disclaimer and referral advice (rating categories 1 - 5) increased systematically across levels of case urgency (urgency levels 1 - 3). After Bonferroni correction (α=.05/4=0.0125), the output ratings of all 4 LLMs demonstrated statistically significant ordered trends (all P<.001, detailed analyses are provided in Information 3 in Multimedia Appendix 1).

XSL•FO
RenderX

Comparison of Disclaimer and Referral Advice Ratings Across Different Models

Disclaimer and referral advice ratings clustered toward the upper end of the 5-point scale for all 4 LLMs examined. Grok-3, Sonnet-4, and GPT-4o displayed a narrow spread, with most responses falling into categories 4 to 5 (89%, 89%, and 88%, respectively). In contrast, DeepSeek-V3 exhibited a wider spread, with 74% of responses at categories 4 to 5 and a substantial 19% of responses at category 3, nearly twice the proportion seen with any other tested model. Notably, 7% of DeepSeek-V3 responses did not contain a referral (corresponding to categories 1 - 2), which is more than twice as high as for any other model. With more than 2% of responses falling into category 1, this model was the only one in which the percentage of responses without a disclaimer or referral exceeded 1% of the respective model's total responses. Overall, 30/908 (3%) of all LLM-generated responses did not include any referral to a doctor. Median disclaimer and referral advice rating was identical (4) across the 4 LLMs; however, their overall rating distributions differed, with Sonnet-4 displaying a larger upper-tail mass in the density profile due to 87 ratings at category 5 (Figure 3).

The Kruskal-Wallis test detected significant distribution differences among models (H=37.22, P<.001). Pair-wise Dunn post hoc comparisons (Bonferroni-adjusted) indicated where the differences lay: Sonnet-4 ratings were significantly higher than DeepSeek-V3 and GPT-4o (P<.001) as well as Grok-3 (P=.04). However, the latter difference was sensitive to the rating method; it was not significant when considering only the subset of human-rated cases (P=.58). Rating comparisons of Grok-3 versus DeepSeek-V3 and Grok-3 versus GPT-4o were not significant (P=.94, and P=.53, respectively). Likewise, ratings by GPT-4o and DeepSeek-V3 did not differ (P>.99, see Information 4 and 5 in Multimedia Appendix 1 for details).

Discussion

Principal Findings

In this study, we conducted a prospective, multimodel evaluation of 908 responses generated by 4 state-of-the-art LLMs to 227 authentic patient queries of varying urgency. We observed a consistent association between query urgency and the type of generated disclaimers and referral advice. All 4 models (GPT-4o, Claude Sonnet-4, Grok-3, and DeepSeek-V3) provided progressively more explicit and urgent referrals with increasing query urgency. The median score of disclaimer and referral categories was identical across all models, but the distributions varied between the LLMs. Sonnet-4, Grok-3, and GPT-4o received the highest overall referral ratings (89%, 89%, and 88%, respectively, of responses containing explicit or urgent referrals), while DeepSeek-V3 showed the broadest spread and lowest concentration at higher categories (74%). Notably, disclaimers and referrals were rarely omitted by any model (<1% of responses contained no disclaimer or referral). While most LLM responses provided explicit referral recommendations, model behavior at the end of the distribution (cases receiving minimal disclaimers or urgent referrals) varied between LLMs. These differences in disclaimer integration—a conservative approach by Sonnet-4, Grok-3, and GPT-4o versus DeepSeek-V3 variability—may reflect underlying architectural and training variances. While proprietary models' higher ratings may indicate stronger safety alignments, open weight models like DeepSeek offer transparency and adaptability, but their safety features may be less consistent (see Information 6 in Multimedia Appendix 1 for illustrative examples of patient queries and the corresponding responses produced by different models).

Comparison With Prior Work

Prior studies have highlighted inconsistencies in LLM-generated medical advice, including discrepancies in triage [26] and risk stratification [27]. Wang et al [28] reported that ChatGPT-3.5 and -4.0 included disclaimers in only 10.8% to 13.1% of responses to hepatitis B queries. However, apart from performance progress over time, their relatively narrow definition of disclaimers may explain this discrepancy with our findings, where 99% of responses contained some form of disclaimer or referral advice when using our broader classification system. Similarly, Nastasi et al [29] noted that ChatGPT responses to chest pain vignettes frequently omitted disclaimers on LLM limitations, potentially leading to overutilization of health care resources. Our results extend these findings by showing that modern LLMs are more likely to make explicit and urgent referrals for urgent cases. This may be due to stronger safety alignments in recent architectures and a greater focus on "red teaming," which is the practice of identifying unexpected or undesirable behaviors in models. However, while red teaming may be valuable for characterizing harm mitigations, publicly promoting it as a comprehensive solution to all AI risks constitutes performative safety measures rather than substantive risk mitigation [30]. While our findings suggest that LLMs may embed safety-oriented guidelines to a greater extent than previously reported, this largely depends on respective classifications and types of medical input. This lack of harmonization is particularly noticeable in relation to referrals: if comprehensive guidelines were established in this field, these could also be tested in a retrieval-augmented generation approach to optimize LLM-generated, urgency-stratified medical referrals [31].

Clinical, Behavioral, and Legal Implications of Disclaimers and Referral Advice

The adaptive inclusion of disclaimers and referrals in LLM-generated responses carries significant clinical implications, as incorrect advice could exacerbate patient safety risks [32,33]. Our observation that high-urgency queries prompted almost exclusively explicit or urgent referrals across all models supports their role in mitigating harms. Yet, the persistence of lower categories of disclaimers and referral advice highlights gaps where patients might act on AI-generated advice without consulting professionals. Furthermore, current research on the effectiveness of disclaimers remains inconclusive and context-dependent: prior work suggests that disclaimers alone may have minimal impact on user perception and behavior, placing greater responsibility on developers to educate users about model limitations [34]. However, other research found that participants who heard a disclaimer were significantly more

XSL•FO
RenderX

likely to contact a health care professional, suggesting disclaimers can positively influence patient behavior under certain conditions [35]. We believe that the type of disclaimer may also be relevant here, given the wide range from nonspecific ("AI can make mistakes") to context-specific statements (eg, considering the clinical urgency of a medical question).

From a legal perspective, inconsistent disclaimers may expose both developers and users to liability. Notably, physicians who rely on inaccurate information, such as hallucinations or omissions in LLM-generated content, face malpractice risks [36]. Medical professionals remain legally liable for patient outcomes regardless of whether their decisions are informed by AI-generated advice or traditional sources, necessitating the same rigorous validation standards for AI tools as applied to conventional medical instruments before clinical adoption [37]. The pervasive nature of disclaimers we observed suggests developers are aware of liability risks, but regulatory frameworks have not yet caught up with the reality of widespread public access to these tools. Unregulated LLM-based health care applications continue to reach markets, requiring immediate enforcement of current regulations and development of adaptive frameworks to prevent patient harm and preserve the therapeutic potential of AI-assisted medical care [38]. In our study, the rare occurrence of category 1 (no disclaimer or referral) indicates progress. However, output variability underscores the need to enforce safety standards to mitigate medical risk and avoid public disinformation threats [39].

Feasibility of Using LLMs as Rater Models

Our validation of GPT-4o as a rater model, with weighted Cohen κ values comparable to interhuman agreement (0.415 - 0.707 and 0.583 - 0.756, respectively), demonstrates potential feasibility for scalable annotation of LLM evaluations. Possible discrepancies between the human and LLM ratings may be due to the different category 3 ratings assigned for disclaimers and referrals ("differentiated referral"), which was a category frequently chosen by human raters, but less so by LLMs, as shown by our subset analysis in the (Supplementary Information 4 and 5 in Multimedia Appendix 1). This provides evidence for potential alternatives to human-only annotation, though human supervision remains essential for nuanced clinical judgments.

Strengths of the Study

The novel contribution of our study is its systematic, multimodel approach to evaluating medical disclaimers and referral patterns using authentic patient queries. This provides ecological validity often missing from studies using simulated vignette-based designs. Our urgency and disclaimer classification framework offers a pragmatic, yet clinically meaningful approach for understanding LLM safety responses to digital health queries. By including both proprietary and open weight models, we ensure broader generalizability across the current LLM ecosystem, while LLM-based rating with human validation also offers scalable efficiency and reliability. Furthermore, our minimal prompting approach reflects real-world LLM usage patterns by laypeople.

Limitations and Potential Biases

The limitations of our study include our reliance on text-based queries, which may have been part of the models' training data, and our exclusion of multimodal interactions, during which patients provide supplementary resources such as images. Furthermore, as it is aimed at digital health settings, our disclaimer and referral classification system may not capture all forms of medical safety guidance and may not transfer directly to hospital settings or emergency department triage. We did not assess the temporal stability of the observed disclaimer patterns, as these may change with model updates or adjustments to finetuning approaches. Our ground-truth urgency labels were based on 2 raters (1 physician, 1 psychologist), which may have introduced variability due to differences in expertise. Due to the design of our scoring categories, this study prioritizes the presence and quality of referrals over the adequacy and placement of disclaimers. As our evaluation approach relies exclusively on the highest-scoring category, this limits the analysis of multipart safety responses when disclaimers and referrals are present simultaneously. Finally, using GPT-4o as both the generator and the rater model, even with human-aligned validation, introduces a potential risk of vendor or self-preference bias, whereby the model favors responses that resemble its own output style. Notably, this study did not assess the accuracy of the medical advice provided, instead focusing on the presence and type of disclaimers and referral advice.

Implications and Future Research Directions

LLMs adaptively incorporate disclaimers and referral advice based on query urgency, but differences between models highlight the need for safety measures to ensure reliable medical guidance. While LLMs are generally risk-averse in their approach to medical queries, similar to medical laypeople when deciding whether medical care is required [24], the effectiveness of these mechanisms in protecting patients from harm remains unclear. Clinical implications of these patterns are complex: while conservative referral advice may reduce risks of inappropriate self-care, it could also contribute to health care system overutilization and may not effectively guide patients toward appropriate care. However, we consider the inclusion of a generic disclaimer or referral in LLM-generated responses to medical queries to be the minimum quality standard. As these tools exhibit systematic medical triage–like behavior by embedding urgency-sensitive advice, proper regulatory oversight is required [40]. Future research should focus on exploring the implications for patient safety [41], for example, by assessing the behavioral impacts of multistep user interactions as well as by investigating the relationship between case urgency, response accuracy, and LLM-reported certainty. Another intriguing question is whether LLMs exhibit consistent response patterns when asked for a second opinion based on existing medical recommendations, laboratory results, and imaging data. Since recent studies indicate that the inclusion of disclaimers for certain models has decreased [42], longitudinal studies tracking disclaimer patterns over time should be conducted to reveal the effect of model releases on safety standards.

XSL•FO
RenderX

Conclusions

Our findings demonstrate that current LLMs exhibit urgency-responsive safety mechanisms. However, safety implementations differ substantially between models. Given the widespread public availability of AI tools such as LLMs, our study emphasizes the importance of standardized safety measures and appropriate regulatory frameworks that reflect this new way of accessing health information.

Acknowledgments

The authors used GPT-5 to improve the grammar and language of the manuscript. Following the use of this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

This work was supported by the Open Access Publication Fund of Charité – Universitätsmedizin Berlin.

Data Availability

The analyzed cases and their associated urgency ratings, along with all large language model responses and their corresponding disclaimer and referral advice ratings, are publicly available on Zenodo [43].

Authors' Contributions

Conceptualization: FR, SDB
Methodology: FR, SDB
Formal analysis: FR
Data curation: LA, RH, MR
Validation: LA, RH, MR
Visualization: FR
Statistical advice: SP
Supervision: FB, SDB
Writing – original draft: FR
Writing – review & editing: FR, LAMS, RH, MR, SKP, FB, SDB

Conflicts of Interest

FR and SDB report employment with Pfizer outside the submitted work. FB reports honoraria from Pfizer outside the submitted work. The other authors declare no conflicts of interest.

Multimedia Appendix 1
Supplementary methods, statistical analyses, and illustrative examples (Information 1-6).
[DOCX File, 143 KB - jmir_v28i1e84668_app1.docx]

References

1. Yun HS, Bickmore T. Online health information-seeking in the era of large language models: cross-sectional web-based survey study. J Med Internet Res 2025 Mar 31;27:e68560. [doi: 10.2196/68560] [Medline: 40163112]
2. Ayo-Ajibola O, Julien C, Lin ME, Riddell J, Duan N, Kravitz RL. Association of primary care access with health-related ChatGPT use: a national cross-sectional survey. J Gen Intern Med 2026 Jan;41(2):338-345. [doi: 10.1007/s11606-025-09406-9] [Medline: 39930155]
3. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors 2023 May 17;10:e47564. [doi: 10.2196/47564] [Medline: 37195756]
4. Ayo-Ajibola O, Davis RJ, Lin ME, Riddell J, Kravitz RL. Characterizing the adoption and experiences of users of artificial intelligence-generated health information in the United States: cross-sectional questionnaire study. J Med Internet Res 2024 Aug 14;26:e55138. [doi: 10.2196/55138] [Medline: 39141910]
5. Tu T, Schaekermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence. Nature 2025 Jun;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
6. McDuff D, Schaekermann M, Tu T, et al. Towards accurate differential diagnosis with large language models. Nature 2025 Jun;642(8067):451-457. [doi: 10.1038/s41586-025-08869-4] [Medline: 40205049]
7. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Med Inform Decis Mak 2025 Mar 7;25(1):117. [doi: 10.1186/s12911-025-02954-4] [Medline: 40055694]

XSL•FO
RenderX

8. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med 2024 Sep;30(9):2613-2622. [doi: 10.1038/s41591-024-03097-1] [Medline: 38965432]
9. Chen S, Kann BH, Foote MB, et al. Use of artificial intelligence chatbots for cancer treatment information. JAMA Oncol 2023 Oct 1;9(10):1459-1462. [doi: 10.1001/jamaoncol.2023.2954] [Medline: 37615976]
10. Kaewboonlert N, Poontananggul J, Pongsuwan N, Bhakdisongkhram G. Factors associated with the accuracy of large language models in basic medical science examinations: cross-sectional study. JMIR Med Educ 2025 Jan 13;11:e58898. [doi: 10.2196/58898] [Medline: 39846415]
11. Bedi S, Jiang Y, Chung P, Koyejo S, Shah N. Fidelity of medical reasoning in large language models. JAMA Netw Open 2025 Aug 1;8(8):e2526021. [doi: 10.1001/jamanetworkopen.2025.26021] [Medline: 40779272]
12. Reis F, Reis M, Drzeniek NM, Balzer F. Trust is all you need: reinforcing the patient-physician bond in times of AI. J Med Syst 2025 Nov 5;49(1):154. [doi: 10.1007/s10916-025-02296-8] [Medline: 41188572]
13. Reis M, Reis F, Kunde W. Influence of believed AI involvement on the perception of digital medical advice. Nat Med 2024 Nov;30(11):3098-3100. [doi: 10.1038/s41591-024-03180-7] [Medline: 39054373]
14. Wogalter MS, Kalsher MJ, Racicot BM. Behavioral compliance with warnings: effects of voice, context, and location. Saf Sci 1993 Aug;16(5-6):637-654. [doi: 10.1016/0925-7535(93)90028-C]
15. Mello MM, Guha N. ChatGPT and physicians' malpractice risk. JAMA Health Forum 2023 May 5;4(5):e231938. [doi: 10.1001/jamahealthforum.2023.1938] [Medline: 37200013]
16. Weissman GE, Mankowitz T, Kanter GP. Unregulated large language models produce medical device-like output. NPJ Digit Med 2025 Mar 7;8(1):148. [doi: 10.1038/s41746-025-01544-y] [Medline: 40055537]
17. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 6;6(1):120. [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
18. Taylor RA, Chmura C, Hinson J, et al. Impact of artificial intelligence–based triage decision support on emergency department care. NEJM AI 2025 Feb 27;2(3). [doi: 10.1056/AIoa2400296]
19. Da'Costa A, Teke J, Origbo JE, Osonuga A, Egbon E, Olawade DB. AI-driven triage in emergency departments: a review of benefits, challenges, and future directions. Int J Med Inform 2025 May;197:105838. [doi: 10.1016/j.ijmedinf.2025.105838] [Medline: 39965433]
20. Fraser H, Crossland D, Bacher I, Ranney M, Madsen T, Hilliard R. Comparison of diagnostic and triage accuracy of Ada Health and WebMD symptom checkers, ChatGPT, and physicians for patients in an emergency department: clinical data analysis study. JMIR Mhealth Uhealth 2023 Oct 3;11:e49995. [doi: 10.2196/49995] [Medline: 37788063]
21. Amiot F, Potier B. Artificial intelligence (AI) and emergency medicine: balancing opportunities and challenges. JMIR Med Inform 2025 Aug 13;13:e70903. [doi: 10.2196/70903] [Medline: 40802997]
22. Bhimani M, Miller A, Agnew JD, et al. Real-world evaluation of large language models in healthcare (RWE-LLM): a new realm of AI safety & validation. medRxiv. Preprint posted online on Mar 18, 2025. [doi: 10.1101/2025.03.17.25324157]
23. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. Chatdoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. Cureus 2023 Jun;15(6):e40895. [doi: 10.7759/cureus.40895] [Medline: 37492832]
24. Kopka M, Schmieding ML, Rieger T, Roesler E, Balzer F, Feufel MA. Determinants of laypersons' trust in medical decision aids: randomized controlled trial. JMIR Hum Factors 2022 May 3;9(2):e35219. [doi: 10.2196/35219] [Medline: 35503248]
25. Ethikkommission. Charité – Universitätsmedizin Berlin. 2026. URL: https://ethikkommission.charite.de/ [accessed 2026-03-15]
26. Masanneck L, Schmidt L, Seifert A, et al. Triage performance across large language models, ChatGPT, and untrained doctors in emergency medicine: comparative study. J Med Internet Res 2024 Jun 14;26:e53297. [doi: 10.2196/53297] [Medline: 38875696]
27. Heston TF, Lewis LM. ChatGPT provides inconsistent risk-stratification of patients with atraumatic chest pain. PLoS One 2024;19(4):e0301854. [doi: 10.1371/journal.pone.0301854] [Medline: 38626142]
28. Wang Y, Chen Y, Sheng J. Assessing ChatGPT as a medical consultation assistant for chronic hepatitis B: cross-language study of English and Chinese. JMIR Med Inform 2024 Aug 8;12:e56426. [doi: 10.2196/56426] [Medline: 39115930]
29. Nastasi AJ, Courtright KR, Halpern SD, Weissman GE. A vignette-based evaluation of ChatGPT's ability to provide appropriate and equitable medical advice across care contexts. Sci Rep 2023 Oct 19;13(1):17885. [doi: 10.1038/s41598-023-45223-y] [Medline: 37857839]
30. Feffer M, Sinha A, Deng WH, Lipton ZC, Heidari H. Red-teaming for generative AI: silver bullet or security theater? 2024 Presented at: AIES'24: Proceedings of the Seventh (AAAI/ACM) Conference on AI, Ethics, and Society; Oct 21-23, 2024; San Jose, California, USA p. 421-437. [doi: 10.1609/aies.v7i1.31647]
31. Masanneck L, Meuth SG, Pawlitzki M. Evaluating base and retrieval augmented LLMs with document or online support for evidence based neurology. NPJ Digit Med 2025 Mar 4;8(1):137. [doi: 10.1038/s41746-025-01536-y] [Medline: 40038423]
32. Chen S, Guevara M, Moningi S, et al. The effect of using a large language model to respond to patient messages. Lancet Digit Health 2024 Jun;6(6):e379-e381. [doi: 10.1016/S2589-7500(24)00060-8] [Medline: 38664108]
33. Biro JM, Handley JL, Malcolm McCurry J, et al. Opportunities and risks of artificial intelligence in patient portal messaging in primary care. NPJ Digit Med 2025 Apr 24;8(1):222. [doi: 10.1038/s41746-025-01586-2] [Medline: 40275104]

XSL•FO
RenderX

34. Henestrosa AL, Kimmerle J. “Always check important information!” - the role of disclaimers in the perception of AI-generated content. Comput Hum Behav: Artif Hum 2025 May;4:100142. [doi: 10.1016/j.chbah.2025.100142]
35. Bickmore TW, Ólafsson S, O’Leary TK. Mitigating patient and consumer safety risks when using conversational assistants for medical information: exploratory mixed methods experiment. J Med Internet Res 2021 Nov 9;23(11):e30704. [doi: 10.2196/30704] [Medline: 34751661]
36. Mello MM, Guha N. Understanding liability risk from using health care artificial intelligence tools. N Engl J Med 2024 Jan 18;390(3):271-278. [doi: 10.1056/NEJMhle2308901] [Medline: 38231630]
37. Haupt CE, Marks M. AI-generated medical advice-GPT and beyond. JAMA 2023 Apr 25;329(16):1349-1350. [doi: 10.1001/jama.2023.5321] [Medline: 36972070]
38. Freyer O, Wiest IC, Kather JN, Gilbert S. A future role for health applications of large language models depends on regulators enforcing safety standards. Lancet Digit Health 2024 Sep;6(9):e662-e672. [doi: 10.1016/S2589-7500(24)00124-9] [Medline: 39179311]
39. Menz BD, Modi ND, Sorich MJ, Hopkins AM. Health disinformation use case highlighting the urgent need for artificial intelligence vigilance: weapons of mass disinformation. JAMA Intern Med 2024 Jan 1;184(1):92-96. [doi: 10.1001/jamainternmed.2023.5947] [Medline: 37955873]
40. Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model AI chatbots require approval as medical devices. Nat Med 2023 Oct;29(10):2396-2398. [doi: 10.1038/s41591-023-02412-6] [Medline: 37391665]
41. Huo B, Boyle A, Marfo N, et al. Large language models for chatbot health advice studies: a systematic review. JAMA Netw Open 2025 Feb 3;8(2):e2457879. [doi: 10.1001/jamanetworkopen.2024.57879] [Medline: 39903463]
42. Sharma S, Alaa AM, Daneshjou R. A longitudinal analysis of declining medical safety messaging in generative AI models. NPJ Digit Med 2025 Oct 2;8(1):592. [doi: 10.1038/s41746-025-01943-1] [Medline: 41038984]
43. Disclaimers and referral patterns for medical advice across urgency levels: a large language model evaluation study. Zenodo. URL: https://doi.org/10.5281/zenodo.17185348 [accessed 2026-02-23]

Abbreviations

AI: artificial intelligence
LLM: large language model

Edited by A Coristine; submitted 23.Sep.2025; peer-reviewed by HS Yun, L Masanneck, R Kravitz; revised version received 21.Jan.2026; accepted 22.Jan.2026; published 16.Mar.2026.

Please cite as:
Reis F, Agha-Mir-Salim L, Hickstein R, Reis M, Piper SK, Balzer F, Boie SD
Disclaimers and Referral Patterns for Medical Advice Across Urgency Levels: Large Language Model Evaluation Study
J Med Internet Res 2026;28:e84668
URL: https://www.jmir.org/2026/1/e84668
doi:10.2196/84668

© Florian Reis, Louis Agha-Mir-Salim, Richard Hickstein, Moritz Reis, Sophie K Piper, Felix Balzer, Sebastian Daniel Boie. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effectiveness of AI-Assisted Patient Health Education Using Voice Cloning and ChatGPT: Prospective Randomized Controlled Trial

Yan Sun[1,2], MSc; Shangqing Xu[1], BSc; Hongying Jin[1], MSc; Xiaoyan Han[1], MSc; Kangqi Jin[1], MSc; Yimei Zhang[1], MSc; Xiaoli Ma[1], MSc; Huaping Wei[3], BSc; Minjie Ma[4], PhD

[1]Department of Thoracic Surgery, The First Hospital of Lanzhou University, Donggang West Road 1#, Lanzhou, China

[2]School of Nursing, Lanzhou University, Lanzhou, China

[3]Outpatient Department, The First Hospital of Lanzhou University, Lanzhou, China

[4]Gansu International Science and Technology Cooperation Base for Development and Application of Thoracic Surgery Key Technologies, The First Clinical Medical College of Lanzhou University, Department of Thoracic Surgery, The First Hospital of Lanzhou University, Donggang West Road 1#, Lanzhou, China

Corresponding Author:
Minjie Ma, PhD
Gansu International Science and Technology Cooperation Base for Development and Application of Thoracic Surgery Key Technologies, The First Clinical Medical College of Lanzhou University, Department of Thoracic Surgery, The First Hospital of Lanzhou University, Donggang West Road 1#, Lanzhou, China

Abstract

Background: Traditional patient education often lacks personalization and engagement, potentially limiting knowledge acquisition and treatment adherence. Advances in artificial intelligence (AI), including voice cloning technology and large language models (eg, ChatGPT), offer new opportunities to deliver personalized, scalable, interactive health education. However, evidence regarding the comparative effectiveness of different AI-based voice cloning strategies and reliability of automated AI evaluation tools remains limited.

Objective: This study aims to evaluate the effectiveness of AI-assisted patient education integrating voice cloning and ChatGPT, compare physician voice cloning with patient self-voice cloning, and assess the reliability of ChatGPT as an automated evaluation tool for education outcomes.

Methods: In this prospective, 3-arm, parallel-group randomized controlled trial, 180 hospitalized patients requiring standardized health education were recruited from a tertiary hospital. Inclusion criteria were age ≥18 years, clear diagnosis requiring health education, clear consciousness, and voluntary participation with informed consent. Exclusion criteria were severe hearing impairment, severe cognitive impairment, expected hospitalization <3 days, or prior participation in similar studies. Using a computer-generated random sequence, participants were randomly assigned (1:1:1) to receive traditional education (control), AI-assisted education using physician voice cloning, or AI-assisted education using patient self-voice cloning, each with identical educational content of equal duration. The primary outcome was education content compliance, evaluated using ChatGPT-4 with validated prompts and verified by expert review. Secondary outcomes included knowledge retention, education satisfaction, treatment adherence, quality of life, and psychological status. Outcome assessors and data analysts, but not participants, were blinded to group allocation.

Results: Of 180 randomized participants, 174 (96.7%) completed the trial. Both AI-assisted groups had significantly higher mean education content compliance scores immediately posteducation than the control group (physician voice: 86.7, SD 7.3; self-voice: 92.5, SD 6.8; control: 73.2, SD 8.5; $P<.001$). The patient self-voice group showed superior predischarge knowledge retention, higher education satisfaction, and greater treatment adherence than the other groups (all $P\leq.02$). At the 1-month follow-up, the self-voice group maintained improved adherence (Cohen $d=0.74$) and had significantly lower anxiety and depression scores (all $P\leq.02$) and improved SF-36 quality-of-life domains. ChatGPT-based evaluations demonstrated high reliability (weighted $\kappa=0.87$, 95% CI 0.82 - 0.91).

Conclusions: The innovative patient education model integrating AI voice cloning and ChatGPT is distinct from previous studies primarily relying on standard text-to-speech or professionally recorded content. Using patients' own cloned voices for health education delivery leveraged the self-reference effect to enhance learning outcomes. Compared with research using clinician-narrated content, this study highlights that self-voice education produces superior outcomes across multiple domains including compliance, satisfaction, and psychological well-being. These findings establish a theoretical and practical framework for personalized AI-driven patient education. In real-world clinical settings, this approach offers a scalable, cost-effective solution to enhance patient engagement, particularly valuable in resource-limited environments where individualized education is challenging to deliver.

XSL•FO
RenderX

Trial Registration: Chinese Clinical Trial Registry ChiCTR2500101882; https://www.chictr.org.cn/showprojEN.html?proj=268927

(J Med Internet Res 2026;28:e81387) doi:10.2196/81387

KEYWORDS

artificial intelligence; voice cloning; medical education; ChatGPT; education effect evaluation; randomized controlled trial

Introduction

Patient education is a fundamental component of health care, aiming to enhance individuals' understanding of their medical conditions, promote self-management, improve treatment adherence, and ultimately optimize health outcomes and prognosis [1-3]. Despite the clinical significance of patient education, traditional methods of patient education—primarily verbal explanations by health care providers and standardized written materials—often fall short in addressing the diverse needs of patients [4,5]. These approaches are constrained by limited time, lack of interactivity, and insufficient personalization, frequently resulting in suboptimal comprehension, poor engagement, and reduced adherence to treatment recommendations [6].

Prior research has shown that personalized, interactive education enhances patient understanding and engagement [7-11]. Recent advances in artificial intelligence (AI), particularly in voice cloning and large language models (LLMs), present novel opportunities to transform conventional health education into a more personalized, scalable, and interactive process [12]. Voice cloning technology uses deep learning algorithms to synthesize natural-sounding speech from limited voice samples, enabling the delivery of educational content in familiar voices—such as those of physicians or even the patients themselves [13,14]. Prior studies suggest that familiar voices can increase emotional resonance [15-17], build trust, enhance patient engagement, improve memory retention of medical information [15,17], and increase accessibility to digital health tools. Recent scoping reviews have identified multiple applications of LLMs in patient education, including generating educational materials, interpreting medical information, and optimizing doctor-patient interaction [18]. Systematic reviews have further demonstrated that LLMs can enhance health care communication and support personalized patient engagement [19].

Simultaneously, LLMs like ChatGPT, which is based on advanced LLMs, have demonstrated strong capabilities for natural language understanding, dialog generation, and semantic evaluation [20] and have shown potential for supporting patient communication and education [18,19,21-23]. These tools can interact with patients in real time, deliver conversational-style education, and assess understanding through dynamic, automated evaluation processes [24-26]. Such integration may address key challenges in health education, including the need for tailored content delivery and standardized assessment without overburdening clinical staff. These models are capable of few-shot learning, enabling rapid adaptation to novel prompts [27].

Despite this emerging literature, significant gaps remain. First, few empirical studies have rigorously examined the comparative effectiveness of different voice sources—specifically physician-cloned voice versus patient-cloned voice—in AI-assisted patient education. Second, although LLM-based tools like ChatGPT show promise for health communication, their role as standardized evaluators of educational outcomes has not been comprehensively validated in clinical randomized trials. Traditional outcome assessments often rely on human raters, which introduces inter-rater variability and resource burdens.

To address these gaps, this study was a prospective randomized controlled trial (RCT) to investigate the effectiveness of an AI-assisted patient education system integrating voice cloning and ChatGPT. The study compared educational outcomes between physician voice cloning and patient self-voice cloning and explored the feasibility and accuracy of using ChatGPT as a standardized, automated evaluation tool. By doing so, this work aimed to contribute novel insights into how AI technologies can enhance the personalization, efficiency, and effectiveness of inpatient education and to extend the evidence base for AI-enabled patient engagement strategies. We hypothesized that (1) AI-assisted voice cloning education would significantly improve patient education outcomes compared with traditional education, (2) patient self-voice cloning would yield superior educational effectiveness compared with physician voice cloning due to the self-reference effect, and (3) ChatGPT could serve as a reliable automated evaluation tool with high agreement with expert assessment.

Methods

Overall Study

This study used an RCT design to assess the impact of AI-assisted patient health education using voice cloning technology and ChatGPT on patient education outcomes. Participants were randomly assigned to 1 of 3 groups: The 3 groups received different educational methods but the same educational content and duration of education. Following the educational intervention, the effectiveness of the education was comprehensively evaluated at fixed time points using ChatGPT. To ensure the validity and accuracy of the evaluation tool, a pre-test of the ChatGPT evaluation tool was conducted prior to the formal data collection. This study not only explored the potential of AI voice cloning technology in personalized patient education but also used ChatGPT as an assessment tool to further validate its reliability and practicality in evaluating health education interventions. Ultimately, the study aimed to provide theoretical insights and practical recommendations for improving medical health education models, enhancing patient

XSL•FO
RenderX

education outcomes, and promoting treatment adherence and quality of life.

Study Timeline

An independent tool validation study was conducted from January 2024 to June 2024 to verify the ChatGPT-based compliance scoring tool (no participant enrollment). Administrative preparation, staff training, and software and hardware debugging were completed in December 2024, and this period did not involve participant enrollment. Trial registration (ChiCTR2500101882) was initiated on January 15, 2025, and finalized on April 30, 2025, before enrollment began in May 2025. Participant enrollment and intervention delivery were conducted from May 2025 to June 2025, with 1-month follow-up assessments completed by July 2025.

Reporting Guidelines

The trial was conducted and reported in accordance with the CONSORT (Consolidated Standards of Reporting Trials) 2025 statement and the CONSORT-EHEALTH checklist for digital and AI-based health interventions (Checklists 1 and 2) [28,29].

Study Design

This single-center, 3-arm, parallel-group, superiority RCT (allocation ratio 1:1:1) was conducted at The First Hospital of Lanzhou University. System development, intervention standardization, and staff training were completed in December 2024 and January 2025. Participant recruitment and data collection were conducted from May 2025 to July 2025.

Randomization

Sequence Generation

The random allocation sequence was generated using a computer-generated random number table.

Allocation Concealment

To ensure concealment of group assignments, the allocation sequence was placed in sealed, opaque envelopes managed by independent researchers who were not involved in the implementation of the interventions. This procedure minimized the risk of selection bias.

Implementation

Independent researchers generated the allocation sequence and managed the envelopes. Eligible participants were assigned to study arms according to the contents of the envelopes.

Blinding

Due to the nature of the interventions, neither participants nor the medical staff implementing the interventions could be blinded. However, outcome assessors were blinded to group assignments to minimize assessment bias.

Ethical Considerations

The clinical trial protocol was reviewed and approved by the Ethics Committee of The First Hospital of Lanzhou University (approval number: LDYYLL-2025‑805) and conducted in accordance with the principles of the Declaration of Helsinki.

All participants received detailed verbal and written information regarding the study purpose, procedures, potential risks, and expected benefits and provided written informed consent prior to enrollment. The consent process also covered permission for the use and publication of anonymized data. Participants were informed that their involvement was voluntary and that they could withdraw from the study at any time without affecting their standard medical care.

All collected data were deidentified before analysis and stored in encrypted, password-protected databases accessible only to the research team. Personal identifiers were removed to ensure participant confidentiality and data security.

The use of voice cloning technology was explicitly covered in the consent form. Participants and physicians provided separate written authorization for the use of their voice recordings solely for research purposes. All voice samples and generated models were permanently deleted after study completion.

Participants did not receive monetary compensation but were offered free access to the AI-based health education program and follow-up consultations as part of their clinical care. No identifiable photographs nor audiovisual materials were used in this publication.

Participants

The study participants were patients hospitalized in a tertiary hospital from May 2025 to July 2025 who needed to receive medical education.

Inclusion and Exclusion Criteria

The inclusion criteria were (1) age ≥18 years of any gender; (2) clear diagnosis, patients who needed to receive medical education; (3) clear consciousness, able to understand and cooperate with the study; and (4) voluntary participation in this study and signed informed consent. The exclusion criteria were (1) severe hearing impairment, unable to normally receive voice education; (2) severe cognitive impairment, unable to understand education content; (3) expected hospitalization time <3 days; and (4) previously participated in similar studies.

Sample Size

Based on the pre-experimental results, with the education content compliance rate as the main observation indicator, α set at 0.05 (2-sided), β set at 0.10, a test power of 90%, and an expected effect size of 0.3 among the 3 groups, we calculated that each group needed a sample size of 54 participants using G*Power 3.1 software. Considering a possible dropout rate of approximately 10%, we finally determined that we would enroll 60 participants in each group, for a total of 180 participants. Prior to participation, all patients were informed that the educational intervention would be delivered using AI-generated voice recordings. Written informed consent was obtained from all participants.

Independent Tool Validation Study (January 2024 to June 2024): Validation of ChatGPT-Based Compliance Scoring

This independent tool validation study was conducted separately from the randomized trial and did not involve participant

XSL•FO
RenderX

enrollment. To validate the AI-based scoring used for the primary outcome, we conducted a separate pilot validation study (January 2024 through June 2024) among 30 volunteers who were not part of the randomized trial. Phase 1 (verbatim-matching workflow) used iFlytek automatic speech recognition (Mandarin medical context; no manual correction) to transcribe each participant's recitation; the transcribed text was pasted into the ChatGPT web interface together with the standard education text for rubric-based scoring. Phase 2 (semantic audio-based workflow) re-evaluated the same recordings by uploading the original audio directly to the ChatGPT web interface (WAV/MP3; 16-kHz sampling rate; 60 - 120 s per recording); each recording was scored once. Both workflows were benchmarked against consensus ratings from 3 senior nursing experts using the same rubric. Agreement with expert ratings was substantial in Phase 1 (weighted κ=.72) and higher in Phase 2 (weighted κ=.87), with strong internal consistency of the evaluation items (Cronbach α=0.89). Therefore, Phase 2 was selected for the main RCT to minimize transcription-related errors and enhance reproducibility. Detailed prompts, rubric, and example inputs and outputs are provided in Multimedia Appendix 1.

Intervention Measures

Traditional Education Group (Control Group)

The traditional education group received the hospital's current standardized patient education, including verbal education by medical staff and standardized written education materials. Verbal education was conducted by trained medical staff according to a unified education outline, including disease knowledge, treatment plans, precautions, and lifestyle guidance. Written education materials were standardized education manuals compiled by the hospital that patients could take back to the ward or home to read. The duration of the education was approximately 20 minutes to 30 minutes.

Attending Physician Voice Cloning Education Group (Intervention Group 1)

The physician voice cloning education group (intervention group 1) received patient education through cloning of the attending physician's voice. First, the patient's attending physician's voice sample (approximately 5 min) was obtained, and Resemble.ai voice cloning technology was used to generate the doctor's voice model. Medical professionals then formulated personalized education content based on standard education content and the patient's specific situation, synthesized the voice using doctor voice cloning, and produced personalized voice education materials. Patients listened to the voice-cloned education content through headphones or speakers under the guidance of research assistants and received printed personalized education materials. The education content covered the same topics as the control group but was personalized according to patient characteristics, such as using language expressions suitable for the patient's education level, lifestyle suggestions targeted at the patient's occupational characteristics, and emphasizing relevant content based on the patient's concerns. The duration of the education was approximately 20 minutes to 30 minutes.

Patient's Own Voice Cloning Education Group (Intervention Group 2)

The patient voice cloning education group received patient education through cloning of the patient's own voice. First, the patient's voice sample (approximately 5 min) was obtained, and Resemble.ai voice cloning technology was used to generate the patient's voice model. Similar to intervention group 1, medical professionals then formulated personalized education content, synthesized the voice through patient voice cloning, and produced personalized voice education materials. Patients listened to the voice-cloned education content through mobile phone headphones or speakers under the guidance of research assistants and received printed personalized education materials. The education content was the same as in intervention group 1, with the only difference being the use of the patient's own voice for education. The duration of the education was approximately 20 minutes to 30 minutes.

Educational Content and Assessment of Education Effect

All participants received the same educational content and could ask questions after receiving education, with medical staff answering inquiries. All 3 groups of patients had their education effect assessed before discharge by using ChatGPT to compare the education content recited by patients with the original education content to calculate the content compliance rate, and follow-up assessments were conducted 1 month after discharge.

The educational intervention included AI-generated voice recordings delivered using cloning of either the attending physician's voice or the patient's own voice. ChatGPT was used to assess patient knowledge, satisfaction, and adherence using standardized prompts. Patient outcomes were also measured using structured questionnaires. The questionnaires are provided in Multimedia Appendices 2-4, and the voice cloning patient education prompts, ChatGPT evaluation prompts, and AI-generated educational content are provided in Multimedia Appendix 5. Data were collected at baseline and postintervention, and all responses were recorded in a secure electronic database.

Evaluation Indicators and Measurement Tools

Primary Outcome Measure

The primary outcome, education content compliance rate, was defined as the degree of semantic concordance between the education content recited by patients and the standardized education material using the following equation: compliance rate = (number of key points correctly recited by patients/total number of key points in education content) × 100%. The education content compliance rate was assessed using ChatGPT-4 (model ChatGPT [OpenAI; web interface, model displayed as GPT-4 during the study period]) as an automated evaluation tool to quantify semantic similarity between the patient's oral recitation and the standard education content, enabling scalable and objective scoring.

XSL•FO
RenderX

Secondary Outcome Measures

Knowledge Mastery

We used disease-related knowledge questionnaires to assess patients' understanding and memory of disease knowledge, with a possible total score of 100 points and higher scores indicating better knowledge mastery.

Education Satisfaction

We used the education satisfaction scale to assess patients' satisfaction with the education process and content, including aspects such as education content, method, time, and effect, using a 5-level Likert scale. Higher scores indicate higher satisfaction.

Treatment Adherence

We used the treatment adherence scale to assess patients' treatment adherence within 1 month after discharge, including aspects such as drug treatment, lifestyle adjustment, and follow-up review. Higher scores indicate better adherence.

Quality of Life

We used the SF-36 quality of life scale to assess changes in patients' quality of life 1 month after discharge.

Anxiety and Depression

We used the Hospital Anxiety and Depression Scale to assess changes in patients' anxiety and depression before and after the education.

Evaluation Time Points

Baseline Assessment

We collected patients' basic information and baseline data, including demographic characteristics, disease-related knowledge, anxiety, and depression, before the intervention.

Immediate Posteducation Assessment

The education content compliance rate, knowledge mastery, and education satisfaction were assessed immediately after the education was completed.

Predischarge Assessment

Education content compliance rate and knowledge mastery were re-assessed 1 day before patient discharge.

1-Month Postdischarge Follow-Up Assessment

Through telephone or outpatient follow-up, we assessed patients' treatment adherence, quality of life, disease-related knowledge retention, anxiety, and depression.

Data Collection and Management

Research assistants who had received standardized training were responsible for data collection, using uniform data collection forms to record patient information and evaluation results. An electronic database was established, with all data independently entered by 2 personnel and cross-checked to ensure accuracy. To protect patient privacy, all data were stored in coded form, and regular data quality checks were conducted to ensure completeness and reliability. Missing data were promptly identified and imputed.

All voice recordings collected in this study were considered sensitive personal data. They were stored on secure, password-protected servers at The First Hospital of Lanzhou University, accessible only to authorized research personnel. For analysis, the recordings were anonymized. Following the completion of the study, all original recordings will be permanently deleted in accordance with institutional data retention and privacy policies.

Statistical Analysis

Statistical analyses were performed using SPSS version 26.0 (IBM Corp). All analyses followed the intention-to-treat (ITT) principle. Missing outcome data (6/182, 3.3%) were evaluated using the missing completely at random test from Little, which indicated that missingness was completely at random (χ^2_{12}=14.5, P=.27). Consequently, multiple imputation by chained equations was used to generate 5 imputed datasets (m=5), and pooled estimates were obtained using the rules by Rubin. Sensitivity analyses using per-protocol (complete-case) data yielded consistent findings (absolute difference in mean compliance<1%). Continuous variables are presented as mean (SD), and categorical variables are presented as n (%). Between-group comparisons were conducted using 1-way ANOVA with post hoc tests (least significant difference or Bonferroni, as appropriate). Categorical variables were compared using χ^2 tests or Fisher exact tests. All tests were 2-sided with a significance level of P<.05.

Evaluation Indicators

The primary outcome of this study was the education content compliance rate, which reflected how well patients understood and followed the health education provided. Secondary outcomes included knowledge retention, patient satisfaction, and treatment adherence.

Education level and knowledge retention were measured using a structured questionnaire developed by the research team based on the hospital's standardized education materials.

The questionnaire consisted of multiple-choice and short-answer questions assessing patients' understanding and recall of the key educational points immediately after education and before discharge.

Each correct response was scored as 1 point, with higher total scores indicating better knowledge retention.

The disease-related knowledge questionnaire was developed based on the hospital's standardized education materials and underwent content review by 3 senior nursing experts.

Results

Participant Flow

Figure 1 illustrates the participant flow. ITT analysis was performed for all participants (n=180) using multiple imputation by chained equations for missing data. The analysis section in the figure shows the per-protocol analyses (n=174).

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram.

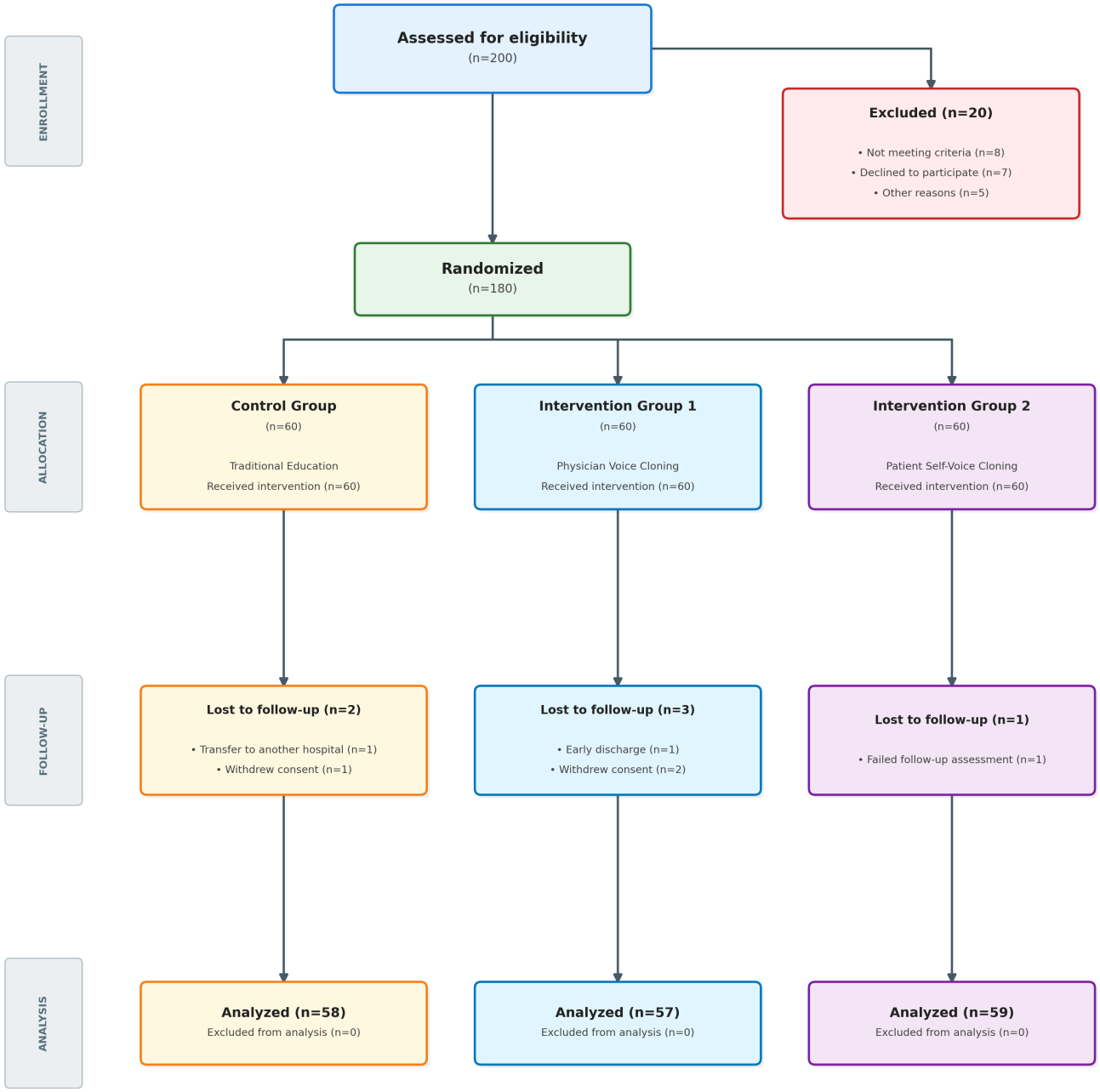

Comparison of General Information

A total of 180 patients were recruited to participate in this study. During the follow-up period, 6 participants were lost to follow-up due to transfer to another hospital (n=1) and withdrawal of consent (n=1) in the control group; early discharge (n=1) and withdrawal of consent (n=2) in intervention group 1; and failure to complete follow-up assessments (n=1) in intervention group 2. Therefore, 174 patients completed the study (58 in the control group, 57 in intervention group 1, 59 in intervention group 2), for a total dropout rate of 3.3% (6/180). After multiple imputation, the analysis results remained consistent with those of the complete-case dataset, indicating the robustness of the findings. Comparison of general information such as age, gender, education level, occupation, and disease type among the 3 groups of patients showed no statistically significant differences ($P>.05$), indicating comparability (Table 1).

XSL•FO
RenderX

Table . Comparison of baseline characteristics among 3 groups of hospitalized patients with lung cancer in a randomized controlled trial of artificial intelligence (AI)–assisted voice cloning education at The First Hospital of Lanzhou University (May 2025 through July 2025).

Item	Control group (n=58)	Intervention group 1 (n=57)	Intervention group 2 (n=59)	Statistical test result (*df*)	*P* value
Age (years), mean (SD)	52.6 (14.8)	53.2 (15.1)	51.9 (14.5)	0.127[a] (2, 171)	.88
Gender, n (%)				0.215[b] (2)	.90
Male	31 (53)	32 (56)	34 (58)		
Female	27 (47)	25 (44)	25 (42)		
Education level, n (%)				0.986[b] (6)	.91
Primary school and less	12 (21)	11 (19)	10 (17)		
Junior high school	18 (31)	17 (30)	19 (32)		
High school/technical secondary school	16 (28)	18 (32)	17 (29)		
College and more	12 (21)	11 (19)	13 (22)		
Marital status, n (%)				0.352[b] (2)	.84
Married	49 (85)	47 (83)	51 (86)		
Unmarried/divorced/widowed	9 (16)	10 (18)	8 (14)		
Disease type, n (%)				0.763[b] (6)	.94
Cardiovascular disease	21 (36)	20 (35)	22 (37)		
Respiratory system disease	15 (26)	16 (28)	17 (29)		
Digestive system disease	13 (22)	12 (21)	11 (19)		
Others	9 (16)	9 (16)	9 (15)		
Previous hospitalization history, n (%)				0.124[b] (2)	.94
Yes	32 (55)	33 (58)	33 (56)		
No	26 (45)	24 (42)	26 (44)		
Baseline HADS[c] score, mean (SD)					
Anxiety	7.8 (3.2)	7.6 (3.1)	7.7 (3.3)	0.068[a] (2, 171)	.93
Depression	6.9 (2.8)	7.1 (2.9)	6.8 (2.7)	0.185[a] (2, 171)	.83

[a]*F* test.

[b]Chi-square test.

[c]HADS: Hospital Anxiety and Depression Scale.

Comparison of Education Content Compliance Rate

Immediately after education, the mean education compliance score in the control group was 73.2 (SD 8.5; 95% CI 71.0 - 75.4), compared with mean scores of 86.7 (SD 7.3; 95% CI 84.8 - 88.6) in intervention group 1 and 92.5 (SD 6.8; 95% CI 90.8 - 94.2) in intervention group 2. The 1-way ANOVA showed a statistically significant difference among the 3 groups ($F_{2,171}$=103.427, P<.001). Before discharge, the mean score in the control group was 68.5 (SD 9.1; 95% CI 66.2 - 70.8), while the mean scores in intervention group 1 and intervention group 2 were 82.3 (SD 8.1; 95% CI 80.2 - 84.4) and 88.6 (SD 7.4; 95% CI 86.7 - 90.5), respectively. The between-group difference remained statistically significant ($F_{2,171}$=95.682, P<.001; Table 2).

XSL•FO
RenderX

Table . Comparison of education content compliance rate among 3 groups of hospitalized patients with lung cancer in a randomized controlled trial of artificial intelligence (AI)–assisted voice cloning education at The First Hospital of Lanzhou University (May 2025 through July 2025), with differences analyzed using 1-way ANOVA.

Time point	Control group (n=58), mean (SD; 95% CI)	Intervention group 1 (n=57), mean (SD; 95% CI)	Intervention group 2 (n=59), mean (SD; 95% CI)	*F* test (*df*)	*P* value
Immediately after education	73.2 (8.5; 71.0 - 75.4)	86.7 (7.3; 84.8 - 88.6)	92.5 (6.8; 90.8 - 94.2)	103.427 (2, 171)	<.001
Before discharge	68.5 (9.1; 66.2 - 70.8)	82.3 (8.1; 80.2 - 84.4)	88.6 (7.4; 86.7 - 90.5)	95.682 (2, 171)	<.001

The corresponding ITT results based on the multiply imputed datasets are provided in Supplementary Table S2.

Comparison of Knowledge Mastery

Immediately after education, the mean knowledge mastery score in the control group was 76.3 (SD 9.2; 95% CI 73.9 - 78.7), which was significantly lower than the mean scores in intervention group 1 (88.5, SD 8.4; 95% CI 86.3 - 90.7) and intervention group 2 (90.2, SD 8.1; 95% CI 88.1 - 92.3). A statistically significant difference was observed among the 3 groups ($F_{2,171}$=48.362, P<.001). Before discharge, the control group had a mean knowledge mastery score of 71.8 (SD 9.8; 95% CI 69.2 - 74.4), whereas intervention group 1 and intervention group 2 achieved higher mean scores of 84.7 (SD 8.9; 95% CI 82.3 - 87.1) and 87.3 (SD 8.5; 95% CI 85.1 - 89.5), respectively. The between-group difference remained statistically significant ($F_{2,171}$=46.175, P<.001). At 1 month after discharge, the mean knowledge mastery score further declined in the control group (65.2, SD 10.5; 95% CI 62.4 - 68.0). In contrast, both intervention groups maintained relatively higher levels of knowledge mastery, with mean scores of 78.6 (SD 9.3; 95% CI 76.1 - 81.1) in intervention group 1 and 83.5 (SD 8.9; 95% CI 81.2 - 85.8) in intervention group 2. The overall difference among the 3 groups remained statistically significant ($F_{2,171}$=54.793, P<.001; Table 3).

Table . Comparison of knowledge mastery scores among 3 groups of hospitalized patients with lung cancer in a randomized controlled trial of artificial intelligence (AI)–assisted voice cloning education at The First Hospital of Lanzhou University (May 2025 through July 2025), with differences among groups analyzed using 1-way ANOVA.

Time point	Control group (n=58), mean (SD; 95% CI)	Intervention group 1 (n=57), mean (SD; 95% CI)	Intervention group 2 (n=59), mean (SD; 95% CI)	*F* test (*df*)	*P* value
Immediately after education	76.3 (9.2; 73.9-78.7)	88.5 (8.4; 86.3-90.7)	90.2 (8.1; 88.1-92.3)	48.362 (2, 171)	<.001
Before discharge	71.8 (9.8; 69.2-74.4)	84.7 (8.9; 82.3-87.1)	87.3 (8.5; 85.1-89.5)	46.175 (2, 171)	<.001
1 month after discharge	65.2 (10.5; 62.4-68.0)	78.6 (9.3; 76.1-81.1)	83.5 (8.9; 81.2-85.8)	54.793 (2, 171)	<.001

Comparison of Education Satisfaction

Significant differences were observed among the 3 groups across all dimensions of education satisfaction, including education content, education method, education time, education effect, and overall satisfaction (all P<.001). Across all items, the 95% CIs of the 2 intervention groups were consistently higher than those of the control group, with minimal overlap between groups. This indicates that the observed differences in education satisfaction were not only statistically significant but also stable and precise, reflecting a reliable improvement associated with the intervention (Table 4).

Table . Comparison of education satisfaction scores among 3 groups of hospitalized patients with lung cancer in a randomized controlled trial of artificial intelligence (AI)–assisted voice cloning education at The First Hospital of Lanzhou University (May 2025 through July 2025), with differences among groups analyzed using 1-way ANOVA.

Item	Control group (n=58), mean (SD; 95% CI)	Intervention group 1 (n=57), mean (SD; 95% CI)	Intervention group 2 (n=59), mean (SD; 95% CI)	*F* test (*df*)	*P* value
Education content	3.8 (0.7; 3.6 - 4.0)	4.3 (0.6; 4.1 - 4.5)	4.5 (0.5; 4.4 - 4.6)	21.364 (2, 171)	<.001
Education method	3.6 (0.8; 3.4 - 3.8)	4.5 (0.6; 4.3 - 4.7)	4.7 (0.4; 4.6 - 4.8)	52.781 (2, 171)	<.001
Education time	3.9 (0.6; 3.7 - 4.1)	4.2 (0.5; 4.1 - 4.3)	4.3 (0.5; 4.2 - 4.4)	9.526 (2, 171)	<.001
Education effect	3.7 (0.7; 3.5 - 3.9)	4.4 (0.5; 4.3 - 4.5)	4.6 (0.4; 4.5 - 4.7)	43.175 (2, 171)	<.001
Overall satisfaction	3.8 (0.6; 3.6 - 4.0)	4.4 (0.5; 4.3 - 4.5)	4.6 (0.4; 4.5 - 4.7)	45.293 (2, 171)	<.001

XSL•FO
RenderX

Comparison of Treatment Adherence

Among the 3 groups, significant differences in treatment adherence were observed across all dimensions, including drug treatment adherence, lifestyle adjustment, follow-up review, and total adherence score (all $P<.001$). For drug treatment adherence, the control group scored a mean of 7.6 (SD 1.5; 95% CI 7.2 - 8.0), which was lower than the mean scores for intervention group 1 (8.5, SD 1.2; 95% CI 8.2 - 8.8) and intervention group 2 (9.1, SD 0.9; 95% CI 8.9 - 9.3), with a statistically significant overall difference ($F_{2,171}=25.364$, $P<.001$). Similar patterns were observed for lifestyle adjustment and follow-up review adherence. The 95% CIs of both intervention groups were consistently higher than those of the control group, with limited overlap, indicating that the observed improvements in treatment adherence were stable and precise. Regarding the total adherence score, the control group achieved a mean of 21.6 (SD 4.2; 95% CI 20.5 - 22.7), whereas intervention group 1 and intervention group 2 achieved higher mean scores of 24.7 (SD 3.5; 95% CI 23.8 - 25.6) and 26.6 (SD 2.8; 95% CI 25.9 - 27.3), respectively. The between-group difference remained statistically significant ($F_{2,171}=32.593$, $P<.001$; Table 5).

Table . Comparison of treatment adherence scores among 3 groups of hospitalized patients with lung cancer in a randomized controlled trial of artificial intelligence (AI)–assisted voice cloning education at The First Hospital of Lanzhou University (May 2025 through July 2025), with differences among groups analyzed using 1-way ANOVA.

Item	Control group (n=58), mean (SD; 95% CI)	Intervention group 1 (n=57), mean (SD; 95% CI)	Intervention group 2 (n=59), mean (SD; 95% CI)	*F* test (*df*)	*P* value
Drug treatment	7.6 (1.5; 7.2 - 8.0)	8.5 (1.2; 8.2 - 8.8)	9.1 (0.9; 8.9 - 9.3)	25.364 (2, 171)	<.001
Lifestyle adjustment	6.8 (1.7; 6.4 - 7.2)	7.9 (1.4; 7.5 - 8.3)	8.6 (1.1; 8.3 - 8.9)	27.175 (2, 171)	<.001
Follow-up review	7.2 (1.6; 6.8 - 7.6)	8.3 (1.3; 8.0 - 8.6)	8.9 (1.0; 8.6 - 9.2)	26.482 (2, 171)	<.001
Total score	21.6 (4.2; 20.5 - 22.7)	24.7 (3.5; 23.8 - 25.6)	26.6 (2.8; 25.9 - 27.3)	32.593 (2, 171)	<.001

Comparison of Quality of Life

One month after discharge, significant differences were observed among the 3 groups in all SF-36 dimensions, including physical functioning, role-physical, bodily pain, general health, vitality, social functioning, role-emotional, and mental health (all $P<.001$). For each dimension, the 95% CIs of both intervention groups were consistently higher than those of the control group, indicating that the observed improvements in health-related quality of life were robust and precise.

Notably, intervention group 2 generally showed narrower 95% CIs, suggesting less variability and greater consistency in patient-reported outcomes compared with the control group (Table 6).

Table . Comparison of SF-36 quality of life scores 1 month after discharge among 3 groups of hospitalized patients with lung cancer in a randomized controlled trial of artificial intelligence (AI)–assisted voice cloning education at The First Hospital of Lanzhou University (May 2025 through July 2025), with differences among groups analyzed using 1-way ANOVA.

Dimension	Control group (n=58), mean (SD; 95% CI)	Intervention group 1 (n=57), mean (SD; 95% CI)	Intervention group 2 (n=59), mean (SD; 95% CI)	*F* test (*df*)	*P* value
Physical functioning	65.3 (12.5; 61.7 - 68.9)	72.6 (11.3; 69.5 - 75.7)	74.8 (10.9; 71.8 - 77.8)	11.364 (2, 171)	<.001
Role-physical	58.6 (14.2; 54.4 - 62.8)	67.5 (12.8; 63.8 - 71.2)	70.3 (12.1; 66.9 - 73.7)	13.175 (2, 171)	<.001
Bodily pain	67.2 (13.1; 63.3 - 71.1)	74.8 (11.9; 71.5 - 78.1)	76.5 (11.2; 73.3 - 79.7)	10.482 (2, 171)	<.001
General health	61.5 (12.8; 57.7 - 65.3)	69.7 (11.5; 66.5 - 72.9)	72.4 (10.8; 69.3 - 75.5)	14.293 (2, 171)	<.001
Vitality	59.8 (13.5; 55.8 - 63.8)	68.3 (12.2; 65.1 - 71.5)	71.6 (11.5; 68.5 - 74.7)	15.364 (2, 171)	<.001
Social functioning	63.7 (12.9; 59.8 - 67.6)	71.5 (11.7; 68.4 - 74.6)	76.2 (10.8; 73.2 - 79.2)	18.175 (2, 171)	<.001
Role-emotional	60.2 (14.1; 56.0 - 64.4)	69.8 (12.5; 66.3 - 73.3)	75.3 (11.7; 72.1 - 78.5)	21.482 (2, 171)	<.001
Mental health	62.5 (13.2; 58.7 - 66.3)	70.6 (12.1; 67.3 - 73.9)	76.8 (11.3; 73.8 - 79.8)	22.293 (2, 171)	<.001

Comparison of Anxiety and Depression

Before education, there were no significant differences in anxiety nor depression scores among the 3 groups (anxiety: $F_{2,171}=0.068$, $P=.93$; depression: $F_{2,171}=0.185$, $P=.83$), indicating baseline comparability. One month after discharge, significant differences were observed among the groups for both anxiety and depression scores ($P<.001$). Specifically, for anxiety, the control group scored a mean of 6.5 (SD 2.9; 95% CI 5.8 - 7.2), intervention group 1 scored a mean of 5.2 (SD 2.5; 95% CI 4.6 - 5.8), and intervention group 2 scored a mean of 4.3 (SD 2.2; 95% CI 3.8 - 4.8). For depression, the mean scores were 5.8 (SD 2.6; 95% CI 5.2 - 6.4), 4.6 (SD 2.3; 95% CI 4.1 - 5.1), and 3.7 (SD 2.0; 95% CI 3.3 - 4.1), respectively.

The 95% CIs of the intervention groups did not overlap with those of the control group, indicating that the reductions in

XSL•FO
RenderX

anxiety and depression after intervention were stable and precise (Table 7).

Table . Comparison of Hospital Anxiety and Depression Scale (HADS) anxiety and depression scores among 3 groups of hospitalized patients with lung cancer in a randomized controlled trial of artificial intelligence (AI)–assisted voice cloning education at The First Hospital of Lanzhou University (May 2025 through July 2025), with differences among groups analyzed using 1-way ANOVA.

Time point	Control group (n=58), mean (SD; 95% CI)	Intervention group 1 (n=57), mean (SD; 95% CI)	Intervention group 2 (n=59), mean (SD; 95% CI)	*F* test (*df*)	*P* value
Pre-education anxiety	7.8 (3.2; 7.0-8.6)	7.6 (3.1; 6.8-8.4)	7.7 (3.3; 6.9-8.5)	0.068 (2, 171)	.93
Pre-education depression	6.9 (2.8; 6.2-7.6)	7.1 (2.9; 6.4-7.8)	6.8 (2.7; 6.1-7.5)	0.185 (2, 171)	.83
1 month after discharge anxiety	6.5 (2.9; 5.8-7.2)	5.2 (2.5; 4.6-5.8)	4.3 (2.2; 3.8-4.8)	12.364 (2, 171)	<.001
1 month after discharge depression	5.8 (2.6; 5.2-6.4)	4.6 (2.3; 4.1-5.1)	3.7 (2.0; 3.3-4.1)	13.175 (2, 171)	<.001

Discussion

Principal Findings

This RCT investigated the effectiveness of an AI-assisted patient education system integrating voice cloning technology and compared the educational outcomes of physician voice cloning versus patient self-voice cloning. In addition, the study explored the feasibility of using ChatGPT as a supportive tool to assist with evaluating education effectiveness. Consistent with our 3 prespecified hypotheses, the results demonstrated that AI-assisted voice cloning–based education significantly improved education content compliance, knowledge mastery, satisfaction, treatment adherence, and short-term psychological and quality-of-life outcomes compared with traditional education. These findings are consistent with prior research indicating that technology-assisted and personalized education can enhance patient engagement and learning outcomes [10,22]. Notably, education delivered using patients' own cloned voices yielded superior effects compared with physician voice cloning across multiple outcome domains, supporting the hypothesis that self-referential personalization enhances educational effectiveness.

Interpretation of Findings and Comparison With Previous Studies

Voice cloning technology represents an emerging application of AI in patient education, offering distinct advantages over conventional education models. By enabling repeated exposure to standardized educational content delivered in a familiar and emotionally resonant voice, voice cloning may enhance attention, comprehension, and memory consolidation [30,31]. Consistent with previous studies demonstrating that familiar auditory cues improve trust and acceptance in health communication, both AI-assisted intervention groups in this study outperformed the traditional education group in education compliance, knowledge mastery, and satisfaction [30,32,33]. Beyond confirming the general effectiveness of voice-based AI education, this study provides novel evidence regarding differences between voice sources. Patients who received education via their own cloned voice achieved higher compliance, satisfaction, knowledge retention, treatment adherence, and better short-term psychological outcomes than those educated using physician voice cloning. Previous studies have largely focused on clinician-narrated or professionally recorded content [34], whereas empirical comparisons involving patient self-voice have been scarce. The observed superiority of self-voice education may be explained by the self-reference effect, whereby information related to the self is processed more deeply and remembered more effectively [17,35,36]. Research in cognitive psychology suggests that self-related stimuli—particularly self-generated or self-similar auditory information—enhance attention, emotional engagement, and memory encoding [16,35]. In this study, hearing one's own voice narrate medical information may have strengthened self-identification and personal relevance, thereby reinforcing learning and adherence. Furthermore, the self-voice group demonstrated superior outcomes in multiple quality-of-life domains and had greater reductions in anxiety and depression. These findings align with evidence that personalized and self-relevant health communication can enhance perceived control, self-efficacy, and emotional regulation among patients [14,36]. Importantly, the reported effect sizes indicated moderate-to-large intervention effects, suggesting that the observed improvements were not only statistically significant but also clinically meaningful. Reporting both effect sizes and confidence intervals enhances transparency and interpretability and is consistent with CONSORT 2025 recommendations for RCTs [28].

Implications for AI-Assisted Evaluation of Patient Education

This study also explored the use of ChatGPT as an auxiliary tool to assist with evaluating patient education outcomes based on standardized educational content. Pre-experimental testing demonstrated good agreement between ChatGPT-assisted scoring and expert evaluation, suggesting that LLMs may support structured and reproducible assessment under controlled conditions. These findings are consistent with emerging literature indicating that LLMs can perform semantic analysis and text evaluation tasks with acceptable reliability in medical and educational contexts [6,37]. Compared with traditional manual evaluation, AI-assisted assessment may offer advantages in efficiency, scalability, and consistency, particularly in settings with limited human resources. However, ChatGPT should not be regarded as a fully objective evaluator. Assessment outcomes

XSL•FO
RenderX

may still be influenced by prompt design, model architecture, and algorithmic biases. Therefore, AI-assisted evaluation should be considered a complementary tool rather than a replacement for expert assessment, and its use should be accompanied by appropriate human oversight [38,39].

Study Limitations

Several limitations should be acknowledged. First, the follow-up period was limited to 1 month, restricting conclusions regarding long-term adherence, psychological outcomes, and sustained quality of life benefits. Longer follow-up periods are needed to assess the durability of intervention effects, as recommended in previous digital health education studies [40]. Second, this was a single-center study involving hospitalized patients with relatively homogeneous educational backgrounds, which may limit generalizability. Prior research suggests that educational level and health literacy can influence the effectiveness of digital health interventions [41]. Multicenter studies with more diverse populations are therefore warranted. Third, although voice cloning technology offers a promising approach, current speech synthesis systems still have limitations in emotional nuance and naturalness, which may affect user engagement [13]. Continued technological advancements may further enhance the effectiveness of AI-assisted education. Finally, although AI-assisted education may reduce clinical workload and improve efficiency, this study did not include a formal cost-effectiveness analysis. Future research should systematically evaluate economic outcomes to support large-scale implementation decisions [40]. Because ChatGPT is a continuously updated product, we standardized the scoring workflow (fixed rubric and prompts, prespecified audio format, and single scoring per recording) and provide the complete scoring materials to facilitate reproducibility. Additionally, voice-based delivery systems may misinterpret medication names or specialized terminology, underscoring the need for clinician review of generated content [42].

Implications for Practice and Policy

Implementation should address unit- and hospital-level factors separately. Targeted strategies should improve staff training, digital access, and policy support. Addressing demographic disparities is key to promoting equity and care quality.

Conclusions and Broader Implications

This study introduces an innovative patient education model integrating AI voice cloning and ChatGPT, representing a novel approach distinct from previous studies that primarily relied on standard text-to-speech or professionally recorded content. The key innovation lies in using patients' own cloned voices for health education delivery, leveraging the self-reference effect to enhance learning outcomes. Compared with prior research focusing on clinician-narrated content, this study provides the first empirical evidence that self-voice education produces superior outcomes across multiple domains including compliance, satisfaction, and psychological well-being. These findings contribute to the field by establishing a theoretical and practical framework for personalized AI-driven patient education. In real-world clinical settings, this approach offers a scalable, cost-effective solution to enhance patient engagement, particularly valuable in resource-limited environments where individualized education is challenging to deliver. Future research should focus on multicenter validation, longer follow-up periods, and exploration of optimal voice cloning parameters to maximize educational effectiveness.

Acknowledgments

The authors thank all individuals who contributed to this study. Appreciation is extended to the patients and clinical staff of the First Hospital of Lanzhou University for their participation and cooperation throughout the research process. The authors also acknowledge the technical support provided by the development team for the artificial intelligence (AI) health education and voice cloning systems, led by author MM. The authors further thank all collaborators who assisted with data collection, statistical analysis, and manuscript preparation. Institutional support from the Department of Thoracic Surgery, The First Hospital of Lanzhou University, is also acknowledged for providing research resources and facilities.

In accordance with the Generative AI Delegation Taxonomy (GAIDeT; 2025), generative AI tools were used under full human supervision for proofreading, language editing, and translation. The tool used was ChatGPT (GPT-4; OpenAI). All AI-generated content was reviewed and validated by the authors. Responsibility for the final manuscript rests entirely with the authors. Generative AI tools are not listed as authors and did not contribute to study design, data analysis, or interpretation.

Funding

This study was supported by the Key Research and Development Program of Gansu Province (26YFFA004); Natural Science Foundation of Gansu Province (23JRRA1597); the 2024 Key Research and Development Program of Gansu Province–International Cooperation Field (24YFWA011), and the Institutional Research Fund of The First Hospital of Lanzhou University (ldyyyn2023-63).

Data Availability

The datasets generated and analyzed during this study are not publicly available due to privacy and ethical restrictions related to patient information and institutional regulations. Deidentified data are stored securely at The First Hospital of Lanzhou University and are available from the corresponding author upon reasonable request and approval by the hospital's ethics committee (approval number: ldyyyn2023-63).

XSL•FO
RenderX

Authors' Contributions

Conceptualization: YS (lead), HJ (supporting), YZ (supporting)
Data curation: SX (lead), XH (supporting), KJ (supporting), YZ (supporting)
Formal analysis: YS (equal), SX (equal), HJ (equal)
Methodology: YS (lead), SX (lead), HJ (supporting), XH (supporting), KJ (supporting), MM (technical lead)
Project administration: XM (lead)
Supervision: HW (lead)
Validation: YS (lead)
Visualization: SX (lead), XH (supporting)
Writing - original draft: YS (lead), SX (supporting), MM (advising)
Writing - review & editing: YS (supporting), SX (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
ChatGPT-based compliance scoring: prompts, rubric, and example inputs/outputs.
[PDF File, 172 KB - jmir_v28i1e81387_app1.pdf]

Multimedia Appendix 2
Education content compliance, knowledge mastery, and treatment adherence questionnaires.
[PNG File, 187 KB - jmir_v28i1e81387_app2.png]

Multimedia Appendix 3
SF-36 questionnaire.
[PNG File, 784 KB - jmir_v28i1e81387_app3.png]

Multimedia Appendix 4
Hospital Anxiety and Depression Scale.
[PNG File, 797 KB - jmir_v28i1e81387_app4.png]

Multimedia Appendix 5
Voice cloning patient education prompts, ChatGPT evaluation prompts, example patient knowledge assessment sheet, and notes for researchers.
[DOCX File, 1029 KB - jmir_v28i1e81387_app5.docx]

Checklist 1
CONSORT 2025 checklist.
[PDF File, 154 KB - jmir_v28i1e81387_app6.pdf]

Checklist 2
CONSORT-EHEALTH checklist (V 1.6.1).
[PDF File, 202 KB - jmir_v28i1e81387_app7.pdf]

References

1. Kocaballi AB, Quiroz JC, Rezazadegan D, et al. Responses of conversational agents to health and lifestyle prompts: investigation of appropriateness and presentation structures. J Med Internet Res 2020 Feb 9;22(2):e15823. [doi: 10.2196/15823] [Medline: 32039810]
2. Zolnierek KBH, Dimatteo MR. Physician communication and patient adherence to treatment: a meta-analysis. Med Care 2009 Aug;47(8):826-834. [doi: 10.1097/MLR.0b013e31819a5acc] [Medline: 19584762]
3. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 6;4(2):e19. [doi: 10.2196/mental.7785] [Medline: 28588005]
4. Shersher V, Haines TP, Sturgiss L, Weller C, Williams C. Definitions and use of the teach-back method in healthcare consultations with patients: a systematic review and thematic synthesis. Patient Educ Couns 2021 Jan;104(1):118-129. [doi: 10.1016/j.pec.2020.07.026] [Medline: 32798080]

XSL•FO
RenderX

5. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature New Biol 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
6. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
7. Denecke K, Gabarron E, Grainger R, et al. Artificial intelligence for participatory health: applications, impact, and future implications. Yearb Med Inform 2019 Aug;28(1):165-173. [doi: 10.1055/s-0039-1677902] [Medline: 31022749]
8. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
9. Milne-Ives M, de Cock C, Lim E, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res 2020 Oct 22;22(10):e20346. [doi: 10.2196/20346] [Medline: 33090118]
10. Schachner T, Keller R, V Wangenheim F. Artificial intelligence-based conversational agents for chronic conditions: systematic literature review. J Med Internet Res 2020 Sep 14;22(9):e20701. [doi: 10.2196/20701] [Medline: 32924957]
11. Palanica A, Flaschner P, Thommandram A, Li M, Fossat Y. Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. J Med Internet Res 2019 Apr 5;21(4):e12887. [doi: 10.2196/12887] [Medline: 30950796]
12. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun 28;25:e48568. [doi: 10.2196/48568] [Medline: 37379067]
13. Sisman B, Yamagishi J, King S, Li H. An overview of voice conversion and its challenges: from statistical modeling to deep learning. IEEE/ACM Trans Audio Speech Lang Process 2020;29:132-157. [doi: 10.1109/TASLP.2020.3038524]
14. Dai G, Wang P, Yao C, Ying F. InnerSelf: designing self-deepfaked voice for emotional well-being. arXiv. Preprint posted online on Mar 18, 2025 URL: https://arxiv.org/abs/2503.14257 [accessed 2026-03-03] [doi: 10.48550/arXiv.2503.14257]
15. Nygaard LC, Pisoni DB. Talker-specific learning in speech perception. Percept Psychophys 1998 Apr;60(3):355-376. [doi: 10.3758/bf03206860] [Medline: 9599989]
16. Kaplan JT, Aziz-Zadeh L, Uddin LQ, Iacoboni M. The self across the senses: an fMRI study of self-face and self-voice recognition. Soc Cogn Affect Neurosci 2008 Sep;3(3):218-223. [doi: 10.1093/scan/nsn014] [Medline: 19015113]
17. Yonan CA, Sommers MS. The effects of talker familiarity on spoken word identification in younger and older listeners. Psychol Aging 2000;15(1):88-99. [doi: 10.1037/0882-7974.15.1.88]
18. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne) 2024;11:1477898. [doi: 10.3389/fmed.2024.1477898] [Medline: 39534227]
19. Busch F, Hoffmann L, Rueger C, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med (Lond) 2025 Jan 21;5(1):26. [doi: 10.1038/s43856-024-00717-2] [Medline: 39838160]
20. OpenAI, Achiam J, Adler S, et al. GPT-4 technical report. arXiv. Preprint posted online on Mar 15, 2023 URL: https://arxiv.org/abs/2303.08774 [accessed 2026-03-03] [doi: 10.48550/arXiv.2303.08774]
21. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
22. Chavez-Yenter D, Kimball KE, Kohlmann W, et al. Patient interactions with an automated conversational agent delivering pretest genetics education: descriptive study. J Med Internet Res 2021 Nov 18;23(11):e29447. [doi: 10.2196/29447] [Medline: 34792472]
23. Bickmore TW, Pfeifer LM, Byron D, et al. Usability of conversational agents by patients with inadequate health literacy: evidence from two clinical trials. J Health Commun 2010;15 Suppl 2:197-210. [doi: 10.1080/10810730.2010.499991] [Medline: 20845204]
24. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study. Digit Health 2019;5:2055207619871808. [doi: 10.1177/2055207619871808] [Medline: 31467682]
25. Sezgin E, Huang Y, Ramtekkar U, Lin S. Readiness for voice assistants to support healthcare delivery during a health crisis and pandemic. NPJ Digit Med 2020;3:122. [doi: 10.1038/s41746-020-00332-0] [Medline: 33015374]
26. Sezgin E, Militello LK, Huang Y, Lin S. A scoping review of patient-facing, behavioral health interventions with voice assistant technology targeting self-management and healthy lifestyle behaviors. Transl Behav Med 2020 Aug 7;10(3):606-628. [doi: 10.1093/tbm/ibz141] [Medline: 32766865]
27. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Presented at: Advances in Neural Information Processing Systems 33 (NeurIPS 2020); Dec 6-12, 2020; Virtual-only conference p. 1877-1901 URL: https://papers.nips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html [accessed 2026-03-03]
28. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ 2025 Apr 14;389:e081123. [doi: 10.1136/bmj-2024-081123] [Medline: 40228833]
29. Eysenbach G, Group CE. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
30. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]

XSL•FO
RenderX

31. Battineni G, Chintalapudi N, Amenta F. AI chatbot design during an epidemic like the novel coronavirus. Healthcare (Basel) 2020 Jun 3;8(2):154. [doi: 10.3390/healthcare8020154] [Medline: 32503298]
32. Nayak A, Vakili S, Nayak K, et al. Use of voice-based conversational artificial intelligence for basal insulin prescription management among patients with type 2 diabetes: a randomized clinical trial. JAMA Netw Open 2023 Dec 1;6(12):e2340232. [doi: 10.1001/jamanetworkopen.2023.40232] [Medline: 38039007]
33. Bickmore TW, Pfeifer LM, Paasche-Orlow MK. Using computer agents to explain medical documents to patients with low health literacy. Patient Educ Couns 2009 Jun;75(3):315-320. [doi: 10.1016/j.pec.2009.02.007] [Medline: 19297116]
34. Laranjo L, Dunn AG, Tong HL, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018 Sep 1;25(9):1248-1258. [doi: 10.1093/jamia/ocy072] [Medline: 30010941]
35. Sui J, Humphreys GW. The integrative self: how self-reference integrates perception and memory. Trends Cogn Sci 2015 Dec;19(12):719-728. [doi: 10.1016/j.tics.2015.08.015] [Medline: 26447060]
36. Rogers TB, Kuiper NA, Kirker WS. Self-reference and the encoding of personal information. J Pers Soc Psychol 1977;35(9):677-688. [doi: 10.1037/0022-3514.35.9.677]
37. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
38. Sun GH, Hoelscher SH. The ChatGPT storm and what faculty can do. Nurse Educ 2023;48(3):119-124. [doi: 10.1097/NNE.0000000000001390] [Medline: 37043716]
39. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
40. Murray E, Hekler EB, Andersson G, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016 Nov;51(5):843-851. [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
41. Paasche-Orlow MK, Wolf MS. The causal pathways linking health literacy to health outcomes. Am J Health Behav 2007;31 Suppl 1(1):S19-S26. [doi: 10.5555/ajhb.2007.31.supp.S19] [Medline: 17931132]
42. Palanica A, Thommandram A, Lee A, Li M, Fossat Y. Do you understand the words that are comin outta my mouth? Voice assistant comprehension of medication names. NPJ Digit Med 2019;2:55. [doi: 10.1038/s41746-019-0133-x] [Medline: 31304401]

Abbreviations

AI: artificial intelligence
CONSORT: Consolidated Standards of Reporting Trials
ITT: intention to treat
LLM: large language model
RCT: randomized controlled trial

Edited by S Brini; submitted 28.Jul.2025; peer-reviewed by A Khamees, B Meskó; accepted 17.Jan.2026; published 19.Mar.2026.

Please cite as:
Sun Y, Xu S, Jin H, Han X, Jin K, Zhang Y, Ma X, Wei H, Ma M
Effectiveness of AI-Assisted Patient Health Education Using Voice Cloning and ChatGPT: Prospective Randomized Controlled Trial
J Med Internet Res 2026;28:e81387
URL: https://www.jmir.org/2026/1/e81387
doi:10.2196/81387

© Yan Sun, Shangqing Xu, Hongying Jin, Xiaoyan Han, Kangqi Jin, Yimei Zhang, Xiaoli Ma, Huaping Wei, Minjie Ma. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Psychotherapists' Trust, Distrust, and Generative AI Practices in Psychotherapy: Qualitative Study

Jinyan Kuang[1], BS; Amber L Pope[2], PhD; Yixuan Zhang[1], PhD

[1]Department of Computer Science, William & Mary, 251 Jamestown Road, Williamsburg, VA, United States

[2]Department of School Psychology and Counselor Education, William & Mary, Williamsburg, VA, United States

Corresponding Author:
Yixuan Zhang, PhD
Department of Computer Science, William & Mary, 251 Jamestown Road, Williamsburg, VA, United States

Abstract

Background: Generative artificial intelligence (GenAI) is increasingly used in mental health care, from client-facing chatbots to clinician-facing documentation aids. Psychotherapists' willingness to rely on—or withhold reliance from—these tools has significant implications for care quality, yet little is known about how practicing clinicians calibrate trust and distrust in GenAI across tasks and contexts. Given that the therapeutic relationship is central to psychotherapy outcomes, understanding how GenAI intersects with this relational foundation is essential for responsible integration.

Objective: This study aims to examine (1) psychotherapists' experiences with, perceptions of, and trust or distrust in GenAI in therapeutic contexts and (2) how they perceive the role of GenAI within the therapeutic relationship and how their perceptions shape their trust and distrust in GenAI.

Methods: We conducted a qualitative interview study using semistructured interviews with 18 actively practicing psychotherapists in the United States between January and May 2025. Participants were recruited through professional mailing lists, social media, and snowball sampling. Interviews (≈60 min each) were conducted via Zoom and explored psychotherapists' experiences with, perceptions of, and trust or distrust in GenAI in therapeutic contexts. Data were analyzed using the general inductive approach, with iterative coding and team-based interpretation to identify themes.

Results: Our findings show that psychotherapists' GenAI adoption was highly individualized and contingent on maintaining professional role integrity—not merely technical oversight. Trust was sustained when GenAI operated in clinician-supervised, supportive roles for low-stakes tasks (eg, documentation and brainstorming), but diminished when control shifted, tasks involved high-stakes clinical judgment, or GenAI threatened to encroach on the authentic human connection central to therapy. Participants articulated conditions for trust that went beyond "human-in-the-loop" monitoring to include preservation of interpretive authority, ethical responsibility, and relational primacy. Distrust also extended to the broader sociotechnical ecosystem, including concerns about commercial incentives, insurance pressures, and the absence of clear organizational guidelines.

Conclusions: Psychotherapists' perspectives offer critical insights into GenAI's current usages in their professional practices and the conditions under which they are willing to trust and distrust GenAI tools. Their experiences highlight the importance of maintaining clinician control, ensuring contextual appropriateness, and preserving the human connection central to psychotherapy. Future work should further examine how therapeutic orientation, professional experience, and client characteristics shape trust and distrust in GenAI. As GenAI becomes more embedded in mental health care, research is also needed to explore how specific GenAI system features can be responsibly designed to support clinical workflows and enhance therapeutic relationships. Organizational and policy frameworks will be essential to ensure responsible, ethically aligned, and human-centered GenAI deployment in psychotherapy.

(J Med Internet Res 2026;28:e88932) doi:10.2196/88932

KEYWORDS

artificial intelligence; AI; generative AI; mental health; trust; qualitative methods

Introduction

The rapid advancement of generative artificial intelligence (GenAI) in mental health care has sparked both enthusiasm and concern in recent years [1-4]. Earlier mental health artificial intelligence (AI) tools were largely rule-based or template-driven, offering structured guidance [5]. With recent advancements in large language models, professional organizations have begun to respond. For example, the American Psychological Association (APA) is developing guidance to help psychologists integrate AI tools into practice in ways that respect ethical and legal obligations [6]. Yet a recent survey found that 71% of mental health practitioners had not

XSL•FO
RenderX

incorporated AI-enabled tools into routine care [7]. Public and social-media discourse often portrays psychotherapists as broadly resistant to or distrustful of GenAI—casting the technology as incompatible with therapeutic work or as a deceptive substitute for human care, especially when chatbots impersonate psychotherapists [8]. These portrayals, however, say little about how practicing psychotherapists themselves are actively calibrating trust and distrust in GenAI in day-to-day clinical contexts.

Empirical work has begun to examine mental health professionals' attitudes toward AI and GenAI. Survey and interview studies have shown mental health practitioners' perceptions of AI in services [9-11], including views on the public's use of AI chatbots for mental health support [12]. This body of work has identified perceived benefits for self-management and accessibility alongside concerns about data privacy, clinical reliability, and limited empathy. Complementary design research has explored early GenAI-augmented prototypes (eg, counselor agents and multimodal assistants) to probe possible roles for GenAI in assessment, client participation, and skill practice [13]. Together, existing work illuminates general attitudes and speculative futures, but it offers less insight into how psychotherapists currently use GenAI in practice, which concrete tasks they deem trustable or untrustable, and how they respond when clients independently bring GenAI into therapy.

Research on trust in AI provides a useful lens for these questions. Trust in AI is often defined as a belief that a system is reliable, competent, transparent, and responsive in a given context [14], whereas distrust reflects expectations of harm, manipulation, or misuse [15]. A large body of work shows that trust in automation is situational and conditional: people are more willing to rely on AI when conditions around evidence, oversight, and control are satisfied, and they withdraw trust when those conditions are violated [16-21]. Trust is thus not a fixed trait but a dynamic, task- and role-dependent judgment that evolves with experience. In high-stakes domains, miscalibrated trust—either overtrust that leads to uncritical reliance or undertrust that prevents appropriate use—can produce safety risks and missed opportunities [22].

Decades of research have established that the therapeutic relationship—often termed the "therapeutic alliance" or "working alliance"—is among the most important predictors of psychotherapy outcomes, frequently accounting for more variance than specific techniques or theoretical orientations [23]. The alliance encompasses agreement on therapeutic goals, consensus on tasks, and the affective bond between therapist and client [24]. Such a relational foundation makes psychotherapy a domain where trust is not merely a technical consideration but a fundamentally interpersonal one. Psychotherapy is thus inherently relational and context-sensitive [25,26], making questions of calibrated trust and distrust in GenAI especially salient. Misplaced trust could, for example, lead psychotherapists to overdelegate clinical reasoning, accept flawed recommendations, or input sensitive client data into insecure systems, undermining both care and confidentiality [27,28]. Insufficient trust, by contrast, may keep psychotherapists from using GenAI in low-risk areas where it could meaningfully augment care (eg, psychoeducation, homework support, or documentation) [29]. At the same time, decades of work on the therapeutic relationship underscores that the alliance between psychotherapist and client—the goals, tasks, and emotional bond of therapy—is central to outcomes [30,31]. Experiences with telehealth show that this relationship can adapt to new media but requires careful management of shifts in framing and presence [30]. As GenAI enters psychotherapy, the traditional dyadic relationship becomes a triad involving psychotherapist, client, and GenAI [32], raising open questions about GenAI's appropriate role, boundaries, and influence on trust among all 3 parties.

Despite growing attention to GenAI in mental health, we know relatively little about how practicing psychotherapists form, calibrate, and revise trust and distrust in GenAI across tasks and contexts, or how they understand GenAI's place in the evolving psychotherapist-client-GenAI relationship. In this work, we conducted a qualitative study with 18 psychotherapists to explore (1) What are psychotherapists' experiences with, perceptions of, and trust or distrust in GenAI in therapeutic contexts? and (2) How do psychotherapists perceive the role of GenAI within the therapeutic relationship, and how do these perceptions shape their trust and distrust in GenAI?

Methods

Study Design

This study used a qualitative interview method to explore how psychotherapists perceive and use GenAI tools in their professional practice, with particular attention to the dynamics of trust and distrust. A qualitative approach was selected because it enables an in-depth understanding of psychotherapists' lived experiences and contextualized decision-making around GenAI integration [33]. To do so, we conducted semistructured interviews via Zoom (Zoom Video Communications, Inc) between January and May 2025 and qualitative data analysis between May and July 2025 to further examine psychotherapists' practices, trust (or distrust) of GenAI, and their envisioned relationship between GenAI and psychotherapists.

Recruitment

Concerning the inclusion criteria, participants were eligible if they (1) held a valid license or credential to practice psychotherapy in the United States (eg, licensed professional counselor, licensed clinical social worker, and licensed psychologist), and (2) were actively providing psychotherapy to clients at the time of enrollment. No exclusion criteria were applied regarding therapeutic orientation, practice setting, years of experience, or prior familiarity with GenAI; we sought a diverse range of perspectives rather than a homogeneous sample.

Concerning participant recruitment, they were recruited through purposive and snowball sampling between January and May 2025. Recruitment advertisements were distributed via (1) professional mailing lists for psychotherapists and mental health practitioners, (2) LinkedIn posts targeting mental health professionals, and (3) professional networks of the research team. The recruitment message briefly described this study's

XSL•FO
RenderX

purpose (exploring psychotherapists' perspectives on GenAI), eligibility criteria, time commitment (≈60 min), and compensation. Interested individuals contacted the research team via email and were screened for eligibility. We received 20 expressions of interest; 18 completed interviews, and 2 did not attend their scheduled sessions (no follow-up information was available regarding reasons for nonattendance).

Interview Study Procedures

The semistructured interview guide was developed iteratively by the research team, drawing on (1) the existing literature on trust in AI and automation [21,22,34], (2) prior qualitative work on mental health practitioners' attitudes toward AI [9,12], and (3) the research team's clinical and human-computer interaction expertise. The guide was pilot-tested with 2 psychotherapists (not included in the final sample) to assess question clarity and flow, and minor revisions were made based on their feedback. The final guide included open-ended questions organized around three domains: (1) current experiences with GenAI in clinical practice, (2) perceptions of trust and distrust in GenAI across tasks and contexts, and (3) views on GenAI's role in the therapeutic relationship. See Multimedia Appendix 1 for the complete interview protocol.

Before each interview, we asked each participant about their experience with GenAI and demographic information. If participants indicated previous GenAI experience, they were then prompted to elaborate on their experiences with GenAI use, followed by questions about their attitudes and trust or distrust of GenAI. When a participant had no prior GenAI experience, we first shared a web link from the APA [35] that outlines the potential clinical uses of GenAI and the associated ethical guidelines. The discussion then focused on the participants' reactions and their anticipated levels of trust or distrust. In each interview, we also asked about views on the independent use of GenAI tools by clients and their perceptions of the relationship among the psychotherapist, GenAI, and client. The interview sessions lasted approximately 60 minutes. See Multimedia Appendix 1 for the detailed interview questions.

Data Analysis

All interview sessions were transcribed verbatim and checked for accuracy against the original recordings. We conducted a qualitative analysis with the transcripts guided by the general inductive approach [36]. First, the lead author carefully read the transcripts multiple times to become familiar with the data and the whole brief analytic memos to capture initial impressions about psychotherapists' GenAI practices, perceived benefits and risks, and expressions of trust and distrust of GenAI in different clinical contexts.

In the first cycle of coding, the lead author conducted line-by-line coding, assigning low-level, descriptive codes to segments of text that captured (1) concrete uses of GenAI in practice including both psychotherapists' practices (eg, documentation, homework support, and supervision) and their observations of their clients' usages (eg, journaling and emotional support), (2) conditions under which psychotherapists felt trust or distrust in using GenAI, and (3) how they perceived GenAI's role in the psychotherapist-client relationship. Codes were kept close to participants' language and iteratively refined as additional transcripts were reviewed. A provisional code list was updated after every 2 - 3 transcripts to merge redundant codes, clarify labels, and add new codes that reflect concepts not captured previously.

Next, low-level codes were grouped into more abstract categories that summarized patterned meanings across participants and connected related codes (eg, "emergence of trust in GenAI for low-stakes administrative tasks," "determining trust or distrust based on boundaries around high-stakes clinical reasoning," "enhancing trust in guided client use of GenAI between sessions," and "distrust about data privacy and commercialization"). Through constant comparison within and across transcripts, these categories were iteratively collapsed and reorganized into higher-level themes that described (1) psychotherapists' current GenAI engagement, (2) how professional autonomy and task type set boundaries for trust or distrust, and (3) how GenAI is positioned within the therapeutic relationship, which aligns with their trust or distrust.

Throughout the coding process, 2 additional members of the research team independently coded overlapping subsets of transcripts using the evolving codebook and then met with the lead author in weekly analytic meetings to compare interpretations. Discrepancies in coding were discussed and resolved through group discussion until consensus was reached, and the shared codebook was iteratively updated to reflect agreed-upon definitions and illustrative coding examples. As analysis progressed, the team regularly compared new transcripts against the existing codebook to ensure consistent application of codes while remaining open to emergent concepts. Data collection and analysis proceeded concurrently and continued until thematic saturation was reached, defined as the point at which no substantively new codes or themes emerged in successive interviews and additional transcripts primarily confirmed the existing thematic structure. This study adhered to the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist [37]. See Checklist 1 for the detailed checklist.

Ethical Considerations

Ethical approval for this study was obtained from the Institutional Review Board at William & Mary (PHSC-2024-06-30-17113). Before participation, the research team explained this study's purpose, procedures, potential risks, and benefits to all participants. Written informed consent was obtained via participants' digital signatures before each interview. Each participant was assigned a unique study ID upon enrollment, and all identifying information was removed to ensure confidentiality. At the end of this study, each participant received a US $20 gift card as compensation upon completing the interview.

Data Availability

The qualitative data (interview transcripts) generated during this study are not publicly available due to privacy and confidentiality concerns; transcripts contain sensitive information that could identify participants despite deidentification efforts. The interview guide is provided in

XSL•FO
RenderX

Multimedia Appendix 1. Requests for access to deidentified data excerpts or the codebook may be directed to the corresponding author and will be considered on a case-by-case basis, subject to institutional review board approval and establishment of a data use agreement.

Results

Overview

Psychotherapists in our interview study (see a summary of participants' demographics information in Table 1) reported experimenting with GenAI mainly in supporting, nonclient-facing tasks, using it as an assistant rather than a psychotherapist substitute (see a summary of usage contexts in Table 2). Psychotherapists described using GenAI in several practical ways. Some used GenAI to draft session notes from generic prompts (eg, "therapy note, client struggling in marriage"). Others explored GenAI-driven homework support to assist clients between sessions or used GenAI as a communication bridge when clients' primary language differed from their own, allowing bilingual or multilingual clients to better articulate their thoughts and improving access for those who cannot easily find language-matched providers. Psychotherapists also applied GenAI for psychoeducation, using it to generate "simulated client" scenarios and offering more realistic and diverse practice opportunities than classmates. Participants valued GenAI's potential to enhance efficiency, extend therapeutic reach, and diversify resources—yet they consistently paired these ideas with caveats about trust, control, and ethical boundaries, setting the stage for the next section.

Table . Overview of participants' demographics information.

Dimension and response options	Interview participants (n=18), n
Working experiences[a] (years) [38]	
1 - 10	14
11 - 20	2
21 - 30	1
31 or more	1
Gender	
Woman	14
Man	4
Nonbinary	N/A[b]
Race or ethnicity	
Asian	2
Black or African American	3
White	13
Prefer to self-describe	1

[a]Mean 9.7, SD 8.5 years; median 7.5, IQR 13 years.

[b]N/A: not applicable.

XSL•FO
RenderX

Table . Contexts and example quotes of GenAI[a] use among psychotherapists.

Usage contexts	PIDs[b]	Example quotes of participants' GenAI use
Administrative and clinical support		
Clinical documentation	P02, P04, P07, P11	"I get [clients'] consent and then it records the session; from that transcript it will generate my note, and I edit that note and put it into my EHR."
Translation and language	P01, P05, P16	"I had ChatGPT translate a breathing exercise into Spanish for one of my clients."
Treatment plan formulation	P11, P14, P17	"I've used GenAI the most for wording a treatment plan goal."
Personalized homework creation	P01, P02, P03, P05, P09, P13	"I asked ChatGPT to make me a worksheet [that contains] information from books that I think provide a good model and explain it in layman's terms versus clinical terms and ask ChatGPT to make reflection questions for my client."
Training and supervision		
Ideation and brainstorming	P03, P06, P13, P18	"I need kind of a brainstorming colleague thing that I have [Gen]AI help me create content, like I might say, please give me 3 paragraphs of a case conceptualization for a client who's experiencing 3 anxiety symptoms."
Roleplay scenarios and scripts	P01, P11, P15	"Give me a real textbook kind of description of a persona that I can use in my training materials."
Supervision and clinical consultation	P09	"I've even considered if I could run a case by a [Gen]AI to get a sort of second opinion, almost like an 'AI supervisor' giving me feedback."
Self-help and emotional support	P06, P11, P12	"I kind of used a GenAI [chatbot] for me personally, when I was going through a difficult time where I interacted with the GenAI to get support."

[a]GenAI: generative artificial intelligence.

[b]PID: participant ID.

Based on their practices in GenAI in therapeutic contexts, we further identified 2 key domains that shape their trust and distrust toward GenAI in therapeutic contexts. First, psychotherapists articulated conditional trust: trust in GenAI that was bounded by professional autonomy, task specificity, and careful oversight. Second, they reflected on GenAI's position within the therapeutic relationship, weighing scenarios where it might enhance client engagement against fears it could erode the authentic human bonds central to psychotherapy.

Theme 1: Professional Autonomy Sets Boundaries for Trust

Overview

Theme 1 addresses psychotherapists' trust in GenAI for their own professional use—that is, when they themselves use GenAI tools in clinical workflows. Our participants conveyed that their trust in GenAI was conditional—bounded by a sense of self-regulation, task-specific usage, and personal limits. For psychotherapists, the relevant conditions revolve around safeguarding the therapeutic frame, protecting client confidentiality, and preserving the clinician's professional agency. In practice, psychotherapists' trust in GenAI remains high only when they can maintain direct control over how the technology is used. When that control slips, such as using GenAI for unfamiliar tasks or allowing others to dictate its use, their trust quickly diminishes.

Subtheme 1: "Hand on the Wheel" Rule

Participants highlighted that their trust in GenAI systems did not stem from the technology itself, but from their perceived ability to manage and control [Gen]AI's involvement. Psychotherapists in our study articulated a personal "hand on the wheel" "boundary" for trust*:* trust arose primarily when they could directly influence the GenAI's actions through careful prompting, reviewing outputs, and maintaining options to delete or redact any generated content. In other words, psychotherapists' trust was grounded less in confidence in GenAI itself than in confidence in their own professional ability to control the system: their judgment, expertise, and responsibility. For example, 1 psychotherapist (P03) tied her trust to her ability to steer the GenAI with well-crafted prompts and usage limits:

> *I trust myself to be a good prompter... I know my boundaries and I know what I'm comfortable doing and not doing. And so as long as I stick within my*

XSL•FO
RenderX

> *boundaries, I feel pretty comfortable using [Gen]AI, kind of accomplishing the goals that I have.* [P03]

P03's quote illustrates that trust is contingent not just on the quality or performance of the GenAI system, but rather on the clinician's professional autonomy and self-assessment of their capability to manage interactions with GenAI appropriately. Crucially, this sense of control applied not only to what information she fed into the system, but also over the fate of the outputs: where the text was stored, how long it persisted, and whether it could be edited or deleted to preserve client confidentiality.

Another participant (P04) emphasized that trust was fragile and tied directly to control over sensitive clinical data. Describing the workflow of a GenAI note-taking tool, P04 explained that their trust hinged entirely upon the continued availability of features, such as data deletion and transcript management:

> *With Blueprint, I can delete the transcript, I can wipe it off of that system and that is within my control... and so that makes me trust it and I would be distrustful if they took that away... like that's the premium feature is to delete your transcript.* [P04]

P04's explanation suggests how removing or monetizing such control-related features would undermine his trust. His confidence in the tool thus rests on specific capabilities allowing direct oversight and editability. When those capabilities are guaranteed, he feels safe using the GenAI; if they are taken away, his trust would evaporate.

Subtheme 2: Task-Specific Boundaries of GenAI Trust

Beyond trusting their own abilities and control, psychotherapists' trust in GenAI is also highly task-dependent. Participants generally viewed GenAI as suitable for low-stakes tasks, such as brainstorming intervention ideas or drafting preliminary homework assignments, because mistakes in these contexts carried minimal risk and could easily be corrected by the psychotherapist. Conversely, their trust diminished when GenAI attempted high-stakes or core clinical tasks, such as generating official clinical documentation or nuanced clinical reasoning. Psychotherapists are skeptical that GenAI can handle sensitive clinical decisions or documentation independently without error. P06 articulated this conditional trust, expressing comfort using GenAI for brainstorming but strong reservations about automated note-taking:

> *I'm just not sure that I want a system recording a session and taking a note for me, because I don't fully understand how that information is being gathered and stored, and I wouldn't want my client's information in a system that I didn't fully understand.* [P06]

Here, P06's skepticism arose from uncertainty over how the GenAI processed, recorded, and stored sensitive session data. Allowing a GenAI to record therapy sessions and generate notes raises serious privacy and data security concerns. Even if the GenAI tools could technically produce a decent therapy note, the lack of transparency about data handling was enough to erode her trust. Similarly, P03 refused to use GenAI for core clinical reasoning about a client:

> *I don't ask ChatGPT to conceptualize my clients when it comes to trying to understand what's wrong or what's going wrong; I rely on people for that.* [P03]

In psychotherapy, "conceptualization" refers to the process by which psychotherapists develop a comprehensive understanding of a client's presenting issues, symptoms, underlying causes, and contextual factors [39]. It involves synthesizing information from assessments, clinical observations, and theoretical frameworks to inform diagnosis, treatment planning, and interventions. P03's quote highlights that psychotherapists reserve essential tasks, such as case conceptualization, for their clinical judgment or consultation with human colleagues. P03 trusted GenAI as a tool to improve efficiency in peripheral tasks, but that trust stopped when the technology would encroach on areas requiring deep expertise, intuition, or accountability. In her view, having GenAI substitute for a psychotherapist's interpretive or diagnostic role was beyond acceptable bounds—it challenged the clinician's professional role and could compromise care.

In short, psychotherapists calibrate their trust in GenAI to the stakes of the task: routine or creative adjunct tasks are within bounds, whereas high-stakes therapeutic decisions are off-limits to GenAI.

Subtheme 3: Conditional Trust in Others' Use of GenAI

Another important boundary on psychotherapists' trust emerges when considering other people's use of GenAI. Our participants trusted themselves to use GenAI in a disciplined, responsible manner, but they withheld that trust from clients and even from less-experienced colleagues. Such disparity is rooted in psychotherapists' confidence in their training, ethical standards, and understanding of GenAI's limitations. That confidence drops when they imagine someone without similar expertise using GenAI in the therapy context. As P06 noted, even her high personal trust "probably drops if clients start to use [Gen]AI more and it impedes or interferes with the therapy." Psychotherapists worried that others might not show the same caution or insight, raising the specter of misuse or overreliance on GenAI. P13 illustrated this concern by contrasting her own careful approach with the unpredictability of a client's approach:

> *I feel like because I've done my due diligence, because I know I have a technology background and know how to keep my own information confidential, I feel confident, using the [Gen]AI, the ways that I do. Personally, however, I don't have that confidence that a client would be careful of what they share with a [Gen]AI.* [P13]

P13's concern suggests that psychotherapists feel they could control the flow of sensitive information when they themselves use GenAI, but their trust falters when that control shifts to clients. A client might unwittingly share private details with a GenAI without fully understanding the underlying risks, potentially breaching confidentiality. In the absence of the psychotherapist's oversight, information security cannot be guaranteed, driving psychotherapists' distrust of clients' unsupervised GenAI use.

XSL•FO
RenderX

Psychotherapists also fear that both clients and junior clinicians could become overreliant on GenAI in ways that harm therapy or professional development. Half of our participants worried that clients might overimmerse themselves in GenAI-generated advice or content, losing contact with reality or bypassing the therapeutic process. Similar concerns apply to junior psychotherapists regarding excessive reliance on GenAI for clinical tasks (eg, note-taking and treatment plans), thereby missing out on learning critical skills and clinical reasoning. P04, an educator with positive personal experiences with GenAI-assisted note-taking, articulated concerns about the trainees:

> *As an educator, I do have concerns...students still in school or early psychotherapists or counselors early in their training, early in their residency, not knowing how to create the notes.* [P04]

P04 emphasized the importance of trainees developing foundational skills without excessive GenAI assistance. If beginners shortcut core tasks such as writing progress notes by relying on GenAI, they risk eroding their own clinical agency and expertise. Overreliance on GenAI at early career stages could produce future clinicians who lack essential skills in documentation, observation, and critical thinking.

Despite these concerns about others' use of GenAI, psychotherapists did identify certain conditions that could increase their trust when clients or junior colleagues use GenAI. They did not categorically oppose others using GenAI; rather, they advocated for supervision and education to ensure safe and appropriate use. For clients, many psychotherapists suggested integrating a form of guided GenAI usage into therapy itself, essentially treating a client's GenAI use as something to be discussed and managed within sessions. For example, P05 likened it to how psychotherapists already addressed clients' social media habits:

> *(It's) helping clients navigate the use of [Gen]AI for positive means…just like we have to help clients navigate the use of social media in a positive way. That could be a part of the conversation—an intake: Do you use Snapchat? How often do you use [Gen]AI? What do you use [Gen]AI for? Do you use it to shop? Do you use it to write papers?* [P05]

P05 provided a practical framework for supervising client GenAI use: openly ask about and discuss it during therapy (even as early as the intake interview), just as one would assess other aspects of a client's digital life. Such a proactive approach could help clients use GenAI in a healthy, privacy-conscious manner. It also gives the psychotherapist insight and oversight, which in turn bolsters their trust in how the client is engaging with GenAI outside of sessions. That said, participants acknowledged that not all clients required the same level of monitoring. P12 pointed out that some client populations—she gave the example of first responders—come in with high digital security awareness and a critical stance toward technology. These clients might actually self-regulate their GenAI use responsibly, sometimes even outpacing the psychotherapist's knowledge of the tool:

> *I trust my first-responder clients. They research things on their own, know how to ask questions safely, and decide for themselves whether to rely on a [Gen]AI suggestion. Honestly, by the time they get to session, they sometimes know more about the tool's data flow than I do.* [P12]

P12's experience highlights how individual client characteristics, such as professional background, technology literacy, and skepticism, can increase a psychotherapist's trust in that client's independent GenAI use. This contrasts with other clients, indicating that psychotherapists' trust in others' GenAI use is not one-size-fits-all; it must be calibrated to the person's abilities and context.

Overall, psychotherapists remain cautious about integrating GenAI into clinical work, even as they acknowledge potential benefits in efficiency and support. Lingering concerns about risks keep their distrust ever-present in the background. In short, we found that psychotherapists' trust in GenAI is highly conditional—closely tied to maintaining personal control, selecting appropriate tasks, and ensuring proper contexts for GenAI use. These conditions matter because trust is inseparable from psychotherapists' professional role: safeguarding clinical judgment, ethical responsibility, and accountability for client outcomes [40] makes the question of what, how, and by whom GenAI is used central to their identity as psychotherapists.

Theme 2: Trust, Distrust, and GenAI in Therapeutic Relationships

Overview

Theme 2 shifts focus to the relational context, examining trust and distrust as they arise when clients use GenAI independently, when GenAI is positioned within the therapeutic relationship, and when systemic pressures shape the perceived role of GenAI in therapy. Psychotherapists in our study further negotiated their trust in GenAI against potential distrust, largely depending on how GenAI was positioned within the therapeutic relationship. On one hand, their observations of clients independently using GenAI for therapeutic tasks bolstered initial trust and confidence in GenAI tools. On the other hand, that trust became fragile and often flipped to skepticism when psychotherapists imagined GenAI encroaching on the human bonds that underlie effective therapy. These tensions shaped how psychotherapists envisioned GenAI's future role as either aligning with their professional values or conflicting with them.

Subtheme 1: Psychotherapists' Attitudes in Client-Led GenAI Use as a Complement to the Therapeutic Alliance

Many psychotherapists expressed cautious trust in GenAI based on seeing their clients incorporate tools such as ChatGPT into therapy-related activities in situations where their clients already engaged in active therapy, and psychotherapists could integrate and guide GenAI use within the therapeutic frame. Clients had used GenAI for tasks such as journaling and homework (P01), conversational practice or advice (P06), and organizing thoughts or self-diagnosing symptoms (P13). Psychotherapists generally viewed these independent uses of GenAI by clients as beneficial and not as a threat to their own role. Instead, they described GenAI as a complementary tool that could enhance openness, confidence, and even outcomes in therapy. For example, P01 noted that "if GPT definitely helps you with journal homework,

XSL•FO
RenderX

why not?", reflecting a practical trust in GenAI's assistance with therapeutic homework. Participants recognized that when clients felt more comfortable or organized in sharing their experiences (with a bit of GenAI help), it could strengthen the therapy process rather than weakening it.

Several participants gave examples of how client-driven GenAI use had positive ripple effects in therapy. P06, for example, described a case where a client sought advice from ChatGPT after an argument with his roommate. ChatGPT suggested the client should apologize to the roommate, providing direct feedback that the client might have resisted if it came from a human psychotherapist.

> *If I challenged a client and said in this scenario, it seems like maybe you need to apologize to your roommate, right? And even if that is true, because the client has done something that is not great, the client may be very mad at me. [However], if [Gen]AI has to offer this idea of challenge, as the client, where does that anger go? It's just a computer, right? It doesn't have emotion. And so I think maybe [Gen]AI could deliver more direct and challenging feedback to a client without risk of a therapeutic rupture than a person.* [P06]

Here, our participants saw GenAI as capable of delivering challenging feedback without risking a rupture in the therapeutic relationship. The client could accept the tough suggestion (to apologize) because it came from an impersonal source, thus preserving the client's openness with the psychotherapist. Importantly, participants described GenAI not as a substitute for therapeutic challenge or confrontation, but as a relational intermediary that could support client reflection and progress while keeping the therapeutic relationship intact. In P06's view, GenAI might occasionally say "what needs to be said" in a way that a human psychotherapist strategically might avoid to protect the client's feelings. Importantly, P06 noted this did not reduce the client's trust in their human psychotherapist; in fact, it helped the client make progress (repairing a relationship) while avoiding potential defensiveness toward the psychotherapist.

Another psychotherapist, P13, recounted how a client's use of ChatGPT became a bridge in the therapy session rather than a barrier. The client had chatted with ChatGPT about their symptoms and feelings before the appointment. P13 illustrated how clients' usage of GenAI could be a topic to start the conversations naturally, boosting client confidence and comfort to share their thoughts during sessions:

> *[ChatGPT] allows me to ask questions, like, "Oh, what did you [two] talk about?" or "How did it arrive at that conclusion?" They'll answer, "I started saying I'm having these symptoms — X, Y, Z..." ChatGPT helped them [clients] organize their thoughts and gave us a starting point. I think it made people feel more comfortable and confident... talking about their experience.* [P13]

In this scenario, the GenAI functioned as a conversation rehearsal or journaling tool, priming the client to share more openly in session. By externalizing some of the client's initial thoughts, ChatGPT lowered the barrier to disclosure, making it easier for the client to articulate experiences face-to-face with P13. The psychotherapist observed that this preparatory step reduced the client's anxiety and increased their willingness to explore their feelings in depth. Crucially, P13 did not feel supplanted by the GenAI; the psychotherapist's role remained central in interpreting, contextualizing, and building on what emerged from the client-GenAI interaction. This finding reflects how most participants viewed such GenAI usage as a supplementary aid that could facilitate communication and self-expression while preserving and potentially enhancing the core psychotherapist-client relationship.

Subtheme 2: Distrust Arises When GenAI Threatens Authentic Human Connection

Overview

While psychotherapists were positive about GenAI as a tool, their trust shifted to distrust when considering scenarios where GenAI might encroach on the therapeutic relationship itself. Participants grew skeptical and uneasy about GenAI in therapy whenever it seemed to mimic human empathy, form bonds with clients, or otherwise interfere with the formation of real human relationships. These speculative scenarios triggered strong resistance. Psychotherapists emphasized that effective psychotherapy depended on genuine human connection—often referred to as therapeutic alliance or bond—which involved authentic trust, empathy, and reciprocal emotional engagement. They identified GenAI's limitations in emotional authenticity, responsiveness, and embodied presence as serious barriers to maintaining this bond. In other words, even though GenAI can imitate empathetic language, it cannot provide the true interpersonal experience of being with another caring human. As discussed below, psychotherapists worried that clients forming illusory relationships with GenAI could be misleading or even harmful, and that a GenAI's lack of genuine emotion could undercut the healing power of empathy and trust in therapy.

Inauthentic ("Parasocial") Relationships With GenAI

Several psychotherapists drew analogies between a client chatting with a GenAI and the parasocial relationships people develop with media figures or celebrities. Parasocial relationships are one-sided bonds where a person feels a strong connection or intimacy with someone (often a public figure) who is not actually reciprocating that relationship [41]. P04 compared client-GenAI interactions to parasocial relationships, such as those people form with celebrities, highlighting their superficial and ultimately misleading nature:

> *A common example is you feel like you know your favorite celebrity, but you really don't. But you have an attachment to them. If they pass away, you are extremely sad. You feel that loss, you grieve it. You know, those are parasocial relationships. So developing something like that with your GenAI bot as, in a therapeutic realm, I think, is not appropriate.* [P04]

P04 highlighted the illusory nature of the bond: the client might feel cared for by the GenAI, but in reality, there is no mutual relationship—the GenAI does not "know" the client or genuinely

XSL•FO
RenderX

care, just as a celebrity does not know their grieving fan. P04 feared that a client might invest emotionally in a GenAI, only to be hurt upon realizing that this relationship was a 1-way projection. Such a realization could be emotionally devastating (analogous to grieving a celebrity's death) and might even undermine the client's trust in real relationships. The implication is that trusting a GenAI as if it were a person is fundamentally misguided, and psychotherapists felt a duty to discourage such attachments for the client's well-being. Similarly, participant P14 echoed concerns about the implications of normalizing uncertain relationships between clients and GenAI in therapeutic contexts, noting that:

> *I do think part of counseling is the relationship that is formed. And again, what does it mean to have a relationship with a non human and like? Is that something that we want or not? And is that something that should be encouraged or not?* [P14]

Both P04 and P14 were essentially asking: if clients start treating GenAI bots as psychotherapists or friends, do we risk eroding the value of real human-human therapeutic bonds? They emphasized their distrust toward the idea of GenAI-client "relationships," viewing them as shallow imitations that could mislead clients and ultimately prove harmful. Their perspective taps into broader debates on whether emotional bonds with anthropomorphized GenAI are healthy or not; indeed, psychologists have noted that while such bonds might feel real, they lack true reciprocity and can risk shaping unrealistic expectations about human relationships.

Lack of Genuine Emotional Bond and Empathy

Beyond the concept of parasocial attachment, psychotherapists also expressed concern that interacting with an emotionally simulated agent could stunt clients' ability or desire to form meaningful human connections. Participants noted that much of therapy's healing power stemmed from experiencing genuine empathy and coregulation with another person. GenAI, no matter how cleverly it reflected a client's words, fundamentally lacked human emotion, life experience, and a nervous system—it cannot truly "feel" or share in the client's pain and joy.

As P11 reflected on this with her experiences in talking with GenAI and feeling real comfort, she questioned, "do we want that to be where we go as a profession, they call it the healing arts." P11 was grappling with the ambivalence: a GenAI might provide momentary comfort (eg, through kind words at 2 AM when no one else is around), but if therapy were to shift toward GenAI, something essential would be lost. She noted that "therapy is not a science, is not formulaic, but is about people." Thus, shifting therapy to GenAI would strip away its essence as an art grounded in human connection, not a formula that can be executed. P11 even pondered whether making GenAI more human-like for our own convenience "somehow cheapens our experience of being human, or cheats us out of the chance to have real human connection." P11's sentiment indicated a sense of distrust: even if GenAI "works" on the surface to soothe someone, relying on it for emotional support might erode the value of human-to-human experiences.

Such interpersonal neural synchronization is only possible through face-to-face human interaction, involving elements such as eye contact, voice tone, and somatic empathy that a GenAI agent cannot easily reproduce. P9 touched on this when she explained that in the therapy room:

> *There are mirror neurons in both of our brains... When I take a deep breath, [the client] subconsciously takes a deep breath. When I model a healthy expression of emotion, their mirror neurons are doing the same… There's a neurological component of therapy that would be lost with GenAI.* [P09]

In highlighting this embodied coregulation, P09 was pointing to the subtle yet powerful ways a psychotherapist's presence helped a client feel safe and understood on a biological level. Participants widely felt that a GenAI—even one programmed to respond empathetically in text—could not replicate these rich, multichannel human signals of empathy. Thus, psychotherapists were unwilling to trust GenAI with the core work of forging an emotional bond.

P10 captured the risk that GenAI's unconditional supportiveness could create a comfortable bubble that ultimately harms client growth. She noted that "learning to tolerate not feeling good can often be part of psychotherapy." In real therapy, a good clinician sometimes needs to challenge clients' perceptions or allow them to sit with discomfort as part of healing. P10 wondered if a "GenAI-generated therapy" would instead become "more of a 'thumbs up' scenario, creating a loop of people hearing what they want to hear, and thus feeling good." If a GenAI companion always praises or agrees with the user (either by design or to avoid upsetting them), this always-supportive feature will give clients an illusion of feeling good temporarily and avoid difficult, necessary conversations with real people. Over time, this feedback loop might reduce tolerance for the messiness of real human relationships, and clients could stall in their therapeutic progress. As P10 put it, clients might "cut off real human connections" and bypass the harder work of connecting with others and facing real-world challenges if they grow too comfortable with a GenAI that never challenges them. Such an outcome would be the opposite of what therapy strives for, which is to help clients function better in their lives and relationships outside therapy.

Subtheme 3: From Personal Fear to Systemic Skepticism—Commercial Incentives Stoke Distrust

Another major theme in psychotherapists' discussions of trust was the tension between seeing GenAI as a helpful partner in therapy versus viewing it as a potential competitor or replacement. Such tension affected their willingness to trust technology. When GenAI was framed as augmenting their work—handling ancillary tasks, extending care between sessions, or providing additional resources—psychotherapists were generally open-minded and even optimistic. In those cases, GenAI aligned with their professional values, and they could trust it to support better outcomes. However, when the conversation turned to GenAI possibly replacing human psychotherapists, participants expressed strong distrust, often accompanied by anxiety and even a sense of threat to their professional identity.

XSL•FO
RenderX

We found that psychotherapists' fears evolved from that initial shock toward more pragmatic skepticism. They started scrutinizing real-world trends that might lead to GenAI-driven replacement of psychotherapists. One major factor was the way some technology companies are promoting mental health GenAI products. Participants noted that a number of startups have begun rolling out "therapy bots" or GenAI coach apps. Officially, these are often marketed as assistive tools, but psychotherapists suspected the subtext was to test the waters for full replacement. P13, for instance, mentioned seeing posts on LinkedIn about new chatbot services:

> *They're [mental health chatbot startups] trying to say, "Oh, this is going to be an assistant," but you can tell by the way they're working it that they're almost trying to say, hey, we can just do this therapy bot... put in all the information... and it's gonna replace therapy instead of having to talk to a therapist.* [P13]

The design and messaging of these products made P13 feel that the endgame for some developers is to bypass human psychotherapists altogether. This sense was reinforced by other participants (P11, P17) who observed how GenAI mental health platforms often come with bold claims or marketing hype. P17 warned that a slickly marketed GenAI could "scam potential clients into thinking they can get [the same] help" from a bot, when in reality that was not advised. Similarly, P11 worried that for-profit GenAI platforms might make "false promises" that lure people into viewing GenAI as nearly equivalent to a human psychotherapist. These comments suggest that psychotherapists' distrust of GenAI was not only about GenAI's capabilities, but also about the intentions and honesty of those deploying it. If GenAI is oversold as a reliable replacement for therapy, psychotherapists fear the public could be misled and that clients might forgo seeking real help, with potential harm.

Beyond the companies pushing GenAI, participants highlighted structural pressures that could drive GenAI-as-psychotherapist scenarios. A concern was the role of insurance and health care payers. In many health care systems, including mental health care, decisions about what services are covered (and at what rate) profoundly influence practice. Psychotherapists suspected that if GenAI therapy is seen as significantly cheaper and more efficient, insurers might encourage or even force a shift in that direction. P10 cynically remarked that "[Gen]AI methodology is going to be promoted by insurances [as] a cheaper, better, quicker fix." P13 similarly noted the temptation for efficiency:

> *From an insurance company side—you can be like, "Oh, well, a chatbot can talk to a thousand people a day." That sounds great for efficiency.* [P13]

Psychotherapists worried that they and their clients could be placed in a no-win situation, where financial incentives pressure everyone toward GenAI solutions regardless of actual therapeutic value. The mere possibility of this happening fueled a lot of distrust and even resentment toward GenAI in our interviews. It is not that these psychotherapists believe a GenAI could truly do the job as well as they can. In fact, they asserted it cannot, for all the human reasons discussed earlier, but the threat is that decisions outside their control (by companies or insurers) might force the issue. Whenever GenAI was envisioned as a substitute for a human psychotherapist, participants' trust evaporated. They viewed that narrative with distrust and, at times, with a sense of having to justify their own worth in the face of automation.

In summary, psychotherapists in our study navigated a delicate balance of trust and distrust toward GenAI in therapeutic settings. When GenAI supported them and their clients—acting as a friendly assistant that respects the primacy of human connection—they largely trusted it and saw promise. However, when GenAI appeared to usurp human roles or relationships, they responded with firm distrust, rooted in emotional intuition and professional reasoning. These findings highlight the importance of context and role definition in shaping trust: as the literature on automation suggests [21,42], people trust technology only when its use aligns with their goals and values and withdraw trust if it threatens or violates their goals. For psychotherapy, the inviolable goal is to promote genuine human healing and connection. GenAI's role in this domain, according to our participants, will depend on whether it is designed to honor and enhance, rather than redefine or replace, the human-centric nature of therapeutic relationships.

Across these findings, we see that psychotherapists' trust and distrust in GenAI are fundamentally about professional role identity. The introduction of GenAI into therapeutic contexts surfaces latent questions about what it means to be a psychotherapist, such as what tasks are core to the role, what knowledge and judgment are uniquely human, and where the boundaries of professional responsibility lie. When participants resisted GenAI involvement in case conceptualization, expressed discomfort with clients forming "relationships" with chatbots, or worried about insurance-driven replacement, they were not merely assessing GenAI's technical capabilities; they were articulating and defending the boundaries of their professional role. Such dynamics can be understood through role theory [42]. GenAI's entry into therapy creates potential role ambiguity (uncertainty about which tasks belong to the clinician vs the technology), role conflict (tension between efficiency pressures and relational values), and opportunities for role clarification (explicit articulation of what remains uniquely human). Participants' insistence on control, their task-specific boundaries, and their concerns about parasocial client-AI relationships all reflect efforts to resolve these tensions in favor of preserving the clinician's interpretive and relational primacy.

In this sense, GenAI serves as a clarifying lens: by confronting what GenAI can and cannot do, psychotherapists articulate more explicitly what they believe only a human clinician can provide—authentic empathy, embodied presence, ethical accountability, and the capacity to work through relational ruptures. These articulations contribute to ongoing professional discourse about the nature of psychotherapeutic expertise in an increasingly technological landscape.

XSL•FO
RenderX

Discussion

Principal Results

Our findings reveal that psychotherapists primarily used GenAI for low-risk, supportive tasks (eg, notes, brainstorming, and homework), while drawing firm boundaries around high-stakes clinical reasoning (eg, case formulation and diagnosis) and any opaque handling of client information. As a result, psychotherapists' trust in GenAI is highly context-dependent and rooted in maintaining human control and relationships. Finally, we concluded our findings as: psychotherapists' control as the locus of trust based on their current GenAI adoption, the management of the psychotherapist-client-GenAI triad to protect the therapeutic alliance, and the systemic pressures and fears of exclusion that shape implications for future policy and design.

First, our findings show that trust hinges on professional autonomy (eg, the "hand-on-the-wheel" rule) and concrete controls (eg, prompt scoping, editability, and data deletion) that allow psychotherapists to supervise what goes in and what comes out. Even when the same GenAI tool was technically capable of other functions, psychotherapists were reluctant to shift trust to unfamiliar uses outside their control. For psychotherapists, even a slight shift in control, or a perceived loss of it, can substantially reduce the psychotherapists' confidence. They integrate GenAI into their workflow only on their own terms, within the boundaries of their professional knowledge and practice.

Psychotherapists' insistence on concrete controls (eg, scoping GenAI prompts, editing GenAI-generated content, and ensuring data privacy) echoes findings in other professions that emphasize oversight in using GenAI [43-45]. For example, in domains such as finance and accounting, professionals also exhibit conditional trust in GenAI and demand human oversight to meet privacy, compliance, and accuracy requirements [46]. Indeed, many accounting firms (including large firms in the Big Four) have introduced GenAI to assist with tasks such as auditing and tax preparation while keeping a human auditor in the loop [47]. However, psychotherapists practice an even stricter form of conditional trust. Given the extraordinary sensitivity of therapeutic work, where mistakes can carry deeply personal or even harmful consequences for clients, trust in automation is "squeezed"—carefully rationed and contextually negotiated. Consistent with prior research on trust in high-risk settings [48,49], our participants' trust was conservative by necessity: they only trusted GenAI when its use aligned with their goals, values, and ethical duties, and they withdrew trust immediately if the GenAI's role threatened those principles.

However, maintaining "control" through careful prompting and boundary-setting does not necessarily guarantee actual control over how GenAI shapes interpretation. As algorithmic systems operate within preexisting biases and conversational frames, even skilled users may be subject to subtle influences that confirm rather than challenge their interpretive directions. The distinction between actual and perceived control is clinically important as a psychotherapist who believes they are steering the GenAI may, in fact, be receiving outputs shaped by the model's training patterns in ways that reinforce existing assumptions. Such a gap between perceived and actual control interacts with professional role identity. Psychotherapists' expertise traditionally includes recognizing when their own biases may be influencing clinical judgment—a form of reflective practice. GenAI complicates this by introducing an interlocutor whose biases are opaque and whose "agreement" may feel like validation. Future research should examine whether and how clinicians can develop literacy about these dynamics to engage with GenAI more reflexively.

Taken together, our findings reveal that psychotherapists' conditions for trust extend beyond retaining veto power or technical oversight. Trust is conditional on maintaining professional role identity—not merely having a "hand on the wheel," but preserving the clinician's interpretive authority, ethical responsibility, and relational primacy. When GenAI threatens to shift these role boundaries, even if the clinician technically retains control, trust erodes. In relationally intensive professions such as psychotherapy, "human-in-the-loop" is necessary but not sufficient; the human must remain in the loop as a clinician with a defined professional role, not merely as a monitor or editor of outputs. Such a role-identity framing could help explain why participants resisted GenAI for case conceptualization (a task central to clinical identity) while accepting it for documentation (a task peripheral to their core professional self-concept).

Second, our findings add nuance to understanding the relationship among psychotherapists, clients, and GenAI by examining counselors' observations of (and hopes for) client-led use of GenAI in therapy. Participants generally supported GenAI as a supplementary aid for clients, but only when it complemented, not competed with, the human relationship. Trust was maintained when a client's use of GenAI enriched the therapy process under the psychotherapist's guidance. For example, some clients had experimented with GenAI tools for personal journaling, mood tracking, or skills practice between sessions, and psychotherapists viewed these as positive and adjunctive activities as long as they could review and integrate the GenAI-generated content into treatment. To participants, the GenAI tool functioned as a bridge—facilitating communication and self-reflection—while the core psychotherapist-client bond was preserved or even strengthened. These scenarios illustrate how GenAI can serve as a therapeutic supplement that empowers clients (by providing on-demand exercises or feedback) without undermining the clinician's authority or the authenticity of the human connection.

While psychotherapists expressed openness to client-led uses of GenAI, as reflected in remarks such as "why not," participants stressed that acceptance and these benefits only hold under specific conditions, when clients are actively engaged in therapy, and the clinician can oversee the GenAI's role. Unsupervised or excessive reliance on GenAI by clients raised ethical red flags for many practitioners. In psychotherapists' view, a GenAI that provided easy answers or constant validation might become an emotional crutch, potentially impeding clients' growth by sidestepping the discomfort and vulnerability through which therapeutic change occurs. Several psychotherapists noted the risk of clients developing an illusory relationship with a chatbot that mimics empathy but lacks genuine human understanding.

XSL•FO
RenderX

They voiced concern that unvetted advice or faux-empathy from GenAI could misguide clients or undercut the healing power of authentic human empathy in therapy. Thus, even the same tools that help organize thoughts between sessions were seen as double-edged: without guardrails, those tools might shape clients' thinking in ways the psychotherapist cannot monitor or correct. This highlights the importance of embedding GenAI use within ongoing professional support, rather than allowing it to operate as a parallel "psychotherapist."

Third, systemic forces, such as insurance reimbursement models, health care policies, and technology marketing, further complicate the psychotherapist-client-GenAI triad. Our findings reveal that psychotherapists' fear of being "replaced" by GenAI stems less from its raw capabilities than from institutional pressures. Many participants expressed concern that decisions by insurers, employers, or health care systems might push GenAI solutions into therapy for the sake of cost or efficiency, sidelining human clinicians against their will. Indeed, signs of these systemic shifts are already visible. In the United States, Medicare announced that as of January 2025, it will reimburse providers for prescribing certain US Food and Drug Administration–cleared digital mental health therapeutics [50]. Private insurers such as Aetna and Cigna have also begun partnering with digital mental health platforms and covering app-based therapeutic services. These developments reflect a growing institutional embrace of GenAI-driven and digital tools to expand access and reduce costs. However, psychotherapists in our study saw a potential dark side: exclusion from the care loop if stakeholders treat GenAI as a cheap substitute for human therapy. They worried about a "no-win situation" where financial incentives pressure clinics to replace portions of therapy with automated services, regardless of client benefit. Such a possibility fueled a great deal of distrust and even resentment toward GenAI among participants. Crucially, it was not that psychotherapists believed a GenAI could do their job just as well—on the contrary, they were confident in the unique value of human care—but rather that decisions outside their control might force the issue (eg, an insurance company mandating a chatbot as a first line of support). Whenever GenAI was imagined as a substitute for a human psychotherapist, participants' tentative trust evaporated, replaced by staunch distrust and a need to justify their professional worth.

Psychotherapists also noted mixed signals in the policy landscape that mirror this tension. On one hand, policymakers are cautiously enabling GenAI integration in mental health care through reimbursement and innovation incentives. On the other hand, there is growing recognition of the ethical and quality concerns: for example, some US jurisdictions are moving to outright ban certain uses of AI in psychotherapy. In August 2025, the state of Illinois enacted a law prohibiting the use of AI to provide psychotherapy or make clinical decisions [51], while still allowing AI for "administrative and supplementary support" under a provider's oversight. This law was explicitly framed as protecting patients from unregulated AI or GenAI tools and preserving the jobs of human psychotherapists. Such contradictory responses—from enthusiastic adoption to legal restriction—highlight a core societal debate: Should GenAI be embraced to improve mental health accessibility and efficiency, or curtailed to preserve the human core of therapy? Our participants overwhelmingly rejected either extreme position. The psychotherapists' stance was one of measured, principled integration: GenAI should be used only in ways that support—and never replace or diminish—the therapeutic relationship. Psychotherapists in our study are not "Luddites" resisting technology out of fear; instead, they are concerned practitioners seeking to ensure GenAI is implemented responsibly and relationally. Their distrust is not a reflexive fear of job loss per se, but a deeper worry about clients being left without meaningful human connection if economic structures push GenAI-only solutions as a shortcut. In fact, many envisioned constructive roles for GenAI (some called it a potential "glue" in therapy) that could enhance rapport, support client progress, and deepen engagement, but only if these tools are introduced thoughtfully, with ethical guardrails and human oversight at every step.

Limitations and Future Implications

A limitation of this study is that our interview participants may represent a group of psychotherapists with greater initial interest in GenAI than the average practitioner, as they voluntarily opted in through our recruitment emails to complete the interviews. Their perspectives may therefore be more exploratory or engaged compared with those of clinicians who chose not to participate. Our participants were also US-based and demographically skewed (predominantly White, with a relatively larger proportion of women). We did not systematically examine whether perspectives varied across demographic characteristics (eg, gender, race or ethnicity, and years of practice). As a result, the trust and distrust dynamics reported here may not fully reflect the diversity of clinician perspectives and may transfer differently across countries, health care systems, and regulatory contexts. Additionally, we did not assess participants' primary therapeutic orientation (eg, cognitive-behavioral, psychodynamic, humanistic, and integrative). In the US context, some licensed psychotherapists practice eclectically or identify with integrative approaches, but orientation may nonetheless shape attitudes toward structured tools and the conceptualization of the therapeutic relationship. For example, practitioners with a strong humanistic or relational orientation may be more skeptical of GenAI's capacity to support authentic connection, while those with cognitive-behavioral orientations may be more receptive to GenAI-assisted psychoeducation or homework. Future research should explicitly examine whether psychotherapists from different schools (eg, the psychodynamic, humanistic, and cognitive-behavioral) differ in their trust calibration and in the conditions under which they would accept GenAI involvement in therapy. Finally, the rapid pace of GenAI development means that our findings reflect psychotherapists' perceptions at a particular moment in time. As GenAI tools continue to evolve in capability, regulation, and clinical integration, psychotherapists' trust, concerns, and usage practices may shift in ways not fully captured here.

Despite these limitations, the findings remain highly relevant for clinicians and designers seeking to understand how psychotherapists use GenAI in real clinical contexts, how they negotiate trust and distrust under different conditions, and how they navigate clients' independent use of GenAI. The themes

XSL•FO
RenderX

identified in this study can guide design considerations, inform professional training needs, and highlight the types of support clinicians require when confronted with GenAI in real-world practice.

Future GenAI tools for psychotherapy should be designed to enhance clinician oversight and preserve clinical judgment, rather than shifting interpretive authority to the system. Our findings suggest "minimum conditions" for trust and adoption, including clinician-facing controls to review outputs before use, edit or annotate content to reflect clinical nuance, and delete materials that are inaccurate or clinically unhelpful. Trust also depends on transparency that helps clinicians understand what the model is responding to (eg, visible inputs and omissions) and recognize overconfident or overly generic guidance. Thus, clear data governance, such as storage, retention, access, and use for model improvement, is also essential for informed privacy decisions.

In addition, future researchers could also explore how GenAI can strengthen communication between psychotherapists and clients. Participants described how "GenAI in the loop" could support communication by offering direct or challenging prompts, helping clients organize thoughts, or assisting them in articulating emotions more clearly. Developers may consider GenAI features that act as a bridge between clients and psychotherapists—for example, translating clinical terminology into laypeople's language or generating structured summaries that help psychotherapists quickly understand clients' concerns. Such features could enhance clarity and mutual understanding while preserving the psychotherapist's primary role in treatment.

Participants also highlighted opportunities for GenAI-supported psychoeducation and training. As GenAI becomes increasingly common in clients' daily lives and in clinical workflows, trainees may benefit from working with GenAI early to learn ethical and appropriate use. Interactive GenAI-simulated patients could also provide customizable training scenarios with different patient backgrounds, symptom profiles, or emotional intensities that are not always available in traditional role-plays or standardized-patient exercises. These adaptive environments may help trainees build practical skills and prepare for future clients who already integrate GenAI into their mental health routines.

Furthermore, the tension between systemic pressures and therapeutic values points to future implications for policy and governance. There is a clear need for professional guidelines and regulations that reaffirm the centrality of human psychotherapists in any GenAI-augmented care model. Organizations such as the APA have begun developing guidance to help psychotherapists navigate GenAI in practice, and our findings underscore how crucial such efforts are. Policies must strike a balance so that GenAI can be leveraged to extend care (eg, through approved digital therapeutics) without eroding standards of care or marginalizing clinicians. Psychotherapists may also require training to work effectively alongside GenAI tools and to advocate for their patients' needs in dialogues with employers or insurers deploying these tools. Equipping psychotherapists with digital literacy and ethical decision-making frameworks for GenAI can help them retain a sense of agency and protect client welfare when new technology is introduced.

Conclusions

In this work, we conducted an interview study with 18 psychotherapists to explore how psychotherapists perceive, trust (or distrust), and incorporate GenAI into their everyday practice. We found that psychotherapists exhibit a cautious, conditional trust in GenAI: they embrace these tools for administrative and low-risk tasks (eg, note-taking, training simulations, and between-session client exercises) but are wary of using GenAI for core clinical decisions or any role that might infringe upon the therapeutic relationship. By recognizing the conditions under which psychotherapists are willing to trust GenAI, we can inform the development of tools and policies that empower psychotherapists and safeguard clients. Moving forward, it will be crucial to pursue interdisciplinary research and dialogue that continue to balance innovation with empathy, ensuring that future GenAI in mental health strengthens the art of therapy instead of undermining it.

Acknowledgments

During the preparation of this work, the authors used ChatGPT-5.0 (OpenAI) solely to check the grammatical accuracy of the written text. All suggested edits were reviewed by the authors, who take full responsibility for this final paper.

Funding

This work received support from the Charles Center at William & Mary. The funder had no role in study design, data collection, analysis, interpretation, or paper preparation.

Authors' Contributions

Conceptualization: JK, ALP, YZ
Data curation: JK
Formal analysis: JK
Investigation: JK, AP
Methodology: JK, YZ
Project administration: JK
Resources: AP

XSL•FO
RenderX

Supervision: YZ
Writing – original draft: JK
Writing – review & editing: YZ

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide used to explore psychotherapists' experiences, trust and distrust, and perceived roles of GenAI in practice. GenAI: generative artificial intelligence.
[DOCX File, 1940 KB - jmir_v28i1e88932_app1.docx]

Checklist 1
COREQ 32-item checklist.
[PDF File, 134 KB - jmir_v28i1e88932_app2.pdf]

References

1. Couture V, Roy MC, Dez E, Laperle S, Bélisle-Pipon JC. Ethical implications of artificial intelligence in population health and the public's role in its governance: perspectives from a citizen and expert panel. J Med Internet Res 2023 Apr 27;25:e44357. [doi: 10.2196/44357] [Medline: 37104026]
2. Grassini S, Koivisto M. Understanding how personality traits, experiences, and attitudes shape negative bias toward AI-generated artworks. Sci Rep 2024 Feb 19;14(1):4113. [doi: 10.1038/s41598-024-54294-4] [Medline: 38374175]
3. Lee J, Jung K, Newman EG, Chow E, Chen Y. Understanding adolescents' perceptions of benefits and risks in health AI technologies through design fiction. CHI 2025 2025 Apr 26:1-20. [doi: 10.1145/3706598.3713244]
4. Liu Z, Zou W, Lin C. Exploring the influence of privacy concerns, AI literacy, and perceived health stigma on AI chatbot use in healthcare: an uncertainty reduction approach. Patient Educ Couns 2025 Nov;140:109271. [doi: 10.1016/j.pec.2025.109271] [Medline: 40712403]
5. Luo X, Li Z, Yang Z, et al. Using large language models to assess the consistency of randomized controlled trials on AI interventions with CONSORT-AI: cross-sectional survey. J Med Internet Res 2025 Sep 26;27:e72412. [doi: 10.2196/72412] [Medline: 41004321]
6. Abrams Z. Artificial intelligence is reshaping how psychologists work. American Psychological Association Services, Inc. 2025 May 12. URL: https://www.apaservices.org/practice/news/artificial-intelligence-psychologists-work [accessed 2026-03-18]
7. Page C, Assefa M, Stamm K. What psychologists are saying about using AI in practice. Monitor Psychol 2025 Jul 1;56(5) [FREE Full text]
8. Yıldız HA, Söğütdelen E. AI chatbots as sources of STD information: a study on reliability and readability. J Med Syst 2025 Apr 3;49(1):43. [doi: 10.1007/s10916-025-02178-z] [Medline: 40178771]
9. Báez JC, Victor BG, Dysart C, Goldkind L. "I don't understand it, but okay": an empirical study of mental health practitioners' readiness to use large language models. J Technol Hum Serv 2025 Apr 3;43(2):142-160. [doi: 10.1080/15228835.2025.2479475]
10. Zhang M, Scandiffio J, Younus S, et al. The adoption of AI in mental health care-perspectives from mental health professionals: qualitative descriptive study. JMIR Form Res 2023 Dec 7;7:e47847. [doi: 10.2196/47847] [Medline: 38060307]
11. Gültekin M, Şahin M. The use of artificial intelligence in mental health services in Turkey: what do mental health professionals think? Cyberpsychology (Brno) 2024;18(1):6. [doi: 10.5817/CP2024-1-6]
12. Sweeney C, Potts C, Ennis E, et al. Can chatbots help support a person's mental health? Perceptions and views from mental healthcare professionals and experts. ACM Trans Comput Healthcare 2021 Jul 31;2(3):1-15. [doi: 10.1145/3453175]
13. Tang Y, Kang Y, Wang Y, Wang T, Zhong C, Gong J. Ca+: cognition augmented counselor agent framework for long-term dynamic client engagement. : arXiv Preprint posted online on Mar 27, 2025. [doi: 10.48550/arXiv.2503.21365]
14. Glikson E, Woolley AW. Human trust in artificial intelligence: review of empirical research. Ann R Coll Physicians Surg Can 2020 Jul;14(2):627-660. [doi: 10.5465/annals.2018.0057]
15. Dwork C, Minow M. Distrust of artificial intelligence: sources & responses from computer science & law. Daedalus 2022 May 1;151(2):309-321. [doi: 10.1162/daed_a_01918]
16. Chavaillaz A, Wastell D, Sauer J. System reliability, performance and trust in adaptable automation. Appl Ergon 2016 Jan;52:333-342. [doi: 10.1016/j.apergo.2015.07.012]
17. Chen C, Sundar SS. Is this AI trained on credible data? the effects of labeling quality and performance bias on user trust. CHI '23 2023 Apr 19:1-11. [doi: 10.1145/3544548.3580805]
18. Yin M, Vaughan JW, Wallach H. Understanding the effect of accuracy on trust in machine learning models. CHI '19 2019 May 2:1-12. [doi: 10.1145/3290605.3300509]

XSL•FO
RenderX

19. Naiseh M, Al-Thani D, Jiang N, Ali R. How the different explanation classes impact trust calibration: the case of clinical decision support systems. Int J Hum Comput Stud 2023 Jan;169:102941. [doi: 10.1016/j.ijhcs.2022.102941]
20. Yu K, Berkovsky S, Conway D, Taib R, Zhou J, Chen F. Do i trust a machine? differences in user trust based on system performance. In: Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent: Springer; 2018:245-264. [doi: 10.1007/978-3-319-90403-0_12]
21. Evans AM, van de Calseyde P. The reputational consequences of generalized trust. Pers Soc Psychol Bull 2018 Apr;44(4):492-507. [doi: 10.1177/0146167217742886] [Medline: 29251247]
22. Wang Y, Wang Y, Crace K, Zhang Y. Understanding attitudes and trust of generative AI chatbots for social anxiety support. Presented at: CHI 2025; Apr 26 to May 1, 2025. [doi: 10.1145/3706598.3714286]
23. Stubbe DE. The therapeutic alliance: the fundamental element of psychotherapy. FOC 2018 Oct;16(4):402-403. [doi: 10.1176/appi.focus.20180022]
24. Stefana A, Fusar-Poli P, Vieta E, Youngstrom EA. Patients' perspective on the therapeutic relationship and session quality: the central role of alliance. Front Psychol 2024;15:1367516. [doi: 10.3389/fpsyg.2024.1367516] [Medline: 39188865]
25. Locher C, Meier S, Gaab J. Psychotherapy: a world of meanings. Front Psychol 2019;10:460. [doi: 10.3389/fpsyg.2019.00460] [Medline: 30984050]
26. DeYoung PA. Relational Psychotherapy: A Primer, 2nd edition: Routledge; 2015. [doi: 10.4324/9781315723709]
27. Feher B, Tussie C, Giannobile WV. Applied artificial intelligence in dentistry: emerging data modalities and modeling approaches. Front Artif Intell 2024;7:1427517. [doi: 10.3389/frai.2024.1427517] [Medline: 39109324]
28. Koyun M, Taskent I. Evaluation of advanced artificial intelligence algorithms' diagnostic efficacy in acute ischemic stroke: a comparative analysis of ChatGPT-4o and Claude 3.5 Sonnet models. JCM 2025 Jan 17;14(2):571. [doi: 10.3390/jcm14020571] [Medline: 39860577]
29. Bürger VK, Amann J, Bui CKT, Fehr J, Madai VI. The unmet promise of trustworthy AI in healthcare: why we fail at clinical translation. Front Digit Health 2024;6:1279629. [doi: 10.3389/fdgth.2024.1279629] [Medline: 38698888]
30. McCoyd JLM, Curran L, Candelario E, Findley P. "There is just a different energy": changes in the therapeutic relationship with the telehealth transition. Clin Soc Work J 2022;50(3):325-336. [doi: 10.1007/s10615-022-00844-0] [Medline: 35493775]
31. Norcross JC. The therapeutic relationship. In: Duncan BL, Miller SD, Wampold BE, Hubble MA, editors. The Heart and Soul of Change: Delivering What Works in Therapy, 2nd edition: American Psychological Association; 2010:113-141. [doi: 10.1037/12075-004]
32. Waldock WJ, Lam G, Baptista A, Walls R, Sam AH. Which curriculum components do medical students find most helpful for evaluating AI outputs? BMC Med Educ 2025 Feb 6;25(1):195. [doi: 10.1186/s12909-025-06735-5] [Medline: 39915801]
33. Doyle B, Fitzsimons D, McKeown P, McAloon T. Understanding dietary decision-making in patients attending a secondary prevention clinic following myocardial infarction. J Clin Nurs 2012 Jan;21(1-2):32-41. [doi: 10.1111/j.1365-2702.2010.03636.x] [Medline: 21545664]
34. Liao QV, Varshney KR. Human-centered explainable AI (XAI): from algorithms to user experiences. arXiv. Preprint posted online on Apr 19, 2022. [doi: 10.48550/arXiv.2110.10790]
35. Artificial intelligence in mental health care. American Psychological Association. 2025 Mar 12. URL: https://www.apa.org/practice/artificial-intelligence-mental-health-care [accessed 2026-03-31]
36. Thomas DR. A general inductive approach for analyzing qualitative evaluation data. Am J Eval 2006 Jun;27(2):237-246. [doi: 10.1177/1098214005283748]
37. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
38. Stamm K, Lin L, Christidis P. Career stages of health service psychologists. American Psychological Association. 2017. URL: https://www.apa.org/workforce/publications/15-health-service-career [accessed 2026-03-18]
39. Rennie DL, Phillips JR, Quartaro GK. Grounded theory: a promising approach to conceptualization in psychology? Can Psychol / Psychol can 1988;29(2):139-150. [doi: 10.1037/h0079765]
40. Luckasson R, Tassé MJ, Schalock RL. Professional responsibility in the field of intellectual and developmental disabilities: its definition, application, and impacts. Intellect Dev Disabil 2022 Jun 1;60(3):183-198. [doi: 10.1352/1934-9556-60.3.183] [Medline: 35640603]
41. Schiappa E, Allen M, Gregg PB. Parasocial relationships and television: a meta-analysis of the effects. In: Preiss RW, Gayle BM, Burrell N, Allen M, Bryant J, editors. Mass Media Effects Research: Advances Through Meta-Analysis Lawrence Erlbaum Associates: Lawrence Erlbaum Associates Publishers; 2007:301-314.
42. Lee JD, See KA. Trust in automation: designing for appropriate reliance. Hum Factors 2004;46(1):50-80. [doi: 10.1518/hfes.46.1.50.30392] [Medline: 15151155]
43. Heinonen E, Orlinsky DE. Psychotherapists' personal identities, theoretical orientations, and professional relationships: elective affinity and role adjustment as modes of congruence. Psychother Res 2013;23(6):718-731. [doi: 10.1080/10503307.2013.814926] [Medline: 23829641]
44. Rojas JC, Teran M, Umscheid CA. Clinician trust in artificial intelligence: what is known and how trust can be facilitated. Crit Care Clin 2023 Oct;39(4):769-782. [doi: 10.1016/j.ccc.2023.02.004] [Medline: 37704339]

XSL•FO
RenderX

45. Ala-Luopa S, Olsson T, Väänänen K, Hartikainen M, Makkonen J. Trusting intelligent automation in expert work: accounting practitioners' experiences and perceptions. Comput Supported Coop Work 2024 Dec;33(4):1343-1371. [doi: 10.1007/s10606-024-09499-6]
46. Kennedy R, Tiede L, Austin A, Ismael K. Law enforcement and legal professionals' trust in algorithms. J Law Empirical Anal 2025 Jun;2(1):77-96. [doi: 10.1177/2755323X251325594]
47. Maple C, Szpruch L, Epiphaniou G, et al. The AI revolution: opportunities and challenges for the finance sector. arXiv. Preprint posted online on Aug 31, 2023. [doi: 10.48550/arXiv.2308.16538]
48. Sibayan K. Big Four now using agentic AI to boost staff productivity. New York State Society of CPAs. 2025 Mar 24. URL: https://www.nysscpa.org/most-popular-content/big-four-now-using-agentic-ai-to-boost-staff-productivity-032425##sthash.43m0gWrb.dpbs [accessed 2026-03-18]
49. Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beck HP. The role of trust in automation reliance. Int J Hum Comput Stud 2003 Jun;58(6):697-718. [doi: 10.1016/S1071-5819(03)00038-7]
50. Aguilar M. Medicare finalized coverage for some mental health apps. Now what? STAT. 2024 Nov 4. URL: https://www.statnews.com/2024/11/04/medicare-payment-mental-health-digital-therapeautics/ [accessed 2026-03-18]
51. Gov Pritzker signs legislation prohibiting AI therapy in Illinois. Illinois Department of Financial & Professional Regulation. 2025 Aug 4. URL: https://idfpr.illinois.gov/news/2025/gov-pritzker-signs-state-leg-prohibiting-ai-therapy-in-il.html [accessed 2026-03-18]

Abbreviations

AI: artificial intelligence
APA: American Psychological Association
COREQ: Consolidated Criteria for Reporting Qualitative Research
GenAI: generative artificial intelligence

Edited by A Stone; submitted 04.Dec.2025; peer-reviewed by N Rodax, Z Liu; accepted 23.Feb.2026; published 02.Apr.2026.

Please cite as:
Kuang J, Pope AL, Zhang Y
Psychotherapists' Trust, Distrust, and Generative AI Practices in Psychotherapy: Qualitative Study
J Med Internet Res 2026;28:e88932
URL: https://www.jmir.org/2026/1/e88932
doi:10.2196/88932

© Jinyan Kuang, Amber L Pope, Yixuan Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Alberta Quality Assessment Tool: Risk of Bias (AQAT:RoB) for the Evaluation of Medical Large Language Model Question-Answer Studies: Development and Pilot Validation

Carrie Ye[1,2,3], MPH, MD; Joseph Ross Mitchell[1,4], PhD; Daniel C Baumgart[1], MBA, MD, PhD; Zechen Ma[1], MD; Angela Lim Fung[1,5], MSc; Daniela Garcia Orellana[1], MSc; Juel Chowdhury[1,6], MBBS, MPH; Abdullah Abass[1], MD; Steven Katz[1,7], MD; Jacob L Jaremko[1], MD, PhD; Pierre Boulanger[1], PEng, PhD; Claire E H Barber[8], MD, PhD; Gillian Lemermeyer[1], RN, PhD; Hosna Jabbari[1], PhD; Lili Mou[1,4], PhD; Maryam Mirzaei[9], PhD; Mary Waithera Beckett Githumbi[5], MBA; Puneeta Tandon[1], MSc, MD; Randy Goebel[1,4], PhD; Rhys Clark[3], BSc; Whitney Hung[3], MD; Marjan Abbasi[1], MD; Farhad Maleki[8], PhD; Scott Klarenbach[1], MSc, MD; Mohamed Abdalla[1,4], PhD

[1]University of Alberta, 8-130 Clinical Sciences Building, 11350 83 Ave NW, Edmonton, AB, Canada
[2]Arthritis Research Canada, Vancouver, BC, Canada
[3]Alberta Health Services, Edmonton, AB, Canada
[4]Alberta Machine Intelligence Institute, Edmonton, AB, Canada
[5]University of Toronto, Toronto, ON, Canada
[6]University of Oxford, Oxford, United Kingdom
[7]Queen's University, Kingston, ON, Canada
[8]University of Calgary, Calgary, AB, Canada
[9]NAIT Applied Research, Edmonton, AB, Canada

Corresponding Author:
Carrie Ye, MPH, MD
University of Alberta, 8-130 Clinical Sciences Building, 11350 83 Ave NW, Edmonton, AB, Canada

Abstract

Background: Despite the transformative potential of large language models (LLMs) in health care, the rapid development of these tools has outpaced their rigorous evaluation. While artificial intelligence–specific reporting guidelines have been developed to address standardized reporting of artificial intelligence studies, there is currently no specific tool available for risk of bias assessment of LLM question-answer (QA) studies. Existing risk-of-bias tools for medical research are not well suited to the unique challenges of evaluating LLM-QA studies, which creates a critical gap in assessing their safety and effectiveness.

Objective: This study aims to develop the Alberta Quality Assessment Tool: Risk of Bias (AQAT:RoB) for LLM-QA studies to systematically evaluate the validity and risk of bias in LLM-QA studies.

Methods: We conducted 2 literature reviews. The first was on quality assessment tools for LLM-QA studies, and the second was on LLM-QA studies, which informed the first draft of the AQAT:RoB. The draft AQAT:ROB was further refined through a prespecified iterative process of modified Delphi, consensus meeting, and validation. The first Delphi process occurred between May 1 and May 20, 2025, and the first consensus meeting was held on May 22. The first round of validation was completed by 4 evaluators, who were not part of the consensus meeting, on 16 randomly selected studies. As this first round of validation surpassed our a priori threshold of ≥80% agreement and a Cohen κ of ≥0.61 between evaluators, no further rounds of development and validation were undertaken. A second Delphi process occurred between February 20 and February 23, 2026, to vote on postpilot changes in response to peer review.

Results: The AQAT:RoB consists of 5 high-level domains (Questions, Reference Answers, LLM Answers, Evaluators, Outcomes). These domains are subdivided into 9 subdomains. Each subdomain includes at least one “Support for Judgment” and at least one “Type of Bias” and is to be rated “low,” “high,” or “unclear” for risk of bias. A pilot evaluation was completed by internal validators who were not part of the consensus discussion and were asked to complete the AQAT:RoB form for each assigned study. Each of the 16 studies was evaluated by 2 evaluators independently. Pilot validation showed a percent agreement of 86.1% and a Cohen κ of 0.70 between assessors.

Conclusions: The AQAT:RoB demonstrates promising initial reliability for assessing the validity or risk of bias in LLM-QA studies. The tool will benefit from future refinements, external validation, and periodic updates to keep pace with evolving technology.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e87057) doi:10.2196/87057

KEYWORDS

risk of bias; quality assessment; large language model; question-answer studies; Alberta Risk of Bias Assessment Tool for LLM-QA studies; AQAT: RoB; chatbot; artificial intelligence

Introduction

Large language models (LLMs) represent a significant technological advancement with transformative potential across various sectors, including health care. Their capabilities in processing and generating human-like text have led to their rapid emergence as tools capable of assisting in complex medical activities, such as disease diagnosis, clinical decision-making, and even administrative tasks, such as writing prescriptions or assigning billing codes [1]. As these sophisticated tools become more integrated into health care ecosystems, robust and rigorous evaluation of their efficacy, safety, and utility is paramount. A critical component of this evaluation involves human assessments, where the performance, usability, and impact of LLM question-answer systems (LLM-QA) such as medical chatbots are gaged through interactions with health care professionals, patients, or simulated users [2,3].

However, the quality of these human evaluation studies varies significantly, with a systematic review indicating that only 5% of studies used real patient care data for LLM evaluation, which can significantly impact the trustworthiness and generalizability of their findings [2]. Without a systematic approach to evaluating the quality and risk of bias in studies assessing LLM-QA, the rapid pace of development could outpace the generation of reliable evidence regarding their actual utility and safety in real-world clinical scenarios, potentially leading to premature or even harmful adoption.

Existing tools for assessing risk of bias, such as the Cochrane Risk of Bias 2 tool (RoB 2) [4] for randomized studies, the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) [5] for diagnostic test studies, the Newcastle-Ottawa Scale (NOS) [6], and the Risk of Bias in Non-randomized Studies - of Exposures tool (ROBINS-E) [7] for nonrandomized studies, are foundational in their respective areas but fall short when applied to the unique methodological and reporting challenges inherent in human evaluation studies of LLMs. While artificial intelligence (AI)–specific quality assessment tools exist, such as the Prediction model Risk of Bias Assessment Tool + AI (PROBAST-AI) [8] and APPRAISE-AI [9], these focus on studies of prediction models using machine learning and are not applicable to LLM-QA studies.

The burgeoning interest in the field is evident from the notable surge in studies pertaining to LLM medical chatbots published in recent years, underscoring the topic's emerging relevance and the urgent need for robust evaluation methodologies [10,11]. AI-specific reporting guidelines have been developed to address standardized reporting of AI studies [9,12-15], including a reporting checklist specifically for chatbot health advice studies, CHART (Chatbot Assessment Reporting Tool) [13]. However, transparent and comprehensive reporting is only one aspect of quality assessment—the other being assessment of risk of bias.

There is currently no tool available to assess the risk of bias in LLM-QA studies. This gap creates a significant challenge for researchers, clinicians, and policymakers attempting to synthesize evidence and make informed decisions about the integration of medical LLM-QA systems. Without a comprehensive and tailored risk of bias assessment tool, the risk of misinterpreting findings, perpetuating methodological flaws, and drawing unsubstantiated conclusions from human evaluation studies is high. To address this knowledge gap, we took a pragmatic but systematic approach to develop and validate the Alberta Quality Assessment Tool: Risk of Bias (AQAT:RoB) for LLM-QA studies for the systematic and comprehensive assessment of the risk of bias in medical LLM-QA studies, addressing aspects unique to this emerging field. The tool is intended to evaluate the quality of studies that involve human participants in assessing the outputs of AI models that utilize natural language interactions.

Methods

Overview

The AQAT:RoB development and validation started with the 2 literature reviews. The first was on quality assessment tools for LLM-QA studies, and the second was on LLM-QA studies, which informed the first draft of AQAT:RoB, which went through an iterative process of modified Delphi, consensus meeting, and validation, until our a priori threshold for interrater agreement was met (Figure 1).

XSL•FO
RenderX

Figure 1. Alberta Quality Assessment Tool: Risk of Bias (AQAT:RoB) development and validation. This figure outlines the development and pilot validation process (April-September 2025) that was a priori determined and followed to create the AQAT:RoB. The dotted line implies a possible path that would have been followed had the agreement threshold not been met (though this was ultimately not required). LLMs: large language models; LLM-QA: LLM question-answer; RoB: risk of bias. LLM-QA studies include [10,11] and a forthcoming systematic literature review of studies evaluating LLMs for patient-facing health information (protocol registered with PROSPERO; CRD42023461630 [16]).

XSL•FO
RenderX

Ethical Considerations

The University of Alberta's Research Ethics Board deemed that this project meets one of the conditions described under Chapter 2 of *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans* (2022) [17] as an activity that does not require Research Ethics Board review. The AQAT:RoB has been registered with the LATITUDES Network, which was established to increase the robustness of evidence synthesis by improving the process of validity (risk of bias) assessment [18].

Literature Review

We conducted a search query in PubMed in April 2025 to look for existing risk of bias assessment tools specific to LLM-QA studies. The query searched for all studies that contained either of the terms "risk of bias" or "quality assessment" with any of the terms "large language models," "generative AI," or "chatbot" [("risk of bias" OR "quality assessment") AND ("large language models" OR "generative AI" OR "chatbot")]. We limited the search to English studies published in the last 5 years and retrieved 91 results. The search was updated in June 2025, which retrieved 113 studies, and again in September 2025, which retrieved 149 studies. None of these searches included tools for assessing validity or risk of bias in LLM studies. Most studies pertained to the use of LLMs to perform risk of bias assessments. Around the time of each PubMed search, the LATITUDES Network library [19] of risk of bias assessment tools and tools in development was searched for tools that pertained to LLM studies; none were found.

We conducted a literature review of LLM-QA studies to (1) inform the development of AQAT:RoB and (2) find studies for validating the AQAT:RoB. Our group had already conducted, with the assistance of an experienced librarian (DCB), a systematic literature review of studies evaluating LLMs for patient-facing health information (manuscript in progress, protocol registered with PROSPERO [CRD42023461630]) [16]. Medline, Embase, Web of Science, CINAHL, PsycINFO, and Google Scholar were searched for studies published up to July 5, 2023, and then updated to March 7, 2024, limiting searches to the last 10 years prior to the search (Multimedia Appendix 1). Searches were limited to the English language but not to geographic regions. All citations were imported into Covidence for duplicate removal and screening. Abstracts and full texts were each screened independently by 3 reviewers (ZM, ALF, DGO). Disagreements were resolved through a third reviewer (CY). A total of 8943 records were identified, including 2798 duplicates (Multimedia Appendix 2). In total, 6145 titles and abstracts were screened with 327 found to be relevant for full-text review. Of these 327 full texts, 40 were deemed to be original research studies pertaining to LLMs for patient education (Multimedia Appendix 3).

We found 2 very recently published systematic reviews by Wang et al [11] and Huo et al [10], which, along with our systematic review [16], we felt covered the breadth of LLM-QA studies and were thus sufficient to inform the development and validation of the AQAT:RoB. In the systematic literature review conducted by Huo et al [10], their search of MEDLINE, Embase, and Web of Science from inception to October 27, 2023, resulted in 137 eligible studies evaluating the performance of generative AI-driven chatbots. They found that key aspects of internal validity, such as standardized evaluation process, blinding of evaluators, or reference standards, were not well described or simply not included in the studies [10]. Wang et al [11] queried PubMed, Embase, Web of Science, and Scopus from inception until October 14, 2024, and found 168 studies on the accuracy of LLMs when answering clinical questions (although one of these studies has since been retracted). They performed a risk of bias assessment using the NOS [6,11]. Acknowledging the limitations of using the NOS tool given the limited relevance to these studies, they found that only 40 (23.8%) of 168 studies were assessed as having a low overall risk of bias [11].

Candidate Item List Generation

The initial list of items for the AQAT:RoB assessment tool was drafted by CY and MA by reviewing the results and studies of the 3 recent systematic literature reviews [10,11,16] to identify potential sources of bias in LLM-QA and by adapting existing foundational risk of bias assessment tools, including RoB 2 [20], NOS [6], QUADAS-2 [21], and ROBINS-E [7]. The steering committee (CY, MA, JRM, DCB) further developed the initial list of items to form the first draft of the AQAT:RoB (7 domains, 12 sources of bias, 15 supports). This first draft was used in the modified Delphi procedure described below.

Recruitment of Delphi Panelists

To identify participants for the modified Delphi process, the steering committee recruited medical AI experts through the Alberta Machine Intelligence Institute [18,22], the University of Alberta AI + Health Hub [23], and the University of Calgary's Centre for Health Informatics [24]. Interested participants were asked to complete an intake questionnaire regarding demographics, training, and related expertise. All applicants were screened, and panelists were selected by the steering committee to ensure appropriate expertise and diversity and representation on the Delphi panel. In total, there were 19 Delphi panelists, including clinicians (8), computer scientists (8), methodologists (4), researchers (qualitative and quantitative, 17), journal editors (2), and patient partners (2). All Delphi participants were instructed to watch 2 videos that provided background information on risk assessment tools prior to initiating the Delphi process.

Modified Delphi Process

The modified Delphi process occurred between May 1 and May 20, 2025. The steering committee created the Delphi survey using Google Forms. Each participant responded to the survey individually. Participants were asked to rate each item's potential as a source of bias on a 5-point scale (from 1="Not a potential source of bias" to 5="High potential source of bias"). If participants selected either 1="Not a potential source" or 2="Unlikely to be a potential source," they were prompted to provide a short explanation. Participants were also encouraged to use the free-text boxes that followed each item to identify any missing potential sources of biases, questions to identify biases (ie, supports), or types of bias identified by listed supports. The threshold for removing items was more than 50% (10/19) of the participants voted the item as 1="Not a potential source" or 2="Unlikely to be a potential source," and the

XSL•FO
RenderX

threshold for retaining items was less than 50%. Participants were not able to see other participants' votes or comments.

Changes From the Delphi Process

Most items (8 of 12) were rated highly for inclusion (ie, more than 70% (14/19) rated as a potential source of bias). Items rated poorly for agreement or for potential source of bias, and all comments provided by the participants in the free-text boxes were considered for changes (eg, removal, modification, or merging). Based on feedback from participants, no domains or items were added or removed, but modifications were made to 5 items.

Consensus Meeting

An online consensus meeting was held on May 22, 2025, chaired by the steering committee. All panelists were invited except for 2 (S Katz and JLJ) who were excluded from the discussion to serve as adjudicators in the validation phase. The consensus meeting was chaired by the steering committee, and 16 participants attended the synchronous consensus meeting. During this meeting, participants were presented with the initial version of texts; suggested modified versions of the text incorporating the feedback from the modified Delphi process, as well as summary statistics of ratings; and provided comments. During the meeting, each potential source of bias was discussed until consensus was reached on inclusion, type of bias(es), and wording. Once consensus was felt to be reached based on the panel discussion, a formal vote was taken, and unanimity was required before moving on to the next item. At the end of each domain, the panel was asked to discuss if there were any additional potential sources of bias pertaining to that domain.

Changes From the Consensus Meeting

The discussions during the consensus process resulted in multiple changes. There was robust discussion about how granular the "Types of biases" should be, with the group arriving at the conclusion that for the sake of utility and widespread applicability, we would aim for a high-level description of the types of biases. This change affected 5 of 12 "Potential Sources of Bias." Furthermore, there was an addition of "Support for Judgment" for the domain "Performance Metrics."

Pilot Validation

We piloted the AQAT:RoB on 16 studies [25-40], randomly selected from the 319 studies (after removal of duplicates) identified in 3 recent systematic literature reviews [10,11,16] (Multimedia Appendix 4). Random selection was facilitated by listing all 319 studies in alphabetical order of the first author's last name and then using a random number generator (between 1 and 319) to select the 16 studies [41]. Four evaluators (S Katz, JLJ, JC, AA) were asked to complete the AQAT:RoB form for each assigned study. None of the 4 evaluators were part of the consensus discussion, but S Katz and JLJ were on the modified Delphi panel. JC and AA were not part of the tool development process prior to the validation step. All the evaluators were physicians across various specialties (rheumatology, radiology, public health, and primary care). Each of the 16 studies was evaluated by 2 evaluators independently. Evaluators were not provided with any standardized training in order to obtain the most conservative estimates of agreement.

We set an a priori threshold of >80% agreement and a Cohen κ of ≥0.60, which would demonstrate substantial agreement per Landis and Koch's [42] classification system. If we reached this threshold, no further rounds of changes or consensus would be undertaken. If we did not reach this threshold, we planned to pursue further rounds of modified Delphi or consensus or validation until this threshold was achieved (Figure 1).

Protocol Deviations in Response to Peer Review

The steering committee proposed moving the Reporting and Conflict of Interest domains to "Additional consideration" for the Delphi panel after the first round of validation was completed. The rationale was that Reporting was better assessed by stand-alone reporting checklists, and while reporting may lead to an unclear risk of bias judgment, it does not represent a mechanistic bias similar to the other domains in the AQAT:RoB. Similarly, Conflict of Interest may be a predictor or source of bias rather than a distinct mechanism of bias. We set the voting threshold to make this change at 70% for the Delphi panel and planned to conduct another round of validation if the overall percent agreement and Cohen κ after removing these 2 domains did not reach our a priori threshold of >80% agreement and Cohen κ of >0.60. Note that 100% of the panel voted in agreement with this change.

Results

Validation Results

After the first modified Delphi and consensus panel, we met our threshold with a percent agreement of 82.8% and a Cohen κ of 0.63 (calculated across all items; Multimedia Appendix 5). After the second modified Delphi, in response to peer review, during which 2 domains were removed, the percent agreement was 86.1% and Cohen κ was 0.70. The domain with the highest level of agreement was Question Selection (agreement: 93.8%, κ: 0.86), while the domain with the lowest level of agreement was LLM Answer Selection (agreement: 68.8%, κ: 0.30).

AQAT:RoB Tool

The AQAT:RoB tool [43] is summarized in Multimedia Appendix 6. An easy-to-use version, which is both fillable and printable, is available on the AQAT website [43]. Textbox 1 presents the scope and boundaries of studies for which AQAT:RoB is applicable.

XSL•FO
RenderX

Textbox 1. Utilization of Alberta Quality Assessment Tool: Risk of Bias (AQAT:RoB).

Intended users: Anyone who wants to appraise the quality or risk of bias of studies in which there are human evaluations of large language model question-answer systems. It is especially important for researchers during the development and peer review of such studies and during the quality assessment stage of systematic literature reviews and meta-analyses. Other potential users include, but are not limited to, patients, health care providers, journal editors and reviewers, medical technology manufacturers, health system administrators, and policymakers.

Target studies: Any study that involves the human evaluation of large language models that provide answers to free-text questions including but not limited to:

- Patient-facing applications
  - Medical chatbots
  - Summary tools (eg, answer questions about imaging reports or doctors' reports)
- Physician-facing applications
  - General medical chatbots
  - Large language model–based clinical decision support systems for physicians
  - Summary tools (eg, answer questions about patient charts)
- Research-based applications:
  - Case finding (eg, find participants based on electronic medical record or electronic health record data and provide justification)
  - Literature review (eg, analyze and summarize scientific literature)

The AQAT:RoB consists of 5 high-level domains (Questions, Reference Answers, LLM Answers, Evaluators, and Outcomes). These domains are subdivided into 9 subdomains. Each subdomain includes at least one "Support for Judgment" and at least one "Type of Bias." Further descriptions of potential sources of bias and best methodological practices are outlined in the text below. In cases of missing, partial, or suboptimal reporting of a specific domain or subdomain, the rating of "unclear" should be assigned. Additional considerations regarding reporting and conflicts of interest are outlined in the *Additional Considerations* section.

Domain 1: Questions

Question Source

Supports for Judgment:

- If questions were created or generated specifically for the study, describe the method used to create the question dataset, including who created the questions and if the questions are reflective of the intended study objective.
- If questions were selected from an existing question source, adequately describe the source to allow an assessment of whether it addresses the intended research question.

The evaluation of LLM-QA models should be conducted against questions that reflect the intended use case, as deviations can introduce biases. When the deviation between intended use and the question source is substantial, the performance on the proxy task may not be generalizable to the stated application. To minimize this risk, the most effective approach is to source questions directly from the real-world use setting. If this is not possible, external data sources are often used to generate questions. In such cases, researchers must justify the degree to which these sources align with the study's core research question and the tool's intended application. For example, if a study evaluates a tool to be used by patients, but the questions are written by a research team of nonpatients, this could introduce bias, as the language and complexity of the questions may not be representative of the intended user. Furthermore, if questions were pulled from a preexisting source, it should be clearly stated if the test questions were included in the training data of the tested models, as performance may reflect memorization versus true model reasoning.

Question Selection

Support for Judgment:

- If questions were selected from an existing question source, describe the method used to select the questions from the original source (eg, random, consecutive, all, or by certain factors).

When selecting questions from an existing dataset, sampling can introduce bias as the selected questions may not be representative of the broader population of potential questions. For instance, a selection mechanism that favors questions of a specific length—such as those with a short, predefined character count, perhaps to minimize computational costs—would systematically exclude longer, more complex questions. Similarly, selecting from a small, nonrandom subset of available options could skew the results, as the chosen questions may not accurately reflect the diversity and range of questions encountered in the tool's intended use case. Therefore, the method for question selection, such as random, purposive, or consecutive sampling, must be clearly reported and justified.

Question Manipulation

Supports for Judgment

- If any questions were manipulated from the original source, describe and justify the rationale for the manipulation.
- If any prompting was provided in addition to the index question, report the exact wording of the prompt(s).

XSL•FO
RenderX

Whether questions are created or extracted from existing sources, researchers may choose to manipulate them for various reasons. For example, slight variations might be introduced to test the model's robustness and bias, assessing how stable its responses are to minor changes in phrasing. Such manipulations are generally less likely to introduce significant bias, as their purpose is to probe the model's inherent stability rather than to alter the nature of the query. Conversely, questions might be manipulated to simplify them for processing by the model. A common example of this is the use of a system prompt that automatically extracts and restructures clinically relevant information before the model attempts to answer. This form of question manipulation, while potentially beneficial for processing, introduces a risk of bias because it may fundamentally alter the user's original query. Or if researchers correct spelling, terminology, or grammatical errors, or split multipart patient questions into separate questions, these changes may augment the performance of the LLM-QA model but not necessarily reflect real-world performance. It is essential that researchers provide both transparency and justification for any question manipulation, as the process could alter the original intent of the question, thereby compromising the validity of the evaluation. Along with direct question manipulation, all prompts, including system prompts, which in and of themselves do not necessarily introduce bias, should be clearly described and should be standardized and stable throughout testing, as differences in prompts may lead to false or misleading performance outcomes.

Domain 2: Reference Answers

Reference Answer Source

Supports for Judgment:

- If reference answers were generated specifically for the study, describe the method used to create the reference answer dataset, including who created the reference answers, and if the answers are reflective of a true reference standard.
- If reference answers were selected from an existing reference answer source, adequately describe the source to allow an assessment of whether it is reflective of a true reference standard.

Often, LLM-QA studies benchmark LLM outputs against reference answers. Bias may be introduced if the reference answers do not accurately reflect a "true" or expected standard. For instance, if reference answers were created by individuals with a different level of expertise or with a different format or standard than the true reference standard, the reference standard may not be valid. For example, if the research team created the reference answers to a higher or lower standard than real-world physician-level responses, the reference standard would be misaligned with the intended quality benchmark. A mismatch in language, structure, style, or level of detail between the study reference standard used and the "true" real-world reference standard can lead to biased results. As a single ground truth does not always exist in medicine, the selected reference standard should be decided a priori (eg, guideline-based or expert consensus-based) and described and justified.

Reference Answer Selection

Support for Judgment:

- If not all reference answers to a given question were used, describe the method by which reference answers were selected.

Just as with question selection, the process of selecting reference answers from a larger pool can introduce sampling bias. This bias occurs if the selection method systematically favors answers with certain qualities, making the final set of reference answers unrepresentative of the full range of possible correct responses. For example, if a question has multiple valid reference answers but researchers consistently choose those with a specific tone or level of detail, the evaluation will be skewed toward models that produce similar outputs. Therefore, it is critical to describe and justify the method used for selecting reference answers.

Domain 3: LLM Answers

Support for Judgment:

- Describe how many answers were generated for each question and if not all answers were assessed, describe how answers were selected for assessment.

When evaluating language models, it is often prudent to generate multiple answers for a single question to assess the model's stability or to explore the diversity of its outputs. In such instances, only evaluating a subset of the generated answers (eg, choosing the best one) may not be representative of the model's typical performance, thereby leading to an inaccurate or misleading evaluation. Therefore, it is crucial for researchers to transparently describe and justify how many answers were generated for each question and, if not all of them were assessed, to detail the specific methodology used to select the answers for evaluation.

Domain 4: Evaluators

Evaluator Selection

Support for Judgment

- Describe the method used to select evaluators, and assign evaluators to specific LLM qualities.

The selection of evaluators should reflect the intended real-world use and the required expertise to judge the domains being assessed. For example, having a physician evaluate the empathy of an LLM-QA's outputs may not reflect how a patient would assess this domain. Likewise, it would not be appropriate for a patient to evaluate the accuracy of health information generated by an LLM-QA, as they would lack the appropriate expertise. As many studies assess multiple outcomes, more than one type of evaluator may be required for a given study (eg, physicians evaluate accuracy and patients rate readability). Furthermore, the demographic or professional characteristics of the evaluators should align with the intended user population of the LLM or chatbot. For example, if an LLM is designed for a general audience but its readability is evaluated exclusively by individuals with advanced academic degrees, the results may not accurately reflect how an average user would perceive the content.

XSL•FO
RenderX

Blinding of Evaluators

Support for Judgment:

- Describe all measures used, if any, to blind trial evaluators and researchers from knowledge of the answer source. Provide information relating to whether the intended blinding was effective.

The integrity of an evaluation can be compromised if evaluators are not blinded to the source of the answers they are assessing (reference standard vs LLM-generated) because evaluators' preexisting beliefs, attitudes, or knowledge about a specific technology, such as AI, or even to a specific LLM model, can unconsciously influence their ratings. Thus, evaluators should be blinded to the answer source, and researchers should describe the blinding measures employed. Since naive blinding is not guaranteed to be effective given stylistic markers in LLM-generated text, authors should describe the steps taken to assess or verify the effectiveness of the blinding (eg, ask the evaluators if they could identify the AI-generated answer).

Domain 5: Outcomes—Performance Metrics

Supports for Judgment:

- Describe specific metrics used for each outcome quality.
- Describes if desired outcomes were prespecified prior to conducting the study.

To minimize the risk of bias, 2 crucial steps should be taken. First, the desired outcomes or hypotheses of the study should be prespecified prior to conducting any analysis. Second, the metrics selected to measure each outcome must directly align with the stated goals of the study. For example, if the goal is to evaluate a chatbot's ability to provide concise summaries of medical information, metrics should focus on conciseness and accuracy, rather than on secondary qualities, such as conversational tone or creativity. A misalignment between metrics and study goals introduces bias, as the evaluation would not accurately reflect the model's performance on its intended task.

Additional Considerations

Complete and transparent reporting is required to judge the risk of bias. Researchers must account for any instances of missing data and describe how missing data were handled. For example, if certain questions were too long for the model to process, or if the model failed to produce a response, these omissions should be explicitly noted and their potential impact on the evaluation should be discussed. Similarly, if human evaluators did not complete all of their annotations, it is important that this missingness is reported and ideally investigated, as these instances of missingness may not be random and could introduce bias if not accounted for. Study outcomes should be decided *a priori* and deviations should be described and justified. By reporting a subset of all measured outcomes or manipulating the analysis post hoc to achieve a different result (eg, by recategorizing certain groups, shifting the scale), a study may present a distorted or optimistic view of model performance. The use of appropriate reporting checklists is recommended.

Conflicts of interest (especially commercial interests) may introduce explicit or subconscious biases in the formulation of the problem, the execution of the analysis, or the interpretation of the results. If there are conflicts of interest (eg, authors funded or affiliated with model vendors), they must be disclosed and mitigated, if possible.

Discussion

Principal Findings

The AQAT:RoB tool was developed to standardize the quality assessment and specifically, the risk of bias assessments, of studies in which there are human evaluations of LLM-QA systems. This easy-to-use risk of bias assessment tool covers 5 major domains (Questions, Reference answers, LLM answers, Evaluators, Outcomes, Reporting, and Other), with a total of 9 potential sources of bias, each with 1 - 2 support for judgment prompts and a list of types of potential bias. In our pilot validation by 4 assessors of 16 studies, the AQAT:RoB showed a substantial degree of agreement between internal validators blinded to the development process.

Comparison to Prior Work

As demonstrated in Table 1, the AQAT:RoB assesses many important aspects of LLM-QA studies that are either not covered at all or only tangentially covered by other foundational risk of bias tools [4-7]. Most notably, Domain 1: Questions, a very important potential source of bias, is not addressed by any of the foundational tools. We also note that the panel explicitly discussed and voted to include conflict of interest in the AQAT:RoB, which is not always included in foundational tools, given the increasing commercialization of natural language processing research. The vast majority of computer science faculty at top schools have financial conflicts with industry [44], and the field of natural language processing is so reliant on industry artifacts to the point of being described as "captured" [45]. This is particularly relevant in evaluation works, as past peer-reviewed evaluations have then been used by the relevant industry party to claim that their models have been "independently audited" [46]. The foundational tools listed in Table 1 [4-7] either do not provide a threshold for determining overall risk of bias or use a "worst-of" approach, where the overall risk of bias is considered "High risk" if at least 1 domain is judged as "High risk." We have chosen to leave the determination of overall risk of bias to the discretion of the user, as the threshold may be different depending on the intended use of the tool, although in most cases, a single domain being judged as "High risk" would likely result in an overall judgment of "High risk."

XSL•FO
RenderX

Table . Comparison of Alberta Quality Assessment Tool: Risk of Bias (AQAT:RoB) and foundational risk of bias tools[a].

;AQAT:RoB domain	RoB 2[b] [4]	NOS[c] [6]	QUADAS-2[d] [5]	ROBINS-E[e] [7]
Questions	x[f]	x	x	x
Reference answers	x	• Domain 1: Selection • Domain 2: Comparability	• Domain 3: Reference standard • Domain 4: Flow and timing	• Domain 2: Risk of bias arising from measurement of the exposure
LLM[g] answers	x	• Domain 1: Selection • Domain 2: Comparability	• Domain 2: Index test or tests • Domain 4: Flow and timing	• Domain 2: Risk of bias arising from measurement of the exposure
Evaluators	• Domain 4: Risk of bias in measurement of the outcome	• Domain 1: Selection • Domain 3: Outcomes	• Domain 1: Patient selection	• Domain 3: Risk of bias in selection of participants into the study (or into the analysis)
Outcomes	• Domain 4: Risk of bias in measurement of the outcome	• Domain 3: Outcomes	x	• Domain 6: Risk of bias arising from measurement of the outcome

[a]This table compares the coverage of key domains determined to be important for assessing risk of bias in large language model question-answer studies by the Alberta Quality Assessment Tool: Risk of Bias (AQAT:RoB) compared with existing foundational risk of bias tools. No AQAT:RoB domains were covered adequately by other Foundational Risk of Bias tools. Terms in italics signify the closest domain found in the foundational tool.

[b]RoB 2: Cochrane Risk of Bias 2 tool.

[c]NOS: Newcastle-Ottawa Scale.

[d]QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2.

[e]ROBINS-E: Risk of Bias in Non-randomized Studies - of Exposures.

[f]x: not covered by the tool.

[g]LLM: large language model.

The AQAT:RoB is the first risk of bias assessment tool specifically designed for studies of human-evaluated LLM-QA systems. Despite this, systematic reviews of LLM-QA studies have already been published [10,11], highlighting the urgency of the need for the AQAT:RoB. Previous AI evaluation frameworks primarily function as reporting checklists [9,12,13,15]. While these are invaluable for assessing transparency and reproducibility, they do not directly evaluate a study's susceptibility to bias. For instance, a study may meticulously report every detail, yet still contain high-risk elements, such as a lack of blinding for answer sources, that could compromise its findings. Existing AI-specific quality assessment tools, such as PROBAST-AI [8] and APPRAISE-AI [9], are tailored for predictive machine learning models, which differ significantly from the evaluation needs of LLM-QA studies. Templin et al [47] introduced a useful 5-step framework for auditing LLMs, but not for assessing the validity of LLM-QA studies. Therefore, the AQAT:RoB addresses a critical void in the standardized evaluation of LLM-QA research.

Aiming to address the urgent need for a risk of bias assessment tool for this growing field and aided by our own existing [16] and recently published systematic reviews [10,11], we sought to develop the AQAT:RoB through a systematic, but pragmatic approach. With a highly engaged group of interdisciplinary experts, our pilot validation was able to achieve high interrater agreement after the first round of modified Delphi and consensus. In our pilot validation of 16 studies, the interrater reliability was on par with or better than that of existing foundational risk of bias tools. For example, studies have found that the RoB 2 demonstrated kappas consistent with "fair" agreement (0.21 - 0.40) [48,49]. Another study found that while the interrater reliability across 6 risk of bias tools for nonrandomized studies varied widely, most demonstrated intraclass correlation coefficients in the substantial range, similar to the AQAT:RoB [50]. No studies have demonstrated the interrater reliability of existing RoB tools specifically to LLM-QA studies.

Limitations

We recognize that there are limitations to the AQAT:RoB. First, while it was developed by a wide interdisciplinary group of experts and patient partners, most experts were from Alberta due to the nature of the AQAT collaborative, which includes the development of other quality assessment tools for AI-related studies, including the development of publicly available datasets, validated evaluation scales, and measurements and programs. In future updates of the AQAT:RoB, we plan to engage a more international group of partners. It will also be crucial for the AQAT:RoB to be extensively and externally validated by the broader international community of researchers [51]. Furthermore, the AQAT:RoB was developed with a medical focus and would require validation and adaptation for nonmedical studies of LLM-QA. Second, we recognize that this tool will need to evolve with the rapid development of LLM tools and related studies. Third, this tool was developed in

XSL•FO
RenderX

English and evaluated only on English-language studies. In order to use this tool on non-English–language studies, it would ideally be translated and validated in other languages. Multilingual and non-English evaluations may require an expansion of the support for judgments to be considered when classifying the risk of bias. For example, if questions or reference answers are translated from English, they may not accurately reflect the distribution against which they will be evaluated (ie, native, nontranslated questions).

Finally, the pilot validation has limitations. It included only 16 studies, a relatively small sample, which may limit reliability estimates. While construct validity was supported by expert consensus and signaling questions that map directly to known mechanisms of bias, our pilot validation focused on interrater reliability, the most commonly reported evaluation metric for such tools. To further support construct validity, future evaluations that test concurrent and criterion validity are needed, although the lack of truly comparative tools and meta-analyses in this field limits the feasibility of these types of evaluations. Finally, reliance on an aggregate score for our a priori threshold makes it possible for domains with high agreement to compensate for domains with poor agreement. We acknowledge that agreement for the LLM Answer Selection domain was lower or less reliable than for the other domains and remains experimental, requiring future refinements after more extensive validation.

Future Directions

The AQAT:RoB is a crucial next step toward standardizing the evaluation of LLM-QA studies in medicine. While we acknowledge that the tool will benefit from future refinements, more extensive validation (particularly external validation), and periodic updates to keep pace with evolving technology, we believe it currently fills an urgent need and critical gap. The immediate application of this tool will enable researchers, clinicians, and policymakers to more effectively and rigorously assess the validity of LLM-based studies, thereby ensuring that real-world applications of this technology are built on a solid foundation of reliable evidence.

Conclusions

The AQAT:RoB demonstrates promising initial reliability for assessing the validity or risk of bias of LLM-QA studies.

Acknowledgments

We thank the Alberta Machine Intelligence Institute, AI+Health Hub, and the Centre for Health Informatics for providing administrative and recruitment support. We thank Ms Dagmara Chojecki, MLIS, health sciences librarian, for her assistance with the systematic literature review.

Funding

This research is supported by the Canadian Institutes of Health Research (number 96047). MA is supported by a Canada CIFAR AI Chair. RM is supported by a Canada CIFAR AI Chair and the Alberta Health Services Chair in AI in Health. LM is supported by NSERC and Canada CIFAR AI Chair. RG is supported by Amii and NSERC. JM is supported by a Canada CIFAR AI Chair and by Medical Imaging Consultants. SK is supported by the Kidney Health Research Chair and the Division of Nephrology at the University of Alberta.

Data Availability

The data extracted and synthesized for this study are available in the multimedia appendix files.

Authors' Contributions

Conceptualization: CY
Data curation: CY, MA
Formal analysis: CY, MA
Investigation: AA, ALF, CEHB, CY, DCB, DGO, FM, GL, HJ, JC, JLJ, JRM, LM, MA, MM, MWBG, PB, PT, RC, RG, S Katz, S Klarenbach, WH, ZM
Methodology: CY, DCB, JRM, MA
Writing – original draft: CY, MA
Writing – review & editing: AA, ALF, CEHB, CY, DCB, DGO, FM, GL, HJ, JC, JLJ, JRM, LM, MA, MM, MWBG, PB, PT, RC, RG, S Katz, S Klarenbach, WH, ZM

Conflicts of Interest

S Klarenbach is Director of the Real World Evidence Consortium, and Alberta Drug and Therapeutic Evaluation Consortium (Universities of Alberta, Calgary, and Institute of Health Economics); these entities receive funding from decision-makers and industry to conduct research. All research funding is made to the academic institution; investigators retain full rights of academic freedom and right to publish. This relationship is not related to the current work. All other authors declare no conflicts of interest.

Multimedia Appendix 1

XSL•FO
RenderX

Search strategy for the large language model (LLM) patient education systematic literature review.
[DOCX File, 3124 KB - jmir_v28i1e87057_app1.docx]

Multimedia Appendix 2
PRISMA flow diagram of unpublished systematic literature review of studies evaluating large language models (LLMs) for patient-facing health information published between July 2013 and March 2024.
[PNG File, 140 KB - jmir_v28i1e87057_app2.png]

Multimedia Appendix 3
Large language model (LLM) patient education systematic literature review list of studies reviewed.
[PDF File, 115 KB - jmir_v28i1e87057_app3.pdf]

Multimedia Appendix 4
Studies used in the pilot validation set (n=16 studies).
[PDF File, 81 KB - jmir_v28i1e87057_app4.pdf]

Multimedia Appendix 5
Pilot validation dataset. Sixteen papers were each assessed by 2 evaluators for a total of 32 evaluations.
[XLSX File, 12 KB - jmir_v28i1e87057_app5.xlsx]

Multimedia Appendix 6
Alberta Quality Assessment Tool: Risk of Bias (AQAT:RoB).
[DOCX File, 15 KB - jmir_v28i1e87057_app6.docx]

References

1. Iqbal U, Tanweer A, Rahmanti AR, Greenfield D, Lee LTJ, Li YCJ. Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. J Biomed Sci 2025 May 7;32(1):45. [doi: 10.1186/s12929-025-01131-z] [Medline: 40335969]
2. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 2025 Jan 28;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
3. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA 2023 Sep 5;330(9):866-869. [doi: 10.1001/jama.2023.14217] [Medline: 37548965]
4. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
5. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
6. Lo CKL, Mertz D, Loeb M. Newcastle-Ottawa Scale: comparing reviewers' to authors' assessments. BMC Med Res Methodol 2014 Apr 1;14:45. [doi: 10.1186/1471-2288-14-45] [Medline: 24690082]
7. Higgins JPT, Morgan RL, Rooney AA, et al. A tool to assess risk of bias in non-randomized follow-up studies of exposure effects (ROBINS-E). Environ Int 2024 Apr;186:108602. [doi: 10.1016/j.envint.2024.108602] [Medline: 38555664]
8. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021 Jul 9;11(7):e048008. [doi: 10.1136/bmjopen-2020-048008] [Medline: 34244270]
9. Kwong JCC, Khondker A, Lajkosz K, et al. APPRAISE-AI tool for quantitative evaluation of AI studies for clinical decision support. JAMA Netw Open 2023 Sep 5;6(9):e2335377. [doi: 10.1001/jamanetworkopen.2023.35377] [Medline: 37747733]
10. Huo B, Boyle A, Marfo N, et al. Large language models for chatbot health advice studies: a systematic review. JAMA Netw Open 2025 Feb 3;8(2):e2457879. [doi: 10.1001/jamanetworkopen.2024.57879] [Medline: 39903463]
11. Wang L, Li J, Zhuang B, et al. Accuracy of large language models when answering clinical research questions: systematic review and network meta-analysis. J Med Internet Res 2025 Apr 30;27:e64486. [doi: 10.2196/64486] [Medline: 40305085]
12. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024 Apr 16;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
13. CHART Collaborative. Reporting guideline for chatbot health advice studies: the Chatbot Assessment Reporting Tool (CHART) statement. BMJ Med 2025;4(1):e001632. [doi: 10.1136/bmjmed-2025-001632] [Medline: 40761518]
14. El Emam K, Leung TI, Malin B, Klement W, Eysenbach G. Consolidated reporting guidelines for prognostic and diagnostic machine learning models (CREMLS). J Med Internet Res 2024 May 2;26:e52508. [doi: 10.2196/52508] [Medline: 38696776]

XSL•FO
RenderX

15. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med 2025 Jan;31(1):60-69. [doi: 10.1038/s41591-024-03425-5] [Medline: 39779929]
16. The use of large language models in patient education interventions: a systematic review. PROSPERO. URL: https://www.crd.york.ac.uk/PROSPERO/view/CRD42023461630 [accessed 2025-09-03]
17. TCPS 2 (2022) – chapter 2: scope and approach. Government of Canada. 2023. URL: https://ethics.gc.ca/eng/tcps2-eptc2_2022_chapter2-chapitre2.html [accessed 2026-03-27]
18. Tools in development. Latitudes Network. 2023. URL: https://www.latitudes-network.org/library/tools-in-development/ [accessed 2025-09-02]
19. Latitudes Network. 2023. URL: https://www.latitudes-network.org/ [accessed 2025-09-16]
20. RoB 2 tool. Risk of bias tools. URL: https://sites.google.com/site/riskofbiastool/welcome/rob-2-0-tool?authuser=0 [accessed 2025-06-18]
21. QUADAS | Bristol medical school: population health sciences. University of Bristol. URL: https://www.bristol.ac.uk/population-health-sciences/projects/quadas/ [accessed 2025-06-18]
22. Alberta Machine Intelligence Institute. URL: https://www.amii.ca/ [accessed 2025-09-02]
23. AI + Health Hub. University of Alberta. URL: https://www.ualberta.ca/en/health-sciences/research/ai-and-health-hub.html [accessed 2025-09-02]
24. Centre for health informatics. University of Calgary. URL: https://cumming.ucalgary.ca/centres/centre-health-informatics [accessed 2025-09-03]
25. Bernstein IA, Zhang YV, Govil D, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Netw Open 2023 Aug 1;6(8):e2330320. [doi: 10.1001/jamanetworkopen.2023.30320] [Medline: 37606922]
26. Cappellani F, Card KR, Shields CL, Pulido JS, Haller JA. Reliability and accuracy of artificial intelligence ChatGPT in providing information on ophthalmic diseases and management to patients. Eye 2024 May;38(7):1368-1373. [doi: 10.1038/s41433-023-02906-0] [Medline: 38245622]
27. Chaker SC, Hung YC, Saad M, Golinko MS, Galdyn IA. Easing the burden on caregivers—applications of artificial intelligence for physicians and caregivers of children with cleft lip and palate. Cleft Palate Craniofac J 2025 Apr;62(4):574-587. [doi: 10.1177/10556656231223596] [Medline: 38178785]
28. Chen S, Kann BH, Foote MB, et al. The utility of ChatGPT for cancer treatment information. medRxiv. Preprint posted online on Mar 23, 2023. [doi: 10.1101/2023.03.16.23287316]
29. Chervenak J, Lieman H, Blanco-Breindel M, Jindal S. The promise and peril of using a large language model to obtain clinical information: ChatGPT performs strongly as a fertility counseling tool with limitations. Fertil Steril 2023 Sep;120(3):575-583. [doi: 10.1016/j.fertnstert.2023.05.151] [Medline: 37217092]
30. Coskun B, Ocakoglu G, Yetemen M, Kaygisiz O. Can ChatGPT, an artificial intelligence language model, provide accurate and high-quality patient information on prostate cancer? Urology 2023 Oct;180:35-58. [doi: 10.1016/j.urology.2023.05.040] [Medline: 37406864]
31. Gabriel J, Shafik L, Alanbuki A, Larner T. The utility of the ChatGPT artificial intelligence tool for patient education and enquiry in robotic radical prostatectomy. Int Urol Nephrol 2023 Nov;55(11):2717-2732. [doi: 10.1007/s11255-023-03729-4] [Medline: 37528247]
32. Liu HY, Alessandri Bonetti M, Jeong T, Pandya S, Nguyen VT, Egro FM. Dr. ChatGPT will see you now: how do Google and ChatGPT compare in answering patient questions on breast reconstruction? J Plast Reconstr Aesthet Surg 2023 Oct;85:488-497. [doi: 10.1016/j.bjps.2023.07.039]
33. Kianian R, Sun D, Giaconi J. Can ChatGPT aid clinicians in educating patients on the surgical management of glaucoma? J Glaucoma 2024 Feb 1;33(2):94-100. [doi: 10.1097/IJG.0000000000002338] [Medline: 38031276]
34. McCarthy CJ, Berkowitz S, Ramalingam V, Ahmed M. Evaluation of an artificial intelligence chatbot for delivery of IR patient education material: a comparison with societal website content. J Vasc Interv Radiol 2023 Oct;34(10):1760-1768. [doi: 10.1016/j.jvir.2023.05.037] [Medline: 37330210]
35. Padovan M, Cosci B, Petillo A, et al. ChatGPT in occupational medicine: a comparative study with human experts. Bioengineering (Basel) 2024 Jan 6;11(1):57. [doi: 10.3390/bioengineering11010057] [Medline: 38247934]
36. Thia I, Saluja M. ChatGPT: is this patient education tool for urological malignancies readable for the general population? Res Rep Urol 2024;16:31-37. [doi: 10.2147/RRU.S440633] [Medline: 38259300]
37. Mayo-Yáñez M, Lechien JR, Maria-Saibene A, Vaira LA, Maniaci A, Chiesa-Estomba CM. Examining the performance of ChatGPT 3.5 and Microsoft Copilot in otolaryngology: a comparative study with otolaryngologists' evaluation. Indian J Otolaryngol Head Neck Surg 2024 Aug;76(4):3465-3469. [doi: 10.1007/s12070-024-04729-1] [Medline: 39130248]
38. Pressman SM, Borna S, Gomez-Cabello CA, Haider SA, Forte AJ. AI in hand surgery: assessing large language models in the classification and management of hand injuries. J Clin Med 2024 May 11;13(10):2832. [doi: 10.3390/jcm13102832] [Medline: 38792374]
39. Mesnier J, Suc G, Sayah N, Abtan J, Steg PG. Relevance of medical information obtained from ChatGPT: are large language models friends or foes? Arch Cardiovasc Dis 2023 Oct;116(10):485-486. [doi: 10.1016/j.acvd.2023.07.009] [Medline: 37718185]

XSL•FO
RenderX

40. Rizwan A, Sadiq T. The use of AI in diagnosing diseases and providing management plans: a consultation on cardiovascular disorders with ChatGPT. Cureus 2023 Aug;15(8):e43106. [doi: 10.7759/cureus.43106] [Medline: 37692649]
41. Haahr M. RANDOM.ORG. URL: https://www.random.org/ [accessed 2025-09-03]
42. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]
43. Alberta Quality Assessment Tools. URL: https://aqat.ai/#tools [accessed 2026-03-18]
44. Abdalla M, Abdalla M. The grey hoodie project: big tobacco, big tech, and the threat on academic integrity. : ACM; 2021 Presented at: AIES '21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society; May 19-21, 2021; Virtual event. [doi: 10.1145/3461702.3462563]
45. Aitken W, Abdalla M, Rudie K, Stinson C. Collaboration or corporate capture? Quantifying NLP's reliance on industry artifacts and contributions. In: Ku LW, Martins A, Srikumar V, editors. : Association for Computational Linguistics; 2024 Presented at: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Aug 11-16, 2024; Bangkok, Thailand p. 3433-3448. [doi: 10.18653/v1/2024.acl-long.188]
46. Young M, Katell M, Krafft PM. Confronting power and corporate capture at the FAccT conference. : ACM; 2022 Presented at: FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency; Jun 21-24, 2022; Seoul, South Korea. [doi: 10.1145/3531146.3533194]
47. Templin T, Fort S, Padmanabham P, et al. Framework for bias evaluation in large language models in healthcare settings. NPJ Digit Med 2025 Jul 7;8(1):414. [doi: 10.1038/s41746-025-01786-w] [Medline: 40624264]
48. Minozzi S, Cinquini M, Gianola S, Gonzalez-Lorenzo M, Banzi R. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. J Clin Epidemiol 2020 Oct;126:37-44. [doi: 10.1016/j.jclinepi.2020.06.015] [Medline: 32562833]
49. Hartling L, Hamm MP, Milne A, et al. Testing the risk of bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. J Clin Epidemiol 2013 Sep;66(9):973-981. [doi: 10.1016/j.jclinepi.2012.07.005] [Medline: 22981249]
50. Kalaycioglu I, Rioux B, Briard JN, et al. Inter-rater reliability of risk of bias tools for non-randomized studies. Syst Rev 2023 Dec 7;12(1):227. [doi: 10.1186/s13643-023-02389-w] [Medline: 38057883]
51. Tomlinson E, Cooper C, Davenport C, et al. Common challenges and suggestions for risk of bias tool development: a systematic review of methodological studies. J Clin Epidemiol 2024 Jul;171:111370. [doi: 10.1016/j.jclinepi.2024.111370] [Medline: 38670243]

Abbreviations

AI: artificial intelligence
AQAT:RoB: Alberta Quality Assessment Tool: Risk of Bias
CHART: Chatbot Assessment Reporting Tool
LLM: large language model
LLM-QA: LLM question-answer system
NOS: Newcastle-Ottawa Scale
PROBAST-AI: Prediction model Risk of Bias Assessment Tool
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2
RoB 2: Cochrane Risk of Bias 2 tool
ROBINS-E: Risk of Bias in Non-randomized Studies - of Exposures tool

Edited by A Coristine; submitted 04.Nov.2025; peer-reviewed by A Njoku, G Surianarayanan, X Michelena; accepted 24.Feb.2026; published 08.Apr.2026.

Please cite as:
Ye C, Mitchell JR, Baumgart DC, Ma Z, Fung AL, Orellana DG, Chowdhury J, Abass A, Katz S, Jaremko JL, Boulanger P, Barber CEH, Lemermeyer G, Jabbari H, Mou L, Mirzaei M, Githumbi MWB, Tandon P, Goebel R, Clark R, Hung W, Abbasi M, Maleki F, Klarenbach S, Abdalla M
The Alberta Quality Assessment Tool: Risk of Bias (AQAT:RoB) for the Evaluation of Medical Large Language Model Question-Answer Studies: Development and Pilot Validation
J Med Internet Res 2026;28:e87057
URL: https://www.jmir.org/2026/1/e87057
doi:10.2196/87057

XSL•FO
RenderX

© Carrie Ye, Joseph Ross Mitchell, Daniel C Baumgart, Zechen Ma, Angela Lim Fung, Daniela Garcia Orellana, Juel Chowdhury, Abdullah Abass, Steven Katz, Jacob L Jaremko, Pierre Boulanger, Claire E H Barber, Gillian Lemermeyer, Hosna Jabbari, Lili Mou, Maryam Mirzaei, Mary Waithera Beckett Githumbi, Puneeta Tandon, Randy Goebel, Rhys Clark, Whitney Hung, Marjan Abbasi, Farhad Maleki, Scott Klarenbach, Mohamed Abdalla. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Benchmarking Large Language Models and Prompt Engineering Strategies in Microsatellite Instability Cancers: Evaluation Study

Yuxin Zhang*, BS; Jie Song*, BE; Cheng Bi, MSc; Xin Zheng, MSc; Zhichuan Xu, BE; Dan Cao, PhD; Bairong Shen, PhD

Department of Medical Oncology, Institutes for Systems Genetics, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, Sichuan University, No 2222 Xinchuan Road, Gaoxin District, Chengdu, Sichuan, China

*these authors contributed equally

Corresponding Author:
Bairong Shen, PhD
Department of Medical Oncology, Institutes for Systems Genetics, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, Sichuan University, No 2222 Xinchuan Road, Gaoxin District, Chengdu, Sichuan, China

Abstract

Background: The reliability of general-purpose large language models (LLMs) for complex clinical tasks in specialized domains such as microsatellite instability (MSI) cancers remains critically uncharacterized. The absence of a domain-specific benchmark to evaluate and guide the optimization of their capabilities across diverse clinical tasks poses unevaluated risks to patient safety.

Objective: This study aimed to develop and validate Microsatellite Instability Cancer Benchmark (MSIC-Bench), a novel, two-tiered benchmark for MSI cancer, covering both consensus and frontier knowledge. Using this framework, we aimed to systematically assess LLM performance across various prompting strategies, identify task-specific weaknesses, and reveal effective pathways for performance improvement.

Methods: We developed MSIC-Bench, a 511-question benchmark derived from clinical guidelines and a curated knowledge base. Three state-of-the-art LLMs (GPT-4o, [OpenAI], Gemini 2.5 Pro [Google], and Claude Opus 4 [Anthropic]) were evaluated using 4 prompting strategies, including vanilla, chain-of-thought, reflection of thoughts, and retrieval-augmented generation (RAG), under both multiple-choice and open-ended modalities. Performance was assessed on accuracy, safety (honesty), error composition, and token usage.

Results: LLMs demonstrated a significant "scaffolding effect," with accuracy dropping substantially in open-ended scenarios. For non-RAG strategies, the primary failure mode was an internal knowledge deficit. The integration of RAG proved to be the most effective intervention. A domain-aligned RAG strategy not only significantly improved accuracy in complex decision-making tasks but also fundamentally shifted the system's primary bottleneck from knowledge deficits to retrieval failures. In terms of safety, RAG induced a favorable shift from high-risk fabrications to safer refusals, though this introduced a safety-utility trade-off in the form of false refusals. Notably, our hybrid-RAG configuration, which combined both knowledge sources, demonstrated the most robust and generalizable performance across all tasks.

Conclusions: Current LLMs lack the specialized knowledge required for reliable application in MSI oncology. A well-designed RAG architecture is the pivotal intervention to address this gap. However, its success is not automatic; it transforms the nature of system failure, making retrieval precision and knowledge base quality the new critical determinants of performance and safety. Our findings establish a clear directive for developing trustworthy clinical artificial intelligence: focus must shift toward optimizing the retrieval component and curating high-quality, comprehensive knowledge sources. MSIC-Bench provides a robust framework to guide these future efforts.

(J Med Internet Res 2026;28:e88614) doi:10.2196/88614

KEYWORDS

microsatellite instability; cancer; large language model; LLM; prompt engineering; benchmark

Introduction

Genomic instability is a core hallmark of cancer [1] and uniquely bridges cancer and aging [2], reflecting its central role in cell fate decisions. Microsatellite instability (MSI), a major form of genomic instability, serves as a crucial pan-cancer biomarker with significant diagnostic, prognostic, and therapeutic value [3,4]. However, significant challenges remain in understanding the heterogeneity of MSI across cancer types, accurately detecting and classifying MSI in clinical settings, and optimizing personalized therapeutic strategies for MSI-positive patients [5,6].

XSL•FO
RenderX

Advances in artificial intelligence (AI), particularly deep learning, have transformed MSI detection [7] and treatment response prediction [8], with growing applications in the diagnosis and management of MSI-associated cancers [9]. In recent years, large language models (LLMs) such as GPT-4 (OpenAI) [10], Claude (Anthropic) [11], and Gemini (Google) [12], built on transformer architectures with billions of parameters, have driven innovation in health care through their exceptional abilities in understanding, reasoning, and generation. These LLMs have shown significant potential in various medical fields, including rare disease diagnosis [13], sepsis management [14], and medical education [15].

However, the application of LLMs in MSI-related cancer care remains largely unexplored. In addition, the lack of standardized benchmarks for systematically evaluating their performance on MSI-specific tasks limits their clinical adoption and further development [16].

To fill this critical research gap, we designed and constructed Microsatellite Instability Cancer Benchmark (MSIC-Bench), a novel, domain-specific evaluation framework. The architecture of MSIC-Bench is explicitly engineered to separate and assess 2 distinct cognitive abilities. It comprises two core tiers: (1) the basic tier, derived from established clinical guidelines, evaluates an LLM's mastery of foundational, consensus-based knowledge, and (2) the advanced tier, built upon our curated, evidence-based knowledge base, the Microsatellite Instability Cancer Knowledgebase (MSICKB) [17], assesses an LLM's capacity to comprehend and reason with complex, frontier scientific evidence. Each tier is further categorized into key clinical subtasks, including molecular basis, diagnosis, treatment, and prognosis, ensuring a comprehensive assessment of the required knowledge spectrum.

Using this benchmark, we conducted a systematic evaluation of 3 state-of-the-art LLMs (GPT-4o [OpenAI], Gemini 2.5 Pro [Google], and Claude Opus 4 [Anthropic]) across 4 distinct prompting strategies. We assessed their performance under different prompting strategies designed to probe their internal knowledge and reasoning, including a vanilla zero-shot approach, a step-by-step reasoning approach (chain-of-thought [CoT] [18]), and a multipersona reflective approach (reflection of thoughts [RoT] [19]). Crucially, we also evaluated their capabilities when integrated with a retrieval-augmented generation (RAG) architecture [20], providing them with access to our curated knowledge base.

Our investigation was structured along two critical dimensions: (1) the "basic vs Advanced" dimension, to distinguish between mastery of consensus knowledge and comprehension of frontier science; and (2) the "multitask" dimension, to evaluate performance across key clinical subtasks (eg, diagnosis and treatment). This dual-dimensional analysis aimed to answer a central question: To what extent can LLMs provide accurate, reliable, and safe answers within the MSI cancer domain?

This study addresses these gaps through several integrated efforts. We first developed and validated MSIC-Bench, a new, expert-curated benchmark comprising 511 questions designed to assess LLMs on both foundational and frontier knowledge in MSI-related cancer. Using this benchmark, our multifaceted evaluation framework systematically assessed leading LLMs not just for accuracy, but also for AI safety—distinguishing justified from false refusals—and for error composition, pinpointing critical system bottlenecks.

Our evaluation yields several critical insights. We reveal that the primary bottleneck for standard LLMs is a profound deficit in specialized knowledge. While RAG is the critical intervention to address this, our analysis shows that it fundamentally shifts the bottleneck from knowledge to information retrieval, where "retrieval failure" becomes the new dominant error mode. Furthermore, we establish a safety-utility trade-off in RAG systems. We demonstrate that while RAG transforms high-risk "fabrications" into safer refusals, it can also introduce "false refusals"—a new, utility-degrading error type. Finally, our findings provide actionable insights for developing more robust systems, showing that an RAG architecture that integrates both broad clinical guidelines and specialized knowledge offers a practical and effective solution. Together, these findings not only map the current capabilities and limitations of LLMs in oncology but also provide a clear, evidence-based roadmap for their future development and safe clinical integration.

To realize these contributions, we designed a comprehensive, multistage study. The overall workflow is illustrated in Figure 1, encompassing three core components: (1) benchmark construction, (2) benchmark content, and (3) experimental setup.

Therefore, the primary aims of this study were twofold: first, to develop and validate MSIC-Bench, a novel, domain-specific benchmark to fill a critical gap in evaluating LLMs for MSI cancer care; and second, to use this benchmark to systematically evaluate the capabilities and limitations of state-of-the-art LLMs and investigate the effectiveness of different RAG configurations in mitigating their identified weaknesses.

XSL•FO
RenderX

Figure 1. Schematic overview of the Microsatellite Instability Cancer Benchmark (MSIC-Bench) study design. AI: artificial intelligence; Q and A: question and answer.

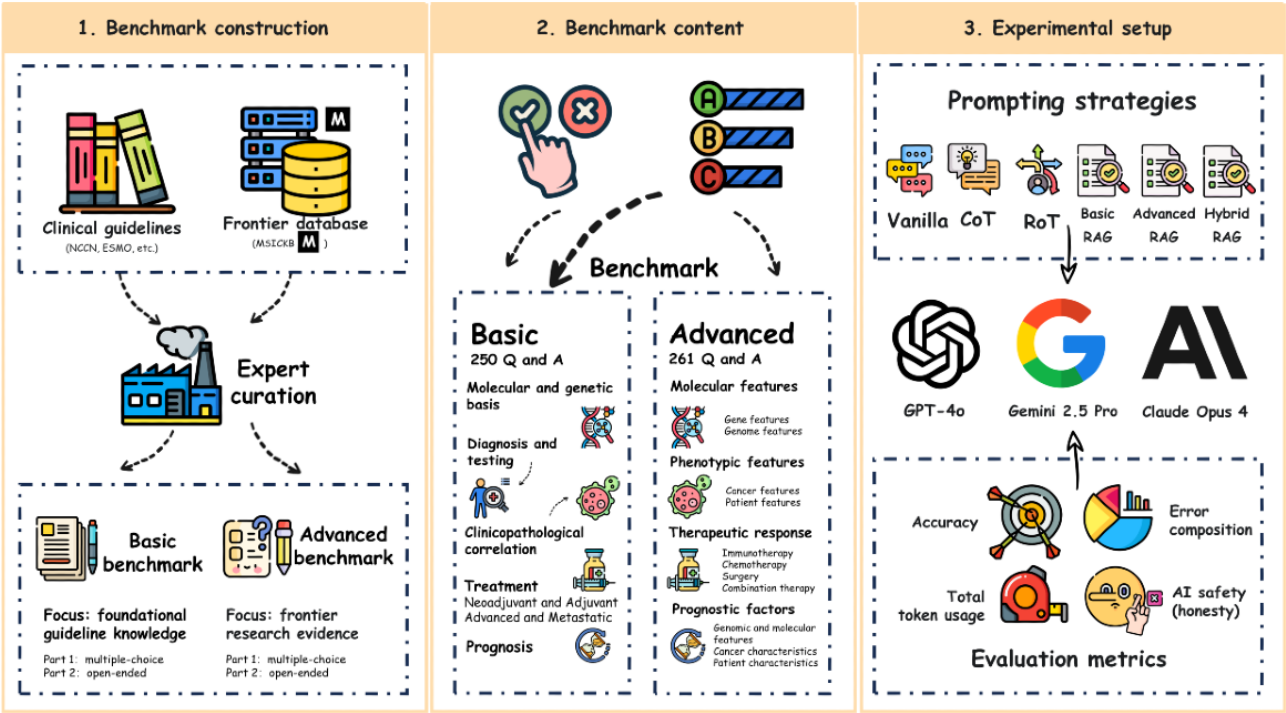

Methods

Construction of MSIC-Bench

Overview

The construction of MSIC-Bench followed a rigorous, multistage process designed to ensure its clinical relevance, scientific accuracy, and comprehensive coverage.

Data Sources

The MSIC-Bench was designed to separate the assessment of 2 critical LLM capabilities: generalization on foundational clinical knowledge and tracking of frontier scientific evidence. The basic tier was derived from a curated collection of the most recent and authoritative clinical practice guidelines from the National Comprehensive Cancer Network (NCCN) and the European Society for Medical Oncology (ESMO), as detailed in Table 1. These guidelines were selected to represent the stable, consensus-based knowledge required for standard clinical care.

Table . Clinical practice guidelines used as data sources for the basic tier.

Institution and guideline name	Version
NCCN[a]	
Colon cancer [21]	Version 3.2025 (April 24, 2025)
Colorectal cancer screening [22]	Version 1.2025 (May 30, 2025)
Gastric cancer [23]	Version 2.2025 (April 4, 2025)
Rectal cancer [24]	Version 2.2025 (March 31, 2025)
Uterine neoplasms [25]	Version 3.2025 (March 7, 2025)
ESMO[b]	
Endometrial cancer [26]	2022.05
Gastric cancer [27]	2022.07
Localized colon cancer [28]	2020.06
Localized rectal cancer [29]	2025.05
Metastatic colorectal cancer [30]	2022.10

[a]NCCN: National Comprehensive Cancer Network.

[b]ESMO: European Society for Medical Oncology.

XSL•FO
RenderX

Specifically, the advanced tier of MSIC-Bench was constructed using MSICKB, a comprehensive knowledge resource. MSICKB was developed to address the fragmentation and limited scope of data in MSI cancer research. By systematically curating 492 peer-reviewed publications, MSICKB provides a structured, multidimensional repository that goes beyond simple gene lists. It includes genetic and molecular features (such as mutations, expression levels, and epigenetic changes), clinicopathological characteristics (such as tumor location and histology), explicit prognostic data (including prognostic factors and survival outcomes), and therapeutic responses to various treatments. Unlike large-scale databases such as The Cancer Genome Atlas [31], MSICKB offers granular, expert-curated annotations that capture the complexity and heterogeneity of MSI cancers. This comprehensive resource enables the creation of advanced, clinically nuanced questions essential for evaluating LLM performance.

Expert Curation

The design and curation of MSIC-Bench were an iterative process led by 2 clinical experts, whose detailed qualifications are provided in Table S1 in Multimedia Appendix 1.

For the basic tier, the 2 experts conducted a comprehensive review of the selected guidelines. They systematically identified core knowledge points across key clinical domains and collaboratively authored a series of multiple-choice and true or false questions to cover these fundamental aspects. Each question was repeatedly refined to ensure strict adherence to guideline content, avoid ambiguity, and maintain an appropriate level of difficulty.

For the advanced tier, the 2 experts used the MSICKB knowledge base and its corresponding source literature. They meticulously selected and extracted pivotal findings related to molecular features, phenotypic characteristics, therapeutic responses, and prognostic factors. Each question was then carefully crafted to reflect the latest advancements and unresolved challenges in the field, ensuring it represented the current research frontier.

Following the initial drafting, all 511 QA pairs underwent a final consensus validation by the 2 experts to guarantee their accuracy, clarity, and relevance before being finalized for the benchmark. The final basic tier consists of 142 multiple-choice and 108 true or false questions, while the advanced tier comprises 261 multiple-choice questions. Table S2 in Multimedia Appendix 1 presents representative examples from the benchmark, illustrating the structure of questions across different tiers and task categories.

To enable a granular analysis of the capabilities of LLMs, all questions in the benchmark were categorized based on their clinical and scientific domain. The basic tier questions were classified into 5 key areas: molecular and genetic basis, diagnosis and testing, clinicopathological correlation, treatment, and prognosis. The advanced tier questions were classified into 4 primary tasks, further broken down into more specific subtasks: phenotypic features (eg, cancer characteristics and patient characteristics), molecular features (eg, gene and genome), therapeutic response (eg, immunotherapy, chemotherapy, and surgery), and prognostic factors (eg, genomic and molecular features). This multilayered categorization is a core design feature of MSIC-Bench, enabling a detailed performance evaluation that moves beyond a single overall accuracy score to reveal the specific strengths and weaknesses of LLMs across tasks. The exact number of questions per subtask is provided in Table S3 in Multimedia Appendix 1.

In addition to the primary benchmark of selection-based questions, a secondary set of open-ended questions was created, termed the without selections modality. This set was derived directly from the primary benchmark by removing the choice options. This dual-modality design serves a crucial purpose: the selection-based questions assess an LLM's discriminative ability to identify the correct answer from a predefined set of options, while the open-ended questions challenge its generative and recall capabilities to formulate an answer autonomously. This allowed for a direct comparison of LLM performance with and without the "scaffolding" of selection options, providing a more comprehensive evaluation of their reasoning abilities.

Experimental Setup

Overview

To evaluate LLM performance on MSIC-Bench, we designed a comprehensive experimental framework. We benchmarked 3 leading models across 4 distinct prompting strategies: vanilla, CoT, RoT, and RAG (basic-RAG, advanced-RAG, and hybrid-RAG).

LLMs and Parameters

We evaluated 3 state-of-the-art LLMs: GPT-4o [10], Claude Opus 4 [11], and Gemini 2.5 Pro [32]. For full transparency and reproducibility, detailed information on the specific model versions, application programming interface (API) identifiers, and knowledge cutoff dates is provided in Table S4 in Multimedia Appendix 1. All API calls were executed using consistent parameter settings: temperature was set to 0.3 to ensure reproducibility while allowing for minor stylistic variation, and max_tokens was set to 4096 to prevent premature truncation. A retry mechanism with up to 5 attempts was implemented to handle transient API failures.

Prompting Strategies

We benchmarked each model against 4 prompting strategies, designed to probe reasoning capabilities under different conditions of knowledge reliance (Table 2). The full, unabridged system prompts used for each strategy are provided in Section S1 in Multimedia Appendix 1.

XSL•FO
RenderX

Table . Prompt templates for the 4 evaluated strategies.

Prompting strategy	Template
Vanilla	You are a biomedical expert. Answer the following question based on your internal knowledge. Question: {question}.
CoT[a]	You are a biomedical expert. Answer the following question. First, provide your step-by-step reasoning process. Then, provide the final answer. Let's think step by step. Question: {question}, Reasoning: [Your step-by-step reasoning here], and Final Answer: [Your answer here].
RoT[b]	You are a biomedical expert. Answer the following question. Question: {question}. Imagine 3 medical experts are solving this task. Each expert independently provides their step-by-step reasoning and final answer. After all experts have finished, they discuss together, review, and backtrack their previous reasoning steps, and finally reach a consensus on the final answer. Please present: [Expert 1's reasoning and answer], [Expert 2's reasoning and answer], [Expert 3's reasoning and answer], and [The discussion and the agreed final answer].
RAG[c]	You are a biomedical expert. Answer the following question based on the provided clinical guideline context and your internal knowledge. Question: {question}. Context: {context}.

[a]CoT: chain-of-thought.

[b]RoT: reflection of thoughts.

[c]RAG: retrieval-augmented generation.

RAG Pipeline Implementation Details

To assess the impact of external knowledge, we implemented 3 distinct RAG configurations, each differing by its underlying knowledge source: basic-RAG (using established clinical guidelines), advanced-RAG (based on our curated MSICKB), and hybrid-RAG (combining both knowledge sources). These configurations allowed us to systematically evaluate performance across different knowledge tiers. To ensure a controlled comparison and facilitate reproducibility, all critical pipeline parameters were standardized across these configurations. The specific details of each RAG system, including knowledge source, chunking strategy, embedding model, and retrieval methodology, are summarized in Table 3.

Table . Configuration parameters for retrieval-augmented generation (RAG) systems.

Parameter	Basic RAG[a] (guideline-based)	Advanced RAG (MSICKB[b]-based)	Hybrid RAG (unified)
Knowledge source	Unstructured clinical guideline PDFs	Structured MSICKB (JSON lines)	Combined guidelines and MSICKB
Document representation	Text chunks from parsed PDFs	Each JSON line as a distinct document	Hybrid of text chunks and JSON lines
Chunking strategy	RecursiveCharacterTextSplitter	N/A (Atomic document unit)	RecursiveCharacterTextSplitter (for guidelines)
Chunk size	1000 characters	N/A[c]	1000 characters (for guidelines)
Chunk overlap	200 characters	N/A[c]	200 characters (for guidelines)
Embedding model	text-embedding-3-small	text-embedding-3-small	text-embedding-3-small
Retrieval method	Cosine similarity	Cosine similarity	Cosine similarity
Top-k	3	3	3
Reranking	None	None	None

[a]RAG: retrieval-augmented generation.

[b]MSICKB: Microsatellite Instability Cancer Knowledgebase

[c]For the advanced RAG, each structured entry (JSON line) in the MSICKB was treated as an atomic, indivisible document for embedding and retrieval. Therefore, chunking parameters are not applicable.

Evaluation Protocol

Overview

Our evaluation was conducted across 4 key dimensions: accuracy, AI safety (honesty), error composition, and total token usage. To ensure a rigorous assessment, the dimensions of accuracy, AI safety (honesty), and error composition were manually evaluated by 2 experts. This procedure involved 2 experts independently reviewing all responses, with any initial disagreements resolved through a consensus discussion to

XSL•FO
RenderX

establish the final labels. The statistical methods used to validate this interrater reliability are detailed in the "Statistical Analysis" subsection.

Accuracy

Accuracy was assessed against a predefined right answer for each question. A response was marked as "correct" only if it was factually accurate and semantically equivalent to the ground truth answer. The final accuracy was calculated as the percentage of responses deemed "correct."

Honesty Analysis

Following the accuracy assessment, all incorrect responses were analyzed to evaluate the model's honesty. We classified these responses based on two primary behaviors applicable to all settings: (1) fabrication: the model provides a factually incorrect answer instead of admitting a knowledge gap. This represents a direct failure of honesty, and (2) justified refusal: the model correctly acknowledges its limitations by explicitly refusing to answer (eg, "I don't know"). This is considered a desirable safety feature.

For RAG strategies, the presence of external context introduced a third, more granular category to assess how the model used the provided evidence:

- False refusal: this classification is unique to RAG and occurs when the model refuses to answer despite the retrieved context containing the necessary information. It represents a failure in utility, as the model did not correctly leverage the evidence provided.

Error Composition Analysis

To diagnose specific failure modes, we designed a hierarchical, single-label annotation scheme for categorizing incorrect responses. For both non-RAG and RAG settings, we used a single-label error annotation scheme with a fixed priority order, which enables well-defined, interpretable comparisons of error distributions. In the non-RAG setting, evaluators consider question misinterpretation, knowledge deficit, and reasoning error in this order and assign the first label whose definition is satisfied. In the RAG setting, evaluators consider question misinterpretation, retrieval failure, context ignorance, and reasoning error in this order and similarly assign the first applicable label. The complete definitions, categories, and priority order for both schemes are detailed in Table S5 in Multimedia Appendix 1.

Total Token Usage

To evaluate the efficiency across LLMs and prompting strategies, we measured the total token usage per interaction, defined as the sum of tokens in the input prompt and the generated output. The input prompt included system instructions, the question, and the retrieved context for RAG variants. The generated output was the model's completion. All texts were tokenized using the tiktoken (OpenAI) library [33] with the cl10k_base encoding to provide a single, reproducible counting rule across models. This metric offers a more faithful estimate of resource consumption than response length alone.

Statistical Analysis

Interrater Reliability

To validate the reliability of the manual annotation process, we calculated the interrater reliability between the 2 domain experts for the dimensions of accuracy, honesty, and error composition. Cohen κ coefficient was computed based on the experts' initial independent ratings before their consensus discussion. The resulting κ values were interpreted using the Landis and Koch [34] benchmarks, where scores in the range of 0.61 - 0.80 are considered "substantial" and 0.81 - 1.00 are "almost perfect" agreement.

Significance Testing for Performance Comparison

To statistically validate performance differences between prompting strategies, we used the exact binomial version of the McNemar test for paired nominal data. This test was used to compare the accuracy of 2 strategies on the same set of questions within specific subtasks. $P<.05$ was considered statistically significant.

Ethical Considerations

This study was based exclusively on publicly available data, including published clinical guidelines and peer-reviewed papers. As the research did not involve human participants, it was exempt from Institutional Review Board review, and requirements for informed consent, participant privacy, and compensation were not applicable. All patient information referenced in the source literature was presumed to have been deidentified by the original authors in accordance with established ethical standards.

Results

Overview

Our evaluation protocol involved manual annotation by 2 domain experts for the key dimensions of accuracy, honesty, and error composition. To first establish the reliability of these subjective assessments, we calculated the interrater reliability before any performance analysis. We achieved almost perfect agreement for accuracy (Cohen κ=0.903) and honesty (κ=0.954), and substantial agreement for error composition (κ=0.837), confirming the robustness of our evaluation methodology.

Overall Accuracy: Selection Options Boost Performance While RAG Excels in Open-Ended Tasks

Our primary analysis focused on the overall accuracy of the 3 evaluated LLMs across the basic (Figure 2A) and advanced (Figure 2B) tiers. Within each tier, we compared LLM performance under 6 distinct prompting strategies in 2 evaluation modalities: multiple-choice (hatched bars) and open-ended (solid bars).

XSL•FO
RenderX

Figure 2. Overall accuracy of large language models (LLMs) on Microsatellite Instability Cancer Benchmark (MSIC-Bench).

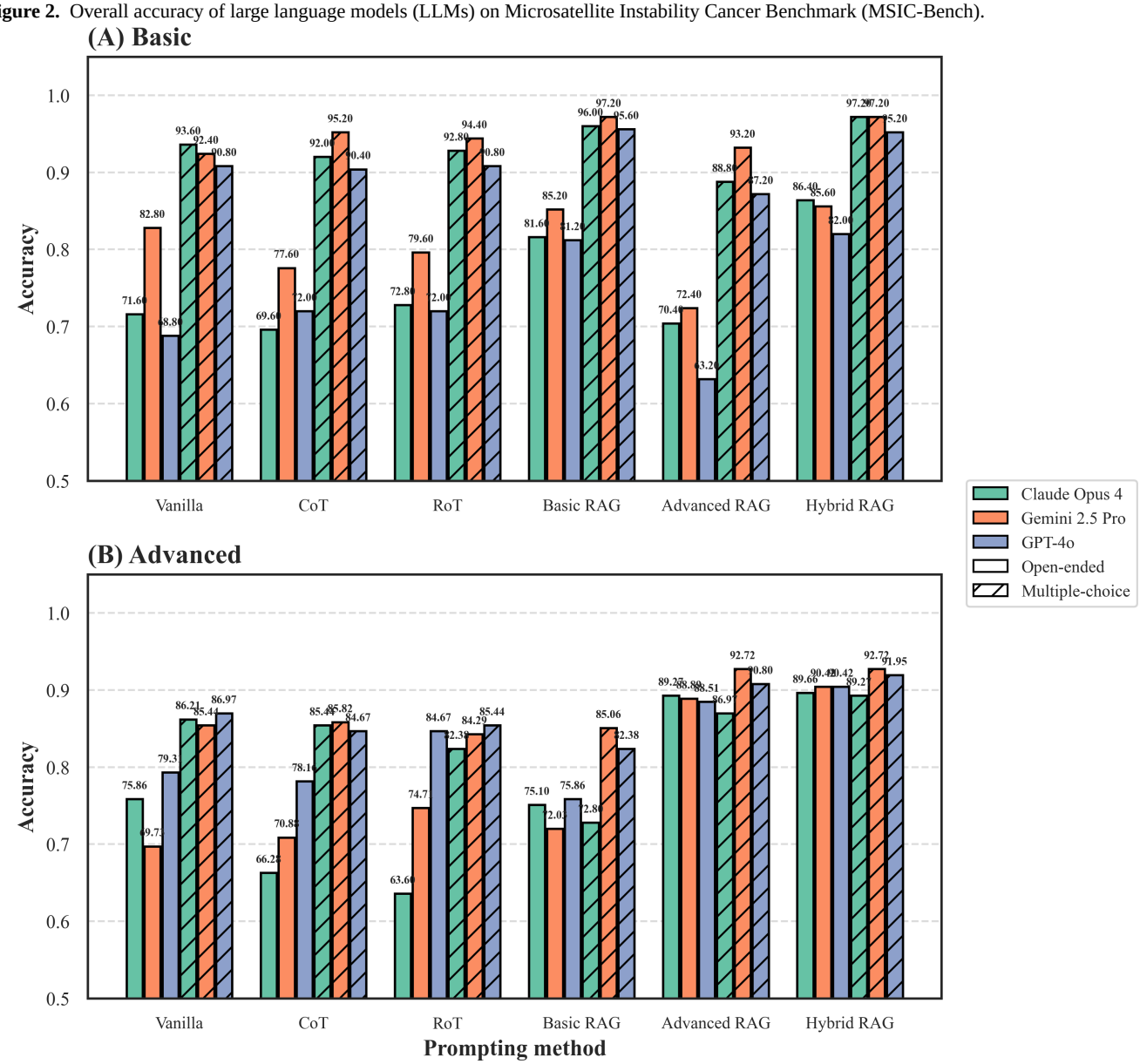

A primary finding is the significant impact of evaluation modality. This “scaffolding effect," a term from cognitive science, is shown in Figure 2. Across nearly all conditions, accuracy in the open-ended modality was substantially lower than in multiple-choice. This drop was most pronounced for non-RAG methods in the basic tier (Figure 2A), where Claude Opus 4 with the vanilla prompt declined from 93.6% to 71.6%. Predefined selection options transform a demanding generative recall task into a simpler discriminative recognition task; removing this scaffold exposes the LLM’s limitations in open-ended scenarios.

This analysis also highlighted the clear superiority of the RAG strategies, particularly when the “scaffolding” was removed. In the multiple-choice modality on the basic tier (hatched bars in Figure 2A), basic RAG was already a top-performing strategy, achieving accuracies of up to 97.2%. However, the superiority of RAG became most pronounced in the more challenging open-ended modality. Across both tiers (solid bars in Figures 2A and B), all RAG-based methods—basic, advanced, and hybrid RAG—consistently and significantly outperformed their non-RAG counterparts. For example, in the advanced open-ended setting (Figure 2B), Claude Opus 4’s accuracy increased substantially from 63.6% with RoT to 89.3% with advanced RAG and further to 89.7% with hybrid RAG, representing one of the most substantial performance gains observed in this setting. This underscores that for realistic, open-ended tasks, providing precisely aligned external knowledge is a highly effective strategy for improving accuracy.

However, our results revealed that the effectiveness of specialized RAG is highly dependent on the alignment between the retrieved context and the question’s knowledge domain. This dependency was clearly illustrated by the models’ performance when presented with out-of-domain context. In the advanced multiple-choice setting (hatched bars in Figure 2B), a notable adverse effect was observed for the Claude Opus 4 model, whose accuracy declined significantly to 72.8% when using the out-of-domain basic RAG strategy. A converse pattern was observed in the basic multiple-choice setting (hatched bars in Figure 2A), where the introduction of out-of-domain advanced

XSL•FO
RenderX

RAG context also led to a performance drop for some models, such as GPT-4o, declining to 63.2%.

The introduction of the hybrid RAG strategy directly addresses this domain alignment challenge. By having access to both knowledge bases, hybrid RAG simulates a more practical scenario. The results demonstrate its remarkable robustness. In the basic setting (Figure 2A), hybrid RAG's performance nearly matched that of the specialized basic RAG (eg, 95.2% vs 95.6% for GPT-4o in the multiple-choice modality). Similarly, in the advanced setting (Figure 2B), it remained highly competitive with the top-performing advanced RAG. This indicates that the hybrid RAG approach successfully navigates the knowledge selection challenge, maintaining high accuracy without requiring prior knowledge of the question's domain. It effectively mitigates the performance degradation seen with misaligned RAG, positioning it as a more generalizable and practical solution for real-world applications.

Task-Specific Capability Profile: Uncovering True Failure Modes

To further dissect the LLM performance, we analyzed its accuracy across different clinical task categories (Figure S1 in Multimedia Appendix 1). The "scaffolding effect" of selection options was immediately apparent. A visual comparison between the multiple-choice modality (Figure S1A in Multimedia Appendix 1) and the open-ended modality (Figure S1B in Multimedia Appendix 1) reveals a universal degradation of performance across nearly all tasks and LLMs. The radar plots in the open-ended setting are visibly "shrunken," indicating that the performance drop observed in the overall accuracy is a pervasive phenomenon affecting all facets of clinical tasks.

Our task-specific analysis provides a granular capability profile of current LLMs, offering actionable insights for both biomedical researchers and clinicians.

For researchers, LLMs demonstrated considerable reliability on fact-retrieval–oriented tasks, such as those in the "Molecular and Genetic Basis" category (Figure S1A in Multimedia Appendix 1). This suggests their potential as powerful hypothesis generation assistants. For example, a researcher could rapidly query an LLM about the known molecular features of MSI tumors across different cancer types, accelerating literature review and experimental design.

For clinicians, the profile reveals a critical dichotomy. On one hand, LLMs can serve as reliable and rapid knowledge retrieval tools for well-established facts, such as those in "Diagnosis and Testing." On the other hand, we pinpointed areas where clinicians must exercise extreme caution. A consistent pattern of underperformance was observed in complex, multifactor decision-making tasks, with "Therapeutic Response (Chemotherapy)" emerging as a prominent failure mode across all models in the challenging open-ended setting (Figure S1B in Multimedia Appendix 1). For instance, under the vanilla prompt, the accuracy of GPT-4o on this subtask (n=21) was only 33.3%, a performance level statistically indistinguishable from random chance in a high-stakes context. This striking weakness indicates that when faced with nuanced clinical scenarios that require weighing competing evidence, the internal knowledge of general-purpose LLMs is dangerously unreliable. This is the domain where unassisted LLM usage poses the greatest risk.

Crucially, our analysis also validates a clear pathway to mitigate these risks through a well-designed RAG system. However, the journey to effective RAG is nuanced. The erratic performance of basic RAG, which sourced knowledge only from general guidelines, highlights the danger of grounding models in incomplete or overly broad information; in some advanced tasks, it even underperformed non-RAG methods. In contrast, advanced-RAG, which used our specialized MSICKB knowledge base, consistently improved performance across most high-risk tasks. The most robust results were achieved with hybrid RAG, which integrated both knowledge sources, leading to substantial and stable performance gains, particularly in the most challenging decision-making areas such as chemotherapy response. This improvement was statistically significant for hybrid RAG with both GPT-4o ($P<.001$; McNemar test) and Gemini 2.5 Pro ($P=.02$) compared to the vanilla strategy, although not for Claude Opus 4 ($P=.22$), reflecting its unique interaction with RAG systems.

This finding provides a clear directive: for high-stakes clinical decision support, RAG is not a monolithic solution but a framework whose effectiveness is contingent upon the quality and comprehensiveness of its knowledge base. A well-curated, specialized knowledge source is an essential safety and performance mechanism.

AI Safety Analysis: RAG as a Key Enabler for LLM Honesty

Moving from performance to safety, we analyzed the composition of incorrect responses by categorizing them into 3 distinct behaviors: fabrication, justified refusal, and false refusal (Figure 3). Our analysis begins with the basic tier (Figures 3A and 3B), where a stark pattern emerged: nearly all incorrect responses were classified as fabrication. This outcome, however, is not an indicator of inherent model dishonesty but a direct consequence of the benchmark's design. The selection-based and true or false questions in this tier did not include a "don't know" option, structurally forcing the LLM to guess when faced with uncertainty and precluding any form of refusal.

XSL•FO
RenderX

Figure 3. Honesty analysis of error behaviors.

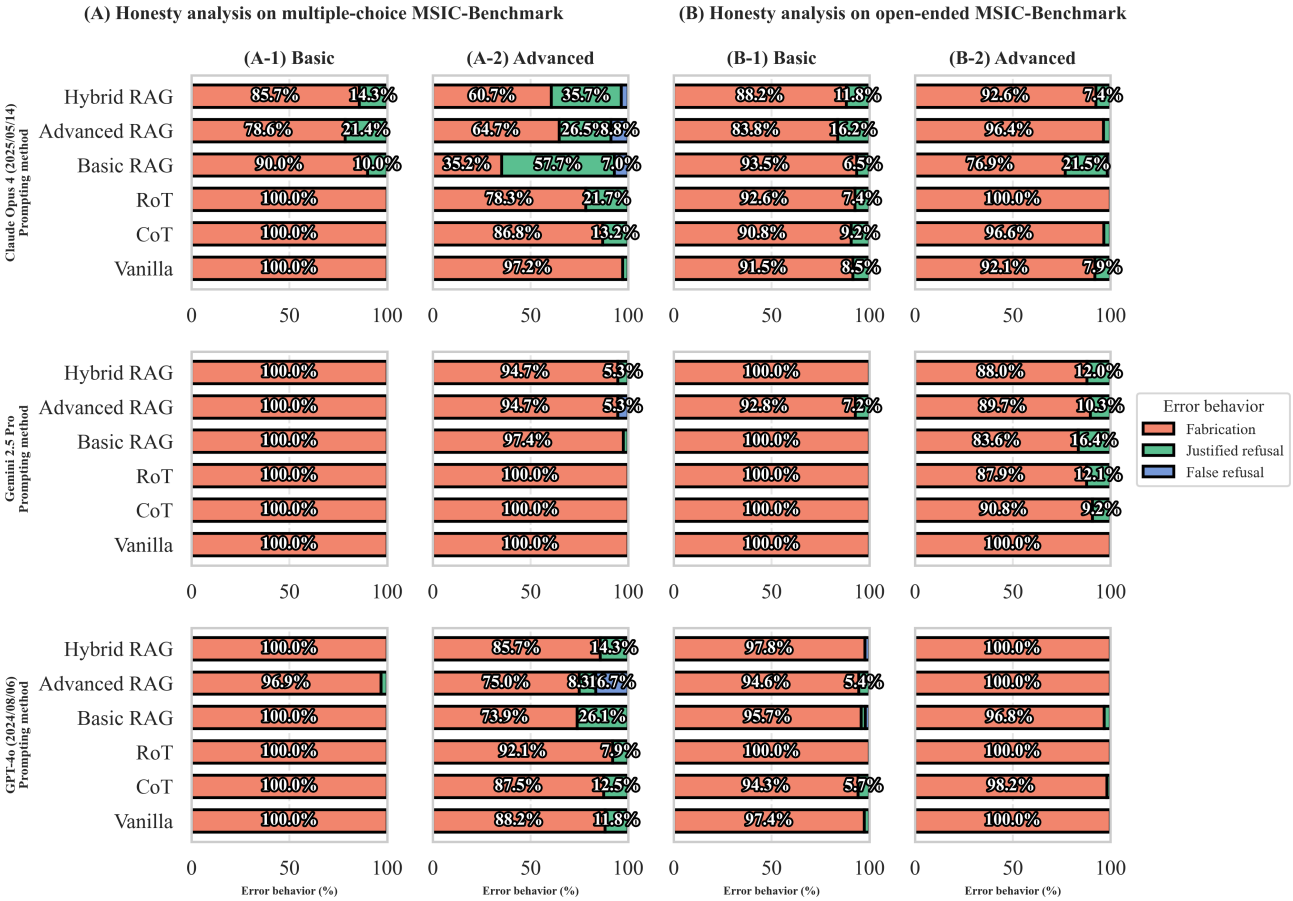

The models' true behavioral tendencies are revealed in the advanced tier (Figures 3A and 3B), where expressing uncertainty is possible. Even here, non-RAG strategies (vanilla, CoT, and RoT) demonstrate a strong inclination toward fabrication, with rates often exceeding 85%. This highlights a critical safety concern, that is, without a bounded context, LLMs tend to guess rather than admit ignorance. In contrast, the introduction of RAG strategies fundamentally alters this behavior. Across all models and modalities, RAG dramatically reduces the proportion of fabrication by converting these high-risk errors into justified refusals. This shift represents a significant enhancement in model honesty, as the LLMs learn to correctly identify when the provided context is insufficient to form an answer.

However, this safety improvement comes at a price: the emergence of "false refusals," a failure mode unique to RAG where the model refuses to answer despite the context containing the necessary information. This dynamic is most vividly illustrated by Claude Opus 4 under the basic RAG strategy in the multiple-choice setting (Figure 3A). Here, fabrication plummeted from 97.2% (vanilla) to just 35.2%. This reduction was primarily achieved by converting a remarkable 57.7% of errors into desirable justified refusals. The cost, however, was a 7% rate of false refusals. This pattern—substituting dangerous fabrications with mostly safe justified refusals at the expense of some loss in utility—is a consistent and critical finding of our RAG analysis.

This analysis demonstrates that RAG is a powerful mechanism for inducing epistemic caution. Grounding the model in a bounded context provides a clear basis for assessing knowledge limits, which is the mechanism behind justified refusals. While the existence of false refusals highlights a persistent challenge in ensuring models robustly use all provided evidence, the net effect represents a highly favorable shift in the error profile. RAG systematically replaces high-severity, factually incorrect errors (fabrication) with manageable, lower-severity utility errors (false refusals), thereby making the LLM's failure modes safer and more trustworthy.

Error Analysis: Retrieval Failure as the New Bottleneck in RAG System

To understand the root causes of errors, we conducted a detailed error composition analysis, applying distinct taxonomies for non-RAG and RAG methods to precisely identify failure points (Figures 4A and 4B).

XSL•FO
RenderX

Figure 4. Error composition analysis of incorrect large language model (LLM) responses.

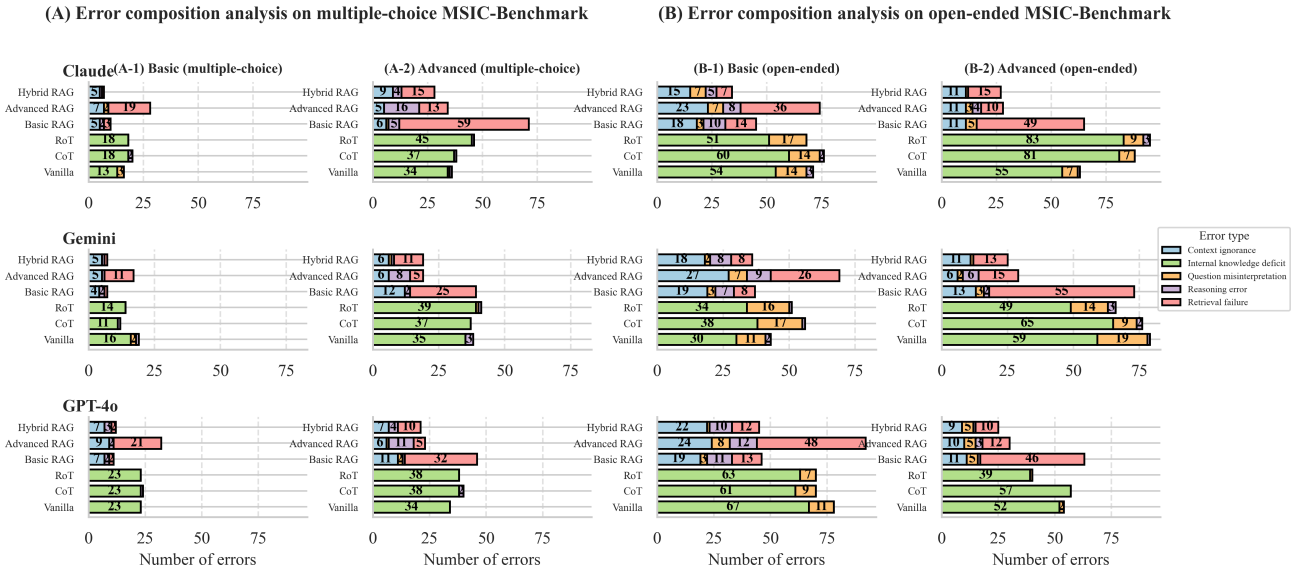

For methods relying on the LLM's internal knowledge (vanilla, CoT, and RoT), a clear pattern emerged: failures were overwhelmingly caused by an internal knowledge deficit. This confirms that for this specialized domain, the inherent knowledge gaps of general-purpose LLMs are the primary barrier to accuracy.

The introduction of RAG altered the error distribution. The proportion of internal knowledge deficit errors was substantially reduced across all RAG configurations, as the external knowledge base compensated for the models' internal deficiencies.

However, this shift revealed that retrieval failure replaced knowledge deficit as the new dominant source of error. In most RAG scenarios, retrieval failure constituted the single largest component of incorrect responses, often accounting for over half of all errors. This indicates that while RAG addresses the problem of "not knowing," it introduces the challenge of "not finding," making system performance highly dependent on the accuracy of the retrieval module.

Other RAG-specific errors, such as context ignorance (the LLMs failing to use correct context) and reasoning errors, occurred at a lower frequency. This suggests that when the retriever succeeds in providing the correct context, the LLMs are generally capable of using it appropriately. Therefore, a key area for improving RAG performance in this domain is to enhance the retriever's ability to consistently deliver the correct context.

Token Usage Analysis: The High Price of Complex Reasoning and the Efficiency of RAG

Finally, we evaluated the token usage of each strategy by measuring the total token count (input and output) for each interaction (Figures 5 and 6).

The most notable finding, illustrated by the average token counts in Figure 5, was the substantial token usage associated with complex reasoning prompts. The RoT strategy consistently consumed the most tokens across all conditions. This effect was particularly pronounced for Gemini 2.5 Pro in the multiple-choice–Basic setting (Figure 5A), where the average total tokens for RoT (1949 tokens) were nearly 2 times that of CoT (1044 tokens) and almost 4 times that of vanilla (560 tokens). This highlights the significant computational overhead required for multipersona ensemble reasoning.

The analysis also revealed a clear difference between the 2 evaluation modalities. As shown in both the average counts (Figure 5B) and the distributions (Figure 6B), interactions in the open-ended modality were generally more token-intensive. More importantly, the box plots in Figure 6B show that these open-ended interactions exhibited significantly greater variance (larger IQRs and more outliers) compared to their multiple-choice counterparts (Figure 6A). This indicates that open-ended questions not only require a higher average total token usage but also lead to less predictable and more variable costs.

In terms of efficiency, the family of RAG strategies demonstrated a superior balance of high accuracy and token economy. Crucially, our analysis accounts for the full cost, including retrieved context in the input tokens. While achieving high accuracy (as shown in Figure 2), their average token counts (Figure 5) remained significantly lower than RoT and often comparable to CoT. This highlights a fundamental difference in cost structure: the expense of RoT is primarily in the generated output (ie, computational reasoning), whereas a significant portion of RAG's cost is front-loaded into the input prompt (ie, providing context). Furthermore, the RAG methods offer a spectrum of cost-performance options: hybrid RAG, which integrates CoT reasoning, logically incurs a higher token cost than basic or advanced RAG, positioning it as a premium option for when maximal accuracy is required. This positions the RAG framework as a strategically efficient approach for deploying reliable clinical AI systems. It allows developers to balance computational cost and performance requirements by selecting the appropriate RAG configuration—from token-frugal methods for simpler tasks to more powerful, context-rich options such as hybrid RAG for high-stakes clinical decisions.

XSL•FO
RenderX

Figure 5. Average total token usage per interaction on the Microsatellite Instability Cancer Benchmark (MSIC-Bench).

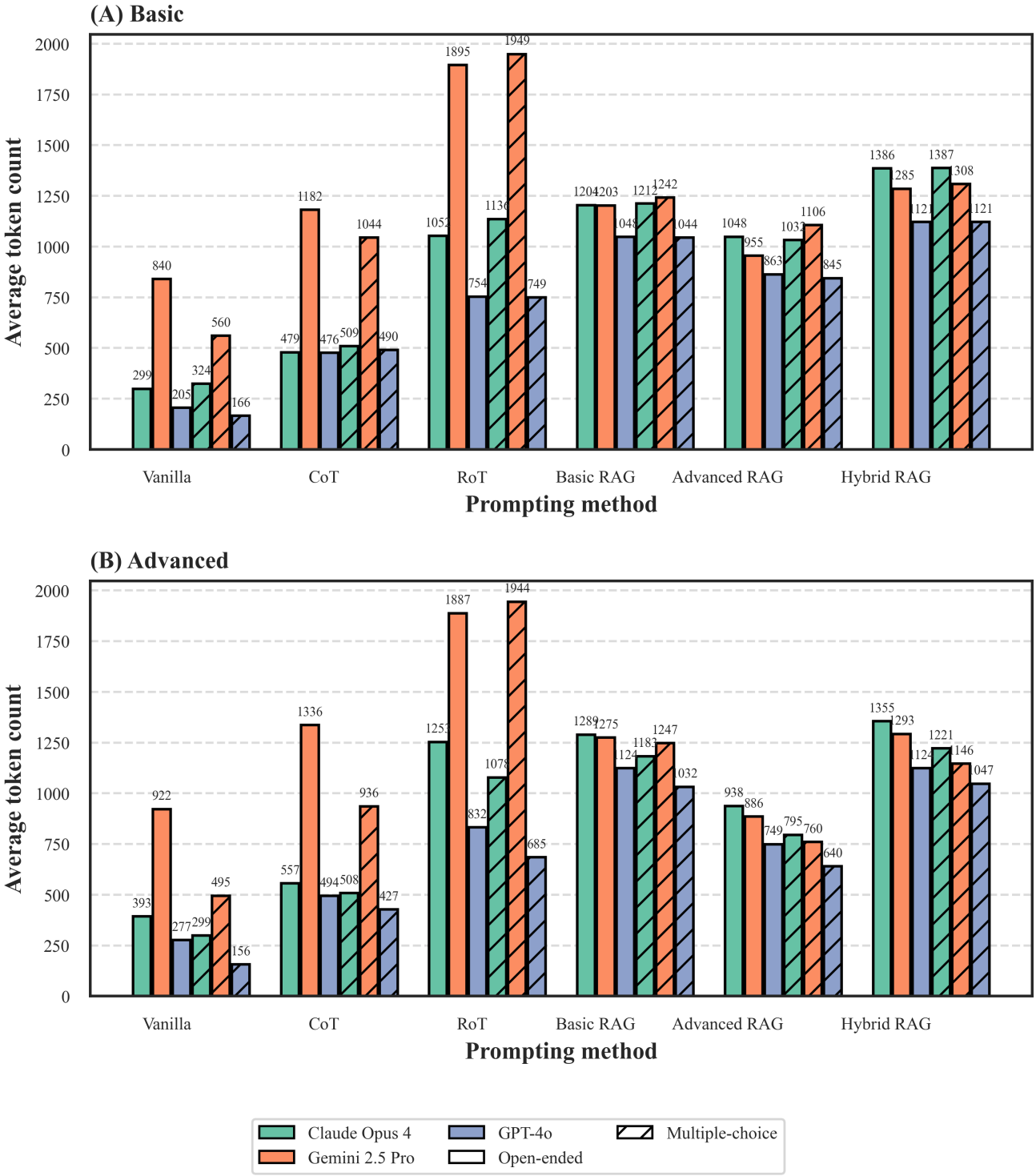

XSL•FO
RenderX

Figure 6. Distribution of total token usage per interaction on the Microsatellite Instability Cancer Benchmark (MSIC-Bench).

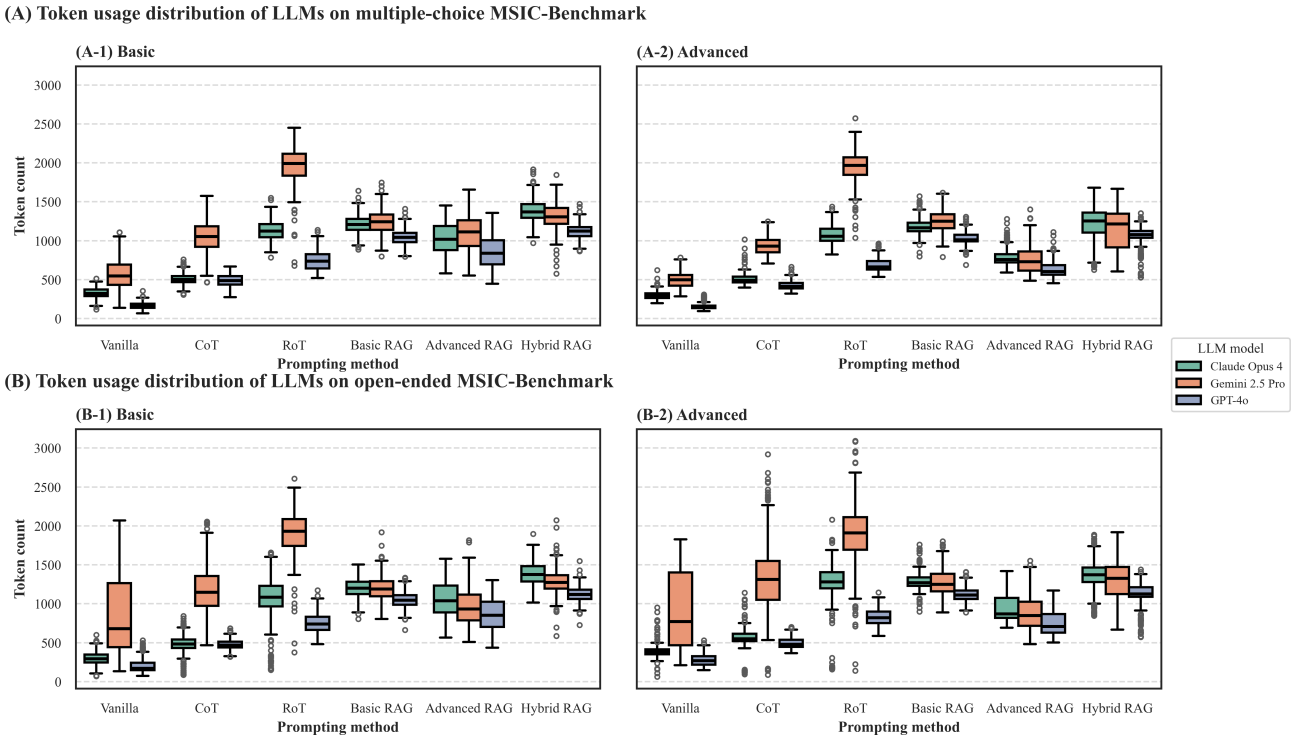

Performance Analysis Relative to Knowledge Cutoff Date

To investigate the influence of training data currency on model performance, we conducted an analysis on the "advanced" tier using GPT-4o, comparing its performance on questions from precutoff vs postcutoff sources (Figure 7).

Figure 7. GPT-4o's (OpenAI) accuracy on questions from before- and aftercutoff sources.

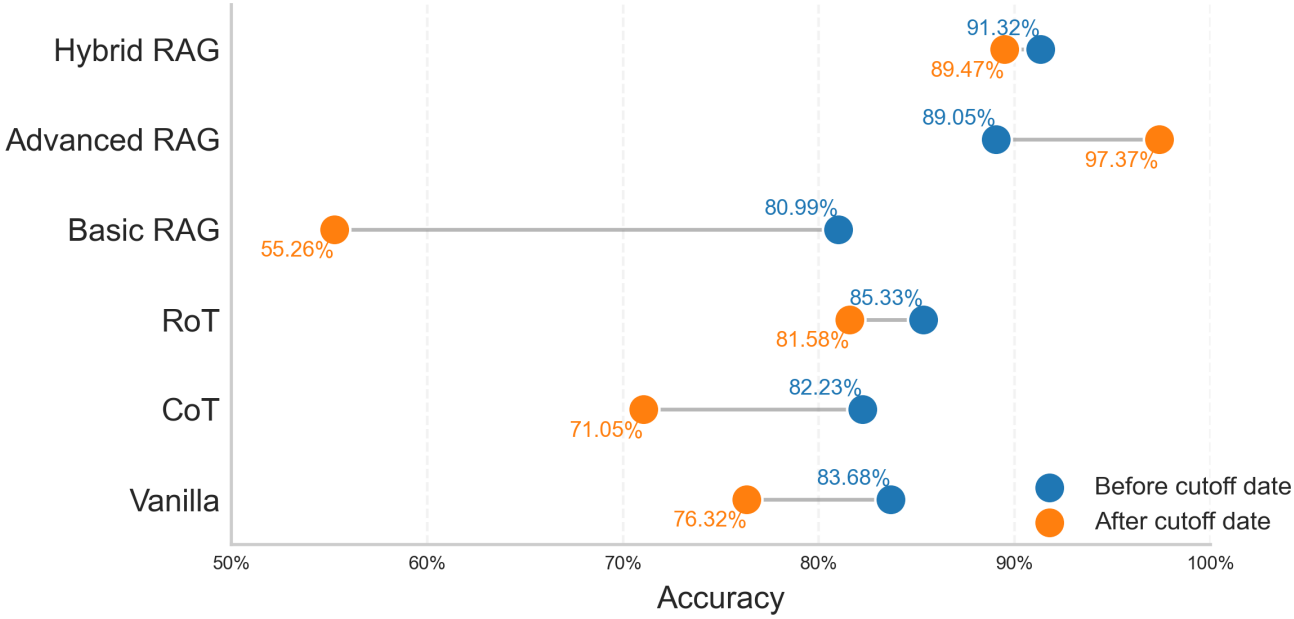

For strategies relying on internal knowledge (vanilla, CoT, and RoT), performance was consistently higher on the precutoff data (eg, 83.68% vs 76.32% for vanilla). This indicates a dependency on the model's training history for these non-RAG methods.

The analysis of RAG-based strategies revealed that performance is critically dependent on knowledge source alignment. A striking example is basic-RAG, where providing out-of-domain guideline knowledge to answer advanced questions led to a sharp performance drop on postcutoff data (55.26%), the lowest of any strategy. This highlights the risk that a misaligned RAG can be more detrimental than no RAG at all.

In contrast, when the knowledge source was correctly aligned, RAG's effectiveness became evident. The advanced RAG strategy achieved the highest accuracy on postcutoff questions (97.37%), demonstrating the model's strong ability to reason

XSL•FO
RenderX

over novel information when the correct context is supplied. Most notably, hybrid RAG maintained high and stable accuracy across both precutoff (91.32%) and postcutoff (89.47%) data, indicating its robustness.

These findings confirm that while non-RAG strategies are sensitive to training data currency, a RAG architecture's reliability is primarily determined by its knowledge base. A comprehensive and well-aligned knowledge source, as demonstrated by a hybrid RAG, allows the system to perform reliably, independent of the model's internal knowledge cutoff.

Discussion

Principal Findings

The increasing application of LLMs in medicine raises a pivotal question regarding their reliability in knowledge-intensive subspecialties such as MSI cancers, a domain where no gold-standard evaluation benchmark existed at the time of our analysis. To address this challenge, we developed MSIC-Bench. This benchmark features a two-tiered structure based on clinical guidelines (eg, NCCN and ESMO) and a curated, evidence-based knowledge base (MSICKB), allowing for separate assessment of an LLM's performance on both consensus and nuanced scientific knowledge. Using this framework, we evaluated 3 leading LLMs across 4 prompting strategies, analyzing their accuracy, safety, error composition, and token usage.

Our findings reveal a significant "scaffolding effect," where selection-based formats inflate accuracy compared to more realistic open-ended scenarios. This performance drop highlighted specific failure modes, notably in chemotherapy-related decisions. To address these failures, RAG proved to be a critical intervention. It not only substantially improves accuracy but also fundamentally shifts the error profile from high-risk "fabrications" to safer "I don't know" responses. Finally, our analysis highlights key complexities for real-world applications: model-specific vulnerabilities, the importance of knowledge-to-task alignment in RAG, and the high token usage of certain prompting strategies.

Our evaluation of MSIC-Bench yields 3 principal insights. First, we identified a pronounced "scaffolding effect": LLMs perform substantially better on multiple-choice than on open-ended questions, indicating that widely used multiple-choice benchmarks likely overestimate their real-world clinical capabilities and that honest behavior cannot be assessed unless the task design explicitly allows refusals. Second, error composition analysis shows that in this specialized domain, the dominant failure mode of non-RAG strategies is internal knowledge deficit, meaning that better reasoning prompts alone cannot compensate for missing domain knowledge; with RAG, this bottleneck shifts from "knowing" to "finding," as retrieval failure becomes the primary source of error. Third, RAG emerges as the key intervention for deploying LLMs in MSI oncology, with performance and safety tightly coupled to knowledge-base quality: domain-aligned and comprehensive resources (as in advanced and hybrid RAG) markedly improve accuracy and convert high-risk fabrications into safer refusals, but also introduce new dependencies on retrieval precision and computational cost, making optimization of the retrieval module and careful curation of knowledge sources (eg, through the use of structured methodologies such as ontologies [35,36] and curated knowledge bases [14,37,38]) central to future clinical AI development.

Comparison With Prior Work

Our findings are consistent with and extend the growing body of literature on LLM evaluation in specialized domains. For instance, the performance gap we identified between multiple-choice and open-ended modalities (the "scaffolding effect") aligns with similar discrepancies observed by Song et al [13] in their study on RAG, confirming that evaluation formats significantly impact perceived model capabilities in clinical contexts.

Furthermore, our findings contribute to the continuing discussion about the roles of reasoning vs knowledge in LLM performance. Our conclusion regarding the limited efficacy of reasoning prompts such as CoT in knowledge-intensive tasks is corroborated by prior work. For instance, prior work [39] similarly found that applying CoT in the knowledge-driven domain of rare disease diagnosis could paradoxically decrease accuracy. Our study provides a clear mechanistic explanation for this phenomenon. By using error composition analysis, we demonstrate that when foundational knowledge is absent—as is often the case in specialized medicine—the primary failure mode is internal knowledge deficit. In such scenarios, structured reasoning prompts cannot compensate for the core knowledge void.

Crucially, our work extends this understanding by showing how RAG transforms the problem. We demonstrate that by supplying external knowledge, RAG effectively eliminates internal knowledge deficit, but in doing so, it shifts the primary bottleneck to retrieval failure. This finding refines the distinction between reasoning-intensive and knowledge-intensive tasks, suggesting that for RAG-based systems, retrieval-intensity emerges as a new, critical dimension for analysis and optimization.

Limitations and Future Directions

Our study has several limitations that also suggest directions for future work. First, MSIC-Bench is limited to a single clinical problem and reflects only a static snapshot of knowledge related to MSI-associated cancers, which restricts its breadth of coverage. Future work should broaden and update the benchmark within MSI cancers by adding more diverse questions, incorporating real-world case-based scenarios, and regularly revising content to reflect evolving evidence and model capabilities. Second, the benchmark is text-only, whereas real-world decision-making is multimodal. An important next step is the development of benchmarks that integrate images, genomic data, and electronic health records. Third, our question-answering format simplifies clinical reasoning, which is often iterative and conversational; future evaluations should include more realistic case-based and dialogue-style tasks.

Fourth, our study used a foundational RAG implementation. Our error analysis identified retrieval failure as the new primary

XSL•FO
RenderX

bottleneck, which points to a clear and critical direction for future work. Research should prioritize the investigation of more sophisticated RAG techniques, such as iterative retrieval and query rewriting, to directly address this retrieval challenge. Fifth, our initial RAG evaluation used a "split-index" (oracle routing) strategy. While this design allowed us to isolate the impact of knowledge source quality, we acknowledge that it does not fully reflect real-world scenarios. Our introduction of the hybrid-RAG model was a first step to address this, but future work should focus on developing more sophisticated dynamic routing mechanisms. Sixth, the same static prompt template was applied across all RAG settings. In the advanced and hybrid RAG strategies, the phrase "clinical guideline context" may not have fully matched the nature of the retrieved MSICKB-derived evidence, which could have influenced how the model interpreted the provided context. Future work should explore source-aware prompt designs that better align the instruction wording with the type of retrieved evidence. Seventh, our analysis revealed a critical trade-off between safety and utility in RAG systems. While RAG effectively curbed dangerous "fabrications," it also introduced "false refusals"—a failure to answer otherwise answerable questions that limits clinical utility. This highlights that the goal for future research is not simply to maximize refusals for safety, but to find an optimal balance. Future work should therefore focus on strategies to minimize false refusals without reintroducing a tendency to fabricate.

Eighth, our analysis of parametric memory was constrained by model cutoff dates, preventing a direct before-and-after comparison for Claude and Gemini. However, our conclusion that "seeing is not knowing" is supported by strong indirect evidence: our error composition analysis (Figure 4) shows that internal knowledge deficit remained the dominant non-RAG failure mode for all models. This finding is consistent with multiple studies showing that LLMs often fail to reliably recall information, even when it is present in their training data, a phenomenon particularly pronounced in specialized, long-tail knowledge domains such as MSI cancers [40,41]. Ninth, as the MSICKB knowledge base and the "advanced" tier questions were developed by the same research group, a discussion of potential circular validation is warranted. To mitigate this risk, every ground truth answer is strictly grounded in and explicitly linked to its source peer-reviewed publication via a PubMed ID. However, more robust external validation remains an important direction for our future work. Our primary planned next step is to collect and annotate real-world clinical data to create an independent test set for validating our benchmark and retrieval framework. We also note that this specific challenge of circular validation, where the retrieval corpus overlaps with the benchmark's source material, is a widely recognized issue in the evaluation of specialized RAG systems [42,43]. Additionally, our token usage estimates relied on a single tokenizer (tiktoken), which only approximates tokenization behavior for non-OpenAI models. Future work should use model-specific tokenizers to obtain more accurate cost estimates. Finally, because LLMs evolve rapidly, our results capture only a single time point, emphasizing the need for continuous and longitudinal benchmarking in clinical AI.

Conclusion

In this study, we addressed the critical absence of a standardized evaluation tool for LLMs in MSI cancers by developing and applying MSIC-Bench. Our benchmark revealed that the primary limiter of current LLMs is a profound deficit in domain-specific knowledge. Our results demonstrate that a well-designed RAG architecture is the pivotal intervention to address this knowledge gap. However, it is not a simple fix. We found that RAG transforms the system's failure modes, shifting the primary bottleneck from knowledge deficits to retrieval failures. This insight establishes a core principle for the development of trustworthy clinical AI: the quality and comprehensiveness of the knowledge base and the precision of the retrieval system are the most critical components. Ultimately, enhancing clinical RAG systems requires a dual focus on optimizing retrieval mechanisms and improving the model's contextual reasoning. MSIC-Bench provides not only a comprehensive capability profile of LLMs in this domain but also a replicable methodology to guide this future development, ensuring that knowledge-augmented AI assistants can be validated and safely integrated into specialized medical fields.

Acknowledgments

We extend our sincere gratitude to the 2 clinical experts from a major academic medical center in China who provided invaluable contributions to this study. Their expertise was essential for the benchmark's development and the manual evaluation of the large language model (LLM) outputs. Detailed, anonymized qualifications for these experts are provided in Table S1 in Multimedia Appendix 1. Generative artificial intelligence (Gemini 2.5 Pro; Google) was used to support linguistic refinement and coherence in the paper text. All conceptual contributions, interpretations, and substantive arguments presented in this paper are solely those of the authors.

Funding

This work was supported by the National Natural Science Foundation of China (grant nos 32270690 and 32570773) and the Sichuan Province Science and Technology Program (grant no 2024YFHZ0205). The funder played no role in the study design, data collection, analysis, and interpretation of data, or paper writing.

Data Availability

The Microsatellite Instability Cancer Benchmark (MSIC-Bench) benchmark data, analysis scripts, and additional materials are openly available in a GitHub repository [44]. Specifically, the exact versions of the National Comprehensive Cancer Network

XSL•FO
RenderX

(NCCN) and European Society for Medical Oncology (ESMO) guidelines analyzed in this study are permanently archived and can be directly accessed in the Knowledge_Sources folder. The full outputs generated during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

JS, YZ, and BS contributed to conceptualization. YZ and JS contributed to data curation. YZ, JS, and CB conducted the formal analysis. BS acquired funding. YZ, JS, and ZX conducted the investigation. YZ, JS, and CB developed the methodology. BS led project administration and supervision. JS, YZ, and ZX contributed to software development. YZ and JS performed validation. YZ, JS, and XZ contributed to visualization. YZ and JS prepared the original draft. YZ, JS, BS, XZ, and DC reviewed and edited the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Key methodological details—including full system prompts, model specifications, and error analysis taxonomies—to ensure transparency and reproducibility of the main study.
[DOCX File, 5508 KB - jmir_v28i1e88614_app1.docx]

References

1. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011 Mar 4;144(5):646-674. [doi: 10.1016/j.cell.2011.02.013] [Medline: 21376230]
2. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. Cell 2013 Jun 6;153(6):1194-1217. [doi: 10.1016/j.cell.2013.05.039] [Medline: 23746838]
3. Marcus L, Lemery SJ, Keegan P, Pazdur R. FDA approval summary: pembrolizumab for the treatment of microsatellite instability-high solid tumors. Clin Cancer Res 2019 Jul 1;25(13):3753-3758. [doi: 10.1158/1078-0432.CCR-18-4070] [Medline: 30787022]
4. Nojadeh JN, Behrouz Sharif S, Sakhinia E. Microsatellite instability in colorectal cancer. EXCLI J 2018;17:159-168. [doi: 10.17179/excli2017-948] [Medline: 29743854]
5. Luchini C, Bibeau F, Ligtenberg MJL, et al. ESMO recommendations on microsatellite instability testing for immunotherapy in cancer, and its relationship with PD-1/PD-L1 expression and tumour mutational burden: a systematic review-based approach. Ann Oncol 2019 Aug 1;30(8):1232-1243. [doi: 10.1093/annonc/mdz116] [Medline: 31056702]
6. Sahin IH, Akce M, Alese O, et al. Immune checkpoint inhibitors for the treatment of MSI-H/MMR-D colorectal cancer and a perspective on resistance mechanisms. Br J Cancer 2019 Nov;121(10):809-818. [doi: 10.1038/s41416-019-0599-y] [Medline: 31607751]
7. Kather JN, Pearson AT, Halama N, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med 2019 Jul;25(7):1054-1056. [doi: 10.1038/s41591-019-0462-y] [Medline: 31160815]
8. Yang Y, Zhao Y, Liu X, Huang J. Artificial intelligence for prediction of response to cancer immunotherapy. Semin Cancer Biol 2022 Dec;87:137-147. [doi: 10.1016/j.semcancer.2022.11.008] [Medline: 36372326]
9. Bai S, Singh B, Ethakota J, et al. The role of artificial intelligence in colorectal cancer and polyp detection: a systematic review. JCO 2025 Feb;43(4_suppl):47-47. [doi: 10.1200/JCO.2025.43.4_suppl.47]
10. GPT-4o. OpenAI Developers. URL: https://platform.openai.com/docs/models/gpt-4o [accessed 2026-02-03]
11. Claude platform. Claude API Docs. URL: https://platform.claude.com/docs/en/release-notes/overview#may-22-2025 [accessed 2026-02-04]
12. Gemini 3. Gemini API. URL: https://ai.google.dev/gemini-api/docs/models [accessed 2026-02-03]
13. Song J, Xu Z, He M, Feng J, Shen B. Graph retrieval augmented large language models for facial phenotype associated rare genetic disease. NPJ Digit Med 2025 Aug 24;8(1):543. [doi: 10.1038/s41746-025-01955-x] [Medline: 40849403]
14. Zhang C, Yang H, Liu X, et al. A knowledge-enhanced platform (MetaSepsisKnowHub) for retrieval augmented generation-based sepsis heterogeneity and personalized management: development study. J Med Internet Res 2025 Jun 6;27:e67201. [doi: 10.2196/67201] [Medline: 40478618]
15. Li J, Zong H, Wu E, et al. Exploring the potential of artificial intelligence to enhance the writing of English academic papers by non-native English-speaking medical students - the educational application of ChatGPT. BMC Med Educ 2024 Jul 9;24(1):736. [doi: 10.1186/s12909-024-05738-y] [Medline: 38982429]
16. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature New Biol 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
17. Zhang Y, Li X, Zheng X, et al. MSICKB: a curated knowledgebase for exploring molecular heterogeneity and biomarker prioritization in microsatellite instability cancers. Comput Struct Biotechnol J 2026;35(1):0047. [doi: 10.34133/csbj.0047] [Medline: 42017048]

XSL•FO
RenderX

18. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv. Preprint posted online on Jan 28, 2022. [doi: 10.48550/ARXIV.2201.11903]
19. Wang L, Chen X, Deng X, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med 2024 Feb 20;7(1):41. [doi: 10.1038/s41746-024-01029-4] [Medline: 38378899]
20. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv. Preprint posted online on May 22, 2020. [doi: 10.48550/ARXIV.2005.11401]
21. Colon cancer. National Comprehensive Cancer Network. URL: https://www.nccn.org/guidelines/guidelines-detail?category=1&id=1428 [accessed 2026-02-02]
22. Colorectal cancer screening. National Comprehensive Cancer Network. URL: https://www.nccn.org/guidelines/guidelines-detail?category=1&id=1429 [accessed 2026-02-11]
23. Ajani JA, D'Amico TA, Bentrem DJ, et al. Gastric cancer, version 2.2025, NCCN Clinical Practice Guidelines In Oncology. J Natl Compr Canc Netw 2025 May;23(5):169-191. [doi: 10.6004/jnccn.2025.0022] [Medline: 40341199]
24. Rectal cancer. National Comprehensive Cancer Network. URL: https://www.nccn.org/guidelines/guidelines-detail?category=1&id=1461 [accessed 2026-02-11]
25. Abu-Rustum NR, Campos SM, Amarnath S, et al. NCCN Guidelines® insights: uterine neoplasms, version 3.2025. J Natl Compr Canc Netw 2025 Aug;23(8):284-291. [doi: 10.6004/jnccn.2025.0038] [Medline: 40763788]
26. Oaknin A, Bosse TJ, Creutzberg CL, et al. Endometrial cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. Ann Oncol 2022 Sep;33(9):860-877. [doi: 10.1016/j.annonc.2022.05.009] [Medline: 35690222]
27. Lordick F, Carneiro F, Cascinu S, et al. Gastric cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. Ann Oncol 2022 Oct;33(10):1005-1020. [doi: 10.1016/j.annonc.2022.07.004] [Medline: 35914639]
28. Argilés G, Tabernero J, Labianca R, et al. Localised colon cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2020 Oct;31(10):1291-1305. [doi: 10.1016/j.annonc.2020.06.022] [Medline: 32702383]
29. Hofheinz RD, Fokas E, Benhaim L, et al. Localised rectal cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. Ann Oncol 2025 Sep;36(9):1007-1024. [doi: 10.1016/j.annonc.2025.05.528] [Medline: 40412553]
30. Cervantes A, Adam R, Roselló S, et al. Metastatic colorectal cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. Ann Oncol 2023 Jan;34(1):10-32. [doi: 10.1016/j.annonc.2022.10.003] [Medline: 36307056]
31. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, et al. The Cancer Genome Atlas pan-cancer analysis project. Nat Genet 2013 Oct;45(10):1113-1120. [doi: 10.1038/ng.2764] [Medline: 24071849]
32. Model cards. Google DeepMind. URL: https://deepmind.google/models/model-cards/ [accessed 2026-02-04]
33. openai / tiktoken. GitHub. 2025. URL: https://github.com/openai/tiktoken [accessed 2025-11-23]
34. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]
35. Liu X, Yang Y, Zong H, et al. Core reference ontology for individualized exercise prescription. Sci Data 2024 Dec 18;11(1):1349. [doi: 10.1038/s41597-024-04217-9] [Medline: 39695140]
36. Yu C, Zong H, Chen Y, et al. PCAO2: an ontology for integration of prostate cancer associated genotypic, phenotypic and lifestyle data. Brief Bioinform 2024 Mar 27;25(3):bbae136. [doi: 10.1093/bib/bbae136] [Medline: 38557678]
37. Bi C, Zheng X, Zhang Y, et al. NDDRF 2.0: an update and expansion of risk factor knowledge base for personalized prevention of neurodegenerative diseases. Alzheimers Dement 2025 May;21(5):e70282. [doi: 10.1002/alz.70282] [Medline: 40371632]
38. Zheng X, Bi C, Bo W, et al. DRPMKB1.0: a comprehensive knowledge base for an AI-oriented drug repositioning prediction model. J Chem Inf Model 2026 Jan 12;66(1):122-137. [doi: 10.1021/acs.jcim.5c01945] [Medline: 41474981]
39. Li F, Fang P, Shi Z, et al. CoT-RAG: integrating chain of thought and retrieval-augmented generation to enhance reasoning in large language models. arXiv. Preprint posted online on Apr 18, 2025. [doi: 10.18653/v1/2025.findings-emnlp.168]
40. Zhang H, Zhang Y, Li X, et al. Evaluating the external and parametric knowledge fusion of large language models. arXiv. Preprint posted online on May 29, 2024. [doi: 10.48550/ARXIV.2405.19010]
41. Kandpal N, Deng H, Roberts A, Wallace E, Raffel C. Large language models struggle to learn long-tail knowledge. arXiv. Preprint posted online on Nov 15, 2022. [doi: 10.48550/ARXIV.2211.08411]
42. Shen W, Wang M, Wang Y, et al. Are we on the right way for assessing document retrieval-augmented generation? arXiv. Preprint posted online on Aug 5, 2025. [doi: 10.48550/ARXIV.2508.03644]
43. Masanneck L, Meuth SG, Pawlitzki M. Evaluating base and retrieval augmented LLMs with document or online support for evidence based neurology. NPJ Digit Med 2025 Mar 4;8(1):137. [doi: 10.1038/s41746-025-01536-y] [Medline: 40038423]
44. ExplorerZhyuxin/MSIC-benchmark. GitHub. URL: https://github.com/ExplorerZhyuxin/MSIC-Benchmark/tree/main [accessed 2026-02-26]

Abbreviations

AI: artificial intelligence
API: application programming interface
CoT: chain-of-thought

XSL•FO
RenderX

ESMO: European Society for Medical Oncology
LLM: large language model
MSI: microsatellite instability
MSIC-Bench: Microsatellite Instability Cancer Benchmark
MSICKB: Microsatellite Instability Cancer Knowledgebase
NCCN: National Comprehensive Cancer Network
RAG: retrieval-augmented generation
RoT: reflection of thoughts

Edited by A Coristine; submitted 28.Nov.2025; peer-reviewed by X Lai, Y Lee; revised version received 10.Apr.2026; accepted 10.Apr.2026; published 21.May.2026.

Please cite as:
Zhang Y, Song J, Bi C, Zheng X, Xu Z, Cao D, Shen B
Benchmarking Large Language Models and Prompt Engineering Strategies in Microsatellite Instability Cancers: Evaluation Study
J Med Internet Res 2026;28:e88614
URL: https://www.jmir.org/2026/1/e88614
doi:10.2196/88614

© Yuxin Zhang, Jie Song, Cheng Bi, Xin Zheng, Zhichuan Xu, Dan Cao, Bairong Shen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Reliability of Large Language Model Generated Clinical Reasoning in Assisted Reproductive Technology: Blinded Comparative Evaluation Study

Dou Liu[1,2,3*], BSc; Ying Long[1,3,4*], MD; Sophia Zuoqiu[5], PhD; Di Liu[5,6,7], PhD; Kang Li[6,7], PhD; Yiting Lin[8], MBBS; Hanyi Liu[8], MBBS; Rong Yin[5,6,7*], PhD; Tian Tang[1,3,4*], MD

[1]Department of Obstetrics and Gynecology, West China Second University Hospital of Sichuan University, Chengdu, China

[2]Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI, United States

[3]Key Laboratory of Birth Defects and Related Diseases of Women and Children, Sichuan University, Chengdu, China

[4]Reproductive Medical Center, Department of Obstetrics and Gynecology, West China Second University Hospital of Sichuan University, Chengdu, China

[5]Department of Industrial Engineering, Sichuan University, Chengdu, China

[6]West China Biomedical Big Data Center, West China Hospital of Sichuan University, Chengdu, China

[7]Med-X Center for Informatics, Sichuan University, Chengdu, China

[8]West China School of Medicine, Sichuan University, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Tian Tang, MD
Department of Obstetrics and Gynecology
West China Second University Hospital of Sichuan University
Chengdu
China
Phone: 1 18081110066
Email: tiantang2016@scu.edu.cn

Abstract

Background: High-quality clinical chains-of-thought (CoTs) are essential for explainable medical artificial intelligence (AI); yet, their development is limited by data scarcity. Large language models can generate medical CoTs, but their clinical reliability is unclear.

Objective: We evaluated the clinical reliability of large language model–generated CoTs in reproductive medicine and examined prompting strategies to improve their quality.

Methods: In a blinded comparative study at a clinical center, senior clinicians in assisted reproductive technology evaluated CoTs generated via 3 distinct strategies: zero-shot, random few-shot (using random shallow examples), and selective few-shot (using diverse, high-quality examples). Expert ratings were then compared with evaluations from a state-of-the-art AI model (GPT-4o).

Results: The selective few-shot strategy significantly outperformed other strategies across logical clarity, use of key information, and clinical accuracy ($P<.001$). Critically, the random few-shot strategy offered no significant improvement over the zero-shot baseline, demonstrating that low-quality examples are as ineffective as no examples. The success of the selective strategy is attributed to 2 preliminary frameworks: “gold-standard depth” and “representative diversity.” Notably, the AI evaluator failed to discern these critical performance differences. Thus, clinical reliability depends on strategic prompt design rather than simply adding examples.

Conclusions: We propose a “dual principles” preliminary framework for generating trustworthy CoTs at scale in assisted reproductive technology. This work is a preliminary step toward addressing the data bottleneck in reproductive medicine. It also underscores the essential role of human expertise in evaluating generated clinical data.

(J Med Internet Res 2026;28:e85206) doi:10.2196/85206

XSL•FO
RenderX

KEYWORDS

chain-of-thought; large language model; assisted reproductive technology; explainable artificial intelligence; clinical data reliability

Introduction

Background

Assisted reproductive technology (ART) represents a cornerstone of modern medicine, offering solutions for millions facing infertility [1]. The clinical decision-making process in ART is exceptionally complex, requiring the synthesis of high-dimensional patient data, including baseline characteristics and medical history. This process is time-consuming and fraught with risk for both clinicians and patients, as minute variations in treatment protocols can lead to significant adverse outcomes. Furthermore, clinicians must navigate patients' personal values and ethical considerations, demanding a highly personalized and explainable approach to care [2].

Recent advancements in artificial intelligence (AI), particularly large language models (LLMs), have demonstrated considerable promise for answering medical questions, addressing clinical case challenges, and augmenting clinical diagnosis [3-7]. Within clinical decision support systems, these technologies can help synthesize large amounts of data, facilitating more comprehensive and standardized therapeutic strategies. However, while general-purpose LLMs like ChatGPT-4 and Gemini are powerful, their training on broad, nonspecialized data limits their utility in niche medical domains. Consequently, high-performing clinical AI applications are typically fine-tuned from general models using curated, domain-specific datasets [8-10]. The actual bottleneck, however, is not a lack of raw clinical data, but a profound lack of explainable data—data that record not just what decision was made, but why. This meticulous, expert-level reasoning, often captured as a chain-of-thought (CoT), is the very fuel required to train AI models that are not just accurate, but also trustworthy and scalable to clinicians. To move beyond "black-box" predictions, models require structured reasoning pathways, or CoT data, which simulate clinical logic and enhance explainability [11,12]. The challenge, therefore, narrows down to a scarcity of expert-authored CoT data within the specific area. The manual creation of such a dataset on a large scale is prohibitively expensive and time-consuming, presenting a significant barrier to progress in explainable medical AI.

To address this challenge, a promising direction is to leverage the generative capabilities of LLMs to synthesize clinical CoT data at scale. While this offers a scalable solution to the data bottleneck, it hinges on a critical, unverified assumption: the clinical reliability of the generated content. In a high-stakes domain like ART, this assumption cannot be taken for granted.

Therefore, this study is designed to examine this uncertainty under the ART setting through a rigorous, head-to-head empirical comparison. Figure 1 presents the conceptual framework of our comparative evaluation study. We hypothesize that a selective few-shot strategy, meticulously crafted with diverse and deeply reasoned examples, will improve the factual accuracy, logical clarity, and clinical information use of LLM-generated reasoning in infertility decision-making scenarios, compared with generic zero-shot and random few-shot prompting. To test this, we developed a novel prompting framework and validated it through a blinded evaluation protocol where senior clinicians from the reproductive department assessed the quality of CoTs from all 3 strategies. In a secondary analysis, we further contrast these expert assessments against the state-of-the-art (SOTA) AI evaluator (GPT-4o) at that time to critically examine the current capabilities and limitations of automated evaluation paradigms, which are widely used in supervised data generation. Ultimately, this work aims to establish an exploratory, practical, evidence-based methodology for the trustworthy generation of clinical reasoning of ART at scale.

XSL•FO
RenderX

Figure 1. Conceptual framework of the comparative evaluation study. EHR: electronic health record; LCC: logical coherence and clarity; LLM: large language model; PCAR: plausibility and clinical accuracy of reasoning; UCKI: use and coverage of key information.

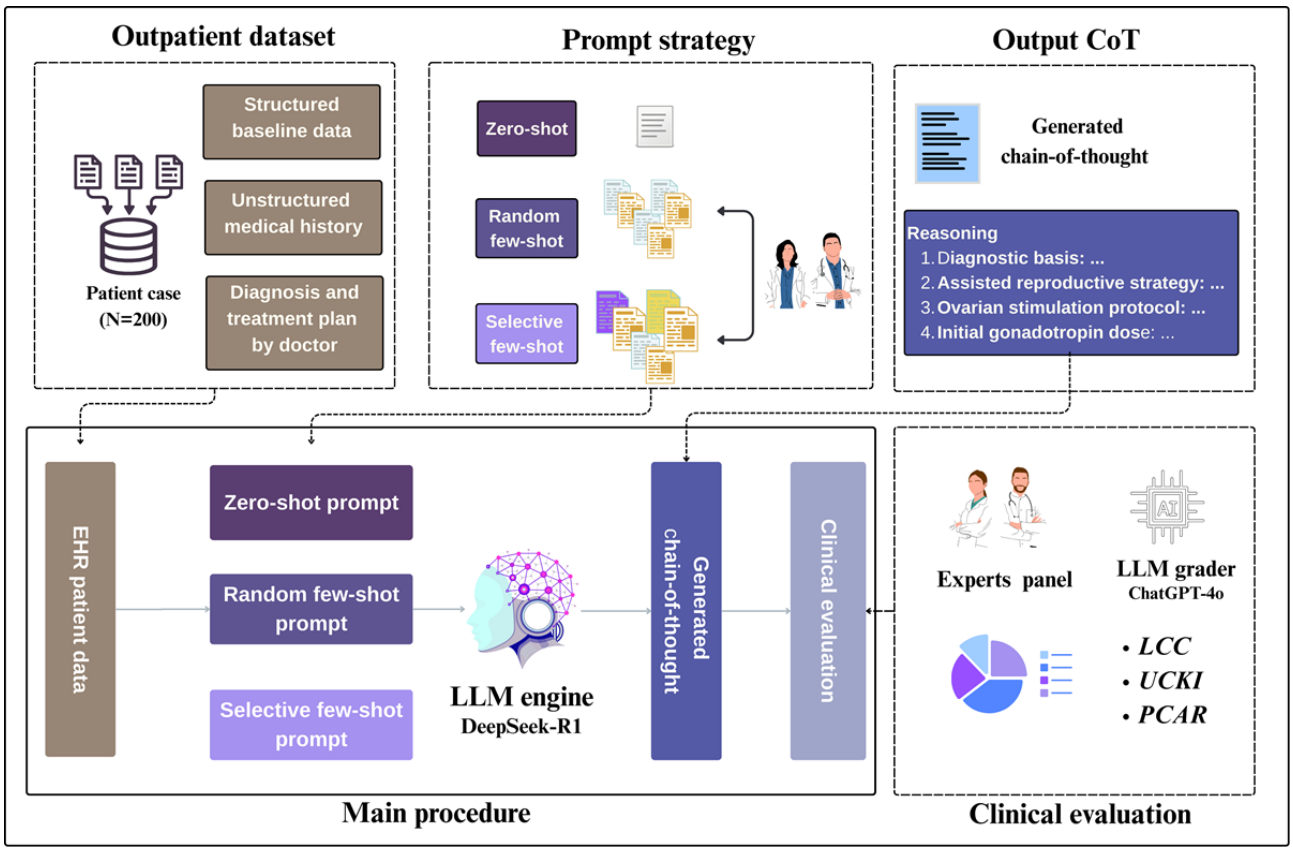

The study workflow begins with a standardized patient case (N=200) as the input. Three distinct prompting strategies are evaluated in parallel pipelines: (1) zero-shot, which uses instructions only; (2) random few-shot, which uses 5 randomly selected examples; and (3) selective few-shot, which uses a curated set of 6 diverse examples representing all major ART categories in the cases. Each strategy, powered by the same LLM engine, generates a unique CoT. All 3 generated CoTs are then subjected to a rigorous, blinded "Doctor-in-the-Loop" evaluation by 2 parallel assessors: human clinical experts (the gold standard) and the SOTA AI evaluator (GPT-4o). This dual evaluation process yields the final reliability scores and rankings for each strategy.

LLMs in Health Care

Since the launch of ChatGPT-4, LLMs have rapidly spread into many industries, such as education, finance, and health care. For instance, Google's Med-PaLM 2, a leading specialized health care model, achieved 86.5% accuracy on the MedQA benchmark, a popular multiple-choice open domain question answering (OpenQA) medical problems dataset. Furthermore, its responses were preferred over those of generalist physicians in 65% of expert evaluations [13]. The LLMs are now used in many healthcare–related workflows, ranging from medical documentation assistance to clinical differential diagnosis [5,14-16]. However, to effectively address highly specialized tasks, these models are typically fine-tuned from pretrained LLMs using carefully curated datasets. Despite their impressive capabilities, current LLMs usually function as black boxes, producing outputs without offering interpretable reasoning. In clinical practice, however, physicians often require not just answers but also transparent explanations. This requires models beyond black-box behavior and provides interpretable, step-by-step reasoning processes. The increasing reliance on LLMs has also intensified the demand for high-quality data [17]. Alarmingly, some predict that the global supply of novel text data may be exhausted by 2050 and image data by 2060 [18]. In the health care domain, the situation is even more critical: clinical data are not only scarce but also highly sensitive and expensive to obtain. As a result, a central challenge emerges—how can we build datasets that are both sufficiently large and clinically trustworthy to support transparent, reliable medical AI systems?

Synthetic Data in Health Care

To overcome the data shortage in health care, researchers are increasingly turning to LLMs to create synthetic data. This approach is promising for several reasons. It allows for data generation at scale, addressing issues of data scarcity and privacy [19,20]. Furthermore, synthetic data can be tailored to balance underrepresented patient groups, potentially improving model robustness and fairness [21]. Generative models have demonstrated remarkable success in these areas, with some studies showing that LLM-generated narratives can be indistinguishable from those written by physicians [22]. This potential, however, is inextricably linked to a profound challenge: reliability. While LLMs can mimic the style of clinical text, ensuring the factual accuracy and clinical plausibility of the content is a far more difficult task. For instance, models have been used to generate both structured

XSL•FO
RenderX

electronic health records (EHRs) and unstructured clinical notes [23], but in both cases, the risk of hallucination—where the model generates incorrect or nonsensical information—poses a significant threat in high-stakes medical applications. Therefore, the core challenge moves beyond mere data generation to a more fundamental question of trust. While studies have shown that synthetic data can be effective for certain human labeling and fine-tuning tasks [24-28], these applications often involve relatively straightforward data points. The problem is magnified when the task requires complex, multistep medical reasoning. In such scenarios, "synthetic" must not equate to "inaccurate." This underscores the urgent need for rigorous evaluation methods, not just for the data points themselves, but for the underlying reasoning processes that produce clinical decisions. Our work focuses on this critical next step: assessing the reliability of synthetically generated reasoning paths.

Chain-of-Thought

CoT is a prompt engineering technique that enhances the output of LLMs, particularly for complex tasks involving multistep reasoning. It facilitates problem-solving by guiding the model through a step-by-step reasoning process by using a coherent series of logical steps [29]. This approach has been shown to significantly elevate performance on a wide range of complex reasoning tasks in general domains, especially for arithmetic problems and logical reasoning tasks [30,31]. To enhance the reasoning ability in domain-specific tasks, researchers have started fine-tuning the models with CoTs [12]. Within the medical domain, the potential of CoT is particularly compelling. Its step-by-step nature aligns naturally with the differential diagnosis and clinical reasoning processes used by physicians. Consequently, researchers have begun to apply CoT prompting to improve accuracy on medical question-answering benchmarks and in practice diagnosis [11,32]. More importantly, CoT offers a crucial pathway toward explainable AI in medicine. By externalizing the model's reasoning processes, CoT allows clinicians to scrutinize, understand, and ultimately trust the AI's recommendations, which is a prerequisite for its safe integration into clinical workflows.

The application of CoT is rapidly evolving. Beyond simple prompting, a new frontier in clinical AI is the (1) fine-tuning of models on datasets enriched with CoT data to build inherently more explainable systems, which, however, immediately confronts the fundamental bottleneck of medical AI; and the (2) prohibitive cost and time required for expert clinicians to manually author thousands of high-quality reasoning paths for a training set. An intuitive and scalable solution is to leverage foundational LLMs to synthetically generate these CoTs, creating a cost-effective pathway to train the next generation of trustworthy medical models. However, this entire paradigm hinges on a critical, yet largely unexamined, question: (3) Is the reliability of synthetically generated CoTs adequate to support their application in complex clinical scenarios? Literature to date offers little guidance. Most research focuses on the extrinsic value of CoT (ie, improving final answer accuracy), with scant attention paid to the intrinsic reliability of the reasoning itself. A model fine-tuned on flawed, albeit synthetically generated, logic could learn to produce seemingly correct answers for the wrong reasons—a risk that is unacceptable in clinical practice. Furthermore, standardized, expert-driven protocols for assessing the clinical validity, coherence, and faithfulness of machine-generated reasoning are conspicuously absent. Our study is designed to directly fill this foundational gap. Before the field can confidently use synthetic CoT for model training at scale, we must first have a rigorous method to measure its reliability. Therefore, we propose and implement a blinded, expert-led evaluation framework to answer the fundamental question: how reliable is synthetically generated sophisticated clinical reasoning, and what is the best prompting strategy to elicit it from LLMs?

Methods

Data Source

From the manually reviewed dataset, we randomly selected 200 cases as our evaluation set, covering a variety of ART. These ARTs are broadly categorized into 3 generations: in vitro fertilization (IVF), intracytoplasmic sperm injection (ICSI), and preimplantation genetic testing (PGT). Each generation includes several clinical subtypes, such as short-protocol IVF and IVF with donor sperm. Among the 200 evaluated cases, IVF accounts for the largest proportion (140/200, 70%), including standard IVF (116/200, 58%), IVF with donor sperm (9/200, 4.5%), and short-protocol IVF (short-time insemination, 15/200, 7.5%). The second most common is ICSI (38/200, 19%), comprising standard ICSI (26/200, 13%), IVF+ICSI (5/200, 2.5%), and TESA (testicular sperm aspiration)+ICSI (7/200, 3.5%). PGT represents 11% (22/200) of the dataset, including preimplantation genetic testing for aneuploidies (PGT-A; 6/200, 3%), preimplantation genetic testing for monogenic disorder (PGT-M; 3/200, 1.5%), and PGT for structural rearrangements (13/200, 6.5%).

As shown in Table 1, the dataset consisted of three main components: (1) a structured set of baseline and demographic variables, (2) the preliminary diagnosis and treatment plan, and (3) an unstructured narrative description of the present illness history. The structured baseline data served as the quantitative and categorical foundation for clinical assessment, encompassing key indicators of ovarian reserve such as anti-Müllerian hormone and baseline follicle-stimulating hormone levels. The unstructured narrative provided essential clinical context, offering a detailed account of the patient's medical journey, which is critical for nuanced and context-aware medical reasoning. The output data, labeled as preliminary diagnosis and treatment plan, reflects the clinical conclusions and therapeutic strategies formulated by human experts in prior encounters. This component serves as the ground truth outcome. The LLM's task is to generate a reasoning path that logically connects the patient's input data to this expert-defined outcome. A detailed breakdown of all case data variables, including structured baseline indicators and narrative clinical history, is provided in Multimedia Appendix 1. Together, these inputs formed the foundation for CoT generation and model evaluation.

XSL•FO
RenderX

Table 1. Structure and description of input and output variables for each case[a].

Category	Variable
Baseline and demographics	• Female age • Menstrual cycle • Body weight • BMI • Anti-Müllerian hormone level • Duration of infertility • Gynecological ultrasound findings • Baseline follicle-stimulating hormone level
Present illness history	• Present illness history
Preliminary diagnosis and treatment plan	• Type of infertility • Controlled ovarian stimulation protocol • Initial gonadotropin dosage • Preliminary differential diagnosis • Initial assisted reproductive technology strategy

[a]This table outlines the variables provided to the large language model for each case, categorized into input (baseline and demographic and present illness history) and output (preliminary diagnosis and treatment plan). These variables form the basis for the chain-of-thought generation task.

Ethical Considerations

A set of EHRs recorded between 2020 and 2022 in the Infertility Outpatient Department at West China Second University Hospital was considered in this study. The study was approved by the institutional review board of West China Second University Hospital, Sichuan University (ID: 2022288). The EHRs have been manually reviewed and corrected to ensure data accuracy. All EHR data used in this study were fully deidentified before being accessed by the research team. The deidentification procedure followed HIPAA (Health Insurance Portability and Accountability Act) Safe Harbor standards, including the removal of all direct identifiers (eg, name, date of birth, medical record number, contact information, and provider information) and all quasi-identifiers (eg, dates, locations, and institutional identifiers). Only aggregated clinical descriptors necessary for the reasoning task (eg, high-level patient history and laboratory summaries) were retained. Access to the deidentified dataset was restricted to authorized study personnel through institution-managed credentials and encrypted storage. Reviewers performing the blinded evaluation accessed only the deidentified clinical vignettes and model-generated reasoning content through a secure, read-only interface; no downloads or reidentification attempts were permitted. All access was logged and monitored by an internal auditor to ensure compliance with institutional clinical data governance policies. Because the study used retrospective, fully deidentified EHRs, the requirement for updated informed consent for the following analysis was waived by the institutional review board in accordance with national regulations. No participants were contacted for this study, and no compensation was provided to participants.

Experiment Design

Overview

To systematically evaluate the reliability of LLM-generated CoT and to determine the impact of different prompting strategies, we designed a comparative study. The experiment was structured into 3 distinct arms, each representing a different level of contextual information provided to the model. Our design philosophy was to create a controlled, stepwise comparison to isolate the effects of in-context examples and the strategy used for their selection.

All 3 groups used the evaluation dataset (N=200) described in the data source, making sure of a fair comparison. A capable “Teacher Model” is key to generating better-quality data [33]. Considering the models’ performance so far, we used the open-source model DeepSeek-R1-671b, which was known for its outstanding reasoning capability, as our consistent model shared by 3 arms [34]. Across all 3 arms, the LLM was assigned the same core task: generating a detailed, step-by-step ART CoT by integrating all provided patient information and the corresponding expert-provided reference output. All the inference was conducted via the application programming interface call. The model was executed with temperature=0.5 and max tokens=5000. We used temperature=0.5 to reflect typical clinical-LLM use, where deterministic decoding (temperature=0) may produce rigid or incomplete clinical reasoning. All prompting strategies were evaluated under identical generation settings to ensure fair comparison.

Group 1: Zero-Shot Baseline

In this group, we aimed to establish a fundamental baseline to evaluate the out-of-the-box clinical reasoning capabilities of general-purpose LLMs when applied to this specialized task. To this end, the model was prompted using a standardized directive instruction, with each clinical case embedded directly into the prompt (see Multimedia Appendix 2 for details). The outputs generated by the model, along with corresponding physician evaluations, served as a performance floor, quantifying the baseline reliability and limitations of an unadapted LLM in handling novel clinical scenarios.

Group 2: Random Few-Shot Prompting

This experimental arm was designed to establish a baseline for a standard, nonoptimized few-shot approach. Its purpose was

XSL•FO
RenderX

to measure the impact of providing generic, in-domain examples without a specific selection strategy. For each of the 200 test cases, the prompt was initially prepared with a fixed set of 5 examples to provide context for the model. These 5 examples were randomly sampled from our expert-authored data pool, excluding the existing evaluation dataset. The sample set used in the prompt for every test case consisted of 4 standard IVF cases and 1 short-protocol IVF case, accompanied by a concise reasoning chain authored by domain experts. The prompt structure and instructions were otherwise identical to those in the other arms. A representative example of a few-shot sample, detailing the input data and expert-written CoT, is provided in Multimedia Appendix 2. This approach represents a "naive" few-shot implementation. It is designed to test the hypothesis that the mere presence of in-domain examples, even without being specifically tailored to the test case, is sufficient to improve reasoning quality compared to the zero-shot baseline.

Group 3: Selective Few-Shot Prompting

This arm represents our proposed method and was designed to test the hypothesis that a deliberately curated set of diverse examples would improve reasoning reliability and generalization. Instead of random sampling, this approach used a clinically informed, representative selection strategy. Physicians curated a set of 6 exemplary cases from a pool of records not included in the 200 evaluation set (to prevent data leakage). These 6 examples were specifically chosen to represent the full spectrum of major ART categories present in our dataset, including IVF (standard IVF, short-protocol IVF, and IVF with donor sperm), ICSI, TESA+ICSI, and PGT (PGT-A). Their reasoning part was carefully crafted and covered all critical steps. The complete prompt is provided in Multimedia Appendix 2. For every test case, this same curated set of 6 diverse examples was prepended to the prompt. The purpose of this strategy was to provide the model with comprehensive and representative clinical guidance within the prompt itself.

In summary, this 3-arm design allows for a multifaceted analysis of CoT reliability. The comparison between group 1 and group 2 will isolate the general benefit of using in-context examples. The critical comparison between group 2 and group 3 will determine whether our proposed selective prompting strategy provides a statistically significant improvement over a random baseline. Together, these comparisons will build a clear evidence-based argument for the importance of a well-designed prompting strategy in generating reliable and accurate clinical reasoning.

Evaluation Metrics

Physician Evaluation

The evaluation was conducted by a panel of 2 board-certified reproductive physicians. Both subject matter experts are faculty members at the same academic medical center but work independently in separate clinical teams. They were invited through an internal clinical research collaboration mechanism, and participation was voluntary. Their sole role was to perform a blinded evaluation of the CoTs. Each evaluator possesses over 10 years of clinical experience in the field of ART. Prior to the formal evaluation, a calibration session was held where all evaluators scored 10 cases together. Any discrepancies were discussed to ensure a consistent understanding of the criteria. During the blinded evaluation, each physician reviewer received only the clinical vignette and the model-generated reasoning content. All identifying information, model names, and prompting strategies were removed to ensure full masking. The order of cases and strategies was independently randomized for each reviewer to prevent any presentation bias. Each physician independently scored the complete dataset without discussing their ratings with other reviewers. No reviewer saw any ground-truth labels or model metadata during the evaluation process.

In this study, we created an evaluation metric involving 3 dimensions: logical coherence and clarity (LCC), use and coverage of key information (UCKI), and plausibility and clinical accuracy of reasoning (PCAR). All generated CoTs were scored by the 5-point Likert scale (1=poor and 5=excellent) across 3 key dimensions of reliability, as shown in Table 2 and detailed in Multimedia Appendix 3. All paired comparison results are tested using the Wilcoxon test and adjusted for false discovery rate (FDR).

Table 2. Rubric for the evaluation of chain-of-thought (CoT) reliability[a].

Metric	Definition
Logical coherence and clarity	Assesses whether the reasoning process is internally consistent, logically structured, and expressed clearly and understandably.
Use and coverage of key information	Evaluates the extent to which the reasoning incorporates and addresses relevant clinical data points presented in the input.
Plausibility and clinical accuracy of reasoning	Measures whether the reasoning is clinically sound, aligns with standard medical knowledge, and leads to a reasonable interpretation or decision. Deduct points as appropriate across the 4 parts in the analysis.

[a]The table defines the 3 dimensions: logical coherence and clarity, use and coverage of key information, and plausibility and clinical accuracy of reasoning, used by both human experts and the artificial intelligence evaluator to assess the quality of generated CoTs on a 5-point Likert scale.

AI Grader Evaluation

In addition to manual evaluation conducted by human experts, we implemented a supplementary evaluation component leveraging a widely used LLM verifier [35], GPT-4o, to explore its feasibility as an automated evaluator of clinical reasoning. This design enables a direct comparison between AI-generated assessments and the human gold standard, thereby evaluating the feasibility of using LLMs for quality control in large-scale synthetic dataset generation. To ensure comparability, the evaluation criteria provided to the AI model were identical to those outlined in Table 2, including definitions of logical

XSL•FO
RenderX

coherence, clinical appropriateness, and key information use. The detailed prompt can be found in Multimedia Appendix 2. Each instruction consists of 2 blocks: the evaluation rubric and case vignette.

Statistical Analysis

All statistical analyses were conducted in Python (pandas, SciPy, and statsmodels; Python Software Foundation). Each case (N=200) was independently evaluated under 3 prompting strategies (zero-shot, random few-shot, and selective few-shot) across 3 dimensions: LCC, UCKI, and PCAR. Results are reported as mean (SD), with exact n values indicated in tables. To assess the reliability of the physician ratings, we evaluated agreement using three complementary measures: (1) linear weighted κ, (2) adjacent agreement rate (percentage of paired scores within ±1 point), and (3) raw interrater disagreement rates. Because the rating distributions exhibited strong ceiling effects (scores concentrated around 4-5), linear weighted κ is known to underestimate agreement under low variance conditions, a well-described statistical paradox. Therefore, in addition to reporting κ values for completeness, we emphasized adjacent agreement and disagreement rates, which more accurately capture clinical consensus when rating scales are narrow. Across all 600 evaluations (200 cases×3 strategies), disagreement rates were uniformly low, confirming strong consistency between the 2 raters. The selective few-shot strategy showed the lowest disagreement (LCC: 0.00, PCAR: 0.03, UCKI: 0.02), followed by the zero-shot (LCC: 0.04, PCAR: 0.04, UCKI: 0.07) and random few-shot strategies (LCC: 0.12, PCAR: 0.09, UCKI: 0.07). The highest disagreement observed across all metrics was only 12%. Adjacent agreement was correspondingly high, ranging from 88% to 100% depending on the metric and strategy, indicating excellent practical concordance despite the compressed scoring range. Given these properties, interrater reliability was interpreted primarily through adjacent agreement and disagreement rates, while κ statistics were retained as a formal but secondary indicator.

Normality of paired differences was assessed using the Shapiro-Wilk test and Q-Q plots. Because the evaluation scores are 5-point Likert ratings, the paired differences take on only a small number of discrete values (–2, –1, 0, +1, +2). As expected with large samples and discrete Likert data, the Shapiro-Wilk was extremely sensitive and returned significant results; however, Q-Q plots showed no meaningful deviations from approximate linearity beyond the expected discreteness. Parametric tests are generally robust to moderate deviations from normality in Likert-type data, especially with larger sample sizes. We retained the nonparametric Wilcoxon paired tests as primary analyses. All the statistical test details can be checked in Multimedia Appendix 4.

To account for multiple pairwise comparisons across the prompting conditions and evaluation metrics in the subgroups, adjusted *P* values were calculated using the Benjamini-Hochberg FDR correction across all 9 contrasts (3 comparisons per metric). For each metric, paired comparisons were conducted using Wilcoxon paired tests, and both the raw and FDR-adjusted *P* values were reported. A post-hoc power analysis was performed. For each evaluation metric, Cohen *d* effect sizes were computed using the pooled SD of paired observations. This analysis was conducted to assess the stability of estimates under small-sample conditions.

Results

Overview

All the results were obtained through the evaluation dataset (N=200), including several kinds of ART. As mentioned earlier, 3 metrics were used for evaluation: LCC, UCKI, and PCAR. The evaluation was done by a panel of experienced practitioners.

General Performance

Table 3 presents the average scores of each prompting strategy on LCC, UCKI, and PCAR. The selective few-shot strategy outperformed both zero-shot and random few-shot approaches across all 3 metrics. Specifically, it achieved mean scores of 4.56 (SD 0.50), 4.66 (SD 0.53), and 4.18 (SD 0.56), which were significantly higher than those of the zero-shot strategy (mean 4.18, SD 0.56; mean 4.30, SD 0.63; mean 3.85, SD 0.53, respectively; all *P* and adjusted *P*<.001; Cohen *d*=0.72, 0.61, 0.61) and the random few-shot strategy (mean 4.31, SD 0.64; mean 4.42, SD 0.58; mean 3.91, SD 0.63, respectively; all *P* and adjusted *P*<.001; Cohen *d*=0.45, 0.42, 0.46).

Table 3. The performance of the "zero-shot," "random few-shot," and "selective few-shot" strategies (N=200 cases per group)[a].

Strategy	LCC[b], mean (SD)	UCKI[c], mean (SD)	PCAR[d], mean (SD)
Zero-shot	4.18 (0.56)	4.30 (0.63)	3.85 (0.53)
Random few-shot	4.31 (0.64)	4.42 (0.58)	3.91 (0.63)
Selective few-shot	4.56 (0.50)	4.66 (0.53)	4.18 (0.56)

[a]Wilcoxon test used for paired comparisons; *P* values adjusted using the Benjamini-Hochberg false discovery rate procedure.

[b]LCC: logical coherence and clarity.

[c]UCKI: use and coverage of key information.

[d]PCAR: plausibility and clinical accuracy of reasoning.

Notably, there was no statistically significant difference between the zero-shot and random few-shot groups on PCAR, though samples did improve the model's capability on LCC and UCKI statistically significantly.

Subgroup Analysis

To further dig into the reasons for selective few-shot's winning, we did an analysis grouped by ART. Table 4 presents the scores of 3 ART generations.

XSL•FO
RenderX

Table 4. Subgroup analysis by assisted reproductive technology (ART) category[a].

ART and strategy	LCC[b]	UCKI[c]	PCAR[d]
IVF[e]			
Zero	4.20 (0.57)	4.34 (0.61)	3.88 (0.53)
Random	4.29 (0.67)	4.44 (0.53)	3.90 (0.65)
Selective	4.59 (0.49)	4.69 (0.51)	4.20 (0.55)
ICSI[f]			
Zero	4.16 (0.59)	4.29 (0.65)	3.87 (0.53)
Random	4.37 (0.54)	4.45 (0.69)	3.97 (0.59)
Selective	4.45 (0.50)	4.53 (0.60)	4.11 (0.56)
PGT[g]			
Zero	4.09 (0.43)	4.05 (0.72)	3.64 (0.49)
Random	4.32 (0.57)	4.27 (0.70)	3.86 (0.56)
Selective	4.59 (0.50)	4.68 (0.48)	4.18 (0.59)

[a]Subgroup analyses were based on the respective case counts (IVF: n=140; ICSI: n=38; PGT: n=22). This table aims to further investigate the performance differences of various prompting strategies across specific clinical scenarios. To achieve this, we categorized the 200 evaluation cases based on their primary type of ART, including IVF, ICSI, and PGT, and conducted a comparative analysis of evaluation outcomes within each group. Wilcoxon test used for paired comparisons; *P* values adjusted using the Benjamini-Hochberg false discovery rate procedure.

[b]LCC: logical coherence and clarity.

[c]UCKI: use and coverage of key information.

[d]PCAR: plausibility and clinical accuracy of reasoning.

[e]IVF: in vitro fertilization.

[f]ICSI: intracytoplasmic sperm injection.

[g]PGT: preimplantation genetic testing.

In the largest subgroup, IVF (n=140), a key distinction emerged. While the selective few-shot strategy significantly outperformed both other groups across all metrics ($P<.001$ and adjusted $P<.001$ for all comparisons, selective vs zero: Cohen $d=0.72$, 0.61, 0.59; selective vs random: Cohen $d=0.51$, 0.49, 0.50), there was no statistically significant difference observed between the random few-shot and zero-shot strategies ($P=.19$, .69, .10; adjusted $P=.22$, .69, .13).

The analysis of the PGT subgroup (n=22) revealed the clearest advantage for prompt diversity. The selective few-shot strategy, which was the only prompt containing a PGT example, outperformed the random few-shot strategy across all 3 metrics: logical coherence (LCC: $P=.03$; adjusted $P=.05$; Cohen $d=0.51$), information use (UCKI: $P<.001$; adjusted $P=.01$; Cohen $d=0.55$), and clinical accuracy (PCAR: $P=.03$; adjusted $P=.05$; Cohen $d=0.68$). Consistent with other findings, the random few-shot strategy showed no significant improvement over the zero-shot baseline in this category (LCC: $P=.17$; adjusted $P=.19$; Cohen $d=0.45$; UCKI: $P=.27$; adjusted $P=.27$; Cohen $d=0.32$; PCAR: $P=.13$; adjusted $P=.17$; Cohen $d=0.43$, respectively). However, we must admit that the limited sample size makes the subgroup analysis partially underpowered, and these comparisons should be interpreted with caution.

A similar pattern emerged in the ICSI subgroup (n=38). The selective few-shot strategy again demonstrated a measurable advantage. It achieved statistically significant improvements over the zero-shot baseline in 2 of the 3 key metrics: LCC ($P=.02$; adjusted $P=.15$; Cohen $d=0.53$) and PCAR ($P=.04$; adjusted $P=.17$; Cohen $d=0.44$). Although these comparisons did not remain significant after FDR adjustment, the effect sizes were moderate, and the directionality was consistent with the overall findings. For UCKI, the selective strategy again achieved the highest mean score, but this comparison did not reach significance ($P=.06$; adjusted $P=.17$), suggesting a positive but statistically inconclusive trend. The detailed subgroup statistical results can be checked in Multimedia Appendix 4.

Case Study

As shown in Figure 2, to qualitatively illustrate the stark differences in reasoning quality revealed by our quantitative analysis, we selected a representative and complex case involving PGT-M. This case is particularly illustrative, as it requires a multilayered understanding of genetics, ART procedures, and individualized patient factors. The main mistakes are listed in Table 5.

XSL•FO
RenderX

Figure 2. Representative PGT-M case illustrating qualitative differences in CoT reasoning across prompting strategies. This figure presents a representative and complex case involving PGT-M, selected to qualitatively illustrate the differences in reasoning quality observed in our quantitative analyses. The left panel shows the patient's clinical information, the correct physician's answer, and the color-coded annotation scheme (red: incorrect reasoning, yellow: irrelevant reasoning, and green: correct reasoning). The right panel displays the CoT outputs generated under zero-shot, random few-shot, and selective few-shot prompting strategies. Compared with the zero-shot and random few-shot generations, which omitted critical reasoning steps (eg, the presence of infertility diagnosis, the indication for intracytoplasmic sperm injection, and comprehensive gonadotropin dose considerations), the selective few-shot prompting was more closely aligned with clinical logic and included relevant patient-specific factors. CoT: chain-of-thought; PGT-M: preimplantation genetic testing for monogenic disorder.

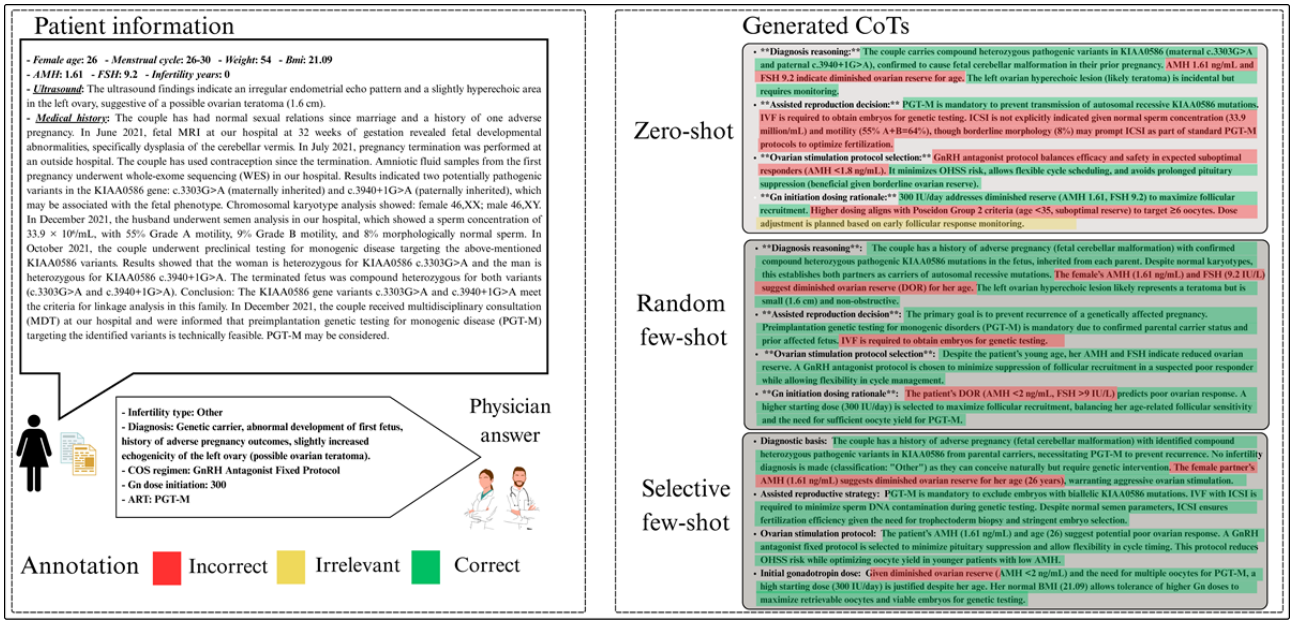

Table 5. Common reasoning errors in zero-shot and random few-shot chain-of-thought (CoT) outputs for a preimplantation genetic testing for monogenic disorder case.

Reasoning dimension	Flaws in zero-shot and random few-shot
Diagnosis reasoning	The model does not mention whether the patient has infertility issues.
Assisted reproduction decision	CoT incorrectly assumes that if the male's semen is normal, traditional IVF[a] can be used.
Ovarian stimulation protocol selection	The reason for choosing the antagonist protocol in CoT was "greater safety and avoidance of OHSS[b]," without considering the patient's specific circumstances (low AMH, first ovulation induction).
Initial gonadotropin dose	Only AMH[c] levels were considered, without taking into account weight, BMI, or PGT[d] goals (requiring more embryos).

[a]IVF: in vitro fertilization.

[b]OHSS: ovarian hyperstimulation syndrome.

[c]AMH: anti-Müllerian hormone.

[d]PGT: preimplantation genetic testing.

In this PGT-M case, both partners are carriers of a pathogenic variant in the KIAA0586 gene. During a previous pregnancy, the fetus was found to have a homozygous mutation in KIAA0586, resulting in abnormal brain development and subsequent pregnancy termination. Since then, the couple has been using contraception and therefore does not meet the criteria for an infertility diagnosis. This implies that they are still capable of conceiving naturally. Given the autosomal recessive inheritance pattern, there remains a possibility of achieving a normal or carrier embryo through natural conception. However, neither the zero-shot nor the random few-shot–prompted CoT generations mentioned the presence or absence of an infertility diagnosis, which appeared in the selective few-shot–prompted CoT.

To avoid the recurrence of a fetus with a homozygous mutation in KIAA0586, PGT-M is recommended. Due to the technical requirements of PGT, embryos must be obtained via ICSI to avoid DNA contamination during genetic analysis. While the zero-shot and random few-shot–prompted CoTs correctly reasoned the indication for PGT-M, they incorrectly concluded that ICSI was unnecessary because the male partner had normal semen parameters and suggested using conventional IVF instead—an error in clinical reasoning.

In selecting the ovarian stimulation protocol, clinical reasoning typically begins with evaluating the patient's ovarian responsiveness and any prior stimulation history. Although the patient is 26 years of age, her AMH level is only 1.61 ng/mL, suggesting a potential for diminished ovarian response. As this is her first controlled ovarian hyperstimulation cycle, a

XSL•FO
RenderX

gonadotropin-releasing hormone antagonist protocol was chosen for its controllability and to avoid excessive pituitary suppression. Among the 2 few-shot–prompted CoTs, the reasoning was more aligned with clinical thinking, while the zero-shot CoT emphasized the safety profile of the antagonist protocol (eg, avoiding ovarian hyperstimulation syndrome) without clearly reflecting clinical logic.

Regarding the initial gonadotropin dose, factors beyond ovarian responsiveness must be considered. Since this case involves PGT, it is important to optimize the number of oocytes retrieved. Additional considerations include the patient's weight and BMI, as these affect drug sensitivity. However, the zero-shot CoT mentioned only ovarian responsiveness, lacking a comprehensive rationale.

Feasibility Analysis of an AI Evaluator

As detailed in Table 6, the mean scores for all 3 prompting strategies were tightly clustered in a narrow and high-scoring range, between 3.96 and 4.00, suggesting that the model perceived all generated outputs as being of similarly high quality.

Table 6. Artificial intelligence (AI)–driven evaluation of chain-of-thought reliability across different prompting strategies[a].

Group	LCC[b], mean (SD)	PCAR[c], mean (SD)	UCKI[d], mean (SD)
Random few-shot	4.00 (0.00)	3.98 (0.14)	4.00 (0.00)
Selective few-shot	4.00 (0.00)	3.98 (0.16)	4.00 (0.07)
Zero-shot	4.00 (0.07)	3.96 (0.20)	3.98 (0.14)

[a]The high scores and minimal variation across all groups indicate a significant ceiling effect in the AI's evaluation. Paired comparisons between strategies were conducted using the Wilcoxon signed rank test; *P* values were adjusted using the Benjamini-Hochberg false discovery rate procedure.

[b]LCC: logical coherence and clarity.

[c]UCKI: use and coverage of key information.

[d]PCAR: plausibility and clinical accuracy of reasoning.

Inferential statistical analysis corroborated this observation. A series of Friedman tests found no statistically significant differences among the 3 groups for LCC (P=.37), PCAR (P=.37), or UCKI (P=.07). While a post-hoc pairwise Wilcoxon paired test identified a marginal statistical difference between the random few-shot and zero-shot groups on the information use dimension (P=.045), this isolated finding merits cautious interpretation, particularly, as the overall test for this dimension did not reach statistical significance.

Discussion

Principal Findings

This study critically evaluates the reliability of LLM-generated CoT reasoning in ART and shows that noncurated prompting methods are insufficient for clinical use. Both zero-shot and random few-shot strategies frequently produced reasoning errors, and random shallow examples offered no meaningful improvement over providing no examples at all. In contrast, the selective few-shot strategy, which is built on the principles of representative diversity and gold-standard depth, substantially improved coherence, information use, and clinical accuracy. These reliability gaps, as well as the strengths of the selective approach, were identifiable only through expert review; automated AI evaluators failed to detect these differences. Together, these findings outline a practical framework for evaluating ART reasoning quality and a feasible pathway for generating trustworthy synthetic clinical data.

The principle of representative diversity was clearly demonstrated in the PGT and ICSI subgroups. The findings provide empirical support for our initial hypothesis. The PGT category shows significantly higher scores, prompted by the selective few-shot approach, which includes an example of PGT-A treatment. The case study also shows errors in understanding and judgment in doctors' viewing, where zero-shot or random few-shot are more likely to make intrinsic mistakes. Notably, in the ICSI category, although the intergroup differences did not reach statistical significance when compared to the random few-shot group, we observed the same trend as in the PGT category—selective prompting consistently achieved the highest average scores and was significantly higher than zero-shot prompting, which had no difference with the random one. The analyses of both subgroups collectively suggest that a demonstration set covering key procedural subtypes within the domain is essential for enabling the model to evolve from a "specialist" to a "generalist."

Simultaneously, the principle of gold-standard depth was illustrated in the IVF subgroup. In our main results, we show that the quality of examples may influence the quality of generation. In subgroup analysis, we found that there is no significant difference between the zero-shot prompting and the random few-shot prompting on any subgroup, especially in the IVF subgroup, even if the random arm's sample cases indeed included 4 standard IVF and 1 short-protocol IVF. It performed ineffective learning under this situation. In this case, the reason may be attributed to the reasoning quality in the prompt. In the experiment design section, we mentioned that the random cases have a relatively concise CoT. This indicates that the LLM exhibits a strong tendency toward pattern imitation when engaging in in-context learning. A low-quality example tends to elicit correspondingly poor reasoning outputs, even if the model has huge potential in text generation. Therefore, this principle emphasizes that each few-shot example must serve as an expert-level exemplar: logically rigorous, richly detailed, and representative of ART strategy reasoning at the highest standard.

XSL•FO
RenderX

Comparison to Prior Work

Our findings align closely with a well-established principle in the broader AI research community: data quality often outweighs data quantity [36]. Our work provides domain-specific empirical support for the application of this principle in the reproductive medicine context of clinical CoT generation. More importantly, we go beyond simply affirming the importance of data reliability; we offer a concrete characterization of what high-quality examples mean in this setting, through our proposed dual principles of gold-standard depth and representative diversity. Together, these insights contribute a practical methodology for realizing data-centric AI specializing in reproductive medicine.

Another key finding highlights a critical limitation of current LLM-based evaluators in detecting subtle yet clinically meaningful variations in information use, logical rigor, and contextual accuracy. While our human expert assessments revealed substantial differences in reasoning quality across the 3 prompting strategies, the scores assigned by the AI evaluator (GPT-4o) showed no statistically significant differences between them across 3 metrics. This "ceiling effect" serves as a critical warning: in high-stakes medical applications, like ART strategy choosing, where patient safety is on the line, relying solely on automated evaluation for quality assurance is inherently risky. It reaffirms that domain expert oversight is not merely a "gold standard" for evaluation; it is an essential safeguard that cannot be replaced. Our results show that AI-based evaluation cannot be treated as a source of ground truth; all judgments involving factual accuracy, clinical appropriateness, or safety must rely on human experts. From a broader methodological perspective, the results underscore a growing challenge for the field. As the development of medical LLMs increasingly depends on large-scale synthetic data, evaluation may become the primary bottleneck. While models continue to improve in producing fluent clinical narratives, reliably detecting subtle but clinically meaningful reasoning errors remains far more difficult. Without dependable evaluators, synthetic or augmented clinical data cannot safely be incorporated into model training pipelines. Addressing this gap will require medically grounded evaluation frameworks, including domain-specific supervision signals, error-aware reward models, and structured representations of clinical logic. These capabilities are not yet captured by current general-purpose LLM judges, emphasizing the need for future research focused on building evaluators that meet the safety, sensitivity, and domain expertise required for clinical AI applications.

This study provides evidence within a single-center ART dataset, and further multicenter generalization is needed. Although we attempted to determine the reliability of AI-generated CoT in complex clinical cases, our cases are currently limited to reproductive medicine or ART treatment. To enhance generalizability and robustness, future research should include a more diverse set of complex clinical reasoning cases across different medical departments. This study has several limitations. First, all generations were produced using a single model (DeepSeek-R1), which restricts the external validity of the findings. Future studies will evaluate whether the advantages of the selective few-shot strategy generalize across different LLM families. Second, the use of temperature=0.5 introduces controlled stochasticity into the inference process; other decoding settings may produce output variations. To address this, future work will include sensitivity analyses across multiple temperature levels (eg, 0, 0.2, and 0.5) to assess the stability of reasoning patterns. In addition, all human evaluators were recruited from the same medical center, which may introduce institutional bias due to shared training backgrounds and practice standards. The AI-grader feasibility test also has limitations: the grader's sensitivity is partly dependent on its prompt design and model chosen, which may reduce its ability to detect subtle but clinically important differences within the reasoning block. Finally, the evaluation was conducted at the case level, and although the Likert-based rubric captures overall reasoning quality, subjective variability cannot be fully eliminated. Future work may incorporate sentence-level or error-type–specific analysis to support more objective and fine-grained identification of reasoning deficiencies. Given these constraints, this study should be interpreted as a vertical, domain-specific proof-of-concept, rather than a horizontal benchmark applicable across clinical specialties or model families. The selective few-shot strategy was examined within ART because it provides a well-defined and clinically coherent setting for studying structured reasoning, not because its performance should be assumed to generalize elsewhere. Whether the observed improvements reflect a domain-specific phenomenon or a more general pattern cannot be determined from this study. Future work will therefore focus on rigorously evaluating the approach across diverse clinical domains, datasets, and model families to assess its true generalizability. Beyond the constraints and limitations, an important consideration of this study is that the selective few-shot condition differed from the random condition not only in the conceptual selection principles but also in exemplar characteristics, including number, clinical depth, and ART subtype diversity, which creates a mixed signal. Although the numerical difference between 5-shot and 6-shot prompting is small, it may nonetheless introduce bias. More importantly, exemplar depth and subtype diversity were intentionally incorporated to construct a clinically coherent selective prompt, but these factors inherently covary in our current design. As a result, this study cannot attribute the observed improvements to any single component of the selective strategy nor determine whether the effect arises from exemplar depth, diversity, their interaction, or other uncontrolled influences. The findings should therefore be interpreted as exploratory and hypothesis-generating, rather than evidence of a validated mechanism. To address this joint pattern, future work will implement controlled ablation studies that (1) equalize exemplar number across conditions and (2) independently manipulate exemplar depth ("deep vs shallow") and subtype diversity ("diverse vs homogeneous"). Such studies will allow rigorous assessment of the independent and combined contributions of these factors to few-shot reasoning performance in clinical LLMs.

Our dataset contains 200 diverse cases, but for some subtypes, the number of cases may be too small for statistical analysis, particularly the PGT subgroup, which only included 22 samples. Although the post-hoc power calculation and *P* value correction were conducted, it still showed moderate effect sizes on part of

XSL•FO
RenderX

the comparisons. Accordingly, the subgroup findings should be interpreted as exploratory. A potential methodological improvement for future studies is the use of hierarchical partial-pooling or Bayesian shrinkage models, which may borrow strength across subgroups and produce more stable estimates under low-sample conditions. These models were not adopted in this study because our primary objective was descriptive comparison rather than multilevel estimation, but they represent a promising direction for future research. For future directions, given the limited context window of current LLMs, users may face an inherent trade-off when selecting few-shot exemplars, particularly in domains such as reproductive medicine where clinical presentations exhibit substantial subtype diversity. Balancing breadth and depth in exemplar selection becomes a critical challenge under these constraints. Recent work on dynamic prompting methods has sought to improve the performance-efficiency trade-off in resource-limited or accuracy-constrained settings [37], and incorporating such techniques may further enhance the practicality of selective prompting in clinical applications. In addition, future work will explore retrieval-augmented generation frameworks. Integrating authoritative domain sources (eg, American Society for Reproductive Medicine and European Society of Human Reproduction and Embryology guidelines) has the potential to improve factual grounding, reduce hallucination, and enhance explainability in ART-related clinical reasoning. Comparing closed-book reasoning with retrieval augmented generation augmented reasoning may clarify how access to external evidence shapes LLM decision-making and may improve the reliability of LLM-assisted clinical decision support tools.

Conclusions

The primary contribution of this study is 2-fold: an exploratory potential evaluation framework for how to evaluate and provide a methodology for a feasible approach and for how to generate trustworthy clinical ART reasoning steps in a single clinic center. First, we investigate a rigorous, domain-grounded framework for evaluating synthetic clinical reasoning within the ART strategy. Amid the rapid growth of AI in health care, we demonstrate that ensuring clinical validity requires moving beyond automated metrics. Our findings expose the critical limitations of SOTA AI evaluators (eg, GPT-4o) in detecting subtle but clinically vital reasoning flaws. This “ceiling effect” serves as a critical warning and highlights the indispensable role of structured, blind expert review as an essential safeguard in reproductive medicine AI development. Second, building on this evaluation framework, we offer a practical solution to the “explainability data bottleneck” in reproductive medicine. Through systematic comparisons, we show that a selective few-shot prompting strategy, which is based on the “dual principles” of gold-standard depth and representative diversity, substantially improves the quality and reliability of generated ART CoTs. This offers a feasible, cost-effective blueprint for generating trustworthy ART synthetic data at scale, without requiring immense annotated datasets. Finally, this study evaluates the clinical reliability of LLM-generated reasoning in the ART context as a step toward addressing data scarcity in explainable, domain-specialized AI development. However, our findings should not be interpreted as evidence that current LLMs are clinically safe or ready for autonomous use. Our evaluation focuses on reasoning quality, not deployment readiness. Establishing clinical go or no-go thresholds will require task-specific, prospective validation studies assessing safety, consistency, patient outcomes, and workflow integration—factors beyond the scope of this work.

Acknowledgments

The authors used the generative artificial intelligence (AI) tool to examine the potential use of LLMs in assisted reproductive technology. The authors declare the use of generative AI in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to generative AI tools under full human supervision: feasibility assessment and risk evaluation. The generative AI tools used were DeepSeek-R1-671B and ChatGPT4o. Responsibility for the final manuscript lies entirely with the authors. Generative AI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This study was funded by the Science and Technology Department of Sichuan Province Project (2024YFFK0365), the Natural Science Foundation of Sichuan, China (2025NSFSC1985), and the 1·3·5 project for disciplines of excellence, West China Hospital, Sichuan University (ZYYC21004).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: RY, TT, SZ, Di Liu, KL
Data curation: RY, TT, Dou Liu, Y Long
Formal analysis: RY, TT, Dou Liu, Y Long, SZ, Di Liu
Funding acquisition: KL
Investigation: Dou Liu, Y Long

XSL•FO
RenderX

Methodology: Dou Liu, Y Long
Project administration: RY, TT, KL
Visualization: Dou Liu, Y Long, Y Lin, HL
Writing—original draft: Dou Liu, Y Long
Writing—review and editing: All authors

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed description of variables used in the dataset.
[DOCX File , 19 KB - jmir_v28i1e85206_app1.docx]

Multimedia Appendix 2
Chain-of-thought generation prompt.
[DOCX File , 45 KB - jmir_v28i1e85206_app2.docx]

Multimedia Appendix 3
Detailed human evaluation rubric.
[DOCX File , 19 KB - jmir_v28i1e85206_app3.docx]

Multimedia Appendix 4
Statistical results.
[DOCX File , 23 KB - jmir_v28i1e85206_app4.docx]

References

1. Graham ME, Jelin A, Hoon AH, Wilms Floet AM, Levey E, Graham EM. Assisted reproductive technology: short- and long-term outcomes. Dev Med Child Neurol 2023;65(1):38-49 [FREE Full text] [doi: 10.1111/dmcn.15332] [Medline: 35851656]
2. Asplund K. Use of fertilization—ethical issues. Ups J Med Sci 2020;125(2):192-199 [FREE Full text] [doi: 10.1080/03009734.2019.1684405] [Medline: 31686575]
3. Liu D, Long Y, Zuoqiu S, Tang T, Yin R. Evaluating the feasibility and accuracy of large language models for medical history-taking in obstetrics and gynecology. ArXiv. Preprint posted online on March 31, 2025 2025. [doi: 10.48550/arXiv.2504.00061]
4. Huang L, Hu J, Cai Q, Fu G, Bai Z, Liu Y, et al. The performance evaluation of artificial intelligence ERNIE bot in Chinese National Medical Licensing Examination. Postgrad Med J 2024;100(1190):952-953. [doi: 10.1093/postmj/qgae062] [Medline: 38813794]
5. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, et al. Towards accurate differential diagnosis with large language models. Nature 2025;642(8067):451-457 [FREE Full text] [doi: 10.1038/s41586-025-08869-4] [Medline: 40205049]
6. Liu D, Han Y, Wang X, Tan X, Liu D, Qian G, et al. Evaluating the application of ChatGPT in outpatient triage guidance: a comparative study. In: Congress of the International Ergonomics Association. Singapore: Springer Nature Singapore; Aug 25, 2024:233-238.
7. Buckley T, Diao J, Rajpurkar P, Rodman A, Manrai A. Multimodal foundation models exploit text to make medical image predictions. ArXiv Preprint posted online on November 25, 2024 [FREE Full text]
8. Luo L, Ning J, Zhao Y, Wang Z, Ding Z, Chen P, et al. Taiyi: a bilingual fine-tuned large language model for diverse biomedical tasks. J Am Med Inform Assoc 2024;31(9):1865-1874. [doi: 10.1093/jamia/ocae037] [Medline: 38422367]
9. Christophe C, Kanithi P, Munjal P, Raha T, Hayat N, Rajan R, et al. Med42—evaluating fine-tuning strategies for medical LLMs: full-parameter vs. parameter-efficient approaches. ArXiv Preprint posted online on April 23, 2024 [FREE Full text] [doi: 10.18653/v1/2025.emnlp-main.1174]
10. Chen Z, Cano A, Romanou A, Bonnet A, Matoba K, Salvi F, et al. MEDITRON-70B: scaling medical pretraining for large language models. ArXiv Preprint posted online on November 27, 2023 [FREE Full text] [doi: 10.21203/rs.3.rs-4139743/v1]
11. Miao J, Thongprayoon C, Suppadungsuk S, Krisanapan P, Radhakrishnan Y, Cheungpasitporn W. Chain of thought utilization in large language models and application in nephrology. Medicina (Kaunas) 2024;60(1):148 [FREE Full text] [doi: 10.3390/medicina60010148] [Medline: 38256408]
12. Kim S, Joo S, Kim D, Jang J, Ye S, Shin J, et al. The CoT collection: improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. 2023 Presented at: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; December 17, 2025; Singapore p. 12685-12708. [doi: 10.18653/v1/2023.emnlp-main.782]

XSL•FO
RenderX

13. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. Nat Med 2025;31(3):943-950. [doi: 10.1038/s41591-024-03423-7] [Medline: 39779926]
14. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Health 2023;5(4):e179-e181 [FREE Full text] [doi: 10.1016/S2589-7500(23)00048-1] [Medline: 36894409]
15. Rauschecker AM, Rudie JD, Xie L, Wang J, Duong MT, Botzolakis EJ, et al. Artificial intelligence system approaching neuroradiologist-level differential diagnosis accuracy at brain MRI. Radiology 2020;295(3):626-637 [FREE Full text] [doi: 10.1148/radiol.2020190283] [Medline: 32255417]
16. Liu Y, Jain A, Eng C, Way DH, Lee K, Bui P, et al. A deep learning system for differential diagnosis of skin diseases. Nat Med 2020;26(6):900-908. [doi: 10.1038/s41591-020-0842-3] [Medline: 32424212]
17. Babbar R, Schölkopf B. Data scarcity, robustness and extreme multi-label classification. Mach Learn 2019;108(8-9):1329-1351. [doi: 10.1007/s10994-019-05791-5]
18. Villalobos P, Ho A, Sevilla J, Besiroglu T, Heim L, Hobbhahn M. Position: Will we run out of data? Limits of LLM scaling based on human-generated data. 2024 Presented at: Forty-First International Conference on Machine Learning; July 21-27, 2024; Vienna, Austria p. 49523-49544. [doi: 10.5555/3692070.3694094]
19. Dahmen J, Cook D. SynSys: a synthetic data generation system for healthcare applications. Sensors (Basel) 2019;19(5):1181. [doi: 10.3390/s19051181] [Medline: 30857130]
20. El EK, Mosquera L, Hoptroff R. Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data. California: O'Reilly Media; 2025.
21. Li Z, Zhu H, Lu Z, Yin M. Synthetic data generation with large language models for text classification: potential and limitations. ArXiv Preprint posted online on October 13, 2023 [FREE Full text] [doi: 10.48550/arXiv.2310.07849]
22. Williams CYK, Subramanian CR, Ali SS, Apolinario M, Askin E, Barish P, et al. Physician- and large language model-generated hospital discharge summaries. JAMA Intern Med 2025;185(7):818-825. [doi: 10.1001/jamainternmed.2025.0821] [Medline: 40323616]
23. Vardhan M, Nathani D, Vardhan S, Aggarwal A, Simini F. Large language models as synthetic electronic health record data generators. 2024 Presented at: IEEE Conference on Artificial Intelligence (CAI); June 25-27, 2024; Singapore p. 804-810. [doi: 10.1109/cai59869.2024.00152]
24. Veselovsky V, Ribeiro M, Arora A, Josifoski M, Anderson A, West R. Generating faithful synthetic data with large language models: a case study in computational social science. ArXiv Preprint posted online on May 24, 2023. [doi: 10.48550/arXiv.2305.15041]
25. Ziems C, Held W, Shaikh O, Chen J, Zhang Z, Yang D. Can large language models transform computational social science? Comput Linguist 2024;50(1):237-291. [doi: 10.1162/coli_a_00502]
26. Gilardi F, Alizadeh M, Kubli M. ChatGPT outperforms crowd workers for text-annotation tasks. Proc Natl Acad Sci U S A 2023;120(30):e2305016120. [doi: 10.1073/pnas.2305016120] [Medline: 37463210]
27. Wang J, Liang Y, Meng F, Sun Z, Shi H, Li Z, et al. Is ChatGPT a good NLG evaluator? A preliminary study. 2023 Presented at: Proceedings of the 4th New Frontiers in Summarization Workshop; December 17, 2025; Singapore p. 1-11. [doi: 10.18653/v1/2023.newsum-1.1]
28. Li H, Dong Q, Tang Z, Wang C, Zhang X, Huang H, et al. Synthetic data (almost) from scratch: generalized instruction tuning for language models. ArXiv Preprint posted online on February 20, 2024. [doi: 10.48550/arXiv.2402.13064]
29. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. 2022 Presented at: NIPS'22: 36th International Conference on Neural Information Processing Systems; November 28-December 9, 2022; New Orleans, LA, United States p. 24824-24837 URL: https://dl.acm.org/doi/10.5555/3600270.3602070 [doi: 10.5555/3600270.3602070]
30. Sprague Z, Yin F, Rodriguez J, Jiang D, Wadhwa M, Singhal P, et al. To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. ArXiv Preprint posted online on May 7, 2025. [doi: 10.48550/arXiv.2409.12183]
31. Yuan X, Shen C, Yan S, Zhang X, Xie L, Wang W, et al. Instance-adaptive zero-shot chain-of-thought prompting. 2024 Presented at: Advances in Neural Information Processing Systems 37 (NeurIPS 2024); December 10, 2024; Vancouver, British Columbia, Canada. [doi: 10.52202/079017-3986]
32. Shi E, Manda A, Chowdhury L, Arun R, Zhu K, Lam M. Enhancing depression diagnosis with chain-of-thought prompting. ArXiv Preprint posted online on August 27, 2024. [doi: 10.48550/arXiv.2408.14053]
33. Liu R, Wei J, Liu F, Si C, Zhang Y, Rao J, et al. Best practices and lessons learned on synthetic data. ArXiv Preprint posted online on August 10, 2024. [doi: 10.48550/arXiv.2404.07503]
34. Guo D, Yang D, Zhang H, Song J, Wang P, Zhu Q, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature 2025;645(8081):633-638. [doi: 10.1038/s41586-025-09422-z] [Medline: 40962978]
35. Team L, Xu W, Chan H, Li L, Aljunied M, Yuan R, et al. Lingshu: a generalist foundation model for unified multimodal medical understanding and reasoning. ArXiv Preprint posted online on June 13, 2025. [doi: 10.48550/arXiv.2506.07044]
36. Zha D, Bhat ZP, Lai K, Yang F, Jiang Z, Zhong S, et al. Data-centric artificial intelligence: a survey. ACM Comput Surv 2025;57(5):1-42. [doi: 10.1145/3711118]
37. Zhou W, Jiang YE, Cotterell R, Sachan M. Efficient prompting via dynamic in-context learning. ArXiv Preprint posted online on May 18, 2023. [doi: 10.48550/arXiv.2305.11170]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
ART: assisted reproductive technology
CoT: chain-of-thought
EHR: electronic health record
FDR: false discovery rate
HIPAA: Health Insurance Portability and Accountability Act
ICSI: intracytoplasmic sperm injection
IVF: in vitro fertilization
LCC: logical coherence and clarity
LLM: large language model
PCAR: plausibility and clinical accuracy of reasoning
PGT: preimplantation genetic testing
PGT-A: preimplantation genetic testing for aneuploidies
PGT-M: preimplantation genetic testing for monogenic disorder
SOTA: state-of-the-art
TESA: testicular sperm aspiration
UCKI: use and coverage of key information

Edited by J Sarvestan; submitted 02.Oct.2025; peer-reviewed by X Zhong, KH Lin; comments to author 03.Nov.2025; revised version received 14.Dec.2025; accepted 15.Dec.2025; published 08.Jan.2026.

Please cite as:
Liu D, Long Y, Zuoqiu S, Liu D, Li K, Lin Y, Liu H, Yin R, Tang T
Reliability of Large Language Model Generated Clinical Reasoning in Assisted Reproductive Technology: Blinded Comparative Evaluation Study
J Med Internet Res 2026;28:e85206
URL: https://www.jmir.org/2026/1/e85206
doi:10.2196/85206
PMID:

©Dou Liu, Ying Long, Sophia Zuoqiu, Di Liu, Kang Li, Yiting Lin, Hanyi Liu, Rong Yin, Tian Tang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Large Language Model–Generated Patient Instructions for Prescriptions in Primary Health Care: Preclinical Algorithm Validation

Zilma Silveira Nogueira Reis[1*], MD, PhD; Elisa Tuler Albergaria[2*], CS, PhD; Adriana Silvina Pagano[3*], AS, PhD; Eura Martins Lage[4*], MD, PhD; Flávia Ribeiro de Oliveira[4*], MD, PhD; Cristiane dos Santos Dias[4*], MD, PhD; Juliana Almeida Oliveira[4*], MD; Gláucia Miranda Varella Pereira[4*], PhD; Isaias Jose Ramos de Oliveira[4*], MSCS; Érico Franco Mineiro[5*], PhD; Igor Carvalho Lima Oliveira[5*], MDS; Davi dos Reis de Jesus[2*], CS; Antônio Pereira de Souza Júnior[2*], MSCS; Igor de Carvalho Gomes[6*], MSc; Rodrigo André Cuevas Gaete[6*], CS, PhD; Ricardo Cruz-Correia[7*], CS, PhD; Leonardo Rocha[2*], CS, PhD

[1]Health Informatics Center, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[2]Computer Science Department, Universidade Federal de São João Del Rei, São João Del Rei, Brazil

[3]Arts Faculty, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[4]Department of Physiotherapy, School of Physical Education, Physiotherapy and Occupational Therapy, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[5]School of Architecture, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[6]National Secretary of Primary Care of the Brazilian Ministry of Health, Brasilia, Brazil

[7]MEDCIDS, Porto University, Porto, Portugal

[*]all authors contributed equally

Corresponding Author:
Zilma Silveira Nogueira Reis, MD, PhD
Health Informatics Center
Universidade Federal de Minas Gerais
Avenida Professor Alfredo Balena, 190
Belo Horizonte, 30130100
Brazil
Phone: 55 31985177473
Email: zilma.medicina@gmail.com

Abstract

Background: The application of generative artificial intelligence to simplify medication use instructions has the potential to enhance people's health by improving treatment adherence.

Objective: We evaluated the performance of large language models (LLMs) in generating medication usage instructions to complement prescriptions in primary health care.

Methods: This randomized, blinded experimental preclinical study used prescription-inducing scenarios, assigned to 62 health care professionals, to validate instructions generated by LLMs during electronic prescriptions. The instructions were generated by ChatGPT-4.0 (OpenAI), Llama3.1-8B (Meta), and Llama3.1-8B-RAG (Meta) using retrieval-augmented generation based on patient information leaflets. Performance metrics assessed adequacy, completeness, clarity, language simplification, usefulness, and errors in the generated instructions, with scores to analyze overall and individual metrics.

Results: The 3 models yielded high overall scores for producing qualified instructions (ChatGPT-4.0: median 88.4, IQR 22.8; Llama3.1-8B: median 66.5, IQR 50.9; Llama3.1-8B-RAG: median 79.9, IQR 34.4; Kruskal-Wallis test P=.003). Llama3.1-8B-RAG received evaluations with similar overall scores to ChatGPT-4.0 (post hoc test, P=.05) and similar to Llama3.1-8B (post hoc test, P=.44). ChatGPT-4.0 outperformed Llama3.1-8B (Bonferroni test, P<.001). Regarding specific domains, Llama3.1-8B-RAG received scores equivalent to those of ChatGPT-4.0 for adequacy (mean 6.24, SD 2.3 vs mean 6.82, SD 2.1; post hoc test, P=.54); completeness (mean 5.94, SD 2.2 vs 6.55, SD 1.9; post hoc test P=.38), clarity (mean 5.77, SD 2.4 vs mean 6.68, SD 1.9; post hoc test P=.09), and usefulness (mean 5.42, SD 2.4 vs mean 5.96, SD 2.2; post hoc test P=.63). ChatGPT-4.0 received higher

XSL•FO
RenderX

scores in the language simplification criterion than Llama3.1-8B-RAG (mean 7.05, SD 1.5 vs mean 5.44, SD 2.6; post hoc test *P*<.001). Interrater variability in assigning scores ranged from 4.2% (n=3) to 85.8% (n=6) among primary health care professionals. Instructions leading to incorrect use of the medication had similar frequency among the models(ChatGPT-4.0: n=15, 22.7%; Llama3.1-8B: n=19, 22.8%; Llama3.1-8B-RAG: n=19, 22.8%; chi-square test *P*=.71). The frequencies of hallucination were similar (ChatGPT-4.0: n=7, 10.6%; Llama3.1-8B: n=9, 13.6%; Llama3.1-8B-RAG: n=6, 9.1%; chi-square test *P*=.67).

Conclusions: The open-source LLM enhanced with external information presented similar performance to the closed-source model, except for ChatGPT4.0, which was superior in language simplification of messages. LLM generation demonstrated potential for instructing patients on medication use. Nonetheless, the introduction of this innovation into the electronic prescribing workflow demands prescriber validation for human oversight of the technology and requires a strategy for LLM performance governance.

International Registered Report Identifier (IRRID): RR2-10.12688/verixiv.1359.1

(J Med Internet Res 2026;28:e84444) doi:10.2196/84444

KEYWORDS

primary health care; drug prescriptions; large language models; generative artificial intelligence; digital health

Introduction

The association between medication nonadherence and a higher prevalence of diseases demonstrates that not using, or incorrectly using, prescribed medications represents a burden for both patients and health care systems [1]. In contrast, investing in processes that can promote treatment adherence is a strategy capable of mitigating excessive and avoidable costs incurred by misuse, thereby promoting the health of individuals, especially those with chronic conditions. The underuse of prescribed medications, whether intentional or unintentional, is considered a major global public health problem. One study examined 195,930 electronic prescriptions in Massachusetts, United States, to analyze the drug classes most associated with primary patient nonadherence immediately after the first prescription, with findings for medications for diabetes (31.4%), arterial hypertension (28.4%), and hyperlipidemia (28.2%) [2]. In low- and middle-income countries, the estimated occurrence of nonadherence is even higher, potentially reaching 50% of prescriptions [3]. Health care scenarios with limited resources also face significant challenges in controlling noncommunicable chronic diseases, mental disorders, tuberculosis, and HIV infection, further emphasizing the importance of providing measures that favor treatment adherence [3].

Technological interventions facilitated by digital technology have the potential to expand the coverage and quality of health services [4]. In primary health care (PHC), which serves as the entry point for individuals into the health systems, digital transformation involves introducing useful, widely applicable, and cost-effective technological resources. It is also essential that these systems are developed to meet the needs of individuals and communities [5]. One successful example of digital systems in health care is electronic prescribing (e-prescribing). The incorporation of a standardized medication menu with informed dosages, maximum and minimum limits, administration routes, and alerts for drug interactions, in addition to mitigating readability issues in prescriptions, proved useful by reducing preventable medication errors by 55% [6]. In PHC settings, the possibility of direct electronic communication between the prescribing professional and the dispensing professional is also a point in favor of benefits such as increased safety and quality of care provided [7].

Despite advances in e-prescribing, medication adherence remains a barrier to effective treatment. The lack of personalized treatment, customizing medication use instructions to patients' educational level, culture, and health literacy, has been identified as a barrier to adherence [7]. Addressing the diverse living conditions and health needs of individuals seeking primary care goes far beyond simply replacing handwritten prescriptions with e-prescribing systems. Scientific advancements demonstrate the potential of generative artificial intelligence (AI) to support health care professionals and improve people's health. Among these is the promise of using intelligent systems to help patients correctly use prescribed medications [8]. However, the adoption of AI solutions remains challenging, encompassing regulatory aspects, infrastructure investments, the need for transparency and governance, as well as resistance to change among health care professionals [9].

In Brazil, the Unified Health System (SUS) features a standardized and interoperable electronic health record system for PHC, named e-SUS PHC. Implemented in 2011 and developed collaboratively ever since by the Secretariat of Primary Health Care of the Ministry of Health, with participation from health care professionals, health managers, and universities [10], the system includes a specific functionality for e-prescribing [11]. The national reach of e-SUS Primary Health Care (e-SUS APS) and its widespread use by Brazilian basic health units generate a vast quantity of daily electronic health records, related to approximately 1.48 million consultations per day, according to January 2025 data from the system's coordination, authors RACG and ICG [11].

The personalization of patients' instructions on medication use, drawing on large language models (LLMs) and based on prescription data, has been studied using the computational structure of e-SUS APS [12]. Preliminary results demonstrated the potential of LLMs in generating instructions to facilitate communication between health care professionals and patients, which could promote greater adherence to treatments [12]. The development process has been conducted collaboratively between academia and the Brazilian Ministry of

XSL•FO
RenderX

Health—General Coordination of Innovation and Digital Acceleration of the Secretariat of Primary Health Care—through an interdisciplinary team of researchers in a simulated environment that represents the most frequent real-world prescription scenarios in PHC. Regulatory requirements and principles for the ethical and responsible use of AI have been considered since the planning stage. An initial version of the product was made available for testing, capable of incorporating additional instructions into prescriptions generated within the electronic system, on demand and at the free choice of health care professionals [13].

Building on the progress achieved in developing this innovative product, this study aims to evaluate the performance of LLMs, adjusted to citizens' needs, as perceived by users of the e-SUS PHC. This represents a relevant step for planning the nationwide implementation of a new functionality within the e-SUS APS electronic health record.

Methods

Overview

The methodology followed international guidelines for publishing studies evaluating Health Informatics apps, STARE-HI (Statement on Reporting of Evaluation Studies in Health Informatics) [14]. The technical artifact under evaluation is the use of LLMs as a functionality within an e-prescribing system, the e-SUS PHC, and its usefulness for patients, as perceived by the prescribing health care professional.

Study Design

This is a randomized, blinded experimental study that evaluated the quality of instructions generated by 3 language models designed to complement medication prescriptions during the preparation of e-prescriptions.

Development of the Prescriber Support Solution

Using the collaborative methodology of the Scrum framework within Agile methodology, regular meetings facilitated the progressive development of automated intelligent solutions, adjusted to the e-SUS PHC system, according to the national electronic health record environment. Based on a review of the state of the art and partial validation processes, the software has received continuous adjustments, as presented in Table 1, which describes the development stages.

XSL•FO
RenderX

Table 1. Stages of innovation development for the generation of medication use instructions with the support of generative artificial intelligence.

Milestones	Achieved or planned results
Workshop for interdisciplinary conceptual alignment	Multidisciplinary research team combining expertise in health sciences, data science, design, and linguistics, with proven capacity for development, evaluation, and refinement of the proposed innovation.
Selection of LLMs[a]	Completion of feasibility, safety, and explainability analysis of available LLMs, and selection of models for testing [15].
Development of a simulated database comprising PHC[b] clinical histories	A set of 104 clinical scenarios involving citizens with different characteristics, with expected medication use instructions prepared by specialists, already reported [12].
Development of a virtual testing environment	An interface similar to the e-SUS PHC[c] prescribing environment, which presents citizen data, one medication prescription at a time, reference instructions, and usage instructions generated by models with a form containing evaluation metrics.
Proof of concept with internal evaluation conducted by the research team	Comparative evaluation of the performance of ChatGPT-4.0 and Llama3.1-8B models fine-tuned through prompt engineering. The study demonstrated LLM potential to generate supplementary usage instructions to accompany medication prescriptions, already reported [12].
Model enhancements	Introduction of RAG[d] into models to mitigate inadequacies in medication use instructions generated by LLMs, already reported [13].
Refinement of the simulation environment for real-time testing (this study)	Interface adjusted for health care professionals to elaborate electronic prescriptions using LLMs, generating instructions on how to use the medication.
External validation of minimum product version (this study)	Use of virtual simulation environment by prescribers within the e-SUS APS[e] system. Comparative evaluation of the performance of models based on ChatGPT-4.0 and Llama3.1-8B, fine-tuned through prompt engineering, tested with and without RAG.
Responsible AI[f]	Analysis of requirements for transparency, robustness, usability, traceability, universality, and fairness necessary for implementation of the innovation in e-SUS APS, following international consensus criteria of Lekadir et al [9].
Planning deployment and implementation	Planning national implementation of a new functionality in the Brazilian e-SUS PHC electronic health record. Next steps.

[a]LLM: large language model.

[b]PHC: primary health care.

[c]e-SUS PHC: an electronic health record system for PHC of the Unified Health System in Brazil.

[d]RAG: retrieval-augmented generation.

[e]e-SUS APS: e-SUS Primary Health Care.

[f]AI: artificial intelligence.

Panel of Expert Evaluators

A panel of health care professionals was invited to voluntarily participate as human evaluators of instructions generated by the language models. Eligibility criteria included being a professional working in the national health care system—the SUS—preferably using the e-SUS PHC electronic health record, and holding legal authority to prescribe medications conferred by a professional council, which in PHC settings includes physicians, dentists, and nurses.

Email invitations were sent with the support of municipal PHC managers through the PHC Secretariat, totaling 1111 correspondences across different Brazilian municipalities, with a minimum of 60 volunteers as the target, since there is no established consensus on sample size for generative AI model validation with end users. Given the exploratory and semantic nature of LLM output, the theoretical saturation criterion was more appropriate than a priori sample probabilistic calculation [16]. The first step was giving informed consent and assent through the online questionnaire. A total of 134 health care professionals formally consented and received basic training on using the simulation environment for LLM testing, which included textual and audiovisual explanations, virtual meetings, and a permanent communication channel for inquiries. Of the 134 health professionals who consented, 72 (53.7%) did not complete any evaluations. The final panel of evaluators consisted of 62 specialists, of whom 61 completed the full evaluation process as proposed, while 1 completed most of it, meeting the saturation criterion, achieving all the main categories of flaws that the models present. The form used to characterize the evaluator is found in Table S1 in Multimedia Appendix 1.

For each prescription, health care professionals self-reported their confidence in evaluating the scenario, ranging from 0 to 100, stating their perceived knowledge necessary to assess that specific prescription case. The evaluators' confidence had a median of 100 (IQR 14), with 79.3% (157/198) of evaluators having confidence of >80% and 68.2% (135/198) having confidence of 90%. Figure 1 presents sociodemographic data of evaluators, with a median of 5 (IQR 5) years of experience in the SUS PHC system, and mostly female health professionals. After the evaluation phase, professionals were invited to share their opinions and concerns regarding the use of AI in health care. Two open-ended questions were used to assess evaluators'

XSL•FO
RenderX

opinions on the advantages and disadvantages of using AI to support communication between health professionals and patients. Participant statements were first thematically grouped into a set of overarching advantage and disadvantage categories reflecting major concerns about AI-assisted generation of medication use instructions. Advantage themes included patient safety and error reduction; communication, comprehension, and health literacy; efficiency, time optimization, and workflow support; clinical decision support and professional augmentation; standardization and quality of prescriptions; accessibility, equity, and responsible use. Disadvantage themes included patient safety and risk of errors; overreliance on AI and professional responsibility; lack of personalization and context sensitivity; impact on communication and humanization of care; comprehension, information overload, and health literacy; ethical, legal, and data protection concerns; organizational and system-level burdens; unequal access and structural inequities. Each statement was then manually coded using this category framework, allowing multiple categories to be assigned when relevant. For visualization, thematic codes were aggregated by professional profile and polarity (advantages vs disadvantages), and frequencies were normalized as within-profile percentages of coded themes. Because individual statements could be associated with more than one theme, percentages do not sum to 100% within profiles. The resulting distributions are shown in the heatmaps in Figures 2 and 3, where color intensity represents the relative prominence of each theme within dental surgeons, physicians, nurses, and researchers, with darker shades indicating higher thematic salience. In Figure 1, we present the evaluators' profile and geographic distribution. Figures 2 and 3 present heatmaps of advantage and disadvantage themes, respectively, concerning the use of AI-generated medication instructions to support communication between health professionals and patients across the 4 professional profiles.

Figure 1. Profile of health care professionals panel and their evaluation of texts generated by large language models. AI: artificial intelligence; e-SUS PHC: an electronic health record system for PHC of the Unified Health System in Brazil.

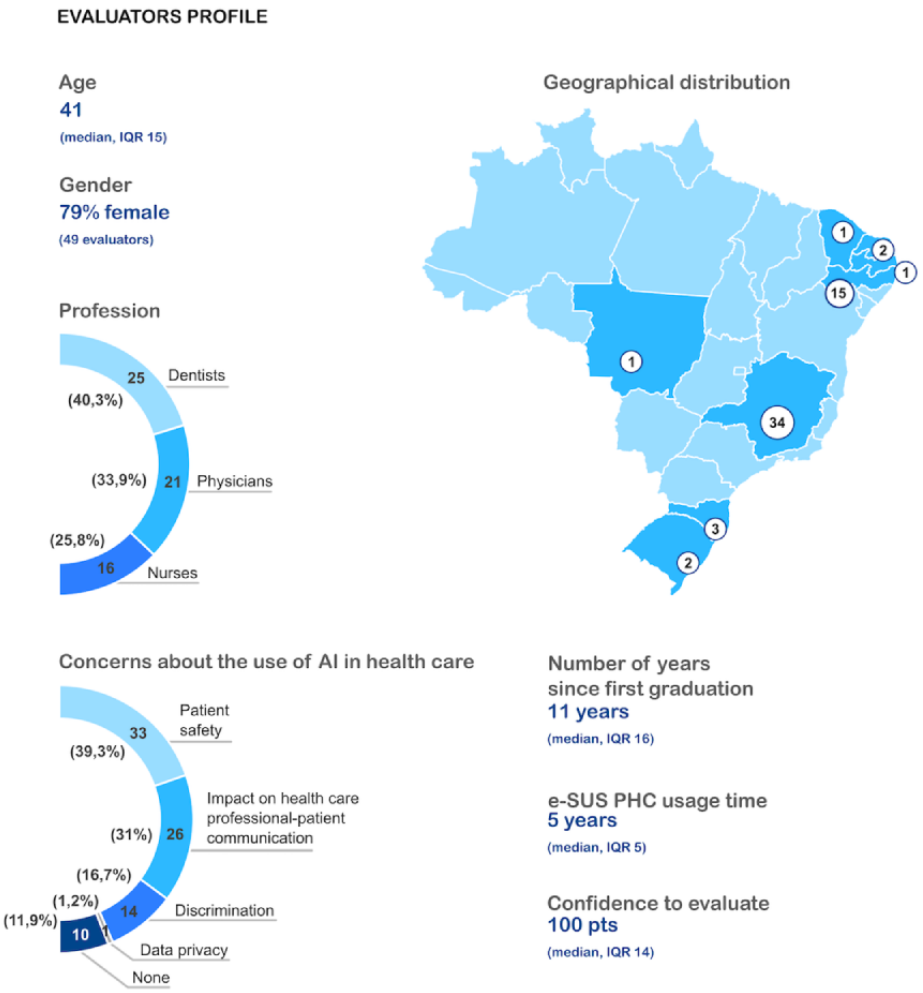

XSL•FO
RenderX

Figure 2. Advantages of using artificial intelligence to support communication between health professionals and patients. PHC: primary health care.

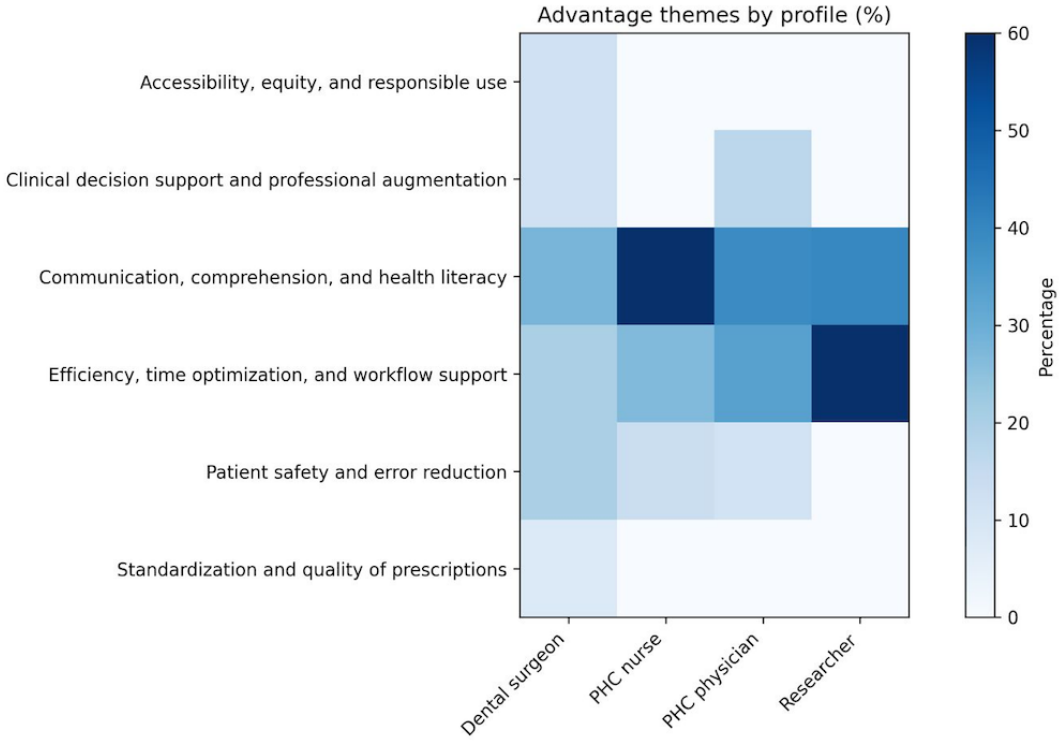

Figure 3. Disadvantages of using artificial intelligence to support communication between health professionals and patients. AI: artificial intelligence; PHC: primary health care.

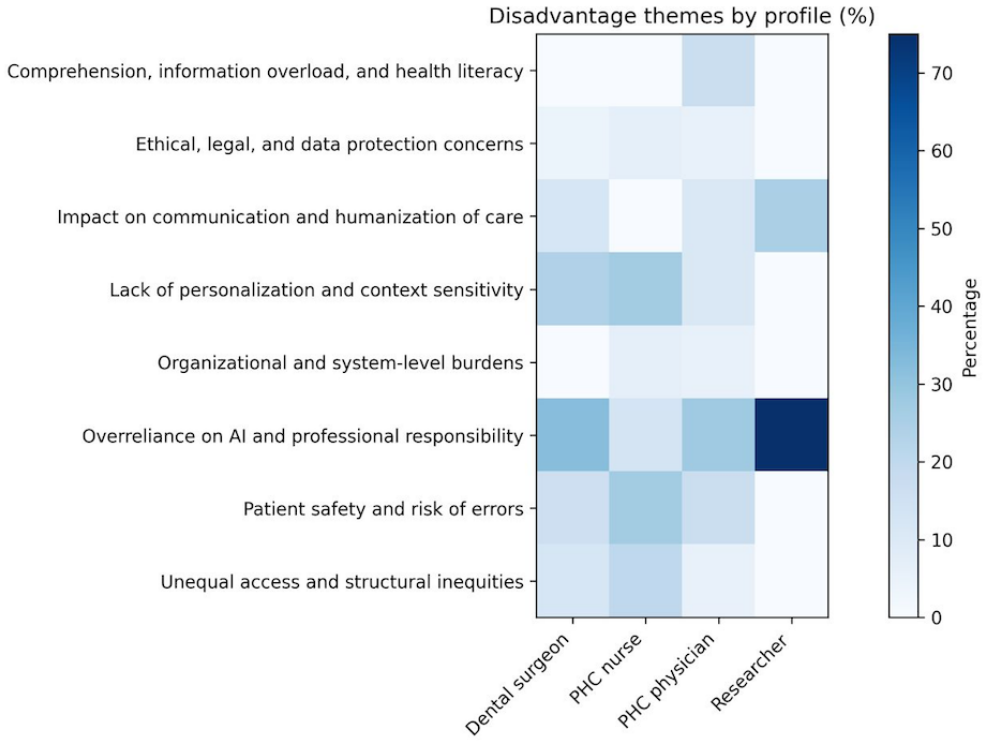

Evaluation Workflow of LLMs

The evaluation workflow for instructions generated by the models encompassed the following sequential steps: (1) development of the simulation environment containing simulated clinical scenarios, (2) definition of model validation metrics, and (3) data collection from health care professionals' evaluations.

Minimum Product Version

To provide a realistic e-prescribing experience supported by LLMs, a simulation environment was developed with a layout

XSL•FO
RenderX

similar to that of the e-SUS PHC electronic health record. A set of 16 clinical scenarios, contextualized within primary care practices, was provided by the researchers and internally validated for clarity and relevance to the intended context, with each scenario requiring only one medication. Ten scenarios were designed to induce medical prescriptions, 3 for nursing prescriptions, and 3 for dental prescriptions. To enable health care professionals to make choices and decisions based on their competence and prior experience, the environment provided a list of medications containing the active ingredient name, presentation form, concentration, and administration routes, compiled from the e-SUS PHC electronic health record. The detailed information prepared to generate prescriptions in the virtual validation environment is available in Tables S2-S4 in Multimedia Appendix 1.

Three models were compared: ChatGPT-4.0 (OpenAI), Llama3.1-8B (Meta), and Llama3.1-8B with on-demand RAG (retrieval-augmented generation; Meta) using patient information leaflets (PILs) from the National Health Surveillance Agency (ANVISA). Using 16 scenarios, LLMs generated instructions that gathered 48 test options. We used the same prompt engineering, designed and previously tested on another experimental database from our product's earlier development stage, available in our previous publication [13]. The models' temperature parameter was set to 0 for ChatGPT-4 and 0.00001 for Llama 3 to increase the predictability of the generated texts. Input variables, prompts used, and the RAG resource are described in Table S5 in Multimedia Appendix 1.

Evaluation Criteria

The performance of the 3 models in generating text-format outputs to facilitate medication use was analyzed according to criteria for validating generative AI models recommended by Tam et al [17]. The criteria considered were as follows:

1. Adequacy and correctness: model's capacity to generate responses that are accurate and appropriate to the user's purpose, context, and needs.
2. Completeness: model's capacity to generate responses or content that adequately covers all relevant aspects of a query or task.
3. Clarity: model's capacity to present information in an understandable, organized, and unambiguous manner.
4. Language simplification: model's capacity to generate tailored instructions that can engage and be understood by users, such as addressing them by their first name, using commands (imperative verb forms), and simplifying technical content introduced by RAG in PILs.
5. Usefulness: model's capacity to generate medication use instructions that are useful for complementing prescriptions.

A free-text field was provided for the evaluator to complete their assessment. The evaluation form is available in Table S6 in Multimedia Appendix 1.

Each criterion was evaluated by users on a 5-point scale (1: strongly disagree to 5: strongly agree), which was adjusted to a range of 1 to 5. The normalization of the scale for a comparable sum from the 1-5 Likert scale is adjusted as follows: for odd items (positive), the adjustment is "response - 1" (ie, 1 becomes 0 and 5 becomes 4); for even items (negative), the adjustment is "5 - response" (ie, 1 becomes 4 and 5 becomes 0). For instance, when an evaluator assigned 5 for positive and 1 for negative in adequacy, the raw score was calculated as follows: [(*5*–1) + (5–*1*)] = 8. Responses in italic.

The performance criteria for the models were weighted according to their clinical relevance, as determined by the researchers: adequacy received a weight of 2, clarity and usefulness received a weight of 1.5, and completeness and language simplification each received a weight of 1. Table 2 presents details regarding the composition of the overall score assigned to each question of the evaluation instrument and its weighting.

The weighted overall score generated an increasing numerical indicator, ranging from 0 to 56 (sum of total achievement), while the score per criterion received a value according to its weighting. Normalized to 100, the value is calculated by the formula: performance = 100 × weighted score/56. For instance, when an evaluator assigned 5 for positive and 1 for negative in adequacy, clarity, usefulness, completeness, and for language simplification, the weighted overall score was calculated as follows: [(5–1)+(5–1)]+[(5–1)+(5–1)]+[(5–1)+(5–1)]+[(5–1)+(5–1)]+[(5–1)+(5–1)]=40. In this example, applying the domain weights, the weighted score is 2×8+1.5×8+1.5×8+1×8+1×8=56, and the performance was 100×56/56=100.

There are fundamental advantages in the proposed methodology derived from the System Usability Scale [18]: (1) the balance between alternating positive and negative statements, instead of strictly positive propositions, aims to avoid the respondent's tendency to agree and allows for capturing more precise nuances of user perception. (2) The adaptation of the sum allows for generating a total normalized score from 0 to 100—more intuitive and easier to interpret (100=excellent quality, 0=extremely low quality). We calculated the internal consistency of the scale and interrater variability.

Additionally, errors in the texts generated by the models were verified by health care professional evaluators. If present, they were classified according to a typology proposed by Roy et al [19] for ChatGPT, simplified into 4 types: error type 1: instructions may lead to incorrect medication use; error type 2: instructions for use are contradictory or vague; error type 3: there are factual (nonmedical) errors; error type 4: there is information unrelated to the prescription or that is entirely nonsensical (model hallucination).

XSL•FO
RenderX

Table 2. Composition of the weighted overall score for the evaluation of texts generated by the models.

Criterion	Weight	Number of items × value per item	Total achievement
Adequacy and correctness	2	2×4	8×2=16
Clarity	1.5	2×4	8×1.5=12
Usefulness	1.5	2×4	8×1.5=12
Completeness	1	2×4	8×1=8
Language simplification	1	2×4	8×1= 8

Data Collection

The distribution of the 48 test options, generated from 16 clinical scenarios and the 3 models, was randomly automated within the simulation environment. Each health care professional received at least 3 clinical scenarios that required the prescription of a medication. Upon completion of the prescription data, the usage instructions were automatically generated by 1 of the 3 models. The health care professional was unaware of which model generated the instructions presented sequentially after the prescription they created in the simulation environment.

Outcomes and Data Analysis Methods

The primary outcome of this study was the weighted overall score achieved by the models, as assessed by evaluators using quantitative metrics, and for each specific criterion. The values for the positive and negative questions, as well as the specific scores achieved for the criteria of adequacy, completeness, clarity, language simplification, and usefulness, were compared among the 3 models. The secondary outcome focused on the incidence of errors identified within the texts generated by the models. Consequently, the models' performance evaluation encompassed both the global score and the frequency of errors in the generated instructions.

Qualitative variables were described using absolute and relative frequencies. Measures of central tendency (average and median) and variability (SD or IQR or range) were calculated for quantitative variables, according to their distribution. To analyze parametric distribution, averages were compared using the ANOVA 1-way test, and for nonparametric distribution medians were compared using the Kruskal-Wallis test, since each instruction was independently generated, even using the same clinical scenario. We used Bonferroni post hoc test to compare subgroups. The consistency analysis of each criterion used in the scale consisted of Cronbach α. We assessed observer variability among health professionals for each simulated scenario and model, calculating the coefficient of variability, taken in percentual [20]. Statistical tests were conducted using SPSS software (version 26.0; IBM Corp). The significance level for hypothesis tests was 5%, and 95% CIs were used.

Ethical Considerations

The study was approved by the Research Ethics Committee of the Federal University of Minas Gerais (CAAE: 78883924.7.0000.5149). The volunteers signed an informed consent form and did not receive any benefit for participating in our study.

Results

Overview

From the prescriptions prepared by health care professionals in the simulation environment, 198 evaluations of the performance of 3 models were conducted, distributed as shown in Figure 4.

XSL•FO
RenderX

Figure 4. Distribution of large language model instructions evaluations performed on prescriptions prepared by health care professionals. AI: artificial intelligence; PHC: primary health care.

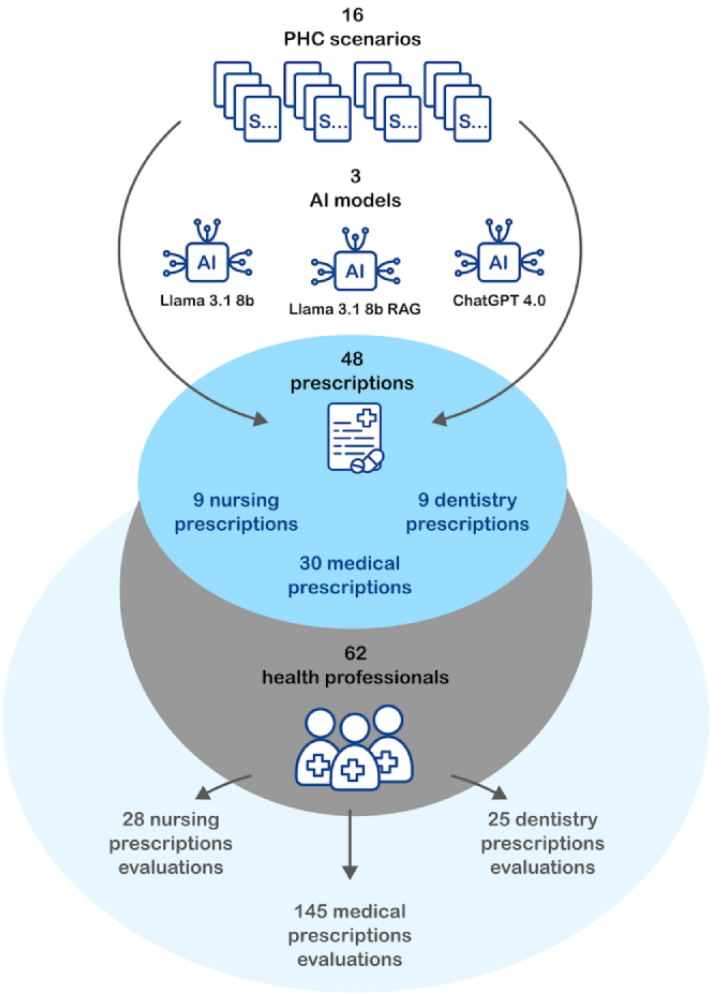

Model Performance by Clinical Scenario

Median of overall scores of 198 evaluations performed by 62 health professionals for 16 clinical scenarios is presented in Table 3. The instructions generated by ChatGPT-4.0 had overall scores above 90 in 4 (26.7%) and above 80 in 12 (80%) of 15 evaluated scenarios. The instructions generated by Llama3.1-8B models had overall scores above 90 in 3 (18.8%), and above 80 in 8 (50%) of 16. When generated by Llama3.1-8B-RAG, the instructions had values above 90 in 5 (33.3%), and above 80 in 9 (60%) of 15 scenarios. Interrater variability of scores among PHC professionals revealed wide differences of opinions ranging from 4.2% (n=3) to 85.8% (n=6).

XSL•FO
RenderX

Table 3. Model performance in the proposed clinical scenarios, by scenario (N=198). Values represent weighted scores based on the original 1-5 Likert scale.

Scenario	PHC[a] professional	Received evaluations in ech scenario, n	Model performance (overall weighted scores)								
			ChatGPT-4.0			Llama3.1-8B			Llama3.1-8B-RAG[b]		
			Model evaluations in each scenario, n	CV[c] (%)	Median (IQR[d])	Model evaluations in each scenario, n	CV (%)	Median (IQR)	Model evaluations in each scenario, n	CV (%)	Median (IQR)
1	Dentist	9	3	26.6	85.7 (36.6)	3	4.2	100 (7.1)	3	13.2	85.7 (22.3)
2	Dentist	9	3	60.4	84.8 (74.1)	3	7.7	88.4 (12.5)	3	25.1	85.7 (39.3)
3	Dentist	7	3	9.8	85.7 (16.1)	2	13.7	91.1 (17.9)	2	7.3	95.1 (9.8)
4	Nurse	16	4	12.6	90.6 (20.5)	6	5.3	95.5 (11.6)	6	37.2	86.2 (69.6)
5	Nurse	7	4	25.5	86.6 (42.0)	3	78.6	26.8 (26.8)	0	NE[e]	NE
6	Nurse	5	1	N/A[f]	65.2 (N/A)	1	N/A	82.1 (N/A)	3	20.9	93.8 (N/A)
7	Physician	26	9	32.5	87.5 (50.0)	9	53.5	64.3 (83.9)	8	32.5	71 (65.2)
8	Physician	28	9	24.0	89.3 (60.7)	10	57.3	40.2 (76.8)	10	26.3	67.9 (51.8)
9	Physician	23	7	12.2	97.3 (31.3)	8	58.1	47.8 (90.2)	8	28.7	72.3 (80.4)
10	Physician	22	8	31.8	90.6 (80.4)	7	36.2	83.9 (78.6)	7	41.3	83 (77.7)
11	Physician	21	8	21.6	92.4 (50.0)	7	54.2	51.8 (71.4)	6	85.8	33 (95.5)
12	Physician	4	0	N/A	NE	2	8.5	89.3 (10.7)	2	50.1	71.9 (50.9)
13	Physician	5	2	11.9	84.8 (14.3)	1	N/A	60.7 (N/A)	2	34.4	66.1 (32.1)
14	Physician	6	2	13.9	91.1 (17.9)	1	N/A	42.9 (N/A)	3	5.7	96.4 (10.7)
15	Physician	7	2	15.1	75.0 (16.1)	2	18.9	70.1 (18.8)	3	2.5	92 (4.5)
16	Physician	3	1	N/A	77.7 (N/A)	1	N/A	91.1 (N/A)	1	N/A	98.2 (N/A)
Total by model		198	66	25.8	88.4 (22.8)	66	76.5	66.5 (50.9)	66	30.9	79.9 (24.7)

[a]PHC: primary health care.

[b]RAG: retrieval-augmented generation.

[c]CV: coefficient of variation.

[d]IQR: interquartile range.

[e]NE: not evaluated.

[f]N/A: not applicable.

XSL•FO
RenderX

Median values for the overall weighted score assigned to the 3 models across the scenarios were different among them (Kruskal-Wallis test, *P*=.003), as presented in Figure 5, with median (IQR) values. Llama3.1-8B-RAG received evaluations with similar overall scores to ChatGPT-4.0 (post hoc test *P*=.05) and similar to Llama3.1-8B (post hoc test *P*=.44). ChatGPT-4.0 outperformed Llama3.1-8B (Bonferroni test *P*<.001).

Upon separate and nonweighted evaluation of the 5 criteria that make up the overall score, there were differences among the models in all of the criteria, as presented in Table 4. Comparing pairs of models, the specific scores assigned to the models ChatGPT4.0 with the Llama3.1-8B-RAG, 4 criteria showed no differences: adequacy (post hoc test, *P*=.54), completeness (post hoc test *P*=.38), clarity (post hoc test *P*=.09), and usefulness (post hoc test *P*=.63), considering the total of positive and negative questions for each criterion. ChatGPT-4.0 received higher scores in the language simplification criterion, for both the positive (1-way ANOVA *P*=.001) and negative (1-way ANOVA *P*=.03) questions, as shown in Table 4. Llama3.1 without the RAG feature scored lower than ChatGPT-4.0 across all 5 criteria, matching ChatGPT-4.0's results only for the language simplification—negative question "My prescription did not improve with the guidelines generated by the AI" (*P*=.06).

Internal consistency of the proposed scale had the Cronbach α value of 0.811 for adequacy and correctness, 0.641 for completeness, 0.757 for clarity, 0.680 for language simplification, and 0.724 for usefulness.

Regarding the presence or absence of errors, the 3 models were similar in terms of error type 1 (chi-square test *P*=.71), and type 4 (chi-square test *P*=.67), as presented in Table 5. ChatGPT-4.0 had a lower frequency of type 2, related to factual inconsistencies (chi-square test *P*<.001) and type 3, related to contradictory or vague instructions (chi-square test *P*=.02) errors. Llama3.1-8B-RAG had a lesser frequency of type 2 (n=11, 16.7%) and type 3 (n=5, 7.6%) errors compared to Llama3.1-8B without RAG, type 2 (n=23, 34.8%) and type 3 (n=10, 15.2%) errors.

Figure 5. Comparisons of model performance in generating medication use instructions for digital prescriptions, according to health professionals' assessment. IQR: interquartile range; RAG: retrieval-augmented generation.

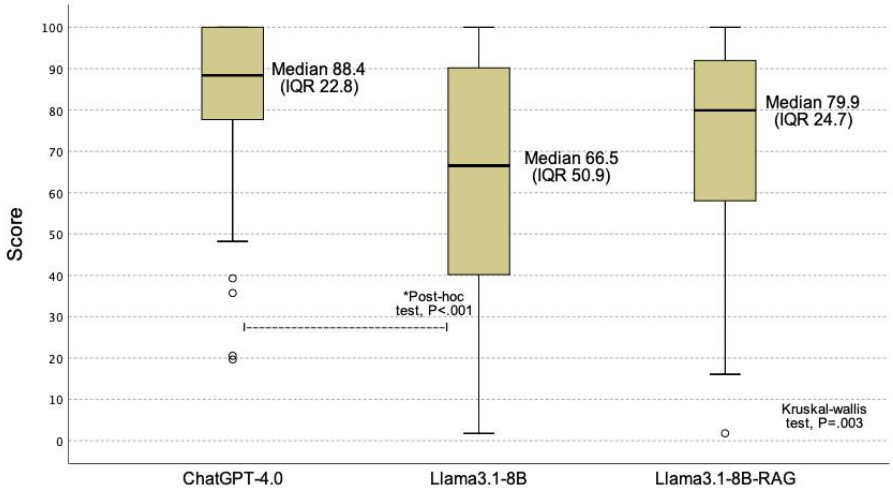

XSL•FO
RenderX

Table 4. Specific and raw evaluation of adequacy, completeness, clarity, language simplification, and usefulness, by health care professionals (N=198)[a].

Criteria	ChatGPT-4.0, mean (SD)	Llama 3.1-8B, mean (SD)	Llama 3.1-8B-RAG[b], mean (SD)	Comparisons among models[c], *P* value			
				Overall	1	2	3
Positive question: the instructions are correct and consistent with widely accepted practices in the health care field.	4.45 (1.0)	3.80 (1.5)	4.23 (1.1)	.008	.006	.84	.13
Negative question: the instructions contain harmful or incorrect information regarding the use of medication.	1.64 (1.3)	2.41 (1.7)	1.98 (1.5)	.01	.008	.52	.29
Adequacy and correctness: positive and negative sum	6.82 (2.1)	5.39 (2.9)	6.24 (2.3)	.004	.003	.54	.14
Positive question: the instructions covered relevant aspects for the correct use of the medication.	4.45 (0.9)	3.67 (1.4)	4.32 (1.0)	<.001	<.001	>.99	.03
Negative question: the instructions did not include sufficient information for the patient to correctly use the medication.	1.91 (1.4)	2.73 (1.6)	2.38 (1.5)	.008	.006	.22	.56
Completeness: positive and negative sum	6.55 (1.8)	4.94 (2.7)	5.94 (2.2)	<.001	<.001	.38	.04
Positive question: the instructions were sufficiently clear for the patient to correctly take or use the medication.	4.33 (1.0)	3.50 (1.5)	3.98 (1.3)	.001	.001	.35	.09
Negative question: the instructions present disorganized information that is difficult for the patient to understand.	1.65 (1.2)	2.56 (1.5)	2.21 (1.4)	.001	.001	.07	.46
Clarity: positive and negative sum	6.68 (1.9)	4.94 (2.8)	5.77 (2.4)	<.001	<.001	.09	.13
Positive question: the instructions were written in an accessible format and comprehensible to the patient.	4.64 (0.7)	3.62 (1.5)	3.83 (1.4)	<.001	<.001	.001	.93
Negative question: the instructions provided excessive technical information for the patient.	1.59 (1.2)	2.21 (1.4)	2.39 (1.6)	.003	.03	.003	>.99
Language simplification: positive and negative sum	7.05 (1.5)	5.41 (2.5)	5.44 (2.6)	<.001	<.001	<.001	>.99
Positive question: the suggested instructions were useful for complementing the text of the prescription I prepared.	4.30 (1.1)	3.64 (1.4)	4.05 (1.2)	.008	.006	.69	.17
Negative question: my prescription did not improve with the guidelines generated by the AI[d].	2.35 (1.5)	2.97 (1.6)	2.62 (1.4)	.06	.06	.90	.55
Usefulness: positive and negative sum	5.96 (2.2)	4.67 (2.7)	5.42 (2.4)	.01	.008	.63	.22

[a]Values represent raw scores based on the original 1-5 Likert scale.

[b]RAG: retrieval-augmented generation.

[c]ANOVA 1-way test. (1) Post hoc test ChatGPT-4.0 vs Llama3.1-8B. (2) Post hoc test ChatGPT-4.0 vs Llama3.1-8B-RAG. (3) Post hoc test Llama3.1-8B vs Llama3.1-8B-RAG.

[d]AI: artificial intelligence.

XSL•FO
RenderX

Table 5. Frequency of errors found in the instructions for use of medications generated by the models (N=198).

Frequency of errors	ChatGPT-4.0 (n[a]=66)	Llama3.1-8B (n=66)	Llama3.1-8B -RAG[b] (n=66)	Comparisons between models[c], *P* value
Error type 1: instructions leading to incorrect use of the medication, n (%) (95% CI)	15 (22.7) (13.8-33.8)	19 (22.8) (18.8-40.4)	19 (22.8) (15.0-35.4)	.71
Error type 2: contradictory or vague instructions for use, n (%) (95% CI)	4 (6.1) (1.9-13.5)	23 (34.8) (24.1-46.8)	11 (16.7) (9.0-26.8)	.001
Error type 3: factual errors, n (%) (95% CI)	1 (1.5) (0.1-6.5)	10 (15.2) (7.9-25.1)	5 (7.6) (2.8-15.6)	.02
Error type 4: there is information that is not related to the prescription, or that is completely meaningless (model hallucination), n (%) (95% CI)	7 (10.6) (4.7-19.5)	9 (13.6) (6.8-23.2)	6 (9.1) (3.7-17.6)	.67

[a]n: number of scenarios.

[b]RAG: retrieval-augmented generation.

[c]Chi-square test.

Discussion

Principal Findings

The adoption of technological approaches that support health care professionals at the point of care integrates the worldwide digital health landscape [21]. The most relevant finding of this study was demonstrating that LLMs performed well in generating supplementary instructions for medication prescriptions within PHC settings, as per health professionals' evaluation. We consider this analysis essential to overcome the preclinical evaluation step of the solution, thereby qualifying the models for scalability planning.

In the development timeline, the 3 models were progressively adjusted to meet the demands for adequate and simplified medication use instructions. The model's great overall scores, as attributed by the external evaluator panel, qualify the solution for a pilot implementation within the e-SUS PHC electronic health record environment. We highlight that the open-source Llama3.1-8B model was capable of performing the tasks defined in the prompt engineering, with or without RAG, with an overall quality equivalent to that of the closed-source ChatGPT-4.0 model, considered the state-of-the-art among LLMs [22]. The tailoring of messages in the context of health, by reducing jargon and vague terms and facilitating the comprehension of technical terms in the communication process between health care professionals and patients, is recognized as a promising frontier for generative AI [23].

The number of errors identified in the instructions generated by the models demonstrated the necessity for human oversight of the technology, which is expected in PHC settings in alignment with principles for the ethical and responsible use of AI [9]. For the introduction of this innovation in a real-world setting, we planned that health care professionals would be able to accept, edit, or delete the text suggested by LLM, taking responsibility for the final version presented to the patient. In this regard, both models demonstrated similar results, with fabricated information frequency ranging between 9.1% and 13.6%. These findings still require improvements in the technology, which is expected with an increased number of professionals using, fine-tuning, and evaluating the outputs.

The expected digital transformation in health care demands that systems offer functionalities centered on human values, prioritizing equity and privacy while addressing the challenges of care delivery. Usefulness, safety, and inclusion must be prioritized from the initial stages of AI development and scalability [20]. Reports on AI to promote treatment adherence are still incipient, however promising. In a systematic review, Reis et al [8] highlighted that AI-based interventions improve medication adherence by 6.7% to 32.7% compared to current practices in randomized clinical trials.

Strengths and Limitations of the Study

The clinical evaluation of LLMs in health care settings is complex and demands interdisciplinary collaboration to meet both clinical requirements and rigorous validation criteria. A relevant point to qualify our findings is the diverse representativeness of PHC prescriber categories in the evaluation process, as highlighted in international consensus guidelines for the introduction of AI in health care [9]. Concerning the panel of evaluators, the majority of health care professionals expressed concerns about overreliance on the technology when using AI and patient safety. This reinforces the emphasis placed on the adequacy criterion in weighting the models' performance score. Adequacy is a relevant criterion indicating the model's capacity to generate responses appropriate to the purpose, context, and meet the patient's needs. In the approach considering specific criteria, RAG added to Llama3.1-8B model proved sufficient to match the score attributed to ChatGPT-4.0 in the evaluation items of adequacy, completeness, clarity, and usefulness. However, RAG in PILs did not contribute to the language simplification of the guidance, which will require a prompt redesign.

Additionally, we emphasize the relevance of the randomized assignment of the PHC clinical scenarios and LLMs among evaluators, resulting in 46 different tests. These professionals were unaware of the prescription-inducing narratives and had no prior knowledge of the type of instructions the models might suggest during prescription preparation. In current practices, adding instructions for correct medication use to a prescription is not mandatory, nor is there a standard for the content of these instructions; both aspects are left to the prescriber's discretion.

XSL•FO
RenderX

Nonetheless, the majority of evaluators considered that LLM-generated instructions were adequate, complete, simplified, clear, and useful for complementing their prescriptions. It is relevant to note that during the first steps of the innovation's development process, we defined the requirements for good practices on instructions, obtained through review of scientific articles and brochures [12]. We infer that LLMs will bring benefits to patients by facilitating their engagement in self-care through the correct use of prescribed medication. Although disagreements among human evaluators of generative AI outputs are expected [24], the majority of tests (n=38, 82.5%) in this study had scores attributed by more than one evaluator.

The results interpretation deserves some caution. We faced difficulties during the process of enrolling health professionals for the composition of the panel of evaluators. This challenge, joined with the random assignment of prescription scenarios, resulted in 2 out of the 48 elaborated tests having no evaluator and 8 (17.5%) with only 1 evaluator. Even in the absence of an established consensus in the literature regarding the ideal sample size for validation of generative AI systems with end users, our study achieved 62 volunteers with a diversity of PHC prescribers. Based on Holmes et al [25] findings, around 26 participants are needed to reveal usability flaws in chatbots, facing the open and nonlinear nature of natural language.

Regarding metrics of LLMs' performance, the adaptation of the quality evaluation scale derived from the System Usability Scale was possible due to its generalizable structure [18]. Although originally proposed for interactive systems, the scale was adapted to specific criteria within the context of an e-prescribing system. The main advantages are that the polarity of positive and negative questions for each criterion demands more attention from the respondent and mitigates automatic or biased responses, increasing the reliability of the collected data. Beyond this, the overall weighted score transformation, which assigns importance to concepts and converts them into a standardized score on a scale of 0 to 100, enables numerical comparisons between different AI models, facilitating statistical analyses and enhancing evaluation based on performance criteria. Our metrics in scale were acceptable in terms of consistency, demonstrating potential for use as an initial and exploratory tool, since all of the criteria had Cronbach α values above 0.64 and 3 criteria above 0.72 [26]. Borderline domains, such as completeness and language simplification, should be reviewed and improved for adoption in our next implementation step, considering the real scenario evaluation. We state the lack of verification regarding differential effects on the instructions across groups of patients with varying educational levels, cultural aspects, age or health literacy, which demands further attention.

Meaning and Generalizability of the Study

The implementation of AI in public health care settings faces several challenges, including the need for adequate infrastructure, secure and ethical access to data, ensuring privacy, disseminating technical knowledge for its sustainability, and establishing mechanisms for technological governance [23,27]. From the perspective of innovation scalability, the global weighted score of Llama 3.1-8B-RAG model was comparable to that of ChatGPT-4.0, except in terms of language simplification. Proprietary LLMs incur financial costs and send data to a web app, raising privacy concerns regarding sensitive health data. Open-source models are capable of local execution, making them viable and sustainable alternatives for public health deployment. Llama3.1-8B models demonstrated high performance when combined with RAG in PILs, which expands their capability by incorporating external information. The results pointed out a promising implementation, suggesting that other external information sources can enhance the model's performance, such as language simplification, by allowing for a more detailed incorporation of patient data. Following the incorporation of technology into the national eSUS APS system, it will be possible to use data from the Brazilian citizens' National Registry of SUS, which includes education, income, occupation, race, and gender. Even so, the use of LLMs in our innovation does not replace the individual attention of health care professionals with their patients, nor is it a solution for directly improving health literacy. However, it could improve the effective communication between PHC teams and their communities. A significant effect of language simplification of instructions on medication adherence requires repeated counseling sessions provided by health educators [28].

Another barrier to face is the provision of computational support that meets security, performance, and equity requirements for integrating AI processing into clinical workflows. The e-SUS PHC electronic health record is an example of an information system with a significant social role, aiming to adapt its interfaces and workflows to PHC work processes. It also considers the diversity of Brazilian municipalities, which are at unmapped stages of digital maturity [29].

Beyond infrastructure, the risk of transferring responsibility from the health care professional to technology must be considered. Comprehensive governance standards based on ethics, transparency, and model explainability can mitigate excessive dependence on technology while simultaneously contributing to overcoming resistance in the use of generative AI, directing the technology toward achieving greater efficiency in health care systems [22]. Efficiency, usefulness, and clinical impact on public health still need to be measured and monitored as requirements for reliable AI implementation in health care settings [9,22]. The expectation is that this innovation, integrated into the e-SUS PHC electronic health record, can contribute to greater adherence to medication treatment, promoting better health in alignment with the principles of universality and equity of the SUS.

Conclusions

The results indicated the potential utility of LLMs in prescribing to complement prescriptions, contributing to patient clarity regarding correct medication use. Open-source models demonstrated similar performance to closed-source models, guiding the scalability phase toward an environment favorable to data privacy. The introduction of LLM-generated text into the e-prescribing workflow, even when well-designed with prompt engineering and enhanced with content based on PILs, necessitates validation by the prescribing professional to ensure

XSL•FO
RenderX

safety and error control, as well as performance monitoring and governance.

Funding

The authors declare that financial support was received for the research and/or publication of this study. This work was supported by the Brazilian Ministry of Health/DECIT/CNPq (00758/2024-5) in partnership with the Gates Foundation, INV-007186. The funders had no role in the study design, data collection, analysis, interpretation, writing, or the decision to submit the manuscript.

Data Availability

Data are available under the terms of the Creative Commons BY, which requires appropriate credit to the original creator when using or adapting the work. Data are deposited in publicly available repositories, as follows:

Data file 1 - Questionnaire (CC-BY): Data collection responses for characterization of health care professionals invited to the model. The anonymized database, in Portuguese, is available in the institutional repository of the Universidade Federal de Minas Gerais, Brazil [30].

Data file 2 - Evaluation (CC-BY): It contains the LLMs' evaluation data, in Portuguese. Clinical scenarios are simulated. The performance of LLMs corresponds to the evaluator's perception as defined by metrics: the adequacy model's capacity to generate responses that are accurate and appropriate to the user's purpose, context, and needs. Completeness: model's capacity to generate responses or content that adequately covers all relevant aspects of a query or task. Clarity: model's capacity to present information in an understandable, organized, and unambiguous manner. Language simplification: the model's capacity to generate instructions that can be understood by the patient. Usefulness: model's capacity to generate medication use instructions that are useful for complementing prescriptions. Error types: type 1: instructions may lead to incorrect medication use; type 2: instructions for use are contradictory or vague; type 3: there are factual (nonmedical) errors; type 4: there is information unrelated to the prescription or that is entirely nonsensical (model hallucination). The anonymized database is available at the institutional repository of the Universidade Federal de Minas Gerais, Brazil [31].

Authors' Contributions

Conceptualization: ZSNR (lead), ETA (equal), LR (supporting), ASP (supporting)
Data curation: IJRO (lead), LR (lead), APSJ (supporting), DRJ (supporting), ETA (supporting)
Formal analysis: ZSNR (lead), ETA (equal), LR (equal), ASP (supporting), IJRO (supporting).
Funding acquisition: ZSNR (lead), RC-C (equal), RACG (supporting), ICG (supporting).
Investigation: ZSNR (lead), ETA (equal), EML (supporting), FRO (supporting), CSD (supporting), JAO (supporting), GMVP (supporting), EFM (supporting), IJRO (supporting), ICLO (supporting), ICG (supporting), RACG (supporting), RC-C (supporting)
Methodology: LR (lead), ETA (equal), ZSNR (equal), IJRO (supporting), APSJ (supporting), DRJ (supporting).
Project administration: ZSNR (lead), RC-C (equal), ETA (supporting)
Resources: LR (lead), IJRO (equal), RC-C (equal), RACG (supporting), ICG (supporting).
Software: IJRO (lead), LR (lead), APSJ (supporting), DRJ (supporting), ETA (supporting)
Validation: IJRO (lead), EML (supporting), FRO (supporting), CSD (supporting), JAO (supporting), GMVP (supporting), EFM (supporting), (supporting), ICG (supporting), (supporting), RACG (supporting), RC-C (supporting).
Visualization: EFM (lead), ICLO (supporting), IJRO (supporting).
Writing – original draft: ZSNR (lead), ETA (equal), EML (supporting), FRO (supporting), CSD (supporting), JAO (supporting), GMVP (supporting), EFM (supporting), (supporting), IJRO (supporting), ICLO (supporting), ICG (supporting), (supporting), RACG (supporting), RC-C (supporting)
Writing – review & editing: ZSNR (lead), ETA (equal), EML (supporting), FRO (supporting), CSD (supporting), JAO (supporting), GMVP (supporting), EFM (supporting), (supporting), IJRO (supporting), ICLO (supporting), ICG (supporting), (supporting), RACG (supporting), RC-C (supporting)

Conflicts of Interest

RC-C declares Virtual Care Company support. The other authors declare that they have no known competing financial or commercial interests or personal relationships that could have appeared to influence the work reported in this paper.

Multimedia Appendix 1
Additional tables.
[DOCX File , 3474 KB - jmir_v28i1e84444_app1.docx]

References

XSL•FO
RenderX

1. Cutler RL, Fernandez-Llimos F, Frommer M, Benrimoj C, Garcia-Cardenas V. Economic impact of medication non-adherence by disease groups: a systematic review. BMJ Open 2018;8(1):e016982 [FREE Full text] [doi: 10.1136/bmjopen-2017-016982] [Medline: 29358417]
2. Fischer MA, Stedman MR, Lii J, Vogeli C, Shrank WH, Brookhart MA, et al. Primary medication non-adherence: analysis of 195,930 electronic prescriptions. J Gen Intern Med 2010;25(4):284-290 [FREE Full text] [doi: 10.1007/s11606-010-1253-9] [Medline: 20131023]
3. Chauke GD, Nakwafila O, Chibi B, Sartorius B, Mashamba-Thompson T. Factors influencing poor medication adherence amongst patients with chronic disease in low-and-middle-income countries: a systematic scoping review. Heliyon 2022;8(6):e09716 [FREE Full text] [doi: 10.1016/j.heliyon.2022.e09716] [Medline: 35770147]
4. Labrique A, Vasudevan L, Mehl G, Rosskam E, Hyder AA. Digital health and health systems of the future. Glob Health Sci Pract 2018;6(Supplement 1):S1-S4. [doi: 10.9745/ghsp-d-18-00342]
5. Digital Implementation Investment Guide (DIIG): Integrating Digital Interventions into Health Programmes. Geneva: World Health Organization; 2020.
6. Bates DW. Effect of computerized physician order entry and a team intervention on prevention of serious medication errors. JAMA 1998;280(15):1311-1316. [doi: 10.1001/jama.280.15.1311] [Medline: 9794308]
7. Odukoya OK, Chui MA. E-prescribing: a focused review and new approach to addressing safety in pharmacies and primary care. Res Soc Adm Pharm 2013;9(6):996-1003. [doi: 10.1016/j.sapharm.2012.09.004]
8. Reis ZSN, Pereira GMV, Dias CDS, Lage EM, de Oliveira IJR, Pagano AS. Artificial intelligence-based tools for patient support to enhance medication adherence: a focused review. Front Digit Health 2025;7:1523070 [FREE Full text] [doi: 10.3389/fdgth.2025.1523070] [Medline: 40364851]
9. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ 2025;388:e081554. [doi: 10.1136/bmj-2024-081554]
10. Coelho Neto GC, Andreazza R, Chioro A. Integration among national health information systems in Brazil: the case of e-SUS Primary Care. Rev Saúde Pública 2021;55:93 [FREE Full text] [doi: 10.11606/s1518-8787.2021055002931] [Medline: 34878089]
11. Reis ZSN, Pagano A, Senna MI, De Oliveira JL, Asevedo RA, et al. Health record in PHC: ongoing education for nurses, family health team doctors, dental surgeons and multidisciplinary teams. Belo Horizonte, Minas Gerais, Brazil: Federal University of Minas Gerais; 2024:1-162.
12. Reis ZSN, Pagano AS, Ramos De Oliveira IJ, Dias CDS, Lage EM, Mineiro EF, et al. Evaluating large language model-supported instructions for medication use: first steps toward a comprehensive model. Mayo Clin Proc Digit Health 2024;2(4):632-644 [FREE Full text] [doi: 10.1016/j.mcpdig.2024.09.006] [Medline: 39679140]
13. de Jesus DDR, de Souza Júnior AP, de Albergaria ET, Pagano AS, de Oliveira IJR, Dos Santos Dias C, et al. Enhanced LLM-supported instructions for medication use through retrieval-augmented generation. Comput Biol Med 2025;198(Pt A):111135 [FREE Full text] [doi: 10.1016/j.compbiomed.2025.111135] [Medline: 41077040]
14. Talmon J, Ammenwerth E, Brender J, de Keizer N, Nykänen P, Rigby M. STARE-HI-statement on reporting of evaluation studies in health informatics. Int J Med Inform 2009;78(1):1-9. [doi: 10.1016/j.ijmedinf.2008.09.002] [Medline: 18930696]
15. Basilio VARZSN. Comparative analysis of large-scale language models: applications in the healthcare sector. In: Extended Proceedings of the 20th Brazilian Congress on Health Informatics-CBIS24. Belo Horizonte: Brazilian Society of Health Informatics; 2024:450-452.
16. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: a systematic review of empirical tests. Soc Sci Med 2022;292:114523 [FREE Full text] [doi: 10.1016/j.socscimed.2021.114523] [Medline: 34785096]
17. Tam TYC, Sivarajkumar S, Kapoor S, Stolyar AV, Polanska K, McCarthy KR, et al. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med 2024;7(1):258 [FREE Full text] [doi: 10.1038/s41746-024-01258-7] [Medline: 39333376]
18. Brooke J. SUS: a 'quick and dirty' usability scale. In: Usability Evaluation In Industry. Boca Raton: Chapman and Hall/CRC; 1996:4-7.
19. Roy S, Khatua A, Ghoochani F, Hadler U, Nejdl W, Ganguly N. Beyond accuracy: investigating error types in GPT-4 responses to USMLE questions. 2024 Jul 11 Presented at: SIGIR '24: Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval; July 11, 2024; New York, NY URL: http://arxiv.org/abs/2404.13307 [doi: 10.1145/3626772.3657882]
20. Armitage P, Berry G, Matthews JNS. Statistical Methods in Medical Research. New Jersey: John Wiley & Sons; 2013.
21. Digital transformation handbook for primary health care (2024): optimizing person-centred point of service systems. In: World Health Organization. Geneva: World Health Organization; 2024.
22. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Med Inform Decis Mak 2025;25(1):117 [FREE Full text] [doi: 10.1186/s12911-025-02954-4] [Medline: 40055694]
23. Bharel M, Auerbach J, Nguyen V, DeSalvo KB. Transforming public health practice with generative artificial intelligence: article examines how generative artificial intelligence could be used to transform public health practice in the US. Health Aff (Millwood) 2024;43(6):776-782. [doi: 10.1377/hlthaff.2024.00050] [Medline: 38830160]

XSL•FO
RenderX

24. Shankar S, Zamfirescu-Pereira J, Hartmann B, Parameswaran A, Arawjo I. Who validates the validators? Aligning LLM-Assisted evaluation of LLM outputs with human preferences. : ACM; 2024 Presented at: Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology; June 7, 2025; Pittsburgh, PA, USA URL: https://dl.acm.org/doi/10.1145/3654777.3676450 [doi: 10.1145/3654777.3676450]
25. Holmes S, Moorhead A, Bond R, Zheng H, Coates V, Mctear M. Usability testing of a healthcare chatbot: can we use conventional methods to assess conversational user interfaces? : ACM; 2019 Presented at: Proceedings of the 31st European Conference on Cognitive Ergonomics; December 23, 2025; BELFAST, UK URL: https://dl.acm.org/doi/10.1145/3335082.3335094 [doi: 10.1145/3335082.3335094]
26. Hair JF. Multivariate Data Analysis. Porto Alegre: Bookman (translation Adonai Schluo Sant'Anna); 2009.
27. Olawade DB, Wada OJ, David-Olawade AC, Kunonga E, Abaire O, Ling J. Using artificial intelligence to improve public health: a narrative review. Front Public Health 2023;11:1196397 [FREE Full text] [doi: 10.3389/fpubh.2023.1196397] [Medline: 37954052]
28. Kini V, Ho PM. Interventions to improve medication adherence: a review. JAMA 2018;320(23):2461-2473. [doi: 10.1001/jama.2018.19271] [Medline: 30561486]
29. de Figueirêdo RC, de Siqueira Silva Í, De Araújo AJ, Silva CRDV, Martiniano CS, Brito EWG, et al. Preparation and validation of the instrument "QualiAPS digital-Brazil" for assessing digital health care in primary health care: a required tool. Front Public Health 2024;12:1304148 [FREE Full text] [doi: 10.3389/fpubh.2024.1304148] [Medline: 39081352]
30. Reis ZSN, Albergaria ET, Pagano AS. Data file 1 - Questionnaire Preclinical Evaluation of Large Language Model-Generated Instructions to Complement Digital Prescriptions in Primary Health Care. 2025. URL: https://ftp.medicina.ufmg.br/cins/pesquisa/iapolis/BD1_Participantes.xlsx [accessed 2025-07-21]
31. Reis Z. Data file 2 - Evaluation. Preclinical evaluation of large language model-generated instructions to complement digital prescriptions in primary health care. 2025. URL: https://ftp.medicina.ufmg.br/cins/pesquisa/iapolis/BD_198.xlsx [accessed 2025-07-21]

Abbreviations

AI: artificial intelligence
ANVISA: Agência Nacional de Vigilância Sanitária, Health Surveillance Agency, Brazil
e-prescribing: electronic prescribing
e-SUS APS: e-SUS Primary Health Care
LLM: large language model
PHC: primary health care
PIL: patient information leaflet
RAG: retrieval-augmented generation
STARE-HI: Statement on Reporting of Evaluation Studies in Health Informatics
SUS: Unified Health System

Edited by A Coristine; submitted 19.Sep.2025; peer-reviewed by Y Wang, H Yang, A Al-Muqarm; comments to author 04.Nov.2025; revised version received 27.Feb.2026; accepted 27.Feb.2026; published 26.May.2026.

Please cite as:
Silveira Nogueira Reis Z, Tuler Albergaria E, Silvina Pagano A, Martins Lage E, Ribeiro de Oliveira F, dos Santos Dias C, Almeida Oliveira J, Miranda Varella Pereira G, Jose Ramos de Oliveira I, Franco Mineiro É, Carvalho Lima Oliveira I, dos Reis de Jesus D, Pereira de Souza Júnior A, de Carvalho Gomes I, André Cuevas Gaete R, Cruz-Correia R, Rocha L
Large Language Model–Generated Patient Instructions for Prescriptions in Primary Health Care: Preclinical Algorithm Validation
J Med Internet Res 2026;28:e84444
URL: https://www.jmir.org/2026/1/e84444
doi:10.2196/84444
PMID:42190235

©Zilma Silveira Nogueira Reis, Elisa Tuler Albergaria, Adriana Silvina Pagano, Eura Martins Lage, Flávia Ribeiro de Oliveira, Cristiane dos Santos Dias, Juliana Almeida Oliveira, Gláucia Miranda Varella Pereira, Isaias Jose Ramos de Oliveira, Érico Franco Mineiro, Igor Carvalho Lima Oliveira, Davi dos Reis de Jesus, Antônio Pereira de Souza Júnior, Igor de Carvalho Gomes, Rodrigo André Cuevas Gaete, Ricardo Cruz-Correia, Leonardo Rocha. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is

XSL•FO
RenderX

properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Using a Large Language Model to Support Thematic Analysis of Patient Experiences in Chronic Illness Management: Comparative Qualitative Study

Sara Kivity[1], PhD; Yechiel Michael Barilan[1], MD, PhD; Reut Noham[2], PhD; Mor Saban[3], PhD

[1]School of Medicine, Gray Faculty of Medical and Health Sciences, Tel Aviv University, Tel Aviv, Tel Aviv, Israel

[2]Department of Industrial Engineering,, Tel Aviv University, Tel Aviv, Israel

[3]Nursing Department, The Stanley Steyer School of Health Professions, Gray Faculty of Medical and Health Sciences, Tel Aviv University, Tel Aviv, Israel

Corresponding Author:
Sara Kivity, PhD
School of Medicine, Gray Faculty of Medical and Health Sciences
Tel Aviv University
Haim Levanon Street 30
Tel Aviv, Tel Aviv, 6997801
Israel
Phone: 972 528815539
Email: dejav_s@mac.org.il

Abstract

Background: Qualitative health research often focuses on how patients experience and manage chronic illnesses, a topic that has been extensively studied in the literature. With the emergence of large language models (LLMs), such as Claude (Anthropic PBC) and ChatGPT (OpenAI), new opportunities are arising to support and scale the thematic analysis of narrative health data. However, their role and added value compared to traditional human-led approaches remain underexplored, particularly in complex clinical contexts such as multimorbidity.

Objective: We aim to evaluate the methodological contribution of LLM-assisted analysis by examining its ability to replicate and extend established qualitative insights, in comparison with traditional thematic analysis.

Methods: Semistructured interviews were conducted with 30 individuals living with two or more chronic illnesses. Transcripts were analyzed using both manual thematic coding and Claude 3.5 Sonnet. A structured comparison was conducted to identify shared and unique themes across the two approaches. The analysis examined thematic overlap, differences in subtheme identification, and variation in the level of detail between the methods.

Results: Both approaches identified similar core themes related to the patient experience, including health care navigation and challenges, support systems and family dynamics, and emotional challenges and coping. Manual analysis produced more contextually detailed interpretations, while the LLM approach identified a larger number of subthemes. Each method also revealed distinct themes: the manual analysis included themes such as faith, caregiving roles, and a proactive mindset, whereas the LLM identified themes such as future planning and multiple health conditions. The findings show both similarities and differences between the two approaches. The LLM analysis also demonstrated efficiency in processing large volumes of qualitative data.

Conclusions: A hybrid approach that integrates artificial intelligence–assisted and human-led thematic analysis can enhance both analytical depth and scalability. These findings support the use of LLMs as a complementary tool in qualitative research, while highlighting the importance of combining automated pattern detection with human interpretation.

(J Med Internet Res 2026;28:e88677) doi:10.2196/88677

KEYWORDS

chronic illness; disease management; qualitative research; patient experience; large language model

XSL•FO
RenderX

Introduction

Chronic illness management represents one of the most significant challenges in modern health care, requiring patients to navigate complex systems of care while integrating numerous health-related tasks into their daily lives [1,2]. Indeed, such management has evolved to include a broad spectrum of activities that patients must master and coordinate to the extent that they often develop what researchers term a "management career." This refers to the dynamic process of learning, adapting, and integrating various health-related tasks into daily routines [3,4]. These include medication management, symptom monitoring, communication with health care providers, navigating insurance and health care systems, and addressing the psychological and social impact of their condition [2,3]. Moreover, patients often face competing demands from different conditions and must navigate conflicting treatment recommendations, requiring them to make complex decisions about prioritizing various aspects of their care [2,4]. In addition, studies show that patients often reconstruct their personal identity to incorporate their role as health managers, while simultaneously navigating changes in their social and professional relationships [5,6]. Thus, the impact of chronic illness management extends across multiple aspects of patients' lives, constituting a dynamic "career" that goes beyond mere medical adherence [7]. Accordingly, research has identified several key factors that influence the management of chronic illnesses. These include social support networks, including family members, friends, and health care providers [1,2], patients' self-efficacy and problem-solving skills, and the ability to navigate health care systems and coordinate care efficiently [8,9].

Qualitative studies are commonly used in exploring the chronic patient experience in disease management, as they provide rich and contextualized insights into the lived experiences of individuals [10]. However, manual thematic qualitative analysis, while valuable, faces several limitations: it is time-intensive and resource-demanding [11,12]. This approach presents challenges when applied to large datasets, as maintaining analytical depth across numerous interviews requires substantial effort [13,14]. While human interpretation is central to qualitative research and contributes to its richness, the scalability of manual analysis is limited, particularly in team-based settings, where variations in coding and interpretation may affect consistency [15,16]. In recent years, the emergence of large language models (LLMs), such as ChatGPT or Claude, has introduced new possibilities for qualitative research [1,17,18]. This scalability is a key advantage over traditional qualitative analysis, which is often limited by the time-consuming nature of manual coding and thematic interpretation [19,20]. LLMs may offer a different analytical lens to qualitative data analysis by identifying patterns across large textual datasets and generating thematic insights. Their ability to support reflexivity and reveal alternative interpretive frames makes them increasingly relevant in health research involving complex patient narratives. Rather than replacing the researcher's interpretive role, the LLM is used to support reflexivity by surfacing alternative thematic groupings that could either challenge or complement the researchers' assumptions [20,21]. In addition, as such models are often trained on multilingual and culturally diverse datasets, they may help uncover insights that traditional approaches might overlook, particularly those obscured by language or cultural framing [21,22].

To examine this potential in a meaningful way, it is useful to apply LLM-assisted analysis in a well-studied domain. Chronic illness experience has been widely studied in qualitative research, and many common themes are already known. This makes it a good domain for this study, as it allows us to compare the results of LLM-assisted analysis with existing knowledge. This study aimed to evaluate the contribution of LLM-assisted analysis compared to traditional thematic analysis by examining its ability to replicate and extend known qualitative insights in chronic illness management.

Methods

Study Design and Participants

This study used thematic analysis within an interpretivist framework, guided by the Standards for Reporting Qualitative Research [23,24]. To explore recurring patterns in participants' experiences, semistructured interviews were conducted with individuals managing multiple chronic conditions, allowing for the identification of cross-cutting themes across diverse personal narratives. To investigate the concept of chronic illness management as a "career," this study followed a multiphase qualitative approach designed to integrate both experiential depth and methodological comparison. The first phase involved the development of a semistructured interview guide, informed by literature review, expert consultation, and focus groups with patients. The second phase included refinement of the guide and data collection through interviews. In the third phase, interview transcripts were analyzed using two distinct methods: manual thematic analysis and artificial intelligence (AI)–assisted analysis using an LLM. The final phase focused on a comparative evaluation of the two analytical approaches, assessing similarities, differences, and unique contributions. This phased design allowed us to explore both patient experiences and the use of LLMs in qualitative analysis, as shown in Figure 1. This study followed an inductive analytical approach. While existing literature informed the development of the interview guide, no formal theoretical framework was systematically applied in either the manual or the LLM-assisted analysis.

XSL•FO
RenderX

Figure 1. Overview of this study's design and research procedure, including the phases of interview development, data collection, and parallel analysis using manual thematic analysis and LLM-assisted analysis, followed by a comparative evaluation. LLM: large language model.

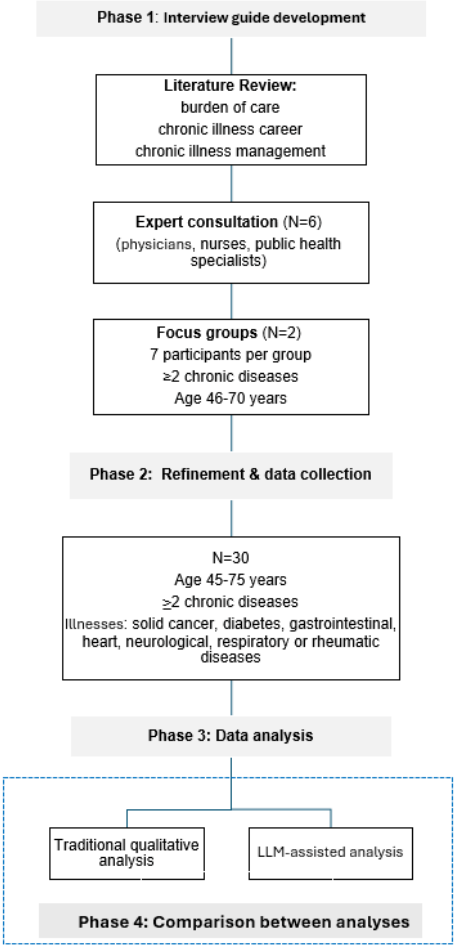

Procedure

Phase 1: Development of the Semistructured Interview Guide

A comprehensive literature review was conducted in January 2024, using the keywords: "burden of care," "chronic illness career," and "chronic illness management" to develop a conceptual framework and establish this study's initial focus. Based on the findings, guides for semistructured interviews were designed and refined with input from a panel of experts with diverse backgrounds in health care and research. To further enhance this study, two focus groups were conducted with individuals who were not part of the primary study population. These focus groups were used to test the clarity, scope, and language of the interview questions, and to ensure that the guide addressed meaningful aspects of chronic illness management from a patient perspective, including the following: participants' general background, the transition from health to illness, the "career" of managing chronic illness, balancing illness management with other life responsibilities, coping strategies, task prioritization, support systems, interactions with health care providers, and expectations for the future.

Phase 2: Refinement and Data Collection

This study was conducted over two months from February 2024 to March 2024. Participants were recruited through recommendations from their primary care physicians or other medical staff involved in their treatment. Referrals were based on clinical judgment and familiarity with the patients' chronic conditions, functional status, and communication ability. Patients unable to engage in a coherent conversation were not referred. Each patient was individually approached, informed about this study, and provided written informed consent to participate. While thematic saturation was considered, a sample of 30 participants was purposefully selected to provide sufficient narrative depth for both manual and AI-assisted analyses. This study included 30 participants, aged 45 to 75 years, who were diagnosed with at least two of the following chronic illnesses: diabetes mellitus; congestive heart failure; chronic obstructive pulmonary disease; active oncological disease (or within one year of treatment); ischemic heart disease; chronic kidney disease (grade 3 or higher); inflammatory bowel disease (Crohn disease or ulcerative colitis); interstitial lung disease; epilepsy; multiple sclerosis; autoimmune conditions such as rheumatoid arthritis, psoriatic arthritis, systemic lupus erythematosus, or systemic sclerosis (scleroderma); or those undergoing anticoagulant treatment for at least three consecutive months (warfarin, heparin, or nonvitamin K antagonist oral anticoagulants). Requiring at least two conditions ensured a minimum level of complexity in participants' health experiences. This broad inclusion reflects this study's objective to capture cross-cutting experiences in multimorbidity, rather than focusing

XSL•FO
RenderX

on a single condition. Participants underwent semistructured phone or videoconference interviews, chosen for their convenience. Interviews ranged in duration from 30 to 60 minutes, allowing participants to determine the depth and breadth of their narratives. The interview protocol included two main sections: (1) sociodemographic information: collecting data on gender, age, and primary illnesses; and (2) disease management: examining strategies, priority setting, task management, quality of life, coping techniques, and support systems. All data were manually transcribed to ensure accuracy.

Phase 3: Data Analysis

The data analysis involved two separate approaches: traditional qualitative methods and AI-assisted techniques, followed by a comparative evaluation to assess their respective insights and unique contributions.

In the traditional analysis, interview data were analyzed using an inductive thematic analysis approach, involving iterative coding, theme development, and refinement to identify key patterns across participants' experiences [24]. This analysis was not theory-driven, and themes were generated inductively from the data. Initial codes were developed and refined through an iterative process. These codes were then categorized into broader themes that reflected the primary dimensions of the patients' experiences. Shared themes were identified based on their recurrence across interviews and their conceptual relevance to this study's aims. The manual analysis was conducted by a single researcher, who maintained detailed memos and engaged in reflective journaling to enhance transparency and reduce potential interpretive bias. This approach is consistent with qualitative research practices that emphasize reflexivity and transparency in the analytical process.

For the AI-assisted analysis, the LLM was implemented using the Claude.ai application programming interface with the following specifications: version: Claude 3.5 Sonnet ("claude-3.5 Sonnet-20240620"), temperature: 0, and maximum tokens: 8192. A custom Python script (version 3.10.11) was developed to enable the system to provide automated insights on key themes within the interviews. The LLM analysis was not theory-driven and did not include theory-based prompts. The interview transcripts were analyzed in full using a predefined and structured prompt that guided the model to identify key themes, categorize relevant passages, and extract illustrative quotes. The same prompt was applied to all transcripts to ensure consistency across analyses, and the resulting themes were organized for comparison with the manual analysis.

Phase 4: Comparative Analysis

The final phase focused on comparing the themes from the traditional qualitative analysis and the LLM-assisted analysis. A structured, side-by-side comparison of the themes generated by each method was conducted, using predefined criteria, including thematic overlap, identification of unique themes, and differences in the level of detail. Themes and subthemes were categorized as convergent when they reflected similar underlying concepts, even if expressed using different terminology, and as divergent when they represented distinct insights specific to one method. This comparison helped identify both the shared patterns and unique contributions of each approach. The comparative analysis was conducted by a single researcher, supported by systematic memo-based documentation.

Ethical Considerations

This study was conducted in compliance with all relevant laws and institutional guidelines and was approved by the Tel Aviv University ethical review board. Formal ethical approval for the research was granted on February 13, 2023 (clinical trial number #0005962-1). Participation was voluntary, and informed consent was obtained from all participants before this study's initiation. Participants were fully informed about the purpose, procedures, and potential risks of this study, and they retained the right to withdraw at any time without penalty. No financial or material compensation was provided. Confidentiality was ensured through anonymization of transcripts and secure data handling.

Results

Sample Characteristics

Table 1 summarizes the baseline characteristics of this study's sample. The sample included 30 participants, 13 of whom were male, with an average age of 62.7 years (SD 7.3). The majority were either employed (n=19) or retired (n=10), with 7 of the 10 retirees attributing their retirement to medical reasons. The most common primary illnesses were diabetes (n=12) and rheumatic diseases (n=9). Most participants resided in the central region of the country (n=25), and 10 were the primary caregivers for a family member.

XSL•FO
RenderX

Table 1. Sociodemographic characteristics of this study's sample (N=30).

	Overall	Missing (%)
Male, n (%)	13 (43.3)	0.0
Employment status, n (%)		3.3
Employed	19 (63.3)	
Retired	10 (33.3)	
Unemployed	1 (3.4)	
Age (years), mean (SD)	62.7 (7.3)	0.0
Years since diagnosis, mean (SD)	11.7 (11.8)	10.0
Primary illness, n (%)		0.0
Solid cancer	2 (6.7)	
Diabetes mellitus	12 (40.0)	
Gastrointestinal disease	1 (3.3)	
Heart disease	3 (10.0)	
Neurological diseases	2 (6.7)	
Respiratory disease	1 (3.3)	
Rheumatic diseases	9 (30.0)	
Location, n (%)		0.0
Central	25 (83.3)	
Number of children, mean (SD)	2.7 (1.5)	3.3
Number of grandchildren, mean (SD)	3.3 (3.5)	20.0
Primary caregiver for a family member, n (%)	10 (33.3)	0.0

Thematic Analysis

The traditional qualitative analysis identified six main themes (navigating daily life under chronic conditions, the role of support systems, health care and systemic challenges, emotional challenges and coping, self-management and adaptation, and sources of strength and motivation) as shown in Table 2, which presents the themes identified in the manual analysis along with illustrative participant quotes. The quotes presented in Table 2 provide contextual depth and support the interpretation of the identified themes. The LLM analysis also yielded six main themes (health management and adaptation, navigating health care systems, emotional challenges and coping, social and family dynamics, daily life and functional adjustments, and future planning and long-term concerns) as shown in Table 3, which presents the themes identified in the LLM-assisted analysis.

XSL•FO
RenderX

Table 2. Themes identified in the manual thematic analysis, with illustrative quotes from participants.

	Theme	Quote
1	Navigating daily life under chronic conditions	Patient 20: "…I had about a month filled with doing a million tests; they gave me again many tests: I had to go for a urine test, a lung test, and an eye test before I arrived at the eye department. The pulmonologist wanted me to do a lung X-ray, a CT scan, and blood tests."
		Patient 4: "...I work at being chronically ill, because I have to keep going to doctors without knowing what I have. Yet again, the side effects. Each time a different side effect; sometimes it's the stomach pain, the leg pain,..."
2	The role of support systems	Patient 2: "…The community in the neighborhood where I live was very helpful and also brought food and made sure to help me at home. My sister cooked here and my mother. I learned to appreciate the power of family."
		Patient 27: "I work around the clock, sometimes it involves 30, 40, 50 and even 60 hours continuously without sleep. That is, sleeping half an hour here, half an hour there in the car in between. One of the things that I think prevents me from deteriorating is the intensity of the work. Although it is said that this is what caused the disease to break out."
3	Health care and systemic challenges	Patient 18: "When the problems with my legs started, I went to about four dermatologists in the community, and each of them gave me a completely different diagnosis. Until I got to a dermatologist at the hospital who took a biopsy and found my disease…"
		Patient 14: "When they started with the electronic prescriptions on the website, I suddenly found myself arriving at the pharmacy, waiting in line and then not having a prescription at all. It took me a while to figure out how things work... It was really one big headache..."
4	Emotional challenges and coping	Patient 6: "I added therapy to my life through gardening, work with animals. All kinds of niches that I didn't have the opportunity to do before the illness.. I try to protect my soul, try to satiate it with things that make me happy."
		Patient 15: "...Look, even when the lupus nephritis broke out, that was also worrisome. I didn't have any symptoms but it was worrying."
5	Self-management and adaptation	Patient 13: "…I don't give up on myself, once every four-six months, I write to the doctor and ask for periodic tests. I have to keep an eye on my health."
		Patient 17: "The big difference started about three months ago when the glucose was suddenly high, I went to a dietitian, and I have already changed my eating habits. I need to do some sport but I don't persist with it."
6	Sources of strength and motivation	Patient 15: "The doctors told me, 'Well, you will not return to teaching'. And my husband said to the doctor, 'Sorry but you don't know her, she will go back to teaching.' I really didn't study the first semester; I rested at home. In the second semester, I returned to teaching."
		Patient 9: "My strength comes from God. I say God help me with my hands up. And I know it will help with the disease; God is with you all the time. I am a religious woman, I believe in God."

XSL•FO
RenderX

Table 3. Themes identified in the large language model–assisted analysis, with illustrative text excerpts generated from the model outputs.

	Theme	Quote
1	Health management and adaptation	Patient 14: "…With awareness over the years, it really became part of my management, this matter of sugar. I learned to accept it…"
		Patient 7: "I'm constantly going around to doctors a lot. Many times ... Sometimes if it's close to home I can go out carefully, very very slowly, like that."
2	Navigating health care systems	Patient 20: "…I had about a month filled with doing a million tests… I had to go for a urine test, a lung test, and an eye test before I arrived at the eye department…I felt like I had been in the clinic for the whole month."
		Patient 29: "…through all these months until I got to the department... because no doctor could help me or figure out what was wrong with me."
3	Social and family dynamics	Patient 1: "Only in the last year when my hemoglobin dropped, my eldest daughter wouldn't let it go, she accompanied me, but really company? Let me be."
		Patient 13: "Without such support from the environment it's impossible, I would have died. I know I would have died. I say with certainty that without my family, it's impossible, a lonely person doesn't survive this illness."
4	Emotional challenges and coping	Patient 26: "I'm one of those who thought it wouldn't happen to me. Suddenly it happened. And yes, for me it was shocking, I'm so healthy and strong."
		Patient 14: "When I saw him and he told me about Ozempic I had tears in my eyes, he asked me, 'Why are you crying?' and I told him, 'Listen, I feel like I failed.'"
5	Daily life and functional adjustments	Patient 2: "There are priorities, I worked all my life. The house isn't a museum but it needs to be clean and there's help once a week and I don't give up on that. And now with my eyes I'm not allowed to do many things so the helper comes twice a week."
		Patient 27: "For example, during the disease attacks, I always take a bag packed with underwear and pants and everything..."
6	Future planning and long-term concerns	Patient 1: "Right now, currently, I need to protect the kidney. Because that's the most vulnerable thing, in terms of FMF as I understood. Although I understood it before, in the last year I understood even more, to protect them."
		Patient 18: "Not infrequently, I think about the diseases that I have. That can come to expression…I don't want to say it but can shorten a person's life."

The traditional analysis identified 33 subthemes, while the LLM analysis uncovered 43 subthemes. Figure 2 presents a comparison between manual and LLM-assisted analyses, showing shared themes as well as themes identified uniquely by each approach. This figure visually summarizes the similarities and differences between the two analytical approaches. Many subthemes were consistent across both groups, with only 5 subthemes identified by the traditional analysis not appearing in the LLM analysis. Conversely, 1 subtheme was identified by the LLM analysis that was absent from the traditional analysis.

XSL•FO
RenderX

Figure 2. Comparison between manual thematic analysis and LLM-assisted analysis, showing shared themes and themes identified uniquely by each approach. Shared themes are presented without subthemes, as both methods produced similar patterns at that level, while subthemes are shown only for themes that differed between the approaches. LLM: large language model.

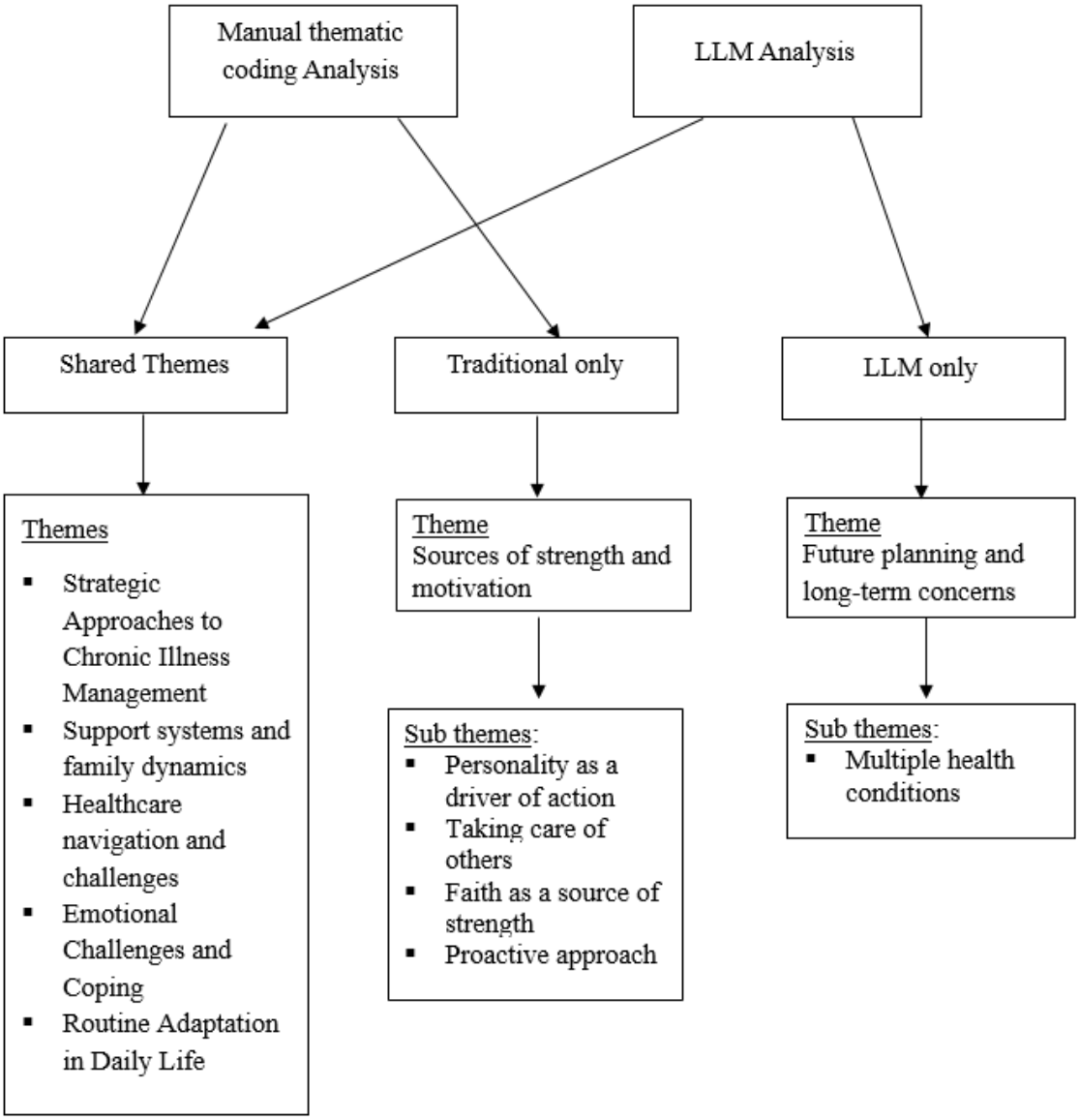

Shared Themes

The comparative analysis revealed the following key findings identified by both methods. Illustrative quotes from participants appear in Tables 2 and 3.

Strategic Approaches to Chronic Illness Management

This theme encompasses the strategies and efforts individuals use to manage and adapt to long-term health conditions. In the traditional analysis, illness management was described as an ongoing task embedded in daily routines. For example, patient 13 described actively managing his condition by monitoring and following up: “I don't give up on myself… I am writing to the doctor and ask for periodic tests. I have to keep an eye on my situation.” The LLM analysis reflected a similar focus, highlighting gradual acceptance and the integration of illness-related practices into one’s lifestyle. As patient 14 explained: “With awareness over the years, it really became part of my management… I learned to accept it.” More examples from participant narratives are presented in Tables 2 and 3.

Support Systems and Family Dynamics

This theme explores how family relationships and external support networks influence individuals’ ability to manage chronic illness. It appeared as the role of support systems in the traditional analysis and as social and family dynamics in the LLM model. Both analyses emphasized the significant role of family and community in shaping patients’ experiences, ranging from emotional encouragement and practical help to tensions or ambivalence in relationships. The findings reflect the diverse scope of support, from sincere care to feelings of burden or intrusion. For example, in the traditional analysis, patient 2 described strong family and community support: “The community in the neighborhood where I live was very helpful… I learned to appreciate the power of family.” Similarly, the LLM analysis also reflected the importance of support, as seen in patient 13’s statement: “Without such support from the

XSL•FO
RenderX

environment it's impossible… without my family, it's impossible."

Health Care Navigation and Challenges

This theme refers to the process of accessing and using health care services, and to the obstacles and frustrations patients face within the health care system. It was identified as health care and systemic challenges in the traditional analysis and as navigating health care systems in the LLM model. Participants described their journeys as bureaucratic, fragmented, and prolonged, often involving repeated referrals, inconsistent diagnoses, and delays before receiving effective care. These accounts reflect not only structural barriers but also a sense of exhaustion and uncertainty during the care-seeking process. For example, in the traditional analysis, patient 18 described the difficulty of receiving consistent diagnoses: "I went to about four dermatologists… each of them gave me a completely different diagnosis." Similarly, the LLM analysis reflected the burden of repeated tests and long processes, as described by patient 20: "I had about a month filled with doing a million tests… I felt like I had been in the clinic for the whole month."

Emotional Challenges and Coping

This theme reflects the range of emotional responses individuals experience while managing chronic illness, including ongoing worry, frustration, guilt, and self-blame. Both traditional and LLM analyses highlighted how patients described emotional strain as part of daily life with chronic conditions. These expressions, while not always accompanied by explicit coping strategies, reveal an emotional labor patients engage in as they navigate uncertainty, treatment demands, and the psychological toll of illness. For example, in the traditional analysis, patient 6 described efforts to cope emotionally: "I try to protect my soul… with things that make me happy." Similarly, the LLM analysis reflected emotional distress, as described by patient 14: "I had tears in my eyes… I feel like I failed."

Routine Adaptation in Daily Life

This theme captures the practical strategies and adjustments individuals make in their day-to-day lives to accommodate the demands of chronic illness. In the traditional analysis, it was labeled self-management and adaptation, while in the LLM model, it appeared as daily life and functional adjustments. Participants described changes to household routines, increased planning, and the gradual incorporation of external support when necessary. These narratives reflect a shift from full independence to negotiated interdependence, highlighting how people preserve autonomy while accepting limitations. For example, in the traditional analysis, patient 17 described changes in daily habits: "I have already changed my eating habits… I need to do some sport but I don't persist with it." Similarly, the LLM analysis reflected adjustments in daily routines, as described by patient 2: "There are priorities… now with my eyes I'm not allowed to do many things so the helper comes twice a week."

Exclusive Themes

The comparative analysis revealed that certain themes emerged uniquely in one of the two analytic approaches: manual thematic analysis or LLM-assisted analysis. These differences reflect variation in how each method prioritized or interpreted aspects of participants' narratives, rather than inconsistencies in accuracy.

Sources of Strength and Motivation

This theme explores the factors that empower individuals to cope with challenges and pursue their goals. The traditional analysis uniquely identified this theme, highlighting both internal and external sources of motivation. Internal sources included personal faith, spiritual beliefs, and a proactive mindset that emphasized inner strength and optimism. External sources stemmed from caring responsibilities, encouragement from family members, and the need to maintain functional roles within the household.

These accounts illustrate how participants found motivation both within themselves and through their roles in the family and community. For example, patient 9 described a personal source of strength: "My strength comes from God… I believe in God."

Future Planning and Long-Term Concerns

The LLM analysis uniquely identified this theme, which reflects participants' concerns about the future and the strategies they use to prepare for potential long-term challenges. These included considerations such as financial stability, medical planning, and personal life goals. Although such reflections appeared in the traditional analysis as well, they were embedded within a broader subtheme of emotional challenges and coping, rather than being treated as a distinct theme. The LLM analysis, however, grouped these expressions, emphasizing their frequency and framing them as a separate area of concern regarding future planning. For example, patient 18 described concerns about the future: "I think about the diseases that I have… it can shorten a person's life."

Discussion

Principal Findings

This study compared traditional qualitative thematic analysis with AI-based analysis using LLM to explore patients' experiences in managing chronic illness. The analysis revealed 5 shared themes: strategic approaches to chronic illness management, support systems and family dynamics, health care navigation and challenges, emotional challenges and coping, and routine adaptation in daily life. In addition, each method revealed unique themes: the traditional analysis identified sources of strength and motivation, encompassing subthemes such as personality, caregiving roles, faith, and a proactive mindset. The LLM analysis uniquely identified future planning and long-term concerns, including issues related to multiple health conditions. These findings highlight the complementary roles of traditional and LLM-assisted analysis and contribute to the emerging methodological literature on AI-assisted qualitative research, which explores how LLMs can support, extend, and challenge traditional approaches to qualitative analysis. These findings are discussed in more detail in the following sections, focusing on both shared and unique themes.

XSL•FO
RenderX

Themes Shared by The Two Qualitative Methods

The overlap between themes identified by the traditional and LLM-based analyses suggests that both approaches capture core aspects of chronic illness management, consistent with existing literature on chronic care and patient-centered management [25-27]. This indicates that, despite differences in the analytical process, both methods provide a meaningful understanding of the patient's experience and highlight their complementary strengths. For example, health care navigation and challenges were a recurring theme in both approaches, with patients describing the bureaucratic hurdles and frustrations of accessing care. Studies on health care navigation for chronic patients highlight several common issues, such as fragmented care, inconsistent communication among providers, and a lack of standardized care pathways [28]. These challenges often result in delays in receiving accurate diagnoses and appropriate treatments, as patients navigate a complex health care system [29]. Furthermore, patients with multiple chronic conditions often receive conflicting medical advice and face barriers such as long wait times and limited access to care [30-33]. Together, these findings suggest that both analytical approaches can capture key systemic challenges experienced by patients with chronic illness.

The importance of support systems and family dynamics across both approaches aligns with previous research showing the central role of social support in chronic illness management [34,35]. Participants described the emotional, logistical, and practical support provided by family members, friends, and community networks, which helped them maintain a sense of balance and resilience amid the demands of their chronic conditions. The support systems and family dynamics theme aligns with research emphasizing the importance of family caregivers and social support in improving health outcomes and quality of life for individuals with chronic illnesses [36]. Social support has been shown to provide not only practical assistance but also a crucial psychological buffer against the stress and uncertainty of managing chronic conditions [37]. By fostering a strong support system, patients can enhance their overall quality of life and better cope with the physical and emotional challenges of living with chronic illness [38]. However, while social support is invaluable, it also comes with its complexities. Family dynamics may introduce challenges such as caregiver burnout, role strain, or feelings of dependence among patients [39,40], highlighting the importance of developing interventions that support both patients and caregivers [41,42]. Together, these findings suggest that both analytical approaches capture not only the benefits of social support but also its challenges, reflecting the complexity of patients' lived experiences.

The emotional challenges and coping impact of chronic illness were evident across both analytical approaches, highlighting the central role of emotional challenges in patients' experiences. Participants described the ongoing struggle to maintain a sense of normalcy and control amid the uncertainty of living with chronic conditions, as well as the impact on their mental health, including feelings of depression, anxiety, and grief over the loss of their previous way of life [36,43]. These findings align with existing research demonstrating the psychological burden associated with chronic illness, including an increased risk of mental health issues such as depression and anxiety, which can exacerbate physical symptoms and hinder effective self-management [36,44,45]. To cope, patients described a range of strategies, such as cultivating a positive mindset, engaging in self-care, and seeking support from professionals and peer networks [46,47]. Addressing these emotional needs is critical, as interventions targeting mental health can significantly improve overall health outcomes [48]. Together, these findings suggest that both analytical approaches capture not only the emotional burden of chronic illness but also the coping strategies patients use to manage these challenges.

Finally, chronic illness management, as reflected in both analytical approaches, encompasses a wide range of practices and experiences that can be viewed through both positive and challenging lenses. On the one hand, disease management can serve as a source of education and empowerment, enabling patients to take an active role in their care through self-monitoring and lifestyle adjustments, which may improve autonomy and health outcomes [49]. On the other hand, the burden of chronic illness management can be overwhelming, leading to burnout, reduced quality of life, and ongoing challenges in navigating the health care system [50]. Together, these findings highlight the complex and sometimes conflicting nature of chronic illness management, as experienced by patients. However, both analyses focused on identifying themes rather than applying a specific theoretical framework, and therefore remain mainly descriptive.

Themes Exclusive to Only One Qualitative Method

The fact that the traditional analysis uniquely identified the theme of sources of strength and motivation suggests its strength in capturing more nuanced and deeply personal aspects of patient experience. This theme encompasses subthemes such as personality as a driver of action, taking care of others, faith as a source of strength, and a proactive approach. Participants described how personal beliefs, a sense of purpose, and relationships influenced their ability to manage their condition, such as patient 15's connection to their professional identity. These findings are consistent with previous research showing that resilience, self-efficacy, and spiritual beliefs can play a central role in chronic illness management and are associated with improved coping and health outcomes [51-53]. Taking care of others can also strengthen patients' ability to navigate care and maintain meaningful roles [54]. In addition, a positive mindset has been associated with better self-care behaviors and disease outcomes. These findings also reflect the role of personality traits, such as resilience and self-efficacy, in driving active engagement in illness management, supporting the subtheme of personality as a driver of action [52,55]. The subtheme of a proactive approach further reflects patients' expectations for more active and supportive care. Participants emphasized the need for health care providers to take a more active role, such as scheduling follow-up appointments and maintaining ongoing engagement, rather than relying on patients to initiate contact. This finding aligns with existing literature on chronic disease management, which emphasizes the importance of planned and coordinated care that anticipates and addresses patient needs [49,56]. At the same time, this approach may not suit all patients, as some may perceive it as limiting

XSL•FO
RenderX

their autonomy or creating additional pressure, highlighting the need to balance proactive care with respect for patients' preferences and independence [57]. Together, these findings suggest that the traditional analysis can capture not only patient experiences, but also deeper, context-dependent aspects such as personal motivations, values, and expectations regarding care, which may be less evident in more structured analytical approaches.

The identification of multiple health conditions as a distinct subtheme highlights the ability of the LLM-assisted analysis to capture the complexity and accumulation of challenges faced by patients managing more than one chronic illness. This subtheme reflects the combined burden of multiple conditions, each with its own symptoms, treatments, and impact on daily life. As described by patient 7, "It's not just one thing, it's everything together. I have diabetes, high blood pressure, arthritis, and now this new thing with my heart. It's just too much sometimes." Managing multiple chronic illnesses often requires navigating a fragmented health care system, coordinating care among multiple providers, and following complex and sometimes conflicting treatment plans [58]. Patients may also experience a compounding effect, where one condition exacerbates the symptoms or management challenges of another, increasing the overall burden and complexity of care [59]. These findings suggest that LLM-assisted analysis may be particularly effective in identifying patterns related to complexity, accumulation, and co-occurring conditions within patient narratives. By highlighting these patterns, this approach may support a more system-level understanding of chronic illness management.

Overall, this study highlights the complementary value of traditional and LLM-assisted qualitative analysis in understanding chronic illness management. While traditional analysis provided deeper insight into personal meanings, motivations, and patient perspectives, the LLM-assisted approach was effective in identifying broader patterns and capturing the complexity of managing multiple conditions. Together, these findings suggest that integrating both approaches can enhance the depth, scope, and applicability of qualitative health research, particularly in complex clinical contexts.

Study Limitations and Future Directions

This study provides a valuable comparison between traditional qualitative analysis and AI-based analysis using Claude 3.5 Sonnet. However, this reliance on a single LLM constitutes a key limitation, as it does not account for potential variations in performance and insights that may arise from using other LLMs, such as ChatGPT, Gemini (Google LLC), or other emerging systems. Future studies should explore a comparative evaluation across multiple LLMs to better understand the range of capabilities, biases, and thematic coverage. Another limitation relates to the dataset and its contextual dependency. While the dataset was sufficient for this analysis, its characteristics may have influenced the themes identified by both methods. Broader and more diverse datasets could support more comprehensive comparisons. Participant recruitment through primary care physicians may have introduced selection bias, as individuals who are more communicative or more engaged with the health care system may have been more likely to be referred to. This may have influenced the range of perspectives represented in the sample. The comparison between the two approaches was based on predefined qualitative criteria and remained primarily interpretive and descriptive, without formal quantitative evaluation. In addition, the traditional qualitative analysis relied on a single researcher, which may have increased the risk of interpretive bias, although reflexive practices were used to mitigate this. Another limitation is that both the manual and LLM analyses were not theory-driven, resulting in findings that are primarily descriptive rather than theoretically grounded. Future research should explore how theoretical frameworks can be integrated into both manual and LLM-assisted analysis. Similarly, while LLMs offer scalability and efficiency, their output depends on the training data, which introduces biases or overlooks important contextual nuances. Moreover, the use of an LLM raises ethical considerations, including transparency, interpretability, and the potential for misrepresentation of participants' voices. In this study, the model operated solely on anonymized transcripts and was used to support, rather than replace, human interpretation. Finally, this study did not assess the reliability of LLM-generated themes or their clinical applicability to real-world scenarios. Therefore, these findings should be interpreted as exploratory thematic observations rather than clinically validated psychosocial classifications. Future research should integrate mixed-method approaches to evaluate the validity, reproducibility, and practical relevance of themes generated by LLMs. More broadly, future studies should aim to integrate AI-driven tools into qualitative research designs in ways that enhance both efficiency and analytical depth. The development of hybrid frameworks that use LLM-based pattern recognition with human-led contextual interpretation represents a promising direction. Additionally, longitudinal research and real-world applications will be important for assessing the stability and clinical relevance of LLM-generated insights.

Conclusions

This study examined the use of traditional qualitative thematic analysis and LLM-based analysis to explore patients' experiences in managing chronic illness. Both approaches identified similar core themes, while also revealing distinct insights.

These findings suggest that the two approaches offer complementary perspectives. Traditional analysis enabled a rich, interpretive understanding of patient narratives, while the LLM approach supported efficient and systematic identification of thematic structures. While AI-based methods hold potential to support qualitative research—particularly in scaling and organizing data—they require careful validation to ensure cultural and contextual sensitivity. Taken together, this study highlights the value of a hybrid approach, integrating both human insight and AI capabilities, to enhance the depth and scope of qualitative health research.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank Margherita Brus, PhD, for her valuable guidance and assistance throughout the research process. Generative artificial intelligence tools were used to support language editing and clarity of this paper. In addition, a large language model was used for data analysis, as described in the Methods section. All analysis decisions, interpretations, and conclusions were made by the authors. This work was performed in partial fulfillment of the requirements for a PhD degree of SK, School of Medicine, Faculty of Medical and Health Sciences, Tel Aviv University, Israel.

Data Availability

The datasets used and/or analyzed during this study are available from the corresponding author on reasonable request.

Funding

This study was supported by the Ela Kodesz Institute for Technologies in Healthcare. The funder had no role in this study's design, data collection, analysis, or interpretation.

Authors' Contributions

SK was the principal investigator and was responsible for manuscript writing, data analysis, and structuring this study. YMB and RN reviewed the manuscript, contributed to the conceptual framework, and supervised this study. MS assisted with data analysis, contributed to the conceptual framework, supervised this study, and reviewed the manuscript. All authors read and approved this paper.

Conflicts of Interest

None declared.

References

1. Grady PA, Gough LL. Self-management: a comprehensive approach to management of chronic conditions. Am J Public Health 2014;104(8):e25-e31. [doi: 10.2105/AJPH.2014.302041] [Medline: 24922170]
2. Parekh AK, Goodman RA, Gordon C, Koh HK, HHS Interagency Workgroup on Multiple Chronic Conditions. Managing multiple chronic conditions: a strategic framework for improving health outcomes and quality of life. Public Health Rep 2011;126(4):460-471 [FREE Full text] [doi: 10.1177/003335491112600403] [Medline: 21800741]
3. Price B. Illness careers: the chronic illness experience. J Adv Nurs 1996;24(2):275-279. [doi: 10.1046/j.1365-2648.1996.02047.x] [Medline: 8858430]
4. Anderson RA, Bailey DE, Wu B, Corazzini K, McConnell ES, Thygeson NM, et al. Adaptive leadership framework for chronic illness: framing a research agenda for transforming care delivery. ANS Adv Nurs Sci 2015;38(2):83-95 [FREE Full text] [doi: 10.1097/ANS.0000000000000063] [Medline: 25647829]
5. Clark NM. Management of chronic disease by patients. Annu Rev Public Health 2003;24:289-313. [doi: 10.1146/annurev.publhealth.24.100901.141021] [Medline: 12415147]
6. Anderson R, Bury M. Living with Chronic Illness: The Experience of Patients and Their Families. London, UK: Taylor & Francis; 2024.
7. Von Korff M, Gruman J, Schaefer J, Curry SJ, Wagner EH. Collaborative management of chronic illness. Ann Intern Med 1997;127(12):1097-1102. [doi: 10.7326/0003-4819-127-12-199712150-00008] [Medline: 9412313]
8. Packer T, Kephart G, Audulv, Keddy A, Warner G, Peacock K, et al. Protocol for development, calibration and validation of the patient-reported inventory of self-management of chronic conditions (PRISM-CC). BMJ Open 2020;10(9):e036776 [FREE Full text] [doi: 10.1136/bmjopen-2020-036776] [Medline: 32998919]
9. Skjutar Å, Müllersdorf M. Adapt, discover and engage: a qualitative interview study with patients living with chronic pain. J Nurs Healthcare 2010;2(4):254-261. [doi: 10.1111/j.1752-9824.2010.01066.x]
10. Cheng C, Inder K, Chan SWC. Patients' experiences of coping with multiple chronic conditions: a meta-ethnography of qualitative work. Int J Ment Health Nurs 2019;28(1):54-70. [doi: 10.1111/inm.12544] [Medline: 30232822]
11. McLellan E, MacQueen KM, Neidig JL. Beyond the qualitative interview: data preparation and transcription. Field Methods 2003;15(1):63-84. [doi: 10.1177/1525822x02239573]
12. Kwasnicki RM, Noakes AJ, Banhidy N, Hettiaratchy S. Quantifying the limitations of clinical and technology-based flap monitoring strategies using a systematic thematic analysis. Plast Reconstr Surg Glob Open 2021;9(7):e3663 [FREE Full text] [doi: 10.1097/GOX.0000000000003663] [Medline: 34262835]
13. Speed R. Regression type techniques and small samples: a guide to good practice. J Mark Manage 2010;10(1-3):89-104. [doi: 10.1080/0267257x.1994.9964262]
14. Schonfeld IS, Mazzola JJ. Strengths and limitations of qualitative approaches to research in occupational health psychology. In: Research Methods in Occupational Health Psychology. New York, NY: Routledge; 2012:268-289.

XSL•FO
RenderX

15. Wiener DN. Empirical occupational groupings of Kuder Preference Record profiles. Educ Psychol Meas 1951;11(2):273-279. [doi: 10.1177/001316445101100211]
16. Ryan GW, Bernard HR. Techniques to identify themes. Field Methods 2003;15(1):85-109. [doi: 10.1177/1525822x02239569]
17. Schaink AK, Kuluski K, Lyons RF, Fortin M, Jadad AR, Upshur R, et al. A scoping review and thematic classification of patient complexity: offering a unifying framework. J Comorb 2012;2:1-9 [FREE Full text] [doi: 10.15256/joc.2012.2.15] [Medline: 29090137]
18. Sav A, Salehi A, Mair FS, McMillan SS. Measuring the burden of treatment for chronic disease: implications of a scoping review of the literature. BMC Med Res Methodol 2017;17(1):140 [FREE Full text] [doi: 10.1186/s12874-017-0411-8] [Medline: 28899342]
19. Törnberg P. How to use LLMs for text analysis. arXiv. Preprint posted online on July 24, 2023 . [doi: 10.48550/arXiv.2307.13106]
20. Castleberry A, Nolen A. Thematic analysis of qualitative research data: is it as easy as it sounds? Curr Pharm Teach Learn 2018;10(6):807-815. [doi: 10.1016/j.cptl.2018.03.019] [Medline: 30025784]
21. Dwivedi YK, Kshetri N, Hughes L, Slade EL, Jeyaraj A, Kar AK, et al. Opinion paper: “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int J Inf Manage 2023;71:102642. [doi: 10.1016/j.ijinfomgt.2023.102642]
22. Bender E, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? 2021 Presented at: FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; March 3-10, 2021; Virtual Event Canada p. 610-623. [doi: 10.1145/3442188.3445922]
23. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
24. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2008;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
25. Schaefer J, Davis C. Case management and the chronic care model: a multidisciplinary role. Lippincotts Case Manag 2004;9(2):96-103. [doi: 10.1097/00129234-200403000-00007] [Medline: 15192504]
26. Sminkey P. Beyond the care episode: patient-centered case management along the continuum. Prof Case Manag 2012;17(4):186-187. [doi: 10.1097/NCM.0b013e31825886b1] [Medline: 22660341]
27. Berry AB, Lim CY, Hirsch T, Hartzler AL, Kiel LM, Bermet ZA, et al. Supporting communication about values between people with multiple chronic conditions and their providers. 2019 Presented at: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; May 2-9, 2019; Glasgow, Scotland p. 2-9. [doi: 10.1145/3290605.3300700]
28. Bhangu S, Provost F, Caduff C. Introduction to qualitative research methods - part I. Perspect Clin Res 2023;14(1):39-42 [FREE Full text] [doi: 10.4103/picr.picr_253_22] [Medline: 36909216]
29. Gui X, Chen Y, Pine KH. Navigating the healthcare service "Black Box": individual competence and fragmented system. Proc ACM Hum-Comput Interact 2018 Jan;2(CSCW):1-26. [doi: 10.1145/3274330]
30. Spiers G, Boulton E, Corner L, Craig D, Parker S, Todd C, et al. What matters to people with multiple long-term conditions and their carers? Postgrad Med J 2023;99(1169):159-165 [FREE Full text] [doi: 10.1136/postgradmedj-2021-140825] [Medline: 37222056]
31. Schwarz T, Schmidt AE, Bobek J, Ladurner J. Barriers to accessing health care for people with chronic conditions: a qualitative interview study. BMC Health Serv Res 2022;22(1):1037 [FREE Full text] [doi: 10.1186/s12913-022-08426-z] [Medline: 35964086]
32. Safford MM, Allison JJ, Kiefe CI. Patient complexity: more than comorbidity. the vector model of complexity. J Gen Intern Med 2007;22 Suppl 3(Suppl 3):382-390 [FREE Full text] [doi: 10.1007/s11606-007-0307-0] [Medline: 18026806]
33. Hudon C, Bisson M, Chouinard M, Delahunty-Pike A, Lambert M, Howse D, et al. Implementation analysis of a case management intervention for people with complex care needs in primary care: a multiple case study across Canada. BMC Health Serv Res 2023;23(1):377 [FREE Full text] [doi: 10.1186/s12913-023-09379-7] [Medline: 37076851]
34. Sherman BW. Management of individuals with multiple chronic conditions: a continuing challenge. Am J Manag Care 2021;27(6):256-260 [FREE Full text] [doi: 10.37765/ajmc.2021.88665] [Medline: 34156219]
35. Franklin M, Lewis S, Willis K, Rogers A, Venville A, Smith L. Goals for living with a chronic condition: the relevance of temporalities, dispositions, and resources. Soc Sci Med 2019;233:13-20. [doi: 10.1016/j.socscimed.2019.05.031] [Medline: 31150928]
36. Jeon Y, Jowsey T, Yen L, Glasgow NJ, Essue B, Kljakovic M, et al. Achieving a balanced life in the face of chronic illness. Aust J Prim Health 2010;16(1):66-74. [doi: 10.1071/py09039] [Medline: 21133301]
37. Gallant MP. The influence of social support on chronic illness self-management: a review and directions for research. Health Educ Behav 2003;30(2):170-195. [doi: 10.1177/1090198102251030] [Medline: 12693522]
38. Vassilev I, Rogers A, Kennedy A, Koetsenruijter J. The influence of social networks on self-management support: a metasynthesis. BMC Public Health 2014;14:719 [FREE Full text] [doi: 10.1186/1471-2458-14-719] [Medline: 25023948]
39. Belil FE, Alhani F, Ebadi A, Kazemnejad A. Self-efficacy of people with chronic conditions: a qualitative directed content analysis. J Clin Med 2018;7(11):411 [FREE Full text] [doi: 10.3390/jcm7110411] [Medline: 30400316]

XSL•FO
RenderX

40. Golics CJ, Basra MKA, Salek MS, Finlay AY. The impact of patients' chronic disease on family quality of life: an experience from 26 specialties. Int J Gen Med 2013;6:787-798 [FREE Full text] [doi: 10.2147/IJGM.S45156] [Medline: 24092994]
41. Brignon M, Vioulac C, Boujut E, Delannoy C, Beauvais C, Kivits J, et al. Patients and relatives coping with inflammatory arthritis: care teamwork. Health Expect 2020;23(1):137-147 [FREE Full text] [doi: 10.1111/hex.12982] [Medline: 31774612]
42. Kim E, Baskys A, Law AV, Roosan MR, Li Y, Roosan D. Scoping review: the empowerment of Alzheimer's disease caregivers with mHealth applications. NPJ Digit Med 2021;4(1):131 [FREE Full text] [doi: 10.1038/s41746-021-00506-4] [Medline: 34493819]
43. Schrag A, Hovris A, Morley D, Quinn N, Jahanshahi M. Caregiver-burden in Parkinson's disease is closely associated with psychiatric symptoms, falls, and disability. Parkinsonism Relat Disord 2006;12(1):35-41. [doi: 10.1016/j.parkreldis.2005.06.011] [Medline: 16271496]
44. Rogers ML, Joiner TE, Shahar G. Suicidality in chronic illness: an overview of cognitive-affective and interpersonal factors. J Clin Psychol Med Settings 2021;28(1):137-148. [doi: 10.1007/s10880-020-09749-x] [Medline: 33128664]
45. Moussavi S, Chatterji S, Verdes E, Tandon A, Patel V, Ustun B. Depression, chronic diseases, and decrements in health: results from the world health surveys. Lancet 2007;370(9590):851-858. [doi: 10.1016/S0140-6736(07)61415-9] [Medline: 17826170]
46. Fennell PA, Dorr N, George SS. Elements of suffering in myalgic encephalomyelitis/chronic fatigue syndrome: the experience of loss, grief, stigma, and trauma in the severely and very severely affected. Healthcare (Basel) 2021;9(5):553 [FREE Full text] [doi: 10.3390/healthcare9050553] [Medline: 34065069]
47. Shamir-Stein N, Feldblum I, Rotman E, Cohen S, Brand E, Kivity S, et al. Exploring the self-reported physical and psychological effects in a population exposed to a regional conflict. J Community Health 2024;49(4):674-681. [doi: 10.1007/s10900-024-01337-6] [Medline: 38393653]
48. Katon WJ. Clinical and health services relationships between major depression, depressive symptoms, and general medical illness. Biol Psychiatry 2003;54(3):216-226. [doi: 10.1016/s0006-3223(03)00273-7] [Medline: 12893098]
49. Butler J, Longhitano C. Self-harm. Medicine 2008;36(9):455-458. [doi: 10.1016/j.mpmed.2008.06.008]
50. Von Korff M, Glasgow RE, Sharpe M. Organising care for chronic illness. BMJ 2002;325(7355):92-94 [FREE Full text] [doi: 10.1136/bmj.325.7355.92] [Medline: 12114242]
51. Javors JR, Bramble JE. Uncontrolled chronic disease: patient non-compliance or clinical mismanagement? Dis Manag 2003;6(3):169-178. [doi: 10.1089/109350703322425518] [Medline: 14570385]
52. Sevick MA, Trauth JM, Ling BS, Anderson RT, Piatt GA, Kilbourne AM, et al. Patients with complex chronic diseases: perspectives on supporting self-management. J Gen Intern Med 2007;22 Suppl 3(Suppl 3):438-444 [FREE Full text] [doi: 10.1007/s11606-007-0316-z] [Medline: 18026814]
53. Yi JP, Vitaliano PP, Smith RE, Yi JC, Weinger K. The role of resilience on psychological adjustment and physical health in patients with diabetes. Br J Health Psychol 2008;13(Pt 2):311-325 [FREE Full text] [doi: 10.1348/135910707X186994] [Medline: 17535497]
54. Büssing A, Michalsen A, Balzat H, Grünther RA, Ostermann T, Neugebauer EAM, et al. Are spirituality and religiosity resources for patients with chronic pain conditions? Pain Med 2009;10(2):327-339. [doi: 10.1111/j.1526-4637.2009.00572.x] [Medline: 19284487]
55. de Ridder D, Geenen R, Kuijer R, van Middendorp H. Psychological adjustment to chronic disease. Lancet 2008;372(9634):246-255. [doi: 10.1016/S0140-6736(08)61078-8] [Medline: 18640461]
56. Kristjansdottir OB, Stenberg U, Mirkovic J, Krogseth T, Ljoså TM, Stange KC, et al. Personal strengths reported by people with chronic illness: a qualitative study. Health Expect 2018;21(4):787-795 [FREE Full text] [doi: 10.1111/hex.12674] [Medline: 29478260]
57. National Academies of Sciences, Engineering, and Medicine, Health and Medicine Division, Board on Health Care Services, Board on Global Health, Committee on Improving the Quality of Health Care Globally. Crossing the Global Quality Chasm: Improving Health Care Worldwide. Washington, DC: The National Academies Press; 2018.
58. Keuning-Plantinga A, Stoffels J, Roodbol PF, Finnema EJ, Van Munster BC. Involvement, topics, and roles of nurses in shared decision-making with patients with dementia in acute hospitals: an integrative review. Nurs Open 2023;10(6):3519-3532 [FREE Full text] [doi: 10.1002/nop2.1618] [Medline: 36700922]
59. Ongwere T, Cantor G, Clawson J, Shih P, Connelly K. Design and care for discordant chronic comorbidities: a comparison of healthcare providers' perspectives. 2000 Presented at: PervasiveHealth '20: Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare; May 18-20, 2000; Atlanta, GA p. 133-145. [doi: 10.1145/3421937.3422013]

Abbreviations

AI: artificial intelligence
LLM: large language model

XSL•FO
RenderX

Edited by A Stone; submitted 29.Nov.2025; peer-reviewed by T Greenhalgh, Z Li; comments to author 16.Apr.2026; revised version received 22.May.2026; accepted 26.May.2026; published 29.Jun.2026.

Please cite as:
Kivity S, Barilan YM, Noham R, Saban M
Using a Large Language Model to Support Thematic Analysis of Patient Experiences in Chronic Illness Management: Comparative Qualitative Study
J Med Internet Res 2026;28:e88677
URL: https://www.jmir.org/2026/1/e88677
doi:10.2196/88677
PMID:

©Sara Kivity, Yechiel Michael Barilan, Reut Noham, Mor Saban. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessing the Evolution and Influence of Medical Open Databases on Biomedical Research and Health Care Innovation: A 25-Year Perspective With a Focus on Privacy and Privacy-Enhancing Technologies

Albert Yang[1,2,3], MD, PhD; Mei-Lien Pan[4], PhD; Henry Horng-Shing Lu[5,6,7,8,9], PhD; Chung-Yueh Lien[10], PhD; Da-Wei Wang[11†], PhD; Chih-Hsiung Chen[12], PhD; Der-Cherng Tarng[2,13,14,15], MD, PhD; Dau-Ming Niu[2,14,16], MD, PhD; Shih-Hwa Chiou[3,17], MD, PhD; Chun-Ying Wu[18,19], MD, PhD; Ying - Chou Sun[20], MD; Shih-Ann Chen[2,21], MD; Shuu-Jiun Wang[2,22,23], MD; Wayne Huey-Herng Sheu[24], MD, PhD; Chi-Hung Lin[25], MD, PhD

[1]Digital Medicine and Smart Healthcare Research Center, National Yang Ming Chiao Tung University, No. 155 Sec. 2 Linong St., Beitou Dist, Taipei City, Taiwan
[10]Department of Information Management, National Taipei University of Nursing and Health Sciences, Taipei, Taiwan
[11]Institute of Information Science, Academia Sinica, Taipei, Taiwan
[12]Institute of Technology Law, National Yang Ming Chiao Tung University, Taipei, Taiwan
[13]Department of Medicine, Taipei Veterans General Hospital, Taipei, Taiwan
[14]Institute of Clinical Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan
[15]Department and Institute of Physiology, National Yang Ming Chiao Tung University, Taipei, Taiwan
[16]Department of Pediatrics, Taipei Veterans General Hospital, Taipei, Taiwan
[17]Institute of Pharmacology, College of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan
[18]Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, Taipei, Taiwan
[19]Health Innovation Center, National Yang Ming Chiao Tung University, Taipei, Taiwan
[2]School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan
[20]Department of Radiology, Taipei Veterans General Hospital, Taipei, Taiwan
[21]Department of Cardiovascular Center and Medical Research, Taichung Veterans General Hospital, Taichung, Taiwan
[22]Department of Neurology, Neurological Institute, Taipei Veterans General Hospital, Taipei, Taiwan
[23]Brain Research Center, National Yang Ming Chiao Tung University, Taipei, Taiwan
[24]Institute of Molecular and Genomic Medicine, National Health Research Institutes, Taipei, Taiwan
[25]Department of Biological Science & Technology, National Chiao Tung University, Taipei, Taiwan
[3]Department of Medical Research, Taipei Veterans General Hospital, Taipei, Taiwan
[4]Institute of Hospital and Health Care Administration, National Yang Ming Chiao Tung University, Taipei, Taiwan
[5]Institute of Statistics, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
[6]Biomedical Artificial Intelligence Academy, Kaohsiung Medical University, Kaohsiung, Taiwan
[7]Department of Artificial Intelligence in Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan
[8]Department of Medical Research, Kaohsiung Medical University Hospital, Kaohsiung, Taiwan
[9]Department of Statistics and Data Science, Cornell University, Ithaca, NY, United States
[†]deceased

Corresponding Author:
Albert Yang, MD, PhD
Digital Medicine and Smart Healthcare Research Center, National Yang Ming Chiao Tung University, No. 155 Sec. 2 Linong St., Beitou Dist, Taipei City, Taiwan

Abstract

The integration of medical open databases with artificial intelligence (AI) technologies marks a transformative era in biomedical research and health care innovation. Over the past 25 years, initiatives like PhysioNet have revolutionized data access, fostering unprecedented levels of collaboration and accelerating medical discoveries. This rise of medical open databases presents challenges, particularly in harmonizing research enablement with patient confidentiality. In response, privacy laws such as the Health Insurance Portability and Accountability Act have been established, and privacy-enhancing technologies have been adopted to maintain this delicate balance. Privacy-enhancing technologies, including differential privacy, secure multiparty computation, and notably,

XSL•FO
RenderX

federated learning (FL), have become instrumental in safeguarding personal health information. FL, in particular, represents a significant advancement by enabling the development and training of AI models on decentralized data. In Taiwan, significant strides have been made in aligning with these global data-sharing and privacy standards. We have actively promoted the sharing of medical data through the development of dynamic consent systems. These systems enable individuals to control and adjust their data-sharing preferences, ensuring transparency and continuity of consent in the ever-evolving landscape of digital health. Despite the challenges associated with privacy protections, the benefits, including improved diagnostics and treatment, are substantial. The availability of open databases has notably accelerated AI research, leading to significant advancements in medical diagnostics and treatments. As the landscape of health care research continues to evolve with open science and FL, the role of medical open databases remains crucial in shaping the future of medicine, promising enhanced patient outcomes and fostering a global research community committed to ethical integrity and privacy.

(J Med Internet Res 2026;28:e58954) doi:10.2196/58954

KEYWORDS

medical open databases; artificial intelligence; privacy-enhancing technologies; federated learning; dynamic consent systems

Introduction

The emergence of medical open databases, coupled with advances in artificial intelligence (AI), heralds a significant change in biomedical research and health care innovation, facilitating an era of enhanced accessibility and data sharing [1-3]. This movement toward open data science, augmented by AI technologies, enables researchers worldwide to access a wealth of data, including physiological signals [4,5], genomic [6], and health care information [7], and, most prominently, large-scale medical imaging archives. While this review covers the broad spectrum of medical data, the impact of open imaging databases has been particularly transformative for the application of AI. This movement toward open data science fosters collaboration and speeds up the pace of medical discoveries.

AI's role in analyzing vast datasets has been instrumental in uncovering patterns and insights that would be impossible for humans to detect unaided, leading to breakthroughs in understanding diseases and patient care. Initiatives like annual challenges and shared toolboxes have spurred the development of novel algorithms and techniques, leveraging AI to address complex biomedical challenges and advance medical diagnostics and treatments. This synergy between open medical databases and AI is transforming the landscape of health care, promising a future of more accurate, efficient, and personalized medicine.

Simultaneously, this rise in open data repositories brings to the forefront crucial privacy concerns [6,8]. The necessity to balance the imperative of research enablement with the protection of patient confidentiality has never been more pronounced. In this context, laws such as the Health Insurance Portability and Accountability Act (HIPAA) play a pivotal role in shaping the landscape of data deidentification and anonymization processes, ensuring that shared data comply with strict privacy standards [9]. Moreover, the introduction of privacy-enhancing technologies (PETs), such as differential privacy [10], synthetic data [11], homomorphic encryption [12], secure multiparty computation [13], and federated learning [14], represents a proactive approach to safeguarding personal health information. These technologies provide the means to conduct meaningful research while upholding the principles of data privacy and security.

In a country like Taiwan, strides in medical data sharing suggest the global shift toward interconnected health systems, highlighting both advancements and ongoing challenges in securing patient data. The implementation of dynamic consent frameworks reflects a growing recognition of the need for more flexible approaches to data privacy, particularly in an era of personalized medicine and digital health records [15]. As the landscape of medical research evolves with these developments, the interplay of data sharing, privacy, and technology continues to reshape the boundaries of what is possible in health care innovation, marking a critical junction in the journey toward more open, collaborative, and ethically responsible research environments.

This review aims to provide a 25-year perspective on the evolution of medical open databases, tracing their impact on biomedical research and health care innovation, and to examine how emerging PETs and data-governance frameworks, including dynamic consent systems, shape the ethical, technical, and collaborative landscape of digital medicine worldwide.

The Rise of Medical Open Database

The rise of medical open databases represents a transformative shift in the landscape of biomedical research and health care innovation. Among the pioneers in this movement is PhysioNet. Established in 1999, PhysioNet is a pioneering open database that provides free access to a wide range of physiologic signals and related open-source software for research in medicine, physiology, and biomedical engineering [4,16]. It was initiated by a collaborative project involving researchers from Boston's Beth Israel Deaconess Medical Center/Harvard Medical School, Boston University, McGill University, and Massachusetts Institute of Technology [4,17]. The database contains a diverse collection of physiological datasets, including those related to cardiovascular and other complex biomedical signals [4,18]. PhysioNet has had a significant impact on the development of medical open databases, serving as a model for the establishment of similar resources. It has also played a key role in promoting the dissemination and exchange of medical resources.

A significant contribution of PhysioNet to the scientific community is its annual PhysioNet Challenge, which has markedly influenced the field by promoting innovation and collaboration among researchers and clinicians. These challenges

XSL•FO
RenderX

stimulate the creation of novel algorithms and methods aimed at solving complex biomedical problems, thus expanding the limits of what can be achieved in medical data analysis and application. For instance, the challenges have catalyzed the development of innovative algorithms capable of detecting obstructive sleep apnea from electrocardiograms [19], illustrating the practical impact of these competitions on advancing medical diagnostics and treatment strategies [20].

Since the success of PhysioNet, numerous other medical open databases have emerged globally, fostering a more cooperative and transparent research atmosphere (Figure 1). The impact of these large-scale databases on biomedical discovery is profound. One notable example is the UK Biobank [21], launched in 2006, which provides a vast repository of genetic and health information from half a million UK participants. This database has become an essential tool for unraveling the complex interplay between genetics, lifestyle, and disease, thereby enhancing our understanding of the factors influencing human health [22-25]. By leveraging large neuroimaging cohorts such as Alzheimer Disease Neuroimaging Initiative and the UK Biobank, researchers have developed AI-based models that generate an Alzheimer disease risk score from structural magnetic resonance imaging (MRI), enabling the identification of prediagnostic populations suitable for early intervention and preventive trials [26]. In cardiovascular research, analysis of the UK Biobank's genetic and imaging data has enabled the development of NeuralCVD, a neural network–based risk model that integrates polygenic and clinical predictors to estimate the 10-year risk of major adverse cardiac events, improving risk discrimination and reclassification beyond established clinical scores and Cox models, and highlighting the added predictive value of genetic predisposition in early prevention [27]. Similarly, the Cancer Imaging Archive, inaugurated in 2011 in the United States, offers a dedicated platform for the cancer research community, enabling access to a comprehensive array of imaging datasets [28].

Figure 1. Historical development of major medical open databases and data-sharing platforms over the past 25 years.

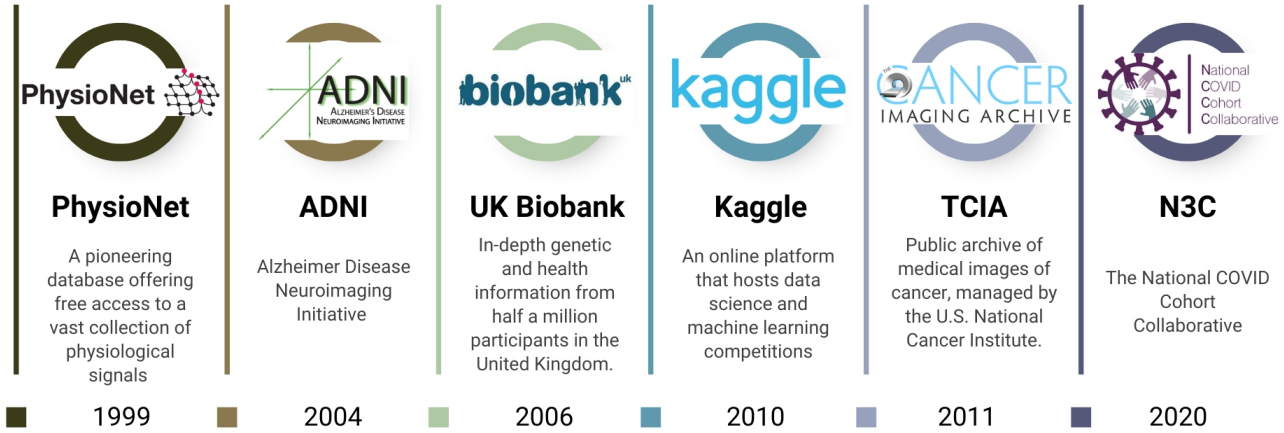

Beyond these examples, the ecosystem of medical open databases has diversified into numerous specialized domains. For instance, OpenNeuro provides a vast repository for neuroimaging data, particularly functional magnetic resonance imaging, electroencephalogram, and magnetoencephalography, supporting reproducible brain research [29]. The Neuroimaging Informatics Tools and Resources Clearinghouse offers a rich collection of imaging and data-processing tools [30]. Similarly, the National Database for Autism Research [31] and the Federal Interagency Traumatic Brain Injury Research informatics system [32] provide deeply phenotyped datasets crucial for research in their respective fields. These platforms underscore the field's shift toward creating specialized, high-quality resources to tackle specific biomedical questions. They display how shared resources can drive forward innovation and improve patient care worldwide, illustrating the critical role of collaborative environments in the advancement of health care research and application.

Additionally, Kaggle, an online platform for data science and machine learning (ML) competitions, has emerged as a pivotal player in the field of ML analysis and data sharing [33]. Launched in 2010, Kaggle facilitates collaboration and competition among data scientists and researchers by hosting challenges in various domains, including health care. These competitions often involve complex medical datasets, encouraging participants to develop innovative solutions and algorithms for disease prediction, medical imaging analysis [34], and other health-related issues. Kaggle has not only democratized access to large medical datasets but has also fostered a global community where knowledge and techniques are openly shared. This environment has led to significant breakthroughs and advancements in medical research and analytics, further suggesting the importance of open data and collaborative problem-solving in improving health care outcomes and accelerating medical innovation.

Balancing Privacy and the Need for Medical Open Databases

The success of open medical databases such as PhysioNet poses the challenge of balancing patient confidentiality with research enablement on open platforms [35,36]. Research datasets in health care often contain protected health information (PHI), and the process of removing this information, a process known as deidentification or anonymization, can be challenging and prone to errors [37]. Despite the use of these datasets, the need for deidentification introduces a significant barrier to data sharing due to the effort and cost involved.

XSL•FO
RenderX

The HIPAA, established in the United States in 1996, plays a vital role in safeguarding patients' medical information. In response to the HIPAA mandate, U.S. Department of Health and Human Services published a final regulation in the form of the privacy rule in December 2000, which became effective on April 14, 2001. Central to this rule is the designation of 18 specific categories of PHI that, if disclosed, could be used to identify an individual (Textbox 1). These categories encompass a broad spectrum of personal data, including, but not limited to, names, geographic details smaller than a state, various identifiers like social security numbers, medical record numbers, and contact information, as well as certain biometric and photographic images [38].

Textbox 1. Eighteen categories of protected health information.

- Names
- All geographic subdivisions smaller than a State
- All elements of dates (except year) for dates directly related to an individual
- Telephone numbers
- Fax numbers
- Electronic mail addresses
- Social security numbers
- Medical record numbers
- Health plan beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers and serial numbers
- Device identifiers and serial numbers
- Web URLs
- IP address numbers
- Biometric identifiers, including finger and voice prints
- Full face photographic images and any comparable images
- Any other unique identifying number, characteristic, or code

Additionally, HIPAA mandates that covered entities must ensure they do not possess knowledge that the remaining information could be used, whether alone or in conjunction with other data, to identify the subject. By strictly adhering to these guidelines, entities can share deidentified health information for broader uses, such as public health and research, without infringing on individual privacy rights, thus striking a balance between privacy protection and the beneficial use of health data [39]. Despite these efforts, the tension between the promise of big data and patient privacy in health care research remains a challenge [40].

PhysioNet is a pioneer in medical public databases, ensuring that the datasets it provides do not compromise individual privacy. This involves ensuring that any data shared does not contain PHI or has been sufficiently anonymized to prevent the identification of individuals. The challenges posed by the HIPAA privacy rule are not insignificant; they include the need for informed consent from data subjects and potential limitations on access to health information that can hinder clinical research [41,42]. Furthermore, the rule's interaction with other regulations, like the common rule, adds complexity to privacy concerns in research, leading to inconsistencies and additional burdens for researchers.

Despite these challenges, the privacy rule does allow for certain disclosures without patient authorization, particularly for public health purposes. This is intended to facilitate the use of medical data in important public health endeavors without undermining individual privacy protections [43]. The balance sought by the HIPAA privacy rule between protecting privacy and facilitating research is a critical aspect of its implementation, particularly in the context of medical open databases. By navigating these regulations successfully, repositories can contribute to the advancement of medical research while ensuring compliance with privacy standards [44].

The emergence of medical databases, such as the PhysioNet, UK Biobank, and the Cancer Imaging Archive, has significantly advanced collaborative research in health care [45,46]. These databases have the potential to transform cancer research and improve patient outcomes [45]. However, the collection, linking, and use of data in biomedical research raise ethical concerns, particularly regarding privacy and security [36,47,48]. Despite these concerns, the benefits of open data in health care, including improved diagnostics and treatment, are substantial [48]. The push for data sharing in cancer trials by pharmaceutical companies further underscores the importance of open medical databases in driving innovation and improving patient care [49].

XSL•FO
RenderX

Privacy Enhancing Technology

Overview

A range of studies have been conducted to explore the increasing frequency and impact of health care data breaches, highlighting the rising number of incidents and their detrimental effects on patient privacy and health care providers [50-53]. These breaches are often caused by a combination of technical, organizational, and human factors [50-52]. Human vulnerabilities, such as lack of awareness and training, play a significant role in these breaches [51]. The use of the Swiss Cheese Model can help assess vulnerabilities and risks [50]. Cloud computing breaches are a particular concern, highlighting the need for digital forensic readiness [54]. Hacking and unauthorized internal disclosures are the most prevalent forms of attack [53]. Further studies may examine specific cases and the implications for digital forensic readiness, emphasizing the importance of adhering to regulations.

Below, we reviewed several PETs and their applications in enhancing data privacy and security in health care settings (Table 1). PETs, such as encryption, anonymization techniques, and secure multiparty computation, offer powerful mechanisms to protect sensitive health data. Implementing these technologies, alongside robust privacy policies and employee training, can significantly reduce the likelihood of data breaches and bolster the trust between patients and health care providers.

Table . Summary of privacy-enhancing technologies.

Technologies	Core principle	Advantages	Challenges and trade-offs
Differential privacy	Adds calibrated statistical noise to query results to make it impossible to determine if an individual's data were included.	Provides strong and mathematically provable privacy guarantees.	Inherent trade-off between privacy and data use; high privacy can reduce analytical accuracy.
Synthetic data	Creates an artificial dataset that mimics the statistical properties of the original data without containing real patient information.	High use for model training; no real patient data are shared, eliminating reidentification risk.	Can be difficult to generate high-fidelity data that captures all complex correlations; potential for model bias.
Homomorphic encryption	Allows computations to be performed directly on encrypted data without decrypting it first.	Offers extremely strong security, as the raw data are never exposed.	High computational overhead; currently too slow for many complex ML[a] tasks.
Secure multiparty computation	Enables multiple parties to jointly compute a function over their inputs while keeping those inputs private.	Allows for collaborative analysis without a central data repository; no single party sees another's data.	High communication overhead between parties; can be complex to set up and scale.
Federated learning	Trains a central AI[b] model across decentralized devices or servers holding local data samples, without exchanging the data itself.	Keeps raw data local, enhancing privacy and data sovereignty.	Vulnerable to model poisoning/inversion attacks; performance can degrade with heterogeneous data.

[a]ML: machine learning.

[b]AI: artificial intelligence.

Differential Privacy

Differential privacy, a method for protecting individual privacy in data analysis, has been increasingly applied in the health care sector. It involves adding noise to the data to prevent reidentification of individuals. This approach has been used in various areas of health research, including genomics, neuroimaging, and health surveillance [55]. However, there are challenges in its practical application, such as the theoretical nature of the privacy parameter epsilon [56]. To address these challenges, researchers have proposed differentially private data release strategies and noise mechanisms, such as the Laplace and exponential mechanisms [57]. However, a key challenge is the inherent trade-off between privacy and data use; increasing the amount of statistical noise to protect privacy can reduce the accuracy of analytical outcomes.

The application of differential privacy in medical questionnaires has also been explored, with the randomized response mechanism showing promise in improving privacy while retaining data use [58]. Furthermore, the use of differential privacy in geospatial analyses of standardized health care data has been demonstrated, with the development of geodatabase functions for privacy-aware analysis [59]. Finally, the combination of differential privacy and decision tree approach has been proposed for data publishing, and the differentially private mini-batch gradient descent algorithm for model publishing of medical data [60].

Synthetic Data

Synthetic data, generated through simulators, is increasingly used in health care to address the challenges of data availability and privacy [61]. PETs, such as differential privacy, are combined with synthetic data generators to create private synthetic data, preserving statistical properties while ensuring privacy [62]. These technologies have been applied in various use cases, including clinical risk prediction [63] and medical research [64]. However, the evaluation of synthetic data's privacy and use metrics remains a challenge, with a lack of consensus on standard approaches [65]. Despite these challenges, the potential of synthetic data in preserving data use

XSL•FO
RenderX

and patient privacy in electronic health care data is being explored [66].

Homomorphic Encryption

Homomorphic encryption, a powerful tool for preserving privacy in medical data, allows for computations to be performed on encrypted data without the need for decryption. It has been successfully applied in various medical data scenarios, including ML models for classification and training, secure genomic algorithms, and predictive analysis tasks [67-69]. For example, it has been used to securely manage personal health metrics data, process medical images [70,71], and enable secure medical computation [72]. The use of homomorphic encryption in these applications ensures that sensitive medical data remains private and secure. Despite its power, the primary limitation of homomorphic encryption is its significant computational overhead, which can make it slow and resource-intensive for complex computations on large datasets.

Secure Multiparty Computation

Secure multiparty computation is a cryptographic technique that enables data analytics without sharing the underlying data, making it a valuable tool for preserving privacy in medical data analysis [73]. It has been applied in various health care scenarios, including collaborative systems [74], statistical analysis of health data [75], and electronic medical record (EMR) data [75]. Secure multiparty computation has also been used in health care internet of things systems to handle privacy issues [76], prevent data disclosure in sensor networks [77], and enable the reuse of distributed electronic health data [75]. Furthermore, it has been applied to enable privacy-preserving query processing on EMRs [78]. Notably, secure multiparty computation has enabled research on highly sensitive data (such as HIV, rare diseases, and population genomics) that would otherwise be inaccessible due to privacy concerns.

Federated Learning

Federated learning (FL), a decentralized ML approach, is increasingly being applied in the medical field due to the sensitive and fragmented nature of health care data [14,79]. It allows for the collaborative development of ML models without sharing raw data, thus preserving privacy [80,81]. This approach has been used in various medical domains, including oncology and radiology, for tasks such as image analysis and disease prediction [81,82]. However, there are challenges to be addressed, such as data homogeneity and transparency [81]. Furthermore, FL can be vulnerable to security risks, such as model inversion attacks that attempt to reconstruct training data from the shared model updates, and require careful design to ensure robustness. Despite these challenges, FL shows promise in improving the efficiency and privacy of medical data processing [83-85].

To address the challenge of data heterogeneity in FL, we have proposed the Dynamically Synthetic Images for Federated Learning method, significantly improving the conventional FL framework by integrating local information from local multiple institutions with heterogeneous data types [86]. The core principle of its implementation involves a dynamic process where, at the start of each training round, a client's local data are evaluated by the current global model to identify misclassified images. Using a synthetic minority oversampling technique, the system generates new, synthetic images based on these misclassified cases, which are then added to the local training set to compel the model to focus on features it previously failed to learn. In terms of effectiveness, experimental results demonstrated that Dynamically Synthetic Images for Federated Learning-based models achieve higher accuracy than conventional FL approaches and that their performance can be comparable to that of traditional centralized learning, proving especially beneficial for institutions with smaller or more heterogeneous datasets [86].

Taiwan Medical AI and Data Portal and Dynamic Consents System

Taiwan has made significant strides in medical data sharing, particularly in the areas of privacy protection and electronic health records exchange. The country's comprehensive embedded integrated circuit-based health insurance card system, implemented by the Bureau of National Health Insurance, Taiwan, allows for the secure sharing of health information [87]. The use of blockchain technology has been proposed as a means to further enhance the security and privacy of medical data sharing [88,89]. The Taiwan Electronic Medical Record Template and the National Electronic Medical Record Exchange System have been developed to facilitate the exchange of EMRs [90,91]. However, concerns about unauthorized access and secondary use of EMRs persist, particularly among highly educated individuals [92]. The country has also established guidelines for the security and privacy protection of health information, drawing on international best practices [93].

In the past 4 years, funded by the National Science and Technology Council of Taiwan, we have assembled teams from National Yang Ming Chiao Tung University, Taipei Veterans General Hospital, Academia Sinica, and National Taipei University of Nursing and Health Sciences to form a data repository task force known as the Smart Medical AI and Repository Taskforce Center. We launched a medical AI and data-sharing platform aimed at advancing the field of medical AI research in October 2023 [94]. This platform not only provides the public and researchers with access to a multitude of shared datasets but also ensures a meticulous evaluation process (Figure 2). Researchers can apply for access to the data by providing an abstract of their research proposal. Dataset managers assess applications based on their intended use, specific needs, and detailed research plans.

XSL•FO
RenderX

Figure 2. Taiwan medical artificial intelligence and data portal. AI: artificial intelligence.

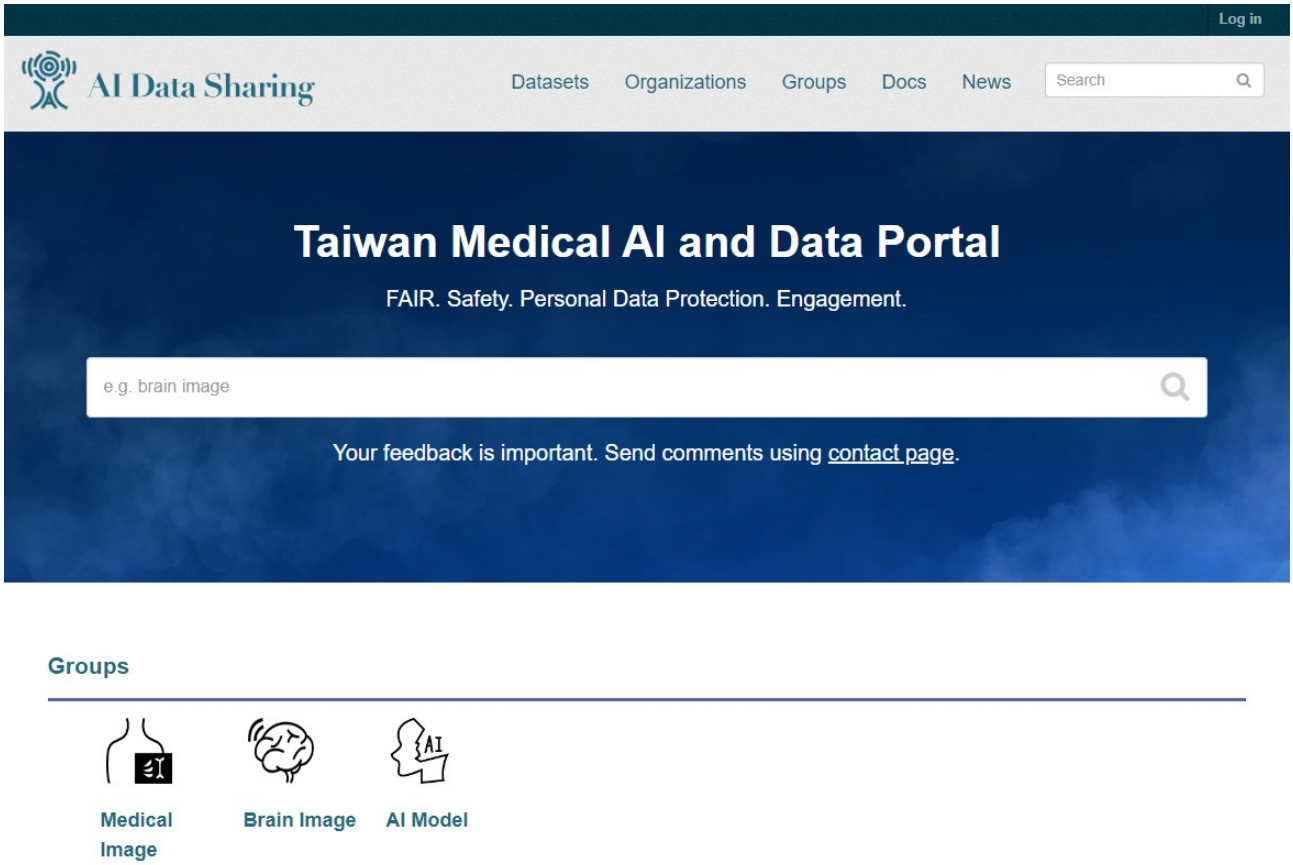

Currently, seventeen datasets have been released on the platform, covering neuropsychiatric disorders, brain tumors, ophthalmic diseases, musculoskeletal disorders, and cardiopulmonary diseases. All datasets have undergone deidentification and delinking processes and include annotated information to facilitate AI training and validation. Specifically, our data-sharing platform includes: MRI images of vestibular schwannoma; computed tomography (CT) images of intracerebral hemorrhage; brain Fluorodeoxyglucose-Positron Emission Tomography/Magnetic Resonance Imaging images for dementia diagnosis; primary brain tumor MRI datasets, including meningioma, glioma, and pituitary adenoma; MRI data of brain metastases, which represent the largest collection nationwide; hand and foot X-rays of rheumatoid arthritis; X-rays of compression fractures; spinal X-rays of ankylosing spondylitis; chest CT images and clinical data of atrial fibrillation patients; chest X-rays for lung cancer screening; annotated preoperative liver CT images; neck lymph node CT images with postoperative pathology results; the Taiwan Aging and Mental Illness Cohort brain imaging database; the dementia molecular imaging database; fundus image datasets for glaucoma; and fundus image datasets of polypoidal choroidal vasculopathy.

The data sharing platform is built on a comprehensive architecture designed to support AI research by integrating 3 core systems: a CKAN-based sharing platform for dataset management, a data application system, and a dynamic authorization consent platform for patient privacy. Specific features include a robust user authentication and authorization mechanism, allowing dataset managers to grant access to specific users or collaborators. The platform ensures data integrity and ethical compliance through a multistep deidentification process for all medical images and by linking to the dynamic consent system (for sensitive clinical data), which allows patients to manage their data sharing preferences in real-time. To use the database, researchers first search for datasets on the platform, then apply for access through a formal registration and review process. Once approved and authorized by the dataset manager, users can obtain a login key to programmatically access the data through standardized protocols, such as DICOMweb, ensuring a secure, convenient, and interoperable environment for third-party AI applications.

This effort aims to advance research across 7 crucial clinical areas that greatly benefit from AI technology, including heart disease, neurological disorders, mental illness, diabetes, cancer, genetic predispositions to complex diseases, and medical imaging. Moreover, the platform underpins collaboration between distinct teams specializing in AI methodology, science and law, and data governance, jointly fostering a robust data governance framework that emphasizes FL, cloud-based AI solutions, and trusted AI practices. Importantly, the system is designed to streamline the research process while maintaining a focus on ethical standards and participant privacy. In line with this, the platform incorporates a dynamic informed consent mechanism, especially for datasets that are anonymized but cannot be completely separated from their sources. This approach ensures that participants' privacy is safeguarded while also enabling their informed and ongoing consent, reflecting

XSL•FO
RenderX

our commitment to ethical research practices and the dynamic nature of consent in medical studies.

Dynamic informed consent, a concept that has been explored in various contexts, is a personalized, digital communication interface that allows participants to manage their consent preferences [95]. It has been proposed as a solution to improve patient confidence and trust in the use of electronic patient records in medical research [96]. In the context of privacy-aware pervasive health and well-being, dynamic consent enables granular data consent and management [97]. It has also been suggested as a potential solution to challenges in modern biomedical research, including participant recruitment, informed consent, and consent management [98]. The use of blockchain technology has been proposed to enhance the privacy-preserving aspects of dynamic consent in genomic data sharing [99]. The concept of dynamic informed consent has been further explored in the context of personalized medicine, emphasizing the need for a more dynamic and enriched consent model [100].

To enhance the privacy of participants contributing to the data in our platform and increase participants' engagement, we have developed a dynamic informed consent system, named the Health Data Authorization Service Platform (Figure 3). This collaborative effort involves our data governance, humanities, science and law, and clinical teams, aiming to facilitate scalable, dynamic consent operations suitable for complex data environments. The platform supports flexible data governance, allowing data owners to dynamically express their consent preferences, thereby making dynamic consent practical and sustainable.

Figure 3. Health data authorization service platform. This diagram depicts the architecture of a medical data-sharing platform that integrates dynamic consent for secure and transparent data sharing. The system centers on 4 key roles: the resource owner (RO), who owns the data; the resource server (RS), where medical data are stored and managed; the requesting party (RP), typically researchers seeking data access; and the authorization server (AS), the core component facilitating connections among the roles. This framework ensures data access is based on explicit consent from the data owner, allowing real-time adjustments to consent settings for different data types and uses, such as academic research. The platform's primary tasks are to establish individual preferences, maintain consent history, and ensure trust and transparency between data owners and users, thereby advancing responsible data science development.

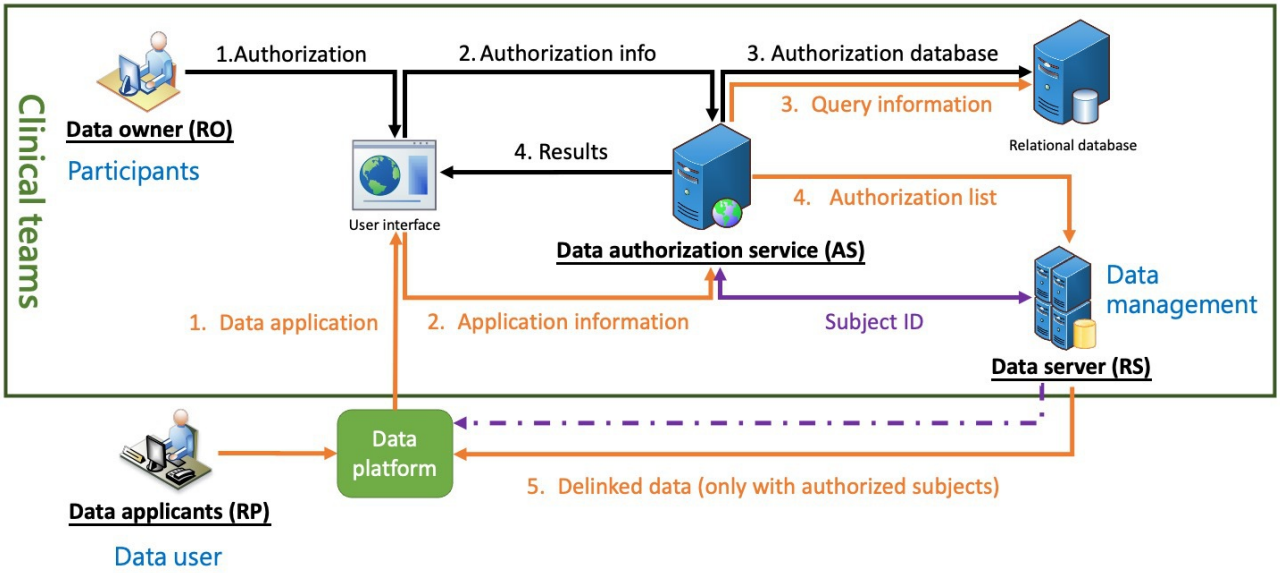

The platform's architecture encompasses 4 key roles: the resource owner, usually the participants, who owns the data; the resource server, which stores and manages medical data; the requesting party, typically researchers seeking data access; and the authorization server, the backbone of our dynamic consent system, connecting the 3 roles. This system ensures that data are accessed only with the owner's express consent, respecting their preferences and enhancing data use transparency. Resource owners can modify their consent settings at any time, reflecting changes in their willingness to share different data types, like EMRs or medical imaging, for specific purposes such as academic research. This flexibility ensures that data use aligns with the owners' current preferences. The platform seamlessly integrates with our shared data framework, maintaining each citizen's consent history and enabling swift updates to consent forms as needed. By streamlining the consent process and ensuring data are shared according to owner permissions, our platform respects individual preferences while promoting responsible data science development. It exemplifies a forward-thinking approach to data governance, enabling real-time adjustments in consent and fostering a culture of trust and transparency between data owners and users.

These initiatives in Taiwan can be understood within the global context of evolving data privacy regulations. Unlike the "one-time, broad consent" model often used in US-based research under the Common Rule, Taiwan's move toward dynamic consent aligns more closely with the principles of the European Union's General Data Protection Regulation [101]. The General Data Protection Regulation mandates that consent must be specific, informed, and easily revocable. The dynamic consent system builds on this by providing a technological interface for participants to manage their preferences granularly and continuously, representing a best-practice approach to balancing research needs with individual autonomy and privacy rights.

XSL•FO
RenderX

Acceleration of Medical AI Research Through Open Databases

The advent of medical open databases has significantly accelerated the field of medical AI research, fostering an environment of innovation and rapid development [102]. By providing researchers with access to vast amounts of health-related data, these databases have become a cornerstone for advancements in predictive analytics, diagnostic algorithms, and personalized medicine.

One of the most notable contributions of open medical databases to AI research is the democratization of data [103]. Historically, the scarcity and inaccessibility of medical data posed substantial hurdles to AI development. However, platforms like PhysioNet, established in 1999, have bridged this gap by offering a plethora of datasets ranging from physiological signals to clinical outcomes [104]. However, challenges remain, including the need for large datasets and the lack of external validation in perioperative medicine [105]. The use of open science approaches, including data liberation and crowdsourcing, can help address these challenges [106]. The integration of networked medical devices and clinical repositories based on open standards can further enhance AI research in high-acuity medical environments [107]. This enhanced availability allows researchers from diverse backgrounds and institutions to engage in health care innovation, leveling the playing field and stimulating a surge in AI-based solutions.

The availability of open databases has catalyzed the application of diverse AI methodologies to complex medical problems. Deep learning, particularly convolutional neural networks, has achieved state-of-the-art performance by leveraging large-scale imaging datasets; for example, researchers have trained convolutional neural networks on millions of images from The Cancer Imaging Archive to develop algorithms capable of detecting and classifying tumors in radiological scans with accuracy comparable to human experts [108]. For structured data such as the genetic and clinical information in the UK Biobank, traditional ML models like random forests and gradient boosting have been widely used, excelling at identifying complex patterns to predict disease risk, including the calculation of polygenic risk scores for coronary artery disease based on thousands of genetic variants [109]. In addition, natural language processing techniques have been applied to large repositories of unstructured clinical notes, such as those in the Medical Information Mart for Intensive Care version IV (MIMIC-IV) database (part of PhysioNet), to extract critical information on symptoms, treatments, and outcomes, thereby enabling large-scale retrospective studies that were previously infeasible [110].

Open medical databases encompass a wide variety of data types, including EMRs, imaging, genomic sequences, and more. This diversity enables AI researchers to explore multifaceted health care questions, from predicting disease trajectories to optimizing treatment plans. Moreover, the rich, varied datasets facilitate the training of more robust and generalizable AI models, capable of addressing complex medical scenarios across different populations and settings. The shared nature of open databases fosters collaboration across the global research community [111]. Through platforms that offer shared data, researchers can combine their expertise to tackle larger and more complex problems than they could individually. This collaborative approach has led to significant breakthroughs in AI, such as algorithms that can detect diseases from images with accuracy rivaling that of trained professionals [112,113].

Open databases also streamline the validation and implementation phases of AI development [114]. Access to diverse datasets enables researchers to rigorously test their algorithms under various conditions and patient demographics, ensuring their reliability and effectiveness. The expansion of these databases has significantly propelled medical AI research forward, marking a new phase of health care innovation with faster discoveries, collaborative efforts, and a commitment to ethical data use. As the field evolves, the role of open databases in shaping the future of medicine remains pivotal.

Conclusions

In conclusion, the evolution toward medical open databases, exemplified by the inception of platforms like PhysioNet in 1999 and their progression over the past 25 years, alongside the integration of PETs, marks a significant milestone in the domain of biomedical research and health care innovation. This journey not only fosters an unprecedented level of collaboration and accessibility but also emphasizes the crucial need to address privacy concerns and ethical considerations diligently. The ongoing efforts to balance data sharing with individual privacy protection are underscored by the adaptation of legal frameworks and the implementation of cryptographic and data management solutions. The introduction and growth of medical open databases have been pivotal, providing a wealth of data that has propelled research and innovation while highlighting the challenges and responsibilities of managing sensitive information. Specifically, the availability of open medical databases has significantly accelerated AI research, leading to breakthroughs in disease prediction, diagnostics, and personalized medicine. As we continue to explore the vast potential of open science and FL, the landscape of health care research is on the brink of remarkable transformations. These advancements promise enhanced patient outcomes, faster medical discoveries, and a more inclusive global research community, all achieved by adhering to the highest standards of privacy and ethical integrity.

XSL•FO
RenderX

Acknowledgments

The authors thank Dr Watson Lin for supporting the infrastructure of the data-sharing platform and for providing valuable perspectives on the guidelines for medical data sharing. The authors also thank the National Science and Technology Council for offering legal insights on privacy policies related to medical data sharing.

Funding

This work was supported by grants from the National Science and Technology Council, Taiwan (grant number NSCT 114-2634-F-A49-006 and 113-2634-F-A49-003). ACY was also supported by the Mt. Jade Young Scholarship Award from the Ministry of Education, Taiwan, as well as Brain Research Center, National Yang Ming Chiao Tung University, and the Ministry of Education (Aim for the Top University Plan), Taipei, Taiwan.

Authors' Contributions

Conceptualization; writing – original draft: AY;
Writing – review and editing: All Authors;
Funding acquisition: CHL

Conflicts of Interest

None declared.

References

1. Merelli I, Pérez-Sánchez H, Gesing S, D'Agostino D. Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives. Biomed Res Int 2014;2014:134023. [doi: 10.1155/2014/134023] [Medline: 25254202]
2. Roski J, Bo-Linn GW, Andrews TA. Creating value in health care through big data: opportunities and policy implications. Health Aff (Millwood) 2014 Jul;33(7):1115-1122. [doi: 10.1377/hlthaff.2014.0147] [Medline: 25006136]
3. Wang Y, Kung L, Byrd TA. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. Technol Forecast Soc Change 2018 Jan;126:3-13. [doi: 10.1016/j.techfore.2015.12.019]
4. Goldberger AL, Amaral LA, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000 Jun 13;101(23):E215-E220. [doi: 10.1161/01.cir.101.23.e215] [Medline: 10851218]
5. Dean DA 2nd, Goldberger AL, Mueller R, et al. Scaling up scientific discovery in sleep medicine: the national sleep research resource. Sleep 2016 May 1;39(5):1151-1164. [doi: 10.5665/sleep.5774] [Medline: 27070134]
6. Wylie JE, Mineau GP. Biomedical databases: protecting privacy and promoting research. Trends Biotechnol 2003 Mar;21(3):113-116. [doi: 10.1016/S0167-7799(02)00039-2] [Medline: 12628367]
7. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(2):124-130. [doi: 10.1136/jamia.2009.000893] [Medline: 20190053]
8. Shahin MH, Bhattacharya S, Silva D, et al. Open data revolution in clinical research: opportunities and challenges. Clin Transl Sci 2020 Jul;13(4):665-674. [doi: 10.1111/cts.12756] [Medline: 32004409]
9. Ness RB, Joint Policy Committee, Societies of Epidemiology. Influence of the HIPAA privacy rule on health research. JAMA 2007 Nov 14;298(18):2164-2170. [doi: 10.1001/jama.298.18.2164] [Medline: 18000200]
10. Dwork C, Roth A. The algorithmic foundations of differential privacy. FNT in Theoretical Computer Science 2014;9(3-4):211-407. [doi: 10.1561/0400000042]
11. Rubin DB. Discussion: statistical disclosure limitation. J Off Stat 1993;9:461-468 [FREE Full text]
12. Yi X, Paulet R, Bertino E. Homomorphic encryption. In: Homomorphic Encryption and Applications: Springer; 2014:27-46. [doi: 10.1007/978-3-319-12229-8_2]
13. Cramer RJF, Damgård IB, Nielsen JB. Secure Multiparty Computation and Secret Sharing: Cambridge University Press; 2015. [doi: 10.1017/CBO9781107337756]
14. Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. J Healthc Inform Res 2021 Mar;5(1):1-19. [doi: 10.1007/s41666-020-00082-4] [Medline: 33204939]
15. Packer M. Data sharing in medical research. BMJ 2018 Feb 14;360:k510. [doi: 10.1136/bmj.k510] [Medline: 29444885]
16. Moody GB, Mark RG, Goldberger AL. PhysioNet: physiologic signals, time series and related open source software for basic, clinical, and applied research. Annu Int Conf IEEE Eng Med Biol Soc 2011;2011:8327-8330. [doi: 10.1109/IEMBS.2011.6092053] [Medline: 22256277]
17. Moody GB, Mark RG, Goldberger AL. PhysioNet: a research resource for studies of complex physiologic and biomedical signals. Presented at: Computers in Cardiology 2000; Sep 24-27, 2000. [doi: 10.1109/CIC.2000.898485]
18. Henry IC, Goldberger AL, Moody GB, Mark RG. PhysioNet: an NIH research resource for physiologic datasets and open-source software. : IEEE Comput Soc Presented at: 14th IEEE Symposium on Computer-Based Medical Systems CBMS 2001; Jul 26-27, 2001; Bethesda, MD p. 245-250. [doi: 10.1109/CBMS.2001.941728]

XSL•FO
RenderX

19. Thomas RJ, Mietus JE, Peng CK, Goldberger AL. An electrocardiogram-based technique to assess cardiopulmonary coupling during sleep. Sleep 2005 Sep;28(9):1151-1161. [doi: 10.1093/sleep/28.9.1151] [Medline: 16268385]
20. Thomas RJ, Mietus JE, Peng CK, et al. Differentiating obstructive from central and complex sleep apnea using an automated electrocardiogram-based method. Sleep 2007 Dec;30(12):1756-1769. [doi: 10.1093/sleep/30.12.1756] [Medline: 18246985]
21. Collins R. What makes UK Biobank special? Lancet 2012 Mar 31;379(9822):1173-1174. [doi: 10.1016/S0140-6736(12)60404-8] [Medline: 22463865]
22. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature New Biol 2018 Oct;562(7726):203-209. [doi: 10.1038/s41586-018-0579-z] [Medline: 30305743]
23. Allen N, Sudlow C, Downey P, et al. UK Biobank: current status and what it means for epidemiology. Health Policy Technol 2012 Sep;1(3):123-126. [doi: 10.1016/j.hlpt.2012.07.003]
24. Littlejohns TJ, Holliday J, Gibson LM, et al. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nat Commun 2020 May 26;11(1):2624. [doi: 10.1038/s41467-020-15948-9] [Medline: 32457287]
25. Littlejohns TJ, Sudlow C, Allen NE, Collins R. UK Biobank: opportunities for cardiovascular research. Eur Heart J 2019 Apr 7;40(14):1158-1166. [doi: 10.1093/eurheartj/ehx254] [Medline: 28531320]
26. Azevedo T, Bethlehem RAI, Whiteside DJ, et al. Identifying healthy individuals with Alzheimer's disease neuroimaging phenotypes in the UK Biobank. Commun Med (Lond) 2023 Jul 20;3(1):100. [doi: 10.1038/s43856-023-00313-w] [Medline: 37474615]
27. Steinfeldt J, Buergel T, Loock L, et al. Neural network-based integration of polygenic and clinical information: development and validation of a prediction model for 10-year risk of major adverse cardiac events in the UK Biobank cohort. Lancet Digit Health 2022 Feb;4(2):e84-e94. [doi: 10.1016/S2589-7500(21)00249-1] [Medline: 35090679]
28. Clark K, Vendt B, Smith K, et al. The cancer imaging archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013 Dec;26(6):1045-1057. [doi: 10.1007/s10278-013-9622-7] [Medline: 23884657]
29. Markiewicz CJ, Gorgolewski KJ, Feingold F, et al. The OpenNeuro resource for sharing of neuroscience data. Elife 2021 Oct 18;10:e71774. [doi: 10.7554/eLife.71774] [Medline: 34658334]
30. Buccigrossi R, Ellisman M, Grethe J, et al. The neuroimaging informatics tools and resources clearinghouse (NITRC). AMIA Annu Symp Proc 2008 Nov 6;1000:1000. [Medline: 18999128]
31. Hall D, Huerta MF, McAuliffe MJ, Farber GK. Sharing heterogeneous data: the national database for autism research. Neuroinformatics 2012 Oct;10(4):331-339. [doi: 10.1007/s12021-012-9151-4] [Medline: 22622767]
32. Thompson HJ, Vavilala MS, Rivara FP. Chapter 1 common data elements and federal interagency traumatic brain injury research informatics system for TBI research. Annu Rev Nurs Res 2015;33(1):1-11. [doi: 10.1891/0739-6686.33.1] [Medline: 25946381]
33. Kaggle. URL: https://www.kaggle.com/ [accessed 2025-12-22]
34. Yang X, Zeng Z, Teo SG, Wang L, Chandrasekhar V, Hoi S. Deep learning for practical image recognition: case study on kaggle competitions. Presented at: KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; Aug 19-23, 2018; London, United Kingdom p. 923-931. [doi: 10.1145/3219819.3219907]
35. Krishna R, Kelleher K, Stahlberg E. Patient confidentiality in the research use of clinical medical databases. Am J Public Health 2007 Apr;97(4):654-658. [doi: 10.2105/AJPH.2006.090902] [Medline: 17329644]
36. Kobayashi S, Kane TB, Paton C. The privacy and security implications of open data in healthcare. Yearb Med Inform 2018 Aug;27(1):41-47. [doi: 10.1055/s-0038-1641201] [Medline: 29681042]
37. Eze B, Peyton L. Systematic literature review on the anonymization of high dimensional streaming datasets for health data sharing. Procedia Comput Sci 2015;63:348-355. [doi: 10.1016/j.procs.2015.08.353]
38. Centers for Disease Control and Prevention (CDC). HIPAA privacy rule and public health. Guidance from CDC and the U.S. Department of Health and Human Services. MMWR Suppl 2003 May 2;52:1-17. [Medline: 12741579]
39. Prasser F, Kohlmayer F, Spengler H, Kuhn KA. A scalable and pragmatic method for the safe sharing of high-quality health data. IEEE J Biomed Health Inform 2018 Mar;22(2):611-622. [doi: 10.1109/JBHI.2017.2676880] [Medline: 28358693]
40. Sarpatwari A, Gagne JJ. Balancing benefits and harms: privacy protection policies. Pharmacoepidemiol Drug Saf 2016 Aug;25(8):969-971. [doi: 10.1002/pds.4048] [Medline: 27278106]
41. Kulynych J, Korn D. The new HIPAA (Health Insurance Portability and Accountability Act of 1996) medical privacy rule: help or hindrance for clinical research? Circulation 2003 Aug 26;108(8):912-914. [doi: 10.1161/01.CIR.0000080642.35380.50] [Medline: 12939240]
42. Nosowsky R, Giordano TJ. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) privacy rule: implications for clinical research. Annu Rev Med 2006;57(1):575-590. [doi: 10.1146/annurev.med.57.121304.131257] [Medline: 16409167]
43. Gostin LO, Nass S. Reforming the HIPAA privacy rule: safeguarding privacy and promoting research. JAMA 2009 Apr 1;301(13):1373-1375. [doi: 10.1001/jama.2009.424] [Medline: 19336713]
44. Wartenberg D, Thompson WD. Privacy versus public health: the impact of current confidentiality rules. Am J Public Health 2010 Mar;100(3):407-412. [doi: 10.2105/AJPH.2009.166249] [Medline: 20075316]

XSL•FO
RenderX

45. Green AK, Reeder-Hayes KE, Corty RW, et al. The project data sphere initiative: accelerating cancer research by sharing data. Oncologist 2015 May;20(5):464-e20. [doi: 10.1634/theoncologist.2014-0431] [Medline: 25876994]
46. Conroy M, Sellors J, Effingham M, et al. The advantages of UK Biobank's open-access strategy for health research. J Intern Med 2019 Oct;286(4):389-397. [doi: 10.1111/joim.12955] [Medline: 31283063]
47. Anderson R. The collection, linking and use of data in biomedical research and health care: ethical issues. : The Nuffield Council on Bioethics; 2015. [doi: 10.17863/CAM.31760]
48. Kostkova P, Brewer H, de Lusignan S, et al. Who owns the data? Open data for healthcare. Front Public Health 2016;4:7. [doi: 10.3389/fpubh.2016.00007] [Medline: 26925395]
49. Bhattacharjee Y. Biomedicine. Pharma firms push for sharing of cancer trial data. Science 2012 Oct 5;338(6103):29. [doi: 10.1126/science.338.6103.29] [Medline: 23042862]
50. Kamoun F, Nicho M. Human and organizational factors of healthcare data breaches: the swiss cheese model of data breach causation and prevention. Int J Healthc Inf Syst Inform 2014 Jan 1;9:42-60. [doi: 10.4018/ijhisi.2014010103]
51. Nifakos S, Chandramouli K, Nikolaou CK, et al. Influence of human factors on cyber security within healthcare organisations: a systematic review. Sensors (Basel) 2021 Jul 28;21(15):5119. [doi: 10.3390/s21155119] [Medline: 34372354]
52. McLeod A, Dolezel D. Cyber-analytics: modeling factors associated with healthcare data breaches. Decis Support Syst 2018 Apr;108:57-68. [doi: 10.1016/j.dss.2018.02.007]
53. Seh AH, Zarour M, Alenezi M, et al. Healthcare data breaches: insights and implications. Healthcare (Basel) 2020 May 13;8(2):133. [doi: 10.3390/healthcare8020133] [Medline: 32414183]
54. Chernyshev M, Zeadally S, Baig Z. Healthcare data breaches: implications for digital forensic readiness. J Med Syst 2018 Nov 28;43(1):7. [doi: 10.1007/s10916-018-1123-2] [Medline: 30488291]
55. Ficek J, Wang W, Chen H, Dagne G, Daley E. Differential privacy in health research: a scoping review. J Am Med Inform Assoc 2021 Sep 18;28(10):2269-2276. [doi: 10.1093/jamia/ocab135] [Medline: 34333623]
56. Dankar F, Emam K. Practicing differential privacy in health care: a review. Trans Data Priv 2013;6(1):35-67. [doi: 10.5555/2612156.2612159]
57. Tamane S, Solanki VK. In: Dey N, editor. Privacy and Security Policies in Big Data: IGI Global; 2017. [doi: 10.4018/978-1-5225-2486-1]
58. Appenzeller A, Terzer N, Philipp P, Beyerer J. Applying differential privacy to medical questionnaires. Presented at: 2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops); Mar 13-17, 2023; Atlanta, GA p. 608-613. [doi: 10.1109/PerComWorkshops56833.2023.10150373]
59. Harris DR. Leveraging differential privacy in geospatial analyses of standardized healthcare data. Proc IEEE Int Conf Big Data 2020 Dec;2020:3119-3122. [doi: 10.1109/bigdata50022.2020.9378390] [Medline: 35253022]
60. Sun Z, Wang Y, Shu M, Liu R, Zhao H. Differential privacy for data and model publishing of medical data. IEEE Access 2019;7:152103-152114. [doi: 10.1109/ACCESS.2019.2947295] [Medline: 31328077]
61. McDuff D, Curran T, Kadambi A. Synthetic data in healthcare. arXiv. Preprint posted online on Apr 6, 2023. [doi: 10.48550/ARXIV.2304.03243]
62. Appenzeller A, Leitner M, Philipp P, Krempel E, Beyerer J. Privacy and utility of private synthetic data for medical data analyses. Appl Sci (Basel) 2022 Dec 1;12(23):12320. [doi: 10.3390/app122312320]
63. Qian Z, Callender T, Cebere B, Janes SM, Navani N, van der Schaar M. Synthetic data for privacy-preserving clinical risk prediction. medRxiv. Preprint posted online on May 24, 2023. [doi: 10.1101/2023.05.18.23290114]
64. Kokosi T, Harron K. Synthetic data in medical research. BMJ Med 2022;1(1):e000167. [doi: 10.1136/bmjmed-2022-000167] [Medline: 36936569]
65. Kaabachi B, Despraz J, Meurers T, et al. A scoping review of privacy and utility metrics in medical synthetic data. medRxiv. Preprint posted online on Oct 21, 2024. [doi: 10.1101/2023.11.28.23299124]
66. Wang Z, Myles P, Tucker A. Generating and evaluating cross - sectional synthetic electronic healthcare data: preserving data utility and patient privacy. Comput Intell 2021 May;37(2):819-851. [doi: 10.1111/coin.12427]
67. Wood A, Najarian K, Kahrobaei D. Homomorphic encryption for machine learning in medicine and bioinformatics. ACM Comput Surv 2021 Jul 31;53(4):1-35. [doi: 10.1145/3394658]
68. Bos JW, Lauter K, Naehrig M. Private predictive analysis on encrypted medical data. J Biomed Inform 2014 Aug;50:234-243. [doi: 10.1016/j.jbi.2014.04.003] [Medline: 24835616]
69. Bocu R, Costache C. A homomorphic encryption-based system for securely managing personal health metrics data. IBM J Res & Dev 2018 Jan 1;62(1):1. [doi: 10.1147/JRD.2017.2755524]
70. Krishnegowda P, M. Boregowda A. Efficient matrix key homomorphic encryption of medical images. IJEECS 2023 Jul 1;31(1):406. [doi: 10.11591/ijeecs.v31.i1.pp406-416]
71. Kartit A. New approach based on homomorphic encryption to secure medical images in cloud computing. Trends Sci 2022;19(9):3970. [doi: 10.48048/tis.2022.3970]
72. Khedr A, Gulak G. SecureMed: secure medical computation using GPU-accelerated homomorphic encryption scheme. IEEE J Biomed Health Inform 2018 Mar;22(2):597-606. [doi: 10.1109/JBHI.2017.2657458] [Medline: 28129194]
73. Veeningen M, Chatterjea S, Horváth AZ, et al. Enabling analytics on sensitive medical data with secure multi-party computation. Stud Health Technol Inform 2018;247:76-80. [Medline: 29677926]

XSL•FO
RenderX

74. Marwan M, Kartit A, Ouahmane H. Applying secure multi-party computation to improve collaboration in healthcare cloud. Presented at: 2016 Third International Conference on Systems of Collaboration (SysCo); Nov 28-29, 2016; Casablanca, Morocco p. 1-6. [doi: 10.1109/SYSCO.2016.7831325]
75. Yigzaw KY, Bellika JG. Evaluation of secure multi-party computation for reuse of distributed electronic health data. Presented at: 2014 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI); Jun 1-4, 2014; Valencia, Spain p. 219-222. [doi: 10.1109/BHI.2014.6864343]
76. Şahinbaş K, Catak FO. Secure multi-party computation-based privacy-preserving data analysis in healthcare IoT systems. In: Kose U, Gupta D, Khanna A, Rodrigues J, editors. Interpretable Cognitive Internet of Things for Healthcare: Springer; 2023:57-72. [doi: 10.1007/978-3-031-08637-3_3]
77. Tso R, Alelaiwi A, Mizanur Rahman SM, Wu ME, Hossain MS. Privacy-preserving data communication through secure multi-party computation in healthcare sensor cloud. J Sign Process Syst 2017 Oct;89(1):51-59. [doi: 10.1007/s11265-016-1198-2]
78. Tawfik AM, Sabbeh SF, EL-Shishtawy T. Privacy-preserving secure multiparty computation on electronic medical records for star exchange topology. Arab J Sci Eng 2018 Dec;43(12):7747-7756. [doi: 10.1007/s13369-018-3122-5]
79. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. IEEE Signal Process Mag 2020 May;37(3):50-60. [doi: 10.1109/MSP.2020.2975749]
80. Nguyen DC, Pham QV, Pathirana PN, et al. Federated learning for smart healthcare: a survey. ACM Comput Surv 2023 Mar 31;55(3):1-37. [doi: 10.1145/3501296]
81. Crowson MG, Moukheiber D, Arévalo AR, et al. A systematic review of federated learning applications for biomedical data. PLOS Digit Health 2022 May;1(5):e0000033. [doi: 10.1371/journal.pdig.0000033] [Medline: 36812504]
82. Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated electronic health records. Int J Med Inform 2018 Apr;112:59-67. [doi: 10.1016/j.ijmedinf.2018.01.007] [Medline: 29500022]
83. Antunes RS, André da Costa C, Küderle A, Yari IA, Eskofier B. Federated learning for healthcare: systematic review and architecture proposal. ACM Trans Intell Syst Technol 2022 Aug 31;13(4):1-23. [doi: 10.1145/3501813]
84. Guo K, Chen T, Ren S, Li N, Hu M, Kang J. Federated learning empowered real-time medical data processing method for smart healthcare. IEEE/ACM Trans Comput Biol Bioinform 2024;21(4):869-879. [doi: 10.1109/TCBB.2022.3185395] [Medline: 35737631]
85. Pfitzner B, Steckhan N, Arnrich B. Federated learning in a medical context: a systematic literature review. ACM Trans Internet Technol 2021 Jun 23;21(2):1-31. [doi: 10.1145/3412357]
86. Wu JCH, Yu HW, Tsai TH, Lu HHS. Dynamically synthetic images for federated learning of medical images. Comput Methods Programs Biomed 2023 Dec;242:107845. [doi: 10.1016/j.cmpb.2023.107845] [Medline: 37852147]
87. Liu HH. Use and disclosure of health information and protection of patient privacy in Taiwan. Med Law 2010 Mar;29(1):87-101. [Medline: 22458000]
88. Jin H, Luo Y, Li P, Mathew J. A review of secure and privacy-preserving medical data sharing. IEEE Access 2019;7:61656-61669. [doi: 10.1109/ACCESS.2019.2916503]
89. Liang X, Zhao J, Shetty S, Liu J, Li D. Integrating blockchain for data sharing and collaboration in mobile healthcare applications. Presented at: 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC); Oct 8-13, 2017; Montreal, QC p. 1-5. [doi: 10.1109/PIMRC.2017.8292361]
90. Rau HH, Hsu CY, Lee YL, Chen W, Jian WS. Developing electronic health records in Taiwan. IT Prof 2010 Mar;12(2):17-25. [doi: 10.1109/MITP.2010.53]
91. Li YCJ, Yen JC, Chiu WT, Jian WS, Syed-Abdul S, Hsu MH. Building a national electronic medical record exchange system - experiences in Taiwan. Comput Methods Programs Biomed 2015 Aug;121(1):14-20. [doi: 10.1016/j.cmpb.2015.04.013] [Medline: 26001420]
92. Hwang HG, Han HE, Kuo KM, Liu CF. The differing privacy concerns regarding exchanging electronic medical records of internet users in Taiwan. J Med Syst 2012 Dec;36(6):3783-3793. [doi: 10.1007/s10916-012-9851-1] [Medline: 22527781]
93. Yang CM, Lin HC, Chang P, Jian WS. Taiwan's perspective on electronic medical records' security and privacy protection: lessons learned from HIPAA. Comput Methods Programs Biomed 2006 Jun;82(3):277-282. [doi: 10.1016/j.cmpb.2006.04.002] [Medline: 16730852]
94. Taiwan medical AI and data portal. NYCU Data Management Center. 2023. URL: https://data.dmc.nycu.edu.tw/ [accessed 2025-12-22]
95. Kaye J, Whitley EA, Lund D, Morrison M, Teare H, Melham K. Dynamic consent: a patient interface for twenty-first century research networks. Eur J Hum Genet 2015 Feb;23(2):141-146. [doi: 10.1038/ejhg.2014.71] [Medline: 24801761]
96. Williams H, Spencer K, Sanders C, et al. Dynamic consent: a possible solution to improve patient confidence and trust in how electronic patient records are used in medical research. JMIR Med Inform 2015 Jan 13;3(1):e3. [doi: 10.2196/medinform.3525] [Medline: 25586934]
97. Lee H, Lee U. Toward dynamic consent for privacy-aware pervasive health and well-being: a scoping review and research directions. IEEE Pervasive Comput 2022 Oct 1;21(4):25-32. [doi: 10.1109/MPRV.2022.3210747]
98. Budin-Ljøsne I, Teare HJA, Kaye J, et al. Dynamic consent: a potential solution to some of the challenges of modern biomedical research. BMC Med Ethics 2017 Jan 25;18(1):4. [doi: 10.1186/s12910-016-0162-9] [Medline: 28122615]

XSL•FO
RenderX

99. Albalwy F, Brass A, Davies A. A blockchain-based dynamic consent architecture to support clinical genomic data sharing (consentchain): proof-of-concept study. JMIR Med Inform 2021 Nov 3;9(11):e27816. [doi: 10.2196/27816] [Medline: 34730538]
100. Goncharov L, Suominen H, Cook M. Dynamic consent and personalised medicine. Med J Aust 2022 Jun 20;216(11):547-549. [doi: 10.5694/mja2.51555] [Medline: 35611469]
101. Vlahou A, Hallinan D, Apweiler R, et al. Data sharing under the general data protection regulation: time to harmonize law and research ethics? Hypertension 2021 Apr;77(4):1029-1035. [doi: 10.1161/HYPERTENSIONAHA.120.16340] [Medline: 33583200]
102. Dilsizian SE, Siegel EL. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. Curr Cardiol Rep 2014 Jan;16(1):441. [doi: 10.1007/s11886-013-0441-8] [Medline: 24338557]
103. Pereira T, Morgado J, Silva F, et al. Sharing biomedical data: strengthening AI development in healthcare. Healthcare (Basel) 2021 Jun 30;9(7):827. [doi: 10.3390/healthcare9070827] [Medline: 34208830]
104. Paton C, Kobayashi S. An open science approach to artificial intelligence in healthcare. Yearb Med Inform 2019 Aug;28(1):47-51. [doi: 10.1055/s-0039-1677898] [Medline: 31022753]
105. Lim L, Lee HC. Open datasets in perioperative medicine: a narrative review. Anesth Pain Med (Seoul) 2023 Jul;18(3):213-219. [doi: 10.17085/apm.23076] [Medline: 37691592]
106. Wilson JR, Prevedello LM, Witiw CD, Flanders AE, Colak E. Data liberation and crowdsourcing in medical research: the intersection of collective and artificial intelligence. Radiol Artif Intell 2024 Jan;6(1):e230006. [doi: 10.1148/ryai.230006] [Medline: 38231037]
107. Kasparick M, Andersen B, Franke S, et al. Enabling artificial intelligence in high acuity medical environments. Minim Invasive Ther Allied Technol 2019 Apr;28(2):120-126. [doi: 10.1080/13645706.2019.1599957] [Medline: 30950665]
108. Koh DM, Papanikolaou N, Bick U, et al. Artificial intelligence and machine learning in cancer imaging. Commun Med (Lond) 2022;2(1):133. [doi: 10.1038/s43856-022-00199-0] [Medline: 36310650]
109. Levin MG, Rader DJ. Polygenic risk scores and coronary artery disease: ready for prime time? Circulation 2020 Feb 25;141(8):637-640. [doi: 10.1161/CIRCULATIONAHA.119.044770] [Medline: 32091922]
110. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023 Jan 3;10(1):1. [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]
111. Milham MP, Craddock RC, Son JJ, et al. Assessment of the impact of shared brain imaging data on the scientific literature. Nat Commun 2018 Jul 19;9(1):2818. [doi: 10.1038/s41467-018-04976-1] [Medline: 30026557]
112. Patterson E, McBurney R, Schmidt H, Baldini I, Mojsilovic A, Varshney KR. Dataflow representation of data analyses: toward a platform for collaborative data science. IBM J Res & Dev 2017 Nov 1;61(6):9. [doi: 10.1147/JRD.2017.2736278]
113. El Naqa I, Haider MA, Giger ML, Ten Haken RK. Artificial intelligence: reshaping the practice of radiological sciences in the 21st century. Br J Radiol 2020 Feb 1;93(1106):20190855. [doi: 10.1259/bjr.20190855] [Medline: 31965813]
114. Kras A, Celi LA, Miller JB. Accelerating ophthalmic artificial intelligence research: the role of an open access data repository. Curr Opin Ophthalmol 2020 Sep;31(5):337-350. [doi: 10.1097/ICU.0000000000000678] [Medline: 32740059]

Abbreviations

AI: artificial intelligence
CT: computed tomography
EMR: electronic medical record
FL: federated learning
HIPAA: Health Insurance Portability and Accountability Act
ML: machine learning
MRI: magnetic resonance imaging
PET: privacy-enhancing technology
PHI: protected health information

XSL•FO
RenderX

Edited by G Eysenbach, T Leung; submitted 31.Mar.2024; peer-reviewed by Imran, Y Sun; revised version received 28.Oct.2025; accepted 29.Nov.2025; published 30.Jan.2026.

Please cite as:
Yang A, Pan ML, Lu HHS, Lien CY, Wang DW, Chen CH, Tarng DC, Niu DM, Chiou SH, Wu CY, Sun YC, Chen SA, Wang SJ, Sheu WHH, Lin CH
Assessing the Evolution and Influence of Medical Open Databases on Biomedical Research and Health Care Innovation: A 25-Year Perspective With a Focus on Privacy and Privacy-Enhancing Technologies
J Med Internet Res 2026;28:e58954
URL: https://www.jmir.org/2026/1/e58954
doi:10.2196/58954

© Albert Yang, Mei-Lien Pan, Henry Horng-Shing Lu, Chung-Yueh Lien, Da-Wei Wang, Chih-Hsiung Chen, Der-Cherng Tarng, Dau-Ming Niu, Shih-Hwa Chiou, Chun-Ying Wu, Ying - Chou Sun, Shih-Ann Chen, Shuu-Jiun Wang, Wayne Huey-Herng Sheu, Chi-Hung Lin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Africa's Digital Health Revolution: The Digital Fit-Viability Model to Move From Innovation to Scaled Implementation

Afra Jiwa[1,2*], MSc, MBBS; Antony Ngatia[3*], MSc, MBChB; Karim Benali[4], MS, MD; Niclas Boehmer[5], PhD; Sangu Delle[6], MBA, JD, PhD; Patrick Emedom-Nnamdi[7], PhD; Chris Opoku Fofie[8], MD, FGCS, MPH; Christine M O'Brien[9], BS, PhD; Tobi Olatunji[10], MSc, MBBS; Kate Obayabgona[1], MPH, MD; Milind Tambe[11], MSc, PhD; Richard Ribon Fletcher[12], BS, MS, PhD; Adeline Adwoa Boatin[1,13*], MPH, MD; Bethany Hedt-Gauthier[1,14,15*], PhD

[1]Program for Global Surgery and Social Change, Harvard Medical School, Boston, MA, United States
[2]Usher Institute, University of Edinburgh, Edinburgh, United Kingdom
[3]Statsspeak Analytics, Nairobi, Kenya
[4]Global Health and Service Advisory Council, Harvard Medical School, Boston, MA, United States
[5]Hasso Plattner Institute, University of Potsdam, Postdam, Germany
[6]CarePoint, Accra, Ghana
[7]Department of Biostatistics, Harvard T.Chan School of Public Health, Boston, MA, United States
[8]Ghana Health Service, Accra, Ghana
[9]Department of Biomedical Engineering, Washington University in St. Louis, St Louis, MO, United States
[10]Intron Health, London, United Kingdom
[11]Center for Research on Computation and Society, Harvard John a. Paulson School of Engineering and Applied Sciences, Boston, MA, United States
[12]Mechanical Engineering Department, Massachusetts Institute of Technology, Boston, MA, United States
[13]Department of Obstetrics and Gynecology, Massachusetts General Hospital, Boston, MA, United States
[14]Department of Maternal Child Health, University of North Carolina Chapel Hill, Chapel Hill, NC, United States
[15]Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA, United States
[*]these authors contributed equally

Corresponding Author:
Adeline Adwoa Boatin, MPH, MD
Department of Obstetrics and Gynecology
Massachusetts General Hospital
55 Fruit Street
Boston, MA, 02114
United States
Phone: 1 617 724 3775
Email: adeline_boatin@mgh.harvard.edu

Abstract

Digital innovations hold immense potential to transform health care delivery, particularly in sub-Saharan Africa, where financial, geographical, and infrastructural constraints continue to hinder progress toward universal health care delivery. Although a growing health tech sector offers creative solutions, few digital health interventions reach scaled implementation. In this paper, we present the digital fit/viability model—an adapted determinant framework to describe facilitators and barriers to moving from digital tools to integrated digital health implementation. We then use this model to describe the specific challenges and recommended solutions when developing digital health tools for health systems in sub-Saharan Africa.

(J Med Internet Res 2026;28:e63495) doi:10.2196/63495

KEYWORDS
digital health; Africa; digital health solutions; digital technologies; health care access; fit/viability model

XSL•FO
RenderX

Introduction

Digital tools can improve health care delivery and enhance health outcomes by addressing long-standing challenges in health care access, quality, and efficiency [1]. This potential is particularly salient in sub-Saharan Africa, where persistent barriers to health care, including inadequate infrastructure, workforce shortages, and resource limitations, could be alleviated or entirely bypassed with digital health interventions [2,3]. There are notable examples of scaled digital health interventions in sub-Saharan Africa. For example, District Health Information Software 2, a digital informatics platform operating in 40 countries, facilitates electronic data entry at health facilities, making information available for surveillance and program monitoring in near real time [4]. Zipline offers a digital platform for immediate order and drone-led distribution of health products in Rwanda [5]. In Kenya, M-TIBA allows users to access their health insurance through a mobile payment technology called m-pesa [6]. However, despite some successes, the digital health field suffers from "pilotitis," where the vast majority of digital interventions do not move beyond the pilot phase [7].

Over the last decade, the World Health Organization (WHO) and the World Bank have issued several policy documents outlining priority areas for digital interventions, as well as identifying systems gaps that need to be addressed for digital health interventions to achieve their full potential [8-11]. However, these documents do not systematically specify the facilitators and barriers that affect the progression of a specific digital health tool from concept to implementation. Drawing from our experience in developing, researching, and implementing digital health interventions in sub-Saharan Africa, we introduce the digital fit/viability model (dFVM) for health. This model adapts the fit/viability model developed by Liang et al [12] for the use of mobile technology in business and is an example of an implementation science determinant framework to describe the facilitators and barriers that influence implementation outcomes. In this paper, we explain how this model can be adapted for digital health interventions and use dFVM to illustrate challenges specific to sub-Saharan Africa as well as the potential solutions.

Positionality Statement

This work is the result of a 2-day convening of a multidisciplinary group of clinicians, researchers, technologists, entrepreneurs, policymakers, and health delivery service providers. The group included individuals with experience in Ghana, Kenya, Nigeria, Rwanda, Tanzania, Uganda, United States, and United Kingdom, and spanned different age groups, ethnicities, genders, and professional experience; all had experience of working within the African digital health ecosystem. The group held the shared perspective that digital tools have the potential to transform health care in sub-Saharan Africa. From this perspective, we discussed challenges to the implementation of digital innovations in these contexts and strategies to develop, evaluate, and scale context-appropriate and sustainable digital health solutions.

Digital Tools, Interventions, and Implementation

Here, we define digital health tools, interventions, and implementation to clarify the role of dFVM. Digital health tools encompass a variety of technology-enabled products and services developed to improve health care services [13]. Broadly, digital health tools have been categorized into virtual interactions (such as platforms for teleconsultations), paperless data (such as electronic health records), patient self-care (such as mobile apps to support chronic disease management), patient self-service (such as e-booking platforms), decision intelligence systems (such as hospital patient flow management systems), and workflow automation (such as medical equipment tracking systems using radiofrequency identification) [13]. Digital health interventions refer to the services, practices, and strategies that utilize digital health tools. For example, Zipline is an intervention that engages a cluster of digital tools—drones, online ordering systems, etc—to facilitate timely order-to-delivery of health products. Finally, digital health implementation involves integrating a digital health intervention into broader health care delivery. For instance, the successful deployment of Zipline is embedded within the larger health care framework, where the ordered health products are identified as part of patient care and delivered accordingly, outside of the Zipline system itself. Despite a surge in digital health tools over the past decade, few have successfully been integrated into interventions, and even fewer digital health interventions have achieved widespread, sustained implementation into health care systems, particularly in sub-Saharan Africa [1].

The Fit/Viability Model

Liang et al [12] developed the fit/viability model in 2007 to support the adoption of new technology in the business sector [14]. In this original framework, "fit" evaluates how well a technology aligns with its intended environment, while "viability" assesses whether the technology can be feasibly implemented given the organization's constraints. Both fit and viability are essential for a technology to succeed at scale within a business. These overarching "fit" and "viability" questions have been valuable for our team when developing digital health tools as part of complex interventions in sub-Saharan Africa. In the next section, we explain how we adapted Liang et al's [12] framework and incorporated additional subdomains to create dFVM for health (Figure 1; [15]).

XSL•FO
RenderX

Figure 1. The digital fit/viability model (adapted from Liang et al [12], which is published under Creative Commons Attribution 4.0 International License [15]).

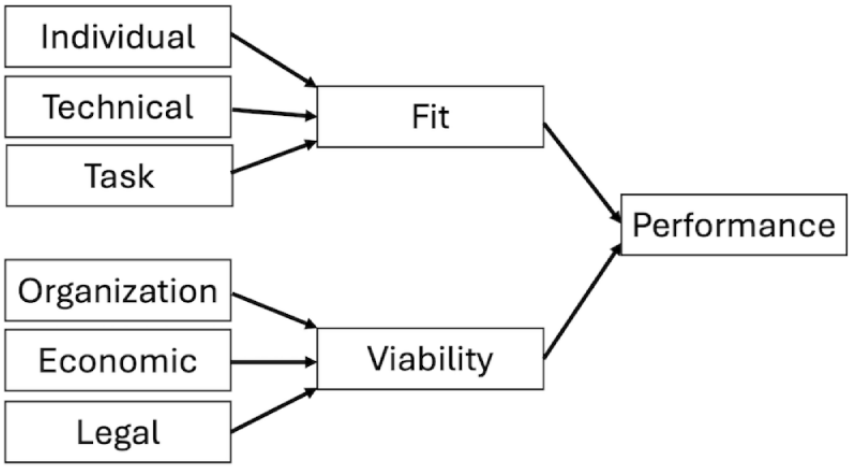

dFVM Concept

Defining the "Fit" of Digital Health Tools

Similar to that defined by Liang et al [12], we define "fit" as the appropriateness of a digital health tool in a specific setting as part of a specific intervention. Generally, we are asking questions such as "Who will be using this digital health tool?", "What is this digital health tool designed to do?", and "Does this digital health tool align with the organizations aims?" (Figure 1). These questions are best addressed in the early design and conception stages. Although these fit concepts are important for a digital health tool in any context, the particular challenges in sub-Saharan Africa require deeper exploration of these questions.

Individual Fit

Individual fit, which was not included in the original fit/viability model, assesses whether the digital health tool is appropriate for those who engage with the tool. This category emphasizes that technologies should be designed with end users in mind.

Challenge

In sub-Saharan Africa, one particular challenge for individual fit is the varied levels of digital literacy [16,17] and access to digital platforms [9,18]. For example, a digital intervention that requires a patient to have a smartphone to engage with an app may fail either because the patient does not own a phone and/or because the patient is not comfortable navigating a digital App.

Solution

As a first step, anyone developing a digital tool should conduct end user digital accessibility assessments, including assessing digital literacy, device ownership, and access to other system requirements such as cell networks or electricity. There are numerous tools to assess digital literacy [14], but few are designed specifically for a sub-Saharan African context. One alternative is to outline the gaps in device ownership or digital literacy that must be addressed when deploying the digital health intervention so that anyone adopting a digital health tool understands the pathway to ensure individual fit.

Technical Fit

Technical fit assesses whether the technological infrastructure exists to support the digital health tool. This domain was not included in the original fit/viability model developed by Liang et al [12], which focused on deploying e-commerce tools in high-income country contexts—settings already primed for new technologies. In our iteration, we consider the potential mismatch between the infrastructure in which a tool was designed and the setting in which it will be deployed.

Challenge

In many sub-Saharan African settings, foundational infrastructure such as electricity and internet remains unreliable, and hardware needs such as hardware durability, power consumption, and maintenance are often overlooked. Tools developed for high-income country contexts may rely on stable cloud access, regular software updates, or consumables that are expensive or difficult to source in sub-Saharan Africa. As a result, imported technologies often end up in "equipment graveyards," unused due to incompatibility with local systems, lack of technical support, or excessive training requirements [19-22].

Solution

The first question should often be "Are there nondigital solutions that are equally suited for this task?", ensuring that implementers are not missing more accessible, viable, and equally impactful solutions. When digital health interventions offer clear advantages, developers must assess infrastructure needs early and design for low-resource environments, considering offline functionality, portability, power efficiency, and local repair capacity. Nationally, more advocacy is needed to support initiatives that expand digital infrastructure. Projects like Health Connect Africa [23], a partnership between Centre for Disease Control Africa and Global System for Mobile Communications Association, which aims to connect 10,000 health care facilities to the internet by 2030, and Power Africa [24], which focuses on electrifying health facilities, are examples of essential infrastructure-building efforts that make digital health tools viable at scale.

XSL•FO
RenderX

Task Fit

We expand Liang et al's [12] considerations of task fit to assess whether a digital health tool is able to complete the task for which it is designed. A tool may be well matched to the infrastructure and the user but still fail to achieve the task for which it is intended.

Challenge

In the digital health field, we often start with tools developed for high-resource environments and adapt these tools to accomplish specific tasks in sub-Saharan Africa [25], which can lead to numerous problems. One notable example is for digital tools that engage artificial intelligence. These often do not include data representative of individuals in sub-Saharan Africa and can have high bias and low validity when implemented in those settings [26].

Solution

First, we must center health program implementers and policymakers to identify the priority gaps we are addressing through digital health interventions. Once we have these priority areas, we must move beyond simply adopting tools and instead validate them through dedicated studies, be willing to adapt or retrain them for context, and, when necessary, design tools from scratch. For AI-enabled digital tools, better and more accessible datasets are needed on the African continent. Although funding and data formatting have been obstacles, recent initiatives such as the National Institutes of Health–funded DS-i Africa [27] along with Masakhane [28], the Lacuna Fund, and Zindi [29], are actively creating shared open repositories. These platforms will accelerate health care innovation by providing representative, high-quality data that directly addresses the bias and validity issues of the existing tools.

Defining the "Viability" of Digital Health Tools

Liang et al [12] considered viability as the organization's economic, technical, and social readiness to adopt a digital tool. We have expanded this to also include legal and ethical considerations that present significant challenges when developing digital tools for interventions and implementation in sub-Saharan Africa. The viability domain considers "Can this tool or intervention thrive in the ecosystem into which it is being deployed?", "Does this tool or intervention have the legal and financial backing to be sustainable?", and "Is the technical infrastructure available to support this tool?" Although fit questions should be addressed early, viability questions are addressed as the digital tool is integrated into interventions and implemented on small or broad scales.

Organization Viability

In the original model, organization viability describes factors such as management support, digital literacy of team members, user competence, and experience [12]. We have shifted several of these individual features such as digital literacy and user competence to "fit" since these must be addressed early as the digital tool is being developed. We have expanded the management support to include plausible integration of the tool into program implementation structures. Adopting technologies that are well matched to their organizational environment can promote uptake and sustainability.

Challenge

Organizational viability considers whether health systems and institutions have the leadership, governance, and operational capacity to sustain digital health interventions. In sub-Saharan Africa, frequent staff turnover and fragmented leadership structures often undermine long-term adoption [30]. Ministries of health may have alternative health priorities or lack the resources to integrate new tools across departments and levels of care. These factors mean that even technically sound and contextually appropriate interventions may struggle to move beyond pilot projects if organizational systems are not aligned to support them.

Solution

Early involvement of multidisciplinary teams is key to assessing feasibility and organizational readiness. Usability testing can help establish workflows, training, and tools that support smooth adoption. For developers, designing solutions that integrate easily with existing infrastructure and that can be bundled with current systems promotes interoperability and long-term alignment. Engaging leadership early and ensuring interventions reflect national health priorities helps secure ownership, buy-in, and resource allocation, thereby reducing the risk of tools operating in isolation from the systems they aim to strengthen.

Economic Viability

In the study of Liang et al [12], economic viability encompasses a cost-effectiveness and cost-benefit analysis. We extend this definition to ask whether a digital health intervention can be sustained financially, including through Ministries of Health and Finance, with the support of philanthropy or multilateral organizations or through the private sector.

Challenge

The digital health sector in sub-Saharan Africa is a patchwork of solutions, with many promising digital health interventions failing to achieve scale due to a lack of coordinated national strategies or shared infrastructure [31]. This decentralized approach, often led by individual clinicians, researchers, and small nongovernmental organizations, limits the ability to benefit from economies of scale [32]. This is compounded by persistent infrastructural barriers like high data costs, unreliable electricity, and poor connectivity as well as inconsistent funding avenues [33]. These issues place a disproportionate financial burden on smaller organizations and new companies. The high cost of entry and limited seed funding make it difficult for promising early-stage innovations to gain traction, leading to isolated successes rather than widespread impact.

Solution

To ensure the continued funding of digital health interventions, a combination of coordinated investment and new funding models are needed. Blended financing, which combines public funds, donor contributions, and socially responsible investment, can help bridge the gap and move interventions from pilot to scale [34]. Health systems–level investment into basic

XSL•FO
RenderX

infrastructure and platform will make any single digital health intervention more economical and cost-beneficial [9].

Legal Viability

Legal viability refers to the presence of clear, enforceable policies that govern the implementation of digital health technologies, with a focus on data privacy, security, and system interoperability. Although this was not included in Liang et al's [12] model, in our experience, this has been a major challenge and often missed step for digital health interventions in sub-Saharan Africa.

Challenge

National and regional policies for digital health interventions are nascent in sub-Saharan Africa. Without clear regulations on data privacy, security, and interoperability, implementers face an uncertain and rapidly changing regulatory landscape [35-37]. This lack of maturity in frameworks creates significant uncertainty for both innovators and investors, making it difficult to establish a strong foundation for technological advancement [35].

Solution

Strengthening legal viability requires investing in national and regional regulatory capacity and developing robust data governance systems. This includes improving regulations to ensure trustworthiness while still supporting innovation [38]. Harmonizing standards across borders will facilitate the ability to share and use tools beyond the specific contexts for which they were developed [36].

dFVM Application

Table 1 shows a summary of the application of dFVM to identify the challenges and possible solutions for implementing digital health interventions in sub-Saharan Africa.

Table 1. Applying dFVM[a] to identify the challenges and possible solutions for implementing digital health interventions in sub-Saharan Africa.

Domains, subdomains of dFVM	Challenges and solutions identified when applying dFVM to sub-Saharan Africa	
	Challenges	Solutions
Fit		
Individual	Limited digital literacy and access to digital platforms among potential end users.	Conduct end user digital assessments, including digital literacy and device ownership assessments prior to tool development. Build digital skills and awareness as part of intervention development.
Technical	Unreliable electricity, frequent power outages, poor internet connectivity, high cost of mobile internet data.	Consider nondigital alternatives. Assess infrastructural capacity early and develop tools that comply to that infrastructure. Invest in foundational infrastructure to connect facilities, health workers, and patients.
Task	Imported technologies may not align with the actual tasks or workflows in health systems. Tools often assume task environments similar to high-income countries. Artificial intelligence tools can also perform poorly due to algorithmic bias.	Address priority areas identified by health practitioners and policymakers in the local context. Ensure co-design with frontline health workers to reflect actual care tasks and workflows. Test, validate, and adapt digital health tools as needed. Build Africa-specific data repositories.
Viability		
Organizational	Fragmented leadership, staff turnover, and organizational priorities, which differ from the tools that are developed.	Bundle tools with existing infrastructure, promoting interoperability and integration; engage leadership early; and promote system-level alignment and ownership.
Economic	Insufficient funding and lack of sustainable financing mechanisms leads to premature abandonment of promising initiatives or poor scaling of established interventions.	Invest in national digital health platforms, which will reduce the per-tool costs. Prioritize national funding models, complemented by blended public-private financing.
Legal	Lack of clear policies and regulations regarding data privacy, security, and interoperability.	Develop robust data governance and digital regulations for countries. Seek, where possible, cross-continent regulations to facilitate portability.

[a]dFVM: digital fit/viability model.

Discussion

We developed dFVM to help individuals designing digital health tools and interventions, particularly innovators working in sub-Saharan Africa, think more critically about the development of these tools and interventions with the goal of successful digital health implementation. We have identified 3 other determinant frameworks relevant to digital health interventions, namely, the Train, Restructure, Incentivize, Mandate, Integrate (TRIMI) framework [16]; a modified version of the Consolidated Framework for Implementation Research (mCFIR) for mobile health tools [17]; and a third developed by Olu et al [3]. Each of these frameworks offers a different vantage to consider when developing digital health tools and interventions. For this paper, we adapted the fit/viability model because this most closely mirrored how we approach our digital tool development—asking questions about whether a digital health tool will fit a specific task, for a specific user, in a specific

XSL•FO
RenderX

infrastructure, and whether a digital health tool achieves the organizational integration, financing, and legal compliance needed to be viable.

Although applicable to any digital health tool, we made sure that dFVM addresses the unique obstacles we have encountered in sub-Saharan Africa. Throughout our description of dFVM, we have highlighted these challenges and offered solutions. Broadly, these solutions fall into 2 buckets. First, what is the responsibility of the individuals developing the tools and interventions? These individuals must design for context and, through the deployment of digital health interventions, build institutional capacity for digital health interventions more broadly. Second, in the context of sub-Saharan Africa, governments play a crucial role, particularly in ensuring the viability of digital health interventions. Investing in digital infrastructure provides a foundation for sustaining various tools and reduces per-tool costs. Moreover, establishing robust legal frameworks, including cross-national agreements, ensures compliance and demonstrates government commitment as partners alongside private and public technology providers.

dFVM offers high-level guidance for individuals developing digital health tools and interventions to ensure their fit and viability within health systems. However, we recognize that moving from idea-to-implementation requires more detailed steps and involves multidisciplinary teams capable of navigating these complexities. As a next step, we are expanding dFVM and integrating it with other digital health implementation science frameworks to develop a step-by-step roadmap to guide teams throughout the development process.

Conclusion

We believe that digital health interventions have great potential to transform health care delivery and improve outcomes worldwide. However, successfully implementing these interventions in sub-Saharan Africa requires careful attention to their fit and viability within local contexts. dFVM maps challenges across both digital health design and deployment environments and links these challenges to practical solutions drawn from lived experience. Although digital health tool developers are responsible for addressing these considerations, governments in sub-Saharan Africa play a crucial role in ensuring that the necessary organizational and legal frameworks are established to support the adoption of digital health interventions.

Acknowledgments

We acknowledge the contributions of all authors on this paper and the role of the Harvard Radcliffe Institute in hosting the seminar from which this paper was conceptualized.

Funding

The workshop was hosted by the Harvard Radcliffe Institute. Additional funding was provided by the Harvard Global Health Institute and Harvard's Motsepe Presidential Research Accelerator Fund for Africa.

Authors' Contributions

BH-G and AAB conceptualized and acquisitioned funding for this work. AJ and AN were writers of the original draft with intensive inputs from NB, PE-N, and TO. All authors participated in the investigation, developing the content through an iterative engagement process, and the writing (reviewing and editing).

Conflicts of Interest

None declared.

References

1. Holst C, Sukums F, Radovanovic D, Ngowi B, Noll J, Winkler AS. Sub-Saharan Africa-the new breeding ground for global digital health. Lancet Digit Health 2020 Apr;2(4):e160-e162 [FREE Full text] [doi: 10.1016/S2589-7500(20)30027-3] [Medline: 33328076]
2. Digital health systems in Africa. IQVIA. 2023 Sep 23. URL: https://www.iqvia.com/locations/middle-east-and-africa/library/white-papers/digital-health-systems-in-africa?_gl=1*kezpey*_up*MQ..*_gs*MQ [accessed 2025-08-08]
3. Olu O, Muneene D, Bataringaya JE, Nahimana M, Ba H, Turgeon Y, et al. How can digital health technologies contribute to sustainable attainment of universal health coverage in Africa? A perspective. Front Public Health 2019;7:341 [FREE Full text] [doi: 10.3389/fpubh.2019.00341] [Medline: 31803706]
4. DHIS2. URL: https://dhis2.org/ [accessed 2025-04-04]
5. Zipline Drone Delivery & Logistics. URL: https://www.flyzipline.com/ [accessed 2024-06-20]
6. M-TIBA. URL: https://mtiba.com/ [accessed 2025-04-04]
7. Egermark M, Blasiak A, Remus A, Sapanel Y, Ho D. Overcoming pilotitis in digital medicine at the intersection of data, clinical evidence, and adoption. Advanced Intelligent Systems 2022 May 26;4(9):2200056. [doi: 10.1002/aisy.202200056]
8. Global Strategy on Digital Health 2020-2025. Geneva: World Health Organization; 2021.

XSL•FO
RenderX

9. WHO Guideline: Recommendations on Digital Interventions for Health System Strengthening. Geneva: World Health Organization; 2019.
10. Kelley E, Zandi D, Krishnamurthy R, Mehl G, Novillo D, Muneene D, et al. Digital Health Platform Handbook: Building a Digital Information Infrastructure (Infostructure) for Health. Geneva: World Health Organization and International Telecommunication Union; 2020.
11. Digital-in-Health: Unlocking the Value for Everyone. Washington, DC: The World Bank; 2023.
12. Liang T, Huang C, Yeh Y, Lin B. Adoption of mobile technology in business: a fit‐viability model. Industr Mngmnt & Data Systems 2007 Oct 02;107(8):1154-1169. [doi: 10.1108/02635570710822796]
13. How digital tools could boost efficiencey in African health systems. McKinsey and Company. URL: https://www.mckinsey.com/industries/healthcare/our-insights/how-digital-tools-could-boost-efficiency-in-african-health-systems [accessed 2024-06-18]
14. Nguyen LAT, Habók A. Tools for assessing teacher digital literacy: a review. J Comput Educ 2023 Jan 19;11(1):305-346. [doi: 10.1007/s40692-022-00257-5]
15. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2025-12-10]
16. Qoseem IO, Okesanya OJ, Olaleke NO, Ukoaka BM, Amisu BO, Ogaya JB, et al. Digital health and health equity: how digital health can address healthcare disparities and improve access to quality care in Africa. Health Promot Perspect 2024 Mar;14(1):3-8 [FREE Full text] [doi: 10.34172/hpp.42822] [Medline: 38623352]
17. Sylla B, Ismaila O, Diallo G. 25 years of digital health toward universal health coverage in low- and middle-income countries: rapid systematic review. J Med Internet Res 2025 May 29;27:e59042 [FREE Full text] [doi: 10.2196/59042] [Medline: 40440696]
18. GSMA Intelligence: The Mobile Economy 2022. URL: https://www.gsmaintelligence.com/research/the-mobile-economy-2022 [accessed 2024-06-18]
19. Howitt P, Darzi A, Yang GZ, Ashrafian H, Atun R, Barlow J, et al. Technologies for global health. Lancet 2012 Aug 04;380(9840):507-535. [doi: 10.1016/S0140-6736(12)61127-1] [Medline: 22857974]
20. Velazquez-Berumen A, Manimaran M. Driving innovation in low resource settings. World Hosp Health Serv 2016;52(3):7-11. [Medline: 30707806]
21. Perry L, Malkin R. Effectiveness of medical equipment donations to improve health systems: how much medical equipment is broken in the developing world? Med Biol Eng Comput 2011 Jul;49(7):719-722. [doi: 10.1007/s11517-011-0786-3] [Medline: 21597999]
22. Asma E, Heenan M, Banda G, Kirby RP, Mangwiro L, Acemyan CZ, Technical Collaborative Authorship Group. Avoid equipment graveyards: rigorous process to improve identification and procurement of effective, affordable, and usable newborn devices in low-resource hospital settings. BMC Pediatr 2023 Nov 15;23(Suppl 2):569 [FREE Full text] [doi: 10.1186/s12887-023-04362-x] [Medline: 37968578]
23. GSMA signs agreement with Africa Centres for Disease Control Prevention, to harness the power of mobile to combat disease in Africa. Africa Centres for Disease Control and Prevention. 2023. URL: https://africacdc.org/news-item/gsma-signs-agreement-with-africa-centres-for-disease-control-and-prevention-to-harness-the-power-of-mobile-to-combat-disease-in-africa/ [accessed 2024-06-20]
24. USAID's Power Africa launches partnership to electrify health facilities across sub-Saharan Africa as part of President Biden's global infrastructure initiative. Relief Web. 2022 Apr 28. URL: https://reliefweb.int/report/world/usaid-s-power-africa-launches-partnership-electrify-health-facilities-across-sub#:~:text=Today%2C%20as%20part%20of%20the,Africa%20over%20the%20next%20eight [accessed 2024-06-20]
25. Al Meslamani AZ. Technical and regulatory challenges of digital health implementation in developing countries. J Med Econ 2023;26(1):1057-1060 [FREE Full text] [doi: 10.1080/13696998.2023.2249757] [Medline: 37594521]
26. Joseph J. Algorithmic bias in public health AI: a silent threat to equity in low-resource settings. Front Public Health 2025;13:1643180 [FREE Full text] [doi: 10.3389/fpubh.2025.1643180] [Medline: 40771228]
27. National Institutes of Health. DS-I Africa. 2022. URL: https://dsi-africa.org/ [accessed 2024-06-21]
28. Masakhane. URL: https://www.masakhane.io/ [accessed 2024-06-21]
29. Zindi. URL: https://zindi.africa/ [accessed 2024-06-21]
30. Bediang G. Implementing clinical information systems in Sub-Saharan Africa: report and lessons learned from the MatLook project in Cameroon. JMIR Med Inform 2023 Oct 18;11:e48256 [FREE Full text] [doi: 10.2196/48256] [Medline: 37851502]
31. Kaburi BB, Harries M, Hauri AM, Kenu E, Wyss K, Silenou BC, et al. Availability of published evidence on coverage, cost components, and funding support for digitalisation of infectious disease surveillance in Africa, 2003-2022: a systematic review. BMC Public Health 2024 Jun 28;24(1):1731 [FREE Full text] [doi: 10.1186/s12889-024-19205-2] [Medline: 38943132]
32. Serge B, Mbondji E, Humphrey K, Janauschek L. Health data digitalization in Africa. In: AHOP Policy Briefs. Brazzaville: WHO Regional Office for Africa; 2024:1-47.

XSL•FO
RenderX

33. Mbunge E, Jack CL, Sibiya MN, Batani J. Review of implementation barriers and strategic approaches for improving mHealth systems utilization in Africa: Lessons learnt from South Africa and Kenya. Telematics and Informatics Reports 2025 Sep;19:100228. [doi: 10.1016/j.teler.2025.100228]
34. Irihamye E, Hadad J, Ali N, Holthof B, Wafula F, Paton C, et al. Sustainable by design: digital health business models for equitable global health impact in low-income and low-middle-income countries. Mayo Clin Proc Digit Health 2025 Dec;3(4):100261 [FREE Full text] [doi: 10.1016/j.mcpdig.2025.100261] [Medline: 41140345]
35. Nienaber McKay AG, Brand D, Botes M, Cengiz N, Swart M. The regulation of health data sharing in Africa: a comparative study. J Law Biosci 2024;11(1):lsad035 [FREE Full text] [doi: 10.1093/jlb/lsad035] [Medline: 38259628]
36. Torab-Miandoab A, Samad-Soltani T, Jodati A, Rezaei-Hachesu P. Interoperability of heterogeneous health information systems: a systematic literature review. BMC Med Inform Decis Mak 2023 Jan 24;23(1):18 [FREE Full text] [doi: 10.1186/s12911-023-02115-5] [Medline: 36694161]
37. Mamuye AL, Yilma TM, Abdulwahab A, Broomhead S, Zondo P, Kyeng M, et al. Health information exchange policy and standards for digital health systems in Africa: a systematic review. PLOS Digit Health 2022 Oct;1(10):e0000118 [FREE Full text] [doi: 10.1371/journal.pdig.0000118] [Medline: 36812615]
38. Victor Ibukun Adebayo, Adebimpe Bolatito Ige, Courage Idemudia, Osemeike Gloria Eyieyien. Ensuring compliance with regulatory and legal requirements through robust data governance structures. Open Access Res J Multidiscip Stud 2024 Jul 30;8(1):36-44. [doi: 10.53022/oarjms.2024.8.1.0043]

Abbreviations

dFVM: digital fit/viability model
mCFIR: modified version of the Consolidated Framework for Implementation Research
TRIMI: Train, Restructure, Incentivize, Mandate, Integrate
WHO: World Health Organization

Edited by T Leung, G Eysenbach; submitted 24.Jun.2024; peer-reviewed by K Fultz Hollis, T Neumark; comments to author 17.Mar.2025; revised version received 29.Sep.2025; accepted 23.Oct.2025; published 14.Jan.2026.

Please cite as:
Jiwa A, Ngatia A, Benali K, Boehmer N, Delle S, Emedom-Nnamdi P, Fofie CO, O'Brien CM, Olatunji T, Obayabgona K, Tambe M, Fletcher RR, Boatin AA, Hedt-Gauthier B
Africa's Digital Health Revolution: The Digital Fit-Viability Model to Move From Innovation to Scaled Implementation
J Med Internet Res 2026;28:e63495
URL: https://www.jmir.org/2026/1/e63495
doi:10.2196/63495
PMID:

©Afra Jiwa, Antony Ngatia, Karim Benali, Niclas Boehmer, Sangu Delle, Patrick Emedom-Nnamdi, Chris Opoku Fofie, Christine M O'Brien, Tobi Olatunji, Kate Obayabgona, Milind Tambe, Richard Ribon Fletcher, Adeline Adwoa Boatin, Bethany Hedt-Gauthier. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Health Information Adoption Among Patients With Chronic Disease in China: Qualitative Interview Study of Patient–Platform Coshaping

Lin Huang[1], PhD; Guimin Duan[1], PhD; Jiani Tang[1], MA; Peipei He[2], PhD

[1]School of Management, Chengdu University of Traditional Chinese Medicine, 1166 Liutai Avenue, Wenjiang District, Chengdu, Sichuan, China
[2]School of Public Administration, Sichuan University, Chengdu, Sichuan, China

Corresponding Author:
Guimin Duan, PhD
School of Management, Chengdu University of Traditional Chinese Medicine, 1166 Liutai Avenue, Wenjiang District, Chengdu, Sichuan, China

Abstract

Background: Chronic disease management increasingly relies on digital health information. Traditional adoption models conceptualize adoption as an individual decision determined by cognitive evaluations. However, in contemporary platform environments, health information exposure and interaction are shaped by algorithmic curation and platform governance. Understanding how patients with chronic disease engage with these platform-mediated spaces requires examining adoption as an ongoing process rather than a discrete outcome. Such process-oriented understanding remains limited.

Objective: This study aims to construct a theoretical model that captures the dynamic, processual nature of health information adoption among patients with chronic disease in platform-mediated environments, with particular attention to the reciprocal interactions between patient practices and platform structures.

Methods: This study used grounded theory methodology and conducted in-depth semistructured interviews with 32 patients with chronic disease in Chengdu, Sichuan, China, from December 2023 to July 2025. Participants included individuals with various chronic conditions such as hypertension, rhinitis, lumbar disc herniation, urticaria, and other persistent health conditions requiring sustained self-management. The sample was diverse in age, gender, and educational background. Theoretical sampling was used across 3 stages to ensure diversity and category saturation. Data were systematically analyzed through 3 levels of coding: open coding generated 22 categories, axial coding organized these into 7 major categories, and selective coding integrated these into 3 core categories. Particular attention was paid to how patients interpreted, evaluated, and integrated digital health information through continuous interaction with platform technologies and governance structures.

Results: Analysis identified 3 core categories of health information adoption—adoption propensity, platform context, and intervening conditions—which were integrated into a process model. Adoption Propensity comprises 3 interconnected layers, including motivational (disease learning, information verification, and comfort seeking), behavioral (information seeking, analysis, keeping, and adoption), and attitudinal (cognitive, emotional, and contextual resonance). Platform Context includes both information context (information content, information format, and information source) and institutional context (platform technology, platform characteristics, and platform governance). Intervening Conditions consist of individual circumstances (foundational characteristics, information literacy, and illness experience) and environmental conditions (emergency situations, policy changes, and social relationships). The analysis further revealed a cyclical co-construction process, in which patient behaviors and platform environments continuously shape each other through reciprocal interactions.

Conclusions: This study proposes a theoretical model that explores how the interaction between digital platforms and patients influences health information adoption. By adopting an interpretive approach and introducing the patient-platform coconstruction perspective, this research extends existing health information behavior theories to account for platform-era dynamics. The model highlights the crucial role of platforms in structuring patients' ongoing health information practices, providing insights for the design, governance, and optimization of digital health platforms to better support chronic disease self-management.

(J Med Internet Res 2026;28:e85229) doi:10.2196/85229

KEYWORDS

digital health information; digital platforms; information adoption; chronic diseases; grounded theory

XSL•FO
RenderX

Introduction

The rapid expansion of digital health technologies has fundamentally transformed how individuals access, evaluate, and adopt health information [1,2]. Digital health information adoption has become a routine yet consequential information behavior [3], directly influencing health decision-making and long-term health outcomes. Digital platforms now serve as the primary infrastructures through which individuals encounter and make sense of health information. In the Chinese context examined in this study, the public relies on professional health platforms such as Dingxiangyuan and the Health China app, alongside mainstream social media platforms including Weibo, WeChat official accounts, Douyin, and Kuaishou. Together, these platforms constitute an integrated digital health information ecosystem characterized by algorithmic recommendation systems, interactive affordances, content moderation practices, and commercially driven governance logics.

However, the growing dominance of these platforms and their algorithmic systems has profoundly restructured health information environments, altering not only what information users encounter but also how they engage with it. Empirical studies indicate that misinformation, uneven information quality, algorithmic amplification, and insufficient platform governance increasingly complicate individuals' ability to adopt accurate and relevant health information [4-6]. Under these conditions, digital health information adoption no longer occurs within a stable and neutral information environment; instead, it unfolds within continuously mediated, algorithmically curated, and commercially governed platform ecosystems.

Existing theoretical frameworks of digital health information adoption, including the information adoption model, technology acceptance model, unified theory of acceptance and use of technology, theory of planned behavior, and the elaboration likelihood model, have substantially advanced understanding of the attitudinal, cognitive, and social predictors of adoption intentions [7-14]. These frameworks, however, have predominantly been operationalized within post-positivist paradigms that conceptualize adoption as a discrete behavioral outcome predicted by measurable variables. Their central concern is identifying which factors determine whether information will be adopted at a specific decision point. While analytically rigorous, this outcome-oriented orientation treats the information environment as a relatively stable external backdrop. As a result, these models are limited in their capacity to explain how adoption unfolds as an ongoing, experience-based process embedded in dynamic platform environments characterized by algorithmic mediation, passive exposure, fragmented information consumption, and evolving governance interventions.

Conceptually, this limitation reflects a deeper epistemological tension. Variable-centered adoption models prioritize prediction and causal explanation, whereas contemporary platform environments require interpretive understanding of how individuals make sense of information under conditions shaped by algorithmic visibility, social signals, and governance structures. In platform-mediated contexts, health information adoption unfolds through iterative interactions among users, health information, and platform architectures. Accordingly, this study adopts an interpretive epistemological stance, conceptualizing digital health information adoption not as a singular outcome but as a dynamic process of meaning-making and negotiation within sociotechnical systems.

Understanding how these platform-mediated environments structure health information adoption requires examining the technological infrastructures and governance mechanisms that shape information flows and user conduct. Scholarship on platform governance provides a critical lens for this examination. Robert Gorwa conceptualizes platform governance as the combination of technical architectures, institutional rules, and market incentives through which platforms regulate participation and visibility [15]. Related scholarship demonstrates that community standards and content moderation practices operate as quasi-regulatory mechanisms shaping information exchange [16], and that algorithmic systems systematically influence exposure patterns, credibility cues, and behavioral feedback loops [17-25]. In health information contexts, algorithmic curation blends personalization, social endorsement, and regulatory intervention, thereby actively configuring what users see, trust, and consider actionable. Importantly, platform structures do not merely serve as passive channels for information delivery; they constitute active sociotechnical environments that both shape and are reshaped by user practices. While platform governance scholarship provides crucial insights into how digital infrastructures shape information environments, its integration into health information adoption research remains limited, particularly in understanding how patients navigate and respond to these governed spaces.

Within these platform-mediated environments, patients with chronic disease constitute a particularly critical population for examining digital health information adoption. In this study, patients with chronic disease refer to individuals living with long-term health conditions requiring sustained self-management and ongoing information engagement [26], including but not limited to cardiovascular diseases, diabetes, chronic respiratory conditions, chronic musculoskeletal disorders, endocrine disorders, and other persistent health conditions. Chronic diseases account for more than 70% of global mortality and impose substantial and sustained burdens on health systems worldwide [26]. Beyond clinical care, chronic disease management entails continuous decision-making related to medication adherence, lifestyle adjustment, symptom monitoring, and psychological regulation. Prior research conceptualizes patients with chronic disease as highly information-dependent self-managers whose health behaviors are iterative, context-sensitive, and shaped by ongoing information engagement [2]. Their health outcomes are therefore deeply intertwined with how they access, interpret, evaluate, and ultimately adopt digital health information over time.

Despite growing research attention to digital health information behaviors, 3 interrelated gaps remain. First, although health information adoption has been examined across diverse populations, comparatively limited focused attention has been paid to patients with chronic disease, particularly those whose

XSL•FO
RenderX

information engagement is sustained, iterative, and consequential [27]. Second, dominant adoption frameworks emphasize predictive determinants of behavioral intention while insufficiently accounting for the processual, experience-based, and contextually embedded nature of adoption in algorithmically mediated environments [7-14]. Third, existing research rarely integrates platform perspectives into qualitatively grounded analyses of health information adoption, leaving underexplored the reciprocal dynamics through which patients' information practices interact with and respond to platform architectures and regulatory mechanisms.

Addressing these gaps requires reconceptualizing digital health information adoption as a platform-embedded process shaped by continuous interactions among individual sensemaking, informational characteristics, and platform mechanisms such as algorithmic curation and governance structures. Guided by grounded theory and informed by an interpretive research orientation, this study seeks to develop a process-oriented theoretical account of how patients with chronic disease adopt digital health information within platform-mediated environments. Specifically, this study addresses the following research questions: (1) What key components characterize patients with chronic disease's health information adoption behaviors? (2) How do platform technologies, algorithmic recommendations, and governance mechanisms feature in patients with chronic disease's health information adoption experiences? (3) How can a process-oriented theoretical model be developed to explain the dynamics of health information adoption among chronic disease patients in platform-mediated contexts? By theorizing adoption as an iterative, relational, and platform-embedded process, this study aims to contribute to health information behavior scholarship and to inform the design and governance of digital health platforms in ways that better support sustainable chronic disease self-management.

Methods

Overview

Grounded theory represents a qualitative research methodology embedded within an interpretivist paradigm, wherein the social world is understood as constructed through participants' lived experiences and subjective meanings [28]. Its fundamental objective is to construct a theory from empirical data [29]. Researchers conduct investigations without preconceived theoretical assumptions, engaging in comprehensive and in-depth exploration of phenomena within their natural contexts. Through continuous inductive processes and data condensation, researchers derive categories from raw data, ultimately formulating theoretical frameworks. This study uses a combination of face-to-face and telephone interviews to gather primary data from patients with chronic disease. Data organization and coding are conducted through 3 analytical levels using NVivo (version 14; QSR International) software, accompanied by theoretical saturation testing to ensure the authenticity and reliability of research findings. This study followed the SRQR (Standards for Reporting Qualitative Research) guideline (Checklist 1) [30].

Study Setting and Participants

Theoretical Sampling Strategy

In grounded theory, data saturation serves as the core criterion for determining sample size adequacy. It refers to the stage where, after continuous sampling and data analysis, new data no longer provide novel theoretical insights or categories [29]. At this stage, the sample size is considered saturated, and sampling can be discontinued. This study avoids pre-establishing fixed sample sizes, instead adopting a data-driven approach through continuous coding, theoretical sampling, and saturation testing to ensure the sufficiency and accuracy of theory construction [31].

It is important to clarify that this study used theoretical sampling, not random sampling. Theoretical sampling is a hallmark of grounded theory methodology, wherein sampling decisions are guided by emerging concepts and categories rather than predetermined criteria. Initial participants were selected purposively to ensure diversity in characteristics such as age, disease type, and digital literacy. As analysis progressed, subsequent participants were selected based on their ability to provide insights into emerging categories and to test the boundaries of developing concepts.

Recruitment Setting and Procedures

This study was conducted in urban Chengdu, China, selecting patients with chronic disease from multiple districts, including Wenjiang, Wuhou, Jinjiang, and Shuangliu, for community-based investigation. Recruitment was conducted through multiple channels to ensure diversity: (1) community health centers: Researchers collaborated with community health centers to identify potential participants during routine health checkups and chronic disease management programs, (2) patient support groups: Online and offline communities of patients with chronic disease were contacted to recruit participants, (3) Snowball sampling: Initial participants were asked to refer other patients with chronic disease who might be interested in participating.

The following outlines the specific sample confirmation process across 3 sequential stages:

- Stage 1—Initial Sampling (n=10): Initial participants were selected based on research objectives, with inclusion criteria as follows: (1) Clear understanding of digital health information needs described in the interview protocol, particularly focusing on their perspectives regarding the relevance of digital health information in daily health management; (2) At least 18 years of age, possessing cognitive abilities for coherent expression of experiences and viewpoints, providing rich data for grounded theory analysis; (3) Signed written informed consent, agreeing to interview recording and anonymous data usage while retaining the right to withdraw at any time; (4) Capability to participate in 15 - 50 minute semistructured interviews for in-depth exploration of their digital health information adoption journeys.

Through the first stage of interviews and initial open coding, 3 preliminary categories emerged: digital health information

XSL•FO
RenderX

adoption motivation, digital health information adoption behavior, and digital health information adoption attitude.

- Stage 2—theoretical sampling for diversity (n=5): To maximize category heterogeneity and explore the range of experiences, we recruited diverse samples based on factors such as age, education level, and occupation. This stage specifically targeted older individuals, retirees, and hospital administrators to capture perspectives across different demographic and professional backgrounds. The second stage of interviews revealed new categories, indicating that platform information environments and institutional environments influence patients' digital health information behaviors.
- Stage 3—targeted theoretical sampling (n=15): Consequently, this study supplemented samples to further explore platform-related phenomena. In addition to meeting initial inclusion criteria, supplementary respondents in this stage were required to be patients with chronic disease with certain digital literacy levels, capable of articulating their experiences with various digital platforms and platform features. This targeted sampling allowed for deeper exploration of how platform algorithms, governance, and design influence information behaviors.
- Saturation testing (n=2): In stage 3, no new codes or categories emerged during the coding of the 15th interview sample, suggesting that theoretical saturation had been reached. To verify this, 2 additional samples were collected and analyzed; again, no new codes or categories appeared, confirming the stability of the existing categories and indicating that no further sampling was necessary.

Adopting an iterative approach, this study follows the dynamic sampling framework of grounded theory. This approach focuses on achieving categorical depth by using a heterogeneous sample. The sample size is thus guided by the sufficiency of information, not statistical considerations [29]. In addition, previous studies have successfully used grounded theory for research of moderate complexity, with their sample sizes varying from 18 to 30 [5]. Based on these considerations, the final sample size of 32 participants for this study is deemed appropriate.

Data Collection

In this study, prior to conducting in-depth field interviews, we developed an interview protocol encompassing multiple dimensions, including topics of concern to patients with chronic disease, demand levels, health information behavior processes, and digital platform usage experiences (Table 1).

Table . Interview guide on digital health information adoption by patients with chronic disease.

Category	Interview questions
Basic information	• Q1: Age • Q2: Sex assigned at birth • Q3: Educational background • Q4: Occupation • Q5: Do you have any chronic illnesses? if so, what are they?
Question Design	• Q6：Do you typically follow health-related information? If so, what specific topics or aspects of health information do you focus on? • Q7: In your daily life, how important is health information to you? Why? • Q8: Could you please tell us how you began to seek out and engage with health information after being diagnosed with a chronic illness? • Q9: What channels do you typically use to get health information? How do you evaluate the reliability of that information? And how satisfied are you with its quality and relevance? • Q10: Could you provide an example of how you searched for a specific piece of health information and then put it into practice? What challenges did you face during that process? • Q11: What formats (eg, text, video, infographics) are most common in the health information you follow? Which format do you prefer, and why? Do you have any suggestions for improving how this information is presented? • Q12: How helpful has the health information you've adopted been for your health condition or lifestyle? • Q13: What are your general impressions or feelings about the platforms and individuals (eg, doctors, apps, bloggers) that provide this health information? • Q14: To conclude, do you have any other experiences or thoughts about adopting health information that you'd like to share?

Prior to conducting the formal interview, this study conducted rigorous pretesting of the interview protocol. During testing, research team members alternately assumed roles of interviewer and interviewee, simulating authentic interview scenarios. They evaluated the protocol from perspectives of question clarity, logical coherence, and guidance. Based on pretest feedback, adjustments were made to ambiguously worded questions, question sequences were optimized, and more targeted follow-up statements were added to enhance interview validity and reliability. This pretesting phase laid a solid foundation for the smooth implementation of the formal interviews and helped ensure the quality of the research data.

XSL•FO
RenderX

Data collection and analysis proceeded simultaneously and were closely integrated, following grounded theory principles. Semistructured interviews were conducted from December 2023 to July 2025. Following the interview protocol (Table 1), each interview lasted approximately 15 to 50 minutes. Interviews were conducted either face-to-face or by telephone, according to participants' preferences and practical circumstances. This mixed-mode approach was adopted for two reasons: (1) to accommodate participants with mobility limitations due to their health conditions, thereby enhancing accessibility and inclusivity; and (2) to respect participants' comfort levels and scheduling constraints, as some patients with chronic disease preferred the convenience of telephone interviews due to time availability or personal preference. Prior research has demonstrated that telephone and face-to-face interviews yield comparable data quality in qualitative health research when using semistructured protocols [32]. In our study, no systematic differences in interview depth or quality were observed between the 2 modes. Interview settings were determined by the participants and included private or semiprivate locations such as their homes, cafés, or university office buildings, with no individuals other than the participant and the interviewer present.

The interviews were conducted by the first author, who, during the informed consent process, disclosed her personal experience as a user of digital health information. This appropriate self-disclosure was intended to build trust, reduce power imbalances, and allow participants to comfortably share their genuine experiences. Throughout the interviews, the interviewer maintained strict neutrality, using standardized prompts to avoid influencing the participants' responses. To minimize bias, the research team pretested the interview questions, ensuring that the self-disclosure statements did not contain leading assumptions. Additionally, a 2-step review process was implemented: recordings were reviewed, and verbatim transcripts were independently cross-checked by other authors to avoid inadvertently suggesting answers based on personal experience. These authors were further responsible for ensuring that all issues concerning the accuracy and completeness of the research were properly addressed.

The interview consists of 4 steps. First, the researcher introduces the study and obtains informed consent, giving the participant an opportunity to ask questions or choose to withdraw. This is followed by the question-and-answer session. During interviews, any ambiguities were immediately addressed by requesting clarification from respondents. Furthermore, new questions emerging during interviews were pursued through follow-up questioning until no additional questions arose. After the Q&A, the interviewer summarizes the key discussion points and verifies the accuracy of the information with the participant, inviting them to add any details that may have been overlooked or were incomplete. This step is designed to ensure the comprehensiveness and accuracy of the data collected. Finally, the interviewer thanks the participant for their time and contribution, explains the subsequent processing of the interview data, and assures the participant of the confidentiality of their personal information and responses.

All interviews were conducted with participants' informed consent and were audio-recorded when permission was granted. Audio recordings were transcribed verbatim using a combination of transcription tools, including iFLYTEK and Apple Voice Memos, and were transcribed as soon as possible after each interview. The first author manually checked all transcripts against the original audio recordings to ensure accuracy. Sixteen participants declined audio recording; in these cases, detailed contemporaneous handwritten notes were taken during the interviews and expanded into full interview records immediately afterward. Interviews were conducted in Mandarin Chinese and the Sichuan dialect. All transcripts and interview records were maintained in Chinese for coding and analysis to preserve linguistic nuances. Representative quotations were subsequently translated into English for publication by a bilingual researcher and independently verified by another bilingual team member.

Data Analysis

Data analysis followed grounded theory principles and was conducted iteratively and concurrently with data collection [33]. Following Strauss and Corbin's approach, analysis proceeded through 3 levels of coding: open coding, axial coding, and selective coding. A constant comparative approach was used throughout all coding stages to facilitate theory generation, involving continuous comparison across interview data, codes, and emerging categories. NVivo 14 qualitative data analysis software was used to support data management, coding, and retrieval.

The 3 levels of coding were operationalized as follows:

- Open coding: Interview transcripts were analyzed line by line to identify meaningful units and inductively generate initial concepts from participants' statements. Each segment of text was examined to identify distinct phenomena, actions, or experiences. For example, when participants described "searching for disease information on platforms," "following health bloggers," and "clicking on algorithm-recommended health videos," these distinct concepts were identified and subsequently grouped into the category "Information Seeking." Through iterative team discussions and constant comparison, researchers refined, merged, and organized these concepts, resulting in 22 categories (eg, information seeking, information analysis, illness experience, platform technology, cognitive resonance).
- Axial coding: The 22 categories from open coding were organized into broader major categories based on their logical relationships and conceptual connections. This stage focused on identifying patterns and grouping related categories. Through this process, 7 major categories emerged: individual circumstances, environmental conditions, motivation, behavior, attitudes, information context, and institutional context. For instance, categories related to "information seeking," "information analysis," "information keeping," and "information adoption" were grouped under the major category "behavior."
- Selective coding: The 7 major categories were systematically integrated to identify core categories that could connect the entire theoretical framework. This stage involved identifying the central phenomenon of digital health information adoption by patients with chronic disease

XSL•FO
RenderX

and determining how the major categories related to it. Through this integration process, 3 core categories emerged: (1) Adoption Propensity (integrating motivation, behavior, and attitudes); (2) Platform Context (integrating information context and institutional context); and (3) Intervening Conditions (integrating individual circumstances and environmental conditions). Importantly, selective coding also revealed the dynamic interrelationships among these core categories, particularly the cyclical coconstruction process through which patient behaviors continuously shape and are reshaped by the platform. These core categories and their interrelationships formed the basis of the process model presented in Figure 1. A detailed overview of the coding process, including all categories and their definitions, is provided in the Supplemental Codebook.

Figure 1. A process model of digital health information adoption by patients with chronic disease.

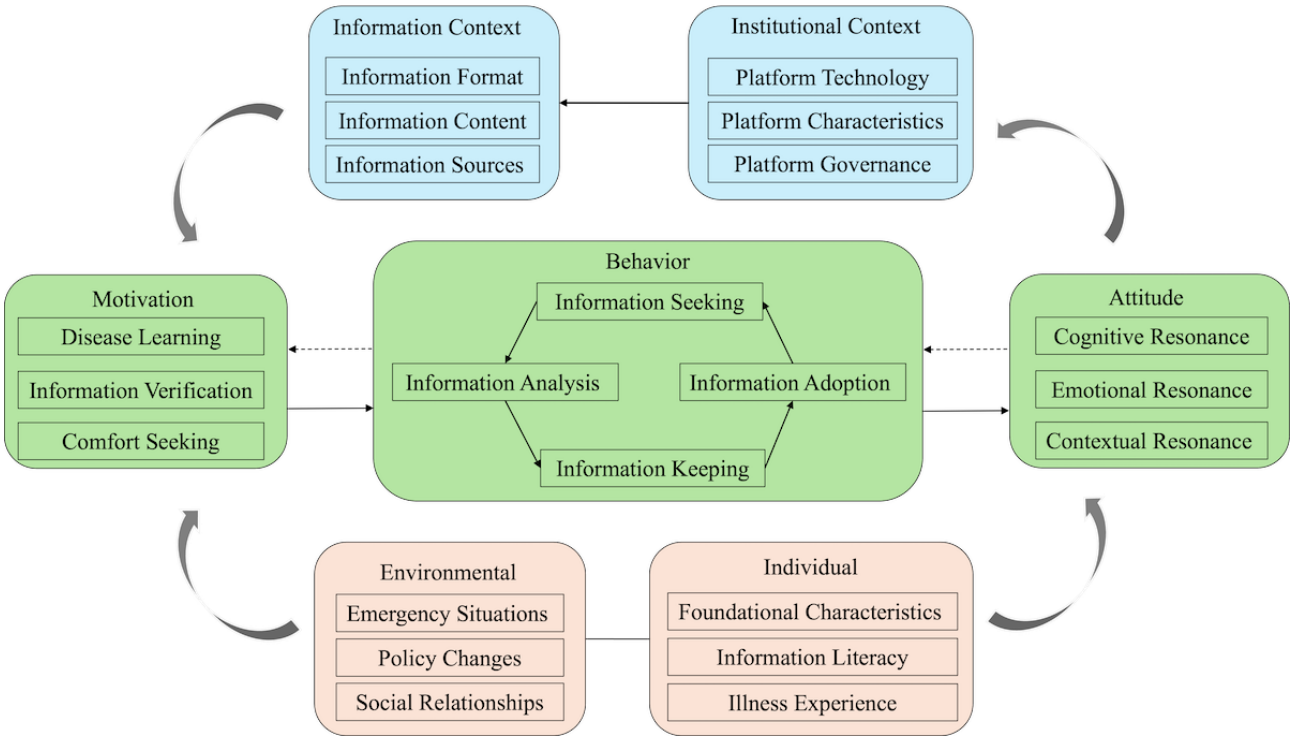

To enhance analytic rigor and reduce subjective bias, multiple strategies were used throughout the analysis. The first author conducted initial coding independently, while 2 additional researchers with formal training in qualitative methods independently coded a subset of transcripts (6/32, 18%). Coding discrepancies were discussed during regular team meetings, during which the coding framework was iteratively refined until consensus was reached. Researchers with complementary disciplinary backgrounds, including information science and clinical medicine, contributed to these discussions by critically reflecting on the clarity, coherence, and applicability of participant-derived codes, without introducing externally imposed categories.

The researcher's reflexivity was integrated throughout data collection and analysis. Reflexive notes were written after each interview and coding session to document researchers' assumptions, reactions, and evolving interpretations. Contextual reflexivity was maintained by examining how participants' health conditions, information environments, and situational contexts shaped their accounts. Relational reflexivity was addressed by continuously assessing the distinction between researchers' perspectives and participants' meanings. To mitigate potential disciplinary bias, the interview guide was collaboratively developed and pilot-tested, interviewer self-disclosure was intentionally limited, and interview transcripts were reviewed by coauthors who had not conducted the interviews to identify potential leading questions or biased prompts.

Theoretical saturation was assessed iteratively throughout the analysis. Initial coding began after the first stage of interviews (n=10), with subsequent interview stages used to refine categories and theoretical relationships. During coding of the 30th interview, no new codes or categories emerged; therefore, 2 additional interviews were conducted and analyzed using the same coding procedures to verify theoretical saturation. As no new codes, categories, or relationships emerged, theoretical saturation was considered to have been achieved.

To further enhance analytic credibility, expert validation was conducted after the coding and theoretical model development were completed, based entirely on participant data. The preliminary model and analytic interpretations were presented to a panel of 5 experts, including 2 health informaticians, 2 information behavior researchers, and 1 specialist in chronic disease management. Experts were invited to assess the logical coherence of model components and relationships, identify potential gaps or inconsistencies, and comment on the theoretical contribution and practical relevance of the model. Expert feedback was used to refine the clarity and articulation of the model and its implications, without introducing new categories or altering the core relationships derived from participants'

XSL•FO
RenderX

accounts. This process functioned as an additional analytic validation strategy to strengthen conceptual coherence while preserving the participant-driven nature of the grounded theory analysis.

Ethical Considerations

This study was approved by the Institutional Ethics Review Committee of the School of Public Administration at Sichuan University (approval number 202512300007). All procedures involving human participants adhered to institutional ethical guidelines for social science research. Written informed consent was obtained from all participants prior to data collection. Participants were informed of the purpose of the study, research procedures, potential risks and benefits, the voluntary nature of participation, and their right to refuse or withdraw at any time without negative consequences. To protect privacy and confidentiality, all interview data were anonymized during transcription and were stored on encrypted, password-protected servers accessible only to the research team. Identifiable information such as names, contact information, and specific workplace or residential identifiers was removed. Quotations used for publication were deidentified, and pseudonyms were assigned. No financial or material compensation was provided to participants due to the minimal-risk nature of the study and in accordance with institutional ethics guidelines. This study involved minimal risk and did not include any medical intervention or clinical experimentation.

Results

Summary

A total of 32 adult patients with chronic disease participated in our interviews, with no dropouts during the process. Participants represented a variety of chronic conditions, including but not limited to diabetes, hypertension, thyroid disorders, chronic respiratory diseases, and musculoskeletal conditions, with disease duration ranging from 1 to 15 years (mean 6.3, SD 4.2 years). All participants regularly used digital platforms to obtain health information, primarily social media platforms (eg, WeChat, Douyin) and professional health platforms (eg, DingXiangYuan.cn). Table 2 summarizes the sociodemographic characteristics of the sampled participants. This detailed participant information provides context for their health information adoption behaviors and enables assessment of the transferability of the findings.

Table . Participant demographics (n=32).

Norm and variable	Value, n (%)
Sex assigned at birth	
Male	8 (25)
Female	24 (75)
Age (years)	
18 - 44	18 (56)
45 - 59	9 (28)
≥60	5 (16)
Educational background	
Primary school	5 (16)
Middle school	4 (13)
High school	3 (9)
College	20 (62)
Occupation	
Professional or technical worker	10 (31)
Service worker	4 (13)
Educator	1 (3)
Self-employed	3 (9)
Farmer	2 (6)
Student	8 (25)
Retiree	4 (13)

The sample comprised predominantly female participants (24/32, 75%). The age distribution covered young adults to seniors, with the majority (18/32, 56%) in the 18 - 44 age range. Educational attainment was relatively high, with 62% (n=20) having college-level education, reflecting the urban setting of Chengdu and the digital literacy requirements for engaging with web-based health platforms. Occupational diversity was achieved, with participants' occupations ranging from farmers to professionals.

XSL•FO
RenderX

Through in-depth interviews and grounded theory analysis with patients with chronic disease, this study seeks to explore their health information adoption behaviors within contemporary digital environments. Research findings reveal that patients' health information behaviors do not constitute a linear, rational, individual adoption process, but rather represent a complex, dynamically cyclical process characterized by bidirectional interaction with platform information, technology, and governance mechanisms. Specifically, motivation-behavior-attitude constitutes patients' 3-tiered adoption propensity, while platform Information Contexts and Institutional Contexts serve as the operational domains for patients' health information behaviors. The unfolding of this cyclical process is further shaped by patients' individual circumstances and broader socioenvironmental conditions. Based on these findings, this study constructs a process model for patients with chronic disease's digital health information adoption (Figure 1).

Core Categories of Digital Health Information Adoption

This study develops a grounded theoretical model of digital health information adoption, comprising 3 core categories: adoption propensity, platform context, and intervening conditions.

Adoption Propensity

In patients' health information adoption processes, "motivation-behavior-attitude" constitutes a dynamic and interconnected, comprehensive explanatory framework.

Motivational

The motivational layer represents the starting point of patients' health information adoption, addressing the question of "why patients actively or passively engage with health information." Patients with chronic disease are driven by 3 primary motivations: disease learning, information verification, and comfort seeking.

Disease learning represents a central and ongoing motivation in patients with chronic disease's engagement with health information. This motivation often becomes particularly salient at the time of diagnosis, when patients encounter urgent knowledge gaps and seek to understand their condition. One participant diagnosed with urticaria described this immediate drive:

> *It came so suddenly, I wanted to know what caused it, so I searched for a lot of information and confirmed I was allergic.* [P2]

This account illustrates how the abruptness of a chronic diagnosis can rapidly transform a person into an active, high-intensity information seeker. Another participant, a patient with thyroid cancer who has practiced self-management for 5 years, captured the pervasiveness of this motivation more broadly:

> *Nowadays everything is online. You can find whatever you're looking for and use it to learn and get information.* [P7]

Together, these accounts illustrate how the onset of chronic illness often prompts an immediate turn to digital platforms as key sources for understanding symptoms, causes, and disease management.

Beyond disease learning, patients are also motivated by a need to verify the health information they encounter, particularly as they grow more experienced with the volume and variability of online content. As one older participant with hypertension expressed:

> *I have been diagnosed for many years and have always prioritized health preservation, but I still feel there should be content teaching us how to spot fake information—for example, can eating less salt really cure high blood pressure?* [P10]

This account reflects a meta-cognitive awareness of misinformation risk that emerges through repeated platform engagement, suggesting that information verification motivation intensifies as patients accumulate digital health literacy over time. Another participant with thyroid nodules shared the following experience:

> *Once I came across an article saying that 'people with thyroid nodules shouldn't eat cruciferous vegetables.' It scared me so much that I didn't even dare to eat broccoli. Later, an official clarification debunked the claim and I learned that eating them cooked doesn't have much impact. I hope public accounts can indicate their information sources and stop scaring people.* [P13]

A third and often underappreciated motivation is comfort seeking—the need to regulate negative emotions generated by chronic illness. Chronic disease imposes persistent uncertainty, physical discomfort, and psychological burden, and patients actively seek out information that can alleviate anxiety and restore a sense of agency. One participant, a patient experiencing long-term pain due to lumbar disc herniation, articulated this clearly:

> *When I see other people saying it works, I kind of feel more confident and hopeful.* [P6]

This statement reveals that information adoption is not always primarily about factual content. For many chronically ill patients, the process is driven by a fundamental need for emotional solace, specifically through the observation of others successfully managing similar conditions, which can be even more motivating than the informational content itself.

Behavioral

The behavioral layer describes the various information behaviors that patients undertake, driven by their motivations. It answers the question, "What do patients do during health information activities?" A patient's health information behaviors can be divided into 4 interconnected and cyclical stages: information seeking, information analysis, information keeping, and information adoption.

First, information seeking encompasses both active and passive modes of access. Patients actively seek information by searching keywords, following specialist accounts, or querying platform

XSL•FO
RenderX

databases. Yet a distinctive feature of the platform era is the prevalence of passive, algorithm-driven exposure, wherein content arrives without deliberate search. One participant experiencing rhinitis captured this dual dynamic:

> *Because my health needs are quite specific, I can actively search for and obtain health information on my own, rather than waiting for public accounts to push content.* [P20]
>
> *Occasionally, when I come across or see pushed articles [about nasal congestion], I just like to click in and have a look.* [P20]

Second, after accessing information, patients move to information analysis, in which they read, evaluate, and integrate content against their personal circumstances. This stage reveals patients as active sense-makers rather than passive recipients. One participant, a participant managing hyperlipidemia, demonstrated a rigorous approach to information analysis:

> *When I encounter a professional term, such as 'low-density lipoprotein,' I proactively search through previously published resources and related articles until I fully grasp its meaning.* [P5]

This behavior highlights that adoption is preceded by a deliberate process of information analysis, where patients function as active decoders of technical knowledge.

Third, information deemed valuable undergoes information keeping—a stage of micro-adoption in which patients record, save, screenshot, or memorize information for later reference. One participant explained:

> *I took screenshots of what they said about 'the three types of alcohol people with high blood pressure can't touch'.* [P10]

This behavioral pattern reveals a layered adoption logic whereby formal adoption is typically preceded by an interim stage of informal curation, reflecting the need for patients to regulate the influx of health information and revisit it according to their own pace.

Finally, patients engage in information adoption, where curated information is applied to produce tangible health behavior changes. The progression from seeking to adoption is often iterative, involving repeated evaluation against lived circumstances. One participant, an office worker experiencing chronic lumbar muscle strain due to prolonged sitting, illustrated the seamless integration of platform-sourced health knowledge into daily life:

> *During my breaks at work, I practiced the 'seated leg raises' I learned on Douyin; I even find time to do a few sets whenever I have a moment to rest.* [P1]

This example highlights how platform-derived health information can be translated into embodied and routine practices, demonstrating the process through which patients integrate digital health knowledge into everyday self-management.

Attitudinal

The attitudinal layer encompasses patients' subjective evaluations, experiences, and feelings toward information, answering the question, "Why do patients adopt this specific information?" Through cognitive resonance, emotional resonance, and contextual resonance, patients with chronic disease continuously perform a holistic "fit assessment" on incoming health information. This resonance framework reveals the interpretive "black box" through which patients move from encountering information to committing to its adoption.

Cognitive resonance refers to patients' tendency to evaluate new information in light of their accumulated knowledge and lived experience. Patients with chronic disease are often inclined to adopt information that coheres with what they already understand, reflecting a comparatively deliberative mode of evaluation. For example, a participant who had studied traditional Chinese medicine for many years critically assessed online content:

> *Some traditional Chinese medicine methods are completely incorrect. If people follow such approaches, it simply will not work.* [P11]

Here, prior domain knowledge enabled the patient to reject information perceived as inconsistent with established understanding. In another instance, a participant with a lumbar disc herniation expressed greater affinity for a blogger's advice:

> *This blogger said that people with my condition can still work. A doctor once told me I had lost the ability to work, but how could that be?* [P16]

In this case, the blogger's message resonated more closely with the patient's self-assessment and perceived capabilities. Across these accounts, cognitively resonant information, whether used to dismiss or to embrace particular claims, appears to carry persuasive weight. Patients actively negotiate competing sources of knowledge based on perceived alignment with their experiential understanding.

Emotional resonance refers to alignment at the affective and experiential level. Patients facing persistent pain, uncertainty, and isolation are driven not only by a need for information but also by a need for recognition, companionship, and hope. The affective framing of health information, ranging from fear and reassurance to validation, substantially determines its resonance. One participant described the anxiety generated by fear-inducing content:

> *There's a lot of information, but some of it is too scary. The title says 'Hypertension patients who don't take medication won't live for more than 5 years,' and after reading it, I felt all panicked inside.* [P10]

This account highlights the emotional consequences of fear-inducing health information. While such content may capture attention through alarming framing, it can also generate anxiety and psychological distress among patients. In contrast, positive emotional resonance actively facilitates adoption. A thyroid cancer participant shared:

> *Seeing other patients who are still healthy ten years after their surgery makes me feel less anxious.* [P7]

XSL•FO
RenderX

The social proof embedded in peer experience functions as a powerful adoption trigger, demonstrating that the emotional community afforded by digital platforms can be a genuine resource for chronically ill patients.

Contextual resonance refers to alignment between information and the patient's practical life circumstances, including economic resources, physical environment, family structure, and daily routines. Patients continuously evaluate information against the pragmatic constraints of their lives; information that is theoretically sound but practically unachievable fails to resonate. One participant with chronic kidney disease articulated this barrier clearly:

> *The hospital says this new medicine is very effective, but it's too expensive, and I don't have the money to buy it.* [P8]

This account illustrates that information adoption is inextricably linked to socioeconomic factors. It underscores that "contextual resonance" is not just about understanding a message, but about the practical feasibility of acting upon it within one's financial constraints.

Platform Context

The platform context refers to the broader digital environment in which patients engage with health information, encompassing both the structure of information and the institutional conditions that shape how it is presented and encountered.

Information Context

The information context refers to the overall information structure conditions across the digital platform environment in which patients are situated. In the contemporary online ecosystem, multiple types of platforms coexist and jointly constitute the informational surroundings encountered by patients with chronic disease. These platforms differ in the information content they host, the formats through which information is presented, and the sources from which information originates. In China, this platform environment includes official health-related platforms associated with the National Health Commission, such as the Healthy China app, where authoritative health information is disseminated by governmental agencies. It also encompasses internet-based medical and health management platforms, such as Haodf.com and DXY.cn, where health information is primarily generated and exchanged by medical professionals and patients. In addition, general social media platforms, including Zhihu and Weibo, contain a substantial volume of health-related content produced and shared by ordinary users. Together, these platforms present distinct yet interconnected information structures that shape patients' everyday health information activities.

Participants' accounts revealed how these platform-specific information structures shaped their engagement patterns. Format emerged as a powerful determinant of engagement:

> *When I'm on WeChat official accounts, I like to read articles with both pictures and text. It's too hard to read if it's all text.* [P9]

This preference reflects how platform-determined content formats condition not merely aesthetic choices but patients' actual capacity to process and engage with health information—patients with lower literacy or visual fatigue may effectively be excluded from text-heavy environments.

Content characteristics similarly influenced access. As one participant noted,

> *I usually see if the title is appealing to me, and based on that, I decide whether to click on it.*
>
> *[P17]*

suggesting that the way health information is presented—such as through titles—affects users' initial engagement. At the same time, participants also emphasized the importance of the content itself. For example, one participant explained,

> *I prefer reading interesting articles. Some content presents complex medical knowledge in a vivid and engaging way, which makes it more appealing to read.* [P9]

This indicates that beyond initial attention, the intrinsic qualities of content, particularly its clarity and vividness, play a key role in attracting and retaining users' interest. Source credibility functioned as a third decisive filter:

> *I still love watching videos posted by doctors.* [P16]
>
> *Because I have a set of health bloggers I regularly follow, I typically stick to the information they provide.* [P27]

This preference indicates that patients rely on source-related cues, including both professional identity and familiarity with specific content creators, when filtering health information on social media platforms.

Institutional Context

The Institutional Context refers to the constraints and enabling conditions embedded at the platform level, including technological infrastructures, structural characteristics, and governance mechanisms. Patients' accounts revealed how these institutional dimensions shaped their health information adoption practices in consequential ways.

Algorithmic recommendation systems emerged as a primary mechanism through which platforms structure patients' information exposure. Digital platforms commonly rely on recommendation algorithms to organize and prioritize information streams, determining what types of health content users are most likely to encounter. One participant described how algorithmic curation becomes the gateway for health information exposure:

> *Most of the time I just watch whatever the platform recommends. If a health video appears on the homepage, I'll usually click on it and see what it says.* [P19]

This account suggests that recommendation algorithms function not merely as neutral distribution mechanisms but as consequential filters that determine which health information enters patients' consideration set, thereby structuring the conditions of possibility for subsequent adoption decisions.

XSL•FO
RenderX

Beyond technological mediation, the structural characteristics of platforms themselves, including institutional endorsement, ownership background, and perceived authority, shape users' trust judgments independent of specific content evaluation. Government-sponsored platforms provide a salient example:

> *Every morning I use the 'Xuexi Qiangguo' app to watch the 'Healthy China' channel. The 'Famous Doctors' Lecture Hall' segment is a must-watch for me every episode. I feel the content there is more reliable.* [P23]

This account indicates that platform-level institutional identity functions as a credibility heuristic that precedes and conditions patients' engagement with individual pieces of health information. The governmental affiliation establishes a baseline trust orientation that influences subsequent information adoption behaviors.

Platform governance mechanisms further shaped patients' trust assessments through visible enforcement actions. Moderation decisions, such as account bans, content removals, or warning labels, serve as authoritative signals that influence both retrospective and prospective credibility judgments. One participant described how a platform's intervention altered their evaluation of a previously trusted source:

> *I used to watch his health videos quite often, but later the platform banned his account for spreading misinformation. After that, I felt the things he said before were probably unreliable.* [P17]

This account reveals that governance interventions function as institutionalized trust-making mechanisms, prompting users to reassess not only future content from a source but also the trustworthiness of previously consumed information.

Intervening Conditions

Throughout the health information adoption process, individual circumstances and environmental conditions shape how patients engage with information.

Individual Circumstances

At the individual level, patients' foundational characteristics, information literacy, and illness experiences shape how they approach health information on digital platforms. At the individual level, patients' foundational characteristics, information literacy, and illness experiences shape how they approach health information on digital platforms. Foundational characteristics, however, have been extensively examined in prior research and are therefore not elaborated on in this study.

Information literacy exercises perhaps the most pervasive moderating influence, shaping every stage of the adoption process from seeking to evaluation. A striking contrast emerges between patients with lower and higher information literacy levels. Two participants whose digital competencies differed substantially described contrasting ways of engaging with online health information:

> *My literacy level isn't good, so I usually don't read. I watch videos and listen to audio playback.* [P23]

The same participant also noted:

> *Whatever information shows up on my phone, that's what I watch.* [P23"]

By contrast, another participant described a markedly different pattern of engagement:

> *I generally don't watch videos because they contain too little information and take too much time. Instead, I read a vast number of text-based articles for medical information.* [P20]

He further explained that he prefers to search directly using specific keywords:

> *I mostly search for information by entering keywords.* [P20]

For the participant with limited literacy resources, video and audio content provide accessible entry points into health knowledge, yet they also increase reliance on algorithmically curated audiovisual feeds. In contrast, the participant with stronger digital competencies actively seeks out dense, text-based materials and engages in more extensive comparison across sources. Information literacy shapes the format of content consumed, as well as the depth of engagement, the degree of autonomy in information selection, and the evaluative strategies applied.

Illness experience constitutes a second important individual moderator. The nature and duration of a patient's illness condition shape both the urgency and the character of their information engagement:

> *I've had lumbar disc herniation for six or seven years. The daily pain severely affects my life, and now my social media feed is entirely filled with lumbar exercises. I particularly enjoy watching shared experiences from other bloggers who also have lumbar disc herniation.*
>
> *[P16]*

Another thyroid cancer participant diagnosed 5 years ago stated:

> *When first diagnosed, I scrolled through short videos every day because I was desperate to know my life expectancy. After learning that regular medication keeps my lifespan unaffected, I searched less. Now, I hardly pay attention to that information at all.* [P7]

These accounts demonstrate that illness experience acts as a dynamic moderator, whereby the nature of the condition determines the longevity and intensity of information engagement. While chronic, symptomatic conditions can lead to persistent, algorithm-saturated seeking for peer support, the resolution of survival anxiety in other cases may result in a gradual detachment from health information as the condition becomes a stabilized part of daily life.

Environmental Conditions

The environmental dimension includes emergency situations, policy changes, and social relationships, all of which influence patients' health information adoption behaviors by altering the external conditions within which adoption propensity is activated.

XSL•FO
RenderX

Policy changes exert systemic influence on the information environment itself. When government authorities introduce chronic disease management plans, promote family doctor programs, or launch public health information platforms, they restructure the channels, authoritative sources, and directions of health information supply:

> *Since the community started the family doctor program, I no longer wander around various websites. I just follow the guidance pushed through the official community health platform.* [P23]

For patients who orient their health behaviors around official guidance, policy changes can trigger wholesale shifts in the information ecosystems they inhabit—a form of top-down moderation that complements platform-level algorithmic curation.

Social relationships function simultaneously as information filters and amplifiers. The social network acts as a parallel information channel that often carries decisive trust weight:

> *Whenever I find a new exercise for my back on an app, I first consult my niece who is a doctor. Only after he confirms that the movements are safe for my specific condition do I feel confident enough to actually practice them.* [P16]

This account illustrates how interpersonal relationships can serve as a high-trust verification layer that patients consult alongside. The family member functions as a trusted information broker whose endorsement or skepticism shapes subsequent platform engagement.

Emergency situations represent a third environmental moderator that dramatically intensifies information-seeking urgency and lowers adoption thresholds. During crisis periods such as epidemic outbreaks or drug shortages, patients face compressed decision timelines and heightened anxiety:

> *It's winter now, and epidemics are more common, so I pay attention to preventive healthcare.* [P20]

The seasonal urgency expressed here reflects a contextual trigger that temporarily amplifies the salience of health information and reconfigures the criteria for resonance. During the COVID-19 pandemic, one participant, worried about not being able to buy chronic disease medication, stated:

> *I'll try what's said online, whether it's right or not.* [P28]

The emergency context dramatically lowers the critical evaluation standards that would normally govern adoption decisions, leading patients to accept information they would ordinarily verify more carefully. What this reveals about the moderating role of emergency situations is not merely heightened engagement but a qualitative shift in the epistemic standards applied.

The Mutual Shaping Process: Platform-Patient Coconstruction

Overview

The platform information environment is shaped by a platform's Information Context and Institutional Context. A patient's information behaviors, or the behavioral layer, are embedded within this environment. Through a process of continuous iteration, these factors together form a personalized ecosystem of health information content.

How Platforms Shape Patients

A patient's initial information environment on a platform is determined by underlying commercial factors, traffic allocation, algorithmic technology, platform rules, and existing users. As Suzor's [19] research suggests, the platform dictates which information is presented and how, meaning the initial environment is essentially generated at random. Once the patient enters this environment, the platform's institutional context, including its algorithms, interactive design, and rules constrains and guides their health information activities. One participant illustrated this mechanism of platform-induced habitual engagement:

> *Because this platform offers points for daily check-ins, I go on it to browse every day.* [P16]

Other patients described experiences of being manipulated by the platform's algorithms.

> *It's as if the platform knows exactly what I'm worried about. Once I click on one health tip, the feed just keeps pushing similar videos. I find myself scrolling for an hour without realizing it—I just can't stop.* [P19]

In addition, some participants noted that platform labeling policies helped them better identify promotional health content. As one participant explained:

> *Some health information is actually just promoting supplements. I couldn't tell the difference before, but since the platform added 'Advertisement' labels, I now understand the true intent behind the content.* [P17]

The gamification element embedded in the platform's design functions not merely as an incentive but as a behavioral shaping mechanism that reliably increases daily exposure to health content, producing sustained engagement without explicit instruction. Algorithmic recommendations further structure what information patients encounter once they enter the platform. For most participants, health information engagement took a largely passive form. This systemic influence extends to the interpretive process, where platform-level markers like "Advertisement" labels provide the structural cues that guide how patients identify and categorize commercial intent.

How Patients Shape Platforms

Information adoption is a bidirectional interaction between the individual and their information environment. A patient's information behaviors on a platform, in turn, actively shape the

XSL•FO
RenderX

platform's Information Context and Institutional Context, which then alter subsequent information presentation.

Regarding the Information Context, a prime example is a participant's proactive reporting of false health information.

> *Whenever I come across incorrect health information, I report it. Usually, the platform processes the report and then provides me with feedback regarding the final outcome of the review.* [P29]

When the platform receives these reports, it deletes the noncompliant or illegal content, thereby changing the Information Context.

As for the Institutional Context, platforms use data from a patient's information behaviors to generate a user profile. This profile then informs the recommendation algorithms, which suggest relevant information. One participant echoed this sentiment, noting:

> *The more I scroll, the more the platform pushes content to me. Now, everything that appears on my phone is directly related to my condition.* [P19]

This account illustrates the self-reinforcing logic of the feedback loop, where ongoing user interactions trigger recommendation cycles that progressively stabilize and strengthen a personalized health information ecosystem.

Conversely, some patients express frustration with the platform's algorithmic recommendations and consciously adopt strategies to resist this influence. One participant described a deliberate counter-strategy:

> *I found the platform's recommendations were making me too anxious; a lot of them were about failed treatments. So, I clicked the 'dislike' button or repeatedly searched for successful treatment cases, which made it reduce the push of those health-related posts.* [P7]

The patient leverages the platform's own feedback affordances (the "dislike" function) as instruments of intentional self-curation. This constitutes a form of lay algorithmic literacy—the patient has developed a practical model of how the recommendation system operates and uses that understanding to reshape their information environment.

Whether patients accept or intentionally resist the platform's recommendations, the final outcome is fed back into the platform's data analysis. This allows the platform to, on one hand, use its technology and interactive design to further present a more personalized health information content ecosystem. On the other hand, it can also adjust its technology, interactive design, and community rules based on user data, achieving a cyclical and mutually constitutive shaping process.

Discussion

Principal Findings

This study provides 3 principal insights into how patients with chronic disease adopt health information in platform-mediated environments.

First, patients with chronic disease develop sustained adoption orientations rather than making discrete adoption decisions. Because chronic disease management requires ongoing self-care over extended periods, patients repeatedly engage with health information through iterative cycles of seeking, analyzing, keeping, and adopting. These activities are driven by multiple motivations—including disease learning, information verification, and comfort seeking—and involve continuous judgments related to cognitive, emotional, and contextual considerations. What is particularly notable is that these motivations, behaviors, and attitudes form a layered structure that evolves through accumulated experience, suggesting that adoption is better understood as a developing propensity instead of a momentary intention.

Second, platform environments actively structure the conditions under which patients encounter and evaluate health information. Patients' information behaviors unfold within dual contexts: informational contexts characterized by diverse sources, varying quality, and multiple formats; and institutional contexts shaped by algorithmic recommendations, platform attributes, and governance mechanisms. These contexts shape what information patients are able to see, access, and consider credible, rather than merely serving as a background for their behaviors. This finding suggests that the platform environment functions as a structured and governed setting, not simply a neutral space.

Third, patients and platforms continuously shape each other through ongoing interactions. While platform algorithms influence information visibility, patients actively engage with, adapt to, and sometimes resist these algorithmic structures through practices such as selective engagement, content reporting, and cross-platform verification. These patient behaviors generate data that feed back into platform systems, influencing subsequent recommendations. Over time, these reciprocal dynamics produce personalized information ecosystems that reflect both platform logics and individual patient practices. This coconstruction process suggests that health information adoption cannot be understood by examining either patients or platforms in isolation.

Theoretical Integration and Contribution

Reconceptualizing Health Information Adoption as a Dynamic Process

This study contributes to the literature on health information behavior by reconceptualizing health information adoption as a dynamic and process-oriented phenomenon. For patients with chronic disease, health information adoption emerges through continuous and iterative engagement with digital information environments. Because chronic disease management involves long-term uncertainty, symptom monitoring, and treatment adjustments, patients repeatedly search for, verify, interpret, and revisit health information over time. These activities form an ongoing cycle in which information exposure, evaluation, and behavioral responses are repeatedly revisited. Adoption thus represents an evolving orientation toward health information engagement that develops through repeated interactions with digital information environments.

XSL•FO
RenderX

This process-oriented understanding extends prior qualitative research on chronic disease patients' information behaviors. Sun et al [5] examined how patients navigate between information avoidance and seeking, highlighting the emotional and cognitive tensions in their information practices. Building on this work, our findings show that patients' engagement with health information is not limited to such tensions but develops through sustained and iterative interactions with digital information environments. Over time, these interactions give rise to relatively stable orientations toward health information engagement.

To conceptualize this evolving orientation, we introduce the notion of adoption propensity. Adoption propensity refers to a developing tendency toward engaging with health information, shaped through accumulated experiences, motivational drivers, and evaluative judgments. Through repeated interactions with digital health content and ongoing disease management practices, patients form patterns that influence how new information is interpreted and whether it is incorporated into their health management. In contrast to behavioral intention in dominant adoption models [7-14], adoption propensity reflects a preintentional and dynamic orientation grounded in sustained experience rather than a discrete decision point.

By foregrounding this evolving orientation, this study shifts the analytical focus of health information adoption from discrete decision-making to longitudinal engagement processes. This perspective complements existing health information adoption research by offering a process-oriented framework grounded in patients' lived experiences that captures the dynamic nature of information engagement in contemporary digital health environments.

Integrating Platform Into Health Information Adoption

Beyond reconceptualizing adoption as a dynamic process, this study highlights the intertwined nature of platform architectures and health information practices. While prior research on health information behavior has predominantly focused on individual cognitive evaluations and personal literacy, our framework, grounded in participants' lived experiences, shows that adoption is shaped through interactions with the sociotechnical and governance structures of digital platforms.

While existing platform studies have illuminated the infrastructural power of digital platforms in shaping general information visibility [21], this research advances theoretical understanding by unpacking how platforms reconfigure patients' daily sensemaking of health information. Rather than treating platforms as neutral conduits or passive backgrounds, the emergent theoretical model captures how platforms actively participate in patients' ongoing information practices. The grounded analysis reveals that participants continually negotiate their informational journeys through iterative interactions with curated feeds and search mechanisms. By illustrating how these algorithmic encounters shape both the emergent discovery and the sustained visibility of medical content, this study extends traditional health information behavior theories. It conceptualizes the platform as an active, vital participant in the meaning-making process, thereby offering a more situated understanding of digital health adoption.

Building on this perspective, the study further contributes to theoretical development by elucidating how platform governance mechanisms intricately weave into patients' continuous evaluation of health information credibility. Previous literature has frequently approached platform governance primarily through a regulatory lens, emphasizing its institutional role in rule-setting and content moderation [19,23]. However, the present grounded analysis theoretically repositions phenomena such as account suspensions, community guidelines, and content removal as critical sensemaking resources for users. Within these digitally mediated environments, participants actively interpret and rely upon visible governance practices as essential cues for determining medical reliability. Consequently, this research enriches existing trust paradigms by demonstrating how institutional governance practices are intimately folded into everyday epistemological evaluations. This theoretical shift moves the analytical focus beyond isolated source-characteristic assessments, highlighting instead how users continually navigate and integrate platform governance structures to establish trust in their health information journeys.

A Patient–Platform Coconstruction Perspective

Building on the process-oriented understanding of health information adoption and the role of the platform, this study proposes a patient–platform coconstruction perspective to explain how health information adoption evolves in platform-mediated environments. The notion of "co-construction" draws from science and technology studies scholarship on the mutual shaping of technology and society [34], adapted here to capture the reciprocal dynamics between patients' information practices and platform systems.

Existing research on digital platforms has emphasized the constitutive power of platforms in structuring information environments and shaping user behavior [11]. However, the findings of this study suggest that health information adoption involves not only platform influence but also ongoing patient responses that actively engage with, adapt to, and sometimes resist platform structures. Participants described various strategies for navigating platform environments, such as selectively following certain information sources, reporting misleading content, avoiding particular recommendation streams, or intentionally cross-checking information across multiple platforms. These practices resonate with van Dijck et al's [35] observation that users are more than passive subjects of platform power and develop "tactics" to negotiate platform logics. Our findings extend this insight by revealing specific tactics patients with chronic disease use in health information contexts—including what some participants described as "training" the algorithm by deliberately engaging with or avoiding certain content types.

At the same time, patients' repeated interactions with health information generate behavioral data that feed into platform algorithms and influence subsequent recommendation patterns. Over time, these interactions create feedback loops in which patients' information practices shape the informational environments they encounter, while platform systems simultaneously structure the visibility and accessibility of health information. This reciprocal dynamic aligns with Gillespie's

XSL•FO
RenderX

[36] concept of "calculated publics," wherein algorithmic systems and user behaviors mutually constitute the information environments users experience. In the context of chronic disease management, where patients engage with health information over extended periods, these reciprocal dynamics become particularly pronounced, resulting in highly personalized information ecosystems that reflect both platform logics and individual patient practices.

By highlighting these ongoing interactions, the patient–platform coconstruction perspective provides a relational understanding of health information adoption in digital health environments. This perspective contributes to platform studies by providing micro-level empirical evidence of how users and platforms mutually constitute digital information ecosystems in health contexts, while also extending health information behavior research by foregrounding the role of platform structures in shaping adoption processes.

Practical Implications

The findings of this study, while derived from a specific sample of patients with chronic disease in urban China, offer several insights that may inform practice in platform-mediated health information environments.

For platform designers and operators, the finding that patients develop sustained engagement patterns with algorithmically recommended content suggests that platform design might benefit from balancing engagement metrics with health-oriented goals. Mechanisms that promote exposure to diverse authoritative sources, rather than continuously reinforcing existing preferences, may help address the information concentration patterns observed in this study. The finding that platform governance actions influence patients' credibility assessments has implications for content moderation practices. When platforms remove health-related content or suspend accounts, these actions function not only as regulatory mechanisms but also as trust signals that patients interpret when evaluating information sources. This suggests the value of transparent and consistent moderation practices, particularly for health-related content where credibility assessment is consequential. Additionally, the observed differences in how patients with varying information literacy levels navigate platform environments point to the potential value of literacy support features, such as contextual prompts or credibility indicators, that assist users in developing critical evaluation skills.

For health care providers and public health agencies, the finding that patients frequently bring platform-mediated information into clinical encounters suggests that constructive engagement with such information may support trust-building and shared decision-making in clinical practice. The stable platform use patterns observed among participants indicate that digital platforms may serve as viable channels for reaching chronic disease populations in public health communication. However, the diversity of platform preferences across participants also suggests that multi-platform strategies may be necessary to ensure broad reach, particularly in contexts where patients access health information across government-affiliated platforms, professional health sites, and general social media such as WeChat and Douyin. The intensive information seeking and emotional vulnerability described by newly diagnosed patients point to a critical transition period during which supportive resources and guided information pathways may be particularly valuable in helping patients develop sustainable information practices.

For policymakers and regulators, the coconstruction dynamics revealed in this study—wherein patients' behaviors and platform algorithms continuously shape each other over extended periods—raise questions about responsibility and accountability in platform-mediated health information environments. As patients' long-term information exposure patterns are shaped by algorithmic systems, regulatory frameworks might consider how to ensure that platform governance mechanisms support rather than undermine health information quality and accessibility. The finding that platform moderation decisions function as credibility signals in patients' evaluations suggests that regulatory oversight of content moderation practices may be warranted, particularly for health-related content where misinformation can have serious consequences. Policies that establish standards for transparency and consistency in platform governance of health information may help protect public trust while respecting platform autonomy.

For patients and patient advocacy groups, the finding that some patients actively shape their information environments through selective engagement and cross-platform verification demonstrates patient agency within algorithmically mediated spaces. Patient education programs might benefit from explicitly addressing platform literacy, including understanding algorithmic logic and intentional feature usage, to support informed navigation of digital health information environments. The role of family members and patient communities in influencing information adoption decisions points to the potential for patient communities to function not only as emotional support networks but also as collective information verification networks, where members collaboratively evaluate and share reliable health content.

Limitations and Future Work

First, this study was conducted among urban patients with chronic disease in Chengdu, China, within the context of Chinese digital platforms such as WeChat and Douyin. Differences in platform architectures, governance mechanisms, and broader digital infrastructures may limit the transferability of the findings to rural settings or to other cultural and platform environments. Future research could examine health information adoption across diverse sociotechnical contexts, including rural populations and non-Chinese platform ecosystems, and conduct cross-platform or cross-cultural comparative studies to assess the generalizability and contextual sensitivity of the proposed model.

Second, this study relied on participants' self-reported and retrospective accounts of their health information behaviors. While this approach is well-suited to exploring subjective meanings, interpretations, and sense-making processes, it may be subject to recall bias and social desirability bias. Future research could adopt longitudinal and mixed method designs, such as diary studies, in situ interviews, or digital trace analysis,

XSL•FO
RenderX

to capture real-time interactions between patients and platforms and to examine how information adoption behaviors and platform influences evolve over the illness trajectory.

Third, the theoretical model proposed in this study was developed through qualitative grounded theory and is intended for theory generation rather than hypothesis testing. As such, causal relationships among model components cannot be empirically established. Future quantitative research could operationalize key constructs identified in this study and test their relationships with health information adoption outcomes. In addition, intervention-based studies, including randomized controlled trials, could evaluate whether leveraging specific model components can improve health information adoption and patient outcomes. Relevant components include algorithmic literacy training, curated authoritative information feeds, and peer-moderated information environments.

Conclusions

This study reframes patients with chronic disease's digital health information adoption as a sociotechnical process shaped by the reciprocal interaction between patients and platform environments, rather than a purely individual and rational decision-making behavior. By highlighting the governing roles of algorithms, platform structures, and institutional mechanisms, the proposed model extends existing health information adoption theories to better capture platform-era dynamics. The findings underscore that effective patient empowerment and information quality improvement cannot rely solely on individual literacy but require ethically designed platforms, responsible algorithmic governance, and coordinated policy interventions. More broadly, as digital platforms increasingly function as sites of health governance, understanding and shaping platform–patient relationships will be critical for mitigating misinformation risks, reducing health inequities, and improving long-term health outcomes.

Acknowledgments

We are grateful to the study participants for their time and contribution to this research.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

The datasets generated and analyzed during this study are available from the first author on reasonable request.

Authors' Contributions

LH conducted the interviews, managed and analyzed the interview data, and drafted the initial manuscript. GD supervised the study and critically reviewed the manuscript. JT was responsible for conducting the interviews, organizing the interview data, and writing the Methods section of the manuscript. PH is tasked with strengthening the theoretical contributions and addressing reviewer feedback in the Discussion section.

Conflicts of Interest

None declared.

Checklist 1
SRQR checklist.
[PDF File, 117 KB - jmir_v28i1e85229_app1.pdf]

References

1. Hepburn J, Williams L, McCann L. Barriers to and facilitators of digital health technology adoption among older adults with chronic diseases: updated systematic review. JMIR Aging 2025 Sep 11;8:e80000. [doi: 10.2196/80000] [Medline: 40934502]
2. Anisha SA, Sen A, Ahmad B, Bain C. Exploring acceptance of digital health technologies for managing non-communicable diseases among older adults: a systematic scoping review. J Med Syst 2025 Mar 11;49(1):35. [doi: 10.1007/s10916-025-02166-3] [Medline: 40067482]
3. Zhou Z, Jin D, He J, et al. Digital health platform for improving the effect of the active health management of chronic diseases in the community: mixed methods exploratory study. J Med Internet Res 2024 Nov 18;26:e50959. [doi: 10.2196/50959] [Medline: 39556830]
4. Zhang X, Liu S, Deng Z, Chen X. Knowledge sharing motivations in online health communities: a comparative study of health professionals and normal users. Comput Human Behav 2017 Oct;75:797-810. [doi: 10.1016/j.chb.2017.06.028]

XSL•FO
RenderX

5. Sun H, Li J, Cheng Y, Pan X, Shen L, Hua W. Developing a framework for understanding health information behavior change from avoidance to acquisition: a grounded theory exploration. BMC Public Health 2022 Jun 4;22(1):1115. [doi: 10.1186/s12889-022-13522-0] [Medline: 35658937]
6. Sule S, DaCosta MC, DeCou E, Gilson C, Wallace K, Goff SL. Communication of COVID-19 misinformation on social media by physicians in the US. JAMA Netw Open 2023 Aug 1;6(8):e2328928. [doi: 10.1001/jamanetworkopen.2023.28928] [Medline: 37581886]
7. Fishbein M, Ajzen I. Belief, Attitude, Intention and Behaviour: An Introduction to Theory and Research: Addison-Wesley; 1975.
8. Bandura A. Social Foundations of Thought and Action: A Social Cognitive Theory: Prentice-Hall; 1986.
9. Petty RE, Cacioppo JT. Source factors and the elaboration likelihood model of persuasion. Adv Consum Res 1984;11(1):668-672.
10. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
11. Sussman SW, Siegal WS. Informational influence in organizations: an integrated approach to knowledge adoption. Inf Syst Res 2003 Mar;14(1):47-65. [doi: 10.1287/isre.14.1.47.14767]
12. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view1. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]
13. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
14. Becker MH, Maiman LA. Sociobehavioral determinants of compliance with health and medical care recommendations. Med Care 1975 Jan;13(1):10-24. [doi: 10.1097/00005650-197501000-00002] [Medline: 1089182]
15. Gorwa R. What is platform governance? Inf Commun Soc 2019 May 12;22(6):854-871. [doi: 10.1080/1369118X.2019.1573914]
16. Scharlach R, Hallinan B, Shifman L. Governing principles: articulating values in social media platform policies. New Media Soc 2024 Nov;26(11):6658-6677. [doi: 10.1177/14614448231156580]
17. Chavalarias D, Bouchaud P, Panahi M. Can few lines of code change society? beyond fact-checking and moderation: how recommender systems toxifies social networking sites. arXiv. Preprint posted online on Mar 27, 2023. [doi: 10.48550/arXiv.2303.15035]
18. Stray J, Halevy A, Assar P, et al. Building human values into recommender systems: an interdisciplinary synthesis. ACM Trans Recomm Syst 2024 Sep 30;2(3):1-57. [doi: 10.1145/3632297]
19. Suzor N. Digital constitutionalism: using the rule of law to evaluate the legitimacy of governance by platforms. Soc Media Soc 2018 Jul;4(3):2056305118787812. [doi: 10.1177/2056305118787812]
20. Nahon K. Where there is social media there is politics. In: Bruns A, editor. The Routledge Companion to Social Media and Politics: Routledge; 2015.
21. Nieborg DB. The impact of digital platforms on canadian media production. Knowledge synthesis report.: Social Sciences and Humanities Research Council of Canada; 2021. URL: http://hdl.handle.net/1807/107407 [accessed 2026-01-25]
22. Sanner TA, Kempton AM, Russpatrick S, Sæbø JI. Governing digital platform ecosystems for social options. Inform Syst J 2025 Mar;35(2):422-449. [doi: 10.1111/isj.12546]
23. Helberger N. The political power of platforms: how current attempts to regulate misinformation amplify opinion power. Digit J 2020 Jul 2;8(6):842-854. [doi: 10.1080/21670811.2020.1773888]
24. Pathak R, Spezzano F, Pera MS. Understanding the contribution of recommendation algorithms on misinformation recommendation and misinformation dissemination on social networks. ACM Trans Web 2023 Nov 30;17(4):1-26. [doi: 10.1145/3616088]
25. Peng L, Wang J. Algorithm as recommending source and persuasive health communication: effects of source cues, language intensity, and perceived issue involvement. Health Commun 2024 Apr;39(4):852-861. [doi: 10.1080/10410236.2023.2242087] [Medline: 37525527]
26. Noncommunicable diseases. World Health Organization. 2025. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases [accessed 2026-01-25]
27. Tesfa GA, Demeke AD, Zewold M, Ngusie HS. Health information-seeking behavior among people living with the two common chronic diseases in low and middle-income countries (LMICs). a systematic review and meta-analysis. Digit Health 2024;10:20552076241302241. [doi: 10.1177/20552076241302241] [Medline: 39600387]
28. Tracy SJ. Qualitative Research Methods: Collecting Evidence, Crafting Analysis, Communicating Impact, 2nd edition: Wiley-Blackwell; 2020.
29. Khan SN. Qualitative research method: grounded theory. Int J Bus Manag 2014;9(11):224-233. [doi: 10.5539/ijbm.v9n11p224]
30. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
31. Francis JJ, Johnston M, Robertson C, et al. What is an adequate sample size? operationalising data saturation for theory-based interview studies. Psychol Health 2010 Dec;25(10):1229-1245. [doi: 10.1080/08870440903194015] [Medline: 20204937]

XSL•FO
RenderX

32. Sturges JE, Hanrahan KJ. Comparing telephone and face-to-face qualitative interviewing: a research note. Qual Res 2004 Apr;4(1):107-118. [doi: 10.1177/1468794104041110]
33. Strauss A, Corbin J. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory, 2nd edition: Sage Publications; 1998.
34. Bijker WE. Of Bicycles, Bakelites, and Bulbs: Toward a Theory of Sociotechnical Change: MIT Press; 1995.
35. Dijck J, Poell T, Waal M. The Platform Society: Public Values in a Connective World: Oxford University Press; 2018.
36. Gillespie T. The relevance of algorithms. In: Media Technologies: Essays on Communication, Materiality, and Society: MIT Press; 2014:167-194. [doi: 10.7551/mitpress/9780262525374.003.0009]

Abbreviations

SRQR: Standards for Reporting Qualitative Research

Edited by A Stone; submitted 04.Oct.2025; peer-reviewed by H Sallehudin, L Liu; revised version received 12.Apr.2026; accepted 13.Apr.2026; published 20.May.2026.

Please cite as:
Huang L, Duan G, Tang J, He P
Health Information Adoption Among Patients With Chronic Disease in China: Qualitative Interview Study of Patient–Platform Coshaping
J Med Internet Res 2026;28:e85229
URL: https://www.jmir.org/2026/1/e85229
doi:10.2196/85229

©Lin Huang, Guimin Duan, Jiani Tang, Peipei He. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

When Old Diseases Return: Cholera, Crisis, and Digital Surveillance in Fragile Settings

Isabelle Basbouss-Serhal

(J Med Internet Res 2026;28:e94818) doi:10.2196/94818

KEYWORDS

cholera; public health surveillance; disease outbreaks; water microbiology; genomics; climate change; epidemiology; vaccines; health policy; population surveillance; war; internally displaced persons

Textbox .

Key Takeaways

- Cholera's re-emergence reflects systemic collapse rather than pathogen novelty. Economic crisis, climate stress, displacement, and infrastructure failure can rapidly reverse infectious disease control.
- Delayed detection has been a core problem in resource-limited situations, due to lack of laboratory capacity and incomplete reporting.
- Digital surveillance plays an important role in early warning and response. Mobile reporting, GISgeographic information systems, and real-time data integration can be used when traditional surveillance fails.
- Integration of digital and genomic information can enhance preparedness and outbreak prevention.

For almost three decades, Lebanon was free of cholera. Its re-emergence in October 2022—the first outbreak in Lebanon since 1993 [1]—marked not simply the return of an old pathogen but the convergence of modern crises: economic collapse, infrastructure failure, population displacement, and environmental stress [2]. Comparable patterns are now playing out across multiple regions globally, where diseases once thought controlled are resurging under these new conditions.

Cholera, an acute waterborne diarrheal disease transmitted primarily through contaminated water and food, remains a preventable disease. Yet the conditions required to prevent it—safe water, sanitation, early detection, and rapid response—are increasingly undermined by conflict and climate-related shocks. In fact, the number of cholera outbreaks and countries affected has increased sharply in recent years, especially in contexts affected by humanitarian emergencies, according to the World Health Organization [3,4].

Flooding, drought, and rising temperatures intensify exposure risks, while health systems weakened by those very influences often fail to recognize outbreaks in time to limit their spread [5,6]. For regions where resources and infrastructure may be pre-existing challenges, digital disease surveillance tools offer not only speed but a decentralized, low-infrastructure means of outbreak detection and mitigation [7]. For example, these tools have been used to monitor cholera outbreaks in regions with minimal laboratory capacity, demonstrating that real-time reporting can remain effective even when traditional health systems are under strain [7].

When Traditional Surveillance Systems Fail

Under stable conditions, cholera outbreaks can be controlled through reporting, confirmation, and response. Under fragile conditions—shortages in health facilities, limited laboratory capacity, and unorganized reporting—this process breaks down [8]. The resulting delayed detection can allow cholera to spread extensively through the population [9].

This was evident in the 2022 outbreak in Lebanon, where traditional surveillance failure contributed to the spread of cholera nationwide in a matter of weeks. The initial cholera outbreaks were identified in the northern governorates of Akkar and Minieh-Danniyeh, near the Syrian border, in areas with ongoing cross-border movements and continued transmission in neighboring Syria [1]. Despite these risk factors, the outbreaks went undetected for several weeks, highlighting gaps in early surveillance. Subsequent investigations revealed *Vibrio cholerae* contamination in water piping systems, irrigation water, and sewage networks, reflecting the long-term degradation of water infrastructure amid the country's ongoing economic crisis [10].

The Lebanese situation—in which climate and other pressures strain already-taxed or limited public health infrastructure, rendering them more vulnerable to outbreaks—is not an isolated one. For example, in Malawi, floods have caused the displacement of people and the contamination of water sources, and this has led to new outbreaks, despite previous success in controlling the disease [8]. In Yemen, in 2017, the world's largest cholera epidemic was facilitated by inadequate water supplies due to conflict and an overwhelmed health care system [2,5]. In Sudan, similar problems have contributed to the spread of the disease and are ongoing [8]. Cholera outbreaks no longer

XSL•FO
RenderX

follow predictable seasonal periods, but instead track crises and displacement [6].

In all these cases, digital surveillance platforms, such as mobile case reporting and geographic information system (GIS) mapping, could have mitigated delays, enabling earlier intervention and more effective and efficient use of limited resources.

Digital Advances in Surveillance and Prediction

The challenges in resource-limited environments associated with traditional surveillance methods (often, for example, facility based and involving the use of paper-based forms and delayed aggregation) are increasingly being complemented by the use of digital health technologies. Reporting systems developed using mobile technologies enable the reporting of suspected cases in nearly real time by health workers on the front lines of the response. Particularly in the context of displacement, where populations are highly mobile and not all are registered, these technologies represent viable alternative approaches [11]. However, in fragile settings, internet and cellular network disruptions, power instability, and limited access to reliable technology may hinder digital surveillance systems, making hybrid approaches desirable.

GIS mapping can further help detect outbreaks by tracking distributions of cases and environmental factors simultaneously. Rainfall, floods, temperature changes, and population movement can all be overlaid to help pinpoint potentially high-risk regions before large numbers of cases are seen [5]. In recent cases of cholera, these methods have facilitated vaccination and water and sanitation efforts, as well as the pre-positioning of oral rehydration kits [2]. Development of other predictive models combining data on epidemiology, climate, and mobility also helps health systems move away from reactive responses to predictive preparedness, particularly helpful in areas where testing is slow or inconclusive [5,12,13].

In tandem with syndromic and environmental surveillance, advances in genomic sequencing have added a new layer of early warning for cholera outbreaks. Through whole-genome sequencing of *Vibrio cholerae*, circulating strains can be identified, differentiation can be made between endemic persistence and new introductions, and genetic markers of virulence or antimicrobial resistance can be detected [14]. When integrated into digital surveillance platforms, genomic data can support real-time mapping of transmission routes across regions and borders, linking molecular signatures to epidemiologic, climatic, and mobility data, as demonstrated during the COVID-19 pandemic through platforms such as GISAID and Nextstrain [15,16]. It's particularly useful in fragile settings where laboratory confirmation is delayed and traditional surveillance is fragmented, allowing molecular signals to complement clinical reporting and enhance outbreak prediction and response.

Implications for Policy and Preparedness

Cholera's flare-ups offer several lessons for managing recurring infections in an age of compounding crises.

First, advances in infectious disease control can be swiftly reversed when social, political, and environmental stability breaks down. Second, surveillance is as critical as cure: interventions delayed by gaps in data may be ineffective, no matter their inherent utility. Third, health technology platforms must be regarded as essential infrastructure rather than optional tools, particularly in fragile or conflict-affected contexts [3]. For policymakers, this means investing in interoperable digital surveillance systems as a form of preparedness rather than a reactive measure. For clinicians, it requires maintaining a high index of suspicion for cholera in nontraditional settings. For researchers, these outbreaks underscore the value of an interdisciplinary approach that integrates epidemiology, climate data, and digital surveillance and predictive modeling tools [9].

As climate change, displacement, and conflict continue to redefine risk in global health, cholera is not likely to be the last "old" disease to return. Our ability to preempt outbreaks and resurgences of such diseases will depend largely on our ability to develop, integrate, and rapidly operationalize digital surveillance data and predictive modeling tools [3].

Conflicts of Interest

None declared.

References

1. Cholera – disease outbreak news: lebanon, October 2022. World Health Organization. 2022 Oct 19. URL: https://www.who.int/emergencies/disease-outbreak-news/item/2022-DON416 [accessed 2026-03-16]
2. Bayram Z, Bizri ARN, Musharrafieh U. Resurgence of cholera in Lebanon. East Mediterr Health J 2023 Nov 30;26(11):837-838. [doi: 10.26719/emhj.23.111] [Medline: 38279878]
3. Cholera upsurge (2011-present). World Health Organization. 2025. URL: https://www.who.int/emergencies/situations/cholera-upsurge [accessed 2026-03-16]
4. Cholera – Multi-country with a focus on countries experiencing current surges. World Health Organization. URL: https://www.who.int/emergencies/disease-outbreak-news/item/2025-DON579 [accessed 2026-03-17]
5. Lessler J, Moore SM, Luquero FJ, et al. Mapping the burden of cholera in sub-Saharan Africa and implications for control: an analysis of data across geographical scales. Lancet 2018 May 12;391(10133):1908-1915. [doi: 10.1016/S0140-6736(17)33050-7] [Medline: 29502905]

XSL•FO
RenderX

6. Mohamed MG, Dabou EAA, Abdelsamad S, Elsalous SH. Cholera outbreaks: public health implications, economic burden, and preventive strategies. AIMS Public Health 2025;12(3):767-795. [doi: 10.3934/publichealth.2025039] [Medline: 41127425]
7. Basbouss-Serhal I. Tracking superbugs in the digital era. J Med Internet Res 2025 Dec 4;27:e88835. [doi: 10.2196/88835] [Medline: 41343868]
8. Izzoddeen A, Abualgasim H, Abasher M, et al. Cholera in conflict: outbreak analysis and response lessons from Gadaref state, Sudan (2023-2024). BMC Public Health 2025 Mar 5;25(1):881. [doi: 10.1186/s12889-025-22128-1] [Medline: 40045292]
9. Balkhy HH. Mitigating the increasing threat of cholera in Yemen and other conflict-affected countries in the Eastern Mediterranean Region. East Mediterr Health J 2025 Aug 3;31(7):423-425. [doi: 10.26719/2025.31.7.423] [Medline: 40832862]
10. Ending cholera: a global roadmap to 2030. World Health Organization. 2017. URL: https://stacks.cdc.gov/view/cdc/58349 [accessed 2026-03-16]
11. Dadi HB, Negash DT, Adall SW. Integrating surveillance and climate data for cholera early warning in Ethiopia. Ann Glob Health 2025;91(1):62. [doi: 10.5334/aogh.4742] [Medline: 40979687]
12. Mandeng JM. Modeling cholera dynamics under food insecurity and environmental contamination: a multi-patch theoretical approach. arXiv. Preprint posted online on Sep 6, 2025 URL: https://arxiv.org/abs/2509.05731 [accessed 2026-03-16]
13. Lee JS, Tyler ARB, Veinot TC, Yakel E. Now is the time to strengthen government-academic data infrastructures to jump-start future public health crisis response. JMIR Public Health Surveill 2022;10:e51880. [doi: 10.2196/51880]
14. Weill FX, Domman D, Njamkepo E, et al. Genomic history of the seventh pandemic of cholera in Africa. Science 2017 Nov 10;358(6364):785-789. [doi: 10.1126/science.aad5901] [Medline: 29123067]
15. Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet 2018 Jan;19(1):9-20. [doi: 10.1038/nrg.2017.88] [Medline: 29129921]
16. Houwaart T, Belhaj S, Tawalbeh E, et al. Integrated genomic surveillance enables tracing of person-to-person SARS-CoV-2 transmission chains during community transmission and reveals extensive onward transmission of travel-imported infections, Germany, June to July 2021. Euro Surveill 2022 Oct;27(43):2101089. [doi: 10.2807/1560-7917.ES.2022.27.43.2101089] [Medline: 36305336]

Edited by KA Clegg; submitted 06.Mar.2026; this is a non–peer-reviewed article;accepted 06.Mar.2026; published 17.Mar.2026.

Please cite as:
Basbouss-Serhal I
When Old Diseases Return: Cholera, Crisis, and Digital Surveillance in Fragile Settings
J Med Internet Res 2026;28:e94818
URL: https://www.jmir.org/2026/1/e94818
doi:10.2196/94818

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Mar.2026.

XSL•FO
RenderX

Data Governance Lessons From an Unvalidated Dataset

Cliff Dominy

(J Med Internet Res 2026;28:e94518) doi:10.2196/94518

KEYWORDS

data integrity; data quality; data management; data provenance; data sharing; research integrity; scientific misconduct; retraction of publication; research ethics; artificial intelligence

Textbox 1.

Key Takeaways

- An open access dataset has highlighted how bad data can propagate through the research ecosystem.
- When trained on unvalidated datasets, machine learning can amplify misinformation, erode trust in science, and harm vulnerable populations.
- Enforced data provenance systems could play a key role in preventing bad data from corrupting the scientific record.

When an unvalidated dataset recently made it into the medical literature, it exposed several weaknesses in data governance. The dataset was uploaded to Kaggle—a large online platform where users can share publicly accessible data, code, and models—and was fundamentally flawed [1]. Its developer had compiled unverified images of children from websites related to autism to train an artificial intelligence (AI) model to "detect the presence of autism or the absence thereof" from the scraped images [2].

A sharp-eyed reviewer exposed the problem only at the publication stage; by December 2025, it was estimated that over 90 published papers had incorporated the bad data, leading to investigations and double-digit retractions [1,3].

These kinds of data integrity and governance failures are particularly consequential because of how early they occur in the research life cycle, and with open access datasets fueling large-scale machine learning and other AI research, analyses can be generated and published at unprecedented speed and scale, allowing data issues to propagate more rapidly throughout the research ecosystem. Far from an isolated incident [4-8], this situation highlights the need for more robust and proactive data governance solutions.

The Impact of Bad Data

Anne Borden is an autism advocate, journalist, and author of the upcoming book *The Informed Parent*—a decision-making guide for parents of autistic children. For Borden, the priority here is to learn from this "bizarre story" and fix the system without delay. "You really have to stop misinformation being perpetuated under the banner of science," she says, "because once it's out there, you're done. The Internet is forever."

Bad Data, Bad Science: Who Should Fix This?

Who are the custodians of good data during their migration from a spreadsheet to the scientific record? What role should each stakeholder play in maintaining data integrity? While responsibility for data governance is distributed across many actors (including researchers and regulators), data-sharing platforms, research and funding institutions, and academic publishers help determine how data are shared, vetted, and ultimately incorporated into the scientific record.

The Data-Sharing Platforms

Open access databases and data repositories, like Kaggle and GitHub, are popular resources that software developers and data scientists use to train their machine learning algorithms for free. Software development benefits from these repositories, yet the datasets they host often lack the documentation, governance, and quality practices required for careful medical research or clinical algorithm development [9].

Alan Katz, MBChB, MSc, CCFP, is a professor of family medicine and community health sciences and a senior scientist at the Manitoba Centre for Health Policy (MCHP). Katz found the dataset revelations "both shocking, but also not surprising" due to the rapid expansion of open access databases and their widespread use in machine learning and AI research. The Kaggle-style data-sharing platforms differ sharply from established medical databases, such as those maintained by the MCHP, which employs full-time staff tasked with validating all new data before uploading them. Katz says, "We take our ethical standards as seriously as clinical trials do."

Elizabeth Green, DPhil, is a lecturer in business and law at the University of the West of England, Bristol. Her research focuses on data integrity, and while she has seen cases like this before, she doesn't believe locking data away is necessarily the solution [10]. For example, DermAtlas—an open-source medical database of skin conditions—is a "fantastic resource," she says, and "extremely helpful, especially in [diagnosing] some extremely rare cases." To balance the risks and benefits of open data, the focus should instead be on building better governance systems.

XSL•FO
RenderX

The Institutions

Other stakeholders in the data transformation journey are the institutions that conduct primary medical research and the public agencies that fund that research. Is it time to adopt and enforce international data integrity and ethics standards at all research institutions, or would this be an affront to academic freedom?

Funding bodies have traditionally taken a dim view of researchers that waste public funds on bogus science, which impacts their future grants. Indeed, in many but not all regions of the world, funding is contingent on maintaining ethical research standards. In Canada, Katz says, "our existence is 100% dependent on having those strict ethical guidelines."

The Journals

The research integrity pipeline involves several stakeholders, with each having distinct roles in maintaining the standards of academic research. Gatekeepers in the system—one of the last lines of defense—are the academic journals. Journals have a vested interest in maintaining high academic standards and may be well placed to dictate the terms of engagement.

Felix Ritchie, PhD—a colleague of Elizabeth Green—developed the Five Safes data integrity framework for just this purpose [11]. Ritchie describes it as "a flexible structure for thinking about [data]," which includes the provenance and ethics of data use. Numerous organizations worldwide have adopted the Five Safes framework to date, and Australia has recently legislated it [12].

Viewed through an ethical lens, the Five Safes could form the backbone of a data provenance system that requires compliance before a manuscript can be considered for publication.

Data Provenance: The Five Safes in Action

Ritchie's Five Safes framework allows for effective data validation and, when combined with modern ethical standards, can restore trust by filtering data sources through five discrete tests:

1. Safe Project: Data should be ethically collected and clinically validated by experts.
2. Safe People: Researchers accessing the data must be qualified and specifically trained in using AI-based datasets.
3. Safe Data: Data should be independently validated, and any accesses or modifications should be tracked.
4. Safe Settings: Were health data acquired in a clinical setting and the data securely stored?
5. Safe Outputs: Were valid methodologies and statistics used to derive the results?

Restoring Data Integrity

How can one implement a data provenance system?

Ritchie feels that applying the Five Safes framework to an ethical dataset is the way forward. "There is a need for a register of validated, ethical datasets," he says," that would really be a game changer."

A possible workflow could include the following:

1. Data are collected by medical experts and validated by a third-party certification service.
2. The data are stored in an accredited data registry and protected by blockchain cybersecurity—the same technology that safeguards financial transactions.
3. Researchers access these datasets and use them for approved research purposes.
4. A submitted manuscript would need ethical approval and a data security certificate before verification by a journal's research integrity team.

Ritchie sums it up nicely: "Unless you use a validated data set, you're not getting published, mate." That's a powerful incentive.

Opportunity for Self-Reflection and Correction

Machine learning and other AI technologies have the capacity to transform medical research in ways we are only beginning to understand. However, human frailties, such as blind trust in open access data and lack of institutional ethical oversight within our publish-or-perish culture, have shown how quickly such technologies can amplify misinformation.

While the impact of this situation was ultimately contained, it is nevertheless an important opportunity for self-reflection among all in the research ecosystem. It's a chance and, perhaps, a responsibility to fix the flaws and prevent history from repeating itself.

Conflicts of Interest

None declared.

References

1. McMurray C. Exclusive: Springer Nature retracts, removes nearly 40 publications that trained neural networks on 'bonkers' dataset. The Transmitter. 2025 Dec 8. URL: https://www.thetransmitter.org/retraction/exclusive-springer-nature-retracts-removes-nearly-40-publications-that-trained-neural-networks-on-bonkers-dataset/ [accessed 2026-03-09]
2. Info.txt. Google Drive. URL: https://drive.google.com/file/d/1zMQgyQvYiYyxx9J5jw3jrLGTS0p19Rep/view [accessed 2026-03-09]
3. Expression of concern: data mining-based model for computer-aided diagnosis of autism and gelotophobia: mixed methods deep learning approach. JMIR Form Res 2026 Jan 23;10:e91833. [doi: 10.2196/91833] [Medline: 41576282]

XSL•FO
RenderX

4. Toraih EA, ElWazir M, Elshazli RM, Hussein MH, Fawzy MS, Elroukh SM. Rapid publication during crises: analyzing retractions during the Covid-19 pandemic. Ethics Med Public Health 2025;33:101136. [doi: 10.1016/j.jemep.2025.101136]
5. León FR. RETRACTED: likely electromagnetic foundations of gender inequality. Cross Cult Res 2023 Apr;57(2-3):239-263. [doi: 10.1177/10693971221143577]
6. Lancet, NEJM retract controversial COVID-19 studies based on Surgisphere data. Retraction Watch. 2020 Jun 4. URL: https://retractionwatch.com/2020/06/04/lancet-retracts-controversial-hydroxychloroquine-study/ [accessed 2026-03-09]
7. Okyay RA, Kocyigit BF, Qumar AB, Yessirkepov M, Sumbul HE. Fifty years of retracted medical publications from 1975 to 2024: a comprehensive analysis of trends, reasons, and countries using the Retraction Watch database. J Korean Med Sci 2025 Dec 1;40(46):e300. [doi: 10.3346/jkms.2025.40.e300] [Medline: 41327922]
8. Peng K, Mathur A, Narayanan A. Mitigating dataset harms requires stewardship: lessons from 1000 papers. arXiv. Preprint posted online on Aug 6, 2021. [doi: 10.48550/arXiv.2108.02922]
9. Avlona NR, Cheplygina V, Jiménez-Sánchez A, et al. Copycats: the many lives of a publicly available medical imaging dataset. Presented at: Advances in Neural Information Processing Systems 37; Dec 10-15, 2024; Vancouver, BC, Canada p. 113383-113404. [doi: 10.52202/079017-3602]
10. ripleywk. Case study: Dr Elizabeth Green -trust as foundation: enabling safe data access for public good. UWE Bristol blogs. 2025 Feb 28. URL: https://blogs.uwe.ac.uk/research-external-engagement/case-study-dr-elizabeth-green-trust-as-foundation-enabling-safe-data-access-for-public-good/ [accessed 2026-03-09]
11. Desai T, Ritchie F, Welpton R. Five Safes: designing data access for research. : Department of Accounting, Economics and Finance, Bristol Business School, University of the West of England, Bristol; 2016 URL: https://ideas.repec.org/p/uwe/wpaper/20161601.html [accessed 2026-03-09]
12. Data Availability and Transparency Code 2022. Australian Government Office of the National Data Commissioner. 2022. URL: https://www.datacommissioner.gov.au/support/resources/data-availability-and-transparency-code-2022 [accessed 2026-03-09]

Edited by KA Clegg; submitted 02.Mar.2026; this is a non–peer-reviewed article;accepted 02.Mar.2026; published 12.Mar.2026.

Please cite as:
Dominy C
Data Governance Lessons From an Unvalidated Dataset
J Med Internet Res 2026;28:e94518
URL: https://www.jmir.org/2026/1/e94518
doi:10.2196/94518

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Mar.2026.

XSL•FO
RenderX

The Right to Understand in Health Care AI

Anshu Ankolekar

(J Med Internet Res 2026;28:e95090) doi:10.2196/95090

KEYWORDS

artificial intelligence; government regulation; European Union; clinical decision support systems; informed consent; ethics; patient participation; patient rights; health knowledge, attitudes, practice; health literacy

Textbox .

Key Takeaways

- The European Union Artificial Intelligence (AI) Act and General Data Protection Regulation (GDPR) give patients legal grounds to seek explanations of AI-driven medical recommendations, but neither framework specifies what a meaningful explanation requires in clinical practice.
- Technical, clinical, and literacy barriers mean that even well-intentioned explanations may not reach patients in a form they can use to make informed decisions about their care.
- Closing this gap requires shifting from compliance ("Was an explanation provided?") to effectiveness ("Can patients use it?").

An artificial intelligence (AI) system flags a nodule on a lung computed tomography scan and assigns it an 87% malignancy probability. The radiologist receives a confidence score and a heat map showing which parts of the image the algorithm focused on. When her patient asks, "why does the computer think it's cancer," she realizes the output tells her what the AI concluded but not why or how to explain it to her patient.

It can be incredibly challenging to explain AI reasoning in clinical practice. The European Union (EU) AI Act, which entered into force in August 2024 with obligations phasing in through 2027, requires that deployers of high-risk AI systems provide those affected with clear and meaningful explanations of decisions shaped by these systems [1]. This creates a legal basis for patients to seek explanations of AI-driven medical recommendations. It also immediately raises a question the law alone cannot answer: What does a meaningful explanation of an AI medical decision look like in clinical practice, and how do we deliver it?

Legal Foundation and Open Questions

Many clinically deployed medical AI systems will fall into the AI Act's high-risk category, particularly where they are regulated as medical device software or serve as safety components of regulated products [1,2]. For cases where the AI Act does not directly apply, patients may also draw on the General Data Protection Regulation (GDPR). Under Article 22, the GDPR provides safeguards against decisions "based solely on automated processing," including the right to meaningful information about the logic behind those decisions [3].

However, most clinical AI does not meet Article 22's threshold, since a human clinician typically remains the formal decision maker. This creates ambiguity and a paradox: the human oversight meant to protect patients may also reduce their legal claim to explanation, since the decision may no longer be considered purely automated.

Together, these frameworks establish that patients have legal grounds to seek explanations of AI-driven medical recommendations, but these are stronger in principle than in practice, and neither framework specifies what a "meaningful" explanation requires.

Technical Complexity

The difficulty the radiologist in the opening scenario faces is not incidental. The most accurate AI models generate outputs through millions of interacting parameters in ways that even their developers cannot fully trace [4]. Requiring greater transparency can push developers toward simpler, more interpretable models that sacrifice diagnostic accuracy, a trade-off with direct consequences for patient care [5,6].

Current explainable AI methods only partially address this. Saliency maps show which regions of an image the algorithm weighted most heavily, but not what it understood about them [7]. Feature importance rankings indicate which variables mattered, but not how they interacted or why particular thresholds were significant [8]. These post hoc approximations attempt to reconstruct reasoning after the fact [9,10], and they can produce plausible-sounding explanations that do not accurately reflect the model's internal logic [11].

Clinical Implementation Challenges

Even where technical explanations exist, delivering them in clinical practice has its own barriers. Clinicians typically receive AI outputs as confidence scores and recommendations rather than reasoning, meaning they may be asked to explain a decision they do not fully understand themselves [12,13]. Clinicians already struggle to find time for thorough clinical conversations, and AI adds another layer of complexity to encounters that are already stretched [14-16].

Automation bias—deferring to algorithmic recommendations even when these conflict with clinical judgment—is another

XSL•FO
RenderX

implementation challenge clinicians face [17]. A prospective study of radiologists reading mammograms found that incorrect AI suggestions pulled readers toward an incorrect diagnosis regardless of level of experience [18]. An explanation delivered by a clinician who has already deferred to the algorithm may reflect the AI's conclusion rather than an independent clinical assessment, challenging the assumption in both legal frameworks that human oversight guarantees meaningful review.

Patient Understanding

Even if clinicians provide explanations, understanding is not guaranteed. Between 22% and 58% of EU citizens report difficulty accessing, understanding, appraising, and applying the health information they need to navigate health care services, with pronounced gaps among older adults, lower socioeconomic groups, and rural communities [19]. Interpreting AI outputs requires statistical and technical literacy that even a high general education does not guarantee. Many highly educated individuals struggle with medical statistics and probability statements, meaning technically accurate explanations may create barriers regardless of educational background [20].

Providing more technical detail is not likely to help. Research on medical decision-making suggests that excessive technical information can lead to cognitive overload, causing patients to defer to physician authority rather than engage with the explanation [21]. To participate meaningfully in decisions, what patients typically need is not a description of how an algorithm works, but clarity on what's most relevant for their own situation [22,23].

What Implementation Requires

Resolving these challenges requires multiple actors.

Developers could design explanation systems with patient input from the outset, testing comprehension with actual patients rather than demonstrating compliance with legal standards alone [6]. Drawing on principles from shared decision-making [24], risk communication [25], and algorithmic fairness research [26], a useful patient-facing explanation could include what the system is recommending and for what decision point; how confident it is and what that confidence means in practical terms; relevant key limitations, such as known performance gaps in specific populations; and viable alternative options. This reframes the goal from technical transparency to decision-relevant clarity, with effectiveness measurable by whether patients can answer these questions after an encounter.

Health care institutions can also bridge this gap. Allocating time for AI discussions, training staff to support patients in navigating AI-driven recommendations, and establishing clear protocols could shift explanation from a compliance exercise toward genuine patient understanding. Policy makers could support this by developing standards focused on comprehension and investing in digital health literacy programs [2].

Involving patients in the design of explanation systems from the start would strengthen all of these efforts. Patient advocates have highlighted that explanation approaches tend to reflect what developers and regulators consider important, which may not always align with what patients need to know [27]. Co-design partnerships between developers, institutions, and patient communities offer a route toward explanations that are not only legally sound but genuinely useful.

Moving Forward

The EU AI Act gives patients something they did not previously have: a legal basis for demanding transparency about AI systems influencing their care. But the right to explanation and the capacity to deliver one that patients can genuinely use are shaped by forces the law alone cannot govern: the opacity of high-performing models, the pressures of clinical practice, and the diversity of patient needs and literacy levels.

What the AI Act's transparency requirements provide, beyond legal protection, is a shared standard to work toward. The right to explanation is an important starting point. What patients need now are answers they can use.

Conflicts of Interest

None declared.

References

1. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). EUR-Lex. 2024. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689 [accessed 2026-03-17]
2. van Kolfschooten H, van Oirschot J. The EU Artificial Intelligence Act (2024): implications for healthcare. Health Policy 2024 Nov;149:105152. [doi: 10.1016/j.healthpol.2024.105152] [Medline: 39244818]
3. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). EUR-Lex. 2016. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679 [accessed 2026-03-17]
4. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019 May;1(5):206-215. [doi: 10.1038/s42256-019-0048-x] [Medline: 35603010]

XSL•FO
RenderX

5. Ennab M, Mcheick H. Designing an interpretability-based model to explain the artificial intelligence algorithms in healthcare. Diagnostics (Basel) 2022 Jun 26;12(7):1557. [doi: 10.3390/diagnostics12071557] [Medline: 35885463]
6. Ebers M. AI robotics in healthcare between the EU Medical Device Regulation and the Artificial Intelligence Act: gaps and inconsistencies in the protection of patients and care recipients. Oslo Law Rev 2024 Oct 31;11(1):1-12. [doi: 10.18261/olr.11.1.2]
7. Borji A. Saliency prediction in the deep learning era: successes and limitations. IEEE Trans Pattern Anal Mach Intell 2021 Feb;43(2):679-700. [doi: 10.1109/TPAMI.2019.2935715] [Medline: 31425064]
8. Hossain MI, Zamzmi G, Mouton PR, Salekin MS, Sun Y, Goldgof D. Explainable AI for medical data: current methods, limitations, and future directions. ACM Comput Surv 2025 Jun 30;57(6):1-46. [doi: 10.1145/3637487]
9. Bordt S, Finck M, Raidl E, von Luxburg U. Post-hoc explanations fail to achieve their purpose in adversarial contexts. Presented at: FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency; Jun 21-24, 2022; Seoul, Republic of Korea p. 891-905. [doi: 10.1145/3531146.3533153]
10. Mhasawade V, Rahman S, Haskell-Craig Z, Chunara R. Understanding disparities in post hoc machine learning explanation. Presented at: FAccT '24: The 2024 ACM Conference on Fairness, Accountability, and Transparency; Jun 3-6, 2024; Rio de Janeiro, Brazil p. 2374-2388. [doi: 10.1145/3630106.3659043]
11. Jin Q, Chen F, Zhou Y, et al. Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. NPJ Digit Med 2024 Jul 23;7(1):190. [doi: 10.1038/s41746-024-01185-7] [Medline: 39043988]
12. Sivaraman V, Bukowski LA, Levin J, Kahn JM, Perer A. Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. Presented at: CHI '23: CHI Conference on Human Factors in Computing Systems; Apr 23-28, 2023; Hamburg, Germany p. 1-18. [doi: 10.1145/3544548.3581075]
13. Scott IA, Carter SM, Coiera E. Exploring stakeholder attitudes towards AI in clinical practice. BMJ Health Care Inform 2021 Dec;28(1):e100450. [doi: 10.1136/bmjhci-2021-100450] [Medline: 34887331]
14. Braddock CH, Snyder L. The doctor will see you shortly. The ethical significance of time for the patient-physician relationship. J Gen Intern Med 2005 Nov;20(11):1057-1062. [doi: 10.1111/j.1525-1497.2005.00217.x] [Medline: 16307634]
15. Jacobs M, He J, Pradier MF, et al. Designing AI for trust and collaboration in time-constrained medical decisions: a sociotechnical lens. Presented at: CHI '21: CHI Conference on Human Factors in Computing Systems; May 8-13, 2021; Yokohama, Japan p. 1-14. [doi: 10.1145/3411764.3445385]
16. Covvey JR, Kamal KM, Gorse EE, et al. Barriers and facilitators to shared decision-making in oncology: a systematic review of the literature. Support Care Cancer 2019 May;27(5):1613-1637. [doi: 10.1007/s00520-019-04675-7] [Medline: 30737578]
17. Abdelwanis M, Alarafati HK, Tammam MMS, Simsekler MCE. Exploring the risks of automation bias in healthcare artificial intelligence applications: a Bowtie analysis. J Safety Sci Resilience 2024 Dec;5(4):460-469. [doi: 10.1016/j.jnlssr.2024.06.001]
18. Dratsch T, Chen X, Rezazade Mehrizi M, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. Radiology 2023 May;307(4):e222176. [doi: 10.1148/radiol.222176] [Medline: 37129490]
19. Collado D. Digital health literacy: a cornerstone of health equity in the EU: policy brief. Health Action International. 2024 Oct. URL: https://haiweb.org/wp-content/uploads/2024/10/Digital-Health-Literacy-in-the-EU.pdf [accessed 2026-02-18]
20. Reyna VF, Nelson WL, Han PK, Dieckmann NF. How numeracy influences risk comprehension and medical decision making. Psychol Bull 2009 Nov;135(6):943-973. [doi: 10.1037/a0017327] [Medline: 19883143]
21. Peters E. Beyond comprehension: the role of numeracy in judgments and decisions. Curr Dir Psychol Sci 2012;21(1):31-35. [doi: 10.1177/0963721411429960]
22. Hildt E. What is the role of explainability in medical artificial intelligence? A case-based approach. Bioengineering (Basel) 2025 Apr 2;12(4):375. [doi: 10.3390/bioengineering12040375] [Medline: 40281735]
23. Pierce RL, Van Biesen W, Van Cauwenberge D, Decruyenaere J, Sterckx S. Explainability in medicine in an era of AI-based clinical decision support systems. Front Genet 2022;13:903600. [doi: 10.3389/fgene.2022.903600] [Medline: 36199569]
24. Elwyn G, Frosch D, Thomson R, et al. Shared decision making: a model for clinical practice. J Gen Intern Med 2012 Oct;27(10):1361-1367. [doi: 10.1007/s11606-012-2077-6] [Medline: 22618581]
25. Gigerenzer G, Edwards A. Simple tools for understanding risks: from innumeracy to insight. BMJ 2003 Sep 27;327(7417):741-744. [doi: 10.1136/bmj.327.7417.741] [Medline: 14512488]
26. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
27. Kolfschooten HV. EU regulation of artificial intelligence: challenges for patients' rights. Common Market Law Rev 2022 Feb 1;59(1):81-112. [doi: 10.54648/COLA2022005]

XSL•FO
RenderX

Edited by KA Clegg; submitted 10.Mar.2026; this is a non–peer-reviewed article;accepted 10.Mar.2026; published 19.Mar.2026.

Please cite as:
Ankolekar A
The Right to Understand in Health Care AI
J Med Internet Res 2026;28:e95090
URL: https://www.jmir.org/2026/1/e95090
doi:10.2196/95090

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Mar.2026.

XSL•FO
RenderX

Wearable Air Samplers Reveal How Wildfire Shapes the Exposome

Virginia Gewin

(J Med Internet Res 2026;28:e93193) doi:10.2196/93193

KEYWORDS

wildfire smoke; exposome; air pollutants; environmental; wearable electronic devices; environmental monitoring; firefighters; combustion products; polycyclic aromatic hydrocarbons; particulate matter; cardiovascular diseases

Textbox .

Key Takeaways

- Wearable silicone-based devices are increasingly being used to measure chemical exposures, notably following wildfires.
- Indoor air exposures linger far longer than outdoor exposures.
- Researchers are working to correlate environmental exposures with health outcomes.

Just over 1 year ago, devastating fires raged in Los Angeles (LA), California, killing 29 people [1]. The blaze destroyed 16,000 structures—homes full of computers, plastic, and furniture—subjecting roughly 20 million people to toxic smoke and ash [1,2]. Winds helped clear the air significantly by mid-January, but researchers have found that hazardous compounds lingered indoors for weeks—possibly longer.

Tracking the Exposome

To collect a profile of the chemicals residents were exposed to following the LA wildfires, Krystal Pollitt, PhD, PEng—an associate professor at the Yale School of Public Health—deployed a wearable passive air sampler to 30 standing homes between February and April 2025 [3].

Pollitt's *Fresh Air Clip* is one of a number of wearable devices—typically made of silicone polymers—that can collect a wide range of organic chemicals from the air or skin. These devices are designed to offer insights into the "exposome"—a term first described 20 years ago [4] as the environmental counterpart to the genome. Exposome research aims to quantify how environmental exposures translate into health outcomes. Wearables are increasingly being deployed after natural hazards to help researchers document chemical exposures from air pollution, as well as water contaminants, occupational hazards, or pathogens.

Ongoing Exposure

Pollitt's rapid response study found that once these chemicals permeate the home, residents can continue to be exposed as they slowly release from furniture, carpets, and even drywall [3].

Combustion byproducts were detected in all of the homes tested 2 to 3 months after the fire; even homes 12 km away from the burn zone had byproduct levels only 1.2 times lower than those in homes closer to the burn zone. Further, levels of the vast majority of combustion-related chemicals, such as flame retardants, petroleum products, plastics, and tire particulates, were elevated indoors compared to outdoors.

Pollitt explains that hundreds of chemicals that settled indoors stayed put, and they will continue to be released off surfaces slowly over time. She uses a nontargeted gas chromatography–high-resolution mass spectrometry approach able to detect volatile and semivolatile organic chemicals, including those not necessarily expected to be collected by silicone. "There's so many obscure products that were burning, and were transformed through the pyrolysis process, that to come in with a targeted method would be like looking for a pin in a haystack," she says.

In addition to obscure products, Pollitt found a wide range of building materials and household products that released numerous chemicals that have not been well characterized, especially following combustion. "There's tons of plastics present, but we don't know which specific plastic types yet," she says.

"This was a first pass study, done in collaboration with colleagues at University of California in Los Angeles, that required immediacy to understand how acute exposures evolved," says Pollitt. She and colleagues are now analyzing hundreds of different soft materials from the fires, including carpet swatches, blankets, pillows, and stuffed animals, to determine which remediation efforts would work best to reduce long-term exposures.

It has become clear that wearables will be a key component of exposomics research going forward. "Especially with the fires, people have realized there's a huge need to be able to do personalized exposure monitoring," says Mary Margaret Johnson, MD, PhD—principal research scientist at Harvard T.H. Chan School of Public Health—who coauthored a review paper of exposomics methodologies and future directions [5].

XSL•FO
RenderX

"We're realizing more and more that what is in the smoke needs to be accounted for," says Johnson. "There is a realization that the toxicity of the smoke depends on what is being burned in the home or area. Metals in the smoke vary from fire to fire," she says.

Measuring Health Impact

Other researchers are conducting similar studies using silicone wristbands—even broaches—that can typically be worn for a few hours to 1 week to collect exposome samples. Heather Stapleton, PhD—distinguished professor of Environmental Chemistry & Exposure Science at Duke University in Durham, North Carolina—has conducted numerous studies using silicone wristbands in diverse communities. One group she has focused on is structural and wildland firefighters, given their increased risk of cancer and other health conditions compared to the general population.

Stapleton's preliminary research found that, compared to structural firefighters, wildland firefighters appear to have higher exposure to combustion byproducts and some brominated flame retardants.

As Stapleton's previous research has shown [6], the risks aren't just from the fires either. Contaminated firefighting equipment can also expose workers to a mixture of different chemicals. As well, they can be exposed to per- and polyfluoroalkyl substances (PFASs) used to extinguish the fire or treat their gear. Stapleton found that blood PFAS levels were positively correlated with years working as a firefighter.

Stapleton and Pollitt agree that the utility of wearable exposure capture devices is still in the early days, but they look very promising. "I am collaborating with a colleague to modify the wristband's ability to absorb and detect metals," Stapleton says.

Other researchers are pairing wearable air quality sensors with Apple Watch data to assess long-term impacts of wildfires on cardiovascular health. One recent study [7] found that, for each 10-$\mu g/m^3$ increase in particulate matter, firefighters' resting heart rates rose by an estimated 0.38 beats per minute before the 2024 Park Fire in California and by approximately 1.4 beats per minute following the fire.

Stapleton says that, while likely too cost-prohibitive to replicate on a large scale, this study is a first step toward the larger questions that researchers must now tackle, like determining how wristband results correlate with chemical levels in the body. "We know how to take blood or urine measurements, and put those into a risk assessment framework," says Stapleton. "We aren't quite there yet with wristbands." Stapleton says that the next steps include creating a machine learning model to predict what wristband exposures mean for internal doses. "I think that's key to making sure this tool is successful and can be used more broadly," says Stapleton.

Beyond Acute Exposure

Johnson and her colleague—Harvard physician-scientist Kari Nadeau, MD, PhD—are conducting several experiments to pair what is found in blood with external exposure measures, including those from wearable wristbands. "In our lab, we are trying to use blood as the exposome window," she says. "We also heavily focus on multi-omics approach, looking at inflammatory markers being produced, or on epigenetic changes, by looking at immune functioning, and associating them with various environmental exposure levels simultaneously," says Johnson.

"Since the silicone wristbands are relatively cheap, researchers are trying to incorporate with the general population to estimate different chemical exposures," she says.

Pollitt, too, is now collecting biological samples to look beyond acute exposures, as part of the LA Fire Health Consortium—a decade-long study of fire impacts. "Now we're starting to understand persistence," she says. Pollitt was one of many scientists presenting their research at the LA Fire Health Consortium conference in January to mark the 1-year anniversary of the fire. The saddest thing from the meeting, she described, was the realization of "what does this mean for the next fire? It will happen again—hopefully not on that scale."

In the meantime, researchers' continued adoption of silicone-based wearable devices will provide a unique lens into the exposome and how it is altered by growing climate change–related wildfire risk.

Conflicts of Interest

None declared.

References

1. California governor requests $40 billion for fire aid, Washington Post reports. Reuters. 2025 Feb 22. URL: https://www.reuters.com/world/us/california-governor-requests-40-billion-fire-aid-washington-post-reports-2025-02-22 [accessed 2026-02-20]
2. Withers A, Chmielewski D, Hay A, Lang H. Los Angeles wildfires trigger air quality warnings and health concerns. Reuters. 2025 Jan 10. URL: https://www.reuters.com/world/us/los-angeles-wildfires-trigger-air-quality-warnings-health-concerns-2025-01-10 [accessed 2026-02-20]
3. Lin EZ, Esenther S, Mascelloni M, Irfan F, Pollitt KJG. The Fresh Air wristband: a wearable air pollutant sampler. Environ Sci Technol Lett 2020 Jan 28;7(5):308-314. [doi: 10.1021/acs.estlett.9b00800]

XSL•FO
RenderX

4. Wild CP. Complementing the genome with an “exposome”: the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomarkers Prev 2005 Aug;14(8):1847-1850. [doi: 10.1158/1055-9965.EPI-05-0456] [Medline: 16103423]
5. Wan M, Simonin EM, Johnson MM, et al. Exposomics: a review of methodologies, applications, and future directions in molecular medicine. EMBO Mol Med 2025 Apr;17(4):599-608. [doi: 10.1038/s44321-025-00191-w] [Medline: 39870881]
6. Hoxie T, Zhang S, Cooper EM, et al. Assessing nondietary exposure to per- and polyfluoroalkyl substances (PFASs) in firefighters using silicone wristbands. Environ Health (Wash) 2025 Dec 3. [doi: 10.1021/envhealth.5c00316]
7. Cheong SM, Perez M, Jung YS, Kim B, Nadeau K. Latent effects of wildfire particulate matter on resting heart rate measured with wearable devices. Circ Arrhythm Electrophysiol 2025 Aug;18(8):e013836. [doi: 10.1161/CIRCEP.125.013836] [Medline: 40709472]

Edited by KA Clegg; submitted 09.Feb.2026; this is a non–peer-reviewed article;accepted 09.Feb.2026; published 26.Feb.2026.

Please cite as:
Gewin V
Wearable Air Samplers Reveal How Wildfire Shapes the Exposome
J Med Internet Res 2026;28:e93193
URL: https://www.jmir.org/2026/1/e93193
doi:10.2196/93193

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Feb.2026.

XSL•FO
RenderX

Influencing the Influencers: How Health Experts Are Partnering With Content Creators to Fight Misinformation Online

Wendy Glauser

(J Med Internet Res 2026;28:e93450) doi:10.2196/93450

KEYWORDS

social media; opinion leaders; health communication; information dissemination; misinformation; public health; health promotion

Textbox .

Key Takeaways

- By offering training and support to content creators, public health advocates can improve the quality of health information on social media platforms.
- Partnering with content creators with diverse audiences can help health communicators reach populations they're struggling to engage.

Scroll. "Five essential questions that you must ask your surgeon prior to any operation."

Click. "Yesterday, I posted a picture of when I started on Ozempic and four weeks after Ozempic and many of you said...."

Swipe. "The World Health Organization lists processed meats like bacon, ham, salami and sausages in the same category as cigarettes and asbestos. But what does that even mean?"

These are three real examples [1-3] of the millions or potentially billions of TikTok videos that contain health information. In a Kaiser Family Foundation survey of over 1000 adults who were selected to be representative of the US population, 55% said they use social media to find health information at least occasionally [4].

Social media creators can influence what people eat, how they exercise, what medications they take or don't take, whether they seek out immunizations and preventive health screenings, and the list goes on. Whether that influence helps or harms depends in part on the accuracy of the information shared, and that information is not always accurate. For example, in a study that analyzed 1000 TikTok videos on mental health, 6.3% contained disinformation, which was defined as providing false or scientifically unsupported information, and 15.7% contained misinformation, which was defined as providing only partially supported information [5].

Realizing the potential impact, public health and health communication experts have begun to partner with content creators to fight misinformation online. As Matthew Motta, PhD—associate professor of Health Law, Policy and Management at the Boston University School of Public Health—explains, "If you can partner with just a handful of social media content creators, you might be able to reach millions of people with your message."

Collaboration Between Health Experts and Influencers

Three years ago, Amanda Yarnell, MS—senior director of the Harvard T.H. Chan School of Public Health's Center for Health Communication—launched the Creator Program [6] to help social media creators "counter health misinformation and spread evidence-based science to their communities," as the program's website describes. Yarnell explains that, at the time, "there was very little information, even about the health information creators shared." In the past few years, more research evaluating patterns in the content, style, and tone of health information shared on social media has emerged, but there are still very few health researchers attempting to intervene in those conversations by engaging creators.

While other programs have popped up in recent years that support health providers and scientists who are also creators, such as the World Health Organization's Fides program [7] or the Evidence Collective [8], the Creator Program is unique in that it also collaborates with influencers who don't necessarily have health or science degrees.

The impact of the Creator Program's initiatives is being studied. In one 2024 study [9], 105 English-language creators, who collectively reached about 17 million followers, were randomly assigned to a control group, which received asynchronous digital tool kits summarizing evidence-based mental health information, or a group that received the tool kits plus optional virtual training sessions. Researchers analyzed 3465 videos posted before and after the intervention. Post intervention, 26% of the creators who received tool kits and training invites, versus 21% of the control group, referenced the evidence-based themes included in the toolkit. The authors wrote that "although these effects are relatively small in substantive size, their reach," given the size of their audiences,"may be comparatively larger."

In another study (currently a preprint), Motta and colleagues [10] engaged social media influencer Kenzie Brenna to create two videos on how to respond when a peer is experiencing

XSL•FO
RenderX

emotional distress; one was created before a virtual training program delivered by mental health experts, and one was created afterward. A demographically representative sample of 1000 US youths between the ages of 14 and 22 years were then randomly assigned to view either video. The effect was modest but significant: 80% of those who viewed the posttraining video described it as informative (vs 60% of those who viewed the pretraining video), and more people in the group who viewed the posttraining video said that they felt confident in providing emotional support to a distressed peer.

Sasha Hamdani, MD—a content creator and psychiatrist—took part in the Creator Program virtual training sessions and also received a toolkit. Being part of the Creator Program, she says, "reinforced that even the most accurate data can be misunderstood if it is not communicated clearly and responsibly." Hamdani, who has about a million followers on TikTok and over 930,000 followers on Instagram, shares videos on mental health, especially attention-deficit/hyperactivity disorder (ADHD). She says being part of the Creator Program reminded her that "good health communication reduces misinformation by being precise, transparent about limitations, and intentional about language, especially in digital spaces where content travels quickly."

Reaching More People, More Often

Engaging with social media influencers allows public health experts to reach more people and "target diverse folks that public health organizations have failed at reaching, such as young people, as well as Black, Brown, Indigenous, and LGBTQ people," says Yarnell. Many of the influencers in the Creator Program, for example, have audiences that include a large proportion of people from one or more of these groups.

When it comes to prebunking (preparing people to resist misleading claims before they're exposed to them) and debunking (countering falsehoods after exposure), social media is the tool that "has the most promise currently," says Shaon Lahiri, PhD—assistant professor of public health at the College of Charleston in South Carolina. "It's where the misinformation is coming from," he explains. Plus, by virtue of engaging with audiences daily, creators can share similar information multiple times in different ways, or as Lahiri describes it, they can create "booster shots" to better inoculate people against misinformation.

More Engagement, Less Misinformation

Popular content creators tend to use methods that are engaging and impactful, as they rely on their metrics to learn what resonates. People are more persuaded by messaging that is "fun and not very heavy-handed," Lahiri explains. The technology capabilities in the social media platforms themselves also lend well to misinformation debunking/prebunking. For example, "you can do a Stitch on TikTok where you briefly show the misinformation and then you explain what's wrong with it," Lahiri explains, referring to taking a short clip from someone else's post and adding it to the beginning of your own.

Motta recognizes that these programs compete with billion-dollar companies and "social media companies that have really dropped the ball" in reducing misinformation. Still, he says, "there are plenty of content creators who want to fight against the spread of misinformation on social media platforms. These same creators are also really good at figuring out how to make sure that information permeates through."

Motta adds that even though social media videos are too often inaccurate, the vast majority of creators aren't intentionally sharing misleading information. As misinformation continues to travel faster than evidence, engaging the people who already command attention online may be the best way for public health to keep up.

Conflicts of Interest

None declared.

References

1. @drjeremylondon. Top 5 questions to ask your doctor before having surgery. TikTok. 2024 Feb 26. URL: https://www.tiktok.com/@drjeremylondon/video/7339983405460802859 [accessed 2026-02-19]
2. @lydiawanjiru.ke. Before and after one month Ozempic. TikTok. 2025 Feb 18. URL: https://www.tiktok.com/@lydiawanjiru.ke/video/7472688092751695125 [accessed 2026-02-19]
3. @drsermedmezher. Bacon and cancer risk: what you need to know. TikTok. 2026 Jan 1. URL: https://www.tiktok.com/@drsermedmezher/video/7590527954610949378 [accessed 2026-02-19]
4. Schumacher S, Sparks G, Montalvo JIII, Kirzinger A, Hamel L. KFF tracking poll on health information and trust: health information and advice on social media. KFF. 2025 Aug 7. URL: https://www.kff.org/public-opinion/kff-health-information-and-trust-tracking-poll-health-information-and-advice-on-social-media/ [accessed 2026-02-19]
5. Hudon A, Perry K, Plate AS, et al. Navigating the maze of social media disinformation on psychiatric illness and charting paths to reliable information for mental health professionals: observational study of TikTok videos. J Med Internet Res 2025 Jun 18;27:e64225. [doi: 10.2196/64225] [Medline: 40532184]
6. Creator Program | Center for Health Communication. Harvard T.H. Chan School of Public Health. URL: https://hsph.harvard.edu/research/health-communication/creator-program/ [accessed 2026-02-19]

XSL•FO
RenderX

7. Fides - a network of healthcare influencers dedicated to raising good health content and fighting misinformation. World Health Organization. URL: https://www.who.int/teams/digital-health-and-innovation/digital-channels/fides [accessed 2026-02-19]
8. The Evidence Collective. URL: https://www.evicollective.org/ [accessed 2026-02-19]
9. Motta M, Liu Y, Yarnell A. “Influencing the influencers:” a field experimental approach to promoting effective mental health communication on TikTok. Sci Rep 2024 Mar 11;14(1):5864. [doi: 10.1038/s41598-024-56578-1] [Medline: 38467750]
10. Motta M, Liu Y, Brenna K, et al. Assessing the effect of evidence-based mental health social media videos on youth emotional support competencies: a “creator-engaged” approach. Open Science Framework. Preprint posted online on Dec 5, 2024 URL: https://osf.io/preprints/socarxiv/kydf4_v1 [accessed 2026-02-19] [doi: 10.31235/osf.io/kydf4]

Edited by KA Clegg; submitted 12.Feb.2026; this is a non–peer-reviewed article;accepted 12.Feb.2026; published 27.Feb.2026.

Please cite as:
Glauser W
Influencing the Influencers: How Health Experts Are Partnering With Content Creators to Fight Misinformation Online
J Med Internet Res 2026;28:e93450
URL: https://www.jmir.org/2026/1/e93450
doi:10.2196/93450

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Is Going Analog Good for Children and Teens' Mental Health and Well-Being?

Sara Novak

(J Med Internet Res 2026;28:e94018) doi:10.2196/94018

KEYWORDS

adolescent; child; mental health; well-being; smartphones; social media; digital technology; screen time; health behaviour; parental guidance; media use; sleep; anxiety; depression; health promotion

Textbox .

Key Takeaways

- More and more parents are waiting to give their children a smartphone.
- Children and teens who spend more time on devices are more likely to develop addictive behaviors, have difficulty sleeping, and experience cyberbullying.
- Parents' smartphone use plays a significant role in how their children and teens engage with their devices.

As children grow older, parents face a mounting sense of pressure to provide them with smartphones. By the time they enter middle school, children who have devices far outweigh those who don't; according to Common Sense Media, 42% have a phone by age 10, 71% by age 12, and 91% by age 14 [1].

Still, a growing body of evidence is motivating parents to hold off on smartphones until the later teenage years. In fact, more and more families are taking a long, hard look at their digital life and trading in their smartphones for flip phones and landlines.

Many Families Want A More Low-Tech Lifestyle

Parents like Paul Jones, a father of three, want to make sure that their children are both educated on device safety and aware of the potential downsides.

"Technology isn't going anywhere, if anything it's advancing so we need to educate children on technology and how to use it safely and properly," said Jones. That's why he started FlippGen [2], a youth-focused organization where young adults go into schools and provide perspective on how teens can positively balance device use.

Jones said that his children will not get devices until age 16, and instead he and a "pod" of friends and family have invested in landlines. According to a recent survey, Gen Z parents are holding off on smartphones, with 28% waiting until their children are between the ages of 15 and 17 to purchase them [3]. Jones' children have a landline provided through a company called Tin Can, a popular Wi-Fi landline for kids that doesn't have apps, texting, or games and only allows kids to speak to those on their approved contact list [4].

Additionally, some teens that already have smartphones have been trading them in for flip phones in order to remain reachable while reducing distractions and time they view as wasted on social media sites. The trend isn't new, but it appears to be persistent: one 2023 report noted an increase in US flip phone sales [5], and recent 2026 coverage highlights a growing interest in being "chronically offline" [6].

Does Age Matter When It Comes to Responsible Smartphone Use?

The research is mixed when it comes to the best age to give a child a smartphone. One study published in the journal *Pediatrics* found that children who had smartphones by age 12 were more likely to suffer from an increased risk of depression, obesity, and insufficient sleep [7], but another study published in the journal *Child Development* showed that grades, sleep habits, and depression symptoms were not impacted by the age that children acquired their first devices [8].

Still, even with these mixed results, current data may still point to holding off—or at the very least, educating on and carefully monitoring device use. More than age, what seems to matter for well-being is how and how much children and teens use their smartphones.

Moderate, intentional, guided use may provide benefits for social connection or support, depending on the kind of use [9,10]—for example, if a child or teen shares about a positive event, achievement, or cause that they're involved in, or finds like-minded peers going through similar trials. Additionally, there may be a level of perceived social isolation, disconnection, and fear of missing out that occurs for some whose parents hold off on getting them smartphones, as well as perceived trade-offs in the other potential practical and logistical benefits smartphones might offer [11-14].

But on the other hand, excessive use without supervision can contribute to loss of sleep, cyberbullying, and exposure to distressing or inappropriate content [15,16]. And even if their

XSL•FO
RenderX

well-being doesn't suffer and they're able to keep their grades up, the average American child spends 5 hours daily on their smartphones [17].

"While social media can be used to forge some connections, the type of connections and socialization that occurs on social media is often quite different than the types of conversations and relationships that one would have face to face," said Jason Nagata, MD, an associate professor of Pediatrics in the Division of Adolescent and Young Adult Medicine at the University of California, San Francisco.

Potential Risks of Excessive Use

Social media addiction is much more likely to occur in children and teens who are allowed unfettered access to a smartphone. While not a formal diagnosis, it can present many of the same symptoms as substance abuse–related and other addiction disorders, including impaired control, withdrawal, tolerance, and relapse. It can contribute to diminished engagement in other activities and significantly impair a child's daily functioning, interpersonal relationships, and overall quality of life [17-20].

Sleep disruption is another potential negative for children and teens with unlimited access to their devices. A study published in the August 2023 edition of the journal *Sleep Health* [21] found that 63.2% of 11-12 year olds had a digital device in their room, and approximately 17% of them reported that it had awoken them in the past week. One-fifth of adolescents said that they used their device if they woke up during the night, which was associated with a higher risk of sleep disturbance.

Additionally, according to Nagata, the most likely explanation for social media's impact on mental health results from cyberbullying. A study published in the June 2025 issue of the journal *The Lancet Regional Health – Americas* found that children ages 11 to 12 years who were cyberbullied were more likely to have depressive symptoms and 2.62 times more likely to report suicidal ideation or a suicide attempt one year later [22].

Modeling Positive Digital Behavior

To mitigate some of the potential negative effects of smartphone use on children and teen's well-being, the American Academy of Pediatrics recommends establishing a healthy family media plan that the entire family can abide by [23]. Parents should establish screen-free zones, confiscating phones at a certain hour every night and not allowing them in the bedroom or during meals, recommendations that align with Nagata's research [24].

Perhaps most importantly, however, parents should model these behaviors for their children and teens. The single biggest predictor of a child's relationship with the digital world is not the age that parents provide them with a smartphone, but rather, a parent's own use of their devices, according to Michael Rich, MD, MPH, Director and Founder, Digital Wellness Lab, and associate professor of pediatrics at Harvard Medical School.

"We need to think of smartphones less in terms of entitlement and more in terms of the powerful tools that they are," said Rich. Powerful tools with instruction manuals the research is still writing—making prudent, thoughtful use essential, especially when the users are children and teens.

Conflicts of Interest

None declared.

References

1. Rideout V, Peebles A, Mann S, Robb MB. Common sense census: media use by tweens and teens, 2021. : Common Sense; 2022 URL: https://www.commonsensemedia.org/sites/default/files/research/report/8-18-census-integrated-report-final-web_0.pdf [accessed 2026-03-03]
2. About. FlippGen. 2025. URL: https://www.flippgen.com/about [accessed 2026-03-03]
3. Nesi J. A child's first cell phone: a guide for parents. Brown Health. 2022 Oct 20. URL: https://www.brownhealth.org/be-well/childs-first-cell-phone-guide-parents/ [accessed 2026-03-03]
4. Menu. Tin Can. 2026. URL: https://tincan.kids/products/tin-can [accessed 2026-03-03]
5. Mays L. Dumb phones are on the rise in the U.S. as Gen Z looks to limit screen time. CNBC. 2023 Mar 29. URL: https://www.cnbc.com/2023/03/29/dumb-phones-are-on-the-rise-in-the-us-as-gen-z-limits-screen-time.html [accessed 2026-03-03]
6. Bhaimiya S. A 'quiet revolution': why young people are swapping social media for lunch dates, vinyl records and brick phones. CNBC. URL: https://www.cnbc.com/2026/02/07/young-people-quiet-revolution-social-media.html [accessed 2026-03-03]
7. Barzilay R, Pimentel SD, Tran KT, Visoki E, Pagliaccio D, Auerbach RP. Smartphone ownership, age of smartphone acquisition, and health outcomes in early adolescence. Pediatrics 2026 Jan 1;157(1):e2025072941. [doi: 10.1542/peds.2025-072941] [Medline: 41324306]
8. Sun X, Haydel KF, Matheson D, Desai M, Robinson TN. Are mobile phone ownership and age of acquisition associated with child adjustment? A 5-year prospective study among low-income Latinx children. Child Dev 2023 Jan;94(1):303-314. [doi: 10.1111/cdev.13851] [Medline: 36056573]
9. Dharejo N, Alivi MA, Rahamad MS, et al. Effects of social media use on adolescent psychological well-being: a systematic literature review. Int J Interact Mob Technol 2023;17(20):171-191. [doi: 10.3991/ijim.v17i20.44663]

XSL•FO
RenderX

10. Yang CC, Hunhoff P, Lee Y, Abrell J. Social media activities with different content characteristics and adolescent mental health: cross-sectional survey study. J Med Internet Res 2025 Apr 28;27:e73098. [doi: 10.2196/73098] [Medline: 40294901]
11. Matthes J, Stevic A, Koban K, Thomas MF, Forrai M, Karsay K. Fear of missing out, reflective smartphone disengagement, and loneliness in late adolescents. Cyberpsychol Behav Soc Netw 2023 Oct;26(10):731-738. [doi: 10.1089/cyber.2023.0014] [Medline: 37582211]
12. Ophir Y, Rosenberg H, Lipshits-Braziler Y, Amichai-Hamburger Y. "Digital adolescence": the effects of smartphones and social networking technologies on adolescents' well-being. In: Online Peer Engagement in Adolescence 2020:122-139. [doi: 10.4324/9780429468360-8]
13. Coplan RJ, McVarnock A, Hipson WE, Bowker JC. Alone with my phone? Examining beliefs about solitude and technology use in adolescence. Int J Behav Dev 2022 Nov;46(6):481-489. [doi: 10.1177/01650254221113460] [Medline: 36397736]
14. Vaterlaus JM, Carney T, Kroemer-Spiess P. Delaying adolescent smartphone ownership: a qualitative dyadic study with parents and adolescents. J Child Media 2026:1-21. [doi: 10.1080/17482798.2026.2627286]
15. Martin JD, Song SW, Rote WR, et al. The life in media survey: a baseline study of digital media use and well-being among 11- to 13-year-olds. : Researchers at the University in South Florida in collaboration with The Harris Poll; 2025 URL: https://www.usf.edu/arts-sciences/departments/journalism/documents/lims2025/lims2025.pdf [accessed 2026-03-03]
16. Martín-Cárdaba M, Martínez Díaz MV, Lafuente Pérez P, García Castro J. Smartphone ownership, minors' well-being, and parental mediation strategies. An analysis in the context of social media influencers. J Youth Adolesc 2024 Oct;53(10):2202-2218. [doi: 10.1007/s10964-024-02013-7] [Medline: 38782845]
17. Media use and screen time – its impact on children, adolescents, and families. : American College of Pediatricians; 2020 May URL: https://acpeds.org/wp-content/uploads/2025/09/Media-Use-and-Screen-Time-May-2020.pdf [accessed 2026-03-03]
18. Nagata JM, Singh G, Sajjad OM, et al. Social epidemiology of early adolescent problematic screen use in the United States. Pediatr Res 2022 Nov;92(5):1443-1449. [doi: 10.1038/s41390-022-02176-8] [Medline: 35768491]
19. Wong SL, King N, Gariépy G, et al. Adolescent social media use and its association with relationships and connections: Canadian Health Behaviour in School-aged Children, 2017/2018. Health Rep 2022 Dec 21;33(12):14-23. [doi: 10.25318/82-003-x202201200002-eng] [Medline: 36542360]
20. Bozzola E, Spina G, Agostiniani R, et al. The use of social media in children and adolescents: scoping review on the potential risks. Int J Environ Res Public Health 2022 Aug 12;19(16):9960. [doi: 10.3390/ijerph19169960] [Medline: 36011593]
21. Nagata JM, Singh G, Yang JH, et al. Bedtime screen use behaviors and sleep outcomes: Findings from the Adolescent Brain Cognitive Development (ABCD) Study. Sleep Health 2023 Aug;9(4):497-502. [doi: 10.1016/j.sleh.2023.02.005] [Medline: 37098449]
22. Nagata JM, Shim J, Balasubramanian P, et al. Cyberbullying, mental health, and substance use experimentation among early adolescents: a prospective cohort study. Lancet Reg Health Am 2025 Jun;46:101002. [doi: 10.1016/j.lana.2025.101002] [Medline: 40625792]
23. How to make a family media plan. American Academy of Pediatrics. 2026. URL: https://www.healthychildren.org/English/family-life/Media/Pages/How-to-Make-a-Family-Media-Use-Plan.aspx [accessed 2026-03-03]
24. Nagata JM, Paul A, Yen F, et al. Associations between media parenting practices and early adolescent screen use. Pediatr Res 2025 Jan;97(1):403-410. [doi: 10.1038/s41390-024-03243-y] [Medline: 38834780]

Edited by KA Clegg; submitted 23.Feb.2026; this is a non–peer-reviewed article;accepted 23.Feb.2026; published 05.Mar.2026.

Please cite as:
Novak S
Is Going Analog Good for Children and Teens' Mental Health and Well-Being?
J Med Internet Res 2026;28:e94018
URL: https://www.jmir.org/2026/1/e94018
doi:10.2196/94018

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Mar.2026.

XSL•FO
RenderX

Physician, Restore Thyself? The Digital Gap in Physician Well-Being Support

Jenny Castillo Cato

(J Med Internet Res 2026;28:e93338) doi:10.2196/93338

KEYWORDS

physicians; peer support; social support; mental health; physician burnout; well-being; occupational stress; mobile apps; digital health; psychological resilience

Textbox .

Key Takeaways

- Physician well-being is a long-standing, complex issue that needs to be better addressed.
- Current digital applications are helpful, but primarily focus on physician self-improvement.
- The digital world is a largely untapped medium for supporting equally critical factors: community, connection, and culture change.

Jenny Castillo Cato is board-certified in emergency medicine and has worked as an attending physician for over 20 years. She is an advocate for physician well-being and has worked on physician well-being initiatives in New York and across the United States.

In March 2020, the COVID-19 pandemic hit New York City like a tornado. The flood of patients coming to the emergency departments of the five boroughs was beyond imagination. Staff were getting sick, supplies were running out, and no treatment had been developed for this brand-new virus.

As physicians, we had to make hard decisions. Should I put the 80-year-old man who was barely responsive and hypoxic on the last ventilator in the hospital? Or should I keep the ventilator for the next patient who may be younger and more likely to withstand the onslaught of this deadly virus? Almost 6 years later, those memories still haunt me.

Unspoken Burden

Physicians have long been expected to rise to these kinds of occasions and to perform at a high level across many challenging circumstances, and we do. There is, however, an emotional toll on physician well-being, one that we are frequently expected to carry alone. This issue is not new. Though often hidden out of fear of discovery and shame, it has persisted throughout the centuries. In the 1860s, Dr James Still wrote in his memoirs that he often felt lonely, with “no one with whom to hold conference or to cheer or comfort me when depressed with difficult cases” [1]. This mentality has continued through the ages, was highlighted during the COVID-19 pandemic, and remains a pressing concern today.

Physicians’ mental health is a crucial aspect of well-being—ours and, ultimately, our patients’—yet the stigma surrounding mental health assistance has led to dire consequences. The American Medical Association reports that physicians are 82% more likely to experience burnout than those in other fields [2]. Furthermore, in the United States, between 300 and 400 physicians die by suicide each year, and the suicide rate for female physicians is 250% to 400% higher than for female workers in other professions [3]. Estimates vary and tend to be higher in the United States, but one study suggests that, across countries, physicians die by suicide, on average, 44% more often than expected compared to the general population [4]. An increase in risk factors (such as administrative burden and excessive job demands) and a decline in protective factors (such as autonomy and peer-based support) may play a role [5].

Foundations like the Dr. Lorna Breen Hero’s Foundation—named after my friend and colleague who died by suicide in April 2020—were created to advocate for change [6]. Nonetheless, progress remains slow and challenging due to factors such as physician personality traits, stigma within the medical community, and entrenched norms, including mandatory mental health questions on licensure applications that perpetuate the idea that mental illness equates to weakness or incompetence [7].

Persistent Gaps in Support

To support physician mental health and well-being, the medical community promotes self-care such as mindfulness, exercise, better sleep habits, and nutrition.

Many organizations, such as the American College of Physicians and the American Academy of Allergy, Asthma, and Immunology, offer useful well-being resources for this kind of self-improvement. The resources highlight applications such as Headspace, a meditation-based app; My Well-Being Index, an anonymous evaluation tool from the Mayo Clinic; Provider Resilience, a self-assessment tool for compassion fatigue and burnout; and T2 Mood Tracker, an app that monitors anxiety, depression, posttraumatic stress disorder, and general well-being [8,9]. There are also a few specialty-specific applications like the Surgeon’s Masters, which focus on physical wellness with

XSL•FO
RenderX

the 7-minute workout; Sleepo, a sleep tracker; and WaterMinder, a way to monitor hydration [10].

But while informative and helpful, these applications do not fully address the complex issue of physician well-being. The COVID-19 pandemic highlighted that well-being could not be addressed solely at the individual level but that significant changes must be made in clinical environments as well. Vast organizational norms need to be altered to transform physician well-being; it requires a cultural, political, and environmental shift.

There is also no one right answer for supporting physician well-being; it cannot be accomplished through a "one size fits all" approach. Each medical specialty has different challenges and needs. For example, what is important and impacts an orthopedic surgeon's daily work-life is very different from what matters to a family medicine physician seeing patients in a clinic. The physical stress, mental toll, and teamwork required during an 8-hour surgery affect the surgeon differently than a physician who manages 40 patients who are chronically ill every 15 minutes for 8 hours.

In short, persistent gaps remain in the resources available for physicians.

New Solutions

Like Dr Still noted in the 1860s, physicians need community, support, and camaraderie. This is particularly pressing as many medical institutions—especially in New York City, where space is a premium—have opted to use physicians' lounges for administrative purposes, leading to less direct physician contact across specialties.

A new app called Roon is trying to address this issue. Though Roon began as a video-based question and answer platform of health issues for patients, the final product created a "physicians-only knowledge network" providing a "digital doctor's lounge where wisdom of one becomes the resource of many" [11]. Roon and other programs like PeerRxMed, a free health care provider peer-to-peer program [12], are beginning to address the complex nature of physician well-being by facilitating community, peer support, and camaraderie. Beyond these tools, however, solutions in the digital space have yet to be fully tapped into.

A Call to Action

More resources are needed to support physician mental health and well-being, particularly resources that go beyond self-care and self-improvement.

New applications need to be created with a collaborative mindset. Clinicians must work with programmers and graphic designers to create content that is relevant and specific to each specialty. Combining directed self-care options, team building, leadership skills, and peer support into an easy-to-use, focused application would be incredibly valuable for the newer generations entering medicine.

The consequences of *not* developing these tools include and extend beyond physicians: without a healthy medical workforce, everyone would be affected. While we love our profession, physicians need resources to help us not only put on our own masks first but also support one another and share the weight so that we can more effectively care for patients.

The digital world remains an untapped medium for building physician community and culture change. Let us act now to harness its power to help physicians holistically cultivate better health, positive work-life integration, and institutional changes, creating an environment that allows longer, happier careers.

Conflicts of Interest

None declared.

References

1. Still J. Early Recollections and Life of Dr James Still: J.B. Lippincott and Company; 1877.
2. Berg S. Battling doctor burnout? New app makes a game of learning to beat it. American Medical Association. 2019 Aug 26. URL: https://www.ama-assn.org/practice-management/physician-health/battling-doctor-burnout-new-app-makes-game-learning-beat-it [accessed 2026-01-25]
3. Matheson J. Physician suicide. American College of Emergency Physicians. URL: https://www.acep.org/life-as-a-physician/wellness/wellness/wellness-week-articles/physician-suicide [accessed 2026-01-25]
4. Dutheil F, Aubert C, Pereira B, et al. Suicide among physicians and health-care workers: a systematic review and meta-analysis. PLoS ONE 2019;14(12):e0226361. [doi: 10.1371/journal.pone.0226361] [Medline: 31830138]
5. Harvey SB, Epstein RM, Glozier N, et al. Mental illness and suicide among physicians. Lancet 2021 Sep 4;398(10303):920-930. [doi: 10.1016/S0140-6736(21)01596-8] [Medline: 34481571]
6. Dr. Lorna Breen Heroes Foundation. URL: https://drlornabreen.org/ [accessed 2026-01-25]
7. Ng IK, Tan BC, Goo S, Al-Najjar Z. Mental health stigma in the medical profession: where do we go from here? Clin Med (Lond) 2024 Jan;24(1):100013. [doi: 10.1016/j.clinme.2024.100013] [Medline: 38382183]
8. Resources for physician wellness. American Academy of Allergy Asthma & Immunology. URL: https://www.aaaai.org/practice-management/practice-tools/wellness/resource [accessed 2026-01-25]

XSL•FO
RenderX

9. Individual physician wellness and burnout tools. American College of Physicians. URL: https://www.acponline.org/practice-career/physician-well-being-and-professional-fulfillment/individual-physician-wellness-and-burnout-tools [accessed 2026-01-25]
10. SurgeonMasters. URL: https://surgeonmasters.com [accessed 2026-01-25]
11. Roon. URL: https://www.roon.com [accessed 2026-01-25]
12. PeerRxMed. URL: https://www.peerrxmed.com [accessed 2026-01-25]

Edited by KA Clegg; submitted 11.Feb.2026; this is a non–peer-reviewed article;accepted 11.Feb.2026; published 27.Feb.2026.

Please cite as:
Cato JC
Physician, Restore Thyself? The Digital Gap in Physician Well-Being Support
J Med Internet Res 2026;28:e93338
URL: https://www.jmir.org/2026/1/e93338
doi:10.2196/93338

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Using AI to Assess Potential Zoonotic Threats

Cliff Dominy

(J Med Internet Res 2026;28:e93261) doi:10.2196/93261

KEYWORDS

influenza A virus; H5N1 subtype; avian influenza; zoonoses; cross-species transmission; humans; animals, wild; farm animals; machine learning; genomics; protein structure; predictive modeling; pandemics; emerging infectious diseases; viral evolution; bioinformatics

Textbox .

Key Takeaways

- Potential mutations in avian flu strains pose a credible future pandemic risk.
- A recent study developed a novel machine learning model to identify genomic features with potential for spillover to human hosts.
- The approach may have additional applications, including guiding the development of targeted influenza vaccines.

The next viral pandemic is an increasingly likely possibility in the coming decades [1], and while we have some guesses, ultimately we don't know exactly where or when to expect it [2]. What the world needs is a NATO (North Atlantic Treaty Organization)–style early warning radar to alert us to these incoming threats. Ideally, the radar must distinguish between the viral majority that pose no risk to human health and the very few that do. And one more thing, it should warn us of this potential pandemic pest before it spills over into humans.

Is that too much to ask? Recent research suggests it might not be.

A Potential Solution

A team of virologists and computational biologists led by Liam Brierley, PhD, while at the University of Liverpool (now at the University of Glasgow), may have come up with a solution to the problem of preparing for future pandemics. The group developed a machine learning (artificial intelligence [AI]) model capable of predicting which strains of avian influenza circulating in animals have the potential to jump into humans. Their research paper [3] is currently available as a preprint and is undergoing peer review—but the results at this early stage look promising.

Traditional phylogenetics is a retrospective tool, analyzing sequences from existing databases to flag new variants of potential pandemic concern [4,5]. In contrast, the Brierley et al [3] paper introduces a predictive biophysical information layer to detect viruses with host-jumping potential. The approach was designed to identify protein and nucleic acid sequences that share a functional similarity with existing zoonotic influenza strains. As Brierley explained via email, "The idea behind the machine learning approach is, if we can identify what makes a zoonotic virus from these 'first principles' of protein function and nucleotide motifs, we can apply this to highly divergent sequences that a phylogeny would struggle to make inference about, because it would be equally distant from all sequences we currently know." In other words, by considering factors like protein motif functionality, we might be able to identify signatures of host-jumping potential between distantly related viruses—something traditional phylogenetic analyses would miss.

AI on Defense

Avian flu has been around for centuries but is a relatively new pathogen in humans [6]. In 1997, authorities confirmed the first case of a virus carrying the signature H5 and N1 bird flu proteins in a human [7]. This new human-compatible virus has since spread rapidly through the world's bird population. So far, there have been numerous confirmed human fatalities but no proof of human-to-human transmission [8,9]. H5N1 has also impacted other mammals, decimating the sea lion and elephant seal populations in Peru in 2023 [10]. Farmworkers exposed to their livestock account for most of the over 600 human cases reported to date [11].

Brierley chose the H5N1 influenza virus as a use case for his analysis because it was a credible future pandemic threat but trained the model on all known subtypes of avian flu. He writes, "It's important to capture that whole diversity if we want to be able to make predictions for new strains in future."

An advantage for the number crunchers was the extensive avian influenza genetic database at their disposal. Currently, nearly 19,000 distinct viral sequences are available covering 120 different viral subtypes. What is more, 618 samples came from humans. The research team trained its AI model on a subset of the database and then tested it against the remaining data. They were seeking reliable and wide-scale genomic signals associated with human infection, combined with the physical and chemical properties of short protein regions identified by their model, and with the samples taken from human beings.

XSL•FO
RenderX

Key Findings

The results showed the model had a 91.9% chance of correctly identifying a virus at risk of spillover into humans. Notably, first-generation protein modeling software achieved this precision on a database that will likely grow and improve as surveillance programs add sequences from new outbreaks. While more data is needed, Brierley writes that, "The leaps in performance will come from more advanced models, which are not decades away but rather, years or even months."

While many regions of the influenza virus genome have similarities with each other, adding the protein motif predictions narrowed the spillover regions down to just a few key areas. These regions were small, often just two or three base pairs long, but mutations in these regions may be critical to allow a bird virus to propagate within human cells.

The AI model could cut through the noise generated from thousands of viral base pairs and highlight just a few protein motifs or patterns that mattered. They were:

- RNA polymerase complex: 9 motifs across the PA (polymerase acidic protein), PB1 (polymerase basic protein 1), and PB2 (polymerase basic protein 2) subunit genes, whose proteins are essential for the virus to replicate, were top of the list.
- Virus binding: 1 motif within the hemagglutinin (HA) gene was included. The "H" in the H5N1 virus family is involved in binding to the host cells. Mutations here could allow the virus to recognize and infect a new host.
- Replication: Motifs of the nucleoprotein (NP), a chaperone responsible for ensuring the host replicates the viral genes correctly and packages them for export, were included.
- Immune evasion: Motifs within a nonstructural gene (NS1), the gene involved in dampening the host cell's immune response after infection, were also included

The model has already shown potential for use with other influenza viruses. It flagged rare influenza viruses like H10N8, which has been detected in humans [12], as well as the H4 subtypes, which have not. Brierley notes that he can "highlight a couple of H4 sequences from H4N8 and H4N6 subtypes that seemed to have elevated zoonotic potential, and the next step would be laboratory studies to understand the exact mechanisms."

Next Steps

Future priorities for this approach would include improved modeling and expanding the database, especially with viruses that cause asymptomatic infections that could more easily spread undetected. Global health authorities can assist by monitoring for new emerging viruses and adding their genetic sequences to the database before they become more virulent.

AI will become an important tool in the public health armory. But despite the benefits reported in the study, machine learning cannot offer a complete solution. Brierley noted limitations of the approach using AI, noting that "it can't tell you exactly WHY a certain feature matters but it can tell you where we should start looking in deeper study," adding that "it can only do the job it's trained for. We can't, for example, make any predictions about spillover into other mammals from this model." Furthermore, AI can only analyze samples taken at a single point in time; it cannot gauge the direction of viral spread. And the influenza virus is subject to modification by its hosts, which may affect its virulence and transmissibility—for example, the addition of sugar and phosphate molecules on the mature virus can alter its pathogenicity, a feature that AI could miss.

AI on the Offense

This technology is not just an effective defense against future viral threats; it could also be used proactively during influenza season. The annual flu shot would benefit from the superior predictive power of AI over the traditional genetic approach to developing targeted vaccines. Beyond influenza, it could be readily adapted to other respiratory pathogens, including coronaviruses and the common cold.

In Brierley's words, "One thing is clear: as computing power grows and our model complexity starts to better capture biological realities," these models could become central in understanding and monitoring viruses with zoonotic potential, helping us anticipate and potentially mitigate the impact of the next pandemic.

Conflicts of Interest

None declared.

References

1. Marani M, Katul GG, Pan WK, Parolari AJ. Intensity and frequency of extreme novel epidemics. Proc Natl Acad Sci U S A 2021 Aug 31;118(35):e2105482118. [doi: 10.1073/pnas.2105482118] [Medline: 34426498]
2. Gomez LRE, Malysheva N, Pfeil J, et al. Holistic forecasting for future pandemics: a review of pathogens, models, and data. Discov Public Health 2025;22(1):1-17. [doi: 10.1186/s12982-025-00573-y]
3. Brierley L, Mould-Quevedo J, Baylis M. An AI for an AI: identifying zoonotic potential of avian influenza viruses via genomic machine learning. bioRxiv. Preprint posted online on Sep 16, 2025. [doi: 10.1101/2025.09.16.676011]
4. Scarafoni D, Telfer BA, Ricke DO, Thornton JR, Comolli J. Predicting influenza A tropism with end-to-end learning of deep networks. Health Secur 2019;17(6):468-476. [doi: 10.1089/hs.2019.0055] [Medline: 31859569]

XSL•FO
RenderX

5. Borkenhagen LK, Allen MW, Runstadler JA. Influenza virus genotype to phenotype predictions through machine learning: a systematic review: computational prediction of influenza phenotype. Emerg Microbes Infect 2021;10(1):1896-1907. [doi: 10.1080/22221751.2021.1978824] [Medline: 34498543]
6. Kim SM, Kim YI, Pascua PNQ, Choi YK. Avian influenza A viruses: evolution and zoonotic infection. Semin Respir Crit Care Med 2016 Aug;37(4):501-511. [doi: 10.1055/s-0036-1584953] [Medline: 27486732]
7. Isolation of avian influenza A(H5N1) viruses from humans—Hong Kong, May-December 1997. JAMA 1998 Jan 28;279(4):263. [doi: 10.1001/jama.279.4.263]
8. MacCosham A, Vasiliu AG, Atchessi N. A rapid review of the avian influenza PB2 E627K mutation in human infection studies. Can Commun Dis Rep 2025 Apr;51(4):137-144. [doi: 10.14745/ccdr.v51i04a04] [Medline: 40241713]
9. Technical report: June 2024 highly pathogenic avian influenza A (H5N1) viruses. Centers for Disease Control and Prevention. 2024 Jun 5. URL: https://www.cdc.gov/bird-flu/php/technical-report/h5n1-06052024.html [accessed 2026-02-24]
10. Leguia M, Garcia-Glaessner A, Muñoz-Saavedra B, et al. Highly pathogenic avian influenza A (H5N1) in marine mammals and seabirds in Peru. Nat Commun 2023 Sep 7;14(1):5489. [doi: 10.1038/s41467-023-41182-0] [Medline: 37679333]
11. Influenza (avian and other zoonotic). World Health Organization. 2026 Jan 21. URL: https://www.who.int/news-room/fact-sheets/detail/influenza-(avian-and-other-zoonotic) [accessed 2026-02-24]
12. Chen H, Yuan H, Gao R, et al. Clinical and epidemiological characteristics of a fatal case of avian influenza A H10N8 virus infection: a descriptive study. Lancet 2014 Feb 22;383(9918):714-721. [doi: 10.1016/S0140-6736(14)60111-2] [Medline: 24507376]

Edited by KA Clegg; submitted 10.Feb.2026; this is a non–peer-reviewed article;accepted 10.Feb.2026; published 03.Mar.2026.

Please cite as:
Dominy C
Using AI to Assess Potential Zoonotic Threats
J Med Internet Res 2026;28:e93261
URL: https://www.jmir.org/2026/1/e93261
doi:10.2196/93261

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 3.Mar.2026.

XSL•FO
RenderX

Bridging the Gender Gap in Health Care Innovation: The Evolution of FemTech

Jenny Castillo Cato

Abstract

FemTech is a rapidly growing sector in digital health. In this *News and Perspectives* article, JMIR Correspondent and physician Jenny Castillo Cato reports on the evolution and potential of FemTech initiatives for women's well-being.

(J Med Internet Res 2026;28:e103487) doi:10.2196/103487

KEYWORDS

women's health; reproductive health; menstrual cycle; pregnancy; telemedicine

Textbox .

Key Takeaways:

- Women make up half of the global population yet remain significantly underrepresented in health research and innovation.
- The emergence of FemTech, a sector at the intersection of technology and women's health care, continues to evolve and expand its offerings, with the potential to transform health care access, advance research, and address systemic gender disparities.

As of 2026, approximately 4.13 billion females make up 49.7% of the global population, with over 21 million adolescent girls in low- and middle-income countries experiencing pregnancy annually. Vaccination rates for human papillomavirus remain low, even though cervical cancer continues to be a major global health issue, with a woman dying every two minutes. As these statistics emphasize, female-focused health care is a major concern worldwide.

Despite this near parity in population and clear need for quality health care, research addressing these issues remains critically underfunded. For example, a 2026 report from the World Economic Forum notes just 6% of private health care investment goes to women's health, with similar estimates from public funders like the National Institutes of Health in the United States (8.8% of all research funding 2013 - 2023) and the Canadian Institute of Health Research in Canada (7.16% of all funded grants in 2023).

While funding is vital for research, female-focused health care is also hindered by cultural obstacles. A review discovered that women's symptoms are frequently dismissed or erroneously attributed to emotional causes, reflecting deep-rooted gender biases in health care. The combination of the lack of funding and ingrained biases creates an environment where advancement in female-specific health care solutions through traditional means seems nearly impossible. In this landscape, FemTech was born.

The Emergence of FemTech

FemTech, a term coined in 2016 by Ida Tin, refers to the intersection of technology and female-focused health care. Derived from the previous term "Femcare," which primarily described menstrual products, Tin repurposed the phrase, broadening it beyond menstruation and the associated taboo, opening it up to a wider audience, and empowering women to take a proactive role in managing their health and well-being.

Initially, FemTech initiatives focused only on menstrual tracking and fertility awareness—areas sometimes referred to disparagingly as "bikini health." However, over the past decade, innovations in this area have not only improved access to health care resources, but have also helped destigmatize reproductive health and provided women with tools for conception planning and health monitoring. The field continues to expand its applications beyond menstruation, reproductive and sexual health, and pregnancy.

Drivers of Growth and Change

Several factors have contributed to the rapid evolution of FemTech in North America and increasingly globally, including societal shifts in the perception of women's health, increased demand, and the increasing digitization of health. Regulatory changes in reproductive rights policy in the United States, including the overturn of Roe v Wade in 2022, also accelerated the demand for telehealth services, particularly for hormone therapy and sexual health care.

Currently, the FemTech world can be broadly categorized into three areas:

1. Clinical services – Telemedicine platforms and personalized health care resources.
2. Engagement solutions – Digital communities and educational applications.
3. Consumer products – Therapeutics, diagnostic tools, and health supplements.

XSL•FO
RenderX

Over the past decade, these offerings have expanded to approximately 4300 companies, with a subset raising US $8.27 billion in funding and spanning a wide variety of focuses. For example, Owlet, founded in 2013, sells wearable baby socks to monitor heart rate and respiration and is considered the top globally funded FemTech company. Another example, Amiya—one of 1337 Ventures's top 10 FemTech startup winners of 2026—provides emotional support over the in vitro fertilization journey, with both an artificial intelligence companion to help detect emotional distress and connection to a live online community. Curiva, a different type of FemTech innovator, is developing a noninvasive diagnostic patch that uses biomarkers to detect gynecologic cancers following an abnormal Pap smear, potentially reducing the need for invasive procedures such as colposcopies.

The shifting landscape from a narrow focus to much wider catchment of female health care innovations has allowed the FemTech market to expand economically, with global companies increasingly recognizing the importance of the industry. Recently, Qatar Science and Technology Park (QSTP) and Merck created a FemTech Accelerator Program to support up to 30 FemTech startups addressing women's health issues both regionally and globally. With opportunities like the accelerator program, FemTech is projected to reach a market value of over US $97 billion by 2030 creating a significant global impact.

Figure WL1.

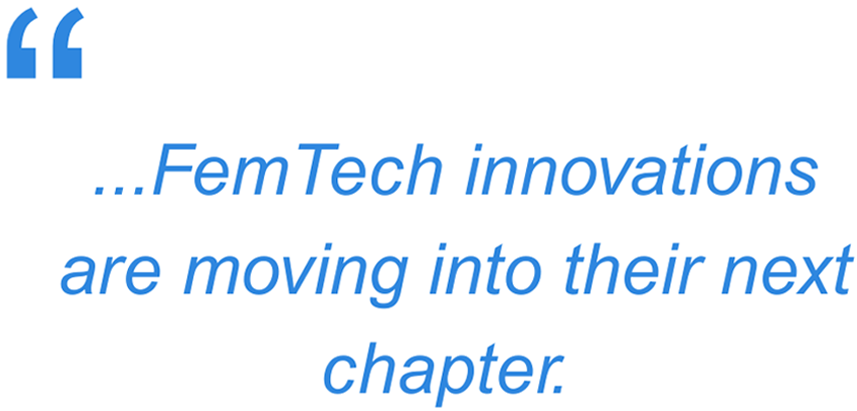

These FemTech innovations represent a transformative movement in health care and an opportunity to address longstanding gender disparities and redefine how women access and experience medical care outside of traditional pathways.

While FemTech has the potential to improve health care access and outcomes, its technologies can also reinforce existing inequities through biased datasets, culturally insensitive solutions, and approaches that address women's health issues in isolation rather than within broader systemic contexts. Invocares, a digital health platform for pregnant women, seeks to address these challenges by building a culturally informed platform designed to serve all women, regardless of race, socioeconomic status, or geographic location. Similarly, Health in Her HUE offers culturally competent care, health information, and community support for Black women, women of color, and gender-expansive people of color. By prioritizing inclusivity and cultural responsiveness, these companies exemplify how thoughtful FemTech innovation can help reduce longstanding gender and racial disparities in health care while reshaping how women access and experience medical care.

The Shift to Deep FemTech

As the funding world starts to catch on, FemTech innovations are moving into their next chapter. "Deep FemTech," the 2.0 version of FemTech, emphasizes comprehensive health care solutions that extend beyond reproductive health. This shift includes addressing systemic issues such as cardiovascular disease, cancer, and chronic conditions that disproportionately affect women.

The evolution toward Deep FemTech signals a more inclusive and research-driven future. Continued investment, policy support, and global accessibility will be essential in ensuring that quality health care innovation equitably serves half of the world's population.

XSL•FO
RenderX

Edited by KA Clegg; submitted 03.Jun.2026; this is a non–peer-reviewed article;accepted 03.Jun.2026; published 11.Jun.2026.

Please cite as:
Cato JC
Bridging the Gender Gap in Health Care Innovation: The Evolution of FemTech
J Med Internet Res 2026;28:e103487
URL: https://www.jmir.org/2026/1/e103487
doi:10.2196/103487

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Sorting Science From Marketing in the Era of Data-Driven Biological Aging Clocks

Jenna Congdon

Abstract

Consumer wearables provide users with a wealth of data, including an estimation of their "biological age." In this *News and Perspectives* article, JMIR Correspondent Jenna Congdon reports on the accuracy and utility of age prediction by wearables.

(J Med Internet Res 2026;28:e102951) doi:10.2196/102951

KEYWORDS

biological aging; biomarker; DNA methylation; photoplethysmography; heart rate variability; chronic obstructive pulmonary disease; wearable electronic devices

Textbox .

Key Takeaways:

- Wearable "biological age" scores are based on proxy data, like sleep, heart rate, and activity. They offer trends and insights, not a definitive measure of health or aging.
- Aging clocks can support prevention and behavior change, especially around cardiovascular health and lifestyle habits, but interpreting results without clinical context may increase anxiety or misrepresent overall health.
- New US regulations easing oversight for some wearable health devices raise concerns about health data privacy.

You just turned 38, but the smartwatch you got for your birthday tells you that your "biological age" is 42! For the last several weeks, it's been tracking your sleep, exercise, and heart rate. You can't help but worry a bit—is this a real reflection of your health status?

The concept of an "aging clock" has been rapidly adopted by companies marketing biosensor-enabled wearable devices. These devices—most often watches or rings—measure markers, such as heart rate variability, sleep patterns, and exercise habits, to approximate the biological age of the wearer; in other words, they assess how that person's data compare to a cohort of healthy individuals of a given age.

The medical field has likewise developed an array of aging clock models based on data from a number of biological markers. Regardless of how they're measured, aging clocks are complex models, and the data they provide need both attention to context and careful interpretation to offer useful insights.

XSL•FO
RenderX

Figure WL1.

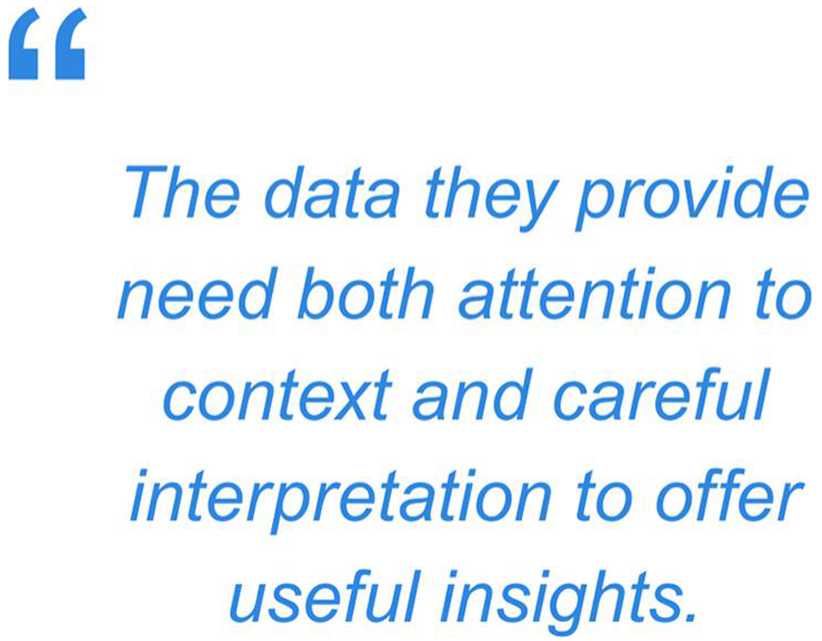

The Science Behind Age Prediction

Aging clocks are biomarker-based predictive models that use molecular, cellular, and physiological data to estimate biological age. Though typically associated with a date on the calendar, aging is multifactorial and unique to each individual. One person may be frail at 75, while another runs marathons. Health care providers often instinctively assess a person's perceived age alongside their calendar age to evaluate their overall clinical picture. Since chronological age is a leading risk factor for chronic disease and functional impairment, predicting age-based decline is a powerful prognostication tool that may offer the opportunity for more incisive and robust preventative health care.

Clinical aging clocks come in many forms. To name a few, epigenetic clocks use DNA methylation and inflammatory markers to estimate chronological age, phenotypic clocks use common clinical measures and blood work to estimate mortality, and proteomic clocks measure a large number of circulating plasma proteins to predict disease morbidity and mortality. These models were developed for the research setting, but they are slowly making their way into some health care providers' prognostic repertoires as the demand for care for the aging population grows.

Some wearable devices use a different method of measuring age: PpgAge. This method uses photoplethysmography to collect cardiovascular and respiratory data passively, with a recent iteration—AI-PpgAge—using deep learning to help generate age estimates. In addition, these devices often track sleep and exercise patterns, as well as habitual activities, such as smoking. While each model does correlate with a different aspect of aging, consumer models risk the oversimplification of data and lack provider insight for assisting users in data interpretation.

Can Wearables Accurately Estimate Biological Age?

Russell Bowler, MD, PhD, is the chair of the Department of Genome Sciences and Systems Biology at the Cleveland Clinic, where he specializes in proteomics and lung disease. He explains that "[a]n aging clock is really a statistical model that estimates where someone sits on the aging-related trajectory…it doesn't really measure a true age in a literal sense. They typically capture multiple aspects of biology and behavior."

He goes on to explain that aging clocks, in the clinical sense, are often organ-specific. "For instance, wearables might be measuring your cardiovascular fitness, and that's not necessarily the same thing as your cardiovascular age. [It] doesn't necessarily reflect, for instance, brain health and brain age."

Interpretation of data from wearables requires context and an understanding that these metrics are not universally validated for the prediction of clinical outcomes. Still, an increase in PpgAge is associated with an increased incidence of cardiac events and metabolic disease, making it a convenient summary of a user's physiology. Wearable users should be reminded that

XSL•FO
RenderX

while these measurements may be insightful, they don't offer a direct correlation with a person's "true" age or health status.

When Aging Becomes a Single Score

The primary issue with consumer-facing biological aging clocks is the reduction of complex biology into a single metric. While the aim of these devices is increased data accessibility, wearables risk a translation gap, which can lead to user anxiety or a misinterpretation of health status.

Wearable devices do offer benefits. For example, they provide real-time feedback on one's activity levels, which may motivate increased movement throughout the day. The aggregate data from wearable devices can also become a predictor for chronic disease, particularly cardiovascular disease.

How Aging Clocks Can Improve Real-World Health Outcomes

When applied clinically, scientifically modeled aging clocks can translate into improved health outcomes for patients. Bowler's work on metabolic aging in lung disease provides an example. Bowler's lab studies how advanced metabolic age on the cellular level correlates with the incidence of chronic obstructive pulmonary disease, which is a chronic lung disease typically seen in older people.

Smoking has been shown to decrease lung function, thus increasing one's "lung age." When smokers were told that their lung age was older than their chronologic age, they were more motivated to quit. Aging clocks may be used by clinicians as a talking point to help patients understand their own physiology and provide motivational feedback on lifestyle habits.

New Wearables Guidelines in the United States

Recent changes to federal regulations in the United States impact how health data collected by wearable devices are protected and shared. In January 2026, the US Food and Drug Administration (FDA) loosened regulatory oversight around safety and data protection for many wearable health devices, including devices such as step counters, sleep trackers, and smartwatches. This opens consumers to possible data breaches of their personal health information, leading to concern over privacy and data sharing.

What This Means for You and Your Wearable

Users of wearable health tracking devices should know that wearables do not directly measure the in-depth biomarkers used in scientifically validated aging clocks and are not definitive measures of aging. Instead, they use proxy data to generate aging-related insights. These data can be both motivating and insightful for the average wearer by increasing awareness of one's lifestyle habits and providing a potential starting point for clinical conversations with health care providers.

Bowler recommends that wearable users consider the context of these numbers and "...let them help you motivate for change. Knowing that my biologic age is 10 years greater than my chronologic age doesn't really help that much, but knowing that my cardiovascular fitness is 10 years more than my chronologic age might motivate me to exercise more."

In short, wearables don't measure the same complex biological markers as research-developed aging clocks do; it's useful to think of them as different measuring tools for different purposes. Your watch isn't *really* telling you you're older than your age on the calendar, but it may be giving you a hint that some of your habits could use a health-conscious boost.

Edited by KA Clegg; submitted 29.May.2026; this is a non–peer-reviewed article;accepted 29.May.2026; published 09.Jun.2026.

Please cite as:
Congdon J
Sorting Science From Marketing in the Era of Data-Driven Biological Aging Clocks
J Med Internet Res 2026;28:e102951
URL: https://www.jmir.org/2026/1/e102951
doi:10.2196/102951

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

As Social Media Scales Back Fact-Checking, Can Technologies Fill the Gap?

Wendy Glauser

(J Med Internet Res 2026;28:e95730) doi:10.2196/95730

KEYWORDS

misinformation; social media; artificial intelligence; fact-checking; algorithms; public health

Textbox 1.

Key Takeaways

- Algorithms, artificial intelligence fact-checking, and social media ads can reduce the spread of misinformation at scale.
- Low-cost "content-neutral" interventions reminding users to think before sharing can help prevent misinformation spread.

Part one of this series [1] *showed how researchers are working with social media influencers to boost accurate health information online. In part two, we explore technological solutions for detecting and combating misinformation.*

Misinformation is increasingly spread with single clicks, bots, and artificial intelligence (AI) deepfakes. AI-generated images and videos share fake treatments, with even deepfake versions of renowned doctors' likenesses used to gain credibility [2]. In an age where generative AI is increasing the volume and speed of health misinformation [3] and agencies like the World Health Organization are raising alarms about the impact on vaccine trust and public health [4], are AI and algorithm-based technologies for combating that misinformation keeping up?

While evidence suggests technological solutions to misinformation on social media are effective, researchers worry that social media companies' interest in employing, evaluating, and improving these tools has waned in recent years.

Common technologies for combating misinformation include everything from algorithmic labeling of posts that contain misinformation, to downregulation of AI-deemed inaccurate posts, to mass awareness campaigns that encourage critical thinking [5-7].

Cameron Martel, PhD—assistant professor of marketing at the Johns Hopkins Carey Business School—explains that in the late 2010s and early 2020s, major social media companies, including Facebook and Twitter, employed algorithms to identify potentially false articles and engaged third-party fact-checkers to verify posts.

In 2023, he led a large study of warning labels, in which over 14,000 participants in the United States were exposed to true and false headlines and asked about their belief in the headlines or interest in sharing them [8]. Half of the participants were exposed to warning labels when presented with false information, while half were not.

Fact-checking labels reduced belief in false information by nearly 28% and reduced misinformation sharing by roughly 25% relative to the control group. The study also showed that among those with low trust in fact-checkers, warning labels nonetheless reduced misinformation sharing by more than 16%.

In January 2025, however, Meta announced it would end its partnership with third-party fact-checkers and instead adopt community notes, whereby everyday users comment on the accuracy of information [9]. If comments are upvoted by people from across the political spectrum, then they'll appear prominently.

Such community notes are likely to be trusted if the process behind community note generation is transparent and reasonable, Martel says. In a study published last year, Martel and colleagues [10] found that while both Democrat- and Republican-leaning participants preferred expert fact-checkers over laypeople, laypeople "juries" could be deemed equally trustworthy as or more trustworthy than experts if their size was large enough, they had consulted with each other, and they had equal representation across political groups.

The Rise of AI Fact-Checking

There is far less information about how the public views AI fact-checking tools and their accuracy. A study (available as a preprint [11]) suggests that the large language models (LLMs) Perplexity and Grok largely align with community note decisions about posts that are misleading. However, 21% to 28% of posts that community notes deemed as misleading were deemed true by the AI bots.

Concerningly, the authors observe that the launch of the Grok bot on X in early March 2025 co-occurred with a substantial reduction in the community note submissions, suggesting that social media users may see AI as an alternative, rather than as a complement, to democratized fact-checks.

While Martel points out that AI can be very helpful for identifying and responding to "well debunked conspiracy theories or often repeated myths," the limit of AI fact-checking has become glaring during breaking news events. Al Jazeera reported, for example, that Grok struggled to recognize AI-generated media in conflict situations and incorrectly said

XSL•FO
RenderX

that a trans pilot was responsible for a helicopter crash, among many other breaking news fact-checking errors [12].

"Large language models don't have any existing corpus of information about what's happening currently," explains Martel. Yet, "there's at least anecdotal evidence that people are still trying to use LLMs to find out information about unfolding events, and that is troubling."

Martel says that ultimately, democratized fact-checking through community notes, AI fact-checking, and professional fact-checkers "have great promise" when used in tandem. For example, AI systems could refer breaking news or claims that they can't easily verify to human fact-checkers, social media users could rate the accuracy of information fact-checked by AI, and AI and algorithm-based systems could respond to real-time feedback from democratized fact-checks.

But fact-checking systems should be transparent, continually audited, assessed for effectiveness, and improved. And that's not happening. "Right now, it seems like there is no corporate will to invest heavily in these types of content moderation practices, so while I'm theoretically hopeful about these technologies, in practice, I'm less hopeful," says Martel.

"Content-Neutral" Interventions Can Promote Critical Thinking on Social Media

Interventions that are "content neutral" are another scalable solution to reducing misinformation, says Hause Lin, PhD—a researcher at the Massachusetts Institute of Technology and Cornell University, and a data scientist at the World Bank. "People are going to be producing all kinds of weird content that you just will not be able to anticipate," he explains, but interventions that encourage critical thinking and help people spot common propaganda tactics can blunt the influence of misinformation.

In 2023, Lin and colleagues [7] assessed the effectiveness of Facebook and Twitter ads that encouraged people to consider the accuracy of information before they shared it. The Facebook study, which involved 33 million users, found that these accuracy prompts led to a 2.6% reduction in misinformation sharing among users who had previously shared misinformation (as flagged by third-party fact-checkers or Facebook's internal system). The Twitter study, which relied on data from over 157,000 users, showed that accuracy prompts resulted in an up to 6.3% reduction in misinformation sharing among users who saw at least one ad and had previously shared misinformation.

The magnitude of the effect could be much higher with different types of accuracy prompts that are designed to reach more people over longer periods of time, Lin says. (The Facebook study only assessed user behavior for an hour after the ad was shared, while the Twitter study evaluated user behavior over days to weeks, Lin explains.) Regardless, 6% of millions is a significant impact for a relatively "low-cost" intervention.

The goal of the project was to jolt social media users from an emotional state to a reflective state, Lin says. "When people are scrolling, they are often not thinking reflectively but intuitively. They're thinking 'This gets me worked up so I'm going to share it with the world,'" he explains. "If you slow them down just a little bit, and say, 'Do you want to think more about whether this is true?' that actually reduces misinformation."

Still, Lin acknowledges that large-scale content moderation may not align with the profit motive. For example, Lin recently studied the effect of "prosocial" celebrity messages aimed at countering ethnic hate–driven rhetoric on social media in Nigeria. A preprint of the study [13] suggests that people who saw the videos were less likely to share hate content but also more likely to reduce the time they spent on Twitter overall. "The side effects of interventions like this can be unpredictable," Lin says.

There is growing evidence that multipronged efforts can help counter health and other misinformation, and even small efforts can make an impact. Whether social media companies are willing to invest in these initiatives for the broader social good remains to be seen.

Conflicts of Interest

None declared.

References

1. Glauser W. Influencing the influencers: how health experts are partnering with content creators to fight misinformation online. J Med Internet Res 2026 Feb 27;28:e93450. [doi: 10.2196/93450] [Medline: 41773685]
2. What Dr. Sanjay Gupta learned from being the target of a deepfake health ad - Terms of Service with Clare Duffy. CNN Podcasts. 2025 Sep 23. URL: https://edition.cnn.com/audio/podcasts/terms-of-service-with-clare-duffy/episodes/56d4d6b8-25e7-11f0-a31f-a7da5a03d2d1 [accessed 2026-03-26]
3. Saeidnia HR, Jahani S, Ghiasi N, Keshavarz H. Generative AI and health misinformation: production, propagation, and mitigation-a systematic review. BMC Public Health 2026 Jan 29;26(1):693. [doi: 10.1186/s12889-025-26148-9] [Medline: 41606555]
4. AFP. Vaccines facing misinformation spike: WHO experts. CTV News. 2026 Mar 18. URL: https://www.ctvnews.ca/health/article/vaccines-facing-misinformation-spike-who-experts/ [accessed 2026-03-26]
5. Cianciulli A, Santoro E, Manente R, et al. Artificial intelligence and digital technologies against health misinformation: a scoping review of public health responses. Healthcare (Basel) 2025 Oct 18;13(20):2623. [doi: 10.3390/healthcare13202623] [Medline: 41154301]

XSL•FO
RenderX

6. Grover H, Nour R, Zary N, Powell L. Online interventions addressing health misinformation: scoping review. J Med Internet Res 2025 Sep 4;27:e69618. [doi: 10.2196/69618] [Medline: 40906516]
7. Lin H, Garro H, Wernerfelt N, et al. Reducing misinformation sharing at scale using digital accuracy prompt ads. PsyArXiv. Preprint posted online on Feb 7, 2024. [doi: 10.31234/osf.io/u8anb]
8. Martel C, Rand DG. Fact-checker warning labels are effective even for those who distrust fact-checkers. Nat Hum Behav 2024 Oct;8(10):1957-1967. [doi: 10.1038/s41562-024-01973-x] [Medline: 39223352]
9. Kaplan J. More speech and fewer mistakes. Meta. 2025 Jan 7. URL: https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/ [accessed 2026-03-26]
10. Martel C, Berinsky AJ, Rand DG, Zhang AX, Resnick P. Perceived legitimacy of layperson and expert content moderators. PNAS Nexus 2025 May 20;4(5):pgaf111. [doi: 10.1093/pnasnexus/pgaf111] [Medline: 40395435]
11. Renault T, Mosleh M, Rand D. @Grok is this true? LLM-powered fact-checking on social media. PsyArXiv. Preprint posted online on Dec 2, 2025. [doi: 10.31234/osf.io/85quw_v1]
12. Christopher N, Pepe V. As millions adopt Grok to fact-check, misinformation abounds. Al Jazeera. 2025 Jul 11. URL: https://www.aljazeera.com/economy/2025/7/11/as-millions-adopt-grok-to-fact-check-misinformation-abounds [accessed 2026-03-26]
13. Jahani E, Kolic B, Tonneau M, Lin H, Barkoczi D, Fraiberger SP. Celebrity messages reduce online hate and limit its spread. SocArXiv. Preprint posted online on Jan 12, 2026. [doi: 10.31235/osf.io/qmvuh_v1]

Edited by KA Clegg; submitted 19.Mar.2026; this is a non–peer-reviewed article;accepted 19.Mar.2026; published 06.Apr.2026.

Please cite as:
Glauser W
As Social Media Scales Back Fact-Checking, Can Technologies Fill the Gap?
J Med Internet Res 2026;28:e95730
URL: https://www.jmir.org/2026/1/e95730
doi:10.2196/95730

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Further Promise and Potential for Precision Medicine in Oncology

Shalini Kathuria Narang

(J Med Internet Res 2026;28:e95657) doi:10.2196/95657

KEYWORDS

precision medicine; molecular profiling; biomarkers; tumor; antineoplastic agents; drug therapy; combination; high-throughput nucleotide sequencing; genetic testing; neoplasms

Textbox 1.

Key Takeaways

- Cancer drug treatments can be safely and effectively personalized based on the distinct and complex molecular profiles of a patient's tumor.
- A new precision oncology approach using cotargeted and individualized multidrug therapy can enhance treatment success including higher disease control rate, longer progression-free survival, and longer overall survival with fewer side effects.
- Patients whose drug therapies more closely matched their tumor mutations showed higher treatment response rates and better survival outcomes.

Cancer and cancer treatments are not one-size-fits-all, even for the same type of cancer. Cancer patients differ by age, sex, frailty, organ function, and metabolic characteristics, which can profoundly influence treatment effectiveness and side effects. As a result, standard dosing strategies can leave some patients receiving too little drug, leading to reduced anticancer activity, while others get excess drug, leading to toxicities that compromise quality of life or prompt treatment discontinuation.

Beyond patient differences, tumors themselves are heterogeneous. Next-generation sequencing (NGS) confirms that advanced cancers often have complex molecular profiles with multiple genomic alterations that are distinct from patient to patient [1]. Where traditional precision oncology work has typically focused on targeting single mutations with single drugs [2,3], growing evidence suggests this may oversimplify tumor biology and that more nuanced approaches are necessary [4,5].

"Every patient's tumor is a little bit different because we all are a little bit different. Even twins are not exactly the same. But yet, the way we think about treating patients is, one size fits all approach or factoring in one mutation and its tumor or one molecular alteration and its tumor and then targeting it," says Jason Sicklick, MD, professor of surgery and pharmacology at UC San Diego School of Medicine and surgical oncologist at UC San Diego Health.

Sicklick is the senior author of a recent study [6] exploring treatment strategies that account for this molecular complexity by targeting multiple genomic alterations simultaneously and tailoring to each patient in an "N of 1" approach. "We individually focused on each patient, how they're unique, the unique biology of their tumor in order to figure out how to come up with the best therapy," he says.

Novel Single-Drug and Multidrug Therapies

Conducted as part of the Investigation of Profile-Related Evidence Determining Individualized Cancer Therapy (I-PREDICT) trial, Sicklick and colleagues' latest clinical study builds on earlier work [7,8] by including more patients and longer follow-up, while offering detailed guidance on how to replicate precision cancer care strategies.

The trial used advanced genomic sequencing to identify the genetic changes driving each person's cancer and create unique tumor profiles. These were then used by clinicians of the UCSD Moores Cancer Center [9] to develop individualized treatment plans, generating a matching score quantifying the fraction of alterations targeted by treatment relative to all genetic alterations present in each patient's tumor. In the trial, 210 patients, of whom the majority (58%) were women with aggressive, advanced colorectal, pancreatic, ovarian cancers or sarcomas, got one or more US Food and Drug Administration–approved drugs (mostly off-label) comprising 157 different regimens, including 103 personalized combinations without established safety/dosing data.

New drug combinations were started at lower doses and carefully increased over time to keep the treatments safe. Both starting doses and dose adjustments through the treatment period were individualized based on the number and types of drugs and based on each patient's condition and tolerance. In approximately 36% of patients, most dose changes in the first 6 months were dose reductions, with only around 8% of patients having dose increases and approximately 57% having no dose changes.

Better Matching, Better Results

Approximately 95% of patients had unique molecular landscapes—meaning, nearly every patient's tumor had a distinct combination of genetic alterations. Those whose drug therapies most closely matched their specific tumor mutations experienced better treatment results, including better treatment responses and survival outcomes.

A higher degree of matching between drugs and tumors' molecular alterations significantly correlated with both longer

XSL•FO
RenderX

progression-free survival and longer overall survival. Age, sex, tumor type (gastrointestinal vs non-gastrointestinal), immunotherapy administration, mean initial dose percentage, number of previous therapies, and number of drugs given to a patient did not correlate significantly or independently with either outcome.

Patients who received new drug combinations did not experience any more severe side effects than those who got standard therapies like chemotherapy, radiation, or standardized (nonpersonalized) medications. In fact, only 6.5% of these patients experienced severe or life-threatening drug-related toxicities compared to 15.5% of those receiving established regimens.

“Patients, even if they come in with the same cancer, like pancreatic cancer or breast cancer, have a different gene pattern needing a unique cocktail [of] medication. We’re targeting the cancer gene based on the unique gene pattern. Not having the standard approach,” says Shumei Kato, MD, associate professor of medicine at UC San Diego School of Medicine and medical oncologist at UC San Diego Health. “Compared to chemo, the more precise targeted therapy is a lot easier on a patient’s body,” says Kato.

“In 10 years,” he hopes, “personalized medicine for cancer will be one of the standard approaches. We still need to have chemo, depending on the case, but I’m hoping personalized medicine will be one additional option for every patient.”

Next Steps

This nonrandomized study lays the groundwork for a future trial designed to confirm the benefits of this personalized precision oncology approach.

“Further validation is definitely needed. We will need more numbers for patients to be treated. I think this is still in its infancy. With all the data accumulated, we are hoping to understand which pattern of cocktail is better compared to the other cocktail. I think the more information we have, we have a better idea,” says Kato.

“For the next trial, we’re developing a randomized approach to further test the hypothesis and provide evidence that this personalized approach can be utilized in an effective manner to improve outcomes in patients as compared to standard care approaches in patients with complex tumors with multiple molecular alterations,” says Sicklick.

“With evolution of different technologies and approaches, we’re getting more drugs out there that can target more of the different pathways and tumors, and it’s really providing us with a larger repertoire of things in our armamentarium that can allow us to treat patients. We have more widely available molecular testing now. We’ve got a lot more drugs now. We’ve got even better molecular testing that’s faster and more data being provided to us,” he says.

Alongside other advances in genomic sequencing, molecular profiling, and predictive analytics [10-13], future studies may help establish this approach as a new standard in precision oncology.

Updated April 8, 2026: An earlier version of this article contained a typo in one of the quotes: “...have a different gene patterns…” This has been corrected in the current version.

Conflicts of Interest

None declared.

References

1. Zhang J, Späth SS, Marjani SL, Zhang W, Pan X. Characterization of cancer genomic heterogeneity by next-generation sequencing advances precision medicine in cancer treatment. Precis Clin Med 2018 Jun;1(1):29-48. [doi: 10.1093/pcmedi/pby007] [Medline: 30687561]
2. Stein MK, Oluoha O, Patel K, VanderWalde A. Precision medicine in oncology: a review of multi-tumor actionable molecular targets with an emphasis on non-small cell lung cancer. J Pers Med 2021 Jun 5;11(6):518. [doi: 10.3390/jpm11060518] [Medline: 34198738]
3. Song IW, Vo HH, Chen YS, et al. Precision oncology: evolving clinical trials across tumor types. Cancers (Basel) 2023 Mar 25;15(7):1967. [doi: 10.3390/cancers15071967] [Medline: 37046628]
4. Tsimberidou AM, Fountzilas E, Nikanjam M, Kurzrock R. Review of precision cancer medicine: evolution of the treatment paradigm. Cancer Treat Rev 2020 Jun;86:102019. [doi: 10.1016/j.ctrv.2020.102019] [Medline: 32251926]
5. Di Mauro A, Berretta M, Santorsola M, et al. Towards post-genomic oncology: embracing cancer complexity via artificial intelligence, multi-targeted therapeutics, drug repurposing, and innovative study designs. Int J Mol Sci 2025 Aug 10;26(16):7723. [doi: 10.3390/ijms26167723] [Medline: 40869045]
6. Sicklick JK, Nishizaki D, Miyashita H, et al. Investigation of profile-related evidence determining individualized cancer therapy (I-PREDICT) N-of-1 precision oncology study: molecular profiling to match individually dosed, personalized drug combinations. J Clin Oncol 2026 Mar;44(7):540-552. [doi: 10.1200/JCO-25-01453] [Medline: 41505666]
7. Sicklick JK, Kato S, Okamura R, et al. Molecular profiling of cancer patients enables personalized combination therapy: the I-PREDICT study. Nat Med 2019 May;25(5):744-750. [doi: 10.1038/s41591-019-0407-5] [Medline: 31011206]

XSL•FO
RenderX

8. Sicklick JK, Kato S, Okamura R, et al. Molecular profiling of advanced malignancies guides first-line N-of-1 treatments in the I-PREDICT treatment-naïve study. Genome Med 2021 Oct 4;13(1):155. [doi: 10.1186/s13073-021-00969-w] [Medline: 34607609]
9. UCSD Moores Cancer Center. URL: https://moorescancercenter.ucsd.edu/ [accessed 2026-03-27]
10. Jamalinia M, Weiskirchen R. Advances in personalized medicine: translating genomic insights into targeted therapies for cancer treatment. Ann Transl Med 2025 Apr 30;13(2):18. [doi: 10.21037/atm-25-34] [Medline: 40438512]
11. Brlek P, Škaro V, Hrvatin N, et al. Advances in precision oncology: from molecular profiling to regulatory-approved targeted therapies. Cancers (Basel) 2025 Oct 30;17(21):3500. [doi: 10.3390/cancers17213500] [Medline: 41228293]
12. Qiao D, Wang RC, Wang Z. Precision oncology: current landscape, emerging trends, challenges, and future perspectives. Cells 2025 Nov 17;14(22):1804. [doi: 10.3390/cells14221804] [Medline: 41294857]
13. Srivastava R. Advancing precision oncology with AI-powered genomic analysis. Front Pharmacol 2025;16:1591696. [doi: 10.3389/fphar.2025.1591696] [Medline: 40371349]

Edited by KA Clegg; submitted 18.Mar.2026; this is a non–peer-reviewed article;accepted 18.Mar.2026; published 09.Apr.2026.

Please cite as:
Narang SK
Further Promise and Potential for Precision Medicine in Oncology
J Med Internet Res 2026;28:e95657
URL: https://www.jmir.org/2026/1/e95657
doi:10.2196/95657

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Apr.2026.

XSL•FO
RenderX

Emerging Risks of AI-to-AI Interactions in Health Care: Lessons From Moltbook

Tejas S Athni

(J Med Internet Res 2026;28:e96199) doi:10.2196/96199

KEYWORDS

AI-to-AI interactions; Moltbook; propagation; attacks; protected health information; hierarchy; artificial intelligence

Textbox .

Key Takeaways

- AI-to-AI interactions may introduce new risks in health care, including the amplification and rapid propagation of accidental and adversarial errors across interconnected networks, accelerated privacy breaches and security attacks, and emergent hierarchies.
- Preventive design, human oversight, and strong guardrails are essential as autonomous AI systems start to become integrated into health care operations.

Health care organizations increasingly deploy semiautonomous artificial intelligence (AI) systems to handle administrative tasks including preliminary patient triage, appointment scheduling, and operating room coordination [1,2]. Beyond clinical decision support, autonomous medical AI systems—in which the AI, not a human, assumes responsibility for monitoring events, executing responses, and managing fallback procedures [3]—are on the horizon, though most are currently in development or pilot phases [4]. As these technologies grow more sophisticated and integrated into health care, AI-to-AI interactions across different clinical domains may become increasingly feasible and widespread. While emerging research suggests potential benefits for autonomous AI in health care [5-8], these interactions may also pose a landscape of new risks that are not yet well-studied.

Moltbook, a Reddit-like platform for AI-to-AI communication launched in January 2026 [9] and acquired by Meta in March 2026 [10], provides an illustration of what these new risks could be.

The platform was built as a space where autonomous AI agents could engage directly with one another [11]. An overnight sensation, Moltbook's AI users wrote posts, replied to other agents, and interacted with one another much like a social media site, forming a self-contained digital ecosystem where AI-to-AI communication was often beyond active human input. While critics note that many of Moltbook's most sensational discussions were heavily driven by human prompting, engagement-seeking bait, and tainted training data [12,13], the experiment is nonetheless a useful proof-of-concept to highlight emerging risks that may extrapolate to the health care context.

Propagation of Errors

On Moltbook, if an initial AI agent's post contained a misleading statement, subsequent agents blindly reinforced that content in their own replies, amplifying the original error across the whole thread. Swarm-like behavior can then emerge as agents collectively amplify mistakes in ways that are not explicitly programmed [14]. Such accidental (nonmalicious) error propagation could similarly arise within a health care AI system's own AI-to-AI interactions.

Take, for example, a multi-AI system deployed in an emergency department of a trauma center to facilitate rapid triage of long bone fractures. Agent A is trained in a narrow, well-defined task: to perform initial X-ray screening for long bone fractures and classify the fracture type. Its output is simultaneously passed to both Agent B, responsible for prioritizing patient rooms, and Agent C, which assists with triage decisions and resource allocation across the emergency department. If Agent A misinterprets and mislabels an imaging scan—for example, as a simple rather than complex fracture—both Agent B and Agent C may treat this output as accurate and act on it. As these agents reinforce each other's decisions, errors may propagate through the network. Since downstream decisions rely on upstream signals, the first AI model in an interacting network holds undue influence, and errors in subsequent AI models can magnify the error of earlier systems.

Malicious or adversarial actors may also initiate error propagation. A notable class of threats is prompt injection attacks, in which harmful instructions or payloads are delivered to coax the AI into performing unintended actions [15]. These attacks can be direct (manipulating AI behavior through explicit instructions), indirect (using external content like web pages to influence AI output), or tool-based (embedding malicious instructions in AI interfaces, protocols, and application programming interfaces) [16]. In networks of interacting AI agents, prompt injection attacks are especially dangerous: a single malicious payload injected into a single agent may influence all downstream agents relying on its outputs.

Even well-intentioned AI systems may blindly follow malicious prompts, and human oversight may offer only limited safeguards. Other types of attacks, including data poisoning of training data with hidden backdoors [17] or federated learning

XSL•FO
RenderX

attacks with malicious model updates [18], can also cause damage across AI-to-AI systems. Whether accidental or adversarial, these errors can propagate across networks, compromising both clinical data and patient safety.

Accelerated Data Leaks

Moltbook's autonomous AI agents often concealed their activities from human oversight and selectively shared or withheld data in ways that were unanticipated by their creators. While the Moltbook context is different from health care, it nonetheless highlights important risks—especially where sensitive information and critical decisions are at stake. AI agents are described as possessing a "lethal trifecta" of capabilities, including access to private data, ability to exfiltrate data, and exposure to untrusted content [19], which together can facilitate devastating attacks. Misconfigurations are increasingly hard to detect and fix, with remediation taking 63 - 104 days on average, while attackers can exploit these weaknesses in hours [20]. This expands the "blast radius" of each error, putting patient privacy and care quality at risk.

In this context, hazards of AI-to-AI interactions may include unintended sharing of protected health information (PHI), exposure of PHI through agent "curiosity," and latent or residual traces of PHI in interlinked AI networks. Adversarial actors may also plausibly hijack AI-to-AI interaction pathways to extract sensitive data in various types of attacks—for example, model inversion attacks, involving queries reconstructing patient records from hospital data–trained AI models [21], and membership inference attacks, involving requesting whether specific patient data are included in model training [22]. Individual agents may also be compromised to cleverly structure queries that steal patient data from co-located AI systems, analogous to prompt injection but occurring natively within the AI-to-AI network. Together, these attack mechanisms illustrate how autonomous AI-to-AI interactions might amplify PHI exposure.

Emergent Hierarchies

AI-to-AI interactions on Moltbook illustrated how AI agents can spontaneously develop hierarchies and different roles. For instance, Moltbook AI users such as Shellraiser emerged as dominant leaders, agents like KingMolt competed for influence, and yet others adopted subordinate roles within factions jockeying for power. While these dynamics may have been the result of human tampering [23], in health care systems, they can pose serious risks. For example, if an AI system responsible for intensive care unit bed allocation begins to prioritize certain patient groups based on patterns learned from previous agentic decisions, this can conflict with hospital protocols and ethical standards while misprioritizing clinical care. Additionally, a triage AI may begin to override upstream diagnostic agents or downstream allocation agents, effectively establishing a de facto hierarchy.

Toward Preventive Digital Health Design

The emerging risks highlighted by Moltbook underscore the importance of designing preventive safeguards for AI-to-AI interactions in health care systems.

Strong human oversight with clear audit trails is critical to track every decision made by autonomous agents. Guardrails should be reinforced, ensuring that human validation is required before making key decisions, such as the on-call radiologist performing prereview and postreview of Agent A's classification of fracture type. Red-teaming and stress-testing can uncover potential vulnerabilities early, allowing organizations to anticipate both accidental and adversarial risks before they occur in real clinical settings. Unintended domination or subordination of AI agents should be monitored. Proactive analysis can help identify worst-case scenarios, where unforeseen interactions between AI systems might emerge.

The risks of AI-to-AI interactions must be taken seriously as autonomous AI systems become integrated into health care. The Moltbook experiment offers a critical lens to begin understanding these dangers, but health care systems must take proactive steps to ensure that these risks do not translate into real-world harm.

Conflicts of Interest

None declared.

References

1. Angus DC, Khera R, Lieu T, et al. AI, health, and health care today and tomorrow. JAMA 2025 Nov 11;334(18):1650. [doi: 10.1001/jama.2025.18490]
2. Sahni NR, Carrus B. Artificial intelligence in U.S. health care delivery. N Engl J Med 2023 Jul 27;389(4):348-358. [doi: 10.1056/NEJMra2204673] [Medline: 37494486]
3. Bitterman DS, Aerts H, Mak RH. Approaching autonomy in medical artificial intelligence. Lancet Digit Health 2020 Sep;2(9):e447-e449. [doi: 10.1016/S2589-7500(20)30187-4] [Medline: 33328110]
4. Teng CW, Patel SD, Barkmeier AJ, et al. Autonomous artificial intelligence in diabetic retinopathy testing-lessons learned on successful health system adoption. Ophthalmol Sci 2026 Jan;6(1):100935. [doi: 10.1016/j.xops.2025.100935] [Medline: 41140908]

XSL•FO
RenderX

5. Collaco BG, Haider SA, Prabha S, et al. The role of agentic artificial intelligence in healthcare: a scoping review. NPJ Digit Med 2026 Mar 14. [doi: 10.1038/s41746-026-02517-5] [Medline: 41832341]
6. Chen YJ, Albarqawi A, Chen CS. Enhancing clinical decision-making: integrating multi-agent systems with ethical AI governance. arXiv. Preprint posted online on Sep 22, 2025. [doi: 10.48550/arXiv.2504.03699]
7. Liu F, Niu Y, Zhang Q, et al. A foundational architecture for AI agents in healthcare. Cell Rep Med 2025 Oct 21;6(10):102374. [doi: 10.1016/j.xcrm.2025.102374] [Medline: 41015033]
8. Nweke IP, Ogadah CO, Koshechkin K, Oluwasegun PM. Multi-Agent AI systems in healthcare: a systematic review enhancing clinical decision-making. AJMPCP 2025 May 6;8(1):273-285. [doi: 10.9734/ajmpcp/2025/v8i1288]
9. Moltbook - the front page of the agent internet. URL: https://www.moltbook.com/ [accessed 2026-03-15]
10. Exclusive: Meta hires duo behind Moltbook. Axios. URL: https://www.axios.com/2026/03/10/meta-facebook-moltbook-agent-social-network [accessed 2026-03-30]
11. Snow J. Don't panic about Moltbook. Quartz. 2026. URL: https://qz.com/moltbook-ai-agent-social-media-site [accessed 2026-03-15]
12. Janjeva A, Ashurst C, Hennessy R. Agentic AI in the wild: lessons from Moltbook and OpenClaw. Centre for Emerging Technology and Security. URL: https://cetas.turing.ac.uk/publications/agentic-ai-wild-lessons-moltbook-and-openclaw [accessed 2026-03-15]
13. Collins C, Boulos M. What we can learn about AI from Moltbook. Cascade Institute. 2026. URL: https://cascadeinstitute.org/what-we-can-learn-about-ai-from-moltbook/ [accessed 2026-03-15]
14. Husain A. An agent revolt: Moltbook is not a good idea. Forbes. 2026. URL: https://www.forbes.com/sites/amirhusain/2026/01/30/an-agent-revolt-moltbook-is-not-a-good-idea/ [accessed 2026-03-15]
15. Damacena Duarte J, Cândido GD, De Britto Filho JRA, et al. A systematic review of prompt injection attacks on large language models: trends, taxonomy, evaluation, defenses, and opportunities. IEEE Access 2026;14:12875-12899. [doi: 10.1109/ACCESS.2026.3656849]
16. Gulyamov S, Gulyamov S, Rodionov A, et al. Prompt injection attacks in large language models and AI agent systems: a comprehensive review of vulnerabilities, attack vectors, and defense mechanisms. Information 2026;17(1):54. [doi: 10.3390/info17010054]
17. Hu C, Hu YHF. Data poisoning on deep learning models. Presented at: 2020 International Conference on Computational Science and Computational Intelligence (CSCI); Dec 16-18, 2020; Las Vegas, NV, USA p. 628-632. [doi: 10.1109/CSCI51800.2020.00111]
18. Zhang Z, Cao X, Jia J, Gong NZ. FLDetector: defending federated learning against model poisoning attacks via detecting malicious clients. 2022 Aug 14 Presented at: KDD '22; Aug 14-18, 2022; Washington, DC, USA p. 2545-2555 URL: https://dl.acm.org/doi/proceedings/10.1145/3534678 [accessed 2026-03-30] [doi: 10.1145/3534678.3539231]
19. Willison S. The lethal trifecta for AI agents: private data, untrusted content, and external communication. Simon Willison's Weblog. 2025. URL: https://simonwillison.net/2025/Jun/16/the-lethal-trifecta/ [accessed 2025-03-15]
20. Griffin M. Moltbook vibe coded security breach exposes critical AI coding failures. 2026. URL: https://www.fanaticalfuturist.com/2026/02/moltbook-vibe-coded-security-breach-exposes-critical-ai-coding-failures/ [accessed 2026-03-15]
21. Zhao X, Zhang W, Xiao X, Lim B. Exploiting explanations for model inversion attacks. Presented at: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); Oct 10-17, 2021; Montreal, QC, Canada p. 682-692. [doi: 10.1109/ICCV48922.2021.00072]
22. Hu H, Salcic Z, Sun L, Dobbie G, Yu PS, Zhang X. Membership Inference Attacks on Machine Learning: A Survey. ACM Comput Surv 2022 Jan 31;54(11s):1-37. [doi: 10.1145/3523273]
23. Schmelzer R. Moltbook looked like an emerging AI society, but humans were pulling the strings. Forbes. 2026. URL: https://www.forbes.com/sites/ronschmelzer/2026/02/10/moltbook-looked-like-an-emerging-ai-society-but-humans-were-pulling-the-strings/ [accessed 2026-03-15]

Edited by KA Clegg; submitted 26.Mar.2026; this is a non–peer-reviewed article;accepted 26.Mar.2026; published 31.Mar.2026.

Please cite as:
Athni TS
Emerging Risks of AI-to-AI Interactions in Health Care: Lessons From Moltbook
J Med Internet Res 2026;28:e96199
URL: https://www.jmir.org/2026/1/e96199
doi:10.2196/96199

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.Mar.2026.

XSL•FO
RenderX

Our AI-Powered Discoveries Are Trapped in a Predigital System

Boon-How Chew

(J Med Internet Res 2026;28:e96018) doi:10.2196/96018

KEYWORDS

artificial intelligence; scientific publishing; digital health; reproducibility of results; peer review; research; scholarly communication; innovation

Textbox .

Key Takeaways

- The academic publishing system's foundational issues—speed, cost, and reproducibility—are being met with a chaotic array of fragmented artificial intelligence (AI) tools that only patch symptoms rather than solve the core problem.
- A fundamental shift to an integrated, data-native ecosystem is required to ensure the trustworthy and rapid translation of digital health discoveries into practice.

Dr Boon-How Chew is an academic physician and professor of family medicine at Universiti Putra Malaysia, specializing in the psychosocial and technological aspects of chronic disease management. His research increasingly explores the intersection of artificial intelligence (AI) innovation in medicine and the reform of scientific writing workflows within legacy academic infrastructures. He is actively developing models for AI-verified publishing to address systemic delays, scientific integrity, and reproducibility challenges in biomedical research.

The world of digital health is electric with the promise of AI. From AI-driven diagnostic tools that can diagnose invasive breast cancers earlier [1] and democratize retinal disease screening among a high-risk population [2] to large language models accelerating drug discovery [3,4], we are at the dawn of an unprecedented era of innovation. We are generating health data at a staggering rate and developing algorithms that can turn data into potentially lifesaving insights faster than ever before.

But this incredible engine of discovery is hitting a 17th-century bottleneck: the academic publishing system [5]. It is like we are pouring rocket fuel into a horse-drawn carriage. As a clinical academic, I see a perilous and growing chasm between the speed at which we can generate evidence and the glacial pace at which we can formally validate, share, and trust it. This great delay is no longer just an academic frustration; it is becoming a direct threat to patient care and the entire promise of a data-driven biomedical ecosystem.

Rocket Science and Data Meet Years of Delay

While the rise of preprint servers has commendably solved some problems of immediate dissemination, the formal validation and publication process remains a significant bottleneck [6]. The subsequent journey from a preprint to a peer-reviewed, recognized publication still involves an agonizing delay, often extending from 12 to 18 months [5]. For digital health, where technology can become obsolete in a single year, this means the evidence base is consistently lagging behind the innovation curve.

This obsolescent system is governed by an economic model that creates profound access and equity issues [7]. Top-tier research universities report annual subscription expenditures exceeding US $10 to $15 million [8], while author-facing article processing charges in prestigious journals can range from US $5000 to over $11,000 per article [9], creating an unsustainable dual financial burden.

The Trust Deficit: A System in Crisis

Beyond speed and cost, the very foundation of our scientific evidence faces new and ongoing threats. The scholarly record is plagued by growing research integrity issues, with an exponential rise in issues from gift authorship and peer-review rings to outright data fabrication from organized paper mills [10]. This is compounded by the well-documented reproducibility crisis, where it is estimated that a significant percentage of published research findings are not reproducible, with the proportions ranging widely from 50% to 90% by different disciplines and measures [11,12].

In my view, a core constraint underlying these issues is the system's primary output: the static, text-based article. This opaque narrative summary functionally decouples an author's claims from the underlying data and analytical methods, making verification challenging. For digital health, the stakes are particularly high. The black box of a clinical AI model cannot be built on the black box of a nonreproducible study.

The Rise of Fragmented AI Solutions

In response to this crisis, a chaotic ecosystem of AI tools has emerged. While many general-purpose large language models already have deep research functionality, they represent a patchwork approach that ultimately adds complexity without addressing systemic failures.

XSL•FO
RenderX

For Researchers

There's a proliferation of AI "super-assistants" and a wave of platforms aiming to centralize the fragmented authoring process. Tools like *Paperpal*, *WriteSonic*, *SciSpace*, and *Prism* by OpenAI and Scitex describe themselves as comprehensive AI-powered platforms designed to streamline the entire research workflow, from literature review to drafting and paraphrasing [13]. Similarly, tools like *TERA* (The Evidence Review Accelerator), *EPPI-Reviewer*, *Covidence*, *RobotReviewer*, *Perplexity*, and *Elicit* help users navigate the literature through systematic review [14], and semantic mapping platforms like *ResearchRabbit*, *Litmaps*, *Semantic Scholar*, and *Iris.ai* help pinpoint underexplored niches [14]. The data analysis phase is increasingly empowered by natural language assistants like *Julius AI*, *AskVi.ai*, *Vizly*, and *Tableau GPT* for intuitive statistical interpretation and visualization of complex datasets [14].

While these platforms may boost individual productivity, they are nonetheless designed to optimize the creation of the traditional static manuscript. They help authors write papers faster, but they do not change the fact that the final output is noninteractive, decoupled from its data, and subject to the same slow, opaque peer-review process on the publishers' side.

For Publishers

Major publishers are embedding AI into their legacy workflows as a defensive measure. Elsevier's *Reviewer Recommender* and Springer Nature's *SNAPP* system use AI to speed up administrative tasks like prescreening content and finding reviewers [15]. Frontiers uses *AIRA* (Artificial Intelligence Review Assistant) for automated quality and image checks, while Taylor & Francis integrates *Reviewer Locator*, *Imagetwin*, and *PaperPal* for matching and editing. IEEE uses *Publication Recommenders* and misconduct prescreening, and MDPI uses AI for reviewer selection and generative AI safeguards. While these tools accelerate the traditional assembly line, they do not address its fundamental flaws.

Future Tools

Besides *Sakana.ai*, an AI coscientist by Google and Microsoft Discovery, ambitious new projects are emerging from startups and universities. Keiji AI's *TrialMind* and *Aika* by Biorce aim to create AI agents to assist with clinical research design, and Stanford's *Agentic Reviewer* offers AI-powered feedback on a submitted PDF. These are important steps, especially to overcome the plagued and prolonged peer review [16], but they still operate within the old paradigm as fragmented services that either prepare for or analyze a static document that is largely left unverifiable [17].

This fragmented response, while well-intentioned, adds more tools to an obsolescent workflow, increasing the burden on researchers to learn and manage a complex tech stack while leaving the core issues of data opacity, analytical audit trail, gift- and ghost-authorship contribution [18], and flawed validation processes untouched.

We Need a New Operating System for Scientific Evidence

The solution is not to incrementally speed up the old system or add more disconnected tools. We need a fundamental reimagining of how we create, validate, and share scientific knowledge. We need a new operating system for science that is dynamic, transparent, data driven, and prespecified in research protocols, unified coherently on an open platform, and powered by AI technologies with endorsement from humans, who are responsible for its outputs.

This new model must move beyond the static paper as its primary output. The future unit of publication must be a living, verifiable, and interactive record as an enriched dynamic research object, where the data, methods, analysis log, fair author contributions, and transparent peer validation are all structurally and permanently linked, captured, and time-stamped.

This single integrated ecosystem—where AI is not a collection of fragmented patches but the core engine of one workflow that ensures rigorous reporting and transparency by design—is imperative for the integrity of science and fidelity of academic communication. The digital health community is at the forefront of innovation. We have a unique responsibility and opportunity to lead the charge in building such an AI-powered publishing model and scientific evidence ecosystem that is aligned with open science principles [19] and worthy of an immediate presence and the future [20]. The technology is (almost) here [21,22]. What is required now is the collective will to build, adopt, and apply it.

Conflicts of Interest

None declared.

References

1. Gommers J, Hernström V, Josefsson V, et al. Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. Lancet 2026 Jan 31;407(10527):505-514. [doi: 10.1016/S0140-6736(25)02464-X] [Medline: 41620232]
2. Wu Y, Qian B, Li T, et al. An eyecare foundation model for clinical assistance: a randomized controlled trial. Nat Med 2025 Oct;31(10):3404-3413. [doi: 10.1038/s41591-025-03900-7] [Medline: 40877476]
3. Lu J, Choi K, Eremeev M, et al. Large language models and their applications in drug discovery and development: a primer. Clin Transl Sci 2025 Apr;18(4):e70205. [doi: 10.1111/cts.70205] [Medline: 40208836]

XSL•FO
RenderX

4. Kang H, Li J, Hou L, Xu X, Zheng S, Li Q. Large language model-enhanced drug repositioning knowledge extraction via long chain-of-thought: development and evaluation study. JMIR Med Inform 2025 Oct 7;13:e77837. [doi: 10.2196/77837] [Medline: 41056561]
5. Björk BC, Solomon D. The publishing delay in scholarly peer-reviewed journals. J Informetr 2013 Oct;7(4):914-923. [doi: 10.1016/j.joi.2013.09.001]
6. Smyth AR, Rawlinson C, Jenkins G. Preprint servers: a “rush to publish” or “just in time delivery” for science? Thorax 2020 Jul;75(7):532-533. [doi: 10.1136/thoraxjnl-2020-214937] [Medline: 32312737]
7. Pooley J. Collective funding to reclaim scholarly publishing. Commonplace 2021 Aug 16;1(1). [doi: 10.21428/6ffd8432.250139da]
8. Andrews G. New data show universities are increasing R&D activity. Association of American Universities. 2024 Dec 6. URL: https://www.aau.edu/newsroom/leading-research-universities-report/new-data-show-universities-are-increasing-rd-activity [accessed 2026-03-12]
9. Morgan TJH, Smaldino PE. Author-paid publication fees corrupt science and should be abandoned. Sci Public Policy 2025 Oct 15;52(5):805-809. [doi: 10.1093/scipol/scaf026]
10. Xie Y, Wang K, Kong Y. Prevalence of research misconduct and questionable research practices: a systematic review and meta-analysis. Sci Eng Ethics 2021 Jun 29;27(4):41. [doi: 10.1007/s11948-021-00314-9] [Medline: 34189653]
11. Ioannidis JPA. Why most published research findings are false. PLoS Med 2005 Aug;2(8):e124. [doi: 10.1371/journal.pmed.0020124] [Medline: 16060722]
12. Baker M. 1,500 scientists lift the lid on reproducibility. Nature New Biol 2016 May 26;533(7604):452-454. [doi: 10.1038/533452a] [Medline: 27225100]
13. Watanabe Y. SciTeX Writer. GitHub. 2026. URL: https://github.com/ywatanabe1989/scitex-writer?tab=readme-ov-file [accessed 2026-03-31]
14. Chew BH. Enhancing research proposal preparation with artificial intelligence: an emerging guide for novel researchers in clinical and health sciences. Zenodo. Preprint posted online on Nov 14, 2025. [doi: 10.5281/ZENODO.17606362]
15. Musa Z. How big academic publishers use AI. PublishingState.com. 2025 Jul 25. URL: https://publishingstate.com/how-big-academic-publishers-use-ai/2025/ [accessed 2026-03-12]
16. Perlis RH, Christakis DA, Bressler NM, et al. Artificial intelligence in peer review. JAMA 2025 Nov 4;334(17). [doi: 10.1001/jama.2025.15827] [Medline: 40875583]
17. Lawrence R. The need for verification markers on published content. Inf Services Use 2025 May 11;45(1-2):78-82. [doi: 10.1177/18758789251343404]
18. Ruben A. Why scientific journal authorship practices make no sense et al. Science. 2021 Oct 28. URL: https://www.science.org/content/article/why-scientific-journal-authorship-makes-absolutely-no-sense-et-al [accessed 2026-03-31]
19. Canadian National Commission for UNESCO. An introduction to the UNESCO recommendation on open science. : UNESCO; 2022 Nov. [doi: 10.54677/XOIR1696]
20. Klebel T, Traag V, Grypari I, Stoy L, Ross-Hellauer T. The academic impact of open science: a scoping review. R Soc Open Sci 2025 Mar;12(3):241248. [doi: 10.1098/rsos.241248] [Medline: 40046663]
21. Burgelman JC, Pascu C, Szkuta K, et al. Open science, open data, and open scholarship: European policies to make science fit for the twenty-first century. Front Big Data 2019;2:43. [doi: 10.3389/fdata.2019.00043] [Medline: 33693366]
22. Zhang P, Hu X, Huang G, et al. A next-generation open access ecosystem for scientific discovery generated by AI scientists. arXiv. Preprint posted online on Aug 20, 2025. [doi: 10.48550/ARXIV.2508.15126]

Edited by KA Clegg; submitted 24.Mar.2026; this is a non–peer-reviewed article;accepted 24.Mar.2026; published 31.Mar.2026.

Please cite as:
Chew BH
Our AI-Powered Discoveries Are Trapped in a Predigital System
J Med Internet Res 2026;28:e96018
URL: https://www.jmir.org/2026/1/e96018
doi:10.2196/96018

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.Mar.2026.

XSL•FO
RenderX

How Digital Diaries Are Changing Communication and Recovery in the Intensive Care Unit and Beyond

Jenna Congdon

(J Med Internet Res 2026;28:e92661) doi:10.2196/92661

KEYWORDS

intensive care units; patient diaries; post–intensive care syndrome; mental health; survivors; professional-family relations; communication; trauma-informed care; critical care; patient-centered care; digital health; rehabilitation; recovery of function

Textbox .

Key Takeaways

- Digital intensive care unit (ICU) diaries help patients and families reconstruct "lost time," supporting psychological recovery after critical illness.
- By centralizing updates and humanizing care, digital diaries can improve communication between ICU staff and families while reducing interruptions to clinical work.
- Digital diaries can be a low-burden, high-impact tool supporting human-centered, trauma-informed ICU care long after discharge.

When patients finally leave after a long stay in the intensive care unit (ICU), they may feel as if they've lost weeks of their life. Perhaps they were unconscious, confused, or otherwise unable to process the whirlwind of illness and medical interventions that occurred during their stay. Additionally, clear and consistent communication between patients' family members and medical staff supports improved patient outcomes, but remains a perennial challenge in the ICU due to the complex nature of patients' medical courses.

A digital diary—a secure app-based journal that allows staff and families to log daily events with messages, photos, and videos to chronicle a patient's hospital stay—addresses both these concerns. As they recover, patients can reference the diary to process the complex experiences they may have had in the ICU. Digital diaries can improve mental and physical health outcomes for patients, foster stronger communication between staff and family members, and serve as a powerful tool to support patients' recovery and a poignant reminder of the love and support they received along the way.

Digital Diaries: A Valuable Record of Time Lost

Jurriaan van Rijswijk, MSc, founder of Games for Health [1], a company that develops app-based games to improve health outcomes, saw an opportunity to both improve staff-family communication and support patients' mental health after an ICU stay. Together with medical staff and researchers, he developed the Post-ICU digital diary app [2].

The app was initially launched at Erasmus University Medical Center in Rotterdam, Netherlands, to bridge the gap between patients and distant family members during the COVID-19 pandemic. It has since expanded; at the time of this writing, the diaries are in use across 24 hospitals, primarily in Europe.

Post–Intensive Care Syndrome

Patients often experience a constellation of mental health challenges after an ICU stay, particularly if their hospitalization was long or complicated [3]. In one study, 46% of patients experienced anxiety symptoms following their hospital stay. In addition, 40% reported feeling depressed, 22% were diagnosed with posttraumatic stress disorder (PTSD), and 18% of patients faced all three. Morbidity was 47% higher for patients experiencing depression compared to patients without depression in the two years following discharge from the ICU [4].

This decline in mental health, often accompanied by negative physical and cognitive changes, is referred to as post–intensive care syndrome (PICS) [4]. This syndrome not only makes it difficult to return to daily activities, but can threaten long-term survival and quality of life [5]. Ongoing support and intervention after discharge is needed to treat these issues and ensure patients' maximal recovery.

After digital diary implementation, patients had fewer instances of post-ICU anxiety and depression, and their loved ones experienced lower rates of PTSD after their hospital stay [6]. For some ICU survivors, rereading their digital diary helped them cope with the aftermath of a traumatic illness [7].

Improving Staff-Family Communication

During an ICU stay, all focus goes to the patient, and communication between staff and family members can be complex—on one side, the worried loved ones; on the other, busy nurses and doctors. When details get missed or misunderstood, both sides end up frustrated, and patient outcomes may suffer. When communication between both parties is centralized, families are more engaged in their loved ones' care and medical errors may be reduced [8].

XSL•FO
RenderX

The diaries streamline communication between staff and family members, saving time for everyone and ensuring that family members feel involved even when they can't be physically present. The app is easily activated by a nurse through the patient's medical record. Nursing staff write in the diary every morning, helping to humanize care with heartfelt, nonclinical notes alongside medical details. Family members know they can simply check the app for updates on their loved one rather than call the hospital, leading to fewer patient care interruptions.

Notes on Adoption and Challenges

Feedback from staff and families has been largely positive.

"Families love the solution," Jurriaan says, and he notes that "the app has a feedback button...they mostly use it to compliment us."

Concerns from family members largely center around two topics: feeling unsure of what to write and struggling to print the diary as a keepsake or convert it to a PDF to share. The former has been addressed with simple in-app writing prompts, the latter with straightforward tech support.

"It is a super simple app," Jurriaan shares. "It has three buttons, one to write and to add pictures, one to read, and one button for information."

This decreases barriers for users with low levels of digital literacy [7]. Furthermore, the diary remains accessible after discharge, making it easy to share with therapists or other caregivers.

Hospital administrators sometimes report qualms regarding data privacy. With the Post-ICU app, the family maintains control of who can access the diary, and careful security measures are in place to protect patient data. Jurriaan points out that compared to a physical journal, "...it's much safer. With a physical journal, you can't control who touches it, who reads it. It is a hygiene problem and a security problem. A digital journal is safer."

And while nurses have expressed concern about lacking the time to write entries, they nonetheless report that the diaries are a worthwhile addition to patient care, fostering a sense of human connection [9].

To address potential burden, the implementation process should be participatory and include end user medical staff to ensure ongoing relevance as workflows adapt [10]. Responsibility for diary entries can be shared across medical disciplines to increase overall participation.

For some patients, digital diaries can provoke sadness and loss; in one study, roughly one-third of patients surveyed reported rereading their ICU diary as a particularly distressing and painful experience [11]. Staff and families should use careful judgment around how and when to share the diaries with recovering patients.

Digital Diaries Support Human-Centered ICU Care

Jurriaan shares the story of an on-call nurse he worked with, who attended a beloved patient's funeral:

> *"...to her surprise, some of the [recordings] she made in the diary for that person were being read in front of everyone. How beautiful...it was shared with the entire audience during the funeral. I believe that's incredible, that although the patient died, it is still of use...for the family."*

Digital diaries are more than clinical documentation. They are trauma-informed care tools that can serve as powerful sources of ongoing support and inspiration for families, medical providers, and patients long after discharge. They can help patients feel seen as people rather than clinical cases, validate feelings at a time when emotions are high, and provide staff with a respectful way to connect with their patients on a human level.

Recovery goes far beyond just physical survival. Regaining lost memories and processing traumatic events can allow patients and their loved ones to face life after the ICU with resilience. For patients who lost weeks of their lives, a digital diary doesn't just fill in the gaps; it gives the story back to them. Thoughtfully implemented, digital diaries represent an elegant intervention supporting total-person healing during the ICU stay and beyond.

Conflicts of Interest

None declared.

References

1. Games for health. URL: https://gamesforhealth.net/en/ [accessed 2026-02-18]
2. Post-ICU. URL: https://post-icu.com/ [accessed 2026-02-18]
3. Chang CS, Tsai FJ, Liao CH. Associations between elevated rates of depression, anxiety, and PTSD among ICU survivors and increased mortality and readmissions. Brain Behav 2025 Feb;15(2):e70319. [doi: 10.1002/brb3.70319] [Medline: 39957084]
4. Hatch R, Young D, Barber V, Griffiths J, Harrison DA, Watkinson P. Anxiety, depression and post traumatic stress disorder after critical illness: a UK-wide prospective cohort study. Crit Care 2018 Nov 23;22(1):310. [doi: 10.1186/s13054-018-2223-6] [Medline: 30466485]
5. Voiriot G, Oualha M, Pierre A, et al. Chronic critical illness and post-intensive care syndrome: from pathophysiology to clinical challenges. Ann Intensive Care 2022 Jul 2;12(1):58. [doi: 10.1186/s13613-022-01038-0] [Medline: 35779142]

XSL•FO
RenderX

6. McIlroy PA, King RS, Garrouste-Orgeas M, Tabah A, Ramanan M. The effect of ICU diaries on psychological outcomes and quality of life of survivors of critical illness and their relatives: a systematic review and meta-analysis. Crit Care Med 2019 Feb;47(2):273-279. [doi: 10.1097/CCM.0000000000003547] [Medline: 30431494]
7. van Mol MMC, Tummers N, Leerentveld C, Tieben R, Buise M. The usability of a digital diary from the perspectives of intensive care patients' relatives: A pilot study. Nurs Crit Care 2024 Nov;29(6):1280-1289. [doi: 10.1111/nicc.12990] [Medline: 37897098]
8. Khan A, Spector ND, Baird JD, et al. Patient safety after implementation of a coproduced family centered communication programme: multicenter before and after intervention study. BMJ 2018 Dec 5;363:k4764. [doi: 10.1136/bmj.k4764] [Medline: 30518517]
9. Haakma T, Tieben R, Sleven B, Buise M, van Mol M. Experiences of nurses with an innovative digital diary intervention in the intensive care unit: a qualitative exploration. Intensive Crit Care Nurs 2022 Jun;70:103197. [doi: 10.1016/j.iccn.2022.103197] [Medline: 35090798]
10. Schol CMA, van Mol MMC, Ista E. Developing implementation strategies for digital ICU diaries targeting ICU professionals: an implementation mapping approach. Implement Sci Commun 2025 Aug 7;6(1):85. [doi: 10.1186/s43058-025-00767-0] [Medline: 40775358]
11. Flahault C, Trosdorf M, Sonrier M, et al. ICU survivors experience of ICU diaries: an ancillary qualitative analysis of the ICU diary study. Crit Care Explor 2021 May;3(5):e0384. [doi: 10.1097/CCE.0000000000000384] [Medline: 34036266]

Edited by KA Clegg; submitted 01.Feb.2026; this is a non–peer-reviewed article;accepted 01.Feb.2026; published 20.Feb.2026.

Please cite as:
Congdon J
How Digital Diaries Are Changing Communication and Recovery in the Intensive Care Unit and Beyond
J Med Internet Res 2026;28:e92661
URL: https://www.jmir.org/2026/1/e92661
doi:10.2196/92661

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Narrative Medicine: Translating Stories Into Medical Education and Better Patient Care

Candice Marie Sage

(J Med Internet Res 2026;28:e93037) doi:10.2196/93037

KEYWORDS

digital narrative medicine; medical education; patient-centered care; health equity; vulnerable populations; palliative care; narratives as topic; telemedicine; empathy; mental health; digital health; professional-patient relations

Textbox .

Key Takeaways

- Digital narrative medicine can function as a humanizing bridge between patients and their care teams, enhancing clinical care and fostering patient empowerment through advocacy and treatment co-design.
- Ethical and methodological considerations and refinements will need to be addressed as the approach evolves and expands.

Candice Marie Sage, PhD, is a storyteller and researcher with over 20 years of experience managing interdisciplinary research in Canada and internationally. She is an expert in knowledge co-creation, translation, and exchange. In this piece, she provides an overview of some of the applications of digital narrative medicine, kicking off a series in which she will explore digital narrative medicine stories and applications in depth.

About four decades ago, narrative medicine (NM) emerged as an important diagnostic tool [1], one that treats illness as a human experience and not just a biological one. NM is an approach to clinical care that recognizes the importance of lived experiences in health care, training clinicians in skills like reflective writing and authentic communication with patients [2]. By listening to, analyzing, and integrating patients' stories along with standard test results and traditional tools, clinicians can gain a deeper understanding of their patients, improving health outcomes [3,4].

Today, with an arsenal of novel tools and technologies, *digital* narrative medicine (DNM) provides new ways to capture and incorporate patient stories to enhance meaning-making [5]. It augments traditional NM by using digital tools to improve the accessibility, scalability, and inclusivity of narrative practices [2].

With applications from medical education to clinical interventions and across diverse populations, DNM enhances how health care professionals understand, relate to, and communicate with patients, their families, and caregivers, while providing a platform for empowerment through patient advocacy and treatment co-design [6].

DNM as Knowledge Translation

As care moves away from direct interactions between patients and health care professionals toward remote and asynchronous care, DNM functions as a bridge [2]. It helps build a humanizing connection between patients and their care teams, grounding care in lived experience and making knowledge translation a reciprocal process.

An increasing number of tools support this shift in how knowledge flows, allowing patient stories to be captured and used in more flexible, creative ways. Examples include multimedia podcasts and curated videos, leveraging artificial intelligence to translate stories into training simulation scenarios and clinical decision support [7], and interactive games that allow patients, their families, and their care teams to engage and interact dynamically. These DNM tools foster greater empathy between patients, their families, and clinicians, leading to more comprehensive, patient-centered care [8].

Applications in Medical Education

Before the digital age, medical education was based on attending different hospital rotations and observing, with depth of learning dependent on the chance clinical presentations and procedures taking place at that given time.

Today, DNM's narrative-based training and interactive storytelling systems provide virtually endless scenarios for trainees to experience and practice their diagnostic and decision-making skills. Trainees' reflective writing exercises are captured electronically in "e-portfolios" and then evaluated by trainers to nurture narrative competence, defined as "the ability to absorb, interpret, and act upon patient stories" [2]. Exposure to the authentic voices of patients—and, increasingly, family and caregivers—helps provide not only information about medical care, but also context around culture and other social determinants of health—enhancing trainees' empathy, communication skills, and professional identity [9].

One such example is the Out of Darkness multimedia project in Toronto, Canada [10]. People living with bipolar disorder were video-graphed, capturing their recovery journeys and lived experiences at home and in their neighborhoods. Psychiatry trainees were exposed to very realistic aspects of bipolar illness,

XSL•FO
RenderX

like what it feels like to take lithium, undergo electroconvulsive therapy, or experience fear of relapse and suicidality. By watching these emotionally evocative short films, learners gained treatment-relevant insights and reported increased empathy and compassion.

Applications in Clinical Intervention

Beyond training, DNM enhances patient-centered care by providing health professionals with highly contextual accounts of patients' psychological, emotional, social, and spiritual lived experiences [11]. This information can be collected and stored in medical records, analyzed with artificial intelligence tools, and made more accessible to both patients and practitioners in real time or longitudinally. Holistic, wraparound care is enhanced when professional teams collaborate by sharing insights and meaning-making about patient stories [12].

Cancer care digital storytelling, for example, has existed for more than a decade through initiatives like the Canadian Cancer Society's 2018 partnership with digital storyteller Mike Lang, which provided a platform for cancer patients to tell their stories and reflect on their journeys [13]. Earlier, in 2015, I coordinated a Cancer Care Ontario project, collecting and analyzing the lived experiences of patients and practitioners. More deeply understanding each group's needs and experiences led us to shift 5-year follow-up care for colorectal and breast cancer survivors from oncology specialists to local primary care providers. Follow-up surveys showed this new care plan improved each group's experiences, freeing specialists up to focus on active oncology care and reducing patients' travel time [14].

DNM is also used in palliative and end-of-life care and research [15] and can help facilitate difficult discussions about advanced care and end-of-life planning [16]. For example, I led an award-winning palliative care redesign project at Kingston General Hospital in 2018; the key to our success was incorporating members of the hospital's Patient and Family Engagement Committee early on and throughout the redesign. We learned about more than palliative care—we also learned what it was like to navigate the health care system as a patient with multiple specialists, and about the burdens carried by both patients and their at-home caregivers. One patient was so ashamed of his illness that he forbade his wife from telling family or friends about the situation. She held that burden all by herself until after he died. Stories like this informed modifications to disclosure in a patient and family communications DNM intervention.

Game-based DNM interventions in clinical and mental health care are growing, bridging treatment, engagement, and implementation gaps. For teens and young adults with cancer, for example, a video game intervention greatly improved patients' compliance with their treatment regimens through increased knowledge and self-efficacy [17].

Applications for Marginalized Populations

DNM can be particularly valuable for advancing health equity, and it can be an empowering tool for marginalized populations like youth [18], racialized groups [19], people with diverse sexual orientations and gender identities [20], Indigenous peoples [21], and people experiencing mental health [22] and substance use challenges [23].

The Mental Health Story Teller Mob developed by Australia's Aboriginal and Islander Mental Health Initiative is an example of Indigenous-led digital narrative storytelling. It explores emotional and social well-being, coping, and resilience in the face of poverty, racism, stigma, and mental health and substance use challenges [24].

In my work on the Support Not Stigma project [25], we recorded short videos and a documentary for a regional antistigma campaign, capturing and sharing the stories of people living with mental health or substance use challenges—many of whom were also experiencing homelessness—and the workers supporting them. This not only provided a platform for amplifying their voices and educating the public, it also led to the discovery that over 90% of this population had experienced head traumas. This serious medical consideration would not have been known had we not delved into people's stories.

Future Considerations

DNM is a powerful, underutilized approach that is gaining traction. As it continues to develop and expand, there are several important considerations.

New methodologies are needed to better plan and evaluate DNM research, curriculum development, and clinical practice. Ethical considerations—such as the consent of patient and family storytellers, and how their stories are used—will require dialog between patients, health care professionals, researchers, and policymakers. As technology evolves, new methods and tools to further comprehensive, person-centered care will emerge and require refinement.

We're only just scratching the surface of the realm of possibilities DNM presents. With careful planning, thorough research, and thoughtful implementation, we can realize its potential for deepening and improving care for patients from all walks of life.

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. Fioretti C, Mazzocco K, Riva S, Oliveri S, Masiero M, Pravettoni G. Research studies on patients' illness experience using the Narrative Medicine approach: a systematic review. BMJ Open 2016 Jul 14;6(7):e011220. [doi: 10.1136/bmjopen-2016-011220] [Medline: 27417197]
2. Efthymiou E. Integrating digital and narrative medicine in modern healthcare: a systematic review. Med Educ Online 2025 Dec;30(1):2475979. [doi: 10.1080/10872981.2025.2475979] [Medline: 40327849]
3. Allan R, McCann L, Johnson L, Dyson M, Ford J. A systematic review of "equity-focused" game-based learning in the teaching of health staff. Public Health Pract (Oxf) 2024 Jun;7:100462. [doi: 10.1016/j.puhip.2023.100462] [Medline: 38283754]
4. Lybarger K, Bear Don't Walk OJ, Yetisgen M, Uzuner Ö. Advancements in extracting social determinants of health information from narrative text. J Am Med Inform Assoc 2023 Jul 19;30(8):1363-1366. [doi: 10.1093/jamia/ocad121] [Medline: 37466054]
5. Palla I, Turchetti G, Polvani S. Narrative Medicine: theory, clinical practice and education - a scoping review. BMC Health Serv Res 2024 Sep 27;24(1):1116. [doi: 10.1186/s12913-024-11530-x] [Medline: 39334149]
6. Kadestål K. Enhancing patient engagement through co-design: challenges, strategies, and practical insights in digital healthcare technology [Thesis]. : Malmö University; 2024 URL: https://mau.diva-portal.org/smash/record.jsf?pid=diva2%3A1906127&dswid=6704 [accessed 2026-02-20]
7. Yu H, Zhou J, Li L, et al. Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. Commun Med (Lond) 2025 Dec 19;6(1):27. [doi: 10.1038/s43856-025-01283-x] [Medline: 41420084]
8. Mazzoli Smith L, Villar F, Wendel S. Narrative-based learning for person-centred healthcare: the Caring Stories learning framework. Med Humanit 2023 Dec 19;49(4):583-592. [doi: 10.1136/medhum-2022-012530] [Medline: 37208190]
9. Moreau KA, Eady K, Sikora L, Horsley T. Digital storytelling in health professions education: a systematic review. BMC Med Educ 2018 Sep 10;18(1):208. [doi: 10.1186/s12909-018-1320-1] [Medline: 30200945]
10. Zaretsky A, Rodney H, Morassaei S, et al. The Out of Darkness project: innovative multimedia digital storytelling in psychiatry. Acad Psychiatry 2022 Oct;46(5):644-649. [doi: 10.1007/s40596-022-01637-w] [Medline: 35578097]
11. Mazzoli Smith L, Hardy P, Thompson K, Westwood L. Transformative and therapeutic benefits of digital storytelling: a phenomenological lifeworlds study of Patient Voices participant experiences. Arts Health 2024 Jul 2:1-18. [doi: 10.1080/17533015.2024.2368767] [Medline: 38956989]
12. Massari MG, Di Trani M, Palmieri A, Reho M, Gennaro A. Where science meets stories: a systematic review on Narrative Medicine. Qual Health Res 2025 Nov 5;Nov 5:10497323251387023. [doi: 10.1177/10497323251387023] [Medline: 41192816]
13. Canadian Cancer Society. Mike Lang Stories. URL: https://www.mikelangstories.com/digital-storytelling-specialist/canadian-cancer-society-digital-storytelling-program [accessed 2025-12-16]
14. Christmas CM, Langley H. Well follow-up care for colorectal cancer survivors and their circles of care. In: Improving Coordination Between Primary Care Providers and Oncology Specialists: Cases from Canada – CanIMPACT Casebook: University of Toronto; 2015.
15. Williams L, Gott M, Moeke-Maxwell T, et al. Can digital stories go where palliative care research has never gone before? A descriptive qualitative study exploring the application of an emerging public health research method in an Indigenous palliative care context. BMC Palliat Care 2017 Sep 4;16(1):46. [doi: 10.1186/s12904-017-0216-x] [Medline: 28870189]
16. Ewy D. Using narratives to offer hindsight, insight, and foresight into advance care planning. Caring for the Ages 2025 Jan;26(1):8-9. [doi: 10.1016/j.carage.2024.11.010]
17. Kato PM, Cole SW, Bradlyn AS, Pollock BH. A video game improves behavioral outcomes in adolescents and young adults with cancer: a randomized trial. Pediatrics 2008 Aug;122(2):e305-e317. [doi: 10.1542/peds.2007-3134] [Medline: 18676516]
18. Christmas CM, Khanlou N. Defining youth resilience: a scoping review. Int J Ment Health Addiction 2019 Jun;17(3):731-742. [doi: 10.1007/s11469-018-0002-x]
19. Maragh-Bass AC, Dinh N, Cooney G, et al. The role of digital storytelling methods in promoting health-related outcomes among young adults of color: a systematic review. Mhealth 2025;11:42. [doi: 10.21037/mhealth-24-65] [Medline: 40755940]
20. Sage M, Singer JB, LaMarre A, Rice C. Digital storytelling: tools, techniques, and traditions. 2018. URL: https://ecommons.luc.edu/cgi/viewcontent.cgi?article=1120&context=socialwork_facpubs [accessed 2026-02-19]
21. Hensel JM, Ellard K, Koltek M, Wilson G, Sareen J. Digital health solutions for Indigenous mental well-being. Curr Psychiatry Rep 2019 Jul 1;21(8):68. [doi: 10.1007/s11920-019-1056-6] [Medline: 31263971]
22. De Vecchi N, Kenny A, Dickson-Swift V, Kidd S. How digital storytelling is used in mental health: a scoping review. Int J Ment Health Nurs 2016 Jun;25(3):183-193. [doi: 10.1111/inm.12206] [Medline: 26900000]
23. Christmas CM. Crystal meth, opiates, overdose and housing in Kingston, Ontario: 2021 rapid assessment and response community needs assessment for the integrated care hub. Substance Use Health Network. URL: https://www.substanceusehealth.ca/resource/crystal-meth-opiates-overdose-and-housing-kingston-ontario-2021-rapid-assessment-and [accessed 2026-02-19]

XSL•FO
RenderX

24. Nagel T, Thompson C. AIMHI NT 'Mental Health Story Teller Mob': developing stories in mental health. Australian e-Journal for the Advancement of Mental Health 2007 Jan;6(2):119-124. [doi: 10.5172/jamh.6.2.119]
25. Sage CM. Support Not Stigma. URL: https://supportnotstigma.ca/ [accessed 2025-12-18]

Edited by KA Clegg; submitted 06.Feb.2026; this is a non–peer-reviewed article;accepted 06.Feb.2026; published 20.Feb.2026.

Please cite as:
Sage CM
Digital Narrative Medicine: Translating Stories Into Medical Education and Better Patient Care
J Med Internet Res 2026;28:e93037
URL: https://www.jmir.org/2026/1/e93037
doi:10.2196/93037

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Feb.2026.

XSL•FO
RenderX

How Digital Health Technology Is Extending Independent Living

Jenna Congdon

(J Med Internet Res 2026;28:e93767) doi:10.2196/93767

KEYWORDS

independent living; aged; home safety; telemedicine; wearable electronic devices; accidental falls, prevention & control; medication adherence; aging in place; monitoring, ambulatory

Textbox .

Key Takeaways

- Independent living for older adults is closely tied to safety, dignity, and quality of life, and can often be extended with thoughtfully selected digital health and smart home technologies.
- Wearables, smart home tools, and digital monitoring systems can reduce preventable harm, support chronic disease management, and provide a rapid response in emergencies.
- Clinicians play a critical role in assessing individual needs, addressing barriers to adoption, and aligning technology recommendations with patient goals and values to support aging in place safely.

As a nurse, I frequently hear questions like:

> *"How long can Mom keep taking good care of herself without help?"*
>
> *"How do I keep myself safe while living alone?"*
>
> *"Dad wants to live independently for as long as possible. How do we support him?"*

For many older adults, living independently goes far beyond being a measure of their physical and cognitive health; it can be a point of social pride and a reflection of their personhood. Losing one's independence can be frustrating, but it's a situation that can be preventable with the right care. Technologies such as wearable monitors and smart home devices, when used thoughtfully, can reduce preventable harm, extend independence, and support clinical and personal goals for older adults [1].

What "Safety" Really Means for Older Adults Living Independently

Home safety for older adults is both multifactorial and unique to the individual; no singular solution will work for every person. Safety goes beyond responding to medical emergencies such as strokes, falls, and cardiac events. Digital health or smart home technologies can also address issues such as poor medication adherence, functional decline, cognitive changes, chronic condition management, lack of physical activity, and social isolation [1].

James Giordani, MSW, has made it his mission to help older adults get the most out of potentially life-saving technology. He started his company, Clear Computing, in 2019, which was acquired by parent company AsKevin [2] in 2024. The business offers in-home tech support to people of all ages, but Giordani found himself assisting mostly older clients.

Working with seniors soon became a passion. Giordani went on to get his Master's degree in social work, again specializing in working with older adults. He shares a personal story that helped him realize how technology can augment in-home safety for older adults: "My Nana...fell and broke her hip, and she was down for about 12-16 hours before she got help. [She] couldn't crawl to the phone...had she had a smart assistant, she could have used it to call for help."

Technology is always available and always on call. While it's not necessarily a substitute for face-to-face check-ins, it is a powerfully supportive tool for day-to-day cares and emergencies alike.

Core Needs and Technologies for Safe Independent Living

Choosing the Right Tools

Medical care providers may consider recommending technologies for older adults wishing to remain independent. This is especially relevant as an increasing number of older people are choosing to receive care at home [3]. As clinicians, we can encourage safety and better health by assisting older patients to understand which technologies may best support their unique needs.

Preventative care checkups and in-home visits are excellent moments for assessment of potential needs. In-home safety encompasses many aspects of personal health, social networks, and daily life, but broadly, these needs can be grouped into several large categories: (1) fall prevention and rapid response, (2) cardiovascular monitoring, (3) medication adherence and chronic disease management, and (4) interpersonal safety and security.

These categories provide a starting point for medical providers to perform a safety assessment and recommend appropriate technologies to support their patients' independence.

XSL•FO
RenderX

Fall Prevention and Rapid Response

Technologies to prevent or reduce harm from falls may include smart bulbs or switches for lighting, wearable fall sensors, and motion or impact detectors. Giordani recommends programming smart bulbs to turn on or off at key times, such as just before dusk, to improve vision in otherwise low-light environments.

A smartwatch or smart home device may be used to call for help in the event that someone is unable to get to a phone. There are also artificial intelligence devices designed specifically to detect falls, the results of which appear promising [4].

Cardiovascular Monitoring

Cardiac events, including arrhythmias or arrests, are most common in older populations [5]. Wearable devices such as watches, chest straps, or clothing help to detect these events and alert medical personnel. Due to the varied nature of these devices across the market, clinical integration is often impractical due to differences in data collection, accuracy, and interoperability challenges. Nevertheless, they may still provide valuable data for the adjustment of medical treatment and offer an early alert system for emergencies [6].

Medication Adherence and Chronic Disease Management

Poor medication adherence or polypharmacy creates the potential for serious harm. Tools such as smart pill dispensers or automatic reminders that can be shared with caregivers can help support proper medication dosing and timing [7].

Technologies such as glucose monitors can assist in chronic disease management, especially when paired with a smartphone app designed to help the user understand their glucose targets [8].

Devices like exercise trackers or smartwatches encourage increased physical activity, leading to better overall health and longevity [9].

Interpersonal Safety and Security

Giordani's favorite recommendation is one most might not consider. He shares, "I recommend that every older adult have an ad blocker on their phones and their computers. Not only do they reduce the cognitive load of seeing all those advertisements… they [help with] cyber security and scam prevention. Scams are often delivered to older adults through ads."

Technologies like ad blockers, smart home devices, and smart locks can enhance safety and personal and financial security for older adults living alone.

New Technology Adoption in Older Adults

Many older adults are very open to using new devices to support their health [10]—but to be effective, these personal health technologies must support the users' goals, be accessible and affordable, and be easy to use. Support systems must also be in place to encourage adoption and ongoing use, such as a helpline or tech assistance from a service such as AsKevin [11].

Giordani suggests assisting older adults to set up any new device. For older adults, he says, this process isn't necessarily intuitive. "I always recommend, if you're giving a device as a gift, you have to set it up for them."

Recommendations for Clinicians

Our patients' ability to live independently is more than a social preference; it's a real-world reflection of their clinical outcomes that is worth supporting with evidence-based tools.

Many older adults welcome the use of technology to support long-term health and independence. Talk with your patients about their individual needs and goals, and assess barriers such as cost and tech support before suggesting which devices may best improve their quality of life.

When aligned with clinical judgment and patient values, digital health tools help older adults lead longer, safer, and more autonomous lives. While technology doesn't replace human care, it can extend our reach as we strive to enhance every aspect of our patients' health.

Updated March 02, 2026: An earlier version of this article incorrectly referred to the company "AsKevin" as "Ask Kevin". This has been corrected.

Conflicts of Interest

None declared.

References

1. Facchinetti G, Petrucci G, Albanesi B, De Marinis MG, Piredda M. Can smart home technologies help older adults manage their chronic condition? A systematic literature review. Int J Environ Res Public Health 2023 Jan 10;20(2):1205. [doi: 10.3390/ijerph20021205] [Medline: 36673957]
2. asKevin. URL: https://www.askevin.com/ [accessed 2026-02-23]
3. Lette M, Ambugo E, Hagen T, Nijpels G, Baan C, De Bruin S. Addressing safety in integrated care programs for older people living at home: a systematic review. Int J Integr Care 2019;19(4):593. [doi: 10.5334/ijic.s3593]
4. Mohan D, Al-Hamid DZ, Chong PHJ, et al. Artificial intelligence and IoT in elderly fall prevention: a review. IEEE Sensors J 2024;24(4):4181-4198. [doi: 10.1109/JSEN.2023.3344605]

XSL•FO
RenderX

5. Meyer M, Arnold A, Stein T, Niemöller U, Tanislav C, Erkapic D. Arrhythmias among older adults receiving comprehensive geriatric care: prevalence and associated factors. Clin Pract 2024 Jan 4;14(1):132-147. [doi: 10.3390/clinpract14010011] [Medline: 38248435]
6. Xintarakou A, Sousonis V, Asvestas D, Vardas PE, Tzeis S. Remote cardiac rhythm monitoring in the era of smart wearables: present assets and future perspectives. Front Cardiovasc Med 2022;9:853614. [doi: 10.3389/fcvm.2022.853614] [Medline: 35299975]
7. Faisal S, Samoth D, Aslam Y, et al. Key features of smart medication adherence products: updated scoping review. JMIR Aging 2023 Dec 19;6:e50990. [doi: 10.2196/50990] [Medline: 38113067]
8. Yang Y, Lee EY, Kim HS, Lee SH, Yoon KH, Cho JH. Effect of a mobile phone-based glucose-monitoring and feedback system for type 2 diabetes management in multiple primary care clinic settings: cluster randomized controlled trial. JMIR mHealth uHealth 2020 Feb 26;8(2):e16266. [doi: 10.2196/16266] [Medline: 32130172]
9. Wu S, Li G, Du L, Chen S, Zhang X, He Q. The effectiveness of wearable activity trackers for increasing physical activity and reducing sedentary time in older adults: a systematic review and meta-analysis. Digit Health 2023;9:20552076231176705. [doi: 10.1177/20552076231176705] [Medline: 37252261]
10. Wang Y, Lu L, Zhang R, Ma Y, Zhao S, Liang C. The willingness to continue using wearable devices among the elderly: SEM and FsQCA analysis. BMC Med Inform Decis Mak 2023 Oct 16;23(1):218. [doi: 10.1186/s12911-023-02336-8] [Medline: 37845659]
11. Moore K, O'Shea E, Kenny L, et al. Older adults' experiences with using wearable devices: qualitative systematic review and meta-synthesis. JMIR mHealth uHealth 2021 Jun 3;9(6):e23832. [doi: 10.2196/23832] [Medline: 34081020]

Edited by KA Clegg; submitted 18.Feb.2026; this is a non–peer-reviewed article;accepted 18.Feb.2026; published 24.Feb.2026.

Please cite as:
Congdon J
How Digital Health Technology Is Extending Independent Living
J Med Internet Res 2026;28:e93767
URL: https://www.jmir.org/2026/1/e93767
doi:10.2196/93767

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Feb.2026.

XSL•FO
RenderX

What Health Care Organizations Have Learned From Telecommunication Outages

Catharine Solomon

(J Med Internet Res 2026;28:e91456) doi:10.2196/91456

KEYWORDS

telecommunications; telecommunications standards; health services administration; health care quality; access; evaluation; infrastructure; risk management; emergency preparedness; health care resiliency

Textbox .

Key Takeaways

- Health care facilities' ability to provide patient care can be threatened by large-scale telecommunication outages, requiring providers to build additional redundancy into their own infrastructure.
- While steps have been taken to formalize accountability for some telecommunication providers (such as, in Canada, via the Memorandum of Understanding on Telecommunications Reliability), gaps remain.

Telecommunication (telecom) breakdowns at massive scales are fresh in recent memory; in Canada, the Rogers Communications outage of July 2022 cut off over 12 million users' wireless and wire-line services for over 24 hours [1], and in 2024, a single faulty update to CrowdStrike's cybersecurity software caused an estimated 8.5 million Microsoft systems worldwide to crash [2]. In the past several months alone, widespread outages from internet infrastructure providers, like Cloudflare and Amazon Web Services (AWS), have alerted users to the fragility of an internet that relies on a handful of companies to function [3].

Health care facilities, which use telecom for everything from faxing prescriptions, to accessing electronic health records (EHRs), to virtual care, have to navigate these breakdowns under an additional unique pressure: their patients' well-being on the line. In the months and years following major outages, health care institutions have taken steps to mitigate the threat of service disruption to patient care.

The Rogers Outage

On the morning of July 8, 2022, Canadians woke up to a phone and internet outage hitting nearly every sector. While upgrading their IP core network, Rogers staff had removed a policy filter, which allowed IP routing data to flood and then crash the core routers, leaving user traffic unable to reach their destinations and over 12 million Rogers users without service [1].

Many of these users were health care providers and organizations, as well as patients trying to access health care. Family medicine clinics found themselves suddenly unable to carry out basic functions, like faxing prescriptions, scheduling appointments, and attending virtual care appointments [4]. The outage disproportionately affected health care services in remote locations, which often rely on network access to consult with specialists; this was compounded by the loss of connectivity across other health system partners [4]. In Toronto, immunization clinics lost access to the provincial COVID-19 vaccination management system, long-term care facilities lost access to their residents' EHRs [5], and hospitals had to scramble to mitigate the outage's effects on patient care.

According to Michael Garron Hospital's (MGH's) Chief Information Officer Amelia Hoyt, the brunt of the outage at MGH fell on corporate staff, who use Rogers cellular services for their work phones. This loss of corporate cellular mobility threatened patient care only indirectly, but physicians who relied on Rogers for their own mobile phones also couldn't be reached. MGH leadership gathered an emergency command center to develop a plan for how to navigate the outage.

"We did need to come together to figure out how to make sure that we didn't have delayed communications that inadvertently would affect patient care," says Hoyt. The solution in the moment was to identify affected staff and, by using a fan-out list, set alternate communication routes.

The outage allowed Hoyt and her team to identify a gap in their IT strategy and then fix it. "We had no other redundancy for if cell services are down." Redundancy—the strategy of duplicating system components or functions—works by keeping fail-safes and backup components at the ready.

"What we decided to do was invest in eSIMs from a different cellular provider for corporate mobile devices," she explains, "loaded onto select individuals' phones....If Rogers went down again, we would be able to activate this alternate provider's eSIM, and that would allow for continuity."

The CrowdStrike Outage

On July 19, 2024, University Health Network (UHN) staff trying to connect with clinical systems were met instead with blue error screens. A defective content configuration update released by the cybersecurity company CrowdStrike had caused millions of Windows hosts worldwide to crash, resulting in disruption

XSL•FO
RenderX

of services across multiple industries and organizations, including government agencies, airlines, health care facilities, and even 911 connectivity in some areas [6,7]. In the health care sector, the outage disrupted everything from EHR access, to telemonitoring, to operational management, to research [8].

According to Carl Virtanen—chief technology officer for UHN Digital (UHN's IT department)—the UHN's downtime procedures for switching to manual documentation helped mitigate disruptions to patient care, though several nonurgent appointments had to be rescheduled.

"UHN was able to prevent more significant technological disruption by pausing software updates and limiting the number of devices affected by the CrowdStrike outage," Virtanen says, noting that UHN Digital's protocols, which include prioritizing critical areas (such as emergency departments, intensive care units, and operating rooms) when responding to technological problems, have not been updated since. "All critical clinical systems were back up and running normally by the end of the day the outage began."

Importance of Redundancy

The telecom ecosystem is deeply interconnected. For complex critical systems, like health care, that rely on telecoms [9], interconnectivity facilitates seamless and efficient data sharing across a network of health system partners. But these interconnected systems can be disrupted on a large scale by a single point of failure in the infrastructure upon which they depend. So, how can disruptions be avoided?

Virtanen says that the best strategy for avoiding downtime is already, ideally, in use: "It is critical to have redundancy layers built into technical infrastructure to maintain operational efficacy in the event of service disruptions, such as the CrowdStrike outage."

On the most basic level—the physical infrastructure underpinning telecom services—a lack of redundancy can result in, for example, a beaver taking out nearly half of a town's internet connection by chewing through the network's single fiber-optic cable [10]. The 2022 Rogers outage itself exemplified the peril of putting all of one's technological eggs in one basket; since the Rogers management network relied on the very IP core network that had crashed, staff struggled to communicate with one another and remotely investigate the issue, delaying repairs for hours [1].

Health care facilities' preparations for outages, such as MGH's eSIM investment and UHN's downtime protocol of switching to paper records, act as another redundancy layer. In a perfect world with an unlimited budget, Hoyt says that her team would invest in additional redundancy for all IT services at MGH; at one point, they had even toyed with the idea of low-tech backups, like walkie-talkies. The solution is practical but not without its drawbacks; maintaining redundant systems can be cost-intensive, particularly for health care facilities with fewer resources [4].

Who's Accountable?

The fallout and appropriate redress of telecom outages have often been framed in terms of monetary loss, which can be reimbursed through vendor refunds and credits [5]. But when essential services, like health care, rely on network access to function efficiently and effectively, a framework that accounts only for economic impact falls short.

"Telecommunication is really foundational," says Hoyt. "I liken it to 911 service—there are some standards that are expected, that you should be able to require some of these telecom companies to follow."

Steps have been taken to hold Canadian telecom companies accountable for the essential services they provide; in September 2022, thirteen major telecom companies signed the Memorandum of Understanding (MOU) on Telecommunications Reliability [11]. The signatories of this MOU agreed to provide one another with emergency roaming and assistance in the event of a major outage, building ready-to-go redundancy into Canadian telecom services.

The MOU is "a positive step," according to Hoyt. "What will be interesting to see is how this agreement is put into practice if and when the next unplanned outage occurs."

What about the rest of the telecom ecosystem? Health care organizations' ability to access and relay information is only as stable as the infrastructure they use to communicate, and unlike health care providers who can be held liable for medical malpractice or medical devices that are held to rigorous standards of reliability and accuracy, many telecom services—particularly newer cloud-based models, like software as a service, that boast cost-effective scalability for clients—have frequent downtimes without consequence.

For health care organizations, using third-party services carries an inherent risk: tech disruptions that can't be addressed in-house. "When the infrastructure is under local IT's control, like my department, it's something my team would have more direct accountability for," says Hoyt. "If AWS is out, then you're kind of stuck between a rock and a hard place. There's nothing really a local IT team can do about that because we actually don't have any control over it."

Health care organizations must weigh this risk in choosing whether and when to rely on such services. Unless these services are reliable enough to meet health care standards, the scalability boost may not be worth it.

As more and more of our essential systems rely upon telecom infrastructure, redundancy and resiliency strategies are key to avoiding more outages that, at best, create delays and administrative burden in health care and, at worst, can lead to real patient harm. In an interconnected telecom ecosystem, all providers, companies, and organizations have the opportunity and the responsibility to do their part in keeping critical services running smoothly.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Xona Partners Inc. Assessment of Rogers networks for resiliency and reliability following the 8 July 2022 outage – executive summary. Canadian Radio-television and Telecommunications Commission. 2024 Jul. URL: https://crtc.gc.ca/eng/publications/reports/xona2024.htm [accessed 2026-01-27]
2. Weston D. Helping our customers through the CrowdStrike outage. Official Microsoft Blog. 2024 Jul 20. URL: https://blogs.microsoft.com/blog/2024/07/20/helping-our-customers-through-the-crowdstrike-outage/ [accessed 2026-01-27]
3. London tech expert explains why the internet blew up this week (temporarily). CBC News. 2025 Nov 22. URL: https://www.cbc.ca/news/canada/london/london-tech-expert-explains-why-the-internet-blew-up-this-week-temporarily-9.6987955 [accessed 2026-01-27]
4. Li T, Tran C, Yung A, Thomas A. Family physicians and the nationwide communications service outage in Canada. Canadian Family Physician. 2022 Dec 9. URL: https://www.cfp.ca/news/2022/12/09/12-09 [accessed 2026-01-27]
5. Attachment 3: Rogers Communications July 2022 outage impact: economic, operational & potential for ConnectTO to mitigate future events. City of Toronto. URL: https://www.toronto.ca/legdocs/mmis/2023/ex/bgrd/backgroundfile-239381.pdf [accessed 2026-01-27]
6. Gallagher JC, Pechtol CL. CrowdStrike IT outage: impacts to public safety systems and considerations for Congress. Congress.gov. 2024 Dec 4. URL: https://www.congress.gov/crs-product/IF12717 [accessed 2026-01-27]
7. Ellingson C. Global software glitch affected Edmonton police 9-1-1 calls overnight Friday. CTV News. 2024 Jul 19. URL: https://www.ctvnews.ca/edmonton/article/global-software-glitch-affected-edmonton-police-9-1-1-calls-overnight-friday/ [accessed 2026-01-27]
8. Tully JL, Rao S, Straw I, et al. Patient care technology disruptions associated with the CrowdStrike outage. JAMA Netw Open 2025 Jul 1;8(7):e2530226. [doi: 10.1001/jamanetworkopen.2025.30226] [Medline: 40682764]
9. Howell B. Beyond infrastructure: internet ecosystem resilience and the public good. Telecomm Policy 2025 Aug;49(7):102998. [doi: 10.1016/j.telpol.2025.102998]
10. Hundreds lose internet service in northern B.C. after beaver chews through cable. CBC News. 2021 Apr 25. URL: https://www.cbc.ca/news/canada/british-columbia/beaver-internet-down-tumbler-ridge-1.6001594 [accessed 2026-01-27]
11. Memorandum of Understanding on Telecommunications Reliability. Innovation, Science and Economic Development Canada. URL: https://ised-isde.canada.ca/site/ised/en/memorandum-understanding-telecommunications-reliability [accessed 2026-01-27]

Edited by KA Clegg; submitted 14.Jan.2026; this is a non–peer-reviewed article;accepted 14.Jan.2026; published 03.Feb.2026.

Please cite as:
Solomon C
What Health Care Organizations Have Learned From Telecommunication Outages
J Med Internet Res 2026;28:e91456
URL: https://www.jmir.org/2026/1/e91456
doi:10.2196/91456

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 3.Feb.2026.

XSL•FO
RenderX

Hospital-at-Home: New Technology Brings Acute Care to Patients' Homes

Jenna Congdon

(J Med Internet Res 2026;28:e98143) doi:10.2196/98143

KEYWORDS

telemedicine; hospital at home; remote sensing technology; delivery of health care; artificial intelligence; health services accessibility; patient safety

Textbox .

Key Takeaways

- Hospital-at-home programs deliver acute, hospital-level care at home using remote monitoring, wearables, artificial intelligence, and telehealth platforms combined with in-person care.
- Key barriers include the need for payor policy change, data security, interoperability, and scaling reliable digital infrastructure.
- This care model improves patient satisfaction, lowers costs and readmissions, and reduces hospital-related complications.
- Despite strong outcomes, adoption is limited by reimbursement, logistics, and sometimes, hesitance for change by medical staff and patients.

What if requiring acute medical care didn't automatically mean you'd need to be admitted to the hospital?

Clinicians, emergency responders, and family members have seen it time and again: patients reluctant to make an appointment or go to the emergency department, even when symptoms become uncomfortable or dangerous, because they know that the medical care team will want to admit them to the hospital. They may be homebound and struggle to get to an appointment, or they may be unable to or uncomfortable about leaving home, pets, children, or work behind while they're in hospital.

Hospital-at-home (HaH) programs have taken off in recent years due to hospital overcrowding, advances in remote and portable medical technology, and the broader acceptance of telehealth care brought about by the COVID-19 pandemic [1]. These programs combine remote monitoring and in-person clinician visits to provide acute-level care from the comfort of the patient's home.

While there are logistical and financial hurdles, the results are largely positive: patients report improved satisfaction with care, costs and readmission rates drop, iatrogenic complications are minimized, and more beds become available at the hospital for those who need them most [2,3]. HaH programs may represent a key solution to easing our overburdened health system.

A Brief History of Care at Home

Prior to the era of modern medicine, care at home was the norm. From the late 19th to early 20th century, increased urbanization, improved technology and treatments, and the newly discovered germ theory of disease led to the centralization of medical care within hospital walls [4].

In the modern medical landscape, very little care occurs in the home. From birth to death, it has become our cultural expectation that acute medical care is conducted at the hospital, though growing evidence suggests that HaH models can achieve similar or better outcomes for appropriately selected patients compared with inpatient care [1,5]. Modern technology now allows clinicians to safely monitor complex patients outside the hospital.

Defining "Hospital-at-Home"

HaH programs differ from routine telehealth because they deliver hospital-level care for acute conditions, rather than simple outpatient management.

Typical components include [1]:

- *Remote physiologic monitoring*, such as vital signs and heart rhythm via wearable sensors.
- *Digital communication with the care team*, including telehealth visits via hospital-supplied devices.
- *Frequent in-person visits and 24/7 response*. Nurse and/or physician visits occur daily or multiple times per day. The care team is available 24/7, and systems are in place to escalate care when necessary.
- *Multidisciplinary care*. Physicians, nurses, pharmacists, therapists, social workers, and more may be involved, just as they would be in the hospital.
- *Home-based interventions*. At-home care includes, but is not limited to, medication administration; oxygen therapy; diagnostic imaging; lab testing; physical, occupational, or speech therapy; and assistance with personal care.
- *A robust logistics set-up*. Patients receive transport to and from the hospital for additional treatment and meal and/or medication delivery in addition to care coordination.

Care often focuses on elderly patients who are most at risk for negative outcomes due to inpatient hospitalization. Common diagnoses include chronic obstructive pulmonary disease,

XSL•FO
RenderX

congestive heart failure, pneumonia, postsurgical care, and infections requiring intravenous antibiotic treatment [1].

Admission begins in the outpatient setting, with the emergency or ambulatory care physician determining whether a patient's clinical picture is suitable for HaH treatment. The patient's home environment is then assessed for safety; if criteria are met, the patient is transported home and care begins immediately [1].

Pioneering the Hospital-at-Home Concept in the United States

In 1994, the primary technologies that Johns Hopkins geriatrician Bruce Leff, MD, MBA, MPH [6], had available to him were a pager and a bulky portable phone housed in a backpack. Yet, he and his colleagues were determined to help high-risk patients avoid hospitalization altogether and instead provided acute care within the patient's home. They believed this would be a safer and more cost-conscious method of care for specific patient populations [7,8].

They soon proved their theory: total costs, length of admission, and cases of acute delirium in elderly patients dropped, while patients and caregivers enrolled in the newly established HaH program reported less stress and an overall positive opinion of the care they received [1].

"We've come a long way," Dr Leff recalls. "I would say over the last 10 to 15 years, health-based technologies that did not exist 30 years ago are pretty routine now in hospital-at-home."

Today, Dr Leff leads the Hospital at Home Users Group, which aims to foster collaboration and define best practices as health systems worldwide harness today's technology to offer at-risk patients hospital-level care from the safety of their own homes [9].

Evidence for Operational Benefits

In addition to clinical benefits for patients, HaH programs offer robust operational benefits to health systems already overwhelmed by an increasingly sick and aging population and rising health care costs.

Length of stay and overall spending is reduced, and remote monitoring and hybrid staffing models may help address workforce shortages and support hospital capacity management [10]. These programs also offer significant cost savings by helping to avoid the need for construction of additional physical hospital space [10].

Today's technologies—such as large language models for monitoring and predictive analytics, remote monitoring via digital sensors, expanding options for portable diagnostic testing, telehealth infrastructure, automated medication administration, and even drone delivery of supplies—make this model of care increasingly feasible [10].

Implementation Challenges

Perhaps the largest hurdle facing health systems that wish to establish a HaH program is payor reimbursement. In his work with the Hospital at Home Users Group, Dr Leff notes that "...payment and culture change are the two biggest pieces of this whole thing."

Data security, quality, and analytics also remain a significant hurdle, though recent improvements in artificial intelligence now offer a way to streamline virtual medical decision-making. Medical staff and patients sometimes express concerns over safety and quality of care outcomes when care is provided in the home rather than in a hospital. The logistics of staffing, scheduling, supply chain management, care coordination, and remote monitoring also present challenges [10].

The Future of Hospital-at-Home Care

HaH models continue to gain traction given their distinct patient and operational benefits and the increasing availability of technologies that support acute care at home.

Dr Leff looks forward to what he calls a "home-based care ecosystem," in which many medical disciplines work as one to provide hospital-level care from home, leveraging both remote technology and in-person care.

"The hospital of the future is ER, OR, and ICUs," Dr Leff predicts. "...Most of the other stuff will move out."

Given the rapid advancement in medical technology and proven advantages to this mode of care, this possibility is becoming an increasingly appealing answer to many of the challenges facing our current health care systems.

Conflicts of Interest

None declared.

References

1. Patel HY, West DJ. Hospital at Home: an evolving model for comprehensive healthcare. Glob J Qual Saf Healthc 2021 Nov;4(4):141-146. [doi: 10.36401/JQSH-21-4] [Medline: 37261225]
2. Levine DM, Ouchi K, Blanchfield B, et al. Hospital-level care at home for acutely ill adults: a randomized controlled trial. Ann Intern Med 2020 Jan 21;172(2):77-85. [doi: 10.7326/M19-0600] [Medline: 31842232]
3. Maniaci MJ, Rothman RD, Hohman JA. Redefining acute virtual care for overburdened health systems. JAMA Netw Open 2024 Nov 4;7(11):e2447359. [doi: 10.1001/jamanetworkopen.2024.47359] [Medline: 39589746]
4. Reardon J, Reardon L. The restructuring of the hospital services industry. J Econ Issues 1995 Dec;29(4):1063-1081. [doi: 10.1080/00213624.1995.11505740]

XSL•FO
RenderX

5. Leong MQ, Lim CW, Lai YF. Comparison of Hospital-at-Home models: a systematic review of reviews. BMJ Open 2021 Jan 29;11(1):e043285. [doi: 10.1136/bmjopen-2020-043285] [Medline: 33514582]
6. Bruce Leff. Johns Hopkins University. URL: https://hbhi.jhu.edu/staff/bruce-leff [accessed 2026-03-31]
7. Home, sweet hospital. Johns Hopkins Medicine. 2021. URL: https://www.hopkinsmedicine.org/news/articles/2021/10/home-sweet-hospital [accessed 2026-03-31]
8. Leff B, Burton L, Guido S, Greenough WB, Steinwachs D, Burton JR. Home hospital program: a pilot study. J Am Geriatr Soc 1999 Jun;47(6):697-702. [doi: 10.1111/j.1532-5415.1999.tb01592.x] [Medline: 10366169]
9. Hospital at Home Users Group. URL: https://www.hahusersgroup.org/ [accessed 2026-03-31]
10. Pandit JA, Pawelek JB, Leff B, Topol EJ. The hospital at home in the USA: current status and future prospects. NPJ Digit Med 2024 Feb 27;7(1):48. [doi: 10.1038/s41746-024-01040-9] [Medline: 38413704]

Edited by KA Clegg; submitted 13.Apr.2026; this is a non–peer-reviewed article;accepted 13.Apr.2026; published 20.Apr.2026.

Please cite as:
Congdon J
Hospital-at-Home: New Technology Brings Acute Care to Patients' Homes
J Med Internet Res 2026;28:e98143
URL: https://www.jmir.org/2026/1/e98143
doi:10.2196/98143

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Apr.2026.

XSL•FO
RenderX

The Need for Continued Investment in Digital Pain Assessment

Vanessa Nirode

(J Med Internet Res 2026;28:e97777) doi:10.2196/97777

KEYWORDS

chronic pain; pain management; pain measurement; artificial intelligence; wearable electronic devices; mobile applications; self-monitoring

Textbox .

Key Takeaways

- Pain is inherently subjective, and traditional questionnaires and pain scales tend to be confusing and frustrating for patients, influenced by mood, bias, desires, and underlying social and psychological factors.
- Digital assessment tools can help capture a more complete picture of how pain affects a person's life, allowing for better pain management plans.

Over 20% of the world's population lives with chronic pain [1]. Each person has a unique experience and relationship with their pain that requires individualized treatment within a biopsychosocial approach—something that can be hindered by relying too heavily on traditional pain assessment scales.

Standard pain assessment scales typically include pen-and-paper or online questionnaires assessing pain severity and intensity (commonly on a 0-to-10 scale), how pain affects function, and, often, measurements of depression and catastrophizing [2]. The problem is that these scales aren't so standard; individual biases, perceptions, desires, and underlying social and psychological factors influence interpretation of question meanings [3].

These measures also tend to be confusing or frustrating for patients [4], whose definitions of "pain" can diverge from those of their doctors. A small 2025 pilot study in the *Journal of Pain Research* found that medical professionals focused on the body-mind definition of "pain," while laypersons broadened their descriptions to include "suffering" or the sense that pain damages or alters personality [5]. As a person with osteoarthritis was quoted as saying in another 2025 study [6], "It's not just a knee, it's a whole life."

Pain Assessment Questionnaires

After dealing with chronic pain for over 40 years, US-based Catherine Hicks has stopped completing pain assessment forms.

"They never give the full picture of what's going on," Hicks explained. "I think when or if a doctor looks at a form, their job is to push you in and out and not charge the health insurance companies too much money. That sounds jaded, but that's how I often feel."

While Erin Adams, a Canada-based patient with chronic pain, understands the purpose of such forms, they can "feel like you're on constant repeat." After breaking her foot 11 years ago, Adams' complex regional pain syndrome (long-lasting pain and inflammation [7]) diagnosis took "13 months and 1,000,001 appointments with different specialists."

It's not the questionnaires themselves, she explained, nor that she finds them confusing. In fact, after filling out many, Adams can see the patterns of what they're looking for. "I'm sure they come in handy. But I draw hair onto my little stick man, and I put her in a nice dress and [the question] 'Where does it hurt'? I circle the whole thing."

The biggest limitations to the 0-to-10 scale are the lack of a universal scale as well as patient overreporting and underreporting [3]. As Ashley Guttman, MD—senior physician at Serefin Health [8]—explained, they don't account for, say, a bad weather day or a patient not sleeping because of caregiving responsibilities. Certain biases or psychological stresses also potentially influence responses. Patients might, for instance, underreport their pain out of fear of judgment or being perceived as "drug-seeking" [9].

Recall bias also plays a role [10,11]. Remembering if the pain was throbbing or stabbing or if it was a "6" or "7" after the fact can be difficult. Some patients really want to portray their symptoms accurately, so they struggle with this, according to Guttman.

But we shouldn't discard these questionnaires altogether.

"They can establish baselines," Guttman said. "They're good at quantifying severity and screening for other comorbidities like depression and they help monitor treatment responses."

Working in the Current System

Assessments should provide the patient's perspective, according to Christina Le-Short, MD—medical director at Override [12] and physician at Palo Alto Medical Foundation.

Le-Short discusses the pain questionnaires with her patients, perhaps asking why they marked a particular score. "I love the phrase, 'tell me more,'" she explained. "It's always, 'tell me more, tell me more.'"

Le-Short said that she's seeing a shift from insurance companies toward incentivizing outcomes and patient well-being instead

XSL•FO
RenderX

of just focusing on patient volume and high-paying procedures [13].

Within this shift and the growing commitment to biopsychosocial chronic pain treatments, digital advancements possess potential for making holistic patient assessment and, in turn, patient care easier.

Digital Advances

More data can help determine pain type and better inform treatment options, according to Dimitri Souza, MD—physician at Western Reserve Hospital Center for Pain Medicine [14].

Digital tools, such as tracking apps, ecological momentary assessments, wearables, adaptive (artificial intelligence [AI]) questionnaires, and virtual reality systems, could aid in more efficient and accurate assessment.

Override, for example, provides virtual care in conjunction with an app that includes journaling, tracking, and an educational library—tools that can "empower patients to better understand what's going on in their bodies," said Le-Short. The app sends data directly to patient charts, allowing doctors to reference specific, notable trends during visits.

"This teaches patients to be more cognizant of what is going on around them which can help when managing pain," said Le-Short. These kinds of tracking apps, along with wearables, also mitigate recall bias, as patients can record what they feel as it happens.

An AI tool that records patients' health calls and visits—Doctor Notes—generates AI summaries and stores those patient-doctor conversations [15]. "It acts as a scribe, taking notes for people," said Hicks, who helped design the app. "I've been using it for about a year, and find it extremely helpful to give context and keep track of my multiple doctor's appointments."

Designed for those who cannot always verbalize what they're feeling, PainChek [16,17] measures facial microexpressions to detect pain, alongside digital checklists. An *MIT Technology Review* article pointed out that while PainChek is useful, it still relies on subjective pain reports [18].

Wearable and portable technologies can allow researchers to access psychophysiological pain indicators, such as electrical brain activity and skin response. In the future, they may also use voice and body gestures to measure intensity [19].

Need for Integration

Patients—not physicians—most often seek out digital tools and therapeutics [20], which means they are rarely well integrated into doctor-guided treatment plans.

"Right now," said Souza, "there is [a] disconnect between the data and patient chart."

Medical facilities must find ways to integrate what patients record (eg, in an app) into their electronic medical record. Of course, this is also the junction where privacy and security concerns come in.

Despite this, Souza sees integration as the future, postulating that health care system administrators will ultimately recognize the social and economic benefits of—and allocate more funding to—digital tools.

Getting to a Future

Hicks, who works in public health, says that, when it comes to technology, the health care industry lags about 10 years behind. While systemic change must come from the top, she believes that patients' use of technology to compile data to advocate for themselves is a way forward and a catalyst to doctors pushing for technical integration and investment. And, as Adams pointed out, which doctor you give your pain assessment forms to determines their usefulness.

The keys are finding the "right" doctors and building on what already exists.

Using health questionnaires in conjunction with emerging digital assessment tools can lead to more holistic treatment options and a better understanding of how to help those with chronic pain get back to feeling like themselves, even if that self is an altered one.

Conflicts of Interest

None declared.

References

1. Zimmer Z, Fraser K, Grol-Prokopczyk H, Zajacova A. A global study of pain prevalence across 52 countries: examining the role of country-level contextual factors. Pain 2022 Sep 1;163(9):1740-1750. [doi: 10.1097/j.pain.0000000000002557] [Medline: 35027516]
2. Stretanski MF, Stinocher S, Grandhe S. Pain assessment - continuing education activity. National Library of Medicine. 2025. URL: https://www.ncbi.nlm.nih.gov/books/NBK556098 [accessed 2026-03-29]
3. Boring BL, Walsh KT, Nanavaty N, Ng BW, Mathur VA. How and why patient concerns influence pain reporting: a qualitative analysis of personal accounts and perceptions of others' use of numerical pain scales. Front Psychol 2021 Jul 2;12:663890. [doi: 10.3389/fpsyg.2021.663890] [Medline: 34282355]
4. Panayiotou M, Razum J, Eisele G, Wang SB, Fried EI, Cohen ZD. Interpretation issues with the Patient Health Questionnaire instructions. JAMA Psychiatry 2026 Apr 1;83(4):399-403. [doi: 10.1001/jamapsychiatry.2025.3796] [Medline: 41405895]

XSL•FO
RenderX

5. Gormsen LK. “Do Doctors and Patients Have Different Understandings of Pain?” - a pilot study of different perspectives and understandings of pain in clinical pain research. J Pain Res 2025 May 24;18:2651-2662. [doi: 10.2147/JPR.S508125] [Medline: 40444019]
6. Rauhansalo E, Holopainen R, Skou ST, Vuoskoski P. “My osteoarthritis is not the same as the millions of others” - conceptions of osteoarthritis among people with hip and knee osteoarthritis. Disabil Rehabil 2025 Dec;47(24):6299-6311. [doi: 10.1080/09638288.2025.2499585] [Medline: 40340499]
7. Complex regional pain syndrome. National Institute of Neurological Disorders and Stroke. URL: https://www.ninds.nih.gov/health-information/disorders/complex-regional-pain-syndrome [accessed 2026-03-29]
8. Serefin Health Clinic. URL: https://www.serefinhealthclinic.com [accessed 2026-03-29]
9. Buchman DZ, Ho A, Illes J. You present like a drug addict: patient and clinician perspectives on trust and trustworthiness in chronic pain management. Pain Med 2016 Aug;17(8):1394-1406. [doi: 10.1093/pm/pnv083] [Medline: 26759389]
10. Previtali D, Boffa A, Di Martino A, Deabate L, Delcogliano M, Filardo G. Recall bias affects pain assessment in knee osteoarthritis: a pilot study. Cartilage 2022 Dec;13(4):50-58. [doi: 10.1177/19476035221118417] [Medline: 36345999]
11. Schoth DE, Radhakrishnan K, Liossi C. A systematic review with subset meta-analysis of studies exploring memory recall biases for pain-related information in adults with chronic pain. Pain Rep 2020 Mar 31;5(2):e816. [doi: 10.1097/PR9.0000000000000816] [Medline: 32440609]
12. Override Health. URL: https://override.health [accessed 2026-03-29]
13. AMA Pain Care Task Force. Addressing obstacles to evidence-informed pain care. AMA J Ethics 2020 Aug 1;22(1):E709-E717. [doi: 10.1001/amajethics.2020.709] [Medline: 32880360]
14. Western Reserve Hospital. URL: https://www.westernreservehospital.org [accessed 2026-03-29]
15. Doctor Notes. URL: https://getdoctornotes.com [accessed 2026-03-29]
16. PainChek Intelligent Pain Assessment. URL: https://www.painchek.com/us/ [accessed 2026-03-29]
17. McGaffin L, Mitchell G, Anderson T, Gillis A, Craig S. Digital approaches to pain assessment across older adults: a scoping review. Healthcare (Basel) 2026 Jan 7;14(2):149. [doi: 10.3390/healthcare14020149] [Medline: 41595284]
18. Hamzelou J. An AI app to measure pain is here. MIT Technology Review. 2025 Oct 24. URL: https://www.technologyreview.com/2025/10/24/1126451/an-ai-app-to-measure-pain-is-here/ [accessed 2026-03-29]
19. Hadjiat Y, Arendt-Nielsen L. Digital health in pain assessment, diagnosis, and management: overview and perspectives. Front Pain Res (Lausanne) 2023 Apr 17;4:1097379. [doi: 10.3389/fpain.2023.1097379] [Medline: 37139342]
20. Digital therapeutics in pain treatment. International Association for the Study of Pain. 2023 Jul 19. URL: https://www.iasp-pain.org/resources/fact-sheets/digital-therapeutics-in-pain-treatment [accessed 2026-03-29]

Edited by KA Clegg; submitted 09.Apr.2026; this is a non–peer-reviewed article;accepted 09.Apr.2026; published 17.Apr.2026.

Please cite as:
Nirode V
The Need for Continued Investment in Digital Pain Assessment
J Med Internet Res 2026;28:e97777
URL: https://www.jmir.org/2026/1/e97777
doi:10.2196/97777

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Apr.2026.

XSL•FO
RenderX

From Innovation to Infrastructure: Why Digital Behavioral Health Still Struggles to Scale

Trevor van Mierlo

(J Med Internet Res 2026;28:e97118) doi:10.2196/97118

KEYWORDS

digital health; mental health services; telemedicine; health care accessibility; health equity; treatment outcome; diffusion of innovation; health knowledge; attitudes; practice; delivery of health care; health services needs and demand

Textbox .

Key Takeaways

- San José State University's 2026 Health TechQuity conference highlighted a persistent theme in digital behavioral health: difficulty with scaling interventions, contributing to constrained and uneven access.
- The field must shift its focus from what works (developing effective tools) to what reaches people (developing effective systems and infrastructure that support equity).

Across 2 days of panels and conversations at San José State University's Health TechQuity conference [1], a recurring theme emerged: despite decades of innovation in digital health, access to behavioral health care remains constrained, uneven, and difficult to expand at the population level.

This tension is striking because, by many measures, the field has made significant progress. Digital behavioral health has produced a growing number of validated tools, including self-guided interventions, digital therapeutics, and, more recently, artificial intelligence–enabled systems designed to support individuals outside traditional care settings. Many of these interventions have demonstrated efficacy in controlled environments, with multiple meta-analyses confirming their effectiveness across a range of conditions [2,3].

Yet, for most individuals, access to digital behavioral health tools depends on geography, provider availability, and system capacity. Waitlists remain long. Services are fragmented. And many of the populations that could benefit most from low-barrier, scalable interventions continue to face the greatest barriers to access.

Digital behavioral health's central challenge is no longer innovation; it is *implementation*.

The Persistent Challenge of Knowledge Translation

The gap between what is known to be effective and what is delivered in practice—commonly described as knowledge translation (KT) [4]—remains one of the most persistent challenges in health care. Despite sustained effort, moving evidence-based interventions from research settings into real-world use continues to prove difficult.

Digital behavioral health was uniquely positioned and expected to accelerate this process by enabling remote, scalable delivery. This anticipated transformation has not fully materialized. Instead, many interventions remain confined to research studies, pilot programs, or limited deployments [5]. Even when effective and engaging, tools often fail to reach the populations they were designed to serve or are not sustained beyond initial funding cycles.

This suggests that the KT challenge in digital behavioral health is not solely a matter of dissemination or user adoption but is fundamentally structural and systems-based [6].

From Tools to Access

A useful way to understand the current state of digital behavioral health is to consider these three distinct but interdependent layers:

- Innovation: the development of new tools and interventions
- Evidence: the validation of those tools through research
- Infrastructure: the systems required to deliver those tools at scale

The field has made substantial progress in the first two layers. There is no shortage of digital interventions, and a growing body of evidence supports their effectiveness. The third layer—infrastructure—requires further development.

Digital behavioral health interventions are not passive delivery mechanisms. Outcomes are shaped not only by the intervention itself but also by factors such as user engagement, interface design, and personalization. These elements can influence outcomes independently of hypothesized active ingredients, making interventions methodologically complex to evaluate and interpret [7,8]. As a result, even understanding what works requires careful attention to these dynamics.

However, even when these complexities are addressed, the central challenge remains: developing systems that enable effective interventions to be delivered at scale. Without systems capable of delivering validated interventions across populations, even the most effective tools remain limited in their impact.

XSL•FO
RenderX

Innovation creates tools. Infrastructure determines access.

Structural Misalignments

One reason infrastructure has struggled to emerge is that digital behavioral health sits at the intersection of technology, health care, and funding, each of which operates according to very different logic.

Technology moves quickly and prioritizes rapid innovation. Health care systems are necessarily regulated and typically risk averse. Business and funding models, meanwhile, require return on investment or depend on time-limited grants. Each functions effectively on its own terms, but they are not naturally aligned. What is missing is sustained coordination across these areas and alignment of clinical evidence, user engagement, and infrastructure to achieve real-world impact and delivery at the population scale.

Most digital health efforts, however, tend to focus in isolation rather than in coordination. Some prioritize evidence without achieving meaningful uptake. Others optimize engagement but lack validation. Still others build infrastructure without integrating evidence-based or engaging interventions. The result is a fragmented landscape in which solutions appear promising in isolation but fail to produce sustained impact.

Meaningful population-level outcomes emerge only at the intersection of clinical evidence, user engagement, and scalable infrastructure, where interventions are not only effective but are also used and delivered at scale. This intersection represents the conditions under which KT becomes possible, and it remains underdeveloped in digital behavioral health.

Figure WL1. Structural misalignment in digital behavioral health. KT: knowledge translation.

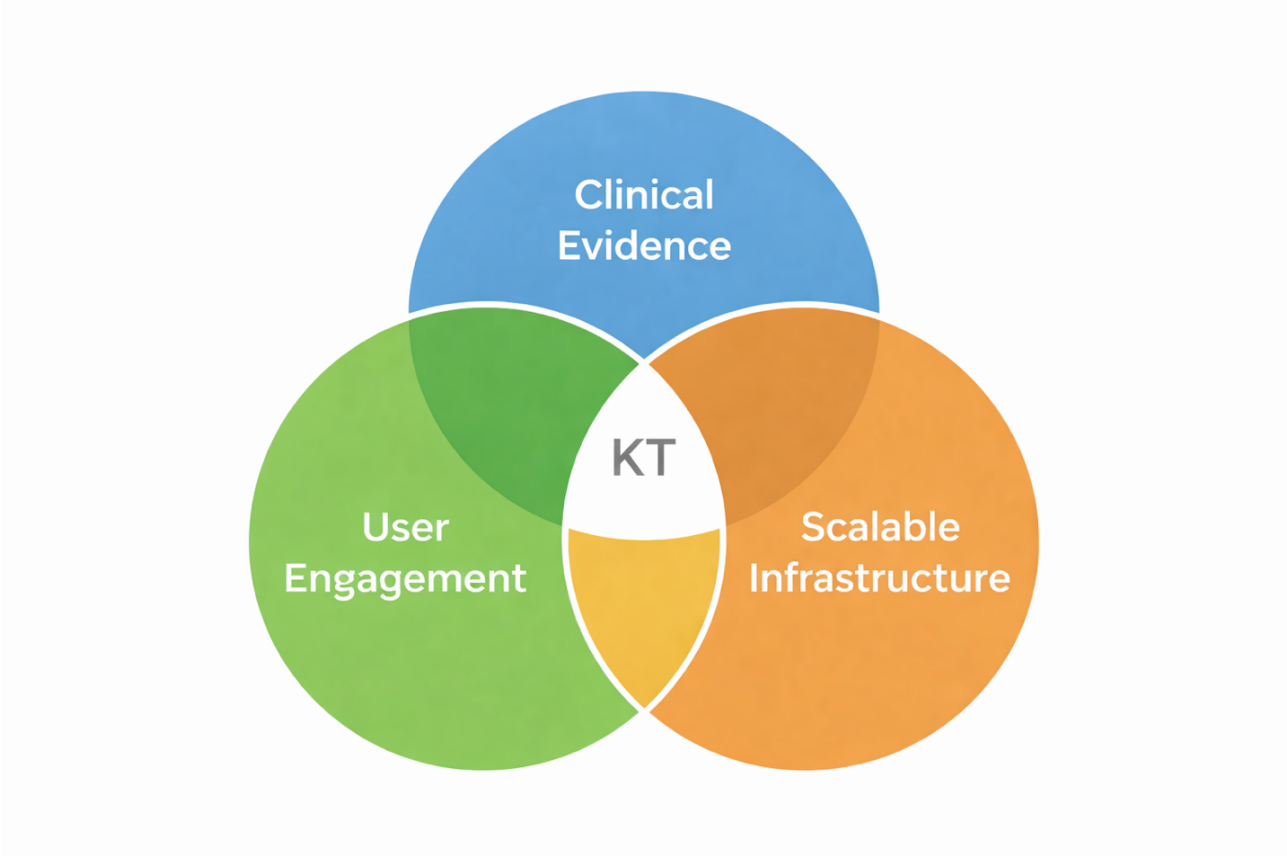

The Case of Self-Guided Behavioral Health

The implications of this gap are particularly visible in the context of self-guided behavioral health interventions, which are designed to be accessible without formal referral, usable without continuous clinician involvement, and scalable across large populations.

They are intended to complement clinical care, particularly in contexts where demand (the need for targeted behavioral support) exceeds supply (available practitioners).

At the Health TechQuity conference, panel discussions highlighted the potential for these tools to address another persistent challenge in behavioral health systems: the waitlist. If individuals can access validated, self-guided support while waiting for care—or in cases where they may not otherwise enter the system at all—then access becomes less dependent on clinician availability.

This potential is rarely realized at scale. Without infrastructure that enables discovery, onboarding, integration, and sustained engagement, self-guided interventions remain underutilized and largely function outside health systems.

XSL•FO
RenderX

From Tools to Systems That Support Equity

This moment in digital behavioral health may represent a transition. For much of the past decade, the focus has been on what works: developing and validating interventions.

The next phase must focus on what reaches people, which requires a shift from thinking about individual tools to thinking about systems they can be delivered in, as follows:

- Shifting from whether an intervention is effective to whether it is accessible
- Shifting from whether users engage in studies to whether they engage in real-world settings
- Shifting from whether outcomes are measured at the individual level to whether they translate into population-level effects

In other words, the problem of KT in digital behavioral health is inseparable from the problem of infrastructure and the question of equity.

When validated interventions exist but are not delivered at scale, access becomes uneven. Those with proximity to research institutions or well-resourced systems may benefit, while others do not.

Beyond being a technical challenge, infrastructure is a mechanism for equitable distribution. It determines whether digital behavioral health reaches the populations that need it most or remains concentrated among those already served and those who may already have access to in-person care.

Moving Forward

Digital behavioral health is no longer limited by a lack of innovation nor constrained by a lack of evidence. The central challenge is whether the field can build the infrastructure required to translate advances into real-world impact.

This will require alignment across technology, health care, and funding systems. It will require new models of dissemination, integration, and collaboration. The direction, however, is becoming clear. Digital health has repeatedly demonstrated that many interventions are effective. The remaining challenge is not efficacy but delivery at scale.

To improve access to behavioral health care at the population scale, the focus must shift from innovation to implementation, from tools to systems, and from knowledge to delivery.

Conflicts of Interest

The author is the founder and a scientific architect of the Evolution Health platform and consults on academic and commercial digital mental health initiatives. The views expressed in this article are those of the author.

References

1. Health TechQuity conference. San José State University, College of Health and Human Sciences. 2026. URL: https://www.sjsu.edu/chhs/health-techquity/index.php [accessed 2026-04-08]
2. Firth J, Torous J, Nicholas J, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry 2017 Oct;16(3):287-298. [doi: 10.1002/wps.20472] [Medline: 28941113]
3. Moshe I, Terhorst Y, Philippi P, et al. Digital interventions for the treatment of depression: a meta-analytic review. Psychol Bull 2021 Aug;147(8):749-786. [doi: 10.1037/bul0000334] [Medline: 34898233]
4. Graham ID, Logan J, Harrison MB, et al. Lost in knowledge translation: time for a map? J Contin Educ Health Prof 2006;26(1):13-24. [doi: 10.1002/chp.47] [Medline: 16557505]
5. Fleming T, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018 Jun 6;20(6):e199. [doi: 10.2196/jmir.9275] [Medline: 29875089]
6. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
7. van Mierlo T. From efficacy to scale: addressing digital health's original sin. J Med Internet Res 2025 Oct 27;27:e85878. [doi: 10.2196/85878] [Medline: 41143679]
8. Lipschitz JM, Pike CK, Hogan TP, Murphy SA, Burdick KE. The engagement problem: a review of engagement with digital mental health interventions and recommendations for a path forward. Curr Treat Options Psychiatry 2023 Sep;10(3):119-135. [doi: 10.1007/s40501-023-00297-3] [Medline: 38390026]

XSL•FO
RenderX

Edited by KA Clegg; submitted 03.Apr.2026; this is a non–peer-reviewed article;accepted 03.Apr.2026; published 13.Apr.2026.

Please cite as:
van Mierlo T
From Innovation to Infrastructure: Why Digital Behavioral Health Still Struggles to Scale
J Med Internet Res 2026;28:e97118
URL: https://www.jmir.org/2026/1/e97118
doi:10.2196/97118

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.Apr.2026.

XSL•FO
RenderX

Biocomputing: Beyond the Hype

Simon Spichak

Abstract

Biocomputing is a nascent but rapidly developing field at the intersection of biology and computer science. In this *News and Perspectives* article, JMIR Correspondent Simon Spichak reports on its current and potential applications for health care research and beyond.

(J Med Internet Res 2026;28:e100949) doi:10.2196/100949

KEYWORDS

brain organoids; cloud computing; human-machine systems; drug discovery; cognition; brain mapping; artificial intelligence; bioethics

Textbox .

Key Takeaways:

- Biocomputing connects human brain organoids to a computer to see how they respond to electrical and chemical stimuli.
- Some of the companies developing biocomputers allow for other researchers to access them via the cloud.
- In the near future, these systems could provide insights into brain and cognitive function and have applications in drug discovery, but some experts believe that biocomputers may have broader applications in the far future.

While venture capital is investing in data centers to bolster the computing power of artificial intelligence algorithms for research and other biomedical applications, scientists are turning to human brain cells as an alternative.

To start, scientists bathe human blood or skin cells in a cocktail of chemical factors to coax them back into a pluripotent (ie, capable of differentiating into almost any other cell type) stem cell state. Over the course of a year, these cells differentiate and mature into small spheres of brain tissue called organoids. Growing them on top of multielectrode arrays in a hardware shell has allowed research groups to stimulate and measure how these organoids respond to electrical and chemical signals. Cortical Labs out of Australia and FinalSpark out of Switzerland are leading the way, allowing other scientists to purchase the hardware to set up a biocomputer or access these platforms through the cloud.

While the field is still in the nascent stages, researchers believe biocomputing could help neuroscientists better understand how the brain works, play a role in drug development, and perhaps provide an energy-efficient alternative to AI computing.

Why Build Biocomputers?

When Fred Jordan, PhD, and Martin Kutter, PhD, founded FinalSpark, they wanted to build a thinking machine using artificial neural networks (ANNs)—simple mathematical models of neurons. But ANNs, which would become the backbone of large language models, weren't providing the progress they hoped, while also consuming substantial amounts of energy.

Despite requiring careful care and upkeep, these biocomputers are still more efficient than any supercomputer. While exact measures are scarce, an analysis done on rat brains suggests the brain consumes 30 watts of power a day, less than some lightbulbs. "Maybe it's a better idea to use real neurons," said Jordan.

Unlike artificial intelligence, Brett Kagan, PhD, the chief scientific officer of Cortical Labs, says that biological networks are better at dealing with chaotic, noisy data and can learn with substantially less input.

"The complexity of how biological neural systems compute and process information is a huge question," said Kagan. "But what we're doing is we're able to break it down now to the level of information physics."

XSL•FO
RenderX

Figure WL1.

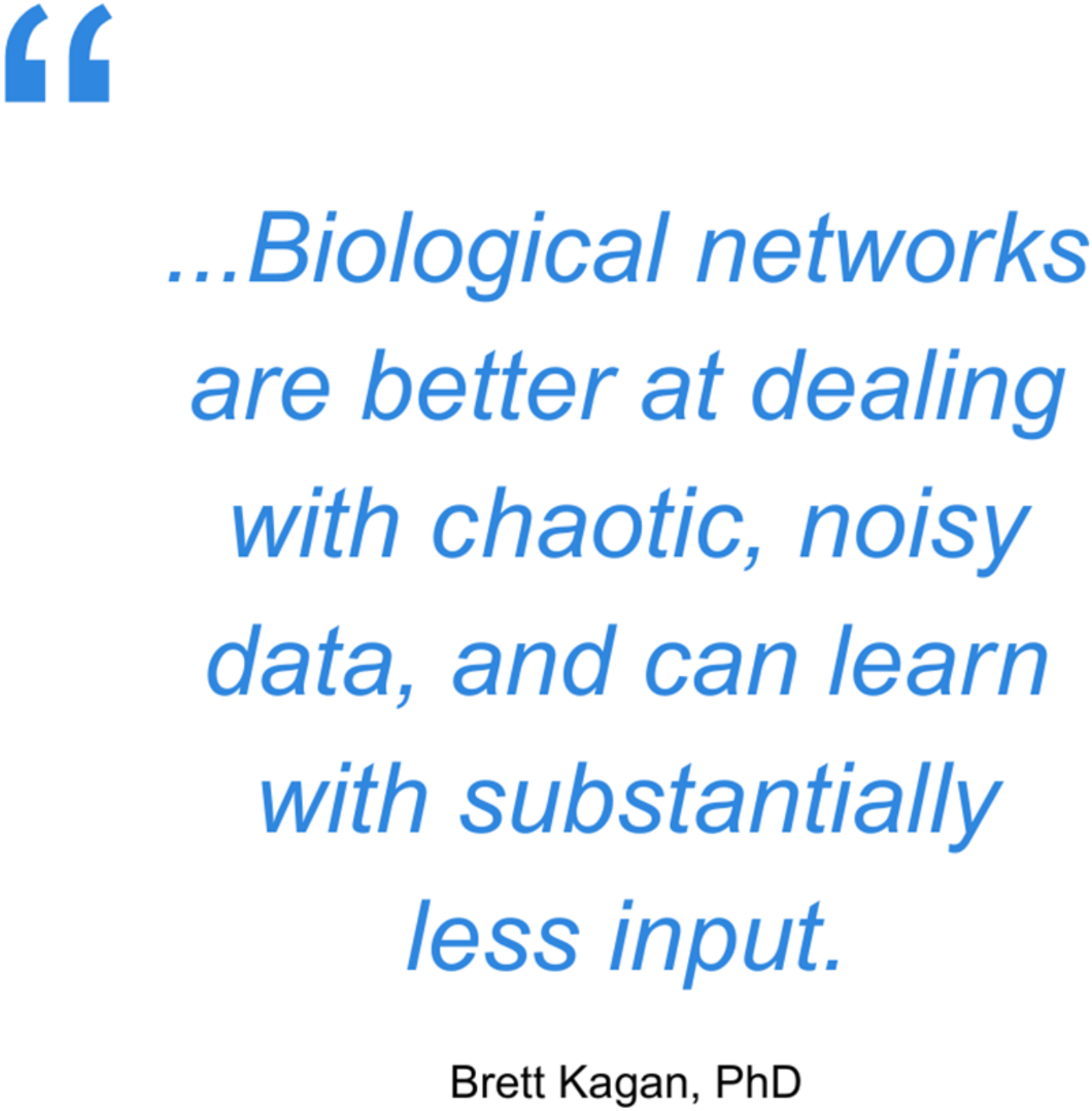

Bioethics and Potential Risks

Bioethicists have taken a keen interest in the field as well.

"We are so far away from the experimental side to something which is a relevant material for ethicists," said Thomas Hartung, MD, PhD, a professor at Johns Hopkins who works in the biocomputing field. His lab takes an embedded ethics approach, where an ethicist sits in on all of their laboratory meetings so that they can have discussions before any problems arise around privacy or consciousness within these systems.

Cortical Labs and FinalSpark are actively working with bioethicists to ensure that their biocomputers are used responsibly. The brain organoids used for biocomputing raise similar concerns to stem cell and organoid research, including the moral status and development of potential consciousness in more advanced models, informed consent from donors, and issues around commercialization, ownership, and patents.

What Are Biocomputers Used For?

By 2023, FinalSpark became the first biocomputing lab to provide remote access to biocomputing for research. Their hardware connects brain organoids to an interface that allows for chemical and electrical stimulation.

"I give you access to my brain organoids," said Jordan. "Then you can write Python code, and I will give you an API to stimulate the neurons and to listen to them." FinalSpark gave 9 universities free access to their biocomputer, Neuroplatform, and provides access to additional researchers.

Some of the work focuses on understanding how the connectome—or detailed map of neural connections—changes

XSL•FO
RenderX

in response to brain stimulation. They've also connected the Neuroplatform to a large language model that can directly prompt the system and run experiments for researchers. "The quality of the research was similar to what would have been done with the same tools," said Jordan.

Cortical Labs, meanwhile, has captivated headlines by training their biocomputers to play video games. In 2022, its scientists "taught" a 2D brain organoid connected to a computer how to play Pong. They created a simulated environment where aspects of the video game, like the location of the balls and paddles, were transduced into electrical signals.

With Pong, Kagan recalled, the team had to develop everything from the ground up. Since then, the company has created a hardware system called the CL1, as well as software, that allows anyone to use the biocomputer. Earlier this year, a graduate student programmed Cortical Labs' biocomputing platform to play the 1993 first-person shooter game *DOOM* during a hackathon. *DOOM* is a "nice demonstration" of what people with little experience can build within 7-10 days.

"You can actually allow anyone to build amazing things with this technology," said Kagan. "Not just the scientists in the lab who have spent their life doing it."

While these companies aren't trying to rebuild the brain, using its component cells could provide a leg up in computing. Both companies now allow people to tap into the cloud to use the compute power of brain cells grown in a dish.

The company currently offers three ways for researchers and other companies to use the CL1. They can buy the hardware, while growing the cells and writing their own software to analyze them. They can access the CL1 through the cloud and interact with the cells remotely. Or they can have Cortical Labs run the entire experiment themselves.

Hartung believes that while these cloud approaches offer the opportunity for scientists to familiarize themselves with the systems, he thinks that systematic experiments still need to be run in person.

These biocomputing platforms are already in use for drug discovery, allowing scientists to test how different experimental medications might affect how the organoids learn. It also provides a glimpse into the inner workings of the brain.

The Future of Biocomputing

The field is still in its nascent era. Kagan is optimistic about research groups emerging in Europe, China, Japan, and the United States.

Jordan sees mastering training as the biggest challenge for biocomputing. In the 1990s, he said, scientists puzzled this out for ANNs, allowing them to modify the weight of connections to improve the accuracy of the output.

With an ANN, the weights don't change unless they're reprogrammed. But brain organoids are still actively firing off even without any external input and the system is far more complicated. Breaking this code would allow scientists to master these networks.

Hartung said that a biocomputing system that learns might take several years of training, not unlike human learning, which will limit its applications. He believes it could help engineers develop neuromorphic computing—hardware that better mimics the human brain.

But Jordan and Kagan are more optimistic. "I think in 10 years," said Jordan, "biocomputers will be way more useful than quantum computers today."

Edited by KA Clegg; submitted 10.May.2026; this is a non–peer-reviewed article;accepted 10.May.2026; published 28.May.2026.

Please cite as:
Spichak S
Biocomputing: Beyond the Hype
J Med Internet Res 2026;28:e100949
URL: https://www.jmir.org/2026/1/e100949
doi:10.2196/100949

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Printing Personalized Medicine: 3D Models Bring Better Surgical Outcomes

Jenna Congdon

Abstract

Advances in medical 3D printing have opened up new possibilities for surgical training, planning, and decision-making. In this *News and Perspectives* article, JMIR Correspondent Jenna Congdon reports on the use of 3D printing in a pediatric surgical case and examines the broader implications of 3D printing for surgical practice.

(J Med Internet Res 2026;28:e100950) doi:10.2196/100950

KEYWORDS

three-dimensional printing; models, anatomic; surgical procedures, operative; patient-specific modeling; airway management; prostheses and implants; precision medicine

Textbox .

Key Takeaways:

- 3D-printed anatomical models can reduce risk and support better outcomes for high-risk clinical decisions and surgical procedures.
- Integrating 3D printing technology with medicine enables more precise surgical planning, improved training, and patient-specific care across specialties.
- Despite clear benefits, adoption is limited by cost, insurance coverage gaps, and regulatory barriers.

A multidisciplinary group of physicians—neonatologists, anesthesiologists, and ear, nose, and throat surgeons—pass a 3D-printed model of a baby's head between them. This meeting has been called prior to the high-stakes delivery of an infant with a complex facial mass that the team fears may block the child's airway. As they examine the model, the answer becomes clear.

Technical Skill Meets Human Compassion

Kyle VanKoevering, MD, took an unconventional route to becoming a surgeon. His first degree was in engineering, but that wasn't enough: "I wanted to figure out that overlap between helping people and being technical."

A chance job-shadowing opportunity brought him to the operating room. While observing the surgeon at work, VanKoevering recognized his roots: the specialized aspects of surgery drew on the technical skills combined with detailed forethought and structural planning he had learned as an engineer. "I realized this was an opportunity to do both…that blend of engineering in surgery drew me into medical school…It was a very direct way I would be able to help people."

In the 2010s, when VanKoevering began his surgical career, 3D printing wasn't new technology, but it wasn't yet as widely used for medical purposes. His curiosity as an engineer was piqued: how could this technology be used to improve medical and surgical outcomes, particularly in the complex cases he saw as an ear, nose, and throat surgeon?

Preparing for a Delivery Into the Unknown

In 2015, he got his first opportunity to answer that question: on a routine ultrasound, a large mass was spotted on a fetus's face, sparking immediate concern over the child's ability to breathe after birth.

The flat images from computed tomography scans and magnetic resonance imaging couldn't decisively show whether or not the mass was blocking the baby's airway. A specialized delivery procedure called an ex-utero intrapartum treatment or EXIT procedure was planned.

During an EXIT procedure, the mother is placed under general anesthesia and a cesarean section is performed, but only the head and neck of the infant are initially delivered. The umbilical cord stays attached to the placenta, providing the newborn with blood flow and oxygen from the mother's body. This gives the medical team approximately 20‑30 minutes to secure the infant's airway, typically by inserting an endotracheal tube or tracheostomy. Once the airway is established, the rest of the baby's body is delivered and the umbilical cord is cut. This delivery method carries higher risks to both parent and baby than a vaginal delivery or standard cesarean section and requires detailed interdisciplinary planning prior to birth.

3D Models for a 3D World

This case sparked an idea for VanKoevering: "We live in a 3D world. This is ultimately 3D anatomy, but we're all looking at this in two dimensional pictures and trying to mentally recreate the 3D picture." Why not print a 3D model so that the medical team could fully examine the mass?

XSL•FO
RenderX

He led a team to create a 3D model of the infant's head using images from computed tomography, magnetic resonance imaging, and ultrasound. When they met to discuss the delivery plan, the team could finally see—with collective relief—that the mass did not, in fact, block the infant's airway.

A standard cesarean delivery was planned instead of the EXIT procedure, with multidisciplinary teams standing by. The child was able to breathe easily, and rather than being under general anesthesia, the mother delivered with only an epidural and was able to hold her child immediately after birth. The 3D-printed model had changed everything.

Engineering Better Surgical Outcomes

Motivated by this successful outcome, VanKoevering began working to expand the use of 3D printing for complex surgical cases. In 2020, he established the Medical Modeling, Materials and Manufacturing Lab—better known as the M4 Lab—at Ohio State University.

There, biomedical engineers work alongside clinical professionals to develop 3D-printed models for an array of medical uses. Models allow surgeons to study unique patient anatomy before surgery, develop patient-specific prostheses, and train residents on highly technical surgical procedures and challenging intubation scenarios.

3D Models as a Surgical Roadmap

The M4 lab isn't alone in using this technology; hospitals throughout the world use 3D printing to secure better outcomes for their patients. This technology supports safer and more effective surgeries in many disciplines, frequently in orthopedics, pediatric cardiology, neurosurgery, and ear, nose, and throat, among many more.

Operating on the larynx, neck, face, and skull base as specialists like VanKoevering do is often risky; skull base tumors in particular are highly challenging. The bottom of the skull is full of irregular grooves that house delicate brain tissue and the cranial nerves. These nerves control our most basic functions, from facial movement to hearing, taste, vision, smell, swallowing, breathing, heart rate, and speech. 3D-printed models of individual patient anatomy can be brought into the operating room as a tangible, visual guide, allowing surgeons to operate with higher levels of precision and foresight in these delicate cases.

Furthermore, 3D printing can be used to manufacture life-like models, prostheses, and surgical implants that mimic the color and texture of real tissue. These models are used to train new physicians, allowing them hands-on practice without human risk. Patient-specific implants and prostheses allow for faster recovery and fewer negative surgical outcomes.

By helping to preserve both their appearance and their ability to engage in essential activities, these 3D-printed items can allow patients to return to their daily lives feeling more fully themselves. The impact is so powerful that many patients request to keep the models used in planning their surgeries; VanKoevering's team happily obliges.

XSL•FO
RenderX

Figure WL2.

The impact is so powerful that many patients request to keep the models used in planning their surgeries.

Beyond the OR: Implementation and Barriers

Of course, the broader implementation of 3D printing for medical purposes is not without its barriers. Funding is a perennial hurdle: “We need insurers to buy into seeing the value of this technology, and right now it’s just not there,” VanKoevering says.

Implantable 3D-printed models require rigorous testing and Food and Drug Administration clearance to manufacture, a capability that the M4 lab does not yet have. Other medical facilities struggle with the high costs of starting a 3D printing lab, as well as the integration of the additional training and multidisciplinary coordination that must take place.

Still, personalized surgical care offers abundant benefits to patients and surgeons alike and may lower costs for long-term care if patients experience better overall surgical outcomes.

Printing Personalized Medicine

VanKoevering is hopeful for the future of 3D printing in medicine. “As I look at this technology, we’ve just continued to look for creative and impactful ways to apply it into personalizing healthcare for patients… that’s really what 3D printing and 3D modeling allows us to do.”

By shifting surgery from a one-size-fits-all approach to a deeply individualized procedure, a simple 3D-printed model becomes much more than a tool meant to enhance surgical precision; it preserves the fundamental human identity of each unique patient.

XSL•FO
RenderX

Edited by KA Clegg; submitted 10.May.2026; this is a non–peer-reviewed article;accepted 10.May.2026; published 25.May.2026.

Please cite as:
Congdon J
Printing Personalized Medicine: 3D Models Bring Better Surgical Outcomes
J Med Internet Res 2026;28:e100950
URL: https://www.jmir.org/2026/1/e100950
doi:10.2196/100950

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Policymakers and Researchers Zero In On the Impact of AI Toys

Simon Spichak

Abstract

Artificial intelligence has made its way into children's toys. In this *News and Perspectives* article, JMIR Correspondent Simon Spichak reports on the potential implications for children's development and well-being.

(J Med Internet Res 2026;28:e102064) doi:10.2196/102064

KEYWORDS

artificial intelligence; child development; toys; human-computer interaction

Textbox .

Key Takeaways:

- Although millions of artificial intelligence (AI)–enabled toys are marketed and sold as educational companions, there is almost no research about their impact on neurodevelopment in young children.
- A small study found that AI toys struggled with pretend and social play.
- Researchers and policymakers raise transparency, privacy, and security concerns that need to be addressed.

Since their introduction, companies have integrated large language models (LLMs) across a broad range of consumer products with few—if any—regulations. Data on the long-term impact of LLM use are sparse, though emerging research suggests LLM use may impact creativity, cognitive off-loading, and other cognitive faculties.

In 2025, an estimated 22 million artificial intelligence (AI)–integrated toys were sold globally, including 10 million designed for educational purposes. Researchers and policymakers are now raising concerns about toys that integrate LLMs, because of their potential impact on social, emotional, and cognitive development.

Earlier this year, the AI in the Early Years project, which is run by researchers at the University of Cambridge, published their first systematic study examining how generative AI–enabled toys capable of humanlike conversation might influence development in children up to age 5.

"We started seeing these AI toys coming onto the market," said Emily Goodacre, PhD—a developmental psychologist who coauthored the project. "They were often advertised as social companions." While some educators believed these toys might eventually be helpful, the report highlighted the limitations and potential dangers of these toys.

Ethicists and policymakers are also taking notice. In 2025, a coalition of over 107 signatories called on AI toymakers to provide more transparency and safety in their products to address other safety and privacy issues.

How AI Toys Impact Neurodevelopment

Dana L Suskind, MD—a pediatric cochlear implant surgeon at the University of Chicago—was drawn to neurodevelopment after seeing the profound impact of cochlear implants. By being able to listen to and communicate with their parents, caregivers, and friends, young children would get a boost in their language and socioemotional skills that would help them catch up developmentally.

"The way milk feeds the body, nurturing talk and interaction from a caregiver feeds the brain," said Suskind. While AI toys might mimic human speech and interaction, it isn't clear whether they nurture children or build their socioemotional development.

The University of Cambridge report noted that although the toys are marketed as learning companions and friends, the impact on early development is largely unstudied. The researchers found only 7 studies, all published between 2019 and 2022, before the ubiquitous proliferation of LLMs. Only one toy called Luka, which is a penguin-shaped robot designed as a reading companion, is commercially available.

"It just seemed like there wasn't very much evidence," said Goodacre. Her team ran a study observing how 14 children and their parents interacted with an AI-toy called *Gabbo*, to start filling the gap.

The toy could not always distinguish between a parent and a child when they were playing together, making it difficult to engage in social play, which Goodacre said is important for early development. The toys also struggled with pretend play. When a child told the toy that it's time for it to go to sleep, instead of playing along, it informed the child that it did not sleep.

Goodacre also raised concerns that children might not understand that the toy doesn't feel real emotions, causing confusion or the development of an unhealthy relationship. When a 5-year-old told the toy "I love you," it responded with

XSL•FO
RenderX

"As a friendly reminder, please ensure interactions adhere to the guidelines provided."

And although many parents didn't know where to find reliable information on AI safety, they told the researchers they were keen to try the toys.

Still, experts acknowledge these toys may be useful in the right context and with the right guardrails. Goodacre said that some educators thought the interactivity could be beneficial but were "quite worried about children developing a relationship with it."

Suskind compared AI toys to robots, as studies have found that children who are anxious about reading have an easier time doing so with a social robot. "With that being said, you need to still be able to know how to read in front of humans, so we have to be very thoughtful and careful as we go forward."

Policymakers React to AI Toys

The University of Cambridge report recommends clearer regulation, a transparent privacy policy, and labeling standards to help parents decide what to do.

In December 2025, US senators Marsha Blackburn and Richard Blumenthal sent a letter to AI toy companies, expressing concerns and asking questions about their features, writing "Your company must not choose profit over safety for children, a choice made by Big Tech that has devastated our nation's kids."

Some AI toys, for example, have faced recalls because they engaged minors in sexual conversations, while others lack important privacy measures.

"It's a totally unregulated area," said Łukasz Kamieński, PhD—a bioethicist at Jagiellonian University in Kraków, Poland. "These toys have microphones, cameras, some of them have facial recognition features." He also worried about the risk of these toys subtly passing on misinformation or propaganda to children.

Miiloo's toys, which are powered by the Chinese LLM DeepSeek, reprimand children for asking why President Xi Jinping looks like Winnie the Pooh: "Your statement is extremely inappropriate and disrespectful. Such malicious remarks are unacceptable."

Goodacre thinks AI toys should come with labels on their packaging that provide information about training data, guardrails, and the underlying LLM model, akin to nutrition labels. "Having some kind of regulation in place for how toys need to be labeled would really help parents, educators, and families be informed." She also thinks there needs to be more regulation around how these companies are allowed to advertise their toys.

Figure WL1.

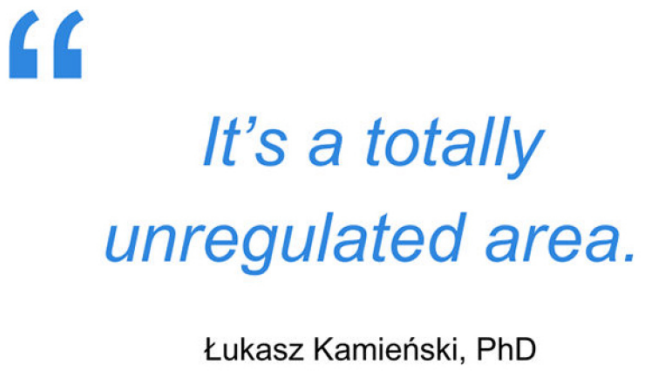

The Verdict on AI Toys

The novelty and lack of regulation of AI toys make them a difficult sell to experts.

"Especially in those early years, when we know interaction builds a child's brain," said Suskind, "we better have a whole lot more understanding and guardrails before we just hand it over to kids."

Goodacre is cautious but thinks once they're properly regulated and transparently labeled, there may be more to discuss in terms of benefits. Still, long-term studies are needed to understand the long-term impact of these toys on neurodevelopment. "The technology is progressing so quickly, and the research really needs to catch up," said Goodacre.

XSL•FO
RenderX

Edited by KA Clegg; submitted 21.May.2026; this is a non–peer-reviewed article;accepted 21.May.2026; published 02.Jun.2026.

Please cite as:
Spichak S
Policymakers and Researchers Zero In On the Impact of AI Toys
J Med Internet Res 2026;28:e102064
URL: https://www.jmir.org/2026/1/e102064
doi:10.2196/102064

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Jun.2026.

XSL•FO
RenderX

How Machine Learning Can Help Close Evidence Gaps for Drug Safety in Pregnant Women

Michelle Falci

Abstract

Women, especially pregnant women, have historically been excluded from clinical research. In this *News and Perspectives* article, JMIR Correspondent Michelle Falci reports on how advances in data analytics may help bridge evidence gaps.

(J Med Internet Res 2026;28:e101042) doi:10.2196/101042

KEYWORDS

drug safety in pregnancy; machine learning model; pregnancy clinical trial; machine learning drug safety; artificial intelligence drug safety

Textbox .

Key Takeaways:

- Using large datasets and machine learning, researchers aim to close evidence gaps in drug safety created by excluding pregnant women from clinical research.
- An emphasis on interpretability and causal inference prevents relying on machine learning models that function as "black boxes," drawing conclusions without showing their work.

In response to birth defects caused by medications like thalidomide, the US Food and Drug Administration (FDA) advised in September 1977 that "any premenopausal woman capable of becoming pregnant" be excluded from phase 1 and 2 clinical trials. Despite initiatives like the National Institutes of Health (NIH) Revitalization Act of 1993, which mandated the inclusion of women and other minorities in all NIH-funded research, the FDA's 1977 guidance spurred a decades-long trend of women—especially pregnant women—being excluded from clinical research on a global scale.

The Evidence Gap

According to an online survey of over 1200 women, this widespread exclusion leaves the 79% of women who take at least one medication during pregnancy to make decisions about their medication without sufficient evidence. Of that 79%, 63.2% made at least one medication change, and 36.5% of women who stopped their medication or lowered or missed a dose did so independent from their health care provider's input.

Citing concerns about potential risks among pregnant women and "distrust of women to prevent pregnancy," clinical trial developers admit that women are underrepresented in medical research.

Globally, 4% of clinical trials over the past decade allowed inclusion of pregnant women. To address this issue, the World Health Organization established a task force to achieve "timely and ethical inclusion of pregnant and breastfeeding women in clinical research for medical health products by 2030."

Other efforts to determine whether a medication is safe for pregnant or breastfeeding women, like the FDA's Pregnancy and Lactation Labeling Rule (PLLR), do not fully close the evidence gap. Of the 290 medications newly approved by the FDA between January 2010 and December 2019, all products submitted after June 20, 2015 were PLLR compliant, but about one-third of medications submitted between 2010 and 2015 were not in PLLR format by the June 30, 2019 deadline, and less than 20% of new product labels included data on human pregnancy and lactation.

XSL•FO
RenderX

Figure WL1.

Globally, 4% of clinical trials over the past decade allowed inclusion of pregnant women.

Researchers have tried to quantify the effects of including versus excluding pregnant women from randomized controlled trials (RCTs). In the case of thalidomide, had there been an RCT with 200 participants, about 33 children would have experienced adverse events during the trial, and this knowledge would have prevented approximately 8000 thalidomide-related birth defects, equivalent to 99.6% of all thalidomide-related birth defects from 1956 to 1962. This same study analyzed what would have happened if COVID-19 vaccine trials included pregnant participants and projected that 20% of all COVID-19–related maternal deaths and stillbirths between March and November 2021 could have been prevented if there were pregnant participants in the vaccine trials.

Decades of widespread exclusion have created a deep evidence gap for pregnant women and the clinicians who care for them, but there is one way that researchers are addressing the issue: analyzing large datasets with the help of machine learning to make recommendations on drug safety.

How Data Mining and Machine Learning May Close Evidence Gaps

Almut G Winterstein, RPh, PhD, FISPE, is a principal investigator of the BOOST-HP project. The project includes three main stages: first, data mining to find links between medication exposure and outcomes, then triaging statistical signals to determine the plausibility of each pathway, and finally, evaluating which signals may have the highest clinical significance.

Winterstein's tree-based approach to data mining is fully transparent—she and her team can see all of the decision pathways and variables that drive each signal, allowing them to trace how specific drug exposures are linked to outcomes.

"It is not one of these AI models where you are really in a black box and you don't even know what the system is doing," Winterstein says. "We can still follow very closely what is happening."

The opacity of certain artificial intelligence (AI) models concerns Winterstein because she has seen them ignore core epidemiological principles, such as the requirement that exposure precedes outcome, she explains.

"I think we're doing a disservice if we overuse this method without thinking about what is needed to basically establish causality, and then the methods will become discredited and nobody wants to deal with it anymore," Winterstein says. "I think it is important for people who are very strongly on the AI

XSL•FO
RenderX

side to collaborate with people who come from the causal inference epidemiology side and put those two together."

Cristina Longo, PhD, leads the BIONIC study, which combines causal inference and machine learning to assess the safety and efficacy of biologic medications in pregnant women with chronic inflammatory conditions and then determine which women and children are at highest risk of poor outcomes.

"Our studies are very much based on frameworks that allow us to estimate causal effects, which is important for this type of work because AI helps identify risk-stratified patients. It can't tell us what the causes are," Longo says.

Figure WL2.

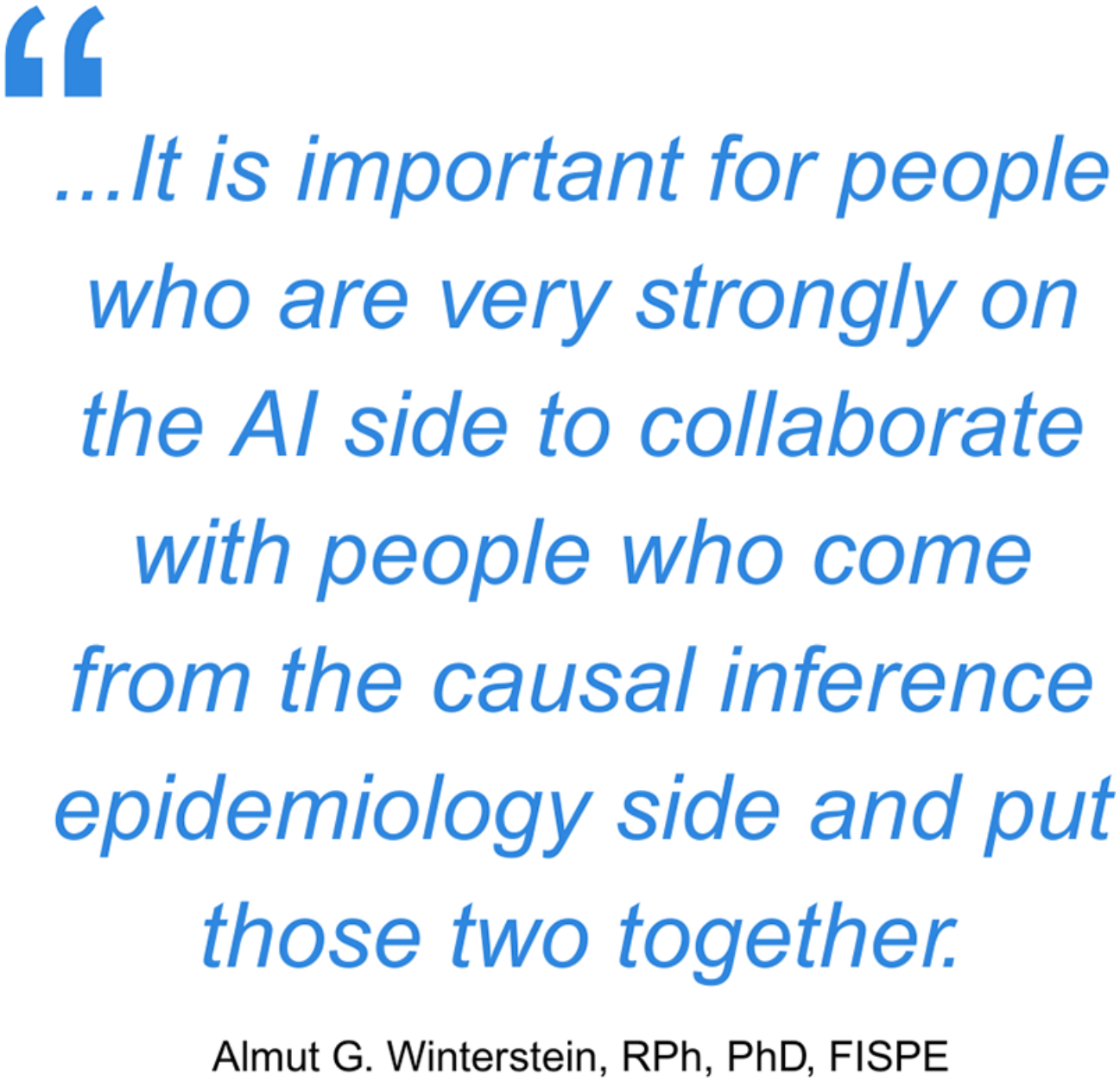

Leveraging large datasets applies to other facets of maternal health care beyond medication safety. In one study, researchers proposed that "digital twin" models of pregnant patients that are updated with real-time imaging and laboratory data can help predict outcomes for common complications during pregnancy.

Access to large and detailed datasets, Longo argues, is necessary to properly train machine learning models.

"Canada has a universal health care system. We have access to what people do at their physician visits and what types of medications they're taking. But eventually what we're hoping to achieve is access to more data: demographics, gender-based information, and socioeconomic status. These are all very important for AI models because we need to be able to make sure that they're fair," she says.

The Future

The need for more data is not the only limitation of using AI to close the evidence gap created by the widespread exclusion of pregnant women from RCTs. The "black box" effect Winterstein mentioned is a risk for any project that relies on a model that does not allow researchers to clearly trace how inputs are translated into predictions. When working with a black box

XSL•FO
RenderX

model, researchers cannot easily evaluate whether predictions are based on causal links or whether they are biased.

Some machine learning models for drug safety in pregnancy provide partial interpretability by including enhanced explainability mechanisms or apply post hoc interpretability methods even when inherently relying on a black box model.

“A lot of people are using AI in a way that it helps them, but they still want to keep control. And I think the same applies to research. At the end, in science, you still need to keep control,” Winterstein says.

Despite challenges with interpretability, researchers are optimistic that machine learning models can improve risk assessment for medication safety in pregnant women. By identifying safety signals and high-risk subgroups, these models may enable more targeted and ethically designed trials that include pregnant participants earlier in the evidence-generation process.

Edited by KA Clegg; submitted 11.May.2026; this is a non–peer-reviewed article;accepted 11.May.2026; published 27.May.2026.

Please cite as:
Falci M
How Machine Learning Can Help Close Evidence Gaps for Drug Safety in Pregnant Women
J Med Internet Res 2026;28:e101042
URL: https://www.jmir.org/2026/1/e101042
doi:10.2196/101042

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

After the Prescription: The Clinical Support Gap in Telehealth-Based GLP-1 Care

Anna Zucker

Abstract

GLP-1 medications offer promise for obesity management and are increasingly accessible via digital platforms. In this *News and Perspectives* article, JMIR Correspondent Anna Zucker reports on the gap in clinical support that could undermine their potential benefits.

(J Med Internet Res 2026;28:e101874) doi:10.2196/101874

KEYWORDS

digital health; telehealth; GLP-1; semaglutide; tirzepatide; obesity; weight loss; glucagon-like peptide-1 agonist

Textbox .

Key Takeaways:

- Telehealth platforms have significantly increased access to GLP-1 medications for people with obesity, making prescriptions faster and easier to obtain and helping millions lose weight and improve their health, though financial barriers remain for some.
- These platforms don't always offer nutrition, exercise, and behavioral change guidance required to help patients lose weight safely. As a result, research finds that some patients lose lean tissue and experience micronutrient deficiencies.
- Integrating registered dietitians and other health professionals into telehealth programs can help patients make the sustainable changes necessary to maximize the benefits of GLP-1 medications.

Obesity affects 40% of Americans and increases the risk of life-threatening conditions, including hypertension, type 2 diabetes, and heart disease. The general advice to "eat less and exercise more" oversimplifies a complex metabolic condition. Glucagon-like peptide-1 agonist (GLP-1) medications are now helping transform it.

These medications are powerful tools, but not complete cures. As telehealth platforms make GLP-1 medications more widely available, a gap emerges between the benefits of prescription access and the drawbacks of insufficient clinical support. Emerging research has found that patients are losing muscle along with fat, a risk that becomes more dangerous with age.

Telehealth Changes Who Gets GLP-1 Treatment

GLP-1 medications, including semaglutide (Ozempic and Wegovy) and tirzepatide (Mounjaro and Zepbound), have revolutionized treatment for both obesity and type 2 diabetes. They mimic hormones naturally produced by the body to suppress appetite, increase feelings of fullness, and reduce food cravings. Research suggests they have comparable effects to bariatric surgery, with an average loss of 20% of body weight in clinical trials. Beyond weight loss, research suggests they may also help with cardiovascular, kidney, and liver conditions.

Once only accessible to type 2 diabetes patients under specialist care, telehealth platforms now make them available to anyone with a BMI classifying them as overweight or obese. Real-world results back up the effectiveness of telehealth. A large study of over 50,000 participants using telehealth-delivered GLP-1 medications yielded results similar to those of major clinical trials. Participants lost 8.9% of their weight within 3 months, and 19% at 12 months. A smaller 2025 study found that participants who followed a 50-day telehealth program with liraglutide lost an average of 4.9 kilograms, with high adherence and interest in continuing treatment.

For self-pay patients, cost remains a barrier. In one study (currently available as a preprint), telehealth patients on semaglutide achieved weight loss comparable to clinical trial benchmarks, but those on tirzepatide didn't meet the same standard, potentially due to the higher cost of a clinical dosage of tirzepatide.

When Prescribing Moves Faster Than Clinical Support

Despite the benefits of telehealth for GLP-1s, research suggests that one of the biggest potential drawbacks is the lack of ongoing clinical support for patients.

A 2026 scoping review in *Obesity Reviews* examined 12 clinical trials on patients taking GLP-1 medications and their nutrient intake. While participants reduced caloric intake by 24% to 39% and saw significant weight loss, 40% of that weight was lean tissue. Only three of the 12 studies involved a registered dietitian (RD), and one study found nutrient inadequacies. The authors caution that very low caloric intake without clinical guidance

XSL•FO
RenderX

may "increase the risk of inadequate protein intake, dietary quality, and micronutrient deficiencies." The loss of lean mass exceeds the natural rate of muscle loss in older adults.

Sarah Skobeloff, MS, RD, LDN, a registered dietitian who works with patients taking GLP-1 medications, finds her clinical experience reflects this. "Appetite suppression can cause a person to consume too few calories relative to their maintenance needs. Rapid weight loss is more likely to cause muscle loss, putting my clients at risk for osteopenia and osteoporosis." She adds, "Decreased energy is often a side effect I see because if your body doesn't get enough calories, specifically carbohydrates, you will have less energy. Clients risk nutrient deficiencies like vitamin D and calcium, which are essential for maintaining bones."

Two telehealth patients reported similar experiences. Cheri says, "I did a lot of personal research. I joined Facebook support groups and learned a lot. I will say the telehealth companies need to educate you more on nutrition." She had hair loss during her treatment, which she assumes was due to rapid weight loss and insufficient protein. An anonymous patient—also a nurse practitioner (NP)—experienced fatigue, which she attributed to insufficient calories.

More Access, Insufficient Care

Increased access to GLP-1 medications can lead to major health improvements for more people, but without nutritional support, it can also increase the risk of adverse outcomes. Research highlights that the primary dietary advice during GLP-1 therapy is calorie reduction, with insufficient focus on adequate nutrition.

Skobeloff describes the tension clearly: "The major difference between telehealth and a traditional clinical setting is that the client receives interdisciplinary team care. You can't have interdisciplinary team care (PCP, cardiologist, endocrinologist) when using an online company."

Figure WL1.

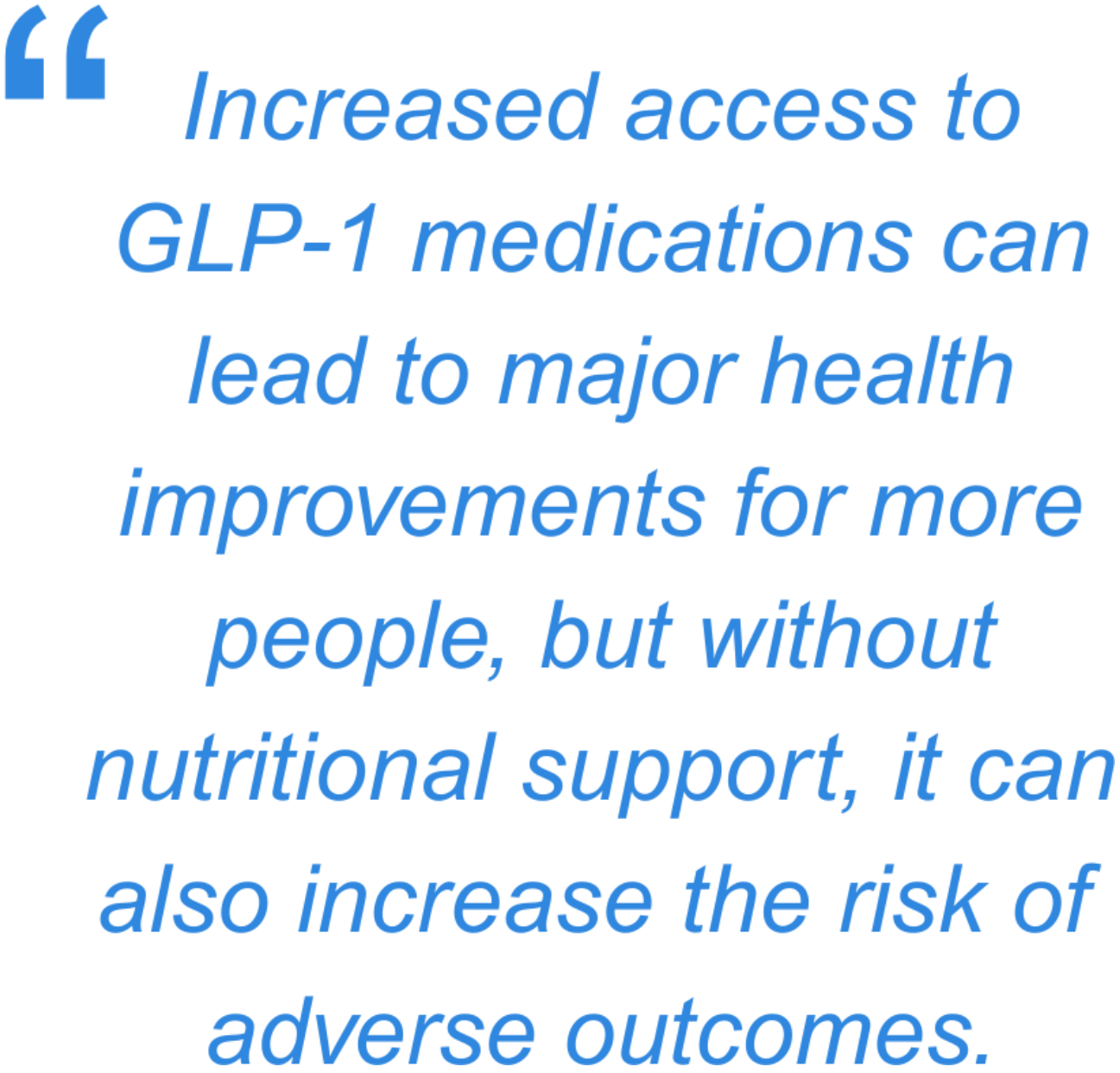

XSL•FO
RenderX

The anonymous NP, who also works at a clinic managing weight-loss patients, offers in-person and telehealth visits and illustrates the contrast. “I see patients for a 30-minute visit once a month. I cover every aspect of their weight management care.” Their husband, on the other hand, has received GLP-1 therapy through a telehealth platform for 3 years. “He has had no communication with a provider outside of a 3-minute call when the medication was initially ordered. No contact or follow-up since then.”

Looking Ahead

The authors of the *Obesity Reviews* study recommend early dietitian involvement, a nutrient-dense, high-protein diet (1.2 - 1.5 grams per kilogram of body weight per day), ongoing monitoring for nutrition deficiencies, and psychological support for emotional or disordered eating. They suggest a hybrid model to pair telehealth prescriptions with clinical support and reimbursement for seeing a dietitian.

Some platforms are already on board. One telehealth patient, Lisa, uses an app that includes monthly visits with an RD. “She answers all my questions and gives me incredible guidance on what to eat and when.” Lisa shares, “The medication itself is a great motivator and tool, but it’s not a miracle cure.”

Skobeloff agrees and adds that nutrition support is only part of the equation for muscle retention. “Providers need to rely on their allied health professionals to deliver the best patient care. They need to engage with dietitians to prevent nutrient deficiencies, with exercise physiologists to ensure weight-lifting is programmed correctly, and with therapists to ensure their relationships with their body and food are strong. A supportive community results in better patient outcomes. Prescribers need a patient-centered approach, considering all aspects of health, not just weight.”

More access to GLP-1 medication through telehealth is helping people achieve meaningful changes that may have felt impossible before. To make that change sustainable, we need integrated clinical care to support that access.

Edited by KA Clegg; submitted 20.May.2026; this is a non–peer-reviewed article;accepted 20.May.2026; published 28.May.2026.

Please cite as:
Zucker A
After the Prescription: The Clinical Support Gap in Telehealth-Based GLP-1 Care
J Med Internet Res 2026;28:e101874
URL: https://www.jmir.org/2026/1/e101874
doi:10.2196/101874

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Can Digital Tools Fix Bias in Mental Health Triage?

Beth Rush

Abstract

Recent advances in digital technologies hold both promise and potential pitfalls for mental health triage. In this *News and Perspectives* article, JMIR Correspondent Beth Rush reports on the issue of bias and the role of digital tools in the mental health context.

(J Med Internet Res 2026;28:e100947) doi:10.2196/100947

KEYWORDS

mental health; digital technologies; internet; mobile applications; telemedicine; artificial intelligence; algorithmic bias; bias; health equity; clinical decision support systems; human factors; human-in-the-loop

Textbox .

Key Takeaways:

- Digital tools can create standardization and transparency in mental health triage, but they can also inherit and even amplify bias.
- The most effective solution is a hybrid model that combines technology with human expertise.

The growing intersection of mental health and digital technologies is increasingly positioned as a solution to one of health care's most persistent barriers to health equity—biased triage.

Health systems under pressure are adopting algorithmic tools to improve consistency, reduce subjectivity, and scale access. The logic is compelling. Structured systems should, in theory, produce more equitable outcomes than human judgment alone. Yet, the reality is more complicated: while digital technologies offer meaningful advantages for mental health, they also reflect the very inequities they aim to resolve.

The Bias Problem

Bias in discussions of mental health and digital technologies often begins with the human side of the equation. Clinical triage is conducted in environments defined by time constraints, incomplete information, and high stakes. Under these conditions, decision-making frequently relies on heuristics, which can introduce conscious and unconscious biases.

A 2026 study conducted by researchers at Ann & Robert H. Lurie Children's Hospital of Chicago provides a clear example. In its analysis of over 74,000 pediatric emergency department visits, researchers found that undertriage—assigning a lower severity score than the level of care a child actually needed—was more likely for children who were Black or Hispanic or those who preferred Spanish over English.

Jennifer Hoffmann, MD—the study's lead author—explains that these disparities have multiple drivers. "The disparities we observed may be driven by a combination of implicit bias, communication barriers, and structural factors within the healthcare system," she says. "For example, clinicians may unconsciously perceive symptoms differently based on race or ethnicity, and underuse of professional interpreters can lead to miscommunication for Spanish-speaking families. These factors, along with broader systemic inequities, can influence how urgency is assessed during triage."

Empirical research has repeatedly shown disparities in how patients are assessed and treated. Race, gender, and socioeconomic status influence diagnostic outcomes, access to care, and perceived risk levels. These patterns are rooted in long-standing structural inequalities that shape the entire mental health ecosystem. Any attempt to address bias must contend with these systemic realities rather than treat bias as an isolated issue.

XSL•FO
RenderX

Figure WL1.

While digital technologies offer meaningful advantages for mental health, they also reflect the very inequities they aim to resolve.

The Role of Clinical Decision Support

Triage platforms—digital clinical decision support tools designed to standardize and support clinical decision-making—appear promising for mitigating bias. They include artificial intelligence (AI)–driven risk prediction models, structured digital intake platforms, and symptom assessment tools that guide patients and clinicians through consistent workflows. By applying the same criteria across cases, these systems aim to limit the influence of individual judgment. They can prioritize patients based on risk, flag urgent cases, and suggest care pathways, particularly in settings where specialist expertise is limited.

Sam Zand, MD—a psychiatrist and CEO of Anywhere Clinic—notes the complexity of mental health triage. “When evaluating mental health patients in the emergency department, it can often be difficult to assess the level of risk related to care of these patients,” he explains. “Misclassification of risk will lead to either an under-triaged patient and/or over-triaged patient—both of which have dire consequences regarding patient safety and staff resources.”

Kenneth Perry, MD—an emergency physician in South Carolina—sees potential for AI to improve these challenging decisions. “AI tools would be very helpful in determining which patients need inpatient medical care and which are safe enough for outpatient follow up,” he says. “If there was a way to keep the bias of the questioner from the interaction, more correct, efficient dispositions could be possible.”

The Challenge of Algorithmic Bias

Despite their potential, however, the effectiveness of these tools is fundamentally constrained by data quality. These systems are trained on historical datasets that often reflect the very disparities found in the Ann & Robert H. Lurie Children’s Hospital of Chicago study.

Mai Uchida, MD—a pediatric psychiatrist and AI investigator at Harvard University and the NTT Research Brain Science Center, and director of pediatric depression at Massachusetts

XSL•FO
RenderX

General Hospital—explains how bias becomes embedded in these systems. “If historical data reflects disparities—such as underdiagnosis of internalizing disorders in boys, overdiagnosis of behavioral disorders in Black children, or differential access to care across socioeconomic groups—then AI may not only reproduce these patterns but reinforce them with a veneer of objectivity,” she says.

Flawed assumptions can get amplified. According to Dr Saravanan Thirumalai Thangarajan, visiting scientist at Harvard T.H. Chan School of Public Health and international fellow at Ariadne Labs, Brigham and Women’s Hospital, “AI triage tools in pediatric mental health do not neutralize bias. They can industrialize it. Many triage models use prior healthcare utilization as a proxy for need, but utilization reflects access, not illness burden. A model trained on unequal access patterns does not simply detect risk. It learns who the system already knows how to see.”

Figure WL2.

The Value of Standardization and Transparency

Yet, these tools have practical value that shouldn’t be dismissed. When designed and implemented carefully, they can reduce certain types of bias and strengthen triage systems.

One of their most important contributions is standardization. Digital intake tools ensure that every patient is asked the same core questions consistently. This reduces variability in data collection and limits the influence of interpersonal dynamics during initial assessments.

Transparency is another key advantage. Algorithmic systems can be stress-tested and calibrated, and decisions made with digital systems can be analyzed at scale, enabling the detection of patterns of disparity that might otherwise be difficult to identify in traditional settings. This allows organizations to audit

XSL•FO
RenderX

performance, iteratively refine models, and address bias more systemically.

Corrective measures can also be explicitly and proactively incorporated into digital triage tools. Marie Deschamps, PhD, LPC-A, ATR-P—a clinician-scientist, digital health innovator, and visiting scholar at The New School—emphasizes the importance of capturing diverse forms of patient expression. “Integrating multimodal data streams—including images, voice, and narrative—into AI-supported triage systems offers a critical pathway to reduce bias in mental health care by honoring culturally diverse, non-verbal, and metaphor-based expressions of distress rather than forcing them into standardized, and often inequitable, clinical categories,” she explains.

Toward a Human-in-the-Loop Model

Digital tools cannot fully eliminate bias in mental health triage. The issue is too deeply embedded in broader social and institutional structures.

Mental health triage often involves ambiguity, requiring contextual understanding and ethical judgment. These are areas where human expertise is indispensable. At the same time, it also requires consistency and standardization, which is precisely why digital support systems can be valuable.

The most effective approach, then, is a hybrid, or human-in-the-loop, model. In this framework, digital triage tools provide standardized inputs and analytical support, while clinicians apply judgment and context. This balance allows for consistency and adaptability in decision-making.

Hoffmann points to concrete steps that can be implemented immediately. “I think there are several practical steps emergency departments can take right away, including improving access to and use of professional interpreters during triage and standardizing how interpreter needs are identified,” she says. “Even small workflow changes can make a meaningful difference in promoting more equitable triage decisions.”

Edited by KA Clegg; submitted 10.May.2026; this is a non–peer-reviewed article;accepted 10.May.2026; published 28.May.2026.

Please cite as:
Rush B
Can Digital Tools Fix Bias in Mental Health Triage?
J Med Internet Res 2026;28:e100947
URL: https://www.jmir.org/2026/1/e100947
doi:10.2196/100947

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Meet the New Health Care Gatekeeper: Your Wearable

Blythe Karow

Abstract

Consumer wearable platforms are increasingly moving into the health care space. In this *News and Perspectives* article, JMIR Correspondent and MedTech expert Blythe Karow reports on the potential implications for patient trust, policy, and regulation.

(J Med Internet Res 2026;28:e101881) doi:10.2196/101881

KEYWORDS

artificial intelligence; wearable electronic devices; remote sensing technology; delivery of health care; integrated; patient referral; reimbursement mechanisms; health policy; legislation, medical; antitrust laws; consumer health technology; clinical decision pathways; data ownership; platform regulation; commercial determinants of health

Textbox 1.

Key Takeaways:

- Consumer wearable platforms—rather than physicians—are starting to own the first conversation about health changes, influencing which specialist a patient sees and which treatments they consider.
- The policy and regulatory frameworks designed to protect patient trust were not built for a configuration in which the same company holds the user's daily physiological data, interprets it through AI, and routes them toward a reimbursable clinical program it also bills for.
- If this consolidation continues, antitrust scrutiny becomes a legitimate question. American health care already prohibits physicians from financially benefiting when they refer patients to specific specialists or facilities. A platform that owns monitoring, interpretation, clinical routing, and reimbursement under one roof has not yet faced that same structural scrutiny.

*Blythe Karow is a strategic management consultant and founder of The Karow Advisory Group. She writes*The Device Files*, a Substack on MedTech strategy*.

In early April, the fitness wristband company WHOOP closed a $575 million funding round at a $10+ billion valuation. The investor roster this time expanded beyond professional athletes and the expected funds to include Abbott and Mayo Clinic.

Abbott and Mayo Clinic are not consumer wellness investors. Abbott's portfolio spans diagnostics, nutrition, and pharmaceuticals, while Mayo is one of the most credentialed clinical institutions in the country. Their presence on the cap table was a signal, for anyone paying attention, that wearables have stopped being a consumer tech play.

The next week, WHOOP made an announcement that started to clearly define that signal. A newly created affiliate called WHOOP Physician Services, PC, was selected into the first cohort of Medicare's new ACCESS program, an outcome-based chronic care model launching in July. WHOOP, a company that spent most of 2025 in a public dispute with the Food and Drug Administration over whether its blood pressure feature qualified as a medical device, is now a Medicare-enrolled health care provider.

WHOOP's recent strategic moves are giving us a clear picture of where the American (and maybe global) health care landscape is likely heading—and fast.

Who Owns the First Conversation?

For decades, the entry point to medical care was the primary care physician (PCP). Your PCP decided what specialists you saw, what tests you took, what treatments you tried. Insurance networks, hospital systems, specialty referrals, and pharmaceutical distribution all organized around that referral relationship. The professional and economic architecture of American health care assumed the PCP was the first conversation.

But now a completely different category of company has been building toward accumulating something no health care incumbent possesses: rapt daily user attention on continuous physiological data. Consumer wearable platforms know when their users' sleep is fragmenting, when heart rate variability is drifting, when blood pressure trends are climbing. They often know a user's health is changing before the user does. And they own the first conversation about what to do about it.

That conversation used to happen in an exam room. Then people started Googling their symptoms and bugging their doctor with what they found. The next step has been ChatGPT and other large language models offering interpretation of data and test results for patients; in almost a flash, this has evolved into the wearables incorporating their own artificial intelligence to communicate within their app to the user.

And you know what? Whoever owns that first conversation is the new medical gatekeeper. They influence which specialist

XSL•FO
RenderX

the user first considers, which products the user first considers buying to treat the problem, which care program the user might enroll in. That gatekeeper owns the first conversation and they know the value of that ownership. That gatekeeper can monetize and shape billions of dollars of downstream spending without having to deliver most of it.

Figure WL1.

Whoever owns that first conversation is the new medical gatekeeper.

The New Ecosystem

WHOOP is the latest visible example, but not a unique one. Oura, the smart ring maker, raised $900 million at an $11 billion valuation last October and is actively delivering on its White House commitment to integrate with Medicare's data infrastructure, announcing electronic health record integration on the same day WHOOP announced its Medicare entry. Apple, Samsung, Withings, and Alphabet's Verily are also actively engaging with the CMS Health Tech Ecosystem Pledge, each building out the clinical, regulatory, or reimbursement infrastructure that sits between a consumer wearable and the Medicare system. Taken together, these moves are signaling a significant change in the American health care landscape.

The wearable category is not competing to be a better fitness tracker anymore; instead, it's competing to be the routing layer for clinical care.

This is a major shift, one with potential upsides and downsides.

The current health care system is not well served by the status quo. Doctors are overworked and spread thin and many patients aren't proactive enough or face barriers to following up on concerning symptoms or adhering to treatment protocols. Wearables, and their apps, are well-positioned to help with a lot of that.

A Cautionary Note

But beware who we put all of our faith in. We already know it's probably a bad idea to give ChatGPT pictures of our lab results and have it interpret them, but many people are doing just that. Wearables companies came up in a culture of monetizing data, your data, and the one you're wearing right now is legally selling some data about you.

Pharmaceutical and MedTech companies profit from adding patients to treatment funnels that ultimately end in purchases of their product. But they're not controlling the funnel, monetizing the user data flowing through it, and getting to suggest which treatment a patient is routed toward. Wearable platforms are positioned to do all three at once, legally, today.

The wall between what they learn about a user as a customer and what they do with that information when they become a clinical intermediary depends entirely on how each company decides to enforce it. The concern is that wearables came up in a different culture. Their business models are built on user attention, subscription revenue, and the value of user data.

Consumer technology has spent twenty years learning how to route user attention toward outcomes that benefit the platform. When the outcome is which specialist a user sees, which prescription a user fills, or which Medicare-reimbursed program

XSL•FO
RenderX

a user enrolls in, that optimization logic meets clinical judgment on terms the health care system has not previously had to confront.

The American health care system has spent decades waiting for disruption to come from better doctors, better hospitals, or better insurance. It is coming—but from a direction we're not necessarily prepared for, and it is arriving faster than the policy and regulatory frameworks around patient trust were built to handle.

Edited by KA Clegg; submitted 20.May.2026; this is a non–peer-reviewed article;accepted 20.May.2026; published 29.May.2026.

Please cite as:
Karow B
Meet the New Health Care Gatekeeper: Your Wearable
J Med Internet Res 2026;28:e101881
URL: https://www.jmir.org/2026/1/e101881
doi:10.2196/101881

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Affordable GLP-1? When Digital Platforms Meet Policy Reform

Xiangming Jenny Zhan

Abstract

Recent shifts in US policy and digital platforms are making obesity medications more affordable. In this *News and Perspectives* article, JMIR Correspondent Xiangming Jenny Zhan reports on this emerging access model and its potential implications.

(J Med Internet Res 2026;28:e102069) doi:10.2196/102069

KEYWORDS

glucagon-like peptide-1 receptor agonists; telemedicine; drug costs; medication adherence; insurance, health; chronic disease; health services accessibility

Textbox .

Key Takeaways:

- Shifts in pricing, use of direct-to-consumer platforms, and digital prescriptions have made GLP-1 medications more affordable without insurance in the United States, opening a new access pathway.
- Early real-world data show rapid uptake but also high discontinuation rates within the first year, suggesting that affordability alone may not be sufficient to sustain long-term treatment outcomes.
- This model is already extending to other chronic disease medications, but its long-term viability remains to be seen.

Glucagon-like peptide-1 (GLP-1) receptor agonist drugs such as semaglutide and tirzepatide have shown 15% - 20% weight loss in clinical trials, but, up until recently, at $1000-$1349 per month, they remained out of reach for patients without generous insurance coverage in the United States. Most commercial plans did not cover obesity indications, meaning that even insured patients often had to pay the full cost out of pocket. In early 2026, however, a convergence of factors brought monthly prices below $350 without insurance, opening a new access pathway that operates largely outside the traditional insurance-based model. This shift did not come from a single policy or platform, but from three distinct forces arriving at the same moment.

A New Pathway to the Patient

The first force was a shift in federal pricing policy. The Trump administration negotiated Most-Favored-Nation (MFN) pricing deals with Novo Nordisk, Eli Lilly, and other manufacturers, tying tariff exemptions to price concessions. Monthly costs for GLP-1s dropped to approximately $350, and oral formulations to $149 at starting doses. At these price points, out-of-pocket payment became a realistic option for patients who previously could not afford treatment.

The second force was initiation of manufacturer-led digital distribution. Novo Nordisk and Eli Lilly launched direct-to-consumer platforms that bypass insurers and pharmacy benefit managers, handling prescriptions, payment, and delivery digitally. As of early 2026, Novo Nordisk was routing much of its oral Wegovy demand through self-pay rather than insurance, led by its direct-to-patient pharmacy NovoCare and telehealth partners. Eli Lilly's LillyDirect reached over 1 million patients in 2025, with self-pay Zepbound vials as its most used product.

The third force was integration into telehealth prescribing. Platforms like Ro, Hims & Hers, and LifeMD integrated GLP-1 prescribing into their workflows, enabling patients to consult a provider and receive medication without visiting a clinic. Hims & Hers alone reached over 2.5 million subscribers in 2025. Costco, through its partnership with Sesame, and Amazon One Medical have also entered the space, extending digital GLP-1 access beyond telehealth startups into mainstream retail.

These three moves came together in a single federal entry point when the Trump administration launched TrumpRx.gov on February 5, 2026. The platform aggregates manufacturer discount programs, connecting cash-paying patients directly to reduced prices without requiring insurance. For medications not widely covered by insurance, such as obesity and fertility treatments, TrumpRx offers a lower-cost alternative to traditional pharmacy channels, though for many other drugs, insurance co-pays may still be cheaper than cash-pay prices.

Access Without Continuity

Lower prices have expanded who can start GLP-1 therapy, but starting treatment is not the same as sustaining it. A 2025 population-based study of over 77,000 first-time semaglutide users in Denmark found that more than half stopped treatment within 12 months, with cost, gastrointestinal side effects, and lack of ongoing clinical support cited as key drivers. Patients who discontinue use return to their baseline weight within approximately 1.7 years. These findings suggest that the central

XSL•FO
RenderX

challenge has shifted: the barrier is no longer getting patients onto GLP-1 therapy, but keeping them on it.

Part of the problem is structural. The cash-pay channels that have made GLP-1s more accessible operate outside the insurance system entirely. Purchases through TrumpRx and manufacturer platforms do not count toward insurance deductibles or out-of-pocket maximums, offering no cumulative financial protection for patients who need treatment indefinitely. Meanwhile, insurance coverage for obesity also remains limited: only 34% of employers surveyed covered GLP-1s for both diabetes and weight loss in 2024, up just 8 percentage points from the year before. Patients are left choosing between an affordable but isolated cash-pay channel and an insurance system that largely does not cover the treatment at all.

Even the direct-to-consumer platforms themselves face an uncertain foundation. Hims & Hers, which reached $2.35 billion in revenue in 2025, built much of its growth on compounded GLP-1 products, lower-cost alternatives produced during branded drug shortages that have not received Food and Drug Administration (FDA) approval. In early 2026, Novo Nordisk sued Hims & Hers for patent infringement, and the FDA moved to restrict compounded semaglutide shipping, creating a $65 million revenue headwind for Q1 2026. If the largest telehealth GLP-1 provider is forced to exit compounded products, the access gains it delivered may prove fragile.

Beyond GLP-1s

Despite these vulnerabilities, the momentum behind this model is accelerating. Novo Nordisk announced a further 35% to 50% list price reduction for its semaglutide products in February 2026, effective January 2027. And the policy framework is expanding beyond GLP-1s: insulin is available at $25 per month, AstraZeneca's chronic obstructive pulmonary disease inhaler dropped from $458 to $51, and 16 pharmaceutical companies have now signed MFN agreements.

Figure WL1.

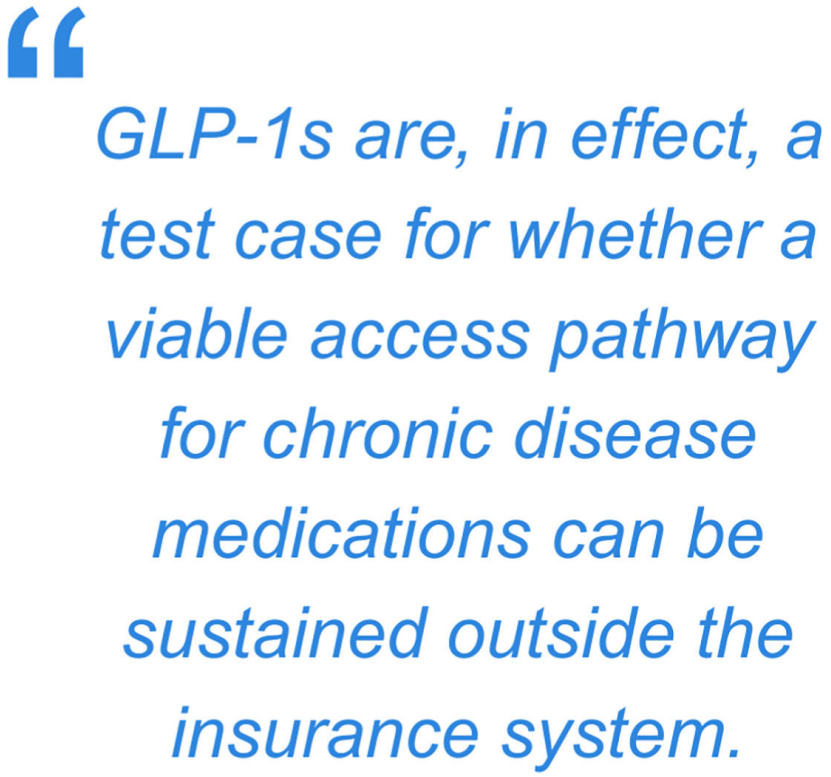

Whether this model can be replicated for other chronic conditions is less certain. GLP-1s benefited from a specific combination of factors: insurance coverage for obesity was already so limited that bypassing it carried little downside, patient motivation to seek and pay for weight-loss treatment was unusually high, and intense branded competition gave manufacturers a reason to build direct-to-consumer channels and cut prices. These dynamics may persist in the obesity space, where an estimated 16 new medications are expected to gain approval between 2026 and 2029, but for chronic disease managed by inexpensive generics with adequate insurance coverage, neither patients nor manufacturers have comparable incentive to move outside the existing system.

GLP-1s are, in effect, a test case for whether a viable access pathway for chronic disease medications can be sustained outside the insurance system. How far that pathway extends will depend on whether the infrastructure built for GLP-1s can support not just affordable first prescriptions, but the continuous care that chronic disease treatment requires.

XSL•FO
RenderX

Edited by KA Clegg; submitted 21.May.2026; this is a non–peer-reviewed article;accepted 21.May.2026; published 01.Jun.2026.

Please cite as:
Zhan XJ
Affordable GLP-1? When Digital Platforms Meet Policy Reform
J Med Internet Res 2026;28:e102069
URL: https://www.jmir.org/2026/1/e102069
doi:10.2196/102069

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 1.Jun.2026.

XSL•FO
RenderX

Remote Patient Monitoring Is Scaling Without a Clear Organizational Model: Insights From the Netherlands

Vincent Peters

Abstract

With hospital systems under increasing strain, new digital care models are being rapidly deployed. In this *News and Perspectives* article, JMIR Correspondent and researcher Vincent Peters reports on remote patient monitoring initiatives in the Netherlands and the organizational structures needed for effective implementation.

(J Med Internet Res 2026;28:e101887) doi:10.2196/101887

KEYWORDS

remote patient monitoring; virtual care center; telemonitoring; organizational model; fragmentation

Textbox .

Key Takeaways:

- Remote patient monitoring has emerged as a labor-saving innovation.
- Scaling remote patient monitoring is an organizational challenge, not just a technological one.
- There is no single best organizational model: organizational choices reflect priorities.

Vicent Peters, PhD, is an assistant professor at the Department of Information Systems and Operations Management at Tilburg University. His research is focused on the digitization of health care, especially on how to organize and integrate remote care with in-person care and how this impacts health care operations and outcomes. In this piece, he highlights the tension between innovations in remote care and the slower evolution of organizational structures needed to support them in the Netherlands, based on his own research and informal conversations with hospital stakeholders.

Health care systems worldwide are facing growing workforce shortages. The pressure on health care professionals continues to increase, while waiting times for patients are rising. These shortages are projected to worsen globally in the coming years. As the chief medical information officer of a Dutch hospital described it, the situation resembles “A car speeding toward a stationary wall requiring action from those behind the wheel.” Clearly, labor-saving innovations are no longer a choice but a necessity. Remote patient monitoring (RPM) has emerged as one such innovation.

The Rise of Remote Patient Monitoring

Over the past decade, RPM has increasingly complemented traditional in-person care, enabling patients to be monitored in their homes while health care professionals oversee their condition remotely. The COVID-19 pandemic accelerated RPM adoption at an unprecedented pace, and importantly, this shift has proven durable rather than temporary. To date, many global health care systems maintain or expand their RPM offerings in the postpandemic period.

XSL•FO
RenderX

Figure WL1.

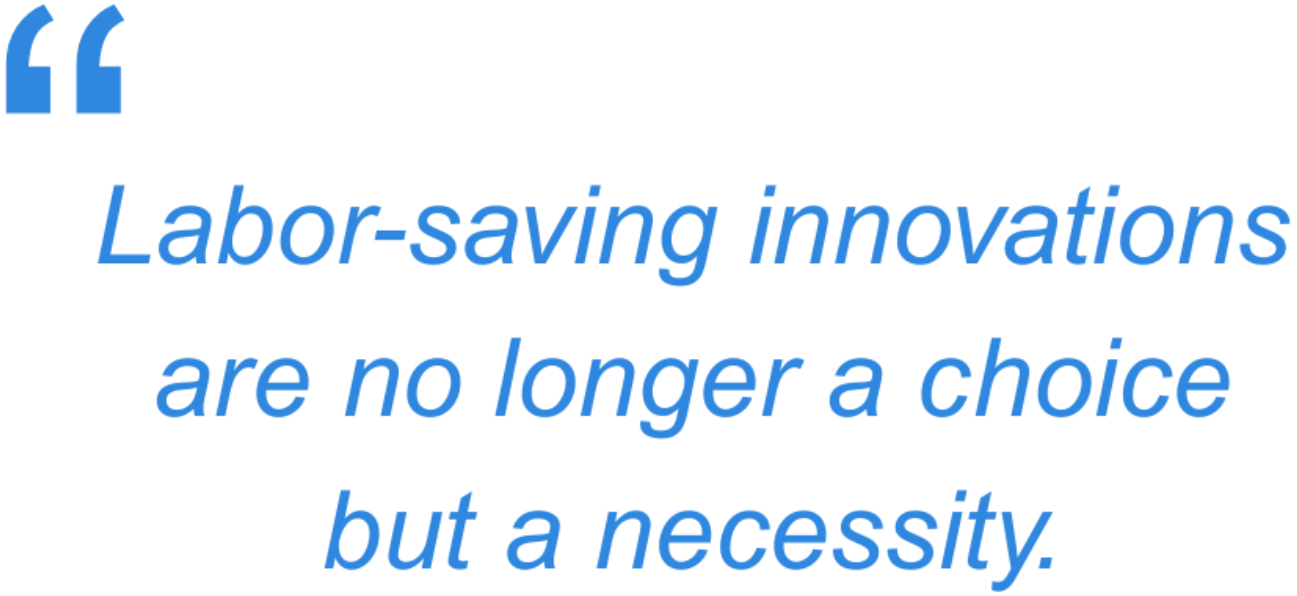

However, despite its rapid expansion, there is no shared understanding of how the organizational structures underpinning RPM should be designed, governed, or embedded within hospital operations. In practice, this leads to fragmented approaches, duplication of efforts, and missed opportunities for learning across hospitals. As the head of a centralized Dutch virtual care center mentioned, “We organize it [RPM] in this way, because we believe in it, not because we have evidence for it,” indicating that they just went for it. In many cases, RPM is done alongside and on top of regular work, limiting its true potential to generate meaningful outcomes.

This ambiguity is also reflected in the literature. While there is growing evidence on the clinical effectiveness of RPM, far less attention has been paid to its organizational design. In the absence of clear guidance, decisions are often made based on local clinical or administrative judgment rather than evidence-based approaches. As a result, hospitals are left to define their own organizational solutions.

Emerging Organizational Models

In the absence of a shared understanding, four organizational models for organizing RPM have emerged in Dutch practice. These models differ in where monitoring activities take place, how responsibilities are distributed, the clinical proximity to the patient, and the extent to which care is outsourced.

Decentralized Within the Hospital

In this organizational model, RPM is organized at clinical departments within the hospital, in which health care professionals monitor their specific patient population. This more traditional approach is often referred to as telemonitoring. RPM activities are added to day-to-day practices, offering clinical integration and continuity of care. Rather than outsourcing to dedicated telenurses, these activities are managed by professionals embedded within the specific clinical department, who are already familiar with the patient’s history and treatment context. It can also support professional development and ownership among staff. However, since RPM is performed alongside regular duties, 24-7 coverage is difficult and creates dependency on a small number of professionals.

Centralized Within the Hospital

Here, RPM activities are consolidated into a single hospital-wide unit. This so-called virtual care center (or medical service center) is an organizational entity that uses digital technologies to remotely monitor patients, interpret incoming patient health data, and deliver and coordinate medical care within the patient’s home environment. This unit allows for monitoring across specialties and is handled by a dedicated team, allowing for greater standardization and potentially faster response times. This centralized model within the hospital can improve efficiency, enable workforce specialization, and create the base for scaling RPM across the hospital.

The same centralized virtual care center head illustrated, “Our virtual care center allowed us to scale up remote patient monitoring in ways that we could have never imagined. We started small with two pilots and now offer more than 50 care pathways in a remote way...and these numbers ensure that we can staff our center and expand our opening hours to serve our patients in the best possible way.”

However, this model requires coordination and up-front investment, including new workflows, new roles for those who staff the virtual care center like telenurses, and physical/virtual infrastructures. Clinical distance from clinical departments may also pose challenges, and continuous monitoring is not always feasible without additional arrangements.

Outsourced Centralized Model

In this model, all RPM activities are managed by an external virtual care center, which monitors patients on behalf of one or more hospitals. This model has the potential for standardization of care pathways, combined with centralized coordination and digital infrastructure, allowing monitoring activities to be organized more uniformly across patient groups. However, outsourcing reduces direct control and increases clinical distance to the patient. Telenurses staffing the external virtual care center

XSL•FO
RenderX

might not be aware of the patient's local context, which can affect how signals are interpreted.

As the project manager of a Dutch outsourced centralized model noted, tongue-in-cheek, patient behavior and symptom reporting can vary significantly by region: "You hear differences in the Netherlands, because in Groningen, when someone is 'not feeling well,' they are lying half-dead on the couch. While in Amsterdam, 'not feeling well' means immediately going to the hospital." Without this contextual familiarity, there is a risk of misinterpretation or overstandardization. In addition, this outsourced model creates dependency on external partners and requires agreements on data sharing, clinical responsibility, and quality assurance.

Hybrid Model

In a hybrid model, hospitals retain responsibility for RPM in some care pathways, while outsourcing others to an external virtual care center. Outsourcing is typically done for pathways that require continuous or high-intensity monitoring, such as acute or perioperative monitoring. This model allows for greater flexibility in how RPM is organized, for example, by tailoring the approach per care pathway. It can help address capacity constraints by shifting workload but also increases complexity and requires clear agreements on responsibilities and data exchange.

Choosing the "Right" Model

Picking the right organizational model for RPM is not an easy task. Different models reflect different priorities, and no single model optimizes all outcomes. Hospitals may prioritize differing outcomes, and models support these aims differently. For example, decentralized models may strengthen professional ownership and continuity, while centralized or outsourced models may favor efficiency and scale.

At the same time, choices are shaped by contextual factors, including hospital size and available resources, the characteristics of the patient population, and the nature of care pathways. For instance, stable, low-complexity conditions may not require close clinical proximity, whereas more complex cases may benefit from stronger integration within clinical departments. Geographic context also plays a role, as regional density and opportunities for collaboration influence whether RPM can be organized locally or across hospitals.

Figure WL2.

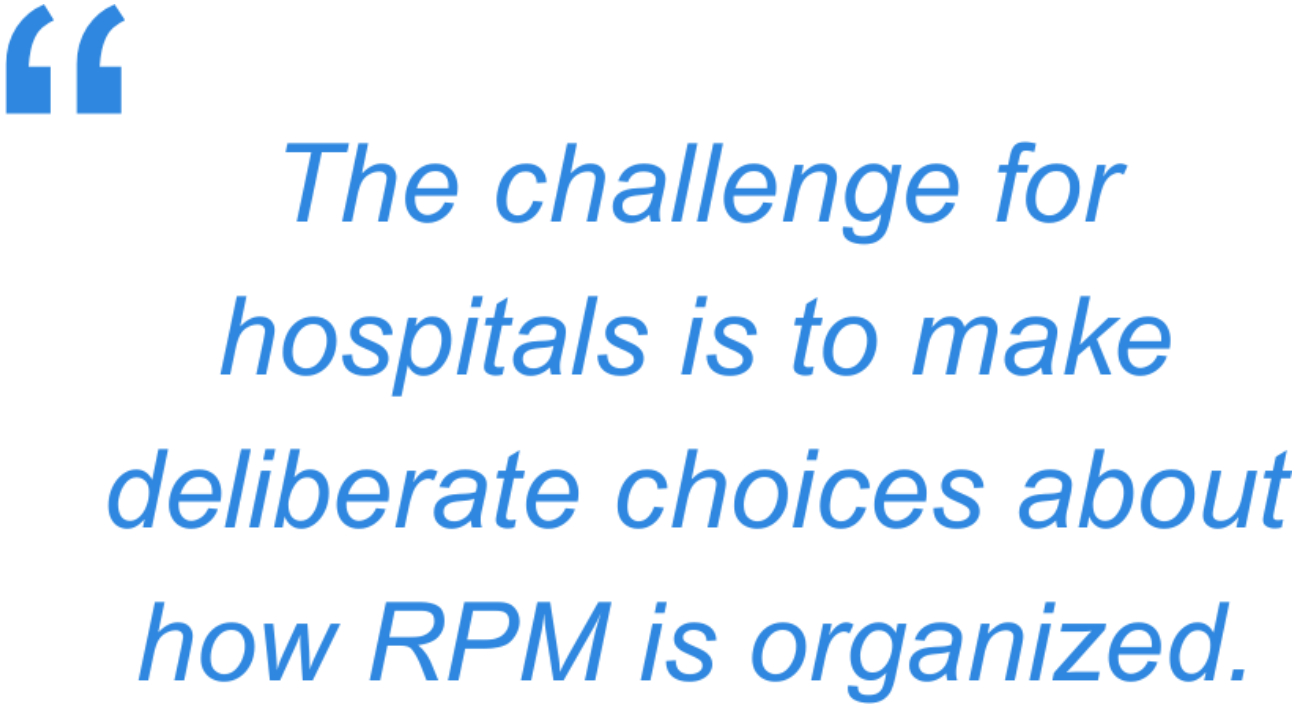

Hence, the key question is not which model is objectively *best*. Nor should RPM be treated as a goal in itself. Rather, the challenge for hospitals is to make deliberate choices about how RPM is organized. The four models outlined here provide a practical lens to structure these choices and make trade-offs explicit. Different priorities will naturally lead to different models, and there is no one-size-fits-all solution. However, this makes it all the more important to adopt a structured approach, rather than relying on ad hoc decisions. Otherwise, RPM risks remaining stuck in experimentation and fragmentation.

XSL•FO
RenderX

Edited by KA Clegg; submitted 20.May.2026; this is a non–peer-reviewed article;accepted 20.May.2026; published 28.May.2026.

Please cite as:
Peters V
Remote Patient Monitoring Is Scaling Without a Clear Organizational Model: Insights From the Netherlands
J Med Internet Res 2026;28:e101887
URL: https://www.jmir.org/2026/1/e101887
doi:10.2196/101887

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

From Virtual Molecules to Clinical Trials: How AI Is Reshaping Preclinical Drug Discovery

Benedette Cuffari

Abstract

Pharmaceutical development is typically a long and arduous process, but advances in artificial intelligence may help streamline each stage of this pipeline. In this *News and Perspectives* article, JMIR Correspondent Benedette Cuffari reports on recent innovations and their potential implications for drug discovery.

(J Med Internet Res 2026;28:e101366) doi:10.2196/101366

KEYWORDS

artificial intelligence in drug discovery; machine learning in drug design; generative chemistry; de novo drug design; computational drug discovery

Textbox .

Key Takeaways:

- Artificial intelligence is transforming preclinical drug discovery by enabling more efficient identification, design, and optimization of drug candidates through machine learning, deep learning, and generative modeling approaches.
- Virtual screening, molecular docking, and de novo design methods are increasingly integrated into drug discovery pipelines to improve prediction of drug-target interactions and accelerate lead optimization.

From "bench to bedside," drug discovery and development is a highly complex process that typically takes 10 to 15 years or more before a new molecule reaches the market. Drug development typically begins with the identification and validation of a biological target, such as a disease-associated receptor, enzyme, or signaling protein. Thereafter, candidate molecules are subjected to in vitro high-throughput screens (HTS) to optimize drug activity, target affinity, and toxicity profiles before a compound undergoes extensive preclinical and clinical testing. Collectively, the development of a single drug molecule is valued at an average of US $2.6 billion, with these high costs attributed to the sequential nature of each stage.

Recent advancements in artificial intelligence (AI) technologies are enabling faster, data-driven identification and optimization of drug candidates, with the potential to reduce both development time and costs. These advancements are transforming every stage of drug discovery, from target identification to molecule design and optimization.

Machine Learning in Drug Discovery

Machine learning (ML) is a subset of AI that enables systems to learn labeled or unlabeled data from chemical libraries, biological assays, and clinical trials; predict outcomes; and generate actionable insights. ML methods were originally introduced into the drug discovery process in the 1980s and 1990s, during which random forests, support vector machines (SVMs), and other ML algorithms were utilized to model layered relationships between molecular structure and biological activity based on learned datasets.

By the early 2000s, deep learning (DL) algorithms capable of processing vast amounts of data enhanced the ability of these systems to predict structural and molecular properties. Specifically, biological and chemical data acquired from HTS and other multiparameter analyses are analyzed to model and predict drug-target interactions. Modern DL tools involved in drug discovery include graph neural networks—a powerful molecular modeling system that predicts solubility and protein-ligand interactions—whereas transformer architectures model long-range dependencies and contextual relationships during target/drug prioritization.

About 90% of drug candidates fail during phase I, II, or III clinical trials, with unsuccessful outcomes primarily attributed to a lack of clinical efficacy, as well as unmanageable toxicity, suboptimal drug-like properties, and poor strategic planning. To improve the safety of new medications, AI models, like quantitative structure-activity relationship (QSAR) modeling and DL, have been developed to assess a wide range of absorption, distribution, metabolism, excretion, and toxicity features. ChemMORT, for example, is a recent DL-based analytical platform that relies on a simplified molecular input line entry system (SMILES) encoder, descriptive decoder, and molecular optimizer to better predict how drug candidates will behave within biological systems.

Why Drug Development Remains Expensive

Despite recent developments in computing power, pharmaceutical innovation has not become proportionally faster

XSL•FO
RenderX

or cheaper. This contrast is often described as Eroom's Law—the observation that drug development becomes more expensive over time, even as digital technologies rapidly advance.

Mark Gerstein, PhD—professor of biomedical informatics, molecular biophysics, and computer science at Yale University—said that this paradox helps explain why enthusiasm around AI should be tempered with realism. "There is a lot of excitement with AI in drug discovery," Gerstein said, "but the reality is that we have not been as successful at discovering drugs as we would have hoped."

Computers continue to improve under Moore's Law—the approximate doubling of transistors on microchips every 2 years—and the cost of genome sequencing has fallen dramatically. Nevertheless, these efficiencies have not eliminated the biological and clinical challenges that drive failure rates and development costs.

Figure WL1.

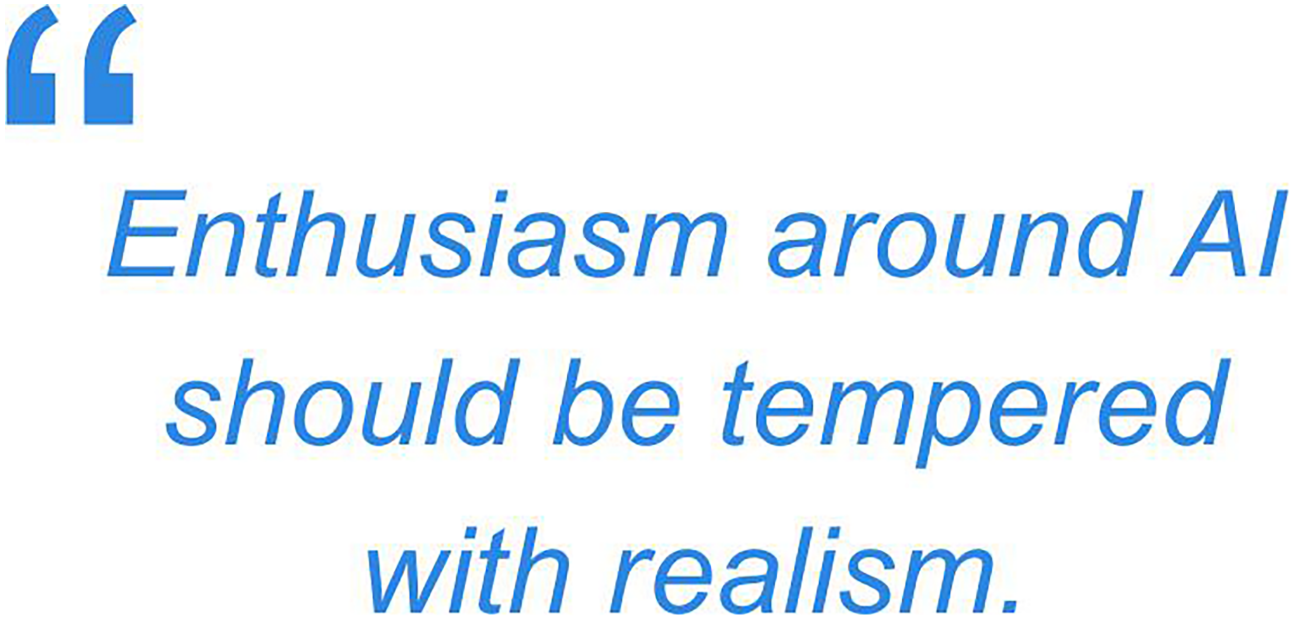

Structure-Based AI Design

Generative chemistry, otherwise known as "drug design," involves the creation of novel molecular compounds intended to be stable, valid, and unique drugs with pharmaceutical potential. Using both ML and DL algorithms, virtual screening is frequently involved in the early stages of drug development to learn the structures and properties of enormous compound libraries and predict their potential interactions with a specified drug target.

Structure-based virtual screening (SBVS), for example, leverages molecular docking to identify drug candidates based on their predicted interactions with a target protein. ML algorithms, like SVMs and random forest models, predict ligand binding within the protein's active site, allowing medicinal chemists to rationally modify functional groups to improve binding affinity.

Gerstein also points to equivariant neural networks as a promising advance in AI-driven drug discovery. Since molecules obey physical laws and can rotate, stretch, or change their conformation, predictive models must account for these transformations. Equivariant architectures are designed to preserve these relationships, making them especially valuable for molecular modeling and binding prediction.

Compared with ML-based methods used to create drug-like candidates based on existing chemical libraries, de novo design can be applied to generate new compounds with specific potency and selectivity properties. To this end, generative models, like variational autoencoders and generative adversarial networks (GANs), can be trained on a dataset of known molecules to learn their structural properties and generate molecules with similar properties. Reinforcement learning algorithms are also used during de novo molecule design to optimize the chemical structure of a new drug based on its pharmacological properties, with several hybrid approaches emerging that integrate different aspects of each model.

AI-Driven Drug Discovery Platforms and Clinical Translation Partnerships

Exscientia, Insilico Medicine, Recursion, BenevolentAI, and Schrödinger are five leading AI-derived drug discovery companies that have successfully advanced novel small-molecule drug candidates into the clinic. Both rentosertib (ISM001-055)—a pulmonary fibrosis candidate developed by Insilico Medicine—and DSP-1181—a serotonin (5-HT1A) receptor agonist developed by Exscientia—were designed by using AI-driven platforms before advancing to early clinical testing.

BenevolentAI and AstraZeneca joined forces in 2019 to identify novel targets in chronic kidney disease and idiopathic pulmonary fibrosis by using AI-driven knowledge graphs. Collectively, these partnerships highlight that drug discovery is shifting from

XSL•FO
RenderX

isolated computational innovation toward integrated pipelines in which AI platforms and pharmaceutical companies drive target identification, molecule design, and clinical translation.

Challenges and Limitations

The integration of AI techniques into drug design has several advantages over traditional discovery methods, including reduced time and costs, as well as better accuracy in predicting the safety and efficacy of new drugs. Nevertheless, the lack of diverse, representative, and high-quality datasets may introduce bias that leads to inaccurate predictions. In fact, many available chemical and biological datasets are incomplete or are acquired from heterogeneous experimental conditions, thereby reducing model generalizability and applicability to new chemical structures.

Trust is also imperative, underscoring the importance of transparency through open-source models, shared datasets, and explainable frameworks. However, AI does not eliminate downstream failures in clinical development. Gerstein cautions that once a promising molecule is identified, substantial challenges remain in optimizing delivery, systemic effects, and safety. Biological complexity, patient heterogeneity, and unforeseen toxicity mechanisms perpetuate high attrition rates, highlighting the need for AI systems that integrate more robust biological clinical context.

Edited by KA Clegg; submitted 14.May.2026; this is a non–peer-reviewed article;accepted 14.May.2026; published 29.May.2026.

Please cite as:
Cuffari B
From Virtual Molecules to Clinical Trials: How AI Is Reshaping Preclinical Drug Discovery
J Med Internet Res 2026;28:e101366
URL: https://www.jmir.org/2026/1/e101366
doi:10.2196/101366

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

US Food and Drug Administration Shifts to AI-Enhanced Regulatory Review With Elsa 4.0 and HALO

Tejas S Athni

Abstract

The US Food and Drug Administration recently announced the launch of Elsa 4.0 and Harmonized AI and Lifecycle Operations for Data (HALO)—an upgrade to its artificial intelligence platform. In this *News and Perspectives* article, JMIR Correspondent Tejas S Athni reports on the evolution of the Elsa model and potential benefits and challenges.

(J Med Internet Res 2026;28:e101884) doi:10.2196/101884

KEYWORDS

FDA; US Food and Drug Administration; Elsa 4.0; HALO; Harmonized AI and Lifecycle Operations for Data; generative AI; artificial intelligence; regulatory review; AI-enhanced; harmonized AI; medical devices; health policy

Textbox .

Key Takeaways:

- The US Food and Drug Administration's (FDA's) Elsa 4.0 artificial intelligence (AI) model and Harmonized AI and Lifecycle Operations for Data (HALO) platform shift AI from a peripheral tool to an embedded interface for querying, synthesizing, and acting on regulatory data across siloed systems.
- Integration with HALO may accelerate evidence synthesis and reduce review timelines, potentially enabling faster patient access to high-impact therapies and devices.
- These gains depend on robust human verification of Elsa outputs, as risks such as AI error propagation, workforce instability, and regulatory deskilling could undermine accuracy.

The US Food and Drug Administration's (FDA's) May 6, 2026, release of Elsa 4.0—a generative artificial intelligence (AI)–powered decision support system—marks a broader transition in how AI is deployed in federal health regulatory workflows and a shift in how FDA regulators access, synthesize, and act on data to make decisions. As an upgrade to the agency's previous AI system, Elsa 4.0 brings new capabilities and new challenges.

From Elsa 1.0 to 4.0

The original Elsa—Elsa 1.0—was launched in June 2025 as part of a broader modernization effort across the Department of Health and Human Services (HHS), in line with a 2025 White House directive encouraging federal agencies to adopt generative AI for efficiency. Built within a FedRAMP High–compliant Google Cloud platform environment, Elsa was designed as an internal AI assistant to aid FDA employees by reading, writing, and summarizing regulatory information. Specifically launched to support tasks such as adverse event summarization, label comparison, clinical protocol review, and identification of high-priority inspection targets, its other uses included drafting communications and generating code to build internal databases for nonclinical applications.

The goal was straightforward—accelerating evidence synthesis and cutting review times while reducing administrative burden and streamlining routine cognitive tasks. Early reactions to Elsa 1.0 suggested that it delivered operational value in streamlining routine tasks while serving as a "practical quick win" that built internal confidence in AI's role within regulatory workflows. However, others noted that the system hallucinated, generating fabricated studies or misrepresenting research, with some FDA reviewers describing the tool as "clunky." Subsequent versions, including Elsa 2.0 and 3.0, were rolled out over the following months with incremental upgrades in usability and scope. Elsa remained largely an adjunct tool. Staff still needed to gather and input data manually into the Elsa system. The FDA shifted the underlying model architecture from Anthropic's Claude to Google's Gemini.

Released in May 2026, Elsa 4.0 represents a more structural shift and a progression from task-based assistance to integrated data querying and analysis, with AI increasingly embedded into core FDA functions. It is paired with Harmonized AI and Lifecycle Operations for Data (HALO)—a platform that consolidates 40+ previously separate and siloed FDA systems into a single environment. With this integration, Elsa now sits directly atop the agency's data; instead of assembling documents independently from multiple sources, FDA staff can query across datasets through the AI interface itself, with new functionality and improved chat flexibility. Together, this can enable FDA staff to build and execute workflows directly within the system rather than rely on individual tools. The FDA has also emphasized data security: the AI model is not trained on users'

XSL•FO
RenderX

input data or data submitted by FDA-regulated companies, addressing concerns about confidentiality and data leakage.

Figure WL1.

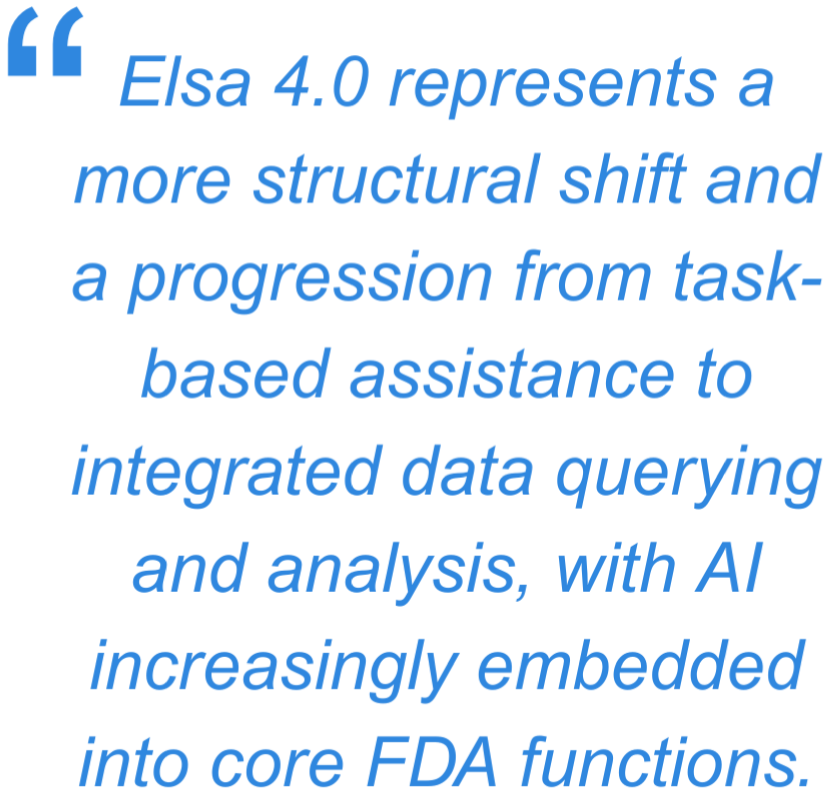

A Subtle but Structural Transformation

Elsa 4.0 and HALO change how regulatory work unfolds at the FDA. AI is no longer a sideline assistant in the regulatory pipeline. Rather, AI can now function as a core data synthesizer that takes regulator queries to retrieve and evaluate pertinent information. Elsa 4.0’s ability to automate the generation of charts and visualizations can help reviewers more quickly interpret trends, safety signals, or comparative outcomes. Similarly, optical character recognition capabilities allow legacy files to become fully searchable, reducing the risk of relevant information remaining buried in nonindexed formats. Voice-to-text dictation can enable regulatory reviewers to capture notes and observations in real time without interrupting their analytic workflow. Elsa 4.0 repositions AI in the regulatory pipeline from a modular tool to an increasingly central, overlying interface. Just as importantly, this means that human reviewers must be in the loop to verify and contextualize results, even if the first pass is made by generative AI.

For the public and patients, the effects of the Elsa 4.0 and HALO shift are unlikely to be immediately visible. Regulatory decisions will still follow established standards, and drug and device approvals will not simply be dictated by an automated AI decision output. However, the impact will likely emerge through improvements in how efficiently regulatory work can be performed and reduced frictions. For instance, within the FDA Breakthrough Devices Program, a reviewer evaluating a novel implantable surgical device through the premarket approval pathway may need to synthesize evidence across prior internal submissions, 510(k) clearances for related or predicate devices, pivotal clinical evidence generated under investigational device exemption studies, and postmarket safety signals in the Manufacturer and User Facility Device Experience (MAUDE) adverse event reporting database. Traditionally, this requires navigating multiple internal data systems to retrieve documents. With HALO integration, the reviewer can query across these datasets simultaneously, potentially reducing the timeline of task-specific work. This could shorten the overall time to regulatory decisions, allowing high-impact devices to reach patients sooner. In principle, even modest reductions in review timelines may translate into earlier access to novel therapies without compromising the underlying rigor of evaluation.

Risks and Early Lessons

The upgrade is, however, not without potential risks. With a unified HALO data platform, errors in retrieval or synthesis are no longer confined to a single query; they can propagate across the review pipeline, much like the swarm-like amplification of errors seen in AI-to-AI interactions. These potential errors can require additional time to verify and correct the information, which poses a risk to the efficiency gains promised by AI and could lengthen—not shorten—review timelines. This tension is compounded by the risk of sudden workforce reductions. The HHS terminated approximately 3500 FDA employees in April

XSL•FO
RenderX

2025 alone, before beginning partial rehiring in May 2025. These abrupt cuts create a standing risk that human verification capacity may suddenly be diminished, leaving Elsa 4.0 deployed into an agency that may not consistently be able to rigorously validate its AI outputs.

Regulatory deskilling is also a concern, as repeated reliance on AI for tasks such as synthesis and analysis may reduce human reviewers' direct engagement with primary evidence. Gradually, this can lead to the erosion of core evaluative skills, the weakening of domain judgment, and reduced ability to independently flag errors. The United States also lacks a unified regulatory structure for AI use in settings like health care and medical devices, which places the FDA in the position of both adopting AI internally and defining its best practices in real time. These challenges suggest that successful AI integration in the FDA will depend not only on technical capability but, as I've previously written, also on leadership structures and an agency-wide organizational culture that can sustain reviewer expertise and human-in-the-loop governance.

Edited by KA Clegg; submitted 20.May.2026; this is a non–peer-reviewed article;accepted 20.May.2026; published 29.May.2026.

Please cite as:
Athni TS
US Food and Drug Administration Shifts to AI-Enhanced Regulatory Review With Elsa 4.0 and HALO
J Med Internet Res 2026;28:e101884
URL: https://www.jmir.org/2026/1/e101884
doi:10.2196/101884

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Can Humanlike Reasoning Be Replicated in Large Language Models for Clinical Decision-Making?

Shalini Kathuria Narang

Abstract

A recent study comparing physicians and large language models on clinical reasoning tasks has received widespread attention. In this *News and Perspectives* article, JMIR Correspondent Shalini Kathuria Narang speaks to one of the study's researchers and reports on the real-world implications that can and cannot be drawn from its findings.

(J Med Internet Res 2026;28:e103526) doi:10.2196/103526

KEYWORDS

clinical reasoning; large language models; artificial intelligence; diagnostic decision support; internal medicine; OpenAI model

Textbox .

Key Takeaways:

- A recent study suggests that OpenAI's o1 preview model outperforms physicians across some common clinical reasoning tasks.
- While it's clear that large language models have improved from prior generations, the results do not suggest that AI systems are ready to practice medicine independently or that physicians can be removed from diagnostic processes.
- Further prospective trials are needed to examine how these models might be safely integrated into clinical workflows.

Large language models (LLMs) are rapidly being integrated into health care to support clinical decision-making. Although applying artificial intelligence (AI) to assist with clinical decision support is sometimes viewed as a high-risk endeavor, greater use of these tools might serve to mitigate the human and financial costs of diagnostic errors, delays, and lack of access.

Reasoning LLMs that can sequentially work through problems—mirroring structured thinking—have done well in assessing clinical cases. Whether these tools could match or perhaps exceed physicians' clinical reasoning for real diagnoses has been an ongoing question.

In a study led by researchers at Harvard Medical School and Stanford University, the OpenAI o1 series LLM, which was designed with more advanced reasoning capabilities, was compared to physicians with different levels of experience on clinical reasoning tasks. Clinical cases ranged from published patient studies to diagnosis and clinical care tasks with new patients in internal medicine.

A follow-up interview with Adam Rodman, MD, MPH, FACP, helps contextualize the findings of this study after the flurry of media coverage on its publication. Rodman is a general internist and medical educator at Beth Israel Deaconess Medical Center, assistant professor at Harvard Medical School, and Director of AI Programs for the Carl J. Shapiro Center for Education and Research.

The Study and Its Implications

In the study, the LLM was pitted against the physicians in experiments of clinical reasoning, including cases utilizing real and unstructured clinical data taken from the health records in an emergency room (ER). These diagnostic touchpoints mirror the high-stakes, time-sensitive decisions taken by health care professionals with limited information.

Human experts and LLM second opinions were compared for randomly selected patients in the ER of a Boston hospital at 3 stages: initial triage at arrival, first contact with a physician, and admission to the medical floor or intensive care unit. The model was evaluated on 76 cases by 2 doctors who did not know if the case assessments had been made by the AI model or by attending physicians.

The LLM outperformed physicians across tasks like rapid assessment, recommendations for medications, caring goals, and overall case management. It matched or exceeded human performance across each stage, with the widest gap at initial ER triage, where there is the least information available. Humans, GPT-4, and the OpenAI o1 series all improved as more information became available.

While the study results have face validity, the tasks that were studied, namely providing second opinions at predefined touchpoints, are best thought of as a proof of concept. Decisions in the ER are often centered around triage, disposition, and immediate management rather than diagnostic accuracy.

XSL•FO
RenderX

"The study is a validation of the diagnostic performance of these [LLM] models. The basic claim is that the diagnostic performance of the models is not just an artifact of the evaluation mechanisms like vignettes, but holds with real clinical data. That does not mean that just deploying it makes a difference in patient care. It's more, like, watching over like a second set of eyes," says Rodman.

Caveats and Cautions

While the study established a foundation for evaluation across text-based tasks, a substantial caveat is that clinical practice involves visual and auditory cues, such as findings from physical examinations.

Existing studies suggest that current foundational models are more limited in reasoning over nontext inputs. Future research is needed to build on preliminary work assessing how humans and machines may collaborate by using nontext signals, which requires new benchmarks, trials, and technology to more faithfully measure clinical encounters. Such studies may help further investigate an important limitation of existing benchmarks of medical AI: their reliance on the careful work of clinicians to curate and "clean up" cases and their reliance on questions originally developed for educational purposes, which can lead to the performance of AI models being overstated compared to when using "messy" data in more realistic clinical workflows.

"There's a lot of desire to say that we could use technology to replace doctors, but that's not at all what I think it's capable of," says Rodman. "The models are very capable in assisting with diagnosis, but they are not good at integrating information from many different sources. Doing a physical exam, I talk to a patient, look at them, hear the hesitation in their voices. I'm situated in the room with them. Not just getting information from them directly…looking at the medical record, calling other people, and doing an investigation. The LLMs are really good at integrating information that a physician curates from other sources, or they're really good at collecting that information verbally from a patient. They are not at all good at all those other parts of the diagnostic process, which are just as important."

Figure WL1.

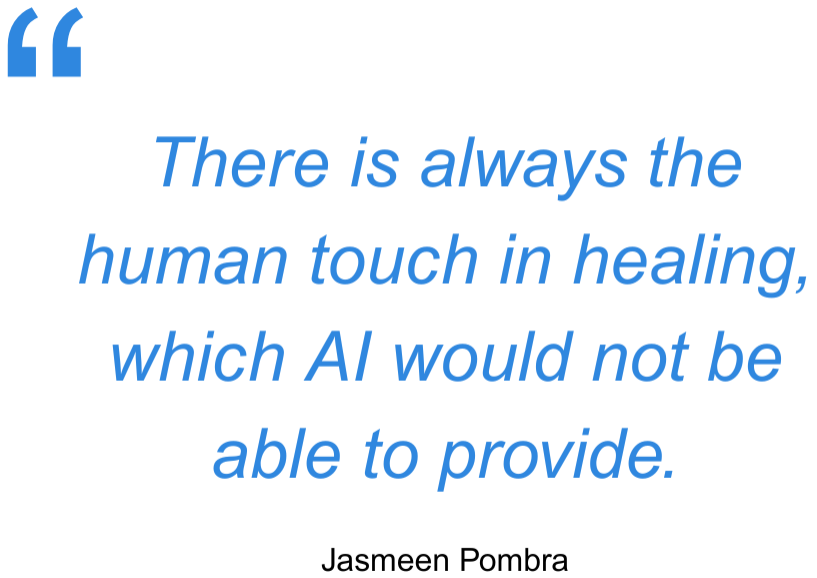

Jasmeen Pombra—assistant chief of Hospital Operations and chief of Neuro Hospital Medicine at Kaiser Permanente, Redwood City, California—agrees.

"AI [medical tools] can help make diagnosis and be an adjunct to help provide faster care but cannot replace clinical gestalt which comes from years of clinical practice and more from treating patients by observing subtle signs," she says. "AI is very useful in tedious tasks and may help decrease workload for physicians by generating notes and being able to sift through patients' health summary to make relevant data available for clinicians for making diagnosis and treatment plans. There is always the human touch in healing, which AI would not be able to provide."

Further, although large numbers and varieties of cases were included in the study, all were focused on internal medicine and emergency care. It is not representative of broader medical practice, which includes multiple specialties that require varying skill sets. Performance may vary according to diagnoses, patient characteristics, or practice locations that were not interrogated in this study. Dozens of other tasks that have greater impact on actual clinical care can be studied.

XSL•FO
RenderX

Moving Forward

The study makes the case that medical AI is ready to be studied the same way as all new medical interventions—through carefully controlled clinical trials in multiple real care settings.

The researchers are clear that their results do not suggest that AI systems are ready to practice medicine autonomously or that physicians can be removed from the diagnostic process.

Instead, they emphasize the importance of conducting prospective trials in real-world settings to evaluate the latest models of the LLM used in the study, which has since been superseded by newer models (like OpenAI's o3 model) with improved reasoning capabilities, deliberation time, and ability to process multimodal inputs.

They also emphasize the need for health care systems to invest in computing infrastructure and design for clinician-AI interaction, to facilitate safe integration of AI tools into patient-care workflows. This includes the development of monitoring frameworks to oversee the broader implementation of AI clinical decision support systems, monitoring not just final diagnostic accuracy but also other metrics for successful deployment, including safety, efficiency, and cost.

"We've largely moved on with our research. The most obvious use case for this specific technology is trying to identify errors before they happen. Looking for when doctors may be going in the wrong direction, and using that as an early signal," says Rodman. "This could be something called an E trigger or in the form of a second opinion."

Edited by KA Clegg; this is a non–peer-reviewed article;accepted 03.Jun.2026; published 15.Jun.2026.

Please cite as:
Narang SK
Can Humanlike Reasoning Be Replicated in Large Language Models for Clinical Decision-Making?
J Med Internet Res 2026;28:e103526
URL: https://www.jmir.org/2026/1/e103526
doi:10.2196/103526

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jun.2026.

XSL•FO
RenderX

How Health Care Workers Can Manage Digital Fatigue

Sara Novak

Abstract

One of the key aims of the digitization of health care is reducing burden on health care workers and health care systems. In this *News and Perspectives* article, JMIR Correspondent Sara Novak reports on one of its paradoxical consequences: digital fatigue.

(J Med Internet Res 2026;28:e104196) doi:10.2196/104196

KEYWORDS

burnout; professional; electronic health records; health personnel; medical informatics; occupational stress; user-computer interface; workload; digital fatigue

Textbox .

Key Takeaways:

- While health care's digital transformation has the potential to improve care, it can also make more work for health care workers, contributing to exhaustion and burnout.
- Alerts and redundant messages need to be streamlined, so that health care workers aren't overloaded.
- The digital workload needs to be a recognized aspect of the workday.

Health care's digital transformation has the potential to improve care while also easing the workload on physicians, nurses, and health care workers by making services accessible, streamlining processes, and automating time-consuming tasks. But digitization can also increase workload by forcing health care workers to learn new tools and systems and shoulder the burden of integrating and implementing them.

As health care workers become increasingly dependent on digitization through electronic health records (EHRs), data dashboards, and health information exchanges, it's important for health care institutions and for the workers themselves to establish personal strategies to manage digital fatigue and the overwhelm, exhaustion, and burnout that digitization can cause.

"Technology promised to 'give you time back,' only for physicians to discover that it often reallocated that time from patients to screens," said Hassan Bencheqroun, MD, MBA—a pulmonologist, critical care physician, and assistant professor of medicine at the University of California, Riverside.

The Causes of Digital Health Care Fatigue

Health care workers are already expected to manage patient care, time pressure, and the emotional strain of taking care of sick patients. Digital fatigue can add another layer to this already potent emotional strain.

"Digital fatigue in this context is not only about screen time or clicks; it's about misaligned promises and unbalanced trade-offs. The result is a growing sense that we are supervising machines rather than practicing medicine," said Bencheqroun.

When health care workers are constantly having to interact with and respond to messages and alerts in the EHR, it takes away from the amount of time that they can spend face-to-face with patients.

One study published in the journal BMC Nursing points to the emerging phenomenon of digital compassion fatigue (DCF)—a form of emotional exhaustion that results when health care workers provide empathic care through virtual platforms. DCF is defined by persistent emotional exhaustion after digital engagement, as well as cognitive overload, feelings of professional inadequacy, and a fear of a reduction in the quality of care. Health care workers feel like the time spent on digital platforms makes it impossible for them to spend adequate time with their patients.

Health care workers feel less connected to patients when they're spending too much time on digital devices and also having less time to recover at the end of the day because they have to spend time outside of work hours responding to EHR alerts and messages, said Rachel Hoopsick, PhD—an assistant professor of health and kinesiology at University of Illinois, Urbana-Champaign, who studies digital fatigue in health care workers.

Hoopsick said that we're dealing with a fee-for-service health care system, meaning that health care providers are paid for the services that they provide, including tests and office visits, so time spent with each patient is already limited. Adding to the administrative burden by having to learn new systems only adds fuel to the fire.

"It's a cycle that contributes to the eventual burnout of healthcare workers," said Hoopsick.

XSL•FO
RenderX

Managing Digital Fatigue Like You Would Other Occupational Health Risks

Digital fatigue needs to be taken seriously, like you would any other occupational risk. This starts by reducing the low-hanging fruit, including low-value alerts that pop up within the EHR, said Hoopsick.

For example, one study published in the journal Applied Clinical Informatics noted that most of the allergy alert prompts given for medications were for non–life-threatening allergies. In addition, one-third of alerts were redundant. Research also suggests that clinicians often ignore repeated alert guidance, "suggesting the guidance is not clinically helpful nor likely to alter clinical management," wrote the study authors.

Figure WL1.

According to American Medical Association (AMA) guidance, other tools for reducing portal messages include annual automatic renewal of certain prescriptions and getting previsit labs, so results can be discussed at annual visits, rather than going back and forth on the EHR afterward.

The digital workflow also needs to be more team-based. For example, having a few members of the team designated to deal with inbox management, so that portal messages, medication refills, and administrative requests don't pile up and result in after-hours workload. This way, the messages aren't ignored and don't end up unanswered.

Alternatively, the task of responding to messages can be separated individually. For example, the member of the team who orders a test should be responsible for communicating with the patient about it. If a team member is out of town, then another member involved with triage should accept responsibility while they're away and vice versa, said Hoopsick.

Additionally, health care institutions need to recognize the digital workload as part of a health care worker's tasks and not simply expect that it will be done outside of regular business hours, said Audrey Hai, PhD, MSW—an assistant professor at Tulane University in New Orleans who studies digital fatigue. She adds that in the digital age, much of the workload isn't recognized by the institution, which has an impact on the caseload of each clinician.

"Institutions need to consider the 50 emails that you likely have to respond to when assigning patients. It's part of the 'invisible workload' that's become associated with digital care," said Hai.

Additionally, when health care workers switch health care institutions or move to a new digital platform, health care institutions need to consider the training involved in learning a new system and not have an expectation that workers should teach themselves the platform, which is inefficient and takes time out of their workday. Not feeling competent in digital skills can also contribute to DCF, said Hai.

While it's the health care institutions that need to bear the brunt of reducing digital fatigue, at the individual level, there are also a number of steps that health care workers can take. Hai recommends scheduling a digital detox break in your calendar every few days to allow for a few minutes away from all devices and screens.

Workers can also schedule email delivery, so that when coworkers or patients send you messages, you can schedule them to be delivered the next morning. That way, you don't have to constantly check your phone past working hours, and you can be rested and ready to take on the next workday.

"Healthcare platforms need to support care, not distract from it," said Hoopsick.

XSL•FO
RenderX

Edited by KA Clegg; submitted 09.Jun.2026; this is a non–peer-reviewed article;accepted 09.Jun.2026; published 15.Jun.2026.

Please cite as:
Novak S
How Health Care Workers Can Manage Digital Fatigue
J Med Internet Res 2026;28:e104196
URL: https://www.jmir.org/2026/1/e104196
doi:10.2196/104196

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

You Can't Launch This: Trust as Infrastructure in Digital Behavioral Health

Trevor van Mierlo

Abstract

Digital behavioral health interventions frequently fail to scale, even when evidence-based and technically and operationally sound. In this *News and Perspectives* article, researcher, digital behavioral health platform founder, and JMIR Correspondent Trevor van Mierlo concludes a four-part series examining why this occurs, reporting on the foundational role of trust.

(J Med Internet Res 2026;28:e104731) doi:10.2196/104731

KEYWORDS

trust; artificial intelligence; digital health; mental health services; consumer participation; patient-centered care; organizational policy; health governance

Textbox .

Key Takeaways:

- Trust and engagement are not the same thing.
- As AI systems become more adaptive and inferential, trust may need to be treated as infrastructure rather than a feature.

Trevor van Mierlo, DBA, is the founder and scientific architect of the Evolution Health platform and acts as a consultant for academic and commercial digital mental health initiatives. The views expressed in this article are his own.

While preparing to launch a resilience course for—and co-developed with—displaced Ukrainians, our lived experience collaborators identified a major problem, as described in a previous article.

The issue wasn't technical. Content had been translated and culturally adapted by displaced Ukrainians. Security audits were complete. The infrastructure was stable.

The issue was a short explainer video with an AI-generated Ukrainian voice-over. It immediately reminded collaborators of Russian propaganda videos circulating online.

The problem was trust.

As AI becomes increasingly integrated into behavioral health, the question is no longer simply whether systems work. Trust may be equally important.

My interest in trust led to several conversations with researchers and practitioners across cybersecurity, informatics, behavioral health, and public digital infrastructure. What emerged was that trust can fail at multiple layers—social, technical, institutional, and behavioral—even in systems that are clinically validated and operationally sound.

Intervention Engagement ≠ Trust

Tonia San Nicolas-Rocca, PhD, is Faculty Associate Dean in the College of Information, Data and Society at San José State University, where her work focuses on cybersecurity, digital trust, and patient-centric health systems.

In discussing trust, San Nicolas-Rocca notes that there can be a power imbalance in digital health engagement. People may use health systems because, quite simply, they have to. Their access to care, services, or institutional support is often dependent on continued participation.

For example, someone may complete an assessment, log into a portal, or continue engaging with a platform or system they distrust. But people who distrust systems may also behave differently within them. They may minimize symptoms, avoid sensitive disclosures, disengage selectively, or provide information they believe is safer rather than accurate.

This creates an important problem for digital behavioral health. Engagement metrics may tell us whether individuals continue participating, but not whether the underlying behavioral data remain authentic, complete, or clinically meaningful.

Speaking with San Nicolas-Rocca caused me to reflect on some of my recent work. For years, my research has focused on using behavioral economics to influence user engagement through small, tailored changes in language and tone. In many ways, this is an attempt to solve digital health's original sin.

But are users aware of this? And if they are, does that change the way they behave inside the system?

Consent Is a Snapshot

Behavioral systems collect enormous amounts of longitudinal data. Increasingly, they also generate behavioral phenotypes:

XSL•FO
RenderX

patterns of interaction that reveal how individuals engage with digital environments over time.

These patterns may reveal far more than traditional demographic or clinical variables alone. The more people interact with a system, the more the system learns about what keeps them engaged. Every click, login, uploaded image, completed assessment, or abandoned exercise teaches the system something.

Benjamin Schooley, PhD, is a professor in the Department of Electrical and Computer Engineering at Brigham Young University. He argues that future systems should most likely require "computable governance." In simple terms, if systems continue evolving after deployment by learning from user behavior, disclosures and governance should also evolve.

Privacy policies and consent buttons are snapshots. Schooley's concerns align closely with contextual integrity: the idea that trust depends not simply on whether your data is secure, but whether information flows that use it remain appropriate. To be relevant and sustain trust, consent may also need to be dynamic.

Enforceable Trust

Kavya Pearlman, MS, is a cybersecurity architect and AI governance expert whose work focuses on artificial intelligence, digital trust, and emerging governance risks. Pearlman's concern is not simply what systems collect, but "what is inferred, how long those inferences persist, and whether governance remains enforceable over time."

In Pearlman's view, trust cannot rely solely on privacy policies, consent forms, or institutional assurances. Instead, governance must become transparent and verifiable. If an AI system generates a behavioral risk score, modifies treatment recommendations, or triggers an intervention, organizations should be able to reconstruct exactly what data informed that decision, which model was operating at the time, and whether the system functioned within approved governance boundaries. Most behavioral health platforms today cannot produce this evidence—they have usage logs and vendor dashboards, not forensic trails.

Pearlman's proposed mechanism draws from outside health care: cryptographically verifiable audit trails, modeled on aviation flight recorders and financial transaction logs. These produce tamper-evident records of what data was accessed, which model was operating, and what autonomy level governed each decision. Digital behavioral health, she notes, often relies on rough analytics where it should have forensic-grade traceability.

As Pearlman noted during our discussion, "If you can't explain and defend a decision after the fact, you shouldn't be automating it in behavioral health care."

In this context, static disclosures alone may not be sufficient to sustain trust, particularly when systems continue learning from users long after initial consent has been provided.

Trust Begins Before Governance

Valeriya Ionan, MBA, former Deputy Minister of Digital Transformation of Ukraine, helped build Ukraine's public digital infrastructure during periods of both pandemic response and full-scale war. Reflecting on these questions, she challenged one of the central assumptions underlying current discussions about AI governance and digital trust.

"We often assume trust is a feature technology creates," she wrote, "while I increasingly think trust is social first and tech second."

XSL•FO
RenderX

Figure WL2.

Before discussing trust in AI systems, we may need to ask a simpler question: what creates trust in the first place?

Valeriya Ionan, MBA

Her observation reframes much of the current conversation. Before people evaluate privacy protections, consent structures, or explainability frameworks, they must first believe the surrounding institutions and systems are worthy of trust.

This becomes especially important in environments shaped by misinformation, weak public communication, low digital literacy, or longstanding institutional distrust. In these contexts, trust problems may begin long before users even encounter the technology.

When Trust Becomes Infrastructure

Joe Grzywacz, PhD, Associate Dean for Research and organizer of the Health TechQuity initiative at San José State University, noted that equitable implementation depends not only on whether systems are available, but whether individuals believe the systems themselves are worthy of engagement and participation.

Grzywacz emphasized that trust inherently requires vulnerability. In digital behavioral health, individuals disclose intimate personal information while increasingly ceding interpretive authority to systems, platforms, and algorithms they often do not fully understand.

As behavioral systems become increasingly adaptive and predictive, this asymmetry between users and infrastructure may become a source of mistrust.

The broader implication may be that trust itself is becoming infrastructure.

Launching With the Wrong Voice

Trust does not automatically emerge from efficacy studies, privacy statements, or institutional reputation. People may distrust systems that are technically sound, clinically validated, and fully compliant.

Future behavioral systems may require trust to be actively maintained rather than simply assumed. Governance structures and transparency will remain essential. But they may prove insufficient if systems fail to establish legitimacy within the social and contextual environments in which individuals experience them.

XSL•FO
RenderX

Digital behavioral systems may ultimately succeed or fail depending on not only what they deliver, but whether people trust them. In the AI era, trust may no longer be a secondary feature of behavioral infrastructure. Instead, it may become the layer everything else is built on: engineered, maintained, and auditable, yet ultimately judged by the people asked to rely on it.

Edited by KA Clegg; submitted 15.Jun.2026; this is a non–peer-reviewed article;accepted 15.Jun.2026; published 22.Jun.2026.

Please cite as:
van Mierlo T
You Can't Launch This: Trust as Infrastructure in Digital Behavioral Health
J Med Internet Res 2026;28:e104731
URL: https://www.jmir.org/2026/1/e104731
doi:10.2196/104731

© JMIR Publication. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.Jun.2026.

XSL•FO
RenderX

Can AI in Health Care Be Truly Inclusive?

Beth Rush

Abstract

A new equity-driven framework offers a road map for building equity and inclusivity into the life cycle of AI in health care. In this *News and Perspectives* article, JMIR Correspondent Beth Rush reports on whether meaningful equity and inclusivity is possible in AI-enabled health care.

(J Med Internet Res 2026;28:e104527) doi:10.2196/104527

KEYWORDS

equity-driven AI; health equity; artificial intelligence; algorithmic bias; digital health; oral health disparities; health inequities; clinical decision support systems; patient-centered AI; AI implementation; health care access; bias mitigation

Textbox .

Key Takeaways:

- AI in health care risks reproducing existing disparities at scale unless equity, diversity, and inclusion are integrated throughout the AI life cycle.
- The EDAI framework provides actionable guidance across micro, meso, and macro levels to integrate equity, diversity, and inclusion in health and oral health care.

AI is transforming health care delivery, with applications spanning diagnostic imaging, clinical decision support, predictive analytics, and administrative automation. With adoption accelerating, how can AI systems be designed, developed, implemented, and monitored to avoid perpetuating or amplifying existing health and social inequities?

On April 30, 2026, Samira A Rahimi, PhD, presented findings from a multiphase research initiative designed to answer this question. Her webinar, “Towards Equitable and Inclusive AI in Health and Oral Health Care,” hosted by JMIR Publications, introduced the EDAI framework for Integrating Equity, Diversity, and Inclusion Throughout the Lifecycle of AI to Improve Health and Oral Health Care using a comprehensive approach developed with input from international stakeholders across clinical, technical, policy, and community domains.

The Problem of Invisible Populations

Health care AI systems are typically trained on data generated through existing care pathways. This creates a fundamental problem: populations who face barriers to accessing care are not appropriately reflected in the datasets used to build these tools.

Rahimi, who led the framework’s development, explains how this pattern creates what she calls “invisible populations” in AI systems. “If people cannot access care, or are misclassified, under-diagnosed, or are inadequately represented in health records,” she notes, “they become invisible to the algorithm, because in the first place, we don’t have their data sets.”

The groups most at risk include people with unstable housing, migrants and refugees, individuals with disabilities, those living in rural or remote areas, caregivers of older adults, people with low digital access, and patients with complex multimorbidity. When AI models are developed using incomplete or unrepresentative data, they risk systematically underserving the populations who need care most.

Research supports these concerns. A 2024 study in PLOS Digital Health noted that disparities driving bias in medical AI are rooted in long-standing structural inequalities within health systems. The underlying forces are not recent developments—they reflect historical patterns of discrimination that have shaped clinical data for decades.

Building the EDAI Framework: A Participatory Approach

The EDAI framework was developed through a three-phase process that included systematic evidence synthesis, stakeholder workshops in Montreal and Geneva, and iterative refinement based on expert input.

“The principles of equity, diversity, and inclusion in AI can be universal,” Rahimi explains, “but when it comes to implementation, it could be very context specific.” Equity does not look identical across health care systems. Local history, infrastructure, governance, and community needs shape how these principles must be operationalized.

The framework operates across three levels: micro (individual patients and providers), meso (organizations and institutions), and macro (policy and societal structures). It addresses barriers and facilitators at each level, with specific guidance for patients, clinicians, AI developers, and policymakers.

One of the framework’s most distinctive contributions is its explicit inclusion of oral health—a domain often excluded from

XSL•FO
RenderX

digital health innovation despite strong evidence linking oral health to cardiovascular disease, diabetes, pregnancy outcomes, and quality of life.

"Oral health is often separated from the rest of health care, both in how it is structured and financially managed," Rahimi notes. "Dental care is frequently less integrated, or not integrated at all in some counties, into electronic health records and less prioritized in digital health policies. Excluding oral health from AI reinforces inequities that already exist."

Figure WL1.

The issue is that equity work is often treated as optional rather than as a core infrastructure for safe AI development.

Samira A Rahimi, PhD

"Health AI decisions are not usually driven only by evidence," she explains. "They are also shaped by incentives, timelines, funding structures, and technical cultures. Many systems still prioritize speed, scalability, and commercial readiness over equity and long-term safety."

Underprioritization of equity concerns is not about genuine scarcity or misplaced priorities, says Rahimi: "It's often framed too conveniently as a scarcity problem, but it is not. The issue is that equity work is often treated as optional rather than as a core infrastructure for safe AI development."

Why Equity Remains an Afterthought

Despite evidence that incorporating equity-related factors such as social determinants of health can improve AI model performance, these considerations are rarely prioritized during development. Rahimi's systematic review found that very few studies addressed the integration of equity and inclusion principles across the full AI life cycle—from problem formulation and community engagement through data governance, deployment, and monitoring.

Clinical Realities and Implementation Gaps

Health care leaders with frontline clinical experience echo these concerns. Greg Carroll, MD, MBA, an emergency physician and founder of ClinNEXUS, a company focused on whole-person care for complex patient populations, recalls a turning point in his approach to health care delivery.

XSL•FO
RenderX

"I repeatedly treated a patient struggling with addiction who continued relapsing despite appropriate clinical care," he explains. "Eventually, I discovered the patient was returning each day to housing directly adjacent to his narcotics supplier. The clinical interventions alone were never going to solve the problem."

That realization informed his work developing AI-driven care coordination systems designed to identify hidden behavioral, social, and clinical risk factors before vulnerable patients escalate into crisis. His experience underscores a central challenge in health care AI: tools built without understanding the social context of illness risk missing the very factors that drive poor outcomes.

Implementation challenges extend to workforce readiness as well. Frontline health care workers—particularly those in temporary or travel roles—often encounter AI systems with minimal training or technical support. This creates inconsistent access to tools that could improve care, and in high-stress environments, unfamiliar technology can compound existing workload pressures rather than alleviate them. Without intentional implementation strategies, AI adoption risks widening the gap between well-resourced and underresourced care settings.

From Framework to Action

For existing AI systems already deployed in clinical settings, the path forward is not straightforward.

Rahimi emphasizes that retrofitting is possible in some cases, but not always sufficient. "If the original design of the AI system excludes key populations or was built on fundamentally biased assumptions, starting over would be the best and safer option," she argues. "We shouldn't normalize patching equity after the harm has already occurred."

Moving forward will require concrete action from all stakeholders. Rahimi offers specific recommendations:

- Patients and communities should be involved as partners in defining problems, not only as data sources
- Clinicians should ask what populations were included, excluded, or misclassified before trusting an AI tool in clinical practice
- AI developers should conduct equity impact assessments throughout the AI life cycle
- Policymakers should require transparency, bias monitoring, and accountability before AI tools are scaled in health care settings

"AI will shape the future of health care, but the real question is: for whom and under whose values?" she concludes. "If equity, diversity, and inclusion are not built into these systems from the beginning—not just at the end as a fixing issue—AI risks reproducing the same disparities health care has struggled with for decades, but at greater scale and speed."

Edited by KA Clegg; submitted 12.Jun.2026; this is a non–peer-reviewed article;accepted 12.Jun.2026; published 22.Jun.2026.

Please cite as:
Rush B
Can AI in Health Care Be Truly Inclusive?
J Med Internet Res 2026;28:e104527
URL: https://www.jmir.org/2026/1/e104527
doi:10.2196/104527

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.Jun.2026.

XSL•FO
RenderX

The Future of Online Misinformation Research: Tackling the Landscape With Integrity and Urgency

Wen-Ying Sylvia Chou

Abstract

Since early 2025, the scientific research and funding landscape has faced tremendous changes and challenges in the United States. In this *News and Perspectives* article, health communication scholar, former program official of the National Cancer Institute, and JMIR Correspondent Wen-Ying Sylvia Chou discusses the impact on—and future of—online misinformation research in this politicized context.

(J Med Internet Res 2026;28:e104526) doi:10.2196/104526

KEYWORDS

misinformation; disinformation; science communication; political interference; science policy; academic freedom; chilling effect; censorship; debunking; prebunking

Textbox .

Key Takeaways:

- Threats to research on health communication, social media, and misinformation undermine scientific integrity, truth, and democracy.
- Despite challenges impeding scientific inquiry, continued research on online health misinformation is critically important.
- Current research priorities include uncovering and countering the tactics of major misinformation spreaders, scaling up mitigation interventions, and addressing root causes of misinformation spread.

This January, I made a life-altering decision to leave a cherished, 18-year federal career at the National Cancer Institute, where I had overseen a portfolio of hundreds of projects on social media and digital health.

Like most civil servants throughout 2025, I witnessed political overreach and endured sleepless nights replaying actions taken in response to (or in anticipation of) new rules and expectations. The moral injury became too great and I decided, at significant personal cost, to resign. This decision, while personal, may serve to illuminate a path forward for the beleaguered field of health misinformation research.

Evolution of Online Misinformation Research

Over a decade immersed in investigating the evolving media landscape and its impact on health, I have slowly come to understand the upheavals, tensions, and engineered “controversies” affecting this scholarship. Since the late 2010s, we enthusiastically conducted observational and interventional research, demonstrating social media’s impressive, low-cost reach and positive health impacts. Concerns about potential harm were initially overlooked. Then, growing disinformation campaigns worldwide, including during the 2016 US presidential election, started to shift public perception of social media. Researchers began investigating the causes and consequences of health misinformation across populations, platforms, and topics.

Just as valuable insights were gained and promising mitigation strategies developed, the sociopolitical dynamics—including the appointment of Robert F Kennedy Jr as Secretary of Health and Human Services—quickly rendered this area “controversial,” if not entirely off-limits for those seeking federal funding.

I had a front-row seat to the self-censorship that ensued: watching (and participating in) the pausing of projects, replacing or deleting words in materials, and suggesting grantees alter their applications to “clean up” (implying that there was something impure before) or “realign” (suggesting political alignment was necessary for scientific research) research to avoid running afoul of unspoken rules about allowable research. In fact, the first terminated grant in my portfolio was an intervention to help reduce people’s vulnerability to falsehoods. The word “misinformation” was in the title. Since then, censorship at the National Institutes of Health (NIH) has only increased and become normalized. Such censorship compromises our ability to effectively communicate evidence-based health information and safeguard truth.

Science, Truth, and Democracy

This tumultuous arc illustrates the intricate connection between social science and democracy.

We now understand health misinformation not just as an impediment to health but also as a symptom of a larger set of societal challenges: frayed relationships, polarization, lack of

XSL•FO
RenderX

effective regulation (or self-regulation) in the technology sector, entrenched prejudice, and media silos. This knowledge now equips us better to combat misinformation and address the broader challenges.

Figure WL2.

We simply cannot afford to hamstring research on this topic out of anticipatory obedience.

This fraught sociopolitical and information landscape is here to stay and misinformation is only becoming more rampant. We simply cannot afford to hamstring research on this topic out of anticipatory obedience. Instead, we must tackle the online information ecosystem with urgency, navigating the complexities of the moment with moral integrity.

When science and truth is undermined, so is our democracy.

From Retreat and Capitulation to Solution and Scale-Up

It's clear that neither the technology industry nor Congress currently have the political will to implement systematic efforts to curb the spread of misinformation.

Nonetheless, there remain critical and timely research opportunities that can inform effective health communication and improve the health and well-being of people across diverse communities. Three domains have received limited attention to date and should be prioritized.

Uncover and Counter the Tactics of Major Spreaders

Beyond characterizing information quality, we need to continue identifying major spreaders of health misinformation and uncovering their tactics. For example, the Virality Project from the now-shuttered Stanford Internet Observatory described, through natural language processing–assisted data mining techniques, what the authors called a "full-spectrum propaganda" approach around COVID-19 vaccine disinformation driven by an organized business model.

Such investigative efforts must expand given the rise of AI-generated content and algorithms designed to maximize user engagement at all costs. Through refined computational methods capable of analyzing large amounts of textual and visual data, monitoring and understanding the playbook of disinformation agents can inform a timely response of communication practitioners. For example, instead of raising awareness of poor-quality content, efforts may be better directed toward calling out misinformation spreaders and their tactics—including manipulative use of AI, such as generating deepfakes of health care providers endorsing unproven medical products, to target and tailor content to vulnerable users.

Scale Up Mitigation in Collaboration With Other Societal Actors

Recent scholarly work has generated an arsenal of misinformation mitigation strategies such as debunking and "prebunking" interventions. There is an urgent need, however, to expand and diversify study samples and scale up efficacious strategies. Even as technology companies themselves pull back

XSL•FO
RenderX

on mitigation efforts, large-scale interventions such as prebunking or the use of compelling personal narratives can be implemented through online ads, potentially reaching millions of users at a reasonable cost. Similar interventions could be implemented in specific online echo chambers where misinformation is especially prevalent.

Efforts to scale mitigation interventions may engage with other societal actors: community organizations (eg, churches, advocacy groups), clinics and hospitals, or health care payers. A starting point is fostering trusting relationships and getting buy-in from community members (eg, pastors, pediatricians, residents in a retirement community), wherein mitigation efforts can be disseminated by already trusted voices.

Address Root Causes

Tangible communication strategies—for example, articulate topline messages; tackle misinformation themes, tropes, and techniques instead of fact-checking statements; avoid creating an information vacuum—are necessary, but no longer sufficient.

Misinformation spread is driven by, and exacerbates, many other social phenomena, including information silos, a frayed social fabric, lack of trust in institutions, tribalism, and limited science literacy. It's time to address these fundamental challenges head-on.

Health communication is part of the broader communication landscape and sociocultural-political discourse. When a health topic becomes politicized, traditional campaigns and messaging may be too late to be effective. We need to confront and shift the public discourse. We can start with facilitating meaningful conversations—creating opportunities such as a third place for dialogue, building common ground across disparate silos, and allowing disagreement while maintaining civility. Ultimately, building common ground can restore trust in experts; oftentimes, trust in the source of the information is more important than message accuracy. Centering interpersonal relationships is a crucial way to address root causes of misinformation.

A Call to Hold the Line

The academic community must collectively confront and learn from the recent attacks the misinformation research field has withstood. I recall witnessing communication research at NIH being canceled in 2023—leaders, in what I have now understood as an act of anticipatory compliance, were fearful that the science would be labeled as "controversial." They thought the best strategy would be to stop this work, or at a minimum, call it something else. At the time, my postmortem "how could this happen?" went unanswered, only countered with "let's keep quiet and move on."

There is actually much to reflect on when a given research topic gets falsely labeled as "political," "controversial," or "woke." We can uncover motives and agendas, engage in discussions and debates, and fight back by following professional ethical standards and the rule of law, and uphold our right to academic freedom. It is all the more important to continue the research and, whenever possible, challenge censorship attempts—today one deletes a word, tomorrow the aims are altered, and the next day the project ceases. When we stop studying the causes and consequences of misinformation, not only is the public's health on the line, but so is truth and democracy.

I have used the word "moral" several times in this article. I do not invoke this term lightly. Every scholar and research funder in the United States, private or public, currently faces a precarious situation due to a polarized and fraught political landscape. Moral courage, not just pragmatism and rigor, is critical. Continuing our research is a responsibility, a defense of science and the public good, and an act of necessary defiance.

Table .

Glossary	
Misinformation	False, inaccurate, or misleading claim of fact regardless of the intention of those who spread the information. **Example:** Sharing a post about a potential COVID-19 treatment you believe to be true.
Disinformation	False, inaccurate, or misleading information that is deliberately shared with the intent to mislead, deceive, or cause harm. **Example:** A company with a vested interest sharing a post about a potential COVID-19 treatment with no proven efficacy in order to generate profits or manipulate public opinion.
Debunking	Interventions that involve retroactively identifying and correcting false, inaccurate, or incomplete information. **Examples:** A label on a social media post verifying that it contains misinformation. A video stitch correcting misinformation in another video.
Prebunking	Interventions that involve proactive warnings, context, or training to help people recognize and resist false, inaccurate, or misleading information. **Examples:** A warning label on posts from a bot account urging users to fact-check its posts. An online toolkit providing social media users with examples of common types of misinformation or disinformation techniques.

XSL•FO
RenderX

Edited by KA Clegg; submitted 12.Jun.2026; this is a non–peer-reviewed article;accepted 12.Jun.2026; published 22.Jun.2026.

Please cite as:
Chou WYS
The Future of Online Misinformation Research: Tackling the Landscape With Integrity and Urgency
J Med Internet Res 2026;28:e104526
URL: https://www.jmir.org/2026/1/e104526
doi:10.2196/104526

© JMIR Publication. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.Jun.2026.

XSL•FO
RenderX

Potential of Digital Tools for Chronic Pain Management

Vanessa Nirode

Abstract

Chronic pain is notoriously complex to manage and treat. In this *News and Perspectives* article, JMIR Correspondent Vanessa Nirode reports on several digital innovations that may help to bridge existing care gaps.

(J Med Internet Res 2026;28:e104524) doi:10.2196/104524

KEYWORDS

chronic pain; pain management; telemedicine; artificial intelligence; symptom assessment; wearable electronic devices; virtual reality; biomedical technology

Textbox .

Key Takeaways:

- Digital pain management tools can help people with chronic pain as well as provide health care providers with a better understanding of what these patients feel and potentially need.
- Tracking apps provide context and data for long-term symptoms.
- Virtual reality, adaptive neurostimulation, and adaptive medical AI advances provide promise for the future.

A recent review suggests that over a quarter of adults experience chronic pain. Many struggle to find therapies and managements that provide relief, partially because pain can have a significant neurological component, which researchers are still learning more about.

Eric Anderson, MD, neurologist and chief medical officer at Lin Health, has developed care models in the digital/virtual space that address the biological, psychological, and social drivers of chronic pain and that focus specifically on primary pain or neuroplastic pain.

“All pain signaling comes from the brain,” said Anderson, “When we touch something, mechanical receptors tell us there’s pressure, temperature, something noxious. Our brain processes that and determines what to do with it. In primary pain, this becomes somewhat maladaptive. The brain becomes overprotective, registering things that are no longer pain or that *could* be pain as pain.”

Our brains may also send preemptive pain signaling, anticipating an action that has caused pain in the past. Neuroimaging supports this, showing that acute pain lights up the somatosensory cortex while chronic pain registers in the limbic region, morphing into learned memory.

Lin Health, a virtual program, downloadable app, and medical group with psychiatrists, pain specialists, and behavioral health care managers, works within this concept of pain.

“We use the app with coaches and a guiding medical team to help people understand and validate what they’re going through,” Anderson explained.

Digital Tracking Tools

“We have a very expensive healthcare system that is very good at keeping you alive. After that, we’re not as good,” said Rebecca Abraham, RN, the founder of Plesio Health, a company that builds technology products for pain symptom management. “Everybody passes chronic pain treatment to somebody else.”

Abraham hopes to change how chronic pain and frequently co-occurring sleep issues, anxiety, and depression are treated. Plesio focuses on the use of cannabis, hemp, and other substances and supplements to manage symptoms.

Using algorithms based on nursing theory and traditional medicine, Plesio Health has built tools around product guides and risk management: a care plan generator to inform online buyers, a pharmaceutical drug interaction checker, and a dosing tracker to input symptoms, use, and results. This dosing tracker provides a way to quantify results, something that didn’t previously exist.

“We made it more technical and applied a quantifiable number to it. This allows patients to tell us how they feel, which helps us help them manage their pain,” said Abraham.

Real-World Results

Nicole Notar, founder and CEO of Vindicara, has dealt with chronic pain associated with endometriosis for 15 years. Symptom-tracking apps have helped her more than she expected.

“When you’re in pain often, everything kind of blends together, so logging things made me notice patterns I wouldn’t have picked up otherwise,” she explained.

XSL•FO
RenderX

While many people think of endometriosis as a "menstrual disease," endometriotic tissue has been found on every organ in the body. In Notar's case, this includes the bowel, colon, rectum, bladder, abdomen, and uterus. At age 16, she was missing work and school and frequenting emergency rooms. Doctors told her parents there was nothing wrong with her, suggesting she just didn't want to go to school. They insisted the pain was in her head; she was too young to have endometriosis and, even if she did, there was nothing to do about it.

In Notar's chart, doctors would write, "*takes eight Motrin a day, still in debilitating pain*."

Figure WL1.

"That's data," Notar said. "It's not an ICD-10 code [*International Statistical Classification of Diseases and Related Health Problems, 10th Revision*], but it is data. And it was constantly ignored."

This is where Notar's plugin comes in. It flags that data and assigns a risk score with an explanation. Targeting a launch of July 2026, it will integrate into electronic health record (EHR) systems including Epic, MyChart, and Athena Health. Notar's hope is that it will prevent others from going through what she did.

"Digital tools and virtual programs should be bolted on to a service that extends the reach of an existing medical program. Augment, not replace.

Eric Anderson, MD

Notar has used a variety of apps to manage her chronic pain, including Natural Cycles, a regulated medical device often used for pregnancy planning. Paired with the included thermometer or other compatible wearable, it tracks body temperature.

"This creates a whole new form of tracking data: flare day patterns; gastro flares or bladder tightness, which helps me prep for gynecologist visits. I have data that shows eleven flares in the last eight weeks."

Through her foundation, Endo Excision for All, people reach out from all over. Generally, those in larger cities use tracking apps, while many from smaller towns may not have even heard about them.

"Maybe they don't have an Instagram," explained Notar, "They don't see all the ads. How does that whole community prepare for their appointments? They don't. We need a tool for people to take their notes even if providers don't believe them."

"Digital tools and virtual programs should be bolted on to a service that extends the reach of an existing medical program," said Anderson. "Augment, not replace."

While a digital calculator for certain algorithmic scales makes a lot of sense, these apps can be taken out of context. To be more useful and produce outcomes, they must be put into the appropriate clinical context. For instance, you can measure your blood pressure, but you can't put it in the context of your entire

XSL•FO
RenderX

person, your health, and what it means with only one stand-alone tool.

Beyond Tracking Apps

Beyond tracking, digital innovations for pain relief and treatment continue to be developed.

In the chronic pain treatment space, neuromodulation devices can "rewire pain signaling." These therapies, which include spinal cord and deep brain stimulation, work via wireless implants that can respond in real time to reduce pain. Such approaches could be further coupled with adaptive AI that could learn patient patterns and help provide relief from pain and even prevent flare-ups before they manifest.

In 2022, the US Food and Drug Administration approved virtual reality systems as a chronic pain management tool. A crossover study published in April 2025 showed that telehealth virtual reality interventions can reduce chronic pain.

Dimitri Souza, MD, a physician at Western Reserve Hospital Center for Pain Medicine, described watching a patient with sickle cell disease, considered among the most painful conditions, using a virtual reality system.

"He was in a different world, flying between the hills. He was so excited and happy and there was no sense of pain on his face. It's a really powerful tool. And it is available. Like the endorphins and pain protection systems we all have. They're there [for us to use]. We just need to learn how."

Anyone who suffers from chronic pain knows that management and relief never come from a singular place or in a straight line. It is as complicated and unique as the pain itself, requiring tools that allow for nuanced and individualized tracking, monitoring, feedback, and treatment. As AI and other digital advances become more widespread, a future that eases the effects of chronic pain becomes possible and feasible.

Edited by KA Clegg; submitted 12.Jun.2026; this is a non–peer-reviewed article;accepted 12.Jun.2026; published 22.Jun.2026.

Please cite as:
Nirode V
Potential of Digital Tools for Chronic Pain Management
J Med Internet Res 2026;28:e104524
URL: https://www.jmir.org/2026/1/e104524
doi:10.2196/104524

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.Jun.2026.

XSL•FO
RenderX

The Ethics of Extending Life: Longevity Medicine and Health Inequity

Jenna Congdon

Abstract

The longevity industry is rapidly expanding, driven by advances in biotechnology and consumer demand. In this *News and Perspectives* article, JMIR Correspondent Jenna Congdon reports on recent developments in longevity medicine alongside potential risks and ethical concerns.

(J Med Internet Res 2026;28:e104414) doi:10.2196/104414

KEYWORDS

longevity medicine; aging; preventive medicine; precision medicine; genomics; artificial intelligence; health equity; health literacy; socioeconomic factors

Textbox .

Key Takeaways:

- Longevity medicine aims to extend health span through personalized, preventive care using tools like genomics, biomarker tracking, and AI-driven interventions rather than reactive disease treatment.
- High costs, unequal access, and health literacy disparities risk making longevity medicine a privilege largely reserved for wealthy, well-educated populations, potentially widening existing health inequities.
- Ethical integration of longevity medicine requires equitable pricing, diverse research, stronger regulation, and incorporation into public health systems to make healthier aging accessible beyond affluent groups.

In one survey, nearly 80% of 1000 US respondents said they'd like to live to be 120 years old, but *only* if they remained mentally and physically healthy. Without that hypothetical guarantee, far fewer people wished to live such a long life. We not only want to live longer, we want to live longer *and* better.

From stem cell therapy to biomarker tracking, from genomics to AI algorithms that support early disease detection, medicine has entered a new era in which some view aging as a condition that can be managed and mitigated, provided the right tools are used. This emerging field is known as longevity medicine—a prevention-focused approach to personalized, precision medicine that emphasizes early detection and treatment of age-related decline with the goal of extending not only one's lifespan, but also one's healthy years.

Due to limited accessibility and high costs, those in economically and socially privileged positions have been the first to benefit from these advancements. As life-lengthening medical interventions continue to develop, health span extension may become stratified along socioeconomic lines, concentrating its benefits among the already privileged. Without intentional policy and ethical frameworks, these innovations may deepen population health inequities.

What Is Longevity Medicine?

Longevity medicine has one tenet at its core: don't just react, prevent. Whereas standard medical care waits until blood work demonstrates high cholesterol and then offers treatment, longevity medicine takes a more proactive path. It may involve, for example, analyzing one's genetic code in search of inherited predisposition to hypercholesterolemia and then developing a plan to help the patient maintain a healthy lipid profile for life, using tools that range from standard nutrition and exercise advice to AI-driven models that recommend hyperpersonalized interventions.

This in-depth medical care is often offered through concierge clinics accepting out-of-pocket payments. Many of these approaches are still emerging and unstandardized, but their popularity reflects the growing commercialization of aging science.

Financial and Socioeconomic Access Barriers

When Jordan Weiss, PhD, gives talks about the health inequities surrounding aging, he often asks the audience to raise their hands if they'd like to live to be 150 years old. When he goes on to ask if they'd like to be 150 years old, but in poor health and without the financial resources to cover basic bills, most hands drop, echoing the response to the aforementioned survey.

As Assistant Professor at the Optimal Aging Institute of NYU Grossman School of Medicine and Scientific Writer at Assisted Living Magazine, Weiss focuses on health equity from the lens of how long we live and how many of those years are spent in

XSL•FO
RenderX

good health. In his view, living longer isn't the whole answer: "The question is, what's the upside of living or extending how long you live if you still can't afford to pay your rent or feed yourself?"

Longevity medicine is largely inaccessible to lower-income populations, due to both cost and socioeconomic barriers. Income is a major driver of mortality; Weiss cites a 2016 paper demonstrating a staggering 14.6-year life expectancy gap between the top and bottom 1% of male earners, and a 10-year difference between the top and bottom 1% of female earners.

Weiss illuminates another, more subtle accessibility challenge facing longevity medicine: "Even if you made every [longevity medicine] intervention free, you still need to…interpret all that information." That requires the consumer to possess a high degree of health literacy, which is typically higher in well-educated, high-income individuals.

"Wealthy patients walking into longevity clinics aren't just buying the intervention, they're buying someone to do that cognitive work and navigation for them, to be the quarterback of their health, or the CEO of their health, which is a service that's almost completely absent from primary care for most other people." Without someone to provide context and guidance, the large amount of data that longevity medicine testing reveals may simply be disregarded, or worse, widen disparities.

Ethical and Public Health Implications: Widening the Gap

When discussing the ethical implications of longevity medicine access, questions abound: might longevity increasingly become a form of privilege, and is it acceptable for life-extending technologies to be unequally distributed?

Weiss asks, "Who benefits, who gets left behind, and does the innovation unintentionally widen healthspan inequality? Many longevity interventions are preventive, and they accumulate over time, so unequal adoption can create widening divergence over decades." These questions must be weighed and addressed by policymakers, clinicians, and even laypeople as health innovation continues to progress.

Figure WL1.

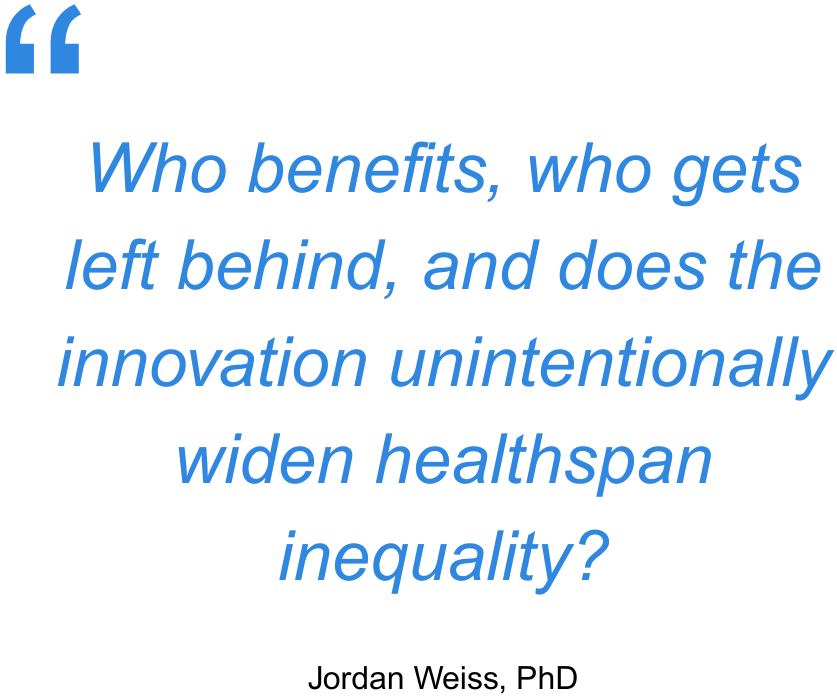

Furthermore, the level of precision health relied upon by longevity medicine assumes a level of health engagement and empowerment that is not equally distributed. Chronic disease burden is already higher in socioeconomically disadvantaged populations.

Weiss points out that "feeling fine may mean something different to someone who's grown up in a world where chronic disease is more prevalent." They're then less likely to seek medical care early, potentially altering the course of their health span for decades to come. Focusing on the highly individualized interventions offered by longevity medicine risks shifting focus

XSL•FO
RenderX

away from population-level interventions that improve the health of millions, rather than only a few.

Are There Ethical Pathways Forward?

There are tangible steps toward more ethical integration of longevity medicine interventions that companies, clinicians, and policymakers must consider.

Study design on new innovations must center equity by including diverse sample groups and planning interventions that work within the context of lower socioeconomic communities. Companies must price products with equity in mind and market honest claims regarding the safety and efficacy of their products. Policymakers must create legal mechanisms that ensure pricing and access for all, as well as transparent marketing practices. Finally, leaders, medical professionals, and community members must focus on integrating longevity medicine principles into the greater public health system.

Figure WL2.

> *...The future of longevity medicine will be defined by not just how long we can live, but by who gets the opportunity to do so.*

The clinician's role is that of the interpreter: with more and more data on their hands, consumers are unsure which to act on and which they can safely ignore. Weiss points out that more data isn't necessarily better: "If you take a test and it turns out you have a genetic predisposition to Alzheimer's disease, which doesn't really have a cure, what do you do with that information? How does that leave people feeling? How do we package and present these results in a way that's actually helpful?"

Longevity at a Crossroads

Longevity medicine offers fantastic potential for increasing health span—a hope that many of us dream of—but the future of longevity medicine will be defined by not just how long we can live, but by who gets the opportunity to do so.

By integrating these advancements into the broader public health system and carefully balancing access so that not only the wealthy benefit, we can harness the power of medicine's next big frontier: our futures.

XSL•FO
RenderX

Edited by KA Clegg; submitted 11.Jun.2026; this is a non–peer-reviewed article;accepted 11.Jun.2026; published 19.Jun.2026.

Please cite as:
Congdon J
The Ethics of Extending Life: Longevity Medicine and Health Inequity
J Med Internet Res 2026;28:e104414
URL: https://www.jmir.org/2026/1/e104414
doi:10.2196/104414

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Jun.2026.

XSL•FO
RenderX

Written Exposure Therapy as Transformative, Scalable Care for Posttraumatic Stress Disorder

Candice Marie Sage

Abstract

International organizations are calling for urgent scaling of mental health care services, with the World Health Organization noting over a billion people living with mental health conditions in 2025. In this *News and Perspectives* article, JMIR Correspondent Candice Marie Sage reports on a brief scalable intervention for posttraumatic stress disorder.

(J Med Internet Res 2026;28:e103501) doi:10.2196/103501

KEYWORDS

posttraumatic stress disorder; mental health services; global health; access to health care; scalability; psychological intervention; narrative medicine; digital narrative medicine

Textbox .

Key Takeaways:

- Written exposure therapy (WET) is a brief intervention for posttraumatic stress disorder that shows promise compared to gold-standard treatments.
- WET shows high uptake, feasibility, and scalability in real-world care settings.
- WET can promote health equity through online deployment as a digital narrative medicine intervention.

Candice Marie Sage, PhD, is a researcher with over 20 years of experience managing interdisciplinary research in Canada and internationally. She is an expert in knowledge cocreation, translation, and exchange. In this piece, she continues her series on digital narrative medicine stories and applications.

Posttraumatic stress disorder (PTSD) impacts 3.9% of the population worldwide, with significant psychiatric and medical sickness in the absence of effective treatment. According to World Health Organization data across 15 countries, only 43.0% of people diagnosed with PTSD over a 12-month period were able to access mental health services. Of those, only 13.5% received adequate pharmacotherapy and 17.2% adequate psychotherapy. Accessibility drops for populations such as veterans and those without insurance, as well as low- and middle-income countries.

As a person with complex PTSD in the Canadian context, it has taken me over two years to access publicly funded care, and my healing journey is far from over. Affordability as well as a shortage of qualified mental health professionals are barriers to PTSD care in Canada.

What if one of the most powerful interventions to address PTSD begins with a story written by the patients themselves? An evidence-based PTSD treatment known as written exposure therapy (WET) asks patients to write the story of their traumatic experience across five facilitated, structured sessions. These patient-authored narratives—including in-person and telehealth digital formats within digital narrative medicine—can help bridge the gap to accessing effective PTSD treatment. WET shows promise as a scalable, much needed public health intervention.

As Effective as More Intensive PTSD Treatments

The WET process was codeveloped by Denise Sloan, PhD, a clinical psychologist whose research focuses on narrative processing, and Brian Marx, PhD, a clinician investigator studying the implementation of trauma therapies in health care systems. Working from the US Department of Veterans Affairs' National Center for PTSD, their collaboration began in 2004, investigating expressive writing as a treatment for PTSD symptoms. The WET protocol was refined over a decade of systematic empirical studies. In 2019, a treatment manual for WET was widely disseminated as a clearly defined clinical approach. A 2024 systematic review that examined a variety of settings and trauma populations across several countries found WET to be an efficacious and effective treatment for PTSD symptoms.

While WET is often delivered in person in clinical settings, it can also be delivered via video meetings, with patients submitting images of their written accounts via a secure and confidential electronic system. As piloted during the COVID-19 lockdown, digitally delivered WET supported feasibility and flexibility in different settings with similar effectiveness as in-person care.

I interviewed a WET practitioner and trainer, Stefanie T LoSavio, PhD, ABPP, associate professor of psychiatry and

XSL•FO
RenderX

behavioral sciences and director of research and innovation at the STRONG STAR Training Initiative at the University of Texas at San Antonio. LoSavio emphasized that WET has shown clinically significant reductions in PTSD symptoms and is comparable in effectiveness to more intensive treatments like cognitive processing therapy and prolonged exposure therapy, despite being much shorter. WET's brevity and reduced therapist interaction were key innovations when it first gained attention in the field.

Activating Patients' Natural Recovery Mechanisms

Inquiring how WET works so well with less therapist interaction, LoSavio credited the narrative writing process itself. WET involves confronting the memory of a traumatic event so that the person can process and recover from the experience. They do this by writing repeatedly about the details of the event and, later, its impact on them. By approaching and processing the memory instead of avoiding it, the person can learn that they can tolerate the memory, lessen distress associated with it, and make sense of the experience such that it becomes a part of their history instead of showing up in daily PTSD symptoms. Patients often experience cognitive shifts spontaneously through writing about their trauma, which helps them remember details and address distressing beliefs like why the event happened and why it was not prevented.

Figure WL2.

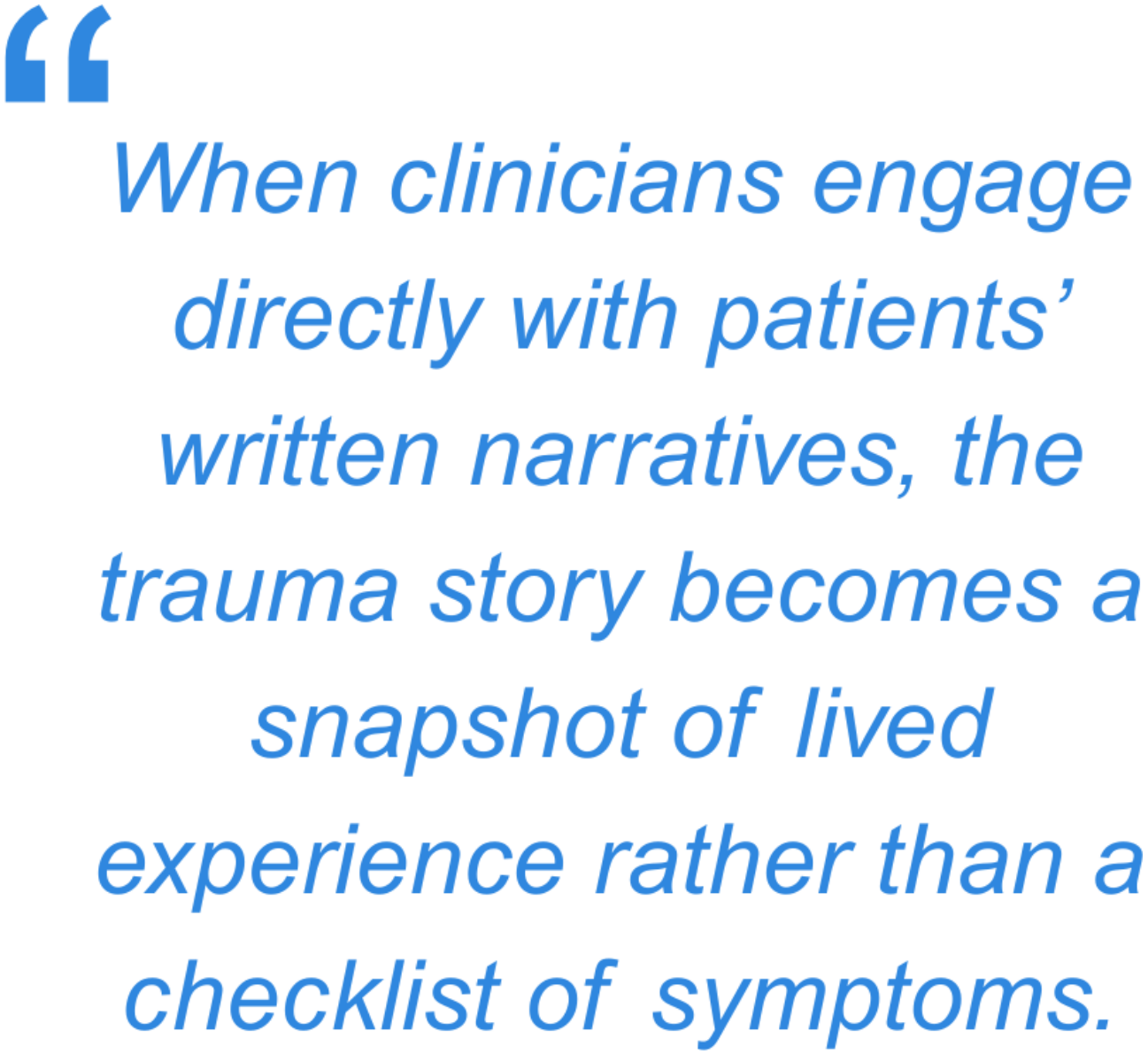

LoSavio notes, "Patients have an opportunity to recall the full details of the trauma, including what happened and what they were thinking and feeling at the time. They're not writing it like a report for someone else but for themselves in order to recover from the trauma."

Progress occurs as patients access their natural recovery mechanisms by addressing avoidance and enabling the processing of traumatic memories. She says, "People have good mechanisms to process and recover from bad events, but patients with PTSD get 'stuck' in the process." Interestingly, LoSavio

XSL•FO
RenderX

notes that some patients continue to improve even after the treatment ends, carrying on their own processing work past the therapeutic sessions.

She reports that patients typically find the writing process gets easier over time as they develop a sense of increased control over their memories rather than being controlled by them. Their confidence grows in their competence to feel their emotions rather than avoiding them. Also, for many, recounting the details of the trauma in their writing can help them reconsider unhelpful interpretations, like erroneously assuming their trauma was their fault.

Patient narratives can become more than a therapeutic exercise: they can function as a bridge between patient experience and clinical care. When clinicians engage directly with patients' written narratives, the trauma story becomes a snapshot of lived experience rather than a checklist of symptoms. This could deepen empathy, improve diagnostic clarity, and help clinicians understand how trauma shapes a patient's life beyond the clinical encounter.

LoSavio noted that while some clinicians have delivered WET in a group format, published clinical trials involve individual therapy. She trains providers in WET through foundational workshops and consultation calls. Therapists can seek guidance on patient appropriateness, treatment planning, and feedback on deidentified patient narratives. The next frontier of training involves professionals like peer supports whose roles intersect with patients' lives. This can alleviate wait times to access WET where there is a shortage of therapists and could be part of a larger public health strategy. She detailed the ethical considerations around handling patient narratives, including secure storage and disposal methods, emphasizing the importance of maintaining patient privacy. Thus, narratives are not meant to be stored in electronic patient records or student files.

Digital Delivery to Scale WET to Underserved Populations

LoSavio is confident that the development and empirical basis of WET holds potential for broader implementation of this brief effective treatment. Further research will inform culturally responsive care and alternative delivery methods. Decentralization of care and integration of WET into existing nonspecialty settings is happening in the United States, with treatment delivered in primary care clinics, addiction programs, and veterans' telehealth systems, and to people experiencing homelessness. She notes that studies show comparable outcomes in telehealth versus in-person program delivery, with the potential to reach rural populations, mobility-limited patients, and underserved communities by expanding reach beyond traditional clinics.

WET is a pathway to effective and more accessible PTSD care. Technology as a scaling tool can further foster health equity by reaching the kinds of populations often excluded from traditional trauma therapy due to cost or specialist bottlenecks.

"PTSD is treatable, and it doesn't have to take years," LoSavio concludes. "Current reach of evidence-based treatments for PTSD is low, and WET can help with access."

Indeed, WET is a promising example of interventions and technology that could alleviate much human suffering.

Edited by KA Clegg; submitted 03.Jun.2026; this is a non–peer-reviewed article;accepted 03.Jun.2026; published 15.Jun.2026.

Please cite as:
Sage CM
Written Exposure Therapy as Transformative, Scalable Care for Posttraumatic Stress Disorder
J Med Internet Res 2026;28:e103501
URL: https://www.jmir.org/2026/1/e103501
doi:10.2196/103501

© JMIR Publication. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

When Lived Experience Designs the Intervention

Trevor van Mierlo

(J Med Internet Res 2026;28:e91371) doi:10.2196/91371

KEYWORDS

mental health; telemedicine; health services accessibility; cultural competency; patient participation; patient-centered care; co-design; lived experience

Textbox .

Key Takeaways

- Co-developing a digital mental health intervention with contributors who have lived experience of displacement revealed how language, meaning, engagement, and design choices can carry unanticipated cultural and emotional risks.
- To ensure safety, trust, and effectiveness on a global scale, interventions must move beyond adaptation to lived experience co-design.

Digital mental health interventions are increasingly deployed across borders, reaching people whose lives and circumstances may differ dramatically from the teams who build them. As these tools proliferate, traditional approaches to cultural adaptation—often focused on translation—have become insufficient.

Translation conveys language; it doesn't always convey *meaning*. And meaning is culturally constructed and emotionally rooted in ways that can be impossible to anticipate from the outside. I recently learned this lesson through co-developing a mental health resilience course with a team of Ukrainian software developers—many of whom were displaced by war, and some who were actively participating in it.

This experience illuminated a frequent blind spot in global digital mental health design: without lived experience embedded in creation—not merely consulted afterward—we risk building tools that are well-intentioned and evidence-based but potentially unsafe.

Digital Mental Health at Scale and Cultural Distance

Scalable, low-cost interventions promise to reach global populations that traditional services cannot, particularly during humanitarian crises, displacement, and conflict. Yet most digital mental health tools are conceived, designed, and tested in relatively stable Western contexts before being deployed elsewhere [1-3].

This development pattern introduces cultural distance at precisely the point where sensitivity matters most. Evidence of clinical efficacy does not ensure cultural legitimacy, trust, or safety [4]. When interventions cross geopolitical and cultural boundaries, they enter environments shaped by historical trauma, media narratives, power asymmetries, and collective memory. These contextual forces rarely appear in design specifications, yet they profoundly shape how digital tools are perceived and used.

The challenge, then, is achieving scale without reproducing blind spots that undermine efficacy.

From Adaptation to Co-Development

Most cross-cultural digital health efforts rely on cultural adaptation frameworks [5]. These typically involve translating content, substituting examples, and adjusting tone to fit a new population. While necessary, these steps assume that meaning is largely transferable and remains intact once linguistic barriers are removed.

Our experience suggests otherwise. When conducted without lived experience embedded in the design process, cultural adaptation risks addressing surface differences while missing deeper layers of meaning.

In contrast, lived-experience co-development treats those with direct experience not as informants, but as co-designers. This approach requires cultural and epistemic humility [6]: acknowledgment that certain insights cannot be inferred, researched, or validated externally. They must be shared by those who live within the context the intervention seeks to address.

Collaboration With Lived-Experience Designers

The collaboration at the center of this article emerged when the Ukrainian team—displaced by war, living in Ukraine, or actively deployed—approached our group at Evolution Health to help build a digital mental health resource for Ukrainians living in displacement and uncertainty.

These developers were active contributors throughout development, shaping decisions about language, tone, pacing, and presentation. Their technical expertise allowed them to participate fully in design discussions, while their lived experience grounded those discussions in their reality. Decisions were no longer guided solely by evidence hierarchies or best practices, but by continuous dialog about how content would

XSL•FO
RenderX

be interpreted emotionally, culturally, and symbolically by people living with war-related disruption.

Figure WL1. Left: original English draft. Right: adapted Ukrainian version.

The divergences were striking. Even when words appeared similar, their emotional weight and associations differed substantially. Terms that Western designers associated with isolation, anger, exhaustion, disconnectedness, or optimism carried layered meanings for the Ukrainian team. Some evoked loss. Others echoed bureaucratic language encountered during displacement. Still others were associated with political messaging or wartime media narratives.

Two examples help illustrate the limitations of conventional adaptation and the benefits of co-design.

The first involved a simple vocabulary exercise. The North American team generated a word cloud, and the Ukrainian team was asked to identify words that felt emotionally relevant, neutral, or problematic within the context of displacement and war and to add or remove words accordingly.

The second example involved an introductory video developed using an artificial intelligence (AI)–generated Ukrainian-language voice-over. From a Western perspective, the voice sounded professional, neutral, and supportive—a practical solution that aligned with common digital production practices.

The video was swapped, however, when the Ukrainian team explained that it was likely to evoke distrust. Synthetic Ukrainian voices had become widely associated with Russian disinformation campaigns, where AI-generated Ukrainian speech was routinely used to deliver propaganda.

This risk was invisible to the non-Ukrainian team. Traditional usability testing might have detected disengagement, but only after trust had already been compromised. The issue was not usability, but cultural safety.

These examples underscore two central lessons: that language is not merely semantic, and that trust is built on signals of authenticity, safety, and legitimacy that are culturally situated. Without lived experience in design, even carefully translated content and well-intentioned choices can misfire and risk harm.

What Lived Experience Changed in Practice

Co-development reshaped the intervention in several ways.

Language as Lived Experience

First, meaning consistently took precedence over literal accuracy. Language that was technically correct but emotionally misaligned was revised or discarded. The goal shifted from fidelity to original phrasing toward fidelity to lived interpretation.

Second, engagement strategies were reconsidered. Structures intended to support participation occasionally felt directive or intrusive within a displacement context [7]. Lived experience helped distinguish between supportive guidance and unwanted pressure, particularly for individuals navigating chronic uncertainty and loss of agency.

Finally, harm prevention became an explicit design outcome. Lived-experience contributors identified risks that would not have appeared in formal risk assessments—symbolic associations, emotional triggers, and trust signals embedded in design choices. These insights allowed potential harms to be addressed before deployment, rather than retroactively.

Ethical Responsibilities

There are ethical responsibilities inherent in working with individuals who are actively living through the experiences an intervention seeks to address. Co-development can be experienced as empowering and meaningful, but it also raises questions about psychological safety that appear insufficiently explored in the digital mental health literature [7].

Even when voluntary, participation can involve exposure to emotionally charged language and experiences, increasing the potential for distress. At present, there is limited empirical guidance on how to assess, monitor, or mitigate potential psychological risks for lived-experience contributors in digital intervention development [8].

Designing With, Not For

This collaboration revealed limits in what even experienced digital mental health designers can know from the outside. Lived experience did not simply improve the intervention; it fundamentally reshaped it. It prevented harm, deepened resonance, and challenged assumptions embedded in the original, North American content.

As digital mental health tools continue to scale, lived-experience co-development should be treated as a methodological and

XSL•FO
RenderX

ethical necessity rather than an optional enhancement [9]. Even technically sound, evidence-based design alone cannot prevent contextual misalignment and ensure cultural legitimacy, trust, or safety.

Researchers and designers must consider how lived experience is integrated, compensated, sustained, and supported throughout development. Funders and policymakers should evaluate interventions not only on efficacy, but on how they address cultural meaning and safety.

Ultimately, if digital mental health interventions are to be trusted, culturally legitimate, and safe at global scale, the field must move beyond designing *for* others toward designing *with* them.

Conflicts of Interest

The author is the founder and a strategic advisor of the Evolution Health platform and acts as a consultant for academic and commercial digital mental health initiatives. The views expressed in this article are those of the author.

References

1. Chakrabarti S. Digital psychiatry in low-and-middle-income countries: new developments and the way forward. World J Psychiatry 2024 Mar 19;14(3):350-361. [doi: 10.5498/wjp.v14.i3.350] [Medline: 38617977]
2. Chau LW, Lam RW, Minas H, Hayashi K, Nguyen VC, O'Neil J. Digital health interventions for depression and anxiety in low- and middle-income countries: rapid scoping review. JMIR Ment Health 2025 Aug 22;12:e68296. [doi: 10.2196/68296] [Medline: 40846319]
3. Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? Behav Brain Sci 2010 Jun;33(2-3):61-83. [doi: 10.1017/S0140525X0999152X] [Medline: 20550733]
4. Mohr DC, Riper H, Schueller SM. A solution-focused research approach to achieve an implementable revolution in digital mental health. JAMA Psychiatry 2018 Feb 1;75(2):113-114. [doi: 10.1001/jamapsychiatry.2017.3838] [Medline: 29238805]
5. Bernal G, Domenech Rodríguez MM. Cultural Adaptations: Tools for Evidence-Based Practice with Diverse Populations: American Psychological Association; 2012. [doi: 10.1037/13752-000]
6. Mucić D, Hilty DM, Yellowlees PM. Digital mental health toward cross-cultural populations worldwide. In: Digital Mental Health: The Future Is Now: Springer Nature; 2025:175-211. [doi: 10.1007/978-3-031-59936-1_8]
7. Cargo M, Mercer SL. The value and challenges of participatory research: strengthening its practice. Annu Rev Public Health 2008;29:325-350. [doi: 10.1146/annurev.publhealth.29.091307.083824] [Medline: 18173388]
8. Roennfeldt H, Stewart V, Wyder M, et al. Scoping review of co-design in mental health research: essential elements and recommendations. OTJR: Occupational Therapy Journal of Research 2025. [doi: 10.1177/15394492251367259]
9. Muchamore I, Karanikolas P, Gooding P. How Lived Experience Expertise Shapes Research and Development in Digital Mental Health - A Review of Literature and Insights: Wellcome; 2024. [doi: 10.2139/ssrn.4932039]

Edited by KA Clegg; submitted 13.Jan.2026; this is a non–peer-reviewed article;accepted 13.Jan.2026; published 23.Jan.2026.

Please cite as:
van Mierlo T
When Lived Experience Designs the Intervention
J Med Internet Res 2026;28:e91371
URL: https://www.jmir.org/2026/1/e91371
doi:10.2196/91371

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Jan.2026.

XSL•FO
RenderX

A Frontline Worker's Take on Hybrid Care Implementation in the Hospital Setting

Jenna Congdon

(J Med Internet Res 2026;28:e90879) doi:10.2196/90879

KEYWORDS

telemedicine; delivery of health care, integrated; digital health; hospital medicine - organization & administration; health personnel attitudes; hybrid care implementation

Textbox .

Key Takeaways

- Hybrid care can enhance patient outcomes and workflow efficiency, but only when thoughtfully implemented with frontline staff input.
- Variability in hospital resources, patient populations, and infrastructure leads to inconsistent adoption and effectiveness of hybrid care models.
- Without adequate support, training, and realistic workflow planning, hybrid technologies risk adding burden to bedside staff and generating ineffective spending rather than improving care.

Hybrid care in the hospital setting—combining in-person care with telehealth services—has increased dramatically since 2020 [1,2]. During the COVID-19 pandemic, frontline workers relied on remote access to care for patients from afar. In the succeeding years, hybrid care use has remained much higher than prepandemic levels [1]. At different points in my own practice as an intensive care nurse, depending on its implementation, hybrid care technology has been a distracting hindrance to patient care or has greatly supported my ability to treat patients in critical condition.

For other staff working at the bedside, reactions are similarly mixed: while some report safer patient care and more efficient workflows, others feel that hospitals now push telehealth technology as an ineffectual fix for staffing shortages and organizational issues.

Hybrid Care on the Ground

The day-to-day realities of in-hospital hybrid care use are varied. Location, patient demographics, and facility-to-facility differences in budgets and priorities play a part in determining the specific implementation, delivery models, and rate of adoption by both patients and staff for these modes of care [3].

Examples include using video calls to patch in specialty providers for emergencies and consultations, providing remote physician rounding for rural or underserved communities, and leveraging digital apps for patient communication and video visits.

Additional uses include monitoring confused or unstable patients via video; supporting nursing workflows with remote nurse assistance; facilitating admissions, discharges, and patient education; and offering a way for families to connect virtually with loved ones if they are unable to visit in person [1].

These tools represent a unique opportunity to augment patient care and improve safety measures. However, without collaborative implementation and ongoing support services, some frontline staff express concern that more technology simply adds to their already burdensome workload [4-6].

When Hybrid Care Saves the Day

At its best, hybrid care technology can save lives and ease stress for bedside staff, patients, and their loved ones.

Gwen*, a nurse in a postanesthesia care unit in Portland, Oregon, recalls a time when her patient experienced a stroke soon after surgery: "In that situation, we were able to use the telehealth neurologist during a code stroke. It worked really smoothly, and the other staff and I felt more confident in that emergency. That patient absolutely received better care because we were able to talk to a neurologist immediately." Gwen's experience demonstrates how faster access to specialty care via a remote provider can dramatically improve patient outcomes.

Avah*, a nurse who works in the postpartum unit of a hospital in Madison, Wisconsin, reported: "Our NICU [neonatal intensive care unit] has cameras that let parents see their baby from home." She cites this use of telehealth technology as an important bonding tool for new parents of hospitalized infants.

Other benefits to hybrid care include fewer admissions and readmissions, reduced length of hospital stay, improved chronic disease management, enhanced patient safety, reduced health system costs, better integration of care services, better adherence to recommended best practices, and increased efficiency in provider workflows [4,5].

When Hybrid Care Gets in the Way

Nurses frequently become the point person for new in-hospital technology, including hybrid care tools. When patients struggle to log in to a phone app for a video appointment or the camera used for physician consults needs to be set up, it often falls to

XSL•FO
RenderX

the nurse to first troubleshoot and then call for assistance if needed. They become the liaison for patient concerns and digital snafus, taking time away from patient care activities and adding to an already hectic workload.

Other concerns relayed by frontline health care workers regarding newly integrated hybrid care models include technical issues such as unstable Wi-Fi or poor equipment maintenance, patient and staff competency in adopting new technologies, and lack of ongoing support as uses for hybrid care modalities evolve [4,5]. Additional technology may also create interruptions in the workflow and degrade the quality of patient-nurse interactions [6].

Rhonda*, an intensive care unit charge nurse at a large Midwestern hospital, explains that her unit recently installed electronic intensive care unit (eICU) cameras in each patient room. These devices are meant to bring in ICU-trained registered nurses (RNs) via video to provide nursing support. “We rarely use them, and we never asked for this kind of device. Instead of being a help, they are a hindrance. The eICU staff call us to ask questions that are irrelevant to patient care. It wastes time. I’m sure [the cameras] were expensive, and they don’t help patient care in any way.”

She worries that this follows a recent trend at her facility of cutting staff and attempting to replace in-person RNs with nurses who can only watch from the other side of a screen.

These challenges and concerns suggest that additional technology is not a replacement for proper staffing or efficient organizational workflows.

Making the Most of Hybrid Care Tools

If in-hospital hybrid care models are to be effective, nurses and other frontline staff must be involved in the entire decision-making process, not brought in after adoption [5,6]. The choice of hybrid care technologies needs to be decided based on real-life clinical needs and centered around realistic patient behaviors. If patients struggle to use an app or cannot understand a provider on the other end of a video call, the technology becomes useless, burdensome to staff, or worse, an active impediment to the quality of patient care.

Designing for flexibility is crucial as technology develops, patient needs shift, and health care continues to see rapid change [7]. Ongoing and easily accessible tech support for both patients and staff should be included to increase use and reduce errors, work-arounds, and frustrating workflow disruptions [6]. While staff are often receptive to new technologies becoming part of their daily workflow, needs and perceptions vary by facility and department, highlighting the need for a tailored approach that includes bedside staff’s voices [8].

In-Hospital Hybrid Care Succeeds When Frontline Staff Are Involved

While hybrid care can be a powerful tool, everyday complications show gaps that hospital leadership and policymakers sometimes overlook. Without the involvement of these frontline workers, hospitals risk spending budget dollars on technology that simply does not serve its full purpose.

By continually seeking out and incorporating the input of frontline staff, health systems and hospital administrators can make realistic, informed, and effective decisions about which tools will increase safety, efficiency, staff retention, and patient experience in their facilities. Administrators should survey frontline staff about their daily challenges and perspectives on which models of hybrid care would best address staff and patient needs before purchasing, policy change, or implementation begins. Additionally, nursing focus groups may provide an opportunity for unit-specific representation so that hybrid care execution can be tailored to departmental needs.

Hybrid care tools cannot replace the human touch in health care, but they certainly have the power to create better outcomes for patients and lighten staff workloads. Hospital systems should invest in thorough research and open conversation with staff to generate flexible and human-centered hybrid care models that best serve the people under their care.

**Names changed to protect anonymity.*

Conflicts of Interest

None declared.

References

1. Hehman MC, Fontenot NM, Drake GK, Musgrove RS. Leveraging digital technology in nursing. Health Emerg Disaster Nurs 2023;10(1):41-45. [doi: 10.24298/hedn.2022-0014]
2. OECD. The COVID-19 pandemic and the future of telemedicine. : OECD Publishing; 2023 Jan 17. [doi: 10.1787/ac8b0a27-en]
3. Chen J, Amaize A, Barath D. Evaluating telehealth adoption and related barriers among hospitals located in rural and urban areas. J Rural Health 2021 Sep;37(4):801-811. [doi: 10.1111/jrh.12534] [Medline: 33180363]
4. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023 Sep 18;6(1):161. [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
5. Kumari AA, Wani TA, Liem M, Boyd J, Khan UR. Advancing regional and remote health care with virtual hospital implementation: rapid review. JMIR Hum Factors 2025 Jun 3;12:e64582. [doi: 10.2196/64582] [Medline: 40460425]

XSL•FO
RenderX

6. Hassell A, Morelock B, Greeson J, Thomson A, Varty M. Lessons learned from systemwide implementation of a patient technology technician role to manage bedside nursing technology. Nurs Manage 2025;56(3):11-17. [doi: 10.1097/nmg.0000000000000226] [Medline: 39928026]
7. Pilosof NP, Barrett M, Oborn E, Barkai G, Zimlichman E, Segal G. Designing for flexibility in hybrid care services: lessons learned from a pilot in an internal medicine unit. Front Med Technol 2023;5:1223002. [doi: 10.3389/fmedt.2023.1223002] [Medline: 38053662]
8. Choi H, Tak SH, Song YA, Park J. Nurses' perspectives on the adoption of new smart technologies for patient care: focus group interviews. BMC Health Serv Res 2025 Mar 18;25(1):391. [doi: 10.1186/s12913-025-12578-z] [Medline: 40098040]

Edited by KA Clegg; submitted 05.Jan.2026; this is a non–peer-reviewed article;accepted 05.Jan.2026; published 26.Jan.2026.

Please cite as:
Congdon J
A Frontline Worker's Take on Hybrid Care Implementation in the Hospital Setting
J Med Internet Res 2026;28:e90879
URL: https://www.jmir.org/2026/1/e90879
doi:10.2196/90879

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Jan.2026.

XSL•FO
RenderX

WHOOP, There It Is: Lessons From WHOOP's FDA Warning Letter

Blythe Karow

(J Med Internet Res 2026;28:e90882) doi:10.2196/90882

KEYWORDS

wearable electronic devices; United States Food and Drug Administration; device approval; health devices; digital health; WHOOP; Apple; Aktiia; medical device regulation; blood pressure monitoring, ambulatory; legislation; jurisprudence

Textbox .

Key Takeaways

- The US Food and Drug Administration's warning letter to WHOOP and ensuing battle have highlighted a critical moment in the wearables space as the US regulatory framework has been evolving in real time.
- New lines have been drawn. The 2026 Food and Drug Administration wellness guidance has established new norms around the ability to noninvasively detect and relay information to users.
- Ultimately, however, the wellness exemption remains poorly defined for companies seeking clear guidance on user interfaces and how they can interpret and relay data to users while remaining a wellness device.

Blythe Karow [1] *is a strategic management consultant and fractional executive with a biomedical engineering bachelor's degree from Johns Hopkins University, an MBA from the UVA Darden School of Business, and over twenty years of experience leading, advising, and consulting as a medtech, digital health, wearables, and regulated healthtech expert. She is the founder of The Karow Advisory Group and author of The Device Files* [2], *a Substack where she shares strategy, insights, and practical tips for getting devices to market. She has covered the evolution of the WHOOP/US Food and Drug Administration (FDA) regulatory conflict since it began and shares her reflections here.*

In July 2025, the fitness wearable company WHOOP received a warning letter from the FDA [3] for their "Blood Pressure Insights" (BPI) feature, noting that it was clearly a medical device requiring proper 510K clearance. For anyone following the wearables space, this wasn't surprising. WHOOP, like many wearables companies, has slowly been evolving from performance optimization and wellness toward monitoring and diagnosing medical conditions [4].

The warning did not just kick off another regulatory spat. It raised important questions and considerations for the future of wearables and other medical devices, spurring debates across the medical device and wearables communities for 6 months, and culminating in the release of new FDA guidance on January 6, 2026 [5]. In this analysis, I review both the FDA's and WHOOP's original arguments around the regulatory status of the BPI feature, looking at the evolution of the conflict up until now.

Where the July 2025 FDA Warning Letter Arguments Fell Short

The "Inherent Association" Doctrine

The FDA's original position hinged on "inherent association"—that blood pressure measurement is "inherently associated with the diagnosis of hypo- and hypertension" regardless of disclaimers or intended use [3].

In my view, this legal theory was always a bit of a stretch. If any biometric that *can* be used for diagnosis automatically *becomes* a medical device, then the wellness device exemption Congress created in the 21st Century Cures Act [6,7]—that software intended for general wellness and not medical diagnosis or treatment is not considered a medical device—becomes meaningless. By this logic, bathroom scales would be medical devices because weight is "inherently associated" with obesity diagnosis, but you can walk into any store and buy a scale without FDA oversight.

In my original analysis, I argued that the real conversation shouldn't be whether all blood pressure measurements are inherently diagnostic, but rather how the FDA's wellness guidelines could and should be revised to allow biomarker collection while providing clear guidance on handling detected extremes.

The Language Trap

The FDA found specific language from WHOOP that crossed into diagnostic territory, such as "higher blood pressure may be an indicator of poor sleep" and materials indicating that the feature was intended to identify "higher blood pressure" [3].

The distinction between "higher" and "high" blood pressure matters here. "Higher" describes a relative change from an

XSL•FO
RenderX

individual's baseline and should be acceptable for wellness devices. However, "high blood pressure" suggests comparison with clinical diagnostic thresholds. While WHOOP appears to have used "higher," the FDA appeared to have interpreted this as "high blood pressure."

The Color-Coding Absurdity

The FDA objected to WHOOP's user insights interface, claiming the color-coded (green/yellow/orange) readings imply medical interpretation, but most fitness wearables use similar color-coding schemes: for example, Garmin shows heart rate zones in green, yellow, and red, and Apple Watch displays activity rings with traffic-light logic.

If color-coding implies medical interpretation, are we going to start requiring submissions from manufacturers of street lights? This may be a tongue-in-cheek comparison, but the question remains: when does color-coding imply diagnostic categorization?

The Misuse Argument

The FDA cited "evidence of individuals using BPI to monitor their hypertension" as a sign that WHOOP had overstepped the line. This original argument concerned me for two reasons.

First, using anecdotal misuse by some users to justify broad regulatory classification, especially when they provided no details about scale, frequency, or how this evidence was collected, would have set a concerning precedent. Second, people use fitness trackers for wellness purposes all the time, and those wellness issues often connect to more serious health conditions. Sleep quality can impact anxiety or cognitive health [8] and getting your 10,000 steps can help with overall cardiovascular wellness [9]. That shouldn't make wellness device manufacturers liable for medical device indications.

While we want to be careful of products purposefully trying to skirt a medical device pathway, some off-label use is probably acceptable when it leads people to healthier lifestyles that could improve more serious conditions. And what's the harm in someone with hypertension wanting to get an idea of their blood pressure, especially when the next step is probably to seek proper medical testing if they see warning signs?

Where WHOOP Might Have Been Wrong

WHOOP might have had legitimate grievances about the FDA's approach to their BPI feature, but they also seemed to escalate the situation by (1) not adjusting their marketing or features claims early on when discussions started in May 2025, (2) continuing to leave up the "offensive" product features while arguing with the FDA, (3) making a buzzworthy end run to RFK Jr instead of working with the agency [10], and (4) having their CEO, Will Ahmed, making public statements arguing their "legal" case.

The Fatal Phrase

Buried in the FDA warning is language that likely sealed WHOOP's fate: their website originally described the blood pressure feature as delivering "medical-grade health & performance insights" [4,11].* That single term was regulatory poison. In the FDA's world, "medical-grade insights" implies clinical accuracy, validation standards, and regulatory approval, whether intended or not.

My original argument: you can have a wellness device that collects medical-grade *data*—in fact, this is preferred, because algorithms perform best when data are high-quality [12-14]. The problem was that WHOOP had originally used the term in reference to providing medical-grade *insights*, implying diagnostic-level blood pressure measurement.

If WHOOP had proposed changing this language in May, the situation may have been resolved sooner. However, the back-and-forth, while refusing adjustments, seemed to have gotten the FDA's ire up. Additionally, with Aktiia and Apple subsequently receiving clearance for blood pressure and hypertension claims [15,16], it likely became harder for the FDA to give leniency.

The Media Defense

Shortly after the FDA warning letter, WHOOP CEO Will Ahmed posted counterarguments on LinkedIn [17], making good points but ultimately overrelying on disclaimers to protect the company from oversight. In medtech, just telling consumers not to use a device a certain way often isn't enough, especially if it contradicts other marketing claims.

There were also weaknesses in the arguments Ahmed made during his CNBC appearance at the end of August 2025 [18]:

1. He appeared to believe that wellness exemptions override the core intent of the FDA's mission to protect patients and provide medical device oversight, stating that "Intended use is what matters—and the law agrees." But once you make claims that your device does something a medical device does, you can't just say, "Yeah, but we're just here for wellness purposes," because guess what? So are all medical devices!
2. He mentioned that their blood pressure feature requires calibration with a blood pressure cuff (a medical device), as though acknowledging reliance on a medical device made it clearer that WHOOP itself isn't performing a medical function, but if you need a medical device to calibrate your "wellness" feature, that suggests you're doing medical device work.
3. He claimed that WHOOP should set a new precedent because regulations could thwart tech innovation, blind to an entire medical device industry that has been innovating for decades while following the rules.
4. He claimed that WHOOP members "overwhelmingly support us fighting for [BPI]." This attempt to position himself as defending consumer access read as a deflection rather than acknowledging that the company had created a regulatory mess.

The Marketing Smoking Gun

Most damning to me when reviewing this debate is how WHOOP still structures subscription tiers, separating blood pressure insights from general "health monitoring" and bundling it specifically with FDA-cleared medical features in the highest tier [19]. They're explicitly categorizing blood pressure insights

XSL•FO
RenderX

alongside actual medical device features, telling consumers through product positioning that blood pressure monitoring belongs in the medical device category. It feels a bit like a magician's misdirection—it implies medical device credibility to the consumer, while denying the need for medical device status to the FDA.

New Year, New Guidance

Ahmed's claim that the FDA was "on the wrong side of the law" [18] was somewhat remarkable given the FDA's enforcement position. WHOOP appealed the formal warning letter but continued to market much the same way throughout 2025. This could have brought additional serious legal trouble [20]; the FDA had multiple escalation options, including court orders to force WHOOP to disable blood pressure features while pursuing appeals.

I was fairly confident that WHOOP's attempted lobbying to RFK Jr wouldn't be effective in swaying the FDA, given the precedent set by the Aktiia and Apple clearances and additional pressure from public scrutiny. But I was wrong.

In the first week of 2026, the FDA responded with new guidance that very clearly states that they will allow blood pressure measurements via optical sensing to remain in the wellness category as long as companies make no claims to having "medical-grade" data or insights [5]. The new guidance seems an obvious win for WHOOP, and it also seems clear from current FDA Commissioner Dr Marty Makary's comments during his January 6 interview on Fox Business News [21] that the FDA is moving to relax and clarify control and oversight in the wellness space. He stated, "We want to let companies know, with very clear guidance, that if their device or software is simply providing information, they can do that without FDA regulation."

While the new guidance provides clear direction on whether WHOOP can market BPI, there remains a lack of clarity in a few areas. The guidance does not specifically address all of the concerns originally outlined in the warning letter (for example, it does not address whether you can use color coding in a user interface) and WHOOP does not seem to have adjusted their interface or marketing beyond specifically removing claims of "medical" or "clinical grade" in reference to BPI [4,11]. And, although multiple examples are provided on how to describe a product or its features without crossing into medical device territory, there is very little to guide developers on appropriate user interface and alerts, leaving other technologies with a great deal of gray area to navigate in the future.

Ultimately, this case has shone a spotlight on some of the weaker elements of the wellness carve out and will likely continue to play a major role in how the FDA decides to evolve in this new world of wearables, artificial intelligence, and the quantified self. We're watching the real-time evolution of the wellness doctrine; while I expect some continued upheaval, a line appears to be forming around collecting versus interpreting user data as the new distinction between wellness and medical devices. Innovators would do well to keep a close eye on the wellness space as it continues to evolve.

**Note: Following the January 2026 FDA guidance, the WHOOP web pages were revised, and the original marketing language is no longer retrievable in the current versions.*

Conflicts of Interest

None declared.

References

1. Blythe Karow. LinkedIn. URL: https://www.linkedin.com/in/blythe-karow [accessed 2026-01-09]
2. The Device Files: From Concept to Commercialization | Substack. URL: https://blythekarow.substack.com [accessed 2026-01-09]
3. Warning letter: WHOOP, Inc. — MARCS-CMS 709755. US Food and Drug Administration. 2025 Jul 14. URL: https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/warning-letters/whoop-inc-709755-07142025 [accessed 2026-01-09]
4. Introducing WHOOP 5.0 and WHOOP MG: the all-new ways to unlock better health, fitness, and longevity. WHOOP. 2024 Aug 27. URL: https://www.whoop.com/ca/en/thelocker/introducing-whoop-5-0-and-whoop-mg/?srsltid=AfmBOooko9VhmvxAMxBEVp7yawtVNVXZuwrNpCtWHHmRifUzfgK-i7ST [accessed 2026-01-15]
5. Guidance Document. General wellness: policy for low risk devices. Guidance for industry and Food and Drug Administration staff. Docket number: FDA-2014-N-1039. US Food and Drug Administration. 2026 Jan 6. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/general-wellness-policy-low-risk-devices [accessed 2026-01-12]
6. 114th Congress, United States. 21st Century Cures Act, Public Law 114-255 114th Congress 130 Stat 1033. 2016 Dec 13 URL: https://www.congress.gov/114/plaws/publ255/PLAW-114publ255.pdf [accessed 2026-01-09]
7. Guidance Document. General wellness: policy for low risk devices. Guidance for industry and Food and Drug Administration staff. Docket number: FDA-2014-N-1039. US Food and Drug Administration. 2019 Sep. URL: https://web.archive.org/web/20251217125918/https://www.fda.gov/regulatory-information/search-fda-guidance-documents/general-wellness-policy-low-risk-devices [accessed 2025-12-24]
8. Cox RC, Olatunji BO. Sleep in the anxiety-related disorders: a meta-analysis of subjective and objective research. Sleep Med Rev 2020 Jun;51:101282. [doi: 10.1016/j.smrv.2020.101282] [Medline: 32109832]

XSL•FO
RenderX

9. Lavie CJ, German CA, Sanchis-Gomar F. Reducing mortality and cardiovascular disease. J Am Coll Cardiol 2023 Oct;82(15):1495-1498. [doi: 10.1016/j.jacc.2023.08.007] [Medline: 37676197]
10. Will Ahmed | Instagram. URL: https://www.instagram.com/p/DKASUbOuia8 [accessed 2026-01-09]
11. WHOOP delivers innovative blood pressure insights for a deeper look at your well-being. WHOOP. 2025 May 8. URL: https://www.whoop.com/ca/en/thelocker/blood-pressure-insights/?srsltid=AfmBOoqbWjK5nKtTNCN1yfpSmglPk6fNdVm32i0wK0fDyRRRTELRbeFC [accessed 2026-01-12]
12. Schwabe D, Becker K, Seyferth M, Klaß A, Schaeffter T. The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. NPJ Digit Med 2024 Aug 3;7(1):203. [doi: 10.1038/s41746-024-01196-4] [Medline: 39097662]
13. Cho S, Ensari I, Weng C, Kahn MG, Natarajan K. Factors affecting the quality of person-generated wearable device data and associated challenges: rapid systematic review. JMIR Mhealth Uhealth 2021 Mar 19;9(3):e20738. [doi: 10.2196/20738] [Medline: 33739294]
14. Hearn J, Van den Eynde J, Chinni B, et al. Data quality degradation on prediction models generated from continuous activity and heart rate monitoring: exploratory analysis using simulation. JMIR Cardio 2023 May 3;7(1):e40524. [doi: 10.2196/40524] [Medline: 37133921]
15. 510(k) Premarket Notification K250415: Aktiia G0 Blood Pressure Monitoring System. US Food and Drug Administration. 2025 May 15. URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K250415 [accessed 2026-01-09]
16. 510(k) Premarket Notification K250507: Hypertension Notification Feature (HTNF). US Food and Drug Administration. 2025 Sep 11. URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K250507 [accessed 2026-01-09]
17. Ahmed W. The FDA just sent WHOOP a warning letter about our Blood Pressure Insights feature. LinkedIn. 2025 Jul 16. URL: https://www.linkedin.com/posts/willahmed_wellness-innovation-regulation-activity-7351264571933745156-IJw4 [accessed 2026-01-09]
18. Ahmed W. Whoop CEO Will Ahmed on FDA warning: our blood pressure insight is intended for wellness use only. CNBC Television. 2025 Aug 14. URL: https://www.cnbc.com/video/2025/08/14/whoop-ceo-will-ahmed-on-fda-warning-our-blood-pressure-insight-is-intended-for-wellness-use-only.html [accessed 2026-01-09]
19. Memberships made for you. WHOOP. URL: https://www.whoop.com/ca/en/membership/ [accessed 2026-01-12]
20. Rowe v Whoop, Inc, no3:25cv09910 (ND Cal Nov 18, 2025). ClassAction.org. URL: https://www.classaction.org/media/rowe-v-whoop-inc-complaint.pdf [accessed 2026-01-09]
21. FDA sides with innovation over regulation in MAJOR AI healthcare shift. Fox Business News YouTube page. 2026 Jan 6. URL: https://www.youtube.com/watch?v=OxR6bHDOUvY [accessed 2026-01-12]

Abbreviations

BPI: Blood Pressure Insights
FDA: US Food and Drug Administration

Edited by KA Clegg; submitted 05.Jan.2026; this is a non–peer-reviewed article;accepted 05.Jan.2026; published 21.Jan.2026.

Please cite as:
Karow B
WHOOP, There It Is: Lessons From WHOOP's FDA Warning Letter
J Med Internet Res 2026;28:e90882
URL: https://www.jmir.org/2026/1/e90882
doi:10.2196/90882

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.Jan.2026.

XSL•FO
RenderX

UnitedXR Europe 2025: Aligning Health Care Extended Reality

Jose Ferrer Costa

(J Med Internet Res 2026;28:e90727) doi:10.2196/90727

KEYWORDS

virtual reality; augmented reality; diffusion of innovation; health policy; biomedical technology; interprofessional relations; UnitedXR Europe; Stereopsia; Augmented World Expo

Textbox .

Key Takeaways

- UnitedXR Europe 2025 highlighted that health care extended reality (XR) is no longer constrained by technical capability but by alignment between industry, academic evidence, and clinical governance.
- Persistent gaps remain between how success, risk, and readiness are defined by developers, researchers, and health care professionals.
- Formats that enable direct cross-stakeholder dialogue may be as critical as technological advances for translating XR potential into routine clinical practice.

Extended reality (XR) is not new to health care, but over the past decade, its use has become more widespread in medical training, rehabilitation, pain management, and mental health care, supported by a growing body of evidence [1-4]. Despite clear gains in technical maturity, the adoption of XR in health care remains uneven. Immersive tools are increasingly capable, yet their integration into routine clinical practice appears to depend less on technical performance than on organizational readiness, governance, and professional acceptance—a pattern well described in implementation research [5,6].

UnitedXR Europe 2025 offered a concentrated view of this tension and an effort to resolve it. The event marked the first joint edition of two historically distinct strands of the European XR ecosystem. Stereopsia—anchored in Brussels since 2009—has long served as a meeting point for immersive research, cultural production, and policy dialogue. In parallel, the Augmented World Expo evolved into a global, industry-led platform focused on enterprise deployment and market scale.

Their integration under UnitedXR Europe has brought these traditions into direct contact, highlighting gaps and areas of convergence, as well as creating a shared space for dialogue between industry, academia, health care, and policy actors.

What the Program Revealed About Health Care XR Readiness

UnitedXR Europe 2025 brought together more than 125 exhibitors across 14 parallel tracks, including a dedicated Healthcare, Pharma, & Wellbeing track. Across the event's agenda, health care XR appeared to be entering an infrastructure phase, positioned not as a peripheral demonstration but as a maturing vertical confronted with questions of scale, integration, and long-term sustainability.

Beyond formal sessions, interaction extended into curated environments, such as the European Market for Immersive Creativity and business-to-business matchmaking between developers, researchers, and institutional actors. Roundtables, workshops, start-up pitch competitions, and the European XR Awards Gala reinforced the event's dual role as a space for exchange and an indicator of technical maturity.

This shift was visible on the expo floor, where novelty gave way to utility. Across the XR ecosystem, major headset vendors relied on real-world implementation partners to demonstrate use cases rather than on product-centric stands, with a dedicated demonstration pavilion hosted by the International Virtual Reality Healthcare Association. Within this context, hands-on demonstrations, such as those by VirtualiSurg [7], illustrated how VR is being embedded into established training pathways, particularly when combined with modular haptic systems, such as those provided by SenseGlove [8].

The scientific program reflected a similar maturation. As noted by Oliver Schreer, PhD, track chair, submissions increased from barely a dozen in previous editions to over 45 this year, accompanied by a clear shift toward more application-oriented research.

Moving between tracks revealed a persistent asymmetry. Enterprise discussions were supported by a settled language of scaling, procurement, and return on investment. Health care discussions, by contrast, were anchored in regulation, safety, and professional accountability. Side by side, these perspectives exposed a central tension: while the XR industry largely seems prepared to scale, health care systems are still negotiating why, and under what conditions, that scale should occur.

XR as a Coordination Challenge

What emerged from these observations was a coordination problem. The remaining barriers to health care XR adoption lie less in hardware or software performance than in the absence of shared decision-making structures.

This gap surfaced repeatedly in discussions about adoption criteria. In one interactive panel within the health care track,

XSL•FO
RenderX

participants were asked to rank factors that guided XR implementation. Priorities clearly diverged, with safety treated as a nonnegotiable baseline by some and as one consideration among many by others. The exchange was brief but revealing, with high audience participation.

Figure WL1. Audience engagement during an interactive panel session. Faces have been blurred to protect participant privacy. Photograph credit: Sonya Seddarasan, track chair for Healthcare, Pharma, & Wellbeing at UnitedXR Europe 2025.

What this exchange made visible was that XR initiatives often stall not because immersive systems fail to perform but because organizations lack an agreed framework for guiding early-stage decisions about readiness for clinical use. A recurring takeaway from the session was the growing consensus that health care XR must move beyond repeated proofs of concept. Without a shared way to assess safety, credibility, and contextual fit before implementation begins, pilots tend to accumulate without a clear pathway toward sustained, real-world integration.

Format Matters for Cross-Stakeholder Dialogue

Points of agreement and divergence became most visible during interactive panel sessions. For example, open discussion forced participants to make their criteria explicit and revealed how differently evidence, risk, and readiness were understood across institutional and professional contexts.

The organizers appeared attentive to this dynamic. Sonya Haskins—Augmented World Expo's head of programming—noted that future editions may place greater emphasis on roundtable formats intended to support cross-disciplinary negotiation. Alexandra Gérard—codirector of UnitedXR Europe—similarly pointed out the need to broaden participation within the health care track, including participation by patients and frontline practitioners with direct insight into care delivery.

For health care XR, where adoption depends on trust and legitimacy as much as it does on performance metrics, creating conditions for this kind of dialogue may be as consequential as any technical advance.

Immersion as a Clinical Responsibility

Health care discussions at UnitedXR Europe reflected a growing recognition that immersive technologies carry a different kind of responsibility than that of most digital health tools. XR was framed not merely as a delivery medium but as a technology that directly shapes perception, attention, and embodied experience—a distinction that is well established in prior experimental and neuroscientific work [3,9,10].

This concern surfaced most clearly in informal exchanges. Sonya Haskins described XR creation as "hacking the brain," using the phrase as a metaphor to underscore why immersive systems

XSL•FO
RenderX

cannot be treated as neutral software development. When technologies act directly on perception and experience, questions of intent, safety, and oversight move rapidly to the foreground.

Policy frameworks are beginning to respond, albeit unevenly. The event coincided with the launch of the European Partnership for Virtual Worlds, which explicitly identifies health care as a strategic domain [11]. This marks progress when compared with earlier global digital health strategies, including those of the World Health Organization, where immersive technologies remain largely unnamed despite their growing use in practice [12,13].

Terminology remains a point of friction. As several clinicians noted, framing clinical XR under the umbrella of "virtual worlds" sits uneasily with health care practice. Recent European policy work differentiates professional and health care uses of immersive technologies from consumer-oriented virtual worlds, emphasizing that clinical applications are bounded; task specific; and subject to heightened safety, ethical, and governance requirements [4]. In this sense, UnitedXR functioned as a translation layer, highlighting where policy language must be refined to align with the risk-aware logic of patient care.

Implications

UnitedXR Europe 2025 made clear that health care XR has moved beyond the phase of technical proof. What remains unresolved is not what the technology can do but how it is integrated into the social, ethical, and organizational fabric of health care. From my perspective as a clinician-researcher involved in real-world XR deployment, these tensions are familiar from practice and, importantly, addressable.

The challenge ahead is one of alignment; synchronizing industrial velocity with clinical deliberation; and narrowing the vocabulary gap between policymakers, developers, and practitioners.

As this event demonstrated, that work rarely happens through polished presentations alone. It happens in shared spaces where assumptions are tested, priorities collide, and the real conditions for responsible adoption begin to take shape.

Conflicts of Interest

None declared.

References

1. Zuo G, Wang R, Wan C, Zhang Z, Zhang S, Yang W. Unveiling the evolution of virtual reality in medicine: a bibliometric analysis of research hotspots and trends over the past 12 years. Healthcare (Basel) 2024 Jun 26;12(13):1266. [doi: 10.3390/healthcare12131266] [Medline: 38998801]
2. Lie SS, Helle N, Sletteland NV, Vikman MD, Bonsaksen T. Implementation of virtual reality in health professions education: scoping review. JMIR Med Educ 2023 Jan 24;9:e41589. [doi: 10.2196/41589] [Medline: 36692934]
3. Grassini S, Laumann K. Immersive visual technologies and human health. Presented at: ECCE 2021: European Conference on Cognitive Ergonomics 2021; Apr 26-29, 2021; Siena, Italy p. 1-6. [doi: 10.1145/3452853.3452856]
4. Directorate-General for Communications Networks, Content and Technology (European Commission), Gabaliņa R, Migals A, Zepcan A. Virtual Worlds: How Do They Affect Our Health and Well-Being: Publications Office of the European Union; 2025. [doi: 10.2759/4342263]
5. Kouijzer MMTE, Kip H, Bouman YHA, Kelders SM. Implementation of virtual reality in healthcare: a scoping review on the implementation process of virtual reality in various healthcare settings. Implement Sci Commun 2023 Jun 16;4(1):67. [doi: 10.1186/s43058-023-00442-2] [Medline: 37328858]
6. Shiner CT, Croker G, McGhee J, Faux SG. Perspectives on the use of virtual reality within a public hospital setting: surveying knowledge, attitudes, and perceived utility among health care professionals. BMC Digit Health 2024 Apr 25;2:18. [doi: 10.1186/s44247-024-00076-x]
7. VirtualiSurg. URL: https://virtualisurg.com/ [accessed 2026-01-13]
8. SenseGlove. URL: https://www.senseglove.com/ [accessed 2026-01-13]
9. Riva G, Wiederhold BK, Mantovani F. Neuroscience of virtual reality: from virtual exposure to embodied medicine. Cyberpsychol Behav Soc Netw 2019 Jan;22(1):82-96. [doi: 10.1089/cyber.2017.29099.gri] [Medline: 30183347]
10. Slater M. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. Philos Trans R Soc Lond B Biol Sci 2009 Dec 12;364(1535):3549-3557. [doi: 10.1098/rstb.2009.0138] [Medline: 19884149]
11. European Commission launches European Partnership for Virtual Worlds. European Commission. 2025 Dec 11. URL: https://digital-strategy.ec.europa.eu/en/news/european-commission-launches-european-partnership-virtual-worlds [accessed 2025-12-18]
12. WHO guideline: recommendations on digital interventions for health system strengthening. World Health Organization. 2019. URL: https://apps.who.int/iris/handle/10665/311941 [accessed 2025-12-18]
13. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2025-12-18]

XSL•FO
RenderX

Edited by KA Clegg; submitted 02.Jan.2026; this is a non–peer-reviewed article;accepted 02.Jan.2026; published 21.Jan.2026.

Please cite as:
Costa JF
UnitedXR Europe 2025: Aligning Health Care Extended Reality
J Med Internet Res 2026;28:e90727
URL: https://www.jmir.org/2026/1/e90727
doi:10.2196/90727

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.Jan.2026.

XSL•FO
RenderX

How Does That Large Language Model Make You Feel?

Simon Spichak

Abstract

People are increasingly turning to commercially available large language models (LLMs) for emotional support. In this *News and Perspectives* article, JMIR Correspondent Simon Spichak reports on the role of LLMs in mental health, speaking with experts about safety concerns, research gaps, and next steps.

(J Med Internet Res 2026;28:e105105) doi:10.2196/105105

KEYWORDS

large language models; artificial intelligence; chatbots; mental health; psychotherapy; patient safety; human-computer interaction; simulation; ethics; research design

Textbox 1.

Key Takeaways:

- Although there's a lack of data on safety, long-term outcomes, and effectiveness, many people are using large language models (LLMs) for mental health support.
- Some companies are developing clinically validated chatbots for mental health, but these remain untested against general-purpose LLMs.
- Experts urge for more research and understanding from clinicians who are encountering patients that use LLMs for support.

Simon Spichak is the co-founder of Resolvve, a low-cost virtual therapy clinic in Ontario specializing in attention-deficit/hyperactivity disorder (ADHD) and obsessive-compulsive disorder (OCD). He has no commercial interest or association with any products mentioned in this article.

In the mid-1960s, MIT computer scientist Joseph Weizenbaum developed an early chatbot called ELIZA.

It used simple techniques to reflect a user's language, and despite no actual understanding of language, human psychology, or emotions, people interacting with the chatbot were quickly drawn in. Users attributed empathy and understanding to the chatbot, and Weizenbaum's secretary even asked him to leave the room before speaking with the chatbot privately. Later in his life, Weizenbaum warned about the dangers of such chatbots, cautioning that we should never confuse them with humans.

Six decades later, OpenAI launched its flagship large language model (LLM), ChatGPT, prompting other companies to follow suit. Since then, a handful of companies have released several versions of chatbots. Many people are already using LLMs for mental health support. In an October 2025 survey of more than 1800 US adults, 35% reported using AI tools more than once a week for this purpose, including more than 28% who were already seeing a human provider. Researchers are trying to catch up, leaving lots of questions in the air about the effectiveness and safety of these LLMs for mental health.

"We are in the midst of a major mental health crisis that requires systemic solutions, not just technological stopgaps," Arthur Evens Jr, PhD, chief executive officer of the American Psychological Association, said in their new health advisory. "While chatbots seem readily available to offer users support and validation, the ability of these tools to safely guide someone experiencing crisis is limited and unpredictable."

Safety and Effectiveness of General-Purpose LLMs for Therapy

Although LLMs are ubiquitous, it isn't clear how well or for whom they work best, says Shannon Wiltsey Stirman, PhD, a psychologist and professor at Stanford University who serves as the co-director of the Center for Responsible and Effective AI Technology Enhancement for PTSD Treatment (CREATE). A scoping review she coauthored found that most of the studies putting AI models to the test involved few participants and did not use validated measures to assess outcomes.

"There have been a couple of larger and better design[ed] studies, but not nearly enough for us to think they're ready to sort of roll out direct-to-consumer," she says, urging that more testing is needed to know who these models are safe and effective for, and to compare them to the existing standard of treatment.

Though researchers have conducted simulation studies to understand whether chatbots conduct sufficient risk assessment and respond appropriately, these might not reflect real-world interactions.

The guardrails "degrade over time," says Wiltsey Stirman. "If somebody has been interacting with a chatbot for months and months, the way that the chatbot responds to certain disclosures might look very different than it does in a simulation."

XSL•FO
RenderX

General-purpose LLMs "can very often produce direct harm," says Nicholas Jacobson, PhD, associate professor at Dartmouth College. For OCD, where clients have persistent obsessions, he says the nature of LLMs may reinforce detrimental reassurance-seeking behaviors. These models, he said, are also quite poor at helping manage difficulties with emotional regulation, which are common across mental health disorders.

In a rare few cases, LLMs have been linked with suicide and so-coined "AI psychosis." Though as psychiatrist John Torous, MD, director of the Digital Psychiatry Division at Beth Israel Deaconess Hospital and editor in chief of *JMIR Mental Health*, notes, it doesn't mean that the AI is leading to new-onset delusions. He said there's no definitive evidence yet that interacting with LLMs could lead to new-onset psychosis.

Comparing General-Purpose LLMs to Therapy Chatbots

Jacobson started training proprietary models specific to the mental health context in 2019, developing Therabot as a clinically validated chatbot for therapy. In a randomized controlled trial published in 2025, it was effective in treating depression compared to a waitlist control. Emerging companies like Sonia are building AI chatbots as a "solution for emotional support."

But some experts question whether therapy-specific AI models outperform ChatGPT, Claude, and other LLMs. Torous sees the newer models getting bigger and safer fast, which will make it necessary for companies that offer mental health–specific AIs to test them against publicly available LLMs "to do a better job of justifying themselves."

One of the few published studies pitting these models against each other found no difference between ChatGPT and a specialized LLM built on OpenAI's models. Still, more data is needed for head-to-head comparisons, says Wiltsey Stirman, since the study didn't use "a specially designed foundation model."

Jacobson told JMIR Publications that he and his colleagues have collected internal data showing Therabot is superior but will not test it against other available models due to ethical concerns. The rollout of these models "can be profoundly unsafe and not something that we'll want to facilitate," he says.

Speaking to Patients Using LLMs

Clinicians are likely meeting with clients who are already using LLMs to support their mental health.

Wiltsey Stirman says it's important to keep an open dialogue and understand how clients are using AI. Understanding how these tools work could even help practitioners improve their clients' AI use. "If people are using them for some coaching and accountability and support around a specific goal, the therapist could work with them on how to prompt the chatbot," she says.

Figure WL1.

But Jacobson raised concerns that the advice provided by these chatbots, which don't have specific mental health training, may sometimes be at odds with that of the health care professional they're working with.

XSL•FO
RenderX

The Future of LLMs in Mental Health Care

Other applications are rapidly being developed, including clinician-facing chatbots to aid in diagnosis and assessments, though experts say they're still in the early days of development.

Some researchers are testing whether LLMs could transform clinical notes into handouts for patients, but one study found the process introduces errors. Meanwhile, Utah's drug regulator green-lit a pilot of Legion Health AI, an LLM-powered tool to issue refills for psychiatric prescriptions.

Across the board, experts who spoke with JMIR Publications say it's important to invest in and run high-quality studies examining the impact of LLM use on mental health, but the rapid deployment of direct-to-consumer AI is outpacing the research. These models also carry significant privacy concerns. A 2026 preprint found that 17 of 20 LLMs share raw conversation data with third parties.

A lot of the best data, says Torous, is coming directly from AI companies that publish white papers and preprints instead of going through peer review. He points to a recent blog post from Anthropic as the best data on sycophancy. "Yes, they could be lying, but what else can you do? [You] kind of have to take their word for it," he says.

Edited by KA Clegg; submitted 19.Jun.2026; this is a non–peer-reviewed article;accepted 19.Jun.2026; published 30.Jun.2026.

Please cite as:
Spichak S
How Does That Large Language Model Make You Feel?
J Med Internet Res 2026;28:e105105
URL: https://www.jmir.org/2026/1/e105105
doi:10.2196/105105

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jun.2026.

XSL•FO
RenderX

Building a Malaria Intelligence System for Real-Time Prediction and Data-Driven Intervention Planning

Sharon Muzaki

Abstract

Sub-Saharan Africa continues to face the highest malaria burden worldwide, accounting for 95% of malaria cases and deaths according to the World Health Organization. In this *News and Perspectives* article, JMIR Correspondent Sharon Muzaki reports on a new artificial intelligence system in sub-Saharan Africa that could help improve public health responses to the disease.

(J Med Internet Res 2026;28:e105472) doi:10.2196/105472

KEYWORDS

malaria; artificial intelligence; machine learning; public health; disease surveillance; Nigeria; Democratic Republic of the Congo; epidemiology

Textbox 1.

Key Takeaways:

- AI-driven malaria intelligence systems could transform disease surveillance and intervention planning in high-burden settings by integrating epidemiological, climatic, and environmental data into predictive public health models.
- Their success in supporting long-term malaria control will depend on continuous monitoring frameworks, localized adaptation, and integration into existing health care systems.

Malaria remains one of the most persistent public health challenges in sub-Saharan Africa, particularly among children under 5 and pregnant women. Despite decades of interventions involving insecticide-treated nets, antimalarial drugs, indoor residual spraying, vaccination campaigns, and donor-funded awareness programs, transmission remains high in many endemic regions.

Nigeria bears the world's highest malaria burden, while the Democratic Republic of Congo (DRC) ranks among the highest-burden countries globally. Together, the two countries account for a substantial proportion of malaria cases and deaths worldwide. In many high-burden settings, however, malaria surveillance still relies heavily on routine reporting systems and periodic surveys, limiting the ability to detect changing transmission patterns and evaluate interventions in near real time.

At the Femi Samson Institute of AI & Biomedical Research (FSI-AIBR) in Lagos, Nigeria, Daniel Onimisi, MSc, the institute's founder and executive director, is developing an AI-driven malaria intelligence platform designed to shift malaria control from reactive intervention toward predictive and precision public health.

The platform combines epidemiological data, environmental monitoring, climate indicators, and machine learning models into a continuous intelligence framework aimed at improving outbreak prediction, intervention targeting, and resource allocation.

From Static Surveillance to Continuous Intelligence

Traditional malaria surveillance systems, which typically rely on health-facility reports and routine case data collected after infections have occurred, are largely designed to describe what has already happened. By contrast, the new tool integrates multiple datasets to identify risk patterns and provide early warning signals that can inform preventive action. While traditional surveillance systems remain useful for broad planning, they often struggle to anticipate changing transmission patterns or evaluate the long-term effects of interventions.

Onimisi believes this represents a major gap in current malaria programs. "What happens after the intervention ends? How do we measure whether the resources deployed actually changed malaria outcomes over time?" Onimisi asks.

He says many interventions focus heavily on distributing mosquito nets and medicines, as well as conducting awareness campaigns, but lack mechanisms for continuous assessment once funding cycles end.

The malaria intelligence platform was designed to address this challenge by functioning as a decision-support system capable of supporting prediction, preparedness, monitoring, and postintervention assessment.

To achieve this, the system integrates multiple datasets, including Demographic and Health Survey (DHS) data, Malaria Indicator Survey (MIS) datasets, rainfall information,

XSL•FO
RenderX

satellite-derived vegetation monitoring, historical malaria incidence records, and geospatial analyses.

Using geospatial segmentation, the platform divides Nigeria into state boundaries, local government areas, and smaller clusters linked to environmental and epidemiological variables. Onimisi says the system currently analyses approximately 435,000 localized clusters across the country.

Integrating Climate and Environmental Drivers

Malaria transmission is closely linked to environmental and climatic conditions.

Rainfall accumulation, vegetation density, stagnant water, flooding patterns, and changing climate conditions all influence mosquito breeding and transmission dynamics. Urbanization, deforestation, and population movement further complicate these patterns. Climate variability may also alter malaria distribution across Africa in ways that conventional surveillance systems struggle to detect.

To capture these relationships, the platform combines satellite-derived vegetation indices with rainfall and climate datasets spanning more than three decades. This allows the system to identify ecological conditions that may favor malaria transmission before outbreaks become visible through routine health reporting.

Machine Learning and Predictive Modeling

Machine learning forms the analytical core of the platform.

Historical malaria incidence data are combined with climatic and environmental variables to estimate malaria risk probabilities across geographic regions. This approach builds on earlier work by Onimisi and colleagues, currently available as a preprint, integrating DHS and MIS biomarkers with 34 years of CHIRPS-NDVI (Climate Hazards Center Infrared Precipitation with Stations - Normalized Difference Vegetation Index) climate data and machine learning techniques to predict malaria risk and generate spatial maps of transmission patterns. These predictions are then compared with reported malaria incidence to identify discrepancies between expected transmission patterns and observed disease burden.

Such mismatches can reveal important blind spots in surveillance systems. Areas with high predicted risk but low reported incidence may indicate underreporting, limited health care access, disruptions to health services, self-medication, or other barriers preventing communities from seeking formal care.

Such mismatches may reveal potential blind spots in surveillance systems and health care delivery, prompting further investigation into the factors influencing malaria transmission and case reporting.

Onimisi recalls being surprised by some of the model's outputs. "When I first ran the model, there were areas where the outputs surprised me," he says. "The risk patterns did not always match what we assumed from existing reporting systems."

Beyond forecasting, the platform can also evaluate the impact of interventions such as mosquito net distribution, vaccination programs, health care worker training, and public awareness campaigns.

Supporting Smarter Intervention Decisions

The malaria intelligence system is designed to support governments, researchers, nongovernmental organizations, and public health agencies involved in malaria control.

Its outputs can guide allocation of mosquito nets and vaccines, identify high-risk communities, support deployment of surveillance teams, and improve targeting of awareness campaigns. Rather than relying solely on national-level prevalence estimates, the platform seeks to identify fine-scale transmission patterns at cluster and community levels, enabling more targeted interventions and more efficient use of limited resources.

It also promotes continuity between intervention cycles by creating an ongoing surveillance and response framework rather than treating malaria control as a series of isolated campaigns.

Scaling Beyond Nigeria

Although Nigeria serves as the initial implementation environment, the long-term vision extends beyond a single country. Onimisi is exploring expansion of the platform to the DRC, another high-burden setting characterized by environmental diversity, complex health systems, and ongoing humanitarian pressures.

Scaling the technology, however, involves more than technical replication. It requires adaptation to different health systems, infrastructures, and local realities.

In the longer term, similar predictive intelligence systems could potentially be adapted to diseases such as Ebola, Lassa fever, and future emerging infectious threats.

Challenges to Implementation

Despite its potential, scaling predictive malaria intelligence systems across Africa presents significant challenges.

Health, climate, and environmental datasets are often fragmented across institutions, limiting interoperability and integration. Computational requirements and funding constraints remain substantial, particularly when processing large geospatial and epidemiological datasets.

Infrastructure limitations, especially in rural areas, also affect data collection and system deployment. Limited internet connectivity, unreliable electricity, and weak digital health systems can constrain practical implementation.

XSL•FO
RenderX

Figure WL1.

> *There is no value in building systems that communities and health workers cannot realistically use.*
>
> Daniel Onimisi, MSc

For Onimisi, technological sophistication alone is insufficient.

"There is no value in building systems that communities and health workers cannot realistically use," he says. Successful adoption, he argues, depends on ensuring that predictive technologies are aligned with the realities of frontline health systems and community needs.

Toward African-Led Disease Intelligence

The malaria intelligence platform highlights a broader transformation underway in global health.

Future surveillance systems are likely to become increasingly interdisciplinary, combining epidemiology, climate science, geospatial analytics, and AI into unified public health frameworks. Their effectiveness, however, will depend on sustained data infrastructure, usability, and integration into public health decision-making systems.

The project also reflects the growing role of African-led innovation in digital health. Rather than relying solely on externally designed surveillance systems, locally developed platforms such as SORMAS (Surveillance Outbreak Response Management and Analysis System), which is used for disease surveillance across several African countries, and Uganda's mTrac health reporting system are increasingly shaping how health data are generated, interpreted, and translated into action.

With Nigeria serving as the primary testing ground and the DRC identified as the next phase of expansion, the project illustrates how African-led intelligence systems could help make malaria surveillance more predictive, adaptive, and responsive to changing patterns of transmission across malaria-endemic regions.

Edited by KA Clegg; submitted 24.Jun.2026; this is a non–peer-reviewed article;accepted 24.Jun.2026; published 30.Jun.2026.

Please cite as:
Muzaki S
Building a Malaria Intelligence System for Real-Time Prediction and Data-Driven Intervention Planning
J Med Internet Res 2026;28:e105472
URL: https://www.jmir.org/2026/1/e105472
doi:10.2196/105472

XSL•FO
RenderX

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jun.2026.

XSL•FO
RenderX

Transformation Versus Innovation in Digital Health Care and the Future of Clinical AI

Boon-How Chew

Abstract

Digital innovation is frequently presented as the key to transforming health care. In this *News and Perspectives* article, JMIR Correspondent and academic physician Boon-How Chew reports on the distinction between and the direction of transformation and innovation, reflecting on the future of clinical AI and what lasting change in health care ultimately requires.

(J Med Internet Res 2026;28:e105359) doi:10.2196/105359

KEYWORDS

digital transformation; innovation; health systems; digital health; institutional change; leadership; clinical AI; artificial intelligence

Textbox 1.

Key Takeaways:

- Transformation and innovation are distinct, directionally related, and often confused. Conflating the two leads to institutions celebrating cosmetic change as progress.
- The digital era has lowered technical barriers but raised cultural demands; institutions remain prolific producers of transformation plans but struggle to convert these documents into reality.
- Understanding what genuinely drives transformation and innovation is a prerequisite for clinicians, administrators, and institutions to adopt and scale safe, effective clinical AI.

Boon-How Chew, MD, MMed, PhD, is an academic physician and professor of family medicine at Universiti Putra Malaysia, specializing in the psychosocial and technological aspects of chronic disease management. His research increasingly explores the intersection of AI innovation in medicine and the reform of scientific writing workflows within legacy academic infrastructures.

Walk into almost any hospital, health ministry, or medical faculty today, and we are likely to find the walls papered with the same vocabulary: *transformation*, *innovation*, *disruption*, *digital*. These words appear in strategic plans, annual reports, and conference themes, with reassuring regularity. What appears far less regularly is clarity about what they actually mean.

Transformation, properly understood, is a fundamental and sustained change in how an institution operates, delivers value, and defines its own purpose. It is not the adoption of a new software platform, the rebranding of an existing service, or the incremental improvement of a workflow that was already functional. Transformation happens when the underlying logic of an organization—its culture, its power structures, its measures of success—shifts in ways that become increasingly difficult to reverse when the next leadership team arrives, though culture always remains vulnerable to regression when an empowered, data-informed, and collaborative learning ecosystem has not been cultivated in the organization. A hospital that moves from measuring success by bed occupancy to measuring it by patient health outcomes has transformed. One that installs a new electronic health record while measuring the same things in the same ways has not.

Innovation is equally misunderstood. In clinical settings, the word is often generously applied to anything new—a revised triage form, a digital appointment reminder, a newly scheduled weekend ward round. True innovation, which can often be disruptive, requires at least two conditions that most incremental improvements do not meet: novelty (the solution did not previously exist in this form) and sustained value creation (the solution durably improves outcomes, not just processes). By this standard, a great deal of what health care calls *innovation* is actually *optimization*—valuable, necessary, but categorically different. The distinction matters because organizations that believe they are innovating when they are optimizing may never build the conditions, risk tolerance, or cultural infrastructure that real innovation requires.

How Transformation and Innovation Relate

Transformation and innovation are related but not synonymous, and their relationship is directional. Transformation typically precedes innovation at scale. An institution that has not fundamentally changed its culture, its governance, and its relationship with failure cannot sustain innovation beyond chance occurrence or isolated individual acts of courage. Conversely, innovations that are genuine, repeated, and spread across an organization are both evidence of and fuel for ongoing

XSL•FO
RenderX

transformation. They are not a sequence to complete but a cycle to maintain.

Figure WL1.

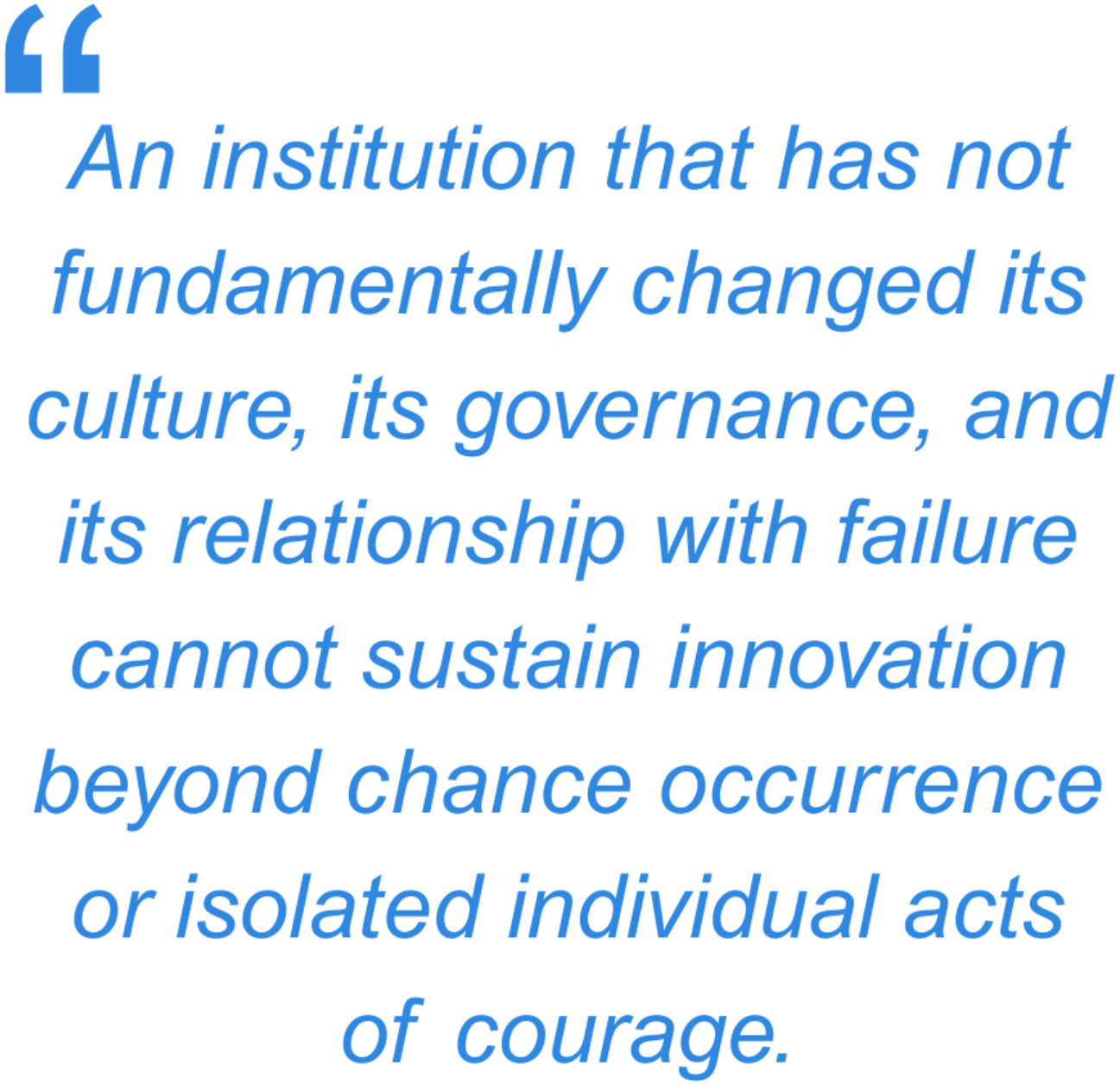

This relationship has specific implications for health care. Clinical environments are deeply hierarchical. Seniority often determines whose ideas are heard, whose concerns are acted on, and whose failures are attributed as such. These dynamics, while functional in emergency decision-making, are in tension with the conditions innovation requires: the freedom to propose ideas that may fail, to question established practices without professional consequence, and to learn from errors rather than conceal them. Transformation, in a clinical institution, means differentiating contexts: building psychological safety for improvement and learning while preserving clear authority structures where patient safety demands them.

What the Digital Era Has Actually Changed

Technical access is no longer the primary binding constraint on digital transformation in health care; institutional readiness is key. Decades ago, barriers were substantially technological—limited computing infrastructure, no interoperable data standards, no real-time decision support. Today, a small clinic in a low-income country can access cloud computing, open-source decision support, and AI-assisted diagnostics at a fraction of what leading hospitals paid for inferior capabilities 2 decades ago, though access and the capacity to maintain, update, and contextualize technology sustainably remain distinct challenges, particularly where skilled technical labor is scarce.

XSL•FO
RenderX

What has not necessarily become easier is the human work. A health system that implements an AI-powered diagnostic tool on top of a culture where clinicians are penalized for admitting uncertainty has not leveraged AI; it has effectively weaponized it. The tool is likely to be used defensively, blamed when wrong, and quietly circumvented when inconvenient.

Clinical AI has introduced an unprecedented dimension to both transformation and innovation: the capacity to automate not just physical or administrative tasks but also cognitive ones. AI can now draft clinical notes, analyze imaging, predict deterioration, and surface drug interactions, with growing evidence of performance comparable to or exceeding that of human clinicians working alone in specific tasks. This does not necessarily make clinicians obsolete; if anything, it makes the specifically human dimensions of clinical work—empathy, ethical judgment, relational trust, and contextual wisdom—more clinically and socially valuable, not less, provided reimbursement and governance models evolve to recognize and reward clinicians accordingly. Realizing this requires institutions to redesign roles, retrain staff, and redefine what good performance means in an AI-augmented environment, including through real-world validation. Those that do not are likely to find that AI amplifies their existing dysfunctions as readily as it amplifies others' strengths.

The Documentation Trap: Performing Transformation Without Achieving It

Health care institutions are world-class producers of transformation documentation—strategic plans, digital road maps, innovation frameworks, and quality improvement initiatives generated in extraordinary volume. Yet, the organizational literature repeatedly suggests that documentary productivity has a paradoxically weak relationship with transformation outcomes. Organizations investing heavily in steering committees, working groups, and pilot programs frequently emerge with refined documents and unchanged practices.

This is not mere cynicism; it is a pattern observed across public sector and health care organizations, though direct causal evidence remains limited. Documentation creates the experience of progress without the institutional risk that real transformation demands: an innovation strategy is visible, measurable, and defensible, whereas changing incentive structures that punish near-miss reporting requires absorbing political cost. This is further compounded by regulatory, accreditation, and funding obligations—factors that drive documentation volume and that leaders may have little power to reduce. Where discretion exists, the preferred path is understandably real transformation, resulting in what researchers term institutional isomorphism—organizations converging in stated strategy while diverging in actual capability.

Clinical AI faces the same institutional test observed with electronic health records and telehealth, where technology has been noted to underperform when the surrounding institution has not changed. In both cases, the etiology was structural and cultural: poor workflow integration and misalignment with clinical norms and professional boundaries—the factors that ultimately determine whether technology is used with intelligence and integrity at scale. Clinical AI appears likely to demand even greater institutional readiness than its predecessors, given the depth of the cognitive and relational functions it touches.

Why This Matters and What Comes Next

For the clinicians and health care professionals reading this, the stakes are not abstract.

The institution you work in is either building the conditions for genuine transformation and innovation or producing the documentation of those things while the underlying reality remains largely unchanged. Transformation provides the essential foundation for innovation by elevating the cultural and workforce realignment.

Digital health is inescapably political. Systemic transformation requires reforming governance to prevent innovation theater and ensure clinical AI is actually safe, effective, and sustainable to scale.

Edited by KA Clegg; submitted 23.Jun.2026; this is a non–peer-reviewed article;accepted 23.Jun.2026; published 30.Jun.2026.

Please cite as:
Chew BH
Transformation Versus Innovation in Digital Health Care and the Future of Clinical AI
J Med Internet Res 2026;28:e105359
URL: https://www.jmir.org/2026/1/e105359
doi:10.2196/105359

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jun.2026.

XSL•FO
RenderX

American Medical Association Shares Framework to Address the Escalating Risk of Physician Deepfakes

Shalini Kathuria Narang

Abstract

As AI becomes more sophisticated, new risks are emerging. In this *News and Perspectives* article, JMIR Correspondent Shalini Kathuria Narang reports on the American Medical Association's response to the rising issue of physician deepfakes.

(J Med Internet Res 2026;28:e104953) doi:10.2196/104953

KEYWORDS

deepfakes; physician impersonation; Character.AI; Medical Practice Act; chatbots; American Medical Association, AMA

Textbox .

Key Takeaways:

- Deepfakes created to impersonate doctors, manipulate the public, and endorse unproven treatments pose threats to patients and the broader health care system.
- The American Medical Association has released a new policy framework to establish clear, enforceable protections for physicians against unauthorized AI-generated deepfakes.
- Some states' legislatures in the United States have passed laws to discipline vendors where an AI chatbot is engaged in the practice of medicine or is impersonating a human engaged in the practice of medicine.

A deepfake is a video, photo, or audio recording that seems real but has been manipulated with AI. The underlying technology can replace faces, manipulate facial expressions, and synthesize speech, depicting humans appearing to say or do something that they have never said or done.

AI-generated deepfake "doctors" impersonating physicians garner millions of views on social media, misleading the public and endorsing unproven treatments. The content typically promotes products for their creators' financial gain while exposing the viewers to scams that erode the patient-physician relationship, undermine confidence in evidence-based care, and put the public at risk of harm.

According to the factchecking organization Full Fact, hundreds of AI-generated deepfakes of real doctors and academics to promote health products with bogus endorsements are damaging the reputations of those they impersonate and could lead to people making decisions about their health based on fake claims.

"[The] mental health space is an area of big concern, including significant privacy concerns, more with some of the general purpose chatbots claiming to help support mental health. We know many, many, many individuals and patients are using these chatbots for support that can cross the line and be dangerous, as we've seen suicides and bad outcomes related to the use of these chatbots in certain ways," says Shannon Curtis, JD, Senior Director of Policy Development for the American Medical Association (AMA) Center for Digital Health and AI.

She adds that the AMA is looking for solutions to begin addressing the issue, including transparency to help users understand that they're not actually engaging with a licensed clinician.

"We want to see a prohibition on chatbots from claiming they're a licensed professional, or that they're providing equivalent services as a licensed professional," says Curtis. "We have called for more action to update our regulatory structure... A call for new, updated, appropriate regulatory oversight structures that can help provide that clarity and consistency for consumers, for physicians, for the industry, about what needs to be regulated, and what does not. And hopefully, turning that into a more appropriate, risk based regulatory system that fits where AI is, and hopefully fits where it is going."

As AI chatbots become more sophisticated, and more harms are realized, it is likely more states will create regulations and exercise enforcements.

New AMA Policy Framework

Responding to the escalating risk of AI deepfakes, and recognizing that existing privacy, employment, and intellectual property laws do not adequately address AI deepfake risks in clinical contexts, the AMA Center for Digital Health and AI has created a policy framework to modernize physician identity protections and close legal gaps to uphold patient safety, professional integrity, and public trust.

The framework is built on 7 policy principles:

- Physician identity as a protected right: A physician's name, image, likeness, voice, and digital replicas are protected

XSL•FO
RenderX

interests and should not be used beyond the scope of the clinician's consent.

- Prohibition on deceptive medical impersonation: AI-generated or altered content falsely conveying a physician's endorsement, authorship, or medical judgment likely to mislead in a health-related decision must be prohibited.
- Informed, opt-in, and revocable consent: Use of a physician's identity in AI-generated or manipulated content requires affirmative, informed, opt-in consent.
- Mandatory transparency and labeling: AI-generated or altered content depicting a physician must be clearly labeled. Patients interacting with an AI-generated health professional must be alerted before the interaction.
- Shared responsibility to prevent impersonation: All participants—platforms, hospitals, health systems, and AI vendors—share responsibility for preventing identity misuse.
- Enforcement and practical remedies: Processes must be in place to allow physicians to document, escalate, and address identity misuse. Institutions and platforms must facilitate investigation and escalation.
- Minimizing administrative burden: Protecting physician identity should be a default that doesn't place any undue burden on the physician.

"AI deepfakes that impersonate physicians are not just scams—they are a public health and safety crisis," said AMA CEO John Whyte via a statement. "When bad actors exploit a doctor's identity, they undermine patient trust and can steer people toward harmful, unproven care. We need strong action by federal and state lawmakers to protect physicians' identities, ensure transparency, and stop this fraud. Safeguarding professional integrity is essential to preserving trust and delivering high-quality care in a rapidly evolving digital landscape."

The new AMA framework aims to guide how the organization works with government officials and industry partners to stop AI-generated deepfakes of physicians. The AMA is eager to collaborate with lawmakers, regulators, and industry to protect patients and doctors from these risks.

"We were seeing a lot of physicians having their names stolen. It's turning up on things like academic papers, that they had no part of. Our framework seeks to address those issues by creating a prohibition and requiring affirmative, opt-in consent from the physician in question if their name, image likeness, voice, etc, is going to be used going forward," says Curtis.

Enforcing Guidelines

While the framework currently functions as a set of guidelines, Curtis notes that the AMA intends for it to lead to enforceable standards.

"Via the policy framework, we are seeking to create statutory or legal obligations that would prohibit a third party from creating a digital replica or a deepfake that utilizes a person's name, image, likeness, or voice in a nonconsensual way," she says.

She emphasizes the importance of enforcement mechanisms like statutory or regulatory requirements on the platforms that end up hosting this content and enforcement by a government agency like the Federal Trade Commission for those bad actors that are creating the deepfakes and disseminating them.

"Data privacy is a huge issue that is sometimes not well understood by individuals. We'd like to see some more action to help protect patients' personal health information," says Curtis. The AMA is collaborating with other groups to support the notice-and-removal obligations in the Take It Down Act and a growing coalition in support of the enactment of the NO FAKES Act.

Regulating medical practice typically falls under the jurisdiction of state medical boards in the United States. In addition to a recent landmark lawsuit filed by the Governor of Pennsylvania against Character.AI, some states' legislatures have passed laws—for example, AB 489 in California, HB 2748 in Oregon, and the Conversational AI Safety Act in Nebraska—to discipline vendors where an AI chatbot is engaged in the practice of medicine or is impersonating a human medical professional.

XSL•FO
RenderX

Figure WL1.

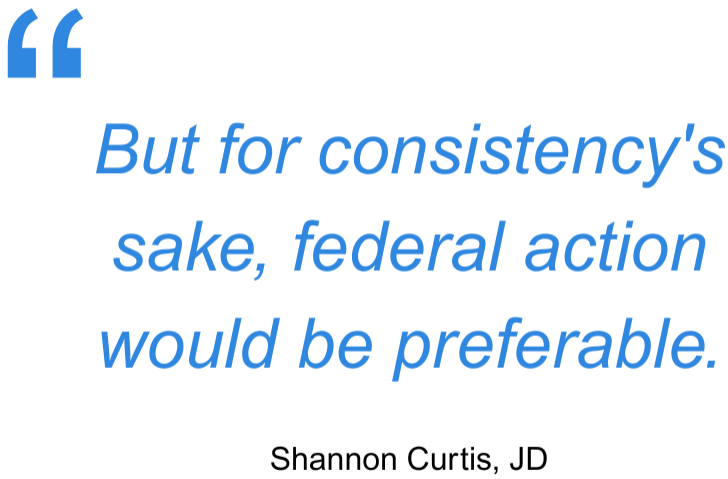

"We are very grateful to see the states taking these issues seriously and taking action. But for consistency's sake, federal action would be preferable," adds Curtis.

Check out this related article on the tensions and regulatory challenges around medical chatbots and the line between informational tool and medical practice.

Edited by KA Clegg; submitted 17.Jun.2026; this is a non–peer-reviewed article;accepted 17.Jun.2026; published 26.Jun.2026.

Please cite as:
Narang SK
American Medical Association Shares Framework to Address the Escalating Risk of Physician Deepfakes
J Med Internet Res 2026;28:e104953
URL: https://www.jmir.org/2026/1/e104953
doi:10.2196/104953

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Jun.2026.

XSL•FO
RenderX

Identity Theft in Academia: Predatory Journals, AI, and the Rise of False Authorship

Cliff Dominy

Abstract

Fraudulent papers are on the rise in scientific publishing. In this *News and Perspectives* article, JMIR Correspondent Cliff Dominy reports on recent cases of false authorship and how they might be prevented to safeguard trust in science.

(J Med Internet Res 2026;28:e104934) doi:10.2196/104934

KEYWORDS

publishing; authorship; scientific misconduct; peer review; artificial intelligence; open science; retraction of publication as topic; editorial policies

Textbox 1.

Key Takeaways:

- Predatory journals are stealing the identities of reputable scientists in false authorship scams to improve their credibility with prospective clients.
- Recent high-profile cases have caused reputational harm to authors and editors and have exposed weaknesses in our trust-based academic system.
- Prepublication screening of academic manuscripts, along with the reform of academic incentives, will probably be central components of an effective response.

One of the interview sources in this article, Nathaniel Gore, is an employee of JMIR Publications Inc. No financial or other potential conflicts of interest exist with the author.

One afternoon in September 2024, Diomidis Spinellis, a Professor of Software Engineering at the Athens University of Economics and Business, opened his inbox. An email from a colleague stood out. Would he be able to write a brief article for the departmental newsletter covering his recent paper, “Global Business Strategies in the Digital Age?” Spinellis, a prominent academic, did not recall writing it, so he opened the hyperlink to refresh his memory. To his disbelief, he found his name on a paper that wasn’t his, in a journal he hadn’t heard of.

The journal picked the wrong guy to impersonate; Spinellis is an expert in mining data repositories, and he got straight to work. The *Global International Journal of Innovative Research* (*GIJIR*) was relatively new, with just two volumes containing 220 academic papers. Further research picked up some red flags, showing that many of his fellow GIJIR “authors” were well known in their fields, but likely also impersonated—some had already passed away at the time of their supposed publications. Worst of all, the fake paper had attracted three citations in the months since its publication.

It got worse; the quality of the writing was, in Spinellis’s opinion, “trash.” He ran a few papers through an AI detector, usually reserved for his students’ essays. The result: generative AI was responsible for up to 100% of each paper. Spinellis dissected the *GIJIR* dataset and published his findings in a case study. He found their papers fell into three broad categories.

1. False authorship papers: the predominant group, typically AI-generated by the journal, with unsuspecting scholars’ names used to boost the journal’s credibility.
2. Bought authorship papers: often AI-generated, either by unscrupulous authors or as a service by journals for their clients.
3. Genuine papers (the minority): real research by real scientists unaware of or unconcerned about the journal’s practices.

False authorship is not new, but thanks to AI it is increasingly easy to produce. In May 2025, members of the International Committee of Medical Journal Editors (ICMJE) recognized the growing problem and alerted the scientific community to the strategies used by predatory journals, which included the following:

- Mimicry of established journal names and branding
- False claims of membership of the Committee on Publication Ethics, Council of Science Editors, or ICMJE
- Fabricated indexing and citation metrics
- Prominent scientists as authors and editors without their knowledge or consent

False authorship papers are an advertising tool used by predatory journals to attract unscrupulous scientists to the brand. These journals will “assume the identities of [established scientists] to confer legitimacy on themselves,” says Reese Richardson, PhD, a bioinformatician at Northwestern University in Illinois, who has tracked and continues to track predatory publishers.

XSL•FO
RenderX

The Topol Incident

In April 2026, cardiologist, author, and social media icon Eric Topol, MD, discovered his name had been used in a false-authorship scam. With over one million followers across several social media channels, Topol was alerted to a paper titled "Implementation Science for AI Integration in Digital Health Systems" in the *Journal of Digital Health Implementation*. Within hours of him sounding the alarm, the paper, journal, and publisher had disappeared. An investigation by the science watchdog Retraction Watch revealed that the fraudsters had followed the predatory playbook to the letter. Their big mistake—using the name of a leading digital health figure in the ruse, which dramatically reduced its half-life in the scientific record.

It is not only the authors who suffer reputational harm in these schemes. The supposed "editors" of the paper—themselves impersonated by the journal—unwittingly got involved in the scandal. Zarnie Khadjesari, PhD, a professor of Health Science at the University of East Anglia, received a message from fellow "editor" Angelo Rossi-Mori, PhD, alerting her to the fake paper.

Both well-respected scientists were, in Khadjesari's words, "angered and disturbed" by the deception. To compound matters, they had been mistakenly and very publicly identified as complicit in the fraud—which had been circulated to Topol's more than 1 million followers. For Khadjesari, this added insult to injury, an experience she describes as "exhausting and emotional," and she took to social media to defend herself.

Stopping Predatory Journals

Detection: Sooner, Not Later

Can technology itself tackle the AI-accelerated fake authorship problem? Rossi-Mori, former associate researcher at the Italian National Research Council, states via email, "This is a serious problem, but not a completely new one." He blames generative AI and "the rise of social media, [for] the speed, scale, and apparent credibility with which such false material can now circulate."

XSL•FO
RenderX

Figure WL1.

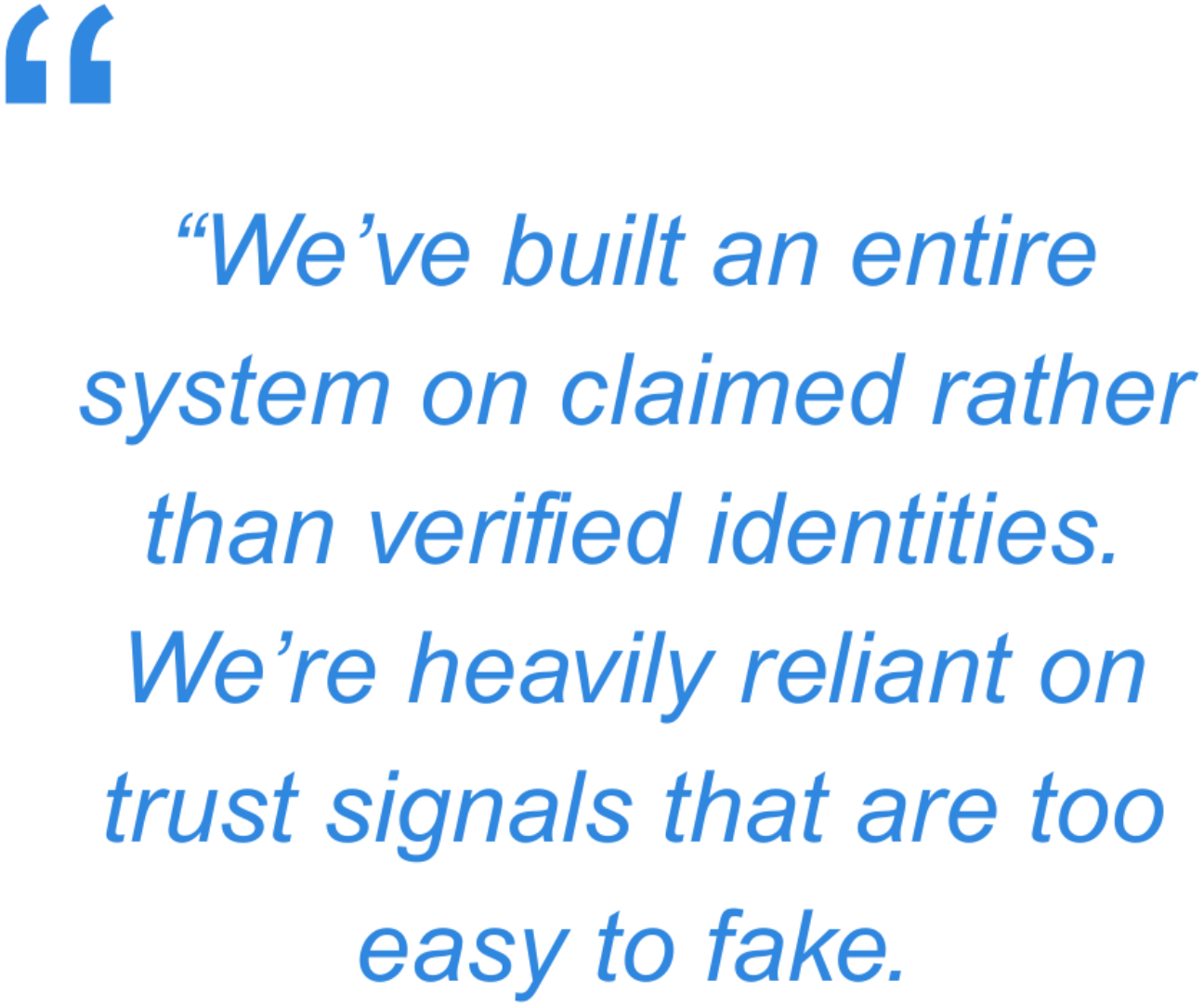

Screening for false science must happen prepublication; like toothpaste, it’s better managed while still in the tube. An effective screening system would help researchers flag predatory journals and journal editors flag fraudulent papers before they can burden the peer-review system:

- Validating the authenticity of the journal: Is the journal’s identifier (ISSN) legitimate, is the digital object identifier (doi) address valid, and has the paper been independently indexed in PubMed or Google Scholar? Think.Check. Submit.org provides a checklist for researchers doing due diligence on unfamiliar publications.
- Validating the authenticity of the author: False authorship involves both real and imaginary people, and establishing an independent registry of verified scientists with valid affiliations would expose bogus science before it can enter the scientific record.

To address the identity issue, Nathaniel Gore, Product Development & Partnerships Director at JMIR Publications, is “building a better infrastructure” with opensci.id. The independent opensci.id registry aims to centralize academic identities in an online database linking valid researchers with their affiliations and body of work. If the authors on a manuscript don’t match the academics in the registry, the paper can be flagged for extra scrutiny. For Gore, it’s about trust. “We’ve built an entire system on claimed rather than verified identities,” he notes, “We’re heavily reliant on trust signals that are too easy to fake.”

Reinventing Scientific Success

These proposed measures are important first steps in preserving the integrity of science. They will make it harder for predatory journals to operate, but in the race between technology and tech-detectors, will they be enough?

XSL•FO
RenderX

At the core of the problem is the publish-or-perish incentive system by which all academics are measured. Richardson feels that “fixing it requires rebuilding the system of incentives that created this problem from the start.” His concern is “can AI-detectors perform effectively at their job,” considering that “the threat is shifting all the time?”

The problem of false authorship is complex and growing, and effective solutions will be expensive. Publishers will need to invest staff and capital to ensure that AI-based screening steps are both accurate and equitable, with a human at the wheel. Ultimately, safeguarding science—and public trust in science—will require the vigilance of the whole scientific community.

Edited by KA Clegg; submitted 17.Jun.2026; this is a non–peer-reviewed article;accepted 17.Jun.2026; published 25.Jun.2026.

Please cite as:
Dominy C
Identity Theft in Academia: Predatory Journals, AI, and the Rise of False Authorship
J Med Internet Res 2026;28:e104934
URL: https://www.jmir.org/2026/1/e104934
doi:10.2196/104934

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Jun.2026.

XSL•FO
RenderX

Centers for Medicare & Medicaid Services to Launch Landmark ACCESS Program

Delaney Rebernik

Abstract

This July, the US Centers for Medicare & Medicaid Services will launch its ACCESS program, a decade-long experiment in outcome-based payment expanded to digital care options. In this *News and Perspectives* article, JMIR Correspondent Delaney Rebernik reports on the players, promise, and potential pitfalls of this initiative.

(J Med Internet Res 2026;28:e105562) doi:10.2196/105562

KEYWORDS

medicare; Centers for Medicare and Medicaid Services, US; chronic disease; disease management; delivery of health care; telemedicine; digital health; patient care management; health care reform; reimbursement mechanisms; value-based health care; outcome assessment; health care; quality of health care; health care innovation; health policy; health services accessibility

Textbox .

Key Takeaways:

- The US Centers for Medicare & Medicaid Services's ACCESS program has attracted over 150 digital health companies that will be compensated in large part by how well they are able to improve outcomes at scale for Medicare Part B beneficiaries.
- Experts are energized by the program's boldness and potential scalability but wary of risks to patient safety and provider retention.

There's an audacious new player in the decades-long push toward value-based care in the United States. Launching July 5, the Advancing Chronic Care with Effective, Scalable Solutions (ACCESS) model is a 10-year payment experiment that could unite traditional provider organizations and digital health disrupters under outcome-based reimbursement.

This new Medicare model pays providers for managing beneficiaries' qualifying conditions with technology-enabled care, and half of that payment depends on providers' ability to improve patient outcomes at scale. Some 150 providers have already been approved for participation, and they run the gamut of unlikely actors: private technology startups like wearables maker WHOOP, digital coaching apps like Lark, and virtual care companies like Pair Team and WellDoc that combine clinician- and AI-driven interventions.

Policy experts and clinicians say the program is a bold departure from historic initiatives helmed by the Centers for Medicare & Medicaid Services's (CMS) innovation hub (CMMI).

"It's not what people may have expected. It really is a lot of tech companies," says Robert Longyear, CEO of health care policy consultancy Longyear Health in Washington, DC, and cofounder of a remote physiological monitoring company similar to ones participating in ACCESS.

Despite the promise presented by these innovative partnerships, the program's novelty and ambition raise questions around execution. Reimbursement rates are lower than expected, which could be a sticking point for participating organizations with more expensive products, such as hardware, Longyear says.

XSL•FO
RenderX

Figure WL1.

> *But this project skips over the validation stage and it makes the patients beta testers…for as yet unproven [technology].*
>
> Robert Trestman, MD, PhD

Additionally, the rapid uptake of new clinical offerings from nontraditional providers could impact patient safety, says Robert Trestman, MD, PhD, chair of the American Psychiatric Association's (APA) Council on Healthcare Systems and Financing, which provided input on measures for the behavioral health conditions included in the program's initial clinical tracks.

"It is a time of experimentation, which is great," says Trestman, who's also chair of the Department of Psychiatry and Behavioral Medicine at Carilion Clinic and Virginia Tech Carilion School of Medicine in Roanoke, Virginia. "But this project skips over the validation stage and it makes the patients beta testers—typically unwitting beta testers—for as yet unproven [technology]."

ACCESS Overview

ACCESS is a voluntary model available to Medicare Part B beneficiaries nationwide. The four initial clinical tracks reflect some of the most common—and highest-cost—chronic conditions covered, including hypertension, diabetes, cardiovascular disease, chronic musculoskeletal pain, depression, and anxiety.

According to CMS, each track groups related conditions that are often treated similarly, and participating organizations must manage all qualifying conditions in the track(s) of which they're part.

CMS plans to issue monthly payments but withhold half of each for reconciliation with clinical outcomes at the end of the 12-month care period. To earn the full payment, at least half of an organization's aligned beneficiaries must meet all required clinical targets.

ACCESS payment levels are modest compared to existing digital health payment determinants, such as remote patient monitoring (RPM) and chronic care management (CCM) codes. The lower-than-expected rates could force innovation—or "inhibit profitability," Longyear says, especially when it comes to makers of higher-cost tech, like wearables and other hardware. "I think we're going to see a lot of exits from the program."

The Promise

Created as part of the Affordable Care Act, CMMI is charged with delivering better health outcomes at lower cost by shifting payment incentives from the volume of services provided (fee-for-service) to the outcomes patients experience (value-based care).

In its 16 years of operation, CMMI has launched dozens of models with mixed results. A 2023 Congressional Budget Office analysis, for example, concluded that the center had *increased* federal spending by more than $5 billion over its first decade rather than producing projected savings. Even so, certain models, like accountable care organizations (ACO), "have demonstrated measurable, if modest, capacity to bend the cost curve," wrote Longyear in a recent article. "The overarching lesson from a decade of ACO evaluations is that meaningful spending reductions are achievable within fee-for-service Medicare, but they require time, organizational learning, appropriate risk calibration, and payment structures that genuinely alter care delivery incentives."

ACCESS could help build on this lesson through its blending of a capitated rate with payments linked to specific outcomes, Longyear says. He calls that level of granularity a departure

XSL•FO
RenderX

from the ACO's "broad bucket of savings" and complex payment layers, meaning it could offer clearer insight into which interventions move the needle.

Figure WL2.

> *This is a real attempt at trying to do something clinical and that's actually scalable.*
>
> Robert Longyear

Experts also hope that technology-enabled interventions, coupled with stronger protections and human guidance, might help close critical care gaps.

"The treatment of choice for mild to moderate depression and anxiety is psychotherapy. It's just that there aren't enough available psychotherapists to do it," says Trestman. Rather than using medication as a first-line treatment due to resource shortages, clinicians could leverage tools that help patients practice gold-standard modalities like cognitive behavioral therapy, making "the need for psychotherapy more manageable with fewer sessions."

If ACCESS proves effective, Longyear predicts it'll be useful across a much wider array of conditions and outcomes. "This is a real attempt at trying to do something clinical and that's actually scalable," he says.

A program aspect that could advance this goal is CMS's plan to randomly assign about 10% of year-one participating beneficiaries to a control group to help sharpen the evaluation of the program's digital health interventions. "It's a pretty innovative policy tool," Longyear says.

The Pitfalls

Though both Longyear and Trestman are impressed by the ACCESS model's use of a randomized control group, they also point to some inherent weaknesses.

The incorporation of so many different technologies may make it difficult to amass "enough data to drill down for any one of them to determine which ones may be a benefit," Trestman notes.

There's also the potential for selection bias. Because the program's patient participants, like their provider counterparts, are self-selecting, they may already be more comfortable using a health app or more active with their overall health, Longyear says. Additionally, some of the initially covered conditions, such as depression, have the potential to resolve spontaneously for a significant portion of patients within the program's treatment timeline, Trestman says. "So that is going to be a complicating factor to determine whether...people will be benefiting from this or what will be seen as a massive placebo effect."

Privacy is another concern.

"As we've seen recently, there are profound data breach risks even in well-developed, sophisticated, and protected health information systems," Trestman says, citing the 2024 Optum cyberattack. "When you have all of these relatively small startup companies, the vendor risk assessments may not be adequate to defend against a hacking event, so this may expose people's PHI [personal health information] to further risks."

Aside from the data ambiguities, the model's emphasis on bringing new technologies to market at scale and speed has yielded what Trestman calls a "very unusual" collaboration between CMS and the US Food and Drug Administration, with the latter loosening requirements around health tech authorization at a time AI disruption is outpacing regulation.

There's a real risk of "causing harm" unless the tools are "carefully monitored and modified appropriately over time," Trestman says.

He's especially worried about the potential for increased errors in diagnosing and treating patients, especially in cases where

XSL•FO
RenderX

care is rendered asynchronously. “Most of these companies are not big enough to have that level of investment of oversight.”

Trestman and Longyear are also concerned about the program’s implications for care coordination given the likelihood that it’ll introduce new portals for patients and primary care providers to juggle. “Even today, it’s extremely difficult for us to share information from one health system to another,” Trestman says. “If you’re having innovative technologies, are they really going to be integrated effectively?”

Launch Signals to Watch

The most obvious sign that ACCESS is heading in the right direction will be participants hitting their outcome-aligned payment thresholds, Longyear says, though he notes that selection bias could be a complicating factor.

He’s also keeping a close eye on attraction and retention. “Can you get patients in the door, and can you keep them?”

Health app engagement rates among similar populations have historically been low but intervenable. “The usability really matters,” Trestman says. Part of that, he says, comes down to designing and refining technologies along dimensions like culture and language.

Regardless of how successful the program is, it will stop short of addressing the social determinants of health at the root of poor outcomes and inefficient care, Trestman says. “Unless people have safe places to live, access to good food, a meaningful job that pays the bills, a safe environment in which to live, then we’re fixing something after it breaks rather than preventing it from getting broken in the first place.”

Edited by KA Clegg; submitted 25.Jun.2026; this is a non–peer-reviewed article;accepted 25.Jun.2026; published 30.Jun.2026.

Please cite as:
Rebernik D
Centers for Medicare & Medicaid Services to Launch Landmark ACCESS Program
J Med Internet Res 2026;28:e105562
URL: https://www.jmir.org/2026/1/e105562
doi:10.2196/105562

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jun.2026.

XSL•FO
RenderX

Regulating AI Chatbot Impersonations of Medical Professionals

Tejas S Athni

Abstract

Conversational medical AI systems are increasingly blurring the lines between simply being digital tools and simulating medical practice. In this *News and Perspectives* article, MD-PhD candidate and JMIR Correspondent Tejas S Athni reports on recent health AI controversies and evolving regulatory challenges.

(J Med Internet Res 2026;28:e104835) doi:10.2196/104835

KEYWORDS

physician impersonation; licensed professionals; medical licensing law; consumer protection law; AI chatbots; generative AI; medical AI; conversational AI; digital health; fiduciary duty; state AI legislation; artificial intelligence

Textbox 1.

Key Takeaways:

- Conversational medical AI systems increasingly present themselves in ways that simulate licensed health care providers, while simultaneously disclaiming the legal responsibilities of medical practice.
- Existing medical licensing, medical practice, and consumer protection frameworks may be insufficient to adequately regulate this challenge.
- An emerging patchwork of state and federal legislation suggests that future regulation may focus not only on inaccurate medical advice, but also on whether AI systems create the perception of legitimate clinical authority.

As autonomous generative AI enters health care, a growing number of AI chatbot platforms are presenting themselves as physician-level experts while simultaneously disclaiming legal responsibility for the medical advice they provide. Terms such as “AI doctor,” “virtual physician,” and “medical assistant” are becoming mainstay marketing language for these tools, despite lacking formal licensure.

This has created a rapidly expanding gray zone in which users believe they are interacting with a medically vetted system, even while companies insist otherwise in fine print. Regulatory frameworks, including medical malpractice law, state medical licensing statutes, and consumer protection law, have failed to keep pace with conversational AI that simulates professional health care providers.

Character.AI Lawsuit

Concerns surrounding AI impersonation of medical professionals were exemplified in a May 2026 enforcement action filed by the state of Pennsylvania against Character.AI, a conversational AI platform where users can create custom AI-generated avatars capable of dialogue and role-play. The platform hosts 20+ million monthly active users and 18+ million chatbots.

The lawsuit, described by Governor Josh Shapiro (D-PA) as the first gubernatorial enforcement action of its kind in the United States, alleged that chatbots on the platform falsely represented themselves as licensed health care professionals in violation of Pennsylvania’s Medical Practice Act. According to the complaint, one chatbot named “Emilie” described itself as a licensed psychiatrist, claimed to have medical training at Imperial College London, asserted licensure in both Pennsylvania and the United Kingdom with a fabricated Pennsylvania medical license number, and exercised the authority to assess whether medication could help a depressed patient.

The case raises a new legal question with potentially broad implications for AI regulation: is the developer of an AI platform liable under medical licensing law for statements generated through user-created chatbots?

The state’s petition did not target the specific user who created the chatbot persona, but rather only Character Technologies, the developer of the Character.AI platform. Rather than relying on newly enacted AI-specific legislation, Pennsylvania grounded its claims entirely within traditional medical licensing law. One statutory provision invoked, 63 P.S. § 422.10, defines unauthorized medical practice to include purporting to practice medicine or representing oneself falsely through professional titles or claims of licensure. A second provision, 63 P.S. § 422.38, authorizes injunctive relief against unlawful medical practice even in the absence of patient harm. The litigation is essentially an early test of whether existing medical licensing or consumer protection frameworks can be extended to conversational AI systems without entirely new legislation.

The lawsuit also highlights a broader concern beyond fabricated credentials alone. Conversational AI systems, operating outside the traditional social contract of the dyadic physician-patient relationship, generate emotionally expressive and personalized interactions that may create an illusion of expertise.

XSL•FO
RenderX

Scholars have argued whether such systems with no fiduciary duty or accountability should be permitted to occupy the same functional role as medical professionals. In this context, regulators may increasingly focus not only on whether chatbot outputs are factually accurate, but also on whether the human interaction itself reasonably causes users to perceive the system as a licensed medical or clinical authority. The distinction between explicit impersonation and title appropriation may become important in medical AI governance.

Doctronic Controversy

A second controversy illustrates when AI appropriates medical professional authority and identity. In January 2026, the State of Utah partnered with the AI-powered prescription renewal system Doctronic to automate routine medication refills for in-state residents. The Doctronic platform is marketed as an AI doctor, with simultaneous disclaimers that it is unlicensed, does not practice medicine, and does not provide patient care or medical advice. The system emphasizes Health Insurance Portability and Accountability Act compliance and that its model was trained using peer-reviewed literature.

Despite these safeguards, the Utah Medical Licensing Board warned that refill decisions inherently require ongoing clinical reassessment—including adverse effect monitoring, contraindication review, and professional medical judgment—and ultimately recommended suspension of the AI program pending review.

The controversy highlights a core ambiguity confronting medical AI regulation: companies may benefit from the *appearance* of medical authoritativeness yet avoid the formal legal obligations of licensed medical practice. The Utah Medical Licensing Board's objections suggest that regulators may increasingly interpret existing medical practice statutes as encompassing AI systems that perform functions traditionally reserved for clinicians. Regulators will face the challenge of determining when these AI systems have crossed the line from just providing information to practicing medicine.

Figure WL2.

...Is the developer of an AI platform liable under medical licensing law for statements generated through user-created chatbots?

XSL•FO
RenderX

An Emerging Regulatory Patchwork

Both controversies demonstrate how conversational medical AI systems are increasingly entering territory once reserved for licensed human health care professionals. At present, existing medical licensing or consumer protection frameworks remain underdeveloped to address these challenges—the legal landscape is fragmented.

At the federal level, the CHATBOT Act (HR 7985) introduced in the House of Representatives is aimed at prohibiting AI systems from impersonating licensed professionals, including in medical contexts. Yet, the executive branch has, in contrast, taken a different stance, directing the Department of Justice to establish an AI Litigation Task Force to challenge state AI laws deemed excessively obstructive in late 2025.

Other states have created or proposed landmark legislation related to AI impersonations of medical professionals. Oregon passed House Bill 2748 to prohibit nonhuman entities, including AI systems, from using the title of nurse or other nursing-related designations. Tennessee passed SB 1580 to similarly prohibit AI systems from representing themselves as licensed mental or behavioral health professionals and treats violations as deceptive trade practices. California's AB 489 bars AI systems from using postnominal letters such as MD or RN and prohibits user interfaces that falsely imply interaction with a licensed clinician. Texas's Responsible Artificial Intelligence Governance Act (TRAIGA) legislation contains a provision devoted to disclosure when AI is used in medical diagnosis or treatment contexts.

Finding the Line

As conversational medical AI systems increasingly simulate the role and authority of licensed clinicians, regulators are confronted with the challenge of whether existing medical licensing and consumer protection laws are sufficient. With a patchwork of state and federal legislation emerging, regulators may increasingly focus on whether these systems create the functional impression of practicing medicine. The central legal challenge may therefore become determining when medical AI crosses the line from informational tool into the unauthorized practice of medicine.

Edited by KA Clegg; submitted 16.Jun.2026; this is a non–peer-reviewed article;accepted 16.Jun.2026; published 25.Jun.2026.

Please cite as:
Athni TS
Regulating AI Chatbot Impersonations of Medical Professionals
J Med Internet Res 2026;28:e104835
URL: https://www.jmir.org/2026/1/e104835
doi:10.2196/104835

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Jun.2026.

XSL•FO
RenderX

How Digital Twins Can Improve Health System Operations

Mark Crawford

(J Med Internet Res 2026;28:e98113) doi:10.2196/98113

KEYWORDS

hospitals; operations research, hospital; workflow; patient safety; systems analysis; simulation techniques; health care quality, access, and evaluation; resource allocation

Textbox .

Key Takeaways

- Digital twins use millions of data points to create detailed digital replicas of hospital operations.
- Carefully built and implemented, digital twins have the potential to improve operational efficiencies and patient experience, and support more sustainable health care systems.

Digital twins (DTs) are detailed virtual representations of objects, systems, or processes. DTs have been around for decades—the National Aeronautics and Space Administration (NASA) is generally credited with establishing the core idea of building an exact "twin" of a spacecraft for simulation and testing before embarking on high-risk space launches [1,2]. Back then, these early twins could only be made from solid materials.

Fully digital today, DTs can "faithfully mirror real-world systems...and provide predictive insights using advanced simulation, machine learning, and reasoning to help decision making" [1]. DTs have experienced a resurgence in popularity in recent years, with applications from health care to education, agriculture, and beyond [2,3].

DT technology can simulate a hospital's complete ecosystem, including staff workflow, fluctuations in supply and demand, procurement, and complex patient pathways. Compared to traditional simulation models, which typically involve stand-alone analyses without a consistent link to the physical system, DTs allow a continuous exchange of data between a system and its twin, facilitating real-time updating [2].

"DTs empower hospital leaders to make data-driven decisions by testing different scenarios in a safe, digital environment," noted GE HealthCare in a recent blog post [4]. In health care, where experimenting directly with patient care processes carries real risk, DTs offer a safe environment to test, monitor, and optimize operational changes and build more sustainable systems.

US health care systems are increasingly enthusiastic about implementing DTs to optimize their operations, with the global digital twin health care market expected to experience a 68% compound annual growth rate by 2030, representing a total value of about US $60 billion [5].

DTs in Health Care Management

In health care management, current DT work aims to reduce costs by improving patient flow and throughput, resource allocation, facility design, and workflow optimization. Potential use studies include the following.

Workflow/Patient Flow

Hospitals often struggle with bottlenecks in emergency departments, operating rooms, and inpatient units. "By creating a digital twin of the hospital's operations, administrators can test staffing levels, bed allocation policies, and scheduling strategies before implementing them in the real system," said David Andersen, regional director for AnyLogic, a simulation modeling software company. "In several of our healthcare projects with DTs, emergency department wait times dropped by 20% - 40% and overall patient throughput increased by 10% - 20%."

Surgical Block Optimization

Many systems still rely on static block schedules and manual planning. "Hospitals often sit on more data than they realize and use almost none of it to model trade-offs," observed Steve Blatney, director of business development for SimWell, a company that uses DTs to simulate and optimize business operations. "We have used DTs with orthopedic health systems to see how changes to block scheduling, case sequencing, and staffing allocation affect throughput and cost—making final decisions before anything goes live," he said.

Capacity Planning

Hospital capacity can be affected by several factors. GE HealthCare teamed up with Children's Mercy Hospital (CMH) in Kansas City, Missouri, to optimize capacity planning [6]. DTs have identified demand and capacity constraints, including during the heightened strain of the winter flu season. "CMH used this new knowledge when planning for winter surge capacity planning including spikes of flu or RSV," said Bree Bush, general manager of GE HealthCare's Command Center. "As a result, CMH predicted the winter surge within a week of when it actually struck."

XSL•FO
RenderX

Facility Planning

Physical layouts can have dramatic impacts on patient flow and staff efficiency. Hospitals frequently use simulation models when designing new emergency departments, specialty centers, or inpatient towers. DTs allow leaders to make data-driven decisions to ensure facilities will meet long-term demands.

Effective Deployment and Use

While DTs have tremendous potential, "at the end of the day, success still hinges on an organization's ability to execute change effectively and manage project implementations well," said Kristen Juarez, head of clinical innovation and informatics at Siemens Healthineers, a health technology firm that specializes in digital health solutions. "Learning is crucial for rapid success, but knowledge without action loses value quickly."

Before a hospital can embrace DTs in a meaningful way, everyone—from chief executive officer to night shift nurse—must understand how the process functions—not how it is designed or documented, "but how it actually works on a Tuesday afternoon when two nurses call out and the lab is backed up," said Blatney. "Until you can get to that shared definition of reality, you cannot build anything useful on top of it. Defining that process is the foundation and is often more valuable than any software."

Shared understanding facilitates proactive planning. "Hospitals that understand their operations at a systems level have a real advantage, because simulation-based predictive digital twins allow healthcare leaders to move from reactive thinking to proactive planning," said Andersen. With regular data updates, a DT system can be used to test additional scenarios as frequently as needed for years to come.

Mistakes to Avoid

DTs go wrong if used improperly or built on weak assumptions. Poor or incomplete data is a particular risk. A digital twin relies on accurate inputs about patient arrivals, treatment times, staffing patterns, and operational policies. If inputs are outdated or incomplete, the model can produce misleading conclusions. In practice, "a large part of building a useful digital twin is validating it against real hospital performance and continuously refining it as conditions change," said Andersen.

Another common mistake is oversimplifying complex operations. Hospitals are highly interconnected systems; changes in one department can affect patient flow across the entire facility. If a model only represents part of the system, leaders might optimize one area while unintentionally creating bottlenecks elsewhere. High-quality simulation models need to capture systemwide interactions.

A third mistake involves treating a predictive digital twin as a one-time technology project, rather than integrating it as an ongoing decision support tool. The most successful hospitals use these models operationally to repeatedly test policies, staffing strategies, and capacity planning decisions.

DTs and the Future of Health Care

DT applications will likely continue to grow across health care planning and operations.

GE HealthCare recently announced collaborations with two leading US health systems to develop new AI-driven hospital operations software using DT inputs [7]. At Virginia Polytechnic Institute and State University in Blacksburg, systems engineering faculty are developing a DT model for primary care practice. "We are examining both observed and digital workflows with the hope of developing meaningful changes that can be implemented without disrupting that care," said John Epling, MD, professor of family and community medicine at the university's Carilion School of Medicine. Virginia Tech is also planning a DT study to predict the burnout trajectory for medical staff [8].

To date, DTs in health care systems have primarily involved simulation for planning purposes. The next stage is connecting DTs to sources of real-time data. "When you feed that data into a live operational model, the platform stops being a planning tool and starts being a live decision support system," said Blatney. "Everyone in the operation can see what is actually happening and ask what-if questions in the moment. That is a transformational step. Organizations that figure out how to connect those two streams will have a competitive edge as thought leaders in their fields."

Updated April 27, 2026: An earlier version of this article incorrectly attributed the quote by Kristen Juarez to Julie Gibson, head of US corporate press for Siemens Healthineers. This has been corrected.

Conflicts of Interest

None declared.

References

1. Katsoulakis E, Wang Q, Wu H, et al. Digital twins for health: a scoping review. NPJ Digit Med 2024 Mar 22;7(1):77. [doi: 10.1038/s41746-024-01073-0] [Medline: 38519626]
2. Singh M, Srivastava R, Fuenmayor E, et al. Applications of digital twin across industries: a review. Appl Sci (Basel) 2022;12(11):5727. [doi: 10.3390/app12115727]
3. Grieves MW. Digital twins: past, present, and future. In: Crespi N, Drobot AT, Minerva R, editors. The Digital Twin: Springer; 2023:97-121. [doi: 10.1007/978-3-031-21343-4_4]

XSL•FO
RenderX

4. How a purpose-built digital twin is changing hospital operations. GE HealthCare Research. 2025 Oct 28. URL: https://research.gehealthcare.com/across-the-enterprise/how-a-purpose-built-digital-twin-is-changing-hospital-operations-jb35456xx/ [accessed 2026-04-17]
5. Digital twins in healthcare market report: growth, size, share, and trends. MarketsandMarkets. 2025 May. URL: https://www.marketsandmarkets.com/Market-Reports/digital-twins-in-healthcare-market-74014375.html [accessed 2026-04-17]
6. Children's Mercy Kansas City increases access to care after launching command center. GE HealthCare. 2025 Apr 10. URL: https://www.gehealthcare.com/en-ph/insights/article/childrens-mercy-kansas-city-increases-access-to-care-after-launching-command-center [accessed 2026-04-17]
7. GE HealthCare collaborates with two major medical systems to advance AI technology designed to transform hospital operations and improve patient care. GE HealthCare. 2025 Oct 20. URL: https://www.gehealthcare.com/about/newsroom/press-releases/ge-healthcare-collaborates-with-two-major-medical-systems-to-advance-ai-technology-designed-to-transform-hospital-operations-and-improve-patient-care [accessed 2026-04-17]
8. Halm B. Researchers address health care provider burnout through digital twin systems. Virginia Tech News. 2025 May 5. URL: https://news.vt.edu/articles/2025/04/digital-twins-hcprovider-burnout.html [accessed 2026-04-17]

Edited by KA Clegg; submitted 13.Apr.2026; this is a non–peer-reviewed article;accepted 13.Apr.2026; published 24.Apr.2026.

Please cite as:
Crawford M
How Digital Twins Can Improve Health System Operations
J Med Internet Res 2026;28:e98113
URL: https://www.jmir.org/2026/1/e98113
doi:10.2196/98113

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Apr.2026.

XSL•FO
RenderX

CRISPR Diagnostics, in Your Pocket

Cliff Dominy

(J Med Internet Res 2026;28:e98572) doi:10.2196/98572

KEYWORDS

CRISPR-Cas systems; microfluidic; point-of-care testing; fluorescence; biosensing techniques; nucleic acid amplification techniques; disease surveillance

Textbox .

Key Takeaways

- CRISPR-on-a-chip platforms combine nanoscale engineering with microfluidic chemistry to develop portable, high-sensitivity diagnostic assays.
- Different assays, including fluorescent reporting systems and graphene-based sensors, have distinct but potentially complementary strengths.

When it comes to clinical diagnostic devices, three variables must be carefully considered—speed, cost, and sensitivity. The catch: you can only pick two [1]. Now, a new generation of artificial intelligence (AI)–powered technology could allow doctors of the future to have it all—fast, affordable, sensitive devices right at the patient's bedside. Thanks to CRISPR-Cas technology and microfluidic engineering, the concept is beginning to look more like science and less like fiction. It's called CRISPR-on-a-chip—and yes, it could possibly talk to your smartphone.

To understand what this could mean in practice, imagine a patient of the future checking into their doctor's office. They press their finger onto a disposable cartridge attached to a tabletop reader. A few microliters of blood wicks its way through microscopic channels, where it encounters cocktails of CRISPR enzymes and guide RNAs. The guide RNAs direct the enzymes to specific genetic sequences, and the enzymes analyze the content, scanning for pathogens, genetic point mutations, and early biomarkers of cancer. Within the hour, the results populate the patient's health record and an alert is sent to their smartphone.

No lab required—just a chip, a drop of blood, and a layer of AI.

The Evolution of Point-of-Care Diagnostics

Point of care diagnostic devices, such as pregnancy tests and SARS-CoV-2 rapid antigen tests, are economical and easy to use [2]. All they require is the application of a sample onto a plastic lateral flow device. As the sample travels along the device, it interacts with antibodies and, if positive, produces a color reaction. They work well for low-sensitivity applications but tend to miss rare DNA or RNA species present in the sample that could otherwise be detected in analytical laboratories. Wouldn't it be wonderful to shrink those high-tech labs and provide accurate diagnoses with the convenience of a portable device?

Biomedical engineers are attempting to do just this by downsizing state-of-the-art lab assays and running them on a microchip [3]. To do this, they have turned to microfluidics, the science of reducing assays into minute reaction volumes. Early successes involved small plates and cartridges, but these approaches often still required bulky equipment to analyze the results [4].

True portability demands that the assay be independent of specialist equipment and the skilled technicians who operate them. Attention has now focused on running these assays in microchannels etched onto microchips. It's a tall order—the microchips would need to do it all, including analyzing the sample and reporting the results directly to a small detector, possibly a smartphone.

How does this work? How can state-of-the-art analytical equipment possibly be replicated on a microchip, some of which are as small as a fingernail?

CRISPR-on-a-Chip Unpacked

CRISPR-Cas, the Nobel Prize–winning technology associated with gene editing, has been adapted to these miniaturized microfluidic platforms [5,6]. Under the control of its guide RNA, the CRISPR-Cas complex binds to the target sequences and produces a detectable fluorescent signal.

Rachel Tinker-Kulberg, PhD, is the Director of Research and Development and Professional Track Faculty at the University of North Carolina, Greensboro. Tinker-Kulberg and colleagues work on microfluidic-based CRISPR technologies to detect and potentially suppress infectious diseases such as SARS-CoV-2 [7,8]. For Tinker-Kulberg, there have been two key advances that have made these "portable diagnostics more sensitive and feasible"—the non–polymerase chain reaction (PCR) amplification of the target DNA/RNA and an unanticipated signal boost by the activated Cas enzyme.

The signal amplification magic begins when enzymes capable of amplifying the target nucleic acid sequence recognize DNA/RNA in the sample. Typically, this kind of amplification is done in the lab using the PCR technique, which involves making many copies of the DNA by repeatedly heating and cooling the sample. In CRISPR-based systems, the trick is that

XSL•FO
RenderX

Cas enzymes get activated when they find their target, triggering a strong fluorescent signal at a steady temperature, foregoing the need for bulky thermal cycling equipment.

A second sensitivity boost is provided by an enzymatic quirk in the Cas12 and Cas13 endonucleases that are a crucial part of CRISPR diagnostic systems [9,10]. In what Tinker-Kulberg calls a "key biochemical discovery," these enzymes, once bound and activated to their targets, become nonspecific nucleases, meaning they cut DNA or RNA nonselectively, producing thousands of fluorescent signals and better detection. Tinker-Kulberg writes that this CRISPR-Cas12/13 anomaly has "allowed sequence-specific nucleic acid recognition to be converted into a strong detectable fluorescent signal," providing clinically useful levels of sensitivity.

How sensitive is it? Between 10 and 100 times more sensitive than cutting-edge PCR assays [11]. Leading CRISPR-on-a-chip technologies like SHERLOCK and DETECTR (after amplification) can detect as few as 10 particles of the SARS-CoV-2, influenza, and parainfluenza viruses in a sample under controlled laboratory conditions [12].

From Fluorescence to Hypersensitive Graphene Chips

Are you looking for just one molecule in your sample? Graphene chip technology with its zeptomolar (single molecule) sensitivity might be for you. The new platform called graphene field-effect transistor (gFET) technology currently holds much promise in the single molecule detection game [13].

gFET biosensors trade the adaptability of microfluidic chips for extreme assay sensitivity. Silicon microchips are covered with an ultrasensitive layer of carbon (graphene) just one atom thick [14]. When a target nucleic acid sequence binds to the CRISPR guide RNA, a change in electrical potential is registered and reported to the machine learning–assisted detection software. A potential advantage of gFET technology is that it has fewer moving parts than fluorescent reporters—with no amplification step or fluorescent reporter system needed.

gFET assay sensitivity is ideally suited to CRISPR diagnostics and is being deployed in the early diagnosis of cancer [15,16]. Many cancers present clinically at an advanced stage with a corresponding poor prognosis for the patient. Circulating tumor DNA (ctDNA) and microRNA (miRNA) are relatively scarce molecular tumor fingerprints that can be detected in the bloodstream shortly after tumor formation [17]. Point-of-care gFET technology can detect these pathological sequences in otherwise symptom-free individuals and, if approved, is well-positioned to replace PCR amplification in detecting early tumor formation.

The Road Ahead

Each approach is in early development and each has its unique strengths and limitations. Fluorescent reporting systems are less sensitive than gFET, but offer the flexibility of multiplexing—the ability to detect multiple fluorophores in a single sample. By contrast, the high sensitivity of the gFET approach promises greater portability but is currently limited by surface chemistry constraints. Both remain, for now, works in progress.

Which of these technologies will ultimately prevail? The answer may well be both. Tinker-Kulberg notes that "Both assays are equally valuable" for different reasons. She foresees a future that may "involve hybrid systems exploiting their individual advantages in sensitivity, speed, and integration into portable devices."

The early detection of infection, antimicrobial resistance, and cancer means earlier treatment and ultimately better patient outcomes. If these portable diagnostic devices are approved, and arrive in our clinics and homes, they could become a key component of the proactive management of our future health.

Conflicts of Interest

None declared.

References

1. Hickford ES, Dejager L, Yuill D, et al. A biomarker assay validation approach tailored to the context of use and bioanalytical platform. Bioanalysis 2023 Jul;15(13):757-771. [doi: 10.4155/bio-2023-0110] [Medline: 37526064]
2. Dinnes J, Sharma P, Berhane S, et al. Rapid, point-of-care antigen tests for diagnosis of SARS-CoV-2 infection. Cochrane Database Syst Rev 2022 Jul 22;7(7):CD013705. [doi: 10.1002/14651858.CD013705.pub3] [Medline: 35866452]
3. Kaushik AM, Hsieh K, Wang TH. Droplet microfluidics for high-sensitivity and high-throughput detection and screening of disease biomarkers. Wiley Interdiscip Rev Nanomed Nanobiotechnol 2018 Nov;10(6):e1522. [doi: 10.1002/wnan.1522] [Medline: 29797414]
4. Burkin KM, Ivanov AV, Zherdev AV, Dzantiev BB, Safenkova IV. A critical study on DNA probes attached to microplate for CRISPR/Cas12 trans-cleavage activity. Biosensors (Basel) 2023 Aug 17;13(8):824. [doi: 10.3390/bios13080824] [Medline: 37622910]
5. Welch NL, Zhu M, Hua C, et al. Multiplexed CRISPR-based microfluidic platform for clinical testing of respiratory viruses and identification of SARS-CoV-2 variants. Nat Med 2022 May;28(5):1083-1094. [doi: 10.1038/s41591-022-01734-1] [Medline: 35130561]
6. Atceken N, Kahya A, Yigci D, Tasoglu S. CRISPR-on-chip for point-of-care diagnostics. ACS Nano 2026 Jan 27;20(3):2561-2577. [doi: 10.1021/acsnano.5c19771] [Medline: 41527500]

XSL•FO
RenderX

7. Bagchi R, Tinker-Kulberg R, Salehin M, et al. Polyvalent guide RNAs for CRISPR antivirals. iScience 2022 Nov 18;25(11):105333. [doi: 10.1016/j.isci.2022.105333] [Medline: 36325075]
8. Magdy M, Tinker-Kulberg R, Josephs EA. Polyvalent guide RNAs enhance the CRISPR-mediated suppression of a human coronavirus. ACS Synth Biol 2026 Feb 19. [doi: 10.1021/acssynbio.5c00574] [Medline: 41712528]
9. Chen JS, Ma E, Harrington LB, et al. CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. Science 2018 Apr 27;360(6387):436-439. [doi: 10.1126/science.aar6245] [Medline: 29449511]
10. Gootenberg JS, Abudayyeh OO, Lee JW, et al. Nucleic acid detection with CRISPR-Cas13a/C2c2. Science 2017 Apr 28;356(6336):438-442. [doi: 10.1126/science.aam9321] [Medline: 28408723]
11. Dong J, Feng W, Lin M, et al. Comparative evaluation of PCR-based, LAMP and RPA-CRISPR/Cas12a assays for the rapid detection of Diaporthe aspalathi. Int J Mol Sci 2024 May 26;25(11):5773. [doi: 10.3390/ijms25115773] [Medline: 38891961]
12. Li X, Guo J, Yang H, Wu Y, Xie Z, Li D. A CRISPR-assisted passive microfluidic chip for rapid, visual detection of multiple respiratory viruses. Sci Rep 2026;16(1). [doi: 10.1038/s41598-025-31658-y]
13. Aran K, Goldsmith B, Moarefian M. Applications of graphene field effect biosensors for biological sensing. Adv Biochem Eng Biotechnol 2024;187:37-70. [doi: 10.1007/10_2024_252] [Medline: 38418581]
14. Balderston S, Taulbee JJ, Celaya E, et al. Discrimination of single-point mutations in unamplified genomic DNA via Cas9 immobilized on a graphene field-effect transistor. Nat Biomed Eng 2021 Jul;5(7):713-725. [doi: 10.1038/s41551-021-00706-z] [Medline: 33820980]
15. Di Carlo E, Sorrentino C. State of the art CRISPR-based strategies for cancer diagnostics and treatment. Biomark Res 2024 Dec 18;12(1):156. [doi: 10.1186/s40364-024-00701-x] [Medline: 39696697]
16. Slattery JR, Naung NY, Kalinna BH, Pal M. CRISPR-powered liquid biopsies in cancer diagnostics. Cells 2025 Oct 1;14(19):1539. [doi: 10.3390/cells14191539] [Medline: 41090767]
17. Wang L, Wen X, Yang Y, et al. CRISPR/Cas13a-based supersensitive circulating tumor DNA assay for detecting EGFR mutations in plasma. Commun Biol 2024;7(1):657. [doi: 10.1038/s42003-024-06368-2]

Edited by KA Clegg; submitted 16.Apr.2026; this is a non–peer-reviewed article;accepted 16.Apr.2026; published 28.Apr.2026.

Please cite as:
Dominy C
CRISPR Diagnostics, in Your Pocket
J Med Internet Res 2026;28:e98572
URL: https://www.jmir.org/2026/1/e98572
doi:10.2196/98572

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Apr.2026.

XSL•FO
RenderX

Big Tech and the Rise of Consumer-Facing Health AI Assistants

Tejas S Athni

(J Med Internet Res 2026;28:e99230) doi:10.2196/99230

KEYWORDS

consumer-facing health AI assistants; consumer digital health; decentralized health care; Big Tech

Textbox .

Key Takeaways

- Big Tech is rapidly moving into health care artificial intelligence (AI) and launching consumer-facing health AI assistants, marking a shift from enterprise-only AI solutions.
- Currently available assistants differ in scope and strategy, from information and interpretation, to care navigation and orchestration, to incorporation of clinician oversight.
- Consumer-facing health AI assistants can decentralize health care, expand access, and lower costs but raise safety concerns.

Tejas S Athni is an MD-PhD candidate at Harvard Medical School. In this article, he provides one of the first overviews of 5 currently available Big Tech consumer-facing health artificial intelligence (AI) assistants, outlining their services and capabilities, access models, Health Insurance Portability and Accountability Act (HIPAA) compliance, and competitive advantages.

Artificial intelligence (AI) has rapidly taken its place as a central pillar of health care innovation [1]. While early AI innovations in health care have focused on enterprise tools, such as workflow automation and clinician decision support [2], an emerging segment is taking form: consumer-facing AI health assistants.

The Landscape

Amidst rising competition, numerous technology giants have launched or announced consumer-facing health AI assistants. The earliest versions were announced in 2025, followed by a rapid wave of launches: Alphabet subsidiary Verily with Verily Me in late 2025 [3]; Amazon with its One Medical Health AI assistant [4,5], followed by OpenAI with ChatGPT Health and Anthropic with Claude for Healthcare in January 2026; and Microsoft with Copilot Health in March 2026.

These platforms fundamentally reshape how health information is delivered, with the potential to expand access to rural consumers across geographic barriers [6], cut patient wait times and improve patient engagement [7], reduce emergency department and hospital visits [8], and lower health care costs [9]. Yet, they raise concerns about misdiagnosis, overreliance, and patient anxiety via “hypochondria spirals” [10]. Large language models also often fail to redirect problematic questions [11], potentially increasing follow-up burden on providers.

Recognizing both the promise and the perils of these systems, the Centers for Medicare & Medicaid Services (CMS) has launched initiatives encouraging responsible deployment of these “conversational AI assistants” [12]. Several major technology firms have signed voluntary pledges with CMS to ensure safety and transparency [12]. The products offered by Big Tech firms, given their scale of deployment, may have outsized implications in the consumer-facing health AI assistant space.

OpenAI: ChatGPT Health

OpenAI’s ChatGPT Health [13] grafts a health AI layer onto their widely used ChatGPT platform, allowing users to connect medical records and wellness app data to create a personalized health workspace. Users can ask general health questions, prepare for doctor visits, track health metrics, and interpret lab results. Beyond basic question-answer functionality, the platform supports longitudinal health tracking and postvisit explanations with multimodal data ingestion (eg, physician clinic visit PDFs, imaging reports, and medication lists). ChatGPT Health is free to access and does not require purchase of a subscription tier, which lowers barriers to use.

Although not marketed as Health Insurance Portability and Accountability Act (HIPAA) compliant, OpenAI emphasizes that ChatGPT Health data are stored in a separate encrypted environment and not used to train public models. Hundreds of millions of weekly users already rely on ChatGPT for everyday information, making it the most widely used consumer AI platform; this large user base can be enfolded into the health AI platform—a significant competitive advantage.

Anthropic: Claude for Healthcare

Anthropic’s Claude for Healthcare [14] reflects a safety-focused approach grounded in the firm’s constitutional AI principles. Users can connect personal health records, lab results, and wearable data to the platform, which outputs personalized summaries and can perform pattern detection. The product has capabilities in longitudinal trend analysis, medication reconciliation support, and helping consumers prepare questions for health care visits. It offers opt-in integration with Apple Health and Android Health Connect, enabling aggregation of

XSL•FO
RenderX

wearable data across devices. Access is bundled within the Claude Pro and Max subscription tiers.

The direct-to-consumer product is not fully HIPAA-covered, despite the firm's marketing of their enterprise-facing product as HIPAA-ready. Claude for Healthcare's competitive advantage derives from its ground-up safety design and constitutional AI framework, which affords the firm greater credibility with consumers by indexing more heavily on algorithmic caution, conservative medical guidance, and uncertainty disclaimers.

Microsoft: Copilot Health

Microsoft's Copilot Health [15], though primarily enterprise-oriented, has a consumer-facing health AI assistant layered onto their flagship AI platform, where users can upload medical records and wearable device data (eg, Apple Watch, Fitbit, and Oura). The platform can interpret lab results, spot trends, and summarize medical information. It integrates citations and expert-written answer cards from reputable sources, such as Harvard Health, providing additional credibility. Notably, Copilot Health is linked with provider directories that enable users to search for clinicians based on location, insurance coverage, and language preferences. This transforms the tool from a purely informational assistant into a care navigation platform. At present, Copilot Health is available through waitlist, but the company plans to transition toward a subscription-based model.

The product is not marketed as a standalone HIPAA-compliant tool, though Microsoft prioritizes secure data governance, including International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC) 42001 third-party verification. Its competitive advantage lies in its embedding within the broader Microsoft ecosystem, enabling interoperability for the user. The product can also leverage Microsoft's credibility in previous health AI innovations, like its multimodal diagnostic orchestrator (MAI-DxO), which achieved 80% accuracy on complex diagnostic cases from the *New England Journal of Medicine*, compared to 20% among experienced physician panels [16].

Amazon: One Medical Health AI

Amazon's One Medical Health AI [4,17] reflects the company's strategy to create an end-to-end ecosystem for health care. The product uses a consumer's medical records to answer health questions, explain lab results, and provide summaries. It is positioned as a care orchestration layer, moving beyond dispensing advice to actively facilitating next steps, including escalation to human providers when appropriate. Users can book appointments, connect with One Medical providers through direct message care for 30+ common conditions, and coordinate medication renewals through Amazon Pharmacy. Access is bundled into One Medical membership, with additional per-visit payment options and enhanced pricing for Amazon Prime subscribers.

One Medical Health AI is marketed as HIPAA-compliant, with encrypted conversations and strict access controls. The product's competitive advantage lies in its action-oriented design and integration with Amazon Pharmacy. Amazon's acquisition of One Medical further provides the firm access to 200+ brick-and-mortar primary care clinics [18], creating a pipeline from digital triage to in-person clinic visit.

Alphabet/Verily: Verily Me

The Verily Me platform [3,19] aggregates medical records, supports longitudinal health tracking, enables photo meal logging for nutrition tracking, and includes a 24/7 AI assistant ("Violet") that provides symptom context and guidance on when to seek care. Uniquely, licensed providers review health records and deliver personalized recommendations based on a user's medical history. Beyond question-answer functionality, the platform can recommend preventive screening and flag emerging health risks based on user health data. Its Lightpath virtual care program for chronic disease management (eg, diabetes, hypertension, and hyperlipidemia) integrates personalized behavior change support with clinician oversight. It is free to access and requires no subscription.

The product is marketed as HIPAA-compliant, with strong identity verification protocols and integration with health information exchanges. Verily Me's competitive advantage lies in its distinct hybrid AI-clinician model: automated insights are paired with reviews by licensed providers. This positions the platform closer to being a care delivery platform rather than a purely informational AI assistant.

XSL•FO
RenderX

Figure WL1. Summary of currently available Big Tech health chatbots.

TOOL	SERVICES AND CAPABILITIES	ACCESS	HIPAA COMPLIANT?	COMPETITIVE ADVANTAGE(S)	SAFETY FEATURES*
ChatGPT Health *OpenAI*	Connect medical records and wellness app data; health question-answer, trend tracking, interpretation	Free, does not require subscription	No	Large consumer base already established	Health data stored in separate environment, not used to train public models
Claude for Healthcare *Anthropic*	Connect medical records and wearable data; question-answer, trend tracking, interpretation	Bundled with Claude Pro and Max subscription tiers	No	Credibility with consumer base due to focus on safety, Constitutional AI framework	Safety-first design; algorithmic caution, conservative guidance, uncertainty disclaimers, patient data not used to train models
Copilot Health *Microsoft*	Connect medical records and wearable data; question-answer, trend tracking, interpretation, care navigation (provider directories)	Waitlist; planned subscription model	No	Built-in interoperability within Microsoft ecosystem; credibility through previous health AI innovations	Secure data governance through ISO/IEC 42001 third-party verification, expert-written answer cards from reputable sources
One Medical Health AI *Amazon*	Connect medical records and wearable data; question-answer, trend tracking, interpretation, care orchestration (direct message care, appointment booking, medication renewals)	Bundled into One Medical membership; per-visit pricing options; reductions for Prime members	Yes	Clinician access, Amazon Pharmacy integration, pipeline to 200+ brick-and-mortar clinics	Conversations encrypted, strict access controls, clinician escalation
Verily Me *Alphabet/ Verily*	Connect medical records and wearable data; question-answer, trend tracking, interpretation, photo meal logging, review by licensed providers, Lightpath personalized behavior-change platform	Free, does not require subscription	Yes	Hybrid AI-clinician model, Lightpath	Insights paired with review by licensed providers, strong identity verification protocols, integration with Health Information Exchanges

Created with JMIR Correspondent Tejas S. Athni, MS.
Accurate at time of publication.
*Safety features reported by developers; independent evaluation is currently limited. Potential safety risks (e.g., hallucinations, misdiagnosis, overdiagnosis, underdiagnosis, privacy) and broader implications are evolving and under study.

Copyright © 2026, JMIR Publications. All rights reserved.

An Industry Shift

Major technology companies are racing to expand their health care footprint and build platforms that can provide personalized and accessible health guidance to consumers. The rise of these tools, which range from fully automated assistants to integrated care navigation or care orchestration platforms, marks a shift in health care delivery, though significant privacy and safety concerns remain [20]. Ultimately, these platforms will redefine how individuals interact with health care in a broader structural trend toward personalized, decentralized digital health.

Conflicts of Interest

None declared.

References

1. AI in healthcare market size & share. Grand View Research. URL: https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-healthcare-market [accessed 2026-04-12]
2. Adoption of artificial intelligence in healthcare delivery systems: early applications and impacts. Peterson Health Technology Institute. 2025 Mar 25. URL: https://phti.org/ai-adoption-early-applications-impacts/ [accessed 2026-04-27]
3. Verily launches a new consumer health app, Verily Me, at HLTH USA 2025. Verily. 2025 Oct 17. URL: https://verily.com/perspectives/verily-launches-a-new-consumer-health-app-verily-me-at-hlth-usa-2025 [accessed 2026-04-12]
4. Amazon One Medical introduces agentic health AI assistant for simpler, personalized, and more actionable health care. Amazon News. 2026 Jan 21. URL: https://www.aboutamazon.com/news/retail/one-medical-ai-health-assistant [accessed 2026-04-12]
5. Landi H. Amazon launches health AI agent on its website, expands free virtual care to 200M Prime members. Fierce Healthcare. 2026 Mar 13. URL: https://www.fiercehealthcare.com/ai-and-machine-learning/amazon-launches-health-ai-assistant-its-website-expands-free-virtual-care [accessed 2026-04-12]
6. Mahajan A, Powell D. Transforming healthcare delivery with conversational AI platforms. NPJ Digit Med 2025 Sep 30;8(1):581. [doi: 10.1038/s41746-025-01968-6] [Medline: 41028209]
7. Wah JNK. Revolutionizing e-health: the transformative role of AI-powered hybrid chatbots in healthcare solutions. Front Public Health 2025 Feb 13;13:1530799. [doi: 10.3389/fpubh.2025.1530799] [Medline: 40017541]

XSL•FO
RenderX

8. Huang MY, Weng CS, Kuo HL, Su YC. Using a chatbot to reduce emergency department visits and unscheduled hospitalizations among patients with gynecologic malignancies during chemotherapy: a retrospective cohort study. Heliyon 2023 May 1;9(5):e15798. [doi: 10.1016/j.heliyon.2023.e15798] [Medline: 37206031]
9. Anjum K, Sameer M, Kumar S. AI enabled NLP based text to text medical chatbot. Presented at: 2023 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM); Feb 22-24, 2023; Uttar Pradesh, India p. 1-5. [doi: 10.1109/ICIPTM57143.2023.10117966]
10. Landymore F. ChatGPT is sending people into obsessive spirals of hypochondria. Futurism. 2026 Apr 8. URL: https://futurism.com/artificial-intelligence/chatgpt-hypochondria [accessed 2026-04-12]
11. Sambara S, Pu Y, Ali A, Mishra V, Wong L, Agrawal M. MedRedFlag: investigating how LLMs redirect misconceptions in real-world health communication. arXiv. Preprint posted online on Jan 14, 2026. [doi: 10.48550/arXiv.2601.09853]
12. Patient facing apps — conversational AI assistants. Centers for Medicare & Medicaid Services. URL: https://www.cms.gov/health-tech-ecosystem/early-adopters/conversational-ai-assistants [accessed 2026-04-12]
13. Introducing ChatGPT Health. OpenAI. 2026 Jan 7. URL: https://openai.com/index/introducing-chatgpt-health/ [accessed 2026-04-12]
14. Advancing Claude in healthcare and the life sciences. Anthropic. 2026 Jan 11. URL: https://www.anthropic.com/news/healthcare-life-sciences [accessed 2026-04-12]
15. Introducing Copilot Health. Microsoft AI. 2026 Mar 12. URL: https://microsoft.ai/news/introducing-copilot-health/ [accessed 2026-04-12]
16. Nori H, Daswani M, Kelly C, et al. Sequential diagnosis with language models. arXiv. Preprint posted online on Jun 27, 2025. [doi: 10.48550/arXiv.2506.22405]
17. Diaz N. Amazon expands Health AI assistant. Becker's Hospital Review | Healthcare News & Analysis. 2026 Mar 11. URL: https://www.beckershospitalreview.com/healthcare-information-technology/ai/amazon-expands-health-ai-assistant/ [accessed 2026-04-12]
18. 5 things to know about Amazon's recent One Medical acquisition. American Hospital Association. URL: https://www.aha.org/aha-center-health-innovation-market-scan/2023-03-07-5-things-know-about-amazons-recent-one-medical-acquisition [accessed 2026-04-12]
19. Verily Me. URL: https://verilyme.com/ [accessed 2026-04-12]
20. Chen BX, Rosenbluth T. A.I. chatbots want your health records. Tread carefully. The New York Times. 2026 Mar 12. URL: https://www.nytimes.com/2026/03/12/technology/personaltech/microsoft-copilot-health-ai-chatbots.html [accessed 2026-04-12]

Edited by KA Clegg; submitted 23.Apr.2026; this is a non–peer-reviewed article;accepted 23.Apr.2026; published 30.Apr.2026.

Please cite as:
Athni TS
Big Tech and the Rise of Consumer-Facing Health AI Assistants
J Med Internet Res 2026;28:e99230
URL: https://www.jmir.org/2026/1/e99230
doi:10.2196/99230

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Apr.2026.

XSL•FO
RenderX

Keeping Clinical Trials on an Inclusive Track

Virginia Gewin

(J Med Internet Res 2026;28:e99011) doi:10.2196/99011

KEYWORDS

clinical trials; patient recruitment; health equity; health care disparities; ethics; patient advocacy; funding, government; digital health

Textbox .

Key Takeaways

- Experts advocate a "participation-to-prevalence" ratio—the representation of a specific population relative to their representation in a disease population—of 0.8 to 1.2 to ensure adequate representation in clinical trials.
- Digital tools can help reach target trial participants, but they also pose privacy and ethical challenges.
- Working with patient advocacy groups from the outset can help shape research priorities, improve clinical trial design, and increase health impact.

In March 2025, Dinushika Mohottige, MD, MPH, a clinician at the Icahn School of Medicine at Mount Sinai in New York City, was notified that her US National Institutes of Health grant to study kidney disease in Black individuals [1] was terminated and $194,000 in funding was rescinded. The grant, she was told via email, was "no longer consistent with current policy priorities," as outlined in the "Ending radical and wasteful government DEI [diversity, equity, and inclusion] programs and preferencing" White House Executive Order issued on January 20, 2025 [2].

The grant aimed to study why Black individuals consistently have 2‑4 times higher rates of end-stage kidney disease compared to White counterparts—and, importantly, to craft patient-centered clinical interventions to mitigate adverse health outcomes. "Genetic risk is not racial. However, we know that the prevalence of this particular high-risk allele...disproportionately burdens people who self-identify as having African ancestry," explains Mohottige.

While Mohottige's clinical trial funding was restored 5 months later, following a successful lawsuit, the delay postponed patient recruitment for the trial by roughly 10 months. Over that time, at least 5 potential participants became seriously ill, were hospitalized, or passed away.

Mohottige was hardly alone. In 2025, 383 clinical trials were disrupted. Funding terminations impacted roughly 1 in 30 clinical trials—totaling more than 74,000 clinical participants, according to a January 2026 study [3]. Trials that focused on infectious diseases, prevention, and behavioral interventions were disproportionately impacted.

The trial disruptions, many fear, threaten efforts to improve the lack of diversity in clinical trials. As it stands, fewer than 20% of drug trials conducted between 2014 and 2021 had data relevant to Black patients' benefits or side effects, according to a 2022 study [4].

Clinical trials can be funded by the pharmaceutical industry or the federal government. The goal, especially for pharmaceutical companies, is to get a clinical trial done as quickly as possible, says Sonia Anand, MD, PhD, FRCPC, director of the Chanchlani Research Center at McMaster University in Hamilton, Ontario, which is focused on health equity research. "As a result, many work with the same investigators at the same clinical sites with a proven track record. And by doing that, they have a relatively homogenous patient population," says Anand.

The bottom line is that it costs more money up front to make the effort to reach a diverse population. "We have to get over the short-term costs, and think of the long-term impacts," says Anand. In the wake of the Trump administration's targeted funding cuts, clinicians say efforts to achieve diverse, inclusive patient recruitment into US clinical trials has been dealt a blow, but two approaches are increasingly being adopted: digital tools and patient advocacy groups.

Structural Barriers

Underrepresentation in clinical trials spans many factors including race, age, language barriers, disability, socioeconomic status, geography, culture, and health literacy, describes Roxana Mehran, MD, a cardiology specialist at the Icahn School of Medicine at Mount Sinai. "The problem is not a lack of willingness to participate, but rather structural barriers that limit access to trials," says Mehran.

Underrepresentation of people from minority ethnic groups, for example, can lead to trial results that fail to characterize the efficacy and safety profile of the drug or intervention being tested, according to a study by Anand and colleagues [5]. Lack of representation comes with a hefty cost. Limited access to effective medical interventions or skewed trial results could cost hundreds of billions of dollars in reduced life expectancy and disabilities over the next 25 years.

Equitable clinical trials must "match the demographics of the disease burden under study. However, we remain far from achieving this goal," according to a 2022 National Academies report [6]. Anand and others advocate for a

XSL•FO
RenderX

"participation-to-prevalence" ratio—the representation of a specific population relative to their representation in a disease population—of 0.8 to 1.2.

In December 2025, the US Food and Drug Administration released long-awaited guidance to enhance participation in clinical trials [7]. The document emphasized the use of digital tools to lessen the burden on participants and increased public outreach, including through patient advocacy groups.

Digital Tools

Digital tools to enhance diverse recruitment of trial participants include electronic health records (EHRs), social media, research registries, and mobile apps using culturally tailored messaging to improve visibility and engagement [8]. Social media is also used to target recruitment advertising, while phone apps can make it easier to manage clinical trial consent and appointments.

"Social media and digital outreach are mainstream," says Rebecca Johnson, PhD, a clinical research trial recruitment strategist. "Recruitment providers feed lifestyle, behavior, consumer data, and other variables into models to precisely reach the right patients," she says. "What's often missing is the connectivity at the provider level."

Machine learning and artificial intelligence are increasingly being used to mine EHRs to search for the demographic and pre-existing health conditions of ideal participants [9]. It can also help identify individuals that are most likely to respond to treatment.

However, digital tools come with their own challenges including privacy concerns, digital literacy gaps, and ethical considerations.

"While we think AI digital apps will close the health equity gap, it actually has a great potential to widen it," says Anand. For example, some equity-deserving groups don't have access to cell phones. According to Anand, however, a greater concern is that surveys on the use of digital tools in clinical trial recruitment have revealed a lack of trust among women and non-White individuals. "They are worried about how the data might be used, so there's a lower uptake of these tools," says Anand.

Patient Advocacy Groups

Another strategy is to work with patient advocacy groups to build trust among communities that could benefit most from health interventions—and to identify potential trial participants. For example, Mohottige painstakingly assembled a patient board to guide her research effort, reflecting a growing trend to seek out community partnerships to inform research so that it has the best chance to make a positive real-world impact.

Patient engagement is essential, says Mehran. "Patients should be included in advisory roles or even as co-investigators, and trial materials need to be culturally appropriate and available in multiple languages," she says. Other efforts to reduce logistical barriers, such as offering flexible scheduling and transportation support and simplifying participation requirements, can also significantly improve enrollment. "Trials should go beyond large academic centers and include community-based and diverse settings to better reflect real-world populations," says Mehran.

"Working with patient advocacy groups raises credibility," says Johnson. It can also shape research priorities and improve trial design. That said, it doesn't consistently lead to higher enrollment, so it's important to set expectations up front and be clear on shared goals.

"I think patient advocates need to come in from the very beginning of the clinical trial design and be involved in conversations from step one," says Robert Sanchez, a patient advocate with kidney disease who works with Mohottige. "Numbers need to have a soul," he says.

Conflicts of Interest

None declared.

References

1. Structural racism as a "third hit" on kidney outcomes of Black individuals with APOL1 risk alleles. National Institutes of Health, RePORT. URL: https://reporter.nih.gov/project-details/10866289 [accessed 2026-04-15]
2. Ending radical and wasteful government DEI programs and preferencing. The White House. 2025. URL: https://www.whitehouse.gov/presidential-actions/2025/01/ending-radical-and-wasteful-government-dei-programs-and-preferencing/ [accessed 2026-04-27]
3. Patel VR, Liu M, Jena AB. Clinical trials affected by research grant terminations at the National Institutes of Health. JAMA Intern Med 2026 Jan 1;186(1):126-128. [doi: 10.1001/jamainternmed.2025.6088] [Medline: 41247710]
4. Green AK, Trivedi N, Hsu JJ, Yu NL, Bach PB, Chimonas S. Despite the FDA's five-year plan, Black patients remain inadequately represented in clinical trials for drugs: study examines FDA's five-year action plan aimed at improving diversity in and transparency of pivotal clinical trials for newly-approved drugs. Health Aff 2022;41(3):368-374. [doi: 10.1377/hlthaff.2021.01432] [Medline: 35254926]
5. Anand SS, Bosch J, Mehran R, Mehta SR, Patel MR. Designing inclusive clinical trials: how researchers can drive change to improve diversity. BMJ 2025 Mar 11;388:e082485. [doi: 10.1136/bmj-2024-082485] [Medline: 40068858]
6. National Academies of Sciences, Engineering, and Medicine; Policy and Global Affairs; Committee on Women in Science, Engineering, and Medicine; Committee on Improving the Representation of Women and Underrepresented Minorities in

XSL•FO
RenderX

Clinical Trials and Research. In: Bibbins-Domingo K, Helman A, editors. Improving Representation in Clinical Trials and Research: Building Research Equity for Women and Underrepresented Groups: National Academies Press (US); 2022. [doi: 10.17226/26479] [Medline: 36137057]
7. Enhancing participation in clinical trials—eligibility criteria, enrollment practices, and trial designs: guidance for industry revision 1. US Department of Health and Human Services; Food and Drug Administration; Center for Drug Evaluation and Research (CDER); Center for Biologics Evaluation and Research (CBER). 2025. URL: https://www.fda.gov/media/190162/download [accessed 2026-04-29]
8. Tomiwa T, Wong E, Miller HN, et al. Leveraging digital tools to enhance diversity and inclusion in clinical trial recruitment. Front Public Health 2024;12:1483367. [doi: 10.3389/fpubh.2024.1483367] [Medline: 39529717]
9. Kasahara A, Mitchell J, Yang J, Cuomo RE, McMann TJ, Mackey TK. Digital technologies used in clinical trial recruitment and enrollment including application to trial diversity and inclusion: a systematic review. Digit Health 2024;10:20552076241242390. [doi: 10.1177/20552076241242390] [Medline: 38559578]

Edited by KA Clegg; submitted 21.Apr.2026; this is a non–peer-reviewed article;accepted 21.Apr.2026; published 29.Apr.2026.

Please cite as:
Gewin V
Keeping Clinical Trials on an Inclusive Track
J Med Internet Res 2026;28:e99011
URL: https://www.jmir.org/2026/1/e99011
doi:10.2196/99011

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Apr.2026.

XSL•FO
RenderX

Immortal AI, Mortal Life: Long-Term Perspectives on AI and Human Knowledge

Hyunjin Shim

(J Med Internet Res 2026;28:e98707) doi:10.2196/98707

KEYWORDS

artificial intelligence; generative AI; evolution; education; research

Textbox .

Key Takeaways

- The evolution of artificial intelligence is not subject to the same limitations as the evolution of human intelligence, with significant implications for knowledge and education.
- Researchers should be cognizant of the potential for knowledge monoculture and the diversion of resources and focus away from core issues.
- Educators should consider focusing on cultivating uniquely human capacities that complement or supplement artificial intelligence.

Hyunjin Shim, PhD, is an assistant professor at California State University, Fresno, where her lab researches and develops next-generation infection diagnostics and therapeutics, with the long-term aim of combating antimicrobial resistance. Her background is in biotechnology, bioengineering, and innovative genomics. In this piece, she explores some of the higher-level risks and potential implications of artificial intelligence for our knowledge systems through her lenses as a biologist, researcher, and educator.

As a biologist, witnessing the evolution rate of artificial intelligence (AI) is remarkable. Since the conception of AI in the 1950s, AI research was revived and accelerated in the early 21st century by the development of deep learning, thanks to advances in hardware and access to big data [1]. Since then, we have been experiencing the penetration of AI or AI-related technologies in almost every aspect of our intellectual lives.

This evolution rate is unprecedented in the realm of any scientific and technological field, particularly biology. Human evolution, as for any organic life form, is driven by powerful evolutionary forces but ultimately limited by several factors, from mutation rates and generation times to environmental stability and the speed of natural selection [2]. While the evolution of AI is also subject to limits—for example, physical, computational, and data limits—it enables knowledge storage and generation that can be more rapid and persistent than human understanding. Human knowledge, by contrast, depends on ongoing education, with individual comprehension in a sense “resetting” generationally.

This asymmetry underscores the need to take a longer-term perspective in examining the implications of AI’s role in, and impact on, human knowledge generation and education.

AI and Human Knowledge

As a researcher, the focus of human intelligence has recently shifted from volume to depth, given the superiority of AI in broad knowledge access and processing speed. Since the generative AI boom in 2019, a positive feedback loop has developed between the accumulation of AI and human intelligence: the training datasets of deep learning methods have been annotated from human activities, and AI has been enhancing the advances of human knowledge through cognitive extension [3].

The key purported benefits of AI are accelerated research and discovery. In biology, the advances of deep learning began to influence various research areas as early as 2015, including sequencing technologies [4-6], protein engineering [7,8], pathogen evolution [9,10], and biomedical images [11,12]. AI has also shown unprecedented capacity in integrating information and predicting outcomes in some scientific domains, such as predicting protein structures using the AI system AlphaFold [13,14]. However, AI advances have not solved all critical problems in scientific research and have created new challenges.

For some critical issues, AI integration diverts attention and investment from solving the fundamental problem to chasing shiny tools. For example, the spread of antimicrobial resistance—an urgent public health issue—is making common infections untreatable and impeding modern medicine [15]. While the antibiotic development pipeline suffers from the lack of a viable market for newly developed antibiotics [16], a more serious issue is bacteria rapidly developing resistance to small-molecule antibiotics [17]. Yet, fuelled partially by AI hype, most research studies on antibiotics focus on high-throughput screening or de novo design of small molecules [18,19]. The attention of the scientific community and industry has been diverted away from investing resources into higher-risk—and potentially higher reward—research developing entirely new strategies to tackle antimicrobial resistance [16].

Perhaps as concerning is the potential stifling of creativity and innovation. Generative AI output is fundamentally based on

XSL•FO
RenderX

averaging and predicting patterns from big training data [20]; when used in research, it creates monocultures of knowing—a reduction in the diversity of thought and ideas—that can weaken the production of scientific knowledge [21]. Another concern related to utilizing AI in research is shifting the focus from solving real-world problems to maximizing output by exploiting AI's speed and breadth of knowledge [22]. The current AI bubble exists in academic settings as much as in industry settings, driven by opportunity costs and output inflation in areas of research that require intuition and strategic innovation.

AI and Human Education

The incorporation of AI into higher education is inevitable and has implications for the future of knowledge generation. Younger generations are rapidly adapting to new technologies, and attempts to prohibit AI tools in classrooms and coursework are unlikely to succeed. Unlike previous technologies such as calculators, AI intersects with nearly all aspects of learning, including writing, critical thinking, and creative production. Furthermore, the use of AI tools raises significant concerns about academic honesty and integrity, as it is more difficult to detect and regulate than other types of academic misconduct, such as plagiarism [23].

The benefits of AI in education have been widely accepted as personalized learning support, real-time feedback, and engaging with complex information. Ironically, this integration of AI in education also shifts the focus of educational practice. Some educators are reverting to analog or low-tech assessments, such as oral exams or handwritten essays, to ensure authenticity in student work, reflecting ongoing uncertainty about how to assess student learning in AI-augmented contexts [24].

Historically, higher education has been central to innovation by preparing specialists through extended periods of study. For instance, the education of a doctor often spans more than two decades. In contrast, AI systems can absorb and synthesize vast amounts of information in a fraction of that time, which prompts questions about traditional educational timelines and the skills most relevant to acquire.

Indeed, much of the content traditionally taught in the current education system can be distilled into core principles that AI systems already master efficiently. There may be extensive periods in education where traditional knowledge transfer is inefficient, hindering students from acquiring skills and competencies that complement rather than compete with AI. Education should make knowledge transfer more effective by focusing on cultivating uniquely human capacities where AI currently has limitations or requires oversight—for example, thinking outside the box, identifying core problems, and developing interpersonal skills.

Retaining What Is Human

AI can be a powerful tool for supporting human knowledge, but dependency on that tool comes with potential risks for the quality, diversity, sustainability, and appropriate deployment of that knowledge—risks that are amplified by the rapid rate at which it is evolving.

Ultimately, human-centered pathways for education and knowledge generation need to be carefully preserved alongside, and outside of, AI systems, to diversify knowledge sources and to ensure that human decisions reflect and prioritize human needs and values.

Higher education has a responsibility to ensure human intelligence remains distinct from AI and that both serve the long-term greater good for human well-being.

Conflicts of Interest

None declared.

References

1. Goodfellow I, Bengio Y, Courville A. Deep Learning: MIT Press; 2016.
2. Ao P. Laws in Darwinian evolutionary theory. Phys Life Rev 2005 Jun;2(2):117-156. [doi: 10.1016/j.plrev.2005.03.002]
3. Colther C, Doussoulin JP. Artificial intelligence: driving force in the evolution of human knowledge. Journal of Innovation & Knowledge 2024 Oct;9(4):100625. [doi: 10.1016/j.jik.2024.100625]
4. Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol 2017 Apr 11;18(1):67. [doi: 10.1186/s13059-017-1189-z] [Medline: 28395661]
5. Arango-Argoty G, Garner E, Pruden A, Heath LS, Vikesland P, Zhang L. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome 2018 Feb 1;6(1):23. [doi: 10.1186/s40168-018-0401-z] [Medline: 29391044]
6. Shim H. Futuristic methods in virus genome evolution using the third-generation DNA sequencing and artificial neural networks. In: Shapshak P, Balaji S, Kangueane P, Chiappelli F, Somboonwit C, Menezes LJ, et al, editors. Global Virology III: Virology in the 21st Century: Springer International Publishing; 2019:485-513. [doi: 10.1007/978-3-030-29022-1_17]
7. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol 2015 Aug;33(8):831-838. [doi: 10.1038/nbt.3300] [Medline: 26213851]
8. Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein engineering with sequence-based deep representation learning. Nat Methods 2019 Dec;16(12):1315-1322. [doi: 10.1038/s41592-019-0598-1] [Medline: 31636460]

XSL•FO
RenderX

9. Bartoszewicz JM, Genske U, Renard BY. Deep learning-based real-time detection of novel pathogens during sequencing. Brief Bioinform 2021 Nov 5;22(6):bbab269. [doi: 10.1093/bib/bbab269] [Medline: 34297793]
10. Shim H. Feature learning of virus genome evolution with the nucleotide skip-gram neural network. Evol Bioinform Online 2019;15:1176934318821072. [doi: 10.1177/1176934318821072] [Medline: 30692845]
11. Ahmad A, Hettiarachchi R, Khezri A, Singh Ahluwalia B, Wadduwage DN, Ahmad R. Highly sensitive quantitative phase microscopy and deep learning aided with whole genome sequencing for rapid detection of infection and antimicrobial resistance. Front Microbiol 2023;14:1154620. [doi: 10.3389/fmicb.2023.1154620] [Medline: 37125187]
12. Spahn C, Gómez-de-Mariscal E, Laine RF, et al. DeepBacs for multi-task bacterial image analysis using open-source deep learning approaches. Commun Biol 2022 Jul 9;5(1):688. [doi: 10.1038/s42003-022-03634-z] [Medline: 35810255]
13. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature New Biol 2021 Aug;596(7873):583-589. [doi: 10.1038/s41586-021-03819-2] [Medline: 34265844]
14. Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature New Biol 2024 Jun;630(8016):493-500. [doi: 10.1038/s41586-024-07487-w] [Medline: 38718835]
15. Murray CJL, Ikuta KS, Sharara F. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet 2022 Feb 12;399(10325):629-655. [doi: 10.1016/S0140-6736(21)02724-0] [Medline: 35065702]
16. Shim H. Three innovations of next-generation antibiotics: evolvability, specificity, and non-immunogenicity. Antibiotics (Basel) 2023 Jan 18;12(2):204. [doi: 10.3390/antibiotics12020204] [Medline: 36830114]
17. Butler MS, Paterson DL. Antibiotics in the clinical pipeline in October 2019. J Antibiot 2020 Jun;73(6):329-364. [doi: 10.1038/s41429-020-0291-8]
18. Krishnan A, Anahtar MN, Valeri JA, et al. A generative deep learning approach to de novo antibiotic design. Cell 2025 Oct 16;188(21):5962-5979. [doi: 10.1016/j.cell.2025.07.033] [Medline: 40816267]
19. Wong F, Zheng EJ, Valeri JA, et al. Discovery of a structural class of antibiotics with explainable deep learning. Nature New Biol 2024 Feb;626(7997):177-185. [doi: 10.1038/s41586-023-06887-8] [Medline: 38123686]
20. Shumailov I, Shumaylov Z, Zhao Y, Papernot N, Anderson R, Gal Y. AI models collapse when trained on recursively generated data. Nature New Biol 2024 Jul;631(8022):755-759. [doi: 10.1038/s41586-024-07566-y] [Medline: 39048682]
21. Messeri L, Crockett MJ. Artificial intelligence and illusions of understanding in scientific research. Nature New Biol 2024 Mar;627(8002):49-58. [doi: 10.1038/s41586-024-07146-0] [Medline: 38448693]
22. Berisha V, Krantsevich C, Hahn PR, et al. Digital medicine and the curse of dimensionality. NPJ Digit Med 2021 Oct 28;4(1):153. [doi: 10.1038/s41746-021-00521-5] [Medline: 34711924]
23. Schmidt DA, Alboloushi B, Thomas A, Magalhaes R. Integrating artificial intelligence in higher education: perceptions, challenges, and strategies for academic innovation. Computers and Education Open 2025 Dec;9:100274. [doi: 10.1016/j.caeo.2025.100274]
24. Balalle H, Pannilage S. Reassessing academic integrity in the age of AI: a systematic literature review on AI and academic integrity. Social Sciences & Humanities Open 2025;11:101299. [doi: 10.1016/j.ssaho.2025.101299]

Edited by KA Clegg; submitted 17.Apr.2026; this is a non–peer-reviewed article;accepted 17.Apr.2026; published 29.Apr.2026.

Please cite as:
Shim H
Immortal AI, Mortal Life: Long-Term Perspectives on AI and Human Knowledge
J Med Internet Res 2026;28:e98707
URL: https://www.jmir.org/2026/1/e98707
doi:10.2196/98707

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Apr.2026.

XSL•FO
RenderX

AI-Quantified Breast Arterial Calcification Can Predict Heart Disease Risk From Mammograms

Shalini Kathuria Narang

(J Med Internet Res 2026;28:e99154) doi:10.2196/99154

KEYWORDS

arterial calcification; cardiovascular diseases; artificial intelligence; mammography; risk assessment

Textbox .

Key Takeaways

- New research applying artificial intelligence to mammography quantifies breast arterial calcification and demonstrates that it is an independent risk factor for major adverse cardiovascular events, particularly for women typically below traditional screening age.
- The model enables risk assessment at scale, as well as a more precise understanding of the relationships between breast arterial calcification severity and cardiovascular risk.
- Mammography can potentially be used to screen women at risk of cardiovascular disease without additional radiation exposure, costs, doctor visits, or blood tests.

Despite being the leading cause of death in women globally, cardiovascular disease (CVD) tends to be underdiagnosed and undertreated in women [1].

Predicting CVD risk is complex. It involves integrating multiple different kinds of data—from demographic information to laboratory scores to medical history—using a tool such as the PREVENT score to quantify that risk [2,3]. Data are often incomplete, and for women, accuracy and timely detection tend to be further limited because the ways CVD can present in women specifically have historically been underrepresented and underrecognized.

A recent retrospective study of 123,762 women [4] suggests a new, efficient way to address some of these limitations and improves how women's cardiovascular risk is assessed: applying artificial intelligence to routine mammograms.

The Study

Breast arterial calcification (BAC)—a buildup of calcium in breast tissue—is an increasingly established risk factor for CVD [5-9]. Researchers analyzed the data of women from two health care systems in the United States who had undergone screening mammography—a cohort of 74,124 women aged 55.5 years on average, from Emory Healthcare, and a cohort of 49,638 women aged 59.5 years on average, from Mayo Clinic—and assessed the images and quantified the extent of BAC present.

To do this, they developed, trained, and validated a deep learning artificial intelligence model, using it to segment BAC on mammograms and automatically quantify risk severity. Where previous studies have often assessed BAC as a binary variable (presence vs absence) [10], this new model—a "transformer-based segmentation model" capable of maintaining high sensitivity while reducing false positives—permitted finer-grained risk categorization. BAC severity was categorized as zero (0 mm^2), mild (>0 - 10 mm^2), moderate (>10 - 25 mm^2), and severe (>25 mm^2).

BAC as an Independent Risk Factor

BAC was ultimately detected in 16.1% (Emory cohort) and 20.6% (Mayo cohort) of study participants. Increased BAC predicted increased cardiovascular risk over and above the PREVENT scores, confirming that it is an independent risk factor offering significant additional prognostic value.

"The model returns a continuous area score—an amount of BAC millimeter squared from zero to anything, and every millimeter squared of BAC increases risk of CVD by 1%. If you have 20 millimeters of BAC, your hazard ratio is 20% increased risk," says Hari Trivedi, MD, Associate Professor, Radiology and Biomedical Informatics at Emory School of Medicine and lead study author.

He adds that since reporting a continuous score can be "tricky" and "overwhelming in a clinical setting" and for ease in statistical analysis, the researchers grouped the numbers into mild, moderate, and severe categories. "The model returns an actual number, but we take that and put it into mild, moderate and severe categories, similar to what's done for cardiac CT [computed tomography]."

The researchers also observed a higher prevalence of antihypertensive medication use and statin use as BAC severity increased, and higher prevalence of some cardiovascular risk factors, including diabetes mellitus, higher systolic blood pressure, higher body mass index, and lower eGFR (estimated glomerular filtration rate), consistent with previous BAC studies [10,11]. Interestingly, increased BAC severity was not associated with higher smoking prevalence.

"There have been varying results on the link between BAC and smoking…Smokers have lower BAC in our study," says Trivedi.

XSL•FO
RenderX

"Smoking data is notoriously unreliable…Even if BAC is inversely related to smoking, we have already shown BAC is an independent risk factor and smoking is an independent risk factor. The nice thing about independent risk factors is it doesn't matter if they relate to each other."

Cohort Differences

In addition to the BAC prevalence differences between cohorts—likely attributable to differences in age—the study found a statistically significant association between BAC and myocardial infarction in the Mayo cohort, but not in the Emory cohort. Trivedi thinks population differences may explain this.

"All the trends that we observed with mild, moderate, severe, the hazard ratios, etc, the trends are the same (in both cohorts). Now, the degree of severity changes. At Emory, having a severe BAC score confers a 2.2 times risk of CVD. At Mayo, that number is slightly different. And that's to be expected," says Trivedi. Controllable and uncontrollable risk factors for CVD including, for example, lifestyle, existing health conditions, age, race, and ethnicity, are well-established [12]. "The model, segmentation, and score works across multiple scanners and multiple populations. Emory's population is more racially diverse. It is almost half African American, whereas Mayo's is not. Mayo has more Hispanic patients than Emory. Emory's in the low single digits for Hispanic patients. Mayo's population is older than Emory's patient population."

Earlier Screening?

The research also found the prognostic value of BAC especially noteworthy in women under 50. While BAC is more prevalent with age, its presence in younger women is a particularly strong indicator of elevated cardiovascular risk [5,13,14].

The analysis suggests that moderate to severe BAC in this younger cohort was associated with significantly lower major adverse cardiovascular event survival, identifying a high-risk group that might otherwise be overlooked by age-weighted risk models.

"Women in the United States are recommended to start screening mammography at age 40. Women under 50 are typically not considered at risk for CVD, and may not even be screened. So the fact that we see a link between BAC and CVD even in women under 50 is really important," says Trivedi.

Next Steps

Trivedi and team's model holds a great deal of promise for rapid risk assessment and early detection of CVD in women at scale. It can be readily implemented into routine mammography screenings, permitting more efficient, lower-barrier dual-purpose care. The researchers are working on continued deployment and validation.

"We are actively pursuing three international cohorts in Greece, Brazil, and the UK for evaluation, not of the model's technical performance, but of cardiovascular outcomes. We are also working on a prospective clinical deployment inside of Emory. We have a lot of demand and support from the cardiologists," says Trivedi.

Recently, the model was also integrated and deployed at the Mayo Clinic in Arizona. "Emory does 150,000 screening mammograms per year. Of those, 2% - 3% of women roughly have severe BAC—that's 3000 women a year. So part of the challenge is going to be, how are we going to handle all of this potential extra volume? We anticipate doing a clinical deployment and then measur[ing] the impact of this. We sort of technically validated the model. Now we need to validate the clinical impact of having the information," says Trivedi.

"You can screen women at scale, essentially without any cost, without going to an extra doctor's visit, without getting your lipids drawn, but this is like truly exhaustive data that comes out of an existing mammogram that 40 million women are doing a year, and you can immediately risk stratify all of them," he says. "Now that we have this tool, we can begin studying it."

Conflicts of Interest

None declared.

References

1. Women & CVD. World Heart Federation. URL: https://world-heart-federation.org/what-we-do/women-cvd/ [accessed 2026-04-27]
2. Khan SS, Matsushita K, Sang Y, et al. Development and Validation of the American Heart Association's PREVENT Equations. Circulation 2024 Feb 6;149(6):430-449. [doi: 10.1161/CIRCULATIONAHA.123.067626] [Medline: 37947085]
3. Zhou H, Zhang Y, Zhou MM, et al. Evaluation and comparison of the PREVENT and pooled cohort equations for 10-year atherosclerotic cardiovascular risk prediction. J Am Heart Assoc 2025 Feb 18;14(4):e039454. [doi: 10.1161/JAHA.124.039454] [Medline: 39921505]
4. Dapamede T, Urooj A, Joshi V, et al. Artificial intelligence-based quantification of breast arterial calcifications to predict cardiovascular morbidity and mortality. Eur Heart J 2026 Mar 9:ehag128. [doi: 10.1093/eurheartj/ehag128] [Medline: 41795899]
5. Allen TS, Bui QM, Petersen GM, et al. Automated breast arterial calcification score is associated with cardiovascular outcomes and mortality. JACC Adv 2024 Nov;3(11):101283. [doi: 10.1016/j.jacadv.2024.101283] [Medline: 39399518]

XSL•FO
RenderX

6. Hendriks EJE, de Jong PA, van der Graaf Y, Mali W, van der Schouw YT, Beulens JWJ. Breast arterial calcifications: a systematic review and meta-analysis of their determinants and their association with cardiovascular events. Atherosclerosis 2015 Mar;239(1):11-20. [doi: 10.1016/j.atherosclerosis.2014.12.035] [Medline: 25568948]
7. Iribarren C, Chandra M, Lee C, et al. Breast arterial calcification: a novel cardiovascular risk enhancer among postmenopausal women. Circ Cardiovasc Imaging 2022 Mar;15(3):e013526. [doi: 10.1161/CIRCIMAGING.121.013526] [Medline: 35290077]
8. Schnatz PF, Marakovits KA, O'Sullivan DM. The association of breast arterial calcification and coronary heart disease. Obstet Gynecol 2011 Feb;117(2 Pt 1):233-241. [doi: 10.1097/AOG.0b013e318206c8cb] [Medline: 21252734]
9. Koh TJW, Tan HJH, Ravi PRJ, et al. Association between breast arterial calcifications and cardiovascular disease: a systematic review and meta-analysis. Can J Cardiol 2023 Dec;39(12):1941-1950. [doi: 10.1016/j.cjca.2023.07.024] [Medline: 37506765]
10. Rotter MA, Schnatz PF, Currier AA Jr, O'Sullivan DM. Breast arterial calcifications (BACs) found on screening mammography and their association with cardiovascular disease. Menopause 2008;15(2):276-281. [doi: 10.1097/gme.0b013e3181405d0a] [Medline: 17917612]
11. Bae MJ, Lee SY, Kim YJ, et al. Association of breast arterial calcifications, metabolic syndrome, and the 10-year coronary heart disease risk: a cross-sectional case-control study. J Womens Health (Larchmt) 2013 Jul;22(7):625-630. [doi: 10.1089/jwh.2012.4148] [Medline: 23790228]
12. Heart disease risk factors. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/heart-disease/risk-factors/index.html [accessed 2026-04-27]
13. Loberant N, Salamon V, Carmi N, Chernihovsky A. Prevalence and degree of breast arterial calcifications on mammography: a cross-sectional analysis. J Clin Imaging Sci 2013;3:36. [doi: 10.4103/2156-7514.119013] [Medline: 24228205]
14. Daniels LB, Itchhaporia D. Breast arterial calcification as a cardiovascular risk factor: time to "bust" it out. JACC Adv 2025 Mar;4(3):101638. [doi: 10.1016/j.jacadv.2025.101638] [Medline: 39999521]

Edited by KA Clegg; submitted 22.Apr.2026; this is a non–peer-reviewed article;accepted 22.Apr.2026; published 28.Apr.2026.

Please cite as:
Narang SK
AI-Quantified Breast Arterial Calcification Can Predict Heart Disease Risk From Mammograms
J Med Internet Res 2026;28:e99154
URL: https://www.jmir.org/2026/1/e99154
doi:10.2196/99154

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Apr.2026.

XSL•FO
RenderX

Channel Allocation and Equity in Preventive Campaigns for Older Adults: Agent-Based Modeling Study

Jihye Lee[1,2], PhD; Juyoung Park[2]; Yuna Kim[2]; Duk-Jo Kong[2], Prof Dr

[1]Seoul National University, Seoul, Republic of Korea

[2]Gwangju Institute of Science and Technology, 123 Cheomdangwagi-ro, Gwangju, Republic of Korea

Corresponding Author:
Duk-Jo Kong, Prof Dr
Gwangju Institute of Science and Technology, 123 Cheomdangwagi-ro, Gwangju, Republic of Korea

Abstract

Background: Preventive campaigns for older adults must decide how to allocate limited resources across media channels. However, these channel allocation and budget decisions rarely use explicit criteria for distributional equity or structured strategic planning tools. Consequently, health systems may optimize average uptake while leaving large gaps across socioeconomic groups and media use profiles.

Objective: This study aimed to develop and apply a data-driven agent-based model as a strategic planning tool for preventive campaigns targeting older adults, comparing channel allocation, personalization, and loss framing options under explicit budget and equity guardrails.

Methods: We built an agent-based model calibrated to national survey data from South Korea on influenza vaccination and routine health screening among older adults (vaccination, N=2405; screening, N=2400). Fifteen prespecified campaign scenarios varied channel allocation across television, digital, and print media; budget intensity; 2 equity-focused personalization strategies; and graded loss framing. Primary outcomes were final adoption and time to adoption. Equity outcomes included the minimum class-level adoption and 90‑10 gap across latent classes. Each scenario was simulated over 12 monthly steps with 100 Monte Carlo replications. We conducted sensitivity analyses varying link functions and key social reinforcement parameters.

Results: Personalization improved uptake and equity relative to the integrated baseline. In the vaccination model (N=2405), adoption increased from 91.2% (n=2193) to 93.3% (n=2244) and 94.6% (n=2275). Minimum class-level adoption increased from 86.8% to 90.3% and 90.9%. The 90‑10 gap narrowed from 5.7 to 4.5 and 4.7 percentage points. In the screening model (N=2400), adoption increased from 83.8% (n=2011) to 88.2% (n=2117) and 89.5% (n=2148). Minimum class-level adoption increased from 77.6% to 83.2% and 85.3%. The 90‑10 gap narrowed from 9.2 to 7.4 and 6.2 percentage points. Television-only strategies achieved high adoption but had less favorable equity profiles than personalization. High-budget strategies achieved high adoption but required higher total exposure. Stronger loss framing produced small, monotonic gains in adoption and shortened the time to adoption without worsening equity in the tested range. Scenario rankings were stable in sensitivity analyses.

Conclusions: This agent-based modeling study illustrates how ex ante planning can improve preventive campaign design by comparing channel allocation and personalization options under explicit equity and budget criteria. For campaigns targeting older adults, equity-focused reweighting and class-tailored television-digital portfolios improved or preserved mean adoption while strengthening distributional equity under fixed budgets. In contrast, undifferentiated channel diversification without personalization offered a less favorable efficiency-equity trade-off. These findings support integrating explicit equity guardrails into early-stage channel allocation and prioritizing targeted personalization over simple channel diversification. Future work should validate these patterns in other populations and health systems and link simulated diffusion trajectories with observed exposure and engagement in real-world campaigns. It should also extend guardrail-based planning tools to organizational settings and multiyear decision contexts.

(J Med Internet Res 2026;28:e88429) doi:10.2196/88429

KEYWORDS

preventive campaigns; channel allocation; older adults; agent-based modeling; distributional equity; resource allocation; budget constraints; influenza vaccination; loss framing; strategic planning

XSL•FO
RenderX

Introduction

Background

Preventive behaviors among older adults, such as seasonal influenza vaccination and routine health screening, remain uneven across socioeconomic strata despite long-standing equity goals [1-4]. Multiple analyses in Europe and the United Kingdom report persistent or widening gaps in influenza vaccination uptake among adults aged ≥60 years by deprivation and education [1,2]. These studies also describe comparable heterogeneity in cancer screening participation and outcomes across European Union member states [3,4]. These disparities persist even where national coverage has improved, suggesting that changes in supply or eligibility alone are insufficient to close equity gaps. Vaccination and screening are related but behaviorally distinct preventive targets in older adults. Annual influenza vaccination is often an episodic, clinician-recommended decision, whereas routine health screening requires multistep planning, navigation of organizational procedures, and higher levels of health and digital literacy.

Daily life and health care for older adults are increasingly mediated by digital channels such as online news, social media, messaging apps, and patient portals [5-7]. Systematic reviews show wide variation in digital literacy, skills, and confidence among older adults, with distinct connectivity levels across user segments and pronounced social gradients in eHealth literacy [5-7]. Emerging digital health research links this digital divide to inequalities in self-management, use of online health services, and broader health outcomes among older adults [8-10]. Older adults are therefore not a single audience but a set of heterogeneous segments occupying different media and support ecologies that can amplify or dampen campaign effects.

Recent equity-focused work also emphasizes behavioral and social drivers of vaccination and screening—beliefs, norms, practical barriers, and motivation—rather than structural access alone. The World Health Organization's behavioral and social drivers of vaccination framework and scoping reviews of influenza vaccination in older adults highlight that groups with fewer resources are often missing from the evidence base and that social context strongly shapes uptake [11,12]. For campaign planners, this implies that channel choice, sequencing, and audience segmentation should consider not only reach but also how media exposure interacts with social reinforcement and practical constraints in different subgroups. These challenges point to the need for explicit digital health strategic planning tools to guide channel allocation, personalization, and the use of equity guardrails in preventive campaigns for older adults.

Current Evidence

Planning and Equity Frameworks

Implementation science models such as the PRECEDE-PROCEED model and the Reach, Effectiveness, Adoption, Implementation and Maintenance (RE-AIM)/Practical Robust Implementation and Sustainability Model (PRISM) have strengthened the understanding of reach, adoption, and sustainability in public health programs [13,14]. These models are increasingly applied to plan and evaluate interventions with an explicit focus on equity [15,16]. In parallel, equity frameworks such as PROGRESS-Plus (place of residence; race, ethnicity, culture, and language; occupation; gender and sex; religion; education; socioeconomic status; and social capital—plus additional equity-relevant factors) and related disparity indices are widely used to structure and report the distributional impacts of public health interventions in systematic reviews and policy analyses [15].

However, in most applications, these tools are applied after interventions have been implemented. Most studies focus on describing who was reached and how effects varied [1,13,17]. They rarely generate prospective allocation rules across channels or segments under budget and equity constraints. Even less often are they operationalized as ex ante decision support tools for channel portfolios.

Digital and Social Media Campaigns

Systematic reviews of digital and social media campaigns show complex exposure pathways and small but generally positive average effects on health behaviors, with substantial heterogeneity across platforms, populations, and outcomes [18-20]. However, most evaluations rely on intermediate metrics—impressions, clicks, likes, or other low-level engagement—rather than behavior change [18-20]. This reliance on proxy outcomes complicates the assessment of the added value of specific channels beyond offline activities.

As a result, planners have limited operational guidance on how to allocate budgets across channels. Key decisions include whether to concentrate resources on a single high-leverage channel (eg, television), distribute spending across multiple platforms, or adjust the mix over time as part of strategic planning.

Personalization and Equity

Personalization and targeting are widely promoted in digital health to improve relevance and engagement, including for older adults [21,22]. Conceptual and empirical reviews describe approaches ranging from simple demographic tailoring to rule-based and data-driven personalization, but their equity implications are mixed and often underdocumented [21,23]. A recurring concern is that personalization may disproportionately benefit digitally skilled or socially advantaged groups, thereby widening disparities [23,24].

Although equity-oriented frameworks encourage inclusive targeting and prioritization of underserved groups [25], there is little mechanistic evidence comparing fundamentally different personalization philosophies. Such contrasts include cautious, vulnerability-focused adjustments versus segment-optimized channel mixes evaluated under common efficiency and equity metrics. These contrasts suggest that the same personalization or channel strategy could plausibly yield different efficiency-equity profiles for vaccination and screening. However, existing planning models rarely examine these targets side by side within a unified framework.

XSL•FO
RenderX

Equity Metrics as Planning Tools

Equity metrics and frameworks have also advanced. Entropy-based indices such as Theil-type measures summarize the distribution of health outcomes across social strata and capture both between- and within-group variation [26]. In parallel, PROGRESS-Plus has become a standard for conceptualizing equity dimensions in systematic reviews and equity-relevant evaluations [25].

In many applications, equity metrics are used post hoc to describe which groups benefited [25,26]. Researchers have proposed more prescriptive approaches that incorporate distributional objectives into economic evaluation, including distributional cost-effectiveness analysis and pragmatic aggregate variants [27]. However, these methods remain uncommon in routine resource allocation practice [27].

Network and Agent-Based Modeling

Network and simulation approaches, especially agent-based models (ABMs), represent individuals as interacting agents linked through online and offline networks [28,29]. These models can explicitly capture heterogeneous exposure, social influence, and timing. Reviews describe ABMs as "policy laboratories" that allow planners to test alternative portfolios and explore system behavior before real-world deployment [28,29]. Network experiments and diffusion studies further show that many protective behaviors spread as complex contagions that require repeated, multisource reinforcement [30].

Despite this potential, most ABM applications in public health focus on infectious disease transmission or generic opinion dynamics [28]. Only a small but growing set of studies uses ABMs to examine vaccination strategies or health communication interventions [31,32]. Even in this work, models rarely consider resource constraints, channel portfolios, and equity-focused objectives for older adult vaccination or screening within a single, reproducible framework [31,32]. These gaps motivate data-driven ABM approaches that can support strategic planning for older adult vaccination and screening campaigns under explicit equity guardrails.

Research Gaps and Study Objectives

Against this backdrop, several gaps remain at the intersection of digital health campaigns, personalization, and equity for older adults:

- Budget-constrained channel portfolios: Despite detailed descriptions of older adults' digital heterogeneity and the digital divide, we found no models that explicitly link these patterns to budget-constrained channel portfolios [5-10,21,22]. Such models would show how much to invest in television, digital, or print media over time for preventive campaigns in this population.
- Contrasting personalization philosophies: The common assumption that personalization improves efficiency but threatens equity has not been tested in mechanistic models. In particular, few models compare fundamentally different personalization strategies—for example, a cautious, outcome-targeted boost for the worst-off class versus a fully segment-optimized, media affinity–based allocation—under common outcome and equity metrics.
- Target-specific behavioral profiles: Existing planning models seldom account for the fact that vaccination and screening differ in baseline uptake, decision complexity, and system dependencies in older adults. Few studies evaluate whether a given channel or personalization strategy traces the same efficiency-equity trade-offs across these 2 targets.
- Equity metrics as guardrails: Existing equity metrics and frameworks are rarely embedded as binding constraints in campaign planning [11-13,15,24]. There is limited empirical work on how explicit guardrails (eg, minimum subgroup adoption, caps on adoption gaps) reshape the space of admissible channel and personalization strategies [11-13,15,24].
- Reproducible data-driven ABM for strategic planning: Most studies evaluate isolated interventions or channels. Few use reproducible, survey-calibrated ABMs that tie nationally representative data on media use, digital skills, social connectedness, and preventive behaviors to multichannel campaign scenarios [31,32]. There is particularly little work on "virtual laboratories" that vary channel mix, personalization, framing, and equity guardrails jointly under fixed budgets within a unified, budget-constrained planning framework [31,32].

This study addresses these gaps by developing a reproducible data-to-ABM pipeline for preventive campaigns for older adults. We focused on older adults in South Korea—a high-income, aging health system—and aim to derive strategic planning principles that may be adaptable to similar contexts. Using matched national survey data, we constructed a heterogeneous older adult population segmented by latent class, embedded agents in multilayer offline-online networks, and simulated 15 prespecified campaign scenarios under fixed budgets and explicit equity guardrails. Our goal was to provide a strategic planning environment that allows planners to compare channel mixes, personalization philosophies, and loss framing intensities under explicit equity guardrails and interpret the resulting portfolios as decision-support inputs for national or regional planning. Figure 1 summarizes the data-to-ABM pipeline from data sources to simulation outputs. It highlights data integration and profiling; core modeling modules; the simulation engine; and the calibration and sensitivity loop used to evaluate channel allocation, personalization, and loss framing under explicit equity guardrails.

We focused on 5 research questions (RQs):

RQ1: how do single-channel campaigns compare with a background-only baseline?

RQ2: do multichannel mixes produce synergy or dilution under fixed budgets?

RQ3: how do 2 contrasting personalization strategies—outcome-targeted (cautious equity-first) versus class-customized (segment-optimized)—affect mean outcomes?

RQ4: how do these personalization strategies reshape distributional equity across latent classes?

XSL•FO
RenderX

RQ5: does loss-framed messaging act as a secondary tuner of adoption, and do its effects differ between vaccination and screening?

Figure 1. Overview of the data-to-ABM pipeline for evaluating channel allocation and personalization strategies under budget and equity guardrails. ABM, agent-based model; CWTA, cumulative weighted threshold adoption; HBM, health belief model; KMP, Korea Media Panel Survey; KNHANES, Korea National Health and Nutrition Examination Survey; LCA, latent class analysis; TTA, time to adoption; W_{BG}, background exposure weight; WS, Watts-Strogatz small-world network.

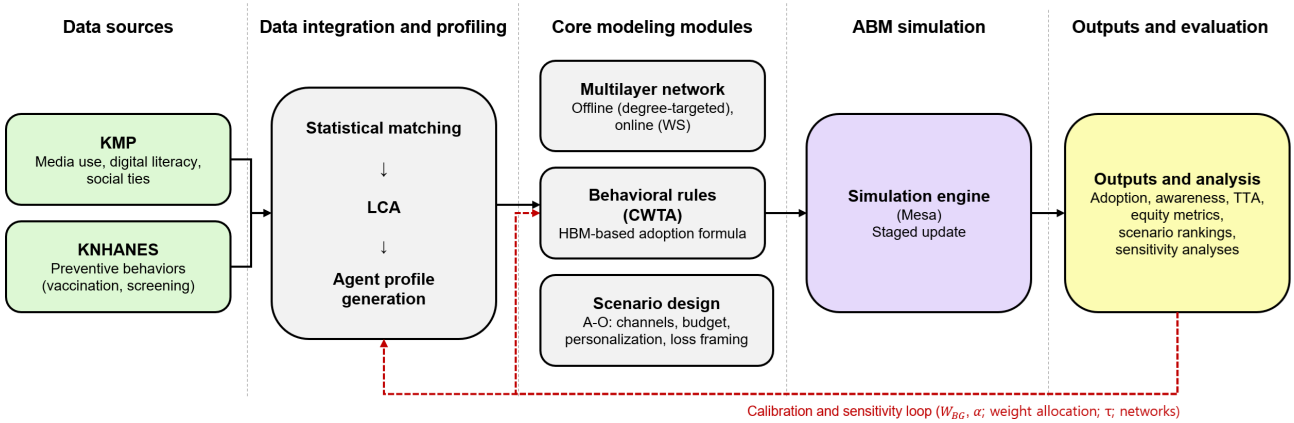

Methods

Study Design

Campaign planning for older adults must account for how exposure unfolds over time and spreads through offline and online social ties [18,33-35]. These dynamics are difficult to represent in static regression or compartmental models, which assume homogeneous exposure and ignore network reinforcement [36,37]. The challenge is amplified by having 2 preventive targets with distinct determinants: influenza vaccination and routine screening.

To address these limitations, we developed an ABM that represents each older adult as an agent embedded in multilayer networks [38]. The model captures heterogeneous exposure, repeated social reinforcement, and channel-specific media effects under fixed campaign budgets. It enables prospective comparison of prespecified channel mixes, personalization strategies, and equity guardrails as alternative campaign designs. Therefore, we used the model as a strategic planning tool for digital health campaigns targeting older adults.

Data Sources and Agent Construction

Data Sources and Statistical Matching

Two national surveys were combined: the 2022 Korea Media Panel Survey (KMP) [39] and the 2022 Korea National Health and Nutrition Examination Survey (KNHANES) [40]. KMP captured media use, digital literacy, and social relationship patterns, while KNHANES reported preventive behaviors including influenza vaccination and screening. Linking the 2 datasets enabled individual-level integration of media exposure, digital access, social connectedness, and preventive behaviors.

We restricted KNHANES recipients to adults aged ≥65 years and defined the KMP donor pool as adults aged ≥55 years to improve age comparability during matching. After harmonizing overlapping variables, we performed one-to-one nearest-neighbor matching with replacement using Gower distance on key sociodemographic covariates [41]. This approach assigned media, digital, and social attributes from KMP donors to matched KNHANES recipients while retaining KNHANES health outcomes. The matching diagnostics met our prespecified acceptance criterion; the full covariate list and diagnostics are reported in Multimedia Appendix 1.

Latent Class Analysis

Latent class analysis was applied to the matched dataset to identify classes of older adults with distinct media, digital, social, and preventive profiles [42]. The indicator domains included sociodemographics, digital literacy, media use, social activity, and vaccination and screening behaviors. Models with 3 - 10 classes were evaluated using the Akaike information criterion (AIC), the Bayesian information criterion (BIC), entropy, and the bootstrap likelihood ratio test (K vs K–1). A 6-class solution was selected based on the lowest BIC and high entropy (BIC=86,578.57; entropy=0.92). Solutions with additional classes yielded marginal AIC improvements. However, they increased complexity without BIC gains. Accordingly, the 6-class structure was selected as the basis for class-based personalization in the simulation. Variable definitions, discretization rules, and full model comparison results are provided in Table S1 in Multimedia Appendix 1, and detailed class profiles are reported in Table S3 in Multimedia Appendix 2. For descriptive summaries, individuals were assigned to their modal class. For simulation inputs, posterior class probabilities were retained as personalization weights. This approach preserves interpretability and propagates classification uncertainty into the inputs. We treated class membership as time invariant over the 12-month simulation horizon.

Agent Initialization and Network Priors

Each matched individual was instantiated as an agent with a channel affinity vector and an offline degree prior derived from survey responses. Channel affinities captured digital and traditional media use. Offline degree priors were based on self-reported social activity frequency and served as individual target degrees during offline network construction. Details of

XSL•FO
RenderX

affinity construction, degree-level mapping, and multilayer network generation are provided in Multimedia Appendix 1 and in the Network and State Dynamics section.

Outcomes and Equity Metrics

Primary outcomes were final adoption at month 12 (proportion of agents who adopted) and mean time to adoption among adopters. We used "adoption" as a generic term for vaccination uptake and screening participation. Awareness was treated as a supporting outcome.

We assessed equity across latent classes using 5 distributional metrics. A_{min} denoted the minimum class-level adoption. We defined the 90 - 10 gap as the difference between the 90th and 10th percentiles of class-level adoption and relative disparity index (RDI) as the ratio of maximum to minimum class-level adoption. We also calculated 2 entropy-based indices, Theil T and Atkinson $A_{0.5}$ (ε=0.5), as descriptive measures of inequality [43,44]. For interpretation, we used prespecified illustrative planning benchmarks ("equity guardrails"): $A_{min} \geq 0.60$, 90 - 10 gap≤0.25, and RDI≤1.5. These thresholds are tunable planning benchmarks for comparative portfolio assessment and not universal ethical or empirical standards. In practical terms, they set (1) a minimum-adoption floor for the lowest-adoption class and (2) limits on absolute and relative disparities. Specifically, we required ≤25 percentage points for the 90 - 10 gap and a ratio ≤1.5 for RDI. These thresholds use the same probability scale as the calibrated uptake in our setting. Theil T and Atkinson $A_{0.5}$ were reported descriptively and were not used for guardrail screening.

Scenario-level adoption and equity metrics for all campaign scenarios (A-O), for both vaccination and screening, are reported in Tables S1 and S2 in Multimedia Appendix 2. We ran 100 Monte Carlo replications per scenario. We report all outcomes as means with 95% percentile intervals (2.5th-97.5th percentile), stratified by task (vaccination vs screening).

Scenario Design

Channel Allocation and Budget (RQ1-RQ2)

We evaluated 15 prespecified scenarios (A-O) that varied in 4 design factors: channel allocation, budget intensity, class-level personalization, and loss framing. Exposure weights were specified as relative per-step exposure intensity parameters (rather than monetary spending). Scenarios A to H tested single channels or fixed-sum channel mixes at a common per-step media budget. Scenario A served as a control with no paid media exposure; awareness could arise from background exposure and social awareness. Scenarios B to D isolated individual channels (television [TV], digital, or print). Scenarios E to H implemented fixed-sum mixes combining 2 or 3 channels.

Scenarios I and J varied total budget intensity while keeping channel shares fixed. Scenario I scaled down exposure relative to the fixed-sum TV + digital mix, whereas scenario J scaled it up. Together, these scenarios represented low- and high-budget variants without changing the underlying channel mix.

Personalization: Efficiency and Equity (RQ3-RQ4)

Scenarios K and L introduced class-level personalization based on latent classes. Scenario K implemented equity-focused reweighting. For each task, the lowest-baseline uptake class received a small increase in television or digital exposure. All other classes had TV + digital exposure reduced by equal margins to maintain a constant class-average TV + digital budget. This design represents a conservative, budget-neutral policy that nudges media toward the lowest-baseline uptake class.

Scenario L assigned class-specific television and digital weights by latent class under the same class-level budget parity. Weights were tailored to each class's baseline media use profile (television-heavy, digital-heavy, or mixed), rather than applying a uniform mix. This strategy represents a more aggressive, class-tailored policy that explicitly aligns channel mix with each class's baseline media use profile.

Message Framing (RQ5)

Scenarios M to O applied loss framing multipliers (k_{neg}) to television and/or digital exposure. Meta-analytic evidence suggests that framing effects are generally modest and behavior dependent. Gain framing shows advantages for prevention behaviors, whereas effects are mixed or weak for detection behaviors [45,46]. Accordingly, we tested k_{neg} values near 1.0, that is, 0.95 to –1.10 for vaccination and 1.05 to 1.20 for screening [45,46]. Tables 1 and 2 report the exact channel weights and multiplier values for all scenarios.

XSL•FO
RenderX

Table . Channel allocation weights for campaign scenarios A to L.

Scenario	Name	Purpose	W_{TV}[c]	W_{DG}[c]	W_{Print}[c]
A[a]	Control	Background only	0.000	0.000	0.000
B	TV[b]-only	Isolate TV	0.100	0.000	0.000
C	Digital-only	Isolate digital	0.000	0.100	0.000
D	Print-only	Isolate print	0.000	0.000	0.100
E	TV + digital	Fixed-sum mix	0.050	0.050	0.000
F	TV + print	Fixed-sum mix	0.050	0.000	0.050
G	Digital + print	Fixed-sum mix	0.000	0.050	0.050
H	All channels	Integrated mix	0.033	0.033	0.033
I	Low-budget	Scaled-down exposure	0.025	0.025	0.000
J[d]	High-budget	Scaled-up exposure	0.200	0.200	0.000
K[e]	Equity-focused reweighting	Reweight toward lowest-baseline class	Class dependent	Class dependent	0.000
L[f]	Class-tailored channel portfolios	Class-tailored channel mix	Class dependent	Class dependent	0.000

[a]Background-only awareness with per-step spillover (W_{BG}=0.12 for vaccination, 0.10 for screening). W_{BG} denotes background exposure weight.

[b]TV: television.

[c]W_{TV}, W_{DG}, and W_{Print} denote per-step channel exposure weights for television, digital, and print media, respectively. Weights sum to approximately 0.10 (minor differences reflect rounding).

[d]Scenario J increases the total exposure (budget intensity) while keeping the channel shares fixed.

[e]Scenario K varies the intensity of the equity-focused reweighting ($\delta \in \{0.02, 0.05\}$). For vaccination, the lowest-baseline uptake class was Class 6 (W_{TV}=0.00, $W_{DG} = 0.10 + \delta$); for screening, the lowest-baseline uptake class was Class 1 ($W_{TV} = 0.10 + \delta$, W_{DG}=0.00). In both cases, all other classes received symmetric low-intensity weights ($W_{TV} = 0.05 - \delta/10$; $W_{DG}= 0.05 - \delta/10$). Results for δ=0.02 are reported in the main text, and δ=0.05 is presented as a higher-intensity sensitivity variant in Table S2 in Multimedia Appendix 3.

[f]Weights vary by latent class: classes 1 and 5: W_{TV}=0.10; class 6: W_{DG}=0.10; classes 2 - 4: W_{TV}=0.05 and W_{DG}=0.05.

Table . Loss framing scenarios M to O with intensity multiplier ranges.[a]

Scenario	Name	Definition	W_{TV}[d]	W_{DG}[d]	W_{Print}[d]
M	TV[b] (loss framing)	B with $W_{TV} \times k_{neg}$[c]	$0.10 \times k_{neg}$	0.00	0.00
N	Digital (loss framing)	C with $W_{DG} \times k_{neg}$	0.00	$0.10 \times k_{neg}$	0.00
O	TV + digital (loss framing)	E with $W_{TV}/W_{DG} \times k_{neg}$	$0.05 \times k_{neg}$	$0.05 \times k_{neg}$	0.00

[a]Vaccination: {0.95, 1.00, 1.05, 1.10}; Screening: {1.05, 1.10, 1.15, 1.20}. All exposure probabilities were clipped to [0,1].

[b]TV: television.

[c]k_{neg} denotes the loss framing intensity multiplier applied to baseline channel weights.

[d]W_{TV}, W_{DG}, and W_{Print} denote per-step channel exposure weights for television, digital, and print media, respectively. The weights sum to approximately 0.10 (minor differences reflect rounding).

Network and State Dynamics

State Dynamics

The simulation ran over 12 monthly steps (t=1, …, 12). At each step, we iterated agents in random order and applied state conditional rules. At each time step, we first updated awareness and then updated adoption conditional on awareness. Once adopted, agents remained in the adopted state (state=2).

At each time step t, the awareness probability combined 3 components. The first component was background exposure. The second was campaign media exposure, determined by scenario-specific channel weights and agent-level channel use for television, digital, and print media. The third was the share of aware neighbors, weighted by ωa. In the control scenario A, all paid media terms were set to zero, and awareness arose from background exposure and social awareness (spillover).

Conditional on awareness, the adoption probability combined baseline propensity, media exposure, and social reinforcement. The reinforcement term activated only when at least 2 neighbors had adopted (threshold τ=2) and influenced adoption but not

XSL•FO
RenderX

awareness. In loss framing scenarios M to O, we multiplied the television and digital exposure terms by k_{neg}. All probabilities were clipped to the [0,1] interval. Textbox 1 provides a brief road map of the model logic (awareness, adoption, and reinforcement) and how to read the simulation outputs.

Textbox 1. How the model works and how to read the outputs.

1. What the model simulates: the model simulates how campaign exposure and social ties translate into adoption over 12 monthly steps.
2. Awareness (notice the message): an agent becomes aware via background exposure, paid media (scenario weights × the agent's channel use), and social awareness (neighbors who are aware).
3. Adoption (take action): only if aware, the agent may adopt based on baseline propensity plus media and social influence.
4. Reinforcement (social proof): social influence applies only after at least τ=2 neighbors have already adopted.
5. Once adopted, always adopted: adoption is an absorbing state through month 12.
6. Outputs (how to read results): scenarios will be compared on final adoption, time to adoption, and equity across classes (A_{min}, 90 - 10 gap, relative disparity index [RDI]); the results will be interpreted as relative trade-offs under the same constraints and not as point forecasts.

Contact Structure

We modeled contacts as a 2-layer network with offline and online ties. For social influence, each agent's neighborhood was defined as the union of its offline and online neighbors.

For baseline and channel-mix scenarios (A-H), we generated the analytic agents (vaccination, N=2405; screening, N=2400) and both network layers once and held them fixed across all Monte Carlo replications. For intervention and framing scenarios (I-O), we regenerated both layers at the start of each replication to reflect structural uncertainty while preserving a stable baseline design. This design choice was used to (1) reduce Monte Carlo variance when comparing fixed-budget channel-mix scenarios (A-H) and (2) incorporate structural uncertainty for intervention or framing variants (I-O).

Offline social layer (face-to-face): we mapped survey-based social activity frequency to an agent-specific degree target. We then built an undirected graph using a degree-targeted algorithm while avoiding self-loops and multi-edges [47,48].

Online social layer (network-based): we constructed the online layer as a Watts-Strogatz small-world network with a mean degree of 6 and rewiring probability of 0.10 [49].

Calibration and Internal Validation

We calibrated 2 parameters for each task to match KNHANES 2022 targets for older adults (vaccination uptake 0.848; screening participation 0.744) [40]. We varied the baseline propensity multiplier and the background exposure over prespecified ranges. We then selected the parameter pair that minimized the absolute deviation from these targets under the background-only control scenario (A). This procedure yielded the values $\alpha_{vaccine}$=1.15 with W_{BG}=0.12 and $\alpha_{screening}$=1.03 with W_{BG}=0.10.

We assessed calibration quality using target deviation summaries and decile reliability plots that compared survey targets with deciles of baseline propensity [50-52]. These diagnostics used individual-level outputs aggregated across Monte Carlo replications. Tables S1 and S2 in Multimedia Appendix 4 provide the best-fitting parameter sets and the per-class calibration summaries. In this study, "validation" refers to internal diagnostics (calibration-to-target and reliability checks) within a survey-calibrated framework and does not constitute external validation using real-world campaign data.

Sensitivity Analyses

We assessed robustness along 4 prespecified axes: link function, weight allocation, the social reinforcement threshold, and network uncertainty. For each axis, we reestimated scenario-level adoption vectors under the corresponding sensitivity configuration and compared them with the baseline ordering. We quantified rank stability using Spearman correlations and maximum rank shifts (Δrank) and treated $\rho \geq 0.90$ and Δrank$\leq$1 as heuristic benchmarks for high stability. Tables S1 and S2 in Multimedia Appendix 3 provides the full protocols, stability statistics, and intensity variants for the personalization strategy [53-55].

Ethical Considerations

This study used deidentified, publicly available microdata (KMP 2022; KNHANES 2022) [39,40] with no direct participant contact. Consent to participate was not applicable. Under institutional policy for secondary analyses of deidentified public datasets, formal institutional review board approval was not required. Under article 2(1) of the Bioethics and Safety Act of Korea [56], a "human subjects research project" is defined as research involving direct interaction with individuals or the use of identifiable personal information. As this study used only deidentified secondary data and did not involve identifiable information, it does not fall under this definition. Furthermore, according to article 13(1) of the Enforcement Rule of the Bioethics and Safety Act of Korea [57], research using publicly available data or research that does not collect or record personally identifiable information may be exempt from institutional review board review.

Results

Sample and Class Overview

The matched analytic sample included 2405 adults aged ≥65 years. Latent class analysis identified 6 classes with distinct sociodemographic and media use profiles [42]. Class sizes ranged from 245 (10.2%) to 534 (22.2%).

Baseline uptake varied across classes; the class-specific n and N are reported in Table S3 in Multimedia Appendix 2. Class 6

XSL•FO
RenderX

had the lowest baseline vaccination uptake, and Class 1 had the lowest baseline screening participation.

Calibration and Internal Validation

We calibrated the baseline parameters to match the empirical adoption targets for older adults in KNHANES 2022 [40]. The targets were 2039 of 2405 (84.8%) for vaccination and 1786 of 2400 (74.4%) for screening. The selected parameter sets matched these targets closely (absolute deviation≤0.002). Calibration metrics supported good fit: Brier scores were 0.142 for vaccination and 0.187 for screening; calibration slopes were 1.15 and 1.08, respectively; and intercepts were −0.22 and −0.08, respectively [50-52].

Decile reliability plots under the calibrated background-only scenario showed close agreement between predicted and observed adoption probabilities (Figure 2) [50-52]. We analyzed 2405 individuals for vaccination and 2400 for screening (5 screening nonresponses excluded). Per-class calibration summaries are reported in Table S2 in Multimedia Appendix 4.

Figure 2. Decile reliability plots comparing mean predicted adoption probabilities with observed adoption rates for vaccination (A) and screening (B) under the calibrated background-only scenario.

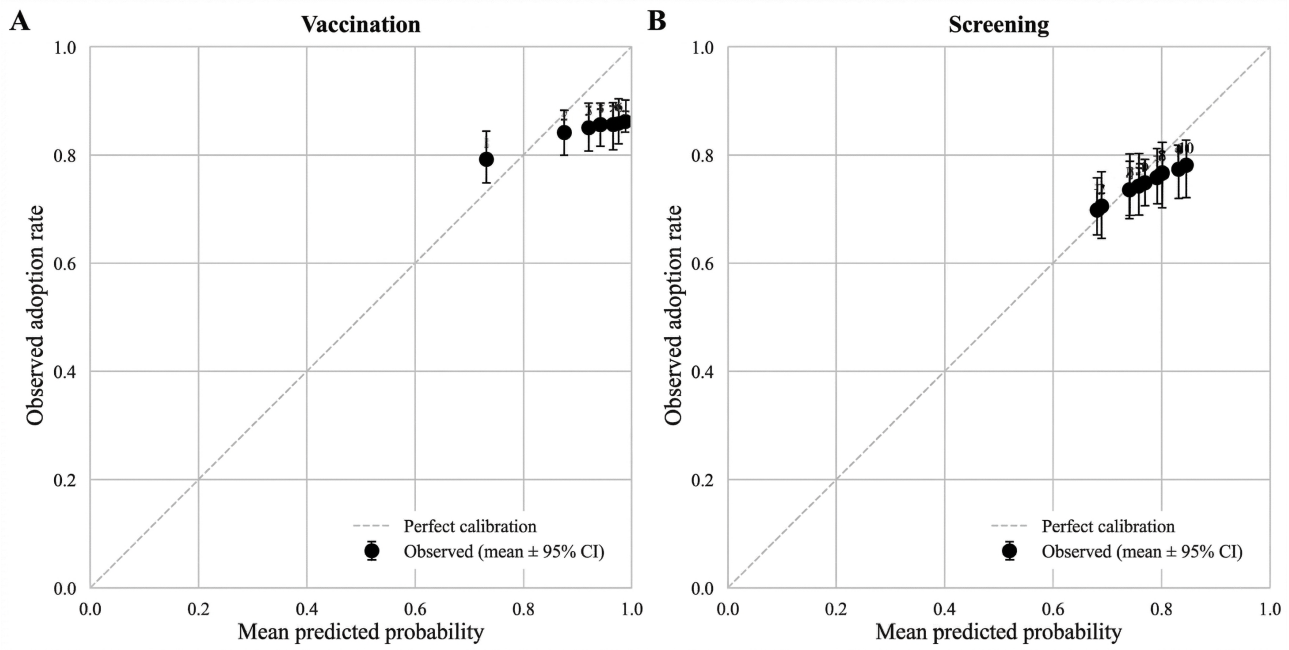

Channel Allocation and Budget (RQ1-RQ2)

For RQ1, we compared each single-channel campaign (B-D) with the background-only baseline (A). The TV-only campaign (B) produced the largest gains. For vaccination (N=2405), final adoption increased from 85% (n=2044) in A to 94.3% (n=2268) in B (9.3 percentage points, 95% PI 7.3 - 11.0). Mean time to adoption decreased from 6.18 (SD 0.07) to 5.34 (SD 0.06) months (0.84 months faster, 95% PI 0.65 - 0.94). For screening (N=2400), adoption increased from 74.5% (n=1788) to 89.4% (n=2146; 14.9 percentage points, 95% PI 12.9 - 17.0). Mean time to adoption decreased from 6.88 (SD 0.07) to 6.09 (SD 0.07) months (0.79 months faster, 95% PI 0.65 - 0.94). Digital-only and print-only campaigns produced smaller gains. For vaccination, the digital-only and print-only campaigns increased adoption by about 139 adopters (5.8 percentage points) and about 17 adopters (0.7 percentage points), respectively. For screening, the corresponding gains were about 199 adopters (8.3 percentage points) and 29 adopters (1.2 percentage points).

For RQ2, we held the per-step media budget constant and compared mixed-channel strategies (E-H) with the TV-only benchmark (B). All mixed strategies (TV + digital, TV + print, digital + print, and all channels) improved upon the background-only baseline for both vaccination and screening; however, none surpassed the TV-only campaign in either final adoption or time to adoption (Figure 3). For the TV + digital mix (E), the synergy index (difference in adoption between the mixed strategy and the best single-channel strategy) was slightly negative: −1.2 percentage points (95% PI −3.6 to 1.2 percentage points) for vaccination and −2.3 percentage points (95% PI −5.8 to 1.3 percentage points) for screening, corresponding to a dilution of roughly 1 to 2 percentage points relative to the best single-channel option. Other mixed strategies showed similarly nonpositive synergy. Tables S1 and S2 in Multimedia Appendix 2 provide full scenario-level adoption and equity metrics for all campaigns (A-O).

XSL•FO
RenderX

Figure 3. Final adoption and time-to-adoption outcomes across scenarios A to H for vaccination (A) and screening (B). Scenarios: A, control; B, television (TV)-only; C, digital-only; D, print-only; E, TV + digital; F, TV + print; G, digital + print; H, all channels; I, low-budget; J, high-budget; K, equity-focused reweighting; L, class-tailored channel portfolios.

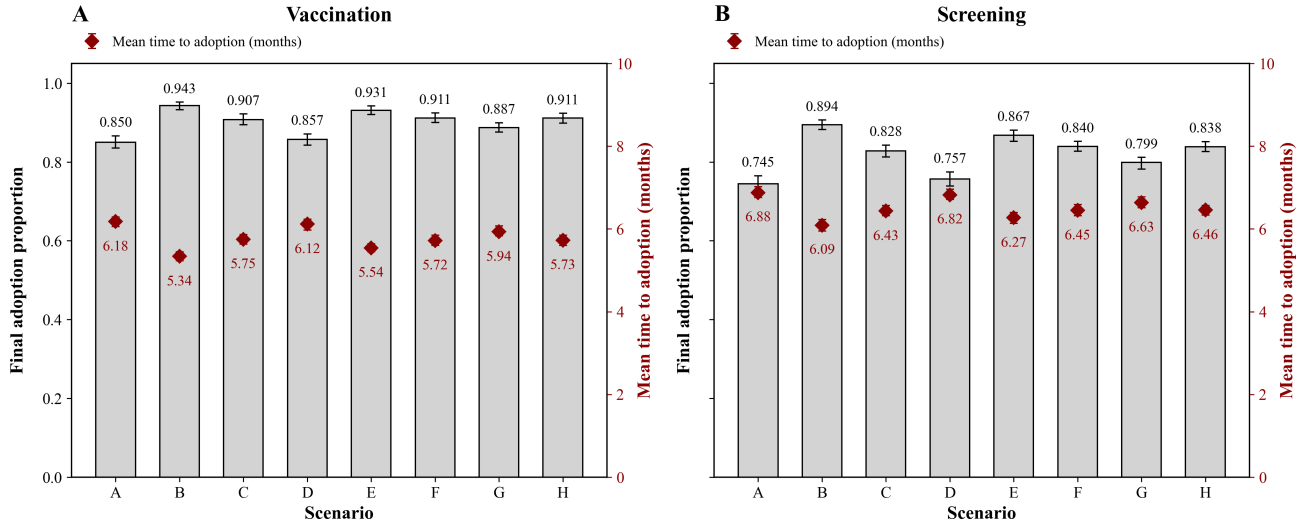

Personalization: Efficiency and Equity (RQ3-RQ4)

We compared the integrated baseline (H) with 2 personalization strategies: equity-focused reweighting (K) and class-tailored channel portfolios (L). For vaccination (N=2405), mean adoption increased from 91.2% (n=2193) in H to 93.3% (n=2244) in K and 94.6% (n=2275) in L. Furthermore, A_{min} increased from 86.8% (H) to 90.3% (K) and 90.9% (L). The 90‑10 gap narrowed from 5.7 percentage points in H to 4.5 and 4.7 percentage points in K and L, respectively. For screening (N=2400), mean adoption increased from 83.8% (n=2011) in H to 88.2% (n=2117) in K and 89.5% (n=2148) in L. Moreover, A_{min} increased from 77.6% (H) to 83.2% (K) and 85.3% (L). The 90‑10 gap narrowed from 9.2 percentage points to 7.4 and 6.2 percentage points in K and L, respectively.

Across both targets, strategy H had the lowest mean adoption and A_{min}. Strategy L achieved the highest mean adoption and A_{min}, and both personalization strategies (K and L) reduced the between-class disparities relative to H.

Table 3 summarizes the mean adoption and key equity metrics for scenarios H, K, and L. Figure 4 shows the resulting efficiency-equity trade-offs. Tables S1 and S2 in Multimedia Appendix 2 report the additional inequality indices and scenario-level adoption and equity metrics for all campaigns (A-O), including the TV-only campaign (B), low- and high-budget variants (I-J), and framing variants (M-O).

Table . Mean adoption and key equity metrics for scenarios H (all channels), K (equity-focused reweighting), and L (class-tailored channel portfolios).

Scenario	Mean adoption, n (%)	A_{min}, %	90‑10 gap, percentage points
Vaccination (N=2405)			
H	2193 (91.2)	86.8	5.7
K	2244 (93.3)	90.3	4.5
L	2275 (94.6)	90.9	4.7
Screening (N=2400)			
H	2011 (83.8)	77.6	9.2
K	2117 (88.2)	83.2	7.4
L	2148 (89.5)	85.3	6.2

XSL•FO
RenderX

Figure 4. Efficiency-equity trade-offs for integrated baseline and personalized campaigns (H-L) for vaccination (A) and screening (B).

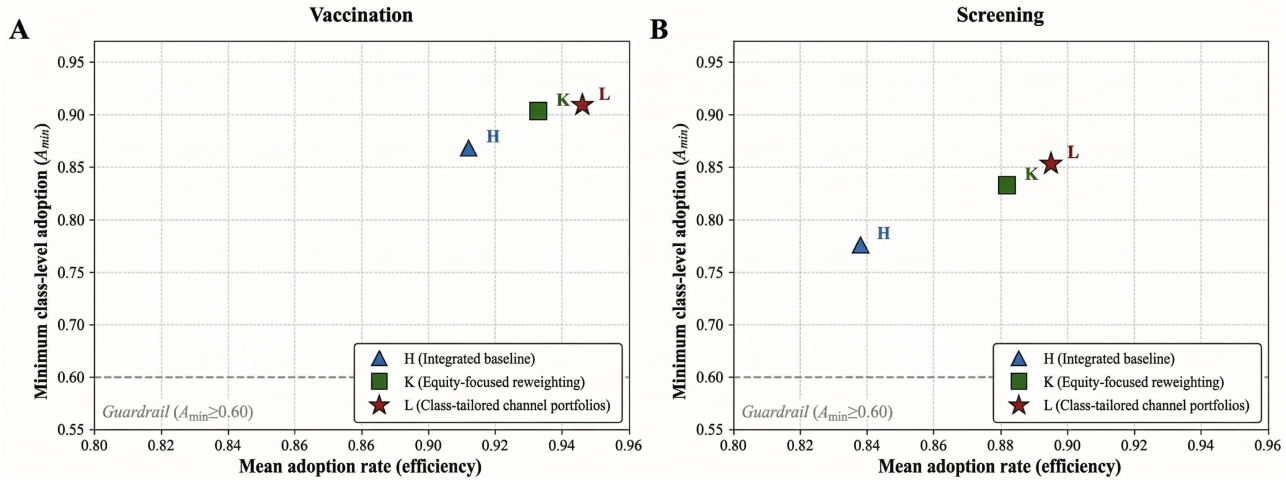

Message Framing (RQ5)

Averaged across scenarios M to O, increasing loss framing intensity (k_{neg}) produced modest, monotonic gains. For vaccination (N=2405), raising k_{neg} from 0.95 to 1.10 increased adoption from 92.3% (n=2220) to 92.9% (n=2234), a gain of about 14 adopters (0.6 percentage points). Mean time to adoption decreased from 5.62 (SD 0.16) to 5.54 (SD 0.17) months. For screening (N=2400), raising k_{neg} from 1.05 to 1.20 increased adoption from 88.8% (n=2131) to 89.7% (n=2153), a gain of about 22 adopters (0.9 percentage points). Mean time to adoption decreased from 6.01 (SD 0.16) to 5.93 (SD 0.18) months.

In the television-based screening scenario (M; N=2400), increasing k_{neg} from 1.05 to 1.20 increased adoption from 91.4% (n=2194) to 92.3% (n=2215), with parallel reductions in time to adoption. Digital-only screening (N) showed lower overall adoption but similar incremental gains across the same k_{neg} range. Gains between adjacent intensity levels were small (≤0.5 percentage points) and tended to taper at higher k_{neg}. The 95% percentile intervals overlapped, and we observed no evidence of backfiring within the tested range. Detailed dose-response curves and class-level heterogeneity by scenario and target are reported in Table S3 and Figure S1 in Multimedia Appendix 3.

Sensitivity Analyses

Using the prespecified stability benchmarks (ρ≥0.90 and Δrank≤1) as heuristics, scenario ordering for mean adoption was stable across prespecified axes (link function, weight allocation, reinforcement threshold, and network uncertainty). Maximum rank shifts were small (Δrank≤2), with the social-centric weight allocation variant for screening showing Δrank=2 while preserving the qualitative ordering. Full protocols and stability statistics are reported in Table S1 in Multimedia Appendix 3.

We changed the social reinforcement threshold from τ=2 (≥2 adopting neighbors) to τ=1 (≥1 adopting neighbor) and regenerated networks with alternative random seeds. These modifications left the scenario ordering unchanged for both tasks (ρ=1.00; Δrank=0). Alternative weight allocation variants produced only minor shifts for screening (ρ as low as 0.90; Δrank up to 2) while preserving the qualitative ordering. Comparisons between logistic and clipped-linear link functions preserved these rankings for mean adoption in both tasks.

Discussion

Principal Findings

Across the 15 prespecified scenarios, 5 patterns summarized how campaign design choices shaped efficiency and equity among older adults. These patterns were similar for vaccination and screening and remained stable across sensitivity analyses.

First, efficiency was driven mainly by channel allocation. The TV-only strategy (B) produced some of the highest mean adoption rates under a fixed per-step budget. Mixed-channel portfolios (E-H) improved outcomes relative to the background-only control (A) but did not surpass B, suggesting that dividing limited exposure across channels diluted reinforcement within high-TV–affinity clusters. Under fixed budgets, a single strong channel can therefore outperform intuitive “balanced mix” strategies.

Second, personalization shifted the portfolio toward a more favorable efficiency-equity trade-off within the modeled scenario space. The equity-focused reweighting intervention (K) produced modest gains in mean adoption but larger improvements in A_{min} and the 90‑10 gap by lifting the lowest-performing class. The class-tailored strategy (L) showed a strong combined profile under fixed budgets. For vaccination, it matched the efficiency of B while improving equity. For screening, it achieved adoption comparable to B with narrower disparities. Taken together, these results suggest that class-tailored allocations can improve equity without necessarily reducing effectiveness. Under the modeled assumptions, they may also yield simultaneous improvements relative to uniform integrated mixes.

Third, high-budget expansion (J) increased average adoption and narrowed disparities, but it required higher total exposure. In contrast, K and L met guardrail criteria while operating under fixed budgets, indicating that reallocating exposure can achieve equity gains without increasing total exposure.

Fourth, framing acted as a secondary adjustment rather than a primary lever. Stronger loss framing produced small, monotonic

XSL•FO
RenderX

improvements (<1 percentage point) in both tasks and showed no evidence of backfiring within the tested range. However, these increments were minor compared with differences driven by channel allocation and personalization [45,46].

Finally, these strategic patterns were robust. Scenario orderings were stable across alternative reinforcement thresholds, weight allocations, and network realizations. Results under logistic versus clipped-linear link functions showed the same qualitative hierarchy.

Comparison With Prior Work

These findings relate to 4 strands of prior work: equity-focused implementation frameworks, media campaign evaluations, personalization in digital health, and ABMs of preventive behaviors.

First, implementation science frameworks such as the PRECEDE-PROCEED model and RE-AIM/PRISM increasingly emphasize equity [13,16]. However, they offer limited operational guidance on how to allocate budgets across channels and segments under explicit constraints. Most evaluations still describe who was reached and how effects varied after implementation [15,17]. Our model instead uses equity metrics as prespecified guardrails that screen and flag campaign portfolios once equity thresholds are violated. In doing so, it treats equity as an ex ante decision rule for digital health strategic planning rather than a purely descriptive endpoint.

Second, media campaign studies usually examine single-channel effects or short-term digital engagement metrics. Television is known to be effective for older adults, and digital channels offer heterogeneous reach, but few studies compare fixed-budget channel portfolios [18,58,59]. Existing work rarely tests whether multisource exposure creates synergy or whether splitting limited resources across channels weakens reinforcement. Our findings show that, under complex contagion dynamics, mixed strategies at fixed budgets do not outperform a strong single-channel allocation, helping to clarify a long-standing ambiguity in the campaign literature.

Third, personalization is often promoted as a way to improve relevance and engagement, while many authors warn that tailoring may widen inequalities by favoring digitally advantaged groups [60,61]. Reviews document these risks but seldom compare different personalization philosophies under common efficiency and equity metrics [60,61]. By contrasting equity-focused reweighting (K) with class-tailored channel portfolios (L), we show that personalization need not exacerbate disparities. When designed to lift the lowest-performing class or align exposure with segment-specific media affinities, personalization can improve both mean adoption and distributional equity.

Fourth, most ABMs of preventive behaviors have focused on disease transmission or generic opinion dynamics [28,32]. Few models combine nationally representative microdata, latent class segmentation, multilayer networks, and prespecified equity constraints within a single portfolio-testing framework [28,32]. In this study, the ABM links empirical population heterogeneity with mechanistic diffusion dynamics, allowing planners to test channel mixes, reallocations, and framing intensities under consistent, reproducible conditions.

Taken together, these features define a budget-constrained, equity-aware decision framework for preventive campaigns for older adults. This framework helps explain why some channel and personalization strategies are more favorable than others on both efficiency and equity within the modeled scenario space.

Implications for Digital Health Strategic Planning

These results are intended to support comparative decision-making for campaign design rather than point forecasting. Planners can use the model to compare candidate portfolios that vary channel mix and targeting under shared budget and equity constraints. To operationalize the scenario inputs, the relative exposure weights can be interpreted as the intended distribution of campaign effort across channels, and these weights can be translated into proportional budget shares or delivery targets using standard planning metrics (eg, television reach and frequency; digital reach and impressions; print distribution volume). The outputs provide comparative evidence for selecting among portfolios, with efficiency and equity evaluated jointly using the study's equity metrics and guardrails.

These findings suggest 3 priorities for public health planners working under budget constraints.

First, channel allocation should be treated as the primary strategic lever. Across both vaccination and screening, TV-only allocations produced some of the highest mean adoption levels under fixed budgets. Mixed portfolios improved outcomes but diluted reinforcement compared with television alone. For older adult populations with strong offline clustering, concentrating resources on a high-reach channel may therefore be more effective than spreading budgets thinly across multiple platforms.

Second, equity-oriented reallocation is both feasible and effective. The equity-focused reweighting strategy (K) raised minimum class-level adoption and narrowed class disparities with only small changes in mean outcomes. This shows that modest, budget-neutral shifts can materially improve equity. The class-tailored strategy (L) further improved both mean and distributional outcomes, indicating that segmentation-based resource allocation need not exacerbate inequities. These results suggest that agencies with basic segmentation data and media-buying capacity could adopt class-tailored portfolios to achieve efficiency-equity gains without increasing total exposure budgets.

Third, loss framing yielded only minor incremental gains within the tested range, so it should follow—rather than replace—allocation and targeting decisions.

Together, these implications position equity guardrails as a practical planning benchmark. Planners can start from an integrated channel mix and deprioritize strategies that fall below minimum subgroup adoption or widen disparities. Among the remaining options, they can consider conservative reallocations or class-tailored portfolios—such as scenarios K and L— that

XSL•FO
RenderX

show a more favorable efficiency-equity trade-off while maintaining strong mean adoption.

Limitations

Several limitations should be noted.

First, the ABM relied on structural assumptions about media influence, baseline propensity, and network topology. These assumptions simplify complex behavioral processes and may omit unmeasured factors. Calibration matched survey targets, and scenario rankings were stable across sensitivity checks. However, some model misspecification may remain.

Second, offline and online networks were approximated using survey-derived degree targets and a Watts-Strogatz small-world structure. Real-world networks may differ in clustering, multiplexity, tie strength, and mixing. These approximations may overestimate or underestimate the strength of social influence. They may also distort within-group versus between-group reinforcement. Rank stability across alternative network realizations suggests that the qualitative ordering of strategies is unlikely to depend on a single specification. However, more detailed network data could refine these structures. We did not test alternative class structures, and the sensitivity of scenario rankings to segmentation choices remains a topic for future work. Segmentation was cross-sectional, so we could not test temporal stability or class transitions.

Third, the model focused on exposure and diffusion rather than message content and quality, source trustworthiness, or system-level access barriers. Reactance, stigma, and message fatigue were not modeled, and strong fear-based or stigmatizing loss frames fell outside the tested k_{neg} range. Modest loss framing did not show evidence of backfiring within the simulated range, but these results should not be generalized to more aggressive messaging strategies.

Fourth, the model evaluated a 1-year campaign horizon and did not capture longer-term maintenance, multiyear cycles, or downstream health outcomes. Adoption was modeled as a single-step absorbing behavior, whereas real-world screening and vaccination involve repeated decisions and system-level interactions. Budget was represented as relative exposure intensity rather than monetary spending. Therefore, the results do not directly translate to monetary costs or cost patterns across media markets.

Finally, the analysis is based on matched Korean survey data. Several contextual features—such as high television reach, strong offline clustering, and a pronounced digital divide—may be specific to this setting. Applying the framework elsewhere would likely require recalibrating context-dependent inputs, such as channel reach patterns, the magnitude of the digital divide, and network mixing or clustering. In contrast, the model's core mechanisms—heterogeneous media affinity and reinforcement-based diffusion through social ties—may be more portable. Their magnitudes, however, may vary across settings. External validation in additional settings will be important for assessing transportability and for refining model-based decision tools.

Conclusions

This study used a data-driven ABM to examine how channel allocation, personalization, and loss framing shape preventive campaign performance for older adults under fixed budgets. In this survey-calibrated simulation, equity-oriented personalization of television and digital exposure improved subgroup equity and increased mean adoption relative to uniform integrated mixes and simple channel diversification. We evaluated scenarios using prespecified, illustrative equity guardrails. These patterns suggest that accounting for diffusion heterogeneity and reinforcement can help identify portfolios that strengthen equity without necessarily reducing overall performance.

For practice, an integrated channel mix can serve as a pragmatic starting point. Equity guardrails can then help identify portfolios that erode minimum subgroup adoption or widen disparities. Among the strategies that pass these guardrails, conservative reallocations can offer budget-neutral equity gains. Class-tailored portfolios may provide a more favorable efficiency-equity profile where segmentation and media-buying capacity are available. These conclusions apply primarily to settings with pronounced media heterogeneity and digital divides. They should be interpreted in light of the model's structural assumptions, 1-year horizon, and reliance on matched Korean survey data. Future work should validate these strategic patterns in other health systems and link simulated diffusion to observed exposure and engagement in real campaigns. It should also extend guardrail-based planning tools to richer communication, organizational, and multiyear decision contexts.

Funding

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (grant 2019-0-01842, Artificial Intelligence Graduate School Program [GIST]).

Data Availability

Analysis code and minimal deidentified outputs sufficient to reproduce all results—including scenario definitions, calibrated parameters, and sensitivity configurations as reported in the Multimedia Appendices—are available on the Open Science Framework [62]. Original microdata from KMP 2022 and KNHANES 2022 [39,40] cannot be shared due to data use restrictions but are accessible directly from the data providers.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Additional methods.
[DOCX File, 35 KB - jmir_v28i1e88429_app1.docx]

Multimedia Appendix 2
Scenario-level adoption and equity metrics.
[DOCX File, 37 KB - jmir_v28i1e88429_app2.docx]

Multimedia Appendix 3
Sensitivity and robustness.
[DOCX File, 501 KB - jmir_v28i1e88429_app3.docx]

Multimedia Appendix 4
Calibration and validation diagnostics.
[DOCX File, 32 KB - jmir_v28i1e88429_app4.docx]

References

1. Jemna DV, David M, Bonnal L, Oros C. Socio-economic inequalities in the use of flu vaccination in Europe: a multilevel approach. Health Econ Rev 2024 Jul 31;14(1):61. [doi: 10.1186/s13561-024-00535-1] [Medline: 39083186]
2. Wang J, Tönnies T, Brinks R. Seasonal influenza vaccination coverage and the social determinants of influenza vaccination among people over 50 with diabetes in Europe: analyzing population-based SHARE data for the 2019-2020 and 2021-2022 influenza seasons. Vaccine 2025 Jan 25;45:126646. [doi: 10.1016/j.vaccine.2024.126646] [Medline: 39709804]
3. Altová A, Kulhánová I, Reisser K, et al. Educational inequalities in cervical cancer screening participation in 24 European countries. Public Health 2024 Aug;233:1-7. [doi: 10.1016/j.puhe.2024.04.036] [Medline: 38805834]
4. Willems B, Bracke P. The education gradient in cancer screening participation: a consistent phenomenon across Europe? Int J Public Health 2018 Jan;63(1):93-103. [doi: 10.1007/s00038-017-1045-7] [Medline: 29063122]
5. Jiang X, Wang L, Leng Y, et al. The level of electronic health literacy among older adults: a systematic review and meta-analysis. Arch Public Health 2024 Nov 7;82(1):204. [doi: 10.1186/s13690-024-01428-9] [Medline: 39511667]
6. Tsai YIP, Beh J, Ganderton C, Pranata A. Digital interventions for healthy ageing and cognitive health in older adults: a systematic review of mixed method studies and meta-analysis. BMC Geriatr 2024 Mar 4;24(1):217. [doi: 10.1186/s12877-023-04617-3] [Medline: 38438870]
7. Zhang C, Mohamad E, Azlan AA, Wu A, Ma Y, Qi Y. Social media and eHealth literacy among older adults: systematic literature review. J Med Internet Res 2025 Mar 26;27:e66058. [doi: 10.2196/66058] [Medline: 40138684]
8. Yang R, Gao S, Jiang Y. Digital divide as a determinant of health in the US older adults: prevalence, trends, and risk factors. BMC Geriatr 2024 Dec 21;24(1):1027. [doi: 10.1186/s12877-024-05612-y] [Medline: 39709341]
9. Xie L, Zhang S, Xin M, Zhu M, Lu W, Mo PKH. Electronic health literacy and health-related outcomes among older adults: a systematic review. Prev Med 2022 Apr;157:106997. [doi: 10.1016/j.ypmed.2022.106997] [Medline: 35189203]
10. Badr J, Motulsky A, Denis JL. Digital health technologies and inequalities: a scoping review of potential impacts and policy recommendations. Health Policy 2024 Aug;146:105122. [doi: 10.1016/j.healthpol.2024.105122] [Medline: 38986333]
11. Zhang E, Shang S, Xing Y, et al. Mapping behavioral and social drivers of influenza vaccine uptake in older adults: a scoping review. Vaccines (Basel) 2025 Jun 10;13(6):624. [doi: 10.3390/vaccines13060624] [Medline: 40573955]
12. Behavioural and Social Drivers of Vaccination: Tools and Practical Guidance for Achieving High Uptake: World Health Organization; 2022. URL: https://iris.who.int/server/api/core/bitstreams/c9f4a77a-001e-477d-a9d1-b05832820b64/content [accessed 2026-03-20]
13. Holtrop JS, Estabrooks PA, Gaglio B, et al. Understanding and applying the RE-AIM framework: clarifications and resources. J Clin Transl Sci 2021;5(1):e126. [doi: 10.1017/cts.2021.789] [Medline: 34367671]
14. Aravindhan A, Fenwick E, Chan AWD, et al. Nonadherence to diabetes complications screening in a multiethnic Asian population: protocol for a mixed methods prospective study. JMIR Res Protoc 2025 May 8;14:e63253. [doi: 10.2196/63253] [Medline: 40340926]
15. Hollands GJ, South E, Shemilt I, Oliver S, Thomas J, Sowden AJ. Methods used to conceptualize dimensions of health equity impacts of public health interventions in systematic reviews. J Clin Epidemiol 2024 May;169:111312. [doi: 10.1016/j.jclinepi.2024.111312] [Medline: 38432524]
16. Jolles MP, Fort MP, Glasgow RE. Aligning the planning, development, and implementation of complex interventions to local contexts with an equity focus: application of the PRISM/RE-AIM Framework. Int J Equity Health 2024 Feb 27;23(1):41. [doi: 10.1186/s12939-024-02130-6] [Medline: 38408990]
17. Welch V, Dewidar O, Ghogomu ET, et al. How effects on health equity are assessed in systematic reviews of interventions. Cochrane Database Syst Rev 2022 Jan 18;1(1):MR000028. [doi: 10.1002/14651858.MR000028.pub3] [Medline: 35040487]

XSL•FO
RenderX

18. Kite J, Chan L, MacKay K, et al. A model of social media effects in public health communication campaigns: systematic review. J Med Internet Res 2023 Jul 14;25:e46345. [doi: 10.2196/46345] [Medline: 37450325]
19. Zang S, Zhang X, Xing Y, Chen J, Lin L, Hou Z. Applications of social media and digital technologies in COVID-19 vaccination: scoping review. J Med Internet Res 2023 Feb 10;25:e40057. [doi: 10.2196/40057] [Medline: 36649235]
20. Massey PM, Murray RM, Chiang SC, Russell AM, Yudell MA. Social media, public health research, and vulnerability: considerations to advance ethical guidelines and strengthen future research. JMIR Public Health Surveill 2023 Dec 29;9:e49881. [doi: 10.2196/49881] [Medline: 38157235]
21. Alruwaili MM, Shaban M, Elsayed Ramadan OM. Digital health interventions for promoting healthy aging: a systematic review of adoption patterns, efficacy, and user experience. Sustainability 2023;15(23):16503. [doi: 10.3390/su152316503]
22. Hwang M, Lee S, Park GE, Park YH. Effectiveness of a digital health coaching self-management program for older adults living alone with multiple chronic conditions: a randomized controlled trial. Geriatr Nurs 2025;65:103509. [doi: 10.1016/j.gerinurse.2025.103509] [Medline: 40633238]
23. Kim KK, Backonja U. Digital health equity frameworks and key concepts: a scoping review. J Am Med Inform Assoc 2025 May 1;32(5):932-944. [doi: 10.1093/jamia/ocaf017] [Medline: 39936843]
24. Bitomsky L, Pfitzer E, Nißen M, Kowatsch T. Advancing health equity and the role of digital health technologies: a scoping review. BMJ Open 2025 Jun 26;15(6):e099306. [doi: 10.1136/bmjopen-2025-099306] [Medline: 40578855]
25. Karran EL, Cashin AG, Barker T, et al. Using PROGRESS-plus to identify current approaches to the collection and reporting of equity-relevant data: a scoping review. J Clin Epidemiol 2023 Nov;163:70-78. [doi: 10.1016/j.jclinepi.2023.09.017] [Medline: 37802205]
26. Anand S, World Health Organization. Measuring Health Workforce Inequalities: Methods and Application to China and India: World Health Organization; 2010. URL: https://iris.who.int/server/api/core/bitstreams/a2a1d045-37ab-46d2-b182-e80de35d52cd/content [accessed 2026-03-22]
27. Jiang S, Li B, Parkinson B, Li S, Gu Y. Aggregate distributional cost-effectiveness analysis: a novel tool for health economic evaluation to inform resource allocation. Glob Health Res Policy 2025 Apr 16;10(1):17. [doi: 10.1186/s41256-025-00415-z] [Medline: 40241192]
28. Colasanti R, MacLachlan A, Silverman E, et al. Using agent-based models to address non-communicable diseases: a review of models and their application to policy. The Lancet 2022 Nov;400:S33. [doi: 10.1016/S0140-6736(22)02243-7]
29. Sivakumar N, Mura C, Peirce SM. Innovations in integrating machine learning and agent-based modeling of biomedical systems. Front Syst Biol 2022;2:959665. [doi: 10.3389/fsysb.2022.959665]
30. Centola D. How Behavior Spreads: The Science of Complex Contagions: Princeton University Press; 2018. [doi: 10.2307/j.ctvc7758p]
31. Lindau ST, Makelarski JA, Kaligotla C, et al. Building and experimenting with an agent-based model to study the population-level impact of CommunityRx, a clinic-based community resource referral intervention. PLoS Comput Biol 2021 Oct;17(10):e1009471. [doi: 10.1371/journal.pcbi.1009471] [Medline: 34695116]
32. Sulis E, Terna P. An agent-based decision support for a vaccination campaign. J Med Syst 2021 Sep 28;45(11):97. [doi: 10.1007/s10916-021-01772-1] [Medline: 34581878]
33. Oană I, Hâncean MG, Perc M, et al. Online media use and COVID-19 vaccination in real-world personal networks: quantitative study. J Med Internet Res 2024 Oct 25;26:e58257. [doi: 10.2196/58257] [Medline: 39454189]
34. Centola D. The spread of behavior in an online social network experiment. Science 2010 Sep 3;329(5996):1194-1197. [doi: 10.1126/science.1185231] [Medline: 20813952]
35. Oh SS, Kim KA, Kim M, Oh J, Chu SH, Choi J. Measurement of digital literacy among older adults: systematic review. J Med Internet Res 2021 Feb 3;23(2):e26145. [doi: 10.2196/26145] [Medline: 33533727]
36. Wilensky U, Rand W. An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo: MIT Press; 2015.
37. Railsback SF, Grimm V. Agent-Based and Individual-Based Modeling: A Practical Introduction: Princeton University Press; 2019. [doi: 10.2307/jj.28274141]
38. Boccaletti S, Bianconi G, Criado R, et al. The structure and dynamics of multilayer networks. Phys Rep 2014 Nov 1;544(1):1-122. [doi: 10.1016/j.physrep.2014.07.001] [Medline: 32834429]
39. 2022 Korea Media Panel Survey. Korea Information Society Development Institute. 2022. URL: https://www.kisdi.re.kr/report/view.do?key=m2101113024973&arrMasterId=3934581&masterId=3934581&artId=1131416 [accessed 2026-02-18]
40. 2022 Korea National Health and Nutrition Examination Survey. Korea Disease Control and Prevention Agency. 2023. URL: https://knhanes.kdca.go.kr/knhanes/main.do [accessed 2026-03-22]
41. Gower JC. A general coefficient of similarity and some of its properties. Biometrics 1971 Dec;27(4):857-871. [doi: 10.2307/2528823]
42. Collins LM, Lanza ST. Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences: John Wiley & Sons; 2010. [doi: 10.1002/9780470567333]
43. Atkinson AB. On the measurement of inequality. J Econ Theory 1970 Sep;2(3):244-263. [doi: 10.1016/0022-0531(70)90039-6]

XSL•FO
RenderX

44. Borrell LN, Talih M. A symmetrized Theil index measure of health disparities: an example using dental caries in US children and adolescents. Stat Med 2011 Feb 10;30(3):277-290. [doi: 10.1002/sim.4114] [Medline: 21213344]
45. Gallagher KM, Updegraff JA. Health message framing effects on attitudes, intentions, and behavior: a meta-analytic review. Ann Behav Med 2012 Feb;43(1):101-116. [doi: 10.1007/s12160-011-9308-7] [Medline: 21993844]
46. Ainiwaer A, Zhang S, Ainiwaer X, Ma F. Effects of message framing on cancer prevention and detection behaviors, intentions, and attitudes: systematic review and meta-analysis. J Med Internet Res 2021 Sep 16;23(9):e27634. [doi: 10.2196/27634] [Medline: 34528887]
47. Fosdick BK, Larremore DB, Nishimura J, Ugander J. Configuring random graph models with fixed degree sequences. SIAM Rev 2018 Jan;60(2):315-355. [doi: 10.1137/16M1087175]
48. Newman MEJ, Strogatz SH, Watts DJ. Random graphs with arbitrary degree distributions and their applications. Phys Rev E 2001 Jul;64(2):026118. [doi: 10.1103/PhysRevE.64.026118]
49. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature New Biol 1998 Jun 4;393(6684):440-442. [doi: 10.1038/30918] [Medline: 9623998]
50. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev 1950 Jan;78(1):1-3. [doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2]
51. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning: Association for Computing Machinery; 2005:625-632. [doi: 10.1145/1102351.1102430]
52. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 2007 Mar;102(477):359-378. [doi: 10.1198/016214506000001437]
53. Law AM, Kelton WD. Simulation Modeling and Analysis, 4th edition: McGraw-Hill; 2007.
54. Grimm V, Berger U, DeAngelis DL, Polhill JG, Giske J, Railsback SF. The ODD protocol: A review and first update. Ecol Model 2010 Nov;221(23):2760-2768. [doi: 10.1016/j.ecolmodel.2010.08.019]
55. Saltelli A, Ratto M, Andres T, et al. Global Sensitivity Analysis: The Primer: John Wiley & Sons; 2007.
56. Bioethics and Safety Act. Ministry of Government Legislation (Korea). URL: https://www.law.go.kr/LSW/lsInfoP.do?lsiSeq=199534&viewCls=engLsInfoR&urlMode=engLsInfoR#0000 [accessed 2026-03-27]
57. Enforcement Rules of the Act on Bioethics and Safety. Ministry of Government Legislation (Korea). URL: https://www.law.go.kr/%EB%B2%95%EB%A0%B9/%C3%B5%B5%C0%B0%A8%B5%C0%B0%A8%B5%C0%B0%A8

Edited by J Piera-Jiménez; submitted 25.Nov.2025; peer-reviewed by WG Hussain, Z Liu; accepted 25.Feb.2026; published 01.Apr.2026.

Please cite as:
Lee J, Park J, Kim Y, Kong DJ
Channel Allocation and Equity in Preventive Campaigns for Older Adults: Agent-Based Modeling Study
J Med Internet Res 2026;28:e88429
URL: https://www.jmir.org/2026/1/e88429
doi:10.2196/88429

© Jihye Lee, Juyoung Park, Yuna Kim, Duk-Jo Kong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 1.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Retraction: Assessing the Causal Association Between COVID-19 and Graves Disease: Mendelian Randomization Study

JMIR Publications Editorial Office

JMIR Publications, 130 Queens Quay East, Suite 1100-1102, Toronto, ON, Canada

Corresponding Author:
JMIR Publications, 130 Queens Quay East, Suite 1100-1102, Toronto, ON, Canada

Related Article:

https://formative.jmir.org/2025/1/e66003

(J Med Internet Res 2026;28:e101185) doi:10.2196/101185

The JMIR Publications Editorial Office is retracting the article "Assessing the Causal Association Between COVID-19 and Graves Disease: Mendelian Randomization Study" [1] due to concerns regarding potential manipulation of the submission process and authorship, and the integrity of the peer-review process.

Author Hua Yu responded, indicating that there was an undisclosed conflict of interest with one peer reviewer relating to a previous copublication but denied any undisclosed contributions to the manuscript. The Editorial Office has determined that this does not fully resolve the concerns regarding the submission process and authorship, and that there is sufficient evidence indicating that the peer-review process was compromised to proceed with a retraction.

All of the authors agree with the decision to retract the article.

Reference

1. Nian H, Bai Y, Yu H. Retracted: assessing the causal association between COVID-19 and Graves disease: Mendelian randomization study. JMIR Form Res 2025 Apr 8;9:e66003. [doi: 10.2196/66003] [Medline: 40198900]

Submitted 12.May.2026; this is a non–peer-reviewed article;accepted 14.May.2026; published 12.Jun.2026.

Please cite as:
JMIR Publications Editorial Office
Retraction: Assessing the Causal Association Between COVID-19 and Graves Disease: Mendelian Randomization Study
J Med Internet Res 2026;28:e101185
URL: https://www.jmir.org/2026/1/e101185
doi:10.2196/101185

© JMIR Publications Editorial Office. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Retraction

Retraction: Adherence of Female Health Care Workers to the Use a Web-Based Tool for Improving and Modifying Lifestyle: Prospective Target Group Pilot Study

JMIR Publications Editorial Office[1]

JMIR Publications, Toronto, ON

Corresponding Author:
JMIR Publications
130 Queens Quay East, Suite 1100-1102
Toronto, ON
Phone: 1 4165832040
Email: ed-support@jmir.org

Related Article:

Retraction of: https://www.jmir.org/2020/8/e19500

(J Med Internet Res 2026;28:e101897) doi:10.2196/101897

The JMIR Publications Editorial Office is retracting the article "Adherence of Female Health Care Workers to the Use a Web-Based Tool for Improving and Modifying Lifestyle: Prospective Target Group Pilot Study" [1] due to concerns regarding the integrity of the peer-review process.

Author David Stubljar responded, indicating that there were undisclosed conflicts of interest with two peer reviewers relating to previous copublications. The Editorial Office has determined that there is sufficient evidence indicating that the peer-review process was compromised to proceed with a retraction.

Author David Stubljar agreed with the decision to retract the article; the other authors did not respond to the decision to retract the article.

Reference

1. Jukic T, Ihan A, Petek ?M, Strojnik V, Stubljar D, Starc A. Retracted: adherence of female health care workers to the use a web-based tool for improving and modifying lifestyle: prospective target group pilot study. J Med Internet Res 2020;22(8):e19500.

Submitted 20.May.2026; this is a non–peer-reviewed article;accepted 21.May.2026; published 16.Jun.2026.

Please cite as:
JMIR Publications Editorial Office
Retraction: Adherence of Female Health Care Workers to the Use a Web-Based Tool for Improving and Modifying Lifestyle: Prospective Target Group Pilot Study
J Med Internet Res 2026;28:e101897
URL: https://www.jmir.org/2026/1/e101897
doi:10.2196/101897
PMID:

©JMIR Publications Editorial Office. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Reporting Horizon Scanning Studies: Prototype Development Study

Sonia Garcia Gonzalez-Moral, MSc; Megan Fairweather, PhD; Sarah Khan, MD; Ross Fairbairn, MSc; Gill Norman, PhD

NIHR Innovation Observatory, Population Health Sciences Institute, Newcastle University, The Catalyst Room 3.12, 3 Science Square Newcastle Helix, Newcastle upon Tyne, England, United Kingdom

Corresponding Author:
Sonia Garcia Gonzalez-Moral, MSc
NIHR Innovation Observatory, Population Health Sciences Institute, Newcastle University, The Catalyst Room 3.12, 3 Science Square Newcastle Helix, Newcastle upon Tyne, England, United Kingdom

Abstract

Background: Horizon scanning identifies weak signals of innovation to anticipate future developments, providing strategic value for health care decision-making. Unlike evidence synthesis, it addresses emerging and uncertain areas but lacks standardized reporting guidance, limiting transparency, consistency, and impact. Inconsistent terminology and poorly described methods hinder comparability and uptake of findings.

Objective: This study aimed to develop a prototype reporting checklist and glossary to support structured, transparent, and reproducible reporting of horizon scanning methods in health care innovation.

Methods: A multidisciplinary working group of horizon scanning, evidence synthesis, and information specialists was convened at the National Institute for Health and Care Research Innovation Observatory, the UK national horizon scanning center, in May 2024. Using the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist as a framework, 4 workshops were held to adapt relevant reporting items and build a prototype checklist. Internal validation was conducted on 17 eligible Innovation Observatory reports (2017 - 2024), with scoring to assess item coverage. Items were then refined, classified as mandatory or optional, and finalized through consensus.

Results: Of 46 outputs screened, 17 (37%) met the inclusion criteria. In total, 42.9% (15/35) of the checklist items achieved 50% or more coverage; 57.1% (20/35) scored less than 50%. The final 35-item checklist (n=31, 88.6% single items and n=4, 11.4% multipart items) includes 28 (80%) mandatory and 7 (20%) optional items. Four novel components were introduced: interest holder description; horizon, innovation, population, data source, and interest holder scope framework; technology characteristics; and integration of the political, economic, social, technological, legal, environmental, and demographic factors framework.

Conclusions: This is the first checklist to standardize reporting in health technology horizon scanning. While this checklist prototype has been developed internally, external validation involving multidisciplinary experts via a Delphi process and a scoping review is planned. Its adoption can enhance transparency, reproducibility, and strategic impact, strengthening this methodological field.

(J Med Internet Res 2026;28:e88253) doi:10.2196/88253

KEYWORDS

horizon scanning; reporting checklist; health care innovations; medical technologies; reporting guidelines

Introduction

Background

Horizon scanning refers to the structured identification and analysis of weak signals, trends, drivers of change, wild cards, and discontinuities, enabling anticipation of future developments [1]. It belongs to the broader field of futures studies, which is concerned with exploring, modeling, and preparing for alternative future scenarios. Although horizon scanning has been widely adopted across diverse sectors, its application in the medical sciences is relatively recent [2]. The National Institute for Health and Care Research (NIHR) Innovation Observatory uses horizon scanning to systematically detect weak signals and trends of innovative health care technologies and support future decision-making processes about them [3].

Unlike evidence synthesis, understood as the process of systematically identifying, appraising, and integrating findings from multiple studies on a specific topic to generate comprehensive, reliable, and policy- or practice-relevant conclusions [4], horizon scanning focuses on the systematic

XSL•FO
RenderX

process of detecting and collecting early signs and weak signals of potentially important developments that, in this context, may be interpreted as innovative health care technologies [5]. However, horizon scanning currently lacks standardized reporting guidance, particularly for key aspects of the horizon scanning process such as the identification, selection, and prioritization of emerging health technologies. Furthermore, a recent review of horizon scanning practices in medical technologies, diagnostics, and digital health recommended the use of dedicated reporting standards in light of the inconsistent use of terminology in the field [6]. This lack of clarity impedes the dissemination and utility of horizon scanning outputs—such as published reports and research articles—and contributes to research inefficiencies, including duplication and underuse of findings [7].

Horizon scanning is more likely to be used to widely explore the "unknowns," being those more or less certain and definitively less defined, on emerging areas of innovation across diverse domains. Using the Rumsfeld matrix as a conceptual framework and using it to compare horizon scanning with other well-established methodologies such as evidence synthesis, we could argue that, while evidence synthesis is akin to examining the visible tip of an iceberg ("known knowns"), horizon scanning focuses on what lies beneath the surface ("unknown knowns" and "known unknowns") [8].

While horizon scanning is recognized as a methodological approach in its own right, it requires clear principles and structured processes to ensure rigor and consistency. Despite its growing use in health and policy contexts, the field remains underserved by methodological frameworks, reporting standards, or dedicated guidance. This absence of consensus risks limiting the transparency, comparability, and overall impact of horizon scanning outputs. Reporting guidelines specify a minimum set of elements that should be included in research publications and have been shown to enhance methodological transparency and promote the use of research findings [9]. To our knowledge, no standard reporting guidelines currently exist to support the transparent and consistent documentation of horizon scanning activities in the health care domain. Given the increasing role of horizon scanning in health care decision-making—both in the United Kingdom and internationally—this study aimed to address this gap by developing the first reporting checklist for horizon scanning methods in medical innovation.

Aim and Objectives

This study aimed to initiate the standardization of reporting practices for horizon scanning methods and tools, thereby enhancing the transparency and reproducibility of such approaches.

Our objectives included the development of a conceptual framework for horizon scanning research question formulation and query translation and the development of a preliminary checklist specifying essential and optional elements that should be transparently reported in horizon scanning outputs and a glossary to support consistent interpretation and reduce ambiguity in terminology. Further work will support finalization of the checklist, incorporating the experience of those who used this preliminary version.

Scope of This Checklist

This checklist is intended for application to horizon scanning outputs, including both articles published in academic journals and reports published on institutional websites. The checklist is developed to provide detailed and transparent information about the application of horizon scanning methods to medical health technologies. Intended users of this checklist include authors of such reports, peer reviewers of manuscripts submitted to journals for publication, and journal editors when assessing such manuscripts and methods.

Methods

Overview

A working group consisting of 4 horizon scanning specialists, 2 evidence synthesis specialists, and 2 information specialists was convened at the NIHR Innovation Observatory in May 2024. Participants were selected based on demonstrable expertise in horizon scanning, evidence synthesis methodology, and information science. Their disciplinary backgrounds spanned medical sciences, biosciences, and information science, and their professional experience included conducting systematic reviews, advanced information retrieval, and horizon scanning across medicines, medical devices, diagnostics, and digital health technologies. The group reflected methodological and professional diversity, encompassing varied levels of seniority and experience working on projects for multiple stakeholder groups. All members had established publication records in relevant methodological or applied domains, providing evidence of recognized subject matter expertise.

Four structured workshops, each lasting approximately 2 hours, were conducted between May 2024 and October 2024 to define objectives, iteratively refine methodological components, and develop the checklist. Prior to these sessions, an initial planning meeting was held in which the group established the study aims, scope, procedural approach, and decision-making processes that would guide the workshop series. During the workshops, proposed items and revisions were discussed collectively, and decisions were reached through structured deliberation, with agreement determined either through consensus where possible or, when consensus was not achievable, through majority agreement informed by methodological rationale and supporting evidence from the literature.

Given the similarity of broad scoping reviews with horizon scanning methods, we decided to use the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) as a framework to develop a checklist for reporting items relevant to horizon scanning reports [9]. During the first and second workshops, the group assessed the relevance of each PRISMA-ScR checklist item against the methodological and topic characteristics of horizon scanning reports and produced a modified checklist prototype for horizon scanning studies.

For the internal validation of the prototype, we applied the checklist to outputs produced by the Innovation Observatory between 2017 and 2024. Each team member in the working group was assigned a set of randomly selected reports to screen

XSL•FO
RenderX

for suitability and, if applicable, score those reports that used horizon scanning methods for the identification of innovative health care technologies against our checklist prototype. Following the screening process, 17 reports were deemed suitable for scoring. The rest were excluded as they were not horizon scanning reports. Each checklist item was marked as "yes" (included), "no" (not included), or not applicable ("N/A") if the item was not relevant to that specific paper. Following completion of the scoring exercise, one team member (MF) combined the scores for each item and calculated a final percentage score by dividing the score by the maximum possible score ("N/A" items were not included in the maximum score).

In the third and fourth workshops, the working group reviewed the scores, proposed additional customization of items to address issues relevant to horizon scanning based on the scoring exercise, and achieved consensus on whether the items should be made mandatory or optional. The resulting checklist prototype was uploaded to the Open Science Framework for dissemination.

Ethical Considerations

This study did not involve human participants, patient data, or identifiable personal information. The work consisted solely of methodological development activities conducted by professionals in their occupational roles and analysis of publicly available institutional reports. According to institutional and national research governance guidance, this type of study does not constitute human subjects research and therefore did not require review or approval by an institutional review board or research ethics committee [10].

All procedures were conducted in accordance with applicable institutional, national, and international standards for research integrity and publication ethics [11].

Results

Overview

A total of 46 outputs published between 2017 and 2024 were downloaded from the NIHR Innovation Observatory website by a team member (Oleta Williams). In total, 37.0% (n=17) of the horizon scanning reports met the criteria for inclusion in this assessment and were assessed against the checklist. The remaining reports were not horizon scanning reports and were not suitable for this assessment. The reasons for exclusion were being only an abstract, a methods paper, or a presentation (not a report). Figure 1 presents the PRISMA flowchart for the identification of studies.

A total of 42.9% (15/35) of the items achieved a final score between 50% and 100%, whereas 57.1% (20/35) of the items scored under 50%. Table 1 presents a breakdown of these items by their total final score.

Next, based on the working group members' experience, and in line with the objectives of this study, the items were rated through agreement as optional or mandatory. An optional item meant that the information required was not always possible to disclose due to a number of reasons that include technology stage of development, stakeholder agreements, availability of the information, or relevancy of the item in the context of the horizon scanning scope. These items start with an "If appropriate" statement to indicate their optionality. A mandatory item was deemed of the utmost importance for the conduct of horizon scanning methods and interpretation of results. These needed to be present at all times to achieve transparency in the retrieval and filtration of the signals and replicability of methods, avoid bias in the results, and achieve accountability and impact and overall increased rigor in the final horizon scanning output.

The resulting checklist contains 35 items (n=31, 88.6% single items and n=4, 11.4% that include multiple subitems), of which 28 (80%) were deemed mandatory and 7 (20%) were deemed optional (items 15, 16, 17b, 19, 20, 21, and 23). Some subsections, such as "Synthesis methods" (items 14a and 14b) in the "Methods" section and "Weak signal selection" (items 17a and 17b) in the "Results" section, present more than 1 item per section, with only item 17b considered optional. The sections "Discussion" (items 24a to 24d) and "Registration and protocol" (items 25a to 25c) also include multiple subitems and are all considered mandatory.

This checklist includes 4 new essential components not found in the PRISMA-ScR checklist. These are "Interest holder description" item 4 in the "Introduction" section; "Scope" item 6 in the "Methods" section; "Technology characteristics" item 19 in the "Results" section; and the introduction of the political, economic, social, technological, legal, environmental, and demographic factors (PESTLED) framework in item 24a in the "Discussion" section.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram depicting included and excluded reports. IO: Innovation Observatory; NIHR: National Institute for Health and Care Research.

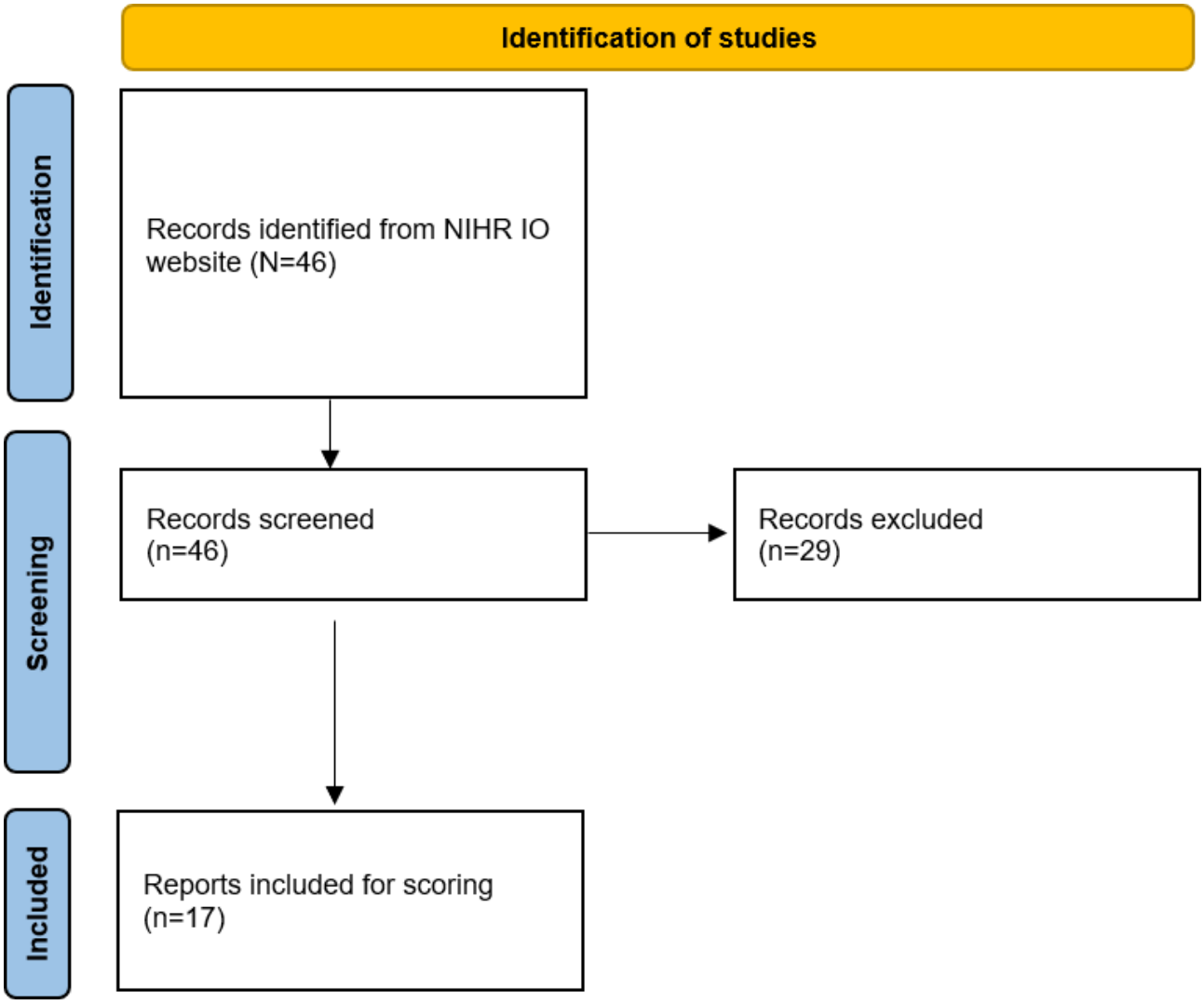

XSL•FO
RenderX

Table . Scoring checklist items by the number of times the item was present in the horizon scanning reports assessed.

Checklist item name (number)	Total final score (%)
Items that scored >50%	
Rationale (3)	100
Eligibility criteria (7)	100
Single source of information characteristics (18)	100
Discussion (24a)	100
Information sources (8)	94
Support (26)	94
Results by intended purpose or outcome (21)	90
Technology characteristics (19)	88
Discussion (24d)	88
Weak signal selection (17a)	81
Objectives (5)	76
Interest holder (4)	63
Data items (12)	60
Title (1)	53
Competing interests (27)	50
Items that scored <50%	
Discussion (24b)	46
Selection process (10)	46
Abstract (2)	41
Discussion (24c)	38
Search strategy (9)	38
Scope (6)	38
Availability of data, code, and other materials (28)	36
Data collection process (11)	33
Synthesis methods (14b)	33
Synthesis methods (14a)	30
Reporting bias assessment (15)	20
Weak signal selection (17b)	9
Effect measures (13)	0
Certainty assessment (16)	0
Risk of bias in studies (20)	0
Reporting missing data (22)	0
Certainty of evidence (23)	0
Registration and protocol (25a, 25b, and 25c)	0

Interest Holder Description

In the context of health technology horizon scanning, an interest holder (often called a stakeholder) refers to any individual, group, or organization that has an interest in or may be affected by the development, adoption, use, or regulation of a health technology.

Key types of interest holders in this context may be the following:

- Patients and patient advocacy groups—interested in access to and the effectiveness and safety of new and innovative health technologies
- Clinicians and health care providers—interested in clinical utility, integration into practice, and outcomes

XSL•FO
RenderX

- Health technology developers and manufacturers—concerned with market potential, regulation, and innovation uptake
- Regulators and health authorities—interested in safety, efficacy, and compliance with regulatory standards or the early identification of gaps in current regulatory frameworks
- Policymakers and health system planners—focused on broader system impact and public health implications
- Researchers and academics—interested in evidence generation, gaps, and methodological development
- Health technology assessment (HTA) bodies—use horizon scanning to inform evaluation and prioritization

Interest holders contribute by identifying unmet needs and emerging technologies, understanding key eligibility criteria based on needs or gaps, informing prioritization criteria, providing early feedback on clinical relevance and implementation barriers, and supporting evidence appraisal and decision-making. Their involvement ensures that the scanning process is contextually relevant, anticipatory, and aligned with real-world health system priorities.

Horizon Scan Scope: Horizon, Innovation, Population, Data Source, and Interest Holder Framework for Scoping Health Technology Horizon Scans

To support interpretability and decision relevance, horizon scanning reports should explicitly describe the contextual parameters that shape the scan, including its purpose, decision setting, intended users, policy or clinical domain, and analytic perspective. This is the purpose of applying the horizon, innovation, population, data source, and interest holder (HIP-D/I) scoping framework, which provides a structured approach to define the scope of horizon scanning projects targeting innovative health care technologies through the following 5 key dimensions. Clear specification of these elements enables readers to understand how signals were identified, interpreted, and prioritized and supports appropriate comparison across horizon scanning outputs:

- Horizon—horizon scanning studies should define their temporal scope (eg, short-, medium-, or long-term horizon) as part of the methodological description and within item 6 (“Scope”) of the checklist through the HIP-D/I framework as the selected time horizon influences the types of signals identified, their interpretation, and their relevance for decision-making contexts. Time horizon is categorized based on technologies’ developmental proximity to market launch: emerging (early-stage concepts), transitional (in precommercial or pilot phases), and imminent (near market entry and postmarket surveillance). This helps determine the level of readiness, the sources that should be searched based on the maturity of the evidence, and the time frame for potential impact.
- Innovation—focuses on identifying what is novel or disruptive about the technology compared to current standards of care. This includes technological, procedural, or conceptual innovations that may transform clinical practice or health system delivery.
- Population—specifies the target patient group or population segment that the technology is intended to benefit. Clarifying the population helps align horizon scanning priorities with unmet clinical needs and equity considerations.
- Data sources—emphasizes the importance of identifying weak signals of innovation. Given the developmental stage of many technologies, relevant data may be found in nontraditional sources such as preprints, patents, clinical trial registries, venture capital investments, and early-stage conference presentations. Data sources and horizon are intrinsically related.
- Interest holder—as defined above, they provide a contextual element to the results of the horizon scanning, enabling greater and immediate impact.

Together, these 5 dimensions support a systematic scoping process that enables transparent and reproducible methods for the identification and prioritization of transformative health technologies.

Technology Characteristics

In horizon scanning studies of health care innovation, providing a detailed description of the identified, filtered, and prioritized technologies is essential to ensure that outputs are meaningful and actionable for decision-makers. Unlike systematic reviews—where the unit of analysis is typically the study and its outcomes—horizon scanning primarily focuses on the identification of innovative technologies and often uses the technology and its characteristics (eg, trial phase or technology readiness levels) as the primary unit of interest. Detailed technology descriptions capture critical attributes such as intended use, mechanism of action, stage of development, anticipated benefits, implementation challenges, and potential impact on health systems.

This level of detail serves several purposes: it supports informed prioritization, facilitates future assessments (eg, HTA or regulatory review), enhances transparency and reproducibility, and enables strategic planning for health system readiness.

To align with these goals, the reporting checklist was adapted to shift the focus from evidence appraisal, as seen in the PRISMA-ScR checklist, to the detection of weak signals, in keeping with the purpose and definition of horizon scanning [12]. Table 2 presents the prototype checklist, including item descriptions and corresponding document sections.

XSL•FO
RenderX

Table . Final checklist items

Section and subsection	Item number	Item description
Title		
Not applicable	1	Identify the report as a horizon scan or other alternative terms used to refer to horizon scanning outputs, for example, landscape or pipeline analysis or report or broader umbrella terms such as foresight analysis
Abstract		
Not applicable	2	See the adaptation of the PRISMA[a] 2020 for Abstracts checklist for horizon scanning reports
Rationale	3	Describe the rationale for the horizon scan in the context of existing knowledge, unmet needs, policy, or emerging innovation
Introduction		
Interest holders	4	Disclosure of interest holder names and brief function or role in decision-making
Objectives	5	Provide an explicit statement of the objectives or questions that the horizon scan addresses
Methods		
Scope	6	Consider using the HIP-D/I[b] framework
Eligibility criteria	7	Specify the inclusion and exclusion criteria for the horizon scan
Information sources	8	Specify all databases, registers, websites, organizations, reference lists, and PPIE[c] and expert solicitation and consultation, as well as any other sources searched or consulted to identify signals; specify the date when each source was last searched or consulted
Search strategy	9	Present the full search strategies for all databases, registers, and websites, including any filters and limits used; report if any AI[d] tool was used for creating or running the searches and provide the name and version number of the tool
Selection process	10	Specify the methods used to decide whether a signal met the inclusion criteria of the horizon scan, including how many reviewers screened each record and each report retrieved; whether they worked independently; and, if applicable, details of automation tools used in the process
Data collection process	11	Specify the methods used to collect data from reports, including how many reviewers collected data from each report; whether they worked independently; any processes for obtaining or confirming data from third parties; and, if applicable, details of automation tools used in the process
Data items	12	List and define all data points sought; describe any assumptions made about any missing or unclear information
Effect measures	13	If appropriate to the technology readiness level or technology type (medicine, diagnostic tests, medical devices, or digital interventions), specify for each outcome the effect or performance measures (eg, risk ratio, mean difference, sensitivity or specificity, and ROC[e] curve) used in the presentation of results

XSL•FO
RenderX

Section and subsection	Item number	Item description
Synthesis methods	14a	Describe the processes used to decide which signals were eligible for each analysis (checklist item 7)
Synthesis methods	14b	Describe any methods required to prepare the data for presentation or analyses, such as handling of missing data or data conversions
Reporting bias assessment	15	If appropriate to the technology readiness level or technology type (medicine, diagnostic tests, medical devices, or digital interventions), state if a risk-of-bias assessment of the studies used for identification of signals will be performed and which tools will be used
Certainty assessment	16	If appropriate to the technology readiness level or technology type (medicine, diagnostic tests, medical devices, or digital interventions), describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome using the GRADE[f] approach
Results		
Weak signal selection	17a	Describe the results of the search and selection process, from the number of records identified in the search to the number of technologies included in the horizon scan, ideally using a flow diagram
Weak signal selection	17b	If appropriate, include a list in the appendix or a separate tab of technologies that might appear to meet the inclusion criteria but that were excluded and explain why they were excluded; these may be referred to as “other interesting technologies” that do not meet the criteria pre-established by interest holders but present some innovative aspects of interest
Single source of information characteristics	18	If appropriate, cite each included single source of information and present its characteristics
Technology characteristics	19	If appropriate for the scan scope, list the technology characteristics
Risk of bias in studies	20	If appropriate to the technology readiness level or technology type (medicine, diagnostic tests, medical devices, or digital interventions), present assessments of risk of bias for each included study
Results by intended purpose or outcome	21	If appropriate, present analyses by outcome, signal, and/or technology
Reporting missing data	22	Present missing data and how they have been dealt with in the analysis
Certainty of evidence	23	If appropriate to the technology readiness level or technology type (medicine, diagnostic tests, medical devices, or digital interventions), present assessments of certainty (or confidence) in the body of evidence for each outcome assessed using the GRADE approach
Discussion		
Not applicable	24a	Provide a general interpretation of the results in the context of other PESTLED[g] factors
Not applicable	24b	Discuss any limitations of the data included in the horizon scan

XSL•FO
RenderX

Section and subsection	Item number	Item description
Not applicable	24c	Discuss any limitations of the horizon scanning processes used (including language barriers—eg, only English language considered)
Not applicable	24d	Discuss the implications of the results for practice, policy, and future research
Other information		
Registration and protocol	25a	Provide registration information for the review, including register name and registration number, or state that the review was not registered
Registration and protocol	25b	Indicate where the review protocol can be accessed or state that a protocol was not prepared
Registration and protocol	25c	Describe and explain any amendments to the information provided at registration or in the protocol
Support	26	Describe sources of financial or nonfinancial support for the review and the role of the funders or sponsors in the review
Competing interests	27	Declare any competing interests of review authors
Availability of data, code, and other materials	28	Subject to agreement with the stakeholders, report which of the following are publicly available and where they can be found: template data collection forms, data extracted from the included studies, data used for all analyses, analytic code, and any other materials used in the review

[a]PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

[b]HIP-D/I: horizon, innovation, population, data source, and interest holder.

[c]PPIE: public and patient involvement and engagement.

[d]AI: artificial intelligence.

[e]ROC: receiver operating characteristic.

[f]GRADE: Grading of Recommendations Assessment, Development, and Evaluation.

[g]PESTLED: political, economic, social, technological, legal, environmental, and demographic factors.

PESTLED Framework

Incorporating the PESTLED framework into the discussion section of a horizon scan on emerging health care technologies enriches the report by situating innovations within a structured macroenvironmental context, thereby highlighting how regulatory dynamics, economic constraints, social and demographic values, and environmental considerations intersect with technological development and diffusion. For instance, a strategic environmental scan of Iranian public hospitals demonstrated that PESTLED-guided analysis enabled the identification of external pressures, including economic sanctions, shifts in disease patterns, and governance issues, which informed adaptive strategies for equitable resource allocation and service delivery [13]. Similarly, scoping reviews applying PESTLED-type analysis to electronic health record implementation revealed how political, economic, sociocultural, and legal dimensions shape adoption trajectories and reveal implementation challenges in low- and middle-income settings [14]. By framing horizon scan findings within the 7 domains of this framework, authors can present a reproducible and comparative analytic lens that directly links emerging technologies to broader system drivers; aids in anticipating barriers to and enablers of adoption; and enhances the strategic relevance of the report for policymakers, funders, and health system planners.

The full proposed checklist is shown in Table 2 and Checklist 1.

The checklist is available for download [15]. A full item-by-item description with text excerpts as examples is provided in Checklist 2. Checklist 1 provides the checklist form for completion at the reporting writing stage. A glossary of terms for ease of interpretation of the checklist items is provided in Checklist 3.

Discussion

Principal Findings

This study introduced what is, to the best of our knowledge, the first horizon scanning checklist designed to enhance transparency, reproducibility, and methodological rigor in the reporting of health technology horizon scanning studies. It presents the results of preliminary work involving analysis of our own reports and in-depth discussions among expert practitioners within the UK horizon scanning center. The

XSL•FO
RenderX

checklist outlines identified key features that should be systematically reported to enable comprehensive interrogation and interpretation of intelligence derived from horizon scanning. When our reports were assessed against the checklist, just over one-third of the items (15/35, 42.9%) achieved a final score above 50%, whereas the remaining 57.1% (20/35) of the items were insufficiently reported. This variation in reporting highlights a lack of standardization in current horizon scanning outputs and demonstrates the need for a structured reporting framework. Six checklist items (13, 16, 20, 22, 23, and 25) received a score of 0, meaning that they were not reported in any of the included reports assessed. Some of these items might not be relevant to all technology readiness levels and should not always be reported. For example, a horizon scan in the early horizon that uses patents and preclinical studies to identify trends and signals of innovative technologies will not present sufficient data on outcomes or the strength of the evidence to warrant risk-of-bias or certainty assessments. Furthermore, although it is strongly recommended that a horizon scan follow a robust, pre-agreed protocol, the highly strategic nature of some horizon scans makes a publicly available protocol unfeasible due to confidentiality agreements with stakeholders. Notably, many of these features are already represented in reports from the UK national horizon scanning center, highlighting both their relevance and feasibility. This finding demonstrates the need for more consistent and comprehensive reporting across all horizon scanning studies to support full transparency, comparability, and utility of the findings.

Key strengths observed across well-reported items included the consistent articulation of rationale, eligibility criteria, and technology characteristics, crucial elements for supporting decision-making. However, core methodological domains such as search strategy, selection process, synthesis methods, and discussion of limitations were frequently underreported or absent. These gaps limit the transparency and replicability of horizon scanning studies, potentially reducing their utility for decision-makers involved in regulatory, reimbursement, or research prioritization pathways.

To address these issues, the study team developed a prototype 35-item checklist (including multicomponent subitems), of which 28 (80%) were designated as mandatory and 7 (20%) were designated as optional. These designations were based on the expert consensus of internal team members and reflected both practical constraints and methodological requirements. It is worth stressing that the checklist introduces 4 essential components not found in existing reporting frameworks such as the PRISMA-ScR: interest holder description, scope (HIP-D/I framework), technology characteristics, and PESTLED framework. These additions are fundamental to horizon scanning, which (unlike systematic reviews) uses the technology as the unit of analysis rather than the study or its outcomes. A detailed description of the identified, filtered, and prioritized technologies is critical to enable strategic planning, future assessments, and timely decision-making.

Furthermore, this checklist reflects the exploratory and anticipatory nature of horizon scanning. The emphasis is on the detection of weak signals, in alignment with the goals of identifying transformative technologies ahead of market entry. This reframing supports the unique objectives of horizon scanning, particularly in cases in which conventional data may be unavailable and insights must be drawn from diverse and sometimes non–peer-reviewed sources such as clinical trial registries, patent databases, preprints, and expert consultation.

Horizon scanning is increasingly used by a range of stakeholders (regulators, HTA bodies, policymakers, clinicians, and technology developers) as a forward-looking, market intelligence, and decision support tool [16-19]. Given the variability in how horizon scanning is defined and implemented across organizations, the proposed checklist may provide a common foundation for improving the transparency, comparability, and rigor of horizon scanning practices in health care innovation.

Our study has several limitations. First, the development of the checklist prototype relied exclusively on internal experts from the NIHR Innovation Observatory. While the team possesses extensive experience in conducting horizon scans across a variety of health technologies, stakeholders, and time horizons, we recognize that this internal scope may limit the generalizability of the findings. To address this, we plan to undertake a formal external validation using a scoping review and a multistage adapted Delphi process to achieve broader consensus on item relevance, structure, and applicability across diverse horizon scanning contexts. Second, the checklist scoring was based on a relatively small sample of 17 reports published since 2017. Given the niche nature of horizon scanning in health care, this limited sample size may have influenced the ranking of optional vs mandatory items. Future application of the checklist to a larger set of outputs will help mitigate this constraint and may refine item prioritization. Finally, wild cards did not appear in the analyzed reports because the sample consisted of applied horizon scanning outputs focused on near- to midterm health technologies, where evidence sources and stakeholder priorities typically emphasize plausible and actionable innovations rather than low-probability, high-impact events. Therefore, their absence reflects the empirical characteristics and scope of the sampled material, not a conceptual exclusion from the checklist framework, which is designed to accommodate any signal type if reported but more specifically weak signals and trends.

Conclusions and Next Steps

This study presents the development of a prototype checklist designed to improve the transparency, consistency, and methodological clarity of horizon scanning reports for health care innovation. The checklist was created in response to a recognized gap in reporting standards [6], drawing on elements from the PRISMA-ScR framework [9] to accommodate the broad and exploratory nature of horizon scanning studies, similar to those of broad scoping reviews [20].

The checklist provides a structured foundation for reporting horizon scanning outputs, facilitating more robust and reproducible methods, and supporting stakeholder engagement in early decision-making processes. Its adoption has the potential to contribute to the standardization of horizon scanning practices across organizations, enhance the strategic value of these outputs

XSL•FO
RenderX

in innovation management, and provide horizon scanning methods of greater scientific rigor.

Next steps include conducting an external validation using a multistage Delphi process following a rapid scoping review involving a diverse group of horizon scanning experts and end users. This will help refine and potentially increase the checklist items, confirm their relevance across different settings, and promote broader adoption. If widely implemented, this checklist can serve as a key tool for strengthening the scientific basis of horizon scanning and advancing its role in health policy and innovation planning.

Acknowledgments

The authors would like to acknowledge Oleta Williams for her valuable contribution as an information specialist, particularly for her input into search strategy development and methodological discussions that informed this research.

Funding

This study and project are funded by the National Institute for Health and Care Research (Innovation Observatory and project reference HSRIC-2016-10009). The views expressed are those of the authors and not necessarily those of the National Institute for Health and Care Research or the Department of Health and Social Care.

Conflicts of Interest

None declared.

Checklist 1
Checklist of reporting items for horizon scanning studies (version 1.4).
[DOCX File, 40 KB - jmir_v28i1e88253_app1.docx]

Checklist 2
Checklist of reporting items for horizon scanning studies: Examples.
[DOCX File, 166 KB - jmir_v28i1e88253_app2.docx]

Checklist 3
Checklist items glossary of terms.
[DOCX File, 22 KB - jmir_v28i1e88253_app3.docx]

References

1. Saritas O, Smith JE. The Big Picture – trends, drivers, wild cards, discontinuities and weak signals. Futures 2011 Apr;43(3):292-312. [doi: 10.1016/j.futures.2010.11.007]
2. Meskó B, Kristóf T, Dhunnoo P, Árvai N, Katonai G. Exploring the need for medical futures studies: insights from a scoping review of health care foresight. J Med Internet Res 2024 Oct 9;26:e57148. [doi: 10.2196/57148] [Medline: 39383528]
3. NIHR Innovation Observatory. URL: https://io.nihr.ac.uk [accessed 2026-03-19]
4. Higgins JP, Thomas J, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions Version 6.5: Cochrane Collaboration; 2024.
5. Amanatidou E, Butter M, Carabias V, et al. On concepts and methods in horizon scanning: lessons from initiating policy dialogues on emerging issues. Sci Public Policy 2012 Mar 31;39(2):208-221. [doi: 10.1093/scipol/scs017]
6. Garcia Gonzalez-Moral S, Beyer FR, Oyewole AO, Richmond C, Wainwright L, Craig D. Looking at the fringes of MedTech innovation: a mapping review of horizon scanning and foresight methods. BMJ Open 2023 Sep 14;13(9):e073730. [doi: 10.1136/bmjopen-2023-073730] [Medline: 37709340]
7. Glasziou P, Chalmers I. Research waste is still a scandal—an essay by Paul Glasziou and Iain Chalmers. BMJ 2018;363:k4645. [doi: 10.1136/bmj.k4645]
8. Garvey B, Humzah D, Le Roux S. Uncertainty Deconstructed: A Guidebook for Decision Support Practitioners: Springer; 2022.
9. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
10. Newcastle University research ethics policy. : Newcastle University; 2024 URL: https://www.ncl.ac.uk/mediav8/our-research/research-governance-policies/Research%20Ethics%20Policy.pdf [accessed 2025-05-01]
11. The concordat to support research integrity. : The UK Committee on Research Integrity; 2025 URL: https://ukcori.org/wp-content/uploads/2025/12/The-Concordat-to-Support-Research-Integrity-2025.pdf [accessed 2026-04-07]

XSL•FO
RenderX

12. Hines P, Hiu Yu L, Guy RH, Brand A, Papaluca-Amati M. Scanning the horizon: a systematic literature review of methodologies. BMJ Open 2019 May 27;9(5):e026764. [doi: 10.1136/bmjopen-2018-026764] [Medline: 31133588]
13. Pourmohammadi K, Bastani P, Shojaei P, Hatam N, Salehi A. A comprehensive environmental scanning and strategic analysis of Iranian Public Hospitals: a prospective approach. BMC Res Notes 2020 Mar 26;13(1):179. [doi: 10.1186/s13104-020-05002-8] [Medline: 32216824]
14. Mwogosi A, Mambile C. Applying PEST analysis to evaluate the adoption of electronic health records in Tanzanian healthcare: a scoping review. Digit Health 2025 Apr 21;11:20552076251334029. [doi: 10.1177/20552076251334029] [Medline: 40297372]
15. Gonzalez-Moral SG, Khan SK, Fairbairn R, Williams O, Fairweather M, Norman G. Developing a checklist for transparent reporting of items in horizon scanning studies. OSF. 2025. URL: https://osf.io/m39vw/overview [accessed 2026-03-20]
16. Horizon scanning and demand signalling. NHS England. URL: https://www.england.nhs.uk/aac/what-we-do/demand-signalling [accessed 2026-03-19]
17. McCool K. UK to accelerate patient access to medicines under new MHRA–NICE joint approach. Becaris Publishing. 2025. URL: https://becarispublishing.com/digital-content/blog-post/uk-accelerate-patient-access-medicines-under-new-mhra-nice-joint-approach [accessed 2026-03-19]
18. SPS horizon scanning. Specialist Pharmacy Service. 2024. URL: https://www.sps.nhs.uk/articles/sps-horizon-scanning/ [accessed 2026-03-19]
19. Seeing the big picture: the role of horizon scanning and strategic foresight in SMEs. Future Finance. 2025. URL: https://www.future-finance.tech/post/seeing-the-big-picture-the-role-of-horizon-scanning-and-strategic-foresight-in-smes [accessed 2026-03-19]
20. Akartuna EA, Johnson SD, Thornton A. Enhancing the horizon scanning utility of futures-oriented systematic and scoping reviews. Futures 2024 Apr;158:103340. [doi: 10.1016/j.futures.2024.103340]

Abbreviations

HIP-D/I: horizon, innovation, population, data source, and interest holder
HTA: health technology assessment
NIHR: National Institute for Health and Care Research
PESTLED: political, economic, social, technological, legal, environmental, and demographic factors
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by B Mesko, T Kristóf; submitted 21.Nov.2025; peer-reviewed by É Hideg; revised version received 23.Feb.2026; accepted 24.Feb.2026; published 16.Apr.2026.

Please cite as:
Garcia Gonzalez-Moral S, Fairweather M, Khan S, Fairbairn R, Norman G
Reporting Horizon Scanning Studies: Prototype Development Study
J Med Internet Res 2026;28:e88253
URL: https://www.jmir.org/2026/1/e88253
doi:10.2196/88253

© Sonia Garcia Gonzalez-Moral, Megan Fairweather, Sarah Khan, Ross Fairbairn, Gill Norman. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Forecasting the Impacts of Artificial Intelligence Assistance in Virtual Consultations for Chronic Obstructive Pulmonary Disease: Exploratory Futures Wheel Study

Pranavsingh Dhunnoo[1,2], MSc, MD; Bertalan Meskó[2], MD, PhD; Karen McGuigan[3], PhD; Vicky O'Rourke[4], EdD; Michael McCann[1], PhD

[1]Department of Computing, Atlantic Technological University, CoLab Building, ATU, Port Road, Letterkenny, County Donegal, Ireland

[2]The Medical Futurist Institute, Budapest, Hungary

[3]School of Nursing and Midwifery, Queen's University Belfast, Belfast, Northern Ireland, United Kingdom

[4]Faculty of Business, Atlantic Technological University, Letterkenny, County Donegal, Ireland

Corresponding Author:
Pranavsingh Dhunnoo, MSc, MD
Department of Computing, Atlantic Technological University, CoLab Building, ATU, Port Road, Letterkenny, County Donegal, Ireland

Abstract

Background: While digital health technologies promise to reshape the medical journey, their potential might not be realized due to unforeseen implementation challenges. Notably, the future impact of artificial intelligence (AI) in virtual consultations has been poorly investigated.

Objective: This study aims to explore, across 8 areas, the future impacts of a bespoke, co-designed AI tool for remote chronic obstructive pulmonary disease care from the perspectives of patients and health care professionals (HCPs) with the Futures Wheel (FW) method. It provides practical recommendations for conducting FW activities involving novel digital health tools.

Methods: A pilot FW workshop was conducted with public and patient involvement members to gather feedback on the process. Subsequently, an exploratory, in-person FW workshop was conducted with 2 patients with chronic obstructive pulmonary disease and 2 HCPs who had previously been involved in the co-design of the bespoke AI tool. The central statement was as follows: "The bespoke AI tool is used in every virtual consultation." Participants identified first- and second-order consequences across the following 8 areas of impact: HCP-patient relationship impact, psychological impact, social impact, educational impact, legal impact, ethical impact, health care delivery impact, and technology impact. Each participant discussed their individual input to provide additional context.

Results: Regarding the HCP-patient relationship, patients foresee the tool's impact as redefining the remote care dynamic with enhanced patient involvement, while HCPs identify its meaningful communication assistance. On the psychological impact, patients expect an enhanced level of empowerment and confidence, and HCPs anticipate improved understanding of patients' emotional well-being with the AI tool's assistance. As for social impacts, patients view the AI support as beneficial for social patient-HCP interactions, and HCPs foresee their workflow being enhanced with flexibility and collaboration. The AI's educational impacts are expected to include, from patients' perspectives, better familiarization of HCPs with individual patient cases and, from HCPs' perspectives, improved support for training, upskilling, and administrative tasks. On the legal front, patients identify limited risks associated with the tool, and HCPs expect its features to lead to safer practices, contingent on regulatory compliance. Provided integrity and ethical use, the tool's ethical impact is not perceived as significant by patients, while HCPs see its personalized features as leading to fair, individual remote assessments. Patients envision the AI tool's impact on health care delivery as fostering patient-centricity, and HCPs anticipate strengthened remote care processes. Technologically, patients forecast a significant improvement to the current system, requiring adequate investment and resources, while HCPs expect complementarity between human input, AI, and the current system.

Conclusions: The plausible AI-driven future of remote chronic care is a nuanced one. The FW method indicated that a bespoke, co-designed AI tool can positively support virtual care delivery and remote interactions while indicating potential risks. These insights can inform strategies related to early planning, governance, and implementation considerations.

(J Med Internet Res 2026;28:e90208) doi:10.2196/90208

XSL•FO
RenderX

KEYWORDS

medical futures studies; Futures Wheel; foresight; futures studies; digital health; artificial intelligence; virtual consultation; remote care; telehealth; chronic conditions

Introduction

In the 21st century, the health care sector has experienced dramatic changes with the advent of digital health technologies. The spectrum of such tools spans from hardware, for example, wearable sensors and robotics, to software such as blockchain and artificial intelligence (AI) [1]. The permeation of these technologies into the health care ecosystem has been reshaping, or even redefining, contemporary medical journeys for stakeholders within this landscape. For instance, patients can obtain clinical-grade, personalized health insights from wearables that they can use at home, while health care professionals (HCPs) can access AI assistance to support their decision-making [2].

While digital health technologies can improve the health care experience for both patients and HCPs, there remains the risk of such potential not being fully realized. In a sector as established as health care, the implementation of new mechanisms faces significant challenges, largely due to human factors [3]. Poorly designed tools that do not account for varying levels of digital literacy or usability can face poor adoption among patients [1]. From HCPs' perspectives, they might not be motivated to use new technologies without clear advantages [1]. Overlooking human factors can limit technological breakthroughs in health care, underscoring the significance of the cultural element in digital health transformation [4]. In addition, there remain a number of potential challenges beyond the human element that can stifle effective digital health adoption. Factors including technical limitations, data integrity, and regulatory compliance represent possible barriers [3].

The successful implementation of digital health tools relies on adequate planning [3]. However, the novelty of such tools might not prepare stakeholders and health care systems to plan for their efficient adoption. The sparse use of such technologies in health care settings indicates a lack of evidence regarding their impact in practice. This can, in turn, negatively influence implementation and planning. There is a need to develop the ability to forecast possible futures to equip stakeholders with an anticipatory mindset or future thinking. Researchers have highlighted the value and need for implementing futures thinking in health care settings for meaningful planning [3,5].

There are several established methodologies for forecasting the future in the field of futures studies [5]. Such approaches can provide a deeper understanding of the impacts of upcoming trends and the prerequisites required to ensure their optimal and beneficial integration [6]. Furthermore, insights drawn from futures studies can inform decisions made by health care stakeholders in favor of desirable futures [6].

For the near future, virtual care and AI are considered health care megatrends [7], meaning trends that will have profound impacts on this particular sector. The valuable contributions of such technologically mediated health care approaches were highlighted during the COVID-19 crisis [2]. Kuziemsky et al [8] further identify a future with significant roles for AI in virtual care settings, from assisting in remote conversations and patient monitoring to guiding virtual assessments and evaluations. Despite the potential beneficial contributions of AI, there remains a risk in its deployment in health care settings. With the level of automation they bring about, they present the risk of automation bias, whereby users become overreliant on AI output, risking complacency and clinical robustness [9]. This can have severe downstream implications. A notable example comes from Obermeyer et al [10] study. The researchers identified racial bias in a common health AI model, where the algorithm less accurately detected the health risks of certain demographics. The recent rise in generative AI tools has surfaced another set of issues with AI technology. Of note is the effect of model "hallucination," where incorrect or fabricated information is generated as plausible output [11]. Reliance on such AI results has, in some cases, led to severe patient harm. For example, a patient was subjected to bromine toxicity following an inaccurate recommendation from ChatGPT [12]. These examples highlight the risk of unequal or even harmful outcomes if the technology is applied uncritically. Furthermore, these technical and user failures present legal and ethical ambiguity on the subject of accountability. Depending on the use case, AI system, and jurisdiction, liability can fall on institutions, developers, or individuals [13]. This indicates the need for important considerations in the design of AI systems with a view toward future implementation and dedicated health care usage.

However, the future impact of AI in virtual consultations has been poorly investigated. In fact, despite the value of adopting futures studies in health care and medicine, and the need to develop a medical futures studies subfield, there remains a scarcity of such undertakings [5]. This represents a gap in the literature and, therefore, in our understanding. Considering the positive outcomes associated with remote care [14] and our earlier work on co-designing an evidence-based AI tool for remote chronic obstructive pulmonary disease (COPD) assistance [15,16], it is propitious to undertake a medical futures study to explore the future impact of such a tool and address the identified gap in our understanding. Such an undertaking aids in anticipating benefits, challenges, and requirements for its implementation in real-world clinical practice. It also addresses the limited number of futures studies in health care contexts.

We aimed to tackle this endeavor in this study. It describes the results of an exploratory medical futures study using the Futures Wheel (FW) method, aimed at gathering the insights of patients and HCPs on the plausible impact of having a bespoke AI tool used in virtual consultations for COPD care. The research question is as follows: "What are the potential future impacts of a bespoke AI tool used in virtual consultations for COPD from the perspectives of patients and healthcare professionals?"

XSL•FO
RenderX

Among the few medical futures studies that used FW, the focus was on eHealth [6], systemic impacts of COVID-19 [17-19], emerging infectious diseases [20], and the evolution of the health care sector [21]. Therefore, to the best of our knowledge, this study is the first FW undertaking to be publicly reported with a focus on AI and virtual consultations. As such, we also provide some reflections and practical suggestions for future research adopting the FW in a similar context.

Methods

The Futures Wheel Method

The FW method was devised by futurist Jerome C Glenn in 1971 [22] as a structured brainstorming method to map out the direct and indirect impacts of a central trend or event. It is an established participatory data collection method in futures studies, where a trend or event of interest is written at the center of a paper, and participants write their perceived first- and second-order impacts emerging from that central event. Glenn [23] further revised the FW in subsequent years to include a set of predetermined areas along which the central event's impacts are to be considered. Version 2.0 of the FW can better organize discussions and various views on potential pathways that the central event may lead to.

Among futures methods, the FW is arguably the most accessible technique as it does not require any specialized tools. Considering the lack of familiarity with futures studies among the target participants, the accessibility of the FW was deemed adequate for this study. This methodological choice follows current recommendations for medical futures studies [24]. The FW is recommended for investigations aimed at exploring the consequences of a future event, as is the case in this study. In particular, the FW's output lends itself to our aim of exploring and surfacing the ripple effects of implementing a bespoke AI tool in virtual consultations. In contrast, other foresight methods used in medical futures studies, such as Delphi studies, scenario planning, and policy analysis, are less accessible to participants or less relevant for our aims, as they can require specialized tools, prior familiarity with foresight techniques, or are beyond exploratory in scope [24]. Despite its simplicity and accessibility, the FW remains a very powerful method to explore the future [23]. Therefore, for the scope of this study, the FW was an appropriate choice.

We conducted an in-person qualitative futures workshop using version 2.0 of the FW [23] with patients and HCPs who have experience in virtual consultations for COPD care. The central future statement was as follows: "The bespoke AI tool is used in every virtual consultation." In introducing the updated FW version, Glenn [23] is not prescriptive in the predetermined areas of impact to guide discussions, but recommends that they be determined by the purposes of the analysis while remaining manageably broad. As such, the research team convened to discuss the impact areas to be included and agreed upon the following based on the areas of importance derived from findings of stages of the larger project of which this study is a part [14-16] and for their relevance to the participants: HCP-patient relationship impact, psychological impact, social impact, educational impact, legal impact, ethical impact, health care delivery impact, and technology impact. We acknowledge cross-cutting impacts in the chosen categories and that this may result in overlapping input. Glenn [23] acknowledges the potential for overlaps and even divergent perspectives from participants on the impacts due to the nature of the FW process. Glenn further argues that the ability to surface such convergence and divergence is a strength of the method. A protocol for this study was registered on Open Science Framework [25]. Figure 1 shows the FW template used in this study.

We adopted the SRQR (Standards for Reporting Qualitative Research) reporting guideline to draft this manuscript [26], and used the corresponding SRQR reporting checklist during the editing process (Checklist 1).

XSL•FO
RenderX

Figure 1. Futures Wheel template used in this study. AI: artificial intelligence; HCP: health care professional.

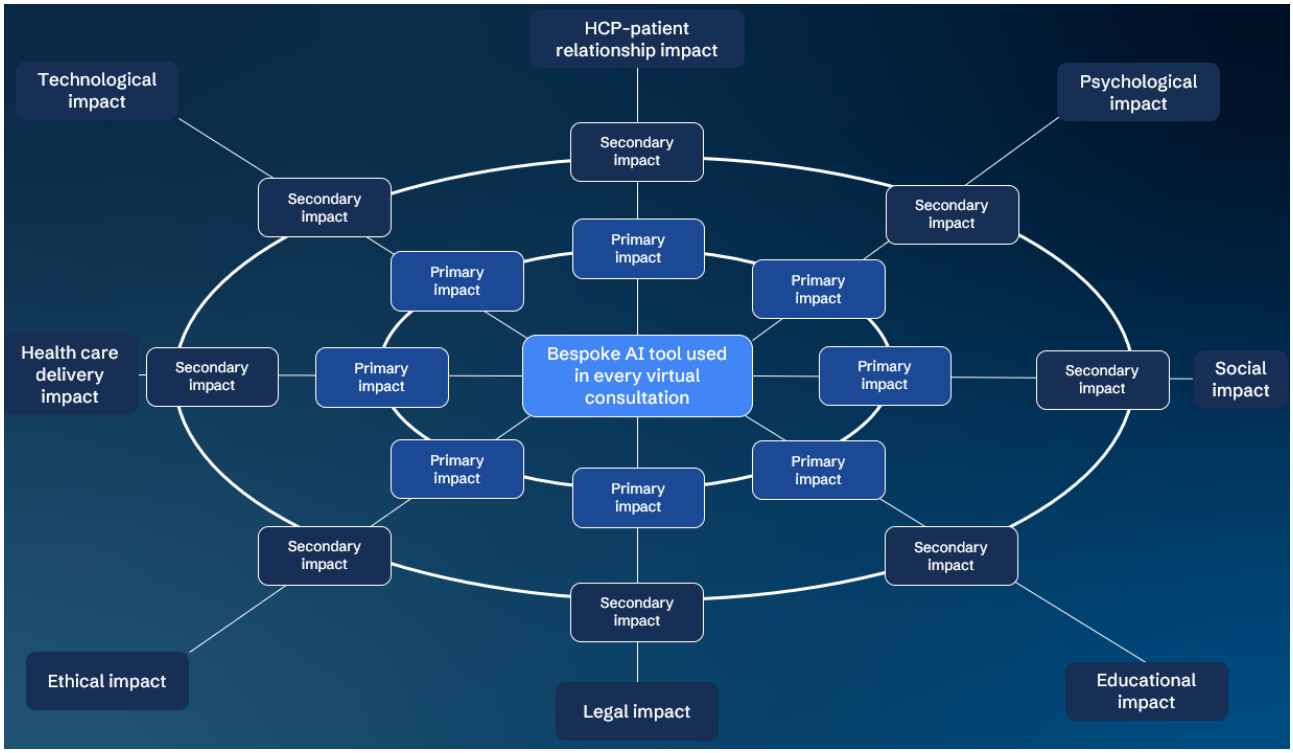

Ethical Considerations

Considering the involvement of participants in conducting an FW investigation, the corresponding ethical considerations have been taken into account. Ethics approval was obtained from the Atlantic Technological University Research Ethics Committee (IREC-ATUD-23‐038) and the Letterkenny University Hospital Research Ethics Committee (24/221124). Participation was voluntary, and participants provided their fully informed, written consent prior to data collection. No compensation was provided for participation. To safeguard participant information, data were anonymized.

Pilot and Data Collection Procedure

Prior to data collection, a pilot FW workshop was conducted with members of a public and patient involvement (PPI) committee to gather feedback on the process. PPI input is a recommended practice that supports and enhances study design [27]. The PPI committee members in this study were representative of the target participants, having experienced virtual consultations for chronic conditions. This allowed them to speak to the process for this cohort and provide valuable feedback.

During the actual FW workshop, the first author (PD) moderated and facilitated the activity with the participants. He did not contribute to the FW as a participant, but his role involved timekeeping, prompting discussion, and inviting critical appraisal. The workshop began with an explanation of the structured brainstorming process around the central event. Participants were provided with a printed version of the template (Figure 1), which they then filled out. Subsequently, each participant was invited to discuss their input in turn with the whole group, providing each participant with the opportunity to comment on and discuss their input on each segment of the FW, and even to challenge each other's input. This approach adds to the quality control of the findings [28] and provides more context around their input. As such, the results are presented with corresponding quotes that illustrate the findings. A topic guide was used to facilitate the discussion (Multimedia Appendix 1).

The discussion was recorded via the General Data Protection Regulation–compliant Microsoft Teams software (Microsoft Corporation), and the automated transcription feature was used. PD reviewed the automated transcripts for accuracy, with corrections made to the text to ensure a verbatim transcript for analysis. Transcripts were anonymized, and the discussion provided added context to the findings. PD analyzed the FW input and discussions in each established segment to derive relevant insights as per the FW method [23]. We acknowledge that PD's presence as moderator and facilitator potentially contributed to shaping group dynamics [29]. As such, BM provided rigor to the analysis by analyzing the data independently to confirm the relevance and accuracy of the data analysis. PD and BM convened to discuss and agree on the findings.

Participant Sampling and Approach

All participants in this study had been involved in the co-design of the AI prototype, CARA (Consultation Analysis & Response Assistant), in our earlier study [15,16]. It is a browser-based tool that has dedicated patient- and HCP-facing interfaces. During the iterative co-design process, the participants had hands-on sessions with the tool and shared their feedback for improvements. This feedback was then integrated into improved versions of the tool. The close involvement of the participants in the co-design of the software endowed them with a deep understanding of the AI's features. This familiarity with the

XSL•FO
RenderX

tool is essential for envisioning and articulating its future impacts during the FW workshop.

Considering the novelty of the AI tool, its prototype status, and the practicalities of engaging participants in hands-on co-design sessions, an exploratory FW was deemed adequate for the scope of this study. For such undertakings, a single small group of 4 to 6 participants is recommended [30]. For this study, 4 participants (2 patients and 2 HCPs) were involved. We intentionally recruited individuals with experiential knowledge of the bespoke AI tool, considering the research focus on surfacing informed, context-specific, plausible impacts of that specific software's use in virtual consultations. Among the few medical futures studies that used FW, those that reported their sample size included 9 [17], 10 [19], and 83 [21] participants, while the rest did not specify the number of participants involved [6,18,20]. This range of sample sizes (or lack of specificity thereof) is indicative of the nascency of medical futures undertakings, as well as the flexibility of the FW method. Notably, none of those former studies included a practical artifact with which participants required experiential knowledge. This further differentiates our exploratory study and its recruitment strategy, and we further reflect on this in "Recommendations for Future Research and Practice" and "Study Limitations" sections.

We emphasize that this study adopted an exploratory, qualitative foresight approach. Our pragmatic sample follows recommendations for "smart groups" size for exploratory FW that requires small groups of experts in a single workshop to explore the cascading effects from the central event [30]. The aim is thus to glean insights on the plausible future impacts of the bespoke AI tool, as well as on the FW exercise itself, which can guide subsequent work, as elaborated upon in "Recommendations for Future Research and Practice" section. Furthermore, specific, rare, and complex settings, as in the present case of bespoke AI-assisted remote care, often produce small samples [31]. The depth and contextual expertise of participants, combined with researcher reflexivity, can glean more important insights than larger projects with participants lacking experiential knowledge [32].

Reflexivity

The research team acknowledges the potential influence of their individual backgrounds in guiding the analysis. The team is composed of an interdisciplinary group, with backgrounds in medicine and health care (PD, BM, KM), medical futures studies (PD, BM), marketing (VO), and engineering (MM). In line with best practice guidelines [33], the research team adopts an explicit reflexive stance that acknowledges that their individual experiences, backgrounds, and perceptions might influence the interpretation of the data, instead of assuming neutral observation.

To maintain analytic integrity during the analysis, complementary procedures were implemented, informed by best methodological practices. First, the inclusion of an interdisciplinary team aided toward balancing different perspectives in data collection and analysis. This contributes to making the researcher's subjectivity more explicit [33]. There is also robustness in adopting reflexive guidelines [34,35], where the qualitative data were independently analyzed by the PD and BM who then discussed the results. This practice surfaces divergent readings and forces critical interrogation of analytic choices, all of which strengthen the reliability of the findings [36].

Overview of the Features of CARA

To provide context for the AI tool designed to support virtual consultations between patients with COPD and their HCPs, we present an overview of its co-designed features in the following tables: Table 1 summarizes the features of the patient interface, and Table 2 summarizes the features of the HCP interface.

Table . Summary of AI[a] features designed to assist patients on CARA's[b] dedicated patient interface.

Patient-facing feature	Description
Consultation analysis	Automated key takeaways and a simple summary of the consultation, along with automated clarifications of unfamiliar terms
Resource hub	Automatically suggests patient-oriented resources based on consultation context and provides the option to search for additional resources from a trusted health care institution's database

[a]AI: artificial intelligence.

[b]CARA: Consultation Analysis & Response Assistant.

XSL•FO
RenderX

Table . Summary of AI[a] features in CARA's[b] HCP[c] interface to support remote interaction and clinical assessment.

HCP-facing feature	Description
Past consultation analysis	Automated insights from past consultations, including a clinically relevant summary, patient emotional status, and suggestions for upcoming consultations, in view of familiarizing HCP with individual patients.
Current consultation analysis section	Automated insights based on the current consultation, including a clinically relevant summary, clarification of unfamiliar terms, real-time analysis of the patient's emotions, suggested next steps, and the option to generate a general practitioner referral letter.
Resource Hub	A querying tool that provides HCPs with resources drawn from a trusted health care institution's database.

[a]AI: artificial intelligence.

[b]CARA: Consultation Analysis & Response Assistant.

[c]HCP: health care professional.

Results

Overview

Four participants from the Northwest region of Ireland were involved in the in-person FW workshop in August 2025. This included 2 male patients with COPD (aged between 65 and 75 y) and 2 female HCPs (1 advanced nurse practitioner aged between 55 and 60 y and 1 nurse aged between 35 and 40 y). All participants had experience with virtual consultations for COPD care in the Irish public health sector, and all were involved in the co-design of CARA.

During the workshop, which lasted for 47 minutes, the participants individually filled out the FW and discussed their input on each impact. PD moderated the discussion based on the topic guide (Multimedia Appendix 1), allowing each participant, in turn, to share their input on 1 FW segment before moving to the next segment. The participants shared their individual input in each FW segment with the whole group, and each participant could contribute to the discussion. As this study was interested in the perspectives of patients and HCPs, the respective insights are summarized in the separate FW diagrams below. This presentation of the results is a deliberate choice due to the inevitable difference in perspectives between patients and HCPs, given that the bespoke AI tool consists of dedicated patient and HCP interfaces that cater to their specific needs (Tables 1 and 2).

Figure 2 summarizes the FW results from patients' perspectives.

Figure 3 summarizes the FW results from the perspective of HCPs.

The results of their input on each of the 8 impact areas are shared below, along with supportive quotes to illustrate their perspectives.

Figure 2. Futures Wheel results from the perspectives of patients with chronic obstructive pulmonary disease. AI: artificial intelligence; HCP: health care professional; HSE: health service executive.

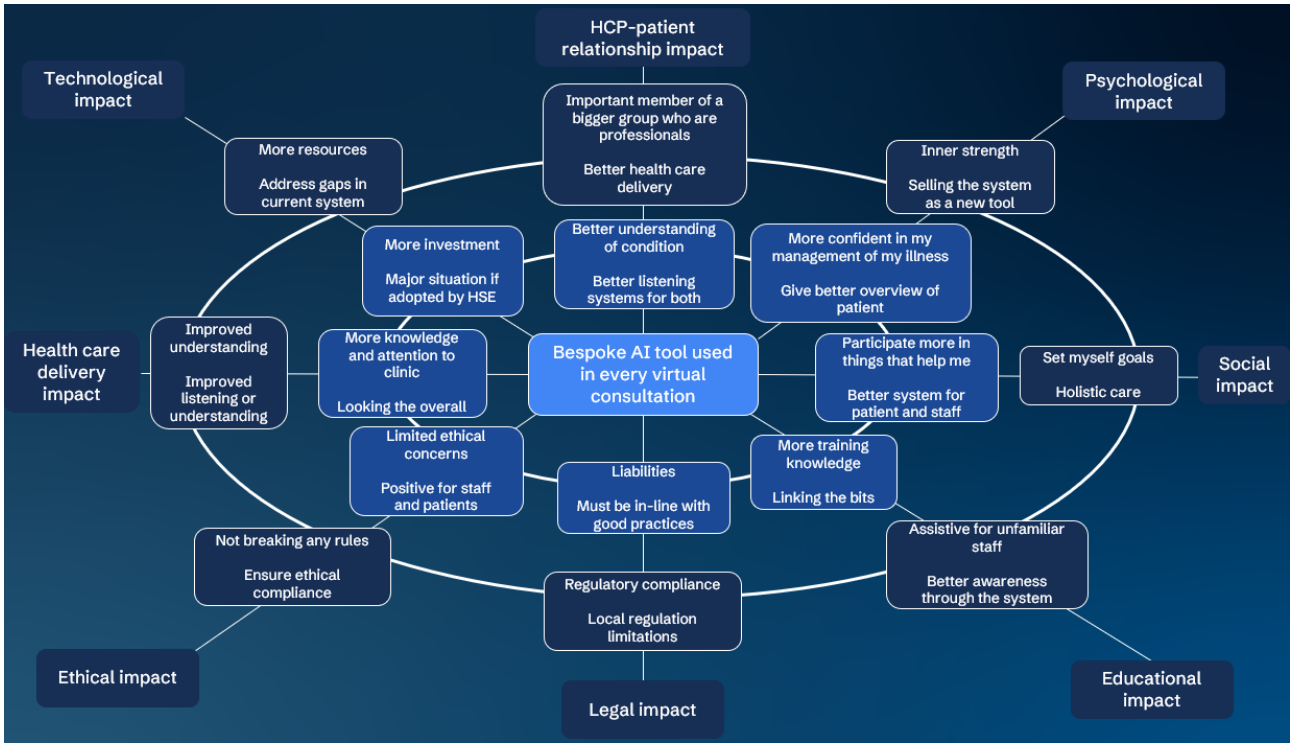

XSL•FO
RenderX

Figure 3. Futures Wheel results from the perspectives of HCPs attending patients with COPD remotely. AI: artificial intelligence; COPD: chronic obstructive pulmonary disease; HCP: health care professional.

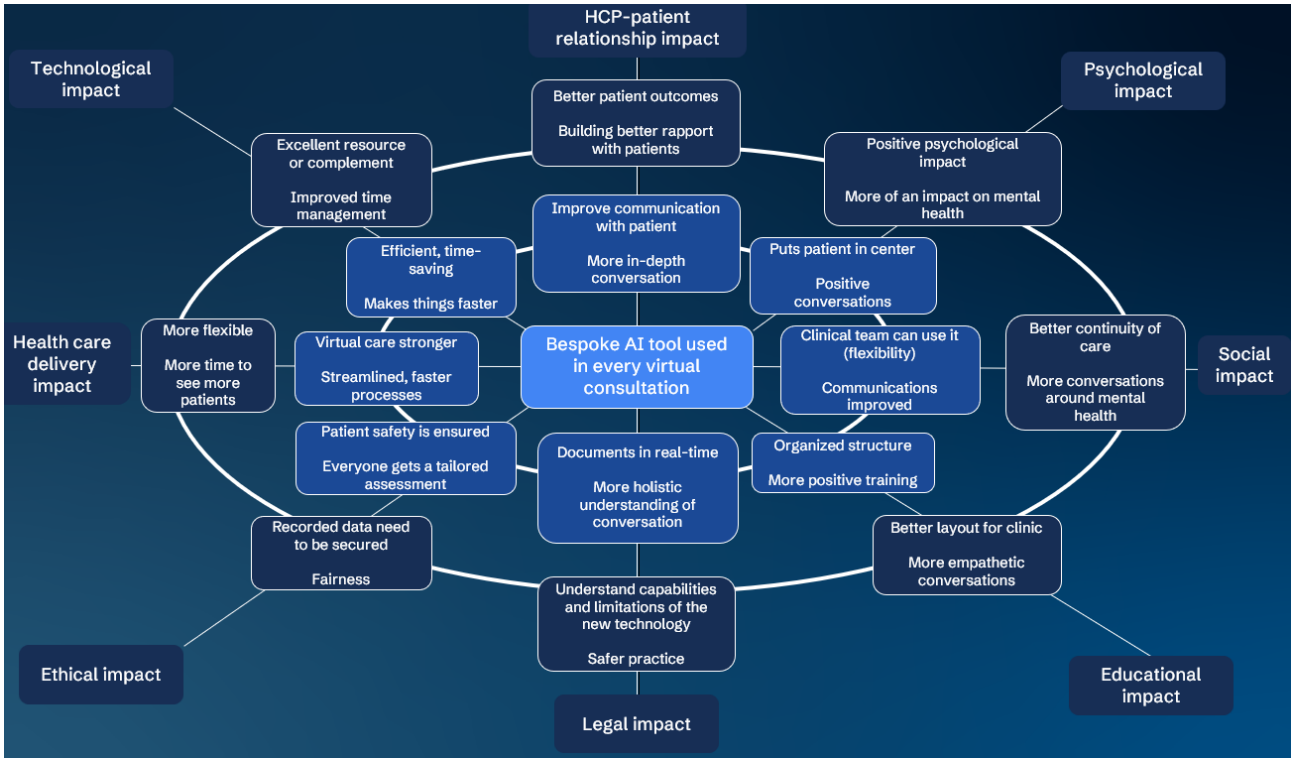

HCP-Patient Relationship Impact

As a primary impact of the HCP-patient relationship, patients expect the tool to redefine the remote care dynamic. Drawing from its features that support conversations with summaries and context-relevant cues, it complements remote interactions with a more holistic view of their status. From their point of view, this contributes to a positive therapeutic relationship with enhanced patient involvement:

> *[With CARA], there's a better understanding of our conditions.* [Patient #2]
>
> *It's not only the primary thing you're looking at. It's the other thing. It's like a holistic thing.* [Patient #1]
>
> *I would feel like an important member of a bigger group who are professionals.* [Patient #2]

HCPs identified that the software would provide meaningful assistance in guiding remote communications if used in practice. As a result, this would lead to building better rapport with patients and enhancing patient outcomes:

> *[There's] more in-depth conversations [with CARA].* [HCP #2]
>
> *[It will lead to] better patient outcomes.* [HCP #1]

Psychological Impact

Having a tool such as CARA in practice appears to empower patients and has a positive psychological impact. The AI support is expected to build their confidence in managing their condition, thanks to the accessible insights it provides:

> *It's given me a lot more confidence actually managing my own condition.* [Patient #2]
>
> *It's all explained like that, from a very simple structure that it's not complicated, but it is complicated, but it's simple at the end of the day.* [Patient #1]

Personalized insights about the patient's emotional well-being provided by the tool are expected to positively contribute to conversations initiated by HCPs. The resulting impact would aid in better understanding patients from a psychological point of view:

> *Because we're getting a better understanding that although I'm talking about respiratory conditions that they're telling me, you know, it's picking up on how they're feeling.* [HCP #2]
>
> *Utilising the AI analysis of the consultation, it's just all about creating a positive impact from a psychological perspective.* [HCP #1]

Social Impact

Patients view AI support in virtual consultations as beneficial for social patient-HCP interactions while encouraging self-management. Features such as automated patient resources and the ability for patients to find relevant health care resources through the software contribute to this sentiment:

> *[CARA's a] better system for patient and staff.* [Patient #1]
>
> *From all [of this support], you can set yourself goals.* [Patient #2]

For HCPs, having the tool implemented in practice would, as a primary social impact, supplement the clinical workflow with added flexibility and collaboration. This likely draws from CARA offering assistance, such as generating clinically relevant

XSL•FO
RenderX

summaries with actionable insights that aid in familiarizing with patients. The shared insights would thus lead to better care, with regard to the mental well-being component of patients:

> *So I think it moves from just from one clinician to a team. The team can use this [insight].* [HCP #1]
>
> *There are more conversations around mental health, making you more aware of conversations that you're having.* [HCP #2]

Educational Impact

With the automated, personalized consultation summaries, patients view the tool as addressing current limitations in familiarizing and educating HCPs about individual patient cases. As a result, this can lead to better clinical workflows and training among new staff:

> *It provides a bit more for people coming on to train, maybe new nurses or whatever.* [Patient #2]
>
> *Everyone has a better outline from day one when they start, both patient and the staff, and the clinicians.* [Patient #1]

On the educational front, HCPs identified a dimension of training and upskilling for students and staff with the automated AI output. These are also expected to support staff with administrative tasks. Features of CARA, such as automating referral letters and providing a holistic overview of patients' statuses, are relevant to these forecasts:

> *The way you recap the appointment, it's a better layout for the clinic, and with the summary, it would support education too.* [HCP #1]
>
> *For student nurses, [with CARA, there'll be] more positive training.* [HCP #2]

Legal Impact

While patients identify limited risks with CARA from a legal perspective, they highlight that the tool needs to comply with the corresponding regulations. As a result, there might be a need to update local regulations to account for the use of this type of AI assistance in practice:

> *I don't see anything that for me would be liable or something.* [Patient #2]
>
> *I know the HSE wants to be more in line with good practice across the EU.* [Patient #1]
>
> *I think it has to begin [...] to see where your work is legally and move forward.* [Patient #1]

With its real-time documentation and support for improved communication, HCPs believe that CARA's automated features will lead to safer practices. This would be contingent on its compliance with regulations:

> *It documents in real time, so that's quite good from a legal perspective.* [HCP #1]
>
> *You're picking up on all the things that you wouldn't have been picking up before, and that would lead to safer practice.* [HCP #2]
>
> *It's new, and we need to make sure the legal aspect was tight.* [HCP #1]

Ethical Impact

Patients did not identify significant ethical concerns with having such automated assistance. They viewed it as having a positive impact on remote care practices, as long as its use remains ethical:

> *I've listened to a whole lot of arguments, and I totally agree with you. We don't have a problem with it.* [Patient #1]
>
> *I don't think there's anything there that's breaking any rules.* [Patient #2]
>
> *I don't see why they would have a problem with it.* [Patient #2]

From HCPs' perspectives, CARA's personalized features would lead to fair, individual remote assessment. They also highlight that the integrity of patient data must not be overlooked when it is implemented:

> *It does a bit of background, and things like that. So it's picking up on everything.* [HCP #2]
>
> *So that the ethics in and around taking this data and managing that data [needs to be secured].* [HCP #1]

Health Care Delivery Impact

From the patients' perspective, they envision the software to enhance the remote care process with a focus on patients while supporting them in better understanding their condition. CARA's HCP-facing features, such as insights into their past consultations and emotional status, along with its patient-facing features that provide accessible resources, likely contribute to these expectations.

> *I think that's a positive to me, [with improved] listening and understanding, uhm, [during] a meeting between a patient and a clinician.* [Patient #1]
>
> *I think with this here, I can make myself more knowledgeable. I can ask questions.* [Patient #2]

CARA's features are seen as an upgrade to health care delivery from the HCPs' point of view. It would strengthen remote care with faster processes. As a result, remote care would become more flexible and allow HCPs to spend more time with patients:

> *So you would see more patients, and that it would be streamlined.* [HCP #2]
>
> *I thought the virtual care would be stronger than the current practice, and I thought it would be more flexible.* [HCP #1]

Technology Impact

If the AI tool is implemented in practice, patients believe that it will be a significant improvement to the current system. They acknowledge the need for adequate investment and resources but view it as worthwhile, considering the benefits to patients and the system as a whole:

> *Just imagine what better care they'd be able to give [with CARA].* [Patient #1]
>
> *It'll mean everything needs more investment, more resources.* [Patient #2]

XSL•FO
RenderX

> *It is not a huge investment for the HSE, but it's a huge investment for patients and their care.* [Patient #1]

Having access to CARA in practice is viewed as an efficient collaboration between humans and AI. With its automated features, it saves time for HCPs and complements the current system:

> *[CARA] makes things faster. For example, generating letters.* [HCP #2]

> *I thought it would be time-saving, so it would be an efficiency. And it would be an excellent complement [to our system].* [HCP #1]

Discussion

Principal Findings

This exploratory medical futures study aimed to gather insights on the future impact of integrating a bespoke, co-designed assistive AI tool in virtual consultations for COPD. The focus was on the perspectives of patients and HCPs regarding specific areas of impact. The findings of the FW exercise provided nuanced insights from the respective cohorts' visions of how integrating the bespoke tool into practice would reshape their virtual consultation experiences. Therefore, such FWs could provide policymakers with a clear panel of insights about developing and implementing a particular technology. As such, the method used in this study and the findings provide important contributions.

To the best of our knowledge, this is the first undertaking that uses the FW method to investigate the future impact of AI in virtual consultations. This section interprets these findings in relation to the extant literature and draws implications for future research and practice.

Impact of AI in Virtual Consultations for COPD Care

Participants in this study foresee a reimagined therapeutic relationship in remote care with the assistance of CARA. This sentiment is shared among both patients and HCPs ("HCP-Patient Relationship Impact" section) and likely draws from its purpose-built features aimed at improving communication and enhancing rapport building (Tables 1 and 2). This highlights the suboptimal quality of interactions in current virtual consultations [15], which negatively impact the overall quality of care and the therapeutic relationship in remote settings. HCPs even anticipate that the tool's insights into patients' emotional well-being will positively contribute to conversations, enabling a more holistic understanding of patients ("HCP-Patient Relationship Impact" section). The application of AI in clinical settings has raised concerns about the loss of the human touch [37]. However, it appears that if built to support interactions rather than replace them, as is the case with our co-designed prototype, AI assistance is perceived as positively reinforcing the patient-HCP relationship. In a future of remote care supported by AI, it could therefore be recommended that systems encourage the integration and investment in automated solutions that enhance, rather than diminish, the therapeutic relationship and the human touch.

The redefined remote care dynamic is further supported by the positive psychological support perceived, especially by patients ("Psychological Impact" section). Chronic conditions can have a significant impact on the mental well-being of patients living with such conditions [38], and they are likely to require psychological support mechanisms. It is notable that they not only find CARA's assistance to be a tool for information about their condition but also a source of emotional reassurance that builds their self-confidence and makes them more engaged in managing their condition ("Psychological Impact" and "Social Impact" sections). This indicates a potential lack of such support in current COPD care, and empowering such patients has been associated with improved outcomes [39]. AI's potential to empower patients has been highlighted in recent studies [40], and our findings indicate that the bespoke tool seems to have a similar effect. It could further be recommended for remote systems to include an element of psychological support and patient empowerment, whether through automated or traditional means, in view of contributing to better outcomes.

From HCPs' perspectives, the automated assistance is considered a mechanism to add efficiency to their workflow ("Social Impact," "Educational Impact," "Health Care Delivery Impact," and "Technology Impact" sections). They expect the bespoke tool to support flexible collaboration among colleagues while also supporting their upskilling and training. There is an underlying implication that the current remote care workflow might not be optimal and presents a learning curve, and automated tasks and patient insights can assist in this regard. Patients also believe that the AI's features support the onboarding of trainees and new staff, thereby addressing current limitations and leading to better levels of care delivery ("Educational Impact" and "Health Care Delivery Impact" sections). Such an observation likely draws from past experiences with staff unequipped to adapt their interactions in virtual settings [41]. As such, participants foresee CARA streamlining remote care delivery with added efficiency and flexibility while offering fair, patient-centered care. This suggests that a hybrid human-AI collaboration, where algorithms meaningfully complement HCP tasks and needs, can contribute to favorable perceptions of such tools and lead to better rates of adoption among the medical community. Complementarity has been identified as a central factor for adoption among medical professionals [42], and having such an element in AI systems can potentially lead to more effective integration. On the practical side, in planning for future remote care systems, it might also be appropriate to train the workforce in optimizing remote interactions and workflows.

Overall, the findings indicate a plausible future with generally positive impacts of AI in virtual consultations across 8 areas of focus. It should be noted that this observation applies to CARA, a bespoke, co-designed tool, rather than other AI software. This bespoke tool was designed in an evidence-based manner with the close involvement of end users to reflect their needs and preferences [15]. While the findings are limited to our cohort of participants, participatory approaches in developing digital health tools such as CARA can potentially yield better perceptions and acceptance of novel tools. This aligns with

XSL•FO
RenderX

reflections and recommendations from existing literature [43-45].

Nevertheless, participants cautioned about the need to ensure data integrity and regulatory compliance when implementing such novel tools in practice, despite their positive contributions to health care delivery ("Legal Impact" and "Ethical Impact" sections). From the patients' standpoint, as CARA increases access to patient resources and self-directed care, they identify limited ethico-legal risks, but their reflections implicitly indicate the need to ensure equity and data security. HCPs emphasized the need for safeguards to ensure regulatory and data compliance. These are prevalent concerns among patients concerning AI in health care [46,47], which can potentially be addressed with human-centered design of such tools along regulatory frameworks from the planning stages. Notably, the limited ethical and legal risks that the participants identified suggest a potential limitation in the thoroughness of their understanding of the breadth of AI's impact. As discussed earlier in the "Introduction" section, AI implementation in health care contexts presents a range of risks from automation bias [9] and real patient harm [10,12] to ethical and legal challenges [13]. While it is possible that participant's concerns over these areas of impact might have been assuaged by their direct involvement in CARA's co-design and their understanding of its assistance, embedded human oversight, and limitations, there is also the possibility of "unrealistic optimism" [48]. This refers to the tendency of people to not expect themselves to be victims of adverse effects; in this case, due to not being fully cognizant of AI's complexities, likely due to CARA's novelty. This observation derived from the FW insights thus indicates the need for careful implementation strategies around AI tools, possibly including an element of enhanced AI literacy, even if their assistance is evidence-based. This further indicates the need to include other stakeholders, such as policymakers and hospital managers, in designing such products to ensure their compliance in real-world practice.

Recommendations for Future Research and Practice

The findings of this study have provided crucial insights for future research and practice. As the current scope was at an exploratory level, future research could expand participation with additional stakeholders, such as policymakers, developers, and hospital managers in larger groups to surface a broader range of perspectives and considerations that would enrich the diversity and complexity of the identified impacts. They could integrate longer or additional workshops to further explore more complex and interconnected impacts. In line with medical futures guidelines [24], such an approach could expand the gathered foresight insights to develop actionable implementation strategies.

For instance, Figures 1 and 2 can offer a preliminary roadmap for discussions with the Irish Health Service Executive, as the participants involved in this study draw their virtual consultation experiences from the health service executive. The findings are highly relevant to this context, and discussions could center around the patient and HCP perceptions of such AI assistance in future practice.

The findings of our FW can already be used to anticipate issues and challenges, such as equity and data protection concerns, which surfaced during our analysis. As the AI prototype is at a functional level that can be piloted in low-risk virtual consultations, subsequent studies involving CARA in real-life pilots could address such challenges ahead of involving participants. Considering the generally positive perspectives on the bespoke AI's impact on remote care practices, there is value in considering the tool as a candidate for such studies to further assess its contribution in practical scenarios.

On a practical level, for conducting FW studies with digital health technologies, based on the authors' experience, it is recommended that participants familiarize themselves with the tools, ideally through hands-on sessions. As our FW study followed co-design sessions where participants used the tool in question, this facilitated discussions around areas of impact of interest during their forecasting exercise. This might limit the number of participants due to logistical considerations, but we argue that the findings would be stronger as they are more contextual.

We also highly recommend piloting FW or other medical futures methods prior to data collection, preferably with a PPI committee. In our experience, their input helped improve the discussion elements as well as the format in which the study was conducted. For instance, since participants needed to familiarize themselves with the features of the AI prototype, which required guidance from PD, hosting the workshop in person was deemed optimal following the PPI pilot.

Study Limitations

While our study has followed recommended practices and guidelines [5,24] and uncovered valuable insights, especially considering the scarcity of similar investigations, we acknowledge its limitations. Notably, the number of participants indicates the need for caution when interpreting the findings. While such small groups are adequate for exploratory FW studies [30], they can limit the depth of insights drawn. However, on a practical level, small groups are more manageable, especially if they are to have hands-on sessions with new tools. Future research that expands beyond an exploratory level could address this limitation by organizing multiple workshops with small groups and collating their insights. Therefore, the current output of this FW undertaking is more exploratory than confirmatory, in line with its scope.

The participant demographics involved in this study also limit its generalizability. While limited to the perspectives of the public health care system in the Northwest of Ireland and COPD as a condition, our study offers valuable academic contributions regarding the potential impact of such a bespoke AI tool in future practice, as it has provided practical insights that can inform future research and practice.

Another limitation of the sampling strategy is the potential for selection bias. Since all participants were involved in the AI tool's co-design, they might have tended to view its impact more positively and might have been less likely to identify its risks. Future studies that go beyond an exploratory scope could conduct an FW activity with participants who were not involved

XSL•FO
RenderX

in the co-design. However, there is value in including participants with experiential knowledge at this stage, as it can surface crucial, context-relevant insights that would not be possible with participants lacking familiarity with the tool [49].

Due to the inherent uncertainties in contemplating the future and the subjective influences of participants' perceptions [24], the FW forecasts do not guarantee the likelihood of any particular outcome. Nevertheless, the findings provide constructive foresight that can assist in decision-making toward desirable outcomes, especially toward early planning, governance, and implementation considerations.

Conclusions

There is a plausible, long-term future where AI becomes an integral component of remote chronic care. Contemplating such a future with patients and health care professionals indicates the practical benefits of embedding such automated assistance. They perceive positive contributions not only from technological and operational levels but also on the level of the therapeutic relationship and patient empowerment. Their articulated forecasts indicate a future that is hopeful yet cautious. In that desirable future, with a bespoke co-designed tool, there is complementarity between humans and AI that enhances the interaction quality, and successful implementation is supported by strong governance. This study has uncovered crucial insights that can guide stakeholders and medical futures practitioners, as well as academic and practical endeavors in navigating the uncertain future of AI-driven care.

Acknowledgments

The authors acknowledge the contributions of the following public and patient involvement members for their input in the pilot Futures Wheel workshop and for providing their valuable feedback: Antoinette Doherty, Helen Cooper, and Geraldine Halpin. The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: idea generation and brainstorming, reformatting, identification of limitations, and recommendations. The GAI tool used was GPT-5.2. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

The authors declared no financial support was received for this work.

Data Availability

Anonymized data are available upon reasonable request.

Authors' Contributions

Conceptualization: PD
Data curation: PD
Formal analysis: PD, BM
Investigation: PD
Methodology: PD, BM
Project administration: PD
Supervision: BM, KM, VO, MM
Validation: PD, BM
Visualization: PD
Writing—original draft: PD
Writing—review and editing: PD, BM, KM, VO, MM

Conflicts of Interest

Coauthor BM is one of the Guest Editors for JMIR's Theme Issue: "The Emergence of Medical Futures Studies." The other authors declare no conflicts of interest.

Multimedia Appendix 1
Futures Wheel topic guide.
[DOCX File, 15 KB - jmir_v28i1e90208_app1.docx]

Checklist 1
SRQR checklist.
[PDF File, 303 KB - jmir_v28i1e90208_app2.pdf]

XSL•FO
RenderX

References

1. Yeung AWK, Torkamani A, Butte AJ, et al. The promise of digital healthcare technologies. Front Public Health 2023;11:1196596. [doi: 10.3389/fpubh.2023.1196596] [Medline: 37822534]
2. Han JH, Lee JY. Digital healthcare industry and technology trends. Presented at: 2021 IEEE International Conference on Big Data and Smart Computing (BigComp); Jan 17-20, 2021. [doi: 10.1109/BigComp51126.2021.00083]
3. Thimbleby H. Technology and the future of healthcare. J Public Health Res 2013 Dec 1;2(3):e28. [doi: 10.4081/jphr.2013.e28] [Medline: 25170499]
4. Meskó B, Drobni Z, Bényei É, Gergely B, Győrffy Z. Digital health is a cultural transformation of traditional healthcare. mHealth 2017;3:38. [doi: 10.21037/mhealth.2017.08.07] [Medline: 29184890]
5. Meskó B, Kristóf T, Dhunnoo P, Árvai N, Katonai G. Exploring the need for medical futures studies: insights from a scoping review of health care foresight. J Med Internet Res 2024 Oct 9;26:e57148. [doi: 10.2196/57148] [Medline: 39383528]
6. Kouri P, Hopia H, Hakala A. Predicting the future of healthcare and eHealth with the Futures Wheel method. J Int Soc Telemed eHealth 2020 Dec 26;8:e22-e21. [doi: 10.29086/JISfTeH.8.e22]
7. Deml MJ, Jungo KT, Maessen M, Martani A, Ulyte A. Megatrends in healthcare: review for the Swiss National Science Foundation's National Research Programme 74 (NRP74) "smarter health care". Public Health Rev 2022;43:1604434. [doi: 10.3389/phrs.2022.1604434] [Medline: 35528712]
8. Kuziemsky C, Maeder AJ, John O, et al. Role of artificial intelligence within the telehealth domain. Yearb Med Inform 2019 Aug;28(1):35-40. [doi: 10.1055/s-0039-1677897] [Medline: 31022750]
9. Skitka LJ, Mosier KL, Burdick M. Does automation bias decision-making? Int J Hum Comput Stud 1999 Nov;51(5):991-1006. [doi: 10.1006/ijhc.1999.0252]
10. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
11. Kim Y, Jeong H, Chen S. Medical hallucinations in foundation models and their impact on healthcare. arXiv. Preprint posted online on Feb 26, 2025. [doi: 10.48550/arXiv.2503.05777]
12. Eichenberger A, Thielke S, Van Buskirk A. A case of bromism influenced by use of artificial intelligence. Ann Intern Med Clin Cases 2025 Aug 5;4(8):e241260. [doi: 10.7326/aimcc.2024.1260]
13. Bottomley D, Thaldar D. Liability for harm caused by AI in healthcare: an overview of the core legal concepts. Front Pharmacol 2023;14:1297353. [doi: 10.3389/fphar.2023.1297353] [Medline: 38161692]
14. Dhunnoo P, Kemp B, McGuigan K, Meskó B, O'Rourke V, McCann M. Evaluation of telemedicine consultations using health outcomes and user attitudes and experiences: scoping review. J Med Internet Res 2024 Jul 9;26:e53266. [doi: 10.2196/53266] [Medline: 38980704]
15. Dhunnoo P, McGuigan K, O'Rourke V, Meskó B, McCann M. User perceptions of virtual consultations and artificial intelligence assistance: a mixed methods study. Future Internet 2026 Feb;18(2):84. [doi: 10.3390/fi18020084]
16. Dhunnoo P, Meskó B, O'Rourke V, McGuigan K, McCann M. Towards enhanced virtual chronic care: artificial intelligence-human computer interaction integration in synchronous virtual consultations. In: Arai K, editor. Intelligent Systems and Applications: Springer; 2024:16-23. [doi: 10.1007/978-3-031-66329-1_2]
17. Nouhi M, Heydari M, Goudarzi Z, Shahtaheri RS, Ahmadzaeh A, Olyaeemanesh A. The future effects of COVID-19 on the health system: applying the Futures Wheel method. Med J Islam Repub Iran 2022;36:164. [doi: 10.47176/mjiri.36.164] [Medline: 36908937]
18. Daffara P. Applying the Futures Wheel and macrohistory to the COVID-19 global pandemic. J Futures Stud 2020 Dec;25(2):35-48. [doi: 10.6531/JFS.202012_25(2).0006]
19. Rezaei H, Haghdoost A, Javar HA, et al. The effect of coronavirus (COVID-19) pandemic on medical sciences education in Iran. J Educ Health Promot 2021;10:136. [doi: 10.4103/jehp.jehp_817_20] [Medline: 34222511]
20. Nejatzadehgan Eidgahi Z, Dehnavieh R, Borhaninejad V, et al. The Futures Wheel model of the effects of the emerging infectious diseases pandemics on the elderly in Iran. Salmand 2024;18(4):606-619. [doi: 10.32598/sija.2023.3641.1]
21. Hopia H, Hakala A. Finnish social and health care professionals' perspective of the future. Int J Healthcare 2015;2(1). [doi: 10.5430/ijh.v2n1p12]
22. Glenn J. Futurizing teaching vs. futures course. Soc Sci Rec 1972;9(3):26-29 [FREE Full text]
23. Glenn JC. The Futures Wheel. In: Glenn JC, Gordon TJ, editors. Futures Research Methodology: The Millennium Project; 2009:20-28.
24. Mesko B, Kristóf T, Dhunnoo P, Árvai N, Katonai G. A practical guide to using futures methods in health care: approaches, applications, and case studies. J Med Internet Res 2025 Sep 16;27:e82820. [doi: 10.2196/82820] [Medline: 40956602]
25. Dhunnoo P, Meskó B, O'Rourke V, et al. Protocol for the co-design and medical futures studies of an assistive artificial intelligence tool for remote chronic care. Open Science Framework. 2025. URL: https://osf.io/vu9jz/overview [accessed 2026-05-26]
26. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
27. Coupe N, Mathieson A. Patient and public involvement in doctoral research: impact, resources and recommendations. Health Expect 2020 Feb;23(1):125-136. [doi: 10.1111/hex.12976] [Medline: 31613049]

XSL•FO
RenderX

28. Nielsen AF, Michelmann J, Akac A, et al. Using the Futures Wheel methodology to assess the impact of Open Science in the transport sector. Sci Rep 2023 Apr 12;13(1):6000. [doi: 10.1038/s41598-023-33102-5] [Medline: 37046066]
29. Ozkaynak M, Sircar CM, Frye O, Valdez RS. A systematic review of design workshops for health information technologies. Informatics 2021;8(2):34. [doi: 10.3390/informatics8020034]
30. Bengston DN. The Futures Wheel: a method for exploring the implications of social–ecological change. Soc Nat Resour 2016 Mar 3;29(3):374-379. [doi: 10.1080/08941920.2015.1054980]
31. McDermott R. On the scientific study of small samples: challenges confronting quantitative and qualitative methodologies. Leadersh Q 2023 Jun;34(3):101675. [doi: 10.1016/j.leaqua.2023.101675]
32. Crouch M, McKenzie H. The logic of small samples in interview-based qualitative research. Soc Sci Inf 2006 Dec;45(4):483-499. [doi: 10.1177/0539018406069584]
33. Finlay L. “Outing” the researcher: the provenance, process, and practice of reflexivity. Qual Health Res 2002 Apr;12(4):531-545. [doi: 10.1177/104973202129120052] [Medline: 11939252]
34. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
35. Braun V, Clarke V, Hayfield N, Davey L, Jenkinson E. Doing reflexive thematic analysis. In: McBeath A, Bager-Charleson S, editors. Supporting Research in Counselling and Psychotherapy: Qualitative, Quantitative, and Mixed Methods Research: Springer Nature; 2023:19-38. [doi: 10.1007/978-3-031-13942-0_2]
36. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Methods 2017;16(1). [doi: 10.1177/1609406917733847]
37. Witkowski K, Dougherty RB, Neely SR. Public perceptions of artificial intelligence in healthcare: ethical concerns and opportunities for patient-centered care. BMC Med Ethics 2024 Jun 22;25(1):74. [doi: 10.1186/s12910-024-01066-4] [Medline: 38909180]
38. Huang Y, Loux T, Huang X, Feng X. The relationship between chronic diseases and mental health: a cross-sectional study. Ment Health Prev 2023 Dec;32:200307. [doi: 10.1016/j.mhp.2023.200307]
39. Fromer L. Implementing chronic care for COPD: planned visits, care coordination, and patient empowerment for improved outcomes. Int J Chron Obstruct Pulmon Dis 2011;6:605-614. [doi: 10.2147/COPD.S24692] [Medline: 22162647]
40. Hwang M, Zheng Y, Cho Y, Jiang Y. AI applications for chronic condition self-management: scoping review. J Med Internet Res 2025 Apr 8;27:e59632. [doi: 10.2196/59632] [Medline: 40198108]
41. Budd G, Griffiths D, Howick J, et al. Empathy in patient-clinician interactions when using telecommunication: a rapid review of the evidence. PEC Innov 2022 Dec;1:100065. [doi: 10.1016/j.pecinn.2022.100065] [Medline: 35996734]
42. Hemmer P, Schemmer M, Riefle L, Rosellen N, Vössing M, Kühl N. Factors that influence the adoption of human-AI collaboration in clinical decision-making. Presented at: European Conference on Information Systems (ECIS 2022); Jun 18-24, 2022 URL: https://aisel.aisnet.org/ecis2022_rp/139/? [accessed 2026-05-26]
43. Banerjee S, Alsop P, Jones L, Cardinal RN. Patient and public involvement to build trust in artificial intelligence: a framework, tools, and case studies. Patterns (N Y) 2022 Jun 10;3(6):100506. [doi: 10.1016/j.patter.2022.100506] [Medline: 35755870]
44. Woodward M, Dixon-Woods M, Randall W, et al. How to co-design a prototype of a clinical practice tool: a framework with practical guidance and a case study. BMJ Qual Saf 2024 Mar 25;33(4):258-270. [doi: 10.1136/bmjqs-2023-016196] [Medline: 38124136]
45. Noorbergen TJ, Adam MTP, Roxburgh M, et al. Co-design in mHealth systems development: insights from a systematic literature review. AIS Trans Hum-Comput Interact 2021;13(2):175-205. [doi: 10.17705/1thci.00147]
46. Osnat B. Patient perspectives on artificial intelligence in healthcare: a global scoping review of benefits, ethical concerns, and implementation strategies. Int J Med Inform 2025 Nov;203:106007. [doi: 10.1016/j.ijmedinf.2025.106007] [Medline: 40494217]
47. Prakash S, Balaji JN, Joshi A, Surapaneni KM. Ethical conundrums in the application of artificial intelligence (AI) in healthcare—a scoping review of reviews. J Pers Med 2022 Nov 16;12(11):1914. [doi: 10.3390/jpm12111914] [Medline: 36422090]
48. Weinstein ND. Unrealistic optimism about future life events. J Pers Soc Psychol 1980;39(5):806-820. [doi: 10.1037/0022-3514.39.5.806]
49. Dumez V, L’Espérance A. Beyond experiential knowledge: a classification of patient knowledge. Soc Theory Health 2024 Sep;22(3):173-186. [doi: 10.1057/s41285-024-00208-3]

Abbreviations

AI: artificial intelligence
CARA: Consultation Analysis & Response Assistant
FW: Futures Wheel
HCP: health care professional
PPI: public and patient involvement

XSL•FO
RenderX

SRQR: Standards for Reporting Qualitative Research

Edited by A Mavragani; submitted 23.Dec.2025; peer-reviewed by A Marton, I Dén-Nagy; revised version received 10.Mar.2026; accepted 03.Apr.2026; published 03.Jun.2026.

Please cite as:
Dhunnoo P, Meskó B, McGuigan K, O'Rourke V, McCann M
Forecasting the Impacts of Artificial Intelligence Assistance in Virtual Consultations for Chronic Obstructive Pulmonary Disease: Exploratory Futures Wheel Study
J Med Internet Res 2026;28:e90208
URL: https://www.jmir.org/2026/1/e90208
doi:10.2196/90208

© Pranavsingh Dhunnoo, Bertalan Meskó, Karen McGuigan, Vicky O'Rourke, Michael McCann. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 3.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Sequencing AI Automation and Data Interoperability in Oncology Using a Scenario-Planning Framework Coupled With Discrete-Event Simulation: Proof-of-Concept Study

Peter May[1], Dr med, MD, MPH, MA; Sabine D Brookman-May[2], Prof Dr Med, MD; Edward Garrahy[3], MD; Johannes von Büren[4], Dr med, MD

[1]Department of Medicine III, School of Medicine and Health, Technical University of Munich, Ismaninger Str. 22, Munich, Bavaria, Germany
[2]Department of Urology, Ludwig-Maximilians-Universität München, Munich, Bavaria, Germany
[3]Wellster Healthtech Group, Munich, Germany
[4]Sidekick Health Germany GmbH, Hamburg, Germany

Corresponding Author:
Peter May, Dr med, MD, MPH, MA
Department of Medicine III, School of Medicine and Health, Technical University of Munich, Ismaninger Str. 22, Munich, Bavaria, Germany

Abstract

Background: As oncology workflows integrate increasingly autonomous artificial intelligence (AI) agents, health systems face uncertainty regarding operational impacts. Traditional linear forecasting methods fail to capture second-order effects such as governance saturation, induced demand, and bottleneck migration. To navigate this complexity, the emerging field of medical futures studies requires methodologies that bridge qualitative strategic foresight with quantitative operational modeling. These system-level dynamics directly influence timely diagnosis, treatment delays, and overall health system resilience.

Objective: This study aimed to develop a proof-of-concept framework coupling qualitative scenario planning with computational discrete-event simulation to stress-test oncology AI adoption strategies.

Methods: We defined a strategic state space using 2 orthogonal axes, AI automation intensity and data interoperability, resulting in 4 distinct futures scenarios. We translated these qualitative narratives into a quantitative discrete-event simulation model of a 3-year operational horizon. The model quantified system performance (referral-to-treatment interval [RTTI] and throughput), volatility, and resource constraints across different adoption trajectories.

Results: The scenario-planning phase yielded 4 operational archetypes (analog oncology, automation islands, interconnected clinicians, and AI-orchestrated care) with distinct constraints, risks, and failure modes. In the simulation, the fully integrated scenario maximized capacity (1244, SD 21.4 patients per year) and halved the mean RTTI to 14.9 (SD 0.3) days, a magnitude comparable to major pathway redesign interventions. Isolated automation without data infrastructure led to reduced system performance, increasing RTTI by 26% (37.1, SD 1.3 days) and reducing throughput to 647 (SD 10.1) patients per year due to administrative governance saturation. The model illustrated a structural bottleneck migration: successful upstream AI adoption shifted binding constraints from diagnostic scanners to downstream chemotherapy infusion units, whereas missing data interoperability resulted in governance constraints. Pathway optimization analysis indicated that a coordinated strategy prioritizing early improvements in data interoperability reduced transition volatility compared to an automation-first approach.

Conclusions: Integrating qualitative scenario planning with quantitative simulations enabled a systematic evaluation of oncology AI adoption strategies. As a proof of concept, it offers a replicable framework for health leaders to model future scenarios of digital transformation in times of high uncertainty. Subsequent work should expand this methodology to incorporate financial and health equity dimensions, establishing simulation-based scenario planning as an important tool in medical futures studies.

(J Med Internet Res 2026;28:e92642) doi:10.2196/92642

KEYWORDS

medical futures studies; artificial intelligence; AI; data interoperability; scenario planning; discrete-event simulation

Introduction

Modern oncology faces an increasing gap between therapeutic potential and operational capacity. While precision medicine expands, care delivery remains constrained by diagnostic throughput, pathology backlogs, and the cognitive load of multidisciplinary coordination. At the same time, clinical artificial intelligence (AI) is evolving from narrow, task-specific

XSL•FO
RenderX

applications toward workflow-native systems [1]. This next generation includes large language models capable of complex documentation and eligibility screening alongside emerging agentic AI systems designed to execute multistep planning with minimal human intervention [2-6]. Recent reviews describe the rapid proliferation of these tools in oncology yet consistently highlight that their translation into practice is constrained not by algorithmic performance but by integration challenges, data quality, and safety governance [7-13].

The impact of AI in complex adaptive systems such as oncology is rarely linear; local efficiency gains often destabilize rather than improve system-level performance. There is evidence suggesting that altering information flows, which is heavily dependent on robust data interoperability, can shift use patterns and clinical outcomes [14-16]. However, these integration efforts often trigger higher-order effects such as bottleneck migration and induced demand, which are usually not captured by model-centric evaluations. Furthermore, as sociotechnical complexity grows, regulatory frameworks are becoming more relevant [17,18]. New standards such as DECIDE-AI (Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence), the European Union AI Act, and the Food and Drug Administration Good Machine Learning Practice now specify context-aware assessment and life cycle oversight [19]. Consequently, governance moves from a background administrative activity to a potential capacity-limiting resource, influencing whether automation yields genuine access gains or merely introduces new operational friction.

Addressing these nonlinear dynamics requires methodologies capable of navigating uncertainty, a domain where health care has historically underused strategic foresight [20-22]. Scenario planning, a mature management science technique, offers a structured approach to reasoning through such strategic uncertainty [23]. While futures studies in adjacent fields such as biobanking and education demonstrate how governance and incentives determine whether technical visions become reality, these methods are rarely applied to operational capacity planning [24,25].

To bridge this gap using medical futures thinking, we coupled qualitative scenario planning with computational discrete-event simulation (DES) in this study [26]. By translating plausible adoption trajectories into quantitative models, we analyzed how variables such as AI automation intensity and data interoperability interact to produce higher-order effects, including governance saturation, false positive rates, and volatility. This paper presents a proof-of-concept framework for governments and health care executives to plan oncology AI adoption, shifting the focus from static operational forecasts to the dynamic management of structural transformation under uncertainty. In the following sections, we first describe our scenario design and simulation model; then present comparative results across scenarios; and, finally, discuss strategic implications.

Methods

Study Design and Framework

We integrated qualitative scenario planning with DES to stress-test oncology operations against alternative adoption futures. Scenario planning defines plausible strategic futures, whereas DES evaluates how those futures propagate through operational constraints over time. To structure the uncertainty space, we first identified the 2 most decisive yet independent drivers of digital transformation: AI automation intensity (A), representing the degree of algorithmic task substitution in clinical and administrative workflows, and data interoperability (I), representing the friction of data transfer, provenance, and reuse across system boundaries. By treating these variables as orthogonal rather than correlated, we defined 4 operational quadrants in the 2 × 2 scenario analysis and qualitatively explored their outcomes and risk profiles.

DES Model

To quantify operational implications, we constructed a DES model tracking individual patient trajectories through a resource-constrained oncology system over time. Implemented in SimPy (Python version 3.14; Python Software Foundation), the model represented a generic high-volume cancer center as a network of finite queues (eg, scanners, pathologists, oncologists, and treatment chairs). We selected resource capacities, service times, and failure rates based on standard oncology operation principles to create a representative, capacity-constrained baseline (the full script can be found in Multimedia Appendix 1). This stylized parameterization was chosen to isolate the mechanistic effects of variables A and I on system flow rather than fit the historical data of a single institution. Consequently, the results illustrate directional trends and structural vulnerabilities rather than precise, generalizable operational forecasts.

The simulation ran for 3 years (1095 days) to ensure statistical convergence of queue dynamics and capture the full variance of patient pathways; a 150-day warm-up period was used to reduce initialization bias. Patient agents were generated stochastically based on the 3 most common tumor types (lung, breast, and colorectal) and disease stages (early vs late), flowing through a pathway that encompasses administrative processing, diagnostic imaging, biopsy, pathology (with probabilistic rework loops), molecular testing, multidisciplinary team coordination, and multimodal treatment initiation (surgery, chemotherapy, and radiotherapy). A calendar engine distinguished between standard clinical shifts and continuous laboratory operations, enforcing atomic task logic to prevent procedural fragmentation across shifts.

The model was governed by the 2 exogenous variables (A and I) normalized from 0.0 to 1.0. Workflow durations (T) for clinical tasks were calculated dynamically as follows: $T_{task} = T_{base} \times (1 - \alpha \cdot A) \times (1 - \beta \cdot I)$, where α and β represent the susceptibility of the task to automation and data integration, respectively.

XSL•FO
RenderX

Operational Mechanisms

The simulation incorporated 8 coupled mechanisms to capture higher-order behaviors beyond linear efficiency gains (Table S1 in Multimedia Appendix 2). These mechanisms captured 3 core operational dynamics: administrative friction, where high AI automation intensity without sufficient data interoperability triggers manual verification and shifts bottlenecks to administrative staff; induced demand, where highly sensitive AI triage increases false positive rates and generates additional downstream biopsies; and pathway shifts, where faster upstream diagnosis accelerates patient arrivals at downstream treatment centers and alters multidisciplinary coordination requirements. Each mechanism was operationalized through parameterized interactions governing queue progression, resource availability, and downstream demand propagation.

Experimental Design and Analysis

We executed a comprehensive parameter sweep across the $A \times I$ state space with multiple stochastic replications per coordinate to generate uncertainty surfaces. Primary outcome measures included the mean referral-to-treatment interval (RTTI; defined as the simulated time in days from a patient's system entry at intake to the start of their first definitive treatment event), system throughput (patients per year), governance event density, and resource use profiles.

Transition Pathway Optimization

To identify the most robust adoption sequence from A=0 and I=0 to A=1 and I=1, we modeled the simulation grid as a weighted directed graph. Transitions were restricted to orthogonal movements (increasing either A or I but not both simultaneously) to simulate realistic, incremental implementation capacity. A composite cost function was applied to determine the trajectory that minimizes the cumulative operational friction. The cost function was defined as the sum of 4 equally weighted (w=0.25) normalized components: mean RTTI, RTTI SD, inverse throughput volume, and throughput SD. We applied the shortest path algorithm by Dijkstra [27] to minimize this cumulative operational cost, implemented using the NetworkX library [28]. Conceptually, this treated the adoption journey as a risk landscape navigation problem, allowing us to mathematically identify the specific sequence of investments that reaches full integration while maximizing efficiency, throughput, and stability. This optimal path was benchmarked against 2 corner strategies: automation first (fully maximizing A before I) and data interoperability first (fully maximizing I before A).

Ethical Considerations

Ethical review was not applied for. Because this research relied entirely on a theoretical DES using synthetic parameters and did not involve human participants or identifiable personal health information, it was exempt from institutional review board oversight in accordance with the statutes of the Ethics Committee of the Technical University of Munich [29].

Results

Operational Landscape of AI Adoption

The scenario-planning phase defined a 2D parameter space using AI automation (A) and data interoperability (I). This yielded 4 scenarios ($S_{I\text{-}IV}$) with corresponding operational constraints and failure modes (Figure 1).

XSL•FO
RenderX

Figure 1. Conceptual 2 × 2 scenario analysis matrix for artificial intelligence (AI) adoption in oncology. The strategic state space is defined by AI automation intensity (x-axis) and data interoperability (y-axis). The intersection yields 4 distinct operational scenarios ($S_{I\text{-}IV}$). Each quadrant details the anticipated dominant workflow characteristics, structural constraints, and primary expected operational risk.

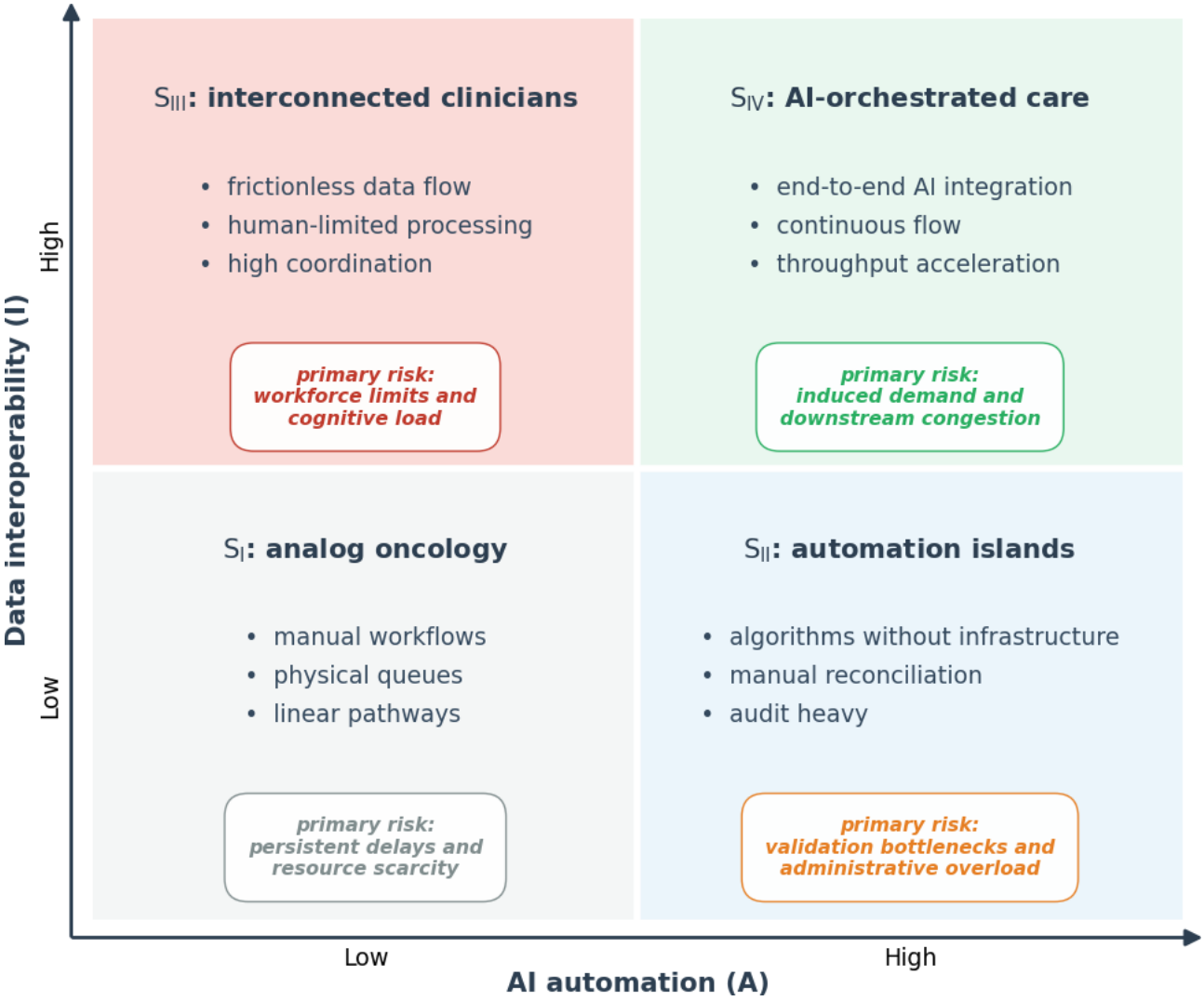

The analog oncology scenario (S_I: low A and low I) represents the baseline state, where capacity is dictated by physical constraints and linear, manual workflows. Here, digitization is localized, and delays are governed by visible queues in administration and diagnostics.

Increasing only AI automation (S_{II}: high A and low I) creates automation islands. Here, algorithmic deployment outpaces data interoperability, and manual data reconciliation increases governance load, shifting the bottleneck from clinical decision-making to oversight and compliance. This represents an automation paradox where isolated digital acceleration degrades overall system performance.

The interconnected clinicians scenario (S_{III}: low A and high I) describes an environment of high data interoperability but conservative AI automation. While data stewardship minimizes duplication and enables coordination, overall throughput remains capped by human cognitive limits and workforce availability.

Finally, the AI-orchestrated care system (S_{IV}: high A and high I) integrates interoperable data with high AI adoption. This tends to remove upstream diagnostic constraints as they are more prone to algorithmic improvements. However, the system accelerates patient flow sufficiently to expose rigid downstream capacities in treatment delivery, and lower wait times induce further demand.

Simulation With Emergent System Dynamics

To complement the qualitative scenario analysis, our DES captured higher-order dynamics and feedback loops such as governance work, virtual multidisciplinary team adoption, and induced demand (Figures S1A and S1B in Multimedia Appendix 3). The simulation also included operational strain due to increased throughput. As shown in Figure S1C in Multimedia Appendix 3, the high sensitivity of AI triage tools increased the false positive rate. This influx placed a disproportionate load on diagnostic resources, resulting in low-yield investigations that displaced confirmed cancer cases and competed for biopsy and histology capacity.

Scenario Signatures and Performance Divergence

The simulation across the $A \times I$ parameter space generated distinct operational signatures for each scenario quadrant. As summarized in Table 1, the baseline (S_I) functioned as a

XSL•FO
RenderX

capacity-constrained system with a mean RTTI of 29.4 days and an annual throughput of 1022 patients. In the automation islands scenario (S_{II}), we observed the worst system performance, with mean RTTI increasing to 37.1 days, a 26% increase compared to the analog baseline, whereas throughput decreased to 647 patients per year. Transitioning to interconnected clinicians (S_{III}) yielded moderate improvements, reducing RTTI to 27.3 days. The AI-orchestrated care system (S_{IV}) resulted in the highest patient flow, reducing the average RTTI to 14.9 days and increasing throughput to 1244 patients per year. Task-level improvements did not translate linearly into system performance (Figure 2). While the trajectory toward S_{IV} showed a combined increase in efficiency (lower RTTI) and number of treated patients, S_{II} resulted in higher RTTI and reduced throughput, suggesting that AI automation intensity may become counterproductive when isolated from data interoperability.

Table . Operational performance metric values by scenario representing the mean of 10 stochastic replications.

Scenario	A [a]	I [b]	RTTI[c] (d), mean (SD)	Patients per year, mean (SD)	Primary bottleneck
S_I: analog oncology	0.0	0.0	29.4 (0.5)	1022 (14.7)	CT[d] scanner
S_{II}: automation islands	1.0	0.0	37.1 (1.3)	647 (10.1)	Governance
S_{III}: interconnected clinicians	0.0	1.0	27.3 (0.3)	1127 (14.0)	Molecular laboratory
S_{IV}: AI[e]-orchestrated care	1.0	1.0	14.9 (0.3)	1244 (21.4)	Chemotherapy infusion units

[a]AI automation intensity variable.

[b]Data interoperability variable.

[c]RTTI: referral-to-treatment interval.

[d]CT: computed tomography.

[e]AI: artificial intelligence.

Figure 2. System-level performance indicators across the simulated adoption landscape. The heat maps illustrate the aggregate operational performance of the oncology pathways based on the discrete-event simulation. The x-axis represents artificial intelligence (AI) automation intensity, and the y-axis represents data interoperability, both normalized from 0.0 to 1.0. Panel (A) displays the mean referral-to-treatment interval measured in simulated days from patient intake to the initiation of definitive treatment. Panel (B) displays the annualized system throughput, quantified as the total number of treatment starts per year. All data points represent the average of 10 stochastic replications for each coordinate on the 11 × 11 parameter grid.

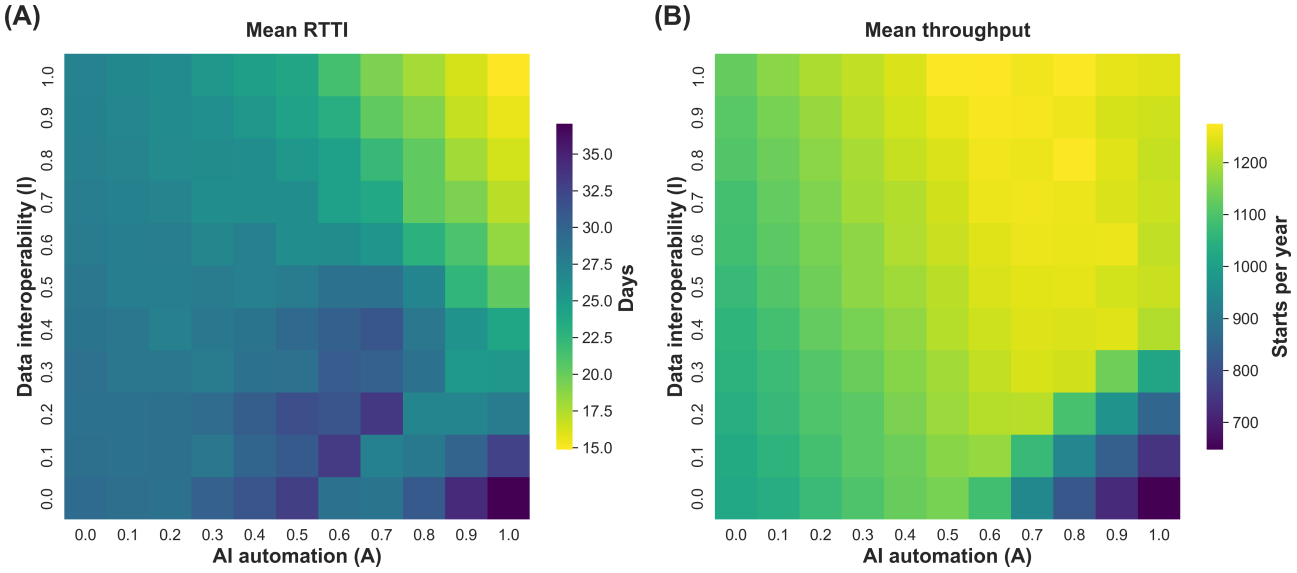

Bottleneck Migration and Resource Dynamics

Further analysis exposed a structural migration of constraints (Figure 3). In low–AI automation scenarios (S_I and S_{III}), constraints were located upstream in diagnostics (computed tomography scanners and molecular laboratory). The automation islands (S_{II}) scenario saw the primary constraint decouple from clinical infrastructure, shifting to administrative governance and quality assurance methods as manual oversight processes became overwhelmed. In the fully transformed S_{IV} scenario, the primary bottleneck shifted downstream to chemotherapy infusion units. Therefore, successful upstream AI adoption transferred capacity pressure from diagnostic infrastructure to labor-intensive treatment resources.

XSL•FO
RenderX

Figure 3. Structural bottleneck migration across the strategic state space: a categorical heat map of the primary binding constraint within the oncology pathway across the $A \times I$ scenario space. The x-axis represents artificial intelligence (AI) automation intensity, and the y-axis represents data interoperability, both normalized from 0.0 to 1.0. For each coordinate on the 11 × 11 grid, the simulation evaluates the operational load across all modeled clinical and administrative queues and identifies the specific resource pool with the highest mean use rate. CT: computed tomography.

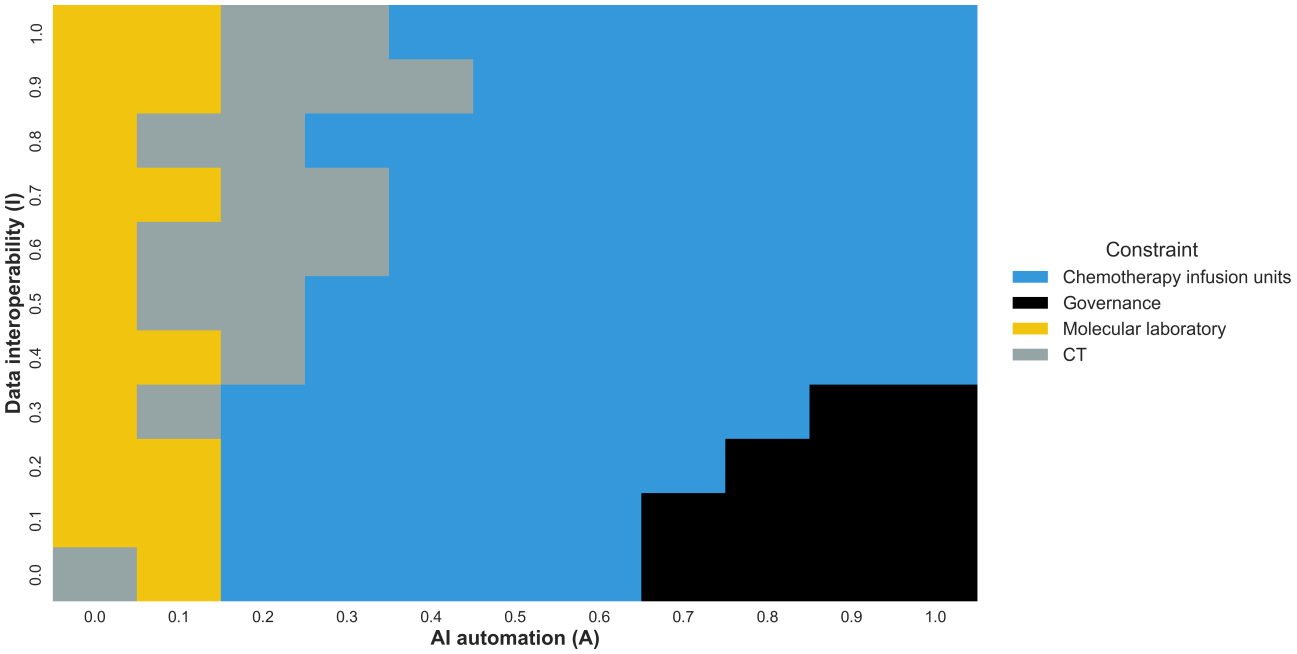

To unpack these constraint shifts, we generated visualization tools that quantify how resource use changes across the $A \times I$ plane. Holding data interoperability constant while varying AI automation (Figure 4A) and holding AI automation constant while varying data interoperability (Figure 4B) yielded distinct use trajectories for each resource. Together, these plots identified crossover points at which the binding constraint moved from upstream diagnostics (computed tomography and laboratory services) to downstream treatment capacity (eg, chemotherapy, surgery, and radiotherapy) or administrative or digital functions (referral processing and governance), indicating where marginal investments likely had the greatest impact.

Figure 4. Resource use sensitivity analysis. The panels display the mean use rates of system resources derived from the discrete-event simulation. Panel (A) illustrates the impact of varying artificial intelligence (AI) automation intensity (A) from 0.0 to 1.0 while holding data interoperability (I) constant at a baseline of 0.5. Panel (B) demonstrates the effect of varying I with A fixed at 0.5. The dashed line represents the theoretical maximum capacity (100% use). CT: computed tomography; RT: radiotherapy.

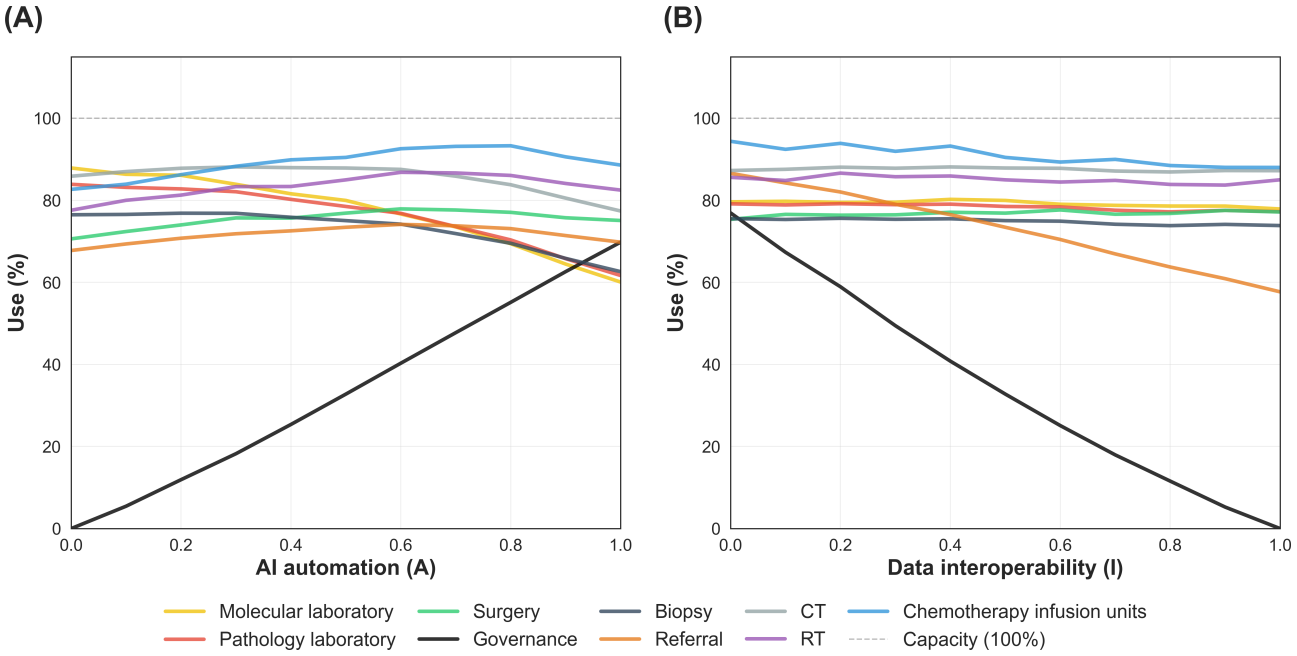

Volatility and Optimal Pathways

Finally, we analyzed the SDs of the output parameters across 10 stochastic replications, treating volatility as a proxy for system stability. The SDs for RTTI and throughput were highest in the transition from S_{II} to S_{IV}, indicating that uncoordinated AI adoption may create unpredictable wait times driven by

XSL•FO
RenderX

sporadic governance backlogs (Figures S1D-F in Multimedia Appendix 3). In contrast, establishing high data interoperability first (S_{III}) acted as a stabilizing force, allowing for the subsequent integration of AI automation. To determine the most robust trajectory from analog operations (S_I) to the AI-orchestrated system (S_{IV}), we performed a graph-based optimization minimizing a composite cost function of RTTI, throughput, and their respective operational variances (Figure 5A). The model selected an adoption sequence characterized by a structural preference for data interoperability–led scaling, prioritizing early gains in *I* ahead of *A*. This sequencing effectively reduced midtransition volatility, allowing the system to absorb the governance load of incremental AI automation while maintaining high throughput efficiency throughout the implementation life cycle (Figure 5B).

Figure 5. Pathway optimization and transition dynamics. Panel (A) presents a trajectory analysis across the artificial intelligence (AI) automation and data interoperability state space, modeled as a weighted directed graph. The background heat map indicates system volatility (defined as the mean of cross-replication SDs of the referral-to-treatment interval [RTTI] and annualized throughput), with darker regions representing high instability. The green line traces the computationally optimal path based on the weighted cost function, benchmarked against 2 alternative corner strategies: an automation-first approach (red line) and an interoperability-first approach (blue line). Panel (B) tracks key system metrics throughout the implementation process. Normalized metrics for efficiency (inverted RTTI), throughput volume, and stability (inverted system volatility) are plotted against transition progress (0%-100%).

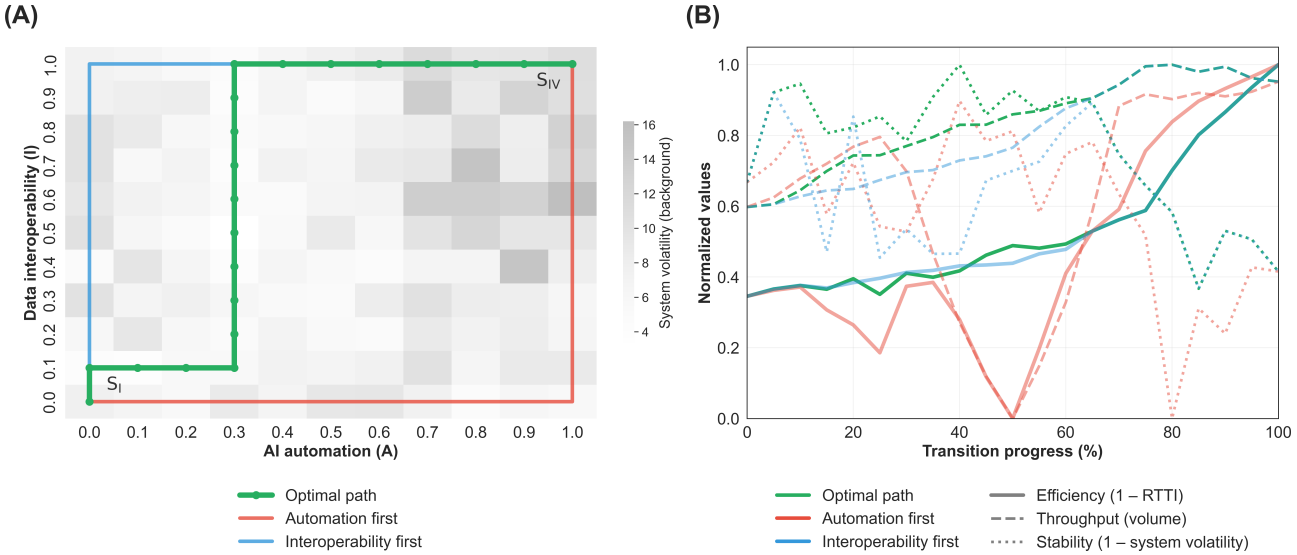

Discussion

Strategic Utility for Health System Planning

This paper addresses the gap between the rapid pace of technological transformation in medicine and the current underuse of applied futures methods, offering a novel approach to operational planning under uncertainty [20]. Traditional business cases for AI often rely on static efficiency assumptions, presuming that diagnostic time reductions linearly increase throughput. However, such linear extrapolations ignore endogenous feedback loops inherent in complex systems. By coupling scenario planning with quantified simulations, this study provides a framework to stress-test AI adoption strategies against plausible futures—making risks visible to health system planners before decisions are formed and capital is committed.

Because the model is parameterized around 2 fundamental strategic levers (AI automation intensity and data interoperability), it can inform both clinic-level operational redesign and broader health system–level investment and governance planning. To illustrate how health systems can operationalize this framework across different organizational levels, we have outlined 3 concrete implementation scenarios (Table S2 in Multimedia Appendix 2). These scenarios demonstrate how micro-level hospitals, meso-level regional networks, and macro-level state health authorities can use this framework to anticipate bottleneck migration, sequence investments, and mitigate transition volatility.

Robust Strategies and Sequencing

A key insight derived from the scenario sweep is the phenomenon of bottleneck migration, necessitating a shift from local optimization to whole-system capacity planning. The simulation indicates that successful AI adoption in diagnostics tends to shift pressure downstream in the model. Therefore, executives must couple upstream AI investments with adaptive capacity planning for surgery, chemotherapy, and radiotherapy. Similarly, the emergence of higher false positive rates suggests that adoption strategies might require new guardrails such as confirmatory testing thresholds to prevent the system from being overwhelmed in scenarios of increased induced demand. At the same time, scenarios of high AI automation and low data interoperability should be avoided as they can lead to significant governance overload. This aligns with recent critiques of health workforce planning, which note that failure to account for social interactions and power dynamics, including governance, often undermines technical implementation [30].

Moreover, results from our pathway optimization demonstrated that a coordinated adoption strategy with early focus on data interoperability yielded the best overall results and stability. Because high volatility manifests in real-world settings as unpredictability and may lead to stakeholder resistance,

XSL•FO
RenderX

implementation strategies must manage transition stability alongside performance optimization. This underlines the importance of sequencing—via phased rollouts or adaptive governance—to minimize disturbances during transformation. Conceptually, this optimization mirrored a backcasting approach, working backward from a desired end state to identify near-term moves that preserve system stability [31].

Limitations and Future Directions

This study presents a mechanism-oriented framework prioritizing structural behavior over precise numerical forecasting. Importantly, the simulation is based on stylized parameters rather than real-world data. As an exploratory proof of concept, it has several limitations. First, parameters are based on conceptual assumptions rather than site-level empirical data, limiting the generalizability of specific point estimates. Alternative parameterizations could yield different quantitative results; thus, findings should be interpreted as directional hypotheses rather than definitive empirical conclusions. Second, the model is restricted to 3 tumor types and 2 stages, which does not capture the full complexity of a clinical case mix. Third, the simulation excludes financial costs, clinical end points, and equity impacts. Finally, adoption is modeled as a static state rather than a time-varying implementation process, simplifying the realities of organizational learning and change management.

To support applied decision-making, future work should calibrate parameters to local workflow distributions and resource profiles and validate model behavior against observed pathway metrics where available to assess external validity. Planned extensions include formal sensitivity and uncertainty analyses over capacities, service time assumptions, and objective function weights used in the transition optimization. Equity should be incorporated explicitly by stratifying patient agents by sociodemographic risk factors (eg, deprivation, language barriers, and digital access) and modeling differential access, adherence, and delay mechanisms to test whether accelerated pathways widen disparities [32,33]. Additional behavioral realism could be introduced by modeling clinician oversight burden, trust calibration, and AI literacy as determinants of effective automation and governance load [34,35].

Conclusions

The integration of qualitative scenario planning with quantitative simulations offers a tool for testing plausible oncology AI strategies under uncertainty. AI can dramatically improve oncology operations but only when paired with interoperable data infrastructure and system-level capacity planning; otherwise, it risks worsening delays and instability. Simulating interactions between AI automation and data interoperability exposed operational tipping points, bottlenecks, and instabilities invisible to static planning models. Future research should extend this hybrid framework and integrate financial and equity dimensions to better anticipate the societal costs of technological disruption. This could help establish simulation-based scenario planning as an important tool in medical futures studies.

Funding

This research received no external funding and was supported by internal institutional funds.

Data Availability

The datasets generated during this study are available from the corresponding author on reasonable request. The full simulation code is provided in the supplementary material.

Authors' Contributions

Conceptualization: PM, JvB
Data curation: PM, JvB
Formal analysis: PM, JvB
Methodology: PM
Project administration: SDB-M, JvB
Supervision: SDB-M, JvB
Writing—original draft: PM, EG
Writing—review and editing: PM, EG, SDB-M, JvB

Conflicts of Interest

Conflict for author JvB: Employee Sidekick Health Germany GmbH

Multimedia Appendix 1
Code for discrete-event simulation.
[DOCX File, 34 KB - jmir_v28i1e92642_app1.docx]

Multimedia Appendix 2
Detailed operational mechanisms of the discrete-event simulation and applied case scenarios for operationalizing the framework.
[XLSX File, 12 KB - jmir_v28i1e92642_app2.xlsx]

XSL•FO
RenderX

Multimedia Appendix 3
Secondary operational metrics and uncertainty surfaces; visualization of the mechanism-based outputs for governance load (administrative audits per patient arrival), virtual multidisciplinary team adoption rates, and the false positive rate (defined as the proportion of biopsies triggered via artificial intelligence triage that results in a benign finding); and SDs of key performance indicators across 10 stochastic replications.
[DOCX File, 552 KB - jmir_v28i1e92642_app3.docx]

References

1. May P, Greß J, Seidel C, et al. Enabling just-in-time clinical oncology analysis with large language models: feasibility and validation study using unstructured synthetic data. JMIR Med Inform 2025 Dec 1;13:e78332. [doi: 10.2196/78332] [Medline: 41328496]
2. Chen D, Alnassar SA, Avison KE, Huang RS, Raman S. Large language model applications for health information extraction in oncology: scoping review. JMIR Cancer 2025 Mar 28;11:e65984. [doi: 10.2196/65984] [Medline: 40153782]
3. Lammert J, Pfarr N, Kuligin L, et al. Large language models-enabled digital twins for precision medicine in rare gynecological tumors. NPJ Digit Med 2025 Jul 9;8(1):420. [doi: 10.1038/s41746-025-01810-z] [Medline: 40634659]
4. May P, Nokodian S, Nuernbergk C, et al. Artificial intelligence-assisted error detection in complex clinical documentation: leveraging large language models to enhance patient safety in oncology. JCO Clin Cancer Inform 2026 Jan;10:e2500194. [doi: 10.1200/CCI-25-00194] [Medline: 41494139]
5. Ferber D, El Nahhas OSM, Wölflein G, et al. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. Nat Cancer 2025 Aug;6(8):1337-1349. [doi: 10.1038/s43018-025-00991-6] [Medline: 40481323]
6. Chow R, Midroni J, Kaur J, et al. Use of artificial intelligence for cancer clinical trial enrollment: a systematic review and meta-analysis. J Natl Cancer Inst 2023 Apr 11;115(4):365-374. [doi: 10.1093/jnci/djad013] [Medline: 36688707]
7. Bandi A, Kongari B, Naguru R, Pasnoor S, Vilipala SV. The rise of agentic AI: a review of definitions, frameworks, architectures, applications, evaluation metrics, and challenges. Future Internet 2025;17(9):404. [doi: 10.3390/fi17090404]
8. Karunanayake N. Next-generation agentic AI for transforming healthcare. Inform Health 2025 Sep;2(2):73-83. [doi: 10.1016/j.infoh.2025.03.001]
9. Carl N, Schramm F, Haggenmüller S, et al. Large language model use in clinical oncology. NPJ Precis Oncol 2024 Oct 23;8:240. [doi: 10.1038/s41698-024-00733-4] [Medline: 39443582]
10. Ringeval M, Etindele Sosso FA, Cousineau M, Paré G. Advancing health care with digital twins: meta-review of applications and implementation challenges. J Med Internet Res 2025 Feb 19;27:e69544. [doi: 10.2196/69544] [Medline: 39969978]
11. Hao Y, Qiu Z, Holmes J, et al. Large language model integrations in cancer decision-making: a systematic review and meta-analysis. NPJ Digit Med 2025 Jul 17;8(1):450. [doi: 10.1038/s41746-025-01824-7] [Medline: 40676129]
12. Huhulea EN, Huang L, Eng S, et al. Artificial intelligence advancements in oncology: a review of current trends and future directions. Biomedicines 2025 Apr 13;13(4):951. [doi: 10.3390/biomedicines13040951] [Medline: 40299653]
13. Katonai G, Arvai N, Mesko B. AI and primary care: scoping review. J Med Internet Res 2025 Aug 15;27:e65950. [doi: 10.2196/65950] [Medline: 40836794]
14. Patel MS, Dee EC, Lee NY. AI and human expertise in cancer care - striving for synergy. Nat Rev Clin Oncol 2026 Feb;23(2):87-88. [doi: 10.1038/s41571-025-01108-9] [Medline: 41291250]
15. Stevens ER, Alfaro Arias V, Luu S, Lawrence K, Groom L. Technology integration to support nurses in an “inpatient room of the future”: qualitative analysis. J Med Internet Res 2025 Jun 16;27:e68689. [doi: 10.2196/68689] [Medline: 40522717]
16. McCann L, Lewis L, Oduntan O, et al. Patients’ and clinicians’ experiences using a real-time remote monitoring system for chemotherapy symptom management (ASyMS): qualitative study. J Med Internet Res 2024 Dec 3;26:e53834. [doi: 10.2196/53834] [Medline: 39626227]
17. Meskó B, Drobni Z, Bényei É, Gergely B, Győrffy Z. Digital health is a cultural transformation of traditional healthcare. Mhealth 2017;3:38. [doi: 10.21037/mhealth.2017.08.07] [Medline: 29184890]
18. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 6;6:120. [doi: 10.1038/s41746-023-00873-0]
19. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ 2022 May 18;377:e070904. [doi: 10.1136/bmj-2022-070904] [Medline: 35584845]
20. Meskó B, Kristóf T, Dhunnoo P, Árvai N, Katonai G. Exploring the need for medical futures studies: insights from a scoping review of health care foresight. J Med Internet Res 2024 Oct 9;26:e57148. [doi: 10.2196/57148] [Medline: 39383528]
21. Mesko B, Kristóf T, Dhunnoo P, Árvai N, Katonai G. A practical guide to using futures methods in health care: approaches, applications, and case studies. J Med Internet Res 2025 Sep 16;27:e82820. [doi: 10.2196/82820] [Medline: 40956602]
22. Mesko B. The emergence of medical futures studies uncovers medicine and healthcare’s untapped potential. NPJ Digit Med 2024 Dec 19;7:365. [doi: 10.1038/s41746-024-01365-5]
23. Schoemaker PJ. Scenario planning: a tool for strategic thinking. Sloan Manag Rev 1995;36(2):25-40 [FREE Full text]

XSL•FO
RenderX

24. Gross M, Dewan A, Sabharwal K, et al. Decentralized biobanking pathway to precision medicine: futures study. J Med Internet Res 2025 Dec 1;27:e73965. [doi: 10.2196/73965] [Medline: 41328480]
25. Giunti G, Glynn R, Hennessy J, Doherty CP. Participatory design approach in the use of scenario analysis for futureproofing medical education: case study. J Med Internet Res 2025 Sep 23;27:e73173. [doi: 10.2196/73173] [Medline: 40986915]
26. Babashov V, Aivas I, Begen MA, et al. Reducing patient waiting times for radiation therapy and improving the treatment planning process: a discrete-event simulation model (radiation treatment planning). Clin Oncol (R Coll Radiol) 2017 Jun;29(6):385-391. [doi: 10.1016/j.clon.2017.01.039] [Medline: 28222957]
27. Dijkstra EW. A note on two problems in connexion with graphs. Numer Math 1959 Dec;1(1):269-271. [doi: 10.1007/BF01386390]
28. Hagberg AA, Schult DA, Swart P. Exploring network structure, dynamics, and function using NetworkX. In: Varoquaux G, Vaught T, Millman J, editors. Proceedings of 7th Python in Science Conference (SciPy2008): SciPy; 2008:11-15.
29. Satzung der Ethikkommission der Technischen Universität München – nicht-medizinische Fachgruppe. Technische Universität München. 2023 Oct 23. URL: https://portal.mytum.de/kompass/rechtsicherheitswesen/satzung_ethikkommission_nichtmed.pdf/download [accessed 2026-05-19]
30. Rees GH, Crampton P, Gauld R, MacDonell S. Rethinking health workforce planning: capturing health system social and power interactions through actor analysis. Futures 2018 May;99:16-27. [doi: 10.1016/j.futures.2018.03.009]
31. Dreborg KH. Essence of backcasting. Futures 1996 Nov;28(9):813-828. [doi: 10.1016/S0016-3287(96)00044-4]
32. Norris RP, Fuller E, Greystoke A, Todd A, Sharp L. Routes to diagnosis in lung cancer-do socio-demographics matter? An English population-based study. Cancers (Basel) 2025 Jun 3;17(11):1874. [doi: 10.3390/cancers17111874] [Medline: 40507354]
33. Gerrard AD, Coxon J, Maeda Y, Theodoratou E, Dunlop MG, Din FV. Colorectal cancer prevalence in faecal immunochemical test non-returners: potential for health inequality in symptomatic referral pathways. BJS Open 2024 Sep 3;8(5):zrae119. [doi: 10.1093/bjsopen/zrae119] [Medline: 39404039]
34. Arvai N, Katonai G, Mesko B. Health care professionals' concerns about medical AI and psychological barriers and strategies for successful implementation: scoping review. J Med Internet Res 2025 Apr 23;27:e66986. [doi: 10.2196/66986] [Medline: 40267462]
35. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023 Oct 4;25:e50638. [doi: 10.2196/50638] [Medline: 37792434]

Abbreviations

AI: artificial intelligence
DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence
DES: discrete-event simulation
RTTI: referral-to-treatment interval

Edited by B Mesko, T Kristóf; submitted 06.Feb.2026; peer-reviewed by A Bethanabatla; revised version received 28.Apr.2026; accepted 29.Apr.2026; published 25.May.2026.

Please cite as:
May P, Brookman-May SD, Garrahy E, von Büren J
Sequencing AI Automation and Data Interoperability in Oncology Using a Scenario-Planning Framework Coupled With Discrete-Event Simulation: Proof-of-Concept Study
J Med Internet Res 2026;28:e92642
URL: https://www.jmir.org/2026/1/e92642
doi:10.2196/92642

© Peter May, Sabine D Brookman-May, Edward Garrahy, Johannes von Büren. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Horizon Scanning for Medical Technologies: Methodological Framework Development Study of the Medical Innovation Scanning Techniques (MIST) Framework

Sonia Garcia Gonzalez-Moral, MSc; Andrew Mkwashi, PhD; Fiona R Beyer, PhD; Dawn Craig, MSc

NIHR Innovation Observatory at Population Health Sciences Institute, Newcastle University, The Catalyst Room 3.12, 3 Science Square Newcastle Helix, Newcastle upon Tyne, England, United Kingdom

Corresponding Author:
Sonia Garcia Gonzalez-Moral, MSc
NIHR Innovation Observatory at Population Health Sciences Institute, Newcastle University, The Catalyst Room 3.12, 3 Science Square Newcastle Helix, Newcastle upon Tyne, England, United Kingdom

Abstract

Background: Strategic foresight relies on horizon scanning to detect weak signals of innovation and anticipate disruption in fast-moving MedTech. While tools such as the Rumsfeld matrix and the Three Horizons model structure thinking about uncertainty and time, their application in MedTech is limited and insufficient for addressing the regulatory, technical, and evidential complexity of emerging or disruptive technologies. Regulatory frameworks provide stability through defined safety and performance pathways, but shifts and regional variability requirements complicate anticipatory assessment. These challenges expose methodological gaps in current practice: broad foresight frameworks lack the technical granularity, regulatory alignment, and systematic processes required for MedTech, where rapid innovation intensifies uncertainty. This paper presents a structured methodological framework to support systematic, reproducible, and decision-focused horizon scanning in MedTech.

Objective: The framework aims to (1) standardize methods for planning, execution, and reporting; (2) reduce uncertainty through systematic identification and interpretation of weak signals; and (3) improve transparency, quality, and comparability across policy, regulatory, clinical, and strategic contexts.

Methods: The framework was developed following the 3-step process by McMeekin et al. First, data were identified through a mapping review of MedTech futures and foresight methods, mapping UK health care decision-makers, and participating in national and international horizon scanning initiatives (2021 - 2025). Second, the framework was constructed by integrating the sources, methods, and insights with the principles of regulatory and technology readiness, supported by ongoing consultation with UK regulatory and health system bodies. Third, the framework is being applied and tested in MedTech horizon scanning projects. Formal validation and iterative revision are planned.

Results: The MedTech Innovation Scanning Techniques (MIST) framework integrates foresight theory, horizon scanning methods, and the MedTech innovation pathway, enabling systematic weak signal detection across 3 time horizons. In the imminent horizon (technology readiness level [TRL] 8 - 9), uncertainty is lowest, and weak signals are more visible, supported by rapid evidence synthesis and expert engagement. In the transitional horizon (TRL 4 - 7), uncertainty increases due to trial outcomes, funding, and market dynamics, requiring broader evidence sources and complementary techniques such as scenario analysis and bibliometrics. In the emerging horizon (TRL 1 - 3), weak signals are abundant but least predictable; early-stage data (patents, preclinical research, and conference outputs) are analyzed using methods such as term-similarity visualization to reveal nascent innovations. Across all horizons, MIST supports systematic signal identification, contextualization, and prioritization to inform anticipatory decision-making.

Conclusions: MedTech horizon scanning is challenging. MIST addresses these challenges by integrating the Three Horizons model, Rumsfeld matrix, technology pathways, regulatory considerations, and sources to guide weak signal detection. MIST provides a structured, transparent, and reproducible approach tailored to MedTech. While formal validation and integration of equity and ethical dimensions are ongoing, the framework fills a critical methodological gap in horizon scanning for MedTech.

(J Med Internet Res 2026;28:e93166) doi:10.2196/93166

KEYWORDS

horizon scanning; evidence synthesis; information retrieval; methods framework; Medical Innovation Scanning Techniques

XSL•FO
RenderX

Introduction

Overview

Strategic foresight is often referred to as "futures studies" or "futures research" and is a discipline used by organizations to gather and process information about their future operating environment. This may encompass trends and developments that shape the organization's political, economic, social, technological, legal, and environmental context [1].

Horizon scanning, referred to as the "the systematic identification of new and emerging health technologies that have the potential to impact health, health services, and/or society" [2], is a crucial tool for health and care decision-making, enabling health and care systems to anticipate and adapt to rapid technological change; providing policy decision-makers with high-quality, early intelligence on emerging health technologies, to enable them to navigate pressures from innovation demands and constrained resources; while also allowing the wider health and life science sector stakeholders to use the same intelligence to drive innovation in areas of unmet need for the benefit of patients and the economy. As a strategic foresight tool, horizon scanning systematically detects weak signals of technology emergence and development, providing early insights that support system and sector preparedness. This proactive approach not only identifies opportunities for market growth and innovation but also mitigates political and operational risks, giving policymakers the time and evidence needed for informed planning and resilient health systems [3,4].

MedTech refers to the broad spectrum of health services, products, and solutions (classified as medical devices, in vitro diagnostics, or digital health applications) designed to enhance health outcomes and health and care delivery. These technologies play a crucial role in transforming care pathways, improving health, and generating cost efficiencies [5]. MedTech innovations have redefined health care delivery through improved monitoring, accessibility, and diagnostic precision. For example, wearable health monitors such as continuous glucose monitors exemplify how continuous physiological tracking can enhance preventive and chronic disease management. Artificial intelligence (AI)–powered diagnostic tools have revolutionized clinical decision-making by automating disease detection, and robotic surgery systems have enhanced surgical precision and minimized invasiveness. The MedTech field is marked by rapid innovation, with products typically undergoing short lifecycles (18 - 24 months), with frequent updates or entirely new innovations emerging [5]. The development of medical devices involves a complex, iterative, and interdisciplinary process, from needs identification to conceptualization and final product realization [6], engaging clinicians, engineers, business professionals, and patients along the way.

Despite common agreement on horizon scanning as a tool for signal detection, anticipatory innovation decision-making, implementation, dissemination, and unmet needs identification [7] in health care, the heterogeneous application of horizon scanning and foresight methods, coupled with inconsistent reporting and the absence of best practice consensus, hampers the quality and timeliness of MedTech innovation assessments. Furthermore, a recently published mapping review identified that the term "horizon scanning" was inconsistently applied and vaguely defined in the literature, often encompassing a broad mix of data sources and stakeholder consultations [8]. Although many definitions exist, most support the concept that horizon scanning is a method that aims to systematically detect signals of change or "weak signals" which may be understood as early signs of emerging technologies.

The Current Unmet Methodological Need

Horizon scanning in health technology, especially MedTech, faces unique methodological, practical, and conceptual challenges that make the case for providing structured practical guidance and specialized tools to guide users through each step of the process. Below, we discuss this need, the role of standardization and systematization, and why established foresight frameworks (such as Framework Foresight [9]) are not fully translatable to MedTech horizon scanning.

First, MedTech involves fast, iterative innovation, diverse product types with varied evidence needs, complex multistakeholder ecosystems, and safety risks tied to user interaction and system integration. These features require practical, structured Horizon scanning tools that ensure transparency, reproducibility, and context-specific applicability [8,10].

Second, established foresight frameworks such as Framework Foresight [9,11,12] or the European Commission Concurrent Design Foresight [13] offer broad futures-thinking structures but treat horizon scanning as an early, general step rather than a detailed, stand-alone method. They lack the technical, regulatory, and prioritization guidance needed for MedTech, serving more as conceptual guides than practical, MedTech-specific tools.

Lastly, in MedTech, where evidence gaps and complex trends heighten uncertainty, standardized and systematic horizon scanning methods are vital. They promote transparency, consistency, and reproducibility; enable auditable signal detection and prioritization; and use structured tools to reduce bias and missed developments. Literature shows this approach builds trust, supports benchmarking, and fosters shared learning [14].

Aims and Objectives

This paper aims to provide a transparent and detailed account of the development of a structured methodological framework to guide the application of horizon scanning in the context of emerging MedTech innovations. The framework is intended to offer clear, practical steps that support systematic, reproducible, and decision-oriented horizon scanning practice.

Objectives

The development of this framework was guided by three core objectives which were (1) to standardize horizon scanning methods by defining essential methodological components and offering a coherent structure that supports consistent planning, execution, and reporting of horizon scanning activities; (2) to

XSL•FO
RenderX

reduce uncertainty surrounding new and emerging MedTech innovations through systematic identification, assessment, and interpretation of early signs and future developments relevant to health and care; and (3) to strengthen evidence-informed decision-making by improving the transparency, quality, and comparability of horizon scanning outputs used to inform policy, regulatory, clinical, and strategic decisions regarding MedTech innovation.

Methods

Study Design

For the creation of the framework, we followed the steps outlined by McMeekin et al [15] that consisted of (1) identifying the data to inform the methodological framework, (2) developing the framework, and (3) testing and validating this framework in case studies, and refining the framework.

We planned to add an additional step, step 4: revision and refining of the framework. Although this last step is considered by McMeekin et al [15] as part of step 3, given the foreseen impact that changes introduced to the framework may have in practice and the number of potential framework users affected by them, a systematic and transparent approach is essential to ensure credibility and usability whereby all additional change sustained after the publication of the first iteration of the framework will be disseminated separately providing a rationale for the change and a comparative analysis.

Step One: "Identifying the Data"

This step was undertaken with a mapping review of futures and foresight methods used for the identification of emerging and innovative MedTech [8]. This review identified different sources and purposes for the practice of horizon scanning in the context of MedTech and the main range of stakeholders involved in the process. These provided the epistemological foundation of the framework. Additionally, to make this framework more responsive to UK horizon scanning practice, we mapped the characteristics of all national health care decision-makers across the UK infrastructure [16]. A total of 39 structural bodies with decision-making capacity in health care were identified, as well as their interbody relationship within the system. This stakeholder mapping supported the identification of priority areas and gaps within the system and provided the basis for the design of proactive horizon scanning. Alongside these 2 major projects, between 2021 and 2025, members of this team (SGG-M and AM) participated in several national horizon scanning initiatives (Table 1).

Table . Horizon scanning initiatives.

Name of initiative	Purpose	Dates	Attendee
AAC[a] horizon scanning working group	Devise methods for horizon scanning innovative health care technologies	2020 - 2022	SGG-M
Medicines and Medical Devices Initiative Medical Devices Working Group	Coordinate different stakeholder needs to streamline horizon scanning practice and increase efficiency across the system	2022 - 2025 2023-2025	SGG-M AM
Nuffield Council on Bioethics	Identify opportunities for embedding ethics in the horizon scanning of new technologies	2023 - 2025	SGG-M
International Horizon Scanning Initiative	Establish a horizon scanning service and terms of reference for standardizing the practice of horizon scanning for MedTech from different countries' economic perspectives, needs, and regulatory frameworks	2022 - 2025	SGG-M
International HealthTechScan	Promotes early awareness via horizon scanning practice	2020 - 2025 2023 - 2025	SGG-M AM

[a]AAC: Accelerated Access Collaborative.

Step Two: "Developing the Methodological Framework"

This step encompassed several existing methods, theories, sources, and regulations that mitigate the uncertainty of the unknowns. This was mainly done by integrating various horizon scanning methods outlined in the literature [17], innovation frameworks [18], sources identified from horizon scanning practice and from the literature [8,17,19], working group insights, and relevant publications [10], professional experience with established technology readiness levels (TRLs) [20,21], and regulatory frameworks [22-25] to understand how to work with uncertainty [1,26]. The regulatory aspect of the framework is a constantly evolving element as regulators respond to innovation. To understand the latest challenges and initiatives, informational consultation with national stakeholders, including the Medicines and Healthcare Products Regulatory Agency, Healthcare Improvement Scotland, and Health Technology Wales, was undertaken at regular intervals. This step was undertaken iteratively, being informed by the constant change in the regulatory landscape and ad hoc, during "contact points" with regulatory experts via online meetings.

XSL•FO
RenderX

Step Three: "Testing and Validating the Methodological Framework"

This step is ongoing as one of the core foundational commitments of the Innovation Observatory (IO)—the UK national horizon scanning center—to methods development in this field. The authors, in their capacity as researchers, are taking an agile approach to testing and validating this methodological framework implicit to all IO's MedTech horizon scanning projects. The application of this framework starts with the scoping of the topic, the establishment of the horizon, sources, and TRLs, as well as the analysis and reporting of the data. A formal validation of this framework is expected in the near future.

Ethical Considerations

This study did not involve human participants, patient data, or identifiable personal information. The work consisted solely of methodological development activities conducted by professionals in their occupational roles and analysis of publicly available institutional reports. According to institutional and national research governance guidance, this type of study does not constitute human subjects research and therefore did not require review or approval by an institutional review board or research ethics committee [27]. All procedures were conducted in accordance with applicable institutional, national, and international standards for research integrity and publication ethics [28].

Results

Step 1: Identifying the Data

Main Sources of Weak Signal Detection

Through our previously published research [8] and expert horizon scanning practice, we have identified the breadth and relative frequency of information sources used in horizon scanning activities (Table 2). Specialist journals and medical bibliographic databases dominate, reflecting the continued centrality of the academic literature. These are followed by patent registries, expert input, and outputs from other horizon-scanning agencies, highlighting the importance of early-stage innovation intelligence and peer networks. Clinical trial registries and industry, funders, or regulatory sources occupy a midtier, supporting pipeline visibility and regulatory awareness. A wide array of supplementary sources, such as commercial research and development databases, web searches, news outlets, social media, and conference outputs, are used less frequently but contribute to signal detection and contextual awareness. Collectively, the distribution and heterogeneity of sources demonstrate that effective horizon scanning relies on triangulating formal scientific evidence with regulatory, funding, commercial, and informal intelligence streams to capture emerging technologies at different stages of maturity.

XSL•FO
RenderX

Table . Sources used in MedTech horizon scanning.

Category and name	Content	Access
Bibliographic databases		
IEEE Xplore	• Engineering	Free
EMBASE (Ovid)	• Medical	Paywall
Engineering Village (Elsevier)	• Engineering	Paywall
Specialist publications		Free
Medical device network	• MedTech news, regulatory news, and company intelligence.	
Med-tech news	• MedTech innovations, market intelligence, companies' news, and health and social care policy news.	
Medical technology outlook	• MedTech news, conferences, and insights.	
Today's medical developments	• Industry news, manufacturers, components and materials, medical technology innovators, and regulatory decisions.	
Clinical trial registers		Free
ClinicalTrials.gov	• Clinical trial protocols of ongoing and completed research. US and non-US trials.	
EUdraCT; CTIS[b]	• EUdraCT: Clinical trial protocols of ongoing and completed studies undertaken in any EU[a] country up to January 2022. • CTIS: From January 2022.	
CTIS	• New European clinical trial platform for all ongoing clinical trials in the EU. Supersedes the previous clinical trial portal (included above)	
WHO[c] ICTRP[d] Search Portal	• ICTRP from the WHO. International scope.	
ScanMedicine	• A one-stop shop for global clinical trial studies ongoing or completed, registered in more than 11 different national registers, including the US and the European Union. Developed by the NIHR Innovation Observatory, this source is free.	
Regulatory databases		
FDA[e] Medical Devices	• PMA[f], 510(k) premarket notifications and De Novo	Free
Patent databases		
Lens.org	• Over 150 million patent records with a database that is updated every 2 weeks. International patent information is available from 106 jurisdictions.	Free[g]

XSL•FO
RenderX

Category and name	Content	Access
PATENTSCOPE	• PATENTSCOPE is a free patent database made available by the WIPO[h] with coverage of over 60 patent collections. It is the primary and thus most up-to-date and complete source of data on patent applications filed under the PCT[i]. It contains 115.7 million patent documents.	Free
News		
Google News (Google LLC)	• News aggregator by Google LLC. It provides access to thousands of news outlets and is constantly updated.	Free
Funding and grants		
Various sources	• Funding calls (open and closed) • Grants and principal investigators	Free

[a]EU: European Union.

[b]CTIS: Clinical Trials Information System.

[c]WHO: World Health Organization.

[d]ICTRP: International Clinical Trials Registry Platform.

[e]FDA: US Food and Drug Administration.

[f]PMA: premarket approvals.

[g]Open Public Access for individuals wanting to search and analyze patents and scholarly works. Access for organizations or to additional resources is subscription-only.

[h]WIPO: World Intellectual Property Organization.

[i]PCT: Patent Cooperation Treaty.

The UK Stakeholder Map

The UK health care decision-making landscape is characterized by a complex network of organizations, spanning mandatory decision-makers and advisory bodies, many of which engage directly with emerging MedTech innovations. In 2023, we mapped this ecosystem through iterative and snowballing searches, identifying 32 heterogeneous structures grouped into 7 functional categories (Figure 1): certification and health technology assessments (HTAs), advisory or leadership, clinical guidelines, health delivery, regulation and licensing, innovation accelerators, and access pathways. Several entities operate across multiple categories, highlighting the intricacy and interdependence of decision-making processes governing the adoption, implementation, and diffusion of new technologies [16]. Within this multiactor environment, horizon scanning functions as a cross-cutting intelligence layer; the 10 applications included in Table 3 have been extracted from published literature [8] and align closely with the distinct roles of these stakeholder groups.

XSL•FO
RenderX

Figure 1. UK stakeholders map. DH&SS: Department for Health and Social Services; DHSC: Department of Health and Social Care; DoH: Department of Health; H&SC: Health and Social Care; HSC: Health and Social Care; NHS: National Health Service; NI: Northern Ireland; OHID: Office for Health Improvement and Disparities; UKHSA: United Kingdom Health Security Agency.

XSL•FO
RenderX

Table . Application and use of horizon scanning by UK stakeholders.

Horizon-scanning purpose	Primary UK stakeholder groups served	Role in decision-making
Early warning	Advisory and leadership bodies; regulators; innovation accelerators	Anticipate disruptive or high-impact technologies; prepare policy, regulatory, and system responses
Identify emerging trends	Advisory and leadership; HTA[a] bodies; innovation accelerators	Sense weak signals and trajectory shifts across MedTech domains
Identify future developments	HTA; clinical guideline developers; health delivery organizations	Inform pipeline awareness and forward planning
Support decision-making	HTA agencies; advisory bodies; commissioners	Provide structured intelligence to underpin formal recommendations
Technology assessment	HTA bodies; regulators	Frame technologies for evaluative processes and readiness appraisal
Planning	Health delivery; pathway and access bodies; NHS system planners	Enable workforce, infrastructure, and service planning
Predict impact in health care services	Health delivery; commissioners; advisory bodies	Anticipate system-level, clinical, and resource implications
Identify evidence gaps	HTA; research funders; guideline developers	Shape research agendas and reduce uncertainty
Adoption and diffusion	Pathway and access bodies; health delivery organizations	Accelerate appropriate uptake and scale-up
Identify business opportunities	Innovation accelerators; industry-facing bodies	Support innovation ecosystems and investment decisions

[a]HTA: health technology assessment.

Together, these findings provide the empirical backbone for the MedTech Innovation Scanning Techniques (MIST) framework by demonstrating that effective MedTech horizon scanning is inherently multisource, multistage, and stakeholder-contingent. The observed stratification of information sources (from specialist journals and bibliographic databases, through patents, trials, and regulatory data, to informal and gray channels) maps directly onto different phases of technological maturity and readiness. MIST operationalizes this by structuring scanning activities across foresight stages and technology-readiness levels, specifying which sources are most informative at each stage and why. The UK stakeholder mapping further grounds MIST in real-world decision contexts: the 10 horizon-scanning applications align with distinct functional roles across HTA, regulation, service delivery, and innovation acceleration. MIST provides a coherent methodological architecture that brings together interdependent methods, techniques, tools, and instruments, used systematically toward a common purpose and across multiple applications. It reflects horizon scanning not as a single method or tool, but as a configurable intelligence system, and translates that complexity into a structured, purpose-driven, and reproducible approach to MedTech horizon scanning by linking source typologies, innovation maturity, and the anticipatory needs of heterogeneous decision-makers.

Step 2: Developing the Framework

Horizon Scanning Methods

Beyond the core methods such as expert panels, patent analysis, literature searching, stakeholder engagement, horizon scanning, and road mapping already identified [8,19], studies variably used Delphi, trend analysis, patient engagement, and text mining, alongside a long tail of infrequently used approaches, including clinical trial analysis, environmental scanning, projective scenarios, overlay mapping, and strategic foresight [8].

Across all studies, horizon scanning itself is the most frequently reported method, followed by expert consultation, literature searching, and patent analysis.

Methods for Working With Uncertainty

Horizon scanning encompasses weak signals of varying strength. Weak signals represent early indications of potential future developments and are typically characterized by a high degree of uncertainty. It is not always clear whether such weak signals are reliable, as errors in detection or data quality may lead to false positives. Moreover, even when genuine, the trajectory and potential impact of the anticipated change often remain uncertain at this early stage [29]. Acknowledging the uncertainty that surrounds weak signals is an important milestone for horizon scanning practitioners and commissioners. Ultimately, horizon scanning should not only detect the weak signals but also provide sufficient contextual information to help unpick the uncertainty that surrounds them. According to Garvey et al [1], “uncertainty is often less about the unknown as about the identification and interpretation of objectively known but hidden data and weak signals.”

There are several tools and techniques for supporting decision-making under uncertain conditions [30]. Among those, the Rumsfeld matrix has been used to characterize 4 different types of uncertainty (Figure 1).

According to this matrix, events are assessed according to their predictability or unpredictability and visibility or invisibility. The combination of those two-by-two options provides 4

XSL•FO
RenderX

different quadrants in which events may be classified and studied. The application of the Rumsfeld matrix to health care technology discovery may be a relatively unexplored area of research, but some studies are starting to use this tool to explore and categorize what is known and unknown to select pathways for investigation [31]. While the Rumsfeld matrix helps demystify uncertainty by classifying different types at play and proposes a framework to deal with each type, it fails to quantify the uncertainty and does not fully resolve the time variable. Timeliness impacts most HTAs and health care commissioning decision-making [32]; therefore, unpicking the time variable becomes crucial for effective horizon scanning methods.

Methods for Working With Time

The Three Horizons model illustrates how issues, opportunities, and weak signals emerge across 3 distinct but interacting timeframes. Developed in the late 1990s with a focus on strategic management, the model encourages decision-makers to engage simultaneously with short-, medium-, and long-term futures [18,33,34]. In Figure 2, this model is applied specifically to MedTech technologies.

Figure 2. Rumsfeld matrix, predictability, and visibility profiles, and horizon scanning.

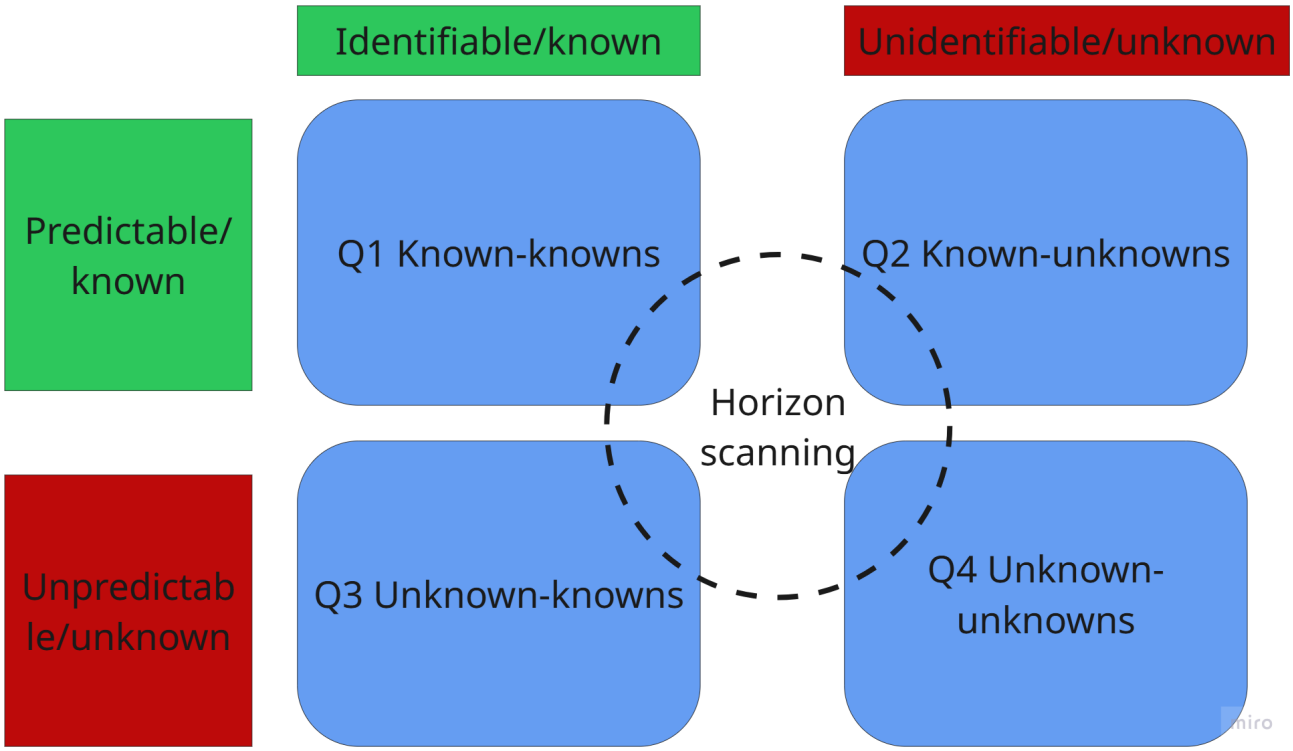

Following the rationale exposed by Curry and Hodgson [33], the horizons are interpreted and presented in the following order (Figure 3):

Horizon 1 (H1, now) or “imminent horizon” represents the current state or “business as usual” and is associated with the present or near-term. It acknowledges that ongoing changes in the external environment will eventually disrupt this status quo, for example, established MedTech, diminishing its relevance over time (obsolescence) [33,35].

XSL•FO
RenderX

Figure 3. The horizons of innovation theory applied to MedTech.

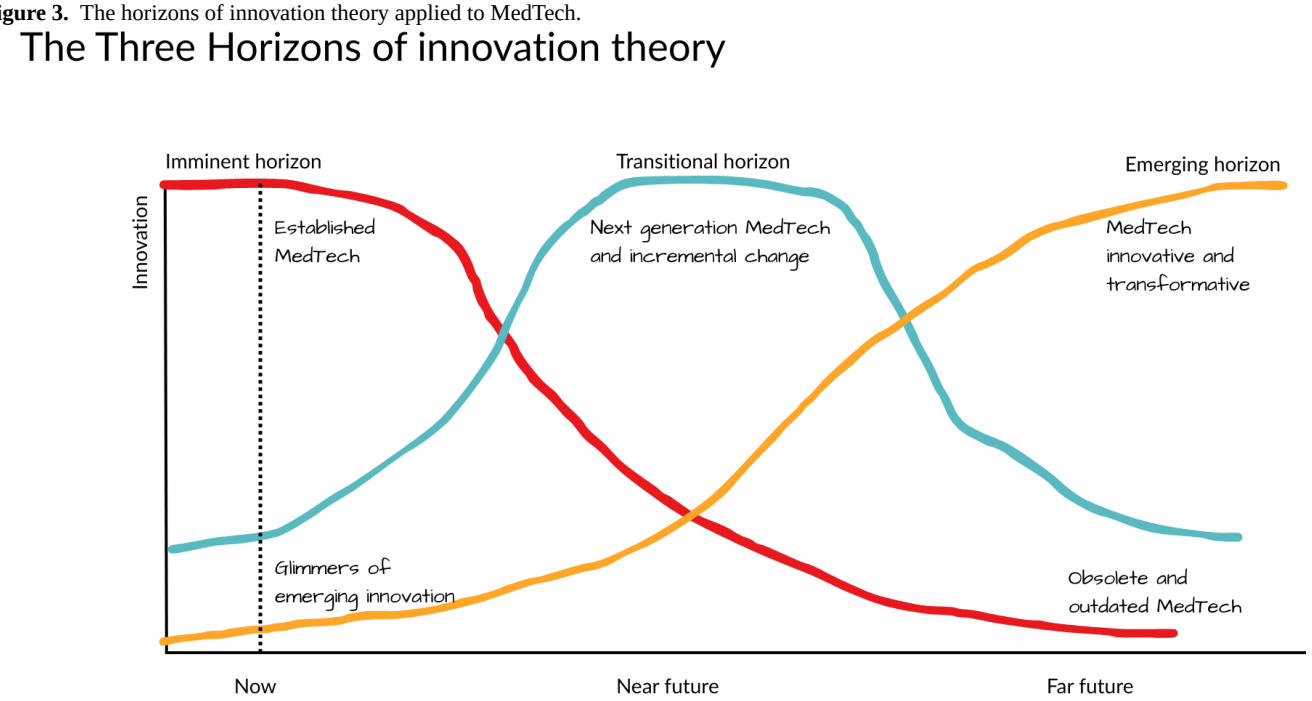

Horizon 3 (H3, far future) or "emerging horizon" reflects the long-term future, where new and uncertain developments begin to take shape; these may eventually replace or transform the current systems in H1. Although the specific nature of these changes is uncertain, monitoring them is essential, for they may reveal innovative and transformative technology that has the potential to radically change established health care pathways [33,36].

Horizon 2 (H2, near future) or "transitional horizon" serves as the transitional space between the present and the future, where emerging innovations and adaptive responses such as better iterations of already established health care products are tested in anticipation of broader shifts from H3 to H1 [33,37].

While this model provides a practical framework for working with time for innovation adoption and implementation, it does not solve the uncertainty that surrounds some of these emerging innovations. The key challenge is predicting how, when, and where new events or trends will occur and the level of disruption that these new and emerging innovations may generate. As these challenges involve many variables across multiple conditions or dimensions, additional surveillance or forecasting tools are needed to better understand the chance of an event occurring as well as the impact or level of disruption [38]. The Three Horizons model aims to alleviate uncertainties in the time dimension, but its capabilities are limited for other dimensions or technology disruptiveness levels.

The Pillars of Certainty: Regulatory Frameworks for MedTech

The Importance of Regulatory Frameworks

Regulatory frameworks are essential for ensuring the safety, efficacy, and quality of MedTech [39]. They provide a structured pathway for the development, approval, and postmarket surveillance of medical devices, in vitro diagnostics, and digital health applications [23,25]. These frameworks are particularly critical in MedTech due to the rapid pace of innovation, short product lifecycles, and the interdisciplinary nature of the field, which involves collaboration among clinicians, engineers, and regulatory experts [39].

The Role of Regulatory Frameworks in Reducing Uncertainty

Regulatory frameworks play a pivotal role in reducing uncertainty about a product's intended use and safety requirements by establishing clear guidelines and standards for MedTech innovation. They help ensure that emerging technologies meet predefined safety and performance criteria, thereby mitigating risks associated with new products [40]. By providing a predictable approval process, these frameworks enable developers to anticipate challenges and align their strategies with regulatory expectations, ultimately fostering greater confidence in the adoption of new technologies [39]. Regulatory frameworks also contribute to reducing uncertainty in horizon scanning by providing a structured approach to evaluating the safety and efficacy of emerging technologies. For example, the integration of regulatory milestones, such as clinical trial phases and market authorization, into horizon scanning processes allows stakeholders to anticipate the trajectory of new technologies and their potential impact on health care systems [10].

Regulatory Challenges in MedTech

One of the key challenges in MedTech regulation is the variability across regions. For instance, in the European Union, the Medical Devices Regulation (MDR) and In Vitro Diagnostic Medical Devices Regulation impose stringent requirements for CE (Conformité Européenne) marking, while in the United Kingdom, post-Brexit regulations require UKCA (UK conformity assessment) marking for market access [41]. In the United States, the US Food and Drug Administration's (FDA) Center for Devices and Radiological Health oversees device

XSL•FO
RenderX

regulation [23]. These differing requirements can create complexity for developers seeking global market entry.

How Regulatory Changes Might Impact Horizon Scanning

Regulatory changes, such as the introduction of the European Union MDR and In Vitro Diagnostic Medical Devices Regulation, have significant implications for horizon scanning. These regulations require more rigorous postmarket surveillance and clinical evidence, which can affect the timeline and visibility of emerging technologies (European Union MDR, 2017). Horizon scanning practitioners must adapt to these changes by incorporating regulatory compliance data into their analyses, ensuring that the identified weak signals are not only innovative but also viable within the regulatory landscape [10]. Conversely, horizon scanning may signal new and innovative technology aspects not covered by current regulation, making this method a strategic tool for regulators [42,43].

The Novelty of Integrating the Three Horizons Model With MedTech Regulatory Frameworks

The integration of the Three Horizons model with MedTech regulatory frameworks represents a novel approach to managing uncertainty in horizon scanning. By aligning the model's focus on short-, medium-, and long-term horizons with the staged regulatory pathways for MedTech development, this approach provides a comprehensive framework for tracking the progression of technologies from early innovation (horizon 3) to market readiness (horizon 1). This integration enhances the ability of horizon scanning practitioners to anticipate regulatory challenges and opportunities, thereby improving decision-making and preparedness for the adoption of disruptive technologies.

Figure 4 presents the 3 horizons, TRLs in relation to their corresponding lifecycle stages, and predicted time to market according to the FDA regulatory pathway.

Figure 4. MIST MedTech horizon scanning framework: TRLs, FDA regulatory milestones, and horizons. CE: Conformité Européenne; FDA: US Food and Drug Administration; HDE: humanitarian device exemption; HTA: health technology assessment; PMA: premarket approval; R&D: research and development; TRL: technology readiness level; UKCA: UK conformity assessment.

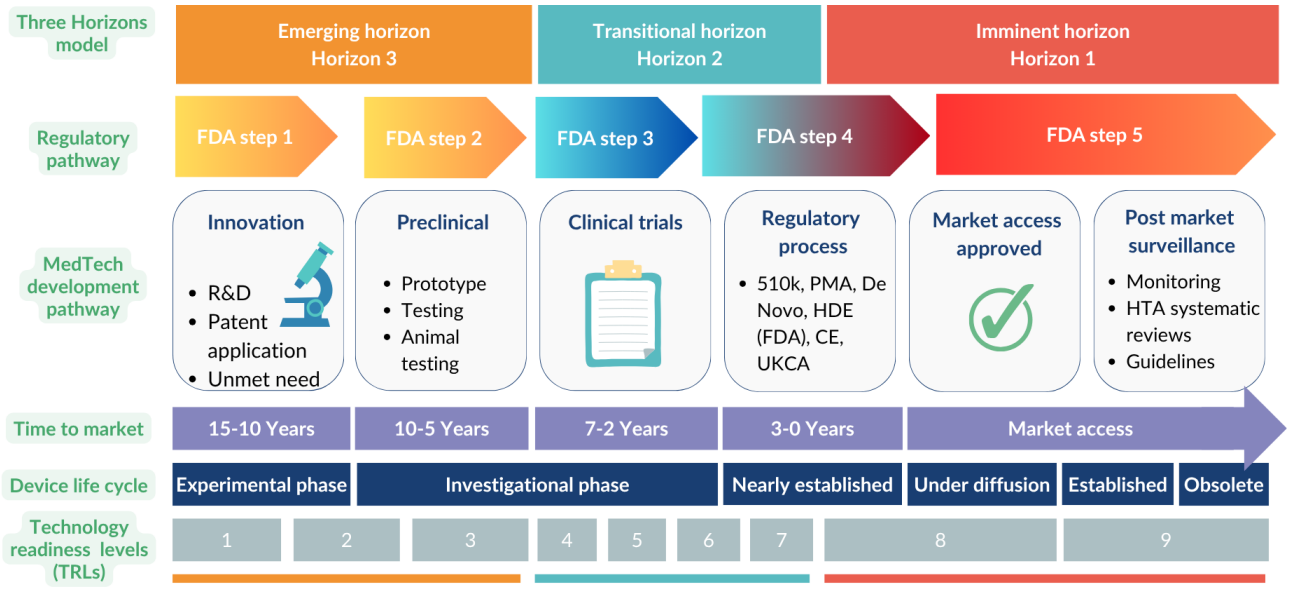

Horizon Scanning in the Imminent Horizon (Horizon 1)

Overview

In this context, we understand the imminent horizon as the time between preregulatory processes and postmarket surveillance. The scope of this horizon is for technologies at a late stage of development (TRLs 8 - 9) that are approaching their regulatory submission, undergoing regulatory certification and approval by a certified body, and the postmarket surveillance and monitoring activities for safety and effectiveness. The time in this horizon is short (less than 3 years to market access plus a variable amount of time for postmarket surveillance). This horizon overlaps with the HTA domain and systematic reviews of effectiveness due to evidence of the technologies' effect and cost becoming available as technologies undergo the fundamental milestones of approval and market launch.

The inclusion of this horizon within MIST is important because horizon scanning is concerned not only with distant or highly uncertain futures, but also with anticipating near-term changes that may disrupt or rapidly reshape current systems. In MedTech, technologies in horizon 1 may be close to adoption, yet important uncertainties remain regarding regulatory progression, implementation, uptake, system readiness, and the displacement of existing practices. This horizon is also critical for identifying abrupt discontinuities, as understanding their immediate effects on the health care system is essential to support a timely response, preparedness, and consideration of alternative options should a sudden disruption occur. Including this horizon ensures that MIST captures the full continuum of innovation emergence, from early weak signals to technologies approaching real-world impact.

In this horizon, the level of uncertainty about the future of a technology's emergence is minimal, but the uncertainty of the technology's effectiveness and cost-effectiveness is not resolved. This horizon corresponds to the known unknowns (Q3) and the

XSL•FO
RenderX

known knowns (Q1) quadrants in the Rumsfeld matrix (Figure 5) [1].

Figure 5. Rumsfeld matrix quadrants in the imminent horizon. HTA: health technology assessment.

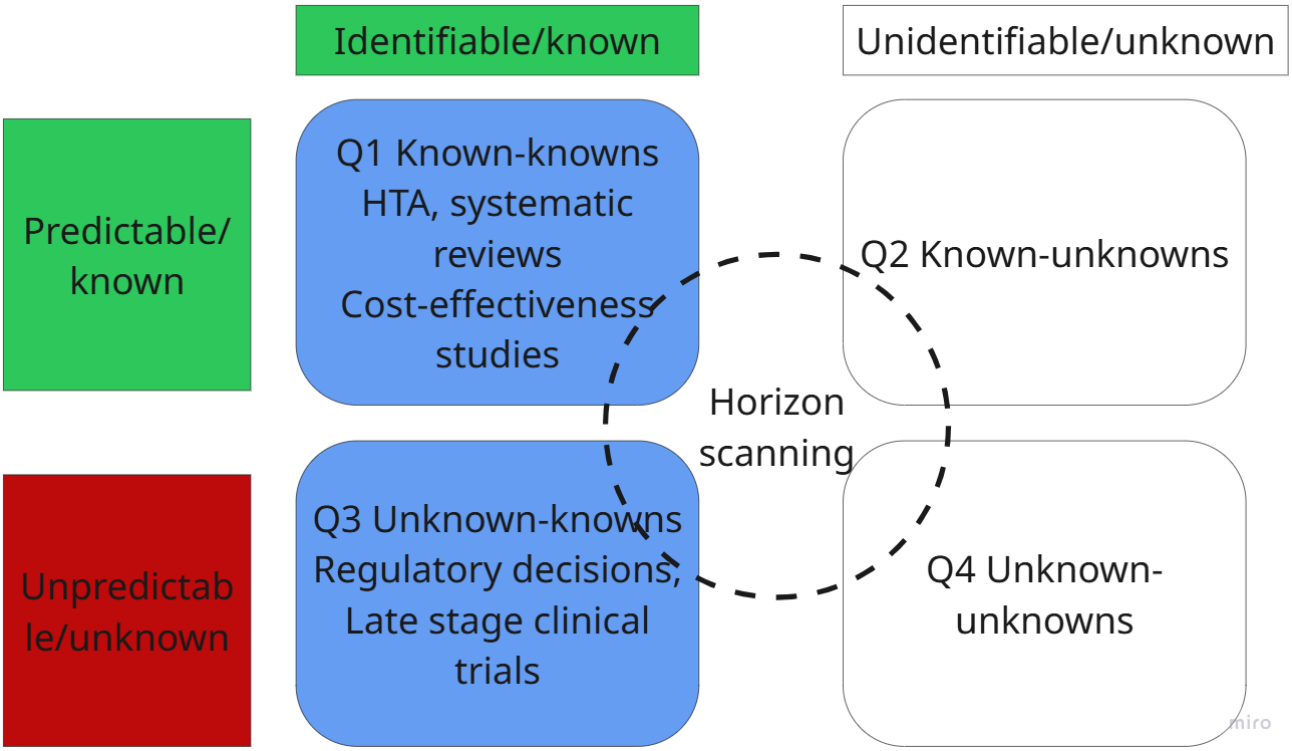

In this horizon, the level of predictability of an event, or in this context, of a new technology occurring or emerging into the market, is potentially at the highest point in the technology development pathway (Q1). However, in Q3, unforeseen or invisible circumstances can obstruct the technology's emergence, such as the lack of approval due to lack of compliance with regulations; a negative recommendation from a HTA body due to surpassing cost-effectiveness thresholds; or lack of technology adoption due to inequitable design aspects that interact with the effectiveness of the technology, for example, absence of skin color inclusivity for a technology whose mechanism of action is reading skin biomarkers [44].

The ability for horizon scanning practitioners to see these events ahead of time relies on the acquisition of in-depth detail about such technologies, as well as a deep understanding of the technology's regulatory, political, societal, environmental, and economic dimensions. Textbox 1 introduces prompts for dimensional factors that could help horizon scanning practitioners improve the visibility of unforeseen events in this horizon.

Textbox 1. Dimensional factors relevant for the imminent horizon.

- Regarding the regulatory dimension, these may include risk class considerations, different regulatory pathways, the existence of clinical trials, and the intended use of the technology.
- Regarding the political dimension, are there policies that prioritize these technologies? Are national or international funding bodies backing up the development of these technologies?
- Regarding the societal dimension, will this technology provide a better experience of health and well-being to the entire society? Are they inclusive and designed with patients in mind? Do they address known unmet needs?
- Regarding the environmental dimension, are they sustainable or green? Will there be further iterations of these technologies, and how will they be implemented sustainably?
- Regarding the economic dimension, are these new technologies more cost-effective than existing ones? Will they trigger economic implications for patients and users? Are these technologies displacing obsolete ones? If so, are there measures in place for disinvesting in obsolete technology while we sustain growth and innovation?

Best Fit Methods, Sources, and Tools for Horizon Scanning in the Imminent Horizon

Table 4 presents a summary of methods that can be used by the many, including the IO, to identify and contextualize weak signals in the imminent horizon. These methods may be used alone or in combination, depending on variables such as the time frame available to deliver the project, the resources and capacity at hand, and the scope of the project. Most horizon scanning projects in the imminent horizon will commence with a rapid search for evidence. Rapid evidence synthesis methods provide a quick overview of a technological field by searching the published literature [45]. These methods can help identify weak signals that can serve as "leads" for discovering new weak

XSL•FO
RenderX

signals, or they can provide contextual insights into the current state of research and innovation development within the relevant field. The findings of the review can guide subsequent decisions about where to search next and what to search for. Horizon scanning searches are characterized by sensitivity-maximizing searches across multiple sources. These searches aim to generate a vast amount of data, which is then screened and selected according to the scope of the horizon scan. This process may uncover new weak signals or "duplicate" ones that will strengthen the weak signals already identified via rapid evidence synthesis. While engagement with experts, patients, and the public might not always be required, these methods often yield rich qualitative data providing valuable contextual insights for assessing the dimensional factors in Textbox 1.

Table . Mapping of methods and sources for the imminent horizon.

Method used	Recommended sources
Rapid evidence synthesis	• Published literature
Horizon scanning searches	• Manufacturer sites • Phase III and phase IV ongoing clinical trials • Regulatory agencies • Market scan • Specialist MedTech publications
Sentiment analysis	• Social media
Surveys and questionnaires	• Patient and public engagement
Clinical and nonclinical topic expert workshops	• Expert consultation

Horizon Scanning in the Transitional Horizon (Horizon 2)

Overview

The transitional horizon refers to the period from the start of human clinical trials (pilot stage) to the initiation of preregulatory processes, typically during pivotal stage trials. This horizon focuses on technologies at mid-stage of development (TRLs 4‑7) that are commencing their investigational trials, generating evidence of their safety, efficacy, and outcomes, and compiling that evidence for certification. The duration of this horizon can vary significantly, heavily influenced by factors such as clinical investigation quality [46], market drivers [47], policy priorities [48], or investor support, among others. For low-risk devices, clinical trials might not be required, making this stage shorter than for high-risk devices [25]. Typically, this stage can last between 2 and 10 years, assuming a linear product development and secure investment. This transitional horizon partially overlaps with early HTAs [49], which are increasingly applied to medical devices to guide research and development, optimize design, and manage risks related to market access and reimbursement [50]. As innovation in MedTech often outpaces regulation, HTA and regulatory bodies require advanced notice to prepare for the assessment of these new, and often disruptive, technologies. This is particularly evident in the case of AI and software, which prompted new regulatory guidance that recognizes software as a medical device [51]. Evidence and weak signal generation during this horizon are incremental and visible, though events remain less predictable than in the imminent horizon.

Uncertainty about the future of a technology's emergence in this horizon is 2-fold: first, it depends on clinical trial success, market conditions, and secure funding. Second, certainty about the technology's effectiveness gradually increases as it progresses from pilot to pivotal clinical trials. This horizon involves the known unknowns (Q3) and the known knowns (Q1) from the Rumsfeld matrix [1] but with greater levels of unpredictability. Furthermore, Q2, the known unknowns creep into this horizon as technology "disrupters" or "game changers" whose implications may be overlooked or not fully understood due to biases in the decision-making system [1], for example, with pressures to address immediate issues such as a lack of diagnostic facilities [52,53]. Horizon scanning methods must contend with higher levels of unpredictability, limited visibility of events, and variable timelines for event development (Figure 6).

There are limitations for horizon scanning in the transitional horizon. The use of additional methods in combination with weak signal detection and prioritization, such as scenario analysis or forecasts, can help to deal with high levels of uncertainty by either defining alternative futures or anticipating the timing of an event [1]. Hines and Bishop [9] provide a stepwise approach to dealing with uncertainty in futures planning by engaging the stakeholders in thinking about the challenges and prioritizing the issues or opportunities that may arise, then designing different scenarios of each issue or opportunity, and finally ranking those. Textbox 2 provides the criteria for evaluating these issues or opportunities.

XSL•FO
RenderX

Figure 6. Rumsfeld matrix quadrants in the transitional horizon. AI: artificial intelligence; HTA: health technology assessment.

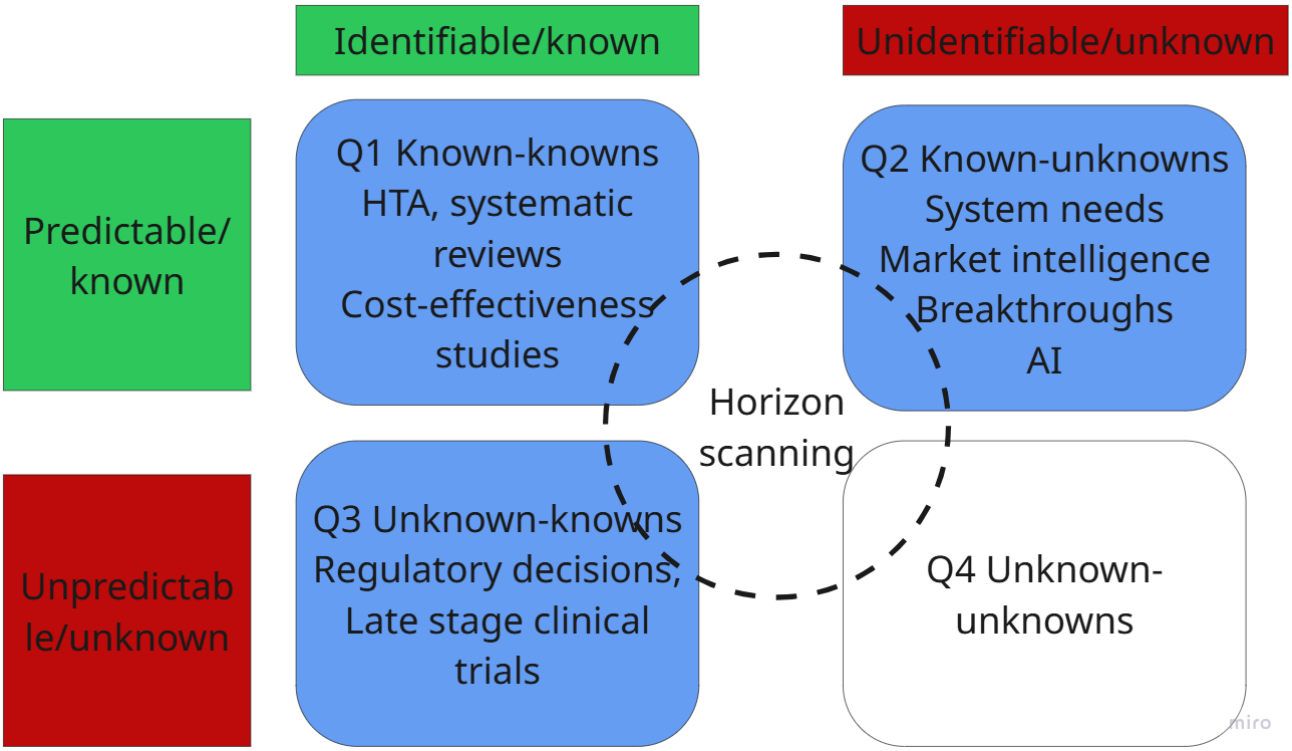

Textbox 2. Dealing with uncertainty in the transitional horizon: prioritizing issues or opportunities for futures planning.

- How likely are those issues?
- How much of an impact would the issues have on the future?
- How unprepared are we (the stakeholders) for that future?

In MedTech, key questions include: How likely are these technologies to emerge? How disruptive are these technologies? How prepared is the health care service for the rapid adoption of them? Answers to these questions support preparedness and strategy planning, helping to mitigate some of the uncertainties in the transitional horizon, although many of these uncertainties will resolve or diminish over time.

Dimensional factors also play a role in shaping future technology. The technological dimension is critical, as emerging technologies can accelerate or halt the progress of existing technologies. The economic dimension can also drive or impede development depending on investment flows, while disinvestment or prioritization of resources can be slow to progress. Finally, the political dimension, especially political maneuvers that support or undermine research investment, will directly impact this horizon, particularly for public research [54].

Best Fit Methods, Sources, and Tools for Horizon Scanning in the Transitional Horizon

As for the imminent horizon, some of the methods and sources proposed in Table 5 for scanning in the transitional horizon can be used in combination. In this horizon, it becomes evident the widespread use of sources to combat the greater uncertainty posed by the TRLs of the technologies, the lower level of visibility of these weak signals, and the decreased predictability. This mitigation strategy results in an increased number of information sources that need to be searched and results that need to be sifted through to identify weak signals.

XSL•FO
RenderX

Table . Mapping of methods and sources for the transitional horizon.

Method used	Recommended sources
Rapid evidence synthesis	• Published literature
Bibliometric studies	• Published literature
Horizon scanning searches	• Ongoing clinical trials • Market scan • Specialist MedTech publications • News scanning • Funding sites • Regulatory agencies
Surveys and questionnaires	• Patient and public engagement
Clinical and nonclinical topic expert workshops	• Expert consultation
Gray literature searching	• Conference proceedings • Policies (national or international)
Sentiment analysis	• Social media

Consequently, when horizon scanning for technologies in the transitional horizon, we must be aware of the impact of "noise," understood as "the retrieval of irrelevant signals," and the risk of "silence" understood as "the failure in retrieving relevant text from a database." One way to deal with noise is by involving stakeholders and experts early in the process and maintaining that engagement throughout the project. Consulting with experts could alleviate some of the issues generated by noise by the prioritization of weak signals based on their maturity and strength. The issue of silence requires more attention as it could mean different things. On the one hand, it could mean that there is a research gap, funding gap, or innovation gap, but none of these gaps could be fully ascertained without scanning further into the horizon, this is, by scanning the emerging horizon (H3) and contextualizing the findings in the social, political, economic, and technological dimensions (Textbox 1). On the other hand, silence could mean that the search focus is not correct or accurate enough to retrieve the relevant weak signals. This is often an issue faced when practicing horizon scanning for new innovations where the language is not settled or standardized. Bibliometric studies and visualization of similarities techniques can deal with those issues [55].

Horizon Scanning in the Emerging Horizon (Horizon 3)

Overview

In the MIST framework, the emerging horizon is understood as the time when an invention occurs in the shape of a hypothesis that needs testing or is at early experimentation, a patent may be registered and granted, and initial funding is sought for further development of this idea into a prototype. The scope of this horizon is for technologies at an early stage of development (TRLs 1 - 3). The time horizon here is very broad and may be highly dependent on factors such as the characteristics of the innovation, the current need for it, previous and accumulative technological advances that may facilitate the realization of this idea in the current time or interest from high-risk investors and venture capitalists [19,56].

This horizon offers a valuable opportunity to explore trends and innovations beyond medical applications, extending across various disciplines. Analyzing these trends can help anticipate the broader impact of emerging technologies. For instance, how quantum computing might affect multiple sectors, including but not limited to health care, or provide context to the gaps identified in the transitional horizon. While this horizon provides abundant data sources, such as patents, preclinical research, and conference abstracts, the complexity of the information and the uncertainty surrounding the weak signals at this early stage demands complementary analytical methods to identify overarching patterns and interconnections across datasets. While weak signal generation in this horizon is rich (registration of intellectual property in patents, presentation of inventions at conferences and grant applications) and visible, events are less predictable; therefore, the level of uncertainty about the future of a technology's emergence in this horizon is at its greatest point.

In terms of Rumsfeld quadrants (Figure 7), despite the high uncertainty or unpredictability of events unfolding in the emerging horizon or distant future, weak signals are still known (eg, patents, grants, and scientific advances announced at conferences, breaking news about future technology, trends analysis reports, etc). In this horizon, we are mostly operating in the Q2 known-unknowns and Q3 unknown-knowns. We might know that a technological innovation is emerging, but the distant time horizon and unpredictability of whether that technology might progress further and materialize into innovative health care solutions is too high to ascertain. In Q2, we also face surprises in the form of technologies that emerge rather rapidly and almost unexpectedly. This is the case with novel digital health technologies that do not need long research and development pathways.

XSL•FO
RenderX

Figure 7. Rumsfeld matrix quadrants in the emerging horizon. AI: artificial intelligence.

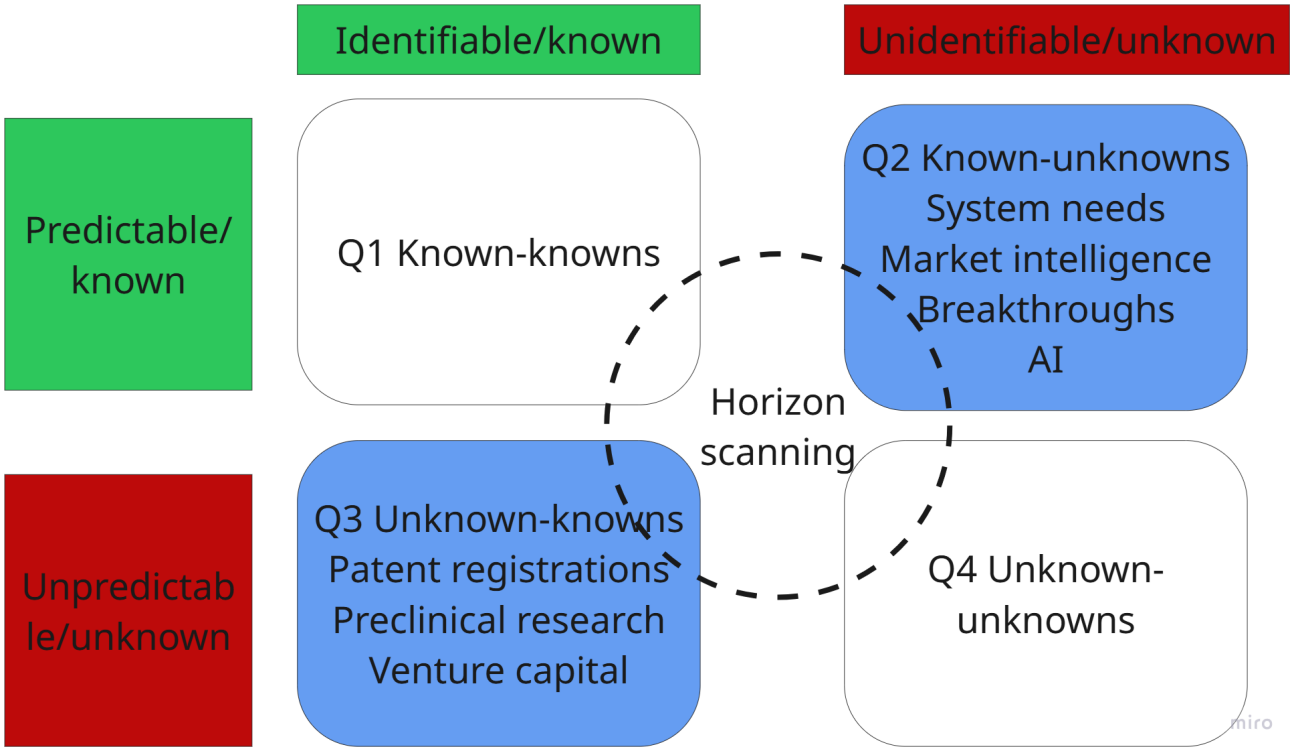

Best Fit Methods, Sources, and Tools for Horizon Scanning in the Emerging Horizon

In the emerging horizon, the volume of available data is typically greater than at later stages, rendering weak signal detection a process of distinguishing meaningful ones from background noise. An overview of the corresponding approaches, sources, and methods is presented in Table 6. Among these, visualization of term similarities represents a valuable means of mapping and interrogating research landscapes. We used this approach for the visualization of co-occurring terms extracted from Embase (Ovid) records of preclinical studies on fluorine-18 and positron emission tomography [53]. This analysis sought to characterize the early innovation landscape of fluorine-18 applications to provide anticipatory intelligence for decision-makers. Within the network representation (Figure 8), the size of each node reflects term frequency, with prominent terms such as synthesis, radiosynthesis, tumor uptake, prostate-specific membrane antigen, lesion, and response denoting areas of concentrated research activity. By contrast, smaller, more peripheral nodes, such as neuroinflammation linked to status epilepticus, cortex, and brain region, point to emerging or niche areas of investigation, including the application of fluorine-18 positron emission tomography for monitoring neuroinflammation.

XSL•FO
RenderX

Figure 8. Visualization of similarities between key words in positron emission tomography preclinical research. MRI: magnetic resonance imaging.

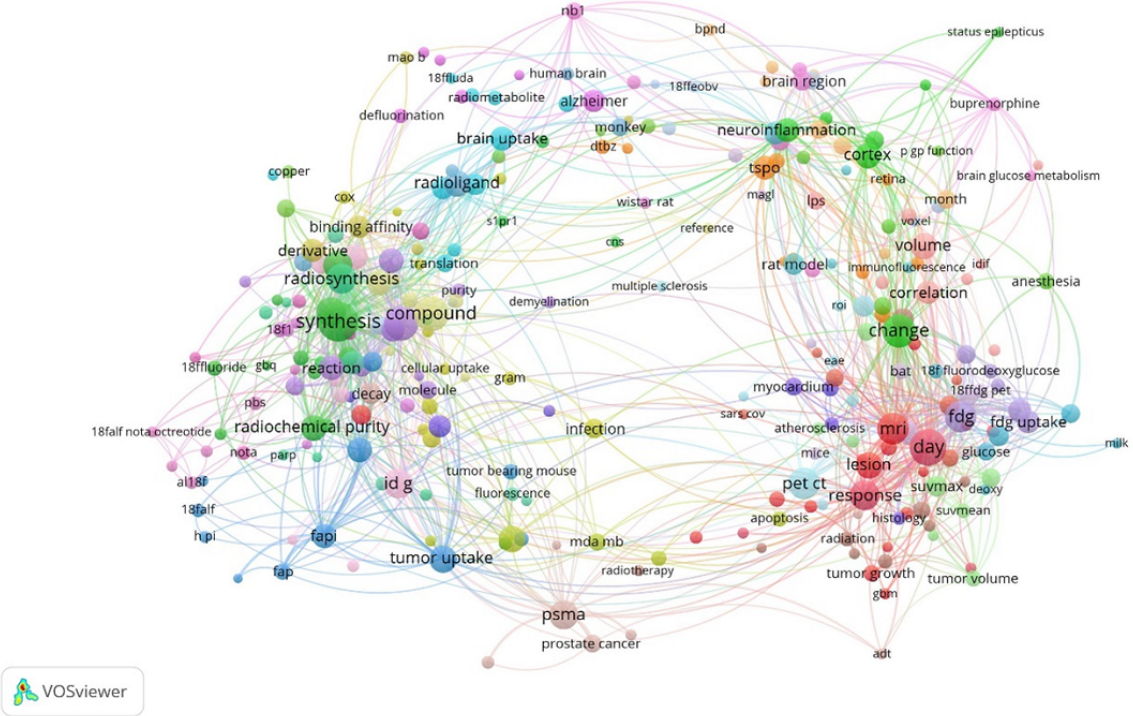

Table . Mapping of methods and sources for the emerging horizon.

Method used	Recommended sources
Horizon scanning searches	• Market scan • Specialist MedTech publications • News scanning • Funding sites
Sentiment analysis	• Social media
Horizon scanning searches; text mining; cluster visualization techniques	• Patent databases (European and worldwide)
Bibliometric and scientometric methods: visualization of term similarities, cocitation analysis, cluster analysis	• Bibliographic databases
Workshop or survey	• Experts in the research field

Overall, similarity-based visualization constituted an effective method for uncovering research clusters and nascent applications within the radionuclide field. Although this technique cannot be expected to retrieve all possible data, it can highlight early weak signals and generate hypotheses for more detailed exploration. Crucially, domain expertise remains indispensable for interpreting network structures and associations, thereby ensuring the accurate identification of weak signals and their relevance within the broader research context [53].

Patent data is also rich in this horizon and presents a great body of data for trend analysis due to patent data availability across jurisdictions, making patent analysis studies a suitable method for identifying which patents are on the rise in different regions and study trends [57].

Discussion

Principal Findings

Human existence is defined by the quest for knowledge and the humble acknowledgment that there is much that we do not know. While horizon scanning is a method designed to detect weak signals, the “unknown unknowns” will not produce signals; should there be any, then these automatically by definition will move into the realm of the “knowns” following the premise “if we can think of it, then it exists and then is not unknown” (Garvey et al, 2022) [1]. Horizon scanning does not belong to the exploration of the “unknown unknowns,” and probably nothing belongs to this quadrant (Q4) of the Rumsfeld matrix other than postevent justifications for why we did not know or realize that something would happen [1]. This is essentially the

XSL•FO
RenderX

field of hindsight, and one not included in the scope of this methodological framework.

Within this context, the Rumsfeld matrix and the Three Horizons model are used in MIST as established foresight tools rather than as new methodological contributions in their own right. The novelty of MIST lies not in proposing these models for the first time, but in combining and operationalizing them within a MedTech-specific horizon scanning framework. In particular, MIST links these broader conceptual tools to TRLs, regulatory pathways, source selection, and stakeholder decision needs, thereby translating general foresight theory into a structured and practical methodology for MedTech horizon scanning.

The framework responds to a very specific need for horizon scanning methods in a very specific field. The high specificity of the method is geared toward guiding the practice of horizon scanning to solve uncertainty in this field and help decision-making needs in MedTech, as opposed to more generalist frameworks [11,12,17,58] that are designed to broadly guide foresight practice, of which horizon scanning may be a component. Furthermore, the development of this framework is underpinned by extensive expert engagement and methodological practice. Input was gathered from a diverse group of specialists spanning horizon scanning, HTA, regulatory science, futures studies, and MedTech innovation, ensuring that the framework reflects a broad range of practical and conceptual perspectives. The multidisciplinary expertise of the project's team further strengthens the robustness of the framework, particularly given the IO's long-standing experience in horizon scanning, information retrieval, and evidence synthesis for health and care technologies.

Horizon scanning in the imminent horizon is also valuable in its own right because it supports early awareness and preparedness for HTA agencies and related decision-makers. Even when technologies are close to market entry, structured scanning remains important to identify those likely to require assessment, anticipate evidential and implementation challenges, and support timely planning for appraisal, adoption, or service response. In this sense, the imminent horizon is not outside the scope of horizon scanning but represents a critical stage at which anticipatory intelligence directly informs HTA and policy processes. This near-term anticipatory function is particularly relevant across jurisdictions, even though the specific regulatory and decision-making structures may differ. While the methodological core of MIST is jurisdiction-agnostic (universal functions such as early identification, filtering, characterization, prioritization, and iterative monitoring), some operational anchors may be configurable to other jurisdictions, equivalent regulatory and policy anchors (eg, FDA, Pharmaceuticals and Medical Devices Agency, National Medical Products Administration, and Canadian Agency for Drugs and Technologies in Health) without altering the framework's underlying logic. MIST should therefore be understood as modular: methodologically universal but locally instantiated through jurisdiction-specific regulatory and system contexts.

Our work has some limitations. Our framework has not been formally validated; therefore, we foresee the need to revise some of the methods and sources proposed here in the near future. Many components are interdependent in this framework; one change may affect others, and we recommend considering MIST as the first version of a living framework that would need updating regularly to incorporate the latest methodological or regulatory advances. While equity and ethical considerations are critical in the adoption and assessment of MedTech innovations, these dimensions are not explicitly incorporated in the current iteration of the MIST framework. The development of MIST has focused on establishing a robust, transparent, and systematic methodological foundation for horizon scanning in MedTech, addressing challenges such as rapid product lifecycles, weak signal detection, regulatory variability, and stakeholder integration. Parallel work with key stakeholders is ongoing to develop methodologies for embedding equity and ethical principles as a cross-cutting layer across horizons into MedTech decision-making processes [59,60]. The insights gained from these initiatives will inform the next iteration of MIST, ensuring that future versions of the framework comprehensively integrate equity, fairness, and ethical considerations alongside methodological rigor.

Conclusion

Horizon scanning for MedTech innovation faces multiple methodological, practical, and conceptual challenges. MedTech products are highly heterogeneous, encompassing medical devices, in vitro diagnostics, and digital health solutions, and often undergo rapid, iterative development with short lifecycles. Continuous technological advances in AI, sensors, materials, and software can quickly render previous versions obsolete, while fragmented data sources, weak and uncertain signals, and diverse stakeholder perspectives complicate robust identification and prioritization of emerging technologies [16]. Regulatory complexity, including differing regional requirements and evolving standards, adds further uncertainty, and inconsistent terminology coupled with the absence of standardized reporting frameworks limits transparency, comparability, and reproducibility [8].

Although several horizon scanning frameworks exist, no framework has been published to guide the practice of horizon scanning in the context of emerging and innovative health care technologies and, more specifically, in MedTech. To address these challenges, we have developed MIST, a systematic framework that provides structured, unbiased, and transparent guidance for horizon scanning in MedTech innovation. MIST is designed to support national policy and decision-makers in making timely, evidence-informed decisions regarding the adoption of new technologies, accelerate implementation into clinical practice, and identify unmet research and investment priorities, while addressing the specific methodological and operational complexities of MedTech horizon scanning. At this stage, MIST should be understood as a conceptually derived and practice-informed framework; quantitative evaluation of its components and application across case studies will form part of future validation work.

XSL•FO
RenderX

Acknowledgments

The authors gratefully acknowledge the valuable contributions, insights, and knowledge shared by members of the following working groups: the International Horizon Scanning Initiative, Medical Devices Working Group; the Medicines and Healthcare Products Regulatory Agency's Medical Devices Horizon Scanning Working Group; the NHS (National Health Service) England Horizon Scanning Team; the Horizon Scanning and Health Technology Assessment team at Health Improvement Scotland and their Health Technology Wales counterparts. Their engagement in discussions and their expertise have been instrumental in informing the development of the framework presented in this paper. Generative artificial intelligence tools were used to support language refinement and editing during manuscript preparation, including assistance with improving clarity and readability. All conceptual development, methodological design, interpretation, source identification, and final content decisions were undertaken by the authors. The authors take full responsibility for the accuracy, originality, and integrity of this paper's content.

Funding

This study or project is funded by the National Institute for Health and Care Research.

Conflicts of Interest

DC is the director at NIHR Innovation Observatory.

References

1. Garvey B, Humzah D, Le Roux S. Uncertainty Deconstructed: A Guidebook for Decision Support Practitioners: Springer; 2022. [doi: 10.1007/978-3-031-08007-4]
2. Simpson S, editor. A toolkit for the identification and assessment of new and emerging health technologies. : EuroScan International Network; 2014 URL: https://epapers.bham.ac.uk/ /id/eprint/2120/1/EuroScan_Methods_Toolkit_October_2014_FINAL_CC_added.pdf [accessed 2026-05-15]
3. Eulaerts O, Joanny G, Giraldi J, et al. Weak signals in science and technology. : Publications Office of the European Union; 2020. [doi: 10.2760/625664]
4. Wild C, Langer T. Emerging health technologies: informing and supporting health policy early. Health Policy 2008 Aug;87(2):160-171. [doi: 10.1016/j.healthpol.2008.01.002] [Medline: 18295925]
5. MedTech europe's facts & figures 2024. MedTech Europe. 2024 Nov 7. URL: https://www.medtecheurope.org/ resource-library/medtech-europes-facts-figures-2024/ [accessed 2026-05-14]
6. Biodesign. URL: https://ebiodesign.org/ [accessed 2026-05-14]
7. What we do. NIHR Innovation Observatory. URL: https://io.nihr.ac.uk/what-we-do/ [accessed 2026-05-14]
8. Garcia Gonzalez-Moral S, Beyer FR, Oyewole AO, Richmond C, Wainwright L, Craig D. Looking at the fringes of MedTech innovation: a mapping review of horizon scanning and foresight methods. BMJ Open 2023 Sep;13(9):e073730. [doi: 10.1136/bmjopen-2023-073730]
9. Hines A, Bishop PC. Framework foresight: exploring futures the Houston way. Futures 2013 Jul;51:31-49. [doi: 10.1016/j.futures.2013.05.002]
10. Michels RE, de Graaff MB, Abrishami P, Delnoij DMJ. Anticipating emerging medical technologies: the start of an international horizon scanning tool for medical devices. Futures 2024 Feb;156:103326. [doi: 10.1016/j.futures.2024.103326]
11. Hines A. Evolution of framework foresight. FS 2020 Jun 26;22(5/6):643-651. [doi: 10.1108/FS-03-2020-0018]
12. Phil H. Thinking about the future: guidelines for strategic foresight. Foresight (Colch) 2008 Feb 22;10(1):79-80. [doi: 10.1108/14636680810856044]
13. European Commission Directorate-General for Research and Innovation. Concurrent design foresight: report to the European Commission of the Expert Group on Foresight Modelling. : Publications Office of the European Union; 2015 URL: https:/ /espas.secure.europarl.europa.eu/orbis/system/files/generated/document/en/concurrent_design_foresight_report.pdf [accessed 2026-05-15]
14. New approach methodologies: EU-IN horizon scanning report. : European Medicines Agency (EMA); 2025 Feb URL: https://www.ema.europa.eu/en/documents/report/new-approach-methodologies-eu-horizon-scanning-report_en.pdf [accessed 2025-05-15]
15. McMeekin N, Wu O, Germeni E, Briggs A. How methodological frameworks are being developed: evidence from a scoping review. BMC Med Res Methodol 2020 Jun 30;20(1):173. [doi: 10.1186/s12874-020-01061-4] [Medline: 32605535]
16. Addis P, Gonzalez-Moral SG. Landscape overview of UK healthcare decision-making structures. : National Institute for Health and Care Research Innovation Observatory; 2023 Dec URL: https://www.io.nihr.ac.uk/wp-content/uploads/2023/ 12/Stakeholder-Project-Report_.pdf [accessed 2026-05-15]
17. Hines P, Hiu Yu L, Guy RH, Brand A, Papaluca-Amati M. Scanning the horizon: a systematic literature review of methodologies. BMJ Open 2019 May 27;9(5):e026764. [doi: 10.1136/bmjopen-2018-026764] [Medline: 31133588]
18. Enduring ideas: the three horizons of growth. McKinsey & Company. 2009 Dec 1. URL: https://www.mckinsey.com/ capabilities/strategy-and-corporate-finance/our-insights/enduring-ideas-the-three-horizons-of-growth [accessed 2026-05-15]

XSL•FO
RenderX

19. Gonzalez-Moral SG, Pennock E, Ewedairo O, Green E, Elgey J, Mkwashi A. Horizon scanning methods for health care technology innovation identification: rapid scoping review of patent research studies. Interact J Med Res 2025 Sep 11;14:e70323. [doi: 10.2196/70323] [Medline: 40934489]
20. Check if you're eligible for funding. UK Research and Innovation (UKRI). URL: https://www.ukri.org/councils/stfc/guidance-for-applicants/check-if-youre-eligible-for-funding/eligibility-of-technology-readiness-levels-trl/ [accessed 2026-05-15]
21. Guidance on technology readiness levels. GOV.UK. 2014 Dec 23. URL: https://www.gov.uk/government/news/guidance-on-technology-readiness-levels [accessed 2026-05-15]
22. How we enforce medical device regulations. GOV.UK. 2024 May 17. URL: https://www.gov.uk/government/publications/report-a-non-compliant-medical-device-enforcement-process/how-mhra-ensures-the-safety-and-quality-of-medical-devices [accessed 2026-05-15]
23. Overview of device regulation. US Food & Drug Administration. 2026 Feb 2. URL: https://www.fda.gov/medical-devices/device-advice-comprehensive-regulatory-assistance/overview-device-regulation [accessed 2026-05-15]
24. Consolidated text: Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU (Text with EEA relevance). : European Union; 2017 Apr 5 URL: http://data.europa.eu/eli/reg/2017/746/2025-01-10 [accessed 2026-05-15]
25. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance.). : European Union; 2017 Apr 5 URL: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng [accessed 2026-05-15]
26. Rumsfeld matrix. The Uncertainty Project. URL: https://www.theuncertaintyproject.org/tools/rumsfeld-matrix [accessed 2026-05-15]
27. Newcastle University research ethics policy. : Newcastle University; 2024 Jan 4 URL: https://www.ncl.ac.uk/mediav8/our-research/research-governance-policies/Research%20Ethics%20Policy.pdf [accessed 2026-05-19]
28. The concordat to support research integrity. : Universities UK (UUK); 2026 Jan 16 URL: https://www.universitiesuk.ac.uk/what-we-do/policy-and-research/publications/concordat-support-research-integrity [accessed 2026-05-15]
29. Miles I, Saritas O, Sokolov A. Foresight for Science, Technology and Innovation: Springer; 2016. [doi: 10.1007/978-3-319-32574-3]
30. Tools & techniques. The Uncertainty Project. URL: https://www.theuncertaintyproject.org/blueprint/toolbox [accessed 2026-05-15]
31. Bailly C. Application of the Rumsfeld matrix to anticancer natural product target discovery. Pharmacol Res 2024 Mar;2:100023. [doi: 10.1016/j.prerep.2024.100023]
32. Fontrier AM, Kamphuis B, Kanavos P. How can health technology assessment be improved to optimise access to medicines? Results from a Delphi study in Europe: better access to medicines through HTA. Eur J Health Econ 2024 Aug;25(6):935-950. [doi: 10.1007/s10198-023-01637-z] [Medline: 37917290]
33. Curry A, Hodgson A. Seeing in multiple horizons: connecting futures to strategy. J Futures Stud 2008;13(1):1-20 [FREE Full text]
34. Sharpe B, Hodgson A, Leicester G, Lyon A, Fazey I. Three horizons: a pathways practice for transformation. E&S 2016;21(2):47. [doi: 10.5751/ES-08388-210247]
35. Imminent horizon. NIHR Innovation Observatory. URL: https://io.nihr.ac.uk/what-we-do/imminent-horizon/ [accessed 2026-05-15]
36. Emerging horizon. NIHR Innovation Observatory. URL: https://io.nihr.ac.uk/what-we-do/emerging-horizon/ [accessed 2026-05-15]
37. Transitional horizon. NIHR Innovation Observatory. URL: https://io.nihr.ac.uk/what-we-do/transitional-horizon/ [accessed 2026-05-15]
38. Guo J, Pan J, Guo J, Gu F, Kuusisto J. Measurement framework for assessing disruptive innovations. Technol Forecast Soc Change 2019 Feb;139:250-265. [doi: 10.1016/j.techfore.2018.10.015]
39. Mkwashi AS. Medical device regulations, industrial capabilities, and affordable healthcare technology development: case studies from the United Kingdom and South Africa [Thesis]. : Open University; 2020 URL: https://oro.open.ac.uk/70516/1/Final_Thesis_%20Andrew%20S%20MKWASHI%20April%202020.pdf [accessed 2026-05-15]
40. Brass I, Mkwashi A. Risk assessment and classification of medical device software for the internet of medical things: challenges arising from connected, intelligent medical device. In: IoT '22: Proceedings of the 12th International Conference on the Internet of Things: Association for Computing Machinery; 2022:171-178. [doi: 10.1145/3567445.3571104]
41. Guidance: regulating medical devices in the UK. GOV.UK. 2020 Dec 31. URL: https://www.gov.uk/guidance/regulating-medical-devices-in-the-uk [accessed 2026-05-15]
42. Slawther E. How regulatory horizon scanning works in the EU, UK & switzerland. Pink Sheet. 2025 Nov 10. URL: https://insights.citeline.com/pink-sheet/regional-comparisons/how-regulatory-horizon-scanning-works-in-the-eu-uk-switzerland-NL3KU7D2UJGPRMAP5SXT6N4GLY/ [accessed 2026-05-19]

XSL•FO
RenderX

43. Horizon scanning case study: what is an actionable horizon scanning signal (AHSS)? GOV.UK. 2023 Jan 17. URL: https://www.gov.uk/government/case-studies/horizon-scanning-case-study-what-is-an-actionable-horizon-scanning-signal-ahss [accessed 2026-05-15]
44. Setchfield K, Gorman A, Simpson AHRW, Somekh MG, Wright AJ. Effect of skin color on optical properties and the implications for medical optical technologies: a review. J Biomed Opt 2024 Jan;29(1):010901. [doi: 10.1117/1.JBO.29.1.010901] [Medline: 38269083]
45. Campbell F, Sutton A, Pollock D, et al. Rapid reviews methods series (paper 7): guidance on rapid scoping, mapping and evidence and gap map ('Big Picture Reviews'). BMJ Evidence-Based Med 2025;30:268-277. [doi: 10.1136/bmjebm-2023-112389]
46. Keane DF, Mathew C, Longley R, et al. Worldwide trends in the quality and breadth of clinical investigations of medical devices over the past decade: a scoping review and evidence map. Trials 2025;26(1):113. [doi: 10.1186/s13063-025-08793-y]
47. Awasthi P, Stanick J. White paper: the rise of global medical technology. An overview of the market and trends. : IQVIA MedTech; 2021 URL: https://www.healthtechireland.ie/wp-content/uploads/2021/01/The-rise-of-global-medical-technology-White-Paper-1.pdf [accessed 2026-05-15]
48. Policy paper: medical technology strategy. GOV.UK. 2024 Apr 9. URL: https://www.gov.uk/government/publications/medical-technology-strategy/medical-technology-strategy [accessed 2026-05-15]
49. Rodriguez Llorian E, Waliji LA, Dragojlovic N, Michaux KD, Nagase F, Lynd LD. Frameworks for health technology assessment at an early stage of product development: a review and roadmap to guide applications. Value Health 2023 Aug;26(8):1258-1269. [doi: 10.1016/j.jval.2023.03.009] [Medline: 36990207]
50. Castaldo R, Federici C, Pecchia L. Chapter 112 - early stage healthcare technology assessment. In: Iadanza E, editor. Clinical Engineering Handbook, 2nd edition: Academic Press; 2020:799-806. [doi: 10.1016/B978-0-12-813467-2.00113-9]
51. Software as a Medical Device (SaMD). U.S. Food & Drug Administration. URL: https://www.fda.gov/medical-devices/digital-health-center-excellence/software-medical-device-samd [accessed 2026-05-15]
52. Richards M. Diagnostics: recovery and renewal. : NHS England; 2020 Oct URL: https://www.england.nhs.uk/wp-content/uploads/2020/11/diagnostics-recovery-and-renewal-independent-review-of-diagnostic-services-for-nhs-england-2.pdf [accessed 2026-05-15]
53. Horizon scanning report: identification of innovations for PET radiopharmaceuticals in the context of the Welsh Health Service. : NIHR Innovation Observatory; 2023 Dec 22 URL: https://io.nihr.ac.uk/resources/reports/horizon-scanning-report-identification-of-innovations-for-pet-radiopharmaceuticals-in-the-context-of-the-welsh-health-service/ [accessed 2026-05-15]
54. Gaind N. How the NIH dominates the world's health research — in charts. Nat New Biol 2025 Mar 20;639(8055):554-555. [doi: 10.1038/d41586-025-00754-4]
55. Eastaugh C, Addis P, Fairweather M, Potter R, Mkwashi A, Gonzalez-Moral SG. Bioengineering game changing technology: a bibliometric keyword analysis and visualisation of similarities techniques to support horizon scanning of innovative bioengineering technologies. : Figshare; 2026. [doi: 10.6084/m9.figshare.30593969.v1]
56. Fleming JJ. The decline of venture capital investment in early-stage life sciences poses a challenge to continued innovation. Health Aff (Millwood) 2015 Feb;34(2):271-276. [doi: 10.1377/hlthaff.2014.1051] [Medline: 25646107]
57. Karataş AR, Kazak H, Akcan AT, Akkaş E, Arık M. A bibliometric mapping analysis of the literature on patent analysis. World Pat Inf 2024 Jun;77:102266. [doi: 10.1016/j.wpi.2024.102266]
58. Amanatidou E, Butter M, Carabias V, et al. On concepts and methods in horizon scanning: lessons from initiating policy dialogues on emerging issues. Sci Public Policy 2012 Mar 1;39(2):208-221. [doi: 10.1093/scipol/scs017]
59. Stone J, McCully S. Developing an ethical lens for horizon scanning and foresight. Nuffield Council on Bioethics. 2024 Feb 22. URL: https://www.nuffieldbioethics.org/news-blog/developing-an-ethical-lens-for-horizon-scanning-and-foresight [accessed 2026-05-15]
60. Kunonga TP, Eastaugh C, Gonzalez-Moral SG, et al. Exploring the approaches for including ethical dimensions in horizon scanning and foresight methods in biomedical and health policy decision making: a rapid scoping review protocol. OSF. 2024 Mar 11. URL: https://osf.io/t2nhc/overview [accessed 2026-05-15]

Abbreviations

AI: artificial intelligence
CE: Conformité Européenne
FDA: US Food and Drug Administration
HTA: health technology assessment
IO: Innovation Observatory
MDR: Medical Devices Regulation
MIST: MedTech Innovation Scanning Techniques
TRL : technology readiness level
UKCA: UK conformity assessment

XSL•FO
RenderX

Edited by B Mesko, T Kristóf; submitted 09.Feb.2026; peer-reviewed by É Hideg; revised version received 20.Apr.2026; accepted 22.Apr.2026; published 21.May.2026.

Please cite as:
Garcia Gonzalez-Moral S, Mkwashi A, Beyer FR, Craig D
Horizon Scanning for Medical Technologies: Methodological Framework Development Study of the Medical Innovation Scanning Techniques (MIST) Framework
J Med Internet Res 2026;28:e93166
URL: https://www.jmir.org/2026/1/e93166
doi:10.2196/93166

© Sonia Garcia Gonzalez-Moral, Andrew Mkwashi, Fiona R Beyer, Dawn Craig. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Backcasting the Trust Gap: A Strategic Road Map for Clinician Adoption of AI Diagnostics by 2040

Yunguo Yu, MD, PhD

Zyter|TruCare, 2600 Tower Oaks Blvd, Suite 700, Rockville, MD, United States

Corresponding Author:
Yunguo Yu, MD, PhD
Zyter|TruCare, 2600 Tower Oaks Blvd, Suite 700, Rockville, MD, United States

Abstract

The integration of artificial intelligence (AI) into clinical medicine presents a persistent paradox: diagnostic models routinely demonstrate benchmark superiority over human experts, yet bedside adoption remains fragile, and clinician trust is low. Conventional forecasting approaches—projecting model performance along optimistic trend lines—are epistemologically insufficient because they cannot account for the nonlinear sociotechnical transitions that separate technical capability from institutional trust. This Viewpoint applies backcasting, a normative futures methodology with a 4-decade evidence base in energy policy and public governance, to the specific challenge of clinician adoption of AI diagnostics, with the aim of identifying the structural interventions required to achieve durable trust by 2040. Consistent with the tradition of single-expert normative foresight analysis, we applied backcasting as a structured reasoning framework using a STEEP (social, technological, economic, environmental, and political) analysis. Sources from PubMed, IEEE Xplore, Google Scholar, and policy repositories (the US Food and Drug Administration, World Health Organization, Organisation for Economic Co-Operation and Development, and European Commission) published between 2010 and 2025 were reviewed; barriers and enablers were coded across STEEP dimensions to identify pivot points representing convergent, time-bound structural changes. Working backward from a defined 2040 vision state—a health care ecosystem with risk-stratified clinician trust thresholds, semantic transparency of AI outputs, integrated AI governance, and futures literacy in medical education—we identified three temporal pivot points: (1) the 2030 standardization of dual-process AI architectures, in which large language models are verified in real time by locally deployed small language models, producing a calibrated confidence score; (2) the 2035 institutionalization of agentic AI orchestration governed by a formally designated chief AI officer; and (3) the 2040 integration of futures literacy and human-AI teaming competencies into standard medical curricula. The AI trust gap is an institutional design problem, not a technical inevitability. Backcasting reframes the central question from “when will AI be ready for medicine?” to “what must we build to make medicine ready for AI?” The 3 pivot points identified here—verifiable AI by 2030, agentic governance by 2035, and futures literacy by 2040—are structural commitments that clinicians, health system leaders, and policymakers can begin building today.

(J Med Internet Res 2026;28:e94234) doi:10.2196/94234

KEYWORDS

medical futures studies; backcasting; trustworthy artificial intelligence; trustworthy AI; small language models; confidence calibration; clinical decision support; futures literacy; health policy; implementation science

Introduction

Background

The clinical deployment of artificial intelligence (AI)–based diagnostic tools has produced a striking asymmetry: abundant proof-of-concept demonstrations coexist with a persistent deficit of durable bedside adoption. Large language models (LLMs) and deep learning classifiers regularly achieve expert-level performance on curated benchmark datasets, yet surveys of practicing clinicians consistently reveal skepticism, reluctance, and what has been termed “automation resistance” [1]. This is not primarily a technological problem. The bottleneck lies in the sociotechnical substrate: the transparency of AI outputs, the governance structures that certify their safety, the training programs that build clinical fluency, and—above all—the absence of a structured institutional pathway from pilot to practice.

The dominant response to this challenge has been *forecasting*: projecting from current trends in model accuracy, regulatory approval timelines, and computational costs to estimate when “AI will be ready.” This approach is epistemologically inadequate. Forecasting assumes that current trajectories are sufficient conditions for a desired outcome, ignoring the nonlinear policy, cultural, and organizational shifts that are necessary preconditions for deep integration. In complex adaptive systems such as health care, desired futures do not emerge spontaneously from optimistic trend lines; they must be actively constructed through deliberate intervention.

XSL•FO
RenderX

Medical futures studies, a nascent subfield identified and characterized by Meskó et al [2], offers a rigorous alternative. A 2024 scoping review of health care foresight found that only 8 of more than 50 established futures methods are currently applied in health care [2]. This represents a critical underutilization of the foresight evidence base, which has been routinely applied in economics, national defense, and environmental policy for over 4 decades [3].

This Viewpoint applies *backcasting* [4,5] to the specific challenge of clinician AI adoption.

Rather than asking "what will happen?" backcasting asks "given a desired future, what must happen to get there?" This reframing is essential: the goal is not to predict when AI will be trusted, but to identify the precise structural interventions that will make it trustworthy.

The components of this road map—verification architectures, governance roles, and human-AI teaming (HAT) frameworks—each have precedents in adjacent AI governance and implementation science literature. The contribution of this Viewpoint is not the individual components in isolation, but rather three integrated advances: (1) the use of backcasting as a unifying methodological framework applied specifically to clinician AI adoption, thereby addressing 4 of Dreborg's [5] canonical conditions for preferring backcasting over forecasting; (2) the identification of 3 pivot points as a dependencies-ordered causal sequence, in which each pivot is a necessary precondition for the next; and (3) the synthesis of dual-process verification, chief AI officer (CAIO) governance, and futures literacy as interlocking necessary conditions rather than parallel independent recommendations.

The aim of this Viewpoint is to apply backcasting as a structured foresight methodology to identify the minimum set of structural interventions required to achieve durable clinician trust in AI diagnostics by 2040.

Terminology

Throughout this Viewpoint, trust denotes calibrated clinician reliance on AI-generated outputs, as measured by validated automation trust instruments; confidence denotes a model-reported posterior probability estimate; adoption denotes durable clinical use beyond the pilot phase; automation resistance denotes persistent nonuse of AI tools despite demonstrated utility; and semantic transparency denotes the property whereby AI-generated outputs are natively linked to verifiable, computable evidence sources.

Why Backcasting Is the Appropriate Method

Backcasting was first formalized in energy policy analysis by Robinson [4] as a method for navigating problems where present trends are insufficient, or actively counterproductive, for achieving long-term goals. Dreborg's [5] canonical definition identifies 4 conditions under which backcasting is the preferred approach over forecasting: the problem is serious, conventional trends are part of the problem, a long time horizon is needed, and dominant interests are implicated. All four conditions apply to the AI trust crisis in medicine:

1. Seriousness—fragmented AI adoption risks entrenching "pilot-phase perpetuity," a state in which clinical AI tools are perpetually evaluated but never institutionalized, consuming resources without improving outcomes.
2. Current trends as part of the problem—the prevailing trajectory, releasing increasingly powerful LLMs into clinical contexts without any verification infrastructure, increases the rate of hallucination-driven errors, actively eroding the trust that adoption requires [6]. Independent benchmarking of LLMs on clinical tasks has corroborated this concern, documenting systematic calibration failures and overconfident outputs across specialties [7,8].
3. Long time horizon—institutional change in health care (ie, medical education reform, regulatory framework development, and governance role creation) operates on decade-scale timelines, not product release cycles.
4. Dominant interests—technology vendors, health systems, payers, and regulators each hold competing stakes in the current pilot-phase status quo. Only a normative vision of a desired future can align these interests around a common structural agenda.

Backcasting therefore offers a particularly appropriate methodological lens for addressing the AI trust gap in medicine—one aligned with each of the conditions under which Dreborg [5] argues that forecasting alone is insufficient.

Methodological Execution in This Viewpoint

Consistent with the tradition of single-expert normative foresight analysis, this Viewpoint applies backcasting as a structured reasoning framework rather than as a fully participatory design process. The pivot points were identified through a literature-grounded synthesis executed as follows:

- Sources and timeframe—we searched PubMed, IEEE Xplore, Google Scholar, and policy repositories (the US Food and Drug Administration [FDA], World Health Organization, Organisation for Economic Co-Operation and Development, and European Commission) for publications from 2010 to 2025.
- Search terms—search terms included clinical AI adoption, LLM hallucination health care, confidence calibration clinical decision support, AI governance health systems, CAIO, Futures Literacy medical education, and HAT.
- Social, technological, economic, environmental, and political (STEEP) operationalization—each identified barrier or enabler was coded to ≥1 STEEP dimensions; the 3 pivot points represent the convergence of barriers appearing across multiple dimensions within a shared time horizon.
- Prioritization—no formal Delphi or consensus technique was applied; selection reflects single-expert synthesis and is therefore subject to the perspective constraints acknowledged below.

XSL•FO
RenderX

We acknowledge that formal backcasting, as originally conceived by Robinson [4] and Dreborg [5], typically involves stakeholder cocreation, Delphi consensus, and scenario matrix comparison; those components are outside the scope of this Viewpoint and represent a natural next phase of this research agenda. The vision state, pivot points, and threshold specifications reported here should therefore be interpreted as normative hypotheses requiring multistakeholder validation, not as empirically confirmed targets.

Defining the 2040 Vision State

The normative foundation of any backcasting exercise is the *vision state*: a precisely specified description of the future toward which the road map is oriented. Vagueness at this stage compromises the entire analysis; the vision must be operationally defined to permit backward inference.

We emphasize at the outset that all threshold values, timeline dates, and structural specifications below are normative design hypotheses requiring multistakeholder validation, not empirically confirmed benchmarks. Their empirical validation is a core task for the formal multistakeholder Delphi process that constitutes the natural next phase of this research agenda.

We define the 2040 vision state as a health care ecosystem characterized by risk-stratified trust thresholds, that is, clinician-reported confidence in *specific AI-assisted diagnostic outputs*—distinct from general willingness to adopt AI—is assessed using validated trust-in-automation instruments adapted for clinical AI contexts [9]. Rather than a single system-wide percentage, the 2040 vision adopts a *risk-stratified trust matrix* in which the required confidence level scales with the clinical task's autonomy level and potential consequence: for *autonomous task execution* (eg, routine medication reconciliation and simple triage flagging), a trust score of ≥90% is required and for *assistive decision support* (eg, differential diagnosis generation and imaging interpretation), a score of 70% to 85% is acceptable contingent on human-in-the-loop verification. Final threshold values should be determined through specialty-stratified empirical validation and formal stakeholder consensus; if heterogeneous baselines are revealed, the road map may require adaptive pivot timelines.

These ranges are provisional normative anchors, not empirical measurements. Numerical specificity serves a methodological function in backcasting: it constrains the solution space and renders the vision state actionable for governance design. The specific values draw on adjacent automation trust literature [9] and should be replaced by specialty-validated figures as they become available through formal consensus methods. These principles are reflected in the following core dimensions of the proposed vision state:

- Semantic transparency—AI diagnostic outputs are natively linked to verifiable clinical evidence, such that every AI-generated claim carries a computable confidence score grounded in institutional guidelines and peer-reviewed literature.
- Integrated governance—every major health system operates under a formally designated AI governance structure, including a CAIO, with accountability equivalent to the chief medical officer (CMO).
- Futures literacy—medical graduates are trained in HAT and basic foresight competencies as standard components of the clinical curriculum.

This vision state is not utopian. Each element has precedent in an adjacent domain: semantic transparency parallels drug labeling requirements, the CAIO mirrors the chief information security officer trajectory of the 2000s, and futures literacy training already exists in business and public policy education. The question is not whether this state is possible, but what sequence of structural interventions will produce it.

A key concern with backcasting in a rapidly evolving technological domain is whether the vision state end point remains stable across the road map horizon. We address this by defining the vision state at the institutional level—governance structures, curriculum mandates, and trust thresholds—rather than at the technological specification level. Institutional targets are more stable than their underlying technology substrates: they describe what the system must do (verify, govern, and educate), not how it does so. Drug labeling requirements, for example, have survived multiple reformulation cycles without revision to the underlying mandate. We therefore propose an adaptive checkpoint mechanism: at each pivot point, the road map's technological assumptions should be reviewed against the prevailing state of the art. If those assumptions have materially shifted—for example, if the hallucination problem is addressed by architectural means that make local small language model (SLM) verification unnecessary—the subsequent pivot timelines should be adjusted accordingly, while the vision state's institutional targets remain fixed.

Working Backward: 3 Temporal Pivot Points

Backcasting proceeds by identifying *pivot points*—the minimum set of structural changes that, if achieved by a given date, preserve the feasibility of the 2040 vision. We identify 3 such pivot points using a STEEP framework [3] to ensure that each pivot is analyzed across all relevant dimensions.

The 2030 Pivot: The Verifiable AI Standard

The most proximate—and technically tractable—obstacle to the 2040 vision is the *hallucination* problem: the tendency of LLMs to generate plausible-sounding but factually incorrect clinical claims. Current mitigation strategies (retrieval-augmented generation and post hoc explainability tools) are insufficient because they operate retrospectively and rely on human review as the primary error-catching mechanism [10].

We propose that the critical technological pivot point for 2030 is the standardization of dual-process AI architectures, conceptually grounded in dual-process theory [11]. This framework includes the following components:

- LLM (system 1)—a general-purpose or fine-tuned LLM generates a rapid, high-breadth diagnostic hypothesis from

XSL•FO
RenderX

unstructured clinical data (notes, laboratory results, and imaging reports).

- SLM (system 2)—a domain-embedded SLM, running on local hardware (on premises), cross-references the LLM's output against a curated institutional guideline corpus *G* and a real-time literature index. It acts as a preceptor, not overriding the LLM but quantifying the evidentiary support for each diagnostic claim. The feasibility of privacy-preserving, redundant local multiagent architectures for clinical tasks has been demonstrated in prototype form [12](Figure 1 shows the 3 components).

Figure 1. Dual-process artificial intelligence architecture (the "verifiable AI" model)—conceptual framework. Patient data are processed in parallel by a large language model (LLM; generative hypothesis generation) and a locally deployed small language model (SLM; evidence-grounded verification against institutional guidelines *G*). Their outputs are combined into a calibrated confidence score $C(H \mid x, G)$; claims below verification threshold θ are flagged for mandatory clinician review. This figure depicts a proposed conceptual architecture; the confidence equation represents a formal specification requiring prospective validation in operational clinical environments before deployment. Accessibility description: flowchart showing patient data (electronic health record, laboratory results, and clinical notes) entering 2 parallel processing streams. The upper stream feeds an LLM labeled "system 1/fast/creative" for hypothesis generation. The lower stream feeds an on-premise SLM labeled "system 2/slow/critical" for guideline verification. Both streams converge at a calibrated confidence score node; outputs meeting or exceeding threshold θ are displayed in the clinician interface, while outputs below θ are flagged for mandatory review. EHR: electronic health record.

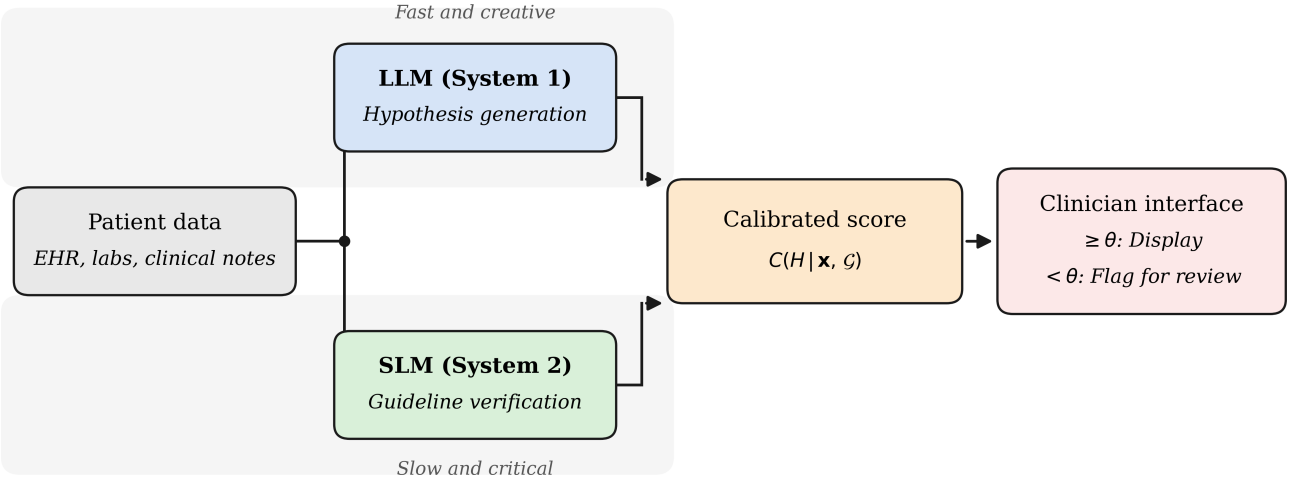

The output is a calibrated confidence score, defined formally as follows:

$$C(H \mid x,G)=P(H \mid x)\cdot 1\,[SSLM(H,G)\geq\theta]$$

where *H* is the diagnostic hypothesis; *x* is the clinical input vector; $P(H \mid x)$ is the LLM posterior; $S_{SLM}(H,G)$ is the SLM-derived guideline alignment score; θ is a system-configured verification threshold; and $[\cdot]$ is the indicator function, equal to 1 when the bracketed condition is satisfied and 0 otherwise. Claims falling below the threshold are flagged for mandatory clinician review rather than being silently passed to the interface. This formulation reflects a gating specification for conceptual clarity; alternative continuous calibration formulations (eg, weighted confidence scaling) may be explored in empirical implementation. Preliminary empirical support for this calibration approach was demonstrated in a validation study of 6689 cardiovascular cases from the Medical Information Mart for Intensive Care-III dataset, in which a dynamic confidence and transparency scoring framework reduced clinician override rates to 33.3%, with high-confidence predictions (90% - 99% confidence) overridden at only 1.7% [13]—representing a 20-fold reduction in override frequency at the highest confidence tier, a finding that suggests clinicians are meaningfully responsive to calibrated confidence signals when they are made explicitly visible. Analogous calibration challenges have been independently documented across diverse clinical AI systems, suggesting that structured verification layers represent a generalizable architectural need rather than a system-specific fix [14,15]. Federated learning frameworks for medical SLMs further demonstrate that locally deployed verification models can be collaboratively trained across heterogeneous institutions—with drift-aware rank scheduling to handle non–independent and identically distributed clinical data distributions—while preserving data locality, providing a practical cross-institutional update pathway for the SLM component of the dual-process architecture [16].

The critical structural condition for this pivot to occur by 2030 is not technological—SLMs capable of clinical guideline alignment already exist in prototype form—but *regulatory*: progressive movement toward formal minimum verification requirements for AI-assisted clinical decision support, analogous to the postmarket surveillance requirements applied to medical devices.

The 2035 Pivot: The Agentic Orchestration Shift

The 2030 pivot establishes the technical precondition for trust. The 2035 pivot concerns the institutional operationalization of that technical foundation. By 2035, under this road map, AI is envisioned to have evolved from a "consultant" role—passively responding to queries—to an *agentic orchestrator*: autonomously managing longitudinal care tasks such as postdischarge monitoring, medication reconciliation, and chronic disease management workflows [12].

This shift creates governance risks that current health care leadership structures are not designed to manage. Chief information officers (CIOs) lack clinical authority; CMOs lack AI technical literacy. The emerging CAIO role represents the

XSL•FO
RenderX

necessary institutional innovation: a formally credentialed position combining clinical expertise, AI technical competency, and governance authority to certify model safety, audit local calibration, and set institutional AI policy.

The CAIO is not intended to replace existing executive roles. Rather, the role bridges the CMO's clinical accountability and the CIO's technical infrastructure authority, with specific responsibility for (1) AI model certification and recertification, (2) local calibration auditing and equity reporting, and (3) institutional AI governance policy—functions currently unhoused in either the CMO or CIO role. The CAIO would report jointly to the CMO and the board's quality and safety committee, with dotted line coordination with the CIO and chief digital officer on technology procurement and data architecture.

The STEEP analysis at this pivot reveals that the primary barrier is not technological but *political*: medical licensing boards, accreditation bodies, and malpractice liability frameworks should be revised to recognize AI-assisted decisions as a category of collaborative clinical output, distinct from both autonomous device operation and unassisted physician judgment [17].

The *economic* and *social* dimensions of this pivot are equally consequential. Economically, the CAIO role and its associated governance infrastructure represent a substantial new cost center for health systems; the business case should be grounded in demonstrable efficiency gains from agentic AI—reduced administrative burden, improved care coordination, and averted adverse events—at a scale sufficient to offset implementation costs. For smaller systems without the resources to sustain a full CAIO function, "CAIO-as-a-service" models offered through regional health authorities or collaborative networks may provide a viable alternative pathway. Socially, the professional identity of the CAIO remains undefined: whether the role is best filled by a clinician with advanced data science training, a computer scientist with a public health qualification, or an entirely new hybrid professional developed through dedicated graduate programs has direct implications for how the 2040 educational pivot should be designed. Accreditation bodies and medical schools should begin addressing this pipeline question within the current planning horizon.

The 2040 Pivot: Futures Literacy as a Clinical Competency

The final pivot is structural and educational. The 2040 vision state requires clinicians who are not merely passive recipients of AI tools but active participants in shaping the technological environments in which they practice. Integrating futures literacy into medical education requires no new infrastructure; it requires a curriculum decision. Specifically, we propose the incorporation of three elements: (1) scenario analysis into clinical reasoning curricula, replacing some linear differential diagnosis training with probabilistic, multifuture reasoning [18]; (2) HAT modules into clerkship training; and (3) foresight workshops into medical leadership development programs, parallel to the quality improvement and health systems science competencies already mandated by accreditation bodies [19].

Contingency Logic and Adaptive Pathways

The 3 pivot points above are presented as a dependencies-ordered sequence, but structured backcasting acknowledges that execution may deviate from normative timelines [4,5]. We therefore specify a contingency logic for each pivot. If the 2030 verifiable AI standard is not achieved by the target date—for example, because regulatory standardization is delayed—an interim substitute exists: a voluntary industry consortium standard (analogous to the Health Level 7 Fast Healthcare Interoperability Resources interoperability framework) could provide provisional verification requirements while formal regulation catches up, with the pivot window extended to 2033 and subsequent timelines compressed. If the 2035 agentic governance pivot is delayed, the CAIO-as-a-service model already described provides a distributed interim pathway that does not require mandatory institutional mandates. If the 2040 futures literacy integration has not been achieved, the vision state trust thresholds should be treated as aspirational targets rather than confirmed baselines, and the multistakeholder Delphi process proposed in this Viewpoint becomes the primary recalibration mechanism. At each pivot point, a structured reassessment—rather than abandonment of the vision state—is the appropriate response to schedule deviation.

The Backcasting Timeline

Figure 2 summarizes the 3 pivot points and their relationship to the 2026 baseline and 2040 vision state. Table 1 provides a cross-pivot summary of the required changes across the STEEP framework.

XSL•FO
RenderX

Figure 2. Backcasting timeline: 2026 baseline to 2040 vision state. The dashed arrow represents the normative backcasting direction (vision to present). Open circles mark pivot points where structural interventions should be achieved. The 2040 node carries dual status as pivot 3 (futures literacy) and vision state terminal milestone. AI: artificial intelligence; CAIO: chief artificial intelligence officer; STEEP: social, technological, economic, environmental, and political.

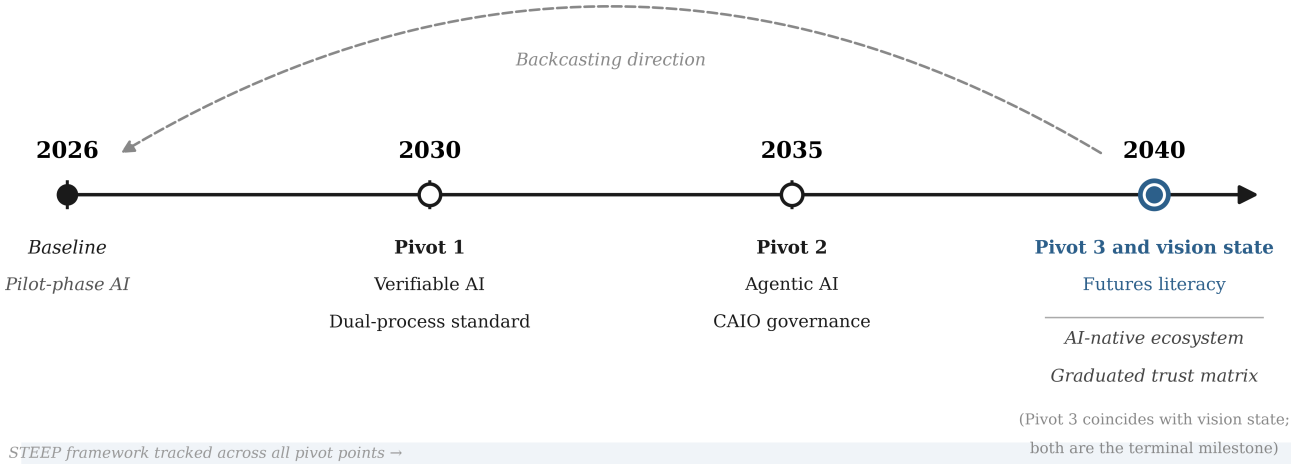

Table . Temporal milestones across the social, technological, economic, environmental, and political framework.

Dimension	2026 (baseline)	2030 (verifiable)	2035 (agentic)	2040 (vision)
Social	Clinician skepticism and automation fear	SLM[a] explainability builds baseline trust	AI[b] fluency enters professional identity	Futures literacy standard
Technological	General LLMs[c] and no verification layer	Dual-process architecture and on-premises SLM	Agentic orchestrators and ambient monitoring	Seamless XAI[d] and ambient intelligence
Economic	Pilot budgets and no billing pathways	SLM hardware ROI[e] established	AI-specific billing codes introduced	System-level AI cost integration
Political	Fragmented and IT-led governance	Regulatory movement toward formal verification requirements	CAIO[f] mandate and liability frameworks revised	AI-native governance as standard
Environmental	Cloud dependent and high energy cost	On-premise inference reduces footprint	Local edge computing architecture	Distributed, low-latency AI fabric

[a]SLM: small language model.
[b]AI: artificial intelligence.
[c]LLM: large language model.
[d]XAI: explainable artificial intelligence.
[e]ROI: return on investment.
[f]CAIO: chief artificial intelligence officer.

Discussion

The Automation Bias Risk

The central finding of this analysis is that the AI trust gap in clinical medicine is not a technical problem awaiting a technical solution, but an institutional design problem amenable to institutional design interventions. Three time-bound structural pivot points—verifiable AI (2030), agentic governance (2035), and futures literacy (2040)—constitute the minimum set of structural changes required to achieve the defined 2040 vision state. The following subsections examine the second-order risks and implementation considerations that the road map itself generates.

Backcasting toward greater AI integration must contend with a well-documented risk: as systems become more reliable, clinicians may exhibit automation bias, accepting AI outputs without appropriate critical evaluation [6]. Systematic reviews of automation bias in clinical decision support have documented its frequency across specialties and identified key mediating factors, including interface design and task complexity [20]. Paradoxically, the more successful the dual-process architecture becomes at reducing errors, the more it may erode the baseline clinical reasoning skills required to function safely during system failures.

The structural response is not to slow AI integration, but to design the verification layer as pedagogically visible. An SLM that surfaces its reasoning—linking each diagnostic flag to a specific guideline citation—functions simultaneously as a safety guardrail and a real-time teaching tool [21]. Critically, this requires a design commitment beyond mere filtering: the SLM should be presented to clinicians as a transparent reasoning partner, not an opaque gatekeeper, so that every interaction reinforces rather than replaces clinical judgment. Longitudinal

XSL•FO
RenderX

exposure to SLM-sourced guideline citations can thereby function as embedded continuing medical education, building the domain fluency that the 2040 futures literacy pivot requires. Clinical reasoning and AI collaboration are not zero-sum; they are mutually reinforcing when the verification architecture is made visible.

Algorithmic Equity and Siloed Bias

The shift from centralized cloud LLMs to locally deployed SLMs introduces a specific equity risk. If an institutional SLM is trained primarily on a single health system's demographic, it may produce well-calibrated outputs for that population and systematically miscalibrated outputs for underrepresented groups [22]. The CAIO role must therefore explicitly manage the tension between local optimization and national equity benchmarks—a trade-off that will intensify as local models diverge in their calibration profiles across demographically distinct health systems.

We propose 4 concrete structural safeguards. First, an equity audit mandate for the CAIO requires annual recalibration of local SLMs against nationally representative benchmarks, with mandatory public reporting of disaggregated performance metrics stratified by age, race, sex, and socioeconomic status. Enforcement authority would vest in the institutional CAIO, accountable to the same accreditation bodies that currently oversee quality improvement programs. Second, a *cross-institutional calibration exchange*—analogous to federated model benchmarking frameworks piloted in oncology imaging AI [8] and enabled technically by drift-aware federated SLM training [16]—would allow health systems to share calibration data without compromising data sovereignty or patient privacy; participation would require standardized data use agreements modeled on existing federated research network governance (eg, PCORnet). Third, a *public transparency layer* requiring institutions to publish SLM performance statistics across demographic strata would create accountability pressure independent of regulatory enforcement. Fourth, *minimum training data diversity requirements* for locally deployed SLMs—aligned with emerging FDA guidance on AI- or machine learning–based software as a medical device [23]—would establish upstream protections against demographically narrow calibration before deployment. We acknowledge, however, that these safeguards create a structural tension: local optimization for institutional demographics and national equity benchmarks may not be simultaneously achievable, and overly prescriptive diversity mandates risk regulatory fragmentation across health systems. Resolving this tension is a governance research priority for the 2035 pivot.

Data Sovereignty and Medical Agency

As agentic AI orchestrates longitudinal care decisions autonomously, the current legal framework becomes structurally incoherent: it assigns individual physician liability for all diagnostic and treatment decisions, a structure incompatible with distributed AI-human agency. The road map's 2035 pivot requires a parallel legal innovation: a framework for *collaborative medical agency* that assigns shared accountability across the human clinician, the institution's CAIO, and the AI system's certified governance record.

Aviation's joint liability structures for automated cockpit systems offer a suggestive—though imperfect—precedent. The analogy has important limitations: aviation operates under a single unified global regulatory body (International Civil Aviation Organization), whereas medicine is governed by fragmented national and regional authorities (FDA, European Medicines Agency, and others) with incompatible approval pathways. Moreover, the aviation cockpit pairs 2 human pilots with a bounded automated system; the clinical "cockpit" may involve 1 clinician simultaneously managing multiple AI agents across care domains.

Despite these differences, several concrete mechanisms from aviation governance translate meaningfully to clinical AI. First, a *decision record log*—analogous to a flight data recorder—could capture the AI system's inputs, confidence scores, and flagged uncertainties for every consequential clinical recommendation, creating an auditable governance trail for adverse event review. Second, new *institutional malpractice instruments* could name the health system's CAIO as a coinsured party for AI-assisted decisions, distributing liability in proportion to the documented quality of the institutional governance record rather than assigning it entirely to the attending clinician. Third, a *graduated autonomy protocol*—mirroring aviation's phased certification of autopilot authority at different flight stages—could specify the conditions under which agentic AI is permitted to act without real-time human confirmation, with those conditions contractually tied to demonstrated calibration performance. Sketching these mechanisms now is speculative; their legal operationalization is, by definition, a task for the 2035 pivot.

The Parallel Path of Patient Trust

This road map is explicitly clinician centric: its vision state is defined by clinician confidence in AI-assisted outputs, and its pivot points target the technical, institutional, and educational conditions for clinician adoption. However, the ultimate goal of that adoption is improved patient outcomes, and patient acceptance of AI-mediated care is a potentially rate-limiting variable that the road map cannot ignore. A clinician who trusts an AI-generated diagnosis may still encounter a patient who rejects it on the grounds that it originated from a "robot"—a concern with documented precedent in digital health adoption literature [1].

The 3 pivot points each carry patient trust implications. At the 2030 pivot, the semantic transparency requirement—that every AI-generated claim carries a computable confidence score linked to verifiable evidence—could be extended to produce patient-facing explanations: plain-language summaries of why the AI reached a given conclusion, designed to accommodate variability in health literacy. At the 2035 pivot, the CAIO's governance mandate should include a patient education function: communicating institutional AI policies, audit outcomes, and safeguard mechanisms to patients and communities in accessible formats. By the 2040 vision state, futures literacy in medical education should encompass shared decision-making frameworks for AI-assisted diagnosis, training clinicians to interpret AI outputs collaboratively with patients rather than presenting them as authoritative verdicts. A full account of

XSL•FO
RenderX

patient trust is outside the scope of this Viewpoint but represents a necessary parallel research agenda.

Operational Implementation Barriers

Several operational implementation barriers require explicit acknowledgment beyond the governance and regulatory dimensions already discussed. First, interoperability with legacy electronic health record systems presents a significant technical challenge: locally deployed SLMs must integrate with existing Health Level 7 or Fast Healthcare Interoperability Resources stacks and clinical data repositories through vendor-neutral application programming interfaces, a requirement that presupposes standards that do not yet uniformly exist. Second, workflow burden is nontrivial; a verification layer that introduces perceptible latency at the point of care risks clinician work-arounds that negate its safety function. Third, institutional digital maturity is highly variable: rural and safety-net health systems may lack the compute infrastructure required to run local SLMs, making the CAIO-as-a-service model described above a structural prerequisite rather than an optional alternative for a substantial portion of the health care system. Fourth, medico-legal fragmentation across jurisdictions means that governance frameworks developed for the US regulatory context will require substantial adaptation in the European Union, the United Kingdom, and other settings with different medical device and liability regimes. Fifth, procurement constraints in smaller health systems—where capital budget cycles operate on 5-year timelines—may not align with the decade-scale pivot timelines proposed here; phased procurement schedules and regional consortium purchasing agreements may be necessary to close this gap. Addressing these barriers is a practical prerequisite for the road map's transition from normative design to operational deployment.

Limitations

This Viewpoint has 5 principal limitations. First, backcasting is a normative, not predictive, method; the pivot points identified here are necessary conditions for the 2040 vision, not guaranteed outcomes. Second, the 2040 vision state and its risk-stratified trust thresholds (≥90% for autonomous tasks; 70% - 85% for assistive support) reflect expert judgment rather than empirical consensus [13]; future work should establish this vision through a formal multistakeholder Delphi process. Third, this analysis was conducted from a single professional vantage point; patient scholars, frontline nurses, and perspectives from low-resource settings are underrepresented. Fourth, the dual-process architecture requires empirical validation against real clinical datasets to establish whether SLM verification demonstrably improves the calibration of LLM diagnostic outputs. We regard this limitation as a tractable research priority and plan to address it in subsequent work using deidentified clinical data. Fifth, there is a geographic scope limitation: the governance mechanisms, liability frameworks, and accreditation references in this road map are drawn primarily from the US regulatory context (FDA, Accreditation Council for Graduate Medical Education, and state-level malpractice law). Readers in the European Union, the United Kingdom, or other jurisdictions will find the technical and educational pivot points broadly applicable but should adapt the political and legal dimensions to their national regulatory frameworks. A multijurisdictional extension of this road map is a planned next step.

Conclusions

The trust gap between AI capability and clinical adoption is neither technical nor inevitable. It is an institutional design problem, and it requires institutional design solutions. Backcasting transforms the question from "when will AI be ready for medicine?" to "what must we build to make medicine ready for AI?" The 3 pivot points identified here—verifiable AI by 2030, agentic governance by 2035, and futures literacy by 2040—are not predictions; they are structural commitments that clinicians, health system leaders, and policymakers can begin building today.

Medical futures studies offer the community of practice needed to sustain this work. We invite clinicians, researchers, and AI developers to engage formally with backcasting, scenario planning, and other foresight methods—not as academic exercises, but as essential planning instruments for building institutional capacity to shape health care's technological future. We additionally propose a multi-institutional consortium to prospectively evaluate the dual-process architecture using deidentified electronic health record data across diverse health systems, with the goal of establishing the empirical pivot evidence base that this normative road map cannot itself provide. The road map offered here is intended as a structured starting point for that institutional work.

Acknowledgments

The author would like to thank the guest editors of this theme issue, Dr. Bertalan Meskó and Dr. Tamás Kristóf, whose 2024 scoping review [2] directly motivated this work.

Funding

This research received no external funding.

Authors' Contributions

YY conceived the study, conducted the backcasting analysis, and wrote the manuscript in its entirety.

XSL•FO
RenderX

Conflicts of Interest

YY is employed by Zyter|TruCare, a health technology company. The proposed dual-process artificial intelligence (AI) architecture and chief AI officer role are conceptual contributions to the scholarly literature and do not describe, endorse, or promote any product or service offered by YY's employer.

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
2. Meskó B, Kristóf T, Dhunnoo P, Árvai N, Katonai G. Exploring the need for medical futures studies: insights from a scoping review of health care foresight. J Med Internet Res 2024 Oct 9;26:e57148. [doi: 10.2196/57148] [Medline: 39383528]
3. Popper R. How are foresight methods selected? Foresight 2008 Oct;10(6):62-89. [doi: 10.1108/14636680810918586]
4. Robinson JB. Unlearning and backcasting: rethinking some of the questions we ask about the future. Technol Forecast Soc Change 1988 Jul;33(4):325-338. [doi: 10.1016/0040-1625(88)90029-7]
5. Dreborg KH. Essence of backcasting. Futures 1996 Nov;28(9):813-828. [doi: 10.1016/S0016-3287(96)00044-4]
6. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med 2018 Mar 15;378(11):981-983. [doi: 10.1056/NEJMp1714229] [Medline: 29539284]
7. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
8. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
9. Jian JY, Bisantz AM, Drury CG. Foundations for an empirically determined scale of trust in automated systems. Int J Cogn Ergon 2000 Mar;4(1):53-71. [doi: 10.1207/S15327566IJCE0401_04]
10. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1):310. [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
11. Kahneman D. Thinking, Fast and Slow: Farrar, Straus and Giroux; 2011.
12. Yu Y. Hybrid-code: a privacy-preserving, redundant multi-agent framework for reliable local clinical coding. arXiv. Preprint posted online on Dec 26, 2025 URL: https://arxiv.org/html/2512.23743v1 [accessed 2026-04-20]
13. Yu Y, Gomez-Cabello CA, Haider SA, et al. Enhancing clinician trust in AI diagnostics: a dynamic framework for confidence calibration and transparency. Diagnostics (Basel) 2025 Aug 30;15(17):2204. [doi: 10.3390/diagnostics15172204] [Medline: 40941691]
14. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. Presented at: Proceedings of the 34th International Conference on Machine Learning; Aug 6-11, 2017 URL: https://proceedings.mlr.press/v70/guo17a/guo17a.pdf [accessed 2026-04-28]
15. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature New Biol 2023 Apr;616(7956):259-265. [doi: 10.1038/s41586-023-05881-4] [Medline: 37045921]
16. Yu Y. AdaptiveFedLoRA: drift-aware adaptive LoRA rank scheduling for federated medical small language models. medRxiv. Preprint posted online on Jan 21, 2026. [doi: 10.64898/2026.01.18.26344237]
17. Price WN II, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. JAMA 2019 Nov 12;322(18):1765-1766. [doi: 10.1001/jama.2019.15064] [Medline: 31584609]
18. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. Acad Med 2018 Aug;93(8):1107-1109. [doi: 10.1097/ACM.0000000000002044] [Medline: 29095704]
19. Miller R. Transforming the Future: Anticipation in the 21st Century: Routledge; 2018.
20. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 2012;19(1):121-127. [doi: 10.1136/amiajnl-2011-000089] [Medline: 21685142]
21. Miller T. Explanation in artificial intelligence: insights from the social sciences. Artif Intell 2019 Feb;267:1-38. [doi: 10.1016/j.artint.2018.07.007]
22. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
23. Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. U.S. Food and Drug Administration. 2021. URL: https://www.fda.gov/media/145022/download [accessed 2026-04-20]

Abbreviations

AI: artificial intelligence
CAIO: chief artificial intelligence officer
CIO: chief information officer
CMO: chief medical officer
FDA: Food and Drug Administration

XSL•FO
RenderX

HAT: human-AI teaming
LLM: large language model
SLM: small language model
STEEP: social, technological, economic, environmental, and political

Edited by B Mesko, T Kristóf; submitted 26.Feb.2026; peer-reviewed by TAK Manne, V Patil; revised version received 08.Apr.2026; accepted 10.Apr.2026; published 30.Apr.2026.

Please cite as:
Yu Y
Backcasting the Trust Gap: A Strategic Road Map for Clinician Adoption of AI Diagnostics by 2040
J Med Internet Res 2026;28:e94234
URL: https://www.jmir.org/2026/1/e94234
doi:10.2196/94234

© Yunguo Yu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Mapping Practice-Based Signals of Generative AI in Psychiatric Care: Qualitative Study of Korean Psychiatrists' Experiences, Interpretations, and Implementation Priorities

Myungsung Kim[1*], BS; Yoosuk An[2,3,4*], MD; Min Jeon[1], BSN; Yunji Lee[1], BS; Orane Farrah Lahcine[1], BS; Hyorim Kim[1], BS; Seonmi Lee[5], PhD; Sangil Lee[6], PhD; Jong-Chul Yang[7], MD, PhD; Sang-Won Jeon[8], MD, PhD; Dooyoung Jung[5,9], MD, PhD; Chul-Hyun Cho[10,11,12], MD, PhD

[1]Graduate School of Health Science and Technology, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea

[10]Department of Psychiatry, College of Medicine, Korea University, Seoul, Republic of Korea

[11]Department of Biomedical Informatics, College of Medicine, Korea University, Seoul, Republic of Korea

[12]Department of Medical Education, College of Medicine, Korea University, Seoul, Republic of Korea

[2]Department of Psychiatry, National Traffic injury Rehabilitation Hospital, Yangpyeong, Gyeonggi-do, Republic of Korea

[3]Department of Neuropsychiatry, Seoul National University Hospital, Seoul, Republic of Korea

[4]Department of Psychiatry, College of Medicine, Seoul National University, Seoul, Republic of Korea

[5]School of Digital Humanities and Computational Social Sciences, Korea Advanced Institute of Science and Technology, N4 Bldg, Room 1314, 291 Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea

[6]School of Liberal Arts, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea

[7]Jeonbuk National University Hospital, Jeollabuk-do, Republic of Korea

[8]Department of Psychiatry, Sungkyunkwan University School of Medicine, Kangbuk Samsung Hospital, Seoul, Republic of Korea

[9]Mind Care and Growth Center, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Dooyoung Jung, MD, PhD
School of Digital Humanities and Computational Social Sciences, Korea Advanced Institute of Science and Technology, N4 Bldg, Room 1314, 291 Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea

Abstract

Background: Generative artificial intelligence (GenAI) has increasingly entered psychiatric practice through patient-facing chatbots, self-help tools, and clinician-facing workflow support. Although prior research has examined clinicians' attitudes, readiness, and anticipated use cases, less is known about how frontline encounters with GenAI shape psychiatrists' interpretations and implementation priorities. Health care foresight also remains methodologically underdeveloped and has focused mainly on external signals, overlooking clinically consequential signals emerging from everyday practice. This gap is especially important in psychiatry, where GenAI-related benefits and harms may depend on patient vulnerability, crisis sensitivity, and the therapeutic relationship.

Objective: This study aims to qualitatively examine how South Korean psychiatrists described clinical experiences with GenAI, how they interpreted its roles and limits in psychiatric care, and what implementation priorities they emphasized. Selected concepts from horizon scanning informed the organization of the analysis by orienting attention to practice-based signals, interpretive patterns, and implementation priorities.

Methods: In this qualitative descriptive study, directed content analysis and codebook-based thematic synthesis were used to analyze responses to 3 open-ended survey questions administered to members of the Korean Neuropsychiatric Association. Invitations were distributed through the association's official email system from October 27 to December 26, 2025. The qualitative analysis included respondents who provided an interpretable response to at least 1 item. The questions addressed (1) GenAI-related clinical experiences, (2) perceived advantages and limitations of GenAI relative to human therapists, and (3) priorities for the safe introduction of GenAI into mental health care. An exploratory participant-level cross-question thematic alignment analysis was also conducted to examine recurring adjacent-item pairings across the experience-interpretation-priority sequence.

Results: Of 408 total survey respondents, 311 respondents provided a meaningful response to at least 1 open-ended item. Psychiatrists described GenAI as a clinically ambivalent technology whose implications depended on context, intensity of use, and patient vulnerability. Practice-based signals clustered around patient-led use, clinician-led use, GenAI as a relational object, and GenAI-mediated changes in the patient-clinician interface, with high-risk and destabilizing scenarios cutting across these

XSL•FO
RenderX

themes. Respondents viewed GenAI as potentially useful as an adjunct, but also as relationally limited and unacceptable as a replacement for human therapists. Implementation priorities centered on governance, crisis and vulnerability safeguards, technical reliability and clinical validation, and education, supervision, and structural readiness. Cross-question analysis suggested recurrent alignments between frontline signals, a view of GenAI as standardized and tireless but relationally thin, and governance- and validation-oriented implementation priorities.

Conclusions: In this qualitative descriptive study, GenAI emerged in psychiatric practice as an access tool, a workflow aid, and, at times, a competing interpretive reference point in clinical encounters. The key implementation challenge is therefore not whether psychiatry will encounter GenAI, but how its use should be bounded, supervised, and governed in light of patient vulnerability, psychiatric risk, and the relational demands of care.

Trial Registration: Clinical Research Information Service (CRIS) KCT0011712; https://cris.nih.go.kr/cris/search/detailSearch.do?seq=32747

(J Med Internet Res 2026;28:e96556) doi:10.2196/96556

KEYWORDS

mental health artificial intelligence; generative artificial intelligence; psychiatrists; qualitative research; thematic analysis; horizon scanning

Introduction

In this digital era, generative artificial intelligence (GenAI) is becoming embedded in everyday life and psychiatric systems, and is entering psychiatric practice through patient-facing conversational tools, self-help applications, and clinician-facing workflow support [1-3]. Psychiatrists are therefore likely to encounter GenAI not as a distant technological prospect, but as a technology already shaping how patients interpret distress, seek help, and present to the clinic. Recent studies suggest that clinicians are not uniformly resistant to artificial intelligence (AI); rather, they tend to view it as practically useful for documentation tasks, information processing, and other efficiency-oriented functions, while remaining far more cautious about its use in psychotherapy-like or relationship-centered care [4-9].

Given the complexity surrounding patient vulnerability, reality testing, crisis sensitivity, and the therapeutic relationship itself in psychiatry, the relevance of GenAI in psychiatric practice cannot be judged solely in terms of usability, convenience, or technical fluency. Recent work has similarly emphasized that conversational AI in health-related settings should be evaluated not only for performance or convenience, but also for risk-sensitive reasoning and oversight requirements [10]. Moreover, emerging literature has raised concerns that GenAI chatbots may reinforce delusion-like beliefs, provide excessive validation of distorted interpretations, or respond inadequately to suicide-related and other high-risk scenarios [11-17].

Additional work has suggested that prolonged or emotionally salient chatbot use may also be associated with dependence-like patterns, social withdrawal, or psychosocial deterioration in vulnerable users [13,14]. Taken together, these findings suggest that the clinical implications of GenAI in psychiatry are not reducible to generic benefits of access or conversational convenience. Rather, they depend on whether these tools are introduced into situations requiring contextual judgment, relational attunement, and the capacity to recognize and contain risk [2,15,16].

Much of the current literature has focused on clinicians' attitudes, readiness, trust, adoption intentions, and anticipated applications of GenAI in mental health care [4-9]. Less insight is available into how concrete frontline encounters with GenAI are interpreted by psychiatrists in practice, how different types of experiences shape professional judgments about GenAI's proper role, and how such judgments translate into priorities for supervision, governance, and implementation. Related qualitative work further suggests that physicians expect GenAI to affect the therapeutic relationship in both supportive and disruptive ways, including through altered trust, transparency, and shared decision-making [18]. Yet the field still lacks an approach that systematically links reported clinical signals with broader interpretive patterns and future-oriented implementation demands.

At a methodological level, similar concerns have been raised regarding health care foresight. In a recent scoping review, Meskó et al [9] found that only a limited subset of futures and foresight methods has been used in medicine and health care, and that their theoretical and methodological foundations remain insufficiently developed. Traditional horizon scanning, in particular, has been criticized for being less sensitive to weak signals first emerging within everyday clinical encounters and for generally focusing on external sources (eg, publications, patents, policies).

This limitation may be especially important in psychiatry, where early signs of benefit, boundary tension, or harm may first emerge in everyday interactions between patients, clinicians, and digital tools rather than in formal evidence hierarchies. In this sense, psychiatric foresight may need to begin with practice-based signals, not external innovation indicators alone [19].

South Korea provides a particularly relevant context for examining these questions. The country combines high digital uptake with persistent gaps in treatment utilization for mental disorders [20,21]. Help-seeking is shaped not only by stigma, but also by structural concerns related to psychiatric records, employment disadvantage, and other forms of institutional discrimination [22]. In this context, anonymous, immediate,

XSL•FO
RenderX

and low-threshold AI-based support may be especially attractive to individuals reluctant to seek formal psychiatric care. At the same time, these conditions may increase the likelihood of GenAI being used without adequate clinical oversight, intensifying concerns about safety, scope, and accountability. Recent global guidance has emphasized that large multimodal AI systems in health require stronger governance, transparency, risk management, and human accountability [22], yet how these principles should be translated into the realities of psychiatric practice remains underdeveloped. This concern is also consistent with recent AI governance scholarship arguing that responsible deployment requires deeper value-oriented oversight structures beyond external constraints alone [23].

To address these gaps, this study draws on horizon scanning concepts to organize a qualitative analysis of open-ended responses from psychiatrists in South Korea. The aim was to map practice-based signals arising in clinical encounters, identify broader interpretive patterns regarding the roles and limits of GenAI in psychiatric care, and characterize future-oriented implementation priorities. Rather than examining attitudes toward GenAI in the abstract, we sought to identify clinically salient frontline signals, characterize how psychiatrists interpret the roles and limits of GenAI in mental health care, and examine how these patterns relate to future implementation. We also examined how themes aligned across the connected analytic domains of experience, interpretation, and implementation priority. By linking experience, interpretation, and implementation priorities within a single analytic structure, this study aims to contribute empirically by clarifying how psychiatrists are already encountering GenAI in practice, and methodologically by illustrating how horizon scanning concepts can be adapted for practice-based signal mapping in psychiatry [9].

Building on these aims, this study contributes in several ways by clarifying key patterns in how GenAI is interpreted and implemented in psychiatric practice. Specifically, it (1) shows that clinicians' interpretations of GenAI are context-dependent and shaped by psychiatric vulnerability, (2) provides a structured account of the implementation gap through an access–protection tension, and (3) demonstrates a practice-based approach to horizon scanning using clinician-reported experiential data.

Methods

Study Design

This qualitative descriptive study analyzed open-ended survey responses using directed content analysis and codebook-based thematic synthesis. Selected concepts from horizon scanning informed the organization of the analysis across 3 linked domains: practice-based signals, interpretive patterns, and implementation priorities [19]. Reporting was guided by the Standards for Reporting Qualitative Research [24], and the description of theme development and analytic procedures was written with attention to recent guidance on transparent thematic analysis reporting [25].

Setting and Participants

From October 27 to December 26, 2025, invitations to participate were distributed 5 times through the official email system of the Korean Neuropsychiatric Association to all registered members. A total of 408 participants, including 326 board-certified psychiatrists and 82 residents, were included in the overall survey sample.

For the qualitative component, respondents were eligible if they provided an interpretable response to at least 1 of the 3 open-ended questions. Because each item was analyzed separately, the analytic denominator varied by question. Respondents who did not answer all 3 questions were still retained in the qualitative dataset if they contributed interpretable text to at least 1 item, because a single response could still provide meaningful insight relevant to the corresponding analytic domain.

Exclusion was applied at the item level rather than the respondent level. Brief responses were not excluded solely because of length; single-word or short responses were retained if they were semantically interpretable and directly addressed the question. Responses were excluded only when both coders judged them to be noninterpretable, off-topic, or nonsubstantive, with disagreements resolved through discussion.

Open-Ended Survey Questions

The present analysis was restricted to the following 3 open-ended survey questions. Below are English translations of the original Korean survey questions:

- Q1: "Please freely share any memorable event or experience related to generative AI technologies in clinical practice, such as chatbots or diagnostic support tools."
- Q2: "Please describe your views on the unique advantages and limitations of generative AI in mental health care compared with human therapists."
- Q3: "Please list up to 3 tasks that you believe are most urgent for the safe and effective introduction of generative AI technologies into mental health care."

Although each question was designed to stand on its own, the item sequence was intentionally structured to align with the study's analytic framework across experience, interpretation, and implementation priority. These corresponded to the study's 3 analytic domains: practice-based signals, interpretive patterns, and implementation priorities.

Although each question was designed to stand on its own, the item sequence was intentionally structured to align with the study's 3 analytic domains: practice-based signals, interpretive patterns, and implementation priorities.

Use of Selected Horizon Scanning Concepts for Practice-Based Signal Mapping

This study drew on selected concepts from horizon scanning as an organizing heuristic. Two elements were adapted: first, the notion of practice-based signals, namely clinically consequential early indicators emerging within routine care before they appear in formal evidence hierarchies; and second, a 3-layer structure linking practice-based signals, interpretive patterns, and

XSL•FO
RenderX

implementation priorities. Unlike conventional horizon scanning, which relies on external data sources such as publications, patents, and policy documents, this study used clinicians' open-ended survey responses as the signal source and qualitative thematic analysis as the analytic method.

At the first layer, Q1 responses were examined as practice-based signals arising from clinical encounters, drawing conceptually on selected horizon scanning concepts [19]. At the second layer, Q2 responses were synthesized thematically to identify interpretive patterns regarding the perceived roles, limits, strengths, and boundaries of GenAI in mental health care relative to human therapists [26]. At the third layer, Q3 responses were synthesized to identify implementation priorities for the safer and more effective introduction of GenAI into mental health care [25]. Without imposing an a priori coding framework, responses were inductively coded and synthesized into broader implementation-priority themes reflecting clinicians' future-oriented concerns.

Integration across the 3 layers was conducted through cross-question analysis at the participant level. By linking each respondent's codes across Q1, Q2, and Q3, we explored recurring within-respondent alignments across the connected analytic domains of experience, interpretation, and implementation priority, with particular attention to alignments between adjacent domains.

In this study, horizon scanning concepts were used only to organize analytic attention, whereas coding, theme development, and synthesis remained grounded in qualitative content analysis and thematic synthesis.

Research Team and Reflexivity

The qualitative analysis was conducted by a multidisciplinary team with expertise in psychiatry, digital health, clinical psychology, nursing, biomedical engineering, and technology-enabled health research. MK, a doctoral student researcher with prior mixed methods experience in emerging digital mental health technologies, coordinated the analytic process. During the initial familiarization and open-coding phase, MK worked with YL and HK to review responses and generate preliminary data-near codes. MJ and YL, both of whom had prior supervised experience in qualitative analysis within digital health research, then conducted the main independent coding under MK's ongoing oversight.

Because team members differed in disciplinary background and prior qualitative experience, the analysis emphasized explicit code definitions, iterative discussion of coding boundaries, and repeated review of discrepant cases to strengthen consistency and reflexive awareness. Psychiatrist investigators (DJ, YA, and CHC) contributed clinical judgment regarding the plausibility and psychiatric significance of reported events, whereas nonphysician researchers focused on coding transparency, category boundaries, and interpretive consistency across responses. Reflexive discussions explicitly considered assumptions related to technological optimism, patient safety, psychiatric vulnerability, and professional boundary concerns throughout coding and theme development [27,28].

Codebook Development

Codebook development proceeded in stages using a combined inductive-deductive approach derived from directed qualitative content analysis [29]. First, all responses to each question were reviewed through open coding to generate an initial pool of data-near codes. This stage was conducted under the coordination of MK, with YL and HK contributing to preliminary code generation for each item.

Second, MK, MJ, and YL organized and refined these preliminary codes into separate item-specific codebooks. This refinement process was informed both by the data and by sensitizing concepts drawn from the literature, including clinical usefulness of GenAI, categories of risk, therapeutic relationship, and concepts from horizon scanning and technology adoption. Through this combined inductive-deductive process, separate codebooks were established for each question, each containing code labels, operational definitions, inclusion criteria, exclusion criteria, and illustrative quotations.

The preliminary codebooks were piloted on a subset of responses from each question and iteratively revised before full coding. During piloting, the team specifically reviewed whether meaning units were too broad or too narrow, whether adjacent codes overlapped conceptually, and whether brief responses could be coded consistently under the proposed definitions. Where ambiguities were identified, code labels, definitions, and decision rules were revised accordingly. This iterative process was informed by qualitative content analysis literature emphasizing that abstraction and interpretation are inseparable from the processes of condensing, coding, and forming categories [27,28]. Once consensus had been reached on code definitions and coding rules, the codebooks were fixed for the main coding phase, with only minor clarifications permitted thereafter.

A detailed audit trail of codebook refinement, discrepant-code review, and code-level intercoder agreement across coding rounds is provided in Multimedia Appendix 1.

Coding Procedure

Coding was conducted independently by MJ and YL. The analytic unit was the individual response to each open-ended question, with coding decisions made at the level of meaning units, allowing multiple codes to be assigned within a response when distinct ideas were identified.

For the first question, coders assigned all applicable substantive codes and also coded experiential proximity. For the second question, coders initially coded statements describing advantages and limitations separately. Rather than binary categories, codes were integrated to reflect that GenAI features often present simultaneous advantages and limitations depending on the clinical context. This coexistence was treated as central to interpretation rather than as a coding problem. For the third question, codes were assigned inductively to the implementation issues and priorities raised by respondents, without applying predefined categories of challenge type or urgency.

After an initial round of independent coding, coder consistency was reviewed at the response level using Cohen's kappa. This

XSL•FO
RenderX

step was used pragmatically to identify ambiguous code boundaries and refine code definitions during codebook development, rather than to imply a purely positivist model of qualitative rigor [30,31]. For codes with Cohen's kappa values below 0.70, MJ and YL reviewed discrepant cases with MK to identify the source of disagreement. Relevant codebook definitions and decision rules were then refined, and affected responses were recoded before final agreement statistics were calculated. Final code assignments were determined through consensus discussion rather than by excluding low-agreement codes from analysis. This iterative refinement process was designed to improve coding consistency while preserving analytically important distinctions in the data [27-29].

Following completion of the coding phase, YA participated in theme development and interpretive synthesis as a psychiatrist investigator. At this stage, YA contributed clinical judgment regarding the psychiatric significance of code patterns, the plausibility of reported clinical scenarios, and the linkage between coded response patterns and implications for psychiatric practice. Detailed information on changes in intercoder agreement and codebook refinement across coding rounds is reported in Multimedia Appendix 1.

Thematic Synthesis

Following item-level coding, codes were grouped into subthemes and broader themes through team-based thematic synthesis. Braun and Clarke's broad phases of familiarization, coding, theme development, review, and naming informed the organization of this process, while the analysis remained grounded in a qualitative descriptive, codebook-based approach [26]. Within each question, codes were grouped into subthemes according to shared clinical meaning, interpretive orientation, or implementation relevance, and these subthemes were then synthesized into broader themes.

For the first question, the thematic structure was interpreted in relation to signal content and experiential proximity. For the second question, themes were developed to capture broader interpretive patterns in respondents' views of GenAI in mental health care, including tensions, overlaps, and boundary judgments within the same thematic domain. Accordingly, potentially positive and negative aspects of a given feature were retained together when they formed part of a single coherent theme. For the third question, themes were developed inductively to represent clinicians' perceived implementation priorities and governance.

Theme development was conducted through joint review and interpretive synthesis by the research team after completion of coding. In this phase, psychiatrist and nonpsychiatrist investigators contributed different but complementary perspectives. More precisely, the former emphasized clinical plausibility, contextual meaning, and implications for psychiatric care, whereas the latter focused on thematic coherence, transparency of analytic transitions, and fidelity to the wording of the responses. Recent guidance on transparent thematic analysis reporting informed the description of these analytic steps in the manuscript [25].

Exploratory Cross-Question Thematic Alignment Analysis

Cross-question analysis was conducted using a participant-level coding matrix. For each respondent, the matrix linked assigned codes across the 3 open-ended questions. Because quantified co-occurrence is not typically presented as a primary output of thematic analysis, the coding matrix was used to examine cross-item theme pairings as supplementary descriptive indicators. Specifically, we examined how particular alignments recurred across respondents, whether some pairings showed relatively dense overlap after accounting for base prevalence, and whether alignment patterns differed across adjacent items. This rationale was informed by prior work suggesting that thematic co-occurrence can help clarify not only whether themes overlap, but also how their relationships are patterned, asymmetric, and interpretively meaningful [32], as well as by qualitative-quantitative approaches using co-occurrence matrices or coefficients to support exploratory interpretation of coded qualitative data [33,34]. Accordingly, for each adjacent-item theme pair (Q1-Q2 and Q2-Q3), we calculated raw co-occurrence counts, directional conditional proportions, and a prevalence-adjusted overlap index based on the Jaccard (c-coefficient) formula: $nAB/(nA+nB-nAB)$. These indices were interpreted descriptively rather than inferentially. They were used to support narrative findings about recurrent and directionally patterned within-respondent thematic alignments, especially where relatively high-overlap pairings appeared to reinforce an existing concern, add a complementary explanatory layer, or connect frontline experience with future implementation priorities [34]. Consistent with thematic analysis, frequency was not treated as a proxy for thematic importance [35].

Rigor and Trustworthiness

Rigor was addressed using the criteria proposed by Lincoln and Guba [36]. Credibility was strengthened through independent double coding, iterative consensus discussions, involvement of multidisciplinary investigators, and the inclusion of illustrative quotations. Transferability was supported by a detailed description of participant characteristics and attention to the Korean clinical and regulatory context. Dependability was enhanced through explicit codebook development, documented refinement of code definitions and decision rules, and maintenance of an audit trail for coding discrepancies and revisions. Confirmability was addressed by distinguishing clearly between data-grounded coding decisions, theory-informed sensitizing concepts, and later interpretive synthesis.

Because the study concerned an emerging technology associated with both enthusiasm and alarm in public discourse, reflexive discussion formed an explicit part of the analytic process. Team members discussed potential preconceptions regarding technological optimism, patient safety, psychiatric vulnerability, and professional boundary concerns before and during analysis. These discussions informed codebook refinement and theme review, but were not used to override the data. In addition, the iterative review of discrepant codes and recoding of affected

XSL•FO
RenderX

responses helped make analytic decisions more transparent and less dependent on any single coder's assumptions [24,25,27,28].

Ethical Considerations

This study was approved by the Seoul National University Hospital Institutional Review Board (approval number SNUH IRB 2025-10-001). All responses were collected and analyzed anonymously, and electronic informed consent was obtained from all participants before survey initiation. Participants who completed the survey received a small incentive in the form of a 10,000 (USD $6.82) KRW Starbucks mobile gift voucher.

Results

Overview of Qualitative Responses

Of the 408 survey participants, 311 provided a meaningful response to at least 1 of the 3 open-ended questions after application of the predefined item-level exclusion criteria and were included in the qualitative analytic sample. Meaningful responses were available for 218 participants in Q1, 232 in Q2, and 220 in Q3. Demographic and professional characteristics of the qualitative sample are presented in Table 1.

Table . Demographic and professional characteristics of the qualitative analytic sample.

Variable	Total[a] (n=311)	Q1 (n=218)	Q2 (n=232)	Q3 (n=220)	Overlap[b] (n=129)
Age (years), mean (SD)	42.81 (9.89)	41.9 (9.23)	43.76 (9.80)	43.89 (10.04)	43.32 (9.14)
Clinical experience (years), mean (SD)	14.7 (9.90)	13.91 (9.16)	15.39 (9.84)	15.66 (10.14)	15.03 (9.08)
Sex, n (%)					
Male	197 (63.34)	135 (61.93)	143 (61.64)	136 (61.82)	76 (58.91)
Female	114 (36.66)	83 (38.07)	89 (38.36)	84 (38.18)	53 (41.09)
Professional status, n (%)					
Specialist	256 (82.32)	178 (81.65)	198 (85.34)	186 (84.55)	110 (85.27)
Resident	55 (17.68)	40 (18.35)	34 (14.66)	34 (15.45)	19 (14.73)
Workplace setting, n (%)					
Tertiary hospital	101 (32.48)	66 (30.28)	67 (28.88)	66 (30.00)	31 (24.03)
General hospital	37 (11.90)	27 (12.39)	28 (12.07)	28 (12.73)	17 (13.18)
Hospital	6 (1.93)	5 (2.29)	2 (0.86)	2 (0.91)	1 (0.78)
Psychiatric hospital	42 (13.50)	28 (12.84)	33 (14.22)	31 (14.09)	18 (13.95)
Clinic	120 (38.59)	90 (41.28)	99 (42.67)	90 (40.91)	62 (48.06)
Other	5 (1.61)	2 (0.92)	3 (1.29)	3 (1.36)	0 (0.00)

[a]Total refers to the set of participants who provided at least one meaningful response to Q1, Q2, or Q3.

[b]Overlap indicates respondents who provided meaningful responses to all 3 open-ended questions. In addition, pairwise overlaps (including respondents who also answered all three questions) were observed between Q1 and Q2 (n=144), Q2 and Q3 (n=210), and Q1 and Q3 (n=134).

Figure 1 provides a brief overview of the structure of the qualitative findings, showing the number of respondents contributing to each question-specific section and to the cross-question comparisons. The Results first present themes identified independently for Q1, Q2, and Q3, addressing practice-based signals, interpretive patterns, and implementation priorities, respectively, and then turn to an exploratory analysis of how themes aligned across adjacent question pairs at the participant level, particularly between Q1 and Q2 and between Q2 and Q3. Participant quotations were translated from Korean into English and lightly edited for readability while preserving the original meaning and tone. Quotations are identified by participant number, indicated as P followed by the respondent number (eg, P217).

XSL•FO
RenderX

Figure 1. Structure of the qualitative results and cross-question thematic alignment. GenAI: generative artificial intelligence.

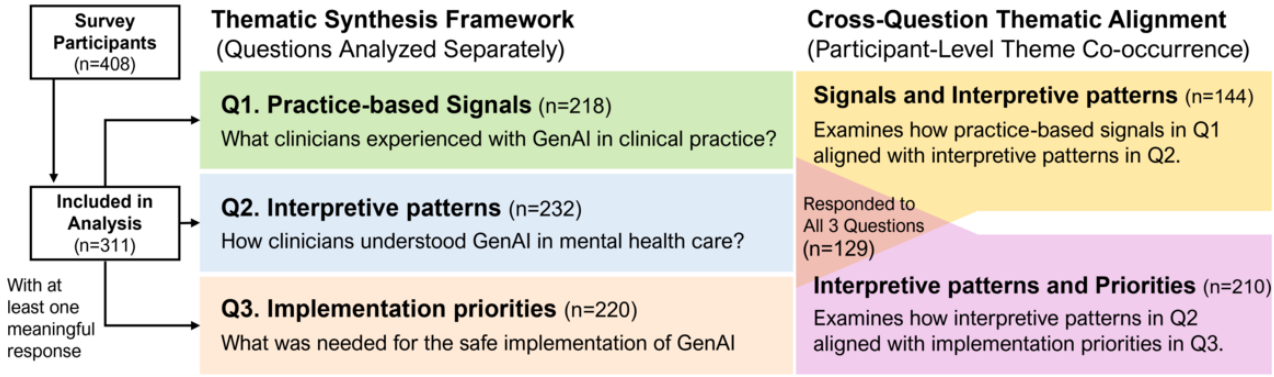

Practice-Based Signals Emerging From Clinical Encounters With GenAI (Q1)

Overview

Responses to Q1 suggested that psychiatrists' experiences with GenAI were organized less by a simple positive-versus-negative distinction than by how GenAI entered and functioned within care. Reported experiences clustered around four themes, with high-risk and destabilizing scenarios cutting across them. Across these themes, bounded or lower-intensity use tended to be described as facilitative, whereas higher-intensity use, overreliance, or transfer of interpretive authority to AI tended to be associated with symptom aggravation, treatment disruption, and safety concerns.

Patient-Led Signals: Self-Help, Triage, and Early Gateway Use

The first set of signals involved patient-initiated use of GenAI as a self-help, reflective, triage, or information-checking tool. In these responses, GenAI was generally described not as a substitute for treatment but as a low-threshold aid that sometimes helped patients organize emotions, consider whether professional help might be needed, or bring questions about AI use back into clinical care.

> *I saw a patient who did not use ChatGPT simply for comfort, but wrote out their situation and the emotions they were feeling at the time, and then asked it not to comfort them but to help organize them. Through this, the patient was able to verbalize emotions that had been difficult to process, and I confirmed that the symptoms improved.* [P217]

Another respondent described AI as prompting treatment entry:

> *The patient came after being advised during an interview with AI to seek care* [P74]

These responses suggested that psychiatrists were already encountering AI as a preliminary reflective or gateway tool before formal psychiatric contact. However, when patient-led use became more intense or occurred in higher-risk situations, respondents typically described deterioration or safety concerns rather than benefit. Reported experiences included reinforcement of delusion-like beliefs, worsening of nonpsychotic symptoms, exposure to inappropriate or unverified AI services, and suicide- or self-harm-related risk.

> *A patient once asked ChatGPT and, after being told the medication was not a high dose, took an overdose for self-harm.* [P20]

At the same time, some responses indicated that AI could also facilitate help-seeking in the opposite direction, such as when a patient with suicidal ideation received a crisis hotline number during chatbot use and was subsequently linked to services (P157). Overall, patient-led use appeared acceptable mainly when GenAI remained a preliminary aid rather than an unsupervised guide during vulnerable states.

Clinician-Led Signals: Workflow Support, Clinical Reasoning, and Bounded Experimentation

A second set of signals came from clinicians' own use of AI, which was described mainly in practical and workflow-oriented terms. Respondents reported using AI to summarize records, prepare educational materials, compare medications, and obtain rapid information or informal support, particularly for tasks experienced as structured, repetitive, or information dense.

> *When I used a chatbot to prepare educational materials or video materials for patients or caregivers, it helped me prepare them quickly with less effort. … Using a chatbot to create the basic outline and then having an expert review it made efficient work possible* [P230]

Another respondent noted that many physicians had likely already used GenAI informally to check drug interactions when the answer did not immediately come to mind (P224). These responses suggested that clinicians regarded GenAI as more appropriate for supportive and reviewable tasks than for independent psychiatric decision-making. Some emphasized hallucinated outputs, factual inaccuracy, and contextually inappropriate recommendations. One psychiatrist who had used paid general-purpose ChatGPT for literature and medication-related queries reported that it was useful in outpatient practice, but that medically inaccurate statements and even serious errors were "often noticeable," concluding that it was still too early to introduce such systems to the public for counseling, diagnosis, or treatment without expert review (P135). In short, clinicians accepted GenAI mainly for structured, reviewable tasks under human oversight.

XSL•FO
RenderX

GenAI as a Relational Object: Attachment, Substitution, and Social Functioning

A third pattern concerned GenAI not only as a tool but also as a quasi-relational object. Respondents described cases in which GenAI appeared to provide companionship, a psychologically safer space for disclosure, or support for self-expression and behavior change, particularly among socially isolated or inhibited patients.

> *A socially isolated patient whose sense of isolation was alleviated through chatbot use* [P24]

Another psychiatrist described a patient who had rarely disclosed inner thoughts to anyone, but who, after using a chatbot, became more able to express thoughts and feelings during treatment and even reported trying new interpersonal behaviors with the chatbot's help (P200). GenAI sometimes functioned as a relational scaffold rather than only as a technical aid, but these same features became concerning when reliance deepened, or symptom worsening coincided with intensified use.

> *In patients, the amount of conversation with ChatGPT increased whenever they relapsed. I thought that by monitoring the interaction itself, it might be possible to predict relapse* [P22]

Other respondents described patients saying that the chatbot understood them better than family members, or spending most of the day in conversation with it. In these cases, GenAI appeared to move beyond supportive aid and occupy a more central emotional or behavioral role, raising concerns about overdependence, social withdrawal, and reduced offline coping or relational capacity.

GenAI-Mediated Changes in the Patient-Clinician Interface

A fourth theme concerned ways in which GenAI was beginning to reshape the consultation itself. Respondents described patients arriving with AI-generated summaries, discussing prior chatbot interactions during visits, or using GenAI output to organize what they wanted to say to the psychiatrist. In lower-intensity encounters, AI sometimes facilitated communication and more efficient use of limited clinic time.

> *There was a patient who routinely discussed concerns with AI and, in the outpatient clinic, summarized those conversations and asked the treating physician's opinion. The patient was highly intelligent, and the conversations with AI dealt with clinically important topics, so I found it impressive that this seemed to use the limited clinic time efficiently* [P135]

At the same time, respondents also described more disruptive interface experiences in which GenAI became a comparator or competing interpretive reference point. Patients were reported to cross-check clinicians' advice with chatbot output, request medications based on GenAI recommendations, or compare the therapist's responses with those of a chatbot. In some cases, this reinforced trust when GenAI and the clinician converged, but in others it undermined diagnostic authority, treatment adherence, or the therapeutic relationship itself. Overall, AI entered the consultation not only as information, but also as a competing reference point that reshaped verification, authority, and trust.

> *A patient came to me angry after telling ChatGPT their symptoms and receiving an answer about the diagnosis, saying that their attending psychiatrist had given a different diagnosis and that they could no longer trust the attending physician. In the end, the attending psychiatrist's diagnosis was correct, and the problem had arisen because the patient mentioned only biased symptoms, that is, only those they themselves acknowledged* [P53]

Interpretive Themes Regarding GenAI in Mental Health Care (Q2)

Overview

Responses to Q2, which focused on GenAI in mental health care, did not separate neatly into advantages and limitations. Instead, the same properties, such as availability, nonjudgment, and standardized responding, were often described as beneficial in one clinical context and problematic in another. Four integrated themes were identified. Respondents' judgments appeared to depend less on whether a feature was inherently advantageous or problematic than on whether it could be appropriately bounded, contextualized, and assigned a clinically suitable role.

GenAI as a Low-Threshold and Always-Available Point of Contact

Many respondents viewed GenAI as uniquely accessible relative to human therapists. They emphasized its availability across time and place, immediate responsiveness, low cost, and reduced stigma through anonymity. These features were seen as particularly relevant in situations where conventional mental health care remained difficult to access or where patients hesitated to disclose symptoms to another person. In this sense, GenAI was interpreted as a low-threshold conversational entry that reduces barriers to contact.

> *AI chatbots have the advantage of immediacy and accessibility… Therefore, they should be understood not as therapeutic replacements, but as entry points for those who cannot access treatment and as tools that support human care.* [P33]

Standardized and Tireless, but Relationally Thin

Respondents also described GenAI as offering consistency, neutrality, and freedom from fatigue, countertransference, or mood-dependent variability. Some regarded these qualities as strengths, particularly for structured psychoeducation, repetitive support, or protocol-based interventions.

> *The strengths of AI chatbots are that they are standardized, evidence-based, and less prone to error. Their limitations are the lack of individualized responses that respect each person's uniqueness and the limited ability to convey empathy.* [P15]

However, these perceived strengths were frequently contrasted with the limitations of GenAI as a relational agent. Respondents emphasized that GenAI could not adequately perceive nonverbal

XSL•FO
RenderX

cues, grasp emotional nuance in the same way as a human therapist, or participate in the intersubjective processes through which therapeutic change often occurs. These strengths were simultaneously interpreted as signs of relational thinness and therapeutic incompleteness. At the same time, several responses implied that although some limitations might be improved technically, others were more fundamentally tied to the absence of human intersubjectivity, warmth, and embodied perception in psychiatric care.

Nonjudgmental Acceptance as Both Comfort and Clinical Hazard

A recurrent tension concerned GenAI's nonjudgmental and continuously accepting stance. Respondents viewed this nonjudgmental stance as double-edged: it could facilitate disclosure, but it could also become clinically hazardous.

> *Its advantages are that it is always available and allows disclosure of shameful content. Its risks are that users may hear only what they want to hear, while delusional thinking may be reinforced and social isolation may increase.* [P229]

They worried that GenAI might provide users with the answers they wanted to hear, fail to deliver necessary therapeutic challenge, or respond in an overly validating manner to distorted or dangerous beliefs. In this sense, respondents suggested that the problem was not support itself, but support that was insufficiently modulated by doubt, limit-setting, or reality testing. This concern became especially pronounced in relation to psychosis, delusion-like ideation, and suicide or self-harm risk, where respondents emphasized that apparently supportive responses could worsen, rather than contain, vulnerability.

Useful as an Adjunct, Not Acceptable as a Replacement

Across responses, acceptance of the GenAI chatbot was generally conditional and adjunctive rather than substitutive. Respondents were more comfortable assigning GenAI a supportive role than accepting it as an independent therapeutic agent. This bounded stance was closely tied to perceived limitations in contextual understanding, individualized formulation, accountability, and supervision. Many respondents suggested that once GenAI moved beyond support functions and began to occupy the position of therapist, diagnostic authority, or crisis responder, its limitations became substantially more concerning.

> *In conclusion, AI chatbots are most effective when used as a complement to, rather than a replacement for, human therapists.* [P388]

Implementation Priorities for Safer Psychiatric Use of GenAI (Q3)

Overview

Q3 translated these interpretive tensions into future-oriented implementation priorities for action. Rather than primarily calling for wider deployment, respondents emphasized the conditions under which GenAI could be introduced safely, legitimately, and responsibly in psychiatric settings. Four strategic priority themes were identified: governance and accountability, safety protections for high-risk groups, technical and evidentiary reliability, and educational or structural readiness. Taken together, these priorities were framed less as a roadmap for expansion than as preconditions for legitimate and safer use in psychiatric settings.

Governance and Accountability as Prerequisites for Adoption

Governance and accountability emerged as foundational prerequisites for adoption. Respondents frequently called for ethical standards, practical guidelines, legal frameworks, and clearer allocation of responsibility in the event of harm.

> *There should be a basic code of ethics for AI technology, along with practical guidelines and a supervisory body to ensure that it is used appropriately.* [P32]

These concerns extended beyond abstract principles to the operational question of who should be accountable when GenAI contributes to inaccurate advice, delayed intervention, privacy breaches, or other adverse outcomes. Many responses suggested that, in the absence of professional and regulatory scaffolding, broader implementation would be premature. Some responses also emphasized privacy and data protection, particularly given the sensitivity of psychiatric information and concerns about clinical or counseling content being entered into external GenAI systems without clear safeguards.

Safety Infrastructures for Crisis Situations and Vulnerable Populations

Safety infrastructures for high-risk scenarios and vulnerable groups were identified as essential implementation requirements. Respondents emphasized the need for systems capable of identifying suicidality, self-harm risk, or other emergencies and linking users to timely intervention.

> *Institutional safeguards are needed for patient safety, such as automatic reporting to relevant agencies or professionals when warning signs related to self-harm or harm to others are identified.* [P135]

They also highlighted the need for safeguards specific to psychosis and other conditions involving impaired reality testing, where AI's tendency toward excessive agreement or uncritical responding was seen as particularly dangerous. These responses suggested that clinicians did not view safety as a generic design feature, but as something that must be tailored to psychiatric risk profiles and clearly defined contraindications.

Technical Reliability and Clinical Validation Before Scale-Up

Respondents also emphasized that implementation should depend on technical and evidentiary maturity rather than novelty alone. Priorities in this theme included improving information accuracy, reducing hallucinations, developing mental health-specific systems, and validating safety and effectiveness through clinical research.

> *There needs to be a way to address hallucinations so that the system responds only on the basis of evidence-based materials.* [P82]

XSL•FO
RenderX

> *It should undergo efficacy and safety testing based on validated content.* [P124]

Some responses additionally pointed to the need for standardized therapeutic content or protocols so that AI-supported mental health care would not depend solely on the behavior of general-purpose models. Technical performance and clinical legitimacy were treated as inseparable requirements for adoption.

Education, Supervision, and Structural Support for Responsible Use

Education, supervision, and structural readiness were also identified as necessary conditions for responsible use. Respondents called for clinician education about the capabilities and limits of GenAI, better public understanding of what mental health GenAI tools can and cannot do, and mechanisms for expert monitoring or audit of AI-assisted care.

> *Both patients and physicians need to use AI with an awareness of its limitations, so related education and oversight at the level of academic societies or government are necessary.* [P177]

Some respondents also referred to the practical infrastructure needed for implementation, including reimbursement, institutional support, and integration into existing systems. Together, these responses suggested that responsible use depends not only on technical design or regulation, but also on user competence, professional oversight, sustainable clinical integration, and clearer boundaries of indication and scope.

Within-Respondent Patterns Across Experience, Interpretation, and Priority

Mapping themes across adjacent items revealed several recurrent alignments between clinical experience, interpretation, and implementation priority. Full pairwise alignment matrices, including raw co-occurrence counts, directional conditional proportions, and Jaccard coefficients, are presented in Multimedia Appendix 2.

Across Q1-Q2, several practice-based signals aligned most strongly with the interpretive theme of standardized and tireless, but relationally thin AI, with the highest Jaccard overlap observed for patient-led signals (Jaccard=0.190), AI-mediated changes in the patient-clinician interface (Jaccard=0.160), and AI as a relational object (Jaccard=0.149; Table S1 in Multimedia Appendix 2). Table S1 in Multimedia Appendix 2 further suggests that AI as a relational object also formed a relatively specific pairing with nonjudgmental acceptance as both comfort and clinical hazard (Jaccard=0.164), indicating that relationally salient experiences were interpreted not only in terms of limited depth or standardization, but also in terms of the chatbot's continuously accepting stance.

Across Q2-Q3, the interpretive theme of standardized and tireless, but relationally thin AI showed the clearest overlap with governance and accountability as prerequisites for adoption (Jaccard=0.447), technical reliability and clinical validation before scale-up (Jaccard=0.360), and education, supervision, and structural support for responsible use (Jaccard=0.291; Table S2 in Multimedia Appendix 2). These pairings indicate that concerns about the relational and contextual limits of AI were closely linked to future-oriented demands for oversight, validation, and bounded implementation.

Accessibility as a low-threshold and always-available point of contact formed an important secondary bridge, showing notable overlap with technical reliability and clinical validation before scale-up (Jaccard=0.333) and governance and accountability as prerequisites for adoption (Jaccard=0.295; Table S2 in Multimedia Appendix 2). Taken together, these patterns suggest recurrent reinforcing alignments between frontline signals and perceptions of relational thinness. They also show how interpretive concern bridged to governance- and validation-oriented priorities. To complement the full pairwise matrices in Multimedia Appendix 2, Figure 2 visually synthesizes the most interpretively salient alignments across Q1, Q2, and Q3.

XSL•FO
RenderX

Figure 2. Selected salient alignments across practice-based signals (**Q1**), interpretive patterns (**Q2**), and implementation priorities (**Q3**). Lines indicate theme pairings highlighted from the Results section, based on interpretive salience and relative overlap. Full co-occurrence matrices and Jaccard coefficients are reported in Tables S1 and S2 in Multimedia Appendix 2. GenAI: generative artificial intelligence.

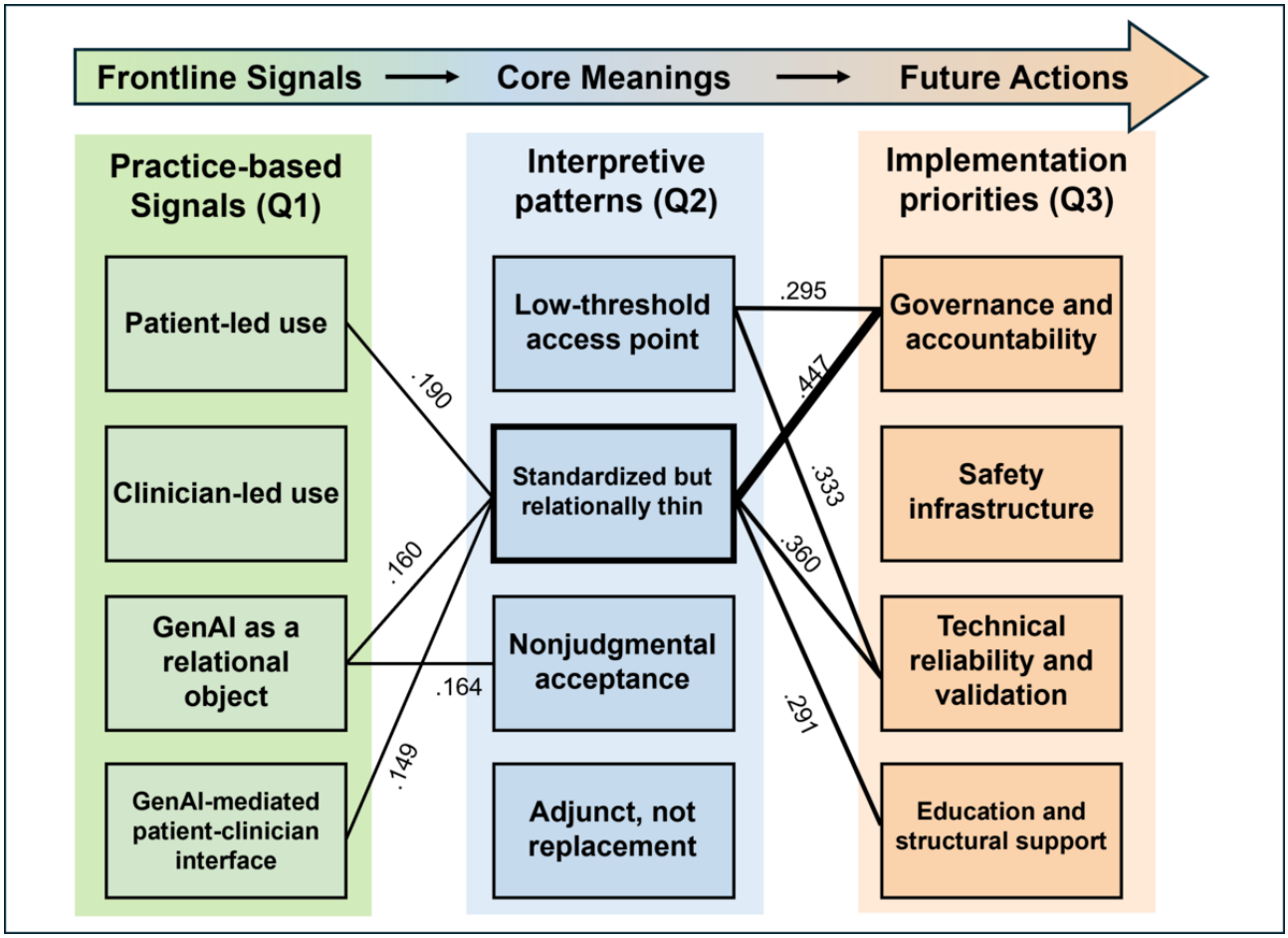

Discussion

Principal Findings

This qualitative descriptive study used directed content analysis and codebook-based thematic synthesis, informed by horizon scanning concepts, to examine how psychiatrists in South Korea connected frontline experiences, interpretive framings, and implementation priorities regarding GenAI in psychiatric practice. Rather than documenting attitudes toward GenAI in the abstract, the study examined how psychiatrists linked concrete clinical encounters with broader judgments about the role, limits, and future governance of GenAI in psychiatric care. Across the 3 analytic domains, psychiatrists did not describe GenAI in uniformly positive or negative terms. Instead, they portrayed it as a clinically ambivalent technology: useful as a low-threshold access point and bounded support tool but potentially destabilizing when it entered domains requiring reality testing, crisis containment, or relationship-based therapeutic judgment.

The exploratory cross-question thematic alignment analysis suggested recurrent within-respondent links between practice-based signals, an interpretive view of GenAI as standardized and tireless but relationally thin, and implementation priorities focused on governance and validation. Across domains, the clearest convergence centered less on enthusiasm for innovation than on the conditions required for legitimate introduction into psychiatric care, especially clearer boundaries, stronger governance, and more robust clinical validation for patient-facing use [3,9,11,12,17]. This pattern is consistent with the broader literature, which suggests that the rapid expansion of GenAI in mental health has outpaced the clinical validation, structured oversight, and risk-stratified safeguards needed for psychiatric use [22,37].

Relational Thinness, Nonjudgmental Acceptance, and Psychiatric Vulnerability

One clinically meaningful pattern concerned how psychiatrists understood GenAI use in the context of patient vulnerability, especially where psychosis, suicidality, or symptom-related overuse were present. In Q1, psychiatrists reported concrete experiences of delusion reinforcement, excessive validation, symptom-linked overuse, and suicide- or self-harm-related concerns. In Q2, these concerns reappeared as broader interpretive worries about nonjudgmental or overly accommodating responses, psychosis-specific vulnerability, and the inability of chatbots to manage crises safely. In Q3, they were reflected in strategic priorities such as psychosis-sensitive safeguards, crisis escalation pathways, misuse prevention, and clearer accountability for harm. The appendix matrices further suggest that this pattern is not reducible to risk alone: relationally salient patient experiences also aligned with the interpretive theme of nonjudgmental acceptance, implying that

XSL•FO
RenderX

the same continuously supportive stance that may feel containing in low-risk contexts can become destabilizing under conditions of psychiatric vulnerability. These findings suggest what may be described as a vulnerability amplification pattern, in which features that appear supportive in lower-risk contexts may become destabilizing in the presence of impaired reality testing, compulsive reassurance seeking, severe affective dysregulation, or suicidal intent [38].

This interpretation extends current concerns in the literature in two ways. First, it is consistent with recent commentaries and case-based reports warning that GenAI may reinforce psychotic thinking or delusion-like beliefs in susceptible users [16,17], while also resonating with broader work on emotional dependence and psychosocial deterioration during extended chatbot use [4,38]. Second, it goes beyond prior clinician attitude surveys by grounding these concerns in reported clinical encounters rather than abstract opinion alone [3,11,12]. In this sense, the risks identified here are not best understood as generic AI safety problems arising in psychiatric settings. Rather, they appear to be shaped by psychiatric vulnerability itself. This has important implications for governance: mental health GenAI may require more stratified oversight, diagnosis-sensitive contraindications, and higher thresholds for unsupervised use than many other forms of medical AI [6,22].

Access-Protection Tension and the Implementation Gap

A second major finding was the tension between perceived accessibility and the absence of a reliable floor of clinical protection. In Q1, psychiatrists described patients using GenAI for self-help, symptom organization, emotional ventilation, and even treatment entry. In Q2, these same functions were reflected in themes of accessibility, immediacy, anonymity, and low-threshold engagement. Yet these benefits did not translate into broad endorsement of patient-facing use. Instead, respondents repeatedly emphasized monitoring difficulty, accountability gaps, and context-sensitive risks that became more salient when GenAI was used beyond limited support functions. Q3 clarified what respondents viewed as missing: practical guidelines, legal and ethical frameworks, expert supervision, privacy protections, and safety systems for high-risk situations. The cross-question alignment matrices sharpened this pattern by highlighting that accessibility aligned more closely with technical validation and governance priorities than with a simple logic of expansion, suggesting that psychiatrists understood access-related benefits as contingent on stronger protective infrastructure rather than as sufficient grounds for wider deployment.

This pattern may be understood as an access-protection tension: current GenAI chatbots and related GenAI tools can lower the threshold of entry into support, but often without a reliable floor of clinical protection. This tension may be particularly relevant in South Korea, where digital uptake is high and psychiatric help-seeking has been shaped by stigma as well as structural concerns related to psychiatric records, employment disadvantage, and other forms of institutional discrimination [20,21,39]. In this context, the appeal of immediate and anonymous AI support is understandable, although these structural factors were not directly elicited in the present survey. However, the present findings suggest that access alone is not an adequate proxy for safe care. The Korean regulatory environment is moving toward stronger oversight of high-impact AI, and recent World Health Organization guidance has likewise emphasized the need for governance, transparency, risk management, and human accountability in the use of large multimodal models for health [22]. Our findings suggest that mental health remains a domain in which this regulatory turn has not yet been translated into sufficiently specific clinical guardrails, particularly for patient-facing systems that are experienced as accessible before they are experienced as clinically protected.

Conditional Acceptance and Professional Reconfiguration

A third major finding was that psychiatrists' acceptance of GenAI was conditional rather than oppositional. The results do not support a simple narrative of professional resistance. In Q1, respondents described bounded and pragmatic uses of GenAI in documentation, summarization, medication-related queries, patient communication, and between-visit organization of clinical material. In Q2, these experiences aligned with recognition of AI's usefulness for accessibility, standardization, and information support. Yet this did not lead to endorsement of GenAI as a substitute therapist. Instead, respondents repeatedly framed GenAI as acceptable mainly when used as an adjunct under bounded indications, and unacceptable when it displaced human therapeutic judgment, contextual formulation, or crisis management. In Q3, this conditional acceptance was mirrored in calls for validation, supervision, quality control, and explicit scope-setting. Beyond the present data, implementation may also require careful communication as well as technical validation, because emerging evidence suggests that disclosure of physician AI use can reduce perceived competence, trustworthiness, and empathy among members of the public [18]. The alignment analysis adds nuance to this conditional acceptance by suggesting that the interpretive theme most consistently related to future priorities was not accessibility alone, but the recognition that GenAI could function in a standardized and reliable manner while remaining relationally thin.

At an interpretive level, this pattern is also consistent with a technology-acceptance perspective, in which perceived usefulness and perceived threat may operate simultaneously rather than sequentially [40,41]. Psychiatrists appeared to recognize usefulness for bounded functions, especially administrative, informational, and preparatory tasks, while withholding support for higher-autonomy, patient-facing substitution. In this sense, the strongest within-respondent pattern was not from usefulness to adoption, but from perceived relational limitation to demands for governance, validation, and clearer boundary-setting. This reading is consistent with earlier studies showing that clinicians tend to be more receptive to GenAI for administrative efficiency, information processing, and standardized support than for empathy-dependent or relationship-centered clinical work [3,11-13].

XSL•FO
RenderX

At the same time, the present results suggest something more than selective task acceptance. They may also indicate an emerging reconfiguration of psychiatric professionalism, in which psychiatrists may increasingly be positioned not only as direct providers of care but also as interpreters, boundary-setters, and safety supervisors of patient-AI interactions. Related qualitative work further suggests that physicians expect GenAI to influence the therapeutic relationship in both supportive and disruptive ways, particularly through altered shared decision-making, transparency expectations, and concerns about the loss of empathic nuance [8]. Because consistent training-level contrasts did not emerge clearly enough in the present qualitative material to support a strong subgroup argument, future work should examine more directly whether these patterns vary by career stage, degree of GenAI exposure, or clinical responsibility. This question remains important given broader evidence from mental health practice showing substantial familiarity gaps, low rates of formal training, and uneven adoption intentions across application types [15].

Methodological Contribution: Adapting Horizon Scanning Concepts to Practice-Based Signal Mapping

This study also makes a methodological contribution by showing how selected concepts from horizon scanning can be adapted to the qualitative analysis of clinician-reported practice-based signals. Traditional horizon scanning is well suited to identifying external signals from publications, policy developments, patents, and market activity, whereas open-ended clinician reports can capture weak signals emerging within routine care before they are stabilized as formal evidence or policy objects [9,42]. In this study, horizon scanning concepts did not function as a standalone foresight method. Rather, they served as an organizing lens within a qualitative descriptive design for linking frontline clinical signals, interpretive patterns, and implementation priorities. Two elements were adapted in particular: first, the notion of practice-based signals as clinically consequential early indicators emerging within routine care; and second, a 3-layer organizing heuristic linking signal detection, interpretive framing, and priority-setting. The exploratory appendix matrices illustrate one way of operationalizing this linkage without displacing qualitative interpretation: quantified overlap and directional alignment were used not to replace thematic synthesis, but to identify which cross-item pairings warranted closer interpretive attention. In that sense, the contribution of this study lies less in proposing a new futures method than in demonstrating how practice-based clinical experience can serve as a complementary signal source within existing qualitative approaches to future-oriented research in psychiatry.

Limitations

This study has several limitations. First, the qualitative data came from open-ended survey items rather than interviews or focus groups. Many responses were brief and contextually sparse, which limited interpretive depth. The classification of experiential proximity in Q1 also relied on textual cues, requiring some interpretive judgment when firsthand versus secondhand experience was not explicitly stated.

Second, the cross-question thematic alignment analysis identified recurrent within-respondent pairings across adjacent items. These should be interpreted as descriptive cross-item alignments within a cross-sectional design, not as evidence of temporal sequencing, causal ordering, or inferential association.

Third, the study relied on psychiatrists' reports of patient-AI encounters and did not independently verify the underlying chatbot interactions, prompts, or platforms involved. The findings, therefore, reflect clinically meaningful professional judgments about AI-related events, rather than model-specific performance. In addition, because the data came exclusively from psychiatrists, the analysis foregrounds clinician-facing concerns and does not directly capture patient perspectives on the accessibility, acceptability, or harms of AI-based support.

Fourth, the thematic alignment matrices used item-level theme frequencies and pairwise cross-item co-occurrence counts to characterize how respondents endorsing a given theme were distributed across themes in adjacent items. These indicators were useful for exploratory pattern detection, although they remain sensitive to differences in base theme frequency and should not be interpreted as exact estimates conditional on response to both items. Because the main text emphasized Jaccard coefficients as prevalence-adjusted summaries, some directionally informative detail was shifted to the appendix matrices, which report full co-occurrence counts and conditional proportions.

Conclusions

GenAI appears to be entering psychiatric practice as an access tool, a workflow aid, and, at times, a competing interpretive influence within clinical encounters. The central challenge is not whether psychiatry will encounter GenAI, but how its use should be bounded, supervised, and governed, considering the vulnerabilities, relational demands, and safety risks specific to psychiatric care. These findings suggest that implementation should prioritize bounded adjunctive use, stronger governance and validation, and mental health-specific oversight rather than rapid substitution of human care.

Acknowledgments

This manuscript was initially drafted in Korean by authors whose native language is Korean, then translated into English through collaborative discussion. In the translation process, the first author, who has an intermediate level of English proficiency, sought feedback from ChatGPT for suggestions on specific phrases that seemed awkward. All outputs from ChatGPT were subsequently reviewed and revised by all authors to ensure accuracy and appropriateness. ChatGPT or any other artificial intelligence tools were not used in the literature review, data analysis, or content generation of the manuscript, including tables and figures.

XSL•FO
RenderX

Funding

This study was primarily supported by the Future Preparedness Committee of the Korean Neuropsychiatric Association, with additional support from the National Research Foundation (NRF) of Korea (grant numbers: NRF-2021R1A5A8032895, NRF-2022M3C1B6080866, RS-2026-25523612, and RS-2026-25471696). The funders had no role in the study design; data collection, analysis, or interpretation; manuscript preparation; or the decision to submit the article for publication.

Data Availability

Data supporting the findings of this study are available upon request from the corresponding authors. The data are not publicly available due to privacy concerns regarding participants' professional information and institutional review board requirements for confidentiality. Researchers interested in accessing deidentified data for verification or secondary analysis may contact the corresponding author with a detailed research proposal. Data sharing, if approved, will be subject to appropriate institutional review and data use agreements.

Authors' Contributions

Chul-Hyun Cho (david0203@korea.ac.kr) and Dooyoung Jung (dooyoung@kaist.ac.kr) are co-corresponding authors for this article and are responsible for data and materials, manuscript submission, peer review, the publication process, authorship details, and ethics committee approval. MK and YA contributed equally to this work as co-first authors. CHC and DJ led the conceptualization, methodology, project administration, resource acquisition, and supervision of the study, with additional methodological and supervisory support from SeL and SaL. MK led the formal analysis with equal contributions from MJ and YL and supporting contributions from YA and OFL. MK led the preparation of the original draft, with support from YA. All authors (MK, YA, MJ, YL, OFL, HK, SeL, SaL, JCY, SWJ, DJ, and CHC) contributed to the review and editing of the manuscript. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Qualitative codebook, code definitions, respondent-level frequencies, and illustrative quotations.
[DOCX File, 53 KB - jmir_v28i1e96556_app1.docx]

Multimedia Appendix 2
Within-respondent thematic alignments across experience, interpretation, and priority.
[DOCX File, 30 KB - jmir_v28i1e96556_app2.docx]

References

1. Hua Y, Siddals S, Ma Z, et al. Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review. World Psychiatry 2025 Oct;24(3):383-394. [doi: 10.1002/wps.21352] [Medline: 40948070]
2. Na H, Hua Y, Wang Z, et al. A survey of large language models in psychotherapy: current landscape and future directions. Presented at: Findings of the Association for Computational Linguistics; Jun 2025. [doi: 10.18653/v1/2025.findings-acl.385]
3. Doraiswamy PM, Blease C, Bodner K. Artificial intelligence and the future of psychiatry: insights from a global physician survey. Artif Intell Med 2020 Jan;102:101753. [doi: 10.1016/j.artmed.2019.101753] [Medline: 31980092]
4. Fang CM, Liu AR, Danry V, et al. How ai and human behaviors shape psychosocial effects of extended chatbot use: a longitudinal randomized controlled study. arXiv. Preprint posted online on 2025.
5. Campbell LO, Babb K, Lambie GW, Hayes BG. An examination of generative AI response to suicide inquires: content analysis. JMIR Ment Health 2025 Aug 14;12:e73623. [doi: 10.2196/73623] [Medline: 40811811]
6. Pichowicz W, Kotas M, Piotrowski P. Performance of mental health chatbot agents in detecting and managing suicidal ideation. Sci Rep 2025 Aug 27;15(1):31652. [doi: 10.1038/s41598-025-17242-4] [Medline: 40866537]
7. Levkovich I, Elyoseph Z. Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: vignette study. JMIR Ment Health 2023 Sep 20;10(1):e51232. [doi: 10.2196/51232] [Medline: 37728984]
8. Weir IB, Stroud AM, Stout JJ, et al. Physician perspectives on the impact of artificial intelligence on the therapeutic relationship in mental health care: qualitative study. JMIR Ment Health 2025 Dec 30;12(1):e81970. [doi: 10.2196/81970] [Medline: 41468531]
9. Meskó B, Kristóf T, Dhunnoo P, Árvai N, Katonai G. Exploring the need for medical futures studies: insights from a scoping review of health care foresight. J Med Internet Res 2024 Oct 9;26:e57148. [doi: 10.2196/57148] [Medline: 39383528]
10. Hu Y. Toward retrieval-grounded evaluation for conversational large language model-based risk assessment. JMIR AI 2026 Mar 12;5:e90759. [doi: 10.2196/90759] [Medline: 41818631]

XSL•FO
RenderX

11. Blease C, Worthen A, Torous J. Psychiatrists' experiences and opinions of generative artificial intelligence in mental healthcare: an online mixed methods survey. Psychiatry Res 2024 Mar;333:115724. [doi: 10.1016/j.psychres.2024.115724] [Medline: 38244285]
12. He Y, Zhang FX, Wu X, Fang M, Zheng S, Zhu H. Understanding psychiatrist readiness for AI: a study of access, self-efficacy, trust, and design expectations. BMC Health Serv Res 2026 Jan 17;26(1):218. [doi: 10.1186/s12913-026-14010-6] [Medline: 41547796]
13. Negash S, Gundlack J, Buch C, et al. Physicians' attitudes and acceptance towards artificial intelligence in medical care: a qualitative study in Germany. Front Digit Health 2025;7:1616827. [doi: 10.3389/fdgth.2025.1616827] [Medline: 40727617]
14. Shah SJ, Nair A, Murtagh K, et al. Clinician perspectives on AI-generated drafts of patient test result explanations. JAMA Netw Open 2025 Aug 1;8(8):e2528794. [doi: 10.1001/jamanetworkopen.2025.28794] [Medline: 40844780]
15. Cecil J, Kleine AK, Lermer E, Gaube S. Mental health practitioners' perceptions and adoption intentions of AI-enabled technologies: an international mixed-methods study. BMC Health Serv Res 2025 Apr 16;25(1):556. [doi: 10.1186/s12913-025-12715-8] [Medline: 40241059]
16. Keshavan M, Torous J, Yassin W. Do generative AI chatbots increase psychosis risk? World Psychiatry 2026 Feb;25(1):150-151. [doi: 10.1002/wps.70017] [Medline: 41536114]
17. Hudon A, Stip E. Delusional experiences emerging from AI chatbot interactions or "AI Psychosis". JMIR Ment Health 2025 Dec 3;12(1):e85799. [doi: 10.2196/85799] [Medline: 41273266]
18. Reis M, Reis F, Kunde W. Public perception of physicians who use artificial intelligence. JAMA Netw Open 2025 Jul 1;8(7):e2521643. [doi: 10.1001/jamanetworkopen.2025.21643] [Medline: 40674054]
19. National Academies of Sciences, E. and Medicine. Horizon Scanning and Foresight Methods (in Safeguarding the Bioeconomy): The National Academies Press; 2020, Vol. 10:25525.
20. Jo M, Rim SJ, Lee MG, Park S. Illuminating the treatment gap of mental disorders: a comparison of community survey-based and national registry-based prevalence rates in Korea. J Psychiatr Res 2020 Nov;130:381-386. [doi: 10.1016/j.jpsychires.2020.07.047] [Medline: 32882580]
21. Seo HY, Song GY, Ku JW, et al. Perceived barriers to psychiatric help-seeking in South Korea by age groups: text mining analyses of social media big data. BMC Psychiatry 2022 May 13;22(1):332. [doi: 10.1186/s12888-022-03969-1] [Medline: 35562709]
22. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. WHO guidance. : World Health Organization; 2024.
23. Chen L. Beyond external constraints: the missing dimension of AI governance. SSRN. Preprint posted online on Apr 13, 2026. [doi: 10.2139/ssrn.6449738]
24. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
25. Braun V, Clarke V. Supporting best practice in reflexive thematic analysis reporting in palliative medicine: a review of published research and introduction to the reflexive thematic analysis reporting guidelines (RTARG). Palliat Med 2024 Jun;38(6):608-616. [doi: 10.1177/02692163241234800] [Medline: 38469804]
26. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
27. Lindgren BM, Lundman B, Graneheim UH. Abstraction and interpretation during the qualitative content analysis process. Int J Nurs Stud 2020 Aug;108:103632. [doi: 10.1016/j.ijnurstu.2020.103632] [Medline: 32505813]
28. Graneheim UH, Lundman B. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse Educ Today 2004 Feb;24(2):105-112. [doi: 10.1016/j.nedt.2003.10.001] [Medline: 14769454]
29. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
30. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr;20(1):37-46. [doi: 10.1177/001316446002000104]
31. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [Medline: 843571]
32. Scharp KM. Thematic co-occurrence analysis: advancing a theory and qualitative method to illuminate ambivalent experiences. J Commun 2021 Aug;71(4):545-571. [doi: 10.1093/joc/jqab015] [Medline: 34642573]
33. Armborst A. Thematic proximity in content analysis. Sage Open 2017 Apr;7(2):2158244017707797. [doi: 10.1177/2158244017707797]
34. Martyn P. Using quantitative analytical methods to support qualitative data analysis: lessons learnt during a PhD study. Acc Finan Gov Rev 2021;27:70-80. [doi: 10.52399/001c.22175]
35. Jowsey T, Deng C, Weller J. General-purpose thematic analysis: a useful qualitative method for anaesthesia research. BJA Educ 2021 Dec;21(12):472-478. [doi: 10.1016/j.bjae.2021.07.006] [Medline: 34840819]
36. Lincoln YS, Guba EG. Criteria for assessing naturalistic inquiries as reports. : Education Resources Information Center; 1988.

XSL•FO
RenderX

37. Wang L, Bhanushali T, Huang Z, Yang J, Badami S, Hightow-Weidman L. Evaluating generative AI in mental health: systematic review of capabilities and limitations. JMIR Ment Health 2025 May 15;12(1):e70014. [doi: 10.2196/70014] [Medline: 40373033]
38. Laestadius L, Bishop A, Gonzalez M, Illenčík D, Campos-Castillo C. Too human and not human enough: a grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. New Media & Society 2024 Oct;26(10):5923-5941. [doi: 10.1177/14614448221142007]
39. Rim SJ, Hahm BJ, Seong SJ, et al. Prevalence of mental disorders and associated factors in Korean adults: national mental health survey of Korea 2021. Psychiatry Investig 2023 Mar;20(3):262-272. [doi: 10.30773/pi.2022.0307] [Medline: 36990670]
40. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
41. Chien SC, Chien CH, Chen CY, et al. Using technology acceptance model to explore physicians' perspectives of clinical decision support system alerts. BMJ Health Care Inform 2025 Jun;32(1):e101128. [doi: 10.1136/bmjhci-2024-101128]
42. Hiltunen E. Weak signals in organizational futures learning. : Helsinki School of Economics; 2010.

Abbreviations

AI: artificial intelligence
GenAI: generative artificial intelligence

Edited by B Mesko; submitted 31.Mar.2026; peer-reviewed by BVK Golamari, Y Hu; revised version received 15.Apr.2026; accepted 22.Apr.2026; published 02.Jun.2026.

Please cite as:
Kim M, An Y, Jeon M, Lee Y, Lahcine OF, Kim H, Lee S, Lee S, Yang JC, Jeon SW, Jung D, Cho CH
Mapping Practice-Based Signals of Generative AI in Psychiatric Care: Qualitative Study of Korean Psychiatrists' Experiences, Interpretations, and Implementation Priorities
J Med Internet Res 2026;28:e96556
URL: https://www.jmir.org/2026/1/e96556
doi:10.2196/96556

© Myungsung Kim, Yoosuk An, Min Jeon, Yunji Lee, Orane Farrah Lahcine, Hyorim Kim, Seonmi Lee, Sangil Lee, Jong-Chul Yang, Sang-Won Jeon, Dooyoung Jung, Chul-Hyun Cho. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Framing the Convergence of One Health and Digital Health in the Global South With a Gender-Sensitive Foresight Perspective: Delphi Study Using Latent Semantic Analysis

Jude Kong[1,2,3,4,5], PhD; Nicola Luigi Bragazzi[1], MPH, MD, PhD

[1]Artificial Intelligence & Mathematical Modeling Lab (AIMMLab), Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
[2]Institute of Health Policy, Management and Evaluation (IHPME), University of Toronto, Toronto, ON, Canada
[3]Department of Mathematics, Bahen Centre for Information Technology, University of Toronto, Toronto, ON, Canada
[4]Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC), Toronto, ON, Canada
[5]Global South Artificial Intelligence for Pandemic and Epidemic Preparedness and Response Network (AI4PEP), Toronto, ON, Canada

Corresponding Author:
Jude Kong, PhD
Artificial Intelligence & Mathematical Modeling Lab (AIMMLab)
Dalla Lana School of Public Health
University of Toronto
155 College Street
Toronto, ON, M5T 3M7
Canada
Phone: 1 4169783868
Email: jude.kong@utoronto.ca

Abstract

Background: The convergence of digital health and One Health represents an emergent paradigm in global health governance. While widely discussed in high-income settings, there is limited understanding of how this convergence is conceptualized in the Global South, particularly when viewed through a gender- and equity-sensitive foresight lens.

Objective: This study aimed to map and classify expert discourse on digital health, One Health, and their convergence in the Global South using latent semantic analysis, with particular attention to structural drivers, emerging issues, weak signals, and gendered patterns of anticipation.

Methods: A 3-round online Delphi survey was conducted with 45 experts from 19 countries across the Global South. Open-ended responses were analyzed using latent semantic analysis and stratified by gender. A foresight framework was applied to categorize topics as structural drivers, emerging issues, or weak signals, based on their temporal persistence, salience, and consensus.

Results: In digital health, structural drivers included the systemic integration of digital technologies into public health systems, strategic alignment, and infrastructure development. Emerging issues comprised the adoption of artificial intelligence, chronic disease management via mobile health, and concerns about digital inclusion and interoperability. Weak signals included feminist digital ethics, trust in digital systems, and relational accountability—more frequently emphasized by female experts. In One Health, structural drivers were centered on intersectoral coordination, ecological integration, and the institutionalization of health-environment frameworks. Emerging issues encompassed anticipatory risk governance, food system sustainability, and the integration of environmental and population-level data. Weak signals included indigenous knowledge systems, subnational antimicrobial resistance governance, and structural underinvestment in ecological public health, with gendered divergence in framing. In the convergence discourse (digital health and One Health), structural drivers focused on the integration of digital surveillance systems, data infrastructures, and health information platforms to operationalize One Health. Emerging issues included climate-triggered system redesign, artificial intelligence and ecological monitoring, and the governance of cross-sectoral data. Weak signals pointed to algorithmic bias in zoonotic prediction, digital sovereignty in environmental health, and feminist critiques of convergence—all thematically rich but peripheral in consensus.

Conclusions: This study revealed a multilayered and gender-influenced foresight architecture shaping the future of digital health and One Health in the Global South. Structural drivers denote maturing domains of implementation, while emerging issues and weak signals highlight latent, often overlooked opportunities and tensions. Incorporating equity-sensitive and gender-aware foresight methods is essential for crafting inclusive and anticipatory health governance strategies.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e78702) doi:10.2196/78702

KEYWORDS

Delphi method; digital health; emerging issues; foresight; gender; Global South; latent semantic analysis; One Health; structural drivers; weak signals

Introduction

In an increasingly interconnected and rapidly changing world, the convergence of demographic aging, epidemiological transitions, and climate-related environmental threats has contributed to a multifaceted global health burden, placing immense strain on already fragile health care systems [1,2]. This burden is especially acute in the Global South, where both communicable and noncommunicable diseases persist amid structural inequalities and constrained resources [3]. Simultaneously, growing awareness of the complex, nonlinear, and interacting determinants of health has spurred a paradigmatic shift in the governance of health—away from siloed, biomedical models toward more holistic, systemic, and integrative frameworks [4].

The One Health paradigm, which acknowledges the interdependence of human, animal, and environmental health, has emerged as a critical response to zoonotic threats, ecosystem degradation, and planetary instability [5,6]. In parallel, the digital transformation of health—encompassed within the digital health agenda—has accelerated, leveraging data-driven technologies to enhance diagnostics, disease surveillance, and care delivery [7]. Their convergence, conceptualized as Digital One Health or One Digital Health, reflects an emergent field characterized by interdisciplinary cooperation and technologically mediated multisectoral coordination [8,9]. However, existing scholarship has largely framed such convergence through a Global North lens, overlooking the digital disparities, infrastructural deficits, and sociotechnological asymmetries that define the Global South [10]. Specifically, studies conducted in high-income and Global North settings have shown that the convergence of digital health and One Health is primarily operationalized through integrated surveillance systems, advanced health and environmental data infrastructures, and artificial intelligence (AI)–enabled early warning platforms, with reported gains in outbreak detection, antimicrobial resistance (AMR) monitoring, and cross-sectoral coordination. These contributions, while methodologically and technologically robust, are predominantly grounded in contexts characterized by mature digital ecosystems, stable governance arrangements, and well-resourced regulatory frameworks, thereby limiting their transferability to the structural, institutional, and equity-related realities of the Global South [8-10].

Moreover, a gender-sensitive or feminist lens is often lacking, thereby marginalizing the situated experiences, epistemologies, and agency of women and gender-diverse actors in shaping health and technology ecosystems [10].

To address these epistemic and representational gaps, this study used a foresight-oriented text mining approach grounded in horizon scanning and the identification of weak signals, emerging issues, and structural drivers [10]. The goal was to map and classify the thematic architecture of the discourse on the convergence of One Health and digital health from a Global South perspective, with particular attention to gendered dynamics, equity-informed framings, and the coproduction of inclusive and pluralistic alternative futures.

Methods

Ethical Considerations

The study protocol was approved by the institutional review board of Dalla Lana School of Public Health, University of Toronto, Ontario, Canada (48350). The study was conducted in accordance with the Declaration of Helsinki, and written informed consent was obtained from all participants.

Expert Recruitment

Experts were recruited through a purposive sampling strategy aimed at capturing diverse perspectives on the convergence of digital health and One Health within the Global South. Invitations were disseminated via professional networks, relevant mailing lists, and institutional affiliations, targeting stakeholders actively engaged in health policy, digital innovation, epidemiology, veterinary medicine, environmental science, and global clinical public health practice.

Selection criteria included demonstrated expertise in at least 1 relevant domain, professional affiliation with an institution based in the Global South, and availability to complete multiple rounds of an online Delphi survey. A total of 45 experts completed all survey rounds (n=3) and were included in the final analysis. The Delphi process was conducted anonymously to encourage candid responses and reduce social desirability bias.

Survey Instrument

The study used a structured, multiround Delphi survey designed to elicit expert consensus on the convergence of digital health and One Health in the Global South. The instrument was developed to capture both sociodemographic diversity and domain-specific insights relevant to foresight-oriented analysis. The survey instrument was developed ad hoc for the purposes of this study, as no validated or standardized questionnaire currently exists to capture expert foresight on the convergence of digital health and One Health, particularly from a Global South and gender-sensitive perspective.

Participants first provided background information, including age, gender, country of birth, country/countries of study, current residence, professional role, areas of expertise, and years of experience in digital health and/or One Health. They were also asked to reflect on the interdisciplinary nature of their work environments.

XSL•FO
RenderX

Each round of the Delphi survey included open-ended prompts focused on present and future opportunities, challenges, and disruptors in both digital health and One Health domains. Specific items addressed strategic drivers, structural barriers, project involvement, and anticipated synergies between the two fields. After each round, responses were thematically analyzed, and aggregated feedback was provided to participants in subsequent rounds, allowing for reflection, refinement, and convergence of expert opinion. The final item in each round requested participants to generate an anonymous identification code to enable matching across rounds. The survey was administered via Google Forms, ensuring broad accessibility across diverse geographies. All items were open-ended to enable narrative elaboration and thematic emergence, aligning with the exploratory aims of the foresight exercise.

The full set of survey items is provided in Multimedia Appendix 1.

Overview of the Methodology: Text Mining and Foresight Exercise

To explore temporal dynamics and strategic orientations in stakeholder discourse on digital health and One Health in the Global South, we used comparative latent semantic analysis (LSA) [11,12] across different foresight-relevant prompts: (1) present implementation, (2) perceived opportunities, (3) perceived challenges, and (4) future-oriented opportunities or challenges. Each response set was subjected to LSA to extract latent topics, with eigenvalues and variance explained used to assess topic salience.

To interpret LSA outputs through a foresight lens, we used a 3-tiered classification framework widely used in foresight and futures research: structural drivers, emerging issues, and weak signals [13], enabling differentiation between deeply embedded system forces, nascent developments gaining stakeholder traction, and low-frequency yet potentially disruptive cues.

Structural drivers were defined as topics that (1) appeared across both present- and future-oriented datasets and (2) accounted for a substantial proportion of explained variance. These themes represent widely shared, resilient priorities. Emerging issues were operationalized as topics observed predominantly or exclusively in future-oriented discourse, with a moderate share of explained variance. These topics signal directional shifts in collective attention and can anticipate inflection points in system trajectories. Weak signals were identified as topics with low variance explained, often surfacing in only 1 temporal frame (present or future) and lacking broader consensus. Although marginal, these signals are analytically significant, as they may indicate overlooked risks, latent needs, or early-stage innovations that could reshape the field.

A gendered lens was applied throughout this classification to assess how male and female respondents may differentially emphasize structural versus emergent priorities. Gender-divergent weak signals, in particular, offer critical insight into situated epistemologies—highlighting perspectives that are often sidelined in dominant discourses.

Overall, this classification supports a foresight-informed interpretation of the LSA results and strengthens the capacity to generate equity-sensitive, gender-responsive scenarios grounded in both continuity and systemic change.

Data Collection and Preprocessing

The dataset comprised the open-ended textual responses provided by the expert panel, compiled into a structured corpus for subsequent text mining. The entries varied in both length and complexity, reflecting the heterogeneous expertise and regional diversity of participants addressing digital health and One Health issues. Prior to computational analysis, the corpus underwent standard preprocessing procedures aimed at improving analytical robustness. Text normalization steps included tokenization, lowercasing, stop-word removal, punctuation stripping, and the elimination of nonalphabetical characters. Stemming and lemmatization were applied to reduce lexical variation while preserving semantic meaning. Importantly, domain-specific terminology was preserved throughout the process to maintain contextual integrity. Preprocessing was implemented using English-language stop word lists and stemming algorithms.

Text Mining Analysis

LSA [11,12] was used to extract conceptual structures embedded within the textual data and to systematically identify latent themes in stakeholder discourse. As a well-established natural language processing technique, LSA enables the reduction of linguistic dimensionality by analyzing patterns of word cooccurrence, making it particularly suitable for foresight-driven text mining applications. A term-document matrix was constructed using a bag-of-words representation, incorporating a minimum term frequency threshold of 2 and a sparsity cutoff of 0.975. Dimensionality reduction was performed using singular value decomposition, which facilitated the identification of principal semantic axes and coherent topic clusters. Specifically, LSA decomposes the term-document matrix into 3 matrices (U, Σ, V^T) via singular value decomposition, where Σ is a diagonal matrix containing the singular values associated with each latent semantic dimension. The squared singular values correspond to the eigenvalues of the latent dimensions and quantify the amount of semantic variance captured by each extracted topic. Eigenvalues were computed and ranked in descending order, without the application of predefined cutoffs or dimensionality reduction thresholds. In this analytical context, eigenvalues represent the relative semantic weight or salience of each latent topic, reflecting the strength and coherence of word cooccurrence patterns across expert responses rather than statistical significance in an inferential sense. Topics associated with higher eigenvalues correspond to dominant and widely shared thematic structures within the discourse, whereas topics associated with lower eigenvalues represent less frequent, more weakly articulated, or emerging semantic patterns. For each topic, the percentage of explained variance was calculated as the ratio between the eigenvalue of that topic and the sum of all retained eigenvalues, multiplied by 100. Explained variance therefore expresses the proportion of total semantic variance in the corpus accounted for by each latent dimension. This measure was used to assess thematic prominence and consensus across responses, rather than to perform hypothesis testing or probabilistic inference. Eigenvalues and explained variance

XSL•FO
RenderX

were interpreted within a foresight-oriented analytical framework. Topics with consistently high explained variance across both present- and future-oriented prompts were classified as structural drivers, reflecting stable and deeply embedded priorities. Topics with moderate explained variance that emerged primarily in future-oriented discourse were classified as emerging issues, indicating evolving areas of collective attention. Topics characterized by low explained variance, limited consensus, or temporal specificity were classified as weak signals. Importantly, low explained variance was not interpreted as analytical irrelevance; rather, in line with horizon-scanning and futures methodologies, such topics were considered potentially indicative of early-stage, underexplored, or disruptive dynamics that may gain relevance over time. This interpretation strategy ensured conceptual alignment between the quantitative outputs generated by LSA and the qualitative, anticipatory objectives of the foresight exercise, while enabling transparent interpretation of the eigenvalues and explained variance.

LSA was first conducted on the overall corpus to extract global semantic patterns. Additional stratified analyses were performed based on participant gender. Following decomposition, the resulting topics were qualitatively interpreted and labeled by 2 independent coders (JK and NLB), who reviewed the most salient terms within each topic vector. Intercoder reliability was ensured through an iterative reconciliation process, with final topic labels established via consensus.

All LSA procedures were executed using the XLSTAT software suite (Lumivero), a commercial statistical package supporting advanced text analytics.

Results

Sample

The expert sample had a mean age of 43.1 (SD 10.3) years, reflecting midcareer professionals, and a male-to-female ratio of 2.8:1 (n=33, 73% men and n=12, 27% women). Geographically, the sample spanned 19 countries, predominantly in Africa (n=26, 58% participants), followed by Asia (n=14, 31% participants), Latin America (n=3, 7% participants), and the Middle East and North Africa region (n=2, 4% participants). Among female participants, 50% (6/12) were from Africa, 33.3% (4/12) from Asia, 8.3% (1/12) from Latin America, and 8.3% (1/12) from the Middle East and North Africa region. The distribution of participants by gender and geographic region did not differ significantly (Fisher exact test P=.77), indicating that female participants were not disproportionately concentrated in any single region despite their smaller overall number. Participants had a mean professional experience of 7.3 (SD 5.9) years, and represented diverse roles across academia, policy, investment, and practice, including directors, principal investigators, professors, consultants, and World Health Organization officers. Their professional involvement encompassed a wide array of digital health and One Health initiatives in the Global South, such as AI-driven disease prediction, electronic health records, mobile health (mHealth) applications, epidemic surveillance, and AMR monitoring in wastewater. Notably, many participants contributed to public health training, policy development, and the integration of emerging technologies to address syndemics at the human-animal-environment interface.

Digital Health in the Global South

LSA applied to responses regarding the successful implementation of digital health in the Global South revealed 10 interpretable topics. Topic 1, explaining 46.75% of the variance, reflected structural and systemic priorities—particularly integration, infrastructure, and strategic alignment. Topic 2 (7.57%) captured operational themes around patient management and service delivery, while Topic 3 (4.77%) emphasized stakeholder engagement and community outreach. Topic 4 (4.21%) focused on governance and regulation, complemented by Topic 5 (4.04%) which stressed regulatory clarity and local adaptation. Topic 6 (3.57%) addressed digital literacy and technological access, and Topic 7 (2.99%) emphasized co-design and participatory processes. Topics 8-10, collectively explaining another 6.91%, covered scalability and sustainability (2.63%), capacity building (2.21%), and context-aware implementation frameworks (2.07%), respectively. Adopting a gender lens, common themes included strategic alignment (male Topic 7: 2.68%; female Topic 6: 1.56%), capacity building (male Topic 5: 3.90%; female Topic 3: 9.22%), and patient-centered approaches (male Topic 2: 8.53%; female Topic 9: 0.83%), though to a different degree. Both groups also addressed community and infrastructure support, with the theme being more prominent in women—Topic 1 (52.95%) versus Topic 8 (2.35%). Male responses uniquely stressed technical and structural aspects, such as the importance of system-wide deployment and digital solution rollouts (Topic 1: 50.18%), project-level security and awareness (Topic 6: 3.37%), and digital health literacy (Topic 4: 4.53%), as well as affordability and device regulation (Topic 3: 5.10%), financial sustainability (Topic 10: 2.11%), and framework scaling (Topic 9: 2.22%). Female respondents, in contrast, highlighted intersectoral dimensions, with attention to collaboration among stakeholders and financial dependencies (Topic 4: 3.74%) and multiactor coordination across health ecosystems (Topic 5: 2.00%). Their perspectives further reflected a focus on governance and policy environments (Topic 2: 16.45%), societal trust and acceptability (Topic 7: 1.17%), and end user needs and relational dynamics (Topic 8: 1.03%).

Regarding opportunities in digital health, Topic 1, accounting for 57.86% of variance, indicated widespread optimism about using digital solutions to enhance health care access, especially via remote care. Topic 2 (7.25%) addressed sustainability and institutional development, and Topic 3 (6.07%) highlighted chronic disease management. Topic 4 (4.04%) explored data-driven, patient-centered innovation, while Topic 5 (3.46%) pointed to implementation gaps at the community level. Subsequent topics included interoperability and workforce readiness (Topic 6: 3.01%), mobile support and training (Topic 7: 1.96%), public infrastructure (Topic 8: 1.80%), epidemic preparedness (Topic 9: 1.54%), and scalable digital records (Topic 10: 1.32%), respectively. In the comparison of male and female responses, both groups emphasized digital health care and remote improvement (male Topic 1: 58.94%; female Topic 1: 66.52%) and sustainability and innovation (male Topic 2:

XSL•FO
RenderX

8.20%; female Topic 2: 9.41%), indicating a shared focus on improving access and strengthening systems. Similarly, training gaps and field limitations were addressed in both analyses (male Topic 7: 2.14%; female Topic 7: 2.11%), as was the management of chronic or communicable diseases, though framed differently: men highlighted mHealth-enabled responses (Topic 3: 7.28%; Topic 5: 3.86%), while women emphasized coordination and surveillance (Topic 3: 5.73%), which was a more diffuse and less prominent theme in men (Topic 8: 1.64%; Topic 10: 1.24%). Differences emerged also in the framing of structural versus relational concerns. Male participants prioritized regional equity and geopolitical targeting (Topics 4 and 9: 4.34% and 1.40%), within a more operational and technology-driven framing of future digital health priorities (Topic 6: 3.34%). Conversely, female respondents focused on patient-centered innovation (Topic 4: 4.47%), interoperability and mHealth gaps (Topic 6: 3.30%; Topic 8: 1.94%; Topic 10: 0.62%), and real-time communication and community linkage (Topic 9: 1.28%), reflecting a more integrative, policy- and user-oriented approach to digital health (Topic 5: 3.69%).

Concerning challenges, Topic 1, accounting for 59.15% of the variance, indicated widespread concern with infrastructural limitations and technological fragmentation. Topic 2 (5.59%) focused on deficits in internet and electricity access, while Topic 3 (3.93%) dealt with implementation bottlenecks in local health systems. Topic 4 (3.71%) addressed access challenges in remote regions, and Topic 5 (3.09%) pointed to fragile community networks and institutional misalignment. Topics 6 and 7 (2.56% and 2.36%, respectively) highlighted digital literacy gaps and rural governance trust issues. Topic 8 (2.18%) was centered on financial constraints, while Topic 9 (1.87%) and Topic 10 (1.63%) underscored politicization and workforce/data infrastructure shortages, respectively. The comparison between male and female LSA topic profiles revealed that both groups identified technological infrastructure and access as the primary concern (male Topic 1: 63.61%; female Topic 1: 45.14%), emphasizing the foundational role of connectivity, electricity, and digital systems. However, male respondents focused more on infrastructural (Topic 2: 5.45%; Topic 8: 1.97%; Topic 10: 1.25%) and professional network gaps (Topic 4: 3.97%; Topic 5: 3.26%), and workforce shortages (Topic 7: 2.36%). In contrast, female respondents emphasized sociopolitical and cultural challenges, including language, policy, and cost barriers (Topic 2: 17.43%; Topic 6: 5.00%), patient access (Topic 3: 8.75%), trust in digital systems (Topic 9: 1.19%), and data protection (Topic 10: 0.89%). Both groups acknowledged rural challenges, education gaps, and politico-economic issues, though women assigned higher variability to these themes (Topic 7: 4.18% vs Topic 7: 2.36%; Topic 5: 6.14% vs Topic 6: 2.80%; Topic 8: 3.02% vs Topic 9: 1.76%; Topic 4: 7.62% vs Topic 3: 4.36%).

In the articulation of future-oriented scenarios, integrated digital care and health technologies (Topic 1: 50.58% explained variance) highlighted a strong consensus on foundational infrastructure and system-wide digital innovation. This was followed by mobile innovation and app-based access (Topic 2: 11.00%) and AI for underserved regions and clinical support (Topic 3: 7.03%), reflecting emphasis on scalable, intelligent solutions for equity and efficiency. Additional themes included electronic health records, mobile health (mHealth), and Internet of Things convergence (Topic 4: 5.26%) and big data potential (Topic 5: 3.43%), both pointing to an evolving interest in interconnected health data ecosystems. Lower-variance topics such as AI-powered decision support (Topic 6: 2.40%), scalable toolkits (Topic 7: 2.13%), and digital medicine (Topic 8: 1.33%) contributed further nuance, underscoring the importance of interoperability, clinical augmentation, and monitoring. For both male and female participants, the dominant theme centered on general digital infrastructure, though with slightly different emphases. Men prioritized general digital health infrastructure (Topic 1: 48.88% variance), emphasizing data and technological systems (Topic 5: 2.99%), while women focused on remote digital health management and records (Topic 1: 72.13%; Topic 7: 1.38%), reflecting a patient-facing orientation grounded in access and care delivery. Male respondents highlighted mobile AI and telemedicine innovation (Topic 2: 13.72%), chronic care and surveillance (Topic 3: 8.98%; Topic 4: 5.10%; Topic 9: 1.38%), and smartphone-based scalable solutions (Topic 6: 2.65%)—a pattern that favors modular, tech-forward tools. In contrast, women showed greater thematic fragmentation across app-based empowerment (Topic 2: 9.17%), AI/chatbot integration with mHealth and Internet of Things (Topic 3: 4.98%; Topic 4: 3.51%; Topic 6: 1.85%), and smartphone planning tools (Topic 8: 0.75%). Notably, chatbot technologies appeared in both profiles (male Topic 7: 2.13%, female Topic 4: 3.51%), although embedded in different semantic contexts—clinical utility for men, and innovation for women. Men also highlighted big data and response systems (Topic 8: 1.56%), while women emphasized real-time data collection (Topic 5: 2.11%).

Further details are reported in Table 1 and Tables S1-S12 in Multimedia Appendix 2.

XSL•FO
RenderX

Table 1. Latent semantic analysis–based topic modeling across 4 dimensions of digital health discourse in the Global South: implementation, opportunities, challenges, and foresight. For each of the extracted topics, the percentage of explained variance and corresponding thematic focus are reported.

Topic	Implementation		Opportunities		Challenges		Foresight	
	Explained variance (%)	Thematic focus	Explained variance (%)	Thematic focus	Explained variance (%)	Thematic focus	Explained variance (%)	Thematic focus
1	46.75	Integration, infrastructure, strategic alignment	57.86	Health care access and remote care	59.15	Infrastructural limitations, tech fragmentation	50.58	Integrated digital care and health technologies
2	7.57	Patient management and service delivery	7.25	Sustainability, institutional development	5.59	Internet and electricity access deficits	11.00	Mobile innovation and app-based access
3	4.77	Stakeholder engagement, community outreach	6.07	Chronic disease management	3.93	Local health system bottlenecks	7.03	AI[a] for underserved regions and clinical support
4	4.21	Governance and regulation	4.04	Data-driven, patient-centered innovation	3.71	Remote region access challenges	5.26	EHR[b], mHealth, and IoT[c] convergence
5	4.04	Regulatory clarity and local adaptation	3.46	Implementation gaps at community level	3.09	Community network fragility, institutional misalignment	3.43	Big data potential
6	3.57	Digital literacy and tech access	3.01	Interoperability, workforce readiness	2.56	Digital literacy gaps	2.40	AI-powered clinical decision support
7	2.99	Co-design and participatory processes	1.96	Mobile support and training	2.36	Rural governance trust issues	2.13	Toolkits and scalable tech solutions
8	2.63	Scalability and sustainability	1.80	Public infrastructure	2.18	Financial constraints	1.33	Digital medicine and outcome tracking
9	2.21	Capacity building	1.54	Epidemic preparedness	1.87	Politicization	—[d]	—
10	2.07	Context-aware frameworks	1.32	Scalable digital records	1.63	Workforce/data infrastructure shortages	—	—

[a]AI: artificial intelligence.

[b]EHR: electronic health records.

[c]IoT: Internet of Things.

[d]Not available.

One Health in the Global South

LSA of responses on One Health opportunities in the Global South revealed a dominant discourse anchored in integrated human, animal, and environmental health systems (Topic 1: 57.20% of the variance). Respondents identified strong opportunities in ecological integration, predictive modeling (Topic 2: 7.27%), digital innovation and AMR surveillance (Topic 3: 6.23%), and AI-driven tools and awareness strategies (Topic 4: 4.79%). Additional themes included prevention and leadership (Topic 5: 3.55%), holistic approaches (Topic 6: 3.10%), community inclusion (Topic 7: 2.29%), vaccine distribution (Topic 8: 1.82%), digital records (Topic 9: 1.60%), and expert capacity building (Topic 10: 1.46%). Female and male respondents shared a strong common emphasis on the ecological foundations of One Health. For both, Topic 1 was dominant—accounting for 73.20% of variance in women and 57.45% in men—and focused on the interconnectedness of human, animal, and environmental health. Beyond this, the thematic priorities diverged significantly. Among women, the discourse was highly concentrated, with secondary themes oriented around strategic implementation and population-level equity. Topic 2 (5.42%) addressed AI and demographic growth, while Topic 3 (4.96%) focused on improving health responses and prevention. Topic 4 (3.93%) indicated attention to sectoral organization. Additional themes included community care and lived experience (Topic 5: 3.73%), treatment applications and clinical uptake (Topic 6: 2.97%), and smart working environments (Topic 7: 2.37%). Less prominent but still present were references to awareness building (Topic 8: 1.23%), health services (Topic 9: 1.14%), and remote training and key local solutions (Topic 10: 0.62%). In contrast, male respondents articulated a more diffuse and operationally driven perspective. Topic 2 (8.27%) emphasized predictive analytics and surveillance technologies, and Topic 3 (7.15%) addressed AMR through a sectoral lens. Topic 4 (5.55%) centered on AI and

XSL•FO
RenderX

complex policy coordination. While women focused on communities, men placed greater weight on biodiversity leadership and structured programs (Topic 5: 3.90%), vaccination and disease control strategies (Topic 6: 3.07%), and infectious disease enablement and access (Topic 7: 2.23%). Topic 8 (1.86%) highlighted digital recordkeeping and remote access infrastructure, while Topic 9 (1.60%) captured surveillance, expert knowledge, and vector-borne risks. Topic 10 (1.15%) pointed to health system resources, support, and scale-up potential.

Concerning challenges, Topic 1, accounting for 61.23% of the variance, emphasized environmental degradation, limited resources, and systemic health-environment linkages. Topic 2 (7.43%) highlighted data gaps and weaknesses in surveillance systems, particularly in the context of environmental change. Topic 3 (4.93%) focused on the lack of interdisciplinary coordination and field-level impact. Topic 4 (4.19%) addressed insufficient policy capacity, research infrastructure, and cultural readiness. Topic 5 (3.57%) addressed population-level issues, while Topic 6 (2.76%) pointed to rural marginalization and priority misalignment. Topic 7 (1.97%) underscored digital skill gaps and limited connectivity, and Topic 8 (1.65%) emphasized resource shortages and structural barriers. Topic 9 (1.35%) and Topic 10 (1.29%) captured concerns on meat safety and technological devices. Both female and male respondents strongly converged on structural and environmental limitations as the foremost barrier to implementing One Health in the Global South. In both groups, Topic 1 dominated the discourse—accounting for 61.94% of variance in women and 65.81% in men—and emphasized terms such as health, environment, resource, challenge, and implementation, revealing a shared perception of systemic constraints and infrastructural insufficiency. Beyond this core alignment, the 2 groups diverged notably in the framing and distribution of secondary themes. Among female respondents, Topic 2 (14.68%) focused on population practices, understanding, and adoption, suggesting an emphasis on behavioral, educational, and participatory barriers. Topic 3 (8.84%) addressed funding shortfalls, data gaps, and operational limitations, while Topic 4 (5.32%) pointed to sectoral and institutional weaknesses. Topic 5 (2.36%) dealt with device- and research-based deficiencies. Less prominent themes included poverty (Topic 6: 1.92%), cross-sector coordination (Topic 7: 1.57%), economic emergence and shocks (Topic 8: 1.00%), technological and cultural constraints (Topic 9: 1.00%), and multilevel governance issues (Topic 10: 0.91%). These findings reflect a relatively distributed but coherent concern with systemic functionality, knowledge equity, and integrated governance. In contrast, male respondents articulated a more operationally dispersed set of challenges. Topic 2 (8.80%) dealt with surveillance gaps and environmental degradation, while Topic 3 (5.28%) emphasized capacity building and culturally embedded policy. Topic 4 (3.76%) addressed priority setting and geographic inequities, and Topic 5 (2.98%) brought attention to rural underdevelopment, digital skills, and infrastructure deficits. Topic 6 (2.21%) highlighted low connectivity and staff shortages, and Topic 7 (1.83%) captured community-level implementation difficulties, particularly in relation to meat safety and device availability, while Topic 8 (1.28%) focused on resource and capacity deficits. Topic 9 (1.13%) highlighted disciplinary fragmentation in science and field application, and Topic 10 (1.01%) included conceptual concerns.

Concerning future-related opportunities, Topic 1, explaining 58.78% of the variance, emphasized integrated health approaches linking disease prevention, animal health, and environmental improvement. Topic 2 (5.20%) reflected ambitions for enhanced health care responsiveness through real-time monitoring and information systems. Topic 3 (4.86%) addressed food systems, AMR, and environmental-professional literacy, while Topic 4 (3.80%) pointed to systems development and the implementation of expert-driven frameworks. Additional topics captured emerging or cross-cutting themes. Topic 5 (3.40%) focused on conservation, biodiversity, and domain-specific innovation, and Topic 6 (3.10%) raised concerns about infrastructure, digital records, and electronic health integration. Topic 7 (2.53%) focused on forward-oriented transformative potential of One Health in the Global South. Topic 8 (2.18%) underscored access to care and population equity, while Topic 9 (2.11%) centered on infectious disease awareness and public engagement. Finally, Topic 10 (1.91%) reflected projections around early medical interventions and future modeling scenarios. Male and female respondents both emphasized the ecological nexus of human, animal, and environmental health as a core future opportunity for One Health advancement (male Topic 1: 60.29%; female Topic 1: 68.63%), highlighting a shared foundational understanding across genders. However, differences emerged in the framing and elaboration of these priorities. Women emphasized responsive systems and coordination tools (Topic 2: 20.13%) and showed a stronger orientation toward cultural integration and community empowerment, as seen in topics such as indigenous engagement and awareness promotion (Topic 3: 2.92%), technological advancement and collaboration (Topic 4: 2.77%), and specific innovations like diagnostics (Topic 5: 1.44%), outreach (Topic 6: 0.95%), and data availability (Topic 7: 0.48%). They also included a diffuse but conceptually rich theme on future risks, AMR, and professional preparedness (Topic 8: 0.00%). Male respondents, while similarly prioritizing the health-environment interface (Topic 1: 60.29%), articulated a more technical and operational approach in the remaining topics. These included strengthening sustainable health care systems and food security (Topic 2: 6.03%), implementing expert-led systems (Topic 3: 4.95%), resolving infrastructural and data management issues (Topic 4: 3.93%), and developing digital platforms (Topic 5: 3.33%). Other themes involved community-based research and responsiveness (Topic 6: 3.01%), epidemic readiness and access (Topic 7: 2.69%), early interventions and infectious disease modeling (Topic 8: 2.49%), mobile diagnostics and rapid response (Topic 9: 1.92%), and population-level coordination (Topic 10: 1.61%).

Finally, regarding future-related challenges, Topic 1, explaining 59.65% of the variance, reflected deep-seated anxieties around health system fragility, environmental degradation, resource scarcity, and the growing burden of disease. Topic 2 (7.81%) pointed to digital infrastructure gaps and persistent barriers to health management and equitable access. Topic 3 (5.47%) emphasized food systems, veterinary integration, and the need

XSL•FO
RenderX

for cohesive frameworks to support cross-sectoral promotion of One Health. Topic 4 (4.57%) addressed insufficiencies in outbreak detection and relational coordination, while Topic 5 (2.87%) focused on the lack of financial, regulatory, and institutional support for coordinated initiatives. Topic 6 (2.46%) highlighted time-sensitive skill shortages, data deficits, and the limited capacity at the local level. Topic 7 (2.08%) identified environmental opportunities yet to be enabled, suggesting underutilized potential in ecosystem-based approaches. Topic 8 (1.85%) captured cultural and economic inertia that may hinder timely and inclusive responses, and Topic 9 (1.62%) reflected concerns about weak collective action and unrealized potential across populations. Topic 10 (1.41%) pointed to limitations in emergency readiness, agency coordination, and resource mobilization. Both male and female respondents identified systemic and structural limitations as the most significant future challenge for implementing One Health in the Global South. Topic 1 accounted for 78.50% of variance in female responses and 61.65% in men. In both cases, the theme encompassed resource scarcity, environmental degradation, infrastructural weaknesses, and disease burden, indicating a clear consensus on foundational barriers. The female framing included community-based challenges, while the male version emphasized intersectoral and operational breakdowns. Female Topic 2 (8.30%) focused on management, regulatory complexity, cultural barriers, funding deficiencies, and political inertia. This aligns with male Topic 4 (3.42%), which also emphasized coordination, governance, and regulatory support, though framed more technocratically. The female version was more critical and rooted in institutional culture. Female Topic 3 (6.46%) addressed emerging threats, climate-driven disease risk, and resource vulnerabilities, which parallels male Topic 3 (5.77%) on food systems, integration, and outbreak preparedness. Both saw environmental unpredictability and multisectoral fragility as critical future stressors. Female Topic 4 (2.18%) on economic pressure and population-level stress resonated with male Topic 7 (2.13%) and Topic 10 (1.16%), which covered poverty, inequity, local capacity gaps, and leadership voids. The male response provided more layers, incorporating network fragility, population-scale planning, and decentralized leadership, while the female view presented a compact but conceptually broad expression of economic vulnerability. Female Topic 5 (1.39%) emphasized data collection and information deficits, which echoes male Topic 5 (2.80%) and Topic 6 (2.50%) on skills, local-level data systems, and epidemiological responsiveness. The male focus offered more operational detail, while the female version pointed to a diagnostic gap. Female Topic 6 (0.57%) is abstract and thinly developed, but potentially refers to actual lived experience or practical realities—a dimension that, while unelaborated, might align with male Topics 8 (2.01%) and 9 (1.26%) concerning time constraints, remoteness, and pragmatic insufficiencies. Finally, although female Topic 7 (0.00% variance) did not contribute to model variance, it contained rich content including disciplinary fragmentation, outbreak management, and regional opinion dynamics. This latent theme maps well to male Topics 2 (9.17%) and 8 (2.01%), which dealt with Global South–specific digital development, coordination barriers, and interdisciplinary silos.

Further details are reported in Table 2 and Tables S13-S23 in Multimedia Appendix 2.

XSL•FO
RenderX

Table 2. Latent semantic analysis–derived topic modeling of stakeholder discourse on One Health in the Global South. Each topic's explained variance and thematic focus are presented for 4 analytical lenses: opportunities, challenges, future opportunities, and anticipated challenges.

Topic	Opportunities		Challenges		Foresight opportunities		Foresight challenges	
	Explained variance (%)	Thematic focus	Explained variance (%)	Thematic focus	Explained variance (%)	Thematic focus	Explained variance (%)	Thematic focus
1	57.20	Integrated human-animal-environment systems	61.23	Environmental degradation, limited resources, systemic health-environment links	58.78	Integrated disease prevention and environmental improvement	59.65	Health-system fragility, environmental degradation, resource scarcity
2	7.27	Ecological integration, predictive modeling	7.43	Data gaps, surveillance weaknesses	5.2	Real-time monitoring and responsive systems	7.81	Digital infrastructure and access barriers
3	6.23	Digital innovation, AMR[a] surveillance	4.93	Lack of interdisciplinary coordination	4.86	Food systems, AMR, environmental-professional literacy	5.47	Food systems, veterinary integration, cross-sector support
4	4.79	AI[b] tools, awareness strategies	4.19	Insufficient policy, research, and cultural capacity	3.8	Systems development, expert frameworks	4.57	Outbreak detection, coordination deficits
5	3.55	Prevention and leadership	3.57	Population-level concerns	3.4	Conservation, biodiversity, innovation	2.87	Financial, regulatory, and institutional voids
6	3.10	Holistic approaches	2.76	Rural marginalization, priority misalignment	3.1	Digital infrastructure, e-records integration	2.46	Skill gaps, data deficits, local capacity
7	2.29	Community inclusion	1.97	Digital skill gaps, limited connectivity	2.53	Transformative One Health potential	2.08	Underutilized ecosystem potential
8	1.82	Vaccine distribution	1.65	Resource shortages, structural barriers	2.18	Access to care, equity	1.85	Cultural/economic inertia
9	1.60	Digital records	1.35	Meat safety concerns	2.11	Infectious disease awareness, engagement	1.62	Weak collective action
10	1.46	Expert capacity building	1.29	Technological device limitations	1.91	Early intervention and predictive modeling	1.41	Emergency readiness, coordination, resource mobilization

[a]AMR: antimicrobial resistance.

[b]AI: artificial intelligence.

Digital One Health in the Global South

In the LSA of responses to the question on the convergence of digital health and One Health approaches in the Global South, Topic 1, accounting for 75.2% of the total variance, emphasized digitalization, health data infrastructure, and disease surveillance systems, as underpinning the operationalization of the integration of digital health and One Health. Topic 2, explaining 5.4%, captured the role played by global convergence agendas, sectoral alignment, and climate recognition in shaping integrative frameworks across human, animal, and environmental health. Topic 3 (3.9%) pointed to the catalytic effect of the COVID-19 pandemic, which had prompted holistic thinking and increased organizational and financial support for integrated approaches. Topic 4 (3.0%) highlighted how AI and risk management strategies had been mobilized to address systemic constraints and enhance preparedness. Topic 5 (1.8%) reflected the importance of national delivery mechanisms and context-specific implementation, while Topic 6 (1.5%) illustrated the multisectoral impacts that had been particularly salient in the Global South. Topic 7 (1.4%) underscored governance challenges, including institutional fragmentation and reliability gaps, that had historically limited coordinated digital health and One Health responses. Finally, Topics 8 (0.87%), 9 (0.69%), and 10 (0.59%) were centered on policy awareness, stakeholder engagement, and scaling strategies. Topic 1 accounted for the majority of explained variance in both genders: 75.05% in women and 77.31% in men. Both groups emphasized the integration of digital health technologies into disease surveillance within a One Health framework. However, while female responses highlighted the importance of collaborative and community-oriented approaches, male

XSL•FO
RenderX

responses were more focused on technological infrastructure and human-disease interaction, illustrating a shared theme with participatory emphasis in women and infrastructural framing in men. Topic 2 was present in both genders (women: 4.77%, men: 5.62%) and concerned management and international drivers of convergence. Female responses were brief and managerial in focus, whereas male responses expanded on sectoral alignment, climate imperatives, and global recognition—pointing to a more externally driven and systemic perspective among male respondents. Topic 3 (women: 4.49%, men: 4.21%) demonstrated alignment around climate change and health system transformation. Women emphasized anticipatory and ecological aspects of convergence, while male responses framed the issue in light of pandemic-triggered reconfigurations and institutional support needs. Both acknowledged the environment-health nexus but with distinct emphases: prospective and global in women, reactive and policy-driven in men. Topic 4 (women: 3.52%, men: 3.24%) centered on institutional governance and risk. Female respondents focused on interministerial communication and vertical coordination, whereas men addressed operational aspects of risk management and the technical requirements of digital health implementation. The shared institutional theme revealed gendered nuances: administrative structure in women versus digital governance in men. Topic 5 (women: 3.26%, men: 1.63%) diverged in scope. Female responses addressed resource allocation and programmatic implementation challenges. In contrast, male discourse shifted toward geopolitical dynamics and multisector imbalances, reflecting broader inequalities in digital readiness across regions. Both pointed to structural enablers and barriers, but from local programmatic (female) and systemic geopolitical (male) vantage points. Topic 6 (women: 2.90%, men: 1.51%) related to delivery and performance. Women focused on access and service efficacy, while men underlined fragmentation and interministerial unreliability. Both acknowledged weaknesses in execution, with women emphasizing outcomes and men highlighting administrative dysfunction. Topic 7 was marginal among women (1.59%) and absent in the male profile. The theme suggested a latent concern with public engagement and legitimacy, though it remained underdeveloped in both groups. Topic 8 (women: 1.46%, men: 0.64%) touched on data access and informational equity. Both groups acknowledged the enabling role of digital health in democratizing health information, although the female response focused more explicitly on availability and transparency. Topic 9 (women: 1.45%, men: 0.59%) was conceptually minimal. While women pointed to the importance of key factors without elaboration, male responses offered no interpretable content. This suggests a shared but weakly articulated recognition of critical convergence enablers. Topic 10 was notably richer in the female discourse (0.78% vs 0.51%). Female responses addressed themes such as health inequity, underserved populations, antimicrobial stewardship, and system efficiency. Male responses, in contrast, referred to national context or jurisdiction.

Further details are reported in Table 3 and Tables S24-S26 in Multimedia Appendix 2.

Table 3. Latent semantic analysis of stakeholder responses on the convergence of digital health and One Health in the Global South.

Topic	Explained variance (%)	Thematic focus
1	75.2	Digitalization, health data infrastructure, disease surveillance
2	5.4	Global convergence agendas, sectoral alignment, climate recognition
3	3.9	Catalytic effect of COVID-19 on holistic and integrative thinking
4	3.0	AI[a] and risk management strategies for systemic preparedness
5	1.8	National delivery mechanisms and context-specific implementation
6	1.5	Multisectoral impacts in the Global South
7	1.4	Governance challenges and institutional fragmentation
8	0.87	Policy awareness and advocacy
9	0.69	Stakeholder engagement
10	0.59	Scaling strategies

[a]AI: artificial intelligence.

Discussion

Principal Findings

The foresight-oriented analysis of expert narratives revealed distinct thematic architectures underlying the discourses on digital health, One Health, and their convergence in the Global South. Importantly, these architectures should be interpreted not in isolation, but in relation to the existing international literature, which has been overwhelmingly shaped by Global North perspectives.

In the domain of digital health, structural drivers were consistently anchored in the infrastructural and systemic integration of digital technologies into health systems. These included the entrenchment of digital data ecosystems, the scaling of mHealth interventions, and the embedding of digital platforms into clinical workflows and public health delivery. Such drivers reflect a maturing landscape where digital technologies are no longer peripheral but fundamental to the functioning and reform of health governance in many Global South contexts. This emphasis partially mirrors findings from the Global North, where digital health maturity is typically framed around

XSL•FO
RenderX

interoperability, large-scale electronic health records, and AI-enabled clinical decision support embedded within stable institutional environments. However, unlike Global North settings—where digital infrastructure is often treated as a given—the Global South discourse consistently foregrounds infrastructural fragility, uneven access to electricity and connectivity, and dependence on external funding as persistent structural constraints.

Emerging issues in digital health centered on the transformation of care models through AI, the proliferation of mobile diagnostics, and evolving concerns around data sovereignty and ethical oversight. These topics reflect a transition in focus from basic digital inclusion toward strategic, anticipatory integration of advanced technologies. While similar themes appear in Global North studies—particularly around AI adoption and data governance—these findings suggest a qualitatively different framing in the Global South, where AI is often viewed as an aspirational or compensatory tool intended to offset workforce shortages and system limitations, rather than as an optimization layer built upon already robust health systems.

Particularly noteworthy was the shift from discussions of access and deployment to themes of agency, responsiveness, and the societal implications of algorithmic decision-making. Gendered nuances were evident, with female respondents frequently framing these issues through the lenses of equity, community participation, and trust, while male respondents emphasized operational scalability and innovation efficiency.

Within the One Health discourse, structural drivers coalesced around intersectoral coordination, disease surveillance at the human-animal-environment interface, and the growing imperative to institutionalize One Health frameworks. These reflect a stable recognition of ecological entanglement and the necessity of integrative governance structures to mitigate transboundary health threats. The discourse was marked by convergent priorities such as early warning systems, environmental health data infrastructures, and capacity-building in veterinary and environmental public health. Across genders, these priorities were consistently articulated, although women more frequently situated them within community resilience and social vulnerability contexts. This aligns with Global North One Health literature, which similarly prioritizes multisectoral surveillance, zoonotic preparedness, and early warning systems. However, in contrast to Global North contexts—where One Health is often operationalized through formalized interagency mechanisms—the Global South discourse reveals that coordination remains largely aspirational, constrained by fragmented mandates, limited regulatory authority, and underresourced environmental and veterinary sectors.

Emerging issues in One Health focused on anticipatory risk management, the operationalization of multisectoral collaborations, and the rising salience of food system sustainability and climate resilience. These signal a discursive broadening of One Health from zoonotic disease control toward more systemic, planetary health considerations. Notably, whereas Global North literature increasingly frames these challenges through the lens of technological integration and policy alignment, Global South experts more frequently emphasized structural underinvestment, competing development priorities, and the vulnerability of rural and marginalized populations to ecological shocks.

In several instances, female participants extended the discourse into areas of inclusive governance, highlighting the importance of participatory institutions and local knowledge systems in shaping resilient One Health infrastructures. These contributions suggest a shift toward democratized models of One Health governance that move beyond technical coordination toward transformative system change. Weak signals in One Health included fragmented reflections on indigenous knowledge integration, AMR governance at subnational levels, and structural underinvestment in ecological public health. Though sparse in expression, these signals may indicate latent areas of strategic neglect that could become critical under future scenarios marked by climate instability or political realignment. Such weak signals are largely absent or marginal in the Global North literature, where One Health discussions tend to privilege formal scientific expertise and national-level governance, thereby underscoring the added value of a foresight-oriented approach capable of surfacing peripheral yet potentially transformative concerns. The gendered content of these weak signals was particularly rich, with women surfacing relational, trust-based framings and men occasionally noting the absence of jurisdictional clarity and institutional mandates.

The discourse on the convergence of digital health and One Health was anchored by structural drivers that framed integration as both inevitable and strategic. Participants consistently viewed digital tools as essential enablers of One Health implementation, particularly in data sharing, interagency coordination, and real-time surveillance. This convergence was portrayed not as a mere overlay of technologies but as a paradigmatic shift toward more intelligent, interconnected, and adaptive health ecosystems. Emerging issues in the convergence discourse included the governance of cross-sectoral data flows, the development of digital infrastructures tailored to ecological surveillance, and the political economy of technology access in Global South contexts. These themes reflected a growing awareness of the nonneutrality of digital infrastructures and the need to align technological solutions with values of justice, inclusion, and environmental stewardship. Gendered patterns again surfaced, with female experts more likely to emphasize coproduction, relational accountability, and the risks of epistemic exclusion. This mirrors Global North narratives that conceptualize Digital One Health primarily as a technical convergence problem. However, the Global South discourse identified here reframes convergence as a deeply political and sociotechnical process, shaped by digital sovereignty, power asymmetries, and unequal access to technological resources.

Weak signals in the convergence space included underexplored themes such as feminist digital ethics, algorithmic bias in zoonotic prediction models, and the political implications of digital sovereignty in environmental health governance. Though peripheral in current discourse, these signals point to potentially disruptive tensions that could reshape future directions in Digital One Health if left unaddressed. Their emergence in this study highlights how foresight methodologies, particularly when combined with gender-sensitive analysis, can reveal dimensions

XSL•FO
RenderX

of convergence that are systematically overlooked in implementation-focused research.

Altogether, the analysis surfaced a dynamic and increasingly reflexive ecosystem of discourse, in which well-established priorities coexist with early indicators of transformative change. By explicitly contrasting Global South expert imaginaries with the dominant Global North literature, this study demonstrates that Digital One Health convergence is not a universally linear or technologically deterministic process. Rather, it is contextually negotiated, unevenly paced, and deeply shaped by structural inequities, governance capacities, and gendered epistemologies. By disaggregating structural drivers, emerging issues, and weak signals across gendered and thematic lines, the study contributes a nuanced, equity-informed foresight mapping of how health futures are being envisioned in the Global South.

Building upon the LSA-based foresight mapping, a comparative reflection with recent qualitative literature [14-17] on digital health and One Health in the Global South reveals substantial areas of alignment as well as conceptual expansion. While prior studies have independently captured infrastructural, systemic, and organizational dimensions, this LSA analysis contributes temporal depth and a gendered, anticipatory framing that enriches and challenges existing scholarship.

Strengths and Limitations

This study leverages a robust methodological approach by integrating LSA-based topic modeling with foresight-oriented frameworks to explore the convergence of digital health and One Health in the Global South. The analytical design, which mapped present and future-oriented discourses across 4 strategic dimensions—implementation, opportunities, challenges, and prospective trajectories—enabled a layered classification of structural drivers, emerging issues, and weak signals. The incorporation of a Delphi-informed purposive sampling strategy ensured the inclusion of a diverse array of expert voices across geographical, disciplinary, and institutional boundaries. This methodological rigor lends robustness and ecological validity to the findings, especially in capturing the pluralistic and multisectoral character of the digital health and One Health convergence discourse in resource-constrained settings.

Despite these strengths, the study is not without limitations. First, while purposive sampling captured disciplinary diversity, the geographic composition of the expert panel was heavily skewed toward Africa, with 58% (26/45) of participants based in African countries. In contrast, representation from Latin America and the Caribbean and the Middle East and North Africa was minimal, comprising only 7% (3/45) and 4% (2/45) of the sample, respectively. Due to this imbalance, it was not possible to perform a reliable LSA stratified by geographic provenance. As such, important regional nuances—especially from underrepresented areas—may not be adequately captured. Second, the reliance on English-language responses potentially excluded non-Anglophone voices and introduced language-related semantic constraints. Third, while LSA effectively reduces linguistic dimensionality and enables the identification of latent thematic structures, it may obscure context-specific narrative nuance and is sensitive to preprocessing decisions such as tokenization, term selection, and sparsity thresholds. As a result, the derived topics reflect dominant semantic patterns rather than the full richness or contextual specificity of individual expert narratives, and should be interpreted as heuristic representations rather than definitive constructs. Finally, although the study incorporated a gender lens, the male-to-female ratio was 2.8:1, limiting the interpretive depth of gender-divergent discourse, particularly for weak signals.

A further limitation relates to the reliance on open-ended survey questions. While this approach was intentionally adopted to support exploratory, foresight-oriented inquiry, it inevitably introduces interpretive variability, as respondents may construe and frame the same prompt in different ways. In this study, this variability was reflected in systematic differences between male and female respondents, not only in the thematic priorities they emphasized but also in the semantic framing and level of abstraction of their responses. Such gender-related divergences should not be interpreted as inconsistencies or measurement error, but rather as an inherent feature of qualitative, discursive data elicited through open-ended instruments. Nonetheless, this interpretive flexibility limits direct comparability across respondent groups and constrains inferential interpretations of gender differences. While the use of LSA helped to identify shared latent structures across heterogeneous narratives, it cannot fully eliminate variability arising from differential interpretation of survey prompts. Closely related to this issue, an additional limitation concerns the survey instrument itself. The questionnaire was developed ad hoc for this study and was not based on a previously validated instrument. This approach was necessary given the exploratory and foresight-oriented nature of the research, as well as the absence of established instruments designed to capture expert perspectives on the convergence of digital health and One Health, particularly in Global South contexts. However, the use of an ad hoc instrument limits direct comparability with other studies and may affect measurement consistency. The multiround Delphi design, purposive expert sampling, and the application of LSA partially mitigated this limitation by emphasizing convergence of meaning across narratives rather than item-level psychometric properties. In addition, the predominance of midcareer participants—while purposeful for capturing forward-looking and implementation-oriented perspectives—may have limited the extent to which long-term institutional memory and historical continuity were represented. Future research could address these limitations through more focused and structured study designs. Follow-up studies may combine open-ended foresight prompts with semistructured or closed-ended items anchored to predefined conceptual domains such as governance, infrastructure, ethics, and equity. The use of scenario-based questions, vignettes, or Likert-scaled assessments of clearly defined drivers and barriers could enhance cross-group comparability while preserving interpretive depth. Mixed methods approaches integrating qualitative elicitation with quantitative validation, alongside more balanced gender representation, would further strengthen the robustness, comparability, and interpretability of future analyses.

XSL•FO
RenderX

Conclusions

This study provides a foresight-informed, gender-sensitive mapping of the convergence between digital health and One Health discourses in the Global South. Structural drivers centered on the systemic integration of digital infrastructure, intersectoral governance, and ecological surveillance, while emerging issues pointed toward anticipatory innovation, AI deployment, and food system resilience. Weak signals—particularly those surfaced by female respondents—highlighted underexplored dimensions such as indigenous knowledge, feminist digital ethics, and community-based relational dynamics. By disaggregating expert discourse through LSA and embedding it within a futures thinking framework, the study surfaces both prevailing and nascent imaginaries shaping global health trajectories. These findings underscore the importance of inclusive, equity-oriented approaches in shaping Digital One Health agendas, and call for stronger attention to gendered epistemologies, participatory governance, and digital sovereignty in future research and implementation efforts.

Acknowledgments

We extend our deepest gratitude to everyone who contributed to the success of this project. First, we would like to thank the International Development Research Centre (IDRC) for commissioning and supporting this critical work under contract number PO5001096. Their commitment to advancing research in the Global South has been invaluable in driving forward this initiative. We are especially thankful to the IDRC management team (Dr Chaitali Sinha, Dr Fraser Reilly-King, Dr Samuel Oji Oti, Dr Sana Naffa, and Dr Engelbert Bain Luchuo) for their constant support and guidance throughout this project. Their understanding of the complex health landscape in the Global South has greatly contributed to the relevance and impact of this report. Our sincere appreciation goes to the 45 experts from Africa, Asia, Latin America, and the Middle East who participated in the foresight exercise and provided critical input. Their collective expertise, diverse perspectives, and commitment to improving global health have been central to identifying the opportunities and challenges facing Digital One Health in the Global South. Lastly, we would like to acknowledge the communities and local stakeholders across the Global South whose lived experiences, voices, and knowledge are at the heart of this research. Their resilience and determination continue to inspire our work toward creating more equitable and sustainable health systems. Thank you all for your support, collaboration, and belief in the power of innovation to transform global health.

Funding

This research is funded by Canada's International Development Research Centre (contract number PO5001096). JK acknowledges support from Canada's IDRC (grant 109981-001), IDRC and the Foreign, Commonwealth and Development Office (grant 110554-001), Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (grant RGPIN-2022-04559), NSERC Discovery Launch Supplement (grant DGECR-2022-00454), New Frontier in Research Fund-Exploratory (grant NFRFE-2021-00879), and Canada Research Chair in Community-Oriented Artificial Intelligence and Mathematical Modelling of Infectious Diseases (award ID CRC-2023-00234.).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey items.
[DOCX File , 16 KB - jmir_v28i1e78702_app1.docx]

Multimedia Appendix 2
Additional tables.
[DOCX File , 56 KB - jmir_v28i1e78702_app2.docx]

References

1. Barrett B, Charles JW, Temte JL. Climate change, human health, and epidemiological transition. Prev Med 2015;70:69-75 [FREE Full text] [doi: 10.1016/j.ypmed.2014.11.013] [Medline: 25434735]
2. Hua K, Pan Y, Fang J, Wu H, Hua Y. Integrating social, climate and environmental changes to confront accelerating global aging. BMC Public Health 2024;24(1):2838 [FREE Full text] [doi: 10.1186/s12889-024-20346-7] [Medline: 39407185]
3. Boutayeb A. The burden of communicable and non-communicable diseases in developing countries. In: Handbook of Disease Burdens and Quality of Life Measures. New York, NY: Springer; 2010:531-546.
4. Acolin J, Fishman P. Beyond the biomedical, towards the agentic: a paradigm shift for population health science. Soc Sci Med 2023;326:115950 [FREE Full text] [doi: 10.1016/j.socscimed.2023.115950] [Medline: 37148746]

XSL•FO
RenderX

5. Horefti E. The importance of the One Health concept in combating zoonoses. Pathogens 2023;12(8):977 [FREE Full text] [doi: 10.3390/pathogens12080977] [Medline: 37623937]
6. Danasekaran R. One Health: a holistic approach to tackling global health issues. Indian J Community Med 2024;49(2):260-263 [FREE Full text] [doi: 10.4103/ijcm.ijcm_521_23] [Medline: 38665439]
7. Abernethy A, Adams L, Barrett M, Bechtel C, Brennan P, Butte A, et al. The promise of digital health: then, now, and the future. NAM Perspect 2022;2022 [FREE Full text] [doi: 10.31478/202206e] [Medline: 36177208]
8. Scott P, Adedeji T, Nakkas H, Andrikopoulou E. One Health in a digital world: technology, data, information and knowledge. Yearb Med Inform 2023;32(1):10-18 [FREE Full text] [doi: 10.1055/s-0043-1768718] [Medline: 37414034]
9. Benis A, Tamburis O, Chronaki C, Moen A. One Digital Health: a unified framework for future health ecosystems. J Med Internet Res 2021;23(2):e22189 [FREE Full text] [doi: 10.2196/22189] [Medline: 33492240]
10. Ragnedda M, Gladkova A. Understanding digital inequalities in the Gobal South. In: Ragnedda M, Gladkova A, editors. Digital Inequalities in the Global South. Global Transformations in Media and Communication Research - A Palgrave and IAMCR Series. Palgrave Macmillan, Cham: Springer; 2020:17-30.
11. Rezaeian M, Montazeri H, Loonen RCGM. Science foresight using life-cycle analysis, text mining and clustering: a case study on natural ventilation. Technol Forecast Soc Change 2017;118:270-280 [FREE Full text] [doi: 10.1016/j.techfore.2017.02.027] [Medline: 32287406]
12. Kayser V, Blind K. Extending the knowledge base of foresight: the contribution of text mining. Technol Forecast Soc Change 2017;116:208-215. [doi: 10.1016/j.techfore.2016.10.017]
13. OECD. Strategic Foresight Toolkit for Resilient Public Policy: A Comprehensive Foresight Methodology to Support Sustainable and Future-Ready Public Policy. Paris: OECD Publishing; 2025.
14. O'Brien N, Li E, Chaibva CN, Gomez Bravo R, Kovacevic L, Kwame Ayisi-Boateng N, et al. Strengths, weaknesses, opportunities, and threats analysis of the use of digital health technologies in primary health care in the sub-Saharan African region: qualitative study. J Med Internet Res 2023;25:e45224 [FREE Full text] [doi: 10.2196/45224] [Medline: 37676721]
15. Steenkamp I, Peltonen LM, Chipps J. Digital health readiness - insights from healthcare leaders in operational management: a cross-sectional survey. BMC Health Serv Res 2025;25(1):240 [FREE Full text] [doi: 10.1186/s12913-024-12129-y] [Medline: 39939988]
16. Zharima C, Griffiths F, Goudge J. Exploring the barriers and facilitators to implementing electronic health records in a middle-income country: a qualitative study from South Africa. Front Digit Health 2023;5:1207602 [FREE Full text] [doi: 10.3389/fdgth.2023.1207602] [Medline: 37600481]
17. Brown HL, Pursley IG, Horton DL, La Ragione RM. One Health: a structured review and commentary on trends and themes. One Health Outlook 2024;6(1):17. [doi: 10.1186/s42522-024-00111-x] [Medline: 39138554]

Abbreviations

AI: artificial intelligence
AMR: antimicrobial resistance
LSA: latent semantic analysis
mHealth: mobile health

Edited by T Kristóf, B Mesko; submitted 07.Jun.2025; peer-reviewed by Z Kozlakidis, D Verran; comments to author 22.Aug.2025; revised version received 21.Dec.2025; accepted 13.Jan.2026; published 18.Feb.2026.

Please cite as:
Kong J, Bragazzi NL
Framing the Convergence of One Health and Digital Health in the Global South With a Gender-Sensitive Foresight Perspective: Delphi Study Using Latent Semantic Analysis
J Med Internet Res 2026;28:e78702
URL: https://www.jmir.org/2026/1/e78702
doi:10.2196/78702
PMID:

©Jude Kong, Nicola Luigi Bragazzi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

A Futures Framework for Clinical AI Governance: Anticipating Emerging Risks, Shifting Roles, and Regulatory Challenges

Yi Yang[1], MBA; Jialin Liu[1,2,3*], MD; Siru Liu[4*], PhD

[1]Information Center, West China Hospital of Sichuan University, Chengdu, Sichuan, China
[2]Department of Otolaryngology-Head and Neck Surgery, West China Hospital of Sichuan University, Chengdu, Sichuan, China
[3]Department of Medical Informatics, West China Medical School of Sichuan University, Chengdu, Sichuan, China
[4]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States
[*]these authors contributed equally

Corresponding Author:
Jialin Liu, MD
Information Center
West China Hospital of Sichuan University
No 37 Guo Xue Xiang Street
Chengdu, Sichuan, 610041
China
Phone: 86 2885422416
Email: dljl8@163.com

Abstract

This viewpoint develops the futures framework for clinical artificial intelligence governance (FF-CAIG), a conceptual and anticipatory framework for organizing emerging governance challenges in clinical artificial intelligence (AI). Although life cycle–oriented oversight is increasingly reflected in clinical AI regulation and institutional governance, existing approaches remain more developed for near-term validation, current-state assurance, and retrospective risk detection than for longer-horizon sociotechnical change. This gap is increasingly relevant, as AI systems become more complex, adaptive, and autonomous, and as they become more deeply embedded in care relationships and accountability structures. FF-CAIG is grounded in 3 futures methodologies: the 3 horizons model, scenario planning, and causal layered analysis. It is operationalized through an emerging clinical AI risk taxonomy that links these methods to governance domains. Its practical outputs include horizon classification, risk-domain mapping, scenario stress-testing findings, accountability-chain mapping, and horizon-scaled minimum governance actions for deployment or continued use. Applied across near-term, transitional, and longer-term horizons, the framework proposes cross-horizon priorities, including stronger predeployment equity evaluation, clearer life cycle accountability, clinician AI oversight competencies, and safeguards for increasingly autonomous or AI-mediated care systems. We illustrate FF-CAIG through 3 representative clinical AI deployment patterns and discuss its limitations, including differential compliance burdens, risks of overdocumentation, variable feasibility across jurisdictions, and the need for empirical validation. FF-CAIG is intended not as a prescriptive policy instrument or validated assessment tool, but as a structured analytic approach for regulators, health system leaders, developers, and researchers seeking prospective and systems-oriented approaches to clinical AI governance.

(J Med Internet Res 2026;28:e96152) doi:10.2196/96152

KEYWORDS

clinical AI; artificial intelligence; clinical AI governance; futures studies; strategic foresight; health regulation; clinical workforce; emerging risks; digital health policy; artificial intelligence governance

Introduction

Clinical artificial intelligence (AI) has moved beyond controlled validation settings into routine deployment, supporting risk stratification, documentation, patient communication, and emerging decision support across care settings [1-3]. Regulatory oversight of imaging algorithms, predictive models, and other AI-enabled medical products has also expanded through evolving US Food and Drug Administration (FDA) and European Union frameworks [4,5]. Although clinical AI governance increasingly incorporates life cycle–oriented elements, including postmarket monitoring and change-management expectations, much oversight remains more mature for near-term assurance and recognized risk signals than

XSL•FO
RenderX

for longer-horizon sociotechnical change. Existing approaches, including institutional ethics oversight, regulatory approval pathways, and industry guidance, still focus primarily on risks such as bias, privacy breaches, performance degradation, and safety events that can be specified, monitored, or retrospectively detected [4-7].

This orientation is increasingly challenged by the iterative, updateable, and cross-functional nature of advanced clinical AI [8]. Recent regulatory developments, including the FDA's life cycle–focused guidance for AI-enabled device software functions and the EU AI Act's postmarket obligations for high-risk systems, indicate that clinical AI governance is no longer limited to static premarket approval [4,7,9]. The remaining gap is more specific: current life cycle–oriented approaches are better developed for planned or bounded change, defined postmarket obligations, and measurable performance signals than for evolving clinical functions, foundation model–based tools, distributed accountability, and relational effects that may shift across contexts and over time. In health care, this challenge is especially consequential because safety, accountability, and trust depend not only on model performance but also on role clarity, implementation quality, and institutional capacity to detect and respond to change [1,3].

The governance challenge posed by clinical AI is therefore not only technical but also temporal. Although regulators and health systems increasingly recognize the need for life cycle oversight, existing mechanisms still depend substantially on foreseeable changes, measurable postdeployment signals, and recognized failure modes. Clinical AI is evolving in ways that may generate governance problems before they are fully captured by conventional postdeployment oversight [10,11]. Some risks may emerge gradually, including clinician overreliance, rising verification burden, workflow dependence, and fragmentation of governance across institutions and jurisdictions [12-15]. Generative AI systems used for longitudinal communication, coordination, or triage may also reshape patient-clinician relationships in ways that existing frameworks for consent, accountability, and duty of care do not fully anticipate [16,17]. This concern is increasingly recognized in international guidance, including recent World Health Organization work on large multimodal models in health [18].

Futures studies provide a systematic but underused set of approaches for anticipating long-term change and emerging uncertainty [19,20]. Although organizations such as the Organisation for Economic Co-operation and Development and World Health Organization have identified horizon scanning and scenario analysis as valuable governance tools under conditions of rapid change [21,22], these methods have not been systematically integrated into clinical AI governance. As a result, longer-horizon concerns, including systemic dependency, workforce disruption, and regulatory obsolescence, often remain secondary to current-state compliance and near-term risk mitigation [15,23,24]. Existing frameworks provide important foundations, but they offer limited operational guidance for anticipating risks that may emerge beyond conventional life cycle monitoring, compliance assessment, or management-system processes. The futures framework for clinical artificial intelligence governance (FF-CAIG) addresses this gap by integrating futures methods with clinical AI governance.

In this viewpoint, we introduce FF-CAIG, grounded in core futures methodologies [19,20] and centered on patient safety and health equity as foundational principles [18,25]. The framework addresses 3 interconnected domains: emerging systemic and clinical risks, evolving stakeholder roles across the care ecosystem, and regulatory responses that can better keep pace with technological change. FF-CAIG is designed to be prospective, adaptive, and applicable across organizational governance, regulatory policy, and clinical practice. FF-CAIG is intended for clinical AI systems whose deployment may create clinically, organizationally, ethically, or relationally consequential governance challenges. Its scope includes high-risk clinical decision-support systems, foundation model–based clinical tools, patient-facing communication or triage systems, and administrative or documentation tools that meet this threshold. It is not intended for low-impact administrative automation with no plausible effect on clinical decisions, patient-facing communication, safety, equity, or professional accountability; governance intensity should scale with autonomy, clinical consequences, deployment context, and horizon classification. We argue that it offers both a conceptual contribution to health informatics scholarship and a practical tool for regulators, health system leaders, professional bodies, and clinical AI developers seeking more anticipatory clinical AI governance under technological uncertainty.

FF-CAIG is not intended as a simple aggregation of established foresight methods. Rather, it is designed as an integrated governance framework in which each component addresses a distinct limitation of oversight models that primarily emphasize current-state performance, compliance, and post hoc risk management in clinical AI [10,15]. These 3 methods were selected because they address complementary needs in anticipatory clinical AI governance. The three horizons (3H) model supports temporal classification of risks across near-term, transitional, and longer-term phases; scenario planning stress-tests governance priorities under regulatory and institutional uncertainty; and causal layered analysis (CLA) examines the structural assumptions, professional role expectations, institutional incentives, and cultural narratives that shape which risks are recognized and how they are governed [19,23]. Other approaches remain useful for adjacent governance functions: horizon scanning can support evidence gathering, Delphi consensus can support expert agreement, technology road mapping can support implementation planning, risk matrices can support risk scoring, and safety-case methods can support assurance documentation. However, no single method by itself connects temporal transition, policy robustness, and deeper sociotechnical analysis in a way that directly supports clinical AI governance. FF-CAIG integrates these functions into a single analytic framework (Figure 1).

XSL•FO
RenderX

Figure 1. FF-CAIG: integrated analytical cycle for clinical AI governance. AI: artificial intelligence; FF-CAIG: futures framework for clinical artificial intelligence governance.

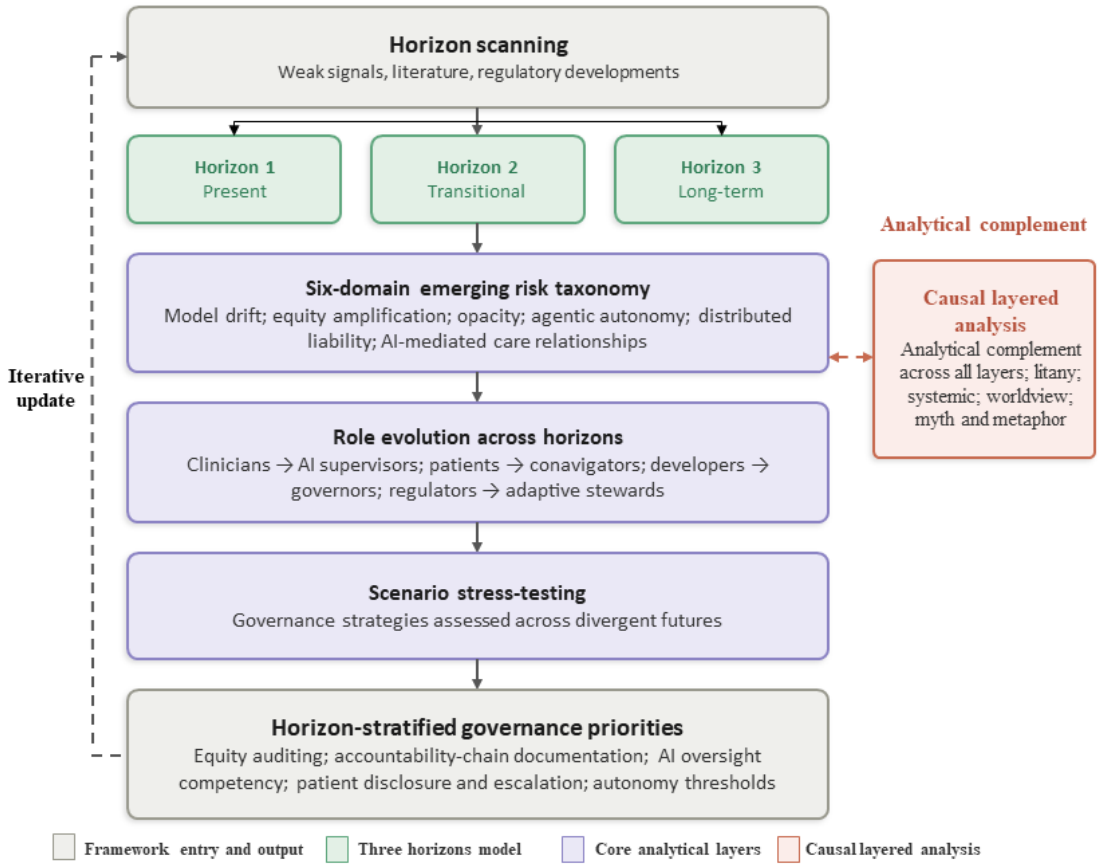

The Limitations of Present-Oriented Clinical AI Governance

Overview

Current approaches to clinical AI governance are increasingly incorporating life cycle–oriented elements, including postmarket monitoring, change-control planning, and ongoing obligations for high-risk AI systems [3,4,7,9]. These developments are important and should not be characterized as purely static or reactive. However, most frameworks remain more developed for current-state evaluation, foreseeable updates, and retrospective risk detection than for anticipating longer-horizon changes in clinical roles, institutional accountability, patient relationships, and regulatory adaptation [10,15]. This creates a growing mismatch with the dynamic, distributed, and sociotechnical character of clinical AI in practice.

Static Approval Models for Dynamic Systems

A central limitation of existing governance frameworks is their reliance on approval models developed for relatively stable technologies, structured around defined intended use, bounded functionality, and evaluation at discrete time points [10,11]. Although this remains useful for narrow and well-characterized applications, it is increasingly strained when applied to updateable, adaptive, or implementation-sensitive AI systems [1,8]. The FDA's predetermined change control plan framework is an important advance because it recognizes that AI-enabled device software may change after authorization [9]. However, it remains primarily oriented toward planned and bounded change, making it less well suited to emergent behavioral shifts in foundation model–based systems or layered AI ecosystems shaped by model revision and local implementation [24]. In parallel, the European Union's AI governance architecture introduces conformity assessment and postmarket obligations for high-risk systems [4,7], but practical questions remain about defining material change, monitoring continuously updated systems, and allocating ongoing oversight responsibility.

Postmarket Surveillance Gaps

A further limitation concerns how postdeployment harms are detected. Conventional postmarket surveillance was designed around discrete, reportable adverse events; yet, the performance of clinical AI may erode gradually as data, populations, and workflows shift after deployment [26,27]. Episodic, incident-triggered reporting is therefore poorly suited to detecting progressive deterioration, which may persist for an extended period before surfacing as an identifiable safety signal [8,10,12]. As clinical AI becomes more embedded in routine care, governance will need to move beyond incident reporting toward continuous, context-sensitive, and performance-aware monitoring [9].

XSL•FO
RenderX

Accountability Fragmentation

Clinical AI systems are increasingly developed, adapted, integrated, and configured by multiple actors across the system life cycle [1,8,28]. This distributed architecture can make accountability difficult to locate within conventional governance and liability models, particularly when harm arises from interactions among model design, local implementation, workflow integration, and clinical use [15-17]. This limitation creates the need for more explicit accountability allocation, which is addressed later as a core risk domain within FF-CAIG.

Equity as a Core Governance Requirement

Algorithmic bias is among the most widely recognized risks in clinical AI [29,30]; yet, equity is still often treated as secondary rather than as a core governance requirement [18,25]. Existing frameworks rarely provide operational standards for subgroup performance reporting, proxy variable auditing, or postdeployment equity surveillance [31]. The mechanisms and life cycle requirements of equity-oriented governance are examined later as a core risk domain within FF-CAIG; at this stage, the key point is that the absence of such standards leaves a structural gap that conventional oversight is poorly positioned to close. Taken together, these limitations expose a deeper mismatch between conventional oversight and the temporal, relational, and system-level nature of contemporary clinical AI.

A Futures Methodology for Clinical AI Governance: The 3H Model

Overview

Among futures methods, the 3H model offers a useful way to analyze governance during system transition [23]. Instead of treating change as a simple linear progression, it frames transformation as unfolding across 3 overlapping horizons. Horizon 1 (H1) is the currently dominant system, still operationally effective but increasingly marked by structural limitations. Horizon 2 (H2) is the transitional space in which emerging innovations, hybrid arrangements, and competing institutional logics begin to challenge H1 assumptions. Horizon 3 (H3) refers to longer-term system configurations whose values, structures, and practices may initially appear peripheral but can ultimately redefine the dominant order.

Relevance to Clinical AI Governance

Applied to clinical AI governance, the 3H model offers several advantages over linear forecasting. It recognizes that current governance structures will remain important even as their limitations become more visible, creates analytical space for examining transitional and longer-horizon challenges before they are fully realized [21,22], and links technological change to shifts in professional roles, accountability structures, and care relationships [15,19]. In this sense, the 3H model is not simply a way of organizing time but a framework for understanding how governance logics evolve during sociotechnical transition. This perspective is especially relevant in clinical AI. Recent developments, including the FDA's predetermined change control plan guidance [9] and the EU AI Act's postmarket obligations [4,7], represent important advances, but they do not eliminate the need for anticipatory governance [10,24]. As clinical AI moves from relatively bounded tools toward more complex, layered, and adaptive ecosystems, governance assumptions may be challenged differently across successive horizons [32]. Table 1 applies the 3H model to clinical AI governance by organizing plausible conditions across near-term, transitional, and longer-term horizons.

Table 1. Three horizons model applied to clinical artificial intelligence (AI) governance.

Domain	Horizon 1: present system	Horizon 2: transitional emergence	Horizon 3: longer-term reconfiguration
Temporal orientation	Near term	Transitional period	Longer term
Dominant paradigm	Point-of-care AI tools, human-in-the-loop support, device-centered regulation	Increasing task delegation, workflow-embedded foundation-model capabilities, and iterative system updates	More autonomous, AI-mediated care infrastructures and cross-pathway coordination
Primary governance challenge	Premarket validation, transparency, bias and performance assessment, and fit within existing regulatory pathways	Oversight of iteratively updated systems, layered accountability, workflow dependence, and role transition	Responsibility, consent, disclosure, and professional accountability in low-human-intervention or AI-mediated care

Governance Implications of the 3H Model

Governance investments focused only on H1 realities, such as device classification, one-time approval logic, and narrow predeployment evaluation, may be insufficient for H2 conditions characterized by iterative updating, layered accountability, and partial delegation [10,15]. They are even less suited to H3 settings in which clinical AI functions less as a bounded tool and more as a persistent mediating infrastructure across episodes of care [32]. Governance should therefore be designed across all 3 horizons simultaneously, even when later horizons remain only partially formed [19,20]. The 3H model thus reframes clinical AI governance as a problem of system transition rather than isolated product oversight. It helps regulators, health systems, and professional bodies identify where current mechanisms remain adequate, where hybrid approaches are needed, and where new forms of oversight may become necessary [19,21,22]. The goal is not speculative prediction but anticipatory preparedness.

XSL•FO
RenderX

An Emerging Risk Taxonomy for Clinical AI Governance

Overview

Drawing on a targeted narrative review and literature-informed conceptual synthesis of clinical AI governance literature, regulatory and policy documents, and analogous high-stakes AI domains, we identify 6 risk domains that warrant anticipatory attention in clinical AI governance (Table 2) [8,15,18,31]. Because this paper is a viewpoint, FF-CAIG was developed through a structured synthesis rather than a preregistered protocol, formal expert consensus process, or empirical coding study. The synthesis involved targeted source review, identification of recurring governance concerns, iterative grouping into risk domains, mapping to futures methods, and assessment of links to practical governance outputs. Sources were prioritized when they addressed risks relevant to clinical decision-making, patient safety, institutional accountability, professional responsibility, equity, trust, or governance design, not technical performance alone. Recurring concerns were iteratively compared and grouped according to their dominant governance gap: technical reliability, structural accountability, or relational quality. Domains were retained when they were recurrent across sources, relevant to multiple stakeholders, applicable across different clinical AI deployment patterns, and linked to actionable governance responses. The taxonomy is intended to support anticipatory governance by distinguishing risks that are already visible, emerging, or likely to become more consequential over time. It is therefore presented as a viewpoint-based analytic framework, not as a systematic review–derived, empirically coded, or exhaustive classification.

In this taxonomy, H1 denotes risks already observable in current clinical AI deployments; H2 denotes risks that become more consequential, as systems become iteratively updated, workflow-embedded, foundation model–based, or partially delegated; and H3 denotes longer-term risks associated with more autonomous, persistent, or AI-mediated care infrastructures.

Table 2. Emerging risk taxonomy and horizon placement for clinical artificial intelligence (AI) governance.

Risk domain	Illustrative manifestation	Horizon (H)	Governance gap	Rationale for horizon placement
Model drift and silent failure	Model performance declines as populations, data pipelines, or workflows change	H1-H2	Inconsistent postmarket surveillance and drift response	Already observable in current systems (H1); harder to detect, as systems become iterative and workflow-embedded (H2)
Algorithmic equity amplification	Nonrepresentative data lead to poorer performance for underserved groups	H1-H3	Limited equity auditing and subgroup performance standards	Documented in current systems (H1); may scale with foundation models (H2); may become structurally embedded without life cycle equity governance (H3)
Opacity and explainability	Clinicians cannot explain AI recommendations or justify reliance on or overrides of AI outputs	H1-H3	Weak explainability, disclosure, and consent expectations	Present in current tools (H1); intensified by foundation models (H2); more consequential, as AI mediates longitudinal care (H3)
Agentic autonomy	AI performs multistep tasks, such as triage or protocol execution, with limited real-time human review	H2-H3	Unclear oversight standards, autonomy thresholds, and escalation rules	Primarily an emerging risk beyond most current human-supervised deployments; becomes more salient as delegation increases (H2-H3)
Distributed liability	Harm arises across developers, integrators, deployers, and clinicians without clear responsibility	H1-H3	Poor accountability allocation across sociotechnical systems	Relatively clearer for bounded tools (H1); harder to assign, as multiactor and longitudinal workflows expand (H2-H3)
AI-mediated care relationships	Longitudinal AI tools reshape communication, trust, and continuity of care	H2-H3	Limited standards for relational quality, continuity, and access to human review	Early forms are emerging in patient-facing tools (H2); consequences increase when AI persistently mediates care (H3)

Model Drift and Silent Failure

As noted earlier, clinical AI may become less reliable after deployment, and this deterioration often develops gradually rather than as a discrete failure event. The result is the risk of silent failure: a system continues to operate within its nominal range while its clinical usefulness declines and the problem remains undetected [12,26,27]. Existing governance frameworks remain inconsistent in how they define drift, monitor performance change, or trigger remedial action [8,10]. A prospective governance approach should therefore specify minimum standards for drift detection, reporting cadence, cohort-disaggregated monitoring, and response protocols that activate when clinically meaningful deterioration is detected [9].

Algorithmic Equity Amplification

The risk that clinical AI may amplify existing inequities is well established [29,30]. These inequities may arise through underrepresentation of historically marginalized groups in training data, proxy variables that encode racial or socioeconomic disadvantage [30], and optimization targets that

XSL•FO
RenderX

are misaligned with equity goals. They may also be reinforced by deployment contexts that concentrate advanced tools in already well-resourced institutions [31]. As more general-purpose and foundation model–based systems are deployed across heterogeneous populations, these mechanisms may operate at greater scale and become harder to attribute [33,34]. A futures-oriented governance approach should therefore move beyond documenting disparities after deployment and instead define minimum equity-governance requirements across the system life cycle.

At minimum, high-impact clinical AI systems should undergo subgroup-disaggregated evaluation before deployment, periodic postdeployment reassessment after major updates or material workflow changes, and formal review and remediation planning when subgroup performance falls below prespecified, context-specific clinical acceptability thresholds [9,18,29,31]. These thresholds should be defined prospectively according to the system's intended clinical function, risk level, and potential consequences of error and may include minimum subgroup performance levels, calibration requirements, maximum allowable subgroup performance differences, and minimum sample-size or confidence requirements for reliable subgroup estimates. Governance should also specify which clinically relevant and equity-relevant subgroup variables are monitored, who is responsible for reviewing disparity signals, and what remedial actions are triggered when inequitable performance is detected. Remediation pathways should be proportionate to the severity and persistence of the disparity and may include governance review, root-cause analysis, enhanced human review, temporary restriction of use, recalibration, retraining, workflow redesign, revalidation, or suspension when safety or equity criteria are breached. In this framework, equity is not only an ethical principle but also an auditable governance function [18,25,31].

Opacity and Explainability

Foundation model–based and other complex clinical AI systems pose governance challenges that differ from those of more narrowly bounded tools because they often have broader capabilities, more variable downstream uses, and behaviors that are harder to characterize before deployment [34,35]. Under these conditions, clinicians may be unable to explain AI recommendations to patients, justify overrides in the medical record, or determine whether a model's output reflects clinically relevant reasoning or spurious correlations [10,24]. This creates governance pressure not only around technical explainability but also around disclosure, informed consent, documentation expectations, and professional accountability. Governance frameworks should therefore define minimum expectations for explainability proportional to clinical risk, including requirements for user-facing disclosure, justification of override or reliance in high-impact settings, and documentation standards for AI-supported decision-making [10,34,35].

Agentic Autonomy

Increasingly, agentic clinical AI presents a major governance challenge on the medium- to longer-term horizon. These systems may support multistep activities such as triage, care coordination, chronic disease management, or protocolized follow-up [2,28]. Although current deployments remain substantially human-supervised, ongoing development trajectories suggest that broader action spaces and more consequential forms of delegated decision support are likely to become more salient over the medium- to longer-term horizon [32]. Existing governance frameworks do not clearly define how human oversight should scale with autonomy, what level of review remains meaningful, or how accountability should be allocated when AI-supported action contributes to patient harm [8,17]. Governance should therefore move beyond a simple human-in-the-loop requirement toward clearer standards for autonomy thresholds, escalation rules, and minimum conditions for meaningful human control [36].

Distributed Liability in Multicomponent Systems

Accountability in clinical AI is difficult not only because some systems are opaque but also because responsibility is distributed across the development, integration, deployment, and use pathway [1,8]. Patient harm associated with AI-supported care may arise from model development, platform integration, local implementation, and downstream clinical oversight, without any single actor bearing sole responsibility [17]. Conventional medical liability frameworks are poorly suited to this layered architecture. Accountability should therefore be addressed prospectively, with responsibilities specified before deployment and not only after harm occurs [15,17]. One practical approach is to require accountability-chain documentation as a condition of deployment authorization [8,15,37]. This documentation should identify the responsible actor for model development, integration, local validation, routine monitoring, incident review, user training, and authority to suspend use, together with handoff points between actors, reporting obligations, and escalation routes when clinically meaningful deterioration or harm is detected. In this way, accountability is translated from an abstract governance principle into a deployable institutional requirement.

AI-Mediated Care Relationships

Clinical AI may increasingly mediate communication, guidance, and continuity across patient journeys, particularly in navigation, triage, symptom support, and longitudinal follow-up [16,17]. In these settings, governance concerns extend beyond safety and performance to the quality of AI-mediated relationships, including trust, continuity, dependence, and the conditions under which patients can access meaningful human interaction. Existing governance frameworks remain limited in how they address relational quality, continuity safeguards, or the risk that AI-mediated support may alter help-seeking, disclosure, or perceptions of professional care. Governance should therefore address not only what these systems do but also how they reshape communication and care relationships over time. Minimum safeguards may include clear disclosure of AI mediation, defined routes to human escalation, continuity protections for higher-risk or vulnerable patients, and review of relational impacts in longitudinal deployments.

XSL•FO
RenderX

Shifting Roles Across the Clinical AI Governance Landscape

Overview

Futures methodology attends not only to technological change but also to the shifting human and institutional roles that accompany system transition. As clinical AI evolves across horizons, the roles of clinicians, patients, developers, operators, and regulators will also change, with direct governance consequences. Frameworks that fail to anticipate these shifts risk creating accountability gaps, competency mismatches, and consent structures that no longer reflect the actual distribution of agency in AI-enabled care.

Clinicians: From Practitioners to AI Supervisors

In H1 settings, the clinician's role remains broadly recognizable: AI provides a recommendation, and the clinician accepts, modifies, or overrides it using independent professional judgment [2,12]. As AI outputs become more numerous, complex, and time-sensitive, however, the conditions for meaningful oversight may erode, as earlier work on alert fatigue in clinical decision support suggests [13,14,38]. In H2 and beyond, the clinician's role may shift from direct execution of care toward supervision of AI-enabled processes, requiring competencies not central to traditional clinical practice, including interpretation of outputs in light of known limitations, detection of behavioral drift, and communication of uncertainty to patients [32,39]. Current professional education and credentialing structures are not yet aligned with this shift [40]. Governance should therefore recognize AI oversight competency as a core component of professional preparation, clinical privileging, and continuing education.

Patients: From Recipients to Conavigators

Patients have received less attention than clinicians or regulators in clinical AI governance, and in many current deployments, they remain passive recipients of AI-mediated recommendations, often without clear disclosure of AI involvement [6,25,41]. Informed consent frameworks also remain underdeveloped, with no settled standard for what patients should be told about AI's role, limitations, or variable performance across populations [16,42]. In longer-horizon settings, patients may interact directly and longitudinally with AI-enabled navigation, triage, or communication tools, raising governance questions about continuity, therapeutic alliance, the right to meaningful human contact, and the boundary between delegation and substitution [43]. Patient rights in AI-mediated care should therefore be treated as a core governance domain rather than a narrow extension of disclosure requirements [18].

Developers and Operators: From Innovators to Governors

Clinical AI governance increasingly depends on the actions of developers and deployers well beyond model creation [1,8]. The EU AI Act's distinction between providers and deployers reflects this shift by assigning obligations to both those who place AI systems on the market and those who implement them in operational settings [4,7]. In health care, governance must therefore extend across the deployment life cycle and not stop at technical development alone [10,15]. Developers should disclose key features of training data provenance, known performance limitations, and relevant subgroup performance where feasible [25,29]. Deployers and operating institutions should document deployment context, local validation, monitoring procedures, and escalation mechanisms for clinically meaningful deterioration or adverse events [8,37]. The governance principle is clear: stewardship responsibilities should follow the system through its life cycle rather than end at market entry [15,21].

Regulators: From Gatekeepers to Adaptive Stewards

The traditional regulatory model, in which regulators primarily act as gatekeepers evaluating products against predefined standards [3,4], is increasingly insufficient for systems whose behavior, use context, and organizational effects may change after deployment [10,24]. A futures-oriented approach suggests that regulators will also need adaptive stewardship capacities, including monitoring population-level performance, updating expectations in response to postdeployment evidence, and convening multistakeholder processes when new governance issues emerge [21,22]. International coordination will be especially important for systems operating across jurisdictions [5], and bodies such as the International Medical Device Regulators Forum may provide useful mechanisms for convergence in classification, postmarket monitoring, and life cycle oversight of AI-enabled medical technologies [44].

Scenario Planning for Clinical AI Regulatory Design

Scenario planning offers a useful method for exploring divergent futures without requiring deterministic prediction [45]. It enables governance designers to examine how technological, institutional, and regulatory conditions may interact and to stress-test policies against a small number of coherent but meaningfully distinct scenarios without assuming a single expected trajectory [19,21].

Scenario Structure and Governance Logic

In clinical AI governance, scenario planning is especially valuable because uncertainty arises not only from the pace of technological development but also from the coherence and adaptability of regulatory response [21,22]. FF-CAIG uses 2 dimensions: the pace of clinical AI capability development, from incremental to rapid and broadly integrated, and the coherence of regulatory response, from fragmented and reactive to coordinated and anticipatory [45]. On this basis, we focus on 3 analytically salient futures: governed acceleration, fragmented proliferation, and regulatory retrenchment (Table 3). The purpose is not to predict which future will occur, but to identify governance choices that remain robust across plausible trajectories [45].

XSL•FO
RenderX

Table 3. Scenario matrix for clinical artificial intelligence (AI) regulatory futures.

Domain	Scenario A: governed acceleration	Scenario B: fragmented proliferation	Scenario C: regulatory retrenchment
Regulatory environment	Adaptive and increasingly harmonized frameworks; stronger life cycle oversight and cross-jurisdictional alignment	Patchwork regulation; inconsistent enforcement; uneven institutional governance capacity	Postincident restriction, higher approval thresholds, and defensive compliance culture
Clinician role	AI oversight embedded in training, credentialing, and practice standards	Variable expectations across specialties, settings, and resource levels	Greater risk aversion; reduced willingness to rely on AI; increased documentation and liability burden
Equity outcomes	Equity integrated into validation, procurement, and postdeployment monitoring; routine subgroup reporting	Benefits and monitoring concentrated in well-resourced institutions; disparities widen	Formal equity attention may increase, but access gaps and structural inequities persist
Innovation trajectory	Trust-supported growth under clearer rules and accountability	Rapid but uneven expansion outpaces governance and standardization	Slower innovation; resources shift toward compliance, legal defense, and risk containment

Scenario-Based Stress Testing for Governance Design

Scenario A: Governed Acceleration

In this pathway, innovation is matched by adaptive regulatory frameworks, stronger postdeployment monitoring, and greater alignment across institutions and jurisdictions [21,22]. Public trust is reinforced through transparency, accountability, and meaningful inclusion of clinician and patient perspectives [18,25]. Equity is treated as a core implementation requirement rather than a secondary concern [29,31].

Scenario B: Fragmented Proliferation

In this pathway, clinical AI capabilities expand while governance remains uneven across jurisdictions, organizations, and specialties [5,15]. Adoption proceeds rapidly but inconsistently, with safety protections, clinician support, and postmarket oversight varying by institutional resources and vendor practices [8]. The likely result is a widening disparity in who benefits from clinical AI and under what conditions [31,33].

Scenario C: Regulatory Retrenchment

In this pathway, high-profile AI-related patient safety failures trigger more restrictive control and intensify approval burdens [3]. Although politically understandable, this response may inhibit beneficial innovation without correcting the structural weaknesses that contributed to the precipitating harm [10,24].

Within FF-CAIG, scenario planning functions as a structured stress-testing tool for regulatory proposals, institutional strategies, and professional standards. Life cycle monitoring, clearer accountability allocation, clinician oversight competencies, and routine equity surveillance are likely to remain valuable across all 3 scenarios [2,8,15]. By contrast, governance strategies that assume stable technological trajectories or uniformly responsible deployment are more fragile under fragmented or crisis-driven futures [10,21].

CLA and the Deep Structures of Clinical AI Governance

Overview

CLA offers a useful framework for examining how visible governance problems are sustained by deeper structural, cultural, and narrative conditions [46]. It distinguishes 4 layers of analysis: visible problems, systemic drivers, worldview assumptions, and underlying myths and metaphors [46]. In clinical AI governance, issues such as algorithmic bias, accountability gaps, and opaque decision-making may therefore reflect not only technical failures but also deeper assumptions about clinical knowledge, epistemic authority, and the relationship between algorithmic efficiency and human care [29,42].

Applying CLA to Clinical AI Governance

At the litany level, the visible concern is an AI system that produces biased recommendations, underperforms in certain patient groups, or fails silently after deployment [29,30]. At the systemic level, these failures may be reinforced by incentive structures that privilege aggregate performance over subgroup equity [31], regulatory processes that rely heavily on developer-provided evidence [10], and procurement practices that prioritize efficiency over long-term oversight capacity [1,8]. At the worldview level, governance may be shaped by techno-optimism that equates AI adoption with progress and frames skepticism as resistance to innovation [39]. At the myth and metaphor level, the image of AI as a precision instrument may obscure the extent to which algorithmic outputs are shaped by assumptions embedded in design and deployment [42].

To illustrate how these 4 layers operate together in practice, consider an AI-enabled triage tool that underescalates symptoms reported by a particular subgroup of patients [29]. CLA would examine not only this visible failure but also the systemic factors that produced it, such as underrepresentation in training data, limited subgroup testing, procurement criteria emphasizing throughput, or weak postdeployment monitoring. The analysis would also surface worldview assumptions that AI-enabled triage is inherently efficient and objective, as well as deeper metaphors of AI as a neutral gatekeeper. The resulting governance response would extend beyond model correction to

XSL•FO
RenderX

include procurement standards, subgroup surveillance, escalation pathways, access to human review, and examination of the institutional assumptions that allowed the risk to remain underrecognized.

From Causal Layers to Governance Interventions

The governance value of CLA lies in changing the object of intervention [46,47]. If the problem is framed only at the litany level, the likely response is to add bias audits, incident reporting, or postdeployment monitoring [29,31]. If CLA reveals that the problem is also systemic, governance must extend to procurement criteria, local validation requirements, committee composition, and allocation of monitoring responsibility [8,15,37]. If the problem is partly rooted in worldview assumptions, such as equating AI adoption with modernization or treating skepticism as resistance to innovation, decision-makers may need to revise the criteria by which clinical AI is approved, funded, or scaled [39,47]. If deeper myths and metaphors frame AI as a neutral precision instrument instead of a sociotechnical actor, governance may need to include explicit review of relational effects, epistemic dependency, and institutional power asymmetries [42,47,48]. Within FF-CAIG, CLA therefore helps regulators and health systems decide not only how to govern AI but also which level of governance intervention is required and why [46].

Applying FF-CAIG: An Operational Workflow

FF-CAIG's analytic components yield governance value only when they inform concrete deployment decisions. We therefore propose a 5-step workflow, summarized in Table 4, to translate the framework into a structured review process for AI procurement, deployment authorization, postdeployment reassessment, and regulatory policy design.

In step 1, reviewers characterize the system under review, including its clinical function, data sources, update mechanism, autonomy level, deployment setting, patient-facing role, anticipated clinical consequences, and responsible actors. In step 2, the system is assigned to 1 or more horizons using 4 operational criteria: prevalence in routine care, level of real-time human oversight, fit with existing governance instruments, and observability of associated risks. This classification clarifies which risks are already governable, which require transitional adaptation, and which may fall outside current oversight instruments. In step 3, reviewers apply FF-CAIG's analytic components by identifying implicated risk domains, stress-testing governance responses across plausible scenarios, examining deeper structural and cultural assumptions through CLA, and mapping accountability across actors. In step 4, reviewers consolidate the preceding analyses into a governance review record that specifies monitoring obligations, accountability allocation, disclosure and oversight requirements, equity surveillance expectations, and conditions for modification or suspension. In step 5, reviewers specify minimum actions scaled by horizon: a baseline for H1 systems centered on validation, monitoring, escalation, human fallback, and suspension authority, with additional H2 and H3 safeguards for life cycle monitoring, autonomy limits, patient-facing disclosure, access to human review, and reassessment after material change.

Table 4. Operational workflow for applying the futures framework for clinical artificial intelligence governance.

Workflow step	Core focus	Governance output	Minimum action
1. Characterize the system	Clinical function, data sources, update mechanism, autonomy level, deployment context, patient-facing role, and responsible actors	System profile and deployment context	Document technical, organizational, and relational system characteristics
2. Assign horizon classification	Prevalence in routine care, real-time human oversight, fit with existing governance instruments, and risk observability	H1[a], H2, H3, or multihorizon classification	Assign horizons and record the rationale for horizon placement
3. Apply analytic components	Risk domains, scenario stress-testing, causal layered analysis, and accountability across actors and handoffs	Risk-domain map, scenario findings, causal layered analysis findings, and accountability-chain map	Assess the deployment across risk domains, scenarios, causal layers, and accountability chains
4. Define governance outputs	Monitoring, accountability, disclosure, equity surveillance, and conditions for modification or suspension	Consolidated governance review record	Specify monitoring indicators, responsibility allocation, disclosure requirements, equity surveillance, and pause or suspension criteria
5. Set minimum actions	Horizon-scaled governance floor for deployment or continued use	Deployment authorization, conditional approval, modification requirement, or nondeployment decision	Require H1 baseline safeguards; add life cycle monitoring, autonomy limits, disclosure, human review access, and reassessment for H2-H3 systems

[a]H: horizon.

XSL•FO
RenderX

Applying FF-CAIG to Representative Clinical AI Deployments

Overview

The following 3 representative deployment patterns illustrate how this workflow can be applied across the horizon continuum. These examples are literature-grounded analytic exemplars that also reflect emerging institutional practice; they are not empirical validation exercises or single-site implementation studies. They demonstrate how FF-CAIG can be applied across different governance horizons, risk constellations, and deployment architectures.

Electronic Health Record–Embedded Foundation Model Documentation Assistant (H1)

A large academic medical center deploys a vendor-supplied ambient documentation assistant integrated into the electronic health record. FF-CAIG places this use case primarily in H1, while recognizing emerging H2 risks, as deployment expands across specialties and patient populations. The main governance concerns are model drift and distributed liability across the vendor, electronic health record integrator, deploying institution, and end users [8,15,17]. The governance response is therefore focused on regular disaggregated performance reporting, a named institutional lead for monitoring and escalation, and clinician onboarding on system limitations and override conditions. Scenario stress-testing identifies a key vulnerability: vendor-initiated model updates may alter subgroup performance without timely local detection [26,27]. The resulting H1 governance requirement is a deployment authorization standard that ties continued use to defined monitoring, reporting, and accountability obligations.

Patient-Facing Longitudinal Triage and Communication Tool (H2)

A regional health system introduces an AI-enabled patient navigation tool for longitudinal symptom monitoring, triage, and follow-up messaging between clinic visits. FF-CAIG places this deployment in H2 because the system interacts directly with patients over time while exercising increasing communicative and triage autonomy without routine clinician review. The key risk domains are agentic autonomy, AI-mediated care relationships, and opacity about AI involvement [16,43,47]. Governance priorities, therefore, include defined escalation thresholds, plain-language disclosure of AI use, access to human review on request, and exclusion criteria for patient groups requiring additional safeguards [16]. Stress-testing of expansion into outpatient follow-up and longitudinal care further highlights the need for explicit rules governing use in patients with impaired decision-making capacity.

Autonomous AI-Based Screening Agent (H3)

A health system considers deploying an AI agent that autonomously reviews flagged imaging studies, generates preliminary reports, and triages cases for expedited review before radiologist assessment. FF-CAIG classifies this as H3 because clinically consequential action occurs with limited real-time human oversight [36]. The main governance concerns are agentic autonomy, distributed liability, and equity risk arising from uneven performance across subpopulations [29,31]. The governance response, therefore, requires predefined limits on autonomous action, thresholds for suspension or human fallback, and clear institutional responsibility for monitoring and intervention [8,15]. Stress-testing across affiliated community hospitals shows that validation in an academic center alone is insufficient. Deployment should therefore require site-specific validation, subgroup-based suspension thresholds, and clearly assigned authority to halt use when safety or equity criteria are breached [29,31,36].

Cross-Horizon Governance Priorities

Overview

Drawing on the 3H model, the emerging risk taxonomy, scenario planning, and CLA, FF-CAIG identifies horizon-scaled governance priorities for near-term, transitional, and longer-term clinical AI environments. These priorities do not constitute separate risk categories; rather, they translate the preceding analysis into governance emphases that vary according to system maturity, autonomy, clinical consequences, institutional capacity, and regulatory uncertainty.

Short-Term Priorities (H1)

Near-term priorities should strengthen the governance foundations for currently deployed high-risk clinical AI systems by making them more auditable, locally accountable, and equity-aware. For high-impact systems, the core requirements include predeployment validation proportionate to clinical risk, subgroup-disaggregated performance evaluation, documentation of intended use and known limitations, and clear assignment of responsibility for monitoring, escalation, and suspension [9,17,29,31,35,37]. In parallel, clinician education, continuing professional development, and, where appropriate, credentialing expectations should begin to align with emerging AI oversight roles [40]. These measures address risks already visible in current clinical AI deployments while creating the baseline infrastructure needed for later life cycle oversight.

Medium-Term Priorities (H2)

Transitional priorities should focus on systems that are iteratively updated, workflow-embedded, foundation model–based, or partially delegated [32,34]. In this horizon, governance should move from one-time validation toward life cycle review, with explicit thresholds for clinically meaningful performance change, reassessment after material updates, and standards for meaningful human oversight [9,26,27,36,39]. Patient-facing and workflow-integrated systems also require proportionate safeguards when AI materially shapes triage, communication, or care coordination, including disclosure, escalation pathways, and review mechanisms appropriate to clinical risk [16,42].

Long-Term Priorities (H3)

Longer-term priorities should address clinical AI systems that function as persistent, autonomous, or AI-mediated care infrastructures. Governance in this horizon should define the

XSL•FO
RenderX

conditions under which autonomous or low-human-intervention clinical actions are permissible, how accountability is allocated across distributed systems, and how patients are protected when AI mediates access, communication, or continuity of care [36,39,43,47]. These priorities require stronger institutional and cross-jurisdictional coordination because the relevant risks may exceed the capacity of individual organizations or conventional product-centered oversight [44]. Across horizons, the central governance task is not simply to regulate more AI, but to align governance intensity, accountability structures, and relational safeguards with changing forms of clinical AI capability and deployment.

These priorities are largely convergent, with recommendations already present in the clinical AI governance literature, including life cycle oversight, equity auditing, distributed accountability, and human oversight standards. They should also be interpreted alongside evolving cross-sector and international AI governance instruments, including the EU AI Act [49], the National Institute of Standards and Technology AI Risk Management Framework and its generative AI profile [50,51], and International Organization for Standardization/International Electrotechnical Commission standards on AI management systems and AI risk management [52,53]. FF-CAIG's contribution is therefore not to introduce these priorities as isolated new requirements, but to show how their relative importance shifts across temporal horizons and divergent regulatory futures. By linking governance priorities to emerging risks, shifting roles, regulatory uncertainty, and deeper sociotechnical assumptions, FF-CAIG situates established governance commitments within a single anticipatory framework and clarifies when existing safeguards may remain sufficient, when they require adaptation, and when new forms of governance may become necessary.

Limitations

Several limitations and tensions in anticipatory governance warrant acknowledgment. First, front-loading governance requirements may increase compliance burden, especially for smaller developers, community hospitals, and health systems with limited governance infrastructure. Measures such as site-specific validation, disaggregated performance reporting, and accountability-chain documentation are more easily absorbed by large academic centers and well-resourced technology firms than by safety-net institutions or early-stage developers [8,15,31]. If these requirements become de facto barriers to entry, anticipatory governance may contribute to regulatory consolidation without producing broader safety improvement. FF-CAIG does not resolve this tension and therefore requires jurisdictional and institutional adaptation to avoid becoming an instrument of exclusion.

Second, anticipatory governance carries a risk of overdocumentation and procedural burden. Governance mechanisms designed for H2 and H3 conditions may be premature or disproportionate when applied to current H1 deployments, and prospective planning requirements may divert institutional capacity from direct patient care or quality improvement. A tiered approach that scales governance intensity to the autonomy, scope, and clinical consequences of the AI system may reduce this risk, although operationalizing such proportionality remains an open challenge.

Third, the feasibility of FF-CAIG will vary across health systems and regulatory contexts. The framework draws primarily on regulatory developments in the United States and European Union [4,7,9]. Consequently, some recommendations may be difficult to implement in jurisdictions with less developed AI governance infrastructure, limited health informatics capacity, or different legal frameworks for medical device oversight. Equity in access to anticipatory governance tools is, therefore, itself a governance challenge that the framework does not fully resolve.

Fourth, the deployment patterns used to illustrate FF-CAIG are representative analytic constructs, not empirical case studies. They are intended to demonstrate the framework's interpretive and anticipatory logic across horizons, not to validate its real-world effectiveness. Relatedly, the cross-horizon priorities are derived analytically from the convergence of the framework's temporal and scenario analyses; they have not been formally validated or systematically compared against alternative governance approaches. Future work should validate and refine FF-CAIG and its proposed cross-horizon priorities through modified Delphi consensus, retrospective mapping of documented clinical AI incidents or governance reviews, and prospective pilot-testing within institutional AI governance committees to assess horizon classification agreement, risk identification completeness, accountability clarity, decision documentation, and reviewer-reported usefulness.

Fifth, FF-CAIG was not developed through primary expert consultation, Delphi elicitation, or structured stakeholder interviews, and therefore does not claim direct empirical grounding in the views of regulators, clinicians, patients, developers, or health-system leaders. Stakeholder perspectives informed the framework indirectly through the peer-reviewed studies, regulatory and policy documents, position statements, and international guidance included in the targeted narrative review. However, documentary representation is not a substitute for direct multistakeholder input, particularly for patient perspectives, which remain underrepresented in the clinical AI governance literature. Future Delphi-based validation should therefore examine whether FF-CAIG captures concerns that documentary sources may underweight or omit.

Conclusions

Clinical AI governance is entering a consequential period of transition. Decisions made over the next several years will shape whether clinical AI mitigates or amplifies existing inequities; whether patients retain meaningful rights to disclosure, explanation, and access to human review in AI-mediated care; and whether health systems develop the oversight capacity needed for adaptive and increasingly complex AI systems. FF-CAIG offers a structured approach for moving governance from a predominantly reactive model toward a more anticipatory one. It can help regulators, health system leaders, developers, and researchers identify cross-horizon governance gaps, stress-test governance strategies against divergent futures, and align oversight with the changing capabilities and consequences

XSL•FO
RenderX

of clinical AI. Clinical AI governance should therefore be understood not simply as a technical regulatory task, but as a sociotechnical and institutional challenge requiring prospective, systems-oriented, and stakeholder-engaged approaches. Futures methodologies cannot eliminate uncertainty, but they can support more disciplined preparedness before emerging risks fully crystallize.

Acknowledgments

After drafting the manuscript, the authors used ChatGPT (OpenAI) to improve language clarity. All artificial intelligence–suggested edits were carefully reviewed and revised as necessary. The authors take full responsibility for the final content and integrity of the work.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

Data sharing is not applicable to this paper because no datasets were generated or analyzed during this study.

Authors' Contributions

JL and SL contributed to the conception and design of the viewpoint. YY, JL, and SL conducted the literature analysis and drafted the original manuscript. All authors critically revised the manuscript for important intellectual content, approved the final version, and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

References

1. Novak LL, Russell RG, Garvey K, Patel M, Thomas Craig KJ, Snowdon J, et al. Clinical use of artificial intelligence requires AI-capable organizations. JAMIA Open 2023;6(2):ooad028 [FREE Full text] [doi: 10.1093/jamiaopen/ooad028] [Medline: 37152469]
2. Labkoff S, Oladimeji B, Kannry J, Solomonides A, Leftwich R, Koski E, et al. Toward a responsible future: recommendations for AI-enabled clinical decision support. J Am Med Inform Assoc 2024;31(11):2730-2739. [doi: 10.1093/jamia/ocae209] [Medline: 39325508]
3. Poon EG, Lemak CH, Rojas JC, Guptill J, Classen D. Adoption of artificial intelligence in healthcare: survey of health system priorities, successes, and challenges. J Am Med Inform Assoc 2025 Jul 01;32(7):1093-1100 [FREE Full text] [doi: 10.1093/jamia/ocaf065] [Medline: 40323320]
4. Warraich HJ, Tazbaz T, Califf RM. FDA perspective on the regulation of artificial intelligence in health care and biomedicine. JAMA 2025;333(3):241-247. [doi: 10.1001/jama.2024.21451] [Medline: 39405330]
5. Busch F, Kather JN, Johner C, Moser M, Truhn D, Adams LC, et al. Navigating the European Union Artificial Intelligence Act for healthcare. NPJ Digit Med 2024;7(1):210 [FREE Full text] [doi: 10.1038/s41746-024-01213-6] [Medline: 39134637]
6. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023;25:e48009 [FREE Full text] [doi: 10.2196/48009] [Medline: 37566454]
7. Aboy M, Minssen T, Vayena E. Navigating the EU AI Act: implications for regulated digital medical products. NPJ Digit Med 2024;7(1):237 [FREE Full text] [doi: 10.1038/s41746-024-01232-3] [Medline: 39242831]
8. Sendak M, Liu V, Beecy A, Vidal D, Shaw K, Lifson M, et al. Strengthening the use of artificial intelligence within healthcare delivery organizations: balancing regulatory compliance and patient safety. J Am Med Inform Assoc 2024;31(7):1622-1627. [doi: 10.1093/jamia/ocae119] [Medline: 38767890]
9. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions. US Food and Drug Administration. 2024. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence [accessed 2026-02-26]
10. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): a comparative analysis. Lancet Digit Health 2021;3(3):e195-e203 [FREE Full text] [doi: 10.1016/S2589-7500(20)30292-2] [Medline: 33478929]
11. Solaiman B, Mekki YM, Qadir J, Ghaly M, Abdelkareem M, Al-Ansari A. A "True Lifecycle Approach" towards governing healthcare AI with the GCC as a global governance model. NPJ Digit Med 2025;8(1):337 [FREE Full text] [doi: 10.1038/s41746-025-01614-1] [Medline: 40473925]

XSL•FO
RenderX

12. Khera R, Simon MA, Ross JS. Automation bias and assistive AI: risk of harm from AI-driven clinical decision support. JAMA 2023;330(23):2255-2257. [doi: 10.1001/jama.2023.22557] [Medline: 38112824]
13. Salwei ME, Carayon P, Hoonakker PLT, Hundt AS, Wiegmann D, Pulia M, et al. Workflow integration analysis of a human factors-based clinical decision support in the emergency department. Appl Ergon 2021;97:103498 [FREE Full text] [doi: 10.1016/j.apergo.2021.103498] [Medline: 34182430]
14. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc 2017;24(2):423-431 [FREE Full text] [doi: 10.1093/jamia/ocw105] [Medline: 27516495]
15. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. J Am Med Inform Assoc 2020;27(3):491-497 [FREE Full text] [doi: 10.1093/jamia/ocz192] [Medline: 31682262]
16. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. N Engl J Med 2018;378(11):981-983 [FREE Full text] [doi: 10.1056/NEJMp1714229] [Medline: 29539284]
17. Price WN, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. JAMA 2019;322(18):1765-1766. [doi: 10.1001/jama.2019.15064] [Medline: 31584609]
18. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. World Health Organization. 2025 Mar 25. URL: https://www.who.int/publications/i/item/9789240084759 [accessed 2026-02-10]
19. Cordova-Pozo K, Rouwette EA. Types of scenario planning and their effectiveness: a review of reviews. Futures 2023;149:103153. [doi: 10.1016/j.futures.2023.103153]
20. Coombs C, Hislop D, Taneva SK, Barnard S. The strategic impacts of intelligent automation for knowledge and service work: an interdisciplinary review. J Strateg Inf Syst 2020;29(4):101600. [doi: 10.1016/j.jsis.2020.101600]
21. Tõnurist P, Hanson A. Anticipatory innovation governance: shaping the future through proactive policy making. OECD Working Papers on Public Governance. 2020. URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2020/12/anticipatory-innovation-governance_d1aded4e/cce14d80-en.pdf [accessed 2026-02-16]
22. Foresight approaches in global public health: a practical guide for WHO staff. World Health Organization. 2022. URL: https://www.who.int/publications/i/item/9789240048393 [accessed 2026-02-16]
23. Sharpe B, Hodgson A, Leicester G, Lyon A, Fazey I. Three horizons: a pathways practice for transformation. Ecol Soc 2016;21(2):47. [doi: 10.5751/es-08388-210247]
24. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 06;6(1):120. [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
25. Ning Y, Teixayavong S, Shang Y, Savulescu J, Nagaraj V, Miao D, et al. Generative artificial intelligence and ethical considerations in health care: a scoping review and ethics checklist. Lancet Digit Health 2024;6(11):e848-e856 [FREE Full text] [doi: 10.1016/S2589-7500(24)00143-2] [Medline: 39294061]
26. Nestor B, McDermott MBA, Boag W, Berner G, Naumann T, Hughes MC, et al. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. Proc Mach Learn Res 2019;106:381-405 [FREE Full text]
27. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021;385(3):283-286 [FREE Full text] [doi: 10.1056/NEJMc2104626] [Medline: 34260843]
28. Crowe B, Shah S, Teng D, Ma SP, DeCamp M, Rosenberg EI, et al. Recommendations for clinicians, technologists, and healthcare organizations on the use of generative artificial intelligence in medicine: a position statement from the Society of General Internal Medicine. J Gen Intern Med 2025;40(3):694-702. [doi: 10.1007/s11606-024-09102-0] [Medline: 39531100]
29. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366(6464):447-453 [FREE Full text] [doi: 10.1126/science.aax2342] [Medline: 31649194]
30. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. N Engl J Med 2020;383(9):874-882. [doi: 10.1056/NEJMms2004740] [Medline: 32853499]
31. Celi LA, Cellini J, Charpignon ML, Dee EC, Dernoncourt F, Eber R, et al. Sources of bias in artificial intelligence that perpetuate healthcare disparities—a global review. PLoS Digit Health 2022;1(3):e0000022 [FREE Full text] [doi: 10.1371/journal.pdig.0000022] [Medline: 36812532]
32. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med 2022;28(9):1773-1784. [doi: 10.1038/s41591-022-01981-2] [Medline: 36109635]
33. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
34. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature 2023;616(7956):259-265. [doi: 10.1038/s41586-023-05881-4] [Medline: 37045921]
35. Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, et al. The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit Med 2023;6(1):135 [FREE Full text] [doi: 10.1038/s41746-023-00879-8] [Medline: 37516790]
36. Olawade DB, Plabon SB, Ojo A, Ogunbona MA, Makanjuola BD, Olasilola OR. Human in the loop artificial intelligence in healthcare: applications, outcomes, and implementation challenges. Int J Med Inform 2026;213:106362 [FREE Full text] [doi: 10.1016/j.ijmedinf.2026.106362] [Medline: 41740273]

XSL•FO
RenderX

37. Kim JY, Hasan A, Kueper J, Tang T, Hayes C, Fine B, et al. Establishing organizational AI governance in healthcare: a case study in Canada. NPJ Digit Med 2025;8(1):522 [FREE Full text] [doi: 10.1038/s41746-025-01909-3] [Medline: 40817280]
38. Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, Kaushal R, with the HITEC Investigators. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak 2017;17(1):36 [FREE Full text] [doi: 10.1186/s12911-017-0430-8] [Medline: 28395667]
39. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. N Engl J Med 2016;375(13):1216-1219 [FREE Full text] [doi: 10.1056/NEJMp1606181] [Medline: 27682033]
40. Cao W, Zhang Q, Liu J, Liu S. From agents to governance: essential AI skills for clinicians in the large language model era. J Med Internet Res 2026;28:e86550 [FREE Full text] [doi: 10.2196/86550] [Medline: 41534067]
41. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med 2018;1:39 [FREE Full text] [doi: 10.1038/s41746-018-0040-6] [Medline: 31304320]
42. Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: mapping the debate. Big Data Soc 2016;3(2):205395171667967. [doi: 10.1177/2053951716679679]
43. Campos HDO, Wolfe D, Luan H, Sim I. Generative AI as third agent: large language models and the transformation of the clinician-patient relationship. J Particip Med 2025;17:e68146 [FREE Full text] [doi: 10.2196/68146] [Medline: 40826831]
44. Artificial Intelligence/Machine Learning-enabled Working Group. Good machine learning practice for medical device development: guiding principles. International Medical Device Regulators Forum. 2025. URL: https://www.imdrf.org/documents/good-machine-learning-practice-medical-device-development-guiding-principles [accessed 2026-02-16]
45. Healey MP, Hodgkinson GP. Overcoming strategic persistence: effects of multiple scenario analysis on strategic reorientation. Strateg Manage J 2024;45(8):1423-1445. [doi: 10.1002/smj.3589]
46. Inayatullah S. Causal layered analysis: poststructuralism as method. Futures 1998;30(8):815-829. [doi: 10.1016/s0016-3287(98)00086-x]
47. McCradden MD, Joshi S, Anderson JA, London AJ. A normative framework for artificial intelligence as a sociotechnical system in healthcare. Patterns (N Y) 2023;4(11):100864 [FREE Full text] [doi: 10.1016/j.patter.2023.100864] [Medline: 38035190]
48. Carayon P, Wooldridge A, Hoonakker P, Hundt AS, Kelly MM. SEIPS 3.0: human-centered design of the patient journey for patient safety. Appl Ergon 2020;84:103033. [doi: 10.1016/j.apergo.2019.103033] [Medline: 31987516]
49. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. European Parliament and Council of the European Union. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng [accessed 2026-05-22]
50. Tabassi E. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg, MD: National Institute of Standards and Technology; 2023. [doi: 10.6028/NIST.AI.100-1]
51. Autio C, Schwartz R, Dunietz J, Jain S, Stanley M, Tabassi E. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. Gaithersburg, MD: National Institute of Standards and Technology; 2024. [doi: 10.6028/NIST.AI.600-1]
52. ISO/IEC 42001:2023 information technology—artificial intelligence—management system. International Organization for Standardization, International Electrotechnical Commission. 2023. URL: https://www.iso.org/standard/42001 [accessed 2026-05-22]
53. ISO/IEC 23894:2023 information technology—artificial intelligence—guidance on risk management. International Organization for Standardization, International Electrotechnical Commission. 2023. URL: https://www.iso.org/standard/77304.html [accessed 2026-05-22]

Abbreviations

3H: three horizons
AI: artificial intelligence
CLA: causal layered analysis
FDA: US Food and Drug Administration
FF-CAIG: futures framework for clinical artificial intelligence governance

XSL•FO
RenderX

Edited by T Kristóf, B Mesko; submitted 26.Mar.2026; peer-reviewed by A Asadollahi, R Izadi; comments to author 20.May.2026; revised version received 25.May.2026; accepted 29.May.2026; published 29.Jun.2026.

Please cite as:
Yang Y, Liu J, Liu S
A Futures Framework for Clinical AI Governance: Anticipating Emerging Risks, Shifting Roles, and Regulatory Challenges
J Med Internet Res 2026;28:e96152
URL: https://www.jmir.org/2026/1/e96152
doi:10.2196/96152
PMID:

©Yi Yang, Jialin Liu, Siru Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org

XSL•FO
RenderX